Aix-Marseille Université

Faculté des Sciences de Luminy

École Doctorale des Sciences de la Vie et de la Santé

N° attribué par la bibliothèque

|_|_|_|_|_|_|_|_|_|_|_|_|_|

THÈSE

Présentée et soutenue publiquement

le 20 décembre 2013 par

Maxime Ulysse Garcia

Né le 14 mai 1982 à Arles (13)

Découverte de biomarqueurs prédictifs en cancer du sein par Intégration Transcriptome-Interactome

Pour l'obtention du grade de

Docteur d'Aix-Marseille Université

Spécialité : Bioinformatique

Jury :

M.	Jacques van Helden	AMU - TAGC	(Président)
M.	Philippe Dessen	Génétiques des Tumeurs	(Rapporteur)
M.	Benno Schwikowski	Institut Pasteur	(Rapporteur)
M.	Pascal Barbry	IPMC	(Examinateur)
M.	François Bertucci	IPC	(Directeur)
M.	Ghislain Bidaut	AMU - CRCM	(Co-directeur)

Les travaux réalisés pendant cette thèse ont été effectués au Centre de Recherche en Cancérologie de Marseille (CRCM) dans la plateforme de Bioinformatique Intégrative du CRCM (Cibi).
Le CRCM est une unité mixte de recherche affiliée à Aix-Marseille Université (AMU) en tant qu'UM105, l'Institut Paoli-Calmettes (IPC), l'Institut National de la Santé et de la Recherche Médicale (Inserm) en tant qu'U1068 et le Centre National de la Recherche Scientifique (CNRS) en tant qu'UMR7258.

Les recherches réalisées pour cette thèse ont été financées par l'Institut National du Cancer (INCa) et l'Inserm. Notre Cluster Beowulf est financé par la Fondation pour la Recherche Médicale (FRM). Maxime U. Garcia a été financé par une bourse de thèse Inserm/Région Provence Alpes Côte d'Azur (PACA) pendant ces travaux.

À ma famille
À mes parents
À Célia

REMERCIEMENTS

Je voudrais tout d'abord exprimer mes plus profonds remerciements à Ghislain sans qui cette thèse n'aurait pas eu lieu, et qui en plus d'un encadrement sans faille depuis le début a été une source de conseils tout au long de cette thèse. Je remercie de même François pour toute l'aide qu'il a pu m'apporter, pour la rapidité, la justesse de ses réponses et ses corrections à toutes heures. Je voudrais également remercier pour les conseils qu'elles m'ont apportées, toutes les personnes avec qui j'ai collaboré pour cette thèse : Daniel, Pascal, Arnaud, Sabrina, Raphaële, Renaud.

Je tiens également à exprimer mes plus sincères remerciements à tous les membres de la plateforme de Bioinformatique Intégrative du CRCM qui m'ont supporté malgré eux tout le long de ma thèse et de mon écriture : Samuel, Olivier, Alexandre, Fanny, Claire, Quentin, Guillaume. Je voudrais également inclure dans ces remerciements l'ensemble du CRCM : Françoise et Jean-Paul qui ont soutenu ce projet dès le début, Caroline pour avoir été un soutien moral sans faille, Sébastien, Will, Sébastien, Marie-José, Yves, Khedidja, Avais, MLA, MLT, Vincent, Javier, Émilie, Marion, Julien, Armelle, Axelle, Pascale, Bernard, Myriam, Julie, François, Laurence D, Laurence L et je présente mes excuses à tous ceux que j'aurais très probablement oublié…

Je souhaite aussi adresser mes remerciements à tous mes amis qui m'ont soutenu ces dernières années, mais qui sont trop nombreux pour les citer tous ici…

J'exprime toute ma gratitude à mes parents Claude et Marie-Claude, mon frère Florent, Karine et leur trois enfants Robin, Thibaud et Tristan ainsi que tout le reste de la famille…

Je conclurai en remerciant de tout mon cœur Célia pour son aide et son soutien.

TABLE DES MATIÈRES

Thèse **i**

Table des Matières **v**

Index des Tableaux **viii**

Index des Figures **ix**

I Introduction générale **2**
I.1 Problématique . . . 3
I.2 La recherche sur le cancer . . . 5
I.2.a Historique . . . 5
I.2.b Rappels sur l'expression des gènes et sa régulation . . . 7
I.2.c Les caractéristiques spécifiques des cancers . . . 13
I.3 L'apport des technologies à haut-débit à la recherche sur le cancer . . . 22
I.3.a L'ère post-génomique et la fin du paradigme *un gène, une maladie* . . . 22
I.3.b La médecine prédictive et la médecine personnalisée . . . 24
I.4 La recherche sur le cancer du sein . . . 26
I.4.a Les caractéristiques du cancer du sein . . . 26
I.4.b Diagnostics et traitements . . . 28
I.4.c Les classifications utilisées dans le cancer du sein . . . 30
I.4.d Intérêts des signatures prédictives dans le cancer du sein . . . 34
I.4.e Les limitations des technologies utilisées . . . 37
I.4.f Les solutions . . . 38
I.5 Conclusion . . . 40

II Méthodes **42**
II.1 Les avantages de l'intégration de données . . . 43
II.1.a Les avantages de l'intégration de données d'expression des gènes et d'interactions protéine-protéine . . . 43
II.1.b L'intégration de données d'expression des gènes et d'interactions protéine-protéine . . . 45
II.2 L'Intégration Transcriptome-Interactome . . . 45
II.2.a L'intégration massive de données d'expression des gènes et de données d'interactions protéine-protéine . . . 45
II.3 Données transcriptome . . . 47
II.3.a Constitution d'un compendium de données transcriptome dans le cancer du sein . . . 47
II.4 Interactions protéine-protéine . . . 51
II.4.a Assemblage d'interactomes humains . . . 51
II.4.b Nature des interactions et bases de données d'interactions utilisées . . . 51
II.5 Données et outils supplémentaires . . . 52
II.6 Présentation de l'algorithme ITI . . . 52
II.6.a Détection des sous-réseaux . . . 52
II.6.b Validation statistique . . . 54
II.6.c Sélection de sous-réseaux . . . 56
II.6.d Création d'une ressource bioinformatique permettant l'analyse des sous-réseaux et la reproductibilité de la recherche . . . 58
II.6.e Utilisation des SVMs pour la classification . . . 59
II.6.f Stratification, organisation des données et classification . . . 60
II.7 Conclusion . . . 62

III Analyse non-supervisée **64**
III.1 Détails de l'analyse non-supervisée . . . 65
III.1.a Organisation des études . . . 65
III.2 Exploration des sous-réseaux . . . 67
III.3 Conclusion . . . 70

IV Analyse supervisée **72**
IV.1 Détails de l'analyse supervisée . . . 73
IV.1.a Organisation des jeux de données transcriptome en deux études . . . 73
IV.2 Performances des signatures obtenues sur la prédiction de la rechute métastatique . . . 73
IV.3 Exploration des sous-réseaux . . . 78
IV.4 Conclusion . . . 81

V Discussion Générale **83**
V.1 Rappels sur les travaux effectués . . . 84
V.2 Importance des données initiales . . . 84

V.3 Caractéristiques des signatures réalisées . . . 85
V.3.a Amélioration de la stabilité, robustesse et reproductibilité . . . 85
V.3.b Significativité biologique relevante . . . 85
V.3.c Taille des sous-réseaux . . . 87
V.4 Création d'une base de donnés de sous-réseaux . . . 87
V.5 Perspectives . . . 88
V.5.a Améliorations de l'algorithme . . . 88
V.5.b Intégration d'autres types de données . . . 88
V.5.c Étude de l'importance de la nature de l'interaction . . . 88
V.6 Conclusion . . . 89

VI Conclusion Générale **91**

Annexes **95**

A Nomenclatures **96**

B Abréviations **97**
B.1 Gènes . . . 97
B.2 Protéines . . . 98
B.3 Institutions . . . 98
B.4 Divers . . . 99

C Publications **101**
C.1 Chapitre *Linking Interactome to Disease* . . . 101
C.2 Article *Interactome–transcriptome integration* . . . 124
C.3 Chapitre *Large Scale Transcriptome-Interactome Integration* . . . 132
C.4 Chapitre *CNV-Interactome-Transcriptome Integration* . . . 151

Références **177**

Colophon **191**

Abstract **192**

INDEX DES TABLEAUX

I.1 Découvertes et événements majeurs dans le domaine du cancer. 6
I.2 Effectif annuel moyen de décès et taux observé (standardisé monde) de mortalité des cancers pour la période 2004-2008. 27

II.1 Liste des jeux de données inclus dans notre compendium de données d'expression dans le cancer du sein. 48
II.2 Liste des jeux de données inclus pour notre analyse non supervisée (cf Section II.7). 49
II.3 Liste des jeux de données inclus pour notre analyse supervisée (cf Section III.3). 50

III.1 Organisation de la validation croisée 66
III.2 Seuils de p-value et valeur de consensus choisie 66
III.3 Nombre de sous-réseaux découverts pour chacune des analyses 66

IV.1 Liste des jeux de données utilisés dans l'analyse supervisée. 74
IV.2 Taille et p-value de la signature retenue pour chacune des études réalisées. . . 74
IV.3 Comparaison des résultats de classification entre ITI et d'autres signatures sur les jeux de données de validation Desmedt et van de Vijver pour les tumeurs ER- et ER+. 76
IV.4 Enrichissement en termes GO des sous-réseaux ER- et ER+ 80

INDEX DES FIGURES

I.1 Représentation schématique de l'expression d'un gène dans une cellule eucaryote. . . . 8
I.2 Profil d'expression de 42 tissus en représentation par cartes auto-organisatrices de Kohonen. . . . 11
I.3 Caractéristiques du Cancer. . . . 16
I.4 Classification en sous-types moléculaires. . . . 33
I.5 Courbes de survie en fonction des sous-types moléculaires. . . . 35

II.1 Avantages de l'intégration de données d'expression des gènes et d'interactions protéine-protéine sur la précision de la classification de la rechute métastatique par rapport à une analyse classique sur des données d'expression des gènes. . . . 44
II.2 Algorithme détaillant l'intégration de données d'expression des gènes et d'interactions protéine-protéine. . . . 46
II.3 Principe de la sélection des sous-réseaux avec Intégration Transcriptome-Interactome. . . . 55
II.4 Distribution des scores des sous-réseaux pour le jeu de données Desmedt. . . . 57
II.5 Workflow complet des données. . . . 61

III.1 Exploration fonctionnelle du sous-réseau 387-4 68

IV.1 Comparatif des courbes de survies des patients. . . . 79
IV.2 Représentation graphique d'une partie du sous-réseau 6693, Étude 1 ER-. . . . 80

V.1 Diagramme de Venn comptabilisant les gènes communs entre les différentes signatures classiques. . . . 86

CHAPITRE

I

INTRODUCTION GÉNÉRALE

Résumé

Ce chapitre introductif présente de façon générale le cancer. Après un court historique des différentes actions entamées contre cette maladie, nous explorerons les caractéristiques biologiques des cancers. Puis, nous détaillerons les spécificités du cancer du sein, avant d'aborder l'intérêt de la médecine prédictive et personnalisée pour les signatures prédictives de l'évolution des cancers, et ce plus spécifiquement dans le cadre du cancer du sein.

Sommaire

I.1 Problématique 3
I.2 La recherche sur le cancer 5
I.3 L'apport des technologies à haut-débit à la recherche sur le cancer 22
I.4 La recherche sur le cancer du sein 26
I.5 Conclusion 40

I.1 Problématique

Le cancer est, dans les pays occidentalisés, la seconde cause de décès après les maladies cardio-vasculaires. Ceci en fait une préoccupation majeure de santé publique. En 1918, avec la création de La Ligue franco-anglo-américaine contre le cancer [1], une action associative est mise en place pour lutter contre le cancer. Les gouvernements s'impliquent eux aussi. Ainsi en 1937 aux États-Unis d'Amérique, le National Cancer Institute (NCI) Act [(1)] a établi le NCI, institut fédéral de recherche contre le cancer, qui fut par la suite renforcé par le président Nixon et le National Cancer Act en 1971.

L'organisation non-gouvernementale European Organisation for Research and Treatment of Cancer (EORTC) a été fondée en 1962 dans le but de stimuler la recherche clinique en Europe. L'Organisation Mondiale de la Santé (OMS) a créé en 1965 l'International Agency for Research on Cancer (IARC), agence intergouvernementale de recherche contre le cancer, dans le but de coordonner la recherche sur les causes du cancer. L'IARC classifie les substances suivant leur cancérogénicité. En France, l'INCa, groupement d'intérêt public fondé en 2005, est chargé de coordonner la recherche scientifique et la lutte contre le cancer. Les Plans Cancer I (2003-2007) et II (2009-2013), plans de lutte gouvernementaux contre le cancer, mettent eux aussi l'accent sur la recherche, et ont ainsi constitué les Cancéropôles, entités supra-régionales dont le but est de coordonner et de mettre en réseau des équipes de recherche.

Le cancer est une maladie génétique, causée par l'acquisition de mutations qui peuvent être déclenchées par plusieurs substances ou agents. Ces facteurs peuvent être chimiques, physiques ou encore biologiques. Une susceptibilité génétique héréditaire est également mise en cause. Tous ces éléments font du cancer une maladie extrêmement complexe, multifactorielle et hétérogène. Les efforts de la recherche se dirigent par conséquent non seulement vers des traitements ciblés, mais aussi vers des méthodes de classification des tumeurs cancéreuses dans le but de trouver des groupes de patients permettant d'affiner et d'adapter le traitement.

1. actuellement La Ligue nationale contre le cancer

Ainsi, dans le domaine du cancer du sein, le système de classification par grade de Scarff-Bloom-Richardson [2], et sa version étendue par les critères de Nottingham [3] se basent sur la similarité microscopique entre les tumeurs et le tissu sain. L'OMS a établi en 2003 une classification histopathologique des tumeurs [4].

D'autre part, depuis les années 1990, des consortiums internationaux étudient le génome humain. Les buts du Projet Génome Humain (Human Genome Project) étaient le séquençage complet du génome humain et l'identification de tous les gènes [5]. Plus récemment, en septembre 2012, le projet Encyclopedia of DNA elements (ENCODE) a permis l'identification des éléments fonctionnels contenus dans l'ADN non-codant [6]. Ces projets permettent de mieux appréhender la complexité du génome humain, et aident ainsi à sa compréhension.

Les puces à ADN permettent la détermination du profil génétique des tumeurs. Cette utilisation comme outil de diagnostic présente l'avantage de pouvoir faire appel à plus de vingt mille sondes pour fournir une signature du type cellulaire étudié. Si l'on considère que chaque type de tumeur présente une signature spécifique, on pourrait ainsi virtuellement distinguer, classer tous les types de tumeurs et donner un traitement approprié.

Le cancer du sein est le cancer le plus répandu et le plus mortel chez la femme. Les patientes sans ganglions à un stade précoce subissent une chimiothérapie adjuvante que l'on pourrait éviter dans 70 à 80 % des cas [7]. Deux études de puces à ADN ont permis d'établir deux signatures : l'une de 70 gènes (van't Veer et al. [8]) et l'autre de 76 gènes (Wang et al. [9]) prédisant la rechute métastatique dans le cancer du sein. Mais seulement 3 gènes sont communs entre ces deux signatures [10]. Il a également été prouvé que plus d'une signature à 70 gènes existait avec le même pouvoir prédictif [11]. De telles signatures présentent donc une instabilité et un manque de reproductibilité et de généralisation. Nous présentons, dans cet ouvrage, une méthode pour palier à ces inconvénients [12–15]. Nous commencerons par décrire les différentes actions entamées historiquement contre le cancer. Nous explorerons ensuite les caractéristiques biologiques des cancers. Puis, nous développerons les spécificités du cancer du sein, avant d'aborder l'intérêt de la médecine prédictive et personnalisée pour les signatures prédictives de l'évolution des cancers, et ce plus spécifiquement dans le cadre du cancer du sein.

I.2 La recherche sur le cancer

I.2.a Historique

HIPPOCRATE DE COS, médecin grec des V^e et IV^e siècles avant JC est connu comme étant le père de la médecine, mais ce n'est pas le premier à décrire le cancer. En Égypte antique, le papyrus Ebers (1 500 ans avant JC) décrit déjà cette maladie. Quelques dizaines d'années avant Hippocrate, Hérodote décrit la tumeur du sein de la femme de Darius I^er , Roi de l'empire Perse. Mais s'il n'est pas le premier à le décrire, c'est bien Hippocrate qui donne son nom au cancer. Le mot *karkinos* qui signifie crabe en grec, désigne pour Hippocrate le crabe dévorant les tissus, et conduisant de manière inéluctable à la mort.

Il faut cependant attendre la fin du XIX^e siècle pour que la technologie permette plus que des descriptions et des théories. En 1585, Ambroise Paré, dans son traité des *tumeurs contre nature* décrit la tumeur du sein d'une dame d'honneur de la reine Catherine de Médicis. En 1693, Houppeville écrit un traité sur *la guérison du cancer du sein*. Il y présente *La théorie infectieuse* qui défend la contagiosité du cancer. Au début du XIX^e siècle Xavier Bichat, puis René Laennec, sont à l'origine de la théorie cellulaire moderne du cancer.

Les premières révolutions apparaissent à la fin du XIX^e siècle avec les travaux de Louis Pasteur permettant le développement de l'asepsie fortement promue par Eugène Koeberlé. La découverte des rayons X par Röntgen (1895) permet également un meilleur contrôle des conditions d'interventions chirurgicales. En découle une amélioration de la survie post-opératoire. En 1914, Theodor Boveri met en évidence l'importance des mutations chromosomiques dans le cancer. Dès le milieu du XX^e siècle, les découvertes de la transmission de l'information cellulaire par l'ADN, et le décodage du code génétique ont posé les jalons des recherches actuelles sur le génome humain. Le Tableau I.1, inspiré par DeVita and Rosenberg [16], nous présente une fresque historique intégrant les différentes découvertes et événements majeurs dans le domaine de la recherche sur le cancer.

Tableau I.1 – Découvertes et événements majeurs dans le domaine du cancer.

Année	*Découverte ou événement*
1863	Origine cellulaire du cancer (*Virchow*)
1889	Hypothèse de la graine et du sol (*Paget*)
1895	Rayons X *(Röntgen)*
1914	Mutations chromosomiques dans le cancer (*Boveri*)
1918	Création de La Ligue franco-anglo-américaine contre le cancer
1924	Hypothèse de Warburg
1937	Fondation du NCI
1944	Transmission de l'information cellulaire par l'ADN (*Avery*)
1950	Disponibilité des drogues contre le cancer via *Cancer Chemotherapy National Service Center*
1953	Structure de l'ADN (*Watson & Crick*)
1961	Décodage du code génétique (*Nirenberg & Matthaei*)
1962	Fondation de l'EORTC
1965	Fondation de l'IARC
1970	Transcriptase inverse
1971	Enzymes de restriction National Cancer Act (*War on cancer*)
1975	Hybridomes et anticorps monoclonaux Suivi des statistiques sur le cancer par le programme SEER
1976	Origine cellulaire des oncogènes rétroviraux
1979	Facteur de croissance épidermique et son récepteur
1981	Suppression de la croissance tumorale par TP53
1984	Protéines G et signalisation cellulaire
1986	Gène *RB1*, cause génétique du rétinoblastome
1990	Première baisse de l'incidence et de la mortalité du cancer
1991	Association entre mutation du gène *APC* et cancer colorectal
1994	Syndromes génétiques du cancer Association entre le gène *BRCA1* et le cancer du sein
2000	Séquençage du génome humain
2002	Épigénétique dans le cancer MicroARN dans le cancer
2003	Plan Cancer I (2003-2007)
2005	Fondation de l'INCa
2005	Première baisse dans le nombre total de morts à cause du cancer
2006	Interaction tumeur et stroma
2009	Plan Cancer II (2009-2013)

Sources : DeVita and Rosenberg[16]

I.2.b Rappels sur l'expression des gènes et sa régulation

Le gène est une unité d'information constituée d'une séquence ADN utilisée pour synthétiser une protéine ou un ARN qui aura un rôle dans le fonctionnement de la cellule. Chez l'humain, le nombre de gènes qui correspondent à une protéine est estimé à environ 30 000 [5] (+/- 10 000 suivant les estimations). La taille du génome humain est de 3 200 000 000 paires de base (pb) [5]. La taille moyenne d'un gène est d'environ 3 000 pb, mais elle peut être très variable. 100 000 000 pb est une approximation rapide de la taille totale de l'ADN correspondant à des protéines. Cette petite portion du génome (environ 3 %) est qualifié de codante. La majeure partie du génome humain est donc constituée de séquences non-codantes, qui correspondent notamment à des régions régulatrices de l'ADN [6].

La séquence ADN d'un gène subit un ensemble de processus au cours desquels l'information contenue dans l'ADN sert à synthétiser des protéines ou des ARNs fonctionnels. Cet ensemble de processus est appelé expression des gènes et comporte plusieurs étapes, comme le montre la Figure I.1 :

La transcription

Étape au cours de laquelle l'ADN est transcrit en ARN par les ARN polymérases. Des facteurs de transcriptions contrôlent la fixation de l'ARN polymérase au promoteur. Plusieurs types d'ARN polymérases interviennent dans la transcription de plusieurs types d'ARNs. Les ARNs codants pour des protéines sont les ARNm. Cette étape se déroule dans le noyau de la cellule.

La maturation de l'ARNm

Étape au cours de laquelle les extrémités de l'ARNm sont modifiées (ajout de la coiffe en 5' et polyadénylation en en 3'). S'il y en a les introns sont épissés (excision des régions non-codantes). Cette étape se déroule dans le noyau de la cellule et est nécessaire pour que l'ARNm puisse sortir du noyau.

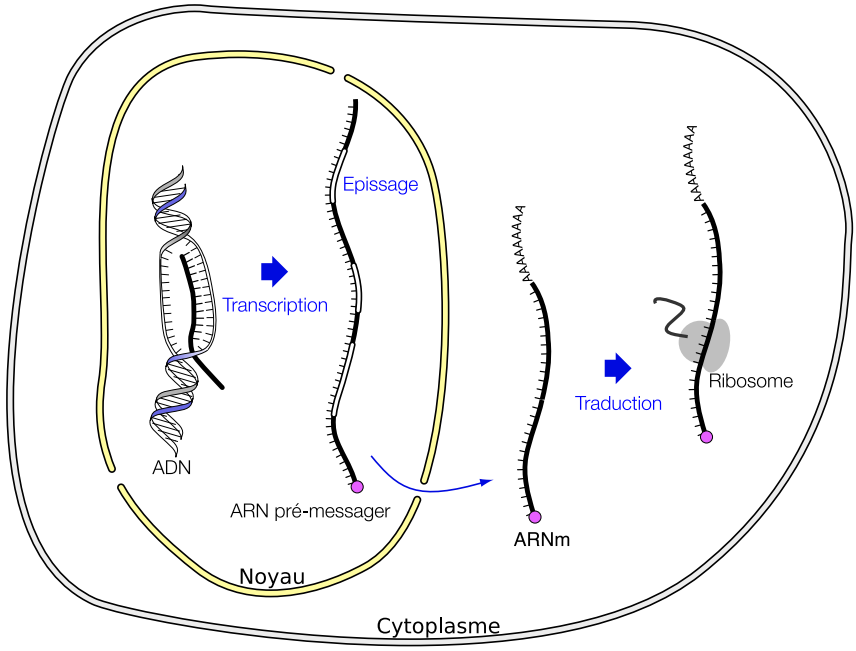


Figure I.1 – Représentation schématique de l'expression d'un gène dans une cellule eucaryote.

La traduction

Étape lors de laquelle l'ARNm mature est traduit en protéine. Cette étape se déroule dans le cytoplasme de la cellule, dans le réticulum endoplasmique et nécessite les ARNt et les ribosomes.

Nous allons explorer ici quelques-uns des mécanismes de la régulation de l'expression des gènes en commençant par le niveau cellulaire.

Les protéines régulatrices de la transcription se lient spécifiquement à des séquences d'ADN et recrutent les co-facteurs ainsi que l'appareil de transcription. Des analyses de type Chromatin immunoprecipitation (ChIP) (ChIP combined with DNA microarray analysis (ChIP-Chip) et ChIP combined with HTS (ChIP-Seq)) ont été utilisées chez l'humain pour identifier les gènes cibles de plusieurs régulateurs de la transcription [17–19].

La famille de facteurs de transcription (*transcription factors)* (TFs) *E2F* a été ainsi impliquée dans le contrôle de la progression du cycle cellulaire, la prolifération, la synthèse de

l'ADN, sa réplication, sa surveillance et sa réparation, la condensation de la chromatine, la ségrégation des chromosomes[20].

Des mécanismes moléculaires épigénétiques[2] participent également à la régulation des gènes. Ils peuvent altérer l'expression des gènes sans en changer la séquence. La méthylation de l'ADN est un phénomène en relation direct avec l'expression des gènes. Une faible méthylation favorise la transcription mais une forte méthylation, au contraire, l'inhibe.

Chez les eucaryotes, lorsque le promoteur d'un gène est méthylé, le gène en aval est en général réprimé et n'est donc plus transcrit en ARNm. Chez les mammifères, des facteurs environnementaux peuvent de plus influencer cette méthylation[21].

D'autres mécanismes épigénétiques peuvent intervenir lors des différents processus de l'expression des gènes : la condensation de la chromatine, la transcription, le transport et la dégradation de l'ARNm, la traduction et la modification post-traductionnelle de la protéine[22–24].

Les micro ARN (*micro RNA (miRNA)*) (miARN) sont des régulateurs post-transcriptionnels capables de bloquer l'expression d'un gène. Leur séquence est complémentaire d'un ARNm cible, et leur appariement conduit donc à la répression post-transcriptionnelle ou à la dégradation de cet ARNm[25,26]. Le génome humain comprendrait environ un millier de gènes de miARN[27], qui cibleraient jusqu'à 60 % des gènes[28,29].

Il y a plusieurs centaines de types cellulaires différents dans le corps humain. Chaque cellule possède le même patrimoine génétique (mis à part les gamètes et les érythrocytes), mais chacune exprime un certain nombre de gènes ce qui la maintient dans une différenciation plus ou moins poussée.

Pendant la différenciation, certains gènes sont exprimés alors que d'autres sont réprimés. Une cellule non-spécialisée se spécialise ainsi en un des nombreux types cellulaires composant le corps. Les mécanismes de la régulation des gènes expliquent comment la cellule différenciée va exprimer une partie spécifique de son génome et développer des structures précises et acquérir certaines fonctions.

2. du grec *epi*, au dessus

La différenciation peut entraîner des changements dans nombre d'aspects de la physiologie de la cellule : sa taille, sa forme, sa polarité, son activité métabolique, sa sensibilité à certains signaux et son expression des gènes. Chaque type cellulaire exprime donc un ensemble de gènes qui lui est propre [30–32].

Un ensemble fonctionnel de cellules semblables, ayant la même origine et contribuant à la même fonction est un tissu. C'est par la régulation de l'expression des gènes que les cellules saines respectent l'homéostasie tissulaire. Leur croissance est contrôlée, et leur mort programmée. Elles conservent leur équilibre de fonctionnement en dépit des contraintes, et permettent la survie de l'organisme.

La Figure I.2 représente les expressions des gènes de 42 types cellulaires différents en utilisant une représentation par cartes auto-organisatrices de Kohonen. Les cartes auto organisatrices de Kohonen définissent une projection d'un espace sur une grille régulière à deux dimensions. Une fonction de voisinage est utilisé lors de la construction de la carte, ce qui préserve les propriétés topologiques de l'espace qui est projeté. Initialement développée pour visualiser des distributions de mesures vectorielles, cette méthode de représentation peut être appliquée pour visualiser tout type de données [33,34].

Les gènes gardant toujours la même position quelle que soit la représentation, cette représentation permet d'illustrer les différences d'expression des gènes entre différents tissus, ainsi que de les situer les uns par rapport aux autres [31].

Les tissus eux-même sont assemblés en organes dont l'activité peut être régulée par les hormones. Comme nous le verrons par la suite dans la Section I.4, les hormones ont un rôle significatif dans le développement du cancer du sein. Nous allons donc faire un rappel à leur sujet. Ce sont des composés chimiques sécrétés par des cellules endocrines qui agissent à distance via des récepteurs situés sur des cellules cibles. Elles sont capables d'agir à très faibles doses et régulent l'activité d'un ou plusieurs organes. Les effets des hormones peuvent être variés :

— Stimulation ou inhibition de la croissance

— Activation ou arrêt de l'apoptose

— Stimulation ou inhibition du système immunitaire

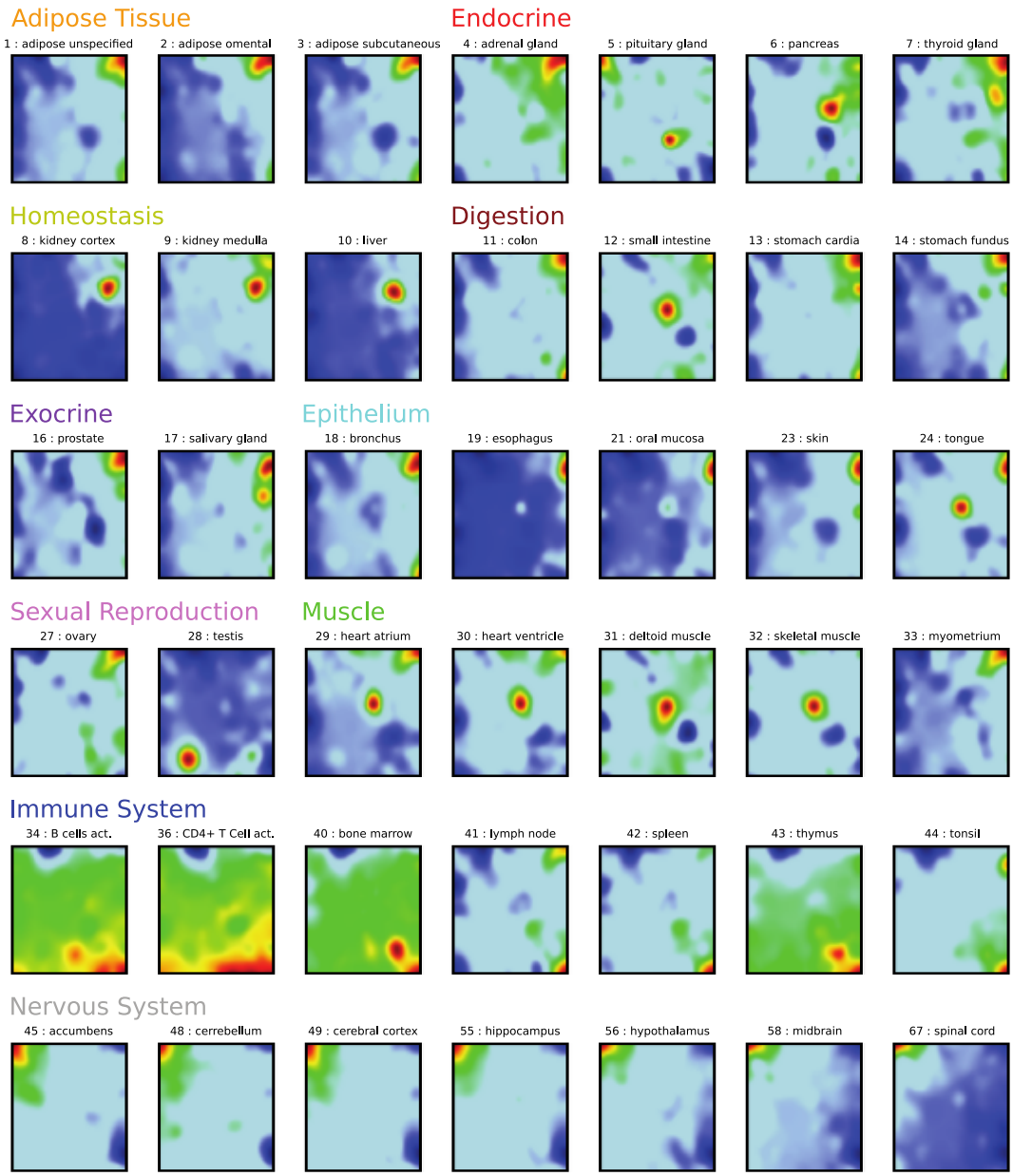


Figure I.2 – Profil d'expression de 42 tissus en représentation par cartes auto-organisatrices de Kohonen.

Figure inspirée de Wirth et al..

— Régulation du métabolisme
— Préparation du corps à la puberté, à la grossesse, à la ménopause
— Contrôle du cycle reproductif
— Sensation de faim
— Variation d'humeur
— Excitation sexuelle

Les hormones peuvent également réguler la production et la libération d'autres hormones. Nous pouvons prendre pour exemple la préparation de l'organisme féminin a une éventuelle grossesse est réalisé par le cycle menstruel. La production d'hormones (œstrogènes et progestérone) par les ovaires est sous l'influence de plusieurs signaux. La GnRH, sécrétée par l'hypothalamus, agit sur l'hypophyse qui sécrète à son tour les hormones FSH et LH qui agissent sur les ovaires pour la production des œstrogènes et de la progestérone.

Les œstrogènes assurent le développement et le maintien des caractères sexuels secondaires :

— Augmentation de volume de l'utérus, du vagin et des organes génitaux externes
— Développement des seins
— Apparition de poils axillaires et pubiens
— Augmentation des dépôts de tissus adipeux sous-cutanés (principalement aux hanches et aux seins)
— Élargissement du bassin
— Début des menstruations

Lors de la puberté, les œstrogènes interviennent dans la poussée de croissance osseuse et maintiennent la solidité de l'os. La progestérone agit en lien avec les œstrogènes lors de l'établissement du cycle menstruel. Elle est synthétisée par les ovaires à partir du cholestérol et assure le maintien et la transformation de la muqueuse utérine. Si l'ovule n'est pas fécondé, la chute de sa concentration induit l'apparition des menstruations. La progestérone prépare également les seins à la lactation.

L'horloge moléculaire contrôlant le cycle circadien est responsable de la régulation de nombreuses hormones. C'est une adaptation de l'organisme à l'alternance du cycle jour / nuit. Cette horloge implique une boucle transcriptionnelle entre les gènes *CLOCK* et *ARNTL*. Le gène *NPAS2* encodant pour le TF NPAS2 en fait également partie[35]. La variation circadienne de l'expression des gènes est également contrôlée par les miARN[36]. Le syndrome de Laron, causant un déficit en *GHR*, confère une résistance au diabète et au cancer[37].

Le rôle des hormones dans le cancer, et ce particulièrement dans le cancer du sein n'est pas négligeable. La régulation de l'expression des gènes est donc le mécanisme fondamental permettant la différenciation cellulaire, la morphogenèse et l'adaptabilité d'un organisme vivant à son environnement. Toutes les cellules interagissent ensemble et sont dépendantes du bon fonctionnement des autres cellules.

Comme nous l'avons vu, les cellules de même type sont réunies en tissus, eux-mêmes formant des organes qui interagissent entre eux à un niveau supérieur. Mais toute cette organisation nécessite une coordination, d'où la nécessité d'un système de communication à tous les niveaux, cellulaire, tissulaire et organique. Nous allons maintenant voir ce qui peut arriver lors qu'un dysfonctionnement dans la communication entre les cellules survient.

I.2.c Les caractéristiques spécifiques des cancers

Le cancer est une maladie très complexe. Il est souvent dit qu'il n'y a pas un cancer, mais des cancers. Dans l'organisme, chaque cellule est une entité vivante qui fonctionne de manière autonome, mais coordonnée avec les autres dans un ensemble dont la survie dépend de la bonne organisation de ses constituants. Le cancer est provoqué par un enchaînement d'événements qui conduisent les cellules saines à ne plus être coordonnées mais à proliférer de façon non-régulée.

Des réseaux de régulations contrôlent la prolifération et l'homéostasie des cellules saines, ce sont ces réseaux qui sont perturbés lors de l'évolution d'une tumeur bénigne en tumeur cancéreuse. Avant de considérer les réseaux et les mécanismes impliqués dans le processus

de cancérisation, nous allons nous intéresser aux gènes. Les gènes associées au cancer ont été classés en deux catégories sur la base de leurs caractères cancérogènes ou protecteurs :

Les oncogènes

L'expression des oncogènes [3] favorise la survenue de cancers. Ces gènes commandent la synthèse de protéines (oncoprotéines) stimulant la division et déclenchent une prolifération désordonnée des cellules.

Plusieurs dizaines d'oncogènes ont été décrits, dont le gène *MYC* codant pour le facteur de transcription MYC qui régule l'expression d'environ 15 % des gènes. Soumis à une sur-expression, il stimule la prolifération des cellules [38].

Les gènes suppresseurs de tumeurs (*tumor suppressor genes*) (TSGs)

Les TSGs agissent en sens inverse des oncogènes. Ce sont des régulateurs négatifs de la prolifération cellulaire. Leur inactivation n'empêchant plus la prolifération cellulaire favorise donc la survenue des cancers. Certains TSGs sont spécifiques de certains cancers. Ainsi le gène *RB1* est impliqué dans le développement du rétinoblastome.

Les gènes *BRCA1* et *BRCA2* sont impliqués dans les cancers du sein [39], *APC* dans les cancers du colon [40], *WT1* dans les cancers du rein [41].

D'autres ont un spectre d'inactivation plus large comme *TP53* ou *CDKN2A* qui sont inactivés dans un grand nombre de types de cancer [42–48].

On peut également ajouter à ces deux catégories de gènes, une troisième facilitant le cancer :

Les gènes de réparation de l'ADN

L'ADN est continuellement soumis aux activités métaboliques intrinsèques à la cellule et à des facteurs environnementaux externes qui portent atteinte à son intégrité. Les facteurs environnementaux peuvent être de nature physique (*ie* rayonnements...), chimique (*ie* radicaux libres, médicaments...) ou biologique (*ie* toxines, virus...).

On estime entre mille et plus d'un million le nombre de lésions par cellule et par jour [49]. Beaucoup de ces lésions provoquent des dommages tels que la cellule elle-

3. du grec *onkos*, signifiant vrac, masse ou tumeur

même ne peut se reproduire ou donne naissance à des cellules-filles non viables. Les gènes de réparation de l'ADN sont capables de détecter et de réparer les lésions de l'ADN et prévenir cet état anormal.

Les mécanismes de réparation de l'ADN garantissent la stabilité du génome. La capacité de réparation de l'ADN d'une cellule est essentielle à l'intégrité de son génome, et donc à son fonctionnement normal et à celui de l'organisme.

Les différents génotypes possibles des cellules cancéreuses ne sont probablement la manifestation d'altérations que de quelques processus essentiels dans la physiologie cellulaire. Ces altérations seraient les caractéristiques spécifiques des cancers. Huit capacités essentielles ont été mises en évidence comme le détaille la Figure I.3 reprenant les publications Hanahan and Weinberg [50,51] :

L'autosuffisance en signaux de croissance

Une grande partie des oncogènes affectent le besoin en signaux de croissance dont les cellules saines sont dépendantes afin d'entrer dans un état de prolifération active. Un tel comportement dévie fortement dans les cellules tumorales qui montrent invariablement une dépendance fortement réduite aux signaux de croissance externes.

Nous pouvons donc en déduire que les cellules tumorales génèrent elles-même leurs propres signaux de croissance et réduisent ainsi leur dépendance vis à vis de la stimulation du micro-environnement qui les entoure. Cette indépendance perturbe totalement le mécanisme d'homéostasie qui gouverne le comportement des cellules dans un tissu. Dans le cas non tumoral, la plupart des signaux de croissance sont produits par un type cellulaire pour permettre la prolifération d'un autre.

Une mutation de *HRAS* par exemple, oncogène de la famille *RAS*, active constamment la protéine HRAS normalement activée par le facteur de croissance EGF et donc augmente la prolifération cellulaire. De par la nature proliférative du cancer, il est probable que les voies des signaux de croissance subissent une dérégulation dans toutes les tumeurs [50].

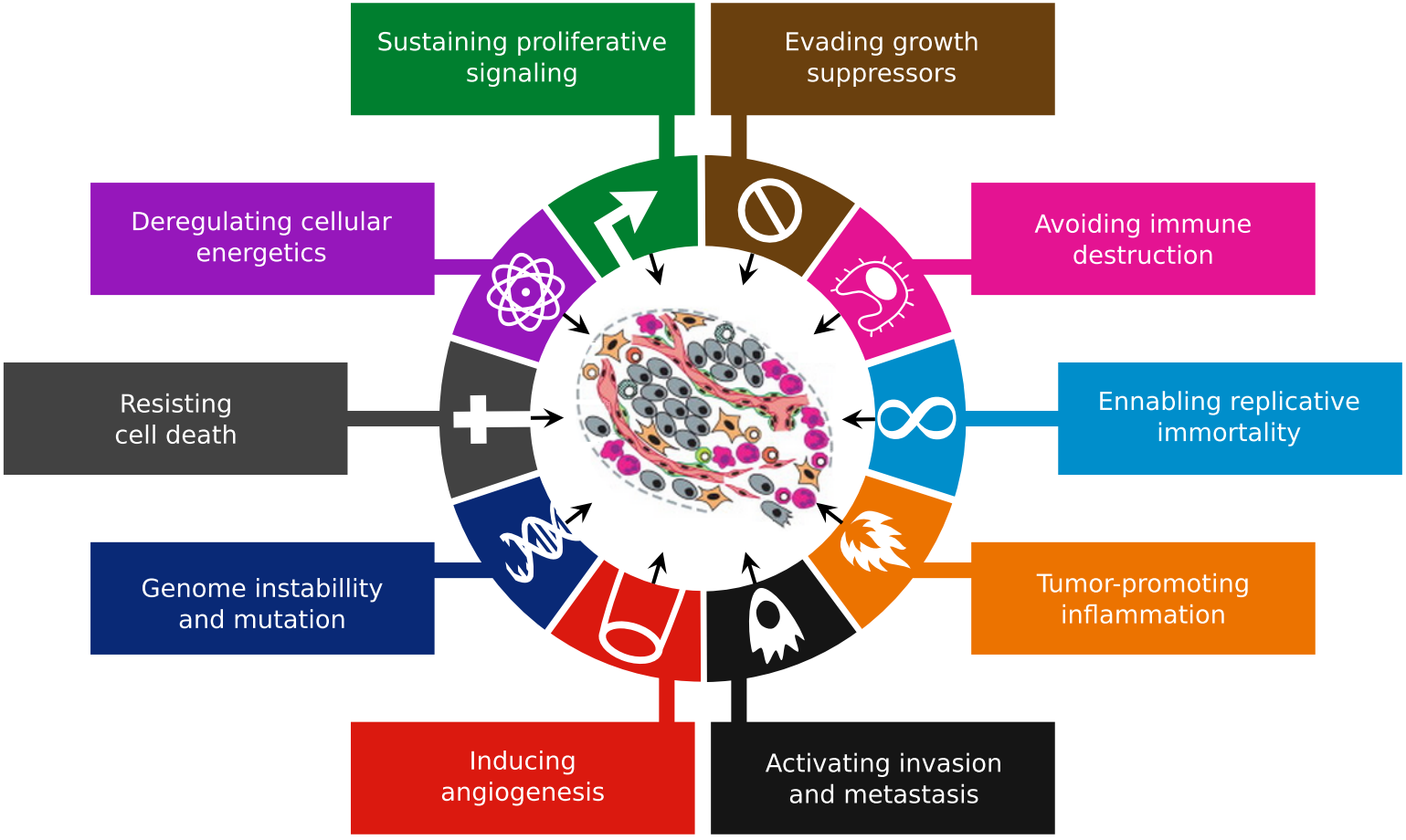


Figure I.3 – Caractéristiques du Cancer.

Figure inspirée de Hanahan and Weinberg, Hanahan and Weinberg.

L'insensibilité au signaux inhibiteurs de la croissance

En parallèle de l'autosuffisance en signaux de croissance, les cellules cancéreuses sont insensibles aux mécanismes d'inhibitions de la prolifération cellulaire. Beaucoup de ces mécanismes dépendent de l'actions de TSGs dont un grand nombre ont été découvert par leurs propriétés d'inactivation dans différentes formes de cancer chez l'homme ou chez l'animal. Nous pouvons citer *TP53* et *RB1* qui sont des *hubs* (gènes interagissent avec un grand nombre de gènes) dans des processus clés de la régulation cellulaire qui contrôle la prolifération et l'apoptose.

RB1 est un point de contrôle pour le cycle de la division cellulaire. Les cellules cancéreuses ayant une perte de *RB1* ne reçoivent plus les signaux inhibiteurs que celui-ci transmet. *TP53* est capable de bloquer la progression du cycle cellulaire, s'il y a des dérégulations dans les niveaux de signaux de croissance, glucose ou d'oxygénation, ou des dégâts trop importants sur le génome, le temps que la situation se normalise. *TP53* peut même déclencher l'apoptose si des dégâts irréparables affectent ces systèmes cellulaires[52–56].

La capacité à échapper à l'apoptose

L'apoptose, ou mort cellulaire programmée, est une des voies possibles de la mort cellulaire. Elle est nécessaire au développement et à la survie des organismes multicellulaires. L'apoptose a un rôle structurel et survient massivement lors du développement embryonnaire, l'exemple le plus parlant étant celui de la formation des doigts lors du développement de l'embryon. L'apoptose a également un rôle de protection de l'organisme et survient lorsqu'une cellule a accumulé trop de dégâts et qu'ils sont devenus irréparables.

Une cellule échappant à l'apoptose survivrait et se multiplierait en dépit d'anomalies génétiques qui peuvent être dommageables à l'organisme entier. Comme vu précédemment, *TP53* étant capable de déclencher l'apoptose, son inactivation implique logiquement la capacité de la cellule cancéreuse à éviter l'apoptose. L'apoptose étant en équilibre constant avec la prolifération cellulaire, l'acquisition de cette capacité induit forcément une prolifération d'autant plus accrue.

La capacité de se répliquer indéfiniment

À l'instar de la capacité à éviter l'apoptose, la capacité à se répliquer à l'infini est inhabituelle pour les cellules qui ne sont normalement autorisées qu'à un certain nombre de cycles de croissance et de division. Il est largement accepté que les cellules cancéreuses requièrent cette capacité pour que les tumeurs puissent se développer suffisamment.

Cette capacité est normalement limité par deux phénomènes barrières : la sénescence, fin de la capacité réplicative des cellules qui conduit à un état stable non-prolifératif, et l'apoptose qui conduit à la mort naturelle programmée des cellules. Les télomères protégeant la fin des chromosomes sont directement impliqués dans cette capacité de prolifération illimitée[57]. La longueur des télomères d'une cellule dirige le nombre de cycles de croissance et de division qu'elle peut subir.

La télomérase, ADN polymérase spécialisée dans l'ajout de segments répétés dans les télomères, pratiquement absente dans les cellules normales est exprimée à des niveaux significatifs dans les cellules cancéreuses. La présence d'une activité télomérase est corrélée avec la résistance à la sénescence et à l'apoptose. La suppression de cette activité entraîne un raccourcissement des télomères et l'activation d'une des deux barrières à la prolifération.

L'induction de l'angiogenèse

L'angiogenèse est le processus qui permet la création de nouveaux vaisseaux sanguins à partir de vaisseaux existants. C'est un processus physiologiquement normal dans le développement de l'embryon ou dans la cicatrisation d'une plaie, mais qui permet à la tumeur de récupérer dans la circulation sanguine l'oxygène et les autres éléments nécessaires à son développement.

Des facteurs de croissance de l'endothélium vasculaire comme *VEGFA* induisent la croissance de capillaires sanguins dans la tumeur[58]. C'est aussi une des étapes nécessaire à la libre circulation des cellules cancéreuses qui va mener à la formation de métastase.

La capacité à former des métastases

Cette capacité nécessite tout d'abord la possibilité d'avoir des cellules libres circulantes, que ce soit par angiogenèse ou par invasion. La perte d'adhésion aux autres cellules ou à la matrice extra-cellulaire par l'inactivation de *CDH1*, gène clé de l'adhésion cellule-cellule par la formation des jonctions adhérentes, est fréquemment observée[59].

Le processus d'invasion est généralement vu comme une première étape vers la métastase. Ce processus commence par l'invasion locale, puis par l'invasion des capillaires sanguins ou lymphatiques proches (intravasion). Il s'en suit la circulation de cellules cancéreuses dans les systèmes lymphatiques et hématogènes et le transfert de ces cellules dans les tissus distants (extravasion). Les métastases se forment ensuite, commençant par des micro-métastases, qui grossissent en tumeurs macroscopiques (colonisation)[60].

La capacité à éviter la destruction immunitaire

Le système immunitaire surveille constamment tous les tissus et organes de l'organisme. Cette surveillance immunitaire identifie et détruit les cellules anormales. Cette réaction immunitaire devient défaillante lorsqu'un cancer se développe, ainsi les patients immunodéprimés sont plus sujets aux cancers[61].

Les tumeurs croissent donc malgré cette surveillance immunitaire et réaction hostile. Divers mécanismes d'échappement sont utilisés : disparition des antigènes pour ne plus être perçues par le système immunitaire, sécrétion de substances immunosuppressives inhibant la réaction immunitaire, détournement de la réaction immunitaire pour produire des facteurs utiles à la tumeur[62].

La dérégulation énergétique de la cellule

Sous conditions aérobies, la glycolyse produit normalement du pyruvate dans le cytoplasme de la cellule, qui via le cycle de Krebs se transforme en CO_2 dans la mitochondrie. Sous conditions anaérobies, la mitochondrie ne peut métaboliser le pyruvate et seul la glycolyse a lieu.

La glycolyse seule a un bilan faible : 2 moles d'ATP sont produites par mole de glucose consommée. Le cycle de Krebs produit quand à lui un équivalent de 30 moles d'ATP par équivalent de mole de glucose consommée, soit un bilan énergétique beaucoup plus riche dans les conditions aérobies. Otto Warburg a observé que même en présence d'O_2 , les cellules cancéreuses modifient et limitent leur métabolisme à la seule glycolyse.

Cette modification du métabolisme peut sembler contre-productive, vu que les cellules cancéreuses doivent compenser pour une production quinze fois moindre d'ATP par rapport au cycle de Krebs. En effet, une plus grande consommation du glucose a été décrite dans de nombreux types tumoraux. Ce mode de fonctionnement a été associé avec l'activation d'oncogènes (*RAS*, *MYC*), et la mutation de TSGs (*TP53*)[(63)].

Des conditions hypoxiques peuvent également sur-activer les transporteurs du glucose et les multiples enzymes de la glycolyse et ainsi renforcer cette dépendance[(63)]. Il apparaît que l'oxygénation, est fluctuante dans les tumeurs, à la fois temporellement et régionalement, et ce probablement sous le résultat de l'instabilité et l'organisation chaotique des vaisseaux nouvellement formés par la tumeur[(64)]. Cette caractéristique de restructuration énergétique du métabolisme cellulaire est directement influencée par des gènes impliqués dans d'autres caractéristiques (*ie RAS* et *MYC* pour l'insuffisance en signaux de croissance, *TP53* pour l'insensibilité aux signaux inhibiteurs de la croissance et l'évitement de l'apoptose)

En plus de ces huit capacités essentielles, Hanahan and Weinberg en décrivent deux autres facilitant le cancer :

L'instabilité du génome et les mutations

Cette capacité découle directement de l'inactivation ou perturbation de l'activité des gènes de réparation de l'ADN précédemment décrits. Et si elle n'est pas la cause première du cancer, elle y contribue fortement.

L'inflammation favorisant les tumeurs

Il est reconnu que certaines tumeurs sont fortement infiltrées par des cellules du système immunitaire et qu'elles génèrent des conditions inflammatoires dans les tissus proches[65]. Une telle réponse immunitaire est généralement comprise comme une tentative du système immunitaire d'éliminer la tumeur.

Mais l'inflammation des tumeurs a également l'effet paradoxal de favoriser la tumorogenèse et la progression tumorale. Les cellules inflammatoires peuvent sécréter des substances chimiques, qui sont activement mutagènes pour les cellules cancéreuses proches, accélérant leur évolution vers des états plus dangereux[66].

Chacun de ces processus met en jeu des gènes différents. Chacun des gènes impliqués dans ces processus est intégré dans un réseau de gènes, une voie de métabolisme ou un réseau de régulation. Nous ne parlons plus alors d'un seul gène en jeu, mais d'un ensemble de gènes impliqués dans des processus. Tout en considérant les réseaux, il nous faut cependant considérer deux catégories de gènes :

Les gènes directeurs (*drivers genes*)

Historiquement, les projets de séquençage des tumeurs recherchaient les gènes qui étaient fréquemment mutés, et qui donc avait donc supposément un rôle dans l'oncogenèse. La difficulté était de différencier les mutations directives (*drivers*), qui déclenchent l'oncogenèse ou le phénotype cancéreux[67–69], des mutations accessoires (*passagères*) qui sont dues au cancer. Par extension, les *drivers genes* sont devenus les gènes porteurs de *drivers mutations*. Nous les considérons comme les gènes ayant une fonction clé dont les perturbations peuvent être la cause des cancers.

Les gènes passagers (*passengers genes*)

Par opposition, les gènes passagers sont donc les gènes qui ne sont pas à l'origine du cancer même s'ils peuvent y participer.

Dans le cas d'un réseau de régulation des gènes. Nous proposons l'hypothèse que les gènes directeurs en amont du réseau, qui subissent une légère perturbation peuvent déclencher de plus grandes perturbations dans les gènes en aval et donc le cancer.

Le cancer peut donc être causé par des oncogènes stimulant la division et déclenchant une prolifération, ou par une inactivation des TSGs ne la régulant plus. Nous avons également exploré les différentes capacités développées par les tumeurs.

Toutes ces capacités n'ont pas nécessairement besoin d'être acquises pour qu'une tumeur deviennent cancéreuse. Mais toutes participent au développement et à la gravité de la tumeur. Nous allons maintenant étudier ce que les technologies à haut-débit ont apporté à la recherche sur le cancer.

I.3 L'apport des technologies à haut-débit à la recherche sur le cancer

I.3.a L'ère post-génomique et la fin du paradigme *un gène, une maladie*

Le paradigme "un gène, une maladie", désignant le fait qu'un gène, et par conséquent ses modifications ou mutations, causerait une maladie avait déjà été bien mis à mal par la découverte des polymorphismes nucléotidiques (*single-nucleotide polymorphisms*) (SNPs) (variations très fréquentes d'une seule paire de bases du génome, entre individus d'une même espèce). Les projets internationaux de séquençage ont définitivement permis de l'enterrer.

Ainsi, le Projet Génome Humain commencé dans les années 1990 par un consortium international s'est étalé sur près de quinze ans et a permis le séquençage complet du génome humain [5]. L'évolution des technologies de séquençage en font des outils de plus en plus utilisés et ce, à des échelles de plus en plus grandes.

De ce fait, le Projet 1000 Génomes commencé début 2008 utilisant des nouvelles technologies plus rapides et moins coûteuses a séquencé en trois ans les génomes d'un millier de personnes appartenant à différents groupes ethniques [70].

De la même manière l'International Cancer Genome Consortium (ICGC) coordonne au niveau international un projet de séquençage à très grande échelle de plus de 25 000 tumeurs sur une cinquantaine de types ou de sous-types cancéreux différents [71]. Le transcriptome est

l'ensemble des ARNs issus de la transcription du génome. Contrairement au génome, qui est généralement fixe (à l'exception des mutations), le transcriptome peut varier grandement.

Une approche transcriptomique mesure le niveau d'expression des ARNm transcrits dans la cellule, ce qui reflète directement les gènes exprimés à un moment donné. Nous utilisons pour cela les puces à ADN, ou plus récemment, le séquençage d'ARN à haut débit RNA-seq.

Cette approche permet l'étude à grande échelle des régulations / dérégulations de gènes dans des conditions diverses. De ce fait, en une seule étude, il est théoriquement possible d'identifier les gènes différentiellement exprimés entre deux expériences portant sur deux conditions biologiques différentes.

Cette technique a été utilisée pour observer l'effet d'une drogue sur les (profils d'expression de gènes (*gene expression profiles*) (GEPs)), pour comparer les tissus sains et les tissus malades, pour étudier les réponses au traitement en comparant les patients traités et ceux non-traités.

De nombreuses maladies ont été étudiées par cette approche : Alzheimer : Ricciarelli et al.[72], le diabète : Kaestner et al.[73] ainsi que plusieurs formes de cancer (leucémie : Golub[74], cancer du colon : Li et al.[75], cancer du sein : Wang et al.[9]) et encore bien d'autres maladies : Munro and Perreau[76].

Dans le contexte du cancer, une maladie particulièrement hétérogène, l'utilisation des GEPs sont utilisés pour prédire la résistance au traitement[77], ou la rechute métastatique dans le cancer du sein[78]. Des études sur le micro-environnement tumoral ont permis de comprendre l'influence du système immunitaire sur la survie des patients[79].

Les puces à ADN font parti des technologies à haut-débit qui ont été introduites en biologie moléculaire à partir des années 1990. Découlant de la technique du southern blot, les puces à ADN permettent en une seule expérience la mesure des niveaux d'expression de plus 30 000 gènes, le tout sur une courte période de temps (de l'ordre de deux jours). Encore très utilisée, il est cependant difficile avec cette technologie d'estimer avec précision la quantification des GEPs[80–84].

Le RNA-seq est une technique de séquençage à haut-débit appliquée au transcriptome. Avec les capacités de couverture et de résolution du séquençage à haut-débit, le RNA-seq permet d'avoir des informations qualitatives, il est par conséquent de plus en plus utilisé et sera probablement amené à remplacer les puces à ADN[85–88].

