Rapport Projet RESYS

Adam Boumessaoud, Océane Li, Thomas Louvet

November 17, 2024

1 Introduction

Le cancer du sein est l'une des principales causes de mortalité chez les femmes à travers le monde, avec des taux de survie qui varient considérablement en fonction des caractéristiques biologiques et génétiques des tumeurs [1]. La personnalisation des traitements pour le cancer du sein est devenue un objectif central, car les différentes sous-catégories de ce cancer présentent des pronostics et des réponses thérapeutiques variées [2]. Les avancées en génomique et en bio-informatique permettent désormais de traiter de grands ensembles de données génétiques et cliniques pour identifier des facteurs déterminants dans la progression de la maladie et la réponse aux traitements [3]. Le jeu de données METABRIC [4] représente une ressource majeure dans ce domaine, comprenant des informations cliniques, des mutations génétiques et des profils d'expression génique pour près de 2000 patients atteints de cancer du sein. Ce projet vise à explorer ce dataset pour identifier comment les mutations génétiques et les variations dans l'expression génique influencent les résultats cliniques, en particulier la survie et la réponse au traitement.

2 Matériels & Méthodes

2.1 Le dataset METABRIC

Le dataset METABRIC que nous avons utilisé est composé de données collectées auprès de 1904 patients atteints de cancer du sein. Il inclut principalement trois types de données : des caractéristiques cliniques, des données d'expression génique et des données sur les mutations des gènes. Les caractéristiques cliniques, représentées par une trentaine de variables, décrivent des éléments essentiels pour le suivi médical et le pronostic des patientes. On y trouve des informations comme le type de cancer, l'âge du patient au moment du diagnostic, le type de thérapie suivi ou encore le stade et la taille de la tumeur. Les données d'expression génique, quantifiées en z-scores, comprennent 489 variables indiquant le niveau d'expression de divers gènes. En parallèle, 173 variables de mutation indiquent la présence ou l'absence de mutations précises, ainsi que leur nature. Par exemple, la mutation S122I désigne une substitution de la Sérine par l'Isoleucine à la position 122 d'un gène.

Toutefois, certaines valeurs de ces variables sont manquantes : pour les variables continues, les valeurs manquantes ont été imputées par la médiane, tandis que pour les variables catégorielles, les lignes concernées ont été retirées. Notre analyse s'est concentrée sur deux axes : d'une part, l'étude des mutations en relation avec des caractéristiques cliniques comme la survie et le type de traitement ; d'autre part, l'analyse des données d'expression génique, en cherchant à identifier des groupes de gènes associés aux résultats cliniques. Ces deux approches visent à explorer l'influence des mutations et de l'expression génique sur la progression et le pronostic du cancer du sein. Dans les deux cas, on a décidé de mettre avec ces données plusieurs features selon le découpage suivant :

Survie du patient : `overall_survival` (vivant ou décédé), `overall_survival_months` (durée de survie après traitement), `nottingham_prognostic_index` (score après chirurgie du pronostic) et `death_from_cancer` (décès lié au cancer ou non). Les indicateurs de survie sont indispensables pour prédire l'issue de la maladie. Ces paramètres aident à identifier les variables influençant le risque de récidive ou de décès liés au cancer, qui sont largement étudiés dans des recherches sur le pronostic du cancer du sein [5][6].

Type de traitement administré : `type_of_breast_surgery` (mastectomie ou conservation de la poitrine), `chemotherapy`, `hormone_therapy` et `radio_therapy`. Les traitements influencent la réponse clinique et la survie, chaque traitement ciblant des mécanismes biologiques spécifiques selon le sous-type de can-

cer. Par exemple, les patients ayant des cancers ER-positifs répondent souvent mieux à l'hormonothérapie (Perou et al., 2000 [7]), tandis que les cas HER2-positifs peuvent bénéficier des thérapies ciblées. L'effet de ces traitements est bien documenté pour améliorer la prise en charge et la survie des patientes [8].

Etat des récepteurs chez les cellules cancéreuses : les variables `er_status` (récepteur à oestrogène), `her2_status` (récepteur du facteur de croissance épidermique) et `pr_status` (récepteur à progestérone) indiquent la probabilité de réponse aux thérapies hormonales. Les cancers ER-positifs ou PR-positifs sont souvent sensibles aux thérapies hormonales (comme le tamoxifène) [9]. Les tumeurs HER2-positives tendent à être plus agressives mais peuvent répondre aux thérapies ciblant HER2 (comme le trastuzumab) [10]. Analyser ces statuts permet de mieux anticiper la réponse aux traitements et d'identifier des sous-groupes de patients avec des pronostics différents.

Facteurs cliniques et pathologiques : `age_at_diagnosis`, `neoplasm_histologic_grade` (grade histologique de la tumeur, de 1 à 3), `lymph_nodes_examined_positive` (nombre de ganglions lymphatiques atteints par le cancer), `tumor_stage` et `tumor_size`. Ces caractéristiques sont des éléments bien établis pour évaluer la progression du cancer et son agressivité. Des études ont montré que les tumeurs de stade et de grade élevés sont souvent associées à un risque accru de métastase et de rechute (Elston & Ellis, 1991 [11]). Ces données fournissent donc un cadre clinique pour les prises de décisions thérapeutiques.

