Application de la biologie des systèmes pour l'identification de marqueurs moléculaires des maladies rénales dans les fluides biologiques

Franck Boizard

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par l'Université Toulouse 3 - Paul Sabatier

Présentée et soutenue par

Franck BOIZARD

Le 8 octobre 2019

Application de la biologie des systèmes pour l'identification de marqueurs moléculaires des maladies rénales dans les fluides biologiques

---

Ecole doctorale : **BSB - Biologie, Santé, Biotechnologies**

Spécialité : **BIO-INFORMATIQUE, GENOMIQUE ET BIOLOGIE DES SYSTEMES**

Unité de recherche :
I2MC - Institut des Maladies Métaboliques et Cardiovasculaires

Thèse dirigée par
Joost Peter SCHANSTRA et Olivier TESTE

Jury

M. Vincent Fromion, Rapporteur
Mme Sandra Bringay, Rapporteure
M. Stanislas Faguer, Examinateur
M. Joost Peter SCHANSTRA, Directeur de thèse

Franck BOIZARD

Application de la biologie des systèmes pour l'identification de marqueurs moléculaires des maladies rénales dans les fluides biologiques

Directeurs de thèse :
Joost-Peter SCHANSTRA, Directeur de recherche – I2MC
Olivier TESTE, Professeur de l'Université Toulouse 2 – Jean Jaurès

Résumé

Les maladies rénales concernent 5 millions de personnes en France et ce nombre ne cesse de croître compte tenu de l'augmentation de l'espérance de vie et de l'évolution de nos modes de vie (sédentarité, alimentation). La prise en charge des malades est actuellement peu efficace en raison d'un diagnostic trop tardif et de notre méconnaissance des mécanismes complexes qui régissent leur progression. L'étude du protéome urinaire s'est imposée comme un excellent moyen pour découvrir des biomarqueurs des néphropathies et ainsi mieux comprendre les mécanismes physiopathologiques. La biologie des systèmes permet d'exploiter l'information moléculaire contenue dans l'urine pour en déduire l'organisation globale des réseaux de régulation dans le tissu rénal malade. C'est dans ce contexte que se situe ce travail de thèse.

Deux problématiques ont été abordées dans cette thèse :

La première relève de la compréhension des mécanismes physiopathologiques : «Comment identifier de nouveaux acteurs clés dans le développement des maladies rénales à partir de l'analyse de la composition moléculaire de l'urine ?» . L'information du protéome urinaire se limitant majoritairement aux protéines excrétées, il est essentiel d'avoir à disposition des méthodes d'analyses bio-informatiques pour "remonter" aux protéines clés présentes dans le tissu rénal, mais non excrétées dans l'urine. Ce type de méthodes étant peu utilisé en néphrologie, nous avons développé un outil méthodologique fiable pour identifier in silico de nouveaux acteurs clés des maladies rénales à partir de l'analyse du protéome urinaire. Ce nouvel outil, appelé PRYNT (PRioritization bY causal NeTwork), repose sur l'utilisation des interactions protéine-protéine associée à une méthode de priorisation pour repérer les protéines du réseau qui interagissent préférentiellement avec les biomarqueurs protéines urinaires.

La seconde problématique s'inscrit dans une démarche de médecine diagnostique, la question étant : «Comment détecter la présence d'une maladie rénale ou prédire son évolution à partir de l'analyse de la composition de l'urine ?». J'ai développé une approche quantitative pour proposer une réponse à cette question. J'ai ensuite appliqué cette approche au métabolome de l'urine et au peptidome du liquide amniotique qui reflètent la fonction rénale. La modélisation et les méthodes statistiques permettent dans ce contexte de prédire la maladie.

Institut des Maladies Métaboliques et Cardiovasculaires – Inserm/UPS UMR 1048
Institut de Recherche en Informatique de Toulouse – UMR 5505 CNRS

Franck BOIZARD

The use of systems biology for the identification of biomarkers of renal diseases in biological fluids

Supervisors:
Joost-Peter SCHANSTRA, Director of research –I2MC
Olivier TESTE, Professor at Toulouse 2 University – Paul Sabatier

Abstract

Kidney disease affects about 5 million people in France mostly due to the increase in life expectancy and the evolution of our lifestyles (sedentary living, diet). Patient management is currently largely ineffective due to late diagnosis and our lack of understanding of the complex mechanisms that govern its progression. The study of the urinary proteome has emerged as an excellent way to discover biomarkers of nephropathies and thus to better understand the underlying pathophysiological mechanisms. Systems biology allows the molecular information contained in urine to be used to understand the overall organization of the regulatory networks in the diseased kidney tissue. In my thesis we have applied systems biology with two aims :

The first aim was to improve the understanding of the pathophysiological mechanisms of kidney disease based on the analysis of urine molecular composition. Since the information in urinary proteome is mainly limited to excreted proteins, it is essential to have bioinformatic analysis methods available to "trace back" the key proteins present in the kidney tissue, but not excreted in the urine. Since this type of method is not widely used in nephrology, I have developed a methodological tool to identify in silico new key actors in kidney disease from the analysis of the urinary proteome. This new tool, called PRYNT (PRioritization bY causal NeTworks), is based on the use of protein-protein interactions with a prioritization method to identify proteins in the network that preferentially interact with urinary protein biomarkers.

The second aim of my thesis was to develop systems biology approaches for the detection and progression of kidney disease using the molecular composition of urine. We developed a quantitative approach to propose an answer to these questions. I then applied this approach to the analysis of the urinary metabolome and amniotic fluid peptidome. Modelling and statistical methods allowed in these contexts to predict the presence of kidney disease and its progression.

Institut des Maladies Métaboliques et Cardiovasculaires – Inserm/UPS UMR 1048
Institut de Recherche en Informatique de Toulouse – UMR 5505 CNRS

Table des matières

Introduction générale

> "Systems biology ... is about putting together rather than taking apart, integration rather than reduction. It requires that we develop ways of thinking about integration that are as rigorous as our reductionist programmes, but different. ... It means changing our philosophy, in the full sense of the term."
>
> *Denis Noble (2006)*

L'Augmentation de la connaissance et la complexité des domaines de recherche ont poussé les scientifiques à l'ultra-spécialisation (Guespin-Michel et Ripoll, 2000). Ce constat est particulièrement vrai en biologie où les disciplines telles que la biochimie, la biologie moléculaire, la biologie cellulaire ou la physiologie par exemple sont souvent étudiées de manière très cloisonnée. Or, pour comprendre la complexité d'un système dynamique multi-échelles, que ce soit une cellule, un tissu ou un organisme, il est nécessaire de mobiliser les connaissances de différentes spécialités. Depuis les années 2000, de nouvelles approches voient ainsi le jour afin d'accompagner cette interdisciplinarité. La biologie des systèmes est l'une d'entre elles. Son but est d'intégrer différents niveaux d'informations pour comprendre les interactions entre les différents composants du système biologique, proposer une modélisation des fonctions et ainsi appréhender le vivant dans sa totalité. La biologie systémique combine une approche expérimentale et une approche théorique dans laquelle les mathématiques et l'informatique occupent une place centrale : le développement de technologies performantes conduit à l'identification simultanée d'un grand nombre de molécules, la généralisation de l'informatique assure la gestion de ces masses d'informations dans des bases de données, l'uniformisation des notations permet d'échanger et de comparer ces données et grâce à l'appui d'outils statistiques et mathématiques, les relations entre les données peuvent être modélisées. Le développement de la biologie des systèmes ouvre ainsi de nouvelles perspectives notamment dans le champ de la médecine pour la compréhension des maladies et le développement de cibles thérapeutiques potentielles.

Les maladies rénales concernent 5 millions de personnes en France et ce nombre ne cesse de croître compte tenu de l'augmentation de l'espérance de vie et de l'évolution de nos modes de vie (sédentarité, alimentation) (Zhang et Rothenbacher, 2008; Hill *et al.*, 2016). Chez les patients concernés, les reins n'assurent plus leurs fonctions essentielles : les déchets ne sont plus suffisamment éliminés hors de l'organisme et la composition en eau et en ions du corps n'est plus maintenue de manière optimale. Dans les cas les plus graves, c'est l'insuffisance rénale

terminale qui nécessite le recours à des traitements lourds de suppléance tels que la dialyse ou la transplantation rénale. La prise en charge des malades est actuellement peu efficace en raison d'un diagnostic trop tardif et de notre méconnaissance des mécanismes complexes qui régissent leur progression.

L'accès au tissu rénal constitue une étape essentielle pour décortiquer la physiopathologie des maladies rénales (Glassock, 2015). Cependant, la biopsie rénale est une intervention chirurgicale, donc invasive, et une source de complications pour le patient (Corapi *et al.*, 2012; Hogan *et al.*, 2015). La recherche médicale en néphrologie se tourne de plus en plus vers l'étude des fluides biologiques (sang, urine) dont le prélèvement est peu invasif et dépourvu de risque (Csősz *et al.*, 2017; Mischak, 2015; Voss *et al.*, 2011). L'urine constitue un liquide de premier choix dans le cas particulier des maladies rénales. En effet, la composition de l'urine est un excellent reflet de la fonction rénale dans la mesure où c'est grâce aux reins qu'elle est synthétisée à partir du plasma sanguin (Barratt et Topham, 2007). De plus, l'évolution des technologies de chimie analytique et notamment l'apparition de la spectrométrie de masse permettent aujourd'hui d'identifier des milliers de molécules, issues de différents strates moléculaires (ARN messagers, micro-ARN, protéines, métabolites ...), dans l'urine de sujets sains ou atteints de maladie rénale. Deux démarches sont généralement utilisées en biologie systémique pour définir les réseaux moléculaires qui reproduisent les comportements fonctionnels du vivant : une approche dite "descendante" (*top down*), dans laquelle les êtres vivants sont simplifiés pour en déterminer les éléments essentiels, et une approche dite "ascendante" (*bottom up*) qui vise au contraire à reconstituer des éléments complexes du vivant à partir d'unités plus simples (Bruggeman et Westerhoff, 2007). Grâce à une démarche de type ascendante, la biologie des systèmes permet d'exploiter l'information moléculaire contenue dans l'urine pour en déduire l'organisation globale des réseaux de régulation dans le tissu rénal malade. C'est dans ce contexte que se situe ce travail.

Deux problématiques ont été abordées au cours de cette thèse :

— La première relève de la compréhension des mécanismes physiopathologiques : « Comment identifier de nouveaux acteurs clés dans le développement des maladies rénales à partir de l'analyse de la composition moléculaire de l'urine ? »

— La seconde s'inscrit plus dans une démarche de médecine diagnostique, la question étant « Comment détecter la présence d'une maladie rénale ou prédire son évolution à partir de l'analyse de la composition de l'urine ? »

Chacune de ces parties fera l'objet d'un chapitre de ce manuscrit.

Première partie

Sélection de protéines importantes dans les maladies rénales

Introduction

Le phénotype d'un organisme découle du fonctionnement d'un grand nombre de molécules, incluant les protéines, les gènes ou encore les métabolites. Mais il dépend également fortement des relations qui s'opèrent entre ces molécules (interactions protéine-protéine, réactions métaboliques, co-expressions...) (Hartwell *et al.*, 1999). L'ensemble de ces interactions intermoléculaires est appelé interactome ; il peut être modélisé sous la forme d'un réseau. L'intérêt pour l'étude des réseaux a fortement évolué au début des années 2000 (Luke et Harris, 2007), avec l'augmentation de la puissance de calcul des ordinateurs et le développement de la biologie des systèmes. D'abord développés chez la levure, les réseaux biologiques sont rapidement passés de quelques dizaines à plusieurs milliers de molécules (Schwikowski *et al.*, 2000; Ho *et al.*, 2002) et la recherche s'est alors focalisée sur l'exploration de leur structure(Albert et Albert, 2004; Maslov, 2002; Jeong *et al.*, 2001; Han *et al.*, 2004). De tels réseaux, bien que plus complexes, ont ensuite été obtenus chez l'homme (Rual *et al.*, 2005) et c'est à partir de ces connaissances qu'est né en 2007 un nouveau concept dans le domaine de la biologie des systèmes, celui de la « médecine par réseaux » (*network medicine*) (Barabási, 2007). Ce concept sous-tend que l'utilisation des réseaux permettrait de modéliser la perturbation simultanée des différentes entités moléculaires au cours des pathologies. Autrement dit, il serait possible de comprendre les processus pathologiques dans leur globalité, et non plus en se limitant à la seule échelle de la mutation génique.

L'étude de la composition de l'urine a connu un progrès spectaculaire grâce à la spectrométrie de masse. Depuis l'étude d'Adachi en 2006 qui fut précurseure dans l'identification à large échelle du protéome urinaire (Adachi *et al.*, 2006), plus de 6000 protéines ont été détectées dans l'urine (Zhao *et al.*, 2017). Cependant, les protéines présentes dans le tissu rénal ne sont pas nécessairement toutes retrouvées dans l'urine et inversement, toutes les protéines urinaires ne sont pas issues du rein. La base de données Human Protein Atlas (Uhlen *et al.*, 2015) répertorie plus de 10000 protéines ayant une expression rénale, mais seulement 4095 d'entre elles sont présentes parmi les 6000 protéines urinaires connues à l'heure actuelle (Figure I.1). Bien que l'urine reste un outil essentiel dans l'identification de protéines associées à des pathologies rénales, il est donc important de se rappeler que l'étude isolée du protéome urinaire implique une vision limitée des mécanismes moléculaires mis en œuvre *in-situ* dans le rein.

Nous verrons dans cette partie comment l'analyse des réseaux d'interactions protéine-protéine peut répondre à notre problématique. Notamment en quoi les analyses de centralité, que nous détaillerons plus loin, permmettent d'identifier les protéines importantes du réseau. En reliant dans un même modèle les protéines urinaires et les protéines rénales, les réseaux d'interactions protéine-protéine pourraient prédire, à partir des modifications du protéome urinaire, de nouvelles protéines importantes dans la maladie mais non-détectées dans l'urine.

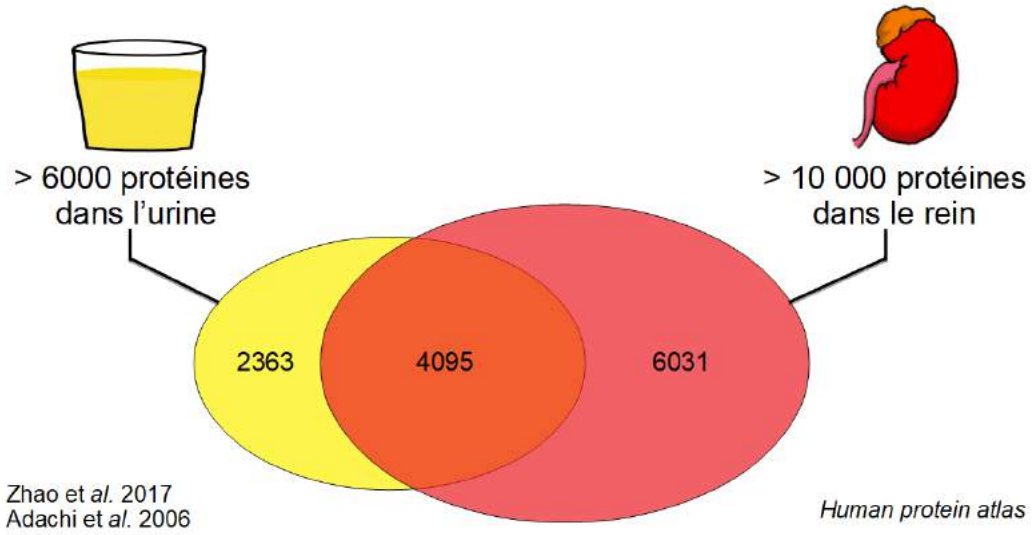

Figure I.1 – Comparaison des protéines détectées dans l'urine et dans le tissu rénal. L'étude d'Adachi *et al.* (Adachi *et al.*, 2006) et l'étude de Zhao *et al.* (Zhao *et al.*, 2017), ont détecté ensemble 6458 protéines différentes dans l'urine de patients sains. La base de données Human protein Atlas (Uhlen *et al.*, 2015), répertoria (au 01.03.2018) 10 126 protéines déjà détectées dans le tissu rénal de patients sains. Seulement 4095 de ces protéines ont déjà été détectées dans l'urine. Si la composition du protéome urinaire est donc un bon reflet du protéome rénal, il n'est donc pas actuellement possible de détecter toutes les protéines exprimées dans le rein grâce à l'urine.

Ce chapitre sera découpé en deux parties. Dans une première partie, je ferai un état de l'art sur les réseaux d'interaction protéine-protéine, les centralités et les méthodes actuelles d'identification des mécanismes pathologiques. Dans la seconde partie, je présenterai les résultats de mes travaux qui proposent une nouvelle méthode (PRYNT) pour prédire les protéines clés des maladies rénales à partir de l'analyse de la composition urinaire.

1 Réseaux d'interactions protéine-protéine et centralités pour l'identification des acteurs clés des maladies

"There are no interactions more interesting and important than those between proteins."

J. L. Oncley et al. *(1952)*

1.1 Réseaux d'interactions protéine-protéine

1.1.1 Interactions protéine-protéine

Les protéines exercent leurs fonctions en coopération les unes avec les autres (Gonzalez et Kann, 2012). Un complexe protéique est l'assemblement de plusieurs protéines par des interactions protéine-protéine (PPI) et on estime que 80 % des protéines exercent leur fonction en formant des complexes protéiques (Berggård *et al.*, 2007). Par exemple, l'ADN polymérase, essentielle à la réplication de l'ADN dans le noyau de toutes les cellules, est un complexe protéique dont la structure et la fonction sont conditionnées par les PPI (Garcia-Diaz et Bebenek, 2007). Le ribosome, impliqué dans la synthèse des protéines dans la cellule, est un complexe composé de pas moins de 80 protéines (Wool, 1979; Ishii *et al.*, 2006). La communication entre les cellules (Pawson, 2004) et le transport des protéines dans le plasma (Kanai *et al.*, 1968; Zanotti *et al.*, 2008) sont également assurés par des PPI. Si l'on veut comprendre l'organisme en tant que système global, il est donc indispensable de comprendre les PPI.

Les PPI peuvent être classées de plusieurs manières selon leurs propriétés structurelles et fonctionnelles (De Las Rivas et Fontanillo, 2010; Keskin *et al.*, 2016; Nooren, 2003; Phizicky et Fields, 1995). Le terme 'interaction' entre protéines est traditionnellement employé pour parler d'un contact physique entre deux protéines. L'interaction est alors qualifiée de directe. L'interaction physique peut être binaire ou macromoléculaire si elle forme respectivement un dimère ou

un polymère. Elle est homo-oligomérique si les protéines impliquées sont les mêmes mais hétéro-oligomérique si les protéines sont différentes. De même, la durée de l'interaction détermine si elle est permanente ou transitoire et sa stabilité la rend obligatoire ou non. Une interaction physique donnée se caractérise alors par la combinaison de ces propriétés. Mais une interaction entre deux protéines n'est pas nécessairement physique (De Las Rivas et de Luis, 2004). Dans ce cas, on parle d'interaction fonctionnelle ou indirecte. Une interaction fonctionnelle existe entre deux protéines si elles sont impliquées dans une même activité biomoléculaire ou si elles sont co-localisées dans un même compartiment cellulaire.

Beaucoup de maladies sont dues à des mutations génétiques qui altèrent les propriétés d'interaction (directe ou indirecte) d'une protéine avec une autre protéine (Schuster-Böckler et Bateman, 2008). La possibilité d'agir sur les PPI constitue donc une piste pour le développement de cibles thérapeutiques (Zinzalla et Thurston, 2009; Arkin et Wells, 2004). Par conséquent, la compréhension des relations qui lient directement ou indirectement les protéines entre elles constituent un enjeu majeur en biologie et en clinique.

1.1.2 Identification des interactions protéine-protéine

Il existe un grand nombre de techniques disponibles pour étudier les interactions protéine-protéine. Snider *et al.* relèvent 4 caractéristiques clés les distinguant : (i) le nombre de PPI détectées, (ii) le type de PPI détectées, (iii) les contraintes de temps et de coûts liées à l'analyse des PPI, (iv) la nature des outils mis en œuvre. Beaucoup de publications présentent un état de l'art des différentes technologies identifiant les PPI (Petschnigg *et al.*, 2011; Stynen *et al.*, 2012; Rao *et al.*, 2014; Snider *et al.*, 2015; Keskin *et al.*, 2016; Miura, 2018). Il est toutefois difficile de comparer les performances de ces différentes méthodes car elles utilisent des technologies et des références différentes. Dans la mesure où elles permettent d'identifier des PPI variées, ces approches doivent plutôt être vues comme complémentaires (Jensen et Bork, 2008). Il existe deux grandes catégories de méthodes d'étude des PPI : les méthodes expérimentales de détection des PPI et les méthodes computationnelles de prédiction. Ce chapitre a pour but de les présenter en insistant sur les avantages et les inconvénients de chacune d'elles.

Méthodes expérimentales de détection des PPI

Plusieurs méthodes expérimentales sont disponibles pour détecter les PPI (Tableau I.1.1). Elles se différencient par (i) le type d'expérience utilisée - à bas débit ou à haut débit (Gonzalez et Kann, 2012; Safari-Alighiarloo *et al.*, 2014), (ii) le contexte de l'expérience - *in vitro* ou *in vivo* (Rao *et al.*, 2014) et (iii) le type d'interactions qu'elles détectent - directe ou indirecte, binaire ou macromoléculaire.

Technique	Principe	Contexte	Type d'interaction		Référence
			direct / indirect	binaire / complexe	
Méthodes bas débit					
Cristallographie aux rayons X	Analyse de la structure des protéines en 3D par rayon X	in vitro	directe	binaire	(Smyth et Martin, 2000)
Spectroscopie résonance magnétique nucléaire (RMN)	Analyse de la structure des protéines en 3D grâce aux magnétisme nucléaire	in vitro	directe	binaire	(O'Connell *et al.*, 2009)
Chromatographie d'affinité	Technique de chromatographie séparant un composé biologique grâce aux PPI	in vitro	directe	binaire	(Belanger, 2009)
Transfert d'énergie entre molécules fluorescentes	Observation au microscope de la proximité des protéines par fluorescence	in vivo	directe	binaire	(Kenworthy, 2001)
Co-immunoprecipitation	Les anticorps se liant avec une protéines en solution précipiteront avec le complexe auquel appartient la protéine. La répétition de ce processus avec des anticorps différent permet de connaitre les différentes protéines constituant un complexe.	in vitro	directe	complexe	(Phizicky et Fields, 1995)
Méthodes haut débit					
Luminescence-based mammalian interactomemapping (LUMIER)	Technique de co-immunoprécipitation mesurant des réactions de bioluminescence	in vivo	directe	binaire	(Barrios-Rodiles *et al.*, 2005)
Purification par affinité en couplé à la spectrométrie de masse	Analyse basée sur le double marquage de la protéine d'intérêt suivi d'un processus de purification et analyse spectroscopique de masse	in vitro	directe	binaire / complexe	(Kaiser *et al.*, 2008)
Technique de double hybride	Détection de l'activité d'un gène rapporteur fixé aux protéines dans une cellule. L'organisme le plus utilisé étant la levure	in vivo	directe	binaire	(Fields et Song, 1989)
Létalité synthétique	Technique analysant la viabilité de la mutation combinatoire des gènes sur la cellule	in vivo	indirecte	binaire / complexe	(Tucker et Fields, 2003)

Tableau I.1.1 – Méthodes expérimentales d'identification des interactions protéine-protéine

Bas débit / haut débit

Les méthodes expérimentales à bas débit sont les méthodes traditionnelles d'identification des PPI. Leur avantage principal réside dans le fait qu'elles permettent de déterminer les caractéristiques des interactions. La cristallographie par diffraction de rayons X (Parker, 2003) et la résonance magnétique nucléaire (RMN) (Wüthrich, 2001) sont deux techniques qui réalisent des structures tridimensionnelles des domaines d'interaction. Fort de cette information, il est alors possible de concevoir des molécules capables de cibler des protéines particulières dans un but thérapeutique (Rao *et al.*, 2014). D'autres techniques comme le FRET ou le BRET, qui reposent sur un transfert d'énergie entre deux protéines en interaction, permettent quant à elles de déterminer la stabilité des PPI (Kenworthy, 2001). De manière générale, les méthodes à bas débit sont utilisées comme standard pour confirmer les interactions déterminées par d'autres méthodes (Miura, 2018). Ces techniques demandent cependant beaucoup d'investissement et elles n'examinent qu'un petit nombre de protéines simultanément.

Les technologies à haut débit grâce à l'automatisation informatique conduisent à la découverte de PPI à plus large échelle. Ce sont des méthodes peu chères, capables d'identifier un grand nombre d'interactions en une seule expérience. L'inconvénient de ces méthodes est le nombre élevé de faux positifs détectés (Huang *et al.*, 2007).

In vitro / in vivo

L'expérimentation *in vitro* se fait dans un environnement contrôlé à l'extérieur de l'organisme vivant. À l'inverse, c'est dans l'organisme vivant que s'applique l'expérience *in vivo*. Les techniques d'identification *in vitro* des PPI sont plus faciles à mettre en œuvre. La spectrométrie RMN et la chromatographie d'affinité permettent notamment de détecter des interactions de faible affinité. Ces méthodes peuvent toutefois détecter une interaction entre deux protéines qui, dans l'organisme, ne seront peut-être jamais en contact.

Les expérimentations réalisées *in vivo*, bien que plus complexes, sont bien-sûr plus attractives car plus proches de la réalité biologique. Les méthodes basées sur des transferts d'énergie entre partenaires d'un complexe protéique permettent même d'observer les interactions en temps réel *in vivo* (Kenworthy, 2001).

Directe / indirecte

Les techniques que nous avons vues jusqu'à présent identifient des PPI physiques (directes). D'autres méthodes expérimentales sont utilisées pour mettre en évidence les liens fonctionnels entre les protéines (indirectes). Par exemple, la technique de létalité synthétique consiste à induire une mutation de plusieurs gênes et à évaluer l'effet de ces mutations, seules ou combinées, sur la viabilité cellulaire (Tucker et Fields, 2003). Si la combinaison des mutations entraîne la mort cellulaire, alors que les mutations individuelles sont sans effet, cela indique que les protéines codées par les gènes considérés exercent leur fonction en coopération. Il s'agit alors de PPI fonctionnelles.

Binaire / macromoléculaire

Certaines méthodes, comme la technique de double hybride, mesurent les interactions physiques entre seulement deux protéines (binaire). D'autres mesurent les interactions macromoléculaires entre un groupe de protéines (De Las Rivas et Fontanillo, 2010). La méthode la plus utilisée est la purification par affinité couplée à la spectrométrie de masse. La protéine d'intérêt est taguée et captée en même temps que les protéines potentiellement attachées. Le groupe de protéines est ensuite purifié et analysé par spectrométrie de masse. L'inconvénient de ces techniques est qu'il est indispensable d'analyser les données, pour savoir spécifiquement quelles protéines sont capables de se fixer ensemble (Hakes *et al.*, 2007).

Méthodes de prédiction des PPI

Les méthodes computationnelles sont utilisées pour prédire *in silico* les PPI (Tableau I.1.2). Elles constituent une approche complémentaire souvent plus rapide et moins coûteuse que les méthodes expérimentales. Elles permettent d'identifier un grand nombre d'interactions potentielles qui pourront par la suite être confirmées par des méthodes expérimentales. Ces méthodes de prédiction sont très diverses en fonction de la nature des données initiales, des algorithmes utilisés ou des concepts sous-jacents.

Empirique / théorique

Les méthodes empiriques de prédiction utilisent des données expérimentales relatives aux PPI pour prédire de nouvelles PPI. Les approches fondées sur le *machine-learning* (« apprentissage automatique ») en sont un très bon exemple. Ces approches utilisent les propriétés déjà connues des PPI, par exemple les séquences en acides aminés ou les domaines impliqués, comme critères pour détecter de nouveaux PPI. Puisque les méthodes empiriques reposent sur l'utilisation de données expérimentales, elles exploitent aussi leurs inexactitudes.

Les méthodes dites théoriques quant à elles tiennent compte d'un grand nombre de concepts biologiques. *Le docking* (Vakser, 2014) étudie la complémentarité entre les structures 3D de deux protéines pour suggérer l'existence d'une interaction. Les méthodes basées sur la coévolution exploitent le fait que la fonction biologique de certaines protéines est conservée au cours de l'évolution malgré la modification de leur séquence en acides aminés : cette conservation de la fonction implique alors que des protéines qui interagissent ensemble n'évoluent pas de manière indépendante. La force des méthodes de prédiction théoriques réside dans le fait qu'elles sont capables de prédire des interactions entre des protéines pour lesquelles peu de choses sont connues.

Finalement, les technologies d'identification des PPI, qu'elles soient expérimentales ou basées sur des stratégies *in silico*, ont largement progressé depuis 20 ans. On estime que l'on ne connaît que 20 % de l'interactome humain (Venkatesan *et al.*, 2009; Stumpf *et al.*, 2008; Hart *et al.*, 2006). Malgré cette connaissance partielle, l'interactome disponible dispose d'une couverture suffisante pour explorer les processus biologiques pouvant lier les maladies aux protéines (Menche *et al.*, 2015).

Technique	Données de départ	Principe	Type d'interaction		Référence
			direct / indirect	binaire / complexe	
Méthodes empiriques					
Fréquence des domaines d'interactions	Signatures de la séquence des protéines	Les signatures des séquences des protéines connues pour interagir peuvent prédire de nouvelles interactions.	directe	binaire	(Sprinzak et Margalit, 2001)
Estimation du maximum de vraisemblance des domaines d'interactions	Informations sur les domaines protéiques	Calcul d'une probabilité d'interaction grâce aux domaines connus de PPI.	directe	binaire	(Deng *et al.*, 2002)
Profil de coexpression des gènes	Données d'expression génique	Des protéines dont les gènes sont exprimés avec des profiles similaire ont de forte chance d'interagir.	indirecte	binaire	(Fraser *et al.*, 2004)
Topologie du réseau	PPI binaires	Deux protéines interagissant avec un grand nombre de protéines communes interagissent probablement entre elles.	directe /indirecte	binaire	(Chua *et al.*, 2006)
Approche par machine-learning	Séquences / propriétés biochimiques / structures ...	Algorithmes utilisant les propriétés des PPI connues avec un modèle de prédiction.	directe /indirecte	complexe / binaire	(Shen *et al.*, 2007)
Méthodes théoriques					
Voisinage génétique	Séquences du génome	Deux protéines ayant des gènes très proches ont de forte chance d'interagir	indirecte	binaire	(Ng et Tan, 2004)
Approche coevolutionnaire	Séquences des protéines entre plusieurs organismes	Utilisation de l'arbre philogénétique pour caractériser l'évolution des protéines	indirecte	binaire	(Gertz *et al.*, 2003)
Fusion génétique	Génome complet de plusieurs organismes	Aussi appelé la méthode pierre de Rosette. Deux protéines fusionnées dans un organisme, interagissent probablement dans un organisme ou elles sont séparées.	indirecte	binaire	(Marcotte et Marcotte, 2002)
Text mining	Littérature scientifique	Evaluation de l'interaction de deux protéines en fonction de leur coexistence dans des textes de la littérature scientifique	indirecte	binaire	(Papanikolaou *et al.*, 2015)
Docking	Structure 3D des protéines	Simulation des orientations possibles entre deux protéines	directe	binaire	(Vakser, 2014)

Tableau I.1.2 – Méthodes computationnelles de prédiction des interactions protéine-protéine

1.1.3 Base de données d'interactions protéine-protéine

Avec l'essor des méthodes d'identification, la génération des données PPI a augmenté de manière exponentielle (Figure I.1.1). Ceci a engendré un nouveau besoin de recherches : la construction de bases de données, connue sous le terme de biocuration, pour collecter, annoter et mettre à disposition l'information biologique afin de pouvoir la réutiliser et la partager de manière efficace (Howe *et al.*, 2008; Snyder, 2009). Dans ce contexte, la littérature scientifique est la principale source d'information, et les données peuvent être extraites des publications soit de manière manuelle, soit de manière automatisée (par exemple par *text-mining*).

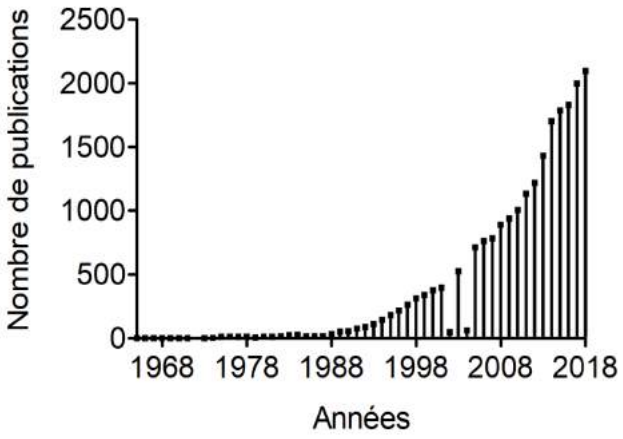

Figure I.1.1 – Augmentation de l'intérêt pour les intéractions protéine-protéine depuis 50 ans. Une recherche sur Pubmed permet de quantifier le nombre de publications scientifiques comportant "*protein-protein interactions*" dans le titre. Ce graphique nous permet d'évaluer l'importance de ce concept dans la littérature du domaine de la biologie médicale.

La construction de la base de données *Yeast protein Database* en 1996 a été l'un des premiers projets collectant à grande échelle des informations en lien avec les protéines (séquence, annotation, localisation et PPI) (Garrels, 1996). Ce travail a démontré qu'il était possible d'assembler un grand nombre de données provenant de sources différentes. Cependant, le challenge du partage des données liées aux PPI s'est rapidement complexifié, du fait de l'augmentation du volume des données à traiter et de la diversité des formats dans lesquels les résultats expérimentaux étaient publiés. La tâche est alors devenue fastidieuse, nécessitant beaucoup de moyens humains et financiers (Howe *et al.*, 2008).

Afin d'améliorer la représentation et l'échange des données PPI, le groupe de travail *Proteomics Standards Initiative*, sous couvert du *Human Proteome Organization* (HUPO-PSI) et soutenu par le consortium *International Molecular Exchange* (IMEx) (Orchard *et al.*, 2012), a développé des directives communautaires standardisées (Hermjakob *et al.*, 2004; Mackay *et al.*, 2007) :

- les directives MIMIx (*Minimum Information about a Molecular Interaction experiment*), afin de guider la publication des données PPI (Orchard *et al.*, 2007).
- le format PS-MI XML, afin de normaliser le partage des données PPI ; ce format est d'ailleurs désormais utilisé par un grand nombre de bases de données.

Base de données	URL	Catégorie	Protéines	PPI
Biogrid	thebiogrid.org	Primaire	23140	473480
DIP	dip.doe-mbi.ucla.edu/ dip/Main.cgi	Primaire	4901	7794
HIPPIe	cbdm-01.zdv.uni-mainz.de/mschaefer/ hippie/index.php	Méta-base	17336	411430
HPRD	hprd.org	Primaire	30047	41327
IntAct	ebi.ac.uk/intact	Primaire	98932	882962
MINT	mint.bio.uniroma2.it	Primaire	11383	48352
PIP	compbio.dundee.ac.uk/www-pips	Méta-base / Prédiction	7751	79441
STRING	string-db.org	Méta-base / Prédiction	16073	3470906

Tableau I.1.3 – Comparaison des bases de données open source de PPI chez l'humain

Le but est de décrire de manière objective toutes les expériences d'interaction moléculaire, en tenant compte de la grande diversité des PPI et de leurs méthodes d'identification et en communiquant les résultats de manière accessible aux outils informatiques. Ces efforts de standardisation sont ainsi principalement centrés sur la description des expériences et l'annotation des données expérimentales : méthode employée pour détecter les PPI, organisme dans lequel les PPI ont été identifiées, liste des molécules participant à l'interaction, etc.

À l'heure actuelle, il existe plus de 300 bases de données spécialisées dans les PPI, dont une centaine chez l'humain. Une liste complète est disponible sur *Pathguide*[1]. Ces bases de données travaillent majoritairement de façon indépendante et peuvent se distinguer en 3 catégories en fonction des méthodes d'identification des PPI ou de curation des données. On trouve ainsi (i) les bases de données primaires, (ii) les méta-bases de données et (iii) les bases de données de prédictions (De Las Rivas et Fontanillo, 2010).

— i) Les bases de données primaires prennent en compte uniquement les PPI démontrées expérimentalement par des méthodes biophysiques. L'extraction et l'intégration des données sont le plus souvent manuelles. Exemples : *Database of Interacting Proteins* (DIP), *Molecular INTeraction database* (MINT) (Licata *et al.*, 2012) et *The Human Protein Reference Database* (HPRD) (Keshava Prasad *et al.*, 2009).
— ii) Les méta-databases combinent les données de plusieurs bases de données primaires. Exemple : *Human Integrated Protein–Protein Interaction rEference* (HIPPIE) (Alanis-Lobato *et al.*, 2017) qui combine les données de 7 bases de données primaires, dont MINT, HPRD et DIP.
— iii) Les bases de données de prédiction incluent, en plus des PPI expérimentales, des PPI prédites grâce aux méthodes de computationnelles. Par ailleurs, la curation peut être manuelle ou automatisée. Exemple : *Search Tool for the Retrieval of Interacting Genes* (STRING) est la plus large base de données PPI de prédiction disponible à l'heure actuelle (Tableau I.1.3).

1. http ://pathguide.org/

Les bases de données primaires, considérées comme fiables, sont souvent utilisées pour valider les PPI prédites par des méthodes computationnelles ou celles extraites par curation automatisée (Yu *et al.*, 2008). De nombreuses bases de données attribuent un score de confiance aux PPI. Plus ce score est haut, plus l'interaction a la probabilité d'exister *in vivo*. Chaque base de données calcule son propre score de notation. Celui-ci tient compte par exemple de la taille de l'expérience, la méthode d'identification (biophysique ou haut-débit, expérimentale ou computationnelle), ou encore le nombre de publications décrivant le PPI. Malgré cela, étant donné les nombreuses méthodes d'identification des PPI et la diversité des bases de données, les bases d'interactions contiennent des informations différentes (Keskin *et al.*, 2016) et ont un faible taux de PPI communes (De Las Rivas et Fontanillo, 2010).

1.1.4 Les réseaux d'interactions protéine-protéine

Les réseaux constituent des ensembles d'éléments (nœuds) interconnectés par des relations particulières (arêtes). Des réseaux existent dans n'importe quel domaine. Un réseau informatique par exemple est un ensemble de machines connectées échangeant des informations entre elles, comme Internet. Un réseau social est un ensemble d'individus entretenant des relations les uns avec les autres, comme un groupe d'amis. En biologie aussi, l'ensemble des interactions intermoléculaires qui s'opèrent dans un organisme est souvent représenté comme un réseau. Dans le cas du réseau modélisant les PPI, les protéines constituent les nœuds du réseau et les liens physiques ou fonctionnels forment les arêtes du réseau.

Les premiers réseaux PPI générés étaient relativement restreints et n'incluaient qu'un certain type de protéines. Par exemple Richter propose en 1975 de modéliser en réseau les PPI impliqués dans la reconnaissance des antigènes par des anticorps (Richter, 1975). Cette modélisation simplifiée des réalités expérimentales lui permet de comprendre quelques phénomènes basiques, à une échelle locale, de la réponse immunitaire. Des réseaux PPI à plus large échelle ont ensuite été développés sur des organismes modèles. Schwikowski *et al.* (Schwikowski *et al.*, 2000) ont ainsi construit un réseau chez la levure comprenant 2358 interactions entre 1548 protéines. Ce n'est qu'en 2005 que des réseaux PPI se focalisant sur l'homme ont vu le jour (Rual *et al.*, 2005; Stelzl *et al.*, 2005). Ces réseaux ont bénéficié des progrès générés dans le domaine des bases de données (Costanzo, 2000; Mewes *et al.*, 2000) et ils fournissent désormais un modèle mathématique de l'interactome humain permettant aux scientifiques de formuler des hypothèses et de les vérifier (Sevimoglu et Arga, 2014). Arthur D Lander (Lander, 2010) estime d'ailleurs que le réseau de l'interactome humain, qu'il représente à l'image d'une pelote de laine (Figure I.1.2), est la nouvelle icône de la biologie du 21ème siècle compte tenu de l'énorme masse de connaissances qu'il renferme et des avancées scientifiques nécessaires à sa construction.

Les réseaux PPI d'aujourd'hui contiennent des milliers de protéines et d'interactions (Tableau I.1.3) et leur visualisation est un véritable challenge (Gehlenborg *et al.*, 2010; Suderman et Hallett, 2007). La plupart des bases de données ont leur propre système d'interrogation permettant la visualisation de réseaux PPI. Certaines sont facilement accessibles aux biologistes. C'est par exemple le cas pour l'interface KUPNetViz adossée à

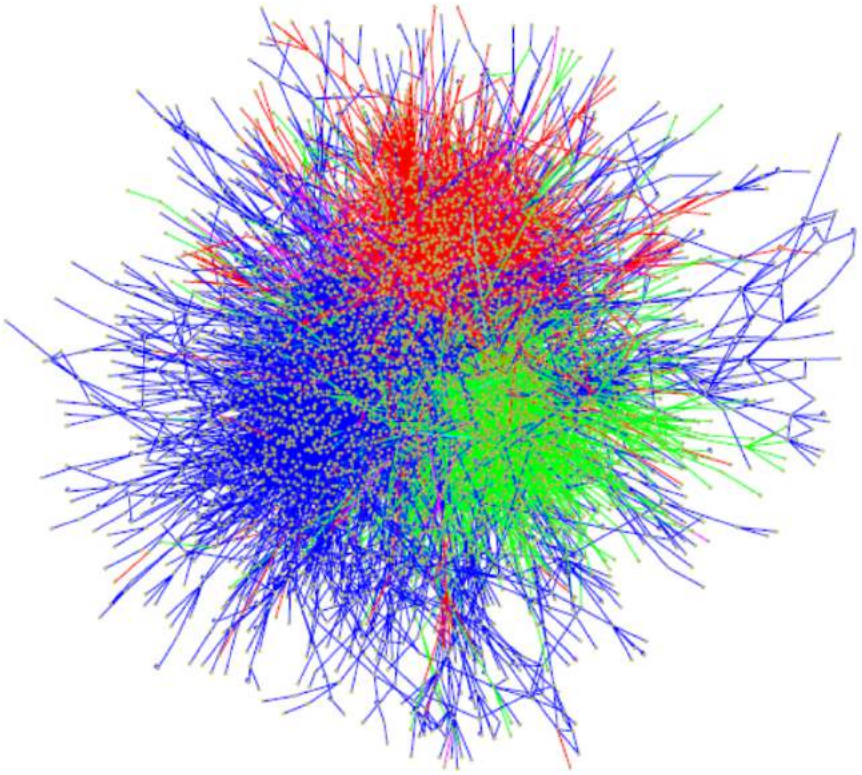

Figure I.1.2 – Représentation de l'interactome humain. Arthur D Lander compare l'interactome humain à une pelote de laine, car les interactions entre les molécules sont aussi intriquées que les fils d'une pelote de laine (Lander, 2010). Les PPI sont une formidable source d'informations mais sa visualisation et son analyse nécessitent des méthodes et des connaissances particulières.

la base de données KUPKB [2] (Klein *et al.*, 2012) ou celle de STRING qui offre la possibilité de visualiser un réseau de 2000 protéines. D'autres en revanche sont populaires chez les bioinformaticiens, mais peu utilisés par les biologistes, du fait de leur complexité. Le logiciel Cytoscape par exemple permet la visualisation de nombreux réseaux, notamment ceux issus de STRING (Doncheva *et al.*, 2018), mais la représentation graphique qu'il donne de la totalité des PPI actuellement connus est très dense et peu informative (Keskin *et al.*, 2016) (Figure I.1.2). L'exploitation des réseaux PPI passera donc par une amélioration des logiciels de visualisation associée à la réduction du fossé qui existe entre biologistes et informaticiens.

Les réseaux PPI constituent un point de départ des recherches en biologie des systèmes. C'est en effet par l'étude de leur organisation et de leur comportement qu'il sera possible de mieux comprendre le fonctionnement des systèmes biologiques.

1.1.5 Structures des réseaux d'interactions protéine-protéine

Les recherches décrivant la structure des réseaux posent les bases de l'analyse moderne de réseaux biologiques. D'une manière générale, tous les réseaux biologiques sont gouver-

2. www.kupkb.org

nés par des lois universelles communes (Barabási et Oltvai, 2004). Les 3 principales lois sont : l'effet petit monde, l'invariance d'échelle et la transitivité.

L'effet petit monde (*small world effect*) (Watts et Strogatz, 1998) est connu des réseaux sociaux sous le nom de la « théorie des 6 poignées de main » (Karinthy, 1929). Appliqué aux PPI, cela signifie que deux protéines sont toujours reliées par un chemin comprenant au maximum 6 interactions. Cet effet petit monde explique pourquoi un organisme ou une cellule sont capables de réagir rapidement et efficacement à une perturbation (Albert, 2005).

L'invariance d'échelle (*scale-free network*) a été définie en 2005 pour exprimer l'idée selon laquelle seulement quelques protéines, appelés « hubs », possèdent beaucoup d'interactions avec d'autres protéines alors qu'inversement, la majorité des protéines n'en possède qu'un nombre réduit (Albert, 2005) (Figure I.1.3). Cette propriété est due au fait que les réseaux grandissent par l'ajout successif de nouveaux nœuds qui se fixent préférentiellement à des nœuds ayant déjà beaucoup d'interactions (Barabasi et Albert, 1999). Deux principales conséquences découlent d'une telle structure. En premier, les réseaux biologiques sont très stables, insensibles à la délétion aléatoire de leurs nœuds. En effet, un réseau restera quasi-entier malgré la suppression aléatoire de 80 % de ses nœuds (Albert *et al.*, 2000) puisque cette suppression affectera principalement les nœuds ayant un petit nombre de relations. En revanche, la structure des réseaux est très vulnérable à la suppression ciblée de ses nœuds essentiels, les hubs. Il a été démontré que la majorité des protéines hubs joue un rôle important dans la survie cellulaire (Jeong *et al.*, 2001) ; on comprend donc aisément que la désorganisation du réseau induite par l'altération de ces hubs sera lourde de conséquences pour les cellules, au point de mettre en péril leur survie.

Enfin, la notion de *transitivité*, étudiée depuis longtemps dans les réseaux sociaux, peut se résumer à l'adage « l'ami de mon ami est mon ami » (Holland et Leinhardt, 1971). Transposée aux réseaux de PPI, cette notion signifie que deux protéines qui interagissent avec une même troisième, ont de fortes chances d'interagir entre elles. Ces agrégations de protéines en interaction constituent des petits groupes appelés modules (Yeger-Lotem *et al.*, 2004; Gavin *et al.*, 2002) (Figure I.1.4). Les protéines de ces modules forment généralement un groupe fonctionnellement cohérent (Hartwell *et al.*, 1999). La réciproque est également vraie puisqu'il a été observé qu'un ensemble de protéines ayant une fonction commune appartiennent à un même bloc de modules dans le réseau (Yook *et al.*, 2004). Ces modules peuvent adopter des formes particulières en fonction du nombre de protéines impliquées et du nombre de relations qui lient ces protéines. Ainsi, un module carré par exemple est constitué de 4 protéines, chacune étant liée à seulement 2 autres protéines du module . Un module dans lequel les protéines, quel que soit leur nombre, sont toutes connectées entre elles est un module particulier appelé clique (Giot, 2003; Albert, 2005). L'existence de ces différentes formes de modules repose sur des phénomènes biologiques particuliers (Yeger-Lotem *et al.*, 2004). Par exemple, la duplication du gène codant une protéine A interagissant avec B et C conduira à la production d'une protéine A', proche de A, qui elle aussi interagit avec B et C ; il se formera de fait le module carré ABA'C dans le réseau (Force *et al.*, 1999). De même, les protéines interagissant les unes avec les autres au sein des complexes macromoléculaires expliquent les cliques.

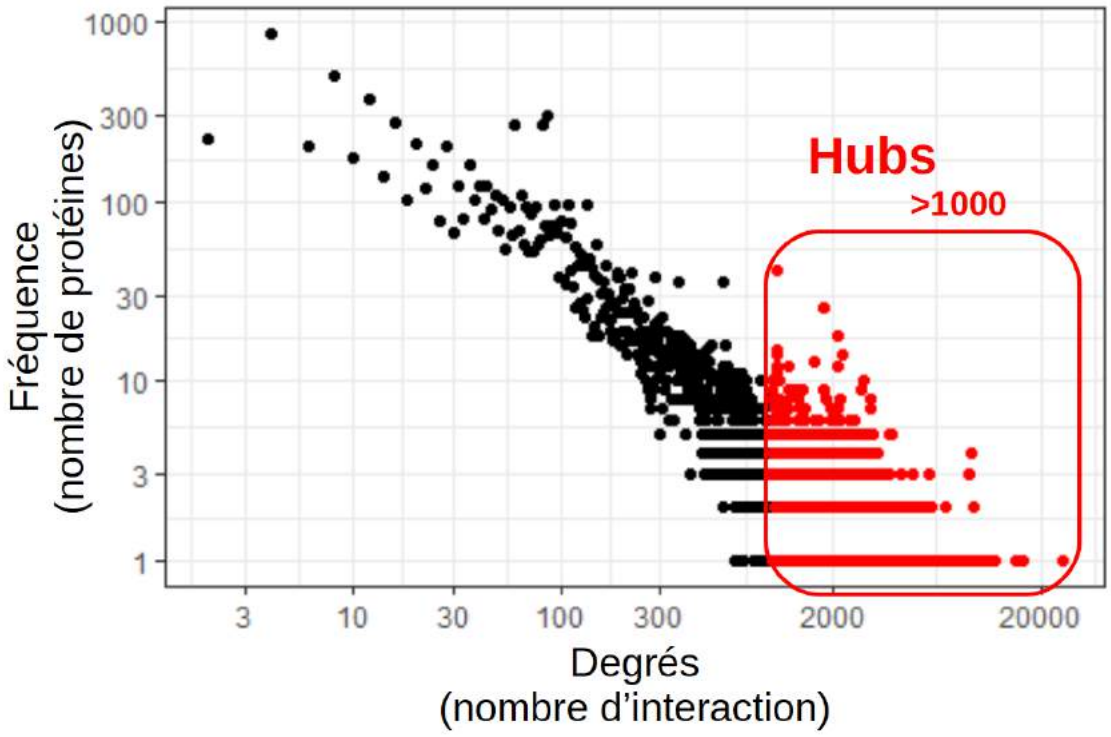

Figure I.1.3 – Distribution des degrés du réseau PPI STRING. La distribution des degrés est représentée par la fréquence du nombre d'interaction par protéine du réseau (les axes suivent une échelle logarithmique). Les hubs sont définis comme les 20 % des protéines ayant le plus grand nombre d'interactions (Yu *et al.*, 2004). La distribution du réseau PPI, ici celle de STRING (Szklarczyk *et al.*, 2019), nous montre que la majeure partie des protéines ont peu d'interactions, la moitié des protéines ont moins de 120 interactions, et une petite partie a beaucoup d'interaction, les hubs (en rouge) ont plus de 1000 interactions.

1.2 Centralités

Une centralité est une mesure capable de quantifier l'importance relative d'un nœud dans un réseau (Kang *et al.*, 2011). Couramment utilisée depuis les années 1970 dans l'étude des réseaux sociaux (Burt, 1976; White *et al.*, 1976; Cook *et al.*, 1983), la centralité est appliquée pour la première fois aux réseaux PPI avec le travail de Jeong en 2001 (Jeong *et al.*, 2001). Il est intéressant de remarquer que Freeman, pionnier de l'analyse des réseaux, qualifie l'évolution de l'utilisation de la centralité comme à contre-courant car il est assez rare qu'une méthode soit transférée des sciences sociales vers les sciences naturelles (Freeman, 2008).

1.2.1 Différents type de centralités

Une centralité permet donc de classer les nœuds selon leur participation à la structure du réseau. Il existe de nombreuses centralités différentes dans la mesure où la notion d'importance d'un nœud dépend de la question que l'on se pose. Certaines centralités,

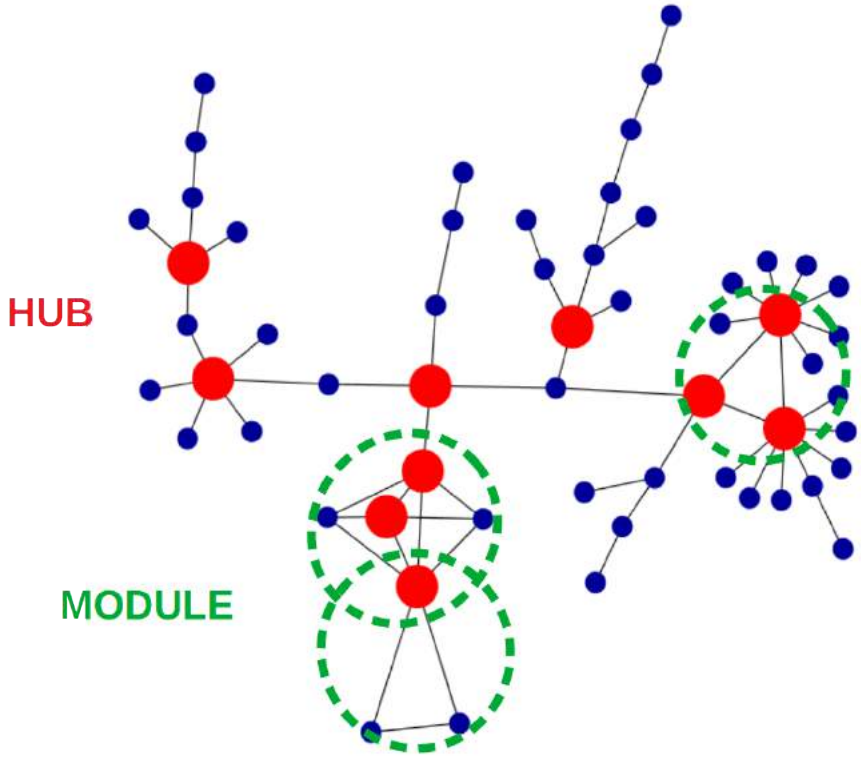

Figure I.1.4 – Hubs et modules d'un réseau modèle. Le modèle Barabasi génère un réseau PPI sans échelle (Barabasi et Albert, 1999) décrivant la plupart des réseaux biologiques. Le réseau jouet représenté ici est généré à partir du package igraph du logiciel R (Csardi et Nepusz, 2006). Les hubs du réseau (en rouge) ont ici plus de 4 interactions. Les modules du réseau sont des groupes de protéines interagissant entre eux.

dites locales, ne prennent en compte que le voisinage direct d'un nœud pour mesurer son importance. Elles incluent la centralité de degré, de sous-graphe, et de vecteur propre. D'autres centralités, dites globales, considèrent tous les nœuds du réseau. Il s'agit des centralités de proximité, d'excentricité et d'intermédiarité.