Les technologies à haut-débit ont grandement augmenté notre connaissance du génome humain. Elles nous permettent également de l'explorer à grande échelle et à moindre frais. Pour reprendre l'exemple du séquençage du génome humain, ce qui a pris pratiquement 15 ans et plusieurs centaines de millions de dollars, actuellement se fait en quelques jours pour quelques milliers de dollars. Nous allons maintenant voir plus en détail ce que de telles évolutions et technologiques ont apporté en médecine et quelles ont été leurs applications.

I.3.b La médecine prédictive et la médecine personnalisée

LA MÉDECINE PRÉDICTIVE est la discipline qui prédit la probabilité de survenue d'une maladie et induit des mesures préventives soit pour la prévenir, soit pour diminuer au maximum son impact pour le patient. Les approches protéomiques même si elles permettent une détection précoce de la maladie, ne détectent généralement des biomarqueurs que parce que la maladie est déjà présente. Par conséquent les approches basées sur la génétique sont celles qui permettent le mieux de prévoir la maladie, et ainsi d'estimer les risques des dizaines d'années avant qu'une maladie apparaisse.

Une femme avec une mutation dans le gène *BRCA1* a ainsi un risque augmenté de 65 % de développer un cancer du sein[89]. Les individus plus susceptibles à une maladie peuvent donc suivre des traitements ou des conseils spécifiques pour améliorer leur hygiène de vie dans le but d'empêcher l'apparition de cette maladie[4].

La médecine personnalisée est la discipline qui attribue à chaque patient des soins spécifiques en se basant sur ses caractéristiques, son mode de vie ou son environnement. Les technologies à haut-débit permettent ainsi de traiter chaque patient en fonction de ses spécificités biologiques et génétiques. L'utilisation de la médecine personnalisée est une des

4. cf le récent cas très médiatisé d'Angelina Jolie

voies les plus prometteuses en cancérologie. Son but est d'améliorer l'efficacité des soins, d'éviter les traitements inutiles et d'améliorer la qualité de vie des patients.

Les technologies à haut-débit permettent de déterminer de façon précise les caractéristiques de chaque tumeur et d'analyser les mécanismes moléculaires en cause. Par conséquent cela permet de préciser le diagnostic, d'identifier les caractéristiques de la tumeur et de proposer si cela est possible une thérapie ciblée, ce qui permet d'avoir moins d'effets indésirables qu'avec les chimiothérapies actuelles.

On appelle biomarqueurs les marqueurs biologiques permettant de caractériser de manière spécifique un tissu, un type cellulaire ou un état anormal. En cancérologie ce sont généralement des molécules, des protéines ou des gènes sur-exprimés ou anormalement absents dans certains types de tumeurs. Un biomarqueur peut servir à évaluer la réponse au traitement ; c'est le cas du récepteur ERBB2, qui permet dans le cancer du sein de prédire la réponse à un traitement hormonal.

Les biomarqueurs peuvent également être utilisés pour choisir un thérapie ciblée. Ainsi les patientes sur-exprimant ERBB2 peuvent avoir de ce fait accès au Trastuzumab, réduisant de 50 % le risque de récidive [(90)]. Actuellement, 17 thérapies ciblées peuvent être prescrites en France pour différents types de cancer.

Les technologies permettent d'acquérir une meilleure connaissance des caractéristiques des tumeurs et de leur évolution. Les biomarqueurs permettent d'affiner au mieux le traitement suivant la caractérisation des facteurs de risques. Nous avons ainsi dans une seule approche combinées les caractéristiques de la médecine prédictive quant à l'évolution de la maladie et de la médecine personnalisée quant à l'adaptation du traitement au patient. Nous allons maintenant aborder la recherche sur le cancer du sein.

I.4 La recherche sur le cancer du sein

I.4.a Les caractéristiques du cancer du sein

LE CANCER DU SEIN est le plus mortel et le plus fréquent chez la femme. Il est en tête de la mortalité devant le cancer colo-rectal et le cancer du poumon comme le montre le Tableau I.2. Néanmoins, le taux de mortalité par cancer du sein chez la femme diminue en France depuis une quinzaine d'années.

Les taux de survie relative à 1, 3 et 5 ans sont respectivement de 97 %, 90 % et 85 %[91]. La survie à 5 ans varie avec le stade du cancer lors du diagnostic. Ils sont respectivement de 98.3 %, 83.5 % et 23.3 % pour les stades locaux, régionaux (envahissement ganglionnaire) et métastatique[92]. Il est donc important de détecter le plus précocement possible le cancer pour augmenter les chances de survie.

La majorité des cas de cancer du sein se trouve chez la femme, mais il existe également un cancer du sein chez l'homme[93,94]. Il est cependant environ 100 fois moins fréquent, mais pour cause de diagnostic généralement plus tardif il a souvent un taux de survie moins élevé. Chez la femme, le cancer du sein est souvent hormono-dépendant. Les facteurs augmentant les taux d'œstrogènes sont des facteurs de risque. Le nombre de cycles menstruels influençant directement les taux d'œstrogènes, les ménopauses tardives ou les pubertés précoces sont des facteurs de risque. Les cellules préalablement différenciées sont moins sensibles aux hormones, la grossesse protège ainsi le sein par différentiation des cellules mammaires. L'âge de la première grossesse est donc également un facteur de risque. Il est estimé qu'environ entre 5 et 10 % des cancers du sein peuvent avoir pour origines des prédispositions génétiques. Les gènes *BRCA1* et *BRCA2* sont reconnus responsables de la moitié des cancers ayant une origine génétique.

Nous allons maintenant explorer quels sont les diagnostics et traitements proposés pour traiter les cancers du sein.

Tableau I.2 – Effectif annuel moyen de décès et taux observé (standardisé monde) de mortalité des cancers pour la période 2004-2008.

Organe	*Homme*		*Femme*	
	Effectif	*TSM p. 100 000*	*Effectif*	*TSM p. 100 000*
Lèvre, cavité buccale et pharynx	3 334	7,1	730	1,2
Œsophage	3 157	6,2	718	0,9
Estomac	3 015	5,2	1 741	1,9
Côlon-rectum	8 759	14,4	7 767	8,3
Foie	5 429	9,9	1 914	2,2
Vésicule biliaire	505	0,8	749	0,8
Pancréas	4 307	7,9	4 012	4,7
Larynx	1 259	2,5	144	0,2
Poumon	21 881	42,3	6 195	9,9
Mésothéliome de la plèvre	583	1,0	236	0,3
Mélanome de la peau	828	1,7	703	1,1
Sein	**-**	**-**	**11 359**	**17,2**
Col de l'utérus	-	-	1 113	1,9
Corps de l'utérus	-	-	1 904	2,3
Ovaires	-	-	3 340	4,8
Prostate	9 012	12,6	-	-
Testicules	94	0,3	-	-
Vessie	3 535	5,6	1 149	1,1
Rein	2 470	4,3	1 264	1,5
Système nerveux central	1 678	3,8	1 313	2,4
Thyroïde	152	0,3	254	0,3
Lymphome malin non hodgkinien	2 236	3,9	1 987	2,2
Maladie de Hodgkin	167	0,4	115	0,2
Myélome multiple	1 367	2,2	1 325	1,4
Leucémies	2 931	5,1	2 412	2,9
Site indéfini ou non précisé	6 634	12,5	4 140	4,8
Autres cancers	4 845	8,6	3 772	4,5
Tous cancers	88 378	158,6	60 359	79,1

TSM : Taux standardisés à la population mondiale
Sources : Institut de Veille Sanitaire (InVS), Centre d'épidémiologie sur les causes médicales de décès (CépiDc) - Inserm, 2011 [91]

I.4.b Diagnostics et traitements

NOUS AVONS VU précédemment qu'une détection précoce augmentait grandement les chances de survie. Dans les années 1980, le dépistage systématique du cancer du sein par mammographie avait été prévu pour réduire fortement la mortalité liée à cette maladie.

Cependant ces mammographies détectent souvent des tumeurs qui n'auraient pas évoluées ou qui n'auraient pas eu besoin d'un traitement lourd. Le sur-diagnostic entraîne généralement un sur-traitement. Ainsi, les patientes sans ganglions à un stade précoce subissent une chimiothérapie adjuvante que l'on pourrait éviter dans 70 à 80 % des cas [(7)].

Il faut néanmoins noter que même si le sur-diagnostic existe, le dépistage est loin d'être inutile et permet d'identifier au plus tôt la tumeur et de la traiter quand sa taille est minimale dans le but d'avoir le meilleur pronostic possible. C'est pourquoi les tumeurs sont analysées cytologiquement ou histologiquement dans le but d'affiner le diagnostic pré-opératoire, et de prévoir le traitement optimal, qui repose sur quatre outils principaux :

La chirurgie

Elle consiste en l'ablation de la tumeur dans le cas de la tumorectomie, de l'ablation d'une partie du sein pour la segmentectomie, ou de l'ablation totale du sein dans le cas de la mastectomie. C'est l'étape indispensable du traitement du cancer du sein, les autres traitements visant à réduire le risque de rechute.

La chimiothérapie

C'est une injection de produits anti-cancéreux ciblant les cellules se divisant trop rapidement, soit en affectant la mitose soit la synthèse de l'ADN. Elle peut être qualifiée d'adjuvante lorsqu'elle suit la chirurgie, ou de néo-adjuvante si elle la précède. Elle permet de réduire le taux de mortalité et de rechute, mais a de nombreux inconvénients pour la patiente (fatigue générale, nausées, vomissement, chute des cheveux).

La radiothérapie

Elle permet de traiter loco-régionalement les cancers, en utilisant des radiations pour

détruire les cellules. L'irradiation a pour but de détruire toutes les cellules tumorales tout en épargnant les tissus sains périphériques. Les séances de radiothérapie sont de courte durée et les effets secondaires moindres que lors d'une chimiothérapie.

L'hormonothérapie

Les cancers du sein étant souvent hormono-dépendants, les tumeurs sont par conséquence souvent hormono-sensibles. Les traitements hormonaux consistent soit à bloquer les récepteurs hormonaux avec des anti-œstrogènes, soit à diminuer le taux d'œstrogènes présent dans le sang et donc la stimulation des récepteurs via des anti-aromatases.

En plus de ces traitements classiques, des thérapies ciblées permettent une action plus précise contre les cellules tumorales avec moins d'effets secondaires. Par exemple, le Trastuzumab, anticorps monoclonal, bloque le récepteur ERBB2 et est ainsi efficace pour les cancers du sein sur-exprimant *ERBB2*. Ces cancers étaient considérés de mauvais pronostic, mais avec un tel traitement le risque de rechute est réduit de moitié, et la mortalité est réduite d'un tiers [90].

Le Lapatinib, inhibiteur intracellulaire, bloque l'activité tyrosine kinase des récepteurs ERBB2 et EGFR [95,96]. Le Bévacizumab, anticorps monoclonal, se fixe sur le facteur de croissance VEGFA et bloque ainsi l'angiogenèse, mais il n'augmente pas le temps de survie, il est essentiellement utilisé sur des patientes ne sur-exprimant pas *ERBB2* en combinaison avec le Paclitaxel [97,98] ou le Docetaxel [99].

Nous venons de voir les différents traitements connus du cancer du sein. Cependant, ce sont les caractéristiques moléculaires de la tumeur à traiter qui vont guider le praticien vers le traitement le plus approprié. C'est pour cela que nous allons approfondir les différents types de classifications des cancers du sein.

I.4.c Les classifications utilisées dans le cancer du sein

LES CANCERS, comme nous venons de le voir ont plusieurs caractéristiques. L'ensemble de ces caractéristiques peuvent suggérer un traitement plus approprié, ainsi qu'un taux de survie ou une probabilité de rechute associé. Le système de classification Tumeur-Ganglion-Métastase (*Tumor-Node-Metastasis*) (TNM) est un des plus courant. Il prend en compte la taille de la tumeur, le nombre de ganglions lymphatiques touchés, et la métastase éventuelle.

Ces trois facteurs sont alors combinés pour obtenir 5 stades :

0 Carcinome canalaire in situ (les cellules sont localisées dans un canal galactophore et n'ont pas migré à l'extérieur) ou carcinome lobulaire in situ (les cellules sont localisées dans la membrane d'un lobule).

I La tumeur est inférieure à 2 cm, et le cancer ne s'est pas propagé aux ganglions.

II La tumeur fait plus de 2 centimètres (sans atteinte ganglionnaire), ou moins de 5 centimètres et le cancer s'est propagé à 1, 2 ou 3 ganglions.

III Le cancer s'est propagé aux ganglions lymphatiques, et peut-être aux tissus voisins du muscle ou de la peau.

IV Le cancer a produit des métastases dans d'autres parties du corps.

La classification Scarff-Bloom-Richardson[(2)], étendue par les critères de Nottingham[(3)] se base sur trois critères histologiques :

Le degré de différentiation architecturale

Ce paramètre évalue le pourcentage de conduits formés par la tumeur. Moins il y en a, plus la structure des tissus est désordonnée.

Le nombre de mitoses

Plus il y a un nombre important de mitoses (signe que les cellules se divisent activement), et plus le cancer est prolifératif.

L'importance du pléiomorphisme nucléaire

Ce paramètre détermine si les noyaux des cellules sont uniformes comme ceux des cellules épithéliales, ou si ils sont plus grands, plus sombres, ou irréguliers (pléomorphe).

Pour chacun de ces critères, un score entre 1 et 3 est attribué. Le score cumulé de ces trois critères permet alors de classer le cancer parmi 3 grades :

1 Cancer possédant des cellules bien différenciées, à croissance lente, avec des risques faibles de propagation.

2 Grade intermédiaire, où le cancer possède des cellules modérément différenciées.

3 Cancer possédant des cellules peu différenciées, d'évolution rapide avec des risques plus élevés de propagation.

Le but de ces classifications est de décrire les cancers du sein pour permettre le choix d'un traitement approprié en maximisant l'efficacité du traitement et les chances de survie et en minimisant la toxicité du traitement. Une connaissance plus précise de la tumeur permet donc d'affiner ce choix. Par conséquent, l'expression des protéines permet d'affiner ces classifications. Ainsi le dosage des 3 récepteurs hormonaux suivants est souvent effectué :

ESR1 récepteur des œstrogènes (souvent dénommé ER)

PGR récepteur de la progestérone (souvent dénommé PR)

ERBB2 récepteur de la famille des récepteurs EGFR (souvent dénommé HER2)

Les cancers possédant les récepteurs ESR1 sont généralement dénommés dans la littérature cancers ER+, ceux ne les possédant pas sont des cancers ER-. C'est la dénomination que nous utiliserons dans ce document.

Les cancers ER+ possédant des récepteurs ESR1 dépendent des œstrogènes pour leur croissance, des traitements bloquant les effets des œstrogènes tels que le Tamoxifen peuvent donc être utilisés. Ils sont généralement de bon pronostic.

Comme nous l'avons vu précédemment, ERBB2 est utilisé comme marqueur pour des thérapies ciblées, ce qui a permis de grandement améliorer le pronostic initialement mauvais des cancers sur-exprimant ce récepteur.

Les cancers ne possédant aucun de ces récepteurs sont appelés triple négatifs. Ce sont généralement des cancers agressifs de petits tailles, dont le pronostic reste le même quelque soit le nombre de ganglions envahis. Ils se distinguent singulièrement dans leur évolution et réponse aux divers traitements.

L'expression des gènes a permis d'affiner encore plus cette classification, et a conduis à la classification des cancers du sein en 5 sous-types moléculaires, sur la base de l'expression de 306 gènes [100,101] (cf Figure I.4) :

Basal-like

Cancers de haut grade, souvent triple négatifs, ils sont généralement agressifs et de mauvais pronostic. Ils ont été nommés ainsi car ils expriment de manière constitutive les gènes exprimés normalement dans les cellules basales du sein.

HER2+

Cancers initialement de mauvais pronostic, le Trastuzumab ou le Lapatinib permettent de l'améliorer. Ils ont été nommés ainsi, car ils sur-expriment les récepteurs ERBB2, souvent dénommés HER2.

Luminal A

Cancers ER+ ayant un grade faible. Ils ont été nommés ainsi par similarité d'expression des gènes avec les cellules luminales du sein.

Luminal B

Cancers ER+ ayant souvent un grade élevé. Comme le sous-type Luminal A, ces cancers ont été nommés ainsi par similarité d'expression des gènes avec les cellules luminales du sein.

Normal-like

Peu caractérisé, il est possible que ce sous-type soit un artefact du à une présence élevée de cellules du stroma. L'expression des gènes de ces cancers se rapproche des cellules normales du sein.

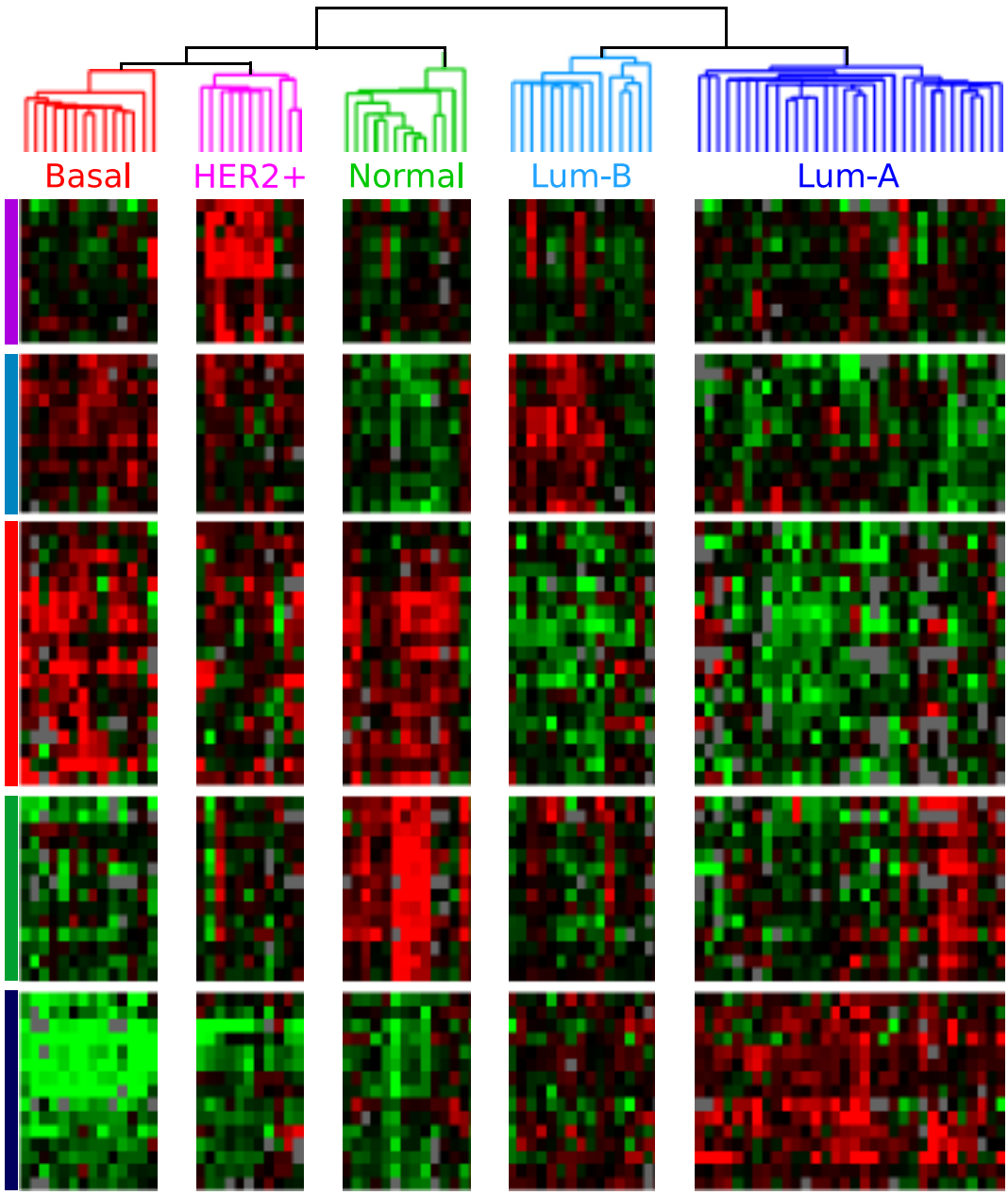


Figure I.4 – Classification en sous-types moléculaires.

Figure inspirée de Perou et al., Sørlie et al..

Ces sous-types correspondent à des niveaux d'expression de gènes différents qui reflètent la diversité et l'hétérogénéité des cancers du sein. Les progrès permis par la recherche et par l'amélioration des technologies permettent de réévaluer ces classifications et de les compléter. Ainsi nous pouvons décrire le sous-type Claudin-low, souvent triple-négatif, mais possédant une faible expression en protéines de jonction cellule-cellule et fréquemment infiltré par des lymphocytes[99,102].

Les classifications par stades, grades, statuts des récepteurs hormonaux ou sous-types moléculaire que nous venons de voir permettent de classifier les cancer dans des groupes ayant la plus grand similarité possible. Mais le cancer du sein est très complexe, et il peut y avoir des différences de taux de survie au sein d'un groupe. C'est pourquoi d'autres classifications peuvent être réalisées pour subdiviser ces sous-types et affiner ainsi la classification. Nous allons voir maintenant l'utilisation de puces à ADN pour classifier les tumeurs dans le cancer du sein en fonction de leur pronostic.

I.4.d Intérêts des signatures prédictives dans le cancer du sein

L'INTÉRÊT DE LA MÉDECINE PRÉDICTIVE est comme nous l'avons vu précédemment (cf Section I.3.b) de pouvoir prédire l'évolution de la maladie, et plus spécifiquement dans le cas qui nous concerne des cancers du sein. Les puces à ADN ont permis l'amélioration des pronostics avec la classification par sous-typage moléculaire qui permet déjà de séparer les patients en groupes ayant des survies différentes comme le montre la Figure I.5[100,101,103].

Cependant, s'ils permettent déjà d'améliorer le pronostic ces sous-types moléculaires sont encore trop hétérogènes, et ce pronostic ne reflète que trop peu la complexité des résultats cliniques.

De ce fait un certain pourcentage de patientes subissent une chimiothérapie qui pourrait être évitée[7] (cf Section I.4).

Deux études fondatrices dans le domaine des signatures prédictives en cancérologie ont établi des signatures liées à la métastase dans le cancer du sein. La signature MammaPrint

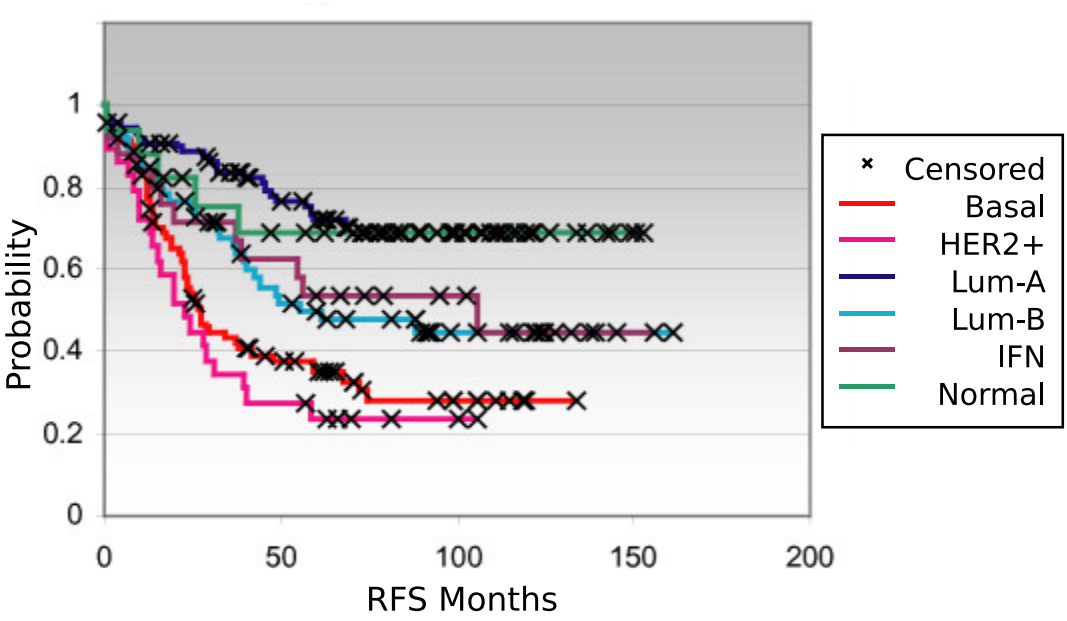


Figure I.5 – Courbes de survie en fonction des sous-types moléculaires.

Figure inspirée de Perou et al., Sørlie et al., Hu et al..

composée de 70 gènes permet de classifier les patientes en groupes de bons et mauvais pronostics[78]. De même Wang et al. ont réalisé une double signature de 76 gènes, spécifique aux statuts des récepteurs aux œstrogènes (*œstrogen receptors*) (ER), c'est à dire 60 gènes pour les ER+ et 16 gènes pour les ER-.

Le regroupement hiérarchique (*hierarchical clustering*) est utilisé pour la découverte de ces signatures. C'est une méthode de classification automatique dont le but est de répartir tous les individus constituant un ensemble dans un certain nombre de classe. Il est nécessaire d'avoir à disposition une mesure de similarité ou de dissimilarité permettant de différencier les individus. Deux approches existent, produisant une hiérarchisation qui donne son nom à cette méthode automatique de groupement :

L'approche ascendante part des individus où chacun d'eux appartiennent à une classe constituée d'un seul individu, puis ces classes sont regroupées en classe de plus en plus grandes, jusqu'à l'obtention d'une classe unique.

L'approche descendante part d'une seule classe que l'on divise de plus en plus pour arriver jusqu'aux individus.

L'approche ascendante est de complexité $O(n^3)$, tandis que l'approche descendante est de complexité $O(2^n)$. C'est pourquoi la variante descendante est généralement peu usitée. Ici, les individus sont les échantillons tumoraux, et la mesure de similarité se fait à partir de l'expression des gènes.

Nous allons détailler la façon dont Van De Vijver et al. et Wang et al. ont organisé leurs études pour la découverte de leurs signatures, ainsi que les caractéristiques des patientes constituant leurs jeux de données :

Van De Vijver et al. [78]

Un premier jeu de données constitué de 98 tumeurs a été utilisé. 78 tumeurs sont utilisées pour sélectionner 5000 gènes (qui sont significativement régulés dans au moins 3 tumeurs sur 78) parmi ceux présents sur la puce à ADN. Un regroupement hiérarchique est alors utilisé pour regrouper les tumeurs sur la base de ces 5000 gènes. 251 gènes sont désignés comme significativement associé avec l'évolution du cancer. Une sélection est ensuite effectuée sur le nombre de gènes suivant leur rang, et au final 70 gènes sont sélectionnés, puis validés sur 19 autres échantillons [8]. Cette signature à 70 gènes est finalement utilisée pour classifier un jeu de données contenant 295 tumeurs (dont une partie avait servi à établir la signature) [78]. Le jeu de données final contient 295 échantillons de tumeurs du sein, 151 sans ganglions et 144 avec. 110 patientes ont reçu un traitement adjuvant (soit 73 %) : 90 ont reçu une chimiothérapie, 20 une hormonothérapie et 20 une chimiothérapie combinée avec une hormonothérapie. 226 patientes ont un statut ER+, et 69 ER-.

Wang et al. [9]

Le jeu de donnée utilisé ici est constitué de 286 échantillons de tumeurs du sein sans ganglions, non traité par chimiothérapie. 248 patientes (soit 87 %) ont été traitées par radiothérapie adjuvante. Les tumeurs sont subdivisées en deux groupes suivants leur statuts ER. 209 sont ER+, et 77 ER-. Chacun de ces groupes est subdivisé en jeu de données d'entraînement et jeu de données de validation. Au final 115 échantillons

ont servi pour l'entraînement (80 ER+ et 35 ER-). 16 gènes sont sélectionnés par regroupement hiérarchique dans les ER+ et 60 dans les ER-. Les deux ensembles de gènes sont alors combinés pour constituer la signature.

Ces deux signatures n'ont que trois gènes en commun, ce qui peut mettre en doute leur fiabilité ainsi que leur reproductibilité. De plus Michiels et al. en réanalysant le jeu de données ayant permis d'établir la signature Mammaprint ont mis en évidence l'importance du jeu de données d'entraînement dans l'établissement d'une signature.

Nous allons donc maintenant voir quelles sont les limitations des techniques de regroupement hiérarchique se basant sur les puces à ADN pour réaliser des signatures prédictives dans le cancer du sein.

I.4.e Les limitations des technologies utilisées

LE BUT DE LA CLASSIFICATION est de fournir un bon modèle prédictif. D'un point de vue biologique, le fait que les signatures ne soient pas reproductibles d'une étude sur l'autre est inacceptable. Cela montre un manque de robustesse dans les méthodes de détection qui pourrait empêcher l'acceptation des technologies de puces à ADN pour des tests clinicaux de routine. Cela s'applique également aux nouvelles méthodes dites Séquençage de nouvelle génération (*Next Generation Sequencing*) (NGS).

Plusieurs raisons simples sont fréquemment citées pour expliquer cette situation :

— L'hétérogénéité des plateformes.

— L'hétérogénéité des différents outils d'analyses utilisés.

— La variabilité génétique inhérente à chaque individu.

— Les différentes méthodes statistiques et les classifications.

Mais ces raisons ne sont suffisantes et les limitations sont principalement de deux origines :

La topologie des données

Le nombre de patients pour une étude étant généralement de l'ordre de la centaine et le nombre de gènes, de l'ordre de la dizaine de milliers, le nombre de patients profilés

est très bas par rapport au nombre de variables. Il y a à la fois trop de variables et pas assez d'échantillons. C'est le double fléau de la dimension et de la parcimonie.

La biologie du cancer

Les puces à ADN sont énormément sensibles aux effets des petites dérégulations des gènes en amont des réseaux de régulation des gènes. Les cancers sont très hétérogènes et peuvent dériver de plusieurs caractéristiques variables. Les gènes directeurs en amont des réseaux de régulation des gènes peuvent causer de fortes perturbations qui sont alors très variables d'un cancer à un autre. Les gènes en aval sont alors fortement dérégulés, et facilement détectables. Mais les gènes en amont qui sont les vraies causes de la condition clinique étudiée et qui ont causé ces perturbations ne sont pas détectés.

Les puces à ADN, ainsi qu'une approche transcriptomique ont donc des limitations inhérentes à la technologie. Il y a également des limitations dues à la maladie étudiée. Nous allons maintenant étudier comment remédier à ces limitations.

I.4.f Les solutions

LES LIMITATIONS, comme nous avons pu le voir ont principalement deux causes. Pour contrebalancer la première limitation causée par la topologie des données, une solution possible est d'augmenter le nombre d'échantillons. D'un point de vue purement théorique, pour conserver une couverture équivalente à celle permise par 100 mesures dans un espace à une dimension, il faudrait 10^{20} mesures dans un espace à 10 dimension. Les puces à ADN mesurent l'expression des gènes, dont on estime le nombre entre 20 000 à 30 000. Ces mesures se réalisent donc dans l'espace des gènes.

Cependant, de nombreux gènes ont leurs expressions liées, ou font partie intégrante d'un réseau de régulation ou d'une voie métabolique. Le nombre de dimension de l'espace des gènes est donc plus réduit que le nombre de gènes, mais il reste néanmoins un espace possédant de multiples dimensions. Le nombre important de gènes entraîne également une augmentation du risque de sur-apprentissage, ce qui à pour effet d'accroître l'erreur de

généralisation.

Ein-Dor et al. estiment qu'il est nécessaire d'avoir plusieurs milliers d'échantillons pour l'obtention d'une signature constituée d'une liste robuste de gènes pour prédire l'évolution du cancer. Les jeux de données utilisés pour les études précédentes[9,78] contiennent au mieux plusieurs centaines d'échantillons. C'est une des raisons généralement avancé pour expliquer le manque de reproductibilité de ces études.

Pour contrebalancer cette limitation, nous proposons de réaliser une méta-analyse de plusieurs jeux de données, et ainsi augmenter le nombre d'échantillons.

Plusieurs méthodes de méta-analyses sont reportés dans la littérature :

— Le test de l'inverse-chi-2 de Fisher : Fisher[105]

— Le regroupement hiérarchique basé sur un test de Student : Gentleman et al.[106]

— La méthode des paires de haut score : Xu et al.[107]

— La méthode des produits de rang : Hong et al.[108]

— La mise en commun bayésienne : Conlon et al.[109]

— Le Binary Matrix Shuffling Filter : Zhang et al.[110]

Ainsi comme nous l'avons vu précédemment, la biologie du cancer elle-même limite aussi l'efficacité des puces à ADN pour ces prédictions. En effet les puces à ADN sont extrêmement sensibles aux effets des variations de l'expression des gènes en amont. Ces deux limitations se combinent, la biologie extrêmement hétérogène du cancer et ses effets sur l'expression des gènes, ainsi que la topologie des données, inhérente à la technologie utilisée.

L'autre solution, qui peut être combinée avec une méta analyse consiste à ajouter des informations supplémentaires (telle les modules (ensemble fonctionnel de gènes)[111] ou des réseaux d'interactions protéines-protéines[10]). Cette dernière méthode, mise en avant par Chuang et al. permet l'obtention de meilleurs résultats par rapport aux méthodes classiques comme nous le détaillerons dans le chapitre suivant.

Cependant, Chuang et al. n'ont utilisé qu'un seul jeu de données pour leur analyse, et nous avons significativement amélioré cette méthode en la combinant avec une approche de méta-analyse. Nous détaillerons notre utilisation de cette méthode d'intégration d'interac-

tion protéines-protéines, ainsi que la façon dont nous l'avons adapté dans nos travaux avec une méta-analyse sur plusieurs jeux de données transcriptomiques dans le chapitre qui suit.

I.5 Conclusion

Nous avons détaillé le rôle central de l'expression des gènes, et de sa régulation, dans le développement et la survie d'un être humain. Nous avons ensuite montré comment le cancer pouvait survenir si l'expression et/ou la régulation des gènes était perturbée. Les approches de médecines personnalisée et prédictive ont démontré l'intérêt de classifier les tumeurs pour pouvoir les traiter de manière spécifique et proposer une thérapie adaptée. L'approche transcriptomique à haut-débit, permettant d'étudier les régulations et dérégulations à grande échelle, est déjà utilisée pour étudier le cancer. Cependant cette approche a des limitations intrinsèques en plus des limitations dues à la maladie elle-même. Pour contrer ces limitations nous utilisons l'ajout d'informations d'interactions protéines-protéines aux données transcriptomiques ainsi qu'une méta-analyse sur plusieurs jeux de données transcriptomiques. Nous allons dans le prochain chapitre détailler cette méthode, puis exposer nos résultats dans les chapitres suivants.

CHAPITRE

II

MÉTHODES

Résumé

Nous détaillerons ici notre méthode d'Intégration Transcriptome Interactome. J'ai choisi d'inclure dans cette section nos chapitres *Linking Interactome to Disease*[12] et *Large Scale Transcriptome-Interactome Integration*[14] qui détaillent plus précisément la méthode. Ces chapitres étant trop longs, ils se trouvent dans les Annexes C.1 et C.3. Les données utilisées, ainsi que l'algorithme ITI seront décrit en détail, et les outils utilisés seront décrits brièvement.

Sommaire

II.1 Les avantages de l'intégration de données 43
II.2 L'Intégration Transcriptome-Interactome 45
II.3 Données transcriptome 47
II.4 Interactions protéine-protéine 51
II.5 Données et outils supplémentaires 52
II.6 Présentation de l'algorithme ITI 52
II.7 Conclusion 62

II.1 Les avantages de l'intégration de données

II.1.a Les avantages de l'intégration de données d'expression des gènes et d'interactions protéine-protéine

NOUS AVONS INTRODUIT dans le chapitre précédent les signatures prédictives dans le cancer du sein, ainsi que les limitations des technologies utilisées (cf Section I.4.e). Ces limitations sont dues d'une part à la topologie des données et du double fléau de la dimension et de la parcimonie. Et d'autres part, à la biologie du cancer et des gènes directeurs en amont qui non seulement sont à l'origine du cancer mais qui dérégulent les gènes en aval et ne sont de ce fait pas détectés. Nous allons maintenant voir quels sont les avantages de l'intégration de données sur les performances de ces signatures. Reprenant les travaux de Van De Vijver et al. et Wang et al. sur l'analyse des jeux de données ayant permis l'établissement de signatures, Chuang et al. rappellent les problèmes soulevés par ces précédentes études.

Les signatures de gènes permettant de prédire la rechute métastatique dans une étude sont moins efficaces, ne permettent pas suffisamment de généraliser, quand il s'agit de prédire la rechute métastique sur le jeu de données de l'autre étude[(11)]. De plus, seulement trois gènes sont communs entre ces deux signatures à 70 et 76 gènes. L'hypothèse des gènes directeurs, à l'origine du cancer, perturbant les gènes en aval est reprise ici pour expliquer ces différences entre ces deux ensembles de gènes. Pour circonvenir à ces inconvénients, l'utilisation de données d'interaction protéine-protéine permet de combiner les mesures d'expression des gènes issus de réseaux communs. Les biomarqueurs permettant d'établir une signature ne sont dans ce cas là plus les gènes ou les protéines, mais des sous-réseaux de protéines interagissant ensemble au sein du réseau des interactions protéines-protéines humain.

Cette méthode a des avantages par rapport aux analyses classiques :

— Les sous-réseaux résultants procurent des modèles des mécanismes moléculaires sous-jacents de la métastase.

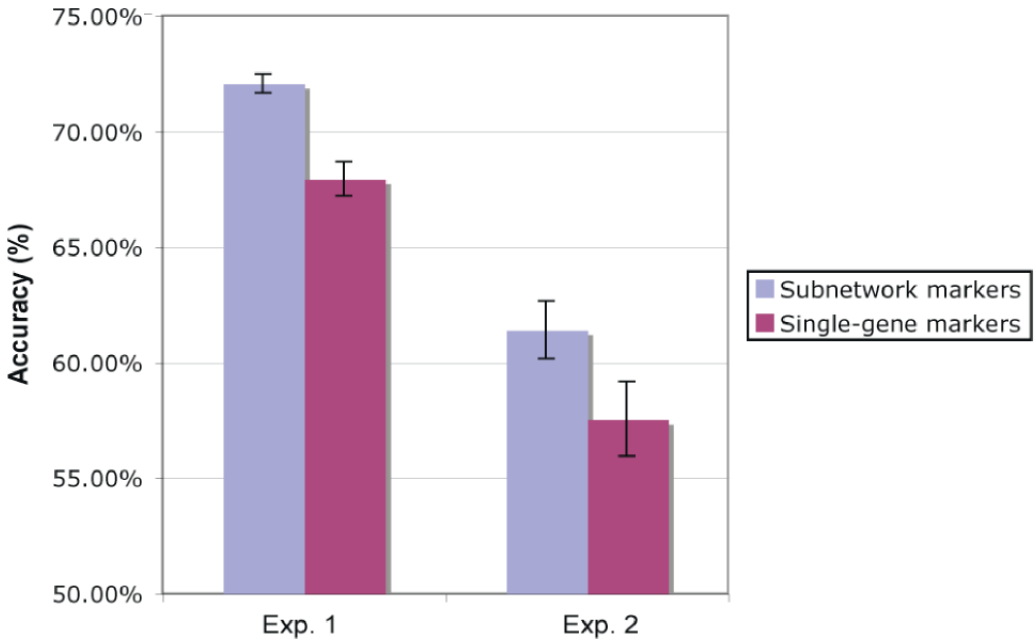


Figure II.1 – Avantages de l'intégration de données d'expression des gènes et d'interactions protéine-protéine sur la précision de la classification de la rechute métastatique par rapport à une analyse classique sur des données d'expression des gènes.

Figure inspirée de Chuang et al..

— Les *hubs* tels que *TP53*, *KRAS*, *HRAS*, *ERBB2*, non détectés par les analyses classiques, jouent un rôle central dans les réseaux en interconnectant un grand nombre de gènes.

— Les sous-réseaux identifiés sont significativement plus reproductibles entre différentes études que des biomarqueurs individuels sélectionnés sans information sur les interactions protéine-protéine.

— Une classification basée sur des réseaux permet d'obtenir une prédiction plus précise en sélectionnant les marqueurs sur un premier jeu de données d'entraînement et en les appliquant à un deuxième jeu de données de validation indépendant (cf Figure II.1).

II.1.b L'intégration de données d'expression des gènes et d'interactions protéine-protéine

COMME NOUS VENONS de le détailler, cette méthode présente des avantages comparé aux méthodes classiques d'analyse de données d'expression des gènes. Nous allons détailler rapidement ici la méthode de Chuang et al.. Cette méthode d'intégration utilise deux types de données. Le premier type est les données d'expression des gènes, ainsi que les données cliniques correspondantes, que nous détaillerons dans la Section II.3. Le second type de données utilisée est les données d'interactions protéine-protéine, que nous nous détaillerons dans la Section II.4. Les conditions cliniques des patients (*ie* métastatique ou non-métastatique) permettent de différencier l'expression des gènes constituant les sous-réseaux pour constituer une matrice d'activité. Elles sont utilisés pour assigner des ensembles de gènes sur des sous-réseaux. Cette matrice d'activité sert à attribuer un score global à chaque sous-réseau, dérivé de l'expression de chacun des gènes le constituant. Des sous-réseaux générés par permutation permettent alors de sélectionner les sous-réseaux discrimants. Les sous-réseaux ainsi sélectionnés sont utilisés pour identifier les gènes causant la maladie, et la matrice d'activité du sous-réseau est utilisé pour entraîner un classifieur.

II.2 L'Intégration Transcriptome-Interactome

II.2.a L'intégration massive de données d'expression des gènes et de données d'interactions protéine-protéine

L'INTÉGRATION MASSIVE DE DONNÉES d'expression des gènes et de données d'interactions protéine-protéine est la solution que nous avons développé pour circonvenir aux inconvénients des approches classiques des méthodes de prédiction de la rechute métastatique dans le cancer du sein.

Nous avons vu précédemment que les approches classiques manquaient de reproductibilité, dépendaient grandement des jeux d'apprentissage et étaient moins performantes utilisées sur un jeu de données indépendant.

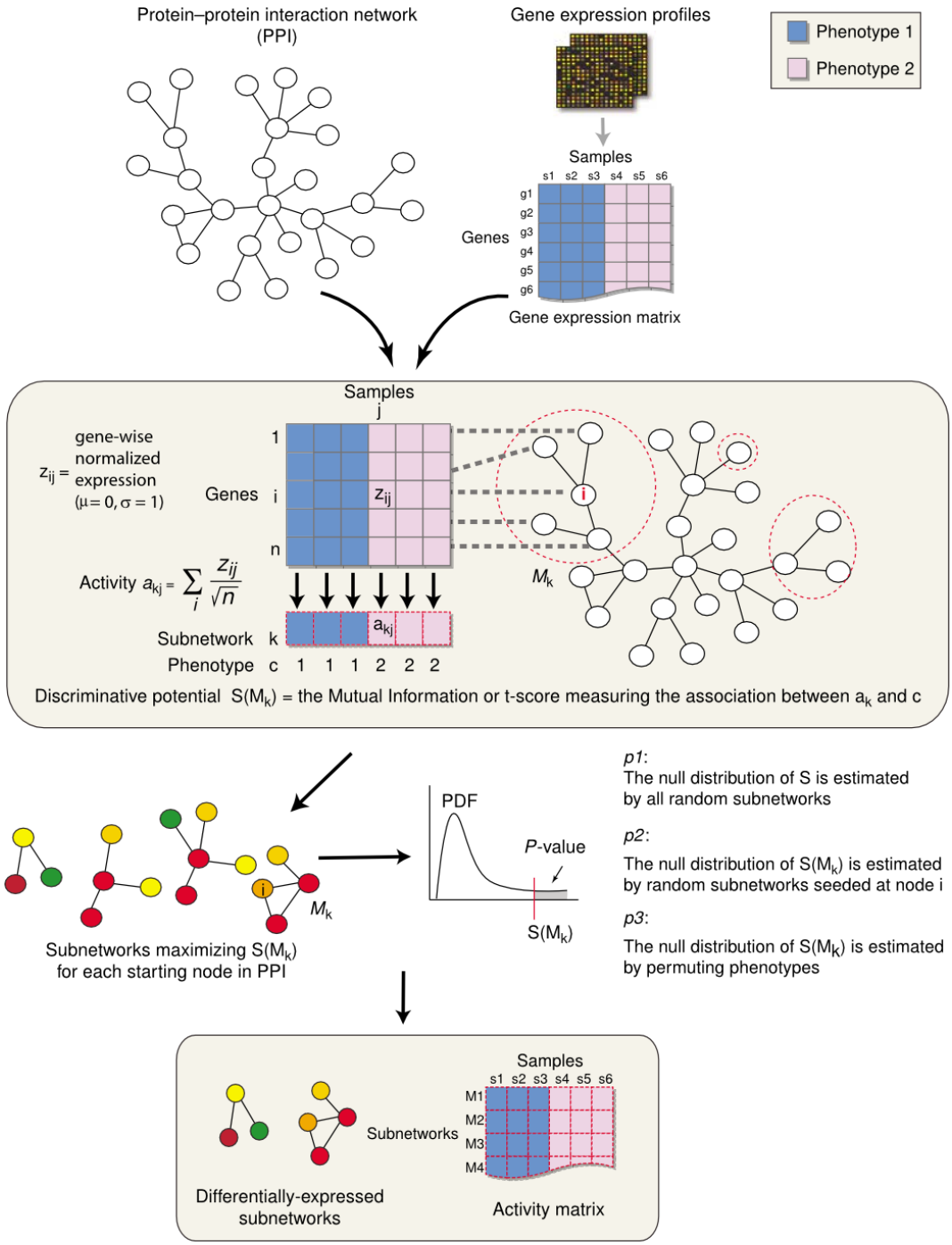


Figure II.2 – Algorithme détaillant l'intégration de données d'expression des gènes et d'interactions protéine-protéine.

Figure inspirée de Chuang et al..

Notre approche est une réimplémentation totale de l'algorithme développé par Chuang et al., avec la capacité supplémentaire de pouvoir prendre en compte plusieurs jeux de données d'expression des gènes pour réduire l'effet du fléau de la dimension par l'augmentation du nombre d'échantillons utilisés.

Nous allons d'abord détailler les données que nous avons utilisées tout au long de ces travaux. Ensuite, nous détaillerons l'algorithme.

II.3 Données transcriptome

II.3.a Constitution d'un compendium de données transcriptome dans le cancer du sein

POUR CONSTITUER UN COMPENDIUM de données d'expression, nous avons exploré les sites de dépôts de données publiques, ainsi que la littérature, et avons sélectionné les jeux de données dont les conditions cliniques étaient disponibles. Nous avons téléchargé l'ensemble des jeux de données sur le dépôt du Gene Expression Omnibus (REF), ou sur le site dédié de l'auteur. Si cela était possible nous avons téléchargé les données brutes, et avons réalisé une normalisation gcrma (package affy du package Bioconductor) sous R.

Nous avons utilisé ces jeux de données d'expression dans le cancer du sein pour deux analyses. Pour la première analyse, non supervisée, nous avons uniquement utilisé comme données cliniques l'événement DMFS. Voici les jeux de données que nous avons sélectionnés pour notre première analyse (cf Section II.7).

Tableau II.1 – Liste des jeux de données inclus dans notre compendium de données d'expression dans le cancer du sein.

Jeu de données	*Plateforme*	*Nombre d'échantillons*	*Présence de données cliniques*
Anders	U95v2	78	Non
Bild	U95v2	158	Non
Campone	UMGC-IRCNA 9k A	150	Non
Chang	cDNA array	50	Non
Chang-Kyu	Merck GEL Breast Tumor Profiles	311	Non
Chanrion	MLRG Human 21K V12.0	155	Non
Desmedt	U133A	198	Oui
Ivshina	U133 Plus 2.0	289	Oui
Jezequel	UMGC-IRCNA 9k A	252	Non
Kreike	NKI-AVL 18K cDNA	59	Oui
Loi	U133A + U133B	327	Oui
	U133 Plus 2.0	87	Oui
Miller	U133A + U133B	251	Oui
Parker	Agilent-011521 1A G4110A	2	Oui
	Agilent-012097 1A G4110B	27	Oui
	Agilent 1A Oligo UNC Custom	196	Oui
Pawitan	U133A + U133B	159	Oui
Perou	SCV	84	Non
Sabatier	U133 Plus 2.0	129	Oui
Schmidt	U133A	200	Oui
Sørlie		85	Non
Sotiriou	U133A	189	Oui
van de Vijver	Agilent whole human genome	295	Oui
van't Veer	Agilent whole human genome	117	Oui
Wang	U133A	286	Oui
Wong	U133A	6	Non
Yu	U133A	341	Non
Zhang	U133A	136	Oui
Zhou	U133Av2	54	Oui
Total	7 différentes	2572	

Tous les jeux de données présentés ici ont été considérés, cependant nous avons gardé seulement les jeux de données accompagnés de données cliniques. Les jeux de données rejetés pourraient cependant être inclus si les données cliniques étaient accessibles.

Tableau II.2 – Liste des jeux de données inclus pour notre analyse non supervisée (cf Section II.7).

Jeu de données		*Nombre d'échantillons*	*Nombre de statuts DMFS positif*	*Nombre de statuts DMFS négatif*
Desmedt	(112)	198	62	136
Ivshina	(113)	249	89	160
Loi	(114)	117	26	91
Parker	(115)	199	45	154
Pawitan	(116)	159	40	119
Schmidt	(117)	200	46	154
Sabatier	(118)	31	9	22
Sotiriou	(119)	179	40	139
van de Vijver	(78)	295	88	207
Wang	(9)	286	107	179
Zhang	(120)	136	20	116
Zhou	(121)	54	9	45
Total		2103	581	1522

L'utilisation de 12 jeux données nous donne l'accès à plus de 2000 tumeurs pour notre analyse non supervisée.

Tableau II.3 – Liste des jeux de données inclus pour notre analyse supervisée (cf Section III.3).

Jeu de données	*Plateforme*	*Nombre d'échantillons*	*Statuts DMFS*	*Statuts ER*
		(Sélectionnés/Total)	*(meta/non meta)*	*(ER-/ER+)*
Desmedt	U133A	190/198	62/128	61/129
Loi	U133A + U133B	101/327	27/74	29/72
Sabatier	U133 Plus 2.0	31/255	9/22	11/20
Schmidt	U133A	182/200	46/136	37/145
van de Vijver	Agilent whole human genome	150/295	56/94	36/114
Wang	U133A	276/286	107/169	72/204
Total	7 différentes	930/1561	307/623	246/684

L'utilisation de 6 jeux de données nous donne l'accès à plus de 1500 tumeurs pour notre analyse supervisée.

Pour notre seconde analyse, supervisée, nous avons choisi de ne sélectionner que les patientes sans traitement supplémentaire par soucis d'homogénéisation, et ainsi éviter de séparer les patientes en fonction de la réponse au traitement. Nous avons également tenu compte de l'expression des récepteurs aux œstrogènes (*œstrogen receptors*). Dans le but d'avoir un ensemble d'échantillons le plus homogène possible, nous les avons soigneusement choisi en nous basant sur les données cliniques accessibles. Les données cliniques qui nous intéressaient, étaient :

— Le statuts DMFS

— Le temps de mesure de ce statuts DMFS

— Le statuts ER

— La présence ou non de traitement, et sa nature éventuelle

Le statuts ER nous a permis de diviser les échantillons en deux groupes pour nos analyses. Nous avons utilisé le statuts DMFS et son temps de mesure, pour contrôler les échantillons et vérifier qu'il n'y avait pas eu des erreurs d'annotations. Les informations sur le traitement nous a permis de sélectionner uniquement les patientes sans traitement supplémentaire.

Après avoir détaillé les jeux de données, leur constitutions, nous allons étudier les interactions protéine-protéine.

II.4 Interactions protéine-protéine

II.4.a Assemblage d'interactomes humains

LA BIOLOGIE MOLÉCULAIRE décrit les différents constituants de la cellule (protéines, ADN, ARN et autres molécules). Mais un organisme vivant est une entité complexe, et il est difficile de le comprendre totalement en analysant des parties spécifiques. C'est pour cela que l'on envisage l'organisme comme un système ou un réseau d'interactions. Les protéines interagissent les unes avec les autres dans une cellule, et ces interactions donnent lieu à des fonctions biologiques et un comportement dynamique du système cellulaire. Généralement ces interactions protéine-protéine sont temporelles, spatiales, ou dépendantes d'une condition spécifique. L'un des plus grands enjeux dans l'ère post-génomique de la biologie est de récolter des informations d'interactions entre protéines, ADNs et autres petites molécules, et de comprendre comment ces interactions sont organisées. Les techniques à haut débit ont permis la génération d'un grand nombre de données d'interactions protéine-protéine. Pour ITI nous avons récupéré différentes bases de données d'interactions protéine-protéine.

II.4.b Nature des interactions et bases de données d'interactions utilisées

LES INTERACTIONS CONTENUES dans les bases de données d'interactions protéine-protéine sont récoltés par différentes techniques, et sont également de différente nature. Nous considérons comme sûres les interactions décrites dans la littérature et celle vérifiées par une technique de double hybride dans la levure. Les interactions de complexes par co-immuno-précipitation ne sont pas directes, mais concernent des protéines qui font parties d'un même complexe. Enfin, les interactions *in silico* sont des prédictions obtenues par divers algorithmes, et ne sont pas validées *in vivo* ou *in vitro*. Elles sont donc moins sûres que les autres interactions.

Human Protein Reference Database (HPRD)[122], INTAct[123], Database of Interacting Proteins (DIP)[124] et Molecular INTeraction database (MINT)[125] contiennent des interactions décrites dans la littérature et vérifiées par double hybride. Comprehensive Resource of Mammalian protein complexes (CORUM)[126] contient des interactions de complexes. Cocite[127] contient des interactions prédites *in silico*.

II.5 Données et outils supplémentaires

NOUS UTILISONS également des données supplémentaires pour les besoins de notre algorithme. Pour l'annotation des différentes plateformes de puces à ADN, nous utilisons les fichiers fournis par la plateforme d'annotation de puces à ADN Resourcerer[128]. Pour l'annotation des gènes, nous utilisons le fichier gene info [1] fourni par le NCBI [2].

Nous avons également utilisé des outils supplémentaires pour la visualisation et les analyses de nos résultats. Pour visualiser nos sous-réseaux, nous avons utilisé le logiciel libre *Graphviz* [3] développé par AT&T Labs Research[129] et le modèle de rendu *neato*. Pour analyser l'enrichissement en termes GO, nous avons utilisé le programme ErmineJ[130]. Pour notre analyse supervisée (cf Section III.3) nous utilisons la librarie libSVM[131] pour classifier nos échantillons (cf Section II.6.e).

II.6 Présentation de l'algorithme ITI

II.6.a Détection des sous-réseaux

LA PREMIÈRE ÉTAPE de notre algorithme est la détection de sous-réseaux. Les données utilisées par ITI en entrée sont d'une part des profils d'expressions des gènes, ainsi que les conditions cliniques correspondantes aux patients, et des données

1. Homo_sapiens.gene_info.gz
2. ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/GENE_INFO/Mammalia/
3. http://graphviz.org/

d'interactions protéine-protéine, détaillés précédemment (données d'expressions des gènes cf Section II.3, données d'interactions protéine-protéine cf Section II.4). Les données d'interactions protéine-protéine sont rassemblées pour ne former qu'un seul interactome. Les auto-interactions (une protéine vers elle) même n'ont pas été gardées. Nous avons gardés les interactions en nous basant sur l'unicité du numéro d'accession gene ID fourni par les fichiers d'annotations du NCBI. Les données d'expressions des gènes sont considérées séparément pour chacun des jeux de données utilisés. Chaque gène est utilisé successivement comme graine pour créer un sous-réseau. Pour accélérer cette étape de détection des sous-réseaux et minimiser les coûts de calcul, ce processus est parallélisé en subdivisant l'interactome, lors de la sélection des graines, sur un cluster de calcul de type Beowulf. Pour chacun des jeux de données d'expression des gènes, la corrélation des conditions cliniques avec les GEP est calculé. Ainsi, si un gène n'a pas d'expression dans un jeu de données particulier, à cause des différences entre les plateformes, il peut quand même être pris en compte pour la constitution d'un sous-réseaux. Récursivement nous considérons le premier voisin du gène graine, et l'ajoutons à notre sous-réseau en construction, si l'ajout de ce gène améliore le score du sous-réseau, suivant l'Équation II.1 :

$$S_{s,d} = \frac{\sqrt{n_d}}{\sqrt{\max n_d(DS)}} \left| corr\left(\frac{1}{n}\sum_{g \in S} e(g,d), cc(d)\right)\right| \tag{II.1}$$

$S_{s,d}$ est le score du sous-réseau s calculé sur le jeu de données d. d est un des jeux de données du compendium DS, de taille NS. $corr$ est la corrélation de Pearson mesurée entre la moyenne de l'expression des gènes $e(g,d)$ contenus dans le sous-réseau s et le vecteur cc contenant les conditions cliniques des patients du jeux de données d. Le score est pondéré par la racine carrée du nombre d'échantillons nd du jeu de données d divisé par le nombre maximum d'échantillons des jeux de données de DS.

Le nombre de gènes ajoutés au sous-réseau influence la corrélation entre la moyenne de l'expression des gènes du sous-réseau et les conditions cliniques, et donc le score du sous-réseau.

Au plus des sous-réseaux sont ajoutés, au moins l'apport d'un nouveau gène au sous-réseau influencera la valeur du score. C'est pourquoi une valeur de seuil nous sert ici pour sélectionner les gènes à ajouter, et ainsi éviter d'ajouter à chaque sous-réseaux la totalité des gènes de l'interactome.