Classification tumorale : `3-gene_classifier_subtype` (sous-type de la tumeur basé sur les gènes ER et HER2), `tumor_other_histologic_subtype` (type de cancer basé sur l'analyse microscopique du tissu cancéreux) et `cancer_type_detailed` (sous-type du cancer). Le sous-type du cancer, défini par l'analyse histologique et des marqueurs biologiques, est utile pour comprendre la biologie de la tumeur et orienter les traitements personnalisés (Schnitt, 2010 [12]). Par exemple, les cancers triple-négatifs, qui manquent d'expression de ER, PR et HER2, sont associés à un pronostic plus défavorable et nécessitent des approches thérapeutiques distinctes.

Ce découpage des variables cliniques repose sur leur rôle démontré dans l'évaluation du pronostic et de la réponse thérapeutique dans le cancer du sein. En structurant notre analyse autour de ces catégories, nous nous appuyons sur des critères qui ont été identifiés dans la littérature comme étant des facteurs déterminants de la survie et de la réponse aux traitements (Curtis et al., 2012 [2]). L'association de chaque catégorie clinique avec les données d'expression génique et de mutations vise à explorer les relations causales potentielles entre ces facteurs cliniques et les profils génétiques, afin de mieux comprendre comment ces caractéristiques influencent l'évolution de la maladie et la réponse aux thérapies.

Pour l'analyse des graphes comportant l'expression des gènes, nous avons fait des groupes de 50 et avons regardé chaque graphe avec les 5 découpages. Pour chacun, nous avons seulement gardé les gènes qui sont directement et fortement connectés avec des features cliniques et fait un graphe final contenant ces gènes sélectionnés. Pour simplifier l'analyse des données de mutations, nous avons converti toutes les mutations, indépendamment de leur type, en une notation binaire : chaque gène est codé par "1" s'il présente une mutation et "0" en l'absence de mutation. Cette transformation rend les données plus interprétables pour l'outil MIIC pour générer des réseaux.

2.2 Construction des réseaux (MIIC)

Nous avons utilisé l'outil MIIC (Multivariate Information-based Inductive Causation) pour construire les réseaux d'interactions. MIIC est une méthode d'inférence causale qui permet de construire des graphes d'interactions entre variables sans faire d'hypothèses a priori. Cette approche est particulièrement utile dans les domaines biologiques où les relations entre les variables sont souvent inconnues ou non linéaires. L'entrée de l'outil est un tableau CSV où chaque colonne représente une feature (variable) et chaque ligne correspond à un individu (patient). MIIC utilise une approche basée sur l'information mutuelle. L'algorithme de MIIC commence par créer un graphe entièrement connecté, où toutes les variables sont liées entre elles. Ensuite, il élimine progressivement les liens considérés comme indirects ou non significatifs, en se basant sur l'information mutuelle, une mesure statistique qui quantifie la dépendance entre deux variables [13]. En outre, MIIC oriente les liens causaux en identifiant des structures en 'V', ce qui permet de déterminer la direction des relations causales entre les variables. Pour chaque lien, MIIC évalue la confiance que l'on peut avoir sur ces derniers, on s'assure alors que les liens conservés dans le graphe sont statistiquement significatifs et robustes [14]. MIIC est un outil simple et permet de construire rapidement des graphes robustes à partir de nos données. Il est particulièrement adapté à l'analyse de grands ensembles de données biologiques, comme les profils d'expression génique, car il permet d'explorer des relations complexes entre gènes sans nécessiter d'hypothèses préalables.

Le choix de MIIC pour la construction de nos réseaux d'interaction se justifie par sa capacité à identifier des relations causales dans un contexte complexe et multidimensionnel. Contrairement à des algorithmes comme igraph, qui se contentent généralement de détecter des relations structurelles (par exemple, en utilisant des critères de corrélation ou des mesures de similarité [15]), MIIC va au-delà en inférant des liens causaux, ce qui peut donner des résultats plus précis en matière de compréhension des mécanismes biologiques sous-jacents. Igraph permet de visualiser et d'analyser des graphes à partir de données structurées, mais il n'infère pas directement des relations causales. Ainsi, bien que les deux outils puissent être utilisés pour la construction de réseaux d'interactions, MIIC offre l'avantage d'une approche plus dynamique et causale.

MIIC donne un score log10 confidence pour chaque arête du réseau. Ce score mesure la robustesse et la fiabilité des liens causaux identifiés entre les variables. Un score élevé indique qu'il existe une forte confiance dans l'existence de l'interaction, car ce score est basé sur l'information mutuelle et la probabilité que cette relation soit observée par hasard. Un score élevé représente donc une forte preuve statistique que la relation entre les variables n'est pas due à un bruit ou à une coïncidence. Pour qu'une interaction soit considérée comme fiable dans les réseaux générés par MIIC, il faut au minimum un score log10 confidence de 10 (soit une p-value $< 10^{10}$, ce qui indique une très forte significativité statistique) [14].

2.3 Clusters de gènes & Gene Ontology

Dans le but de vérifier les résultats obtenus avec MIIC, nous avons utilisé l'outil STRING afin de reconstruire les réseaux en se basant cette fois-ci sur les interactions connues et prédites des protéines des gènes étudiés. Les réseaux obtenus sont non dirigés mais nous permettent l'application aisée d'algorithmes de clustering comme la MCL ainsi qu'une analyse fonctionnelle protéines et donc aux clusters obtenus.

