Soutenance de thèse de Thomas Denecker

La soutenance va bientôt commencer

La présentation est disponible en ligne

https://thomasdenecker.github.io/thesisWebsite/

https://fr.slideshare.net/ThomasDENECKER

Thèse de doctorat

université PARIS-SACLAY

Bioinformatique et analyse de données multiomiques :

Principes et applications chez les levures pathogènes *Candida glabrata* et *Candida albicans*

Thomas DENECKER

Sous la direction de Gaëlle LELANDAIS

Thèse de doctorat de l'université Paris-Saclay présentée et soutenue à Orsay, le 16/09/2020

École doctorale n°577

Structure et dynamique des systèmes vivants (SDSV)

Spécialité de doctorat : sciences de la vie et de la santé

Unité de recherche : Institut de Biologie Intégrative de la Cellule

Référent : Faculté des sciences

INTRODUCTION

Évolution des objets d'étude en Bioinformatique

Période 2000 – 2005
(Thèse G. Lelandais)

Période 2015 – 2020
(Thèse T. Denecker)

Bioinformatics - BMC Bioinformatics - Brefings in Bioinformatics - Journal of bioinformatics and computational biology

Vous avez dit « Data / Donnée » ?

« *Un élément brut qui n'a pas encore été interprété, mis en contexte* » (Chaudet 2009)

« *Collectée par observations* » (*Glossary of statistical terms*)

Données structurées

Généralement organisées dans une base de données (GEO, SRA, Pride,…)

	logFC 1	…	*logFC m*
Gène 1	2.05	…	1.85
Gène 2	1.85	…	0.57
Gène 3	0.02	…	-0.06
…	…	…	…
Gène n	-3.59	…	-2.46

Données non structurées

Plus complexes et à traiter pour les organiser

Exemple : description du gène *FET3* dans la SGD

"Ferro-O2-oxidoreductase; multicopper oxidase that oxidizes ferrous (Fe^{2+}) to ferric iron (Fe^{3+}) ***for subsequent cellular uptake by transmembrane permease Ftr1p****; […]"*

Une observation sans interprétation

La différence entre donnée et information

Une information est une donnée associée à une interprétation

Condition A

	Replicat 1	*Replicat 2*	*Replicat 3*
Gène 1			
Gène 2			
…	…	…	…
Gène n			

Condition B

	Replicat 1	*Replicat 2*	*Replicat 3*
Gène 1			
Gène 2			
…	…	…	…
Gène n			

Analyse différentielle

	logFC	Valeur P
Gène 1	2.05	…
Gène 2	0.35	…
…	…	…
Gène n	-3.59	…

Données

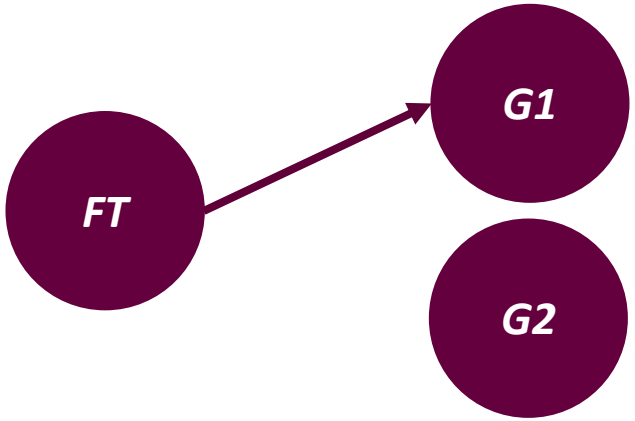
Informations

L’objectif final : la connaissance

« Information comprise, c’est-à-dire assimilée et utilisée qui permet d’aboutir à une action » (Chaudet 2009).

Connaissance explicite

Formalisée et transmissible sous forme de documents réutilisables

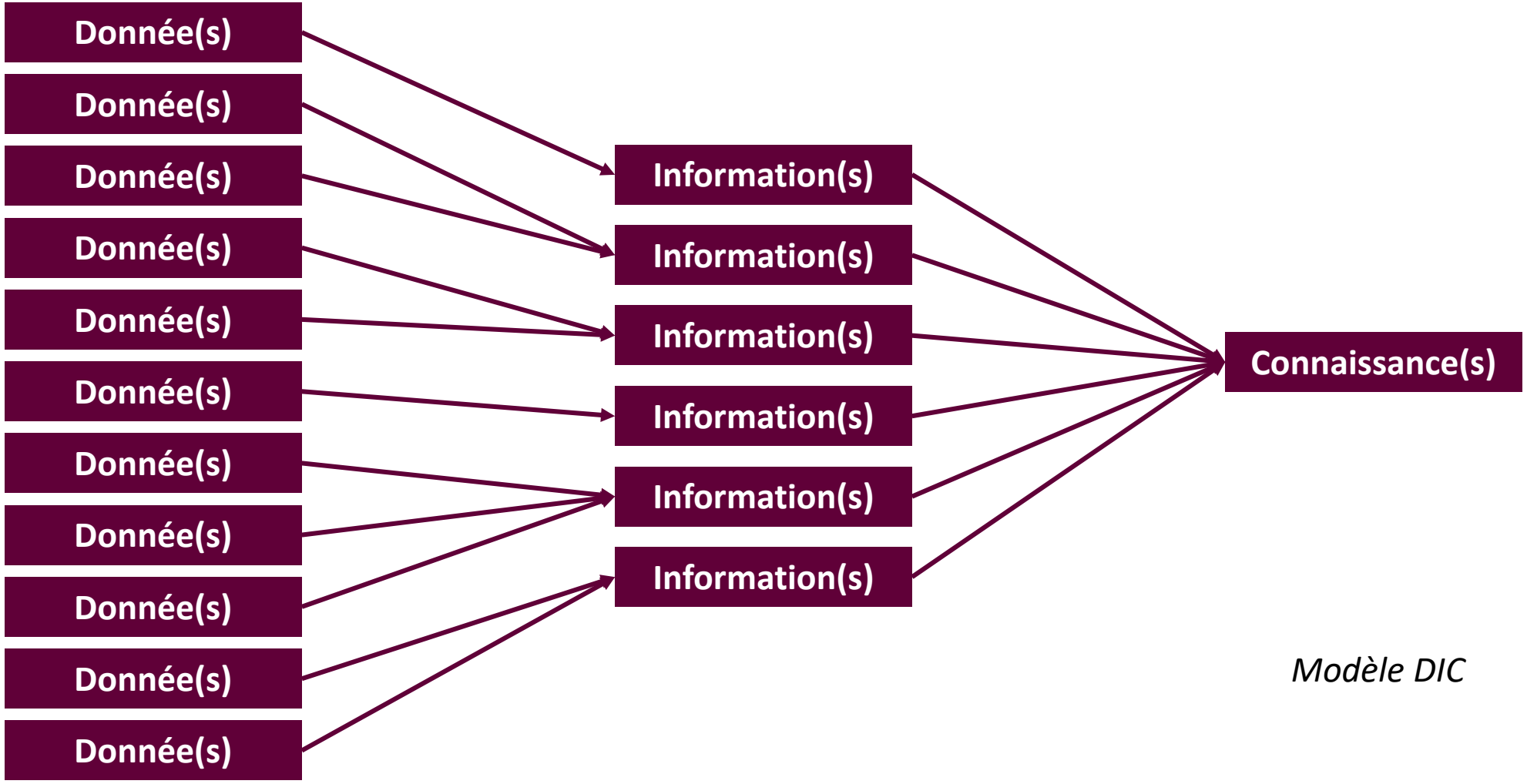


Connaissance tacite

Non formalisée et difficilement transmissible

« Pourquoi utilises-tu cette méthode de *clustering* ? »
« Parce que c’est celle qui donne les meilleurs résultats »

Données, informations, connaissances



Modèle DIC

Collecter → Organiser → Résumer → Analyser → Synthétiser → Décider

Analyse de données ?

Transition entre données, informations et connaissances

« *Processus d'inspection, de nettoyage, de transformation et de modélisation des données, dans le but de découvrir des informations utiles, d'éclairer la conclusion et d'appuyer la prise de décision* » (Wikipédia)

Processus cyclique en 6 grandes étapes

(Peck et al, 2016)

1. Formulation de la question scientifique
2. Recherche et collecte des données
3. Préparation des données
4. Exploration et analyses préliminaires
5. Formulation d'hypothèses statistiques
6. Interprétation et conclusion

Formulation de la question scientifique

Étape clé pour le déroulement complet du cycle d'analyse

Poser une question précise et explicite pour créer un type d'information

(Ne pas viser immédiatement la connaissance)

Exemple : « *Quels sont les gènes différentiellement exprimés entre les conditions A et B ?* »

(≠ « Comment la cellule s'adapte au changement de condition A vers B ? »)

Recherche et collecte des jeux de données

Nombreuses données disponibles librement ou disponibles dans les équipes expérimentales

Utiliser seulement les données nécessaires pour répondre à la question

(d'autant plus facile qu'elle est précise et explicite)

Problématique des données structurées ou non

Préparation des données

Plus ou moins importante

Fastidieuse mais essentielle

En moyenne, 60 % du temps d'une analyse de données

(CrowdFlower 2016)

Exploration et analyses préliminaires

Faire connaissance avec les données

« Quick and dirty »

(R. Peng)

Réalisation de nombreux graphiques, de calculs descriptifs, … le plus facilement possible

Formulation d'hypothèses statistiques

Mise en place d'un plan d'analyse
(méthodes, tests, …)

Rigoureuse et bien documentée

Par exemple « *La fonction F est-elle plus représentée dans la liste de gènes qu'attendue par le hasard ?* »

Interprétation et conclusion

Mise en forme des résultats, rédaction de rapports et réalisation d'infographies

Importance d'une expertise dans le domaine scientifique

De nouveaux questionnements scientifiques ?

Problématiques liées à l'analyse de données

3 problématiques principales rencontrées lors de la thèse

Choix des données

Face à un déluge de données

Big Data

Faut-il toujours plus de données ? Oui mais …

Hétérogénéité des données, de la qualité des données, de l'annotation, etc.

Reproductibilité

Nouvelles problématiques
(informatique, bioinformatique, biologie)

Nouvelles pratiques

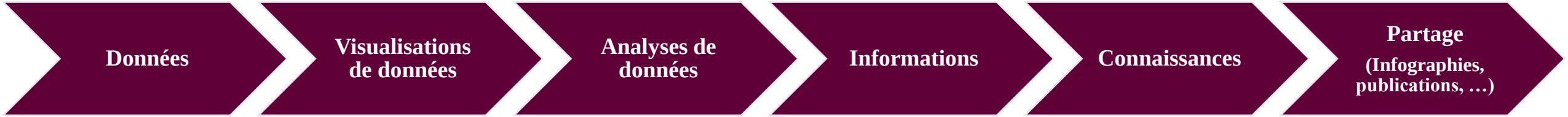
FAIR bioinfo

Représentations des données

Visualisation
Procédure exploratoire

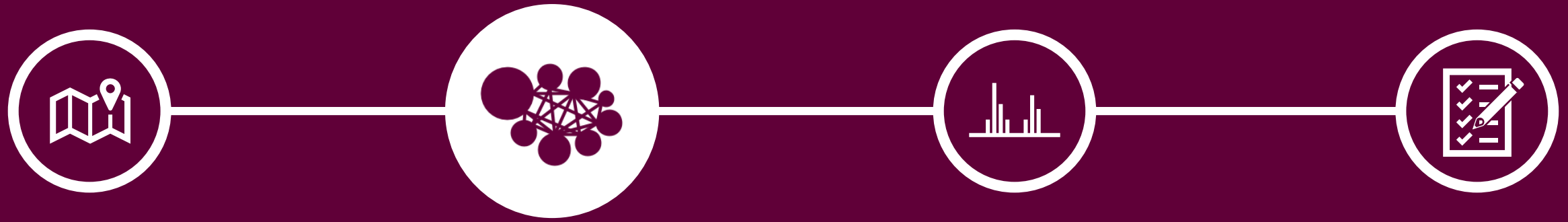
Infographie
Objectif de synthèse et vulgarisation des connaissances

En résumé



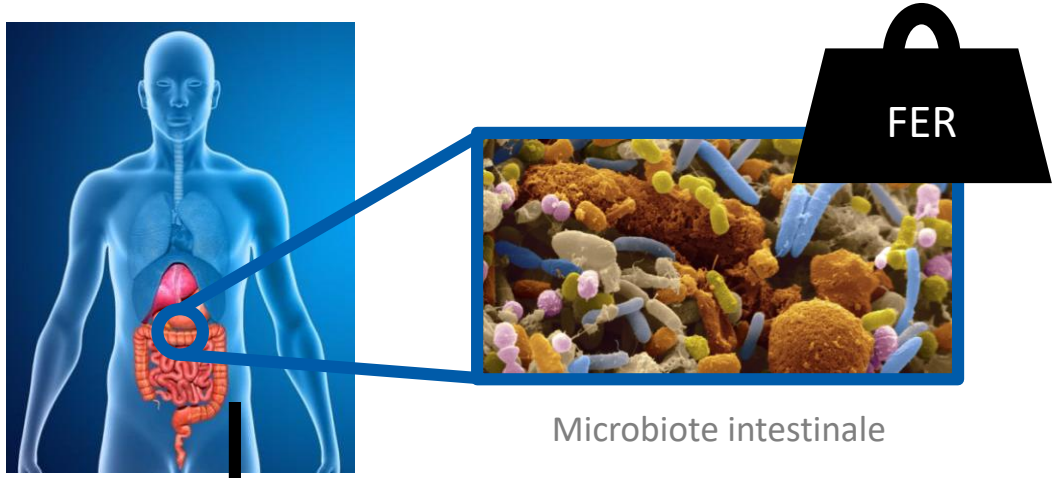
En pratique

Projet 1	Étude de l'homéostasie du fer chez la levure pathogène *Candida glabrata*
Projet 2	Étude de l'impact de la prise en compte systématique des modifications post-traductionnelles lors de l'identification de protéines chez la levure pathogène *Candida albicans*

