UNIVERSITÉ DU QUÉBEC À MONTRÉAL

CARACTÉRISATION DES RÉSEAUX D'INTERACTIONS FONCTIONNELLES DES GÈNES CHEZ LE NÉMATODE ET L'HOMME GRÂCE À LA GÉNOMIQUE INTÉGRATIVE

THÈSE

PRÉSENTÉE

COMME EXIGENCE PARTIELLE

DU DOCTORAT EN BIOCHIMIE

PAR

BENJAMIN BOUCHER

NOVEMBRE 2017

REMERCIEMENTS

Je tiens à remercier chaleureusement toutes les personnes qui m'ont aidé pendant l'élaboration de ma thèse et notamment, ma directrice Sarah Jenna, pour son soutien, sa bravoure et sa ténacité tout au long de cette longue aventure.

Ce travail n'aurait pu être mené à bien sans le soutien et l'aide de mon co-directeur Abdoulaye Baniré Diallo et de son équipe de recherche que j'ai appréciée grandement. Je veux aussi remercier le professeur Michael Hallett de l'Université McGill, pour m'avoir accueilli dans son laboratoire de recherche. Un merci tout spécial aux membres de son équipe de recherche que j'ai côtoyé et sans qui ce projet n'aurait pas eu lieu, en particulier Anna Y. Lee.

Finalement, je remercie celles et ceux qui me sont chers. Leur soutien, attention et encouragements m'ont accompagné tout au long de ces années.

DÉDICACE

À mes enfants,

Élie et Justine.

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

LISTE DES ABRÉVIATIONS, SIGLES ET ACRONYMES

ADN	Acide désoxyribonucléique
ARN	Acide ribonucléique
ARNdb	Acide ribonucléique double brin
AUC	Area under the curve
BasalL	Basal-like breast cancer
BRCA1/2	Breast cancer type 1 susceptibility protein
Cas9	CRISPR associated protein 9
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
DNA	Desoxyribonucleic acid
EGF	Epidermal growth factor
E-MAP	Epistatic miniarray profiles
ERBB2	Human epidermal growth factor receptor 2
GI	Genetic interaction
GO	Gene Ontology
GWAS	Genome-wide association studies
HER2	ERBB2-enriched breast cancer
IF	Interaction fonctionnelle
IG	Interaction génétique
IPP	Interaction protéine-protéine
KEGG	Kyoto Encyclopedia of Genes and Genomes
Log	Logarithm
LOOCV	Leave-one-out cross-validation
LumA	Luminal A breast cancer
LumB	Luminal B breast cancer
mRNA	Messenger ribonucleic acid

MAP	Mitogen-activated
Min	Minimum
NormL	Normal-like breast cancer
OXPHOS	Oxydative phosphorylation
PARP	Poly (adenosine diphosphate)-ribose polymerase
PDS	Protein-protein interactions dense subnetworks
PPI	Protein-protein interaction
RNA	Ribonucleic acid
RNAi	Ribonucleic acid interference
RNA-seq	Ribonucleic acid sequencing
ROC	Receiver operating characteristic
RUNX	Runt-related transcription factor
SGA	Synthetic genetic arrays
shRNA	Short hairpin ribonucleic acid
siRNA	Small interfering ribonucleic acid
SSL	Synthetic sick or lethal
TCGA	The Cancer Genome Atlas
TP53	Tumor protein p53

RÉSUMÉ

Les interactions fonctionnelles (IFs) entre les gènes décrivent en quelque sorte comment les gènes s'influencent mutuellement pour régir le dynamisme des systèmes biologiques. La caractérisation des IFs et, en particulier, de leurs réseaux, est donc d'une importance capitale pour élucider les mécanismes moléculaires complexes liant les gènes chez les organismes modèles et chez l'homme. Afin de mieux comprendre le fonctionnement de ces réseaux, nous avons développé une approche de génomique intégrative pour caractériser les IFs, comprenant les interactions génétiques (IGs), des organismes vivants. Parallèlement, nous avons aussi investigué la prédiction des IGs par des méthodes *in silico* et sommes venus à la conclusion qu'une meilleure compréhension des liens entre différents niveaux de données serait bénéfique pour la prédiction des IGs. D'abord, la caractérisation d'un réseau d'IGs du nématode *Caenorhabditis elegans* nous a permis de mettre en évidence une structure formée de six classes d'IGs ayant des propriétés distinctes. Cette étude révèle deux classes d'IGs, la première apparaissant au centre des complexes d'interactions protéine-protéine et l'autre, à l'extérieur de ces complexes. Notre étude suggère aussi que la coordination des modules fonctionnels, définis par ces deux classes d'IGs, impliquerait des gènes hautement connectés dans le réseau et liés à leurs partenaires fonctionnels via deux autres classes d'IG. Les gènes définis par ces différentes classes d'IG présentent des caractéristiques distinctes lorsque l'on considère leur connectivité au sein du réseau d'IG, mais aussi au sein du réseau d'interactions protéine-protéine, leur pléiotropie et leur redondance fonctionnelle. Nous avons également montré qu'il était possible de transposer notre méthode d'intégration chez l'humain. Plus précisément nous avons utilisé notre approche afin de caractériser un réseau d'IFs pour deux sous-types de cancer du sein. Nous avons ensuite identifié, dans ce réseau, différentes classes d'IFs ayant des caractéristiques similaires aux classes d'IGs de *C. elegans*. Nous avons utilisé ces classes d'IFs, ainsi que leur niveau de plasticité entre les sous-types de cancer du sein, pour construire un modèle prédictif servant à découvrir des gènes potentiellement impliqués dans la reprogrammation métabolique des cellules cancéreuses, et ce, à l'échelle du génome. Dans l'ensemble, le modèle d'interactome fonctionnel, construit par notre étude, pourrait permettre de mieux comprendre le rôle des IFs au niveau de la robustesse des systèmes biologiques, et le contrôle qu'elles exercent sur les processus biologiques. Notre approche pourrait aussi permettre de découvrir de nouvelles cibles thérapeutiques pour le cancer du sein.

MOTS-CLÉS : interaction fonctionnelle, interaction génétique, génomique intégrative, *Caenorhabditis elegans*, cancer du sein.

ABSTRACT

Functional interactions (FIs) between genes describe how genes influence each other to govern the dynamics of biological systems. The characterization of FIs, and their networks, is therefore very important to elucidate the complex molecular mechanisms linking genes in model organisms and in humans. To better understand the functioning of these networks, we have developed an integrative genomic approach to characterize FIs, including genetic interactions (GIs), of living organisms. At the same time, we have also investigated the prediction of GIs using *in silico* methods and have concluded that a better understanding of the links between different levels of data would be beneficial for the prediction of GIs. First, the characterization of a GI network in the nematode *Caenorhabditis elegans* showed a structure formed of six classes of GIs with distinct properties. This study reveals two classes of GIs, one appearing at the center of the protein-protein interactions complexes and the other, appearing outside of these complexes. Our study also suggests that the coordination of functional modules, defined by these two classes of GIs, would involve highly connected genes in the network linked to their functional partners via two other GI classes. The genes defined by these different GI classes have distinctive characteristics when considering their connectivity within the GI network, but also within the protein-protein interactions network, their pleiotropy and their functional redundancy. Also, we have shown that it is possible to transpose our method of integration to humans. More precisely, we used our approach to characterize a network of FIs for two subtypes of breast cancer. We then identified, in this network, different FI classes having similar characteristics with the GI classes of *C. elegans*. We used these classes of FIs, as well as their level of plasticity between breast cancer subtypes, to construct a predictive model to discover genes potentially involved in the metabolic reprogramming of cancer cells on a genome-wide scale. Overall, the functional interactome model, constructed in our study, could be useful to understand the role of FIs in the robustness of biological systems and their control over biological processes. Our approach could also reveal new therapeutic targets for breast cancer.

KEYWORDS: functional interaction, genetic interaction, integrative genomic, *Caenorhabditis elegans*, breast cancer.

CHAPITRE I

INTRODUCTION

1.1 Les interactions génétiques

Les interactions génétiques entre deux ou plusieurs gènes sont des relations entre ces gènes dont le mécanisme n'est pas toujours connu ou est issu d'une cascade d'interactions moléculaires (contact physique entre les produits des gènes) qui ont un effet sur le phénotype. Pour se rendre jusqu'à l'expression d'un phénotype, une cellule emploie une machinerie abondamment complexe dont le point de départ est l'information contenue dans le génotype. Cette information est fixe et héritée alors que du côté du phénotype, il est variable et pourrait n'avoir aucun lien intuitif avec l'information provenant du génotype. Ainsi, une des plus grandes difficultés de la génétique moderne consiste en la compréhension de la relation entre le génotype et le phénotype d'un individu. Cette relation peut être illustrée par la génétique des maladies humaines puisque les pathologies ne sont pas réellement mendélienne, *i.e.* les maladies ne sont pas causées par un seul gène dont l'altération le rend inopérant. Le phénotype principal d'une maladie résulte plus souvent de l'effet combiné d'une multitude de gènes, mais aussi d'autres facteurs (*e.g.* sexe, ethnicité, habitudes de vie) en plus des mécanismes épigénétiques qui modulent l'expression des gènes (Gallati, 2014). Mais alors que les mécanismes épigénétiques sont réversibles et grandement influencés par l'environnement, il est encore très difficile de définir leurs rôles et leurs contributions dans la progression d'une maladie.

À l'inverse, comme le génome d'un individu ne change pas, l'effet combiné de plusieurs gènes mutés (ou variants) sur la sévérité de la maladie peut être mesuré pour identifier de nouveaux acteurs. Par exemple, la mutation du gène CFTR associée a longtemps été vue comme l'unique responsable des symptômes cliniques de la fibrose kystique, mais il y a aujourd'hui plus d'une trentaine de mutations de gènes modificateurs (gènes qui modulent l'expression d'autres gènes) donnant lieu à des aggravations ou atténuations des symptômes de la maladie (Gallati, 2014). Par ailleurs, dans le cancer du sein, la communauté scientifique a longtemps cru que les patients ayant une mutation du gène BRCA1 ou BRCA2, une forme souvent héréditaire du cancer du sein, aient une survie beaucoup plus courte que ceux ne possédant pas de mutation. Or, une revue systématique de 66 études sur le sujet a montré qu'avec la prise d'adjuvant (utilisé en chimioprévention), les patients aux prises avec une mutation BRCA1/2 ont une survie à peu près égale à celle des autres patients atteints de cancer du sein prenant ou pas d'adjuvant (van den Broek *et al.*, 2015). Ceci démontre bien à quel point, même en présence d'une mutation d'un gène associé à la maladie, il est possible de balancer l'effet négatif de la mutation par la modulation d'un ou plusieurs autres gènes modificateurs.

Le lien qui uni les gènes modificateurs avec la modulation d'un autre gène est communément appelé interaction génétique (IG). Une IG survient lorsque le produit de deux ou plusieurs gènes variants est un phénotype inattendu (Bateson *et al.*, 1905). De façon plus générale, une IG est identifiée s'il y a une différence entre les phénotypes mesurés pour le double-mutants et ceux attendus pour chaque mutation individuelle (DixonCostanzo, *et al.*, 2009). Bien sûr, selon la nature de cette différence, une IG sera considérée comme négative si le phénotype du double mutant est plus marqué, à l'inverse, elle sera positive si le phénotype est moins sévère. Plus précisément, une interaction négative peut, à l'extrême, être létale si la combinaison de deux mutations dans des gènes compris dans deux voies compensatoires (« Between pathway, nonessential pathways », Figure 1.1a), ou à l'intérieur d'une voie essentielle (« Within

pathway, essential pathways », Figure 1.1a) conduit à un phénotype non-viable (DixonCostanzo, *et al.*, 2009). De façon similaire, une interaction positive survient par perte-de-fonction si la mutation dans un deuxième gène d'une voie annule la létalité induite du gène en amont (« Loss-of-function », Figure 1.1a), ou encore par gain-de-fonction, si le deuxième gène muté d'une voie est situé en aval de celle-ci et qu'elle annule la létalité d'un gène en amont (« Gain-of-function », Figure 1.1a). Mais le concept entourant une IG peut devenir sensiblement plus complexe lorsque le phénotype étudié ne résulte pas en létalité. En effet, le niveau quantitatif d'un phénotype en particulier découle d'un apport multifactoriel de perturbations génétiques hautement dynamiques et interconnectés entre les différents réseaux biologiques responsables, par exemple, du développement, de la transcription, du métabolisme ou de la signalisation (Fox, 1993). Il est donc nécessaire de bien définir le rôle que jouent les IG au sein de la relation entre le génotype et le phénotype d'un organisme ainsi que ses différents liens avec les autres sources de perturbations génétiques.

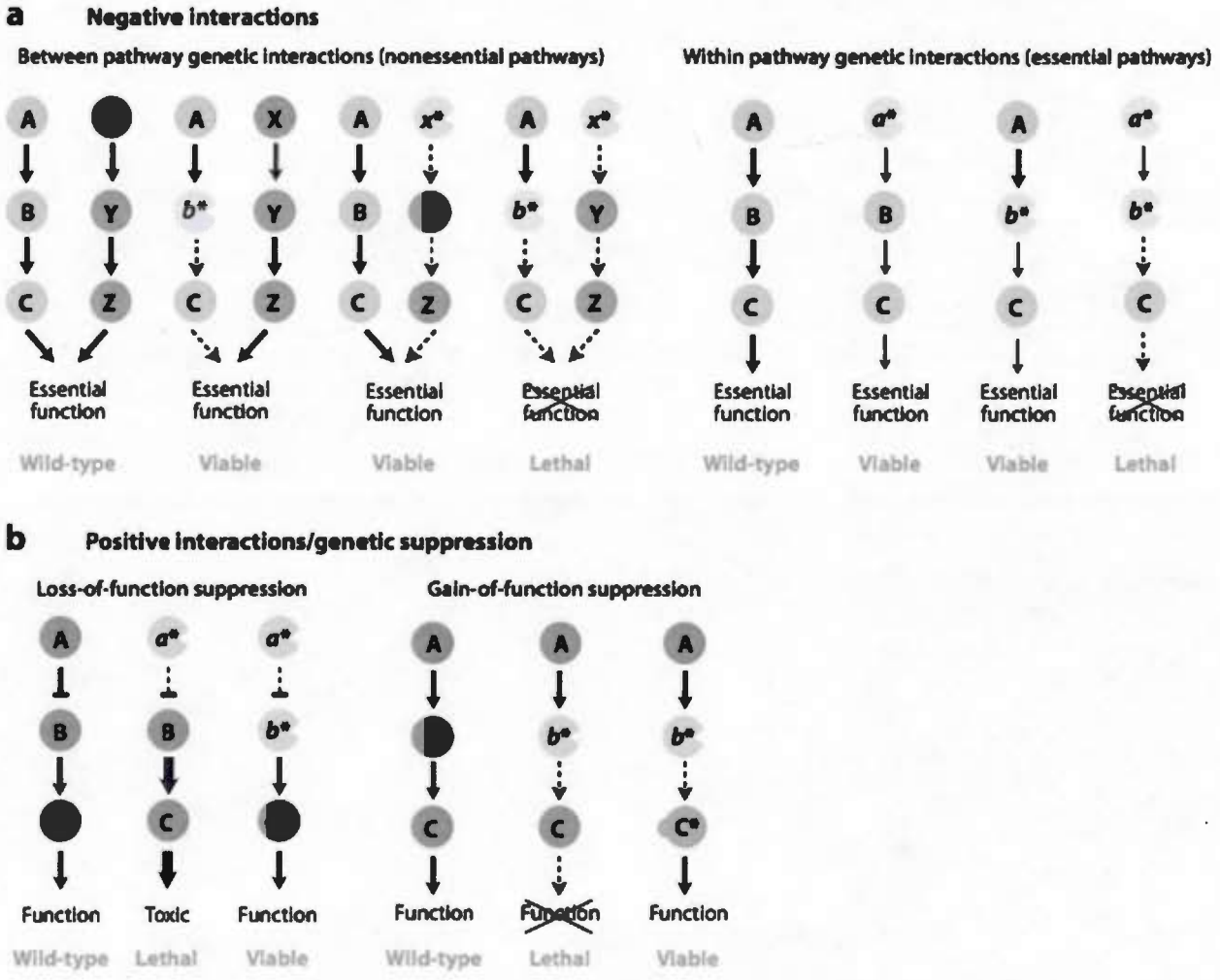


Figure 1.1 Modèles moléculaires pour définir les interactions génétiques.

(a) Les interactions négatives entre deux gènes peuvent se produire à l'intérieur d'une même voie essentielle (« Within pathway ») ou entre deux voies parallèles non-essentielles (« Between pathway »). **(b)** La mutation d'un inhibiteur d'une voie cytotoxique entraîne l'accumulation de produits toxiques pour la cellule. Dans ce cas, une interaction positive survient si une deuxième mutation annule ou diminue l'activité de la voie cytotoxique par perte-de-fonction (« Loss-of-function »). Autrement, une interaction positive par gain-de-fonction (« Gain-of-function ») survient lorsque la mutation d'un gène situé en aval d'une voie, annule la létalité d'un gène en situé en amont. Image tirée de (DixonCostanzo, *et al.*, 2009).

1.2 Caractérisation des réseaux d'interactions des systèmes biologiques

Les réseaux (structure de graphes) sont régulièrement exploités comme support graphique pour l'étude et la modélisation des interactions biologiques (Winterbach *et al.*, 2013). Cette représentation permet d'avoir une vue d'ensemble permettant d'analyser un petit sous-ensemble d'interactions en tenant compte des propriétés et des informations globales qui sont habituellement plus facile à expliquer (Barabasi et Albert, 1999). Ainsi, la caractérisation des réseaux biologiques, aussi appelée biologie des réseaux, s'effectue à travers une combinaison de techniques et principes issues de la biologie des systèmes, la théorie des graphes et les analyses statistiques et computationnelles (Barabasi et Oltvai, 2004). Cette caractérisation est associée à la description de la topologie et la définition des principales propriétés des réseaux d'interactions biologiques. Parmi ces propriétés, l'une des plus notables est la distribution des degrés des nœuds du réseau qui suit une loi de puissance donnant, au final, un réseau invariant d'échelle (« Scale-free », Figure 1.2b) dont la structure se situe entre le réseau au hasard (« Random », Figure 1.2a) et le réseau hiérarchique (« Hierarchical », Figure 1.2c). Les réseaux invariants d'échelle ont été décrits pour plusieurs systèmes biologiques dont le métabolisme du nématode *Caenorhabditis elegans* (Jeong *et al.*, 2000), la co-expression des gènes chez *Saccharomyces cerevisiae* (van Noort *et al.*, 2004) et les interactions protéine-protéine (IPP) chez *Homo sapiens* (Stelzl *et al.*, 2005). Dans ce type de structure de réseau, la majorité des noeuds ont seulement une ou deux connections alors que quelques noeuds ont un grand nombre de connections (Figure 1.2b). À l'inverse, les réseaux construits au hasard sont homogènes avec une majorité de noeuds ayant à peu près le même nombre de connections (Figure 1.2a). L'avantage du réseau invariant d'échelle par rapport au réseau construit au hasard est qu'il offre une robustesse contre les perturbations faites sur la plupart des noeuds alors que seulement quelques points (les plus connectés) sont plus sensibles aux

perturbations (Barabasi et Oltvai, 2004). À l'opposée, les noeuds d'un réseau construit au hasard sont tous susceptibles d'être fragiles face aux perturbations.

Ensuite, le fait que les réseaux biologiques soient invariants d'échelles met en lumière une autre propriété qui est l'existence de noeuds hautement connectés appelés « Hub ». Ce sont d'ailleurs ces Hubs qui sont les points faibles des réseaux biologiques, démontré, entre-autre, par leur propension à être létales lorsque leurs gènes sont mutés dans le réseau d'IPP de la levure (Goymer, 2008), ou encore à être oncogénique lorsque mutés dans le cancer (Collavin *et al.*, 2010).

Enfin, une autre propriété commune des réseaux biologiques est le phénomène du petit monde (« small-world ») où chacune des paires de noeuds ont relativement un court chemin entres eux (Barabasi et Oltvai, 2004). Par exemple, les métabolites sont, la plupart du temps, éloignés l'un de l'autre par seulement quelques réactions enzymatiques (Jeong *et al.*, 2000). Cette dernière est aussi intimement liée à une autre propriété des réseaux biologiques nommée « modularité ». Dans celle-ci, il est montré que les réseaux ont un haut degré de regroupement (« clustering ») et que des modules (motifs densément connectés) existent dans des régions locales du réseau (Barabasi et Oltvai, 2004). En effet, en absence de modularité, le degré de regroupements des réseaux réels (*e.g.* sociaux, transports, télécommunications) est comparable à celui des réseaux construits au hasard (Barabasi et Oltvai, 2004). Il est donc raisonnable de déduire que les réseaux biologiques ont une modularité élevée puisqu'un haut degré de regroupement a été observé dans les réseaux métaboliques (Ravasz *et al.*, 2002), neuronaux (Sporns, 2014) et d'IPP (Wagner, 2001).

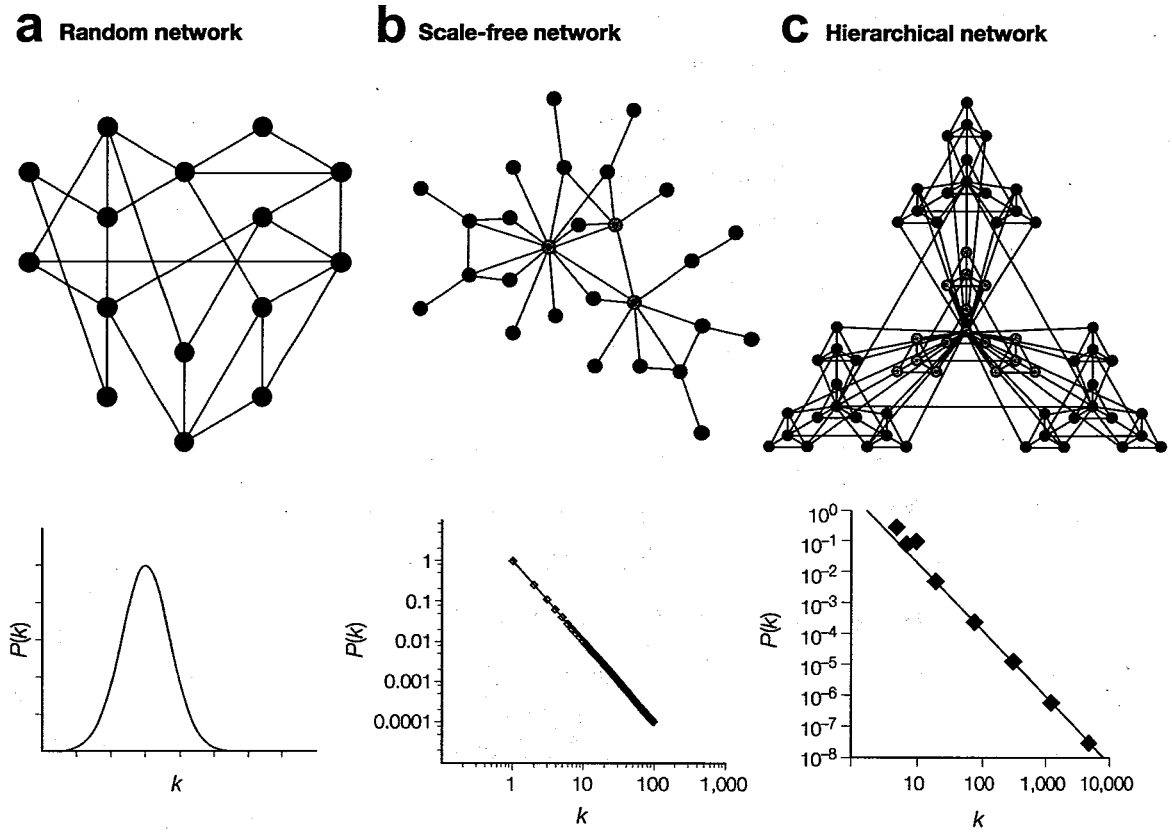


Figure 1.2 Schémas et distributions des degrés de connections dans les réseaux.

La distribution *P* des degrés *k* des nœuds du réseau suit une distribution de Poisson dans les réseaux construits au hasard **(a)** et une loi de puissance dans les réseaux à échelles invariants **(b)**. **(c)** Les réseaux hiérarchiques sont des exemples extrêmes de réseaux invariants d'échelles. Image tirée de (Barabasi et Oltvai, 2004).

1.2.1 Propriétés des réseaux d'interactions protéine-protéine

En plus des propriétés communes aux réseaux biologiques, les réseaux d'IPP possèdent des caractéristiques spécifiques associées au contexte dans lequel ils sont étudiés.

Parmi ces caractéristiques, les protéines hautement connectées (ou Hubs) dans le réseau d'IPP de la levure ont été montrées comme étant codées, la plupart du temps, par des gènes essentiels (Jeong *et al.*, 2001; Yu *et al.*, 2008). Aussi, pour mieux définir ces protéines, qui occupent des positions centrales dans l'interactome protéine-protéine, Yu *et al.* (2007) ont introduit la notion de protéines « bottleneck ». Ces dernières sont des protéines avec un haut degré de centralité intermédiaire (ou « betweenness »). En analogie avec le rôle d'un pont ou d'un tunnel dans le réseau routier, les protéines « bottleneck » représentent les points par lesquels il y a un maximum de chemins les plus courts entres ces derniers et l'ensemble des noeuds dans le réseau d'interactions protéine-protéine (Figure 1.3) (Yu *et al.*, 2007). De manière générale, les protéines « bottleneck » sont un meilleur indicateur d'essentialité que le degré d'IPP et sont plus faiblement co-exprimés avec leurs partenaires que les protéines « non-bottleneck » (Yu *et al.*, 2007). De plus, cette notion permet de classer certaines protéines du réseau d'IPP en 4 types de nœuds (Figure 1.3) (Yu *et al.*, 2007);

i. selon qu'ils soient des Hubs et des protéines « bottleneck » (« Hub-bottleneck »);
ii. ou pas (« Hub-non-bottleneck »);
iii. qu'ils ne soient pas des Hubs mais des protéines « bottleneck » (« Non-hub-bottleneck ») ou
iv. pas (« Non-hub-non-bottleneck »).

Ce modèle offre la possibilité d'utiliser le degré de centralité intermédiaire pour découvrir des protéines qui ne sont pas des Hubs, mais qui jouent un rôle clé entre deux processus biologiques importants. Plus récemment, Wuchty (2014) a étendu cette notion à la connexion entre deux protéines « bottleneck ». Il a montré que le degré de centralité intermédiaire de l'interaction entre deux protéines « bottleneck » est elle aussi supérieur à la moyenne des autres interactions (Wuchty, 2014). De plus, en confirmant cette propriété dans le réseau d'IPP de l'humain, il a trouvé que les protéines qui partagent une interaction à haut degré de centralité intermédiaire sont plus souvent

associées avec le cancer et sont aussi plus souvent ciblées par des virus que les protéines qui ne partagent pas ce type d'interaction (Wuchty, 2014).

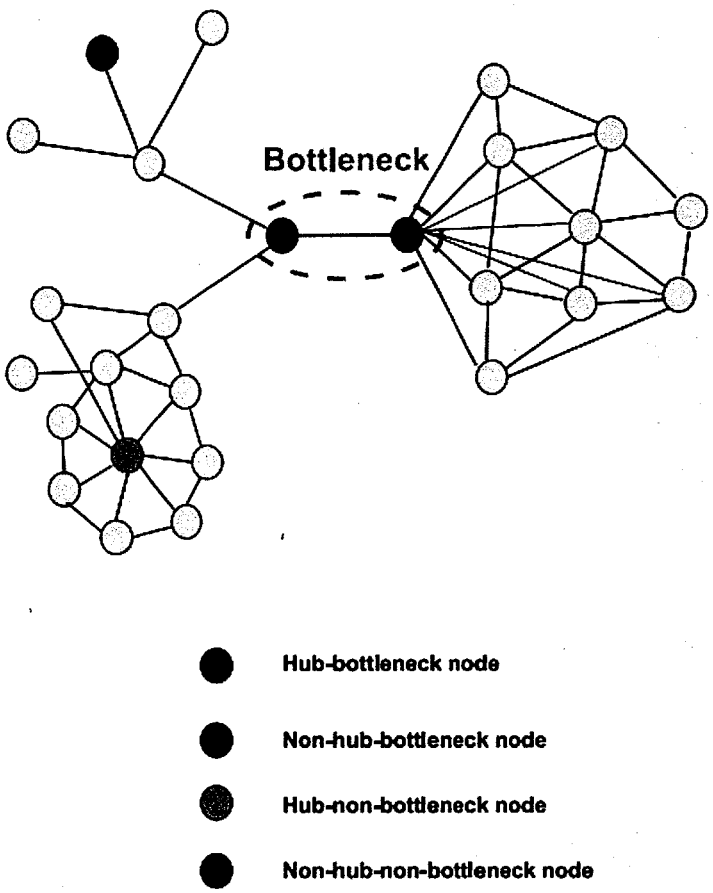


Figure 1.3 Schéma représentant quatre catégories de noeuds dans un réseau.

Les quatre points définis selon le degré de connections et de centralité intermédiaire sont représentés par différentes couleurs. Les protéines « bottleneck » sont encerclées en pointillé. Image tirée de (Yu *et al.*, 2007).

1.2.2 Propriétés des réseaux d'interactions génétiques

En premier lieu, la topologie du réseau d'IG chez la levure, grâce aux nombreux efforts concertés de criblage des IG à grande échelle, est semblable à celui des IPP et des autres réseaux biologiques et présente une structure invariante d'échelle avec des gènes hautement connectés jouant le rôle de Hubs (Costanzo *et al.*, 2010; Tong *et al.*, 2004). De ce fait, le degré de connectivité des Hubs d'interactions génétiques corrèle avec le degré d'interactions dans le réseau d'IPP (Costanzo *et al.*, 2010). Les Hubs d'IG ont aussi une propension à être des gènes essentiels et à être multi-fonctionnels (associés à une multitude de fonctions biologiques), tout comme ceux de l'interactome protéine-protéine (Costanzo *et al.*, 2010).

À plus petite échelle, le réseau d'interactions génétiques de la levure, basé sur des interactions négatives (létales, voir section 1.1), est modulaire et les interactions sont concentrées entre des gènes qui partagent des fonctions biologiques semblables (Costanzo *et al.*, 2010). Cette organisation modulaire des IG, qui implique fréquemment des modules denses d'interactions composés en majorité d'interactions négatives, a abouti à la notion de monochromaticité des IG (Segre *et al.*, 2005) (Connections en rouge; Figure 1.4). La monochromaticité des IG est observée lorsqu'un module dense d'interactions est composé en majorité, soit d'IG positives ou d'IG négatives (Segre *et al.*, 2005). Chez la levure, cette particularité du réseau d'IG a aussi été observée avec un nombre élevé d'interactions positives à l'intérieur de certains modules fonctionnels (Connections en vert; Figure 1.4) (Fiedler *et al.*, 2009; Schuldiner *et al.*, 2005).

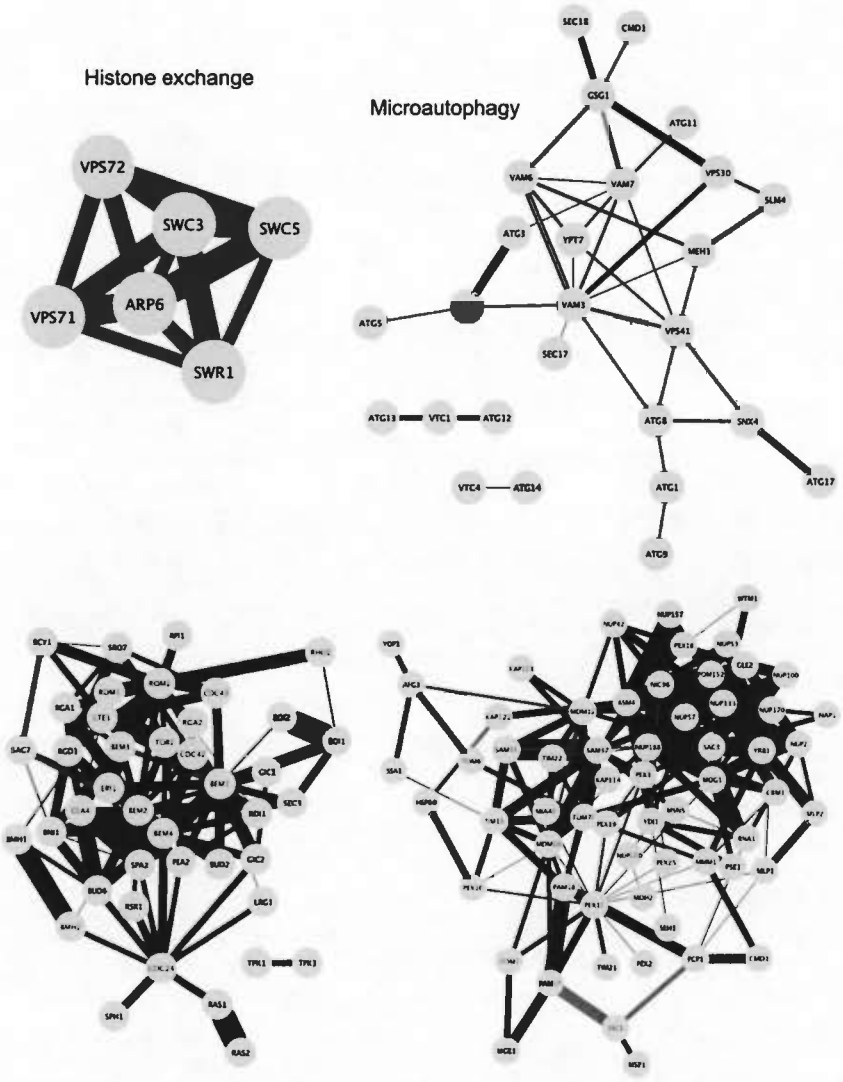


Figure 1.4 Exemples de modules fonctionnels monochromatiques pour les interactions génétiques chez la levure *Saccharomyces cerevisiae*.

Chaque noeud représente un gène lié par une ou des interactions génétiques de type positives (vert) ou négatives (rouge). Image tirée de (Michaut *et al.*, 2011).

En combinant les IG et les interactions protéine-protéine de la levure, Kelley et Ideker (2005) ont précisé la notion de monochromaticité selon deux modèles décrivant la mécanistique des interactions génétiques. Le premier modèle d'IG, nommé « between-

pathways » (*i.e.* IG concentrées entre les voies), décrit plusieurs interactions génétiques connectant les membres de deux voies de signalisation ou de métabolisme différentes (Figure 1.5a) (Kelley et Ideker, 2005). Ce modèle est prépondérant pour les IG négatives de la levure, et donc monochromatique, avec une majorité d'interactions négatives ayant été observées entre les membres de deux voies distinctes dans de nombreuses études qui ont appliqué ce modèle dans le réseau d'IG de la levure (Figure 1.5b) (Bandyopadhyay *et al.*, 2008; Bellay *et al.*, 2011; Kelley et Ideker, 2005; Szappanos *et al.*, 2011; Ulitsky et Shamir, 2007). Le deuxième modèle, nommé « within-pathways » (*i.e.* à l'intérieur des voies), décrit un groupement d'IPP ou un complexe protéique qui est également densément connecté par des interactions génétiques (Figure 1.5a) (Kelley et Ideker, 2005). Ce modèle s'appliquerait aux interactions positives retrouvées entre des gènes non-essentiels puisque leurs mutations individuelles ne sont pas létales et peuvent, au contraire, mener à une atténuation du phénotype observé (*e.g.* le complexe protéique DEFGH; Figure 1.5b). Par contre, cette hypothèse s'est avérée majoritairement fausse pour l'ensemble des interactions positives chez la levure, et c'est plutôt leur prévalence à connecter des modules fonctionnels distants ou indépendants qui a été observé (*e.g.* IG positives entre le complexe DEFGH et IJKL; Figure 1.5b) (Bellay *et al.*, 2011; Kelley et Ideker, 2005; Ulitsky et Shamir, 2007).

L'une des raisons qui pourrait expliquer la supériorité du modèle « between-pathway » sur celui du « within-pathway » réside dans le fait que la plupart de ces études ont considéré les voies de signalisation/métabolisme (ou modules) comme des motifs denses d'interactions protéine-protéine ou protéine-ADN (Bandyopadhyay *et al.*, 2008; Bellay *et al.*, 2011; Kelley et Ideker, 2005; Szappanos *et al.*, 2011; Ulitsky et Shamir, 2007). En réalité, les voies sont définies comme un groupe de molécules fonctionnant ensemble, avec ou sans interactions, pour contrôler un processus biologique donné (Kanehisa *et al.*, 2004; Vastrik *et al.*, 2007).

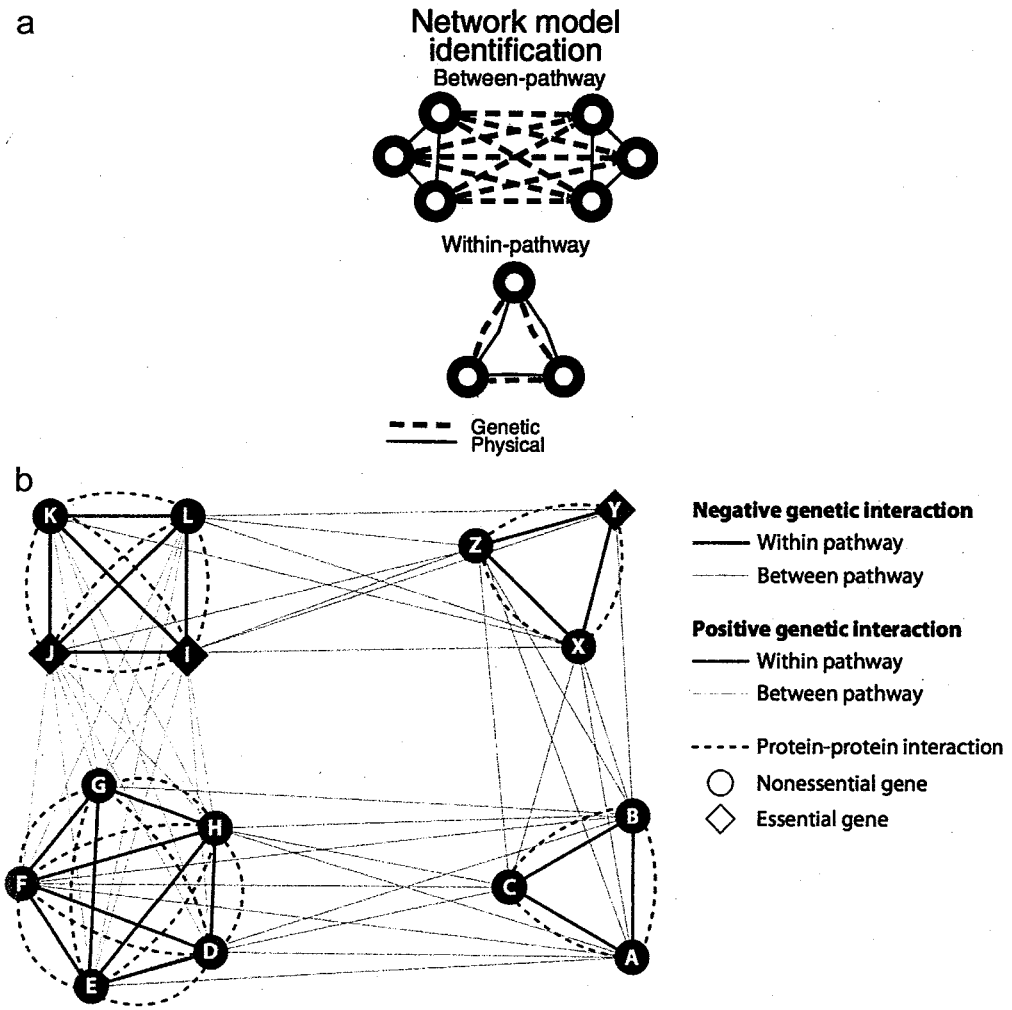


Figure 1.5 Modèles pour illustrer à la fois la modularité et la monochromaticité des interactions génétiques chez la levure.

(a) Les modèles de base proposent de répartir les interactions génétiques (lignes rouge pointillées) selon qu'elles connectent deux modules denses d'interactions protéine-protéine (lignes bleu, « Between-pathway ») ou qu'elles apparaissent à l'intérieur d'un module (« Within-pathway »). **(b)** La monochromaticité des modèles est expliquée selon une prépondérance d'interactions négatives (liens en rouge) entre les complexes d'interactions protéine-protéine (lignes noir pointillées) ou d'interactions positives (liens en vert) apparaissant à l'intérieur d'un complexe protéique formé par des gènes non-essentiels. Image tirée de (Kelley et Ideker, 2005) et (Baryshnikova *et al.*, 2013).

Du côté des organismes multi-cellulaires, il n'y pas d'interactome génétique à grande échelle semblable à la levure dû aux difficultés de criblage expérimental et à la taille du génome de la plupart de ces espèces. Toutefois, des études à petite échelle chez le nématode *C. elegans* montrent un réseau dont la topologie est similaire à celle observée chez la levure avec un réseau invariant d'échelle et la présence de Hubs (Byrne *et al.*, 2007; Lehner *et al.*, 2006). De plus, les Hubs auraient aussi cette tendance à être des gènes essentiels, mais cette observation a été faite dans un réseau d'IG de type négatif uniquement (Byrne *et al.*, 2007). Aussi, au niveau de la modularité du réseau d'IG de *C. elegans*, le modèle « within-pathway » semble mieux s'appliquer à une majorité d'interactions génétiques chez le nématode, à l'inverse de la levure pour qui le modèle « between-pathway » prédomine (Lehner, 2007). Une des raisons pour expliquer cette différence résiderait dans la nature des méthodes utilisées pour cribler les IG chez chacun de ces organismes, mais on ne peut pas exclure la possibilité que cette différence soit due à la nature même de ces espèces, *i.e.* uni- versus multi-cellulaires.

1.3 Le criblage *in silico* des interactions génétiques

Malgré les avantages du criblage à haut-débit des IG en laboratoire, la quantification nécessaire pour évaluer l'effet de deux mutations sur un phénotype en particulier requiert une quantité astronomique de mesures qui augmentent selon la taille du génome de l'espèce (*e.g.* la bactérie *Escherichia coli* ~10 millions de paires de gènes X 10 mesures = 100 millions de mesures, chez l'humain ~200 millions de paires de gènes X 10 mesures = 2 milliards de mesures) (Onami et Kitano, 2006). De plus, le nombre total d'IG identifié par des études à grande échelle représente moins de 1% de toutes les paires de gènes testées chez la levure et le nématode (Costanzo *et al.*, 2010; Lehner *et al.*, 2006). Le rendement en interactions identifiées et la couverture en

nombre de gènes sont donc généralement faibles en laboratoire (« systematic screens », en bleu, Figure 1.6). Une façon de surmonter cette difficulté est d'utiliser une approche *in silico* dans laquelle une paire de gènes en particulier est ciblée (Onami et Kitano, 2006).

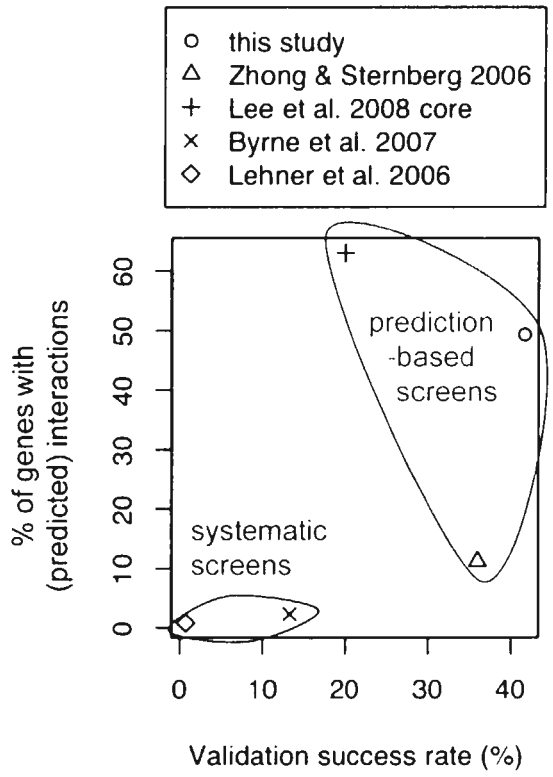


Figure 1.6 Comparaison entre les approches expérimentales de criblages et les méthodes prédictives d'interactions génétiques chez *C. elegans*.

Les comparaisons sont faites selon le pourcentage de gènes ayant au moins une interaction (« % of genes with interactions ») et la fraction de paires de gènes testées montrant une interaction (rendement ou « Validation success rate »). Le criblage systématique en laboratoire (« systematic screens », (Lehner *et al.*, 2006) et (Byrne *et al.*, 2007)) teste seulement quelques paires de gènes dû aux nombreux facteurs limitants des manipulations expérimentales. Le rendement est aussi généralement faible dû à la rareté des interactions génétiques par rapport au nombre de paires testées. À l'inverse, la prédiction *in silico* d'interactions génétiques (« prediction-based screens ») peut se faire pour l'ensemble des paires de gènes selon la couverture des données utilisées ((Lee, I. *et al.*, 2010), (Lee, A. Y. *et al.*, 2010) et (Zhong et Sternberg, 2006)). De plus, étant donné la hiérarchisation des paires de gènes, seules celles présentant une très forte probabilité d'interactions sont testées en laboratoire, augmentant ainsi le rendement. Image tirée de (Lee, A. Y. *et al.*, 2010).

Cette méthode consiste à d'abord prédire des IG pour l'ensemble des paires de gènes et de valider en laboratoire par la suite celles qui présentent des grandes chances d'être de vraies interactions génétiques. Les rendements et couvertures en nombre de gènes obtenus sont donc supérieurs grâce à cette approche (« prediction-based screens » en rouge, Figure 1.6) (Lee, I. *et al.*, 2010).

Pour prédire des IG chez un organisme, il faut d'abord disposer de données, idéalement à grande échelle, provenant de sources multiples (*e.g.* expression des gènes, interactions protéine-protéine, annotations fonctionnelles) (« Data Sources », Figure 1.7) (Madhukar *et al.*, 2015). Ensuite, les données peuvent être transformées (*e.g.* Co-expression des gènes, enrichissement en annotations communes pour deux gènes) ou utilisées directement (*e.g.* IPP, pourcentage d'homologie entre deux gènes) pour calculer des scores entre les différentes paires de gènes qui sont communément appelés attributs (« Feature extraction », Figure 1.7) (Madhukar *et al.*, 2015). Comme chacune des paires de gènes se retrouve avec un score pour chacun des attributs, un réseau peut être construit en appliquant un seuil minimum sur les scores (Lee, A. Y. *et al.*, 2010). Ce réseau peut être utilisé par la suite pour construire d'autres attributs (Lee, A. Y. *et al.*, 2010; Madhukar *et al.*, 2015). Les attributs sont ensuite combinés dans une matrice pour toutes les paires de gènes (Fig.1.7). Bien que facultatif, ces attributs peuvent aussi être évalués individuellement pour leur potentiel prédictif envers les IG afin d'utiliser seulement ceux présentant des liens forts avec les IG (Lee, A. Y. *et al.*, 2010). Bien que cela permet de construire un modèle plus spécifique au contexte des IG, différents algorithmes de sélection automatique existent pour utiliser les attributs les plus performants avant l'étape de l'apprentissage machine (Madhukar *et al.*, 2015). Viens ensuite le choix d'un ensemble d'entraînement qui, après intégration, servira de modèle de base pour la prédiction d'IG (Fig.1.7) (Madhukar *et al.*, 2015; Onami et Kitano, 2006). Cet ensemble de référence doit être représentatif et contient idéalement des IG en nombre suffisant pour être séparé en deux afin de conserver un certain nombre d'IG pour les phases de tests (« Training » et « Test », Fig.1.7) (Madhukar *et al.*, 2015).

L'ensemble de tests, constituer aléatoirement avec des IG, servira à évaluer la performance et la pertinence du modèle prédictif construit avec le reste des IG Madhukar *et al.*, 2015).

Pour le moment, seule la levure dispose d'un nombre suffisant d'IG pour appliquer cette approche. Autrement, lors des phases de tests, le modèle est évalué pour sa pertinence envers d'autres IG par différentes méthodes d'évaluation en considérant uniquement l'ensemble d'entrainement non fragmenté (Madhukar *et al.*, 2015). La méthode la plus utilisée, appelée validation croisée, consiste à séparer l'ensemble d'entrainement au hasard en k sous-ensembles équivalents et de générer les prédictions pour le $k^{ième}$ sous-ensemble en utilisant les k-1 sous-ensembles comme données d'entraînements (Schrynemackers *et al.*, 2013). On applique généralement, de façon aléatoire, plus de 5 séparations pour s'assurer d'un échantillonnage approprié. À l'extrême, toutes les séparations sont effectuées avec $k = n$ sous-ensembles dans une technique appelée « leave-one-out cross-validation » (Madhukar *et al.*, 2015). À la fin, les prédictions sont évaluées sur un ensemble d'IG réel (positif) et un ensemble négatif (*e.g.* paires de gènes pris au hasard), les performances de classification pour chaque modèle sont calculées selon les ratios d'IG et de négatifs correctement prédits par rapport à la taille de l'ensemble de départ (Madhukar *et al.*, 2015). Dans l'ensemble, un modèle qui offre de bonnes performances, basées sur un ensemble d'entrainement suffisamment large, peut être utilisé ensuite sur l'ensemble des paires de gènes d'un organisme pour prédire des IG et cibler ou caractériser celles qui portent un intérêt particulier.

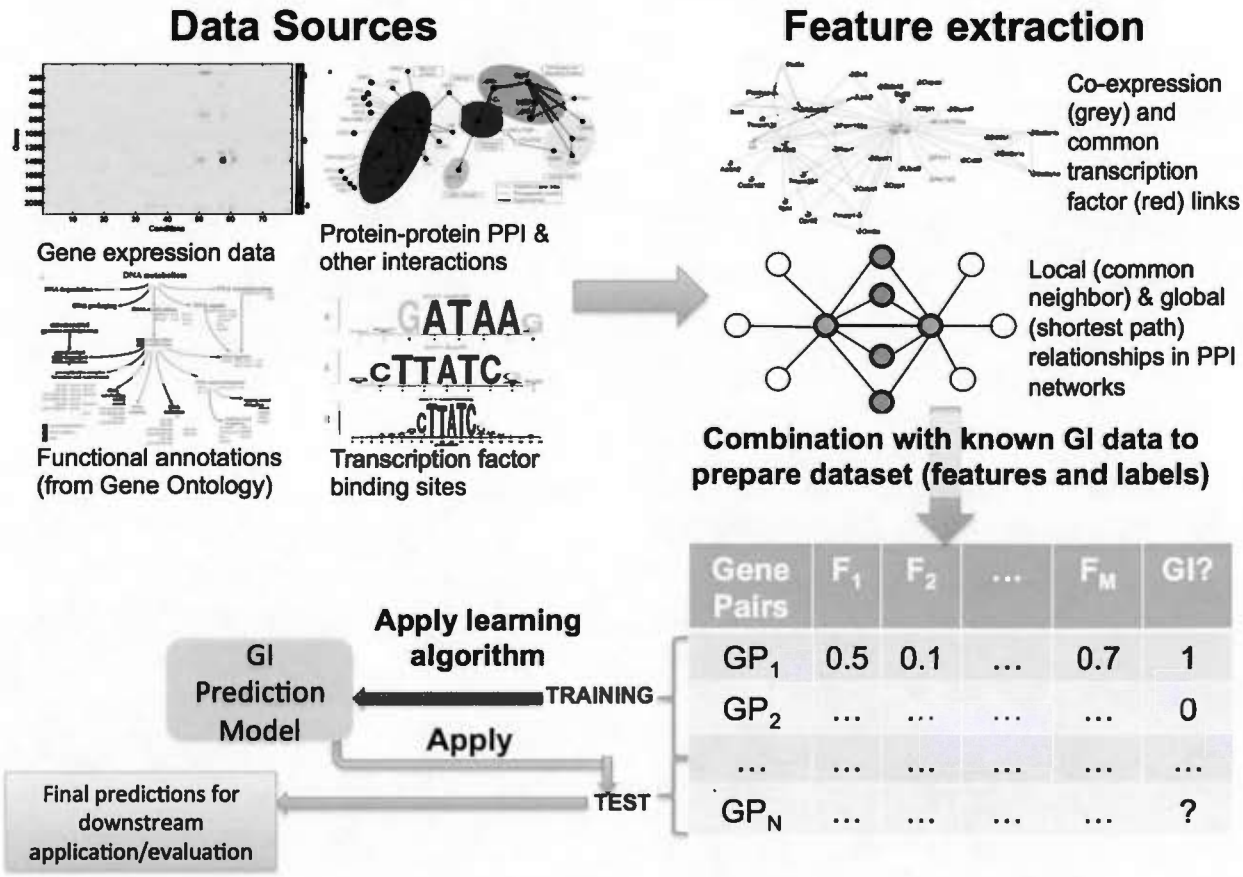


Figure 1.7 Survol de l'approche la plus utilisée pour la prédiction d'interactions génétiques.

Divers types de données à grande échelle sont extraits des bases de données pour être utilisés directement ou encore, être transformés pour construire les attributs (ou « feature »). Ces attributs sont ensuites combinés dans une matrice pour l'ensemble des paires de gènes. Finalement, un modèle prédictif est construit grâce à un ensemble d'entraînement ayant passé les phases de tests pour servir à la prédiction d'interactions génétiques entre des paires de gènes d'intérêts. Image tirée de (Madhukar *et al.*, 2015).

1.3.1 Chez la levure *S. cerevisiae*

Chez les eucaryotes, la levure *S. cerevisiae* est l'espèce qui a bénéficié le plus d'efforts concertés afin de tester plus de cinq millions de paires de gènes expérimentalement (Baryshnikova *et al.*, 2010; Costanzo *et al.*, 2010). Mais comme ces progrès sont relativement récents, la levure a eu son lot d'études *in silico* auparavant, en particulier avant 2010, afin d'augmenter la couverture des gènes avec des IG grâce aux outils prédictifs. Ces études ont aussi mis de l'avant plusieurs méthodes qui sont encore couramment utilisées dans la prédiction des IG (Madhukar *et al.*, 2015).

Brièvement, Tong *et al.* (2004) ont montré qu'environ 20% des voisins d'un gène dans l'interactome génétique partagent aussi des interactions génétiques entre eux. Cette propriété peut donc servir à prédire des IG pour des gènes qui sont peu caractérisés (Tong *et al.*, 2004). Néanmoins, cette approche, généralement appelée règle d'association, est couramment employée comme technique d'inférence mais nécessite une quantité adéquate d'information (des IG en l'occurrence) et l'information doit idéalement être dépourvue de biais expérimentaux (*e.g.* un seul gène est testé systématiquement contre 1000 autres plutôt que toutes les combinaisons possibles de paires de gènes) (Gillis et Pavlidis, 2012). Pour pallier à ce problème (récurrent chez les organismes autre que la levure), Pandey *et al.* (2010) ont construit 152 attributs, dérivés de types de données allant de l'expression des gènes aux annotations fonctionnelles, pour prédire des IG sans avoir recours directement au réseau d'IG pour la construction des attributs. Pour ce faire, ils ont eu recours à une approche d'arbres de décisions et à cinq autres types de prédicteurs différents (*e.g.* analyse discriminante linéaire, réseaux de neurones) qu'ils ont combiné ensemble pour maximiser le potentiel prédictif de chaque type de données (Pandey *et al.*, 2010). La combinaison de plusieurs prédicteurs permet aussi, au final, d'obtenir un consensus au niveau des prédictions. Ils ont appliqué cette approche pour prédire des interactions génétiques entre les facteurs de transcription connus chez *S. cerevisiae* et découvert un grand nombre de nouvelles

interactions entre ces facteurs. Leur effort a permis de générer des IG prédites pour environ 7,5 millions de paires de gènes non-essentiels de la levure (Pandey *et al.*, 2010). Ces prédictions n'ont toutefois pas été validées expérimentalement, probablement en raison des difficultés que cela comporte en termes de temps et d'argent.

Une autre manière de prédire des IG consiste à exploiter la relation entre les réseaux d'interactions protéine-protéine et génétiques. Comme la topologie et la connectivité du réseau des interactions protéine-protéine (IPP) est semblable à celui des IG (*e.g.* réseau invariant d'échelle, présence de hubs), cette information pourrait être utilisée pour prédire des IG entre deux gènes qui codent pour des protéines qui interagissent ensemble physiquement (Chipman et Singh, 2009; Kelley et Ideker, 2005; Paladugu *et al.*, 2008).

Dans l'ensemble, la prédiction d'IG chez *S. cerevisiae* a permis d'instaurer les bases de la prédiction d'IG, en particulier chez les organismes supérieurs, mais n'est évidemment plus nécessaire chez la levure dû aux criblages à grande échelle qui ont permis de tester près de la moitié de l'ensemble des paires de gènes de cet organisme. De plus, la piètre conservation des IG entre les espèces ne permet pas d'utiliser directement les IG chez la levure pour les transposer à d'autres espèces par orthologie de gènes (*i.e.* homologie qui tient compte de la divergence entre les espèces), il est donc nécessaire de développer des outils prédictifs basés sur des données rattachées à l'espèce, et ce, pour chacune des espèces étudiées.

1.3.2 Chez *C. elegans*

Contrairement à la levure, le nématode *Caenorhabditis elegans* n'a que quelques milliers d'IG de disponibles pour les études à moyen ou large spectre. La prédiction

d'IG est donc nécessaire pour cette espèce. *C. elegans* possède tout de même une multitude de phénotypes bien documentés dans les bases de données, de même que des données d'expression et d'IPP (Howe *et al.*, 2016).

La méthodologie générale pour prédire des IG chez *C. elegans* emploie habituellement seulement des données expérimentales, mais peut comprendre aussi des données non-expérimentales comme la co-citation de gènes dans la littérature (Lee, I. *et al.*, 2010). Bon nombre d'études de prédictions d'IG à grande échelle chez *C. elegans* utilise quatre principaux types de données, l'expression des gènes, les annotations fonctionnelles et/ou phénotypes, les IPP et les IG (Chipman et Singh, 2009; Hoehndorf *et al.*, 2013; Lee, A. Y. *et al.*, 2010; Lee, I. *et al.*, 2010; Zhong et Sternberg, 2006). Mais à l'exception de l'expression des gènes (facilitée par les micropuces à ADN ou « gene microarrays »), bien peu de types de données couvrent les quelques 18 000 gènes de *C. elegans*. Aussi, il est important de se rappeler que ce ver n'a que quelques milliers d'IG validées expérimentalement, empêchant par le fait même son utilisation comme réseau dans des analyses prédictives d'IG semblables à celles chez la levure. La plupart des types de données ne couvrent donc qu'une petite partie du génome du nématode, laissant beaucoup de données manquantes dans les matrices de paires de gènes utilisées pour l'intégration de données (Zhong, W. et Sternberg, 2007).

Pour surmonter cet obstacle, Zhong et Sternberg (2006) ont utilisé deux méthodes, la première consiste à combiner des données d'un même type, provenant de différentes sources couvrant chacune une partie de l'ensemble des gènes, afin d'optimiser la couverture du génome. La deuxième se base sur le principe que si deux gènes interagissent ensemble (pour une espèce), leurs orthologues ont aussi plus de chance d'interagir génétiquement que des paires de gènes prises au hasard (Zhong et Sternberg, 2006). Les deux auteurs ont donc intégré cinq types de données provenant de trois espèces différentes (*C.elegans*, *S. cerevisiae*, *D. melanogaster*) pour obtenir des probabilités d'IG pour environ 10% des paires d'orthologues chez le nématode (Zhong et Sternberg, 2006). La faible couverture en nombre de gènes avec au moins une IG

prédite peut s'expliquer en partie par le fait qu'un nombre restreint de gènes chez *C. elegans* ont des orthologues chez une des deux autres espèces (Sonnhammer et Östlund, 2015). Un autre problème, cette fois au niveau de la spécificité de leur modèle, vient du fait que les auteurs ont utilisé les IG de la levure et de la drosophile comme attributs prédictifs pour l'ensemble des paires de gènes de *C. elegans* (Zhong et Sternberg, 2006). Il a été montré que sorti d'un contexte en particulier (*e.g.* IG entre gènes essentiels ou IG à l'intérieur d'une voie de métabolisme), les IG sont peu ou pas conservées entre *C. elegans* et *S. cerevisiae* et ont aussi peu de chance d'être conservées entre *C. elegans* et *D. melanogaster* (Tischler *et al.*, 2008). L'utilisation des IG multi-espèces comme attribut prédictif, en comparaison avec des IG spécifiques à l'espèce, est d'ailleurs l'une des raisons de la faible performance de ce prédicteur face à d'autres (Chipman et Singh, 2009). Par contre, cela n'a pas empêché les auteurs d'obtenir un bon rendement en validations expérimentales tel que démontré par un haut ratio de paires de gènes validées par criblage et d'interactions prédites par leur modèle pour ces mêmes paires de gènes (Zhong et Sternberg, 2006).

Une autre problématique rencontrée lors de la prédiction d'IG chez le nématode concerne l'ensemble d'entrainement utilisé dans le modèle prédictif. Comme mentionné précédemment, le nématode dispose que de quelques milliers d'IG et l'utilisation d'un ensemble d'entrainement jugée trop petit peut mener au phénomène de surapprentissage (« overfitting ») où le modèle prédictif n'a plus la flexibilité requise pour prédire avec succès des nouvelles instances (Okser *et al.*, 2013). La règle est donc de privilégier, si possible, une grande quantité d'instances validées (exemples positifs) pour maximiser les chances d'avoir un modèle prédictif couvrant la majorité des cas non-étudiés (Okser *et al.*, 2013). De plus, les instances positives doivent, dans l'idéal, couvrir un maximum d'exemples d'IG, selon leurs profils de valeurs des attributs, afin d'obtenir un ensemble d'entraînement équilibré (Okser *et al.*, 2013). Pour surmonter ces obstacles, Zong et Sternberg (2006) ont utilisé, en plus des quelques 1 800 IG, les 2 878 interactions protéine-protéine (IPP) de *C. elegans* validées par double hybride

(« yeast two-hybrid ») en se basant sur le fait que deux gènes qui interagissent ensemble ont plus de chance de coder pour des protéines qui interagissent physiquement que deux gènes pris au hasard (Tong *et al.*, 2004). Cette propension des paires de gènes qui interagissent génétiquement, à interagir physiquement aussi, n'est toutefois pas généralisée car dans l'ensemble, moins de 10% des IG partagent aussi des IPP (Tong *et al.*, 2004). L'utilisation d'IPP en tant qu'interactions génétique a donc pour effet d'introduire potentiellement un nombre non négligeable de faux positifs dans l'ensemble d'entraînement.

Une manière différente de contourner le manque d'IG chez *C. elegans* pour la construction de l'ensemble d'entrainement est d'utiliser les annotations fonctionnelles (*e.g.* termes « gene ontology », voies de signalisation) pour former des paires de gènes qui partagent une ou plusieurs annotations en commun (Lee, I. *et al.*, 2010). Cette approche permet d'avoir un ensemble d'entrainement comme modèle qui couvre la majorité des paires de gènes du nématode. Par exemple, en utilisant les annotations « Gene Ontology » (GO), Lee *et al.* (2010a) ont construit un ensemble d'entrainement composé de plus de 600 000 paires de gènes pour générer un modèle capable d'identifier des paires de gènes impliquées dans les mêmes processus biologiques (co-fonctionalité). Ils ont ensuite découvert que ce modèle permettait aussi d'identifier des IG (Lee, I. *et al.*, 2010). Comme l'objectif principal était de prédire des liens fonctionnels entre les gènes (ou impliqués dans les même processus), les rendements en validation expérimentale pour les IG ont été inférieurs à ceux obtenus pour les approches utilisant les IG comme ensemble d'entrainement (Kelley et Ideker, 2005; Lee, A. Y. *et al.*, 2010; Lee, I. *et al.*, 2010; Wong *et al.*, 2004; Zhong, W. et Sternberg, 2007). Ce qui est mis en évidence ici témoigne d'un concept important en apprentissage machine qui survient lorsque des ensembles d'entrainements très volumineux sont utilisés. En effet, plus on ajoute d'instances dans l'ensemble d'entrainement, plus on augmente le risque d'introduire des instances qui s'éloignent de la vérité en raison de plusieurs facteurs qui s'ajoutent (*e.g.* expérimentateurs, méthodes, données

manquantes, bruit de fond) (Raykar *et al.*, 2009). Dans l'idéal, un ensemble d'entrainement bien balancé et sélectionné avec soin permet de générer des prédictions avec beaucoup plus de précisions (Batista et Meira, 2004; Cano *et al.*, 2007). En utilisant seulement les IG de *C. elegans* comme ensemble d'entrainement, nous avons obtenu les meilleurs rendements en validations expérimentales à ce jour. Le succès de l'approche adoptée par notre laboratoire ne repose pas uniquement sur le choix d'un ensemble d'entrainement composé d'IG validées expérimentalement, mais aussi sur l'utilisation d'attributs capable d'exploiter à la fois les réseaux d'IPP/co-expression des gènes et les associations gènes-phénotypes (Lee, A. Y. *et al.*, 2010).

Plus spécifiquement, pour construire ces attributs, des données d'expression des gènes et des IPP sont colligées pour construire les réseaux de co-expressions et d'IPP (Figure 1.9A). Pour le réseau d'IPP, le degré d'enrichissement en partenaires communs de chaque protéine dans le réseau est calculé pour construire un attribut prédictif d'IG (Figure 1.9B). En effet, il a été montré que deux gènes enrichis en partenaires communs dans le réseau d'IPP ont plus de chance d'interagir génétiquement que les autres (Lee, A. Y. *et al.*, 2010). Ensuite, comme les phénotypes ne couvrent qu'une partie de l'ensemble des gènes, cette couverture pourrait être étendue en explorant le voisinage (les partenaires directs) de chaque gène dans le réseau combiné de co-expressions et d'IPP qui lui, couvre plus de 80% des gènes. On applique ensuite le même principe d'enrichissement dans le voisinage, mais cette fois-ci, on calcule l'enrichissement d'un phénotype dans le voisinage (nombre de gènes associés au phénotype) de deux gènes (Figure 1.9C). Finalement, deux gènes enrichis dans leurs voisinages pour un phénotype donné ont plus de chance d'interagir génétiquement que les autres, d'autant plus si les deux gènes sont aussi associés au phénotype en question (Figure 1.9D) (Lee, A. Y. *et al.*, 2010).

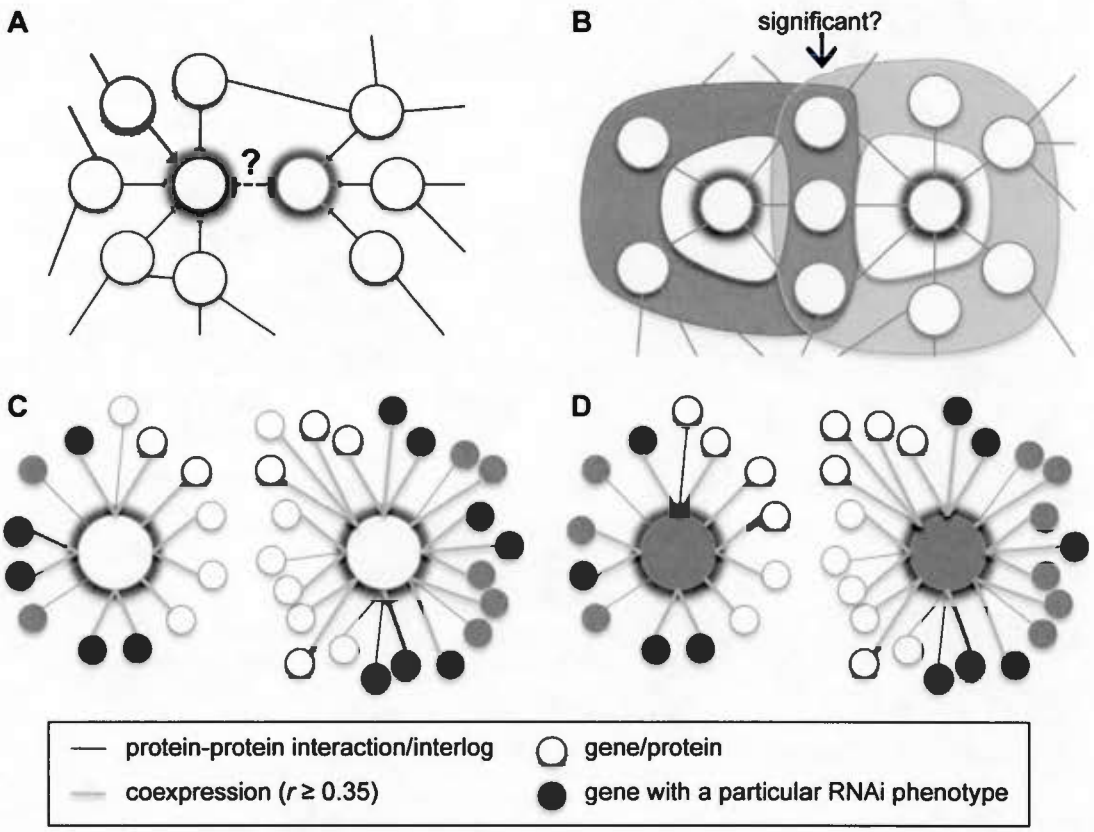


Figure 1.8 Attributs utilisés pour la prédiction d'interactions génétiques.

(A) Les réseaux d'interactions protéine-protéine (IPP) et de co-expressions (non-montrés) peuvent être combinés pour étendre la couverture génomique. **(B)** On peut mesurer le degré d'enrichissement en partenaires communs entre deux gènes dans le réseau d'IPP. **(C)** Le réseau combiné de co-expressions et d'IPP est utilisé pour définir le voisinage de deux gènes afin de calculer l'enrichissement en nombre de voisins associés à un phénotype donné. **(D)** Pareillement à (C), mais avec l'ajout que les deux gènes d'intérêt soient aussi associés au phénotype. Image tirée de (Lee, A. Y. *et al.*, 2010).

En résumé, chez les organismes multi-cellulaires, *C. elegans* est l'espèce qui a le plus bénéficié des études de prédiction d'IG grâce, entre autres, à la quantité d'information disponible dans les bases de données. Mais aussi grâce aux nombreuses études sur la génétique du nématode qui ont permis de rassembler une quantité appréciable d'IG validées expérimentalement pouvant servir à l'entrainement et/ou à la validation de

modèles prédictifs pour les IG. Au final, les différentes approches employées pour la prédiction d'IG chez *C. elegans* ont le potentiel d'être appliquées chez d'autres organismes multi-cellulaires disposant d'un nombre restreint d'IG.

1.3.3 Chez l'humain

Malgré la quantité impressionnante de données à grande échelle et la disponibilité d'IG découvertes dans des études *in vitro* chez l'humain, peu d'études *in silico* ont vu le jour pour tenter d'explorer le sujet. Parmi les causes possibles, citons la découverte des IG dites « contexte-spécifiques » (*i.e.* qui sont détectables dans un seul contexte donné) (voir section 1.2.2) (Chari et Dworkin, 2013). En effet, pour prédire des IG dans le contexte d'une maladie, par exemple, il faut une compréhension approfondie des phénotypes, des voies de signalisation et des mécanismes d'altération propres à cette maladie, et les données rattachées sont parfois non disponibles (*e.g.* maladies rares). Néanmoins, quelques études ont tenté de prédire des IG chez l'humain, notamment dans le cancer (Jerby-Arnon *et al.*, 2014; Lu *et al.*, 2015).

Dans la première, les auteurs ont intégré des données d'expression de gènes et de criblage d'IG provenant de plus d'une centaine de lignées cellulaires de différents types de cancer (Fig. 1.9) (Jerby-Arnon *et al.*, 2014). Ils ont ensuite combiné ces résultats avec ceux provenant du séquençage d'ADN (*e.g.* nombre de copie d'un gène, mutations) pour construire des réseaux d'interactions prédites spécifiques à chaque type de cancer, et pouvant être utilisés dans le pronostic de la maladie ou dans la réponse à un médicament (Fig. 1.9) (Jerby-Arnon *et al.*, 2014). Dans l'élaboration de leur modèle prédictif, les auteurs se sont basés sur le principe de coupable par association pour construire les réseaux d'IG prédites. Plus précisément, en analysant le réseau d'interaction connu, ils ont émis trois remarques:

i. Les paires de gènes qui interagissent négativement ont un taux de co-mutations (ou co-délétions) plus faible que les paires choisies au hasard (puisque le contraire serait défavorable, voir létal, pour la cellule cancéreuse).

ii. Les interactants négatifs d'un gène X peuvent être détectés si la sous-expression ou la perte de ces gènes induit l'essentialité du gène X dans les criblages shRNA.

iii. Les gènes qui interagissent négativement œuvrent dans la même voie de signalisation/métabolique et ont donc tendance à être co-exprimés.

Ils ont par la suite construit trois modèles statistiques basés sur chaque observation, les IG négatives identifiées expérimentalement formant les ensembles de référence (« gold standard » en anglais), et ont identifié les paires de gènes pour qui les trois tests statistiques sont significatifs, comme interagissant génétiquement (Jerby-Arnon *et al.*, 2014). Au final, ils ont obtenu un réseau combiné de 2816 interactions négatives prédites couvrant 2077 gènes. Leur approche a aussi été utilisée avec succès pour découvrir 44 interactants synthétiques létals du suppresseur de tumeur VHL grâce à des validations par ARN interférant (Jerby-Arnon *et al.*, 2014).

Pour étendre cette couverture à plus de 15 000 gènes, Lu *et al.* (2015) ont utilisé des données d'expression de gènes pour mesurer leurs fluctuations (expressions différentielles) dans un réseau d'IG négatives identifiées expérimentalement provenant de plusieurs types de cancer. Ils ont remarqué que lorsqu'un gène interagit négativement avec un autre, l'un d'eux est sur-exprimé pendant que l'autre est, au contraire, sous-exprimé (Lu *et al.*, 2015). Cette constatation leur a permis de construire un modèle prédictif pour les paires de gènes qui partagent un motif d'expression similaire afin de prédire des IG négatives (Lu *et al.*, 2015). Ces prédictions n'ont toutefois pas été validées expérimentalement.

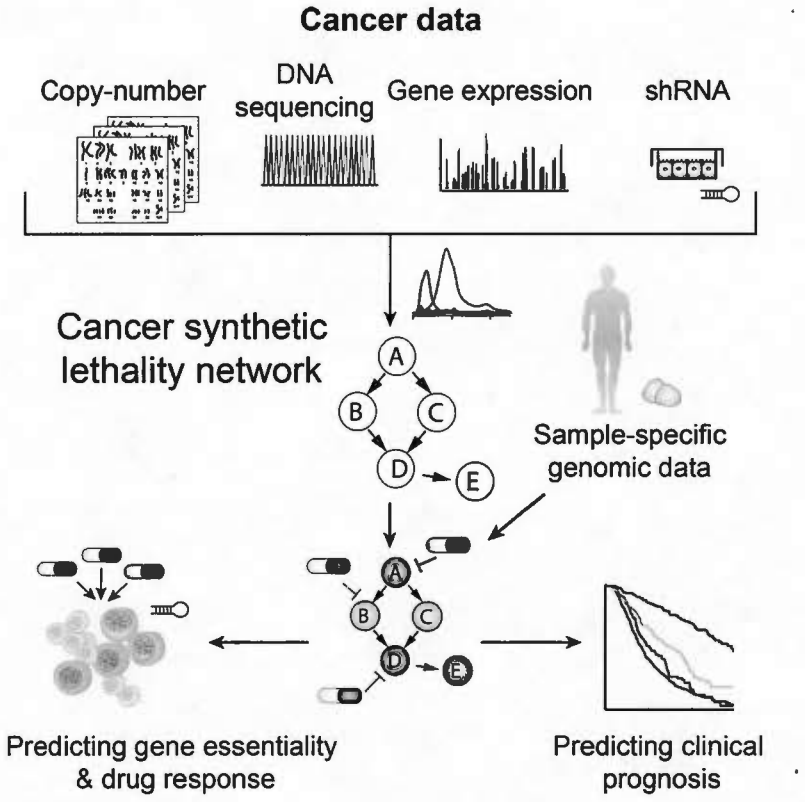


Figure 1.9 Schéma de la prédiction d'interactions génétiques négatives chez l'humain.

Différents types de données sont intégrés avec les interactions génétiques dites « synthétiques létales » (SL). L'information résultante peut être utilisée ensuite, par le principe de coupable par association, pour construire un réseau d'interactions SL prédites. Ces interactions prédites peuvent éventuellement conduire à un meilleur pronostic ou à une meilleure réponse à un médicament en ciblant un interactant SL en particulier avec une autre molécule thérapeutique. Image tirée de (Jerby-Arnon *et al.*, 2014).

1.4 La génomique intégrative dans le cancer du sein

Dans le cadre de cette thèse, j'ai travaillé sur les réseaux d'interactions fonctionnelles du nématode *C. elegans*. Les principales conclusions de nos études nous ont mené à

s'interroger sur la génétique du cancer, plus précisément au niveau du cancer du sein. Les prochaines sections introduisent les notions de bases de cette maladie et exposent des résultats importants provenant de différentes études intégratives.

1.4.1 Caractéristiques des tumeurs mammaires

Le cancer du sein tient son origine des glandes mammaires, lesquelles sont formées par une couche extérieure de cellules myoépithéliales qui entourent les cellules luminales (Figure 1.10). À l'intérieur, se trouve des cellules souches en position basale qui donneraient lieu, après différentiation, aux cellules progénitrices (Figure 1.10) (Tiede et Kang, 2011).

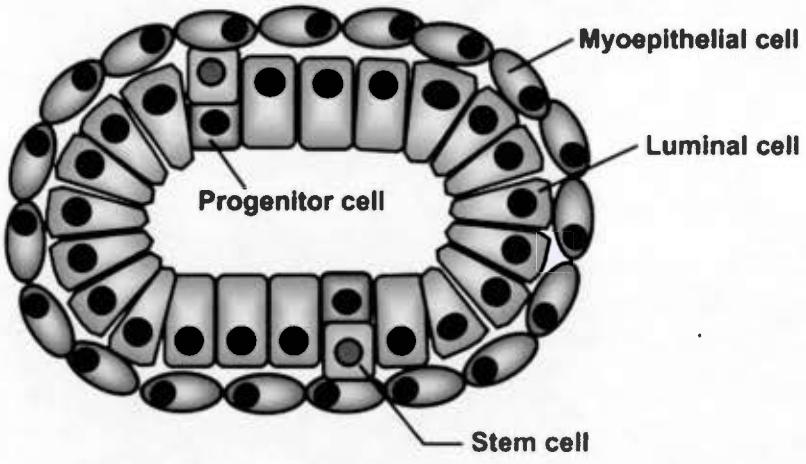


Figure 1.10 Structure cellulaire des glandes mammaires.

Les cellules myoépithéliales (« Myoepithelial ») entourent les cellules luminales (« Luminal ») dont l'origine proviendrait de cellules souches basales (« Stem cell ») différentiées en progénitrices (« Progenitor »). Image tirée de (Tiede et Kang, 2011).

