Approches d'intégration de données haut débit sur l'interactome humain.

Ghislain Bidaut

Aix*Marseille université

École doctorale des Sciences de la Vie et de la Santé

Approches d'intégration de données haut débit sur l'interactome humain

THÈSE

présentée et soutenue publiquement le 13 Mars 2020

pour l'obtention d'une

Habilitation à Diriger les Recherches de l'Université Aix-Marseille

(mention bio-informatique)

par

Ghislain A. M. Bidaut

Composition du jury

Président :	Pr Jacques Van Helden	Université Aix-Marseille
Rapporteurs :	Dr Pascal Barbry	Institut de Pharmacologie Moléculaire et Cellulaire
	Dr Anaïs Baudot	Marseille Medical Genetics
	Pr Gaëlle Lelandais	Université Paris-Saclay
Examinateurs :	Dr Laurence Calzone	Institut Curie
	Dr Benno Schwikowski	Institut Pasteur

Centre de Recherche en Cancérologie de Marseille - Inserm U1068 - CNRS 7528 - Université Aix-Marseille - Institut Paoli Calmettes

Mis en page avec la classe thesul.

Remerciements

Je tiens à remercier chaleureusement mes mentors de master et de doctorat qui m'ont formé à l'analyse de données, les Pr Christophe Garcia (INSA Lyon) et Dr Michael Ochs (Fox Chase Cancer Center),

Le Dr Chris Stoeckert qui m'a accueilli en stage post-doctoral au CBIL à Université de Pennsylvanie à Philadelphie.

Le Dr Benno Schwikowski, qui m'a accueilli en stage post-doctoral à l'Unité de Biologie Systémique à l'Institut Pasteur, Paris.

Les Pr Jean-Paul Borg et Dr Françoise Birg, qui m'ont accueilli et recruté comme responsable de plateforme bioinformatique au CRCM de Marseille.

Merci à mon tuteur, le Pr Jacques Van Helden du TAGC, aux rapporteurs et examinateurs, ainsi qu'aux membres du Jury d'avoir relu et porté un jugement sur ce manuscrit.

Je dédie cette thèse
à ma chère épouse, Wahiba
à mes filles, Leila et Nora
à mes parents,
Et, bien sur, à mes amis.

Sommaire

Préambule

Cette thèse a pour objet de présenter mes travaux de recherche depuis mon entrée au Centre de Recherche en Cancérologie de Marseille en vue de l'obtention de l'Habilitation à Diriger des Recherches. Dans une première partie, je détaille l'ensemble de mes éléments de carrière, à savoir une présentation de mon parcours, la liste de mes activités d'encadrement et enseignements scientifiques. Ensuite, j'expose mes activités de recherche en deux volets principaux.

D'une part, je présente le bilan de mes recherches en intégration de données et biologie des systèmes et leur application en cancérologie. Plusieurs pipelines d'analyse de réseaux de gènes sont décrites, tels que ITI, CITI et HTS-Net.

D'autre part, des projets purement bioinformatiques sont décrits, notamment la mise en place d'un programme d'organisation des données scientifiques, Djeen.

L'introduction générale décrit plus en avant ces travaux et leur insertion dans le contexte scientifique général de recherche en cancérologie.

Sute à ce préambule, je dresse un bilan qualitatif de mon activité durant cette période, avec une liste de mes publications et des travaux effectués avec les étudiants et les ingénieurs que j'ai encadrés au CRCM.

Table des figures

Parcours professionnel

Mon parcours initial

Après avoir intégré l'Ecole Nationale Supérieure de l'Electronique et de ses Applications (ENSEA), j'ai eu l'occasion de suivre un cursus très intéressant en mathématiques appliquées, et notamment en analyse de données multidimensionnelles. Ensuite, j'ai suivi le Diplôme d'Etude Approfondies en traitement d'image (DEA Traitement de l'Image et du Signal) de l'Université de Cergy-Pontoise. C'est par cette voie détournée que je suis entré dans le domaine de bioinformatique après mon DEA. J'ai eu la chance de faire mon stage de DEA avec le Pr Christophe Garcia, au Computer Vision and Robotics Laboratory, le laboratoire de vision par ordinateur dirigé par le Pr George Tziritas en Grèce. Ces différentes compétences m'ont permis d'être recruté comme étudiant en thèse avec Dr Michael Ochs au Fox Chase Cancer Center, à Philadelphie, dans son équipe de bioinformatique nouvellement créée.

J'ai eu l'opportunité, une fois encore, de faire de la bioinformatique sous l'angle de l'analyse de données multidimensionnelles, en y ajoutant l'aspect bayésien. J'ai pu utiliser un nouvel algorithme, la Décomposition Bayésienne pour l'analyse de données d'expression génomique en cancérologie et de profils phylogénétiques pour faire de la comparaison fonctionnelle de génomes bactériens. Dans le cadre de mon doctorat, j'ai participé activement aux échanges scientifiques du Centre. J'ai également mis en pratique une approche traitement du signal, à base d'analyse par Transformée en Ondelettes, pour l'analyse de données de puces à ADN.

J'ai approfondi ces compétences lors de mon stage postdoctoral au Center for Informatics and Computational Biology de l'Université de Pennsylvanie à Philadelphie, avec la mise en place d'un projet d'intégration de données à grande échelle dans le transcriptome (11 lignées de cellules souches adultes, profilées sur 6 laboratoires distincts du consortium Stem Cell Genome Analysis Project). C'est à ce moment-là que j'ai pu professionnaliser mon parcours, en travaillant au sein d'un consortium, et en développant mes compétences technologiques.

Suite à ce stage postdoctoral, j'ai entamé un second postdoctorat à l'institut Pasteur, Paris, dans l'unité de Biologie des Systèmes dirigée par Benno Schwikowski. J'ai pu intégrer une équipe très motivante avec des compétences très larges et travailler sur un projet de bioinformatique très intéressant : faire de l'intégration de données interactions protéine-protéine-expression des gènes avec Cytoscape. C'est un projet géré par un consortium avec des participants de tous horizons et de plusieurs nationalités qui m'a permis de comprendre les enjeux d'un travail en équipe, sur lequel il est à la fois nécessaire de se faire une place tout en participant à la construction d'un projet très large.

Mise en place de la plateforme Cibi au CRCM

Après mon passage à l'Institut Pasteur en 2007, j'ai candidaté à l'Inserm et obtenu avec succès un financement Inserm-INCa, pour la mise en place d'une nouvelle équipe de bioinformatique au Centre de Recherche en Cancérologie de Marseille, dirigé à l'époque par Françoise Birg et aujourd'hui par Pr Jean-Paul Borg.

Cette nouvelle affectation m'a permis de m'installer sur Marseille et de monter mon groupe de recherche, dont l'activité a été également valorisée par la mise en place de la plateforme Bioinformatique Intégrative du CRCM (*CRCM Integrative Bioinformatics - Cibi*). J'ai monté des projets collaboratifs au CRCM et au niveau national. J'ai notamment monté un projet d'intégration de données avec l'équipe du Pr Daniel Birnbaum, projet soumis et financé par l'Inserm, ce qui m'a permis d'accueillir Maxime Garcia en thèse. Maxime a été mon premier thésard et a terminé sa thèse avec succès sur plusieurs publications. J'ai également travaillé avec Olivier Stahl sur la mise en place d'un système de gestion de données scientifiques, Djeen, ainsi qu'à la mise en place d'une structure de calcul. J'ai ensuite été recruté en tant qu'Ingénieur de Recherche par l'Université Aix-Marseille, puis rejoint par Samuel Granjeaud, Ingénieur de recherche Inserm.

Ces différentes réalisations m'ont apporté de la reconnaissance et permis d'être invité en tant que partenaire sur un projet Plan Cancer porté par Sophie Vasseur, alors chargée de recherche dans l'équipe Inserm de stress cellulaire et sur un projet INCa porté par Christophe Ginestier, chargé de recherche au CRCM. Sur ces deux projets, j'ai encadré des ingénieurs expérimentés (Claire Rioualen, Quentin Da Costa) dont l'un a ensuite été recruté à l'Institut Paoli Calmettes, et que je continue d'encadrer dans le cadre du développement des outils de séquençage à haut débit appliqué au diagnostique clinique. J'ai pu ensuite recruter une seconde doctorante (Lucie Khamvongsa-Charbonnier) qui travaille sur un sujet d'intégration chip-seq et RNA-Seq. Plus récemment, je viens d'être recruté en tant que partenaire sur un projet d'étude de données d'expression dans les gliomes sur lequel je vais recruter et encadrer un ingénieur d'étude. Plusieurs stages de master ont eu pour sujet les projets évoqués plus avant.

Enfin, je viens d'entamer un projet d'intégration à grande échelle de donnés omiques en cancérologie pour répondre à l'appel H2020 WP 2018-2019 qui finance la médecine personnalisée et la mise en place des thérapies du futur.

J'ai donc pu mettre en place une équipe de bioinformatique au CRCM et la lancer avec des projets innovants en intégration de données. Lors de cette mise en place, j'ai encadré plusieurs stagiaires de master (principalement en master M2 BBSG d'Aix-Marseille Université), mais aussi PolyTech'Nice. J'ai été amené à lever mes propres financement pour obtenir mon autonomie sur les projets, ce qui m'a permis de financer plusieurs ingénieurs et de développer différents pipelines bioinformatiques.

Titre des travaux

Ghislain Bidaut, *Ingénieur de Recherche*
— **Responsable de la plateforme Cibi** du Centre de Recherche en Cancérologie de Marseille (Institut Paoli-Calmettes, Université Aix- Marseille U105, Inserm U1068, CNRS UMR 7258)
— **Coordonnateur bioinformaticien** de l'Institut Paoli-Calmettes

Contact

— CRCM, 27 Boulevard Leï Roure CS30059 13273 Marseille Cedex 09
— Téléphone : (+33) 4 86 97 73 01
— Fax : (+33) 4 91 26 03 64
— Courriel : `ghislain.bidaut@inserm.fr`

Cursus Universitaire

— Sept. 2000 – Déc. 2004 : Doctorat en Bioinformatique : Université Aix-Marseille II. Diplôme obtenu avec mention très honorable. Thèse intitulée *Analyse de profils phylogéniques bactériens et de profils d'expression génétique par Décomposition Bayésienne.*
— Sept. 1999 – Sept. 2000 : Diplôme d'Etudes Approfondies Traitement des Images et du Signal (TIS) : Université de Cergy-Pontoise, France.
— Sept. 1995 – Sept. 1998 : Diplôme d'ingénieur ENSEA : Ecole Nationale Supérieure de l'Electronique et de ses Applications, Cergy-Pontoise, France.
— Sept. 1994 – Sept. 1995 : Classe préparatoire mathématiques spéciales T.S., Lycée Jean-Jaurès, Argenteuil.
— Sept 1992 – Sept 1994 : Brevet de Technicien Supérieur, Lycée A. Chérioux, Vitry sur Seine.
— Juin 1992 : Baccalauréat Technologique F3 : mention *assez bien*, Lycée A. Chérioux, Vitry sur Seine.

Expérience Professionnelle

— 2010 – Présent : Ingénieur de Recherche en Analyse de données Biologiques.
— Déc. 2007 - présent : Lauréat Inserm-INCa – Responsable de la plateforme Bioinformatique Intégrative du Centre de Recherche en Cancérologie de Marseille, Institut Paoli-Calmettes.
— Mai 2007 – Déc. 2007 : Post-doctorat en Bioinformatique au laboratoire de biologie systémique de l'institut pasteur, Paris, France (Laboratoire du Dr. Schwikowski).

— Mai 2005 – Avril 2007 : Post-doctorat en Bioinformatique : Computational Biology and Informatics Laboratory (CBIL), University of Pennsylvania School of Medicine – Philadelphie (Laboratoire du Dr. Christian D. Stoeckert)
— Déc. 2004 – Avril 2005 : Post-doctorat en Bioinformatique : Fox Chase Cancer Center – Philadelphie (Laboratoire du Dr. Michael F. Ochs) :
— Déc. 2000 – Déc. 2004 : Thèse en Bioinformatique : en cotutelle au Fox Chase Cancer Center – Philadelphie (Laboratoire du Dr. Michael F. Ochs), et au Laboratoire d'Information Génomique et Structurale.
— Avril 2000 – Sept. 2000 : Stage de DEA en analyse et traitement d'image : Institute of Computer Science-Foundation for Research and Technology-Hellas - ICS-FORTH, Héraklion (Laboratoire du Dr. Georges Tsiritas).

Enseignement et encadrement

Thèses de doctorat et de master

— 2009 : Maxime Garcia, Master 2 BBSG Aix Marseille Université. *Recherche de biomarqueurs robustes dans le cancer du sein par intégration transcriptome-interactome.*
— 2009-2012 : Maxime Garcia, Thèse de Doctorat. Ecole Doctorale des Sciences de la Vie et de la Santé. Aix Marseille Université. *Découverte de biomarqueurs prédictifs en cancer du sein par Intégration Transcriptome-Interactome.*
— 2009 : Frédéric Garcia, Master 2 BBSG Aix Marseille Université. *Recherche de signatures moléculaires par intégration à grande échelle de dépôt publiques dans le transcriptome – Projet MAT (Meta Analysis in the Transcriptome).*
— 2012 : Fanny Blondin, PolyTech'Nice Département Génie Biologique 5ème année, option bioinformatique et Modélisation pour la Biologie. *Développement de nouvelles fonctionnalités et Intégration de pipelines d'analyses au composant Djeen (Database for Joomla !'s Extensible ENgine).*
— 2012 : Raphaëlle Millat-Carus, Master 2 BBSG Aix Marseille Université. *Recherche de marqueurs de métastase en cancérologie par intégration de données d'expression, d'interaction et d'altération génomiques (projet CITI).*
— 2012 : Claire Rioualen, Master 2 BBSG Aix Marseille Université. *Analyse à haut débit en cytométrie en flux.*
— 2014 : Quentin Da Costa, Master 2 BBSG Aix Marseille Université. *Déploiement du pipeline ITI sous un portail MobileNet.*
— 2014 : Projet Informatique Appliqué à la Biologie en co-direction avec le Master BBSG Université Aix-Marseille. *Déploiement d'une boîte à outils transcriptome-proteome sous un portail Mobyle.*
— 2016- Lucie Khamvongsa-Charbonnier. Thèse de doctorat. Ecole Doctorale des Sciences de la Vie et de la Santé. Aix Marseille Université. *Analyse intégrative de la régulation combinant cistrome, épigénome et transcriptome.*
— 2018 : Victoria Vidal Tomas, Master 2 Universitat de València, Erasmus. *Interactome analysis of large scale gene expression data.*
— 2018 : Benoit Goutorbe, 5ème année, Département de Bioinformatique & Data Sciences INSA Lyon. *Recherche de signatures moléculaires en métagénomique par analyse de données mutifactorielles.*

Encadrement d'Ingénieurs sur projet

— 2009 - 2012 - Olivier Stahl, Ingénieur d'études en bioinformatique - Financement Inserm-INCa.
— 2012 - Samuel Granjeaud, Ingénieur de recherche en bioinformatique - permanent Inserm.
— 2012 - 2013, Guillaume Tiberi, Ingénieur d'études en bioinformatique. Financement Action Bioinformatique Cancéropôle.
— 2013-2015 - Claire Rioualen, Ingénieur d'études en bioinformatique. INCa - Projet Libre en Biologie II.
— 2014- Quentin Da Costa, Ingénieur d'études en bioinformatique. Projet Plan Cancer II.
— 2014-2016 - Abdessamad El Kaoutari, Ingénieur de recherche en bioinformatique - Financement Action Bioinformatique Cancéropôle.

Cours

— 2015 : PhD Program in data integration - Centro Interdipartimentale Molecular Systems Biology (Turin, Italy) - Responsable Dr Michele de Bortoli.
— 2013 - présent : Programme de Master professionnel d' Oncologie (Responsable : Pr François Bertucci et Pr Emmanuelle Charaffe) (Aix-Marseille Université), *Approches bioinformatiques à haut débit.*
— 2004 : T.A. (Teaching Assistant) à Drexel University (Philadelphie, U.S.A), Cours de Bioinformatique niveau DEA avec Dr. M.F. Ochs : Microarray and Differential Gene Expression, Pattern recognition and gene expression profiles, Gene and Protein Network.

Prix et financements

— Juillet 2017 : Obtention d'un financement INCa (80k€).
— Juin 2016 : Obtention d'un financement INCa pour la mise en place du NGS à visée diagnostique.
— Juillet 2013 : Obtention d'un financement Inserm ITMO Cancer (140k€). Juillet 2012 : Obtention d'un financement PL-Bio INCa (100k€).
— Décembre 2009 : Obtention d'un financement BQR Préciput de 17k€.
— Juin 2009 : Obtention d'un financement pour un chercheur de niveau postdoctoral (Université de la Méditerranée).
— Juin 2009 : Obtention d'une bourse de financement de trois ans pour un thésard Inserm-Région PACA.
— Novembre 2008 : Obtention d'une bourse d'installation de la Mairie de Marseille.
— Février 2008 : Obtention d'un financement de 60k€pour la réponse à l'appel d'offres Nouvelles Equipes lancé par la Fondation pour la Recherche Médicale.
— Juillet 2007 : Obtention d'un financement de 100k€sur 5 ans en réponse à l'appel d'offres Inserm-INCa pour l'établissement d'une équipe de bioinformatique au sein du Centre de Recherche en Cancérologie de Marseille, Institut-Paoli-Calmettes, France.

Présentation à des conférences nationales et internationales

Présentations Orales

— 2009 : Pacific Symposium of Biocomputing (Hawaii) : *Characterization of unknown adult stem cell samples by large scale data integration and artificial neural network.*
— 2013 : Network Biology Symposium (Institut Pasteur Paris) : *CNV-Interactome-Transcriptome Integration to Detect Driver Genes in Breast Cancer.*
— 2015 : Lecture on Doctorate Program for Advanced studies. University of Turin. *Big data accessibility, integra1on and analysis in life sciences.*
— 2013 : Word Biotechnology Congress (Boston) : *CNV-Interactome-Transcriptome Integration to Detect Driver Genes in Cancer.*
— 2013 : John Hopkins Seminar Series : *CNV-Interactome-Transcriptome Integration to Detect Driver Genes in Cancer.*
— 2007 : Institut Pasteur Paris : *High throughput analysis deciphers signaling pathway behavior and isolates developmental biomarkers in stem/progenitor cells.*
— 2007 CRCM : *High throughput analysis deciphers signaling pathway behavior and isolates developmental biomarkers in stem/progenitor cells.*
— 2011 : IPMC : *Interactome-Transcriptome Integration : An integrated network approach to decipher metastasis molecular basis in breast cancer.*
— 2010 : Retraite CRCM : *Intégration de données haut débit en cancérologie.*
— 2014, Retraite CRCM : *Integration of public omics databases to Detect Driver Genes in Breast Cancer.*
— 2012 CRCL : *ITI - Interactome-Transcriptome Integration An integrated network approach to decipher metastasis molecular basis in breast cancer.*
— 2008 : Journées Mathématiques et Cancer (CRCM) : *Methods for robust cancer markers detection through large scale data integration.*

Posters

— Journées Ouvertes en Biologie, Bioinformatique et Mathématiques 2011 (Institut Pasteur Paris, France). Olivier Stahl, Arnaud Guille, Fanny Blondin, Pascal Finetti, Samuel Granjeaud and Ghislain Bidaut. Djeen : A high throughput multi-technological Research Information Management System for the Joomla ! CMS
— Journées Ouvertes en Biologie, Bioinformatique et Mathématiques 2012 (Montpellier, France). Maxime Garcia, Pascal Finetti, François Bertucci, Daniel Birnbaum and Ghislain Bidaut. Interactome-Transcriptome Integration Uncovers More Stable and Better Performing Biomarkers in Breast Cancer.
— 11th European Conference on Computational Biology (Bâle, Suisse). Garcia, M., Millat-Carus, R., Bertucci, F., Finetti, P., Birnbaum, D., and Bidaut, G. Interactome-Transcriptome Integration Uncovers More Stable and Better Performing Biomarkers in Breast Cancer.
— 7th Santorini Conference “Biologie Prospective” Systems Medicine, Personalized Health and Therapy (Thira, Greece). C Rioualen, R El-Helou, E Charafe-Jauffret, C Ginestier, G Bidaut. Interactome–regulome–transcriptome integrative approach as a mean to disclose cancer stem cells regulatory circuits.
— 13th European Conference on Computational Biology (Strasbourg, France). C Rioualen, Q Da Costa, G Pinna, A Harel-Bellan, E Charafe-Jauffret, C Ginestier and G Bidaut. Interactome–regulome integrative approach for genome-wide screening data analysis in a breast cancer stem cells study.

— Retraite CRCM 2014 : C Rioualen, Q Da Costa, R El-Helou, E Charafe-Jauffret, C Ginestier, S Vasseur, F Guillaumond-Marchai, E Mas, I Crenon, G Pinna 6 , A Harel- Bellan, G Bidaut. Using a network approach to unravel biological pathways involved in cancer.
— 15th European Conference on Computational Biology (Strasbourg, France) : Q Da Costa, F Guillaumond, C Rioualen, S Beloribi-Djefaflia, V Gouirand, J Roques, I Crenon, E Mas, S Vasseur, G Bidaut. Time-dependent interactome – transcriptome analysis of PDAC tumorigenesis.
— Assises de Génétique 2018 (Nantes, France). Quentin Da Costa, Tetsuro Nogutchi, Audrey Remenieras, Violaine Bourdon, Cornel Popovici, Ghislain Bidaut, Hagay Sobol. Impact de la norme ISO-15189 sur la structuration des analyses bioinformatiques en Oncogénétique.
— JOBIM2018 (Marseille, France) Claire Rioualen, Quentin Da Costa, Bernard Chetrit, Emmanuelle Charafe-Jauffret, Christophe Ginestier, Ghislain Bidaut. Systems biology analysis of interaction and regulation networks in siRNA high throughput screenings.
— JOBIM2018 (Marseille, France) Quentin Da Costa, Samuel Granjeaud, Ghislain Bidaut. The CRCM Integrative Bioinformatics (Cibi) platform, an offer of service in latest bioinformatics technologies for large scale data analysis in biology.
— JOBIM2018 (Marseille, France) Benoit Goutorbe, Anne Plauzolles, Ghislain Bidaut, Philippe Halphon. Microbiote et santé : impact des nouveaux pipelines bio-informatiques en métagénomique ciblée.

Expertise

— Membre nommé du conseil de Laboratoire du Centre de Recherche en Cancérologie de Marseille (2008-2017).
— Membre de l'Action Bioinformatique Scructurante du Cancéropole PACA (2012-)
— Expert nommé par la Fondation sur la Recherche Médicale pour l'appel d'offres 2016 *Evaluation de l'impact des objets connectés sur la santé.*
— External Faculty for PhD Programme in Advanced Systems. Michele de Bortoli. University of Turin.
— Co-organisateur des séminaires *Mardis des Technologies Gourmandes* du CRCM avec Bernard Chetrit.

Revue de manuscrits

— Revue de manuscrits pour *BMC Bioinformatics*
— Revue de manuscrits pour *BMC Systems Biology*
— Revue de manuscrits pour *Bioinformatics*
— Revue de manuscrits pour *Genome Biology*
— Revue de manuscrits pour *PSB (Pacific Symposium in Biocomputing conférence* – sessions 2008 - 2009).
— Revue de manuscrits pour *IEEE/ACM Transactions on Computational Biology and Bioinformatics*
— Revue de manuscrits pour *Faculty of 1000*
— Revue de manuscrits pour *Frontiers*
— Revue de manuscrits pour *JOBIM 2018*

Publications

Publications à comité de lecture

— Arcangeli, M.-L., Bardin, F., Frontera, V., Bidaut, G., Obrados, E., Adams, R.H., Chabannon, C., and Aurrand-Lions, M. (2014). Function of Jam-B/Jam-C interaction in homing and mobilization of human and mouse hematopoietic stem and progenitor cells. Stem Cells 32, 1043–1054.

— Arnaud, C., Sebbagh, M., Nola, S., Audebert, S., Bidaut, G., Hermant, A., Gayet, O., Dusetti, N.J., Ollendorff, V., Santoni, M.-J., et al. (2009). MCC, a new interacting protein for Scrib, is required for cell migration in epithelial cells. FEBS Lett. 583, 2326–2332.

— Bekhouche, I., Finetti, P., Adelaïde, J., Ferrari, A., Tarpin, C., Charafe-Jauffret, E., Charpin, C., Houvenaeghel, G., Jacquemier, J., Bidaut, G., et al. (2011). High-resolution comparative genomic hybridization of inflammatory breast cancer and identification of candidate genes. PLoS ONE 6, e16950.

— Bidaut, G. (2007). Gene function inference from gene expression of deletion mutants. Methods Mol. Biol. 408, 1–18.

— Bidaut, G., and Ochs, M.F. (2004). ClutrFree : cluster tree visualization and interpretation. Bioinformatics 20, 2869–2871.

— Bidaut, G., and Stoeckert, C.J. (2009a). Characterization of unknown adult stem cell samples by large scale data integration and artificial neural networks. Pac Symp Biocomput 356–367.

— Bidaut, G., and Stoeckert, C.J. (2009b). Large scale transcriptome data integration across multiple tissues to decipher stem cell signatures. Meth. Enzymol. 467, 229–245.

— Bidaut, G., Suhre, K., Claverie, J.-M., and Ochs, M.F. (2005). Bayesian decomposition analysis of bacterial phylogenomic profiles. Am J Pharmacogenomics 5, 63–70.

— Bidaut, G., Manion, F.J., Garcia, C., and Ochs, M.F. (2006a). WaveRead : automatic measurement of relative gene expression levels from microarrays using wavelet analysis. J Biomed Inform 39, 379–388.

— Bidaut, G., Suhre, K., Claverie, J.-M., and Ochs, M.F. (2006b). Determination of strongly overlapping signaling activity from microarray data. BMC Bioinformatics 7, 99.

— Bonacci, T., Audebert, S., Camoin, L., Baudelet, E., Bidaut, G., Garcia, M., Witzel, I.-I., Perkins, N.D., Borg, J.-P., Iovanna, J.-L., et al. (2014). Identification of new mechanisms of cellular response to chemotherapy by tracking changes in post-translational modifications by ubiquitin and ubiquitin-like proteins. J. Proteome Res. 13, 2478–2494.

— De Grandis, M., Bardin, F., Fauriat, C., Zemmour, C., El-Kaoutari, A., Sergé, A., Granjeaud, S., Pouyet, L., Montersino, C., Chretien, A.-S., et al. (2017). JAM-C Identifies Src Family Kinase-Activated Leukemia-Initiating Cells and Predicts Poor Prognosis in Acute Myeloid Leukemia. Cancer Res. 77, 6627–6640.

— El Helou, R., Pinna, G., Cabaud, O., Wicinski, J., Bhajun, R., Guyon, L., Rioualen, C., Finetti, P., Gros, A., Mari, B., et al. (2017). miR-600 Acts as a Bimodal Switch that Regulates Breast Cancer Stem Cell Fate through WNT Signaling. Cell Rep 18, 2256–2268.

— Garcia, M., Millat-Carus, R., Bertucci, F., Finetti, P., Birnbaum, D., and Bidaut, G. (2012). Interactome-transcriptome integration for predicting distant metastasis in breast cancer. Bioinformatics 28, 672–678.

— Garcia, M., Finetti, P., Bertucci, F., Birnbaum, D., and Bidaut, G. (2014). Detection of driver protein complexes in breast cancer metastasis by large-scale transcriptome-interactome integration. Methods Mol. Biol. 1101, 67–85.

— Gondois-Rey, F., Granjeaud, S., Rouillier, P., Rioualen, C., Bidaut, G., and Olive, D. (2016). Multi-parametric cytometry from a complex cellular sample : Improvements and limits of manual versus computational-based interactive analyses. Cytometry A 89, 480–490.
— Gondois-Rey, F., Chéret, A., Granjeaud, S., Mallet, F., Bidaut, G., Lécuroux, C., Ploquin, M., Müller-Trutwin, M., Rouzioux, C., Avettand-Fenoël, V., et al. (2017a). NKG2C+ memory-like NK cells contribute to the control of HIV viremia during primary infection : Optiprim-ANRS 147. Clin Transl Immunology 6, e150.
— Gondois-Rey, F., Chéret, A., Mallet, F., Bidaut, G., Granjeaud, S., Lécuroux, C., Ploquin, M., Müller-Trutwin, M., Rouzioux, C., Avettand-Fenoël, V., et al. (2017b). A Mature NK Profile at the Time of HIV Primary Infection Is Associated with an Early Response to cART. Front Immunol 8, 54.
— Grant, J.D., Somers, L.A., Zhang, Y., Manion, F.J., Bidaut, G., and Ochs, M.F. (2004). FGDP : functional genomics data pipeline for automated, multiple microarray data analyses. Bioinformatics 20, 282–283.
— Guillaumond, F., Bidaut, G., Ouaissi, M., Servais, S., Gouirand, V., Olivares, O., Lac, S., Borge, L., Roques, J., Gayet, O., et al. (2015). Cholesterol uptake disruption, in association with chemotherapy, is a promising combined metabolic therapy for pancreatic adenocarcinoma. Proc. Natl. Acad. Sci. U.S.A. 112, 2473–2478.
— Ochs, M.F., Moloshok, T.D., Bidaut, G., and Toby, G. (2004). Bayesian decomposition : analyzing microarray data within a biological context. Ann. N. Y. Acad. Sci. 1020, 212–226.
— Ochs, M.F., Peterson, A.J., Kossenkov, A., and Bidaut, G. (2007). Incorporation of gene ontology annotations to enhance microarray data analysis. Methods Mol. Biol. 377, 243–254.
— Rioualen, C., Da Costa, Q., Chetrit, B., Charafe-Jauffret, E., Ginestier, C., and Bidaut, G. (2017). HTS-Net : An integrated regulome-interactome approach for establishing network regulation models in high-throughput screenings. PLoS ONE 12, e0185400.
— Secq, V., Leca, J., Bressy, C., Guillaumond, F., Skrobuk, P., Nigri, J., Lac, S., Lavaut, M.-N., Bui, T.-T., Thakur, A.K., et al. (2015). Stromal SLIT2 impacts on pancreatic cancer-associated neural remodeling. Cell Death Dis 6, e1592.
— Stahl, O., Duvergey, H., Guille, A., Blondin, F., Vecchio, A.D., Finetti, P., Granjeaud, S., Vigy, O., and Bidaut, G. (2013). Djeen (Database for Joomla !'s Extensible Engine) : a research information management system for flexible multi-technology project administration. BMC Res Notes 6, 223.
— Tiberi, G., Pekowska, A., Oudin, C., Ivey, A., Autret, A., Prebet, T., Koubi, M., Lembo, F., Mozziconacci, M.-J., Bidaut, G., et al. (2015). PcG methylation of the HIST1 cluster defines an epigenetic marker of acute myeloid leukemia. Leukemia 29, 1202–1206.
— Zangari, J., Partisani, M., Bertucci, F., Milanini, J., Bidaut, G., Berruyer-Pouyet, C., Finetti, P., Long, E., Brau, F., Cabaud, O., et al. (2014). EFA6B antagonizes breast cancer. Cancer Res. 74, 5493–5506.

Chapitres de livre

— Garcia M, Millat-Carus R, Bertucci F, Finetti P, Guille A, Adelaïde J, Bekhouche I, Sabatier R, Chaffanet M, Birnbaum D, Bidaut G, CNV-Interactome-Transcriptome Integration to detect driver genes in cancerology in Microarray Image and Data Analysis : Theory and Practice, CRC Press (2014) 12 : 331-338.
— M. Garcia, O. Stahl, P. Finetti, D. Birnbaum, F. Bertucci, and G. Bidaut (2011). Lin-

king interactome to disease : a network-based analysis of metastatic relapse in breast cancer. Handbook of Research on Computational and Systems Biology : Interdisciplinary Applications pp 406-427. L.A. Liu, D Wei, Y. Li, H. Lei Editors. IGI Global.

— G. Bidaut (2010) : Interpreting and comparing clustering experiments though graph visualization and ontology statistical enrichment with the ClutrFree package. Biomedical Informatics for Cancer Research. MF. Ochs, J.T. Casagrande, R.V. Davuluri Editors.

— G. Bidaut and C.J. Stoeckert (2009) : Large Scale Transcriptome Data Integration across Multiple Tissues to Decipher Stem Cell Signatures. Methods Enzymol. 2009 ;467 :229-45

— G. Bidaut (2007). Estimating gene function in gene deletion mutant, in Methods in Molecular Biology : Microarray Data Analysis, M. Korenberg, Editor, Humana Press, Totowa.

— A.V. Kossenkov, G. Bidaut, and M.F. Ochs (2005) : Estimating Cellular Signaling from Microarray Data, in K.A. Do, P. Mueller, M. Vannucci, Editors, Bayesian Inference for Gene Expression and Proteomics, Cambridge University Press.

— Kossenkov, G. Bidaut, and M.F. Ochs (2004) : Genes associated with prognosis in adenocarcinomas across studies at multiple institutions., in J. Shoemaker and S. Lin, Editors, Methods of Microarray Data Analysis IV, 239-253, Kluwer Academic, Boston.

— M.F. Ochs and G. Bidaut (2002) : Microarray Data Normalization, in Microarray Image Analysis - Nuts and Bolts, S. Shah and G. Kamerova, Editors, DNA Press, LLC : 131-154, London.

— G. Bidaut, J.D. Grant, T.D. Moloshok, F. J. Manion, M. F. Ochs (2002) : Bayesian Decomposition analysis of gene expression in yeast deletion mutants, in Methods of Microarray Data Analysis II, K. Johnson and S. Lin, Editors, Kluwer Scientifc, Boston.

Articles de Conférences

— G. Bidaut, K. Suhre, J.-M. Claverie, M. F. Ochs (2003) : Analysis of phylogenetic profiles using Bayesian Decomposition, in Proceedings of the 2003 IEEE Computer Society Bioinformatics Conference, The IEEE Computer Society.

— M. F. Ochs, T. D Moloshok, G. Bidaut, G. Toby (2003) : Bayesian Decomposition : Analyzing microarray data within a biological context, Proceedings of the National Cancer Institute Applications of Bioinformatics in Cancer Detection Workshop, Annals of the New York Academy of Sciences 1020 :212-26.

L'analyse et le traitement de données à haut débit en biologie moléculaire

1

Analyse du transcriptome et intégration de données

Sommaire

1.1 Introduction

L'analyse de la transcription des gènes à grande échelle est la clé de voûte de la compréhension de leur fonction et de la fonction des protéines qu'elles traduisent. Elle a représenté un intérêt majeur pour la bioinformatique depuis les années 2000, tout en servant de fondation pour les technologies d'analyse utilisées pour l'ensemble des *omiques* que nous étudions aujourd'hui. C'est particulièrement important en cancérologie, domaine où la transcription des gènes a été très étudiée et où de nombreux jeux de données ont été mis à disposition des chercheurs et cliniciens.

Des jeux de données d'organismes modèles ont été mis à disposition dès 2000, avec le jeux de données Hughes *et al.* [74]. Ce jeu de données contient les profils transcriptomiques de 300 levures ayant subi une simple ou double délétion de gènes. Ce type d'expérimentation a montré qu'une analyse unique permet de comprendre plusieurs centaines de fonctions moléculaires et génomiques simultanément. En cancérologie, nous avons assisté au développement de jeux de données d'expression issus de lignées cellulaires (*National Cancer Institute 60 Cell Line* (NCI60), *Cancer Cell Line Encyclopedia* et autres, interrogeables par l'outil *CellMiner* [143]). Des consortiums internationaux ont été formés dès 2005, avec notamment *The Cancer Genome Atlas* [30] et *The International Cancer Genome Consortium* [76]. Ils nous ont apporté de très larges jeux de données non seulement en expression mais aussi sur les autres domaine de régulation des gènes (analyse de la méthylation et des miRNA) et de leur produit (protéomique), avec des données cliniques, et ce pour un grand nombre de types de cancer, en ayant pour objectif de lier le devenir clinique des patients avec la dérégulation/mutations génomiques qu'ils portent.

1.2 Analyse de données à grand échelle par la bioinformatique

Je m'intéresse plus particulièrement aux méthodes d'analyse de données en transcriptomique. Dans ce domaine, les premières analyse se sont focalisées sur les approches non supervisées [137]. Les analyses de statistiques multidimensionnelles ont été largement mises à profit, notamment la classification ascendante hiérarchique. Celle-ci consiste à séparer les échantillons et/ou les gènes en classes, de façon non supervisée, en se basant sur la proximité (par exemple sur une mesure de corrélation) de leur profils d'expression. Par cette méthode, Alizadeh *et al.* [3] ont montré que des échantillons de lymphome analysés par microarrays se séparent en deux classes distinctes de lymphome à large cellules B, précédemment inconnues. Chacune de ces classes est représentative des étapes de différenciation des cellules B, l'une germinale, et l'autre exprimant les gènes normalement exprimées dans les cellules B du sang périphérique. L'analyse a montré que le pronostique des patients était lié à leur appartenance à l'une de ces deux classes.

Bien que ce type d'analyse non supervisée soit utile, il est plus intéressant d'un point de vue clinique de pouvoir mettre en évidence des marqueurs pronostiques ou de diagnostique qui peuvent guider le traitement. Golub *et al.* [58] ont montré qu'une analyse supervisée de profils d'expression pouvait permettre la classification de patients atteints de différentes formes de leucémies. Golub et collègues ont analysé des profils d'expression provenant de leucémies lymphoblastiques aiguës et de leucémies myéloblastiques aiguës et ont découvert que les profils d'expression obtenus correspondaient à ces deux maladies. Cette analyse a été le point de départ de la classification des patients et de la prédiction de leur pronostic par l'utilisation de profils d'expression. Avec un système de vote pondéré pour chaque gène, en fonction du pouvoir discriminant du profil d'expression de chaque gène à séparer les deux classes de maladies, les tumeurs des patients ont pu être assignées correctement aux deux classes, et il est donc possible d'influencer le traitement à partir de ces données.

C'est ce type d'études qui ont montré le fort lien entre l'état génomique du patient et sa situation clinique. Cela a conduit au développement de dizaines de méthodes ayant pour objectif l'analyse des profils d'expression pour l'identification de biomarqueurs. Les méthodes ne concernent pas seulement l'analyse de données mais aussi l'essor phénoménal de la biologie moléculaire, pour l'exploration des autres facettes du génome par l'étude des mutations et des réarrangements de grande taille dans l'ADN (par CGH-Arrays [130]), ainsi que l'étude de la régulation et de l'épigénome (par ChIP-Chip [133]). Les bioinformaticiens ont adapté et utilisé les méthodes précédemment développée en analyse de données, que ce soit en statistique avec la classification ascendante hiérarchique, en intelligence artificielle avec les réseaux de neurones [87] [166], les méthodes matricielles telles que la Factorisation Matricielle Non-négative (*Non-Negative Matrix factorisation* - NNF [28]) ou la Décomposition Bayésienne (supervisée [90] ou non [111, 124]).

1.3 Méthodes d'analyse d'enrichissement

Une fois la liste de gènes d'intérêt établie, il reste à en tirer toute l'information biologique pertinente. Les gènes ont été annotées lors de leur découvertes et analyses successives. Or, bien que les annotations manuelles soient très utiles pour l'étude des gènes, il apparaît rapidement que ces annotations sont limitées lorsque l'on travaille à l'échelle génomique. En effet, leur absence de structuration empêche de faire des requêtes sur la nature ou la fonction des gènes sans risquer de manquer une partie des gènes recherchés. Il en est de même lorsque l'on cherche à intégrer des données provenant d'expérimentation diverses. Une recherche de mutants par délétion dans

une base de données publiques type GEO [10] ou ArrayExpress [89] ne manquera pas de monter rapidement que trouver l'ensemble de ces expérimentation de façon exhaustive est complexe, sauf à utiliser plusieurs termes distincts pour le même objet biologique (Exemple pour *Knock-out, Knockout, Gene deletion mutant, KO, K.O..*, etc...).

Les annotations par vocabulaire contrôlé, ou *Ontologies* montrent rapidement tout leur intérêt dans ce contexte. Elles ont été établies rapidement pour les gènes et les protéines, notamment par le *Gene Ontology Consortium*[6]. Après ma thèse, en 2005, j'ai développé une méthode de classification de données d'expression pour la caractérisation de cellules souches et la compréhension de leur potentiel de différenciation dans le cadre d'une collaboration avec le consortium SCGAP (Stem Cell Genome Anatomy Project). Nous souhaitions trouver des signatures d'expression de gènes en fonction de leur potentiel de différenciation. L'analyse a donc nécessité la mise en place d'un vocabulaire contrôlé pour définir précisément le potentiel de différenciation des échantillons profilés pour chaque organe. J'ai donc mis en place une annotation normalisée pour l'ensemble des puces profilées sur les différents tissus en collaboration avec le consortium SCGAP.

Plus récemment, un effort similaire a conduit à la base de données *Gene Expression Atlas* [82]. Cette base de données regroupe de façon homogène l'ensemble des données grâce à des annotations expérimentales normalisées sur le type de tissu, l'organisme étudié et la condition expérimentale (pathologie). Le but de ce type de base de données est de pouvoir identifier quels gènes sont exprimés dans un tissu et de servir de référence.

Une fois ces ontologies mises en place, il est utile de mesurer un enrichissement des processus biologiques pour les gènes détectés lors d'une expérience. La méthode la plus simple consiste à comparer les distributions de chaque processus dans l'expérimentation avec celles d'une référence (le génome) par comptage des gènes correspondant à ces processus, ramenés au nombre de gène dans chaque groupe comparé. Ce comptage suit une loi hypergéométrique X, puisque ce calcul revient à faire une expérience aléatoire dont le processus consiste en un tirage aléatoire sans remise de n gènes (un groupe de gènes d'intérêt, une signature) dans un génome de taille N avec une probabilité p d'obtenir un gène du processus biologique testé. X est donc défini par $X\ H(N, n, p)$. Ce test nous donne une p-valeur à associer avec les valeurs d'enrichissement. Malgré sa simplicité, il permet néanmoins d'obtenir une première analyse fonctionnelle d'une liste de gènes et donc du résultat d'une expérience.

Plus récemment ont été développés des tests plus avancés, tels que le *Gene Set Enrichment Analysis* (GSEA)[160], qui permet l'analyse d'une liste de gènes différentiellement exprimés entre deux phénotypes, tout en tenant compte du fait que certains gènes faiblement dérégulés pouvaient avoir un impact sur un processus biologique systémique. C'est ce qui a été montré dans la publication originale GSEA [113] dans ce cadre du diabète de type II.

1.4 Présentation des chapitres de cette thèse

Je vais lister ici les différents chapitres qui composent cette thèse et qui constituent les jalons de la recherche que j'ai conduite depuis mon doctorat et au Centre de Recherche en Cancérologie de Marseille.

Le travail exposé chapitre 2, *Recherche d'une signature de différenciation somatique par intégration de données* présente un projet d'intégration de données d'expression. Lors de ma thèse, la plus grosse partie des méthodes développées et des efforts d'analyse portaient sur de l'analyse d'expression. Maintenant, ces méthodes ont naturellement évoluées en passant de méthodes focalisées sur les puces à ADN vers des méthodes d'intégration multi-technologies, ne portant pas

nécessairement sur de l'analyse d'expression mais sur des méthodes d'intégration multi-omiques. Lors de mon stage postdoctoral se posait la question de l'analyse intégrée de plusieurs jeux de données hétérogènes (plateforme et organismes hétérogènes). Cela passe par une normalisation des données qui soit adaptée à l'ensemble des données étudiées. Cette normalisation, associée à un apprentissage statistique par réseau de neurones a permis de déterminer une signature génomique de différenciation cohérente sur l'ensemble des cellules souches adultes.

Dans le chapitre 3, travail développé avec Maxime Garcia, le premier étudiant en thèse que j'ai eu l'opportunité d'encadrer, j'ai utilisé les réseaux d'interactions pour intégrer différents type des données. Ce chapitre, intitulé *Découverte de modules dans l'interactome humain liés aux mécanismes de métastase*, a permis de montrer que l'intégration verticale de plusieurs niveaux de données biologiques (ici, transcriptome et interactions protéines-protéines) associée à l'intégration horizontale de plusieurs jeux de données du même type biologique permettait d'améliorer la classification des patients sur une question biologique difficile telle que l'apparition de récidive métastatique dans le cancer du sein. Nous avons fait un apprentissage par Support Vector Machine des sous-réseaux différentiellement exprimés à partir de 930 tumeurs profilées sur 5 jeux de données Affymétrix et testé indépendamment le classifieur obtenu sur 130 tumeurs du jeu de données van De Vijver [175] et Desmedt [44]. ITI a montré un taux de succès supérieur à plusieurs signatures établies telles que le Genomic Grade Index, la signature Mammaprint, et la signature à 76 gènes de Wang *et al.* (le test de Kaplan-Meyer sur la survie sans métastase donne : $p < 1.10^{-5}$).

Dans le chapitre 4 de cette HDR, je présente un projet développé également avec Maxime Garcia et une autre étudiante, Raphaëlle Millat-Carus, alors en stage de M2 recherche, intitulé *Découverte de gènes drivers par analyse intégrée Interactome, transcriptome et génome*. Nous avons étendu le concept de recherche de sous-réseaux impliqués dans les mécanismes du cancer pour la recherche de gènes drivers et ajouté un niveau d'analyse supplémentaire, le nombre de copie d'ADN pour distinguer les gènes drivers (qui sont sélectionnes favorablement par la tumeur) des gènes passagers (dont l'expression est dérégulée de façon collatérale). La superposition des données d'expression, des données d'interactions et des nombre de copies permet d'identifier des modules de gènes (dans l'interactome) dérégulés (dans le transcriptome) par des gènes amplifiés ou effacées (dans le génome). De cette manière, nous avons pu identifier des gènes drivers sur deux sous types d'intérêt dans le cancer du sein (Luminal A, 80 tumeurs) et Basal (68 tumeurs).

Le chapitre 5 présente un projet original développé en tant que partenaire sur un projet INCa avec Dr Christophe Ginestier du CRCM, sur lequel j'ai encadré 2 ingénieurs d'études, Claire Rioualen et Quentin Da Costa, intitulé *Analyse de données de screening siRNA par HTS-Net : High-Throughput Screening - Network*. Ce projet a consisté à combiner l'expertise acquise sur les deux précédents projets en l'étendant aux mécanismes de régulation. Nous avons utilisé deux types de réseaux d'interactions. Des réseaux type interactome (interactions de données PPI comme dans les deux chapitres précédents) et un réseau de régulation TF-gènes (régulome). Avec ce dispositif et le programme HTS-Net, nous avons analysé des données de type siRNA pour comprendre les mécanismes de différenciation de cellules souches cancéreuses ainsi que les gènes impliqués dans les interactions virus-hôte dans l'hépatite C.

La gestion des données est une partie importante de la bioinformatique. Le Chapitre 6, *Djeen : Gestion de données génomiques simplifiées sur le Web* présente un système d'information de données de recherche développé avec Samuel Granjeaud et Olivier Stahl, ingénieurs sur la plateforme que je dirige. C'est un système permettant la gestion de données génomiques et de leurs annotations. Il comporte une vue par échantillons, une vue projet et une gestion fine des permissions utilisateurs. Ce travail a été publié et a fait l'objet d'un dépôt à l'Agence de la Protection des Programmes.

Le chapitre 7, *Interactome dans les cellules AML* présente un exemple d'application des analyse interactome à un projet d'analyse de donnés en cancérologie pour la compréhension des interactions entre le stroma et les cellules tumorales dans les leucémie aiguë myéloblastique (LAM). Ce travail a été développé avec Dr Michel Aurrand-Lyons et Dr Stéphane Mancini (CIML, CRCM) et présente comment un filtrage fin des données et une intégration réseaux PPI et expression dans plusieurs sous-types moléculaires permets de comprendre la mécanique des interaction entre gènes et la mise en place de modèles d'interactions entre cellules.

Chapitre 8, *Détermination des mécanismes de régulation génomique par intégration de données cistrome, épigénome et transcriptome* présente le projet de thèse de Lucie Charbonnier-Khamvongsa, que je co-encadre avec Pr Jacques van Helden du TAGC (Techniques avancées du Génome et de la Clinique) et de l'IFB (Institut Français de Bioinformatique). Ce projet consiste à concevoir un nouvel algorithme d'intégration de données RNA-Seq et ChIP-Seq pour découvrir des mécanismes de régulation impliqués dans le développement.

Enfin, le lecteur trouvera une conclusion à la fin de cette thèse.

2

Recherche d'une signature de différentiation somatique par intégration de données

Sommaire

Ce chapitre reprend mon travail sur l'identification d'une signature cellule souches universelle. Cette analyse transcriptomique intégrée, que j'ai développé durant mon stage postdoctoral à l'université de Pennsylvanie, au Laboratoire CIBL dirigé par Dr Chis Stoeckert, mon mentor de post-doc, a été publiée dans deux articles [18][22]. Elle a été à l'origine du développement du programme ITI (Intégration Transcriptome-Interactome).

2.1 Introduction

Une grande variété de cellules souches a récemment été découverte dans un large nombre d'organes et on suspecte leur présence dans la plupart des tissus. Les types les plus connus sont les cellules souches hématopoïétiques, les cellules souches neuronales, les progéniteurs myogéniques ainsi que d'autres ayant un potentiel plus restreint, telles que les cellules souches présentes dans les intestins et la peau [29]. Le consensus actuel est que les décisions moléculaires déterminant le sort des cellules sont déclenchées par plusieurs mécanismes distincts entre différents types de cellules souches dans le même organisme [110] ou entre différentes espèces[139].

Deux questions clés restent non ou mal résolues à ce jour. D'une part, bien que leur existence soit connue, la localisation exacte des cellules souches ainsi que leur capacité exacte de différenciation (quels sous types moléculaires sont-il capables d'engendrer ?) ne sont pas bien définis pour la plupart des organes adultes. Il faudrait donc mettre au point des méthodes de détection de cellules souches ou progéniteurs. Par exemple, on ne sait pas si les cellules bêta pancréatiques résident dans l'épithélium ductal, les petites cellules pancréatiques ou les acinus, ou l'ensemble de ces régions [152]. D'autre part, la liste des gènes *drivers* à l'origine de la différenciation ne fait pas l'objet d'un consensus.

Dans ce chapitre, je présente un pipeline bioinformatique permettant la découverte des mécanismes de différenciation communs entre plusieurs types cellulaires à partir de données de type transcriptome. Ce pipeline comporte deux parties principales. D'une part, l'intégration des données microarrays de différent tissus obtenues à partir des données collectées lors d'études transcriptomiques faites sur différents tissus, et d'autre part l'apprentissage d'un réseau de neurones artificiel pour l'extraction d'une liste de gènes marqueurs de différenciation et impliqués dans les prises de décision du sort cellulaire. L'hypothèse que nous souhaitons vérifier est que les voies de signalisation sont au moins partiellement conservées entre des progéniteurs ou cellules souches de type distincts, formant une signature moléculaire liée à la plasticité des cellules souches embryonnaires et adultes. Cette propriété est appelée *stemness* en anglais. Cette hypothèse, décrite dans plusieurs publications [78], est supportée notamment par les données du transcriptome [134]. La découverte d'une signature conservée pourrait aider à caractériser les propriétés de différenciation de cellules de types inconnu ou mal connu, ainsi que de caractériser leur potentiel exact ou leur état de différenciation. De plus, la publication d'un catalogue de marqueurs de différenciation associés à chaque type cellulaire et à leur potentiel de différenciation serait une ressource à haute valeur ajoutée pour les chercheurs en biologie du développement et en cellules souches.

Pour extraire les signatures moléculaires propres aux différentes étapes de différenciation pour différents types de cellules souches ou progéniteurs par analyse de profils d'expression, nous avons effectué l'apprentissage d'un réseau de neurones artificiel (RNA) multiclasses simple couche. Cet apprentissage a permis ensuite de caractériser des échantillons inconnus et de les positionner dans une hiérarchie s'étalant des cellules souches totipotentes au cellules souches pleinement différenciées. Ces prédictions faites par notre réseau de neurones ont ensuite été testées sur deux tissus partiellement caractérisés (Figure 2.2). L'épithélium de l'estomac de la souris et la prostate dans l'humain. La capacité de notre système à généraliser à d'autres types cellulaires a été évaluée par une procédure de validation croisée sur le jeu de données d'apprentissage.

Les réseaux de neurones artificiels (RNA) représentent une classe d'algorithmes d'apprentissage ayant été appliqués à la résolution d'un large nombre de problèmes ouverts. Leur structure basique est inspirée de la neurobiologie et prends la forme d'un réseau de noeuds (les neurones) à boucle de rétroaction et caractérisés par une fonction de transition. Typiquement, une topologie de réseaux est initialement choisit en tenant compte de la nature du problème. Les neurones font ensuite l'objet d'un apprentissage sur plusieurs époques, permettant de générer un modèle de réseau neuronal, qui est en retour appliqué à la classification de réseaux inconnus. Une époque est définie par la soumission (dans un ordre aléatoire) de l'ensemble des éléments du jeu d'apprentissage au réseau. En biologie, ils ont été appliqués pour la classification d'échantillons tumoraux ainsi que pour la découverte de biomarqueurs [61] [60].

Pour faire l'apprentissage de ce système de classification, nous avons dû normaliser les différentes sources de données, labelliser de façon consistante les échantillons de cellules souches ainsi que leurs propriétés de différenciation en utilisant un vocabulaire contrôlé [169]. Dans cette optique, la hiérarchie des cellules souches hématopoïétiques [78] a été utilisée comme modèle.

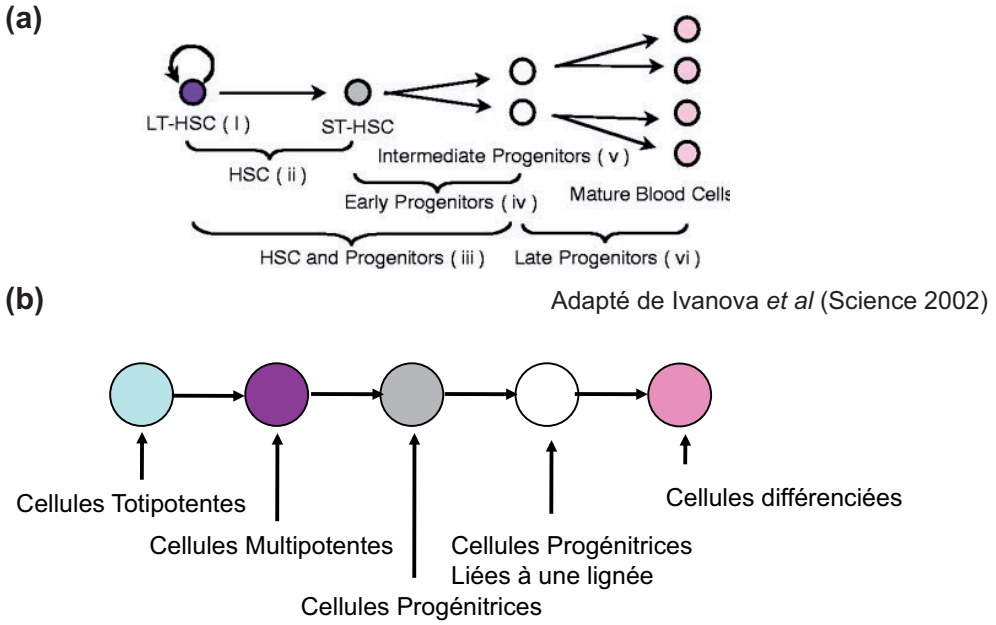


FIGURE 2.1 – (A) Hiérarchie des cellules souches dans le système hématopïétique. Les cellules les moins différenciées (Long-Term Hematopietic Stem Cells) possèdent le potentiel de générer tous les types cellulaires hématopoïétiques et s'auto-renouveler. Les ST-HSC et autres progéniteurs ont des capacités de génération d'un nombre restreint de type cellulaires mais ne peuvent s'auto-renouveler. Ces progéniteurs se différencient en différents types de cellules matures présentant le phénotype final. (B) Le modèle précédent a été généralisé à l'ensemble des tissus, et la hiérarchie a été définie précisément en fonction de leur potentiel de renouvellement. Ceci a donné lieu à un vocabulaire contrôlé permettant l'intégration de plusieurs jeux de données.