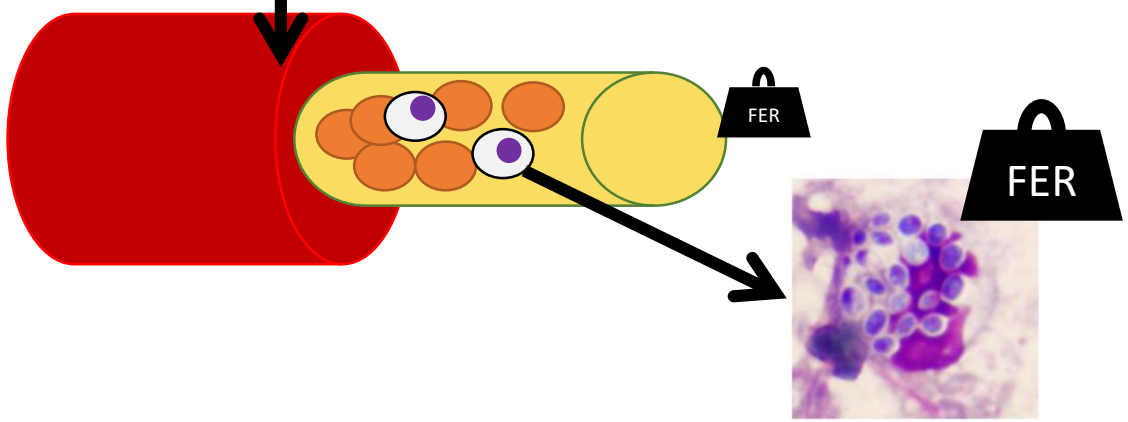


ÉTUDE DE L'HOMÉOSTASIE DU FER CHEZ LA LEVURE PATHOGÈNE *CANDIDA GLABRATA*

Le fer est un élément indispensable aux organismes vivants



Microbiote intestinale



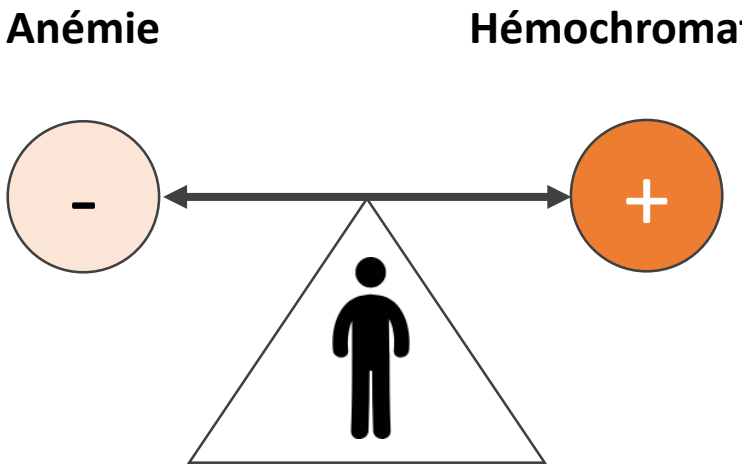
Macrophage avec des *C. glabrata*
Crédit : A. Angoulvant

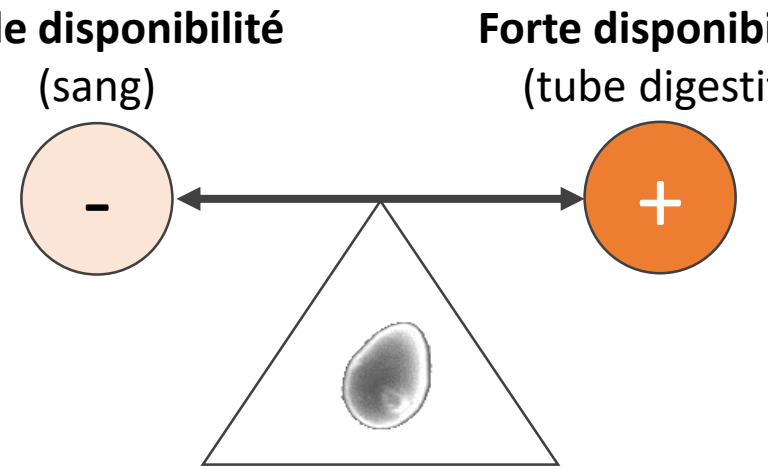
Le fer est un composé essentiel de la **relation hôte / micro-organismes**

Mécanisme de défense de l'hôte **privation du fer**

Stratégies originales pour adapter leur métabolisme à des conditions de vie dans des **environnements pauvres en fer**

Un équilibre complexe à trouver





Homéostasie du fer

Maintien d'un environnement interne dans un état d'équilibre constant, malgré les changements externes

Mécanismes génomiques

Processus de régulation

Carte d'identité de *Candida glabrata*

13 chromosomes - 5293 ORFs - Haploïde
(CGD Genome Snapshot CBS138, Février 2020)

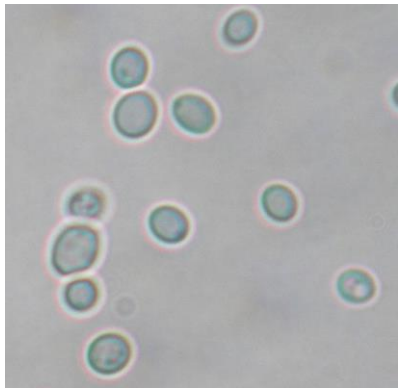
Présent dans la flore commensale
Cavité buccale ou des tractus gastrointestinal et urogénital
(Underhill et al. 2014; Cho et al. 2012; Cui et al. 2013)

Pathogène opportuniste

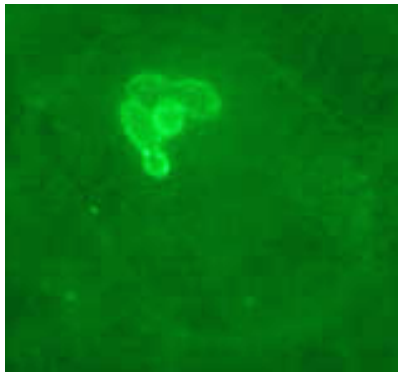
Cause majeure de morbidité et de mortalité dans les structures de soins
(Pfaller *et al*, 2012)

Touche principalement des patients immunodéprimés (cancer, transplantation,…)
(Pfaller *et al*, 2007 ; Goemaere *et al*, 2018)

2ème cause la plus fréquente d'infection à Candida
(Horn *et al*, 2009)



Culture sur milieu Sabouraud
Crédit : Adela Angoulvant



Levures adhérentes à un entérocyte Caco2
Crédit : Adela Angoulvant

Deux types d'infections

Candidose

Au niveau de la peau, de la cavité buccale et du tractus uro-génital

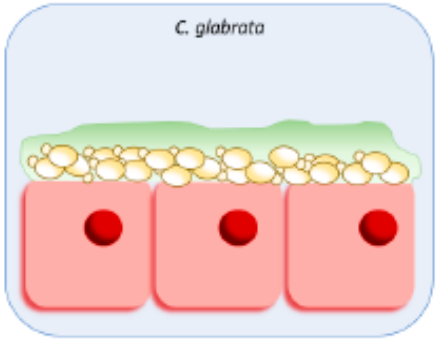
Taux de guérison très élevé

Candidose vaginale

75% des femmes au cours de leur vie, récidive de 50%

Candidose oropharyngée

Muguet chez les jeunes enfants, Infection la plus courante chez les patients atteints par le VIH (Fidel 2006)



(Galocha et al. 2019)

Candidémie

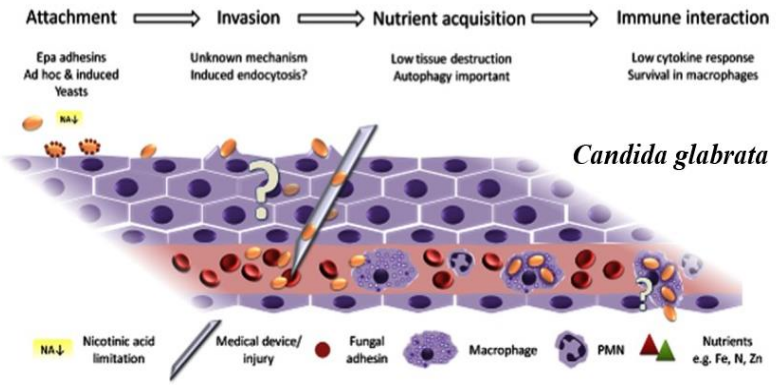
Elévation anormale de la température du corps, accélération du rythme cardiaque et respiratoire, rigidité musculaire, etc.

Infection sanguine très difficile à diagnostiquer

(pas d'état fébrile dans 50% des cas (O Leroy et al. 2008; Olivier Leroy et al. 2016))

Taux de mortalité proche de 50%

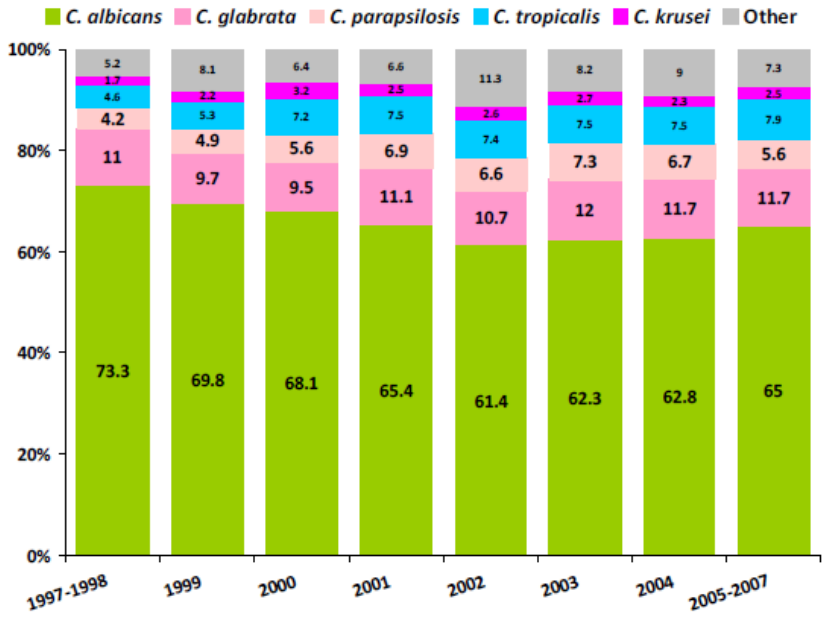
(Jaillette et al. 2016)



(Brunke et al. 2013)

Une annotation très inégale et une homéostasie peu décrite

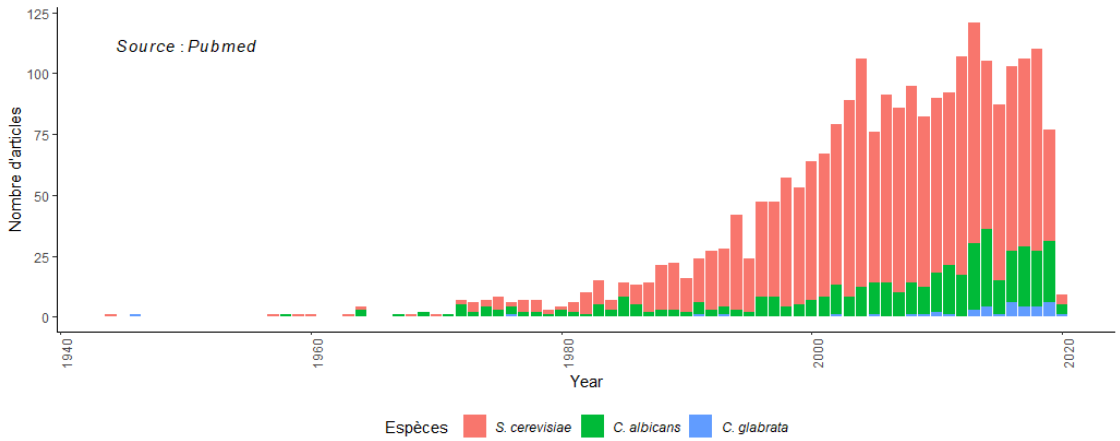
2ème levure pathogène
50% de mortalité (candidémie)
Pas de régression



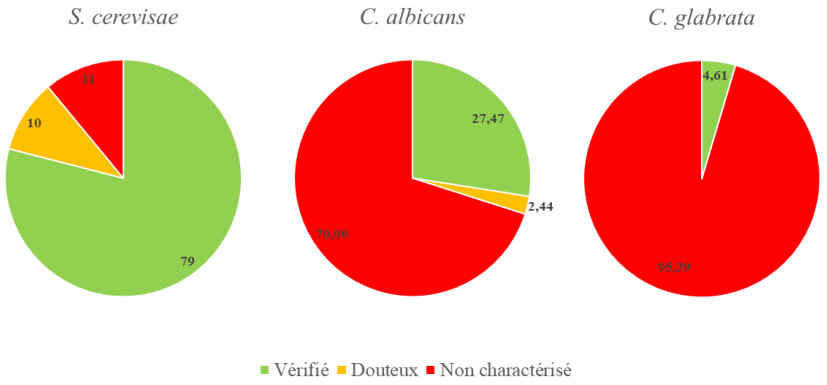
Données ARTEMIS DISK - Guinea et al. - 2014

Et pourtant

Peu de publications



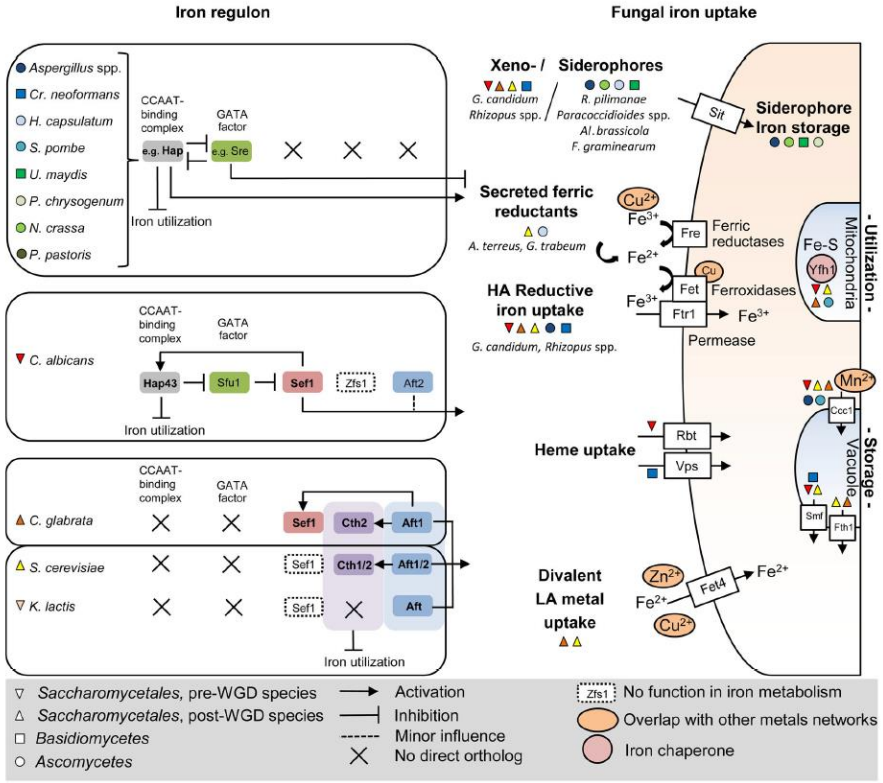
Annotation fonctionnelle pauvre



L'homéostasie du fer encore peu décrite

Beaucoup de transferts d'annotation

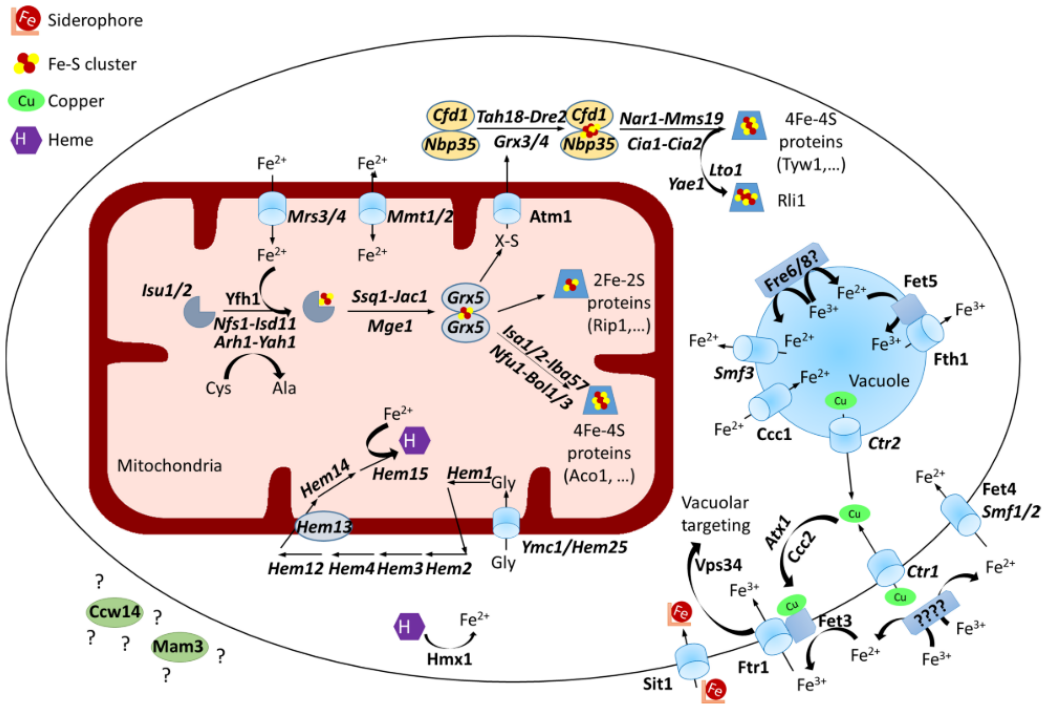
C. glabrata conserve des régulateurs « classiques » par rapport à *S. cerevisiae*…



(Gerwien et al. 2018)

Quelques gènes ont été décrits dans la littérature

… et a remodelé ses propres réseaux fonctionnels pour maintenir l'homéostasie du fer.