Un score global, non utilisé pour le calcul, mais pour simplifier l'affichage des résultats est calculé suivant l'Équation II.2 :

$$S_s = \frac{1}{NS} \sum_{d \in DS} S(s, d) \tag{II.2}$$

S_s est le score global du sous-réseau s. La moyenne des scores est calculé en sommant pour chaque jeu de données s du compendium DS, le score $S_{s,d}$ du sous-réseau s sur le jeu de données s Et en divisant cette somme par NS, le nombre de jeux de données dans le compendium DS.

Reprenant cette méthode de construction des sous-réseau lors de cette détection, nous construisons, dans le but de valider statistiquement les sous-réseaux que nous venons de détecter, des distributions aléatoires de sous-réseaux.

II.6.b Validation statistique

DANS LE BUT DE DÉFINIR des distributions aléatoires de scores, nous permettant de vérifier l'hypothèse nulle, reliant l'expression des gènes et les réseaux d'interactions, nous utilisons trois méthodes pour générer des sous-réseaux aléatoires Ces méthodes se basent sur le principe de construction de sous-réseaux expliqué précédemment. Premièrement, nous mélangeons les conditions cliniques. Secondement, la décision d'ajout d'un gène à un sous-réseau ne dépend plus de la corrélation des conditions cliniques avec les GEP, mais est aléatoire. Troisièmement, nous mélangeons nos interactions protéines-protéines.

Ces trois méthodes différentes nous permettent de générer trois distributions aléatoires de scores qui vont servir pour valider statistiquement les véritables sous-réseaux détectés précédemment. Pour garder les ensembles des sous-réseaux aléatoires comparables avec les

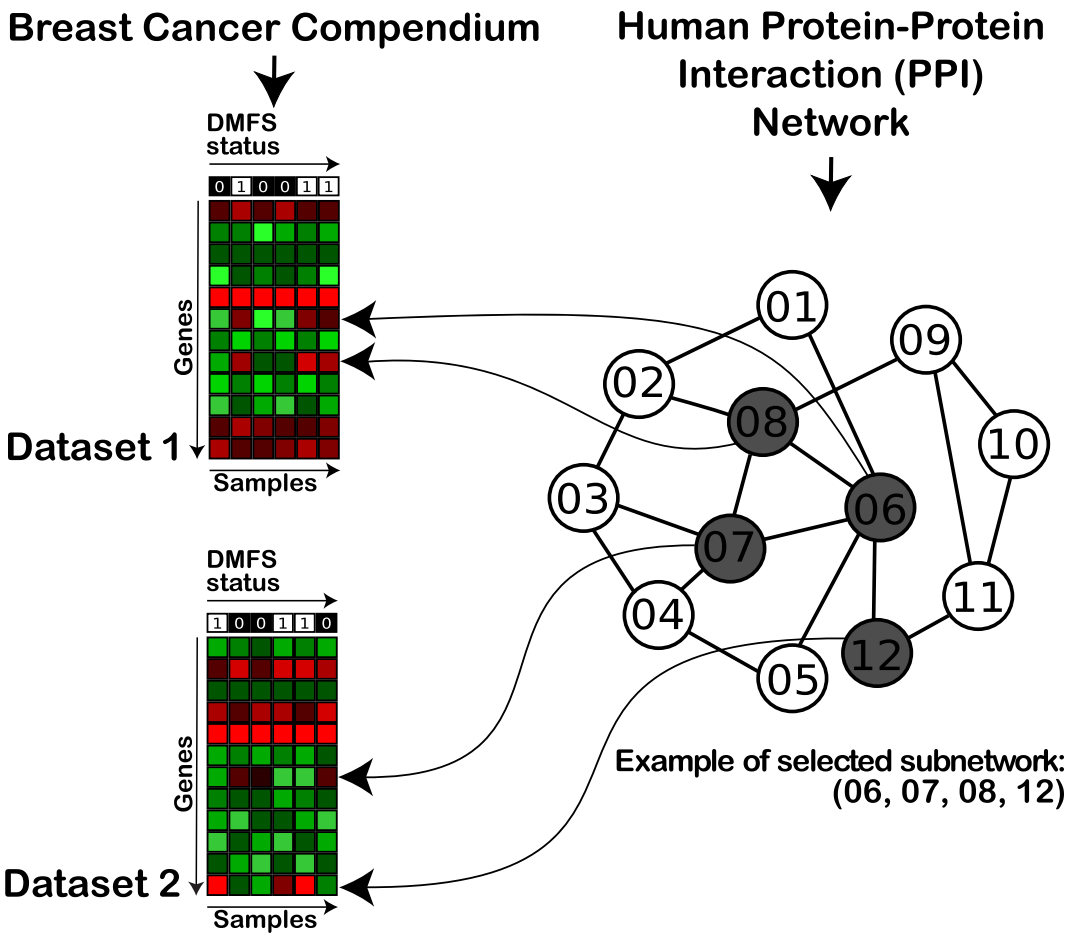


Figure II.3 – Principe de la sélection des sous-réseaux avec Intégration Transcriptome-Interactome.

Figure inspirée de Garcia et al..

sous-réseaux détectés, les distributions de leurs tailles sont forcés pour correspondre à celle des sous-réseaux détectés par modèle gaussien. Les distributions des scores des sous-réseaux aléatoires sont modélisés par mixture de deux distributions gaussiennes. Ces distributions sont utiliser pour fixer des seuils sur les scores, indépendamment des jeux de donnés d'expression, et ainsi leurs attribuer des p-values. Le mélange des interactions protéine-protéines ne permet pas de générer un nombre important de sous-réseaux, confirmant l'importance du lien entre les interactions protéines-protéines et le niveau d'expression des gènes. La validation statistique est réalisée avec Matlab Statistical Toolbox R2010b.

II.6.c Sélection de sous-réseaux

LES P-VALUES CALCULÉES lors de l'étape de validation statistique sont utilisées pour filtrer les sous-réseaux détectés statistiquement significatifs. Les distributions des trois ensembles de sous-réseaux aléatoires nous ont permis d'attribuer 3 p-values différentes pour chacun des sous-réseaux. Nous sélectionnons un seuil de p-value pour chacune des distributions aléatoires et l'utilisons pour filtrer les sous-réseaux détectés et générons trois ensembles de sous-réseaux statistiquement significatifs suivant chacune des méthodes précédemment expliquées. Nous réalisons alors l'intersection de ces trois ensemble pour constituer un ensemble de sous-réseaux dont la significativité est validé par trois distributions aléatoires. Pour faciliter l'affichage et l'interprétation des résultats, nous combinons les trois différentes p-values avec la méthode de Fisher [105]. Cet ensemble final de sous-réseaux constitue une signature avec laquelle nous pourrons classifier des échantillons et ainsi comparer la performance des signatures trouvées avec notre méthode ITI et les autres méthodes existantes. Nous détaillerons ces résultats dans la Section III.3. Nous allons maintenant détailler la réalisation d'une ressource bioinformatique contenant les différents sous-réseaux trouvés lors de nos analyses.

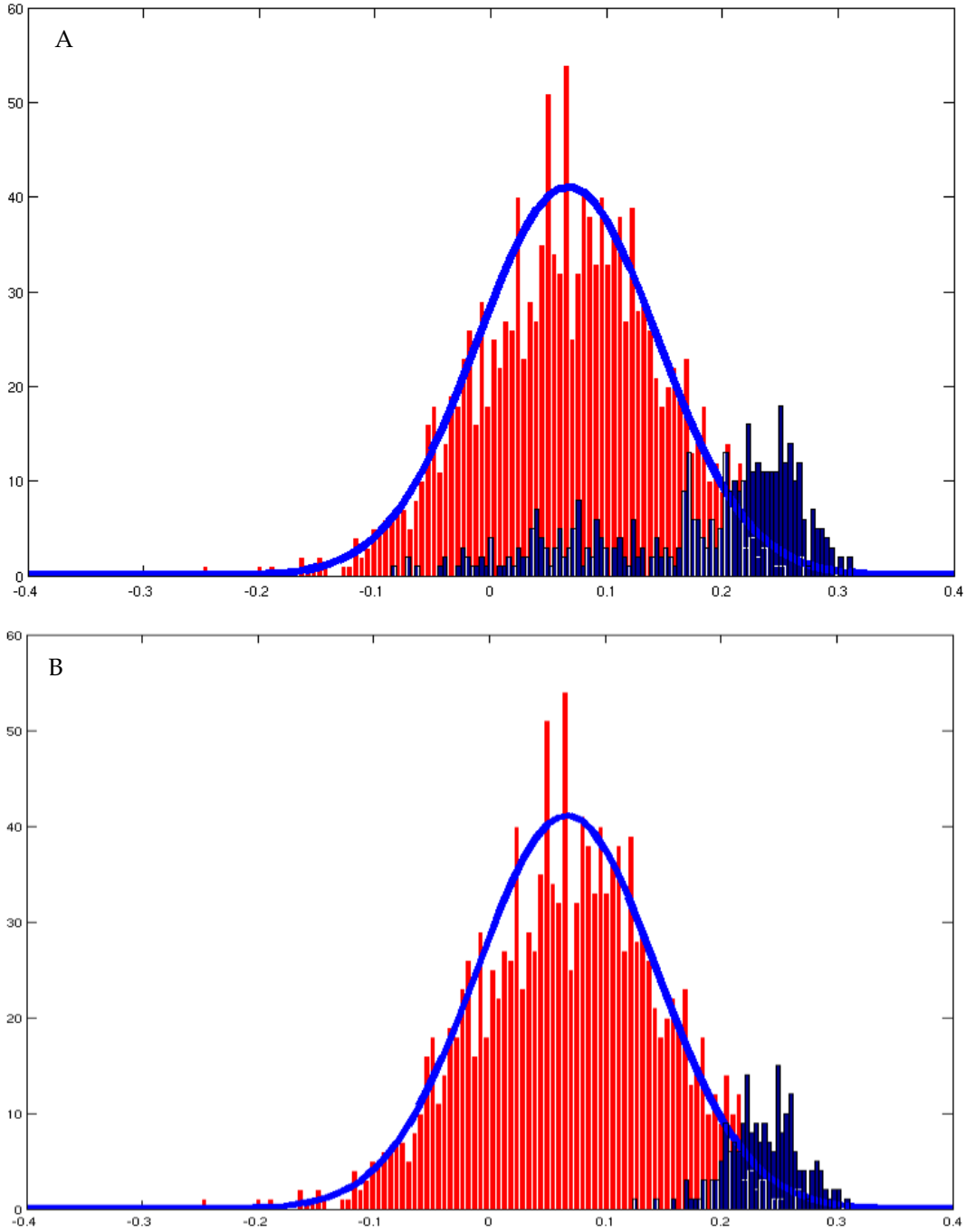


Figure II.4 – Distribution des scores des sous-réseaux pour le jeu de données Desmedt.

Histogramme rouge : distribution aléatoire de sous-réseaux. Courbe bleue : distribution normale (moyenne = 0,0669, σ = 0,0777. Histogramme bleu : distribution réelle des sous-réseaux (en A l'ensemble des sous-réseaux avant filtrage, en B les sous-réseaux obtenus après filtrage).

II.6.d Création d'une ressource bioinformatique permettant l'analyse des sous-réseaux et la reproductibilité de la recherche

POUR PERMETTRE LE PARCOURS des différents ensembles de sous-réseaux trouvés lors de nos analyses, nous avons créé une ressource bioinformatique permettant, l'affichage, l'interprétation et l'analyse des sous-réseaux. Cette ressource est accessible par internet, et constitue le site internet compagnon des publications [4]. Dans un soucis de reproductibilité des résultats, le code source du projet Intégration Transcriptome-Interactome (ITI) est disponible sous licence CeCILL sur sourceforge [5]. L'algorithme ITI nécessitant un cluster de calcul de type Beowulf pour fonctionner, nous avons choisi de l'intégrer à Mobyle, portail permettant de réaliser des analyses bioinformatique. Cette intégration permettra de faciliter l'usage d'ITI. De plus, les dépendances d'ITI à d'autre outils, comme Matlab nécessitant une licence propriétaire payante, ou ermineJ et Resourcerer, vont être remplacés par des scripts R ou perl, ce qui simplifiera à la fois le pipeline et son déploiement sur d'autres machines.

Les résultats de chaque analyse sont présents sur cette ressource bioinformatique. Pour chaque analyse nous avons accès aux sous-réseaux détectés statistiquement significatifs. Une page présente chacun des sous-réseaux. Nous avons utilisé le modèle de rendu neato du logiciel GraphViz pour permettre l'affichage des sous-réseaux. Une image en format png est générée par jeux de données d'expression pour chacun des sous-réseaux. Les scores et les p-values du sous-réseau en fonction du jeu de données d'expression permet de vérifier la significativité des sous-réseaux. Pour chacun des gènes du sous-réseau se trouve la valeur de la corrélation des conditions cliniques avec son expression pour chacun des jeux de données d'expression. Pour chacun des gènes, un lien vers la page du gène sur le site du NCBI est fournie, pour une analyse plus détaillée. Pour permettre plus facilement la navigation d'un sous-réseau à un autre, pour chaque gène il existe une page listant les sous-réseaux dans lesquels il apparaît.

4. http://iti.sourceforge.net/
5. http://sourceforge.net/projects/iti/

Enfin, pour chacun des sous-réseaux un enrichissement en termes GO est calculé, par jeux de données d'expression, se basant sur l'apport des gènes du sous-réseaux par rapport à l'ensemble des gènes présents sur la plateforme de puce à DNA utilisée pour constituer le jeu de données.

II.6.e Utilisation des SVMs pour la classification

Les machines à vecteurs de support (*Support Vector Machines*) (SVMs), ou Séparateurs à Vaste Marge, généralisation des classifieurs linéaires, sont un ensemble de techniques d'apprentissage supervisé destinées à résoudre des problèmes de discrimination et de régression. Ils reposent sur deux idées clés : la notion de marge maximale et la notion de fonction noyau. La marge est la distance entre la frontière de séparation et les échantillons les plus proches, appelés vecteurs de support. La frontière de séparation est choisie comme celle qui maximise la marge et ainsi minimise la capacité (complexité de la classification). Pour trouver cette frontière séparatrice optimale, à partir d'un ensemble d'apprentissage, le problème est reformulé comme un problème d'optimisation quadratique, pour lequel des solutions existent déjà.

La deuxième idée clé est la fonction noyau. Pour résoudre les problèmes de discrimination non-linéaire, l'espace de représentation des données d'entrées est transformé en un espace de plus grande dimension (potentiellement infinie), dans lequel il est probable qu'il existe une séparatrice linéaire. Une fonction noyau permet de réaliser cela, et a l'avantage de ne pas nécessiter la connaissance explicite de la transformation à appliquer pour le changement d'espace. Elle permet d'éviter la transformation coûteuse d'un produit scalaire dans un espace de grande dimension, en une simple évaluation ponctuelle d'une fonction.

Les SVMs permettent de traiter des problèmes de discrimination non-linéaire et sont capables de gérer des données de grandes dimensions, c'est pourquoi nous les utilisons pour classifier nos échantillons.

II.6.f Stratification, organisation des données et classification

POUR LES BESOINS DE NOTRE ANALYSE SUPERVISÉE (cf Section III.3), nous avons laissé de coté un jeu de données d'expression, dans le but d'effectuer une classification finale indépendante. Avec les autres jeux de données d'expression, nous avons réalisé une stratification à dix couches, utilisant 90% des données pour apprentissage, et les 10% restant pour classification.

Nous avons effectué notre stratification en sélectionnant les échantillons de manière à respecter la même proportion entre les différentes couches en échantillons des différents éléments des populations (événement DMFS négatif ou positif avec un statuts ER+ ou un statuts ER-).

Nous avons utilisé notre algorithme ITI pour sélectionner des sous-réseaux sur chacune de ces dix couches. Nous utilisons la librarie libSVM pour créer des modèles SVM pour chacun des jeux de données d'expression utilisés (cf Section II.6.e). Chaque liste de sous-réseaux sélectionnée après validation statistique a été utilisée pour trouver la taille maximisant les performances de la classification. Pour combiner les différents modèles SVM, nous avons réalisé un vote à la majorité pondéré par la taille de la population du jeux de données.

Chacun de ces ensembles de sous-réseaux a été utilisé pour classifier les 10% restant. Ces dix classifications nous ont permis d'arriver à un ensemble de sous-réseaux optimaux que nous avons alors regroupés. Chacun des sous-réseaux a été comparé avec les autres, et s'il y avait une superposition entre deux jeux de données, seul celui avec le meilleur score était gardé. Ce dernier ensemble de sous-réseaux a finalement été utilisé pour classifier le jeu de donnés d'expression préalablement mis de coté, et donc indépendant. La Figure II.5 détaille cette organisation des données. Les résultats seront exploités dans la Section III.3.

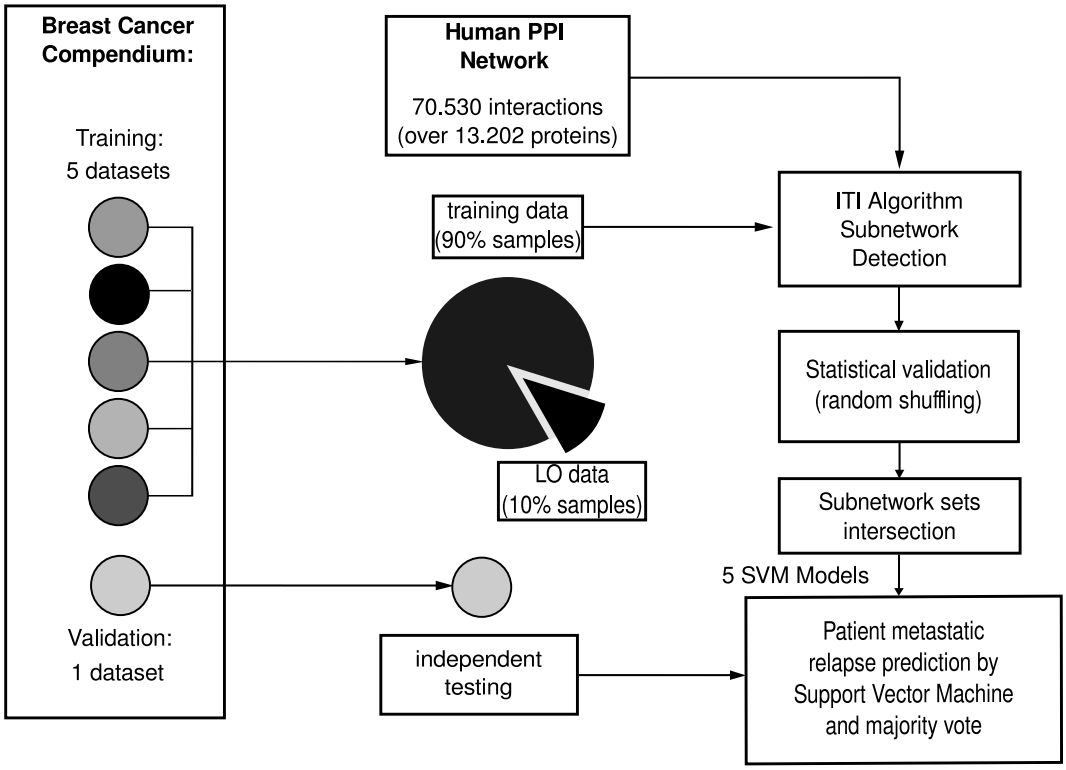


Figure II.5 – Workflow complet des données.

Figure inspirée de Garcia et al..

II.7 Conclusion

APRÈS AVOIR DÉTAILLÉ notre méthode, nous allons maintenant présenter les résultats que nous avons obtenus et publiés Garcia et al.[12,13]. Tout d'abord, nous allons détailler nos résultats sur une analyse non-supervisée (cf Section II.7), puis sur une analyse supervisée (cf Section III.3).

CHAPITRE

III

ANALYSE NON-SUPERVISÉE

Résumé

Dans ce chapitre nous allons présenter les résultats que nous avons obtenus lors de l'utilisation d'ITI sur une analyse non-supervisée. Ces résultats sont détaillés dans notre chapitre *Linking Interactome to Disease* Garcia et al.[12] présent dans les Annexes C.1 de cet ouvrage.

Sommaire

III.1 Détails de l'analyse non-supervisée **65**
III.2 Exploration des sous-réseaux **67**
III.3 Conclusion **70**

III.1 Détails de l'analyse non-supervisée

III.1.a Organisation des études

POUR COMPRENDRE L'IMPACT des différents jeux de donnés transcriptome sur la génération de sous-réseaux et les résultats, nous avons détecté des sous-réseaux suivant plusieurs combinaisons de nos jeux de données d'entraînement. Quatre combinaisons ont été réalisées (A1, A2, B1 et B2 cf Tableau III.1) à partir des jeux de données présentés précédemment (cf Tableau II.2) et correspondent aux quatre études que nous avons réalisées dans cette analyse. L'étude A1 contient une combinaison de tous les jeux de données transcriptome, sauf Van De Vijver et al.[78]. L'étude B1 contient une combinaison de tous les jeux de données transcriptome, sauf Wang et al.[9]. L'étude A2 contient une combinaison de tous les jeux de données utilisant une plateforme Affymetrix. L'étude B2 contient une combinaison de tous les jeux de données utilisant une plateforme Affymetrix, sauf Wang et al.[9].

Pour chacune de ces études, les sous-réseaux ont été validés en utilisant des p-values calculées suivant notre méthode de validation statistique (cf Section II.6.b). Les sous-réseaux obtenues dans l'étude A1 ont été validés par utilisation de p-values de seuils calculées à partir des trois méthodes pour générer des sous-réseaux aléatoires. Nous avons sélectionnés les sous-réseaux avec une p-value de seuil à 1.10^{-2} sur au moins 2 jeux de donnés transcriptome pour la première méthode, une p-value de seuil de 1.10^{-1} sur au moins 11 jeux de données pour la seconde méthode, et une p-value de seuil de 1.10^{-1} sur au moins un jeu de données pour la troisième méthode. Les autres p-values de seuil, et valeur de consensus choisies sont résumées dans le Tableau III.1. Les p-values peuvent sembler faibles mais ceci peut être justifié par le fait que l'analyse intégrée est réalisée sur plusieurs jeux de données simultanément. Il est alors possible de combiner ces p-values en une seule, plus significative (cf Section II.6.c).

Après avoir filtré les sous-réseaux lors de l'étape de validation statistique, ceux-ci sont stockés sur une ressource bioinformatique permettant l'exploration et l'analyse (cf Section II.6.d).

Tableau III.1 – Organisation de la validation croisée

Étude	*Jeu de données d'entraînement*
A1	Tous sauf van de Vijver
B1	Tous sauf Wang
A2	Tous ceux sur plateforme Affymetrix
B2	Tous ceux sur plateforme Affymetrix sauf Wang

Deux études sont réalisées à partir de tous les jeux de données de notre compendium sauf un (A1, B1), tandis que deux autres études sont réalisés à partir des jeux de données profilés sur plateforme Affymetrix pour comparer les performances de notre algorithme entre différentes plateformes.

Tableau III.2 – Seuils de p-value et valeur de consensus choisie

Étude	*Type 1*		*Type 2*		*Type 3*	
	p-value	consensus	p-value	consensus	p-value	consensus
A1	1.10^{-2}	2	1.10^{-1}	11	1.10^{-1}	1
B1	1.10^{-1}	8	1.10^{-2}	2	1.10^{-1}	1
A2	1.10^{-1}	11	1.10^{-2}	2	1.10^{-1}	2
B2	1.10^{-1}	6	1.10^{-2}	2	1.10^{-1}	1

Seuils de p-value et valeur de consensus choisie pour chacune des études représentant une différente combinaison des jeux de données transcriptome.

Tableau III.3 – Nombre de sous-réseaux découverts pour chacune des analyses

Étude	*Nombre de sous-réseaux*	*Nombre de gènes*
A1	119	406
B1	127	236
A2	103	306
B2	100	190

Nombre de sous-réseaux et nombre de gènes obtenus suivant les différents combinaison des jeux de données transcriptome.

L'examination des sous-réseaux obtenus, montre comme nous allons le voir dans la partie suivante peu de divergences parmi les sous-réseaux découverts. Mais nous allons commencer d'abord par une exploration biologique des sous-réseaux découverts.

III.2 Exploration des sous-réseaux

LA BIOLOGIE INTRINSÈQUE des 119 sous-réseaux extraits pour l'étude A1 a été analysée en utilisant les informations des annotations de la base de donnés EntrezGene du NCBI et de la base de donnés du Gene Ontology Consortium. Comme nous avons expliqué précédemment, la biologie intrinsèque des gènes inclus dans chacun des sous-réseaux a été calculée par enrichissement en termes GO II.5. Nous avons trouvés que les sous-réseaux formaient des complexes fonctionnelles supportant la maladie étudiée. Le métabolisme, le contrôle du cycle cellulaire, la prolifération, les adhésions cellules-cellules ainsi que la réponse immunitaire sont des mécanismes connus des différentes caractéristiques du cancer (cf Section I.2.c). L'exploration du sous-réseau s'effectue à l'aide de notre ressource (cf Section II.6.d), la Figure III.1 détaille la page permettant l'exploration du sous-réseau 387-4.

Le sous-réseau ayant le meilleur rang, 387-4 (score S=0.283), montre un enrichissement significatif pour "actin filament bundle formation" (GO :0051017), qui est un processus directement lié au développement de la cellule et à la polarité. Le sous-réseau 291-3 (score S=0.279) montre un enrichissement pour "activation of caspase activity by cytochrome C" (GO :0008635), qui est relié à l'apoptose. Ce sous-réseau montre aussi un enrichissement pour "B cell lineage commitment" (GO :002326), qui révèle une réponse immunitaire à la métastase. Le troisième sous-réseau, 2810-3 (score S=0.278), montre lui aussi un enrichissement pour une fonction similaire. Plus loin dans la liste, le sous-réseau 58-7 (score S=0.271) montre un enrichissement pour des fonctions reliées à la formation de micro-tubules : "microtubule organizing center organization" (GO :0031023) et "regulation of centrosome cycle" (GO :0046605). Le sous-réseau 29959-4 (score S=0.270) est relié au métabolisme avec un enrichissement pour les termes "glucose catabolic process" (GO :0006007), "fructose metabolic process" (GO :0006000) et "alditol metabolic process" (GO :0019400).

Subnetwork 387-4

score

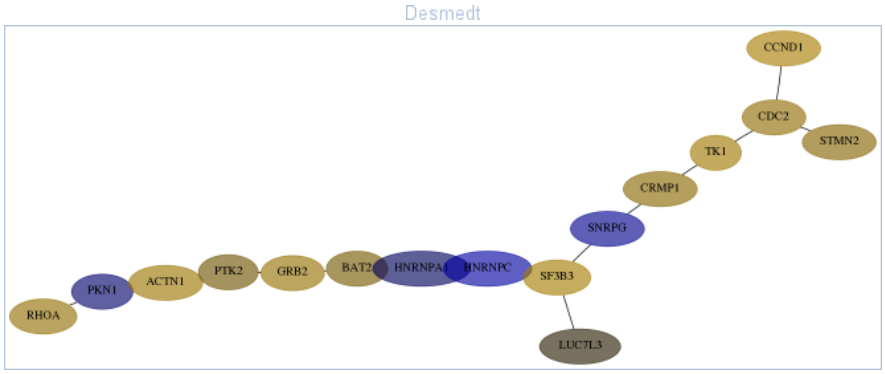
Dataset	Score	P-val 1	P-val 2	P-val 3
Desmedt	0.1892	1.058e-02	1.266e-02	1.034e-01
IPC-NIBC-129	0.2991	3.164e-02	3.902e-02	3.266e-01
Ivshina_GPL96-GPL97	0.2455	9.358e-02	1.153e-01	7.529e-01
Loi_GPL570	0.1361	3.124e-02	3.097e-02	1.782e-01
Loi_GPL96-GPL97	0.1380	0.000e+00	1.000e-06	0.000e+00
Parker_GPL1390	0.2913	1.142e-01	1.151e-01	5.759e-01
Parker_GPL887	0.3647	1.194e-01	1.427e-01	1.032e-01
Pawitan_GPL96-GPL97	0.3242	5.622e-02	3.746e-02	3.540e-01
Schmidt	0.3004	2.918e-03	4.385e-03	4.786e-02
Sotiriou	0.3898	1.961e-03	5.320e-04	5.222e-02
Wang	0.2473	1.573e-01	1.167e-01	4.856e-01
Zhang	0.3266	3.943e-02	2.148e-02	1.013e-01
Zhou	0.4275	8.611e-02	9.390e-02	4.960e-01

Expression data for subnetwork 387-4 in each dataset

Desmedt | IPC-NIBC-129 | Ivshina_GPL96-GPL97 | Loi_GPL570 | Loi_GPL96-GPL97 | Parker_GPL1390 | Parker_GPL887 | Pawitan_GPL96-GPL97 | Schmidt | Sotiriou | Wang | Zhang | Zhou |

Subnetwork structure each dataset

- Desmedt IPC-NIBC-129 Ivshina_GPL96-GPL97 Loi_GPL570 Loi_GPL96-GPL97 Parker_GPL1390 Parker_GPL887 Pawitan_GPL96-GPL97 Schmidt Sotiriou Wang Zhang Zhou



Score for each gene insubnetwork 387-4 in each dataset

Gene Symbol	Links	Frequency	Frequency Rank	Subnetwork score rank	Global rank	Desmedt	IPC-NIBC-129	Ivshina_GPL96-GPL97	Loi_GPL570	Loi_GPL96-GPL97	Parker_GPL1390	Parker_GPL887	Pawitan_GPL96-GPL97	Schmidt	Sotiriou	Wang	Zhang	Zhou
crnp1		2	104	1	18	0.065	0.155	-0.056	0.060	0.087	0.134	-0.257	0.009	0.138	0.086	-0.043	0.030	0.098
stxn2		29	8	1	3	0.057	0.106	0.005	0.052	0.110	0.071	-0.099	0.006	0.034	0.103	0.075	0.113	-0.039
rhoa		1	164	1	31	0.080	0.054	0.083	-0.107	0.073	0.065	-0.190	0.138	-0.177	0.077	-0.026	-0.032	0.160
bat2		1	164	1	31	0.039	0.182	-0.001	-0.013	0.062	undef	-0.335	0.109	0.186	0.099	-0.078	0.059	0.146
luc7l3		7	38	1	5	0.006	0.111	-0.046	-0.016	0.121	0.071	-0.318	-0.084	0.159	0.098	0.155	0.067	0.065
ccnd1		102	2	1	2	0.127	-0.183	0.062	0.190	0.157	-0.002	-0.190	0.105	-0.107	0.174	0.106	0.144	0.290
pkn1		1	164	1	31	-0.031	0.056	0.096	0.119	0.098	0.233	-0.317	-0.037	-0.013	0.127	-0.006	0.168	0.245
hnrnpa1		4	60	1	9	-0.021	0.209	0.088	-0.020	0.087	0.131	undef	0.082	0.036	0.089	0.077	0.083	0.195
sf3b3		7	38	1	5	0.119	0.241	0.143	-0.019	0.082	0.222	0.007	0.271	0.168	0.060	0.017	-0.077	0.372
tk1		6	43	1	7	0.109	0.149	0.190	-0.017	0.131	0.105	-0.100	0.257	0.204	0.142	0.064	0.164	0.222
cdc2		103	1	1	1	0.073	0.247	0.206	0.172	0.247	0.153	-0.131	0.282	0.226	0.221	0.067	0.297	0.336
snrpg		1	164	1	31	-0.068	0.217	0.120	0.108	0.092	0.134	-0.077	0.265	0.253	0.069	0.112	0.291	0.174
ptk2		1	164	1	31	0.035	0.139	-0.059	0.033	0.130	-0.066	-0.055	0.070	0.150	0.131	0.116	0.073	0.196
actn1		4	60	1	9	0.100	0.223	0.031	-0.052	0.053	0.156	0.031	0.019	0.024	0.027	0.124	-0.008	0.241
hnrnpc		1	164	1	31	-0.110	0.101	0.116	0.114	0.102	0.147	0.000	0.106	0.159	0.100	0.173	0.037	0.334
grb2		9	27	1	4	0.084	-0.013	0.086	-0.031	0.162	-0.038	0.003	0.065	0.158	0.127	-0.097	0.194	-0.005

GO Enrichment output for subnetwork 387-4 in each dataset

Desmedt file

Name	Accession Number	Link	P-val	Corrected P-val
actin filament bundle formation	GO:0051017	GO	2.465E-06	5.669E-03
hyperosmotic response	GO:0006972	GO	3.696E-05	0.04250404
positive regulation of cyclin-dependent protein kinase activity	GO:0045737	GO	5.169E-05	0.03963112
positive regulation of NF-kappaB import into nucleus	GO:0042346	GO	5.169E-05	0.02972334
positive regulation of transcription factor import into nucleus	GO:0042993	GO	1.104E-04	0.05080294
actin filament organization	GO:0007015	GO	1.126E-04	0.04318098
cell-matrix adhesion	GO:0007160	GO	1.244E-04	0.04088774
focal adhesion formation	GO:0048041	GO	1.348E-04	0.03876934
cell-substrate adhesion	GO:0031589	GO	1.574E-04	0.0402185
positive regulation of protein import into nucleus	GO:0042307	GO	1.909E-04	0.04389846
regulation of NF-kappaB import into nucleus	GO:0042345	GO	1.909E-04	0.03990769

Figure III.1 – Exploration fonctionnelle du sous-réseau 387-4

Figure inspirée de Garcia et al..

Ce sous-réseau est aussi fonctionnellement impliqué dans la migration cellulaire par la formation de structure cellulaire telles que des lamellipodia ou des filopodia (terme GO "substrate-bound celle migration, cell extension" GO :0006930).

Le sous-réseau 581-7 (score S=0.267) est impliqué dans l'adhésion cellulaire avec un enrichissement pour le terme "focal adhesion formation" (GO :0048041). La différentiation cellulaire est aussi fonctionnellement représentée, le sous-réseau 1452-1 (score S=0.247) montre une impliquation des gènes impliqués dans la voie *WNT* montre un enrichissement pour le terme "regulation of Wnt receptor signaling pathway" (GO :0030111). Le sous-réseau impliqué dans la prolifération cellulaire est 5155-5 (score S= 0.254), et il présente un enrichissement pour les termes "positive regulation of endothelial cell proliferation" (GO :0001938) et "establishment or maintenance of epitelial cell apical/basal polarity" (GO :0045197). La liste globale des enrichissements des termes GO est également stockée dans notre ressource (cf Section II.6.d).

Au niveau des gènes, plusieurs marqueurs montrent des liens évidents vers le cancer et des implications dans le cycle cellulaire, la proliferation, l'adhésion cellulaire et d'autres mécanismes biologiques impliqués dans cette maladie. La protéine codé par *CDK1* est une sous-unité catalytique du complexe MPF, qui est essentiel pour les phases de transitions G1/S et G2/M du cycle cellulaire eucaryote. D'autres gènes impliqués dans ce processus sont également trouvés, comme *CCND1*. *GRB2* est un facteur de croissance associés à plusieurs types de cancer, et pourrait avoir un rôle dans la métastase Yu et al.[132]. *TK1* est connu pour être un marqueur de la prolifération dans le cancer du sein, et sa sur-expression a été lié aux cancers de la thyroïde. *TSC1* est connu pour jouer un rôle central dans la régulation de la survie cellulaire et signaux de prolifération. D'autres gènes d'intérêt ont été également trouvés, dont *LAMA4* qui a des rôles *in vitro* dans la migration et *in vivo* dans la tumorogénicité des cellules cancéreuses de la prostate, et *PGK1* qui a été relié au cancer de la prostate.

III.3 Conclusion

L'ANALYSE non-supervisée réalisée avec ITI nous a permis d'explorer la biologie des sous-réseaux découverts. Cela nous a permis de confirmer l'utilité d'une telle approche en reliant directement des sous-réseaux à des processus cellulaires caractéristiques de la maladie étudiée. Les sous-réseaux découverts sont de plus disponible pour permettre d'identifier des gènes d'intérêt, qu'ils soient non reliés précédemment avec le cancer du sein, et donc des oncogènes ou des TSGs putatifs ou des cibles thérapeutiques potentielles. Nous allons maintenant exposer les résultats obtenus lors de notre analyse supervisée, et notamment les performances lors de la classification.

CHAPITRE

IV

ANALYSE SUPERVISÉE

Résumé

Dans ce chapitre nous allons présenter les résultats que nous avons obtenus lors de l'utilisation d'ITI sur une analyse supervisée. Ces résultats sont détaillés dans notre article *Interactome–transcriptome integration* Garcia et al.[13] présent dans les Annexes C.2 de cet ouvrage.

Sommaire

IV.1 Détails de l'analyse supervisée **73**
IV.2 Performances des signatures obtenues sur la prédiction de la rechute métastatique **73**
IV.3 Exploration des sous-réseaux **78**
IV.4 Conclusion **81**

IV.1 Détails de l'analyse supervisée

IV.1.a Organisation des jeux de données transcriptome en deux études

NOUS AVONS ORGANISÉ notre analyse supervisée en deux études. Ces deux études se basent sur le même compendium de donnés transcriptomique (cf Tableau II.3), en mettant de coté, pour validation indépendante, un jeu de données différent à chaque fois (Desmedt ou van de Vijver, cf Tableau IV.1). Pour chacune de ces deux études nous avons de plus séparé les échantillons suivant leurs statuts ER pour permettre une plus grande homogénéité des données. Pour éviter un sur-entraînement lors de la détection des sous-réseaux, nous avons organisé une stratification à 10 couches. La stratification a été réalisé en conservant entre les différentes populations des couches de la stratification la même proportion en individus, se basant sur les statuts ER et les conditions cliniques (cf Section II.6.f).

La détection des sous-réseaux a été ensuite réalisée comme présenté dans la Section II.6 et a mené à la détection de 165 sous-réseaux (pour les échantillons ER-) et de 6 sous-réseaux (pour les échantillons ER+) pour la première étude, et de 122 sous-réseaux (pour les échantillons ER-) et de 14 sous-réseaux (pour les échantillons ER+) pour la seconde étude. Le nombre de sous-réseaux est moins important pour les échantillons ER+, cela reflète une plus grande homogénéité des échantillons. Ces résultats sont détaillés dans le Tableau IV.2.

Nous allons maintenant passer à l'analyse des performances de signatures obtenues avec ces différentes études sur la prédiction de la rechute métastatique.

IV.2 Performances des signatures obtenues sur la prédiction de la rechute métastatique

DEUX SIGNATURES SÉPARÉES ont été générées pour les sous-types ER+ et ER- pour nos deux différentes études. Nous avons donc obtenus au final quatre ensembles de sous-réseaux. La taille optimale retenue dans le Tableau IV.2 est celle qui

Tableau IV.1 – Liste des jeux de données utilisés dans l'analyse supervisée.

Jeu de données	*Plateforme*	*Nombre d'échantillons*	*Statuts DMFS*	*Statuts ER*
		(Sélectionnés/Total)	*(meta/non meta)*	*(ER-/ER+)*
Desmedt	**U133A**	**190/198**	**62/128**	**61/129**
Loi	U133A + U133B	101/327	27/74	29/72
Sabatier	U133 Plus 2.0	31/255	9/22	11/20
Schmidt	U133A	182/200	46/136	37/145
van de Vijver	**Agilent whole human genome**	**150/295**	**56/94**	**36/114**
Wang	U133A	276/286	107/169	72/204
Total	7 différentes	930/1561	307/623	246/684

Deux études ont été réalisées en utilisant différentes combinaisons pour les jeux de donnés d'entraînement et ceux de validation (**en gras**). Dans l'étude 1 les échantillons provenant de Desmedt ont été mis de côté pour validation indépendante, et l'entraînement a eu lieu avec les autres jeux de données. Les échantillons provenant de van de Vijver ont été pareillement mis de côté pour validation indépendante dans l'étude 2.

Tableau IV.2 – Taille et p-value de la signature retenue pour chacune des études réalisées.

Étude	*Statuts*	*seuil de P-value*	*Nombre de sous-réseaux*	*Nombre de gènes*
Étude 1	ER-	$1e^{-4}$	165	2310
Étude 1	ER+	$1e^{-4}$	6	175
Étude 2	ER-	$1e^{-4}$	122	1481
Étude 2	ER+	$1e^{-4}$	14	272

Le nombre optimal de sous-réseaux pour une classification dépend des jeux de données utilisés pour l'apprentissage. Le fait qu'il soit plus faible pour les ER+ reflète une plus grande homogénéité des échantillons.

maximise la précision moyenne sur les dix couches de la stratification pour chacune des études. Pour l'étude 1, les sous-réseaux discriminatifs retenus avaient un score moyen de 0.49 (ER+) et 0.54 (ER-) confirmant la haute corrélation entre la co-expression et la proximité dans l'interactome. La taille des signatures était respectivement de 6 (ER+) et 165 sous-réseaux (ER-). Pour l'étude 2, la signature ER+ avait un score de classification optimal pour 14 sous-réseaux, et la signature ER- pour 122 sous-réseaux. Ces sous-réseaux correspondent respectivement à des listes de 175 (Étude 1, ER+), 2310 (Étude 1, ER-), 272 (Étude 2, ER+) et 1481 (Étude 2, ER-) gènes. Un grand nombre de gènes étant représentés dans plusieurs sous-réseaux. Ces nombre sont plus larges que ceux reporté pour les autres signatures. Ceci suggère que nous avons détecté un nombre important de gènes significativement liés à la rechute métastatique, reflétant de façon réaliste à la fois l'empreinte biologique de la métastase et l'ampleur des perturbations au niveau de l'expression des gènes. La redondance des gènes dans les sous-réseaux peut être expliquée par la haute connectivité de certains hubs (comme *TP53*), ce qui augmente leur probabilité d'être intégré dans plusieurs sous-réseaux.

Pour évaluer la performance des signatures découvertes avec ITI, nous les avons comparées avec des signatures déjà établies. Les 128 sondes du GGI[133], la signature Mammaprint à 70 gènes[78] et la signature statuts ER spécifique à 76 gènes[9] ont été testées. La performance a été mesuré sur les mêmes échantillons (jeux de données Desmedt et van de Vijver), séparément sur les tumeurs ER+ et ER-. La méthode de classification originale des signatures a été utilisée. Pour la signature de van de Vijver, les distances aux centroïdes moyens entre les groupes avec et sans rechute ont été calculées[78]. Pour la signature de Wang, un score de rechute est calculé pour chaque patient par combinaison linéaire de l'expression des gènes pondéré par des coefficients de Cox standardisés[9]. Les signature GGI et Mammaprint étant sondes-spécifiques, les tests ont donc été réalisés avec les sondes présentes dans le jeu de données de validation. Les résultats et les mesures des performances sont détaillés dans le Tableau IV.3.

Ces résultats montrent que notre algorithme ITI possède une performance largement plus généralisable que les autres signatures précédemment publiées. La classification par le

Tableau IV.3 – Comparaison des résultats de classification entre ITI et d'autres signatures sur les jeux de données de validation Desmedt et van de Vijver pour les tumeurs ER- et ER+.

Statuts	ER-								ER+							
Jeux de Données	*Desmedt*				*van de Vijver*				*Desmedt*				*van de Vijver*			
Signature	*GGI*	*70g*	*76g*	***ITI (165)***	*GGI*	*70g*	*76g*	***ITI (122)***	*GGI*	*70g*	*76g*	***ITI (6)***	*GGI*	*70g*	*76g*	***ITI (14)***
N	61	61	61	**61**	36	36	36	**36**	129	129	129	**129**	114	114	114	**114**
VN	6	0	14	**22**	3	2	12	**17**	63	28	53	**86**	57	39	50	**49**
FP	28	34	20	**12**	16	17	7	**2**	31	66	41	**8**	18	36	25	**26**
VP	23	27	9	**11**	14	17	8	**2**	21	25	25	**9**	20	32	22	**10**
FN	4	0	18	**16**	3	0	9	**15**	14	10	10	**26**	19	7	17	**29**
ACC	0.46	0.42	0.38	**0.54**	0.47	0.53	0.56	**0.53**	0.65	0.41	0.60	**0.74**	0.68	0.62	0.63	**0.52**
SV	0.85	1	0.33	**0.41**	0.82	1	0.57	**0.12**	0.60	0.71	0.71	**0.26**	0.51	0.82	0.56	**0.26**
SP	0.18	0	0.41	**0.65**	0.16	0.11	0.63	**0.90**	0.67	0.30	0.56	**0.92**	0.76	0.52	0.67	**0.65**

Les quatre signatures ont été utilisés pour mesurer la performance de classification d'ITI (**en gras**). Les abréviations suivantes ont été utilisées : N - nombre de tumeurs à classifier ; VN - Vrai Négatif ; FP - Faux Positif ; VP - Vrai Positif ; FN - Faux Négatif ; ACC - Justesse ; SV - Sensibilité ; SP - Spécificité. La performance de la classification basée sur les sous-réseaux est supérieure à la classification basée sur l'expression des gènes pour prédire la métastase dans le jeu de données de Desmedt, et ce de façon similaire pour le jeu de données de van de Vijver.

GGI montre une précision maximale entre 47 et 68 %, la signature à 70 gènes entre 41 et 62 % et la signature à 76 gènes entre 37 et 63%. ITI a une meilleure précision, comparée à la signature de Wang sur les données de Desmedt (ER+) : une précision de 74% (spécificité de 92%) a été obtenue contre une précision de 60% (spécificité de 56%) pour la signature à 76 gènes. ITI donne aussi des résultats plus performants sur les échantillons ER- des données Desmedt avec une précision de 54% (spécificité de 65%) contre une précision de 38% (spécificité de 41%) contre la signature de Wang. Ceci reste également vrai pour la signature Mammaprint à 70 gènes qui marche essentiellement sur le jeu de données van de Vijver. ITI montre une performance de 53% associée à une spécificité de 90% sur les données van de Vijver (ER-) et une précision de 52% avec une spécificité de 65% sur les données van de Vijver (ER-). Cette performance est inférieure à celle obtenu sur l'étude 1 et pourrait refléter un biais vers les plateformes Affymetrix introduit par l'apprentissage sur le compendium. Le signature Mammaprint est caractérisée par une précision inférieure (41% sur les tumeurs ER+ et de 42% sur les tumeurs ER- du jeu de données Desmedt). De la même manière, ITI a montré une performance supérieure à la classification GGI sur les patients ER-. Globalement, ITI a été capable de mieux généraliser avec valeur minimale pour la précision de 52%.

À titre de comparaison, Chuang et al.[10] ont achevé une performance de 48.8% sur les échantillons issus du jeu de données van de Vijver en les entraînant sur ceux du jeu de données Wang et 55.8% réciproquement. Les contributions spécifiques des données d'interactions ou des données d'expression des gènes ne sont pas quantifiées, elles sont difficilement séparables dans la configuration actuelle. Cependant, Chuang et al.[10] ont déjà démontré que qu'une approche intégrative augmentait la robustesse de la signature, et plusieurs études ont démontré que des méta-analyses de données d'expression des gènes augmentaient également les performances de classification[107,134]. Nous avons réalisé une analyse du temps de survie entre les groupes de bon et mauvais pronostics dans l'étude 1 (cf Figure IV.1). Un test Log-rank donne une p-value de 4.89×10^{-5}, suggérant une bonne séparation entre les deux groupes. Cette valeur est plus élevée que les p-values obtenues avec les autres signatures (la signature de Wang donne $P = 4.11 \times 10^{-3}$ et celle du GGI donne

P = 1.34x10^{-3}). La signature Mammaprint n'a pas été capable de séparer significativement dans des groupes les patients issus du jeu de données de Desmedt. Même si ITI n'a pas été spęcifiquement prévu pour obtenir un bon score au test Log-rank, il a été capable de séparer les patients avec une haute espérance de survie de ceux avec une espérance de survie faible. Une alternative aurait pu être de calculer directement les scores des sous-réseaux en se basant sur les P-values du log-rank des gènes.

IV.3 Exploration des sous-réseaux

Dous avons examiné les enrichissements en termes Gene Ontology pour la catégorie "biological process" pour les sous-réseaux obtenus dans l'étude 1. La Table IV.4 montre plusieurs enrichissements en termes GO pour les deux signatures ER+ et ER-. Les termes GO trouvés dans les sous-réseaux discriminatifs sont reliés à des processus regulationnels perturbé dans le cancer (cycle cellulaire, contrôle des dommages à l'ADN) et dans la métastase (système immunitaire, prolifération cellulaire, adhésion focale, migration cellulaire et organisation du cytosquelette) à la fois dans les tumeurs ER+ et dans les ER-. Comme exemple, nous décrivons ici un sous-réseau significativement associé avec la métastase dans l'étude 1 (ER-), le sous-réseau 6693 (cf Figure IV.2). Ce sous-réseau contient des gènes avec des fonctions connues pour leur implication dans les cancers du sein ER- et la métastase, comme le TSG *TP53* et les récepteurs *ERBB2* et *EGFR*. Ce sous-réseau contient également plusieurs kinases et régulateurs du cycle cellulaire (*CDK2*, *CDKN1A*, *CDKN2A*, *NQO1*), dont l'altération de l'expression a été précédemment associée avec plusieurs types de cancer. *PIN1* est présent dans ce sous-réseau et il a été récemment trouvé qu'il promouvait l'agressivité des tumeurs dans le cancer du sein. Le récepteur à l'insuline est également présent ; la dérégulation de son expression est corrélée avec une mauvaise réponse aux traitements anti-*IGFR* dans les cancers du sein triple négatif. Ce sous-réseau contient également de nombreux oncogènes et des gènes non-précédemment reliés au cancer, mais qui pourrait agir comme des gènes directeurs du cancer du sein.

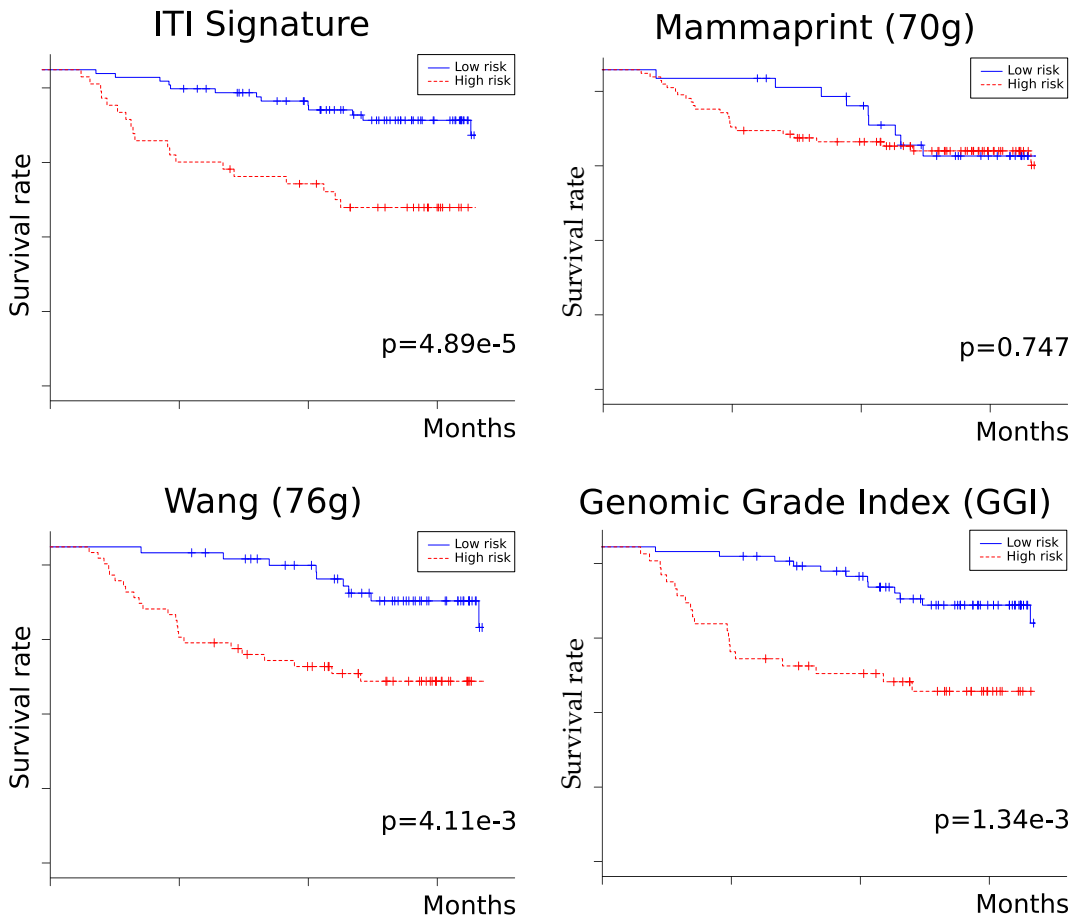


Figure IV.1 – Comparatif des courbes de survies des patients.

Tableau IV.4 – Enrichissement en termes GO des sous-réseaux ER- et ER+

	Terme GO	*GOID*	*P-value corrigée*
ER-	Natural killer cell-mediated imunity	GO :0002228	$293e^{-06}$
	Positive regulation of MAP kinase activity	GO :0043406	$476e^{-10}$
	Muscle cell development	GO :0055001	$106e^{-11}$
	Interphase of mitotic cell cycle	GO :0051329	$408e^{-11}$
	Wnt receptor signaling pathway through β-catenin	GO :0060070	$622e^{-10}$
ER+	mRNA cleavage	GO :0006379	$125e^{-08}$
	Regulation of growth hormone secretion	GO :0060123	$218e^{-07}$
	Positive regulation of cytoskeleton organization	GO :0051495	$206e^{-04}$
	Regulation of insulin secretion	GO :0050796	$155e^{-05}$
	Regulation of chemotaxis	GO :0050920	$429e^{-07}$

Plusieurs enrichissements en termes GO pour les sous-réseaux extraits dans l'étude 1 (ER- et ER+) sont reliés au cancer.

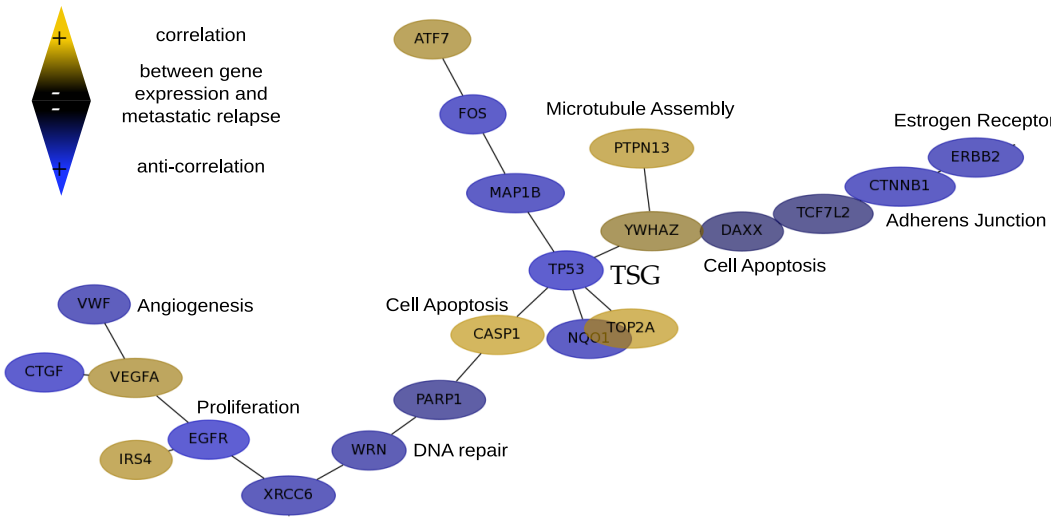


Figure IV.2 – Représentation graphique d'une partie du sous-réseau 6693, Étude 1 ER-.

Ce sous-réseau discriminatif a été découvert sur les données issues de Sabatier et al.[(118)]. Les nœuds et les arcs correspondent respectivement aux gènes codant pour des protéines et aux PII. Les couleurs jaunes et bleues des nœuds montrent respectivement une sur-expression et une sous-expression en comparant les patients avec métastases et ceux sans.

IV.4 Conclusion

L'ANALYSE supervisée réalisée avec ITI nous a permis de confirmer la significativité biologique des sous-réseaux découverts. La classification indépendantes des échantillons nous a permis de comparer les signatures obtenues avec ITI et les signatures précédemment établies. Nous avons donc pu confirmer l'avantage de l'ITI sur la robustesse des performances de la signature ainsi que sa reproductibilité sur un jeu de données indépendant. Nous allons maintenant discuter des autres avantages de cette approche.

CHAPITRE

V

DISCUSSION GÉNÉRALE

Résumé

Après un bref récapitulatif des travaux effectués lors de cette thèse, nous traiterons dans cette conclusion l'importance que peuvent avoir les données initiales et la question biologique posée sur la performance des signatures, la taille des sous-réseaux et la significativité biologique des signatures. Nous analyserons ensuite les caractéristiques des signatures obtenues avec ITI en les comparant avec les signatures précédemment citées. Nous finirons cette discussion en exposant des perspectives sur l'évolution de l'algorithme et des analyses futures.

Sommaire

V.1 **Rappels sur les travaux effectués** 84
V.2 **Importance des données initiales** 84
V.3 **Caractéristiques des signatures réalisées** 85
V.4 **Création d'une base de donnés de sous-réseaux** 87
V.5 **Perspectives** 88
V.6 **Conclusion** 89

V.1 Rappels sur les travaux effectués

NOUS AVONS DANS CET OUVRAGE exploré les caractéristiques biologiques des cancers et les spécificités du cancer du sein. Nous avons ensuite montré qu'une perturbation de l'expression et/ou la régulation des gènes pouvait être à l'origine du cancer. Nous avons rappelé l'intérêt des approches de médecines personnalisée et prédictive de classifier les tumeurs pour pouvoir les traiter de manière spécifique et proposer une thérapie adaptée. Après avoir exposé les limitations de l'approche transcriptomique à haut-débit pour étudier le cancer, nous avons détaillé notre méthode ITI, puis nous avons exposé les résultats d'une analyse non-supervisée et d'une analyse supervisée.