Le cancer du sein est aujourd'hui considéré comme une maladie hétérogène que l'on peut classer selon le profil d'expression des gènes en quatre catégories principales (Perou *et al.*, 2000; Sorlie *et al.*, 2001; Sorlie *et al.*, 2003). La première, luminale A, est basée sur la sur-expression du gène codant pour le récepteur à œstrogène (ESR1, Figure 1.11a) et la tumeur retrouvée chez les patients est généralement associée à un pronostic positif (Figure 1.11b). Une autre catégorie, nommée luminale B, exprime aussi l'ESR1 (Figure 1.11a), mais est relativement plus agressive dans son potentiel métastatique avec un grade de tumeur relativement plus élevé (Figure 1.11b). Les tumeurs de type luminale seraient, par hypothèse, composées majoritairement de cellules myoépithéliales différentiées (Figure 1.11c) La troisième catégorie, HER2, exprime fortement le gène ERBB2 (un récepteur pour les facteurs de croissance épidermiques) et est associée à un pronostic moins favorable (Ignatiadis *et al.*, 2009). Enfin, la dernière catégorie, appelée triple négative, n'exprime pas l'ARNm des récepteurs des autres catégories en plus du récepteur à progestérone (PR) et se divise normalement en trois sous-types. Le premier, appelée basal (ou « basal-like »), exprime des marqueurs spécifiques aux cellules basales (*e.g.* cytokératines CK5) en plus de sur-exprimer beaucoup de gènes peu ou pas exprimés dans les autres catégories (*e.g.* la famille RUNX, Figure 1.11a) (Sotiriou et Pusztai, 2009). Le pronostic du sous-type basal, et par le fait même des autres sous-types de la catégorie triple négative, est généralement très défavorable (Figure 1.11b). Toujours dans cette catégorie, le sous-type normal (« normal-like », Figure 1.11a) possède un profil d'expression des gènes similaire à celui du tissu mammaire normal. Ce sous-type n'est toujours pas reconnu officiellement comme catégorie en raison du fait qu'il n'a jamais pu être isolé par micro-dissection et qu'il pourrait être le résultat de contamination par du tissu sain dans les mesures d'expression des gènes (Weigelt *et al.*, 2010). Faisant aussi partie de la catégorie des tumeurs triple-négatives, le dernier sous-type est caractérisé par l'expression de gènes propres aux cellules souches et mésenchymateuses et à une faible expression des gènes de la famille des claudines (« Claudin-low », Figure 1.11c) (Prat *et al.*, 2009).

Bien qu'il existe aujourd'hui plusieurs outils prédictifs pour aider dans le pronostic du cancer du sein, les différentes catégories et les sous-types du cancer du sein ne sont pas encore utilisés comme signatures d'expression des gènes afin de mieux cibler les traitements à administrer pour les patients dont le profil d'expression des gènes de leur tumeur correspond à l'une de ces catégories (Dai *et al.*, 2015). Avant d'utiliser des outils basés sur les signatures d'expression, il est primordial de mieux définir les catégories et sous-types de tumeurs du sein sur la base de types de données complémentaires et de méthodes intégrant différents niveaux de données (Kristensen *et al.*, 2014).

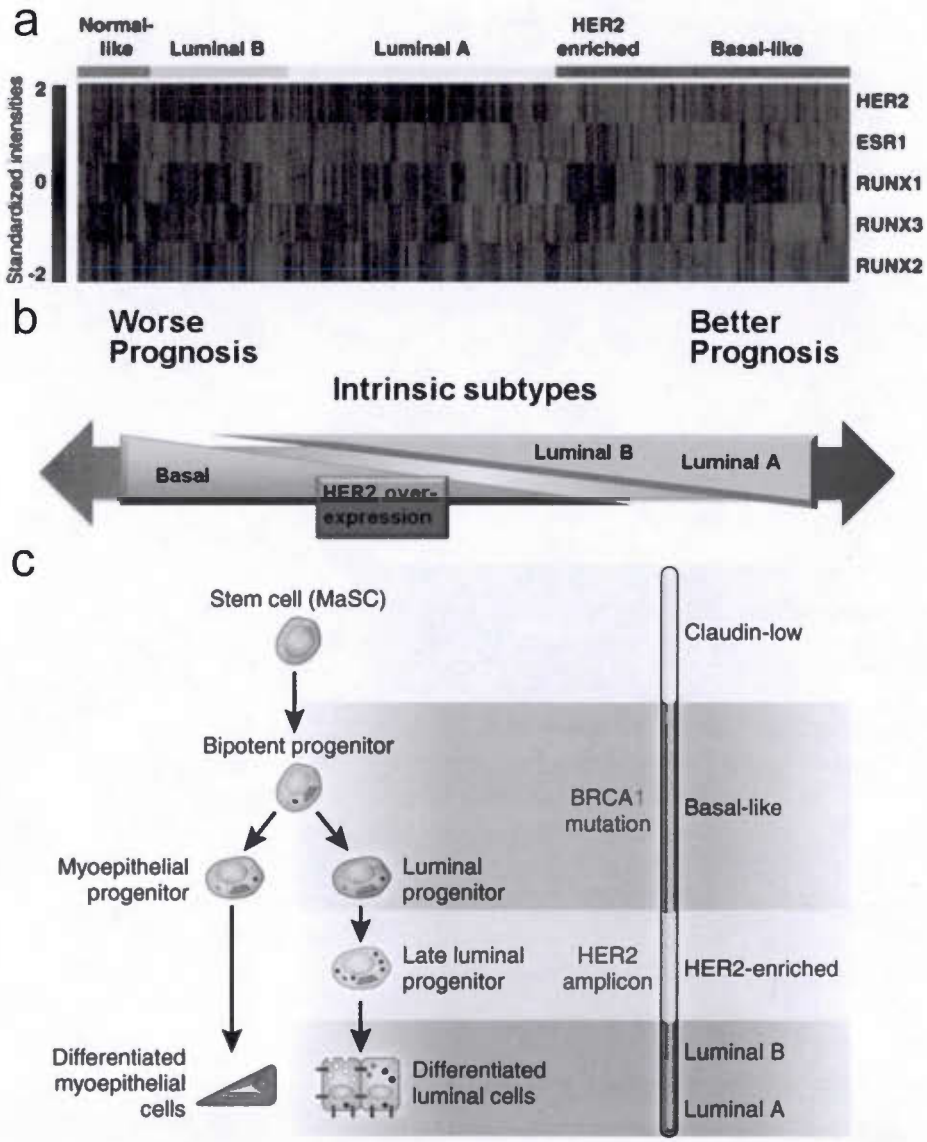


Figure 1.11 Caractéristiques des différentes catégories et sous-types de cancer du sein.

(a) Heatmap d'expression de marqueurs spécifiques à chacune des catégories de cancer du sein. Le gène ERBB2 (alias HER2) est sur-exprimé dans les tumeurs de types « HER2-enriched » alors que le récepteur à œstrogène ESR1 est sur-exprimé dans les catégories luminales (A et B). Les tumeurs de la catégorie triple négative comprennent les sous-types « Normal-like », qui ne sur-expriment pas de marqueur particulier, et les « Basal-like » qui expriment ou sur-expriment des gènes associés aux cellules épithéliales de type basale (ici la famille RUNX). Image tirée de (Chimge et Frenkel, 2013). **(b)** Pronostic des différentes catégories de cancer du sein. Image tirée de (Dai *et al.*, 2015). **(c)** Hypothèse d'associations des catégories de cancer du sein avec les différents stades évolutifs des cellules épithéliales. Image tirée de (Prat *et al.*, 2009).

1.4.2 Approches existantes pour l'analyse fonctionnelle des tumeurs mammaires

Une façon de caractériser davantage les catégories de cancer du sein est d'intégrer plusieurs types de données afin de complémenter l'information fournie par la quantification du niveau d'expression des gènes par RNAseq ou puces à ADN (Dai *et al.*, 2015). Pour ce faire, plusieurs bases de données sont disponibles. D'un côté, les données dites « primaires » proviennent directement des tumeurs de patients et couvrent l'essentiel de ce qui est mesurable dans une cellule (*e.g.* niveau d'accumulation des ARNm, des protéines, des phosphoprotéines et des métabolites, nombre de copie des gènes dans les chromosomes, niveau de méthylation de la chromatine). De l'autre côté, les données « secondaires » ou dérivées), proviennent soit d'un traitement statistique des données primaires (*e.g.* co-expression, profil mutationnel), soit d'une multitude d'annotations n'ayant pas, à priori, d'association avec le cancer (*e.g.* voies métaboliques, IPP, associations gène-maladie) (Kristensen *et al.*, 2014).

Pour les données primaires, le consortium TCGA a mis en évidence le fait que les différentes catégories de cancer du sein peuvent être différentiées non seulement sur la base de l'expression des gènes, mais aussi sur celle de l'expression des protéines, du nombre de copies des gènes et sur le profil mutationnel de certains gènes (The Cancer Genome Atlas Research *et al.*, 2013). Ces types de données peuvent donc être intégrés de façon complémentaire pour mieux caractériser les différentes catégories de tumeurs mammaires et permettent d'offrir, dans le futur, des tests de pronostic basés sur d'autres types de données que l'expression des gènes. L'intégration de plusieurs types de données a permis d'identifier des sous-types à l'intérieur des principales catégories de cancer du sein (Nasser *et al.*, 2011; Netanely *et al.*, 2016). En effet, en intégrant les données d'expression et de méthylation des gènes, Nasser *et al.* (2011) ont démontré l'existence de deux sous-types dans les tumeurs de genre « basal-like » qui sont associés à des taux de survie différents. Plus récemment, la même approche

d'intégration a été appliquée à des données provenant de tumeurs de la catégorie luminale A et a montré aussi une différence significative au niveau de la survie des patients de deux sous-types potentiels de cette catégorie (Netanely *et al.*, 2016). Ces résultats montrent que les stratégies d'intégration de plusieurs types de données pour identifier les principales catégories de tumeurs sont plus informatives que l'utilisation de données d'expression uniquement.

Une autre façon de faire consiste à dériver les données primaires (données expérimentales à grande échelle) en réseaux de gènes. Dans ces réseaux, les gènes sont liés si la force d'un type de données en particulier (*e.g.* co-expression, co-mutation, méthylation) est supérieure à la moyenne. On intègre ensuite des données secondaires (données dérivées ou annotations) afin de conférer des propriétés fonctionnelles à l'ensemble des gènes d'un réseau (Kristensen *et al.*, 2014). Enfin, on construit des modèles prédictifs par apprentissage machine ou des signatures de données permettant de distinguer les gènes clés (impliqués dans la progression tumorale) de ceux qui ne sont pas impliqués. De la même façon, les voies de signalisation dominantes ou actives lors de la progression tumorale peuvent être distinguées. Ces informations peuvent servir éventuellement à l'élaboration de thérapies spécifiques à une catégorie ou au pronostic de la maladie (Kristensen *et al.*, 2014). La prochaine section couvrira en détails certaines approches intégratives ayant mené à l'identification de gènes ou de voies clés pour les différentes catégories de cancer du sein.

1.4.3 Identifier des gènes conducteurs ou des voies de signalisation clés dans la progression des tumeurs mammaires

Plusieurs études ont montré l'avantage d'utiliser ou de dériver une ou plusieurs données primaires et les intégrer aux informations fonctionnelles fournies par les voies

de signalisation et les fonctions biologiques (Dutta *et al.*, 2012; Natrajan *et al.*, 2010; Suo *et al.*, 2015; Tian *et al.*, 2014). Cette intégration est faite afin d'identifier des gènes clés ou des voies de signalisation importantes de la maladie dans chacune des catégories de tumeurs du sein (Dutta *et al.*, 2012; Natrajan *et al.*, 2010; Suo *et al.*, 2015; Tian *et al.*, 2014). Les gènes clés peuvent être considérés comme des gènes « conducteurs » car leurs mutations affectent directement le processus d'oncogenèse (Stratton *et al.*, 2009) À l'opposé, les gènes dits « passagers » n'interviennent pas dans la progression des tumeurs (Stratton *et al.*, 2009). La définition peut être étendue à l'altération, par des mécanismes épigénétique (*e.g.* méthylation) ou encore par duplication, du gène entraînant la progression tumorale qui peut aussi être considéré comme un gène conducteur (De Carvalho *et al.*, 2012; Dutta *et al.*, 2012). Il est à noter qu'il n'existe pas de consensus au sein de l'appellation de gènes conducteurs et passagers dans la littérature.

L'identification de ces gènes nécessite habituellement de déceler des mutations dans ces gènes, provenant des échantillons de tumeurs, par séquençage de prochaine génération. La découverte de ces mutations se fait en comparant la séquence des gènes dans les tumeurs à celle de tissus sains du même patient (Figure 1.12a). L'expression des gènes, mesurée dans chacune des catégories de cancer du sein, est aussi utilisée dans ces études afin d'identifier des gènes mutés et sur-exprimés dans ces catégories de tumeurs (Figure 1.12a). Enfin, les gènes sont classés hiérarchiquement selon leurs enrichissements fonctionnels dans le réseau de sur-expression des gènes (dérivé à partir des gènes différentiellement exprimés entre le tissu normal et la tumeur) pour identifier les gènes conducteurs (Figure 1.12a) (Suo *et al.*, 2015).

Une autre manière d'identifier les gènes conducteurs dans le cancer est de se baser sur leurs rôles fonctionnels dans la cellule (Chen *et al.*, 2013; Dutta *et al.*, 2012) ou sur leurs propriétés topologiques dans le réseau d'IPP (Jonsson et Bates, 2006; Piñero *et al.*, 2016; Porta-Pardo *et al.*, 2015). Toujours sur la base des différentes catégories de cancer du sein, Dutta *et al.* (2012) ont montré qu'il était possible d'intégrer

simultanément les données d'expression et de nombre de copies des gènes dans les réseaux d'IPP, de régulation transcriptionnelle et de signalisation pour identifier des gènes conducteurs pour chacune des catégories de tumeurs. Pour ce faire, ils ont d'abord sélectionné les gènes pour lesquels le nombre de copies corrèle avec son expression dans chacune des catégories de tumeurs. Ils ont ensuite appliqué deux filtres statistiques dont l'un est basé sur l'expression différentielle entre une catégorie de patients et les autres catégories dans le but de sélectionner les groupes de gènes qui sont sur- ou sous-exprimés. Le deuxième filtre consiste à retenir les sous-groupes de gènes qui sont co-exprimés dans chacune des catégories (Dutta *et al.*, 2012). Ils ont ensuite qualifié ces sous-groupes de réseaux conducteurs desquels ils ont identifié les Hubs (gènes hautement connectés) comme étant des gènes conducteurs. Finalement, ils ont validé expérimentalement 30 gènes conducteurs versus 10 gènes passagers potentiels, pris au hasard, dans 13 lignées cancéreuses en regardant la viabilité des cellules après un traitement avec des siRNA dirigés contre ces gènes. Aussi, ces gènes conducteurs étaient, pour les tumeurs de type luminale, impliqués dans l'apoptose et responsable de la résistance aux traitements hormonaux et associés à la transition épithéliale-mésenchymateuse (EMT) dans les tumeurs de catégorie triple-négative (Dutta *et al.*, 2012).

De façon plus générale, l'utilisation du réseau d'IPP, pour identifier des gènes conducteurs dans le cancer du sein ou autres cancers, a été exploité dans plusieurs études grâce à la position centrale qu'occupent ces gènes à l'intérieur du réseau d'interactions (Jonsson et Bates, 2006; Piñero *et al.*, 2016; Porta-Pardo *et al.*, 2015). En effet, les gènes conducteurs sont plus nombreux à coder pour des Hubs (protéines hautement connectées) dans le réseau d'IPP (Jonsson et Bates, 2006; Piñero *et al.*, 2016; Porta-Pardo *et al.*, 2015). Ils ont aussi tendance à coder pour des protéines nommées « bottlenecks » (Piñero *et al.*, 2016; Porta-Pardo *et al.*, 2015). Les protéines « bottlenecks » sont associées à un haut degré de centralité intermédiaire ou « betweenness » où beaucoup de courts chemins traversent ce noeud. Les

« bottlenecks » tendent de plus à être codés par des gènes essentiels (voir section 1.2.1) (Yu *et al.*, 2007). Ces caractéristiques topologiques des gènes conducteurs dans le réseau d'IPP pourraient donc servir éventuellement à complémenter les méthodes d'intégration existantes pour découvrir des nouveaux gènes conducteurs dans les différentes catégories de cancer du sein.

Dans un autre ordre d'idée, une des difficultés rencontrées dans l'application clinique de la découverte de listes de gènes conducteurs est que ces gènes en question ne sont généralement pas de bonnes cibles thérapeutiques, en particulier dans les tumeurs de catégorie triple-négative. En effet, comme leurs mutations conduisent souvent à la progression tumorale (*e.g.* TP53) (Stratton *et al.*, 2009), une molécule dirigée contre l'un de ces gènes pourraient évidemment aggraver la situation (Stratton *et al.*, 2009). Il est donc préférable d'identifier les réseaux fonctionnels entourant les gènes conducteurs afin de trouver des cibles thérapeutiques secondaires (Dutta *et al.*, 2012). Une des façons de faire est de mettre en évidence les voies de signalisation/métabolisme clés de chacune des catégories de tumeurs selon l'état d'activation de ces voies (régulées à la hausse ou à la baisse) (Figure 1.12b) (Dutta *et al.*, 2012; Natrajan *et al.*, 2010; Tian *et al.*, 2014). L'approche générale consiste à déterminer, pour chacun des sous-types de tumeurs, si certains gènes d'une même voie sont corrélés pour leurs niveaux d'expression et leurs nombres de copies (Dutta *et al.*, 2012; Natrajan *et al.*, 2010), ou encore s'ils sont sur-exprimés ou dupliqués par rapport à des données de tissus sains (Tian *et al.*, 2014). On peut ensuite calculer des enrichissements en voies de signalisation dans le voisinage du réseau d'IPP (Dutta *et al.*, 2012), ou encore utiliser directement les annotations provenant des voies de signalisation/métabolisme pour déterminer l'appartenance de chaque groupe de gènes identifiés (Natrajan *et al.*, 2010; Tian *et al.*, 2014). L'avantage de cette approche est qu'elle permet non seulement de découvrir des gènes conducteurs à l'intérieur des voies clés, mais aussi d'identifier les voies qui sont différentiellement activées dans chaque sous-type de tumeurs.

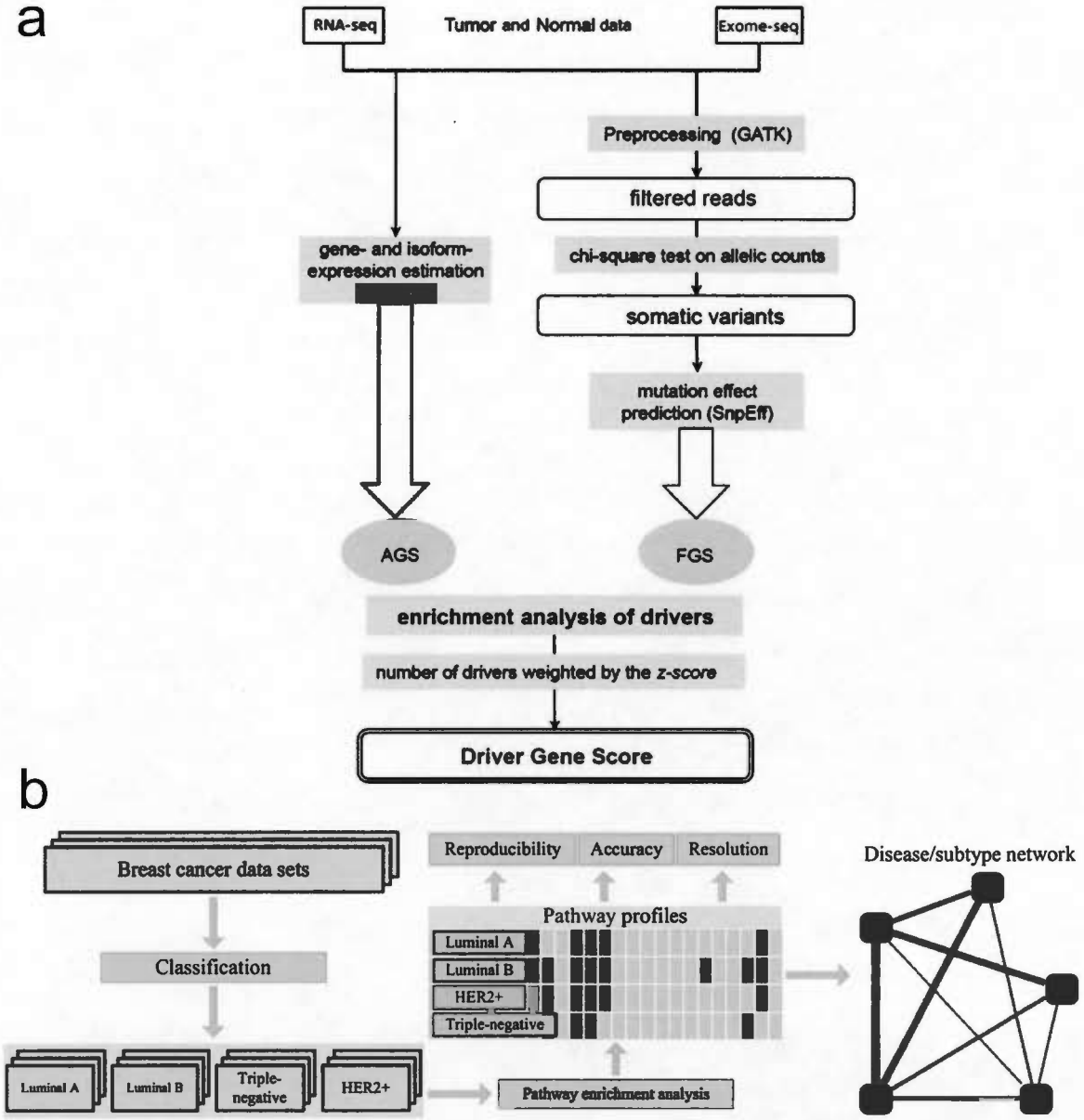


Figure 1.12 Exemples d'approches intégratives pour identifier des gènes conducteurs et des voies de signalisations clés.

(a) Identification de gènes conducteurs en comparant l'expression des gènes (« RNA-seq ») et le profile mutationnel (« Exome-seq ») des tumeurs de chaque catégorie et des tissus normaux suivi par l'analyse fonctionnelle (« enrichment analysis ») afin d'identifier les gènes conducteurs dans le réseau dérivé des expressions selon un score minimum (« Driver Gene Score »). Image tirée de (Suo *et al.*, 2015). **(b)** Analyse des voies de signalisation responsables de la tumorigénèse de chaque catégorie de cancer du sein suite à l'intégration des voies (« Pathway profiles ») et à la classification des données d'expression des gènes. Image tirée de (Tian *et al.*, 2014).

Des études ont montré que le niveau d'expression des gènes conducteurs ne corrèle pas avec le nombre de copies de ces gènes pour plus de la moitié des gènes conducteurs testés (Gatza *et al.*, 2014; Myhre *et al.*, 2013). L'utilisation de cette corrélation pour identifier des gènes et des voies clés dans les différentes catégories de tumeurs n'est donc pas, a priori, idéale au niveau de la sensibilité des modèles prédictifs des gènes conducteurs.

Globalement, les analyses intégratives existantes pour découvrir des voies clés et/ou des gènes conducteurs dans les différentes catégories de cancer du sein ont un potentiel certain au niveau thérapeutique mais sont tout de même limitées pour plusieurs raisons (Kristensen *et al.*, 2014). Parmi celles-ci, les conclusions de la plupart de ces études sont généralement partielles en raison des nombreux exemples (*e.g.* gènes potentiellement conducteurs) qui passent inaperçus par manque de sensibilité. Pour augmenter la sensibilité, il est préférable d'analyser un sous-ensemble de données qui couvre au maximum les exemples, appelé « ensemble de référence » (ou « gold standard » en anglais), afin de traiter l'ensemble des exemples par différents modèles prédictifs ou signatures de données adaptées (Kristensen *et al.*, 2014).

Une autre raison réside dans le fait que peu de méthodes existantes sont validées *in vivo*. En effet, bien que des validations *in silico* sur des ensembles de données externes de patients soient faites dans la plupart de ces études (reproductibilité), peu d'entre-elles valident de façon *ex vivo*, et encore moins valident leurs découvertes sur des organismes vivants avec les techniques modernes (*e.g.* criblage par ARN interférant, édition CRISPR/cas9) (Kristensen *et al.*, 2014). Finalement, d'un point de vue statistique, beaucoup d'études intégratives ont utilisé des ensembles de données à grande échelle provenant de plusieurs plateformes expérimentales ou encore, une multitude de types de données différents qui font tous augmenter le problème de la « dimensionnalité » (*i.e.* pendre en compte trop de niveaux de données dans l'analyse tend à augmenter la dimension du problème et les erreurs). Autrement dit, il vaut mieux utiliser les données provenant d'une même plateforme pour diminuer la variabilité et

évaluer, de façon individuelle, chacun des types de données pour leur pertinence face aux buts visés (Kristensen *et al.*, 2014). En suivant ces recommandations, les découvertes *in silico* mèneront éventuellement à de meilleurs pronostics et à des thérapies ciblées pour chacune des catégories de cancer du sein.

1.4.4 Identification de cibles thérapeutiques dans le cancer du sein

L'une des thérapies les plus prometteuses récemment identifiée pour le cancer du sein est la découverte des inhibiteurs PARP (une enzyme clé dans la réparation de l'ADN) dont plusieurs molécules sont actuellement en essais cliniques pour traiter les tumeurs arborant la mutation BRCA1/2 (Sistigu *et al.*, 2016). La particularité de cette découverte est qu'elle repose sur l'existence d'une interaction génétique négative (synthétique létale, voir section 1.1) entre PARP-1 et BRCA1/2. En effet, l'inhibition de PARP-1 empêche d'abord le recrutement des facteurs de réparation de l'ADN au niveau du simple brin. Ensuite, en l'absence de la protéine BRCA1/2 fonctionnelle, le double brin d'ADN après réplication n'est pas réparé non plus et conduit donc à la mort de la cellule cancéreuse (Figure 1.13) (Sistigu *et al.*, 2016). Mais comme elles ont l'habilité, tout comme les autres types de cancer, de réorganiser leurs mécanismes de réparation de l'ADN grâce au « switch » (changement rapide) de la fonction de certain gènes conducteurs (Gavande *et al.*, 2016). Le développement de thérapies combinées ciblant plusieurs voies de réparation de l'ADN est donc à prioriser dans l'avenir (Gavande *et al.*, 2016).

Pour ce faire, il faut d'abord développer des méthodes pour mieux caractériser les voies et les réseaux fonctionnels qui régulent et qui influencent les mécanismes de réparation de l'ADN (Gavande *et al.*, 2016). L'intégration de données pour découvrir les voies clés et les gènes conducteurs impliqués dans la régulation des mécanismes de

réparation de l'ADN offre donc la possibilité de mieux cibler les thérapies par les inhibiteurs de PARP en les combinant avec d'autres médicaments.

L'exemple précédent illustre bien le phénomène du revirement des gènes conducteurs rencontré dans le cancer, *i.e.* plusieurs gènes semblent adopter une fonction particulière dépendamment du contexte dans lequel ils se retrouvent (Stepanenko *et al.*, 2013). Un autre exemple du « switch » des gènes conducteurs dans le cancer, mais peu documenté cette fois, concerne le métabolisme du glucose. Brièvement, il a été montré que les cellules cancéreuses, à l'inverse des cellules normales, utilise la glycolyse comme source principale d'énergie même en présence d'oxygène (Warburg, 1956). Ce phénomène impliquerait que certains gènes contrôlent le degré d'activation des voies impliquées dans le métabolisme du glucose avec d'autres voies pour favoriser la glycolyse aérobique (Elf et Chen, 2014). Par exemple, le gène PKM2, exprimé durant le développement embryonnaire, converti le phosphoénolpyruvate en pyruvate favorisant ainsi la glycolyse (Christofk *et al.*, 2008). Ce gène est exprimé dans les cellules cancéreuses et serait aussi impliqué dans des voies de signalisation (non-métabolique) en tant que facteur de transcription de plusieurs oncogènes incluant Oct-4, qui joue un rôle dans le maintien de la cellule dans un état de pluripotence (non-différentié) (Lee, J. *et al.*, 2008). PKM2 régulerait aussi la transcription du facteur d'induction de l'hypoxie HIF-1 (Luo *et al.*, 2011) et la β-caténine (Yang *et al.*, 2011). De plus, comme ce gène n'est pas exprimé dans toutes les cellules, il représente une bonne cible thérapeutique par des molécules dont plusieurs essais ont été effectués avec succès (Elf et Chen, 2014). Dans l'ensemble, étant donné que plusieurs gènes qui jouent un rôle dans le métabolisme énergétique des cellules cancéreuses interfèrent aussi avec d'autres processus biologiques. Il est nécessaire d'identifier, pour chacun de ces gènes, les voies et les processus dans lesquelles ils sont impliqués ainsi que leurs différents réseaux fonctionnels avec d'autres gènes. Ces réseaux de gènes pourraient mener à la découverte de thérapies combinées pour cibler le métabolisme particulier des cellules cancéreuses.

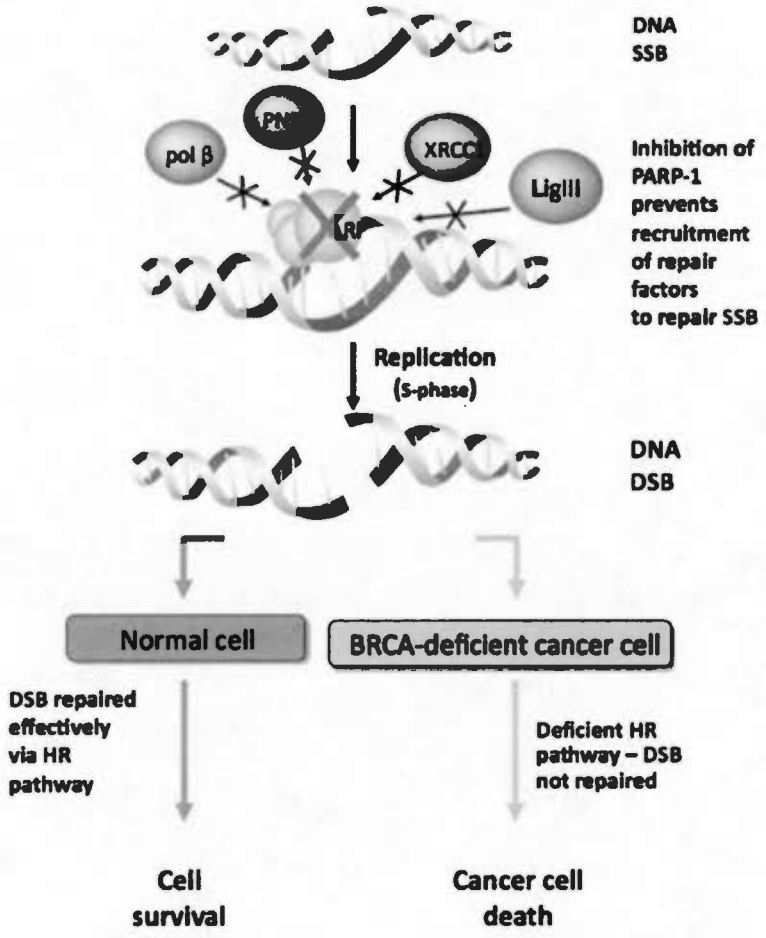


Figure 1.14 L'inhibition de PARP-1 comme thérapie utilisée dans les tumeurs déficientes pour les gènes de la famille BRCA.

L'inhibition de PARP-1 empêche d'abord le recrutement des facteurs de réparation de l'ADN au niveau du simple brin. Ensuite, en l'absence de la protéine BRCA1/2 fonctionnelle, le double brin d'ADN après réplication n'est pas réparé non plus et conduit à la mort de la cellule cancéreuse. Image tirée de (Ashworth, 2008).

1.5 Fondement de la thèse

Il est primordial de mieux comprendre les mécanismes et les contextes dans lesquels les gènes s'influencent entre eux pour régir le dynamisme moléculaire dans les cellules des organismes vivants. D'un côté, ces interactions entre les gènes sont au centre de la robustesse et de l'adaptabilité des organismes vivants à leur environnement, mais de l'autre, elles contribuent aussi aux désordres causés par les maladies chez l'humain.

Dans cette thèse, je fais l'analyse en profondeur d'un interactome génétique chez le nématode *Caenorhabditis elegans* et fais la description d'une approche de génomique intégrative qui permet de prédire des liens fonctionnels impliquant des gènes clés dans la reprogrammation métabolique du cancer du sein.

Dans le Chapitre II, il est question d'un retour sur l'ensemble des connaissances acquises au cours des dernière années sur les interactions génétiques et comment une meilleure compréhension du contexte entourant les différents types de données rattachées peuvent aider à prédire des nouvelles interactions génétiques.

Dans le Chapitre III, je présente la caractérisation en profondeur d'un interactome génétique de *C. elegans* liant des gènes se trouvant à l'intérieur des voies de signalisation/métabolique. Cette caractérisation permet de mettre en évidence une structure formée de six classes d'interactions génétiques ayant des propriétés distinctes. Ce modèle de réseau pourrait permettre de mieux comprendre le rôle des interactions génétiques au niveau de la robustesse des systèmes et le contrôle qu'elles exercent sur les processus biologiques.

Dans le Chapitre IV, je montre d'abord qu'il est possible de transposer à l'humain, une méthode d'intégration utilisée pour prédire des interactions génétiques chez le nématode *C. elegans*, afin de construire un réseau d'interactions fonctionnelles pour des sous-types de cancer du sein. Ce réseau, présentant des caractéristiques similaires à celui qui est présenté dans le Chapitre précédent, a été utilisé dans un modèle prédictif

pour découvrir des gènes potentiellement impliqués dans la reprogrammation métabolique des tumeurs mammaires.

Dans l'ensemble, l'approche et les résultats décrits dans cette thèse ouvre la voie à la caractérisation d'interactomes fonctionnels d'une multitude de maladies chez l'humain possédant des données d'expression des gènes. Ces caractérisations, basées sur nos résultats, pourraient aussi permettre de découvrir des nouveaux gènes impliqués dans différents contextes d'une maladie.

CHAPITRE II

LES RÉSEAUX D'INTERACTIONS GÉNÉTIQUES : MIEUX COMPRENDRE POUR MIEUX PRÉDIRE

L'ensemble de ce chapitre a été publié en décembre 2013 dans la revue *Frontiers in Genetics*, doi: 10.3389/fgene.2013.00290.

Boucher B, Jenna S. (2013) Genetic interaction networks: better understand to better predict. *Frontiers in Genetics*. 4 : 290.

Dans ce deuxième chapitre, il est question de la relation entre le génotype et le phénotype d'un individu. En informatique, la complexité de l'architecture interne d'un ordinateur peut être simplifiée en la représentant en niveaux d'abstraction. De façon similaire, pour mieux comprendre la relation entre le génotype et le phénotype, il est possible de diviser les systèmes biologiques en six niveaux d'abstraction. On montre que les données provenant de chaque niveau d'abstraction ont été utilisées pour prédire des interactions génétiques de même que les divers liens existant entre chaque niveau. On montre aussi que les méthodes existantes pour prédire des interactions génétiques chez *Caenorhabditis elegans* n'arrivent pas à couvrir l'ensemble des interactions mais seulement des portions de celui-ci selon la technique utilisée et qu'il est donc primordial de mieux définir la structure de l'interactome génétique avant de construire d'autres modèles prédictifs.

CONTRIBUTIONS

Pour cet article, j'ai assemblé, organisé et analysé toutes les données nécessaires afin de comparer les outils *in silico* existants pour la prédiction d'interactions génétiques chez *Caenorhabditis elegans*. J'ai aussi compilé toutes les statistiques et les principales conclusions provenant des études en génomique intégrative pour prédire des interactions génétiques ainsi que les différentes techniques utilisées.

Sarah Jenna a conceptualisé les niveaux d'abstraction des systèmes biologiques et a dirigé l'ensemble des analyses effectuées dans cet article.

Nous nous sommes partagés la rédaction de l'article sur ces différentes parties.

RÉSUMÉ

Une interaction génétique (IG) entre deux gènes indique généralement que le phénotype d'un double mutant diffère de ce qui est attendu de chaque mutant individuellement. Dans la dernière décennie, les études quantitatives à grande échelle, portant sur les interactions génétiques, ont été effectuées à l'aide du criblage synthétique par tableau génétique (ou « synthetic genetic array » en anglais) et par ARN interférant dans la levure et *Caenorhabditis elegans*. Ces études ont soulevé des questions concernant l'interprétation fonctionnelle des IGs, la relation entre les réseaux d'interactions génétiques et moléculaires, l'utilité des réseaux d'IGs pour déduire la fonction des gènes et leur co-fonctionnalité, la conservation évolutive des IGs, etc.

Alors que les IGs ont été utilisées pendant des décennies pour clarifier les voies de signalisation dans les modèles génétiques, leur interprétation fonctionnelle n'est pas toujours triviale. L'existence d'une IG entre deux gènes ne signifie pas nécessairement que ces deux gènes codent pour des protéines qui interagissent ou que les deux gènes sont exprimés dans la même cellule. En effet, une IG implique seulement que les deux

gènes partagent une relation fonctionnelle. Ces deux gènes peuvent être impliqués dans le même processus biologique ou dans la même voie ; ils peuvent autrement être impliqués dans des voies compensatoires qui n'ont pas, à première vue, de liens évident. Compte tenu de l'opportunité de mieux comprendre la fonction des gènes, la relation génétique, la robustesse et l'évolution fournie par une cartographie complète du génome par des IGs, plusieurs approches *in silico* ont été utilisées pour prédire des IGs dans les organismes unicellulaires et multicellulaires. La plupart de ces méthodes utilise l'intégration de données de manière pondérée. Dans cet article, nous passerons en revue les connaissances acquises sur les réseaux d'IGs chez les métazoaires en regardant de plus près leur relation avec les voies, les processus biologiques et complexes moléculaires, mais aussi dans leur modularité et leur organisation. Nous allons également examiner les différentes méthodes *in silico* développées pour prédire des IGs et verrons comment les connaissances acquises sur les réseaux d'IGs peuvent être utilisées pour concevoir des outils prédictifs avec de meilleures performances.

2.1 Abstract

A genetic interaction (GI) between two genes generally indicates that the phenotype of a double mutant differs from what is expected from each individual mutant. In the last decade, genome scale studies of quantitative genetic interactions were completed using mainly synthetic genetic array (SGA) technology and RNA interference in yeast and *C. elegans*. These studies raised questions regarding the functional interpretation of GIs, the relationship of genetic and molecular interaction networks, the usefulness of GI networks to infer gene function and co-functionality, the evolutionary conservation of GI, etc.

While GIs have been used for decades to dissect signaling pathways in genetic models, their functional interpretations are still not trivial. The existence of a GI between two genes does not necessarily imply that these two genes code for interacting proteins or that the two genes are even expressed in the same cell. In fact, a GI only implies that the two genes share a functional relationship. These two genes may be involved in the same biological process or pathway; or they may also be involved in compensatory pathways with unrelated apparent function. Considering the powerful opportunity to better understand gene function, genetic relationship, robustness and evolution, provided by a genome-wide mapping of GIs, several *in silico* approaches have been employed to predict GIs in unicellular and multicellular organisms. Most of these methods used weighted data integration. In this article, we will review the later knowledge acquired on GI networks in metazoans by looking more closely into their relationship with pathways, biological processes and molecular complexes but also into their modularity and organisation. We will also review the different *in silico* methods developed to predict GIs and will discuss how the knowledge acquired on GI networks can be used to design predictive tools with higher performances.

Keywords: genetic interaction, network, conservation, prediction, *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, genomics

2.2 What is a genetic interaction?

2.2.1 General definition

The term genetic interaction (GI) covers a group of functional relationships between genes. One kind of these relationships, called epistasis, was first defined by W. Bateson in 1909 (Bateson et Mendel 1909). Biological epistasis was then described as the effect of one allele masking the effect of another one (Moore 2003). Nine years later statistical

epistasis, originally called "epistacy", was described by R. A. Fisher as a significant deviation of the phenotype of a double mutant from what is expected considering the phenotypes of the single mutants (Fisher 1919).

This statistical epistasis enabled the identification of an array of different GIs. One popular classification of these GIs consists of dividing them in two main classes: the negative and the positive interactions. The negative GIs, called also aggravating or synergistic interactions, refer to an observed phenotype higher than expected when considering the phenotypes of single mutants and assuming that the mutated genes function independently one from the other (Figure 2.1). A synthetic lethal interaction, which is an extreme case of negative GI, occurs when both single mutants are viable but the double mutant is lethal (Figure 2.1). At the opposite, the positive GIs can be subdivided in buffering/alleviating interactions where the biological effect of an allele is mitigated by a second one, and also the suppressor interactions in which the double mutant is healthier than the sickest single mutant (Figure 2.1).

As mention above, identification of statistical epistasis depends on the calculation of the expected phenotype of the double mutant considering the phenotype of the single mutants and assuming a functional independency of the two mutated genes. Several models exist and are used to estimate this expected value. For developmental and population geneticists, the quantitative assessment of a phenotype involves the statistical assessment of its penetrance – the statistical occurrence of a phenotype in a group of known genotypes – considering its expressivity. A threshold is then usually set for the expressivity of the phenotype – the degree to which the phenotype expression differs among individuals – to measure the penetrance (Miko 2008).

The development of additive, multiplicative, Min and Log models to calculate the expected phenotype of double mutants was mostly motivated by the development of systematic and large-scale screening of GIs, especially in the yeast *Saccharomyces cerevisiae* (Tong *et al.* 2001; Collins *et al.* 2007; Jasnos et Korona 2007; Costanzo *et al.* 2010). These studies identified GIs based on fitness measurements (Figure 2.1B), a

class of phenotype that is measured in terms of population allele frequency (Wolf, Brodie et Wade 2000; Otto et Lenormand 2002; Puniyani, Liberman et Feldman 2004), growth rate, or number of progeny of mutant strain relative to wild-type (Elena et Lenski 1997; Szafraniec *et al.* 2003; Segre *et al.* 2005; Sanjuan et Elena 2006; St Onge *et al.* 2007). The additive and multiplicative models, originally used by developmental geneticists (Figure 2.1A) and fitness measurements in yeast (Figure 2.1B) respectively, consider the expected phenotype of a double mutant to be the sum (or the product) of the phenotypes measured for the single mutants if the two mutated genes function independently one from the other (Mani *et al.* 2008). The Log model has been specifically designed to identify GIs from measurements on a logarithmic fitness scale (Mani *et al.* 2008). The Min model considers that for non-interacting genes, the fitness of the double mutant should be similar to the fitness of the less-fit single mutant. Although these models agree under certain circumstances, they often diverge dramatically (Mani *et al.* 2008). For example, while the Min model appears to be highly suitable for pairs of genes with more extreme single-mutant defects, this model is clearly not ideal for defining alleviating interactions and more particularly, several epistatic interactions for which a double mutant phenotype is similar to that of the single mutant with the most severe phenotype (St Onge *et al.* 2007). Unfortunately, GIs identified using this model account for a large part of all GIs found in interaction databases. This tends to bias the yeast genetic interactome against this later kind of GIs (Mani *et al.* 2008). Identification of GIs considering several of these models would then be an appropriate approach to enable fair comparison and integration of GIs from different screening pipeline into a homogeneous GI interactome.

2.2.2 Levels of abstraction in biological systems

Mapping of GI networks is an endeavour that attracted more attention with the emergence of network and systems biology approaches. Network biology consists in simplifying complex biological systems into different layers of graphical representations in which nodes correspond to physical elements (genes, protein, metabolites, RNA, etc.) and edges refer to different relationships between these elements. Systems biology, and more particularly integrative genomics, aims to better understand the structure and the functioning of the system through integration of these different networks (Ge *et al.* 2001).

In computer sciences, organization of systems into several abstraction levels aims to hide a certain level of detail to allow the programmer to focus on a given problem. For a computer, the lower level of abstraction would contain details on the hardware while the higher level will represent the logic of the program. In agreement with this approach, a systems biologist will consider a biological system with all its complexity and identify, from the genomic sequence to the phenotype, different levels of abstractions. At the lower level of this conceptual structure, we would find several networks representing the physical structure and organisation of the genome. In these networks, nodes could be genes/coding sequences, single-nucleotide polymorphisms (SNPs) or coding sequences linked by edges representing their physical proximity and organisation within chromosomes, their homology etc. (Figure 2.2, level I). The second level of abstraction would represent the expression of that genome into physical components: proteins and RNA. Edges between these elements would indicate that they are co-expressed in different contexts or that their expression profiles throughout multiple experimental conditions are highly correlated (Figure 2.2, level II) (Ge *et al.* 2001; Vidal, Cusick et Barabasi 2011). The third level of abstraction would represent physical interactions between different elements – protein-protein (PPI), protein-DNA (PDI) or protein-RNA (PRI) interactions (Figure 2.2, level III) (Vidal, Cusick et

Barabasi 2011). The fourth level of abstraction will allow the visualisation of the functional relationships linking these physical elements. This level would contain GI networks, signaling and metabolic pathways (Figure 2.2, level IV). The fifth level would represent biological processes. This level would contain networks where proteins implicated in the same biological process would be linked by an edge (Figure 2.2 level V). The sixth and last level of abstraction would represent phenotypes and show the relationships between elements associated with similar phenotypes and diseases (Figure 2.2, level VI). Breaking down through the different levels of abstraction aims to understand the molecular basis of higher levels. A huge amount of effort is being made to enable such a breaking down and to establish the links and the dynamics underlying the relationships between networks located at the different levels. The relationship between the second (gene expression) and the third level (mainly PPI and PDI) has been well documented. Some studies showed that interacting proteins are more likely to be encoded by genes with similar expression profiles than non-interacting proteins (Ge *et al.* 2001; Grigoriev 2001; Mrowka, Patzak et Herzel 2001; Jansen, Greenbaum et Gerstein 2002; Kemmeren *et al.* 2002). Similarly, expression profiles can be used to understand the organisation and dynamics of protein interaction networks through functional characterisation of highly-connected nodes (Hubs). For example, Hubs have been divided into "party" and "dating" Hubs. The former class of Hubs corresponds to proteins that tend to be co-expressed with their protein partners while the later ones are not (Han et al., 2004). Party Hubs have then been proposed to interact with all their protein partners in all biological conditions, while dating Hubs may interact with subgroups of their protein partners in certain conditions and/or environments (Han *et al.* 2004). PPIs and PDIs can also be used to understand the molecular basis of co-expression (Lee *et al.* 2002; Segal *et al.* 2003; Yu *et al.* 2003; Luscombe *et al.* 2004).

The link between the third (molecular interactions) and the fourth level (functional interactions) has also been investigated. Notably, signalling and metabolic pathways

were shown to be enriched in PPIs and PDIs (Vidal, Cusick et Barabasi 2011). It is important to notice that, as detailed in the section 2.3 of this review, the term pathway has been assimilated in several papers as PPI and PDI modules – PPI/PDI subnetworks with a high density of links – or as dense GI network structures (Kelley et Ideker 2005; Bellay *et al.* 2011). Here, signaling and metabolic pathways will be described as a group of molecules functioning together and most of the time, in cascade to control a biological function. As detailed in the following sections, GI networks are also linked to PPI and PDI networks (see section 2.3). This link is however less evident than the link between PPI/PDI networks and signalling/metabolic pathways (see section 2.3).

The relationship existing between the level six (phenotypes and diseases) and the level four (functional interactions) motivated the construction of pathway databases such as Reactome (Joshi-Tope *et al.* 2005) or the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et Goto 2000), and is at the forefront of the research effort to identify therapeutic targets and pharmaceutical compounds (Yuryev 2012).

The link between the levels four (functional interactions) and five (biological processes) is clear for signaling and metabolic pathways. Each signaling pathway, for example the EGF receptor / Ras / MAP kinase pathway, involves proteins that can be grouped based on their implication in the control of various biological processes, e.g. endocytosis, Ras regulation, actin cytoskeleton remodeling, kinase activity/phosphorylation, etc.

Abstractions levels can also be linked to distant levels. For example, GIs are shown to be enriched in co-expressed genes (Zhong et Sternberg 2006; Lee, A.Y. *et al.* 2010) (link between the fourth and the second level). Similarly, integration of the sixth level (phenotype) to the third (PPI) permitted the construction of the human disease interactome. This interactome was proposed to support the existence of disease specific functional modules and to help the molecular characterization of the protein products of disease genes (Goh *et al.* 2007).

Integration of different networks within or across abstraction levels brings substantial information on the structure of the system, and to some extent, information about its dynamics (Han *et al.* 2004). These pieces of information constitute, as described in this review, the baseline for the construction of predictive tools used to enrich and complete sparse networks.

We will focus, in this review, on the fourth level and more particularly, on GI networks. While this kind of functional relationship is linked to higher and lower levels of abstraction, most of these links appear much less clear than those involving signaling and metabolic pathways. We can then wonder if mapping such a network is of biological interest: would it bring complementary information to those brought from pathways dissection and significantly help understanding the functioning of the system?

2.2.3 Why constructing a catalog of genetic interactions?

There are two main reasons why mapping GI networks is of biological interest. The first one is to understand the mechanisms underlying the robustness of biological systems. How the system compensate for the loss or alteration of a biological function or the alteration of its environment?

Unnecessary genes do not exist in biological systems and would be eliminated through evolutionary processes (Stern et Orgogozo 2009). So, why 73% of these necessary genes appears not to be essential (Giaever *et al.* 2002)? Because compensatory relationships exist between genes, pathways and biological processes. Therefore, mapping of GIs appears to be the best way to identify these compensatory phenomena. In addition to the high contribution this mapping will bring to basic sciences, it is also

of high interest for translational research. Biological robustness is indeed, a major problem in the pharmaceutical industry with the development of resistance to therapeutic agents, particularly to anti-cancer chemotherapies (Edelman, Eddy et Price 2010). Identification of compensatory relationships between genes and pathways, through mapping of GIs, appears then as an avenue that needs to be explored in parallel with the dissection of the pathways themselves.

The second reason is associated with the still mysterious relationship existing between genotype and phenotype. Population geneticists highlighted the intricate complexity of genetic variations and how positive and negative relationships between alleles influence phenotypical outcome (Gibson 2010). Cancer modifier loci, including "susceptibility" or "resistance" alleles, are good examples of genetic variations affecting a patient phenotype, here the aggressiveness of the tumor phenotype (Dragani 2003). Similarly, GIs and more particularly digenic synthetic GIs may underlie many common diseases that are familial but not Mendelian in their inheritance, such as glaucoma, type II diabetes, lupus erythematosus and schizophrenia (Tong *et al.* 2004). Exploring GI networks in model organisms, through screening of low order (between two alleles) and high-order interactions (between more than two alleles), would then help understanding the genetic networks underlying phenotypical variations and multigenic diseases (Lehner 2011).

2.3 Mapping genetic interactomes in model organisms

2.3.1 In yeast

Quantitative studies of synthetic sick or lethal (SSL) interactions in the baker's yeast *Saccharomyces cerevisiae* represent most of the GIs screens done to date. The

existence of mutation libraries for both essential and non-essential genes is regarded as the main reason for the development of large-scale GI studies (Giaever *et al.* 2002). Non-essential gene mutant libraries contain strains where single gene coding sequences are substituted by a drug-resistance marker (Giaever *et al.* 2002), while essential genes mutant libraries consist in a collection of conditional alleles (Tong *et al.* 2001; Davierwala *et al.* 2005; Schuldiner *et al.* 2005; Costanzo *et al.* 2010). These libraries have been extensively used in an automated methodology called synthetic genetic array (SGA) (Tong *et al.* 2001; Tong *et al.* 2004). SGA screening consists in using single mutated yeasts as query against a whole deletion library for the construction of double mutants in a high-throughput fashion (Tong *et al.* 2001; Tong *et al.* 2004). The fitness defects of double mutants are then scored to uncover SSL interactions for non-essential genes (Tong *et al.* 2004; Sharifpoor *et al.* 2012) and essential genes (Tong *et al.* 2001; Davierwala *et al.* 2005; Schuldiner *et al.* 2005; Costanzo *et al.* 2010).

In parallel, the epistatic mini-array profile (E-MAP) – another variant of SGA – takes colony size measurements (based on imaging) as a basis for the detection of GIs (Schuldiner *et al.* 2005). GIs are then identified through measurement of a slower (SSL, negative GIs) or faster (alleviating, positive GIs) growth rate of the double mutants than what is expected from each single mutant growth rate. This allowed the identification of both positive and negative GIs while SGA was set originally to detect negative SSL GIs only. E-MAP was also used to map GIs in different yeast species such as *Schizosaccharomyces pombe* (Ryan *et al.* 2012).

Among the other high-throughput methods to discover GIs in yeast, diploid-based synthetic lethality analysis with microarrays (dSLAM), uses a library of barcoded mutants and barcode microarrays to measure the relative abundance of each barcoded double mutants in pooled populations to identify digenic SSL interactions (Pan *et al.* 2006; Lin *et al.* 2008). Optical density measurements (St Onge *et al.* 2007), biomass quantification analysis termed flux balance analysis (FBA) (Segre *et al.* 2005), quantitative phenotype (Drees *et al.* 2005) and gene expression data (Van Driessche *et*

al. 2005) have also been employed to map GIs in specific biological processes. However, these studies remain restricted in terms of genome coverage.

2.3.2 In *C. elegans*

Screening a large amount of GIs in the nematode requires the utilisation of RNA interference (RNAi) through soaking animals in a solution containing RNAi molecules or feeding them with *E. coli* strains expressing the RNAi (Maeda *et al.* 2001; Timmons, Court et Fire 2001). This approach induces a downregulation of the expression of targeted gene, instead of a deletion. This has to be taken into consideration when comparing the *C. elegans* and yeast genetic interactomes (Lehner 2007; Dixon *et al.* 2009). To identify a GI, either both genes are targeted using RNAi or a genetic mutant strain containing either a hypomorphic or a null allele can be submitted to RNAi targeting the other gene (Kamath *et al.* 2003; Lehner *et al.* 2006; Byrne *et al.* 2007). Both approaches have been used to map a quite limited area of the *C. elegans* genetic interactome (< 2,000 GIs) when compared to genetic studies in yeast (>200,000 GIs) (Lehner *et al.* 2006; Byrne *et al.* 2007; Tischler, Lehner et Fraser 2008; Costanzo *et al.* 2010).

2.3.3 In human

To identify GIs in human, apart from the RNAi treatment of specifically mutated cell lines (reviewed in Dixon et al., 2009), Lin *et al.* suggested an interesting method that uses radiation hybrid (RH) genotyping data sets (Lin *et al.* 2010). This approach, while

being fast and inexpensive, is different than standard RNAi screening in that RH panels are used in order to "simulate" a double mutations. The simulation is done with medium-selected cells that possess extra copies of two genes and "attractive" or "repulsive" interactions are then identified whether the promoting effect of the extra copies is death or survival of the cell line respectively. The results obtained using this approach could not be easily compared to negative and positive interactions observed through gene deletion and/or expression reduction. By joining several data sets of RH panels, a network of ~6.7 million potential GIs were extracted and covered ~3.4% of all human gene pairs (Lin *et al.* 2010).

2.4 *In silico* mapping of GIs

Only few organisms, mainly unicellular, are amenable to an experimental mapping of GIs through genome-wide screening. Mapping of genetic interactomes in higher organisms requires development of predictive tools that allow a significant reduction of the number of gene pairs to be tested experimentally.

During the last decade, numerous strategies have been used to infer GIs in unicellular and multicellular organisms (reviewed in Steen 2012). However, to date, only *S. cerevisiae* and *C. elegans* genetic networks have gained substantial information from large-scale machine learning studies. Numbers of tools were developed to predict PPIs, co-essentiality, genes with similar functions, genes functioning in the same molecular complex and GIs. The design of these tools highlighted the intimate link existing between different networks – GI networks being used to infer PPIs and co-functionality (Tong *et al.* 2004; Ye *et al.* 2005) and inversely PPI networks, phenotypic profiles and GO annotations being used to predict GIs as detailed below. These different predictors present also cross-specificities – GIs occurring to some extend between genes coding

for interacting or non-interacting proteins, between or within-pathways/molecular modules, between genes involved in the same biological process or being involved in different and compensatory processes as discussed below.

Intuitively, we expect that the GI world constitutes a patchwork of functional relationships with distinctive properties. Predictive tools capturing different properties will then be able to identify a portion of the GI interactome and will be complementary one to another. Ultimately, acquiring a good knowledge on the molecular particularities of subclasses of GIs will lead to the design of specific and accurate predictors. To make an informed choice on the different elements that could be employed to design these predictors, we will review here the different structural and functional particularities of GIs, and detail how they have been used or could be used to generate predictor for GIs.

2.4.1 Exploiting the protein-protein and genetic interaction network density and structure

A primary attribute of biological interaction networks, including GI networks, is a scale-free/power law distribution of connections, where most nodes are sparsely connected ("non-Hub" nodes) and few ones are highly connected ("Hub" nodes) (Watts et Strogatz 1998; Jeong *et al.* 2001; Wagner 2001; Tong *et al.* 2004). These networks appear also to exhibit a small-world organization – dense interacting modules are sparsely connected to other modules but with a short average path length (Watts et Strogatz 1998; Jeong *et al.* 2001; Wagner 2001).

There is a clear connection between PPI- and GI-Hubs since a protein with many interactions in the physical network (PPI-Hub) typically has also many interactions in the genetic network (GI-Hub) (Ozier, Amin et Ideker 2003; Kafri *et al.* 2008). Both kinds of Hubs tend to be essential or associated with severe fitness defects, and to

genetically interact with each other (Ozier, Amin et Ideker 2003; Davierwala *et al.* 2005; Lehner *et al.* 2006; Goh *et al.* 2007; Baryshnikova *et al.* 2010; Costanzo *et al.* 2010; Sharifpoor *et al.* 2012). Intuitively, we may see essential Hubs as a direct association with human diseases. However, it is important to notice that, while PPI-Hubs tend to be ubiquitously expressed, disease genes (such as inherited disease genes) tend to encode for PPI-non-Hubs and to be tissue specific (Goh *et al.* 2007; Vidal, Cusick et Barabasi 2011).

Comparative analysis of the yeast interactome networks also revealed that the "non-essential" SSL network is at least four times denser than the PPI network (Tong *et al.* 2004), while the "essential" SSL network is five times denser than the "non-essential" SSL (Tong *et al.* 2001; Tong *et al.* 2004; Davierwala *et al.* 2005). The higher density of essential when compared to non-essential GI networks, suggests that essential genes are highly connected Hubs within GI networks, and that essential pathways may be connected to number of compensatory pathways (Davierwala *et al.* 2005; Costanzo *et al.* 2010). Given that 18% of all yeast genes are essential (Giaever *et al.* 2002; Christie *et al.* 2004), this also suggests that most yeast GIs may involve at least one essential gene (Davierwala *et al.* 2005). The higher density of GI network, when compared to PPI network, may reflect the fact that in the case of two compensatory pathways, PPIs may occur between proteins of a linear pathway, while any member of each pathway may genetically interact with any component of its own pathway or of its compensatory pathway (Tong *et al.* 2004).

As shown for PPI networks, the interaction density is not homogenously distributed within GI networks that are composed of dense modules (Tong *et al.* 2004). These structures, as detailed above and in the following sections, are enriched in interactions occurring within functional modules (such as signaling pathways or protein complexes) or between functional modules. This property of dense GI modules could directly be used to predict novel GIs within a non-saturated network. Tong *et al.*, showed for three specific GI modules, that ~20% of genes that interact with a high number of common

partners – being part of the same dense GI module – also genetically interact one with the others (Tong *et al.* 2004). This was significantly higher than what was measured in random networks (approximately 1%) (Tong *et al.* 2004). Qi *et al.* extended this network analysis by including neighbors of interacting genes from any distances and by classifying those distances by the parity of the path lengths (Qi *et al.* 2008). They employed a graph diffusion kernel that uses weighted sums for different path lengths and found that odd-length kernels were better at predicting GIs while even-length kernels were more effective in finding new PPI partners (Qi *et al.* 2008).

Several methods have been developed to dissect complex networks into functionally meaningful modules. Using various clustering techniques, some studies reordered the GI matrix to sort genes according to the similarity of their GI profiles. Congruent genes are then defined as genes with similar GI profiles (Schuldiner *et al.* 2005; Ye *et al.* 2005; Collins *et al.* 2007; Costanzo *et al.* 2010; Costanzo *et al.* 2011). The resulting map has a modular structure that distinguishes between major biological processes, such as transcription and chromatin remodelling, DNA replication and repair or sister chromatid segregation. These genetic interaction profiles then provide a powerful way to identify sets of genes functioning in the same biological process (Tong *et al.* 2004; Schuldiner *et al.* 2005; Ye *et al.* 2005; Pan *et al.* 2006). Some of these methods have used the complex and pathway (COP) scores for finding sets of genes that are both highly correlated and that lack an aggravating GI (Schuldiner *et al.* 2005; Collins *et al.* 2006; Collins *et al.* 2007). The top-scoring gene pairs using this method included several sets of known complex or linear pathway components, as well as several predictions of novel ones (Schuldiner *et al.* 2005). Mutual clustering coefficient (MCC) was also employed to measure the neighbourhood sharing of connections in the GI network – called congruence score (Ye *et al.* 2005; Ye *et al.* 2005). A high score indicates that two genes share more GI partners than expected by chance. The resulting scores are then used as weight for non-directed edges linking genes within a congruence network (Ye *et al.* 2005). By comparing path lengths in three types of

networks (genetic interaction, genetic congruence, and protein interaction), they showed that high genetic congruence exhibits correlation with direct PPI linkage and also exhibits proportionate distance with the PPI network (Ye *et al.* 2005). This congruent score can then be used to predict PPIs.

Altogether, these studies showed that the structure of the GI network contains enough information to predict novel GIs and also to predict novel PPI, highlighting the intricate relationship existing between PPI and GI networks.

By further exploiting the relationship between PPI and GI networks, Paladugu *et al.* showed that PPI network graph-theory properties could also be used to predict GIs. They showed that proteins coded by SSL gene pairs, as compared to non-SSL ones, tend to have higher average degree, closeness centrality, information centrality and number of mutual neighbors within PPI network (Paladugu *et al.* 2008). When combined, these graph-theory properties of PPI network provided a powerful tool to predict SSL GIs (Paladugu *et al.* 2008). Moreover, this approach showed that the PPI network alone contains enough valuable information to predict SSL interactions. This approach appears particularly useful to predict GIs in higher organisms which are hardly amenable to systematic screening of GIs while having their PPIs at least partially mapped.

Few methods used GI and PPI networks to observe the distribution of GIs within or between dense modules of physical interactions (PPI and PDI), called in these studies "pathways" (Figure 2.3A and 3B) (Kelley et Ideker 2005; Ulitsky et Shamir 2007). Canonical "within and between pathway models" were originally identified by Kelley and Ideker (Kelley et Ideker 2005). They found that the "between pathway model", consisting of GIs occurring between dense modules of molecular interactions (Figure 2.3B), can explain three-and-a-half times as many GIs as the "within pathway" involving GIs within dense molecular interaction modules (Figure 2.3A) (Kelley et Ideker 2005). Further arguments for the prevalence of between-pathway GIs in yeast were given by Ye *et al.* and Pan *et al.*, who postulated that genes in the same pathway

are expected to share common GI partners (Ye *et al.* 2005; Pan *et al.* 2006). The between and within pathway models were however shown to explain only 40% of all yeast GIs (Kelley et Ideker 2005). Ulitsky and Shamir extended this interactome coverage by defining "pathways" as connected subnetworks within the physical interaction network rather than a dense interaction module (Figure 2.3C) (Ulitsky et Shamir 2007). This study provided a 2.7-fold increase from the number of interactions explained by the Kelley and Ideker models (Ulitsky et Shamir 2007).

Kelley and Ideker used their within and between pathway models to predict novel GIs (Kelley et Ideker 2005). A five-fold cross validation technique was used to investigate the accuracy of predicting GIs using both the "within pathway model" – genes within a given pathway genetically interact more frequently than expected by chance – or using the "between pathway model" – genes in one pathway genetically interact with many of the same partners in a second pathway. They showed that both models are efficient for predicting GIs while the "between-pathway" model appears to outperform the "within-pathway model" (Kelley et Ideker 2005).

Deeper studies on the "between and within pathways models" showed that they were often monochromatic, meaning that they were composed almost exclusively of a single type of GIs, either all negatives or all positives (Segre *et al.* 2005; Costanzo *et al.* 2010; Michaut *et al.* 2011). Monochromatic patterns have been used to identify biological processes and other functional modules (Segre *et al.* 2005; Pu *et al.* 2008; Jaimovich *et al.* 2010). Monochromatic processes are functionally diverse, but also biased (Michaut *et al.* 2011; Szappanos *et al.* 2011). For instance, microautophagy and histone exchange are monochromatic positives whereas protein import and small GTPase mediated signal transduction are monochromatic negatives (Michaut *et al.* 2011). Importantly, those studies showed that protein complexes are often monochromatic (Bandyopadhyay *et al.* 2008; Costanzo *et al.* 2010) and that monochromatic patterns, identified within and between biological processes, are mainly dependant on protein complexes (Michaut *et al.* 2011). The distinction between negative and positive

interactions, when considering the relationship between PPIs and GIs, has not yet been exploited to predict GIs to the best of our knowledge. The monochromaticity and the functional bias of this monochromaticity pattern have not been exploited neither.

In contrast to what was shown in yeast, the "within pathway model" tends to be more prevalent when compared to the "between pathway model" in the *C. elegans* interactome (Lehner *et al.* 2006; Lehner 2007). It was suggested that this difference might come from experimental screening methodologies employed to generate the GI interactomes in different organisms (Lehner 2007). While in yeast most of the mutations used to disrupt genes are null, in *C. elegans*, they are mainly hypomorphic. The highest number of "within pathway" interactions in *C. elegans* when compared to yeast may then be explained by the fact that hypomorphic alterations of genes functioning within the same protein complex or signaling pathway, may lead to a significant aggravation of the phenotype (synthetic interaction) while this would not be the case for null mutations (Lehner 2007). Also, we cannot exclude the possibility that this difference might come from the intrinsic difference existing between unicellular and multicellular organisms. "Within and between-pathway models" have not been used directly to predict novel GIs in the nematode.

While it is clear that signaling pathways are enriched in molecular interaction modules, it is important to notice the potential ambiguity created by the denomination of GIs occurring between dense molecular interaction modules as "between pathways" interactions. To the best of our knowledge, it has not been clearly proved that two densely connected molecular networks may not participate to the same signaling pathway – defined as a cascade of molecular events controlling a biological function. This possibility is supported by the fact that a high number of "pathways"/molecular interaction modules defined by Ideker and Kelly as well as Ulitsky and Shamir, are very small (Ma, Tarone et Li 2008). Consequently, we cannot exclude the possibility that some "between pathways/molecular modules" interactions may actually occur within signaling or metabolic pathways. This taken into consideration, the fact that

most GIs in yeast occurs between molecular modules and presumably pathways constitutes a golden avenue to identify compensatory pathways responsible for the cellular homeostasis and development of resistance to therapeutic agents (Tucker et Fields 2003; Szappanos *et al.* 2011). This hypothesis was validated experimentally using, for example, the Cdc14 early anaphase release (FEAR) and the mitotic exit network (MEN), two parallel pathways required for the release of the essential protein phosphatase Cdc14p from nucleolus during yeast cell cycle (Stegmeier, Visintin et Amon 2002).

Different approaches were used to study the modularity of GI networks. The decomposition of these networks using a biclustering technic recalled the idea of congruence. This technic was used to clusters groups of genes based on their GI profiles. However, in addition to clustering, biclustering helped the identification of two kinds of motif within the GI network: bicliques and q-cliques. This decomposition of the GI networks in absence of any integration of molecular networks gave also a bright new perspective to the within/between pathway models (Bellay *et al.* 2011). In this study, the between pathway model implies that GIs occurs in "bicliques" – defined as biclusters in which the query genes (first cluster of genes) and the array genes (set of genes interacting with the query genes) do not overlap (Figure 2.3E). Following the same reasoning, within pathway interactions occur in "cliques/quasi-cliques/q-cliques" – defined as biclusters in which query and array genes have significant overlap (Figure 2.3D) (Bellay *et al.* 2011). Interestingly, both positive and negative interactions were mainly found in bicliques (Bellay *et al.* 2011), similarly to what was shown using the canonical "between pathway" model (Costanzo *et al.* 2010). In addition, negative q-cliques – q-cliques composed of negative interactions – which corresponded to only 9% of negative biclusters (versus 91% of negative bicliques), did not appear to represent single protein complexes or pathways (Bellay *et al.* 2011). This constitutes a major difference with the canonical "within pathway" model defined by the overlap of genetic and molecular modules (Kelley et Ideker 2005). The genes found in negative

q-cliques were found to be expressed in a coordinated manner and to be enriched for chromosome segregation and cell cycle processes (Bellay *et al.* 2011). Bellay *et al.* suggested that this particular functional enrichment might arise due to general sensitivity to perturbation in fragile systems such as cell division.

Altogether, these studies support the idea that different techniques used to decompose GI networks help revealing different categories of GIs. They suggest that predictive tools developed based on any of these models (the canonical "within /between pathway" model or the "biclique/q-clique" model) may be complementary to models built on the other one. The functional bias observed for different GI modules also suggests that predictive tools may gain in performance if they specifically target GIs associated to a subset of biological functions alongside homogenous particularities with respect to GI network modularity.

Network decompositions using biclustering techniques also help to provide critical information on duplicated genes (Kelley et Ideker 2005). Duplicate genes were previously shown to display negative GIs with each other and exhibit fewer GIs than other genes because they tend to buffer one another functionally (VanderSluis *et al.* 2010). They were also shown to exhibit numerous unique GIs, suggesting that duplicated genes are functionally redundant but have divergent roles (Ihmels *et al.* 2007; VanderSluis *et al.* 2010). While, we would expect duplicated genes to be part of the isolated group of GIs within the biclustering array, a significant amount of them were fund to exhibit negative GIs with each other as part of larger modular structures (biclusters) (Bellay *et al.* 2011). Interestingly, this subgroup of duplicates was significantly more divergent in terms of sequence identity. It was suggested by Bellay *et al.*, that duplicates with a high degree of functional similarity specifically compensate for the loss of one another (isolated GIs in biclustering array), while in the second case, they appeared to have diverged into entirely different functional modules with compensatory properties (GIs being part of large biclusters) (Bellay *et al.* 2011).

This study opens the door to predictive avenues that consider using protein sequence homology to identify compensatory genes and modules.

2.4.2 Exploiting relationships between networks at different abstraction levels

Networks at different abstraction levels were used to infer GIs in yeast and *C. elegans* as detailed in Table 2.1 and below. These studies also brought a deeper understanding of the molecular basis of GIs (Avery et Wasserman 1992; Guarente 1993; Thomas 1993).

GIs in yeast, *C. elegans* and in human, were significantly more abundant between genes sharing mutant phenotypes (abstraction level VI) or Gene Ontology (GO) annotations (level V), and between genes encoding proteins in the same subcellular localization (level V) and/or within the same protein complex (level III) or pathway (level IV) (Lee *et al.* 2004; Tong *et al.* 2004; Kelley et Ideker 2005; Lee *et al.* 2008; Lin *et al.* 2010). In agreement with the general idea that synthetic GIs may occur between genes with redundant functions, the SSL yeast network was also found to be enriched in gene pairs encoding homologous proteins (level I).

A link between two genes or their protein products within networks located at different levels of abstraction is then informative of a potential GI. An important class of predictive methodologies used these diverse sources of data to discriminate interacting from non-interacting genes. The first of these studies used decision tree learning to integrate various types of data along with a "2hop" network topology assessment for various genomic relationships (Table 2.1) (Wong *et al.* 2004). The "2hop" method considers gene pairs linked to a common partner by a functional relationship (e.g. physical interaction and sequence homology) to be informative of a potential SSL

interaction between them in yeast. In total, 123 functional relationships (26 "major" categories) were used (Wong *et al.* 2004). The most powerfull predictive informations were selected using a Bayesian Information Criterion (BIC; similar to the Akaike Information Criterion (AIC)).

For multicellular organisms, Zhong and Sternberg integrated multiple types of data from yeast, fly and nematodes to predict 18,183 genetic interactions in the nematode *Caenorhabditis elegans* (Table 2.1) (Zhong et Sternberg 2006). Here, a logistic regression was used to integrate features (or "attributes") defined as the relative weight of a single type of data according to its predictive power. The positive set of elements used to train the model consisted in 1,816 validated GIs and 2,878 PPIs while negative examples were made of 3,296 paired *cis* markers. These markers are used in genetic mapping experiments and are assumed to have less probability of interacting together than pairs of genes randomly picked from the genome. The utilization of yeast/fly data to obtain greater genome coverage for a multitude of data sources appears to positively contribute to the predictive power of the developed tool (Zhong et Sternberg 2006). We will discuss the limitation brought by data from other organisms in the following section considering evolutionary conservation of PPI and GI networks. In this study, the predictive interaction network was submitted to experimental validation using as bait *let-60*/Ras and *itr-1*/ITPR (two human disease-related genes) with a high success rate – 44% and 60% of true positive predictions respectively (Zhong et Sternberg 2006). Although it is still unpublished, a new version of Zhong and Sternberg predictor, called "GeneOrienteer", is available online (geneorienteer.org). This model employs a naïve Bayes classifier and integrates more than 20 features to predict GIs in several species.

Another approach, developed by Chipman and Singh, used a random walks algorithm to calculates the topological similarity of two genes in many types of biological networks, including genetic and physical interactions, co-expression and GO annotation networks, for both *S. cerevisiae* and *C. elegans* (Table 2.1) (Chipman et

Singh 2009). This topological similarity is then used to predict negative GIs. In this study, the decision tree classifier was shown to outperform the logistic regression classifier (Chipman et Singh 2009). The good performances of this approach, tested using cross-validation technics, was unfortunately not supported by any experimental validations.

Other studies using the likelihood scoring of gene pairs for the prediction of GIs in the nematode *C. elegans* were generated soon after (Table 2.1) (Lee, I. *et al.* 2010; Lee, A.Y. *et al.* 2010). The first approach, called "WormNet", is used to infer the shared function of two genes, which is also indicative of a possible GI (Lee, I. *et al.* 2010). This model was trained on thousands of gene pairs sharing GO annotations. A second version of this model, called "WormNet2", employs a weighted sum instead of a naïve Bayes classifier and integrates many "updated" features derived from log likelihood scores of various functional data (Lee, I. *et al.* 2010). Contrarily to Zhong and Sternberg methodology where functional data are more intuitive (e.g. co-expression of genes), WormNet2 included some "less-common" types of data (e.g. co-citation of gene names) as features to infer shared functions (Lee, I. *et al.* 2010). Although they did not use any feature selection methodology (e.g. BIC or AIC), several examples of resulting predicted interactions by the weighted sum model showed that most features contributed to the final scores. They also succeeded in validating several GI for three signaling genes via RNAi screening but the validation success rate for individual genes appears to be low ranged from only 4% to a maximum of 15% achieved for the gene *vab*-1 (Lee, I. *et al.* 2010).

Considering the environment of genes/proteins in networks at different level of abstractions, we built an additional model: "GIFinder" (Table 2.1) (Lee, A.Y. *et al.* 2010). This tool used logistic regression and six features to predict GIs with a positive training set composed uniquely of validated GIs. This model also used novel attributes that consider the enrichment of phenotypic features in the co-expression/physical network environment of a gene. This kind of attribute integrates data from three

abstraction levels (level II, III and VI) to assess whether two genes may be part of the same functional module instead of relying only on evidences of direct interactions. These attributes also reduced the negative effect of using biological datasets with poor genome coverage and were shown to highly contribute to the overall performance of the predictor (Lee, A.Y. *et al.* 2010). This approach would be appropriate when trying to integrate sparse data such as tissue expression profiles and subcellular localization, to other datasets with high genome coverage such as expression data. Experimental validations of predicted GIs for *gdi-1*/GDI1 – a Rho GTPase regulator associated with non-syndromic forms of mental retardation in human – supported the idea that such methodology could be useful to identify therapeutic targets for monogenic diseases from predictive GI networks of lower organisms (Lee, *et al.* 2010). With a success rate of at least 42%, the performance in experimental validations was comparable to that of similar approaches.

Recently, Hoehndorf *et al.* created a predictor of GIs for 4 different species by inferring the function of many genes using semantic similarity measurements of phenotypes and gene ontology annotations (Hoehndorf *et al.* 2013). The semantic similarity – a measure of the distance or relatedness between two terms – was done using the Jaccard index. Unfortunately, the GIs obtained from their inferred gene functions were not validated experimentally. This later methodologies exploit only biological information located at the highest level of abstraction (level V and VI). We expect that this methodology – ignoring co-expression and molecular interaction levels – would then be able to predict GIs occurring between genes controlling a given biological process from distant environments (cell non-autonomous interactions). However, this possibility has not been investigated by the authors (Hoehndorf *et al.* 2013).

When trying to compare the relative performances of predictive tools, it is important to note, that while experimental validation of predictions highly contribute to the demonstration of the validity of the method, the heterogeneity of link density within the GI network and the experimental methods used to validate the interactions may

highly influence the success rate of the validation. Therefore, it is extremely difficult to compare the relative performance of individual methods just by comparing the success rate of validation experiments, using one or two genes as bait, and different validation methods (mutant and RNAi, mutant and double mutant, or RNAi and double RNAi)

To assess how different integration designs impact the prediction of GIs for a given organism, we compared the predictions obtained for GeneOrienteerv2.12, GIFinder and WormNet2. Interestingly, these predictors appear to be highly complementary with more than 90% of predicted interactions by the three models being unique – i.e. predicted by only one approach (Figure 2.4A). This suggests that these three predictors capture different areas of the GI interactome covered by sets of experimentally identified GIs leaving more than 57% of it untouched (Figure 2.4B). GeneOrienteerv2.12 performed extremely well when tested on a set of 1,514 GIs obtained from interaction databases. This set of GIs, being used as a predictive feature or training sets in GIFinder and GeneOrienteerv2.12 (see "geneorienteer.org"; (Lee, A.Y. *et al.* 2010)), we tested the three models on a set of recently published interactions (curated manually and absent from the databases) and observed a significant reduction in the performance of GeneOrienteer when compared to the two other models (Figure 2.4C). The deprived overlaps of predictions generated using the three predictors could be explained by the different integration methodologies used to generate the predictors (naïve Bayes classifier vs. linear regression) or by the different training sets used. The major difference of GIFinder when compared to other tools comes from the utilization of validated GIs as the only positive training examples as opposed to the two other ones that also employed physical interactions or functional annotations (Zhong et Sternberg 2006; Lee, I. *et al.* 2010; Lee, A.Y. *et al.* 2010). While PPI and GI networks may have some overlap (some interactions occurring within protein modules), training a model using PPIs as a positive training set may bias the model towards within protein module GIs. Similar reasoning would be also valid for functional annotations. While functional

annotations, such as GO annotations, are enriched between interacting genes, a large number of GIs are expected to occur between genes with different functions as discussed earlier. Interestingly, and as discussed in the following section, within protein module and within biological process GI appear to be more conserved that between modules or process GIs. We may then postulate that the bias induced through training the models using PPIs and GO annotations may increase the rate of evolutionary conserved interactions in the predictions. This taken into consideration, the fact that the training sets, constituted by the union of GIs and PPIs and/or pairs of genes with similar functions, is larger than validated sets of GIs only may improve the performance of predictive models using machine-learning techniques (Babyak 2004).