(Devaux et al. 2019)

Constitution d'un jeu de données original

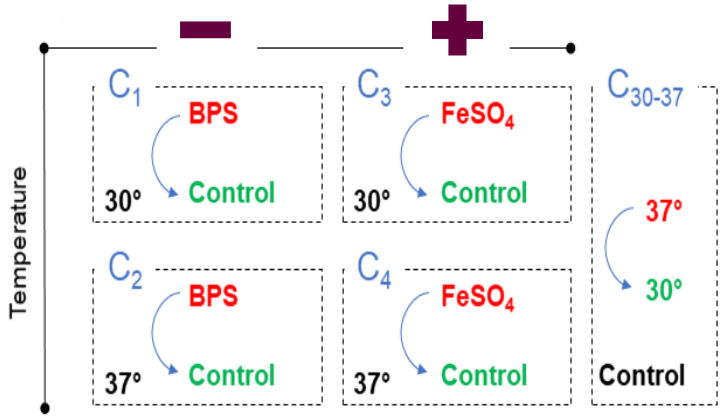
Données qualitatives

Candida Genome Database

Genome Resources for GRYC Yeast Chromosomes

GENEONTOLOGY Unifying Biology

Plan expérimental combinant des milieux pauvres et riches en fer



Souche ATCC 2001 (CBS 138)

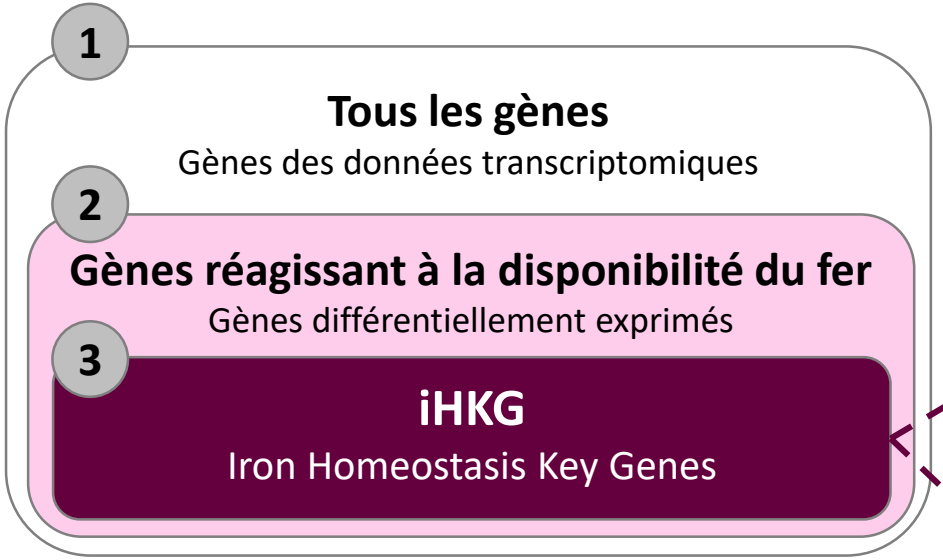
(Denecker *et al*, 2020)

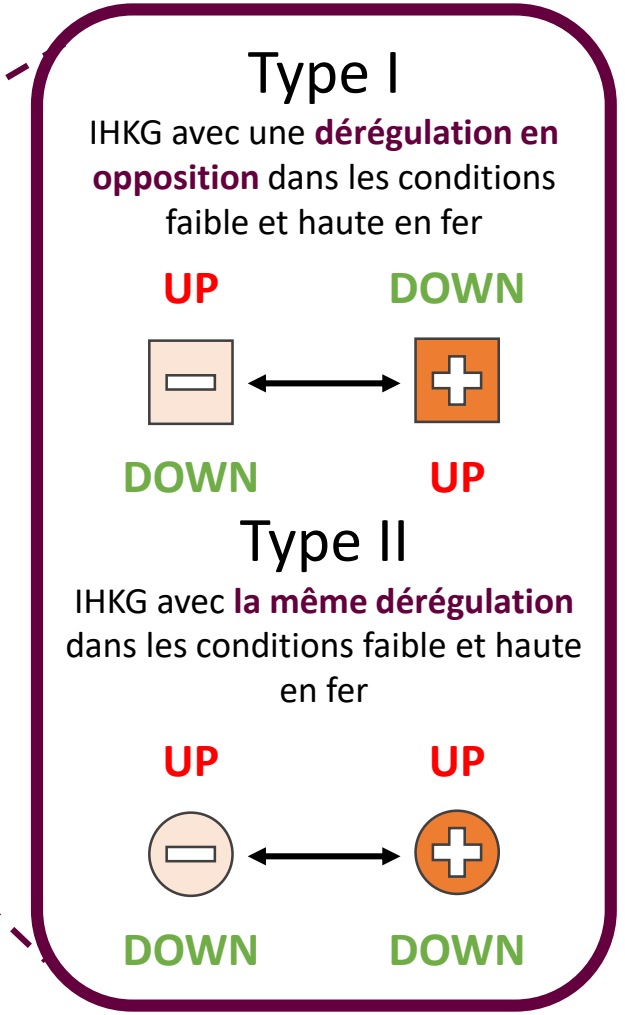
PIXEL

A content management platform for quantitative omics data

(Denecker *et al*, 2019)

Définitions et hypothèse de travail





Hypothèse

Gènes ayant des fonctions cellulaires importantes pour contrebalancer les fluctuations externes de la disponibilité du fer (en carence **et** en surcharge)

iHKG classés en deux sous-groupes : Type I et Type II

214 gènes sélectionnés

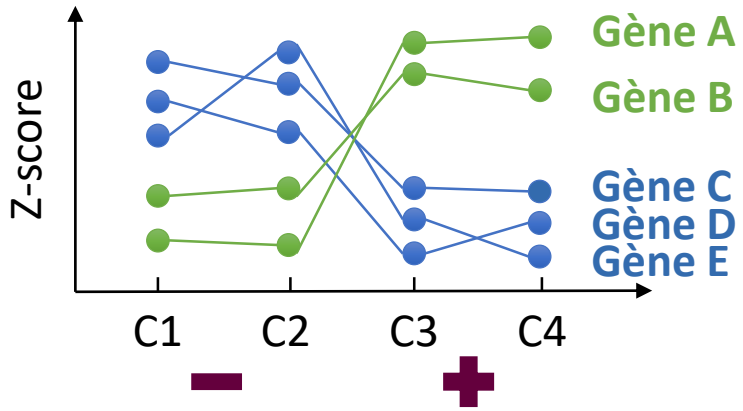
Pertinence biologique des iHKGs

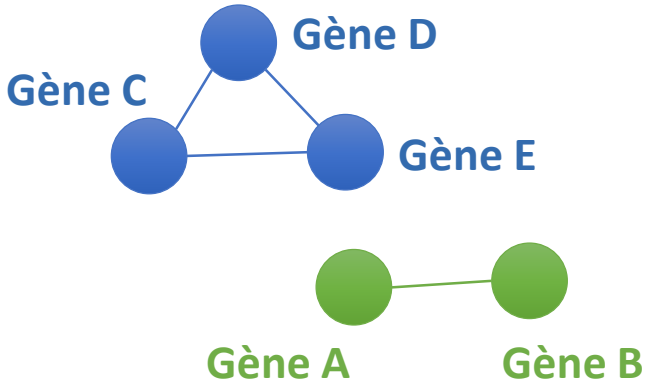
Des gènes connus chez *S. cerevisiae (environ 50-100 gènes)*

- **Fonctions cellulaires dépendantes du fer (respiration,…)** : QCR2, QCR6, QCR7, QCR10, COX4, COX5B, COX6, COX7, COX9, COX12, COX15, ACO1, COX23
- **Des gènes codant des métalloprotéines** : SDH2, CCP1, RIP1, CYT1, LIA1, CYC1, GLT1, YHB1, RLI1, ILV3
- **Des gènes impliqués dans l'autophagie** : ATG19, ATG32, ATG41
- **Dans les clusters Fe-S** : ISA1, CGD1, GRX4, HEM4, HEM15
- **Dans le transport du fer** : FTR1, FET3
- …

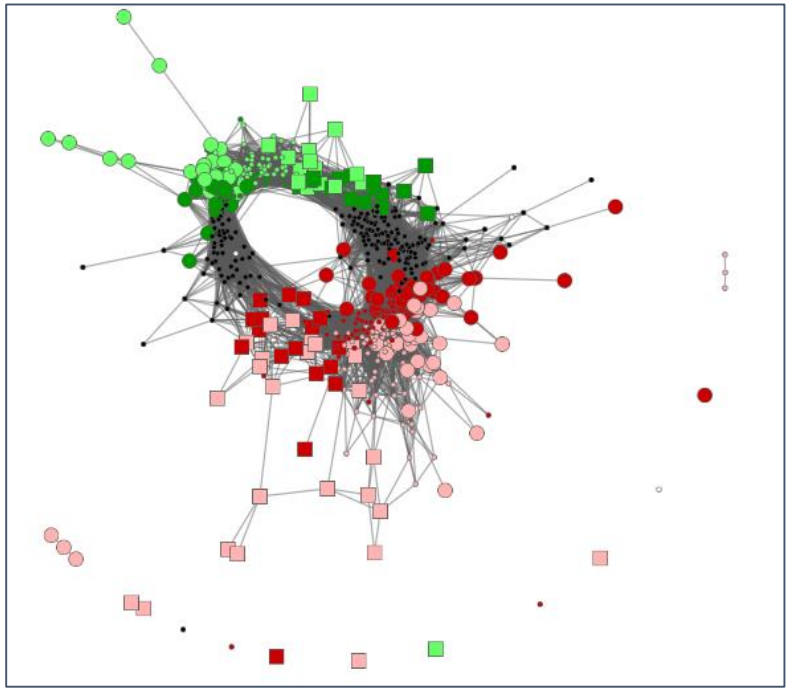
Mais qu'en est-il des autres gènes ?

Réseaux de co-expression des gènes réagissant au fer

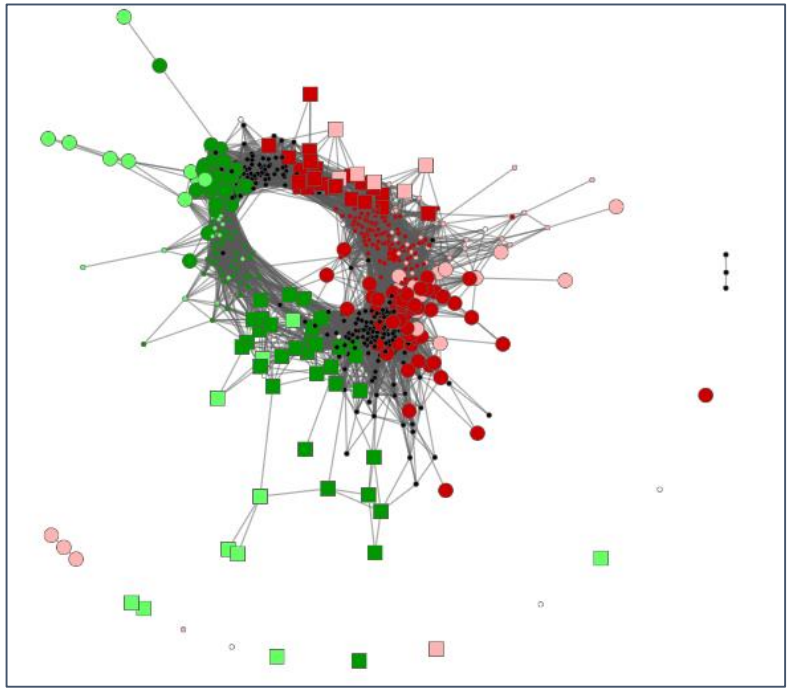


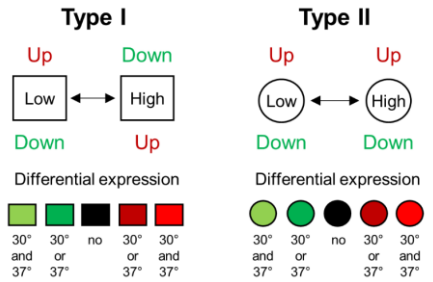


*637 gènes en **carence** en fer*



*637 gènes en **surcharge** en fer*





Pour aller plus loin

Séparation en sous-réseaux fonctionnels de gènes co-exprimés

Comment créer des sous-réseaux fonctionnels de gènes ?