Centralités locales

Degré

La centralité de degré mesure l'importance d'un nœud selon le nombre de voisins directs (Freeman, 1978). Dans les réseaux PPI, seul un petit nombre de protéines possède une grande centralité de degré, les hubs (Figure I.1.3). L'expérience de Jeong, *Lethality and centrality in protein networks*, utilisait sans la nommer une centralité de degré dans le réseau PPI de la levure pour prouver que les protéines hubs ont un rôle important pour la survie cellulaire. Cette observation est connue depuis sous le nom de la règle de centralité-mortalité (*the centrality-lethality rule*) (He et Zhang, 2006).

Sous-graphe

La centralité de sous-graphe mesure la participation des nœuds aux modules du réseau.

Cette centralité donne plus d'importance aux petits groupes de protéines formant des modules de 3 ou 4 protéines.

$$C_{sg}(x) = \sum_{k=0}^{\infty} \frac{(A)^k}{k!} \tag{1.1}$$

1.1 – Équation de la centralité de sous graphe
Avec $C_{sg}(x)$ la centralité de sous graphe du nœud x
k la longueur chemin
A la matrice d'adjacence
Une boucle est un chemin partant d'une protéine x, passant par k protéines distinctes dans le réseaux et revenant à la protéine x. La centralité de sous-graphe d'une protéine est grande si elle participe à un grand nombre de boucles. Le nombre de boucle est divisé par le factoriel de sa longueur k. Ce qui signifie que plus la boucle est longue moins elle fera augmenter la centralité de sous-graphe. En pratique, la centralité de sous-graphe est surtout basée sur le nombre de triangles (boucle de longueur 3) ou de carrés (boucle de longueur 4).

Inventée en 2005 par Ernesto Estrada et Juan Rodríguez-Velázquez (Estrada et Rodríguez-Velázquez, 2005), la centralité de sous-graphe permet d'identifier plus de protéines essentielles à la survie cellulaire que ne le fait la centralité de degré (Jeong *et al.*, 2001). Ce résultat prouve que le caractère indispensable d'une protéine dans le réseau PPI est une conséquence de son imbrication dans des modules, plus que de ses interactions (Yeger-Lotem *et al.*, 2004).

Vecteur propre

La centralité de vecteur propre classe les nœuds selon l'importance de ses voisins (Bonacich, 1987). Le concept de « prestige » dans les réseaux sociaux illustre bien la centralité de vecteur. Le prestige d'une personne ne se résume pas au nombre de personnes qu'elle connaît ; il tient compte aussi de l'importance des personnes avec qui elle est connectée (Rusinowska *et al.*, 2011).

$$C_{vp}(x) = \frac{1}{\lambda} \sum_{y \in V(x)} C_{vp}(y) \tag{1.2}$$

1.2 – Équation de la centralité de vecteur propre
Avec $C_{vp}(x)$ la centralité de vecteur propre du nœud x
λ la valeur propre, une constante
$V(x)$ l'ensemble des voisins directs du nœud x
La centralité de vecteur propre est définie comme la somme des centralités de ses voisins et se calcule grâce à un algorithme de convergence.

Centralités globales

Les mesures de centralités globales se réfèrent à l'ensemble des nœuds du réseau pour définir l'importance d'un nœud donné. Elles mobilisent la notion de distance séparant deux nœuds. La distance la plus utilisée est celle du plus court chemin, aussi appelée la distance géodésique. L'hypothèse est alors que l'information entre deux nœuds se propage uniquement par les plus courts chemins. La distance de la marche aléatoire (*random walk*) est aussi beaucoup utilisée (Newman, 2005). Dans ce cas, la distance entre deux

protéines est la probalité partant d'une protéine d'atteindre la deuxième suivant des chemins sélectionnés aléatoirement (Ghasemi *et al.*, 2014).

Proximité

La centralité de proximité indique quels nœuds peuvent communiquer rapidement avec les autres nœuds du réseau (Borgatti, 2005).

$$C_c(x) = \sum_{y \neq x} \frac{1}{dist(x,y)} \tag{1.3}$$

1.3 – Équation de la centralité de proximité
Avec $C_c(x)$ la centralité de vecteur propre du nœud x
$dist(x,y)$ la distance entre x et y
La centralité de proximité est définie comme l'inverse de la somme des distances entre le nœud x et les autres nœuds du réseau. Plus le nœud x sera proche de beaucoup de nœuds du réseau plus sa centralité de proximité sera proche. Les nœuds éloignés du nœud x n'auront pas beaucoup d'influence sur sa centralité.

Cette mesure a contribué à définir le centre des réseaux biologiques (Wuchty et Stadler, 2003). En effet, une protéine avec une grande centralité de proximité sera située à une faible distance des autres protéines du réseau. Elle occupera donc une position centrale dans le réseau et aura plus de chance d'être une protéine essentielle (Hahn et Kern, 2005).

Excentricité

La centralité d'excentricité mesure l'accessibilité d'un nœud par les autres nœuds. Sa définition est très proche de celle de la centralité de proximité. Traditionnellement cette centralité est utilisée dans les problèmes d'emplacement d'installation (*facilities location problem*). Un hôpital par exemple doit être localisé de manière à minimiser le trajet maximum en cas d'urgence (Krnc *et al.*, 2018).

$$C_e(x) = \frac{1}{max(dist(x,y))} \tag{1.4}$$

1.4 – Équation de la centralité d'excentricité
Avec $C_e(x)$ la centralité d'excentricité du nœud x
$max(dist(x,y))$ la distance maximale entre x et y, toutes autres protéines du réseau
La centralité d'excentricité est définie par l'inverse de la distance maximale entre x et les autres nœuds du réseau. Plus le nœud x sera proche de tous les nœuds du réseau, plus sa centralité d'excentricité sera grande. Les nœuds les plus proches du nœud x n'ont aucune influence sur la centralité d'excentricité.

Dans les réseaux PPI, les protéines avec une grande centralité d'excentricité sont les protéines les plus proches du centre géométrique du réseau (Jalili *et al.*, 2016). Par cette position centrale, elles sont facilement accessibles à d'autres composants du réseau et peuvent donc percevoir les changements dans le réseau (Pavlopoulos *et al.*, 2011).

Intermédiarité

La centralité d'intermédiarité identifie les nœuds qui sont des intermédiaires indispensables pour la communication entre les nœuds du réseau.

$$C_e(x) = \sum_{(y,z)\neq x} \frac{chemin_{x,y}(x)}{chemin_{x,y}} \tag{1.5}$$

1.5 – Équation de la centralité d'intermédiarité
Avec $C_b(x)$ la centralité d'intermédiarité du nœud x
y, z deux autres nœuds du réseau
$chemin_{y,z}$ sont tous les chemins entre y et z
$chemin_{y,z}(x)$ sont tous les chemins entre y et z passant par x
La centralité d'intermédiarité de x, est définie comme le nombre de chemins passant par x. Pour chaque couple de nœuds, il est nécessaire de définir le chemin les reliant. Si l'on considère la distance géodésique, alors on considérera le ou les chemins les plus courts entre les 2 nœuds. Dans le cas des marches aléatoires, on considérera plusieurs chemins générés aléatoirement entre les 2 nœuds. Un nœud avec une grande centralité d'intermédiarité sera présent dans beaucoup de chemins reliant les autres nœuds du réseau.

Les protéines avec une grande centralité d'intermédiarité sont notamment à l'origine d'une structure de goulot d'étranglement (*bottleneck*) (Yu *et al.*, 2007) (Figure I.1.5). Ces protéines relient plusieurs groupes d'autres protéines qui ne peuvent pas interagir sans elles. Les protéines avec une grande centralité d'intermédiarité apparaissent ainsi comme des protéines essentielles à la survie cellulaire (Joy *et al.*, 2005; Hahn et Kern, 2005).

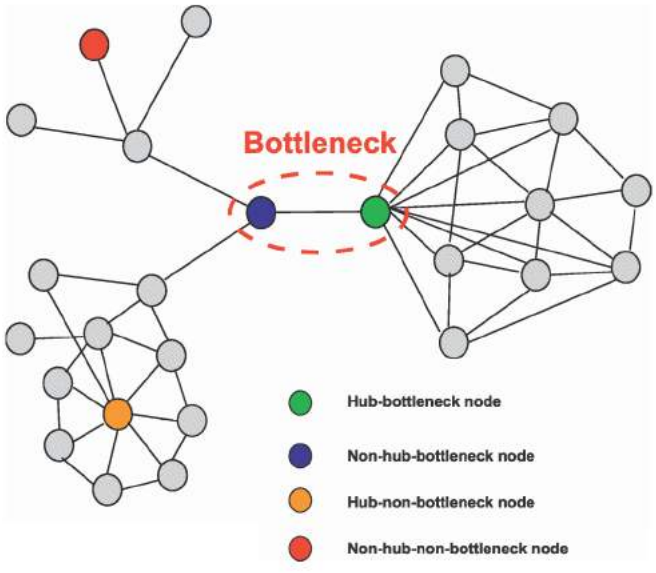

Figure I.1.5 – Illustration de la structure en goulot d'étranglement (Yu *et al.*, 2007). Une protéine qui a une grande centralité d'intermédiarité. Ce sont des protéines essentielles à la survie cellulaire. Cette figure montre qu'une protéine avec une grande centralité d'intermédiarité n'est pas forcément un hub du réseau, et un hub n'a pas nécessairement une grande centralité d'intermédiarité.

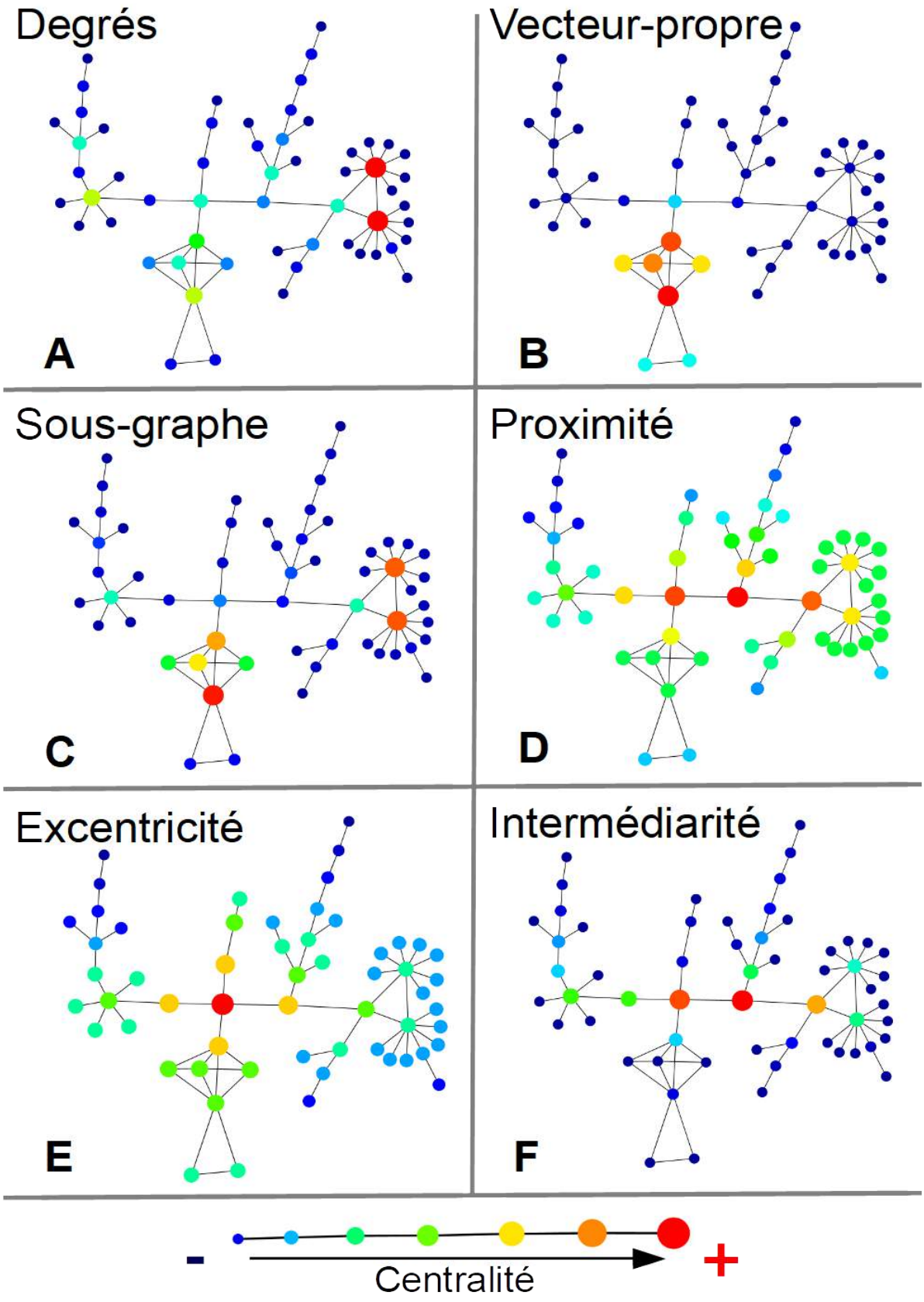

Figure I.1.6 – Calcul des centralités dans un réseau modèle. Une centralité est une mesure de l'importance d'une protéine dans un réseau. J'ai relevé 6 centralités très utilisées dans les réseaux PPI : (A) centralité de degré, (B) centralité de vecteur propre, (C) centralité de sous-graphe, (D) centralité de proximité, (E) centralité d'excentricité et (F) centralité d'intermédiarité. Cette figure montre bien que selon la centralité, les protéines considérées comme importantes, en rouge, ne sont pas toujours les mêmes.

1.2.2 Application des centralités aux réseaux d'interactions protéine-protéine

Toutes les centralités précédemment énoncées peuvent être utilisées seules pour évaluer l'importance relative d'une protéine au sein des réseaux PPI. En effet, toutes ont prouvé leur efficacité puisque les protéines qu'elles placent au centre du réseau sont des protéines essentielles, c'est-à-dire dont la présence est indispensable à la survie de l'organisme (Jeong *et al.*, 2001; Estrada, 2006; Estrada et Rodríguez-Velázquez, 2005; Zotenko *et al.*, 2008). Les centralités peuvent également être utilisées en association pour obtenir de meilleures performances (Jalili *et al.*, 2016). Par exemple, Mistry et collègues combinent l'équation mesurant la centralité de degré avec celle mesurant la centralité de vecteur propre pour générer une nouvelle équation encore plus performante dans l'identification de protéines essentielles (Mistry *et al.*, 2017). Wang *et al.* quant à eux considèrent que les protéines importantes sont celles qui possèdent une centralité élevée quelle que soit la méthode de mesure utilisée (Wang *et al.*, 2014).

Il est donc difficile de préconiser à l'avance l'application d'une centralité plutôt que d'une autre, de manière isolée ou combinée tout dépend du contexte. Premièrement, les performances des différents types de centralité sont dépendantes du réseau étudié. Zotenko *et al.* calculent les centralités (dont celles de degré, de sous-graphe, de vecteur-propre et d'intermédiarité) sur 6 réseaux PPI différents pour identifier les protéines essentielles à la survie cellulaire. Même si elles sont toujours meilleures qu'une sélection aléatoire, les centralités présentent des performances relatives qui diffèrent suivant les réseaux. Deuxièmement, les centralités ont des liens entre elles, mais ces derniers évoluent en fonction du réseau étudié (Figure I.1.7) (Wuchty et Stadler, 2003; Estrada et Rodríguez-Velázquez, 2005; Koschutzki et Schreiber, 2004; Ashtiani *et al.*, 2018). De manière générale, les centralités n'ont pas de corrélation négative entre elles (Estrada et Ross, 2018), ce qui suggère qu'elles ne sont jamais fondamentalement opposées. De plus, des corrélations élevées existent entre la centralité de degré et celle de proximité (Estrada et Rodríguez-Velázquez, 2005), ce qui signifie que les protéines avec une haute centralité de degré sont situées à une faible distance des autres nœuds du réseau. En revanche, des liens peuvent exister dans un réseau donné alors qu'ils n'existent pas dans un autre : par exemple, une forte corrélation entre la centralité de vecteur propre et la centralité de degré a été observée dans le réseau PPI humain (Ashtiani *et al.*, 2018) mais pas dans le réseau PPI de levure (Koschutzki et Schreiber, 2004).

Ainsi, en pratique, lorsqu'on veut analyser un réseau particulier, il est recommandé de tester dans un premier temps l'ensemble des centralités à disposition, seules puis en association, puis de comparer dans un second temps ces méthodes, à l'aide par exemple d'une analyse ACP, afin de choisir la méthodologie la plus appropriée au réseau considéré (Ashtiani *et al.*, 2018).

1.2.3 Propriétés des centralités des protéines pathologiques

La biologie a permis d'identifier un grand nombre de gènes, nommés « gènes pathologiques » qui contribuent à des maladies chez l'homme (Jimenez-Sanchez *et al.*, 2001) et la connaissance des liens qui existent entre ces gènes permet de mieux comprendre les mécanismes pathologiques. L'analyse de ces « gènes pathologiques », avec leurs relations,

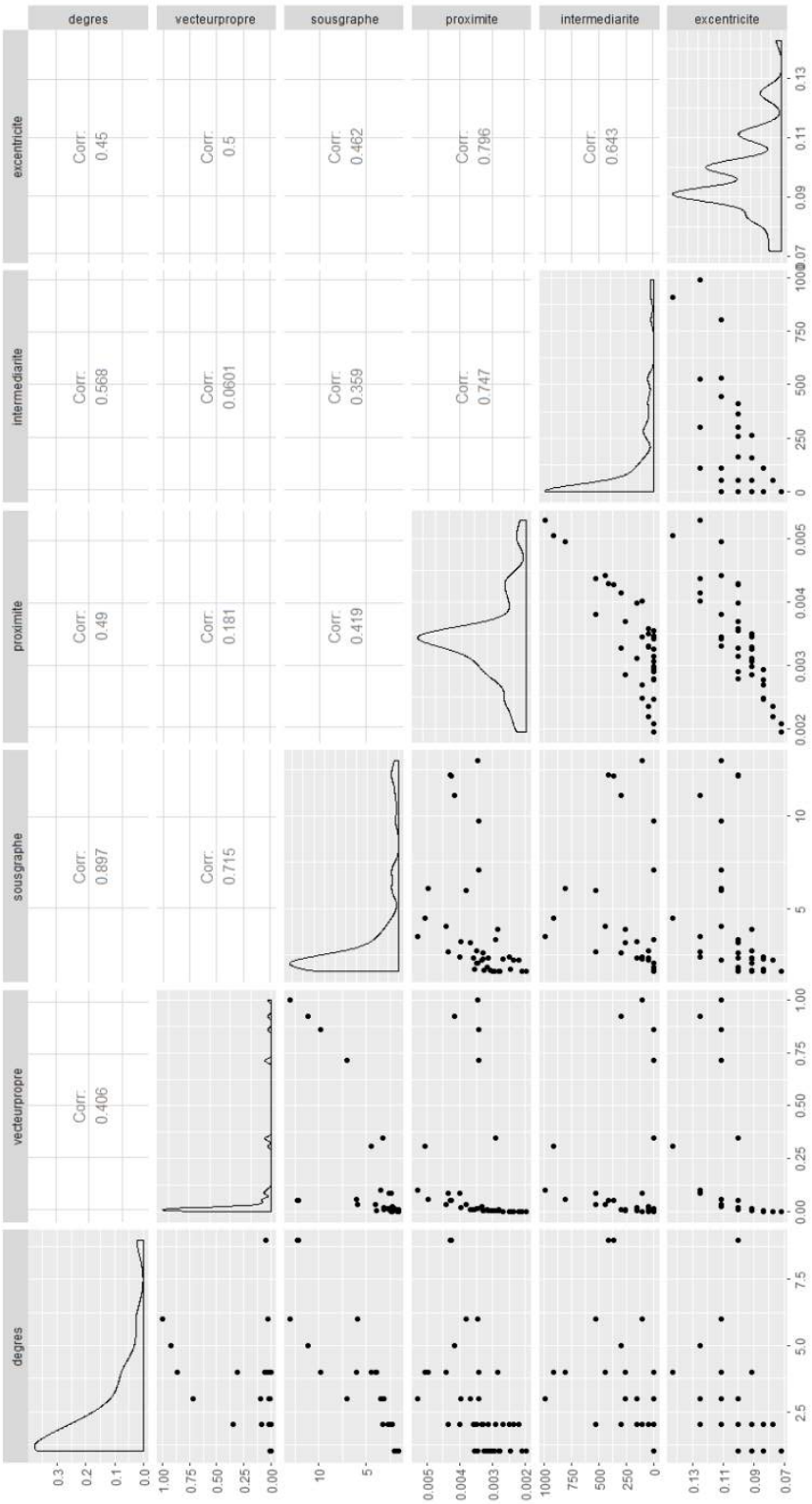

Figure I.1.7 – Corrélations des centralités entre elles dans un réseau modèle. Cette figure est un tableau à double entrée représentant les corrélations entre les 6 centralités présentées précédemment. Cette représentation met en avant les liens qui existent entre les différentes mesures de centralités dans le réseau modèle présenté précédement (Figure I.1.6). La centralité de degré et celle de sous graphe ont la plus grande corrélation : une protéine avec une grande centralité de degré aura aussi une grande centralité de sous graphe, inversement une protéine avec une petite centralité de degré aura une petite centralité de sous graphe. Il est intéressant de remarquer que malgré leur différence aucune des centralités ne sont fondamentalement opposées (elles ne sont jamais corrélées négativement.

peut être élargie à celle des protéines pour lesquelles ils codent (nommées par analogie « protéines pathologiques »), associées à leurs PPI. L'analyse du réseau PPI est alors un formidable outil pour améliorer la compréhension des maladies (Kann, 2007).

L'analyse de la centralité des protéines pathologiques du réseau PPI a permis de mettre en avant 3 propriétés (Figure I.1.8) :

i) Plus les protéines pathologiques ont des centralités élevées, plus elles sont liées à un nombre élevé de maladies.

Quel que soit le réseau, les protéines pathologiques ont toujours plus d'interactions que les autres protéines du réseau (Barrenas *et al.*, 2009; Chavali *et al.*, 2010; Goh *et al.*, 2007). De plus, les protéines impliquées dans beaucoup de maladies occupent des positions centrales dans le réseau, comparée à des protéines liées à une seule maladie, que ce soit en termes de degré, de proximité, d'intermédiarité ou d'excentricité (Chavali *et al.*, 2010).

ii) Plus les protéines pathologiques ont des centralités élevées, plus elles affectent l'organisme dans sa globalité.

En séparant les protéines selon la diversité des phénotypes dans lesquels elles sont impliquées, l'expérience de Chavali *et al.* a en plus prouvé que les protéines pathologiques impliquées dans des maladies très différentes sont plus au centre du réseau que les protéines impliquées dans des maladies semblables. L'explication de ce phénomène vient du fait que les protéines pathologiques de centre du réseau sont exprimées dans plus d'organes différents (Goh *et al.*, 2007). Les protéines pathologiques centrales touchent donc l'organisme de façon plus globale. Les protéines pathologiques ont également tendance à interagir entre elles (Goh *et al.*, 2007; Feldman *et al.*, 2008). Ce phénomène est encore plus fort avec les protéines impliquées dans des maladies possédant des phénotypes semblables (Barrenas *et al.*, 2009). Celles-ci créent des modules distincts fonctionnellement cohérents qui sont exprimés dans les mêmes organes (Goh *et al.*, 2007).

iii) Plus les protéines pathologiques ont des centralités élevées, plus elles sont liées à des maladies sévères.

Barrenas *et al.* ont classé les protéines en fonction du type de maladies auxquelles elles sont liées (Barrenas *et al.*, 2009). Les maladies peuvent en effet être monogéniques. Dans ce cas, la mutation d'un seul gène est suffisante pour générer un phénotype pathologique. Les maladies peuvent être au contraire multigéniques. Cela signifie que des mutations dans plusieurs gènes sont requises pour provoquer la maladie. L'analyse des centralités de proximité et d'excentricité ont montré que les protéines associées aux pathologies monogéniques occupent dans le réseau une position plus centrale que celles des maladies multigéniques. Cela montre que la perturbation d'une protéine centrale mène plus sûrement vers un phénotype pathologique que ne le fait l'altération d'une protéine située en périphérie. Ce résultat est confirmé par le fait que les protéines essentielles à la survie cellulaire sont plus centrales que les protéines non-essentielles (Goh *et al.*, 2007). Il est aussi corroboré par l'observation que les protéines responsables des cancers ont aussi une place

bien particulière dans le réseau PPI, avec bien plus d'interactions et engagées dans plus de modules que les protéines des autres maladies (Jonsson et Bates, 2006; Wang *et al.*, 2011).

En conclusion, l'analyse de la centralité des protéines pathologiques a permis de relier la position occupée par les protéines au sein du réseau PPI avec différents types de maladies. La perturbation d'une protéine centrale aura donc plus d'impact sur le réseau dans son ensemble et mènera vers un état pathologique plus certain.

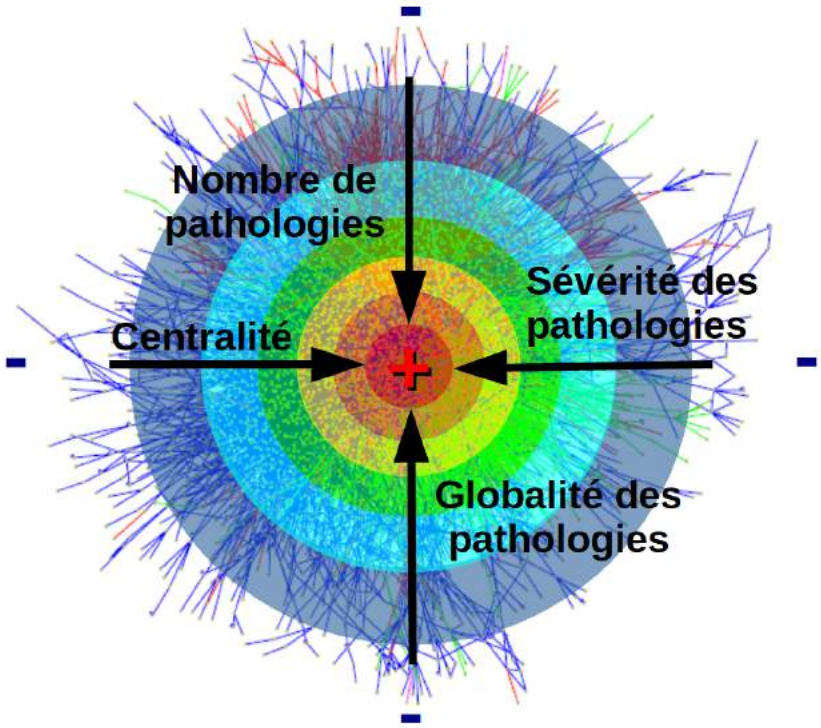

Figure I.1.8 – Illustration de la centralité des protéines pathologiques. La topologie des protéines pathologiques a été étudiée dans plusieurs publications grâce aux mesures de centralités (Barrenas *et al.*, 2009; Chavali *et al.*, 2010; Goh *et al.*, 2007). Il a été remarqué que plus la centralité d'une protéine pathologique est grande, plus cette protéine est liée (i) à un nombre élevé de maladies, (ii) de sévérité importante et (iii) touchant un large panel d'organes.

1.3 Identification des acteurs clés des maladies rénales

Les maladies rénales se manifestent par une perte progressive de la fonction rénale (Romagnani *et al.*, 2017). Les mécanismes biologiques à l'origine de ces maladies et de leur progression ne sont pas toujours bien compris (Cijiang He *et al.*, 2012; Brosius et Ju, 2018; Nicoll *et al.*, 2018). Il est donc important d'identifier de nouveaux acteurs moléculaires des maladies rénales. Dans ce but, 5 types d'approches sont actuellement utilisées : (i) une approche simple basée uniquement sur l'expérience, et 4 approches plus complexes, qui associent les résultats expérimentaux à (ii) l'utilisation d'un logiciel commercial (Ingenuity Pathway Analysis) ou l'analyse de réseaux PPI (iii) focalisés sur les molécules différentiel-

lement exprimées, (iv) spécifiques au tissu rénal ou (v) prenant en compte l'interactome dans sa globalité.

1.3.1 Méthodes basées sur l'expérimentation

Les études expérimentales demeurent de nos jours les approches les plus utilisées pour décortiquer les mécanismes des maladies, que ce soit dans un contexte de maladies rénales ou dans un contexte plus général (Cijiang He *et al.*, 2012). La majorité de ces études compare l'expression de molécules entre une population d'individus malades et un groupe d'individus sains en considérant que les molécules les plus différentiellement exprimées (DE) constituent des acteurs importants de la maladie. Grâce au développement des analyses omiques qui permettent de mesurer simultanément l'abondance de milliers de composés dans un échantillon donné, un grand nombre d'acteurs clés des maladies rénales a ainsi pu être mis en évidence. Ces études expérimentales se limitent cependant souvent à l'étude d'un seul niveau moléculaire (ARNm, protéines, ...). De plus elles identifient les molécules clés uniquement sur la base de leur expression alors qu'une molécule peut avoir un rôle clé dans la pathologie sans modification de son expression, via par exemple ses propriétés de liaison avec d'autres partenaires ou son activité enzymatique. Enfin, certains acteurs clés peuvent être absents de l'échantillon dans lequel les mesures sont réalisées (par exemple certaines protéines du tissu rénal ne seront jamais excrétées dans l'urine).

1.3.2 Méthodes utilisant Ingenuity Pathway Analysis (IPA)

Le logiciel commercial *Ingenuity Pathway Analysis* (IPA) est souvent utilisé pour analyser les données omiques issues de plusieurs niveaux moléculaires. IPA propose par exemple l'analyse *Canonical Pathway* qui permet de mettre en avant les voies métaboliques auxquelles appartiennent les molécules différentiellement exprimées. Cette approche descriptive identifie donc des fonctions pathologiques importantes et les molécules clés qui leur sont associées. Largement utilisée dans l'étude génomique du tissu rénal (Parikh *et al.*, 2015) et dans l'étude protéomique de l'urine (Hogan *et al.*, 2014; Davalieva *et al.*, 2015), cette méthode est toutefois restreinte aux molécules présentes dans l'échantillon analysé et dont l'expression varie au cours de la pathologie.

IPA propose également un deuxième type d'analyse (*Causal analysis approaches* (Krämer *et al.*, 2013)) dont l'objectif est de comprendre quelles sont les origines biologiques des effets observés expérimentalement. L'algorithme principal, *Upstream regulator analysis*, est particulièrement intéressant dans la recherche d'acteurs clés des maladies. En effet il permet de prédire les molécules qui peuvent potentiellement expliquer la perturbation des molécules DE en cherchant les partenaires les plus directs de ces molécules DE (Figure I.1.9). La force de cet algorithme réside donc dans sa capacité à mettre en avant des composés non-DE ou encore non mesurés dans l'échantillon. C'est grâce à ce type d'analyse que Nair *et al.* ont identifié IL1β comme une protéine clé de l'inflammation dans la néphropathie diabétique alors même que son expression n'était pas modifiée dans le tissu rénal des patients (Nair *et al.*, 2018). A part dans le travail de Nair, cette méthodologie a été cependant peu utilisée pour la découverte de nouvelles molécules clés dans le contexte des maladies rénales. De plus, le fonctionnement du logiciel IPA repose

sur une large base de données, *Ingenuity Knowledge Base*, qui intègre les relations entre les gènes, les protéines et les métabolites, mais dont l'accessibilité n'est pas ouverte.

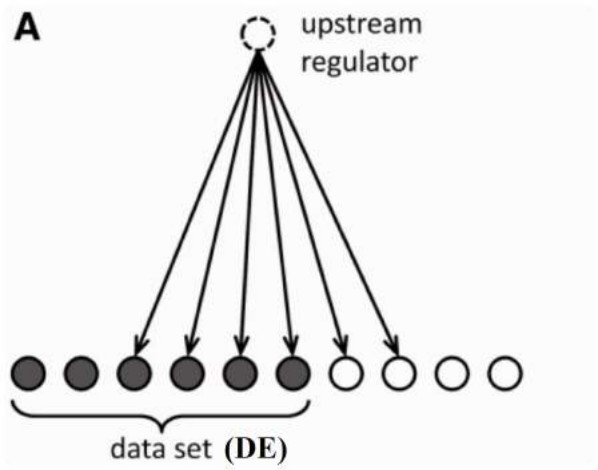

Figure I.1.9 – llustration de l'algorythme *Upstream Regulator* de IPA (Krämer *et al.*, 2013). Le réseau *Ingenuity Knowledge Base* intègre les relations entre les gènes, les protéines et les métabolites. Une molécule est identifiée comme un régulateur de la maladie si elle est liée à beaucoup de molécules différentiellement exprimées (en gris).

1.3.3 Méthodes basées sur les réseaux des gènes différentiellement exprimés

Des travaux récents (Tableau I.1.4) dans le domaine de la néphrologie exploitent les réseaux PPI et les centralités pour hiérarchiser les gènes différentiellement exprimés (DE) identifiés par des expériences de transcriptomique. Ces différents travaux suivent une démarche commune avec (i) la construction du réseau PPI des gènes DE, (ii) l'identification des hubs et des modules du réseau, (iii) l'analyse des fonctions biologiques des gènes pathologiques nouvellement identifiés.

Publication	Maladie	Base de données	Nbre gènes DE	Enrichissement	Centralités	Nbre de hubs	Modules
(Abedi et Gheisari, 2015)	Néphropathie diabétique	STRING	49	+	degré, proximité, intermédiarité	34	-
(Fu *et al.*, 2015)	Néphropathie diabétique	STRING	416	-	degré	12	+
(Ma *et al.*, 2017)	Néphropathie diabétique	STRING	426	-	degré	5	+
(Jia *et al.*, 2018)	BK virus en transplantation rénale	STRING	524	-	degré, intermédiarité, sous-graphe	22	-
(Rabieian *et al.*, 2017)	Maladies rénales chroniques	STRING	280	+	degré, proximité, intermédiarité	122	-
(Zhou *et al.*, 2018)	Maladies rénales chroniques	BioGRID, DIP, HRPD, STRING	226	-	degré	10	+
(Jin *et al.*, 2018)	Carcinome à cellule claire	STRING	1799	-	degré	194	+
(Wei *et al.*, 2019)	Carcinome à cellule claire	STRING	472	-	degré, sous-graphe	34	-

Tableau I.1.4 – Méthodes basées sur les réseaux des gènes différentiellement exprimés

i) Construction d'un réseau PPI des gènes DE

En utilisant une base de données PPI (Section I.1.1.3), la plus utilisée étant STRING (Szklarczyk *et al.*, 2019), le réseau des gènes DE est construit. Dans certains cas, en particulier lorsque le nombre de molécules DE est faible, les réseaux sont très peu connectés. Il est alors possible d'ajouter d'autres composés au réseau afin de l'enrichir. Prenons par exemple l'étude d'Abedi *et al.* (Abedi et Gheisari, 2015) qui ré-analyse des données d'expression génique dans le rein de patients atteints de néphropathie diabétique. Le réseau initial, formé de 49 gènes DE, ne comptait que 5 interactions (Figure I.1.10 A). Pour avoir un réseau plus fourni, Abedi *et al.* y ont ajouté 88 gènes qui correspondent aux voisins des gènes DE dans le réseau PPI STRING (Figure I.1.10 B). Le réseau ainsi enrichi comprenait 137 gènes fortement liés les uns aux autres. Tous, qu'ils soient DE ou non, étaient bien liés à la néphropathie diabétique d'après la littérature.

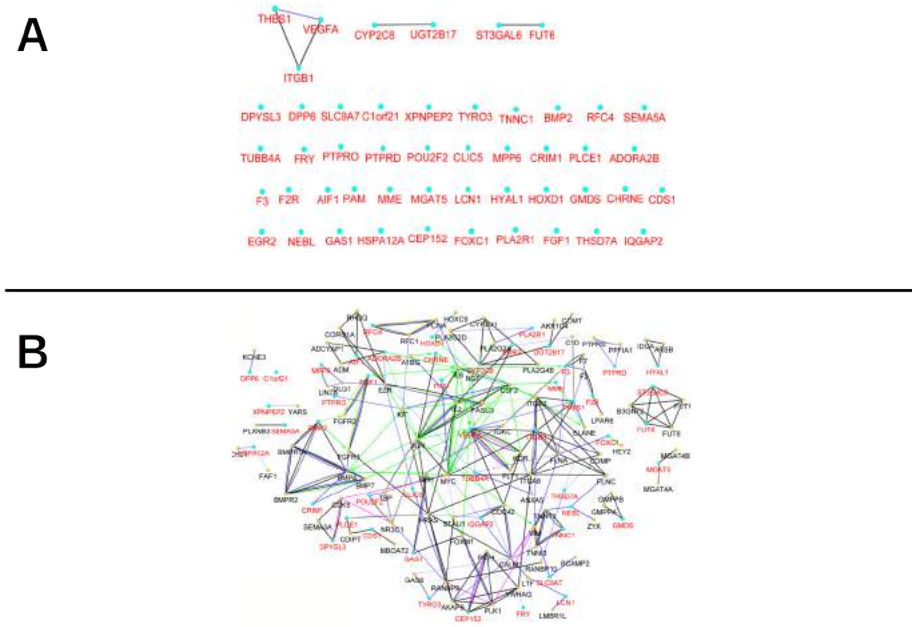

Figure I.1.10 – Enrichissement du réseau des gènes DE (Abedi et Gheisari, 2015). Abedi *et al.* ré-analysent l'expression de gènes de tissu rénal de patients atteints de néphropathie diabétique. Ils ont identifié 49 gènes DE (en rouge) liés par 5 interactions (A). Ils ont enrichi le réseau en sélectionnant les gènes directement adjacents (en noir) sur le réseau STRING (avec un maximum de deux par gènes DE). Ils ont obtenu un réseau de 137 gènes liés par un nombre important d'interactions (B).

ii) Identification des hubs et des modules

L'identification des gènes DE permet de prédire des gènes pathologiques sur la base de leur expression différentielle. L'analyse des hubs et modules dans le réseau PPI permet en plus de mettre en avant les molécules DE qui exercent un rôle particulièrement important dans la pathologie indépendamment de la variation de leur expression.

Les hubs jouent en effet un rôle clé dans les réseaux PPI (Section I.1.1.5). De ce fait, les hubs des réseaux des molécules DE devraient exercer une fonction importante dans les mécanismes pathologiques. Ma *et al.* par exemple ont ré-analysé les résultats d'une étude identifiant les gènes DE de la néphropathie diabétique (Ma *et al.*, 2017). En mesurant la

centralité de degré dans le réseau des gènes DE, ils ont identifié VEGFA, ACTN4, FYN, COL1A2 et IGF1 comme étant des hubs. Leur analyse suggère donc que ces 5 gènes sont des acteurs clés de la néphropathie diabétique, ce que les auteurs ont ensuite confirmé grâce à une étude bibliographique.

Les molécules appartenant à des modules sont également considérées comme des protéines importantes du réseau (Section I.1.1.5). Les gènes DE insérés dans des modules devraient donc être des acteurs essentiels des maladies. Puisque les modules sont composés de protéines fonctionnellement très liées, leur identification permet d'accéder aux mécanismes biologiques associés à la maladie de manière plus complète que ne le permet l'analyse des hubs (Ideker et Sharan, 2008). C'est par exemple via cette approche associée à l'implémentation de MCODE de Cytoscape (Bader et Hogue, 2003) que Fu *et al.* ont identifié 2 modules impliqués dans la prolifération cellulaire et jouant un rôle important de la néphropathie diabétique (Fu *et al.*, 2015).

iii) Analyse des fonctions biologiques des nouveaux gènes pathologiques

La cohérence biologique des gènes pathologiques prédits est ensuite évaluée grâce à des sources d'annotations qui permettent de faire le lien entre un gène et ses fonctions biologiques. La *Gene Ontology* (GO) est utilisée pour annoter les gènes selon 3 catégories : fonction moléculaire, processus biologique et compartiment cellulaire. La *Kyoto Encyclopedia of Gene and Genomes* (KEGG) répertorie quant à elle des gènes et les voies métaboliques (*pathway*) qui leur sont associées. L'annotation des fonctions liées aux gènes pathologiques met en évidence un enrichissement pour certaines voies métaboliques qui constituent de potentiels nouveaux processus contribuant aux maladies. Par exemple, l'étude de Ma *et al.* a identifié un module enrichi en gènes liés au système immunitaire, ce dernier ayant été associé à la néphropathie diabétique par des études antérieures (Ma *et al.*, 2017). De manière intéressante, les résultats d'Abedi *et al.* et de Rabieian *et al.* montrent qu'il est plus informatif d'étudier les voies associées uniquement aux gènes centraux (hubs / modules) plutôt que celles liées à l'ensemble des gènes DE (Abedi et Gheisari, 2015; Rabieian *et al.*, 2017). Par exemple, dans le travail de Rabieian *et al.*, l'analyse des gènes centraux conduit à l'identification d'un ensemble de 78 *pathways* en accord les uns avec les autres alors que la prise en compte additionnelle des gènes DE réduit le nombre de voies mises en avant (34), celles-ci étant de surcroît seulement faiblement liées sur le plan fonctionnel.

L'analyse du réseau PPI basée sur les molécules DE, telle que nous l'avons décrite, est une méthode à fort potentiel ; elle n'a toutefois pas encore porté ses fruits en recherche médicale pour la compréhension des maladies, notamment rénales. Premièrement, l'approche ne s'est développée que récemment (2015-2019) grâce à la disponibilité des données, rendue possible par STRING et GEO par exemple, et celle des outils notamment Cytoscape. Son utilisation est à ce jour encore très limitée aux (bio)informaticiens si bien que la validation expérimentale des nouveaux gènes et processus pathologiques que la méthode a prédit n'a, mise à part une exception (Chen *et al.*, 2018), jamais été réalisée. Deuxièmement, dans le domaine des maladies rénales, la méthodologie n'a exploité que des données génomiques. Or il serait judicieux de l'étendre aux données issues de la protéomique, comme cela a été fait pour d'autres contextes pathologiques, dans la mesure où les protéines sont plus

proches du phénotype que ne sont les gènes. Chen *et al.* par exemple ont identifié deux protéines cibles (SLC2A4 et TUBB2C) pour le traitement du cancer de la prostate (Chen *et al.*, 2016) grâce au calcul des centralités de proximité et d'intermédiarité sur un réseau PPI construit à partir des protéines DE dans le tissu prostatique cancéreux. Enfin, même si les hubs et modules constituent une liste intéressante de molécules pathologiques servant de point de départ pour des études complémentaires, l'approche limite la possibilité de détecter de nouveaux acteurs car elle se restreint aux molécules dont l'expression est mesurable dans l'échantillon étudié, ou au mieux à leurs interacteurs directs dans le réseau si celui-ci a été enrichi.

1.3.4 Méthodes basées sur des réseaux spécifiques au tissu rénal

L'interactome, dans son intégralité, est constitué d'interactions qui ne s'appliquent pas forcément dans tous les contextes (Edwards *et al.*, 2002; Ideker et Krogan, 2012). Il peut donc être judicieux de contextualiser le réseau en fonction de la maladie d'intérêt. Cela passe par la constitution de réseaux PPI spécifiques dont toutes les protéines sont sélectionnées car exerçant une fonction biologique dans un tissu ou un type cellulaire particulier. Trois réseaux PPI spécifiques au rein ont ainsi été générés : GlomNet, PodNet et GCNet.

GlomNet

Le glomérule rénal est la structure cellulaire du néphron permettant la filtration du plasma et la formation de l'urine primitive. L'objectif de GlomNet est d'identifier des acteurs clés dans la biologie du glomérule pour ensuite décrire les mécanismes clés des maladies glomérulaires et par extension des maladies rénales puisque la majorité d'entre elles sont dues à un dysfonctionnement du glomérule.

Le réseau GlomNet compile les gènes considérés par 5 études (réalisées chez l'homme ou la souris) comme étant enrichis dans le glomérule (He *et al.*, 2008). Ces gènes ont été étiquetés tels quels car leur expression dans le glomérule rénal d'individus/animaux sains est différente de celle dans les autres parties du rein. Les PPI liant ces gènes ont été sélectionnés via la base de données HPRD (spécifique à l'homme) (Keshava Prasad *et al.*, 2009). GlomNet contient ainsi 1407 gènes au total mais seulement 543 d'entre eux sont en interaction. He *et al.* ont calculé qu'environ 11 % des gènes présents dans GlomNet sont différentiellement exprimés dans la néphropathie diabétique (maladie d'origine essentiellement glomérulaire), cette proportion étant significativement plus grande que dans l'interactome global (1,8 %) (He *et al.*, 2008). Ils remarquent aussi que les gènes impliqués dans la néphropathie diabétique ont tendance à interagir directement entre eux et que la néphrine, dont le rôle dans les maladies rénales a déjà été prouvé, est située en position centrale dans le réseau. Cela leur permet de conclure que GlomNet est un outil pertinent pour la recherche des mécanismes des pathologies glomérulaires/rénales.

Le réseau PPI GlomNet est désormais vu comme une référence dans l'étude des glomérules (Lindenmeyer *et al.*, 2010). Tomaszewski et al. par exemple utilisent GlomNet pour suggérer que l'association qu'ils mesurent entre l'expression rénale de FGF-1 et l'hypertension pouvait prendre son origine dans les glomérules (Tomaszewski *et al.*,

2015). Perisic *et al.* combinent les interactions qu'ils observent expérimentalement avec d'autres que GlomNet décrit pour lister les partenaires fonctionnels d'une protéine d'intérêt (Plekhh2) (Perisic *et al.*, 2012). Malheureusement, le réseau PPI GlomNet n'a jamais fait l'objet d'une analyse des centralités, ce qui aurait été pourtant intéressant pour comprendre les pathologies glomérulaires dans leur globalité.

PodNet et XPodNet

Les podocytes sont les cellules formant la barrière de filtration du glomérule. PodNet et XPodNet sont des réseaux PPI spécifiques aux podocytes chez la souris (Warsow *et al.*, 2013).

Warsow *et al.* ont construit le réseau PodNet en examinant manuellement la littérature traitant des données d'expression des podocytes. La nature exacte des protéines insérées dans le réseau (exprimées ou enrichies dans les podocytes ?) est toutefois assez floue car les détails de leurs recherches ne sont pas accessibles. Comme très peu de molécules étaient connectées entre elles, Warsow *et al.* ont ensuite enrichi le réseau (XPodNet) en y ajoutant des relations et des protéines issues de STRING et capables de relier les protéines déjà présentes. PodNet contient ainsi 315 protéines et 223 interactions alors que le XPodNet possède 839 protéines à l'origine de 1048 interactions.

Dans l'étude de Warsow *et al.*, le transcriptome podocytaire de la souris adulte est comparé in vitro avec celui de la souris embryonnaire et le réseau XPodNet est utilisé pour mettre en avant les interactions (et non les protéines) qui différencient l'adulte de l'embryon (Warsow *et al.*, 2010). Sasaki *et al.* quant à eux ont d'abord cherché expérimentalement des gènes potentiellement impliqués dans la glomérulonéphrite chez la souris ; ils se sont ensuite servis de XPodNet pour confirmer l'intérêt porté aux candidats et identifier leurs interacteurs (Sasaki *et al.*, 2014). Quelques autres travaux exploitent de manière très limitée les réseaux PodNet et XPodNet (Grgic *et al.*, 2014; Rinschen *et al.*, 2018) mais là encore, comme dans le cas GlomNet, sans calcul des centralités pour quantifier l'importance relative des protéines du réseau.

GCNet

Le cytosquelette est responsable de la structure spatiale des cellules, celle-ci étant essentielle dans le mécanisme de filtration glomérulaire. Ding *et al.* ont proposé un réseau PPI spécifique, *Glomerular Cytoskeleton Network* (GCNet), qui concerne les protéines du cytosquelette des cellules glomérulaires (Ding *et al.*, 2016).

La sélection des gènes exprimés dans le glomérule repose sur 9 études comparant l'expression des gènes dans le glomérule et dans les autres parties du rein, comme cela avait été fait pour GlomNet (He *et al.*, 2008). 2929 gènes ont ainsi été trouvés enrichis dans le glomérule. La sélection des gènes liés au cytosquelette a été faite avec Gene Ontology. 2030 gènes ont ainsi été annotés au cytosquelette. GCNet est l'intersection de ces deux types de données, soit 426 gènes. Les PPI les reliant ont été téléchargées via STRING et HRPD.

Ding et *al.* ont ensuite comparé les gènes inclus dans le réseau avec les gènes associés à 5 maladies rénales provoquant une fuite urinaire de protéines (marqueur de maladies

rénales) : la hyalinose segmentaire et focale, la glomérulonéphrite extra-membraneuse, le syndrome néphrotique idiopathique, la néphropathie diabétique et la néphropathie à IgA. Ils remarquent que GCNet est enrichi en gènes DE dans ces maladies, ce qu'ils considèrent comme une validation du réseau. Ils calculent également la centralité de degré et la centralité d'intermédiarité du réseau mais ils l'utilisent dans un but de visualisation (Figure I.1.11) et non pour prioriser les nouveaux gènes pathologiques.

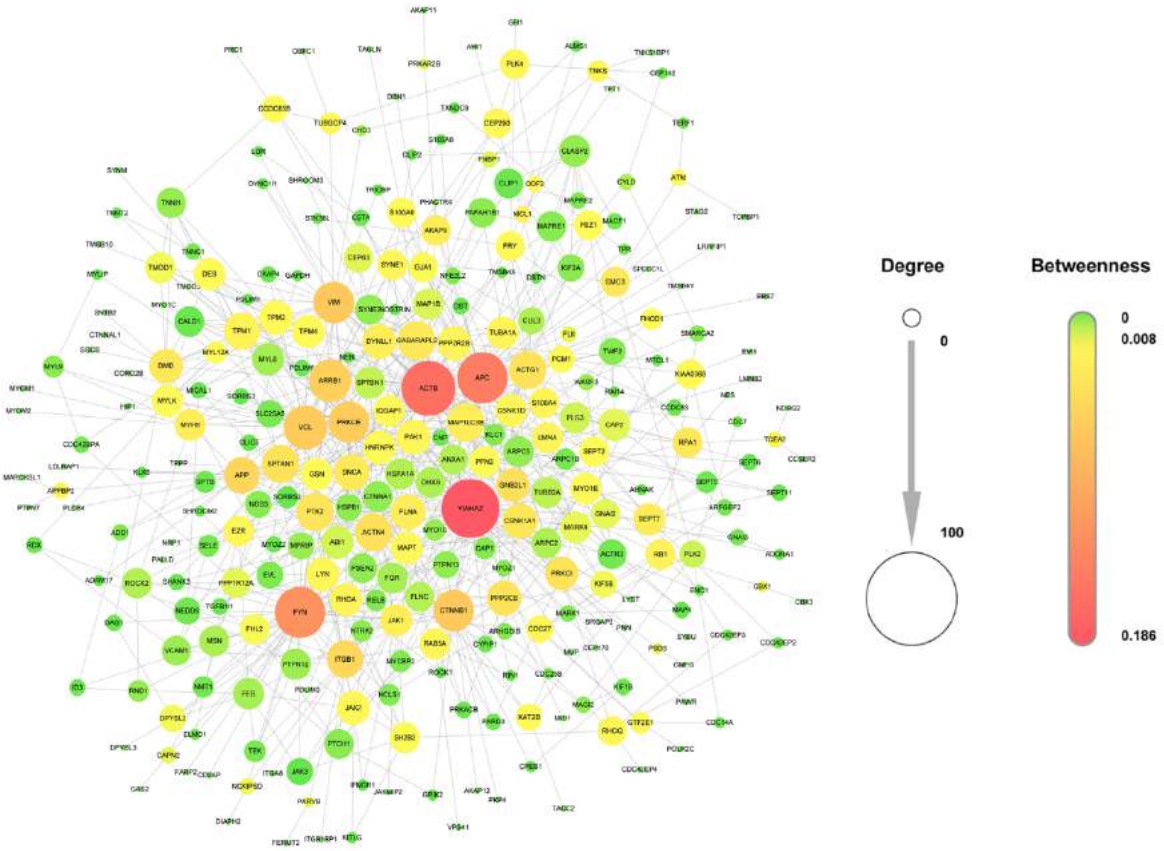

Figure I.1.11 – Réseau *Glomerular Cytoskeleton Network* (GCNet) (Ding *et al.*, 2016). Il est constitué de protéines ayant une fonction particulière dans le cytosquelette des podocytes. La centralité de degré est représentée par la taille des protéines et la centralité d'intermédiarité est représentée par la couleur. Les centralités ont été calculées dans l'objectif de mieux visualiser le réseau.