While the existence of an edge between two genes/proteins in a network at a given level of abstraction is now confirmed as a useful information to infer a missing edge between these two genes/proteins at another level, it is important to realize that the conservation of links between two genes/proteins in different networks is a relatively rare event. For example, approximately 1% of SSL pairs (0.4% of negative and 0.5% of positive genetic interactions in E-MAP datasets) codes for physically linked proteins (conservation of links between networks at levels III and IV) and 1% for homologous proteins (conservation of links between networks at levels I and IV) (Tong *et al.* 2004; Costanzo *et al.* 2010). Cumulating these evidences of direct links between genes and proteins may increase the sensitivity of predictive tools for GIs. Considering only these direct links may also contribute to their relative poor performances. These tools would then gain in performance if integrating attributes that consider the environment of the genes in networks and the network modularity as shown by GIFinder (Lee, A.Y. *et al.* 2010).

2.4.3 Considering evolution of protein-protein and genetic interaction networks

Several tools used data from evolutionary distant species to predict GIs. The evolutionary conservation of these data along with the structure of interaction networks between species is then of a critical interest when considering using this information to design a powerful predictive tool. In addition, while GI interactomes are extensively mapped in certain organisms such as yeast, the utilisation of these networks to predict GIs in higher organisms mainly depends on the evolutionary conservation of GIs and of the GI network structure.

GIs are known to play a critical role in evolutionary processes (Yukilevich *et al.* 2008; Stern et Orgogozo 2009). In opposition to what was initially thought, all genes are not equal in the eyes of evolution, and evolutionarily relevant mutations tend to accumulate in hotspot genes at specific positions of these genes (Stern et Orgogozo 2009). A mutation in a gene, having a high number of GI partners, would not be advantageous in a context of adaptive evolution since it will increase the phenotypic variance associated with this mutation and therefore, will cause an increased fitness fluctuation dependent on the genetic background (Stern et Orgogozo 2009). In addition, mutations generating specific phenotypic changes are more likely to contribute to adaptive evolution than pleiotropic mutations altering several seemingly unrelated traits (Stern et Orgogozo 2009). Genetic Hubs, being by definition connected to a large number of genes and highly enriched for pleiotropic and multifunctional genes (Costanzo *et al.* 2010), would then be less touched by mutations associated with adaptive evolution. As expected, GI-Hubs are highly evolutionary conserved (Bellay *et al.* 2011) as are PPI-Hubs (Wuchty, Barabasi et Ferdig 2006).

When considering PPIs, interactions within modules are conserved at a higher level than interactions occurring outside modules (Zinman, Zhong et Bar-Joseph 2011). This suggests that there might be a much higher selective pressure to maintain interactions

within a single module than between modules (Zinman, Zhong et Bar-Joseph 2011). PPI networks from distant species were used in number of studies to predict GIs (Table 2.1) (Zhong et Sternberg 2006; Chipman et Singh 2009; Lee, I. *et al.* 2010; Lee, A.Y. *et al.* 2010; Hoehndorf *et al.* 2013). These studies, however, did not discriminate dense modules of PPIs from non-modular interactions. Since within complex/modules PPIs were shown to be more conserved than extra-modular PPIs, it would be interesting to assess whether the utilisation of modular components of PPI interactomes from distant species, instead of the complete interactome, would improve performances of predictive tools for GIs.

While the evolutionary conservation of PPI- and GI-Hubs, as well as PPIs within protein complexes/modules has been clearly established, the overall conservation of GIs between evolutionary distant species is still controversial. Comparison of the *S. cerevisiae* and *S. pombe* E-MAPs showed that negative and positive GIs of two yeast species, distant of approximately 400 million years, were significantly conserved (Sipiczki 2000). Also, essentiality in genes appears to be highly conserved between the yeast and nematode (Kamath *et al.* 2003), the extent of the GI conservation between these organisms appears to be very low, and not reported as significant in all studies (Pan *et al.* 2004; Lehner 2007; St Onge *et al.* 2007; Mani *et al.* 2008; Tischler, Lehner et Fraser 2008). The difference in methodologies used to generate the GI networks between yeast and nematodes, the fact that some GIs in nematodes may not be cell autonomous because of its multi-cellularity and the poor genome coverage of *C. elegans* vs. yeast genetic interactomes may be part of the reasons behind the poor conservation of GIs observed between these organisms.

Since we expect the majority of GIs not to be conserved across species, GI-Hubs, on the other hand, appear to be well conserved throughout evolution (Lehner *et al.* 2006; Costanzo *et al.* 2010). Predicting genetic Hubs are of biological importance because of their tendency to influence fitness defects when they are individually mutated (Costanzo *et al.* 2010). Some high-end methodologies have been developed to predict

GI degrees – the number of GIs involving a given gene – in the yeast, *S. cerevisiae* (Szappanos *et al.* 2011; Koch *et al.* 2012). The first study successfully predicts negative and positive interaction degrees for genes implicated in yeast metabolism (Szappanos *et al.* 2011). Using only SSL and positive genetic interactions as training sets, they showed that only a small fraction of interacting genes shares the majority of the interactions in both empirical and *in silico* data. They also provided a mechanistic explanation for genetic "Hubs" in relation with their tendency to be multifunctional and found that the predicted negative interaction degree of a gene correlates with its multifunctionality (Szappanos *et al.* 2011). In another work, Koch et al. drove the analysis furthermore to predict the GI degrees of many genes in *S. cerevisiae* and also in the distantly related species *Schizosaccharomyces pombe* (Koch *et al.* 2012). They integrated 16 features; covering mRNA expressions, GO terms, protein-protein interactions and other functional data, via a decision-tree learning to predict GI degrees with only interacting genes as training sets. Among some interesting findings, they confirmed the general consensus that the GI network structure is conserved across species (Koch *et al.* 2012). In fact, they found retaining high conservation of GI degrees between *S. cerevisiae* and *S. pombe* for specific genes sharing a significant amount of functional information. It would be extremely interesting to carry on such study to assess whether, despite the poor conservation of GIs between yeast and nematodes, the GI network structures may also be conserved between the two organisms.

As the conservation of GI-degrees, conservation of GIs between *S. cerevisia* and *S. pombe* was significantly increased when the analysis was restricted to genes that shared the same functional annotations and when the analysis was restricted to pairs of genes coding for interacting proteins (Roguev *et al.* 2008). This indicates that GIs between two genes is more evolutionary conserved if these two genes are also linked in networks located at lower and higher abstraction levels. Several studies also suggested that both positive and negative GIs within functional modules (protein complexes, gene belonging to the same biological process) are significantly more conserved between *S.*

cerevisiae and *S. Pombe*, than wiring between these modules (Dixon *et al.* 2008; Roguev *et al.* 2008; Ryan *et al.* 2012). This suggests that not only the dependencies, but also the buffering relationships within complexes are highly conserved (Ryan *et al.* 2012).

While the conservation of GIs between functional modules/biological processes appears to be limited, the overall number of GIs between biological processes appears to be highly retained (Ryan *et al.* 2012). For example, while a significantly high number of GIs links genes controlling chromatin/transcription and those controlling mitosis and chromosome segregation in distant species, the level of conservation for individual interactions between these processes remains low (Ryan *et al.* 2012). This suggests that, although there is flexibility at the level of individual GIs and consequently significant rewiring between functional modules/processes in distant species, there may exist a biological selective pressure and requirements for the conservation of a high of low linking strength between particular processes (Ryan *et al.* 2012). Importantly, biological processes interacting with a larger amount of biological processes than expected – called here "process-Hubs" – suggest that these processes are important for mediating cross-process connections in genetic networks of several organisms (Lehner *et al.* 2006; Costanzo *et al.* 2010). For example, process-hubs such as chromatin/transcription, secretion and membrane trafficking, have been identified in *S. cerevisiae* (Costanzo *et al.* 2010) and *C. elegans* (Lehner *et al.* 2006). Conversely, some processes, such as amino acid metabolism and trans-membrane transport, have very few GIs linking them to other processes, suggesting a high degree of functional independency among these modules with less impact on other cellular processes than process-Hubs (Ryan *et al.* 2012).

Altogether, these data suggest that the conservation of the overall structure of GI networks still needs further characterization in distant organisms to identify the selective pressure applied on GI networks, not necessarily at the level of individual genes, but at the level of functional modules. Conclusions from such studies would

bring important information that could be exploited in order to use GI networks from lower organisms to predict GIs in higher ones.

2.5 Conclusion and perspectives

Mapping of GI networks and extensive study of their structures, conservation in different species and relationships with other functional and molecular interaction networks has already provided us with a better understanding of the biological robustness and phenotypical manifestation of genomic codes. Some of these pieces of information have also been exploited to generate predictors for GIs as detailed in this review. However, to date, these tools show limited performances and gave predictions, for example in *C. elegans*, for less than 50% of the expected GI interactome. These studies also opened some paths that could be followed to improve predictive tools for GIs.

The first path suggests that tools should consider GIs in their structural context instead of considering them in isolation. This comes from several observations. The first one showed that similarity of GI profiles of two genes is more indicative of a co-functionality (sharing GO annotations, involvement in the same protein complex, etc.) than a direct GI between these genes. This comes along with the other observations that – irrespective of the method used to decompose the genetic interactome into modules – GI tends to segregate into two categories following either a "within-"or a "between-pathway" model (Figure 2.5). These two kinds of GIs, based on structural properties of the network, have also different particularities. The "between-pathway" GIs tend to be less evolutionary conserved than the "within-pathway" GIs. Similarly, at a lower level of abstraction, "between protein modules" PPIs tend to be less conserved than "within protein modules" PPIs. Overall, these data suggest that "within and between pathways"

GIs may have to be assessed using different approaches. This also suggests that data used to predict GIs, such as PPIs, may also have to be considered in their modular context.

The second path of improvement for predicting GIs consists in considering GIs from a higher level of abstraction when attempting to predict GIs using data from distant species. This comes from the observation that the overall level of GIs between biological processes appears to be much more conserved between distant species than independent GIs between genes involved in different processes (Figure 2.5). Considering GIs at the level of the biological processes (abstraction level V) instead of individual genes (abstraction level IV), may then significantly improve our ability to accurately predict functional relationships between genes and group of genes. Such approach may also open exciting opportunities. Studying the monochromaticity of GI modules also showed that the monochromatic within and between pathways interactions were biologically biased. This suggests that biological processes have either compensating or synergistic relationships one with another, but also that many components of a given biological process have predominantly either compensating or synergistic relationships. These data suggest that considering GIs from a higher level of abstraction may also be a good avenue to specifically identify synergistic and compensating/antagonistic relationships between functional biological modules. This avenue is particularly attractive when considering the need of such predictive tools in translational research and more particularly when trying to identify compensatory mechanisms leading to therapeutic drug resistance.

The last proposed path to improve GI predictions, in particular in higher organisms, is to try to better understand the structural differences that may exist between lower/unicellular and higher organisms. The fact that the within pathway model may be prevalent over the between pathway model in *C. elegans*, as opposed to yeast, need to be confirmed and the reason why this trend might be different in several organisms needs to be explained. In conclusion, while an extensive characterization of smaller

genetic networks in yeast has brought precious information about the still mysterious genetic interactome, its apparent plasticity requires similar studies to be done in higher organisms. These studies would then open the door to the design of well-informed and highly performing predictors for GIs in higher organisms such as human.

2.6 References

Avery, L. and S. Wasserman (1992). Ordering gene function: the interpretation of epistasis in regulatory hierarchies. *Trends Genet* 8(9): 312-316.

Babyak, M. A. (2004). What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosom Med* 66(3): 411-421.

Bandyopadhyay, S., R. Kelley, et al. (2008). Functional maps of protein complexes from quantitative genetic interaction data. *PLoS Comput Biol* 4(4): e1000065.

Baryshnikova, A., M. Costanzo, et al. (2010). Quantitative analysis of fitness and genetic interactions in yeast on a genome scale. *Nat Methods* 7(12): 1017-1024.

Bateson, W. and G. Mendel (1909). *Mendel's principles of heredity, by W. Bateson.* Cambridge [Eng.], University Press.

Bellay, J., G. Atluri, et al. (2011). Putting genetic interactions in context through a global modular decomposition. *Genome Res* 21(8): 1375-1387.

Bellay, J., S. Han, et al. (2011). Bringing order to protein disorder through comparative genomics and genetic interactions. *Genome Biol* 12(2): R14.

Byrne, A. B., M. T. Weirauch, et al. (2007). A global analysis of genetic interactions in Caenorhabditis elegans. *J Biol* 6(3): 8.

Chipman, K. C. and A. K. Singh (2009). Predicting genetic interactions with random walks on biological networks. *BMC Bioinformatics* 10: 17.

Christie, K. R., S. Weng, et al. (2004). Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms. *Nucleic Acids Res* 32(Database issue): D311-314.

Collins, S. R., K. M. Miller, et al. (2007). Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature* 446(7137): 806-810.

Collins, S. R., M. Schuldiner, et al. (2006). A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. *Genome Biol* 7(7): R63.

Costanzo, M., A. Baryshnikova, et al. (2010). The genetic landscape of a cell. *Science* 327(5964): 425-431.

Costanzo, M., A. Baryshnikova, et al. (2011). Charting the genetic interaction map of a cell. *Curr Opin Biotechnol* 22(1): 66-74.

Davierwala, A. P., J. Haynes, et al. (2005). The synthetic genetic interaction spectrum of essential genes. *Nat Genet* 37(10): 1147-1152.

Dixon, S. J., M. Costanzo, et al. (2009). Systematic mapping of genetic interaction networks. *Annu Rev Genet* 43: 601-625.

Dixon, S. J., Y. Fedyshyn, et al. (2008). Significant conservation of synthetic lethal genetic interaction networks between distantly related eukaryotes. *Proc Natl Acad Sci U S A* 105(43): 16653-16658.

Dragani, T. A. (2003). 10 years of mouse cancer modifier loci: human relevance. *Cancer Res* 63(12): 3011-3018.

Drees, B. L., V. Thorsson, et al. (2005). Derivation of genetic interaction networks from quantitative phenotype data. *Genome Biol* 6(4): R38.

Edelman, L. B., J. A. Eddy, et al. (2010). In silico models of cancer. *Wiley Interdiscip Rev Syst Biol Med* 2(4): 438-459.

Elena, S. F. and R. E. Lenski (1997). Test of synergistic interactions among deleterious mutations in bacteria. *Nature* 390(6658): 395-398.

Fisher, R. A. (1919). XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh* 52(02): 399-433.

Ge, H., Z. Liu, et al. (2001). Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae. *Nat Genet* 29(4): 482-486.

Giaever, G., A. M. Chu, et al. (2002). Functional profiling of the Saccharomyces cerevisiae genome. *Nature* 418(6896): 387-391.

Gibson, G. (2010). Hints of hidden heritability in GWAS. *Nat Genet* 42(7): 558-560.

Goh, K. I., M. E. Cusick, et al. (2007). The human disease network. *Proc Natl Acad Sci U S A* 104(21): 8685-8690.

Grigoriev, A. (2001). A relationship between gene expression and protein interactions on the proteome scale: analysis of the bacteriophage T7 and the yeast Saccharomyces cerevisiae. *Nucleic Acids Res* 29(17): 3513-3519.

Guarente, L. (1993). Synthetic enhancement in gene interaction: a genetic tool come of age. *Trends Genet* 9(10): 362-366.

Han, J. D., N. Bertin, et al. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* 430(6995): 88-93.

Hoehndorf, R., N. W. Hardy, et al. (2013). Systematic analysis of experimental phenotype data reveals gene functions. *PLoS ONE* 8(4): e60847.

Ihmels, J., S. R. Collins, et al. (2007). Backup without redundancy: genetic interactions reveal the cost of duplicate gene loss. *Mol Syst Biol* 3: 86.

Jaimovich, A., R. Rinott, et al. (2010). Modularity and directionality in genetic interaction maps. *Bioinformatics* 26(12): i228-236.

Jansen, R., D. Greenbaum, et al. (2002). Relating whole-genome expression data with protein-protein interactions. *Genome Res* 12(1): 37-46.

Jasnos, L. and R. Korona (2007). Epistatic buffering of fitness loss in yeast double deletion strains. *Nat Genet* 39(4): 550-554.

Jeong, H., S. P. Mason, et al. (2001). Lethality and centrality in protein networks. *Nature* 411(6833): 41-42.

Joshi-Tope, G., M. Gillespie, et al. (2005). Reactome: a knowledgebase of biological pathways. *Nucleic Acids Res* 33(Database issue): D428-432.

Kafri, R., O. Dahan, et al. (2008). Preferential protection of protein interaction network hubs in yeast: evolved functionality of genetic redundancy. *Proc Natl Acad Sci U S A* 105(4): 1243-1248.

Kamath, R. S., A. G. Fraser, et al. (2003). Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature* 421(6920): 231-237.

Kanehisa, M. and S. Goto (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 28(1): 27-30.

Kelley, R. and T. Ideker (2005). Systematic interpretation of genetic interactions using protein networks. *Nat Biotechnol* 23(5): 561-566.

Kemmeren, P., N. L. van Berkum, et al. (2002). Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Mol Cell* 9(5): 1133-1143.

Koch, E. N., M. Costanzo, et al. (2012). Conserved rules govern genetic interaction degree across species. *Genome Biol* 13(7): R57.

Lee, A. Y., R. Perreault, et al. (2010). Searching for signaling balance through the identification of genetic interactors of the Rab guanine-nucleotide dissociation inhibitor gdi-1. *PLoS ONE* 5(5).

Lee, I., S. V. Date, et al. (2004). A probabilistic functional network of yeast genes. *Science* 306(5701): 1555-1558.

Lee, I., B. Lehner, et al. (2008). A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans. *Nat Genet* 40(2): 181-188.

Lee, I., B. Lehner, et al. (2010). Predicting genetic modifier loci using functional gene networks. *Genome Res* 20(8): 1143-1153.

Lee, T. I., N. J. Rinaldi, et al. (2002). Transcriptional Regulatory Networks in Saccharomyces cerevisiae. *Science* 298(5594): 799-804.

Lehner, B. (2007). Modelling genotype-phenotype relationships and human disease with genetic interaction networks. *J Exp Biol* 210(Pt 9): 1559-1566.

Lehner, B. (2011). Molecular mechanisms of epistasis within and between genes. *Trends Genet* 27(8): 323-331.

Lehner, B., C. Crombie, et al. (2006). Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways. *Nat Genet* 38(8): 896-903.

Lin, A., R. T. Wang, et al. (2010). A genome-wide map of human genetic interactions inferred from radiation hybrid genotypes. *Genome Res* 20(8): 1122-1132.

Lin, Y. Y., Y. Qi, et al. (2008). A comprehensive synthetic genetic interaction network governing yeast histone acetylation and deacetylation. *Genes Dev* 22(15): 2062-2074.

Luscombe, N. M., M. M. Babu, et al. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* 431(7006): 308-312.

Ma, X., A. M. Tarone, et al. (2008). Mapping genetically compensatory pathways from synthetic lethal interactions in yeast. *PLoS ONE* 3(4): e1922.

Maeda, I., Y. Kohara, et al. (2001). Large-scale analysis of gene function in Caenorhabditis elegans by high-throughput RNAi. *Curr Biol* 11(3): 171-176.

Mani, R., R. P. St Onge, et al. (2008). Defining genetic interaction. *Proc Natl Acad Sci U S A* 105(9): 3461-3466.

Michaut, M., A. Baryshnikova, et al. (2011). Protein complexes are central in the yeast genetic landscape. *PLoS Comput Biol* 7(2): e1001092.

Miko, I. (2008). Phenotype Variability: Penetrance and Expressivity. *Nature Education* 1(1).

Moore, J. H. (2003). The ubiquitous nature of epistasis in determining susceptibility to common human diseases. *Hum Hered* 56(1-3): 73-82.

Mrowka, R., A. Patzak, et al. (2001). Is there a bias in proteome research? *Genome Res* 11(12): 1971-1973.

Otto, S. P. and T. Lenormand (2002). Resolving the paradox of sex and recombination. *Nat Rev Genet* 3(4): 252-261.

Ozier, O., N. Amin, et al. (2003). Global architecture of genetic interactions on the protein network. *Nat Biotechnol* 21(5): 490-491.

Paladugu, S. R., S. Zhao, et al. (2008). Mining protein networks for synthetic genetic interactions. *BMC Bioinformatics* 9: 426.

Pan, X., P. Ye, et al. (2006). A DNA integrity network in the yeast Saccharomyces cerevisiae. *Cell* 124(5): 1069-1081.

Pan, X., D. S. Yuan, et al. (2004). A robust toolkit for functional profiling of the yeast genome. *Mol Cell* 16(3): 487-496.

Pu, S., K. Ronen, et al. (2008). Local coherence in genetic interaction patterns reveals prevalent functional versatility. *Bioinformatics* 24(20): 2376-2383.

Puniyani, A., U. Liberman, et al. (2004). On the meaning of non-epistatic selection. *Theor Popul Biol* 66(4): 317-321.

Qi, Y., Y. Suhail, et al. (2008). Finding friends and enemies in an enemies-only network: a graph diffusion kernel for predicting novel genetic interactions and co-complex membership from yeast genetic interactions. *Genome Res* 18(12): 1991-2004.

Roguev, A., S. Bandyopadhyay, et al. (2008). Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. *Science* 322(5900): 405-410.

Ryan, C. J., A. Roguev, et al. (2012). Hierarchical modularity and the evolution of genetic interactomes across species. *Mol Cell* 46(5): 691-704.

Sanjuan, R. and S. F. Elena (2006). Epistasis correlates to genomic complexity. *Proc Natl Acad Sci U S A* 103(39): 14402-14405.

Schuldiner, M., S. R. Collins, et al. (2005). Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell* 123(3): 507-519.

Segal, E., M. Shapira, et al. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* 34(2): 166-176.

Segre, D., A. Deluna, et al. (2005). Modular epistasis in yeast metabolism. *Nat Genet* 37(1): 77-83.

Sharifpoor, S., D. van Dyk, et al. (2012). Functional wiring of the yeast kinome revealed by global analysis of genetic network motifs. *Genome Res* 22(4): 791-801.

Sipiczki, M. (2000). Where does fission yeast sit on the tree of life? *Genome Biol* 1(2): REVIEWS1011.

St Onge, R. P., R. Mani, et al. (2007). Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions. *Nat Genet* 39(2): 199-206.

Stegmeier, F., R. Visintin, et al. (2002). Separase, polo kinase, the kinetochore protein Slk19, and Spo12 function in a network that controls Cdc14 localization during early anaphase. *Cell* 108(2): 207-220.

Stern, D. L. and V. Orgogozo (2009). Is genetic evolution predictable? *Science* 323(5915): 746-751.

Szafraniec, K., D. M. Wloch, et al. (2003). Small fitness effects and weak genetic interactions between deleterious mutations in heterozygous loci of the yeast Saccharomyces cerevisiae. *Genet Res* 82(1): 19-31.

Szappanos, B., K. Kovacs, et al. (2011). An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nat Genet* 43(7): 656-662.

Thomas, J. H. (1993). Thinking about genetic redundancy. *Trends Genet* 9(11): 395-399.

Timmons, L., D. L. Court, et al. (2001). Ingestion of bacterially expressed dsRNAs can produce specific and potent genetic interference in Caenorhabditis elegans. *Gene* 263(1-2): 103-112.

Tischler, J., B. Lehner, et al. (2008). Evolutionary plasticity of genetic interaction networks. *Nat Genet* 40(4): 390-391.

Tong, A. H., M. Evangelista, et al. (2001). Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* 294(5550): 2364-2368.

Tong, A. H., G. Lesage, et al. (2004). Global mapping of the yeast genetic interaction network. *Science* 303(5659): 808-813.

Tucker, C. L. and S. Fields (2003). Lethal combinations. *Nat Genet* 35(3): 204-205.

Ulitsky, I. and R. Shamir (2007). Pathway redundancy and protein essentiality revealed in the Saccharomyces cerevisiae interaction networks. *Mol Syst Biol* 3: 104.

Van Driessche, N., J. Demsar, et al. (2005). Epistasis analysis with global transcriptional phenotypes. *Nat Genet* 37(5): 471-477.

VanderSluis, B., J. Bellay, et al. (2010). Genetic interactions reveal the evolutionary trajectories of duplicate genes. *Mol Syst Biol* 6: 429.

Vidal, M., M. E. Cusick, et al. (2011). Interactome networks and human disease. *Cell* 144(6): 986-998.

Wagner, A. (2001). The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol Biol Evol* 18(7): 1283-1292.

Watts, D. J. and S. H. Strogatz (1998). Collective dynamics of 'small-world' networks. *Nature* 393(6684): 440-442.

Wolf, J. B., E. D. Brodie, et al. (2000). *Epistasis and the Evolutionary Process*, Oxford University Press.

Wong, S. L., L. V. Zhang, et al. (2004). Combining biological networks to predict genetic interactions. *Proc Natl Acad Sci U S A* 101(44): 15682-15687.

Wuchty, S., A. L. Barabasi, et al. (2006). Stable evolutionary signal in a yeast protein interaction network. *BMC Evol Biol* 6: 8.

Ye, P., B. D. Peyser, et al. (2005). Gene function prediction from congruent synthetic lethal interactions in yeast. *Mol Syst Biol* 1: 2005 0026.

Ye, P., B. D. Peyser, et al. (2005). Commensurate distances and similar motifs in genetic congruence and protein interaction networks in yeast. *BMC Bioinformatics* 6: 270.

Yu, H., N. M. Luscombe, et al. (2003). Genomic analysis of gene expression relationships in transcriptional regulatory networks. *Trends Genet* 19(8): 422-427.

Yukilevich, R., J. Lachance, et al. (2008). Long-term adaptation of epistatic genetic networks. *Evolution* 62(9): 2215-2235.

Yuryev, A. (2012). Contextual data integration in drug discovery. *Expert Opin Drug Discov* 7(8): 659-666.

Zhong, W. and P. W. Sternberg (2006). Genome-Wide Prediction of C. elegans Genetic Interactions. *Science* 311(5766): 1481-1484.

Zinman, G. E., S. Zhong, et al. (2011). Biological interaction networks are conserved at the module level. *BMC Syst Biol* 5: 134.

TABLEAU 2.1 *In silico* methodologies for the predictions of genetic interactions.

Reference	Type	Data	Training	Method	Number of predictive features	Experimental validation	Note
Tong et al. (2004)	Yeast SSL	Genetic interactions (GI)	~4,000 GIs	Network connectivity	1	No	GIs for ~20% of query genes
Wong et al. (2004)	Yeast SSL	Protein function and localization, gene expression, protein–protein interactions (PPI), phenotype, sequence homology	795,732 gene pairs (incl. 4,598 SSL)	Decision tree	26	Yes	Network topology and "2hop" relationships
Kelley and Ideker (2005)	Yeast SSL	PPI, protein-DNA, protein-reaction	4,849 SSLs and 27,604 PPIs for "naïve predictions"	Network connectivity	2	No	Between-pathway and within-pathway models
Zhong and Sternberg (2006)	*C. elegans* GIs	GIs/PPIs orthologs from yeast/fly; anatomical expression, phenotype, GO term, mRNA co-expression along with with orthologs (yeast/fly)	1,816 GIs; 2,878 PPIs; 3,296 *cis* markers as negatives	Logistic regression	5	Yes	18,183 high-confidence GIs covering 2,254 genes
Qi et al. (2008)	Yeast SSL	GIs	13,022 SSL	Graph diffusion kernel	1	Yes	Non-restricting distance measures to find new interacting partners
Paladugu et al. (2008)	Yeast SSL	PPIs	6,074 SSL; 400,473 negatives	Support vector machine	13	No	Graph-theoretic features using only PPIs as a network
Chipman and Singh (2009)	SSL for *S. cerevisiae* and *C. elegans*	Yeast: GO, PPI; worm: PPI with yeast/human orthologs, gene expression	Yeast: 22,432 SSL and 726,457 negatives; worm: 3,863 SSL and 58,579 negatives	Decision tree	5–6	No	Random walks algorithm to achieve topological similarity measurements between gene pairs
Lee et al. (2010b)	*C. elegans* GIs	mRNA co-expression, PPI, cocitation of genes name, phylogenetic profile analysis	626,342 GO-annotated gene pairs	Weighted sum	21	Yes	Functional network-guided predictions of genetic modifiers
Lee et al. (2010a)	*C. elegans* GIs	mRNA co-expression, PPI, phenotype	1,522 GIs	Logistic regression	6	Yes	Extended multi-species interactome and new phenotype/PPI network-based features
Szappanos et al. (2011)	*S. cerevisiae GI degree*	*GIs*	3,572 SSL and 1,901 positive GIs	Flux balance analysis	1	Yes	Prediction of GI connectivity for metabolic genes
Koch et al. (2012)	*S. cerevisiae GI degree*	*Several types including mRNA expression, GO, PPI, copy number, phylogenetic profile analysis*	~3,500 GIs with 63.2% of all genes used in training	Decision tree	16	Yes	Prediction of GI degrees in *S. pombe*
Hoehndorf et al. (2013)	Fish, yeast, fly, worm, mouse GIs	GO, GI, PPI, phenotype	GIs and PPIs as positives	Semantic similarity (with the Jaccard index)	2	No	Prediction of GIs from infererred gene functions

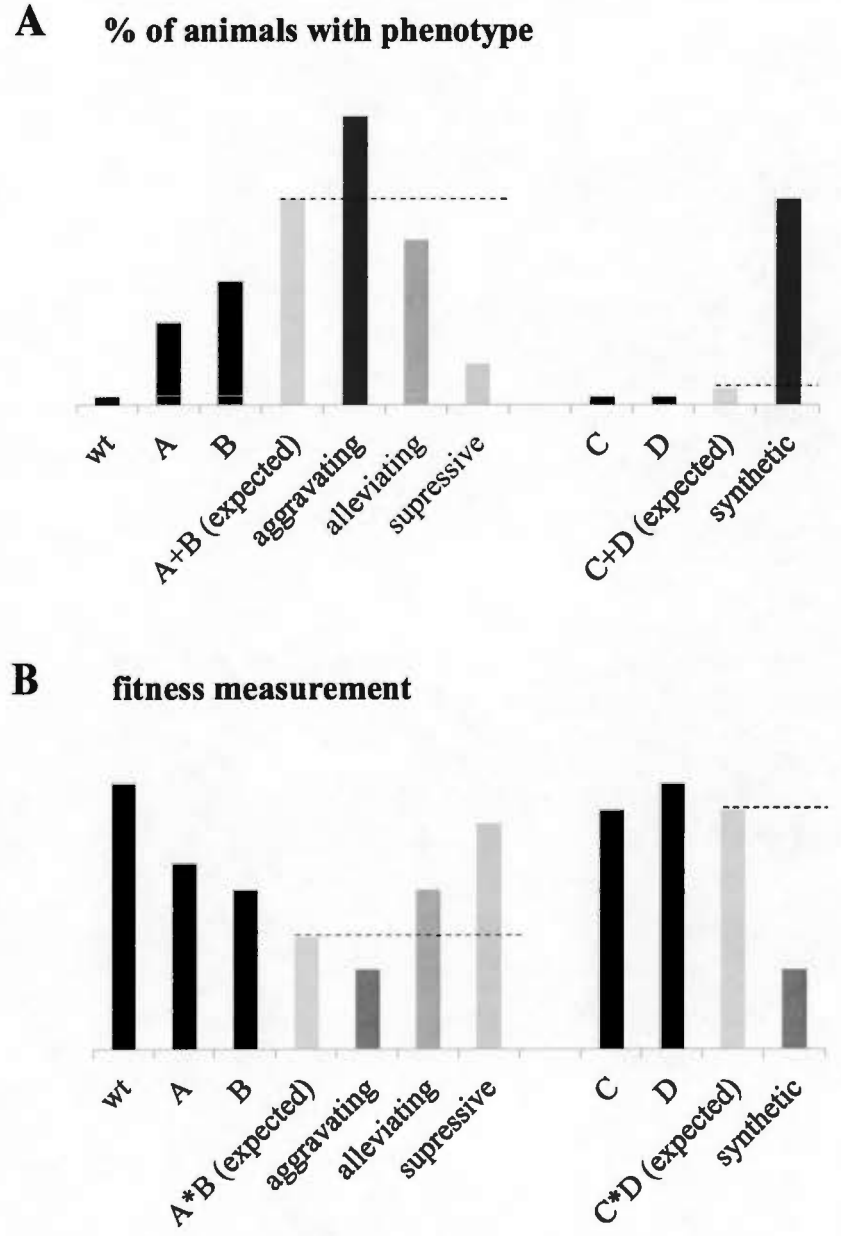


Figure 2.1 Statistical epistasis.

(A) When considering the penetrance of a given phenotype as the percentage of animals expressing this phenotype at a given "significative" level, genetic interactions (GIs) are usually identified using the additive model. Considering the phenotype of wild-type (*wt*) animals, close to zero, the expected phenotype of the double mutant AB corresponds to the sum of the phenotypes of mutant A and B. An aggravating GI between A and B is then identified if the phenotype of AB is significantly higher than the expected. An Alleviating GI is identified if the phenotype of AB is significantly lower than expected. A suppressive interaction is identified if the phenotype of AB is lower than the single mutant with the highest penetrance. When considering two mutants C and D with no observable phenotype, a synthetic interaction is identified if the double mutant CD expresses a significant phenotype. **(B)** When fitness is measured as a phenotype, the *wt* animals present high fitness rate, the expected phenotype of the double mutant AB is calculated using the multiplicative phenotype (it could also be the Log or Min) as the product of the fitness level of A and B. An aggravating interaction is then identified if AB is significantly lower than expected. Alleviating is identified if the fitness of AB is significantly higher than expected. Suppressive interaction is identified or if the double mutant is more viable than the sickest single mutants. A synthetic interaction is identified if the double mutant presents a significant fitness defect while the two single mutants are fit.

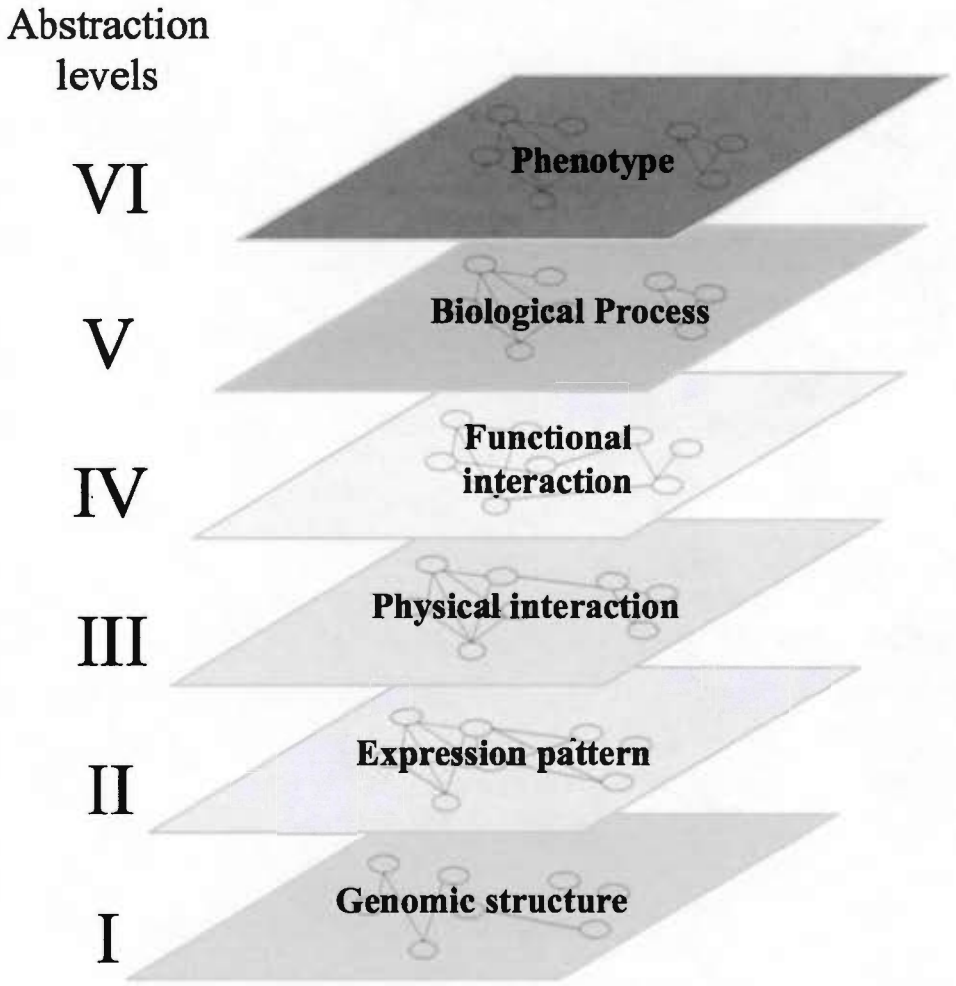


Figure 2.2 Representation of the six levels of abstraction in biological systems.

Note that, while each gene/protein can be followed from one abstraction level to another, the relationships linking it with its neighbours are different at each level. The conservation of links between two levels of abstraction in a given system and between orthologous genes/proteins in different systems are discussed in the main text of this review.

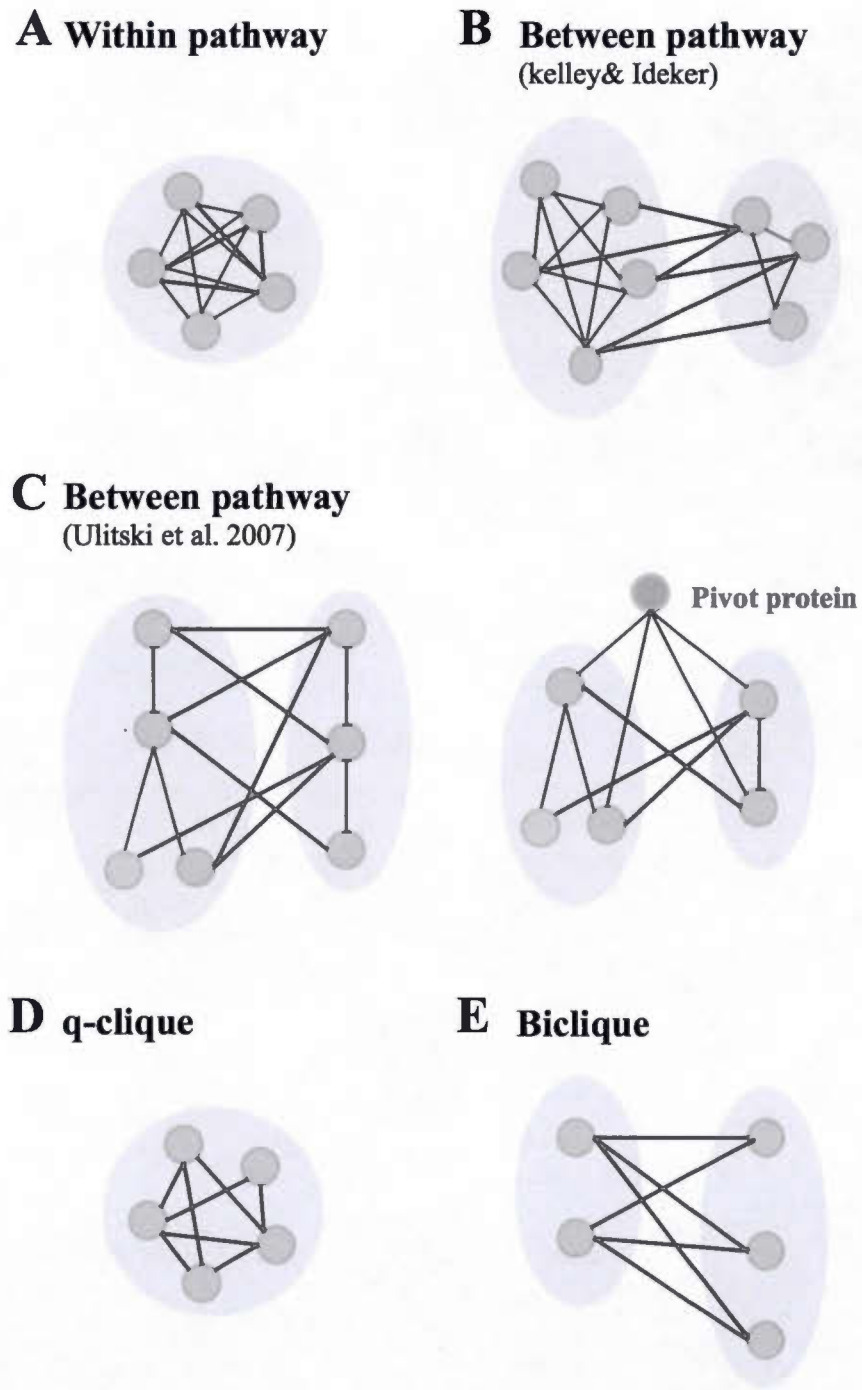


Figure 2.3 Canonical and non-canonical within/between pathway models.

(A) The canonical within pathway model as described by Kelley and Ideker, consists in genetic interactions (GIs, red edges) occurring within a dense module of protein interactions (bleu edges). **(B)** The canonical between pathway models, as described by Kelley and Ideker, consists in GIs occurring between dense modules of protein interactions. **(C)** The canonical between pathway models, as defined by Ulitski *et al.*, consist in GIs occurring between connected subnetworks/graph modules of protein interactions. This study also identified pivot proteins as proteins highly connected at the molecular level with component of two subnetworks connected through between pathway GIs. **(D)** non-canonical within pathway model, are quasi-cliques (q-cliques), biclusters of highly connected genes. **(E)** The non-canonical between pathway models consist in bicliques – biclusters in which prays and target genes of GIs do not overlap.

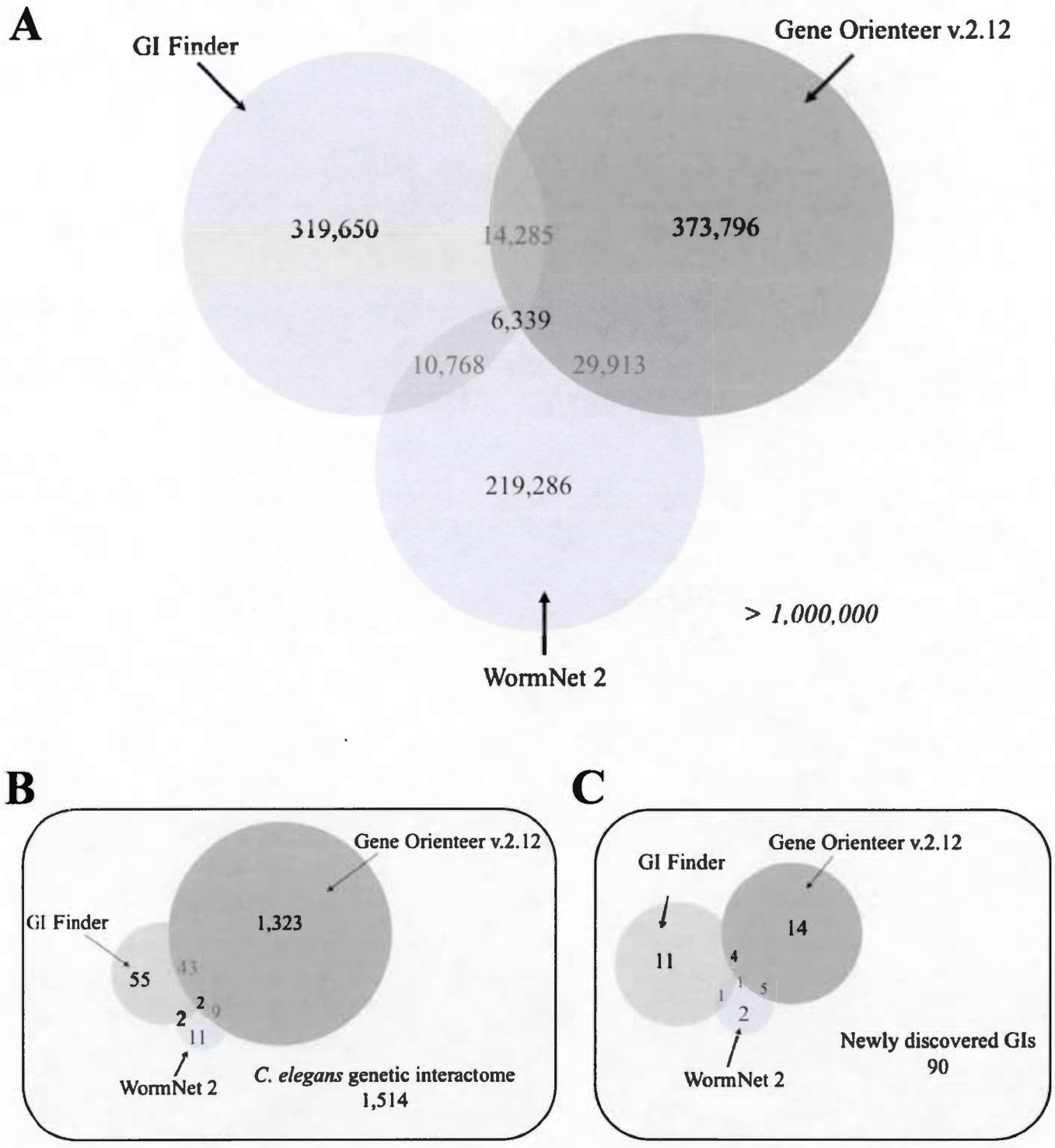


Figure 2.4 Venn diagrams of *C. elegans* predicted genetic interactions from three different approaches.

(A) Genome-wide predictions. **(B)** Experimentally validated genetic interactions taken from Lee, A.Y. *et al.*, 2010 **(C)** Experimentally validated genetic interactions (GIs) collected from recent studies (2009-2012). Numbers in red indicates statistically significant overlaps ($P < 0.05$), evaluated using an exact hypergeometric probability test. The selected score thresholds used to predict GIs yield the same false positive rate (FPR) for all three predictors. Each FPR was evaluated using a negative set consisting of 10,000 random gene pairs free of any gene present in validated interactions. Predicted GIs or functional links, for GeneOrienteer and WormNet respectively, were downloaded in October 2010.

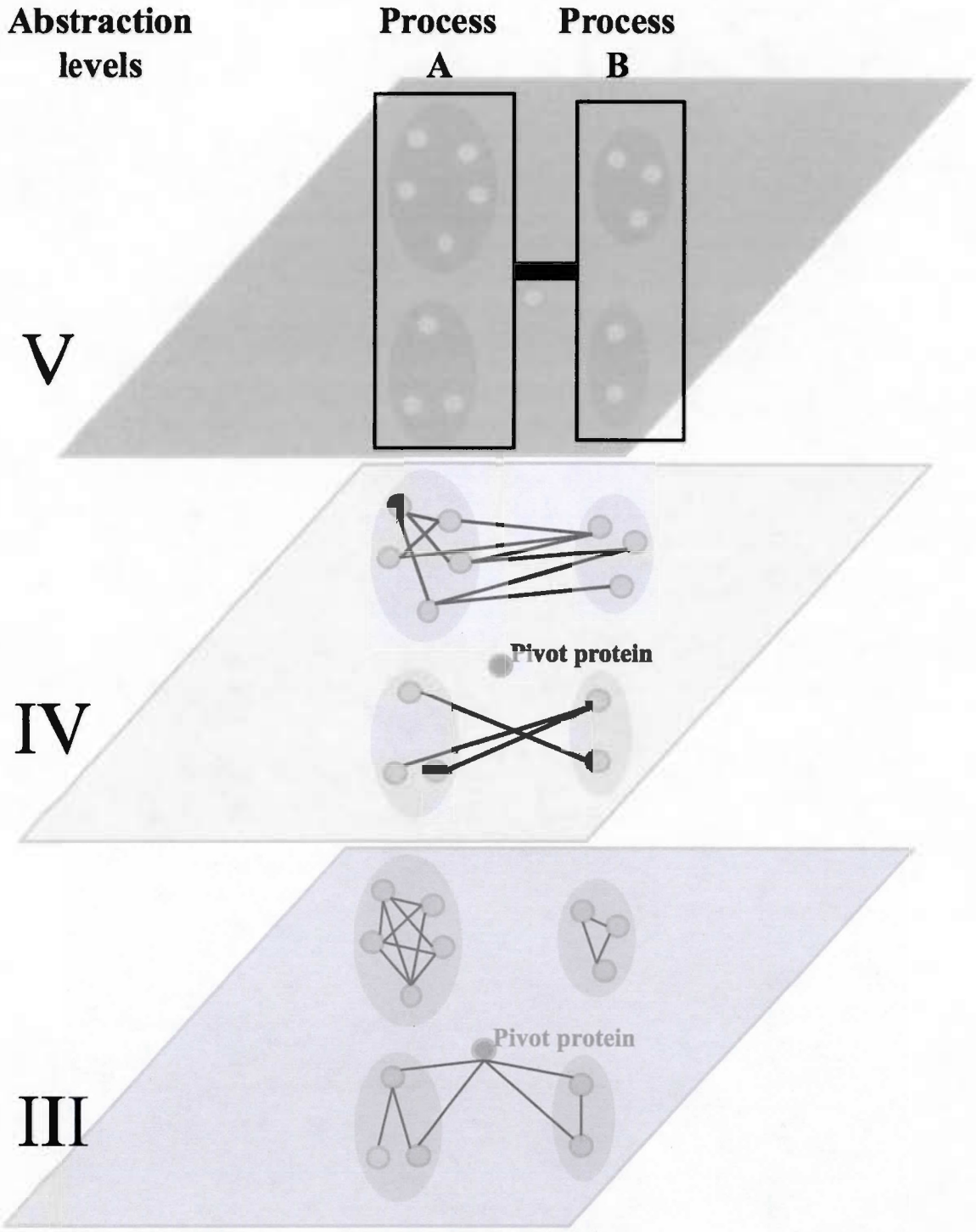


Figure 2.5 Integration of the abstraction level III, IV and V.

Abstraction level III shows protein-protein interactions (PPIs, blue edges) within highly connected protein interaction modules. It represents also a pivot proteins highly connected with proteins of two dense modules. The abstraction level IV shows the connection of dense protein modules through genetic interactions (GIs, red edges, between pathway model). It shows also the approximate rate of within pathway and between pathways GIs observed in yeast. The level V shows the clustering of dense modules in biological processes and the link brought by GIs between these processes. The strength of that link is more evolutionary conserved than individual GIs at the abstraction level IV.

CHAPITRE III

CARACTERISATION STRUCTURALE ET FONCTIONNELLE D'UN RESEAU D'INTERACTIONS GÉNÉTIQUES A L'INTERIEUR DES VOIES DE SIGNALISATION ET DE MÉTABOLISME DU NÉMATODE *CAENORHABDITIS ELEGANS*

L'ensemble de ce chapitre a été publié en février 2016 dans la revue *PLoS Computational Biology*, doi: 10.1371/journal.pcbi.1004738.

Boucher B, Lee AY, Hallett M, Jenna S. (2016) Structural and Functional Characterization of a Caenorhabditis elegans Genetic Interaction Network within Pathways. *PLoS Computational Biology*. 12(2):e1004738.

Pour ce chapitre III, il est question d'une caractérisation en profondeur des interactions génétiques chez le nématode *Caenorhabditis elegans*. Pour ce faire, nous avons utilisé des données se situant à différents niveaux d'abstraction et avons exploité certains liens entre ces niveaux pour montrer que les signatures de données, correspondant aux interactions génétiques, ne sont pas homogènes. Nous avons ensuite classé les interactions en sous-groupes homogène et sélectionné six classes. Nous avons ensuite identifié les différentes propriétés de ces classes pour construire un modèle de réseau génétique dans lequel chaque classe d'interactions joue un rôle particulier. Ce modèle ouvre la voie à une meilleure compréhension des liens entourant les interactions génétiques et les autres niveaux d'abstraction. Le modèle propose aussi une explication pour la robustesse des systèmes biologiques et l'évolution des espèces.

CONTRIBUTIONS

Pour cet article, j'ai (BB) rassemblé l'ensemble des données nécessaires à la caractérisation des classes d'interactions génétiques et effectué la majorité des analyses statistiques. J'ai validé plus de 50% des interactions génétiques manuellement dans la littérature. J'ai élaboré les algorithmes de randomisation des réseaux et conduit les analyses de fréquence subséquentes dans l'interactome génétique. J'ai aussi conçu l'algorithme de sélection des classes d'interactions génétiques basé sur les signatures de données des gènes qui interagissent ensemble.

Anna Y. Lee (AL) a fourni les données de base couvrant chaque paire de gènes du génome de *Caenorhabditis elegans*. Elle a aussi fourni plusieurs codes servant aux analyses statistiques.

Michael Hallett (MH) a effectué un suivi technique du projet.

Sarah Jenna (SJ) a conçu et dirigé le projet. Elle a effectué les analyses topologiques des gènes qui interagissent ensemble dans le réseau d'interactions protéine-protéine. Elle a classé manuellement l'ensemble des phénotypes utilisés dans cette étude pour définir l'index pléiotropique. Elle a aussi contribué à l'analyse statistique des index pléiotropiques.

J'ai rédigé la section Matériel et méthode, les légendes des figures ainsi qu'une partie des résultats de l'article. Sarah Jenna a rédigé le reste.

Détail des contributions par figure :

Figure 3.1: BB (collecte de données, analyse, design et réalisation des figures), AL (traitement des données, aide technique), MH (supervision technique), SJ (conception et design).

Figure 3.2 : BB (collecte de données, analyse, design et réalisation des figures), AL (aide technique), MH (supervision technique), SJ (conception et design).

Figure 3.3 : BB (analyse), SJ (conception, collecte de données, analyse, design et réalisation des figures).

Figure 3.4 : BB (collecte de données, analyse, design et réalisation des figures), AL (aide technique), MH (supervision technique), SJ (conception et design).

Figure 3.5 : BB (collecte de données, analyse, design et réalisation des figures), AL (aide technique), MH (supervision technique), SJ (conception et design).

Figure 3.6 : BB (collecte de données, analyse, design et réalisation des figures), AL (aide technique), MH (supervision technique), SJ (conception et design).

Figure 3.7 : BB (collecte de données, analyse, design et réalisation des figures), SJ (conception et design).

Figure 3.8 : BB (collecte de données, analyse, design et réalisation des figures), SJ (conception et design).

Figure 3.9 : SJ (conception, design et réalisation de la figure).

RÉSUMÉ

Une interaction génétique (IG) est définie lorsque la mutation d'un gène modifie l'expression phénotypique associé à la mutation d'un second gène. Les efforts pour cartographier les IG du génome en entier de la levure ont révélé des propriétés structurales et fonctionnelles d'un réseau d'IGs. Cela a donné un aperçu des mécanismes sous-jacents de la robustesse de la levure à des perturbations génétiques et environnementales, mais aussi au niveau du lien existant entre le génotype et le phénotype. Même si une conservation importante des IGs et de la structure du réseau d'IGs a été rapportée entre des espèces de levures lointaines, une telle conservation n'a toujours pas été observée entre les organismes unicellulaires et multicellulaires. La

caractérisation structurale et fonctionnelle d'un réseau d'IGs dans ces derniers organismes est par conséquent, d'un grand intérêt. Dans cette étude, nous présentons une caractérisation en profondeur de ~1.5K IGs du nématode *Caenorhabditis elegans*. Nous avons identifié et caractérisé six classes distinctes d'IGs en examinant un large éventail de propriétés structurales et fonctionnelles des gènes et du réseau, y compris la co-expression, les manifestations phénotypiques, la relation avec les sous-réseaux denses d'interactions protéine-protéine (SDI) et les voies, les fonctions moléculaires et biologiques, l'essentialité des gènes et la pléiotropie. Notre étude montre que les classes d'IGs sont liées à des gènes présents à l'intérieur des voies et présentent des propriétés distinctes, spécifiquement envers les SDI. Il suggère un modèle dans lequel les voies sont composées d'IGs SDI-centriques et SDI-indépendants coordonnées par des machines moléculaires à travers deux classes spécifiques d'IGs, les connecteurs pléiotropiques et non-pléiotropiques. Notre étude décrit la première caractérisation en profondeur d'un réseau d'IGs à l'intérieur des voies d'un organisme multicellulaire. Il suggère également un modèle pour mieux comprendre comment les IGs contrôle la robustesse du système et l'évolution.

3.1 Abstract

A genetic interaction (GI) is defined when the mutation of one gene modifies the phenotypic expression associated with the mutation of a second gene. Genome-wide efforts to map GIs in yeast revealed structural and functional properties of a GI network. This provided insights into the mechanisms underlying the robustness of yeast to genetic and environmental insults, and also into the link existing between genotype and phenotype. While a significant conservation of GIs and GI network structure has been reported between distant yeast species, such a conservation is not clear between

unicellular and multicellular organisms. Structural and functional characterization of a GI network in these latter organisms is consequently, of high interest. In this study, we present an in-depth characterization of ~1.5K GIs in the nematode *Caenorhabditis elegans*. We identify and characterize six distinct classes of GIs by examining a wide-range of structural and functional properties of genes and network, including co-expression, phenotypical manifestations, relationship with protein-protein interaction dense subnetworks (PDS) and pathways, molecular and biological functions, gene essentiality and pleiotropy. Our study shows that GI classes link genes within pathways and display distinctive properties, specifically towards PDS. It suggests a model in which pathways are composed of PDS-centric and PDS-independent GIs coordinating molecular machines through two specific classes of GIs involving pleiotropic and non-pleiotropic connectors. Our study provides the first in-depth characterization of a GI network within pathways of a multicellular organism. It also suggests a model to understand better how GIs control system robustness and evolution.

3.2 Introduction

The behaviour of biological systems and their adaptation to environmental changes depend on many factors on the path from genomic structure, through gene expression, molecular and functional interactions, to phenotypic manifestations. To simplify studies of these different levels of information, systems biologists may build a theoretical framework where biological systems are decomposed into six abstraction levels (Boucher et Jenna 2013): the genome structure (level I), the gene expression (level II), the physical interaction between systems elements (protein, DNA, RNA, etc. level III), the functional relationship between these elements (level IV), their biological and molecular function (level V) and the phenotypical manifestations (level VI).

Within this framework, genetic interactions (GIs) are located at the level IV together with signaling and metabolic pathways (Boucher et Jenna 2013).

The identification of a GI between two genes reveals that a mutation on the first one alters the biological consequences (the phenotype) associated to a mutation on the second one. Mapping GIs represents an important approach in understanding the link between genotype and phenotype. It is also a critical step to understand the robustness of biological systems – i.e. how the system compensates for the alteration of a function. Mapping GIs in human also recently emerged as a necessity to identify biomarkers from Genome-wide association studies (GWAS) and consequently, move the medical field towards a more personalized practice (Gibson 2010).

To date, only few (primarily unicellular) organisms have been amenable to experimental genome-wide screening approaches for mapping GIs. Thus, most of our information on the structure and the function of GI networks has been restricted to yeast (reviewed in (Boucher et Jenna 2013) and (Dixon *et al.* 2009)). Extensive studies on GIs in these systems showed clear relationships between GI networks and networks located at other abstraction levels. These studies revealed the relationship of GIs with signaling and metabolic pathways (Davierwala *et al.* 2005; Costanzo *et al.* 2010; Bellay *et al.* 2011; Szappanos *et al.* 2011), between co-expressed genes, and between genes coding for interacting proteins (Kelley et Ideker 2005; Ulitsky et Shamir 2007). They also identified the relationship between GIs, bioprocesses and phenotypes (Tong *et al.* 2004; Drees *et al.* 2005; Schuldiner *et al.* 2005). They characterized the degree of connectivity of genes within GI networks and assessed their enrichment in genes with high connectivity (GI-Hubs) as well as in multifunctional and essential genes (Ozier, Amin et Ideker 2003; Tong *et al.* 2004; Davierwala *et al.* 2005; Szappanos *et al.* 2011). Importantly, these studies identified dense subnetworks within the GI network (GDS) (Tong *et al.* 2004). They showed that GDS tend to lay between molecular machines, that we will define in this study as dense subnetworks of protein-protein and protein-DNA interactions (Kelley et Ideker 2005; Ulitsky et Shamir 2007; Szappanos *et al.*

2011). They also showed that GDS are monochromatic, i.e. they are composed of either positive (suppressive/alleviating) or negative (synergistic/aggravating) GIs (Segre *et al.* 2005; Pu *et al.* 2008; Jaimovich *et al.* 2010; Michaut *et al.* 2011) and are functionally biased (Szappanos *et al.* 2011; Sharifpoor *et al.* 2012). These studies connect four abstractions levels (levels II to V), showing that GI networks coordinate molecular machines within bioprocesses.

These studies using yeast as a model brought also precious information on the role played by GIs in genomic robustness and evolutionary processes (Roguev *et al.* 2008; Bellay *et al.* 2011; Ryan *et al.* 2012). For example, these studies identified two separate groups of duplicated genes within distinct GDS: a group composed of redundant genes playing an important role for the genomic robustness of the system and a group of redundant genes with divergent biological functions and with expected reduced impact on robustness (Bellay *et al.* 2011). In addition, they revealed that positioning of GIs within or between PPI-dense subnetworks (PDS) had an impact on their evolutionary conservation: GIs within PDS being more conserved than GIs between PDS (Roguev *et al.* 2008; Ryan *et al.* 2012).

To date, the structure and the function of GI networks are still largely unknown in multicellular organisms. Characterizing these networks is therefore required to better understand the genomic robustness of these systems and also how functional relationships between alleles influence phenotypical outcomes and evolutionary processes in multicellular contexts. To address this problem, we provide the first deep characterization of a network composed of ~600 GIs in a multicellular organism, the nematode *Caenorhabditis elegans*. This study aims to identify functional properties associated with GIs and groups of GIs and to understand better how the structural and functional organization of a GI network links molecular machines (abstraction level III) to bioprocesses, phenotypes and diseases (abstraction levels V and VI) in a multicellular context. Our results indicate that GIs form a heterogeneous group of entities when considering biological data located at different abstraction levels in *C.*

elegans. We describe the specific characteristics of GI classes with respect to the connectivity degree within the GI- and the PPI-networks, their relationship with protein-protein interaction dense subnetworks (PDS), signaling and metabolic pathways and with phenotypic manifestations (essentiality, pleiotropy). We also discuss the impact of this structure on *C. elegans* genome robustness and evolution.

3.3 Results

3.3.1 Defining six classes of genetic interactions

Considering that the function and the structure of genetic interaction networks are mainly unknown for multicellular organisms while being of increasing interest, we characterized a network composed of ~1,500 GIs of the nematode *C. elegans*. To do so, we first investigated whether GIs constitute a heterogeneous group of entities in this organism and consequently, whether we could identify several GI classes with distinctive biological properties in this network.

We retrieved 1,514 genetic interactions (GIs) from Wormbase, Biogrid and the literature as described in the Methods section and in Table 3.S1 [en ligne: (Boucher *et al.* 2016)]. This set of GIs, called GIs-all, is composed of 750 (49.5%) interactions identified as experimentally validated GIs by either Wormbase and Biogrid curation systems (see Methods) or manually curated from the literature in our laboratory (Lee *et al.* 2010). The remaining 764 GIs (50.46%) were identified using Textpresso, an automatic text mining system (Muller, Kenny et Sternberg 2004). To test the false-positive rate in this later GI set, we manually curated 261 of these interactions and found that 252 of them (96.55%) were true-positives (experimentally validated GIs; Table 3.S1 [en ligne: (Boucher *et al.* 2016)]). Overall, GIs-all is predicted to contain at

least 98% of validated GIs.

Statistical attributes using expression, protein-protein interaction (PPI) and phenotypic data were previously described as powerful tools to segregate GIs-all from a set of gene-pairs randomly selected from the genome (Lee *et al.* 2010). GIs being shown as rare events (Tong *et al.* 2004), we expect the latter set of gene-pairs to be mostly composed of "true" negative examples of GIs (see Methods). Attributes used in this study capture the level of co-expression between interacting genes (Exp, Figure 3.1a, b), the enrichment of shared phenotypes (Ph, Figure 3.1a, b), their ability to encode proteins that interact physically (I, Figure 3.1a, b) and/or have more common interacting-partners than expected by chance alone (CI, Figure 3.1a, b). Two attributes were also designed to identify GIs within functional modules (N and NPh, Figure 3.1a, b). The attribute called "Neighborhood" (N) identifies the enrichment of phenotypes in the neighborhoods of gene-pairs within a network where genes are linked when they are significantly co-expressed, or code for interacting proteins (see Methods for details; (Lee *et al.* 2010)).The attribute "Neighborhood with Phenotype" (NPh) identifies pairs of genes with a positive value for the attribute N and the association of both genes with the phenotype enriched in their neighborhood (see Methods for details; (Lee *et al.* 2010). We used these attributes to assess whether GIs display distinctive dependency towards data located at different abstraction levels. To do so, attribute values were used to cluster GIs-all together with 1,500 gene-pairs randomly chosen in the *C. elegans* genome. If GIs were forming an homogeneous group of entities with unique distinctive properties when considering negative sets of examples, we would expect them to cluster as a group from these negative examples. However, GIs-all tend to cluster into groups of GIs (using Euclidean distance-based dissimilarities; see Methods) with some dispersion among the negative examples (Figure 3.1a). This suggests that GIs-all can be subdivided into GIs groups with potential distinctive biological properties.

In an attempt to characterize the function of these GI groups, we used a cluster selection algorithm to identify GI classes from GIs-all (Figure 3.1b), while controlling the

robustness of this classification (see Supplementary Information, Text S3.1 (voir Annexe A); Figure 3.S1). This study identified ten GI classes, six of them (C1-C6) being significantly different from gene pairs randomly selected from the genome (see Supplementary Information, Text S3.1 (Voir Annexe A); Figure 3.S2). Analysis of the proportion of missing data supporting each class revealed that missing expression, phenotype and/or PPI data could not explain the classification by itself. We conclude from this analysis that GI classes with similar combination of attribute values would still be identified from a GI set exempt of missing data.

3.3.2 GI classes have different functions

Two main classes of GIs have been identified in the yeast genome: positive (alleviating interactions) and negative (synergistic interactions) (Tong *et al.* 2001; Kelley et Ideker 2005; Schuldiner *et al.* 2005; Collins *et al.* 2007; St Onge *et al.* 2007). Monochromaticity of GI dense subnetworks (GDS), enriched in either positive or negative interactions, has also been identified in yeast (Bellay *et al.* 2011; Sharifpoor *et al.* 2012). These monochromatic GDS were described as being functionally biased - i.e. they tend to be associated with specific biological functions (Bellay *et al.* 2011; Sharifpoor *et al.* 2012). This characterization of GI networks in yeast was critical to understand better the role of GIs in coordinating gene function within biological processes.

To characterize the potential distinctive biological properties of GI classes and their ability to control different bioprocesses, we first assessed whether GIs tend to form dense subnetworks (GDS) in *C. elegans*. We also tested whether these GDS may be monochromatic or multichromatic - if they were enriched in one or several GI classes. Lastly, we assessed whether these GI classes are functionally biased.

This analysis revealed that GIs from GIs-all form few GDS (see Supplementary Information, Text S3.1 (Voir Annexe A); Figure 3.S3a). However, while the number of GDS formed is too small to assess their enrichment in unique class or combination of classes as done in yeast, clustering of classes based upon the GDS composition, revealed that GI classes tend to form GDS in a biased manner: C1 with C2, C4 with C5 and C3 with C6 (see Supplementary Information, Text S3.1 (Voir Annexe A); Figure 3.S3a). Analysis of the gene composition of GI classes also revealed that C4 GIs share more genes with C5 than with the other classes (see Supplementary Information, Text S3.1 (Voir Annexe A); Figure 3.S4) supporting the hypothesis that GI classes may assemble into GDS in a biased way.

Enrichment of GI classes in Gene Ontology annotations (GO, (Ashburner *et al.* 2000)) was also measured. GO annotations were found enriched for all GI classes except C2 (Figure 3.2a; Table 3.S2 [en ligne: (Boucher *et al.* 2016)]; Figure 3.S5-3.S11). This analysis was done first by considering gene repetition within classes (some genes being involved in more than one GI in a given class) and secondly, without considering gene repetition (asterisk indicates enrichment observed without considering gene repetitions; Figure 3.2a; Table 3.S2 [en ligne: (Boucher *et al.* 2016)]). Similar results were obtained for both analyses, suggesting that GI classes are functionally biased as detailed below.

C1 interactions were enriched in genes involved in cell division ($P = 0.01$, Fisher's exact test; Figure 3.2a; Table 3.S2 [en ligne: (Boucher *et al.* 2016)]). For example, C1 GIs were identified between *spd-2*, *spd-5* and *dhc-1* controlling centrosome assembly and maturation (Kemp *et al.* 2004) (Figure 3.S5). We expect genes involved in cell division to be essential. To test a potential enrichment of essential genes in C1, we measured the enrichment of two gene sets of 294 and 1,259 essential genes identified using systematic RNA silencing by Kamath *et al.* (Kamath *et al.* 2003) and Sonnichsen *et al.* (Sonnichsen *et al.* 2005) in GI classes when compared to GIs-all using a hypergeometric test (bars above the red line indicate a significant enrichment; $P < 0.05$;

Figure 3.2b). These data showed that C1 and C2 display the highest enrichment level in essential genes when compared to GIs-all and other classes. These data support the functional bias of C1 towards cell division and also associate GDS enriched in C1 and C2 interactions to essential biological functions. Enrichment analysis of GO terms within GI classes also revealed that C4 and C5 were enriched in genes coding for kinases ($P < 0.0001$, Fisher's exact test; Figure 3.2a; Table 3.S2 [en ligne: (Boucher *et al.* 2016)]). A kinase-centric study of the GI network in yeast showed that genes coding for kinases are often redundant and are involved in either kinase-kinase (K-K) or kinase-substrate (K-S) GIs (Sharifpoor *et al.* 2012). We assessed whether GI classes were enriched in these two kinds of interactions when compared to GIs-all. In this study, kinases were identified based on their GO annotation (Table 3.S2 [en ligne: (Boucher *et al.* 2016)]) and substrates were identified as non-kinases. This analysis revealed that C4 is significantly enriched in K-K ($P = 0.004$, Fisher's exact test) and K-S ($P = 0.03$, Fisher's exact test). Considering that K-K interactions were primarily observed between redundant kinases (Sharifpoor *et al.* 2012), we assessed whether C4 was also enriched in GIs between redundant genes. We thus measured the enrichment in GI classes of 306 genes identified as redundant (linked by a synthetic sick or lethal interaction) and evolutionarily conserved genes (Tischler *et al.* 2006). This analysis revealed that both C4 and C5 were enriched in evolutionarily conserved redundant genes (Bars above the red line indicate significant enrichments; $P < 0.05$; Figure 3.2c). In order to assess whether C4 and C5 were enriched in interactions involving only redundant kinases or redundant genes at large, we measured this enrichment after removing kinases from both the Tischler *et al.* list of conserved redundant genes and from GI classes. This analysis revealed that only C4 was enriched in GIs involving non-kinase redundant genes (Fisher's exact test; $P = 9 \times 10^{-6}$). This suggests that C5 GIs involve mainly conserved redundant kinases while C4 GIs involve conserved redundant genes coding for kinases or not. Examples of C4 GIs between conserved genes coding for redundant kinases and non-kinases are the interactions between the type I and type II TGF-beta receptor coding genes *daf-1* and *daf-4* (Georgi, Albert et

Riddle 1990; Estevez *et al.* 1993; Gunther, Georgi et Riddle 2000); and between the three Rho GTPases *ced-10, mig-2* and *rac-2* shown to control cell migration (Lundquist *et al.* 2001) (Figure 3.S5). Moreover, GO annotation enrichments revealed that C3 was enriched in genes involved in cell signaling ($P = 0.01$, Fisher's exact test; Figure 3.3a) and C6 in small GTPase signaling ($P = 0.001$, Fisher's exact test; Figure 3.3a).

Altogether, these data suggest that GI classes may assemble in GDS in a "class-dependent" manner and also display a functional bias. Genes interacting through C1 and C2 GIs tend to be essential genes, more particularly involved in cell division. Genes interacting through C4 or C5 GIs tend to code for evolutionarily conserved redundant kinases and non-kinases. This study also suggests a potential function for genes interacting through C3 and C6 GIs in cell signaling.

3.3.3. GI classes have different positions in the PPI network

The coordination of molecular machines by GIs has been intensively investigated in yeast through characterization of the relationship existing between GIs and protein-protein interaction (PPI) networks (Kelley et Ideker 2005; Ulitsky et Shamir 2007; Costanzo *et al.* 2010). Four out of the six attributes used in this study are built, even partially, on PPI data (N, NPh, CI, I; see Methods). GI classes, displaying positive values for different combinations of attributes (Figure 3.1b), may exhibit distinct relationship with the PPI network. To assess the relationship between GI classes and molecular machines we measured the average PPI-degree of proteins coded by genes within GI classes and GIs-all (Figure 3.3a). This revealed that C4 and C5 GIs involve genes with a median PPI-degree significantly higher than the other classes and GIs-all ($P < 10^{-3}$ and $P < 10^{-6}$ respectively, Wilcoxon rank-sum test; Figure 3.3a). We then assessed whether this was due to a significant enrichment of genes coding for PPI-Hubs

- defined as the 20% proteins with the highest PPI-degree within the PPI-network (Figure 3.3b; Figure 3.S12). This revealed that C4 and C5 were indeed significantly enriched in PPI-Hubs when compared to GIs-all ($P < 10^{-4}$ and $P < 10^{-11}$ respectively, Fisher's exact test; Figure 3.3b).

The structure of the PPI networks has been extensively studied in unicellular and multicellular organisms (Chang *et al.* 2013). As shown for GIs, PPI tend to assemble into PPI-dense subnetworks (PDS) orchestrated around nodes. These nodes can either be Hubs or non-Hubs, with distinct structural functions within the PPI network (Yu *et al.* 2007; Chang *et al.* 2013). The betweenness centrality metric was previously defined to identify network nodes playing a critical role in the PPI-network organization (called "bottleneck nodes") (Yu *et al.* 2007). These nodes were characterized by their positioning in many shortest paths within the PPI network ([Yu *et al.* 2007], Figure 3.3c). Measurement of the betweenness centrality of PPI-non-Hubs and PPI-Hubs in the different GI classes revealed that C5 non-Hub proteins tend to be bottlenecks with a significantly higher betweenness centrality than GIs-all ($P < 10^{-9}$, Wilcoxon rank-sum test; Figure 3.3d). The betweenness centrality of PPI-Hubs was not significantly different between GI classes and GIs-all (Figure 3.3e).

Altogether, these data clearly show that GI classes have distinctive properties with respect to the PPI network and may consequently coordinate differently molecular machines in biological processes. They identify C4 and C5 interactions as enriched in genes coding for PPI-Hubs that tend to be non-bottlenecks (green node; Figure 3.3c), suggesting that these Hubs may be embedded within PDS rather than at their periphery. C5 also appear to be enriched in non-Hub Bottleneck nodes (Blue nodes, Figure 3.3c) thought to play a critical role in PDS coordination (Yu *et al.* 2007). Therefore, these data suggest, C4 and C5 interactions may play a critical role in controlling the assembly and the coordination of PDS.

3.3.4 GI classes are either PDS-centric or PDS-independent

The data presented above suggest that C4 and C5 GIs link genes coding for proteins within or between PDS while C1, C2, C3 and C6 may link genes coding for proteins outside PDS. To test this hypothesis we used the Cytoscape "MINE" plugin (Rhrissorrakrai et Gunsalus 2011) to identify PDS within the PPI-network as previously done (Kelley et Ideker 2005; Ulitsky et Shamir 2007) (see Methods). This approach identified 106 PDS containing at least four proteins. We then assessed whether GIs-all and GI classes were enriched in genes coding for proteins: (i) in the same PDS (GI within-PDS; red lines / bars; Figure 3.4a and b); or (ii) between PDS (pairs of proteins located in different PDS; GI between-PDS; blue lines / bars; Figure 3.4a, b). To evaluate this enrichment level, we compared the frequencies obtained for the different scenarios (within-PDS and between-PDS) with those for randomized GI networks of similar size and structure, and computed log-ratio transform (LR) scores (see Methods, Figure 3.4b). Consistently with previous studies (Lehner 2007), GIs-all was enriched in within-PDS connections while being depleted in between-PDS connections (Figure 3.4b), confirming that GIs in *C. elegans* are more frequently observed within-PDS than between-PDS. Interestingly, C2 and C3 GIs tended to occur between-PDS, and C4 and C5 GIs occur both within- and between-PDS, while GIs in C1 and C6 appear to be independent from PDS (Figure 3.4b). We assessed whether this PDS-independency results from a dependency of C1 and C6 towards connected 3-protein triangles, called bistable motifs, which are not considered as PDS in our study (see Methods). However, no bistable motif was found in GIs-all and consequently, suggest that C1 and C6 do not form any bistable motif GIs and are independent from PDS.

These data confirm our hypothesis that C4 and C5 GIs are enriched in within- and

between-PDS. We will consequently, characterize them as PDS-centric interactions, and C1 and C6 GIs as PDS-independent. Moreover, C2 and C3 GIs appeared to be enriched in between-PDS connections suggesting some kind of functional relationship with the PDS-centric interactions. Taken together these data suggest that GI classes display distinct relationships towards PDS and consequently, may coordinate differently molecular machines within biological processes.

3.3.5 PDS and pathways constitute different functional modules

In the yeast *S. cerevisiae*, subgroups of GIs have been identified and characterized based on their relationship with physical interaction networks – mainly protein-protein (PPI) and protein-DNA (PDI) interactions (Kelley et Ideker 2005; Ulitsky et Shamir 2007; Costanzo *et al.* 2010). In these studies, a " pathway " was defined as a dense subnetwork or a connected graph module of PPIs (defined as PDS in our study) and PDIs. These studies showed that GIs tend to occur rather between- than within-pathways (Kelley et Ideker 2005; Ulitsky et Shamir 2007; Costanzo *et al.* 2010). Interestingly, a similar study was done on GI networks in *C. elegans* and showed the opposite – GIs occur more frequently within- than between-pathways (Lehner *et al.* 2006; Lehner 2007). These results are in agreement with data obtained in this study for GIs-all (Figure 3.4b).

The pathway described above is quite different from the definition used by developmental geneticists for whom a pathway consists in a group of genes functioning together to control a given biological process (McNeill et Woodgett 2010) (Figure 3.5a). Characterization of the relationship between PDS and pathways is consequently required to clarify the relative positioning of GIs and pathways (both located at the abstraction level IV) between molecular machines (level III) and biological processes

(level V). To do so, 61 pathways were retrieved from the KEGG database (Kanehisa *et al.* 2004) along with 33 pathways controlling the embryonic and larval development of *C. elegans* that we manually curated from the literature (Figure 3.S3 [en ligne: (Boucher *et al.* 2016)]; see Methods). We then assessed whether gene-pairs involved in a given pathway (within-pathway interactions) tend to code for proteins, which are part of the same PDS (within-PDS; orange lines / bars; Figure 3.5a and b) or of two distinct PDSs (between-PDS; green lines / bars; Figure 3.5a and b). We compared the frequencies obtained for these scenarios with those resulting from randomization of pathways (see Methods, Figure 3.5b). Interestingly, proteins coded by gene-pairs in the same KEGG or pathways from the literature (within-pathway gene-pairs, KEGG and Lit. respectively; Figure 3.5b) are mainly found in within- and between-PDS interactions (orange and green bars respectively; Figure 3.5b). We also confirmed that this enrichment of within- and between-PDS observed within-pathways did not depend on the topology of the PDS network but instead depend upon the interactions themselves (see Supplementary Information; Text S3.1 (Voir Annexe A); Figure 3.S14).