Dans cette classification, les différents types cellulaires identifiés grâce à des études fonctionnelles ont été positionnés de façon hiérarchique par rapport à leur potentiel de différenciation. Le haut de la hiérarchie est occupé par les cellules à haut potentiel de différenciation (Cellules souches hématopoïétiques), qui peuvent générer l'ensemble des types cellulaires sanguin, y compris s'auto-renouveler, tandis que le bas est occupé par les cellules matures complètement différenciées (Figure 2.1 (a)). Pour placer les autres types cellulaires, nous avons généralisé ce modèle pour y inclure notamment les cellules souches adultes ou les cellules souches embryonnaires. Une hiérarchie à 5 étages, stable sur l'ensemble des tissus étudiés, a été mise en place, comme précisé Table 2.1 (Figure 2.1 (b)).

Le pipeline fonctionne en trois étapes, et est détaillé figure 2.2. D'abord, nous avons généré un jeu de données d'apprentissage par l'intégration de plusieurs jeux de données d'expression généré par le consortium *Stem Cell Genome Anatomy Project* (SCGAP). Ces jeux de données ont été générés dans différent tissus et ont été re-normalisés et placés sur la hiérarchie à 5 niveau. Ensuite, nous avons projeté ces données dans un espace prédéfinis par des vecteurs formant une base non-orthogonale utilisant la technique de projection vectorielle[151], permettant de détecter les gènes ayant des profils de différenciation similaires. Les vecteurs de base choisis (voir figure) permettent de détecter les gènes propres à chaque étape de différenciation. Ce jeu de données intégré a été utilisé pour l'apprentissage d'un réseau de neurones multiclasses simple couche dans l'objectif de placer les cellules souches inconnues dans une des catégories prédéfinies.

La validation croisée sur le jeu de données d'apprentissage a permis l'identification de 31

Code	Type de Cellule Souche	Propriétés
A	Cellules Souches Totipotentes	Capables de s'auto-renouveler et de générer tous les types cellulaires
B	Cellules Souches Multipotentes	Capables de s'auto-renouveler et de générer la plupart des types cellulaires
C	Cellules Souches Progénitrices	Capables de générer plusieurs types cellulaires
D	Cellules Progénitrices Déterminées pour une lignée	Capables de générer un seul ou un nombre restreint de types cellulaires
E	Cellules Différentiées	Cellules caractérisées par leur phénotype final

TABLE 2.1 – Propriétés de la hiérarchie des cellules souches et définition du vocabulaire contrôlé utilisé pour leur classification.

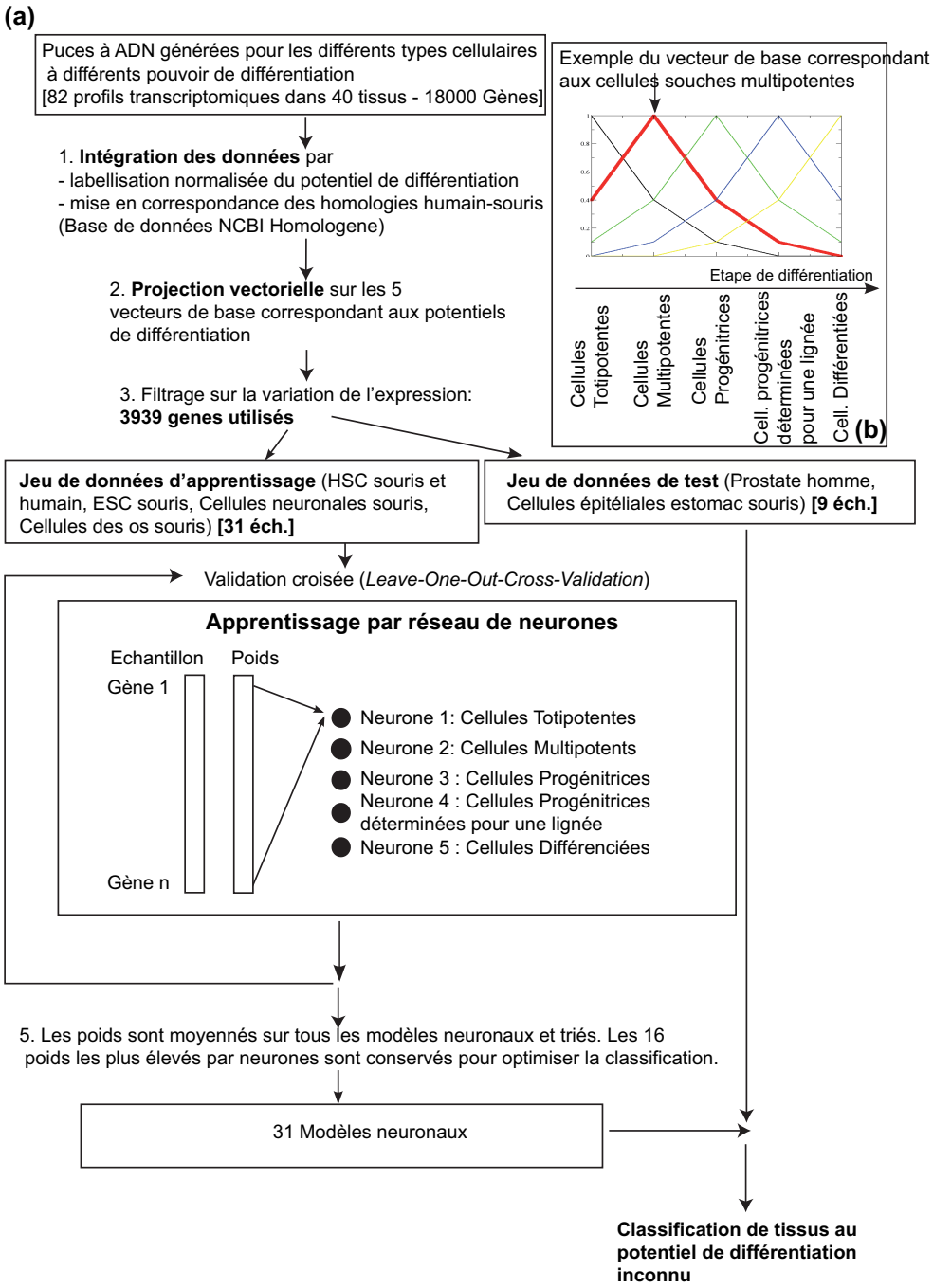


FIGURE 2.2 – (A) Architecture du réseau de neurones utilisé dans l'analyse. Les données sont d'abord annotées de façon normalisée par vocabulaire contrôlé puis par la mise en place d'une table de correspondance entre gènes homologues humain et souris. Le jeu de données est ensuite divisé entre partie apprentissage et jeu de données de test. Une procédure par validation croisée (*Leave-one-out-cross-validation*) est utilisée pour généraliser le modèle. Les modèles obtenus par validation croisées sont moyennés et ensuite testés sur les profils transcriptomiques de progéniteurs souris estomac et cellules de prostate humaine pour les caractériser. (B) Cette figure montre les vecteurs de bases utilisé pour la projection vectorielle. Chaque vecteur décrit un profil "idéal" de transcription adapté à détecter les gènes étant le plus exprimés à un niveau donné.

Premier auteur de la publication ou de l'étude.	Tissu cellulaire	Plateforme transcriptome (microarray)	Couverture génomique (Nombre de gènes)	Catégories (Potentiel de différenciation)
[125]	Cellules du Foie Embryonnaires (Souris)	Affy. 430A	12798	C,E
Rowe *et al.* (unpublished)	Cellules Progénitrices des Os (Souris)	Affy. U74 Av2, Bv2, Cv2	15127	C,D,E
[78]	Cellules Fœtales du Foie (Humain) (HSCs)	Affy. U95 Av2,B,C,D,E	17024	B,D,E
[78]	Cellules Fœtales du Foie (Souris) (HSCs)	Affy. U74 Av2, Bv2, Cv2	15127	B,D,E
[78]	Cellules Adultes de la Moelle Épinière (Souris)	Affy. U74 Av2, Bv2, Cv2	15127	B,C,D,E
[78]	Cellules Souches Embryonnaires (ESCs)	Affy. U74 Av2, Bv2, Cv2	15127	A
[78]	Cellules Souches Neurales (Souris) (NSCs)	Affy. U74 Av2, Bv2, Cv2	15127	B
Ivanova *et al.* (unpublished)	Cellules du cordon Ombilical (Homme) (HSCs)	Affy. U133 A,B	17275	B,D
Ivanova *et al.* (unpublished)	Cellules Adultes de la Moelle Épinière (Homme) (HSCs)	Affy. U133 A,B	17275	B,D
Ivanova *et al.* (unpublished)	Cellules Adultes de la Moelle Épinière (Souris) (HSCs)	Affy. 430 A,B	18626	B,C
[128]	**Cellules Progénitrices Prostate (Humain)**	Affy. U133plus2.0	18806	X,E
[106]	**Cellules Progénitrices Estomac (Souris)**	Affy. Mu11K A,B	6975	X,E
Total : 5 distinct	12 distinct	6 distinct	18720 gènes	5 distinct

TABLE 2.2 – Jeux de données utilisés dans l'analyse, avec les informations concernant les type de tissus profilés, les plateformes transcriptome, ainsi que les couvertures génomiques correspondantes. Les deux derniers tissus marqués en gras sont utilisés pour les tests. Les catégories de cellules souches (correspondant aux potentiels de différenciation cellulaires définies Table 2.1) sont marquées avec le code correspondant (De A à E).

modèles de réseaux neuronaux indépendants, par apprentissage sur le jeu de données avec censure itérative sur un des tissus. Le nombre optimal de gènes à utiliser pour la classification a été trouvé en refaisant l'apprentissage itérativement tout en réduisant le nombre de gènes. A chaque itération, les marqueurs pertinents ont été identifiés par classement des poids du réseau neuronal et gardés pour l'itération suivante. L'erreur de classification est minimale pour 63 gènes, et c'est cette liste de gènes qui a été retenue comme signature minimum représentant la propriété de *stemness* partagées par l'ensemble des cellules souches présentes dans les données. Finalement, nous avons testé le pouvoir prédictif de ces 31 modèles sur les deux tissus de progéniteurs de l'estomac[106] (souris) et prostate [128] (humain).

Les différents jeux de données sont de type microarray, souris ou humain, mesurant l'expression de cellules souches ou progéniteurs à différentes étapes de différenciation. La couverture génomique varie suivant le type de puce (de 6000 à 18000 gènes, voir table 2.2). Les données ont été d'abord intégrées par l'utilisation de la base NCBI *Homologene* ([121]) puis labellisées sur le vocabulaire défini Table 2.1. La matrice résultante contient 18,720 gènes profilés sur 82 échantillons au total. Les 82 échantillons sont groupés par tissus, et un tissu unique est caractérisé par un groupe d'au moins deux puces, ceci pouvant aller jusqu'à 5 puces, correspondant aux catégories A-E définies Table 2.1. C'est la condition minimale pour faire une projection vectorielle, telle que définie dans [151]. Les échantillons manquants sont simplement ignorés lors de la projection. Nous avons donc une table finale de projection de 40 tissus.

L'architecture des réseaux utilisés permettent l'exploration de deux paramètres importants. D'une part, classer les poids des gènes pour chaque étape de différenciation permet l'extraction des gènes reportées par le système pour être des marqueurs liés à cette étape. D'autre part, classer des gènes sur une valeur y résultants de la projection de l'expression multipliée par le poids correspondant à une population cellulaire donnée ($y_n = w_n.p_n$, w étant le poids le plus élevé et p l'expression de la projection) permet d'identifier les profils d'expression qui sont critiques pour la classification de ce tissu. De plus, nous avons fait une analyse d'enrichissement fonctionnel des gènes de rang le plus élevé et statistiquement enrichies pour chacune des étapes de différenciation.

Une couche cachée supplémentaire n'a pas été ajoutée au réseau neuronal, bien qu'elle aurait pu réduire l'erreur de classification, car elle nous aurait fait perdre la capacité d'interprétation des poids du réseau de neurone associées à chaque étape de différenciation. De même, d'autre types de classifieurs ont été considérés mais finalement mis de coté car ils ne nous auraient pas permis de faire une analyse fine des marqueurs liés à chaque étape de différenciation.

2.2 Matériel et Méthodes

2.2.1 Jeu de données

Le jeu de données est une version intégrée des données générées par les 7 membres du consortium Stem Cell Genome Anatomy Project (SCGAP). Il a permis l'apprentissage et le test indépendant du système. Chaque membre de SCGAP a généré un jeu de données microarray (Affymetrix, Santa Clara, CA, USA) dans un tissu particulier sur deux organismes distincts : *Mus Musculus* et *Homo Sapiens*, après purification spécifique (Table 2.2).

Les données sont disponibles pour recherche interactive sur le portail Web du consortium SCGAP [1]. Les données de chaque membre du consortium SCGAP sont également disponibles.

1. http ://www.scgap.org

Les mêmes données sont disponibles sous forme intégrée a partir du site compagnon de cette étude (http://scann.sourceforge.net). Ce site contient également le code source du pipeline.

2.2.2 Procédure d'intégration de données

Les données ont été intégrée à deux niveaux pour tenir compte de la nature des différentes plateformes microarray utilisées dans SCGAP. les données ont été d'abord normalisées (normalisation *lowess*, package Bioconductor [2] *Affy*). La méthode de summarisation *mas* nous a apporté les valeurs d'expression. Nous avons ensuite labellisé les différents échantillons d'apprentissage en utilisant le vocabulaire contrôlé décrivant les 5 étapes de différenciation.

2.2.3 Projection vectorielle

La projection vectorielle est une technique qui a été précédemment utilisée pour l'extraction de profils de gènes particuliers dans des données d'expression [151]. Cette technique permet l'identification rapide de gènes à partir d'une série de profils d'expression en modélisant le comportement (et donc le profil d'expression) des gènes que nous souhaitons isoler et que nous considérons comme plus représentatives. Dans le cadre de cette étude, l'emploi de la projection vectorielle a rendu possible la capture de gènes dont le profil d'expression comporte un pic lors des différentes étapes de différenciation pour un couple tissu/organisme donné (Figure 2.2). La Définition mathématique de la projection vectorielle est le produit scalaire d'une profil d'expression de gènes avec un vecteur modèle . Plus un profil d'expression est similaire à un vecteur modèle, plus le produit scalaire correspondant sera élevé. A l'inverse, un profil d'expression très différent d'un profil de vecteur modèle donnera un produit scalaire proche de zéro. Nous isolons donc les gènes ayant un produit scalaire élevé pour chaque vecteur de base dans chaque couple tissu/organisme. Ces valeurs de projections ont été utilisées pour le filtrage de gènes dont l'expression ne varie pas suffisamment dans les différents tissus. Après filtrage, ce sont ces valeurs d'expression qui ont été présentées au réseau de neurones. Il serait possible d'étendre cette technologie de projection vectorielle pour extraire les gènes dont le profil d'expression présente un pic dans 2 ou plusieurs populations, par exemple pour sortir les gènes exprimées durant la totipotence, puis sous exprimées durant certain stages de différenciation preliminaires et exprimés de nouveau durant les stages finaux de différenciation.

2.2.4 Valeurs manquantes et organisation des données

Un aspect critique des données analysés dans ce chapitre - et dans toute approche d'intégration de données - est le problème des valeurs manquantes. Les jeux de données à intégrer ont été générés sur des plateformes distinctes (fournisseur ou génération de puces différentes) dont le recouvrement génomique est très différent. Certains gènes n'ont donc pas été profilés sur certains tissus/organismes, où le profil de certain gènes n'est pas complet. C'est une question qui doit être soigneusement traitée car différents algorithmes peuvent donner des résultats très différents. Dans cette analyse, nous calculons la projection vectorielle quand au moins deux valeurs de profil de gènes sont disponibles. les valeurs manquantes sont ignorées et le profil du vecteur est renormalisé à 1 sur les valeurs présentes, ce qui permet d'obtenir un produit scalaire d'amplitude correcte.

Après projection, le jeu de données complet est organisé par tissus/organismes et valeur de projections par étape de différenciation. Ces valeurs sont ensuite soumises au réseau de neurones.

2. Bioconductor Project : http://www.bioconductor.org

2.2.5 Architecture du système et paramètres

Le classifieur utilisé est une extension de la mémoire associative à simple couche 2.2 [71]. Elle est caractérisée par une simple couche neuronale contenant les poids d'entrée à apprendre à partir des données. Cinq neurones au total ont été chacun associés à chacune des étapes de différenciation. La sortie de chaque neurone est le produit vectoriel du vecteur constitué des valeur de projection de chaque gène avec le vecteur des poids.

$$y_i = \sum_{n=1}^{N} w_{i,n}.p_n, \tag{2.1}$$

y_i étant la sortie du neurone i, n l'index des gènes, N le nombre total de gènes utilisés pour la classification, $w_{i,n}$ la n-ième valeur du poids w du neurone i, et p_n la n-ième valeur de la projection à classer. A chaque étape de validation croisée, un tissu est sélectionné dans le pool pour test et les autres utilisés pour apprentissage. un modèle de réseau de neurone est créé (les poids sont initialisés) et le jeu d'apprentissage est présenté au réseau pour 200 époques. Pour chaque époque, l'ensemble des tissus du jeu d'apprentissage est présenté dans un ordre aléatoire au réseau, et les poids sont mis à jours avec une règle classique de type descente de gradient :

$$\Delta = a(k)[y_n - yd_n], \tag{2.2}$$

w_n étant la valeur courante du poids, $a(k)$ le taux d'apprentissage, y_n la sortie, et yd_n la sortie désirée. $a(n)$ est le coefficient d'apprentissage, et la formule $a(k) = 1/(k+1)$, k étant l'indexe d'époque, a été utilisée ici, ce qui représente un taux d'apprentissage décroissant lors de la progression. Chaque modèle de réseau neuronal est conservé et utilisé lors de l'étape final de classification d'échantillons inconnus. Le nombre de poids conservé ensuite est ajusté pour l'obtention d'un taux d'erreur minimum en validation croisée.

Nous avons ensuite fait une validation croisée et estimé le nombre de gènes optimal pour la classification des tissus par réduction progressive de leur nombre et mesure des statistiques de classification. La somme des échantillons mal classés en fonction du nombre de gènes a été représentées figure 2.3. Il est minimal pour 63 gènes. (16 gènes par neurone).

2.2.6 Algorithme de classification

Pour un tissu donné, que nous souhaitons classer dans une des cinq catégories pré-définies, la classification a été faite de la façon suivante : Le tissu testé doit être profilé en respectant plusieurs conditions. Au moins deux échantillons doivent être profilés sur puce, l'un provenant de cellules non différenciées (échantillon à caractériser par le système), et l'autre de cellules différenciées (connues *a priori*). Le système caractérise les échantillons inconnus en les assignant à une des catégories associées à un potentiel de différenciation pour chacun des modèles ANN générés durant l'apprentissage. Ensuite, une classe est attribuée par un vote à la majorité sur les 31 modèles ANN. Pour comprendre les processus biologiques activés durant les différentes étapes de différenciation, nous avons examiné les gènes discriminant pour chaque catégorie et mesuré leur enrichissement fonctionnel avec ClutrFree ([21], [19]).

2.3 Classification

La figure 2.3 expose les performances du classifieur. En (a) est représentée un graphique de performance globale de classification en fonction du nombre de gènes. L'erreur quadratique de

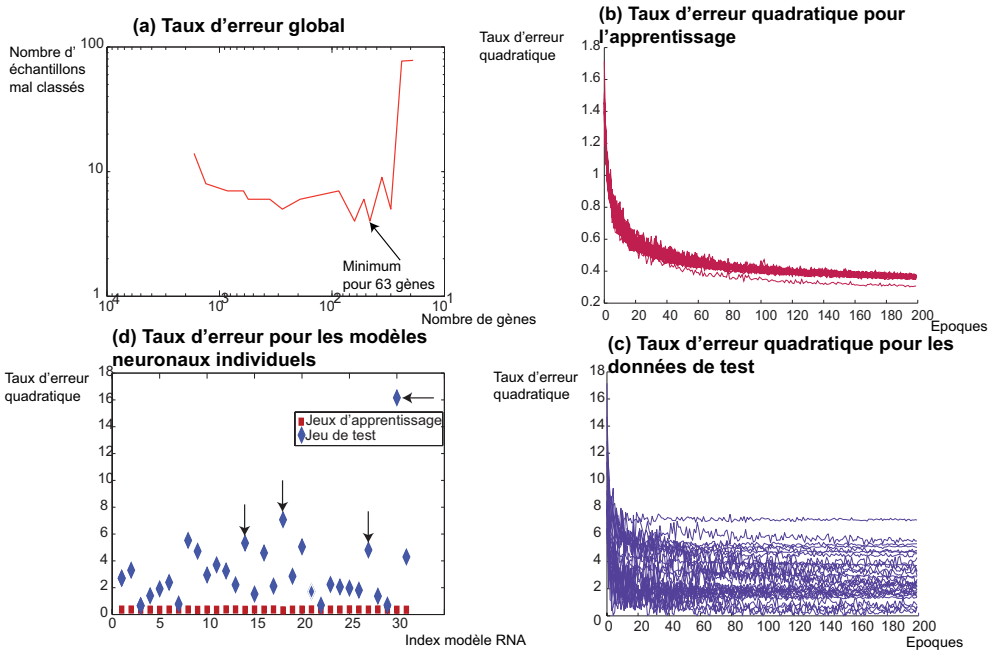


FIGURE 2.3 – Mesure d'erreur quadratique et de taux d'erreur sur les modèles de réseaux neuronaux individuels et sur le classifieur obtenur par vote majoritaire.

classification sur les jeux d'apprentissage et de test sont représentés respectivement en (b) et (c). Le taux d'erreur de chaque modèle ANN avec apprentissage sur 63 gènes est représenté en (d) pour le jeu d'apprentissage et le jeu de test pour chaque étape de validation croisée. Les tissus mal classés sont identifiés par une flèche. Le taux d'erreur minimal a été obtenu pour 63 gènes. Ce jeu de 63 gènes représente une signature de différenciation globale des cellules souches totipotentes aux cellules complètement différenciées à la fois dans *H. sapiens* et *M. musculus*. La progression des courbes d'erreurs montre une décroissance rapide et ne révèle pas de surapprentissage suspect, bien que nous ayons poursuivi l'apprentissage sur 200 époques. Seul un échantillon provenant de tissus neuronaux (cellules souches neuronales échantillons 2 - non représenté ici) n'a pas pu être classé correctement.

Nous avons ensuite utilisé le réseau pour classifier deux échantillons contenant potentiellement des progéniteurs (un échantillon humain de cellules souches de prostate [128] et un échantillon souris de cellules souches épithéliales de cavités dans l'estomac [106]). Les modèles de réseaux neuronaux nous ont donc non seulement permis de caractériser les tissus inconnus mais également d'identifier les gènes impliqués dans les processus de différenciation. Ces gènes peuvent être triés par ordre d'importance sur la base de leur poids - un poids supérieur étant synonyme d'implication supérieure de ce gène dans la classification. 5 jeux de gènes, correspondants aux

Tissu à tester	Résultat du vote majoritaire.
Cellules progénitrices dans l'estomac (Souris)	Cellules souches progénitrices
Cellules progénitrices dans la prostate (Humain)	Cellules souches multipotentes

TABLE 2.3 – Résultat de classification par vote majoritaire sur les 31 modèles de réseaux neuronaux sur les jeux de données indépendants avec 63 gènes.

5 étapes de différenciation ont été établis. L'analyse des gènes marqueurs a été faite sur le premier modèle donné par la validation croisée. Les 10 gènes ayant le poids le plus élevé dans la classification sont représentés Table 2.4.

Cell. totipotentes	Cell. multipotentes	Cell. progenitrices	Cell. progénitrices associées à une lignée	Cellules différenciées
Dbn1	Procr	Letmd1	Coq3	Aqp1
1110001A23Rik	Gprasp2	Lrp8	Mgst1	Rhced
BC053917	AI661017	Kpna3	PDZK8	Rhbdl4
5830405N20Rik	Ctso	Ptprm	Cybb	MGI :1933403
Iqwd1	Gkap1	Rbp4	Rbbp9	Wnt11
5730420B22Rik	Lrrc16	Ass1	Cst7	Eif2ak3
Rbp4	Nrbp2	Itgb2	4932441K18	Nfe2
Rpp40	Adam8	Med6	Fli1	Fech
Nfe2	Irak1bp1	Cd109	Anxa1	AI661017
Rbbp9	MGI :1916782	5830405N20Rik	Gpr124	Gzma

TABLE 2.4 – Top 10 gènes associés aux 5 étapes de différenciation, triés par poids.

Rbbp9, gène connu pour jouer un rôle dans la différenciation et la prolifération cellulaire, a été identifié par notre modèle caractérisant la population totipotente. Dans la population multipotente, le gène *Hopx* a été découvert. - exprimé dans le myocardium embryonnaire mais pas dans l'endocardium ou dans les vaisseaux larges. *Hopx* est également exprimé pendant le développement cardiaque et dépend de la protéine *Nkx2.5*. L'inactivation de *Hopx* dans la souris par recombinaison homologue conduit à un phénotype mortel partiellement pénétrant avec des défaut de développement cardiaques incluant le myocardium.

Dans des tissus contenant des cellules souches de type progéniteurs, nous avons trouvé plusieurs marqueurs intéressants. *Letmd1* est une protéine ayant un rôle dans le développement du cancer du sein à travers la régulation négative de *P53*. La protéine encodée par un autre marqueur identifié, *Ptprm* est une tyrosine phosphatase qui est connue pour faire partie de molécules de signalisation régulant plusieurs types de molécules, régulant la croissance cellulaire, la différenciation, le cycle de la mitose, et des transformations favorisant l'apparition des cancers. Le marqueur de surface *CD109* a également été identifié. C'est une protéine connue pour être exprimée dans les cellules souches hématopoïétiques et progéniteurs [116].

D'autres marqueurs ont également été découvert dans deux autres catégories. Dans les progéniteurs de lignées, *Fiat* est exprimé. C'est un régulateur transcriptionnel régulant la différenciation des ostéoblastes. *Anxa3* encode un membre de la famille des annexines. Cette famille est impliqué dans la croissance cellulaire, et cette protéine particulière pourrait être impliquée dans l'anti-coagulation. *Sfrp4* est un inhibiteur le la voie *Wnt* et joue un rôle central dans les décisions de l'engagement sur une voie de différenciation cellulaire. Nous avons également identifié *CD34*, le marqueur connu de cellules souches hématopoïétiques utilisé comme marqueur de purification.

L'exploration des gènes résultant de l'analyse a également permis d'identifier *Socs2*, gènes connu pour être exprimé de façon ubiquitaire dans un grand nombre de tissus dans la souris. Il a été montré qu'elle jouait un rôle dans le développement neuronal à travers la différenciation des cellules souches [173]. Il a été également prouvé qu'elle jouait un rôle d'inhibiteur de la prolifération des cellules souches épithéliales de l'intestin. Enfin, elle a été également reconnue

pour jouer un rôle dans le développement des glandes mammaires et pour réguler l'engagement de la voie de différenciation des cellules souches mésenchymateuses [129], qui incluent précisément les cellules à l'étude ici. Le phénotype de *Socs2* inclut l'élargissement des os et des muscles squelettiques. La liste complète des marqueurs est disponible sur le site du projet (`http://scann.sourceforge.net`).

2.4 Discussions

L'approche présentée dans ce chapitre et publiée au niveau théorique [18] mais aussi pour sa mise en œuvre méthodologique [22] est nouvelle de par la technologie d'intégration utilisée. Nous avons intégré des données issues de puces à ADN hétérogènes en utilisant une technologie de projection vectorielle associée à un réseau de neurones pour caractériser les propriétés de différenciation de cellules souches de types inconnus *a priori* et identifier une signature de différenciation. Il apparaît que la différenciation utilise des mécanismes de recrutement à grande échelle et résulte de changements d'expression subtiles sur l'ensemble du génome. Cependant, il reste un travail important de validation biologique des marqueurs trouvés. Nous avons été capables de classifier des tissus inconnus (progéniteurs de l'estomac dans la souris et cellules progénitrices de la prostate) avec une signature de 63 gènes de différenciation. Une extension naturelle de cette étude serait d'inclure d'autres échantillons disponibles dans les dépôts publiques tels que GEO ou ArrayExpress. Sur le plan technologique, il pourrait être possible d'améliorer les performances du classifieur par des techniques de boosting. Il serait intéressant de le rendre accessible à travers un serveur Web publique. D'autres types de classifieurs pourraient être également considérés.

2.5 Conclusion du chapitre

Cette étude montre plusieurs limitations, inhérentes à notre approche et des résultats en découlant. Ces limitations pourrait nous permettre d'envisager des améliorations pour des développements futurs. D'une part, le nombre d'échantillons étudié est trop faible pour en généraliser les résultats de façon fiable, surtout que la diversité des échantillons est extrêmement large. On a donc une première limitation inhérente à la nature des données. Les limitations inhérentes aux données mesurées par microarrays sont également connues [65].

Sur la question posée elle-même : Il existe une controverse sur l'existence d'une signature de différenciation universelle, au vu du peu de recouvrement entre les signatures publiées. Chaque étape de différenciation fait appel à des mécanismes distincts qui ne sont pas généralisables au niveau des gènes mais plutôt à un niveau système [185]. Enfin, sur la méthodologie et au vu des résultats obtenus, plusieurs limitations apparaissent. La question de la capacité réelle des données à généraliser est posée. Le nombre de gènes (63) est à rapprocher du nombre d'échantillons du jeu de données (89). J'étais resté dans un modèle linéaire pour éviter au maximum le surapprentissage qui aurait pu apparaître par l'utilisation de réseaux de neurones plus complexes.

Ceci dit, ce travail a permis la mise en place d'un protocole pour l'intégration de données hétérogènes, et notamment en ce qui concerne leur annotation. Cela a servi de base de départ pour l'intégration de plusieurs jeux de données sur l'analyse d'expression pour les analyses présentées dans les chapitres suivants.

3

Découverte de modules dans l'interactome humain liés aux mécanismes de métastase

Sommaire

Ce chapitre présente les données générées dans les papiers publié lors de la thèse de Maxime Garcia. [51] et [50]. Ce travail a été initialement développé durant son stage de Master et également publié [53] sous ma supervision. Il concerne le développement d'un compendium de données

d'expression Cancer du Sein et la création d'un algorithme d'intégration de données d'interactions - données d'expression appelé Interactome-Transcriptome Integration. Cet algorithme a été développé juste après mon arrivée au Centre de Recherche en Cancérologie de Marseille et la mise en place de l'équipe Cibi avec Maxime Garcia qui a effectué son stage de Master M2 (Master Biochimie, Biologie Structurage et Génomique de l'Université Aix-Marseille) ainsi que sa thèse de doctorat dans mon laboratoire.

3.1 Introduction

L'introduction des nouvelles technologies post-génomiques qui ont suivi les efforts de séquençage du génome humain nous apportent la possibilité de décoder l'origine génomique des maladies telles que le cancer. Dans cette optique, l'analyse de l'expression des gènes par puces à ADN nous a permis d'améliorer la classification et la pronostication de plusieurs types de cancer, notamment les cancer du sein [154] [175]. Cette approche a également permis la prédiction de la rechute métastatique ainsi que son issue sur plusieurs jeux de tumeurs. En cancérologie du sein, les marqueurs histologiques couramment utilisés sont limités et ne permettent pas la prédiction de ce type de récidive. La conséquence est que de nombreux patients (entre 70 et 80%) reçoivent une chimiothérapie adjuvante qui n'est pas nécessaire. Les outils génomique de diagnostic, tels que celui présenté dans ce chapitre donnent l'opportunité d'affiner le pronostique tout en améliorant les stratégies de traitement pour le cancer du sein.

Plusieurs études ont produit des signatures liées à la rechute métastatique distante [157]. La signature à 70 gènes Mammaprint [175] a permis de classer des patients sur leur pronostique de rechute. Wang et collègues [182] ont reporté une signature à 76 gènes spécifique aux récepteurs aux œstrogènes (60 gènes pour classifier les patients ER+ et 16 gènes pour classifier les patients ER-). Ces deux signatures ont seulement 3 gènes en commun, ce qui a donné des doutes sur leur capacité à généraliser les résultats. Michiels et collègues [104] ont réanalysé les données produites par van De Vijver et ont conclu que les signatures obtenues dans ce type d'études sont instables et dépendent du jeu de données d'apprentissage. Il y a donc un problème de généralisation de ces signatures. D'un point de vue purement mathématique, tout classifieur fonctionne bien à partir du moment où il le démontre par ses performances. Par contre d'un point de vue purement scientifique ou clinique, le contenu et la stabilité des signatures sont d'une importance cruciale, puisqu'il est nécessaire de comprendre leur contenu génétique qui peut éventuellement nous diriger vers de nouvelle cibles thérapeutiques.

Il existe plusieurs raisons pour l'instabilité inhérente des signatures génétiques, plusieurs d'entre elles étant couramment invoquées [47] [17] : Le bruit induit par les variations expérimentales, les variations dues à l'échantillonnage des patients et le biais de la plateforme utilisée (type de puce ou technologie, oligonucléotide ou nylon, fournisseur, et plus récemment séquençage). Hors, l'instabilité des signatures s'explique par d'autres raisons [46] plus complexes.

D'une part les mesures faites à l'échelle génomique souffrent du fléau de la dimension [105]. Le fléau de la dimension est un problème connu des statisticiens dû à la structure inhérente des mesures faites en génomique de façon générale et sur les puces à ADN de façon plus spécifique : Typiquement, on analyse beaucoup trop peu d'échantillons au regard du nombre de variables mesurées. C'est également un problème sur les technologies plus récentes d'analyse d'expression (type RNA-Seq) qui n'en sont pas exemptes [188].

D'autre part, la nature biologique de l'expression des gènes est en cause. Les puces à ADN permettent de mesurer l'effet ou le lien d'un phénotype sur les changements de niveau de transcription des ARN messagers. Cependant, les gènes ne sont pas indépendants et travaillent de concert

à travers un réseau d'interaction protéines-protéines. Notre hypothèse est que les phénotypes tels que les maladies oncogéniques résultent de faibles et subtiles perturbation dans l'expression (par accumulation de charge mutationelle dans le génome). Cette charge mutationelle va provoquer de faibles changements dans les gènes drivers (en amont du réseau d'interaction gènes-gènes) qui vont eux-même provoquer de larges changements dans les gènes en aval du réseau d'interactions [35]. Superposer un réseau d'interactions sur les données d'expression des gènes permet donc de détecter les gènes drivers porteurs qui ne sont différentiellement exprimés que de façon plus subtile. Nous montrons que ces gènes, utilisés comme biomarqueurs, ont montré une plus grande robustesse dans la prédiction de la rechute métastatique dans les cancer du sein profilés sur des plateformes hétérogènes, comparés à des gènes détectés sans informations d'interactions.

Plusieurs approches d'analyse de type réseau ont été proposées pour les analyses transcriptome sur puces à ADN. Elles comprennent la génération de réseau dépendant des conditions expérimentales, sans l'utilisation d'information a priori sur les réseaux, ce qui limite leur intérêt biologique [55]. Le co-clustering de données d'expression et de graphes a également été proposé avec la construction d'une distance basée à la fois sur l'expression des gènes et les réseaux d'interaction [66]. Les machines à vecteurs de support (Support Vector Machines, SVM) en combinaisons avec une approche de débruitage par décomposition spectrale ont également été proposées pour l'analyse de la réponse transcriptionnelle dans la levure [140]. Une approche réseau a été proposée pour la détection de gènes différentiellement exprimés dans un réseau d'interactions par Chuang et collègues [35]. Par l'utilisation d'un environnement statistique strict, une variantr de l'approche par machines à vecteur de support (SVM) pour l'utilisation directe des données d'interactions dans un classifieur a été proposée pour la classification de données de puces à ADN [195]. Dao et al ont proposé Optdis, un algorithme permettant d'identifier des sous-réseaux discriminant entre plusieurs classes d'échantillons, et ce, de façon optimale [41].

Ces méthodes ont permis de répondre aux problèmes biologiques décrits précédemment, c'est à dire la découverte des gènes drivers et leur identification par rapport aux gènes passagers. Par contre, elles ne résolvent pas le problème de la dimensionnalité, les analyses étant faites sur un nombre d'échantillons restreint.

Dans ce chapitre, je propose une extension multi-jeux de données de la méthode proposée initialement par Chuang et collègues [35], en y ajoutant l'intégration de plusieurs jeux de données avec une application à la recherche de gènes biomarqueurs signant la rechute métastatique dans le cancer du sein. Nous démontrons la pertinence de cette méthode, appelée Intégration Interactome-Transcriptome (ITI) sur un large compendium de données publiques, que nous avons construit et nommé Compendium Cancer du Sein. Pour éviter les biais possibles dans l'identification de sous réseaux, nous avons inclus une validation croisée à 10 niveaux et combiné les réseaux obtenus. Ensuite, nous avons validé les sous réseaux par la classification des patients dans deux jeux de données indépendant, van de Vijver et collègues [175] et Desmedt et collègues [44]. Par cette approche, nous avons augmenté la performance de classification de façon significative, comparées à trois signatures publiées précédemment, tout en abaissant la dépendance de ces signatures sur les données d'apprentissage. La classification sur ces deux jeux de données indépendant a donné 53% et 74% de précision sur les jeux de données van de Vijver et Desmedt.

Le fonctionnement de l'algorithme est détaillé dans la section suivante, avec la méthode de validation statistique. Les résultats de classification détaillés sont également reportés, ainsi que la validation biologique des sous réseaux identifiés.

3.2 Matériel et Méthodes

Pour détecter que des protéines interagissant sous forme de modules (sous-réseaux) répondent de façon coordonnées aux changements d'expression, nous avons superposé un réseau d'interactions protéines-protéines (Protein-Protein Interactions, PPI) sur un compendium de données d'expression dans le cancer du sein. La stratégie implémenté dans ITI consiste à détecter les sous-réseaux dont l'expression est significativement corrélée avec le critère clinique DMFS (*Distant metastasis Free Survival*) mentionné dans l'ensemble des jeux de données étudiés. Ensuite, ces réseaux sont validés par permutation aléatoire de l'expression et des données d'interaction. Pour faire l'apprentissage et tester le système, six jeux de données provenant des bases publiques ont été choisis suivant les critères décrit section 3.2.2. 4 analyses ont été faites (deux jeux d'analyses ont été sélectionnés pour la validation, et deux analyses séparées ont été faites pour les patients classifiés positifs pour les récepteurs aux œstrogènes (ER+) et négatifs pour ces récepteurs (ER-), le but étant de comprendre et de mesurer l'impact des données d'apprentissage sur les sous réseaux détectés et d'évaluer leur potentiel de généralisation. Pour chaque étude, une validation croisée a été appliquée en stratifiant proprement les données d'apprentissage et de tests. Le but de la stratification est d'équilibrer chaque jeu de données sur les patients ER+/ER- st le status DMFS pour garder des proportions équivalentes de chaque catégories lors de chaque itération de la validation croisée.

3.2.1 Mise en place de l'interactome humain

Pour la mise en place des données d'interaction, nous avons intégré plusieurs bases publiques. Parmi celles ci, nous trouvons la base Human Protein Resource Database (HPRD) Version 9 [86], Molecular Interaction Database (MINT) [32], INTAct [5], la Database of Interacting Proteins (DIP) [150], et l'interactome humain généré *in silico* par l'algorithme Cocite [138]. L'ensemble des données a été téléchargé en août 2010 (9/8/2010) et reformaté pour supprimer les auto-interactions, les interactions dupliquées et les protéines référencées sous "Unknown" (inconnues). Les auto-interactions ont été supprimées des fichiers car elles ne sont pas utilisées ou quantifiées par l'algorithme. Les différentes bases d'interactions ont été intégrées par similitude du numéro d'accession NCBI (Entrez Gene, National Center for Biotechnological Information). Les annotations ont été homogénéisée pour l'ensemble des bases pour qu'elles puissent être exploitées correctement par le système. L'agrégation des bases a donné un interactome final de 70530 interactions pour 13202 protéines.

3.2.2 Compendium de données d'expression dans le cancer du sein

Les jeux de données publics (Voir Table 3.1) qui ont permis la construction de notre Compendium Cancer du Sein ont été sélectionnés sur les critères suivants : Cancer du sein apparu tôt, disponibilité des informations relatées à la métastase, (information liée à l'événement et délai entre le diagnostique du cancer et la rechute ou le dernier suivi), statut des récepteurs aux œstrogènes déterminé par immunohistochimie (statut ER+/ER-), et absence de chimiothérapie post adjuvante. De plus, les échantillons dont le suivi est de moins de 5 ans ont été censurés. Un total de 930 tumeurs a été retenu pour l'analyse à partir du pool initial de 1561 tumeurs sur 6 jeux de données. La taille de l'échantillon, les plateformes puces à ADN utilisées (fournisseur) sont détaillées Table 3.1.

Les données d'expression brutes (Fichiers CEL pour les platformes Affymetrix) ont été téléchargées à partir du dépôt GEO (Gene Expression Omnibus) du NCBI [9], si disponibles,

et normalisés par l'algorithme GCRMA (GC-Robust Microarray Average) sous Bioconductor (Package *gcrma*). Le jeu de données van de Vijver a été téléchargé en tant que matériel supplémentaire de la publication originale [175]. Les jeux de données ont ensuite subi une conversion des données de sondes vers les gènes en retenant les sondes ayant le profil d'expression caractérisé par la médiane maximum et la suppression des sondes marquées 'nx_at', comme décrit par Reyal [144]. Les données profilées sur les plateformes HG-U133A et HG-U133B ont été traitées comme provenant d'une plateforme virtuelle résultant de l'intégration des annotations correspondant à ces deux plateformes. Il n'a pas été nécessaire de faire d'étapes de normalisation supplémentaires car seule la corrélation de Pearson - Statut DMFS a été utilisée dans la suite.

3.2.3 Stratification des données d'apprentissage et validation croisée à 10 niveaux

Pour identifier les sous-réseaux différentiellement exprimés tout en évitant le sur-apprentissage, un système de validation croisée a été mise en place en construisant différent jeux de données d'apprentissage et de test tout en tenant compte du statut clinique de la tumeur. La stratification mise en place a été concue pour équilibrer le statut ER+/ER- ainsi que la proportion de patients ayant subi une rechute métastatique dans les jeux de données et d'apprentissage. La préservation de la proportion des statuts moléculaires et cliniques dans ces jeux de données a permis d'accroitre l'homogénéité des jeux de test et d'apprentissage ainsi que d'éviter les biais moléculaires. Pour chaque couple de données d'apprentissage et de test, les sous réseaux ont été identifiés avec l'algorithme ITI (section 3.2.4) et validés par permutations aléatoire des interactions protéines-protéines et des profils d'expression (section 3.2.5). donnant 5 listes de sous réseaux, chacune correspondant à un jeu de données. Ces 5 listes sont ensuite combinées en une signature unique, dont la puissance de classification est évaluée sur le jeu de données de test indépendant.

3.2.4 Algorithme Integration-Transcriptome-Interactome - construction des sous réseaux

Les sous réseaux dont l'expression moyenne est liée à la rechute métastatique sont identifiés dans les jeux de données d'apprentissage avec l'algorithme ITI, défini ici. Cet algorithme est dérivé de l'algorithme publié par Chuang et collègues [35] avec la fonctionnalité supplémentaire de travailler dans un compendium de données (Figure 3.1). ITI a été implémenté sous la forme d'un pipeline Perl / Bash (en open source sous licence CeCILL). La validation statistique a été implémentée sous Matlab Statistical Toolbox R2010b (The Mathworks ©Natick, MA USA). La détection des sous réseaux a été parallélisée et implémentée sur un cluster Beowolf pour réduire le temps d'exécution, au vu du coût calculatoire du parcours d'une telle carte d'interactions. La visualisation des sous-réseaux a été obtenue avec le package GraphViz (AT&T Research, USA). Pour détecter les réseaux discriminants, la corrélation entre les profils d'expression et le statut DMFS a été calculé pour chaque jeu de données. Ensuite, l'interactome humain a été intégralement parcouru à la recherche de régions discriminantes (différentiellement exprimées entre les deux conditions cliniques - voir Figure 3.1), en considérant chaque nœud comme une graine (centre de réseau) potentiel et en agrégeant les voisins de façon récursive sur la base du score. Les nœuds voisins sont ajoutés au réseau s'ils permettent l'amélioration du score défini

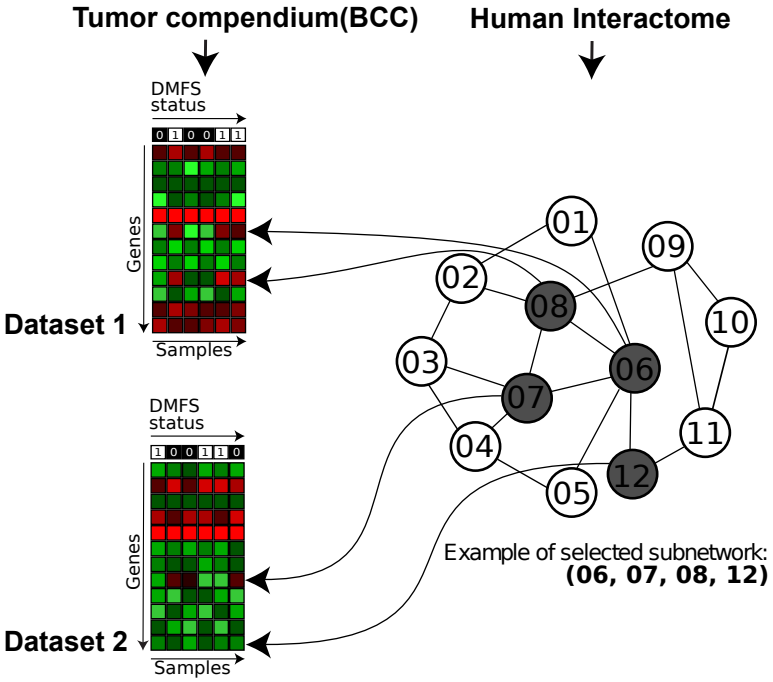


FIGURE 3.1 – Algorithme d'Intégration Transcriptome-Interactome (ITI). Deux types de données sont fournis à l'algorithme, les 5 jeux de données d'expression d'apprentissage avec un vecteur binaire d'annotations cliniques (ici, la rechute avec métastases) et un jeu d'interaction PPI dans humain, qui tient lieu d'interactome. L'expression est analysée simultanément sur les 5 jeux de données pour la constitution de sous réseaux d'interactions discriminant, c'est à dire permettant de séparer les patients suivant le critère clinique retenu - rechute métastatique.

(3.1).

$$S_{s,d} = \frac{\sqrt{n_d}}{\sqrt{\max n_d(DS)}} \left| \text{Corr} \frac{1}{n} \sum_{g \in S} e(g,d), cc(d) \right| \tag{3.1}$$

$$S_s = \frac{1}{NS} \sum_{d \in DS} S_{s,d} \tag{3.2}$$

$S_{s,d}$ est le score du sous réseau s calculé sur le jeu de données d du Compendium Cancer du Sein contenant NS jeux de données DS. Corr est la corrélation au sens de Pearson mesurée entre l'expression moyenne $e(g,d)$ du gène g sur le jeu de données d pour les gènes appartenant au sous réseau s avec le vecteur de condition clinique cc, pondéré par la racine carrée du nombre d'échantillons dans le jeu de données n_d divisé par le nombre maximum d'échantillons dans l'ensemble des jeux de données dans DS. S_s est le score global du sous réseau s calculé par moyennage des sous réseaux $S_{s,d}$.

3.2.5 Validation statistique et filtrage des sous-réseaux

Pour valider statistiquement les sous réseaux obtenus, nous avons comparé leurs scores à une distribution aléatoire de sous-réseaux pour faire un test d'hypothèse. La première distribution aléatoire permet de mesurer la significativité de l'algorithme d'extraction des sous réseaux. Elle

a été obtenue par une sélection aléatoire de sous réseaux dans l'interactome humain, c'est à dire considérer si un sous réseaux est discriminant ou non sans examen de son expression. La seconde distribution a mesuré si le lien biologique entre les interactions protéines-protéines et l'expression des gènes est statistiquement valide. Il a été obtenu par permutation aléatoire des labels des échantillons. Pour garder les sous réseaux aléatoires comparables aux sous-réseaux identifiés, la distribution de leur taille a été forcée à l'identique à celle des sous réseaux réels. Ensuite les distributions aléatoires ont été modélisées par mixture de gaussiennes. Une fois obtenues, ces distributions modélisée ont permis de fixer des seuils sur les scores de sous-réseaux de façon indépendante sur chaque jeu de données à un niveau significatif de p-valeur, et donc de valider les sous réseaux sur les deux types de distributions. Un troisième type de distribution aléatoire a été généré par permutations des interactions dans l'interactome humain tout en maintenant sa topologie (le degré des gènes a été conservé). Cependant, il n'a pas été utilisé dans l'étude car aucun sous-réseau identifié n'a donné de score apparaissant significatif par rapport à un test d'hypothèse sur cette distribution, confirmant le lien extrêmement fort entre expression des gènes et interactions physiques protéines-protéines. Au final, nous avons filtré les sous réseaux identifiés avec un score plus élevé que ce que l'on pouvait obtenir par chance sur les sous-réseaux aléatoires($p < 1.10^{-4}$ sur deux jeux de données au moins) et la permutation sur les labels ($p < 1.10^{-4}$ sur deux jeux de données au moins).

3.2.6 Construction d'une signature sous-réseaux commune

Par l'emploi de ce filtre, 10 jeux de sous-réseaux ont été obtenus pour chaque étape de validation croisée. Ensuite, ces jeux ont été combinés pour l'obtention d'une signature finale. Pour ce faire, nous avons examiné les sous-réseaux deux à deux et les avons combinés s'ils avaient plus de 50% de gènes commun dans un sens ou l'autre (A vers B ou B vers A). Par cette méthode, des clusters de sous réseaux ont été crées. La liste de sous-réseaux finale a été construite en ne conservant que les sous-réseaux apparaissant au moins 2 fois. Pour un cluster donné, seul le sous-réseau ayant le score le plus élevé a été conservé. Les jeux de sous-réseaux finaux sont listés Table 3.2.

3.2.7 Classification des tumeurs et prédiction de la rechute métastatique sur les type ER+ et ER- sur deux jeux de données indépendants

La liste des sous-réseaux obtenue section 3.2.6 a ensuite été utilisée pour la prédiction de la rechute avec métastase sur l'étude 1 (Jeux de données Desmedt utilisé comme jeu de données indépendant) et l'étude 2 (Jeux de données van de Vijver utilisé comme jeu de données indépendant). Sur chaque étude l'apprentissage a été fait indépendamment sur l'ensemble des jeux de données moins le jeu de test, donnant 5 modèles SVM (Support Vector Machine). La classification finale en sortie des 5 machines SVM est faite par vote à la majorité.

L'organisation complète est montrée figure 3.2.

Pour utiliser les sous-réseaux comme variable d'entrée SVM, l'expression de chaque sous réseaux a été obtenue par moyennage de l'expression sur les gènes et utilisée comme profil discriminant à la fois pour l'apprentissage et le test. Plusieurs modèles SVM ont été testés avec un nombre croissant de sous réseau et la liste finale des sous-réseaux retenue est celle qui maximise la précision de classification. Les résultats de classification (précision, sélectivité et sensibilité) ont été reportés Table 3.3 avec comparaison sur les classifieurs existants.

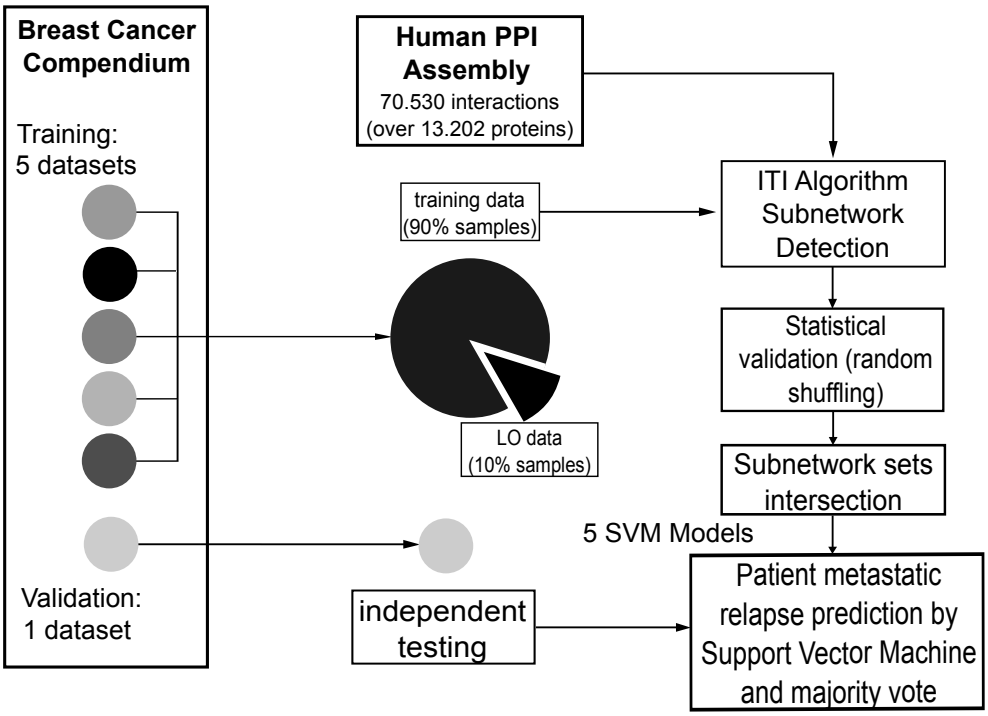


FIGURE 3.2 – Organisation de la validation croisée. L'interactome est assemblé à partir de plusieurs sources. Les jeux de données d'expression formant le Compendium Cancer du Sein sont agrégés pour former le jeu de données d'apprentissage. Ensuite, 10 groupes de patients ont été formés sur la base d'une classification à 10% d'omission. Les sous-réseaux ont été détectés sur l'interactome humain sur chaque jeu d'apprentissage avec ITI et validés statistiquement deux fois par permutation aléatoire des sous-réseaux et de l'expression, comme décrit section 3.2.5. Les sous-réseaux retenus ont été combinés pour l'apprentissage d'une machine à vecteur de support (SVP), par jeu de données, soit 5 machines. Le jeu de données final a été ensuite combiné comme jeu de marqueurs génétiques pour la classification sur des données indépendantes par vote à la majorité sur les 5 modèles SVM.