Ainsi, les réseaux spécifiques sont contextualisés pour des maladies d'intérêt ; il est donc très probable que les gènes pathologiques qu'ils permettront de proposer soient effectivement des acteurs clés des pathologies. De plus, ils ne restreignent pas aux seules molécules DE. Pour ces raisons, les réseaux spécialisés sont plus souvent cités et utilisés par des travaux expérimentaux que les réseaux de gènes DE et leur développement devrait s'intensifier grâce aux outils disponibles sur le web tels que TissueNet [3] (Basha *et al.*, 2017) et *Integrated Interactions Database* [4] (Kotlyar *et al.*, 2019) qui permettent de créer un PPI spécifique à un tissu à partir d'une liste restreinte de molécules. L'analyse des réseaux spécifiques a déjà prouvé son intérêt pour la découverte des nouveaux acteurs clés des pathologies, en particulier dans le domaine du cancer. En utilisant la notion de centralité pour décrire un réseau PPI spécifique au cancer suivie d'une validation expérimentale,

3. netbio.bgu.ac.il/tissuenet/
4. iid.ophid.utoronto.ca/

Kar *et al.* ont par exemple démontré le rôle important de 2 nouvelles protéines (ERBB3 et RAF1) dans les processus cancéreux (Kar *et al.*, 2009). En revanche, comme indiqué précédemment, les trois réseaux PPI spécifiques au rein n'ont qu'une portée limitée pour comprendre les maladies rénales dans la mesure où ils sont peu étudiés du point de vue de leur centralité.

1.3.5 Méthodes basées sur l'utilisation du réseau d'interactions protéine-protéine global

Pour finir, la dernière approche actuellement utilisée pour identifier de nouveaux acteurs moléculaires des maladies rénales se base sur l'exploration de l'interactome dans son intégralité. Il s'agit dans ce cas de repérer des gènes prometteurs dans le réseau PPI global puis de les classer en fonction de leur potentiel. Les nouveaux candidats ainsi hiérarchisés pourront alors servir de point de départ pour des expériences futures (Bromberg, 2013). Cette stratégie relève d'une problématique de priorisation (Tranchevent *et al.*, 2011).

Méthodes de priorisation

Pour prédire de nouveaux gènes pathologiques, les méthodes de priorisation utilisent soit les gènes pathologiques déjà connus soit les gènes DE dans le réseau PPI. L'hypothèse dite *guilt by association* (coupable par association) en est le point de départ (Oliver, 2000; Uetz *et al.*, 2000) et les centralités occupent une place importante dans la méthodologie. Il existe principalement 4 types de méthodes de priorisation qui diffèrent les unes des autres par le type de centralité qu'elles utilisent : degré, vecteur propre, proximité et intermédiarité (Tableau I.1.5).

Centralité	Publication	Données de départ	Reseau PPI	Distance	Validation
degré	(Oti, 2006)	Gènes pathologiques	HPRD	-	Validation croisée
vecteur propre	(Zhao *et al.*, 2011)	Gènes DE	STRING	-	Gènes pathologiques
vecteur propre	(Zhu *et al.*, 2012)	Gènes pathologiques	HPRD	-	Validation croisée
vecteur propre	(Vanunu et Sharan, 2010)	Gènes pathologiques	HPRD	-	Validation croisée
proximité	(Köhler *et al.*, 2008)	Gènes pathologiques	STRING	marche aléatoire	Validation croisée
proximité	(Hsu *et al.*, 2011)	Gènes pathologiques	STRING, DIP, BOND, IntAct, MINT, MIPS, HPRD BioGRID	plus court chemin	Validation croisée
proximité	(Erten *et al.*, 2011)	Gènes pathologiques	HPRD, BioGrid, et BIND	marche aléatoire	Validation croisée
proximité	(Li *et al.*, 2014)	Gènes pathologiques	HPRD	plus court chemin	Validation croisée
intermédiarité	(Dezső *et al.*, 2009)	Gènes DE	MetaCore	plus court chemin	Gènes pathologiques
intermediarité	(Simões *et al.*, 2015)	Gènes pathologiques et DE	HPRD, Inact, Mint	plus court chemin	Gènes pathologiques
degré, proximité	(Ren *et al.*, 2019)	Protéines pathologiques et DE	HINT	marche aléatoire	Validation croisée

Tableau I.1.5 – Méthodes de priorisation basées sur le réseau PPI global

Méthodes basées sur la centralité de degrés

Ces méthodes sont basées sur le voisinage direct et reposent sur l'hypothèse qu'un bon gène candidat est un gène qui interagit directement avec beaucoup de gènes pathologiques (Barabási *et al.*, 2011; Oti, 2006) (Figure I.1.12).

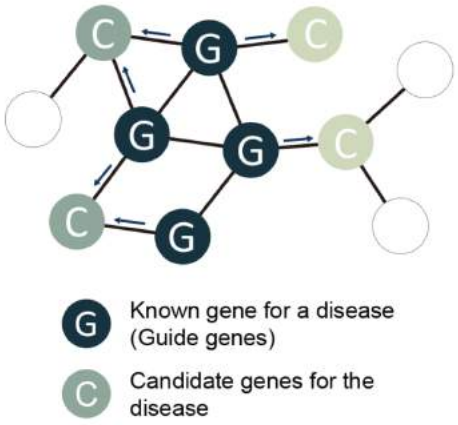

Figure I.1.12 – Illustration de l'hypothèse *guil by association*. Les gènes interagissant directement avec un nombre élevé de gènes pathologiques connus dans le réseau PPI ont plus de chances d'être eux-mêmes des gènes pathologiques. Les gènes candidats notés en gris foncé, liés à deux gènes pathologiques connus, sont donc de meilleurs candidats que les gènes candidats notés en gris pâle qui, eux, n'interagissent qu'avec un seul gène pathologique (Hwang *et al.*, 2019)

Méthodes basées sur la centralité de vecteur propre

Le postulat sous-jacent est le suivant : un bon gène candidat est un gène fortement DE ou un gène proche des gènes fortement DE dans le réseau PPI (Zhao *et al.*, 2011). Dans ce cas donc, ce sont les gènes DE qui sont utilisés comme données de départ à la place des gènes pathologiques déjà connus.

Méthodes basées sur la centralité de proximité

L'idée ici est que plus un gène est proche d'un grand nombre de gènes pathologiques, plus il a de chances d'être lui-même un gène pathologique (Lan *et al.*, 2015).

Pionniers dans le domaine, Köhler *et al.* ont appliqué par exemple une méthode de priorisation basée sur les marches aléatoires pour mettre en avant de nouveaux gènes pathologiques impliqués dans différentes maladies dont la néphronophtise, une maladie rénale évoluant rapidement vers l'insuffisance rénale terminale (Köhler *et al.*, 2008). Ils remarquent également que cette approche de marche aléatoire est meilleure pour prédire de nouveaux gènes pathologiques, au moins dans un contexte de cancer, que deux autres méthodes de priorisation testées, l'une basée sur la centralité de proximité utilisant les plus courts chemin et l'autre basée sur la centralité de degrés (Köhler *et al.*, 2008). D'autres méthodes ont par la suite utilisé le même principe. L'algorithme SPGOranker par exemple a apporté des améliorations en intégrant une source d'information supplémentaire, Gene Ontology (Li *et al.*, 2014). Dans ce cas, un gène est prédit comme un

nouveau gène pathologique s'il est à la fois proche des gènes pathologiques connus et annoté de façon similaire. Erten *et al.* quant à eux proposent de prendre en compte le nombre d'interactions pour pondérer la centralité (Erten *et al.*, 2011) car les méthodes de priorisation favorisent les candidats ayant beaucoup d'interactions (Erten *et al.*, 2011). Grâce à la méthode *Degree-Aware Disease Gene Prioritization* (DADA) qui pénalise les gènes ayant un grand nombre d'interactions, les auteurs montrent qu'ils corrigent ce biais, ce qui leur permet d'identifier des gènes pathologiques ayant peu d'interactions.

Méthodes basées sur la centralité d'intermédiarité

Ces méthodes s'appuient sur le principe de parcimonie (Barabási *et al.*, 2011) qui stipule que les chemins les plus courts entre des gènes pathologiques ou des gènes DE dans le réseau PPI correspondent souvent aux pathways pathologiques. Les gènes situés sur ces plus courts chemins sont donc potentiellement eux aussi des gènes pathologiques.

Dezső *et al.* proposent de calculer un score pour chaque gène qui tient compte du nombre de plus courts chemins entre 2 gènes DE dans lesquels il s'insère (Dezső *et al.*, 2009). Plus le score d'un gène est haut, plus sa probabilité d'être un gène pathologique est grande. Les auteurs intègrent aussi un test de significativité de ce score, en le comparant avec la centralité d'intermédiarité dudit gène. En pénalisant les gènes ayant une grande centralité d'intermédiarité dans le réseau, ils cherchent ainsi à mettre en avant des gènes pathologiques spécifiques à la maladie étudiée et non des gènes pathologiques jouant un rôle à plus grande échelle. Ce type d'approche peut se révéler intéressante pour identifier des cibles thérapeutiques dont la modulation génèrera peu d'effets secondaires.

Validation des méthodes de priorisation

Les 4 méthodes énoncées précédemment ont toutes fait l'objet d'une validation : les gènes qu'elles prédisent comme étant des gènes pathologiques sont effectivement bien des gènes pathologiques. Deux procédures de validation ont été utilisées en fonction du type de données utilisées. Toutes les deux recourent à des bases de données comme OMIM (Hamosh, 2004) qui répertorient un grand nombre de gènes pathologiques afférent à un grand nombre de maladies.

Validation croisée

La validation croisée concernait les méthodes utilisant les gènes pathologiques connus comme données de départ. Dans ce cas, l'ensemble des gènes pathologiques connus pour une maladie a été divisé en 2 sets : un set resté caché, et l'autre set utilisé par l'algorithme de priorisation. La validation a été considérée comme acquise si l'algorithme est parvenu à prédire les gènes pathologiques cachés. L'opération a été répétée plusieurs fois pour évaluer la performance de la méthode. C'est de cette manière que Oti *et al.*(Oti, 2006) ont validé leur approche de priorisation (basée sur la centralité de degré) avec le PPI de 4 espèces (Homme, mouche, vers et levure) et les gènes pathologiques connus dans 383 maladies. Une variante de la validation croisée consiste à utiliser une procédure de *leave-one-out* (Vanunu et Sharan, 2010). Dans ce cas, l'algorithme de priorisation a été appliqué à l'ensemble des gènes pathologiques connus, sauf un, et la prédiction était

bonne si la méthode réussissait à prédire comme pathologique le gène mis de côté. Cette procédure a été reproduite pour chaque gène connu afin d'évaluer la performance de la méthode de priorisation.

Validation basée sur la liste de gènes pathologique connus

Ce second type de validation a été utilisé pour les méthodes de priorisation utilisant les gènes DE comme données de départ. Dans ce cas, la méthode de priorisation a été considérée comme valide s'il y avait conformité entre les gènes prédits comme pathologiques et les gènes pathologiques déjà connus. Zhao *et al.* (2011) ont utilisé ce type de procédure pour valider leur approche de priorisation (basée sur la centralité de vecteur propre) sur 40 jeux de donnés d'expression transcriptomique dont celui du carcinome à cellules rénales.

Les 2 approches de validation que nous venons d'énoncer se sont toutefois heurtées à la même limite. Lorsque les gènes candidats ne correspondent à aucun gène pathologique connu, la méthode de priorisation est considérée comme non performante. Or il n'est toutefois pas exclu que le gène candidat corresponde réellement à un gène pathologique, mais nouveau car encore non découvert comme tel. La validation croisée s'est affranchie de ce biais en multipliant l'utilisation des sets de gènes pathologiques connus pour valider l'approche de priorisation ; la validation basée sur la liste de gènes pathologiques connus a quant à elle contourné ce biais en élargissant l'accès des gènes pathologiques connus au-delà de la base OMIM, via une recherche bibliographique approfondie ou des test d'enrichissement utilisant GEO ou KEGG.

Choix d'une méthode de priorisation

À l'instar de ce qui a été constaté pour les méthodes de centralité il est difficile de choisir une méthode particulière de priorisation. Les méthodes de priorisation reposant sur des hypothèses différentes, elles prédisent des gènes pathologiques différents. C'est ce qu'observent Navlakha et Kingsford (Navlakha et Kingsford, 2010) qui listent les gènes prédits comme pathologiques par plusieurs méthodes de priorisation (notamment celles de Oti (2006), Vanunu et Sharan (2010) et Köhler *et al.* (2008)) et constatent qu'il n'y a quasiment aucune superposition entre les différentes méthodes. De plus, la performance d'une méthode est différente suivant le réseau sur lequel on s'appuie comme le montrent Hsu *et al.* (2011). Dans cette étude, plusieurs méthodes de priorisation sont appliquées (en particulier Köhler *et al.* (2008) et Vanunu et Sharan (2010)) sur 2 réseaux PPI différents, l'un composé uniquement de PPI directs (réseau interactions), l'autre prenant en plus en compte les liens fonctionnels (réseau fonctionnel), sélectionné à partir de STRING. En utilisant la procédure de validation croisée, les auteurs constatent que les performances de chacune des méthodes sont meilleures dans le réseau fonctionnel. Ils ajoutent également que la combinaison de plusieurs méthodes de priorisation génère une plus-value.

Il n'existe pas actuellement de consensus quant au choix d'une méthode de priorisation. La meilleure approche reste de tester plusieurs méthodes pour évaluer leurs pertinences et leurs performances dans le contexte d'intérêt (Guala et Sonnhammer, 2017).

Nom de l'outils	Lien	Publication
MaxLink	maxlink.sbc.su.se/	(Östlund *et al.*, 2010)
DADA	compbio.case.edu/omics/software/dada/ (modules matlab)	(Erten *et al.*, 2011)
PRINCE	(cs.tau.ac.il/bnet/software/PrincePlugin/ (module cytoscape)	(Vanunu et Sharan, 2010)
HumanNet v2	inetbio.org/humannet/	(Hwang *et al.*, 2018)
PINTA	esat.kuleuven.be/	(Nitsch *et al.*, 2011)

Tableau I.1.6 – Outils de priorisation disponibles

Application de la priorisation en néphrologie

Les stratégies de priorisation explorant l'interactome dans son intégralité sont très prometteuses pour identifier de nouveaux acteurs moléculaires des maladies dans la mesure où elles sont capables de mettre en avant non seulement des composés dont l'expression n'est pas modifiée dans la maladie étudiée (non-DE) mais également des molécules non détectables dans l'échantillon testé. Malgré ce fort potentiel, ces méthodes de priorisation sont très peu utilisées pour la recherche médicale dont celle en néphrologie. Cela est essentiellement dû à la difficulté qu'ont les biologistes à s'approprier les méthodes décrites en termes d'équations par les informaticiens / mathématiciens et, inversement, à la difficulté qu'ont ces derniers à simplifier ces outils pour les rendre accessibles à plus grande échelle. Des efforts sont toutefois faits pour réduire le fossé entre les disciplines, ce qui se traduit notamment par l'implémentation de quelques outils de priorisations sous forme d'applications Web (Tableau I.1.6).

Conclusion

Nous avons vu qu'il existe plusieurs méthodes pour identifier, à partir de données expérimentales, de nouveaux acteurs clés des pathologies et améliorer de ce fait notre compréhension des mécanismes à l'origine des maladies.

Ces approches n'ont cependant pas encore été très utiles pour les recherches sur les pathologies rénales du fait de trois limites principales que nous avons déjà évoquées. Premièrement, certaines méthodes proposent uniquement des candidats dont l'expression est modifié dans la pathologie d'intérêt. Or l'implication d'une molécule en physiopathologie peut être indépendante de son expression. De plus, dans la mesure où l'organisme humain est cloisonné (compartiment vasculaire, tissu, compartiment intracellulaire, organite ...), il est possible que certains acteurs clés des maladies soient absents de l'échantillon utilisé. Deuxièmement, lorsque les méthodes ont été appliquées au domaine de la néphrologie, c'est la plupart du temps en partant d'analyses transcriptomiques, relatives donc à l'expression des gènes. Or l'angle de vue des analyses protéomiques pourrait se révéler plus intéressant dans la mesure où l'expression des protéines est considérée comme mieux corrélée au phénotype des individus que ne l'est l'expression des gènes. Troisièmement, Il existe encore une certaine distance entre le monde de la bio-informatique, qui a généré ces méthodes analytiques des acteurs pathologiques, et le monde de la néphrologie qui souhaite appliquer ces méthodes au service de leur problématique médicale. Les deux domaines auraient pourtant beaucoup à gagner dans une coopération plus forte. En effet,

le développement des méthodes de bio-informatique présentées ici repose toujours sur l'exploitation de données déjà générées, le plus souvent par des biologistes. De plus la validation de ces méthodes doit toujours être soutenue, à plus ou moins long terme, par des travaux expérimentaux. D'un autre côté, la biologie médicale et plus particulièrement la néphrologie doit s'approprier ces outils en complément d'approches expérimentales pour réduire les coûts en termes de temps et d'argent.

C'est donc dans ce contexte que se place ce travail de thèse avec pour objectif de proposer une méthode pour identifier, à partir d'échantillons de fluides biologiques, de nouveaux acteurs clés des maladies rénales. Nous nous sommes particulièrement intéressés aux méthodes de priorisation puisqu'elles sont capables de mettre en avant des molécules non détectables dans l'échantillon. Nous pensons que cette méthode serait bénéfique à la recherche en néphrologie, en dépassant les contraintes expérimentales.

2 PRYNT, une méthode de priorisation du protéome urinaire au service des maladies rénales - Résultats

> "L'image que j'ai en tête est celle d'un ensemble de points qui sont reliés par des lignes. Les points de cette image sont des individus, ou parfois des groupes, et les lignes indiquent quelles sont les personnes qui interagissent les unes avec les autres."
>
> *John Arundel Barnes (1954)*

L'urine constitue un liquide de premier choix dans le cas particulier des maladies rénales, sa composition étant un très bon reflet de la fonction rénale et du fait de sa facilité d'obtention. Les protéines sont les effecteurs des actions biochimiques et le protéome, qui représente le contenu global en protéines, est un bon indicateur des processus cellulaires. L'analyse du protéome urinaire peut donc s'avérer fortement utile pour mieux comprendre la physiologie rénale et surtout ses dérèglements en pathologie. Toutefois, toutes les protéines rénales ne sont pas détectables dans l'urine. Pour avoir une image complète des fonctions pathologiques rénales, il peut donc être judicieux de compléter l'approche de protéomique urinaire par une analyse du réseau PPI. Les méthodes de priorisation sont particulièrement intéressantes dans ce contexte puisqu'elles permettent d'identifier de nouvelles molécules clés en pathologie alors même que leur expression n'est pas mesurable dans l'échantillon utilisé. J'ai donc développé une nouvelle méthode, nommée PRYNT (PRioritization bY causal NeTwork), qui exploite les méthodes de priorisation du réseau PPI global à partir des données de protéome urinaire différentiel. J'ai ainsi montré qu'il est possible d'utiliser l'urine comme une source d'information pour identifier de nouvelles protéines importantes dans les maladies rénales.

Prioritization of disease candidates from proteomics data using a combination of shortest-path and random walk algorithms.

Franck Boizard, Bénédicte Buffin-Meyer, Julien Aligon, Olivier Teste, Joost P. Schanstra, Julie Klein.

en cours de publication :
Bioinformatics

Abstract

Motivation

Urine has been shown as a promising pool of biomarkers of kidney disease. However, the molecular changes observed in urine only partially reflect the deregulated mechanisms within the kidney tissue.

Results

In order to improve on the mechanistic insight based on urinary molecular traits, we developed a new prioritization strategy called PRYNT (PRioritization bY causal NeTwork) that employs a combination of two closeness-based algorithms, shortest-path and random walk, and protein-protein interaction (PPI) network. In order to assess the performance of our approach, we evaluated both precision and specificity of PRYNT in prioritizing known kidney disease candidates. Using four urinary proteome datasets associated to kidney diseases, PRYNT prioritization performed better than the standard reference method used by biologists that prioritizes experimental observations based on their p-value. Moreover, PRYNT performed to a similar, but complementary, extent compared to the upstream regulator analysis from the commercial IPA software. In conclusion, PRYNT appears to be a valuable freely accessible disease candidate prioritization tool for omics data and could be applied to other biofluids and diseases.

Introduction

Kidney diseases can be defined as any chronic or acute disorder that affects renal structure and function (Levey *et al.*, 2013). In their most severe form, they are associated with a variety of complications, such as anemia, mineral bone disease or cardiovascular disease, leading to overall increased mortality (Thomas *et al.*, 2008). Causes of renal failure are highly variable and sometimes unknown (Levey et Coresh, 2012). Some kidney diseases are monogenic, resulting from modifications in a single gene. Others are more complex and can result from a multifactorial combination of genetic, environmental and additional modifiers such as age, diabetes, smoking or hypertension. This multifaceted pathology creates challenges for drug and biomarker discovery approaches. Current therapeutic strategies mostly target signs and symptoms of the disease (e.g. inhibitors of the renin–angiotensin system targeting blood pressure and cardiovascular mortality in chronic kidney disease). However, these drugs only slow down the progression of the disease and alternative therapies directly targeting the affected molecular pathways are still necessary. Therefore, identifying key molecular changes in kidney diseases represent an urgent challenge.

Depending on the cause, the severity and the rate of progression of renal failure, typical omics experiments produce a large number of differentially expressed molecular traits. Furthermore, major advances in the field of omics analyses, improving their accessibility, have

led to an exponential increase in available experimental data. While genomics is frequently used to unravel specific mutations in the genome that can increase the risk of developing certain diseases, disease activity is best captured by transcriptome or proteome analysis, as these traits are closer to the phenotype. For example, protein changes can accurately reflect the actual pathophysiological mechanisms associated with kidney diseases such as long-term modifications of extracellular matrix accumulation (i.e. fibrosis) or short-term modifications of cytokine production (i.e. inflammation) (Filip *et al.*, 2014). Ideally, such changes should be studied directly using human kidney tissue. However, kidney biopsies are rarely available for omics-oriented research purposes as this is an invasive procedure associated with hemorrhagic and infectious risks. Alternatively, urine can be collected easily and non-invasively. Urinary proteins predominately originate (70 %) from kidney and urinary tract by mechanisms of secretion and cellular shedding (Decramer *et al.*, 2008; Jia *et al.*, 2009; Pieper *et al.*, 2004). Studies focusing on the urinary proteome composition have led to the identification of more than 6000 urinary proteins (Adachi *et al.*, 2006; Zhao *et al.*, 2017). For this reason urine can be considered as a "liquid biopsy" of the kidney and urogenital tract and a number of studies have performed urinary proteome analysis in order to identify biomarkers of kidney diseases (Bakun *et al.*, 2012; Chen *et al.*, 2018; Lacroix *et al.*, 2014; Mischak, 2015; Rauniyar *et al.*, 2018). Most of these studies considered urinary proteins showing most prominent changes, either based on fold change or p-value, as new promising disease-related candidates. However, not all renal proteins can be found in urine and not all urinary proteins originate from the kidney. Hence, ranking disease proteins solely based on observed urinary changes might limit the complex view of the disease and insight in its pathophysiology.

To help decipher the picture of the deregulated molecular networks and prioritize disease candidates, computational methods and tools have been proposed. Some of these approaches rely on the generation of causal networks and use protein-protein interaction (PPI) networks to elucidate upstream biological causes that can explain the observed molecular changes (Chindelevitch *et al.*, 2012; Babur *et al.*, 2018). Such network-based approaches present the advantage of identifying disease candidates that were absent from the initial set of experimental observations, filling the gaps in the molecular pathophysiological pathways. Ingenuity Pathway Analysis (IPA) is a commercial software used by biologists in order to interpret high-throughput expression data. In IPA, the "Upstream Regulator Analysis" (URA) algorithm prioritizes disease candidates using in-house causal network approach based on a PPI network containing millions of structured, manually curated experimental observations (Krämer *et al.*, 2014). One of the main limitations hampering the use of IPA is that the software is proprietary and therefore its use cannot be broadly generalized to the biology community. Apart from this tool, many efforts have been made to develop open-source prioritization approaches. Some methods prioritize candidates based on whether they directly interact with known disease genes, following the principle of "guilt-by-association" (Oti, 2006; Ren *et al.*, 2019). Other methods, such as shortest-path (Simões *et al.*, 2012) or random walk (Köhler *et al.*, 2008) algorithms, further consider the closeness between candidates and known disease genes in a network considering both direct and indirect relationships. Previous studies have shown that closeness-based approaches outperformed direct neighbour-based methods and that

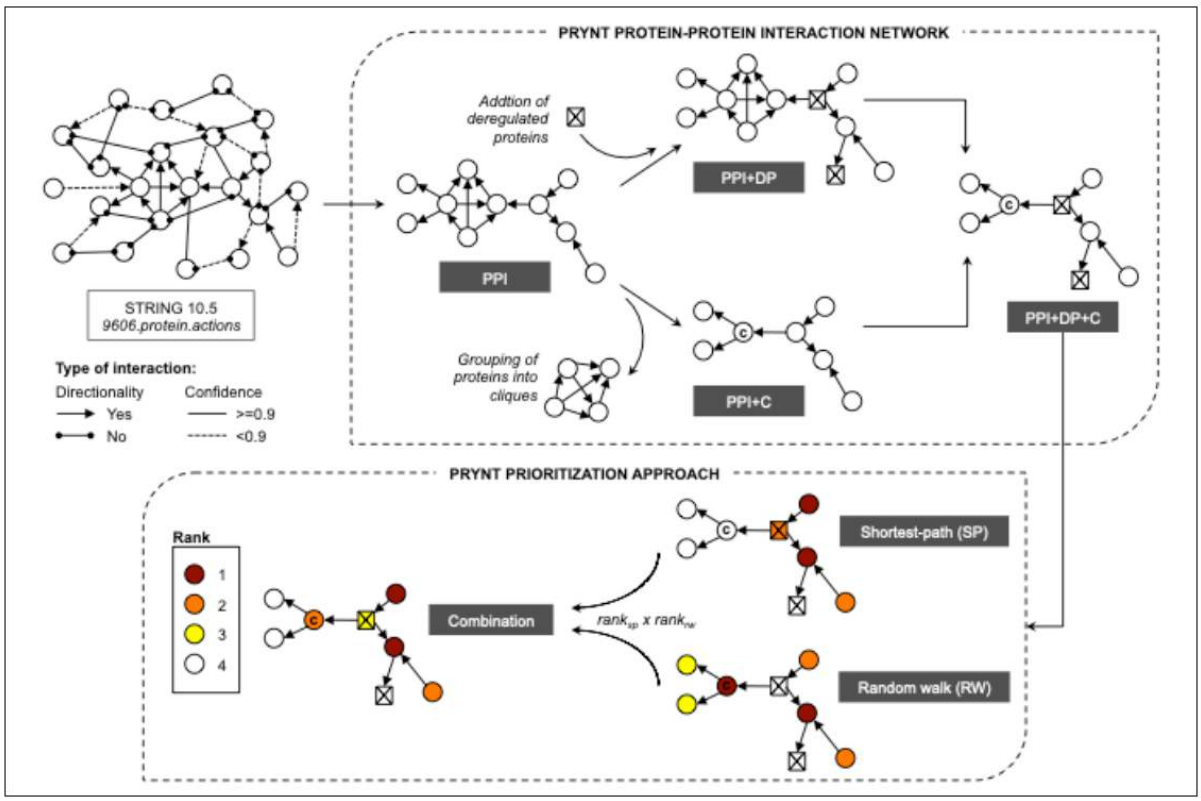

Figure I.2.1 – Description of PRYNT algorithm PRYNT PPI network was based on STRING 10.5 protein.actions restricted to Homo sapiens (*9606.protein.actions*), and only directional interaction with confidence >=0.9 were selected. The raw PPI network was further contextualized by adding the deregulated proteins regardless of their confidence level and by grouping the proteins within cliques (PPI+DP+C). PRYNT prioritization approach was based on the combination of shortest-path and random walk algorithms and was achieved by multiplying the rank of the protein with the shortest-path ranking strategy ($rank_{sp}$), and the rank of the protein with the random walk strategy ($rank_{rw}$). DP: deregulated proteins; C: clique.

combining closeness-based approaches further improved disease candidate prioritization (Ren *et al.*, 2019; Hsu *et al.*, 2011).

In this study, we developed PRYNT (PRioritization bY causal NeTwork), a new approach that could help identify and prioritize disease candidates using a combination of shortest-path and random walk algorithms. Integrating a PPI network and differentially expressed urinary protein profiles, we used PRYNT in the context of human kidney disease with two proteome datasets associated with autosomal dominant polycystic kidney disease (ADPKD) and two proteome datasets associated with ureteropelvic junction obstruction (UPJ), as prototypic monogenic and complex kidney diseases respectively. In order to assess the performance of our approach, we evaluated both precision and specificity of PRYNT in prioritizing known ADPKD and UPJ disease candidates. Finally, we compared the performance of PRYNT to two main reference prioritization methods currently used by biologists: prioritization based on experimental results and prioritization based on IPA's URA algorithm.

	Proteomic dataset	STRING v10.5			PPI	
		Absent	No 'network.actions'	No 'highest confidence'	Raw	+DP
ADPKD1	155	1	27	49	78 (50%)	127 (82%)
ADPKD2	69	0	11	15	43 (62%)	58 (84%)
UPJ1	174	12	33	47	82 (47%)	129 (74%)
UPJ2	186	12	56	31	87 (47%)	118 (63%)

Table I.2.1 – Number of deregulated urinary proteins (DP) from proteomic datasets present in raw PPI and in contextualized PPI+DP networks.

Results

Contextualization of PRYNT PPI network

Approximately 50-60 % of the deregulated urinary proteins from ADPKD and UPJ proteomic datasets were present in STRING PPI network (Table I.2.1). This rather low percentage could be explained either because part of the deregulated proteins were absent from STRING database, or because they did not share any 9606.protein.actions interactions with other proteins in the network, or because their interactions did not reach the highest confidence level (Table I.2.1). Moreover, 56 % (3569 proteins) of the 6391 proteins present in the network were grouped in 265 cliques, which are sets of proteins that all interact with each other and often share similar biological functions. In order to assess the impact of the missing biological input and of the presence of clique subgraphs in the network, we modified the raw PPI network into three additional contextualized PPI networks (Figure I.2.1). The first contextualization consisted in generating a PPI network where the deregulated urinary proteins were added regardless of their confidence level (Figure I.2.1, PPI+DP). In this PPI+DP network, 60-80 % of the deregulated proteins were now present (Table I.2.1). The second contextualization consisted in generating a PPI network where cliques were taken into account (Figure I.2.1, PPI+C). The last network combined both contextualization strategies (Figure I.2.1, PPI+DP+C). We applied the prioritization strategy combining shortestpath and random walk on the four different PPI networks on the four proteomics datasets (Table I.2.2 and Figure I.2.1). We compared the ranked lists to a list of 500 known disease candidates of ADPKD for ADPKD1 and ADPKD2, and of UPJ for UPJ1 and UPJ2. The precision was plotted (Figure I.2.2 A-D) and the areas under the precision curves (AUC) were compared (Figure I.2.2 E-H). Compared to the raw PPI, the use of the contextualized PPI+DP and PPI+C networks slightly increased the AUC of the precision. However, in the four datasets, the combined PPI+DP+C showed much better performance in terms of prioritizing disease candidates. Based on these results, we generated a contextualized PRYNT PPI network combining both the addition of the deregulated proteins and the management of the cliques (Figure I.2.1).

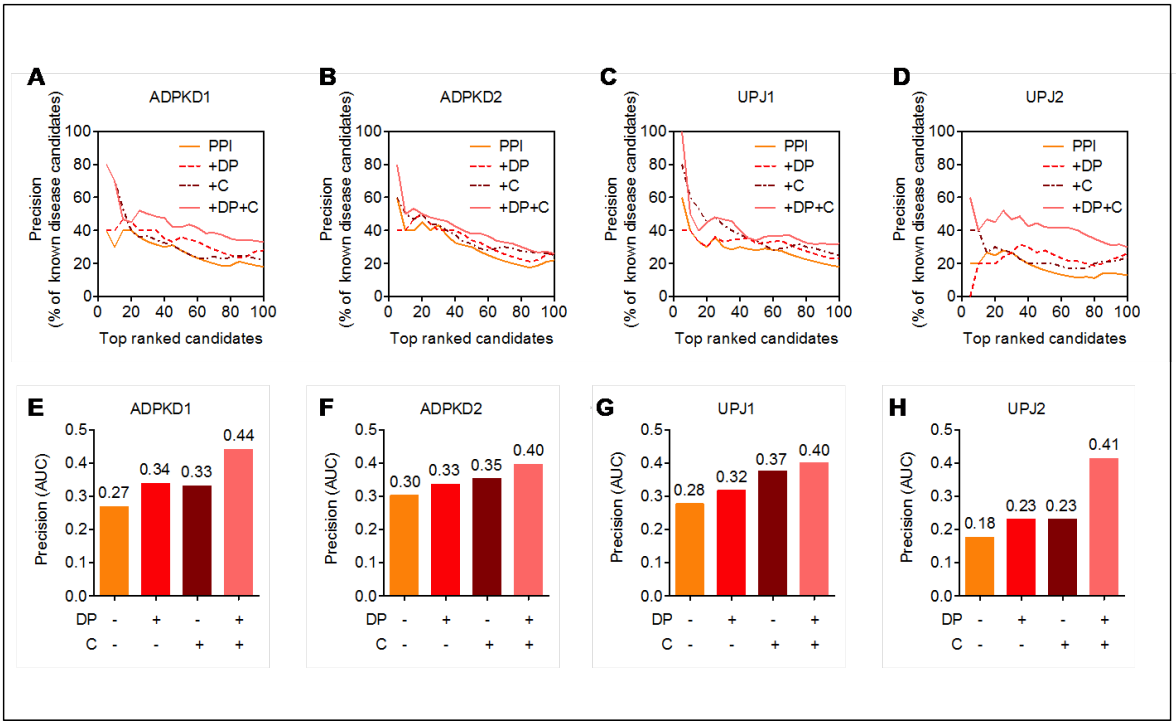

Figure I.2.2 – Comparison of the performance of PRYNT depending on PPI network contextualization. (A-D) The precision was calculated based on the percentage of known ADPKD (A-B) or UPJ (C-D) disease candidates that were prioritized in the top 100 candidates ranked by PRYNT in the four datasets using either the raw PPI network or the PPI networks contextualized by the addition of deregulates urinary proteins regardless of their confidence level (+DP), by the management of clique subgraphs (+C) or by the combination of both (+DP+C). **(E-H)** The corresponding area under the precision curve (AUC) was calculated for network in the four datasets.

Precision of PRYNT compared to reference approaches

PRYNT performance was then compared to a prioritization strategy based on URA algorithm and prioritization based of experimental results (Exp), two reference approaches commonly used by biologists. In the four datasets, PRYNT showed increased performance to prioritize known disease candidates compared to URA and Exp, with better precision and superior AUC (Figure I.2.3). We also compared PRYNT to direct, shortest-path and random walk ranking algorithms, three common prioritization strategies and found that PRYNT displayed higher precision in 10/12 of the cases (Figure I.2.4). We next analyzed the overlap of known disease candidates ranked in the top 100 by PRYNT, URA and Exp in the four datasets (Figure I.2.5). We observed that only a minority of known disease candidates prioritized by PRYNT and URA were commonly (59-70 % uniquely prioritized by PRYNT and 48-64 % uniquely prioritized by URA). For Exp, not only the number of prioritized known disease candidates was very low, but also it showed very poor overlap with URA and no overlap with PRYNT.

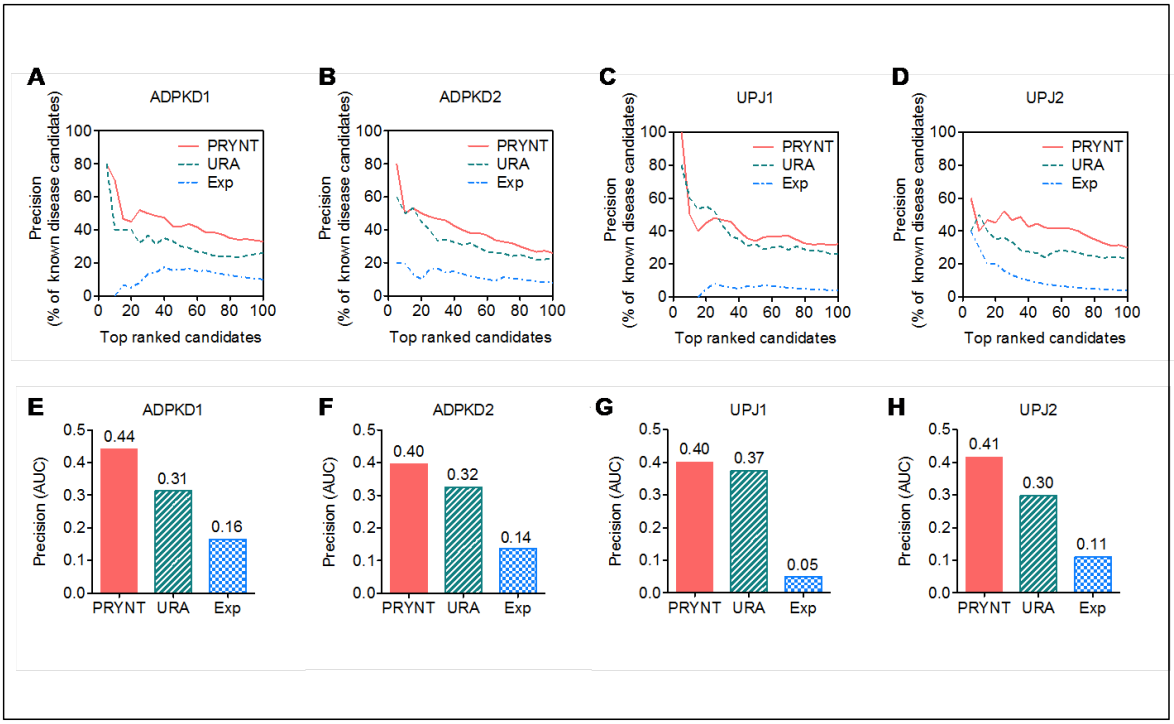

Figure I.2.3 – Performance of PRYNT compared to reference approaches. **(A-D)** The precision was calculated based on the percentage of known ADPKD **(A-B)** or UPJ **(C-D)** disease candidates that were prioritized in the top 100 candidates ranked by the different strategies in the four datasets. **(E-H)** The corresponding area under the precision curve (AUC) was calculated for each prioritization strategy in the four datasets. Exp: experimental, URA: upstream regulator analysis.

Specificity of PRYNT compared to reference approaches

We next assessed how the different prioritization strategies ranked candidates that were specific to the disease under study. First, we studied cross-specificity by analyzing whether prioritization in ADPKD datasets was better for specific ADPKD known disease candidates compared to non-specific UPJ known disease candidates, and conversely for UPJ datasets. For ADPKD1 and ADPKD2, all prioritization strategies showed similar cross-specificity, with the AUC for specific candidates being superior to the AUC for non-specific candidates (Figure I.2.6 A-B). However, for UPJ1 and UPJ2, only PRYNT displayed adequate cross-specificity in both datasets (Figure 4C-4D). We next compared overall specificity of the approaches by comparing the AUC of the specific disease to the AUC of 80 non-specific diseases. For APDKD datasets, overall specificity was similar for all strategies in ADPKD1 with the AUC of the specific disease being in the top 15 out of 80 non-specific diseases (Figure I.2.7A). In ADPKD2, PRYNT showed better performance compared to URA and Exp (rank of specific AUC of 14/80, 34/80 and 21/80 for PRYNT, URA and Exp respectively) (Figure I.2.7 B). For UPJ datasets, overall specificity was lower compared to ADPKD datasets and in both datasets, PRYNT prioritization showed best specificity, with a rank of specific AUC of 21/80 and 27/80 for UPJ1 and UPJ2 respectively (Figure I.2.7 C-D). In UPJ2, Exp showed the lowest specificity with the specific AUC being ranked 65/80 (Figure I.2.7D).

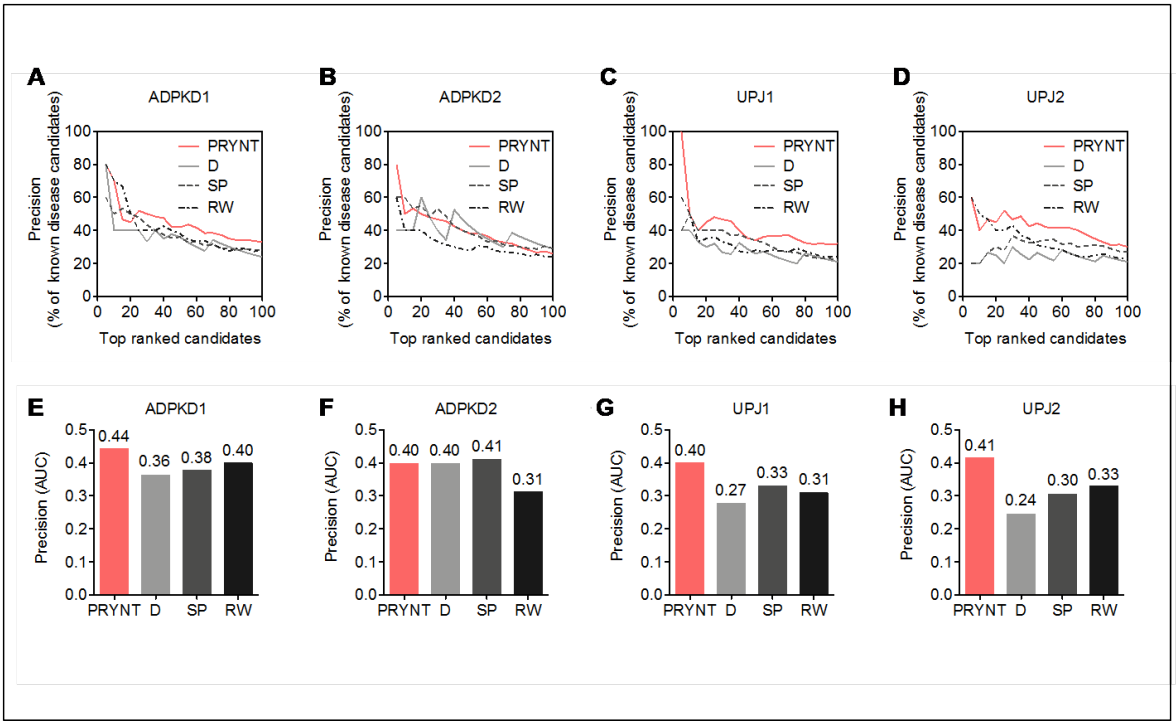

Figure I.2.4 – Performance of PRYNT compared to common prioritization strategies. (A-D) The precision was calculated based on the percentage of known ADPKD **(A-B)** or UPJ **(C-D)** disease candidates that were prioritized in the top 100 candidates ranked by the different strategies in the four datasets. **(E-H)** The corresponding area under the precision curve (AUC) was calculated for each prioritization strategy in the four datasets. D: direct, SP: shortest-path, RW: random walk.

Pathway annotation

Finally, we used KEGG pathway enrichment analysis to assess the biological relevance of the disease candidates prioritized by PRYNT. For ADPKD, the 500 known disease candidates were associated to 166 pathways. Approximately 85 % of these pathways were also enriched with the top 100 ranked candidates prioritized by PRYNT (141/166 and 139/166 for ADPKD1 and ADPKD2 respectively) (Figure I.2.8 A-B) whereas enrichment was 67-72 % for URA top 100 (112/166 and 119/166 for ADPKD1 and ADPKD2 respectively) and dropped to approximately 5 % for Exp (9/166 and 10/166 for ADPKD1 and ADPKD2 respectively)(Figure I.2.8 A-B). Similarly for UPJ, PRYNT results showed higher number of enriched pathways and more overlapping pathways associated to the known UPJ candidates compared to URA and Exp (Figure I.2.8 C-D).

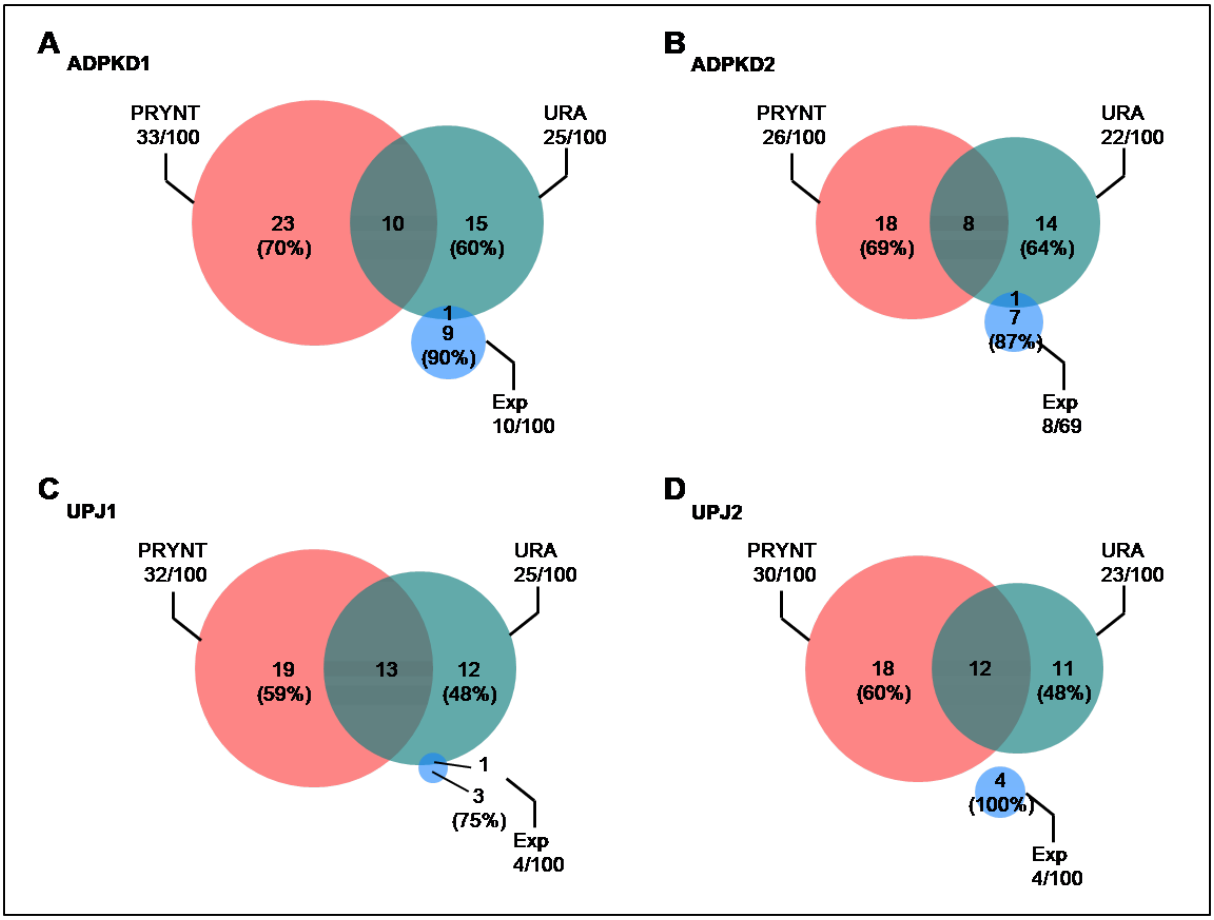

Figure I.2.5 – Overlap of known disease candidates prioritized in the top 100 by PRYNT, URA or Exp. Prioritization by PRYNT, URA or from the experimental urinary proteomic candidates (Exp) was applied and known ADPKD **(A-B)** and UPJ **(C-D)** disease candidates ranked in the top 100 were compared in ADPKD1 **(A)**, ADPKD2 **(B)**, UPJ1 **(C)** and UPJ2 **(D)**. Exp: experimental, URA: upstream regulator analysis.

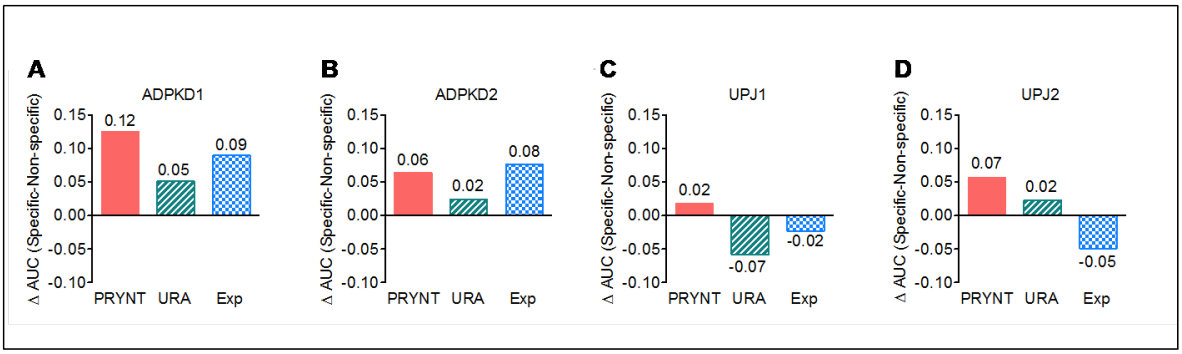

Figure I.2.6 – Cross-specificity of PRYNT compared to reference approaches. **(A-D)** Cross-specificity of the prioritization strategies between ADPKD and UPJ was assessed for the four datasets by calculating the difference between the AUC of the precision curve for specific disease candidates (ADPKD candidates for ADPKD datasets and UPJ candidates for UPJ datasets) and the AUC of non-specific disease candidates (UPJ candidates for ADPKD datasets and ADPKD candidates for UPJ datasets).

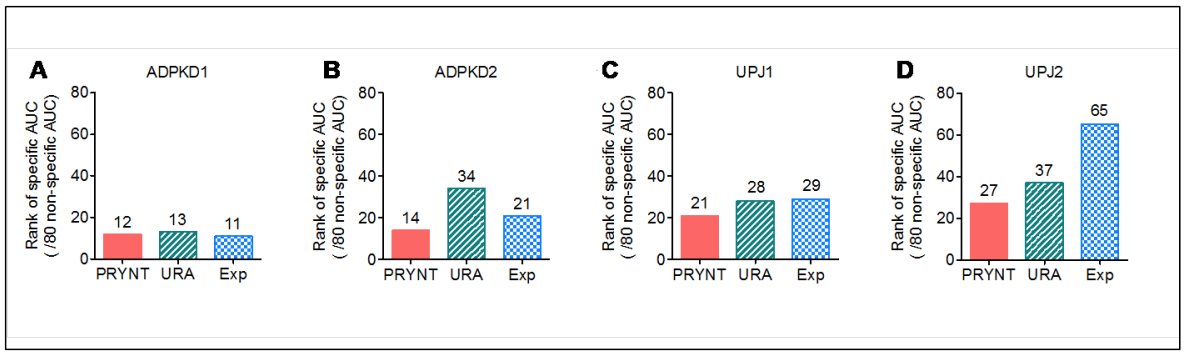

Figure I.2.7 – Overall specificity of PRYNT compared to reference approaches. (A-D) Overall specificity of the prioritization strategies was assessed for the four datasets by comparing the rank of the AUC of the precision curve for specific known disease candidates to the rank of the AUC of known candidates from 80 non-specific diseases, including 40 diseases associated to urogenital tract and 40 diseases from other origin. Exp: experimental, URA: upstream regulator analysis.

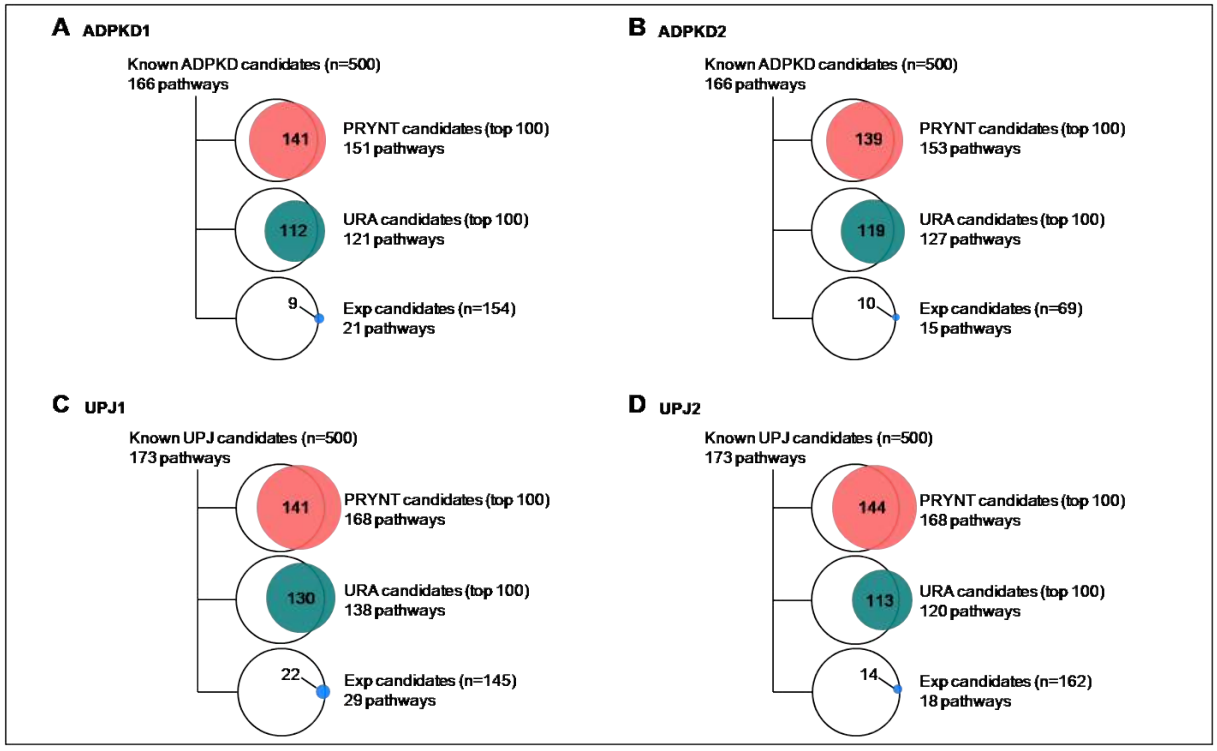

Figure I.2.8 – Pathway annotation. KEGG pathway enrichment analysis was applied to the 500 known ADPKD **(A-B)** and UPJ **(C-D)** disease candidates, and compared to the pathways enriched from top 100 ranked candidates by PRYNT or URA or from the experimental urinary proteomic candidates (Exp) in ADPKD1 (A), ADPKD2 (B), UPJ1 (C) and UPJ2 (D).