En combinaison avec STRING, nous avons utilisé l'outil gProfiler, un autre outil de Gene Ontology, pour comparer les résultats d'analyse que nous avons obtenu avec STRING. Utiliser plusieurs outils nous permet de vérifier la robustesse des résultats. Grâce à gProfiler, nous pouvons obtenir des informations sur la fonction moléculaire des protéines codées par les gènes étudiés, ainsi que sur les processus biologiques dans lesquels elles interviennent.

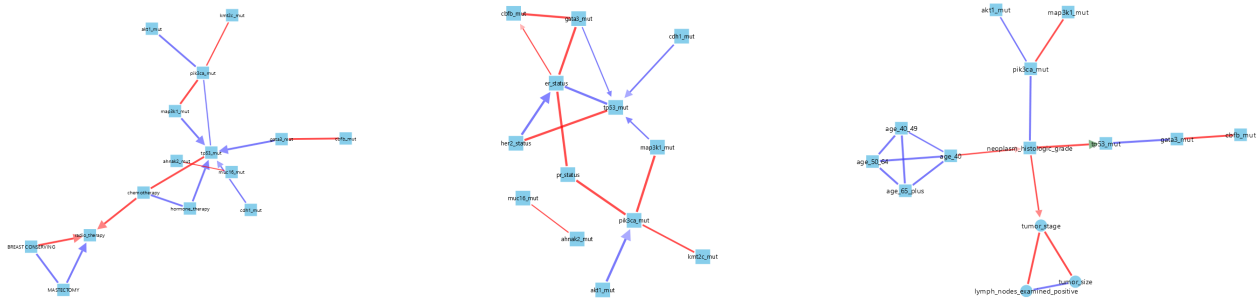
3 Résultats

Dans les graphes générés par MIIC, chaque noeud représente une variable, et chaque arête indique un lien de dépendance (association ou causalité). Les arêtes sans flèches - comme entre un bon pronostic Nottingham et la mutation du gène `tp53` - indiquent une association (dépendance statistique), sans présomption de causalité directe. Les arêtes avec flèches indiquent des relations causales. On distingue :

- Les flèches vertes - comme entre le décès dû au cancer et un mauvais pronostic Nottingham - désignent des relations causales "genuine", dont on est sûr de la causalité.
- Les flèches colorées (rouge, bleu, gris) - comme entre les mutations des gènes `tp53` et `gata3` - représentent des causalités "putatives", c'est-à-dire incertaine quant à l'origine du lien.

Les arêtes pointillées bidirectionnelles - comme entre les mutations des gènes `tp53` et `map3k1` - signalent la présence d'une variable latente non observée qui cause les deux noeuds liés. La couleur des arêtes a une signification particulière également : la couleur rouge indique une association positive, la bleue désigne une association négative, et la couleur gris signifie qu'il s'agit d'une variable catégorielle (ex : type de cancer).

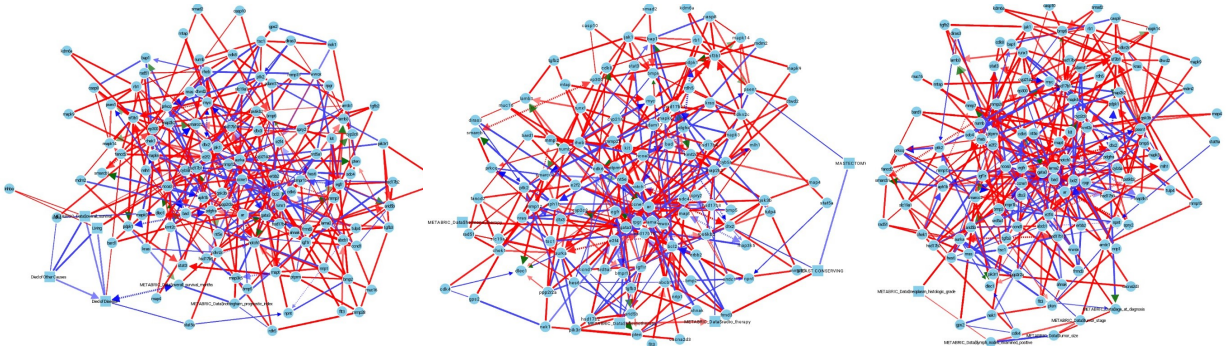
Par rapport aux graphes d'expression génique, les graphes avec les mutations (Figure.1) sont moins fournis car beaucoup de liens entre les features disparaissent avec l'augmentation du seuil de confiance. Les features cliniques tendent à se rejoindre entre elles : ce qui est attendu puisque certaines d'entre elles représentent la même idée, par exemple `overall_survival` et `Living`, et d'autres sont mutuellement exclusives, comme `breast_conserving` et `mastectomy`. D'autres liens sont évidents si on connaît les traitements du cancer du sein : par exemple la chirurgie conservatrice qui agit positivement sur la radiothérapie. Cette dernière est systématiquement utilisée pour éviter le risque de récidive dans le sein restant. Les liens entre les gènes et les features sont intéressants à étudier puisqu'ils sont susceptibles d'avoir un grand rôle et une importance cruciale dans la maladie. C'est sur ces liens que nous allons nous pencher.