3.2.8 Ressource en ligne ITI - Enrichissement GO

Pour détecter les voies de signalisation associées avec la rechute métastatique, nous avons calculé l'enrichissement des processus biologiques GO par l'utilisation du programme ErmineJ [3]. Ce programme donne les enrichissement et les p-valeurs corrigées pour les tests multiples pour l'ensemble des termes GO. Ceux-ci ont été systématiquement calculés pour chacun des sous-réseaux pour les associer avec les processus moléculaires GO.

Les données obtenues ont été organisées dans une ressource en ligne [4]. cette ressource décrit les sous-réseaux et donne une description organisée des gènes inclus dans l'analyse présente. Des fichiers décrivant les sous-réseaux et les listes de gènes sont téléchargeables pour analyse. Les p-valeurs associées aux sous-réseaux calculées avec les distributions aléatoires (section 3.2.5) sont également incluses, avec les scores de Fisher [70]. Les gènes été annotées avec les numéro d'accession NCBI et des liens vers Entrez Gene et entre les sous-réseaux sont fournis. Pour comprendre les changements d'expression des gènes inclus dans les sous-réseaux, des graphes

3. http ://erminej.msl.ubc.ca/
4. http ://iti.sourceforge.net

colorés sont données, avec corrélation expression - DMFS superposée sur chaque sous-réseau. Le score de corrélation est donné séparément pour chaque jeu de données afin de comprendre l'hétérogénéité ou l'homogénéité des résultats.

3.3 Résultats

3.3.1 Établissement de deux jeux de sous réseaux discriminants pour les sous-types ER+ et ER- à partir d'un compendium de 930 tumeurs

Deux signatures indépendantes ont été générées pour les sous-types ER+ et ER- (Tumeurs présentant ou non les récepteurs aux œstrogènes) sur 2 études distinctes. Dans la première, les données Desmedt ont servi de test indépendant, dans la seconde, les données Van de Vijver ont servi de test indépendant. On a donc obtenu quatre jeux de sous-réseaux qui constituent nos signatures pronostiques (Table 3.2).

La taille optimale pour les signatures retenue est celle qui maximise la précision moyenne sur les 10 jeux de données pour chaque analyse. Pour l'étude 1, les sous-réseaux discriminants ont une corrélation de 0,49 (tumeurs ER+) et 0.54 (tumeurs ER-) confirmant la corrélation élevée entre expression et proximité dans l'interactome pour les sous-réseaux détectés. La taille de signature est respectivement de 6 sous réseaux (ER-) et de 165 sous -réseaux (ER-). Pour l'étude 2, la signature ER+ a donné une classification optimale pour 14 sous-réseaux et la signature ER- pour 122 sous-réseaux. Ils correspondent à une liste de 175 gènes (ER+), 2310 (ER-) pour l'étude 1, et de 272 gènes (ER+) et 1481 gènes (ER+) pour l'étude 2, respectivement, plusieurs gènes étant présents dans plusieurs sous-réseaux.

Ces chiffres sont plus larges que les signatures habituelles. Ceci suggère que nous avons identifié un large panel de gènes significativement lié à la rechute, et reflétant de façon réaliste l'empreinte biologique des métastases et l'échelle des perturbations qu'elles provoquent au niveau de l'expression. La redondance des gènes dans les sous-réseaux peut être expliquée par la forte connectivité de plusieurs gènes (Par exemple TP53), ce qui les rend vraisemblablement présents dans plusieurs sous-réseaux.

3.3.2 La classification par sous réseaux sur des données indépendantes montre la supériorité d'ITI par rapport à une classification sans interactome

Pour quantifier la performance des signatures réseaux construites avec ITI, nous les avons comparées avec plusieurs signatures précédemment établies, les 128 sondes du Genomic Grade Index (GGI) [157], les signatures à 70 gènes Mammaprint [175] et la signature à 76 gènes spécifique aux status ER dans le sein [182] ont été testées. Les performances de classification ont été testées sur les mêmes tumeurs que celles testées avec ITI (Jeux de données Desmedt et Van de Vijver) séparément sur les tumeurs ER+ et ER-. Les méthodes de classification originellement associées avec les signatures citées ont été respectivement appliquées. Pour la signature van de Vijver, les distances aux centroïdes moyens pour les tumeurs avec rechutes et les tumeurs sans rechutes sont calculées [175], permettant de déterminer à quel groupe appartient chaque patient. Pour la signature Wang, un score de rechute est calculé pour chaque patient par une combinaison linéaire de l'expression des gènes pondérées par les coefficients standards de Cox [182]. Puisque les signatures CGI et Mammaprint sont définies par un jeu de sondes microarrays, les analyses ont été effectuées avec les sondes présentes dans les jeux de données de test. Les résultats sur les performances de sorties sont détaillés table 3.3. Ils montrent que la performance de généralisation d'ITI est supérieure aux signatures publiés précédemment. La classification GGI a montré

Auteur(s)	No d'accession GEO	Plateforme	Nombre d'échantillons (Filtré/Non Filtré)	Statut DMFS (méta, non méta)	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	N/A	Agilent HumanGenome	150/295	56 /94	36/114
Wang et al., 2005	GSE2034	HG-U133A	276/286	107/169 72/204	
Total : 6 Jeux distincts	6 disponibles	4 plateformes	930/1561	307/623	246/684

TABLE 3.1 – Les jeux de données inclus dans le Compendium Cancer du Sein. Deux apprentissages différents ont été faits avec différentes combinaisons de test et d'entraînement (jeux de données notés en gras). Sur l'étude 1, le jeu de données Desmedt a été retenu comme jeu de test indépendant, tandis que sur l'étude 2, c'est le jeu de données van de Vijver qui a été retenu comme jeu de test.

Dataset	seuil de p-valeur	Nombre de jeux de données	Nombre de sous-réseaux	Nombre de gènes
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

TABLE 3.2 – Différents seuils de l'algorithme ITI utilisés pour chaque analyse (seuil de p-valeur et nombre de jeux de données) et taille des signatures obtenues (dont le nombre de sous réseaux et nombre de gènes).

Statut	ER-								ER+							
Jeu de données	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
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	0.475	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

TABLE 3.3 – Comparaison des performances de la classification pour ITI et d'autres signatures sur les jeux de données Desmedt et van de Vijver pour les tumeurs ER+ et ER-. Les 4 jeux de sous-réseaux ont été utilisés pour mesurer les performances de classification (SP=spécificité, SV=sensibilité, VP=Vrais Positifs, FP=Faux Positifs, VN=Vrais Négatifs, FN=Faux Négatifs, ACC=Précision). La classification par Support Vector Machine associée aux sous-réseaux est supérieure à la classification par signature simple de gènes sur le jeu de données Desmedt (Etude 1) et sensiblement identique pour le jeu de données van de Vijver (Etude 2).

la plus forte précision (sur un intervalle [47%-68%]), la signature Mammaprint sur un intervalle [41%-62%] et la signature à 76 gènes sur un intervalle [37%-63%]. ITI a montré une précision plus élevée que la signature Wang sur le jeu de données Desmedt en ER+ : Une précision de 74% (avec une spécificité de 92%) a été obtenue, contre une précision de 60% (spécificité de 56%) avec la signature Wang. ITI a également donnée une précision supérieure sur les tumeurs ER- Desmedt avec une précision de 54% (spécificité de 65%) contre une précision de 38% (spécificité de 41%) pour la signature Wang. Cela est également vrai pour la signature Mammaprint à 70 gènes qui fonctionne mieux pour les patients du jeu de données van de Vijver. ITI a montré une précision de 53% associée avec une spécificité de 90% sur les patients ER- van de Vijver et une précision de 52% avec une spécificité de 65% sur les patients ER+. Cette performance est inférieure à ce qui a été obtenu sur l'Etude 1 et reflète probablement un biais sur Affymetrix, cette plateforme étant majoritaire sur le Compendium Cancer du Sein. La signature Mammaprint a montré une performance largement inférieure sur le jeu de données Desmedt avec 41% de précision sur les tumeurs ER+ et de 42% sur les tumeurs ER-. De façon similaire, ITI a montré une performance supérieure sur la signature GGI pour les patients ER-. En tout, ITI a été capable de généraliser mieux avec une limite de performance inférieure à 52% sur nos jeux de tests. Sur une autre base de comparaison, Chuang et collègues [35] ont montré une précision de 41% sur le jeu de données van de Vijver en utilisant les données Wang pour l'apprentissage et 55,8% réciproquement. Il est difficile de mesurer la contribution spécifique des données d'interactions et des données d'expression puisqu'elles ne sont pas facilement séparables dans le déroulement d'ITI, sauf à écrire un nouvel algorithme. Ceci dit, Chuang et collègues ont précédemment démontré qu'une approche de type analyse réseau augmentait la robustesse de la signature et plusieurs études ont démontrée l'impact positif d'une intégration de plusieurs jeux de données sur les performances de classification [186, 48].

Nous avons fait une analyse de survie entre les groupes de bon et mauvais pronostiques sur les patients ER+ de l'étude 1 (Résultats présentés Figure 3.3. Le test de log-rang a donnée une p-valeur de $4,89.10^{-5}$, suggérant une excellente séparation entre les deux groupes. C'est plus élevé que les p-valeurs obtanues avec les autres signatures (Wang a donné une p-valeur de $4,11.10^{-3}$ et GGI a donné une signature de $1,34.10-5$. La signature Mammaprint n'a pas été capable de séparer les patients Desmedt sur des groupes significatifs. Même si ITI n'a pas été spécifiquement conçu pour optimiser la séparation sur la base de la survie, il a été capable de séparer les patients entre groupes de bon et mauvais pronostics. Une alternative aurait été de calculer les scores de sous réseaux directement sur les scores du log-rang des gènes et d'y associer une p-valeur.

3.3.3 Les signatures obtenues avec ITI montrent une stabilité supérieure sur différent jeux de données

Les signatures van de Vijver et Wang n'ont que 3 gènes en commun, ce qui représente moins de 5% de l'ensemble des gènes des signatures. Nous avons comparé les deux signatures obtenues avec ITI pour les tumeurs ER+ et ER- sur les deux études Desmedt et van de Vijver. Un total de 937 gènes communs a été comptabilisé entre les deux études pour les tumeurs ER- et de 46 gènes pour les ER+. Cela représente respectivement un recouvrement de 32,8% (ER-) et de 11,5% (ER+). Ces valeurs restent néanmoins relativement faibles, et reflètent les biais des plateformes d'apprentissage respectives. Ceci dit, cela reste largement supérieur aux 3% de gènes communs entre les signatures Wang et van de Vijver. La stabilité des signatures ITI pourrait être augmentée par l'utilisation d'un jeu de données d'apprentissage plus large.

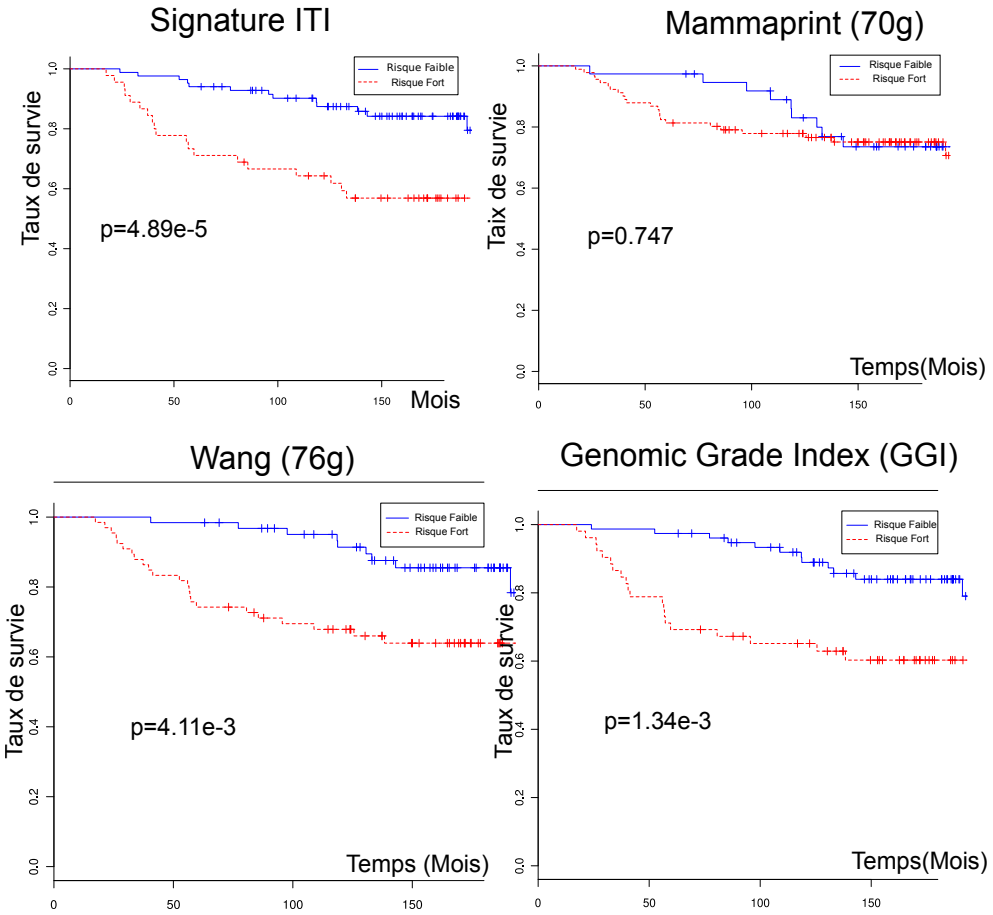


FIGURE 3.3 – Résultats de l'estimation du pronostique de survie par modèle Kaplan-Meyer pour les groupes de bon pronostique (non associés à la rechute métastatique) et de mauvais pronostique (associé à la rechute métastatique) définis par ITI, Mammaprint la signature Wang et le GGI pour les patients Desmedt ER+. ITI a donné la p-valeur la plus faible de $4,89.10^{-5}$ sur l'ensemble des tests de log-rang sur l'ensemble des signatures testées.

3.3.4 La biologie des sous-réseaux est plus facilement interprétable que celle des signatures classiques

Nous avons examiné l'enrichissement en annotations Gene Ontology (GO) [54] pour les sous réseaux obtenus sur l'étude 1 (Desmedt). La Table 3.4 montre plusieurs termes enrichis respectivement pour les signatures ER+ et ER-. Les fonctions enrichies dans dans les sous-réseaux discriminants sont liés aux processus de régulations perturbés dans le cancer du sein (cycle cellulaire, réparation de l'ADN) et particulièrement dans le cas de l'apparition de métastases (système immunitaire, prolifération cellulaire adhésion focale, migration cellulaire, organisation du cytosquelette) à la fois dans les tumeurs ER+ et ER-.

Comme exemple, nous avons pris un sous-réseau significativement associé à l'Étude 1 (ER-), le sous-réseau 6693, représenté Figure 3.4. Ce sous-réseau contient des gènes avec des fonctions connues pour être perturbées dans les tumeurs ER-, les cancer du sein en général et les tumeurs ayant produit des métastases, comme le suppresseur de tumeur TP53 ainsi que les récepteurs aux tyrosine kinases ERBB2 et EGFR.

Les sous-réseaux identifiés contiennent également plusieurs kinases de signalisation et de régu-

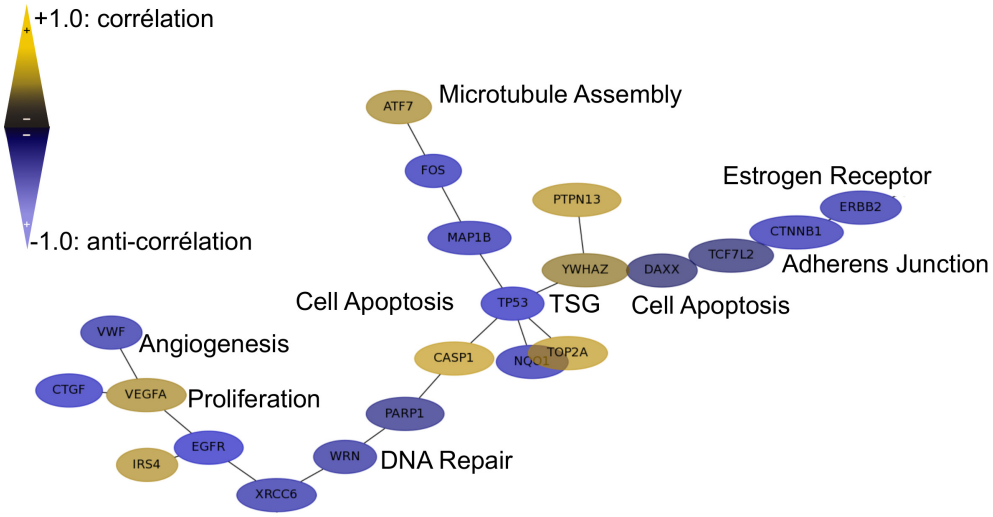


FIGURE 3.4 – Représentation graphique d'une partie du sous-réseau 6693 (Étude 1, ER-). Ce graphe représente un sous-réseau discriminant. Les valeurs de corrélation représentées sont celles du jeu de données Sabatier et collègues, qui fait partie du Compendium Cancer du Sein. Les nœuds et liens correspondent aux gènes codant une protéine et aux interactions protéines-protéines. Aux couleurs bleues et jaunes correspondent respectivement une valeur de sur-expression et de sous-expression du gène, parmi les patients ayant subi une rechute métastatique distante, comparé avec ceux qui n'en ont pas subi.

lation du cycle cellulaire (CDK2, CDKN1A, CDKN2A). NQO1, dont l'altération de l'expression a été associée avec plusieurs types de cancer [161] dont le sein[131] . PIN1 est présent dans les sous-réseau, et a été récemment identifié comme promouvant l'agressivité du cancer du sein[149]. Le récepteur aux insuline est également présent. Son expression de dérégulation corrèle avec la faible réponse aux thérapies anti IGF-FR dans les cancer du sein triple négatifs [98]. Les sous réseaux identifiés contiennent également plusieurs oncogènes connus ainsi que des gènes non connus pour être liés au cancer mais dont la dérégulation pourraient le favoriser.

3.4 Discussion

Nous avons conçu un algorithme d'analyse réseau, ITI (Intégration Transcriptome-Interactome) pour l'identification de signatures pronostiques généralisables sur plusieurs études, multiples et hétérogènes. Cet algorithme est organisé en deux étapes. D'une part, il permet l'intégration d'un Compendium de données Cancer du Sein, et d'autre part, il permet la découverte de sous-réseaux, c'est à dire de groupes de gènes interagissant, dont l'expression génique permet la discrimination de deux conditions d'intérêt. Les sous réseaux identifiés sont ensuite validés statistiquement par deux méthodes, par permutation de l'interactome humain d'une part, et par permutation de l'expression génique d'autre part.

Nous avons appliqué ITI sur la question fondamentale de la découverte de marqueurs signant la rechute métastatique distante dans le cancer du sein, domaine pour lequel un large corpus de données est disponible. Notre approche illustre la faisabilité d'intégrer un compendium de données d'expression (930 tumeurs ont été intégrées) et une carte d'intégration protéines-protéines. ITI représente donc un outil pour l'analyse des grands dépôts de données d'expression. Il inclut

Terme Gene Ontology	GO	p-valeur corrigée
ER-		
mRNA cleavage	GO :0006379	1 ,25E-08
regulation of growth hormone secretion	GO :0060123	2,18E-07
positive regulation of cytoskele-ton organization	GO :0051495	2,06E-04
regulation of insulin secretion	GO :0050796	1,55E-05
regulation of chemotaxis	GO :0050920	4,29E-07
ER+		
natural killer cell mediated immunity	GO :0002228	2,93E-06
positive regulation of MAP kinase activity	GO :0043406	4,76E-10
muscle cell development	GO :0055001	1,06E-11
interphase of mitotic cell cycle	GO :0051329	4,08E-11
Wnt receptor signalling path-way through beta-catenin	GO :0060070	6,22E-10

TABLE 3.4 – Enrichissement des annotations GO pour les sous-réseaux ER+ et ER-. Plusieurs termes enrichis pour les sous-réseaux extraits dans l'étude 1 (ER+ et ER-) sont liés au cancer.

l'intégration de données supplémentaires telles que les données d'interaction protéines-protéines (PPI), ainsi que des données cliniques.

Avec ITI, nous avons produit des signatures spécifiques au statut des récepteurs aux oestrogènes des tumeurs étudiées (Statut ER+ ou ER-) que nous avons validées sur deux jeux de données indépendants (Études Desmedt et van de Vijver). L'application de l'analyse ITI sur ces études a permis de montrer qu'ITI avait une précision plus élevée que que celle des classifieurs publiés précédemment (74% de précision pour Desmedt (ER-) et 53% pour van de Vijver (ER+)). Notre signature, basée sur une analyse réseau, reflète la grande empreinte biologique de la métastase et est par conséquent plus grande (en nombre de gènes) que les signatures publiées. Le classifieur obtenu avec ITI a montré une sensibilité moindre aux biais de plateformes microarrays que les signatures précédentes, puisque la performance est identique sur les deux jeux d'apprentissage. ITI a également montré une très forte spécificité, ce qui est critique pour éviter un sur-traitement systémique adjuvant inutile pour le patient.

A l'issue de ce travail, nous avons projeté de compléter l'algorithme ITI pour y inclure d'autres types de données, telles que des données de variation du nombre de copies dans l'ADN (Ce point est développé chapitre 4). Les données de mutation (SNPs) et de méthylation sont également envisagées pour la recherche de sous-réseaux possédant des taux de mutations plus élevés ou des profils de méthylation différents. La capacité d'ITI à contenir le problème du fléau de la dimension le rend utilisable à la détection de biomarkers données par les technologies de séquencages, qu'elles soient adaptées au génome (DNA-Seq) ou transcriptome (RNA-Seq). Dans les prochaines versions, le type d'interaction PPI sera pris en compte lors des étapes d'intégration de l'interactome et de l'agrégation des gènes composant les sous-réseaux. De plus, les performances de classification sont intimement liées à la nature moléculaire des échantillons à classer et un

sous-typage plus fin serait nécessaire pour rendre la technologie d'ITI utilisable en clinique. Un accroissement significatif des performances de classification pour le sous-type ER- a été observé par séparation des patients rechutant tôt et tard (données non montrées).

3.5 Remerciements

Cette recherche a été principalement financée par l'Institut National du Cancer, la Ligne Nationale contre le Cancer et l'Institut National de la Santé et de la Recherche Médicale. Le cluster de calcul utilisé pour cette étude a été financé par la Fondation Pour la Recherche Médicale. Maxima Garcia a été financé par une bourse de thèse Région Provence-Alpes Côte d'Azur - Institut de la Santé et de la Recherche Médicale. Un très grand merci à Sabrina Carpentier (Ipsogen, Marseille, France) pour son aide précieuse sur les aspects techniques et statistiques et Dr Francoise Birg et Wahiba Gherraby pour leur lecture du manuscrit.

Code et données sont mis à disposition sur le site web d'ITI : [5].

3.6 Conclusions du chapitre

Ce travail a été le centre de la thèse de Maxime Garcia. Il s'accorde pleinement à d'autres projets décrits dans cette thèse et publiés : la version non supervisée d'ITI ([53], le chapitre décrivant l'implémentation technique d'ITI 2.0 sur cluster Beowulf ([50], un chapitre sur l'extension de ce travail pour l'analyse de copies d'ADN (CGH - *Comparative Genomics Hybridization* [52]). Ce dernier travail est décrit dans le chapitre 4 de cette HDR. Il me semble que c'est une voie de recherche importante, qui représente la première étape de la rétro-ingénierie du vivant par l'utilisation de la biologie des systèmes. C'est par ce genre d'approche, qui inclut des connaissance biologiques directement dans l'analyse que nous pourrons tirer le meilleur parti des analyse génomiques par les technologies mise au point ces 10 dernières années. Au moment de l'écriture de ce chapitre, cela était vrai pour les technologies de puces, et c'est maintenant tout à fait valable pour les technologies de séquençage. Il n'y aura pas d'algorithme simple d'analyse sans l'interrogation des bases de connaissances existantes.

5. http://iti.sourceforge.net

4

Découverte de gènes drivers par analyse intégrée Interactome, transcriptome et génome

Sommaire

Ce chapitre reprend les données publiées référence [52]. C'est une extension du travail développé en 2012 par Raphaëlle Millat-Carus et Maxime Garcia, qui ont fait leur stage de M2

Bioinformatique, Biologie Structurale et Génomique de L'université Aix-Marseille (BBSG) sur la plateforme Cibi sous ma responsabilité, suivi d'une thèse de doctorat pour Maxime Garcia.

4.1 Introduction

4.1.1 Sous types moléculaires dans le cancer du sein

Le cancer du sein est une maladie hétérogène. Cela explique pourquoi les traitements de fonctionnement pas de manière égale sur tous les patients. Cette hétérogénéité est difficile à déchiffrer avec les critères histocliniques de la tumeur [176] qui sont utilisés pour prédire le pronostique. De fait, de nombreux patients subissent un sur-traitement [17] [16]. Le défit actuel est est de concevoir de nouveaux classifieurs pour i) séparer les sous types moléculaires et ii) prédire finement le pronostique associé à chacun de ces sous types. L'ère post-génomique a permis la mise en place de plusieurs technologies et outils pour gagner une plus grande connaissance de la nature moléculaire du cancer et notamment du cancer du sein. Parmi ces outils, la technologie des puces à ADN (microarrays) a représenté une avancée essentielle pour l'analyse des ARN messager et de l'ADN (par les puces de types array-Comparative Genomics Hybridization - aCGH) et la découverte de marqueurs pour affiner les classification moléculaires dans les cancers. Les puces à ADN ont permis l'identification de 4 sous types moléculaires dans le cancer du sein (luminal, basal, ERBB2 et normal)[132] , ce qui a été confirmé ensuite sur d'autres études [155] qui a permis de raffiner cette classification en divisant les luminaux en deux sous-types (luminaux A et luminaux B) [33]. Cette classification a évolué avec la découverte d'autres sous groupes pertinents [64] [40], [100], avec notamment le sous type des claudin-low [136]. Les études de corrélation entre la situation clinique et le profil moléculaire ont permis d'assigner des pronostiques spécifiques aux sous-types [156] [177]. Cependant, ces corrélations sont liées à une puissance statistique limitée de part le faible nombre d'échantillons utilisés. Ceci dit, la classification en 5 sous-types apparaît comme étant très robuste à travers différentes études et Hu et collègues ([72] ont validé une liste de 306 gènes qui font maintenant référence pour ces sous types. Les sous-types luminaux A et basal font partie donc des 5 sous-types majeurs, et ont des caractéristiques opposées à la fois au niveau clinique et au niveau génomique. Les cancers du sein luminaux A sont les plus fréquents (45%), ils sont à faible grade (c'est un indicateur clinique qui mesure l'agressivité du cancer), ce sont des tumeurs différenciées qui expriment les récepteurs hormonaux et les gènes de différenciation ESR1 et GATA3. Ils sont en général associés à un pronostique relativement favorable de par leur réponse à l'hormonothérapie. Le sous type basal représente 15% du nombre de cancer du sein, mais ils sont de grade élevé, ce sont des tumeurs prolifératives qui ne possèdent pas de récepteurs hormonaux et qui sont associées avec un pronostique défavorable. Bien que ces tumeurs soient relativement sensibles à la chimiothérapie, l'effet de ce type de traitement reste limité.

4.1.2 Intégration de données génomiques pour la découverte de gènes drivers dans les sous-types moléculaires luminal A et basal

Un problème fondamental pour la caractérisation systématique des cancers en général et du cancer du sein en particulier est la découverte de gènes drivers et de marqueurs. Bien qu'il ait été établi que le cancer du sein soit caractérisé par 5 sous-types moléculaires (Les 4 sous types principaux et la sous division des luminaux en A et B), les profils génomiques montrent que les tumeurs sont extrêmement diverses et que la plupart présentent des caractéristiques génomiques uniques. Il apparaît donc que les lésions génomiques et les anomalies de régulation et d'expression

qui sont détectées dans les tumeurs ne sont pas à l'origine de la maladie et que le challenge est de distinguer les gènes drivers de la tumeur (ceux qui provoquent le prolifération, la résistance au traitement et avec une implication primaire dans les processus de métastase) des gènes passagers (ceux dont les changements sont le résultat des dérégulations dans les gènes drivers et qui n'ont pas d'impact sur la maladie). Les drivers doivent être identifiés pour chaque sous-type [2]. De plus, la division en sous-types, bien que déjà très fournie, pourrait éventuellement être affinée encore ([40]).

Le cancer du sein apparaît comme le résultat d'une expansion de tumeurs conduite par des cellules souches tumorales qui acquièrent l'immortalité et un avantage de survie sur les cellules dites normales à travers des mutations dans leur ADN et des changements d'expression. Les drivers sont les gènes qui apportent spécifiquement cet avantage de sélection qui augmentant l'empreinte biologique des processus liés au cancer et notamment la prolifération ainsi que la résistance à la chimiothérapie. D'un autre coté, les gènes passagers, bien qu'également dérégulés sont neutres par rapport à cet avantage sélectif et sont seulement associés à des voies de signalisation secondaires au cancer [39]. La question est donc de découvrir une métrique pour séparer les drivers des passagers sur la vraisemblance qu'ils sont réellement sélectifs pour le cancer.

Le développement des technologies microarrays CGH a permis l'identification d'altérations génomiques à grande échelle sur le génome [167]. L'amplification du nombre de copies (Copy Number Amplification - CNA) a été associés avec les sous-types moléculaires et la situation clinique du patient [13]. Cependant, l'abondance et l'hétérogénéité des régions génomiques ayant des aberrations significatives complexifie la recherche de marqueurs biologiques viables et de cibles thérapeutiques [2]. Pour ce faire, les approches utilisées s'orientent vers la fréquence des altérations. Si une altération apparaît plus souvent dans un gène donné, elle représente vraisemblablement une altération clé dans le cancer. Par exemple, l'algorithme GISTIC (Genomic Identification of Significant Target Gene in Cancer) identifie les régions génomiques qui apparaissent aberrantes à une fréquence significative [15] [14].

Cependant, l'analyse des aberrations seules n'est pas suffisante pour la détection des gènes drivers. Les régions altérées dues au ré-arrangements de grande taille détectées par CGH sont habituellement larges et comportent plusieurs gènes. La plupart sont passagers, mais ne dont pas distinguable des gènes drivers sauf analyse supplémentaire. De plus, cette approche ne permet pas de distinguer l'importance physiologique ou fonctionnelle de la région identifiée. Ces limitations soulignent un besoin de nouvelles méthodes plus intégrées qui tiennent compte de plusieurs types d'information pour identifier les gènes drivers. En ce sens, l'analyse d'expression peut apporter un sens biologique non négligeable à ce problème. Par contre, la méthode d'intégration n'est pas triviale et doit être conçue avec précaution pour tirer le maximum de chaque type de données.

4.1.3 CNV-Intégration Transcriptome Interactome (CITI)

Le postulat de base est que l'expression génétique et les aberrations génomiques des drivers sont corrélés à un moment ou à un autre dans l'histoire de la tumeur. Ensemble, l'expression et les mutations génomique forment une emprunte génétique [2] que nous souhaitons identifier et comprendre au sens biologique, et ce, pour chaque sous type moléculaire dans le cancer du sein. Le but est double : d'une part identifier la biologie de la maladie et identifier les cibles thérapeutiques, et d'autre part, prédire le devenir du patient pour adapter le traitement à sa maladie. Plusieurs approches ont été développées pour superposer les informations génomiques et l'expression génétique pour l'identification des gènes drivers. Parmi celle-ci, Akavia et collègues [2] ont développé un algorithme (CONEXIC) pour identifier les gènes drivers localisé dans les régions avec des changements génomiques récurrents. Dans cette approche, chaque gène driver est

associé avec un module génétique dérégulé par le driver. Cette méthode a été validée sur un jeu de données mélanome de 62 tumeurs contenant des mesures appairées expression et amplification chromosomique [96]. Cette analyse a permis la confirmation de plusieurs driver connus et leur connections avec plusieurs cibles. De plus, de nouvelles cibles ont été prédites et confirmées expérimentalement. Cette méthodologie est détaillée plus en détail dans le présent chapitre. Une autre approche intégrative a été proposée [14]. Cette méthode, basée sur la détection de de gènes drivers par analyse expression-aberration génomique est également détaillée dans ce chapitre.

Même si le jeu de marqueurs détectés par ce type de méthode contient bien un certain nombre de gènes drivers, le bruit de mesure inhérent aux puces à ADN et l'hétérogénéité des tumeurs représentent deux difficultés majeures pour l'obtention de marqueurs fiables. Il a été établi clairement que les signatures à base de puces à ADN sont intrinsèquement instables au regard de leur application à la construction d'une signature qui servira de prédicteur [35]. Par exemple, deux jeux de données de référence pour l'étude du cancer du sein et la prédiction de la rechute avec métastase, décrivant respectivement 198 tumeurs [182] et 98 tumeurs [178], avec une signature validée ensuite sur 198 tumeurs [175] ont produit des signatures (76 gènes pour Wang et 70 gènes pour van de Vijver) fondamentalement différentes (Seulement 3 gènes communs), intrinsèquement instables (2% de stabilité, comme décrit dans [35] et qui ne permettent pas la classification de données indépendantes de façon fiable [51].

Les signatures basées sur les puces à ADN sont instables pour 2 raisons. La première est purement statistique, et trouve son origine dans la structure topologique des mesures faites à l'échelle du génome, qui est favorable au fléau de la dimensionnalité. Le nombre de variables est largement supérieur au nombre de tumeurs, ce qui empêche l'utilisation directe d'outils statistique conçus pour un faible nombre de mesures pour un grand nombre d'échantillons ([35]). La seconde est due à la nature même des données que nous souhaitons mesurer. Le cancer est provoqué par les gènes driver qui sont mutés ou subtilement dérégulés et ces modifications provoquent des changements secondaires dans un grand nombre de gènes. Puisque tous les gènes sont placés au même niveau par les puces à ADN, l'information de causalité est perdue. En fait, une analyse par puce à ADN permet la détection des gènes avec la statistique la plus favorable, c'est à dire celle qui sont le plus différentiellement exprimées. Ce ne sont pas celles que nous désirons détecter en priorité mais la conséquence de la dérégulation des gènes drivers.

Pour retrouver les gènes drivers, il est d'abord nécessaire de réduire le phénomène du fléau de la dimensionnalité en accroissant le nombre d'échantillons. Cela peut être fait par analyse simultanée de plusieurs jeux de données appairés expression-amplification génomique. De plus, il faut inclure de l'information biologique sur la causalité des interactions pour différencier les gènes drivers des gènes passagers. Cela peut être fait en superposant un large réseau protéines-protéines (l'interactome humain) sur les données d'expression. Ensuite, au lieu d'associer séparément chaque gène à une statistique indépendante, les gènes drivers pourraient être détectés au niveau de l'interactome à travers la détection des sous-réseaux (groupes de gènes interagissant) dérégulées ensemble pour une condition clinique particulière ou un sous type moléculaire dans le cancer du sein. Plusieurs méthodes ont été proposées pour l'analyse réseau de l'expression des gènes (Voir Garcia et collègues [53] [51] pour une liste de références). Ceci dit, la structure à plusieurs niveaux de l'information biologique n'a pas complètement été prise en compte puisque aucune méthode ne propose l'intégration directe des données d'aberration génomiques.

Dans ce chapitre, nous décrivons l'application d'une évolution marquée de l'algorithme ITI (Integration Transcriptome Interactome), décrit chapitre 3. Cette méthode a été précédemment appliquée à l'analyse d'expression pour la prédiction de la rechute dans le cancer du sein [53, 51, 50]. Je rappelle rapidement ici les résultats obtenus avec ITI décrits en détail chapitre 3. ITI a montré une amélioration significative sur l'existant par la détection de signatures spécifiques

en fonction du statut ER (récepteurs aux oestrogènes). L'apprentissage s'est fait sur un jeu de 900 tumeurs et un jeu de validation indépendant. ITI a permis d'obtenir une précision plus élevée (74%) sur les tumeur ER+, 54% sur les tumeurs ER-) et une plus grande stabilité (32% sur les tumeurs ER+ et 11% sur les tumeurs ER-, obtenues par permutation entre les jeux d'apprentissage et de tests), à comparer aux 3% entre la signature Mammaprint [178] et Wang [182]. Le principe de base d'ITI est de superposer l'interactome humain connu et de détecter une liste de gènes candidats (graines) sur une analyse d'expression différentielle.

Le voisinage de ces graines est exploré récursivement et les gènes interactant sont agrégés en cas de co-expression avec la graine. ITI renvoie des régions de l'interactome qui sont différentiellement exprimées. Nous désignons ces régions sous le nom de sous-réseaux. Une seconde passe confronte ces réseaux à une validation statistique par comparaison de leur score à une distribution de score de sous-réseaux aléatoires obtenus par permutations des échantillons et des interactions protéines-protéines. Les sous-réseaux sont ensuite soumis à une analyse fonctionnelle par enrichissement sur les termes de l'ontologie Gene Ontology (GO) [54]. Le calcul d'enrichissement basé sur un test hypergéométrique, [168] associé aux annotations de la base de données Entrez Gene [119] du National Center for Biotechnological Information (NCBI, Bethesda, MA). L'enrichissement des termes GO est définie classiquement de la façon suivante :

$$en(GO, S) = \frac{g_{GO,t} \in S/|S|}{g_{GO,t} \in Genome/|Genome|} \tag{4.1}$$

$en(GO, S)$ étant l'enrichissement ou la déplétion du terme GO, t dans le sous-réseau S, $|S|$ étant le nombre de gènes dans le sous-réseau étudié, $g_{GO,t}$ étant le nombre de gènes annotés avec le terme GO, t, $Genome$ est l'ensemble des gènes de l'organisme étudié dans l'analyse et $|Genome|$ le nombre de gènes de ce génome (Voir Rivals *et al.* [146] pour les détails du tests statistique associé à cette métrique). La méthode décrite dans Garcia *et al.* [51] a été grandement modifiée pour inclure les mesures d'amplification de nombre de copie (Copy Number Amplification - CNA) directement dans l'analyse pour permettre la détection de gènes drivers. Nous l'avons implémenté sous forme d'un pipeline bioinformatique établi par le chaînage de l'algorithme décrit précédemment pour sélectionner des gènes drivers candidats dans les données CNA et l'algorithme ITI qui a été utilisé pour établir des modules candidat autour de ces gènes et établir une ressource bioinformatique dédiée pour l'exploration des résultats. Ce pipeline est appelé Copy Number Variations-Interactome-Transcriptome Integration (C-ITI) dans le suite du chapitre. Les résultats de cette analyse sont disponibles sur la page web dédiée sz C-ITI [6]

Nous avons appliqué la méthode C-ITI à la recherche de biomarqueurs spécifiques pour les sous types basaux et luminaux A. Grâce à cette analyse, nous avons identifié 123 sous -réseaux qui incluent des marqueurs connus pour la biologie des sous types et de leur interactions avec les modules fonctionnels dérégulés dans les tumeurs basales et luminal A. La superposition de l'information sur le nombre de copie a permis de distinguer les gènes drivers des gènes passagers. Pour comprendre l'implication fonctionnelle de ces modules nous avons confronté la liste des gènes drivers à la biologie connue en cancérologie [16].

6. http://iti.sourceforge.net/citi

4.2 Matériel et Méthodes

4.2.1 Pipeline C-ITI

Cette section décrit le workflow global utilisé pour cette analyse et détaille les jeux de données intégrés. Cette analyse intègre les profils de nombre de copies génomiques (ADN), l'expression des gènes (ARN) et les interactions protéines-protéines. Le pipeline C-ITI utilise l'ensemble de ces données pour détecter les gènes drivers d'un phénotype, celui-ci étant la distinction entre le sous type luminal A et basal dans le cancer du sein. D'abord, nous détectons les gènes significativement amplifiées par puces CGH à haute résolution pour l'ensemble des patients sur les deux sous types. Ces gènes, considérées comme étant potentiellement drivers sur les sous types A et basaux sont ensuite analysées dans un contexte réseau d'interactions de gènes avec ITI [51]. L'analyse réseau permet l'analyse simultanée de l'expression des gènes et des cartes d'interactions PPI (Voir chapitre 3) pour replacer les gènes candidats dans un contexte biologique et déterminer le lien entre expression génique et amplification du nombre de copie. Ces étapes ont été schématisées figure 4.1. Les sections suivantes détaillent l'ensemble des jeux de données (Interactome et microarrays en expression et altérations) utilisées dans notre analyse.

4.2.2 Bases de données d'Interactions Protéines-Protéines

Avant d'utiliser ITI, une carte d'interactions protéines-protéines de référence doit être définie et construite. Ce jeu de données d'interactions doit avoir certaines propriétés, notamment être à une échelle compatible avec les données d'expression que nous souhaitons analyser. Si le jeu d'interactions est trop restreint, il est possible d'utiliser des technologies développées localement en laboratoire pour étoffer la liste d'interactions disponibles. Parmi ces technologies, le système du double hybride dans la levure [189] permet la découverte d'interactions à une échelle compatible avec d'autres types d'expériences génomiques, par exemple en analyse d'expression par puces à ADN. Son principe est basé sur le fait que les facteurs de transcription (Transcription Factors - TF) peuvent interagir avec les domaines d'activation présents à proximité sans interaction directe. Les deux protéines dont nous souhaitons tester l'interaction (*proie* et *appât*) sont introduites dans une levure mutante qui ne peut pas synthétiser certains nutriments sans l'expression d'un gène reporter et qui ne survivra pas dans un milieu sans ces nutriments. Deux types de plasmides sont produits. L'un est fusionné avec le domaine d'attache d'une protéine pour laquelle on souhaite découvrir de nouveaux interacteurs (désignée sous le nom d' *appât*). L'autre est fusionné avec le domaine d'attache d'une protéine (ou d'une librairie de protéines) que l'on souhaite tester (désignée sous le nom de *proie*. Les deux plasmides sont ensuite introduits dans la levure. Si les deux protéines interagissent, le gène reporteur peut être transcrit, et dans le cas contraire ce gène reporteur ne sera pas transcrit et la levure ne pourra pas survivre. Parmi les autres approches disponibles, citons la co-précipitation par affinité par spectrométrie de masse[26], qui consiste à cibler une protéine donnée avec un anticorps pour isoler des complexes à partir d'une solution (technologie appelée *Pull-down*)à. Ces technologies sont maintenant très bien maîtrisées et peuvent être appliquées dans un grand nombre de laboratoires à un coût raisonnable. Cependant, un certain nombre de désavantages doivent être cités. Il y a un nombre très large de faux positifs et de faux négatifs, dus à la technologie elle même. Les faux négatifs peuvent être provoqués par le fait que le screening double hybride a lieu dans le noyau de la cellule et que les protéines mesurées ne vont pas interagir si elles ne sont pas présentent habituellement à cet endroit faute de l'activation correcte des signaux de localisation. D'autre part, des faux positifs peuvent être détectés pour des protéines n'interagissant jamais car non présentes simultanément dans le même compartiment cellulaire *in vivo*.

Pour cette analyse, nous n'avons pas généré les données d'interactions protéines-protéines au laboratoire, mais intégré plusieurs jeux de données disponibles sur plusieurs bases de données. Les bases de données publiques ont été largement alimentées par les méthodes citées précédemment (double hybrides ou autre technologie à haut débit) et donc sont caractérisées par les même limitations que nous venons d'énoncer. Construire un pipeline basé sur ce type d'expérimentation implique des précautions quand aux données. Ceci dit, la plupart des bases de données sont construites autour d'un cœur d'interactions qui ont été validées *in vivo* et considérées aussi fiables. Ces bases sont ensuite complétées par un large nombre d'interactions prédites *in silico*. Ces prédictions permettent d'étendre l'empreinte génomique des bases de façon significatives. Pour analyser spécifiquement les sous-réseaux luminaux A et basaux, nous avons utilisé les bases de données d'interaction reportées dans Garcia *et al.* [51]. Cela inclut la base de données Human Protein Référence Database [86], IntAct [5], The Molecular Interactions database [32], the Database of Interacting Proteins [150], et un jeu de données d'interaction prédites *in silico* par l'algorithme Cocite [138]. Nous obtenons un total de 70 530 interactions binaires parmi 13 202 protéines (voir table 12.1). Toutes les données ont été téléchargées sous forme de fichiers plats à partir des sites Web respectifs des bases de données suivi par des étapes de normalisation et de transformation similaires. Celle-ci incluent la suppression des protéines inconnues (marqué comme telles dans le fichier original), la suppression des auto-interactions et les protéines non humaine en plus du remplacement des divers identifiants utilisés par les numéros d'accessions Entrez Gene du National Center for Biotechnological Information [7], pour permettre la mise en correspondance des gènes entre les différents jeux de données d'interactions d'une part et avec les données CGH et d'expression d'autre part.

4.2.3 Profils d'expression puces à ADN

Pour comprendre les différences moléculaires au niveau de l'expression des gènes entre les sous types luminaux A et basaux, plusieurs jeux de données de puces sont été analysés et superposés aux données d'interactions. Nous avons utilisés les données publiques mises à disposition à l'Institut Paoli-Calmettes sur les puces à oligonucléotides de type pan-génomes Affymetrix HG-U133 Plus 2.0. L'intérêt de cette technologie réside dans l'obtention de profils d'expression sur la quasi totalité des gènes humain en une seule expérimentation. Après quantification des données, les données d'expression brutes ont été normalisées avec la méthode RMA (Robust Microarray Average) disponible en standard dans Bioconductor [8].

La lecture des données et la normalisation RMA ont été faites avec la librairie logicielle Bioconductor *Affy*. Les tables de correspondances identifiants de sondes-identifiants de gènes Entrez Gene [9] (ProbeIDs-GeneIDs) ont été générés par les annotations spécifiques Affymetrix disponibles dans Resourcerer [171] et les profils de sondes ont été combinés (étape de *collapse* par la méthode proposée par Reyal et collègues [144]. Pour chaque jeu de sondes Affymetrix correspondant au même gène, les sondes portant l'extension `"_x_at"`, sont filtrées et la valeur de la sont ayant la plus forte médiane dans son profil d'expression est allouée au gène. Dans le cas d'un jeu de sonde entièrement labellisé `"_x_at"`, la règle de la médiane s'applique sans filtrage. Sur la base de la signature des sous types moléculaires dans le cancer du sein proposé par Hu et collègues [72] , nous avons attribué un sous type moléculaire pour chaque tumeur étudiée. Parmi celles ci, 68 et 80 ont été labellisées respectivement en basal et luminal A.

7. http://www.ncbi.nlm.nih.gov/gene
8. http ://www.bioconductor.org
9. `http://www.ncbi.nlm.nih.gov/gene`

4.2.4 Profils d'amplification de copies

Les mêmes tumeurs ont été profilées par Hybridation Génomique Comparative (array-CGH - Copy Number Hybridization array). Cependant, ceci n'est pas un pré-requis spécifique pour l'approche présentée, puisque les données ont été traitées séparément et intégrées sous la forme de statistiques. Après hybridation, numérisation et acquisition de données, une analyse bioinformatique standard a été appliquée, incluant filtrage initial et normalisation LOESS à l'aide du logiciel Feature Extraction Package (Agilent Technologies, Santa Clara CA USA). Les données ont été extraites sous forme de log ratios (tumeur/contrôle) avec CGH Analytics 3.4 (Agilent Technologies, Santa Clara, CA USA). Ensuite, le calcul du nombre de copies a été fait avec une segmentation binaire circulaire (*Binary Circular Segmentation*[126]). Ensuite nous avons identifié les gènes significativement altérés dans les groupes basaux et luminaux A avec l'algorithme GISTIC version 2 (Genomic Identification of Significant Targets in Cancer [15]) (Voir implémentation spécifique dans Bekhouche et collègues [12]). GISTIC tient compte à la fois de l'amplitude et de la fréquence des altérations dans le jeux de données de tumeurs pour attribuer une p-value à la valeur d'amplification/délétion de chaque gène.

4.2.5 Détection des gènes candidats initiaux par intégration primaire des données d'expression et d'altération génomique.

Pour comprendre l'impact des gènes drivers sur la dérégulation de l'expression dans les basaux et luminaux A, l'information associant la dérégulation de l'expression des gènes et les altérations génomique a été intégrée. Ensuite, l'algorithme ITI a été appliqué pour comprendre l'impact des gènes *drivers* en permettant l'identification de régions de l'interactome humain dérégulées (c'est à dire avec un changement d'expression significatif). La liste initiale des candidat drivers a d'abord été établie par l'adaptation de l'approche décrite en [12], schématisée figure 4.2. Dans un premier temps, nous avons sélectionné un jeu de gènes candidats ayant un score GISTIC significativement différent entre les deux sous types. En utilisant le protocole décrit dans [12], nous avons sélectionné $n = 471$ gènes significativement dérégulés entre les tumeurs luminal A et basales avec un taux de faux positifs (False Discovery Rate) $FDR = 1.10^{-3}$ (Correction Benjamini-Hotchberg). D'autres critères ont été ajoutés pour la sélection de candidats : (i) La fréquence d'altération doit être différente (Test exact de Fisher avec $p \leq 0,05$). (ii) Leur expression et nombre d'amplification de copie doivent être corrélés (Test de Student $p \leq 0,05$). (iii) Ces gènes doivent être différentiellement exprimés entre les sous types basaux et luminaux A en plus de leurs différences génomiques (Test de Student $p \leq 0,05$). La liste des gènes candidats a ensuite été soumise à ITI en tant que graines d'initialisation de l'algorithme.

4.2.6 Autres bases de données d'annotations et éléments de pipeline

En plus des bases de données du NCBI et Resourcerer pour la conversion sonde-gènes, d'autres bases de données ont été utilisées dans le pipeline pour l'annotation de sous-réseaux. En particulier, l'information GO (Gene Ontology[54]) a été employée à des fin d'assignations de fonctions biologiques pour les sous réseaux, à travers une mesure d'enrichissement. Cela fut fait avec le logiciel ErmineJ [91], qui mesure de façon spécifique l'enrichissement GO par un calcul statistique basé sur un test hypergéométrique associée à une correction pour test multiples. De plus, la liste des facteurs de transcription humains a été générée avec la version libre d'accès TRANSFAC[101] (Version au moment de l'analyse : 7.0). Pour afficher les sous réseaux identifiés, le logiciel libre

GraphViz [10] (AT&T Research, Florham Park, NJ, USA) a été intégré au pipeline. La validation statistique des sous réseaux a été implémentée avec la Toolbox Statistiques Matlab (The Mathworks, Natick, MA, USA).

4.2.7 Algorithme ITI d'Intégration Transcriptome Interactome

L'algorithme ITI, illustré figure 4.3 fonctionne par examen simultané de données d'interactions et de données d'expression pour identifier des sous réseaux différentiellement exprimés, c'est à dire des régions de l'interactome dont les gènes présentent un différentiel d'expression sur deux conditions expérimentales (voir chapitre 3). Pour diminuer le coût calculatoire de C-ITI, la détection a été parallélisée par séparation des jeux de données d'entrée (carte d'interactions) et distribution sur les différents nœuds de calcul du cluster Beowolf du CRCM. ITI fonctionne en deux étapes principales, une première étape de détection de sous réseaux suivie d'une étape de validation statistiques. Pour la détection des sous réseaux différentiellement exprimés, la corrélation entre les phénotypes ou variables cliniques des deux conditions est calculée. Ensuite le jeu d'interactions est parcouru de façon exhaustive pour la recherche de régions discriminantes en considérant chaque gène comme graine de sous réseau potentiel et en agrégeant récursivement ses voisins s'il accroissent le score du sous réseau courant.

Le score S du sous-réseau s sur un jeu de données d est calculé suivant la formule présentée équation 4.2.7. Le score est calculé par corrélation de Pearson entre l'expression moyenne et un vecteur représentant le phénotype. Les voisins des nœuds sont examinés de façon récursive et inclus dans le sous-réseau courant si leur expression permet d'accroître le score d'une valeur minimale. Le seuil minimal du score avant validation statistique est de $S = 0.3$ et le seuil minimale d'accroissement du score est $r = 0.03$. Une fois que le score ne peut plus être amélioré, on ne considère pas de gènes supplémentaires pour le sous réseau courant.

Ce score mesure la corrélation entre l'expression moyenne d'un sous réseau et le sous type moléculaire. Un facteur de normalisation est appliqué par l'utilisation du nombre total d'échantillons n_d. Ce facteur n'est pas utilisé pour l'analyse d'un jeu de données unique mais permet d'appliquer une échelle corrective lorsque l'on compare plusieurs jeux de données, un jeu de données ayant un faible nombre de conditions ayant tendance à produire de plus fortes valeurs de corrélations.

Lorsque plusieurs jeux de données sont analysés, le score S_s est calculé par moyennage des scores individuels sur chaque jeu de données d (Équation 4.2.7 sur la liste des jeux de données DS de longueur NS.

$$S_{s,d} = \frac{\sqrt{n_d}}{\sqrt{\max n_d(DS)}} corr\left(\frac{1}{n}\sum_{g\in S} e(g,d), cc(d)\right) \tag{4.2}$$

$$S_s = \frac{1}{NS}\sum_{d\in DS} S(s,d) \tag{4.3}$$

Les sous-réseaux qui ont des gènes communs avec les sous réseaux déjà détectés sont traités de la façon suivante : Le chevauchement du réseau courant avec les sous réseaux existants est calculé comme étant l'inclusion maximale de du réseau A dans B ou de B dans A. L'inclusion de A dans B est calculée par comptage du nombre de gènes communs entre les sous réseau A et B suivi d'une division par le nombre total de gènes contenus dans le sous réseau A. En pratique,

10. http ://http ://www.graphviz.org

les sous réseaux présentant un chevauchement de plus de 50% avec les sous réseaux existants ne sont pas retenus.