Discussion

In this study we assessed the performance of PRYNT, a new network-based approach using urinary proteomics profiles to prioritize disease candidates in the context of kidney disease. Recent advances in high-resolution analytical omics technologies have resulted in major advances in the elucidation of diverse molecular pathophysiological mechanisms. However, the remaining challenge associated with such analysis is that these techniques require time-consuming validation experiments to try precisely pinpointing the most probable disease candidate from a list of hundreds of potential candidates. Moreover, while urine has been known for a very long time as a very informative and non-invasive source of potential candidates in the context of kidney disease, the molecular changes observed in urine partially reflect the deregulated mechanisms within kidney tissue. In order to move from this status quo, we developed an approach that could help expand and fill the gaps of the molecular view, and short-list new and most probable candidates of kidney disease using urinary proteomics data.

The prioritization strategy that was developed for PRYNT was a combination of shortest-path and random walk, two closeness-based algorithms as it has been previously shown in the literature that this method outperformed other computational methods (Ren *et al.*, 2019; Hsu *et al.*, 2011). This was confirmed in this study as PRYNT showed better performance compared to direct ranking, or to shortest-path and random walk ranking alone in the majority of the cases (Figure I.2.4). Most importantly, we further validated that PRYNT showed better precision and better specificity in prioritizing known disease candidates compared to IPA's URA and to ranking based on experimental results, two references approaches commonly used by biologists.

In order to assess the performance of the method, we compared the top 100 ranked candidates to a list of disease candidates already known to be associated to the disease under study. As proof of concept, we selected two prototypic kidney diseases: ADPKD is a well-characterized monogenic kidney disease, and UPJ is a congenital kidney disease probably resulting from a complex multifactorial combination of genetic and environmental factors. The complexity of UPJ mechanisms was highlighted by the fact that almost 18000 known disease candidates were associated to UPJ in CTDbase, while only 500 were associated to ADPKD. This could explain why all the prioritization methods were less specific in the context of UPJ compared to ADPKD, as a larger spectrum of deregulated mechanisms is expected to overlap with more disease conditions.

Prioritization of experimental results based on p-value displayed very low performance in ranking known disease candidates. Moreover, pathway enrichment based on PRYNT or URA top 100 was more informative than enrichment with the initial set of experimental results as not only more pathways were associated to the short-listed candidates, but most importantly, identified pathways were more representative of the pathophysiological mechanisms known to be associated with the disease under study. This tends to confirm that although urine could be a very promising source of biomarker of kidney disease, complementary methods are still required to allow a closer look into pathophysiology. In this context, PRYNT has proven to be a valuable alternate method compared to commercially available tools such as IPA's URA. Interestingly, PRYNT and URA appear as

	Reference	Type of kidney disease	Controls	Cases	Deregulated proteins
ADPKD1	Bakun et al. 2012	Monogenic	30	30	155
ADPKD2	Rauniyar et al. 2018	Monogenic	18	14	69
UPJ1	Lacroix et al. 2014	Complex	5	5	174
UPJ2	Chen et al. 2018	Complex	23	23	186

Table I.2.2 – Dataset description

two independent, complementary prioritization strategies as the known disease candidates short-listed in the top 100 showed rather poor overlap.

In conclusion, using PRYNT PPI network and prioritization strategy combining random-walk and short path algorithm could be of great benefit to identify new key proteins associated to renal diseases from urinary proteomic datasets obtained non-invasively. Such approach, that could be applied to any other form of biological fluid and generalized to any other disease, will help fill the gaps and generate the missing links necessary to better understand the deregulated molecular networks, identify new potential biomarkers or develop alternative therapeutic strategies.

Material and methods

Urinary proteomic datasets

In order to test PRYNT approach, four urinary proteome datasets were used: two associated with ADPKD (ADPKD1 and ADPKD2) and two associated with UPJ (UPJ1 and UPJ2) (Table I.2.2). The first study by Bakun *et al.* (Bakun *et al.*, 2012)(ADPKD1) analysed urine protein composition from 30 ADPKD patients and 30 healthy volunteers identifying 155 differentially abundant proteins. A second study by Rauniyar *and al.* (Rauniyar *et al.*, 2018)(ADPKD2) compared 14 urine samples from ADPKD patients to 18 normal controls and identified 69 significantly deregulated proteins. Lacroix *and al.* (Lacroix *et al.*, 2014) explored the urinary proteome of newborns with UPJ and discovered 174 differentially abundant proteins between 5 individuals with severe UPJ and 5 healthy individuals. Chen *et al.* (Chen *et al.*, 2018)(UPJ2) analysed the proteome of urine from 23 infants with UPJ and 23 controls and identified 186 proteins with different urinary abundance between the two groups.

Protein-protein interaction network

Protein-protein interaction network

We constructed a PPI network based on the Search Tool for the Retrieval of Interacting (STRING, version 10.5) (Szklarczyk *et al.*, 2015). STRING is a comprehensive database of PPI based on experimental evidence as well as interactions predicted by comparative genomics and text mining. In the present study, we used STRING 10.5 protein.actions restricted to Homo sapiens (*9606.protein.actions*), which compiles physical interactions

such as reaction, binding, catalysis, inhibition and activation (Figure 1). Each interaction has a confidence score between 0 and 1 according to the number and the type of source that was used to describe the interaction. Only interactions with the highest confidence level (score greater than 0.9) were selected for PRYNT (Figure 1). Moreover, directionality of the interaction could be applicable or not to its physical action. Only directional interactions were considered for PRYNT analysis as ranking strategies use directionality (Figure I.2.1). After removing duplicates and self-linked interactions, we obtained 353643 interactions between 6391 proteins. The raw PPI network was contextualized by adding the deregulated urinary proteins regardless of their confidence level and by removing cliques (Figure I.2.1). In the PPI network, 3569 proteins of the 6391 were grouped in 265 cliques, each clique containing on average 13.5 proteins. Using R igraph package (Csardi et Nepusz, 2006), we grouped proteins that were part of cliques and selected for each clique the protein candidate with best ranking following prioritization. This led to a PPI network containing 21051 interactions between 3109 nodes (proteins or cliques).

Prioritization approach

Prioritization was based on the combination of two closeness-based approaches, namely shortest-path (Simões *et al.*, 2012) and random walk (Köhler *et al.*, 2008) algorithms. Shortest path between a disease candidate and a differentially abundant urinary protein was defined by the distance between any protein in the network and the differentially abundant proteins, taking into account the direction of interactions. The shortest-path score (SP) of a protein x was calculated as the reciprocal of the sum of the length of the shortest-path between x and the deregulated proteins (y) in the network:

$$SP = \frac{1}{\sum d(x,y)} \tag{2.1}$$

where d(x,y) is the minimum number of interaction from x to y. Disease candidates are ranked from higher to lower SP ($rank_{sp}$) (Figure I.2.1).

Random walk with restarts (Köhler *et al.*, 2008) simulates a random walker starting on differentially abundant urinary proteins and moving to their immediate neighbors randomly at each step. Each protein in the graph is prioritized by the probability of the random walker reaching it. The random walk score (RW) corresponds to the probability of a protein to be reached by the walker at the next step t+1 and can be formally described as follows:

$$RW = (1-r)AP_t + rP_0 \tag{2.2}$$

where A is the column-normalized adjacency matrix; r the restart probability (set to 0.7 as the default parameters); P_0 the initial probability of the random walk, i.e. the inverse of the number of deregulated protein for a deregulated protein and 0 for other proteins in the network; and P_t the probability after the t-th round of the step. Prioritization based on random walk was calculated using the R package RandomWalkRestartMH (Valdeolivas *et al.*, 2019). Disease candidates are ranked from higher to lower RW ($rank_{rw}$) (Figure I.2.1).

For each disease candidate, a combined score (CS) was calculated as:

$$CS = rank_{sp}.rank_{rw} \tag{2.3}$$

where $rank_{sp}$ is the rank of the protein in the shortest-path ranking strategy, and $rank_{rw}$ in the random walk strategy (Hsu *et al.*, 2011) (Figure I.2.1).

Reference prioritization methods

Prioritization based on experimental results

For prioritization based on experimental results, differentially abundant proteins from the four proteomics datasets were ranked based on their p-value (from smallest to largest).

Prioritization based on URA algorithm

Prioritization based on URA algorithm was performed using IPA software (content version release date 2017-12-07). This analysis examines how many known targets of each upstream regulator are present in the experimental dataset. Disease candidates (limited to proteins) were ranked based on the overlap p-value. The overlap p-value, calculated using Fisher's Exact Test, measures whether there is a statistically significant overlap between the experimental dataset and the known targets that are under control of the upstream regulator.

List of known disease candidates

Disease candidates already known to be associated to ADPKD and UPJ were collected from Comparative Toxicogenomics Database (CTDbase) (Davis *et al.*, 2019) using the R package CTDquerier (Hernandez-Ferrer et Gonzalez, 2018). For ADPKD, 504 disease candidates were found to be associated to the term "Polycystic Kidney, Autosomal Dominant". For UPJ, 17786 disease candidates were associated to the term "Ureteral Obstruction". In order to obtain comparable results with ADPKD, we selected the first 500 known disease candidates according to their inference score in UPJ. Moreover, to assess overall specificity of the prioritization strategies, we also collected known disease candidates from 80 other diseases (40 associated to the term "Urogenital disease" and 40 associated to other type of diseases). For each of these diseases, we selected the first 500 known disease candidates according to their inference score.

Precision measurement

In order to evaluate the performance of PRYNT and the reference approaches, we compared the ranked lists to the list of known disease candidates obtained from CTDbase and calculate the precision of each method. The precision of the prioritization is the percentage of known disease candidates in the ranking. The precision curve represents the precision depending on the size of the ranking taken into account. The area under the precision curve (AUC) was estimated using the trapezoidal rule.

Specificity assessment

Cross-specificity

Cross-specificity was assessed by calculating the difference between precision AUC for known specific disease candidates and precision AUC for known non-specific disease candidates. A positive difference was expected to be associated with specific approach while a negative difference was expected to be in favor of a lack of specificity. For ADPKD datasets, specific precision AUC was calculated based on prioritization of known ADPKD candidates, and non-specific AUC was calculated based on prioritization of known UPJ candidates. Conversely, specific precision AUC for UPJ datasets was calculated based on prioritization of known UPJ candidates, and non-specific AUC was calculated based on prioritization of known ADPKD candidates.

Overall specificity

Overall specificity was assessed by ranking precision AUC for known specific disease candidates and precision AUC of known candidates from the 80 non-specific diseases. Specific prioritization method was expected to be associated with specific AUC being in the top ranked AUCs.

Pathway enrichment analysis

KEGG pathway enrichment analysis was performed using the R package limma (Ritchie *et al.*, 2015). A pathway was considered associated to the set of candidates if its p-value was under 0.05.

Availability and implementation

The source code is available on GitHub at: github.com/Boizard/PRYNT.

Conclusion

De nombreuses avancées dans le domaine de la protéomique ont été réalisées depuis plusieurs années avec le développement de techniques de plus en plus rapides et précises qui permettent l'identification et la quantification d'un grand nombre de protéines urinaires. La composition de l'urine est un bon reflet de la fonction rénale. Cependant, toutes les protéines rénales n'y sont pas détectables. L'objectif de mon travail a donc été de mettre au point une méthode pour prédire les protéines importantes dans les maladies rénales à partir de l'information incomplète contenue dans le protéome urinaire.

La position des protéines dans le réseau d'interactions protéine-protéine (PPI) est liée à son rôle biologique et les calculs de centralités sont des mesures qui permettent d'identifier les protéines importantes de ce réseau. Très peu développé aujourd'hui en néphrologie, ce type d'approche a servi de support pour l'élaboration de la nouvelle méthode de priorisation PRYNT (Priorization bY causal NeTwork). PRYNT projette en effet le protéome urinaire d'une maladie sur le réseau PPI et explore le réseau par la combinaison de deux mesures de centralité (algorithmes du plus court chemin et de marche aléatoire). L'application de PRYNT à quatre études du protéome urinaire pathologique a permis d'identifier un grand nombre de protéines pathologiques non détectées dans l'urine, validant de ce fait la méthode. Les performances de PRYNT étaient également meilleures que celles des approches actuellement disponibles (approches expérimentales et IPA). PRYNT est donc une nouvelle méthode pour la compréhension des mécanismes pathologiques des maladies rénales.

Limites et perspectives

PRYNT utilise le réseau PPI comme un modèle représentant les comportements des protéines dans l'organisme. Même s'il est cependant impossible de construire un modèle idéal lorsque l'on étudie les comportements d'un système biologique (Iris *et al.*, 2009), PRYNT comporte comme tout modèle mathématique des hypothèses qui pourraient éloigner de la réalité biologique les interprétations que la méthode génère.

— Le réseau PPI qui a servi au développement de PRYNT prend en compte toutes les interactions quel que soit le contexte dans lequel elles ont été identifiées. Or, nous savons que les interactions se produisent seulement dans un contexte spécifique (un tissu ou un type cellulaire) et sont rarement généralisables à l'ensemble de l'organisme (Kotlyar *et al.*, 2019). La construction d'un réseau PPI contextualisé pour chaque pathologie devrait permettre à la méthode PRYNT d'identifier des processus biologiques plus spécifiques des maladies considérées (Ideker et Krogan,

2012).

— PRYNT ne prend pas en compte la nature des différents types d'interactions qui composent le réseau PPI. Or, les interactions n'ont pas toutes la même signification biologique. Des méthodes comme les réseaux booléens proposent de prendre en compte cette diversité en différenciant les interactions activatrices et inhibitrices (Wang *et al.*, 2012) ce qui pourrait expliquer les niveaux d'abondance des protéines dans les réseaux PPI. Cependant, la complexité de ce type de modèle ne permet aujourd'hui son application qu'à des réseaux restreints (plusieurs dizaines de molécules) (Poret et Guziolowski, 2018; Trairatphisan *et al.*, 2016), bien plus simples que le réseau PPI.

— PRYNT a été développé à partir des données de protéome. Or, si les protéines sont proches du phénotype, elles ne suffisent pas à l'expliquer à elles seules. D'autres niveaux moléculaires (métabolites, peptides, micro-ARN...) jouent également un rôle important dans les processus biologiques et pourraient ainsi participer au phénotype pathologique. La prise en compte de ces composés dans leur globalité est un objectif qu'on doit se fixer. Cela passera par l'obtention de « data-omes » pathologiques, la construction de réseaux multi-omiques et le développement de méthodes capables de projeter ces data-omes sur ces réseaux en vue de les analyser.

Malgré ces limites, PRYNT doit être considéré comme un nouvel outil performant au service de la biologie pour mieux comprendre les processus pathologiques du vivant.

Deuxième partie

Identification de nouveaux biomarqueurs des maladies rénales dans les fluides biologiques

Introduction

Un diagnostic précoce est indispensable pour ralentir la progression des maladies rénales. En effet, plus la maladie est détectée tôt, moins un patient a de chance d'atteindre l'insuffisance rénale terminale (Romagnani *et al.*, 2017). Les maladies rénales sont en générales asymptomatiques dans leurs premières étapes de développement et ainsi peu diagnostiquées aux premiers stades de la maladie (Dousdampanis *et al.*, 2012). Le diagnostic précoce est aussi un enjeu important dans les maladies rénales congénitales (Reznik et Budorick, 1995). Un diagnostic prénatal précoce peut améliorer les conditions de vie de l'enfant atteint grâce à des traitements in utero qui peuvent empêcher des lésions supplémentaires et la perte de la fonction rénale (Hindryckx et De Catte, 2011).

Un biomarqueur est un indicateur objectif, précis et reproductible d'un processus biologique normal ou pathogène (Strimbu et Tavel, 2010). Plusieurs biomarqueurs sont utilisés couramment pour le diagnostic de la fonction rénale. La créatinine sérique pour estimer le débit de filtration glomérulaire ou l'albumine dans l'urine comme indice de sélectivité de la barrière de filtration sont les plus utilisés. Cependant ces approches ont des défauts notables. L'augmentation de la créatinine est un signe tardif de la perte de la fonction rénale (Stevens et Levey, 2009). La mesure de l'albumine dans l'urine est quant à elle très variable d'un individu à l'autre (Zachwieja *et al.*, 2010) et également un signe tardif de la présence d'une maladie rénale. La créatinine et l'albuminurie sont des marqueurs de dysfonction rénale mais ne permettent pas de procéder à un diagnostic d'une maladie en particulier. Certains biomarqueurs provenant du tissu rénal peuvent être de bon marqueurs de maladies rénales. L'immunoglobuline A par exemple se dépose dans le mésangium durant la néphropathie de Berger. Ainsi sa mise en évidence à la biopsie est un marqueur diagnostique de la maladie. Mais l'accès à ces biomarqueurs tissulaires nécessite une biopsie qui n'est prescrite quand les premiers symptômes sont déjà apparus (Mischak *et al.*, 2015). Le même constat existe concernant les moyens de pronostics prénataux des malformations congénitales rénales (Spaggiari *et al.*, 2017a; Aulbert et Kemper, 2016a). La mesure de β2-microglobuline dans le sérum/urine fœtal est une mesure prédictive seulement à un âge gestationnel avancé et les approches non invasives, utilisant de l'imagerie (la visualisation par ultrasons), manque de la valeur prédictive (Morris *et al.*, 2009a).

Les avancées expérimentales, notamment en spectrométrie de masse (Thomas *et al.*, 2016), permettent désormais d'identifier une très grande variété de molécules dans des échantillons aussi complexes que les fluides biologiques, et ce dans des temps d'analyse toujours plus réduit (Fliser *et al.*, 2007; Maizi, 2017). Les études omiques des fluides biologiques pathologiques ont déjà identifié un grand nombre de biomarqueurs potentiels

des maladies (Schrohl *et al.*, 2008; Ahn et Simpson, 2007) et il a été montré que l'urine est un excellent réservoir de biomarqueurs potentiels (peptides, protéines et métabolites) pour un grand nombre de maladies rénales (Brown *et al.*, 2015; Decramer *et al.*, 2006; Klein *et al.*, 2013).

La découverte de biomarqueurs dans les fluides biologiques suscite beaucoup d'intérêt et d'espoir chez les cliniciens car elle ouvre des perspectives de développement de nouveaux tests cliniques (Csősz *et al.*, 2017). Ils peuvent conduire à la conception de modèles incluant un grand nombre de biomarqueurs qui sont mieux adaptés que un simple biomarqueur pour décrire les mécanismes physiopathologiques complexes (Coffman et Richmond-Bryant, 2015). Nous verrons dans cette partie de la thèse comment nous avons mis en place un pipeline de détection des maladies rénales grâce à la modélisation statistique des molécules présentes dans les fluides biologiques. Je vais d'abord décrire brièvement l'outil, appelé *La Boize*, que j'ai développé à l'usage de l'équipe. Par la suite, le chapitre 2 décrit l'étude du métabolome urinaire utilisant cet outil pour la prédiction de a sévérité de l'obstruction de la jonction pyélo-urétérale. Le chapitre 3 présente notre étude sur la prédiction de l'évolution de la fonction rénale des fœtus porteurs d'une anomalie du développement rénal à partir de la composition peptidique du liquide amniotique.

1 *La Boize*, développement d'un outil de diagnostic à partir de données omiques

> "The advent of high-throughput multi-platform genomics technologies providing whole-genome molecular summaries of biological samples has revolutionalized biomedical research. These technologiees yield highly structured big data, whose analysis poses significant quantitative challenges. The field of Bioinformatics has emerged to deal with these challenges, and is comprised of many quantitative and biological scientists working together to effectively process these data and extract the treasure trove of information they contain."
>
> *Morris et Baladandayuthapani (2017)*

L' équipe 12 de l'I2MC a développé une expertise dans l'analyse de la composition des fluides biologiques. Grâce au développement technique de l'électrophorèse capillaire couplé à la spectrométrie de masse ils ont pu analyser différentes liquides biologiques (urine, urine fœtal, liquide amniotique (Klein *et al.*, 2013; Schanstra *et al.*, 2015; Desveaux *et al.*, 2016)) à plusieurs niveaux moléculaires (peptide, métabolite). Ces études génèrent des données volumineuses et viennent avec leurs problématiques propres, les données manquantes par exemple. Le volume et la complexité des données générées ne permettent pas aux biologistes une exploitation optimale de ces informations (Eckel-passow *et al.*, 2009). C'est pour cette raison que j'ai développé un outil bio-informatique facile d'accès, appelée *La Boize* (Figure II.1.1), qui permet d'analyser les données omiques sans connaissance préalable de bio-informatique.

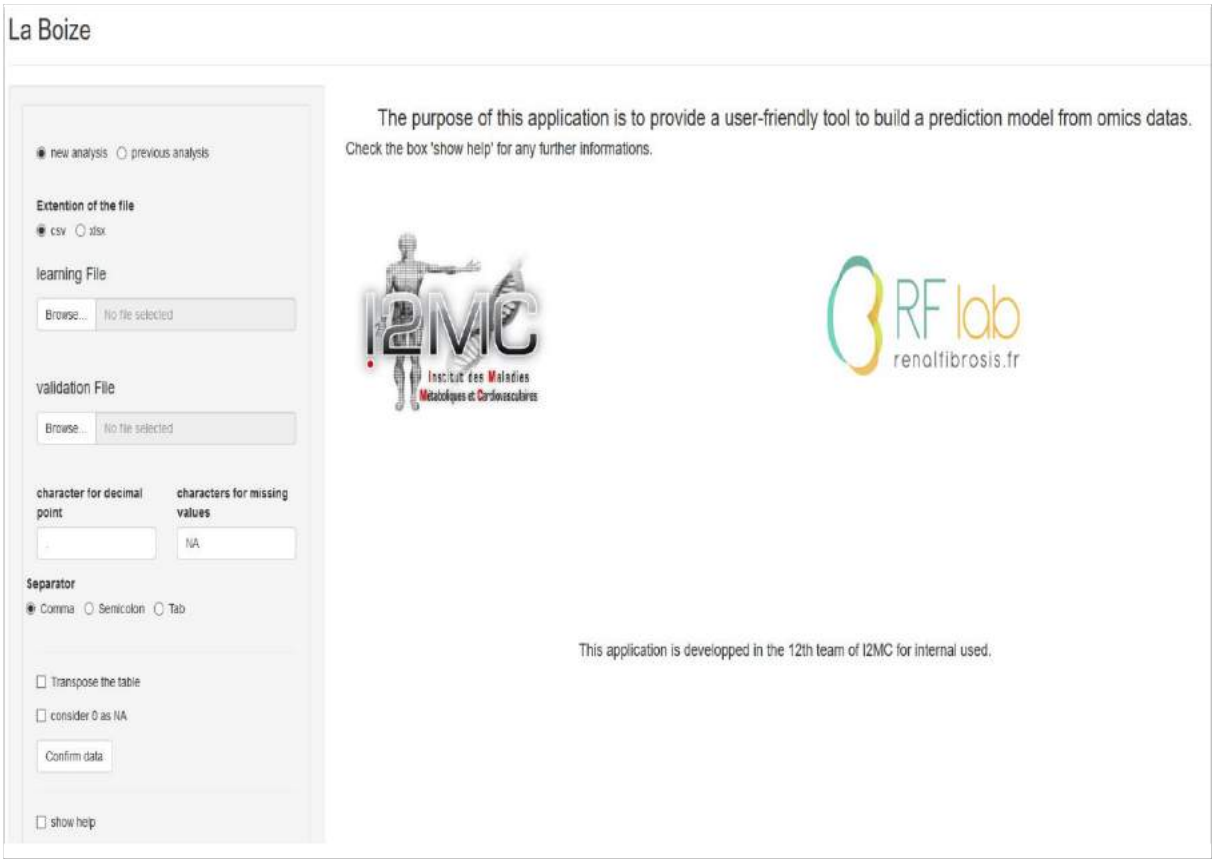

Figure II.1.1 – Page d'accueil de *La Boize*. Le lancement de l'application se fait grâce à deux lignes de code dans l'application R studio. Son utilisation ne nécessite pas de connaissance en programmation.

1.1 Faciliter l'accès aux biologistes

L'objectif principal de cet outil est d'être pris en main directement par les biologistes. Son utilisation ne devrait donc pas nécessiter de connaissance en programmation. Je l'ai développé grâce au package R *Shiny* (Chang *et al.*, 2019), permettant de créer une interface simple. Le script de cette application est disponible sur Github [1]. L'outil est adapté aux besoins et aux habitudes de ce domaine d'application. De ce fait de nombreux choix dans le développement de cet outil ont été faits pour se rapprocher des routines en biologie. Par exemple l'importation des données et le téléchargement des résultats peuvent être faits par des formats communément utilisé comme *Excel*. De plus il est important de pouvoir s'assurer de la bonne compréhension et utilisation de chaque étape du pipeline. Il m'a paru donc important que chaque étape soit pourvue de sortie graphique et que les résultats (intermédiaires ou finaux) soient téléchargeables par l'utilisateur de l'application.

L'une des préoccupations principales dans son développement est de créer un outil adaptable à différentes études. Il n'existe pas de traitement de référence avec ce type de données, en termes de normalisation, de sélection, ou même de modélisation. La conception de cet outil a donc été faite avec la perspective de s'adapter à toutes les utilisations par la mise à disposition de tous les paramètres, réglables directement par l'utilisateur.

1. github.com/Boizard/Laboize

1.2 Les données

Les données manquantes sont un enjeu important pour les données de spectrométrie de masse (Wei *et al.*, 2018; Lazar *et al.*, 2016). La principale problématique est qu'il n'est souvent pas possible de faire la différence entre les différents motifs pouvant mener à la non-détection d'une molécule dans un échantillon : problème technique, présence de la molécule en petite quantité (sous le niveau de détection de la machine) ou réelle absence de l'échantillon. Suivant l'interprétation de ses données manquantes, il peut être nécessaire d'imputer et de sélectionnées les molécules de façon différente. Une partie de mon travail dans le développement de *La Boize* a été d'envisager plusieurs possibilités d'imputation (par la moyenne, par la moyenne du groupe, par zéro, ou imputation par ACP (Dray et Josse, 2015)) et de sélection des données pour s'adapter à l'utilisation dans plusieurs études.

De même une transformation des mesures d'abondance peut être parfois obligatoire suivant la distribution des données. J'ai donc implémenté plusieurs méthodes de transformation. La transformation logarithmique permet de normaliser des données suivant une transformation logarithmique, c'est le cas des données ayant des valeurs extrêmes. La transformation arcsine permet de normaliser les données de proportion, entre 0 et 1. Le centrage-réduction permet d'obtenir des variables indépendantes des unités.

1.3 Identification et validation des biomarqueurs

Les données de départ utilisé par *La Boize* sont deux jeux de données décrivant les abondances des molécules, chez des individus sains (*controls*) et malades (*cases*) (Figure II.1.2). Les données omiques, en particulier les résultats obtenus par spectrométrie de masse, sont de données d'abondance relatives à un grand nombre de molécules qu'il est difficile de gérer avec des outils conventionnels, du type tableur. L'utilisation des outils statistiques comme R est souvent indispensable.

L'association observée entre la maladie et les molécules peut être très spécifique aux individus sélectionnés. Pour éviter ce biais les études d'identification de biomarqueurs séparent souvent les individus en 2 groupes qui formeront la cohorte d'apprentissage et celle de validation. La cohorte d'apprentissage permettra d'identifier les biomarqueurs et de créer le modèle et les performances sont ensuite évaluées sur la cohorte de validation indépendante. Ce procédé permet de s'assurer de la généralisation de l'observation (Mischak *et al.*, 2010a; Moons *et al.*, 2012).

1.3.1 Identification statistique des biomarqueurs

Unes des étapes importantes de ces études est l'identification des biomarqueurs associés aux pathologies.

La Boize intègre les outils statistiques basiques pour identifier les molécules différentiellement abondantes entre les deux groupes (Figure II.1.2). Les tests de Wilcoxon et de Student sont principalement utilisés. Une des questions importantes à se poser dans le choix du test statistique à utiliser est la vérification des hypothèses de normalité et

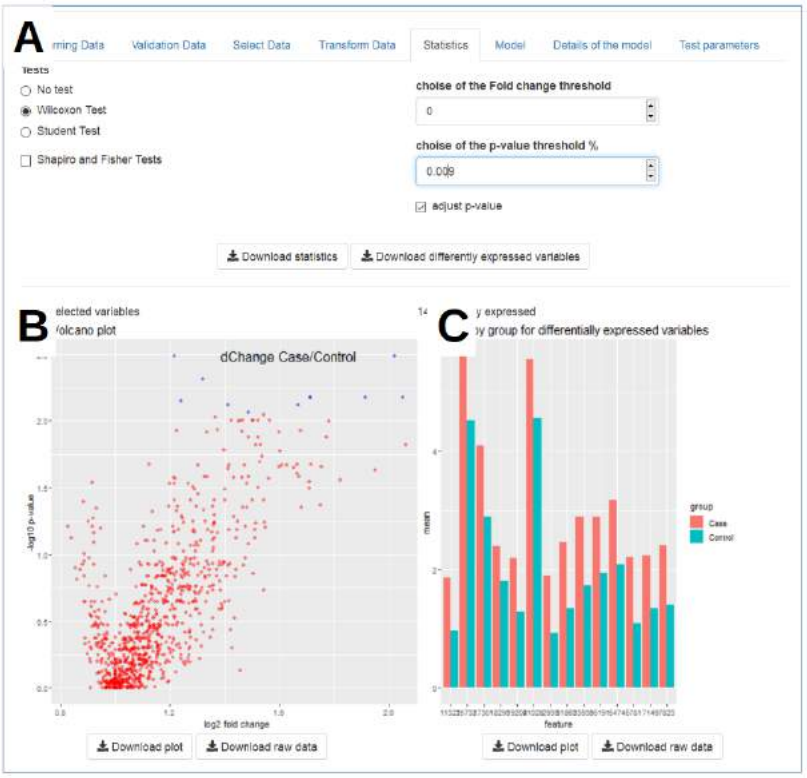

Figure II.1.2 – Identification statistique des biomarqueurs. *La Boize* intègre un grand nombre d'outils statistiques pour identifier les biomarqueurs. L'application permet à l'utilisateur, grâce à des formulaires simples (**A**), de choisir les paramètres adaptés à son analyse. La *volcano plot* (**B**) représente les potentiels biomarqueurs selon leur *p-value* et leur *foldchange*. Le diagramme en barre (**C**) représente l'expression moyenne par groupe (*controls* et *cases*) des biomarqueurs sélectionnés.

d'hétéroscédasticité. J'ai donc intégré le test de Shapiro et de Fisher pour vérifier les hypothèses statistiques. De plus la multiplicité des tests augmente le risque d'identifier des faux positifs. Dans ce cas il est nécessaire d'ajuster les résultats des tests pour prendre en compte ce biais. J'ai donc implémenté la correction pour des tests multiples de Benjamini Hochberg (Benjamini et Hochberg, 1995) qui est une méthode commune pour ajuster la *p-value*.

1.3.2 Construction d'un modèle de prédiction

Les biomarqueurs sont ensuite utilisés ensemble dans un modèle de prédiction, classant au mieux les individus en deux groupes : sains ou malades.

J'ai implémenté deux types de modèles connus dans ce type de problématique. Le modèle svm (*support vector machine*) est basé sur la recherche de l'hyperplan optimal séparant le plus possible les groupes (Meyer *et al.*, 2019). Les modèles de forêt aléatoire sont basés sur un grand nombre de tirages d'arbre de décision (Liaw et Wiener, 2002). La construction de ces modèles est accompagnée par des sorties graphiques décrivant les performances prédictives sur la cohorte d'apprentissage ou celle de validation (Figure II.1.3). L'application présente aussi certaines données pour aider dans le choix des paramètres

du modèle. Notamment dans le choix du seuil, qui peut être un enjeu majeur suivant l'application de l'étude.

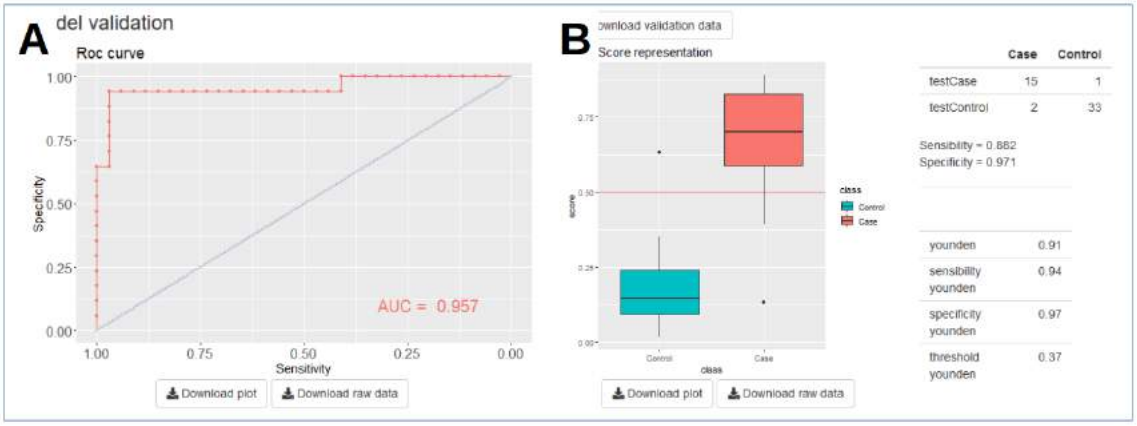

Figure II.1.3 – Performance de prédiction sur une cohorte de validation. La courbe ROC (**A**) est la manière la plus populaire de représenter les résultats d'un outil de classification. Il permet de calculer l'AUC (ici de 0,957) qui permet de comparer facilement les résultats de 2 méthodes concurrentes. Cette représentation n'est malheureusement pas très compréhensible. La représentation en boîtes à moustache (**B**) permet de mieux appréhender les résultats. Les résultats présentés ici ont été obtenus à partir d'analyse peptidomique du liquide amniotique (en cours de publication).

1.4 Application du modèle à de nouvelles données

L'objectif de la construction d'un modèle de prédiction est d'être ensuite utilisé sur des individus nouveaux servant ainsi d'outil clinique d'aide à la décision.

L'implémentation de *La Boize* a été construite avec cette perspective. Une fois construit le modèle peut être enregistré et utilisé dans un outil plus léger ne présentant que les résultats s'intéressant à une nouvelle cohorte (Figure II.1.4).

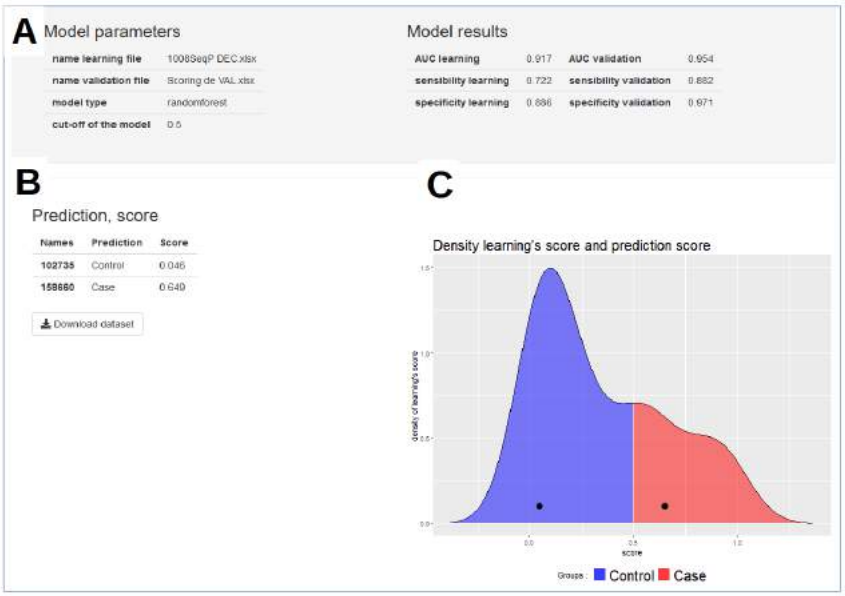

Figure II.1.4 – Utilisation de *La Boize* sur de nouveaux individus. La dernière partie de l'application est l'outil d'aide à la décision. Cette étape ne s'intéresse qu'à l'application du modèle sur des nouveaux individus. Le modèle précédemment construit (**A**) permet d'attribuer un score (entre 0 et 1) permettant de classer les nouveaux individus en fonction du seuil choisi (ici 0,5)(**B**). Au-dessous de ce score il est prédit comme sain (*control*). Au-dessus, le nouvel individu est prédit comme malade (*case*). Outre le classement de ces nouveaux individus, cette application présente les performances du modèle sur les individus utilisés pour construire le modèle. Le graphe présente la densité des scores des données d'apprentissage (**C**). Cela permet de mettre en perspectives les résultats obtenus sur les nouveaux individus (les points noirs).

Conclusion

Une des grandes problématiques des projets pluridisciplinaires est de faire rencontrer les connaissances et le savoir-faire de plusieurs disciplines. Sur des études comme celles-ci mêlant statistique et biologie, le challenge est de fournir des outils pertinent et opérationnel en biologie. C'est d'ailleurs une des problématiques majeures que j'ai soulevée dans mon précédent chapitre (I.1.3.5). Mon travail dans ce domaine, aboutissant à la création de La Boize, s'inscrit dans cette problématique comme une solution possible. Cet outil permet l'analyse des données omiques et la création de modèles de prédiction sans l'utilisation de langage de programmation.

Cette application a été la base de l'étude statistique des publications intégrées dans mon manuscrit de thèse dans les 2 chapitres suivants. La première décrit l'application d'une méthodologie robuste permettant d'identifier les métabolites urinaires associés à la sévérité des patients atteint d'obstruction de la jonction pyélo-uretérale (Boizard *et al.*, 2016). La deuxième étude se focalise sur l'analyse le peptidome du liquide amniotique pour prédire la fonction rénale post-natale des fœtus porteurs d'une anomalie du développement rénal (en cours de publication). Par ailleurs ce pipeline fait d'ores et déjà l'objet de tests cliniques par les cliniciens de l'équipe.

2 Analyse du métabolome urinaire de l'obstruction de la jonction pyélo-uretérale - Résultats

> "Connection between sweet tasting urine and diabetes was made as early as 600 BC, and writings of Hippocrates mention theuse of appearance, colour, and consistency of urine in diagnosis and prognosis. [...] Even today, with explosion of knowledge and availability of newer sophisticated techniques, urinalysis still remains the indispensable first step in the evaluation of renal disease."
>
> *K Abirami (2001)*

Le diagnostic précis des maladies rénales sans passer par des méthodes invasives comme la biopsie est un enjeu de santé important. Ainsi, une simple analyse d'urine grâce à laquelle nous pourrions diagnostiquer de façon précise une maladie rénale participerait à l'amélioration de la prise en charge des patients. La métabolomique, définie comme l'analyse des composés de faible poids moléculaire contenu dans un échantillon ($<$ 1500 Da), offre des avantages par rapport aux autres omiques. Les métabolites sont les molécules les plus proches du phénotype puisqu'ils intègrent l'information du génome, du transcriptome, du protéome et réagissent aux variations mineures comme l'alimentation, les médicaments ou les changements pendant le développement ou la progression de la maladie. L'objectif de ce projet est d'optimiser le pipeline bioinformatique et statistique en aval de l'électrophorèse capillaire couplé à la spectrométrie de masse pour l'analyse du métabolome urinaire comme un outil de diagnostic des maladies rénales. Pour la validation de cette nouvelle approche, nous nous sommes concentrés sur son utilisation dans l'identification des biomarqueurs métaboliques d'une anomalie rénale, l'obstruction de la jonction pyélo-uretérale, chez les nouveau-nés.

SCIENTIFIC REPORTS

OPEN

Received: 20 July 2016
Accepted: 14 September 2016
Published: 03 October 2016

A capillary electrophoresis coupled to mass spectrometry pipeline for long term comparable assessment of the urinary metabolome

Franck Boizard[1,2,*], Valérie Brunchault[1,2,*], Panagiotis Moulos[3], Benjamin Breuil[1,2], Julie Klein[1,2], Nadia Lounis[4], Cécile Caubet[1,2], Stéphanie Tellier[5], Jean-Loup Bascands[1,2], Stéphane Decramer[1,2,5], Joost P. Schanstra[1,2] & Bénédicte Buffin-Meyer[1,2]

Although capillary electrophoresis coupled to mass spectrometry (CE-MS) has potential application in the field of metabolite profiling, very few studies actually used CE-MS to identify clinically useful body fluid metabolites. Here we present an optimized CE-MS setup and analysis pipeline to reproducibly explore the metabolite content of urine. We show that the use of a beveled tip capillary improves the sensitivity of detection over a flat tip. We also present a novel normalization procedure based on the use of endogenous stable urinary metabolites identified in the combined metabolome of 75 different urine samples from healthy and diseased individuals. This method allows a highly reproducible comparison of the same sample analyzed nearly 130 times over a range of 4 years. To demonstrate the use of this pipeline in clinical research we compared the urinary metabolome of 34 newborns with ureteropelvic junction (UPJ) obstruction and 15 healthy newborns. We identified 32 features with differential urinary abundance. Combination of the 32 compounds in a SVM classifier predicted with 76% sensitivity and 86% specificity UPJ obstruction in a separate validation cohort of 24 individuals. Thus, this study demonstrates the feasibility to use CE-MS as a tool for the identification of clinically relevant urinary metabolites.

'Omics'-based strategies appear to be promising tools for the identification of diagnostic and prognostic biomarkers of disease. They can lead to the design of multimarker models which are potentially better suited than single biomarkers to describe complex pathophysiological mechanisms[1–3]. Metabolomics, defined as the analysis of the low-molecular-weight compound (<1500 Da) content of a sample, offers advantages compared to the other omics traits. Indeed, being the downstream products of cellular function, metabolites represent a sensitive measure of the actions of upstream molecular species such as genes, transcripts, and enzymes, including the effects of disease, drugs, toxicity, and the environment[4,5]. However sensitivity to these many perturbants also contributes to potential issues about the high variability in metabolome exploration[6].

Analysis of urine plays a central role in clinical diagnostics as it can be collected non-invasively, often in large quantities, and requires minimal sample pre-treatment due to its low complexity and protein content. In addition, we and others have already shown that urine is an excellent reservoir of biomarkers (peptides, proteins and metabolites) of many diseases[7–17].

Metabolomics studies mostly use NMR spectroscopy and liquid chromatography coupled to mass spectrometry (LC-MS) that provide complementary readouts[4]. NMR spectroscopy allows both identification and quantification of metabolites. It is a highly reproducible and non-destructive method which requires minimal sample preparation thereby minimizing contamination and maintenance issues and enabling the routine and

[1]Institut National de la Santé et de la Recherche Médicale (INSERM), U1048, Institut of Cardiovascular and Metabolic Disease, Equipe 12, 1 avenue Jean Poulhès, BP 84225, 31432 Toulouse Cedex 4, France. [2]Université Toulouse III Paul-Sabatier Toulouse, France. [3]HybridStat Predictive Analytics, Athens, Greece. [4]Unité de Recherche Clinique Pédiatrique, Module Plurithématique Pédiatrique, Centre d'Investigation Clinique - Hôpital des Enfants, Toulouse, France. [5]CHU Toulouse, Hôpital des Enfants, Service de Néphrologie – Médecine Interne – Hypertension Pédiatrique, Toulouse, France. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to J.P.S. (email: joost-peter.schanstra@inserm.fr) or B.B.-M. (email: benedicte.buffin-meyer@inserm.fr)

high-throughput analysis of hundreds to thousands of samples[4,18]. The inherent low sensitivity of NMR, however, restricts the detection limit to about 1 μM[18,19]. Moreover, the interpretation of NMR data is challenging[18,20]. In contrast, LC-MS allows the detection, quantification and structure elucidation of metabolites in the picomolar to nanomolar range of several thousand metabolites in a single measurement[5]. Unfortunately, the coupling of chromatographic separations with MS platforms requires an elevated level of maintenance, as the samples come in direct contact with many components of these platforms, contaminate surfaces and cause drift in the measured response and retention time over relatively short analysis periods[4,5], thereby preventing the comparison of large numbers of samples. Relevant progress in the field of LC-MS was made with the introduction of ultra high performance liquid chromatography (UPLC) leading to improvement of analysis speed as well as sensitivity and resolution[19,21,22]. In particular, the potential of miniaturized UPLC-MS, based on the optimized use of microbore columns, was recently demonstrated for large-scale metabolomic studies[23,24].

Until approximately ten years ago, capillary electrophoresis coupled to mass spectrometry (CE-MS) has only been rarely used for metabolome analysis. This was potentially due to issues related to stable coupling of CE to the MS instrument and the limited loading capacity of CE capillaries. However the significantly increased sensitivity of modern mass spectrometers and optimized methods for coupling of CE to MS have transformed CE-MS into a potential appropriate tool for profiling of disease associated metabolites in clinical relevant body fluid samples[20,25–29]. A number of recent studies now report the use of CE-MS for metabolome analysis of clinically relevant samples, with in particular those recently conducted by Soga and coworkers[30–32], the first group to develop CE-MS for the comprehensive profiling of metabolites in biological samples[33]. However, the use of CE-MS for the discovery and validation of clinically relevant metabolic markers of human disease requires evaluation of its performance in terms of long term reproducibility and comparability.

Here, we present an optimized CE-MS setup and data analysis pipeline. Using a normalization procedure based on a set of "housekeeping" metabolites, this method allows to compare the metabolite content in urine samples analyzed over a period of several years. As proof of concept, we demonstrate the clinical relevance of this pipeline for the urinary metabolome based-detection of obstructive nephropathy in infants.

Results

Identification of metabolite internal standards for CE-MS normalization. As a first measure towards improved comparison of large numbers of clinical samples over time, we developed a new method that allows to normalize the metabolite content of a biofluid sample. This method is based on the use of a set of persistent and stable metabolites across disease and healthy urine samples. In order to identify these so-called stable endogenous metabolites, 54 CE-MS runs of urine obtained from various kidney and urinary tract pathologies together with 21 control CE-MS runs of urine from healthy patients (Supplementary Table S1) were processed using the Bioconductor package xcms[34]. Each metabolite feature was identified by a unique identifier (ID) on the basis of the specific mass-to-charge ratio and migration time with a peak height representing the relative abundance. After preprocessing of the mass spectra (including mass calibration and migration time window restriction), the xcms pipeline (see Materials and Methods) identified 9642 distinct molecule features in terms of m/z and migration time pairs across all 75 samples. From this initial list, only features present (no-null abundance) in at least 50% of the total samples were considered for further analysis. The 6044 remaining metabolite features spanned a CE migration time from 16 to 50 min and a m/z range from 30–650. This reference dataset of 6044 metabolite features was then interrogated for the presence of stable molecule features, in terms of intensity, that would comprise the basis for a set of CE-MS internal normalization standards. For this, several established algorithms from the 'rank invariant' family of normalization methods present in the DNA microarray literature were deployed. Specifically, the Rank Invariant normalization method implemented in the dChip algorithm[35], the Rank Invariant normalization algorithms for Illumina BeadArrays implemented in the lumi Bioconductor package[36] and the GRSN algorithm[37] were tested. However, each one of these suffered from several drawbacks, including among others unstable housekeeping sets because of their selection algorithm (dChip), selection preference in higher (dChip), lower (lumi) or medium (GRSN) intensities instead of spanning the whole metabolite abundance range, very high number of metabolites to achieve proper normalization (lumi) or poor normalization efficiency (dChip). The failure of present methodologies (partially due to the different nature of CE-MS data as compared to microarrays) to detect a stable set of metabolites led to the development of two new different internal standard selection strategies. Specifically, the first approach used the residuals of Robust Linear Regression models[38,39] to identify sets of metabolites presenting low variability across samples and the second, more geometrical than statistical, approach was based on the Euclidean distance of each metabolite abundance vector from the identity 'hyperline' in the sample space. The final set of stable metabolites for each method was derived using a Forward Selection procedure with the purpose of finding the smallest possible subset of metabolites with the greater normalization power (detailed description of the methods in the 'Materials and Methods' section). The method that was finally followed was the geometrical approach as it was found to yield more robust results in terms of metabolite intensity coverage, normalization power, smaller number of stable metabolites and its application did not require any assumptions for a baseline as compared to the RLM approach which requires a baseline. This led to the identification of 267 endogenous housekeeping metabolic features among the 6044 features detected (Supplementary Table S2) which spanned a CE migration time from 17 to 36 min and a m/z range from 82 to 650. These stable endogenous metabolite features were implemented in the CE-MS normalization pipeline. Hence, the CE migration time is normalized in a first step (Fig. 1A) followed by normalization of the metabolite abundance using the endogenous housekeeping metabolic features, as exemplified on a random selection of six samples (Fig. 1B).

Use of a beveled capillary improves the sensitivity of metabolite detection. CE coupling to MS via electrospray ionization (ESI) can be performed using either a sheathless or a sheath flow interface[40]. The use

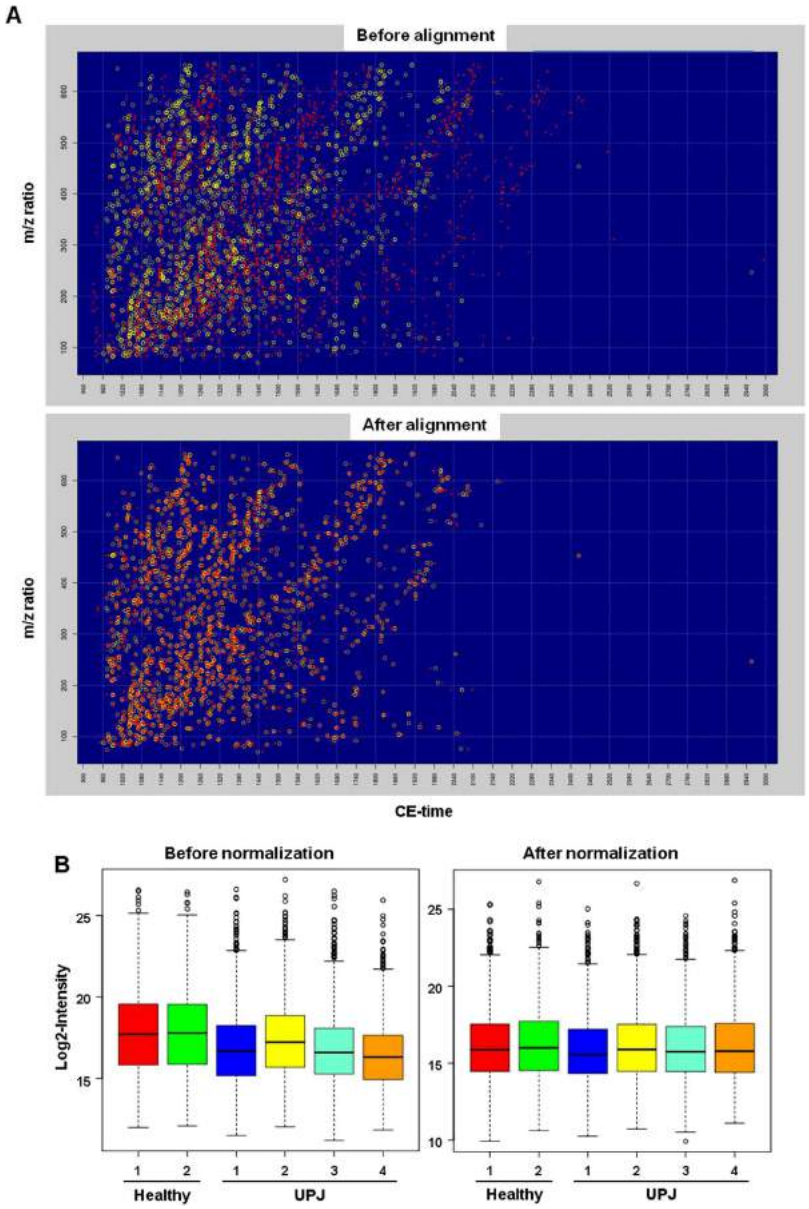

Figure 1. Processing and normalization of samples. Urinary samples were analyzed in CE-MS, processed and then normalized using the stable endogenous metabolites-based procedure described in the Materials and Methods section. (**A**) Representative distribution profile of urinary metabolite features before and after migration time alignment against reference dataset. Each circle is a unique peak processed with xcms. Red: metabolite features detected in a random urine sample and matching the reference; yellow: equivalent features in the reference dataset. (**B**) Box-whisker plot for metabolite abundance of exemplary healthy (2) and UPJ obstruction (4) patients before and after intensity normalization.

of sheathless systems is promising. In particular, the potential usefulness of a sheathless porous tip interface for CE-MS has been recently demonstrated for the analysis of the urinary metabolome[28,29]. Nevertheless this porous tip has not yet been adopted as a routine method for CE-MS coupling. So far, the sheath flow interface has been most widely used for CE-MS in metabolomics[26,40,41]. This type of coupling is stable and provides good sensitivity, its implementation is relatively easy and allows using a wide range of buffers. However, the CE-effluent is diluted in this configuration, thereby reducing the achievable sensitivity of the method[28,29,40,41]. As part of a continuous effort to improve the interface between CE and MS, Tseng *et al.*[41] have developed a beveled tapered tip emitter in order to reduce the sheath flow leading to decreased sample dilution. By analyzing synthetic drugs and triazine mixtures, they demonstrated that the use of beveled tip provides better sensitivity for detection than conventional sheath liquid interface which uses flat capillary tips[41].