These data suggest that pathways and PDSs are distinct functional modules. They also suggest that pathways are composed of several PDS and that proteins involved in a given pathway may be part of the same PDS or of different PDSs. We confirmed these assumptions through a close examination of genes/proteins involved in pathways and PDSs as shown in 3.S4 [en ligne: (Boucher *et al.* 2016)] and detailed in supplementary information (Text S3.1 (Voir Annexe A)).

3.3.6 GI classes are enriched in within-pathway interactions

Considering that pathways and PDS are distinct functional modules and that GI classes

have distinct relationships towards PDS, we were interested to test whether this would also be the case for pathways. We thus investigated if GI classes and GIs-all were enriched in within-pathways (pairs of genes functioning in at least one common pathway, see Methods) and/or between-pathways (pairs of genes involved in at least one pathway but not involved in any common pathway, see Methods; Figure 3.6a). As detailed for PDS in the previous sections, we compared the frequencies obtained for the different scenarios (within-pathways and between-pathways) with those for randomized GI networks (Figure 3.6b and c, see Methods). GI classes and GIs-all were enriched in within-pathway interactions for both pathways retrieved from the KEGG database and from the literature (Figure 3.6b and c respectively). Remarkably, GI classes were depleted in between-pathway interactions while GIs-all was enriched in this kind of interactions (Figure 3.6b and c). These results suggest that non-selected clusters, C7 to C10 (Figure 3.1b), are enriched in between-pathway interactions. Characterization of these later GI classes revealed that it was indeed the case: except for C10 that was enriched only in within-pathway interactions, C7, C8 and C9 GIs were enriched in both within- and between-pathways interactions (Figure 3.S13)

Together, these data show that C1/C6 PDS-independent, C4/C5 PDS-centric and C2, C3 GIs are enriched in within-pathways interactions. This implies that pathways involve both functional interactions that are organized around PDS and other that are independent from PDS.

3.3.7 C3 and C6 GIs involve connectors

Highly connected genes called connectors, or GI-Hubs, are genes whose alteration impacts on a large number of genes. Their function is consequently expected to be central within pathways and bioprocesses (Lehner *et al.* 2006; Costanzo *et al.* 2010;

Szappanos *et al.* 2011). GI-Hubs are defined here as the 20% genes with the highest GI degree within the GI-network (Figure 3.S16). In order to understand better the function of GI classes within pathways, we characterized the distribution of GI-Hubs/connectors within GIs-all and their potential enrichment in GI classes. Therefore, we computed the average GI degree of interacting genes in GI classes and GIs-all (Figure 3.7a), and also measured their enrichment in GI-Hubs (Figure 3.7b). This analysis revealed that C3 and C6 showed a median GI-degree significantly higher than GIs-all and are enriched in GI-Hubs (Figure 3.7a and b). These results suggest C3 and C6 are enriched in connectors that may play a critical role in the coordination of gene functions within pathways. Such a function for C3 and C6 GIs is consistent with their expected involvement in cell signaling (Figure 3.2a).

3.3.8 Identifying pleiotropic and non-pleiotropic GIs

Our study supports a model in which four out of the six classes of GIs coordinate the function of genes either in a PDS-centric or in a PDS-independent manner. They also identify two additional classes of GIs with central coordination function through the involvement of connectors/GI-Hubs. The *C. elegans*, being a multicellular organism, one may ask whether some parts of this organization may be organ or process specific while others may be ubiquitous. To answer this question, we assessed whether GI classes are enriched in pleiotropic genes – i.e. genes whose genetic alteration is associated to multiple phenotypic expressions and consequently, whose function is required in several organs and/or at different developmental stages of the animal. To do so, we divided *the C. eleg*ans phenotype ontology into 22 groups of phenotypes (Figure 3.S15). We computed a pleiotropic index (PI) for each gene as the number of phenotype groups containing at least one phenotype expressed upon genetic alteration

of the gene of interest (see Methods). The PI median genome-wide (2 ± 1.48 MAD) is smaller than in GIs-all (5 ± 2.97 MAD; Figure 3.S16b), and its distribution revealed that a large number of genes displayed a low PI while a small number of genes have a high PI (Figure 3.S16b). Using this distribution of PI, we defined a class of highly pleiotropic genes (High-PI) as the 20% genes displaying the highest PI genome-wide (Figure 3.S16b).

We measured the PI of genes involved in GI classes and found that only genes involved in C1, C2 and C6 GIs displayed significantly higher median PIs than GIs-all ($P < 10^{-4}$, Wilcoxon rank-sum test; dark grey boxes; Figure 3.7c) and are significantly enriched in High-PI genes ($P < 10^{-5}$; Fisher's exact test; Figure 3.7d). We further characterized the distribution of GI classes at different PI ranges and the ability of these classes to link genes between ranges (see Supplementary Information, Text S3.1 (Voir Annexe A); Figure 3.S17). This study revealed that while C5 connects genes only within an average PI range, other classes connect genes across PI ranges, especially C3 and C6 that tend to link genes within an average PI range to either genes with low (C3 class) or high PI (C6 class).

These data suggest that PDS-independent GIs tend to involve pleiotropic genes while PDS-centric GIs involve genes with an average to low pleiotropy. They also further characterized the properties of connectors involved in C3 and C6 GIs, linking genes within different pleiotropic ranges: C6 and C3 GIs link genes from an average to a high-pleiotropy and from a low to an average pleiotropy respectively. These data show that the organization of the characterized GI network within pathways is more complex when considering the multicellular nature of *C. elegans*. They imply that genetic mutations within a given pathway may either lead to pleiotropic or tissue/developmental stage-specific phenotypic manifestations. This also implies that such mutation may be compensated by a mutation in a genes involved in the same pathways but with pleiotropic or non-pleiotropic effect.

3.3.9 Integration of different kind of data is required to identify GI classes

PDS-centric GIs (C4 and C5) are GI classes displaying positive values for I and CI attributes (Figure 3.1b), which are attributes mainly built on PPI network. We, consequently, wondered whether these two classes of GIs could have been identified considering PPI data alone. In order to test the value of the data integration strategy used in this study to identify the GI classes, we assessed whether functional characteristics observed for each class may depend on one or a combination of the attributes used for the classification (Figure 3.1a and b). To answer this question, we defined a threshold value for each attribute [en ligne: (Boucher *et al.* 2016)], allowing us to identify groups of GIs associated to a positive or a negative value per attribute. We subsequently assessed the enrichment in GI-Hubs, PPI-Hubs, redundant genes, essential genes, genes with high PI (High-PI; Figure 3.8a and 3.S18), within- and between-PDS (W-PDS and B-PDS; Figure 3.8b and 3.S19) as well as within- and between-pathway interactions (W-Path and B-Path respectively, Figure 3.8b) in these GI groups. We clustered GI groups and GI classes (C1-C6) based on these enrichment levels using Euclidean distances (Figure 3.8a and b). This analysis revealed that the enrichment of a given GI property is not associated to a positive or a negative value for a single attribute but for several of them (Figure 3.8a and b). For example, enrichment of genes with High-PI was observed in groups of GIs with positive or negative values for Ph (enrichment of phenotypic manifestations), with negative value for CI (high number of common partners within the PPI network), and positive value for NPh (enrichment of a phenotype associated to interacting genes in their respective neighborhoods). This suggests that integration of a set of attributes defines the biological functions of identified GIs.

This analysis also revealed that the biological characteristics observed for GI-classes are not found in GI groups associated to either a positive or a negative value for an attribute. For instance, when comparing C4 and C5 with the GI group positive for CI (interacting genes coding for proteins sharing a significantly high number of common PPI partners), these GI groups/classes displayed similar enrichment of redundant genes, PPI-Hubs and essential genes (blue square; Figure 3.8a) but have different enrichment profiles when considering their relative positioning towards pathways and PDS (blue square; Figure 3.8b). This suggests that the enrichment of essential genes, PPI-Hubs and redundant genes in C4 and C5 may be significantly influenced by a positive value for the CI attribute (Figure 3.1b). Positive value for this attribute does not, however, explain the depletion of C4 and C5 in GIs between-pathways and their enrichment in GI between-PDS (Figure 3.8b).

Altogether, this study demonstrates that the biological characteristics identified for GI classes depend on the combination of attributes identified through data integration strategy used for this study. Importantly, it indicates which combination of attributes is appropriate in integration to identify classes of GI with specific biological functions.

3.4 Discussion

Extensive mapping and characterization of the yeast genetic interaction network has revealed precious information on the still mysterious genetic interactome. These studies identified multiple classes of GIs with distinctive properties when considering the protein-protein interaction network and biological processes (bioprocesses) (Kelley et Ideker 2005; Costanzo *et al.* 2010; Bellay *et al.* 2011). Therefore, these studies suggested that GIs display a certain level of heterogeneity. Whether such heterogeneity could also be observed in higher organisms was however still unknown.

In this study, we characterized the structure of a GI network composed of ~1,500 GIs in the nematode *Caenorhabditis elegans*. We showed that GIs form a heterogeneous group of entities with respect to attributes computed from expression, protein-protein interaction (PPI) and phenotype data. From this set of GIs, we identified six classes covering ~600 GIs, named C1 to C6 that cluster apart from gene-pairs randomly picked from the genome. Characterization of these GI classes revealed that they are either centered on protein-protein interaction dense subnetworks (PDS) and called PDS-centric interactions, or independent of these PDS (PDS-independent interactions) (Figure 3.9). The current study shows that PDS-centric interactions are composed of C4 and C5 GIs, and tend to involve redundant genes, particularly kinases, PPI-Hubs and non-Hub bottlenecks, which also display average to low pleiotropy (Figure 3.9, Figure 3.S20). PDS-independent interactions are mainly composed of C1 GIs. They tend to involve essential and pleiotropic genes and also genes involved in cell division (Figure 3.9, Figure 3.S20). Our data also suggest that PDS-centric and PDS-independent interactions coordinate gene functions within-pathways and that two classes of GIs, C3 and C6 are highly involved in this coordination due to their association with connectors/GI-Hubs. These later GI classes were shown to connect genes across ranges of pleiotropy. C6 GIs tend to connect genes with an average to genes with a high pleiotropy while C3 GIs connect genes with an average to genes with a low pleiotropy. We call GI-Hubs involved in C6 and C3 GIs, pleiotropic connectors (PC) and non-pleiotropic connectors respectively (NPC; Figure 3.9; Figure 3.S20). Our study built a structural and functional model for GI-networks in *C. elegans* mainly composed of GIs within signaling and metabolic pathways. This network may be representative for a fraction of the GI network genome-wide (the union of C1-C6 GI corresponds to ~40% of GIs-all). Our data showed that GIs within C7-C10 are composed of both within and between pathway interactions that could not be distinguished from gene-pairs randomly selected from the genome using our classification strategy. Another classification strategy may consequently, be required to characterize these prominent classes of GIs. Nevertheless, our characterization of

within-pathway GI classes raises critical questions relative to the coordination of genes and their protein products within pathways as well as on the role played by the individual GI classes on genomic robustness and on network evolution as discussed below.

3.4.1 Coordination of molecular machines by GIs within-pathways

We showed in this study that PDS and pathways are different structural/functional modules, consistent with their respective positions at abstraction levels III (physical interactions) and IV (functional interactions) (Boucher et Jenna 2013). Furthermore, we showed that both PDS-centric and PDS-independent GIs contribute to pathways. While, the function of PDS and protein complexes within pathways has been extensively studied (reviewed in (Wu, Zhao et Chen 2009)), the function of PDS-independent GIs is quite unexplored. The apparent independence of C1 interactions towards PDS is intriguing. It may be explained by the possibility that physical interactions between protein products of genes involved in C1 GIs have not been identified yet. Interestingly, physical interactions between proteins controlling early embryogenesis, including that controlling cell division have been the subject to special attention (Boxem *et al.* 2008), suggesting that this portion of the PPI network may be less prone to missing data than other parts. This supports the idea that cell division-associated interactions found in C1 define true PDS-independent GIs. PDS-independence of C1 GIs may also result from a potential enrichment of transcriptional regulatory elements in this class (and consequently, protein-DNA interactions instead of PPI). As an example of this possibility, C1 GIs include interactions between the chromatin modifying enzymes coding genes *mes-4*, *mes-2/3/6* and *mep-1*, which were shown to differentially modify histones and consequently, to bind to DNA and not

physically to each other (Strome et Wood 1983) (Figure 3.S6).

Our study also revealed that C3 and C6 GI classes are enriched in GI-Hubs and may play a critical role in coordinating molecular machines within pathways. GI-Hubs have been previously proposed to constitute connectors with high modifier potential – i.e. the ability to modify the expression of phenotypes resulting from genetic alterations of multiple genes – both in *C. elegans* and in yeast (Lehner *et al.* 2006). Our study also characterized C6 GIs as involving pleiotropic connectors (PC, Figure 3.9) and C3 GIs as involving non-pleiotropic connectors (NPC, Figure 3.9). The involvement of PC in the coordination of molecular machines within pathways is consistent with a study of pleiotropic genes *in C. elega*ns suggesting that genes involved in early embryogenesis are organized into partially overlapping functional modules and that pleiotropic genes represent connectors between these modules (Li, Yuan et Zhang 2010). Our data, while supporting this model, suggest that non-pleiotropic genes may also have an organizational role within pathways through C3 interactions.

Intriguingly, genes interacting using C3 and C6 GIs, while being associated individually to several phenotypes did not present any significant correlation in their phenotypic profiles. This is the major distinctive property of C3 and C6 GIs, when compared to other classes. A study investigating the tissue specificity of PPI-network identified both housekeeping/pleiotropic and "local" PPI-Hubs (Bossi et Lehner 2009). It also documented the way housekeeping/pleiotropic Hubs interact with tissue-specific proteins, being consequently involved in different biological processes than other housekeeping protein (Bossi et Lehner 2009). Such a scenario transposed to GI-networks may partially explain how a connector may be associated to a panel of phenotypes which is significantly different than that of their partners.

Overall, our study establishes an organizational model for pathways suggesting that they are built around both pleiotropic PDS-independent and non-pleiotropic PDS-centric GIs coordinating molecular machines together with both pleiotropic and non-pleiotropic connectors. Such an organization suggests that mutations in genes of a

given pathway may lead to either pleiotropic or non-pleiotropic effect. These effects may then be modulated at the organism or at the tissue level, through selection of compensative mutations (Pavlicev et Wagner 2012) or through buffered pleiotropy effect (Fernandez *et al.* 2014). Further characterization of the GI-network structure within and between pathways will thus certainly provide a better understanding of the abundant pleiotropy effect found in human complex diseases and traits (Sivakumaran *et al.* 2011).

3.4.2 GI-network modularity and genomic robustness

C5 and C4 GIs involve genes with an average to a low pleiotropy. This is consistent with their enrichment in genes coding for kinases (Figure 3.2a) and also evolutionarily conserved redundant genes that are mostly kinases in C5 and either kinases or non-kinases in C4. While these genes may be involved in more than one bioprocess, their redundancy may highly limit the number of phenotypes expressed upon perturbation, thus reducing their apparent pleiotropy. While compensatory functions between duplicated genes were shown to be evolutionarily unstable for most of the duplicated gene-pairs (Lynch et Conery 2000), a small fraction of them are submitted to natural selection that stabilizes their functional overlap (Li, Yuan et Zhang 2010). These selected pairs tend to have a high propensity of clustering into the same protein complexes, and share common interaction partners (Li, Yuan et Zhang 2010). Protein complexes were identified form PPI networks as dense subnetworks with methods similar to that used in this study to identify PDS (Chen *et al.* 2014). Altogether, these data are in agreement with C4 and C5 GIs displaying positive values for the CI attribute (gene pairs coding for proteins with a significantly high number of common PPI partners; Figure 3.1b). They are also consistent with C4 beingbeing enriched in

evolutionarily conserved redundant genes. Gene duplication, when associated to functional redundancy, was associated to genomic robustness – i.e. increased resistance of the system to genetic alterations (Woods *et al.* 2013). The PDS-centric within-pathway GIs involving redundant and conserved genes within- or between-PDS may then contribute to the robustness of *C. elegans* genome.

3.4.3 GI-network modularity and evolution

The present study shows that C1 to C6 GIs classes are enriched in within-pathway interactions while C7 to C9 classes are enriched in both within- and between-pathways GIs. Interestingly, these later GI classes could not efficiently be dissociated from gene pairs randomly picked from the genome based on the six attributes used for the classification. While attribute values may be highly influenced by missing data for these GI classes, it is intriguing to observe that the vast majority of between-pathways interactions lay in these classes. It was shown in yeast that both positive and negative GIs within functional modules (protein complexes, gene belonging to the same biological process) are significantly more conserved between *S. cerevisiae* and *S. pombe*, than wiring between these modules (Roguev *et al.* 2008; Dixon, Andrews et Boone 2009; Ryan *et al.* 2012). Considering that a pathway is a functional module, it would then be interesting to assess whether C1 to C6 GIs would be more evolutionarily conserved than C7 to C9. Similarly, it would be interesting to assess whether interactions within-PDS would be more evolutionarily conserved than between or outside PDS interactions.

The organization of the within-pathway GI-network described here is consistent with the evolution theory model called selection, pleiotropy and compensation (SPC) recently built from quantitative genetics studies (reviewed in (Pavlicev et Wagner

2012)). This model predicts that adaptive change in one character (through functional alterations of a pleiotropic gene) is associated with deleterious pleiotropy in others and subsequent selections to compensate for these pleiotropic effects. This compensation involves the genetic alteration of non-pleiotropic/"private" gene(s). This model relies upon the existence of functional interaction between pleiotropic and non-pleiotropic genes within pathways targeted by evolution. Interestingly, such interactions have been identified in tissue-specific PPI-networks (Bossi et Lehner 2009) and are also consistent with the model proposed here, in which pleiotropic Hubs are connected to genes displaying average pleiotropy through C6 GIs within pathways.

We showed that integration of attributes based on data located at different abstraction levels (in this study, levels II, III and VI), identified GI classes with distinctive biological properties from a GI-network. Similar classification strategy applied to gene-pairs found within-pathways of systems inappropriate to a genome-wide experimental mapping of GIs would be interesting to interrogate the evolutionary conservation of functional interactions in multicellular organisms. Integrative genomics approaches using a selected subset of statistical attributes may also be used to improve predictors for specific classes of GIs such as connectors/GI-Hubs, which are of high interest for health-oriented research.

3.5 Conclusions

Our study provides the first deep structural and functional characterization of a GI-network enriched in within-pathways interactions in a multicellular organism. It proposes a model in which PDS-centric and PDS-independent interactions coordinate molecular machines within pathways together with pleiotropic and non-pleiotropic connectors. This study demonstrates the value of integrative genomics approaches,

using data from several abstraction levels to characterize genetic interaction networks, their relationship with networks located at different abstraction levels and to study the systems basis of complex biological phenomena, including genomic robustness, pleiotropic effects and adaptive evolution.

3.6 Methods

3.6.1 Dataset of genetic interactions

The complete set of genetic interactions (GIs-all) consists of 1,514 GIs retrieved from Wormbase, Biogrid and/or curated manually from the literature ([Lee *et al.* 2010] and this study; Table 3.S1, en ligne: [Boucher *et al.* 2016]). Five GIs were curated using the curation and blind re-curation procedure of BIOGRID (Chatr-Aryamontri *et al.* 2015); 689 GIs (45%) were curated by an author-based curation approach used by Wormbase to insure the accuracy of their data (Yook *et al.* 2012); 56 and 261 GIs were manually curated in our laboratory in the context of (Lee *et al.* 2010) and this study respectively (Table 3.S1 [en ligne: (Boucher *et al.* 2016)]). The 261 GIs curated in this study cover GIs identified from the literature by Textpresso (Muller, Kenny et Sternberg 2004) and found in GI classes (C1 to C6) (Table 3.S1 [en ligne: (Boucher *et al.* 2016)]). References to an experimental validation were found in the literature for 252 of these GIs, leaving nine of them without experimental evidences (3.57%). Less than 3.3% of GIs lay in that category of non-validated interactions per class. GIs-all was used to build a GI network analyzed and visualized using Cytoscape v2.8.2 (Shannon *et al.* 2003). GI degree was computed for each gene using the network statistic tool.

3.6.2 Attributes used to cluster GIs into classes

Attributes and data used to compute them have been described previously (Lee *et al.* 2010). Briefly, The co-expression attribute *Exp*(A, B) is the *P*-value derived for the Pearson correlation of genes A and B across 514 microarray experiments (retrieved from (Kim *et al.* 2001)) relative to the empirically estimated probability distribution of correlation for all gene pairs (i.e. a fitted normal). The co-phenotype attribute *Ph*(A, B) uses 107 phenotypes (retrieved from WormBase release WS141) and measures the statistical significance of the number of shared phenotypes between the two genes (A and B) via a standard Fisher's exact test (N = the number of phenotypes observed for at least two genes). The multispecies PPI network was taken from (Lee *et al.* 2010). Briefly, PPIs were obtained from all *C. elegans*, *Saccharomyces cervisiae*, *Drosophila melanogaster*, and *Homo sapiens* yeast two-hybrid datasets stored in BioGRID v2.0.37 (http://www.thebiogrid.org/) and from two additional yeast two-hybrid datasets (Formstecher *et al.* 2005; Stelzl *et al.* 2005). To create a multi-species PPI network, we used the orthology mappings generated by InParanoid v1.35 (O'Brien, Remm et Sonnhammer 2005) (non-default parameters: score cutoff 10, in-paralog confidence cutoff 0.025, sequence overlap cutoff 0.2). The interaction attribute *I*(A, B) indicates whether the proteins encoded by A and B exhibit a PPI in the multi-species PPI network as defined in (Lee *et al.* 2010). Similarly, the common interactors attribute, *CI*(A, B), considers the statistical significance of the observed number of common physical interactors of the proteins encoded by A and B, in the multi-species PPI network. The attribute *CI* is then assigned a *P*-value derived from a one-tailed Fisher's exact test (N = the number of genes encoding proteins that are in the multi-species PPI network). For the other attributes, a biological network called the PhEP was created where two genes A and B are connected by an edge if the Pearson correlation coefficient of their gene expression exceeds 0.35 or if their gene products exhibit a PPI in the multi-species

PPI network. For both A and B, we measured how surprising it is to witness the observed number of their neighbors (i.e. genes connected to it by one edge) in the PhEP network labeled with a specific phenotype identified by RNAi in *C. elegans*. This was measured using a one-tailed Fisher's exact test (N=the number of genes with some assigned phenotype). If the derived *P*-value is less than or equal to 0.05 for A and B for at least one phenotype, we assign a value of 1 to a categorical variable *N*(A, B), and 0 otherwise. Similarly, if A and B exhibit a phenotype that is also enriched in both their neighborhoods in the PhEP network, a value of 1 is assigned to a categorical variable *NPh*(A, B), and 0 otherwise. Missing values for any of the derived attributes (due to missing values in the underlying datasets) were replaced with the expected value (i.e. the sample mean) of the attribute before training. See (Lee *et al.* 2010) for additional information regarding the attributes and datasets used to derive them.

3.6.3 Heatmaps

Unless otherwise specified, heatmaps were the result of hierarchical clustering using Ward's agglomerative method with a distance metric between x and y based on Euclidean distance and is given by:

$$d = \sqrt{\sum_i \left(x_i - y_i\right)^2}$$

The Canberra distance is given by:

$$d_{x,y} = \sum_i \frac{\left|x_i - y_i\right|}{\left|x_i\right| - \left|y_i\right|}$$

where $d_{x,y}$ is the Canberra distance between two GI classes x and y, x_i and y_i is the

number of genes that occur in the classes. The binary distance is defined as the fraction of the n genes that is present in only one of classes x and y.

3.6.4 GI classes identification

GIs-all was hierarchically clustered, and uncertainty in the resulting clusters was judged with approximately unbiased (AU) *P*-values acquired after a multiscale bootstrap resampling of the data (10,000 resamples) using the R package "pvclust" v1.2-2 with the default parameters (Suzuki et Shimodaira 2006). The dendrogram was cut at each height with the R function cutree. Clusters with AU >0.95 were retained for further analysis. For each remaining cluster of genetic GIs (positive set), a logistic regression model was tested using leave-one-out cross-validation (LOOCV) against a negative set (randomly selected gene pairs) of equal size with the requirement that one of the genes found in a negative pair was included in the cluster of genetic interactions. For all clusters, the six attributes were used in the regression models. True and false positive rates were computed at 20 equally spaced model score cutoffs in [0,1], resulting in a receiver operating characteristic (ROC) curve for each model. The area under the ROC curve (AUC) was used as an indicator of how well a classifier model could discriminate GIs found in our positive training sets from negative examples.

3.6.5 Enrichment calculation

Gene Ontology (GO) term, essential gene, redundant genes, PPI-Hubs, GI-Hubs and High-PI genes enrichments were evaluated using a one-tailed Fisher's exact test.

Measurement of enrichment in groups of GIs defined by positive or negative values for a single attribute required the identification of a threshold above which the attribute value is considered as positive. These thresholds are indicated in the Table 3.S6 [en ligne: (Boucher *et al.* 2016)] as well as the number of GIs within GIs-all associated to a positive value to the attribute. For GO term enrichments, the reference universe N was constituted of all terms associated to genes (with or without repetition) found in genetic interactions of GIs-all (see Dataset of genetic interactions section). Note that certain genes are involved in more than one GI within GI classes. As indicated in the result section, GO enrichment was done considering the gene frequency within classes (each repetition being considered as an independent gene) or without considering the gene frequency (each gene is used only once per class and in GIs-all to calculate the enrichment). The universe N for the other enrichment tests contained all genes (with repetition) found in GIs-all.

3.6.6 Monochromaticity index

To evaluate the monochromatic index of each GI class, we first partitioned the GIs-all network into several dense subnetworks using the Cytoscape plugin "MINE" v1.5 (Rhrissorrakrai et Gunsalus 2011) with the default parameter values. The resulting network contained a total of 42 GIs subnetworks (GDS) (S3a Fig). To assess the proportion of interactions from a GI class within a particular GDS, we calculated a monochromatic score (MS) in a similar way than described previously (Szappanos *et al.* 2011). Let B_R represent the ratio of GIs from a given class within GIs-all and M_R, the ratio of GIs from the same GI class within a GDS. The monochromatic scores of a GI class and for a GDS is given by:

$$if\ M_R > B_R, MS = \frac{\left(M_R - B_R\right)}{\left(1 - B_R\right)}$$

$$if\ M_R = B_R, MS = 0$$

$$if\ M_R < B_R, MS = \frac{\left(M_R - B_R\right)}{B_R}$$

3.6.7 Protein-protein interaction (PPI) network analysis

PPI-Hubs were identified as the 20% proteins with the highest PPI-degree ($k \geq 22$) as previously done (Yu *et al.* 2007). The degree and the betweenness centrality were assessed for all genes, Hubs and non-Hubs using the network statistic tool in Cytoscape v2.8.2 (Shannon *et al.* 2003). Distribution of interaction degrees and betweenness centralities was computed for all the genes (considering the frequency of involvement for each gene in GIs of the class) in a given set of GIs.

3.6.8 PPI dense subnetworks and pathways

The modular partitioning of PPI-networks was done using the Cytoscape plugin "MINE" v1.5 (Rhrissorrakrai et Gunsalus 2011) with the default parameter values. Significant PPI dense subnetworks (PDS) were selected by taking all complexes with a score (Density * Number of Proteins) ≥ 4, giving a total of 106 PDS covering 1,760 proteins connected by 17,430 edges. The size distribution of all PDS is given in Figure 3.S14a. Significant GI-dense modules (GI-modules) were selected by taking all

complexes with a score (Density * Number of Proteins) ≥ 3. KEGG pathways were retrieved from the Kyoto Encyclopedia of Genes and Genomes database 61.1 release (Kanehisa *et al.* 2004). Pathways from the literature were manually curated from (Riddle 1997) (Table 3.S3 [en ligne: (Boucher *et al.* 2016)]).

3.6.9 Within- and between- PDS/pathway assessments

To measure the enrichment of within-PDS/pathway and between-PDS/pathway within classes of GIs and pathways, we defined several networks. Networks U_{Ci} are composed of genes and interactions found in GI classes C*i*. A network W_{Pa} is composed of genes found in pathways, which are linked by an edge if these genes are found in at least one common pathway. We also defined a network W_{PDS}, which is composed of proteins and PPI found within PDS. The networks B_{Pa} and B_{PDS} are composed of nodes found in W_{Pa} and W_{PDS} respectively and all possible edges between these nodes from which were removed the edges found respectively in W_{Pa} and W_{PDS}. (note that B_{PDS} do not overlap with the PPI-network). We then computed the frequency *F* in these networks with respect to the mean frequency calculated for a random network *V* as follows:

$$F_{X,V} = \frac{\sum(T_X \cap T_Y)}{\sum T_X} \Big/ \frac{\sum(T_V \cap T_Y)}{\sum T_V}$$

where T_X, T_Y and T_V represent all edges in network *X*, *Y* and *V* respectively.

To compute the frequency of within-PDS and between-PDS interactions found in pathways, we defined the following: X = W_{Pa}, Y = W_{PDS} and B_{PDS} respectively, and V is a random network with the same structure as W_{Pa} (detailed in the Network randomization section below).

To compute the frequency of within-PDS and between-PDS found in GI classes, we defined the following: X = U_{Ci}. Y = W_{PDS} and B_{PDS} respectively, and V is a random network with the same structure as U_{Ci}.

To compute the frequency of within-pathways and between-pathways found in GI classes, we defined the following: X = U_{Ci}, Y = W_{Pa} and B_{Pa} respectively and V is a random network with the same structure as U_{Ci}.

We then computed the $\log_{10}$-ratio (*LR*) transform score of the frequency for network *X* relative to randomized network *V*. To avoid the log-ratio of a zero value, we used a simple transform that took care of any undetermined possibilities as follows:

$$LR\left(F_{X,V}\right)=\begin{cases}\log_{10} \; if \; F_X, F_V > 0 \\ -1 \quad if \; F_X > 0 \,\&\, F_V = 0 \\ 0 \quad if \; F_X = 0 \,\&\, F_V = 0\end{cases}$$

3.6.10 Network randomization

The following randomization procedure was used to randomize GI networks (Figure 3.4 and 3.6; Figure 3.S13) and pathways (Figure 3.5). To do so, for each network being randomized, all connected gene-pairs were split in two groups. The order of genes and the number of edges in the first group were kept unchanged. Genes in the second group were randomly reordered. This aims to preserve the degree of connectivity for any gene present in the network and its randomized version. Restriction was applied to make sure that a pair of gene could not be composed of twice the same gene. The number of randomized networks generated was determined giving Hoeffding's inequality (Hoeffding 1963). Basically, by increasing the number (n) of random networks, we minimize the relative error ρ and get a better estimation of p, the real mean frequency

of edges in within- or between-pathways/PDS. Since the calculated mean μ is an estimation of the real mean p and $\varepsilon = \sqrt{\ln 2 - \ln(1-c)/2n}$, let $\mu^- = \max\{0, \mu - \varepsilon\}$ and $\mu^+ = \max\{1, \mu + \varepsilon\}$ and with probability $c = 0.99$, $\mu^- < p < \mu^+$. And if $\mu^- > 0$, we get with probability c, the maximum value of ρ:

$$\rho = \frac{|\mu - p|}{p} \leq \frac{\max\{|\mu - \mu^-|, |\mu - \mu^+|\}}{\mu^-}$$

For all networks being randomized, 100,000 randomizations were sufficient to obtain a reasonably small value of the relative error ρ with a probability of 0.99. As a consequence, error bars cannot be seen (because they are too small) and are not indicated on bar graphs in Figure 3.4-6.

The second randomization method, used to validate within- and between-PDS relationships (Figure 3.S14), aimed at randomizing all the PDS node labels to create new PDS (Random; Figure 3.S14d) with the exact same topology than that extracted from the PPI network (Original; Figure 3.S14c), but with different node labels. In short, for a given PDS, we permuted the node labels by randomly selecting labels from a list of nodes present in another PDS and not already been reassigned. The procedure was done iteratively until more than 90% of labels in a given PDS were permuted. Edges were unchanged to preserve the degree distribution, PDS size, within- and between-PDS connectivity (as seen in Figure 3.S14c and 3.S14d). After the randomization process, the resulting network contained less than 3% overlapping edges with the original PDS newtork. Note that all PPI used in our study have been generated using yeast -two-hybrid systems which use protein bait to identify preys. However, the bait/prey orientation of PPIs was not considered in the present study.

3.6.11 Distribution pleiotropic indices (PIs)

IDs of observed phenotypes for every gene found in the *C. elegans* genome, and their hierarchical relationships, were downloaded from WormBase (release WS220-bugfix). Relationships between phenotypes were visualized in Cytoscape, where a node represents a phenotype, and an edge between two nodes, the hierarchical relationship between two phenotypes. Groups of phenotype corresponding to the 22 most general phenotypes, covering in 1 step the entire network, were identified (Figure 3.S15). The pleiotropic index (PI) of a gene was computed as the number of these 22 classes containing at least one phenotype associated with the gene. This strategy was used to ensure that we identify the involvement of a gene in different tissues and at different developmental stages without being biased by the extensive characterization of certain developmental stages and/or biological processes. As seen in Figure 3.S15, some groups of phenotypes, for example, "Developmental variant" and "morphology variant" are associated to a much larger number of specific phenotypes than other groups. Several specific phenotypes of highly populated groups may be attributed to individual genes. Our strategy avoids having those genes being pleiotropic if not associated to other phenotypic groups.

Each distribution of PIs was computed from a given set of genes, e.g. all *C. elegans* genes, or genes in a given set of GIs. Odds ratios (*OR*), used to measure the enrichment of GIs between genes within or across certain PI ranges, are defined as:

$$\log_{10}(OR) = \log_{10}\left[\frac{n_{x,i}/n_{all,i}}{N_x/N_{all}}\right]$$

where $n_{x,i}$ = number of class x GIs (e.g., C1 GIs) in subnetwork i (i being for example a subnetwork of GIs between genes with pleiotropic index >8); $n_{all\text{-}i}$ = total number of GIs in subnetwork i; N_x = total number of class x GIs (e.g. C1 GIs) and N_{all} = total

number of GIs in the genetic interactome. Significant enrichments of GI classes in each subnetwork were estimated using a one-tailed Fisher's exact test. Only *OR* associated with a *P*-value <0.05 were represented in Figure 3.S17.

3.6.12 Wilcoxon rank-sum test

Wilcoxon rank-sum test was done according to Hollander and Wolfe (1972) using the wilcoxon.test function in R. This test is used when the population cannot be assumed to be normally distributed.

3.6.13 *P*-values

All computed *P*-values were adjusted using the Benjamini and Hochberg (1995) method for controlling the false discovery rate (specifically, with the p.adjust function in R).

3.7 References

Ashburner, M., C. A. Ball, et al. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25(1): 25-29.

Bellay, J., G. Atluri, et al. (2011). Putting genetic interactions in context through a global modular decomposition. *Genome Res* 21(8): 1375-1387.

Bossi, A. and B. Lehner (2009). Tissue specificity and the human protein interaction network. *Mol Syst Biol* 5: 260.

Boucher, B. and S. Jenna (2013). Genetic interaction networks: better understand to better predict. *Front Genet* 4: 290.

Boucher, B., A. Y. Lee, et al. (2016). Structural and Functional Characterization of a Caenorhabditis elegans Genetic Interaction Network within Pathways. *PLoS Comput Biol*. En ligne: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004738

Boxem, M., Z. Maliga, et al. (2008). A protein domain-based interactome network for C. elegans early embryogenesis. *Cell* 134(3): 534-545.

Chang, X., T. Xu, et al. (2013). Dynamic modular architecture of protein-protein interaction networks beyond the dichotomy of 'date' and 'party' hubs. *Sci Rep* 3: 1691.

Chatr-Aryamontri, A., B. J. Breitkreutz, et al. (2015). The BioGRID interaction database: 2015 update. *Nucleic Acids Res* 43(Database issue): D470-478.

Chen, B., W. Fan, et al. (2014). Identifying protein complexes and functional modules--from static PPI networks to dynamic PPI networks. *Brief Bioinform* 15(2): 177-194.

Collins, S. R., K. M. Miller, et al. (2007). Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature* 446(7137): 806-810.

Costanzo, M., A. Baryshnikova, et al. (2010). The genetic landscape of a cell. *Science* 327(5964): 425-431.

Davierwala, A. P., J. Haynes, et al. (2005). The synthetic genetic interaction spectrum of essential genes. *Nat Genet* 37(10): 1147-1152.

Dixon, S. J., B. J. Andrews, et al. (2009). Exploring the conservation of synthetic lethal genetic interaction networks. *Commun Integr Biol* 2(2): 78-81.

Dixon, S. J., M. Costanzo, et al. (2009). Systematic mapping of genetic interaction networks. *Annu Rev Genet* 43: 601-625.

Drees, B. L., V. Thorsson, et al. (2005). Derivation of genetic interaction networks from quantitative phenotype data. *Genome Biol* 6(4): R38.

Estevez, M., L. Attisano, et al. (1993). The daf-4 gene encodes a bone morphogenetic protein receptor controlling C. elegans dauer larva development. *Nature* 365(6447): 644-649.

Fernandez, A. G., E. K. Mis, et al. (2014). Uncovering buffered pleiotropy: a genome-scale screen for mel-28 genetic interactors in Caenorhabditis elegans. *G3 (Bethesda)* 4(1): 185-196.

Formstecher, E., S. Aresta, et al. (2005). Protein interaction mapping: a Drosophila case study. *Genome research* 15(3): 376-384.

Georgi, L. L., P. S. Albert, et al. (1990). daf-1, a C. elegans gene controlling dauer larva development, encodes a novel receptor protein kinase. *Cell* 61(4): 635-645.

Gibson, G. (2010). Hints of hidden heritability in GWAS. *Nat Genet* 42(7): 558-560.

Gunther, C. V., L. L. Georgi, et al. (2000). A Caenorhabditis elegans type I TGF beta receptor can function in the absence of type II kinase to promote larval development. *Development* 127(15): 3337-3347.

Hoeffding, W. (1963). Probability Inequalities for Sums of Bounded Random Variables. *Journal of the American Statistical Association* 58(301): 13-30.

Jaimovich, A., R. Rinott, et al. (2010). Modularity and directionality in genetic interaction maps. *Bioinformatics* 26(12): i228-236.

Kamath, R. S., A. G. Fraser, et al. (2003). Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature* 421(6920): 231-237.

Kanehisa, M., S. Goto, et al. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res* 32(Database issue): D277-280.

Kelley, R. and T. Ideker (2005). Systematic interpretation of genetic interactions using protein networks. *Nat Biotechnol* 23(5): 561-566.

Kemp, C. A., K. R. Kopish, et al. (2004). Centrosome maturation and duplication in C. elegans require the coiled-coil protein SPD-2. *Dev Cell* 6(4): 511-523.

Kim, S. K., J. Lund, et al. (2001). A gene expression map for Caenorhabditis elegans. *Science* 293(5537): 2087-2092.

Lee, A. Y., R. Perreault, et al. (2010). Searching for signaling balance through the identification of genetic interactors of the Rab guanine-nucleotide dissociation inhibitor gdi-1. *PLoS ONE* 5(5).

Lee, I., B. Lehner, et al. (2010). Predicting genetic modifier loci using functional gene networks. *Genome Res* 20(8): 1143-1153.

Lehner, B. (2007). Modelling genotype-phenotype relationships and human disease with genetic interaction networks. *J Exp Biol* 210(Pt 9): 1559-1566.

Lehner, B., C. Crombie, et al. (2006). Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways. *Nat Genet* 38(8): 896-903.

Li, J., Z. Yuan, et al. (2010). The cellular robustness by genetic redundancy in budding yeast. *PLoS Genet* 6(11): e1001187.

Lundquist, E. A., P. W. Reddien, et al. (2001). Three C. elegans Rac proteins and several alternative Rac regulators control axon guidance, cell migration and apoptotic cell phagocytosis. *Development* 128(22): 4475-4488.

Lynch, M. and J. S. Conery (2000). The evolutionary fate and consequences of duplicate genes. *Science* 290(5494): 1151-1155.

McNeill, H. and J. R. Woodgett (2010). When pathways collide: collaboration and connivance among signalling proteins in development. *Nat Rev Mol Cell Biol* 11(6): 404-413.

Michaut, M., A. Baryshnikova, et al. (2011). Protein complexes are central in the yeast genetic landscape. *PLoS Comput Biol* 7(2): e1001092.

Muller, H. M., E. E. Kenny, et al. (2004). Textpresso: an ontology-based information retrieval and extraction system for biological literature. *PLoS Biol* 2(11): e309.

O'Brien, K. P., M. Remm, et al. (2005). Inparanoid: a comprehensive database of eukaryotic orthologs. *Nucleic Acids Research* 33(Database issue): D476-480.

Ozier, O., N. Amin, et al. (2003). Global architecture of genetic interactions on the protein network. *Nat Biotechnol* 21(5): 490-491.

Pavlicev, M. and G. P. Wagner (2012). A model of developmental evolution: selection, pleiotropy and compensation. *Trends Ecol Evol* 27(6): 316-322.

Pu, S., K. Ronen, et al. (2008). Local coherence in genetic interaction patterns reveals prevalent functional versatility. *Bioinformatics* 24(20): 2376-2383.

Rhrissorrakrai, K. and K. C. Gunsalus (2011). MINE: Module Identification in Networks. *BMC Bioinformatics* 12: 192.

Riddle, D. L. (1997). *C. elegans II. 2nd edition.* Cold Spring Harbor (NY), Cold Spring Harbor Laboratory Press.

Roguev, A., S. Bandyopadhyay, et al. (2008). Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. *Science* 322(5900): 405-410.

Ryan, C. J., A. Roguev, et al. (2012). Hierarchical modularity and the evolution of genetic interactomes across species. *Molecular cell* 46(5): 691-704.

Schuldiner, M., S. R. Collins, et al. (2005). Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell* 123(3): 507-519.

Segre, D., A. Deluna, et al. (2005). Modular epistasis in yeast metabolism. *Nat Genet* 37(1): 77-83.

Shannon, P., A. Markiel, et al. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13(11): 2498-2504.

Sharifpoor, S., D. van Dyk, et al. (2012). Functional wiring of the yeast kinome revealed by global analysis of genetic network motifs. *Genome Res* 22(4): 791-801.

Sivakumaran, S., F. Agakov, et al. (2011). Abundant pleiotropy in human complex diseases and traits. *Am J Hum Genet* 89(5): 607-618.

Sonnichsen, B., L. B. Koski, et al. (2005). Full-genome RNAi profiling of early embryogenesis in Caenorhabditis elegans. *Nature* 434(7032): 462-469.

St Onge, R. P., R. Mani, et al. (2007). Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions. *Nat Genet* 39(2): 199-206.

Stelzl, U., U. Worm, et al. (2005). A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 122(6): 957-968.

Strome, S. and W. B. Wood (1983). Generation of asymmetry and segregation of germ-line granules in early C. elegans embryos. *Cell* 35(1): 15-25.

Suzuki, R. and H. Shimodaira (2006). Pvclust: an R package for assessing the uncertainty in hierarchical clustering. *Bioinformatics* 22(12): 1540-1542.

Szappanos, B., K. Kovacs, et al. (2011). An integrated approach to characterize genetic interaction networks in yeast metabolism. *Nat Genet* 43(7): 656-662.

Tischler, J., B. Lehner, et al. (2006). Combinatorial RNA interference in Caenorhabditis elegans reveals that redundancy between gene duplicates can be maintained for more than 80 million years of evolution. *Genome Biol* 7(8): R69.

Tong, A. H., M. Evangelista, et al. (2001). Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* 294(5550): 2364-2368.

Tong, A. H., G. Lesage, et al. (2004). Global mapping of the yeast genetic interaction network. *Science* 303(5659): 808-813.

Ulitsky, I. and R. Shamir (2007). Identification of functional modules using network topology and high-throughput data. *BMC Systems Biology* 1(1): 8.

Woods, S., A. Coghlan, et al. (2013). Duplication and retention biases of essential and non-essential genes revealed by systematic knockdown analyses. *PLoS Genet* 9(5): e1003330.

Wu, Z., X. Zhao, et al. (2009). Identifying responsive functional modules from protein-protein interaction network. *Mol Cells* 27(3): 271-277.

Yook, K., T. W. Harris, et al. (2012). WormBase 2012: more genomes, more data, new website. *Nucleic Acids Res* 40(Database issue): D735-741.

Yu, H., P. M. Kim, et al. (2007). The importance of bottlenecks in protein networks: correlation with gene essentiality and expression dynamics. *PLoS Comput Biol* 3(4): e59.

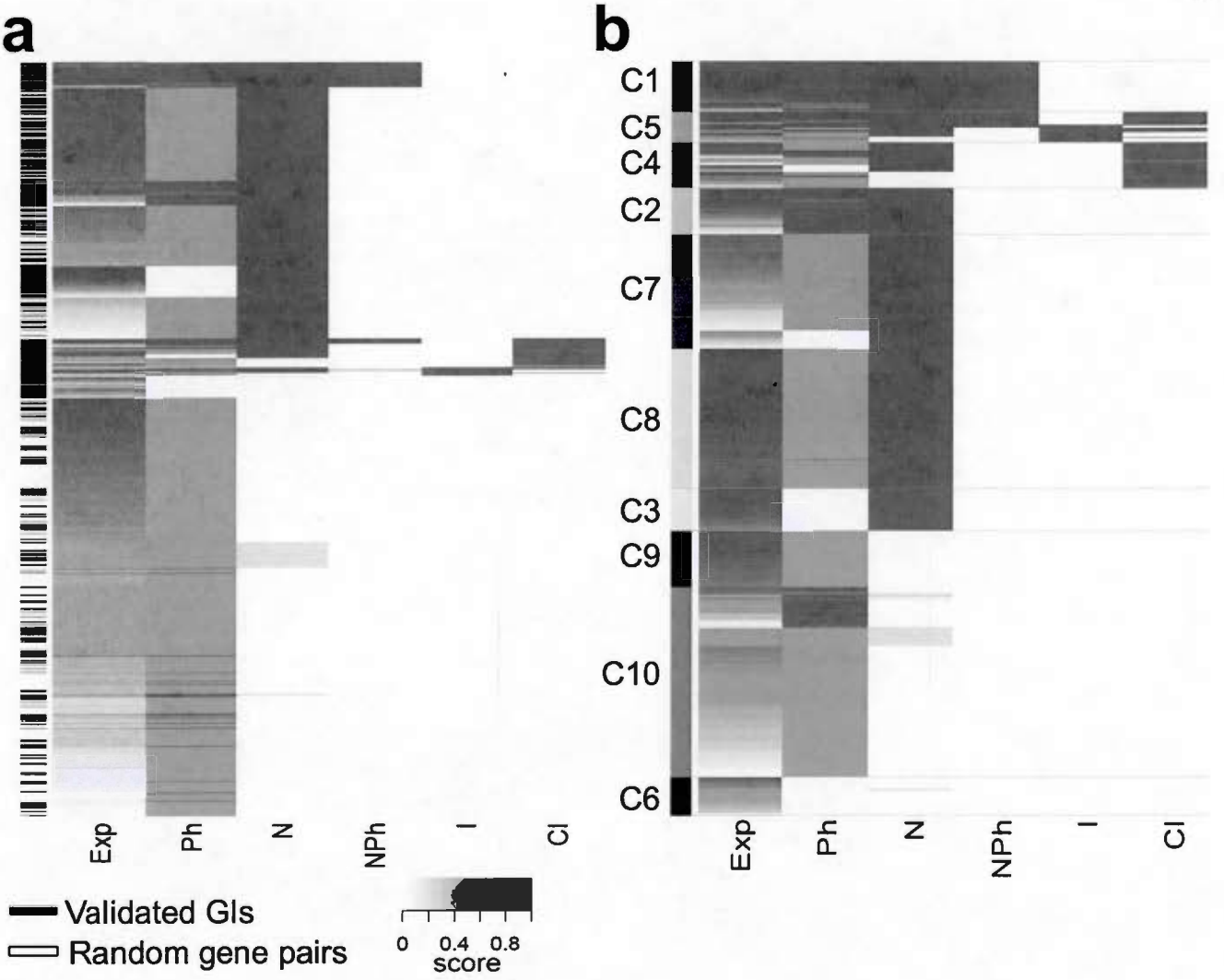


Figure 3.1 Identifying GI classes.

(a) Positive (GIs, black lines) and negative (Random gene pairs, white lines) examples of genetic interactions were clustered based on their attribute scores using unsupervised methods. Columns show values for the six attributes used to predict interactions in (Lee et al. 2010). Each individual attribute is either a measure of genes co-expression levels (Exp) or enrichment of shared phenotypes (Ph). They are also indicator for whether the neighborhoods of the genes of interest are enriched with the same phenotype (N). Here we define the neighborhood of a given gene as the set of genes that show significant co-expression with it ($P \leq 0.05$, see Methods) and/or encode proteins that exhibit a PPI with the product of this gene. NPh is an indicator like N with the additional requirement that the genes of interest themselves must also exhibit the phenotype enriched in their neighborhoods. Attributes I and CI indicate that interacting genes code for interacting or for proteins sharing a significantly high number of common protein-protein interaction partners. Scores correspond to the following valuation: 1 or 0 (on a binary system) for N, NPh and I attributes. (1 – (P-value <0.05)) for Exp, Ph and CI attributes. **(b)** Positive examples of genetic interactions were clustered and identified by the color code on the left.

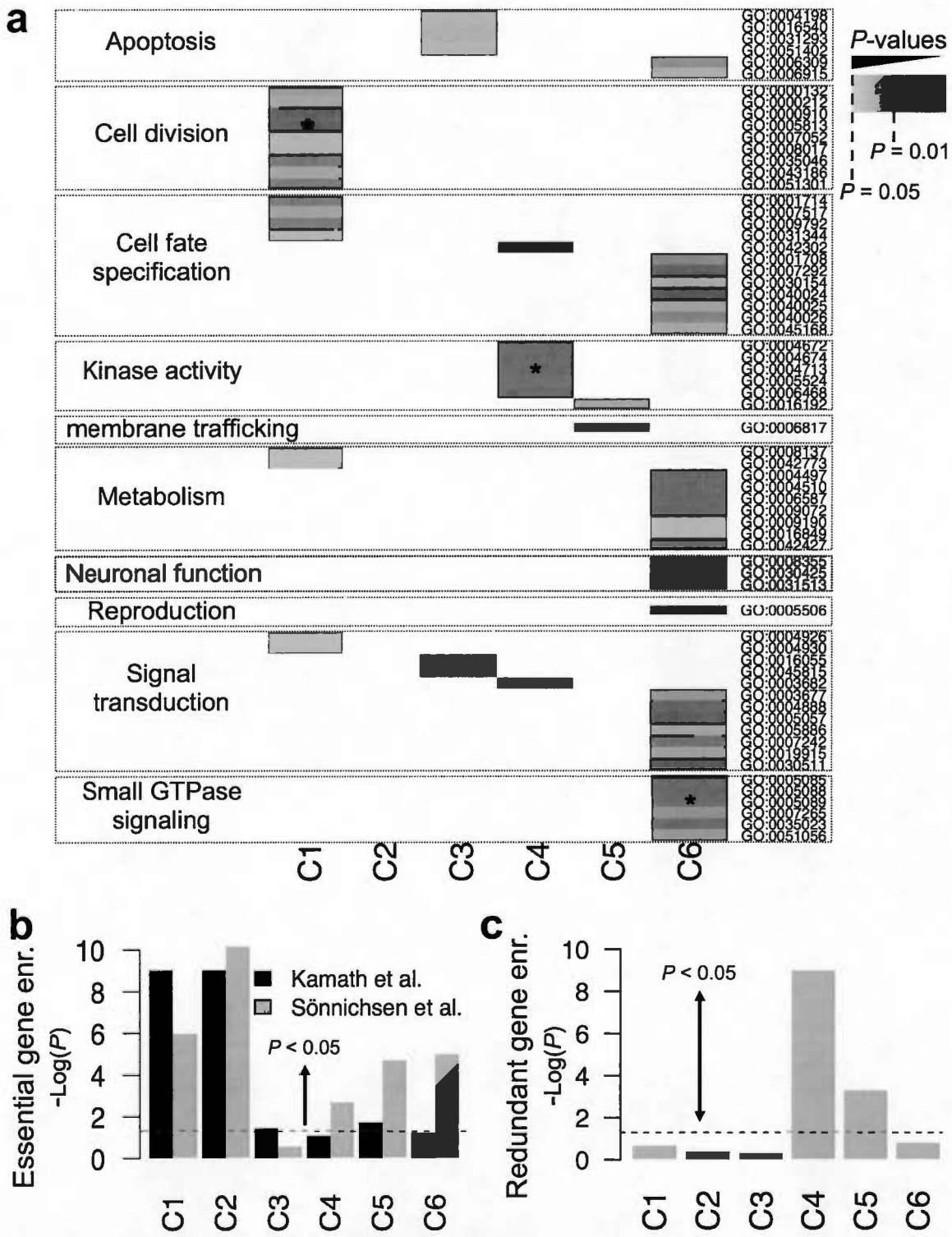


Figure 3.2 GI classes have different functions.

(a) Summary of Gene Ontology (GO) term enrichments for interacting genes in individual genetic interactions (GIs) classes when compared to interacting genes from GIs-all. Only the statistically significant enrichments are shown with adjusted P-values. An asterisk (*) means that the GI class is also enriched with the functional annotation when gene repetitions were not considered (Table 3.S2 [en ligne: (Boucher et al. 2016)]). **(b)** Enrichments of essential gene sets (Kamath et al. and Soninchsen et al. taken from (Kamath et al. 2003) and (Sonnichsen et al. 2005)) in GI classes when compared to GIs-all. (c) Enrichment of redundant and evolutionarily conserved genes in GI classes when compared to GIs-all. The redundant gene list was made of 306 genes taken from (Tischler et al. 2006). **(b-c)** -Log10 of P-values obtained using Fisher's tests are indicated. The area over the red dashed line indicates significant enrichment ($P < 0.05$).

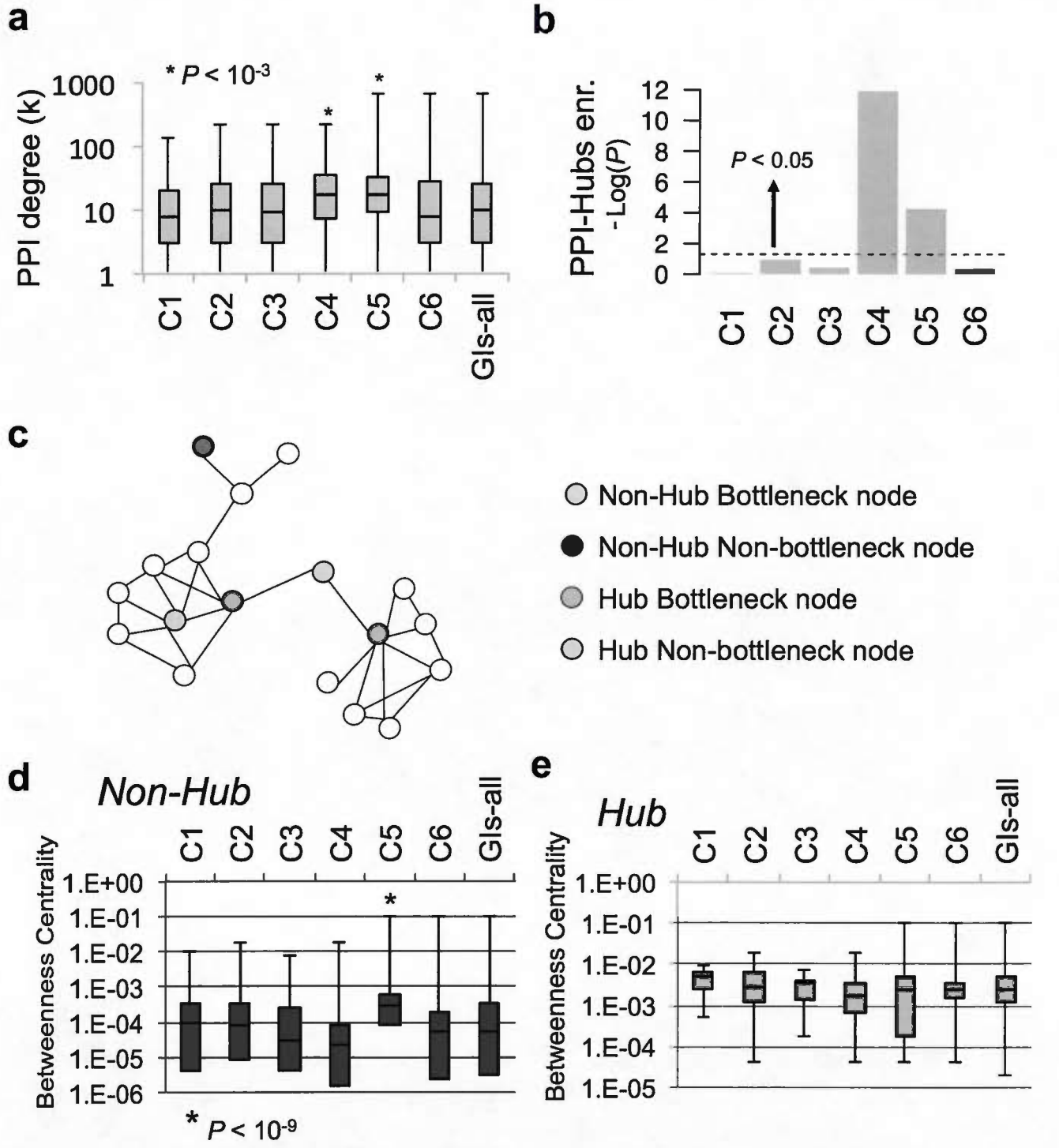


Figure 3.3 GI classes have different position in PPI-network.

(a) Physical interaction degree of gene products in the protein-protein interaction (PPI) network (only genes coding for proteins with at least one interacting partner were considered). *P < 10-3, Wilcoxon rank-sum test, indicates a significant difference when compared to GIs-all and other GI classes. **(b)** Enrichment of PPI-Hubs in GI classes when compared to GIs-all. - Log10 of P-values obtained using Fisher's tests are indicated. The area over the red dashed line indicates significant enrichment (P < 0.05). **(c)** Illustration of bottlenecks and the different type of nodes in the protein-protein interaction (PPI) network (adapted from (Yu et al. 2007)). **(d-e)** Betweenness centrality of non-Hub and Hub gene products. *P < 10-3, Wilcoxon rank-sum test, indicates significant difference when compared to other GI classes. The box plots **(a, d-e)** represent the min, max, 25th, 50th (median) and 75th percentile of either PPI Degree (k) or betweenness centrality values.

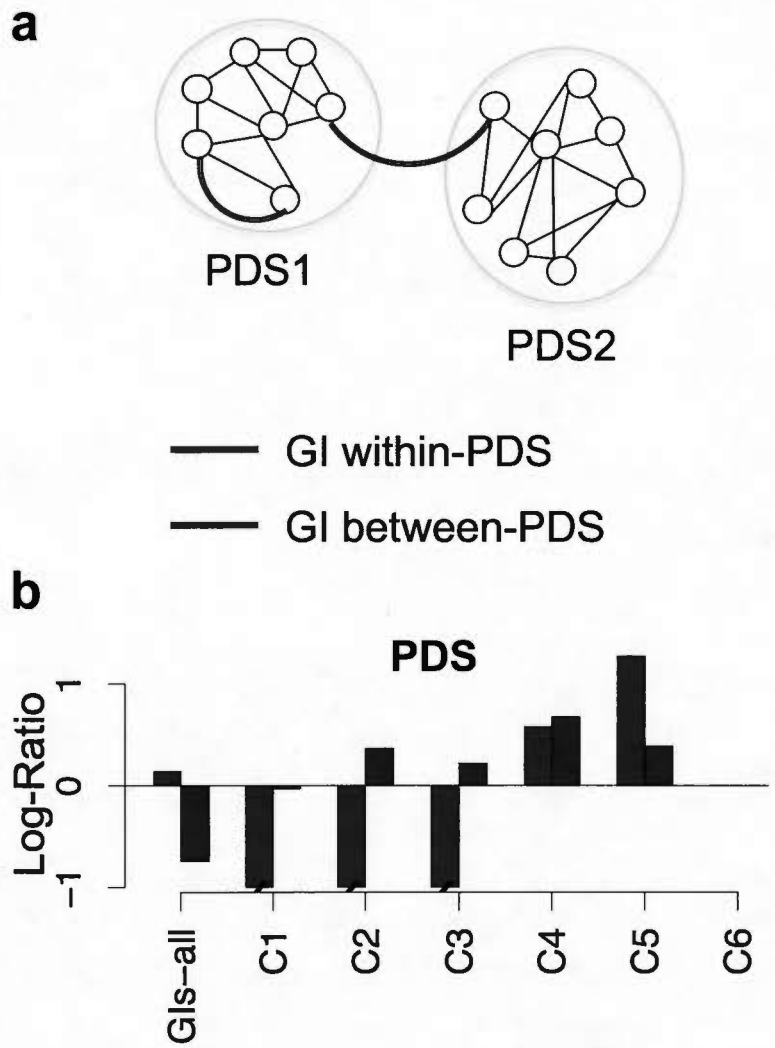


Figure 3.4 GI classes define PDS-depend and -independent modules.

(a) Schematic representation of possible relationships between genetic interactions (GIs) and protein-protein dense subnetworks (PDS). Each node represents a gene being part of PDS1 or PDS2. A red line indicates a within-PDS GI and a blue line, a between-PDS GI. **(b)** Log-Ratio scores for within-PDS (red bars) and between-PDS (blue bars) relationships observed between genes interacting through different GI classes, or present in GIs-all. A positive Log-Ratio score means that the frequency of within- or between-PDS GIs occurring in the GI class is significantly higher than the frequency of similar situations witnessed in relevant randomized GI networks with a probability of 0.99 (see Methods).

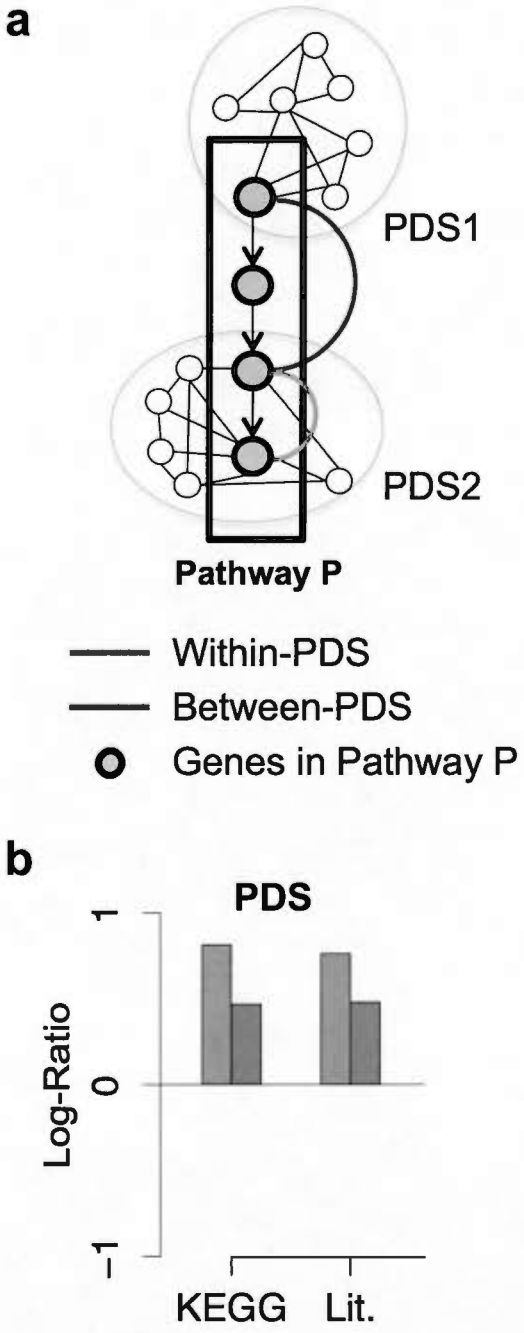


Figure 3.5 PDS and pathways are different functional modules.

(a) Schematic representation of possible relationships between Pathways and protein-protein interaction dense subnetwork (PDS). Each node represents a gene product connected by a PPI (black lines) or by a functional relationship within a pathway P (back arrows). The grey nodes represent gene products involved in a pathway P. An orange line linking two grey nodes indicates that these two genes/proteins interact within-pathway and also within-PDS. A green line linking two grey nodes indicates that these two genes/proteins interact within-pathway and between-PDS. **(b)** Log-Ratio scores for within-PDS and between-PDS relationships occurring within-pathways. A positive Log-Ratio score means that the frequency of within-pathway relationships occurring also within-PDS (orange bars) or within-pathway relationships occurring between-PDS (green bars) is higher than the frequency of similar situations witnessed in relevant randomized pathway networks.

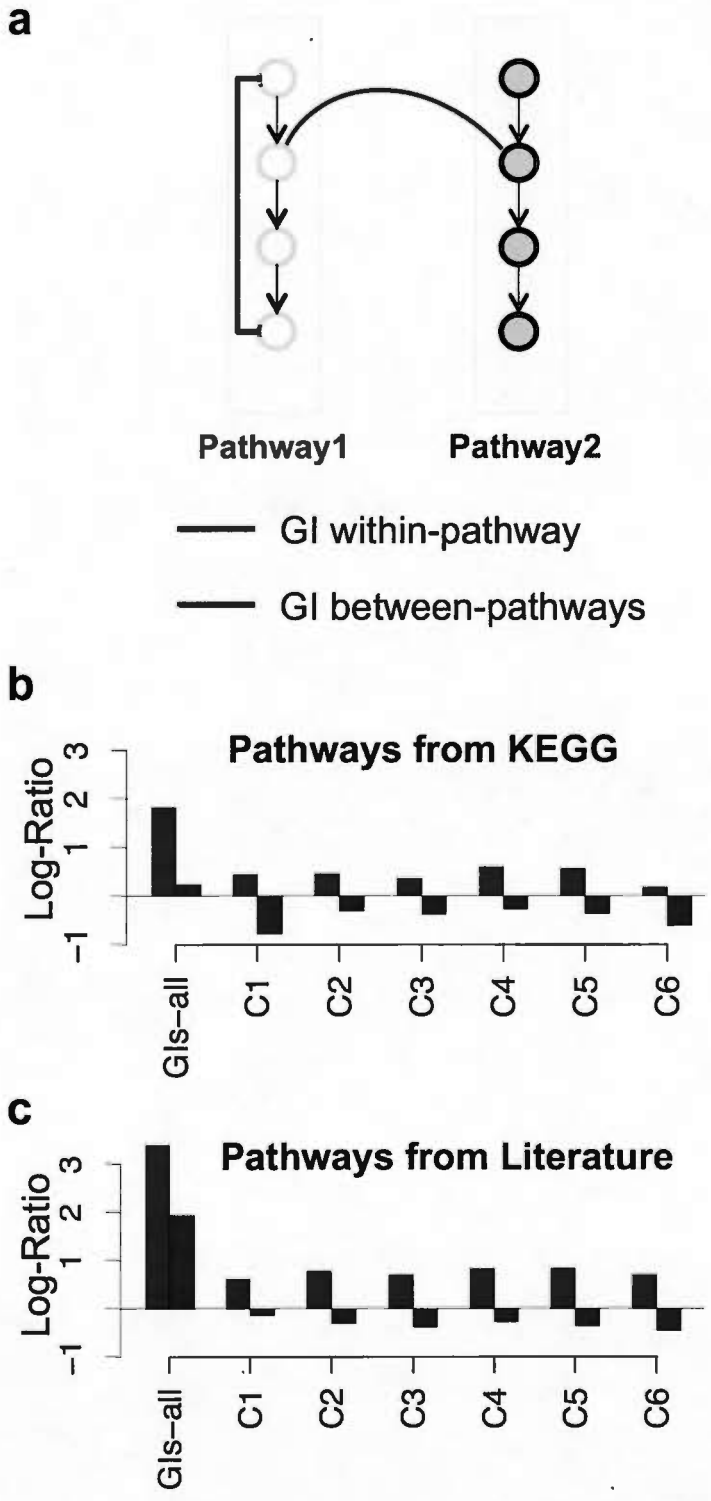


Figure 3.6 GI classes are enriched in within-pathways interactions.

(a) Schematic representation of possible relationships between genetic interactions (GIs) and pathways. Each node represents a gene being part of pathways 1 or 2. A red line indicates a within-pathway GI and a blue line, a between-pathway GI. **(b-c)** Log-Ratio scores for within-pathway (red bars) and between-pathway (blue bars) relationships observed between genes interacting through different GI classes, or present in GIs-all. A positive Log-Ratio score means that the frequency of within- or between-pathways GIs occurring in the GI class is significantly higher than the frequency of similar situations witnessed in relevant randomized GI networks with a probability of 0.99 (see Methods).

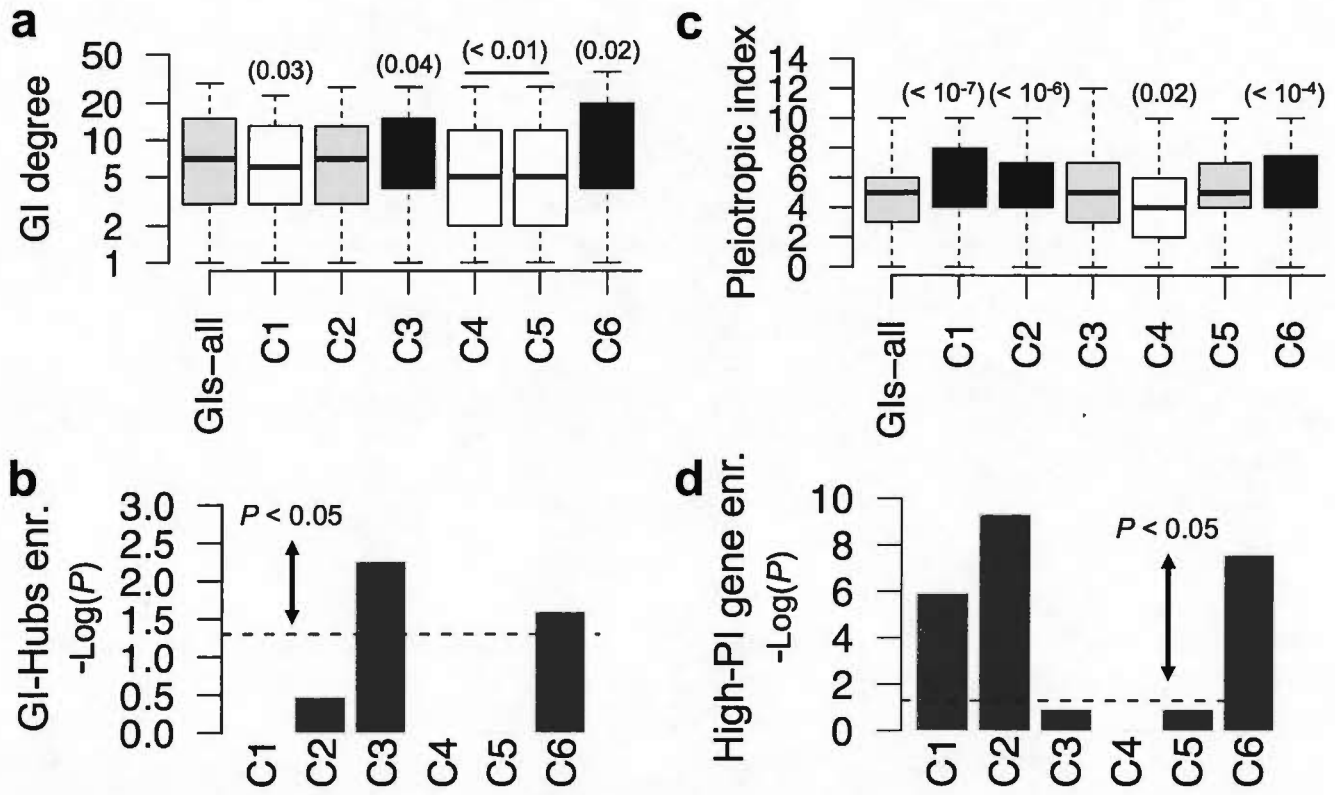


Figure 3.7 Defining pleiotropic and non-pleiotropic connectors.

(a) Box plots showing distributions of genetic interaction (GI) degrees for GIs-all and the six classes of GIs. Distributions were found to be significantly higher (Dark grey boxes) or lower (white boxes) than GIs-all ($P < 0.05$, Wilcoxon rank-sum test). Light grey boxes indicate classes with GI degree not significantly different than GIs-all. **(b)** Enrichment in GI-Hubs (defined as the 20% most connected genes in GIs-all) for GI classes when compared to GIs-all. **(c)** Distribution of pleiotropic indices (PIs) for GIs-all and the six classes of GIs. Dark grey, white and light grey boxes are designated as in panel (a). **(d)** Enrichment in High-PI (defined as the top 20% of genes with the highest PIs genome-wide) for GI classes when compared to GIs-all. The box plots **(a, c)** represent the min, max, 25th, 50th (median) and 75th percentile of either GI Degree (a) or pleiotropic indices (c). **(b, d)** -Log10 of P-values obtained using Fisher's tests are indicated. The area over the red dashed line indicates significant enrichment ($P < 0.05$).

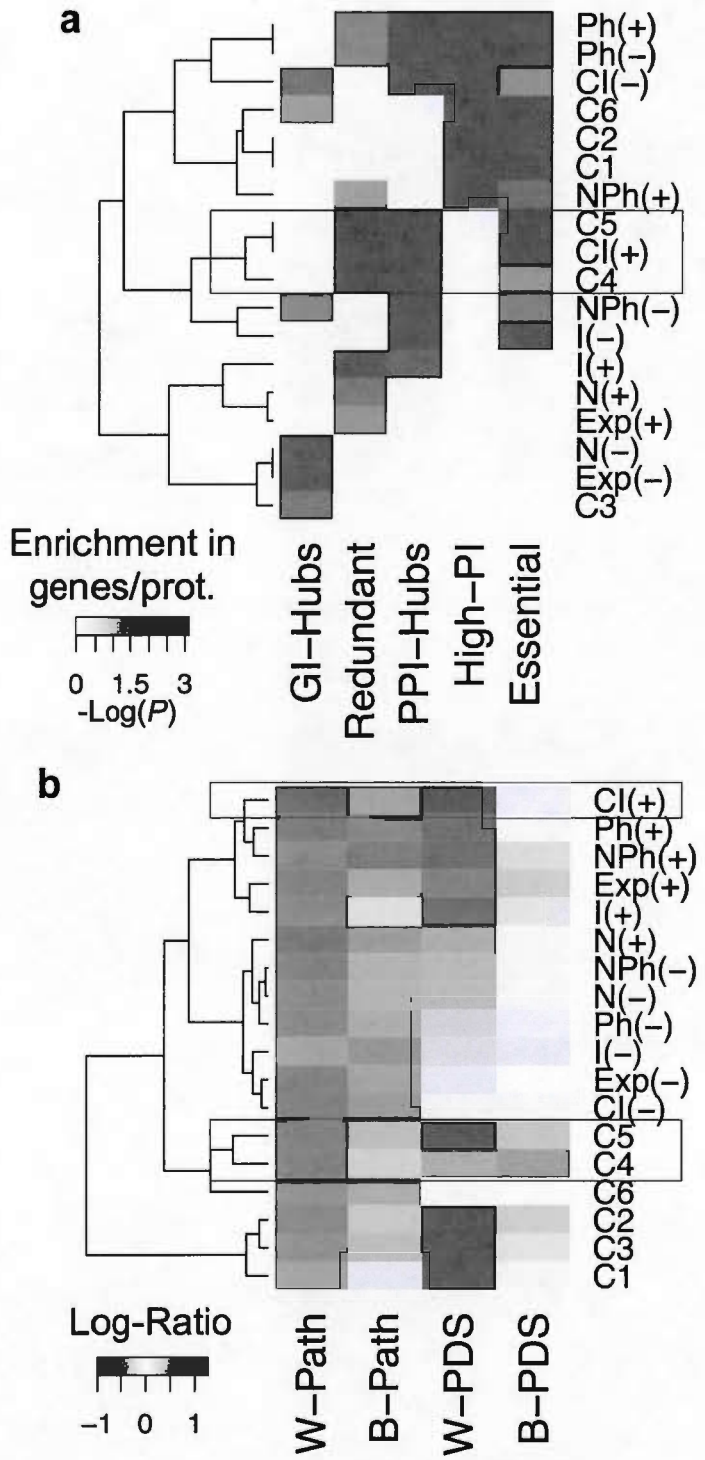


Figure 3.8 Data integration is required to identify GI classes.

Hierarchical clustering of **(a)** Enrichment in GI classes and GI groups of Genetic interaction-Hubs (GI-Hubs), redundant genes, protein-protein interaction-Hubs (PPI-Hubs), Highly pleiotropic genes (High-PI) and essential genes -log of P-values from Fisher's exact test are indicated. GI groups are associated to a positive value (+; above a threshold) or a negative value (-; below the threshold) for indicated attributes. Threshold values for each attributes are indicated Table 3.S6 [en ligne: (Boucher et al. 2016)]. **(b)** Log-Ratios profiles for GI classes and GI groups associated to a positive (+) or a negative (-) values for indicated attributes. Blue boxes indicate enrichment of biological characteristics for C4 and C5 GI classes and CI(+) GI group.