Une fois les sous réseaux détectés, ils doivent être validés de façon statistique. Ceci est fait par génération de deux distributions pour la mise en place de seuils de score validés statistiquement et associés avec une p-valeur. La première distribution permet de valider la pertinence de l'algorithme ITI lui même. Elle est obtenu par sélection aléatoire de sous-réseaux, c'est à dire remplacement de l'algorithme de décision par une agrégation aléatoire autour d'une graine elle même choisie aléatoirement. La seconde distribution permet de vérifier si le lien interactome-expression des gènes est validé statistiquement. Cette seconde distribution est obtenue par mélange des sous types luminaux A et basaux. Pour garder les sous réseaux détectés aléatoirement comparables en taille par rapport aux sous-réseaux détectés précédemment, leur distribution de taille a été forcée pour suivre une distribution Gaussienne modélisée de façon identique à celle des vrais réseaux. Après génération des deux distributions, celles-ci sont modélisées par un modèle de mixture Gaussien. Ce modèle est ensuite utilisé pour établir un seuil significativement statistique sur les scores des sous-réseaux. Seuls les sous-réseaux passant la validation statistique sur les deux distributions sont retenus. Les sous réseaux chevauchant sont groupés grâce au calcul de recouvrement précédemment défini s'ils se recouvrent sur un seuil supérieur à 0_s spécifié en tant que proportion. Pour chaque cluster, le sous-réseau ayant le score le plus élevé est conservé.

4.3 Analyse

Dans cette section, nous détaillons les étapes d'analyse des données ainsi que l'analyse des gènes drivers obtenus avec le pipeline C-ITI.

4.3.1 Détection des gènes drivers avec intégration transcriptome-interactome

Pour détecter séparément les sous-réseaux exprimés dans les sous-types basaux et luminaux A, nous avons appliqué deux analyses C-ITI séparées sur le jeu de tumeurs spécifique à chaque groupe. Ces deux analyses ont pris en entrée la liste des 471 gènes candidats qui passent les filtres CGH initiaux pour les utiliser comme cœurs de réseaux candidats. Pour générer des sous-réseaux sur l'interactome humain complet, deux types de données ont été générés. D'une part la base de données d'interactions précédemment assemblée a été utilisée pour l'exploration. D'autre part, l'expression génique pour 148 tumeurs a été superposée sur l'ensemble de l'interactome humain. Ensuite, la corrélation entre les sous types moléculaires et l'expression a été calculée et les sous-réseaux extraits et validés statistiquement sur les deux distributions aléatoires par p-value $Pval_1 = 1.10^{-3}$ et $Pval_2 = 1.10^{-3}$ respectivement. Après la suppression des sous-réseaux chevauchant (seuil de chevauchement établi à $O_s = 60\%$), 123 sous-réseaux ont été retenus comme étant différentiellement exprimés et utilisés pour la suite de l'analyse, avant un total de 541 gènes. Parmi ceux-ci, 62 (279) gènes ont été détectés comme étant exprimés dans les sous-types basaux et 61 sous-réseaux (262 gènes) ont été détectés dans les sous-types luminaux A. Une analyse séparée de l'expression globale et CNA a donné 5000 gènes différentiellement exprimés et 1000 gènes présentant des fréquences d'altérations distinctes entre les deux sous types [1].

4.3.2 Visualisation des sous-réseaux

La visualisation des données de cette analyse est une étape complexe car elle impose une intégration des données d'interactions et d'expression sur un grand nombre de tumeurs sur deux

sous types distincts, ainsi que les information de CNA et des annotations de gènes. Pour accommoder l'ensemble de ces sous types, nous avons modifié les routines de visualisation du pipeline original ITI pour superposer les valeurs GISTIC. La figure 4.4 illustre un exemple de sous-réseau obtenu avec C-ITI (Donnée accessible à partir du site web d'ITI). Le sous-réseau présenté Figure 4.4.A est exprimé dans le sous-type luminal, alors que le sous-réseau présenté Figure 4.4.B est exprimé dans le sous-type basal. Pour la visualisation, deux figures additionnelles pour chaque sous réseaux sont générées en plus de l'expression pour la visualisation des scores GISTIC (avec un code couleur approprié pour refléter les pertes ou gain homozygotes ou hétérozygotes). Les sous-items A.1 et B.1 représentent les scores d'expression obtenus avec ITI (bleu= corrélation avec le sous-type luminal A, rouge=corrélation avec le sous-type basal). Les scores GISTIC sont représentés sur les nœuds du réseau avec un bord de couleur (rouge = gain, vert = perte) pour les sous-types luminal A (A.2 et B.2) et basal (A.3 et B.3). Le niveau d'altération ou d'amplification est représenté avec un code couleur. Le nœud est laissé blanc quand le score GISTIC n'est pas considéré significatif ($p - valeur < 0.05$). Une capture d'écran du rapport d'analyse C-ITI est présenté Figure 4.5.

4.3.3 Analyse fonctionnelle des sous-réseaux

Les sous-réseaux sont ensuite analysés pour leur pertinence biologique. Les gènes connus pour être spécifiquement exprimés dans les deux sous types sont identifiés. Pour le sous type luminal A, le gène le plus fréquemment trouvé dans les sous-réseaux est, sans surprise ESR1 (trouvé dans 9 sous-réseau). FOXA1 est également fréquemment trouvé par ITI (3 sous-réseaux). ERG est présent dans 1 sous-réseau. Au contraire, les récepteurs aux œstrogènes ou ERBB2 n'ont pas été détectés dans les tumeurs basales. Puisque ces tumeurs sont hautement prolifératives, les gènes liées au cycle cellulaire (cyclines en général et les CDKs) ont été retrouvées [16]. De plus, parmi les 62 sous-réseaux exprimés dans les basaux, CDK6 a été la plus fréquemment trouvée (8 occurrences). Elle régule le suppresseur de tumeur RB1. Les gènes codant pour la cycline E1 (CCNE1) et CDK2 apparaissent respectivement dans 5 et 3 sous-réseaux.

Bertucci *et al.* [16] ont construit une liste de gènes discriminants les sous types luminaux et basaux à partir de la littérature. Nous avons croisé cette liste avec celle détectée avec C-ITI pour vérifier si les gènes précédemment associés avec ces phénotypes ont été retrouvés. Parmi ceux-ci, Cyclin D1, MYB, un facteur de transcription important pour la tumorigenèse [103], SMAD3 (Un composant de la signalisation TGF-β [187]) étaient présents dans les luminaux A. Le facteur de transcription RUNX3 était présent dans les tumeurs basales. C'est un suppresseur de tumeur régulant la carcinogénèse dont l'expression est réduite dans les cancer du sein[123]. La protéine kynase LYN est également exprimée dans les sous-type basaux. Elle encode une protéine tyrosine kinase connue pour être impliquée dans les voies de signalisation activées dans les sous types basaux [38]. CDK6 est connue pour son expression élevée dans plusieurs sous types [147]. Plus de 5000 gènes sont ainsi différentiellement exprimées entre les deux sous types, ce qui rend l'analyse complexe d'un point de vue pratique. L'analyse intégrée présentée ici propose une réduction drastique du nombre de gènes drivers ($n = 472$) et leurs interacteurs dans l'interactome humain, ce qui réduit la liste de candidats potentiels, puisque seulement 114 gènes ont été retenus comme cœur de sous-réseau. La liste complète est présentée Figure 4.6.

4.3.4 Identification des gènes mutés, suppresseurs de tumeurs et oncogènes

La figure 4.6 révèle les gènes identifiés par C-ITI, avec une liste détaillée des candidats et la liste des gènes sur-exprimées respectivement dans les luminaux A et basaux. Les score GISTIC

ont révélé plusieurs pertes génomiques significatives dans les deux sous types, particulièrement avec les basaux ($n = 131$), mais également avec les luminaux A ($n = 78$). Les amplification génomiques sont également significatives ($n = 78$ et $n = 56$) pour les basaux et luminaux A, respectivement. Bien que les aberrations génomiques soient spécifiques à chaque sous type, il est également possible de différencier les deux catégories de tumeurs sur le nombre d'altérations, qui sont largement supérieures dans les tumeurs basales. Lorsque l'on analyse uniquement sur les graines de réseau, 60% montrent une perte génomique dans la population basale, et seulement 21% dans la population luminale A. Les gains ont été moins fréquents et observés pour les gènes *CPB1*, *IL20RA*, *TNIK*, *EPB4L2*, *MRC1* pour les tumeurs basales et pour *CREBBP*, *TSC2*, *SPAG5* et *TNFSRF17* pour les luminaux A.

Ces pertes ont un impact significatif pour l'expression des transcrits correspondants et des gènes avec lesquels ils interagissent. Ces gènes sont donc considérés drivers pour les basaux ($n = 75$) et pour les luminaux A ($n = 30$ drivers). Les gènes associées avec les amplifications pourraient être également associées avec des drivers potentiels, bien qu'avec un impact moindre, car ils influencent également l'expression dans le voisinage de l'interactome.

Pour comprendre le rôle de ces gènes, nous avons examiné leurs annotations de façon manuelle. Plusieurs gènes identifiés par C-ITI n'avaient pas été précédemment liés à la tumorigenèse, et sont donc d'un intérêt primordial. Par exemple, *CPB1*, les kinases *TRAF2* et *TNIK*, impliquées dans la signalisation *JNK*, ainsi que *EPB41L2* n'ont pas de rôle connu dans la carcinogenèse mais ont bien été détectées comme étant différentiellement exprimées.

Un total de 28 facteurs de transcriptions a été détecté par C-ITI, 8 d'entre eux en tant que cœurs de réseaux. Ceux-ci sont *C16orf80*, *LM04*, *GTF2B* et *SOX10* (perdu et sous-exprimé dans les tumeurs luminal A), ainsi que *FOXA1*, *EGR1*, *HIF1A*, *YBX1* et *C16orf80* (perdu et sous exprimé dans les tumeurs basales).

Le gène *EGR1* est connu comme étant un gène suppresseur de tumeur. *FOXA1* est connu pour promouvoir la croissance de tumeurs dans plusieurs types de cancer [84] et il a été établi que les sites contenant des SNPs associés aux cancer du sein [83].

4.3.5 Autres méthodes d'intégration destinées à la détection de gènes drivers en cancérologie

Pour illustrer la diversité des approches pour la détection de gènes drivers en cancérologie, nous détaillons ici deux autres méthodes pour la détection de gènes dans les mélanomes (CONEXIC [2]) et dans les carcinome du rein [14]). Au moment de la publication de Garcia *et al.* (2013) [52], il n'existait pas de méthode définitive permettant l'intégration des données d'interactions protéines-protéines et nombre de copies, et donc celles présentées ici sont considérées comme étant les plus pertinentes pour notre analyse.

4.3.6 Méthode CONEXIC par Akavia *et al.*

Le postulat de base d'Akavia est que les mutations de type drivers sont corrélées avec l'expression des gènes. Un gène driver dérégule un module de gènes qui provoque la tumorigenèse. Un gène driver peut également être dérégulé sans altération significative de sa séquence. La méthode d'Akavia consiste à intégrer l'expression des gènes et l'amplification du nombre de copies pour trouver des signatures contenant vraisemblablement des gènes drivers. Cette méthode est basée sur un algorithme publié précédemment, *Module Networks* [153] qui permet de rechercher des modules de gènes co-régulés à partir de données d'expression et de relation de régulation connues. COpy Number and Expression in Cancer (CONEXIC) étends *Module Networks* pour

permettre la détection de gènes drivers à partir de données d'expression. CONEXIC utilise une recherche basée sur un score proportionnel au changement de nombre de copies pour détecter les modules différentiellement exprimés et ayant le score le plus élevé dans les régions amplifiées ou effacées. La sortie est une liste classée de modulateurs ayant un score élevé et corrélant avec les modules différentiellement exprimés et étant localisé dans des régions altérées du génome. Les modules eux-mêmes sont modulés, ce qui indique que les gènes divers affectent l'expression des modules auxquels ins sont rattachés. Puisque ces altérations sont trouvées de façon récurrente dans un nombre significatif de tumeurs, il est vraisemblable que modulateurs soient liés à la tumorigenèse. Pour vérifier ce lien, CONEXIC a été appliqué à l'étude d'échantillons dans le mélanome. D'abord une liste de drivers candidats a été générée par l'analyse des 101 profils de nombre de copies (CNA) disponibles sur les 101 échantillons tumoraux par GISTIC. Ensuite CONEXIC a été appliqué à 62 échantillons appairés pour sélectionner les drivers de modules. Un total de 64 modulateurs a été sélectionné par l'algorithme pour expliquer le comportement de 7869 gènes. Les 30 modulateurs ayant le score le plus élevé sont considérés comme étant des gènes driver de la tumorigenèse. Pour annoter les gènes drivers et les gènes modulés par ces drivers, une procédure automatisée a été développées par les auteurs (LitVan), qui permet de connecter les gènes à la liste des articles concernés sur la base *PubMed* du NCBI.

4.3.7 Méthode de Beroukhim *et al.*

Beroukhim *et al.* [14] ont développé une analyse intégrée (Données d'altérations génomique et d'expression) sur des carcinomes du rein à cellules claires de forme sporadique (*sproradic clear cell kidney carconoma* - ccRCC) et la maladie de von-Hippel Lindau (VHL). Un total de 90 tumeurs a été analysé et les altérations génomiques significatives identifiées. Le lien entre les régions amplifiées et les changements d'expression a permis l'identification de plusieurs gènes de grand intérêt, dont CDK2NA, CDK2NB et MYC, parmi d'autres. L'analyse est articulée de la façon suivante. : D'une part, les donnés sont générées par extraction d'ADN et d'ARN des tumeurs ccRCC et VHL. Après hybridation de l'ADN sur puces CGH, l'ARN a été hybridé sur puces à ADN. GISTIC a identifié 7 régions amplifiées et 7 régions effacées dans les tumeurs ccRCC validées par une $q - value < 0.25$. Pour tenir compte du fait que plusieurs régions peuvent avoir été déplacées par des événements passagers, des régions plus robustes (appelées régions à pic larges) ont été identifiées avec une méthode de validation croisée robuste à la suppression aléatoire d'une tumeur. Les auteurs ont ensuite assumé que les oncogènes dérégulés par les phénomènes d'amplifications d'ADN sont activés par sur-expression. Ensuite, une analyse intégrée a permis de prioriser les gènes dans les régions amplifiées à pics larges. Spécifiquement, les gènes significativement exprimées dans les tumeurs ont été sélectionnées par permutations de labels de tumeurs et calcul d'un rapport signal sur bruit. Parmi les gènes situées dans les régions à pics larges pour lesquels des sondes existent, 23 sont significatives. Par exemple, MYC est amplifié dans l'ensemble des tumeurs avec une amplification 8q24. Une approche similaire pour la détection des gènes sous-exprimées a permis de sortis plusieurs gènes suppresseurs de tumeurs candidats. Parmi celles-ci, citons CDKN2A, un gène suppresseur de tumeurs connus, a été identifié.

4.4 Discussion

Le cancer du sein est une maladie extrêmement hétérogène, avec des tumeurs caractérisée par des événements génomiques qui sont à la fois communs et unique, et associés avec des changements d'expression très différents entre tumeurs. Cela rend l'identification de marqueurs

pronostiques difficile. Nous sommes particulièrement intéressés par deux sous-types présentant des caractéristiques génomiques, transcriptomique et cliniques opposées, les luminaux A (cellules différenciées) et basaux (cellules prolifératives).

Pour s'attaquer à l'hétérogénéité de cette maladie, des méthodes intégratives doivent être appliquées pour tenir compte de l'information disponibles à plusieurs niveaux biologiques et permettre la séparation des *gènes drivers* (gènes à l'origine de la maladie et sélectionnée comme étant avantageuse à la tumorigenèse) des *gènes passagers* (gènes dérégulés ou altéré comme résultat collatéral des changements d'état des gènes drivers). Les méthodes intégrant l'information liée aux CNA (mesurées par CGH-arrays) et l'expression des gènes (mesurée par les puces à ADN) ont été développées avec des exemples d'applications sur différent types de cancer. Elles permettent de de détecter des marqueurs donc le nombre de copies est soit réduit et associé avec une baisse d'expression, soit amplifié avec une sur expression. En particulier, la méthode de Beroukhim a permis la détection de nouveaux oncogènes dans les tumeurs ccRCC. la méthode d'Akavia, CONEXIC, utilise les informations de régulation pour associer la détection des drivers à une notion de modules et séparer les évènements drivers/passagers. L'application à un jeu de données dans le mélanome a permis l'identification de nouveaux candidats drivers confirmés par la présence d'oncogènes connus.

L'application aux sous types moléculaires luminal A et basal donne des centaines de candidats potentiels. Une information biologique supplémentaires doit être prise en compte par les algorithmes pour proprement distinguer gènes drivers et passagers, ainsi que les modules de gènes connus dérégulés et ce, séparément pour chaque sous type. Nous avons combiné la méthode d'intégration de Bekhouche *et al.* [12] et l'algorithme ITI pour construire le pipeline C-ITI. Un jeu de 471 gènes candidats a d'abord été sélectionné par une étape d'intégration CNA-expression. Ces gènes ont ensuite été soumises à C-ITI en tant que graines de sous-réseaux pour déterminer si elles sont liées à des modules dérégules en expression. Après analyse, C-ITI a donné une liste de sous-réseaux différentiellement exprimés, après un parcours complet de l'interactome humain et une validation statistique. Seuls 24% des 471 gènes initialement soumis à C-ITI ont été retenus dans 123 sous-réseaux validés. 61 sous-réseaux exprimés dans le sous-type luminal A et 62 sous-réseaux exprimés dans le sous-type basal ont été identifiés. Les marqueurs connus sont retrouvés (*ESR1* pour les luminaux A, les cyclines et kinases pour les basaux) et de nouveaux oncogènes et gènes suppresseurs de tumeurs ont été découverts. Cette approche a réduit drastiquement le nombre de marqueurs candidats, ce qui a accru leur significativité statistique et leur valeur biologique.

4.5 Conclusion du chapitre

Dans ce chapitre, nous avons décrit l'algorithme intégré C-ITI (Copy Number Variation-Interactome-Transcriptome Intégration) destiné à la détection de gènes drivers en cancérologie. Ce pipeline fonctionne en deux étapes. D'une part, il détecte les marqueurs candidats par intersection des gènes différentiellement exprimés et des gènes ayant un nombre de copie significativement altéré. Ensuite, ces candidats ont été soumis à l'algorithme C-ITI pour la recherche de sous-réseaux différentiellement exprimés dans l'interactome humain. Les gènes retenus sont ceux confirmés comme étant impliqués dans les sous réseaux différentiellement exprimés, puisqu'ils sont sélectionnés comme étant les modules spécifiquement dérégulés dans les sous-types étudiés. Ensuite, l'ensemble des données est stocké dans une ressource bioinformatique pour visualisation et analyse. Nous avons également amélioré la partie visualisation/génération de rapports disponible de base dans ITI en permettant l'ajout de l'information des altérations génomiques

sur les sous-réseaux. Comme exemple illustratif, nous présentons l'analyse de deux sous-types moléculaires dans le cancer du sein, basal et luminal A, pour lesquels nous avons identifié les gènes drivers. Le pipeline C-ITI pourrait être appliqué pour obtenir une vue intégrée de la masse de données générées par les différents consortiums (*The Cancer Genome Atlas* [30], *The International Cancer Genome Consortium* [76] et autres). Cela permettrait l'obtention de marqueurs fiables non seulement statistiquement significatifs mais aussi avec une information biologique complète sous la forme d'un jeu d'interactions qui nous permettraient de réellement valider l'impact de ces drivers sur le développement des maladies étudiées. Ces analyses intégrées sont une étape nécessaire pour comprendre les mécanismes de tumorigenèse en cancérologie. Celle présentée ici pourrait être étendue pour intégrer d'autres niveau d'information tels que les mutations ponctuelles et la régulation par les micro-ARNs.

4.6 Remerciements

Nous remercions les sources de financement qui ont permis de mener à bien ce travail. Cette recherche a été financée par l'Institut National du Cancer (INCa), La Ligne Nationale contre le Cancer (Label Daniel Birnbaum). le support pour l'infrastructure de calcul a été obtenu par un financement de la Fondation pour la Recherche Médicale (FRM) . Maxime Garcia a été financé par un financement Institut National de la Santé et de la Recherche Médicale - Région Provence-Alpes Cote d'Azur, Le support pour Raphaëlle Millat-Carus a été obtenu par l'Institut National du Cancer. Je remercie Sabrina Carpentier pour les discussions qui ont permis la mise au point de la version originale d'ITI ainsi que Wahiba Gherraby pour la lecture du manuscrit de Garcia *et al.* [52].

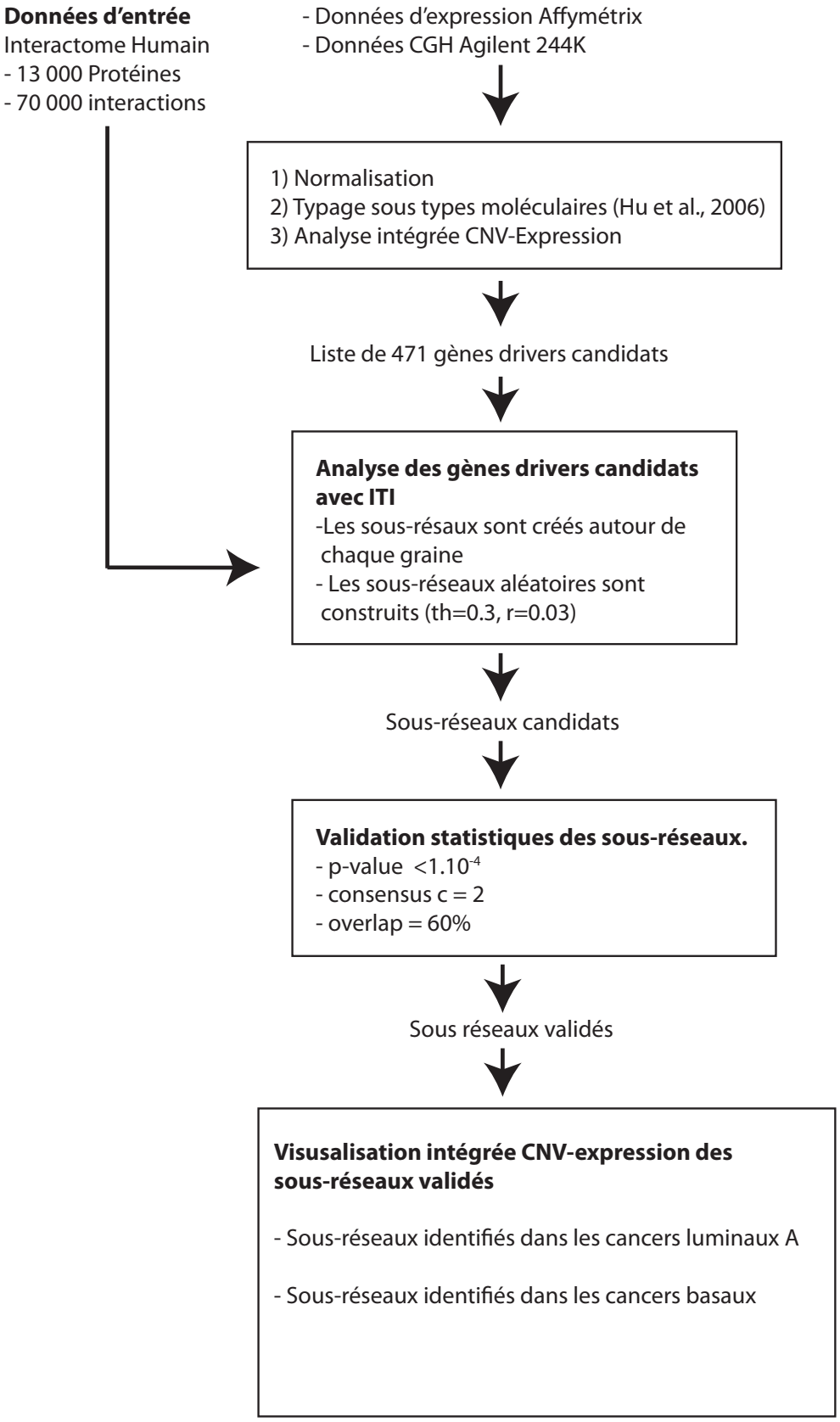


FIGURE 4.1 – Algorithme C-ITI et organisations des données. Les 471 gènes sélectionnés comme graines lors de l'étape précédente sont analysés avec ITI. ITI prends en entrée une carte d'interactions PPI (interactome), et construit des sous-réseaux autour des gènes préselectionnés pour les sous-types basaux et luminaux. Ensuite, ces sous-réseaux sont validés statistiquement.

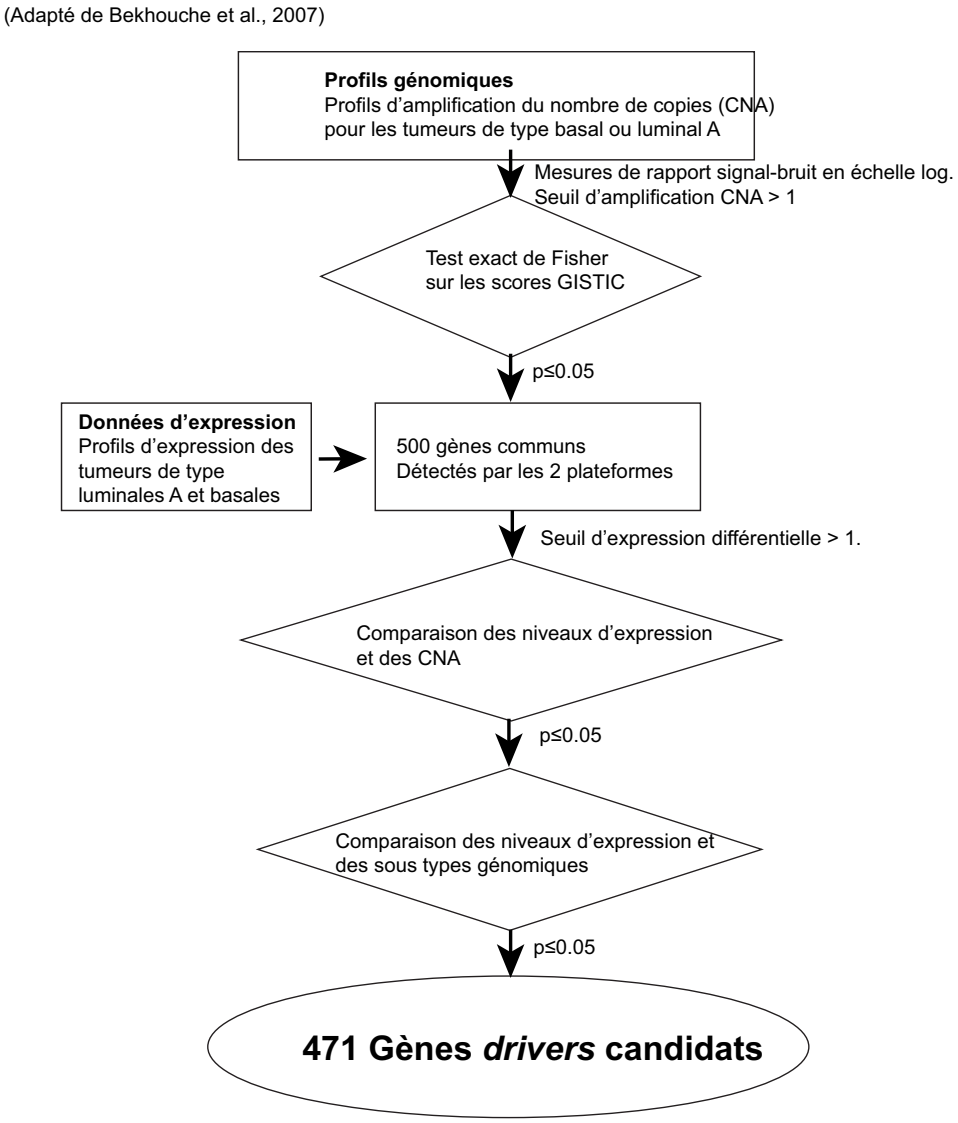


FIGURE 4.2 – Détection initiale des gènes candidats initiaux par intégration primaire des données d'expression et d'altération génomique.

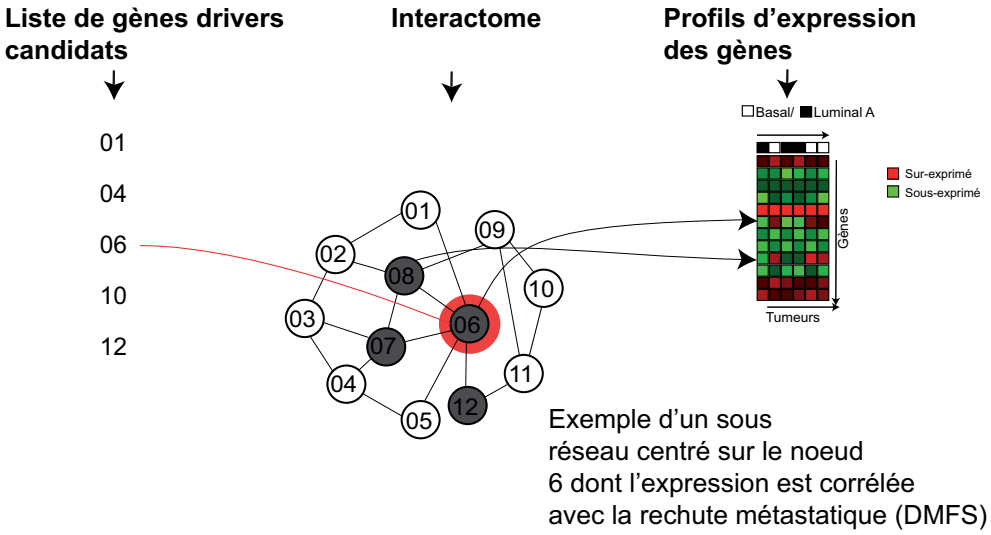


FIGURE 4.3 – Algorithme ITI appliqué aux 471 gènes drivers candidats sélectionnés lors de l'étape précédente. Les sous-réseaux sont agrégés récursivement autour de ces graines si leur expression est corrélée avec les sous types moléculaires.

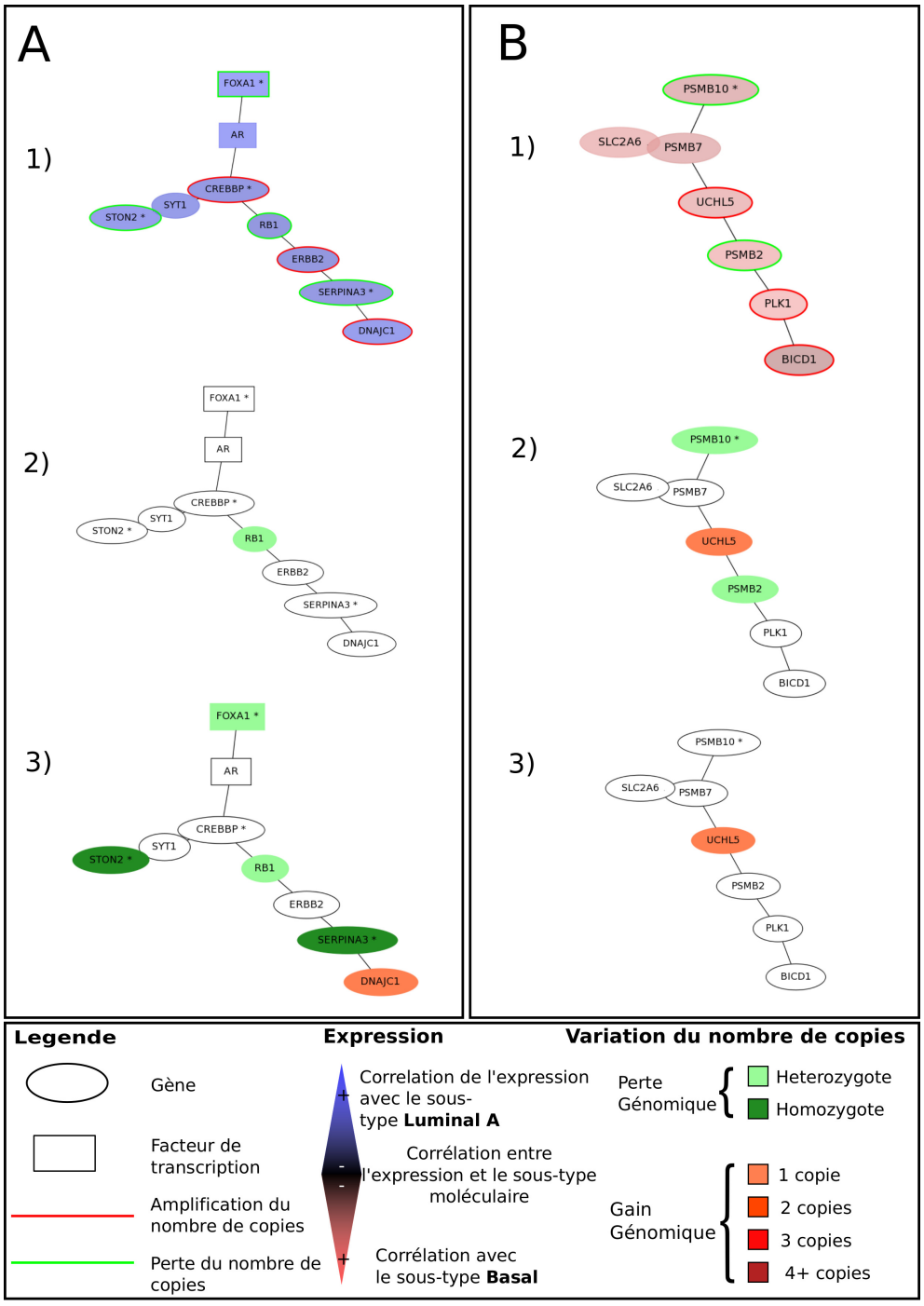
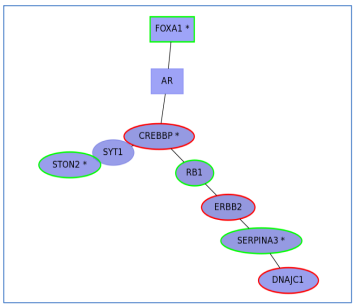


FIGURE 4.4 – Visualisation des données générés par C-ITI. Deux exemples de sous-réseaux détectés avec ITI sont représentés ici, l'un exprimé dans le sous type basal, et l'autre dans le sous-types luminal A. En 1) est représenté l'expression(Rouge=corrélation avec le sous type basal, bleu=corrélation avec le sous type luminal A. L'information sur le nombre de copie pour chaque gène est également incluse. En 2) figurent les changements de nombre de copies dans les tumeurs basales et en 3) les changements de nombre de copies dans les tumeurs luminal A.

Dataset	Score	P-val 1	P-val 2	P-val 3
IPC-255	0.8978	1.603e-03	1.810e-04	8.053e-02

expression (LuminalA versus Basal)	IPC-255
aCGH-LuminalA	IPC-255
aCGH-Basal	IPC-255



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

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 4.5 – Rapport d'analyse C-ITI. Les différents éléments de visualisation associés avec les sous-réseaux sont représentés : Les sous-réseaux et leurs p-values, la structure en graphe des sous-réseaux et leurs interactions, superposée avec l'expression des gènes et les valeurs de nombre de copie (Score GISTIC). Le score individuel ITI pour chaque gène inclus dans chaque sous-réseau, ainsi que les liens vers la base d'annotations *Entrez Gene* du NCBI. Les annotations Gene Ontology GO et leur mesure d'enrichissement sont également incluses.

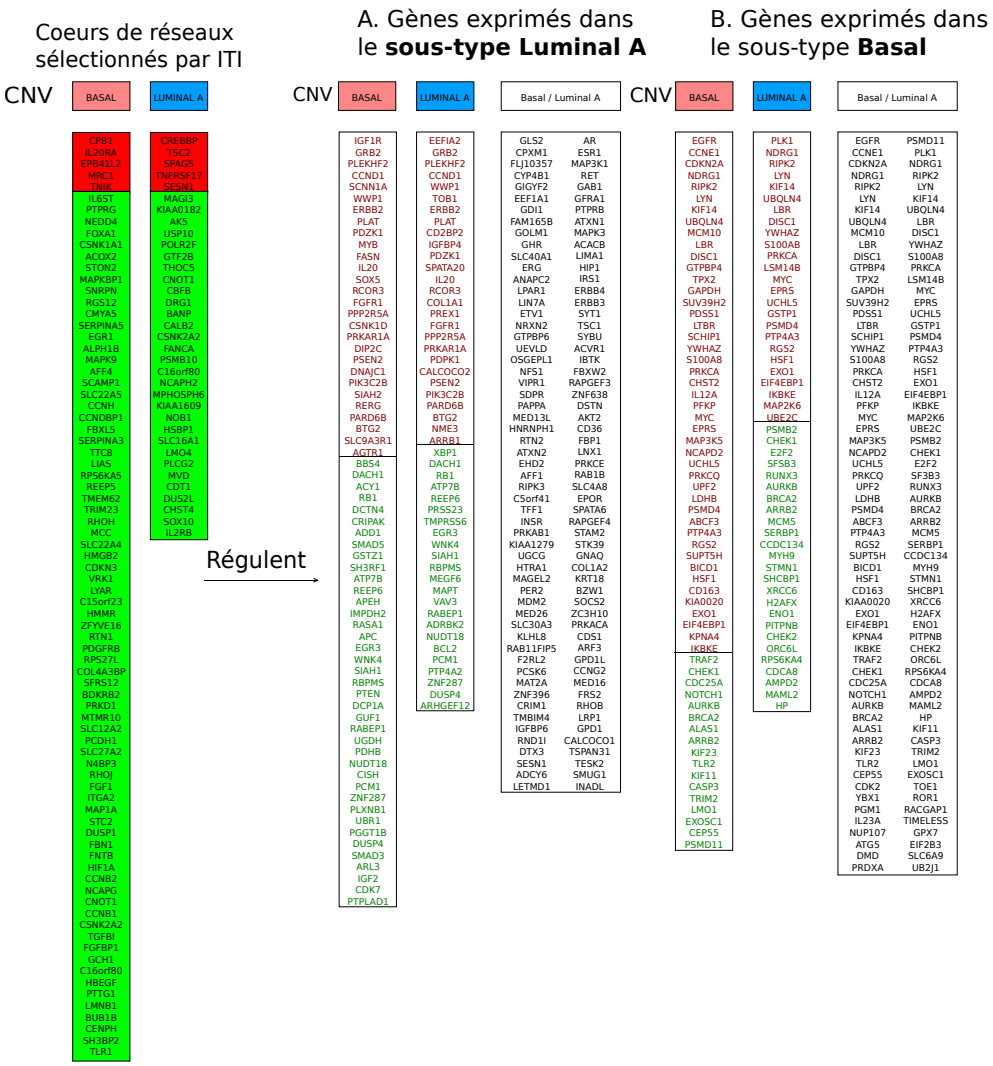


FIGURE 4.6 – Listes des gènes drivers détectés avec le pipeline C-ITI. Sur la gauche sont représentés la liste des gènes spécifiquement exprimés dans les sous types luminaux A et basaux et qui ont été retenus comme cœur de réseau par ITI. Le code couleur représente la nature de l'aberration chromosomique de ces gènes (rouge=amplifié, vert=perdu). Ces gènes sont considérés comme étant des drivers putatifs, c'est à dire suppresseur de tumeurs ou oncogènes. Sur la droite est représentée la liste des gènes régulés ou interagissant avec ces drivers. Dans A, la liste des gènes trouvées dans les sous-réseaux exprimées dans les luminaux A. En B, la liste des gènes trouvées dans les tumeurs basales. Pour les listes A et B, nous avons détaillé la liste des gènes mutées le cas échéant dans les tumeurs basale ou luminal. Dans chacune de ces listes, les gènes marqués rouges sont amplifiées, et les gènes marqués verts sont effacés.

5

Analyse de données de screening siRNA par HTS-Net : *High-Throughput Screening - Network*

Sommaire

Ce chapitre reprend les données générées dans l'article [145] que j'ai publié dans PlosOne en 2017 en tant que dernier auteur. C'est une extension du travail développé en 2013 par Claire Rioualen et Quentin Da Costa qui ont fait leur stage de M2 spécialité Bioinformatique, Biologie

Structurale et Génomique de L'université Aix-Marseille (BBSG) sur la plateforme Cibi sous ma responsabilité, suivi de contrats ingénieurs durant lesquels ce projet a été développé.

Claire Rioualen a développé la partie algorithmique liée à l'analyse de régulation, l'intégration des bases de données de régulation, et la production de rapport d'analyse. Quentin Da Costa a développé l'algorithme de parcours de l'interactome et la validation statistique.

5.1 Introduction

Les études fonctionnelles par les librairies de screening RNAi (RNA interférence) ont profité d'améliorations techniques significatives en termes de vitesse, qualité des données et couverture génomique, grâce aux améliorations des méthodes biochimiques de perturbation des mécanismes de transcription. Les screenings RNAi et la construction de librairies pour les petits ARN interférant associées (*Small Interfering RNA* - siRNA) permettent une compréhension beaucoup plus fine de la fonction des gènes à l'échelle du génome. En parallèle, l'amélioration des technologies de microscopie analytique et le développement d'outils d'analyse de contenu ont permis aux scientifiques d'accéder à des analyse très riches au niveau cellulaire. Mises ensemble, ces technologies ont marque la mise en place des screening à haut débit (*High Throughput Screenings*, HTS) avec des sorties cellulaires sophistiquées à l'échelle génomique. Les gènes cibles détectés par ce type d'expérience peuvent rapidement lier une protéine unique à la fonction biologique ou au phénotype étudié [108] [109].

L'analyse de données primaires, incluant la normalisation [114], le contrôle qualité et la sélection de gènes cibles [192], est bien développée dans la littérature, et des logiciels stables ont commencé à émerger. Alors que les statistiques classiques sont largement utilisées pour reporter les gènes cibles, (z-score, rapport signal à bruit), de nouvelles statistiques ont été proposées, comme la mesure de différence de moyenne strictement standard (*strictly standardized mean difference* - SSMD) [191]. Gotkug et collègues [57] ont proposé un système intégré, GUITar, permettant d'effectuer l'analyse primaire et la détection des gènes cibles dans un environnement doté d' une interface simple d'utilisation. Cependant, il est clair qu'une simple liste classée de gènes n'est pas suffisante pour décrire précisément les processus biologiques affectés par un screening RNAi.

Puisqu'une fonction cellulaire découle de plusieurs interactions construites entre les réseaux de gènes des méthodes sophistiquées d'analyse réseaux sont nécessaires pour analyser ces interactions et comprendre comment elles affectent la biologie. De manière générale, les analyses de type réseau évitent l'utilisation de jeux de gènes pré-définis et identifient *sans biais* de nouveaux pathways d'intérêt sur un phénomène biologique donné à partir d'une carte d'interaction des gènes. Cela permet la découverte potentielle de pathways qui jouent un rôle dans la fonction biologique étudiée, puisque l'analyse se fait à partir de la connaissance de base des pathways connus vers des gènes non assignés. L'intégration des z-scores et des données d'interaction permet également de diminuer le bruit, c'est à dire de réduire le nombre de faux-positifs et de faux négatifs, par exemple pour la recherche d'une signature génomique en apportant un niveau de validation supplémentaire [51]. Dans le cas de la recherche d'une cible thérapeutique, une analyse réseau permet d'identifier un sous-réseau de gènes au lieu d'une protéine isolée et donc d'être plus spécifique en terme d'efficacité thérapeutique. Ce type d'approche a été appliquée avec succès sur des données d'expression pour de la classification tumorale [63, 51], la prédiction de la réponse d'un médicament [41], et de la stratification de tumeurs à partir de variants identités par des technologies de séquençage de nouvelle génération (Next Generation Sequencing, NGS) [69].

Gonzalez et Zimmer [59] ont été parmi les pionniers des approches d'analyse de screening

RNAi basé sur les réseaux d'interactions. Ils ont développé une méthode basée sur le co-clustering [66] qui a été appliquée à la découverte des facteurs liés à l'hôte du virus de l'hépatite C (Hepatitis C Virus, HCV). Ils ont pu identifier des modules de gènes en calculant une matrice de distance entre gènes, combinant leur z-score RNAi et leur chemin le plus court dans l'interactome. Ils ont ensuite fait une classification ascendante hiérarchique à lien moyen. Une fois les groupes de gènes formés, leur taille a été contrôlée par taux de faux positifs (*False Discovery Rate* - FDR) (FDR). L'avantage de cette méthode est qu'elle évite la mise en place d'une heuristique de recherche dans l'interactome. Cependant, elle utilise les p-values associées aux z-scores, qui ne sont pas toujours disponibles. De plus, il est difficile de trouver des modules d'intérêt dans l'arbre de classification hiérarchique de part sa taille.

Wang *et al.* [181] ont conçu un algorithme pour estimer la sensibilité et sélectivité d'un screen RNAi en faisant une analyse complète de 24 screens RNAi dans la drosophile avec les interactions présentent dans STRING [162]. Bankhead et collègues [7] ont proposé une approche robuste par enrichissement des voies canoniques, la *protein interaction permutation analysis* (PIPA) qui tient compte de l'information reliée aux interactions entre gènes. Leur méthode se concentre sur les pathways ciblés par le screening RNAi analysé.

Murali et collègues [115] ont proposé une méthode d'analyse réseau pour l'identification de nouveaux facteurs de dépendance au virus HIV (HIV Dependency Factors - HDF). Cependant, cette méthode dépend de la connaissance préalable d'autres HDFs à partir de trois études de type RNAi déjà publiées. Une autre méthode a été publiée, basée sur un modèle de Markov caché, appelée *Signaling and Dynamic Regulatory Event Miner* (SDREM) [56]. Cet algorithme intègre l'information de régulation des facteurs de transcriptions et les interactions protéines-protéines avec des données d'expression en séries temporelles spécifiques à différentes conditions expérimentales, et a été enrichi par de nouvelles fonctionnalités d'analyse spécifiques aux données de type RNAi et aux études d'association pan-génomique (genome-wide association study - GWAS).

Plus récemment, Amberkar et Kaderali [4] ont proposé une méthode basée sur l'analyse de données PPI. Ils ont construit un interactome en agrégeant les données de plusieurs bases publiques et ont ensuite utilisé un algorithme *"glouton"* non supervisé pour rechercher des modules fonctionnels basé sur la topologie du réseau d'interaction. Ensuite, ils regardent l'enrichissement pour les hits RNAi dans les sous-modules découverts. Bien qu'intéressante, cette stratégie pourrait potentiellement laisser de côté les gènes qui ne sont pas impliqués dans ces modules hautement connectés. De plus, leur méthode ne tient pas compte du fait que les données d'interactions PPI ne sont pas exhaustives.

Une autre approche a été développée par Wilson et al [183], de façon à identifier les pathways impliqués dans la progression des Leucémies lymphoblastiques aiguës (Acute Lymphoblastic Leukemia- ALL), en utilisant un screening shRNA à l'échelle du génome. Ils ont construit un modèle de réseau en intégrant plusieurs bases de données 'omics, des bases de données d'interactions PPI et de régulation de facteurs de transcription basée sur des données CHiP-Seq. Dans leur modèle les arêtes du réseau et les nœuds sont pondérés par rapport à la qualité des données d'interactions et des protéines associées au shRNA utilisé dans le screening. Cette méthode combine plusieurs types de données, mais pourrait laisser de coté les hits plus faibles (z-score relativement bas mais à la limite de la validité statistique).

Tuncbag et collègues [172] ont développé un logiciel, OmicsIntegrator, qui permet également d'intégrer différent types de données. Il inclut un outil, Forest, qui identifie les sous-réseaux d'intérêt en utilisant une ou plusieurs listes de gènes définies par l'utilisateur (liste de hits RNAi, facteurs de transcriptions, métabolites, etc). Cet outil possède la particularité d'inclure les soi-disant preuves négatives : en effet, un poids négatif peut être appliqué aux nœuds qui peuvent

représenter un biais dans la sélection des sous-réseaux. comme les protéines avec une forte connectivité dans l'interactome (hubs). Une fonctionnalité intéressante est la possibilité d'ajouter des *nœuds de Steiner*, c'est à dire des nœuds qui n'avaient pas été retenus au début mais qui pourraient être biologiquement liés à la condition expérimentale étudiée.

HotNet2 (HotNet Diffusion-oriented subnetworks) [94] est un outil d'analyse réseau conçu initialement pour la découverte de gènes et voies métaboliques ou de signalisation significativement dérégulées ou mutées dans les cancers. Les auteurs ont effectué une analyse pan-génome sur un large jeu de données du Cancer Genome Atlas (TCGA) [11] [31]. Ils ont découvert 16 voies de signalisation contenant des voies déjà connues pour leur implication dans les cancer mais aussi des voies qui n'avait pas été liés à la maladie jusqu'alors. Cette approche réseau permet donc d'identifier des combinaisons de mutations rares. HotNet2 utilise un processus de diffusion de chaleur pour découvrir simultanément les mutations locales des gènes et leur topologie d'interactions. HotNet2 démarre avec les valeurs de fréquences de mutation initiales représentées sous forme de chaleur. Ensuite, il exécute un processus de diffusion permettant de simuler la diffusion de chaleur à travers le réseau. Une matrice d'échange de chaleur est ensuite calculée, permettant d'identifier les sous réseaux "*chauds*", considérés biologiquement intéressants. Il est également possible d'initialiser le programme avec un score de tout type qui sera converti en mesure de chaleur, et donc d'utiliser HotNet2 pour tout type d'analyse génomique. Le test statistique donné apporte la significativité sur le nombre et la taille des sous-réseaux trouvés. Nous avons testé HotNet2 sur des données RNAi et comparé le résultat obtenu avec HTS-Net.

Dutta et collègues *Dutta2016* ont proposé CARD, une application web interactive dédiée spécifiquement aux analyses de données RNAi. CARD comprend un pipeline étendu, qui permet d'aller de la normalisation incluant le contrôle qualité par plaque jusqu'à l'analyse fonctionnelle avancée de type réseau. Chaque étape est dissociée de la précédente. Le programme RNACut [81] constitue la première étape de l'analyse réseau proposée par CARD. Il combine les données de screening et les données d'interactions PPI en comparant la probabilité de la connectivité obtenue d'un hit identifié avec celle espérée par chance. Ce programme permet donc de fixer un seuil sur les z-scores pour les hits pour les étapes suivantes de l'analyse. L'analyse réseau du screening RNAi arrive ensuite. L'algorithme fonctionne en positionnant les gènes dans un réseau d'interactions PPI (interactome) pour l'identification des connections entre les hits candidats. Cette étape permet l'inclusion de gènes connecteurs entre hits, ayant un z-score plus faible mais néanmoins apparaissant pertinent pour l'analyse. L'application CARD permet également de chercher les réseaux obtenus en enrichissement de termes GO (Gene Ontology) et voies canoniques.

Dans ce chapitre, nous présentons une méthode nouvellement publiée [145], une adaptation substantielle et amélioration algorithmique du programme Interactome-Transcriptome Integration (ITI)[51], (Chapitres 3 et 4), pour les analyses réseau de données RNAi. L'algorithme ITI a été profondément repris pour optimiser la détection de réseau en investiguant tous les voisins de chaque graine de réseau [51], au lieu d'étendre l'exploration de réseau à partir d'un voisin particulier. De plus, nous avons intégré les données de régulation dans l'analyse, au lieu de se limiter aux interactions protéines-protéines, ce qui a donné l'algorithme HTS-Net (High Throughput Screening-Network). La plupart des méthodes citées s'appuient uniquement sur les données PPI et ne permettent donc pas de porter des conclusions en ce qui concerne la régulation.

Notre méthode fonctionne par la découverte de sous-réseaux qui travaillent conjointement sur le domaine des interactions protéines-protéines (interactome) et du réseau de régulation (Interactions facteur de transcription-ADN, appelé régulome). En pratique, nous superposons les z-scores RNAi des gènes sur les cartes d'interactions et de régulations de façon séparée. Chacun

11. Site Web : http ://www.tcga.org

des réseaux est analysé pour l'identification de régions à score élevé, qui sont ensuite extraites et reportées sous formes de sous-réseaux. Ensuite, nous intégrons les sous-réseaux obtenus pour former des méta-sous-réseaux qui contiennent à la fois l'information de régulation et les données d'intégration protéines-protéines. Cette approche permet de prioriser les gènes hits d'un screening RNAi en les replaçant dans leur contexte biologique comprenant leurs interacteurs physiques et leur régulateurs. Nous avons fait notre analyse en utilisant l'ensemble des données d'interactions disponibles et avons prouvé sa pertinence sur 3 jeux de données screening RNAi. Pour chaque analyse, nous reportons les marqueurs nouvellement trouvés, leur enrichissement en termes GO, ainsi qu'une comparaison avec les analyses originales. Nous avons également comparé notre méthode aux algorithmes existants. HTS-Net a été rendu disponible à la communauté à travers le serveur Mobyle HTS-Net [12].

5.2 Matériel et Méthodes

L'algorithme HTS-Net fonctionne de la manière suivante : Les données RNAi sont obtenues sous la forme de liste de gènes associées à un z-score (résultats d'une expérience de siRNA ou de shRNA montrant la variation de l'abondance de la population cellulaire mesurée). Deux réseaux sont construits, interactome (relatif aux interactions PPI) et régulome (relatif aux relations de régulations TF-ADN). Les z-scores sont superposés sur chaque réseau séparément. Ensuite, les régions d'intérêt avec des scores élevés sont identifiées et validés statistiquement par comparaison avec une distribution de score aléatoire. Les sous-réseaux identifiés dans l'interactome et le régulome sont ensuite fusionnés par l'utilisation de gènes communs, donnant des méta sous-réseaux.

5.2.1 Assemblage des données réseau

La découverte de sous-réseaux s'appuie sur la qualité de l'interactome et du régulome de départ. Nous avons décidé d'analyser séparément ces deux types de réseaux. Nous les construisons donc séparément par assemblage de plusieurs bases de données publiques d'interactions. un interactome et un régulome (Figure 5.1).

Les bases de données d'interactions protéines-protéines contiennent majoritairement des données générées à haut débit (Y2H, spectrométrie de masse, ou puces à protéines), agrégées à des données générées *à faible débit* (données manuellement assemblées à partir de la littérature). Pour construire notre interactome humain, nous avons téléchargé, analysé et combiné toutes les interactions présentes dans les bases publiques majeures : La base *Database of Interacting Proteins (DIP)* [150], la base *Human Protein Reference Database (HPRD)* [86], la base *Interologous Interaction Database (I2D)* [27], la base IntAct [85], la base *Molecular INTeraction database (MINT)* [95] et la Human ProteinPedia [80]. Un total de 14423 gènes et de 171146 interactions ont été regroupés dans l'interactome.

Les interactions facteurs de transcription-facteurs de transcription (TF-TF)) ont été rassemblées à partir de TRANSFAC, de l'*Atlas of combinatorial transcriptional regulation* (L'Atlas de la régulation transcriptionnelle combinatoire [141]), des interactions prédites par Myšičková et Vingron [118], ainsi que celles prédites par Yu et collègues [190]. Ces interactions ont été incluses dans l'*interactome humain*.