Therefore in an attempt to optimize the sensitivity of the detection of urinary metabolites, we compared the performance of a standard flat tip and a beveled tip sheath-liquid ESI interface. A QC urine sample was analyzed by CE-MS using either a standard (ten consecutive runs) or beveled capillary (ten consecutive runs) for CE. Of

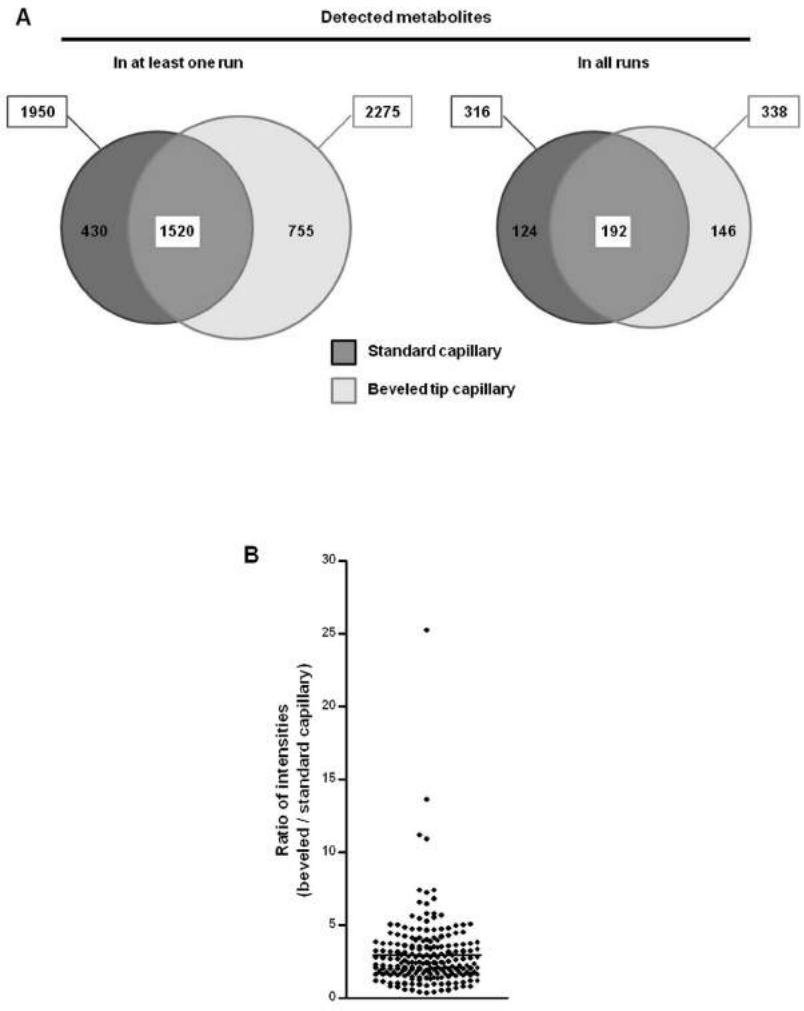

Figure 2. Effect of the capillary on sensitivity of metabolite detection. The same sample was analyzed in CE-MS using either standard (10 times) or beveled tip capillary (10 times) for CE. (**A**) Euler diagrams showing for each capillary the number of metabolite features detected at least once (left) or every time (right). Dark gray: standard (flat tip) capillary; light gray: beveled tip capillary. (**B**) For each metabolite detected in every run and with both types of capillaries (n = 192), the mean intensity was calculated and then the ratio between intensity measured with beveled tip capillary and intensity measured with classical capillary was calculated. Graph shows the mean ratio ± SEM, indicating that metabolite detection was more sensitive with beveled tip than with standard capillary.

note, the previously described 267 stable endogenous metabolites required for normalization procedure were identified using a beveled tip. After normalization, 2275 and 1950 distinct molecule features were detected in at least one run using the beveled tip and standard flat tip, respectively (Fig. 2A). Moreover, 338 and 316 metabolite features were detected consistently in all ten runs using the beveled tip and conventional tip, respectively (Fig. 2A). Although the absolute number of features detected is only slightly higher using the beveled tip, comparison of the intensities of 192 features detected in all runs with both types of capillary revealed a significant 3 fold gain in sensitivity using the modified capillary (Fig. 2B). Of note, robustness of the beveled tip was not decreased compared to flat tip (resisting to 40–50 runs [data not shown]). Therefore, the use of beveled tip as sheath-flow interface for CE-MS displays increased sensitivity towards the detection of urinary metabolites. We used the beveled tip for the remainder of the experiments.

QC-based validation of CE-MS pipeline for urine metabolome profiling. In order to estimate the analytical variability of the CE-MS pipeline, a set of experiments for validation was performed: repeatability (intra-assay precision), postpreparation stability, postdilution stability, and long-term (intermediate) precision were evaluated.

Repeatability expresses the precision under the same operating conditions over a short interval of time. Repeatability of the CE-MS pipeline was examined by analyzing the QC urine sample in five consecutive runs, covering a total run time of ≈8 h. Among 6044 potential metabolites, 1342 (22%) features were detected on average in each run. Figure 3A shows a typical plot of a CE-MS analysis of a QC sample, giving an indication of the distribution of mass-to-charge ratio and CE migration times encountered for this typical sample. To obtain

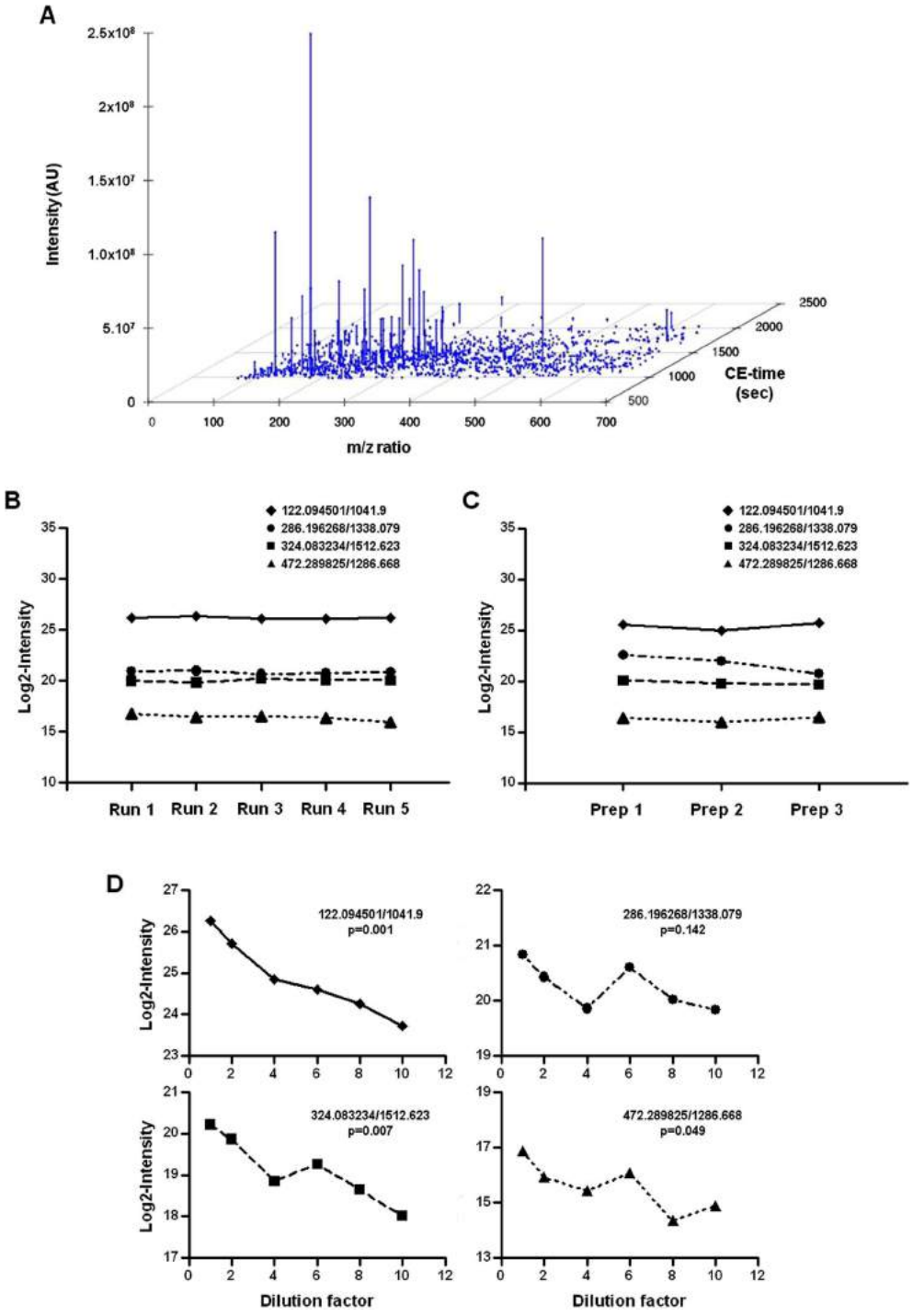

Figure 3. Short term performance characteristics of metabolomic CE-MS platform. The data from QC analyses were investigated to assess intra-assay precision, postpreparation stability and postdilution stability for molecule intensities. (**A**) Typical plot from the CE-MS analysis of the QC sample: Each metabolite was identified by a unique identifier (ID) on the basis of the specific mass-to-charge ratio and migration time. Graph shows the distribution of metabolite mass-to-charge ratio (*m*/*z*) with CE-migration time for a representative QC injection. (**B**) Short term precision: The QC was analyzed in five consecutive runs and the intensity in each run was shown for four exemplary randomly selected metabolite features. The coefficient of variance (CV) for amplitude was between 0.7 and 1.9% for these individual features, thereby demonstrating the repeatability in peak height. (**C**) Variability according to preparation: QC sample was prepared on three different dates using different lots of buffer, and then analyzed in consecutive runs. The intensity in each run was shown for four exemplary randomly selected metabolite features. The obtained CV for abundance was between 1.1 and 4.3%, showing a stability depending of the preparation. (**D**) Stability according to dilution: QC sample was prepared at different concentrations and then analyzed in consecutive runs. The intensity of four exemplary randomly selected metabolite features was plotted against the dilution factor.

information on the run-to-run precision, four metabolite features were randomly selected for evaluation of intensity variation. The abundance variation of these four metabolite features was found to be negligible (Fig. 3B), with coefficient of variation (CV) values less than 2%, thereby indicating high performance of CE-MS platform in terms of repeatability. Next, the effect of different sample preparations was studied (post-preparation stability). We prepared QC sample according to the same procedure but using three different lots of buffer before CE-MS analysis in 3 consecutive runs. As shown in Fig. 3C, the intensity of the four exemplary selected metabolite features was constant in preparations, with a low CV, below 4.3%. Third, in order to test linearity of detection, the QC sample was prepared at six different concentrations and then analyzed by CE-MS in consecutive runs. Figure 3D depicts the abundance of the four randomly selected molecule features as a function of the dilution factor of a urine sample. For three of them, a significant negative correlation was observed between dilution and abundance whereas only a trend was observed for the fourth (Fig. 3D), thereby suggesting the relative stability of CE-MS platform when urine samples are diluted.

Finally, we evaluated intermediate precision of CE-MS platform which expresses the precision within laboratory variations. This assay involved analysis of QC urine metabolites at different days by different operators over a long period of time. It included different lot numbers of buffers, solvents and chemicals and also implies annual maintenance service of both CE and MS devices. This evaluation is important in the field of clinically useful metabolite biomarkers where durable use of CE-MS is necessary. For the long-term stability assay, the QC sample was analyzed repeatedly 128 times over a range of 4 years (from 2011 to 2014). Among 6044 potential metabolite features, 1389 (23%) were detected on average in each run, this result being similar to the previously reported value. A mean of 67.7% of all metabolite features and 30.5% of the stable endogenous metabolites in the QC samples from these 128 runs matched against the reference dataset. The analysis of our data set revealed that the distribution of intensities is bimodal, with a strong proportion of values at a point-mass at zero (*point-of-mass values* [PMVs] corresponding to missing values [NaN], zero intensity data being treated as missing data) and a continuous component (Fig. 4A). The occurrence of zero component in the data matrix is a recurrent issue encountered in MS data[42]. The origin of PMVs may either be biological, eg absence of a specific metabolite in biological sample, or technical, eg the inability of the mass spectrometer to detect the specific metabolite or of the algorithm to identify the peak. Next, as it is recommended that the coefficient of variation should not exceed 15%[4,43], we examined CE-MS results using similar acceptance criteria as a means of determining the quality of the data. For this, the abundance of four exemplary randomly chosen molecule features was plotted over time (Fig. 4B). The statistical spread for these metabolite features was between 2.2 and 8.6%, indicating that CE-MS platform exhibits long-term stability. In addition, CV of intensities was calculated for all metabolite features across the QC samples. A data subset was considered including features which were detected in at least one of the 128 QC injections (4879 entities) and different filters of selected metabolites were considered to evaluate improvement of the proportion of peaks being acceptable. Using this subset, we observed that 4487 (92%) of the 4879 molecule features displayed a variation of $\leq$10%, whilst 2892 (59%) exhibited a variation of $\leq$5% level. Altogether, these results demonstrated the long-term stability of CE-MS platform and thus suggest that the optimized CE-MS setup and analysis pipeline allows to compare the metabolite content in urine samples regardless of the time of analysis.

CE-MS for clinical metabolomics: application to diagnosis of UPJ obstruction. Next we analyzed the capacity of the aforementioned pipeline in clinical research for the identification of diagnostic/prognostic biomarkers of disease. Newborns with UPJ obstruction were chosen for our proof of principle study. Two different cohorts of infants were employed: one discovery cohort (n = 49) for the identification of urinary metabolite biomarkers of UPJ obstruction (15 healthy newborns and 34 patients with UPJ obstruction; Table 1 and Supplementary Table S3) and one cohort (n = 24) for the blinded validation of urinary biomarkers (7 healthy newborns and 17 patients with UPJ obstruction; Table 2 and Supplementary Table S4). All urine samples were analyzed by CE-MS for their metabolite content and normalized using the above developed stable endogenous metabolites-based normalization procedure.

Metabolic profiling of urine samples from patients with UPJ obstruction and healthy children. The urinary metabolome of the discovery cohort, composed of 15 healthy children and 34 patients with severe UPJ obstruction (Table 1 and Supplementary Table S3) was studied by CE-MS. A mean of 42.0% of the stable endogenous metabolites in urine samples matched against the reference dataset. Among 6044 potential metabolite features, 1889 (31%) were detected on average in each sample. Only the features detected in at least 75% of the urine samples in each group (healthy and UPJ) were further investigated. This noise-filtering process reduced the number of features to 388 entities (Fig. 5A). The distribution of the metabolite intensities for all the 388 selected metabolite features showed, as for QC sample data, a bimodal distribution characterized by a proportion of PMVs (Fig. 5B) and a continuous component. In order to explore the origin of PMVs, metabolite features with consonant or dissonant differences were quantified. In the former case, the group with the higher proportion of PMVs has the smaller mean in the continuous part, while in the latter case the group with the higher PMV proportion also has the higher mean. An example of each type is shown in Fig. 5C. Although this definition does not distinguish between technical and biological PMVs, technical PMVs naturally correspond to consonant compounds whereas biological PMVs generally allow for both types[44,45]. The data employed here contains 357 (92%) consonant compounds, 15 (4%) dissonant and 16 (4%) without point-mass component. The high proportion of consonant markers associated with the low number of dissonant markers suggests that PMVs in present metabolomics data originated from technical considerations rather than biological (Fig. 5D).

Identification of urinary metabolites associated to UPJ obstruction. Comparing urinary metabolites from UPJ and healthy patients led to the identification of 32 adjusted (Benjamini and Hochberg[46]) differentially excreted

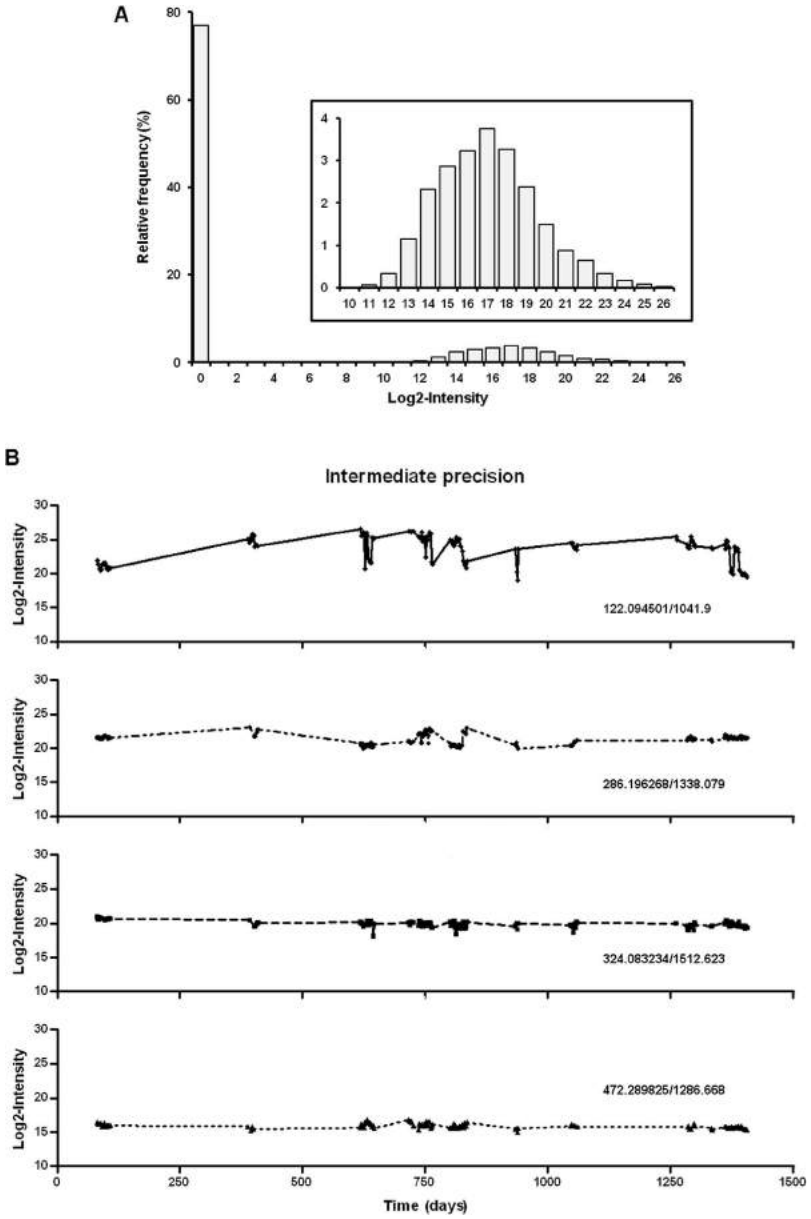

Figure 4. Long term performance characteristics of metabolomic CE-MS platform. The data from QC analyses were investigated to evaluate intermediate precision for molecule intensities. (**A**) Histograms of the distribution of abundance: The mean frequency of all features in QC sample was plotted against the logarithm (2) of the intensity. Profiles show a point-mass at zero and a continuous component. The zero component arises because the molecule features are either absent or their concentration is below the detection limit. Insert: magnification of the continuous distribution. (**B**) Long term variability: The QC sample was analyzed 128 times between 2011 and 2014. The intensity of four exemplary randomly selected metabolite features was plotted against the time.

metabolite features (Fig. 6A,B and Supplementary Table S5). Matching 32 features against databases (HMDB, ChEBI and KEGG) led to determination of real mass for 9 metabolite features; 5 of 9 were annotated for chemical formulas (Table 3). Of note, abundances of two compounds (227.111791/989.758 and 228.114334/990.108) corresponding to the same annotation were highly correlated ($R^2 = 0.94$, $p < 0.0001$, data not shown). The 32 metabolite features of interest were then used to develop a support vector machine (SVM) discrimination model that we called "UPJMetab32". Scoring the patients from the discovery cohort with the UPJMetab32 classifier clearly separated UPJ from healthy patients (Fig. 6C).

Validation of UPJMeta32 in a separate, blinded cohort. In the next step, following the recommendations for biomarker identification[47], the UPJMetab32 model was validated in a separate, blinded study using urine from 7 healthy and 17 UPJ patients not used in the discovery cohort (Table 2 and Supplementary Table S4). These urine samples were analyzed by CE-MS and scored using the UPJMetab32 model (Supplementary Table S4).

	All patients	Healthy	UPJ obstruction
n	49	15	34
Gender			
M	46 (93.9%)	15 (100%)	31 (91.2%)
F	3 (6.1%)		3 (8.8%)
Age			
Mean (months)	2.25 +/− 0.27	2.29 +/− 0.62	2.22 +/− 0.29
Median (months)	1.45 (range 0 to 7.0)	1.58 (range 0 to 6.1)	1.45 (range 0.7 to 7.0)

Table 1. Discovery cohort.

	All patients	Healthy	UPJ obstruction
n	24	7	17
Gender			
M	21 (87.5%)	7 (100%)	14 (82.4%)
F	3 (12.5%)	0	3 (17.6%)
Age			
Mean (months)	2.51 +/− 0.58	1.35 +/− 1.21	2.99 +/− 0.65
Median (months)	1.28 (range 0 to 8.6)	0.03 (range 0 to 8.6)	1.61 (range 0.8 to 8.6)

Table 2. Validation cohort.

A UPJMetab32 score >0 predicts patients with UPJ obstruction. These predictions were compared to the clinical criteria based status. The UPJMetab32 classifier diagnosed clinical status (healthy versus UPJ) with a sensitivity of 76.5%, a specificity of 85.7%, and an area under the curve (AUC) of 0.90 [95% CI: 0.707 to 0.984] (Fig. 7A). The UPJMetab32 model predicted 13 out of 17 UPJ cases correctly, showing the efficacy of the model to detect patients with severe UPJ. In addition, it predicted 6 out of 7 control cases correctly. The distribution of the UPJMetab32 scores for the validation cohort showed significant separation of the two patient populations (Fig. 7B).

Discussion

We have explored the use of CE-MS and endogenous stable urinary metabolites for long-term, reproducible and comparable analysis of the urinary metabolome. The developed pipeline allowed comparison of urinary metabolite content analyzed over a 4 year timespan. As proof-of-concept we have used this pipeline to discover and validate urinary metabolites associated to a frequently encountered renal pathology in newborns.

Clinical metabolomics aims at the detection of clinically useful metabolites that can be extracted from a diverse range of sample types. Amongst those samples, easily accessible bodyfluids like urine and blood are most suited for clinical use. Although the field of metabolomics has advanced significantly in the past 10 years[4], there has been little progress in the identification of clinically useful urinary metabolite biomarkers. To enable the discovery and the validation of diagnostic/predictive biomarkers, medium-to-large-scale epidemiological studies are required in order to take into account the substantial diversity observed in physiology/physiopathology, metabolic status and lifestyle in the general human population. This involves the use of analytical methods able to analyze large numbers of samples over periods of many months or years with both high reproducibility and high sensitivity[4]. We explored the potential of CE which offers multiple advantages: (i) as CE separates compounds on the basis of their charge and size[25], it demonstrates high-resolution power for separation of small ionogenic metabolites which are important constituents of the urinary metabolome; (ii) CE separations require a low sample volume and consume very little solvent[25], thereby reducing the matrix effect that can cause ion suppression and then insufficient ionization and lower peak intensity in MS; (iii) CE displays high reproducibility when analyzing large numbers of samples since no gradients are applied. Indeed, we observed high stability of urinary metabolite abundance when analyzing the same sample nearly 130 times over a range of 4 years. A few studies report stability evaluation of pipelines, such as for example over 535 runs covering a timespan of 5 months (GC-TOF-MS[48]) or over 120 runs covering a timespan of 3 years (UPLC-TOF-MS[49]). However, such a long term assessment of reproducibility and comparability is only rarely performed. Hence our 4 years proof of stability of the developed pipeline, associated with its use in the UPJ obstruction, validates its potential use in the clinic field.

Establishing long term stability has therefore been a major objective of the study. Although CE-MS is a reproducible analytical tool, some variations induced by sample concentration (especially for urine where individual urine outputs are dependent of water uptake, diet, …), interfering compounds and injection volume differences might still be observed. Several normalization strategies, such as normalization to creatinine, osmolarity and total area normalization are frequently employed in urine metabolomics studies. However, these commonly used normalization methods are not well adapted. For example, the creatinine level can be impacted by factors such as kidney function impairment, gender difference, and lean body mass[50,51]. The osmolarity normalization procedure is often affected by insoluble components, such as urine particles[50,52]. Adjusting the total peak area might yield biased results since the background noise and ion suppression due to the matrix may greatly interfere with the

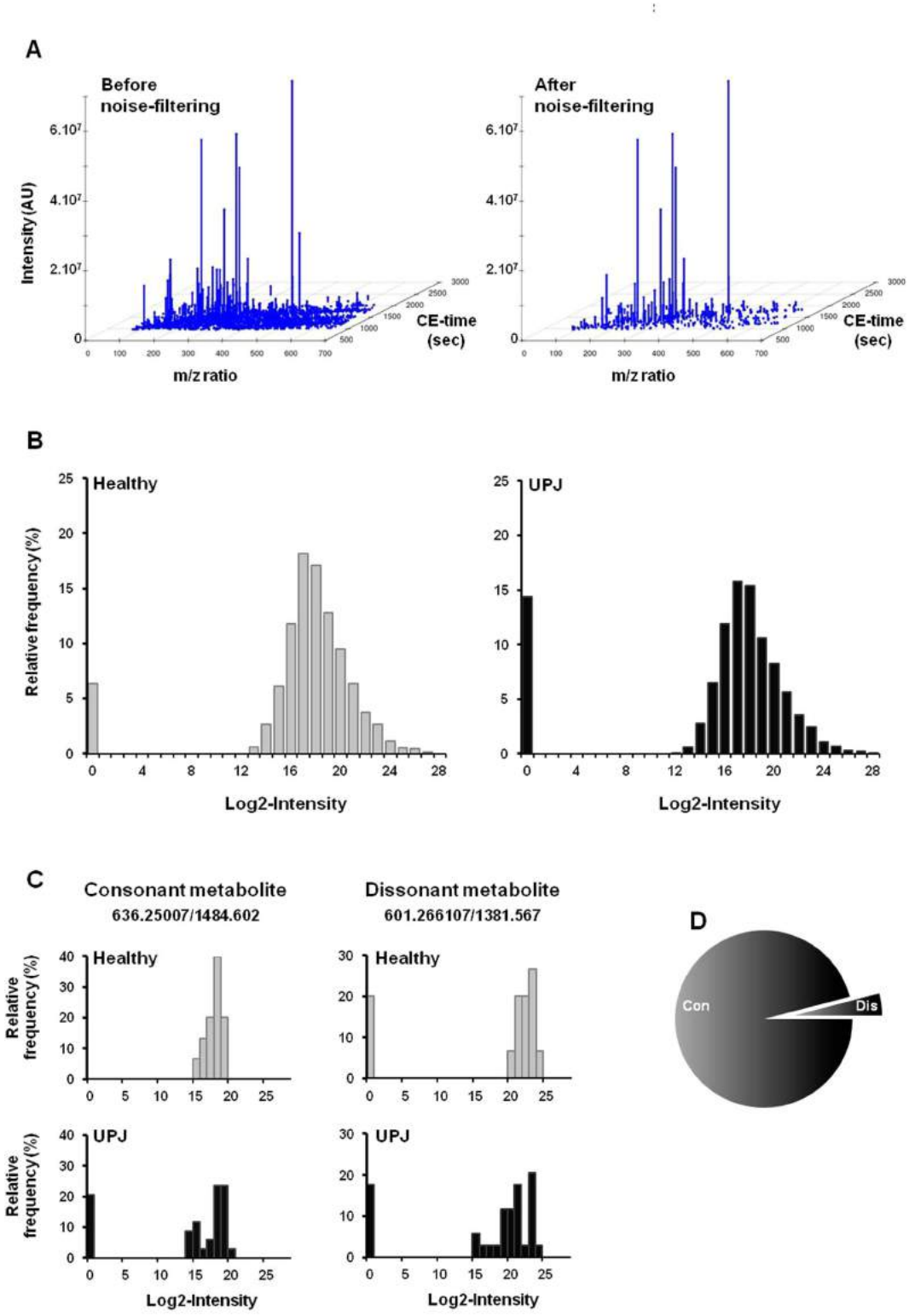

Figure 5. Metabolomic CE-MS analysis in urine of patients with UPJ. The urine metabolome of 15 healthy and 34 UPJ patients of the discovery cohort was analyzed. (**A**) Representative figure showing abundance of the CE-MS detected-urinary metabolite features: on the left, before application of a filter; on the right: after selection of features present in at least 75% of the samples in each group. (**B**) Histograms of distribution: The frequency of all metabolite features in healthy and UPJ samples was plotted against the logarithm (2) of the intensity. As for QC sample data, profiles show a point-mass at zero and a continuous component. (**C**) Histograms of distribution of two selected metabolite features from example dataset. Metabolite feature ID: 636.25007/1484.602 (left): consonant; metabolite feature ID: 601.266107/1381.567 (right): dissonant. (**D**) Repartition of compounds with consonant and dissonant differences between healthy and UPJ obstruction groups.

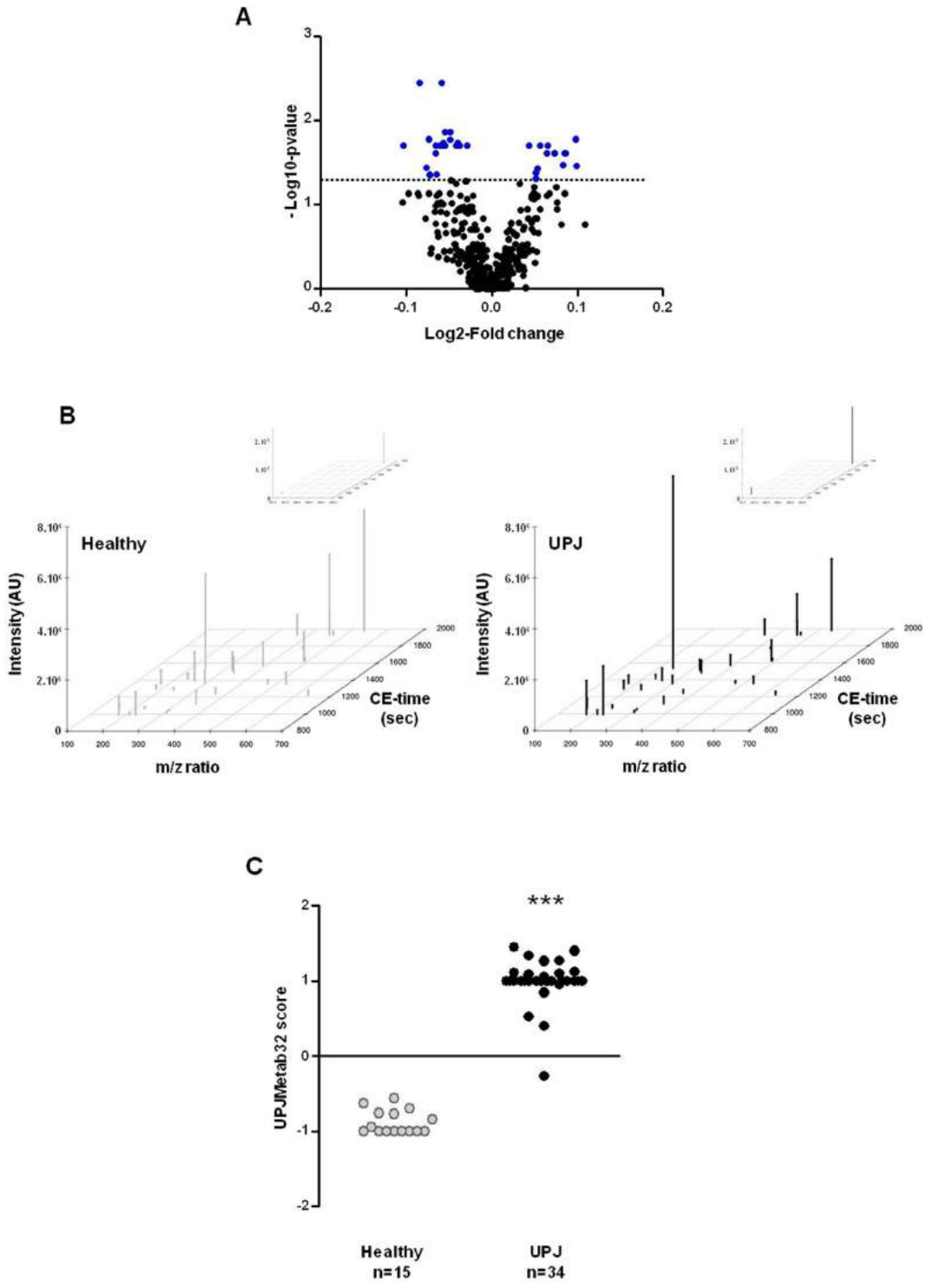

Figure 6. Identification of a classifier: UPJMeta32. The urine metabolome of 15 healthy and 34 UPJ patients (discovery cohort) was analyzed. (**A**) Volcano plot showing fold-changes (Log2) between UPJ obstruction and healthy groups as well as statistical significance (-Log 10 of p-value) for 388 considered metabolite features. The dashed line shows where $p = 0.05$. Points above the line had $p < 0.05$ and corresponding metabolite features (32) have been considered as significantly differentially excreted by UPJ patients. (**B**) Compared abundance of the 32 urine metabolite features which were identified as differentially excreted between UPJ patients and healthy subjects in the discovery cohort. Insert: two strongly abundant metabolite features (**C**) Cross-validation score of an SVM metabolite model, called UPJMetab32, consisting of 32 differentially excreted metabolite features. ***$p < 0.0001$ versus healthy subjects. Mann-Whitney test for independent samples.

total signal. Furthermore, the total signal for samples with different metabolite distributions does not reflect the total concentration differences as ionization efficiency is compound dependent[50]. Variations can also be corrected

ID	Isotope	Adduct	Real Mass	Database	Database code	Proposed Formula	Proposed Name
227.111791/989.758	[56][M]+	[M+H]+	226.104515	HMDB	HMDB00033	C9H14N4O3	Carnosine
228.114334/990.108	[56][M+1]+			KEGG	cpd:C00386	C9H14N4O3	Carnosine
				HMDB	HMDB12482	C9H14N4O3	Hydroxypterin
				HMDB	HMDB00245	C10H14N2O4	Porphobilinogen
				KEGG	cpd:C00931	C10H14N2O4	Porphobilinogen
				KEGG	cpd:C02345	C15H14O2	(2S)-Flavan-4-ol
				KEGG	cpd:C15598	C15H14O2	Favan-3-ol
				KEGG	cpd:C09757	C15H14O2	7-Hydroxyflavan
				KEGG	cpd:C10276	C15H14O2	Pinosylvinmethylether
				KEGG	cpd:C10325	C15H14O2	Deoxylapachol
				KEGG	cpd:C13632	C15H14O2	4,4′-Dihydroxy-alpha-methylstilbene
				KEGG	cpd:C07205	C14H14N2O	Metyrapone
				ChEBI	55316	C7H16BrNO2	Acetylcholine bromide
				ChEBI	50426	H4O6P2S2	Disulfanediylbis(phosphonic acid)
229.117309/1322.695	[8][M]+	[M+H]+	228.110033	HMDB	HMDB06695	C10H16N2O4	Prolylhydroxyproline
				KEGG	cpd:C13733	C10H16N2O4	(S)-ATPA
				KEGG	cpd:C10371	C15H16O2	MansononeC
				KEGG	cpd:C13624	C15H16O2	BisphenolA
				KEGG	cpd:C15210	C15H16O2	1,1-Bis(4-hydroxyphenyl)propane
				KEGG	cpd:C17424	C15H16O2	Lindenenone
				ChEBI	58089	C5H11NO7P	5-phosphonato-D-ribosylaminium(1−)
				ChEBI	58681	C5H11NO7P	5-phospho-β-D-ribosylaminium(1−)
				KEGG	cpd:C18436	C9H16N4OS	Tebuthiuron
				ChEBI	53648	C7H4N2O7	2-hydroxy-3,5-dinitrobenzoic acid
355.071351/1117.064	[306][M]+		354.064075	KEGG	cpd:C01268	C9H15N4O9P	5-Amino-6-(5′-phosphoribosylamino)uracil
				KEGG	cpd:C02927	C15H14O10	2-Caffeoylisocitrate
				KEGG	cpd:C07952	C17H19ClN2S. HCl	Chloropromazinemonohydrochloride
				KEGG	cpd:C12600	C19H14O5S	Phenolsulfonphthalein
488.133087/1622.375	[453][M]+		487.125811	KEGG	cpd:C02555	C26H21N3O5S	Luciferylsulfate
				KEGG	cpd:C18429	C18H22FN5O8S	Flucetosulfuron
366.599792/1929.853	[461][M]+		365.592516				
438.677677/1763.369	[443][M]+		437.670401				
474.701959/1359.882	[359][M]+		473.694683				
526.161131/1357.627	[160][M]+		525.153855				

Table 3. Annotation for potential chemical formulas and names. HMDB: Human Metabolome Data Base; KEGG: Kyoto Encyclopedia of Genes and Genomes, ChEBI: Chemical Entities of Biological Interest.

by addition of exogenous standards but this method assumes that those are representative of the thousands of injected metabolites[53]. In the present study, we have opted for the selection of a set of most stable endogenous metabolites observed in a range of samples. This method offers several advantages. Firstly, for the selection of these stable endogenous compounds, we have chosen 75 urine samples potentially representing the diversity of (pediatric) diseases to be encountered in future studies. Therefore, we anticipate that the 267 derived stable endogenous metabolites can be used for the discovery of metabolite-based biomarkers in a number of pediatric diseases of the kidney and the urinary tract. Secondly, such a high number of stable endogenous metabolites for normalization spanning a CE migration time from 17 to 36 min and a m/z range from 82–650 allows that signal normalization can be performed 'locally', using metabolites with comparable ionization efficiency since close in terms of CE migration time and m/z ratio. In addition, this inclusion ensures that for every new sample, there will be a sufficient number of endogenous internal standards so as to span the whole intensity range of the new sample. Thirdly, as a result of this high number of stable endogenous metabolites, we observed that significant numbers of metabolite features are available for robust normalization in nearly all cases (we identified a mean of 42.0% of stable endogenous metabolites in the UPJ experiments). A potential drawback of the use of these endogenous metabolites for normalization could be that those are stable in the specific case of kidney disease and are excluded for the selection of biomarkers of disease. Selection of novel endogenous stable metabolites might thus be required in order to discover biomarkers for disease affecting other organs than the kidney/urinary tract.

Analysis of urinary metabolome is extremely attractive since changes reflect modifications of the entire organism in its equilibrium with the environment including particularly contributions from nutritive substances, drugs and gut microbial activities[19]. However, the variability induced by these factors can introduce a day-to-day intrapersonal variability as well as interpersonal differences, being a major drawback in studies aiming at disease diagnosis/prognosis. In order to address the sources of urinary metabolome variation throughout the day, Kim

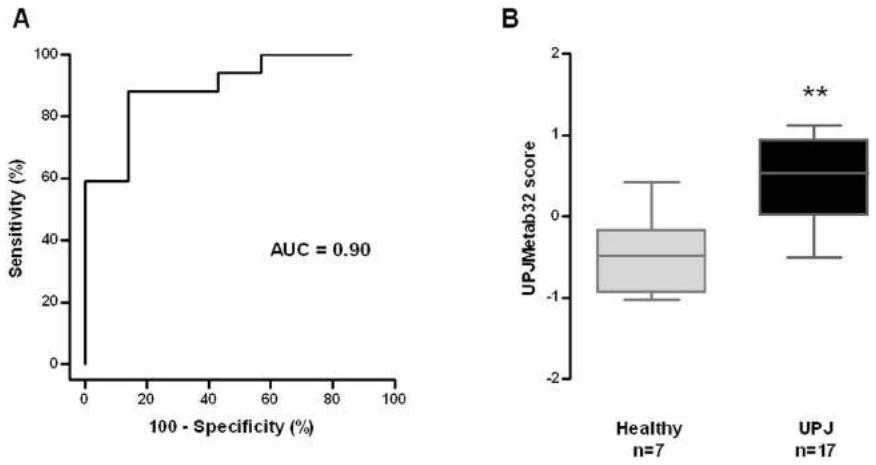

Figure 7. Validation of urinary metabolite classifier UPJMetab32 in a separate population. The diagnostic value of the UPJMetab32 model was tested in an independent cohort (7 healthy subjects and 17 UPJ patients) by a blinded analysis. (**A**) ROC curve for the UPJMetab32 classifier. (**B**) Box-whisker plot for classification of healthy and UPJ patients in the validation set according to the UPJMetab32 score. $^{**}p < 0.005$ versus healthy subjects. Mann-Whitney test for independent samples.

et al.[6] have performed LC-MS metabolomics analysis of urine in subjects receiving a standardized and weight-based diet. The largest source of instability was attributable to technical issues such as sample preparation and analysis; to a lesser extent, an inconstancy subject-to subjects as well as intrapersonal variability due to meals and time of day were observed; day-to-day fluctuation was minimal[6]. Despite that, several studies suggest the existence of a stable part (time scale: months to years) of the urine metabolomic profile which seems to be specific to each individual[54,55]. Under unrestricted lifestyle conditions, multiple collections of urine samples can be used to reduce the metabolic noise and retrieve the individual phenotype[56]. In the current study, differences in alimentation are most likely not a confounding factor since alimentation of newborns/infants is significantly less variable than in adults.

We have show-cased the use of the pipeline in a frequently encountered renal pathology in newborns[11,57]. We were able to identify 32 metabolic features associated to UPJ obstruction. Combination of the 32 metabolite features in a SVM classifier predicted with 76% sensitivity and 86% specificity UPJ obstruction in a separate validation cohort, thereby demonstrating the efficacy of the model to detect patients with UPJ obstruction. Increased carnosine excretion in UPJ was attributed to two highly correlating isotopes of a same metabolite. Carnosine is a dipeptide synthesized from alanine and histidine by the carnosine synthase in muscle, brain and other tissues such as kidney. It is degraded by the carnosinase predominantly in the liver but also in kidneys. Carnosine from animal food can also be absorbed in the small intestine, and at least part of it enters the blood intactly upon oral ingestion. Finally, kidneys filter plasma carnosine, reabsorb a part of carnosine *via* specific transporters and excrete the remaining in urine[58,59]. In order to understand the origin of the elevated urinary level of carnosine from UPJ obstruction patients, further experiments measuring expression of carnosine related-enzymes and transporter proteins in both obstructed and contralateral kidneys should be performed. The dipeptide possesses also strong antioxidant and free radical scavenging activities[58]. Interestingly, protective effects of carnosine have been demonstrated in rodent models of kidney disease[60–62] and in patients with diabetic nephropathy[63] or children with glomerulopathies[64]. Thus, increased urinary excretion of carnosine in UPJ obstruction could be an adaptive rather than a deteriorating mechanism.

In conclusion, we have developed a robust setup and analysis pipeline for the exploration by CE-MS of the metabolite content of urine and found that the long-term reproducibility of the metabolite data generated was excellent. As proof of concept, we demonstrated the feasibility to use CE-MS as a tool for the identification of clinically relevant urinary metabolites.

Materials and Methods

Patients and urine collection. *Samples used for optimization of the CE-MS normalization procedure.* Fifty-four urinary samples from various kidney and urinary tract pathologies together with 21 control CE-MS samples from healthy patients (Supplementary Table S1) were used. We considered that these samples represent the potential diversity to be encountered in clinical samples and hence used those samples for the development of CE-MS normalization procedure.

Quality control (QC). The QC sample was a mixture of urine samples of 9 healthy individuals (3 females and 6 males, mean age 34.1 ± 2.8 years).

Ureteropelvic junction (UPJ) obstruction and healthy patients. UPJ obstruction patients ($n = 51$) and healthy individuals ($n = 22$) of less than one year old were recruited in Toulouse Hospital and included in our study. The UPJ obstruction group was composed of patients scheduled for pyeloplasty with a pelvic dilatation of at least 16 mm and grade 3 and 4 hydronephrosis. Renographies were performed as soon as possible after birth,

generally between week 3 and 6 to establish baseline differential renal function (DMSAscan) and washout pattern (MAG3-scan). Healthy and UPJ obstruction patients were randomly divided into two cohorts: a discovery cohort (n = 49; Table 1 and Supplementary Table S3) and a blinded cohort for validation (n = 24; Table 2 and Supplementary Table S4). Mann Whitney analysis revealed no significant difference in the age of healthy and UPJ obstruction newborns included in both discovery and validation cohorts ($p = 0.26$). In addition, the use of Chi Squared test also revealed no gender bias ($p = 0.45$). Urine from newborns was collected in the morning during 30 min using a sterile pediatric urine collection pouch (B. Braun, Boulogne, France) during hospital consultation. Urine from healthy controls was collected from newborns in the maternity hospital and at home using the same sterile collection bags and a pair of gloves. Care was taken to not take the first morning urine. After collection, all urines were frozen within the hour at −20 °C both in the hospital (dedicated −20 °C freezer in the clinic) and at home. Transport was done using ice blocks in both cases and the samples were finally stored at −80 °C in the laboratory. The UPJ study was performed in accordance with the ethical principles in the Declaration of Helsinki and Good Clinical Practice. The study and its experimental protocols were approved by the ethics committee of the French Ministry of National Education, Higher Education and Research (number DC-2008-452). Written informed consent was obtained from all participants (parents of the newborns).

Sample Preparation. A 170 μl aliquot of urine was diluted with the same volume of a denaturing solution composed of 2 M urea, 0.0125% NH_4OH, 100 mM NaCl and 0.01% SDS. To remove higher molecular mass proteins, the sample was ultrafiltered using a Centristat 20 kDa cut-off centrifugal filter device (Satorius, Göttingen, Germany) at 2000 × *g* for 45 min at 4 °C. In order to remove urea, electrolytes and SDS, 200 μl of filtrate was applied onto a NAP5 gel filtration column (GE Healthcare Bio Sciences, Uppsala, Sweden), washed and then eluted with 700 μl of 0.01% NH_4OH. Finally, all samples were lyophilized in a Savant speedvac SVC100H connected to a Virtis 3L Sentry freeze dryer (Fischer Scientific, Illkirch, France). At this step, samples can be stored at 4 °C until use and re-suspended in HPLC grade water shortly before CE-MS analysis. The resuspension volume was adjusted to yield 1 μg/μl protein as measured by BCA assay (Pierce Biotechnology, Rockford, USA).

CE-MS analysis. CE-MS analyses were performed as previously described[11,12,65] using a Beckman Coulter Proteome Lab PA800 capillary electrophoresis system (Beckman Coulter, Fullerton, USA) on-line coupled to a micrOTOF II MS (Bruker Daltonic, Bremen, Germany). The electro-ionization sprayer (ESI, Agilent Technologies, Palo Alto, CA, USA) was grounded, and the ion spray interface potential was set between –4 and –4.5 kV. The CE separation buffer contained 20% (v/v) acetonitrile and 250 mM formic acid (Sigma-Aldrich) in HPLC-grade water. The CE-system was equipped with a 95 cm (internal diameter: 50 μm) bare fused silica capillary. Two types of CE-ESI-MS interfaces were tested (see results section); either a flatted or a tapered and beveled needle surrounding the capillary terminus. Data and MS acquisition methods were automatically controlled by the CE via contact-close-relays. Spectra were accumulated every 2 s, over a range of m/z 30 to 650.

CE-MS sample preprocessing for stable endogenous metabolites identification. After mass calibration using the measurement of sodium formate salts at the start of each run, the raw MS-data were converted into NetCDF format (http://www.unidata.ucar.edu/software/netcdf/) through the Bruker software (DataAnalysis version 4.0). The NetCDF files were filtered by excluding spectra corresponding to a migration time less than 520 or greater than 3650 seconds prior to preprocessing using the Bioconductor package *xcms*[34] as previously described[25]. All the standard xcms pipeline parameters were kept to their defaults apart from *steps* which was set to 3 and *bw* which was set to 20. In addition, the total number of migration time alignment iterations was set to 5, using the LOESS approach of xcms. The resulting molecule features derived from the execution of the xcms pipeline (in terms of m/z and migration time pairs) were further filtered for their presence across samples by including only those molecule features present in at least 50% of the total samples. The latter ensured the robustness of the initial set of molecule intensities which would be later interrogated for the presence of stable (in terms of intensity) molecule features that would serve as a set of CE-MS internal normalization standards.

Stable endogenous metabolites identification. The final filtered set of xcms preprocessed and identified m/z – migration time pairs was further interrogated for the potential presence of a set of 'housekeeping' metabolites with stable intensity across pathologies and spanning the whole intensity range. To this end, a subset of 'rank invariant' family of normalization algorithms from the DNA microarray literature was applied with the purpose of identifying stable molecule features that would represent the 'invariant set' as referenced in the microarray bibliography[66]. Specifically, the algorithms described in[35] (*dChip* algorithm)[36], (*lumi* Bioconductor package) and[37] (*GRSN* algorithm) were applied and sets of rank invariant metabolite abundances were retrieved. However, graphical assessment of the performance of these algorithms (see main text) revealed that the nature of CE-MS data prohibited the usage of these algorithms for the identification of a set of internal standards. Therefore, the following two strategies were applied:

1. The first strategy is based on the assumption that the majority of identified metabolites do not present differential abundance across samples (a similar assumption made for the normalization methods in the DNA microarray literature) and as a result, the relationship among different sample abundances is close to linear, after xcms preprocessing. Specifically, this approach includes the fitting of a set of Robust Linear Regression models[38,39], either among all possible sample pairs, or for all samples against a baseline (e.g. the median metabolite abundances across samples) and the calculation of each model residuals. The set of stable

metabolites is iteratively constructed by aggregating those ones whose abundance presented very low residuals in each model, implying low divergence from the model and subsequently, among samples.

2. The second strategy does not make any assumptions about the differential abundance distribution of the metabolites but requires noise preprocessing, as performed by the xcms pipeline and is based on the geometrical distance of each metabolite abundance vector in the sample space from the identity 'hyperline'. Specifically, this approach includes the construction of the identity 'hyperline' $\vec{Y} = \vec{X}$, in the n-dimensional sample space, where $\vec{Y} = (y_1,\ y_2,\ \ldots,\ y_n)\vec{X} = (x_1,\ x_2,\ \ldots,\ x_n)$, and n is the number of samples. Then for each metabolite abundance vector $\vec{A_i} = (a_{i1},\ a_{i2},\ \ldots,\ a_{in})$, the Euclidean distance d_i from an equally spaced grid distributed along $\vec{Y} = \vec{X}$ is calculated. The set of stable metabolites is constructed by aggregating metabolites with small d_i which imply both very high correlation as well as low inter-sample variability.

In both strategies (i) and (ii), the optimal number of metabolites with stable abundances is selected according to the normalizing potential of forward selected subsets of metabolites. The Forward Selection approach was selected as the number of stable metabolites should be kept to the minimum possible also for later purposes (exploration of prognostic or diagnostic values). Specifically, the initial candidate list is constructed by retrieving the first 1000 metabolites with the smallest Euclidean distance from the identity hyperline (or with the smallest residual value from an RLM) and sorting it in ascending order (of distance or residual value). Then, starting from a minimum set S of 10 metabolites, the whole dataset is normalized by fitting a LOESS curve L in this set and using it as the normalization reference. In each iteration one member of the stable metabolite candidate list is added to S, L is recalculated, the whole dataset is normalized and the following dataset variability metric is calculated:

$$M = MAD(\widetilde{X_1}, \widetilde{X_2}, \ldots, \widetilde{X_n}),\ \text{where} = (x_{1i}, x_{2i}, \ldots, x_{mi})$$

and x_{ij} the normalized abundance of metabolite i in sample j, $i = 1, \ldots, m$ (m the number of metabolites in the total dataset), $j = 1, \ldots, n$ (n the number of CE-MS samples). M reflects the total variability of the normalized intensity matrix by firstly summarizing each column (sample) by taking its median value and then calculating the variability of this summarization, by taking the Median Absolute Deviation (MAD) of the column medians distribution. The final number of the stable metabolites is the size of S that minimizes M and has thus the best normalizing potential while at the same time being as small as possible.

Processing and normalization of new samples. New CE-MS urine samples are preprocessed up to filtering (exclusion of spectra corresponding to a CE-time less than on average 840 [sodium salts] or greater than 3000 seconds) and peak-picking (no migration time alignment) with xcms as described above. Then, the masses of the new samples are matched against the reference dataset (consisting of 75 disease and control runs as described above) with a tolerance of 0.01 mass units, and the molecule features that do not match the reference are excluded from further analysis. The migration time alignment of the new samples is performed with an iterative procedure, similarly to the one followed by the xcms package but using the urine specific internal standards instead of the ones that are identified for independent datasets by xcms. Specifically, the migration times of the internal standards subset which is specific to the new sample (identified as described above) and span the whole range of the new sample's migration times, are used as seeds for the creation of migration time clusters using *k-means* clustering with the k parameter equal to the number of matching internal standards. Then, a LOESS curve is fitted to each cluster and used as a reference for the alignment of migration times in each cluster. The intensity normalization of the new samples is performed as described above ('stable endogenous metabolites identification' section), using the proper subset of the internal standards set according to the aforementioned mass match procedure.