Contraintes fortes

Un gène

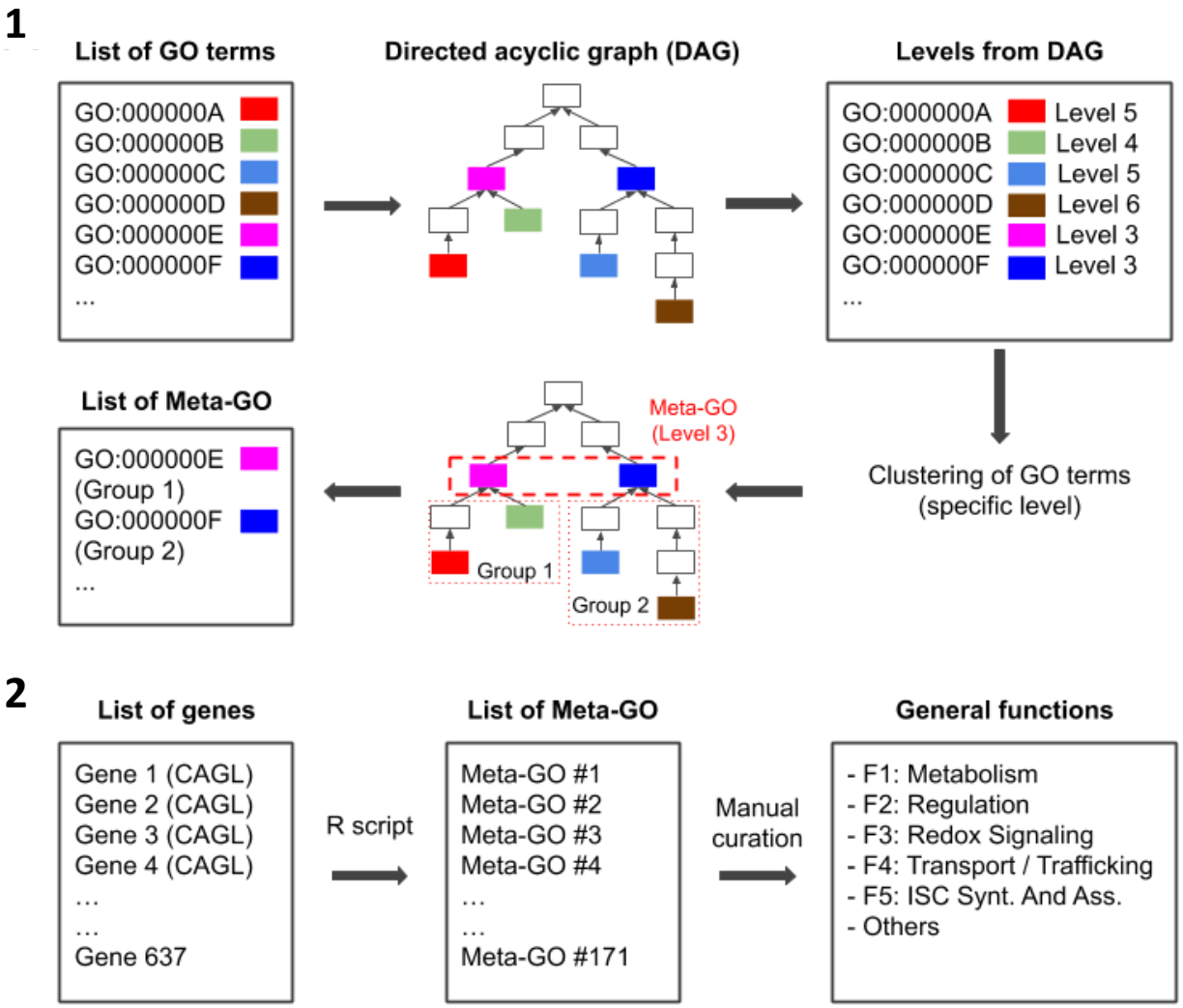
=

Une fonction

(un seul sous-réseau)

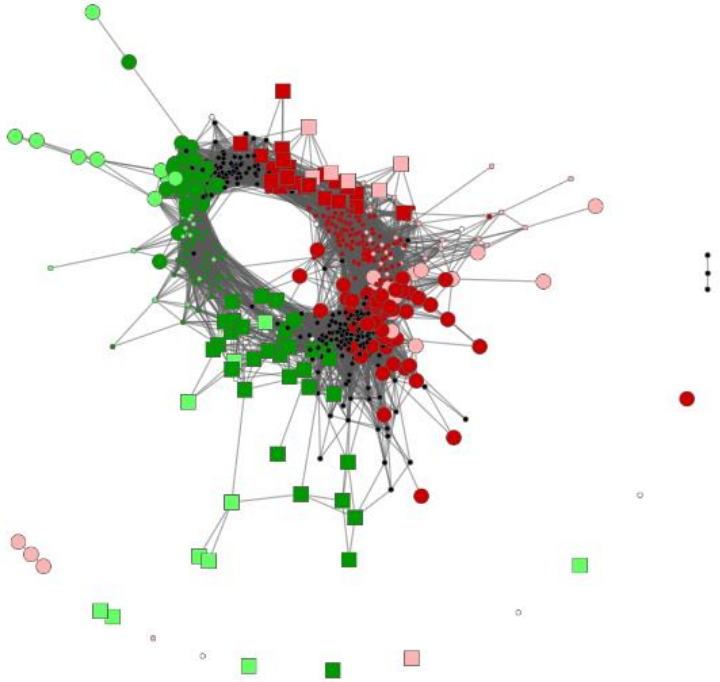
Nombre limité de sous-réseaux fonctionnels

Méthode semi automatique avec curation manuelle



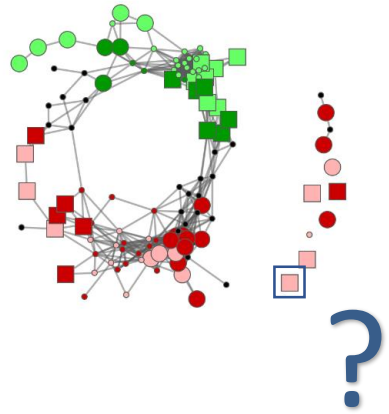
Exploration des sous-réseaux fonctionnels de gènes

637 gènes



Redox signaling

118 gènes



?

Comment exploiter au maximum ces réseaux,
résultat d'une intégration de données hétérogènes ?

iHKGviewer

iHKGviewer

Exploration simplifiée par une interface web

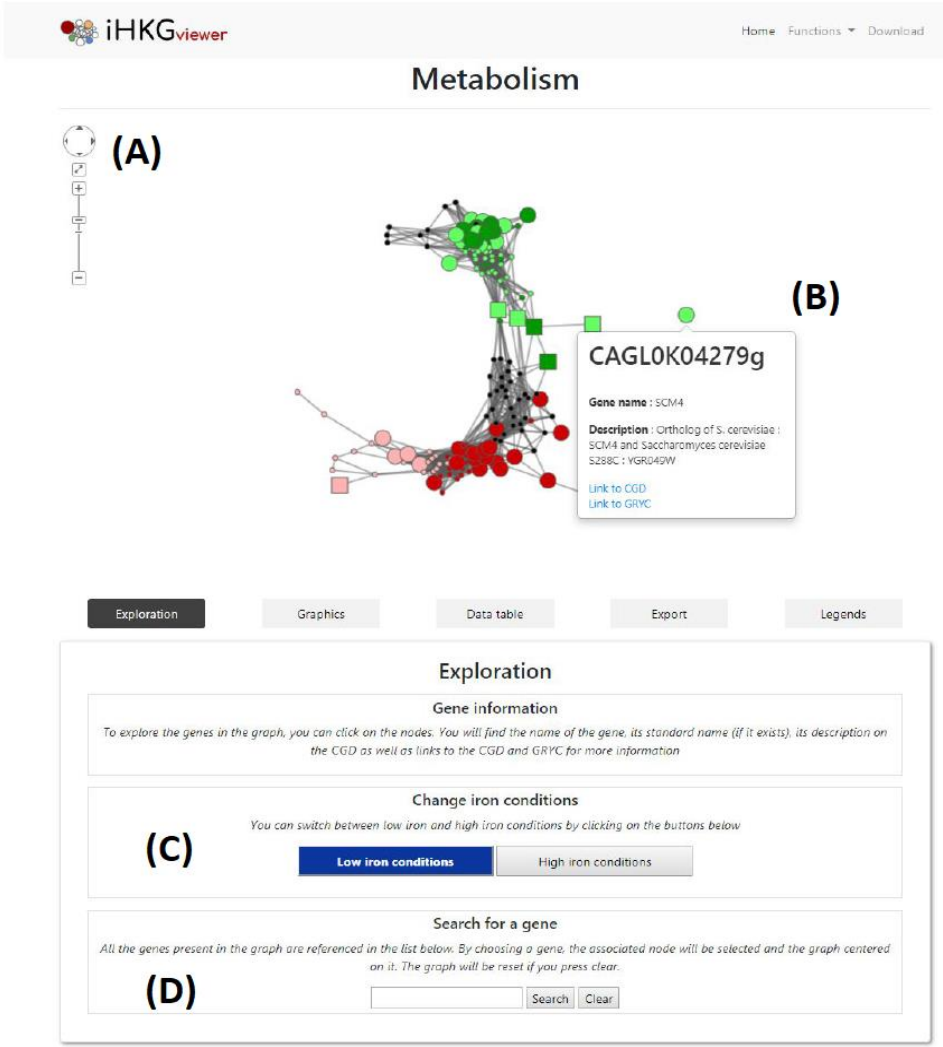
https://thomasdenecker.github.io/iHKG/

(A) Possibilité de zoomer sur le graphique

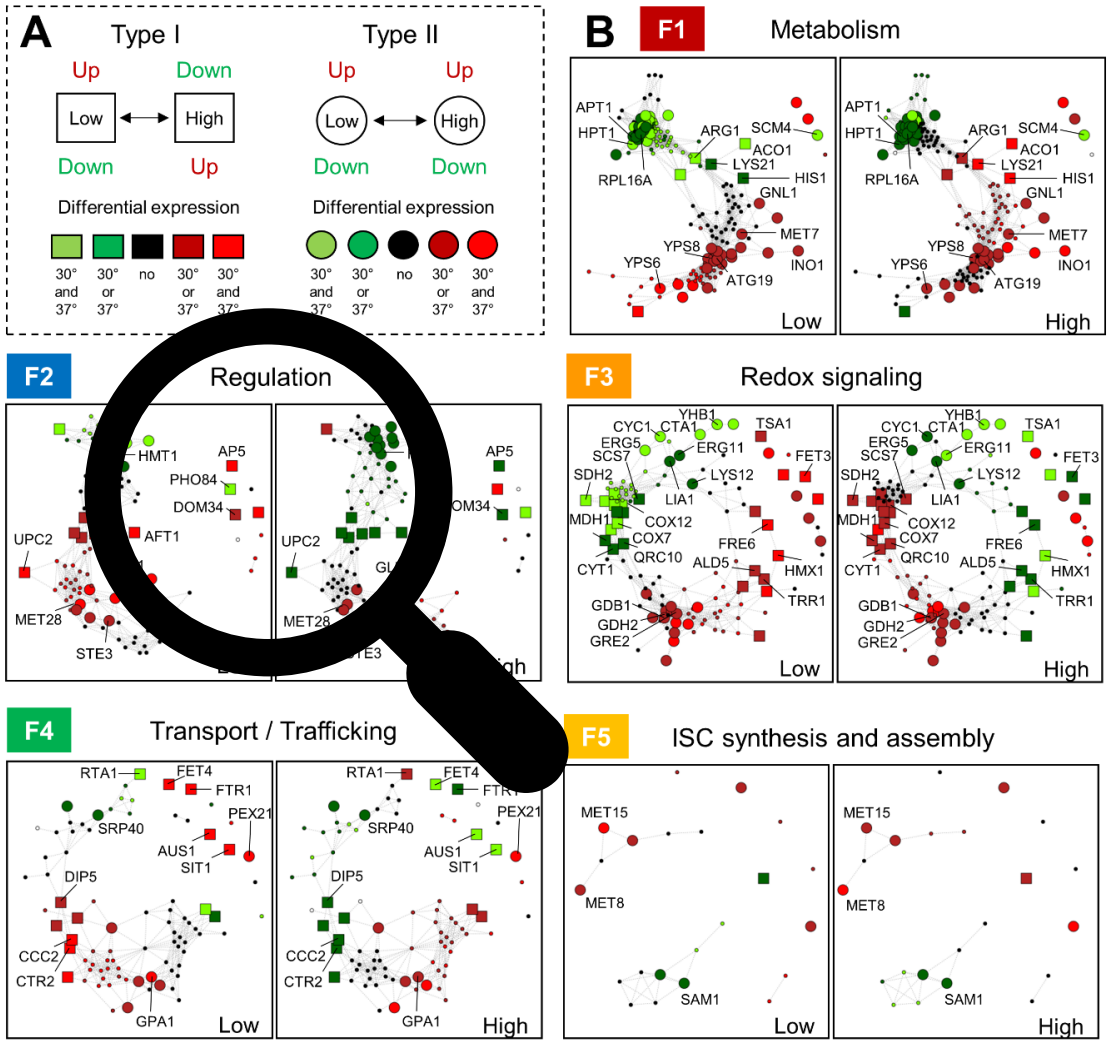
(B) Possibilité de cliquer sur un nœud avec la souris pour obtenir le nom du gène, sa description et des liens web directs vers les bases de données CGD et GRYC

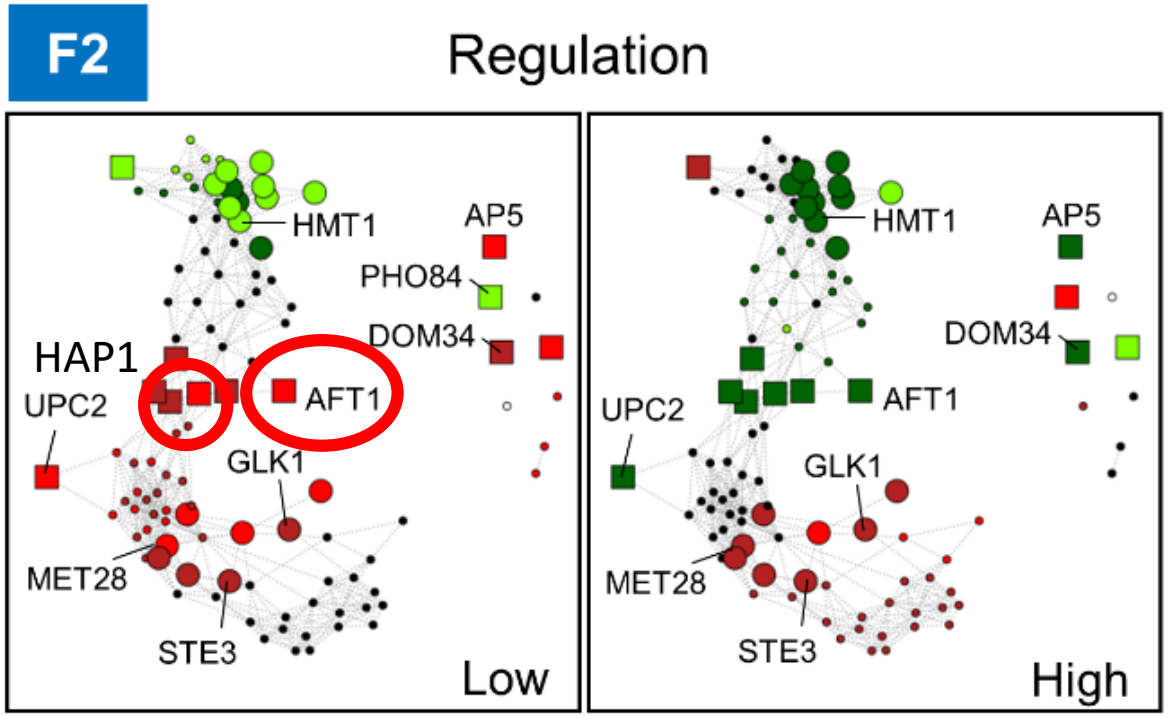
(C) Possibilité de passer d'une condition de fer faible à une condition de fer élevé