Les bases de données de régulation contiennent des données de régulation prédites des facteurs de transcription (TF), pour des gènes exprimés dans des tissus spécifiques. ITFP (*The Integrated Platform of mammalian Transcription Factors*) reporte 4105 facteurs de transcription potentiels

12. (http ://htsnet.marseille.inserm.fr/

et 69496 couples facteurs de transcription-cibles potentielles dans l'humain [194]. TRED (*The Transcriptional Regulatory Element Database*) reporte 7479 gènes cibles [79]. TRANSFAC, une base de données commerciale, reporte 707 facteurs de transcription et 8712 gènes cibles [101]. ORegAnno, la ressource ouverte pour les annotations des liens de régulations, reporte 465 facteurs de transcriptions et 3853 gènes cibles en utilisant un système d'annotations communautaire [62]. Le système PAZER, une base de données d'information de régulation, reporte 1284 gènes régulés pour 708 facteurs de transcription [135]. Au final, un total de 9755 gènes et 68703 interactions ont été consolidées dans notre carte du régulome humain.

Pour ces deux réseaux (régulome et interactome), l'ensemble des identifiants des protéines a été mappé sur des numéros d'accession standards de type EntrezGene, en utilisant les tables de correspondances disponibles sur le site FTP du NCBI (date de téléchargement du fichier *gene_ info.gz* : 5 novembre 2014). De plus, les facteurs de transcriptions ont été annotés dans les deux réseaux.

5.2.2 Détection de sous-réseaux avec HTS-Net

Les sous-réseaux sont associés à un score dans HTS-Net. Ce score est formellement défini comme étant la moyenne des *z-scores* pour l'ensemble des gènes appartenant à ce sous-réseau, comme suit :

$$S(sn) = \sum g \in snz - score(g) \tag{5.1}$$

sn étant le sous réseau analysé courant, $S(sn)$ son score, g le gène courant, et $z - score(g)$ le score normalisé du screening RNAi du gène courant. Pour chaque jeu de données étudié, l'ensemble des gènes a été investigué avec HTS-Net (Figure 5.1). Pour chaque réseau, interactome ou régulome (Figure5.1A-I), chaque nœud est testé comme étant une graine putative (Figure 5.1A-II - 5.1B-II). Pour développer un sous réseau autour de ces graines, nous avons décidé d'utiliser un algorithme identique pour chacun des réseaux. Dans le cas du régulome, même si seulement une partie des gènes agit en tant que régulateur (les facteurs de transcription), nous avons choisi de tester chaque gène, y compris les gènes régulés, comme graine putative. Cela permet la détection de sous réseaux avec des gènes ayant un *z-score* élevé régulés par un facteur de transcription qui pourrait être caractérisé par un score plus faible. D'un autre coté, cela permet l'identification de gènes caractérisées par un *z-score* plutôt faible mais tous régulés par un facteur de transcription caractérisé par un score élevé.

A partir de l'ensemble des graines potentielles, les voisins sont ensuite testés indépendamment. Cela signifie que chaque voisin de la première couche est testé indépendamment pour l'amélioration du score et marqué comme étant conservé ou supprimé dans le sous-réseau courant. Un nœud est conservé si son inclusion permet l'augmentation du score du sous-réseau au delà d'un seuil d'amélioration de score *th*.

Ensuite, les voisins qui améliorent le score sont ajoutés simultanément au sous-réseau et le nouveau score du sous réseau est calculé. Puis, les voisins des nœuds nouvellement ajoutés sont explorés récursivement, couche après couche *en largeur d'abord*. C'est une forte amélioration par rapport aux premières versions d'ITI où les nœuds étaient explorés *en profondeur d'abord*, ce qui provoquait un biais sur le résultat final par rapport à l'ordre de l'exploration des voisins.

Le seuil d'amélioration du score *th* a un impact limité sur le nombre de sous-réseaux et de gènes détectés (Figures 5.2A et 5.2B). En faisant varier *th* entre 0 (moins strict) jusqu'à 0.1 (plus strict), le nombre de sous-réseaux passe de 60 à 110. A première vue, c'est un comportement contraire à ce qui est espéré intuitivement. Cependant, une valeur de *th* plus faible diminue le

score des sous-réseaux produits, qui sont ensuite filtrés avec la validation statistique. Le même phénomène est observé avec les gènes, dont le nombre varie entre 160 jusqu'à 190 pour la même variation de th sur l'intervalle $[0-0.1]$. Pour les analyses, nous avons utilisé un seuil de $th = 0.05$, qui est un bon compromis entra obtenir un nombre de sous-réseau trop grand pour pouvoir le gérer et une approche trop stricte.

Dans le jeu de sous-réseaux identifiés, il peut y avoir beaucoup de nœuds communs. Ce phénomène est provoqué parce que chaque gène de chaque réseau est considéré comme étant une graine potentielle. Deux graines distantes peuvent interagir avec un grand nombre de gènes communs aux *z-score* élevés. Avant de faire la validation statistique des sous-réseaux, les sous-réseaux redondants devraient être filtrés, en tenant compte de leur tailles et scores. Nous comparons les sous-réseaux par paires, disons A et B, de la manière suivante. Nous calculons la proportion de nœuds dans le sous réseau A présents dans le sous réseau B (recouvrement 1) et vice versa, la proportion de nœuds du sous réseau B également présent dans le sous réseau A (recouvrement 2). Si les deux recouvrements sont au delà d'un seuil spécifique (80% par défaut), le sous réseau avec le score le plus faible est effacé. Si seul un recouvrement est au delà du seuil, il est vraisemblable qu'il y ait une différence importante de taille entre les sous-réseaux comparés. Dans ce cas, si le sous réseau plus petit a également un score plus faible, il est éliminé. Sinon, les deux sous-réseaux sont conservés. Nous avons posé cette condition, qui est une évolution par rapport à la version de l'algorithme ITI, pour éviter de supprimer des sous-réseaux de grande taille, et donc de l'information potentiellement importante. Après cette étape de filtrage, le sous-réseaux restants sont considérés pour le passage à la validation statistique.

5.2.3 Validation statistique

Pour mesurer la signifiance statistique des sous-réseaux détectés, leurs scores sont comparés avec 3 distributions aléatoires de sous-réseaux, suivant la méthode décrite par Garcia et collègues [51]. La première distribution est obtenue par mélange des *z-scores* du screen RNAi avant détection des sous-réseaux. Ensuite, une p-value pour les sous réseaux est obtenue par comparaison de cette distribution avec les scores des sous-réseaux. C'est la p-value I. Cette p-value permet de mesurer le lien statistique entre les z-scores RNAi et les sous-réseaux, et montre l'intérêt d'étudier un screening RNAi avec l'algorithme HTS-Net. La p-value II est obtenue par détection des sous-réseaux sur un réseau aléatoire. L'ensemble des arêtes du réseau étudié (interactome ou regulome) est mélangé en gardant le degré (nombre de connexions) de chaque nœud. Cette seconde p-value renforce le lien entre les sous-réseaux et de réseau global. La p-value III démontre la valeur statistique de l'algorithme par comparaison des sous-réseaux avec des sous-réseaux de la même taille choisis aléatoirement, sans considérer la topologie du réseau étudié ou les *z-scores*.

Ces 3 distributions sont modélisées avec une mixture de 2 gaussiennes (Figure 5.3) à l'aide du package R *nor1mix*. La sélection des sou-réseaux est ensuite faite par seuillage des p-values à un niveau statistiquement significatif, après correction de Bonferroni pour les tests multiples. Pour l'étude de l'interactome et du régulome, seuls les sous-réseaux statistiquement significatifs pour les 3 p-values sont retenus. Les seuils retenus pour les 3 p-values sont reportés Table 5.2.3. Nous avons choisi ces seuils de façon à obtenir un ordre de grandeur de nombre de nous-réseau similaire aux 3 jeux de données étudiés. A titre d'exemple, une p-value de 1.10^{-2} pour le jeu de données Chia a donné 766 sous réseaux pour le régulome. La figure 5.2C montre que le nombre de sous-réseaux s'accroît exponentiellement quand le seuil de p-value II est augmenté. On obtient des résultats similaires pour les p-values de type I et III et sur d'autres jeux de données. Fixer des p-values sur une valeur fortement significative de 1.10^{-4} nous a donné un nombre de sous-réseau raisonnable pour le jeu de données Chia. Les différences du nombre de sous-réseaux trouvés sont

Jeu de données	Seuil th	P-valeur de type I, II et III. (interactome, regulome)	Nombre de sous-réseaux (inte
Chia	0.05	5.10^{-3} ; 1.0^{-2}	134 ; 130
Tai	0.05	5.10^{-3}; 5.10^{-3}	32 ; 28
Wolf	0.05	5.10^{-2}; 1.10^{-2}	178 ; 54

TABLE 5.1 – Choix des paramètres et scores obtenus pour les 3 analyses. Les colonnes décrivent les valeurs prises par les paramètres suivants : th est le seuil d'accroissement minimum du score à l'ajout d'un gène dans un sous réseau. Il a été fixé à $th = 0.05$ pour les 3 jeux de données analysés. Les P-values de type I, II et II ont été décrites section 5.2.3. La même valeur a été utilisée pour les 3 p-valeurs indépendamment du régulome et de l'interactome. L'intervalle des scores interactome et régulome correspond aux scores obtenues après validation statistique. L'intervalle présente les scores minimum et maximum pour chaque analyse interactome et régulome. n est le seuil du nombre de nœuds communs utilisé pour intégrer les sous réseaux régulome et interactome. Entre parenthèses est donné le nombre correspondant de méta-sous-réseaux.

dues aux différences entre les topologies des jeux de données et aux différences de corrélation entre les *z-scores* et les réseaux étudiés. Comme dans tout algorithme basé sur une heuristique, les paramètres sont choisis et optimisés à travers plusieurs tests.

5.2.4 Intégration des sous-réseau interactome et regulome

De façon à identifier les pathways de régulation incluant plusieurs niveaux de régulation (par exemple, découvrir un sous-réseaux dans le regulome qui régule plusieurs sous-réseaux dans l'interactome), HTS-Net inclus combine les sous-réseaux identifiés dans l'interactome avec ceux identifiés dans le régulome en trouvant les protéines communes entre ces deux types de sous)-réseaux. Pour chaque jeu de données analysé, nous examinons individuellement tous les sous réseaux regulome et interactome et connectons les sous-réseaux présentant un recouvrement d'au moins n nœuds communs pour générer les *meta-sous-réseaux*. La valeur n choisie dépend de la convergence des sous-réseaux détectés et est spécifique à chaque jeu de données.

Nous avons étudié la variation du nombre de sous-réseaux obtenus en fonction de la valeur n choisie (Figure 5.2 D). Ce nombre est grandement affecté par le seuil n de nombre de protéines communes et varie d'environ 200 méta-sous-réseaux obtenus pour $n = 1$ protéines communes à 0 pour un seuil de $n = 6$ protéines communes. Cela révèle la nature modulaire des sous-réseaux régulome et interactome, qui ne sont liés que par un nombre très limité de protéines en commun. Pour le cas pratique du jeu de données Chia, pour lequel nous avons représenté l'effectif de méta-sous-réseaux obtenus Figure 5.2 D, nous observons une décroissance rapide du nombre de méta-sous-réseaux pour une valeur de 4 protéines, ce qui donne un nombre relativement gérable de méta-sous-réseaux de 12. La table 5.2.3 donne la valeur des paramètres choisis pour chaque analyse.

5.2.5 Rapport d'analyse et base de données de sous-réseaux générées par HTS-Net

La nature hétérogène des sources de donnés, la problématique de la conversion des numéro d'accession des gènes, et la complexité de la structure des données générées par HTS-Net nécessitent un système avancé de visualisation de façon à interpréter la nature des sous-réseaux. En particulier, plusieurs type d'interactions et de types de sous réseaux (interactome et régulome) doivent être visualisés de façon particulière et comparé à la biologie connue. Le pipeline HTS-Net

contient un système de génération de rapports permettant la visualisation de réseaux et de nature d'interactions génétiques de différent types (PPI et régulation de gènes). Tous les sous-réseaux détectés sont rendus graphiquement en utilisant le package GraphViz (AT&T Research, Austin, TX USA) et le style du site web a été développé avec KickStart 0.99 en HTML5 et CSS3. En utilisant cette interface, tous les sous-réseaux peuvent être comparés et analysés. Tous les gènes sont également listés et connectés à leur sous-réseaux respectifs. Les liens vers la page correspondante du site Entrez Gene du NCBI et l'analyse en enrichissement GO sont également fournis. Une analyse avec Cytoscape est directement faisable car tous les sous-réseaux sont stockés dans des fichiers au format standard Network Nested Format [13]. Toutes les données générées sont disponibles sur fichier plats à partir du site web compagnon de la publication HTS-Net [14].

5.2.6 Analyse d'enrichissement en termes Gene Ontology (GO)

.

Les gènes ont été annotés avec l'ontologie Gene Ontology (GO). L'analyse en enrichissement GO a été faite pour les gènes identifiés dans les sous-réseaux détectés pour chaque jeu de données. Nous avons employé un test statistique standard de type hypergéométrique associé à une correction pour tests multiple de type Benjamini-Hotchberg (BH). Nous avons fixé la valeur du False Discovery Rate de la correction à 5%. L'enrichissement en termes GO des l'ensemble des sous-réseaux pour les jeux de données testés. L'enrichissement GO obtenu est disponible pour chaque réseau détecté pour chaque jeu de donnée et récapitulé dans un fichier unique sur le site web compagnon de la publication [15]. Les résultats obtenus en enrichissement Go sont discutés dans la section résultat pour les trois jeux de données analysés.

5.2.7 Sources de données publiques

De façon à évaluer la capacité de notre pipeline HTS-Net à identifier des réseaux de gènes d'intérêt, nous avons réanalysés trois screening RNAi, à savoir, le jeu de données publié par Chia et collègues [34], le jeu de données publié par Wolf et collègues [184], et le jeu de données publié par Tai et collègues [164]. Chia et collègues [34] ont mis en place un screening RNAi pour déterminer les gènes signant pour les cellules souches embryonnaires (hESC). Ils ont d'abord conçu un modèle de lignée cellulaire, H1 hESC ([170]) en introduisant un reporteur GFP activé par la région amont régulatrice de POU5F1, est un marqueur de cellules hESC non-différenciées. Le modèle de lignée cellulaire a été ensuite filtrée sur une librairie de 21212 gènes humains en réplication. La moyenne des *z-scores* obtenus pour la réduction de fluorescence GFP (*Fluorescence reduction* - FaV) et réduction du nombre de noyaux (*Nuclei number reduction*, Nav) ont été calculés. Les auteurs ont identifiés tous les gènes ayant un score $Fav > 2$ comme étant un marqueur potentiel de l'identité des cellules hESC, ce qui a donné un total de 566 gènes. Ils ont ensuite fait un screening secondaire sur le top 200 gènes. Le second jeu de données analysé concerne une étude de Wolf et collègues ([184]) ou a été testé la possibilité de population de cellules souches cancéreuses (CSC) définie par CD44+CD24-/Low de former des mammosphères (un marqueur fonctionnel permettant d'identifier la propriété d'auto renouvellement des CSCs), après un shRNAi (*short hairpin interference assay*) pour identifier le régulateurs des CSC sur une librairie de 5045 gènes. Ils ont comparé cette analyse avec les mêmes cellules cultivées sous

13. http://manual.cytoscape.org/en/3.4.0/Nested_Networks.html
14. http://cibi.marseille.inserm.fr/htsnet
15. http ://htsnet.marseille.inserm.fr/supplementary-data.html

différentes conditions (conditions de culture permettant la différentiation cellulaire) pour éliminer les shRNAi ne dépendant pas des effets liés spécifiquement aux cellules CSC.

Wolf et collègues ont initialement sélectionné 1051 gènes dont le *z-score* adhérent a une p-valeur de < 0.01 pour trouver les gènes bloquant la croissance cellulaire. Ensuite, ils ont fait une analyse d'enrichissement sur un jeu de 392 gènes liés à la formation des mammosphère, pour filtrer sur une p-valeur < 0.01 sur le *z-score* mammosphère et sur une p-valeur > 0.1 sur le *z-score* adhérent, ceci permettant de sélectionner les gènes signant pour les CSC. Ils ont ensuite choisi de conduire l'analyse sur les gènes ATG4A et ATG4B ayant des *z-scores* respectifs de 0.51 et 0.11.

Le dernier jeu de données analysé est celui de Tai et collègues [164], qui a recherché les protéines hôtes impliquée dans la réplication du virus de l'hépatite C (Hepaticis C virus - HCV) en utilisant le réplicon subgénomique Hu7/Rep-Feo. Ce réplicon encode les protéines HCV non structurelle qui set de médiateur pour la transcription virale ainsi qu'une protéine de fusion de la luciole utilisée pour la quantification du signal. Cette lignée cellulaire a été filtrée sur 21094 siRNAs ciblant le génome humain avec une réplication.

Le rapport web disponible sur le site web compagnon montre comment les différents gènes et les sous-réseaux associés détectés par HTS-Net se positionnent par rapport aux gènes et z-scores obtenus dans l'analyse originelle. La figure 5.4 montre les diagrammes de Venn obtenus entre les analyses originelles et les gènes identifiés par HTS-Net, et sont commentés dans la section Résultats.

5.2.8 Comparaison des résultats obtenus avec HTS-Net, HotNet2, CARD et la méthode de Gonzalez&Zimmer

Nous avons comparé HTS-Net avec trois autres programmes, à savoir HotNet2 [94], CARD [45] et la méthode réseau proposée par Gonzalez & Zimmer [59] sur le jeu de données Tai. Nous avons choisi ce jeu de données, à savoir Tai, car c'est ce jeu qui a été testé par Gonzalez & Zimmer, qui ne proposent pas de code source avec leur publication mais qui ont publié leur résultats, ce qui donne une bonne base de comparaison. HotNet2 est distribué sous forme de code source. Il a donc été installé sur une machine de calcul avec la collaboration de la plateforme de calcul du CRCM [16] après compilation et installation des dépendances. Nous l'avons ensuite exécuté avec le réseau HPRD [86] comme référence après avoir généré les matrices d'influence correspondantes à ce réseau (matrices qui mesurent le valeur de propagation de la chaleur à travers le réseau). Le réseau HPRD a été choisi car il est distribué (et donc formatté) avec HotNet2 de façon officielle avec HotNet2. HotNet2 est initialisé avec deux paramètres, Beta et Delta. Beta est un paramètre de diffusion de chaleur qui permet de fixer comment la valeur d'un nœud interagit avec ses voisins. Nous l'avons laissé à sa valeur par défaut. Delta représente le poids des arêtes et a un impact sur la taille des sous-réseaux détectés. Après exécution, HotNet2 propose plusieurs valeurs de Delta sur l'intervalle $[2, 44.10^{-3} - 4.92.10^{-3}]$, et nous avons choisi $Delta = 2.44.10^{-3}$, qui nous a donné un nombre raisonnable de sous-réseaux à interpréter (288 sous-réseaux, p=0.46). Nous avons ensuite utilisé la fonction d'exploration pour obtenir une liste de gènes comparable avec les autres programmes testés.

CARD est accessible en tant que service web. Nous avons donc créé un compte et formaté le jeux de données Tai au format CARD. Chrome ®44 est officiellement recommandé, mais nous avons ou le faire tourner sans problème notable sous la version plus récente numérotée 55. Sous CARD, il faut rentrer les données mesurées pour chaque plaque, et il n'y a pas moyen de spécifier

16. http ://disc.marseille.inserm.fr

que nous avons déjà une version normalisée des données à notre disposition. Nous avons donc donné des données aléatoires pour les identifiants de plaque et spécifié 'no normalisation' dans les paramètres d'exécution, ce qui nous a permis de préserver la normalisation initiale. Ensuite, nous avons déterminé le niveau optimal de signification statistique pour les *z-score*. La valeur maximum pour le $log(p-valeur)$ apparaît pour le $z-score = 2.786$ ($p-valeur = 0.004$). Nous avons ensuite fait l'analyse réseau avec ce paramètre. Nous avons sélectionné tous les réseaux disponibles sous CARD, à savoir HPRD, Bind et BioGRID. Enfin, nous avons téléchargé la liste de gènes résultat sous forme de fichier CSV pour comparaison avec les autres pipelines.

En ce qui concerne la méthode Gonzalez & Zimmer, le code n'est pas disponible, nous avons donc réutilisé les résultats sur les données Tai données dans l'article.

5.3 Analyse et Résultats

5.3.1 Un nouveau pipeline pour comprendre les interactions et les phénomènes de régulations impactés dans un screen HTS

Nous proposons un nouveau pipeline, HTS-Net, pour l'analyse réseau de screening HTS à grande échelle, adaptés à plusieurs situations d'analyse. Le pipeline est maintenu et accessible à travers un portail Mobyle [122] déployé localement [17]. Un tutoriel est disponible à partir de la même adresse. Une analyse peut être conduite une fois les données normalisées, les réplicats combinés et stockés dans un fichier au format d'entrée HTS-Net. Le format est très simple, c'est un fichier délimité par des tabulations à deux colonnes. La première contient des identifiants de gènes (Type Entrez Gene ID), et la seconde colonne les *z-scores* associés du screening. Si plusieurs réplicats existent, ils doivent être moyennés ou combinés en un score unique, puisque HTS-Net n'est pas conçu pour travailler avec plusieurs scores simultanément. Une information additionnelle doit être donnée, à savoir l'organisme étudié et les paramètres HTS-Net, à savoir les p-valeur pour les seuils de score. Après avoir tourné, HTS-Net génère une archive contenant les rapports d'analyse en HTML, format compatible avec tout navigateur Web standard.

De façon à comprendre les résultats obtenus avec le pipeline HTS-Net, nous avons comparé les gènes identifiés par HTS-Net avec les gènes publiés originellement. Nous avons également produit des analyses en enrichissement GO pour les gènes communs entre les gènes identifiés par HTS-Net et les gènes spécifiquement identifiés avec une des deux méthodes (Les résultats sont sur le site web compagnon).

La figure 5.5 montre la distribution des *z-scores* pour le jeu de données Tai avec le seuil retenu pour les gènes 'Hits', superposés à la distribution des scores des gènes retenus par HTS-Net. Ces distributions montrent que HTS-Net est moins stringent avec les gènes qui ont un score bas mais connecté à des sous-réseaux pertinents et qui auraient été rejetés par une analyse classique basée sur une statistique tenant compte uniquement du *z-score*.

5.3.2 Analyse réseau des déterminant de la différentiation des cellules souches embryonnaires humaines [34]

Nous avons refait l'analyse de Chia et collègues [34] avec la valeur Fav, telle que dans l'article original, et l'avons soumis à HTS-Net sous forme de score. Le seuil minimal d'accroissement du score th a été fixé à 0.05. 21023 gènes ont été mappées sur l'interactome et sur le régulome, de façon à détecter les sous-réseaux propres à la régulation de la différentiation cellulaires. Nous

17. http://htsnet.marseille.inserm.fr

avons appliqué l'algorithme HTS-Net avec un seuil sur les p-valeurs de 5.10^{-3} pour l'analyse interactome et un seuil de 1.10^{-2} pour le régulome de façon à obtenir un nombre de sous-réseaux équivalent tout en gardant une bonne significance statistique pour les deux réseaux. HTS-Net a été configuré pour détecter les gènes avec des scores négatifs avec une valeur absolue élevée, comme dans la publication originale. Nous avons obtenu 134 sous-réseaux interactome et 130 réseaux régulome qui ont ensuite été intégrés en 12 méta-sous-réseaux après intégration sur 4 noeuds communs. Ces sous-réseaux contiennent 294 gènes pour l'interacome, 330 pour le régulome, et 113 gènes dans les méta-sous-réseaux. Cela donne 32 gènes communs entre les 113 gènes identifiés par HTS-Net et les 126 gènes par Chia (Figure 5.4).

Chia s'est concentré sur les régulateurs transcriptionnels PRDM14, NFRKB and YAP1. tous étaient présents dans les sous-réseaux a score élevés. PDRM14 a été retrouvé dans 7 sous-réseaux régulome. Le meilleur sous-réseaux régulome contenant PRDM14 possède un score de 2.412. NFRKB a été détecté dans 48 sous réseaux interactome, 3 sous réseaux regulome et retenu dans 3 méta-sous-réseaux. Le meilleur sous-réseau contenant NFRKB a un score de 3.11. YAP1 a été détecté dans 94 sous réseaux interactome et 12 méta-sous-réseaux. Le meilleur sous réseau contenant YAP1 a un score de 3.54.

Grâce au pipeline HTS-Net, nous avons pu identifier d'autres gènes d'intérêt. Plusieurs gènes du 2^{e} complexe répressif Polycomb (PRC2 *Polycomb Repressive Complex 2* ont été identifiés dans des sous-réseaux à score élevé par HTS-Net et laissés de côté par la méthode de Chia. Ces régulateurs de chromatine sont connus pour leur fonction dans l'établissement et le maintien de la mémoire épigénétique durant le développement. Le gène EZH2, connu pour jouer un rôle dans l'auto-renouvellement des cellules souches en maintenant l'identité cellulaire ([24]) possède un z-score de -1.63 et n'a pas été retenu par Chia. Il a cependant été détecté par HTS-Net et intégré à 2 sous-réseaux interactome et 5 réseaux régulome. Il est présent dans le sous-réseau `int-snw-2146` en tant que graine, interagissant avec un autre membre du complexe PRC2, SUZ12, ainsi qu'avec un nombre de gènes détectés par Chia (Figure 5.4A). SUZ12 est impliqué dans la différentiation cellulaire et a été détecté comme étant impliqué dans la progression d'un grand nombre de cancers ([36]. Il a également été détecté dans les sous-réseaux `int-snw-121536` ayant AEBP2 comme graine, ce gène ayant également un score plus bas que le seuil Chia. AEBP2 est requis pour optimiser l'activité enzymatique du complexe PRC2 ([179]). Dans le sous-réseau `int-snw-54857`, plusieurs gènes identifiés par Chia sont présents (YAP1, NFRKB) mais la graine de sous-réseau GDPD2 et le gène HEXDC ont été spécifiquement identifiés par HTS-Net. GDP2 code pour une enzyme qui pourrais jouer un rôle dans la différentiation des ostéoblastes ([37]). La liste complète des paramètres utilisés pour initialiser HTS-Net sur cette analyse ont été reportés Table 5.2.3.

L'analyse en enrichissement GO (voir les données supplémentaires) pour les gènes trouvés avec HTS-Net et les gènes identifiées dans le screening secondaire de Chia, montrent plusieurs termes GO spécifiquement enrichis dans l'analyse HTS-Net (voir fichiers supplémentaires sur le site compagnon). Les processus généraux ont été détectés par les deux approches, tels que 'gene expression' et 'protein binding'. Par contre, des processus spécifiques ont été détectés comme étant hautement enrichies par HTS-Net, tels que 'viral process' (16 gènes - 9 trouvés par Chia et collègues). Les gènes localisés dans le cytosol (38 gènes - 26 trouvés par Chia et collègues) auxquels s'ajoutent les gènes du complexe médiateur (5 gènes et aucun détectés par Chia). HTS-Net a permis également l'identification de gènes enrichies pour le terme 'poly(A) RNA biending' (27 gènes, seules 17 détectées par Chia). D'autres enrichissement spécifiques détectés dans le screening secondaire Chia ont été détectés à la limite de la significativité statistique avec HTS-Net comme 'oligodendrocyte development', 'cellular response to stimulus', ainsi que les processus 'negative regulation of transcription' et 'negative regulation of translational initiation in response to stress'.

5.3.3 Analyse réseau de la formation des mammosphères dans les cellules cancéreuses dans le sein [184]

L'analyse Wolf [184] a été faite sur 5045 gènes que nous avons réanalysés avec HTS-Net. Ces gènes, reportés dans le matériel supplémentaire de la publication Wolf, ont été convertis en 4948 numéros d'accession Entrez Gene pour assurer la compatibilité HTS-Net. Nous avons mis en place une analyse intégrée avec à la fois les données du screening obtenus avec les cellules enrichies en mammosphères et les cellules enrichies en cellules adhérentes. De façon à différentier les sous-réseaux affectés dans les deux conditions, nous avons calculés les ratios mammosphère/adherent pour chaque KO et analysé les données obtenues avec un seuil minimal d'amélioration du seuil de $th = 0.05$ pour les données interactome et $the = 0.01$ pour le regulome. La valeur absolue du ratio est retenue de façon à détecter à la fois les sous-réseaux spécifiques au mammosphères et ceux spécifiques aux cellules adhérentes en un run HTS-Net unique.

HTS-Net a identifié 239 gènes, organisés en 136 méta-sous-réseaux, 178 sous-réseaux interactome et 54 sous-réseaux régulome, et a permis une analyse intégrées des cellules enrichies en mammosphères et des cellules adhérentes. Le diagramme de Venn (Figure 5.4) révèle que nous avons surtout détectés des cellules adhérentes, car le *z-score* des cellules adhérentes tend a être plus élévé.

Le gène ATG4A n'a pas été détecté par HTS-Net, mais ATG4B a été identifié comme graine d'un sous réseau globalement sur-exprimé dans les cellules mammosphères (`int-snw-23192`, score =0.993) tandis que ATG4B est lui, sous exprimé dans ces cellules. Bien qu'ATG4B soit de degré élevé (le nombre de ses connections est de 159 dans notre interactome), il a été ajouté à un seul sous réseau, le sous réseau interactome `int-snw-23192`, connecté au sous-réseau régulome `reg-snw-2958` pour former une structure régulatrice retrouvée dans le méta-sous-réseau `meta-reg-snw-23192` (Figure 5.6B) et dans une structure plus large, `meta-reg-2958`, composée de 31 sous-réseaux. Le méta-sous-réseau `meta-reg-snw-23192` contient 18 gènes, parmi lesquelles TP53, un nombre de gènes modulant la formation des microtubules (TUBA1C et ACTB), et des gènes participant à l'élongation (EEF2). D'autres protéines inclus dans les ribosomes sont présentes dans ce sous-réseau.

Nous avons trouvé qu'un grand nombre des 239 gènes identifiées par HTS-Net avaient été détectées adhérentes, alors que seulement 5 ont été identifiées comment étant enrichies dans les mammosphères. Cette observation s'explique par le fait que nous avons configuré le run HTS-Net pour détecter les ratios $log_2(z - score_{Mammosphre}/z - score_{Adherent})$ les plus larges possibles en valeur absolue, et que la plupart des gènes tendent à avoir un score plus large pour les cellules adhérentes. Un nombre d'autres fonctions reliées au cycle cellulaire et à la différentiation cellulaire ont également été détectées spécifiquement par HTS-Net, avec par exemple 'mitotic cell cycle' (32/145 gènes détectés avec HTS-Net et 15/515 dans la publication originale). 'DNA damage response, signal transduction by p53 class mediator resulting in cell cycle arrest' (13/68 gènes détectés avec HTS-Net et seulement 1/68 dans la publication originale).

5.3.4 Analyse réseau des co-facteurs hôtes de réplication du virus de l'hépatite C [164]

Deux analyses ont été faite précédemment avec ce jeu de données : Celle de la publication originale de la part de Tai et collègues et une seconde analyse, de type réseau, faite par Gonzalez & Zimmer [59], qui nous a donné l'opportunité de faire des comparaisons et de comprendre les avantages et limitations de chaque approche. Nous avons fait une analyse avec HTS-Net, reportée dans cette section.

HTS-Net a mappé 17692 gènes de l'analyse sur des identifiants Entrez Gene, 12906 sur l'interactome humain, et 8990 sur le régulome humain. 141 gènes ont été retenus comme étant significatifs à la fois dans l'interactome et le régulome et ont été trouvés dans respectivement 32 et 28 sous-réseaux correspondants et 55 méta-sous-réseaux. Des 141 gènes identifiées avec HTS-Net, 2" sont communs avec ceux identifiés par Tai.

Toutes les études ont identifié gènes dans le COPI. Ces gènes ont été classifiés comme étant significatifs dans le screening et ont été également favorisés dans l'analyse clustering de Gonzalez & Zimmer. Nous avons identifié un nombre de gènes qui sont par de ce complexe avec HTS-Net, incluant COPA, COPB1, COPB2, COPG1 et COPZ1. l'activité COPI et été confirmée comme étént essentielle dans la réplication HCV par Tai et collègues. CDC42, ayant un rang élevé dans le screening, et attaché à ce complexe, a été également détecté avec HTS-Net, Tai et Gonzalez.

Certains gènes ont été spécifiquement identifiés par HTS-Net. Parmi celles-ci, nous avons découvert BRCA2, un facteur de transcription connu pour son apport dans la stabilité du génome et comme gène de prédisposition dans les cancers du sein et des ovaires. Un grand nombre de gènes reliée à la machinerie du ribosome ont également été identifiées par HTS-Net (RPS11, RPS13 et autres).

Nous avons concentré notre analyse sur le sous-réseau `meta-reg-snw-7133`. Ce méta-sous-réseau résulte de l'agrégation de 10 sous-réseau et contient un nombre d'élément régulant. Son centre est matérialisé par le sous-réseau interactome `int-snw-7133` lui-même régulé par EIF3M, présent dans trous sous-réseau régulome. ce facteur de transcription est nécessaire à la synthèse de protéines mais est également connue pour être un homologie de Tango7 dans la drosophile, implique dans le transport Golgo ([8]). Bien qu'ayant un score relativement haut, ces gènes n'ont pas été détectés originellement par Tai et collègues parmi les gènes les mieux classés. Comme facteur de transcription, ce gènes pourrais représenter une cible HCV très valable.

Tai n'a pas retenu PIP5K1A, qui a un score moyennement significatif (-1.820). Ceci dit, il appartient à plusieurs méta-sous-réseaux ayant un score élevé (incluant le méta-sous-réseau ayant le score le plus élevé de l'étude = 4.751). Nous avons trouvé cette kinase dans un seuil sous-réseau interactome, interagissant avec COPA, PSMD1 et la machinerie ribosomale (Figure 5.6C) et 4 méta-sous-réseaux.

HAC3D (PTPLAD1, BIND1) a également été spécifiquement trouvé par HTS-Net. C'est un régulateur de la réplication virale [163]. Tai n'a pas identifié ce gène car il a un score plus faible (0.875), mais son mécanisme d'action a été révélé par HTS-Net. Ce gène est présent dans deux méta-sous-réseaux en tant que graine du sous-réseau de régulation `reg-snw-51495`, qui a un score élevé (4.504).

Nous avons également regardé l'enrichissement GO des sous-réseaux HTS-Net. Parmi ceux-ci, `int-snw-4381` est enrichi de façon significative pour les termes 'viral transcription' ($p = 2.07.10^{-27}$) et 'viral life cycle' ($p = 2.95.10^{-27}$. l'enrichissement GO des tous les gènes identifiés avec HTS-Net est donné dans le matériel supplémentaire sur le site compagnon.

5.3.5 Comparaison de HTS-Net avec les algorithmes existants

Pour comprendre les différences et points communs entre HTS-Net et les approches existantes, nous avons fait une analyse complète du jeu de données Tai avec 3 autres programmes. Le programme d'analyse de Gonzalez & Zimmer [59], spécifiquement conçu pour les analyses RNAi, le programme d'analyse réseau à usage général HotNet2 ([94]), et la suite logicielle CARD ([45]), spécialisée en analyse de données siRNA (de l'analyse bas niveau jusqu'à l'interprétation fonctionnelle).

Nous avons spécifiquement utilisé le jeu de données Tai pour la comparaison des différents

programmes d'analyse car l'analyse de Gonzalez & Zimmer n'est disponible que pour ces données. La liste des cibles identifiées par chaque méthode est donnée sur le site web compagnon de HTS-Net en tant que matériel supplémentaire. Les résultats clés de la comparaison ont été données Table 5.3.5. Les diagrammes de Venn comparant les jeux de gènes obtenus avec les différents algorithmes sont donnés figure 5.7.

HTS-Net

Le temps d'exécution est de 20 minutes en faisant tourner la version parallèle du code sur notre cluster de calcul. La version mise à disposition du public tourne en 1h48. HTS-Net a permis l'identification de PIP5K1A en tant que hit. Ni Tai ni Gonzalez n'ont trouvé ce gène dans leur analyse. De façon intéressante, il a été découvert comme étant impliqué dans les mécanismes d'infection virale du virus HCV et dans sa réplication. Les autres éléments spécifiques sont détaillés section 5.3.4.

Méthode de Gonzalez & Zimmer

La méthode de Gonzalez et Zimmer a détecté 128 gènes. Le nombre de sous-réseaux n'est pas décrit, car les auteurs les ont combinés en un sous-réseau unique. Le coût calculatoire n'est pas spécifié. Gonzalez a détecté 23 gènes initialement détectés par Tai, tandis que 46 gènes ont également été détectés par HTS-Net (Figure 5.7), montrant une topologie de résultats plus proche avec HTS-Net, qui possède également 23 gènes communs avec Tai. Par exemple, EIF3M possède un score relativement élevé (4,41) mais n'avait pas été identifié initialement par Tai *et al.*. Hors, il a été détecté par Gonzalez et HTS-Net. En tant que facteur de transcription, il représente une cible très intéressante pour les mécanismes d'invasion du virus HCV.

CARD

En utilisant spécifiquement les modules RNAiCut (détection des gènes cibles) et analyse Réseau de CARD, nous avons obtenu 119 gènes (sous 3 sous-réseaux) pour des réseaux de gènes cibles ($z - score > 1.0$) et 493 avec les non-cibles ($z - score < 1.0$). Le système permet de télécharger les réseaux de gènes cibles obtenus, ce qui a permis la mise en place de la comparaison avec les autres méthodes explorées dans ce chapitre. Le temps d'exécution a été d'environ 40 minutes pour RNAiCut et de quelques secondes pour la partie analyse réseau. L'utilisation de CARD sur le jeu de données Tai a révélé 19 gènes déjà détectés par Tai *et al.* et 61 gènes ont été trouvés commun avec les résultats données par HTS-Net (Figure 5.7B). Parmi ceux-ci ont été identifiés NCOA6, un co-activateur transcriptionnel qui peut interagir avec les récepteurs hormonaux pour améliorer les fonction d'activation transcriptionnelle. COPA, COPB1, COPB2, COPZ1 (les protéines de la sous unité alpha du complexe coatomère, de la sous unité béta 1, de la sous unité béta 2 et zéta 1) ont été identifiées par CARD. Tai a initialement détecté ces protéines dans le screening primaire, et leur rôle important dans le trafic des protéines a été détaillé dans leur publication. Cependant, HTS-Net et la méthode de Gonzalez ont également été parfaitement capables de détecter ces gènes importants. SERPB1 a été identifié (également par HTS-Net) de concert avec les protéines ribosomales et EIF3M, connu pour interagir avec le point d'entrée du ribosome du virus de l'héptatite C (Internal Ribosome entry Site - IRES) [99]. RAN (RAS-related Nuclear protein) a été détecté par l'ensemble des méthodes avec l'exception de HotNet2. C'est une petite protéine à domaine GTP appartement à la superfamille RAS, qui est essentielle pour la translocation de l'ADN et les protéines à travers le complexe formant le pore nucléaire connu pour être un anticorps du virus HCV [43].

HotNet2

Nous avons effectué un run HotNet2 après calcul des matrices d'influences pour le réseau HPRD et généré 100 matrices d'influence aléatoires pour l'analyse statistique, puis calculé la matrice de chaleur pour le jeu de données Tai après formatage des données.

Après son exécution, HotNet2 a retenu 4 analyses correspondantes à 4 valeurs du paramètre Delta. Les changements de Delta ont un impact conséquent sur la taille et le nombre des sous-réseaux trouvés. De manière à obtenir une taille de réseau raisonnable, nous avons choisi $Delta = 0.0244$, ce qui a donné 36 sous-réseaux de taille 5 ou plus. Le temps d'exécution est de 40h et 45 min pour la génération des matrices d'influences pour le réseau HPRD seul (cette étape n'est faite qu'une seule fois). Le run HotNet2 sur les données Tai n'a pris que 12min 32s sur une machine virtuelle 4 cœurs.

Seuls 4 gènes communs entre HotNet2 et HTS Net ont été identifiés, et aucun gène commun avec Tai (Figure 5.7C). Ceci peut être dû à des différences dans les métriques utilisés, car HotNet2 est un logiciel conçu à l'origine pour travailler sur des données de mutation et donne un grand poids au connections entre gènes lors du processus de diffusion de chaleur, en minimisant l'influence des scores liés aux gènes. De plus, le programme inclut seulement la base d'interactions HPRD dans sa distribution actuelle, ce qui pourrait être un facteur limitant. HotNet2 donne la possibilité d'inclure d'autre génomes, mais le coût calculatoire en est prohibitif.

Nous avons examiné les résultats obtenus pour le sous-réseau ayant le score le plus élevé. Il contient PPP1CB, qui est un gène intéressant, bien que non trouvé par les autres programmes, puisqu'il a été identifié comme un des 40 gènes régulés dans les échantillons carcinomes hépatocellulaires liés au HCV étudiés par De Giorgi et collègues [42]. Ceci dit, aucune des cibles détectées par Tai n'ont été retenues.

Méthode	Type	Interactome(s) disponible(s)
HTS-Net	serveur web+code source et exécutable	HPRD, BIND, DIP, I2D,IntAct, MINT, Prot
Gonzalez & Zimmer	Code source non fourni	STRING
CARD	Serveur web	HPRD, BIND, BioGrid
HotNet2	Installation à partir du code source	HPRD, iRefIndex

5.4 Discussion

Nous avons développé un nouvel outil, HTS-Net, qui s'appuie sur une approche biologie des système, pour la détection non biaisée de candidats potentiels dans les screening RNAi. HTS-Net replace les gènes dans leur contexte de régulation et d'interactions pour la découverte de modules de gènes interagissant qui sont biologiquement liés à l'expérimentation en cours. Nous avons démontré d'utilité de notre approche sur 3 applications biologiques : régulation des cellules souches embryonnaire ([34]), différenciation de cellules cancéreuses dans le sein ([184]), et identification des mécanismes de régulations dans les interactions entre le virus HCV (Virus de l'Hépatite C) et sont hôte humain ([164]).

Chia *et al.* ont mis au point un screening de type RNAi pour découvrir l'identité des cellules hESC. Notre approche a permis d'identifier un certain nombre de gènes qui n'avaient pas été trouvé par Chia. Parmi celles-ci, nous avons identifié PRC2 en tant que régulateur maître de la différentiation hESC.

Wolf et al ont construit un screening RNAi pour comprendre les régulateurs clés des cellules souches cancéreuses (CSC - Cancer Stem Cell) dans le sein. La ré-analyse avec HTS-Net a permis d'identifier le rôle du régulateur d'autophagie ATGA4 dans la maintenance de la sous-population

CSC. Bien qu'HTS-Net n'ai pas retenu directement ATG4A, gène jouant un rôle sur la régulation des CSC, il a retenu ATG4B et ses voisins dans les réseaux d'interaction.

L'analyse Tai s'intéresse à l'identification des protéines humaines interagissant et supportant les interactions du virus HCV avec son hôte. Nous avons comparé les résultats de la publication originelle et la ré-analyse par Gonzalez avec HTS-Net. L'ensemble des études a permis d'identifier les gènes dans le manteau protéique COPI. Notre méthode a été spécifiquement capable d'identifier les kinases telles que PIP5K1A.

La comparaison avec les autres programmes de références sur le jeu de données Tai a démontré la capacité d'HTS-Net a retrouver un large nombre de cibles identifiées dans la publication originelle avec l'identification de nouvelles cibles non trouvées avec des approches réseaux. HTS-Net offre un environnement biologique complet pour les analyse RNAi en utilisant les données de régulation. Sur plusieurs aspects, HTS-Net est plus proche d'autres approches réseaux que de l'analyse originelle de Tai (Recouvrement plus élevé). Cela est du aux similarités des approches réseaux en général qui vont pouvoir inclure des cibles ayant un score plus faible, mais une plus forte connectivité.

HTS-Net permet d'obtenir des scores plus proche de CARD et la méthode Gonzalez & Zimmer, tandis que HotNet2 a rendu des résultats éloignés, puisque cette méthode n'a pas été conçue pour les analyses RNAi. En ce qui concerne l'implémentation, HTS-Net est fourni avec son code source Perl assorti d'une implémentation sur cluster de calcul (exécution très rapide), ou en tant que service Web Mobyle, qui ne nécessite pas d'installation mais demande un temps d'exécution plus long. CARD existe uniquement sous forme de service Web (On est donc tributaire du réseaux et de la maintenance du serveur), tandis que HotNet demande une installation locale. Gonzalez & Zimmer n'ont fourni ni service Web, ni code source.

5.5 Conclusion du chapitre

Nous proposons le pipeline HTS-Net pour l'analyse de données de screening RNAi. Le pipeline est disponible à partir de notre serveur Mobyle, et a été testé avec 3 jeux de données et comparé avec d'autres approches, certaines spécifiquement conçues pour l'analyse de données RNAi (CARD) [45], la méthode proposée par Gonzalez & Zimmer [59], et d'autres permettant tout type d'analyse de données omiques (HotNet2 - [94]).

Notre approche a montré de nombreux avantages comparées avec d'autres méthodes. Sur l'analyse du jeu de données Tai, il a identifié un nombre de gènes comparable (Gonzalez), ou supérieur (HotNet2, CARD), ce qui montre que l'information biologique fondamentale est retrouvée, tout en permettant la découverte de nouvelles cibles intéressantes. La production de rapport exhaustifs a permis l'exploration des réseaux de gènes, puisque les données de régulation sont inclues dans le pipeline, ce qui n'est pas proposé par d'autres logiciels pour l'instant (avec l'exception de HotNet2). Les temps d'exécution restent stable sur un serveur web, et une exécution ultra rapide peut être obtenue par installation sur un cluster de calcul Linux.

L'algorithme HTS-Net évite de définir des seuils sur des gènes individuels, permettant de surpasser les limitations des analyse classiques, tout en permettant de découvrir des composants clés pour les systèmes biologiques étudiés, tels que les régulateurs maîtres. Nos approche représente une première réponse à l'analyse de la régulation dans les systèmes étudiés dans les screening RNAi. Parmi les améliorations possibles, il serais intéressant de tenir compte de la qualité des interactions lors de la construction des sous-réseaux, par l'inclusion d'arêtes pondérées dans les réseaux reflétant la nature des interactions et les preuves biologiques la confortant. Tandis que les expérimentations à grande échelle deviennent de plus en plus utilisés dans les laboratoires,

les outils tels que HTS-Net seront de plus en plus indispensables pour les analyses avancées de ces données.

5.6 Remerciements

Ce travail a été financé par une bourse de l'Institut National du Cancer (INCa) numéro INCA_5911 à CG et GB.

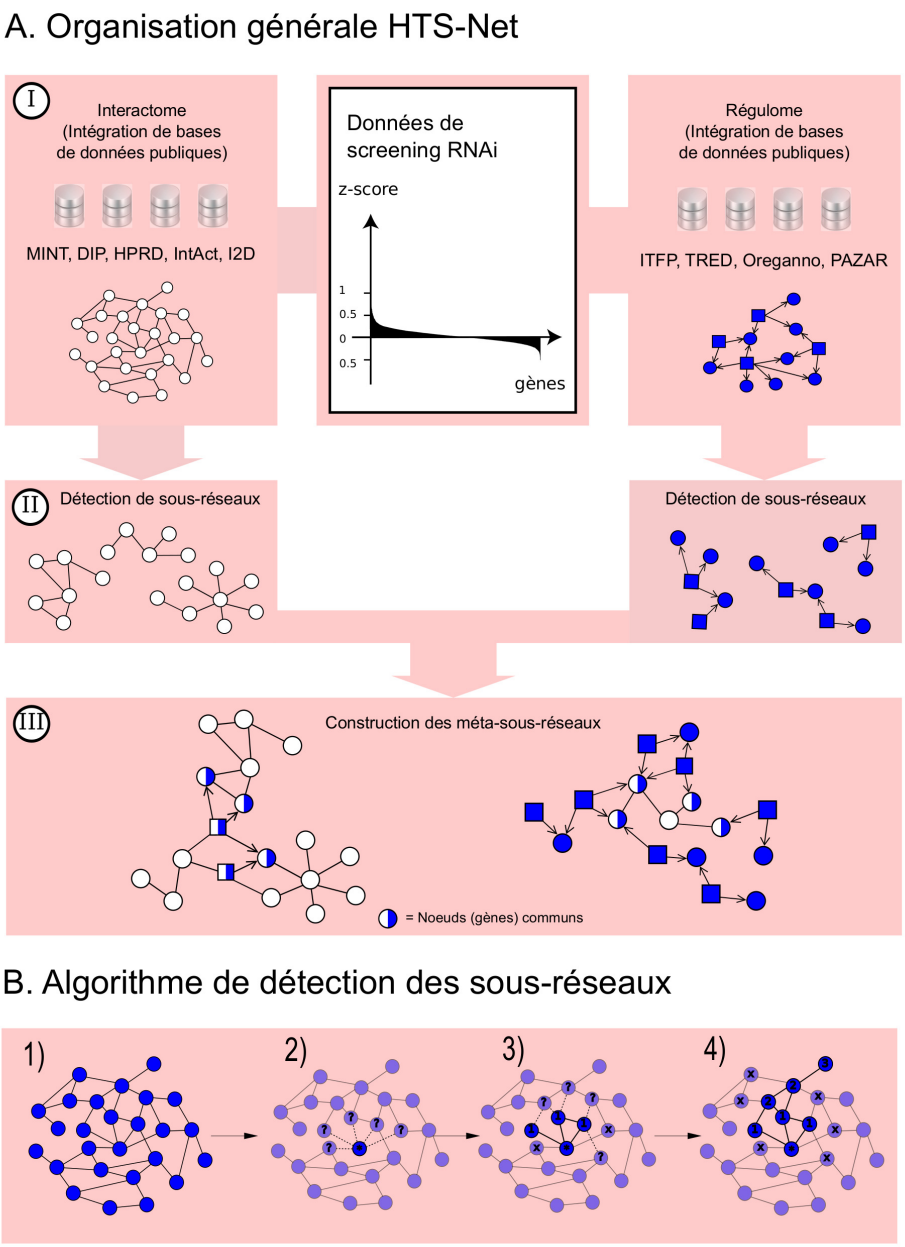


FIGURE 5.1 – 1A. Organisation générale de l'algorithme HTS-Net en 3 étapes, pour les données interactome, et les données régulome. Étape I : assemblage des données pour l'interactome (gènes représentés sous formes de nœuds blancs dans l'interactome) et le régulome (gènes représentés sous formes de nœuds bleus). Étape II : Détection des sous réseau faites séparément dans l'interactome et le régulome. Étape III : Intégration des sous réseaux issus du régulome et de l'interactome dans les méta-sous-réseaux. Les facteurs de transcriptions sont représentés sous forme de carrés et leur interactions sont dirigées (la flèche marque le sens TF vers gène cible). 1B. Détail de la détection des sous-réseaux. 1. Chaque nœud du réseau (ici, l'interactome est utilisé pour la représentation mais ce schéma reste valable pour le régulome) est utilisé comme graine potentielle. 2. Ses voisins sont testés et agrégés s'ils accroissent le score du sous réseau au delà d'un seuil *th*. Sinon, le voisin et et le chemin auquel il est associé sont fermés. 3. Ensuite, les nouveaux voisins sont explorés de manière itérative. 4. Une fois l'ensemble des voisin testés et acceptés ou rejetés, le sous-réseau est considéré comme étant complet et nous pouvons ensuite tester la graine suivante.

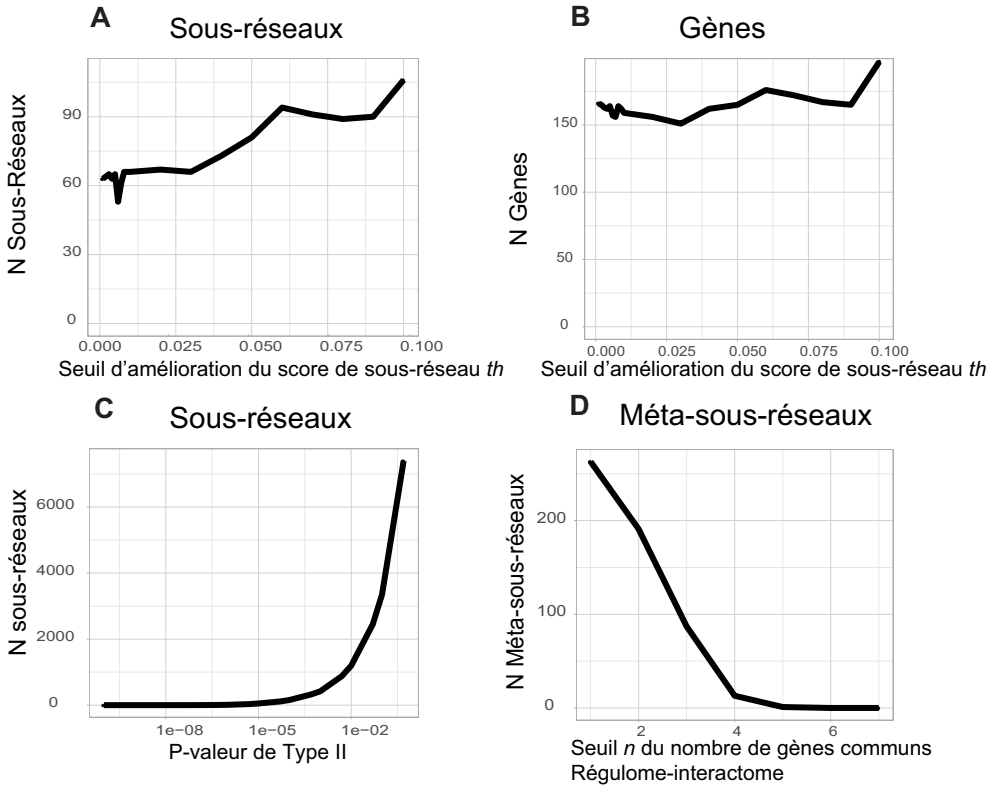


FIGURE 5.2 – Analyse de l'impact de plusieurs paramètres HTS-Net sur le nombre et la taille des sous-réseaux et méta-sous-réseaux dans le jeu de données Chia [34]. Nous avons testé l'impact du seuil minimal d'amélioration du score *th* sur le nombre de sous-réseaux et de gènes, l'impact du seuil de p-value de type II sur le nombre de sous-réseaux et l'impact de seuil d'intégration regulome-interactome sur le nombre de méta-sous-réseaux.

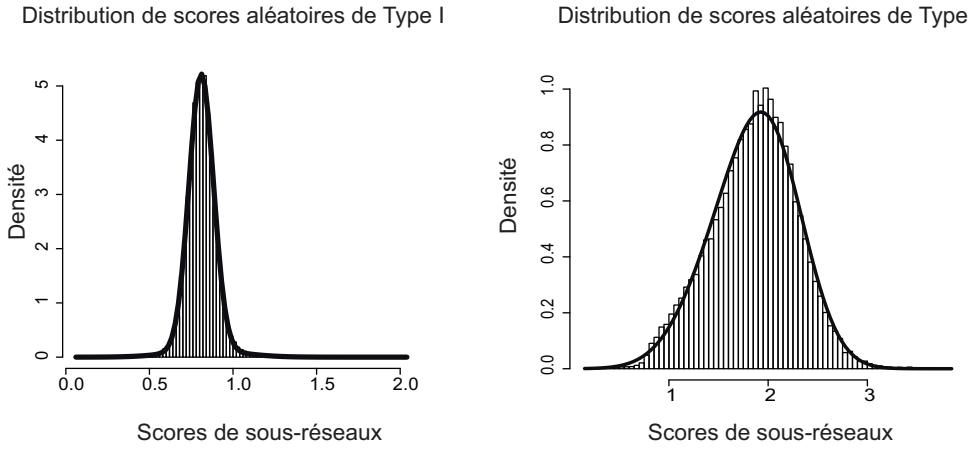


FIGURE 5.3 – Exemples de distributions de scores de type I (mélange et II obtenues avec HTS-Net pour le jeu de données Chia. L'histogramme des scores est ainsi représenté. Ces distributions sont modélisée avec une mixture de 2 gaussiennes (trait noir épais).

