Développement d'une approche hybride ensembliste de sélection de variables pour l'élaboration d'un atlas de biomarqueurs de cancers

Elsa Claude

THÈSE EN COTUTELLE PRÉSENTÉE POUR OBTENIR LE GRADE DE

DOCTEURE DE

L'UNIVERSITÉ DE BORDEAUX (UB) et de L'UNIVERSITÉ LAVAL (UL)

ÉCOLE DOCTORALE DE MATHÉMATIQUES ET D'INFORMATIQUE (UB) et FACULTÉ DE MÉDECINE (UL)

Par **Elsa Claude**

SPÉCIALITÉS : INFORMATIQUE (UB) et MÉDECINE MOLÉCULAIRE (UL)

Développement d'une approche hybride ensembliste de sélection de variables pour l'élaboration d'un atlas de biomarqueurs de cancers

---

Soutenue le 18 Décembre 2024

Membres du jury :

Mme. Emmanuelle Becker, Pr.	Univ. Rennes	Présidente, Rapporteuse
Mme. Farida Zehraoui, MCF .	Univ. Evry, Paris-Saclay	Rapporteuse
M. François Coste, CR	Univ. Rennes	Examinateur
M. Raoul Santiago, Pr.	Univ. Laval	Examinateur
Mme. Patricia Thébault, Pr. .	Univ. Bordeaux	Co-directrice
M. Arnaud Droit, Pr.	Univ. Laval	Co-directeur

Membres invités :

M. Mickaël Leclercq, PPR ...	Univ. Laval	Co-encadrant
Mme. Raluca Uricaru, MCF ..	Univ. de Bordeaux	Co-encadrante

Résumé

La recherche biomédicale s'appuie sur les données omiques, telles que la transcriptomique, pour déchiffrer la complexité de maladies comme le cancer. Les méthodes traditionnelles d'identification des biomarqueurs transcriptomiques associés à un phénotype font appel à des techniques de sélection de variables. Les méthodes hybrides ensemblistes de sélection de variable (Hybrid Ensemble Feature Selection, HEFS) sont de plus en plus employées dans la dernière décennie pour leur capacité à garantir la robustesse des variables sélectionnées à travers des perturbations fonctionnelles et des données. Toutefois, leur conception reste un défi. Notre première contribution vise à évaluer la stratégie de type HEFS en analysant en détail 4 scénarios appliqués à l'identification de biomarqueurs dans divers cancers à partir de données de transcriptomique RNA-Seq. Ces scénarios explorent les combinaisons de deux méthodes de réduction des variables (gènes différentiellement exprimés et variance) avec deux stratégies de rééchantillonnage (repeated holdout par distribution équilibrée stratifiée et stratifiée aléatoire) pour la sélection de variable en aval de l'agrégation de signatures de milliers de modèles d'apprentissage machine enveloppés. Nos résultats soulignent les avantages de l'utilisation des approches HEFS pour identifier des biomarqueurs de maladies complexes, étant donné leur capacité à produire des résultats généralisables et stables. En outre, nous soulignons les considérations critiques essentielles pour la conception de telles stratégies. Notre deuxième contribution est le développement d'un atlas de signatures de biomarqueurs transcriptomiques de multiples cancers à des stades d'avancement variés. L'étude intégrée de cet atlas avec des données d'annotation biologique éprouvées (ex : REACTOME) doit permettre l'identification de gènes d'intérêt pour divers phénotypes oncologiques.

Abstract

Biomedical research relies on omics data, such as transcriptomics, to decipher the complexity of diseases such as cancer. Traditional methods for identifying transcriptomic biomarkers associated with a phenotype make use of feature selection approaches. Hybrid Ensemble Feature Selection (HEFS) methods have become increasingly popular in the last decade for their ability to guarantee the robustness of selected variables across functional and data perturbations. However, their design remains a challenge. Our first contribution aims to evaluate the HEFS strategy by analyzing in detail 4 scenarios applied to the identification of biomarkers in various cancers from RNA-Seq transcriptomics data. These scenarios explore combinations of two variable reduction methods (differentially expressed genes and variance) with two resampling strategies (repeated holdout by stratified distribution-balanced and random stratified) for variable selection downstream of the aggregation of thousands of wrapped machine learning models signatures. Our results highlight the advantages of using HEFS approaches to identify biomarkers of complex diseases, given their ability to produce generalizable and stable results. In addition, we highlight critical considerations essential for the design of such strategies. Our second contribution is the development of an atlas of transcriptomic biomarker signatures from multiple cancers at various stages of progression. The integrated study of this atlas with proven biological annotation data (e.g. REACTOME) should enable the identification of genes of interest for various oncological phenotypes.