V.2 Importance des données initiales

LE JEU DE DONNÉES D'APPRENTISSAGE a une grande importance lors de l'établissement d'une signature prédictive, comme l'a montré Michiels et al.. Ce phénomène est dû à la fois à la topologie des données et à la biologie du cancer (cf Section I.4.e). Notre approche intégrant simultanément plusieurs jeux de données transcriptome permet de moins dépendre des jeux de données initiaux. De plus, le fait d'homogénéiser les échantillons permet de minimiser la variation biologique dûe au cancer. Ainsi lors de la détection du sous-réseau, un gène n'ayant pas de variation d'expression dans un jeu de données peut être pris en compte pour la constitution d'un sous-réseaux, si son expression varie suivant les conditions cliniques dans un autre jeu de données (cf Section II.6.a). Avec un compendium de plusieurs jeux de données recouvrant 930 tumeurs, nous obtenons une performance de classification supérieure sur un jeu de données totalement indépendant comme nous l'avons montré dans l'exploration des résultats de notre analyse supervisée (cf Table IV.3). Nous avons comparé les signatures réalisées avec ITI et les signatures issues de la littérature[9,78,133].

V.3 Caractéristiques des signatures réalisées

V.3.a Amélioration de la stabilité, robustesse et reproductibilité

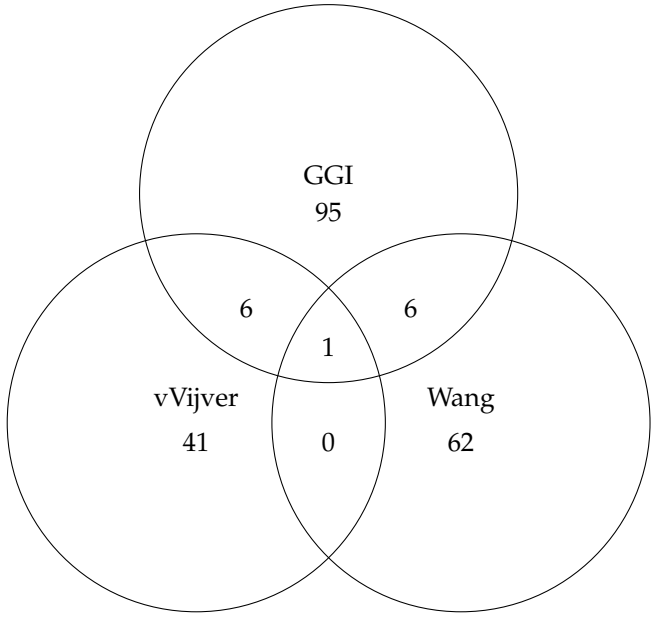
POUR COMPARER LA PERTINENCE de nos sous-réseaux aux signatures déjà publiées, nous avons récupéré dans les articles originaux, les listes des sondes constituant la signature Mammaprint à 70 gènes[78], la signature statuts ER spécifique à 76 gènes[9], et la signature du GGI[133]. Nous avons utilisé les annotations les plus récentes des sondes, enlevé les doublons, et n'avons pas considéré les EST. Les annotations des sondes étant constamment mises à jour, il peut y avoir une variation sur le nombre de gènes, c'est pourquoi nous trouvons un seul gène commun entre les signatures Van De Vijver et al.[78] et Wang et al.[9] alors que Chuang et al.[10] en trouvaient 3. C'est également la raison pour laquelle la signature GGI[133] que nous avons réalisé contient 108 gènes (et non 97 comme précisé dans la littérature). La Figure V.1 compare les différentes signatures et notamment les gènes en communs.

Il apparaît que très peu de gènes sont communs entre ces signatures (moins de 5% de la totalité des gènes des signatures deux à deux). À titre de comparaison, nous avons croisé les deux signatures ITI obtenues sur les échantillons ER+ et ER- avec nos deux études de notre analyse supervisée. Un total de 937 gènes sont communs entre nos deux signatures pour les échantillons ER-, et 46 gènes sont communs pour les signatures réalisées sur les échantillons ER+. Cela représente un recouvrement de respectivement 32.8% (ER-) et 11.5%(ER+). Ces valeurs relativement basses reflètent les biais dus aux jeux de données et aux plateformes de puces à ADN. Cependant, ce recouvrement est largement supérieur aux quelques gènes communs entre les autres signatures. Il pourrait être probablement augmenté en utilisant un ensemble de jeux de données d'entraînement plus large.

V.3.b Significativité biologique relevante

REPRENANT LES TROIS SIGNATURES ÉTABLIES dans la littérature (la signature Mammaprint à 70 gènes[78], la signature statuts ER spécifique à 76 gènes[9], et la signa-

Figure V.1 – Diagramme de Venn comptabilisant les gènes communs entre les différentes signatures classiques.



ture du GGI[133]), Haibe-Kains et al. supposent que les pronostics sur le devenir des patients sont basés sur la représentation de processus biologiques qui se recouvrent largement. Thomassen et al. ont comparé 9 signatures pronostiques, et ont trouvé que le cycle cellulaire et la prolifération cellulaire étaient les termes GO les plus représentés. Yu et al. ont conduit des analyses sur les différentes voies de régulation de 5 signatures pronostiques publiées avaient un grand nombre de voies en commun, comme le cycle cellulaire, la régulation du cycle cellaire, la mitose, l'apoptose, etc …Haibe-Kains et al. ont étudié dans une méta-analyse à grande échelle des données publiques d'expression des gènes et ont trouvé que la prolifération était une force directrice commune d'un grand nombre de signatures pronostiques. Nous retrouvons dans nos analyses ces processus biologiques impliqués dans le cancer et la métastase, mais également d'autres processus, comme le contrôle des dommages à l'ADN et le système immunitaire qui sont également important.

V.3.c Taille des sous-réseaux

LA QUESTION BIOLOGIQUE posée a son importance. Dans le cadre d'un cancer très hétérogène comme le cancer du sein, un nombre de gènes important est nécessaire pour réaliser une classifieur sur une question complexe comme la rechute métastatique. Si la question posée est la différentiation entre deux types de cancer où beaucoup de gènes sont différentiellement exprimés, comme pour la leucémie myéloïde aiguë et la leucémie lymphoblastique aiguë, peu de gènes peuvent suffire pour différencier les deux conditions cliniques Dobbin et al.[138]. L'homogénéité des données a également son rôle dans la taille finale des sous-réseaux, comme nous l'avons vu précédemment.

V.4 Création d'une base de donnés de sous-réseaux

LA CRÉATION D'UNE RESSOURCE bioinformatique permettant l'exploration des sous-réseaux. Cette ressource, a été mise en place, accessible par Internet, contient pour chacune de nos analyses les sous-réseaux détectés statistiquement significatifs. Pour chacun des sous-réseaux, un enrichissement en termes GO est calculé, par jeux de données d'expression, en se basant sur l'apport des gènes du sous-réseaux par rapport à l'ensemble des gènes présents sur la plateforme de puce à DNA utilisée pour constituer le jeu de données. Cette ressource permet de relier l'interactome à la biologie de la maladie étudiée, ici dans le cadre du cancer du sein. Les sous-réseaux stockés étant discriminatifs pour la rechute métastatique dans le cancer du sein, leur exploration, facilitée par des analyses en enrichissement en termes GO permet la découverte de gènes d'intérêts, non-liés précédemment au cancer du sein. Ces gènes d'intérêts peuvent être des cibles thérapeutiques potentielles, des TSGs ou des oncogènes putatifs, ainsi que des gènes directeurs potentiels.

V.5 Perspectives

V.5.a Améliorations de l'algorithme

LORS DE CETTE THÈSE nous avons pensé améliorer l'algorithme de détection de sous-réseaux, qui, en partant d'un gène graine, vérifie successivement et récursivement chacun des gènes voisins lors de la décision d'ajout d'un gène ou non sur le sous-réseau. L'algorithme explore donc l'interactome de façon linéaire, et l'ajout d'un gène voisin peut alors influencer la décision d'ajout d'un gène à la même distance de la graine. Nous pouvons modifier cet étape pour réaliser une exploration circulaire de l'espace interactome. Toujours en partant d'un gène graine, l'algorithme vérifierait donc simultanément tous les gènes voisins de même rang, et ajouterai ou non ceux qui améliorent le score du sous-réseaux. L'ajout des gènes de même rang se faisant de façon simultanée, il ne pourrait donc plus y avoir d'influence à l'ajout d'un gène sur l'ajout d'un autre gène de même rang.

V.5.b Intégration d'autres types de données

NOUS AVONS RÉALISÉ des analyses pour inclure des informations de variabilité du nombre de copies des gènes dans le chapitre "*CNV-Interactome-Transcriptome Integration to detect driver genes in cancerology*"[(15)]. Nous allons prochainement explorer la biologie particulière de certains sous-types du cancer du sein, comme les Claudin-low. Des travaux seront également réalisés pour inclure des données épigénétiques (méthylation de l'ADN), ainsi que des informations sur les miRNAs.

V.5.c Étude de l'importance de la nature de l'interaction

DANS LE BUT D'ÉTUDIER l'impact de la nature des différents types d'interactions sur la nature des sous-réseaux et la performance des signatures ainsi obtenues, nous allons réaliser des analyses en variant les jeux de données d'interactions utilisés. Le lien étroit entre interactions protéines-protéines et le niveau d'expression des gènes, déjà évoqué lors de l'explication de la validation statistique (cf Section II.6.b) laisse

supposer que l'utilisation d'un interactome plus bruité (*ie* un interactome contenant en plus des informations fausses sur les interactions protéines-protéines) aura peu d'impact. En effet l'utilisation d'un interactome aléatoire ne permet la création que d'un faible nombre de sous-réseaux. Nous supposons de plus qu'un interactome plus complet aura un impact positif sur la détection des sous-réseaux et sur la performance des signatures.

V.6 Conclusion

APRÈS UN RAPPEL des travaux réalisés, nous avons expliqué l'impact des données initiales et de la complexité de la question biologique posée sur les caractéristiques des signatures découvertes avec ITI. Nous avons également comparé les signatures découvertes avec ITI avec les signatures déjà publiées[9,78,133]. Nous allons maintenant aborder dans la dernière partie notre conclusion générale.

CHAPITRE

VI

CONCLUSION GÉNÉRALE

Conclusion Générale

Nous avons conçu un algorithme basé sur une approche réseau (ITI) pour identifier des signatures génomiques généralisables sur plusieurs jeux de données transcriptomiques de différentes origines. Cet algorithme fonctionne en deux étape : tout d'abord, il intègre des données d'un compendium de jeux de données de puces à ADN dans le cancer du sein, et il permet de détecter des sous-réseaux, *ie* des groupes de gènes interagissant ensemble, dont l'expression discrimine deux conditions d'intérêt. Les sous-réseaux sont filtrés par validation statistique. Nous avons appliqué l'algorithme ITI à la question complexe de la rechute métastatique dans le cancer hétérogène qu'est le cancer du sein pour lequel un grand nombre de données publiques sont disponibles.

Notre approche démontre la faisabilité de l'intégration d'un large compendium de données d'expression des gènes (2103 tumeurs du sein ont été intégrées dans notre analyse non-supervisée et 930 dans notre analyse supervisée) et un réseau d'interactions protéine-protéine à grande échelle. ITI représente un outil potentiellement utile pour explorer les sites de dépôts de données d'expression des gènes. Dans l'étude de la rechute métastatique dans le cancer du sein, nous avons produit deux signatures statut ER spécifique qui ont été validées sur des jeux de données indépendants. Nous avons obtenu une meilleure performance de classification que les précédents classifieurs publiés (74% pour Desmedt (ER+) et 53% pour van de Vijver (ER+)). Nos signatures basées sur les réseaux reflètent la large empreinte biologique de la métastase et est par conséquence plus large que les signatures précédemment publiées. Le classifieur obtenu avec ITI est moins sensible que les classifieurs précédemment publiés aux biais des plateformes, puisque la performance de la signature ITI reste similaire sur les deux compendium d'entraînement. Nos signatures montrent également une spécificité forte, ce qui est critique dans le cas de prise de décision pour éviter un traitement adjuvant systémique inutile.

L'algorithme ITI est actuellement étendu pour incorporer d'autres types de données (CNV [14], méthylation de l'ADN, miRNAs). ITI est capable d'atténuer le fléau de la dimensionnalité, rendant possible la détection de biomarqueurs par des analyses de type

NGS. Dans les prochaines versions, la nature de l'interaction protéine-protéine sera prise en compte lors de l'étape de détection des sous-réseaux. Les performances de classification sont inhérentes aux sous-types moléculaires, et un sous-typage plus fin est nécessaire pour permettre l'utilisation de cette technologie pour des usages cliniques. Une future validation clinique pourrait être envisagée avec un essai clinique utilisant des puces à ADN. L'utilisation de plusieurs sources de données en entrée pourraît nourrir une intégration multiple et massive aboutissant à la découverte d'une signature encore plus robuste et généralisable. La performance de l'algorithme est prouvée dans le cadre de la question complexe de la rechute métastatique dans le cadre hétérogène du cancer du sein, il serait intéressant de transposer cet algorithme non seulement à d'autres questions biologiques, telle la réponse au traitement où la différentiation entre sous-types moléculaires, mais aussi à d'autres types de cancer ainsi qu'à d'autres maladies.

ANNEXES

Détail

Dans cette section sont regroupées des informations complémentaires : la nomenclature employée pour nommer les gènes et les protéines, des listes ordonnées des abréviations, noms de gènes et noms des protéines utilisés ; ainsi que les publications présentées dans cette thèse.

ANNEXE

A

NOMENCLATURES

Nous utilisons les nomenclatures dans ces pages.

Gènes

La nomenclature HGNC est utilisée pour le nom des gènes. Le symbole officiel du gène est utilisé tout au long de ce document. Pour écrire le symbole du gène nous utilisons la nomenclature usuelle avec le symbole du gène en majuscule et en italique (exemple : *TP53*). Le nom complet officiel est détaillé dans les abréviations B.1.

Protéines

Le symbole officiel du gène dont est issue la protéine est utilisé tout au long de ce document. Pour écrire le symbole de la protéine, nous utilisons la nomenclature usuelle avec le symbole du gène en majuscule sans italique (exemple : TP53). Le nom complet recommandé par le consortium UniProt est détaillé dans les abréviations B.2.

ANNEXE

B

ABRÉVIATIONS

B.1 Gènes

APC adenomatous polyposis coli

ARNTL aryl hydrocarbon receptor nuclear translocator-like

BRCA1 breast cancer 1, early onset

BRCA2 breast cancer 2, early onset

CCND1 cyclin D1

CDH1 cadherin 1, type 1, E-cadherin (epithelial)

CDK1 cyclin-dependent kinase 1

CDK2 cyclin-dependent kinase 2

CDKN1A cyclin-dependent kinase inhibitor 1A (p21, Cip1)

CDKN2A cyclin-dependent kinase inhibitor 2A

CLOCK clock circadian regulator

E2F E2F gene family

EGFR epidermal growth factor receptor

ERBB2 v-erb-b2 erythroblastic leukemia viral oncogene homolog 2, neuro/glioblastoma derived oncogene homolog (avian)

GHR growth hormone receptor

GRB2 growth factor receptor-bound protein 2

HRAS v-Ha-ras Harvey rat sarcoma viral oncogene homolog

KRAS v-Ki-ras2 Kirsten rat sarcoma viral oncogene homolog
IGFR insulin-like growth factor 1 receptor
LAMA4 laminin, alpha 4
MYC v-myc myelocytomatosis viral oncogene homolog (avian)
NCK1 NCK adaptor protein 1
NPAS2 neuronal PAS domain protein 2
NQO1 NAD(P)H dehydrogenase, quinone 1
PDGFB platelet-derived growth factor beta polypeptide
PGK1 phosphoglycerate kinase 1
PIN1 peptidylprolyl cis/trans isomerase, NIMA-interacting 1
RAS rat sarcoma viral oncogene homolog
RB1 retinoblastoma 1
TK1 thymidine kinase 1, soluble
TSC1 tuberous sclerosis 1
TP53 tumor protein p53
VEGFA vascular endothelial growth factor A
WT1 Wilms tumor 1
WNT wingless-type MMTV integration site family

B.2 Protéines

EGF Pro-epidermal growth factor
EGFR Epidermal growth factor receptor
ERBB2 Receptor tyrosine-protein kinase erbB-2
ESR1 Estrogen receptor
HRAS GTPase HRas
MYC Myc proto-oncogene protein
NPAS2 Neuronal PAS domain-containing protein 2
PGR Progesterone receptor
TP53 cellular tumor antigen p53
VEGFA Vascular endothelial growth factor A

B.3 Institutions

AMU Aix-Marseille Université
Cibi plateforme de Bioinformatique Intégrative du CRCM
CépiDc Centre d'épidémiologie sur les causes médicales de décès

CNRS Centre National de la Recherche Scientifique
CRCM Centre de Recherche en Cancérologie de Marseille
ENCODE Encyclopedia of DNA elements
EORTC European Organisation for Research and Treatment of Cancer
FRM Fondation pour la Recherche Médicale
HUGO Human Genome Organisation
HGNC HUGO Gene Nomenclature Committee
IARC International Agency for Research on Cancer
ICGC International Cancer Genome Consortium
INCa Institut National du Cancer
Inserm Institut National de la Santé et de la Recherche Médicale
InVS Institut de Veille Sanitaire
IPC Institut Paoli-Calmettes
IPMC Institut de Pharmacologie Moléculaire et Cellulaire
NCBI National Center for Biotechnology Information
NCI National Cancer Institute
OMS Organisation Mondiale de la Santé
PACA Provence Alpes Côte d'Azur
SEER Surveillance Epidemiology and End Results Program
TAGC Technological Advances for Genomics and Clinics

B.4 Divers

ADN acide désoxyribonucléique (*DNA*)
ARN acide ribonucléique (*RNA*)
ARNm ARN messager (*mRNA*)
ARNt ARN transfert (*tRNA*)
ATP adénosine triphosphate
CNV Variation du nombre de copies (*Copy Number Variation*)
ChIP Chromatin immunoprecipitation
ChIP-Chip ChIP combined with DNA microarray analysis
ChIP-Seq ChIP combined with HTS
CORUM Comprehensive Resource of Mammalian protein complexes
DNA desoxyribonucleic acid
DIP Database of Interacting Proteins
DMFS Survie sans rechute métastatique (*Distant Metastasis Free Survival*)
ER récepteur aux œstrogènes (*œstrogen receptor*)

ER+ possédant le recepteur aux œstrogènes ESR1
ER- ne possédant pas le recepteur aux œstrogènes ESR1
GEP profil d'expression de gènes (*gene expression profile*)
GGI Genomic Grade Index
GO Gene Ontology
HTS High-Throughput Sequencing
HPRD Human Protein Reference Database
HER2 récepteur ERBB2 de la famille des récepteurs EGFR
HER2+ possédant le recepteur ERBB2
ITI Interactome-Transcriptome Interaction
ITI Intégration Transcriptome-Interactome
miARN micro ARN (*miRNA*)
miRNA micro RNA
MINT Molecular INTeraction database
MPF facteur de promotion de la maturation (*Maturation-promoting factor*)
mRNA messenger ribonucleic acid (RNA)
NGS Séquençage de nouvelle génération (*Next Generation Sequencing*)
pb paires de base
PII interaction protéine-protéine (*protein-protein interaction*)
PR récepteur à la progestérone (*Progesteron receptor*)
RNA ribonucleic acid
SNP polymorphisme nucléotidique (*single-nucleotide polymorphism*)
SVM machine à vecteurs de support (*Support Vector Machine*)
TF facteur de transcription (*transcription factor*)
TNM Tumeur-Ganglion-Métastase (*Tumor-Node-Metastasis*)
tRNA transfer RNA
TSG gène suppresseur de tumeurs (*tumor suppressor gene*)

ANNEXE

C

PUBLICATIONS

C.1 Chapitre *Linking Interactome to Disease*

Dans ce chapitre intitulé "*Linking Interactome to Disease : A network-based analysis of Metastatic Relapse in Breast Cancer*"[12], nous décrivons en détail l'algorithme ITI ainsi que la façon dont on l'utilise pour réaliser une analyse non-supervisée.

Abstract
www.igi-global.com/chapter/linking-interactome-disease/52327

Base de données des sous-réseaux
http://iti.sourceforge.net/unsupervised-10-datasets/index.html

Documentation
http://sourceforge.net/p/iti/wiki/Home/

Code Source
http://sourceforge.net/projects/iti/files/Source%20Code/iti-1.0.tar.gz

406

Chapter 19
Linking Interactome to Disease:
A Network-Based Analysis of Metastatic Relapse in Breast Cancer

Maxime Garcia
Inserm, Paoli Calmettes Institute, France

Olivier Stahl
Inserm, Paoli Calmettes Institute, France

Pascal Finetti
Inserm, Paoli Calmettes Institute, France

Daniel Birnbaum
Inserm, Paoli Calmettes Institute, France

François Bertucci
Inserm, Paoli Calmettes Institute, France

Ghislain Bidaut
Inserm, Paoli Calmettes Institute, France

ABSTRACT

The introduction of high-throughput gene expression profiling technologies (DNA microarrays) in molecular biology and their expected applications to the clinic have allowed the design of predictive signatures linked to a particular clinical condition or patient outcome in a given clinical setting. However, it has been shown that such signatures are prone to several problems: (i) they are heavily unstable and linked to the set of patients chosen for training; (ii) data topology is problematic with regard to the data dimensionality (too many variables for too few samples); (iii) diseases such as cancer are provoked by subtle misregulations which cannot be readily detected by current analysis methods. To find a predictive signature generalizable for multiple datasets, a strategy of superimposition of a large scale of protein-protein interaction data (human interactome) was devised over several gene expression datasets (a total of 2,464 breast cancer tumors were integrated), to find discriminative regions in the interactome (subnetworks) predicting metastatic relapse in breast cancer. This method, Interactome-Transcriptome

DOI: 10.4018/978-1-60960-491-2.ch019

Copyright © 2011, IGI Global. Copying or distributing in print or electronic forms without written permission of IGI Global is prohibited.

Integration (ITI), was applied to several breast cancer DNA microarray datasets and allowed the extraction of a signature constituted by 119 subnetworks. All subnetworks have been stored in a relational database and linked to Gene Ontology and NCBI EntrezGene annotation databases for analysis. Exploration of annotations has shown that this set of subnetworks reflects several biological processes linked to cancer and is a good candidate for establishing a network-based signature for prediction of metastatic relapse in breast cancer.

INTRODUCTION

Since introduction of high-throughput technologies in molecular biology in the late nineties, a number of technologies for deciphering the genomic origin of several diseases has flourished. Among these, cDNA microarrays (Schena et al., 1995) have allowed measuring Gene Expression Profiles (GEP) at the genome scale and have shed light on large scale gene regulation/misregulation under varied conditions. Many diseases, including several forms of cancer [leukemia (Golub et al., 1999), colon cancer (Li et al., 2001), breast cancer (Wang et al., 2005)], diabetes (Kaestner et al., 2003), and others (Munro & Perreau. 2009) have been studied that way. Of particular interest in the context of cancer, a particularly heterogeneous disease, is the use of GEPs to either predict drug resistance (de Lavallade et al, 2010), or the metastatic recurrence, for instance in breast cancer (van de Vijver et al., 2002). Tumor microenvironment studies have allowed understanding the influence of immune system on patient outcome (Pagès et al., 2009).

There is an increasing number of controversial cases for the use of systemic adjuvant therapy due to the clinical and pathological heterogeneity of the disease to treat. In node-negative early breast cancer, most patients undergo adjuvant chemotherapy even though 70-80% of them would have survived without it (Bertucci & Birnbaum, 2008). The refinement of current prognostic histopathological methods using molecular diagnostics can also lead to increase of detection of disease subtypes that necessitate specific treatments, such as T1 breast cancer (Mook et al., 2010). In all cases, the goal is to refine and individualize treatment and lead the way to personalized medicine for a growing number of pathologies.

In cancer, the understanding of molecular basis of metastasis is of primary importance. Several studies have attempted to obtain a molecular portrait for a large number of patients using DNA microarray analysis, performed supervised analysis and published list of genes predicting patient outcome. Two of these signatures are currently under clinical trials in breast cancer: the MINDACT trial, based on the 70-genes Mammaprint signature [van't Veer et al. (2002) van de Vijver et al. (2002), Bueno-de-Mesquita et al. (2007)]), and the TAYLORx trial, an RT-PCR-based 21-genes OncotypeDX signature (Paik et al., 2006).

However, most prognostic signatures reported for breast cancer show very little or no overlap, and do not appear generalizable from one study to another, and this un-reproducibility was widely criticized (Chuang et al., 2007, Bertucci et al., 2008). Two studies in particular are often cited for their lack of agreement, although they addressed similar questions, which were the two breast cancer studies performed by van de Vijver et al. (2002) and Wang et al. (2005), who reported two prognosis signatures for metastatic relapse in breast cancer. Two different signatures comprising respectively 70 and 76 genes predictive of breast cancer patient outcome were reported but presented only three genes in common. Even more concerning was the study by Michiels et al. (2005) which showed that hundreds of 70-genes signatures with equal classification power can be drawn by shuffling the training and test sets in the van't Veer et al. (2002) dataset, showing the

instability and dependency of the resulting gene lists on the training data.

Microarray technology itself was blamed at first for these inconsistencies, and DNA microarrays were suspected to be extremely noisy and leading to non reproducible results. Once the stability and inherent reproducibility were demonstrated by comparing several platforms in the Microarray Quality Control project (Shi et al., 2006), the reasons for the lack of uniqueness in gene signatures had to be found elsewhere.

This chapter deals with addressing the problem of signature instability and proposes a new computational model, the Interactome-Transcriptome Integration (ITI), which simultaneously integrates multiple datasets, to compensate for data dimensionality, and uses the human interactome to include genes with weaker signal in the signature.

BACKGROUND

Instability of Signatures

The aim of any classification study is to provide a good prediction model. From that viewpoint (for pure classification and prognosis prediction), the gene set/classifier does not have to be unique (Dobbin et al., 2008), especially since it has been shown that established signatures have a high rate of concordance in regard to other datasets (Fan et al., 2006). However, from a biological point of view, the fact that signatures are not repeatable among studies is not acceptable. This shows a lack of robustness in detection methods that could prevent widespread acceptance of DNA microarray profiling methods for routine clinical use (Ein-Dor et al., 2006). The case holds for Next Generation Sequencing profiling, for which similar issues are likely to appear.

The trivial and most simplistic reasons are frequently cited and are insufficient to explain this situation. Are the discrepancies due to the heterogeneity of platforms and analytic tools used by microarray core among different institutions, or the genetic background inherent to each patient, or to the variations among statistical methods and classifiers?

Besides these, reasons behind the lack of generalization are twofold: (i) DNA microarray data structure and (ii) the biology of cancer. The data topology - too few patients profiled on too many variables - prevents any classifier to be trained according to proper statistical standards. Analysis is therefore suffering from a double curse (Fishel et al., 2007): *the curse of dimensionality* (too many variables), and the *curse of sparsity* (too few samples).

In addition, Chuang et al. (2007) showed that microarray data are highly sensitive to subtle misregulation of a few genes. Highly differentiated genes from an experiment are therefore resulting from subtle misregulation (or mutation) of a smaller set of genes that are at the origin of the disease. These genes are the true perpetrator of the clinical condition studied we wish to predict, but are not detected. For instance, mutation in RAS family of oncogenes, which are very common in multiple types of tumors, have a disastrous effect on regulation of a mitogen-activated protein kinases which in turn phosphorylates several receptors which can be potentially very different from one patient to another (Goodsell 1999). Availability of large scale protein-protein interaction data gives the opportunity of retrieving a large number of these hubs, critical for breast cancer relapse, whose activity is measured on several datasets.

Meta-Analysis Methods

Several solutions have been envisioned to tackle the curse of dimensionality issue. The most practical methods are based on a meta-analysis of several datasets, to increase sample size computationally. Meta-analysis consists of comparing gene statistics inferred from several datasets and combining them to obtain an integrated gene list. Methods of meta-normalization could also be envisioned,

where datasets could be combined before analysis, but data heterogeneity prevents the adoption of these methods (Iziarry et al., 2005). Application to real data showed that meta-analysis in general achieves higher reproducibility of results than independent studies. Hong & Breitling (2008) have reported a comparison of three previously published methods: the T-based hierarchical clustering, rank products, and Fisher's Inverse chi-2 test. Fisher's Inverse chi-2 test [Fisher (1925), Bioconductor package GeneMeta, Gentleman et al.,(2004)] is a straightforward method combining *p*-values measured on different datasets to obtain a combined score for each gene. The T-based hierarchical clustering relies on the measurement of individual *t*-statistics on each dataset and the assessment of intra and inter-study variation by hierarchical modeling, also implemented in GeneMeta (Gentleman et al., 2004). Rank products method (Breitling et al., 2004) is also available as a Bioconductor package [RankProd, Hong et al. (2006)].

Other methods have been proposed. Conlon et al. (2006) proposed a Bayesian model to pool multiple independent studies and provided a Bayesian model of False Discovery Rates. Top-scoring pairs (TSP) method was used for cancer data integration (Xu et al., 2005). Van Vliet et al. (2007) proposed the use of meta-features (for instance, modules related to functional grouping of genes), which, by reducing the number of input variables, helps alleviate the curse of dimensionality. This is an extension of the model proposed by Segal et al. (2004) with training on a cancer compendia and inclusion of a classification system validated on independent data. Recently, more advanced techniques have been proposed, such as a neural network-based multi conditional classifier applied to developmental biology: In our previous studies, (Bidaut & Stoeckert, 2009a; Bidaut & Stoeckert, 2009b) we combined several stem cell profiles using vector projection technique to discover a multi-stage reproducible signature. Fishel et al. (2005) proposed a predictor-based approach (repeatability-based gene list, RGL) to find a stable lung cancer differentiation signature.

In addition, several reports have developed methods to answer the critical question of estimating the necessary sample size. Ein-Dor et al. (2006) suggested that several thousand patients are needed to obtain a signature that is robust among several studies, i.e., independent on the training set. Dobbin et al. (2008) tempered that conclusion and stated that such a high number of samples is not necessary when expression measurement information such as the largest standardized fold change, and the proportion of samples in each class are taken into account. They proposed a formula to calculate sample size based on a minimal set of information including the largest standardized fold change, the number of features and the data structure, i.e., the proportion of cases and control in the data.

The use of prior biological information helps reduce data dimensionality (Bidaut el al., 2006). In the past five years, several alternative approaches using network analysis have been proposed. For instance, Chuang et al. (2007) superimposed GEP over human interactome to generalize a signature for breast cancer metastasis relapse but using only a single dataset for training, as opposed to our approach. Wachi et al. (2005) showed that genes expressed in lung cancer tissues have a higher connectivity and are centrally located in the protein network. For a review of responsive functional modules identification in PPI networks, see the review by Wu et al. (2009).

INTERACTOME-TRANSCRIPTOME INTEGRATION

To discover a stable and robust signature predicting breast cancer metastatic potential and to infer genes subtly misregulated but crucial for such prediction, we proceeded by linking the human interactome on the largest body of available breast tumors profiled on DNA microarray, using a

framework named as Interactome-Transcriptome Integration (ITI). Basically, we created a breast cancer compendium from several DNA microarray datasets and superimposed it over the human interactome. The compendium was built by selecting individual datasets on the basis of clinical information availability and large overlap with existing protein-protein interaction data. Several DNA microarray platforms are represented in the compendium (7 distinct platforms in total, see Table 1) in order to avoid platform biases. This gives the ability to recover common subtly differentiated genes correlated with distant metastatic relapse. A signature is searched on all data simultaneously by parsing the interactome and aggregating subsets of nodes correlated with distant metastatic relapse in a number of datasets (See ITI Algorithm section for details). No extra normalization step was necessary to integrate individual datasets, as gene expression is not used directly. Correlation of gene expression profiles with clinical situation was rather superimposed on the interaction data.

After superimposition of expression information over the interactome, the list of interactions was searched for with consistent agreement of discrimination power over multiple datasets, leading to a database of subnetworks linked to metastasis in breast cancer. This database is available from the ITI web site main page. Several expression datasets combinations were tested to assess platform bias, as shown in Table 2, to compare discriminative subnetworks assessed from all datasets with subnetworks derived only with datasets profiled on Affytmetrix platforms (data not shown).

Human Interactome: Combining Several Sources of Large Scale Interaction Data

To build our set of interaction data, we integrated two existing human interactomes. The first is a recent version of the Human Protein Reference Database [HPRD version 8, released on June 7th, 2009, (Prasad et al. (2009)]. We used the flat file version available from the HPRD web site (http://www.hprd.org) after registration for non-commercial use. This file includes 35880 binary interactions between 8769 proteins after removal of unidentified interactors. The second set of interactions is the *in silico* predicted interactome described in Ramani et al. (2005). The Ramani interactome is available as a flat file and including 31609 interactions between 7500 proteins. We chose to omit self interactions present in HPRD (already filtered out in the Ramani interactome for benchmarking reasons) as they are not quantified in the subnetwork search process. Both interactomes were integrated by uniqueness of NCBI EntrezGene identifiers, leading to a final set of 57991 interactions among 10943 proteins (with an overlap of 7165 interactions between them).

Building a Breast Cancer Compendium

We integrated 12 distinct DNA microarray datasets of breast tumor profiles in our compendium, after examination of about two tens of datasets for clinical information availability. Most datasets are accessible from NCBI Gene Expression Omnnibus (Barrett et al., 2009) repository with exception of the van de Vijver data, available from the original publication supplementary web site. Each dataset was downloaded either as raw data from GEO and normalized within Bioconductor using Affy and GCRMA packages, or directly loaded from GEO as a GSE file when raw data were unavailable, or from the author's web site. Correspondence tables between gene IDs and probe IDs were constructed with the method described in Reynal et al. (2005). Briefly, one probe was kept per gene by filtering out all probes carrying "x_at" extension and keeping the probe with highest expression profile median.

Table 1 summarizes the breast cancer compendium datasets, links to GEO, platforms, publications and sample size. Thereafter, we name

Linking Interactome to Disease

Table 1. List of datasets included in the Breast Cancer Compendium for training. Datasets in light grey were considered but not included because of lack of clinical data or lack of platform information, but are potentially includable if such information could be gathered. Some tumors were filtered out if they were already present in other datasets (for instance van't Veer dataset is filtered out since it has been included in van de Vijver). The compendium result from integration of white datasets, resulting in a total set of 2464 untreated tumors annotated with e.DFS or e.DMFS. Datasets spanning over multiple platforms (133A and B) were integrated into one (see Methods).

Dataset	NCBI Accession number (if available)	Platform	Number of samples before filtering	Number of samples after filtering	Presence of clinical information (e.DFS or e.DMFS)
Anders	GSE7849	U95v2	78	78	No
Bild	GSE3143	U95v2	158	158	No
Campone	GSE7017	UMGC-IRCNA 9k A	150	150	No
Chang	GSE3945	cDNA array	50	50	No
Chang-Kyu	GSE2845	Merck GEL Breast Tumor Profiles	311	311	No
Chanrion	GSE9893	MLRG Human 21K V12.0	155	155	No
Desmedt	GSE7390	U133A	198	198	Yes
Finetti		U133 Plus 2.0	129	129	Yes
Ivshina	GSE4922	U133 Plus 2.0	289	249	Yes
Jezequel	GSE11264	UMGC-IRCNA 9k A	252	252	No
Kreike	GSE4913	NKI-AVL 18K cDNA	59	59	Yes
Loi	GSE6532	U133A + U133B	327	293	Yes
		U133 Plus 2.0	87	87	
Miller	GSE3494	U133A + U133B	251	251	Yes
Parker	GSE10886	Agilent-011521 1A G4110A	2	2	Yes
		Agilent-012097 1A G4110B	27	22	
		Agilent 1A Oligo UNC Custom	196	177	
Pawitan	GSE1456	U133A + U133B	159	159	Yes
Perou	GSE61	SCV	84	84	No
Schmidt	GSE11121	U133A	200	200	Yes
Sorlie	GSE3193		85	85	No
Sotiriou	GSE2990	U133A	189	179	Yes
van de Vijver		Agilent whole human genome	295	295	Yes
van't Veer		Agilent whole human genome	117	117	Yes
Wang	GSE2034	U133A	286	286	Yes
Wong	GSE7930	U133A	6	6	No
Yu	GSE5364	U133A	341	341	No
Zhang	GSE12093	U133A	136	136	Yes
Zhou	GSE7378	U133Av2	54	54	Yes
Total: 12		7 distinct	2572	2464	

Table 2. Cross validation training organization. Two series were trained on all data from the breast cancer compendium but one (A1, B1), whereas two other series were trained only on datasets profiles on Affymetrix platforms (A2, B2) to assess inter-platform subnetwork stability.

Run	Training datasets
A1	All but van de Vijver
B1	All but Wang
A2	All Affymetrix platforms
B2	All Affymetrix Platforms but Wang

datasets after the corresponding paper's first author name. Since we are building a prognostic classifier for metastatic relapse, we gathered the clinical information related to e.DMFS (Distant Metastasis-Free Survival) or e.DFS (Disease-Free survival) for every dataset when available. Availability of this information was required to include a given dataset within our analysis. For the Parker, Pawitan and Wang studies, distant metastatic relapse information was not mentioned, and disease relapse information was used instead (variable e.DFS). The initial dataset from Ivshina contained 58 samples from the Pawitan study that were removed to avoid duplicates, and datasets profiled on Affymetrix HG-U133A and B platforms were merged (Ivshina, Loi and Pawitan) by creating a virtual platform annotation file of 44692 probes using the methodology previously employed for probe to gene expression conversion. Tumors without distant metastatic relapse information were further removed, leading to a final compendium of 2464 tumors. Clinical information was binarized in order to compute Pearson correlations with GEPs. Annotations were gathered for each platform from the Resourcerer database (Tsai et al. (2001), Data downloaded on Nov. 1st 2008). The Gene_info file was downloaded from NCBI the On Sept. 1st 2009 and was used as a table of correspondence between NCBI geneID accession numbers and NCBI gene Symbols (Sayers et al., 2010).

ITI Algorithm

To superimpose physical interaction data (human interactome) to several transcriptome datasets, we constructed an algorithm named ITI, Interactome-Transcriptome Integration, derived from the one described in Chuang et al. (2007) but extended to perform the analysis on several datasets. This algorithm allows one to superimpose GEP from several datasets to a map of physical interactions and to extract subnetworks that consistently discriminate two opposite clinical conditions across a number of datasets. To do so, a heuristic method examines all nodes present in interaction data and tries to construct a subnetwork by recursive aggregation of neighboring nodes. Aggregation is done on the basis of consistency of gene expression across subnetwork and high correlation of the whole subnetwork with the clinical condition. This is quantified by a subnetwork score computed as the absolute value of average correlation of gene expression profiles of genes included in the subnetwork with a numerical vector representing clinical situation for each dataset.

Several variables are set before starting the algorithm: *th* is the minimal score threshold that a subnetwork must meet to be accepted, *mi* is the minimal score increase when adding a new node to an existing subnetwork, and *c* (consensus) is the minimal number of datasets on which a gene must meet conditions on *th* and *mi* to be added on the subnetwork. The following formula *Sc(S, d)* details score calculation of a subnetwork *S* over a single dataset *d*. A global subnetwork score is computed for each subnetwork by averaging scores obtained over all datasets for information (see supporting web site at http://bioinformatics.marseille.inserm.fr/iti).

$$Sc(S,d) = Pearson_corr(\frac{1}{p}\sum_{g \in S} GEP(g)), Cc(d))$$

S being the current subnetwork, *Sc(S)* the score, *d* the current dataset index, *p* the number of genes contained in the subnetwork *S*, *g* the current gene, *GEP(g)* the gene *g* expression profile of gene *g*, and *Cc*, the numerical vector representing clinical condition (1 = relapse, 0 = healthy).

To construct subnetworks, the following recursive algorithm is used. The subnetwork is first constructed from a candidate seed, and a recursive method aggregates neighbors' nodes if the score *Sc* stays above threshold *th* over at least *c* datasets and does not vary below the minimal value *mi* (minimal increase). The following pseudo code details the method, also represented in Figure 1. *Nc(S)* is the current consensus value for subnetwork *S*.

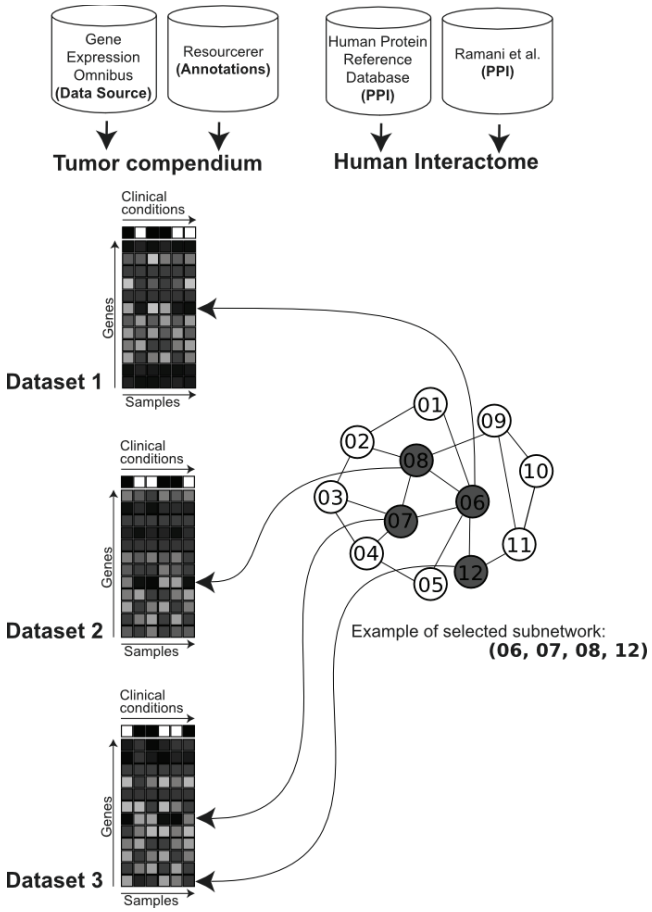
```
testSubnetwork = Subnetwork = empty;
Routine constructNode For Each node In
interactome
testSubnetwork = concatenate(node,
Subnetwork)
Nc=0;
For Each dataset
          Sc(testSubnetwork)= Com-
pute-subnetwork_score(testSubnetwork,
dataset)
          If Sc(Subnetwork,
dataset)>th and (Sc(testSubnetwork)
- Sc(Subnetwork))
>mi Then Nc(Subnetwork) ++
          End
End
If Nc(Subnetwork)>=c Then Subnetwork =
testSubnetwork
          For Each node In
neighbor(node) constructNode(node) End
Else break
End
End
```

Choice of variables has an impact on detected subnetworks' number and size. Obviously, lowering *th* and *c* will increase the number of detected subnetwork, and lowering *mi* increases subnetwork size. However, lowered scored subnetwork will be filtered out by the statistical validation step. Parameters have been set to the following values: *mi* = 0.01, *th* = 0.05 and *c* = 6 to obtain a reasonably sized subnetwork set. Subnetworks overlapping by more than 80% were removed. To tackle the computational cost of this algorithm, we parallelized it on a 96 cores [12 nodes] Beowulf cluster. Parallelization was done by partitioning the interactome over the nodes, leading to a 45 minutes approximate execution time, including random distribution drawing for statistical validation (see following section). As a point of comparison, execution time on a single CPU is about 10 hours.

Subnetworks Statistical Validation

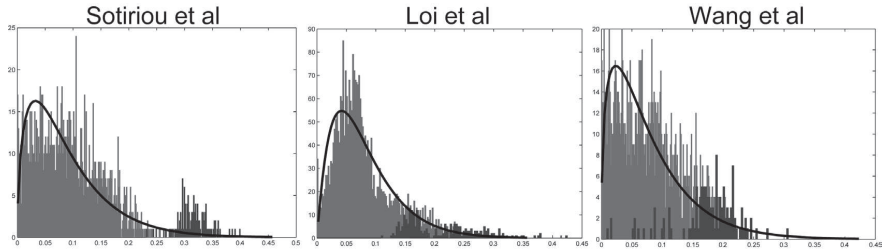
Subnetworks are validated over three *p*-values, related to (i) the node aggregation decision (type 1 *p*-values), (ii) the link between expression data and interaction data (type 2 *p*-values), and (iii) the network topology (type 3 *p*-values). Each *p*-value is computed by drawing a random distribution of scores, and setting up a score threshold. In this framework, three random distributions are computed for each dataset. The first is computed by randomizing aggregation decision (nodes are added randomly until subnetwork size reaches a normally randomly distributed value having the same distribution as the regularly detected subnetworks). The second is computed by shuffling experimental conditions over datasets. The third is computed by shuffling all interactome interactions while keeping the original connectivity ±1 for each node. This allows drawing a *p*-value distribution evaluating only the link between protein-protein interactions and co-expression while conserving

Figure 1. Scheme represents the ITI Algorithm of DNA microarray datasets integration, their superimposition over the human interactome, and the construction of discriminating subnetwork. Briefly, all nodes are considered as a seed, and neighboring nodes are aggregated if discriminative over clinical condition over a number of datasets.



the human interactome power law distribution. *P*-values are then computed from these distributions by generating 3000 random subnetworks within these three settings. As an argument to the program, distribution model type can be given to properly model the random distributions (Gamma distribution, Normal distribution, and bimodal normal distribution). Examples of random distributions are presented in Figure 2. In the case of bimodal distribution, distributions were separated (Matlab © *gmdistribution* object, Statistical Toolbox, Matlab ©, The Mathworks), and the upper score distribution could be used. In this report, a Gamma distribution model was used.

Figure 2. Type 1 Random Distributions for three datasets for A1 configuration. The light grey histogram represents random distribution of scores, black histogram is the actual subnetworks score distribution, and black cure is the gamma model.



Additionally, functional biology is inferred by measurement of enriched Gene Ontology terms. For each subnetwork over each dataset, Benjamini-Hotchberg corrected *p*-values were calculated for Gene Ontology terms (GO, The Gene Ontology Consortium, 2009) using an hypergeometric distribution. Version 2.1.18 of ErmineJ (Lee et al., 2005) was used (GO data downloaded on Sept 1st, 2009).

Constructing a Gene Signature

To construct a gene signature reflecting the subnetwork information, a list of discriminative genes must be extracted from the subnetworks set. Two metrics are available: co-occurrence (number of times a gene appears in the subnetwork set) and correlation (Pearson correlation of its expression profile with clinical condition vector *Cc*). However, some genes appear in several subnetworks and have a high occurrence rate, but a low discriminative power, whereas some genes have a lower occurrence rate but a high correlation with the studied clinical situation. Therefore, a ranked metric is needed to reflect both the number of gene occurrences in subnetworks as well as the relative discriminative power of each gene in each subnetwork. To equitably rank genes from these two situations, a so-called 'general rank' is computed as the average rank obtained with co-occurrence and correlation ranking. Co-occurrence ranking is computed by counting gene occurrence on a subnetwork set, and correlation ranking is produced by ranking genes according to GEP-*Cc* Pearson correlation (see section ITI Algorithm). For instance, *LUC7L3* is ranked only 38th in our signature according to its occurrence in the subnetworks, but ranked 5th by general ranking, as it belongs to the subnetwork ranked 1st. Genes ordered by these different metrics are reported in the database.

Several Training Sets were used to Test Subnetwork Stability

To understand the impact of each dataset on results and generalization of subnetworks, discriminating subnetworks for different combinations of input datasets and validation datasets from the breast cancer compendium were generated. Four combinations were used, named A1, A2, B1, and B2 (See Table 2). The combination run A1 is using all but van de Vijver datasets for input. Combination B1 is using all but Wang datasets for input, combination A2 is using datasets profiled on Affymetrix platforms, and B2 is using all Affymetrix datasets but Wang. For each run, subnetworks were validated using *p*-values computed section Subnetwork statistical validation. Subnetwork set constructed with A1 datasets was validated by

Table 3. P-value thresholds and consensus chosen for the three random distribution types for each input dataset configuration

Run	Type 1 *p*-values	Type 2 *p*-values	Type 3 *p*-values
A1	1.10^{-2} on at least 2 datasets	1.10^{-1} on at least 11 datasets	1.10^{-1} on at least 1 datasets
B1	1.10^{-1} on at least 8 datasets	1.10^{-2} on at least 2 datasets	1.10^{-1} on at least 1 datasets
A2	1.10^{-1} on at least 11 datasets	1.10^{-2} on at least 2 datasets	1.10^{-1} on at least 2 datasets
B2	1.10^{-1} on at least 6 datasets	1.10^{-2} on at least 2 datasets	1.10^{-1} on at least 1 datasets

keeping subnetwork meeting type 1 *p*-values of 1.10^{-2} over at least 8 datasets, type 2 *p*-values of 1.0^{-2} over at least 9 datasets, and type 3 *p*-values of 1.10^{-1} over at least 2 datasets, yielding a final set of 119 discriminative subnetworks containing 406 genes. Lower *p*-values yielded by type 3 random subnetworks are discussed in results section. Subnetworks with more significant *p*-values on individual datasets were present but not consistent over all datasets and thus not retained. *P*-values thresholds for A2, B1, and B2 configurations were summarized in Table 3.

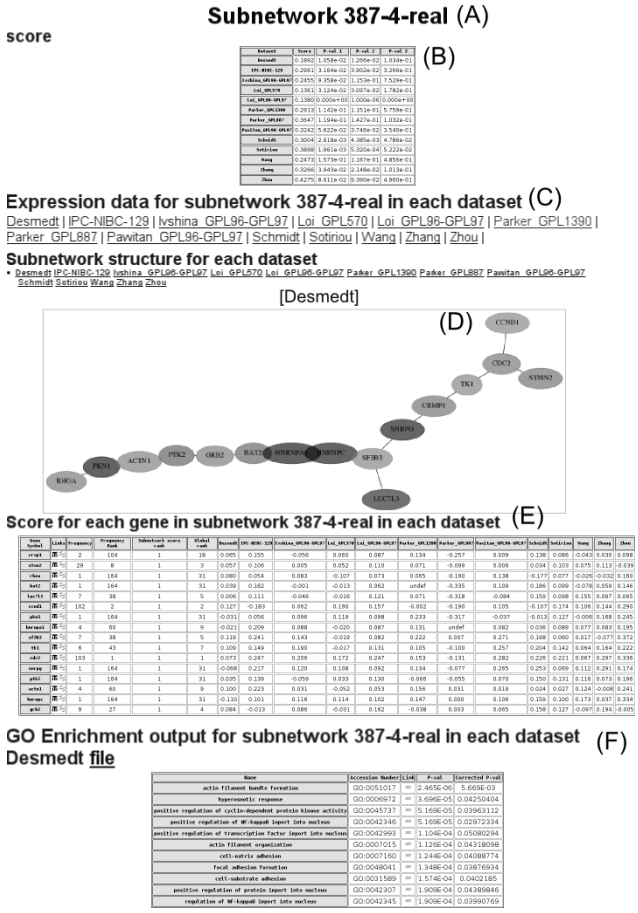
Examination of subnetworks found for each run (see ITI subnetwork Database in following section) shows little discrepancies among subnetwork sets found.

ITI Subnetwork Database

We stored the ITI algorithm results – discriminative subnetworks for a given clinical condition into a relational database, the ITI database. This database, publicly available on the web (http://bioinformatique.marseille.inserm.fr/iti), is the first one to explicitly link gene interaction models to disease, as opposed to many others, which store either plain gene lists, such as the Candidate List of yoUr Biomarker (CLUB, Lee et al., 2008), or subnetworks of interests but with no explicit link to the disease (CellCircuits database, Ideker Lab). CLUB database allows for sharing and comparison of putative candidate biomarkers, including lists of genes along with the protocol that obtained them. This database is cross-platforms to allow repository of studies made in heterogeneous platforms, such as proteomics, DNA microarrays, etc. Other gene list repositories exist, such as List Of Lists Annotated (LOLA) which allows inter-studies comparisons. CellCircuits contains network models extracted from different studies with little possibility of data mining in its current version. Its main interest is to have a platform for data sharing that allows searching for gene products and related Gene Ontology terms.

In our ITI implementation, the code produces web pages which can be placed on a plain web server and navigated through with a web browser. Web pages were generated for each runs, which allow qualitative comparison of subnetwork sets. Figure 3 shows information displayed by the database for each subnetwork. Briefly, each subnetwork is identified by a unique number (e.g. 387-4) and is described on a unique web page containing score and *p*-values for each dataset, heatmaps of expression data, subnetwork layout with superimposed correlation with the biological question asked, correlation and annotation for each gene contained in the subnetwork, and GO enrichment for each dataset. In parallel, one can find complete gene lists and annotations links to NCBI EntrezGene. A complete list of enriched Gene Ontology terms (Hypergeometric distribution) is also provided for the whole set of subnetworks. Figure 3 details the report page for subnetwork 387-4.

Figure 3. ITI Database Web Interface. (A) is the subnetwork ID (here 387-4), B is a table representing scores and p-values for each dataset, (C) is a series of links to gene expression heatmaps for each dataset, (D) shows the network topology as well as correlation (light grey= correlation, dark grey = anticorrelation) with clinical condition for each node. (E) is the individual genes score table, and (F) shows the top 10 enriched GO terms for each dataset, with Benjamini-Hotchberg-corrected p-values.



Dataset	Score	P-val_1	P-val_2	P-val_3
Desmedt	0.1892	1.858e-02	1.295e-02	1.038e-01
IPC-NIBC-129	0.2991	3.164e-02	3.802e-02	3.266e-01
Ivshina_GPL96-GPL97	0.2455	9.358e-02	1.353e-01	7.529e-01
Loi_GPL570	0.1361	3.124e-02	3.067e-02	1.782e-01
Loi_GPL96-GPL97	0.1380	0.000e+00	1.000e-06	0.000e+00
Parker_GPL1390	0.2913	1.142e-01	1.151e-01	5.759e-01
Parker_GPL887	0.3647	1.194e-01	1.427e-01	1.032e-01
Pawitan_GPL96-GPL97	0.3242	5.622e-02	3.745e-02	3.503e-01
Schmidt	0.3004	2.918e-03	4.385e-03	4.789e-02
Sotiriou	0.3898	1.961e-03	5.320e-04	5.222e-02
Wang	0.2473	1.573e-01	1.187e-01	4.855e-01
Zhang	0.3266	3.943e-02	2.148e-02	1.013e-01
Zhou	0.4275	8.611e-02	9.390e-02	4.960e-01

Gene Symbol	Links	Frequency	Frequency Rank	Subnetwork score rank	Global rank	Desmedt	IPC-NIBC-129	Ivshina_GPL96-GPL97	Loi_GPL570	Loi_GPL96-GPL97	Parker_GPL1390	Parker_GPL887	Pawitan_GPL96-GPL97	Schmidt	Sotiriou	Wang	Zhang	Zhou
crnn1		2	104	1	18	0.065	0.155	-0.056	0.060	0.087	0.134	-0.257	0.009	0.138	0.086	-0.043	0.036	0.098
ctnnd1		29	8	1	3	0.057	0.106	0.005	0.052	0.110	0.071	-0.099	0.006	0.034	0.103	0.075	0.113	-0.039
cfl1		1	164	1	31	0.080	0.054	0.083	-0.107	0.073	0.065	-0.190	0.138	-0.177	0.077	-0.026	-0.032	0.180
hat2		1	164	1	31	0.039	0.182	-0.001	-0.013	0.062	undef	-0.335	0.109	0.186	0.099	-0.078	0.059	0.146
lsm7l		7	38	1	5	0.006	0.111	-0.046	-0.016	0.121	0.071	-0.318	-0.084	0.159	0.098	0.155	0.067	0.065
ccnd1		102	2	1	2	0.127	-0.183	0.062	0.190	0.157	-0.002	-0.190	0.105	-0.107	0.174	0.108	0.144	0.290
pkm1		1	164	1	31	-0.031	0.056	0.086	0.119	0.098	0.233	-0.317	-0.037	-0.013	0.127	-0.006	0.168	0.245
hnrnpa1		4	68	1	9	-0.021	0.209	0.088	-0.020	0.087	0.131	undef	0.082	0.036	0.089	0.077	0.083	0.195
actb1		7	38	1	5	0.119	0.241	0.143	-0.019	0.082	0.222	0.007	0.271	0.168	0.060	0.017	-0.077	0.372
tk1		6	43	1	7	0.109	0.149	0.190	-0.017	0.131	0.105	-0.180	0.257	0.204	0.142	0.094	0.164	0.222
cdc2		103	1	1	1	0.073	0.247	0.206	0.172	0.247	0.153	-0.131	0.282	0.226	0.221	0.057	0.297	0.336
snrpg		1	164	1	31	-0.068	0.217	0.120	0.108	0.092	0.134	-0.077	0.265	0.253	0.069	0.112	0.291	0.174
ptk2		1	164	1	31	0.035	0.139	-0.059	0.033	0.130	-0.068	-0.055	0.076	0.150	0.131	0.116	0.073	0.180
actn1		4	68	1	9	0.100	0.223	0.031	-0.052	0.053	0.156	0.031	0.019	0.024	0.027	0.124	-0.008	0.241
hnrpc		1	164	1	31	-0.110	0.101	0.116	0.114	0.102	0.147	0.000	0.106	0.159	0.100	0.173	0.037	0.334
gnb2		9	27	1	4	0.084	-0.013	0.086	-0.031	0.162	-0.038	0.093	0.065	-0.158	0.127	-0.097	0.194	-0.005

Name	Accession Number	Link	P-val	Corrected P-val
actin filament bundle formation	GO:0051017	∞	2.465E-06	5.669E-03
hyperosmotic response	GO:0006972	∞	3.696E-05	0.04250404
positive regulation of cyclin-dependent protein kinase activity	GO:0045737	∞	5.169E-05	0.03963112
positive regulation of NF-kappaB import into nucleus	GO:0042346	∞	5.169E-05	0.02972334
positive regulation of transcription factor import into nucleus	GO:0042993	∞	1.104E-04	0.05080294
actin filament organization	GO:0007015	∞	1.126E-04	0.04318098
cell-matrix adhesion	GO:0007160	∞	1.244E-04	0.04088774
focal adhesion formation	GO:0048041	∞	1.348E-04	0.03876934
cell-substrate adhesion	GO:0031589	∞	1.574E-04	0.0402185
positive regulation of protein import into nucleus	GO:0042307	∞	1.909E-04	0.04398846
regulation of NF-kappaB import into nucleus	GO:0042345	∞	1.909E-04	0.03990769

Implementation and Code Availability

A Matlab © license and a Matlab Statistics Toolbox © license are necessary to compute *p*-values.