Metabolite features annotation. The final set was matched against HMDB[67], ChEBI[68] and KEGG[69] for known molecules and annotated for potential chemical formulas using the *CAMERA* Bioconductor package[70]. From the two aforementioned methods, RLM and identity hyperline, the latter was selected as it was found to yield more robust results in terms of metabolite intensity coverage (S contained features spanning a sufficient range of intensities), normalization power, cardinality of S and its application did not require any assumptions for a baseline.

Statistical analysis. *Biomarker identification and modelling.* For the identification of potential metabolite biomarkers, the normalized levels of urinary metabolite features were compared between the healthy and UPJ obstruction patient groups. Only molecule features that were detected with a minimal frequency of 75% in every of the discovery groups were investigated for statistical analysis. Missing values (recorded as "Not a Number" [NaN]) from the discovery cohort were replaced by the average of the metabolite intensities found in the corresponding group (UPJ obstruction patients or healthy newborns). However, in the validation cohort where the belonging of the sample is unknown, we used the mean abundance of all patients from discovery set as imputation methods for missing values. Of note, zero values were considered as missing values. P-values were calculated for the comparison between healthy and UPJ obstruction patient groups using the Wilcoxon test followed by adjustment for multiple testing using the method described by Benjamini and Hochberg[46]. Only metabolite features with a corrected $p < 0.05$ were considered significant. Using an in-house developed tool, we next used a support vector machine (SVM)-based approach (SVM package e1071 of R)[71] to generate a prognostic biomarker classifier based on 32 biomarkers associated with UPJ obstruction. The parameters of the radial kernel function

(type C) for the multi-dimensional hyperplane were: cost parameter (C) of 1 and kernel width (γ) of 0.03125. Sensitivity and specificity were calculated using receiver operating characteristic (ROC) plots *via* the software R.

Comparison of svm scores. Statistical analyses were performed using GraphPad Prism 5.0 for Windows (GraphPad Software Inc) and comparisons between two groups were assessed using a Mann-Whitney test for independent samples. $p < 0.05$ was considered as statistically significant.

References

1. Decramer, S. *et al.* Urine in clinical proteomics. *Mol Cell Proteomics.* **7,** 1850–62 (2008).
2. Frantzi, M. *et al.* Developing proteomic biomarkers for bladder cancer: towards clinical application. *Nat Rev Urol.* **12,** 317–30 (2015).
3. Mischak, H. *et al.* Clinical proteomics: A need to define the field and to begin to set adequate standards. *Proteomics Clin Appl.* **1,** 148–56 (2007).
4. Dunn, W. B. *et al.* Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nat Protoc* **6,** 1060–83 (2011).
5. Gowda, G. A. & Djukovic, D. Overview of mass spectrometry-based metabolomics: opportunities and challenges. *Methods Mol Biol.* **1198,** 3–12 (2014).
6. Kim, K. *et al.* Mealtime, temporal, and daily variability of the human urinary and plasma metabolomes in a tightly controlled environment. *PLoS One.* **9,** e86223 (2014).
7. Brown, C. E. *et al.* Urinary proteomic biomarkers to predict cardiovascular events. *Proteomics Clin Appl.* **9,** 610–7 (2015).
8. Metzger, J. *et al.* Diagnosis of subclinical and clinical acute T-cell-mediated rejection in renal transplant patients by urinary proteome analysis. *Proteomics Clin Appl.* **5,** 322–33 (2011).
9. Metzger, J. *et al.* Urine proteomic analysis differentiates cholangiocarcinoma from primary sclerosing cholangitis and other benign biliary disorders. *Gut.* **62,** 122–30 (2013).
10. Zimmerli, L. U. *et al.* Urinary proteomic biomarkers in coronary artery disease. *Mol Cell Proteomics.* **7,** 290–8 (2008).
11. Decramer, S. *et al.* Predicting the clinical outcome of congenital unilateral ureteropelvic junction obstruction in newborn by urinary proteome analysis. *Nat Med.* **12,** 398–400 (2006).
12. Klein, J. *et al.* Fetal urinary peptides to predict postnatal outcome of renal disease in fetuses with posterior urethral valves (PUV). *Sci Transl Med.* **5,** 198ra106 (2013).
13. Schanstra, J. P. *et al.* Diagnosis and Prediction of CKD Progression by Assessment of Urinary Peptides. *J Am Soc Nephrol.* **26,** 1999–2010 (2015).
14. Schonemeier, B. *et al.* Urinary Peptide Analysis Differentiates Pancreatic Cancer From Chronic Pancreatitis. *Pancreas* (2016).
15. Posada-Ayala, M. *et al.* Identification of a urine metabolomic signature in patients with advanced-stage chronic kidney disease. *Kidney Int.* **85,** 103–11 (2013).
16. Zhao, X. *et al.* Metabonomic fingerprints of fasting plasma and spot urine reveal human pre-diabetic metabolic traits. *Metabolomics.* **6,** 362–374 (2010).
17. Wang, X. *et al.* Urine metabolomics analysis for biomarker discovery and detection of jaundice syndrome in patients with liver disease. *Mol Cell Proteomics.* **11,** 370–80 (2012).
18. Emwas, A. H. *et al.* Standardizing the experimental conditions for using urine in NMR-based metabolomic studies with a particular focus on diagnostic studies: a review. *Metabolomics* **11,** 872–894 (2015).
19. Nicholson, J. K. *et al.* Metabolic phenotyping in clinical and surgical environments. *Nature* **491,** 384–92 (2012).
20. Ramautar, R. Capillary Electrophoresis-Mass Spectrometry for Clinical Metabolomics. *Adv Clin Chem.* **74,** 1–34 (2016).
21. Gika, H. G., Macpherson, E., Theodoridis, G. A. & Wilson, I. D. Evaluation of the repeatability of ultra-performance liquid chromatography-TOF-MS for global metabolic profiling of human urine samples. *J Chromatogr B Analyt Technol Biomed Life Sci.* **871,** 299–305 (2008).
22. Novakova, L., Matysova, L. & Solich, P. Advantages of application of UPLC in pharmaceutical analysis. *Talanta* **68,** 908–18 (2006).
23. Gray, N. *et al.* Development of a Rapid Microbore Metabolic Profiling Ultraperformance Liquid Chromatography-Mass Spectrometry Approach for High-Throughput Phenotyping Studies. *Anal Chem.* **88,** 5742–51 (2016).
24. Gray, N., Lewis, M. R., Plumb, R. S., Wilson, I. D. & Nicholson, J. K. High-Throughput Microbore UPLC-MS Metabolic Phenotyping of Urine for Large-Scale Epidemiology Studies. *J Proteome Res.* **14,** 2714–21 (2015).
25. Ramautar, R., Somsen, G. W. & de Jong, G. J. CE-MS for metabolomics: developments and applications in the period 2010–2012. *Electrophoresis* **34,** 86–98 (2012).
26. Ramautar, R., Somsen, G. W. & de Jong, G. J. CE-MS for metabolomics: developments and applications in the period 2012–2014. *Electrophoresis* **36,** 212–24 (2015).
27. Kuehnbaum, N. L., Kormendi, A. & Britz-McKibbin, P. Multisegment injection-capillary electrophoresis-mass spectrometry: a high-throughput platform for metabolomics with high data fidelity. *Anal Chem.* **85,** 10664–9 (2013).
28. Ramautar, R., Busnel, J. M., Deelder, A. M. & Mayboroda, O. A. Enhancing the coverage of the urinary metabolome by sheathless capillary electrophoresis-mass spectrometry. *Anal Chem.* **84,** 885–92 (2012).
29. Zhang, W., Hankemeier, T. & Ramautar, R. Next-generation capillary electrophoresis-mass spectrometry approaches in metabolomics. *Curr Opin Biotechnol.* **43,** 1–7 (2016).
30. Harada, S. *et al.* Metabolomic profiling reveals novel biomarkers of alcohol intake and alcohol-induced liver injury in community-dwelling men. *Environ Health Prev Med.* **21,** 18–26 (2016).
31. Kami, K. *et al.* Metabolomic profiling of lung and prostate tumor tissues by capillary electrophoresis time-of-flight mass spectrometry. *Metabolomics* **9,** 444–453 (2013).
32. Kimura, T. *et al.* Identification of biomarkers for development of end-stage kidney disease in chronic kidney disease by metabolomic profiling. *Sci Rep.* **6,** 26138 (2016).
33. Soga, T. *et al.* Simultaneous determination of anionic intermediates for Bacillus subtilis metabolic pathways by capillary electrophoresis electrospray ionization mass spectrometry. *Anal Chem.* **74,** 2233–9 (2002).
34. Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R. & Siuzdak, G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal Chem.* **78,** 779–87 (2006).
35. Li, J. & Wong, L. Emerging patterns and gene expression data. *Genome Inform* **12,** 3–13 (2001).
36. Du, P., Kibbe, W. A. & Lin, S. M. lumi: a pipeline for processing Illumina microarray. *Bioinformatics* **24,** 1547–8 (2008).
37. Pelz, C. R., Kulesz-Martin, M., Bagby, G. & Sears, R. C. Global rank-invariant set normalization (GRSN) to reduce systematic distortions in microarray data. *BMC Bioinformatics* **9,** 520 (2008).
38. Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. & Stahel, W. A. *Robust Statistics: The Approach Based on Influence Functions.* (New York: Wiley) (1986).
39. Huber, P. J. *Robust Statistics.* (New York: Wiley) (1981).
40. Maxwell, E. J. & Chen, D. D. Twenty years of interface development for capillary electrophoresis-electrospray ionization-mass spectrometry. *Anal Chim Acta* **627,** 25–33 (2008).

41. Tseng, M. C., Chen, Y. R. & Her, G. R. A beveled tip sheath liquid interface for capillary electrophoresis-electrospray ionization-mass spectrometry. *Electrophoresis* **25,** 2084–9 (2004).
42. Gleiss, A., Dakna, M., Mischak, H. & Heinze, G. Two-group comparisons of zero-inflated intensity values: the choice of test statistic matters. *Bioinformatics* **31,** 2310–7 (2015).
43. FDA & Industry, G. f. Bioanalytical Method Validation, Food and Drug Administration: A Guidance. *Centre for Drug Valuation and Research (CDER)* (2001).
44. Taylor, S. & Pollard, K. Hypothesis tests for point-mass mixture data with application to 'omics data with many zero values. *Stat Appl Genet Mol Biol.* **8,** Article 8 (2009).
45. Dakna, M. *et al.* Addressing the challenge of defining valid proteomic biomarkers and classifiers. *BMC Bioinformatics* **11,** 594 (2010).
46. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Statist. Soc. B.* **57,** 289–300 (1995).
47. Mischak, H. *et al.* Recommendations for biomarker identification and qualification in clinical proteomics. *Sci Transl Med.* **2,** 46ps42 (2010).
48. Begley, P. *et al.* Development and performance of a gas chromatography-time-of-flight mass spectrometry analysis for large-scale nontargeted metabolomic studies of human serum. *Anal Chem.* **81,** 7038–46 (2009).
49. Zelena, E. *et al.* Development of a robust and repeatable UPLC-MS method for the long-term metabolomic study of human serum. *Anal Chem.* **81,** 1357–64 (2009).
50. Chen, G. Y., Liao, H. W., Tseng, Y. J., Tsai, I. L. & Kuo, C. H. A matrix-induced ion suppression method to normalize concentration in urinary metabolomics studies using flow injection analysis electrospray ionization mass spectrometry. *Anal Chim Acta.* **864,** 21–9 (2015).
51. Waikar, S. S., Sabbisetti, V. S. & Bonventre, J. V. Normalization of urinary biomarkers to creatinine during changes in glomerular filtration rate. *Kidney Int.* **78,** 486–94 (2010).
52. Chadha, V., Garg, U. & Alon, U. S. Measurement of urinary concentration: a critical appraisal of methodologies. *Pediatr Nephrol.* **16,** 374–82 (2001).
53. Wu, Y. & Li, L. Determination of total concentration of chemically labeled metabolites as a means of metabolome sample normalization and sample loading optimization in mass spectrometry-based metabolomics. *Anal Chem.* **84,** 10723–31 (2012).
54. Assfalg, M. *et al.* Evidence of different metabolic phenotypes in humans. *Proc Natl Acad Sci USA* **105,** 1420–4 (2008).
55. Bernini, P. *et al.* Individual human phenotypes in metabolic space and time. *J Proteome Res.* **8,** 4264–71 (2009).
56. Wallner-Liebmann, S. *et al.* The impact of free or standardized lifestyle and urine sampling protocol on metabolome recognition accuracy. *Genes Nutr.* **10,** 441 (2015).
57. Decramer, S., Bascands, J. L. & Schanstra, J. P. Non-invasive markers of ureteropelvic junction obstruction. *World J Urol.* **25,** 457–65 (2007).
58. Boldyrev, A. A., Aldini, G. & Derave, W. Physiology and pathophysiology of carnosine. *Physiol Rev* **93,** 1803–45 (2013).
59. Peters, V. *et al.* Intrinsic carnosine metabolism in the human kidney. *Amino Acids.* **47,** 2541–50 (2015).
60. Kurata, H. *et al.* Renoprotective effects of l-carnosine on ischemia/reperfusion-induced renal injury in rats. *J Pharmacol Exp Ther.* **319,** 640–7 (2006).
61. Riedl, E. *et al.* Carnosine prevents apoptosis of glomerular cells and podocyte loss in STZ diabetic rats. *Cell Physiol Biochem.* **28,** 279–88 (2011).
62. Yay, A. *et al.* Antioxidant effect of carnosine treatment on renal oxidative stress in streptozotocin-induced diabetic rats. *Biotech Histochem.* **89,** 552–7 (2014).
63. Janssen, B. *et al.* Carnosine as a protective factor in diabetic nephropathy: association with a leucine repeat of the carnosinase gene CNDP1. *Diabetes.* **54,** 2320–7 (2005).
64. Peters, V. *et al.* CNDP1 genotype and renal survival in pediatric nephropathies. *J Pediatr Endocrinol Metab* (2016).
65. Desveaux, C. *et al.* Identification of Symptomatic Fetuses Infected with Cytomegalovirus Using Amniotic Fluid Peptide Biomarkers. *PLoS Pathog.* **12,** e1005395 (2016).
66. Tseng, G. C., Oh, M. K., Rohlin, L., Liao, J. C. & Wong, W. H. Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects. *Nucleic Acids Res.* **29,** 2549–57 (2001).
67. Wishart, D. S. *et al.* HMDB 3.0–The Human Metabolome Database in 2013. *Nucleic Acids Res.* **41,** D801–7 (2013).
68. Hastings, J. *et al.* The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res.* **41,** D456–63 (2013).
69. Kanehisa, M. *et al.* Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.* **42,** D199–205 (2014).
70. Kuhl, C., Tautenhahn, R., Bottcher, C., Larson, T. R. & Neumann, S. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal Chem.* **84,** 283–9 (2012).
71. Meyer, D., Dimitriadou, E., Hornik, K. A. W. & F., L. E1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien. R package version 1.6-7. http://CRAN.R-project.org/package=e1071 (2015).

Acknowledgements

This work was supported in part by a research grant from the French Ministry of Education, Research and Technology for VP, by the "Programme Hospitalier de Recherche Clinique" (PHRC) [number No. 06 223 01 - No. RCB 2007-A00854-49] project for NL, ST, SD, and by the "European Consortium for High-Throughput Research in Rare Kidney Diseases" [EURenOmics, GA2012-305608) for JK, SD, JK, BBM.

Author Contributions

J.-L.B., S.D., J.P.S. and B.B.-M. designed the experiments; F.B., V.B., P.M. and B.B. performed the experiments; F.B., V.B., P.M., J.K., J.P.S. and B.B.-M. carried out analysis and interpretation of data; B.B., N.L., C.C., S.T. and S.D. collected clinical data and banked human urines; P.M., J.P.S. and B.B.-M. wrote the manuscript. All authors have reviewed the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/srep

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Boizard, F. *et al.* A capillary electrophoresis coupled to mass spectrometry pipeline for long term comparable assessment of the urinary metabolome. *Sci. Rep.* **6**, 34453; doi: 10.1038/srep34453 (2016).

3 Analyse du peptidome du liquide amniotique des anomalies congénitales du rein - Résultats

> "Clearly, however, before termination is considered, it is essential that every effort be made to ensure that the diagnosis be accurate."
>
> *Benson et Sej (1971)*

Les anomalies congénitales du rein et des voies urinaires sont la première cause d'insuffisance rénale chez l'enfant. Les outils de pronostic des maladies du développement rénal basés sur de l'imagerie prénatal sont aujourd'hui des outils de prédiction imprécis et tardifs. Il est toutefois essentiel pour le conseil parental prénatal. Cette étude décrit l'identification d'une signature peptidique du liquide amniotique à l'aide d'une grande cohorte de 178 grossesses impliquant des fétus porteurs d'une pathologie du développement rénal. Nous avons construit grâce à cette signature un modèle peptidique pouvant prédire le devenir postnatal rénal de ces fétus. Les performances obtenues avec cette signature sur une cohorte indépendante sont meilleures que les prévisions basées sur les mesures cliniques actuellement utilisées. Cette approche a l'avantage de fournir un score clair et objectif, ce qui en fait un vrai outil prénatal d'aide à la décision pour les médecins.

Amniotic fluid peptides predict postnatal renal survival in congenital anomalies of the kidney and the urinary tract

Julie Klein, Bénédicte Buffin-Meyer, Franck Boizard, Nabila Moussaoui, Ophélie Lescat, Benjamin Breuil, An Hindryckx, Luc Decatte, Elena Levtchenko, Anke Raaijmakers, Christophe Vayssière, Martine Dugué-Maréchaud, Emanuelle Descombes, Franck Perrotin, Sylvie Cloarec, Alexandra Benachi, Marie-Christine Manca-Pellissier, Hélène Laurichesse Delmas, Lucie Bessenay, Claudine Le Vaillant, Emma Allain-Launay, Jean Gondry, Bernard Boudailliez, Elisabeth Simon, Fabienne Prieur, Marie-Pierre Lavocat, Anne-Hélène Saliou, Loic De Parscau, Laurent Bidat, Catherine Noel, Corinne Floch, Guylène Bourdat-Michel, Romain Favre, Anne-Sophie Weingertner, Jean-François Oury, Véronique Baudouin, Jean-Paul Bory, Christine Pietrement, Maryse Fiorenza, Jérôme Massardier, Sylvie Kessler, Nadia Lounis, Françoise Conte Auriol, Pascale Marcorelles, Sophie Collardeau-Frachon, Petra Zürbig, Harald Mischak, Pedro Magalhães, Jean-Loup Bascands, Franz Schaefer, Stéphane Decramer, Joost P. Schanstra

en cours de publication : Journal of the American Society of Nephrology

Abstract

Background

Bilateral congenital anomalies of the kidney and urinary tract (CAKUT) are the first cause of end stage renal disease (ESRD) in children. Ultrasound-based prenatal clinical advice on postnatal renal survival in CAKUT pregnancies is far from accurate. We aimed to develop a novel procedure based on amniotic fluid (AF) peptides to improve the prediction of postnatal renal outcome.

Methods

Two-hundred pregnant women, expecting 178 CAKUT and 22 healthy fetuses, participated in this prospective observational multicenter cohort study. AF peptides were analyzed with capillary electrophoresis/mass spectrometry and ELISA. The primary outcome was chronic kidney disease stage 3-5 before two years of age or perinatal death due to ESRD or TOP with fetopathology confirming severe renal maldevelopment.

Results

Among the 7,000 AF peptides, 98 were associated with compromised renal outcome (corrected p-values <0.05) in a training cohort of 53 CAKUT pregnancies. The majority of these peptides were fragments from extracellular matrix proteins and thymosin-β4. Modification of AF thymosin-β4 abundance was confirmed with ELISA. Use of the 98 peptide-signature in an independent validation set of 51 CAKUT pregnancies led to the prediction of postnatal renal outcome with an AUC of 0.96 (95 %CI: 0.87-1.00), outperforming ultrasound. The signature was validated in a geographically different setting (AUC 1.00, n=12) and displayed high specificity (82 % to 94 %) in non-CAKUT settings (n=69).

Conclusion

The introduction of this AF peptide signature in the diagnostic workup of prenatally detected CAKUT can provide a long-sought evidence base for accurate management of the CAKUT disorder that is currently unavailable.

Introduction

Congenital anomalies of the kidney and the urinary tract (CAKUT) represent 20-30 % of all inborn malformations (Nicolaou *et al.*, 2015). Whereas prognosis is generally favorable in unilateral disease, bilateral CAKUT are the predominant causes of chronic kidney disease (CKD) in childhood (Calderon-Margalit *et al.*, 2018) and account for 50 % of pediatric and young adult end stage renal disease (ESRD) cases (Wühl *et al.*, 2013).

Bilateral CAKUT display a wide spectrum of outcomes ranging from death in utero to normal renal function after birth. Current routine ultrasound-based prenatal clinical advice on the postnatal renal outcome to parents expecting a child with CAKUT is far from accurate (Morris *et al.*, 2009b; Mehler *et al.*, 2018). Alternatively, invasive testing such as assessing fetal serum β2-microglobulin (Berry *et al.*, 1995) has been used, but is controversial due to the absence of clear cutoff values and the fact that only measurements at advanced gestational age are predictive (Spaggiari *et al.*, 2017b; Aulbert et Kemper, 2016b).

This predictive uncertainty has particularly serious implications for prenatal counseling of the parents confronted with the decision of continuation or elective termination of pregnancy. Such uncertainty has led to documented situations where half of the cases of severe bilateral CAKUT for whom termination of pregnancy was considered but not performed had normal postnatal renal function (Hogan *et al.*, 2012). In addition, knowledge of the precise outcome would allow anticipating dialysis, transplantation or palliative care in the case parents decide to continue the pregnancy.

Modern-era approaches such as clinical omics, allowing a precise molecular description of disease, is expected to be better linked to postnatal renal survival. However, although knowledge of the underlying genetic causes in CAKUT is increasing, the absence of a clear genotype-phenotype correlation in CAKUT (Nicolaou *et al.*, 2015; Decramer *et al.*, 2007) suggests that searching markers of progression should focus on traits beyond the genotype, closer to the phenotype including proteome and metabolome approaches. This has already been shown in adult CKD where a FDA-supported urinary peptide panel outperformed routine clinical measurements in prediction of disease progression (Nkuipou-Kenfack *et al.*, 2017). Therefore we hypothesized that application of clinical amniotic fluid peptidomics could change the status quo in CAKUT.

Here, using the specifically designed and largest prenatal prospective cohort in this renal developmental disease and a 2 year postnatal follow-up (Clinicaltrials.gov: NCT02675686), we developed and validated an amniotic fluid peptide signature and evaluated the value of adding this signature to prenatal management.

Results

Characteristics of the study population

During follow-up of the prospectively included 178 bilateral CAKUT patients, 34 pregnancies were excluded for medical or sampling reasons. Four other samples were withdrawn since the amniotic fluid peptidome could not be analyzed (Figure II.3.1). For the remaining 140 patients, the major etiologies were hyperechogenic kidneys (40/140) and lower urinary tract obstruction (29/140) representing 49 % of the patients (Table II.3.1). Forty nine percent (69/140) of the fetuses had normal or moderately reduced renal function (eGFR>60 ml/min/1.73m2) at 2 years postnatally. Etiologies mostly associated to normal outcome were non-obstructive urinary tract anomalies and upper urinary tract obstruction. In contrast, 71/140 (51 %) of the fetuses developed early-onset renal failure (CKD3-5, eGFR<60 ml/min/1.73m2 at 2y) or perinatal death due to ESRD or were

subjected to TOP. Non-functioning kidneys and lower urinary tract obstruction were the main etiologies associated to these compromised outcomes.

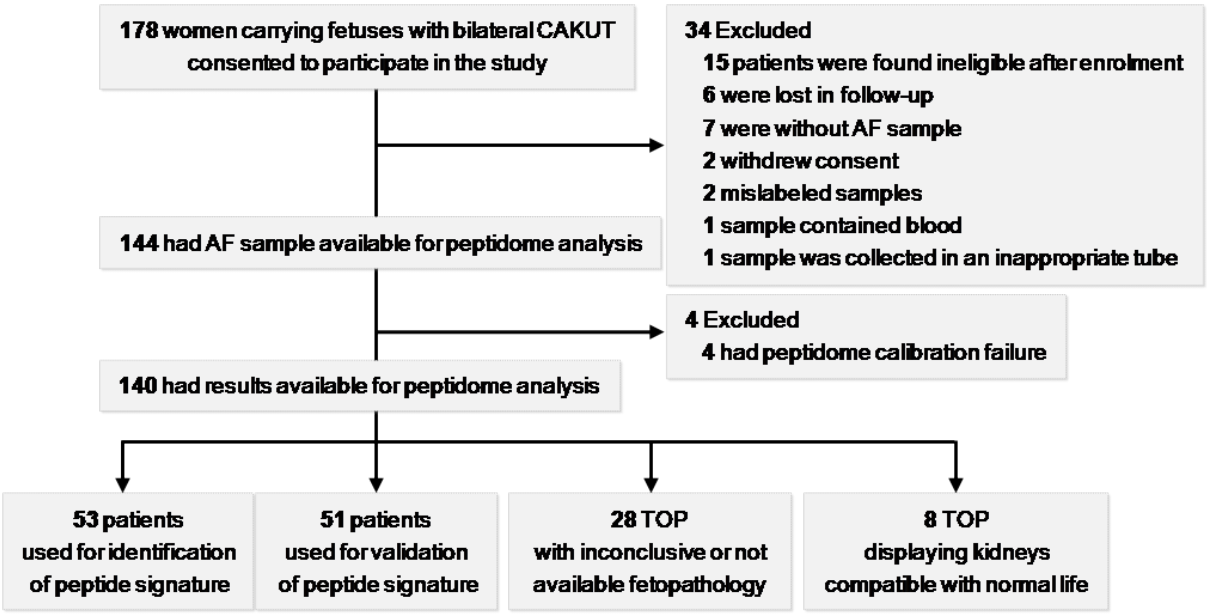

Figure II.3.1 – Participant flow in the prospective multicenter fetal bilateral CAKUT study(clinicaltrials.gov NCT02675686) specifically designed for the evaluation of the added value of AF peptides in the prediction of postnatal renal survival.

	N	Gender	Gestational age	Amniotic fluid	Outcome		
		(f/m)[1]	(w)[2]	(n.a, n, o, a)[3]	GFR>60	GFR<60 or death	TOP
Total cases	140	40/82	25.68 +/- 0.50	18, 68, 42, 12	69	11	60
Bilateral hyperechogenic kidneys							
normal size	17	04/12/19	26.06 +/- 1.29	2, 12, 3, 0	10	0	7
enlarged	23	14/08/19	27.35 +/- 1.30	4, 12, 7, 0	10	2	11
Lower urinary tract obstruction							
PUV	25	0/20	24.64 +/- 1.40	3, 3, 16, 3	5	4	16
others	4	01/02/19	17.50+/- 1.19	1, 3, 0, 0	1	0	3
Abnormal solitary kidney*							
agenesis	9	02/07/19	27.44 +/- 1.40	3, 2, 3, 1	4	1	4
MCDK	12	06/05/19	28.17 +/- 1.11	1, 8, 3, 0	8	1	3
Upper urinary tract obstruction**	17	01/11/19	25.12 +/- 1.15	2, 14, 1, 0	15	1	1
Bilateral hypoplasia	10	04/06/19	25.00 +/- 1.56	1, 4, 4, 1	5	0	5
Nonfunctioning kidneys***	8	03/05/19	20.75 +/- 1.77	0, 0, 1, 7	0	0	8
Non obstructive urinary tract anomalies****	7	02/01/19	23.43 +/- 2.16	0, 6, 1, 0	7	0	0
Bilateral dysplasia	5	01/04/19	28.80 +/- 2.82	1, 3, 1, 0	3	1	1
One hypoplastic and one dysplastic kidney	3	02/01/19	33.67 +/- 2.73	0, 1, 2, 0	1	1	1

Table II.3.1 – Antenatal cohort characteristics of 140 CAKUT patients of which the amniotic fluid peptidome was analyzed * One nonfunctional (agenesis or multicystic dysplastic kidney (MCDK)) kidney and one kidney with either ureteropelvic junction obstruction (UPJ) with parenchymal lesions or dysplasia or hypoplasia or hyperechogenecity or combinations thereof; ** Bilateral UPJ with bilateral parenchymal lesions; *** Bilateral agenesis or MCDK; **** Vesicoureteral reflux, duplex collecting system, megaureter;[1] Gender of fetus, female/male (18 missing values); [2] Gestational age plus or minus standard error in weeks; [3] Amniotic fluid volume: n.a, not available; n, normal; o, oligoamnios; a, anhydramnios; g Postnatal pregnancy outcome at two years of age: GFR>60, normal renal function or moderately reduced renal function (eGFR>60 ml/min/1.73m^2); GFR<60 or death, eGFR<60 ml/min/1.73m^2 or death due to renal dysfunction; TOP, termination of pregnancy; Abbreviation: PUV, posterior urethral valves.

Identification of an amniotic fluid peptide signature associated to postnatal renal outcome

For identification of peptides associated to postnatal renal outcome, a training cohort of 53 CAKUT patients (Figure II.3.1) including 35 with normal outcome (eGFR>60 ml/min/1.73m^2 at age 2 years) and 18 with compromised outcome (early-onset renal failure, perinatal death due to ESRD, or TOP with fetopathology showing severe renal maldevelopment) were used. A total of 7,000 peptides were detected in AF, for 1,008 of which sequence information could be obtained. Ninety-eight peptides with significantly different abundance (corrected p-values <0.05) and multi-fold changes (up to 100 fold) were associated to compromised postnatal renal outcome. The majority of the peptides were fragments of various collagens (88 %, Figure II.3.2 A). Other peptides included fragments of thymosin-β4, inter α trypsin inhibitor heavy chain H4 or osteopontin (3 %, 2 % and 1 %, respectively, Figure II.3.2 A). Increased abundance of a thymosin-β4 peptide was confirmed using a commercial enzyme-linked immunosorbent assay in a subset of patients (Figure II.3.2 B).

The 98 AF peptides were included in a random forest mathematical model (called the bCAKUTPep signature), which was optimized for the classification of the 53 patients of the training set leading to a predictor with an area under the receiver operator curve (AUC) of 0.92 (95 %CI: 0.85-1.00), a 78 % (95 %CI: 52-94) sensitivity and a 94 % (95 %CI: 81-99) specificity using a cutoff score of 0.47 (Figure II.3.2 C and D).

Independent validation of the amniotic fluid peptide signature

In a first step to clinical application of the bCAKUTPep signature, and in accordance with clinical proteomics guidelines, it is essential to confirm that predictive biomarkers are generalizable to 'similar but different' individuals outside the set of patients in which they were identified (Moons *et al.*, 2012; Desveaux *et al.*, 2016). Therefore in the next step we validated the signature in a new set of 51 CAKUT patients (Figure II.3.1) composed of 34 fetuses with normal renal outcome and 17 patients with compromised renal outcome. This resulted in prediction of postnatal renal outcome with an AUC of 0.96 (95 %CI: 0.90-1.00), a 88 % (95 %CI: 64-98) sensitivity and a 97 % (95 %CI: 85-100) specificity (Figure II.3.3 A and B). The predictive capacity of the peptide signature significantly outperformed routinely used clinical parameters. Indeed, both reduced AF volume (oligohydramnios/anhydramnios) and gestational age at sampling predicted postnatal renal outcome with significant (P(one-sided)=0.03 and 0.001, respectively) lower AUCs of 0.84 (95 %CI: 0.69-0.98) and 0.72 (95 %CI: 0.56-0.88), respectively (Figure II.3.3 C).

We next assessed whether the predictive performance of the peptide-based signature could be improved by adding these clinical parameters. Combination of the peptides with either AF volume or gestational age or both did not significantly improve the AUC (0.98 (95 %CI: 0.94-1.00), 0.97 (95 %CI: 0.94-1.00) and 0.97 (95 %CI: 0.94-1.00), respectively) compared to bCAKUTPep alone (0.96 (95 %CI: 0.90-1.00), Figure II.3.3 C).

In our study 60 out of 140 (43 %) CAKUT pregnancies were terminated (Table II.3.1) but for 28/60 fetopathology was absent (usually due to parental non-consent) or inconclusive (no definite status as to the severity of the renal lesions, see Methods for severity

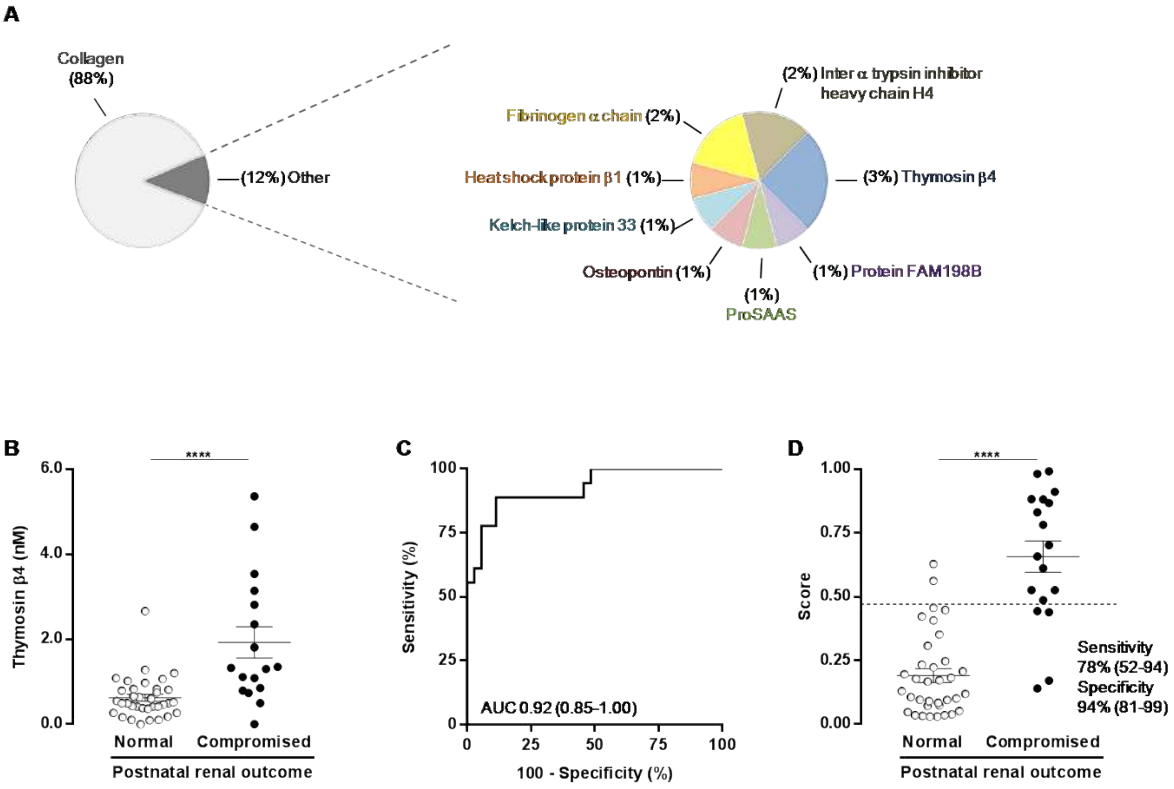

Figure II.3.2 – Identification of amniotic fluid peptides predictive of postnatal renal outcome in bilateral developmental renal disease A) Identities of the 98 AF peptides that displayed a significantly different abundance between 35 CAKUT fetuses with normal postnatal renal outcome (eGFR>60 ml/min/1.73m2 at age 2 years) and 18 CAKUT fetuses with compromised renal outcome (eGFR<60 ml/min/1.73m^2 at age 2 years, perinatal death due to ESRD, or TOP with fetopathology showing severe renal maldevelopment) from the training cohort. **B)** Enzyme-linked immunosorbent assay analysis (ELISA) of Thymosin β4 in AF of a subset of CAKUT patients. The abscissa indicates the clinical end-point at 2 years; ****P $<$ 0.0001, Mann-Whitney test for independent samples; Data are means plus or minus standard errors; normal, normal or moderately reduced postnatal renal function (eGFR>60 ml/min/1.73m2 at age 2 years); compromised, eGFR<60 ml/min/1.73m2 at age 2 years, perinatal death due to ESRD, or TOP with fetopathology showing severe renal maldevelopment.**C)** ROC curve of performance of the bCAKUTPep signature based on the random forest mathematical combination of the 98 peptides in the training cohort of 53 CAKUT patients. Confidence intervals given in brackets for AUC are two-sided 95 %CI. D) Scores of the bCAKUTPep signature in the training cohort of 53 CAKUT patients. The dotted horizontal line indicates the cutoff score of 0.47, which allowed to better discriminate CAKUT patients with normal outcome and CAKUT fetuses with compromised renal outcome. A bCAKUTPep score above the 0.47 cutoff predicts severely compromised postnatal renal outcome. The abscissa indicates the clinical end-point at 2 years; ****, P $<$ 0.0001, Mann-Whitney test for independent samples; Data are means plus or minus standard errors; normal, normal or moderately reduced postnatal renal function (eGFR>60 ml/min/1.73m^2 at age 2 years); compromised, eGFR<60 ml/min/1.73m^2 at age 2 years, perinatal death due to ESRD, or TOP with fetopathology showing severe renal maldevelopment.

assessment). The application of the bCAKUTPep signature to these 28 fetuses led to the prediction of 9 fetuses (32 %,Figure II.3.3 D) with a compromised outcome, thereby suggesting that the present sub-cohort exhibited an important number of TOP where postnatal renal outcome could have been compatible with normal life. Finally, for 8/60

TOP, triple expert reading fetopathology concluded that kidneys were compatible with normal life (Figure II.3.1). The bCAKUTPep signature predicted compromised outcome for only 2 of them (Figure II.3.3 E).

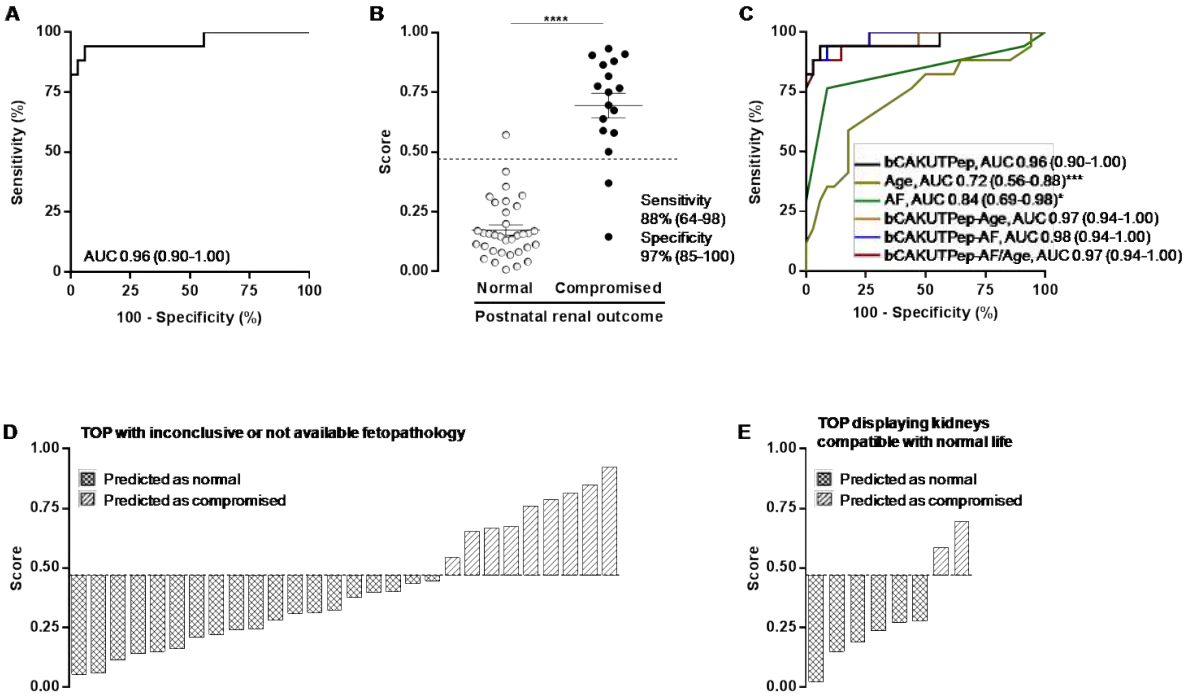

Figure II.3.3 – Validation of the performance and comparison to clinical parameters of the AF peptide signature. A) ROC curve of the bCAKUTPep signature in the validation cohort of 51 additional CAKUT patients. Confidence intervals given in brackets for AUC are two-sided 95 %CI. **B)** Scores of the bCAKUTPep signature in the validation cohort of 51 CAKUT patients. ****, $P < 0.0001$, Mann-Whitney test for independent samples; Data are means plus or minus standard errors; normal, normal or moderately reduced postnatal renal function (eGFR>60 ml/min/1.73m^2 at age 2 years); compromised, an eGFR<60 ml/min/1.73m2 at age 2 years, perinatal death due to ESRD, or TOP with fetopathology showing severe renal maldevelopment. Confidence intervals given in brackets for sensitivity and specificity are two-sided 95 %CI. **C)** ROC curve of the bCAKUTPep signature compared to clinical parameters or to its combination with those clinical parameters in the validation hold-out cohort of 51 CAKUT patients. Age, gestational age at AF sampling; AF, amniotic fluid volume; bCAKUTPep-Age, combination of the bCAKUTPep signature with gestational age at sampling; bCAKUTPep-AF, combination of the bCAKUTPep signature with AF volume; bCAKUTPep-AF/Age, combination of the bCAKUTPep signature with both gestational age at sampling and AF volume. * P(one-sided)=0.03 and ***P(one-sided)=0.001 versus bCAKUTPep using Hanley's method for comparing the areas under receiver operating characteristic curves derived from the same cases. Confidence intervals given in brackets for AUC are two-sided 95 %CI. **D)** Prediction of postnatal renal outcome by the bCAKUTPep signature of 28 TOPs in CAKUT pregnancies where fetopathology was inconclusive or not available. E) Prediction of postnatal renal outcome by the bCAKUTPep signature of 8 termination of pregnancies (TOPs) in bilateral CAKUT pregnancies where fetopathology, analyzed by three independent pathologists, displayed a renal phenotype type that appeared compatible with normal life. A bCAKUTPep score above the 0.47 cutoff suggests severely compromised postnatal renal outcome.

Robustness of the peptide-based signature

To further evaluate a potential routine application and the robustness of the signature we evaluated the performance of the peptides using other mathematical approaches including support vector machine (SVM), a k-nearest neighbor (KNN) or linear models. These models performed similarly well (AUCs of 0.96 (95 %CI: 0.90-1.00), 0.86 (95 %CI: 0.71-1.00) and 0.88 (95 %CI: 0.75-1.00) respectively, (Figure II.3.4 A), confirming the robustness of the selected peptides. Furthermore the AF peptide signature appears geographically independent since it showed excellent prediction (AUC: 1.00 (95 %CI: 1.00-1.00), Figure II.3.4 B) in a subset of patients from the validation cohort, i.e. 12 patients with CAKUT from Belgium (Belgium was not included in the training phase). Finally we performed a domain validation of the signature to test the specificity of the signature in individuals having a very different clinical status than CAKUT, even if the bCAKUTPep signature is not intended to be applied in those individuals. For this validation, we used AF samples of 22 healthy fetuses and of 47 fetuses with primary maternal CMV infection associated with different grades of neurological lesions 14. The latter cohort was used to preclude for the risk of a prediction only correlating with general fetal damage. The bCAKUTPep signature predicted normal postnatal outcome with a specificity of 82 % (95 %CI: 60-95) and 94 % (95 %CI: 82-99) in the two cohorts, respectively (Figure II.3.4 C).

Discussion

We developed and separately validated a novel method for the prediction of postnatal renal survival in developmental renal disease based on an AF peptide signature in the largest prospective bilateral CAKUT cohort to date. This signature (bCAKUTPep), assigning a numerical score with a clear-cut cutoff, predicted postnatal renal outcome with high sensitivity and specificity and significantly outperformed routine clinical ultrasound-based measures. The different antenatal CAKUT etiologies included in our study are similar to the CAKUT populations encountered in other but retrospective studies (Danziger *et al.*, 2016; Mehler *et al.*, 2018), suggesting that the signature will be applicable to CAKUT in general. We believe that the excellent predictive performance of the signature will be maintained with wider application. Indeed this study was conducted in strict adherence to the clinical proteomics guidelines (Mischak *et al.*, 2010b) for a rare disease, using separate and large training and validation multicenter cohorts and stringent statistics. Furthermore the signature was validated at a small scale in a geographically different setting, in another developmental non-renal (neurologic) disease and in healthy fetuses thereby demonstrating its robust applicability.

Counseling of future parents with a fetus with bilateral CAKUT is emotion loaded as it involves the consideration to terminate the pregnancy in the face of a highly uncertain prognosis ranging from largely normal postnatal kidney function to perinatal death or life-long end-stage kidney disease. The established AF peptide signature provides for the first time an unambiguous prediction of postnatal renal outcome with significantly higher accuracy compared to conventional methods. In addition, the measurement of AF peptides is not subject to personal interpretation, which can be the case for sonographic imaging (Mehler *et al.*, 2018; Loos et Kemper, 2018; Linder *et al.*, 2018). In case a high-risk

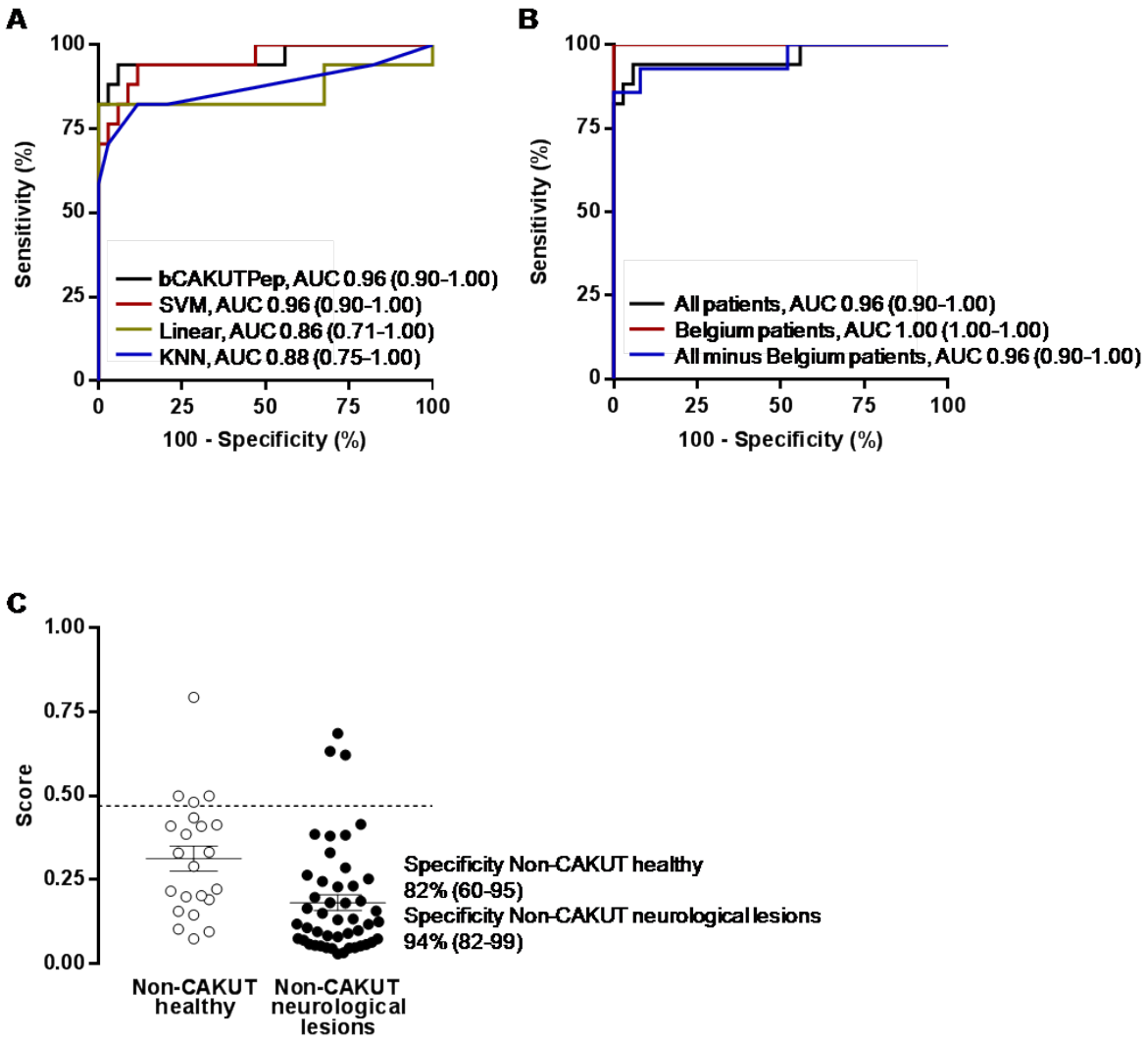

Figure II.3.4 – Robustness and wider application of the signature.A) ROC curves in the validation set of 51 CAKUT patients of the 98 peptides combined in different mathematical models. SVM, a support vector machine model; Linear, a linear model; KNN, a k-nearest neighbors model. **B)** ROC curves for the geographical validation of the bCAKUTPep signature. All patients, all patients in the validation set; Belgium patients, 12 patients from the validation set with a distinct geographical origin; All minus Belgium patients, the validation set without the Belgium patients.**C)** Domain validation using 22 healthy fetuses and 47 fetuses with primary maternal CMV infection displaying neurological lesions 14. The dotted horizontal line indicates the cutoff score of 0.47, previously defined in the training cohort using the random forest mathematical combination of the 98 peptides, above which a patient is predicted to display severely altered postnatal renal outcome. Confidence intervals given in brackets for AUC and specificity are two-sided 95 %CI.

phenotype is diagnosed, clear-cut knowledge of the outcome of the disease will give time to the future parents to psychologically accept (Stevenson et Goldworth, 1998) the fact that they will have a child with chronic, potentially severe disease and decide whether they would like their newborn to be offered palliative care or renal replacement therapies (RRT, i.e. dialysis or transplantation) (Lantos et Warady, 2013). Although RRT has improved over the last decade long term survival for young infants (<1 y), they still represent a major burden on the quality of life of the children and their families 20. Therefore, the

decision of initiating RRT must be weighed and anticipated carefully.

Sixty out of 140 (43 %) CAKUT pregnancies were terminated. This rate is slightly lower than in previous European studies where the rate of pregnancy termination was 55-62 % (Spaggiari *et al.*, 2017b; Jouannic *et al.*, 2003; Grijseels *et al.*, 2011; Ryckewaert-D'Halluin *et al.*, 2011), but close to a recent retrospective study from the US (45 % (32/71)) (Danziger *et al.*, 2016). Rarely put in the open, and as also observed in our study, a number of CAKUT pregnancies are terminated while subsequent fetopathology analysis reveals fetal kidneys with normal appearance (Hogan *et al.*, 2012). The added value of the AF peptide-based signature in this context is evident from the fact that the bCAKUTPep predicted a normal outcome for 6 out of the 8 terminated pregnancies in which fetopathology showed kidneys that appeared compatible with normal life. Although shown on a small number, application of such a signature in the management of CAKUT might thus have avoided 75 % of TOP displaying kidneys compatible with normal life.

The excellent predictive capacity of the signature potentially resides in the fact that early molecular modifications (i.e. changes in AF peptide abundance) precede structural and functional changes. We have recently observed that the presence of specific urinary collagen peptides is related to the degree of in situ kidney fibrosis in adult CKD (Magalhães *et al.*, 2017) and that these peptides are predictive of disease progression (Schanstra *et al.*, 2015). Similarly, we speculate that a focus on the AF peptides may allow assessing the early underlying molecular changes of CAKUT such as connective tissue turnover (collagen fragments) leading to hypo/dysplasia and hyperechogenicity, inflammation (osteopontin, inter α trypsin inhibitor heavy chain H4) and repair (thymosin β4).

A limitation for clinical implementation is that the analysis is mass spectrometry-based. However, such approach is indispensable since it is currently impossible to simultaneously analyze 98 peptides using immunological detection. In addition, we have shown in previous studies that samples can be frozen in the clinic, shipped and analyzed in specialized laboratories equipped with CE-MS technology (Klein *et al.*, 2013; Mischak *et al.*, 2013) with a total turnaround time of less than one week, an acceptable timeframe for clinical decision-making in CAKUT pregnancies. In terms of regulatory support, the CE-MS technology for body fluid peptide profiling was supported by the FDA for a postnatal application (Nkuipou-Kenfack *et al.*, 2017). Another limitation is that compared to ultrasound, AF sampling is invasive. However, recent studies show that the increased risk for miscarriage and stillbirth in women undergoing amniocentesis is almost solely due to the fact that this procedure is exclusively performed in at risk pregnancies (Malan *et al.*, 2018; Wulff *et al.*, 2016). We also did not identify adverse outcomes due to AF sampling in the current study. Therefore, the added value of accurate prediction of postnatal renal survival by the AF signature appears to outweigh the invasive procedure associated to AF sampling.

In conclusion, based on an objective numerical score, the introduction of the bCAKUT-Pep signature in the diagnostic workup of prenatally detected CAKUT provides a long-sought evidence base for rational management of the challenging CAKUT disorder that is currently unavailable.