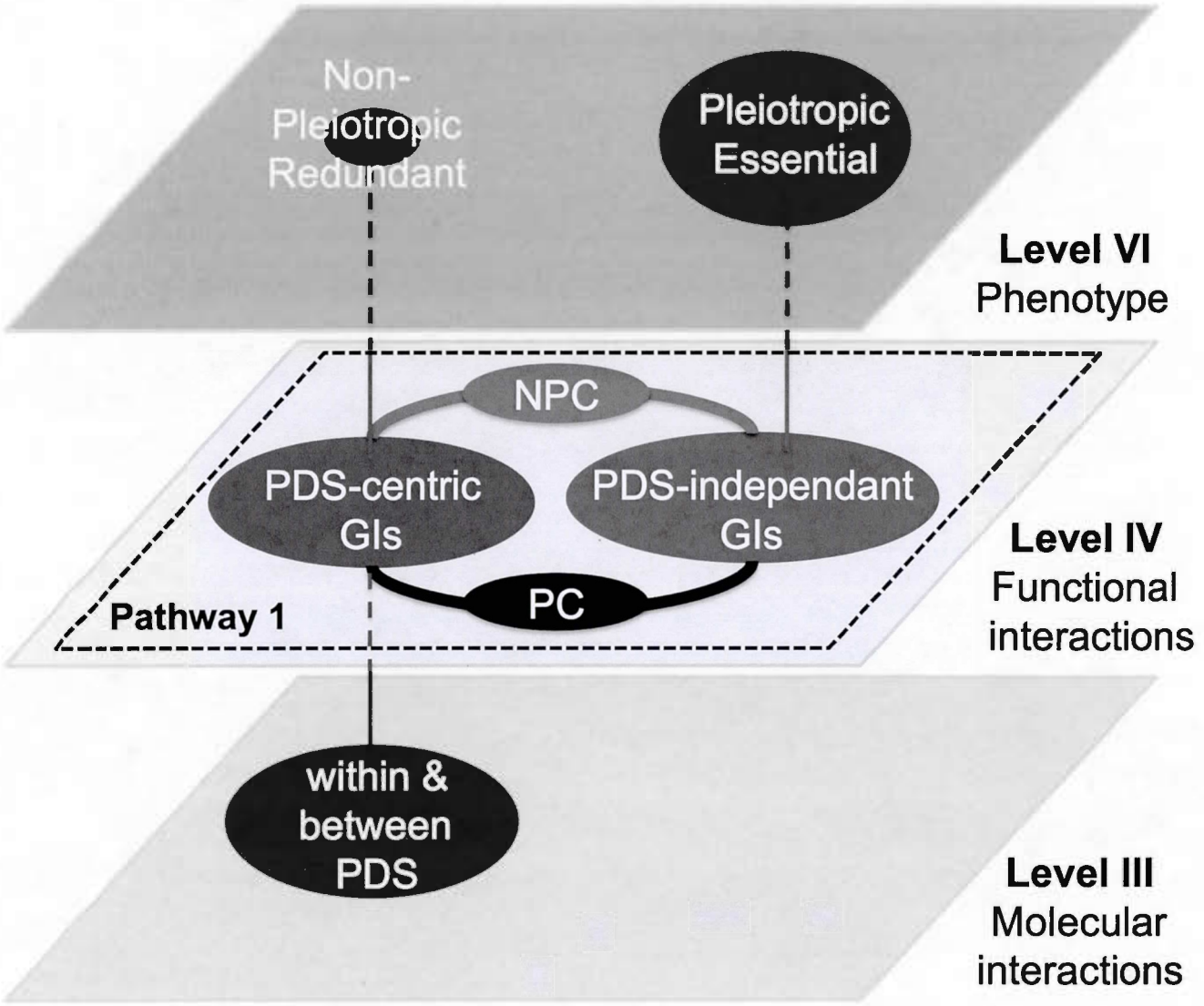


Figure 3.9 Structure and function of a GI-network within-pathway in Caenorhabditis elegans.

Three abstraction levels are represented as plan from the bottom to top: Level III, level IV and level VI. PDS-centric interactions (C4 and C5 classes) are organized around Protein-protein interaction dense subnetworks (PDS) located at level III. They tend to involve non-pleiotropic and redundant genes (level VI). PDS-independent interactions (C1 and C2 classes) tend to involve pleiotropic and essential genes. GIs involving pleiotropic (PC; C6 class) and non-pleiotropic connectors/GI-Hubs (NPC; C3 class). They tend to link genes from an average to a high range of pleiotropy and to an average to a low range of pleiotropy respectively.

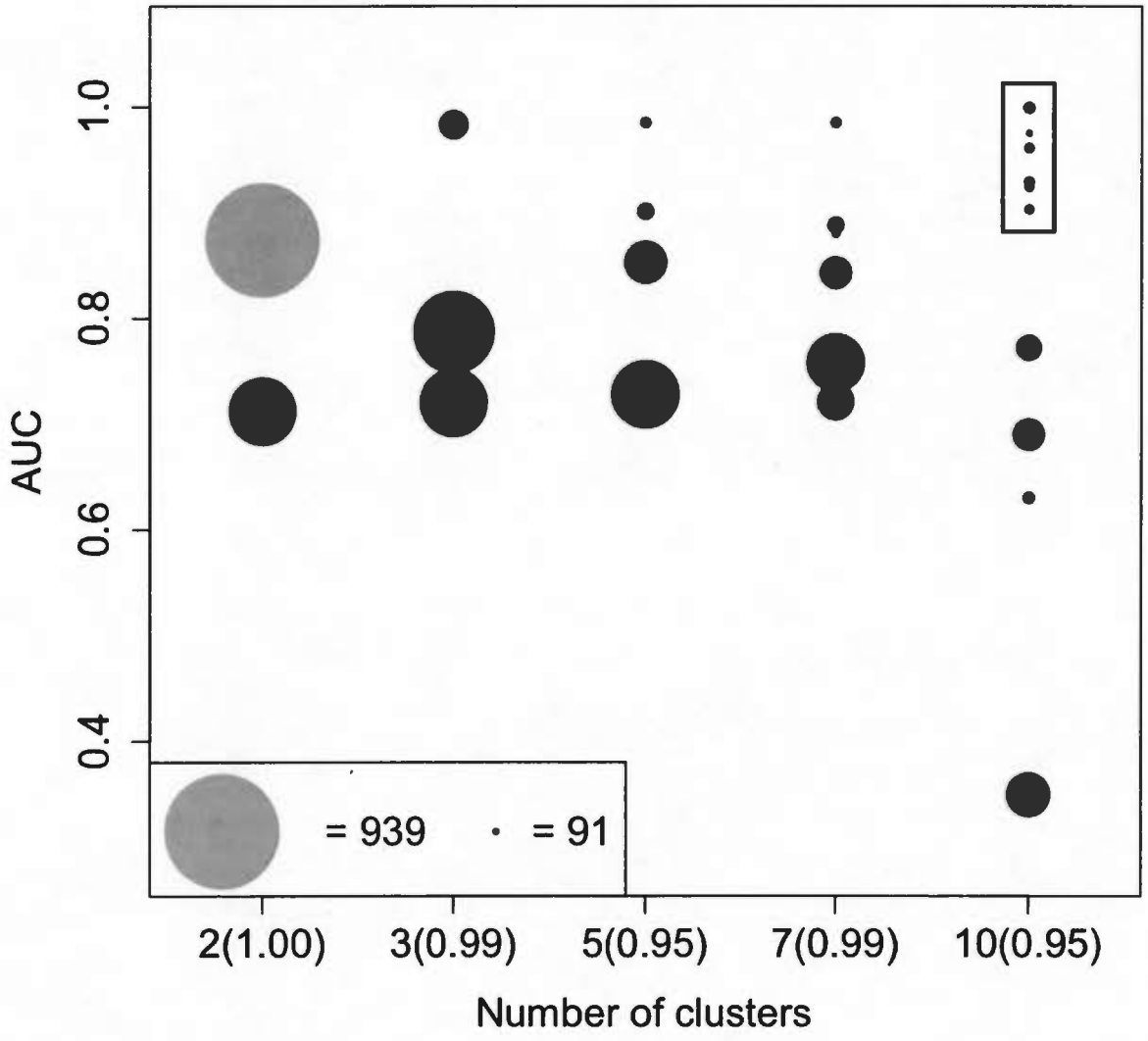


Figure 3.S1 Cluster selection within GIs-all.

The x-axis shows the number of clusters with approximately unbiased P-value (higher values indicate greater significance, see Methods) greater than or equal to the threshold indicated in the parentheses. The y-axis shows area under the curve (AUC) values following cross-validation analysis of each cluster-based model. Each dot represents one cluster and its magnitude represents the number of GIs within the cluster. Dots inside the red rectangle correspond to the 6 clusters selected for further analysis based on their high AUC values and comparable sizes.

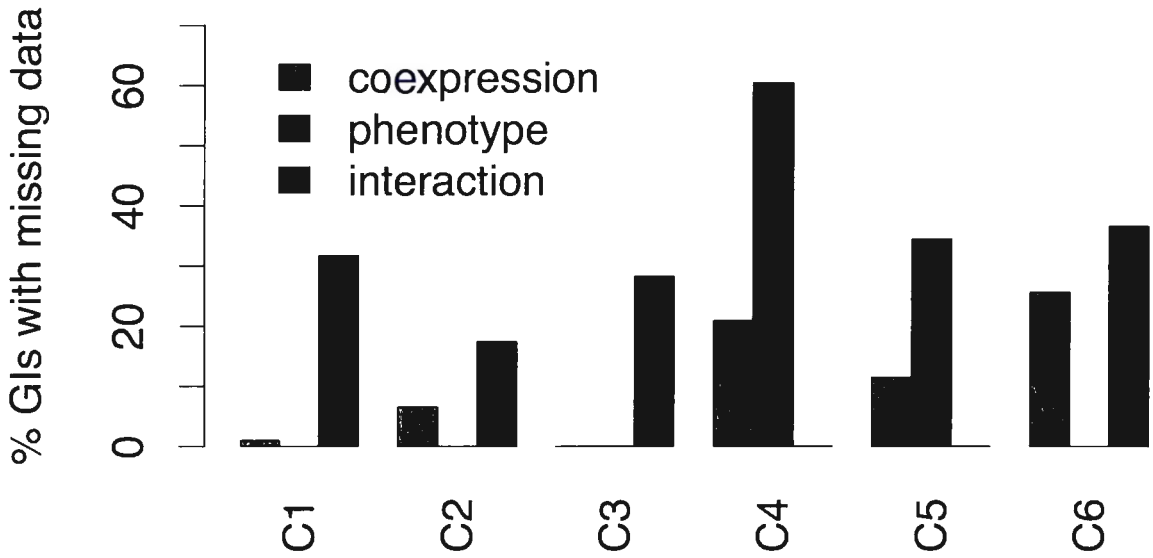


Figure 3.S2 Genetic interaction classes and associated missing data.

Percentage of interacting gene pairs missing co-expression, phenotype or protein-protein interaction data are indicated.

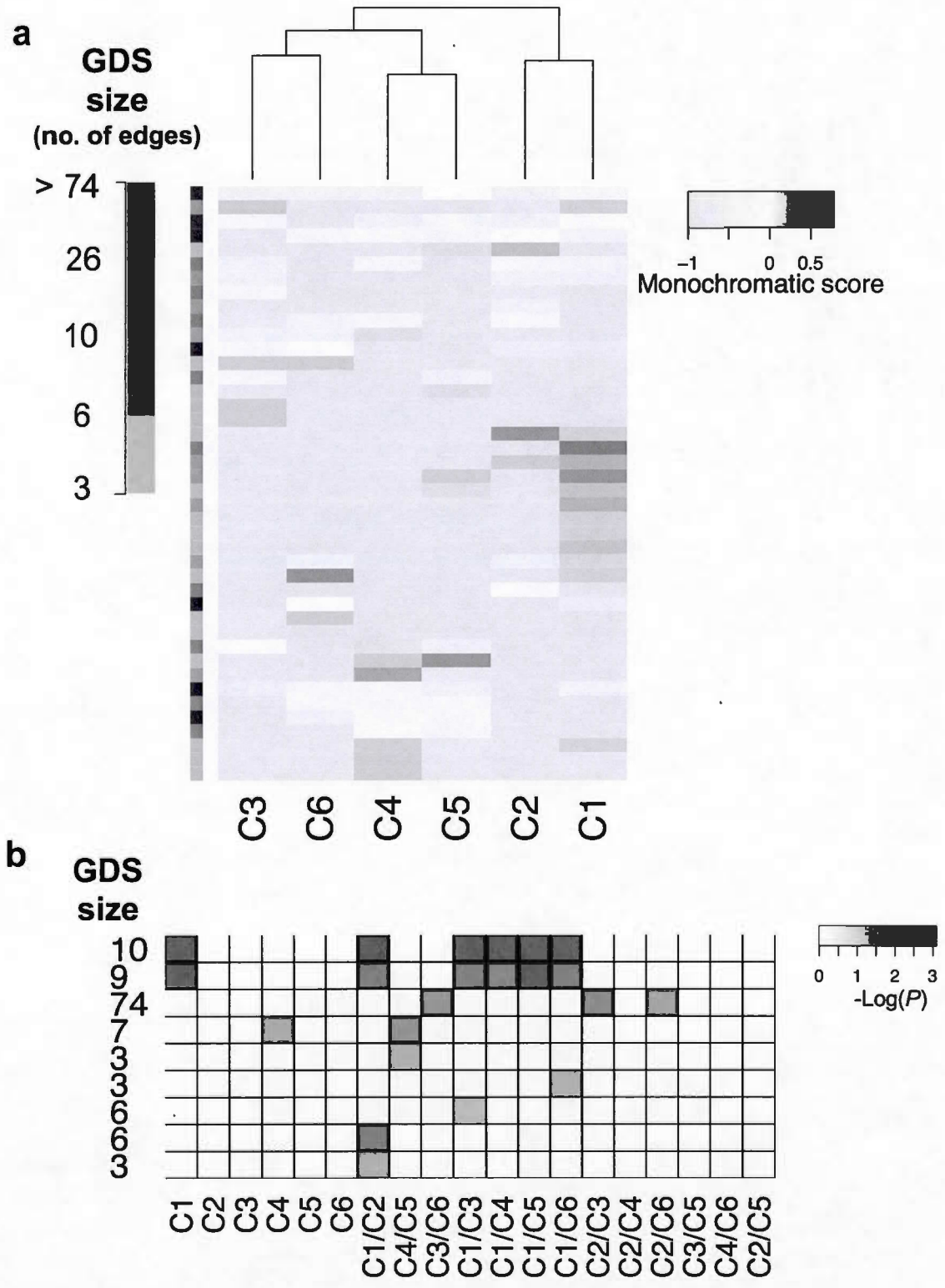


Figure 3.S3 GI classes tend to form GDS in a class-dependent manner.

(a) Hierarchical clustering of monochromatic indices of GI classes. Each row represents a GDS extracted from GIs-all using the MINE tool. GDS sizes are indicated by the blue shaded legend. **(b)** Enrichments of GIs classes and pair combinations of classes in GDS when compared to GIs-all. Only the statistically significant enrichments are shown with -Log10 adjusted P-values ($P < 0.05$, see Methods). The GDS size indicates the total number of GIs (edges) of the subnetwork.

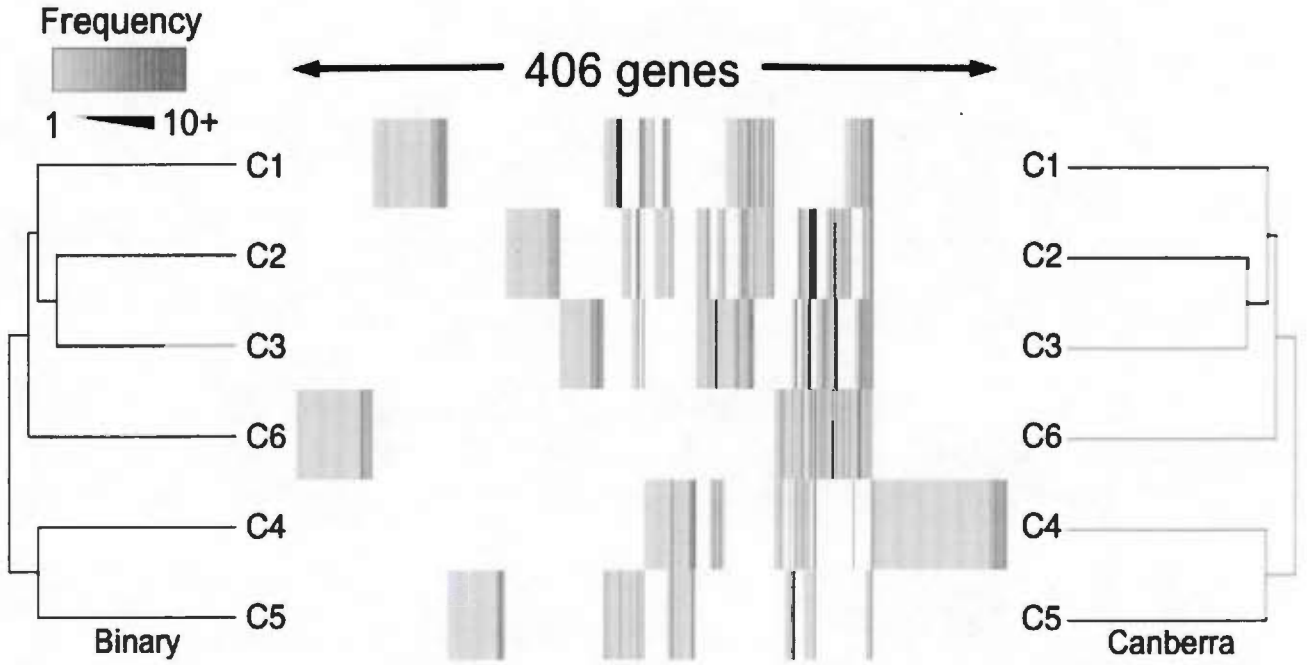


Figure 3.S4 Hierarchical clustering of GI classes based on their genes frequencies.

Gene frequencies are clustered using Binary (left dendrogram) and Canberra (right dendrogram) distance metrics (see Methods).

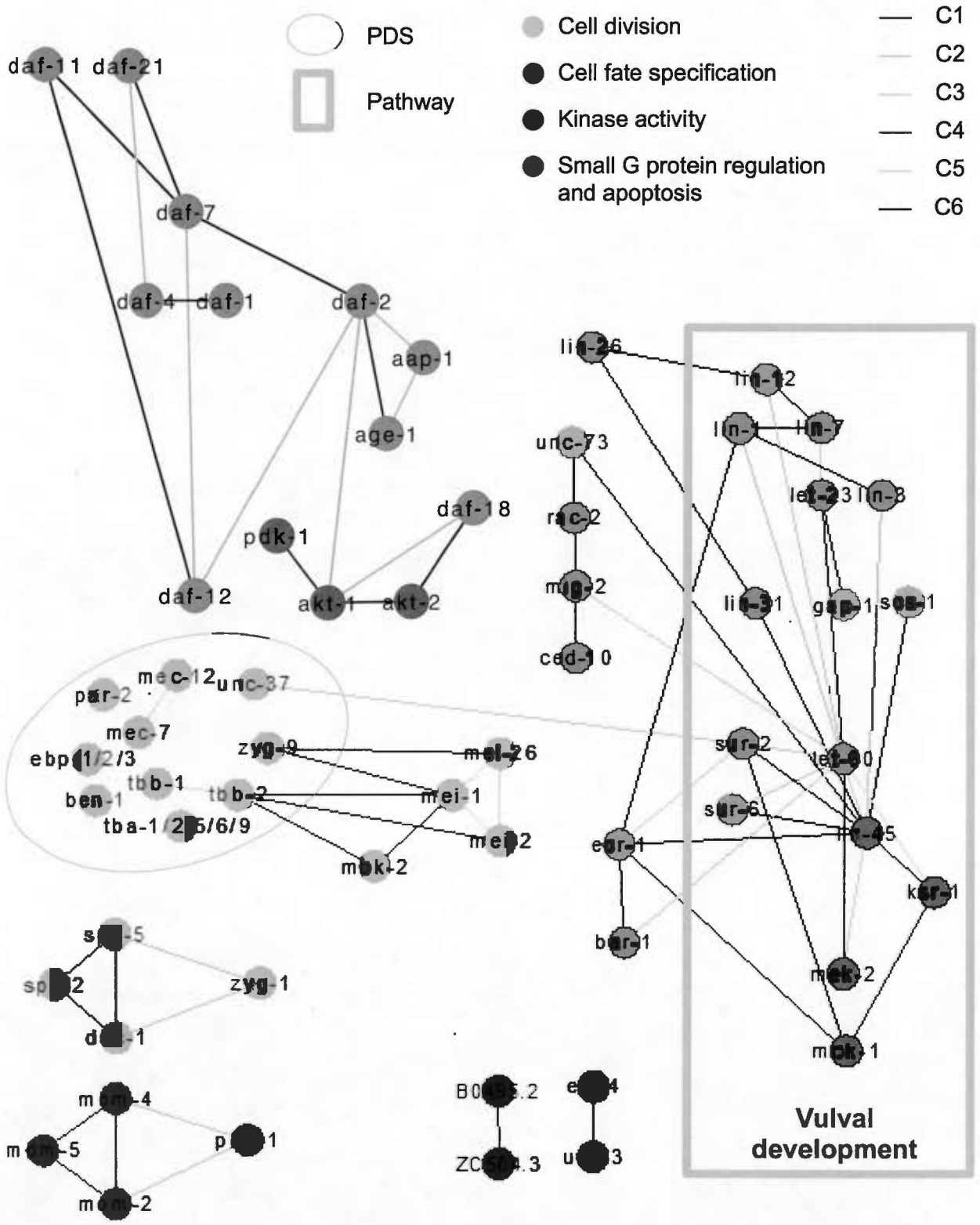


Figure 3.S5 Examples of genetic interactions from C1 to C6 classes.

Nodes represent genes connected by genetic interactions from the six classes of GIs. Are also indicated, protein-protein interaction dense subnetworks (PDS) and genes involved in C1 to C6 GIs and also involved in signalling pathways controlling vulval development (Table 3.S3 [En ligne: (Boucher et al. 2016)])

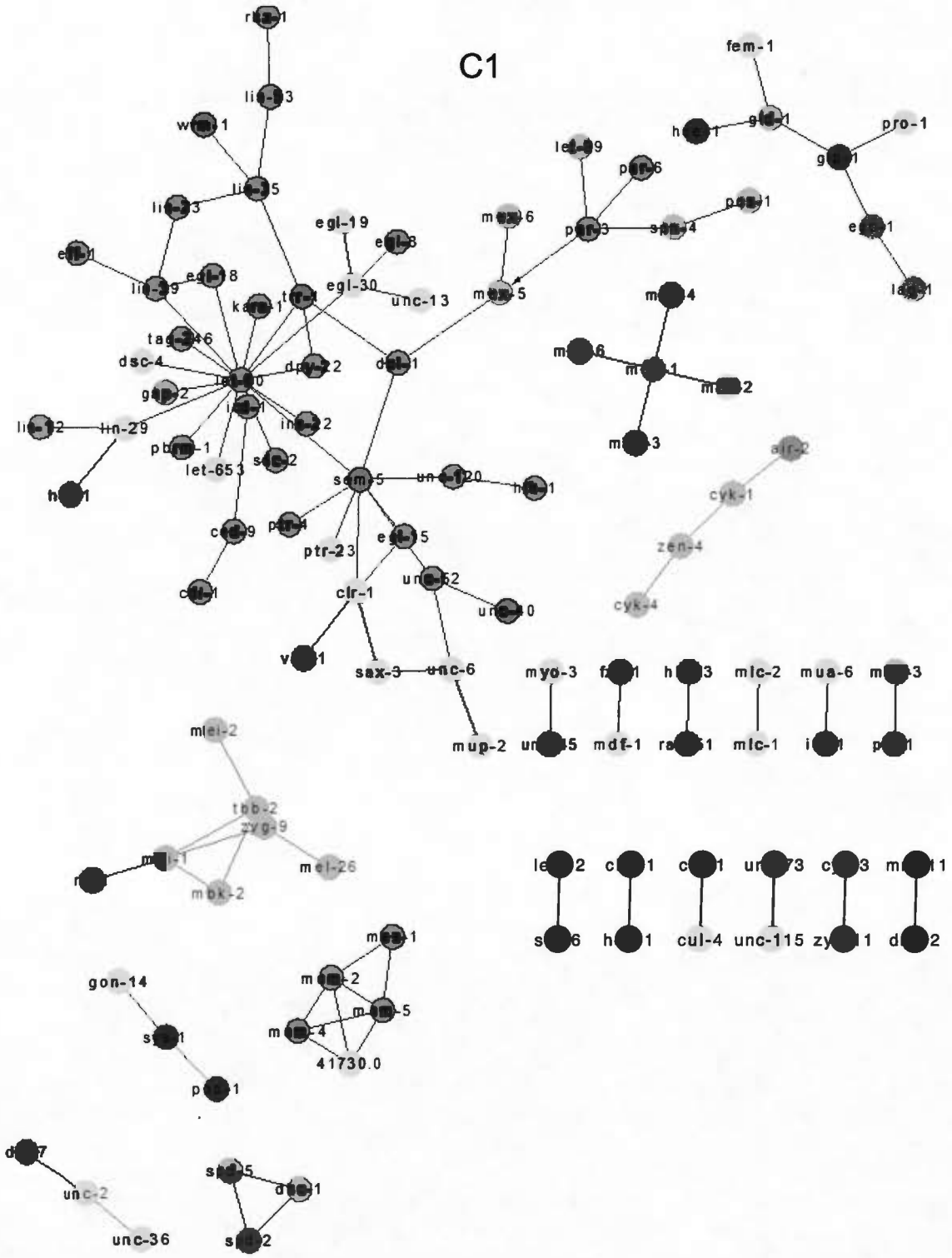


Figure 3.S6 Genetic interactions found in C1 GIs.

Refer to Figure 3.S5 for the nodes and edges descriptions.

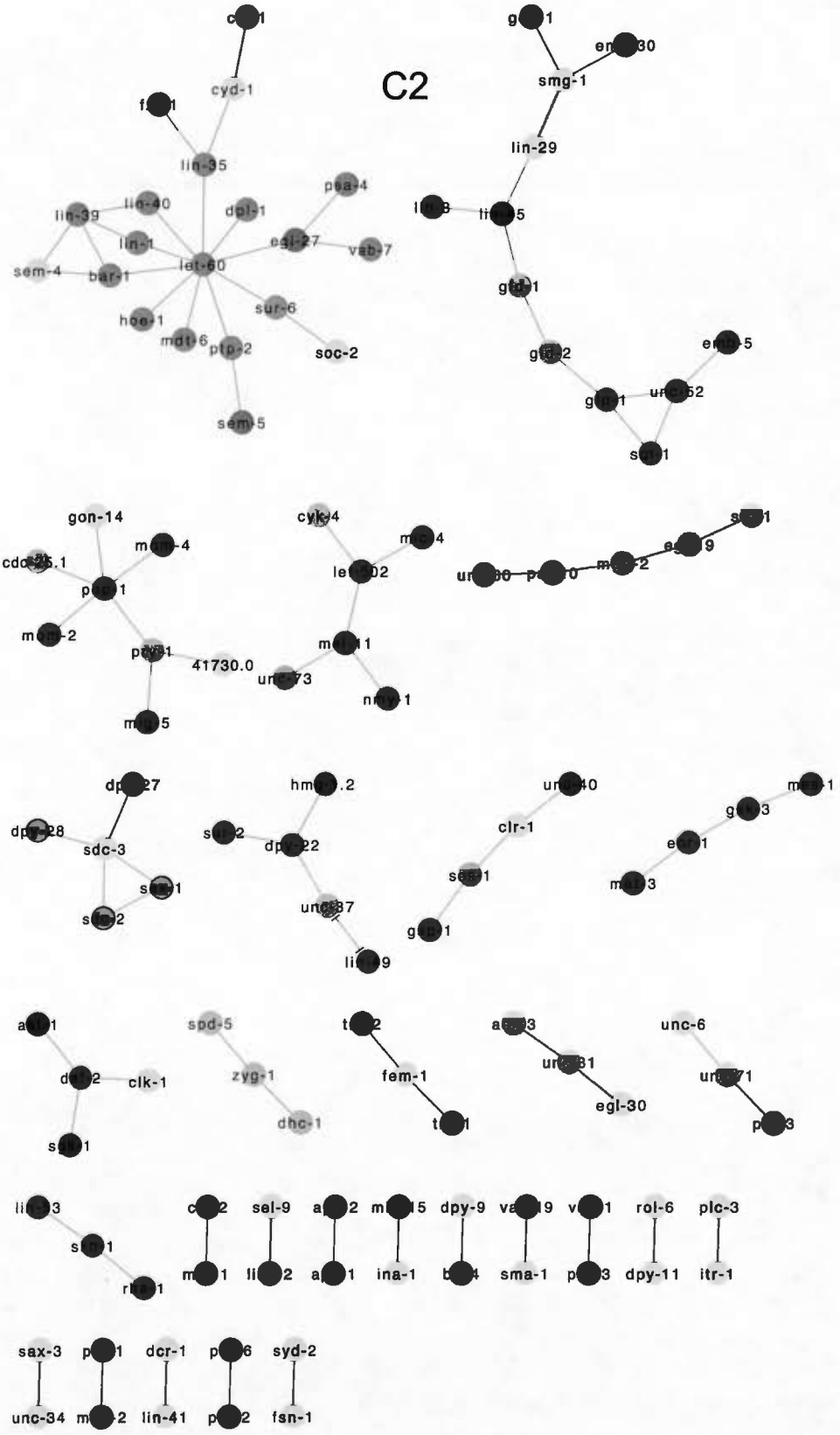


Figure 3.S7 Genetic interactions found in C2 GIs.

Refer to Figure 3.S5 for the nodes and edges descriptions.

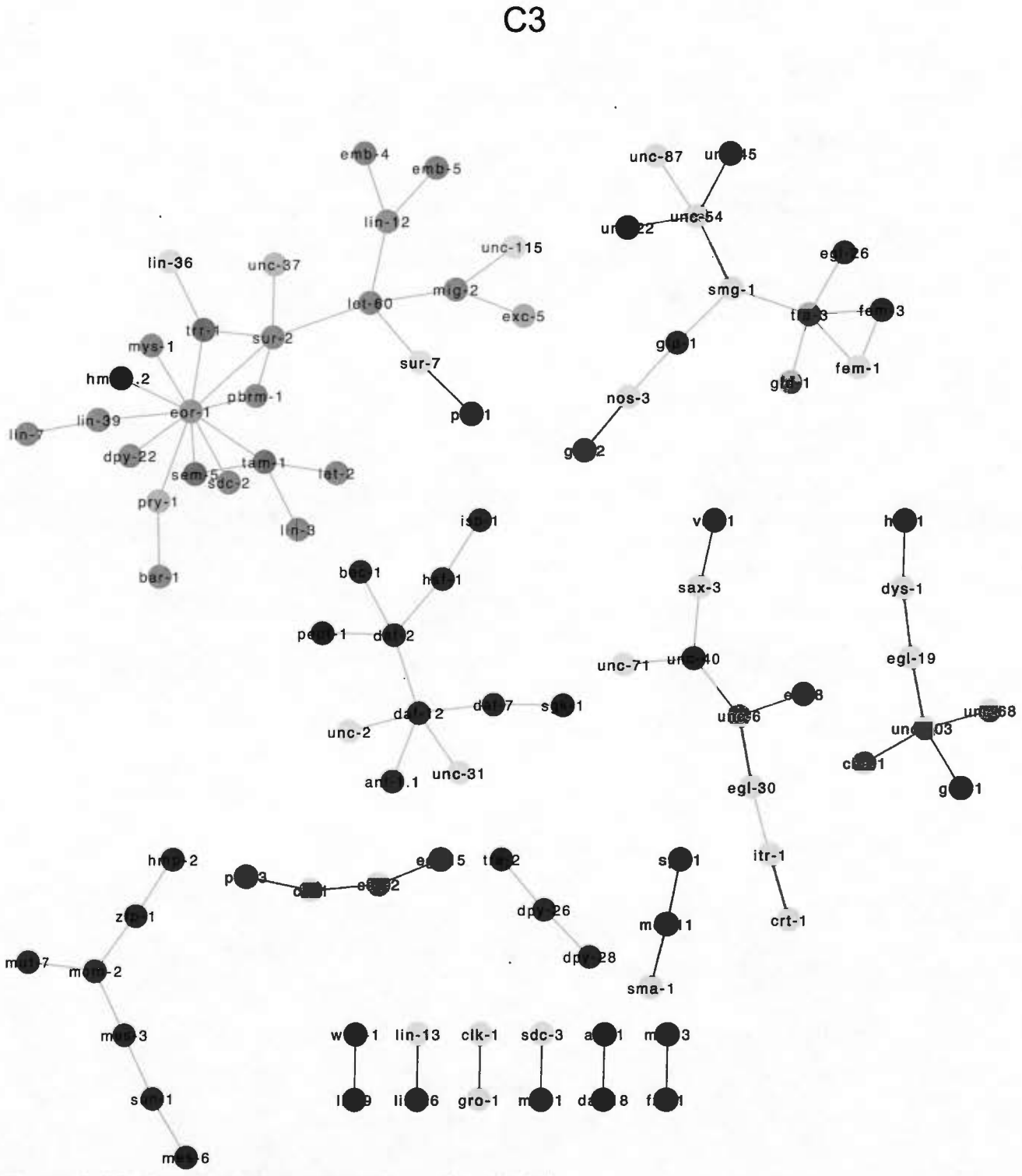


Figure 3.S8 Genetic interactions found in C3 GIs.

Refer to Figure 3.S5 for the nodes and edges descriptions.

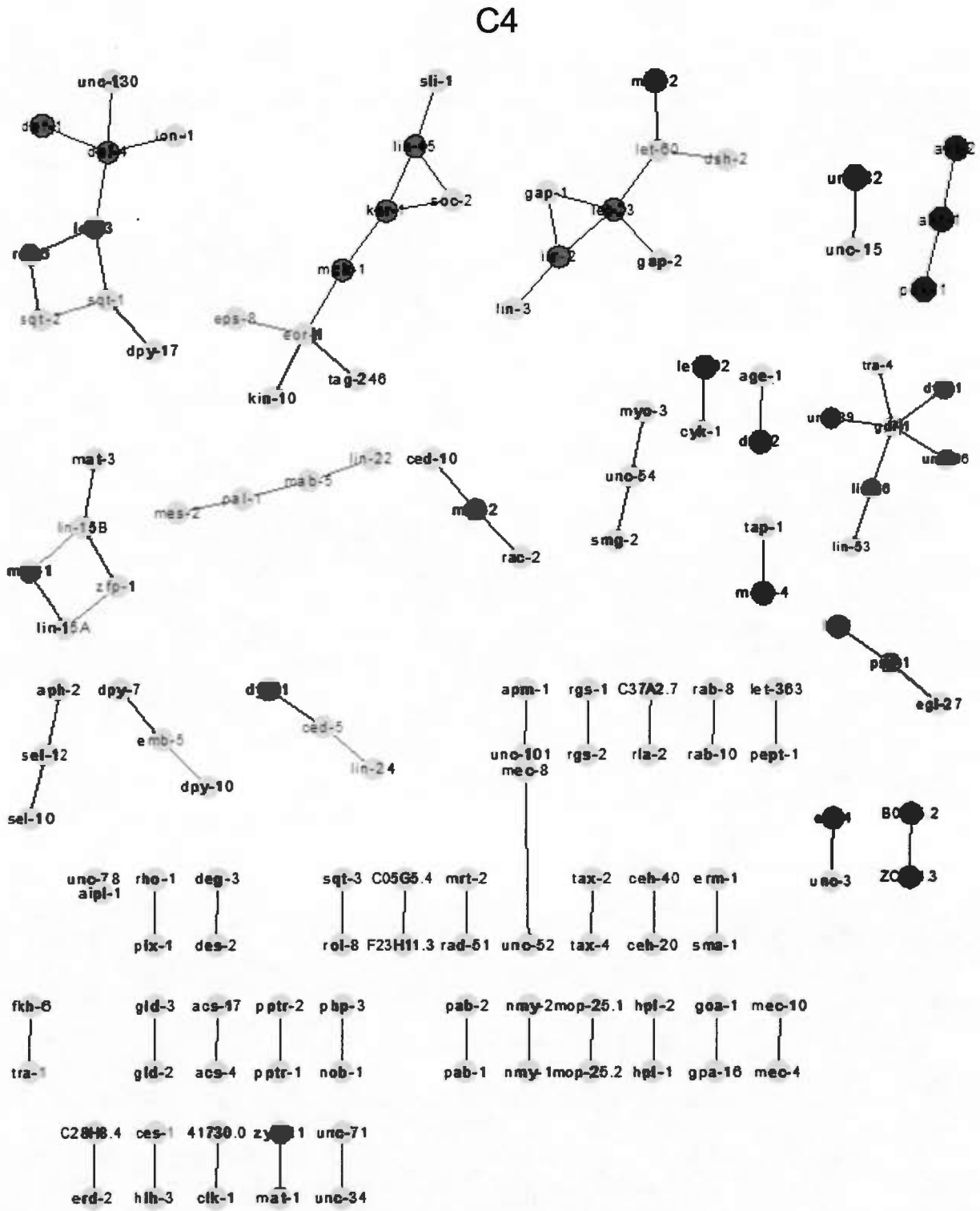


Figure 3.S9 Genetic interactions found in C4 GIs.

Refer to Figure 3.S5 for the nodes and edges descriptions.

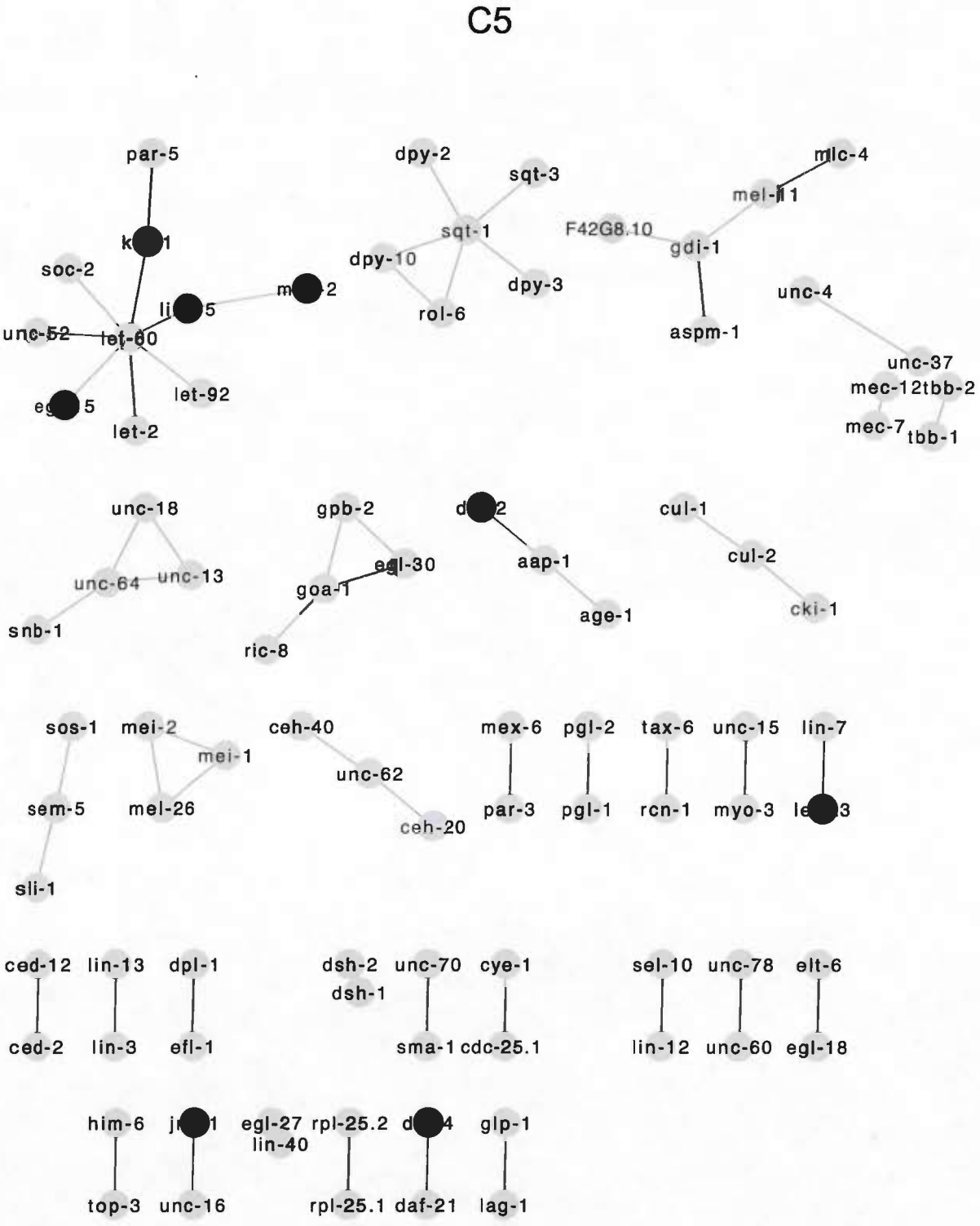


Figure 3.S10 Genetic interactions found in C5 GIs.

Refer to Figure 3.S5 for the nodes and edges descriptions.

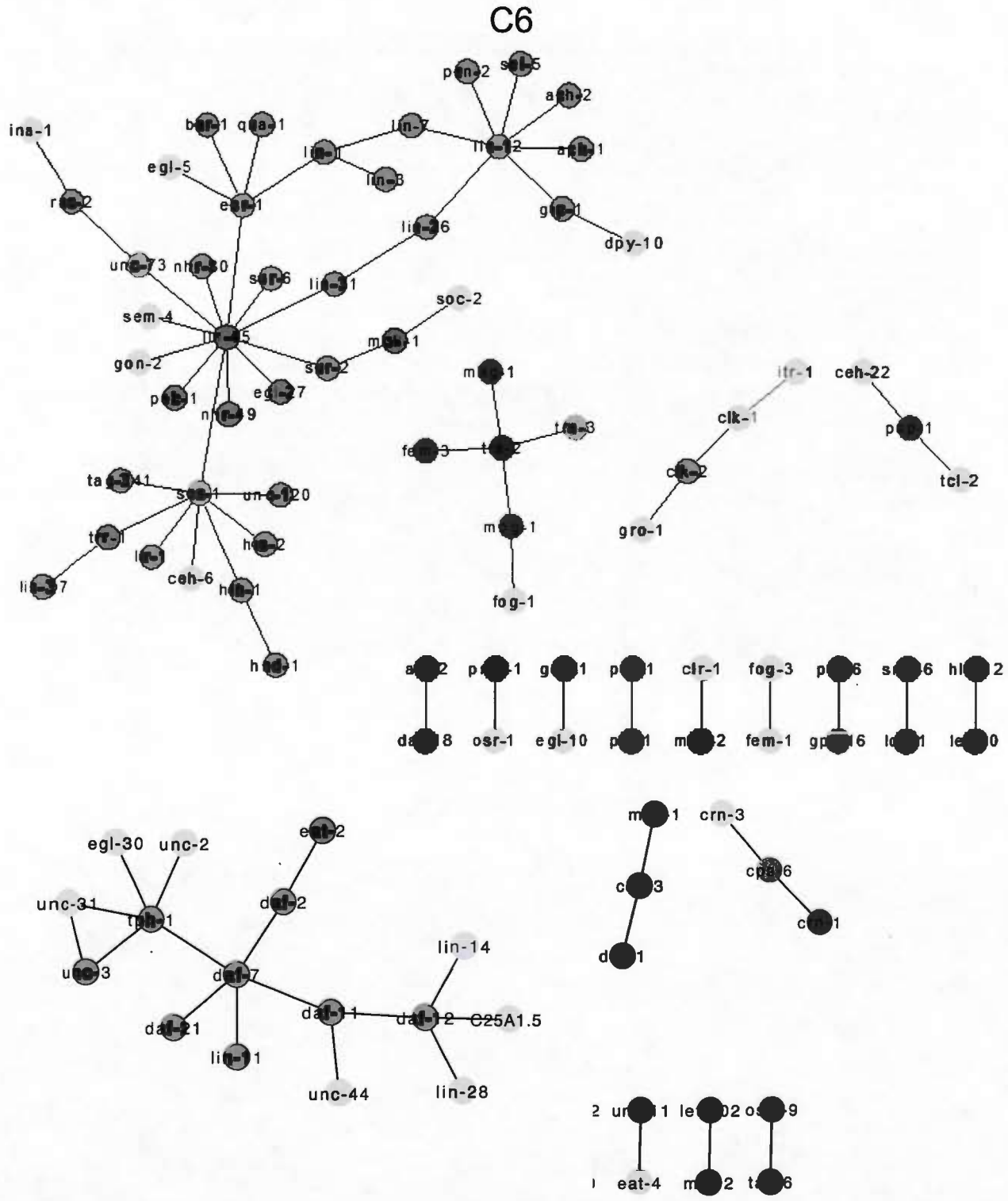


Figure 3.S11 Genetic interactions found in C6 GIs.

Refer to Figure 3.S5 for the nodes and edges descriptions.

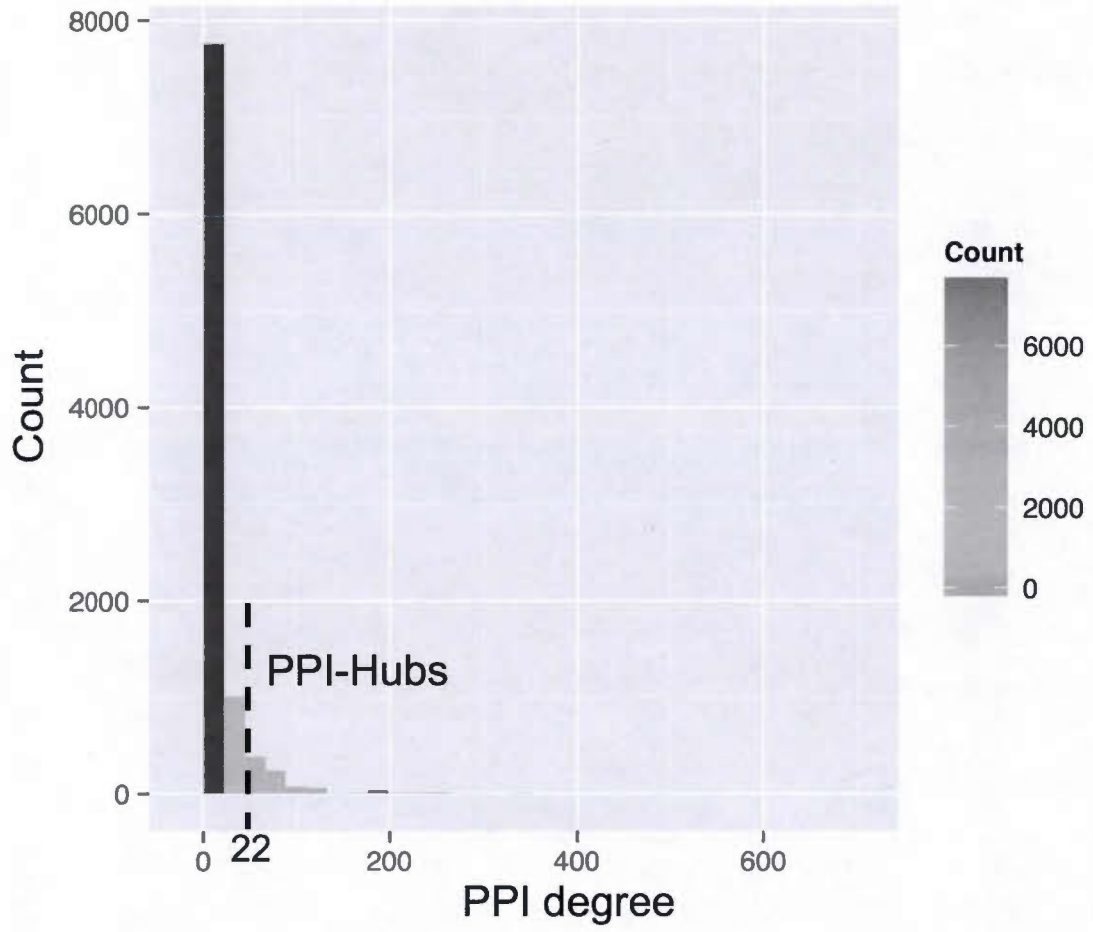


Figure 3.S12 Protein-protein interactions degree distribution for the multi-species Caenorhabditis elegans interactome.

PPI-Hubs are identified as the top 20% most connected proteins in the PPI network indicated by a dashed line.

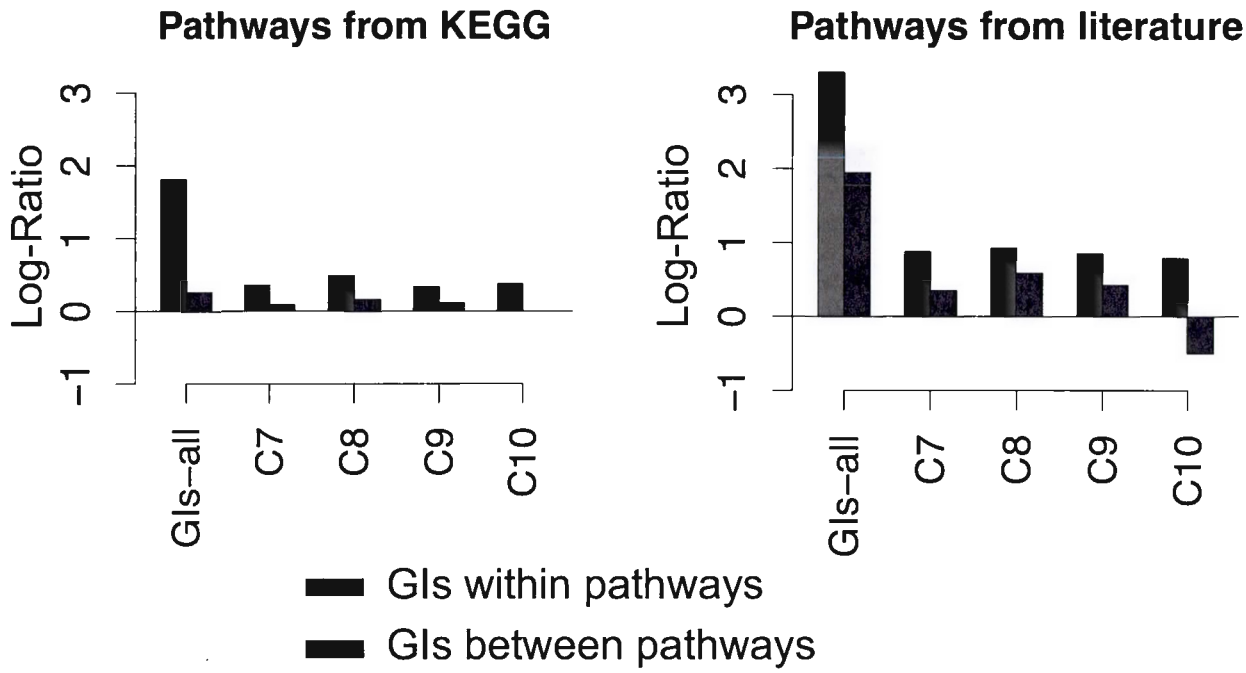


Figure 3.S13 C7 to C9 classes are enriched in between-pathways interactions.

Log-Ratio scores for within-pathway (red bars) and between-pathway (blue bars) relationships observed between genes interacting through unselected GI classes (C7-C10) or present in GIs-all. A positive Log-Ratio score means that the frequency of within- or between-pathway GIs occurring in GI classes is significantly higher than the frequency of similar situations witnessed in relevant randomized GI networks with a probability of 0.99.

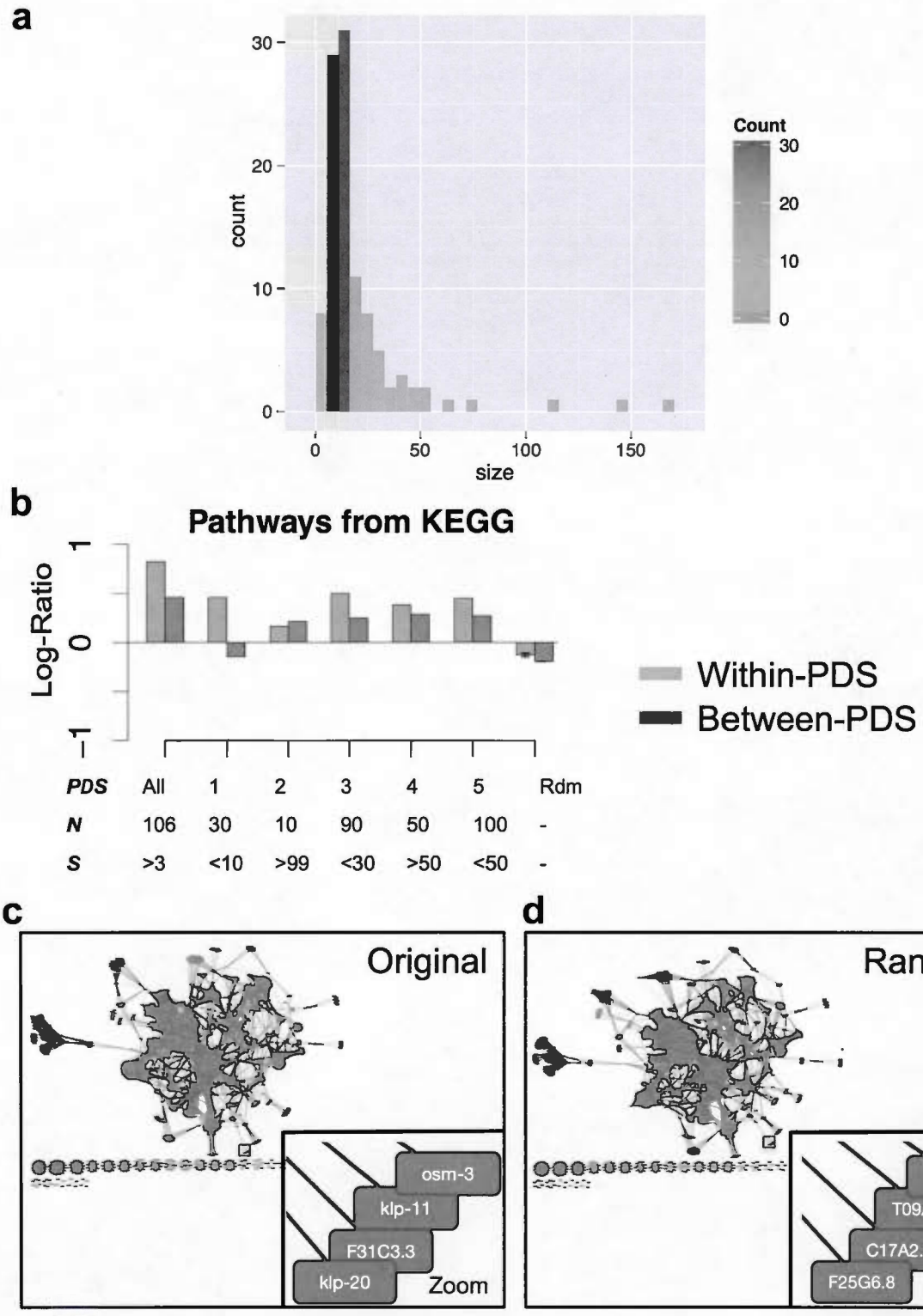


Figure 3.S14 Enrichment of within- and between-PDS relationship within pathways is not significantly influenced by the topology of PDS.

(a) Size distribution of 106 protein-protein interaction dense subnetworks (PDS). **(b)** Log-Ratio scores for within-PDS and between-PDS relationships occurring within-pathways. Different PDS networks were built by varying the number (N) and size (S) of the PDS. The original network "All" correspond to the union of PDS used for the study presented in Figure 3.6. Mean of Log-Ratio are indicated for Randomized PDS networks (Rdm) (n=100) have the exact same topology than the original PDS network. Error bar indicate standard deviation of Log-ratio obtained across the 100 Rdm. **(c-d)** Depiction of the original and the random PDS networks topology.

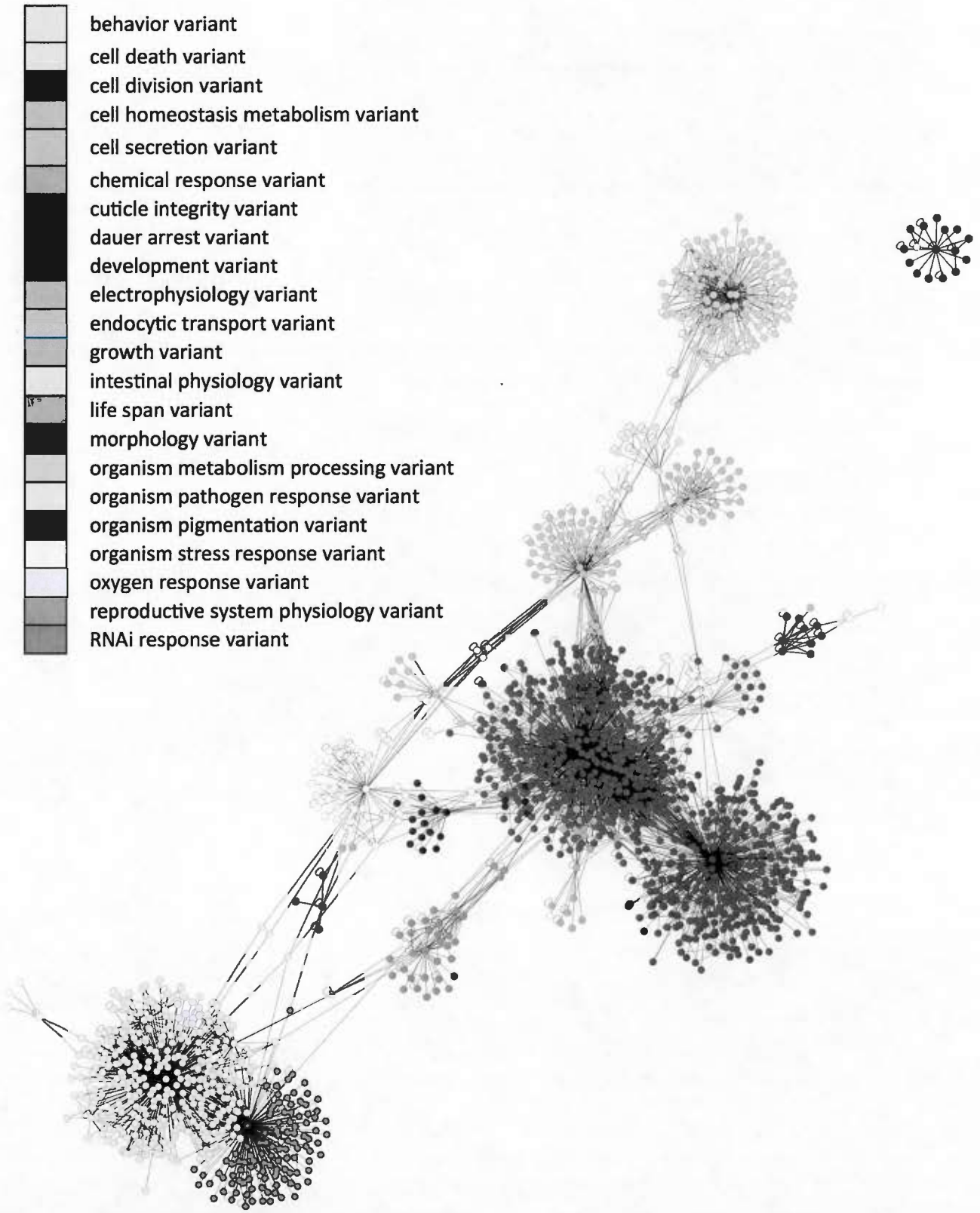


Figure 3.S15 Representation of the 22 classes of phenotypes identified in C. elegans.

Phenotype IDs retrieved from WormBase (release WS220-bugfix) are represented by nodes and their hierarchical relationships represented by edges. Groups of phenotypes corresponding to the 22 most general phenotypes, and their first neighbors in the network were identified by different node colors.

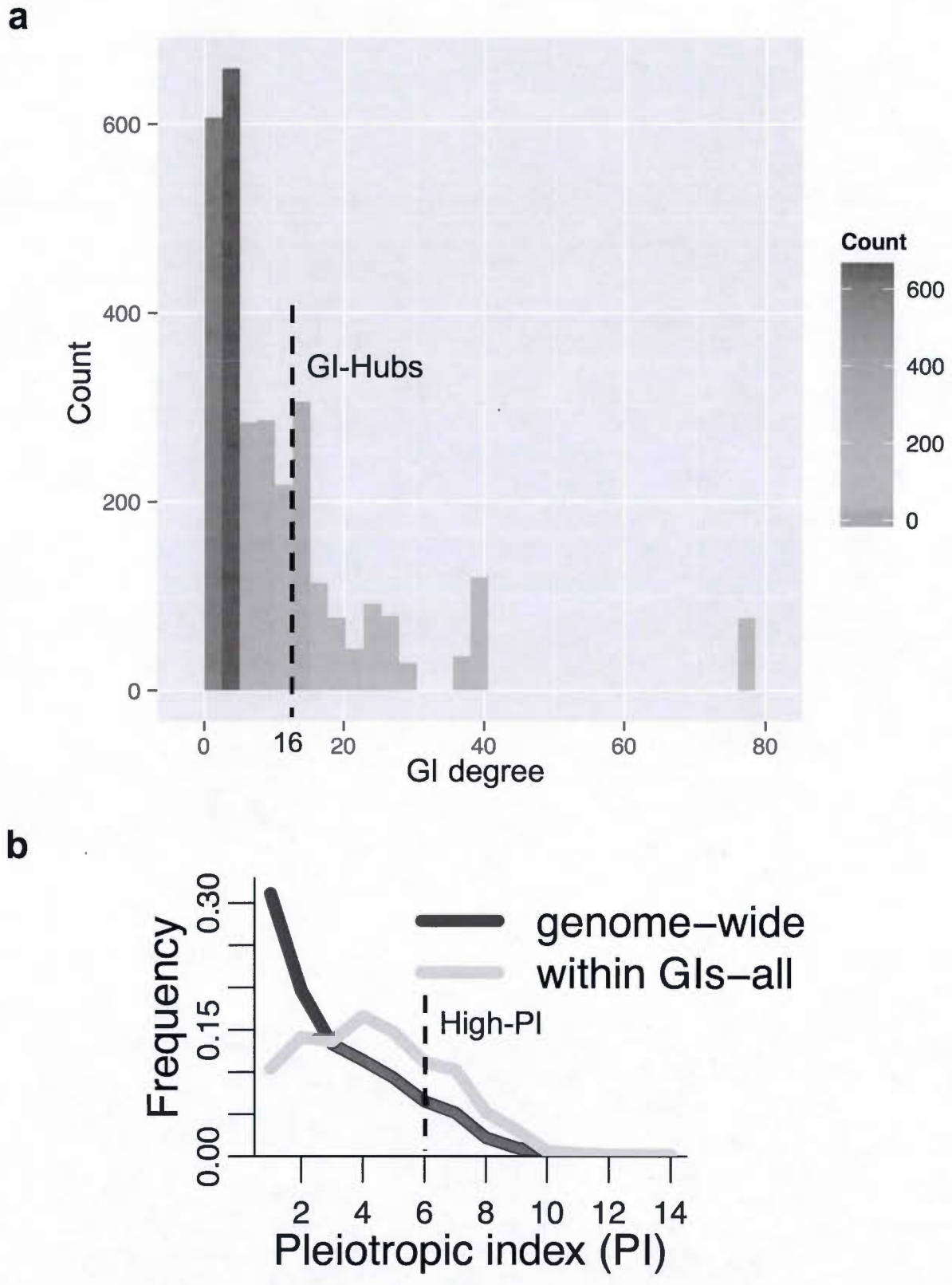


Figure 3.S16 Distribution of GI degree and pleiotropic index (PI).

(a) Degree distribution for GI degrees in GIs-all. GI-Hubs, corresponding to the 20% GIs with the highest GI degrees, are located on the right side of the dashed line (GI degree ≥ 16). **(b)** PI distribution for all genes with PI > 0 in the C. elegans genome and for all interacting genes in GIs-all. High-PI genes, corresponding to the 20% genes with the highest PI, are located on the right side of the dashed line (PI ≥ 6).

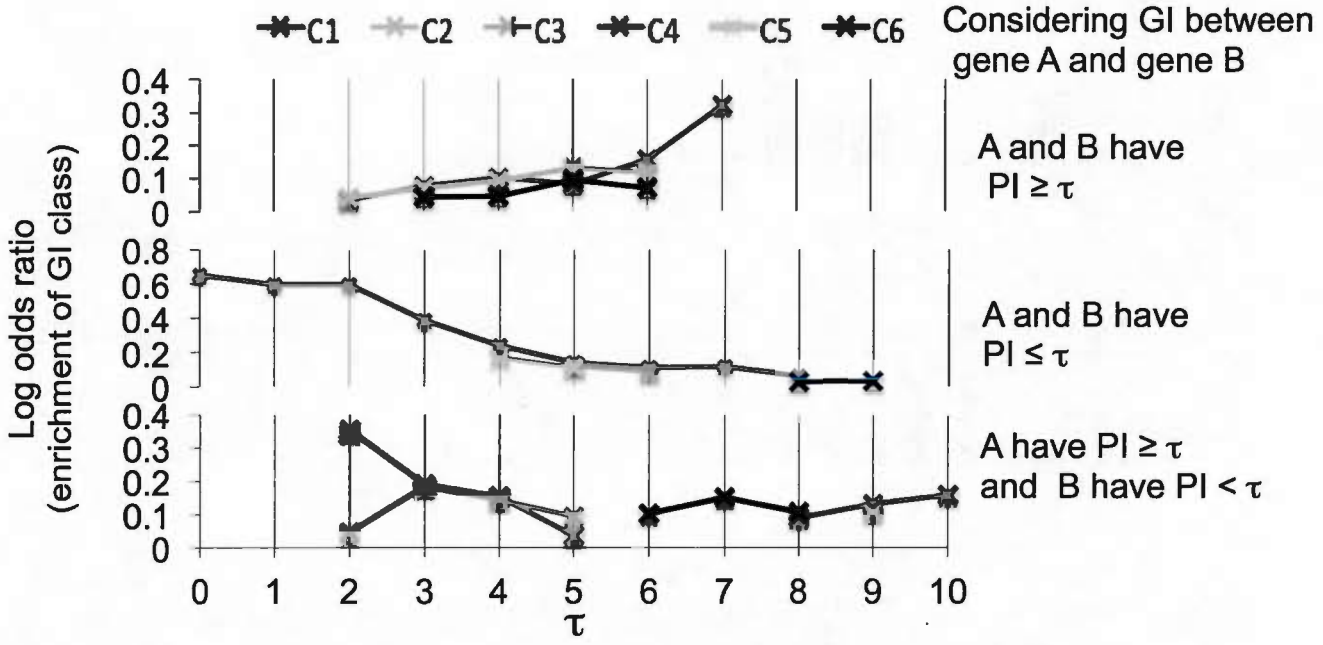


Figure 3.S17 Interaction between gene within or across pleiotropic ranges.

Log odds ratios of GI classes enriched in interactions between genes displaying the same range of PI higher or equal to a given threshold (τ) (upper panel); or lower or equal to τ (middle panel). GI classes enriched in interactions in which one partner (gene A) displays a PI ≥ τ and the other (gene B) display a PI < τ are also indicated. τ is indicated by the x-axis. Only significant enrichments of GI classes are indicated (Fisher's exact test, $P < 0.05$). High log odds ratio indicates high enrichment of GI classes within or between indicated PI ranges.

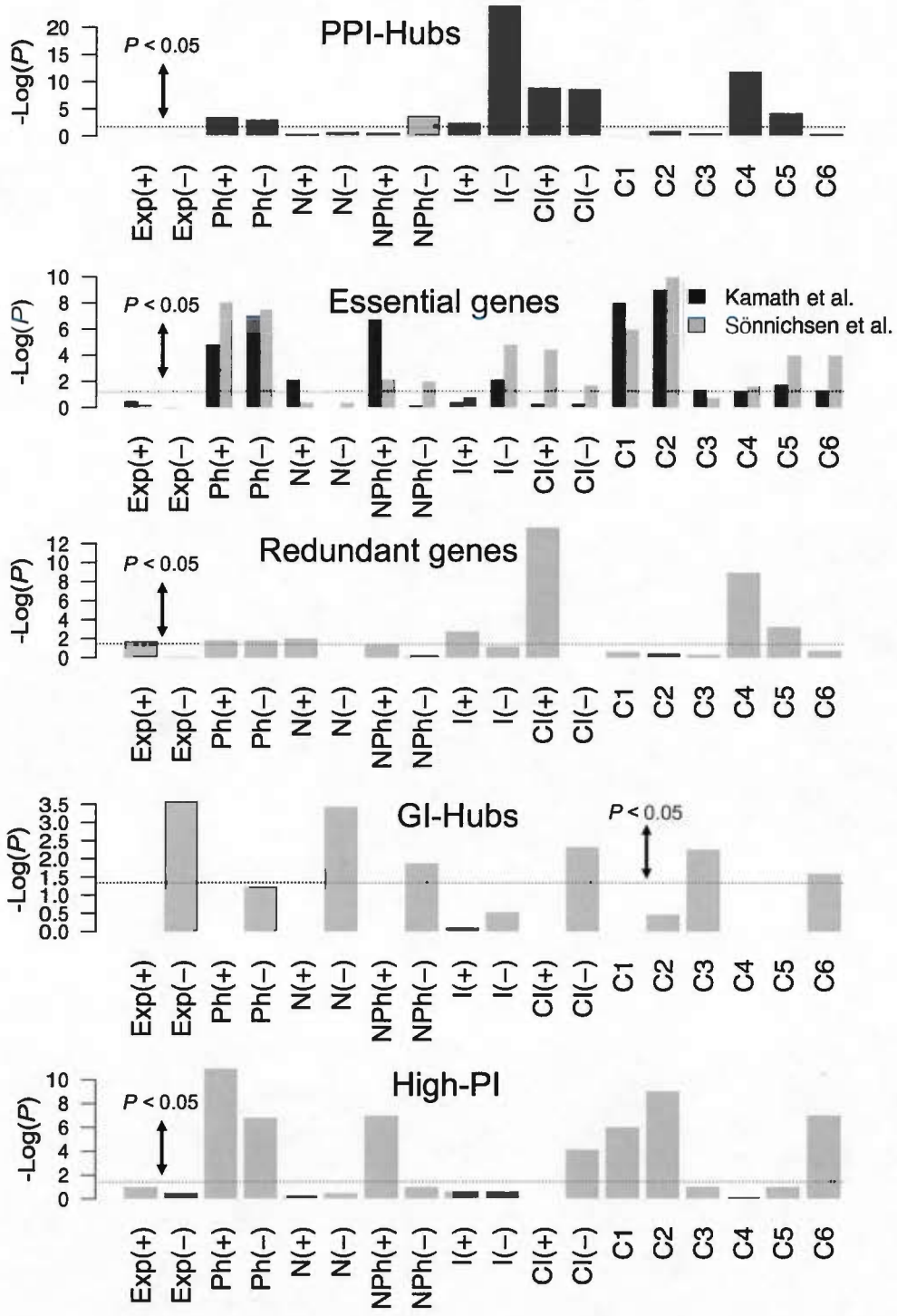


Figure 3.S18 GI groups and GI classes enrichments for different biological characteristics.

Enrichment in GI classes and GI groups of Genetic interaction-Hubs (GI-Hubs), redundant genes, protein-protein interaction Hubs (PPI-Hubs), Highly pleiotropic genes (High-PI) and essential genes. -Log of P-values from Fisher's exact test are indicated. GI groups are associated to a positive value (+; above a threshold) or a negative value (-; below the threshold) for indicated attributes. Threshold values for each attributes are indicated Table 3.S6 [en ligne: (Boucher et al. 2016)].

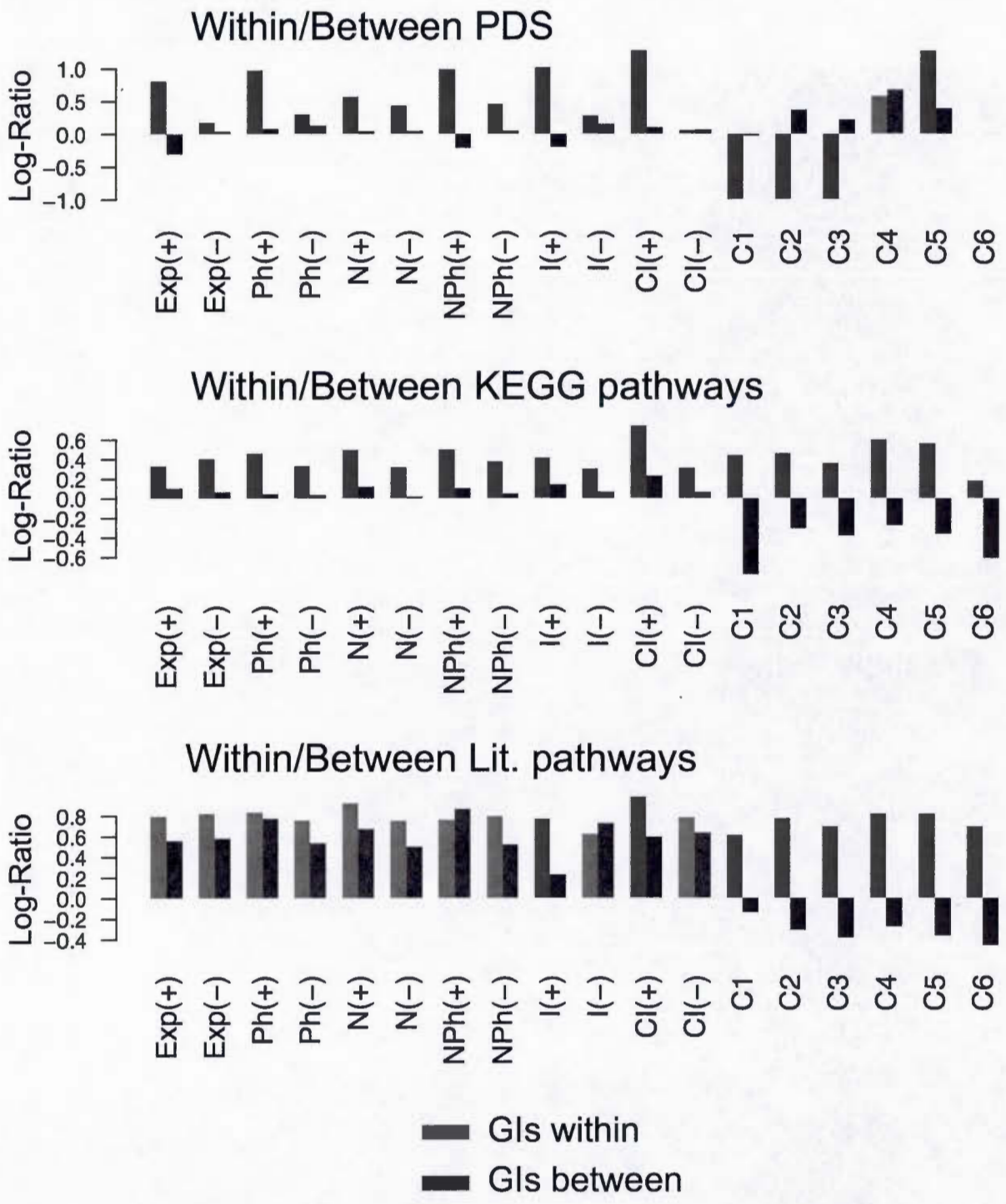


Figure 3.S19 Log-Ratios for GI groups and GI classes occurring within or between PDS or pathways.

Log-Ratios profiles for GI classes and GI groups associated to a positive (+) or a negative (-) values for indicated attributes. Blue boxes indicate enrichment of biological characteristics for C4 and C5 GI classes and CI(+) GI group.

		PDS-centric modules	PDS-independ. modules	Non-pleiotropic connectors	Pleiotropic connectors
Level VI Phenotype	**Pleiotropy**	Low to Average	High	Average	High
	Essential genes	+	++	-	+
Level V Biological/ molecular Function	**Functional enrichment**	Kinases	Cell division	-	wide spectrum of functions
	Redundant genes	++	-	-	-
Level IV Functional Interactions	**GI classes**	C4-C5	C1-C2	C3	C6
	GI degree	Average	Average	High	High
	Pathways	Within	Within	Within	Within
Level III Molecular Interactions	**PDS**	Within and between	Between (C2 only)	Between	-
	PPI degree	High	Average	Average	Average
Level II Gene expression	**Co-expression**	Average	High	Average	Low

Figure 3.S20 Summary of the distinctive characteristics identified for modules and connectors.

CHAPITRE IV

LA GÉNOMIQUE INTÉGRATIVE POUR IDENTIFIER LES GÈNES CONTRÔLANT LA REPROGRAMMATION MÉTABOLIQUE DES TUMEURS DU SEIN CHEZ L'HUMAIN

Boucher B, Diallo AB, Jenna S. (2017) Integrative genomics identifies genes controlling metabolic reprogramming in human breast tumours. Manuscrit en préparation finale pour la soumission.

Après avoir caractérisé et mis en évidence des classes d'interactions génétiques à l'intérieur des voies chez *Caenorhabditis elegans*, nous avons décidé d'explorer cette approche chez l'humain et plus particulièrement, dans le cancer du sein.

Brièvement, on sait que les cellules cancéreuses, au contraire des cellules saines, ont plus souvent recourt à la glycolyse pour produire de l'ATP et ce, même en présence d'oxygène. Cette particularité du métabolisme des cellules cancéreuses serait aussi associée à des formes plus agressives de cancer. Dans les tumeurs du sein, cela s'expliquerait en partie par une sur-activation de la glycolyse et une surutilisation du glucose dans des types de cellules associés à des formes agressives de cancer du sein.

Pour ce chapitre, nous avons décidé d'utiliser les connaissances acquises et décrites dans le chapitre précédent pour identifier les gènes clés des réseaux fonctionnels du métabolisme des cellules du cancer du sein. Nous allons ici proposer une liste de gènes prédits comme ayant un rôle clé dans le contrôle de la glycolyse et dans le développement des cellules du cancer du sein.

CONTRIBUTIONS

Dans ce troisième article, j'ai (BB) rassemblé l'ensemble des données provenant des bases de données chez l'humain. J'ai conçu l'ensemble des algorithmes et construit les attributs pour l'ensemble des paires de gènes du génome humain. J'ai effectué les analyses *in silico* d'expression et de co-expressions des gènes et appliqué les différents filtres. J'ai aussi conçu l'algorithme pour le calcul des proportions d'interactions de chacune des classes d'interactions fonctionnelles pour les gènes situés à l'intersection de voies de signalisation/métabolisme. Finalement, j'ai fait l'ensemble des tests statistiques d'enrichissement.

Abdoulaye Baniré Diallo (ABD) a contribué à la conception et au design de l'étude. Il a fait aussi fait un suivi technique du projet.

Sarah Jenna (SJ) a conçu et dirigé le projet. Elle a effectué l'ensemble des analyses statistiques des gènes associés à l'effet Warburg. Elle a élaboré le concept de force contextuelle et défini la plasticité entre les sous-types de cancer du sein. Elle a conçu les modèles prédictifs des gènes Warburg. Elle a aussi contribué aux analyses statistiques et a effectué la comparaison des méthodes de prédiction de gènes impliqués dans l'effet Warburg.

J'ai rédigé le Matériel et méthode, les légendes des figures, ainsi qu'une partie des résultats de l'article. Sarah Jenna a rédigé le reste.

Détail des contributions par figure :
Figure 4.1: BB (design et réalisation des figures), SJ (design).
Figure 4.2: BB (collecte de données, analyse, design et réalisation des figures), ABD (supervision technique), SJ (design).

Figure 4.3: BB (collecte de données, analyse, design et réalisation des figures), ABD (supervision technique), SJ (design et réalisation des figures).

Figure 4.4: BB (collecte de données, analyse), ABD (supervision technique), SJ (design).

Figure 4.5: BB (analyse), ABD (supervision technique), SJ (collecte de données, analyse, design et réalisation des figures).

Figure 4.6: SJ (collecte de données, analyse, design et réalisation des figures).

Figure 4.7: SJ (collecte de données, analyse, design et réalisation des figures).

Figure 4.8: SJ (collecte de données, analyse, design et réalisation des figures).