(D) Possibilité de rechercher un gène particulier dans le réseau

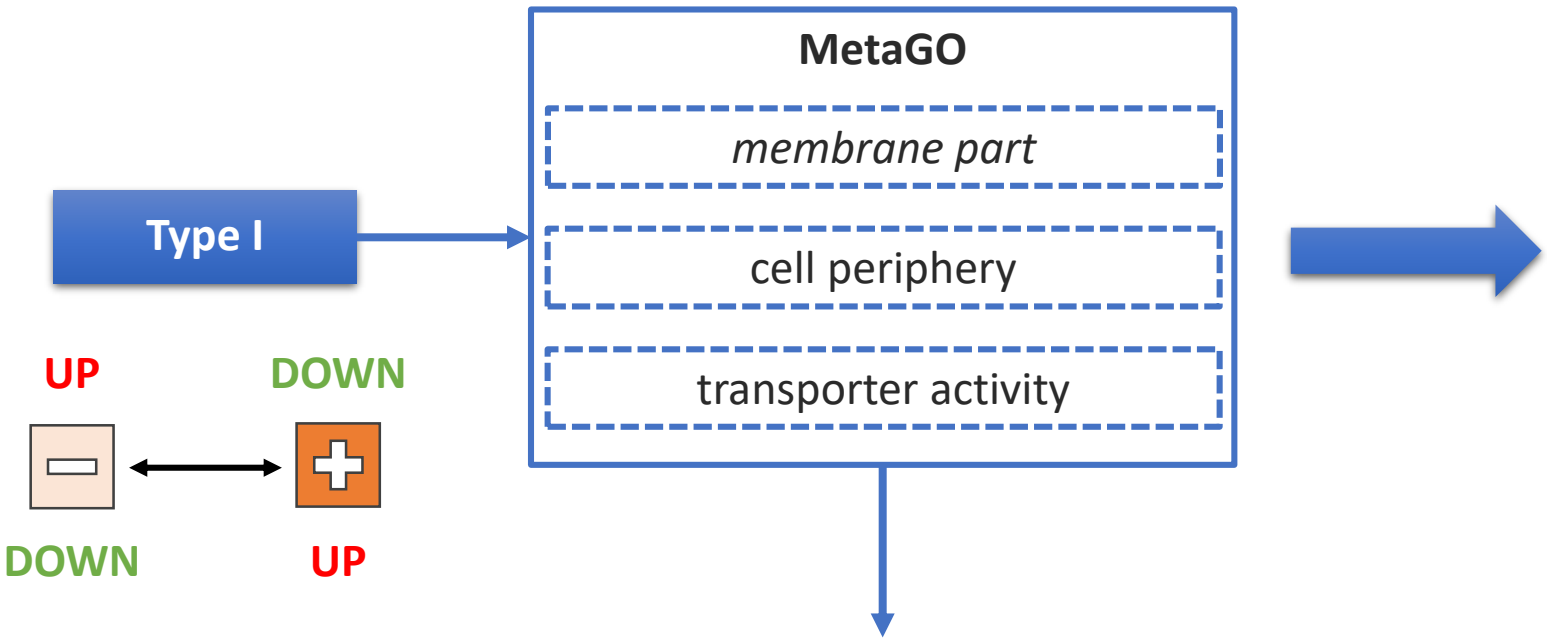


Réseaux fonctionnels de gènes co-exprimés

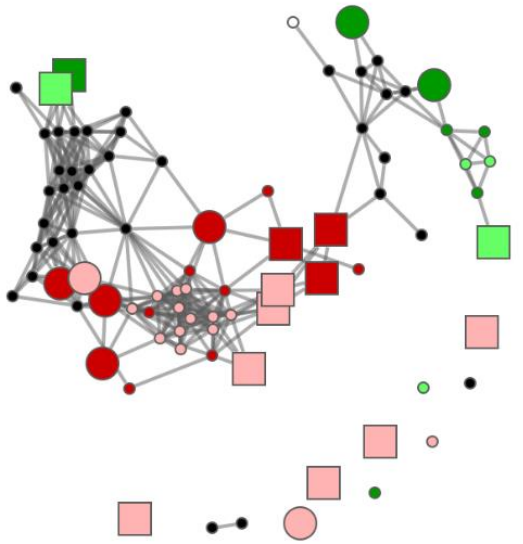




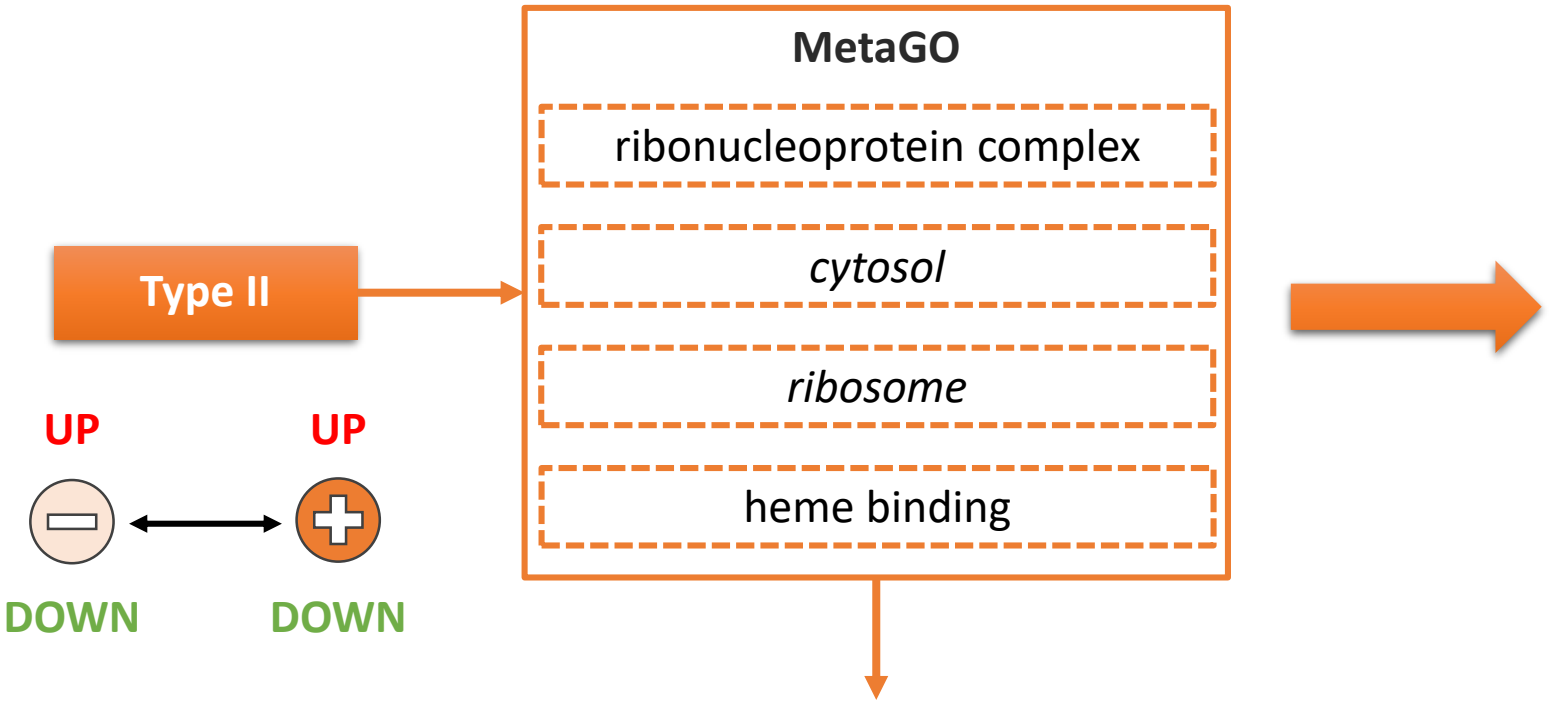
Enrichissement fonctionnel



Transport / trafficking

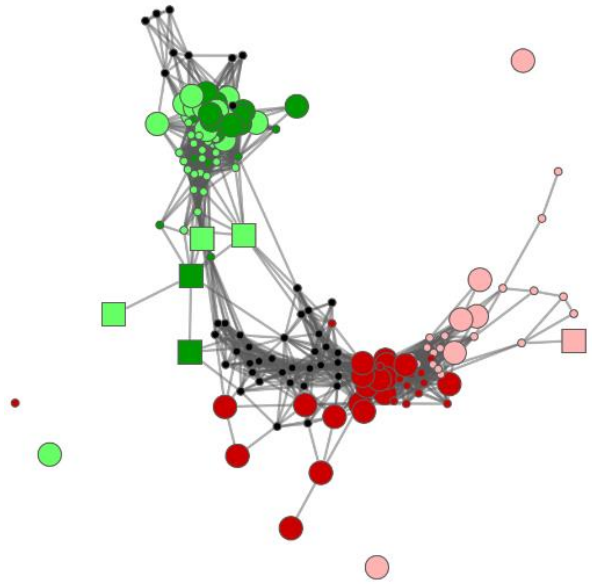


Enrichissement fonctionnel

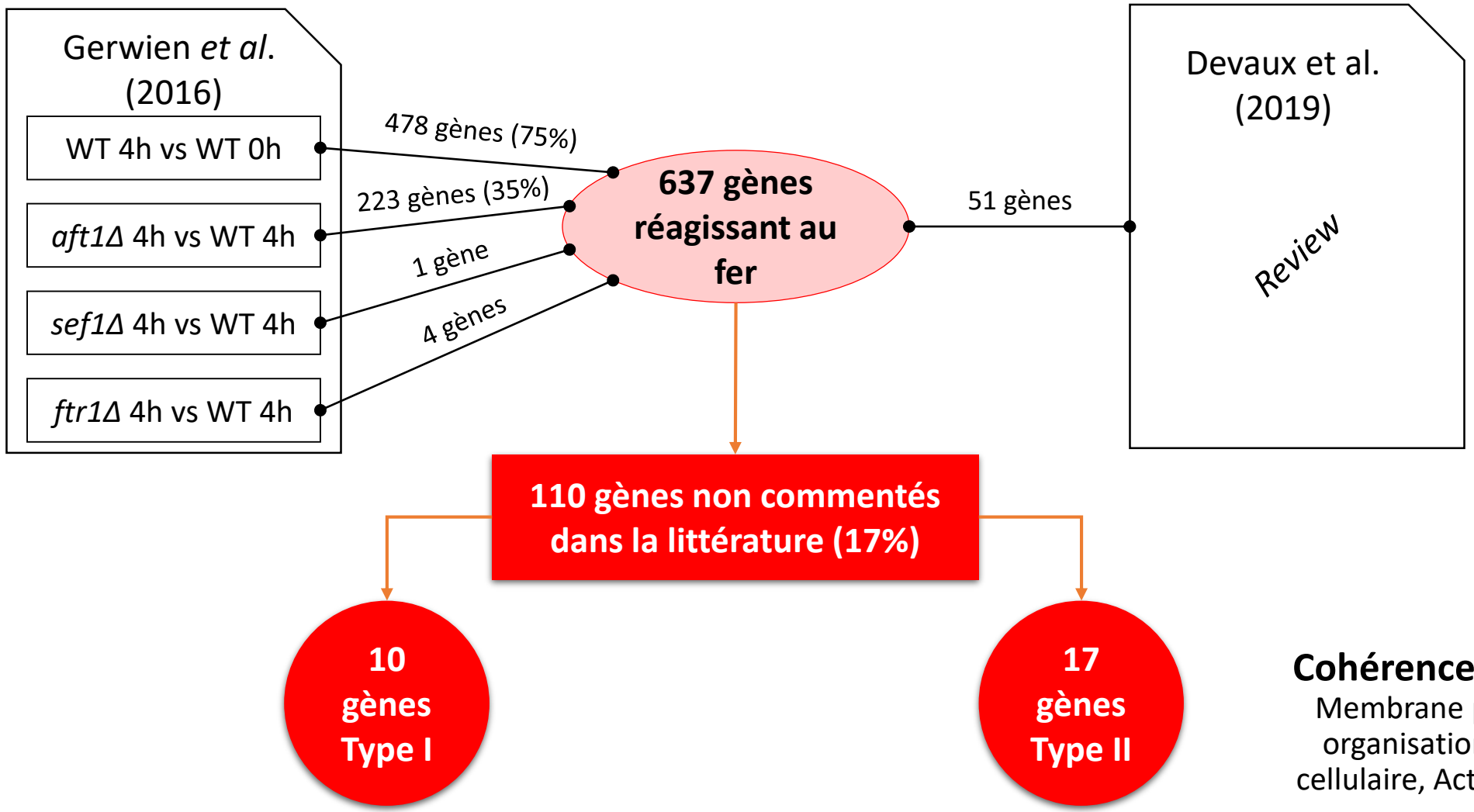


Réponse au stress et stabilisation des processus clés impliquant du fer

Metabolism

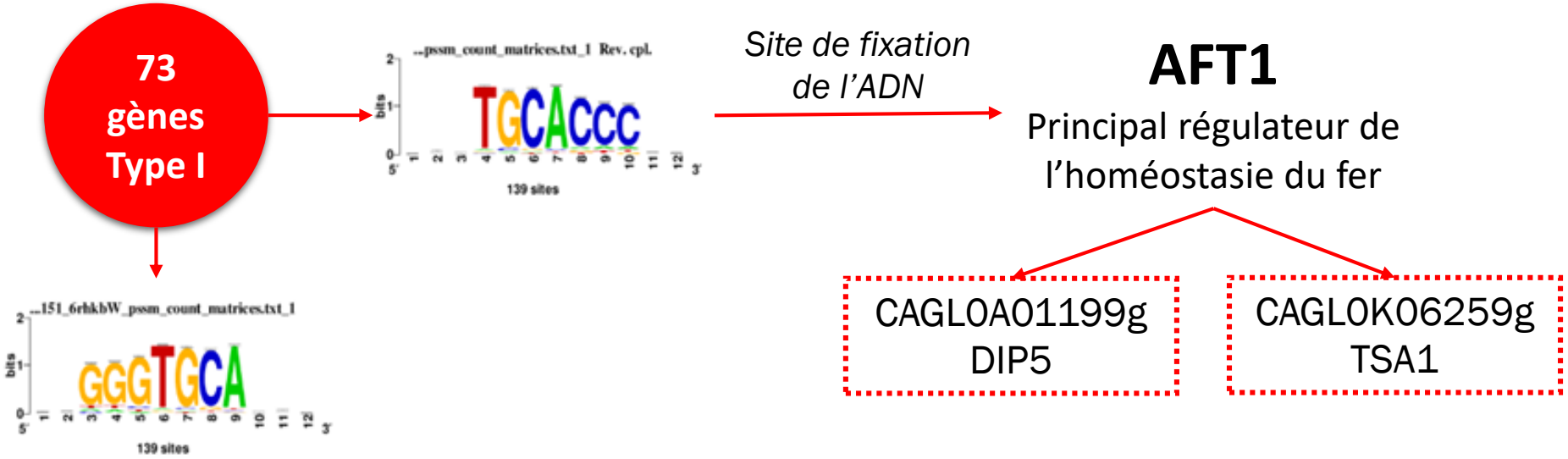


Nouvelles annotations fonctionnelles de gènes



Cohérence biologique
Membrane plasmique et organisation de la paroi cellulaire, Activité Redox, …

Nouvelles annotations fonctionnelles de gènes



"Régulateur du fer" – Premières descriptions fonctionnelles pour ces gènes

sur la base d'expériences menées directement chez *C. glabrata* sans transfert d'informations des levures modèles *S. cerevisiae* et *C. albicans*