(a) **(b)** **(c)**

Figure 1: Visualisation des graphes MIIC pour les relations entre (a) les traitements suivis et mutations, (b) les états des récepteurs (hormonaux et HER2) et les mutations, et (c) les variables cliniques, en lien avec les mutations.

Un gène que l'on trouve approximativement au centre des trois graphes ci-dessus est `tp53` : c'est donc un bon candidat pour trouver les gènes les plus probables d'avoir un rôle dans le mécanisme du cancer du sein. Les mutations de tp53 sont liées, dans les graphes, par une association négative avec un bon pronostic (par le score de Nottingham), par une association positive avec la chimiothérapie, par un lien causal négatif (présumé) partant de l'hormonothérapie et par un lien causal positif partant du grade histologique.



(a) **(b)** **(c)**

Figure 2: Visualisation des graphes MIIC représentant les relations clés entre les gènes exprimés et (a) la survie des patientes, (b) les traitements administrés, et (c) les variables cliniques.

De la même manière que dans les graphes avec les données de mutations, chaque nœud est une variable, représentant soit une caractéristique clinique mesurée, soit un gène dont on a mesuré l'expression sur les patients. On peut déjà remarquer que les features cliniques se regroupent souvent entre elles. La plupart des liens ne sont pas fléchés (donc représentant des associations entre deux variables).

Contrairement aux graphes avec les mutations où les liens disparaissent rapidement avec l'augmentation du seuil de la confiance, les liens dans les graphes d'expression génique sont très robustes (la plupart d'entre eux ont une log10 de confiance supérieure à 10). On peut se pencher sur certains liens entre des gènes et les caractéristiques cliniques. Dans le premier graphe [fig 2a], le gène `hsd17b11` est associé positivement aux caractéristiques `overall_survival` et `Living`. On peut donc supposer que ce gène, si fortement exprimé, agit dans l'intérêt de la survie du patient. Une analyse de Gene Ontology de ce gène peut nous aider à étayer ou non cette hypothèse. De plus, ce gène `hsd17b11` est au centre de ce réseau (il est lié à 12 autres gènes). Il serait intéressant de savoir quelle est la fonction de ce dernier.

Un gène intéressant dans la figure 2b est le gène `gata3`. Tout comme `hsd17b11`, il est central au réseau (il est connecté à 23 gènes). De plus, il est associé négativement à la chimiothérapie et positivement à l'hormonothérapie. A nouveau, une analyse de Gene Ontology peut nous permettre d'identifier pourquoi on a ces liaisons et de potentiellement identifier une cible thérapeutique.

Enfin dans la figure 2c, `pdgfra` cause négativement sur l'âge au diagnostic. On peut penser que ce gène, si exprimé, avance l'apparition de la maladie. Nous avons construit d'autre graphe qui ne sont pas montrés ici. Dans ces derniers, les gènes mentionnés plus haut ont les mêmes caractéristiques. Par exemple, `hsd17b11` est au centre des réseaux.

Une affichage plus claire et nette des réseaux (Figure 1 et 2) sont disponibles dans les pages Annexes.

4 Discussion

Les graphes d'interaction générés par MIIC, basés sur les données d'expression génique et de mutations dans le cancer du sein, révèlent des liens cruciaux entre certaines caractéristiques cliniques et génétiques. Ils mettent en lumière des associations fortes et spécifiques qui peuvent aider à comprendre le mécanisme moléculaire de la maladie et identifier des cibles thérapeutiques potentielles. Une analyse approfondie de ces liens suggère plusieurs hypothèses intéressantes.

Dans les graphes basés sur les mutations, le gène `tp53` se distingue par sa position centrale dans le réseau, soulignant son rôle potentiel dans la progression du cancer du sein. En effet, `tp53` est bien connu pour son implication dans les mécanismes de suppression des tumeurs et est régulièrement altéré dans de nombreux types de cancers, y compris le cancer du sein [16]. Les mutations de ce gène sont associées à des pronostics plus sévères et à une réponse accrue à la chimiothérapie, soulignant son influence sur la progression tumorale et les options thérapeutiques [17]. En lien négatif avec le statut "ER-positif" (récepteurs aux oestrogènes fonctionnels), `tp53` pourrait donc être lié à des sous-types de cancers plus agressifs et moins sensibles à l'hormonothérapie, ce qui en fait une cible pertinente pour de nouvelles recherches et stratégies thérapeutiques dans les cancers ER-négatifs [18].

Contrairement aux graphes de mutations, les graphes d'expression génique restent robustes même avec des seuils de confiance élevés, suggérant que l'expression génique est moins variable et souvent plus liée aux processus cellulaires fondamentaux. Le gène `hsd17b11` est particulièrement intéressant, car il a déjà été repéré pour avoir eu un lien avec une survie sans rechute du cancer [19]. Cela suggère que les gènes connectés à `hsd17b11` dans les réseaux pourraient être explorés comme cibles thérapeutiques ou indicateurs des mécanismes de survie. Le gène `hsd17b11` joue un rôle dans le métabolisme des androgènes en participant à la stéroïdogenèse. Il peut agir en métabolisant des composés qui stimulent la synthèse de stéroïdes et/ou en générant des métabolites qui l'inhibent. Il code pour un antigène associé aux tumeurs dans le lymphome cutané à cellules T. Sa fonction enzymatique peut influencer la balance hormonale, ce qui est pertinent pour les cancers du sein où la prolifération tumorale est souvent hormonodépendante. Des études montrent que les enzymes de la famille HSD, dont `hsd17b11`, impactent potentiellement la croissance cellulaire en modifiant la disponibilité des stéroïdes, et pourraient ainsi influencer la réponse aux thérapies hormonales ciblées pour les cancers hormonodépendants [20].