RÉSUMÉ

Le cancer du sein humain est une maladie hétérogène dans laquelle la progression tumorale pourrait dépendre d'un changement dans les programmes métaboliques. Cette reprogrammation métabolique implique l'utilisation de la glycolyse aérobique comme source d'énergie et est connue sous le nom de l'effet Warburg. En outre, le maintien de ces programmes pourrait être contrôlé par plusieurs gènes qui ne sont pas directement impliqués dans ces voies métaboliques. Afin d'identifier ces gènes associés à l'effet Warburg, nous avons développé une nouvelle approche de génomique intégrative pour prédire les réseaux d'interactions fonctionnelles (IFs) des sous-types de tumeurs du sein Luminal A et Basal-like, qui sont caractérisés par un effet Warburg faible et élevé respectivement. Nous avons identifié les propriétés distinctives de 28 gènes associés à l'effet Warburg dans le réseau prédit du cancer du sein et construit un modèle prédictif basé sur ces propriétés pour identifier de nouveaux gènes Warburg dans le génome humain. Fait intéressant, le modèle prédictif résultant a significativement surpassé l'identification des gènes Warburg basée uniquement sur des données d'expression. Au total, 70 gènes Warburg prédits seront soumis à des validations expérimentales qui

pourraient conduire à la découverte de nouvelles cibles pour le traitement anti-tumoral et anti-métastatique pour le cancer du sein.

4.1 Abstract

Human breast cancer is an heterogeneous disease in which tumor progression could rely on a change in metabolic programs. This metabolic reprogramming involves the use of aerobic glycolysis as an energy source and is known as the Warburg effect. Moreover, the control of these programs could be controlled by several genes that may not be directly implicated in these metabolic pathways. To identify these Warburg effect associated genes, we have developed a novel integrative genomics approach to predict functional interaction (FI) networks of Luminal A and Basal-like breast tumor subtypes which are characterized by low and high Warburg effect respectively. We identified the distinctive properties of 28 Warburg-associated genes within predicted breast cancer FI networks and built a predictive model based on these properties to identify novel Warburg effect genes in the human genome. Interestingly, the resulting predictive model significantly outperformed the identification of Warburg effect genes based on expression data only. In total, 70 predicted Warburg genes will be submitted to experimental validations that could lead to the discovery of novel targets for anti-tumoral and anti-metastatic therapy for breast cancer.

4.2 Background

Cells and organisms need to generate biocompatible energy to live, develop and reproduce. The generation of this energy involves the production of ATP from

metabolic processes consuming macromolecules such as lipids, carbohydrates and proteins in the presence or in the absence of oxygen. In aerobic conditions (presence of oxygen), ATP production involves oxidative phosphorylation (OXPHOS) occurring at mitochondria membranes (Hinkle, 2005). In anaerobic conditions or upon mitochondrial dysfunction, glycolysis and lactic fermentation (resulting in the production of lactate) become the main energy providers of the cells (Cairns *et al.*, 2011). Cancer cells extensively use glycolysis/fermentation to produce ATP even in aerobic conditions, a biological phenomenon called the Warburg effect (Lincet et Icard, 2015). While TCA cycle and OXPHOS still function to some extent in cancer cells, an energetic metabolism heavily relying on glycolysis provides the cells with several growth advantages (DeBerardinis, Sayed *et al.*, 2008). For instance, it allows faster synthesis of ATP with reduced production of reactive oxygen species (ROS). In addition, glycolysis produces large number of precursors for biosynthetic pathways (DeBerardinis, Lum *et al.*, 2008; Ward et Thompson, 2012).

Breast cancer cells can be divided in subtypes based on their morphology and molecular signatures including Luminal A cells, which display an epithelial morphology and are characterized by active oestrogen and progesterone signaling and Basal-like subtype, which is characterized by mesenchymal morphology and inactivation of these pathways (Blick *et al.*, 2008; Rakha *et al.*, 2009). Recent studies revealed that aerobic glycolysis was not adopted evenly as a main energetic metabolic pathway by all cancer cells (Guppy *et al.*, 2002; Martin *et al.*, 1998; Pasdois *et al.*, 2003; Rodríguez-Enríquez *et al.*, 2006). In fact, Basal-like breast cancer cells were shown to use aerobic glycolysis at a higher rate than Luminal A cells (Brauer *et al.*, 2013; Palaskas *et al.*, 2011). Metabolic reprograming of cells which involves the activation of aerobic glycolysis has been correlated with increased aggressiveness and poor outcome in patients (Gillies *et al.*, 2008). Also, this reprogramming was recently associated with the ability of breast cancer cells to invade tissues and form secondary tumors (Dupuy *et al.*, 2015). For instance, formation of secondary tumor in liver was

recently shown to be associated with an increase in aerobic glycolysis, while the invasion of lung by cells from the same breast primary tumor was associated with a metabolism relying mainly on OXPHOS (Dupuy *et al.*, 2015). Metabolic reprogramming during cancer progression is therefore recognized as an important hallmark of aggressive tumors and attracts a growing attention for the development of cancer therapeutics (DeBerardinis, Sayed *et al.*, 2008).

Genes controlling tumor progression are usually identified through measurement of differential expression of these genes in tumor versus normal tissues using expression data (DNA-chip and RNA-seq) or protein and phosphoprotein quantification using proteomic analyses (Mertins *et al.*, 2016; Mocellin et Rossi, 2007). Integrative genomics approaches have also been successfully used to identify driver mutations, as well as prognostic and diagnostic genetic markers for breast cancer (Malinowski, 2007; Normanno *et al.*, 2009; Tian *et al.*, 2014). Still, no approach has been developed to specifically target genes controlling metabolic reprogramming. Consequently, the development of such approaches is still needed to develop novel anti-tumoral and anti-metastatic therapeutic strategies.

We recently developed a novel integrative genomics approach using statistical assessment of the context of genes within functional networks using *C. elegans* as a model system (Boucher, B. *et al.*, 2016; Lee, A. Y. *et al.*, 2010). This approach was proven efficient to predict genetic interactions in the nematode (Lee, A. Y. *et al.*, 2010) and to identify classes of genetic interactions (GIs) defining functional modules with distinctive properties within a GI network highly enriched in functional interactions between genes occurring within signalling and metabolic pathways (Boucher, B. *et al.*, 2016). This study also identified the genomics signatures of genes with expected critical coordination function within functional interaction networks (Boucher, B. *et al.*, 2016). We consequently hypothesized that transposition of this integration strategy to Luminal A and Basal-like breast cancer subtypes, characterized by low and high

levels of aerobic glycolysis respectively, will allow us to identify genes controlling the metabolic reprogramming associated to tumour progression.

We tested this hypothesis and revealed that a set of 28 Warburg associated genes displayed distinctive properties within predicted breast cancer functional interaction networks. These properties were then used to build a predictive model to identify Warburg effect genes in a genome-wide manner. We tested these predictions *in silico* using a list of genes co-cited in the literature with the term "Warburg effect" from which were remove genes used to build the model. This validation demonstrated that the predictive model built in this study significantly outperformed identification of Warburg effect genes based on expression data only. A set of 70 genes predicted by our model will be submitted to experimental validation using both *C. elegans* and breast cancer cells. This validation may identify novel targets for anti-tumoral and anti-metastatic therapy for breast cancer, especially for Basal-like subtypes which are associated with poor outcome for patients.

4.3 Results

4.3.1 Overview of the integration strategy

We previously showed that integrating expression, protein-protein interactions (PPI) and phenotype data from the nematode allowed the identification of different classes of functional interactions between genes with distinctive structural and functional characteristics within *C. elegans* signaling and metabolic pathways (Boucher, B. *et al.*, 2016). In the current work, we hypothesized that Warburg effect genes could display distinctive properties within the functional network that could lead to the generation of a predictive model to identify these genes genome-wide. To test this hypothesis, we

transposed the integration strategy previously developed in *C. elegans* to breast cancer. We retrieved, from public databases, breast cancer expression data, PPI and disease annotations for human genes (Data; Figure 4.1A). We then used these data to build statistical attributes derived from single data-type (Data derived-attributes; Figure 4.1A and B) or multi-data networks (Network derived attributes; Figure 4.1A and B). We also built a reference set of "true" functional interactions between genes on which the predictive power of each attribute to identify true functional interactions was tested (Reference Set and Class validation; Figure 4.1A). We subsequently used statistical attributes with high predictive power to generate a predicted genome-wide breast cancer functional interaction networks (Genome-wide set; filtering stage and Predictive network; Figure 4.1A). Next, we used Luminal A and Basal-like functional networks to identify genes interacting with both glycolysis and OXPHOS machineries (Predictive network; Figure 4.1A). Using these subnetworks, we extracted the distinctive properties of 28 Warburg effect genes and used these properties to build a model that ranked all genes interacting with both glycolysis and OXPHOS machineries based on their probability of being a Warburg effect genes (Predictive network; Figure 4.1A). We subsequently tested these predictions *in silico*. The detailed description of this integration strategy is presented below.

4.3.2 Building the Reference set of functional interactions

We first identified a set of experimentally validated functional interactions within the human genome. To our knowledge, no comprehensive screen has been carried out to identify genetic interactions in human breast cancer cells. We previously showed that classes of genetic interactions (GIs) identified in *C. elegans* were enriched in within-pathways GIs. Consequently, we used within-pathways gene pairs in human as a

reasonable starting point to build a set of "true" functional interactions. Human metabolic and signaling pathways (3,424 genes; 203 pathways) were retrieved from WikiPathways (June 2015; (Kutmon *et al.*, 2016)). Combinations within individual pathways allowed the identification of 245,365 gene-pairs within pathways (Reference set; Figure 4.1A).

To identify within-pathway gene pairs with the highest probability of being "true" functional interactions, we selected those that are composed of genes co-cited in more than two articles (Figure 4.2A), and retrieved from the PubMed database in June 2015 (16,544 gene-pairs; Reference set; Figure 4.1A). Similar methods have been successfully employed to uncover functional associations between genes (Karadeniz *et al.*, 2015; Lee, I. *et al.*, 2008; Lee, I. *et al.*, 2010; Sintupisut *et al.*, 2013) and proteins (Franceschini *et al.*, 2013). Also, since genome-wide studies tend to cite many unrelated genes, we excluded from this study PubMed IDs citing more than 15 genes (which correspond to ~1% of all PubMed entries; Figure 4.2B). In total, 7,118 gene pairs composed of co-cited genes in more than two independent non-genome-wide studies constituted the reference set of "true" functional interactions (Reference set; Figure 4.1A).

4.3.3 Generation of co-expression attributes

Expression data from 1,058 breast invasive carcinoma were downloaded from the TCGA Data Portal (see Methods). Twelve samples from male patients were removed and the resulting data collection from 1,036 tumors were classified into five breast cancer (BC) subtypes (Luminal A, Luminal B, HER2, Basal-like and Normal-like) using the AIMS tool (Paquet et Hallett, 2015): 321 tumors were classified as Luminal A (LumA), 306 as Luminal B (LumB), 102 as ERBB2 positive (HER2E), 162 as Basal-

like (BasalL) and 145 as Normal-like (NormL) (Supplementary Table 4.S1; voir Annexe C).

We used the Pearson correlation coefficient (PCC) of expression values for each pair of genes across tumour samples for each BC subtype to build the co-expression attribute (*CoE*; Figure 4.1A and B). We identified significantly co-expressed gene-pairs as pairs with an expression PCC value of r higher than a threshold, relative to the top 5% gene-pairs with the highest r values (see Methods; CoE; Data-derived attributes; Figure 4.1A and B).

Considering that our reference set of functional interactions covers pathways associated to all human organs, including several pathways that may not be activated in breast cancer cells, we assessed whether removing expected non-expressed genes from breast cancer (BC) expression data would significantly enrich the set of co-expressed genes in interactions associated with BC. Since all the male patients were removed from the mRNA expression data, we used the 30^{th} percentile expression value (1^{st} value over 0) on the Y chromosome to filter the expression data from potentially unexpressed genes per BC subtype. A gene was identified as not expressed if expression values reported for this gene were below this threshold for at least 95% of tumor samples for a given BC subtype (Figure 4.2C) (see Methods). On average 1,600 genes were considered as non-expressed for each BC subtype and removed from our expression data sets (number of non-expressed genes are below the dashed line per BC subtype; Figure 4.1C). The resulting expression datasets covered a minimum of 14,000 genes for each BC subtype.

Co-expressed gene-pairs were subsequently identified in the set of "true" functional interactions (Refset) before and after removing non-expressed genes. We also identified pairs of genes co-cited with the term "Breast cancer" in the PubMed database (February 2016) and computed the enrichment of these BC co-cited gene-pairs in the reference sets after removing pairs with non-expressed genes via a one-tailed Fisher's exact test (N = All gene pairs in Refset; Figure 4.2D). This revealed that the reference

sets for each BC subtypes, which are composed of ~5,400 gene pairs after removing "non-expressed genes", were significantly enriched in BC co-cited gene-pairs (Fisher's exact test, $P < 0.001$; Figure 4.2D).

4.3.4 Generation of protein-protein interaction-based attributes

Protein-Protein interactions (PPI) were downloaded from the BioGRID (April 2015, (Oughtred *et al.*, 2016)) and IntAct (April 2015, (Orchard *et al.*, 2014)) databases. To reduce the false positive rate of these datasets, we selected 19,928 PPI validated by at least two independent experimental methods. The resulting filtered interactions set covered a total of 7,060 proteins.

For each pair of proteins A and B, we assigned a value 1 to the binary attribute *I(A, B)* if the proteins A and B exhibited a direct PPI and 0 otherwise (Data-derived attributes; Figure 1A and B). We also computed an attribute *CI (A, B)* (Common Interactors), considering the statistical significance of the observed number of common physical interactors of the proteins encoded by A and B, in the PPI network (Data-derived attributes; Figure 1A and B). *CI* is assigned a *P*-value derived from a one-tailed Fisher's exact test (*N* = the number of genes encoding proteins that are in the considered PPI network), which is discretized into binary values: 1 for $P < 0.05$ and 0 otherwise.

4.3.5 Generation of co-disease attribute

Disease-gene annotations were downloaded from the DISEASES database (August 2015; (Pletscher-Frankild *et al.*, 2015)) and curated to minimize false positive annotations and redundancy of annotations for highly characterized diseases (see Methods). This selection identified a set of 653 annotations, which were associated with at least two genes and provided annotations for 15,116 genes.

The co-disease attribute *Di(A, B)* measured the statistical significance of the number of shared diseases annotations between two genes (A and B) via a standard Fisher's exact test (N = the number of curated diseases annotations observed for at least two genes; Data-derived attributes; Figure 4.1A and B). We discretized the resulting P-values to obtain three groups of relationships between genes based on the *Di* attribute: a group with significant enrichment of disease annotations between genes A and B (*Di (A,B)* = 1 for $P < 0.05$); a group with significant depletion of disease annotations between genes A and B (*Di(A,B)* = 0 for $P > 0.95$); and a group with no significant enrichment or depletion of disease annotations (*Di(A,B)* = 0.5).

4.3.6 Generation of network-derived attributes

To build the networks-derived attributes, a total of five networks were created for all BC subtypes. In these networks, two genes A and B are connected by an edge if they are co-expressed (significant *CoE*; Pearson correlation coefficient $\geq$ r; Table 4.1), or if their gene products exhibit a PPI (I = 1; Figure 4.1A and B). For both genes A and B, we measured the enrichment of at least one common disease annotation within their first neighbours (i.e. genes connected to gene A or B by one edge) in the above network. This enrichment was measured using a one-tailed Fisher's exact test (N = the number of genes with disease annotation). If the derived P-value from this test was less than or equal to 0.05 for A and B, and for at least one phenotype, we assigned a value of 1 to

a categorical variable *N(A, B)*, and 0 otherwise (Network-derived attributes; Figure 4.1A and B). Similarly, if A and B were both associated with the disease annotation that was also enriched in both their neighborhoods, a value of 1 was assigned to a categorical variable *NDi(A, B)*, and 0 otherwise (Network-derived attributes; Figure 4.1A and B).

4.3.7 Identification of functional interaction classes within-pathways

Using the statistical attributes described above, we identified classes of functional interactions (FIs) as previously done using *C. elegans* data (Boucher, B. *et al.*, 2016). To do so, FIs of the reference sets (Refset) were clustered hierarchically for Luminal A and Basal-like tumor subtypes based on discretized values of the six statistical attributes (CoE, Di, N, NDi, I and CI; Figure 4.3A) using Ward's agglomerative method with Euclidean distance metric. This strategy identified eight major classes of FIs within the Refset of each subtypes of BC. We named FI classes according to the combination of values observed for the statistical attributes as previously done (Boucher, B. *et al.*, 2016). Three groups of classes could be identified based on values for PPI-associated (*I* and *CI*) and Disease annotations-associated (*Di*) attributes. Classes with positive values for *I* or *CI* comprised pairs of genes that code for proteins that directly interact or that share more common PPI partners than expected by chance and are identified as PPI-centered interactions (PC; Figure 4.3A). From those that are not associated with positive values for *I* or *CI*, we identified a group of FIs involving pairs of genes which are significantly depleted in common disease annotations ($Di = 0$; Figure 4.3). In *C. elegans*, this class of FIs was suggested to play a major coordination function within the FI network (Boucher, B. *et al.*, 2016). Consequently, we called them Connectors (C; Figure 4.3). The remaining FI classes were named based on their

observed independence from PPI and were named PPI-independent interactions (PI; Figure 4.3). Next, each one of these three FIs classes was subdivided in three subclasses based on value of network-based attributes *N* and *NDi*. These attributes assess the enrichment in the neighbourhood of gene pairs of genes associated to a given Disease annotation (*N* =1; Figure 4.3B) and the association of genes of this pair with the enriched annotation (*NDi*=1; Figure 4.3B). We hypothesized that gene pairs with both positive *N* and *NDi* attributes could act in a stronger functional context than those with null values for either or both attributes. Based on this assumption, we defined FIs with *N*=1 and *NDi*=1 as displaying a strong (s) contextual strength (CS). We also identified FIs with *N*=1 and *NDi*=0 as displaying a medium (m) CS and FIs with *N*=0 and *NDi*=0 as displaying a weak (w) CS (Figure 4.3B). Overall, this classification allowed us to identify three main classes of FIs that are also subdivided into three subclasses (Figure 4.3A).

4.3.8 Optimization of the co-expression attributes to identify BC-associated gene-pairs

The attributes generated for this study are similar to those designed in our previous study in *C. elegans* as is the clustering strategy we employed in both systems to identify FI classes (Boucher, B. *et al.*, 2016). We therefore compared the range of values supporting the identification of FI classes in the two studies. All attributes, except *CoE*, were discretized and associated to binary or trinary values as detailed above. The only difference in attribute values observed between FI classes in human and *C. elegans* consists in the distribution of *CoE* values (Figure 4.4A). While most gene pairs of PI

and PC classes display *CoE* higher than 0.8 in *C. elegans*, the *CoE* of gene-pairs of corresponding classes in human were more homogenously distributed along the range of positive values (Figure 4.4A). The major difference between gene expression data used in our studies consists in the biological samples used to generate these data, one being whole animals for *C. elegans* expression data, and breast tumor tissues for human. Since within pathway FIs used to generate our training sets are involved in pathways that may not be active in breast cancer cells, we hypothesized that such pairs might display low *CoE* when considering breast cancer expression data only. To filter-out these gene-pairs from the 7,118 pairs composing the Refset (Refset; Figure 1A), we identified a threshold on *C. elegans CoE* data for each FI class as the *CoE* values for 99% of gene-pairs present in the *C. elegans* reference set (Boucher, B. *et al.*, 2016). This threshold was used to filter-out 1,718 gene pairs. We tested whether this filtering step (Filtering stages; Figure 4.1A) further enriched the Refset, composed of ~5,400 FIs, of gene-pairs reported in the literature on breast cancer (Figure 4.4B). The resulting set of gene pairs were significantly enriched in BC gene-pairs for some FI classes, depending on breast cancer subtypes (Figure 4.4B), suggesting that this filtering is beneficial to increase the specificity of the Refset towards "true" FIs involve in BC biology (Figure 4.4B).

4.3.9 Generation of a predicted network of functional interactions for Luminal A and Basal-like breast cancer

To test the potential of the integration strategy used in this study to build a predicted functional interaction networks enriched in "true" FIs, we assessed whether models considering the 9 subclasses of FIs individually allowed an efficient segregation of gene-pairs taken from the Refset ("true" FIs), from those of a negative set of FIs. This

negative set of FIs was generated through randomizations of genome-wide gene-pairs as previously done (Boucher, B. *et al.*, 2016). This strategy insures that both positive and negative training sets were composed of equivalent ratios of missing data. Leave-one-out cross-validation test was then used to evaluate the performance of logistic regression models trained for each FIs class to identify the positive set of FIs (Refset) against the negative set (see Methods). The receiver operating characteristics (ROC) at 20 equally spaced model score cut-offs ranging from 0 to 1 was calculated for each model. The true positive rates (TPR) obtained for FI class-specific models at a false positive rate (FPR) of 0.5% were compared with that of a unique model trained on the entire Refset (Table 4.2). The sum of pairs retrieved from the Refset by class-specific models for each subtype of BC (2,820 gene-pairs; 53% for Luminal A; Table 4.2) was higher than the number of gene-pairs retrieved by the models trained on the complete Refset (2,119 gene-pairs; 40% for Luminal A; Table 4.2). This suggests that the integrative strategy considering classes if FIs individually displays high performances to identify FIs from gene-pairs randomly selected from the genome and is also performing better than a general integrative strategy, ignoring intrinsic differences existing between FIs classes in human, as already shown in *C. elegans* (Boucher, B. *et al.*, 2016).

Considering the predictive power of our models, we applied the integration strategy described above to pairs of genes identified in the human genome in orderÀ to build a genome-wide predicted functional interactions networks for Luminal A and Basal-like breast cancer. From the 300 million possible pairs of genes, 70 million pairs with expression and disease annotations were retained (Genome-wide set; Figure 4.1A). From this list, approximately 30 million gene-pairs were filtered out due to their involvement of non-expressed genes or their associated *CoE* values inferior to the filtering threshold identified for each FI class (Filtering stages; Figure 4.1A). Networks were then built for Luminal A and Basal-like breast tumours and were composed of

10,947 genes and approximately 40 million of interactions subdivided into three main classes and 9 sub-classes of FIs.

4.3.10 Rewiring of FIs between subtypes of breast tumors

A significant amount of rewiring was observed between predicted breast cancer FI networks. Interestingly, this rewiring occurred within FI classes, which tend to change contextual strength (CS) from one network to another (Figure 4.5A). For instance, while 41,8% of FIs are rewired/plastic in the Refset, only 3,65% FIs change classes (between PI and C). The remaining 38.2% change CS within classes. Interestingly, the three classes of FIs (PI, PC and C) did not display the same plasticity level. Indeed, only C interactions had a significant enrichment of plastic interactions between Basal-like and luminal A subtypes of BC when considering genome-wide predictive networks (Fisher's exact test, P-value < 10^{-200}; Figure 4.5B). PI and PC classes were significantly depleted in plastic interactions (Fisher's exact test, P-value < 10^{-200}; Figure 4.5B). These data suggest that FI classes display distinctive plasticity properties.

4.3.11 Warburg effect genes display distinctive properties within predicted breast cancer FI networks

The Warburg effect is a metabolic reprogramming of cells. Recent data showed that this effect was stronger in Basal-like when compared to luminal A breast tumours (Brauer *et al.*, 2013; Palaskas *et al.*, 2011). We consequently assessed whether a set of 28 genes, shown to control this reprogramming (Gogvadze *et al.*, 2010; Hensley *et al.*,

2013; Lunt et Vander Heiden, 2011; Samudio *et al.*, 2009; Semenza, 2013; White, 2013), displayed distinctive properties within predicted FI networks built for Luminal A and basal-like tumours.

To do so, we identified indirect interactions linking a set of genes identified in the glycolysis machinery (Supplementary Table 4.S2; voir Annexe D) to genes associated with the OXPHOS (Supplementary Table 4.S3; voir Annexe E) within predicted BC FI networks. This analysis identified a network involving 10,450 genes, including 23 of the 28 Warburg associated genes (training set). We assessed, for each one of these genes, the distribution of PI, PC and C interactions in Luminal A and Basal-like tumours (Figure 4.6A and C). We also assessed the percentage of plastic interactions linking these genes with glycolysis and OXPHOS genes (Figure 4.6B and C). This revealed that genes of this training set displayed slightly, but significantly less plastic interactions than the other genes interacting with both glycolysis and OXPHOS genes (Student's *t*-test, *P*-value = 0.043; Figure 4.6C). It also revealed that these genes were interacting with glycolysis and OXPHOS genes through a higher rate of PI and a lower rate of C than the other genes of this indirect interaction network (Student's *t*-test, *P*-value < 0.0001; Figure 4.6C). The significant enrichment of PI and depletion of C between genes controlling the Warburg effect and their partners in the glycolysis and OXPHOS machinery is intriguing. Previous study characterizing the genetic interaction network of *C. elegans* suggested that genes linked through C interactions to their partners may play major coordination role within functional interaction networks (Boucher, B. *et al.*, 2016). Comparing the structural function of PI, PC and C classes of functional interaction in human breast cancer and *C. elegans* revealed that in both systems C and to a lesser extent PI interactions play a critical structural function at the intersection of functional modules (see Supplementary data (voir Annexe B); Supplementary Figure 4.S1).

4.3.12 Generation and *in silico* validations of predictive models for Warburg effect genes

Gene interacting with both glycolysis and OXPHOS genes (10,450 genes) are identified and called the indirect binding set. We used the distinctive properties of the training set of 23 Warburg effect genes (training set) to generate an unweighted linear model providing each gene *A* of the indirect binding set with a value *V(A)* such as :

$V(A) = PI_{lum}(A) + PI_{bas}(A) - Plastic(A) - C_{lum}(A) - C_{bas}(A)$; with $PI_{lum}(A)$,

and $PI_{bas}(A)$ being the percentage of PI interactions linking the gene A with Glycolysis and OXPHOS genes in Luminal A and Basal-like predicted FI networks respectively; $C_{lum}(A)$ and $C_{bas}(A)$ being the percentage of C interactions linking the gene A with Glycolysis and OXPHOS genes in Luminal A and Basal-like predicted FI networks respectively and *Plastic(A)* being the percentage of plastic interactions linking gene A with glycolysis and OXPHOS genes.

Groups of genes associated with the top 25%, 5%, 1%, 0.5% and 0.1% *V(A)* values of the indirect binding set were identified. The percentages of the training set genes found in these groups were represented with respect to the percentage of the indirect binding set also found in each group (Unweighted model; orange; Figure 4.7A). The expected percentage of genes from training set picked when randomly picking genes from the indirect binding set is also indicated (Random; black curve; Figure 4.7A). Fold enrichment of training set genes over indirect binding set and the number of all sets of genes contained per group were also computed (Figure 4.7D and E respectively). This analysis revealed that the unweighted linear model built in a supervised manner on the distinctive characteristics of the training set enables the identification of genes from this training set with high efficiency from the indirect binding set (up to 69-fold enrichment for the top 0.1% *V(A)* value group; Figure 4.7D).

To further test the performance of this model to identify genes controlling the Warburg effect, we identified testing sets of genes. To do so, identified human genes co-cited with the "Warburg effect" term in the PubMed database, from which we removed the 28 genes from our training set. We also divided these genes into two sets: a set containing genes co-cited with the "Warburg effect" term in at least one article (testing set 1), and a set of genes co-cited in at least two articles (testing set 2). We measured the percentage of testing sets 1 and 2 (orange lines; Figure 4.7B and C respectively) with respect to the percentage of indirect binding set identified for *V(A)* values within the top 25% to 0.1%. We also measured the enrichment of testing set in groups of genes with *V(A)* values within the top 25% to 0.1% (Figure 4.7D) and the number of genes of each set per group (Figure 4.7E). This revealed that the unweighted linear model built in this study displayed similar performances to identify genes from the training and testing sets. It also revealed that the highest performance of the model was observed at a threshold corresponding to the top 1% V(A) values which display 35.4-, 17.4- and 29.4-fold enrichment of the training, testing sets 1 and 2 respectively when compared to the actual representation of these sets genome-wide (Figure 4.7D). At this threshold, 108 genes were selected from the indirect binding set, 71 of them being genes never associated to the Warburg effect in the literature.

4.3.13 Model built on integrative genomic approaches outperforms prediction of Warburg effect genes driven from expression data only

One hallmark of genes associated to the Warburg effect is a reduced level of plasticity (Figure 4.6C). As discussed in the supplementary data section, very low plasticity levels are detrimental for the predictions, suggesting that interaction between Warburg effect genes and glycolysis/OXPHOS genes still display a certain level of rewiring

(Supplementary Data (voir Annexe B); Supplementary Figure 4.S2). Other Hallmarks of Warburg effect genes revealed by our study consist in an elevated level of PI and a low level of C interactions linking these genes to glycolysis and OXPHOS associated genes (Figure 4.6C). PI interactions are characterized by high co-expression of interacting genes as well as a significant enrichment of common disease annotations (Figure 4.2A), while C interactions contain gene pairs that are characterized by a lower co-expression level and a significant depletion of common disease annotations (Figure 4.2A). We tested whether the co-expression of training and testing set genes with Glycolysis and OXPHOS genes alone would support the predictions obtained using the unweighted linear model built in this study. To do so, we used the *CoE* attribute to identify genes significantly co-expressed with Glycolysis and OXPHOS genes. Co-expressed genes were ranked based on their co-expression level and the number of glycolysis and OXPHOS genes they were significantly co-expressed with. To do so, each gene A was associated a *CoESum(A)* value such as:

$$CoESum(A) = \sum_{i=1}^{n} (1 - CoE(Pval)_i)$$

where n is the number of genes i of the glycolysis or OXPHOS machineries significantly co-expressed with the gene A, and *Pval* is the *P*-value lower than 0.05 supporting the significance of the co-expression (see Methods).

Groups of genes associated to the top 25%, 5% and 1% of *CoESum* for the Basal-like and Luminal A subtype of tumors were identified. The enrichment of training and testing set 1 and 2 over respective representation of these sets genome-wide were measured in each group of genes (Figure 4.8A and B). As expected, considering that the Warburg effect has been shown to be higher in Basal-like than Luminal A BC subtypes, a higher enrichment is observed for training and testing sets in groups of genes co-expressed with glycolysis and OXPHOS in Basal-like when compared to Luminal A tumours (compare top 1%, Figure 4.8A and B). Enrichment of training and

testing sets were, however, more than 6-fold lesser to those obtained when using V(A) value from the unweighted linear model resulting from the integrative genomic strategy described in this study (compare top 1%, Figure 4.8A-B and Figure 4.7D).

We also measured the overlap of predictions obtained using co-expression data and unweighted linear model. This overlap was null when considering groups of genes identified with a threshold corresponding to the top 1% *V(A)* and *CoESum(A)* values. They were lower than 0.4% when using threshold corresponding to top 5% *V(A)* and *CoESum(A)* (Figure 4.8C and D). This suggests that predictions obtained by the unweighted linear model generated in this study do not rely exclusively on the co-expression attribute.

Co-expression is a commonly used metric to identify genes sharing similar functions than a group of seed genes (Lee, H. K. *et al.*, 2004; van Noort *et al.*, 2003). Therefore, we compared the performance of the unweighted model generated in this study with a classical approach and identified genes significantly co-expressed with genes in the training set using the *CoESum* metrics described above. Enrichments of genes in testing sets 1 and 2, when compared to their respective representation genome-wide, were measured in groups of genes co-expressed with the training set and displaying top 25%, 5% and 1% of *CoESum* values (Figure 4.8A and B). A maximum of 4-fold enrichment of testing sets was reached for genes co-expressed with the training set (top 1%; Figure 4.8B) versus a maximum of 29.4-fold enrichment when using the unweighted model (top 1%; Figure 4.7D). The overlap between predictions based on co-expression with the training set and the unweighted model is also lower than 1% (Figure 4.8C and D).

Taken together, these data reveal that the predictive model, generated through the integrative genomic strategy described in this study, outperforms predictions of Warburg effect genes using co-expression data only. It also suggests that predictions using the two methods are complementary.

4.4 Discussion

Metabolic reprogramming constitutes a hallmark of cancer development. Recent studies showed that this reprogramming controls tumour progression and metastasis. This identifies genes controlling the increase of aerobic glycolysis in cancer cells, the Warburg effect, as potent anti-tumoral and metastatic targets. In this study, we developed an integrative genomics strategy that involves the generation of predictive functional interaction networks for Luminal A and Basal-like subtypes of breast cancer and the identification of distinctive feature of genes shown to control the Warburg effect in these networks. These distinctive features were used to build an unweighted linear model. The performance of this model to identify genes associated the Warburg effect in the literature was tested and was shown to be good and to outperform predictions made using expression data only.

We consider the strategy used to generate the predictive model described in this study as preliminary. Several unsupervised modeling approaches, such as linear regression, naive Bayesian or random forest, need to be evaluated to generate a robust model to be used prior to experimental validation.

The study presented here identifies the Warburg effect genes as genes that tend to relate with genes of the glycolysis and the OXPHOS machineries through PPI-independent interactions (PI) and through interactions that are on average less plastic than expected. These results are surprising when considering our recent characterization of *C. elegans* genetic interactome which suggested that genes linked through "connector" interactions are critical coordinator of biological processes. This point is discussed in further detail in the next chapter.

In addition to provide a better understanding of the relative positioning of Warburg effect genes in the functional interaction network with respect to glycolysis and

OXPHOS genes, this study also highlights the utility of integrative genomics and interactomic approaches to identify therapeutic targets for human diseases.

It is important to note that the integrative genomics strategy developed in this study uses expression data which are tissue specific (or tumor specific) and PPI data and Disease annotations that are relevant to any tissue and human pathologies. The integration strategy described here could consequently be transposed to any human disease model and tissue for which expression data are available. The superior performance of that method, when compared to the identification of co-expressed genes alone, suggests that this method may highly improve predictions of genes with a given function or co-functionality of genes when there are only expression data available. This tool would be particularly useful for pathologies with limited access to enough biological material to generate proteogenomics studies.

4.5 Methods

4.5.1Generation of the co-expression attribute

Expression data (Level 3 TCGA RNAseqV2) from 1,058 breast invasive carcinoma were downloaded from the TCGA Data Portal (February 2016; (The Cancer Genome Atlas Research *et al.*, 2013)). Twelve samples from male patients were removed and the resulting data collection from 1,046 tumors were classified into five breast cancer (BC) subtypes (Luminal A, Luminal B, HER2, Basal and Normal) using the AIMS tool (Paquet et Hallett, 2015). Considering that the minimum posterior probability of a correctly classified sample was at least 0.90 for each predicted subtype, we could classify 1,036 tumors from their associated expression data as followed: 321 tumors were classified as Luminal A (LumA), 306 as Luminal B (LumB), 102 as ERBB2

positive (HER2E), 162 as Basal-like (BasalL) and 155 as Normal-like (NormL) (Supplementary Table 4.S1; voir Annexe C).

To remove "non-expressed genes" from breast expression data based on expression level of the Y chromosome, a RNA-Seq by Expectation-Maximization value (RSEM) of 0.3 was considered as the minimal expression threshold per sample and gene. Similar strategies have been successfully used in the past to filter out unexpressed genes quantified by Reads Per Kilobase of transcript per Million mapped reads (Łabaj *et al.*, 2011; Ramsköld *et al.*, 2009; Sam *et al.*, 2011). For each BC subtype, a gene was identified as not expressed if at least 95% of its expression values (Figure 4.2C) were below this threshold (0.3 RSEM).

We computed the Pearson correlation of expression values for each pair of genes across tumour samples from each BC subtype. We then built the Co-expression (CoE) attribute using P-values relative to the empirically estimated probability distribution of correlations for all gene pairs fitted on a normal distribution. Correlations greater than r and characterized by a P-value<0.05 (one-tailed probability) were indicative for significantly co-expressed genes (Table 4.1; CoE; Data-derived attributes; Figure 4.1A and B).

4.5.2 Co-disease attributes

Disease-gene annotations were downloaded from the DISEASES database (August 2015; (Pletscher-Frankild *et al.*, 2015)). Disease annotations of genes were collected from the literature through text-mining and manual curation or experimental evidences (from COSMIC and DistiLD database; (Forbes *et al.*, 2008 and Pallejà *et al.*, 2012)). We eliminated from this dataset the associations retrieved through text-mining and

experimental evidences with less than two stars confidence scores. We then did a manual curation of disease annotations using the Disease Ontology tool (Kibbe *et al.*, 2015). The goal of this curation was to remove redundancy of annotations for highly characterized diseases. We tested the efficiency of our manual curation of Disease annotations through computation of the number of genes overlapping between two disease annotations (277,885 annotation pairs) and found only 0.1% of these combinations have more than 30% of genes in common (Supplementary Table 4.S4; voir Annexe F). From the resulting 746 curated disease annotations, we selected 653 annotations which were associated with at least two genes, and provided annotations for 15,116 genes.

4.5.3 Leave-one-out cross validation on trained models

Leave-one-out cross-validation test was used to evaluate the performance of logistic regression models trained for each FIs class to identify the positive set of FIs (Refset) against the negative set. The receiver operating characteristics (ROC) at 20 equally spaced model score cut-offs from 0 to 1 was calculated for each model. The true positive rates (TPR) obtained for FI class-specific models at a false positive rate (FPR) of 0.5% were compared with that of a unique model trained on the entire Refset (Table 4.2). The 0.5% FPR was chosen as previously done (Boucher, B. *et al.*, 2016; Lee, A. Y. *et al.*, 2010), as probability of finding in yeast a genetic interaction by chance within a set of gene-pair randomly picked from the genome (Tong *et al.*, 2004).

4.6 References

Blick, T., Widodo, E., Hugo, H., Waltham, M., Lenburg, M.E., Neve, R.M. et Thompson, E.W. (2008). Epithelial mesenchymal transition traits in human breast cancer cell lines. *Clin Exp Metastasis, 25*(6), 629-642. doi: 10.1007/s10585-008-9170-6

Boucher, B. et Jenna, S. (2013). Genetic interaction networks: better understand to better predict. *Frontiers in Genetics, 4*, 290. doi: 10.3389/fgene.2013.00290

Boucher, B., Lee, A.Y., Hallett, M. et Jenna, S. (2016). Structural and Functional Characterization of a Caenorhabditis elegans Genetic Interaction Network within Pathways. *PLoS Comput Biol, 12*(2), e1004738. doi: 10.1371/journal.pcbi.1004738

Brauer, H.A., Makowski, L., Hoadley, K.A., Casbas-Hernandez, P., Lang, L.J., Romàn-Pèrez, E., D'Arcy, M., Freemerman, A.J., Perou, C.M. et Troester, M.A. (2013). Impact of tumor and epithelial phenotypes on metabolism in breast cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research, 19*(3), 571-585. doi: 10.1158/1078-0432.ccr-12-2123

Cairns, R.A., Harris, I.S. et Mak, T.W. (2011). Regulation of cancer cell metabolism. *Nat Rev Cancer, 11*(2), 85-95. doi: 10.1038/nrc2981

DeBerardinis, R.J., Lum, J.J., Hatzivassiliou, G. et Thompson, C.B. (2008). The Biology of Cancer: Metabolic Reprogramming Fuels Cell Growth and Proliferation. *Cell Metabolism, 7*(1), 11-20. doi: http://dx.doi.org/10.1016/j.cmet.2007.10.002

DeBerardinis, R.J., Sayed, N., Ditsworth, D. et Thompson, C.B. (2008). Brick by brick: metabolism and tumor cell growth. *Current Opinion in Genetics & Development, 18*(1), 54-61. doi: http://dx.doi.org/10.1016/j.gde.2008.02.003

Dupuy, F., Tabaries, S., Andrzejewski, S., Dong, Z., Blagih, J., Annis, M.G., Omeroglu, A., Gao, D., Leung, S., Amir, E., Clemons, M., Aguilar-Mahecha, A., Basik, M., Vincent, E.E., St-Pierre, J., Jones, R.G. et Siegel, P.M. (2015). PDK1-Dependent Metabolic Reprogramming Dictates Metastatic Potential in Breast Cancer. [Research Support, Non-U.S. Gov't]. *Cell Metab, 22*(4), 577-589. doi: 10.1016/j.cmet.2015.08.007

Forbes, S.A., Bhamra, G., Bamford, S., Dawson, E., Kok, C., Clements, J., Menzies, A., Teague, J.W., Futreal, P.A. et Stratton, M.R. (2008). The Catalogue of Somatic Mutations in Cancer (COSMIC). *Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.], CHAPTER*, Unit-10.11. doi: 10.1002/0471142905.hg1011s57

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C. et Jensen, L.J. (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research, 41*(Database issue), D808-D815. doi: 10.1093/nar/gks1094

Gillies, R.J., Robey, I. et Gatenby, R.A. (2008). Causes and consequences of increased glucose metabolism of cancers. *J Nucl Med, 49 Suppl 2*, 24S-42S. doi: 10.2967/jnumed.107.047258

Gogvadze, V., Zhivotovsky, B. et Orrenius, S. (2010). The Warburg effect and mitochondrial stability in cancer cells. *Mol Aspects Med, 31*(1), 60-74. doi: 10.1016/j.mam.2009.12.004

Guppy, M., Leedman, P., Zu, X. et Russell, V. (2002). Contribution by different fuels and metabolic pathways to the total ATP turnover of proliferating MCF-7 breast cancer cells. [Article]. *Biochemical Journal, 364*(1), 309-315.

Hensley, C.T., Wasti, A.T. et DeBerardinis, R.J. (2013). Glutamine and cancer: cell biology, physiology, and clinical opportunities. *The Journal of Clinical Investigation, 123*(9), 3678-3684. doi: 10.1172/JCI69600

Hinkle, P.C. (2005). P/O ratios of mitochondrial oxidative phosphorylation. *Biochim Biophys Acta, 1706*(1-2), 1-11. doi: 10.1016/j.bbabio.2004.09.004

Karadeniz, İ., Hur, J., He, Y. et Özgür, A. (2015). Literature Mining and Ontology based Analysis of Host-Brucella Gene–Gene Interaction Network. *Frontiers in Microbiology, 6*, 1386. doi: 10.3389/fmicb.2015.01386

Kibbe, W.A., Arze, C., Felix, V., Mitraka, E., Bolton, E., Fu, G., Mungall, C.J., Binder, J.X., Malone, J., Vasant, D., Parkinson, H. et Schriml, L.M. (2015). Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Research, 43*(Database issue), D1071-D1078. doi: 10.1093/nar/gku1011

Kutmon, M., Riutta, A., Nunes, N., Hanspers, K., Willighagen, Egon L., Bohler, A., Mélius, J., Waagmeester, A., Sinha, Sravanthi R., Miller, R., Coort, S.L., Cirillo, E., Smeets, B., Evelo, Chris T. et Pico, A.R. (2016). WikiPathways: capturing the full diversity of pathway knowledge. *Nucleic Acids Research, 44*(Database issue), D488-D494. doi: 10.1093/nar/gkv1024

Łabaj, P.P., Leparc, G.G., Linggi, B.E., Markillie, L.M., Wiley, H.S. et Kreil, D.P. (2011). Characterization and improvement of RNA-Seq precision in quantitative

transcript expression profiling. *Bioinformatics, 27*(13), i383-i391. doi: 10.1093/bioinformatics/btr247

Lee, A.Y., Perreault, R., Harel, S., Boulier, E.L., Suderman, M., Hallett, M. et Jenna, S. (2010). Searching for signaling balance through the identification of genetic interactors of the Rab guanine-nucleotide dissociation inhibitor gdi-1. *PLoS One, 5*(5), e10624. doi: 10.1371/journal.pone.0010624

Lee, H.K., Hsu, A.K., Sajdak, J., Qin, J. et Pavlidis, P. (2004). Coexpression Analysis of Human Genes Across Many Microarray Data Sets. *Genome Research, 14*(6), 1085-1094. doi: 10.1101/gr.1910904

Lee, I., Lehner, B., Crombie, C., Wong, W., Fraser, A.G. et Marcotte, E.M. (2008). A single gene network accurately predicts phenotypic effects of gene perturbation in Caenorhabditis elegans. *Nat Genet, 40*(2), 181-188. doi: 10.1038/ng.2007.70

Lee, I., Lehner, B., Vavouri, T., Shin, J., Fraser, A.G. et Marcotte, E.M. (2010). Predicting genetic modifier loci using functional gene networks. *Genome Research, 20*(8), 1143-1153. doi: 10.1101/gr.102749.109

Lincet, H. et Icard, P. (2015). How do glycolytic enzymes favour cancer cell proliferation by nonmetabolic functions[quest]. [Review]. *Oncogene, 34*(29), 3751-3759. doi: 10.1038/onc.2014.320

Lunt, S.Y. et Vander Heiden, M.G. (2011). Aerobic glycolysis: meeting the metabolic requirements of cell proliferation. *Annu Rev Cell Dev Biol, 27*, 441-464. doi: 10.1146/annurev-cellbio-092910-154237

Malinowski, D.P. (2007). Multiple biomarkers in molecular oncology. II. Molecular diagnostics applications in breast cancer management. *Expert Rev Mol Diagn, 7*(3), 269-280. doi: 10.1586/14737159.7.3.269

Martin, M., Beauvoit, B., Voisin, P.J., Canioni, P., Guérin, B. et Rigoulet, M. (1998). Energetic and morphological plasticity of C6 glioma cells grown on 3-D support; Effect of transient glutamine deprivation. [Article]. *Journal of Bioenergetics and Biomembranes, 30*(6), 565-578.

Mertins, P., Mani, D.R., Ruggles, K.V., Gillette, M.A., Clauser, K.R., Wang, P., Wang, X., Qiao, J.W., Cao, S., Petralia, F., Kawaler, E., Mundt, F., Krug, K., Tu, Z., Lei, J.T., Gatza, M.L., Wilkerson, M., Perou, C.M., Yellapantula, V., Huang, K.-l., Lin, C., McLellan, M.D., Yan, P., Davies, S.R., Townsend, R.R., Skates, S.J., Wang, J., Zhang, B., Kinsinger, C.R., Mesri, M., Rodriguez, H., Ding, L., Paulovich, A.G., Fenyö, D., Ellis, M.J., Carr, S.A. et Nci, C. (2016). Proteogenomics connects somatic mutations

to signalling in breast cancer. [Article]. *Nature, 534*(7605), 55-62. doi: 10.1038/nature18003

Mocellin, S. et Rossi, C.R. (2007). Principles of gene microarray data analysis. *Adv Exp Med Biol, 593*, 19-30. doi: 10.1007/978-0-387-39978-2_3 Récupéré de http://link.springer.com/chapter/10.1007%2F978-0-387-39978-2_3

Normanno, N., De Luca, A., Carotenuto, P., Lamura, L., Oliva, I. et D'Alessio, A. (2009). Prognostic applications of gene expression signatures in breast cancer. *Oncology, 77 Suppl 1*, 2-8. doi: 10.1159/000258489

Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N.H., Chavali, G., Chen, C., del-Toro, N., Duesbury, M., Dumousseau, M., Galeota, E., Hinz, U., Iannuccelli, M., Jagannathan, S., Jimenez, R., Khadake, J., Lagreid, A., Licata, L., Lovering, R.C., Meldal, B., Melidoni, A.N., Milagros, M., Peluso, D., Perfetto, L., Porras, P., Raghunath, A., Ricard-Blum, S., Roechert, B., Stutz, A., Tognolli, M., van Roey, K., Cesareni, G. et Hermjakob, H. (2014). The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic acids research, 42*(Database issue), D358-363. doi: 10.1093/nar/gkt1115

Oughtred, R., Chatr-aryamontri, A., Breitkreutz, B.J., Chang, C.S., Rust, J.M., Theesfeld, C.L., Heinicke, S., Breitkreutz, A., Chen, D., Hirschman, J., Kolas, N., Livstone, M.S., Nixon, J., O'Donnell, L., Ramage, L., Winter, A., Reguly, T., Sellam, A., Stark, C., Boucher, L., Dolinski, K. et Tyers, M. (2016). BioGRID: A Resource for Studying Biological Interactions in Yeast. *Cold Spring Harb Protoc, 2016*(1), pdb top080754. doi: 10.1101/pdb.top080754

Palaskas, N., Larson, S.M., Schultz, N., Komisopoulou, E., Wong, J., Rohle, D., Campos, C., Yannuzzi, N., Osborne, J.R., Linkov, I., Kastenhuber, E.R., Taschereau, R., Plaisier, S.B., Tran, C., Heguy, A., Wu, H., Sander, C., Phelps, M.E., Brennan, C., Port, E., Huse, J.T., Graeber, T.G. et Mellinghoff, I.K. (2011). 18F-fluorodeoxy-glucose positron emission tomography (18FDG-PET) marks MYC-overexpressing human basal-like breast cancers. *Cancer research, 71*(15), 5164-5174. doi: 10.1158/0008-5472.can-10-4633

Pallejà, A., Horn, H., Eliasson, S. et Jensen, L.J. (2012). DistiLD Database: diseases and traits in linkage disequilibrium blocks. *Nucleic Acids Research, 40*(Database issue), D1036-D1040. doi: 10.1093/nar/gkr899

Paquet, E.R. et Hallett, M.T. (2015). Absolute assignment of breast cancer intrinsic molecular subtype. *J Natl Cancer Inst, 107*(1), 357. doi: 10.1093/jnci/dju357

Pasdois, P., Deveaud, C., Voisin, P., Bouchaud, V., Rigoulet, M. et Beauvoit, B. (2003). Contribution of the Phosphorylable Complex I in the Growth Phase-Dependent Respiration of C6 Glioma Cells in Vitro. [Article]. *Journal of Bioenergetics and Biomembranes, 35*(5), 439-450. doi: 10.1023/a:1027391831382

Pletscher-Frankild, S., Palleja, A., Tsafou, K., Binder, J.X. et Jensen, L.J. (2015). DISEASES: text mining and data integration of disease-gene associations. *Methods, 74*, 83-89. doi: 10.1016/j.ymeth.2014.11.020

Rakha, E.A., El-Sayed, M.E., Reis-Filho, J. et Ellis, I.O. (2009). Patho-biological aspects of basal-like breast cancer. *Breast Cancer Res Treat, 113*(3), 411-422. doi: 10.1007/s10549-008-9952-1

Ramsköld, D., Wang, E.T., Burge, C.B. et Sandberg, R. (2009). An Abundance of Ubiquitously Expressed Genes Revealed by Tissue Transcriptome Sequence Data. *PLOS Computational Biology, 5*(12), e1000598. doi: 10.1371/journal.pcbi.1000598

Rodríguez-Enríquez, S., Vital-González, P.A., Flores-Rodríguez, F.L., Marín-Hernández, A., Ruiz-Azuara, L. et Moreno-Sánchez, R. (2006). Control of cellular proliferation by modulation of oxidative phosphorylation in human and rodent fast-growing tumor cells. [Article]. *Toxicology and Applied Pharmacology, 215*(2), 208-217. doi: 10.1016/j.taap.2006.02.005

Sam, L.T., Lipson, D., Raz, T., Cao, X., Thompson, J., Milos, P.M., Robinson, D., Chinnaiyan, A.M., Kumar-Sinha, C. et Maher, C.A. (2011). A Comparison of Single Molecule and Amplification Based Sequencing of Cancer Transcriptomes. *PLOS ONE, 6*(3), e17305. doi: 10.1371/journal.pone.0017305

Samudio, I., Fiegl, M. et Andreeff, M. (2009). Mitochondrial uncoupling and the Warburg effect: Molecular basis for the reprogramming of cancer cell metabolism. *Cancer research, 69*(6), 10.1158/0008-5472.CAN-1108-3722. doi: 10.1158/0008-5472.CAN-08-3722

Semenza, G.L. (2013). HIF-1 mediates metabolic responses to intratumoral hypoxia and oncogenic mutations. *J Clin Invest, 123*(9), 3664-3671. doi: 10.1172/jci67230

Sintupisut, N., Liu, P.-L. et Yeang, C.-H. (2013). An integrative characterization of recurrent molecular aberrations in glioblastoma genomes. *Nucleic Acids Research, 41*(19), 8803-8821. doi: 10.1093/nar/gkt656

The Cancer Genome Atlas Research, N., Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C. et Stuart, J.M.

(2013). The Cancer Genome Atlas Pan-Cancer analysis project. [Commentary]. *Nat Genet, 45*(10), 1113-1120. doi: 10.1038/ng.2764

Tian, F., Wang, Y., Seiler, M. et Hu, Z. (2014). Functional characterization of breast cancer using pathway profiles. *BMC Medical Genomics, 7*(1), 45. doi: 10.1186/1755-8794-7-45

Tong, A.H., Lesage, G., Bader, G.D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G.F., Brost, R.L., Chang, M., Chen, Y., Cheng, X., Chua, G., Friesen, H., Goldberg, D.S., Haynes, J., Humphries, C., He, G., Hussein, S., Ke, L., Krogan, N., Li, Z., Levinson, J.N., Lu, H., Menard, P., Munyana, C., Parsons, A.B., Ryan, O., Tonikian, R., Roberts, T., Sdicu, A.M., Shapiro, J., Sheikh, B., Suter, B., Wong, S.L., Zhang, L.V., Zhu, H., Burd, C.G., Munro, S., Sander, C., Rine, J., Greenblatt, J., Peter, M., Bretscher, A., Bell, G., Roth, F.P., Brown, G.W., Andrews, B., Bussey, H. et Boone, C. (2004). Global mapping of the yeast genetic interaction network. *Science, 303*(5659), 808-813. doi: 10.1126/science.1091317

van Noort, V., Snel, B. et Huynen, M.A. (2003). Predicting gene function by conserved co-expression. *Trends Genet, 19*(5), 238-242. doi: 10.1016/s0168-9525(03)00056-8

Ward, Patrick S. et Thompson, Craig B. (2012). Metabolic Reprogramming: A Cancer Hallmark Even Warburg Did Not Anticipate. *Cancer Cell, 21*(3), 297-308. doi: http://dx.doi.org/10.1016/j.ccr.2012.02.014

White, E. (2013). Exploiting the bad eating habits of Ras-driven cancers. *Genes Dev, 27*(19), 2065-2071. doi: 10.1101/gad.228122.113

Table 4.1: Pearson correlation coefficient thresholds.

Subtype	*r*
LumA	0.32
LumB	0.24
HER2E	0.28
BasalL	0.26
NormL	0.44

Table 4.2: The true positive rates obtained for functional interactions class-specific models at a given false positive rate.

	Class	FPR	Threshold	TPR	nb. of pairs
LumA	All	0.005	0.95	0.4	2119
	PIs	0.005	0.95	0.6	728
	PIm	0.005	0.95	0.58	385
	PIw	0.005	0.95	0.55	251
	Cs	0.005	0.95	0.5	165
	Cm	0.005	0.9	0.48	69
	Cw	0.005	0.9	0.43	58
	PCs	0.005	0.95	0.51	614
	PCm	0.005	0.95	0.49	424
	PCw	0.005	0.9	0.44	126
BasalL	All	0.005	0.95	0.41	2201
	PIs	0.005	0.95	0.61	706
	PIm	0.005	0.95	0.58	359
	PIw	0.005	0.95	0.54	252
	Cs	0.005	0.95	0.51	176
	Cm	0.005	0.9	0.47	72
	Cw	0.005	0.9	0.42	56
	PCs	0.005	0.95	0.52	628
	PCm	0.005	0.95	0.48	373
	PCw	0.005	0.9	0.42	156

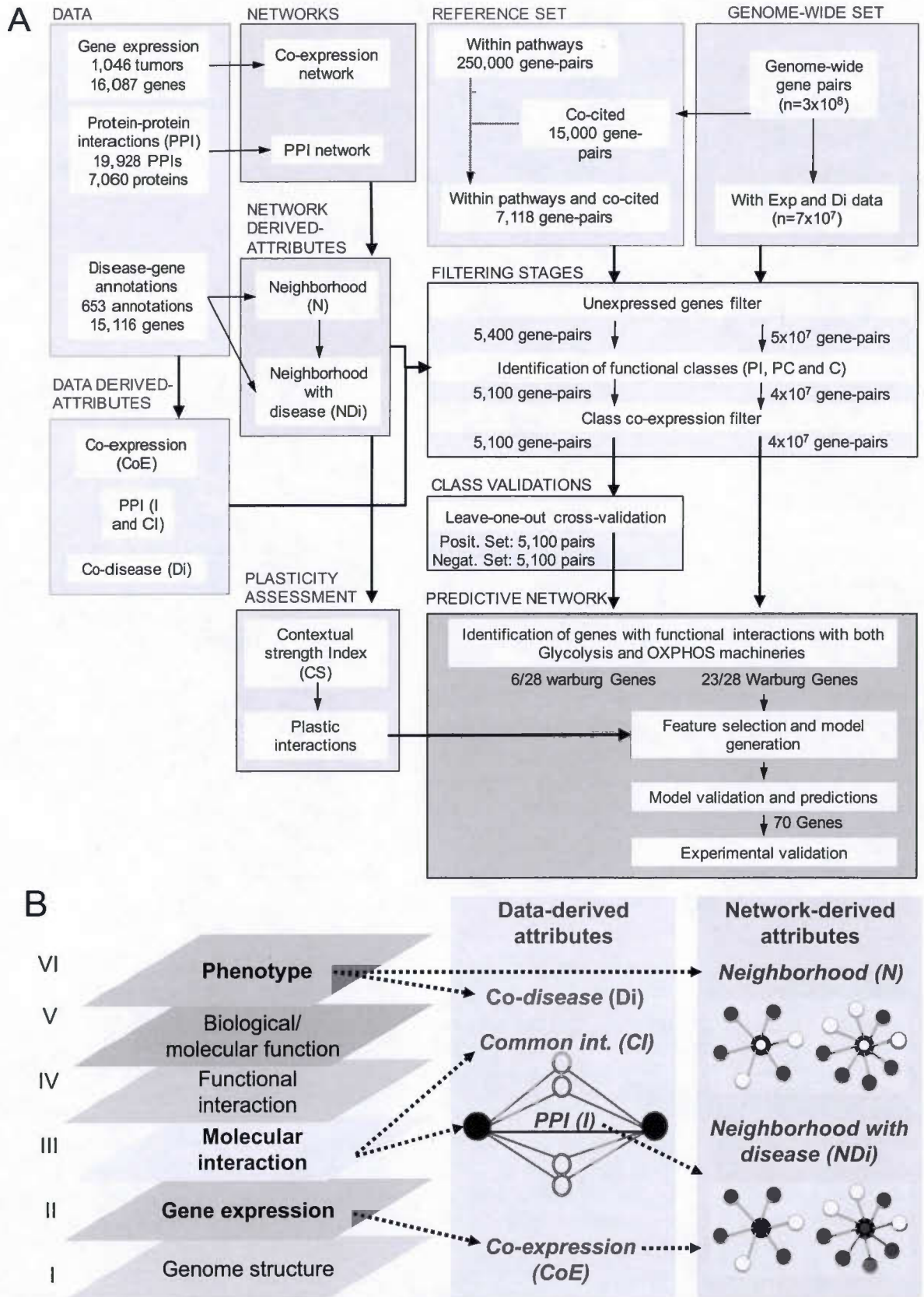


Figure 4.1. Workflow of the integration strategy and description of the statistical attributes.

(A) Workflow of the integration strategy is divided is group of tasks and chronological dependencies are indicated by arrows. **(B)** Statistical attributes are derived from either single type of data (*CoE*, *D*, *I* and *CI*) or from network built on several kinds of data (*N*, *NDi*). Each attribute is built with data corresponding (arrows) to the different abstraction levels of biological systems (see (Boucher, B. et Jenna, 2013)).

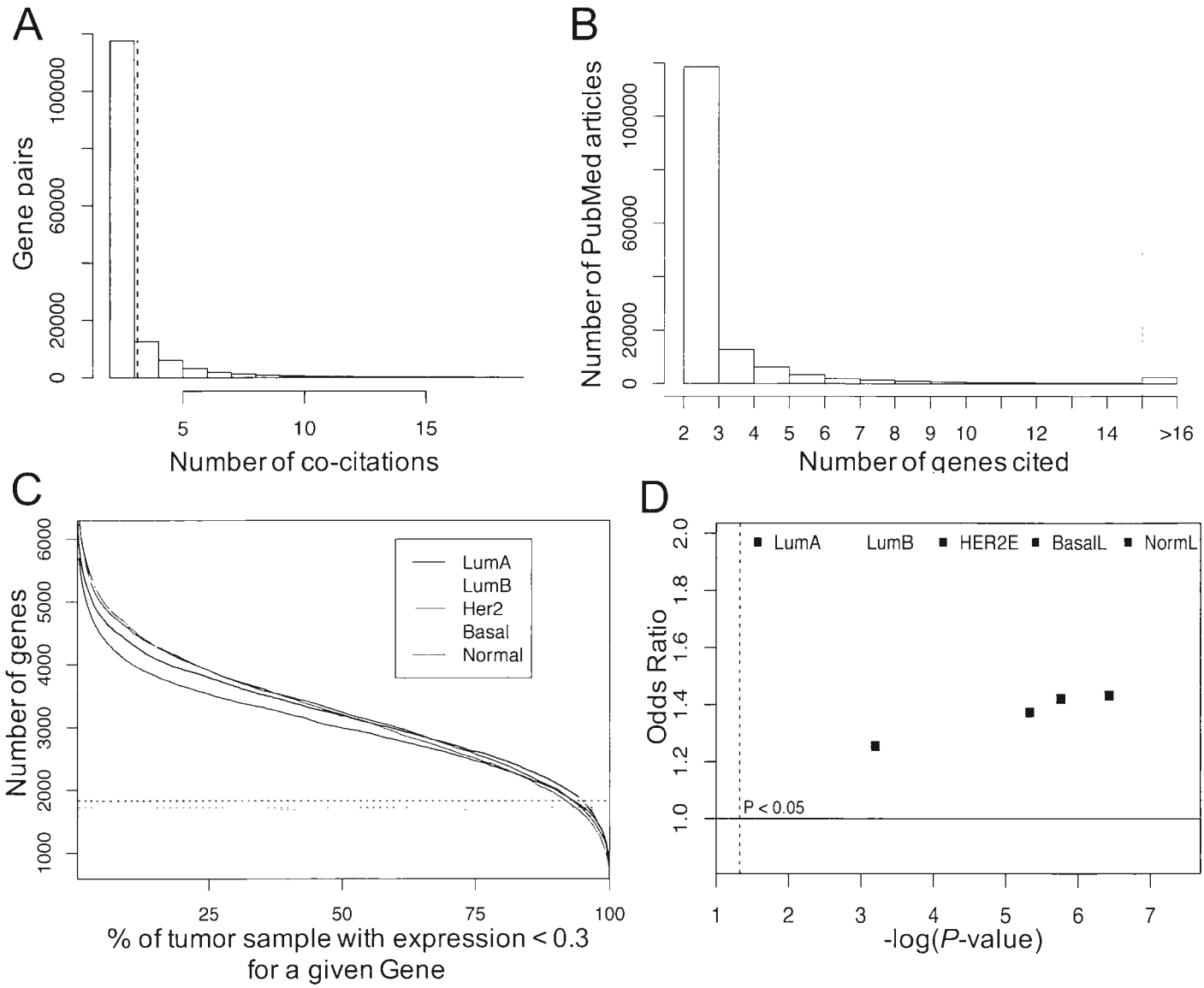


Figure 4.2. Filtering stages of co-expression attributes.

(A) Distribution of gene pairs according to the number of time they are co-cited in PubMed database. The PubMed gene co-citations threshold (3 co-citations) is indicated by a red dashed line. **(B)** Distribution PubMed articles according to the number of genes they cite. PubMed articles citing more than 15 genes are located at the right of the red dashed line. **(C)** Distribution of genes considering the percentage of breast tumor samples in which their expression level is lower than minimal expression limit (0.3). The 95% cutoff corresponding to the fraction of tumor samples is indicated by the vertical black dotted line. Horizontal dashed lines indicate the number of non-expressed genes per tumor sub-type. **(D)** Enrichment of co-expressed gene pairs co-cited with "breast cancer" term when considering CoE attribute after removing non-expressed genes when compared to the attribute before filtering. Odds ratio and *P*-value from one-tailed Fisher's exact test (N= All gene pairs in the original reference set) are indicated.

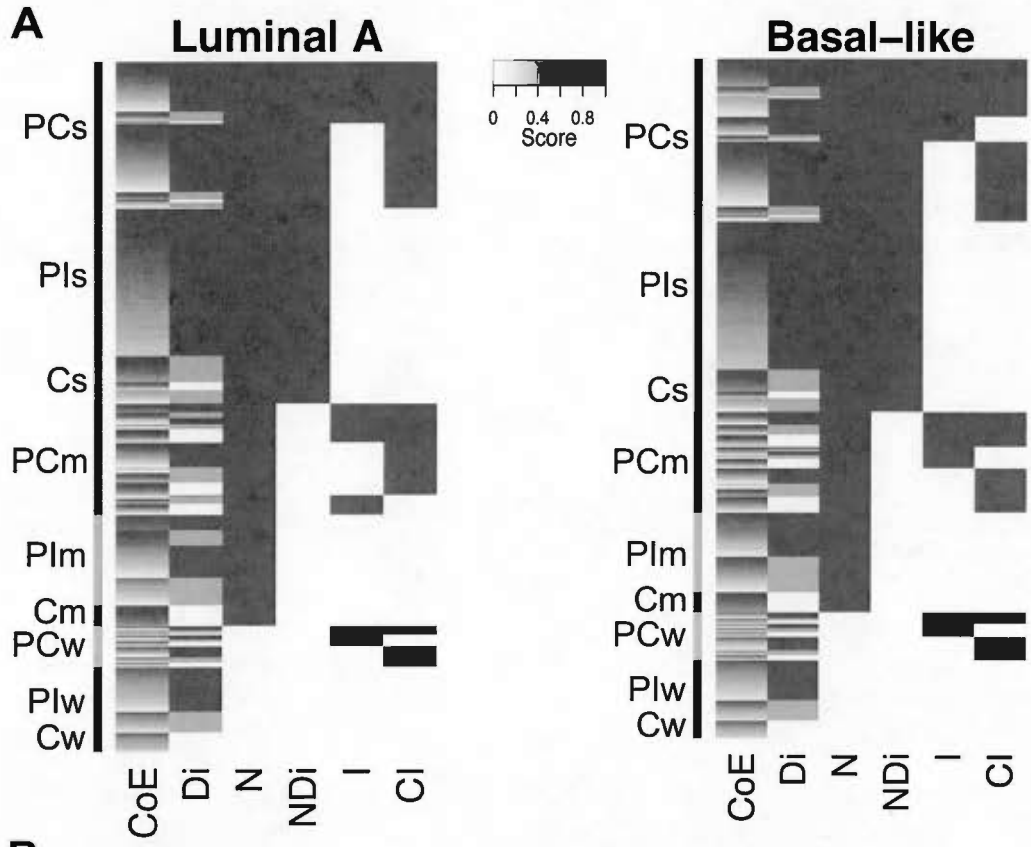


Figure 4.3. Identification of classes and subclasses of functional interactions.

(A) Functional interactions classes identified with hierarchical clustering. Heatmap showing the hierarchical clustering, using Ward's agglomerative method with a distance metric based on the Euclidean distance, of the Luminal A and Basal-like reference gene pairs signatures. Classes are defined as: Protein-Protein interaction (PPI) independent (PI), PPI-centric (PC) and Connectors (C). Six columns show values for statistical attributes. genes co-expression levels (CoE; high score = (1 –(*P*-value <0.05)) indicate high co-expression); enrichment of shared disease annotations (Di; score=1 indicate enrichment; 0.5 indicate not significant enrichment; 0 indicate significant depletion of shared annotation); enrichment of the neighborhoods of gene-pairs in a given disease annotation (N; score =1; 0 indicate no enrichment), enrichment of the neighborhoods of gene-pairs in a given disease annotation and association of both genes of the considered pair to this enriched annotation (NDi; score =1; 0 otherwise); interacting genes code for interacting proteins (I; score =1; 0 otherwise); interacting genes code for proteins sharing a significantly high number of common protein-protein interaction partners (CI; score =1; 0 otherwise). For N and NDi attribute, the neighborhood of a given gene is considered as the set of genes that show a significant co-expression ($P \leq 0.05$, see Methods) and/or encode proteins that exhibit a PPI with the product of this gene. **(B)** Description of the contextual strength (CS). The metric is based on the binary values from the neighborhood (N) and neighborhood with disease (NDi) attributes. The different gene characteristics are described on the right.

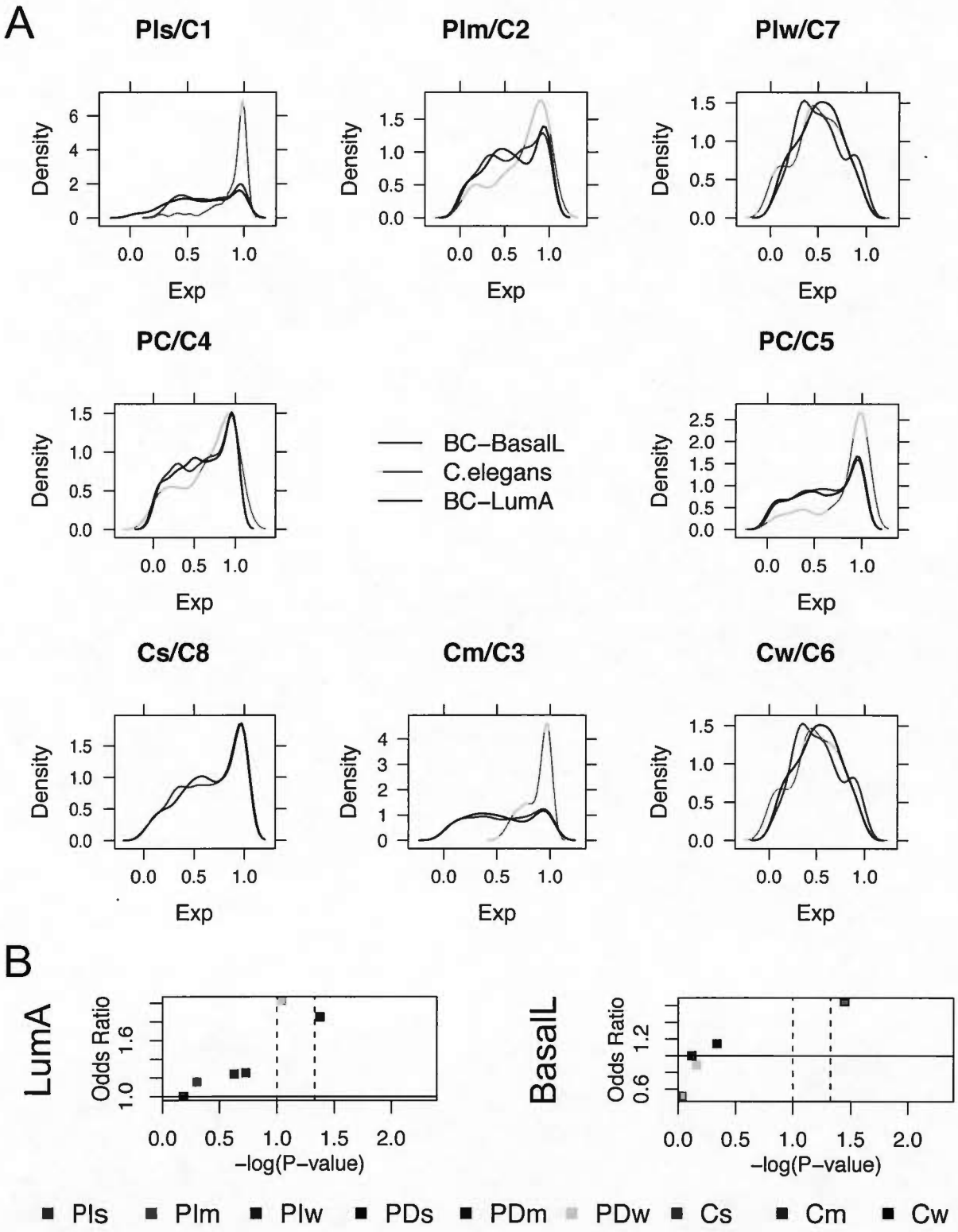


Figure 4.4. Comparative distribution of Co-expression distributions between *C. elegans* and breast cancer subtypes.

(A) Distribution of co-expression values of functional interactions classes for the Basal-like (BC-BasalL) and Luminal A (BC-LumA) breast cancer subtypes compared with co-expression distributions of corresponding genetic interactions classes in *Caenorhabditis elegans* (see (Boucher, B. *et al.*, 2016)). Co-expression values are given as (1-(*P*-value)) of a normal fitted distribution of all values. **(B)** Enrichment of gene pairs co-cited with breast cancer term after vs before filtering stage. Enrichment is shown by a positive odds ratio and the statistical significance is calculated via a one-tailed Fisher's exact test and is given as -log(base10)(P-value). Two significance levels are shown: 0.10 (red dashed line) and 0.05 (black dashed line).

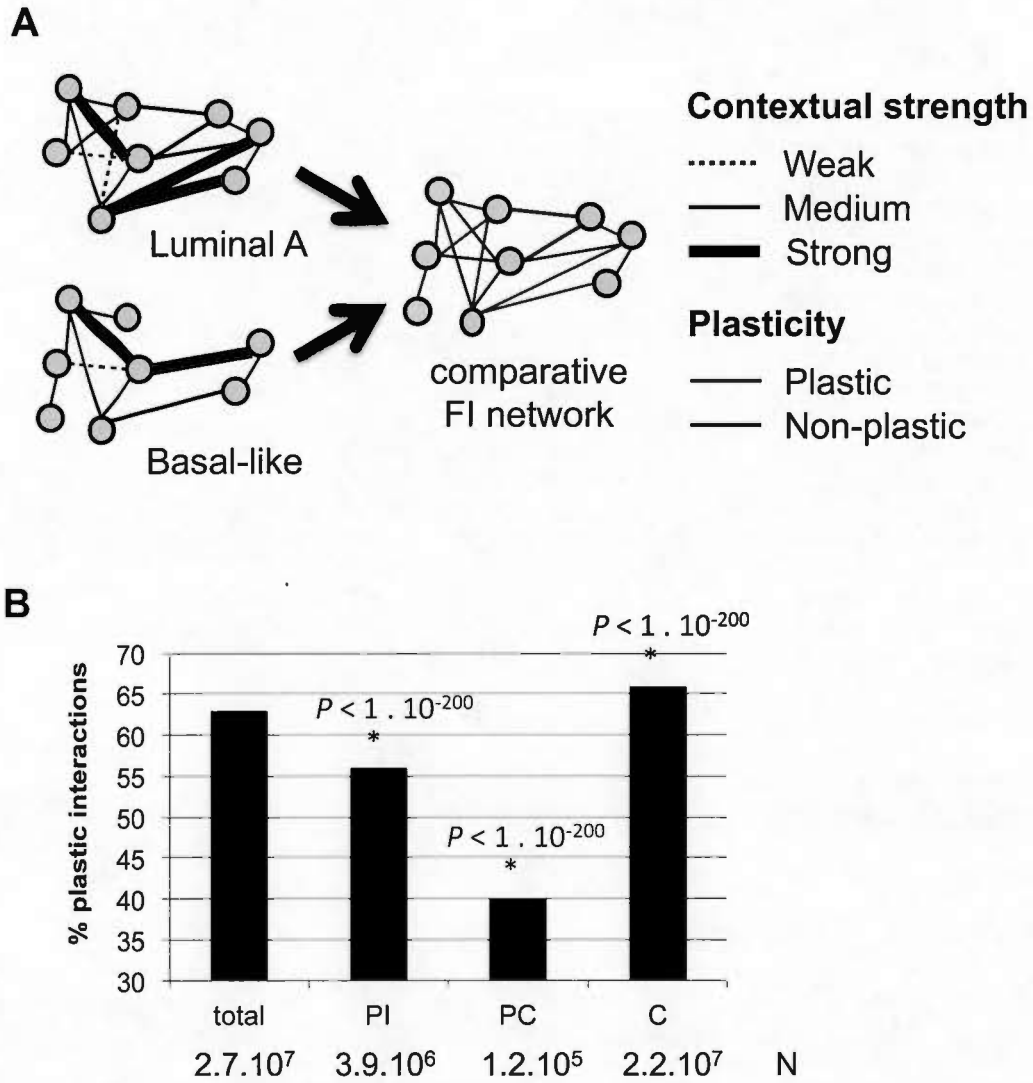


Figure 4.5. Plasticity of the predicted functional interactions networks of of Luminal A and Balsal-like breast cancer subtypes.

(A) Contextual strength (weak, medium and strong) is indicated for interactions in a schematic representation of Luminal A and Basal-like predicted functional interactions networks. The plasticity is defined by a change in the contextual strength of an interaction or by the presence/absence of the interaction. (B) Percentage of plastic functional interactions for the three main classes when comparing Luminal A and Basal-like networks. N indicates the number of predicted interactions. enrichment of plastic interactions in each class of functional interaction is indicated by a low P-values given by a two-tailed Fisher's exact test when considering the total population of interactions N.

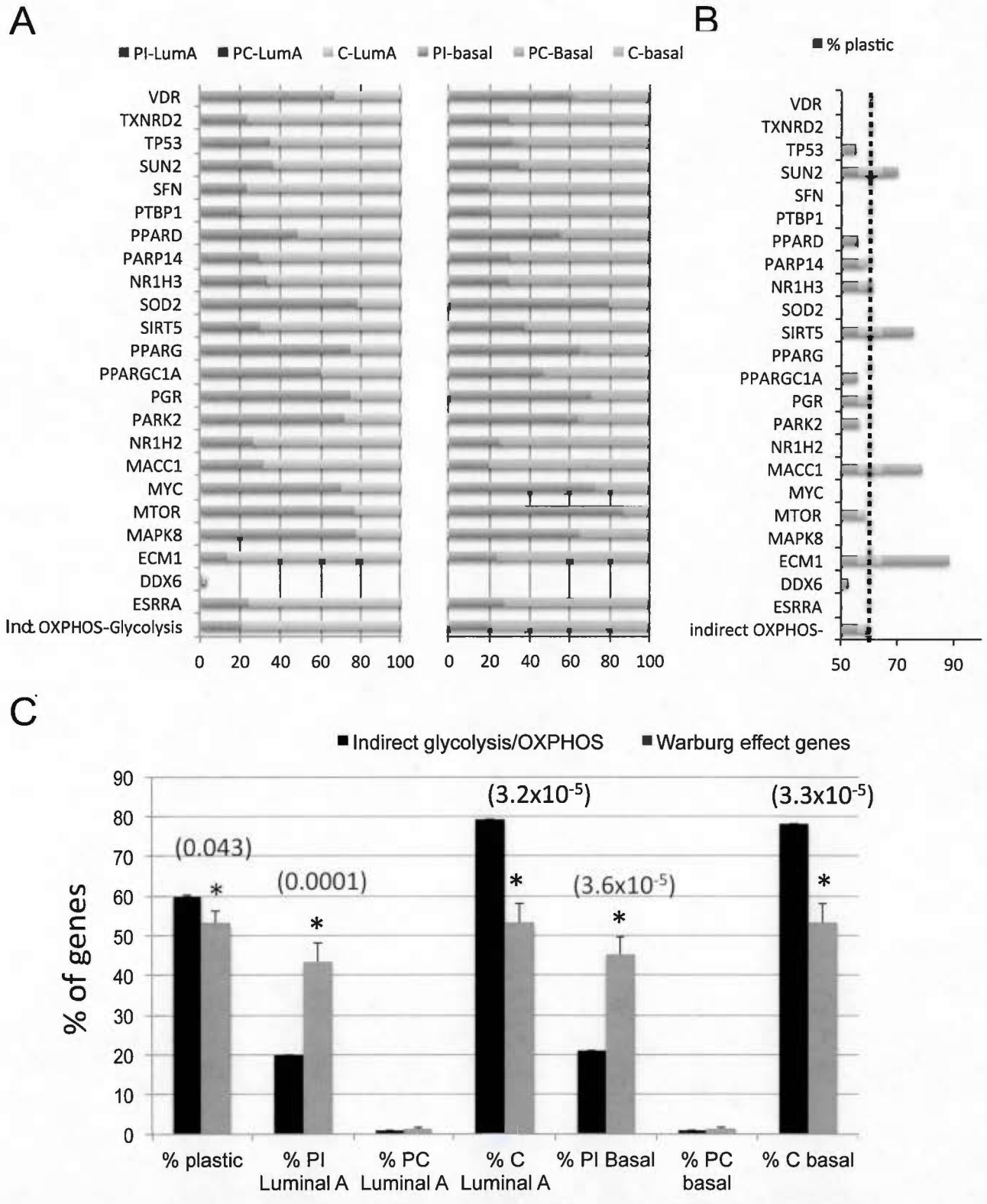


Figure 4.6. Properties of Warburg effect associated genes within functional interactions classes.