Des pistes à explorer expérimentalement

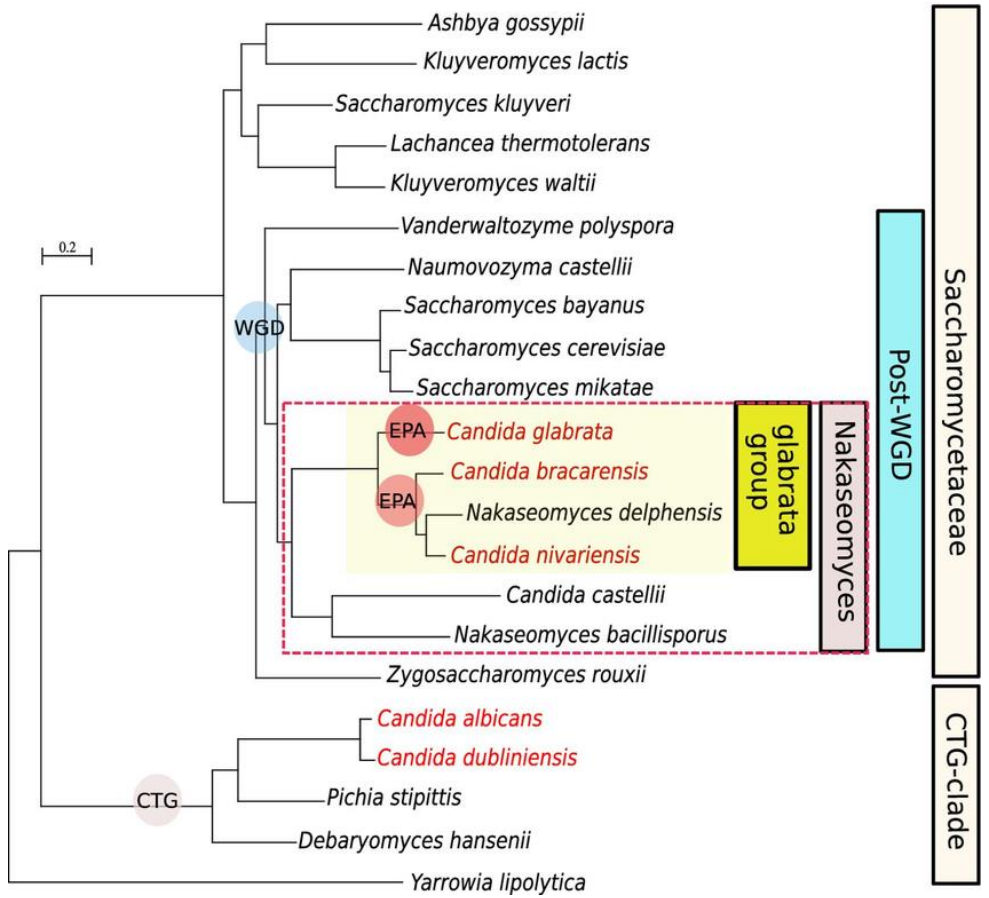
Conclusion et perspectives

Méthodologie originale d'intégration et d'exploration des données

637 gènes réagissent aux changements de concentration en fer

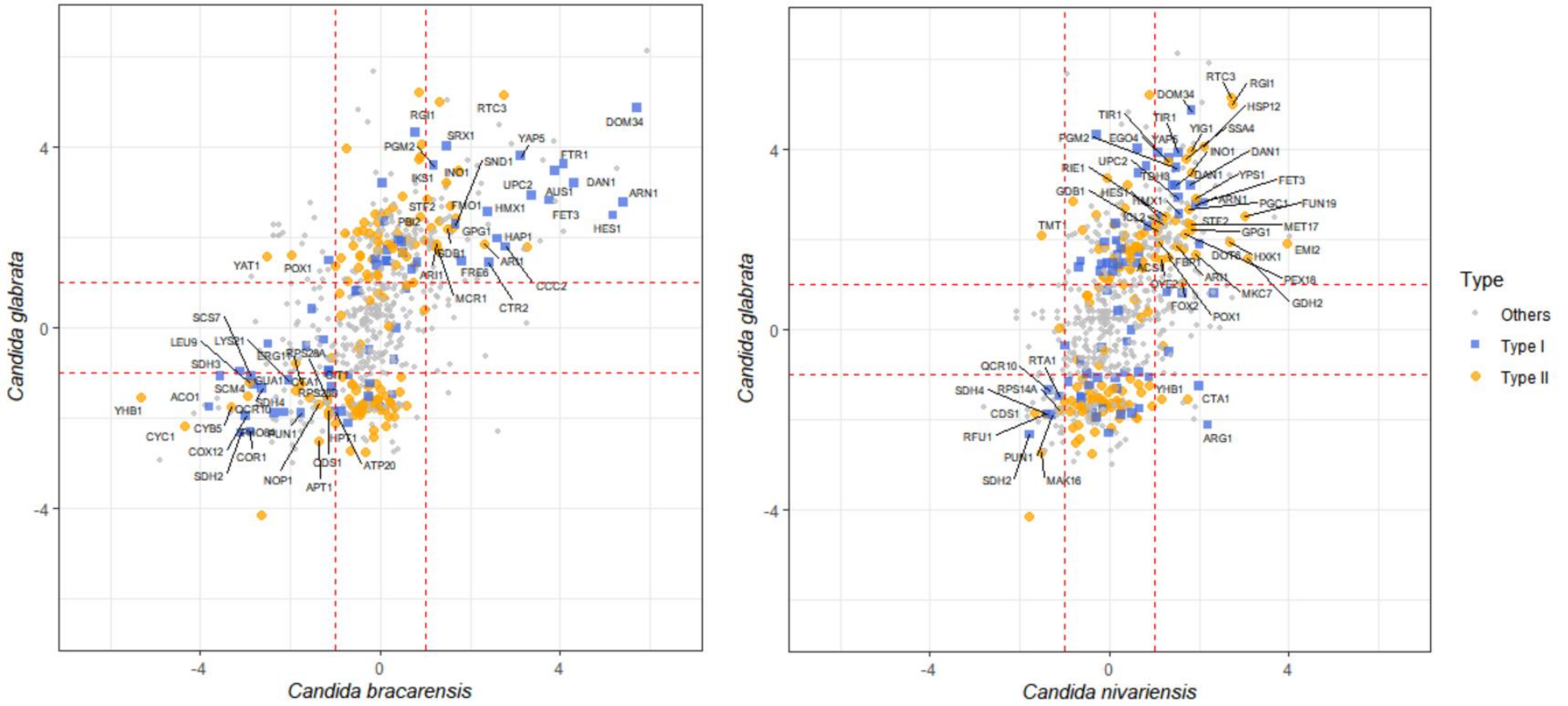
214 gènes étant de très bons candidats dans l'homéostasie du fer :

- Peuvent être une aide dans l'amélioration de l'annotation de la CGD (seulement 5% des ORFs sont vérifiées)
- Peuvent constituer un point de départ pour une étude comparative avec des espèces proches phylogénétiquement (clade des *Nakaseomyces*) dont la pathogénie
- Peuvent permettre de mieux comprendre l'évolution des réseaux de régulation de l'homéostasie du fer chez les levures



(Gabaldón *et al*, 2016)

Résultats préliminaires

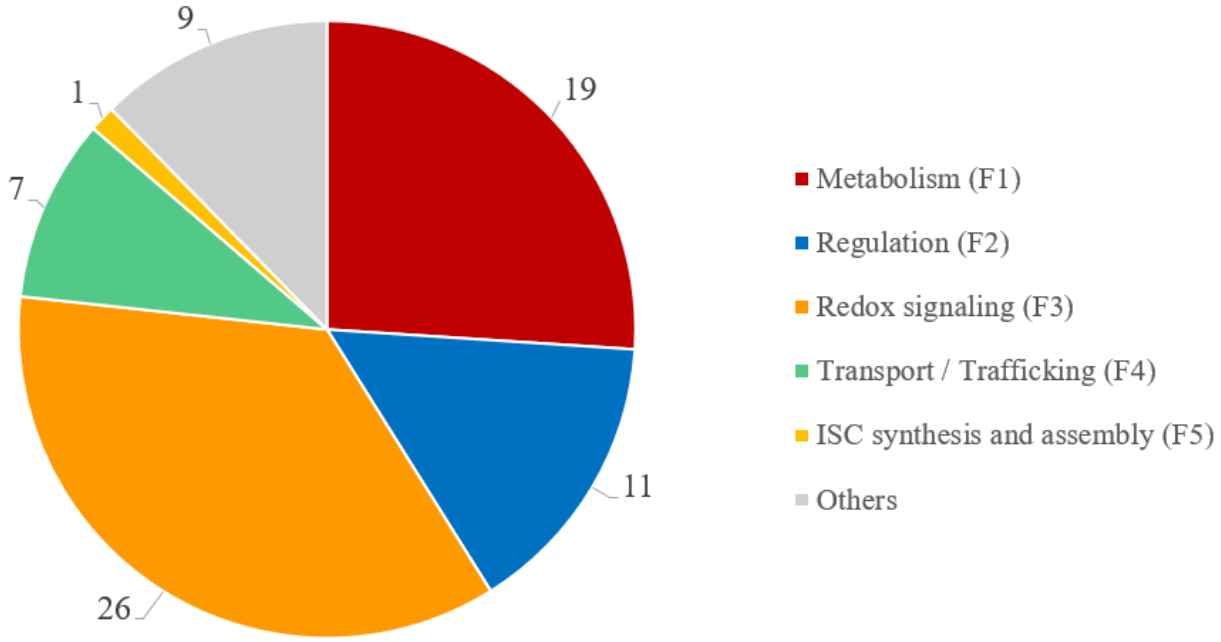


Mise en évidence de gènes très bien décrits chez *C. glabrata* dont les orthologues au sein du clade sont différentiellement exprimés de façon similaire en condition de carence en fer

Résultats préliminaires

73 gènes partagés

Entre *C. glabrata* et *C. bracarensis* et *C. nivariensis* dont les fonctions générales sont dominées par des fonctions clés dans l'homéostasie du fer



ÉTUDE DE L'IMPACT DE LA PRISE EN COMPTE SYSTÉMATIQUE DES MODIFICATIONS POST-TRADUCTIONNELLES LORS DE L'IDENTIFICATION DE PROTÉINES CHEZ LA LEVURE PATHOGÈNE *CANDIDA ALBICANS*

Constat sur la plateforme de protéomique de l'IJM

50 %

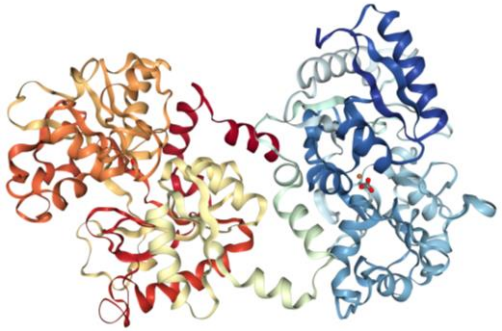
des spectres de masse ne conduisent pas à l'identification d'une protéine par spectrométrie de masse MS/MS en approche Bottom Up sur la plateforme

Perte considérable !

Hypothèse : Les modifications post-traductionnelles

Pourquoi ?

Spectrométrie de masse LC-MS/MS - Approche Bottom up



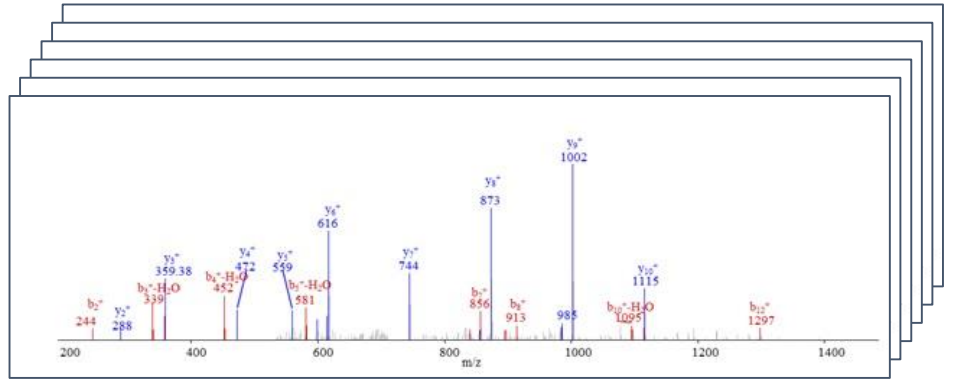
Protéine inconnue

Digestion trypsique

Peptides

Chromatographie liquide

Spectrométrie de masse

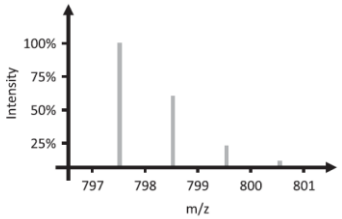


Bioinformatique : Identification des peptides

Protéine identifiée

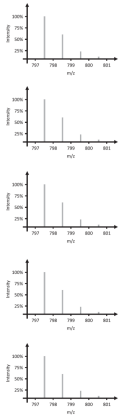
Identification des peptides à partir des spectres de masse

Spectre expérimental



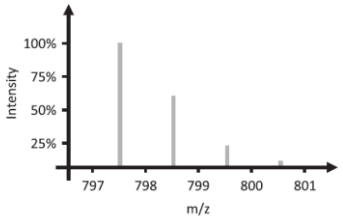
Séquences théoriques

Création d'une banque spectrale théorique



Approche 1
Comparaison des spectres

Spectre expérimental



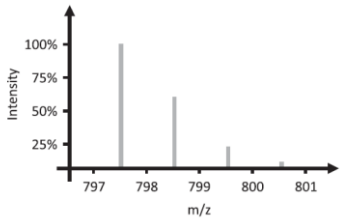
Déduction de la séquence expérimentale

FIAVGYK

Séquences théoriques

Approche 2
Comparaison des séquences

Spectre expérimental

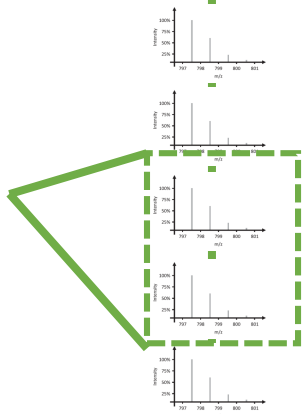


Déduction de la séquence expérimentale

FIAVGYK

Séquences théoriques

Création d'une banque spectrale théorique



Filtrage de la banque

Approche 3
Comparaison hybride

Création d'une banque spectrale théorique

> Protéine *TOTO*
MYELNNEEVLRKRKERFSKFGKEAIINDPLRDVALLSRSGESNTIIDLKINHDKRSEMVS
MLKLLFYDEKQLTTVEHGLRKLREVFMSIRQDHRDEDESFWKQASEVYKLSYDFLLRHGQ
YNKLGGLVLNAIHEWFPLQYRKPYAKIYALYLSHIEKDVPKCVDFLQYSSVSQSESLDII
NMANIYVLKSESPRIWFHYCKNLKDDELNFLELSSVMQVMINRTDNLLQLCYNQLSVKVA
QQLWFGDHFTSNLETRIKDKYDMRAGTDIILFKKRQIKG

Simulation d'une digestion trypsique

MYELNNEEVLR
~~K~~
~~R~~
~~K~~
~~ER~~
~~FSK~~
~~FGK~~
EAIINDPL
...