Le gène `gata3`, essentiel dans le développement des lymphocytes T, est corrélé positivement à un statut ER positif [21]. Ce gène central dans les réseaux d'expression pourrait renforcer la sensibilité des cellules cancéreuses à l'hormonothérapie, ouvrant la possibilité d'une meilleure réponse dans les cancers ER positifs. Son rôle pourrait être exploré pour ajuster les stratégies d'hormonothérapie.

Enfin, `pdgfra`, une protéine de la famille des récepteurs tyrosine kinase, joue un rôle majeur dans les voies de signalisation cellulaire. La fixation de facteurs de croissance à ces récepteurs provoque l'activation d'une cascade de signalisation qui joue un rôle dans la gastrulation et le développement de nombreux systèmes d'organes. Des recherches montrent que ce gène `pdgfra` (Platelet-derived Growth Factor Receptor Alpha) a un rôle significatif dans la progression des cancers du sein, notamment en étant associé à une prolifération cellulaire accrue et au statut HER2-positif. L'expression de `pdgfra` a été observée dans des sous-types agressifs de cancers, et sa présence est liée à une évolution tumorale défavorable [22]. Ces liens suggèrent que `pdgfra` pourrait être une cible prometteuse pour des thérapies visant des mécanismes de prolifération tumorale dans le cancer du sein.

L'analyse avec l'outil gProfiler nous confirme que les 3 gènes (`hsd17b11`, `gata3` et `pdgfra`) jouent un rôle dans le métabolisme hormonal. Cela établit un lien entre les mutations de ces gènes et le statut ER de certains cancers du sein. Le gène `tp53` semble avoir l'effet inverse : on observe dans la figure 1a une corrélation négative entre le statut ER positif et `tp53_mut`. Il en est de même entre `tp53_mut` et `gata3_mut`. Ce contraste observé pourrait indiquer des voies de développement distinctes dans les sous-types de cancer du sein, associant `tp53` à des cancers plus résistants à l'hormonothérapie.

En conclusion, ces résultats illustrent la complexité des interactions entre les mutations, l'expression génique et les caractéristiques cliniques du cancer du sein. L'analyse des gènes `tp53`, `hsd17b11`, `gata3` et `pdgfra` révèle des pistes prometteuses pour la compréhension des mécanismes tumoraux, et met en évidence leur potentiel en tant que cibles thérapeutiques. Cette approche pourrait ainsi orienter le développement de traitements personnalisés - adaptés au profil génétique et aux caractéristiques d'expression de chaque patient - ce qui améliore la qualité de prise en charge clinique.

5 Future work

Afin de mieux interpréter les résultats réseaux complexes obtenus à partir des données transcriptomiques, nous pourrions appliquer divers algorithmes de clustering pour peut-être déceler des liens de causalités plus précis entre les gènes. En outre, pour étudier de manière plus précise la place des features dans le réseau de cooccurrences des mutations, nous pourrions nous créer des réseaux contenant uniquement une feature et toutes les mutations. Cela nous permettrait d'éviter les liens forts entre features les poussant à former des clusters entre eux. A la place, nous aurions potentiellement des liens directs entre mutations et features que nous n'aurions pas pu voir avant. Nous pourrions par exemple appliquer ce principe à l'analyse du lien entre les mutations `tp53_mut`, `gata3_mut` et la présence de récepteurs hormonaux sur les cellules tumorales.

References

[1] Rebecca L. Siegel, Kimberly D. Miller, Heather E. Fuchs, and Ahmedin Jemal. Cancer statistics, 2021. CA: A Cancer Journal for Clinicians, 71(1):7–33, 2021. PMID: 33433946, https://doi.org/10.3322/caac.21654.

[2] 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, Rachael Russell, Shannon McKinney, the METABRIC Group, Anita Langerød, Andrew Green, Elena Provenzano, Gordon Wishart, Sarah Pinder, Peter Watson, Florian Markowetz, Lorraine 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):346–352, 2012. PMID: 22522925, PMCID: PMC3440846, https://doi.org/10.1038/nature10983.

[3] Joel S. Parker, Michael Mullins, Maggie C. U. Cheang, Samuel Leung, David Voduc, Tracey Vickery, Sian Davies, Christophe Fauron, Xiaping He, Zhiyuan Hu, John F. Quackenbush, Isaac J. Stijleman, Joe Palazzo, J. S. Marron, Andrew B. Nobel, Elaine Mardis, Torsten O. Nielsen, Matthew J. Ellis, Charles M. Perou, and Philip S. Bernard. Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal of Clinical Oncology, 27(8):1160–1167, 2009. PMID: 19204204, PMCID: PMC2667820, https://doi.org/10.1200/JCO.2008.18.1370.

[4] Metabric dataset from cbioportal. https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric/data.