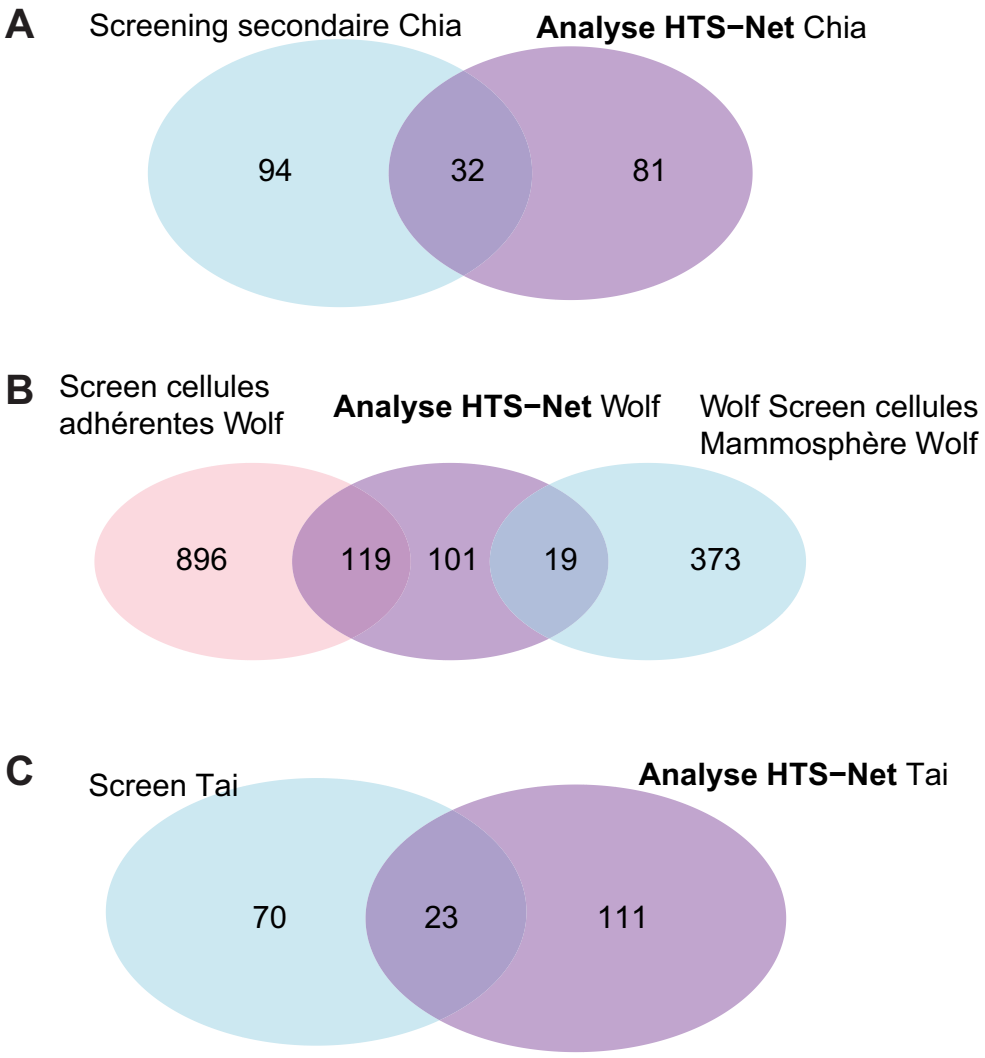


FIGURE 5.4 – Cette figure représente les diagrammes de Venn pour les listes de gènes obtenues entre les analyses HTS-Net sur les jeux de données Chia, Wolf et Tai, et les analyses originelles. En A est représenté le diagramme de Venn pour les listes de gènes obtenues par l'analyse secondaire de Chia et HTS-Net. En B est représenté le diagramme de Venn obtenu entre les listes de gènes enrichies pour les cellules adhérentes, enrichies pour les cellules formant des mammosphères et celles obtenues avec HTS-Net pour le jeu de données Wolf. Enfin, en C est représenté le diagramme de Venn obtenu par l'analyse des liste de gènes identifié par Tai et par HTS-Net.

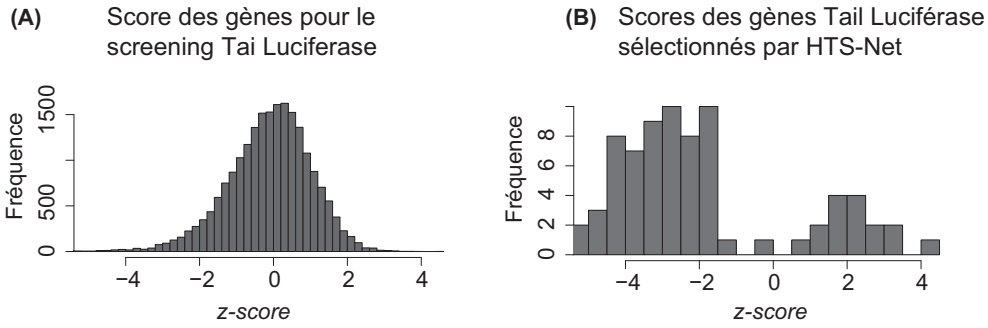


FIGURE 5.5 – Les distribution des *z-scores* de RNAi screening pour le jeu de données Tai (A) et des *z-scores* des gènes identifiées par HTS-Net (B) montrent que plusieurs gènes retenus par HTS-Net ont un *z-score* en deçà du score retenu par une validation statistique classique.

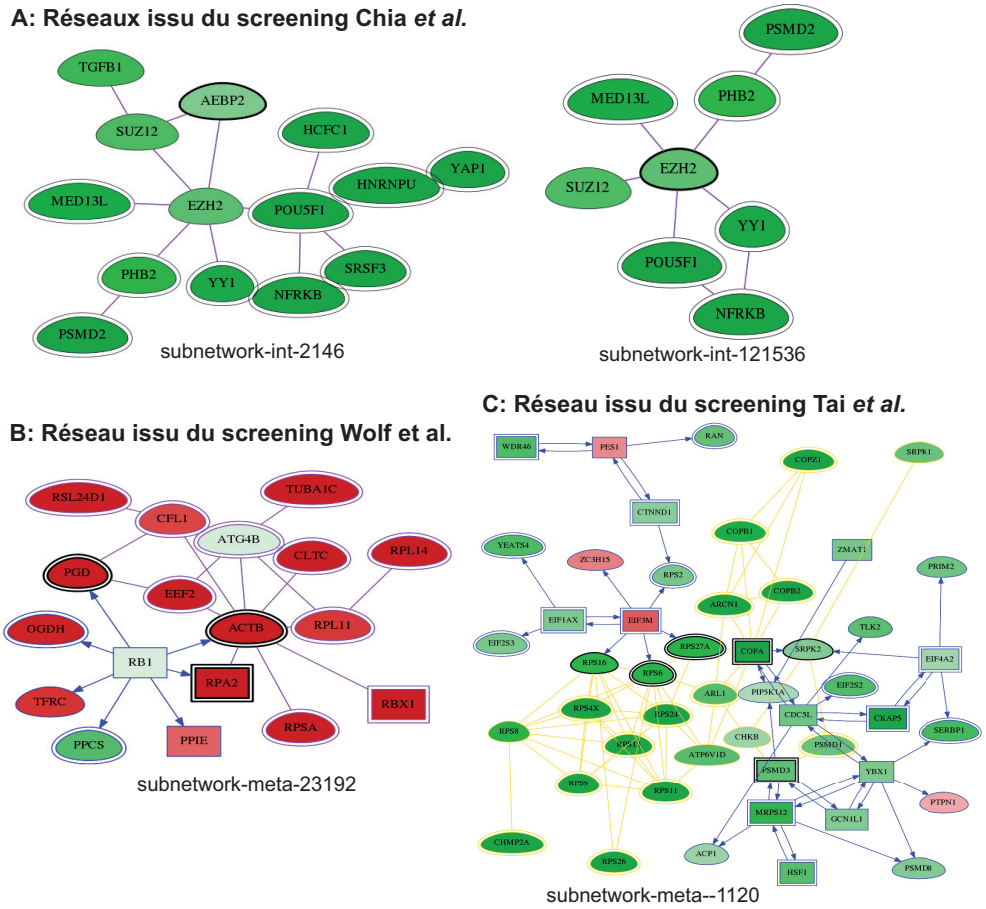


FIGURE 5.6 – Exemples de sous-réseaux détectés dans chaque jeu de données par HTS-Net. (A) Analyse Chia. Les sous-réseaux incluant EZH2 sont représentés. Le vert foncé correspond à un score plus élevé. (B) Analyse Wolf. Le méta-sous-réseau relaté à ATG4B est représenté. Le code couleur correspond au ratio score adhérent/score mammosphère. Au rouge correspondent les gènes dont le score est plus élevé dans les mammosphères (donc faible dans les adhérent), au vert correspondent les gènes dont le score est plus élevé dans les adhérent (soit non présent dans les mammosphères, comme espéré dans les ATG4B). (C) Analyse Tai montrant les interactions PIP5K1A incluant OPA, PMD1 et la machinerie ribosomale (rouge = *z-score* négatif correspondant à l'inhibition de la machinerie HCV, vert = *z-score* positif.

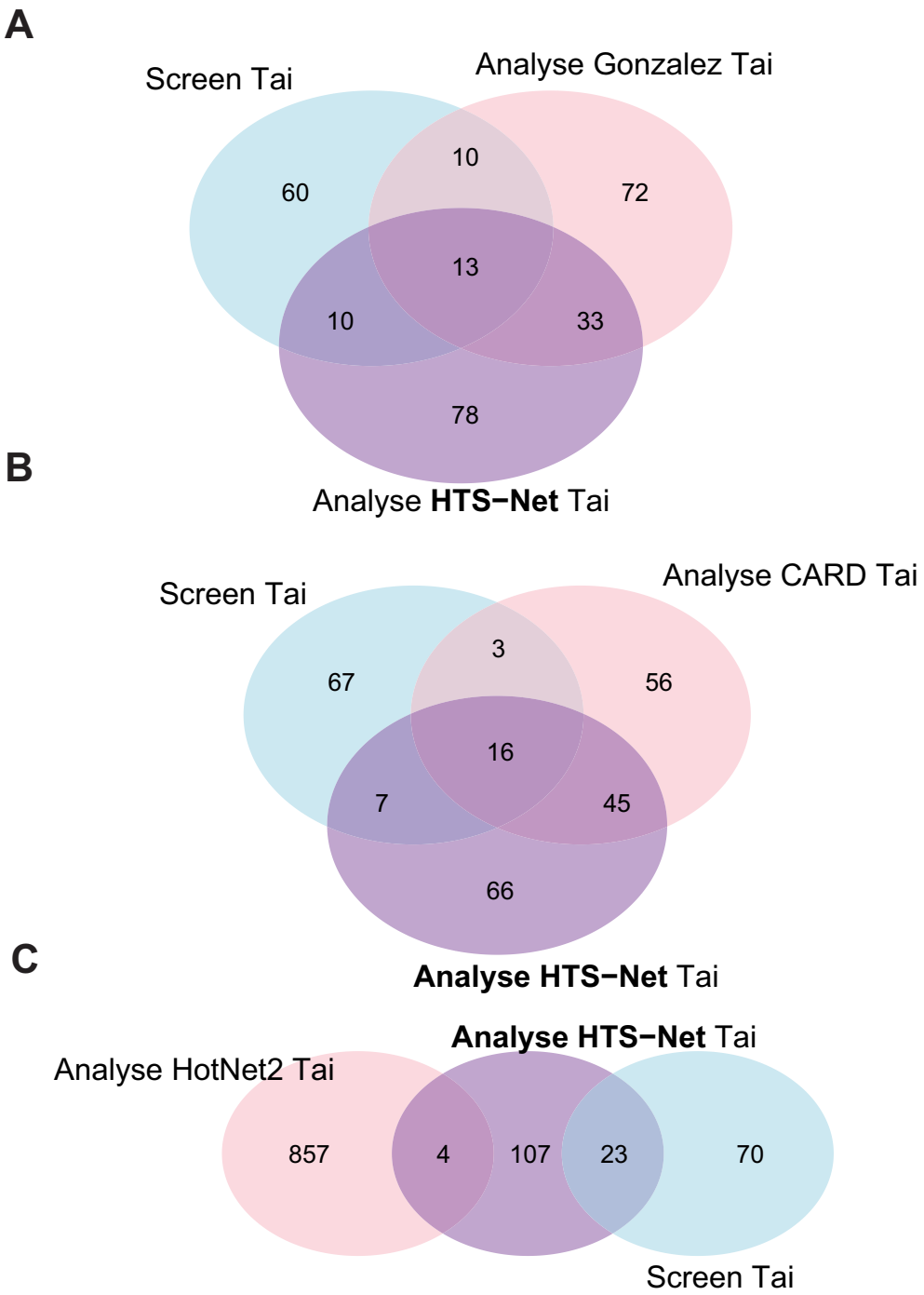


FIGURE 5.7 – Les diagrammes de Venn des gènes communs entre l'analyse originelle de Tai, l'analyse HTS-Net et d'autres méthodes. En A) est repprésent le recouvrement entre la méthode de Gonzalez & Zimmer, en B) le recouvrement entre CARD et HTSNet et en C) le recouvrement avec HotNet2.

6

Djeen : Gestion de données génomiques simplifiées sur le Web

Sommaire

Ce chapitre décrit le système de gestion de données recherche Djeen, développé au Centre de Recherche en Cancérologie de Marseille depuis mon arrivée en 2008, et publié dans Stahl *et al.* [158]. Le système Djeen a été développé conjointement avec la Fonctional Proteomics Platform (FPP) avec la collaboration d'Oana Vigy. Le développeur principal du projet est Olivier Stahl. Nous avons pu mettre au point un programme très intéressant au regard de ce qui est habituellement proposé par les spécialistes du domaine.

6.1 Introduction

L'avènement des technologies post-génomiques ont permis l'adoption et la mise en place de nouvelles techniques en biologie, notamment par un apport considérable de la biochimie, et ont radicalement changé les façons de produire, collecter partager et publier des données. Les technologies de puces à ADN adoptées dans les années 2000 et les générations successives de nouvelles technologies de séquençage mises en place depuis 2006 permettent à tout laboratoire

de disposer d'un appareillage générant des données à très grande de échelle sur un large nombre d'échantillons à des coûts toujours plus bas.

For de ces technologies, les analyses biologiques à haut débit se font maintenant d'une manière multicentrique à travers plusieurs laboratoires générant des données ayant besoin d'être finement intégrées, annotées et surtout, partagées de la manière la plus simple qui soit [22]. De fait, il est également crucial que des standards de qualité d'annotation soient utilisés et respectés. Il existe des standards d'annotations définis pour la plupart des technologies utilisés en biologie, et notamment pour les microarrays (*Minimum Information About a Microarray Experiment* (MIAME)[25]), les données de screening [148], la cytométrie en flux (MIFlowCyt) [92], et la protéomique (MIAPE) [23], dont certains sont utilisés largement pour les dépôt de données publiques. Ces standards consistent à identifier et lister l'ensemble des informations nécessaires à la reproductibilité d'une expérimentation à haut débit et font donc partie inhérente du processus d'analyse scientifique. Elles sont essentielles à la validation et la valorisation des résultats produits en laboratoire, notamment par les expérimentations haut débit en biologie. Ils servent notamment à stocker un large nombre d'échantillons dans des dépôts de données publics, tels que ArrayExpress [89] ou NCBI GEO [10].

Au laboratoire, ces standards sont également indispensables, même à usage interne, dès qu'il existe le besoin de stocker des données dans un volume conséquent et de les retrouver en les identifiants simplement. En pratique, ces standards ont été implémentés en partie dans des systèmes d'acquisition et de gestion de données liés aux appareillages, notamment les LIMS (*Laboratory Information Management System* - Système de gestion de données de laboratoire). Ces systèmes sont des logiciels de stockage de données liés à un appareil de mesure ou équipement particulier. Bien qu'ils soient indispensable au fonctionnement des équipements eux mêmes, ils présentent de nombreuses limites.

Premièrement, ils ne permettent pas de passer au niveau supérieur d'intégration de données en proposant un système d'annotations et d'intégration adapté. La plupart ont été intrinsèquement conçus pour abriter des données issues d'une unique technologie, que ce soit puces à ADN [88] [174], screening à haut débit [165] ou protéomique [67]. Par exemple, il n'est pas possible de coupler des données de chimiogramme (réponse d'un échantillon cellulaire à un traitement) avec des données issues d'un séquenceur pour plusieurs centaines de patients sans développement ad-hoc, ce qui est pourtant une priorité en les centres anti-cancéreux. En pratique, puisque les LIMS sont "mono-technologies" et donc physiquement séparés, cela complique fortement les étapes d'intégration de données et d'analyse. Cette complexification a un impact non négligeable sur le contrôle qualité [75].

Deuxièmement, chaque évolution technologique impliquant une réorganisation des données demande une re-conception et l'adaptation ou l'adoption d'un nouveau LIMS. La plupart des LIMS sont conçus pour correspondent aux besoins du laboratoire à un instant donné sans tenir compte du fait que la technologie, les formats et architecture de données, ainsi que les besoins évoluent constamment, ce qui restreint d'autant l'utilisation de ces LIMS sur le long terme.

Finalement, chaque installation d'un nouvel équipement dans un laboratoire implique de refaire l'apprentissage du système informatique de gestion de données associé puisque chaque LIMS est conçu avec une nouvelle interface utilisateur (UI), alors que le besoin fondamental reste d'associer des données mesurées à un échantillon, quel qu'il soit, qu'il soit purement biologique ou liée à un patient avec une situation clinique. La fragmentation qui en découle, impactant systèmes et ressources, a un impact négatif sur le contrôle qualité. De plus, la gestion de LIMS distincts est une tâche complexe et compliquée qui ne peut être assumée par une structure de taille modeste (comme une équipe de recherche) qui n'a pas les structures de développement informatique et bioinformatique nécessaires pour adapter ou développer des outils à ses besoins

spécifiques. De part ses limites, les LIMS typiquement proposés ou fournis ne permettent ni la collaboration translationnelle (multicentriques) et/ou multidisciplinaires (plusieurs technologies pour effectuer des mesures sur des échantillons identiques), et empêché l'intégration d'information et du transfert sur différent laboratoires.

Dans ce chapitre, je présente le système *Database for Joomla !'s Extensible Engine(Djeen)*. C'est une nouvelle génération de LIMS qui fait l'objet d'une conception simple, tout en ayant une forte capacité d'adaptation pour être directement utilisé dans un laboratoire donné ou en s'adaptant à de nouvelles technologies. Le schéma de base de données de Djeen a été conçu de façon simple de façon à maximiser la généralisation de son utilisation et des applications possibles.

Seules l'organisation des données et les métadonnées (annotations) sont stockés dans une base de données, tandis que les données expérimentales sont stockés dans le système de fichier. De plus, le programme Djeen est adaptable à tout type de technologie de laboratoire puisque aucune sémantique liée à une technologie particulière n'a été utilisé dans son code. Djeen peut donc évoluer avec les besoins des utilisateurs et du laboratoire. Puisque plusieurs technologies peuvent être gérées avec Djeen, cela permet l'unification de l'ensemble des systèmes d'information du laboratoire.

Djeen possède un grand nombre de fonctionnalités (import-export, implémentation multi-technologiques, gestion avancées des permissions utilisateurs) et est conceptuellement plus avancé que plusieurs LIMS disponibles. Djeen est un *Research Information Management System* (RIMS)[117], et en tant que tel, réponds à quatre problèmes fondamentaux en gestion de données biomédicales :

— Organisation des données (Avec mise en place de requêtes)
— Partage des données (Gestion des droits d'accès)
— Collaboration (co-contribution à un corpus de données)
— Publication (Mise en place de ressources fiables pour la publication)

L'organisation des données consiste à permettre la recherche simple du contenu des données sans la présence de la personne qui les a générées. Les données doivent donc être stockées de façon sécurisés en utilisant des standards. Le partage des données consiste à permettre à certains collaborateurs à accéder aux même données, tout en évitant leur duplication ou leur reformatage. La collaboration est un aspect important de la contribution à un jeu de données. Cet aspect permet à plusieurs collaborateurs de maintenir et de gérer les mêmes données, ce qui signifie que leur organisation n'est pas assurée par un seul individu. Enfin, la publication des données inclus le partage sécurisé des données pour un grand groupe d'utilisateurs (par Internet ou l'intranet).

6.2 État de l'art

L'état de l'art comprend une offre de solutions technologique pléthorique. Beaucoup de LIMS ont été développés par de grandes entreprises pour répondre à des spécifications très précises pour satisfaire un besoin. Il existe également des LIMS génériques qui permettent la customisation à des besoins précis. Certains ont été confrontés à des situations réelles sur une grande plateforme de production cellulaire [148]. L'ensemble du personnel de la plateforme de production cellulaire ont évalué plusieurs LIMS sur des critères précis et ont finalement adopté un LIMS commercial. Ensuite, Russom *et al.* ont dû embaucher un ingénieur informaticien consultant sur une année pour la mise en place et l'adaptation du LIMS, avec le développement de modules spécifiques. Ce travail nécessite des moyens considérables et a un coût significatif pour le laboratoire. D'autre part, il n'y a aucune garantie que le système obtenu soit compatible avec les LIMS existant et permette le croisement de données.

On voit donc qu'un LIMS générique commercial à sources fermées n'est pas un produit facilement adaptable et nécessite une installation très complexe spécifique à l'utilisation qui en sera faite. De plus, il existe une très forte dépendance sur le logiciel du vendeur et sa capacité à supporter le produit et assurer sa longévité. Hors, sur un produit à sources fermées, seul le vendeur a, par définition, la capacité de développer et supporter ledit produit, et ce sans garantie pour l'utilisateur sur le long terme. Cela implique une perte de données précieuse si le logiciel ne permet pas l'exportation sur un format lisible par d'autres systèmes.

Un nombre de LIMS open-sources ont été publiées par la recherche académique et sont disponibles au téléchargement, et répondent partiellement aux besoins décrits, mais souffrent également de limites. Beaucoup de LIMS sont limités à une technologie unique (Protéomique, ms_lims [67] ou microarrays (système BASE [174], LAD (Longhorn Array Database, [88]), ce qui est une limitation très sérieuse pour la plupart des utilisateurs qui ont besoin d'un système identique pour stocker l'ensemble des données générées. D'autres LIMS permettent la gestion de plusieurs technologies. MADMAX [97] permet le stockage de plusieurs jeux de données hétérogènes dans un même dépôt. Cependant, son implémentation et installation sont très complexes et basées sur un système de gestion de bases de données (SGBD) Oracle, dont le prix de la licence seule est un facteur limitant pour beaucoup de laboratoires. SIGLa (Systema Integrado de Gerência de Laboratoriós [102]) est basé sur un outil de gestion de flux d'information (workflow) qui le rend également utilisables sur plusieurs types de technologies. Le concept de flux d'information n'est pas forcément un besoin pour l'ensemble des utilisateurs et complexifie grandement le processus de saisie des données. BonsaiLIMS [11] est un LIMS de conception plus légère (utilisation faible des ressources serveur) qui a été initialement conçu pour centraliser des données patient pour une application en recherche translationnelle. Cependant, BonsaiLIMS possède seulement un modèle de schéma de bases de données très limité qui doit être adapté à chaque utilisation et qui nécessite le développement d'une interface utilisateur. Cette solution n'est donc pas immédiatement utilisable.

L'information scientifique générée par les laboratoires de recherche en biologie, que ce soit en immunologie, ou cancérologie ou autres possède une immense valeur pour les patients. Cependant, seule une partie limitée de cette information est utilisée par les biologistes et cliniciens confrontés directement avec la maladie. Une des raisons de cette limitation est la complexité de l'extraction de connaissances biomédicale à partir de bases de données existantes, et l'annotation poussée des données générées par les équipes, que ce soit sur des aspects clinique ou phénotypique. Aux données existantes, s'ajoute la problématique de l'évolution très rapide des technologies en biologie qui demande de nouveaux types de traitement informatique et qui apportent de l'information à résolution beaucoup plus élevée, par exemple en biologie moléculaire, notamment en séquençage ou en protéomique. L'inclusion plus large de patients dans des essais cliniques, permet également de générer de plus en plus de données biologiques en lien avec des formes cliniques précises, mas tout en augmentant la complexité et les demandes en analyse et traitement de données, sans compter le volume de données.

6.3 Implémentation

6.3.1 Objectifs

Djeen est un Système de Gestion de l'Information appliqué à la Recherche (Research Information Management System - RIMS) développé pour répondre aux quatre objectifs suivants :

Organisation des données

Djeen permet la structuration des données dans une hiérarchie, permettant de créer et de gérer de structures de données complexes tout en faisant appliquer des standards propre au laboratoire ainsi que les standards de type *Minimum Information* (MI). Djeen est capable de gérer des projets contenant des données hétérogènes tout en permettant la maintenance d'annotations diverses à travers l'utilisation (optionnelle) de gabarits (*templates*).

Partage des données

Le partage des données est géré à travers des permissions permettant de fixer les différents niveaux d'accès aux données. Plusieurs rôles peuvent êtres associés aux utilisateurs en fonction du niveau de leurs responsabilité. Cela permet de gérer des données confidentielles, de partager les données avec plusieurs groupes ou utilisateurs ou de faire un partage public sur internet.

Collaboration entre utilisateurs

La collaboration va plus loin que le partage des données. Elle implique que les jeux de données sont construits par plusieurs collaborateurs, tandis que la structure générale et les annotations sont sous le contrôle d'un chef de projet. Dans Djeen, une fois que les permissions ont été mises en place aux utilisateurs et groupes appropriées, l'administrateur de projet peut contrôler l'homogénéité des annotations avec des gabarits. Les gabarits permettent de de faire respecter l'intégrité des données et leur homogénéité, ainsi que le contrôle qualité. Leur rôle est détaillé section 6.3.4.

Publication

De façon à publier les jeux de données et à répondre aux besoins en partage de données et collaborations, Djeen a été développé en tant qu'application Web.

Administration

Un cinquième point a été pris en compte lors de la conception de Djeen. Djeen a été conçu comme une application Web simple à installer, à apprendre et à administrer. Une instance Djeen est simple à maintenir par un groupe de recherche avec peu ou pas de ressources humaines dédiées à la bioinformatique, et utilisable avec un apprentissage minimal.

6.3.2 Système de Gestion de Contenu

Djeen a été développé en tant qu'extension au système de gestion de contenu Joomla ! (Content Management Systems - CMS). Les CMS sont une catégorie d'applications permettant la gestion et la publication de contenu, le "contenu" étant défini de manière plutôt large. Mooney *et al.* [112] donne une revue des CMS utilisés en bioinformatique. Dans notre cas, le contenu réfère à de l'information scientifique, et plus précisément à des jeux de données et leur annotations. Les CMS sont très attractifs puisqu'ils sont conçus de façon modulaire et permettent la création de nouvelles applications pour visualiser du contenu. La réutilisation des outils présents dans un CMS pour le développement d'applications web simplifie potentiellement le développement et permet de se concentrer sur les fonctionnalités de l'application au lieu de passer du temps à re-développer des fonctionnalités forcément présentes (telles que l'apparence visuelle ou les systèmes d'authentification).

Comparé à une application développée entièrement en interne, développer une application Web comme une extension d'un CMS nous a permis de maximiser la sécurité en prenant avantage du système d'authentification existant. De plus, la sécurité ne dépends plus de l'application elle même mais du CMS, qui doit être choisi maintenu et suivi, ainsi que de la configuration locale du réseau.

Le choix approprié pour un CMS est crucial, surtout quand on considère les problèmes de sécurité [112]. Plusieurs critères ont prévalu au choix du CMS utilisé dans ce projet, notamment le support, la présence d'une API (Application Programming Interface) documentée, la qualité de l'interface, et notamment la présentation visuelle et le comportement de l'interface graphique, et enfin la simplicité d'installation et d'administration. Au regard de ces considérations, nous avons décidé de développer Djeen en tant que composant *Joomla !*.

Joomla ! (version utilisée au moment de la publication : 1.5) est un CMS open-source [18] qui présente une API documentée pour créer des extensions avancées basé sur un modèle de développement de type Modèle-Vue-Contrôleur (MVC) [93]. De plus, il est développé par une large communauté de programmeurs qui corrigent rapidement les trous de sécurité et mettent à jour la documentation.

6.3.3 Intégration de Djeen avec Joomla !

Dans cette section, je vais exposer comment Djeen interagit avec Joomla !, et comment ses différent éléments sont architecturés. La figure 6.2 montre l'architecture trois tiers, basée sur une infrastructure MVC, pour séparer la structure des données (Modèles) la visualisation des données (Vue), et les actions (Contrôleur). La raison d'être de la couche *Framework* est de gérer les données avec Joomla !. Elle inclut trois différentes classes et librairies pour accéder à la base de données et au système de fichier. La couche *Application* gère l'interface utilisateur (*User Interface* UI), telles que le panneau d'administration et la partie site web publique. Finalement, la couche *Extension* inclut des gabarits pour gérer le rendu des interfaces et des extension pour changer ou étendre les fonctionnalités Joomla !. On distingue deux types d'extensions : Les *Modules* sont utilisés pour pour afficher des données supplémentaires dans des boîtes prévues à cet effet. Les *Composants* sont une forme d'extension plus avancée qui permettent de développer des applications complètes en ayant accès à l'API Joomla !. Djeen a été développé comme un Composant et implémente sa propre base de données et un système de fichier dédié.

Par l'utilisation de l'API Joomla !, Djeen est déployé et géré à partir du panneau d'administration Joomla !. Aucun logiciel tierce partie ou client n'est nécessaire pour travailler avec Djeen et la compatibilité a été pleinement testé avec les navigateurs Mozilla Firefox ®et Google Chrome ®(version utilisée au moment de la publication). L'application est contenue dans l'interface Joomla ! à travers la couche application.

La sécurité des données est assurée par la séparation complète des bases de données Djeen et Joomla ! pour simplifier les sauvegardes et garder Djeen indépendant à tous les niveaux. Les objets Djeen sont stockés dans une base de données spécifique (au choix : MySQL ®ou PostgreSQL ®sont proposés) tandis que les fichiers sont simplement stockés sur le système de fichier. La connexion à la base de données externe Djeen est assurée par un second niveau de sécurité qui utilise un mot de passe encrypté stocké dans le fichier de configuration du composant Djeen. Plusieurs instances peuvent être maintenues et chaque instance possède son mot de passe.

18. http://www.joomla.org

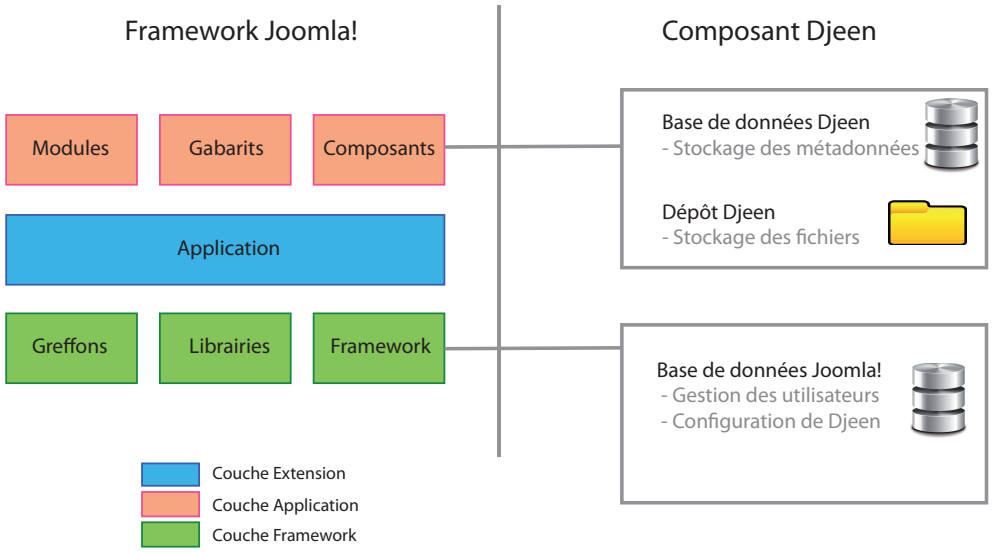


FIGURE 6.1 – API (Application Programming Interface) de Joomla ! et implémentation de Djeen. Cette figure détaille l'API Joomla et les mécanismes d'interactions entre Djeen et Joomla ! La partie droite de la figure représente le système trois tiers (adapté à partie de la documentation Joomla ! (http :// docs.joomla.org/Framework/1.5#Packages_and_Classes). Cette architecture définit le logiciel depuis l'accès aux données jusqu'à l'affichage final à travers un modèle de développement Modèle-Vue-Contrôleur (MVC). La partie gauche de la figure présente les principaux composante de Djeen (Base de données Djeen, dépot de fichier) ainsi que les composants propres à Joomla ! (Base de données Joomla !, contenant les informations utilisateur et la configuration des composants), réutilisés par Djeen. Djeen utilise la couche Framework pour interagir avec Joomla ! et gérer l'authentification des utilisateurs et la couche Applications pour l'affichage.

6.3.4 Organisation des données et éléments de Djeen

Dans cette section, les objets spécifiquement Djeen (Projets, Gabarits) sont identifiés avec une lettre capitale. Ce n'est pas le cas quand ces termes sont utilisés de façon non spécifique. La figure 2 6.2 exporte le modèle de données Djeen et la dualité fichiers-base de données. L'objet Fichier est la pièce centrale de l'information gérée par Djeen, puisqu'il représente la granularité la plus faible de l'information dans un grand nombre d'analyse biologique (par exemple un fichier .CEL.gz pour une expérimentation Affymetrix ®). Ce modèle et le code associé ont été conçus pour être indépendant de la technologie et libre de toute sémantique spécifique pour rester adaptable à tout type de données et généralisable autant que possible. En ce qui concerne la base de données, la principale entité gérée par Djeen est le Projet. Les Projets sont organisés en hiérarchie et un projet peut référencer plusieurs sous-projets. Cette hiérarchie est reflétée dans l'organisation en répertoire du système de fichier. Les Projets sont liés aux Fichiers venus de l'expérimentation ou des analyses par l'utilisation de l'organisation en répertoires. Les Projets sont équivalents aux Répertoires et les fichiers sons simplement stockés dans ces répertoires. Cette organisation est très intuitive et simple à sauvegarder, migrer ou dupliquer.

Les Fichiers et Projets sont annotés par deux types de Métadonnées, les Annotations et les Caractéristiques. Les Annotations sont liées aux Fichiers et les Caractéristiques sont liées au Projets. Les Annotations permettent de décrire les informations spécifiques aux échantillons (Fichiers) et les Caractéristiques permettent de stocker les informations spécifiques aux Projets, donc à plusieurs échantillons. Les Gabarits, qui correspondent aux Annotations ou Caractéris-

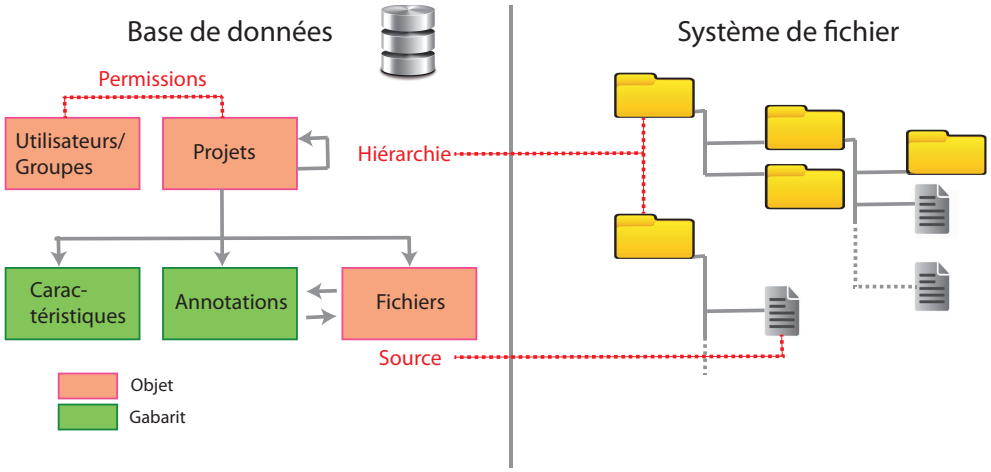


FIGURE 6.2 – Modèle d'organisation des données dans Djeen ; Djeen sépare la hiérarchie des projets. Les Gabarits et les Annotations sur un coté (ils sont stockés dans la base de données relationnelle de Djeen) et les données elle-mêmes (les données expérimentales sont stockées sur le système de fichier). Les données stockées dans la base de données sont réparties sur plusieurs objets, chacun d'eux étant représenté par une table de base de données, liée à une classe dans l'architecture MVC. Les Projets sont le principal élément organisationnel de Djeen. Il permettent la construction d'un arbre organisationnel avec des projets et des sous-projets et contiennent les Métadonnées sur lui-même et les fichiers qui lui sont liés. L'objet Fichier est le plus simple objet et représente le degré de granularité le plus fin dans Djeen dans les analyses à haut débit. Cette hiérarchie projet/sous-projet/fichiers est dupliquée dans le système de fichier. Les Métadonnées peuvent être converties en Gabarits, permettant aux utilisateurs de les réutiliser dans d'autre projets. L'objet Utilisateurs est une table spécifique contenant les informations utilisateurs et groupes. Il a préséance sur le système d'utilisateurs Joomla ! en ajoutant des informations spécifiques à Djeen permettant de gérer des permissions et groupes spécifiques tout en permettant la réutilisation de la plupart des fonctionnalités déjà mise en place dans le CMS, telles que le système d'authentification ou la messagerie électronique.

tiques, ne peuvent être modifiés une fois qu'elles sont utilisés dans un projet, de façon à éviter la perte d'annotations involontaire.

Djeen n'impose pas de contraintes sur la structure des Projets ou le type d'Annotation utilisées. Cela permet l'utilisation de plusieurs type de données dans la même instance de Djeen. Cependant, les gestionnaire de projets peuvent utiliser les Gabarits pour faire respecter l'intégrité des Annotations à fournir par l'expérimentaliste et garantir l'ensemble des information minimum (MI). La provenance des échantillons, par exemple, peut être encodée en tant que champ obligatoire. Les gabarits permettent également de structurer les projets, pour permettre d'intégrer des données plus simplement même après que la collecte de données pour le projet ait commencée. En plus des annotations, Djeen permet de faire ressortir (avec un drapeau) l'annotation qui correspond à la question biologique posée dans un projet.

Djeen permet le partage de données en utilisant un système avancé de gestion de permissions ayant préséance sur le système natif de Joomla !. Quatre rôle ont été définis dans Djeen (Superadministrateur, Administrateur, Modérateur et Utilisateur), des droits spécifiques étant accordés à chacun sur les données, la gestion des Gabarits et des droits utilisateurs. Un Projet donné est géré par un Administrateur ou un Modérateur ayant le rôle de gestionnaire de Projet. Il contrôle les annotations et les droits utilisateurs sur les fichiers. Les droits sur les Projets

et Sous-Projets peuvent être donné à des groupes d'utilisateurs spécifiques pour gérer finement l'accès à des projets collaboratifs de grande ampleur.

6.3.5 Fonctionnalités et interface utilisateur

L'interface de Djeen a été conçue pour faciliter la navigation et la manipulation de données, aussi intuitivement que possible, et peut s'apprendre très facilement. Pour optimiser l'interopérabilité avec Joomla ! et optimiser sa maintenabilité, Djeen a été principalement développé en PHP. Un chargeur de donné a été développé en Perl pour permettre des transfert de données massifs (Plusieurs giga-octets) pour l'importation de grands jeux de données dans Djeen. Tous les rendus d'interface et l'interaction avec les utilisateurs ont été implémenté sur une couche Ajax/Javascript avec la librairie Motools [19]. De plus, une boite de dialogue customisable à été mise à disposition pour les administrateurs pour afficher les informations légales (Licences ou Conditions d'Utilisations) liée à une instance de base de données particulières. Les principaux éléments de l'interface Djeen ont été représentés Figure 6.3. Les icônes principaux (Figure 6.3A) permettent un accès rapide aux projets, expérimentations et fonction recherche. Les autres icônes (Figure 6.3B) permettent la gestion des gabarits, la gestion des projets, des options liées à l'utilisateur, et connexion/déconnexion au système. Les fils d'Ariane permettent de naviguer dans le hiérarchie des projets. Le panneau (6.3D) montre la principale vue du projet. Cette vue contient la description générale du projet, le nom de l'administrateur du projet, ainsi que la date de création, le gabarit utilisé, ainsi que la table des caractéristiques (6.3H). L'onglet (6.3E) permet la gestion des sous-projets, fichiers, annotations, historique du projet et permissions utilisateurs spécifiques. L'onglet 'Fichiers' ouvre un panneau (superposé sur la figure), affichant une liste de fichiers liée au projet courant, annotés avec la table d'annotations (6.3G). De plus, les utilisateurs peuvent générer des captures de l'état courant du projet en utilisant la fonction *"History"* dans (6.3E). Une capture de ce type permet à un administrateur de projet d'extraire un jeu de données à partir de la version précise d'un projet en vue d'une publication. Une zone de notification affiche les messages du système (6.3I) de façon élégante et non intrusive. Les icônes (6.3I) renvoient à des actions générales présentes dans la plupart des tables, notamment *"Print'* et *"Export"*. Les Icônes (6.3J) renvoient à des actions générales liée à la gestion des projets, notamment *"Print"*, *"New Project"*, *"Edit"*, *"Copy as template"*, *"Copy in Clipboard"* and *"Clear clipboard"*. Les annotations peuvent être soit saisies par l'interface Djeen ou importée par fichier CSV. La saisie directe d'annotations est facilitée par un outil intégré d'édition de données tabulées. Plusieurs fichiers de données peuvent être importées en une fois (Par l'utilisation de fichiers au format zip ou tar), et des valeurs d'annotations par défaut spécifiées de façon à accélérer la saisie des données. Dans un jeu d'annotations, les variables qui représentent la question biologique peuvent être marquées pour identifier de façon explicite ce qu'est la question biologique.

6.4 Résultats et discussions

Pour démontrer les fonctionnalités de Djeen et son utilité dans un environnement scientifique, nous détaillons la procédure complète de gestion d'un jeu de données de type puces à ADN (Affymetrix ®) mesurant l'expression des gènes de patients atteint d'un cancer du sein, avec des annotations cliniques. Bien que cet exemple spécifique soit lié aux microarrays, ce qui est décrit ici est bien sur applicable à d'autres types de données. Des tutoriels spécifiques détaillant la gestion des données microarrays et cytométrie en flux sons accessibles à partir de la documentation

19. http ://www.motools.net

Djeen. Le jeu de données utilisé dans l'exemple qui va suivre [44] est publiquement disponible sur la base de données Gene Expression Omnibus (NCBI GEO [20]) sous le numéro d'accession "GSE7390". Ce jeu de données contient 198 fichiers Affymetrix ®au format CEL, chacun d'eux correspondant à un échantillon unique de tumeur dans le sein. Pour être capable d'être cohérent dans le format d'annotations sur plusieurs projets, nous avons conçu et employé un gabarit Djeen pouvant être réutilisé sur plusieurs projets. Ce projet fera partie d'une hiérarchie et définit comme étant un sous projet d'un projet plus global appelé "Cancer du Sein". Cela permet l'inclusion de nouveaux projets en cancérologie au même niveau dans Djeen.

6.4.1 Préparer la hiérarchie des projets dans Djeen

D'abord, les données doivent être stockées au bon endroit dans Djeen. Cela dépends de l'organisation locale des données, ainsi que du type de données à traiter. Dans tous les cas, il faut définir la hiérarchie soigneusement, de façon à pouvoir stocker de nouveaux projets sur le long terme. Pour notre exemple, nous allons créer un méta-projet *"breast_cancer"*, puis un sous-projet *"Desmedt et al. Local Copy"* pour le projet que nous sommes en train de stocker, celui-ci étant une copie du travail publié par Desmedt *et al.*. C'est simplement fait en cliquant sur l'onglet général Project, puis sur le bouton d'ajout (+). Le nom et la description du projet peuvent être ajoutés à ce moment, par exemple *"Meta_Breast_Cancer_Transcriptome_Project"* et *"Meta Breast Cancer Transcriptome project grouping multiple datasets"*. Les permissions peuvent être définies pour limiter l'accès aux données. Puisque c'est un jeu de donnée publiques, nous pouvons le partager avec les personnes enregistrées comme utilisateur du système. Les changements sont enregistrés par un clic sur le bouton "Save". De façon similaire, un sous-projet peut être défini pour stocker et décrire les données. Cela se fait en cliquant sur l'onglet "sous-projet", ceci créant un sous-projet appelé *"GSE7390 (Desmedt et al.)"*. Nous n'avons pas défini de gabarit à cette étape. De nouveau, une description peut être ajoutée. Les permissions sont également laissées à leur valeurs par défaut : accès aux utilisateurs enregistrés. Les changements peuvent être sauvés par un clic sur "Save". Ces étapes ont donc créé la hiérarchie suivante : *"Meta_Breast_Cancer_Transcriptome'->'GSE7390 (Desmedt et al)"*.

6.4.2 Importation de données microarray(Fichiers CEL)

Les jeux de données de tests complets peuvent être directement téléchargés à partir de GEO et du numéro d'accession GSE7390. Après avoir téléchargé l'archive complète, tous les fichiers CEL sont importés dans Djeen par un clic sur l'icône "File", sous le projet "GSE7390 (Desmedt et al)". Djeen possède une capacité de décompression intégré pour directement importer les archives zip, tar.gz et tar. Après importation des données, tous les fichiers sont accessibles à partir de l'onglet "File".

6.4.3 Ajouter et formatter un gabarit au standard MIAME

. Il reste à annoter les échantillons que nous venons d'importer. A cette étape un gabarit MIAME va être appliqué sur les données. Ce gabarit remplit deux fonctions : D'une part, il permet à un administrateur de projet de s'assurer que les données sont bien conformes au standard MIAME, et d'autre part, de réutiliser les mêmes annotations pour d'autres jeux de données et ainsi de garder une cohérence entre les données.

20. http://www.ncbi.nlm.nih.gov/geo

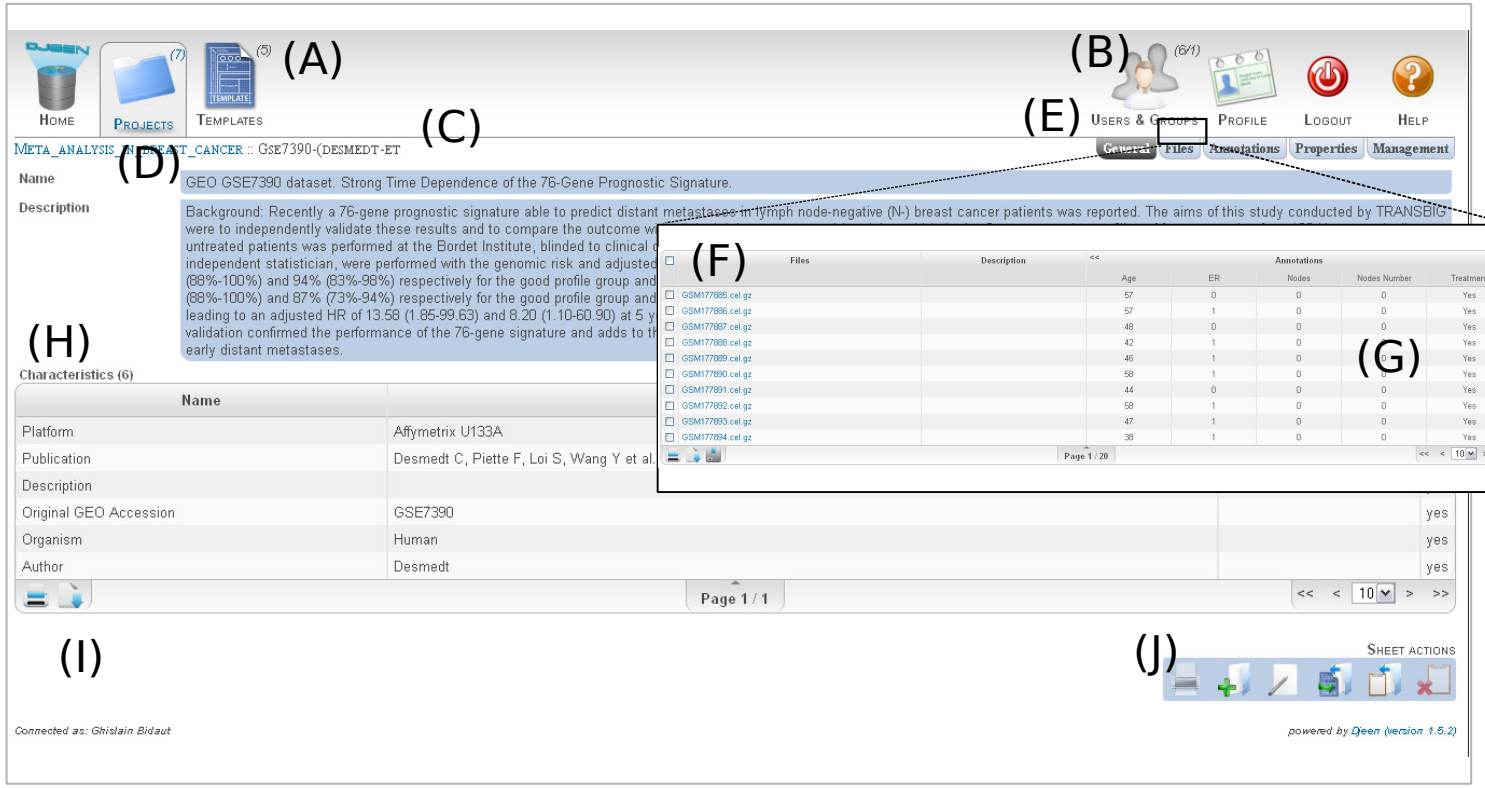


FIGURE 6.3 – Interface Web de Djeen. Cette figure montre l'interface Web de Djeen ouverte sur le vue Projet, telle que l'on peut la voir en étant connecté comme utilisateur avec permissions de lecture. L'interface est contenue dans celle de Joomla ! (Non représentée ici), et présente tous les éléments reliée à cette vue particulière et plusieurs éléments communs à d'autres vues, dont la vue *Home* (Liée aux éléments de l'utilisateur courant) la vue Projets, et la vue Gabarits. Les icônes en B, également accessible à partir des autres vues, sont liée aux tâches administratives, c'est à dire Administration des utilisateurs et groupes, la gestion du Profil de l'utilisateur courant, l'icône de connexion/déconnexion au système , et l'accès à l'aide. Le fil d'Ariane (C) permet de se localiser dans la hiérarchie globale, tout comme dans un système de fichiers classique. En D sont les éléments généraux d'identification du projet, soit son nom et sa description. En E nous voyons une sous vue du Projet, qui est la vue "Générale" (celle actuellement affichée), les fichiers (représentés dans la vue F superposée sur la figure), les Annotations, qui contiennent le détail de chaque annotation représentée en G. Les propriété et la gestion permettent la gestion des permissions utilisateurs et autres variables techniques (acronyme du projet, identifiant du gabarit). La table G dans la vue Fichiers montre les valeurs d'annotations pour chaque échantillons. La table H liste les caractéristiques des projets. En I se trouvent les icônes des fonctions d'impression et d'export, qui sont communes à chaque table. Les icônes d'actions en J permettent les actions générales sur le projet courant, telles que l'impression, créer un nouveau projet, éditer un projet, définir un gabarit à partir du projet courant, copier dans le presse-papier, et effacer le presse-papier.

Les gabarits sont ajoutés à partie de l'onglet *"Template"* de Djeen. Dans notre exemple, nous allons nommer notre gabarit *"Breast Cancer Transcriptome Template"*. Le système permet de lier des mots-clés à cet gabarit. Nous donnons les mots-clés suivants : *"Transcriptome Breast Cancer Clinical Data"*. En tant que description de ce gabarit, nous spécifions *"Template for describing microarray of breast cancer samples"*. Les changements sont sauvés par *"Save"*. Un exemple concret de gabarit est donné Figure 6.4.

Ensuite, les deux éléments clés du gabarit doivent être remplis. Les Caractéristiques (Variables spécifiques au Projet) et les Annotations (variables spécifiques au Échantillons). Dans notre cas, les Caractéristiques sont *"Author", "Original GEO ArrayExpress Accession Number", "Organism", "Platform Vendor" and "Platform Type"*. Des listes de valeurs et des descriptions peuvent être ajoutés si nécessaire pour chaque Caractéristique. Les Annotations sont entrées de façon similaire, tel que spécifié Figure 6.4. Les champs *"e. dmfs" and "t.dmfs"* sont spécifiés comme question expérimentale puisque ce jeu de données permet de lier la survie à l'expression des gènes dans les tumeurs.

Ensuite le gabarit que nous venons de créer doit être lié au projet *"GSE7390 (Desmedt et al.)"* . Cela se fait à partir de la vue projet en éditant le champ gabarit. Une fois cette opération faite, 5 Caractéristiques et 6 Annotations se retrouvent ajoutées au projet.