(A) Warburg effect genes associated proportion of each functional interactions (FI) classes from Luminal A (LumA) and Basal-like (Basal) breast tumours predicted networks between oxidative phosphorylation (OXPHOS) and glycolysis associated genes. The proportion for all FI (indirect-OXPHOS) is also indicated. **(B)** Percentage of plastic FIs between each Warburg associated genes and its partners in the OXPHOS and glycolysis machineries. All plastic FI (indirect-OXPHOS) is also indicated. **(C)** Comparison of the proportion of plastic and class-specific interactions between all indirect genes and the training set (Warburg effect genes) within the glycolysis-OXPHOS FI indirect networks of Luminal A and Basal-like (basal) breast cancer. P-values are calculated via a two-tailed Student's t-test.

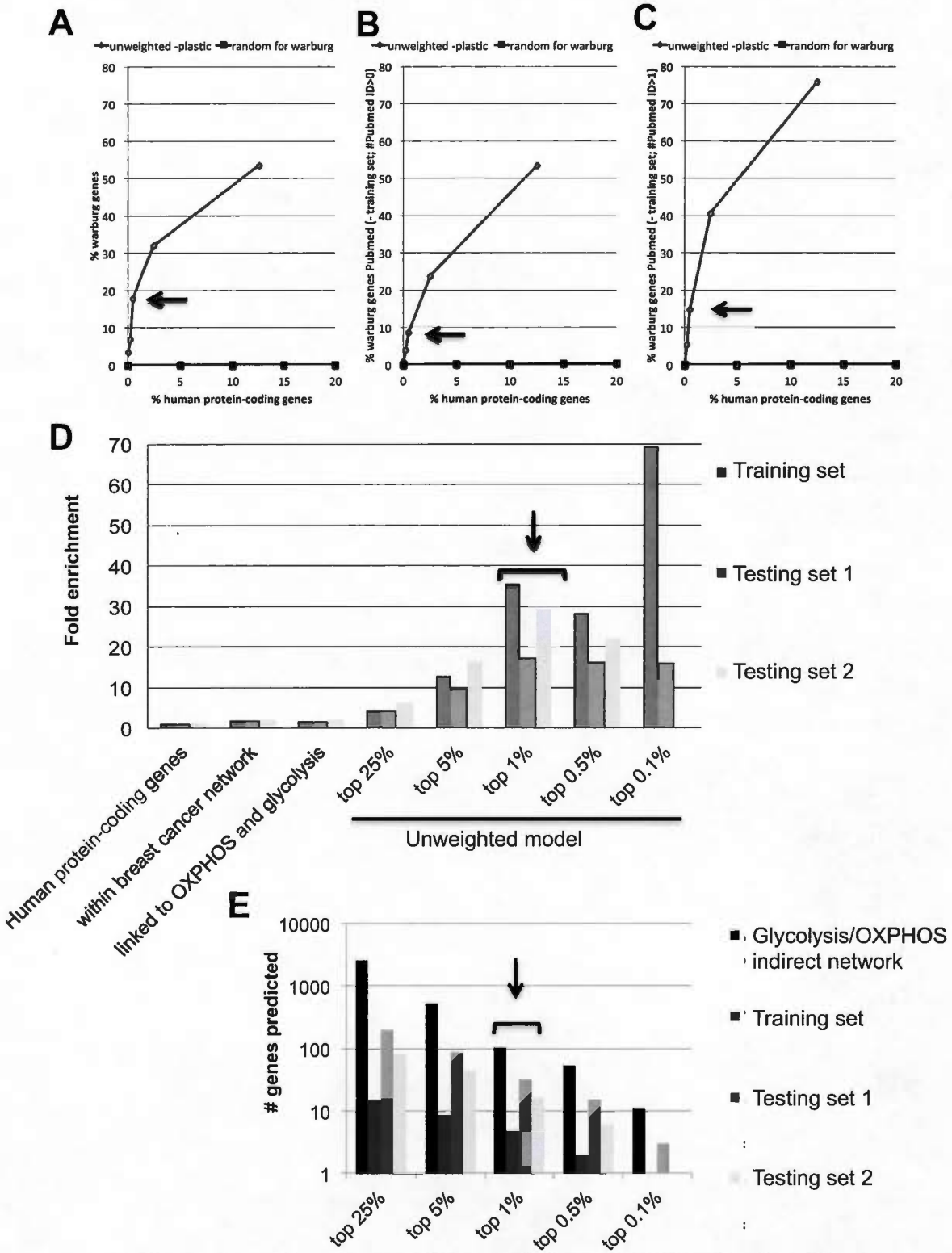
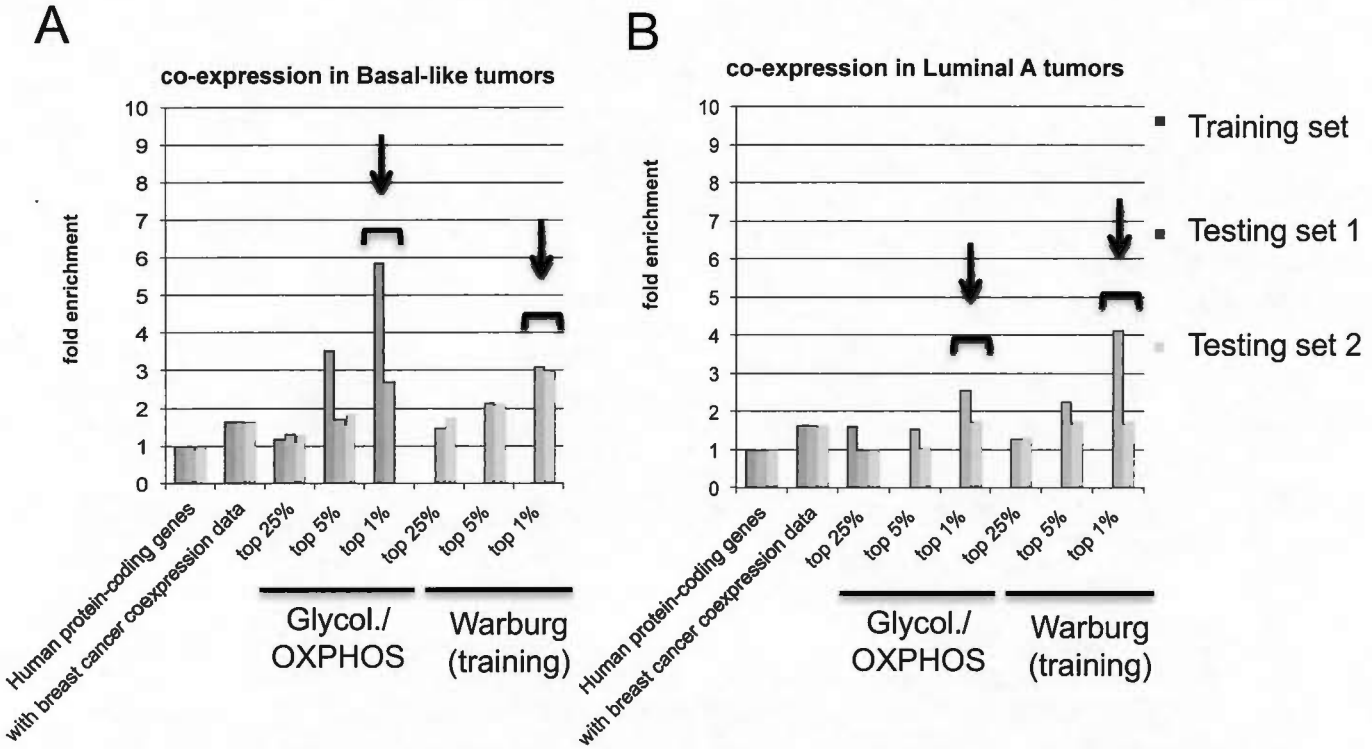


Figure 4.7. Overall performances of a prediction model trained on Warburg effect associated genes.

Performances of the unweighted-linear models on the training set **(A)**, the testing set 1 **(B)** and the testing set 2 **(C)**. The relative representation of each set genome-wide is indicated by a black line close to the X axis (Random for Warburg). **(D)** Enrichment of training and testing set 1 and 2 in subgroups of the OXPHOS-Glycolysis indirect interaction network with V(A) value provided by the unweighted model within the top 25%, 5%, 1%, 0.5% and 0.1% when compared to their representation in human protein coding genes. Enrichment in total predicted functional interactions (within breast cancer network) and in the indirect interaction network between OXPHOS and Glycolysis are also indicated. **(E)** Number (#) of genes predicted by the unweighted model at different thresholds.



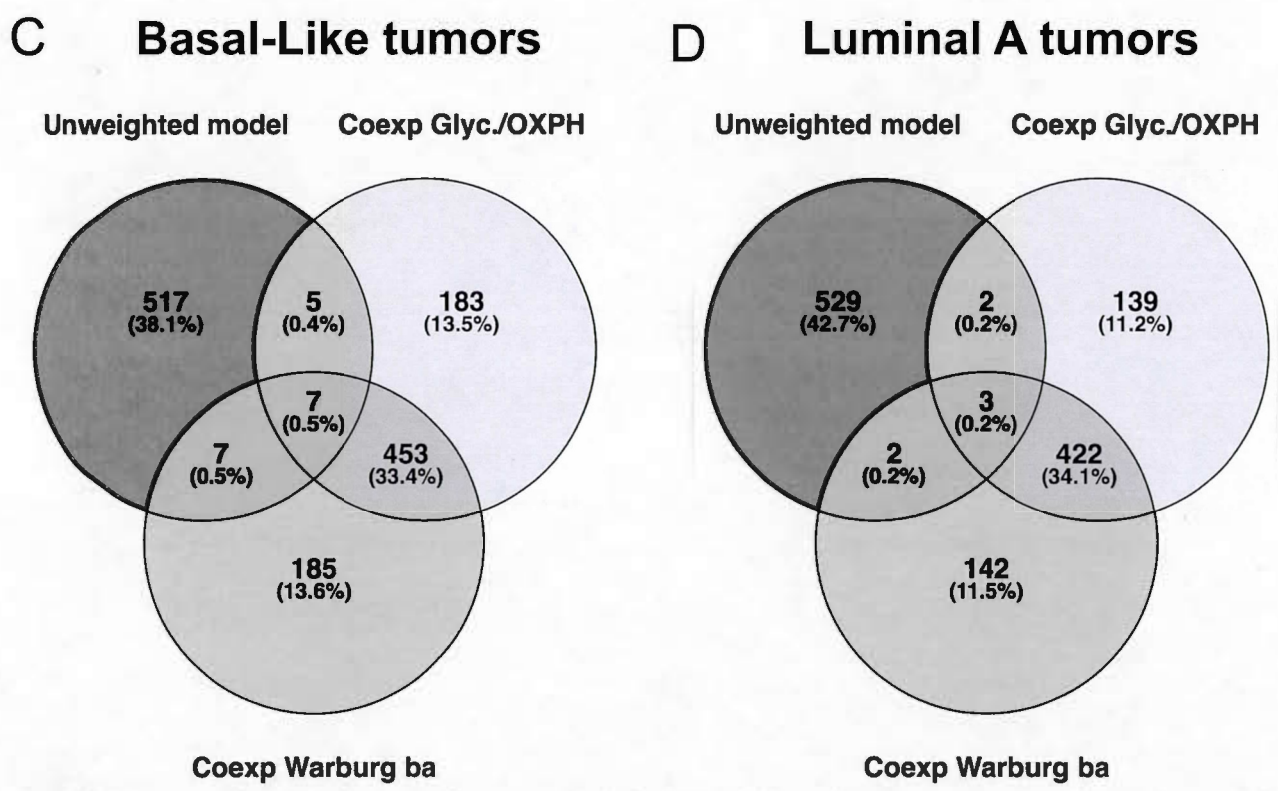
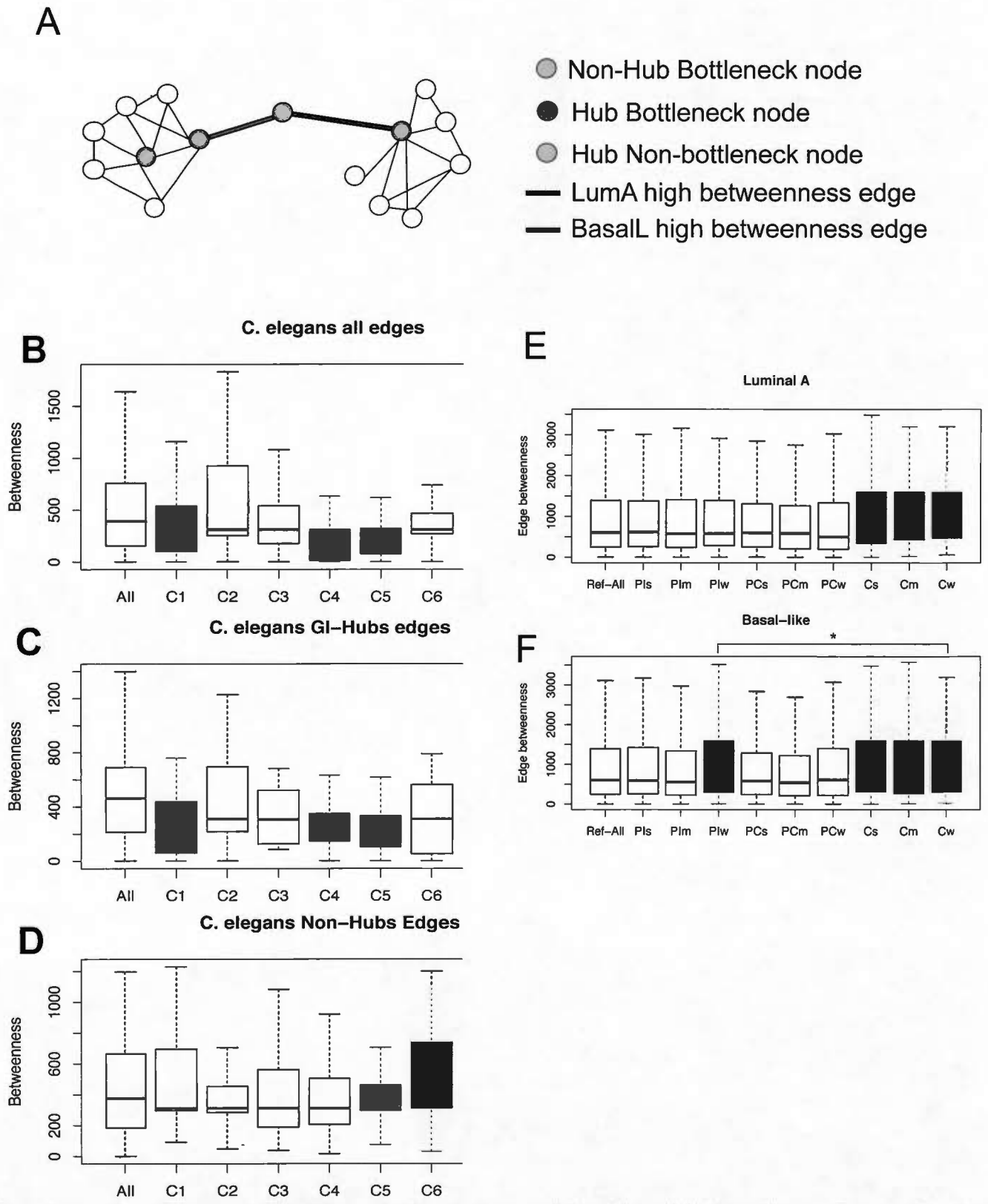


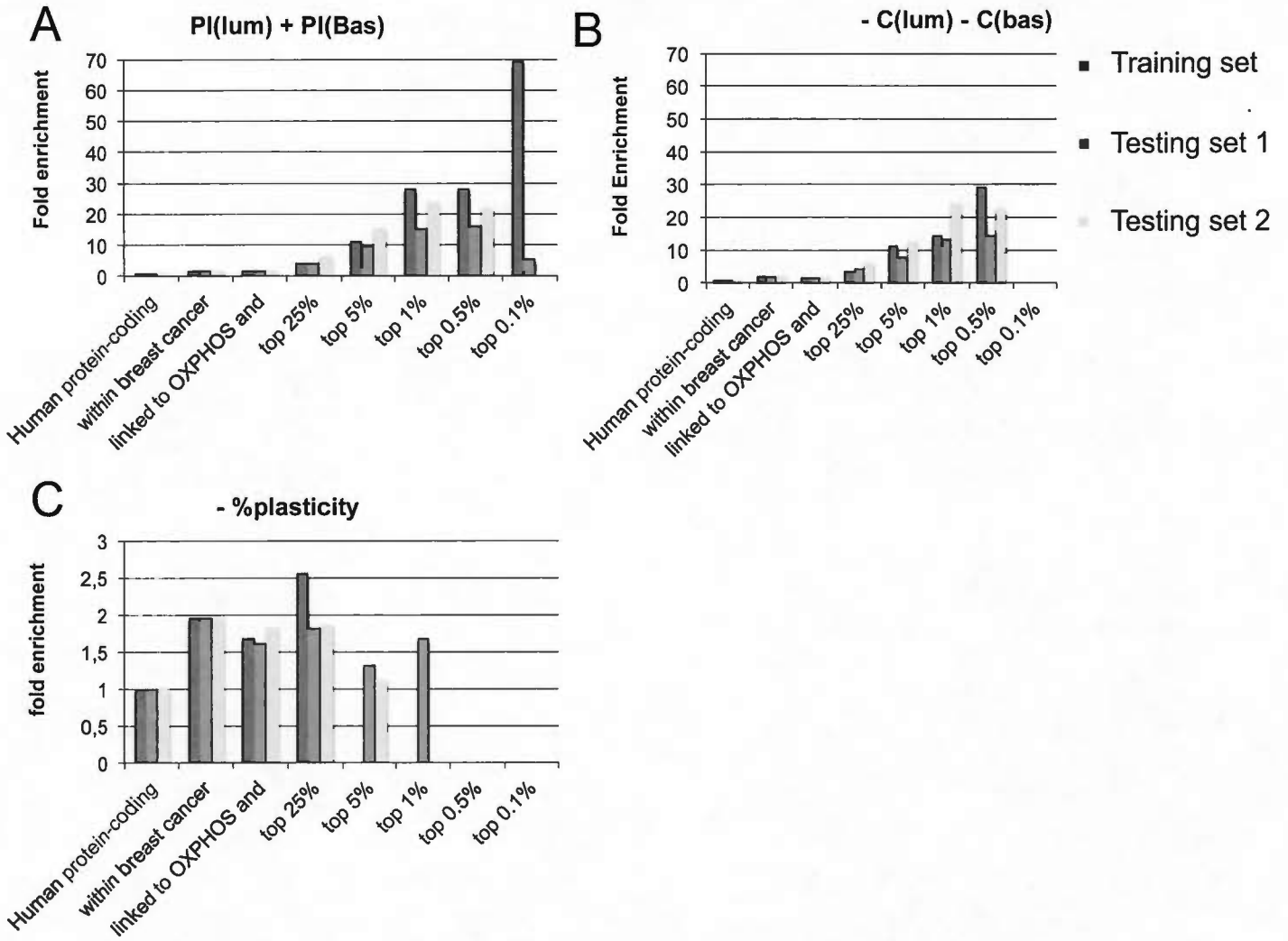
Figure 4.8. Comparison between the unweighted model and co-expression based predictions of Warburg effect genes.

Fold enrichment of training and testing sets in groups of genes identified based on their co-expression with genes of the glycolysis and OXPHOS machineries (Glyco./OXPHOS) or with the training set (Warburg (training)) using Basal-like **(A)** and Luminal A **(B)** breast tumors co-expresssion data (filtered *CoE* attributes). Groups are identified based on to 25%, 5% and 1% of *CoESum* values. (C)-(D) Venn diagrams showing the overlaps between predictions based on the unweighted model (top 5% V(A) value for gene A) and co-expression with Glycolysis/OXPHOS genes or the Warburg (top 5% CoESum(A) for gene A) when considering Basal-like **(C)** and Luminal A **(D)** data.



Supplementary Figure 4.S1. Edges betweenness of functional interactions classes for breast cancer subtypes and genetic interactions in Caenorhabditis elegans.

(A) schematic representation of functional (genetic) interaction networks showing interactions (edges) with high betweenness linking Hub and non-Hub bottlenecks. Boxplots showing betweenness values distributions in C. elegans (B-D) and human breast cancer (E-F). Light grey (lower) and dark grey (higher) boxes represent statistically significant differences ($P < 0.05$) between interactions classes and union of interactions classes (All). White boxes indicate not statistically significant differences with All. Differences are assessed with a two-tailed Wilcoxon signed-rank test.



Supplementary Figure 4.S2: contribution of attribute to the unweighted linear model.

Three models have been generated using either the sum of percentage of PI interactions in Luminal **(A)** (PI(lum)) and Basal-like (PI(Bas)) subtypes (A) or the of C interactions (C(Lum) and C(Bas) respectively). **(B)** A third model use only low plasticity value for predictions. **(C)** Enrichment of training and testing sets 1 and 2 in groups of gene with score value from the model within the top 25%, 5%, 1%, 0.5% and 0.1% are indicated when considering their respective representation genome-wide.

CHAPITRE V

DISCUSSION

5.1 Résumé de la thèse

Connaître et comprendre la structure des réseaux fonctionnels est d'une importance capitale pour élucider les mécanismes moléculaires complexes liant les gènes chez les organismes modèles et chez l'homme. Dans cette thèse, j'ai décrit une méthode d'intégration de données qui a permis d'élucider la structure d'un interactome génétique chez le nématode *Caenorhabditis elegans* et ai aussi montré qu'une structure semblable chez l'humain pouvait servir à prédire des liens fonctionnels impliquant des gènes clés dans la progression du cancer du sein.

Dans le Chapitre II, nous avons montré que pour mieux prédire des interactions génétiques (IGs) chez *C. elegans*, il était d'abord nécessaire de mieux comprendre ce type d'interaction fonctionnelle encore mal décrit dans la littérature (Boucher et Jenna, 2013). Pour ce faire, nous avons présenté une étude comparative de plusieurs méthodes existantes démontrant que les techniques et approches actuelles n'arrivaient pas à prédire l'ensemble des IGs. Dans les faits, l'analyse exhaustive exposée dans ce chapitre montre que la prédiction d'IGs pourrait bénéficier grandement d'une meilleure compréhension des liens qui existent entre les différents niveaux d'abstraction des systèmes biologiques (Boucher et Jenna, 2013).

Dans le Chapitre III, j'ai présenté la caractérisation d'un interactome génétique de *C. elegans* liant des gènes se trouvant à l'intérieur des voies de signalisation/métabolisme (Boucher *et al.*, 2016). Cette caractérisation a permis de mettre en évidence une structure de ce réseau formé de six classes d'IGs ayant des propriétés distinctes. Cette étude révèle que les voies de signalisation/métabolisme sont composées d'IGs apparaissant au centre des modules denses d'interactions protéine-protéine (PDS) appelées PDS-centric GIs et d'autres étant indépendantes de ces mêmes modules (« PDS-indépendant GI »). Notre étude suggère que la coordination des modules fonctionnels définis par ces deux classes d'IGs impliquerait des gènes hautement connectés dans le réseau (IG-Hub) et liés à leurs partenaires fonctionnels via deux autres classes d'IG (« Connector involving GI »). Les gènes définis par ces différentes classes d'IG présentent des caractéristiques distinctives lorsque l'on considère leur connectivité au sein du réseau IG mais aussi au sein du réseau d'interactions protéine-protéine, leur pléiotropie et leur redondance fonctionnelle (Boucher *et al.*, 2016). Le modèle de l'interactome génétique construit par notre étude pourrait permettre de mieux comprendre le rôle des IGs au niveau de la robustesse des systèmes et le contrôle qu'elles exercent sur les processus biologiques.

Dans le Chapitre IV, j'ai montré qu'il était possible de transposer à l'humain, la méthode d'intégration développée chez le nématode *C. elegans*. Plus précisément, nous avons démontré dans l'étude présentée à ce chapitre qu'il était possible, avec notre approche, de caractériser un réseau d'interactions fonctionnelles (IFs) pour des sous-types de cancer du sein (Boucher *et al.*, 2017, manuscrit en préparation). Nous avons ensuite identifié, dans ce réseau, différentes classes d'IFs ayant des caractéristiques similaires aux classes d'IGs étudiées précédemment chez *C. elegans*. Nous avons utilisé ces classes d'IFs ainsi que leur niveau de plasticité entre les sous-types de cancer du sein Luminal A et Basal-like, pour construire un modèle prédictif servant à découvrir des gènes potentiellement impliqués dans l'effet Warburg, et ce, à l'échelle du génome. La comparaison des performances *in silico* de ce modèle et d'autres approches, utilisant

des données d'expression seulement, démontre clairement l'avantage de l'approche de génomique intégrative pour identifier des gènes impliqués dans la reprogrammation métabolique des tumeurs mammaires (Boucher *et al.*, 2017, manuscrit en préparation). La validation expérimentale des prédictions de notre modèle sera effectuée dans les prochains mois permettant la soumission de cet article pour publication.

5.2 Le degré de centralité intermédiaire des réseaux d'interactions fonctionnelles

Un des points importants de cette thèse concerne la structure de l'interactome génétique chez *C. elegans*, il serait très intéressant de savoir si cette structure partage des similitudes avec le réseau d'interactions fonctionnelles du cancer du sein chez l'humain. Par exemple, lorsque l'on considère le degré de centralité intermédiaire (« betweenness », Chapitre IV, Figure 4.S1), les IFs de types connecteurs pour les sous-types de tumeurs du sein chez l'humain ont des « betweenness » significativement supérieures à l'ensemble des IFs. Du côté du nématode, les classes d'interactions de type connecteur (C3, C6) du réseau d'IGs ont des valeurs de « betweenness » semblables à l'ensemble du réseau, mais supérieures aux autres classes (Figure 4.S1). Sur ce point précis, je vais discuter du lien unissant la classe des connecteurs et le degré de centralité intermédiaire élevé de leurs interactions.

5.2.1 Comparaison entre la levure et le nématode

L'organisation des IGs en PDS-independant, PDS-centric, ainsi qu'en connecteur, particulièrement les C6 (enrichis en « GI-Hub ») et présentant une valeur moyenne de

« betweenness » élevée, suggère le modèle suivant (Figure 5.1). Les interactions PDS-independant et PDS-centric pourraient former des modules hyperconnectés (bien que cela n'ait pu être testé) et les liens impliquant des connecteurs de type C6 feraient le lien entre ces deux types de modules fonctionnels à l'intérieur d'une même voie de signalisation ou de métabolisme, ce qui expliquerait leur degré élevé (GI-Hub) au sein de l'interactome génétique et leur valeur de Betweenness élevée (« Pathway 1 », Figure 5.1).

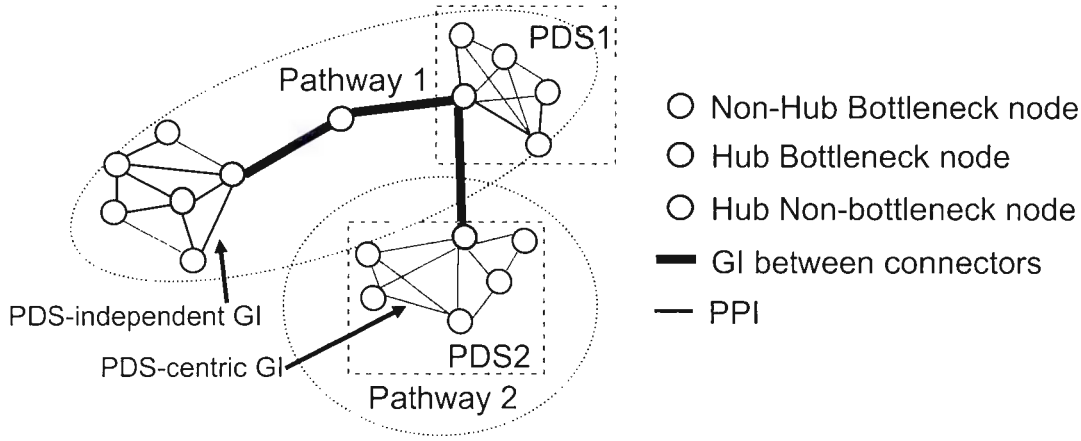


Figure 5.1 Schéma représentant la structure possible du réseau d'interactions génétique de *Caenorhabditis elegans*.

Différents types de points nodaux sont connectés par des interactions génétiques (« GI ») ou des interactions protéine-protéine (« PPI »). Les motifs denses d'interactions protéine-protéine (« PDS ») sont à l'intérieur des cadres en gris pointillé. Les différentes classes sont indiquées par des flèches (« PDS-independent » et « PDS-centric ») ou par des liens en gras (« GI between connectors »).

Ce modèle s'appuie par analogie sur l'existence de protéines « bottleneck » servant de ponts entre les modules fonctionnels dans l'interactome protéine-protéine de la levure (Joy *et al.*, 2005; Yu *et al.*, 2007). Ces protéines ont été montrées comme étant codées, la plupart du temps, par des gènes essentiels qui sont faiblement co-exprimés avec leurs partenaires (Yu *et al.*, 2007). Une propriété similaire est d'ailleurs attribuée chez *C.*

elegans aux gènes connecteurs liés par des interactions de type C6, qui sont faiblement co-exprimés avec leurs partenaires (Chapitre III, Figure 3.1b) (Boucher *et al.*, 2016).

Néanmoins, des études sont nécessaires sur le degré de centralité intermédiaire des réseaux d'IGs chez la levure et aussi, chez *C. elegans*, dont différentes méthodes de prédictions d'IGs pourraient être mises à contribution.

5.2.2 Comparaison entre le nématode et l'humain

Dans le cancer du sein, chez l'humain, notre réseau d'interactions fonctionnelles (IFs) entre paires de gènes appartenant à une même voie de signalisation/métabolisme présente des classes ayant un degré de centralité intermédiaire semblable aux classes retrouvées dans le réseau d'IGs caractérisé de *C. elegans* (Figure 5.2) (Boucher *et al.*, 2017, manuscrit en préparation). En effet, le degré de centralité intermédiaire des liens de types connecteurs est supérieur aux autres types de liens, et ce, autant chez le nématode (C3, C6; chapitre IV, Figure 4.S1) que chez l'humain (Cs, Cm, Cw; Chapitre IV, Figure 4.S1) (Boucher *et al.*, 2017, manuscrit en préparation). Spécifiquement, le « betweenness » moyen des classes connecteurs des réseaux d'IFs chez l'humains est significativement différent de la valeur moyenne lorsqu'on considère l'union des classes d'IFs (Chapitre IV, Figure 4.S1). Par contre, chez *C. elegans*, le « betweenness » des classes connecteurs n'est pas significativement différent du réseau d'IGs en entier (Chapitre IV, Figure 4.S1). L'apparente différence de « betweenness » entre les classes des connecteurs chez le nématode et l'humain pourrait s'expliquer, bien évidemment, par la nature différente des liens utilisés (IG vs IF). En effet, l'étude chez le nématode décrit la caractérisation d'IGs validées expérimentalement et donc enrichies en interaction fonctionnelles, identifiées à l'intérieur des voies de signalisation (Boucher *et al.*, 2016). Étant donné la non disponibilité de telles données pour l'homme,

particulièrement dans les tumeurs mammaires, nous avons utilisé des IFs consistant en des paires de gènes prises à l'intérieur de voies de signalisation en plus d'être co-citées à plus d'une reprise dans la littérature (Boucher *et al.*, 2017, manuscrit en préparation). Dans le passé, la co-citation a été utilisée comme attribut pour prédire des interactions protéine-protéine (Chowdhary *et al.*, 2009; Ramani *et al.*, 2005) et génétiques (Lee, I. *et al.*, 2010; Lee, I. *et al.*, 2007). Nous postulons donc que cet ensemble de référence contienne peu de faux positif et qu'il soit enrichi en IGs, quoique ce postulat soit difficilement testable.

De façon intéressante, une étude a montré que des réseaux fonctionnels construits à partir de voies de signalisation et de données d'expression des gènes provenant de ces deux espèces sont sensiblement différents dans leurs topologies (Korcsmaros *et al.*, 2010). Cette étude montre en effet que le réseau fonctionnel de l'humain présente beaucoup plus de gènes communiquant entre les voies de signalisation que celui du nématode et de la mouche *Drosophila melanogaster* (Korcsmaros *et al.*, 2010). Ceci supporte l'hypothèse selon laquelle la différence observée entre les degrés de centralité intermédiaire du réseau fonctionnel de *C. elegans* et celui de l'humain pourraient être due à des différences biologiques et non expérimentales. Par analogie, ces gènes communiquant entre les voies de signalisation sont aussi, fort possiblement, des « bottlenecks » au niveau du réseau fonctionnel complet. Bien que dans cette étude le réseau fonctionnel soit constitué de paires de gènes co-exprimés à l'intérieur des voies de signalisation (Korcsmaros *et al.*, 2010), il demeure encore bien éloigné dans sa nature par rapport à un réseau d'IGs validées expérimentalement. Des études précises sur ce point sont donc nécessaires pour confirmer cette hypothèse chez l'humain et d'autres animaux.

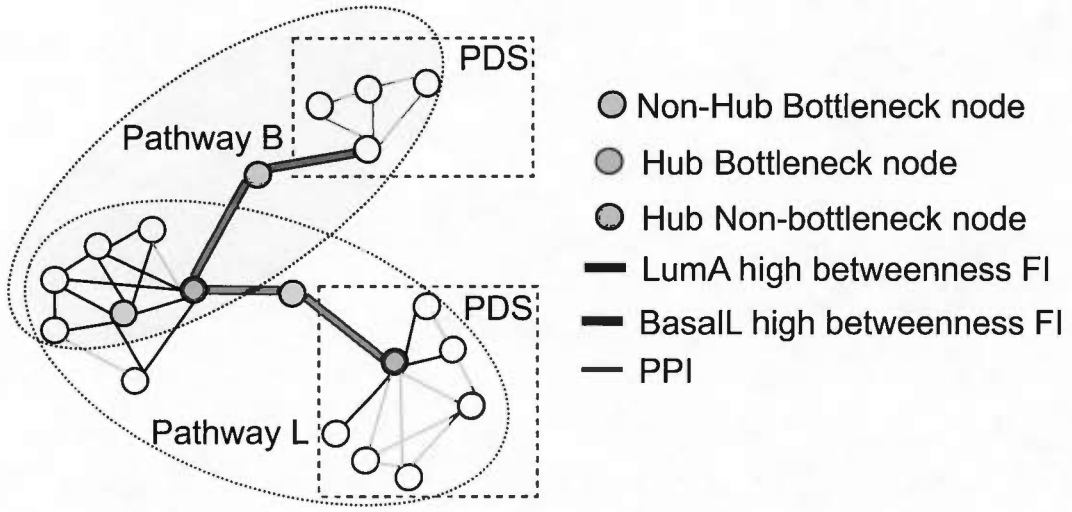


Figure 5.2 Schéma représentant la structure possible des réseaux d'interactions fonctionnelles de deux sous-types de cancer du sein chez l'humain.

Différents types de points nodaux sont connectés par des interactions fonctionnelles (« FI ») ou des interactions protéine-protéine (« PPI »). Les motifs denses d'interactions protéine-protéine (« PDS ») sont à l'intérieur des cadres en gris pointillé. Deux voies (« Pathway ») sont montrées pour les deux sous-types de cancer du sein, les Basal-like (B) et les Luminal A (L).

Dans l'ensemble, l'évolution et la plasticité des réseaux fonctionnels pourraient dépendre, en partie, d'une certaine quantité de nouveaux gènes acquis ou de gènes communs, codant pour de nouvelles isoformes d'une même protéine. Ces gènes pourraient alors permettre une coordination différente des modules fonctionnels nécessaires à la survie et au développement des cellules saines chez *C. elegans*, ou encore à ceux des cellules cancéreuses chez l'humain. L'identification de gènes ayant un « betweenness » élevé représente donc une voie possible dans la recherche de cibles thérapeutiques adaptées au différents contextes fonctionnels d'une maladie.

5.3 Distinguer les interactions fonctionnelles impliquant des interactions protéine-protéine des autres

Une des conclusions importantes du Chapitre III concerne la distinction entre les IGs concentrées dans les modules denses d'interactions protéine-protéine (PDS) et les IGs comprises dans les modules fonctionnels dépourvus de PDS. Contrairement à beaucoup d'études intégratives existantes, notre méthode d'intégration traite séparément les paires de gènes codant pour des protéines qui interagissent ensemble ou qui partagent un nombre significatif de partenaires en commun de celles qui ne semblent pas impliquées dans un complexe protéique (« PDS-independent ») (Boucher *et al.*, 2016) (Boucher *et al.*, 2017, manuscrit en préparation). Chez *C. elegans*, les IGs associées aux PDS, appartenant à la classe des « PDS-centric », sont enrichies en gènes codants pour des protéines hautement connectées dans le réseau d'interactions protéine-protéine (IPP) (« PPI-Hubs ») (Chapitre III, Figure 3.3b) (Boucher *et al.*, 2016).

Les « PPI-Hubs » ou protéines hautement connectées dans le réseau d'IPP, sont d'un intérêt particulier au niveau thérapeutique. En effet, beaucoup de cibles thérapeutiques chez l'humain sont des protéines hautement connectées dans le réseau d'IPP et sont aussi conservées chez plusieurs espèces en plus d'être codées par des gènes essentiels (Lv *et al.*, 2016). Cela a évidemment motivé beaucoup d'études à utiliser le degré d'IPP pour découvrir des cibles thérapeutiques (Fu *et al.*, 2015 ; Ivanov *et al.*, 2013 ; Padiadpu *et al.*, 2010). Bien sûr, cibler des gènes essentiels n'a évidemment pas que des avantages. En effet, il a été montré que les effets secondaires de nombreux médicaments soient liés au fait que leurs principales cibles soient des « Hubs » essentiels dans l'interactome protéine-protéine (Wang *et al.*, 2013). Dans le troisième Chapitre, on a montré que les IGs « PDS-centric » étaient peu, ou pas enrichies en gènes essentiels (Figure 3.2b) mais enrichies en « PPI-Hubs » (Figure 3.3b) (Boucher *et al.*, 2016). Cette particularité, avec leur pléiotropie réduite (Figure 3.7c et 3.7d), suggère que cette classe pourrait contenir des interactions entre gènes pouvant être de

bonnes cibles thérapeutiques et spécifiques à un tissu. À l'inverse, tout en étant enrichie en gènes essentiels, la classe des interactions « PDS-independent » n'est pas enrichie en « PPI-Hubs » (Figure 3.2b et 3.3b), mais est enrichies en gènes pléiotropiques (Figure 3.7c et 3.7d) (Boucher *et al.*, 2016). Cela suggère que cette classe contienne des IG entre des gènes qui ont peu de chance d'être de bonnes cibles thérapeutiques. Comme les gènes associés à l'effet Warburg sont majoritairement retrouvés dans la classe « PDS-independent » (Chapitre IV), ceux-ci auraient donc aussi peu de chance d'être de bonnes cibles pour des traitements anti-cancéreux.

L'avantage de notre méthode d'intégration est qu'elle permet de traiter séparément les paires de gènes associées fortement aux IPP et d'éviter de concentrer notre étude sur les interactions génétiques associées à des complexes protéiques, qui font actuellement l'objet de la plupart des études utilisant des approches de génomique intégrative (Mitra *et al.*, 2013).

5.4 La coordination des processus biologiques par différentes classes d'interactions fonctionnelles selon le contexte

Dans le dernier chapitre, le modèle prédictif pour identifier des gènes potentiellement impliqués dans l'effet Warburg utilise les liens de la classe PI comme un attribut fortement prédictif. À l'inverse, les interactions fonctionnelles (IFs) de la classe des connecteurs C sont utilisées négativement dans le modèle prédictif final. Aussi, les IFs associées aux IPP, appartenant à la classe PC, sont complètement absentes du modèle prédictif (Boucher *et al.*, 2017, manuscrit en préparation). Nous avons été surpris par ces résultats puisque l'étude faite sur le réseau d'interaction génétique du nématode a permis de qualifier les gènes impliqués dans des interactions de type connecteurs (C3 et C6) comme étant des gènes ayant des propriétés de coordination des processus

biologiques, principalement due à leur enrichissement en « GI-Hubs » et à leur pléiotropie (Chapitre III ; C6, Figure 3.7b et c). Notre hypothèse semblaient aussi se confirmer par le fait que de nombreux gènes de la famille "Lineage deficient" Lin (Chapitre III ; lin-1, lin-12; Figure 3.S5), qui sont des gènes clés déterminant la spécification du lignage cellulaire au cours du développement du nématode, étaient liés à leur partenaires fonctionnels via des interaction de type connecteur pléiotropiques, (C6; Figure 3.S5, (Boucher *et al.*, 2016)). Nous nous attendions donc à ce que des gènes contrôlant la reprogrammation métabolique des tumeurs mammaires fasse partie de cette catégorie de gènes et interagissent avec leurs partenaires de la voie de la glycolyse et de la phosphorylation oxydative (OXPHOS), en majorité via des interactions de type connecteur. Les résultats obtenus indiquent un appauvrissement de ce type d'interactions entre les gènes associés à l'effet Warburg et les gènes de glycolyse/OXPHOS, nous laissant imaginer: (1) que les interactions fonctionnelles que nous avons appelées connecteur (C) dans notre interactome fonctionnel humain n'ont pas de fonction de coordination des processus biologiques; ou (2) qu'il y a plusieurs niveaux de coordination des processus biologiques dont certains impliquent des interactions de type « PDS-independent » (PI) et d'autres, des interactions de type "connecteur" (C).

Le fait que les liens C aient une valeur moyenne de « betweenness » élevée dans les interactomes fonctionnels, et ce à la fois chez le nématode et l'humain, suggère que les gènes liés par des interactions de ce type jouent un rôle structural et potentiellement important dans les réseaux d'interactions fonctionnels de ces deux espèces. Ceci va donc à l'encontre de l'hypothèse (1) énoncée plus haut.

Nous considérons donc l'hypothèse (2) comme la plus probable. En prenant l'ensemble des processus biologiques, les voies métaboliques constituent un ensemble de gènes que nous imaginons très interconnectés. Des gènes impliqués dans la reprogrammation métabolique seraient donc des gènes à l'intersection d'éléments ou de sous-groupe de gènes à l'intérieur de cet ensemble. Nous pourrions donc considérer les gènes associés

à l'effet Warburg comme des gènes coordonnant des modules fonctionnels situés à l'intérieur de l'ensemble des voies métaboliques. Ces gènes auraient donc des fonctions de coordination entre des modules fonctionnels situé à "proximité" les uns des autres dans l'interactome fonctionnel du tissue considéré. Suivant ce même raisonnement, nous émettons l'hypothèse que les gènes coordonnant le métabolisme avec d'autres processus biologiques, tels que la machinerie de polarité ou de migration cellulaire, serait des gènes montrant des fonctions de coordination entre des processus biologiques plus distants dans les réseaux d'interactions fonctionnelles et interagiraient via des interactions de type Connecteurs.

Cette hypothèse est supportée par les caractéristiques distinctives des interactions de la classe PI, équivalent à la classe « PDS-independent » de l'interactome génétique chez *C. elegans* (C1 et C2; Figure 3.1b) (Boucher *et al.*, 2016), et de ceux de classe C équivalent aux classes C3 et C6 chez *C. elegans* (Figure 3.1b) (Boucher *et al.*, 2016). La classe PI est caractérisée par des gènes qui sont très souvent co-exprimés et qui sont associés à plusieurs phénotypes ou maladies en communs (Boucher *et al.*, 2017, manuscrit en préparation) (Boucher *et al.*, 2016). Les interactions fonctionnelles de type PI, chez l'humain et le nématode, seraient donc concentrées au sein de modules fonctionnels associés aux même phénotypes et maladies. Il fait du sens que les altérations du métabolisme énergétique associées à la perturbation de la glycolyse, ou à celle de la respiration mitochondriale, soient associées aux mêmes pathologies. Notre étude suggère donc que les gènes associés à l'effet Warburg soient des gènes présentant des fonctions de coordination de processus de proximité à l'intérieur de l'interactome fonctionnel des tumeurs mammaires humaines. Il serait aussi intéressant de voir si une approche similaire nous permettrait d'identifier une classe de gènes ayant une fonction de coordination de processus distant via l'identification d'interactions de type connecteurs.

5.5 Pertinence de notre approche intégrative pour la découverte de gènes impliqués dans différents contextes d'une maladie

Comme démontré dans le dernier chapitre, le modèle prédictif basé sur notre approche intégrative est beaucoup plus performant que celui basé sur des données d'expressions seulement. Il est donc raisonnable de croire que notre approche, grâce aux différents attributs dont certains dérivés à partir des données d'expression, permette de caractériser les réseaux fonctionnels avec beaucoup plus de précision qu'en utilisant l'expression des gènes seulement. Bien évidemment, il est primordial de tester d'autres approche de prédiction que la régression linéaire simple présenté dans le dernier chapitre.

Dans le même ordre d'idée, un des champs d'études parmi les plus prometteurs ayant mené à la caractérisation de réseaux fonctionnels de plusieurs types de cancer est la protéogénomique (Locard-Paulet *et al.*, 2016; Mertins *et al.*, 2016; Zhang, B. *et al.*, 2014; Zhang, H. *et al.*, 2016). Ce champ de recherche combine non seulement la protéomique et la phosphoprotéomique (Locard-Paulet *et al.*, 2016), mais intègre aussi souvent des données d'expression des gènes et de nombre de copies des gènes (Mertins *et al.*, 2016; Zhang, B. *et al.*, 2014; Zhang, H. *et al.*, 2016). Ces dernières études ont non seulement permis d'identifier des gènes conducteurs potentiels (dont leur altération est favorable à la progression tumorale), mais ont aussi permis d'identifier les voies de signalisation, et donc les contextes fonctionnels qui sont altérés par les mutations de ces gènes (Mertins *et al.*, 2016; Zhang, B. *et al.*, 2014; Zhang, H. *et al.*, 2016). Il serait donc intéressant de voir où se situe notre méthode d'intégration et notre caractérisation des réseaux fonctionnels du cancer du sein par rapport aux résultats obtenus par l'entremise de la protéogénomique. Un avantage certain de notre approche vis-à-vis cette dernière est le fait que notre approche ne nécessite que l'obtention de données d'expression pour le tissu et échantillon de tumeur. En effet, à l'inverse des données d'expression qui sont accessibles et peu coûteuses, les données de protéomique sont

beaucoup moins accessibles et sont plutôt coûteuses à obtenir pour de larges cohortes de patients (Di Meo *et al.*, 2014; White, 2011). Notre méthode pourrait s'avérer être une alternative peu coûteuse et très accessible pour identifier des cibles thérapeutiques pour différentes pathologies incluant tous les types de cancer.

Enfin, toute chose n'étant pas parfaite, notre méthode d'intégration possède tout de même quelques faiblesses. Parmi celles-ci, les interactions fonctionnelles caractérisées ou prédites par notre méthode ne sont pas directionnelles. Autrement dit, il n'est pas possible de prédire si une interaction fonctionnelle donnée a un effet positif (e.g. activation) ou négatif (e.g. inhibition) dans le contexte étudié. Par exemple, la mesure de la force contextuelle (FC) décrite pour chaque interaction et la moyenne des forces contextuelles des interactions impliquant un gène dans un contexte A versus un contexte B, ne nous permet pas de prédire si ce gène est plus "actif" ou "inactif" dans chacun de ces contextes. Par exemple, si le gène X a des interactions présentant un FC plus fort dans le contexte A que dans le B, ceci n'indique pas que ce gène soit plus actif dans le contexte A. En effet, ce gène pourrait être impliqué dans un plus grand nombre d'interactions inhibitrices dans le contexte A et sera donc plus inactif que dans le contexte B.

Pour remédier à cette faiblesse, il serait possible d'identifier, chez l'humain, les interactions fonctionnelles inhibitrices et activatrices dans les voies de signalisation/métabolisme afin de caractériser leurs signatures basées sur nos attributs et/ou d'autres propriétés. Ces signatures pourraient ensuite servir à départager, dans le contexte d'une maladie humaine, les interactions fonctionnelles ayant potentiellement un effet positif de celles ayant un effet négatif et de les utiliser comme ensembles d'entrainements dans un modèle prédictif. Ce modèle prédictif peut ensuite servir à classifier, à l'échelle du génome, les paires de gènes selon leur potentiel inhibiteur ou activateur. Cette dernière information pourrait aussi être intégrée au calcul de la force contextuelle, vu dans le dernier chapitre, afin de mieux caractériser les voies et modules

fonctionnels activés ou désactivés, et d'identifier avec une meilleure précision la plasticité des réseaux entre des cohortes de patients.

5.7 Directions futures et suivi expérimental

Grâce à l'information provenant de l'analyse en profondeur d'interactomes génétiques et au modèle prédictif permettant d'identifier des gènes potentiellement associés à l'effet Warburg, plusieurs possibilités de suivi expérimental pourraient être réaliser dans les prochaines années.

Parmi celles-ci, notons la validation expérimentale des prédictions, provenant du dernier chapitre, chez le nématode *Caenorhabditis elegans*. En utilisant les orthologues des gènes prédits comme jouant un rôle dans l'effet Warcburg dans le cancer du sein chez l'humain (chapitre IV). Il serait possible de caractériser, chez le nématode, les rôles joués par ces orthologues dans des processus clés du développement cellulaire de cet animal. Il a été montré, dans plusieurs études récentes, que *C. elegans* pourrait utiliser l'effet Warburg lorsque la respiration mitochondriale est compromise par inactivation chimique (Jang *et al.*, 2016; Luz *et al.*, 2016). De plus, les essais de développement et de locomotion de larves développés dans ces études pourraient servir à évaluer le rôle énergétique critique joué par des gènes prédits par notre modèle pour contrôler l'effet Warburg chez l'humain. Ces tests seraient réalisés chez des animaux présentant des déficiences de la respiration mitochondriale et soumis à un ARN interférant (ARNi) ciblant ces gènes. Par la suite, les gènes identifiés dans le contrôle de l'effet Warburg chez *C. elegans* seraient testés pour leur fonction pendant la prolifération cellulaire et la migration dans les cellules cancéreuses du sein chez l'humain. Pour ce faire, des cellules associées aux sous-types de tumeurs mammaires luminal et basal, affichant un faible et un haut niveau de glycolyse aérobie

respectivement (Brauer *et al.*, 2013; Palaskas *et al.*, 2011), seraient utilisées. Des résultats concluants permettraient d'établir *C. elegans* comme modèle pour l'étude de l'effet Warburg et de valider notre modèle prédictif.

ANNEXE A

Text S3.1

Supplementary information

Robustness of GIs-all clustering into classes

We classified GIs from GIs-all using a cluster selection algorithm (Figure 3.1b; see Methods). Briefly, for each pre-selected cluster obtained by cutting the dendrogram at a certain height, we assessed the information content with the approximately unbiased bootstrap probability, which indicates how data support the clusters (see Methods). Secondly, we evaluated the classification performance of a logistic regression model using the cluster information content as a positive learning set and computed the performance with a leave-one-out cross-validation (LOOCV) test. A receiver operating characteristic (ROC) curve was then built following LOOCV results, which gave AUC scores between 0 and 1 (0.5 for a random prediction and 1 for a perfect prediction; Figure 3.S1). We judged that equivalently sized clusters with high AUCs were preferable over disparately sized clusters to ensure objective comparisons in further analysis. Thus, clustering the positive set of GIs identified 10 clusters and the six with AUC scores higher than 0.8 (C1 to C6) were subsequently analyzed to identify distinctive properties of GI classes (Figure 3.S1).

Impact of missing data on GI clustering

Despite the ever-growing amount of data coming from genome-wide studies, certain types of data, such as PPIs and phenotypical characterization of genetic alterations, still present very poor genome coverage. In addition, while all genes have expression data available, missing data is reported for the co-expression attribute for a certain number

of genes that have expression values for less than 80% of the data used in our study. These data were used to compute five out of six of our attributes (see Methods; Lee *et al.* 2010). The sixth one (Exp) relies on expression data, which present high genome-coverage (Lee *et al.* 2010). Before further characterizing the different classes of GIs, we asked whether the genomic signatures associated with each class might result from the sparse nature of genomic data used in the clustering. To evaluate this possibility, we calculated the percentage of GIs in each class missing expression data, PPI or phenotypical characterizations (Figure 3.S2). While a certain amount of data were missing in each classes of GIs, detailed analysis revealed that missing expression, phenotype and/or PPI data did not significantly influence the distinctive properties observed for the six selected classes of GIs identified by our clustering technique. For example, GIs found in C1 and C2 segregate from C5 and C4 based on the absence of direct interaction between protein products of interacting genes (Int = 0; see Methods; Figure 3.1b) and the absence of an unexpectedly high number of common protein partners for these proteins within the PPI network (CI $P \geq 0.05$; see Methods; Figure 3.1b). At most 31.7% of GIs clustered in C1 and C2 lack PPI data (Figure 3.S2), suggesting that up to 31.7% of these interactions may segregate into C4 and C5 if using more complete protein interaction data. While this constitutes a significant portion of the GIs found in C1 and C2, these data suggest that the majority of C1 and C2 (more than 68%) are significantly distinct from C5 and C4 and consequently, that C1 and C2 constitute a real distinct class from C4 and C5.

C5 is composed of 51.7% of GIs expressing the phenotype enriched in their neighbourhood (*NPh* =1, see Methods; Figure 3.1b) and expressing proteins that share a large number of common interactors within the PPI network (*CI* < 0.05, see Methods; Figure 3.1b). These interactions could then segregate from C4 GIs based on the values of the *NPh* attribute (this value being null for C4 GIs). 60.4% of C4 GIs does not have any assigned phenotype upon genetic alteration. The lack of phenotypic data is then responsible for the null values of *NPh* for 60.4% of C4 GIs indicating that at least one

of the interacting genes is not essential to any biological function assayed during genome-wide RNAi screening, or that the RNAi experiment used was not efficient enough to identify any phenotype. The cumulative false negative proportion (i.e. the frequency of genes having no phenotype association because of technical artefact) of genome-wide RNAi screenings reported in WormBase (release WS180) is impossible to estimate and may be responsible for the clustering of GIs in C4 instead of C5. C4 GIs are also distinguished from the second half of C5 GIs based on *I* and *CI* attributes. No missing PPI data are observed for C4 and C5 GIs suggesting that the segregation of these two clusters cannot be explained by missing PPI data.

C2 GIs were distinct from C1 based on the fact that C2 gene pairs do not express the common phenotype(s) enriched in their neighborhoods (*NPh* = 0, see Methods; Figure 3.1b). However, no missing phenotypical data is observed for C1 and C2 GIs, suggesting that different *NPh* values are a real distinctive property for differentiating C2 from C1 GIs. Missing expression data in C2 (Figure 3.S2) may also partially explain the lower average coexpression value observed in C2 when compared to C1, (Figure 3.1b).

C3 and C6 segregated from the other clusters based on the absence of a significant enrichment of similar phenotypical profiles (*Ph* =1; see Methods; Figure 3.1b). However, none of the GIs in C3 and C6 lack phenotypical data (Figure 3.S2), suggesting that the absence of a significant enrichment of similar phenotypical profiles is a real distinctive property of these two GI classes.

C3 and C6 distinguished themselves from the other clusters based on low coexpression values of interacting genes in C6 and on the lack of phenotype enriched in their neighbourhood ($N \geq 0.05$; see Methods, Figure 3.1b). Expression data are missing in 25.6% of GIs from C6 (Figure 3.S2). These data play also an important role in the computation of the *N* attribute since the neighbourhood of a gene is defined by its significantly coexpressed genes and the PPI partners of its protein product. This suggests that up to 26% of GIs from C6 may redistribute to C3, given more complete

data sets. This also implies that the majority of C6 (more than 74%) retains distinctive properties when compared to C3.

Overall, these data show that missing expression, phenotype and/or PPI data cannot explain the classification of GIs-All in six classes with distinctive properties.

GI classes assembly into GI dense subnetworks

We tested whether GIs may assemble into GI dense subnetworks (GDS). To test this possibility, we identified GDS within GIs-all using the cytoscape "MINE" plugin (Rhrissorrakrai et Gunsalus 2011). We identified 42 GDS (Figure 3.S3). For each GDS, we computed a monochromacity score (MS) adapted from similar study done on yeast GI network (Szappanos *et al.* 2011). Positive and negative MS identified GDS enriched or depleted respectively in a given GIs class (see Methods). Clustering of GI classes using Euclidian distance based on MS revealed that GI classes tend to cluster by pair within GDS: C1 with C2, C4 with C5 and C3 with C6 (Figure 3.S3). Measurement of the enrichment of combination of classes using hypergeometric test also revealed that C1 is enriched alone or in combination with C2, C3 or C6 in a few GDS (Figure 3.S3b). Similarly, C4 was found enriched alone or in combination with C5 (Figure 3.S3b). These data suggest that while GDS number is too small to clearly measure a monochromaticity of GDS, GI classes tend to assemble in GDS in a biased manner.

We supported this assumption through measurement of the amount of overlapping genes in the selected GI classes. The rational behind this measurement is that GI classes that had a propensity to assemble in GDS together may also share common genes. To do so, we hierarchically clustered gene frequencies and used the resulting dendrograms as an indication of proximity between GI classes (Figure 3.S4). Two distance metrics

were used to distinguish GI classes: Canberra distance was used to cluster GI classes when considering repetition of genes within classes (due to the potential involvement of genes in more than one GI). We used also Binary distance to cluster these classes base on gene composition without considering frequency. This study revealed that gene sets appear to be different between GI clusters (Binary; Figure 3.S3), and that the most connected genes are also largely different from one cluster to another (Canberra; Figure 3.S3). Interestingly, irrespective of the clustering method used, GI classes distributed in three groups, C1/C2/C3, C4/C5 and C6 based on gene overlaps (Figure 3.S4).

Enrichment in within- and between-PDS within pathways does not depend on the PPI network topology

We tested whether the size and the number of PDS may influence the enrichment of within- and between-PDS, within pathways. To do so, we analyzed the distribution of PDS identified from our PPI network (Figure 3.S14a), cut this distribution and generated several networks with different combination of small and large PDS. As seen in Figure 3.S14b, these networks are composed of:

PDS-1: 30 PDS of less than 10 proteins

PDS-2: 10 PDS of more than 99 proteins

PDS-3: 90 PDS of less than 30 proteins

PDS-4: 50 PDS of more than 50 proteins

PDS-5: 100 PDS of less than 50 proteins

PDS-All: As a control we used the network composed of all PDS found in our PP-interactome

PDS-Rdm: In order to strengthen this study, we used node permutation to generate 100 random networks displaying the same topology than PDS-all (same number of interactions, same number of PDS, same PDS size-distribution, same between PDS interactions; Figure 3.S14c and 3.S14d). These random networks are called PDS-Rdm (Figure 3.S14b) and contain less than 3% overlapping edges with the PDS-all.

We then used these networks to assess the enrichment in within- and between-PDS interaction found within pathway as detailed in the Method section. This analysis showed that except for PDS-1 containing only PDS of less than 10 proteins, PDS-2 to -5 and PDS-all display a significant enrichment in within- and between-PDS interactions within pathways (Figure 3.S14b). In addition, PDS-Rdm with the same topology than PDS-all did not display any enrichment neither in within- nor in between-PDS and even appears significantly depleted in these two kinds of interactions (Figure 3.S14b). These data suggest that the enrichment of within- and between-PDS observed within-pathways did not depend on the topology of the PDS network but instead depends on the interactions themselves.

These data suggest that pathways and PDS are distinct functional modules. They also suggest that pathways are composed of several PDSs and that proteins involved in the same pathway may be part of the same PDS or of different PDS. We confirmed these assumptions through a close examination of KEGG pathways and PDS as shown in Table 3.S4 (en ligne: Boucher *et al.* 2016). We identified for each KEGG pathways and pathways from the literature genes involved in the same PDS (same-pathway & same-PDS) and genes involved in the same pathway and in different PDS (same pathway & different PDS). This study revealed that our assumption is exact and that up to 6 PDS may contain proteins involved in the same pathway and up to 4 proteins of the same PDS may be also involved in the same pathway. For example, Vulval development is a pathway involving genes coding for proteins present in 5 different PDS. two genes LIN-45 and LET-60 are part of the same PDS, while 4 other genes, LIN-3, SEM-5, KSR-1 and LET-23 code for proteins that are part of 4 different PDS

(Table 3.S4; en ligne: Boucher *et al.* 2016).

Distribution of GI classes within and between ranges of Pleiotropy

We looked at the distribution of GI classes at different ranges of pleiotropy. To do so, we measured the enrichment of GI classes in GI subnetworks composed of genes within a similar PI range, either PI higher or equal to a certain threshold (τ) (upper panel, Figure 3.S17) or PI lower or equal to τ (middle panel, Figure 3.S17). Log Odds ratio and hypogeometric *P*-values are indicated in Table 3.S5 (en ligne: Boucher *et al.* 2016). Interestingly, we found that C1, C2 and C6 GIs are enriched in interaction between genes with high PI ($PI \geq 7$ for C1 and $PI \geq 6$ for C2 and C6; upper panel, Figure 3.S17). On the other side, C4 is highly enriched in interactions between genes with lower PIs ($PI \leq 2$; middle panel, Figure 3.S17). C5 and C3 are enriched in interaction between genes with an average PI ($4 \leq PI \leq 5$ and $5 \leq PI \leq 6$, respectively).

GIs may also link genes in different PI ranges. To test this possibility, we measured the enrichment of GI classes in sets of interactions linking genes with a PI higher or equal to a certain threshold ($PI \geq \tau$) with genes having a PI below this threshold ($PI < \tau$). We tested this for every τ between 1 and 10 (lower panel, Figure 3.S17). These data showed that C1, C2 and C6 link genes with average to high PIs (between 4 and 7; higher panel, Figure 3.S17) to genes with an even higher PI (from 6 to 10; lower panel, Figure 3.S17). These data also show that C4 links genes with low PIs (from 0 to 2; middle panel, Figure 3.S17) to genes having an average PI (between 2 and 5; lower panel, Figure 3.S17). C3 appeared to do the opposite of C4, linking genes of average to high PIs (from 5 to 6; middle panel, Figure 3.S17) to genes with lower PI (from 2 to 5; lower panel, Figure 3.S17). C5 interactions appeared not to be enriched across ranges of PIs, suggesting that C5 interactions link genes mainly within an average PI range (from 4

to 5; middle panel, Figure 3.S17). Note that GI between genes that have very different PI will be considered to compute the Log Odds ratio more than once (meaning at different threshold (τ)). For example an interaction between a gene with PI=9 and a gene with PI=7 will be identified with $\tau = 9$ and 8.

Altogether these data further characterize PDS-centric (C4-C5) functional modules composed of elements with average and low PIs, and PDS-independent (C1-C2) functional modules linking genes with high PIs. These data also further distinguish C3 from C6 connectors, with C6 linking genes from an average to a high level of pleiotropic and C3 interactions linking genes from an average to a low pleiotropic level.

References

Lee AY, Perreault R, Harel S, Boulier EL, Suderman M, et al. (2010) Searching for signaling balance through the identification of genetic interactors of the Rab guanine-nucleotide dissociation inhibitor gdi-1. PLoS ONE 5.

Rhrissorrakrai K, Gunsalus KC (2011) MINE: Module Identification in Networks. BMC Bioinformatics 12: 192.

Szappanos B, Kovacs K, Szamecz B, Honti F, Costanzo M, et al. (2011) An integrated approach to characterize genetic interaction networks in yeast metabolism. Nat Genet 43: 656-662.

ANNEXE B

Supplementary data

Structural function of classes of FI in C. elegans and human

"Connector" interactions (C) were enriched in GI hubs in the *C. elegans* GI network and were proposed to have coordination function within this network (Boucher *et al.*, 2016). Due to the lack of GI data in human breast cancer, we addressed the potential coordination function of C through measurement of their betweeness indices of FIs identified in each network. This study is used to identify edges within a network that are used in more paths than expected by chance (Suppl. Fig. 1A). Edges with high betweenness are called bottleneck edges. The analogy for these edges are bridges linking dense networks of roads within a city located on an island to those located on the shore (Yu *et al.*, 2007). Measurement of betweenness for GIs of the *C. elegans* network characterized in our previous study identified Cm and Cw interactions (C3 and C6 respectively; Suppl. Fig. 1B-D) and PIm (C2; Suppl. Fig. 1B-D) as bottleneck edges of the networks, particularly for interactions involving non-Hub genes (Suppl.l. Fig. 1D). Similarly, Cs, Cm, Cw and PIw displayed significantly higher betweenness than other classes of FIs in both Luminal A and Basal-like (Suppl. Fig. 1E and F respectively). All together, these data suggest that FI classes display distinctive plasticity properties and structural functions within FI networks. These data also suggest that in both human and *C. elegans*, C, and to a lesser extent, PI interactions tend to be bottleneck interactions, suggesting that genes linked to their partners through these classes of interactions may display more critical coordination function between biological processes that those that do not.

Characterisation of predictive model of Warburg effect genes

To better characterize the model generated to predict the Warburg effect genes described above, we considered the respective contribution of the percentage of PI, C and plastic interaction to this model. We generated three models using: only percentage of PI (Suppl. Fig 2A); only the percentage of C (Suppl. Fig. 2B); and only the percentage of plastic interactions (Suppl. Fig. 2C). This analysis revealed that none of these models were as performant as the linear model combining these three groups of attributes. It also revealed that set of genes identified with the lowest percentage of plastic interactions are depleted in Warburg effect genes, suggesting that while lower plasticity value than average display a significant predictive power to identify Warburg effect genes, a minimal plasticity is still an hallmark of this group of genes.

References

Boucher, B., Lee, A.Y., Hallett, M. et Jenna, S. (2016). Structural and Functional Characterization of a Caenorhabditis elegans Genetic Interaction Network within Pathways. *PLoS Comput Biol, 12*(2), e1004738. doi: 10.1371/journal.pcbi.1004738

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), e59. doi: 10.1371/journal.pcbi.0030059

ANNEXE C

Luminal A	Luminal B	HER2-E	Basal-like	Normal-like
TCGA-BH-A0EA-01	TCGA-D8-A3Z6-01	TCGA-E9-A1RF-11	TCGA-AR-A5QQ-01	TCGA-BH-A1EW-11
TCGA-B6-A0I5-01	TCGA-AC-A5XU-01	TCGA-BH-A0DD-01	TCGA-C8-A131-01	TCGA-E2-A1BC-11
TCGA-AQ-A54N-01	TCGA-BH-A0BT-11	TCGA-E9-A1NC-01	TCGA-A2-A3XS-01	TCGA-BH-A0HX-01
TCGA-E9-A1R6-01	TCGA-Z7-A8R6-01	TCGA-AO-A1KQ-01	TCGA-AR-A1AO-01	TCGA-BH-A1FJ-11
TCGA-B6-A0X4-01	TCGA-A2-A04Q-01	TCGA-E2-A1LH-11	TCGA-BH-A0DT-01	TCGA-OL-A66K-01
TCGA-D8-A145-01	TCGA-BH-A1FM-11	TCGA-BH-A0BW-01	TCGA-D8-A147-01	TCGA-AN-A0AT-01
TCGA-E2-A1II-01	TCGA-AR-A24R-01	TCGA-A8-A06Y-01	TCGA-E9-A1RD-11	TCGA-AC-A3EH-01
TCGA-GM-A2DO-01	TCGA-A2-A25E-01	TCGA-E2-A14V-01	TCGA-AR-A5QM-01	TCGA-AO-A0JM-01
TCGA-B6-A0WV-01	TCGA-BH-A1FU-11	TCGA-A2-A0T6-01	TCGA-AC-A2BK-01	TCGA-C8-A1HG-01
TCGA-A1-A0SP-01	TCGA-E9-A1NG-01	TCGA-E2-A1LB-11	TCGA-E2-A1IO-01	TCGA-AO-A1KT-01
TCGA-BH-A0HP-01	TCGA-BH-A0DS-01	TCGA-AR-A2LR-01	TCGA-LL-A5YO-01	TCGA-C8-A12V-01
TCGA-BH-A0DZ-01	TCGA-BH-A0BS-11	TCGA-B6-A401-01	TCGA-OL-A5DA-01	TCGA-D8-A1XA-01
TCGA-E9-A1NA-01	TCGA-AR-A0U4-01	TCGA-B6-A0RN-01	TCGA-D8-A27R-01	TCGA-E2-A1IU-01
TCGA-EW-A1PG-01	TCGA-AO-A0JB-01	TCGA-AN-A0FS-01	TCGA-E9-A227-01	TCGA-3C-AALI-01
TCGA-AC-A6IX-01	TCGA-PE-A5DC-01	TCGA-A2-A4S3-01	TCGA-E9-A1RF-01	TCGA-AN-A0FK-01
TCGA-BH-A18P-01	TCGA-E9-A1RD-01	TCGA-A8-A08Z-01	TCGA-AR-A0TQ-01	TCGA-E2-A150-01
TCGA-AN-A0AK-01	TCGA-A2-A25C-01	TCGA-C8-A1HE-01	TCGA-A8-A0A2-01	TCGA-PL-A8LX-01
TCGA-E9-A22D-01	TCGA-XX-A89A-01	TCGA-BH-A0HU-01	TCGA-A2-A1G1-01	TCGA-AN-A03X-01
TCGA-BH-A0DK-11	TCGA-BH-A0B5-11	TCGA-A8-A0A1-01	TCGA-A2-A1FV-01	TCGA-S3-A6ZH-01
TCGA-BH-A0BF-01	TCGA-BH-A0E7-01	TCGA-EW-A1P1-01	TCGA-E2-A573-01	TCGA-AN-A049-01
TCGA-D8-A1XS-01	TCGA-D8-A1JS-01	TCGA-D8-A27I-01	TCGA-D8-A27V-01	TCGA-OL-A6VQ-01
TCGA-A2-A25A-01	TCGA-XX-A899-01	TCGA-S3-AA15-01	TCGA-A7-A3IZ-01	TCGA-AR-A251-01
TCGA-AO-A0J7-01	TCGA-E2-A15I-01	TCGA-AR-A1AV-01	TCGA-A2-A0T4-01	TCGA-BH-A1FG-11
TCGA-E9-A243-01	TCGA-BH-A1FL-01	TCGA-EW-A1PD-01	TCGA-EW-A1OV-01	TCGA-LL-A7SZ-01
TCGA-BH-A0AY-01	TCGA-E9-A248-01	TCGA-E2-A15D-01	TCGA-GM-A2DL-01	TCGA-A8-A06P-01
TCGA-BH-A18S-01	TCGA-OL-A5RU-01	TCGA-A2-A25F-01	TCGA-AC-A2FF-01	TCGA-D8-A1X6-01
TCGA-E2-A158-11	TCGA-A2-A0ET-01	TCGA-C8-A1HI-01	TCGA-A7-A425-01	TCGA-AC-A2FF-11
TCGA-D8-A27N-01	TCGA-C8-A138-01	TCGA-A8-A07W-01	TCGA-BH-A0BA-11	TCGA-E2-A15P-01
TCGA-A8-A06Q-01	TCGA-EW-A1P5-01	TCGA-B6-A0IN-01	TCGA-BH-A0DV-11	TCGA-E2-A14Y-01
TCGA-AO-A0JE-01	TCGA-E9-A5FK-01	TCGA-E9-A1N6-11	TCGA-A8-A06N-01	TCGA-EW-A1PH-01
TCGA-AR-A2LL-01	TCGA-BH-A1FN-11	TCGA-BH-A0H9-01	TCGA-A2-A0CT-01	TCGA-E2-A1LH-01
TCGA-C8-A12Q-01	TCGA-D8-A1XT-01	TCGA-A7-A0CJ-01	TCGA-BH-A0DH-11	TCGA-AC-A5EH-01
TCGA-A2-A0CW-01	TCGA-A8-A09W-01	TCGA-A1-A0SM-01	TCGA-A2-A0CY-01	TCGA-E2-A3DX-01
TCGA-BH-A0W4-01	TCGA-EW-A6SD-01	TCGA-D8-A1J8-01	TCGA-AC-A3W7-01	TCGA-A7-A0DC-11
TCGA-D8-A1J9-01	TCGA-AO-A12C-01	TCGA-A8-A07O-01	TCGA-E2-A15A-01	TCGA-B6-A0RS-01
TCGA-E2-A15K-01	TCGA-C8-A273-01	TCGA-OL-A66N-01	TCGA-B6-A408-01	TCGA-E9-A1RH-01
TCGA-A2-A0YE-01	TCGA-E9-A1NE-01	TCGA-A8-A075-01	TCGA-A2-A3KD-01	TCGA-A8-A07F-01
TCGA-BH-A0B8-11	TCGA-JL-A3YX-01	TCGA-OL-A5D6-01	TCGA-AO-A0JG-01	TCGA-LL-A441-01
TCGA-BH-A0DP-01	TCGA-E2-A2P5-01	TCGA-D8-A1JC-01	TCGA-E2-A1AZ-01	TCGA-D8-A1JP-01
TCGA-BH-A204-01	TCGA-AR-A2LM-01	TCGA-A8-A08G-01	TCGA-A2-A0CV-01	TCGA-S3-A6ZF-01
TCGA-E9-A1RI-01	TCGA-AO-A12A-01	TCGA-A2-A0SY-01	TCGA-AR-A2LE-01	TCGA-A8-A09I-01

TCGA-E9-A1RH-11	TCGA-E2-A14N-01	TCGA-BH-A1FH-11	TCGA-EW-A6S9-01	TCGA-D8-A73U-01
TCGA-AC-A3HN-01	TCGA-JL-A3YW-01	TCGA-BH-A42T-01	TCGA-AC-A8OQ-01	TCGA-A8-A095-01
TCGA-BH-A8FY-01	TCGA-OL-A5RV-01	TCGA-A2-A0YG-01	TCGA-BH-A1FU-01	TCGA-A2-A0ES-01
TCGA-A2-A04X-01	TCGA-A8-A097-01	TCGA-BH-A0DK-01	TCGA-BH-A0BZ-11	TCGA-AR-A0U0-01
TCGA-B6-A0IP-01	TCGA-WT-AB41-01	TCGA-C8-A133-01	TCGA-LL-A6FP-01	TCGA-BH-A0B3-11
TCGA-AO-A128-01	TCGA-E2-A15E-06	TCGA-E9-A1RI-11	TCGA-E2-A14P-01	TCGA-A7-A26I-01
TCGA-C8-A12M-01	TCGA-D8-A1Y3-01	TCGA-OL-A66L-01	TCGA-E9-A54Y-01	TCGA-AR-A2LH-01
TCGA-AC-A23G-01	TCGA-A2-A0EU-01	TCGA-A7-A5ZX-01	TCGA-LD-A66U-01	TCGA-EW-A1J6-01
TCGA-A8-A06X-01	TCGA-BH-A0EI-01	TCGA-E9-A2JS-01	TCGA-A7-A3J0-01	TCGA-AR-A0TS-01
TCGA-A8-A09V-01	TCGA-AN-A0XT-01	TCGA-AC-A6IW-01	TCGA-BH-A18R-01	TCGA-AR-A2LJ-01
TCGA-A8-A07L-01	TCGA-D8-A142-01	TCGA-E2-A1L7-11	TCGA-D8-A1JL-01	TCGA-AC-A3QP-01
TCGA-D8-A27M-01	TCGA-AO-A0JL-01	TCGA-3C-AALJ-01	TCGA-BH-A0DH-01	TCGA-E9-A226-01
TCGA-LL-A440-01	TCGA-AC-A2QJ-01	TCGA-BH-A208-11	TCGA-AO-A124-01	TCGA-E9-A3HO-01
TCGA-Z7-A8R5-01	TCGA-E9-A54X-01	TCGA-B6-A0IO-01	TCGA-E2-A10A-01	TCGA-BH-A18N-01
TCGA-OL-A5D8-01	TCGA-LL-A6FQ-01	TCGA-AO-A03M-01	TCGA-AR-A1AU-01	TCGA-AC-A7VC-01
TCGA-E2-A10F-01	TCGA-BH-A0HK-11	TCGA-LL-A7T0-01	TCGA-D8-A1X5-01	TCGA-D8-A1JN-01
TCGA-E2-A108-01	TCGA-BH-A1F5-01	TCGA-BH-A0BA-01	TCGA-AN-A0FY-01	TCGA-E9-A1R2-01
TCGA-AR-A255-01	TCGA-BH-A18P-11	TCGA-A2-A04N-01	TCGA-D8-A27T-01	TCGA-AR-A256-01
TCGA-GM-A4E0-01	TCGA-AO-A1KO-01	TCGA-E2-A1L6-01	TCGA-D8-A1XB-01	TCGA-BH-A0H6-01
TCGA-PE-A5DE-01	TCGA-A2-A25D-01	TCGA-E9-A5UO-01	TCGA-A2-A0YF-01	TCGA-BH-A0B7-11
TCGA-PL-A8LY-01	TCGA-AR-A1AW-01	TCGA-A2-A3XZ-01	TCGA-BH-A0BJ-11	TCGA-GM-A5PX-01
TCGA-AC-A2FB-11	TCGA-A2-A0D0-01	TCGA-OL-A5RY-01	TCGA-AO-A0JF-01	TCGA-A8-A09D-01
TCGA-A2-A4S1-01	TCGA-A7-A0CG-01	TCGA-BH-A18R-11	TCGA-BH-A0DT-11	TCGA-AR-A0U3-01
TCGA-E2-A1LG-01	TCGA-D8-A3Z5-01	TCGA-OK-A5Q2-01	TCGA-AN-A0G0-01	TCGA-GI-A2C9-11
TCGA-E2-A106-01	TCGA-BH-A1EN-11	TCGA-EW-A423-01	TCGA-D8-A1JT-01	TCGA-A1-A0SQ-01
TCGA-AC-A2FE-01	TCGA-B6-A0X0-01	TCGA-E9-A3Q9-01	TCGA-E2-A1IK-01	TCGA-BH-A0H5-11
TCGA-BH-A1EU-01	TCGA-A7-A26E-01	TCGA-AQ-A7U7-01	TCGA-A7-A6VW-01	TCGA-D8-A27F-01
TCGA-LL-A5YL-01	TCGA-BH-A1ES-06	TCGA-LD-A74U-01	TCGA-A8-A0A9-01	TCGA-E2-A14O-01
TCGA-A7-A4SF-01	TCGA-E2-A15M-01	TCGA-BH-A0B6-01	TCGA-A1-A0SB-01	TCGA-BH-A0C7-01
TCGA-BH-A0C3-01	TCGA-E9-A1N8-01	TCGA-E9-A3QA-01	TCGA-E9-A1RA-01	TCGA-A7-A4SB-01
TCGA-B6-A0I1-01	TCGA-C8-A8HP-01	TCGA-BH-A8G0-01	TCGA-OL-A5RZ-01	TCGA-E2-A15R-01
TCGA-BH-A1F0-01	TCGA-E9-A1ND-11	TCGA-AC-A3W5-01	TCGA-4H-AAAK-01	TCGA-D8-A1JA-01
TCGA-A7-A13E-11	TCGA-EW-A1J3-01	TCGA-C8-A26X-01	TCGA-EW-A2FR-01	TCGA-A7-A0DC-01
TCGA-C8-A12T-01	TCGA-BH-A0B3-01	TCGA-E2-A1IL-01	TCGA-A8-A09C-01	TCGA-A2-A3XU-01
TCGA-A1-A0SD-01	TCGA-A2-A3Y0-01	TCGA-BH-A0HL-01	TCGA-D8-A1JD-01	TCGA-D8-A1XK-01
TCGA-A7-A26J-01	TCGA-C8-A26Z-01	TCGA-D8-A141-01	TCGA-E2-A153-01	TCGA-BH-A1FG-01
TCGA-AO-A0JJ-01	TCGA-AR-A24P-01	TCGA-AN-A0XS-01	TCGA-BH-A202-01	TCGA-E2-A1IG-11
TCGA-GM-A3XG-01	TCGA-A8-A07U-01	TCGA-LL-A8F5-01	TCGA-A2-A0T2-01	TCGA-B6-A0RQ-01
TCGA-A8-A0AD-01	TCGA-AR-A24T-01	TCGA-A2-A3XY-01	TCGA-A2-A0YC-01	TCGA-A2-A0YJ-01
TCGA-EW-A1J2-01	TCGA-A2-A0D3-01	TCGA-BH-A1FM-01	TCGA-BH-A0DO-11	TCGA-EW-A1IZ-01
TCGA-AR-A1AR-01	TCGA-A8-A083-01	TCGA-B6-A0RL-01	TCGA-AN-A0AM-01	TCGA-B6-A409-01
TCGA-BH-A0B2-01	TCGA-D8-A1JI-01	TCGA-A8-A084-01	TCGA-BH-A0C0-11	TCGA-E2-A15T-01