Material and methods

Study design and patients

We performed a prospective multicenter observational study. Patient inclusion started on January 2011 and ended on September 2014. Follow up was continued until December 2016. Two-hundred women consented to participate in the study, including 178 originally identified as having a pregnancy with a fetus presenting bilateral CAKUT (Figure II.3.1) and 22 from non-CAKUT pregnancies. The 178 CAKUT pregnancies were recruited in 28 multidisciplinary prenatal diagnosis centers in France and Belgium. In these 2 countries, termination of pregnancies (TOP) can be considered in cases when severe fetal abnormality has been diagnosed and even in the third trimester of pregnancy. Centers were Amiens, Angers, Besançon, Brest, Caen, Clamart, Clermont-Ferrand, Grenoble, Leuven (Belgium), Limoges, Lyon, Marseille, Montpellier, Nancy, Nantes, Nice, Nîmes, Paris (R. Debré and Necker), Poitiers, Pontoise, Reims, Rennes, Saint-Etienne, Strasbourg, Talence, Toulouse and Tours. The 22 non-CAKUT patients were recruited in Toulouse and Poitiers. Inclusion criteria for bilateral CAKUT were a singleton pregnancy with bilateral developmental nephropathies with or without urinary tract malformations and a signed consent form. Kidney lesions on ultrasound were defined by hypoplasia (<2 SD) or hyperplasia (>2SD) and/or hyperechogenicity and/or the presence of cysts and/or reduced or absence of corticomedullary differentiation. Non-inclusion criteria were urinary tract anomalies without bilateral involvement of the kidneys, syndromic CAKUT, chronic (VIH, hepatitis B or C) or acute (e.g. chorioamnionitis) infection and refusal of the parents to participate in the study. During follow-up of the 178 CAKUT patients, 34 pregnancies were excluded (Figure II.3.1) . Fifteen were found ineligible after enrolment (presence of chromosomal abnormalities (n=6); polymalformative syndromes or digestive stenosis (n=8) and fetal death independent of end stage of renal disease (ESRD) (n=1)). Six were lost in follow-up. Seven AF samples were unavailable. Two families withdrew consent. Two AF samples were mislabeled. One AF sample contained blood and another was collected in an inappropriate clot activator tube. Four AF additional samples were excluded as they did not pass quality control after peptidome analysis (low number of peptides (<800), impossible to calibrate and normalize) (Figure II.3.1) . As a result 140 were considered for further analysis. The trial was performed in accordance with the Declaration of Helsinki (hel, 1997) and with Good Clinical Practice guidelines. The study endpoint was the renal status after 2 years of postnatal follow-up, obtained in each center for all patients. For live born children, renal function was estimated at 2 years of life using serum creatinine concentrations according the Schwartz method (Schwartz *et al.*, 2009). A fetopathological analysis was performed to confirm the severity of renal damage in case of perinatal death or in case of TOP. The 22 samples from non-CAKUT fetuses were obtained from pregnancies tested, but being negative, for chromosomal abnormalities. Furthermore, follow-up until birth did not reveal any abnormalities in these 22 pregnancies. The research was approved by national ethics committees (N ° RCB 2010-AO1151-38, France and S 55406 and B32220096569, Belgium) and informed consent was obtained from each participant.

Ultrasound observations

Ultrasound observations obtained within 2 weeks of amniotic fluid sampling leading to the presence of specific ultrasound annotations were the following: (i) Oligohydramnios/anhydramnios: amniotic fluid (AF) volume was estimated using the amniotic fluid index (AFI) which is defined as the sum of the largest vertical fluid pocket in each of the four quadrants of the maternal abdomen. An AFI of <5 indicated the presence of oligohydramnios, an AFI of 0 was considered as absence of AF (anhydramnios); (ii) Hyperechogenicity was defined by a right kidney brighter than the liver and a left kidney brighter than the spleen; (iii) Dysplasia was defined by the absence or reduced corticomedullary differentiation, (iv) Hypoplasia was defined by a kidney size of <2SD and, (v) Hyperplasia was defined by a kidney size of >2SD.

Fetopathology after termination of pregnancy

Post-TOP fetopathology was reviewed and assessed by 3 independent pathologists for attribution of a severity renal score: HS, high severity, defined by extensive dysplasia and/or hypoplasia; S, severe, segmental dysplasia and/or hypoplasia with alternations between healthy and pathological areas; LS, low severity, corresponding to kidneys with nearly normal parenchyma or little segmental dysplasia and/or hypoplasia. Dysplasia was defined by alteration of the renal structure with both glomerular and tubular lesions, persistence of primitive medullar tubules surrounded by fibromuscular cells and cartilaginous islets; hypoplasia was histologically defined by a reduction of structurally normal nephron number. At least one HS score without any LS score was interpreted as fetuses with renal lesions incompatible with normal life. At least two LS scores without any HS score was interpreted as compatible with normal life. All other combinations of scores or absence of fetopathology data were considered as inconclusive.

Training and validation sets

The prospective multicenter cohort of 104 bilateral CAKUT fetuses for which we had definite endpoint data was divided in independent training and validation sets for developing and then testing the signature performances, respectively. To be free from potential unintentional inter center differences in both eligibility criteria and outcome, we tried to distribute patients from each center (except for Belgium, see below) in both the training and validation sets. The two sets were balanced in terms of gender, gestational age at sampling, AF volume, postnatal renal outcome and antenatal etiology. All patients from Belgium (n=12) were excluded from the training set in order to evaluate the model performance in a small subset of patients enrolled in another country (geographical validation). The 22 healthy fetuses from pregnancies of healthy women were included together with 47 fetuses with primary maternal CMV infection (Desveaux *et al.*, 2016) in another independent set in order to perform a validation in individuals having a very different clinical status than CAKUT (domain validation).

Sample collection

Amniotic fluid (AF) was collected according to local management under ultrasound guidance and frozen at -20 ° C locally. Samples in case of anhydramnios were obtained after amnioinfusion. Samples were shipped on dry ice to the central laboratory (Inserm U1048, Toulouse, France).

Sample preparation

Briefly, immediately before preparation, amniotic fluid aliquots kept at –80 ° C were thawed and 150 µl aliquots were diluted with the same volume of 2 M urea, 10 mM NH4OH containing 0.2 % SDS. Subsequently, samples were passed over aCentristat 20-kDa cut-off centrifugal filter device (Sartorius) in order to eliminate high molecular weight compounds. The filtrate was desalted using a NAP-5 gel filtration column (GE Healthcare) to remove urea and electrolytes. Lyophilisation of the sample was performed using a Savant speedvac SVC100H connected to a Virtis 3L Sentry freeze dryer (Fisher Scientific) and stored at 4 ° C until use.

CE-MS analysis

Shortly before CE-MS analysis, the samples were re-suspended in 10µL of HPLC grade H2O. CE-MS analyses were centrally performed in Toulouse, France (INSERM) between May 2015 and October 2017, as previously described (Theodorescu *et al.*, 2006), using a Beckman Coulter Proteome Lab PA800 capillary electrophoresis system (Beckman Coulter) on-line coupled to a micrOTOF II MS (Bruker Daltonic). The electro-ionization sprayer (Agilent Technologies) was grounded, and the ion spray interface potential was set between –4 and –4.5 kV. Data and MS acquisition methods were automatically controlled by the CE via contact-close-relays. Spectra were accumulated every 3 s, over a range of m/z 350 to 3000.

Data processing

Mass spectral ion peaks representing identical molecules at different charge states were deconvoluted into single masses as described (Desveaux *et al.*, 2016). Normalization of the amplitude of the amniotic fluid peptides was based on sequenced endogenous "housekeeping" peptides that varied little among the samples (Theodorescu *et al.*, 2006).

ELISA measurement of Ac-SDKP/Thymosinβ4

N-acetyl-seryl-aspartyl-lysyl-proline (Ac-SDKP) is a natural tetrapeptide released from thymosin-β4. Ac-SDKP was measured in 50 µL amniotic fluid (1X) using a commercially available, highly specific, competitive enzyme-linked immunosorbent assay (ELISA) kit (Bertin Pharma, France) as described previously (Kumar *et al.*, 2016).

Adherence to guidelines

For the entire study we aimed to adhere to "Transparent reporting of a multivariable prediction model" (TRIPOD (Moons *et al.*, 2015)).

Data and materials availability

All data in this study will be accessible with publication of the manuscript. This includes the deidentified clinical data, the normalized amniotic fluid peptide data of each patient and the link between the clinical and peptidome data. The data has been deposited on Mendeley (data.mendeley.com) with the following reserved doi:10.17632/pnk73m22xn.1 and link available for review for referees (url to be copy/pasted in browser and account created): data.mendeley.com/datasets/pnk73m22xn/draft?a=a7b5bdf8-8014-4d8e-b6e2-b1184a833517

Statistics

Development of signatures: Considering only AF peptides detected in at least 75 % of the samples in a group, significant peptides were selected by Wilcoxon analysis followed by correction for multiple testing using the method of Benjamini-Hochberg(Benjamini et Hochberg, 1995). The prognostic 'bCAKUTPep' peptide signature was generated using the Random Forest (RF)-package (Liaw et Wiener, 2002) of R. The number of trees was fixed to 1000 and the other parameters were kept as default values. For combination with gestational age at AF sampling (bCAKUTPep-Age), data were previously arcsin-transformed. For combination with AF volume (bCAKUTPep-AF), we first assigned a score of 1 for normal AF volume or polyhydramnios, 2 for oligohydramnios and 3 for absence of AF. A score of 0 was attributed in case of missing value (8 and 4 cases in training and validation cohorts, respectively). For modeling with both gestational age at sampling and AF volume (bCAKUTPep-AF/Age), two peptides that did not influence the accuracy of the model in the total cross-validation of the training data were left out from the final bCAKUTPep-AF/Age model. In all cases, the number of trees was fixed to 1000 and the other parameters were kept as default values. To evaluate the robustness of bCAKUTPep, the 98 selected peptides were modelled using other mathematical models than RF including support vector machines (SVM), k-nearest neighbors (KNN) and linear models. For the SVM model (SVM package of R (Meyer *et al.*, 2019), the parameters of the radial kernel function (type C) were 1 (cost parameter) and 0.01020408 (kernel width). For the development of the KNN signature we used the KNN function of the R software (class package (Venables et Ripley, 2002)). The k parameter, tuning to get the best AUC on training dataset, was set to 27. To perform the linear model we used the method previously described (Klein *et al.*, 2013). For all models, we choose the score as cut-off based on the best outcome of the training set. Predictive performance was assessed by calculating sensitivity, specificity, area under the receiver-operating-characteristic curve (AUC) using Medcalc (Version 14.12.0).

Comparisons: Characteristics and signature based-scores of CAKUT case patients were compared with CAKUT control patients using a Mann-Whitney test. To assess the discriminatory ability of clinical parameters or signature, we tested the hypothesis that the

AUC is 0.5 (Hanley et McNeil, 1982). Comparisons of sensitivities and specificities between bCAKUTPep and currently used clinical methods or between the different signatures were performed using the McNemar test for paired proportions whereas comparisons of AUC were performed according the Hanley's method 40. Two-tailed tests were conducted excepted when we evaluated i) the added-value of bCAKUTPep model versus currently used clinical parameters and ii) the added-value of models combining peptides with gestational age and/or AF volume versus the bCAKUTPep signature. In all cases, $p<0.05$ was considered as statistically significant.

Conclusion

PAR sa facilité d'accès et la diversité des molécules qui s'y trouvent, l'urine est un fluide particulièrement intéressant pour la recherche de biomarqueurs des pathologies rénales. Ces biomarqueurs s'avèrent essentiels pour diagnostiquer précocement les maladies, prédire leur évolution et le risque de complications. Les modèles diagnostiques et pronostiques basés sur une combinaison de biomarqueurs ont de meilleures performances cliniques que ceux qui n'utilisent qu'un biomarqueur unique (Coffman et Richmond-Bryant, 2015). Toutefois, la construction des modèles multimarqueurs nécessite le recours à des méthodes bio-informatique qu'il est difficile de s'approprier sans une maîtrise des logiciels de programmations statistiques. C'est pour répondre à ce besoin de prise en main des outils statistiques dans la recherche en biologie que j'ai développé *La Boize*.

En donnant accès à un grand nombre de méthodes statistiques et grâce à une interface graphique simple d'utilisation, *La Boize* est un nouvel outil pour analyser des données omiques et créer des modèles de prédiction sans utilisation du langage de programmation. Le logiciel *La Boize* est désormais utilisé quotidiennement au laboratoire et son application a été valorisée dans deux études détaillées dans ce manuscrit. La première décrit l'application d'une méthodologie robuste pour identifier les métabolites urinaires associés à la sévérité des patients atteint d'obstruction de la jonction pyélo-uretérale ; la deuxième étude se focalise sur l'analyse du peptidome du liquide amniotique afin de prédire la fonction rénale post-natale des fœtus porteurs d'une anomalie du développement rénal.

Limites et perspectives

Les outils statistiques mis en œuvre dans *La Boize* sont actuellement relativement simples. Il existe plusieurs axes d'amélioration de l'outil qui passeront par l'intégration de méthodes plus complexes.

— Les sets de biomarqueurs sont identifiés dans *La Boize* par une approche monovariée, où chaque biomarqueur est considéré comme indépendant. Les approches multivariées peuvent avoir une meilleure capacité de prédiction (Robotti *et al.*, 2013) en prenant en compte la structure de corrélation des variables. Il pourrait donc être utile à des fins diagnostiques d'ajouter à *La Boize* des méthodes multivariées.

— La Boize a été développée à partir d'études caractérisant deux groupes d'individus (malade ou sain). Cependant, beaucoup d'études différencient un nombre plus élevé de groupes, en prenant en compte par exemple la sévérité de la maladie. Il pourrait donc être utile d'implémenter dans le logiciel la possibilité de prendre en compte

plus de deux groupes.

— Les données prises en compte par *La Boize* ne concernent qu'une seule strate moléculaire. La prise en compte de plusieurs niveaux moléculaires, ou bien de données cliniques pourrait améliorer les performances prédictives (Ge *et al.*, 2018). Sachant que chaque type de données peut nécessiter un type de traitement particulier (Zhan *et al.*, 2019; Pirola et Sookoian, 2018), il pourra être nécessaire d'y inclure l'utilisation d'outils statistique comme la PLS-DA (Singh *et al.*, 2016).

Néanmoins, l'utilisation de méthodes et de données plus complexes pourraient constituer une difficulté supplémentaire. Or l'objectif de *La Boize* est d'être à l'interface entre le domaine des statistiques et celui de la biologie. Il est donc particulièrement important de faire attention à sa maniabilité en biologie tout en répondant aux besoins de la recherche.

Conclusion générale

> "There is no good reason to think that only one discipline has all the answers."
>
> *Chris Eliasmith (2015)*

Mon travail de thèse s'est intéressé à l'analyse de la composition des fluides biologiques et leurs utilisations dans le diagnostic et la compréhension des mécanismes pathologiques rénaux par des approches de biologie des systèmes.

« Comment identifier de nouveaux acteurs clés dans le développement des maladies rénales à partir de l'analyse de la composition moléculaire de l'urine ? »

Nous nous sommes tout abord intéressés à l'identification de nouveaux acteurs clés dans le développement des maladies rénales à partir de l'analyse de la composition moléculaire de l'urine. Le protéome urinaire est connu comme un excellent reflet de modifications physiopathologiques *in situ* dans le rein. Cependant, toutes les protéines exprimées dans le rein ne sont pas détectées dans l'urine. En s'intéressant aux interactions protéine-protéine (PPI), il est toutefois possible de compléter ces observations grâce à la représentation des processus biologiques dans leur ensemble. Le développement massif des techniques de détection moléculaire et de stockage des données des PPI, nous permet aujourd'hui de construire l'interactome humain de manière satisfaisante. L'analyse de la structure du réseau PPI humain via les centralités, nous permet d'identifier des protéines essentielles de l'organisme et également importantes dans les processus pathologiques. Plusieurs types d'approches présentées dans cette thèse se sont intéressés au réseau PPI dans l'étude des maladies rénales. Cependant, peu de ces méthodes ont réussi à identifier de nouveaux acteurs clés des maladies rénales. Dans ce but j'ai proposé une méthode, appelé PRYNT (Priorization bY causal NeTwork). PRYNT prédit les protéines clés des maladies rénales à partir de l'analyse de la composition moléculaire urinaire. La validation de cette approche, sur quatre jeux de données concernant deux maladies rénales, démontre qu'il est

possible d'identifier de nouveaux acteurs clés des maladies rénales grâce à la protéomique urinaire combinée à une approche de biologie des systèmes.

« Comment détecter la présence d'une maladie rénale ou prédire son évolution à partir de l'analyse de la composition moléculaire de l'urine ? »

Le deuxième volet de mon travail a consisté en une approche diagnostique des maladies rénales. Pour cela, j'ai développé un outil appelé *La Boize*, permettant l'analyse des résultats des données omiques. Cette application a vocation à être utilisée facilement et sans connaissance bio-informatique préalable. Cet outil permet la construction d'un modèle de prédiction des maladies à partir de données omiques. Cette approche a fait l'objet de deux validations : comme outil de diagnostic à partir du métabolome urinaire de nouveaux nés atteints d'une obstruction de la jonction pyélo-urétérale, et également comme outil de prognostic des maladies du développement rénal à partir du peptidome du liquide amniotique. Ces deux études montrent l'intérêt de l'étude de la composition moléculaire des liquides biologiques comme outil d'aide à la décision clinique.

La multidiciplinarité : une complexité nécessaire.

La plupart des maladies ont un développement complexe : elles impliquent des interactions entre des composés moléculaires différents (gènes, protéines, métabolites...) et à différentes échelles (cellules, tissues, organes...). L'étude de ces différentes composantes de l'organisme nécessite des technologies variées (spectrométrie de masse, séquençage, imagerie...), qui génère des données dont le traitement et l'analyse statistique et biologique nécessitent des compétences là encore différentes. C'est leur mise en interaction qui pourra permettre la compréhension des processus biologiques dans leur intégralité. L'ultra-spécialisation n'est donc pas la meilleure réponse pour faire face à la croissance de la complexité de la recherche scientifique. De nouvelles découvertes et des solutions innovantes seront possibles seulement si des chercheurs de différentes disciplines mettent en commun leurs connaissances et leurs compétences (Mabry *et al.*, 2008).

La multidisciplinarité constitue pourtant toujours un challenge que j'ai pu expérimenter au cours de ma thèse. C'est à mon sens, une des principales difficultés à laquelle se confronte aujourd'hui la recherche médicale. Nous avons vu ici qu'un certain nombre de méthodes se développent dans un domaine, en informatique ou en bioinformatique par exemple, et qui ne sont pourtant pas appliquées directement en biologie. En effet, la transmission inter-disciplinaire des connaissances n'est pas aisée : discordance des attentes, incompréhension ou outils non-disponibles ou non-utilisables. La principale contribution de mon travail est la mise en commun des connaissances de plusieurs domaines. À mon niveau je donne ici des exemples de projet faisant appel à plusieurs domaines et menant à des approches qui s'appuient sur la bioinformatique, la statistique et la biologie des systèmes, pour s'étendre à une application clinique.

C'est le développement de ce type d'approche multidisciplinaire qui a fait naître le

concept de médecine personnalisé. Ce nouveau modèle de médecine tend à proposer à chaque patient un suivi et un traitement qui correspond à son profil moléculaire (génétique, protéique, métabolique, etc ...) mais aussi à ses antécédents cliniques (Ruiz et Philippe, 2012). Le développement de la médecine personnalisée passera par un développement des méthodes d'analyses omiques : création d'outils d'analyse omiques utilisable en clinique ; augmentation des dépistages non-invasif pour prendre en charge les patients à risque avant les premiers symptômes de la maladie ; la collection de profils d'un grand nombre d'individus pour plusieurs strates moléculaires. La principale difficulté à l'heure actuelle est la constitution de bases de données suffisamment complètes pour développer permettre l'évolution vers la médecine personnalisée de demain.

Glossaire

ACP	Analyse en Composante Principale
ADN	Acide DésoxyriboNucléique
ADPKD	Autosomal Dominant Polycystic Kidney Disease
AF	Amniotic Fluid
AFI	Amniotic Fluid Index
ARN	Acide RiboNucléique
AUC	Area Under the Curve
BIOGRID	BIOlogical General Repository for Interaction Datasets
BRET	Bioluminescence Resonance Energy Transfer
CAKUT	Congenital Anomalies of the Kidney and Urinary Tract
CE-MS	Capillary electrophoresis - Mass Spectrometry
CE	Capillary electrophoresis
ChEBI	Chemical Entities of Biological Interest
CI	Confidence Interval
CKD	Chronic Kidney Disease
CMV	Cytomegalovirus
CTD	Comparative Toxicogenomics Database
CV	Coefficient of Variance
DE	Différentiellement Exprimé
DIP	Database of Interacting Proteins
eGFR	estimated Glomerular Filtration Rate
ESRD	End Stage Renal Disease
FRET	Förster Resonance Energy Transfer
GEO	Gene Expression Omnibus
GO	Gene Ontology
GRSN	Global rank-invariant set normalization algorithm
HIPPIE	Human Integrated Protein-Protein Interaction rEference
HMDB	Human Metabolome Database
HPRD	Human Protein Reference Database
I2MC	Institut des Maladies Métaboliques et Cardiovasculaires
IPA	Ingenuity Pathway Analysis
KEGG	Kyoto Encyclopedia of Genes and Genomes
KNN	K-Nearest Neighbors
KUPKB	Kidney and Urinary Pathway Knowledge Base
KUPNetViz	Kidney and Urinary Pathway Network Vizualizer

LC-MS	Liquid chromatography - Mass Spectrometry
MCDK	MultiCystic Dysplastic Kidney
MIMIx	Minimum Information about a Molecular Interaction experiment
MINT	Molecular INTeraction database
MS	Mass Spectrometry
NaN	Not a Number
PMV	Point-of-Mass Values
PPI	Protein-Protein Interaction
PRYNT	PRioritization bY causal NeTwork
PUV	Posterior Urethral valve
QC	Quality Control
RMN	Résonance Magnétique nucléaire
RRT	Renal Replacement Therapies
SD	Standart deviacion
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
SVM	Support Vector Machines
TOP	Termination Of Pregnancy
UPJ	UreteroPelvic Junction obstruction
NMR	Nuclear Magnetic Resonance spectroscopy
OMIM	Online Mendelian Inheritance in Man
PIP	human Protein-protein interaction Prediction
PLS-DA	Partial Least Squares Discriminant Analysis
URA	Upstream Regulator Analysis

Bibliographie

"Moi, j'ai appris à lire, et ben je souhaite ça à personne !"
Léodagan de Carmélide (2006)

(1997). World Medical Association Declaration of Helsinki : Recommendations Guiding Physicians in Biomedical Research Involving Human Subjects. *JAMA*, 277(11):925–926. Cité 1 fois, p. 100.

Abedi, M. et Gheisari, Y. (2015). Nodes with high centrality in protein interaction networks are responsible for driving signaling pathways in diabetic nephropathy. *PeerJ*, 3:e1284. Cité 3 fois, p. 30, 31 et 32.

Abirami, K. (2001). Urinalysis in Clinical Practice (Akin to Liquid Kidney Biopsy). 2(1):12. Cité 1 fois, p. 71.

Adachi, J., Kumar, C., Zhang, Y., Olsen, J. V. et Mann, M. (2006). The human urinary proteome contains more than 1500 proteins, including a large proportion of membrane proteins. *Genome Biology*, 7(9):R80. Cité 3 fois, p. 5, 6 et 45.

Ahn, S.-M. et Simpson, R. J. (2007). Body fluid proteomics : Prospects for biomarker discovery. *PROTEOMICS – CLINICAL APPLICATIONS*, 1(9):1004–1015. Cité 1 fois, p. 64.

Alanis-Lobato, G., Andrade-Navarro, M. A. et Schaefer, M. H. (2017). HIPPIE v2.0 : Enhancing meaningfulness and reliability of protein–protein interaction networks. *Nucleic Acids Research*, 45(D1): D408–D414. Cité 1 fois, p. 14.

Albert, I. et Albert, R. (2004). Conserved network motifs allow protein-protein interaction prediction. *Bioinformatics*, 20(18):3346–3352. Cité 1 fois, p. 5.

Albert, R. (2005). Scale-free networks in cell biology. *Journal of Cell Science*, 118(21):4947–4957. Cité 1 fois, p. 17.

Albert, R., Jeong, H. et Barabási, A.-L. (2000). Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382. Cité 1 fois, p. 17.

Arkin, M. R. et Wells, J. A. (2004). Small-molecule inhibitors of protein–protein interactions : Progressing towards the dream. *Nature Reviews Drug Discovery*, 3:301. Cité 1 fois, p. 8.

Ashtiani, M., Salehzadeh-Yazdi, A., Razaghi-Moghadam, Z., Hennig, H., Wolkenhauer, O., Mirzaie, M. et Jafari, M. (2018). A systematic survey of centrality measures for protein-protein interaction networks. *BMC Systems Biology*, 12(1). Cité 1 fois, p. 24.

Aulbert, W. et Kemper, M. J. (2016a). Severe antenatally diagnosed renal disorders : Background, prognosis and practical approach. *Pediatric Nephrology*, 31(4):563–574. Cité 1 fois, p. 63.

Aulbert, W. et Kemper, M. J. (2016b). Severe antenatally diagnosed renal disorders : Background, prognosis and practical approach. *Pediatric Nephrology*, 31(4):563–574. Cité 1 fois, p. 91.

Babur, Ö., Luna, A., Korkut, A., Durupinar, F., Siper, M. C., Dogrusoz, U., Aslan, J. E., Sander, C. et Demir, E. (2018). Causal interactions from proteomic profiles : Molecular data meets pathway knowledge. Preprint, Systems Biology. Cité 1 fois, p. 45.

Bader, G. D. et Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, page 27. Cité 1 fois, p. 32.

Bakun, M., Niemczyk, M., Domanski, D., Jazwiec, R., Perzanowska, A., Niemczyk, S., Kistowski, M., Fabijanska, A., Borowiec, A., Paczek, L. et Dadlez, M. (2012). Urine proteome of autosomal dominant polycystic kidney disease patients. *Clinical Proteomics*, 9(1):13. Cité 2 fois, p. 45 et 54.

Barabási, A.-L. (2007). Network Medicine — From Obesity to the "Diseasome". *New England Journal of Medicine*, 357(4):404–407. Cité 1 fois, p. 5.

Barabasi, A.-L. et Albert, R. (1999). Emergence of Scaling in Random Networks. 286:5. Cité 2 fois, p. 17 et 19.

Barabási, A.-L., Gulbahce, N. et Loscalzo, J. (2011). Network medicine : A network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68. Cité 2 fois, p. 38 et 39.

Barabási, A.-L. et Oltvai, Z. N. (2004). Network biology : Understanding the cell's functional organization. *Nature Reviews Genetics*, 5(2):101–113. Cité 1 fois, p. 17.

Barnes, J. A. (1954). Class and Committees in a Norwegian Island Parish. *Human Relations*, 7(1):39–58. Cité 1 fois, p. 43.

Barratt, J. et Topham, P. (2007). Urine proteomics : The present and future of measuring urinary protein components in disease. *Canadian Medical Association Journal*, 177(4):361–368. Cité 1 fois, p. 2.

Barrenas, F., Chavali, S., Holme, P., Mobini, R. et Benson, M. (2009). Network Properties of Complex Human Disease Genes Identified through Genome-Wide Association Studies. *PLoS ONE*, 4(11):e8090. Cité 2 fois, p. 26 et 27.

Barrios-Rodiles, M., Brown, K. R., Ozdamar, B., Bose, R., Liu, Z., Donovan, R. S., Shinjo, F., Liu, Y., Dembowy, J., Taylor, I. W., Luga, V., Przulj, N., Robinson, M., Suzuki, H., Hayashizaki, Y., Jurisica, I. et Wrana, J. L. (2005). High-Throughput Mapping of a Dynamic Signaling Network in Mammalian Cells. *Science*, 307(5715):1621–1625. Cité 1 fois, p. 9.

Basha, O., Barshir, R., Sharon, M., Lerman, E., Kirson, B. F., Hekselman, I. et Yeger-Lotem, E. (2017). The TissueNet v.2 database : A quantitative view of protein-protein interactions across human tissues. *Nucleic Acids Research*, 45(D1):D427–D431. Cité 1 fois, p. 35.

Belanger, K. D. (2009). Using Affinity Chromatography to Investigate Novel Protein–Protein Interactions in an Undergraduate Cell and Molecular Biology Lab Course. *CBE—Life Sciences Education*, 8(3):214–225. Cité 1 fois, p. 9.

Benjamini, Y. et Hochberg, Y. (1995). Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society : Series B (Methodological)*, 57(1):289–300. Cité 2 fois, p. 68 et 103.

Benson, D. P. F. et Sej, L. (1971). Enzyme activity in the amniotic fluid resides both in amniotic cells and in the cell-free supernatant. Current data indicate that cell-free amniotic fluid is not a reliable index of the enzyme status of the fetus (Nadler et al. 1970). *Section of Obstetrics & Gynacology*, 64:3. Cité 1 fois, p. 89.

Berggård, T., Linse, S. et James, P. (2007). Methods for the detection and analysis of protein–protein interactions. *PROTEOMICS*, 7(16):2833–2842. Cité 1 fois, p. 7.

Berry, S., Smith, R., Dombrowski, M., Puder, K., Cotton, D., Johnson, M., Kithier, K., Lecolier, B., Bercau, G. et Bidat, L. (1995). Predictive value of fetal serum $B2$-microglobulin for neonatal renal function. *The Lancet*, 345(8960):1277–1278. Cité 1 fois, p. 91.

Boizard, F., Brunchault, V., Moulos, P., Breuil, B., Klein, J., Lounis, N., Caubet, C., Tellier, S., Bascands, J.-L., Decramer, S., Schanstra, J. P. et Buffin-Meyer, B. (2016). A capillary electrophoresis coupled to mass spectrometry pipeline for long term comparable assessment of the urinary metabolome. *Scientific Reports*, 6(1). Cité 1 fois, p. 70.

Bonacich, P. (1987). Power and Centrality : A Family of Measures. *American Journal of Sociology*, 92(5):1170–1182. Cité 1 fois, p. 20.

Borgatti, S. P. (2005). Centrality and network flow. *Social Networks*, 27(1):55–71. Cité 1 fois, p. 21.

Bromberg, Y. (2013). Chapter 15 : Disease Gene Prioritization. *PLoS Computational Biology*, 9(4): e1002902. Cité 1 fois, p. 36.

Brosius, F. C. et Ju, W. (2018). The Promise of Systems Biology for Diabetic Kidney Disease. *Advances in Chronic Kidney Disease*, 25(2):202–213. Cité 1 fois, p. 27.

Brown, C. E., McCarthy, N. S., Hughes, A. D., Sever, P., Stalmach, A., Mullen, W., Dominiczak, A. F., Sattar, N., Mischak, H., Thom, S., Mayet, J., Stanton, A. V. et Delles, C. (2015). Urinary proteomic biomarkers to predict cardiovascular events. *PROTEOMICS - Clinical Applications*, 9(5-6):610–617. Cité 1 fois, p. 64.

Bruggeman, F. J. et Westerhoff, H. V. (2007). The nature of systems biology. *Trends in Microbiology*, 15(1):45–50. Cité 1 fois, p. 2.

Burt, R. S. (1976). Positions in Networks*. *Social Forces*, 55(1):93–122. Cité 1 fois, p. 18.

Calderon-Margalit, R., Skorecki, K. et Vivante, A. (2018). History of Childhood Kidney Disease and Risk of Adult End-Stage Renal Disease. *New England Journal of Medicine*, 378(18):1750–1752. Cité 1 fois, p. 90.

Chang, W., Cheng, J., Allaire, J. J., Xie, Y. et McPherson, J. (2019). *Shiny : Web Application Framework for R*. Cité 1 fois, p. 66.

Chavali, S., Barrenas, F., Kanduri, K. et Benson, M. (2010). Network properties of human disease genes with pleiotropic effects. *BMC Systems Biology*, 4(1):78. Cité 2 fois, p. 26 et 27.

Chen, C., Shen, H., Zhang, L.-G., Liu, J., Cao, X.-G., Yao, A.-L., Kang, S.-S., Gao, W.-X., Han, H., Cao, F.-H. et Li, Z.-G. (2016). Construction and analysis of protein-protein interaction networks based on proteomics data of prostate cancer. *International Journal of Molecular Medicine*, 37(6):1576–1586. Cité 1 fois, p. 33.

Chen, G., Li, Y., Su, Y., Zhou, L., Zhang, H., Shen, Q., Du, C., Li, H., Wen, Z., Xia, Y. et Tang, W. (2018). Identification of candidate genes for necrotizing enterocolitis based on microarray data. *Gene*, 661:152–159. Cité 3 fois, p. 32, 45 et 54.

Chindelevitch, L., Ziemek, D., Enayetallah, A., Randhawa, R., Sidders, B., Brockel, C. et Huang, E. S. (2012). Causal reasoning on biological networks : Interpreting transcriptional changes. *Bioinformatics*, 28(8):1114–1121. Cité 1 fois, p. 45.

Chua, H. N., Sung, W.-K. et Wong, L. (2006). Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics*, 22(13):1623–1630. Cité 1 fois, p. 12.

Cijiang He, J., Chuang, P. Y., Ma'Ayan, A. et Iyengar, R. (2012). Systems biology of kidney diseases. *Kidney International*, 81(1):22–39. Cité 2 fois, p. 27 et 28.

Coffman, E. et Richmond-Bryant, J. (2015). Multiple biomarker models for improved risk estimation of specific cardiovascular diseases related to metabolic syndrome : A cross-sectional study. *Population Health Metrics*, 13(1):7. Cité 2 fois, p. 64 et 105.

Cook, K. S., Emerson, R. M., Gillmore, M. R. et Yamagishi, T. (1983). The Distribution of Power in Exchange Networks : Theory and Experimental Results. *The American Journal of Sociology*, 89(2):275–305. Cité 1 fois, p. 18.

Corapi, K. M., Chen, J. L., Balk, E. M. et Gordon, C. E. (2012). Bleeding Complications of Native Kidney Biopsy : A Systematic Review and Meta-analysis. *American Journal of Kidney Diseases*, 60(1):62–73. Cité 1 fois, p. 2.

Costanzo, M. C. (2000). The Yeast Proteome Database (YPD) and Caenorhabditis elegans Proteome Database (WormPD) : Comprehensive resources for the organization and comparison of model organism protein information. *Nucleic Acids Research*, 28(1):73–76. Cité 1 fois, p. 15.

Csardi, G. et Nepusz, T. (2006). The igraph software package for complex network research. page 9. Cité 2 fois, p. 19 et 55.

Csősz, É., Kalló, G., Márkus, B., Deák, E., Csutak, A. et Tőzsér, J. (2017). Quantitative body fluid proteomics in medicine — A focus on minimal invasiveness. *Journal of Proteomics*, 153:30–43. Cité 2 fois, p. 2 et 64.

Danziger, P., Berman, D. R., Luckritz, K., Arbour, K. et Laventhal, N. (2016). Severe congenital anomalies of the kidney and urinary tract : Epidemiology can inform ethical decision-making. *Journal of Perinatology*, 36(11):954–959. Cité 2 fois, p. 97 et 99.

Davalieva, K., Kiprijanovska, S., Komina, S., Petrusevska, G., Zografska, N. C. et Polenakovic, M. (2015). Proteomics analysis of urine reveals acute phase response proteins as candidate diagnostic biomarkers for prostate cancer. *Proteome Science*, 13(1). Cité 1 fois, p. 28.

Davis, A. P., Grondin, C. J., Johnson, R. J., Sciaky, D., McMorran, R., Wiegers, J., Wiegers, T. C. et Mattingly, C. J. (2019). The Comparative Toxicogenomics Database : Update 2019. *Nucleic Acids Research*, 47(D1):D948–D954. Cité 1 fois, p. 56.

De Las Rivas, J. et de Luis, A. (2004). Interactome Data and Databases : Different Types of Protein Interaction. *Comparative and Functional Genomics*, 5(2):173–178. Cité 1 fois, p. 8.

De Las Rivas, J. et Fontanillo, C. (2010). Protein–Protein Interactions Essentials : Key Concepts to Building and Analyzing Interactome Networks. *PLoS Computational Biology*, 6(6):e1000807. Cité 4 fois, p. 7, 11, 14 et 15.

Decramer, S., de Peredo, A. G., Breuil, B., Mischak, H., Monsarrat, B., Bascands, J.-L. et Schanstra, J. P. (2008). Urine in Clinical Proteomics. *Molecular & Cellular Proteomics*, 7(10):1850–1862. Cité 1 fois, p. 45.

Decramer, S., Parant, O., Beaufils, S., Clauin, S., Guillou, C., Kessler, S., Aziza, J., Bandin, F., Schanstra, J. P. et Bellanné-Chantelot, C. (2007). Anomalies of the *TCF2* Gene Are the Main Cause of Fetal Bilateral Hyperechogenic Kidneys. *Journal of the American Society of Nephrology*, 18(3):923–933. Cité 1 fois, p. 91.

Decramer, S., Wittke, S., Mischak, H., Zürbig, P., Walden, M., Bouissou, F., Bascands, J.-L. et Schanstra, J. P. (2006). Predicting the clinical outcome of congenital unilateral ureteropelvic junction obstruction in newborn by urinary proteome analysis. *Nature Medicine*, 12(4):398–400. Cité 1 fois, p. 64.

Deng, M., Mehta, S., Sun, F. et Chen, T. (2002). Inferring Domain–Domain Interactions From Protein–Protein Interactions. *Protein Interactions*, page 9. Cité 1 fois, p. 12.

Desveaux, C., Klein, J., Leruez-Ville, M., Ramirez-Torres, A., Lacroix, C., Breuil, B., Froment, C., Bascands, J.-L., Schanstra, J. P. et Ville, Y. (2016). Identification of Symptomatic Fetuses Infected with Cytomegalovirus Using Amniotic Fluid Peptide Biomarkers. *PLOS Pathogens*, 12(1):e1005395. Cité 4 fois, p. 65, 94, 101 et 102.

Dezső, Z., Nikolsky, Y., Nikolskaya, T., Miller, J., Cherba, D., Webb, C. et Bugrim, A. (2009). Identifying disease-specific genes based on their topological significance in protein networks. *BMC Systems Biology*, 3(1):36. Cité 2 fois, p. 37 et 39.

Ding, F., Tan, A., Ju, W., Li, X., Li, S. et Ding, J. (2016). The Prediction of Key Cytoskeleton Components Involved in Glomerular Diseases Based on a Protein-Protein Interaction Network. *PLOS ONE*, 11(5):e0156024. Cité 2 fois, p. 34 et 35.

Doncheva, N. T., Morris, J. H., Gorodkin, J. et Jensen, L. J. J. (2018). Cytoscape stringApp : Network analysis and visualization of proteomics data. *bioRxiv*. Cité 1 fois, p. 16.

Dousdampanis, P., Trigka, K. et Fourtounas, C. (2012). Diagnosis and Management of Chronic Kidney Disease in the Elderly : A Field of Ongoing Debate. *Aging and Disease*, 3(5):13. Cité 1 fois, p. 63.

Dray, S. et Josse, J. (2015). Principal component analysis with missing values : A comparative survey of methods. *Plant Ecology*, 216(5):657–667. Cité 1 fois, p. 67.

Eckel-passow, J. E., Oberg, A. L., Therneau, T. M. et Bergen, H. R. (2009). An insight into high-resolution mass-spectrometry data. *Biostatistics*, 10(3):481–500. Cité 1 fois, p. 65.

Edwards, A. M., Kus, B., Jansen, R., Greenbaum, D., Greenblatt, J. et Gerstein, M. (2002). Bridging structural biology and genomics : Assessing protein interaction data with known complexes. *Trends in Genetics*, 18(10):529–536. Cité 1 fois, p. 33.

Eliasmith, C. (2015). Not really a philosopher. Cité 1 fois, p. 107.

Erten, S., Bebek, G., Ewing, R. M. et Koyutürk, M. (2011). DA DA : Degree-Aware Algorithms for Network-Based Disease Gene Prioritization. *BioData Mining*, 4(1). Cité 3 fois, p. 37, 39 et 41.

Estrada, E. (2006). Virtual identification of essential proteins within the protein interaction network of yeast. *PROTEOMICS*, 6(1):35–40. Cité 1 fois, p. 24.

Estrada, E. et Rodríguez-Velázquez, J. A. (2005). Subgraph centrality in complex networks. *Physical Review E*, 71(5):056103. Cité 2 fois, p. 20 et 24.

Estrada, E. et Ross, G. J. (2018). Centralities in simplicial complexes. Applications to protein interaction networks. *Journal of Theoretical Biology*, 438:46–60. Cité 1 fois, p. 24.

Feldman, I., Rzhetsky, A. et Vitkup, D. (2008). Network properties of genes harboring inherited disease mutations. *Proceedings of the National Academy of Sciences*, 105(11):4323–4328. Cité 1 fois, p. 26.

Fields, S. et Song, O.-k. (1989). A novel genetic system to detect protein–protein interactions. *Nature*, 340(6230):245–246. Cité 1 fois, p. 9.

Filip, S., Pontillo, C., Schanstra, J. P., Vlahou, A., Mischak, H. et Klein, J. (2014). Urinary proteomics and molecular determinants of chronic kidney disease : Possible link to proteases. *Expert Review of Proteomics*, 11(5):535–548. Cité 1 fois, p. 45.

Fliser, D., Novak, J., Thongboonkerd, V., Argilés, À., Jankowski, V., Girolami, M. A., Jankowski, J. et Mischak, H. (2007). Advances in Urinary Proteome Analysis and Biomarker Discovery. *Journal of the American Society of Nephrology*, 18(4):1057–1071. Cité 1 fois, p. 63.

Force, A., Lynch, M., Pickett, F. B., Amores, A., Yan, Y.-l. et Postlethwait, J. (1999). Preservation of Duplicate Genes by Complementary, Degenerative Mutations. page 16. Cité 1 fois, p. 17.

Fraser, H. B., Hirsh, A. E., Wall, D. P. et Eisen, M. B. (2004). Coevolution of gene expression among interacting proteins. *Proceedings of the National Academy of Sciences*, 101(24):9033–9038. Cité 1 fois, p. 12.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239. Cité 1 fois, p. 19.

Freeman, L. C. (2008). Going the Wrong Way on a One-Way Street :Centrality in Physics and Biology. page 15. Cité 1 fois, p. 18.

Fu, F., Wei, X., Liu, J. et Mi, N. (2015). Bioinformatic analysis of specific genes in diabetic nephropathy. *Renal Failure*, 37(7):1219–1224. Cité 2 fois, p. 30 et 32.

Garcia-Diaz, M. et Bebenek, K. (2007). Multiple Functions of DNA Polymerases. *Critical Reviews in Plant Sciences*, 26(2):105–122. Cité 1 fois, p. 7.

Garrels, J. (1996). YPD-A database for the proteins of Saccharomyces cerevisiae. *Nucleic Acids Research*, 24(1):46–49. Cité 1 fois, p. 13.

Gavin, A.-C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A.-M., Cruciat, C.-M., Remor, M., Höfert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.-A., Copley, R. R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G. et Superti-Furga, G. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147. Cité 1 fois, p. 17.

Ge, S., Wang, Y., Song, M., Li, X., Yu, X., Wang, H., Wang, J., Zeng, Q. et Wang, W. (2018). Type 2 Diabetes Mellitus : Integrative Analysis of Multiomics Data for Biomarker Discovery. *OMICS : A Journal of Integrative Biology*, 22(7):514–523. Cité 1 fois, p. 106.

Gehlenborg, N., O'Donoghue, S. I., Baliga, N. S., Goesmann, A., Hibbs, M. A., Kitano, H., Kohlbacher, O., Neuweger, H., Schneider, R., Tenenbaum, D. et Gavin, A.-C. (2010). Visualization of omics data for systems biology. *Nature Methods*, 7(S3):S56–S68. Cité 1 fois, p. 15.

Gertz, J., Elfond, G., Shustrova, A., Weisinger, M., Pellegrini, M., Cokus, S. et Rothschild, B. (2003). Inferring protein interactions from phylogenetic distance matrices. *Bioinformatics*, 19(16):2039–2045. Cité 1 fois, p. 12.

Ghasemi, M., Seidkhani, H., Tamimi, F., Rahgozar, M. et Masoudi-Nejad, A. (2014). Centrality Measures in Biological Networks. *Current Bioinformatics*, 9(4):426–441. Cité 1 fois, p. 21.

Giot, L. (2003). A Protein Interaction Map of Drosophila melanogaster. *Science*, 302(5651):1727–1736. Cité 1 fois, p. 17.

Glassock, R. J. (2015). Con : Kidney biopsy : An irreplaceable tool for patient management in nephrology. *Nephrology Dialysis Transplantation*, 30(4):528–531. Cité 1 fois, p. 2.

Goh, K.-I., Cusick, M. E., Valle, D., Childs, B., Vidal, M. et Barabasi, A.-L. (2007). The human disease network. *Proceedings of the National Academy of Sciences*, 104(21):8685–8690. Cité 2 fois, p. 26 et 27.

Gonzalez, M. W. et Kann, M. G. (2012). Chapter 4 : Protein Interactions and Disease. *PLoS Computational Biology*, 8(12):e1002819. Cité 2 fois, p. 7 et 8.

Grgic, I., Hofmeister, A. F., Genovese, G., Bernhardy, A. J., Sun, H., Maarouf, O. H., Bijol, V., Pollak, M. R. et Humphreys, B. D. (2014). Discovery of new glomerular disease–relevant genes by translational profiling of podocytes in vivo. *Kidney International*, 86(6):1116–1129. Cité 1 fois, p. 34.

Grijseels, E. W. M., van-Hornstra, P. E., Govaerts, L. C. P., Cohen-Overbeek, T. E., de Krijger, R. R., Smit, B. J. et Cransberg, K. (2011). Outcome of pregnancies complicated by oligohydramnios or anhydramnios of renal origin. *Prenatal Diagnosis*, 31(11):1039–1045. Cité 1 fois, p. 99.

Guala, D. et Sonnhammer, E. L. L. (2017). A large-scale benchmark of gene prioritization methods. *Scientific Reports*, 7(1). Cité 1 fois, p. 40.

Guespin-Michel, J. et Ripoll, C. (2000). La pluridisciplinarité dans les sciences de la vie : Un nouvel obstacle épistémologique, la non-linéarité. *Aster*, (30). Cité 1 fois, p. 1.

Hahn, M. W. et Kern, A. D. (2005). Comparative Genomics of Centrality and Essentiality in Three Eukaryotic Protein-Interaction Networks. *Molecular Biology and Evolution*, 22(4):803–806. Cité 2 fois, p. 21 et 22.

Hakes, L., Robertson, D. L., Oliver, S. G. et Lovell, S. C. (2007). Protein Interactions from Complexes : A Structural Perspective. *Comparative and Functional Genomics*, 2007:1–5. Cité 1 fois, p. 11.

Hamosh, A. (2004). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Research*, 33(Database issue):D514–D517. Cité 1 fois, p. 39.

Han, J.-D. J., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., Dupuy, D., Walhout, A. J. M., Cusick, M. E., Roth, F. P. et Vidal, M. (2004). Evidence for dynamically organized modularity in the yeast protein–protein interaction network. *Nature*, 430(6995):88–93. Cité 1 fois, p. 5.

Hanley, J. A. et McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36. Cité 1 fois, p. 104.

Hart, G. T., Ramani, A. K. et Marcotte, E. M. (2006). How complete are current yeast and human protein-interaction networks? *Genome Biology*, 7(11):120. Cité 1 fois, p. 11.

Hartwell, L. H., Hopfield, J. J., Leibler, S. et Murray, A. W. (1999). From molecular to modular cell biology. *Nature*, 402(6761supp):C47–C52. Cité 2 fois, p. 5 et 17.

He, L., Sun, Y., Takemoto, M., Norlin, J., Tryggvason, K., Samuelsson, T. et Betsholtz, C. (2008). The Glomerular Transcriptome and a Predicted Protein–Protein Interaction Network. *Journal of the American Society of Nephrology*, 19(2):260–268. Cité 2 fois, p. 33 et 34.

He, X. et Zhang, J. (2006). Why Do Hubs Tend to Be Essential in Protein Networks? *PLoS Genetics*, 2(6):9. Cité 1 fois, p. 19.

Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., von Mering, C., Roechert, B., Poux, S., Jung, E., Mersch, H., Kersey, P., Lappe, M., Li, Y., Zeng, R., Rana, D., Nikolski, M., Husi, H., Brun, C., Shanker, K., Grant, S. G. N., Sander, C., Bork, P., Zhu, W., Pandey, A., Brazma, A., Jacq, B., Vidal, M., Sherman, D., Legrain, P., Cesareni, G., Xenarios, I., Eisenberg, D., Steipe, B., Hogue, C. et Apweiler, R. (2004). The HUPO PSI's Molecular Interaction format—a community standard for the representation of protein interaction data. *Nature Biotechnology*, 22:177. Cité 1 fois, p. 13.

Hernandez-Ferrer, C. et Gonzalez, J. R. (2018). *CTDquerier : Package for CTDbase Data Query, Visualization and Downstream Analysis*. Cité 1 fois, p. 56.

Hill, N. R., Fatoba, S. T., Oke, J. L., Hirst, J. A., O'Callaghan, C. A., Lasserson, D. S. et Hobbs, F. D. R. (2016). Global Prevalence of Chronic Kidney Disease – A Systematic Review and Meta-Analysis. *PLOS ONE*, 11(7):e0158765. Cité 1 fois, p. 1.

Hindryckx, A. et De Catte, L. (2011). Prenatal diagnosis of congenital renal and urinary tract malformations. *Facts Views Vis Obgyn*, 3(3):165–174. Cité 1 fois, p. 63.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., Adams, S.-L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A. R., Sassi, H., Nielsen, P. A., Rasmussen, K. J., Andersen, J. R., Johansen, L. E., Hansen, L. H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sùrensen, B. D., Matthiesen, J., Hendrickson, R. C., Gleeson, F., Pawson, T., Moran, M. F., Durocher, D., Mann, M., Hogue, C. W. V., Figeys, D. et Tyers, M. (2002). Systematic identi®cation of protein complexes in Saccharomyces cerevisiae by mass spectrometry. 415:4. Cité 1 fois, p. 5.

Hogan, J., Dourthe, M.-E., Blondiaux, E., Jouannic, J.-M., Garel, C. et Ulinski, T. (2012). Renal outcome in children with antenatal diagnosis of severe CAKUT. *Pediatric Nephrology*, 27(3):497–502. Cité 2 fois, p. 91 et 99.

Hogan, J. J., Mocanu, M. et Berns, J. S. (2015). The Native Kidney Biopsy : Update and Evidence for Best Practice. *Clinical Journal of the American Society of Nephrology*, 11(2):354–362. Cité 1 fois, p. 2.

Hogan, M. C., Johnson, K. L., Zenka, R. M., Cristine Charlesworth, M., Madden, B. J., Mahoney, D. W., Oberg, A. L., Huang, B. Q., Leontovich, A. A., Nesbitt, L. L., Bakeberg, J. L., McCormick, D. J., Robert Bergen, H. et Ward, C. J. (2014). Subfractionation, characterization, and in-depth proteomic analysis of glomerular membrane vesicles in human urine. *Kidney International*, 85(5):1225–1237. Cité 1 fois, p. 28.

Holland, P. W. et Leinhardt, S. (1971). Transitivity in Structural Models of Small Groups. page 18. Cité 1 fois, p. 17.

Howe, D., Costanzo, M., Fey, P., Gojobori, T., Hannick, L., Hide, W., Hill, D. P., Kania, R., Schaeffer, M., St Pierre, S., Twigger, S., White, O. et Yon Rhee, S. (2008). The future of biocuration : Big data. *Nature*, 455(7209):47–50. Cité 1 fois, p. 13.

Hsu, C.-L., Huang, Y.-H., Hsu, C.-T. et Yang, U.-C. (2011). Prioritizing disease candidate genes by a gene interconnectedness-based approach. *BMC Genomics*, 12(Suppl 3):S25. Cité 5 fois, p. 37, 40, 46, 53 et 56.

Huang, H., Jedynak, B. M. et Bader, J. S. (2007). Where Have All the Interactions Gone ? Estimating the Coverage of Two-Hybrid Protein Interaction Maps. *PLoS Computational Biology*, 3(11):20. Cité 1 fois, p. 10.

Hwang, S., Kim, C. Y., Yang, S., Kim, E., Hart, T., Marcotte, E. M. et Lee, I. (2018). HumanNet v2 : Human gene networks for disease research. *Nucleic Acids Research*, 47(D1):D573–D580. Cité 1 fois, p. 41.

Hwang, S., Kim, C. Y., Yang, S., Kim, E., Hart, T., Marcotte, E. M. et Lee, I. (2019). HumanNet v2 : Human gene networks for disease research. *Nucleic Acids Research*, 47(D1):D573–D580. Cité 1 fois, p. 38.

Ideker, T. et Krogan, N. J. (2012). Differential network biology. *Molecular Systems Biology*, 8. Cité 2 fois, p. 33 et 59.

Ideker, T. et Sharan, R. (2008). Protein networks in disease. *Genome Research*, 18(4):644–652. Cité 1 fois, p. 32.

Iris, F., Gea, M., Lampe, P.-H. et Santamaria, P. (2009). Modélisation intégrative prédictive et biologie expérimentale : Un processus synergique remarquablement efficace au service de la recherche médicale. *médecine/sciences*, 25(6-7):608–616. Cité 1 fois, p. 59.

Ishii, K., Washio, T., Uechi, T., Yoshihama, M., Kenmochi, N. et Tomita, M. (2006). Characteristics and clustering of human ribosomal protein genes. *BMC Genomics*, page 16. Cité 1 fois, p. 7.