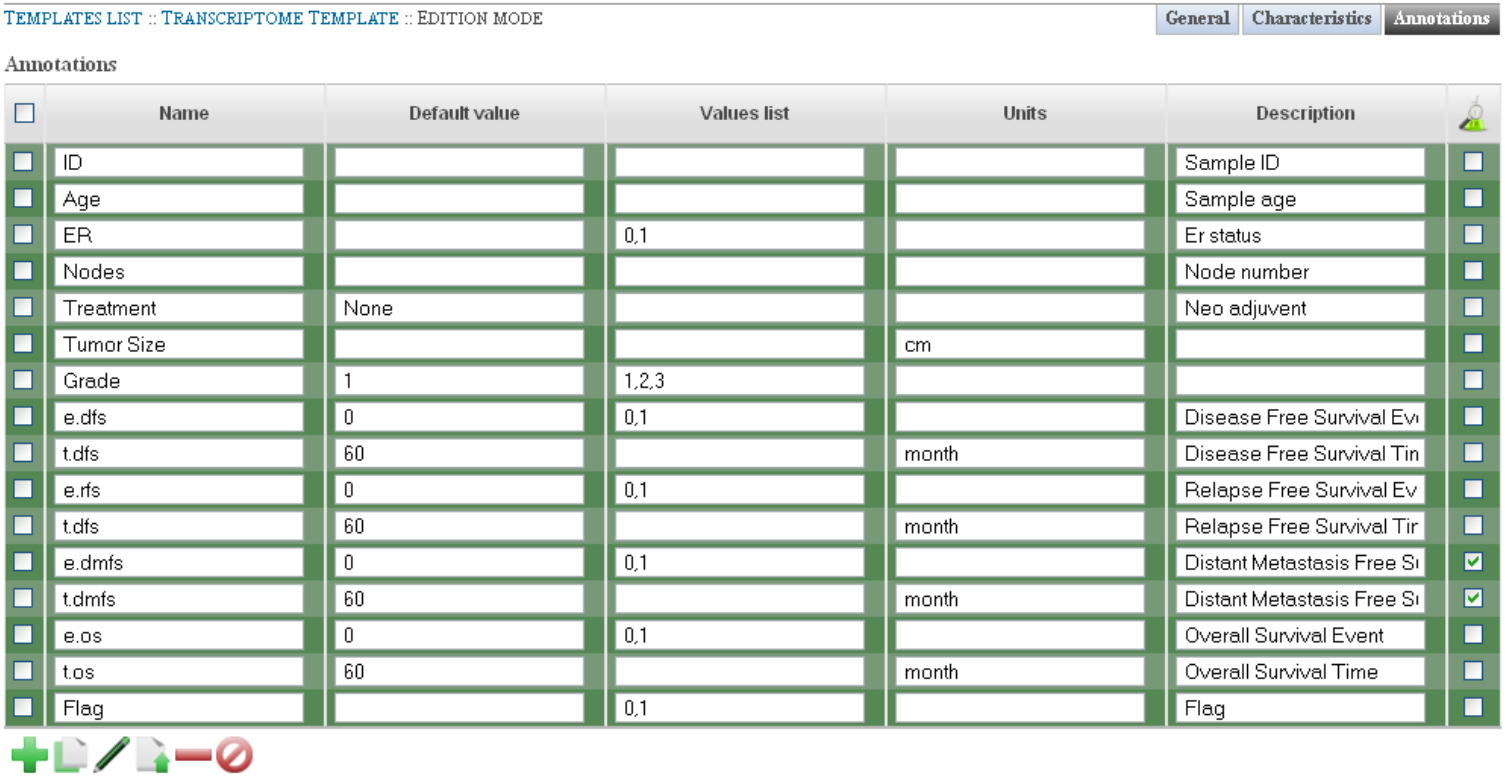
L'étape finale consiste à remplir ces champs. le projet contient 198 échantillons, ce qui rend le processus d'annotations manuelle plutôt long, surtout que GEO contient déjà l'ensemble des annotations. Pour faciliter ce processus, nous avons écrit un fichier d'annotations qui peut être importé directement dans Djeen (Ce fichier est disponible à partir de la page de documentation http ://sourceforge.net/projects/ djeen/files). Cependant, il doit être formaté pour être importé, spécialement pour correspondre aux annotations existantes. Cela se fait en générant un fichier d'annotations vide en exportant les annotations courantes du projet par "File"-> "Save the file to annotation.txt". Ensuite, il faut éditer "annotations.txt" avec un tableur et importer les données du fichier Desmedt-annotations fourni dans la documentation de Djeen. Ce fichier peut être directement importé avec les fonctions "Import" et "Save". Le message suivant doit apparaître : *"198 file(s) correctly edited - 3168 file an- notation(s) updated. "*. A cette étape, le projet a été annoté correctement, le structure de l'annotation, le gabarit peut être réutilisé dans d'autres projets si nécessaire.

La dernière étape consiste à régler correctement les permissions pour partager les jeux de données. Les paramètres *"Read"* et *"Edition"* peuvent être gérés à partir de l'onglet *"Permissions"*. De plus, des paramètres *"Red"* et *"Edition"* spécifiques peuvent être liées pour le groupe des utilisateurs.

Le projet est maintenant complètement chargé dans Djeen et peut être édité et partagé. Les données du projet peuvent être téléchargées à partir de Djeen. Les données liées à un échantillon individuel peuvent être également téléchargés en passant par l'onglet *"File"*, tandis que le jeu de données complet peut être obtenu en faisant une capture de sauvegarde ("Snapshot") puis en téléchargeant l'archive.

6.5 conclusion

Djeen est un système de gestion des données de la recherche, conçu pour répondre aux besoins des laboratoires. Il permet le partage, le stockage sécurisé, l'annotation et la gestion de jeux de données hétérogènes, facilite les opérations d'annotations répétitives et permet l'importation de données existantes. Djeen organise les données dans une hiérarchie naturelle similaire à cette que l'on trouve dans les systèmes d'exploitation couramment utilisés. Les gabarits (utilisés



Name	Default value	Values list	Units	Description
ID				Sample ID
Age				Sample age
ER		0,1		Er status
Nodes				Node number
Treatment	None			Neo adjuvent
Tumor Size			cm	
Grade	1	1,2,3		
e.dfs	0	0,1		Disease Free Survival Ev
t.dfs	60		month	Disease Free Survival Tin
e.rfs	0	0,1		Relapse Free Survival Ev
t.dfs	60		month	Relapse Free Survival Tir
e.dmfs	0	0,1		Distant Metastasis Free Su
t.dmfs	60		month	Distant Metastasis Free Su
e.os	0	0,1		Overall Survival Event
t.os	60		month	Overall Survival Time
Flag		0,1		Flag

FIGURE 6.4 – Interface d'édition des gabarits. Cette figure montre l'interface d'édition des gabarits. Cette interface permet d'éditer la liste des annotations et de leurs valeurs correspondantes. Deux mécanismes ont été implémentés pour contrôler les valeurs permises pour une annotation donnée. Si une valeur par défaut est mentionnée, elle sera utilisé lorsque l'on importe des données et que l'on ne spécifie pas d'autres valeurs. D'autre part, une liste de valeurs peut être spécifiée pour limiter la liste des valeurs acceptables. Lors d'un import de données, la première valeur sera utilisée si aucune autre valeur n'est mentionnée. Les métadonnées (caractéristiques et annotations) sont éditées avec une interface similaire.

optionnellement) permettent l'automatisation et la gestion des MI (Minimum Information). Ces fonctionnalités sont démontrées dans ce chapitre avec la description d'un vrai cas d'usage avec des données de puces à ADN. Djeen possède une procédure d'installation simple et une interface web. Il est conçu comme un composant Joomla !, possède une courbe d'apprentissage rapide et peut être déployé sur MySQL ou PostgreSQL. Les développements futurs incluent une interface de recherche avancée sur les données, un module de développement et d'exécution de script pour programmer Djeen directement à partir de l'interface , et un système d'authentification par serveur LDAP.

6.6 Conclusion du chapitre

Par ce projet, j'ai pu développer un système de gestion de données applicable en biologie mais aussi à d'autre domaines. C'est un travail complémentaire aux travaux d'analyse et d'intégration présentés aux chapitres précédents.

6.7 Remerciements

Ce projet a été financé par un financement INCaInserm alloué à Ghislain Bidaut. Le Cancéropole PACA et l'Université Aix-Marseille ont apporté un support supplémentaire pour OS. Le serveur Djeen est supporté par un financement de la Fondation pour le Recherche Médicale à GB. Le support pour HD est donnée en partie par le contrat GEPETOS 2007 de la Région Languedoc-Roussillon. OV est supporté par Le CNRS. GB est supporté par l'Université-Aix Marseille.

7

Interactome dans les cellules LAM

Sommaire

Ce chapitre reprend l'analyse interactome que j'ai effectuée sur la période 2013-2018 pour le projet d'identification des interactions de cellules souches stromales supportant LT-HSC et différentiation cellulaire. Ce travail a été fait en collaboration avec l'équipe de Stéphane Mancini et Michel Aurrand-Lions du CRCM et a fait l'objet du travail de thèse de Marielle Balzano, et est actuellement en cours de publication. Il marque l'utilisation du travail produit par Maxime Garcia à des fins autres que la classification de tumeurs et à une application de compréhension mécanistique des interactions entre un type cellulaire et son environnement.

7.1 Introduction

La question biologique centrale à ce travail de recherche, et qui fait l'objet de la thèse de Marielle Balzano et de la publication en cours de soumission est liée à l'étude de cellules souches stromales. La partie bioinformatique de cette analyse, qui est l'aspect qui nous intéresse plus particulièrement ici, a permis l'étude du réseau d'interactions entre la niche des cellules stromales péri-sinusoidales et plusieurs sous types de cellules hématopoïétiques. Cette approche a permis l'identification de la connexion entre paires ligands-récepteurs qui ont été validés expérimentalement. C'est un exemple d'intégration de données générées au laboratoire avec des données publiques obtenues par interrogation de la base de données *Gene Expression Omnibus* (NCBI GEO) [10]. Ensuite, les ligands ont été détectés par superposition de données d'interactions binaires également issues de bases publiques.

7.2 Matériel et Méthodes

7.2.1 Analyse d'expression par microarrays

Les cellules stromales ont été triées au CRCM et profilées sur puces Affymetrix ®430 2.0 et MoGene 1.0 ST. Les autres types cellulaires ont été téléchargés sur GEO et ont été profilés sur Affymetrix ®MoGene 10 ST. Avant intégration, chaque jeu de données a été normalisé individuellement pour les CELs qui étaient nécessaires pour l'analyse avec la méthode *Robust Multichip Average* (RMA) [77] par l'utilisation du package *oligo* sous Bioconductor [73]. La version 3.2.0 de R a été utilisée dans le reste de l'analyse.

Le contrôle qualité de l'hybridation ainsi que la compatibilité entre les différentes sources de données ont été faire également sous *oligo*. Elles ont consisté en l'examen visuel des images scannées après hybridation, de la visualisation des distributions de l'expression des sondes et de leur comparaison entre échantillons par boxplots, et des analyses suivantes. Les sondes n'atteignant pas de valeur plus élevée que la valeur définie par examen des distribution de l'expression à travers l'ensemble des échantillons sont été considérés comme non exprimés.

7.2.2 Analyse interactome

Le pipeline bioinformatique pour l'analyse interactome a consisté de quatre étapes principales, illustrées figure 7.1. 1) Collecter les données d'interaction dans les bases publiques, 2) construire une carte d'interactions de référence par croisement des données interactome, 3) normalisation globale des données d'expression microarrays avec conversion des identifiants, et comparaison des gènes présents dans différentes populations.

Construire une carte d'interactions de référence

Pour cette étape, nous avons réutilisés les outils développés pour le pipeline ITI [51], (Chapitres 3 et 4). Pour l'analyse interactome, plusieurs bases de données d'interactions protéines-protéines ont été téléchargées, analysées et intégrées par superpositions des interactions trouvées. Les PPI obtenues à partie de la Database of Interacting Proteins (DIP) [150], de la *Human Protein Reference Database* (HPRD) [86], d'IntAct [85], et de la Molecular INTeraction database (MINT) [95] ont été intégrées pour l'obtention de l'interactome pour cette étude. Un total de 13202 protéines et 70530 interactions ont été retenues.

Tous les identifiants de protéines ont été mappé sur numéro d'accession Entrez Gene en utilisant les tables de correspondantes disponibles au téléchargement sur le site du FTP du *National Center for Biotechnological Information* [120]. Les données publiques ont été téléchargées sous forme brute (fichiers CEL) à partir des liens fournis sur *Gene Expression Omnibus*. Les échantillons fournis par le consortium Immgen ont donné les populations hématopoïétiques et stromales (GSE15907 [68], GSE58589 [180]. Tous les jeux étudiés ont été profilés sur plateforme Affymetrix MoGene 10 ST (GPL 6246). Les données ont été automatiquement lues, ont fait l'objet d'un contrôle qualité et normalisés sous R/Bioconductor à l'aide du package *oligo*. Les sondes ont été mappées aux identifiants *EntrezGene* par valeur médiane maximum de leur profil d'expression [144]. De façon à comparer des données d'expression pour plusieurs jeu de données, nous avons classé les gènes comme étant exprimées/non exprimées pour chaque échantillon en utilisant la méthode de traitement Immgen [68], illustrée figure 7.2. Brièvement, les données d'expression du jeu de données complet ont été modélisées par une mixture de Gaussienne (Nombre de gaussiennes choisi=4) par l'utilisation du package R *Nor1Mix*[21]. Nous avons considéré que la

21. https ://cran.r-project.org/web/packages/nor1mix/index.html

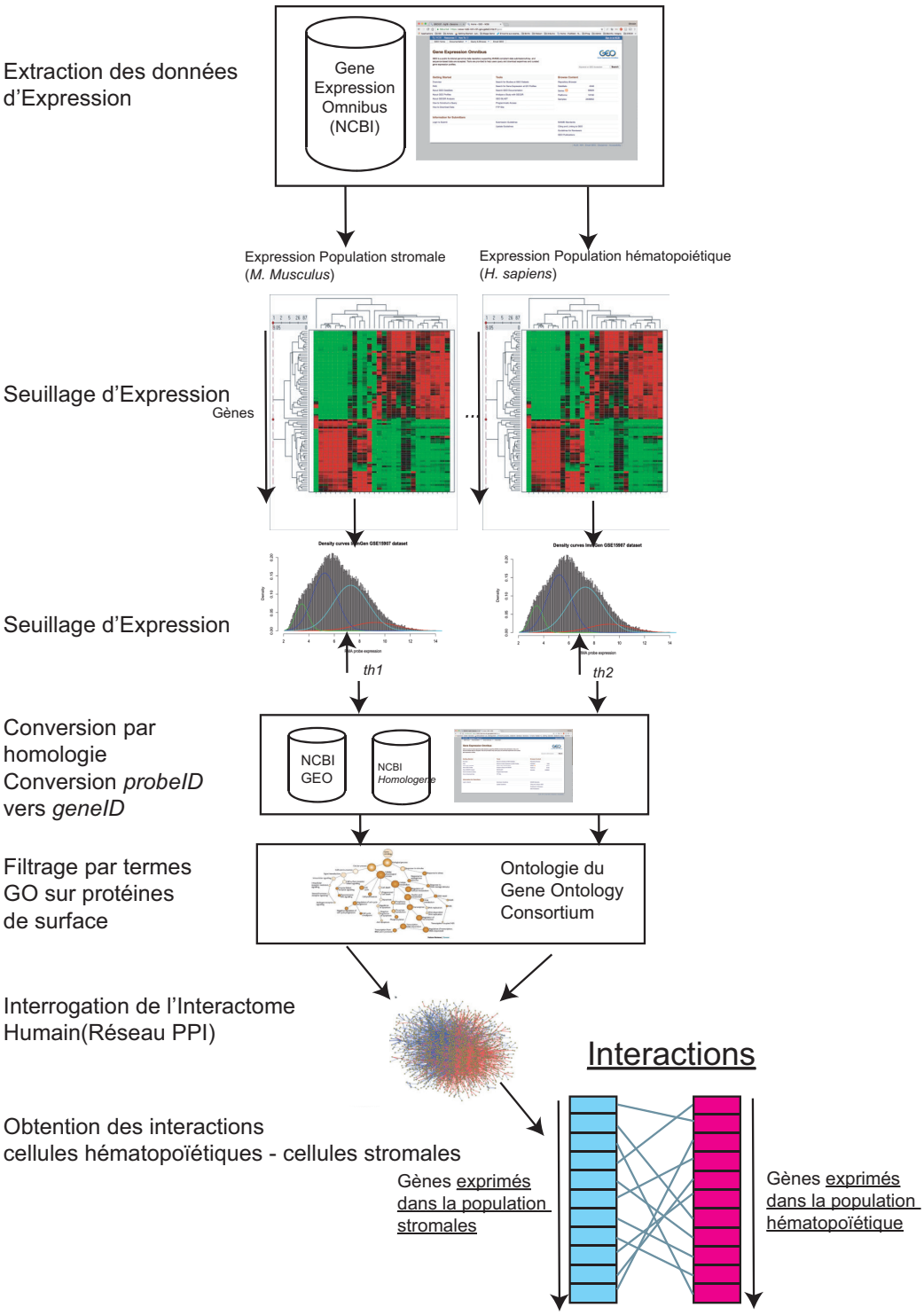


FIGURE 7.1 – Structure du pipeline d'analyse mis en place pour l'identification des interactions entre le type cellulaire stromal et le type cellulaire hématopoïetique

seconde gaussienne en partant des faibles valeurs d'expression modélisait les gènes non exprimées. Le seuil a été fixé à $th = 95\%$ de la seconde gaussienne en partant des valeurs peu exprimées. Les seuils suivants ont été retenus pour l'ensemble des jeux de données : GSE15907, $th = 6.89$;

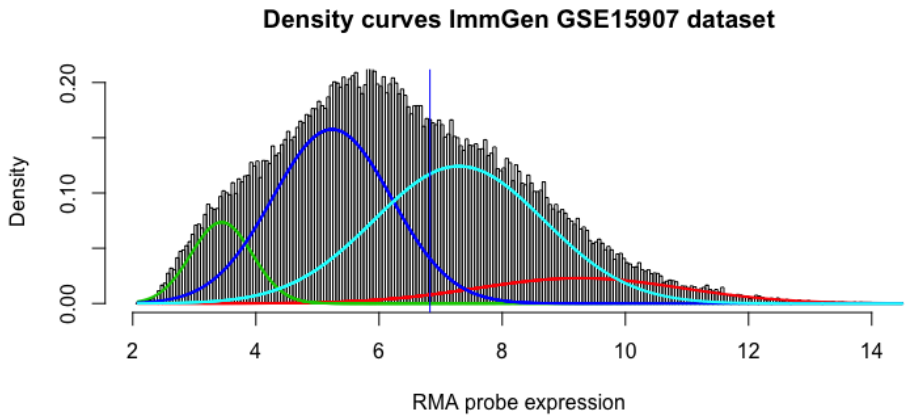


FIGURE 7.2 – Distribution des valeurs d'expression des sondes pour le jeu de données Immgen (GSE15907) et mise en place du modèle à 4 gaussiennes permettant de trouver le seuil d'expression. Le seuil est fixé comme étant le seuil à 95% de la seconde gaussienne, qui modélise les sondes faiblement ou non exprimées.

GSE58859, $th = 5.65$; données des populations stromales générées dans cette étude, $th = 4.77$.

7.2.3 Conversion d'identifiants entre organismes

Pour comparer les populations cellulaires dans la souris en utilisant notre carte d'interactions PPI générées dans l'humain, les identifiants GeneID doivent être homogénéisés. Pour passer des numéro d'accessions *Mus musculus* vers les numéro d'accession *Homo sapiens*, une conversion automatiése a été implémentée, basée sur la base de données Homologene du NCBI [22]. La base a été téléchargée sous forme de fichier plat à partir du site FTP du NCBI [23]. Nous avons ensuite fait une conversion vers les Gene Symbols en utilisant la table de correspondance contenus dans le fichier `Gene_Info.gz` disponible à partir de la même source [24].

7.2.4 Filtrage sur les termes Gene Ontology

Les gènes ont été filtrés sur plusieurs termes GO de façon à concentrer l'analyse et éviter la détection de faux positifs sur des gènes considérés comme n'étant pas des marqueurs de surface ou des marqueurs extracellulaires. Les identifiants GO suivants ont été utilisés : *GO :0009897, GO :0005887, GO :0030246, GO :0031012, GO :0005581, GO :0005923, GO :0048535, GO :0030595* et *GO :0008083*.

7.2.5 Comparaison des populations cellulaires

Les interactions des cellules stromales YFP+CD54+BP1+ à partie de la souris IL7-Cre/Rosa-eYFP avec les sous types cellulaires LT-HSC, ST-HSC, pre-pro-B, pro-B ou pre-B ont été faites en utilisant le jeu d'interactions précedemment défini. Les interactions *cis* ont été considérées comme étant de faux positives et seules les interactions *trans* ont été retenues. Les signatures de

22. https://www.ncbi.nlm.nih.gov/homologene
23. ftp://ftp.ncbi.nlm.nih.gov/pub/HomoloGene/current
24. ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/gene_info.gz

gènes obtenues pour chaque sous type hématopoïétiques ont été ensuite comparées de la façon suivante : LT-HSC contre ST-HSC, pre-pro-B contre pre-B et pro-B contre pre-B. Pour comparer chaque population, la procédure suivante a été utilisée : Si un gène est présent dans la population avec une expression différentielle plus faible que 2 fold, le gène est considéré commun. Sinon, il est assigné à la population ayant l'expression la plus élevée. Les gènes sont ensuite comptés et des diagrammes de Venn générés (Figure).

7.2.6 Enrichissement en termes GO

L'enrichissement en termes GO (Gene Ontology) a été calculé pour chaque population par test hypergéométrique associé avec la correction pour test mlultiple adaptée implémentée dans la version R/Bioconductor de gProfiler [142]. Le seuil de signifiance statistique retenu pour les p-values est de 5%, et la taille du jeu de référence pour les gènes fixé au domaine des gènes annotés.

7.2.7 Enrichissement en groupes de gènes *Gene Set Enrichment Analysis*(GSEA)

Les groupes de gènes signant pour les cellules CD54+BP1+, exprimées dans leur interaction spécifiques avec LT-HSC, pre-pro-B and pro-B ont été analysées avec le logiciel d'enrichissement *Gene Set Enrichment Analysis* (GSEA) [159]. C'est une alternative intéressante à l'enrichissement simple par distribution hypergéométrique qui permet une mesure sur l'ensemble des gènes.

7.3 Conclusion et directions futures

Lors de ce projet, j'ai pu mettre en œuvre l'expertise mise en place lors des projets interactome et cellules souches décrits chapitres 2 et 3, sur l'intégration de données hétérogènes. Mon expertise sur l'utilisation de données publiques a également été très utile. Le projet, initialement développé en Perl, est actuellement migré dans un package R/Bioconductor, pour la mise à disposition à l'ensemble de la communauté. Une publication est en cours de soumission.

8

Détermination des mécanismes de régulation génomique par intégration de données cistrome, épigénome et transcriptome

Sommaire

Ce chapitre présente le sujet de thèse que j'ai mis en place avec Pr Jacques Van Helden du laboratoire TAGC, et pour lequel Lucie Khamvongsa-Charbonnier a obtenu un financement de 3 ans de l'Ecole Doctorale de la vie et de la Santé de l'Université Aix-Marseille.

8.1 Introduction

L'analyse du cistrome (l'ensemble des éléments génomique d'attachement des facteurs de transcription) est essentielle pour la détermination de l'ensemble des éléments régulant les gènes. L'immunoprécipitation de chromatine associée aux puces (ChIP-chip) ou au séquençage de dernière génération (ChIP-Seq) est la technologie de choix pour l'identification du cistrome, que ce soit pour les facteurs de transcription que pour l'étude des marques d'histones. La technologie ChIP-Seq apparaît comme étant une alternative plus intéressante au Chip-chip (rapport signal-bruit supérieur, coût moindre, dynamique de détection des pics accrue) mais au prix d'une analyse de données plus complexe et coûteuse consistant en l'alignement des séquences puis à la normalisation du signal ChIP, et enfin à la détection des régions enrichies (pour les analyses d'histones) ou pics (pour les analyses de facteurs de transcription). Chaque étape de l'analyse bioinformatique ChIP-Seq comporte des heuristiques qui ont une influence non négligeable sur le résultat, notamment les étapes de normalisation et de détection des pics, mais aussi l'alignement. L'étape de normalisation du signal ChIP est nécessaire pour la comparaison correcte de l'échantillon ChIP contre ADN total et la détection des pics, et la quantification du signal. L'approche la plus naturellement utilisée consiste à faire une normalisation globale, c'est à dire la division du signal par le nombre total de séquences, ce qui est proposé par le programme MACS [193]. Hors, il apparaît qu'une normalisation locale peut être indispensable dans certaines

conditions de modification du contrôle par un traitement. Furusawa *et al.* [49] proposent une normalisation basée sur le comptage de séquence sur différentes régions d'un gène de ménage. Le défaut de ce type d'approche est la dépendance du résultat final à la mesure d'un seul gène, ce qui en limite la fiabilité. La méthode de normalisation par l'utilisation d'un épigénome exogène [127] est une alternative de choix pour l'analyse et permet la détection de pics qui ne pourraient pas être identifiés par l'approche classique. Cependant, elle ne peut être appliquée aux données existantes des bases publiques que l'on pourrait être amené à analyser. Enfin, les analyses combinées expression/ChIP sont typiquement réalisées de façon triviales par analyse séparées des données d'expression/ChIP suivi d'un simple croisement de listes de gènes. Il est évident que des approches beaucoup plus fines d'intégration de données d'expression (issues typiquement du RNA-Seq) sont nécessaire par la mise en place d'un modèle statistique rigoureux.

8.2 Matériel et Méthodes

Pour ce sujet, nous proposons la mise en place d'une approche combinée en analyse RNA-Seq/Chip-Seq qui permettra de limiter la dépendance à la normalisation et la détection de faux positifs/négatifs par :

- Une approche d'intégration de données d'expression sur les profils ChIP pour améliorer l'identification des zones d'enrichissement, qui pourra être complétée par l'analyse simultanée de plusieurs échantillons.
- Une analyse de type traitement du signal et l'utilisation des ondelettes [20] pour une modélisation du signal ChIP.

Dans un premier temps, l'analyse par transformée en ondelettes permettra l'exploration du signal ChIP sur plusieurs échelles de façon combinée [107], pour la détection de régions enrichies quelle que soit leur taille (typiquement étroites pour les facteurs de transcription ou larges comme dans le cas des histones), ce qui évite la fixation de seuils de taille de régions purement arbitraires. Ensuite, l'intégration de données d'expression a priori permettra de confirmer les gènes enrichis. Pour ce faire, nous proposons l'introduction d'un modèle bayésien, qui identifiera les gènes régulés à partir des données d'expression RNA-Seq et des régions enrichies identifiés par l'étape précédente et utilisées en tant qu'information *a priori*. La thèse proposée consistera à évaluer la méthode proposée et la comparer avec les approches existantes, sous forme de *benchmark* sur des jeux de données de référence connus et évalués. Enfin, nous appliquerons les méthodologies développées en collaboration avec plusieurs équipes du Centre de Recherche en Cancérologie de Marseille et du TAGC sur des projets stratégiques de l'étude de plusieurs facteurs de transcription et marques d'histones en cancérologie.

Conclusion Générale

Conclusion

Discussions sur l'intégration de données

Lors de l'écriture de ce mémoire, destiné à soutenir et obtenir mon Habilitation à Diriger les Recherches, j'ai pris un grand plaisir à me replonger dans les projets auxquels j'ai participé, certains étant de ma propre initiative, d'autres des collaborations scientifiques. Dans tous les cas, il a fallu développer une expertise, la mettre en œuvre et faire une analyse de donnée poussée et adaptée au problème posé, ce qui est un travail passionnant intellectuellement mais aussi dans ses aspect d'interactions humains, que ce soit avec ses collègues de travail ou dans l'encadrement.

Je vais profiter de cette conclusion pour aborder mon ressenti sur ces travaux, et comment j'envisage le futur de la plateforme Cibi à la lumière de ces projets.

Lors de la rédaction de cette thèse, j'ai remarqué que deux grandes catégories de problèmes posés à la bioinformatique se dégageaient. Le premier axe est d'utiliser des données non directement conçues pour la question posée initialement, pour poser de nouvelles questions. C'est le cas sur l'aspect d'intégration de jeux de données d'expression dans les cellules souches (chapitre 2) pour lequel les jeux de données initiaux répondaient à des questions beaucoup plus spécialisées à chaque organe. Il a fallu constamment filtrer et adapter les données à disposition à la question posée, comme le travail présenté dans le chapitre 7 sur l'analyse de données d'interactions. Le second axe est l'innovation et l'apport de l'informatique (au sens traitement de l'information) au domaine de la bioinformatique. C'est de cette manière que j'ai implémenté ITI (Chapitre 3), CITI (Chapitre 4), HTS-net (Chapitre5), et Djeen (Chapitre6), sans uniquement ré-appliquer ce qui se fait déjà couramment mais en se forçant à constamment remettre en question ses acquis pour adapter tel ou tel algorithme ou traitement à telle question posée. L'aspect intégratif est primordial car il n'est pas possible d'aborder l'analyse de données biologique sans aspect intégration de plusieurs sources de données, puisque c'est inscrit dans le dogme même de la biologie moléculaire.

C'est sur ces directions que j'ai monté la plateforme Cibi du CRCM. Une plateforme de bioinformatique portant l'intitulé *Intégrative* ne peut être une simple plateforme de service standardisés, d'une part parce que l'offre technologique en biologie moléculaire ne cesse d'évoluer mais aussi parce que les questions posées sont pointues et propres à chaque chercheur. Cela rend tout traitement standard impossible, bien que l'on peut mettre à disposition certaines briques de base telles qu'un pipeline ChIP-Seq ou d'analyse d'expression différentielle. Bref, c'est un melting-pot de questions posées dont la trame de font reste l'intégration de données disparates, dont chacune apporte sa contribution permettant au scientifique de construire une analyse et porter une étude à sa conclusion.

J'ai construit la plateforme sur deux axes, l'un que j'ai nommé *Service*, bien que je ne crois pas que ce soit réellement un service au sens strict du terme, c'est à dire correspondant à une prestation standardisée. J'ai centré cet activité sur des projets courts pour une analyse de données brève pour une publication non centrée sur la bioinformatique mais qui nécessitent néanmoins

un développement spécifique.

En ce sens, la plateforme Cibi telle qu'elle est répond à un besoin très concret de demande d'analyse personnalisé. Il m'est arrivé un très grand nombre de fois de répondre à des demandes de ré-analyses pour des données précédemment analysées par des plate-formes tierces. Le travail n'avait pas à être mis en cause, mais l'absence de dialogue entre le bioinformaticien et le biologiste avait fait que le message contenu dans le compte-rendu d'analyse n'était pas passé. Lors de ces ré-analyses, je n'ai pas, dans un grand nombre de cas, fait tourner de pipeline complet, mais plutôt repris les résultats et essayer d'en sortir une interprétation biologique. Cela en dit long sur ce que doit être une plateforme de bioinformatique : On ne peux délivrer de résultats de manière froide et détachée (c'est pourtant faisable sur des analyses que l'on considère standardisées, telles qu'une recherche de régions enrichies pour un facteur de transcription par exemple, ou une analyse d'expression différentielle) mais cette analyse doit impérativement être accompagnée d'un dialogue pour que chaque partie se comprenne au mieux. A mon sens, Cibi ne peut être une plateforme mais doit quasiment constituer une équipe de recherche.

L'autre axe est la partie *Recherche*, qui n'est pas séparable de la première. Vu la nécessité d'avoir une veille technologique permanente, une connaissance profonde des technologies d'analyse de données, ainsi qu'une culture prononcée des algorithmes, et méthodes d'analyse existantes, il est absolument nécessaire d'avoir un axe recherche au sein de la plateforme. Cette axe recherche a pour devoir de publier des publications sur des sujets précis, ce qui est indispensable pour crédibiliser la plateforme et asseoir son expertise sur le long terme. C'est ce qui est fait sur notre plateforme pour l'intégration de données omiques, les analyses de réseaux de gènes et le développement de bases de données à usage des biologistes.

Pour le développement futur, j'envisage de poursuivre mes travaux sur l'intégration de données sur les sujets de recherche suivants :

En ce qui concerne la biologie des réseaux, plusieurs avenues doivent être explorées.

D'abord, j'envisage de porter l'analyse interactome sur le versant de la régulation. Nous disposons de bases de données de régulation (catalogues de cibles de facteurs de transcription) qui peuvent être intégrées à ITI (Voir la version HTS-Net d'ITI chapitre 5). Ensuite, il est nécessaire d'améliorer les outils de visualisation au vu du nombre croissant d'information que nous cherchons superposer et à visualiser. La mise en œuvre de logiciels de visualisation tels que *Cytoscape.js* permettraient au biologistes de vraiment interagir avec l'interface portant les réseaux tout en permettant une lecture de l'ensemble des informations liée à un gène ou une protéine ou à la nature d'une interaction. C'est l'interactivité apportée par ce type d'outil qui permettra d'apporter des réponses concrète au biologiste sur ses interrogations sur les mécanismes de régulation des gènes.

Comme elle a permis de trouver des réponses et une profonde connaissance de la biologie moléculaire des cancers du sein, la bioinformatique doit maintenant apporter des réponses à l'étude de cancers pour lesquels des études sont encore nécessaires, tels que le cancer du pancréas et les gliomes. L'étude de ces cancers pose des problèmes spécifiques en analyse de données parce que le nombre d'échantillons à disposition est faible alors que la tumorigénèse est un phénomène complexe.

Une autre voie d'analyse intéressante pour l'étude des réseaux est l'intégration de l'information des Ontologies (*Gene Ontology*) sur les interactions au niveau des gènes et facteurs de transcription. Cela implique une analyse multi-échelle qui n'est pas encore faisable par la majorité des outils actuels. Ceci dit, un type de visualisation intéressant est celle des Sankey plot, qui permettent de visualiser une hiérarchie, donc à la fois une organisation globale tout en révélant des détails de plus en plus fins.

Les analyses multi-omique ou multi-bloc prennent également tout leur sens au regard des don-

nées disponibles et commencent à être appliquées au sein des projets construits sur la plateforme. En pratique, on se heurte souvent à des limitations sur le nombre d'échantillons disponibles et sur la complexité de l'interprétation et là encore, des outils de visualisation doivent être développés.

Un autre aspect essentiel, que j'ai abordé avec le chapitre consacré au logiciel Djeen, est le développement et les applications de bases de données en biologie. La complexité des données et leur débit rend le développement de ces bases plus qu'évident. Hors, certains aspect sont loin d'être résolus, à commencer par leur conception et leur utilisation. Là encore, on demande trop souvent au bioinformaticien d'intervenir une fois les données générées, donc déjà stockées sans gestion de format ou de contraintes de cohérence alors qu'il est primordial que les développeurs soient intégrés dans le processus de développement et d'analyse en amont, et que chacun joue un rôle d'écoute réciproque pour que l'on comprenne les attentes, que l'on se familiarise avec le vocabulaire de chacun, et de la culture scientifique de chaque partie. Les décideurs pensent que le développement d'une DB est un projet purement informatique mais c'est pratiquement une expérimentation sociologique ! Par exemple, la notion de saisie et d'import de donné est est systématiquement minimisée, alors que c'est une des clés du succès. Quelle est l'utilité d'une base si on ne peu y rentrer de données ? C'est un domaine qui me passionne et je souhaite continuer à m'investir dedans. J'ai initié plusieurs projets de base de données au CRCM ont été commencés alors que je rédigeais cette HDR.

Je ne conçois pas de diriger une plateforme scientifique sans une certaine indépendance intellectuelle et financière. Cela passe par le dépôt et l'acceptation de projet auprès des agences de financement. La bioinformatique faisant plus que jamais partie intégrante de l'ensemble des projets de biologie, cela me permet d'asseoir les thématiques et directions futures que je souhaite développer.

Le futur proche consistera à trouver des partenaires qui partagent des buts et une vision scientifique commune et de répondre aux appels à projets ensemble sur des thématiques essentielles en cancérologie : intégration et visualisation de données complexes. Il est dans mon ambition de répondre à ces appels avec des biologistes, et ce, pour plusieurs raisons. D'une part, je pense que l'adoption des outils développés par les bioinformaticiens doivent être immédiatement applicables pour une application concrète. Un projet (au sens appel à projet) pose un cadre idéal pour cette application et me permet d'être dans la conception du projet comme tout autre acteur. Vu la haute complexité des analyses, il n'est pas concevable qu'un bioinformaticien ne soit associé qu'à la fin d'un projet encore une fois car les analyses ne se font pas de façon standardisée ou déconnectée de la biologie. L'équilibre pour la plateforme que je dirige est donc de répondre à des problèmes concrets posés par les biologistes de façon synergique, ce qui me permettra de continuer à recruter des ingénieurs et de former des étudiants et d'avancer sur les thématiques importantes.

Je n'ai pas parlé de l'encadrement, qui est le cœur de l'obtention d'une HDR. Ce que je retiens de mes années passées à monter et animer des projets est que le succès dépend de deux facteurs : D'une part il est crucial que tout le monde s'entende sur le plan humain, et où chaque participant apporte une compétence bien précise et complémentaire, et d'autre part, un projet qui apporte une satisfaction personnelle et qui éveille la passion, ce qui apporte motivation et indépendance à chaque participant, et donne envie d'innover et d'apporter le meilleur de soi même. C'est ce qui m'a amené à faire de l'analyse de réseaux et du développement de base de données - par opposition à faire du service "de base".

Une voie est donc de mettre à profit cette HDR pour co-encadrer des étudiants en thèse avec des équipes purement biologistes du CRCM. Cela permettrait à la fois d'être impliqué dans la génération des données en amont des projet et de former des futurs biologistes à la bioinformatique, ou en tout cas de leur permettre d'avoir les clés pour communiquer avec d'autres

bioinformaticiens et d'avoir une formation beaucoup plus transversale. Le CRCM est en train de mettre en place la technologie Single-Cell. Nul doute que les algorithmes développés dans cette HDR seront utiles pour ce type d'études.

Finalement, je voulais ajouter que j'ai eu un immense plaisir de conduire ces projets et j'espère avoir pu par ce travail, apporter le goût de la recherche aux stagiaires et étudiants que j'ai encadré dans mon équipe. C'est finalement toute la finalité de ce travail au jour le jour : la transmission des connaissances et leur restitution au travers de nouveaux algorithmes et publications.

Merci pour votre lecture.

Bibliographie

[1] José Adélaïde, Pascal Finetti, Ismahane Bekhouche, Laetitia Repellini, Jeannine Geneix, Fabrice Sircoulomb, Emmanuelle Charafe-Jauffret, Nathalie Cervera, Jérôme Desplans, Daniel Parzy, Eric Schoenmakers, Patrice Viens, Jocelyne Jacquemier, Daniel Birnbaum, François Bertucci, and Max Chaffanet. Integrated profiling of basal and luminal breast cancers. *Cancer Res*, 67(24) :11565–75, Dec 2007.

[2] Uri David Akavia, Oren Litvin, Jessica Kim, Felix Sanchez-Garcia, Dylan Kotliar, Helen C Causton, Panisa Pochanard, Eyal Mozes, Levi A Garraway, and Dana Pe'er. An integrated approach to uncover drivers of cancer. *Cell*, 143(6) :1005–17, Dec 2010.

[3] A A Alizadeh, M B Eisen, R E Davis, C Ma, I S Lossos, A Rosenwald, J C Boldrick, H Sabet, T Tran, X Yu, J I Powell, L Yang, G E Marti, T Moore, J Hudson, Jr, L Lu, D B Lewis, R Tibshirani, G Sherlock, W C Chan, T C Greiner, D D Weisenburger, J O Armitage, R Warnke, R Levy, W Wilson, M R Grever, J C Byrd, D Botstein, P O Brown, and L M Staudt. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403(6769) :503–11, Feb 2000.

[4] Sandeep S. Amberkar and Lars Kaderali. An integrative approach for a network based meta-analysis of viral RNAi screens. *Algorithms for molecular biology : AMB*, 10 :6, 2015.

[5] B Aranda, P Achuthan, Y Alam-Faruque, I Armean, A Bridge, C Derow, M Feuermann, A T Ghanbarian, S Kerrien, J Khadake, J Kerssemakers, C Leroy, M Menden, M Michaut, L Montecchi-Palazzi, S N Neuhauser, S Orchard, V Perreau, B Roechert, K van Eijk, and H Hermjakob. The IntAct molecular interaction database in 2010. *Nucleic Acids Research*, 38(Database issue) :D525–531, January 2010.

[6] M Ashburner, C A Ball, J A Blake, D Botstein, H Butler, J M Cherry, A P Davis, K Dolinski, S S Dwight, J T Eppig, M A Harris, D P Hill, L Issel-Tarver, A Kasarskis, S Lewis, J C Matese, J E Richardson, M Ringwald, G M Rubin, and G Sherlock. Gene ontology : tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1) :25–9, May 2000.

[7] Armand Bankhead, Iliana Sach, Chester Ni, Nolwenn LeMeur, Mark Kruger, Marc Ferrer, Robert Gentleman, and Carol Rohl. Knowledge based identification of essential signaling from genome-scale siRNA experiments. *BMC systems biology*, 3 :80, 2009.

[8] Frederic Bard, Laetitia Casano, Arrate Mallabiabarrena, Erin Wallace, Kota Saito, Hitoshi Kitayama, Gianni Guizzunti, Yue Hu, Franz Wendler, Ramanuj Dasgupta, Norbert Perrimon, and Vivek Malhotra. Functional genomics reveals genes involved in protein secretion and Golgi organization. *Nature*, 439(7076) :604–607, February 2006.

[9] Tanya Barrett, Dennis B Troup, Stephen E Wilhite, Pierre Ledoux, Dmitry Rudnev, Carlos Evangelista, Irene F Kim, Alexandra Soboleva, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Rolf N Muertter, and Ron Edgar. NCBI GEO :

archive for high-throughput functional genomic data. *Nucleic Acids Research*, 37(Database issue) :D885–890, January 2009.

[10] Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, Andrey Yefanov, Hyeseung Lee, Naigong Zhang, Cynthia L Robertson, Nadezhda Serova, Sean Davis, and Alexandra Soboleva. Ncbi geo : archive for functional genomics data sets–update. *Nucleic Acids Res*, 41(Database issue) :D991–5, Jan 2013.

[11] Timothy G Bath, Selcuk Bozdag, Vackar Afzal, and Daniel Crowther. LimsPortal and BonsaiLIMS : development of a lab information management system for translational medicine. *Source Code for Biology and Medicine*, 6 :9, 2011.

[12] Ismahane Bekhouche, Pascal Finetti, José Adelaïde, Anthony Ferrari, Carole Tarpin, Emmanuelle Charafe-Jauffret, Colette Charpin, Gilles Houvenaeghel, Jocelyne Jacquemier, Ghislain Bidaut, Daniel Birnbaum, Patrice Viens, Max Chaffanet, and François Bertucci. High-resolution comparative genomic hybridization of inflammatory breast cancer and identification of candidate genes. *PloS one*, 6(2) :e16950, 2011.

[13] Anna Bergamaschi, Young H Kim, Pei Wang, Therese Sørlie, Tina Hernandez-Boussard, Per E Lonning, Robert Tibshirani, Anne-Lise Børresen-Dale, and Jonathan R Pollack. Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. *Genes, chromosomes & cancer*, 45(11) :1033–1040, November 2006.

[14] Rameen Beroukhim, Jean-Philippe Brunet, Arianna Di Napoli, Kirsten D Mertz, Apryle Seeley, Maira M Pires, David Linhart, Robert A Worrell, Holger Moch, Mark A Rubin, William R Sellers, Matthew Meyerson, W Marston Linehan, William G Kaelin, Jr, and Sabina Signoretti. Patterns of gene expression and copy-number alterations in von-hippel lindau disease-associated and sporadic clear cell carcinoma of the kidney. *Cancer Res*, 69(11) :4674–81, Jun 2009.

[15] Rameen Beroukhim, Gad Getz, Leia Nghiemphu, Jordi Barretina, Teli Hsueh, David Linhart, Igor Vivanco, Jeffrey C Lee, Julie H Huang, Sethu Alexander, Jinyan Du, Tweeny Kau, Roman K Thomas, Kinjal Shah, Horacio Soto, Sven Perner, John Prensner, Ralph M Debiasi, Francesca Demichelis, Charlie Hatton, Mark A Rubin, Levi A Garraway, Stan F Nelson, Linda Liau, Paul S Mischel, Tim F Cloughesy, Matthew Meyerson, Todd A Golub, Eric S Lander, Ingo K Mellinghoff, and William R Sellers. 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*, 104(50) :20007–20012, December 2007.

[16] François Bertucci, Pascal Finetti, Nathalie Cervera, Emmanuelle Charafe-Jauffret, Max Buttarelli, Jocelyne Jacquemier, Max Chaffanet, Dominique Maraninchi, Patrice Viens, and Daniel Birnbaum. How different are luminal a and basal breast cancers ? *Int J Cancer*, 124(6) :1338–48, Mar 2009.

[17] François Bertucci, Pascal Finetti, Nathalie Cervera, Dominique Maraninchi, Patrice Viens, and Daniel Birnbaum. Gene expression profiling and clinical outcome in breast cancer. *Omics : A Journal of Integrative Biology*, 10(4) :429–443, 2006.

[18] G. Bidaut and CJ Stoeckert, Jr. Characterization of unknown adult stem cell samples by large scale data integration and artificial neural networks. *Pac Symp Biocomput*, pages 356–367, 2009.

[19] Ghislain Bidaut. Gene function inference from gene expression of deletion mutants. *Methods Mol Biol*, 408 :1–18, 2007.

[20] Ghislain Bidaut, Frank J Manion, Christophe Garcia, and Michael F Ochs. Waveread : automatic measurement of relative gene expression levels from microarrays using wavelet analysis. *J Biomed Inform*, 39(4) :379–88, Aug 2006.

[21] Ghislain Bidaut and Michael F. Ochs. Clutrfree : cluster tree visualization and interpretation. *Bioinformatics*, 20(16) :2869–2871, Nov 2004.

[22] Ghislain Bidaut and Christian J Stoeckert, Jr. Large scale transcriptome data integration across multiple tissues to decipher stem cell signatures. *Methods Enzymol*, 467 :229–245, 2009.

[23] Pierre-Alain Binz, Robert Barkovich, Ronald C Beavis, David Creasy, David M Horn, Randall K Julian, Sean L Seymour, Chris F Taylor, and Yves Vandenbrouck. Guidelines for reporting the use of mass spectrometry informatics in proteomics. *Nature Biotechnology*, 26(8) :862–862, August 2008.

[24] Laurie A. Boyer, Kathrin Plath, Julia Zeitlinger, Tobias Brambrink, Lea A. Medeiros, Tong Ihn Lee, Stuart S. Levine, Marius Wernig, Adriana Tajonar, Mridula K. Ray, George W. Bell, Arie P. Otte, Miguel Vidal, David K. Gifford, Richard A. Young, and Rudolf Jaenisch. Polycomb complexes repress developmental regulators in murine embryonic stem cells. *Nature*, 441(7091) :349–353, May 2006.

[25] A Brazma, P Hingamp, J Quackenbush, G Sherlock, P Spellman, C Stoeckert, J Aach, W Ansorge, C A Ball, H C Causton, T Gaasterland, P Glenisson, F C Holstege, I F Kim, V Markowitz, J C Matese, H Parkinson, A Robinson, U Sarkans, S Schulze-Kremer, J Stewart, R Taylor, J Vilo, and M Vingron. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nature genetics*, 29(4) :365–371, December 2001.

[26] Edward N Brody, Larry Gold, Richard M Lawn, Jeffrey J Walker, and Dom Zichi. High-content affinity-based proteomics : unlocking protein biomarker discovery. *Expert Rev Mol Diagn*, 10(8) :1013–22, Nov 2010.

[27] Kevin R Brown and Igor Jurisica. Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome biology*, 8(5) :R95, 2007.

[28] Jean-Philippe Brunet, Pablo Tamayo, Todd R Golub, and Jill P Mesirov. Metagenes and molecular pattern discovery using matrix factorization. *Proc Natl Acad Sci U S A*, 101(12) :4164–9, Mar 2004.

[29] Kevin D Bunting and Robert G Hawley. Integrative molecular and developmental biology of adult stem cells. *Biol Cell*, 95(9) :563–578, Dec 2003.

[30] Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, 455(7216) :1061–8, Oct 2008.

[31] Cancer Genome Atlas Research Network, John N Weinstein, Eric A Collisson, Gordon B Mills, Kenna R Mills Shaw, Brad A Ozenberger, Kyle Ellrott, Ilya Shmulevich, Chris Sander, and Joshua M Stuart. The cancer genome atlas pan-cancer analysis project. *Nat Genet*, 45(10) :1113–20, Oct 2013.

[32] Arnaud Ceol, Andrew Chatr Aryamontri, Luana Licata, Daniele Peluso, Leonardo Briganti, Livia Perfetto, Luisa Castagnoli, and Gianni Cesareni. MINT, the molecular interaction database : 2009 update. *Nucleic Acids Research*, 38(Database issue) :D532–539, January 2010.

[33] Emmanuelle Charafe-Jauffret, Christophe Ginestier, Florence Monville, Samira Fekairi, Jocelyne Jacquemier, Daniel Birnbaum, and François Bertucci. How to best classify breast cancer : conventional and novel classifications (review). *Int J Oncol*, 27(5) :1307–13, Nov 2005.

[34] Na-Yu Chia, Yun-Shen Chan, Bo Feng, Xinyi Lu, Yuriy L. Orlov, Dimitri Moreau, Pankaj Kumar, Lin Yang, Jianming Jiang, Mei-Sheng Lau, Mikael Huss, Boon-Seng Soh, Petra Kraus, Pin Li, Thomas Lufkin, Bing Lim, Neil D. Clarke, Frederic Bard, and Huck-Hui Ng. A genome-wide RNAi screen reveals determinants of human embryonic stem cell identity. *Nature*, 468(7321) :316–320, November 2010.

[35] 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.

[36] Eric Conway, Evan Healy, and Adrian P. Bracken. PRC2 mediated H3k27 methylations in cellular identity and cancer. *Current Opinion in Cell Biology*, 37 :42–48, October 2015.

[37] Daniela Corda, Takahiro Kudo, Pasquale Zizza, Cristiano Iurisci, Eri Kawai, Norihisa Kato, Noriyuki Yanaka, and Stefania Mariggiò. The developmentally regulated osteoblast phosphodiesterase GDE3 is glycerophosphoinositol-specific and modulates cell growth. *The Journal of Biological Chemistry*, 284(37) :24848–24856, September 2009.

[38] David R Croucher, Falko Hochgräfe, Luxi Zhang, Ling Liu, Ruth J Lyons, Danny Rickwood, Carole M Tactacan, Brigid C Browne, Naveid Ali, Howard Chan, Robert Shearer, David Gallego-Ortega, Darren N Saunders, Alexander Swarbrick, and Roger J Daly. Involvement of lyn and the atypical kinase sgk269/peak1 in a basal breast cancer signaling pathway. *Cancer Res*, 73(6) :1969–80, Mar 2013.

[39] Giuseppe Curigliano. New drugs for breast cancer subtypes : targeting driver pathways to overcome resistance. *Cancer Treat Rev*, 38(4) :303–10, Jun 2012.

[40] Christina Curtis, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M Rueda, Mark J Dunning, Doug Speed, Andy G Lynch, Shamith Samarajiwa, Yinyin Yuan, Stefan Gräf, Gavin Ha, Gholamreza Haffari, Ali Bashashati, Roslin Russell, Steven McKinney, METABRIC Group, Anita Langerød, Andrew Green, Elena Provenzano, Gordon Wishart, Sarah Pinder, Peter Watson, Florian Markowetz, Leigh Murphy, Ian Ellis, Arnie Purushotham, Anne-Lise Børresen-Dale, James D Brenton, Simon Tavaré, Carlos Caldas, and Samuel Aparicio. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403) :346–52, Apr 2012.

[41] Phuong Dao, Kendric Wang, Colin Collins, Martin Ester, Anna Lapuk, and S Cenk Sahinalp. Optimally discriminative subnetwork markers predict response to chemotherapy. *Bioinformatics*, 27(13) :i205–13, Jul 2011.

[42] Valeria De Giorgi, Alessandro Monaco, Andrea Worchech, Marialina Tornesello, Francesco Izzo, Luigi Buonaguro, Francesco M. Marincola, Ena Wang, and Franco M. Buonaguro. Gene profiling, biomarkers and pathways characterizing HCV-related hepatocellular carcinoma. *Journal of Translational Medicine*, 7 :85, October 2009.

[43] Matsuo Deguchi, Masanori Kagita, Naoko Yamashita, Takasi Nakano, Kazuko Tahara, Seishi Asari, and Yoshinori Iwatani. [Comparison of eight screening tests for ant-HCV antibody]. *Kansenshogaku Zasshi. The Journal of the Japanese Association for Infectious Diseases*, 76(9) :711–720, September 2002.

[44] Christine Desmedt, Benjamin Haibe-Kains, Pratyaksha Wirapati, Marc Buyse, Denis Larsimont, Gianluca Bontempi, Mauro Delorenzi, Martine Piccart, and Christos Sotiriou. Biological processes associated with breast cancer clinical outcome depend on the molecular

subtypes. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, 14(16) :5158–5165, August 2008.

[45] Bhaskar Dutta, Alaleh Azhir, Louis-Henri Merino, Yongjian Guo, Swetha Revanur, Piyush B. Madhamshettiwar, Ronald N. Germain, Jennifer A. Smith, Kaylene J. Simpson, Scott E. Martin, Eugen Buehler, Eugen Beuhler, and Iain D. C. Fraser. An interactive web-based application for Comprehensive Analysis of RNAi-screen Data. *Nature Communications*, 7 :10578, 2016.

[46] 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) :5923–5928, April 2006.

[47] Cheng Fan, Daniel S Oh, Lodewyk Wessels, Britta Weigelt, Dimitry S A Nuyten, Andrew B Nobel, Laura J van't Veer, and Charles M Perou. Concordance among gene-expression-based predictors for breast cancer. *The New England Journal of Medicine*, 355(6) :560–569, August 2006.

[48] Irit Fishel, Alon Kaufman, and Eytan Ruppin. Meta-analysis of gene expression data : a predictor-based approach. *Bioinformatics (Oxford, England)*, 23(13) :1599–1606, July 2007.

[49] Yukihiro Furusawa, Takaho A Endo, Yuuki Obata, Osamu Ohara, Hiroshi Ohno, and Koji Hase. Pitfalls in global normalization of chip-seq data in cd4(+) t cells treated with butyrate : A possible solution strategy. *Genom Data*, 2 :176–80, Dec 2014.

[50] Maxime Garcia, Pascal Finetti, Francois Bertucci, Daniel Birnbaum, and Ghislain Bidaut. Detection of driver protein complexes in breast cancer metastasis by large-scale transcriptome-interactome integration. *Methods Mol Biol*, 1101 :67–85, 2014.

[51] Maxime Garcia, Raphaelle Millat-Carus, François Bertucci, Pascal Finetti, Daniel Birnbaum, and Ghislain Bidaut. Interactome-transcriptome integration for predicting distant metastasis in breast cancer. *Bioinformatics*, 28(5) :672–8, Mar 2012.

[52] Maxime Garcia, Raphaelle 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 Luis Rueda, editor, *Microarray Image and Data Analysis : Theory and Practice Microarray Image and Data Analysis : Theory and Practice Microarray Image and Data Analysis : Theory and Practice*. CRC Press, 2014.

[53] Maxime 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*, pages 406–427. Hershey, New York, igi global edition, 2011.

[54] Gene Ontology Consortium. The gene ontology in 2010 : extensions and refinements. *Nucleic Acids Res*, 38(Database issue) :D331–5, Jan 2010.

[55] Ryan Gill, Somnath Datta, and Susmita Datta. A statistical framework for differential network analysis from microarray data. *BMC Bioinformatics*, 11 :95, 2010.

[56] Anthony Gitter and Ziv Bar-Joseph. Identifying proteins controlling key disease signaling pathways. *Bioinformatics (Oxford, England)*, 29(13) :i227–236, July 2013.