[5] Y. Mao, Q. Qu, X. Chen, O. Huang, J. Wu, and K. Shen. The prognostic value of tumor-infiltrating lymphocytes in breast cancer: A systematic review and meta-analysis. PLoS One, 11(4):e0152500, 2016. PMID: 27073890, PMCID: PMC4830515, https://doi.org/10.1371/journal.pone.0152500.

[6] I. Balslev, C.K. Axelsson, K. Zedeler, B.B. Rasmussen, B. Carstensen, and H.T. Mouridsen. The nottingham prognostic index applied to 9,149 patients from the studies of the danish breast cancer cooperative group (dbcg). Breast Cancer Res Treat, 32(3):281–290, 1994. PMID: 7865856, https://doi.org/10.1007/BF00666005.

[7] 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, 2000. PMID: 10963602, https://doi.org/10.1038/35021093.

[8] P. Advani, L. Cornell, S. Chumsri, and A. Moreno-Aspitia. Dual her2 blockade in the neoadjuvant and adjuvant treatment of her2-positive breast cancer. Breast Cancer (Dove Med Press), 7:321–335, 2015. PMID: 26451122, PMCID: PMC4590321, https://doi.org/10.2147/BCTT.S90627.

[9] Hormone therapy for breast cancer, 2023. https://www.cancer.gov/types/breast/breast-hormone-therapy-fact-sheet.

[10] M. Harries and I. Smith. The development and clinical use of trastuzumab (herceptin). Endocr Relat Cancer, 9(2):75–85, 2002. PMID: 12121832, https://doi.org/10.1677/erc.0.0090075.

[11] C.W. Elston and I.O. Ellis. Pathological prognostic factors in breast cancer. i. the value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology, 19(5):403–410, 1991. PMID: 1757079, https://doi.org/10.1111/j.1365-2559.1991.tb00229.x.

[12] S.J. Schnitt. Classification and prognosis of invasive breast cancer: from morphology to molecular taxonomy. Mod Pathol, 23 Suppl 2:S60–S64, 2010. PMID: 20436504, https://doi.org/10.1038/modpathol.2010.33.

[13] Thomas M. Cover and Joy A. Thomas. Elements of Information Theory. Wiley-Interscience, 2006.

[14] Nadir Sella, Louis Verny, Guido Uguzzoni, Séverine Affeldt, and Hervé Isambert. Miic online: a web server to reconstruct causal or non-causal networks from non-perturbative data. Bioinformatics, 34(13):2311–2313, 2018. https://doi.org/10.1093/bioinformatics/btx844.

[15] Gabor Csardi and Tamas Nepusz. The igraph software package for complex network research. InterJournal of Complex Systems, 1695(5):1–9, 2006.

[16] A. Shahbandi, H. David Nguyen, and J. G. Jackson. Tp53 mutations and outcomes in breast cancer: Reading beyond the headlines. Trends Cancer, 6(2):98–110, Feb 2020. PMID: 32061310, PMCID: PMC7931175, https://doi.org/10.1016/j.trecan.2020.01.007.

[17] E. Panatta, C. Zampieri, G. Melino, and I. Amelio. Understanding p53 tumour suppressor network. Biology Direct, 16:14, 2021. https://doi.org/10.1186/s13062-021-00298-3.

[18] Amber N. Hurson, Mustapha Abubakar, Alina M. Hamilton, Kathleen Conway, Katherine A. Hoadley, Michael I. Love, Andrew F. Olshan, Charles M. Perou, Montserrat Garcia-Closas, and Melissa A. Troester. Tp53 pathway function, estrogen receptor status, and breast cancer risk factors in the carolina breast cancer study. Cancer Epidemiol Biomarkers Prev, 31(1):124–131, Jan 2022. https://doi.org/10.1158/1055-9965.EPI-21-0661.

[19] A. K. Corbet, E. Bikorimana, R. I. Boyd, D. Shokry, K. Kries, A. Gupta, A. Paton, Z. Sun, Z. Fazal, S. J. Freemantle, E. R. Nelson, M. J. Spinella, and R. Singh. G0s2 promotes antiestrogenic and pro-migratory responses in er+ and er- breast cancer cells. Transl Oncol, 33:101676, Jul 2023. PMID: 37086619, PMCID: PMC10214302, https://doi.org/10.1016/j.tranon.2023.101676.

[20] Mirja Rotinen, Joaquín Villar, Jon Celay, Irantzu Serrano, Vicente Notario, and et al. Transcriptional regulation of type 11 17-hydroxysteroid dehydrogenase expression in prostate cancer cells. Molecular and Cellular Endocrinology, 2011. hal-00708529.

[21] David Voduc, Maggie Cheang, and Torsten Nielsen. Gata-3 expression in breast cancer has a strong association with estrogen receptor but lacks independent prognostic value. Cancer Epidemiol Biomarkers Prev, 17(2):365–373, Feb 2008. PMID: 18268121, https://doi.org/10.1158/1055-9965.EPI-06-1090.

[22] I. Carvalho, F. Milanezi, A. Martins, and et al. Overexpression of platelet-derived growth factor receptor in breast cancer is associated with tumour progression. Breast Cancer Res, 7:R788, 2005. https://doi.org/10.1186/bcr1304.