Code availability: the Perl code is available from the ITI database web site ([1]http://bioinformatique.marseille.inserm.fr/iti/iti-1.0.tar.gz), and has been licensed under the CeCILL public license (French GPL extension developed by a consortium of French Research Organisms).

RESULTS

We build a database containing links between interactome and metastatic relapse in breast cancer using results from the ITI algorithm applied on our breast cancer compendium. The ITI database will be extended over time, as we refine the algorithm and process data from other datasets from public repositories. In the meantime, we are using the database to understand the biology of the signature found under the form of a subnetwork set.

Biology of Extracted Subnetwork is Meaningful

Intrinsic biology of the 119 extracted subnetworks (containing 406 genes) from A1 combination was examined using annotation information from NCBI EntrezGene database and Gene Ontology Consortium database linked directly to the ITI database. First, the biology of each gene group included in a subnetwork was assessed by statistical enrichment of Gene Ontology terms (see section Methods and ITI web site). We found that subnetworks formed complexes functionally supporting the studied disease for metabolism, cell cycle control, proliferation, cell-cell adhesion and immunological response, which are known mechanisms of the hallmarks of cancer and metastatic process. The first ranked subnetwork (387-4, score S=0.283) shows significant enrichment for 'actin filament bundle formation' (GO: 0051017), which is a biological process linked to cell development and polarity. The second ranked subnetwork shows enrichment for 'activation of caspase activity by cytochrome C' (GO 0008635) which is linked to apoptosis. It also shows enrichment for 'B cell lineage commitment' (GO:0002326), which reveals immune response to metastasis. The third ranked subnetwork (2810-3) has a similar function (score S=0.278). Lower in the list, subnetwork 58-7 (Score S=0.271, ranked 6th) shows enrichment for functions related to microtubule formation: 'microtubule organizing center organization' (GO:0031023) term is significantly enriched. Functional 'regulation of centrosome cycle' is also significantly enriched (GO:0046605). Subnetwork 29959-4 (ranked 7th, score S=0.270) is functionally linked to metabolism, as seen with the terms 'glucose catabolic process' (GO:0006007), 'fructose metabolic process' (GO:0006000) and 'alditol metabolic process' (GO:0019400). This subnetwork is also functionally involved in cell migration through the formation of cell surface protrusions, such as lamellipodium or filopodium, at the leading edge of a migrating cell (Gene Ontology term 'substrate-bound cell migration, cell extension', GO:0006930). Subnetwork 581-7 (Score S=0.267) is involved in cell adhesion: 'focal adhesion formation' GO term is significantly enriched (GO:0048041). Cellular differentiation is functionally represented by subnetwork 1452-7 through enrichment of genes involved in the Wnt pathway signaling (GO term 'regulation of Wnt receptor signaling pathway' - GO:0030111). Subnetwork involved in cell proliferation is 5155-5 (S=0.254) having the GO terms 'positive regulation of endothelial cell proliferation' (GO:0001938) and 'establishment or maintenance of epithelial cell apical/basal polarity' (GO: 0045197) significantly enriched. Other subnetworks are of course of potential interest. Global list of enriched ontology terms is also stored in the database.

At the gene level, several markers show obvious links to cancer and involvement in cell cycle, proliferation, cell adhesion and other biological mechanisms involved in the disease. We examined the gene list ordered by 'mixed ranks' (see methods).

CDK1 (cyclin-dependent kinase 1) is the highest ranked gene. The protein encoded by this gene is a catalytic subunit of the highly conserved protein kinase complex known as M-phase promoting factor (MPF), which is essential for G1/S and G2/M phase transitions of the eukaryotic cell cycle (EntrezGene). Other genes from this process are also found, such as CCND1 (Cyclin D1, ranked

second). GRB2 (ranked 4th) is a growth factor associated with several cancer types and may have a role in metastasis (Yu et al., 2008). TK1 (thymidine kinase 1, soluble) is known as a proliferation marker in breast cancer, and its overexpression has been linked to thyroid carcinogenesis. TSC1 (tuberous sclerosis 1, ranked 8th) is known to play a central role in regulating cell survival and proliferation signaling pathways. Other genes of interest are present, including LAMA4 (laminin alpha-4), which has roles in *in vitro* migration and *in vivo* tumorigenicity of prostate cancer cells and others, and PGK1 (phosphoglycerate kinase 1), with proven involvement in prostate cancer, and many others.

Extracted Subnetworks Shows Identical Gene Expression Trend over Several Datasets

Figure 4 represents the top scoring subnetwork for A1 configuration: the subnetwork 387-4. It shows that gene expression of its components is consistent for several datasets (superimposition of 387-4 over Sotiriou, Finetti, Desmedt and Wang is shown) with high significance for most, showing a high power of discrimination and a high confidence of the correlation of gene expression of this subnetwork with the clinical condition. *P*-values obtained over several datasets were examined and are represented Figure 4. This subnetwork has significant scores for Finetti (p-values $<5.10^{-2}$), Sotiriou. (p-values $<=1.10^{-3}$) and Loi datasets (p-values $<=1.10^{-6}$) (see all obtained p-values Table 4). This subnetwork constitutes an interacting complex of proteins including the oncogene cyclin D1 (CCND1), and the Ras homolog gene family member (this protein may regulate the invasion and metastasis of breast cancer cells as an upstream signaling of Ezrin). The subnetwork also contains the Protein Tyrosine Kinase 2 (PTK2). It has been shown that increased PTK2 levels due to mutations of p53 are associated with breast and colon

Figure 4. Subnetwork 387-4 with correlation measured for each node and represented as light grey (correlated) or dark grey (anticorrelated). Global score for each dataset is shown as well as p-values, illustrating the expression consensus obtained for most genes within the network across different datasets.

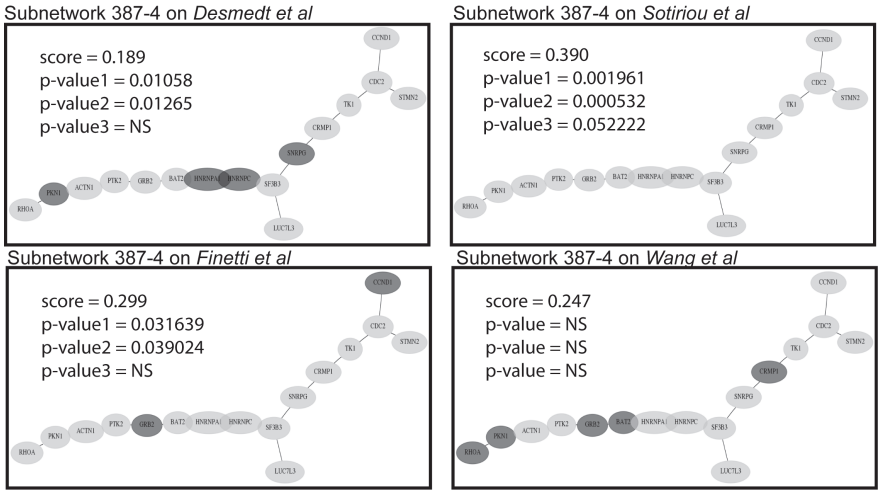


Table 4. Types 1, 2 and 3 p-values measured for subnetwork 387-4 for each dataset. P-values show high significance over several datasets, demonstrating the consensus of gene expression data over different datasets.

Dataset	Type 1 *p*-value	Type 2 *p*-value	Type 3 *p*-value
Desmedt	0.010584	0.012657	0.103381
IPC-NIBC-129	0.031639	0.039024	0.326597
Ivshina GPL96-GPL97	0.093584	0.115279	0.752917
Loi GPL570	0.031238	0.030974	0.178192
Loi GPL96-GPL97	$<1.10^{-6}$	1.10^{-6}	$<1.10^{-6}$
Parker GPL1390	0.114192	0.115112	0.575853
Parker GPL887	0.119422	0.142705	0.103176
Pawitan GPL96-GPL97	0.056216	0.037460	0.353977
Schmidt	0.002918	0.004385	0.047856
Sotiriou	0.001961	0.000532	0.052222
Wang	0.157285	0.116712	0.485642
Zhang	0.039429	0.021484	0.101313
Zhou	0.086106	0.093900	0.496010

cancers. Other subnetworks can be examined in the same way from the ITI database web interface.

CONCLUSION AND FUTURE RESEARCH DIRECTIONS

We present an Interactome-Transcriptome Integration framework (ITI) to isolate prognostic signatures generalizable over multiple datasets of breast cancer. We performed large scale integration of 15 DNA microarray datasets to create a breast cancer compendium. We also constructed a large coverage human interactome by integrating two existing human protein-protein interaction datasets (HPRD and Ramani dataset). These data, used conjointly with a discriminative subnetwork detection algorithm and significance scoring, allowed extraction of subnetworks linked with metastatic potential in breast cancers. Isolated subnetworks functionally cover biological functions related to metastasis and breast cancer, such as cell differentiation, cell cycle signaling, cell adhesion and proliferation, as well as functional links to immune response.

This database is the first of its kind to allow linking a human interactome to diseases or clinical situations. This resource can be mined for isolating potential drug targets as well as prognostic signatures for metastasis of breast cancer as well as other diseases. It has the potential of becoming the starting point to establish finer disease models by systems biology techniques. Improvement of public resources, such as extension of data repositories, refinement of platform annotations, and increased coverage of interaction data will help improve the resource. For instance, interaction data could be extended by inclusion of canonical pathways from the Kyoto Encyclopedia of Genes and Genomes (Kanehisa & Goto, 2000).

The ITI algorithm has also the potential of aiding in mining other large scale repositories such as ArrayExpress (Parkinson et al., 2008), and Stanford Microarray Database, (Hubble et al., 2009), and could be used conjointly with other technologies as well, such as proteomics (PRIDE database, Vizcaíno et al., 2009). Other diseases,

especially pathologies with limited number of available samples (prostate cancer for instance) are planned for further analysis with ITI.

Future developments include algorithm improvements, such as a better network parsing heuristics to find regions of interest. For instance, interactions with high reported confidence should have a higher probability to be included within a subnetwork than *in silico* predicted ones. Error rates in interaction data must also be taken into account for future developments, especially as coverage of interaction database grows. We are also considering integration with promoter specific methylation events, and genomic alteration data (Comparative Genomic Hybridization arrays).

An obvious extension of the presented framework will be the inclusion of a classification system (like Support-Vector Machine) to predict clinical outcome on independent data and to compare signature robustness with previous studies.

ACKNOWLEDGMENT

Research is funded by the Institut National du Cancer and the Institut National de la Santé et de la Recherche Médicale. Code development and calculation were performed on a Beowulf cluster funded by a grant from Fondation pour la Recherche Médicale. Maxime Garcia is funded by a Région Provence-Alpes-Côte d'Azur Fellowship. We thank Françoise Birg and Wahiba Gherraby for their suggestions for improving the manuscript.

REFERENCES

Anders, C. K., Acharya, C. R., Hsu, D. S., Broadwater, G., Garman, K., & Foekens, J. A. (2008). Age-specific differences in oncogenic pathway deregulation seen in human breast tumors. *PLoS ONE*, *3*(1), e1373. doi:10.1371/journal.pone.0001373

Barrett, T., Troup, D. B., Wilhite, S. E., Ledoux, P., Rudnev, D., & Evangelista, C. (2009). NCBI GEO: Archive for high-throughput functional genomic data. *Nucleic Acids Research*, *37*(Database issue), D885–D890. doi:10.1093/nar/gkn764

Bertucci, F., Finetti, P., Cervera, N., & Birnbaum, D. (2008). Prognostic classification of breast cancer and gene expression profiling. *Medecine Sciences*, *24*(6-7), 599–606.

Bidaut, G., & Stoeckert, C. J., Jr. (2009). Characterization of unknown adult stem cell samples by large scale data integration and artificial neural networks. *Pacific Symposium on Biocomputing*, 356-367.

Bidaut, G., & Stoeckert, C. J. Jr. (2009). Large scale transcriptome data integration across multiple tissues to decipher stem cell signatures. *Methods in Enzymology*, *467*, 229–245. doi:10.1016/S0076-6879(09)67009-9

Bidaut, G., Suhre, K., Claverie, J. M., & Ochs, M. F. (2006). Determination of strongly overlapping signaling activity from microarray data. *BMC Bioinformatics*, *7*, 99. doi:10.1186/1471-2105-7-99

Bild, A. H., Yao, G., Chang, J. T., Wang, Q., Potti, A., & Chasse, D. (2006). Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, *439*(7074), 353–357. doi:10.1038/nature04296

Breitling, R., Armengaud, P., Amtmann, A., & Herzyk, P. (2004). Rank products: A simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Letters*, *573*(1-3), 83–92. doi:10.1016/j.febslet.2004.07.055

Bueno-de-Mesquita, J. M., van Harten, W. H., Retel, V. P., van't Veer, L. J., van Dam, F. S., & Karsenberg, K. (2007). Use of 70-gene signature to predict prognosis of patients with node-negative breast cancer: A prospective community-based feasibility study (RASTER). *The Lancet Oncology*, *8*(12), 1079–1087. doi:10.1016/S1470-2045(07)70346-7

Campone, M., Campion, L., Roche, H., Gouraud, W., Charbonnel, C., & Magrangeas, F. (2008). Prediction of metastatic relapse in node-positive breast cancer: Establishment of a clinicogenomic model after FEC100 adjuvant regimen. *Breast Cancer Research and Treatment*, *109*(3), 491–501. doi:10.1007/s10549-007-9673-x

Chang, H. Y., Sneddon, J. B., Alizadeh, A. A., Sood, R., West, R. B., & Montgomery, K. (2004). Gene expression signature of fibroblast serum response predicts human cancer progression: Similarities between tumors and wounds. *PLoS Biology*, *2*(2), E7. doi:10.1371/journal.pbio.0020007

Chanrion, M., Negre, V., Fontaine, H., Salvetat, N., Bibeau, F., & MacGrogan, G. (2008). A gene expression signature that can predict the recurrence of tamoxifen-treated primary breast cancer. *Clinical Cancer Research*, *14*(6), 1744–1752. doi:10.1158/1078-0432.CCR-07-1833

Chuang, H. Y., Lee, E., Liu, Y. T., Lee, D., & Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Molecular Systems Biology*, *3*, 140. doi:10.1038/msb4100180

Conlon, E. M., Song, J. J., & Liu, J. S. (2006). Bayesian models for pooling microarray studies with multiple sources of replications. *BMC Bioinformatics*, *7*, 247. doi:10.1186/1471-2105-7-247

de Lavallade, H., Finetti, P., Carbuccia, N., Khorashad, J. S., Charbonnier, A., & Foroni, L. (2010). A gene expression signature of primary resistance to imatinib in chronic myeloid leukemia. *Leukemia Research*, *34*(2), 254–257. doi:10.1016/j.leukres.2009.09.026

Desmedt, C., Piette, F., Loi, S., Wang, Y., Lallemand, F., & Haibe-Kains, B. (2007). Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series. *Clinical Cancer Research*, *13*(11), 3207–3214. doi:10.1158/1078-0432.CCR-06-2765

Dobbin, K. K., Beer, D. G., Meyerson, M., Yeatman, T. J., Gerald, W. L., & Jacobson, J. W. (2005). Interlaboratory comparability study of cancer gene expression analysis using oligonucleotide microarrays. *Clinical Cancer Research*, *11*(2 Pt 1), 565–572.

Ein-Dor, L., Zuk, O., & Domany, E. (2006). Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(15), 5923–5928. doi:10.1073/pnas.0601231103

Fan, C., Oh, D. S., Wessels, L., Weigelt, B., Nuyten, D. S., & Nobel, A. B. (2006). Concordance among gene-expression-based predictors for breast cancer. *The New England Journal of Medicine*, *355*(6), 560–569. doi:10.1056/NEJMoa052933

Fishel, I., Kaufman, A., & Ruppin, E. (2007). Meta-analysis of gene expression data: A predictor-based approach. *Bioinformatics (Oxford, England)*, *23*(13), 1599–1606. doi:10.1093/bioinformatics/btm149

Fisher, R. A. (1925). *Statistical methods for research workers*. London: Edinburg.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., & Dudoit, S. (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biology*, *5*(10), R80. doi:10.1186/gb-2004-5-10-r80

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., & Mesirov, J. P. (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, *286*(5439), 531–537. doi:10.1126/science.286.5439.531

Goodsell, D. S. (1999). The molecular perspective: The ras oncogene. *The Oncologist*, *4*(3), 263–264.

Hong, F., & Breitling, R. (2008). A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics (Oxford, England)*, *24*(3), 374–382. doi:10.1093/bioinformatics/btm620

Hong, F., Breitling, R., McEntee, C. W., Wittner, B. S., Nemhauser, J. L., & Chory, J. (2006). RankProd: A bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics (Oxford, England)*, *22*(22), 2825–2827. doi:10.1093/bioinformatics/btl476

Hubble, J., Demeter, J., Jin, H., Mao, M., Nitzberg, M., & Reddy, T. B. (2009). Implementation of GenePattern within the Stanford Microarray Database. *Nucleic Acids Research*, *37*(Database issue), D898–D901. doi:10.1093/nar/gkn786

Irizarry, R. A., Warren, D., Spencer, F., Kim, I. F., Biswal, S., & Frank, B. C. (2005). Multiple-laboratory comparison of microarray platforms. *Nature Methods*, *2*(5), 345–350. doi:10.1038/nmeth756

Ivshina, A. V., George, J., Senko, O., Mow, B., Putti, T. C., & Smeds, J. (2006). Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Research*, *66*(21), 10292–10301. doi:10.1158/0008-5472.CAN-05-4414

Jezequel, P., Campone, M., Roche, H., Gouraud, W., Charbonnel, C., & Ricolleau, G. (2009). A 38-gene expression signature to predict metastasis risk in node-positive breast cancer after systemic adjuvant chemotherapy: A genomic substudy of PACS01 clinical trial. *Breast Cancer Research and Treatment*, *116*(3), 509–520. doi:10.1007/s10549-008-0250-8

Kaestner, K. H., Lee, C. S., Scearce, L. M., Brestelli, J. E., Arsenlis, A., & Le, P. P. (2003). Transcriptional program of the endocrine pancreas in mice and humans. *Diabetes*, *52*(7), 1604–1610. doi:10.2337/diabetes.52.7.1604

Keshava Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., & Mathivanan, S. (2009). Human Protein Reference Database-2009 update. *Nucleic Acids Research*, *37*(Database issue), D767–D772. doi:10.1093/nar/gkn892

Kreike, B., Halfwerk, H., Kristel, P., Glas, A., Peterse, H., & Bartelink, H. (2006). Gene expression profiles of primary breast carcinomas from patients at high risk for local recurrence after breast-conserving therapy. *Clinical Cancer Research*, *12*(19), 5705–5712. doi:10.1158/1078-0432.CCR-06-0805

Lee, B. T., Liew, L., Lim, J., Tan, J. K., Lee, T. C., & Veladandi, P. S. (2008). Candidate List of yoUr Biomarker (CLUB): A Web-based platform to aid cancer biomarker research. *Biomarker Insights*, *3*, 65–71.

Lee, H. K., Braynen, W., Keshav, K., & Pavlidis, P. (2005). ErmineJ: Tool for functional analysis of gene expression data sets. *BMC Bioinformatics*, *6*, 269. doi:10.1186/1471-2105-6-269

Li, L., Weinberg, C. R., Darden, T. A., & Pedersen, L. G. (2001). Gene selection for sample classification based on gene expression data: Study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics (Oxford, England)*, *17*(12), 1131–1142. doi:10.1093/bioinformatics/17.12.1131

Loi, S., Haibe-Kains, B., Desmedt, C., Wirapati, P., Lallemand, F., & Tutt, A. M. (2008). Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen. *BMC Genomics*, *9*, 239. doi:10.1186/1471-2164-9-239

Michiels, S., Koscielny, S., & Hill, C. (2005). Prediction of cancer outcome with microarrays: A multiple random validation strategy. *Lancet*, *365*(9458), 488–492. doi:10.1016/S0140-6736(05)17866-0

Miller, L. D., Smeds, J., George, J., Vega, V. B., Vergara, L., & Ploner, A. (2005). An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(38), 13550–13555. doi:10.1073/pnas.0506230102

Mook, S., Knauer, M., Bueno-de-Mesquita, J. M., Retel, V. P., Wesseling, J., & Linn, S. C. (2010). Metastatic potential of T1 breast cancer can be predicted by the 70-gene MammaPrint signature. *Annals of Surgical Oncology*, *17*(5), 1406–1413. doi:10.1245/s10434-009-0902-x

Munro, K. M., & Perreau, V. M. (2009). Current and future applications of transcriptomics for discovery in CNS disease and injury. *Neuro-Signals*, *17*(4), 311–327. doi:10.1159/000231897

Pages, F., Galon, J., Dieu-Nosjean, M. C., Tartour, E., Sautes-Fridman, C., & Fridman, W. H. (2009). Immune infiltration in human tumors: A prognostic factor that should not be ignored. *Oncogene*, *29*(8), 1093–1102. doi:10.1038/onc.2009.416

Paik, S., Tang, G., Shak, S., Kim, C., Baker, J., & Kim, W. (2006). Gene expression and benefit of chemotherapy in women with node-negative, estrogen receptor-positive breast cancer. *Journal of Clinical Oncology*, *24*(23), 3726–3734. doi:10.1200/JCO.2005.04.7985

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., & Vickery, T. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, *27*(8), 1160–1167. doi:10.1200/JCO.2008.18.1370

Parkinson, H., Kapushesky, M., Kolesnikov, N., Rustici, G., Shojatalab, M., & Abeygunawardena, N. (2009). ArrayExpress update-from an archive of functional genomics experiments to the atlas of gene expression. *Nucleic Acids Research*, *37*(Database issue), D868–D872. doi:10.1093/nar/gkn889

Pawitan, Y., Bjohle, J., Amler, L., Borg, A. L., Egyhazi, S., & Hall, P. (2005). Gene expression profiling spares early breast cancer patients from adjuvant therapy: Derived and validated in two population-based cohorts. *Breast Cancer Research*, *7*(6), R953–R964. doi:10.1186/bcr1325

Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., & Rees, C. A. (2000). Molecular portraits of human breast tumours. *Nature*, *406*(6797), 747–752. doi:10.1038/35021093

Ramani, A. K., Bunescu, R. C., Mooney, R. J., & 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. doi:10.1186/gb-2005-6-5-r40

Reyal, F., Stransky, N., Bernard-Pierrot, I., Vincent-Salomon, A., de Rycke, Y., & Elvin, P. (2005). Visualizing chromosomes as transcriptome correlation maps: Evidence of chromosomal domains containing co-expressed genes-a study of 130 invasive ductal breast carcinomas. *Cancer Research*, *65*(4), 1376–1383. doi:10.1158/0008-5472.CAN-04-2706

Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., & Canese, K. (2010). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *38*(Database issue), D5–D16. doi:10.1093/nar/gkp967

Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*, *270*(5235), 467–470. doi:10.1126/science.270.5235.467

Schmidt, M., Bohm, D., von Torne, C., Steiner, E., Puhl, A., & Pilch, H. (2008). The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Research*, *68*(13), 5405–5413. doi:10.1158/0008-5472.CAN-07-5206

Segal, E., Friedman, N., Koller, D., & Regev, A. (2004). A module map showing conditional activity of expression modules in cancer. *Nature Genetics*, *36*(10), 1090–1098. doi:10.1038/ng1434

Shi, L., Reid, L. H., Jones, W. D., Shippy, R., Warrington, J. A., & Baker, S. C. (2006). The MicroArray Quality Control (MAQC) project shows inter- and intra-platform reproducibility of gene expression measurements. *Nature Biotechnology*, *24*(9), 1151–1161. doi:10.1038/nbt1239

Sorlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., & Johnsen, H. (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

Sotiriou, C., Wirapati, P., Loi, S., Harris, A., Fox, S., & Smeds, J. (2006). Gene expression profiling in breast cancer: Understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute*, *98*(4), 262–272. doi:10.1093/jnci/djj052

The Gene Ontology Consortium. (2009). The gene ontology's reference genome project: A unified framework for functional annotation across species. *PLoS Computational Biology*, *5*(7), e1000431. doi:10.1371/journal.pcbi.1000431

van de Vijver, M. J., He, Y. D., van't Veer, L. J., Dai, H., Hart, A. A., & Voskuil, D. W. (2002). A gene-expression signature as a predictor of survival in breast cancer. *The New England Journal of Medicine*, *347*(25), 1999–2009. doi:10.1056/NEJMoa021967

van 't Veer, L. J., Dai, H., van de Vijver, M. J., He, Y. D., Hart, A. A., & Mao, M. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, *415*(6871), 530–536. doi:10.1038/415530a

van Vliet, M. H., Klijn, C. N., Wessels, L. F., & Reinders, M. J. (2007). Module-based outcome prediction using breast cancer compendia. *PLoS ONE*, *2*(10), e1047. doi:10.1371/journal.pone.0001047

Vizcaino, J. A., Cote, R., Reisinger, F., Foster, J. M., Mueller, M., & Rameseder, J. (2009). A guide to the Proteomics Identifications Database proteomics data repository. *Proteomics*, *9*(18), 4276–4283. doi:10.1002/pmic.200900402

Wachi, S., Yoneda, K., & Wu, R. (2005). Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics (Oxford, England)*, *21*(23), 4205–4208. doi:10.1093/bioinformatics/bti688

Wang, Y., Klijn, J. G., Zhang, Y., Sieuwerts, A. M., Look, M. P., & Yang, F. (2005). Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, *365*(9460), 671–679.

Wong, S. Y., Haack, H., Kissil, J. L., Barry, M., Bronson, R. T., & Shen, S. S. (2007). Protein 4.1B suppresses prostate cancer progression and metastasis. *Proceedings of the National Academy of Sciences of the United States of America*, *104*(31), 12784–12789. doi:10.1073/pnas.0705499104

Wu, Z., Zhao, X., & Chen, L. (2009). Identifying responsive functional modules from protein-protein interaction network. *Molecules and Cells*, *27*(3), 271–277. doi:10.1007/s10059-009-0035-x

Xu, L., Geman, D., & Winslow, R. L. (2007). Large-scale integration of cancer microarray data identifies a robust common cancer signature. *BMC Bioinformatics*, *8*, 275. doi:10.1186/1471-2105-8-275

Yu, G. Z., Chen, Y., Long, Y. Q., Dong, D., Mu, X. L., & Wang, J. J. (2008). New insight into the key proteins and pathways involved in the metastasis of colorectal carcinoma. *Oncology Reports*, *19*(5), 1191–1204.

Yu, K., Ganesan, K., Tan, L. K., Laban, M., Wu, J., & Zhao, X. D. (2008). A precisely regulated gene expression cassette potently modulates metastasis and survival in multiple solid cancers. *PLOS Genetics*, *4*(7), e1000129. doi:10.1371/journal.pgen.1000129

Zhang, Y., Sieuwerts, A. M., McGreevy, M., Casey, G., Cufer, T., & Paradiso, A. (2009). The 76-gene signature defines high-risk patients that benefit from adjuvant tamoxifen therapy. *Breast Cancer Research and Treatment*, *116*(2), 303–309. doi:10.1007/s10549-008-0183-2

Zhou, Y., Yau, C., Gray, J. W., Chew, K., Dairkee, S. H., & Moore, D. H. (2007). Enhanced NF kappa B and AP-1 transcriptional activity associated with antiestrogen resistant breast cancer. *BMC Cancer*, *7*, 59. doi:10.1186/1471-2407-7-59

ADDITIONAL READING

Alexe, G., Bhanot, G., Venkataraghavan, B., Ramaswamy, R., Lepre, J., & Levine, A. J. (2005). A robust meta-classification strategy for cancer diagnosis from gene expression data. *Proceedings / IEEE Computational Systems Bioinformatics Conference, CSB. IEEE Computational Systems Bioinformatics Conference*, 322–325.

Barrett, A. B., Phan, J. H., & Wang, M. D. (2008). Combining multiple microarray studies using bootstrap meta-analysis. *Conference Proceedings; ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference*, *2008*, 5660–5663.

Bertucci, F., & Birnbaum, D. (2009). Distant metastasis: not out of reach any more. *Journal of Biology*, *8*(3), 28. doi:10.1186/jbiol128

Cahan, P., Rovegno, F., Mooney, D., Newman, J. C., St Laurent, G. III, & McCaffrey, T. A. (2007). Meta-analysis of microarray results: challenges, opportunities, and recommendations for standardization. *Gene*, *401*(1-2), 12–18. doi:10.1016/j.gene.2007.06.016

Choi, H., Shen, R., Chinnaiyan, A. M., & Ghosh, D. (2007). A latent variable approach for meta-analysis of gene expression data from multiple microarray experiments. *BMC Bioinformatics*, *8*, 364. doi:10.1186/1471-2105-8-364

Choi, J. K., Choi, J. Y., Kim, D. G., Choi, D. W., Kim, B. Y., & Lee, K. H. (2004). Integrative analysis of multiple gene expression profiles applied to liver cancer study. *FEBS Letters*, *565*(1-3), 93–100. doi:10.1016/j.febslet.2004.03.081

DeConde, R. P., Hawley, S., Falcon, S., Clegg, N., Knudsen, B., & Etzioni, R. (2006). Combining results of microarray experiments: a rank aggregation approach. *Stat Appl Genet Mol Biol, 5*, Article15.

Dobbin, K. K., & Simon, R. M. (2007). Sample size planning for developing classifiers using high-dimensional DNA microarray data. *Biostatistics (Oxford, England)*, *8*(1), 101–117. doi:10.1093/biostatistics/kxj036

Dobbin, K. K., Zhao, Y., & Simon, R. M. (2008). How large a training set is needed to develop a classifier for microarray data? *Clinical Cancer Research*, *14*(1), 108–114. doi:10.1158/1078-0432.CCR-07-0443

Ma, S., & Huang, J. (2009). Regularized gene selection in cancer microarray meta-analysis. *BMC Bioinformatics*, *10*, 1. doi:10.1186/1471-2105-10-1

Park, T., Yi, S. G., Shin, Y. K., & Lee, S. (2006). Combining multiple microarrays in the presence of controlling variables. *Bioinformatics (Oxford, England)*, *22*(14), 1682–1689. doi:10.1093/bioinformatics/btl183

Pihur, V., & Datta, S. (2008). Finding common genes in multiple cancer types through meta-analysis of microarray experiments: a rank aggregation approach. *Genomics*, *92*(6), 400–403. doi:10.1016/j.ygeno.2008.05.003

Rhodes, D. R., Barrette, T. R., Rubin, M. A., Ghosh, D., & Chinnaiyan, A. M. (2002). Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer. *Cancer Research*, *62*(15), 4427–4433.

Warnat, P., Oberthuer, A., Fischer, M., Westermann, F., Eils, R., & Brors, B. (2007). Cross-study analysis of gene expression data for intermediate neuroblastoma identifies two biological subtypes. *BMC Cancer*, *7*, 89. doi:10.1186/1471-2407-7-89

KEY TERMS AND DEFINITIONS

Gene Signature: list of genes correlated with a given phenotype.

Interactome: Large scale gene interaction map that can be physical or functional, and inferred experimentally or by in silico analysis.

Meta-Analysis: integrated simultaneous analysis of multiple datasets.

Metastatic Relapse: Relapse of cancer after treatment with spreading to distant organs.

Robustness: Stability, Repeatability.

C.2 Article *Interactome–transcriptome integration*

Dans cet article intitulé "*Interactome–transcriptome integration for predicting distant metastasis in breast cancer*"[13], nous décrivons en détail l'algorithme ITI, ainsi que la façon dont on l'utilise pour réaliser une analyse supervisée.

Abstract
http://bioinformatics.oxfordjournals.org/content/28/5/672.abstract

Base de données des sous-réseaux
http://iti.sourceforge.net/supervised-5-datasets/index.html

Documentation
http://sourceforge.net/p/iti/wiki/Home/

Code Source
http://sourceforge.net/projects/iti/files/Source%20Code/iti-2.0.tar.gz

Site web compagnon
http://iti.sourceforge.net

BIOINFORMATICS **ORIGINAL PAPER**

Vol. 28 no. 5 2012, pages 672–678
doi:10.1093/bioinformatics/bts025

Gene expression

Advance Access publication January 11, 2012

Interactome–transcriptome integration for predicting distant metastasis in breast cancer

Maxime Garcia[1,2,3,4,*], Raphaelle Millat-Carus[1,2,3,4], François Bertucci[1,2,3,4], Pascal Finetti[1,2,3,4], Daniel Birnbaum[1,2,3,4] and Ghislain Bidaut[1,2,3,4,*]

[1]Aix-Marseille Univ, F-13284 Marseille, [2]Inserm, U1068, Centre de Recherche en Cancérologie de Marseille, [3]Institut Paoli-Calmettes and [4]CNRS, UMR7258, Centre de Recherche en Cancérologie de Marseille, F-13009 Marseille, France

Associate Editor: Trey Ideker

ABSTRACT

Motivation: High-throughput gene expression profiling yields genomic signatures that allow the prediction of clinical conditions including patient outcome. However, these signatures have limitations, such as dependency on the training set, and worse, lack of generalization.

Results: We propose a novel algorithm called ITI (interactome–transcriptome integration), to extract a genomic signature predicting distant metastasis in breast cancer by superimposition of large-scale protein–protein interaction data over a compendium of several gene expression datasets. Training on two different compendia showed that the estrogen receptor-specific signatures obtained are more stable (11–35% stability), can be generalized on independent data and performs better than previously published methods (53–74% accuracy).

Availability: The ITI algorithm source code from analysis are available under CeCILL from the ITI companion website: http://bioinformatique.marseille.inserm.fr/iti.

Contact: maxime.garcia@inserm.fr; ghislain.bidaut@inserm.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on October 10, 2011; revised on December 28, 2011; accepted on January 8, 2012

1 INTRODUCTION

The advent of post-genomic technologies provided the opportunity to potentially decipher the genomic origin of human diseases, including cancer. Thus, gene expression analysis using DNA microarrays allowed improving the classification and prognostication of several types of cancer, including breast cancer (Sørlie *et al.*, 2001; van de Vijver *et al.*, 2002). This approach can also help predict the metastatic recurrence and outcome (Wang *et al.*, 2005). In breast cancer (BC), the current prognostic features poorly reflect the heterogeneous clinical outcome. The consequence is that many patients (70–80%) receive unnecessary adjuvant systemic chemotherapy. Genomic tools could provide an opportunity to refine prognosis and improve treatment strategy and lay down foundation of personalized medicine in BC.

Several studies have produced signatures linked to BC distant metastasis (Sotiriou *et al.*, 2006). The Mammaprint 70-gene signature (van de Vijver *et al.*, 2002) classified BC patients in either good or poor prognosis groups. Wang *et al.* (2005) reported a 76 gene signature specific to estrogen receptor (ER) status (60 genes for ER+ patients, and 16 for ER−). These two signatures have an overlap of only three genes, which raised concerns about their reliability. Michiels *et al.* (2005) reanalyzed the van de Vijver dataset and concluded that the signatures obtained in such studies are unstable and dependent on the patients training set. From a classification standpoint, any gene classifier can be a good one as long as it performs and generalizes well (Dobbin *et al.*, 2008). However, from either a scientific or clinical standpoint, both the content and stability of signatures are of primary importance, so as to decipher their molecular basis, to reinforce their robustness and widespread acceptance of their routine clinical use and eventually to lead to new therapeutic targets.

Reasons for inherent instability of gene-based signatures have been previously enumerated (Bertucci *et al.*, 2006; Fan *et al.*, 2006). Besides experimental variability, variation in patient sampling and microarray platform bias, other reasons explain the lack of stability of signatures (Ein-Dor *et al.*, 2006). Reasons best explaining this instability are (i) the curse of dimensionality and (ii) the biological nature of gene expression measurements. The curse of dimensionality is well known of statisticians and is due to the inherent microarray data topology (too few samples for too many variables). The biological nature of instability is the following. Microarrays measure messenger RNA transcript abundance. To the extent that perturbations linked to a particular phenotype are reflected by changes in messenger RNA transcript levels, microarrays may be useful for measuring perturbations linked to a particular phenotype. Genes, however, are not independent but their products act in concert through protein–protein interaction (PPI) network(s). Our hypothesis is that phenotypes such as cancer result from isolated and subtle molecular perturbations (changes in gene expression and/or mutations for example) in driver genes that may provoke expression changes of greater amplitude in downstream genes (Chuang *et al.*, 2007). Statistics for differential expression detect changes of greater amplitude and reveal only these downstream genes. Superimposing an interaction network to expression changes can detect driver genes associated with more subtle expression changes (Chuang *et al.*, 2007). Factors cited previously may be more problematic for markers for subtle changes in differential expression, but we expect their effect to be attenuated by combination of several datasets. Such genes,

*To whom correspondence should be addressed.

© The Author 2012. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com

Downloaded from http://bioinformatics.oxfordjournals.org/ at INIST-CNRS on August 30, 2012

used as biomarkers, have proved to be more robust in predicting distant metastasis of breast tumors profiled on heterogeneous platforms than genes detected without network information. Several network-based approaches have been published for microarray analysis. They include generating condition-dependent networks on differential expression, where no prior information on interaction data is used, which somewhat limits the biological relevance of the results (Gill *et al.*, 2010). Co-clustering expression and graph data were proposed earlier by constructing a novel distance based on expression and network interactions (Hanisch *et al.*, 2002). Support vector machines (SVM) in combination with spectral decomposition data denoising was proposed for analyzing transcriptional response in yeast (Rapaport *et al.*, 2007). A network-based method was proposed to detect differentially expressed subnetworks in existing PPI data by local subnetwork aggregation (Chuang *et al.*, 2007). Using a stricter statistical framework, an SVM variation for directly using interaction data within a classifier was applied to microarray classification (Zhu *et al.*, 2009).

These methods addressed the biological issues mentioned before. However, the data dimensionality issue was still not taken into account because training and testing were done on a single dataset with a relatively low number of samples.

We propose here a multidataset re-implementation of the method proposed by Chuang *et al.* (2007) to integrate analysis of several gene expression datasets so as to extract subnetworks discriminating BC distant metastasis. We demonstrate the performance of our method, called interactome–transcriptome integration (ITI) on a large compendium of publicly available data. To avoid potential bias in subnetwork selection, we performed a stratified 10-fold cross-validation and combined the obtained networks. Validation was then done on two independent BC gene expression datasets (Desmedt *et al.*, 2008; van de Vijver *et al.*, 2002). Using this approach, we significantly increased the classification performance as compared with three previously published signatures while lowering the dependence of the signature on the training set. Independent classification on two studies by van de Vijver *et al.* (2002) and Wang *et al.* (2005) achieved 53 and 74% accuracy, respectively. We detail here our ITI algorithm and report statistical validation, patient classification results, as well as biological validation of the subnetworks thus defined.

2 METHODS

To detect protein complexes with subtle expression changes, we superimposed a large-scale PPI map to a compendium of BC expression profiles. The strategy implemented in ITI consists in detecting interactome subsets (subnetworks) whose expression is significantly correlated with distant metastasis-free survival (DMFS) in several datasets simultaneously. These subnetworks are then validated subsequently by shuffling interaction data and gene expression data. To train and test the system, six public datasets were chosen according to the criteria Section 2.2. Four analyses were performed [two different validation datasets held out for independent testing for Study 1 (Desmedt's dataset) and Study 2 (van de Vijver's dataset) and separate analysis according to positive or negative ER tumor status] to assess the impact of training data on the detected subnetworks and to understand their generalization capability. For each study, a 10-fold cross-validation was performed by carefully stratifying the training (90% of samples) and test sets. The aim of stratification was to properly balance each of the 10 training sets to keep the same ER+/ER− and DMFS event proportions in each of the 10 iterations.

2.1 PPI data integration

The following interaction datasets were used: Human Protein Resource Database release 9 (HPRD Keshava Prasad *et al.*, 2009), Molecular INTeraction database (MINT, Ceol *et al.*, 2010), INTAct (Aranda *et al.*, 2010), Database of Interacting Proteins (DIP, Salwinski *et al.*, 2004) and the human interactome generated *in silico* with the Cocite algorithm (Ramani *et al.*, 2005). All data were downloaded on September 8, 2010, and parsed to remove self-interactions, duplicates and proteins marked as 'unknown'. Self-interactions were removed from the files as they are not quantified by the algorithm, and interaction maps were integrated by uniqueness of NCBI gene ID accession numbers. Annotations were homogenized within datasets for proper display within the system. Resulting interactions obtained by crossing all interaction data totaled a number of 70 530 single interactions among a total of 13 202 proteins.

2.2 Breast cancer compendium

The public datasets (Table 1) were selected and included in a Breast Cancer Compendium (BCC) on the basis of the following criteria: early BC, availability of clinical information related to metastasis (event information and delay between the BC diagnosis, and the relapse or the last follow-up) and immunohistochemical ER status (ER+, ER−) and absence of post-operative adjuvant chemotherapy. A total of 930 tumors were retained for analysis from the initial pool of 1561 tumors through six datasets. Sampling size, platform types and ER status are detailed in Table 1. DMFS status was censored if follow-up was <5 years for all datasets.

Raw datasets were downloaded from National Center for Biotechnological Information (NCBI) Gene Expression Omnibus repository (Barrett *et al.*, 2009) when available, and normalized using the GCRMA method from Bioconductor (http://www.bioconductor.org/packages/release/bioc/html/gcrma.html). The van de Vijver dataset was downloaded as Supplementary Material from the publication (van de Vijver *et al.*, 2002). Datasets were collapsed from probe expression to gene expression as described in Reyal *et al.* (2005). When multiple probesets were available for a gene, we used the probeset having the highest median signal. Following this, 'nx_at' marked probes were removed. HG-U133A and HG-U133B were integrated as a virtual combined platform.

2.3 Dataset stratification, imbricated 10-fold cross-validation and independent testing

To detect discriminative subnetwork while avoiding over-fitting, cross-validation was performed by building training/testing sets while taking into account the clinical and molecular status of the tumors. Hence, stratification was done to balance ER+/ER− and distant metastasis rate between training and testing sets, leading to 10 randomly selected training sets. Preservation of both molecular and clinical status proportions in each dataset allowed increasing training and testing sets homogeneity and avoided molecular bias.

For each test/train set, subnetworks were detected with the ITI algorithm (Section 2.4) and validated by gene expression and PPI shuffling (Section 2.5), yielding five subnetwork lists. The lists were combined into a single signature (Section 2.6) whose discriminative power was tested on datasets held apart, as described in Section 2.7.

2.4 Interactome Transcriptome Integration—Constructing subnetworks

Each couple of training/testing set was searched for discriminative subnetworks whose average expression was linked to distant metastasis using the ITI algorithm. The latter is derived from the algorithm of .Chuang *et al.* (2007), with the added capability of detecting subnetworks on a compendium (Fig. 1). ITI was implemented as a pipeline developed with open source interpreters Perl and Bash and statistical validation was implemented with Matlab Statistical Toolbox R2010b [The Mathworks (c) Natick, MA, USA]. Subnetwork detection was parallelized and implemented on a Beowulf

Downloaded from http://bioinformatics.oxfordjournals.org/ at INIST-CNRS on August 30, 2012

M.Garcia et al.

Table 1. Datasets included in the BCC

Author(s)	GEO accession	Platform	Samples (FilteredInitial)	DMFS status (meta, non meta)	ER−/ER+
Desmedt *et al.* (2008)	**GSE7390**	**HG-U133A**	**190/198**	**62/128**	**61/129**
Sabatier *et al.* (2011)	GSE21653	HG-U133Plus2.0	31/255	9/22	11/20
Loi *et al.* (2008)	GSE6532	HG-U133A and B	101/327	27/74	29/72
Schmidt *et al.* (2008)	GSE11121	HG-U133A	182/200	46/136	37/145
van de Vijver *et al.* (2002)	**NA**	**Agilent HumanGenome**	**150/295**	**56/94**	**36/114**
Wang *et al.* (2005)	GSE2034	HG-U133A	276/286	107/169	72/204
Total: Six distinct sets	Six publicly available	Four distinct platforms	930/1561	307/623	246/684

Two trainings (Study 1 and 2) were performed using different combinations of training and testing data (on bold): On the Study 1, Desmedt tumors were held out for independent testing, and training was done on the rest. Van de Vijver dataset was respectively held out for the Study 2.

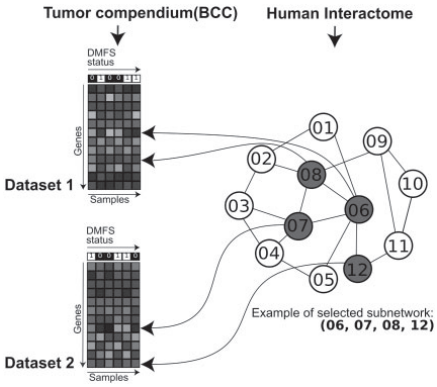


Fig. 1. ITI algorithm. Two data types were fed to the algorithm, the five training BC datasets and an interactome. Expression was simultaneously inspected on several datasets to aggregate discriminative subnetworks, i.e. discriminative regions in the interactome, as described in Section 2.4.

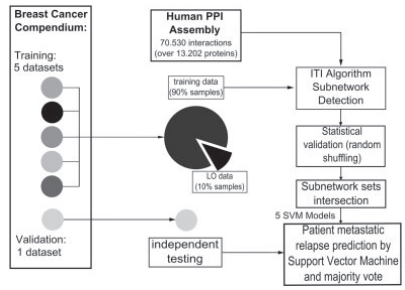


Fig. 2. Complete data workflow. Interactome is assembled from multiple sources (Section 2.1). Gene expression datasets forming our BCC assembled in Section 2.2 are pooled to form a training dataset. Ten groups were then formed on a 10% leave-out basis (Section 2.3). Subnetworks were detected on human interactome on each training set using ITI (Section 2.4) and validated twice by shuffling interactions and expression, as described in Section 2.5. Retained subnetworks were combined (Section 2.6) to train a SVM (Section 2.7). Final set was then used as a set of markers for classifying independent data by majority vote on the five SVM models (Section 2.7).

Cluster to reduce computing costs. Subnetworks visualization was obtained with the graph layout package GraphViz (AT&T Research, USA).

To detect discriminative subnetworks, correlations between clinical status and gene expression were computed for each dataset. Then, the interactome was exhaustively explored for discriminative regions (Fig. 1) by individually considering each node as a potential seed and aggregate recursively neighbors on the basis of a score measuring correlation of expression with DMFS status [Equation (1)]. Neighbors were aggregated until subnetwork score could not be improved above a certain threshold (improvement score threshold = 0.03). Then, the following node in interactome was processed. Parallelization was done by subdividing interactome over available scores. Subnetworks overlapping by >50% with already detected subnetworks were rejected. Overlap between subnetworks A and B was calculated by maximum inclusion score of subnetwork A in B and B in A. Inclusion score of A to B was measured by counting common genes included in subnetwork A to B and dividing by the total number of genes contained in subnetwork A. As an example, with a minimal threshold score of 0.3, analysis led to a total of 2986 subnetworks for Study 1 (ER+) −01 (run where the Desmedt dataset was held for independent testing, and subnetworks were detected on training stratification 01—see Fig. 2).

Each subnetwork was characterized by a score $S_{s,d}$ [Equation (1)] on each dataset measuring absolute correlation between the averaged subnetwork gene expression and the clinical information for this dataset. A global score S_s, defined by Equation (2), was computed by averaging individual scores over all datasets (not used for further computation).

$$S_{s,d} = \frac{\sqrt{n_d}}{\sqrt{\max\ n_d(\mathrm{DS})}} \left| \mathrm{corr}\left(\frac{1}{n} \sum\nolimits_{g \in s} e(g,d), cc(d) \right) \right| \quad (1)$$

$$S_s = \frac{1}{\mathrm{NS}} \sum\nolimits_{d \in \mathrm{DS}} S_{s,d} \quad (2)$$

S_s is the global score of subnetwork s, computed on the dataset d from the compendium (groups of datasets DS of size NS), corr is the Pearson's correlation measured between the averaged gene expression $e(g,d)$ for genes belonging to s with the binary vector cc containing labels linked to clinical conditions of patients in datasets d, weighted appropriately by the square root of the number n_d of samples in dataset d divided by the maximum number of samples in all datasets in DS.

2.5 Validating subnetworks—filtering

To validate subnetworks statistically, two random distributions of score were drawn. The first random distribution assessed the significance of algorithm that extracts subnetworks. It was obtained by randomly selecting subnetworks, i.e. by randomly accepting whether a subset of the interactome is a subnetwork without taking gene expression data into account. The second distribution assessed statistical significance of the biological link

Downloaded from http://bioinformatics.oxfordjournals.org/ at INIST-CNRS on August 30, 2012

Table 2. *P*-values thresholds and signature size for the four training configurations (Study 1 = all BCC but Desmedt, Study 2 = all BCC but van de Vijver)

Dataset	*P*-value threshold – *n* datasets	No. of subnetworks	No. of genes
Study 1 (ER−)	1e-4 – 2	165	2310
Study 1 (ER+)	1e-4 – 2	6	175
Study 2 (ER−)	1e-4 – 2	122	1481
Study 2 (ER+)	1e-4 – 2	14	272

The optimal number of subnetworks for classification depends on the training set and is lower for ER+ tumors, which reflects a higher homogeneity.

expression PPI. It was obtained by shuffling clinical conditions. To keep random subnetworks comparable to detected subnetworks, their distribution of size was forced to match that of the selected subnetworks by Gaussian modeling. Next, the distributions of random subnetwork scores were modeled by mixture of two Gaussian distributions. Once obtained, these distributions were used to fix score thresholds independently over all datasets at significance levels of *P*-values, and filtered statistically significant subnetworks. Shuffling random interactions to obtain a random interactome did not yield subnetworks at reasonable score levels, confirming the strong link between gene expression levels and protein–protein physical interaction(s). Finally, we kept only the subnetworks with a score higher than expected by chance both on subnetwork randomization ($P < 1.10^{-4}$ on two datasets) and shuffling of expression ($P < 1.10^{-4}$ on two datasets).

2.6 Constructing a common subnetwork signature for each training set

Using this filter, 10 subnetworks sets for each training tumor subset were generated. Next, these sets were combined by examining subnetworks pair by pair across datasets and combining them if overlap was larger than 50%. Using this method, clusters of overlapping subnetworks were built. Finally, a subnetwork list was constructed from the list of subnetwork clusters by keeping only subnetworks appearing at least twice. For a given cluster, only the subnetwork with the highest score was kept. Final subnetwork sets size are detailed in Table 2.

2.7 Tumor classification and distant metastasis prediction of ER+ and ER− tumors on two independent datasets

The subnetwork list obtained in Section 2.6 was used for independent classification using two different settings, namely Study 1 (in which Desmedt's data were held out) and 2 (in which van de Vijver's dataset were held out). In each setting, training was performed separately on tumors from all datasets except the held out dataset, yielding five SVM models. Classification on the validation sets was done by majority vote (weighted by sample size for each dataset) on the five SVM models. A complete organization chart is presented in Figure 2.

To use subnetworks as unique SVM input variables, gene expression within a subnetwork was averaged over genes and used as a discriminative profile for both training and testing. Several SVM models were tested for increasing number of subnetworks. A final subnetwork list was retained by maximizing accuracy.

Classification results (accuracy, true and false positives) are reported in Section 3, along with a comparison with previously published classifiers.

2.8 ITI on-line resource—Gene Ontology category enrichment

To detect pathways associated with BC distant metastasis, we computed enrichment of biological process gene ontology in each subnetwork detected by ITI using the ErmineJ program (Gillis *et al.*, 2010) and the reference list of Biological Process from Gene Ontology (Ashburner *et al.*, 2000). ErmineJ provided corrected *P*-values for enrichment of ontological terms computed with hypergeometric distribution. These were systematically computed for all subnetworks to associate them to known molecular processes defined in the Gene Ontology.

The resulting data were organized in a dedicated on-line resource (http://bioinformatique.marseille.inserm.fr/iti). This resource describes subnetworks detected with ITI and gives a thorough description of the included genes. Subnetworks and gene lists are downloadable for further processing. Subnetworks *P*-values calculated according to random distributions described in Section 2.5 were also included, along with combined Fisher scores (Hong and Breitling, 2008). Genes were annotated with direct NCBI EntrezGene links and links to other subnetworks are provided. To understand expression changes of genes included in subnetworks, color-coded gene graphs are provided, with correlation expression/DMFS status superimposed on subnetworks. The correlation score is provided for all datasets separately.

3 RESULTS

3.1 Establishment of two discriminative subnetworks sets (ER+ and ER−) from a joined compendium of 930 tumors

Two separate signatures were generated for ER+ and ER− BC subtypes for two studies. In Study 1, Desmedt's data (Desmedt *et al.*, 2008) were held out, and in Study 2, van de Vijver's data (van de Vijver *et al.*, 2002) were held out, as described in Section 2.3. Thus, four sets of subnetworks were assessed (Table 2).

The optimal signature size retained in Table 2 is the one that maximizes the average accuracy on the 10 training sets for each analysis. For the Study 1, discriminative subnetworks had an average score of 0.49 (ER+) and 0.54 (ER−) confirming the high correlation of co-expression and proximity in the PPI network. Signature size was respectively of 6 (ER+) and 165 subnetworks (ER−). For the Study 2, the ER+ signature yielded an optimal classification score on independent data for 14 subnetworks, and the ER− signature for 122 subnetworks. They correspond to lists of 175 (Study 1, ER+), 2310 (Study 1, ER−), 272 (Study 2, ER+) and 1481 (Study 2, ER−) genes, respectively, many genes being represented in several subnetworks. These numbers are larger than what is reported for other signatures. This suggests that we detected a large panel of genes significantly linked to distant metastasis, realistically reflecting both the biological footprint of metastasis and the scale of perturbations at the gene expression level. Redundancy of genes within subnetworks may be explained by the high connectivity of several hubs (for instance TP53), which makes them likely to be included in several subnetworks.

3.2 Classification results on independent data show superiority of subnetwork-based classification over independent gene signatures

To assess the performance of signatures constructed with ITI, we compared them with previously established signatures. The 128 probes Genomic Grade Index (GGI) (Sotiriou *et al.*, 2006), the

Downloaded from http://bioinformatics.oxfordjournals.org/ at INIST-CNRS on August 30, 2012

Table 3. Benchmark classification results comparison for ITI and other signatures on the two test datasets of Desmedt (Dt) and van de Vijver (vdV), for ER+ and ER− tumors

Status	ER−								ER+							
Dataset	Desmedt				van de Vijver				Desmedt				van de Vijver			
Signature	GGI	70 g	76 g	**ITI (165)**	GGI	70 g	76 g	**ITI (122)**	GGI	70 g	76 g	**ITI (6)**	GGI	70 g	76 g	**ITI (14)**
N	61	61	61	**61**	36	36	36	**36**	129	129	129	**129**	114	114	114	**114**
TN	6	0	14	**22**	3	2	12	**17**	63	28	53	**86**	57	39	50	**49**
FP	28	34	20	**12**	16	17	7	**2**	31	66	41	**8**	18	36	25	**26**
TP	23	27	9	**11**	14	17	8	**2**	21	25	25	**9**	20	32	22	**10**
FN	4	0	18	**16**	3	0	9	**15**	14	10	10	**26**	19	7	17	**29**
ACC	0.475	0.442	0.377	**0.541**	0.472	0.528	0.556	**0.528**	0.651	0.411	0.604	**0.736**	0.675	0.623	0.632	**0.518**
SV	0.852	1	0.333	**0.407**	0.823	1	0.471	**0.118**	0.600	0.714	0.714	**0.257**	0.512	0.821	0.564	**0.256**
SP	0.176	0	0.411	**0.647**	0.157	0.106	0.632	**0.895**	0.670	0.298	0.563	**0.915**	0.760	0.520	0.667	**0.653**

The four subnetworks sets defined in Section 2.3 were used to measure ITI classification performance, highlighted in bold. The following code was used: *N*, number of tumors to classify; TN, true negative; TP, true positive; FP, false positive; TP, true positive; ACC, accuracy; SV, sensitivity; SP, specificity; FPR, false positive rate. Subnetworks classification accuracy was superior to gene expression classification for metastasis prediction for Desmedt's dataset and around the same level for van de Vijver's dataset.