[57] Asli N. Goktug, Su Sien Ong, and Taosheng Chen. GUItars : a GUI tool for analysis of high-throughput RNA interference screening data. *PloS One*, 7(11) :e49386, 2012.

[58] T R Golub, D K Slonim, P Tamayo, C Huard, M Gaasenbeek, J P Mesirov, H Coller, M L Loh, J R Downing, M A Caligiuri, C D Bloomfield, and E S Lander. Molecular classification of cancer : class discovery and class prediction by gene expression monitoring. *Science*, 286(5439) :531–7, Oct 1999.

[59] Orland Gonzalez and Ralf Zimmer. Contextual analysis of RNAi-based functional screens using interaction networks. *Bioinformatics*, 27(19) :2707–2713, January 2011.

[60] Braden Greer and Javed Khan. Online analysis of microarray data using artificial neural networks. *Methods Mol Biol*, 377 :61–74, 2007.

[61] Braden T Greer and Javed Khan. Diagnostic classification of cancer using dna microarrays and artificial intelligence. *Ann N Y Acad Sci*, 1020 :49–66, May 2004.

[62] Obi L. Griffith, Stephen B. Montgomery, Bridget Bernier, Bryan Chu, Katayoon Kasaian, Stein Aerts, Shaun Mahony, Monica C. Sleumer, Mikhail Bilenky, Maximilian Haeussler, Malachi Griffith, Steven M. Gallo, Belinda Giardine, Bart Hooghe, Peter Van Loo, Enrique Blanco, Amy Ticoll, Stuart Lithwick, Elodie Portales-Casamar, Ian J. Donaldson, Gordon Robertson, Claes Wadelius, Pieter De Bleser, Dominique Vlieghe, Marc S. Halfon, Wyeth Wasserman, Ross Hardison, Casey M. Bergman, Steven J. M. Jones, and Open Regulatory Annotation Consortium. ORegAnno : an open-access community-driven resource for regulatory annotation. *Nucleic Acids Research*, 36(Database issue) :D107–113, January 2008.

[63] Jin Gu, Yang Chen, Shao Li, and Yanda Li. Identification of responsive gene modules by network-based gene clustering and extending : application to inflammation and angiogenesis. *BMC systems biology*, 4 :47, 2010.

[64] M Guedj, L Marisa, A de Reynies, B Orsetti, R Schiappa, F Bibeau, G MacGrogan, F Lerebours, P Finetti, M Longy, P Bertheau, F Bertrand, F Bonnet, A L Martin, J P Feugeas, I Bièche, J Lehmann-Che, R Lidereau, D Birnbaum, F Bertucci, H de Thé, and C Theillet. A refined molecular taxonomy of breast cancer. *Oncogene*, 31(9) :1196–206, Mar 2012.

[65] Saad Haider and Ranadip Pal. Integrated analysis of transcriptomic and proteomic data. *Curr Genomics*, 14(2) :91–110, Apr 2013.

[66] Daniel Hanisch, Alexander Zien, Ralf Zimmer, and Thomas Lengauer. Co-clustering of biological networks and gene expression data. *Bioinformatics (Oxford, England)*, 18 Suppl 1 :S145–154, 2002.

[67] Kenny Helsens, Niklaas Colaert, Harald Barsnes, Thilo Muth, Kristian Flikka, An Staes, Evy Timmerman, Steffi Wortelkamp, Albert Sickmann, Joël Vandekerckhove, Kris Gevaert, and Lennart Martens. ms_lims, a simple yet powerful open source laboratory information management system for MS-driven proteomics. *Proteomics*, 10(6) :1261–1264, March 2010.

[68] Tracy S. P. Heng, Michio W. Painter, and Immunological Genome Project Consortium. The Immunological Genome Project : networks of gene expression in immune cells. *Nature Immunology*, 9(10) :1091–1094, October 2008.

[69] Matan Hofree, John P. Shen, Hannah Carter, Andrew Gross, and Trey Ideker. Network-based stratification of tumor mutations. *Nature Methods*, 10(11) :1108–1115, November 2013.

[70] Fangxin Hong and Rainer Breitling. A comparison of meta-analysis methods for detecting differentially expressed genes in microarray experiments. *Bioinformatics (Oxford, England)*, 24(3) :374–382, February 2008.

[71] J J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proc Natl Acad Sci U S A*, 79(8) :2554–8, Apr 1982.

[72] 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, Apr 2006.

[73] Wolfgang Huber, Vincent J Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton S Carvalho, Hector Corrada Bravo, Sean Davis, Laurent Gatto, Thomas Girke, Raphael Gottardo, Florian Hahne, Kasper D Hansen, Rafael A Irizarry, Michael Lawrence, Michael I Love, James MacDonald, Valerie Obenchain, Andrzej K Oleś, Hervé Pagès, Alejandro Reyes, Paul Shannon, Gordon K Smyth, Dan Tenenbaum, Levi Waldron, and Martin Morgan. Orchestrating high-throughput genomic analysis with bioconductor. *Nat Methods*, 12(2) :115–21, Feb 2015.

[74] T R Hughes, M J Marton, A R Jones, C J Roberts, R Stoughton, C D Armour, H A Bennett, E Coffey, H Dai, Y D He, M J Kidd, A M King, M R Meyer, D Slade, P Y Lum, S B Stepaniants, D D Shoemaker, D Gachotte, K Chakraburtty, J Simon, M Bard, and S H Friend. Functional discovery via a compendium of expression profiles. *Cell*, 102 :109–126, July 2000.

[75] Carl Hull, Bruce Wray, Ford Winslow, and Mark Vilicich. Tracking and controlling everything that affects quality is the key to a quality management system. *Combinatorial Chemistry & High Throughput Screening*, 14(9) :772–780, November 2011.

[76] International Cancer Genome Consortium, Thomas J Hudson, Warwick Anderson, Axel Artez, Anna D Barker, Cindy Bell, Rosa R Bernabé, M K Bhan, Fabien Calvo, Iiro Eerola, Daniela S Gerhard, Alan Guttmacher, Mark Guyer, Fiona M Hemsley, Jennifer L Jennings, David Kerr, Peter Klatt, Patrik Kolar, Jun Kusada, David P Lane, Frank Laplace, Lu Youyong, Gerd Nettekoven, Brad Ozenberger, Jane Peterson, T S Rao, Jacques Remacle, Alan J Schafer, Tatsuhiro Shibata, Michael R Stratton, Joseph G Vockley, Koichi Watanabe, Huanming Yang, Matthew M F Yuen, Bartha M Knoppers, Martin Bobrow, Anne Cambon-Thomsen, Lynn G Dressler, Stephanie O M Dyke, Yann Joly, Kazuto Kato, Karen L Kennedy, Pilar Nicolás, Michael J Parker, Emmanuelle Rial-Sebbag, Carlos M Romeo-Casabona, Kenna M Shaw, Susan Wallace, Georgia L Wiesner, Nikolajs Zeps, Peter Lichter, Andrew V Biankin, Christian Chabannon, Lynda Chin, Bruno Clément, Enrique de Alava, Françoise Degos, Martin L Ferguson, Peter Geary, D Neil Hayes, Thomas J Hudson, Amber L Johns, Arek Kasprzyk, Hidewaki Nakagawa, Robert Penny, Miguel A Piris, Rajiv Sarin, Aldo Scarpa, Tatsuhiro Shibata, Marc van de Vijver, P Andrew Futreal, Hiroyuki Aburatani, Mónica Bayés, David D L Botwell, Peter J Campbell, Xavier Estivill, Daniela S Gerhard, Sean M Grimmond, Ivo Gut, Martin Hirst, Carlos López-Otín, Partha Majumder, Marco Marra, John D McPherson, Hidewaki Nakagawa, Zemin Ning, Xose S Puente, Yijun Ruan, Tatsuhiro Shibata, Michael R Stratton, Hendrik G Stunnenberg, Harold Swerdlow, Victor E Velculescu, Richard K Wilson, Hong H Xue, Liu Yang, Paul T Spellman, Gary D Bader, Paul C Boutros, Peter J Campbell, Paul Flicek, Gad Getz, Roderic Guigó, Guangwu Guo, David Haussler, Simon Heath, Tim J Hubbard,

Tao Jiang, Steven M Jones, Qibin Li, Nuria López-Bigas, Ruibang Luo, Lakshmi Muthuswamy, B F Francis Ouellette, John V Pearson, Xose S Puente, Victor Quesada, Benjamin J Raphael, Chris Sander, Tatsuhiro Shibata, Terence P Speed, Lincoln D Stein, Joshua M Stuart, Jon W Teague, Yasushi Totoki, Tatsuhiko Tsunoda, Alfonso Valencia, David A Wheeler, Honglong Wu, Shancen Zhao, Guangyu Zhou, Lincoln D Stein, Roderic Guigó, Tim J Hubbard, Yann Joly, Steven M Jones, Arek Kasprzyk, Mark Lathrop, Nuria López-Bigas, B F Francis Ouellette, Paul T Spellman, Jon W Teague, Gilles Thomas, Alfonso Valencia, Teruhiko Yoshida, Karen L Kennedy, Myles Axton, Stephanie O M Dyke, P Andrew Futreal, Daniela S Gerhard, Chris Gunter, Mark Guyer, Thomas J Hudson, John D McPherson, Linda J Miller, Brad Ozenberger, Kenna M Shaw, Arek Kasprzyk, Lincoln D Stein, Junjun Zhang, Syed A Haider, Jianxin Wang, Christina K Yung, Anthony Cros, Anthony Cross, Yong Liang, Saravanamuttu Gnaneshan, Jonathan Guberman, Jack Hsu, Martin Bobrow, Don R C Chalmers, Karl W Hasel, Yann Joly, Terry S H Kaan, Karen L Kennedy, Bartha M Knoppers, William W Lowrance, Tohru Masui, Pilar Nicolás, Emmanuelle Rial-Sebbag, Laura Lyman Rodriguez, Catherine Vergely, Teruhiko Yoshida, Sean M Grimmond, Andrew V Biankin, David D L Bowtell, Nicole Cloonan, Anna deFazio, James R Eshleman, Dariush Etemadmoghadam, Brooke B Gardiner, Brooke A Gardiner, James G Kench, Aldo Scarpa, Robert L Sutherland, Margaret A Tempero, Nicola J Waddell, Peter J Wilson, John D McPherson, Steve Gallinger, Ming-Sound Tsao, Patricia A Shaw, Gloria M Petersen, Debabrata Mukhopadhyay, Lynda Chin, Ronald A DePinho, Sarah Thayer, Lakshmi Muthuswamy, Kamran Shazand, Timothy Beck, Michelle Sam, Lee Timms, Vanessa Ballin, Youyong Lu, Jiafu Ji, Xiuqing Zhang, Feng Chen, Xueda Hu, Guangyu Zhou, Qi Yang, Geng Tian, Lianhai Zhang, Xiaofang Xing, Xianghong Li, Zhenggang Zhu, Yingyan Yu, Jun Yu, Huanming Yang, Mark Lathrop, Jörg Tost, Paul Brennan, Ivana Holcatova, David Zaridze, Alvis Brazma, Lars Egevard, Egor Prokhortchouk, Rosamonde Elizabeth Banks, Mathias Uhlén, Anne Cambon-Thomsen, Juris Viksna, Fredrik Ponten, Konstantin Skryabin, Michael R Stratton, P Andrew Futreal, Ewan Birney, Ake Borg, Anne-Lise Børresen-Dale, Carlos Caldas, John A Foekens, Sancha Martin, Jorge S Reis-Filho, Andrea L Richardson, Christos Sotiriou, Hendrik G Stunnenberg, Giles Thoms, Marc van de Vijver, Laura van't Veer, Fabien Calvo, Daniel Birnbaum, Hélène Blanche, Pascal Boucher, Sandrine Boyault, Christian Chabannon, Ivo Gut, Jocelyne D Masson-Jacquemier, Mark Lathrop, Iris Pauporté, Xavier Pivot, Anne Vincent-Salomon, Eric Tabone, Charles Theillet, Gilles Thomas, Jörg Tost, Isabelle Treilleux, Fabien Calvo, Paulette Bioulac-Sage, Bruno Clément, Thomas Decaens, Françoise Degos, Dominique Franco, Ivo Gut, Marta Gut, Simon Heath, Mark Lathrop, Didier Samuel, Gilles Thomas, Jessica Zucman-Rossi, Peter Lichter, Roland Eils, Benedikt Brors, Jan O Korbel, Andrey Korshunov, Pablo Landgraf, Hans Lehrach, Stefan Pfister, Bernhard Radlwimmer, Guido Reifenberger, Michael D Taylor, Christof von Kalle, Partha P Majumder, Rajiv Sarin, T S Rao, M K Bhan, Aldo Scarpa, Paolo Pederzoli, Rita A Lawlor, Massimo Delledonne, Alberto Bardelli, Andrew V Biankin, Sean M Grimmond, Thomas Gress, David Klimstra, Giuseppe Zamboni, Tatsuhiro Shibata, Yusuke Nakamura, Hidewaki Nakagawa, Jun Kusada, Tatsuhiko Tsunoda, Satoru Miyano, Hiroyuki Aburatani, Kazuto Kato, Akihiro Fujimoto, Teruhiko Yoshida, Elias Campo, Carlos López-Otín, Xavier Estivill, Roderic Guigó, Silvia de Sanjosé, Miguel A Piris, Emili Montserrat, Marcos González-Díaz, Xose S Puente, Pedro Jares, Alfonso Valencia, Heinz Himmelbauer, Heinz Himmelbaue, Victor Quesada, Silvia Bea, Michael R Stratton, P Andrew Futreal, Peter J Campbell, Anne Vincent-Salomon, Andrea L Richardson, Jorge S Reis-Filho, Marc van de Vijver, Gilles Thomas, Jocelyne D Masson-Jacquemier, Samuel Aparicio, Ake Borg, Anne-Lise Børresen-Dale, Carlos Caldas,

John A Foekens, Hendrik G Stunnenberg, Laura van't Veer, Douglas F Easton, Paul T Spellman, Sancha Martin, Anna D Barker, Lynda Chin, Francis S Collins, Carolyn C Compton, Martin L Ferguson, Daniela S Gerhard, Gad Getz, Chris Gunter, Alan Guttmacher, Mark Guyer, D Neil Hayes, Eric S Lander, Brad Ozenberger, Robert Penny, Jane Peterson, Chris Sander, Kenna M Shaw, Terence P Speed, Paul T Spellman, Joseph G Vockley, David A Wheeler, Richard K Wilson, Thomas J Hudson, Lynda Chin, Bartha M Knoppers, Eric S Lander, Peter Lichter, Lincoln D Stein, Michael R Stratton, Warwick Anderson, Anna D Barker, Cindy Bell, Martin Bobrow, Wylie Burke, Francis S Collins, Carolyn C Compton, Ronald A DePinho, Douglas F Easton, P Andrew Futreal, Daniela S Gerhard, Anthony R Green, Mark Guyer, Stanley R Hamilton, Tim J Hubbard, Olli P Kallioniemi, Karen L Kennedy, Timothy J Ley, Edison T Liu, Youyong Lu, Partha Majumder, Marco Marra, Brad Ozenberger, Jane Peterson, Alan J Schafer, Paul T Spellman, Hendrik G Stunnenberg, Brandon J Wainwright, Richard K Wilson, and Huanming Yang. International network of cancer genome projects. *Nature*, 464(7291) :993–8, Apr 2010.

[77] Rafael A Irizarry, Bridget Hobbs, Francois Collin, Yasmin D Beazer-Barclay, Kristen J Antonellis, Uwe Scherf, and Terence P Speed. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics*, 4(2) :249–64, Apr 2003.

[78] Natalia B Ivanova, John T Dimos, Christoph Schaniel, Jason A Hackney, Kateri A Moore, and Ihor R Lemischka. A stem cell molecular signature. *Science*, 298(5593) :601–604, Oct 2002.

[79] C. Jiang, Z. Xuan, F. Zhao, and M. Q. Zhang. TRED : a transcriptional regulatory element database, new entries and other development. *Nucleic Acids Research*, 35(Database issue) :D137–D140, January 2007.

[80] Kumaran Kandasamy, Shivakumar Keerthikumar, Renu Goel, Suresh Mathivanan, Nandini Patankar, Beema Shafreen, Santosh Renuse, Harsh Pawar, Y. L. Ramachandra, Pradip Kumar Acharya, Prathibha Ranganathan, Raghothama Chaerkady, T. S. Keshava Prasad, and Akhilesh Pandey. Human Proteinpedia : a unified discovery resource for proteomics research. *Nucleic Acids Research*, 37(Database issue) :D773–781, January 2009.

[81] Irene M Kaplow, Rohit Singh, Adam Friedman, Chris Bakal, Norbert Perrimon, and Bonnie Berger. Rnaicut : automated detection of significant genes from functional genomic screens. *Nat Methods*, 6(7) :476–7, Jul 2009.

[82] Misha Kapushesky, Ibrahim Emam, Ele Holloway, Pavel Kurnosov, Andrey Zorin, James Malone, Gabriella Rustici, Eleanor Williams, Helen Parkinson, and Alvis Brazma. Gene expression atlas at the european bioinformatics institute. *Nucleic Acids Res*, 38(Database issue) :D690–8, Jan 2010.

[83] Madhumohan R Katika and Antoni Hurtado. A functional link between foxa1 and breast cancer snps. *Breast Cancer Res*, 15(1) :303, Feb 2013.

[84] Masuko Katoh, Maki Igarashi, Hirokazu Fukuda, Hitoshi Nakagama, and Masaru Katoh. Cancer genetics and genomics of human fox family genes. *Cancer Lett*, 328(2) :198–206, Jan 2013.

[85] 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) :D841–846, January 2012.

[86] 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) :D767–772, January 2009.

[87] J Khan, J S Wei, M Ringnér, L H Saal, M Ladanyi, F Westermann, F Berthold, M Schwab, C R Antonescu, C Peterson, and P S Meltzer. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med*, 7(6) :673–9, Jun 2001.

[88] Patrick J Killion and Vishwanath R Iyer. Microarray data visualization and analysis with the Longhorn Array Database (LAD). *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]*, Chapter 7 :Unit 7.10, December 2004.

[89] Nikolay Kolesnikov, Emma Hastings, Maria Keays, Olga Melnichuk, Y Amy Tang, Eleanor Williams, Miroslaw Dylag, Natalja Kurbatova, Marco Brandizi, Tony Burdett, Karyn Megy, Ekaterina Pilicheva, Gabriella Rustici, Andrew Tikhonov, Helen Parkinson, Robert Petryszak, Ugis Sarkans, and Alvis Brazma. Arrayexpress update–simplifying data submissions. *Nucleic Acids Res*, 43(Database issue) :D1113–6, Jan 2015.

[90] Andrew V Kossenkov and Michael F Ochs. Matrix factorization for recovery of biological processes from microarray data. *Methods Enzymol*, 467 :59–77, 2009.

[91] Homin K Lee, William Braynen, Kiran Keshav, and Paul Pavlidis. Erminej : tool for functional analysis of gene expression data sets. *BMC Bioinformatics*, 6 :269, Nov 2005.

[92] Jamie A Lee, Josef Spidlen, Keith Boyce, Jennifer Cai, Nicholas Crosbie, Mark Dalphin, Jeff Furlong, Maura Gasparetto, Michael Goldberg, Elizabeth M Goralczyk, Bill Hyun, Kirstin Jansen, Tobias Kollmann, Megan Kong, Robert Leif, Shannon McWeeney, Thomas D Moloshok, Wayne Moore, Garry Nolan, John Nolan, Janko Nikolich-Zugich, David Parrish, Barclay Purcell, Yu Qian, Biruntha Selvaraj, Clayton Smith, Olga Tchuvatkina, Anne Wertheimer, Peter Wilkinson, Christopher Wilson, James Wood, Robert Zigon, International Society for Advancement of Cytometry Data Standards Task Force, Richard H Scheuermann, and Ryan R Brinkman. Miflowcyt : the minimum information about a flow cytometry experiment. *Cytometry A*, 73(10) :926–30, Oct 2008.

[93] A. Leff and J.T. Rayfield. Web-application development using the Model/View/Controller design pattern. In *Enterprise Distributed Object Computing Conference, 2001. EDOC '01. Proceedings. Fifth IEEE International*, pages 118 –127, 2001.

[94] Mark D. M. Leiserson, Fabio Vandin, Hsin-Ta Wu, Jason R. Dobson, Jonathan V. Eldridge, Jacob L. Thomas, Alexandra Papoutsaki, Younhun Kim, Beifang Niu, Michael McLellan, Michael S. Lawrence, Abel Gonzalez-Perez, David Tamborero, Yuwei Cheng, Gregory A. Ryslik, Nuria Lopez-Bigas, Gad Getz, Li Ding, and Benjamin J. Raphael. Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature Genetics*, 47(2) :106–114, February 2015.

[95] Luana Licata, Leonardo Briganti, Daniele Peluso, Livia Perfetto, Marta Iannuccelli, Eugenia Galeota, Francesca Sacco, Anita Palma, Aurelio Pio Nardozza, Elena Santonico, Luisa Castagnoli, and Gianni Cesareni. MINT, the molecular interaction database : 2012 update. *Nucleic Acids Research*, 40(Database issue) :D857–861, January 2012.

[96] D Lin, Z Shkedy, T Burzykowski, R Ion, H W H Göhlmann, A De Bondt, T Perer, T Geerts, I Van den Wyngaert, and L Bijnens. An investigation on performance of Significance Analysis of Microarray (SAM) for the comparisons of several treatments with one control in the presence of small-variance genes. *Biometrical journal. Biometrische Zeitschrift*, 50(5) :801–823, October 2008.

[97] Ke Lin, Harrie Kools, Philip J de Groot, Anand K Gavai, Ram K Basnet, Feng Cheng, Jian Wu, Xiaowu Wang, Arjen Lommen, Guido J E J Hooiveld, Guusje Bonnema, Richard G F Visser, Michael R Muller, and Jack A M Leunissen. MADMAX - Management and analysis database for multiple ˜omics experiments. *Journal of Integrative Bioinformatics*, 8(2) :160, 2011.

[98] Beate C Litzenburger, Chad J Creighton, Anna Tsimelzon, Bonita T Chan, Susan G Hilsenbeck, Tao Wang, Joan M Carboni, Marco M Gottardis, Fei Huang, Jenny C Chang, Michael T Lewis, Mothaffar F Rimawi, and Adrian V Lee. High IGF-IR activity in triple-negative breast cancer cell lines and tumorgrafts correlates with sensitivity to anti-IGF-IR therapy. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, 17(8) :2314–2327, April 2011.

[99] Henry Lu, Weiqun Li, William Stafford Noble, Donald Payan, and D. C. Anderson. Riboproteomics of the hepatitis C virus internal ribosomal entry site. *Journal of Proteome Research*, 3(5) :949–957, October 2004.

[100] Richard Marcotte, Azin Sayad, Kevin R Brown, Felix Sanchez-Garcia, Jüri Reimand, Maliha Haider, Carl Virtanen, James E Bradner, Gary D Bader, Gordon B Mills, Dana Pe'er, Jason Moffat, and Benjamin G Neel. Functional genomic landscape of human breast cancer drivers, vulnerabilities, and resistance. *Cell*, 164(1-2) :293–309, Jan 2016.

[101] V Matys, O V Kel-Margoulis, E Fricke, I Liebich, S Land, A Barre-Dirrie, I Reuter, D Chekmenev, M Krull, K Hornischer, N Voss, P Stegmaier, B Lewicki-Potapov, H Saxel, A E Kel, and E Wingender. Transfac and its module transcompel : transcriptional gene regulation in eukaryotes. *Nucleic Acids Res*, 34(Database issue) :D108–10, Jan 2006.

[102] Alexandre Melo, Alessandra Faria-Campos, Daiane DeLaat, Rodrigo Keller, Vinícius Abreu, and Sérgio Campos. SIGLa : an adaptable LIMS for multiple laboratories. *BMC Genomics*, 11(Suppl 5) :S8, 2010.

[103] Rebecca Yu Miao, Yvette Drabsch, Ryan Stanley Cross, Dane Cheasley, Sandra Carpinteri, Lloyd Pereira, Jordane Malaterre, Thomas J Gonda, Robin L Anderson, and Robert G Ramsay. Myb is essential for mammary tumorigenesis. *Cancer Res*, 71(22) :7029–37, Nov 2011.

[104] Stefan Michiels, Serge Koscielny, and Catherine Hill. Prediction of cancer outcome with microarrays : a multiple random validation strategy. *Lancet*, 365(9458) :488–492, February 2005.

[105] Stefan Michiels, Andrew Kramar, and Serge Koscielny. Multidimensionality of microarrays : statistical challenges and (im)possible solutions. *Mol Oncol*, 5(2) :190–6, Apr 2011.

[106] Jason C Mills, Niklas Andersson, Chieu V Hong, Thaddeus S Stappenbeck, and Jeffrey I Gordon. Molecular characterization of mouse gastric epithelial progenitor cells. *Proc Natl Acad Sci U S A*, 99(23) :14819–14824, Nov 2002.

[107] Apratim Mitra and Jiuzhou Song. Waveseq : a novel data-driven method of detecting histone modification enrichments using wavelets. *PLoS One*, 7(9) :e45486, 2012.

[108] Stephanie E. Mohr and Norbert Perrimon. RNAi screening : new approaches, understandings, and organisms. *Wiley interdisciplinary reviews. RNA*, 3(2) :145–158, April 2012.

[109] Stephanie E. Mohr, Jennifer A. Smith, Caroline E. Shamu, Ralph A. Neumüller, and Norbert Perrimon. RNAi screening comes of age : improved techniques and complementary approaches. *Nature Reviews. Molecular Cell Biology*, 15(9) :591–600, September 2014.

[110] Anna V Molofsky, Ricardo Pardal, and Sean J Morrison. Diverse mechanisms regulate stem cell self-renewal. *Curr Opin Cell Biol*, 16(6) :700–707, Dec 2004.

[111] T D Moloshok, R R Klevecz, J D Grant, F J Manion, W F Speier, 4th, and M F Ochs. Application of bayesian decomposition for analysing microarray data. *Bioinformatics*, 18(4) :566–75, Apr 2002.

[112] Sean D Mooney and Peter H Baenziger. Extensible open source content management systems and frameworks : a solution for many needs of a bioinformatics group. *Briefings in Bioinformatics*, 9(1) :69–74, January 2008.

[113] Vamsi K Mootha, Cecilia M Lindgren, Karl-Fredrik Eriksson, Aravind Subramanian, Smita Sihag, Joseph Lehar, Pere Puigserver, Emma Carlsson, Martin Ridderstråle, Esa Laurila, Nicholas Houstis, Mark J Daly, Nick Patterson, Jill P Mesirov, Todd R Golub, Pablo Tamayo, Bruce Spiegelman, Eric S Lander, Joel N Hirschhorn, David Altshuler, and Leif C Groop. Pgc-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet*, 34(3) :267–73, Jul 2003.

[114] John-Patrick Mpindi, Swapnil Potdar, Dmitrii Bychkov, Jani Saarela, Khalid Saeed, Krister Wennerberg, Tero Aittokallio, Päivi Östling, and Olli Kallioniemi. Impact of normalization methods on high-throughput screening data with high hit-rates and drug testing with dose-response data. *Bioinformatics (Oxford, England)*, August 2015.

[115] T. M. Murali, Matthew D. Dyer, David Badger, Brett M. Tyler, and Michael G. Katze. Network-based prediction and analysis of HIV dependency factors. *PLoS computational biology*, 7(9) :e1002164, September 2011.

[116] L. J. Murray, E. Bruno, N. Uchida, R. Hoffman, R. Nayar, E. L. Yeo, A. C. Schuh, and D. R. Sutherland. Cd109 is expressed on a subpopulation of cd34+ cells enriched in hematopoietic stem and progenitor cells. *Exp Hematol*, 27(8) :1282–1294, Aug 1999.

[117] Sahiti Myneni and Vimla L Patel. Organization of Biomedical Data for Collaborative Scientific Research : A Research Information Management System. *International Journal of Information Management*, 30(3) :256–264, June 2010.

[118] Alena Myšičková and Martin Vingron. Detection of interacting transcription factors in human tissues using predicted DNA binding affinity. *BMC genomics*, 13 Suppl 1 :S2, 2012.

[119] NCBI Resource Coordinators . Database resources of the national center for biotechnology information. *Nucleic Acids Res*, 46(D1) :D8–D13, Jan 2018.

[120] NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 44(D1) :D7–D19, January 2016.

[121] NCBI Resource Coordinators. Database resources of the national center for biotechnology information. *Nucleic Acids Res*, 45(D1) :D12–D17, 01 2017.

[122] Bertrand Néron, Hervé Ménager, Corinne Maufrais, Nicolas Joly, Julien Maupetit, Sébastien Letort, Sébastien Carrere, Pierre Tuffery, and Catherine Letondal. Mobyle : a new full web bioinformatics framework. *Bioinformatics (Oxford, England)*, 25(22) :3005–3011, November 2009.

[123] Y-H Nicole Tsang, X-W Wu, J-S Lim, C Wee Ong, M Salto-Tellez, K Ito, Y Ito, and L-F Chen. Prolyl isomerase pin1 downregulates tumor suppressor runx3 in breast cancer. *Oncogene*, 32(12) :1488–96, Mar 2013.

[124] Michael F Ochs, Lori Rink, Chi Tarn, Sarah Mburu, Takahiro Taguchi, Burton Eisenberg, and Andrew K Godwin. Detection of treatment-induced changes in signaling pathways in gastrointestinal stromal tumors using transcriptomic data. *Cancer Res*, 69(23) :9125–32, Dec 2009.

[125] Scott A Ochsner, Hélène Strick-Marchand, Qiong Qiu, Susan Venable, Adam Dean, Margaret Wilde, Mary C Weiss, and Gretchen J Darlington. Transcriptional profiling of bipotential embryonic liver cells to identify liver progenitor cell surface markers. *Stem Cells*, 25(10) :2476–87, Oct 2007.

[126] Adam B Olshen, E S Venkatraman, Robert Lucito, and Michael Wigler. Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, 5(4) :557–72, Oct 2004.

[127] David A Orlando, Mei Wei Chen, Victoria E Brown, Snehakumari Solanki, Yoon J Choi, Eric R Olson, Christian C Fritz, James E Bradner, and Matthew G Guenther. Quantitative chip-seq normalization reveals global modulation of the epigenome. *Cell Rep*, 9(3) :1163–70, Nov 2014.

[128] Asa J Oudes, Dave S Campbell, Carrie M Sorensen, Laura S Walashek, Lawrence D True, and Alvin Y Liu. Transcriptomes of human prostate cells. *BMC Genomics*, 7 :92, 2006.

[129] Xinshou Ouyang, Minoru Fujimoto, Reiko Nakagawa, Satoshi Serada, Toshio Tanaka, Shintaro Nomura, Ichiro Kawase, Tadamitsu Kishimoto, and Tetsuji Naka. Socs-2 interferes with myotube formation and potentiates osteoblast differentiation through upregulation of junb in c2c12 cells. *J Cell Physiol*, 207(2) :428–436, May 2006.

[130] Laetitia Padovani, Carole Colin, Carla Fernandez, Andre Maues de Paula, Sandy Mercurio, Didier Scavarda, Frederic Frassineti, Jose Adélaïde, Anderson Loundou, Dominique Intagliata, Corinne Bouvier, Gabriel Lena, Daniel Birnbaum, Nadine Girard, and Dominique Figarella-Branger. Search for distinctive markers in dnt and cortical grade ii glioma in children : same clinicopathological and molecular entities ? *Curr Top Med Chem*, 12(15) :1683–92, 2012.

[131] Qiliu Peng, Yu Lu, Xianjun Lao, Zhiping Chen, Ruolin Li, Jingzhe Sui, Xue Qin, and Shan Li. The nqo1 pro187ser polymorphism and breast cancer susceptibility : evidence from an updated meta-analysis. *Diagn Pathol*, 9 :100, May 2014.

[132] C M Perou, T Sørlie, M B Eisen, M van de Rijn, S S Jeffrey, C A Rees, J R Pollack, D T Ross, H Johnsen, L A Akslen, O Fluge, A Pergamenschikov, C Williams, S X Zhu, P E Lønning, A L Børresen-Dale, P O Brown, and D Botstein. Molecular portraits of human breast tumours. *Nature*, 406(6797) :747–752, August 2000.

[133] Smitha Pillai and Srikumar P Chellappan. Chip on chip and chip-seq assays : genome-wide analysis of transcription factor binding and histone modifications. *Methods Mol Biol*, 1288 :447–72, 2015.

[134] A. C. Piscaglia, T. Shupe, A. Gasbarrini, and B. E. Petersen. Microarray rna/dna in different stem cell lines. *Curr Pharm Biotechnol*, 8(3) :167–175, Jun 2007.

[135] Elodie Portales-Casamar, David Arenillas, Jonathan Lim, Magdalena I. Swanson, Steven Jiang, Anthony McCallum, Stefan Kirov, and Wyeth W. Wasserman. The PAZAR database of gene regulatory information coupled to the ORCA toolkit for the study of regulatory sequences. *Nucleic Acids Research*, 37(Database issue) :D54–60, January 2009.

[136] Aleix Prat, Joel S Parker, Olga Karginova, Cheng Fan, Chad Livasy, Jason I Herschkowitz, Xiaping He, and Charles M Perou. Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res*, 12(5) :R68, 2010.

[137] J Quackenbush. Computational analysis of microarray data. *Nat Rev Genet*, 2(6) :418–27, Jun 2001.

[138] 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) :R40, 2005.

[139] Mahendra Rao. Conserved and divergent paths that regulate self-renewal in mouse and human embryonic stem cells. *Dev Biol*, 275(2) :269–286, Nov 2004.

[140] Franck Rapaport, Andrei Zinovyev, Marie Dutreix, Emmanuel Barillot, and Jean-Philippe Vert. Classification of microarray data using gene networks. *BMC Bioinformatics*, 8 :35, 2007.

[141] Timothy Ravasi, Harukazu Suzuki, Carlo Vittorio Cannistraci, Shintaro Katayama, Vladimir B. Bajic, Kai Tan, Altuna Akalin, Sebastian Schmeier, Mutsumi Kanamori-Katayama, Nicolas Bertin, Piero Carninci, Carsten O. Daub, Alistair R. R. Forrest, Julian Gough, Sean Grimmond, Jung-Hoon Han, Takehiro Hashimoto, Winston Hide, Oliver Hofmann, Atanas Kamburov, Mandeep Kaur, Hideya Kawaji, Atsutaka Kubosaki, Timo Lassmann, Erik van Nimwegen, Cameron Ross MacPherson, Chihiro Ogawa, Aleksandar Radovanovic, Ariel Schwartz, Rohan D. Teasdale, Jesper Tegnér, Boris Lenhard, Sarah A. Teichmann, Takahiro Arakawa, Noriko Ninomiya, Kayoko Murakami, Michihira Tagami, Shiro Fukuda, Kengo Imamura, Chikatoshi Kai, Ryoko Ishihara, Yayoi Kitazume, Jun Kawai, David A. Hume, Trey Ideker, and Yoshihide Hayashizaki. An atlas of combinatorial transcriptional regulation in mouse and man. *Cell*, 140(5) :744–752, March 2010.

[142] Jüri Reimand, Tambet Arak, and Jaak Vilo. g :Profiler–a web server for functional interpretation of gene lists (2011 update). *Nucleic Acids Research*, 39(Web Server issue) :W307–315, July 2011.

[143] William C Reinhold, Margot Sunshine, Hongfang Liu, Sudhir Varma, Kurt W Kohn, Joel Morris, James Doroshow, and Yves Pommier. Cellminer : a web-based suite of genomic and pharmacologic tools to explore transcript and drug patterns in the nci-60 cell line set. *Cancer Res*, 72(14) :3499–511, Jul 2012.

[144] Fabien Reyal, Nicolas Stransky, Isabelle Bernard-Pierrot, Anne Vincent-Salomon, Yann de Rycke, Paul Elvin, Andrew Cassidy, Alexander Graham, Carolyn Spraggon, Yoann Désille, Alain Fourquet, Claude Nos, Pierre Pouillart, Henri Magdelénat, Dominique Stoppa-Lyonnet, Jérôme Couturier, Brigitte Sigal-Zafrani, Bernard Asselain, Xavier Sastre-Garau, Olivier Delattre, Jean Paul Thiery, and François Radvanyi. 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, February 2005.

[145] Claire Rioualen, Quentin Da Costa, Bernard Chetrit, Emmanuelle Charafe-Jauffret, Christophe Ginestier, and Ghislain Bidaut. Hts-net : An integrated regulome-interactome approach for establishing network regulation models in high-throughput screenings. *PLoS One*, 12(9) :e0185400, 2017.

[146] Isabelle Rivals, Léon Personnaz, Lieng Taing, and Marie-Claude Potier. Enrichment or depletion of a go category within a class of genes : which test ? *Bioinformatics*, 23(4) :401–7, Feb 2007.

[147] Patrick J Roberts, John E Bisi, Jay C Strum, Austin J Combest, David B Darr, Jerry E Usary, William C Zamboni, Kwok-Kin Wong, Charles M Perou, and Norman E Sharpless. Multiple roles of cyclin-dependent kinase 4/6 inhibitors in cancer therapy. *J Natl Cancer Inst*, 104(6) :476–87, Mar 2012.

[148] Diana Russom, Amira Ahmed, Nancy Gonzalez, Joseph Alvarnas, and David DiGiusto. Implementation of a configurable laboratory information management system for use in cellular process development and manufacturing. *Cytotherapy*, 14(1) :114–121, January 2012.

[149] Alessandra Rustighi, Alessandro Zannini, Luca Tiberi, Roberta Sommaggio, Silvano Piazza, Giovanni Sorrentino, Simona Nuzzo, Antonella Tuscano, Vincenzo Eterno, Federica Benvenuti, Libero Santarpia, Iannis Aifantis, Antonio Rosato, Silvio Bicciato, Alberto Zambelli, and Giannino Del Sal. Prolyl-isomerase pin1 controls normal and cancer stem cells of the breast. *EMBO Mol Med*, 6(1) :99–119, 01 2014.

[150] Lukasz Salwinski, Christopher S Miller, Adam J Smith, Frank K Pettit, James U Bowie, and David Eisenberg. The Database of Interacting Proteins : 2004 update. *Nucleic Acids Research*, 32(Database issue) :D449–451, January 2004.

[151] L Marie Scearce, John E. Brestelli, Shannon K. McWeeney, Catherine S. Lee, Joan Mazzarelli, Deborah F. Pinney, Angel Pizarro, Christian J Stoeckert, Jr, Sandra W. Clifton, M Alan Permutt, Juliana Brown, Douglas A. Melton, and Klaus H. Kaestner. Functional genomics of the endocrine pancreas : the pancreas clone set and pancchip, new resources for diabetes research. *Diabetes*, 51(7) :1997–2004, Jul 2002.

[152] Karen L Seeberger, Jannette M Dufour, Andrew M James Shapiro, Jonathan R T Lakey, Ray V Rajotte, and Gregory S Korbutt. Expansion of mesenchymal stem cells from human pancreatic ductal epithelium. *Lab Invest*, 86(2) :141–153, Feb 2006.

[153] Eran Segal, Michael Shapira, Aviv Regev, Dana Pe'er, David Botstein, Daphne Koller, and Nir Friedman. Module networks : identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet*, 34(2) :166–76, Jun 2003.

[154] T Sørlie, C M Perou, R Tibshirani, T Aas, S Geisler, H Johnsen, T Hastie, M B Eisen, M van de Rijn, S S Jeffrey, T Thorsen, H Quist, J C Matese, P O Brown, D Botstein, P Eystein Lønning, and A L 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) :10869–10874, September 2001.

[155] T Sørlie, C M Perou, R Tibshirani, T Aas, S Geisler, H Johnsen, T Hastie, M B Eisen, M van de Rijn, S S Jeffrey, T Thorsen, H Quist, J C Matese, P O Brown, D Botstein, P E Lønning, and A L Børresen-Dale. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A*, 98(19) :10869–74, Sep 2001.

[156] Christos Sotiriou, Soek-Ying Neo, Lisa M McShane, Edward L Korn, Philip M Long, Amir Jazaeri, Philippe Martiat, Steve B Fox, Adrian L Harris, and Edison T Liu. 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*, 100(18) :10393–10398, September 2003.

[157] 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 Nuyten, Marc Buyse, Marc J Van de Vijver, Jonas Bergh, Martine 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) :262–272, February 2006.

[158] Olivier Stahl, Hugo Duvergey, Arnaud Guille, Fanny Blondin, Alexandre Del Vecchio, Pascal Finetti, Samuel Granjeaud, Oana Vigy, and Ghislain Bidaut. Djeen (database for joomla !'s extensible engine) : a research information management system for flexible multi-technology project administration. *BMC Res Notes*, 6 :223, Jun 2013.

[159] Aravind Subramanian, Heidi Kuehn, Joshua Gould, Pablo Tamayo, and Jill P Mesirov. Gsea-p : a desktop application for gene set enrichment analysis. *Bioinformatics*, 23(23) :3251–3, Dec 2007.

[160] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, and Jill P Mesirov. Gene set enrichment analysis : a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, 102(43) :15545–50, Oct 2005.

[161] Zihan Sun, Yuling Cui, Jing Pei, and Zhiqiang Fan. Association between nqo1 c609t polymorphism and prostate cancer risk. *Tumour Biol*, 35(8) :7993–8, Aug 2014.

[162] Damian Szklarczyk, Andrea Franceschini, Stefan Wyder, Kristoffer Forslund, Davide Heller, Jaime Huerta-Cepas, Milan Simonovic, Alexander Roth, Alberto Santos, Kalliopi P. Tsafou, Michael Kuhn, Peer Bork, Lars J. Jensen, and Christian von Mering. STRING v10 : protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Research*, 43(Database issue) :D447–452, January 2015.

[163] Shuhei Taguwa, Toru Okamoto, Takayuki Abe, Yoshio Mori, Tetsuro Suzuki, Kohji Moriishi, and Yoshiharu Matsuura. Human butyrate-induced transcript 1 interacts with hepatitis C virus NS5a and regulates viral replication. *Journal of Virology*, 82(6) :2631–2641, March 2008.

[164] Andrew W. Tai, Yair Benita, Lee F. Peng, Sun-Suk Kim, Naoya Sakamoto, Ramnik J. Xavier, and Raymond T. Chung. A Functional Genomic Screen Identifies Cellular Cofactors of Hepatitis C Virus Replication. *Cell Host & Microbe*, 5(3) :298–307, March 2009.

[165] David Tai, Rathnam Chaguturu, and Jianwen Fang. K-Screen : a free application for high throughput screening data analysis, visualization, and laboratory information management. *Combinatorial Chemistry & High Throughput Screening*, 14(9) :757–765, November 2011.

[166] P Tamayo, D Slonim, J Mesirov, Q Zhu, S Kitareewan, E Dmitrovsky, E S Lander, and T R Golub. Interpreting patterns of gene expression with self-organizing maps : methods and application to hematopoietic differentiation. *Proc Natl Acad Sci U S A*, 96(6) :2907–12, Mar 1999.

[167] David S P Tan and Jorge S Reis-Filho. Comparative genomic hybridisation arrays : high-throughput tools to determine targeted therapy in breast cancer. *Pathobiology*, 75(2) :63–74, 2008.

[168] S Tavazoie, J D Hughes, M J Campbell, R J Cho, and G M Church. Systematic determination of genetic network architecture. *Nat Genet*, 22(3) :281–5, Jul 1999.

[169] Chris F Taylor, Dawn Field, Susanna-Assunta Sansone, Jan Aerts, Rolf Apweiler, Michael Ashburner, Catherine A Ball, Pierre-Alain Binz, Molly Bogue, Tim Booth, Alvis Brazma,

Ryan R Brinkman, Adam Michael Clark, Eric W Deutsch, Oliver Fiehn, Jennifer Fostel, Peter Ghazal, Frank Gibson, Tanya Gray, Graeme Grimes, John M Hancock, Nigel W Hardy, Henning Hermjakob, Randall K Julian, Matthew Kane, Carsten Kettner, Christopher Kinsinger, Eugene Kolker, Martin Kuiper, Nicolas Le NovËre, Jim Leebens-Mack, Suzanna E Lewis, Phillip Lord, Ann-Marie Mallon, Nishanth Marthandan, Hiroshi Masuya, Ruth McNally, Alexander Mehrle, Norman Morrison, Sandra Orchard, John Quackenbush, James M Reecy, Donald G Robertson, Philippe Rocca-Serra, Henry Rodriguez, Heiko Rosenfelder, Javier Santoyo-Lopez, Richard H Scheuermann, Daniel Schober, Barry Smith, Jason Snape, Christian J Stoeckert, Keith Tipton, Peter Sterk, Andreas Untergasser, Jo Vandesompele, and Stefan Wiemann. Promoting coherent minimum reporting guidelines for biological and biomedical investigations : the mibbi project. *Nature biotechnology*, 26 :889–896, August 2008.

[170] J. A. Thomson, J. Itskovitz-Eldor, S. S. Shapiro, M. A. Waknitz, J. J. Swiergiel, V. S. Marshall, and J. M. Jones. Embryonic stem cell lines derived from human blastocysts. *Science (New York, N.Y.)*, 282(5391) :1145–1147, November 1998.

[171] J Tsai, R Sultana, Y Lee, G Pertea, S Karamycheva, V Antonescu, J Cho, B Parvizi, F Cheung, and J Quackenbush. Resourcerer : a database for annotating and linking microarray resources within and across species. *Genome Biol*, 2(11) :SOFTWARE0002, 2001.

[172] Nurcan Tuncbag, Sara J. C. Gosline, Amanda Kedaigle, Anthony R. Soltis, Anthony Gitter, and Ernest Fraenkel. Network-Based Interpretation of Diverse High-Throughput Datasets through the Omics Integrator Software Package. *PLoS computational biology*, 12(4) :e1004879, April 2016.

[173] Ann M. Turnley. Role of socs2 in growth hormone actions. *Trends Endocrinol Metab*, 16(2) :53–58, Mar 2005.

[174] Johan Vallon-Christersson, Nicklas Nordborg, Martin Svensson, and Jari Häkkinen. BASE–2nd generation software for microarray data management and analysis. *BMC Bioinformatics*, 10 :330, 2009.

[175] 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. *The New England Journal of Medicine*, 347(25) :1999–2009, December 2002.

[176] 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*, 188(4-5) :405–9, Jun 1992.

[177] Laura J van 't Veer, Hongyue Dai, Marc J van de Vijver, Yudong D He, Augustinus A M Hart, René Bernards, and Stephen H Friend. Expression profiling predicts outcome in breast cancer. *Breast Cancer Res*, 5(1) :57–8, 2003.

[178] 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, Ron M Kerkhoven, Chris Roberts, Peter S Linsley, René Bernards, and Stephen H Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871) :530–6, Jan 2002.

[179] Pedro Vizán, Malte Beringer, Cecilia Ballaré, and Luciano Di Croce. Role of PRC2-associated factors in stem cells and disease. *FEBS Journal*, 282(9) :1723–1735, May 2015.

[180] Queenie P. Vong, Wai-Hang Leung, Jim Houston, Ying Li, Barbara Rooney, Martha Holladay, Robert A. J. Oostendorp, and Wing Leung. TOX2 regulates human natural killer cell development by controlling T-BET expression. *Blood*, 124(26) :3905–3913, December 2014.

[181] Li Wang, Zhidong Tu, and Fengzhu Sun. A network-based integrative approach to prioritize reliable hits from multiple genome-wide RNAi screens in Drosophila. *BMC genomics*, 10 :220, 2009.

[182] 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 M 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) :671–679, February 2005.

[183] Jennifer L. Wilson, Simona Dalin, Sara Gosline, Michael Hemann, Ernest Fraenkel, and Douglas A. Lauffenburger. Pathway-based network modeling finds hidden genes in shRNA screen for regulators of acute lymphoblastic leukemia. *Integrative Biology : Quantitative Biosciences from Nano to Macro*, 8(7) :761–774, July 2016.

[184] Jonas Wolf, Dyah Laksmi Dewi, Johannes Fredebohm, Karin Müller-Decker, Christa Flechtenmacher, Jörg D. Hoheisel, and Michael Boettcher. A mammosphere formation RNAi screen reveals that ATG4a promotes a breast cancer stem-like phenotype. *Breast cancer research : BCR*, 15(6) :R109, 2013.

[185] David J Wong, Eran Segal, and Howard Y Chang. Stemness, cancer and cancer stem cells. *Cell Cycle*, 7(23) :3622–4, Dec 2008.

[186] 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 (Oxford, England)*, 21(20) :3905–3911, October 2005.

[187] Jianfei Xue, Xia Lin, Wen-Tai Chiu, Yao-Hui Chen, Guanzhen Yu, Mingguang Liu, Xin-Hua Feng, Raymond Sawaya, René H Medema, Mien-Chie Hung, and Suyun Huang. Sustained activation of smad3/smad4 by foxm1 promotes tgf-beta-dependent cancer metastasis. *J Clin Invest*, 124(2) :564–79, Feb 2014.

[188] Chuanping Yang and Hairong Wei. Designing microarray and rna-seq experiments for greater systems biology discovery in modern plant genomics. *Mol Plant*, 8(2) :196–206, Feb 2015.

[189] K. H. Young. Yeast two-hybrid : so many interactions, (in) so little time... *Biology of Reproduction*, 58(2) :302–311, January 1998.

[190] Xueping Yu, Jimmy Lin, Donald J. Zack, and Jiang Qian. Computational analysis of tissue-specific combinatorial gene regulation : predicting interaction between transcription factors in human tissues. *Nucleic Acids Research*, 34(17) :4925–4936, 2006.

[191] Xiaohua Douglas Zhang. A pair of new statistical parameters for quality control in RNA interference high-throughput screening assays. *Genomics*, 89(4) :552–561, April 2007.

[192] Xiaohua Douglas Zhang, Xiting Cindy Yang, Namjin Chung, Adam Gates, Erica Stec, Priya Kunapuli, Dan J. Holder, Marc Ferrer, and Amy S. Espeseth. Robust statistical methods for hit selection in RNA interference high-throughput screening experiments. *Pharmacogenomics*, 7(3) :299–309, April 2006.

[193] Yong Zhang, Tao Liu, Clifford A Meyer, Jérôme Eeckhoute, David S Johnson, Bradley E Bernstein, Chad Nusbaum, Richard M Myers, Myles Brown, Wei Li, and X Shirley Liu. Model-based analysis of chip-seq (macs). *Genome Biol*, 9(9) :R137, 2008.

[194] Guangyong Zheng, Kang Tu, Qing Yang, Yun Xiong, Chaochun Wei, Lu Xie, Yangyong Zhu, and Yixue Li. ITFP : an integrated platform of mammalian transcription factors. *Bioinformatics (Oxford, England)*, 24(20) :2416–2417, October 2008.

[195] Yanni Zhu, Xiaotong Shen, and Wei Pan. Network-based support vector machine for classification of microarray samples. *BMC Bioinformatics*, 10 Suppl 1 :S21, 2009.

Résumé

Depuis le début des années 2000, la biologie moléculaire s'est engagée sur un nouveau paradigme et bascule dans le domaine des sciences expérimentales à grande échelle. Les technologies développées depuis le séquençage du génome humain ont considérablement contribué à faire évoluer l'ensemble des disciplines en biologie moléculaire qui s'orientent maintenant sur un passage à grande échelle des mesures faites sur le vivant.

Cette démocratisation de la mesure de l'information biologique a provoqué un besoin croissant en analyse de données d'un type nouveau. En effet, deux problèmes majeurs se posent : D'une part, bien que le nombre de données et variables mesurées soient très large (par exemple l'ensemble des éléments régulateurs d'un génome), le nombre d'échantillons reste limité (accès à relativement peu de patients), ce qui oblige à sortir du socle de la statistique classique et de réfléchir à de nouvelles méthodologies applicables à ce type de contexte. D'autre part, les processus biologiques résultent de l'enchaînement de la modification ou de l'interaction de cascades de molécules qui peuvent être détectées et mesurées par des technologies différentes (génomique, transcriptomique, protéomique), il est donc nécessaire d'intégrer des données hétérogènes générées par des technologies distinctes pour pouvoir les interpréter et reconstituer le contexte biologique complet de l'étude.

Dans ce mémoire, je présente des approches pour l'analyse et l'intégration de données hétérogènes, et l'amélioration des analyses existantes. Je présente d'abord un système d'analyse de données d'expression multiplateformes pour l'identification d'une signature de cellules souches par réseau neuronal et vocabulaire contrôlé. Pour aller plus loin, j'ai utilisé l'interactome humain pour améliorer les signatures de prédiction de la rechute en cancérologie. De cette manière, j'ai pu déterminer des signatures qui séparent mieux les patients de bon et mauvais pronostique dans le cancer du sein. Puis, j'ai développé plus avant cette technologie en y a joutant des mesures d'amplification de l'ADN. Plus récemment, je me suis intéressé aux réseaux de régulations. J'ai développé un système, HTS-Net qui permet l'analyse de réseaux de régulations identifiés dans des screening à haut début de type siRNA ou shRNA.

Ces approches vont plus loin que les analyses fonctionnelles classiques d'enrichissement car elles permettent d'identifier des réseaux d'interactions intéressants sans biais. Grâce à cette expertise, j'ai pu mettre en place des analyses réseaux pour la communauté, au travers de nombreuses collaborations. Ces approches sont la base de la biologie des systèmes, qui est un outil indispensable pour comprendre et intégrer la masse d'information générée par la communauté biomédicale.

Mots-clés: réseaux de gènes, réseaux de régulation, puces à ADN, interactome, bioinformatique, cancérologie

Abstract

Since the early 2000, molecular biology has engaged on a new paradigm and has shifted into the field of large-scale experimental sciences. The technologies that have been developed since the human genome sequencing have considerably contributed in the evolution of all molecular biology topics which are now moving on large scale analysis of biological phenotypes.

This advent of large-scale biology made critical the need for new data analysis strategies. Classical statistics cannot be directly applied in biological data because of the curse of dimensionality. On one hand, we measure too many variables (several thousands for a typical whole genome experiment, more for sequence analysis), for too little samples (several hundred for the best cases). On the other hand, biological phenomenon results from the cascade of several events measured by heterogeneous technologies (transcriptomics, sequencing, proteomics). Therefore, it became necessary to integrate heterogeneous data in order to interpret them in a global and causal context.

In this thesis, I present several approaches for the analysis and integration of heterogeneous data and the improvement of existing analysis. First, I introduce an approach based on neural network classification and controlled vocabulary to identify a stem cell differentiation signature. To go further, I established a network-based approach by analyzing the human interactome and integrate it with gene expression in order to improve genes signature that separate good and bad prognosis in breast cancer patients. Then, I talk about developing this technology further by adding DNA amplification data.

Lastly, I describe my latest interest in regulation networks, which let me to develop a new system, HTS-Net, for the analysis of siRNA/shRNA high throughout screenings.

These approaches are going further than classical functional enrichment-based approaches since they allow identification of functional networks without biases. Thanks to this expertise, I was able to put into production network analysis through several collaborations. These approaches are the basis for systems biology, an essential tool to analyze the body of information generated in biology.

Keywords: gene networks, regulation networks, microarrays, interactome, bioinformatics, oncology