6 Annexes

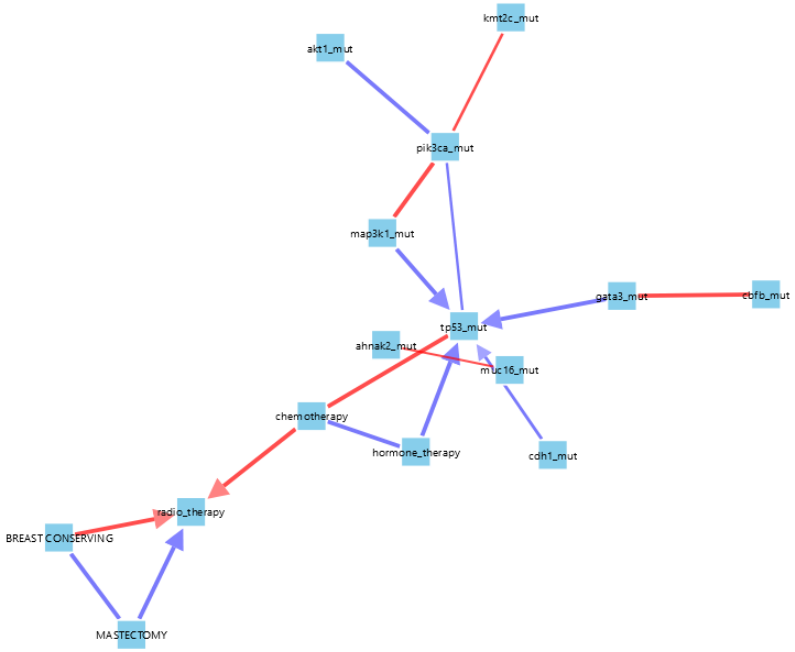


Figure 3: Relations entre les types de traitement et les mutations

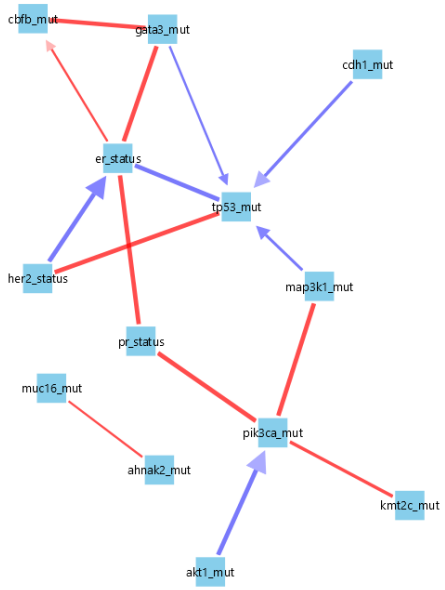


Figure 4: Relations entre les états des récepteurs et les mutations

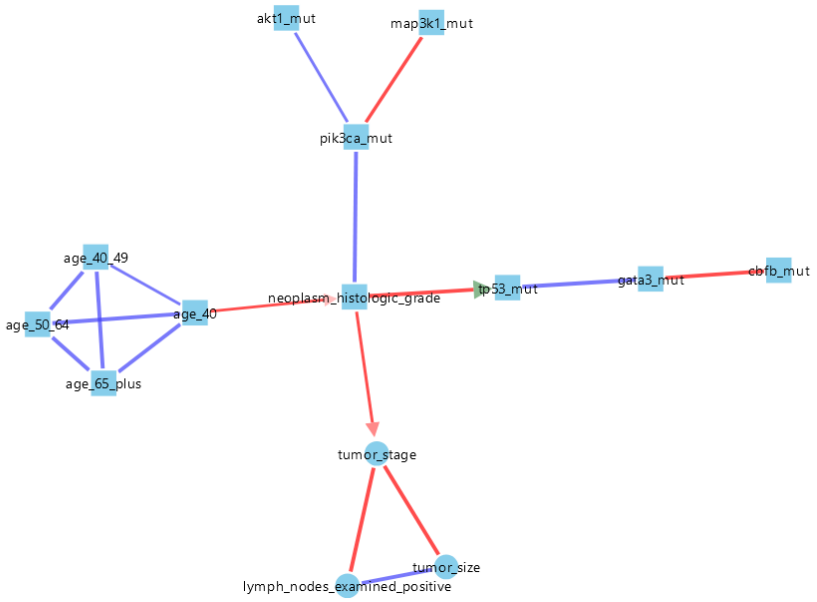


Figure 5: Relations entre les variables cliniques et les mutations