Mammaprint 70-gene signature (van de Vijver *et al.*, 2002) and the 76-gene ER status-specific signature (Wang *et al.*, 2005) were tested. Performance was measured on the same tumors (Desmedt and van de Vijver datasets), separately on ER+ and ER− tumors. The classification methods from the respective original publications were used for each signature. For van de Vijver's signature, distances to mean centroïds from relapse and non-relapse groups are calculated (van de Vijver *et al.*, 2002). For Wang's signature, a relapse score is calculated for each patient by a linear combination of gene expression weighed by standardized Cox's coefficients (Wang *et al.*, 2005). Because the GGI and Mammaprint signatures are probe-specific, the tests were done with the probes present in the test dataset. Results and performance measurements are detailed in Table 3. They show that ITI generalization performance is vastly superior to previously published signatures. The GGI classification showed the highest accuracy on the (47–68%) range, the 70 gene signature on the (41–62%) range and the 76 gene signatures on the (37–63%) range.

ITI gave a better accuracy as compared with the Wang signature on Desmedt's data (ER+); an accuracy of 74% (specificity of 92%) was obtained versus an accuracy of 60% (specificity of 56%) with the 76 gene signature. ITI gave superior results also on Desmedt's ER− tumors with an accuracy of 54% (specificity of 65%) versus an accuracy of 38% (specificity of 41%) for the Wang signature.

This held true for the Mammaprint 70 gene signature, which works mostly for van de Vijver patients. ITI showed an accuracy of 53% associated with a specificity of 90% on van de Vijver's data (ER−) and an accuracy of 52% with a specificity of 65% on van de Vijver's ER+ tumors. This performance is inferior to what was obtained on Study 1 and may reflect a bias toward Affymetrix induced by the training compendium. The Mammaprint signature had a lower performance of 41% on ER+ and of 42% on ER− Desmedt tumors. Similarly, ITI showed performance superiority over the GGI for ER− patients. Overall, ITI was able to generalize better with a lower accuracy bound of 52%.

On a different comparison basis, Chuang *et al.* (2007) achieved 48.8% accuracy on van de Vijver samples with training on Wang samples and 55.8% reciprocally.

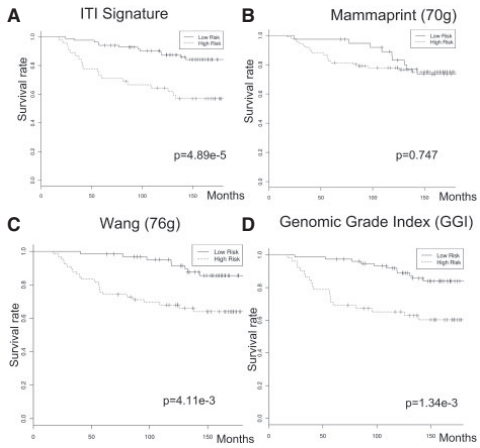


Fig. 3. Kaplan–Meier estimator of good prognosis (lower risk of distant metastasis) and poor prognosis groups (higher risk of distant metastasis) survival rates as defined by (**A**) ITI, (**B**) Mammaprint, (**C**) Wang's signature and (**D**) GGI for the ER+ Desmedt dataset. ITI gave the lowest *P*-value of $4.89.10^{-5}$ with a log-rank test among all tested signatures.

Specific contributions of the interaction data or gene expression data are not quantified, since they are not easily separable in the current setting. However, Chuang *et al.* (2007) already demonstrated that an expression approach increased signature robustness, and several studies showed that gene expression meta analysis also increased classification performance (Fishel *et al.*, 2007; Xu *et al.*, 2005).

We performed a survival analysis between good and poor prognosis groups in Study 1 (ER+) (Fig. 3). Log-rank test gave a *P*-value of 4.89×10^{-5}, suggesting good separation between the two groups. This is higher than *P*-values obtained with other

Downloaded from http://bioinformatics.oxfordjournals.org/ at INIST-CNRS on August 30, 2012

signatures (Wang signature gave $P=4.11\times10^{-3}$ and GGI gave $P=1.34\times10^{-3}$).

The Mammaprint signature was not able to separate Desmedt's patients in significant groups. Even though ITI was not specifically designed to obtain good log-rank score, it was able to separate patients with higher survival and patients with lower survival expectancy. An alternative could have been to compute subnetwork score directly on genes log-rank *P*-values.

3.3 Signatures obtained with ITI show a stability of 11.5–32.8% for different training sets

Wang and van de Vijver signatures have only three genes in common, which represent <5% of all the genes in the two signatures. We compared the two signatures obtained with ITI for ER+ and ER− samples with the Desmedt and van de Vijver tumors. A total of 937 genes were found in common between the Desmedt and the van de Vijver signature for ER− samples, and 46 genes between the Desmedt and van de Vijver signatures for ER+ samples. This represents an overlap of, respectively, 32.8% (ER−) and 11.5% (ER+). These relatively low values reflect the fact that datasets and platforms are biased. However, this is largely superior to the three common genes between the Wang and van de Vijver signatures. This overlap between subnetwork sets could probably be improved by using a larger training compendium.

3.4 Biology of the discriminative subnetwork set is meaningful

We examined the enriched annotations from the Gene Ontology biological process for the subnetworks obtained in Study 1. Table 4 shows several enriched GO terms for both ER+ and ER− signatures. Ontology terms found in discriminative subnetworks are linked to regulatory processes disrupted in cancer (cell cycle, DNA damage checkpoint) and in metastasis (immune system, cell proliferation, focal adhesion, cell migration and cytoskeleton organization) in both ER+ and ER− tumors.

As an example, we describe here a subnetwork significantly associated with metastasis in Study 1 (ER−) (subnetwork 6693, represented in Fig. 4). Subnetwork 6693 contained genes with well-known function in ER− BCs and metastasis, such as the tumor suppressor gene (TSG) TP53 and the tyrosine kinase receptors ERBB2 and EGFR.

The subnetwork contained also several cell cycle kinases and regulators (CDK2, CDKN1A, CDKN2A), NQO1, whose altered expression has been associated with various forms of cancer. PIN1 is present in the subnetwork, and was recently found to promote aggressiveness in BC. Insulin receptor was also present; its deregulated expression correlates with poor response to anti IGF-FR therapy in triple negative BC. It also contained several well-known oncogenes and genes not previously linked to cancer, but which may be acting as BC driver genes.

4 DISCUSSION

We conceived a network-based algorithm (ITI) to identify prognostic genomic signatures generalizable over multiple and heterogeneous microarray datasets. This algorithm works in two steps: first it integrates data from a compendium of BC microarray datasets, and second it finds subnetworks, i.e. interacting gene complexes, whose expression discriminates two conditions of interest. Subnetworks are filtered by statistical validation.

We applied the ITI algorithm to the particularly important but still unresolved question of finding markers for BC distant metastasis for which a large body of public data is available.

Our approach illustrates the feasibility of integrating gene expression data compendia (930 BC tumor samples were integrated) and large-scale PPI data; it represents a potential data mining tool for gene expression repositories. It features inclusion of prior data under the form of PPIs and clinical annotations.

We produced two ER status-specific signatures that were validated on independent datasets held out from training. Repeating the experiments for two datasets (Desmedt and van de Vijver) yielded higher classification performance than previously published classifiers in both cases [74% for Desmedt (ER+) and 53% for van de Vijver (ER+)]. Our subnetwork-based signatures reflect the large biological footprint of metastasis and is consequently larger that previously published signatures. The classifier obtained with ITI subnetworks was less sensitive to platform bias than previously

Table 4. Enriched Gene Ontology annotations of ER+ and ER− subnetworks

Gene Ontology	GO	Corrected *P*-value
ER+		
mRNA cleavage	GO:0006379	125E-08
Regulation of growth hormone secretion	GO:0060123	218E-07
Positive regulation of cytoskeleton organization	GO:0051495	206E-04
Regulation of insulin secretion	GO:0050796	155E-05
Regulation of chemotaxis	GO:0050920	429E-07
ER−		
Natural killer cell-mediated immunity	GO:0002228	293E-06
Positive regulation of MAP kinase activity	GO:0043406	476E-10
Muscle cell development	GO:0055001	106E-11
Interphase of mitotic cell cycle	GO:0051329	408E-11
Wnt receptor signaling pathway through β-catenin	GO:0060070	622E-10

Several enriched ontologies for subnetworks extracted in Study 1 (ER−) and Study 1 (ER+) studies are related to cancer.

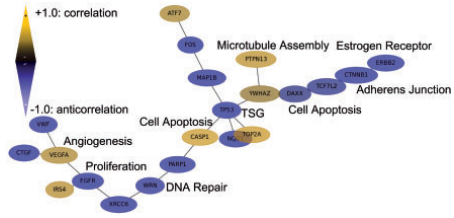


Fig. 4. Graphical representation of part of subnetwork 6693 (Study 1, ER−). This illustrates a discriminative subnetwork from the Sabatier and coworkers dataset. Nodes and edges correspond to genes encoding proteins and PPIs, respectively. Yellow and blue nodes denote an overexpression and an underexpression, respectively, among patients with distant metastasis compared with the other ones.

Downloaded from http://bioinformatics.oxfordjournals.org/ at INIST-CNRS on August 30, 2012

published classifiers, since performance obtained was similar on the two training compendia. It also showed high specificity, which is critical to make a decision on avoiding unnecessary adjuvant systemic treatment.

The ITI algorithm is currently extended to incorporate other data types, including DNA copy number variation data [SNPs, Comparative Genomic Hybridization arrays (CGH) and DNA methylation profiles]. ITI capability to handle the curse of dimensionality makes it suitable to detect biomarkers yielded by deep sequencing analysis. In next versions, PPI interaction type will also be taken into account at the interactome integration and subnetwork aggregation steps. Also, classification performance is inherently tied to molecular subtypes and finer subtyping is necessary to render this technology suitable for clinical use. A significant increase in ER− classification was observed by separating early and lately relapsing patients (data not shown). Further clinical validation could be envisioned through a phase-2 clinical trial with customized microarrays for adjuvant chemotherapy treatment decision making.

ACKNOWLEDGEMENTS

We thank Sabrina Carpentier for helpful discussions on the method, and Dr Françoise Birg and Wahiba Gherraby for proofing the manuscript.

Funding: Institut National du Cancer and Institut de la Santé et de la Recherche Médicale Grant 08/3D1616/Inserm-03-01/NG-NC (to G.B.); Ligue Nationale contre le Cancer (label D.B.). Support for the Beowulf cluster was obtained from Fondation pour la Recherche Médicale Young Team grant (To G.B.); Institut National de la Santé et de la Recherche Médicale - Région Provence-Alpes Côte d'Azur Doctoral Fellowship (to M.G.).

Conflict of Interest: none declared.

REFERENCES

Aranda,B. *et al.* (2010) The IntAct molecular interaction database in 2010. *Nucleic Acids Res.*, **38**, D525–D531.

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

Barrett,T. *et al.* (2009). NCBI GEO: archive for high-throughput functional genomic data. *Nucleic Acids Res.*, **37**, D885–D890.

Bertucci,F. *et al.* (2006) Gene expression profiling and clinical outcome in breast cancer. *OMICS*, **10**, 429–443.

Ceol,A. *et al.* (2010) MINT, the molecular interaction database: 2009 update. *Nucleic Acids Res.*, **38**, D532–D539.

Chuang,H.-Y. *et al.* (2007) Network-based classification of breast cancer metastasis. *Mol. Syst. Biol.*, **3**, 140.

Desmedt,C. *et al.* (2008) Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. *Clin. Cancer Res.*, **14**, 5158–5165.

Dobbin,K.K. *et al.* (2008) How large a training set is needed to develop a classifier for microarray data? *Clin. Cancer Res.*, **14**, 108–114.

Ein-Dor,L. *et al.* (2006) Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proc. Natl Acad. Sci. USA*, **103**, 5923–5928.

Fan,C. *et al.* (2006) Concordance among gene-expression-based predictors for breast cancer. *N. Engl. J. Med.*, **355**, 560–569.

Fishel,I. *et al.* (2007) Meta-analysis of gene expression data: a predictor-based approach. *Bioinformatics*, **23**, 1599–1606.

Gill,R. *et al.* (2010) A statistical framework for differential network analysis from microarray data. *BMC Bioinformatics*, **11**, 95.

Gillis,J. *et al.* (2010) Gene function analysis in complex data sets using ErmineJ. *Nat. Protoc.*, **5**, 1148–1159.

Hanisch,D. *et al.* (2002) Co-clustering of biological networks and gene expression data. *Bioinformatics* **18** (Suppl. 1), S145–S154.

Hong,F. and Breitling,R. (2008) A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics*, **24**, 374–382.

Keshava Prasad,T.S. *et al.* (2009) Human Protein Reference Database–2009 update. *Nucleic Acids Res.*, **37**, D767–D772.

Loi,S. *et al.* (2008) Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen. *BMC Genomics*, **9**, 239.

Michiels,S. *et al.* (2005) Prediction of cancer outcome with microarrays: a multiple random validation strategy. *Lancet*. **365**, 488–492.

Ramani,A.K. *et al.* (2005) Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biol.*, **6**, R40.

Rapaport,F. *et al.* (2007) Classification of microarray data using gene networks. *BMC Bioinformatics*, **8**, 35.

Reyal,F. *et al.* (2005) Visualizing chromosomes as transcriptome correlation maps: evidence of chromosomal domains containing co-expressed genes–a study of 130 invasive ductal breast carcinomas. *Cancer Res.*, **65**, 1376–1383.

Sabatier,R. *et al.* (2011) A gene expression signature identifies two prognostic subgroups of basal breast cancer. *Breast Cancer Res. Treat.*, **126**, 407–420.

Salwinski,L. *et al.* (2004) The Database of Interacting Proteins: 2004 update. *Nucleic Acids Res.*, **32**, D449–451.

Schmidt,M. *et al.* (2008) The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Res.*, **68**, 5405–5413.

Sørlie,T. *et al.* (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl Acad. Sci. USA*, **98**, 10869–10874.

Sotiriou,C. *et al.* (2006) Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J. Natl Cancer Inst.*, **98**, 262–272.

van de Vijver,M.J. *et al.* (2002) A gene-expression signature as a predictor of survival in breast cancer. *N. Engl. J. Med.*, **347**, 1999–2009.

Wang,Y. *et al.* (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, **365**, 671–679.

Xu,L. *et al.* (2005) Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data. *Bioinformatics*, **21**, 3905–3911.

Zhu,Y. *et al.* (2009) Network-based support vector machine for classification of microarray samples. *BMC Bioinformatics*, **10** (Suppl. 1), S21.

Downloaded from http://bioinformatics.oxfordjournals.org/ at INIST-CNRS on August 30, 2012

C.3 Chapitre *Large Scale Transcriptome-Interactome Integration*

Dans ce chapitre intitulé "*Detection of driver protein complexes in breast cancer metastasis by large scale transcriptome-interactome integration*"[14], nous détaillons point par point la façon dont il faut utiliser ITI pour réaliser une analyse supervisée.

Documentation
http://sourceforge.net/p/iti/wiki/Home/

Site web compagnon
http://iti.sourceforge.net

Author's Proof

Chapter 5

Detection of Driver Protein Complexes in Breast Cancer Metastasis by Large-Scale Transcriptome–Interactome Integration

Maxime Garcia, Pascal Finetti, Francois Bertucci, Daniel Birnbaum, and Ghislain Bidaut

Uncorrected Proof

Abstract

With the development of high-throughput gene expression profiling technologies came the opportunity to define genomic signatures predicting clinical condition or cancer patient outcome. However, such signatures show dependency on training set, lack of generalization, and instability, partly due to microarray data topology. Additional issues for analyzing tumor gene expression are that subtle molecular perturbations in driver genes leading to cancer and metastasis (masked in typical differential expression analysis) may provoke expression changes of greater amplitude in downstream genes (easily detected). In this chapter, we are describing an interactome-based algorithm, Interactome–Transcriptome Integration (ITI) that is used to find a generalizable signature for prediction of breast cancer relapse by superimposition of a large-scale protein–protein interaction data (human interactome) over several gene expression datasets. ITI extracts regions in the interactome whose expressions are discriminating for predicting relapse-free survival in cancer and allow detection of subnetworks that constitutes a generalizable and stable genomic signature (Garcia et al., Handbook of research on computational and systems biology: interdisciplinary applications, pp 406–427, 2012). In this chapter, we describe the practical aspects of running the full ITI pipeline (subnetwork detection and classification) on six microarray datasets.

Key words Bioinformatics, Systems biology, Interactome, Module analysis, Microarray classification, Support vector machine, Breast cancer

1 Introduction

High-throughput transcriptome measurement technologies (microarrays) have been in use for many years to decipher links between molecular activity and disease outcome. They have been used in all areas of biology as a high-throughput technology to measure the expression of thousands of genes in order to observe the variation of their expression under different experimental conditions for different clinical status, in the case of patient samples. This versatile technology

Michael F. Ochs (ed.), *Gene Function Analysis*, Methods in Molecular Biology, vol. 1101,
DOI 10.1007/978-1-62703-721-1_5, © Springer Science+Business Media, LLC 2014

Author's Proof

was then applied to cancer, and extensively in Breast Cancer (BCa). Several studies have proven the link between disease outcome and gene expression [2], or metastatic relapse and gene expression [16].

Practically, a microarray consists in spotting, in an ordered way, identified fragments (called probes, of up to 70 oligonucleotides) of DNA on a support (silicon chips are used for high-density oligonucleotide microarrays), each fragment being associated with a single known gene. Sample mRNA is isolated, labeled, and hybridized to the DNA immobilized on the chip. Chip images are then acquired by scanning the chip with a laser scanner. Spot intensity is then measured and quantified for each probe, leading to a datasheet of expression for the whole chip.

Several bioinformatics low-level steps are then required to produce data that is interpretable by biologists and clinicians. For data generated on an Affymetrix platform (Santa Clara, CA, USA), the open-source *affy*, *oligo*, and *gcrma* packages from Bioconductor (http://www.bioconductor.org/) can perform the essential transformation steps from raw data to a gene expression table, such as data parsing, probe combination, and normalization.

At this point, it seems that microarrays are an ideal tool for understanding molecular processes and predicting patient outcome in clinical settings and that standard bioinformatics tools exist to interpret them. However, recent studies have demonstrated several drawbacks in microarray data classification and gene signatures. Signatures appear to generalize quite well [6]. However, it has been established that they are heavily dependent on the sample set used for training [11], and that a large number of signatures could easily perform as good as or better than the 70-gene signatures generated by van de Vijver et al. [16].

Signatures drawn from microarray analysis of patient samples have two fundamental flaws. First, the inherent topology of microarray analysis suffers from a major concern regarding curse of dimensionality. Typical experiments only include 50–300 patients (in the best cases) over platforms measuring 40K+ variables (probes). Many of these variables are not independent and provide a global picture of a complete interacting network of genes. However, this brings the second fundamental flaw, which is biological. The inherent nature of the data measured by microarrays is inherently unstable. The reason behind this instability is the following: Microarrays measure abundance of messenger RNA, under the hypothesis that phenotypic changes are reflected in the messenger transcript levels. Genes are not independent and act in concert through the interactome. In cancer, the hypothesis is that some phenotypes are the result of a subtle change (small expression change or mutation) in several driver genes that provoke changes on a large scale in the whole interactome [4, 7]. Typical gene expression differential analysis is not designed to isolate the genes that are provoking the disease. They rather produce a statistic that has the tendency to rank first the most differentially expressed genes, i.e., the genes that are downstream of the drivers.

Uncorrected Proof

Author's Proof

To detect driver genes connected to a particular phenotype in a microarray experiment, Chuang et al. proposed to structure the whole analysis from the human interactome [4]. They superimposed gene expression on a large human protein–protein interaction (PPI) map to extract differentially expressed subnetworks to detect patients who had metastasis. These subnetworks have the property of being differentially expressed for two conditions that we wish to separate. In Garcia et al., we proposed to increase the statistical power of this approach by integrating several BCa datasets and to use the expression of a larger set of patients to decipher differentially expressed subnetworks [7, 8]. The proposed algorithm is called Interactome–Transcriptome Integration (ITI). Its basic principle of action is detailed in Fig. 1.

this figure will be printed in b/w

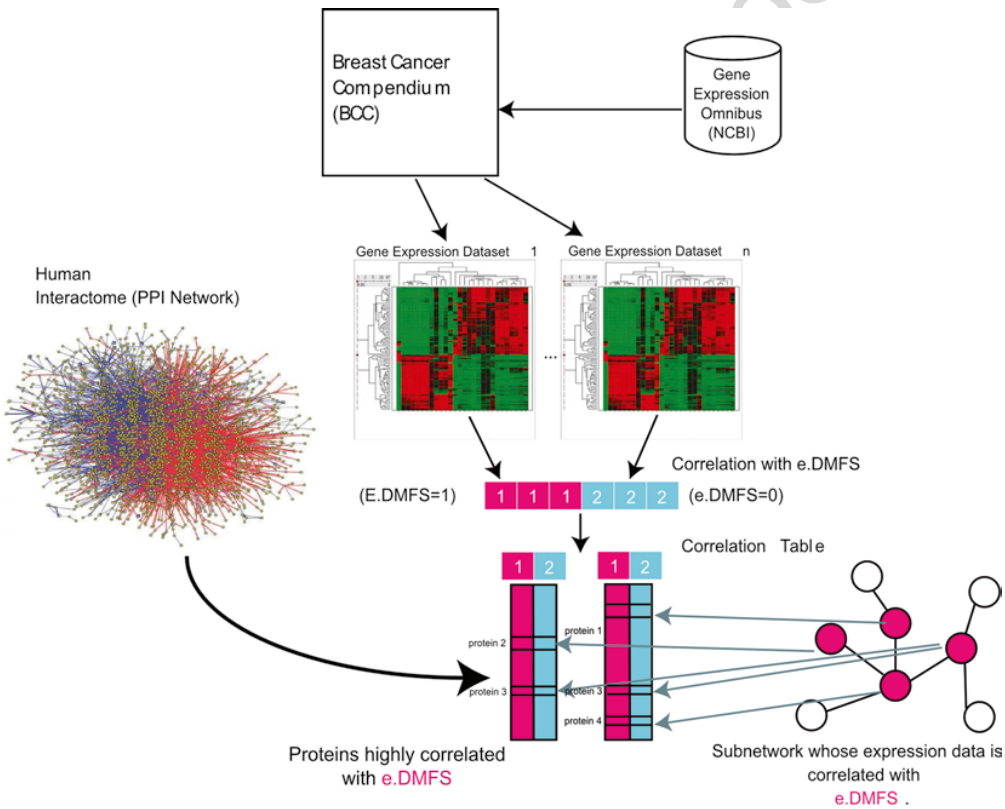


Fig. 1 ITI basic principle. This is the basic working principle of subnetwork detection with ITI. A Breast Cancer Compendium (BCC) is constructed from local data (Institut Paoli Calmettes [14]) and publicly available datasets from Gene Expression Omnibus Repository (GEO) as a set of tumor profiles (*see* Table 2—only two datasets are represented in this diagram). An interactome is constructed from several public protein–protein interaction (PPI) databases. Gene expression is then correlated with clinical condition (in this example, the Distant Metastasis Free Survival event). The interactome is then searched for interacting sets of genes whose expression is globally correlated to DMFS event in one or several datasets

Author's Proof

Maxime Garcia et al.

In this chapter, we detail the specific steps to apply ITI on five datasets spanning 900 samples over an interaction map of 65,000 interactions to separate good prognosis and bad prognosis groups in ER+ patients. To prove the robustness of ITI, we test its classification performance on an independent dataset kept aside during training [5].

2 Materials

2.1 Prerequisites on the Computing Environment: Beowulf Cluster

We have several prerequisites on the computing environment itself. ITI uses a large amount of computing resources, including memory, disk space, and CPU power. We implemented it on a Beowulf cluster configured with 14 nodes. Each node is a Bull R420 with a dual 3 GHz Intel processor (8 cores), 16GB RAM (2GB/core) connected to the head node server through a standard Gigbit Ethernet switch. The cluster nodes are running Linux CentOS 5.6 on the stock kernel. The head node is a Bull R460 server configured with a dual Intel CPU (8 cores), 24GB RAM, and 2TB storage shared to the nodes over NFS. Storage is a RAID 5 build on a set of SATA disk using the internal server bays. It is formatted with the standard ext3 filesystem, providing reasonably fast access to data. The head node runs CentOS 6.0 on the stock kernel. /home and /opt are shared to all nodes with NFS for access to data and ITI scripts.

Several pieces of software are required to run ITI. Some are provided as standard under most Linux distributions, such as Perl, Bash, and standard binaries. Third party software is also needed, such as Matlab (The Mathworks, Inc., Natick, MA, USA) and the Statistical toolbox, GraphViz (AT&T Labs Research, Florham Park, NJ, USA), LibSVM (version 2.9, http://www.csie.ntu.edu.tw/~cjlin/libsvm/), and ErmineJ [9].

Specific job queue management software has to be up and running. We are using the Sun Grid Engine 6.0, but any PBS-compatible program should be fine. As an example, the *qhost* command on our server returns:

```
$ qhost
HOSTNAME   ARCH   NCPU   LOAD   MEMTOT   MEMUSE   SWAPTO   SWAPUS
-------------------------------------------------------------------------------
global     -             -     -      -        -        -      -
frontal    lx24-amd64    8     -      23.5G    -        2.0G   -
node001    lx24-amd64    8     0.00   15.7G    1012.8M  2.0G   0.0
node002    lx24-amd64    8     0.00   15.7G    981.7M   2.0G   0.0
node003    lx24-amd64    8     0.00   15.7G    986.0M   2.0G   0.0
node004    lx24-amd64    8     0.00   15.7G    989.1M   2.0G   0.0
node005    lx24-amd64    8     0.00   15.7G    996.9M   2.0G   0.0
node006    lx24-amd64    8     0.00   15.7G    988.8M   2.0G   0.0
```

Uncorrected Proof

Author's Proof


```
node007   lx24-amd64   8    0.00   15.7G   989.2M    2.0G    0.0
node008   lx24-amd64   8    0.00   15.7G   990.1M    2.0G    0.0
node009   lx24-amd64   8    0.00   15.7G   974.6M    2.0G    0.0
node010   lx24-amd64   8    0.00   15.7G   956.3M    2.0G    0.0
node011   lx24-amd64   8    0.00   15.7G   933.6M    2.0G    0.0
node012   lx24-amd64   8    0.00   15.7G   936.0M    2.0G    0.0
node013   lx24-amd64   12   0.00   70.7G   456.8M    2.0G   549.4M
node014   lx24-amd64   12   0.00   70.7G     1.2G    2.0G   340.8M
```

2.2 Get and Install ITI

ITI is a suite of Perl and bash scripts. A tar archive of all ITI version 2.0 scripts and data can be downloaded from the ITI wiki: http://iti.sourceforge.net/download/index.html.

First, make a dedicated subdirectory in your home space:

```
$ mkdir iti-main
$ cd iti-main
```

Then, download and untar the ITI source code with the following command line:

```
$ wget http://sourceforge.net/projects/iti/files/Source%20Code/iti-2.0.tar.gz
$ tar xvzf iti_2.0.tar.gz
```

2.3 ITI Distribution

The ITI distribution is structured as follows:

```
$ ls -l iti-20
-rw-r--r- 1 bidaut bidaut 91799 févr. 10 2010 CeCILL_V2_en.pdf
-rw-r--r-- 1 bidaut bidaut 93627 févr. 10 2010 CeCILL_V2_fr.pdf
drwxrwxr-x 3 bidaut bidaut 4096 août 26 2010 css
drwxrwxr-x 2 bidaut bidaut 4096 oct. 6 2011 icons
drwxrwxr-x 2 bidaut bidaut 4096 oct. 6 2011 javascript
drwxrwxr-x 2 bidaut bidaut 4096 mai 7 11:31 others
drwxrwxr-x 2 bidaut bidaut 4096 mai 7 10:38 pipeline
-rw-rw-r-- 1 bidaut bidaut 6938 mai 7 11:31 readme.txt
```

The readme.txt file contains CeCILL license agreement and directory structure details. The detailed CeCILL license text is available in the pdf file. The pipeline directory contains the main ITI scripts. The others directory contains the master scripts that are calling the pipeline scripts. The *css*, *icons*, and *javascripts* subdirectories contain files used for the ITI Database (ITIDB) Web site.

Uncorrected Proof

Author's Proof

2.4 Annotation Data

Three types of annotation data are needed.

- The datasets sample annotation: provided in part from Gene Expression Omnibus, and in part from a dedicated Web site for the van de Vijver dataset.
- The datasets platform annotation, i.e., Platform types and probes annotations. These are provided by the Resourcerer site. All annotation files are available from the Resourcerer FTP server (ftp://ftp.tigr.org/pub/data/tgi/Resourcerer).
- Various gene annotation data files. We use the human gene_info.gz and the gene2go.gz flat files provided by the National Center for Biotechnology Information (NCBI) available from ftp://ftp.ncbi.nih.gov/gene/DATA/GENE_INFO/Mammalia/Homo_sapiens.gene_info.gz

For the sake of simplicity and reproducibility of results for this chapter, we compiled all microarray annotation files (from Resourcerer) in a single archive. To download it, issue the following commands:

```
$ wget http://sourceforge.net/projects/iti/files/Book%20Chapter%20Supplementary%20Material/resourcerer.tar.gz
```

Uncompress it with:

```
$ tar xvzf resourcerer.tar.gz
```

NCBI annotations are available with:

```
$ wget http://sourceforge.net/projects/iti/files/Book%20Chapter%20Supplementary%20Material/annotations-iti-ncbi.tar.gz
$ tar xvzf annotations-iti-ncbi.tar.gz
```

2.5 Expression Data: Construction of a Breast Cancer Compendium

All expression data was downloaded from the public repository Gene Expression Omnibus, with the exception of the van de Vijver dataset, obtained from the author's Web site (http://bioinformatics.nki.nl/data.php). All Affymetrix datasets were initially downloaded as raw data, renormalized with Bioconductor GCRMA, and the resulting "exprs" object containing expression measurement for individual probes were saved on disk under tab-delimited format. The following commands were used:

```
$ cd <CEL_directory>
$ R
(under R prompt)
> library(affy)
> library(gcrma)
> d=ReadAffy()
> e=gcrma(d)
> write.exprs(e, file="Desmedt_gcrma.txt")
> q()
```

Uncorrected Proof

Author's Proof

t1.1 **Table 1**
t1.2 **Platforms details for Gene expression datasets used in the study**

Dataset	NCBI accession number (if available)	Platform	Number of samples before filtering	Number of samples after filtering	Patient follow-up
Desmedt[a]	GSE7390	U133A	198	198	Yes
Finetti		U133 Plus 2.0	129	129	Yes
Loi	GSE6532	U133A + U133B U133 Plus 2.0	327 87	293 87	Yes
Schmidt	GSE11121	U133A	200	200	Yes
van de Vijver		Agilent whole human genome	295	295	Yes
Wang	GSE2034	U133A	286	286	Yes

t1.14 [a]Desmedt is used for independent testing

For sake of reproducibility and ease we generated a complete archive containing all normalized expression datasets. To download and uncompress, please issue the following commands:

```
$ wget http://sourceforge.net/projects/iti/files/Book%20Chapter%20Supplementary%20Material/breast_cancer_2012.tar.gz
$ tar xvzf breast_cancer_2012.tar.gz
```

All expression datasets are detailed in Table 1.

2.6 Protein–Protein Interaction Data

To build the largest possible interaction dataset, the following publicly available interaction sets were used:

- Human Protein Reference Database release 9 (HPRD, [10]).
- The Molecular Interaction database [3].
- INTAct [1].
- The Database of Interacting Protein [15].
- The database generated in silico with the Cocite Algorithm [12].

All PPI datasets are detailed in Table 2.

To be usable by ITI, all data was parsed and slightly adapted. All self-interactions were removed, as well as all duplicated interactions. All interactions involving unidentified proteins (marked as *Unknown*) were also removed. All parsing scripts are placed under the "parsing" ITI distribution. The parsing is not detailed here.

All interaction dataset files must be placed in a single directory whose path is given as an argument to ITI. Dataset files are tab-delimited and must be formatted as follows.

Uncorrected Pro

Author's Proof

t2.1 **Table 2**
t2.2 **This table represents the details of all protein–protein interaction (PPI) dataset used for training**

t2.3 t2.4 Resource	Number of proteins	Number of binary interactions	Nature
t2.5 HPRD [Human Protein t2.6 Resource Database] [10] t2.7	9,386	36,577	Y2H In vitro In vivo
t2.8			
t2.9 Cocite [12]	6,349	15,705	In silico [Cocite algorithm]
t2.10 DIP [Database of Interacting t2.11 Proteins]	918	810	In vitro/manually curated
t2.12 MINT [Molecular Interactions t2.13 Database] [3]	5,559	12,143	Manually curated from literature
t2.14 INTAct t2.15	7,471	25,616	Large-scale assays (Y2H, CoIP, pull-down)
t2.16 Total	13,203	70,530	3 Types

248 `geneID1<tab>Gene_Symbol1<tab>`
249 `GeneID2<tab>Gene_Symbol2<tab>Interaction_`
250 `Annotations`
251 The last field of annotation interaction is an optional field that
252 can contain additional information on the nature of interaction
253 (not used in ITI 2.0). Data is verified later on during parsing.
254 Again, ready to be parsed example data can be downloaded as
255 follows:
256 `$ wget http://sourceforge.net/projects/iti/`
257 `files/Book%20Chapter%20Supplementary%20`
258 `Material/interactomes-2012.tar.gz`
259 `$ tar xvf interactomes-2012.tar.gz`
260 A total of 70,530 interactions among 13,202 proteins are
261 available in the integrated PPI dataset.

2.7 Initial Set Up

262 This consists mostly in verifying the presence of all data elements
263 (annotation, scripts) and setting the proper path to data and script
264 directories.
265 Initial scripts must be downloaded by
266 `$ wget http://sourceforge.net/projects/iti/`
267 `files/Book%20Chapter%20Supplementary%20`
268 `Material/iti-study.tar.gz`
269 `$ tar xvzf iti-study.tar.gz`
270 Scripts must be put in the current iti-main directory:
271 `$ cp iti-study/* .`

Uncorrected Proof

Author's Proof

The ITI output directory has to be created. Since this study is concerned with an analysis on Desmedt's dataset, it will be called:

```
$ mkdir runs-Desmedt
```

Scripts have to be edited for correct paths:

- In the script, generateTranscriptomeConds.sh the following variables have to be set.

```
itiPath=/home/bidaut/iti-main/iti-2.0
transcriptomedataPath=/home/bidaut/iti-main/breast-cancer-1007
```

The script has to be executed:

```
$ ./generateTranscriptomeConds.sh
```

- In the script generateAllArgs.pl the following paths have to be properly edited: They correspond to the data that was previously downloaded.

The first variable specifies the first meta script path.

```
my $scriptExecutableName = "/home/bidaut/iti-main/iti-2.0/others/generateScriptsList1.sh";
```

The outDir variable specifies the path where all resulting files are stored.

```
my $outDir = "./runs-Desmedt";
```

This is the ITI scripts path.

```
my $binPath = "/home/bidaut/iti-main/iti-2.0/";
```

This specifies the interaction data directory

```
my $interactomeDirectory = "/home/bidaut/iti-main/interactomes/all";
```

This specifies the expression data directory for the training and testing for the cross-validation. Therefore, all data with the exception of Desmedt's dataset will be used.

```
my $transcriptomeDirectoryTrain = "/home/bidaut/iti-main/breast-cancer-1007/Desmedt-less/";
my $transcriptomeDirectoryTest = "/home/Bidaut/iti-main/breast-cancer-1007/Desmedt-less/";
```

These arguments are the cross-validation data path, they are left to their default.

```
my $conditionDirectoryTrain = "$cwd/conditions-Desmedt-less-ER-pos/run-%03d-train/";
my $conditionDirectoryTest = "$cwd/conditions-Desmedt-less-ER-pos/run-%03d-test/";
```

Uncorrected Proof

Author's Proof

This specifies the Resourcerer annotation directory:

```
my $resourcererDirectory = "/home/Bidaut/iti-main/resourcerer/";
```

This specifies the gene_info file path.

```
my $geneInfoFile = "/home/bidaut/iti-main/annotations/ncbi/Homo_sapiens.gene_info.gz";
```

Once properly edited, generateAllArgs.pl must be executed.

```
$ ./generateAllArgs.pl
```

Extra archive files can finally be removed:

```
$ rm -rf *tar.gz
```

3 Methods

3.1 Clinical Annotation Formatting and Filtering

In order to perform training on a homogeneous set of patients, all patient profiles must be carefully chosen on the basis of their clinical information.

All clinical information was downloaded alongside the expression data, either from Gene Expression Omnibus (Affymetrix data) or from the author's Web site (van De Vijver's dataset). For proper parsing by the ITI annotation filtering script, it has to be reformatted as a tab-delimited file containing several columns. As an example, this is the phenotype ".pheno" file for the Desmedt dataset.

```
$ head -n 1 breast-cancer-1007/Desmedt/Desmedt.pheno
ID AGE ER NODE NODE_NUMBER TREATMENT TUMOR_SIZE GRADE E.DFS T.DFS E.RFS T.RFS E.DMFS T.DMFS E.OS T.OS FLAG
```

Most authors are using different names for metastatic relapse, such as MR and DMFS, so there is a strong need for homogeneizing annotations. All correct files ready for ITI input were downloaded in Subheading 2.

In Breast Cancer two groups of patients have been defined that have deeply different molecular profiles (ER+ and ER–) and necessitate separate analysis.

The whole ITI pipeline will be applied separately on each group of patient, that is only ER+ or ER– patients from each dataset will be selected. Then, the patients having less than 5 years follow-up (60 months) are filtered out, as well as patients that underwent treatment. Finally, the separation criteria that is used for classification must be mentioned, which is the Distant Metastasis Free Survival (DMFS) criteria.

3.2 Patient Training Stratification

To properly train the system, we would like to perform a tenfold cross-validation training (Fig. 2). In practice, this is done by performing ten independent training on randomly selected subsets of patients. These subsets have to be carefully chosen in order to avoid biases, this process is called *stratification*. Again,

Uncorrected Proof

Author's Proof

this figure will be printed in b/w

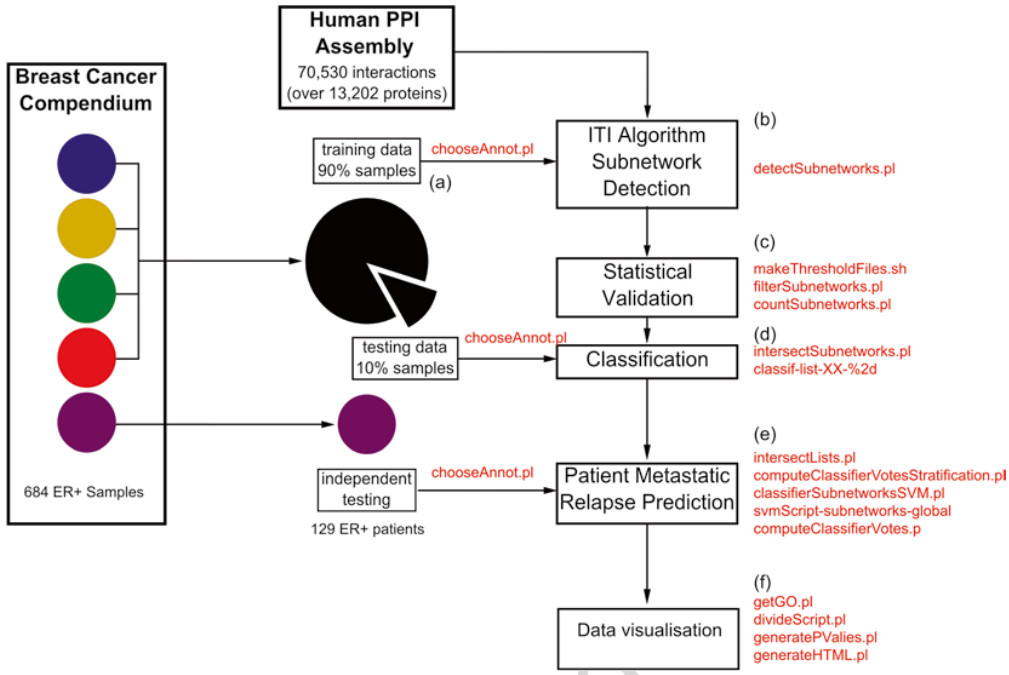


Fig. 2 ITI Framework. This figure represents the analysis structure presented in this chapter. The expression data is selected on its main subtype (ER+ patients are analyzed here) and pooled in ten training/testing sets to perform a cross-validated training (**a**). ITI recursive correlated subnetworks detection algorithm is then separately applied for each training set (**b**) with interactome as in input, and statistically validated (**c**). Classification is performed for each testing subnet (**d**). Optimally detected subnetworks are then intersected and tested on the 129 ER+ tumors profiles kept for independent testing (**e**). Finally, a set of web pages is constructed for visualization (**f**)

Uncorrected

the generateTranscriptomeConds.sh script performs stratification over the parameters specified earlier, that is ER±, keeps each dataset proportion and keeps each dataset proportioned for DMFS+ or – patients. If a dataset does not have enough patients to keep the DMFS± balance, patient samples are duplicated.

The tenfold cross-validation is also specified as an argument of the chooseAnnot.pl script.

Other data files containing "classification vectors", that is, vectors associating clinical conditions for each patient have to be generated for each stratification level. This is also done by the generateTranscriptomeConds.sh script.

All these criteria of filtering and cross-validation are specified in the generateTranscriptomeConds script, as chooseAnnot.pl arguments.

```
$itiPath/pipeline/chooseAnnot.pl -path $itiPath/
pipeline -d $transcriptomedataPath/Desmedt-less -e
"DMFS" -l 10 -s ER=1 -t 60 -s TREATMENT=None -o
conditions-Desmedt-less-ER-pos
```

Author's Proof

This returns a set of condition files that specifies which patient are used for the study (if patient is listed in the file) and their DMFS status.

3.3 ITI Algorithm

Each couple of training/testing set must then be analyzed for subnetworks. The detectsubnetworks script performs

- Interactome data parsing
- Transcriptome data parsing, including condition files
- Subnetwork detection
- Random subnetwork generation with three methods.
  - Random Interactome (shuffled interactome)
  - Random subnetwortk (subnetworks detected with a random decision over aggregation)
  - Shuffled clinical condition

Each step is detailed below:

3.3.1 Interactome Data Parsing

At this step each interactome in the interactome directory is parsed. A global interactome is created on the fly by unification of all interactions. Internally, a hash table for each protein is created, each value referencing an array containing a list of interactors to the protein. To homogenize the code, each protein is referenced by its corresponding gene NCBI accession number.

3.3.2 Transcriptome Data Parsing, Including Condition Files

Each expression dataset is parsed as follows. Expression files are parsed and "collapsed" using the corresponding platform file. The collapse procedure consists in switching the dataset from the probe universe to the gene universe by using the proper identifiers. When several probes are related to the same gene, the one with highest median signal is selected [13]. Once the condition file has been parsed, random conditions vectors are generated, and correlations expression-clinical condition are computed.

3.3.3 Subnetwork Detection

Subnetworks are characterized by several scores, each of them corresponding to a dataset. Each score is calculated by computing correlation of the average expression of all the genes belonging to the subnetwork with clinical condition. The score is weighted with the number of conditions in the given dataset according to Eq. 1.

$$S_{s,d} = \frac{\sqrt{n_d}}{\sqrt{\max n_d\,(DS)}} \left| corr\left(\frac{1}{n} \sum_{g \in s} e(g,d), cc(d) \right) \right| \qquad (1)$$

$S_{s,d}$ being the score of subnetwork S for dataset d. n_d is the number of conditions in the dataset, and max n_d (DS) is the maximum number of conditions in all the datasets. $e(g,d)$ is the vector

Uncorrected Proof

Author's Proof

of expression of gene g in dataset d, and cc(d) is a vector representing clinical condition for dataset d. The corr function represents the Pearson correlation.

After data parsing, subnetworks are then detected as follows. Each gene in the interactome is considered as a potential seed. Neighbors are aggregated if, after merging in the current subnetwork, their score is higher than threshold th in c datasets, and increases by a rated higher than r. In the current implementation, values for these parameters are, respectively, set to th = 0.3, $c = 2$, and $r = 0.03$. These thresholds are simply set for the initial detection and do not play any role in the statistical validation.

Subnetwork detection is started for each cross-validation layer by running the following scripts in the runs-Desmedt subdir:

```
$ cd runs-Desmedt
$./job-Desmedt-01
$ ./job-Desmedt-02
…
$ ./job-Desmedt-10
```

After running the script on the queue manager on the cluster, ten directories named "out-0.3-01-0.03-300-Desmedt-*" are created, corresponding to the different parameter values for th, c, r, and nRandom (nRandom is the number of random subnetworks generated on each cluster node). In this subdirectory, four directories were created, containing the detected subnetworks, and random subnetworks of category 1, 2, and 3 (*see* ref. 7, 8). Subnetworks are statistically validated in the following steps.

To test if all analysis went well, one must check if the jobDesmedt*.e*.* are empty:

```
$ cat jobDesmedt*.e*.*).
```

An example of detected subnetwork is given Fig. 3.

3.3.4 Filtering and Validating

After detecting subnetworks, p-values are computed for all subnetworks with the SCRIPTSLIST1-Desmedt-*.sh scripts. This script is based on Matlab and calculates p-values on the basis of a null distribution.

This is done by running all SCRIPTLIST1 scripts, as follows:

```
$ ./ SCRIPTSLIST1-Desmedt-01.sh
….
$ ./SCRIPTSLIST1-Desmedt-10.sh
```

p-values are generated for the three null-distributions defined previously and histograms are produced over p-values intervals.

To test the number of subnetwork that fall within all intervals, we run

```
$ ./generateScript2-Desmedt-01.sh
```

For the present study, the following table is obtained, with each sub-table correspond to a specific null distribution type:

Uncorrected Proof

Author's Proof

this figure will be printed in b/w

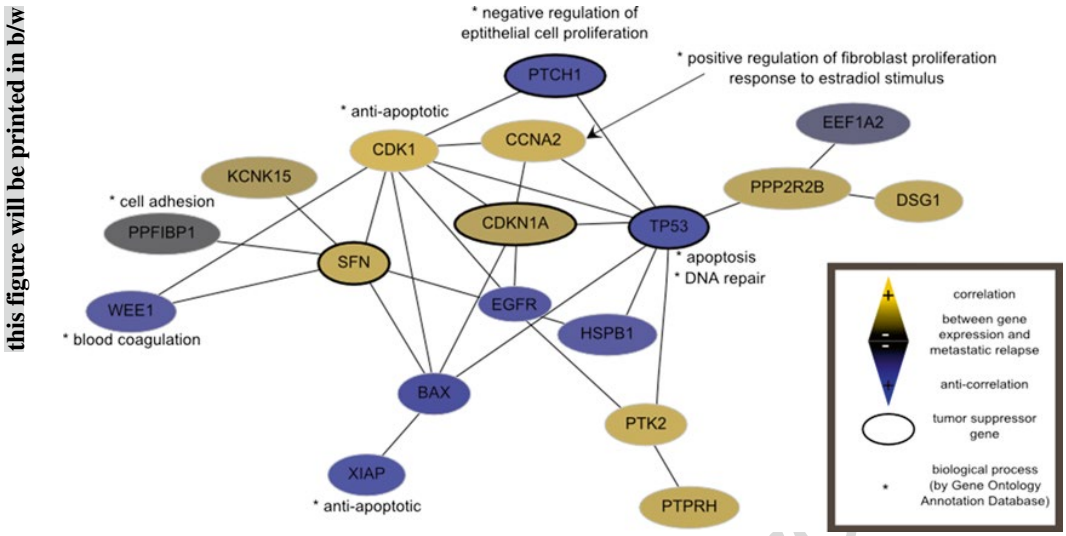


Fig. 3 Example of subnetworks detected with ITI analysis (adapted from [7]). Nodes and edges correspond, respectively, to genes coding proteins and protein interactions. Annotated genes are implicated in cancer progression and metastatic relapse. In its current state, only edges used for the detection are represented by ITI, but all interactions were represented in this figure. The pertinent disease marker is not the individual genes but rather the complete subnetwork expression

ThresholdFile\ThresholdNumber	1	2	3	4	5
subnets-kept_th1-pval-1-e-03	1029	592	187	32	3
subnets-kept_th1-pval-5-e-04	1006	553	126	19	2
subnets-kept_th1-pval-1-e-04	759	248	26	1	0
subnets-kept_th1-pval-5-e-05	490	91	5	0	0
ThresholdFile\ThresholdNumber	1	2	3	4	5
subnets-kept_th2-pval-1-e-03	1038	672	309	69	7
subnets-kept_th2-pval-5-e-04	1024	651	258	50	3
subnets-kept_th2-pval-1-e-04	873	446	110	15	0
subnets-kept_th2-pval-5-e-05	727	278	46	2	0
ThresholdFile\ThresholdNumber	1	2	3	4	5
subnets-kept_th3-pval-1-e-03	100	0	0	0	0
subnets-kept_th3-pval-5-e-04	77	0	0	0	0
subnets-kept_th3-pval-1-e-04	49	0	0	0	0

We then use this table to select a reasonably high number of subnetworks while keeping significantly low p-values and the highest possible number of datasets (parameter c) for each null distribution type. On our configuration, we obtain 248 subnetworks for a p-value of 1×10^{-4} on two datasets for random distribution 1 and 446 subnetworks for a p-value of 1×10^{-4} on two datasets for random distribution 2, while the third distribution was not used further (too much stringency).

Once the appropriate p-value is chosen, it can be applied to all cross-validation runs by editing the files generateScript2-

Uncorrected Proof

Author's Proof

Desmedt-*.sh and specifying the list of subnetworks to use by adding the options

```
--listToIntersect /home/bidaut/iti-main/runs-Desmedt/out-0.3-01-0.03-300-Desmedt-01/lists/subnets-kept_th1-pval-1-e-04-02.txt /home/bidaut/iti-main/runs-Desmedt/out-0.3-01-0.03-300-Desmedt-01/lists/subnets-kept_th2-pval-1-e-04-02.txt.
```

If this option is left unspecified, the system will prompt the *p*-value choice for each cross-validation run. Alternatively, these options can be specified in generateAllArgsScript2List.pl. Using this filter, ten subnetwork lists are generated separately for each cross-validation level. Next, these lists need to be combined by overlap of genes among subnetworks. For the purpose of this study, subnetworks overlapping by more than 50 % are combined as follows. Two subnetworks A and B are considered overlapping if more than 50 % of genes in subnetwork A are included in B and reciprocally. Overlapping subnetworks are grouped and the subnetwork with the highest score is kept as a representative for each group. Others are deleted.

This last step is performed in the main classification script (next section).

3.3.5 Classification

Obtained subnetworks are then used as a genomic signature to predict metastasis in a set of patients. To properly test the generalization capability of the subnetworks and derived signature and optimize the signature size (number of subnetworks to be used), we tested it as a classifier on a dataset not used during training (ER+ patients from Desmedt's dataset).

Subnetworks are intrinsically used as markers. First, the expression is each subnetwork is calculated by averaging the expression of each gene member (similar to the method used during training). Then, these are used for training a Support Vector Machine (based on LibSVM), that will separate good and poor prognosis patients on the basis of subnetwork expression and generate an SVM model. Several SVM models are generated successively with an increasing number of subnetworks, and the subnetwork set maximizing accuracy on the training set is kept. This is done by using the script *independentclassification.sh*, which performs all classification steps: Before running it, all paths must be properly set by editing the script, as follows.

```
$itiPath=/home/bidaut/iti-main/iti-2.0
$outPath=/home/bidaut/iti-main/runs-Desmedt
$transcriptomeDirectoryTrain=/home/bidaut/iti-main/breast-cancer-1007/Desmedt-less/
$transcriptomeDirectoryTest=/home/bidaut/iti-main/breast-cancer-1007/Desmedt/
$conditionDirectoryTrain=/home/bidaut/iti-main/conditions-Desmedt-less-ER-pos
```

Uncorrected Proof

Author's Proof


```
   $conditionDirectoryTest=/home/bidaut/iti-main/conditions-Desmedt/
   $resourcererDirectory=/home/bidaut/iti-main/resourcerer/
   $geneInfoFile=/home/bidaut/iti-main/annotations-ncbi/Homo_sapiens.gene_info.gz
```

Results are given in the file result-SVM-subnetwork.txt.

```
$ cat result-SVM-subnetwork.txt
NB    TN    FP    TP    FN    ACC                  SV        SP    FPR
001   94     0     0    35    0.728682170542636    0         1      0
002   94     0     0    35    0.728682170542636    0         1      0
003   84    10    10    25    0.728682170542636    0.285714285714286
0.893617021276596    0.106382978723404
004   83    11    10    25    0.720930232558139    0.285714285714286
0.882978723404255    0.117021276595745
005   85     9    13    22     0.75968992248062    0.371428571428571
0.904255319148936    0.0957446808510638
006   86     8     9    26    0.736434108527132    0.257142857142857
0.914893617021277    0.0851063829787234
007   89     5     7    28    0.744186046511628    0.2
0.946808510638298    0.0531914893617021
008   88     6     7    28    0.736434108527132    0.2
0.936170212765957    0.0638297872340425
009   87     7     8    27    0.736434108527132    0.228571428571429
0.925531914893617    0.074468085106383
```

The final accuracy on independent testing is obtained for six subnetworks, which gave an accuracy of 73.6 % (Column labeled ACC). The file also details false positive (FP), True Negative (TN), True positive (TP), False negative (FN), Sensitivity (SV), Specificity (SP), and False-Positive Rate (FPR).

3.4 Functional Explorations of Subnetworks with ITIDB

To explore subnetworks found, we can generate a set of web pages on the fly using the formatHTMLSubnet.sh script. Within the script, the following variables must be properly set:

```
top=6
inputDir=/home/bidaut/iti-main/runs-Desmedt/
suffix=Desmedt
```

This script will generate a set of web pages in the "out-0.3-01-0.03-300-Desmedt" directory that allows for functional exploration of scripts.

The result from the present study is available at http://bioinformatique.marseille.inserm.fr/iti-runs/supervised-5-datasets/iti-html-Desmedt-ER-pos/index.html. In Fig. 4 is represented the ITIDB interface for functional exploration of obtained subnetworks. Among the possibilities of ITIDB, one can mention the ability of visualizing subnetworks, rank subnetworks according to their significance to the biological question studied, and analyze individual genes present in subnetworks using classic bioinformatics tools.

Uncorrected Proof

Author's Proof

this figure will be printed in b/w

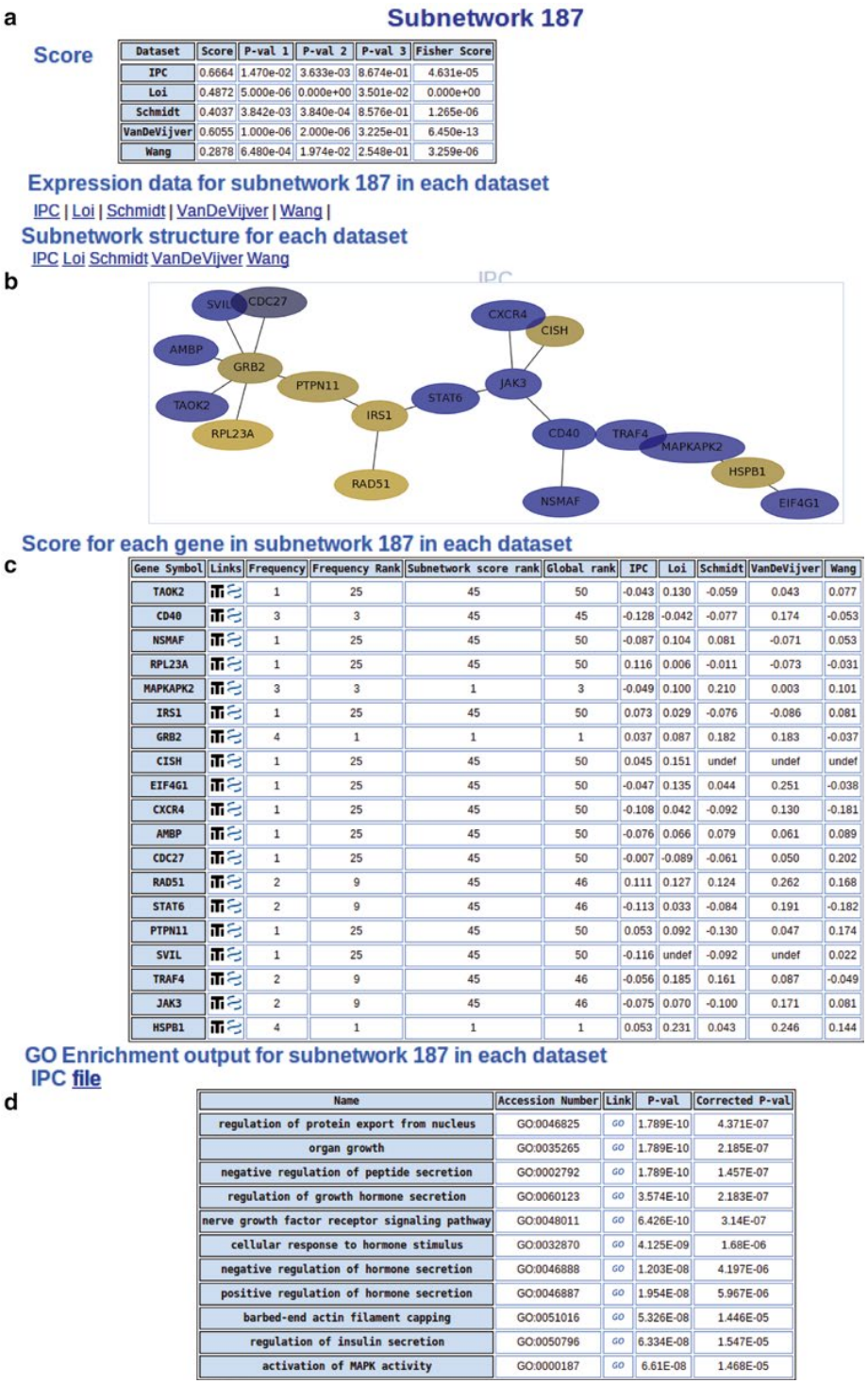


Fig. 4 ITIDB interface. This figure represents different views of the ITIDB interface. The main elements for a specific subnetwork (here subnetwork with accession number 187) are represented here. In (**a**), the subnetwork score and *p*-values for each dataset. In (**b**), the subnetwork topology with nodes (genes) and edges (PPI interaction). In (**c**), the details for each genes present in the subnetwork, including links to NCBI EntrezGene database and specific correlation values for each dataset. Finally, the table (**d**) represents subnetwork-specific GO enrichment computed with the ErmineJ program

Author's Proof

3.5 Conclusion

We presented all the practical steps for running the Interactome–Transcriptome pipeline, version 2.0. Most steps have been scripted, with the exception of a remaining manual threshold choice for the *p*-values. This pipeline is one of the only of its kind to be freely accessible under an open-source license (CeCILL).