TCGA-BH-A0EE-01	TCGA-BH-A1EY-01	TCGA-AR-A0U2-01	TCGA-AR-A250-01	TCGA-AO-A0JI-01
TCGA-A2-A0YL-01	TCGA-BH-A0BR-01	TCGA-GI-A2C8-11	TCGA-E9-A1RB-11	TCGA-E2-A15M-11
TCGA-A2-A0EW-01	TCGA-GI-A2C8-01	TCGA-BH-A0RX-01	TCGA-BH-A42V-01	TCGA-A2-A3KC-01
TCGA-BH-A0HO-01	TCGA-AC-A8OP-01	TCGA-BH-A1EO-11	TCGA-A2-A0SV-01	TCGA-D8-A1XG-01
TCGA-S3-AA12-01	TCGA-AR-A1AY-01	TCGA-E2-A15H-01	TCGA-D8-A1XL-01	TCGA-A2-A04R-01
TCGA-A2-A04U-01	TCGA-E9-A229-01	TCGA-A8-A0A6-01	TCGA-C8-A8HQ-01	TCGA-OL-A6VO-01
TCGA-E2-A15J-01	TCGA-D8-A1XQ-01	TCGA-E9-A22A-01	TCGA-A7-A13D-01	TCGA-E9-A1R5-01
TCGA-AR-A5QP-01	TCGA-AR-A5QN-01	TCGA-E9-A1NG-11	TCGA-C8-A3M8-01	TCGA-E9-A1ND-01
TCGA-EW-A1PF-01	TCGA-UL-AAZ6-01	TCGA-E9-A1RE-01	TCGA-BH-A0HY-01	TCGA-D8-A143-01
TCGA-BH-A0B9-01	TCGA-BH-A1EV-11	TCGA-E2-A9RU-01	TCGA-E2-A15E-01	TCGA-BH-A209-01
TCGA-A2-A0CP-01	TCGA-BH-A5J0-01	TCGA-AO-A1KS-01	TCGA-BH-A18Q-11	TCGA-BH-A0BD-01
TCGA-WT-AB44-01	TCGA-AR-A24S-01	TCGA-A8-A09B-01	TCGA-A2-A1FW-01	TCGA-E2-A14Z-01
TCGA-C8-A26V-01	TCGA-BH-A1FN-01	TCGA-D8-A1Y0-01	TCGA-PL-A8LZ-01	TCGA-A1-A0SN-01
TCGA-C8-A135-01	TCGA-BH-A0H7-11	TCGA-A7-A6VY-01	TCGA-EW-A1P4-01	TCGA-BH-A1FB-11
TCGA-BH-A18U-11	TCGA-AN-A0FD-01	TCGA-B6-A0RP-01	TCGA-GM-A2DH-01	TCGA-C8-A1HJ-01
TCGA-C8-A12O-01	TCGA-C8-A12P-01	TCGA-AC-A3QQ-01	TCGA-E2-A1LE-01	TCGA-AO-A03N-01
TCGA-BH-A0H9-11	TCGA-BH-A1EW-01	TCGA-A2-A04P-01	TCGA-BH-A0BZ-01	TCGA-BH-A0E6-01
TCGA-D8-A1X7-01	TCGA-AO-A03O-01	TCGA-LL-A5YP-01	TCGA-E9-A1N9-01	TCGA-B6-A0IM-01
TCGA-C8-A1HF-01	TCGA-AR-A24Q-01	TCGA-BH-A1FJ-01	TCGA-EW-A424-01	TCGA-BH-A0HI-01
TCGA-GM-A2DC-01	TCGA-BH-A18L-11		TCGA-A7-A4SC-01	TCGA-E9-A228-01
TCGA-D8-A27K-01	TCGA-A8-A0A4-01		TCGA-A2-A0EV-01	TCGA-A7-A4SA-01
TCGA-BH-A0BW-11	TCGA-D8-A1Y2-01		TCGA-D8-A1XF-01	TCGA-A2-A0YH-01
TCGA-BH-A1F0-11	TCGA-B6-A0WY-01		TCGA-B6-A0IC-01	TCGA-BH-A0BT-01
TCGA-A8-A07I-01	TCGA-GM-A2DF-01		TCGA-A7-A3J1-01	TCGA-BH-A0DQ-01
TCGA-AO-A03U-01	TCGA-D8-A1XV-01		TCGA-D8-A1X8-01	TCGA-BH-A0BM-11
TCGA-A7-A13G-01	TCGA-AR-A2LO-01		TCGA-LD-A7W6-01	TCGA-E2-A152-01
TCGA-BH-A0DQ-11	TCGA-BH-A1EO-01		TCGA-B6-A0RH-01	TCGA-OL-A66P-01
TCGA-AR-A1AP-01	TCGA-A7-A3IY-01		TCGA-C8-A12L-01	TCGA-BH-A0AZ-01
TCGA-D8-A1XR-01	TCGA-E2-A14T-01		TCGA-BH-A0E9-01	TCGA-B6-A1KC-01
TCGA-A2-A0EP-01	TCGA-AN-A0FV-01		TCGA-A8-A076-01	TCGA-AN-A0FN-01
TCGA-D8-A27P-01	TCGA-5L-AAT1-01		TCGA-AN-A0XU-01	TCGA-AO-A0J2-01
TCGA-A8-A09R-01	TCGA-AO-A03R-01		TCGA-BH-A208-01	TCGA-B6-A0RM-01
TCGA-E2-A154-01	TCGA-A8-A0AB-01		TCGA-LL-A50Y-01	TCGA-AN-A0AR-01
TCGA-D8-A1Y1-01	TCGA-AO-A0J5-01		TCGA-A1-A0SO-01	TCGA-C8-A1HO-01
TCGA-BH-A1F6-01	TCGA-B6-A0WZ-01		TCGA-EW-A2FW-01	TCGA-D8-A1JE-01
TCGA-B6-A0RU-01	TCGA-A8-A09N-01		TCGA-BH-A0DZ-11	TCGA-E2-A1IG-01
TCGA-E9-A1N5-01	TCGA-D8-A146-01		TCGA-B6-A0RE-01	TCGA-BH-A18K-01
TCGA-A2-A0CQ-01	TCGA-AR-A24X-01		TCGA-BH-A18J-11	TCGA-A2-A3XV-01
TCGA-S3-AA11-01	TCGA-A8-A06Z-01		TCGA-E2-A1B4-01	TCGA-AR-A2LQ-01
TCGA-C8-A12W-01	TCGA-C8-A12N-01		TCGA-AN-A0FT-01	TCGA-A2-A0CM-01
TCGA-C8-A26W-01	TCGA-AC-A62V-01		TCGA-BH-A18K-11	TCGA-C8-A12U-01
TCGA-AO-A0J9-01	TCGA-D8-A1JK-01		TCGA-AO-A1KP-01	TCGA-LL-A5YM-01

TCGA-AR-A24V-01
TCGA-BH-A18Q-01
TCGA-AQ-A0Y5-01
TCGA-AO-A03V-01
TCGA-AR-A1AL-01
TCGA-BH-A0B0-01
TCGA-A2-A0CR-01
TCGA-AC-A3TM-01
TCGA-A8-A08H-01
TCGA-E2-A1BD-01
TCGA-A8-A09M-01
TCGA-AC-A23H-11
TCGA-AN-A03Y-01
TCGA-A7-A26G-01
TCGA-AN-A0AL-01
TCGA-AO-A12F-01
TCGA-A2-A0SU-01
TCGA-BH-A0BV-11
TCGA-A8-A099-01
TCGA-C8-A27A-01
TCGA-E2-A14W-01
TCGA-D8-A1XU-01
TCGA-A8-A09E-01
TCGA-BH-A1ES-01
TCGA-AO-A03T-01
TCGA-AO-A125-01
TCGA-D8-A1XJ-01
TCGA-A7-A0DB-11
TCGA-A2-A0SX-01
TCGA-E2-A1LS-01
TCGA-AQ-A04J-01
TCGA-AR-A24N-01
TCGA-AC-A2QH-01
TCGA-MS-A51U-01
TCGA-BH-A0DX-01
TCGA-AN-A0XL-01
TCGA-B6-A0WX-01
TCGA-EW-A1OZ-01
TCGA-A1-A0SF-01
TCGA-A2-A0ER-01
TCGA-D8-A1XZ-01
TCGA-A2-A0CL-01
TCGA-AC-A23H-01
TCGA-OL-A97C-01
TCGA-OL-A66H-01
TCGA-AR-A24U-01
TCGA-OL-A5RX-01
TCGA-LL-A740-01
TCGA-E2-A1IF-01
TCGA-E9-A5FL-01
TCGA-BH-A18U-01
TCGA-BH-A203-01
TCGA-D8-A1JG-01
TCGA-E2-A15O-01
TCGA-AC-A2QI-01
TCGA-C8-A275-01
TCGA-E2-A153-11
TCGA-B6-A0IG-01
TCGA-A2-A0T0-01
TCGA-AQ-A54O-01
TCGA-EW-A1J5-01
TCGA-A2-A1FX-01
TCGA-BH-A0GY-01
TCGA-E9-A1RB-01
TCGA-E9-A1R4-01
TCGA-E2-A15C-01
TCGA-BH-A1FE-01
TCGA-EW-A2FS-01
TCGA-BH-A1FC-01
TCGA-BH-A0HB-01
TCGA-AC-A2FK-01
TCGA-E2-A158-01
TCGA-AR-A1AH-01
TCGA-E2-A15S-01
TCGA-BH-A1F2-01
TCGA-AR-A24O-01
TCGA-B6-A0WT-01
TCGA-BH-A0AW-01
TCGA-BH-A0BC-11
TCGA-BH-A0DG-11
TCGA-E9-A1N4-01
TCGA-AN-A041-01
TCGA-A2-A0T5-01
TCGA-E2-A109-01
TCGA-E2-A159-01
TCGA-E2-A1LI-01
TCGA-BH-A18V-06
TCGA-AO-A0J4-01
TCGA-BH-A0B1-01
TCGA-BH-A0BS-01
TCGA-E9-A1NI-01
TCGA-E9-A6HE-01
TCGA-B6-A400-01
TCGA-E9-A244-01
TCGA-AC-A2FM-11
TCGA-BH-A0BQ-01
TCGA-GM-A2D9-01
TCGA-BH-A18J-01
TCGA-A1-A0SH-01
TCGA-BH-A0DL-11
TCGA-A8-A093-01
TCGA-EW-A1P6-01
TCGA-BH-A18I-01
TCGA-BH-A0H7-01
TCGA-BH-A18V-11
TCGA-AR-A1AN-01
TCGA-A2-A0CX-01
TCGA-C8-A12Y-01
TCGA-BH-A0BO-01
TCGA-E2-A1IJ-01
TCGA-D8-A1JB-01
TCGA-A8-A08S-01
TCGA-B6-A0RI-01
TCGA-C8-A12X-01
TCGA-AC-A2FM-01
TCGA-BH-A6R8-01
TCGA-AR-A0TR-01
TCGA-AC-A62X-01
TCGA-AC-A62Y-01
TCGA-AR-A1AK-01
TCGA-AC-A2B8-01
TCGA-AC-A4ZE-01
TCGA-BH-A0B4-01
TCGA-S3-A6ZG-01
TCGA-BH-A0E1-11
TCGA-BH-A0BJ-01
TCGA-AN-A0FX-01
TCGA-C8-A132-01
TCGA-BH-AB28-01
TCGA-E9-A1N9-11
TCGA-E2-A572-01
TCGA-E2-A105-01
TCGA-B6-A2IU-01
TCGA-BH-A5IZ-01
TCGA-AR-A0TV-01
TCGA-A8-A08C-01
TCGA-V7-A7HQ-01
TCGA-LL-A5YN-01
TCGA-S3-AA14-01
TCGA-BH-A1EV-01
TCGA-A2-A0CO-01
TCGA-BH-A0E0-11
TCGA-E9-A1NF-11
TCGA-AO-A0JC-01
TCGA-A2-A1FZ-01
TCGA-GM-A2DI-01
TCGA-OL-A66I-01
TCGA-AN-A0XW-01
TCGA-A7-A0CE-01
TCGA-A7-A6VX-01
TCGA-A8-A086-01

TCGA-EW-A1PB-01	TCGA-OL-A66J-01
TCGA-B6-A1KF-01	TCGA-AR-A24K-01
TCGA-A2-A0EO-01	TCGA-E9-A1NF-01
TCGA-B6-A40B-01	TCGA-E2-A56Z-01
TCGA-BH-A1FR-01	TCGA-AR-A24L-01
TCGA-AO-A0JA-01	TCGA-S3-AA17-01
TCGA-A2-A0EM-01	TCGA-D8-A1JJ-01
TCGA-A2-A3XX-01	TCGA-E2-A155-01
TCGA-AQ-A1H2-01	TCGA-D8-A13Z-01
TCGA-A7-A13G-11	TCGA-E9-A1R0-01
TCGA-A2-A1G6-01	TCGA-A7-A3RF-01
TCGA-A8-A09K-01	TCGA-BH-A0AU-11
TCGA-BH-A0DD-11	TCGA-A2-A0CS-01
TCGA-BH-A1FD-01	TCGA-A2-A25B-01
TCGA-A8-A085-01	TCGA-B6-A0RT-01
TCGA-AN-A0FZ-01	TCGA-BH-A1EN-01
TCGA-A7-A0CD-01	TCGA-C8-A12K-01
TCGA-AO-A0J6-01	TCGA-S3-AA0Z-01
TCGA-A8-A07Z-01	TCGA-A1-A0SG-01
TCGA-EW-A1PA-01	TCGA-GM-A2DA-01
TCGA-BH-A0AV-01	TCGA-A8-A082-01
TCGA-E9-A2JT-01	TCGA-A7-A5ZV-01
TCGA-B6-A0X1-01	TCGA-A7-A56D-01
TCGA-AO-A0J8-01	TCGA-BH-A0AZ-11
TCGA-A7-A0DB-01	TCGA-A7-A13F-01
TCGA-AC-A6IX-06	TCGA-A7-A5ZW-01
TCGA-E9-A1R7-11	TCGA-A7-A0CE-11
TCGA-BH-A1F2-11	TCGA-EW-A1P0-01
TCGA-E2-A15L-01	TCGA-C8-A8HR-01
TCGA-A2-A3XW-01	TCGA-D8-A1XY-01
TCGA-E9-A1RC-01	TCGA-AC-A3BB-01
TCGA-AO-A0JD-01	TCGA-A7-A13E-01
TCGA-GM-A3NY-01	TCGA-BH-A8FZ-01
TCGA-B6-A1KI-01	TCGA-A2-A4S2-01
TCGA-BH-A1F8-11	TCGA-AR-A24W-01
TCGA-A8-A07G-01	TCGA-BH-A0HA-01
TCGA-LL-A9Q3-01	TCGA-AN-A0AJ-01
TCGA-GM-A2DK-01	TCGA-A2-A0YD-01
TCGA-AR-A1AX-01	TCGA-D8-A27E-01
TCGA-E9-A1N4-11	TCGA-B6-A0X7-01
TCGA-AO-A126-01	TCGA-AO-A129-01
TCGA-BH-A0DO-01	TCGA-A8-A08F-01

TCGA-AN-A0XO-01	TCGA-D8-A27W-01
TCGA-A2-A4RW-01	TCGA-A8-A09G-01
TCGA-E2-A1IH-01	TCGA-AC-A23E-01
TCGA-AO-A12B-01	TCGA-E9-A3X8-01
TCGA-A8-A081-01	TCGA-A2-A04V-01
TCGA-E9-A1NH-01	TCGA-E2-A10C-01
TCGA-EW-A2FV-01	TCGA-EW-A1IX-01
TCGA-D8-A27L-01	TCGA-E9-A22B-01
TCGA-AR-A0TZ-01	TCGA-GM-A3NW-01
TCGA-C8-A27B-01	TCGA-BH-A18N-11
TCGA-A8-A06R-01	TCGA-BH-A0W7-01
TCGA-BH-A0C1-01	TCGA-BH-A0DE-01
TCGA-B6-A0IB-01	TCGA-C8-A278-01
TCGA-D8-A1XD-01	TCGA-BH-A18G-01
TCGA-A7-A0DA-01	TCGA-E2-A574-01
TCGA-OL-A5D7-01	TCGA-D8-A73X-01
TCGA-BH-A0DP-11	TCGA-AR-A1AS-01
TCGA-C8-A26Y-01	TCGA-D8-A1XM-01
TCGA-A2-A4RY-01	TCGA-HN-A2OB-01
TCGA-E2-A107-01	TCGA-AC-A2FG-01
TCGA-EW-A1PC-01	TCGA-BH-A1FE-06
TCGA-AO-A12E-01	TCGA-AQ-A04H-01
TCGA-A2-A0T3-01	TCGA-E9-A249-01
TCGA-E9-A245-01	TCGA-B6-A0I2-01
TCGA-A8-A08X-01	TCGA-BH-A1ET-01
TCGA-A8-A07J-01	TCGA-A8-A08B-01
TCGA-A8-A094-01	TCGA-BH-A18S-11
TCGA-EW-A1P3-01	TCGA-A1-A0SJ-01
TCGA-EW-A6SB-01	TCGA-BH-A0DL-01
TCGA-BH-A209-11	TCGA-AR-A1AI-01
TCGA-C8-A1HL-01	TCGA-A2-A0D4-01
TCGA-B6-A1KN-01	TCGA-E9-A22E-01
TCGA-BH-A1FE-11	TCGA-E2-A1IN-01
TCGA-BH-A1FR-11	TCGA-D8-A1JH-01
TCGA-AQ-A04L-01	TCGA-EW-A1PE-01
TCGA-AC-A7VB-01	TCGA-BH-A0WA-01
TCGA-AR-A24Z-01	TCGA-AR-A1AT-01
TCGA-BH-A0HW-01	TCGA-A8-A091-01
TCGA-UU-A93S-01	TCGA-BH-A0W3-01
TCGA-A2-A04W-01	TCGA-D8-A73W-01
TCGA-BH-A0B7-01	TCGA-LL-A442-01
TCGA-BH-A0DI-01	TCGA-E2-A15G-01

TCGA-E2-A2P6-01	TCGA-EW-A1P8-01
TCGA-3C-AAAU-01	TCGA-BH-A1EU-11
TCGA-A8-A0A7-01	TCGA-AR-A1AM-01
TCGA-BH-A1FC-11	TCGA-BH-A204-11
TCGA-BH-A1ET-11	TCGA-A8-A07S-01
TCGA-EW-A1P7-01	TCGA-HN-A2NL-01
TCGA-BH-A0BV-01	TCGA-AO-A12G-01
TCGA-GM-A2DB-01	TCGA-BH-A0HK-01
TCGA-A1-A0SI-01	TCGA-E9-A1N6-01
TCGA-B6-A0I9-01	TCGA-A8-A07R-01
TCGA-E9-A1N5-11	TCGA-E9-A295-01
TCGA-A8-A092-01	TCGA-A8-A096-01
TCGA-EW-A1IW-01	TCGA-A8-A08O-01
TCGA-BH-A1F8-01	TCGA-BH-A42U-01
TCGA-E9-A1RG-01	TCGA-BH-A18M-11
TCGA-A2-A0EN-01	TCGA-AC-A2FO-01
TCGA-A7-A13F-11	TCGA-LL-A73Y-01
TCGA-BH-A0BQ-11	selected_samples
TCGA-BH-A1EX-01	TCGA-AC-A23C-01
TCGA-E2-A1B5-01	TCGA-BH-A18H-01
TCGA-A8-A07B-01	TCGA-EW-A1OY-01
TCGA-B6-A0RG-01	TCGA-A8-A09T-01
TCGA-BH-A6R9-01	TCGA-A8-A08T-01
TCGA-OL-A6VR-01	TCGA-BH-A0E2-01
TCGA-A7-A0CH-01	TCGA-E2-A1B0-01
TCGA-E2-A1BC-01	TCGA-E2-A14S-01
TCGA-A7-A0CH-11	TCGA-BH-A0C3-11
TCGA-AR-A24M-01	TCGA-A2-A0EX-01
TCGA-5L-AAT0-01	TCGA-B6-A0WW-01
TCGA-A2-A0CZ-01	TCGA-E2-A570-01
TCGA-BH-A1FD-11	TCGA-C8-A137-01
TCGA-A8-A08P-01	TCGA-BH-A0H0-01
TCGA-AC-A8OR-01	TCGA-BH-A0C0-01
TCGA-B6-A3ZX-01	TCGA-E2-A1LS-11
TCGA-BH-A0BL-01	TCGA-AR-A252-01
TCGA-E9-A22G-01	TCGA-3C-AALK-01
TCGA-A2-A0T7-01	TCGA-A7-A4SE-01
TCGA-AC-A6NO-01	TCGA-LD-A7W5-01
TCGA-C8-A3M7-01	TCGA-AR-A1AJ-01
TCGA-A7-A2KD-01	TCGA-BH-A0H5-01
TCGA-OL-A66O-01	TCGA-A8-A07E-01
TCGA-C8-A134-01	TCGA-BH-A18F-01

TCGA-C8-A12Z-01
TCGA-C8-A1HK-01
TCGA-AO-A0J3-01
TCGA-EW-A3E8-01
TCGA-LL-A73Z-01
TCGA-A8-A09Q-01
TCGA-A1-A0SK-01
TCGA-LL-A6FR-01
TCGA-EW-A1OW-01
TCGA-D8-A1X9-01
TCGA-E2-A576-01
TCGA-BH-A0BP-01
TCGA-E2-A15K-11
TCGA-EW-A6SA-01
TCGA-BH-A0E0-01
TCGA-BH-A0HN-01
TCGA-BH-A0DG-01
TCGA-AN-A0XN-01
TCGA-AR-A24H-01
TCGA-A8-A06O-01
TCGA-C8-A1HM-01
TCGA-A2-A4RX-01
TCGA-AC-A6IV-01
TCGA-BH-A0HF-01
TCGA-AN-A0XP-01
TCGA-GM-A3XL-01
TCGA-BH-A28Q-01
TCGA-A2-A04T-01
TCGA-AN-A04D-01
TCGA-BH-A0B8-01
TCGA-BH-A0EB-01
TCGA-AN-A0FL-01
TCGA-AR-A2LK-01
TCGA-A8-A09Z-01
TCGA-EW-A3U0-01
TCGA-A8-A07P-01
TCGA-AN-A0FF-01
TCGA-A8-A08J-01
TCGA-BH-A0GZ-01
TCGA-AR-A254-01
TCGA-A8-A08L-01

ANNEXE D

MDH1	CRYL1	PGD	UBB
GOT2	DCXR	PGLS	UBC
SLC2A3	EPM2A	PGM1	UGP2
PFKM	G6PC3	PGM2	XYLB
PFKL	G6PD	PHKA1	
PGAM2	GAA	PHKA2	
GCK	GALE	PHKB	
SLC2A1	GALK1	PHKG1	
MDH2	GALT	PHKG2	
HK2	GBE1	POM121	
ENO3	GYG1	PPP2CA	
DLD	GYG2	PPP2CB	
GPI	GYS1	PPP2R1A	
PGAM1	KHK	PPP2R1B	
PC	MGAM	PPP2R5D	
PDHA1	PPP1R3C	PRKACA	
FBP1	NHLRC1	PRKACB	
GAPDH	NUP107	PRPS1	
ENO2	NUP133	PRPS2	
HK3	NUP153	PYGB	
ALDOA	NUP155	PYGL	
HK1	NUP160	PYGM	
DLAT	NUP188	RAE1	
LDHB	NUP205	RANBP2	
SLC2A4	NUP210	RPE	
PDHB	NUP214	RPIA	
PGK1	NUP35	RPS27A	
LDHA	NUP37	SLC25A10	
PDHX	NUP43	SLC25A11	
GOT1	NUP50	SLC25A12	
PFKP	NUP54	SLC25A13	
ENO1	NUP62	SLC25A1	
ALDOC	NUP85	SLC37A4	
SLC2A5	NUP88	SLC5A1	
AAAS	NUP93	SLC5A3	
ADPGK	NUP98	SLC5A9	
AGL	NUPL2	SORD	
AKR1A1	PCK2	TALDO1	
AKR1B1	PFKFB1	TKT	
AMY2B	PFKFB2	TPI1	
B4GALT1	PFKFB3	TPR	
CALM1	PFKFB4	UBA52	

ANNEXE E

ATP5B
ATP5C1
ATP5D
ATP5E
ATP5F1
ATP5G1
ATP5G2
ATP5G3
ATP5H
ATP5I
ATP5J
ATP5J2
ATP5L
ATP5O
ATP5S
ATPIF1
COX11
COX15
COX17
COX4I1
COX5A
COX5B
COX6A1
COX6B1
COX6C
COX7A1
COX7A2
COX7A2L
COX7B
COX7C
COX8A
NDUFA1
NDUFA10
NDUFA2
NDUFA3
NDUFA4
NDUFA5
NDUFA6
NDUFA7
NDUFA8
NDUFA9
NDUFAB1
NDUFB1
NDUFB10
NDUFB2
NDUFB3
NDUFB4
NDUFB5
NDUFB6
NDUFB7
NDUFB8
NDUFB9
NDUFC1
NDUFC2
NDUFS1
NDUFS2
NDUFS3
NDUFS4
NDUFS5
NDUFS6
NDUFS7
NDUFS8
NDUFV1
NDUFV2
NDUFV3
SCO1
SDHA
SDHB
SDHC
SDHD
SLC25A14
SLC25A27
SLC25A4
SLC25A5
SLC25A6
SURF1
UCP2
UCP3
UQCRB
UQCRC1
UQCRC2
UQCRFS1
ATP5A1
ATP6AP1
ATP6AP2
GZMB
NDUFA11
UQCRH

ANNEXE F

Disease A	Disease B	Overlap %
developmental disorder of mental health	specific developmental disorder	86.1
b cell deficiency	blood protein disease	82.8
aagenaes syndrome	bolivian hemorrhagic fever	55.6
multiple endocrine neoplasia type 2a	multiple endocrine neoplasia type 2b	55.3
carbohydrate metabolism disease	nutrition disease	50.4
heart disease	muscle tissue disease	47.9
female reproductive organ cancer	intestinal cancer	45.7
dermatitis	psoriasis	45.6
blau syndrome	muckle-wells syndrome	45.0
intestinal cancer	lung cancer	44.7
carbohydrate metabolism disease	kidney disease	44.2
leiomyoma	reproductive organ benign neoplasm	44.1
hypersensitivity reaction type i disease	upper respiratory tract disease	44.1
neurodegenerative disease	vascular disease	43.7
breast cancer	hematologic cancer	43.6
infertility	male reproductive system disease	42.7
breast cancer	female reproductive organ cancer	42.7
dermatitis	hypersensitivity reaction type i disease	42.6
hypersensitivity reaction type ii disease	systemic lupus erythematosus	42.6
melanoma	skin cancer	42.4
breast cancer	lung cancer	42.4
hypersensitivity reaction type ii disease	lower respiratory tract disease	42.3
female reproductive organ cancer	lung cancer	42.2
epidermolytic hyperkeratosis	pachyonychia congenita	41.9
liver cancer	lung cancer	41.8
murray valley encephalitis	st. louis encephalitis	41.7
kidney disease	lower respiratory tract disease	41.6
hepatitis b	hepatitis c	41.3
hypersensitivity reaction type ii disease	kidney disease	41.0
breast cancer	intestinal cancer	40.6
breast cancer	liver cancer	40.6
bone disease	hypersensitivity reaction type ii disease	40.4
mood disorder	psychotic disorder	40.3
female reproductive organ cancer	liver cancer	40.2
intestinal cancer	liver cancer	40.0
meier-gorlin syndrome	ogden syndrome	40.0
monilethrix	pachyonychia congenita	40.0
cockayne syndrome	xeroderma pigmentosum	39.8
congenital ichthyosiform erythroderma	ichthyosis vulgaris	39.7

hypersensitivity reaction type i disease	leukocyte disease	39.7
blood coagulation disease	leukocyte disease	39.7
bone disease	lower respiratory tract disease	39.5
hypersensitivity reaction type ii disease	intestinal disease	39.5
hepatobiliary disease	kidney disease	38.8
connective tissue cancer	sarcoma	38.8
intestinal cancer	male reproductive organ cancer	38.8
hematologic cancer	lung cancer	38.7
intestinal disease	lower respiratory tract disease	38.5
hepatobiliary disease	hypersensitivity reaction type ii disease	38.4
kidney disease	vascular disease	38.3
bone disease	vascular disease	38.3
female reproductive organ cancer	male reproductive organ cancer	38.1
borjeson-forssman-lehmann syndrome	coffin-siris syndrome	38.1
ichthyosis vulgaris	x-linked ichthyosis	38.1
klippel-trenaunay syndrome	proteus syndrome	38.1
hepatobiliary disease	primary bacterial infectious disease	37.8
kidney disease	nutrition disease	37.6
brain disease	specific developmental disorder	37.6
angioedema	complement deficiency	37.4
heart disease	kidney disease	37.4
hematologic cancer	liver cancer	37.3
bone disease	hematologic cancer	37.2
hematologic cancer	vascular disease	37.2
breast cancer	vascular disease	36.8
carbohydrate metabolism disease	lower respiratory tract disease	36.8
hepatobiliary disease	intestinal disease	36.7
hereditary mucosal leukokeratosis	pachyonychia congenita	36.7
short qt syndrome	timothy syndrome	36.6
carbohydrate metabolism disease	vascular disease	36.5
lung cancer	male reproductive organ cancer	36.4
bone disease	kidney disease	36.4
breast cancer	male reproductive organ cancer	36.3
female reproductive organ cancer	melanoma	36.3
dermatitis	upper respiratory tract disease	36.0
carbohydrate metabolism disease	heart disease	36.0
brain disease	neurodegenerative disease	36.0
brain disease	developmental disorder of mental health	35.8
klippel-trenaunay syndrome	multiple cutaneous and mucosal venous malformations	35.7

lower respiratory tract disease	vascular disease	35.7
intestinal disease	primary bacterial infectious disease	35.5
intestinal cancer	melanoma	35.4
diarrhea	leukocyte disease	35.3
hematologic cancer	intestinal cancer	35.3
female reproductive organ cancer	hematologic cancer	35.3
leukocyte disease	upper respiratory tract disease	35.3
bone disease	eye and adnexa disease	35.3
lower respiratory tract disease	primary bacterial infectious disease	35.2
intestinal disease	kidney disease	35.2
carbohydrate metabolism disease	pancreas disease	35.2
intestinal cancer	pancreatic cancer	35.0
epidermolytic hyperkeratosis	monilethrix	34.9
hepatobiliary disease	lower respiratory tract disease	34.8
heart disease	lower respiratory tract disease	34.8
female reproductive system disease	infertility	34.5
carbohydrate metabolism disease	hypersensitivity reaction type ii disease	34.5
epidermolytic hyperkeratosis	ichthyosis vulgaris	34.5
eye and adnexa disease	vascular disease	34.4
eye and adnexa disease	kidney disease	34.4
spinal cord disease	tropical spastic paraparesis	34.4
malignant glioma	pancreatic cancer	34.3
carbohydrate metabolism disease	hepatobiliary disease	34.2
bardet-biedl syndrome	nephronophthisis	34.2
multiple cutaneous and mucosal venous malformations	proteus syndrome	34.2
dermatitis	leukocyte disease	34.2
sickle cell anemia	thalassemia	34.1
lung cancer	melanoma	34.0
parasitic protozoa infectious disease	primary bacterial infectious disease	33.8
dementia	toxic encephalopathy	33.8
male reproductive organ cancer	melanoma	33.6
bone disease	carbohydrate metabolism disease	33.5
bannayan-riley-ruvalcaba syndrome	cowden disease	33.3
chickenpox	herpes zoster	33.3
cardiofaciocutaneous syndrome	costello syndrome	33.3
malignant glioma	peripheral nervous system neoplasm	33.3
angelman syndrome	prader-willi syndrome	33.3
hypersensitivity reaction type ii disease	primary bacterial infectious disease	33.3
dermatitis	mouth disease	33.2

eye and adnexa disease	lower respiratory tract disease	33.0
hematologic cancer	hypersensitivity reaction type ii disease	32.9
nutrition disease	vascular disease	32.9
adenoma	pancreatic cancer	32.8
human immunodeficiency virus infectious disease	leukocyte disease	32.8
liver cancer	male reproductive organ cancer	32.8
ichthyosis vulgaris	netherton syndrome	32.8
heart disease	vascular disease	32.7
bone disease	intestinal disease	32.7
liver cancer	melanoma	32.7
epidermolytic hyperkeratosis	hereditary mucosal leukokeratosis	32.6
male reproductive organ cancer	pancreatic cancer	32.5
human immunodeficiency virus infectious disease	psoriasis	32.4
brain disease	eye and adnexa disease	32.4
human immunodeficiency virus infectious disease	systemic lupus erythematosus	32.3
hematologic cancer	neurodegenerative disease	32.3
hermansky-pudlak syndrome	oculocutaneous albinism	32.3
mumps	rubella	32.3
heart disease	nutrition disease	32.2
trichorhinophalangeal syndrome type i	trichorhinophalangeal syndrome type ii	32.1
hematologic cancer	lower respiratory tract disease	32.1
dermatitis	human immunodeficiency virus infectious disease	32.1
bardet-biedl syndrome	ciliopathy	32.1
muscle cancer	sarcoma	32.0
cardiofaciocutaneous syndrome	leopard syndrome	32.0
female reproductive organ cancer	pancreatic cancer	32.0
kidney disease	primary bacterial infectious disease	31.9
bone disease	neurodegenerative disease	31.8
heart disease	hepatobiliary disease	31.8
branchiootic syndrome	branchiootorenal syndrome	31.7
anemia	blood coagulation disease	31.7
hypersensitivity reaction type i disease	psoriasis	31.6
bannayan-riley-ruvalcaba syndrome	proteus syndrome	31.6
developmental disorder of mental health	pervasive developmental disorder	31.6
hereditary mucosal leukokeratosis	monilethrix	31.4
mesenchymal cell neoplasm	mixed cell type cancer	31.4

asphyxiating thoracic dystrophy	ellis-van creveld syndrome	31.3
anemia	parasitic protozoa infectious disease	31.2
hematologic cancer	melanoma	31.2
breast cancer	neurodegenerative disease	31.2
anemia	hepatobiliary disease	31.2
factor xi deficiency	factor xii deficiency	31.1
mouth disease	psoriasis	31.1
breast cancer	melanoma	31.0
anemia	leukocyte disease	30.9
adrenal gland disease	endocrine organ benign neoplasm	30.9
edwards syndrome	patau syndrome	30.9
head and neck carcinoma	oral cavity cancer	30.9
lipid metabolism disorder	lysosomal storage disease	30.9
hypersensitivity reaction type iv disease	lymph node disease	30.9
kidney disease	muscle tissue disease	30.8
brain disease	vascular disease	30.7
lung cancer	pancreatic cancer	30.7
hypersensitivity reaction type ii disease	melanoma	30.7
bone disease	breast cancer	30.6
esophageal cancer	head and neck carcinoma	30.5
melanoma	pancreatic cancer	30.5
brain disease	carbohydrate metabolism disease	30.5
lower respiratory tract disease	melanoma	30.5
collagen disease	systemic lupus erythematosus	30.5
lipid metabolism disorder	pancreas disease	30.5
hematologic cancer	male reproductive organ cancer	30.4
hepatobiliary disease	nutrition disease	30.3
carbohydrate metabolism disease	intestinal disease	30.3
eye and adnexa disease	neurodegenerative disease	30.2
kidney disease	male reproductive organ cancer	30.1
psoriasis	upper respiratory tract disease	30.1

RÉFÉRENCES

Agarwal, S., Deane, C.M., Porter, M.A. et Jones, N.S. (2010). Revisiting date and party hubs: novel approaches to role assignment in protein interaction networks. *PLoS Comput Biol, 6*(6), e1000817. doi: 10.1371/journal.pcbi.1000817

Andreani, J. et Guerois, R. (2014). Evolution of protein interactions: From interactomes to interfaces. *Archives of Biochemistry and Biophysics, 554*, 65-75. doi: http://dx.doi.org/10.1016/j.abb.2014.05.010

Bandyopadhyay, S., Kelley, R., Krogan, N.J. et Ideker, T. (2008). Functional Maps of Protein Complexes from Quantitative Genetic Interaction Data. *PLoS Comput Biol, 4*(4), e1000065. doi: 10.1371/journal.pcbi.1000065

Barabasi, A.L. et Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*(5439), 509-512.

Barabàsi, A.L. et Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. *Nat Rev Genet, 5*(2), 101-113. doi: 10.1038/nrg1272

Baryshnikova, A., Costanzo, M., Dixon, S., Vizeacoumar, F.J., Myers, C.L., Andrews, B. et Boone, C. (2010). Synthetic genetic array (SGA) analysis in Saccharomyces cerevisiae and Schizosaccharomyces pombe. *Methods Enzymol, 470*, 145-179. doi: 10.1016/S0076-6879(10)70007-0

Baryshnikova, A., Costanzo, M., Myers, C.L., Andrews, B. et Boone, C. (2013). Genetic interaction networks: toward an understanding of heritability. *Annu Rev Genomics Hum Genet, 14*, 111-133. doi: 10.1146/annurev-genom-082509-141730

Bateson, W., Saunders, E., Punnett, R. et Hurst, C. (1905). Reports to the Evolution Committee of the Royal Society, report II. *London, UK: Harrison and Sons*.

Batista, L.V. et Meira, M.M. (2004) Texture classification using the Lempel-Ziv-Welch algorithm. *: Vol. 3171. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (pp. 444-453).

Bellay, J., Atluri, G., Sing, T.L., Toufighi, K., Costanzo, M., Ribeiro, P.S., Pandey, G., Baller, J., VanderSluis, B., Michaut, M., Han, S., Kim, P., Brown, G.W., Andrews, B.J., Boone, C., Kumar, V. et Myers, C.L. (2011). Putting genetic interactions in context through a global modular decomposition. *Genome Res, 21*(8), 1375-1387. doi: 10.1101/gr.117176.110

Boucher, B. et Jenna, S. (2013). Genetic interaction networks: better understand to better predict. *Front Genet, 4*, 290. doi: 10.3389/fgene.2013.00290

Boucher, B., Lee, A.Y., Hallett, M. et Jenna, S. (2016). Structural and Functional Characterization of a Caenorhabditis elegans Genetic Interaction Network within Pathways. *PLoS Computational Biology, 12*(2), e1004738. doi: 10.1371/journal.pcbi.1004738

Butland, G., Peregrin-Alvarez, J.M., Li, J., Yang, W., Yang, X., Canadien, V., Starostine, A., Richards, D., Beattie, B., Krogan, N., Davey, M., Parkinson, J., Greenblatt, J. et Emili, A. (2005). Interaction network containing conserved and essential protein complexes in Escherichia coli. *Nature, 433*(7025), 531-537.

Byrne, A.B., Weirauch, M.T., Wong, V., Koeva, M., Dixon, S.J., Stuart, J.M. et Roy, P.J. (2007). A global analysis of genetic interactions in Caenorhabditis elegans. *Journal of Biology, 6*(3), 8-8. doi: 10.1186/jbiol58

Cano, J.R., Herrera, F. et Lozano, M. (2007). Evolutionary stratified training set selection for extracting classification rules with trade off precision-interpretability. [Article]. *Data and Knowledge Engineering, 60*(1), 90-108. doi: 10.1016/j.datak.2006.01.008

Chang, X., Xu, T., Li, Y. et Wang, K. (2013). Dynamic modular architecture of protein-protein interaction networks beyond the dichotomy of 'date' and 'party' hubs. [Research Support, N.I.H., Extramural
Research Support, Non-U.S. Gov't]. *Sci Rep, 3*, 1691. doi: 10.1038/srep01691

Chari, S. et Dworkin, I. (2013). The Conditional Nature of Genetic Interactions: The Consequences of Wild-Type Backgrounds on Mutational Interactions in a Genome-Wide Modifier Screen. *PLoS Genetics, 9*(8), e1003661. doi: 10.1371/journal.pgen.1003661

Chatr-Aryamontri, A., Breitkreutz, B.J., Oughtred, R., Boucher, L., Heinicke, S., Chen, D., Stark, C., Breitkreutz, A., Kolas, N., O'Donnell, L., Reguly, T., Nixon, J., Ramage, L., Winter, A., Sellam, A., Chang, C., Hirschman, J., Theesfeld, C., Rust, J., Livstone, M.S., Dolinski, K. et Tyers, M. (2015). The BioGRID interaction database: 2015 update. *Nucleic Acids Res, 43*(Database issue), D470-478. doi: 10.1093/nar/gku1204

Chen, Y., Hao, J., Jiang, W., He, T., Zhang, X., Jiang, T. et Jiang, R. (2013). Identifying potential cancer driver genes by genomic data integration. *Scientific Reports, 3*, 3538. doi: 10.1038/srep03538

Chimge, N.O. et Frenkel, B. (2013). The RUNX family in breast cancer: relationships with estrogen signaling. *Oncogene, 32*(17), 2121-2130. doi: 10.1038/onc.2012.328

Chipman, K.C. et Singh, A.K. (2009). Predicting genetic interactions with random walks on biological networks. *BMC Bioinformatics, 10*, 17. doi: 10.1186/1471-2105-10-17

Chowdhary, R., Zhang, J. et Liu, J.S. (2009). Bayesian inference of protein–protein interactions from biological literature. *Bioinformatics, 25*(12), 1536-1542. doi: 10.1093/bioinformatics/btp245

Christofk, H.R., Vander Heiden, M.G., Harris, M.H., Ramanathan, A., Gerszten, R.E., Wei, R., Fleming, M.D., Schreiber, S.L. et Cantley, L.C. (2008). The M2 splice isoform of pyruvate kinase is important for cancer metabolism and tumour growth. *Nature, 452*(7184), 230-233. doi: 10.1038/nature06734

Collavin, L., Lunardi, A. et Del Sal, G. (2010). p53-family proteins and their regulators: hubs and spokes in tumor suppression. *Cell Death Differ, 17*(6), 901-911.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E.D., Sevier, C.S., Ding, H., Koh, J.L., Toufighi, K., Mostafavi, S., Prinz, J., St Onge, R.P., VanderSluis, B., Makhnevych, T., Vizeacoumar, F.J., Alizadeh, S., Bahr, S., Brost, R.L., Chen, Y., Cokol, M., Deshpande, R., Li, Z., Lin, Z.Y., Liang, W., Marback, M., Paw, J., San Luis, B.J., Shuteriqi, E., Tong, A.H., van Dyk, N., Wallace, I.M., Whitney, J.A., Weirauch, M.T., Zhong, G., Zhu, H., Houry, W.A., Brudno, M., Ragibizadeh, S., Papp, B., Pal, C., Roth, F.P., Giaever, G., Nislow, C., Troyanskaya, O.G., Bussey, H., Bader, G.D., Gingras, A.C., Morris, Q.D., Kim, P.M., Kaiser, C.A., Myers, C.L., Andrews, B.J. et Boone, C. (2010). The genetic landscape of a cell. *Science, 327*(5964), 425-431. doi: 10.1126/science.1180823

Costanzo, M., Baryshnikova, A., Myers, C.L., Andrews, B. et Boone, C. (2011). Charting the genetic interaction map of a cell. [Review]. *Curr Opin Biotechnol, 22*(1), 66-74. doi: 10.1016/j.copbio.2010.11.001

Dai, X., Li, T., Bai, Z., Yang, Y., Liu, X., Zhan, J. et Shi, B. (2015). Breast cancer intrinsic subtype classification, clinical use and future trends. *American Journal of Cancer Research, 5*(10), 2929-2943.

(dir.). (2006). *The relationship between Precision-Recall and ROC curves.* ICML 2006 - Proceedings of the 23rd International Conference on Machine Learning.

De Carvalho, D.D., Sharma, S., You, J.S., Su, S.-F., Taberlay, P.C., Kelly, T.K., Yang, X., Liang, G. et Jones, P.A. (2012). DNA methylation screening identifies driver

epigenetic events of cancer cell survival. *Cancer Cell, 21*(5), 655-667. doi: 10.1016/j.ccr.2012.03.045

Di Meo, A., Diamandis, E.P., Rodriguez, H., Hoofnagle, A.N., Ioannidis, J. et Lopez, M. (2014). What Is Wrong with Clinical Proteomics? [10.1373/clinchem.2014.225185]. *Clinical Chemistry, 60*(10), 1258.

Dixon, S.J., Andrews, B.J. et Boone, C. (2009). Exploring the conservation of synthetic lethal genetic interaction networks. *Commun Integr Biol, 2*(2), 78-81.

Dixon, S.J., Costanzo, M., Baryshnikova, A., Andrews, B. et Boone, C. (2009). Systematic mapping of genetic interaction networks. *Annu Rev Genet, 43*, 601-625. doi: 10.1146/annurev.genet.39.073003.114751

Dutta, B., Pusztai, L., Qi, Y., Andre, F., Lazar, V., Bianchini, G., Ueno, N., Agarwal, R., Wang, B., Shiang, C.Y., Hortobagyi, G.N., Mills, G.B., Symmans, W.F. et Balazsi, G. (2012). A network-based, integrative study to identify core biological pathways that drive breast cancer clinical subtypes. *Br J Cancer, 106*(6), 1107-1116. doi: 10.1038/bjc.2011.584

Elf, S.E. et Chen, J. (2014). Targeting glucose metabolism in patients with cancer. *Cancer, 120*(6), 774-780. doi: 10.1002/cncr.28501

Fawcett, T. (2006). An introduction to ROC analysis. [Article]. *Pattern Recognition Letters, 27*(8), 861-874. doi: 10.1016/j.patrec.2005.10.010

Fiedler, D., Braberg, H., Mehta, M., Chechik, G., Cagney, G., Mukherjee, P., Silva, A.C., Shales, M., Collins, S.R., van Wageningen, S., Kemmeren, P., Holstege, F.C.P., Weissman, J.S., Keogh, M.-C., Koller, D., Shokat, K.M. et Krogan, N.J. (2009). Functional Organization of the S. cerevisiae Phosphorylation Network. *Cell, 136*(5), 952-963. doi: 10.1016/j.cell.2008.12.039

Fox, R.F. (1993). Review of Stuart Kauffman, The Origins of Order: Self-Organization and Selection in Evolution. *Biophysical Journal, 65*(6), 2698-2699.

Fu, Y., Guo, Y., Wang, Y., Luo, J., Pu, X., Li, M. et Zhang, Z. (2015). Exploring the relationship between hub proteins and drug targets based on GO and intrinsic disorder. *Comput Biol Chem, 56*, 41-48. doi: 10.1016/j.compbiolchem.2015.03.003

Gallati, S. (2014). Disease-modifying genes and monogenic disorders: experience in cystic fibrosis. *The Application of Clinical Genetics, 7*, 133-146. doi: 10.2147/TACG.S18675

Gatza, M.L., Silva, G.O., Parker, J.S., Fan, C. et Perou, C.M. (2014). An integrated genomics approach identifies drivers of proliferation in luminal subtype human breast cancer. *Nature genetics, 46*(10), 1051-1059. doi: 10.1038/ng.3073

Gavande, N.S., VanderVere-Carozza, P.S., Hinshaw, H.D., Jalal, S.I., Sears, C.R., Pawelczak, K.S. et Turchi, J.J. (2016). DNA repair targeted therapy: The past or future of cancer treatment? *Pharmacol Ther, 160*, 65-83. doi: 10.1016/j.pharmthera.2016.02.003

Gillis, J. et Pavlidis, P. (2012). "Guilt by association" is the exception rather than the rule in gene networks. *PLoS Computational Biology, 8*(3). doi: 10.1371/journal.pcbi.1002444

Goymer, P. (2008). Network biology: Why do we need hubs? [10.1038/nrg2450]. *Nat Rev Genet, 9*(9), 650-651.

Han, J.D., Bertin, N., Hao, T., Goldberg, D.S., Berriz, G.F., Zhang, L.V., Dupuy, D., Walhout, A.J., 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. doi: 10.1038/nature02555

Hoehndorf, R., Hardy, N.W., Osumi-Sutherland, D., Tweedie, S., Schofield, P.N. et Gkoutos, G.V. (2013). Systematic analysis of experimental phenotype data reveals gene functions. *PLoS ONE, 8*(4), e60847. doi: 10.1371/journal.pone.0060847

Howe, K.L., Bolt, B.J., Cain, S., Chan, J., Chen, W.J., Davis, P., Done, J., Down, T., Gao, S., Grove, C., Harris, T.W., Kishore, R., Lee, R., Lomax, J., Li, Y., Muller, H.-M., Nakamura, C., Nuin, P., Paulini, M., Raciti, D., Schindelman, G., Stanley, E., Tuli, M.A., Van Auken, K., Wang, D., Wang, X., Williams, G., Wright, A., Yook, K., Berriman, M., Kersey, P., Schedl, T., Stein, L. et Sternberg, P.W. (2016). WormBase 2016: expanding to enable helminth genomic research. *Nucleic Acids Research, 44*(Database issue), D774-D780. doi: 10.1093/nar/gkv1217

Hsu, C.H., Wang, T.Y., Chu, H.T., Kao, C.Y. et Chen, K.C. (2011). A quantitative analysis of monochromaticity in genetic interaction networks. *BMC Bioinformatics, 12 Suppl 13*, S16. doi: 10.1186/1471-2105-12-s13-s16

Ignatiadis, M., Desmedt, C., Sotiriou, C., de Azambuja, E. et Piccart, M. (2009). HER-2 as a target for breast cancer therapy. *Clin Cancer Res, 15*(6), 1848-1852. doi: 10.1158/1078-0432.CCR-08-1844

Ivanov, A.A., Khuri, F.R. et Fu, H. (2013). Targeting protein-protein interactions as an anticancer strategy. *Trends in pharmacological sciences, 34*(7), 393-400. doi: 10.1016/j.tips.2013.04.007

Jaramillo-Lambert, A., Fuchsman, A.S., Fabritius, A.S., Smith, H.E. et Golden, A. (2015). Rapid and Efficient Identification of Caenorhabditis elegans Legacy Mutations Using Hawaiian SNP-Based Mapping and Whole-Genome Sequencing. *G3: Genes|Genomes|Genetics, 5*(5), 1007-1019. doi: 10.1534/g3.115.017038

Jeong, H., Mason, S.P., Barabasi, A.L. et Oltvai, Z.N. (2001). Lethality and centrality in protein networks. [10.1038/35075138]. *Nature, 411*(6833), 41-42.

Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. et Barabasi, A.L. (2000). The large-scale organization of metabolic networks. *Nature, 407*(6804), 651-654. doi: 10.1038/35036627

Jerby-Arnon, L., Pfetzer, N., Waldman, Y.Y., McGarry, L., James, D., Shanks, E., Seashore-Ludlow, B., Weinstock, A., Geiger, T., Clemons, P.A., Gottlieb, E. et Ruppin, E. (2014). Predicting cancer-specific vulnerability via data-driven detection of synthetic lethality. *Cell, 158*(5), 1199-1209. doi: 10.1016/j.cell.2014.07.027

Jonsson, P.F. et Bates, P.A. (2006). Global topological features of cancer proteins in the human interactome. *Bioinformatics (Oxford, England), 22*(18), 2291-2297. doi: 10.1093/bioinformatics/btl390

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. doi: 10.1155/JBB.2005.96

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. et Hattori, M. (2004). The KEGG resource for deciphering the genome. *Nucleic Acids Res, 32*(Database issue), D277-280. doi: 10.1093/nar/gkh063

Kelley, R. et Ideker, T. (2005). Systematic interpretation of genetic interactions using protein networks. *Nat Biotechnol, 23*(5), 561-566. doi: 10.1038/nbt1096

Koch, E.N., Costanzo, M., Bellay, J., Deshpande, R., Chatfield-Reed, K., Chua, G., D'Urso, G., Andrews, B.J., Boone, C. et Myers, C.L. (2012). Conserved rules govern genetic interaction degree across species. *Genome Biol, 13*(7), R57. doi: 10.1186/gb-2012-13-7-r57

Korcsmaros, T., Farkas, I.J., Szalay, M.S., Rovo, P., Fazekas, D., Spiro, Z., Bode, C., Lenti, K., Vellai, T. et Csermely, P. (2010). Uniformly curated signaling pathways reveal tissue-specific cross-talks and support drug target discovery. *Bioinformatics, 26*(16), 2042-2050. doi: 10.1093/bioinformatics/btq310

Kristensen, V.N., Lingjaerde, O.C., Russnes, H.G., Vollan, H.K.M., Frigessi, A. et Borresen-Dale, A.-L. (2014). Principles and methods of integrative genomic analyses in cancer. [Review]. *Nat Rev Cancer, 14*(5), 299-313. doi: 10.1038/nrc3721

Lee, A.Y., Perreault, R., Harel, S., Boulier, E.L., Suderman, M., Hallett, M. et Jenna, S. (2010). Searching for signaling balance through the identification of genetic interactors of the Rab guanine-nucleotide dissociation inhibitor gdi-1. *PLoS ONE, 5*(5).

Lee, I., Lehner, B., Vavouri, T., Shin, J., Fraser, A.G. et Marcotte, E.M. (2010). Predicting genetic modifier loci using functional gene networks. *Genome Res, 20*(8), 1143-1153. doi: 10.1101/gr.102749.109

Lee, I., Li, Z. et Marcotte, E.M. (2007). An Improved, Bias-Reduced Probabilistic Functional Gene Network of Baker's Yeast, Saccharomyces cerevisiae. *PLoS ONE, 2*(10), e988. doi: 10.1371/journal.pone.0000988

Lee, J., Kim, H.K., Han, Y.M. et Kim, J. (2008). Pyruvate kinase isozyme type M2 (PKM2) interacts and cooperates with Oct-4 in regulating transcription. *Int J Biochem Cell Biol, 40*(5), 1043-1054. doi: 10.1016/j.biocel.2007.11.009

Lehner, B. (2007). Modelling genotype-phenotype relationships and human disease with genetic interaction networks. *J Exp Biol, 210*(Pt 9), 1559-1566. doi: 10.1242/jeb.002311

Lehner, B., Crombie, C., Tischler, J., Fortunato, A. et Fraser, A.G. (2006). Systematic mapping of genetic interactions in Caenorhabditis elegans identifies common modifiers of diverse signaling pathways. *Nat Genet, 38*(8), 896-903. doi: 10.1038/ng1844

Lewis, A.C.F., Jones, N.S., Porter, M.A. et Deane, C.M. (2012). What Evidence Is There for the Homology of Protein-Protein Interactions? *PLoS Computational Biology, 8*(9), e1002645. doi: 10.1371/journal.pcbi.1002645

Locard-Paulet, M., Lim, L., Veluscek, G., McMahon, K., Sinclair, J., van Weverwijk, A., Worboys, J.D., Yuan, Y., Isacke, C.M. et Jørgensen, C. (2016). Phosphoproteomic analysis of interacting tumor and endothelial cells identifies regulatory mechanisms of transendothelial migration. [10.1126/scisignal.aac5820]. *Science Signaling, 9*(414), ra15.

Lu, X., Megchelenbrink, W., Notebaart, R.A. et Huynen, M.A. (2015). Predicting Human Genetic Interactions from Cancer Genome Evolution. *PLoS ONE, 10*(5), e0125795. doi: 10.1371/journal.pone.0125795

Luo, W., Hu, H., Chang, R., Zhong, J., Knabel, M., O'Meally, R., Cole, R.N., Pandey, A. et Semenza, G.L. (2011). Pyruvate Kinase M2 is a PHD3-stimulated Coactivator for Hypoxia-Inducible Factor 1. *Cell, 145*(5), 732-744. doi: 10.1016/j.cell.2011.03.054

Lv, W., Xu, Y., Guo, Y., Yu, Z., Feng, G., Liu, P., Luan, M., Zhu, H., Liu, G., Zhang, M., Lv, H., Duan, L., Shang, Z., Li, J., Jiang, Y. et Zhang, R. (2016). The drug target genes show higher evolutionary conservation than non-target genes. *Oncotarget, 7*(4), 4961-4971. doi: D - NLM: PMC4826257 OTO - NOTNLM

Madhukar, N.S., Elemento, O. et Pandey, G. (2015). Prediction of Genetic Interactions Using Machine Learning and Network Properties. [Review]. *Frontiers in Bioengineering and Biotechnology, 3*(172). doi: 10.3389/fbioe.2015.00172

Mertins, P., Mani, D.R., Ruggles, K.V., Gillette, M.A., Clauser, K.R., Wang, P., Wang, X., Qiao, J.W., Cao, S., Petralia, F., Kawaler, E., Mundt, F., Krug, K., Tu, Z., Lei, J.T., Gatza, M.L., Wilkerson, M., Perou, C.M., Yellapantula, V., Huang, K.-l., Lin, C., McLellan, M.D., Yan, P., Davies, S.R., Townsend, R.R., Skates, S.J., Wang, J., Zhang, B., Kinsinger, C.R., Mesri, M., Rodriguez, H., Ding, L., Paulovich, A.G., Fenyö, D., Ellis, M.J., Carr, S.A. et Nci, C. (2016). Proteogenomics connects somatic mutations to signalling in breast cancer. *Nature, 534*(7605), 55-62. doi: 10.1038/nature18003

Michaut, M., Baryshnikova, A., Costanzo, M., Myers, C.L., Andrews, B.J., Boone, C. et Bader, G.D. (2011). Protein complexes are central in the yeast genetic landscape. *PLoS Comput Biol, 7*(2), e1001092. doi: 10.1371/journal.pcbi.1001092

Mirzarezaee, M., Araabi, B.N. et Sadeghi, M. (2010). Features analysis for identification of date and party hubs in protein interaction network of Saccharomyces Cerevisiae. *BMC Syst Biol, 4*, 172. doi: 10.1186/1752-0509-4-172

Mitra, K., Carvunis, A.-R., Ramesh, S.K. et Ideker, T. (2013). Integrative approaches for finding modular structure in biological networks. *Nature reviews. Genetics, 14*(10), 719-732. doi: 10.1038/nrg3552

Myhre, S., Lingjaerde, O.C., Hennessy, B.T., Aure, M.R., Carey, M.S., Alsner, J., Tramm, T., Overgaard, J., Mills, G.B., Borresen-Dale, A.L. et Sorlie, T. (2013). Influence of DNA copy number and mRNA levels on the expression of breast cancer related proteins. *Mol Oncol, 7*(3), 704-718. doi: 10.1016/j.molonc.2013.02.018

Nasser, S., Cunliffe, H.E., Black, M.A. et Kim, S. (2011). Context-specific gene regulatory networks subdivide intrinsic subtypes of breast cancer. *BMC Bioinformatics, 12*(Suppl 2), S3-S3. doi: 10.1186/1471-2105-12-S2-S3

Natrajan, R., Weigelt, B., Mackay, A., Geyer, F.C., Grigoriadis, A., Tan, D.S., Jones, C., Lord, C.J., Vatcheva, R., Rodriguez-Pinilla, S.M., Palacios, J., Ashworth, A. et Reis-Filho, J.S. (2010). An integrative genomic and transcriptomic analysis reveals molecular pathways and networks regulated by copy number aberrations in basal-like, HER2 and luminal cancers. *Breast Cancer Res Treat, 121*(3), 575-589. doi: 10.1007/s10549-009-0501-3

Netanely, D., Avraham, A., Ben-Baruch, A., Evron, E. et Shamir, R. (2016). Expression and methylation patterns partition luminal-A breast tumors into distinct prognostic subgroups. [journal article]. *Breast Cancer Research, 18*(1), 74. doi: 10.1186/s13058-016-0724-2

Okser, S., Pahikkala, T. et Aittokallio, T. (2013). Genetic variants and their interactions in disease risk prediction - Machine learning and network perspectives. [Review]. *BioData Mining, 6*(1). doi: 10.1186/1756-0381-6-5

Onami, S. et Kitano, H. (2006). Genome-wide prediction of genetic interactions in a metazoan. *Bioessays, 28*(11), 1087-1090. doi: 10.1002/bies.20490

Padiadpu, J., Vashisht, R. et Chandra, N. (2010). Protein–protein interaction networks suggest different targets have different propensities for triggering drug resistance. *Systems and Synthetic Biology, 4*(4), 311-322. doi: 10.1007/s11693-011-9076-5

Paladugu, S.R., Zhao, S., Ray, A. et Raval, A. (2008). Mining protein networks for synthetic genetic interactions. *BMC Bioinformatics, 9*, 426. doi: 10.1186/1471-2105-9-426

Pandey, G., Zhang, B., Chang, A.N., Myers, C.L., Zhu, J., Kumar, V. et Schadt, E.E. (2010). An integrative multi-network and multi-classifier approach to predict genetic interactions. [Article]. *PLoS Computational Biology, 6*(9). doi: 10.1371/journal.pcbi.1000928

Park, S. et Lehner, B. (2015). Cancer type-dependent genetic interactions between cancer driver alterations indicate plasticity of epistasis across cell types. [10.15252/msb.20156102]. *Molecular Systems Biology, 11*(7).

Paul, A. et Paul, S. (2014). The breast cancer susceptibility genes (BRCA) in breast and ovarian cancers. *Frontiers in bioscience (Landmark edition), 19*, 605-618.

Perkins, J.R., Diboun, I., Dessailly, B.H., Lees, J.G. et Orengo, C. (2010). Transient Protein-Protein Interactions: Structural, Functional, and Network Properties. *Structure, 18*(10), 1233-1243. doi: http://dx.doi.org/10.1016/j.str.2010.08.007

Perou, C.M., Sorlie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Rees, C.A., Pollack, J.R., Ross, D.T., Johnsen, H., Akslen, L.A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S.X., Lonning, P.E., Borresen-Dale, A.L., Brown, P.O. et Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature, 406*(6797), 747-752. doi: 10.1038/35021093

Piñero, J., Berenstein, A., Gonzalez-Perez, A., Chernomoretz, A. et Furlong, L.I. (2016). Uncovering disease mechanisms through network biology in the era of Next Generation Sequencing. *Scientific Reports, 6*, 24570. doi: 10.1038/srep24570

Porta-Pardo, E., Garcia-Alonso, L., Hrabe, T., Dopazo, J. et Godzik, A. (2015). A Pan-Cancer Catalogue of Cancer Driver Protein Interaction Interfaces. *PLoS Computational Biology, 11*(10), e1004518. doi: 10.1371/journal.pcbi.1004518

Prat, A., Karginova, O., Fan, C. et Perou, C.M. (2009). Notch-associated expression profiles in basal-like and claudin-low breast cancer molecular subtypes. *J Clin Oncol, 27*(15_suppl), 11017.

Qian, W., He, X., Chan, E., Xu, H. et Zhang, J. (2011). Measuring the evolutionary rate of protein–protein interaction. *Proceedings of the National Academy of Sciences, 108*(21), 8725-8730. doi: 10.1073/pnas.1104695108

Ramani, A.K., Bunescu, R.C., Mooney, R.J. et Marcotte, E.M. (2005). Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biology, 6*(5), R40-R40. doi: 10.1186/gb-2005-6-5-r40

Ravasz, E., Somera, A.L., Mongru, D.A., Oltvai, Z.N. et Barabasi, A.L. (2002). Hierarchical organization of modularity in metabolic networks. *Science, 297*(5586), 1551-1555. doi: 10.1126/science.1073374

Raykar, V.C., Yu, S., Zhao, L.H., Jerebko, A., Florin, C., Valadez, G.H., Bogoni, L. et Moy, L. (2009). Supervised learning from multiple experts: whom to trust when everyone lies a bit. Proceedings of the 26th Annual International Conference on Machine Learning (p. 889-896). Montreal, Quebec, Canada : ACM.

Roguev, A., Bandyopadhyay, S., Zofall, M., Zhang, K., Fischer, T., Collins, S.R., Qu, H., Shales, M., Park, H.O., Hayles, J., Hoe, K.L., Kim, D.U., Ideker, T., Grewal, S.I., Weissman, J.S. et Krogan, N.J. (2008). Conservation and rewiring of functional modules revealed by an epistasis map in fission yeast. *Science, 322*(5900), 405-410. doi: 10.1126/science.1162609

Ryan, C.J., Roguev, A., Patrick, K., Xu, J., Jahari, H., Tong, Z., Beltrao, P., Shales, M., Qu, H., Collins, S.R., Kliegman, J.I., Jiang, L., Kuo, D., Tosti, E., Kim, H.S., Edelmann, W., Keogh, M.C., Greene, D., Tang, C., Cunningham, P., Shokat, K.M., Cagney, G., Svensson, J.P., Guthrie, C., Espenshade, P.J., Ideker, T. et Krogan, N.J. (2012). Hierarchical modularity and the evolution of genetic interactomes across species. *Molecular cell, 46*(5), 691-704. doi: 10.1016/j.molcel.2012.05.028

Schrynemackers, M., Küffner, R. et Geurts, P. (2013). On protocols and measures for the validation of supervised methods for the inference of biological networks. [Review]. *Frontiers in Genetics, 4*(DEC). doi: 10.3389/fgene.2013.00262

Schuldiner, M., Collins, S.R., Thompson, N.J., Denic, V., Bhamidipati, A., Punna, T., Ihmels, J., Andrews, B., Boone, C., Greenblatt, J.F., Weissman, J.S. et Krogan, N.J. (2005). Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell, 123*(3), 507-519. doi: 10.1016/j.cell.2005.08.031

Segre, D., Deluna, A., Church, G.M. et Kishony, R. (2005). Modular epistasis in yeast metabolism. *Nat Genet, 37*(1), 77-83. doi: 10.1038/ng1489

Shalem, O., Sanjana, N.E. et Zhang, F. (2015). High-throughput functional genomics using CRISPR-Cas9. [Review]. *Nat Rev Genet, 16*(5), 299-311. doi: 10.1038/nrg3899

Sistigu, A., Manic, G., Obrist, F. et Vitale, I. (2016). Trial watch – inhibiting PARP enzymes for anticancer therapy. *Molecular & Cellular Oncology, 3*(2), e1053594. doi: 10.1080/23723556.2015.1053594

Song, J. et Singh, M. (2013). From hub proteins to hub modules: the relationship between essentiality and centrality in the yeast interactome at different scales of organization. *PLoS Comput Biol, 9*(2), e1002910. doi: 10.1371/journal.pcbi.1002910

Sonnhammer, E.L.L. et Östlund, G. (2015). InParanoid 8: Orthology analysis between 273 proteomes, mostly eukaryotic. [Article]. *Nucleic Acids Research, 43*(D1), D234-D239. doi: 10.1093/nar/gku1203

Sorlie, T., Perou, C.M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M.B., van de Rijn, M., Jeffrey, S.S., Thorsen, T., Quist, H., Matese, J.C., Brown, P.O., Botstein, D., Lønning, P.E. et Børresen-Dale, A.-L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America, 98*(19), 10869-10874. doi: 10.1073/pnas.191367098

Sorlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J.S., Nobel, A., Deng, S., Johnsen, H., Pesich, R., Geisler, S., Demeter, J., Perou, C.M., Lønning, P.E., Brown, P.O., Børresen-Dale, A.-L. et Botstein, D. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences of the United States of America, 100*(14), 8418-8423. doi: 10.1073/pnas.0932692100

Sotiriou, C. et Pusztai, L. (2009). Gene-expression signatures in breast cancer. *N Engl J Med, 360*(8), 790-800. doi: 10.1056/NEJMra0801289

Sporns, O. (2014). Contributions and challenges for network models in cognitive neuroscience. *Nat Neurosci, 17*(5), 652-660. doi: 10.1038/nn.3690

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., Toksoz, 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.

Stepanenko, A.A., Vassetzky, Y.S. et Kavsan, V.M. (2013). Antagonistic functional duality of cancer genes. *Gene, 529*(2), 199-207. doi: 10.1016/j.gene.2013.07.047

Stratton, M.R., Campbell, P.J. et Futreal, P.A. (2009). The cancer genome. *Nature, 458*(7239), 719-724. doi: 10.1038/nature07943

Suo, C., Hrydziuszko, O., Lee, D., Pramana, S., Saputra, D., Joshi, H., Calza, S. et Pawitan, Y. (2015). Integration of somatic mutation, expression and functional data reveals potential driver genes predictive of breast cancer survival. *Bioinformatics, 31*(16), 2607-2613. doi: 10.1093/bioinformatics/btv164

Szappanos, B., Kovacs, K., Szamecz, B., Honti, F., Costanzo, M., Baryshnikova, A., Gelius-Dietrich, G., Lercher, M.J., Jelasity, M., Myers, C.L., Andrews, B.J., Boone, C., Oliver, S.G., Pal, C. et Papp, B. (2011). An integrated approach to characterize

genetic interaction networks in yeast metabolism. *Nat Genet, 43*(7), 656-662. doi: 10.1038/ng.846

The Cancer Genome Atlas Research, N., Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., Shmulevich, I., Sander, C. et Stuart, J.M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. [Commentary]. *Nat Genet, 45*(10), 1113-1120. doi: 10.1038/ng.2764

Tian, F., Wang, Y., Seiler, M. et Hu, Z. (2014). Functional characterization of breast cancer using pathway profiles. *BMC Medical Genomics, 7*, 45-45. doi: 10.1186/1755-8794-7-45

Tiede, B. et Kang, Y. (2011). From milk to malignancy: the role of mammary stem cells in development, pregnancy and breast cancer. *Cell Research, 21*(2), 245-257. doi: 10.1038/cr.2011.11

Tischler, J., Lehner, B. et Fraser, A.G. (2008). Evolutionary plasticity of genetic interaction networks. *Nat Genet, 40*(4), 390-391. doi: 10.1038/ng.114

Tong, A.H., Evangelista, M., Parsons, A.B., Xu, H., Bader, G.D., Page, N., Robinson, M., Raghibizadeh, S., Hogue, C.W., Bussey, H., Andrews, B., Tyers, M. et Boone, C. (2001). Systematic genetic analysis with ordered arrays of yeast deletion mutants. [Research Support, Non-U.S. Gov't]. *Science, 294*(5550), 2364-2368. doi: 10.1126/science.1065810

Tong, A.H., Lesage, G., Bader, G.D., Ding, H., Xu, H., Xin, X., Young, J., Berriz, G.F., Brost, R.L., Chang, M., Chen, Y., Cheng, X., Chua, G., Friesen, H., Goldberg, D.S., Haynes, J., Humphries, C., He, G., Hussein, S., Ke, L., Krogan, N., Li, Z., Levinson, J.N., Lu, H., Menard, P., Munyana, C., Parsons, A.B., Ryan, O., Tonikian, R., Roberts, T., Sdicu, A.M., Shapiro, J., Sheikh, B., Suter, B., Wong, S.L., Zhang, L.V., Zhu, H., Burd, C.G., Munro, S., Sander, C., Rine, J., Greenblatt, J., Peter, M., Bretscher, A., Bell, G., Roth, F.P., Brown, G.W., Andrews, B., Bussey, H. et Boone, C. (2004). Global mapping of the yeast genetic interaction network. *Science, 303*(5659), 808-813. doi: 10.1126/science.1091317

Ulitsky, I. et Shamir, R. (2007). Identification of functional modules using network topology and high-throughput data. *BMC Systems Biology, 1*(1), 8.

van Dam, T.J.P. et Snel, B. (2008). Protein Complex Evolution Does Not Involve Extensive Network Rewiring. *PLoS Comput Biol, 4*(7), e1000132. doi: 10.1371/journal.pcbi.1000132

van den Broek, A.J., Schmidt, M.K., van 't Veer, L.J., Tollenaar, R.A.E.M. et van Leeuwen, F.E. (2015). Worse Breast Cancer Prognosis of BRCA1/BRCA2 Mutation Carriers: What's the Evidence? A Systematic Review with Meta-Analysis. *PLoS ONE, 10*(3), e0120189. doi: 10.1371/journal.pone.0120189

van Noort, V., Snel, B. et Huynen, M.A. (2004). The yeast coexpression network has a small-world, scale-free architecture and can be explained by a simple model. *EMBO Rep, 5*(3), 280-284. doi: 10.1038/sj.embor.7400090

Vastrik, I., D'Eustachio, P., Schmidt, E., Joshi-Tope, G., Gopinath, G., Croft, D., de Bono, B., Gillespie, M., Jassal, B., Lewis, S., Matthews, L., Wu, G., Birney, E. et Stein, L. (2007). Reactome: a knowledge base of biologic pathways and processes. [journal article]. *Genome Biology, 8*(3), 1-13. doi: 10.1186/gb-2007-8-3-r39

Wagner, A. (2001). The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol Biol Evol, 18*(7), 1283-1292.

Wang, X., Thijssen, B. et Yu, H. (2013). Target Essentiality and Centrality Characterize Drug Side Effects. *PLOS Computational Biology, 9*(7), e1003119. doi: 10.1371/journal.pcbi.1003119

Warburg, O. (1956). On respiratory impairment in cancer cells. *Science, 124*(3215), 269-270.

Weigelt, B., Mackay, A., A'Hern, R., Natrajan, R., Tan, D.S.P., Dowsett, M., Ashworth, A. et Reis-Filho, J.S. (2010). Breast cancer molecular profiling with single sample predictors: A retrospective analysis. [Article]. *The Lancet Oncology, 11*(4), 339-349. doi: 10.1016/S1470-2045(10)70008-5

White, F.M. (2011). The potential cost of high-throughput proteomics. *Sci Signal, 4*(160), pe8. doi: 10.1126/scisignal.2001813

Winterbach, W., Mieghem, P.V., Reinders, M., Wang, H. et Ridder, D.d. (2013). Topology of molecular interaction networks. *BMC Systems Biology, 7*, 90-90. doi: 10.1186/1752-0509-7-90

Wong, S.L., Zhang, L.V., Tong, A.H.Y., Li, Z., Goldberg, D.S., King, O.D., Lesage, G., Vidal, M., Andrews, B., Bussey, H., Boone, C. et Roth, F.P. (2004). Combining biological networks to predict genetic interactions. [Article]. *Proceedings of the National Academy of Sciences of the United States of America, 101*(44), 15682-15687. doi: 10.1073/pnas.0406614101

Wuchty, S. (2014). Controllability in protein interaction networks. *Proceedings of the National Academy of Sciences of the United States of America, 111*(19), 7156-7160. doi: 10.1073/pnas.1311231111

Wuchty, S. et Almaas, E. (2005). Peeling the yeast protein network. *Proteomics, 5*(2), 444-449. doi: 10.1002/pmic.200400962

Yang, W., Xia, Y., Ji, H., Zheng, Y., Liang, J., Huang, W., Gao, X., Aldape, K. et Lu, Z. (2011). Nuclear PKM2 regulates β-catenin transactivation upon EGFR activation. *Nature, 480*(7375), 118-122. doi: 10.1038/nature10598

Ye, P., Peyser, B.D., Pan, X., Boeke, J.D., Spencer, F.A. et Bader, J.S. (2005). Gene function prediction from congruent synthetic lethal interactions in yeast. *Molecular systems*

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. doi: 10.1126/science.1158684

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 Comput Biol, 3*(4), e59. doi: 10.1371/journal.pcbi.0030059

Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M.C., Zimmerman, L.J., Shaddox, K.F., Kim, S., Davies, S.R., Wang, S., Wang, P., Kinsinger, C.R., Rivers, R.C., Rodriguez, H., Townsend, R.R., Ellis, M.J.C., Carr, S.A., Tabb, D.L., Coffey, R.J., Slebos, R.J.C., Liebler, D.C. et the, N.C. (2014). Proteogenomic characterization of human colon and rectal cancer. *Nature, 513*(7518), 382-387. doi: 10.1038/nature13438

Zhang, H., Liu, T., Zhang, Z., Payne, S.H., Zhang, B., McDermott, J.E., Zhou, J.Y., Petyuk, V.A., Chen, L., Ray, D., Sun, S., Yang, F., Chen, L., Wang, J., Shah, P., Cha, S.W., Aiyetan, P., Woo, S., Tian, Y., Gritsenko, M.A., Clauss, T.R., Choi, C., Monroe, M.E., Thomas, S., Nie, S., Wu, C., Moore, R.J., Yu, K.H., Tabb, D.L., Fenyo, D., Bafna, V., Wang, Y., Rodriguez, H., Boja, E.S., Hiltke, T., Rivers, R.C., Sokoll, L., Zhu, H., Shih Ie, M., Cope, L., Pandey, A., Zhang, B., Snyder, M.P., Levine, D.A., Smith, R.D., Chan, D.W. et Rodland, K.D. (2016). Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell, 166*(3), 755-765. doi: 10.1016/j.cell.2016.05.069

Zhong, W. et Sternberg, P.W. (2006). Genome-Wide Prediction of C. elegans Genetic Interactions. *Science, 311*(5766), 1481-1484

Zhong, W. et Sternberg, P.W. (2007). Automated data integration for developmental biological research. [Review]. *Development, 134*(18), 3227-3238. doi: 10.1242/dev.001073

Zinman, G.E., Zhong, S. et Bar-Joseph, Z. (2011). Biological interaction networks are conserved at the module level. *BMC Systems Biology, 5*, 134-134. doi: 10.1186/1752-0509-5-134

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 Comput Biol, 4*(8), e1000140. doi: 10.1371/journal.pcbi.1000140