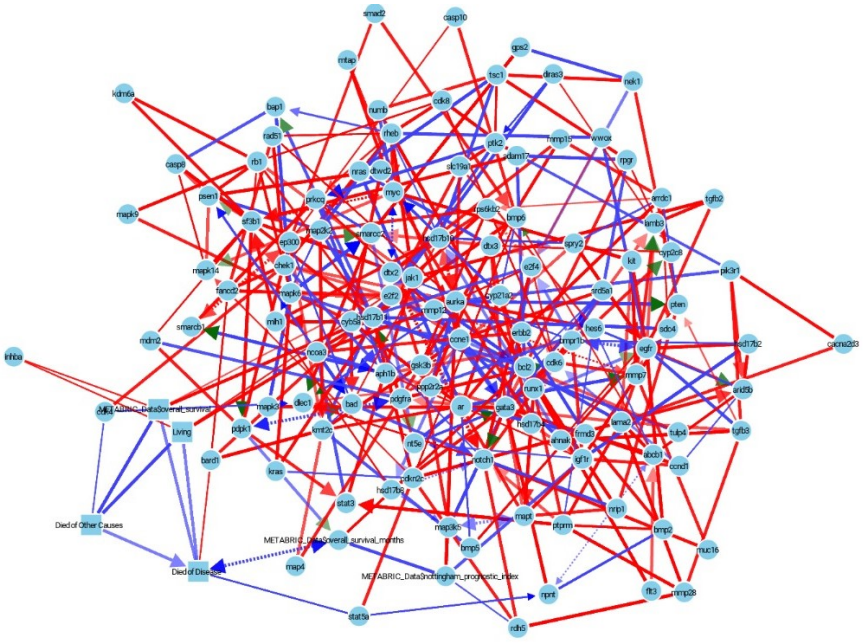


Figure 6: Relations entre la survie des patients et les expressions géniques

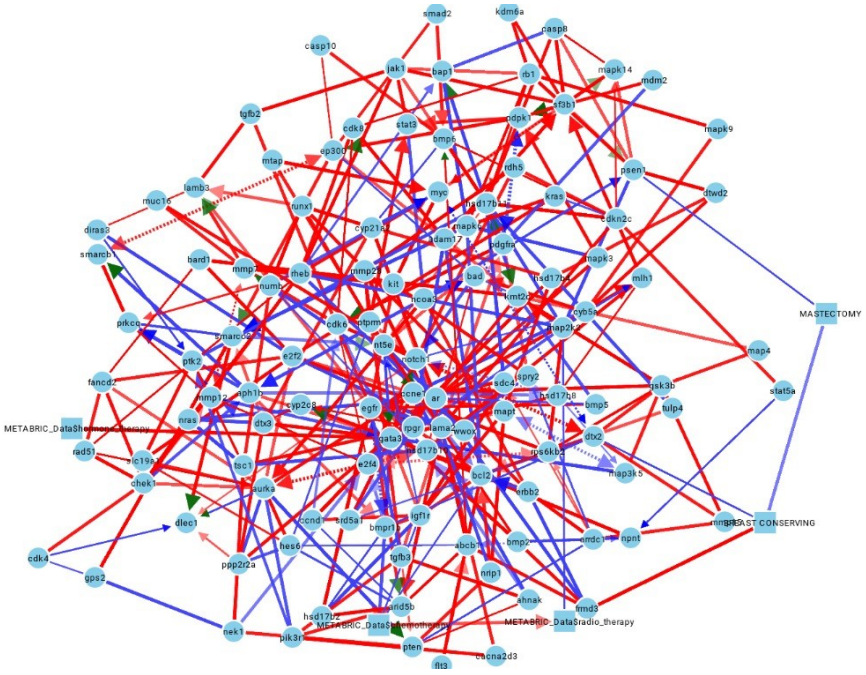


Figure 7: Relations entre les types de traitement et les expressions géniques

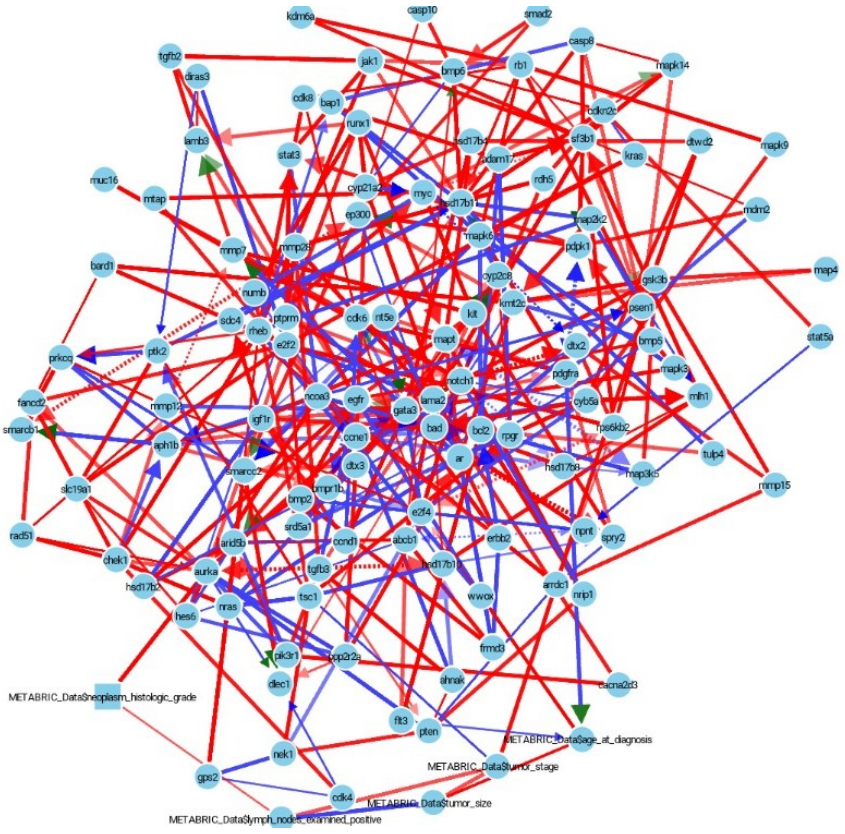


Figure 8: Relations entre les variables cliniques et les expressions géniques