Future developments include building an instance with a web interface, improving the information content on ITIDB, and include other data types. We will also propose other interactome maps readily formatted for ITI. Also, we are planning the direct inclusion of Gene Expression Omnibus data by simply mentioning GEO accession numbers during pipeline initialization.

References

1. Aranda B, Achuthan P, Alam-Faruque Y, Armean I, Bridge A, Derow C, Feuermann M, Ghanbarian AT, Kerrien S, Khadake J et al (2010) The IntAct molecular interaction database in 2010. Nucleic Acids Res 38:D525–D531
2. Bertucci F, Finetti P, Cervera N, Birnbaum D (2008) Prognostic classification of breast cancer and gene expression profiling. Med Sci (Paris) 24:599–606
3. Ceol A, Chatr Aryamontri A, Licata L, Peluso D, Briganti L, Perfetto L, Castagnoli L, Cesareni G (2010) MINT, the molecular interaction database: 2009 update. Nucleic Acids Res 38:D532–D539
4. Chuang H-Y, Lee E, Liu Y-T, Lee D, Ideker T (2007) Network-based classification of breast cancer metastasis. Mol Syst Biol 3:140
5. Desmedt C, Haibe-Kains B, Wirapati P, Buyse M, Larsimont D, Bontempi G, Delorenzi M, Piccart M, Sotiriou C (2008) Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. Clin Cancer Res 14:5158–5165
6. Dobbin KK, Zhao Y, Simon RM (2008) How large a training set is needed to develop a classifier for microarray data? Clin Cancer Res 14:108–114
7. Garcia M, Millat-Carus R, Bertucci F, Finetti P, Birnbaum D, Bidaut G (2012) Interactome-transcriptome integration for predicting distant metastasis in breast cancer. Bioinformatics 28:672–678
8. Garcia M, Stahl O, Finetti P, Birnbaum D, Bertucci F, Bidaut G (2011) Linking interactome to disease: a network-based analysis of metastatic relapse in breast cancer. In: Handbook of research on computational and systems biology: interdisciplinary applications, pp 406–427
9. Gillis J, Mistry M, Pavlidis P (2010) Gene function analysis in complex data sets using ErmineJ. Nat Protoc 5:1148–1159
10. Keshava Prasad TS, Goel R, Kandasamy K, Keerthikumar S, Kumar S, Mathivanan S, Telikicherla D, Raju R, Shafreen B, Venugopal A et al (2009) Human protein reference database—2009 update. Nucleic Acids Res 37: D767–D772
11. Michiels S, Koscielny S, Hill C (2005) Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet 365:488–492
12. Ramani AK, Bunescu RC, Mooney RJ, Marcotte EM (2005) Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. Genome Biol 6:R40
13. Reyal F, Stransky N, Bernard-Pierrot I, Vincent-Salomon A, de Rycke Y, Elvin P, Cassidy A, Graham A, Spraggon C, Désille Y et al (2005) Visualizing chromosomes as transcriptome correlation maps: evidence of chromosomal domains containing co-expressed genes—a study of 130 invasive ductal breast carcinomas. Cancer Res 65: 1376–1383
14. Sabatier R, Finetti P, Cervera N, Lambaudie E, Esterni B, Mamessier E, Tallet A, Chabannon C, Extra J-M, Jacquemier J et al (2011) A gene expression signature identifies two prognostic subgroups of basal breast cancer. Breast Cancer Res Treat 126:407–420
15. Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D (2004) The database of interacting proteins: 2004 update. Nucleic Acids Res 32:D449–D451
16. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AAM, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ et al (2002) A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 347:1999–2009

[AU1]

Uncorrected Proof

C.4 Chapitre *CNV-Interactome-Transcriptome Integration*

Dans ce chapitre intitulé "*CNV-Interactome-Transcriptome Integration to detect driver genes in cancerology*"[(15)], nous explorons l'intégration supplémentaires des CNV.

Documentation

http://sourceforge.net/p/iti/wiki/Home/

Site web compagnon

http://iti.sourceforge.net

12 CNV-Interactome-Transcriptome Integration to Detect Driver Genes in Cancerology

MAXIME GARCIA*, RAPHAËLE MILLAT-CARUS*,
FRANÇOIS BERTUCCI, PASCAL FINETTI, ARNAUD GUILLE,
JOSÉ ADÉLAÏDE, ISMAHANE BEKHOUCHE, RENAUD SABATIER,
MAX CHAFFANET, DANIEL BIRNBAUM AND GHISLAIN BIDAUT

*These authors contributed equally to this work.

12.1 ABSTRACT

The development of high-throughput gene-expression profiling technologies allows the definition of genomic signatures that help predict clinical condition or cancer patient outcome. However, such signatures show dependency on the training set and thus, suffer from lack of generalization and instability. This is the consequence of the microarray data topology and the fact that cancer is provoked by a small number of *drivers* genes provoking changes to *passengers* genes. Driver genes are the genomics elements whose deregulation are provoking the disease. Passenger genes have their expression affected because of misregulations and expression changes on the drivers but these changes have no impact on the disease. Separating drivers and passengers is of primary importance for the understanding of the disease and deciphering of molecular subtypes that exists for most cancers. Detecting these genes are a difficult process since cancer tumors are highly heterogeneous. In this chapter, we describe an interactome-based approach, Copy-Number-Variation-Interactome-Transcriptome Integration (CNV-ITI) that is used to detect driver genes that are specific to molecular subtypes in Breast Cancer (BC) by superimposition of a large scale protein-protein interactions (PPI) dataset (human interactome) over several gene expression datasets and array Comparative Genomic Hybridization (aCGH) datasets. The algorithm extracts interactome regions, so-called subnetworks, that allow for predicting relapse-free survival in cancer and detection of driver genes. As an illustrative example, we specifically applied it to Basal and Luminal A BC subtypes. Two other methods of CGH-Gene Expression profiles integration for detecting driver

genes in cancerology are described and compared to CNV-ITI.

12.2 INTRODUCTION

12.2.1 BREAST CANCER MOLECULAR SUBTYPES

Breast cancer (BC) is an heterogeneous disease. This explains why standard treatment does not work equally on all patients. This heterogeneity is difficult to decipher with the histoclinical criteria [2] that are used to predict prognosis. Consequently, many patient undergo overtreatment [3, 4]. The current challenge is to build better classifiers to i) separate breast cancer subtypes and ii) finely predict the prognosis associated with each subtype. The post-genomic era has seen the appearance of several tools that help gain a deeper knowledge of the molecular nature of the disease. Among these tools, microarray technology has contributed to understand the molecular biology of cancer at the mRNA level and at the DNA level (array-Comparative Genomic Hybridization - aCGH). These technologies allow for the discovery of markers that would refine disease classification. Using gene expression microarrays, four main BC Subtypes (luminal, basal, ERBB2-like and normal-like) were identified [5], which were confirmed in an independent study [6]) that refined this classification by splitting the luminal group in two subtypes (luminal A and luminal B) [7]. This classification has evolved with the discovery of other pertinent subgroups [8, 9], including Claudin-low subtype [10]. Correlation studies between clinical outcome and gene expression profiles were done, and specific prognosis were assigned to the identified subtypes [11, 12]. However, these correlations lack power due to a low number of studied samples. Nevertheless, the classification in five major molecular subtypes proved very robust and Hu *et al.* [13] validated a list of 306 genes that are now the reference for establishing these main subtypes.

Luminal A and basal subtypes are two of the major subtypes that are characterized by opposite features both clinically and at the genomic level [4]. Luminal A BCs are the most frequent (45% total occurrence). They are low-grade, differentiated tumors that express hormonal receptors and the *ESR1* and *GATA3* luminal differentiation genes. They are usually associated with a relatively favorable prognosis due to their response to hormonal therapy. The basal subtype represents 15% of all BCs. Basal BCs are high-grade, proliferative tumors that are hormone receptor negative and associated with poor prognosis [4]. While basal tumors are relatively chemosensitive, the effectiveness of chemotherapy remains limited.

12.2.2 INTEGRATION OF GENOMIC DATA FOR DISCOVERING DRIVER GENES IN LUMINAL A AND BASAL BREAST CANCER MOLECULAR SUBTYPES

A fundamental issue for systematic characterization of disease in general and BC in particular is the discovery of driver genes or markers. While it has been established that BC is characterized by five major subtypes (the four main subtypes and the subdivision of luminal subtype into A and B), genomic profiles show that tumors

are extremely diverse and most of them present unique characteristics. Therefore, it appears that most of the genetic lesions and gene misregulations that are found in tumors are not necessarily all at the origin of the disease at hand and that the challenge is to distinguish the driver (that provoke proliferation, treatment resistance, and metastasis at a primary level) from the passenger genes (genes whose changes are a by-product of drivers deregulations and that do not have a direct impact on the disease). These drivers have to be established for each subtype [14]. In addition, subtyping, although already quite extensive [9], may still need to be further defined.

Breast cancer arises as a result of expansion driven by cells that acquire immortality and a survival advantage through specific mutations and expression changes. The drivers are the genes that specifically provide for this selective advantage that enhance cancer hallmarks, including their involvement in pathways that favor proliferation and chemotherapy resistance. On the other hand, passengers, while also deregulated, are neutral to this selection process and are only involved in pathways of secondary importance in regards to cancer [15]. The main challenge is therefore to find a metric to separate drivers from passengers on the likelihood that they are hence driving the cancer.

The development of high resolution array CGH technology allowed the identification of genomic alterations on the whole genome [16]. Copy Number Aberrations (CNA) were associated with molecular subtypes and clinical outcome in BC [17]. However, the abundance and heterogeneity of genomic regions with significant CNA make the search for viable biological markers or therapeutics targets difficult [14]. To detect recurrent markers among tumors, most approaches are based on the frequency of alterations. If an alteration occurs more often within a gene in a set of tumors, it is likely that the gene represents a cancer key factor. For example, the Genomic Identification of Significant Targets in Cancer (GISTIC) algorithm identifies genomic regions that are aberrant more often than would be expected by chance. GISTIC was applied to several types of cancer [18, 19].

However, the use of CNA alone is not sufficient for detecting drivers. Amplified or deleted regions detected by array-CGH are usually large and cover multiple genes. Many of them are just passengers, but are indistinguishable from driver genes. Also, these approaches cannot determine the physiological or functional importance of the detected regions. These limitations highlight the need for more advanced, integrated approaches that take into account multiple types of biological information to identify drivers. In this regard, gene expression can give fundamental insight. However, the integration methodology is not obvious and must be carefully defined.

12.2.3 SCIENTIFIC GOAL

The basic postulate is that gene expression is fundamentally correlated with driver gene mutations at one time or the other of tumor history. Together, expression and gene mutation form a *genomic footprint* [14] that we wish to understand and establish for all BC subtypes. The goal is dual: (i) understand the disease biology and identify drug targets (druggable footprint), and (ii) predict patient outcome to adapt treatment to the disease and to the patient (actionable footprint). Several approaches have

been developed to superimpose genomic information and gene expression deregulation to help detect driver genes. Among these, Akavia *et al.* [14] developed an algorithm (CONEXIC) to identify driver genes located in regions with recurring genomic changes. In their approach, each driver is associated with a gene module that is deregulated by the driver. This method was validated on a melanoma dataset of 62 tumors with paired measurement of gene expression and CNA, initially published by Lin *et al.* [20]. It confirmed many previously known gene drivers of melanoma and connected them with many of their targets, to identify their biological functions. In addition, new targets were predicted and experimentally confirmed. This integrative method is further detailed in the chapter. Another integrative approach was proposed by Beroukhim *et al.* [19]. This method, based on the detection of driver genes by analysis of their gene expression and CNA profiles is also detailed.

However, even if the set of markers that can be detected is geared toward driver genes by an integrated analysis, microarray measurement noise and the tumoral heterogeneity are still major hurdles to obtain reliable markers. It has been established that DNA microarray signatures are inherently unstable in regard to their application to gene signature prediction [21]. For instance, two datasets of reference for breast cancer metastasis prediction, respectively studying 198 tumors [22] and 98 tumors [12], validated later on 298 tumors [23] produced signatures (76 genes for Wang, 70 genes for van de Vijver) that are different (only 3 genes were found to be common between the two signatures, and hence about 2% stability, as described by Chuang *et al.* [21]) and that do not classify independent data (i.e., predict patient outcome) reliably [24].

The reasons behind instability of expression-based signatures are two-fold. The first is purely mathematical and finds its origin in the topology of genome-based measurement, which are highly prone to the curse of dimensionality. The number of measured variables is vastly superior to the number of tumors, which prevent direct use of tools from classical statistics. Second, the variability is due to the very nature of the biological information to be measured. Cancer is provoked by driver genes that are subtly deregulated or mutated and these prompt secondary changes in a vast array of genes. Since all genes are placed at the same level by microarray analysis, causality information is removed. In fact, a microarray-based analysis detects genes with the most favorable statistics, i.e. the ones that are the most differentially expressed. These are not necessarily the ones we wish to detect in priority but are rather a byproduct of driver deregulation.

To properly retrieve driver genes, it is first necessary to reduce the curse of dimensionality by increasing the number of studied samples. This could be done by filtering the genome-based information (CNAs) by the gene expression data. Second, one has to include biological and causal information to differentiate driver genes from their passengers. This could be done by including a large protein-protein interactions (PPI) map over gene expression data and, instead of measuring independently all genes by individual statistics, driver genes could be detected at a global interactome level by identification of interacting regions that are deregulated together for a particular BC clinical condition or subtype. Several methods have already been proposed for net-

work analysis of gene expression [25, 24]. However, the multi-level nature of biology was not taken into account in the proposed methods since no integration with CNA was done.

In the present chapter, we describe the application of a variant of the Interactome-Transcriptome Integration (ITI) method previously applied to metastasis prediction in BC [24, 25]. ITI showed significant improvement on previously published classification methods by detecting Estrogen Receptor (ER)-based signatures using an integrated analysis of five gene expression datasets covering more than 900 tumors and an independent validation dataset. ITI achieved greater metastasis prediction accuracy (74% on ER-positive tumors, 54% on ER-negative tumors) and a higher stability (11% on ER-negative tumors and 32% on ER-positive tumors, obtained by permutation between training and testing datasets) compared to previously published statistics. The basic principle of ITI is to superimpose the known human interactome and detect a list of candidate seeds based on differential expression analysis. The neighborhood of these seeds are recursively explored and interacting genes are aggregated if they are co-expressed with it. ITI yields interactome regions that are differentially expressed, called subnetworks. In a second pass, subnetworks are statistically validated by confrontation to null distributions of scores. Subnetworks are then interrogated for biological function by Gene Ontology (GO) term enrichment [26] using an hypergeometric test [27] and annotations from the National Center for Biotechnological Information (NCBI) Entrez Gene database [28]. GO Enrichment is defined as follows 12.1:

$$en(GO,S) = \frac{g_{GO,t} \in S/|S|}{g_{GO,t} \in Genome/|Genome|}, \tag{12.1}$$

with $en(GO,S)$ being the enrichment (or depletion) of Gene Ontology term GO,t in subnetwork S, $|S|$ being the number of genes in studied subnetwork, $g_{GO,t}$ being the number of genes annotated with the GO term GO,t, $Genome$ the set of genes in the studied organism complete genome and $|Genome|$ the number of genes in this set. See [29] for the statistical test definition associated with this metric.

The original method described in [24] was heavily modified to include CNA measurement to focus the analysis on detection of driver genes. This was done by constructing a pipeline established by chaining sequentially the algorithm previously described [1] to select a first list of candidate drivers and the ITI algorithm that was used to build functional modules around these and visualize results by creating a dedicated bioinformatics resource. All results from this analysis are available on the ITI Web site[1]. This pipeline is called Copy Number Variations-Interactome-Transcriptome Integration (CNV-ITI) in the following sections.

Here, we applied the CNV-ITI method for the search of specific biomarkers for the basal and luminal A BC subtypes. This search yielded 123 subnetworks that included known markers for the biology of these subtypes as well as the interaction with functional modules deregulated in luminal A and basal tumors. The superimposition

[1] `http://iti.sourceforge.net/citi`

of CNA information helped us to distinguish passengers of the potential drivers. To understand the functional implication of these modules, we confronted the list of driver genes to the known biology of cancer [4].

12.3 MATERIAL AND METHODS

12.3.1 GLOBAL CNV-ITI WORKFLOW

This section describes the global workflow used for this analysis and details the integrated datasets. This analysis includes data from genomic CNV profiles (DNA) and gene expression (mRNA) profiles, as well as PPI.

The CNV-ITI workflow integrates these data to detect driver genes behind a phenotype, namely understanding the genomics difference between basal and luminal A BC subtypes. First, the amplified genes detected by high resolution array CGH at a high significant copy number levels. Then, these genes are considered to be putative driver genes behind the luminal A/basal subtype differentiation, and are further analyzed in a network context with ITI [24, 25]. The network analysis performs a simultaneous analysis of gene expression and PPI maps to replace the candidate genes in a biological context and hence determine the link between amplification and gene expression. These steps are shown in Figure 12.1.

The following sections detail all datasets (PPI and microarray datasets for measurement of gene expression and genomics alteration) used in the analysis.

12.3.2 PROTEIN-PROTEIN INTERACTION DATABASES

Before using ITI, a reference set of PPI must be defined. This set of interaction data must have certain properties, including being at a scale compatible with the expression data at hand. If the interaction set is too restricted, the ITI algorithm will not be able to detect pertinent sets of modules. To build this set of interaction data, it is possible to use locally-developed technologies and establish a PPI map. Among these technologies, the 2-hybrid screening in yeast [30] allows the discovery of PPI at a scalability that is compatible with other types of genomic assays, such as gene expression microarrays. Its principle is based on the fact that transcription factor (TF) binding can work with the activation and binding domains only at proximity without direct binding. The two proteins we wish to interrogate (usually called prey and bait) are introduced in a specifically designed mutant yeast that lack the biosynthesis of certain nutrients and thus will not survive on a medium without these. Two types of plasmids are engineered to produce on one side a protein product (that is typically a known protein the investigator is using to identify new binding partners, referred to as bait) in which the binding domain (BD) has been fused in, and on the other side, a protein product (that is either a single known protein or a library of known or unknown proteins, referred to as a prey) in which the activation domain (AD) has been fused in. The two plasmids are then introduced in the yeast. If the two proteins interact, the transcription of a reporter gene may occur. If the two proteins do not interact, the reporter gene is not activated and the yeast fails to survive. Among

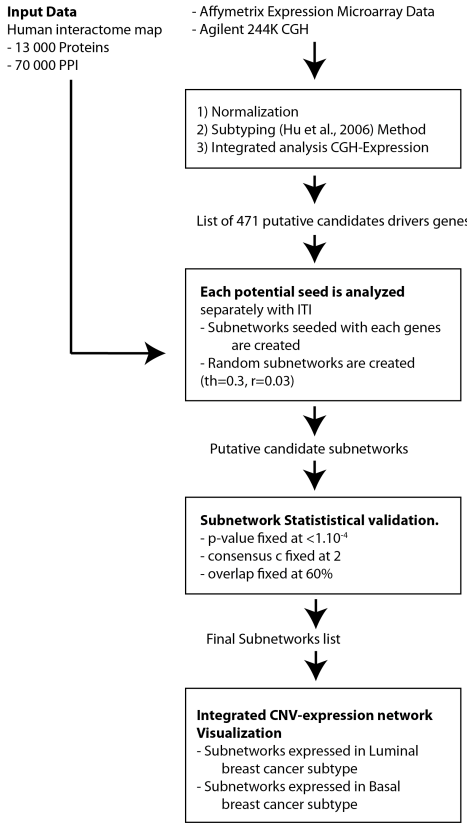


FIGURE 12.1 CNV-ITI general workflow. Gene expression data and CGH profiles help define a list of putative drivers. These are then superimposed to the human interactome for further analysis with ITI. Subnetworks regulated by the candidates are then detected and statistically validated. A visualization module allows the identification of subnetworks and their associated expression and CNA values.

other approaches, coaffinity purification followed by mass spectrometry [31], which consists in targeting a given protein with an antibody to pull simultaneously entire complexes out of solutions (technology referred to as “pull-down”). These technologies are now well-establish and can be carried out in a large number of laboratories at a reasonable cost. However, these still have a number of disadvantages. There are

large numbers of false positives and false negatives, due to the very nature of the screen. In the 2-hybrid screen, the error rate is as high as 70% [32]. Reasons for leading false-positive may include (among others) the fact that the 2-hybrid screen takes place in the nucleus and the measured proteins will not interact if they do not usually interact there, due to the lack of proper localization signals. Also, some proteins may interact if simultaneously expressed in the yeast although they will never be present in the cell simultaneously.

For this report, we did not generate the PPI data in-house, but integrated publicly available datasets in various databases using bioinformatics methods. Publicly available databases are largely built using the previously cited technologies (yeast 2-hybrid or other high-throughput technologies) and therefore carry the same type of limitations that were just described. Building a pipeline that is based on this type of assays implies to take specific precautions in handling the data. However, many databases are built on a strong core of interactions that have been validated *in vivo* and are considered as reliable. Additionally, these sets are completed by a large number of *in silico* predicted interactions. These predictions allow the extension of the interaction data to a larger genomic footprint. For deciphering specific luminal A and basal subnetworks, we used the interaction databases reported in [24]. Our final interaction set is built from five interaction databases, including the Human Protein Reference Database [33], INTAct [34], the Molecular Interactions Database [35], the Database of Interacting Proteins [36], and an *in-silico* interaction set generated by the Cocite algorithm [37]. These total 70,530 binary interactions among 13,202 proteins (see Table 12.1).

All data were downloaded as flat files from all databases respective Web sites, followed by similar normalization and transformation steps. These include the removal of unknown proteins (marked as such in the original file), the removal of self interactions and non-human proteins as well as the replacement of various identifiers by standard National Center for Biotechnology Information (NCBI) Entrez Gene[2] identifiers to allow the mapping between different PPI datasets on the one hand, and between interaction and expression data on the other hand.

12.3.3 MICROARRAY GENE EXPRESSION PROFILES

To understand the molecular differences at the gene expression level between luminal A and basal BC subtypes, microarray gene expression data were analyzed and superimposed to the interaction data. We made use of publicly available expression and genomic data generated from a large pool of tumors maintained at the Institute Paoli-Calmettes on pangenomic Affymetrix microarray HG-U133 Plus 2.0. The main advantage of using this technology is to obtain expression profiles on a very large gene set that covers most of the human genome in a single experiment. Expression data were normalized with the Robust Multi-array Average (RMA) standard method available in Bioconductor[3]. Data parsing and RMA normalization were done

[2] `http://www.ncbi.nlm.nih.gov/gene`

[3] `http://www.bioconductor.org`

TABLE 12.1
Summary of Protein-Protein Interactions Databases

Database	#Proteins	#Interactions	Technology
HPRD [33]	9,386	36,577	Y2H In vitro In vivo
Cocite [37]	6,349	15,705	In silico (Cocite algorithm)
DIP [36]	918	810	In vitro/manually curated
MINT [35]	5,559	12,143	Manually curated from literature
INTAct [34]	7,471	25,516	Large scale assays (Y2H, CoIP, pull-down)

TABLE 12.2
Summary of Microarray Datasets (Expression and CGH) used in the analysis

Dataset	GEO Accession	Platform Type	Basal/ luminal A
Sabatier [40]	GSE21653	Affymetrix HG-U133 Plus2.0	80/68
Bellkhouche [1]	GSE21653	Agilent Technologies Hu 244A	80/68

with the Bioconductor *Affy* package. The same tumors were used to generate CGH profiles as described in section 12.3.4. Correspondence tables for probes IDs-NCBI Entrez Gene[4] accession numbers were generated by specific Affymetrix annotations files available from Resourcerer [38] and probes were combined using the method described by Reyal *et al.* [39]. For each set of Affymetrix probes corresponding to the same gene, probes carrying the "`x_at`" extension were filtered out and probes having the highest median signal were retained. In case of having only "`x_at`" marked probes, we only applied the median-based rule. On the basis of the BC subtype classifier by Hu *et al.* [13], all tumors were attributed one of the five major subtypes commonly used. Among all tumors, 68 and 80 were labeled as basal and luminal A, respectively.

12.3.4 COMPREHENSIVE GENE HYBRIDIZATION PROFILES

aCGH and gene expression profiles were established for the same tumors. However, this is not a specific requirement of the presented approach, as data were separately processed and integrated under the form of statistics. After hybridization, scanning and data acquisition, standard bioinformatics analysis was applied, including initial filtering and LOESS normalization with the Feature Extraction package (Agilent Technologies, Santa Clara, CA, USA). Data were visualized and extracted under the

[4] `http://ncbi.nlm.nih.gov/gene`

form of log ratios using CGH Analytics 3.4 (Agilent Technology, Santa Clara, CA, USA). Then, copy number calculation was done after a Binary Circular Segmentation [41]. Then, the detection of significantly altered genes in basal and luminal A groups were done by GISTIC [18] (for specific implementation, see [1]). GISTIC takes into account both the amplitude and frequency of alteration in the tumor dataset to attribute a *p*-value to each gene.

12.3.5 INITIAL CANDIDATE GENE DETECTION BY PRIMARY INTEGRATION OF GENE GENOMIC ALTERATION AND EXPRESSION

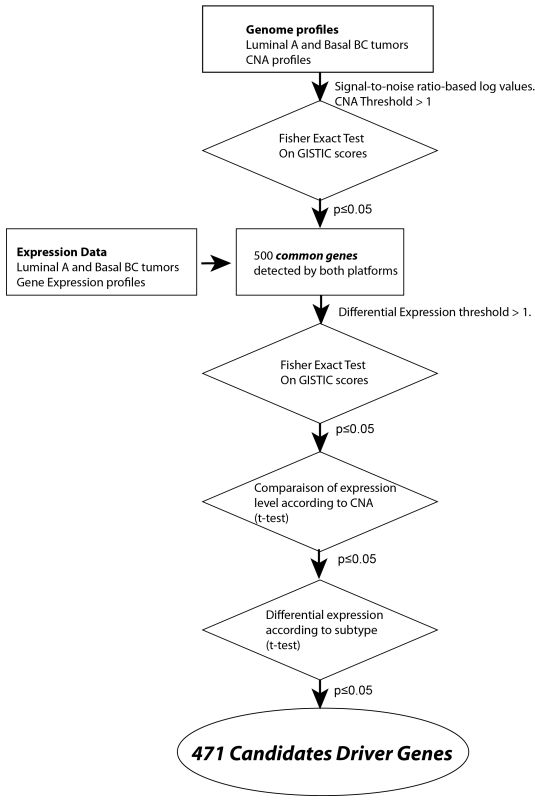


FIGURE 12.2 Initial candidate driver gene detection by Comparative Genomic Hybridization and Gene Expression analysis. Protocol adapted from Bekhouche *et al.* [1].

To understand the impact of driver genes on the deregulation of gene expression in basal and luminal A, the information associating gene expression deregulation and genomic alterations was integrated. ITI was then used to understand the impact of driver genes on the human interactome by detecting interactome regions that are deregulated by them. The list of candidate genes was first detected by adaptation of the approach described in [1]. First, we selected a subset of candidate genes with a significantly different GISTIC score among the two populations. Using the protocol described in Bekhouche *et al.* [1], we initially selected $n = 471$ genes significantly deregulated between luminal A and basal tumors with a False Discovery Rate $FDR = 1.10^{-3}$ (Benjamini-Hotchberg method). Other criteria have been added to select the initial gene candidates (Figure 12.2). (i) the respective frequency of genomic alteration between the two groups must be different (Fisher exact test with $p \leq 0.05$. (ii) Their expression and CNA must be correlated (student t-test with $p \leq 0.05$). (iii) These genes have to be differentially expressed between the basal and luminal A subtypes in addition to their genomic alteration differences (student t-test with $p \leq 0.05$). The final candidate gene list was submitted to ITI as initial seeds.

12.3.6 OTHER ANNOTATION DATABASES AND PIPELINE ELEMENTS

In addition to the NCBI Entrez gene and Resourcerer databases for identifier-probe conversion for microarray analysis, several other databases were used throughout the pipeline for subnetwork annotations. In particular, Gene Ontology information [26] was used to assign biological functions to subnetworks through enrichment measurement. This was done with the ErmineJ package [42]. ErmineJ specifically measures GO enrichment with a calculation of statistical significance with hypergeometric distribution and multiple testing correction. In addition, the list of human transcription factors were generated with the freely accessible for academic version of Transfac [43] (current version at the time of analysis: 7.0). To display generated subnetworks, the open-source GraphViz[5] network and graph visualization package (AT &T Research, Florham Park, NJ, USA) was used. Subnetwork statistical validation was implemented with the Matlab Statistical Toolbox (The Mathworks, Natick MA, USA).

12.3.7 INTERACTOME-TRANSCRIPTOME INTEGRATION ALGORITHM

The interactome-Transcriptome Integration (ITI) algorithm works by simultaneously examining interaction data and expression data to detect differentially expressed subnetworks, that is, interactome regions whose expression is globally differentially expressed among two experimental conditions. To accelerate detection and lower computing costs, the detection was parallelized by separating input datasets onto subsets on a Beowulf Cluster. ITI works in two main steps, an initial subnetwork detection and a statistical validation step of subnetworks. To detect differentially expressed subnetworks, correlation between clinical or phenotypic status of the two

[5] `http://ww.graphviz.org`

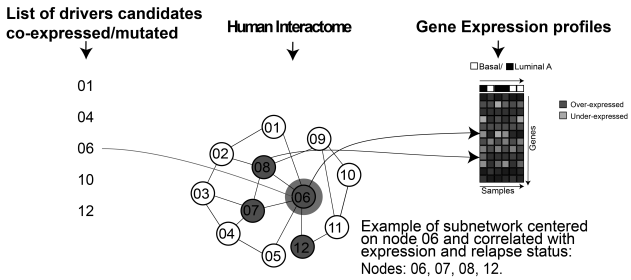


FIGURE 12.3 ITI Algorithm and Data organization. The 471 genes selected as seeds by the previous step are tested by ITI. Subnetworks are aggregated recursively around these seeds if their expression is correlated with the subtype.

conditions one wishes to analyze are computed. The set of interaction is then exhaustively searched for discriminative regions (Figure 12.3) by individually considering each node as potential seed and recursively aggregating neighbors if they increase the subnetwork score. The score is calculated by Pearson correlation between average gene expression of all the genes belonging to the subnetwork and a vector representing the phenotype (see Equation 12.2). In [24], this was done with patient Distant Metastasis Free Survival Status (DMFS) clinical condition in order to predict relapse in BC. Neighbors of nodes are examined recursively and merged to the current subnetwork if their expression allows to increase the score by a minimal rate. The current minimal threshold score for a subnetwork to be retained before statistical validation is $S = 0.3$ and the minimal rate is $r = 0.03$. Once the score cannot be improved, no more nodes are considered for the current subnetwork.

The subnetwork score S of a subnetwork s on a dataset d is calculated by the score $S_{s,d}$ defined in Equation 12.2. This score measures the correlation (in the present case, this is not the absolute correlation, contrary to what was done in [24]) between the subnetwork average gene expression and the BC molecular subtype. A normalization term is calculated by using the total number of conditions in the subnetwork n_d. This term is not useful when analyzing a single dataset but allows to scale appropriately scores when comparing multiple datasets. When analyzing multiple datasets simultaneously, the score S_s is computed by averaging individual scores over the datasets d (Equation 12.3) over the datasets list DS of size NS.

$$S_{s,d} = \frac{\sqrt{n_d}}{\sqrt{\max n_d(DS)}} \left| corr\left(\frac{1}{n}\sum_{g\in S} e(g,d), cc(d)\right)\right| \tag{12.2}$$

$$S_s = \frac{1}{NS}\sum_{d\in DS} S(s,d) \tag{12.3}$$

Subnetworks that are overlapping with subnetworks that have already been de-

tected are not retained. Overlapping between two subnetworks A and B is calculated by maximum inclusion of B in A and A in B. The inclusion score is calculated by counting common genes included in subnetwork A to B and dividing by the total number of genes contained in subnetwork A. In practice, subnetworks overlapping by more than 50% are removed.

Once subnetworks are detected, they have to be statistically validated. This is done by drawing two types of null distribution for score to asses significant thresholds. The first random distribution assesses the pertinence of ITI. It is obtained by randomly selecting subnetworks, i.e. by replacing the recursive aggregation method detailed above by a random aggregation around a seed. The second null distribution assesses whether the link between gene expression and PPI is biologically sound and valid. This second distribution was obtained by shuffling luminal A and basal molecular subtypes. To keep random subnetworks comparable (in terms of size) to previously detected subnetworks, their distribution of size (modeled as Gaussian) was forced to be comparable to the subnetworks detected in the above step. After drawing the two distributions, they were modeled by a Gaussian Mixture Distribution. We then used this model to determine an appropriate threshold score and filter out subnetworks with scores that were considered not significant. Overlapping subnetworks are then clustered according to the inclusion score previously defined if they overlap by more than a threshold O_s specified as a percentage. For each cluster, the subnetwork having the largest score is kept.

12.4 ANALYSIS

In this section, we detail the analysis steps as well as analysis of driver genes obtained with the CNV-ITI pipeline.

12.4.1 DRIVER GENE DETECTION WITH INTERACTOME-TRANSCRIPTOME INTEGRATION

To detect separately subnetworks expressed in the luminal A and basal BC subtypes, we performed two separate ITI runs on the specific tumor sets. These two runs took as input the list of 471 candidate genes that met the initial filters and select them as seed to generate candidate subnetworks. To generate subnetworks in the whole human interactome at hand, two types of data were used. First, the interaction database previously assembled (Section 12.3.2) was parsed and used as a basis for network exploration. The expression data from the 148 tumors were superimposed to the network on a gene basis. Then, correlation between the BC molecular subtype and gene expression is computed and subnetworks were detected by using ITI with p-values fixed at $Pval_1 = 1.10^{-3}$ and $Pval_2 = 1.10^{-3}$ for the two null distribution of scores, respectively. After removal of overlapping networks (the overlapping score was fixed at $O_s = 60\%$, 123 subnetworks were detected to be differentially expressed and further retained for analysis, totaling 541 genes. Among these, 62 subnetworks (279 genes) were detected to be expressed in basal subtype, and 61 subnetworks (262 genes) in luminal A subtype. A separate global analysis of gene expression and

CNA gave, respectively, 5,000 genes differentially expressed and 1,000 genes with distinct frequency of alteration among the two studied molecular subtypes [44].

12.4.2 SUBNETWORK VISUALIZATION

Data visualization is a quite complex step as it involves data integration among PPI; gene expression on a large number of tumors encompassing two different subtypes; CNA information measured on the same tumor set, and a large body of annotation data for microarray probes, transcription factors and genes (Symbols and NCBI Entrez Gene accession number). To accommodate all these data types, we heavily modified the visualization routines of the original ITI pipeline [24], especially to superimpose the GISTIC values.

Figure 12.4 illustrates an example of subnetworks obtained with ITI and an integrated visualization (these data are accessible from the ITI web site[6]). The subnetwork presented in Figure 12.4.A is expressed in luminal A subtype, while the one represented in Figure 12.4.B is expressed in basal subtype. For visualization, two additional figures of each subnetwork are generated for visualizing copy number variation (homozygous or heterozygous loss or gain). Sub-items 1, 2 and 3 of Figure 12.4 represent the ITI subnetwork score (blue=correlation with luminal A, red=correlation with basal subtype). GISTIC scores are directly represented on subnetworks with a red or green edge (red=gain, green=loss) for luminal A subtype (A.2 and B.2) and basal subtype (A.3 and B.3). The level of alteration or amplification is represented with a color code. The node is left white when the GISTIC score is not statistically significant. A screenshot of the complete CNV-ITI Web site is presented in Figure 12.5.

12.5 ANALYSIS AND RELEVANCE OF DETECTED SUBNETWORKS WITH RESPECT TO THE LITERATURE

The detected subnetworks are then analyzed for their biological relevance. Genes known to be specifically expressed in the two studied subtypes are found. In luminal A subtype, the gene that is the most frequently found in all subnetworks is without surprise ESR1 (found in 9 subnetworks). FOXA1 is also frequently found by ITI (3 subnetworks). ERG is found in a subnetwork. To the contrary, estrogen receptors or ERBB2 were not detected among basal tumors. Since these tumors are highly proliferative, genes related to cell cycle (cyclins in general, and CDKs) were found [4]. Also, among the 62 subnetworks expressed in basal, the Cyclin-dependent Kinase 6 was the most frequently found (in 8 subnetworks). CDK6 is known to regulate tumor suppressor RB1. Genes coding for cyclin E1 (CCNE1) and the Cyclin-dependent Kinase 2 (CDK2) were also found in 5 and 3 subnetworks, respectively.

Bertucci *et al.* [4] constructed a list of discriminant genes in luminal A or basal subtypes from literature. We crossed the list of genes documented by Bertucci *et al.*

[6] `http://iti.sourceforge.net/citi`

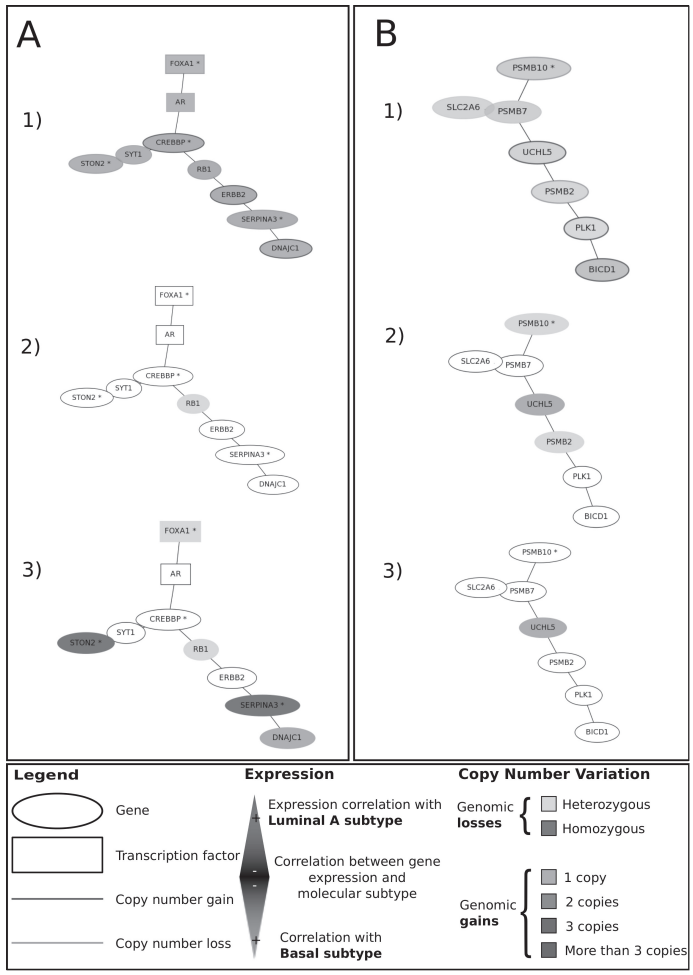


FIGURE 12.4 Examples of subnetworks detected with ITI, expressed in the luminal A (A) and basal (B) subtypes. Sub-items 1,2 and 3 represent respectively the following values: 1) the gene expression correlation score with the phenotype (Blue=correlated with luminal A subtype, Red=correlated with basal subtype, 2) and 3) items represent respectively CNA information with respct to luminal A (2) and basal (3) subtypes with a color code.

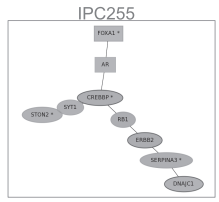
Subnetwork 12

Score

Dataset	Score	P-val 1	P-val 2	P-val 3
IPC-255	0.8978	1.603e-03	1.810e-04	8.053e-02

Subnetwork structure for each dataset

expression (LuminalA versus Basal)	IPC-255
aCGH-LuminalA	IPC-255
aCGH-Basal	IPC-255



Score for each gene in subnetwork 12 in each dataset

Gene Symbol	Links	Frequency	Frequency Rank	Subnetwork score rank	Global rank	IPC-255
CREBBP		3	5	1	3	0.375
RB1		2	20	1	4	0.341
DNAJC1		1	47	65	69	0.580
SERPINA3		1	47	65	69	0.380
AR		14	1	1	1	0.847
ERBB2		2	20	1	4	0.336
SYT1		2	20	65	64	0.499
STON2		2	20	65	64	0.551
FOXA1		3	5	63	58	0.871

GO Enrichment output for subnetwork 12 in each dataset

Name	Accession Number	Link	P-val	Corrected P-val
prostate gland development	GO:0030850	GO	8.269E-09	2.02E-05
gland development	GO:0048732	GO	3.429E-07	4.189E-04
enucleate erythrocyte differentiation	GO:0043353	GO	5.998E-06	4.885E-03
N-terminal protein amino acid acetylation	GO:0006474	GO	8.394E-06	5.127E-03
regulation of lipid metabolic process	GO:0019216	GO	1.054E-05	5.152E-03
regulation of lipid kinase activity	GO:0043550	GO	1.119E-05	4.555E-03
regulation of T cell differentiation in the thymus	GO:0033081	GO	1.119E-05	3.905E-03
phosphoinositide 3-kinase cascade	GO:0014065	GO	1.119E-05	3.417E-03
urogenital system development	GO:0001655	GO	1.206E-05	3.273E-03
positive regulation of myeloid leukocyte differentiation	GO:0002763	GO	1.438E-05	3.513E-03
regulation of macrophage differentiation	GO:0045649	GO	1.438E-05	3.193E-03

FIGURE 12.5 ITI Visualization and organization. This shows visualization elements associated with subnetworks; subnetwork score and *p*-values; subnetwork graph structure and interaction, superimposed with gene expression data and CNA measurements; Independent score measurement for each gene included in the subnetwork with links to annotation databases; GO enrichment information and *p*-values.

[4] and the genes detected by ITI to see if CNV-ITI was able to retrieve genes with a known biological link with the phenotype. Among these, cyclin D1, MYB, a transcription factor that may play a role in tumorigenesis, SMAD3, a major component of TGF signaling, were all present in luminal A. RUNX3 transcription factor was expressed in basal tumors. It is a tumor supressor gene that regulates carcinogenesis and that is silenced in breast cancer. The protein kinase LYN was also expressed in basal subtype. It encodes a tyrosine protein kinase that is known to be involved in basal signaling pathways [45]. CDK6 is known to be overexpressed in multiple cancer types [46]. As such, more than 5,000 genes were differentially expressed among the two subtypes [4], which makes the list of potential candidates for further analysis very large and potentially unpractical. This integrated analysis allows the reduction on a lower number of potential drivers ($n = 472$) and their interactors in the human interactome, which reduces the list of candidates, since only 114 genes were included in subnetworks. The complete gene list is presented in Figure 12.6.

12.5.1 IDENTIFICATION OF MUTATED GENES, TUMOR SUPPRESSORS AND ONCOGENES

Figure 12.6 shows genes detected by the whole pipeline, with a detailed list of initial candidate genes and the list of genes found overexpressed in basal and overexpressed in luminal A subtypes. The GISTIC scores revealed many significant genomic losses in the two subtypes, particularly with basal ($n = 131$) tumors but also with luminal A tumors ($n = 78$). Genomic amplifications were also significant ($n = 78$ and $n = 56$, for basal and luminal A tumors, respectively). Although the chromosomal aberrations found are specific to each subtype, it is also possible to differentiate the two categories of tumors with the number of alterations, which are vastly superior in the basal populations. When we focused only on subnetwork seeds, 60% showed a genomic loss in the basal populations while only 21% in luminal A. Gains were less frequent and were observed for *CPB1*, *IL20RA*, *TNIK*, *EPB41L2* and *MRC1* for basal tumors and for *CREBBP*, *TSC2*, *SPAG5* and *TNFSRF17* for luminal A.

These losses have a significant impact on expression of corresponding transcripts and with genes with which they interact; this is why they are part of detected subnetworks (defined as groups of genes with significant expression changes throughout the interactome). These genes could be considered as drivers for basal ($n = 75$ drivers) and luminal A ($n = 30$ drivers). Genes associated with gains could also be considered as potential drivers, even though with a lower impact, as they also greatly influence gene expression in their neighborhood. To understand the role of these genes, their annotations were manually examined. Several genes were detected by CNV-ITI while not been previously known to be involved in tumorigenesis. The carboxypeptidase B1 (*CPB1*), the *TRAF2* and *NCK* interacting kinase (*TNIK*), involved in *JNK* signaling, and *EPB41L2* have no known major role in carcinogenesis but were detected as differentially expressed.

A total of 28 transcription factors were detected by CNV-ITI, while 8 of them were used as subnetwork seeds. These are *C16orf80*, *LMO4*, *GTF2B* and *SOX10* (lost and under-expressed in luminal A tumors) and *FOXA1*, *EGR1*, *HIF1A*, *YBX1* and

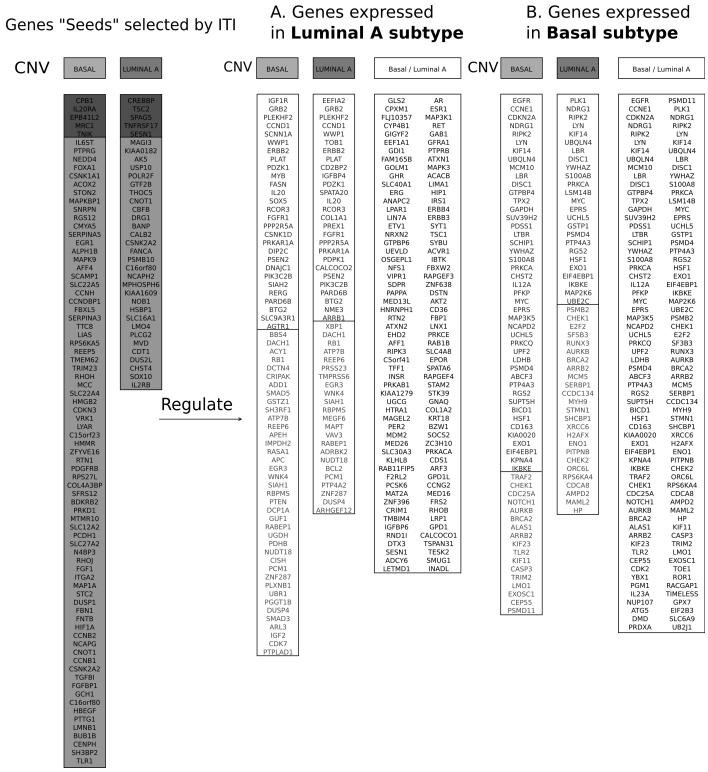


FIGURE 12.6 Detected driver genes with the CNV-ITI pipeline. On the left are represented the list of genes specifically expressed in luminal A and basal subtypes that have been retained as seeds by ITI (red=amplified, green=deleted). These are considered as putative driver genes, either tumor supressors or oncogenes. On the right, the list of genes regulated by these genes. In A, the list of genes found in subnetworks expressed in luminal A tumors. In B, the list of genes found in subnetworks expressed in basal tumors. For the A and B lists, we detailed the list of genes mutated in basal and luminal A tumors, or neither. In each of these sublists, genes marked as red are amplified and genes marked as green are deleted.

C16orf80 (lost and under-expressed in basal tumors). The Early Growth Response 1 gene (*EGR1*) is known as a tumor suppressor gene. *FOXA1* is known to promote tumor growth in several types of cancer [47] and it has been established that SNP sites associated with breast cancer are enriched for *FOXA1* biding sites [48].

12.6 OTHER CNA-GENE EXPRESSION INTEGRATION METHODS FOR DETECTING DRIVER GENES IN CANCER

To illustrate the variety of approaches for detecting driver in cancerology, we detail here two other methods, applied in the field of cancer for the detection of driver genes in melanoma (CONEXIC [14]), and in kidney carcinoma [19]. There were no methods previously developed for expression-CNV-PPI integration data, such as CNV-ITI, and hence these are considered to be the most relevant to our analysis.

12.6.1 CONEXIC METHOD BY AKAVIA ET AL.

The basic postulate by Akavia *et al.* [14] is that driving mutations are correlated with gene expression. A driver usually deregulates gene module expression that provokes tumorigenesis. A driver can also be deregulated without significant sequence alteration. Therefore, the method by Akavia *et al.* integrates gene expression and CNA to find signatures that are likely to contain drivers. The method is based on a previously published algorithm, Module Networks [49] that searches for regulated modules from gene expression data and precompiled lists of candidate control genes. COpy Number and Expression in Cancer (CONEXIC) extended Module Networks to make it suitable for driver detection in gene expression. CONEXIC uses a score-based search to detect gene modules that are differentially expressed and have the highest score within amplified or deleted regions. The output is a ranked list of modulators with high score that correlate with differentially expressed modules in gene expression and that are located in significantly altered regions of the genome. The modules themselves are modulated or deleted, which indicates that they are likely to control gene expression of the corresponding modules. Because these alterations are recurrently found in a significant number of tumors, it is very likely that these modulators are driving tumorigenesis. This method was applied to melanoma samples as follows. First, a list of candidate drivers was generated using CNA profiles available from 101 melanoma samples by the GISTIC method. Then, CONEXIC was applied to 62 paired samples to select driver modulators. A total of 64 modulators were selected by the algorithm that explain the behavior of 7,869 genes. The top 30 were considered as drivers. To annotate the drivers and the genes they are modulating, an automated procedure was developed (LitVan) that connects genes to the complete text of published articles in the NCBI Pubmed Database.

12.6.2 METHOD BY BEROUKHIM *ET AL.*

Beroukhim *et al.* [19] performed an integrated analysis (CN changes and expression) on sporadic clear cell kidney carcinomas (ccRCC) and von-Hippel Lindau (VHL) disease. A total of 90 tumors were analyzed and searched for significant copy number changes. Amplified regions were then searched for consistent expression in gene expression data, which led to the identification of several relevant genes, including CDK2NA, CDK2NB and MYC, among others. The analysis framework is articulated as follows. First, data were generated by extracting DNA and RNA from ccRCC tu-

mors and VHL tumors. Significant regions were determined by the GISTIC method. Then, DNA from VHL and ccRCC tumors were profiled on CGH arrays. RNA was hybridized on DNA microarrays. Using GISTIC, 7 amplified regions and 7 deleted regions were identified in ccRCC with a q-value < 0.25. To take into account the fact that some peak regions may have been displaced by passengers events, more robust regions (called wide peak regions) were identified with boundaries that resisted a leave-one-out analysis that is robust to the iterative removal of one tumor. The first assumption of the authors is that the oncogene targets driven by CNAs are activated by overexpression. The search was done to prioritize genes in the peak amplified regions under the form of an integrated analysis. Specifically, genes being significantly expressed in amplified tumors (p-values were drawn by tumor label permutation) were selected, using a signal-to-noise ratio. Among the genes in the peak regions for which probes exist on the expression arrays, 23 were significant. MYC was consistently overexpressed in tumors with 8q24 amplification. A similar approach to detect under-expressed genes in detected regions yielded several Tumor Suppressor Genes (TSG). This approach identified CDKN2A, a known TSG, among others.

12.7 DISCUSSION

BC is a vastly heterogeneous disease, with tumors characterized by genomic events that are both common and unique, and associated with very heterogeneous gene expression changes. This makes the detection of relevant markers and predictive genes very difficult. We are specifically interested in the study of two subtypes which present opposite features, luminal A (differentiated) and basal (proliferative) and that are extremely different at the genomic, transcriptomic and clinical levels.

To tackle the disease heterogeneity, integrative methods must be applied to take into account the information available at multiple biological levels and separate *driver* (genes that are the origin of the disease) from *passenger* (genes deregulated or altered as a collateral change to driver) genes. Methods integrating CNA information (measured by array-CGH) and gene expression were developed to various types of cancer. They were able to detect markers that are either expressed and amplified or under-expressed and deleted. In particular, the method from Beroukhim *et al.* [19] allowed the detection of new oncogenes in ccRCC tumors. The method by Akavia *et al.* [14], CONEXIC, made use of regulatory information to bring the detection at the level of modules, and separate driver from passengers events. Application to melanoma patients resulted in the identification of new candidate drivers as well as previously known oncogenes.

Applied to luminal A and basal BC subtypes, these methods still yield hundreds of potential candidates. Additional biological information must be taken into account by the detection algorithm to distinguish drivers and associated gene modules that are acting on the system, and this, separately for each subtype. We combined the integration method by Bekhouche *et al.* [1] and the ITI algorithm to build the CNV-ITI pipeline. A set of 471 candidate genes were first selected by a CNA-expression integration. These genes were submitted to ITI as subnetwork seeds to determine if

they drive modules of interacting genes which are differentially expressed. ITI gave a list of differentially expressed subnetworks after a complete interactome parsing and score-based statistical validation. Only 24% of the 471 initially identified genes were finally retained in the 123 validated subnetworks,. The method was able to focus the analysis on relevant sets of genes. A total of 61 subnetworks expressed in the luminal A subtype and 62 expressed in the basal subtype were identified. Known markers were detected (*ESR1* for luminal A, cyclins and kinases for basal) and new potential oncogenes and TSGs were identified. Finally, the number of candidate drivers was significantly reduced, which increased their statistical power as markers.

12.8 CONCLUSION

In this chapter, we describe the Comprehensive Genomic Hybridization-Interactome-Transcriptome Integration (CNV-ITI) pipeline for for the detection of drivers genes in cancerology. This pipeline works in two steps. First, it detects candidate markers by overlapping CNA profiles and gene expression profiles. Then, these candidates are submitted to the Interactome-Transcriptome Integration (ITI) pipeline for validation by searching for differentially expressed subnetworks in the human interactome. Retained drivers are the ones confirmed for being involved in differentially expressed subnetworks, since they are selected for modules that are driving the disease. All data is then stored in a bioinformatics resource for visualization and further analysis. We also improved previously published ITI visualization by superimposing deletion and amplification information in addition to the gene expression on subnetworks. As an illustrative example, we performed an analysis of specific subtypes in breast cancer with CNV-ITI. Two specific BC subtypes (luminal A (80 tumors) and basal (68 tumors)) were searched for driver and passenger genes. The CNV-ITI pipeline could potentially be applied to obtain an integrated view of the massive amount of data generated by international cancer consortia (The Cancer Genome Atlas, TCGA [50], the International Cancer Genome Consortium, ICGC, [51] or others). This would help obtain reliable markers not only significant, but also with the complete biological information and background under the form of gene interaction that would prove that these are hence drivers of the disease. These integrated analyses are a necessary step towards understanding the mechanisms that favor tumorigenesis in cancerology. It can be extended to integrate other levels of information such as point mutations and miRNA expression.

12.9 ACKNOWLEDGEMENTS

We would like to thank our funding sources. This research was funded by the Institut National du Cancer, the Ligue Nationale Contre le Cancer (label DB) and the Institut National de la Santé et de la Recherche Médicale. Support for the computational infrastructure was obtained from a Fondation pour la Recherche Médicale grant. Maxime Garcia is funded by the Institut National de la Santé et de la Recherche Médicale - Région Provence-Alpes Cote dAzur Fellowship. Support for Raphaële Millat-Carus was obtained from the Institut National du Cancer Grant. Thanks to

Sabrina Carpentier (Ipsogen, Marseille, France) for helpful discussions on the original ITI method, and Wahiba Gherraby for proofing the manuscript.