Jalili, M., Salehzadeh-Yazdi, A., Gupta, S., Wolkenhauer, O., Yaghmaie, M., Resendis-Antonio, O. et Alimoghaddam, K. (2016). Evolution of Centrality Measurements for the Detection of Essential Proteins in Biological Networks. *Frontiers in Physiology*, 7. Cité 2 fois, p. 21 et 24.

Jensen, L. J. et Bork, P. (2008). BIOCHEMISTRY : Not Comparable, But Complementary. *Science*, 322(5898):56–57. Cité 1 fois, p. 8.

Jeong, H., Mason, S. P., Barabási, A.-L. et Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, 411(6833):41–42. Cité 5 fois, p. 5, 17, 18, 20 et 24.

Jia, L., Fu, W., Jia, R., Wu, L., Li, X., Jia, Q. et Zhang, H. (2018). Identification of potential key protein interaction networks of BK virus nephropathy in patients receiving kidney transplantation. *Scientific Reports*, 8(1):5017. Cité 1 fois, p. 30.

Jia, L., Zhang, L., Shao, C., Song, E., Sun, W., Li, M. et Gao, Y. (2009). An Attempt to Understand Kidney's Protein Handling Function by Comparing Plasma and Urine Proteomes. *PLoS ONE*, 4(4): e5146. Cité 1 fois, p. 45.

Jimenez-Sanchez, G., Childs, B. et Valle, D. (2001). Human disease genes. *Nature*, 409(6822):853–855. Cité 1 fois, p. 24.

Jin, S., Wu, J., Zhu, Y., Gu, W., Wan, F., Xiao, W., Dai, B., Zhang, H., Shi, G., Shen, Y., Zhu, Y. et Ye, D. (2018). Comprehensive Analysis of *BAP1* Somatic Mutation in Clear Cell Renal Cell Carcinoma to Explore Potential Mechanisms *in Silico*. *Journal of Cancer*, 9(22):4108–4116. Cité 1 fois, p. 30.

Jonsson, P. F. et Bates, P. A. (2006). Global topological features of cancer proteins in the human interactome. *Bioinformatics*, 22(18):2291–2297. Cité 1 fois, p. 27.

Jouannic, J.-M., Hyett, J. A., Pandya, P. P., Gulbis, B., Rodeck, C. H. et Jauniaux, E. (2003). Perinatal outcome in fetuses with megacystis in the first half of pregnancy. *Prenatal Diagnosis*, 23(4): 340–344. Cité 1 fois, p. 99.

Joy, M. P., Brock, A., Ingber, D. E. et Huang, S. (2005). High-Betweenness Proteins in the Yeast Protein Interaction Network. *Journal of Biomedicine and Biotechnology*, 2005(2):96–103. Cité 1 fois, p. 22.

Kaiser, P., Meierhofer, D., Wang, X. et Huang, L. (2008). Tandem Affinity Purification Combined with Mass Spectrometry to Identify Components of Protein Complexes. *In* Walker, J., Starkey, M. et Elaswarapu, R., éditeurs : *Genomics Protocols*, volume 439, pages 309–326. Humana Press, Totowa, NJ. Cité 1 fois, p. 9.

Kanai, M., Raz, A. et Goodman, D. S. (1968). Retinol-binding protein : The transport protein for vitamin A in human plasma. *Journal of Clinical Investigation*, 47(9):2025–2044. Cité 1 fois, p. 7.

Kang, U., Papadimitriou, S., Sun, J. et Tong, H. (2011). Centralities in Large Networks : Algorithms and Observations. *In Proceedings of the 2011 SIAM International Conference on Data Mining*, pages 119–130. Society for Industrial and Applied Mathematics. Cité 1 fois, p. 18.

Kann, M. G. (2007). Protein interactions and disease : Computational approaches to uncover the etiology of diseases. *Briefings in Bioinformatics*, 8(5):333–346. Cité 1 fois, p. 26.

Kar, G., Gursoy, A. et Keskin, O. (2009). Human Cancer Protein-Protein Interaction Network : A Structural Perspective. *PLoS Computational Biology*, 5(12):e1000601. Cité 1 fois, p. 36.

Karinthy, F. (1929). CHAIN-LINKS. Cité 1 fois, p. 17.

Kenworthy, A. K. (2001). Imaging Protein-Protein Interactions Using Fluorescence Resonance Energy Transfer Microscopy. *Methods*, 24(3):289–296. Cité 2 fois, p. 9 et 10.

Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., Balakrishnan, L., Marimuthu, A., Banerjee, S., Somanathan, D. S., Sebastian, A., Rani, S., Ray, S., Harrys Kishore, C. J., Kanth, S., Ahmed, M., Kashyap, M. K., Mohmood, R., Ramachandra, Y. L., Krishna, V., Rahiman, B. A., Mohan, S., Ranganathan, P., Ramabadran, S., Chaerkady, R. et Pandey, A. (2009). Human Protein Reference Database–2009 update. *Nucleic Acids Research*, 37(Database):D767–D772. Cité 2 fois, p. 14 et 33.

Keskin, O., Tuncbag, N. et Gursoy, A. (2016). Predicting Protein–Protein Interactions from the Molecular to the Proteome Level. *Chemical Reviews*, 116(8):4884–4909. Cité 4 fois, p. 7, 8, 15 et 16.

Klein, J., Jupp, S., Moulos, P., Fernandez, M., Buffin-Meyer, B., Casemayou, A., Chaaya, R., Charonis, A., Bascands, J.-L., Stevens, R. et Schanstra, J. P. (2012). The KUPKB : A novel Web application to access multiomics data on kidney disease. *The FASEB Journal*, 26(5):2145–2153. Cité 1 fois, p. 16.

Klein, J., Lacroix, C., Caubet, C., Siwy, J., Zurbig, P., Dakna, M., Muller, F., Breuil, B., Stalmach, A., Mullen, W., Mischak, H., Bandin, F., Monsarrat, B., Bascands, J.-L., Decramer, S. et Schanstra, J. P. (2013). Fetal Urinary Peptides to Predict Postnatal Outcome of Renal Disease in Fetuses with Posterior Urethral Valves (PUV). *Science Translational Medicine*, 5(198):198ra106–198ra106. Cité 4 fois, p. 64, 65, 99 et 103.

Köhler, S., Bauer, S., Horn, D. et Robinson, P. N. (2008). Walking the Interactome for Prioritization of Candidate Disease Genes. *The American Journal of Human Genetics*, 82(4):949–958. Cité 5 fois, p. 37, 38, 40, 45 et 55.

Koschutzki, D. et Schreiber, F. (2004). Comparison of Centralities for Biological Networks. page 8. Cité 1 fois, p. 24.

Kotlyar, M., Pastrello, C., Malik, Z. et Jurisica, I. (2019). IID 2018 update : Context-specific physical protein–protein interactions in human, model organisms and domesticated species. *Nucleic Acids Research*, 47(D1):D581–D589. Cité 2 fois, p. 35 et 59.

Krämer, A., Green, J., Pollard, J. et Tugendreich, S. (2013). Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics*, 30(4):523–530. Cité 2 fois, p. 28 et 29.

Krämer, A., Green, J., Pollard, J. et Tugendreich, S. (2014). Causal analysis approaches in Ingenuity Pathway Analysis. *Bioinformatics*, 30(4):523–530. Cité 1 fois, p. 45.

Krnc, M., Sereni, J.-S., Škrekovski, R. et Yilma, Z. B. (2018). Eccentricity of networks with structural constraints. *Discussiones Mathematicae Graph Theory*. Cité 1 fois, p. 21.

Kumar, N., Nakagawa, P., Janic, B., Romero, C. A., Worou, M. E., Monu, S. R., Peterson, E. L., Shaw, J., Valeriote, F., Ongeri, E. M., Niyitegeka, J.-M. V., Rhaleb, N.-E. et Carretero, O. A. (2016). The anti-inflammatory peptide Ac-SDKP is released from thymosin-$B4$ by renal meprin-α and prolyl oligopeptidase. *American Journal of Physiology-Renal Physiology*, 310(10):F1026–F1034. Cité 1 fois, p. 102.

Lacroix, C., Caubet, C., Gonzalez-de-Peredo, A., Breuil, B., Bouyssié, D., Stella, A., Garrigues, L., Le Gall, C., Raevel, A., Massoubre, A., Klein, J., Decramer, S., Sabourdy, F., Bandin, F., Burlet-Schiltz, O., Monsarrat, B., Schanstra, J.-P. et Bascands, J.-L. (2014). Label-free Quantitative Urinary Proteomics Identifies the Arginase Pathway as a New Player in Congenital Obstructive Nephropathy. *Molecular & Cellular Proteomics*, 13(12):3421–3434. Cité 2 fois, p. 45 et 54.

Lan, W., Wang, J., Li, M., Peng, W. et Wu, F. (2015). Computational approaches for prioritizing candidate disease genes based on PPI networks. *Tsinghua Science and Technology*, 20(5):500–512. Cité 1 fois, p. 38.

Lander, A. D. (2010). The edges of understanding. *BMC Biology*, 8(1). Cité 2 fois, p. 15 et 16.

Lantos, J. D. et Warady, B. A. (2013). The evolving ethics of infant dialysis. *Pediatric Nephrology*, 28(10):1943–1947. Cité 1 fois, p. 98.

Lazar, C., Gatto, L., Ferro, M., Bruley, C. et Burger, T. (2016). Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies. *Journal of Proteome Research*, 15(4):1116–1125. Cité 1 fois, p. 67.

Levey, A. S. et Coresh, J. (2012). Chronic kidney disease. *The Lancet*, 379(9811):165–180. Cité 1 fois, p. 44.

Levey, A. S., Levin, A. et Kellum, J. A. (2013). Definition and Classification of Kidney Diseases. *American Journal of Kidney Diseases*, 61(5):686–688. Cité 1 fois, p. 44.

Li, M., Li, Q., Ganegoda, G. U., Wang, J., Wu, F. et Pan, Y. (2014). Prioritization of orphan disease-causing genes using topological feature and GO similarity between proteins in interaction networks. *Science China Life Sciences*, 57(11):1064–1071. Cité 2 fois, p. 37 et 38.

Liaw, A. et Wiener, M. (2002). Classification and Regression by randomForest. *R News*, 2(3):18–22. Cité 2 fois, p. 68 et 103.

Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A., Nardozza, A. P., Santonico, E., Castagnoli, L. et Cesareni, G. (2012). MINT, the molecular interaction database : 2012 update. *Nucleic Acids Research*, 40(D1):D857–D861. Cité 1 fois, p. 14.

Lindenmeyer, M. T., Eichinger, F., Sen, K., Anders, H.-J., Edenhofer, I., Mattinzoli, D., Kretzler, M., Rastaldi, M. P. et Cohen, C. D. (2010). Systematic Analysis of a Novel Human Renal Glomerulus-Enriched Gene Expression Dataset. *PLoS ONE*, 5(7):e11545. Cité 1 fois, p. 33.

Linder, E., Burguet, A., Nobili, F. et Vieux, R. (2018). Neonatal renal replacement therapy : An ethical reflection for a crucial decision. *Archives de Pédiatrie*, 25(6):371–377. Cité 1 fois, p. 97.

Loos, S. et Kemper, M. J. (2018). Causes of renal oligohydramnios : Impact on prenatal counseling and postnatal outcome. *Pediatric Nephrology*, 33(4):541–545. Cité 1 fois, p. 97.

Luke, D. A. et Harris, J. K. (2007). Network Analysis in Public Health : History, Methods, and Applications. *Annual Review of Public Health*, 28(1):69–93. Cité 1 fois, p. 5.

Ma, F., Sun, T., Wu, M., Wang, W. et Xu, Z. (2017). Identification of key genes for diabetic kidney disease using biological informatics methods. *Molecular Medicine Reports*, 16(6):7931–7938. Cité 3 fois, p. 30, 31 et 32.

Mabry, P. L., Olster, D. H., Morgan, G. D. et Abrams, D. B. (2008). Interdisciplinarity and Systems Science to Improve Population Health. *American Journal of Preventive Medicine*, 35(2):S211–S224. Cité 1 fois, p. 108.

Mackay, J., Sunde, M., Lowry, J., Crossley, M. et Matthews, J. (2007). Protein interactions : Is seeing believing ? *Trends in Biochemical Sciences*, 32(12):530–531. Cité 1 fois, p. 13.

Magalhães, P., Pejchinovski, M., Markoska, K., Banasik, M., Klinger, M., Švec-Billá, D., Rychlík, I., Rroji, M., Restivo, A., Capasso, G., Bob, F., Schiller, A., Ortiz, A., Perez-Gomez, M. V., Cannata, P., Sanchez-Niño, M. D., Naumovic, R., Brkovic, V., Polenakovic, M., Mullen, W., Vlahou, A., Zürbig, P., Pape, L., Ferrario, F., Denis, C., Spasovski, G., Mischak, H. et Schanstra, J. P. (2017). Association of kidney fibrosis with urinary peptides : A path towards non-invasive liquid biopsies ? *Scientific Reports*, 7(1):16915. Cité 1 fois, p. 99.

Maizi, M. (2017). *Le protéome urinaire : caractérisation et intérêt pour la recherche de biomarqueurs de pathologies*. Thèse de doctorat, Université de Grenoble. Cité 1 fois, p. 63.

Malan, V., Bussières, L., Winer, N., Jais, J.-P., Baptiste, A., Le Lorc'h, M., Elie, C., O'Gorman, N., Fries, N., Houfflin-Debarge, V., Sentilhes, L., Vekemans, M., Ville, Y., Salomon, L. J. et for the SAFE 21 Study Group (2018). Effect of Cell-Free DNA Screening vs Direct Invasive Diagnosis on Miscarriage Rates in Women With Pregnancies at High Risk of Trisomy 21 : A Randomized Clinical TrialEffect of Cell-Free DNA Screening on Miscarriage in Women With Pregnancies at High Risk of Trisomy 21Effect of Cell-Free DNA Screening on Miscarriage in Women With Pregnancies at High Risk of Trisomy 21. *JAMA*, 320(6):557–565. Cité 1 fois, p. 99.

Marcotte, C. J. V. et Marcotte, E. M. (2002). Predicting functional linkages from gene fusions with confidence. *Applied Bioinformatics*, page 8. Cité 1 fois, p. 12.

Maslov, S. (2002). Specificity and Stability in Topology of Protein Networks. *Science*, 296(5569):910–913. Cité 1 fois, p. 5.

Mehler, K., Gottschalk, I., Burgmaier, K., Volland, R., Büscher, A. K., Feldkötter, M., Keller, T., Weber, L. T., Kribs, A. et Habbig, S. (2018). Prenatal parental decision-making and postnatal outcome in renal oligohydramnios. *Pediatric Nephrology*, 33(4):651–659. Cité 2 fois, p. 91 et 97.

Menche, J., Sharma, A., Kitsak, M., Ghiassian, S. D., Vidal, M., Loscalzo, J. et Barabasi, A.-L. (2015). Uncovering disease-disease relationships through the incomplete interactome. *Science*, 347(6224):1257601–1257601. Cité 1 fois, p. 11.

Mewes, H. W., Frishman, D., Gruber, C., Geier, B., Haase, D., Kaps, A., Lemcke, K., Mannhaupt, G., Pfeiffer, F., Schüller, C., Stocker, S. et Weil, B. (2000). MIPS : A database for genomes and protein sequences. page 4. Cité 1 fois, p. 15.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A. et Leisch, F. (2019). *E1071 : Misc Functions of the Department of Statistics, Probability Theory Group (Formerly : E1071), TU Wien.* Cité 2 fois, p. 68 et 103.

Mischak, H. (2015). Pro : Urine proteomics as a liquid kidney biopsy : No more kidney punctures! *Nephrology Dialysis Transplantation*, 30(4):532–537. Cité 2 fois, p. 2 et 45.

Mischak, H., Allmaier, G., Apweiler, R., Attwood, T., Baumann, M., Benigni, A., Bennett, S. E., Bischoff, R., Bongcam-Rudloff, E., Capasso, G., Coon, J. J., D'Haese, P., Dominiczak, A. F., Dakna, M., Dihazi, H., Ehrich, J. H., Fernandez-Llama, P., Fliser, D., Frokiaer, J., Garin, J., Girolami, M., Hancock, W. S., Haubitz, M., Hochstrasser, D., Holman, R. R., Ioannidis, J. P. A., Jankowski, J., Julian, B. A., Klein, J. B., Kolch, W., Luider, T., Massy, Z., Mattes, W. B., Molina, F., Monsarrat, B., Novak, J., Peter, K., Rossing, P., Sanchez-Carbayo, M., Schanstra, J. P., Semmes, O. J., Spasovski, G., Theodorescu, D., Thongboonkerd, V., Vanholder, R., Veenstra, T. D., Weissinger, E., Yamamoto, T. et Vlahou, A. (2010a). Recommendations for Biomarker Identification and Qualification in Clinical Proteomics. *Science Translational Medicine*, 2(46):46ps42–46ps42. Cité 1 fois, p. 67.

Mischak, H., Allmaier, G., Apweiler, R., Attwood, T., Baumann, M., Benigni, A., Bennett, S. E., Bischoff, R., Bongcam-Rudloff, E., Capasso, G., Coon, J. J., D'Haese, P., Dominiczak, A. F., Dakna, M., Dihazi, H., Ehrich, J. H., Fernandez-Llama, P., Fliser, D., Frokiaer, J., Garin, J., Girolami, M., Hancock, W. S., Haubitz, M., Hochstrasser, D., Holman, R. R., Ioannidis, J. P. A., Jankowski, J., Julian, B. A., Klein, J. B., Kolch, W., Luider, T., Massy, Z., Mattes, W. B., Molina, F., Monsarrat, B., Novak, J., Peter, K., Rossing, P., Sanchez-Carbayo, M., Schanstra, J. P., Semmes, O. J., Spasovski, G., Theodorescu, D., Thongboonkerd, V., Vanholder, R., Veenstra, T. D., Weissinger, E., Yamamoto, T. et Vlahou, A. (2010b). Recommendations for Biomarker Identification and Qualification in Clinical Proteomics. *Science Translational Medicine*, 2(46):46ps42–46ps42. Cité 1 fois, p. 97.

Mischak, H., Delles, C., Vlahou, A. et Vanholder, R. (2015). Proteomic biomarkers in kidney disease : Issues in development and implementation. *Nature Reviews Nephrology*, 11(4):221–232. Cité 1 fois, p. 63.

Mischak, H., Vlahou, A. et Ioannidis, J. P. (2013). Technical aspects and inter-laboratory variability in native peptide profiling : The CE–MS experience. *Clinical Biochemistry*, 46(6):432–443. Cité 1 fois, p. 99.

Mistry, D., Wise, R. P. et Dickerson, J. A. (2017). DiffSLC : A graph centrality method to detect essential proteins of a protein-protein interaction network. *PLOS ONE*, 12(11):e0187091. Cité 1 fois, p. 24.

Miura, K. (2018). An Overview of Current Methods to Confirm Protein- Protein Interactions. *Protein & Peptide Letters*, 25(8):728–733. Cité 2 fois, p. 8 et 10.

Moons, K. G., Altman, D. G., Reitsma, J. B., Ioannidis, J. P., Macaskill, P., Steyerberg, E. W., Vickers, A. J., Ransohoff, D. F. et Collins, G. S. (2015). Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) : Explanation and ElaborationThe TRIPOD Statement : Explanation and Elaboration. *Annals of Internal Medicine*, 162(1):W1–W73. Cité 1 fois, p. 103.

Moons, K. G. M., Kengne, A. P., Grobbee, D. E., Royston, P., Vergouwe, Y., Altman, D. G. et Woodward, M. (2012). Risk prediction models : II. External validation, model updating, and impact assessment. *Heart*, 98(9):691–698. Cité 2 fois, p. 67 et 94.

Morris, J. S. et Baladandayuthapani, V. (2017). Statistical contributions to bioinformatics : Design, modelling, structure learning and integration. *Statistical Modelling : An International Journal*, 17(4-5):245–289. Cité 1 fois, p. 65.

Morris, R., Malin, G., Khan, K. et Kilby, M. (2009a). Antenatal ultrasound to predict postnatal renal function in congenital lower urinary tract obstruction : Systematic review of test accuracy : Antenatal ultrasound to predict postnatal renal function in LUTO. *BJOG : An International Journal of Obstetrics & Gynaecology*, 116(10):1290–1299. Cité 1 fois, p. 63.

Morris, R., Malin, G., Khan, K. et Kilby, M. (2009b). Antenatal ultrasound to predict postnatal renal function in congenital lower urinary tract obstruction : Systematic review of test accuracy : Antenatal ultrasound to predict postnatal renal function in LUTO. *BJOG : An International Journal of Obstetrics & Gynaecology*, 116(10):1290–1299. Cité 1 fois, p. 91.

Nair, V., Komorowsky, C. V., Weil, E. J., Yee, B., Hodgin, J., Harder, J. L., Godfrey, B., Ju, W., Boustany-Kari, C. M., Schwarz, M., Lemley, K. V., Nelson, P. J., Nelson, R. G. et Kretzler, M. (2018). A molecular morphometric approach to diabetic kidney disease can link structure to function and outcome. *Kidney International*, 93(2):439–449. Cité 1 fois, p. 28.

Navlakha, S. et Kingsford, C. (2010). The power of protein interaction networks for associating genes with diseases. *Bioinformatics*, 26(8):1057–1063. Cité 1 fois, p. 40.

Newman, M. J. (2005). A measure of betweenness centrality based on random walks. *Social Networks*, 27(1):39–54. Cité 1 fois, p. 20.

Ng, S.-K. et Tan, S.-H. (2004). DISCOVERING PROTEIN–PROTEIN INTERACTIONS. *Journal of Bioinformatics and Computational Biology*, 01(04):711–741. Cité 1 fois, p. 12.

Nicolaou, N., Renkema, K. Y., Bongers, E. M. H. F., Giles, R. H. et Knoers, N. V. A. M. (2015). Genetic, environmental, and epigenetic factors involved in CAKUT. *Nature Reviews Nephrology*, 11(12): 720–731. Cité 2 fois, p. 90 et 91.

Nicoll, R., Robertson, L., Gemmell, E., Sharma, P., Black, C. et Marks, A. (2018). Models of care for chronic kidney disease : A systematic review : Models of care for chronic kidney disease. *Nephrology*, 23(5):389–396. Cité 1 fois, p. 27.

Nitsch, D., Tranchevent, L.-C., Goncalves, J. P., Vogt, J. K., Madeira, S. C. et Moreau, Y. (2011). PINTA : A web server for network-based gene prioritization from expression data. *Nucleic Acids Research*, 39(suppl):W334–W338. Cité 1 fois, p. 41.

Nkuipou-Kenfack, E., Zürbig, P. et Mischak, H. (2017). The long path towards implementation of clinical proteomics : Exemplified based on CKD273. *PROTEOMICS - Clinical Applications*, 11(5-6):1600104. Cité 2 fois, p. 91 et 99.

Noble, D. (2006). *The Music of Life : Biology beyond the Genome*. Seuil. Cité 1 fois, p. 1.

Nooren, I. M. (2003). NEW EMBO MEMBER'S REVIEW : Diversity of protein-protein interactions. *The EMBO Journal*, 22(14):3486–3492. Cité 1 fois, p. 7.

O'Connell, M. R., Gamsjaeger, R. et Mackay, J. P. (2009). The structural analysis of protein–protein interactions by NMR spectroscopy. *PROTEOMICS*, 9(23):5224–5232. Cité 1 fois, p. 9.

Oliver, S. (2000). Guilt-by-association goes global. 403:3. Cité 1 fois, p. 36.

Oncley, J. L., Ellenbogen, E., Gitlin, D. et Gurd, F. R. N. (1952). Protein–Protein Interactions. *Protein Interactions*, 56:8. Cité 1 fois, p. 7.

Orchard, S., Kerrien, S., Abbani, S., Aranda, B., Bhate, J., Bidwell, S., Bridge, A., Briganti, L., Brinkman, F. S. L., Cesareni, G., Chatr-aryamontri, A., Chautard, E., Chen, C., Dumousseau, M., Goll, J., Hancock, R. E. W., Hannick, L. I., Jurisica, I., Khadake, J., Lynn, D. J., Mahadevan, U., Perfetto, L., Raghunath, A., Ricard-Blum, S., Roechert, B., Salwinski, L., Stümpflen, V., Tyers, M., Uetz, P., Xenarios, I. et Hermjakob, H. (2012). Protein interaction data curation : The International Molecular Exchange (IMEx) consortium. *Nature Methods*, 9(4):345–350. Cité 1 fois, p. 13.

Orchard, S., Kerrien, S., Jones, P., Ceol, A., Chatr-aryamontri, A., Salwinski, L., Nerothin, J. et Hermjakob, H. (2007). Submit Your Interaction Data the IMEx Way : A Step by Step Guide to Trouble-free Deposition. *PROTEOMICS*, 7(S1):28–34. Cité 1 fois, p. 13.

Östlund, G., Lindskog, M. et Sonnhammer, E. L. L. (2010). Network-based Identification of Novel Cancer Genes. *Molecular & Cellular Proteomics*, 9(4):648–655. Cité 1 fois, p. 41.

Oti, M. (2006). Predicting disease genes using protein-protein interactions. *Journal of Medical Genetics*, 43(8):691–698. Cité 5 fois, p. 37, 38, 39, 40 et 45.

Papanikolaou, N., Pavlopoulos, G. A., Theodosiou, T. et Iliopoulos, I. (2015). Protein–protein interaction predictions using text mining methods. *Methods*, 74:47–53. Cité 1 fois, p. 12.

Parikh, S. V., Malvar, A., Song, H., Alberton, V., Lococo, B., Vance, J., Zhang, J., Yu, L. et Rovin, B. H. (2015). Characterising the immune profile of the kidney biopsy at lupus nephritis flare differentiates early treatment responders from non-responders. *Lupus Science & Medicine*, 2(1): e000112–e000112. Cité 1 fois, p. 28.

Parker, M. (2003). Protein Structure from X-Ray Diffraction. *Journal of Biological Physics*, 29(4):341–362. Cité 1 fois, p. 10.

Pavlopoulos, G. A., Secrier, M., Moschopoulos, C. N., Soldatos, T. G., Kossida, S., Aerts, J., Schneider, R. et Bagos, P. G. (2011). Using graph theory to analyze biological networks. *BioData Mining*, 4(1):10. Cité 1 fois, p. 21.

Pawson, T. (2004). Specificity in Signal Transduction. *Cell*, 116(2):191–203. Cité 1 fois, p. 7.

Perisic, L., Lal, M., Hulkko, J., Hultenby, K., Önfelt, B., Sun, Y., Dunér, F., Patrakka, J., Betsholtz, C., Uhlen, M., Brismar, H., Tryggvason, K., Wernerson, A. et Pikkarainen, T. (2012). Plekhh2, a novel podocyte protein downregulated in human focal segmental glomerulosclerosis, is involved in matrix adhesion and actin dynamics. *Kidney International*, 82(10):1071–1083. Cité 1 fois, p. 34.

Petschnigg, J., Snider, J. et Stagljar, I. (2011). Interactive proteomics research technologies : Recent applications and advances. *Current Opinion in Biotechnology*, 22(1):50–58. Cité 1 fois, p. 8.

Phizicky, E. M. et Fields, S. (1995). Protein-Protein Interactions : Methods for Detection and Analysis. *MICROBIOL. REV.*, 59:30. Cité 2 fois, p. 7 et 9.

Pieper, R., Gatlin, C. L., McGrath, A. M., Makusky, A. J., Mondal, M., Seonarain, M., Field, E., Schatz, C. R., Estock, M. A., Ahmed, N., Anderson, N. G. et Steiner, S. (2004). Characterization of the human urinary proteome : A method for high-resolution display of urinary proteins on two-dimensional electrophoresis gels with a yield of nearly 1400 distinct protein spots. *PROTEOMICS*, 4(4):1159–1174. Cité 1 fois, p. 45.

Pirola, C. J. et Sookoian, S. (2018). Multiomics biomarkers for the prediction of nonalcoholic fatty liver disease severity. *World Journal of Gastroenterology*, 24(15):1601–1615. Cité 1 fois, p. 106.

Poret, A. et Guziolowski, C. (2018). Therapeutic target discovery using Boolean network attractors : Improvements of kali. *Royal Society Open Science*, 5(2):171852. Cité 1 fois, p. 60.

Rabieian, R., Abedi, M. et Gheisari, Y. (2017). Central Nodes in Protein Interaction Networks Drive Critical Functions in Transforming Growth Factor Beta-1 Stimulated Kidney Cells. *CELL JOURNAL*, 18(4):18. Cité 2 fois, p. 30 et 32.

Rao, V. S., Srinivas, K., Sujini, G. N. et Kumar, G. N. S. (2014). Protein-Protein Interaction Detection : Methods and Analysis. *International Journal of Proteomics*, 2014:1–12. Cité 2 fois, p. 8 et 10.

Rauniyar, N., Yu, X., Cantley, J., Voss, E. Z., Belcher, J., Colangelo, C. M., Stone, K. L., Dahl, N., Parikh, C., Lam, T. T. et Cantley, L. G. (2018). Quantification of Urinary Protein Biomarkers of Autosomal. *Proteomics clinical application*, 12(5). Cité 2 fois, p. 45 et 54.

Ren, J., Shang, L., Wang, Q. et Li, J. (2019). Ranking Cancer Proteins by Integrating PPI Network and Protein Expression Profiles. *BioMed Research International*, 2019:1–8. Cité 4 fois, p. 37, 45, 46 et 53.

Reznik, V. et Budorick, N. (1995). Prenatal detection of congenital renal disease. *The Urologic clinics of North America*, 22(1):21—30. Cité 1 fois, p. 63.

Richter, P. H. (1975). A network theory of the immune system. *European Journal of Immunology*, 5(5):350–354. Cité 1 fois, p. 15.

Rinschen, M. M., Huesgen, P. F. et Koch, R. E. (2018). The podocyte protease web : Uncovering the gatekeepers of glomerular disease. *American Journal of Physiology-Renal Physiology*, 315(6):F1812–F1816. Cité 1 fois, p. 34.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W. et Smyth, G. K. (2015). Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47–e47. Cité 1 fois, p. 57.

Robotti, E., Manfredi, M. et Marengo, E. (2013). Biomarkers Discovery through Multivariate Statistical Methods : A Review of Recently Developed Methods and Applications in Proteomics. *Journal of Proteomics & Bioinformatics*, s3. Cité 1 fois, p. 105.

Romagnani, P., Remuzzi, G., Glassock, R., Levin, A., Jager, K. J., Tonelli, M., Massy, Z., Wanner, C. et Anders, H.-J. (2017). Chronic kidney disease. *Nature Reviews Disease Primers*, 3:17088. Cité 2 fois, p. 27 et 63.

Rual, J.-F., Venkatesan, K., Hao, T., Hirozane-Kishikawa, T., Dricot, A., Li, N., Berriz, G. F., Gibbons, F. D., Dreze, M., Ayivi-Guedehoussou, N., Klitgord, N., Simon, C., Boxem, M., Milstein, S., Rosenberg, J., Goldberg, D. S., Zhang, L. V., Wong, S. L., Franklin, G., Li, S., Albala, J. S., Lim, J., Fraughton, C., Llamosas, E., Cevik, S., Bex, C., Lamesch, P., Sikorski, R. S., Vandenhaute, J., Zoghbi, H. Y., Smolyar, A., Bosak, S., Sequerra, R., Doucette-Stamm, L., Cusick, M. E., Hill, D. E., Roth, F. P. et Vidal, M. (2005). Towards a proteome-scale map of the human protein–protein interaction network. *Nature*, 437(7062):1173–1178. Cité 2 fois, p. 5 et 15.

Ruiz, J. et Philippe, J. (2012). La médecine personnalisée et la face cachée de la lune. *Revue Médicale Suisse*, 8:1195–1196. Cité 1 fois, p. 109.

Rusinowska, A., Berghammer, R., De Swart, H. et Grabisch, M. (2011). Social Networks : Prestige, Centrality, and Influence. *In* de Swart, H., éditeur : *Relational and Algebraic Methods in Computer Science*, volume 6663, pages 22–39. Springer Berlin Heidelberg, Berlin, Heidelberg. Cité 1 fois, p. 20.

Ryckewaert-D'Halluin, A., Le Bouar, G., Odent, S., Milon, J., D'Hervé, D., Lucas, J., Rouget, F., Loget, P., Poulain, P., Le Gall, E. et Taque, S. (2011). Diagnosis of fetal urinary tract malformations : Prenatal management and postnatal outcome. *Prenatal Diagnosis*, 31(11):1013–1020. Cité 1 fois, p. 99.

Safari-Alighiarloo, N., Taghizadeh, M., Rezaei-Tavirani, M., Goliaei, B. et Peyvandi, A. A. (2014). Protein-protein interaction networks (PPI) and complex diseases. page 15. Cité 1 fois, p. 8.

Sasaki, H., Sasaki, N., Nishino, T., Nagasaki, K.-i., Kitamura, H., Torigoe, D. et Agui, T. (2014). Quantitative Trait Loci for Resistance to the Congenital Nephropathy in Tensin 2-Deficient Mice. *PLoS ONE*, 9(6):e99602. Cité 1 fois, p. 34.

Schanstra, J. P., Zürbig, P., Alkhalaf, A., Argiles, A., Bakker, S. J., Beige, J., Bilo, H. J., Chatzikyrkou, C., Dakna, M., Dawson, J., Delles, C., Haller, H., Haubitz, M., Husi, H., Jankowski, J., Jerums, G., Kleefstra, N., Kuznetsova, T., Maahs, D. M., Menne, J., Mullen, W., Ortiz, A., Persson, F., Rossing, P., Ruggenenti, P., Rychlik, I., Serra, A. L., Siwy, J., Snell-Bergeon, J., Spasovski, G., Staessen, J. A., Vlahou, A., Mischak, H. et Vanholder, R. (2015). Diagnosis and Prediction of CKD Progression by Assessment of Urinary Peptides. *Journal of the American Society of Nephrology*, 26(8):1999–2010. Cité 2 fois, p. 65 et 99.

Schrohl, A.-S., Würtz, S., Kohn, E., Banks, R. E., Nielsen, H. J., Sweep, F. C. G. J. et Brünner, N. (2008). Banking of Biological Fluids for Studies of Disease-associated Protein Biomarkers. *Molecular & Cellular Proteomics*, 7(10):2061–2066. Cité 1 fois, p. 64.

Schuster-Böckler, B. et Bateman, A. (2008). Protein interactions in human genetic diseases. *Genome Biology*, 9(1):R9. Cité 1 fois, p. 8.

Schwartz, G. J., Muñoz, A., Schneider, M. F., Mak, R. H., Kaskel, F., Warady, B. A. et Furth, S. L. (2009). New Equations to Estimate GFR in Children with CKD. *Journal of the American Society of Nephrology*, 20(3):629–637. Cité 1 fois, p. 100.

Schwikowski, B., Uetz, P. et Fields, S. (2000). A network of protein–protein interactions in yeast. *Nature Biotechnology*, 18(12):1257–1261. Cité 2 fois, p. 5 et 15.

Sevimoglu, T. et Arga, K. Y. (2014). The role of protein interaction networks in systems biomedicine. *Computational and Structural Biotechnology Journal*, 11(18):22–27. Cité 1 fois, p. 15.

Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., Li, Y. et Jiang, H. (2007). Predicting protein-protein interactions based only on sequences information. *Proceedings of the National Academy of Sciences*, 104(11):4337–4341. Cité 1 fois, p. 12.

Simões, S. N., Martins, D. C., Pereira, C. A., Hashimoto, R. F. et Brentani, H. (2015). NERI : Network-medicine based integrative approach for disease gene prioritization by relative importance. *BMC Bioinformatics*, 16(S19):S9. Cité 1 fois, p. 37.

Simões, S. N., Martins-Jr, D. C., Brentani, H. et Fumio, R. (2012). Shortest Paths Ranking Methodology to Identify Alterations in PPI Networks of Complex Diseases. *In Proceedings of the ACM Conference on Bioinformatics, Computational Biology and Biomedicine*, BCB '12, pages 561–563. ACM. Cité 2 fois, p. 45 et 55.

Singh, A., Gautier, B., Shannon, C. P., Rohart, F., Vacher, M., Tebutt, S. J. et Le Cao, K.-A. (2016). DIABLO : From multi-omics assays to biomarker discovery, an integrative approach. Preprint, Bioinformatics. Cité 1 fois, p. 106.

Smyth, M. S. et Martin, J. H. J. (2000). X Ray crystallography. *Molecular Pathology*, 53(1):8–14. Cité 1 fois, p. 9.

Snider, J., Kotlyar, M., Saraon, P., Yao, Z., Jurisica, I. et Stagljar, I. (2015). Fundamentals of protein interaction network mapping. *Molecular Systems Biology*, 11(12):848–848. Cité 1 fois, p. 8.

Snyder, M. (2009). Untangling the protein web. 460:4. Cité 1 fois, p. 13.

Spaggiari, E., Faure, G., Dreux, S., Czerkiewicz, I., Stirnemann, J. J., Guimiot, F., Heidet, L., Favre, R., Salomon, L. J., Oury, J. F., Ville, Y. et Muller, F. (2017a). Sequential fetal serum $B2$-microglobulin to predict postnatal renal function in bilateral or low urinary tract obstruction : Sequential fetal serum β 2-microglobulin. *Ultrasound in Obstetrics & Gynecology*, 49(5):617–622. Cité 1 fois, p. 63.

Spaggiari, E., Faure, G., Dreux, S., Czerkiewicz, I., Stirnemann, J. J., Guimiot, F., Heidet, L., Favre, R., Salomon, L. J., Oury, J. F., Ville, Y. et Muller, F. (2017b). Sequential fetal serum $B2$-microglobulin to predict postnatal renal function in bilateral or low urinary tract obstruction : Sequential fetal serum β 2-microglobulin. *Ultrasound in Obstetrics & Gynecology*, 49(5):617–622. Cité 2 fois, p. 91 et 99.

Sprinzak, E. et Margalit, H. (2001). Correlated sequence-signatures as markers of protein-protein interaction. *Journal of Molecular Biology*, 311(4):681–692. Cité 1 fois, p. 12.

Stelzl, U., Worm, U., Lalowski, M., Haenig, C., Brembeck, F. H., Goehler, H., Stroedicke, M., Zenkner, M., Schoenherr, A., Koeppen, S., Timm, J., Mintzlaff, S., Abraham, C., Bock, N., Kietzmann, S., Goedde, A., Toksöz, E., Droege, A., Krobitsch, S., Korn, B., Birchmeier, W., Lehrach, H. et Wanker, E. E. (2005). A Human Protein-Protein Interaction Network : A Resource for Annotating the Proteome. *Cell*, 122(6):957–968. Cité 1 fois, p. 15.

Stevens, L. A. et Levey, A. S. (2009). Measured GFR as a Confirmatory Test for Estimated GFR. *Journal of the American Society of Nephrology*, 20(11):2305–2313. Cité 1 fois, p. 63.

Stevenson, D. K. et Goldworth, A. (1998). Ethical dilemmas in the delivery room. *Seminars in Perinatology*, 22(3):198–206. Cité 1 fois, p. 98.

Strimbu, K. et Tavel, J. A. (2010). What are biomarkers ? :. *Current Opinion in HIV and AIDS*, 5(6):463–466. Cité 1 fois, p. 63.

Stumpf, M. P. H., Thorne, T., de Silva, E., Stewart, R., An, H. J., Lappe, M. et Wiuf, C. (2008). Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences*, 105(19):6959–6964. Cité 1 fois, p. 11.

Stynen, B., Tournu, H., Tavernier, J. et Van Dijck, P. (2012). Diversity in Genetic In Vivo Methods for Protein-Protein Interaction Studies : From the Yeast Two-Hybrid System to the Mammalian Split-Luciferase System. *Microbiology and Molecular Biology Reviews*, 76(2):331–382. Cité 1 fois, p. 8.

Suderman, M. et Hallett, M. (2007). Tools for visually exploring biological networks. *Bioinformatics*, 23(20):2651–2659. Cité 1 fois, p. 15.

Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., Kuhn, M., Bork, P., Jensen, L. J. et von Mering, C. (2015). STRING v10 : Protein–protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43(D1):D447–D452. Cité 1 fois, p. 54.

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., Jensen, L. J. et von Mering, C. (2019). STRING v11 : Protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Research*, 47(D1):D607–D613. Cité 2 fois, p. 18 et 31.

Theodorescu, D., Wittke, S., Ross, M. M., Walden, M., Conaway, M., Just, I., Mischak, H. et Frierson, H. F. (2006). Discovery and validation of new protein biomarkers for urothelial cancer : A prospective analysis. page 11. Cité 1 fois, p. 102.

Thomas, R., Kanso, A. et Sedor, J. R. (2008). Chronic Kidney Disease and Its Complications. *Primary Care : Clinics in Office Practice*, 35(2):329–344. Cité 1 fois, p. 44.

Thomas, S., Hao, L., Ricke, W. A. et Li, L. (2016). Biomarker discovery in mass spectrometry-based urinary proteomics. *PROTEOMICS - Clinical Applications*, 10(4):358–370. Cité 1 fois, p. 63.

Tomaszewski, M., Eales, J., Denniff, M., Myers, S., Chew, G. S., Nelson, C. P., Christofidou, P., Desai, A., Büsst, C., Wojnar, L., Musialik, K., Jozwiak, J., Debiec, R., Dominiczak, A. F., Navis, G., van Gilst, W. H., van der Harst, P., Samani, N. J., Harrap, S., Bogdanski, P., Zukowska-Szczechowska, E. et Charchar, F. J. (2015). Renal Mechanisms of Association between Fibroblast Growth Factor 1 and Blood Pressure. *Journal of the American Society of Nephrology*, 26(12):3151–3160. Cité 1 fois, p. 33.

Trairatphisan, P., Wiesinger, M., Bahlawane, C., Haan, S. et Sauter, T. (2016). A Probabilistic Boolean Network Approach for the Analysis of Cancer-Specific Signalling : A Case Study of Deregulated PDGF Signalling in GIST. *PLOS ONE*, 11(5):e0156223. Cité 1 fois, p. 60.

Tranchevent, L.-C., Capdevila, F. B., Nitsch, D., De Moor, B., De Causmaecker, P. et Moreau, Y. (2011). A guide to web tools to prioritize candidate genes. *Briefings in Bioinformatics*, 12(1):22–32. Cité 1 fois, p. 36.

Tucker, C. L. et Fields, S. (2003). Lethal combinations. *Nature Genetics*, 35(3):204–205. Cité 2 fois, p. 9 et 10.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S. et Rothberg, J. M. (2000). A comprehensive analysis of protein–protein interactions inSaccharomyces cerevisiae. *Nature*, 403(6770):623–627. Cité 1 fois, p. 36.

Uhlen, M., Fagerberg, L., Hallstrom, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Sivertsson, A., Kampf, C., Sjostedt, E., Asplund, A., Olsson, I., Edlund, K., Lundberg, E., Navani, S., Szigyarto, C. A.-K., Odeberg, J., Djureinovic, D., Takanen, J. O., Hober, S., Alm, T., Edqvist, P.-H., Berling, H., Tegel, H., Mulder, J., Rockberg, J., Nilsson, P., Schwenk, J. M., Hamsten, M., von Feilitzen, K., Forsberg, M., Persson, L., Johansson, F., Zwahlen, M., von Heijne, G., Nielsen, J. et Ponten, F. (2015). Tissue-based map of the human proteome. *Science*, 347(6220):1260419–1260419. Cité 2 fois, p. 5 et 6.

Vakser, I. A. (2014). Protein-Protein Docking : From Interaction to Interactome. *Biophysical Journal*, 107(8):1785–1793. Cité 2 fois, p. 11 et 12.

Valdeolivas, A., Tichit, L., Navarro, C., Perrin, S., Odelin, G., Levy, N., Cau, P., Remy, E. et Baudot, A. (2019). Random walk with restart on multiplex and heterogeneous biological networks. *Bioinformatics*, 35(3):497–505. Cité 1 fois, p. 55.

Vanunu, O. et Sharan, R. (2010). A Propagation-based Algorithm for Inferring Gene-Disease Associations. page 10. Cité 4 fois, p. 37, 39, 40 et 41.

Venables, W. N. et Ripley, B. D. (2002). Modern Applied Statistics with S. page 504. Cité 1 fois, p. 103.

Venkatesan, K., Rual, J.-F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K.-I., Yildirim, M. A., Simonis, N., Heinzmann, K., Gebreab, F., Sahalie, J. M., Cevik, S., Simon, C., de Smet, A.-S., Dann, E., Smolyar, A., Vinayagam, A., Yu, H., Szeto, D., Borick, H., Dricot, A., Klitgord, N., Murray, R. R., Lin, C., Lalowski, M., Timm, J., Rau, K., Boone, C., Braun, P., Cusick, M. E., Roth, F. P., Hill, D. E., Tavernier, J., Wanker, E. E., Barabási, A.-L. et Vidal, M. (2009). An empirical framework for binary interactome mapping. *Nature Methods*, 6(1):83–90. Cité 1 fois, p. 11.

Voss, J., Young Ah Goo, Cain, K., Woods, N., Jarrett, M., Smith, L., Shulman, R. et Heitkemper, M. (2011). Searching for the Noninvasive Biomarker Holy Grail : Are Urine Proteomics the Answer ? *Biological Research For Nursing*, 13(3):235–242. Cité 1 fois, p. 2.

Wang, J., Chen, G., Li, M. et Pan, Y. (2011). Integration of breast cancer gene signatures based on graph centrality. *BMC Systems Biology*, 5(Suppl 3):S10. Cité 1 fois, p. 27.

Wang, P., Lu, J. et Yu, X. (2014). Identification of Important Nodes in Directed Biological Networks : A Network Motif Approach. *PLOS ONE*, 9(8):15. Cité 1 fois, p. 24.

Wang, R.-S., Saadatpour, A. et Albert, R. (2012). Boolean modeling in systems biology : An overview of methodology and applications. *Physical Biology*, 9(5):055001. Cité 1 fois, p. 60.

Warsow, G., Endlich, N., Schordan, E., Schordan, S., Chilukoti, R. K., Homuth, G., Moeller, M. J., Fuellen, G. et Endlich, K. (2013). PodNet, a protein–protein interaction network of the podocyte. *Kidney International*, 84(1):104–115. Cité 1 fois, p. 34.

Warsow, G., Greber, B., Falk, S. S., Harder, C., Siatkowski, M., Schordan, S., Som, A., Endlich, N., Schöler, H., Repsilber, D., Endlich, K. et Fuellen, G. (2010). ExprEssence - Revealing the essence of differential experimental data in the context of an interaction/regulation net-work. *BMC Systems Biology*, 4(1):164. Cité 1 fois, p. 34.

Watts, D. J. et Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. 393:3. Cité 1 fois, p. 17.

Wei, R., Wang, J., Su, M., Jia, E., Chen, S., Chen, T. et Ni, Y. (2018). Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. *Scientific Reports*, 8(1):663. Cité 1 fois, p. 67.

Wei, W., Lv, Y., Gan, Z., Zhang, Y., Han, X. et Xu, Z. (2019). Identification of key genes involved in the metastasis of clear cell renal cell carcinoma. *Oncology Letters*. Cité 1 fois, p. 30.

White, H. C., Boorman, S. A. et Breiger, R. L. (1976). Social Structure from Multiple Networks. I. Blockmodels of Roles and Positions. *American Journal of Sociology*, 81(4):730–780. Cité 1 fois, p. 18.

Wool, I. G. (1979). The Structure and Function of Eukaryotic Ribosomes. *Annual Review of Biochemistry*, 48(1):719–754. Cité 1 fois, p. 7.

Wuchty, S. et Stadler, P. F. (2003). Centers of complex networks. *Journal of Theoretical Biology*, 223(1):45–53. Cité 2 fois, p. 21 et 24.

Wühl, E., van Stralen, K. J., Verrina, E., Bjerre, A., Wanner, C., Heaf, J. G., Zurriaga, O., Hoitsma, A., Niaudet, P., Palsson, R., Ravani, P., Jager, K. J. et Schaefer, F. (2013). Timing and Outcome of Renal Replacement Therapy in Patients with Congenital Malformations of the Kidney and Urinary Tract. *Clinical Journal of the American Society of Nephrology*, 8(1):67–74. Cité 1 fois, p. 90.

Wulff, C. B., Gerds, T. A., Rode, L., Ekelund, C. K., Petersen, O. B., Tabor, A. et the Danish Fetal Medicine Study Group (2016). Risk of fetal loss associated with invasive testing following combined first-trimester screening for Down syndrome : A national cohort of 147 987 singleton pregnancies : Procedure-related risk of fetal loss. *Ultrasound in Obstetrics & Gynecology*, 47(1):38–44. Cité 1 fois, p. 99.

Wüthrich, K. (2001). The way to NMR structures of proteins. *nature structural biology*, 8(11):3. Cité 1 fois, p. 10.

Yeger-Lotem, E., Sattath, S., Kashtan, N., Itzkovitz, S., Milo, R., Pinter, R. Y., Alon, U. et Margalit, H. (2004). Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *Proceedings of the National Academy of Sciences*, 101(16):5934–5939. Cité 2 fois, p. 17 et 20.

Yook, S.-H., Oltvai, Z. N. et Barabási, A.-L. (2004). Functional and topological characterization of protein interaction networks. *PROTEOMICS*, 4(4):928–942. Cité 1 fois, p. 17.

Yu, H., Braun, P., Yildirim, M. A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., Hao, T., Rual, J.-F., Dricot, A., Vazquez, A., Murray, R. R., Simon, C., Tardivo, L., Tam, S., Svrzikapa, N., Fan, C., de Smet, A.-S., Motyl, A., Hudson, M. E., Park, J., Xin, X., Cusick, M. E., Moore, T., Boone, C., Snyder, M., Roth, F. P., Barabasi, A.-L., Tavernier, J., Hill, D. E. et Vidal, M. (2008). High-Quality Binary Protein Interaction Map of the Yeast Interactome Network. *Science*, 322(5898):104–110. Cité 1 fois, p. 15.

Yu, H., Greenbaum, D., Lu, H. X., Zhu, X. et Gerstein, M. (2004). Genomic analysis of essentiality within protein networks. *Trends in Genetics*, 20(6):227–231. Pagination error in this issue, see Publisher's note in Vol. 21 issue 1 p. 36. Cité 1 fois, p. 18.

Yu, H., Kim, P. M., Sprecher, E., Trifonov, V. et Gerstein, M. (2007). The Importance of Bottlenecks in Protein Networks : Correlation with Gene Essentiality and Expression Dynamics. *PLoS Computational Biology*, 3(4):8. Cité 1 fois, p. 22.

Zachwieja, J., Soltysiak, J., Fichna, P., Lipkowska, K., Stankiewicz, W., Skowronska, B., Kroll, P. et Lewandowska-Stachowiak, M. (2010). Normal-range albuminuria does not exclude nephropathy in diabetic children. *Pediatric Nephrology*, 25(8):1445–1451. Cité 1 fois, p. 63.

Zanotti, G., Folli, C., Cendron, L., Alfieri, B., Nishida, S. K., Gliubich, F., Pasquato, N., Negro, A. et Berni, R. (2008). Structural and mutational analyses of protein-protein interactions between transthyretin and retinol-binding protein : Transthyretin-retinol-binding protein interactions. *FEBS Journal*, 275(23):5841–5854. Cité 1 fois, p. 7.

Zhan, X., Zhou, T., Cheng, T. et Lu, M. (2019). Recognition of Multiomics-Based Molecule-Pattern Biomarker for Precise Prediction, Diagnosis, and Prognostic Assessment in Cancer. *In Bioinformatics Tools for Detection and Clinical Interpretation of Genomic Variations [Working Title]*. IntechOpen. Cité 1 fois, p. 106.

Zhang, Q.-L. et Rothenbacher, D. (2008). Prevalence of chronic kidney disease in population-based studies : Systematic review. *BMC Public Health*, 8(1):117. Cité 1 fois, p. 1.

Zhao, J., Yang, T.-H., Huang, Y. et Holme, P. (2011). Ranking Candidate Disease Genes from Gene Expression and Protein Interaction : A Katz-Centrality Based Approach. *PLoS ONE*, 6(9):e24306. Cité 3 fois, p. 37, 38 et 40.

Zhao, M., Li, M., Yang, Y., Guo, Z., Sun, Y., Shao, C., Li, M., Sun, W. et Gao, Y. (2017). A comprehensive analysis and annotation of human normal urinary proteome. *Scientific Reports*, 7(1). Cité 3 fois, p. 5, 6 et 45.

Zhou, L.-T., Qiu, S., Lv, L.-L., Li, Z.-L., Liu, H., Tang, R.-N., Ma, K.-L. et Liu, B.-C. (2018). Integrative Bioinformatics Analysis Provides Insight into the Molecular Mechanisms of Chronic Kidney Disease. *Kidney and Blood Pressure Research*, 43(2):568–581. Cité 1 fois, p. 30.

Zhu, C., Kushwaha, A., Berman, K. et Jegga, A. G. (2012). A vertex similarity-based framework to discover and rank orphan disease-related genes. *BMC Systems Biology*, 6(Suppl 3):S8. Cité 1 fois, p. 37.

Zinzalla, G. et Thurston, D. E. (2009). Targeting protein–protein interactions for therapeutic intervention : A challenge for the future. *Future Medicinal Chemistry*, 1(1):65–93. Cité 1 fois, p. 8.

Zotenko, E., Mestre, J., O'Leary, D. P. et Przytycka, T. M. (2008). Why Do Hubs in the Yeast Protein Interaction Network Tend To Be Essential : Reexamining the Connection between the Network Topology and Essentiality. *PLoS Computational Biology*, 4(8):e1000140. Cité 1 fois, p. 24.

Liste des figures

"Les chiffres, c'est pas une science exacte figurez-vous!"
Karadoc de Vannes (2007)

Liste des tables

"Réduire sa vie à un tableau de bord chiffré est une fuite du rapport complexe que l'on peut entretenir avec la qualité."

Alain Damasio (2014)