< 3000 DA and ≤ 30 AA

Création des spectres théoriques

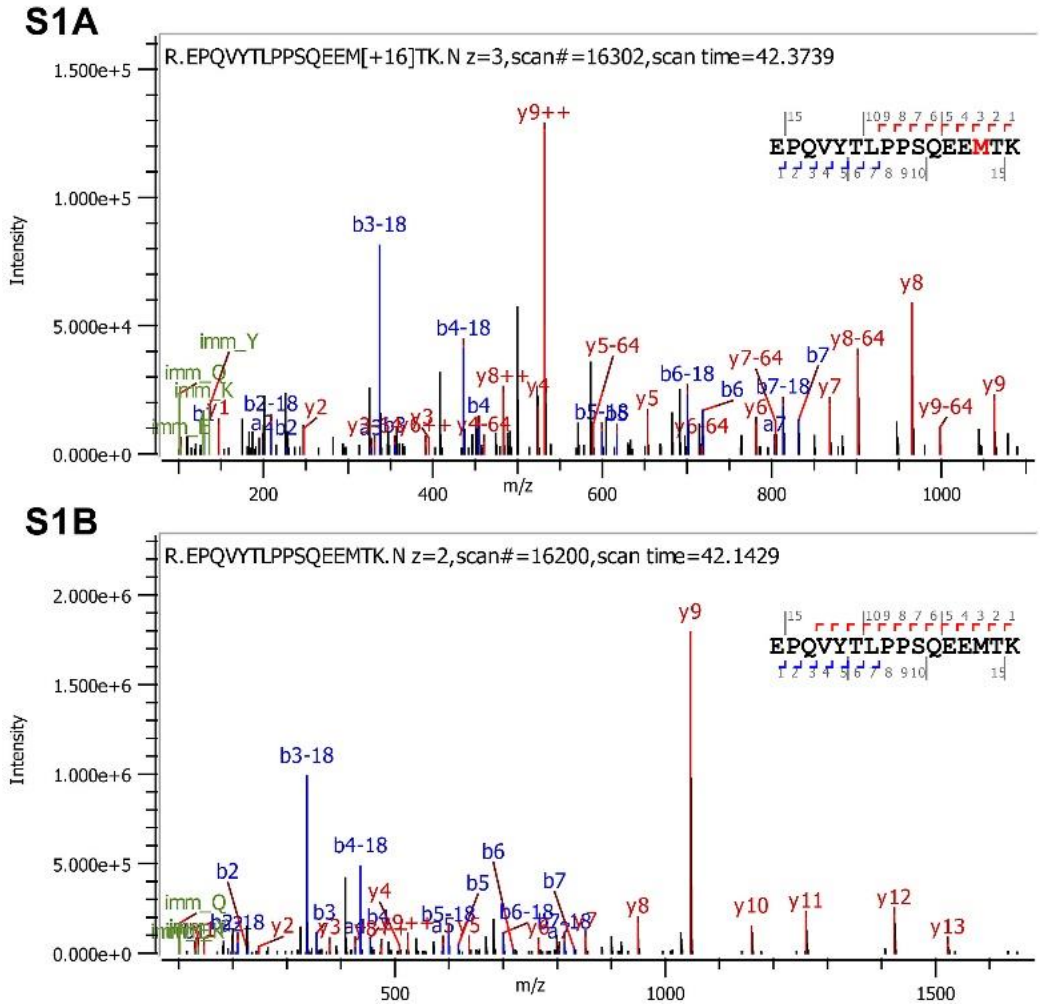


...

Spectres théoriques pour la protéine TOTO

Si aucune modification post-traductionnelle n'est indiquée lors de la construction de la banque spectrale théorique, alors aucun spectre ne contiendra de modification post-traductionnelle

Avec / sans modifications post-traductionnelles



(Xu et al. 2019)

Si la banque ne contient pas de spectres avec des modifications post-traductionnelles

Absence d'identification

Questionnement scientifique

Est-il possible d'améliorer le taux d'identification des protéines en prenant en compte de façon systématique les modifications post-traductionnelles ?

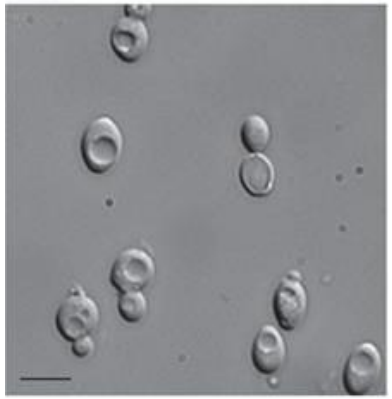
Aujourd'hui, cette recherche est trop longue par les approches classiques d'identification

(Mascot : 1-3 heures pour seulement 3 modifications post-traductionnelles / 1500 possibles)

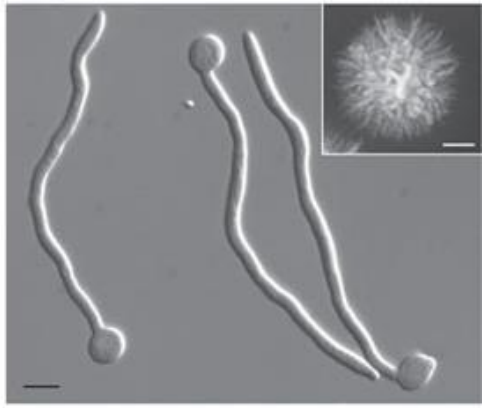
Collecte des données : *Candida albicans*

30 fichiers RAW

15 fichiers dans la forme levure



15 fichiers dans la forme hyphe



Sudbery *et al*, 2011 – DOI : 10.1038/nrmicro2636

8 chromosomes -12 405 ORFs (diploïde)

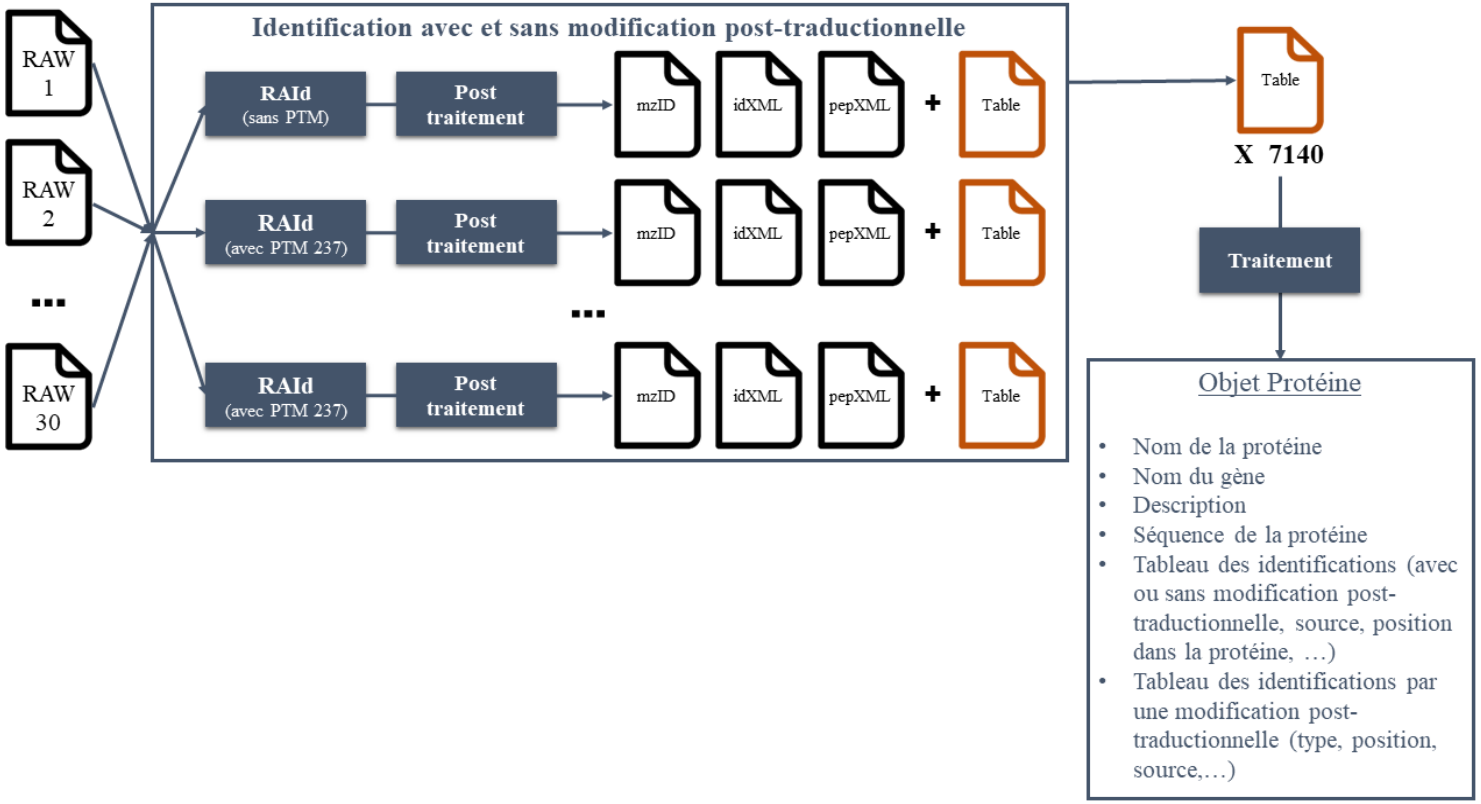
Organisme commensal
des muqueuses humaines

Cause majeure de mortalité dans
les structures de santé

1ère cause d'infection à *Candida*

Défi informatique

Mise en place d'une nouvelle approche systématique utilisant le logiciel RAId pour prendre en compte un maximum de modifications post-traductionnelles



Rapide

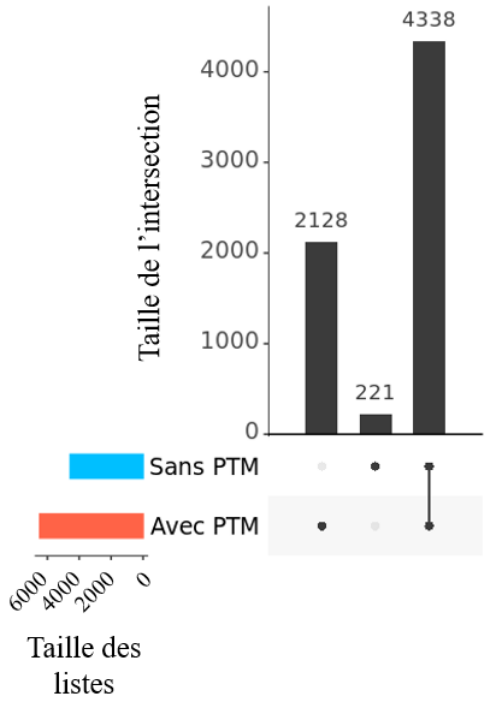
En seulement 14h

Reproductible

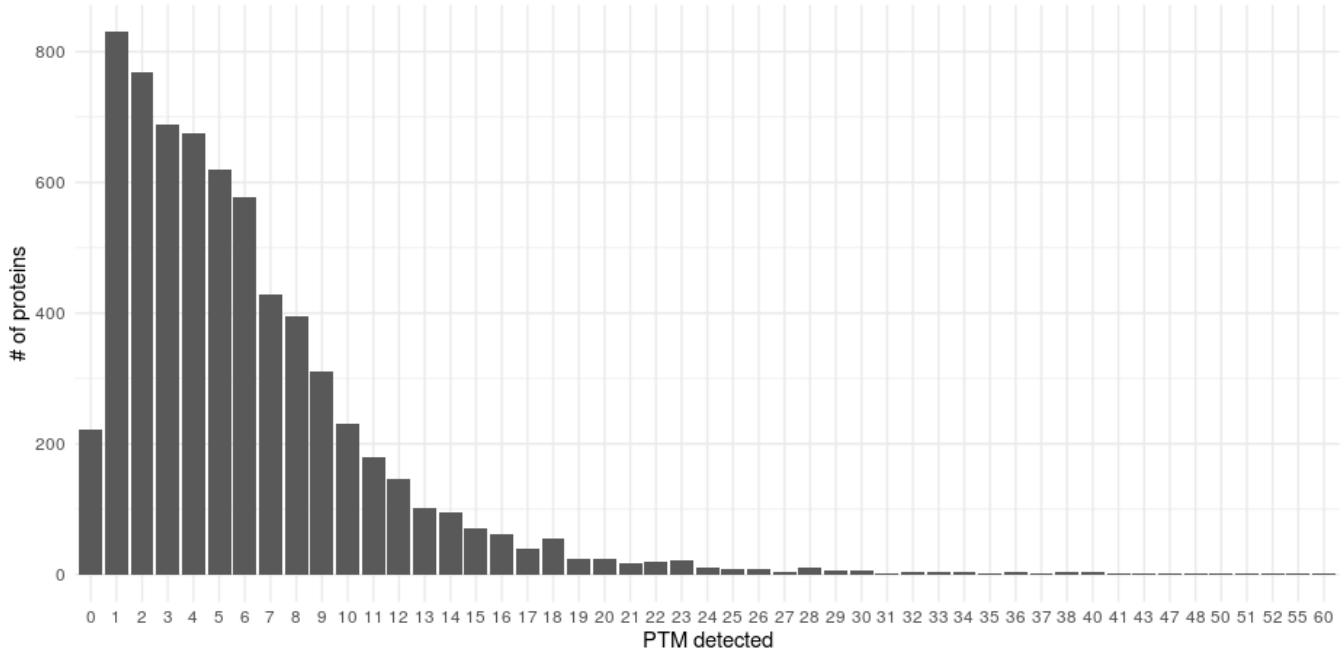
GitHub docker

Résultats de la prise en compte des PTMs

Nombre de protéines identifiées avec ou sans modifications post-traductionnelles

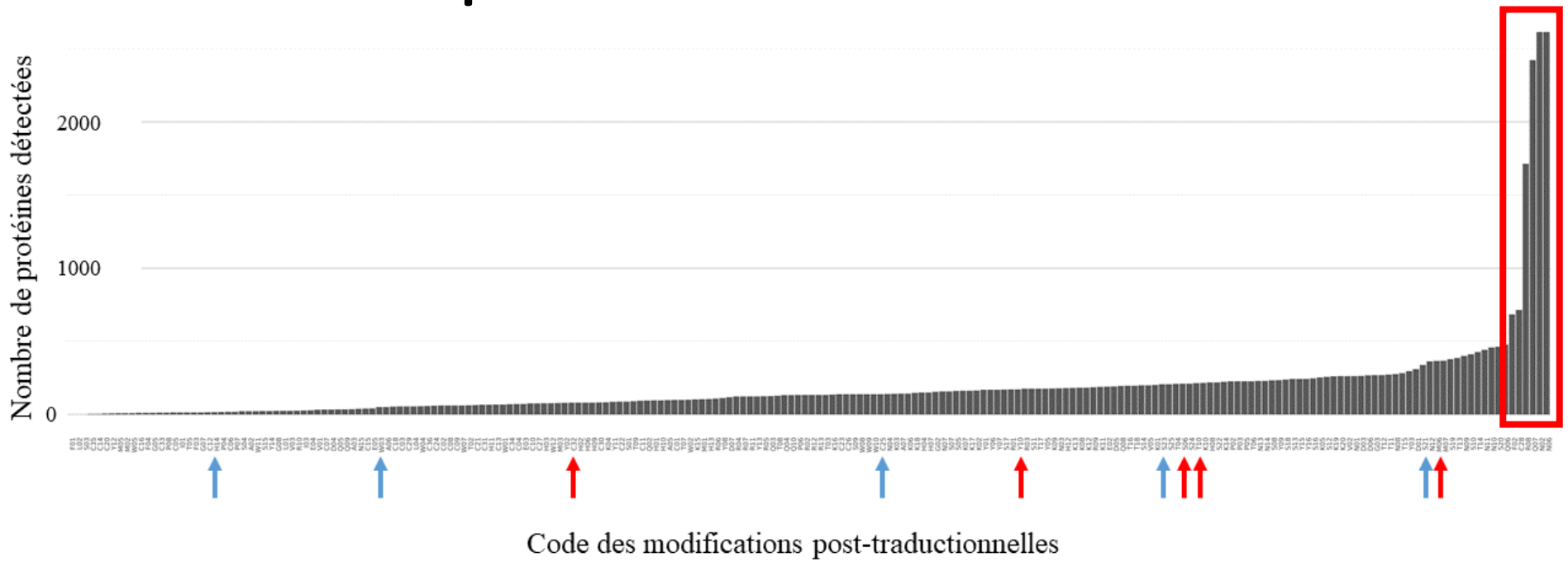


Nombre de protéines identifiées en fonction du nombre de modifications post-traductionnelles détectées