REFERENCES

1. I. Bekhouche, P. Finetti, J. Adelaide, *et al.*, "High-resolution comparative genomic hybridization of inflammatory breast cancer and identification of candidate genes.," *PLoS One*, vol. 6, no. 2, p. e16950, 2011.
2. P. J. van Diest, J. A. Belien, and J. P. Baak, "An expert system for histological typing and grading of invasive breast cancer. first set up.," *Pathol Res Pract*, vol. 188, pp. 405–409, Jun 1992.
3. F. Bertucci, P. Finetti, N. Cervera, D. Maraninchi, P. Viens, and D. Birnbaum, "Gene expression profiling and clinical outcome in breast cancer," *Omics: A Journal of Integrative Biology*, vol. 10, no. 4, pp. 429–443, 2006. PMID: 17233555.
4. F. Bertucci, P. Finetti, N. Cervera, *et al.*, "How different are luminal A and basal breast cancers?," *Int J Cancer*, vol. 124, pp. 1338–1348, Mar 2009.
5. C. M. Perou, T. Sorlie, M. B. Eisen, *et al.*, "Molecular portraits of human breast tumours.," *Nature*, vol. 406, pp. 747–752, Aug 2000.
6. T. Sorlie, C. M. Perou, R. Tibshirani, *et al.*, "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.," *Proc Natl Acad Sci U S A*, vol. 98, pp. 10869–10874, Sep 2001.
7. E. Charafe-Jauffret, C. Ginestier, F. Monville, *et al.*, "How to best classify breast cancer: conventional and novel classifications (review).," *Int J Oncol*, vol. 27, pp. 1307–1313, Nov 2005.
8. M. Guedj, L. Marisa, A. de Reynies, *et al.*, "A refined molecular taxonomy of breast cancer.," *Oncogene*, vol. 31, pp. 1196–1206, Mar 2012.
9. C. Curtis, S. P. Shah, S.-F. Chin, *et al.*, "The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups.," *Nature*, vol. 486, pp. 346–352, Jun 2012.
10. A. Prat, J. S. Parker, O. Karginova, *et al.*, "Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer.," *Breast Cancer Res*, vol. 12, no. 5, p. R68, 2010.
11. C. Sotiriou, S.-Y. Neo, L. M. McShane, *et al.*, "Breast cancer classification and prognosis based on gene expression profiles from a population-based study," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, pp. 10393–10398, Sept. 2003. PMID: 12917485.
12. L. J. van 't Veer, H. Dai, M. J. van de Vijver, *et al.*, "Expression profiling predicts outcome in breast cancer.," *Breast Cancer Res*, vol. 5, no. 1, pp. 57–58, 2003.
13. Z. Hu, C. Fan, D. S. Oh, *et al.*, "The molecular portraits of breast tumors are conserved across microarray platforms.," *BMC Genomics*, vol. 7, p. 96, 2006.
14. U. D. Akavia, O. Litvin, J. Kim, *et al.*, "An integrated approach to uncover drivers of cancer.," *Cell*, vol. 143, pp. 1005–1017, Dec 2010.
15. G. Curigliano, "New drugs for breast cancer subtypes: targeting driver pathways to overcome resistance.," *Cancer Treat Rev*, vol. 38, pp. 303–310, Jun 2012.
16. D. S. P. Tan and J. S. Reis-Filho, "Comparative genomic hybridisation arrays: high-throughput tools to determine targeted therapy in breast cancer.," *Pathobiology*, vol. 75, no. 2, pp. 63–74, 2008.

17. A. Bergamaschi, Y. H. Kim, P. Wang, *et al.*, "Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer.," *Genes Chromosomes Cancer*, vol. 45, pp. 1033–1040, Nov 2006.
18. R. Beroukhim, G. Getz, L. Nghiemphu, *et al.*, "Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, pp. 20007–20012, Dec. 2007. PMID: 18077431.
19. R. Beroukhim, J.-P. Brunet, A. Di Napoli, *et al.*, "Patterns of gene expression and copy-number alterations in von-hippel lindau disease-associated and sporadic clear cell carcinoma of the kidney.," *Cancer Res*, vol. 69, pp. 4674–4681, Jun 2009.
20. W. M. Lin, A. C. Baker, R. Beroukhim, *et al.*, "Modeling genomic diversity and tumor dependency in malignant melanoma.," *Cancer Res*, vol. 68, pp. 664–673, Feb 2008.
21. H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis.," *Mol Syst Biol*, vol. 3, p. 140, 2007.
22. Y. Wang, J. G. M. Klijn, Y. Zhang, *et al.*, "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," *Lancet*, vol. 365, pp. 671–679, Feb. 2005. PMID: 15721472.
23. M. J. van de Vijver, Y. D. He, L. J. van't Veer, *et al.*, "A gene-expression signature as a predictor of survival in breast cancer," *The New England Journal of Medicine*, vol. 347, pp. 1999–2009, Dec. 2002. PMID: 12490681.
24. M. Garcia, R. Millat-Carus, F. Bertucci, P. Finetti, D. Birnbaum, and G. Bidaut, "Interactome-transcriptome integration for predicting distant metastasis in breast cancer.," *Bioinformatics*, vol. 28, pp. 672–678, Mar 2012.
25. M. Garcia, O. Stahl, P. Finetti, D. Birnbaum, F. Bertucci, and G. Bidaut, "Linking interactome to disease : A network-based analysis of metastatic relapse in breast cancer," in *Handbook of Research on Computational and Systems Biology: Interdisciplinary Applications*, 2011.
26. M. Ashburner, C. A. Ball, J. A. Blake, *et al.*, "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.," *Nat Genet*, vol. 25, pp. 25–29, May 2000.
27. S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church, "Systematic determination of genetic network architecture.," *Nat Genet*, vol. 22, pp. 281–285, Jul 1999.
28. E. W. Sayers, T. Barrett, D. A. Benson, *et al.*, "Database resources of the National Center for Biotechnology Information.," *Nucleic Acids Res*, vol. 40, pp. D13–D25, Jan 2012.
29. I. Rivals, L. Personnaz, L. Taing, and M.-C. Potier, "Enrichment or depletion of a GO category within a class of genes: which test?," *Bioinformatics*, vol. 23, pp. 401–407, Feb 2007.
30. K. H. Young, "Yeast two-hybrid: so many interactions, (in) so little time...," *Biology of Reproduction*, vol. 58, pp. 302–311, Feb. 1998.
31. E. N. Brody, L. Gold, R. M. Lawn, J. J. Walker, and D. Zichi, "High-content affinity-based proteomics: unlocking protein biomarker discovery.," *Expert Rev Mol Diagn*, vol. 10, pp. 1013–1022, Nov 2010.
32. C. M. Deane, L. Salwinski, I. Xenarios, and D. Eisenberg, "Protein interactions: Two methods for assessment of the reliability of high throughput observations," *Molecular & Cellular Proteomics*, vol. 1, pp. 349–356, May 2002.
33. T. S. Keshava Prasad, R. Goel, K. Kandasamy, *et al.*, "Human Protein Reference Database-2009 update.," *Nucleic Acids Res*, vol. 37, pp. D767–D772, Jan 2009.

34. B. Aranda, P. Achuthan, Y. Alam-Faruque, *et al.*, "The IntAct molecular interaction database in 2010.," *Nucleic Acids Res*, vol. 38, pp. D525–D531, Jan 2010.
35. A. Ceol, A. Chatr Aryamontri, L. Licata, *et al.*, "MINT, the molecular interaction database: 2009 update.," *Nucleic Acids Res*, vol. 38, pp. D532–D539, Jan 2010.
36. L. Salwinski, C. S. Miller, A. J. Smith, F. K. Pettit, J. U. Bowie, and D. Eisenberg, "The Database of Interacting Proteins: 2004 update.," *Nucleic Acids Res*, vol. 32, pp. D449–D451, Jan 2004.
37. A. K. Ramani, R. C. Bunescu, R. J. Mooney, and E. M. Marcotte, "Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome," *Genome Biology*, vol. 6, no. 5, p. R40, 2005. PMID: 15892868.
38. J. Tsai, R. Sultana, Y. Lee, *et al.*, "RESOURCERER: a database for annotating and linking microarray resources within and across species.," *Genome Biol*, vol. 2, no. 11, p. SOFTWARE0002, 2001.
39. F. Reyal, N. Stransky, I. Bernard-Pierrot, *et al.*, "Visualizing chromosomes as transcriptome correlation maps: evidence of chromosomal domains containing co-expressed genes–a study of 130 invasive ductal breast carcinomas.," *Cancer Res*, vol. 65, pp. 1376–1383, Feb 2005.
40. R. Sabatier, P. Finetti, N. Cervera, *et al.*, "A gene expression signature identifies two prognostic subgroups of basal breast cancer," *Breast Cancer Research and Treatment*, vol. 126, pp. 407–420, Apr. 2011. PMID: 20490655.
41. A. B. Olshen, E. S. Venkatraman, R. Lucito, and M. Wigler, "Circular binary segmentation for the analysis of array-based DNA copy number data," *Biostatistics (Oxford, England)*, vol. 5, pp. 557–572, Oct. 2004. PMID: 15475419.
42. H. K. Lee, W. Braynen, K. Keshav, and P. Pavlidis, "ErmineJ: tool for functional analysis of gene expression data sets.," *BMC Bioinformatics*, vol. 6, p. 269, 2005.
43. V. Matys, O. V. Kel-Margoulis, E. Fricke, *et al.*, "TRANSFAC and its module TRANSCompel:transcriptional gene regulation in eukaryotes.," *Nucleic Acids Res*, vol. 34, pp. D108–D110, Jan 2006.
44. J. Adelaide, P. Finetti, I. Bekhouche, *et al.*, "Integrated profiling of basal and luminal breast cancers.," *Cancer Res*, vol. 67, pp. 11565–11575, Dec 2007.
45. D. R. Croucher, F. Hochgrafe, L. Zhang, *et al.*, "Involvement of Lyn and the atypical kinase SgK269/PEAK1 in a basal breast cancer signaling pathway.," *Cancer Res*, vol. 73, pp. 1969–1980, Mar 2013.
46. P. J. Roberts, J. E. Bisi, J. C. Strum, *et al.*, "Multiple roles of cyclin-dependent kinase 4/6 inhibitors in cancer therapy.," *J Natl Cancer Inst*, vol. 104, pp. 476–487, Mar 2012.
47. M. Katoh, M. Igarashi, H. Fukuda, H. Nakagama, and M. Katoh, "Cancer genetics and genomics of human FOX family genes.," *Cancer Lett*, vol. 328, pp. 198–206, Jan 2013.
48. M. R. Katika and A. Hurtado, "A functional link between FOXA1 and breast cancer SNPs.," *Breast Cancer Res*, vol. 15, p. 303, Feb 2013.
49. E. Segal, M. Shapira, A. Regev, *et al.*, "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data.," *Nat Genet*, vol. 34, pp. 166–176, Jun 2003.
50. Cancer Genome Atlas Research Network, "Comprehensive genomic characterization defines human glioblastoma genes and core pathways.," *Nature*, vol. 455, pp. 1061–1068, Oct 2008.
51. The International Cancer Genome Consortium, "International network of cancer genome projects.," *Nature*, vol. 464, pp. 993–998, Apr 2010.

Index

aCGH, 2
Affymetrix, 9
Agilent Technology, 10

Basal subtype, 2
Breast cancer, 2
Breast cancer subtype, 2

Comparative Genomic Hybridization, 1, 2
Copy Number Variation profile, 10
Copy Number Variation-Interactome-Transcriptome Integration, 5

Driver gene, 2

Gene expression profile, 9
Gene Ontology, 5
Genomic Identification of Significant Targets in Cancer, 3
GISTIC score, 14
GO Enrichment, 5
GraphViz package, 14

Human interactome, 1, 6
Hypergeometric distribution, 5

Interactome-Transcriptome Integration, 5
ITI Algorithm, 12

Luminal A subtype, 2

Passenger gene, 3
PPI database integration, 8
Protein-Protein Interaction, 6

Reference interactome, 6

Subnetwork, 5, 11
Subnetwork functional analysis, 17
Subnetwork score, 12, 13
Subnetwork seed, 11
Subnetwork visualization, 14

Transcription factor, 8

Yeast 2-Hybrid, 8

RÉFÉRENCES

[1] National Cancer Act of 1937, 1937.

[2] H J Bloom and W W Richardson. Histological grading and prognosis in breast cancer ; a study of 1409 cases of which 359 have been followed for 15 years. *British journal of cancer*, 11(3), 1957.

[3] C W Elston and I O Ellis. The value of histological grade in breast cancer : experience from a large study with long-term follow-up. *Histopathology*, 19(5), 1991.

[4] Fattaneh A Tavassoli and Peter Devilee. Tumours of the Breast and Female Genital Organs, 2003.

[5] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822), 2001.

[6] The ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), 2012.

[7] François Bertucci, Rémi Houlgatte, Samuel Granjeaud, Valéry Nasser, Béatrice Loriod, Emmanuel Beaudoing, Pascal Hingamp, Jocelyne Jacquemier, Patrice Viens, Daniel Birnbaum, and Catherine Nguyen. Prognosis of breast cancer and gene expression profiling using DNA arrays. *Annals of the New York Academy of Sciences*, 975, 2002.

[8] Laura J van't Veer, Hongyue Dai, Marc J van de Vijver, Yudong D He, Augustinus A M Hart, Mao Mao, Hans L Peterse, Karin van der Kooy, Matthew J Marton, Anke T Witteveen, George J Schreiber, René Bernards, and Stephen H Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(345), 2002.

[9] Yixin Wang, Jan G M Klijn, Yi Zhang, Anieta M Sieuwerts, Maxime P Look, Fei Yang, Dmitri Talantov, Mieke Timmermans, Marion E Meijer-van Gelder, Jack Yu, Tim Jatkoe,

Els MJJ J J Berns, David Atkins, and John A Foekens. Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, 365 (9460), 2005.

[10] Han-Yu Chuang, Eunjung Lee, Yu-Tsueng Liu, Doheon Lee, and Trey Ideker. Network-based classification of breast cancer metastasis. *Molecular Systems Biology*, 3(140), 2007.

[11] Liat Ein-Dor, Or Zuk, and Eytan Domany. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 103(15), 2006.

[12] Maxime U Garcia, Olivier Stahl, Pascal Finetti, Daniel Birnbaum, François Bertucci, and Ghislain Bidaut. Linking Interactome to Disease : A Network-Based Analysis of Metastatic Relapse in Breast Cancer. In *Handbook of Research on Computational and Systems Biology : Interdisciplinary Applications*. 2011.

[13] Maxime U Garcia, Raphaelle Raphaëlle Millat-Carus, François Bertucci, Pascal Finetti, Daniel Birnbaum, and Ghislain Bidaut. Interactome-Transcriptome integration for predicting distant metastasis in breast cancer. *Bioinformatics*, 28(5), 2012.

[14] Maxime U Garcia, Pascal Finetti, François Francois Bertucci, Daniel Birnbaum, and Ghislain Bidaut. Detection of driver protein complexes in breast cancer metastasis by large scale transcriptome-interactome integration. In *Methods in Molecular Biology - Gene Function Analysis*. 2013.

[15] Maxime Ulysse Garcia, Raphaëlle Millat-Carus, François Bertucci, Pascal Finetti, Arnaud Guille, José Adélaïde, Ismahane Bekhouche, Renaud Sabatier, Max Chaffanet, Daniel Birnbaum, and Ghislain Bidaut. CNV-Interactome-Transcriptome Integration to detect driver genes in cancerology. In *Microarray Image and Data Analysis : Theory and Practice*.

[16] Vincent T DeVita and Steven A Rosenberg. Two Hundred Years of Cancer Research. *New England Journal of Medicine*, 366(23), 2012.

[17] B Ren, F Robert, J J Wyrick, O Aparicio, E G Jennings, I Simon, J Zeitlinger, J Schreiber, N Hannett, E Kanin, T L Volkert, C J Wilson, S P Bell, and R A Young. Genome-wide location and function of DNA binding proteins. *Science*, 290(5500), 2000.

[18] Peter J Park. ChIP-seq : advantages and challenges of a maturing technology. *Nature Reviews Cancer*, 10(10), 2009.

[19] Trong Nguyen-Duc, Eveline Peeters, Serge Muyldermans, Daniel Charlier, and Gholamreza Hassanzadeh-Ghassabeh. Nanobody(R)-based chromatin immunoprecipitation/micro-array analysis for genome-wide identification of transcription factor DNA binding sites. *Nucleic acids research*, 41(5), 2013.

[20] Bing Ren, Hieu Cam, Yasuhiko Takahashi, Thomas Volkert, Jolyon Terragni, Richard A Young, and Brian David Dynlacht. E2F integrates cell cycle progression with DNA repair, replication, and G2/M checkpoints. *Genes & Development*, 16, 2002.

[21] Moshe Szyf. The implications of DNA methylation for toxicology : toward toxicomethylomics, the toxicology of DNA methylation. *Toxicological Sciences*, 120(2), 2011.

[22] Wolf Reik. Stability and flexibility of epigenetic gene regulation in mammalian development. *Nature*, 447(7143), 2007.

[23] Jeffrey a Rosenfeld, Zhibin Wang, Dustin E Schones, Keji Zhao, Rob DeSalle, and Michael Q Zhang. Determination of enriched histone modifications in non-genic portions of the human genome. *BMC Genomics*, 10, 2009.

[24] G Jia, Y Fu, X Zhao, Q Dai, and G Zheng. N6-Methyladenosine in nuclear RNA is a major substrate of the obesity-associated FTO. *Nature Chemical Biology*, 7(12), 2012.

[25] B Kusenda, M Mraz, J Mayer, and S Pospisilova. MicroRNA biogenesis, functionality and cancer relevance. *Biomedical papers of the Medical Faculty of the University Palacký, Olomouc, Czechoslovakia*, 150(2), 2009.

[26] David P Bartel. MicroRNAs : target recognition and regulatory functions. *Cell*, 136(2), 2009.

[27] Isaac Bentwich, Amir Avniel, Yael Karov, Ranit Aharonov, Shlomit Gilad, Omer Barad, Adi Barzilai, Paz Einat, Uri Einav, Eti Meiri, Eilon Sharon, Yael Spector, and Zvi Bentwich. Identification of hundreds of conserved and nonconserved human microRNAs. *Nature Genetics*, 37(7), 2005.

[28] Benjamin P Lewis, Christopher B Burge, and David P Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1), 2005.

[29] Robin C Friedman, Kyle Kai-How Farh, Christopher B Burge, and David P Bartel. Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, 19 (1), 2009.

[30] Harald H H Göring. Tissue specificity of genetic regulation of gene expression. *Nature Genetics*, 44(10), 2012.

[31] Henry Wirth, Markus Löffler, Martin von Bergen, and Hans Binder. Expression cartography of human tissues using self organizing maps. *BMC Bioinformatics*, 12(1), 2011.

[32] Jie Li, Xu Hua, Martin Haubrock, Jin Wang, and Edgar Wingender. The architecture of the gene regulatory networks of different tissues. *Bioinformatics*, 28(18), 2012.

[33] Teuvo Kohonen. Self-organized formation of topologically correct feature maps. *Biological cybernetics*, 69, 1982.

[34] Teuvo Kohonen and Timo Honkela. Kohonen network. *Scholarpedia*, 2(1), 2007.

[35] Nobuya Koike, Seung-Hee Yoo, Hung-Chung Huang, Vivek Kumar, Choogon Lee, Tae-Kyung Kim, and Joseph S Takahashi. Transcriptional architecture and chromatin landscape of the core circadian clock in mammals. *Science*, 338(6105), 2012.

[36] Neel Mehta and Hai-Ying M Cheng. Micro-Managing the Circadian Clock : The Role of microRNAs in Biological Timekeeping. *Journal of Molecular Biology*, 2012.

[37] Jaime Guevara-Aguirre, Priya Balasubramanian, Marco Guevara-Aguirre, Min Wei, Federica Madia, Chia-Wei Cheng, David Hwang, Alejandro Martin-Montalvo, Jannette Saavedra, Sue Ingles, Rafael de Cabo, Pinchas Cohen, and Valter D Longo. Growth hormone receptor deficiency is associated with a major reduction in pro-aging signaling, cancer, and diabetes in humans. *Science translational medicine*, 3(70), 2011.

[38] Zirong Li, Sara Van Calcar, Chunxu Qu, Webster K Cavenee, Michael Q Zhang, and Bing Ren. A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells. *Proceedings of the National Academy of Sciences of the United States of America*, 100 (14), 2003.

[39] J. Hall, M. Lee, B Newman, J. Morrow, L. Anderson, B Huey, and M. King. Linkage of early-onset familial breast cancer to chromosome 17q21. *Science*, 250(4988), 1990.

[40] I Nishisho, Y Nakamura, Y Miyoshi, Y Miki, H Ando, A Horii, K Koyama, J Utsunomiya, S Baba, and P Hedge. Mutations of chromosome 5q21 genes in FAP and colorectal cancer patients. *Science*, 253(5020), 1991.

[41] M H Little, J Prosser, A Condie, P J Smith, V Van Heyningen, and N D Hastie. Zinc finger point mutations within the WT1 gene in Wilms tumor patients. *Proceedings of the National Academy of Sciences of the United States of America*, 89(11), 1992.

[42] C Caldas, S A Hahn, L T da Costa, M S Redston, M Schutte, A B Seymour, C L Weinstein, R H Hruban, C J Yeo, and S E Kern. Frequent somatic mutations and homozygous deletions of the p16 (MTS1) gene in pancreatic adenocarcinoma. *Nature genetics*, 8(1), 1994.

[43] De-Ke Jiang, Lei Yao, Wei-Hua Ren, Wen-Zhang Wang, Bo Peng, and Long Yu. TP53 Arg72Pro polymorphism and endometrial cancer risk : a meta-analysis. *Medical oncology*, 28(4), 2011.

[44] Jing-Jun Wang, Yuan Zheng, Liang Sun, Li Wang, Peng-Bo Yu, Jian-Hua Dong, Lei Zhang, Jing Xu, Wei Shi, and Yu-Chun Ren. TP53 codon 72 polymorphism and colorectal cancer susceptibility : a meta-analysis. *Molecular biology reports*, 38(8), 2011.

[45] Jin-Mei Piao, Hee Nam Kim, Hye-Rim Song, Sun-Seog Kweon, Jin-Su Choi, Woo-Jun Yun, Young-Chul Kim, In-Jae Oh, Kyu-Sik Kim, and Min-Ho Shin. p53 codon 72 polymorphism and the risk of lung cancer in a Korean population. *Lung cancer*, 73(3), 2011.

[46] Shafika Alawadi, Lina Ghabreau, Mervat Alsaleh, Zainab Abdulaziz, Mohamed Rafeek, Nizar Akil, and Moussa Alkhalaf. P53 gene polymorphisms and breast cancer risk in Arab women. *Medical oncology*, 28(3), 2011.

[47] Takayuki Sonoyama, Akiko Sakai, Yuichiro Mita, Yukiko Yasuda, Hirofumi Kawamoto, Takahito Yagi, Masao Yoshioka, Tetsushige Mimura, Kei Nakachi, Mamoru Ouchida, Kazuhide Yamamoto, and Kenji Shimizu. TP53 codon 72 polymorphism is associated with pancreatic cancer risk in males, smokers and drinkers. *Molecular medicine reports*, 4(3), 2011.

[48] H Igaki, H Sasaki, T Kishi, H Sakamoto, Y Tachimori, H Kato, H Watanabe, T Sugimura, and M Terada. Highly frequent homozygous deletion of the p16 gene in esophageal cancer cell lines. *Biochemical and biophysical research communications*, 203(2), 1994.

[49] B N Ames, M K Shigenaga, and L S Gold. DNA lesions, inducible DNA repair, and cell division : three key factors in mutagenesis and carcinogenesis. *Environmental health perspectives*, 101 Suppl, 1993.

[50] Douglas Hanahan and Robert A Weinberg. The Hallmarks of Cancer. *Cell*, 100, 2000.

[51] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer : the next generation. *Cell*, 144(5), 2011.

[52] Philip J Stephens, Patrick S Tarpey, Helen Davies, Peter Van Loo, Chris D Greenman, David C Wedge, Serena Nik-Zainal, Sancha Martin, Ignacio Varela, Graham R Bignell, Lucy R Yates, Elli Papaemmanuil, David Beare, Adam P Butler, Angela Cheverton, John Gamble, Jonathan Hinton, Mingming Jia, Alagu Jayakumar, David Jones, Calli Latimer, King Wai Lau, Stuart McLaren, David J McBride, Andrew Menzies, Laura Mudie, Keiran Raine, Roland Rad, Michael Spencer Chapman, Jon Teague, Douglas F Easton, Anita Langerød, Ming Ta Michael Lee, Chen-Yang Shen, Benita Tan Kiat Tee, Bernice Wong Huimin, Annegien Broeks, Ana Cristina Vargas, Gulisa Turashvili, John Martens, Aquila Fatima, Penelope Miron, Suet-Feung Chin, Gilles Thomas, Sandrine Boyault, Odette Mariani, Sunil R Lakhani, Marc J van de Vijver, Laura van 't Veer, John A Foekens, Christine Desmedt, Christos Sotiriou, Andrew Tutt, Carlos Caldas, Jorge S Reis-Filho, Samuel a J R Aparicio, Anne Vincent Salomon, Anne-Lise Børresen-Dale, Andrea L Richardson, Peter J Campbell, P Andrew Futreal, and Michael R Stratton. The landscape of cancer genes and mutational processes in breast cancer. *Nature*, 486(7403), 2012.

[53] Martin Peifer, Lynnette Fernández-Cuesta, Martin L Sos, Julie George, Danila Seidel, Lawryn H Kasper, Dennis Plenker, Frauke Leenders, Ruping Sun, Thomas Zander, Roopika Menon, Mirjam Koker, Ilona Dahmen, Christian Müller, Vincenzo Di Cerbo, Hans-Ulrich Schildhaus, Janine Altmüller, Ingelore Baessmann, Christian Becker, Bram de Wilde, Jo Vandesompele, Diana Böhm, Sascha Ansén, Franziska Gabler, Ines Wilkening, Stefanie Heynck, Johannes M Heuckmann, Xin Lu, Scott L Carter, Kristian Cibulskis, Shantanu Banerji, Gad Getz, Kwon-Sik Park, Daniel Rauh, Christian Grütter, Matthias Fischer, Laura Pasqualucci, Gavin Wright, Zoe Wainer, Prudence Russell, Iver Petersen, Yuan Chen, Erich Stoelben, Corinna Ludwig, Philipp Schnabel, Hans Hoffmann, Thomas Muley, Michael Brockmann, Walburga Engel-Riedel, Lucia A Muscarella, Vito M Fazio, Harry Groen, Wim Timens, Hannie Sietsma, Erik Thunnissen, Egbert Smit, Daniëlle A M Heideman, Peter J F Snijders, Federico Cappuzzo, Claudia Ligorio, Stefania Damiani, John Field, Steinar Solberg, Odd Terje Brustugun, Marius Lund-Iversen, Jörg Sänger, Joachim H Clement, Alex Soltermann, Holger Moch, Walter Weder, Benjamin Solomon, Jean-Charles Soria, Pierre Validire, Benjamin Besse, Elisabeth Brambilla, Christian Brambilla, Sylvie Lantuejoul, Philippe Lorimier, Peter M Schneider, Michael Hallek, William Pao, Matthew L Meyerson, Julien Sage, Jay Shendure, Robert Schneider, Reinhard Büttner, Jürgen Wolf, Peter Nürnberg, Sven Perner, Lukas C Heukamp, Paul K Brindle, Stefan Haas, and Roman K Thomas. Integrative genome analyses identify key somatic driver mutations of small-cell lung cancer. *Nature Genetics*, 44(10), 2012.

[54] Vladimir Lazar, G Bianchini, N Ueno, R Agarwal, B Wang, Bhaskar Dutta, Lajos Pusztai, Y Qi, Fabrice Andre, and G Bala. A network-based, integrative study to identify core biological pathways that drive breast cancer clinical subtypes. *British Journal of Cancer*, 2012.

[55] Zemin Zhang. Genomic landscape of liver cancer. *Nature Genetics*, 44(10), 2012.

[56] Orit Rozenblatt-Rosen, Rahul C. Deo, Megha Padi, Guillaume Adelmant, Michael a. Calderwood, Thomas Rolland, Miranda Grace, Amélie Dricot, Manor Askenazi, Maria Tavares, Samuel J. Pevzner, Fieda Abderazzaq, Danielle Byrdsong, Anne-Ruxandra Carvunis, Alyce a. Chen, Jingwei Cheng, Mick Correll, Melissa Duarte, Changyu Fan, Mariet C. Feltkamp, Scott B. Ficarro, Rachel Franchi, Brijesh K. Garg, Natali Gulbahce, Tong Hao, Amy M. Holthaus, Robert James, Anna Korkhin, Larisa Litovchick, Jessica C. Mar, Theodore R. Pak, Sabrina Rabello, Renee Rubio, Yun Shen, Saurav Singh, Jennifer M. Spangle, Murat Tasan, Shelly Wanamaker, James T. Webber, Jennifer Roecklein-Canfield, Eric Johannsen, Albert-László Barabási, Rameen Beroukhim, Elliott Kieff, Michael E. Cusick, David E. Hill, Karl Münger, Jarrod a. Marto, John Quackenbush, Frederick P. Roth, James a. DeCaprio, and Marc Vidal. Interpreting cancer genomes using systematic host network perturbations by tumour virus proteins. *Nature*, 2012.

[57] Maria a Blasco. Telomeres and human disease : ageing, cancer and beyond. *Nature Reviews Genetics*, 6(8), 2005.

[58] Kan V Lu, Jeffrey P Chang, Christine a Parachoniak, Melissa M Pandika, Manish K Aghi, David Meyronet, Nadezda Isachenko, Shaun D Fouse, Joanna J Phillips, David a Cheresh, Morag Park, and Gabriele Bergers. VEGF inhibits tumor cell invasion and mesenchymal transition through a MET/VEGFR2 complex. *Cancer Cell*, 22(1), 2012.

[59] Geert Berx and Frans van Roy. Involvement of members of the cadherin superfamily in cancer. *Cold Spring Harbor perspectives in biology*, 1(6), 2009.

[60] James E Talmadge and Isaiah J Fidler. AACR centennial series : the biology of cancer metastasis : historical perspective. *Cancer Research*, 70(14), 2010.

[61] Claire M Vajdic, Limin Mao, Marina T van Leeuwen, Philippa Kirkpatrick, Andrew E Grulich, and Sean Riminton. Are antibody deficiency disorders associated with a narrower range of cancers than other forms of immunodeficiency ? *Blood*, 116(8), 2010.

[62] Masoud H Manjili, Nejat Egilmez, Keith L Knutson, Senthamil R Selvan, and Julie R Ostberg. Tumor Escape and Progression under Immune Pressure. *Clinical & Developmental Immunology*, 2012, 2012.

[63] Ralph J DeBerardinis, Julian J Lum, Georgia Hatzivassiliou, and Craig B Thompson. The biology of cancer : metabolic reprogramming fuels cell growth and proliferation. *Cell metabolism*, 7(1), 2008.

[64] Matthew E Hardee, Mark W Dewhirst, Nikita Agarwal, and Brian S Sorg. Novel imaging provides new insights into mechanisms of oxygen transport in tumors. *Current Molecular Medecine*, 9(4), 2009.

[65] Harold F Dvorak. Tumors : Wounds That Do Not Heal. *New England Journal of Medicine*, 315, 1986.

[66] SI I Grivennikov, FR R Greten, and Michael Karin. Immunity, inflammation, and cancer. *Cell*, 140(6), 2010.

[67] Chris Greenman, Richard Wooster, P Andrew Futreal, Michael R Stratton, and Douglas F Easton. Statistical analysis of pathogenicity of somatic mutations in cancer. *Genetics*, 173(4), 2006.

[68] Tobias Sjöblom, Siân Jones, Laura D Wood, D Williams Parsons, Jimmy Lin, Thomas D Barber, Diana Mandelker, Rebecca J Leary, Janine Ptak, Natalie Silliman, Steve Szabo, Phillip Buckhaults, Christopher Farrell, Paul Meeh, Sanford D Markowitz, Joseph Willis, Dawn Dawson, James K V Willson, Adi F Gazdar, James Hartigan, Leo Wu, Changsheng Liu, Giovanni Parmigiani, Ben Ho Park, Kurtis E Bachman, Nickolas Papadopoulos, Bert Vogelstein, Kenneth W Kinzler, and Victor E Velculescu. The consensus coding sequences of human breast and colorectal cancers. *Science*, 314 (5797), 2006.

[69] Laura D Wood, D Williams Parsons, Siân Jones, Jimmy Lin, Tobias Sjöblom, Rebecca J Leary, Dong Shen, Simina M Boca, Thomas Barber, Janine Ptak, Natalie Silliman, Steve Szabo, Zoltan Dezso, Vadim Ustyanksky, Tatiana Nikolskaya, Yuri Nikolsky, Rachel Karchin, Paul a Wilson, Joshua S Kaminker, Zemin Zhang, Randal Croshaw, Joseph Willis, Dawn Dawson, Michail Shipitsin, James K V Willson, Saraswati Sukumar, Kornelia Polyak, Ben Ho Park, Charit L Pethiyagoda, P V Krishna Pant, Dennis G Ballinger, Andrew B Sparks, James Hartigan, Douglas R Smith, Erick Suh, Nickolas Papadopoulos, Phillip Buckhaults, Sanford D Markowitz, Giovanni Parmigiani, Kenneth W Kinzler, Victor E Velculescu, and Bert Vogelstein. The genomic landscapes of human breast and colorectal cancers. *Science*, 318(5853), 2007.

[70] Gonçalo R Abecasis, David Altshuler, Adam Auton, Lisa D Brooks, Richard M Durbin, Richard a Gibbs, Matt E Hurles, Gil a McVean, and The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319), 2010.

[71] The International Cancer Genome Consortium. International network of cancer genome projects. *Nature*, 464(7291), 2010.

[72] Roberta Ricciarelli, Cristina D'Abramo, Sara Massone, Umberto Marinari, Maria Pronzato, and Massimo Tabaton. Microarray analysis in Alzheimer's disease and normal aging. *IUBMB life*, 56(6), 2004.

[73] Klaus H Kaestner, Catherine S Lee, L Marie Scearce, John E Brestelli, Athanasios Arsenlis, Phillip Phuc Le, Kristen a Lantz, Jonathan Crabtree, Angel Pizarro, Joan Mazzarelli, Deborah Pinney, Steve Fischer, Elisabetta Manduchi, Christian J Stoeckert, Gerard Gradwohl, Sandra W Clifton, Juliana R Brown, Hiroshi Inoue, Corentin Cras-Méneur, and M Alan Permutt. Transcriptional program of the endocrine pancreas in mice and humans. *Diabetes*, 52(7), 2003.

[74] T. R. Golub. Molecular Classification of Cancer : Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286(5439), 1999.

[75] Leping Li, CR R Weinberg, TA A Darden, and LG G Pedersen. Gene selection for sample classification based on gene expression data : study of sensitivity to choice of parameters of the GA/KNN method. *Bioinformatics*, 17(12), 2001.

[76] Kathryn M Munro and Victoria M Perreau. Current and future applications of transcriptomics for discovery in CNS disease and injury. *Neuro-Signals*, 17(4), 2009.

[77] Hugues de Lavallade, Pascal Finetti, Nadine Carbuccia, Jamshid S Khorashad, Aude Charbonnier, Letizia Foroni, Jane F Apperley, Norbert Vey, François Bertucci, Daniel Birnbaum, and Marie-Joëlle Mozziconacci. A gene expression signature of primary resistance to imatinib in chronic myeloid leukemia. *Leukemia Research*, 34(2), 2010.

[78] Marc J Van De Vijver, Yudong D He, Laura J Van't Veer, Hongyue Dai, Augustinus A M Hart, Dorien W Voskuil, George J Schreiber, Johannes L Peterse, Chris Roberts, Matthew J Marton, Mark Parrish, Douwe Atsma, Anke Witteveen, Annuska Glas, Leonie Delahaye, Tony Van Der Velde, Harry Bartelink, Sjoerd Rodenhuis, Emiel T Rutgers, Stephen H Friend, and René Bernards. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347(25), 2002.

[79] F Pagès, J Galon, M-C Dieu-Nosjean, E Tartour, C Sautès-Fridman, and W-H Fridman. Immune infiltration in human tumors : a prognostic factor that should not be ignored. *Oncogene*, 29(8), 2010.

[80] Michael J Heller. DNA microarray technology : devices, systems, and applications. *Annual Review of Biomedical Engineering*, 4, 2002.

[81] Gary Hardiman. Microarray platforms - comparisons and contrasts. *Pharmacogenomics*, 5(5), 2004.

[82] Huixia Wang, Xuming He, Mark Band, Carole Wilson, and Lei Liu. A study of inter-lab and inter-platform agreement of DNA microarray data. *BMC Genomics*, 6, 2005.

[83] Stanislav O Zakharkin, Kyoungmi Kim, Tapan Mehta, Lang Chen, Stephen Barnes, Katherine E Scheirer, Rudolph S Parrish, David B Allison, and Grier P Page. Sources of variation in Affymetrix microarray experiments. *BMC Bioinformatics*, 6, 2005.

[84] S Draghici, P Khatri, Aron C Eklund, and Zoltan Szallasi. Reliability and reproducibility issues in DNA microarray measurements. *Trends in Genetics*, 22(2), 2006.

[85] Christopher A Maher, Chandan Kumar-Sinha, Xuhong Cao, Shanker Kalyana-Sundaram, Bo Han, Xiaojun Jing, Lee Sam, Terrence Barrette, Nallasivam Palanisamy, and Arul M Chinnaiyan. Transcriptome sequencing to detect gene fusions in cancer. *Nature*, 458(7234), 2009.

[86] Xing Fu, Ning Fu, Song Guo, Zheng Yan, Ying Xu, Hao Hu, Corinna Menzel, Wei Chen, Yixue Li, Rong Zeng, and Philipp Khaitovich. Estimating accuracy of RNA-Seq and microarrays with proteomics. *BMC Genomics*, 10, 2009.

[87] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq : a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1), 2009.

[88] Alina Sîrbu, Gráinne Kerr, Martin Crane, and Heather J Ruskin. RNA-Seq vs Dual- and Single-Channel Microarray Data : Sensitivity Analysis for Differential Expression and Clustering. *PloS one*, 7(12), 2012.

[89] A Antoniou, P D P Pharoah, S Narod, H A Risch, J E Eyfjord, J L Hopper, N Loman, H Olsson, O Johannsson, Å ke Borg, B Pasini, P Radice, S Manoukian, D M Eccles, N Tang, E Olah, H Anton-Culver, E Warner, J Lubinski, J Gronwald, B Gorski, H Tulinius, S Thorlacius, H Eerola, Heli Nevanlinna, K Syrjäkoski, Olli-P Kallioniemi, D Thompson, C Evans, J Peto, F Lalloo, D G Evans, and Douglas F Easton. Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history : a combined analysis of 22 studies. *American Journal of Human Genetics*, 72(5), 2003.

[90] Clifford a Hudis. Trastuzumab–mechanism of action and use in clinical practice. *New England Journal of Medicine*, 357(1), 2007.

[91] Inca. La situation du cancer en France en 2011. Technical report, 2011.

[92] N Howlader, AM M Noone, M Krapcho, N Neyman, R Aminou, SF F Altekruse, CL L Kosary, J Ruhl, Z Tatalovich, H Cho, A Mariotto, MP P Eisner, Lewis, HS S Chen, EJ J Feuer, and KA A Cronin. SEER Cancer Statistics Review, 1975-2009 (Vintage 2009 Populations). Technical report, 2011.

[93] Berhan Genç, Aynur Solak, Neslin Sahin, and Aşkın Gülşen. Metastasis to the male breast from squamous cell lung carcinoma. *Case reports in oncological medicine*, 2013, 2013.

[94] Sumit Shah, Samir Bhattacharyya, Arnab Gupta, Apurb Ghosh, and Samindranath Basak. Male breast cancer : a clinicopathologic study of 42 patients in eastern India. *Indian journal of surgical oncology*, 3(3), 2012.

[95] H. A. Burris. Dual Kinase Inhibition in the Treatment of Breast Cancer : Initial Experience with the EGFR/ErbB-2 Inhibitor Lapatinib. *The Oncologist*, 9(suppl_3), 2004.

[96] Gerald M Higa and Jame Abraham. Lapatinib in the treatment of breast cancer. *Expert review of anticancer therapy*, 7(9), 2007.

[97] Kathy Miller, Molin Wang, Julie Gralow, Maura Dickler, Melody Cobleigh, Edith A Perez, Tamara Shenkier, David Cella, and Nancy E Davidson. Paclitaxel plus bevacizumab versus paclitaxel alone for metastatic breast cancer. *The New England journal of medicine*, 357(26), 2007.

[98] Alberto J Montero, Kiran Avancha, Stefan Glück, and Gilberto Lopes. A cost-benefit analysis of bevacizumab in combination with paclitaxel in the first-line treatment of patients with metastatic breast cancer. *Breast cancer research and treatment*, 132(2), 2012.

[99] David W Miles, Arlene Chan, Luc Y Dirix, Javier Cortés, Xavier Pivot, Piotr Tomczak, Thierry Delozier, Joo Hyuk Sohn, Louise Provencher, Fabio Puglisi, Nadia Harbeck, Guenther G Steger, Andreas Schneeweiss, Andrew M Wardley, Andreas Chlistalla, and Gilles Romieu. Phase III study of bevacizumab plus docetaxel compared with placebo plus docetaxel for the first-line treatment of human epidermal growth factor

receptor 2-negative metastatic breast cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 28(20), 2010.

[100] Charles M Perou, Therese Sørlie, M B Eisen, M Van De Rijn, Stefanie S Jeffrey, C A Rees, Jonathan R Pollack, D T Ross, H Johnsen, Lars A Akslen, O Fluge, A Pergamenschikov, C Williams, S X Zhu, P E Lønning, Anne-Lise Børresen-Dale, Patrick O Brown, and David Botstein. Molecular portraits of human breast tumours. *Nature*, 406(6797), 2000.

[101] Therese Sørlie, Charles M Perou, Robert Tibshirani, T Aas, S Geisler, H Johnsen, T Hastie, M B Eisen, M van de Rijn, Stefanie S Jeffrey, T Thorsen, H Quist, J C Matese, Patrick O Brown, David Botstein, P Eystein Lønning, and Anne-Lise Børresen-Dale. 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), 2001.

[102] J Chuck Harrell, Adam D Pfefferle, Nicole Zalles, Aleix Prat, Cheng Fan, Andrey Khramtsov, Olufunmilayo I Olopade, Melissa A Troester, Andrew C Dudley, and Charles M Perou. Endothelial-like properties of claudin-low breast cancer cells promote tumor vascular permeability and metastasis. *Clinical & experimental metastasis*, 2013.

[103] Zhiyuan Hu, Cheng Fan, Daniel S Oh, J S Marron, Xiaping He, Bahjat F Qaqish, Chad Livasy, Lisa a Carey, Evangeline Reynolds, Lynn Dressler, Andrew Nobel, Joel Parker, Matthew G Ewend, Lynda R Sawyer, Junyuan Wu, Yudong Liu, Rita Nanda, Maria Tretiakova, Alejandra Ruiz Orrico, Donna Dreher, Juan P Palazzo, Laurent Perreard, Edward Nelson, Mary Mone, Heidi Hansen, Michael Mullins, John F Quackenbush, Matthew J Ellis, Olufunmilayo I Olopade, Philip S Bernard, and Charles M Perou. The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics*, 7(96), 2006.

[104] Stefan Michiels, Serge Koscielny, and Catherine Hill. Prediction of cancer outcome with microarrays : a multiple random validation strategy. *Lancet*, 365(9458), 2005.

[105] R A Fisher. *Statistical methods for research workers*. Number no 5. 1925.

[106] Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean Y H Yang, and Jianhua Zhang. Bioconductor : open software development for computational biology and bioinformatics. *Genome Biology*, 5(10), 2004.

[107] Lei Xu, Aik Choon Tan, Daniel Q Naiman, Donald Geman, and Raimond L Winslow. Robust prostate cancer marker genes emerge from direct integration of inter-study microarray data. *Bioinformatics*, 21(20), 2005.

[108] Fangxin Hong, Rainer Breitling, Connor W McEntee, Ben S Wittner, Jennifer L Nemhauser, and Joanne Chory. RankProd : a bioconductor package for detecting differentially expressed genes in meta-analysis. *Bioinformatics*, 22(22), 2006.

[109] Erin M Conlon, Joon J Song, and Jun S Liu. Bayesian models for pooling microarray studies with multiple sources of replications. *BMC Bioinformatics*, 7, 2006.

[110] Hongyan Zhang, Haiyan Wang, Zhijun Dai, Ming-Shun Chen, and Zheming Yuan. Improving accuracy for cancer classification with a new algorithm for genes selection. *BMC Bioinformatics*, 13(1), 2012.

[111] Martin H van Vliet, Christiaan N Klijn, Lodewyk F a Wessels, and Marcel J T Reinders. Module-based outcome prediction using breast cancer compendia. *PloS One*, 2(10), 2007.

[112] Christine Desmedt, Benjamin Haibe-Kains, Pratyaksha Wirapati, Marc Buyse, Denis Larsimont, Gianluca Bontempi, Mauro Delorenzi, Martine J Piccart, and Christos Sotiriou. Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. *Clinical Cancer Research*, 14(16), 2008.

[113] Anna V Ivshina, Joshy George, Oleg Senko, Benjamin Mow, Thomas C Putti, Johanna Smeds, Thomas Lindahl, Yudi Pawitan, Per Hall, Hans Nordgren, John E L Wong, Edison T Liu, Jonas Bergh, Vladimir a Kuznetsov, and Lance D Miller. Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Research*, 66(21), 2006.

[114] Sherene Loi, Benjamin Haibe-Kains, Christine Desmedt, Pratyaksha Wirapati, Françoise Lallemand, Andrew M Tutt, Cheryl Gillet, Paul Ellis, Kenneth Ryder, James F Reid, Maria G Daidone, Marco A Pierotti, Els MJJ J J Berns, Maurice Phm Jansen, John A Foekens, Mauro Delorenzi, Gianluca Bontempi, Martine J Piccart, and Christos Sotiriou. Predicting prognosis using molecular profiling in estrogen receptor-positive breast cancer treated with tamoxifen. *BMC Genomics*, 9(1), 2008.

[115] Joel S Parker, Michael Mullins, Maggie C U Cheang, Samuel Leung, David Voduc, Tammi Vickery, Sherri Davies, Christiane Fauron, Xiaping He, Zhiyuan Hu, John F Quackenbush, Inge J Stijleman, Juan Palazzo, J S Marron, Andrew B Nobel, Elaine Mardis, Torsten O Nielsen, Matthew J Ellis, Charles M Perou, and Philip S Bernard. Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 27(8), 2009.

[116] Yudi Pawitan, Judith Bjöhle, Lukas Amler, Anna-Lena Borg, Suzanne Egyhazi, Per Hall, Xia Han, Lars Holmberg, Fei Huang, Sigrid Klaar, Edison T Liu, Lance Miller, Hans Nordgren, Alexander Ploner, Kerstin Sandelin, Peter M Shaw, Johanna Smeds, Lambert Skoog, Sara Wedrén, and Jonas Bergh. Gene expression profiling spares early breast cancer patients from adjuvant therapy : derived and validated in two population-based cohorts. *Breast Cancer Research : BCR*, 7(6), 2005.

[117] Marcus Schmidt, Daniel Böhm, Christian von Törne, Eric Steiner, Alexander Puhl, Henryk Pilch, Hans-Anton Lehr, Jan G Hengstler, Heinz Kölbl, and Mathias Gehrmann. The humoral immune system has a key prognostic impact in node-negative breast cancer. *Cancer Research*, 68(13), 2008.

[118] Renaud Sabatier, Pascal Finetti, Nathalie Cervera, Eric Lambaudie, Benjamin Esterni, Emilie Mamessier, Agnès Tallet, Christian Chabannon, Jean-Marc Extra, Jocelyne Jacquemier, Patrice Viens, Daniel Birnbaum, and François Bertucci. A gene expression

signature identifies two prognostic subgroups of basal breast cancer. *Breast Cancer Research and Treatment*, 126(2), 2011.

[119] Christos Sotiriou and Lajos Pusztai. Gene-expression signatures in breast cancer. *New England Journal of Medicine*, 360(8), 2009.

[120] Yi Zhang, Anieta M Sieuwerts, Michelle McGreevy, Graham Casey, Tanja Cufer, Angelo Paradiso, Nadia Harbeck, Paul N Span, David G Hicks, Joseph Crowe, Raymond R Tubbs, G Thomas Budd, Joanne Lyons, Fred C G J Sweep, Manfred Schmitt, Francesco Schittulli, Rastko Golouh, Dmitri Talantov, Yixin Wang, and John A Foekens. The 76-gene signature defines high-risk patients that benefit from adjuvant tamoxifen therapy. *Breast cancer research and treatment*, 116(2), 2009.

[121] Yamei Zhou, Christina Yau, Joe W Gray, Karen Chew, Shanaz H Dairkee, Dan H Moore, Urs Eppenberger, Serenella Eppenberger-Castori, and Christopher C Benz. Enhanced NF kappa B and AP-1 transcriptional activity associated with antiestrogen resistant breast cancer. *BMC cancer*, 7, 2007.

[122] T S Keshava Prasad, Renu Goel, Kumaran Kandasamy, Shivakumar Keerthikumar, Sameer Kumar, Suresh Mathivanan, Deepthi Telikicherla, Rajesh Raju, Beema Shafreen, Abhilash Venugopal, Lavanya Balakrishnan, Arivusudar Marimuthu, Sutopa Banerjee, Devi S Somanathan, Aimy Sebastian, Sandhya Rani, Somak Ray, C J Harrys Kishore, Sashi Kanth, Mukhtar Ahmed, Manoj K Kashyap, Riaz Mohmood, Y L Ramachandra, V Krishna, B Abdul Rahiman, Sujatha Mohan, Prathibha Ranganathan, Subhashri Ramabadran, Raghothama Chaerkady, and Akhilesh Pandey. Human Protein Reference Database–2009 update. *Nucleic Acids Research*, 37(Database issue), 2009.

[123] Samuel Kerrien, Bruno Aranda, Lionel Breuza, Alan Bridge, Fiona Broackes-Carter, Carol Chen, Margaret Duesbury, Marine Dumousseau, Marc Feuermann, Ursula Hinz, Christine Jandrasits, Rafael C Jimenez, Jyoti Khadake, Usha Mahadevan, Patrick Masson, Ivo Pedruzzi, Eric Pfeiffenberger, Pablo Porras, Arathi Raghunath, Bernd Roechert, Sandra Orchard, and Henning Hermjakob. The IntAct molecular interaction database in 2012. *Nucleic acids research*, 40(Database issue), 2012.

[124] I Xenarios, D W Rice, Lukasz Salwinski, M K Baron, Edward M Marcotte, and David Eisenberg. DIP : the database of interacting proteins. *Nucleic Acids Research*, 28(1), 2000.

[125] Andreas Zanzoni, Luisa Montecchi-Palazzi, Michele Quondam, Gabriele Ausiello, Manuela Helmer-Citterich, and Gianni Cesareni. MINT : a Molecular INTeraction database. *FEBS Letters*, 513(1), 2002.

[126] Andreas Ruepp, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Goar Frishman, Corinna Montrone, Michael Stransky, Brigitte Waegele, Thorsten Schmidt, Octave Noubibou Doudieu, Volker Stümpflen, and H Werner Mewes. CORUM : the comprehensive resource of mammalian protein complexes. *Nucleic acids research*, 36(Database issue), 2008.

[127] Arun K Ramani, Razvan C Bunescu, Raymond J Mooney, and Edward M Marcotte. Consolidating the set of known human protein-protein interactions in preparation for large-scale mapping of the human interactome. *Genome Biology*, 6(5), 2005.

[128] Jennifer Tsai, Razvan Sultana, Yudan Lee, Geo Pertea, Valentin Antonescu, Jennifer Cho, Babak Parvizi, and John Quackenbush. RESOURCERER : a database for annotating and linking microarray resources within and across species. *Genome Biology*, 2 (11), 2001.

[129] AT&T LabsResearch. Graphviz System for Network Visualization, 1988.

[130] Jesse Gillis, Meeta Mistry, and Paul Pavlidis. Gene function analysis in complex data sets using ErmineJ. *Nature Protocols*, 5(6), 2010.

[131] Chih-chung Chang and Chih-jen Lin. LIBSVM - A Library for Support Vector Machines, 2007.

[132] Kun Yu, Kumaresan Ganesan, Lay Keng Tan, Mirtha Laban, Jeanie Wu, Xiao Dong Zhao, Hongmin Li, Carol Ho Wing Leung, Yansong Zhu, Chia Lin Wei, Shing Chuan Hooi, Lance Miller, and Patrick Tan. A precisely regulated gene expression cassette potently modulates metastasis and survival in multiple solid cancers. *PLoS genetics*, 4 (7), 2008.

[133] Christos Sotiriou, Pratyaksha Wirapati, Sherene Loi, Adrian Harris, Steve Fox, Johanna Smeds, Hans Nordgren, Pierre Farmer, Viviane Praz, Benjamin Haibe-Kains, Christine Desmedt, Denis Larsimont, Fatima Cardoso, Hans Peterse, Dimitry SA A Nuyten, Marc Buyse, Marc J van de Vijver, Jonas Bergh, Martine J Piccart, and Mauro Delorenzi. Gene expression profiling in breast cancer : understanding the molecular basis of histologic grade to improve prognosis. *Journal of the National Cancer Institute*, 98(4), 2006.

[134] Irit Fishel, Alon Kaufman, and Eytan Ruppin. Meta-analysis of gene expression data : a predictor-based approach. *Bioinformatics*, 23(13), 2007.

[135] Benjamin Haibe-Kains, Christine Desmedt, Fanny Piette, Marc Buyse, Fatima Cardoso, Laura Van't Veer, Martine Piccart, Gianluca Bontempi, and Christos Sotiriou. Comparison of prognostic gene expression signatures for breast cancer. *BMC Genomics*, 9, 2008.

[136] M Thomassen, Q Tan, F Eiriksdottir, M Bak, S Cold, and TA Kruse. Comparison of gene sets for expression profiling : prediction of metastasis from low-malignant breast cancer. *Clin Cancer Res*, 18 Pt 1(13), 2007.

[137] Jindan Jianjun Yu, Daniel R Rhodes, Scott A Tomlins, Xuhong Cao, Guoan Chen, Rohit Mehra, Xiaoju Wang, Debashis Ghosh, Rajal B Shah, Sooryanarayana Varambally, Kenneth J Pienta, and Arul M Chinnaiyan. A polycomb repression signature in metastatic prostate cancer predicts cancer outcome. *Cancer Research*, 67(22), 2007.

[138] Kevin K Dobbin, Yingdong Zhao, and Richard M Simon. How large a training set is needed to develop a classifier for microarray data ? *Clinical Cancer Research*, 14(1), 2008.

COLOPHON

Ce document a été préparé à l'aide du logiciel de composition typographique LaTeX et de l'éditeur de texte Sublime Text 3. Il a été compilé via pdfTeX 3.1415926-1.40.10 (TeX Live 2009/Debian) sur un système GNU/Linux Ubuntu 13.04. Le template utilisé pour formater cette thèse est basé sur un template originel de Robert Castelo distribué sous licence GNU/GPL copyleft. La version actuelle est disponible sur http://github.com/MaxUlysse/myThesis/. Le pdf est disponible au téléchargement sur http://phd.ithake.eu/Thesis-M-Garcia.pdf.

Titre

Découverte de biomarqueurs prédictifs en cancer du sein par Intégration Transcriptome-Interactome

Résumé

L'arrivée des technologies à haut-débit pour mesurer l'expression des gènes a permis l'utilisation de signatures génomiques pour prédire des conditions cliniques ou la survie du patient. Cependant de telles signatures ont des limitations, comme la dépendance au jeu de données d'entrainement et le manque de généralisation. Nous proposons un nouvel algorithme, ITI (Garcia et al.[13]) pour extraire une signature generalisable prédisant la rechute métastatique dans le cancer du sein par superimposition d'un très large jeu de données d'interaction protèine-protèine sur de multiples jeux de données d'expression des gènes. Cette méthode ré-implemente l'algorithme Chuang et al.[10], avec la capacité supplémentaire d'extraire une signature génomique à partir de plusieurs jeux de donnés d'expression des gènes simultanément. Une analyse non-supervisée et une analyse supervisée ont été réalisés sur un compendium de jeux de donnés issus de puces à ADN en cancer du sein. Les performances des signatures trouvées par ITI ont été comparé aux performances des signatures préalablement publiées (Wang et al.[9], Van De Vijver et al.[78], Sotiriou et al.[133]). Nos résultats montrent que les signatures ITI sont plus stables et plus généralisables, et sont plus performantes pour classifier un jeu de données indépendant. Nous avons trouvés des sous-réseaux formant des complexes précédement relié à des fonctions biologiques impliquées dans la nétastase et le cancer du sein. Plusieurs gènes directeurs ont été détectés, dont *CDK1*, *NCK1* et *PDGFB*, certains n'étant pas déjà relié à la rechute métastatique dans le cancer du sein.

Mots-clés

Transcriptome; Interactome; Intégration de données; Réseaux de gène; classification; SVM; Signature; Biomarqueurs; Cancer; Cancer du Sein

Title

Biomarkers discovery in breast cancer by Interactome-Transcriptome Integration

Abstract

High-throughput gene-expression profiling technologies yeild genomic signatures to predict clinical condition or patient outcome. However, such signatures have limitations, such as dependency on training set, and lack of generalization. We propose a novel algorithm, Interactome-Transcriptome Interaction (ITI) (Garcia et al.[13]) to extract a generalizable signature predicting breast cancer relapse by superimposition of a large-scale protein-protein interaction data over several gene-expression data sets. This method re-implements the Chuang et al.[10] algorithm, with the added capability to extract a genomic signature from several gene-expression data sets simultaneously. A non-supervised and a supervised analysis were made with a breast cancer compendium of DNA microarray data sets. Performances of signatures found with ITI were compared with previously published signatures (Wang et al.[9], Van De Vijver et al.[78], Sotiriou et al.[133]). Our results show that ITI's signatures are more stable and more generalizable, and perfom better when classifying an independant dataset. We found that subnetworks formed complexes functionally linked to biological functions related to metastasis and breast cancer. Several drivers genes were detected, including *CDK1*, *NCK1* and *PDGFB*, some not previously linked to breast cancer relapse.

Keywords

Transcriptome; Interactome; Data Integration; Gene Networks; classification; SVM; Signature; Biomarkers; Cancer; Breast Cancer