Importance de la prise en compte des modifications post-traductionnelles

Une première liste pour *C. albicans*



- Modifications post-traductionnelles recherchées actuellement en routine
- Modifications post-traductionnelles permettant d'identifier de nouvelles protéines uniquement grâce à elles

Une liste spécifique à explorer pour *C. albicans*

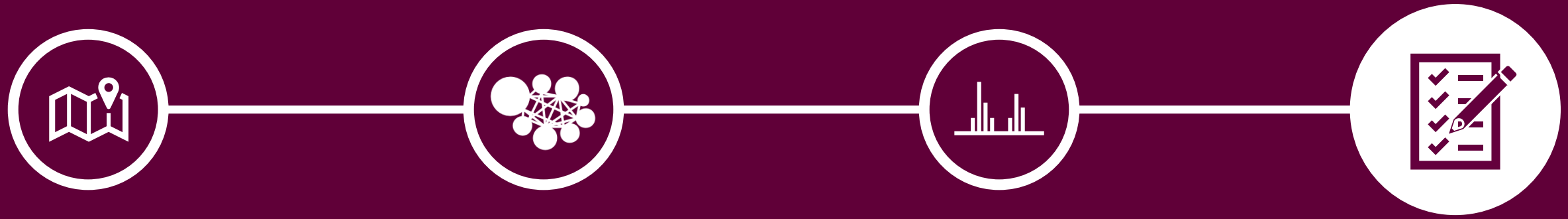
- Glutathionylation (Modification post-traductionnelle très étudiée au laboratoire)

Conclusion

1. **Proposition d'un nouveau protocole d'identification des protéines plus rapide et plus efficace**
2. **Confirmation de l'impact de l'étude des modifications post-traductionnelles dans le taux d'identification des protéines**

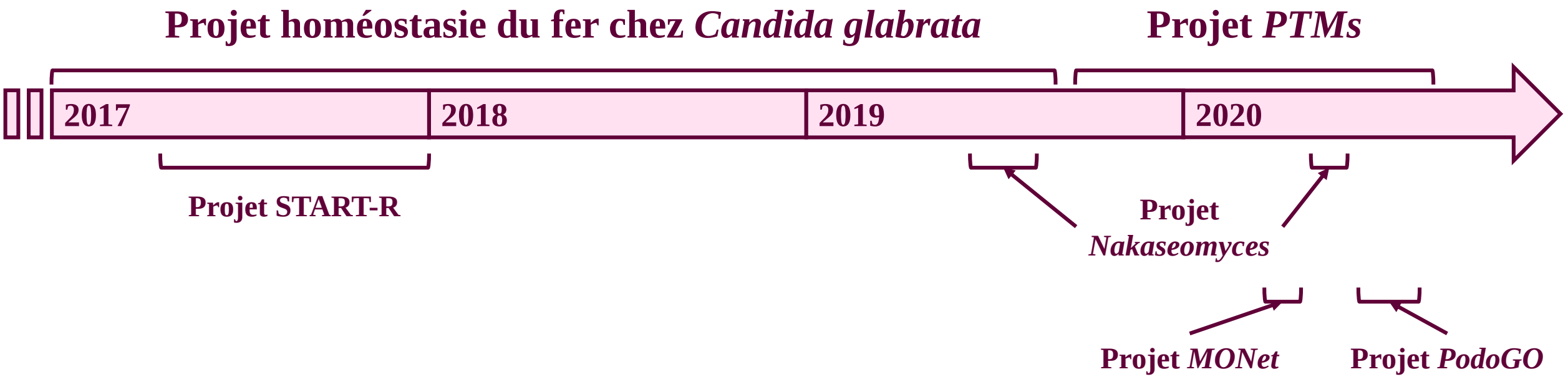
Perspectives

1. **Augmenter la liste des modifications post-traductionnelles recherchées systématiquement**
2. **Réaliser la même étude sur d'autres organismes (données disponibles chez *C. glabrata*)**

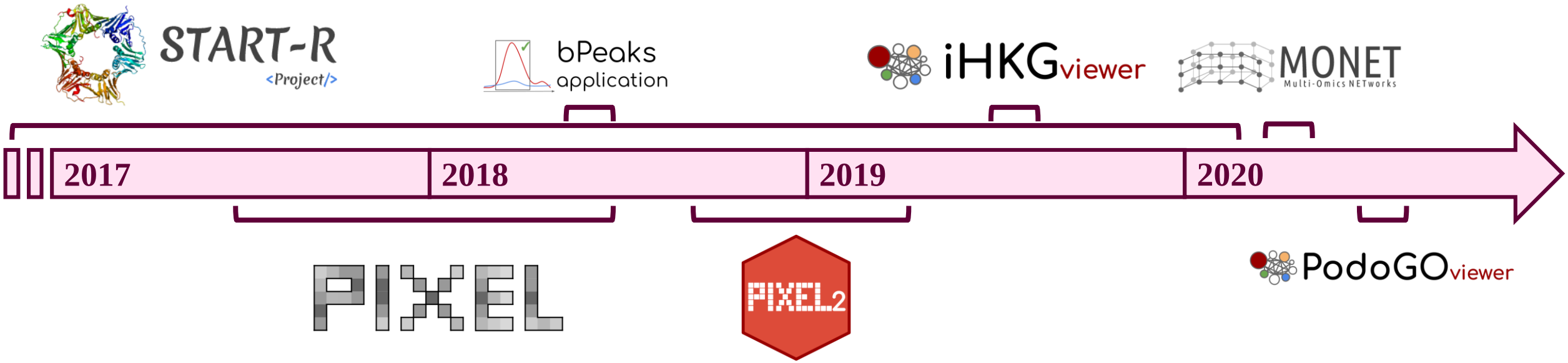


BILAN

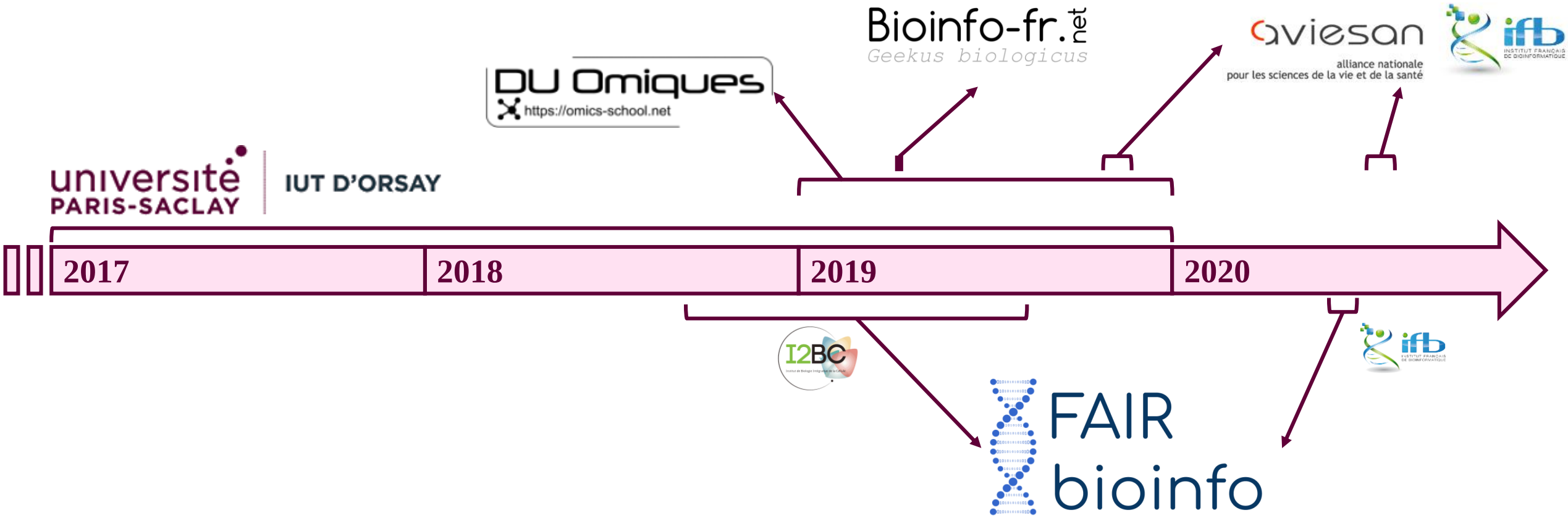
Une thèse variée en projets de recherche et collaborations



Une thèse variée en développement informatique



Une thèse variée en formations



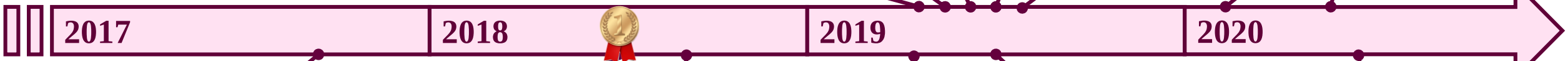
Des travaux de thèse communiqués

Orales

Multi-Omics Data Integration to Model Iron homeostasis in pathogenic yeast *Candida glabrata*

FAIR_Bioinfo : La reproductibilité au service de la biologie computationnelle

Réseaux fonctionnels de gènes co-exprimés pour explorer l'homéostasie du fer chez la levure pathogène *Candida glabrata*



Affichées

Study of the adaptation to iron deficiency of the pathogenic yeast Candida glabrata : bioinformatics analyses of multi-omics data

Pixel: an open source solution for annotation, storage, mining and integration of multi-omics data in biology

Transcriptomics data explorations to decipher iron homeostasis in the pathogenesis yeast *Candida glabrata*

Multi-Omics Data Integration to Model Iron homeostasis in pathogenic yeast *Candida glabrata*

Systematic Analysis of Protein Post-translational Modifications at a Proteomic Scale in the pathogenic yeast *Candida albicans*

-

Functional networks of co-expressed genes to explore iron homeostasis processes in the pathogenic yeast *Candida glabrata*

-

FAIR-bioinfo : disseminate the tools accompanying the reproducibility of analyses

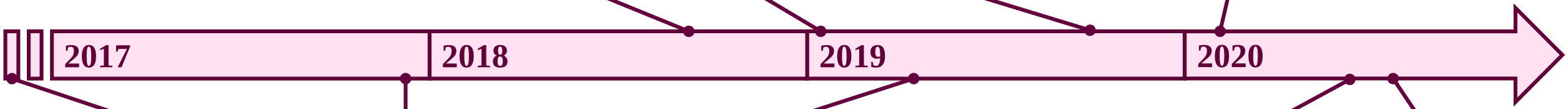
Des travaux de thèse publiés

Pixel: a content management platform for quantitative omics data.
Denecker *et al.* - PeerJ

Functional networks of co-expressed genes to explore iron homeostasis processes in the pathogenic yeast *Candida glabrata*
Denecker *et al.* - NAR Genomics and Bioinformatics

Empowering the detection of ChIP-seq "basic peaks" (bPeaks) in small eukaryotic genomes with a web user-interactive interface.
Denecker *et al.* - BMC Data Note

Label-free quantitative proteomics in Candida yeast species: technical and biological replicates to assess data reproducibility.
Lelandais, Denecker *et al.* - BMC Data Note



Characterization of the replication timing program of 6 human model cell lines
Hadjadj D, Denecker T *et al.*
Genomic Data

A hypothesis-driven approach identifies CDK4 and CDK6 inhibitors as candidate drugs for treatments of adrenocortical carcinomas.
Hadjadj *et al.* - Aging

Rendre ses projets R plus accessibles grâce à Shiny
Denecker - Bioinfo-fr.net

Efficient, quick and easy-to-use, DNA replication timing analysis with START-R suite.
Hadjadj, Denecker *et al.*
NAR Genomics and Bioinformatics

FAIR_Bioinfo: a turnkey training course and protocol for reproducible computational biology
Denecker et Toffano-Nioche
HAL

Merci pour ces belles collaborations !

iHKGviewer — **Projet *Nakaseomyces***

Equipe Fairhead
Adela Angoulvant
Monique Bolotin-Fukuhara
Cécile Fairhead
Laetitia Maroc
Youfang Zhou-Li

MONET Multi-Omics NETworks — **Projet *PTMs***

Equipe Camadro
Jean-Michel Camadro
Véronique Legros
Laurent Lignières
Pierre Poulain
Nicolas Senecaut
Samuel Terrier

START-R <Project/>

Equipe Cadoret
Giuseppe Baldacci
Jean-Charles Cadoret
Anne-Lise Haenni
Fabien Fauchereau
Su-Jung Kim
Chrystelle Maric-Antoinat

PodoGOviewer

Equipe Malagnac
Pierre Grognet
Fabienne Malagnac
Damien Remy

FAIR bioinfo

I2BC / IFB
Claire Toffano-Nioche
Céline Hernandez
Hélène Chiapello
Jacques van Helden

Et nos testeurs
Stéphane Demais et Pauline François

PIXEL

Entreprises TailorDev et Biorosetics

ifb INSTITUT FRANÇAIS DE BIOINFORMATIQUE

Task force
Gildas Le Corguillé
Julien Seiler

Merci aussi

Aux personnes qui ont rendu l'administratif facile

Marie-Hélène Sarda, Jeanne Triki et Sandrine Le Bihan

Aux membres du jury

Sarah Cohen Boulakia, Bertrand Cosson, Marie-Agnès Dillies, Stéphane Le Crom, Hélène Chiapello, Jean-Michel Camadro et Pierre Poulain

À ma directrice de thèse Gaëlle Lelandais

À mes proches

Informations importantes

N’hésitez pas à poser des questions dans le chat, Pierre Poulain se chargera de me les poser à la fin.

Merci pour votre écoute !