Remerciements

Depuis petite, je me suis toujours laissée porter par ma curiosité. Et comme celle-ci est insatiable, elle est devenue une composante majeure de ma voie éducative et professionnelle. Me voici ainsi, non sans surprise, à la fin d'un doctorat. Cette aventure à la recherche de réponses, a été la plus éprouvante et gratifiante qu'il m'a été de vivre à ce jour. Un grand huit sur trois ans (et quelques mois, mais qui compte ?) où l'on sait qu'on va perdre nos repères, mais on ne sait juste pas quand ni comment. J'en ressors changée, pleines de réponses et... de doutes. Toujours est-il que ce doctorat n'aurait eu ni début, ni fin sans la participation de nombreuses personnes. Merci à elles et eux.

Je remercie les membres du jury, Farida Zehraoui, Emmanuelle Becker, Raoul Santiago et François Coste d'avoir accepté d'évaluer mes travaux et pour leurs retours très enrichissants.

J'adresse aussi mes remerciements à Bordeaux INP, au Labri, au CRCHUL, à l'Université de Bordeaux et l'Université Laval qui ont rendu cette thèse possible par leurs contributions financières, matérielles et humaines.

Je souhaite remercier mes encadrantes et encadrants, Patricia Thébault, Raluca Uricaru, Arnaud Droit et Mickaël Leclercq pour m'avoir donné l'opportunité de réaliser ce doctorat. Votre écoute et votre investissement ont été des piliers fondamentaux de ma réussite dans de bonnes conditions. Par votre expérience, j'ai appris à voir le positif dans les périodes les plus dures de cette grande aventure. Patricia, Raluca, j'aspire à adopter votre positivisme, merci pour tout.

Merci à Élodie et Frédéric pour leur expertise et leurs conseils et sans qui mon projet de recherche n'aurait pas connu une fin aussi appréciable.

Cette thèse en co-tutelle internationale m'a amenée sur les deux côtes de l'océan Atlantique, en France et au Québec. Un pied dans chaque pays, j'y ai tissé de solides liens amicaux et professionnels.

Au LaBRI, le bureau 328 a vu passer moult doctorantes et doctorants en trois ans. Merci à Myriam dit Mimi (la pro de la brioche sèche) pour son soutien sans failles et nos discussions sans fin, pour son amitié entière et rassurante. Merci à Claire et Gala (qui ne partagent pas leur gâteau au chocolat) pour leur écoute et les bonnes blagues. Merci à Yanis, que j'ai rencontré peu avant ma mobilité outre-atlantique et qui avec le temps est devenu un point d'ancrage au Labri par son écoute

attentive et son optimisme à tout épreuve (même à l'épreuve de mon perfectionnisme à tendance pessimiste). Merci à Luc pour nos discussions si enrichissantes et ton humour pince-sans-rire. Je souhaite remercier aussi l'équipe BKB du Labri dans son ensemble et chaque membre particulièrement pour nos moments d'échanges en GT ou au détour d'un couloir. Au Labri, sans bureau fixe après mon séjour au Canada, j'ai été accueilli au bureau 325. J'y ai rencontré des personnes formidables qui ont été de véritables appuis en cette fin de thèse. Merci à Théo, Clara, Rémi, Gabriel et Françoise. Merci Théo pour nos belles discussions où on refaisait le monde, merci Gabriel pour ton écoute et ta patience dans ce bureau survolté, merci Clara, Françoise et Rémi pour votre bonne humeur, j'aurais aimé mieux vous connaître toutes et tous plus tôt encore.

Au LaBRI comme à l'EDMI, j'ai été accompagnée par des femmes et des hommes formidables du personnel administratif qui ont permis à cette thèse de voir le jour, de se poursuivre et de prendre fin avec une écoute et un humanisme certains, merci à elles et eux.

À l'ADLab, arrivée après plus d'un an de doctorat, je débarquais un peu perdue. Certains anciens du laboratoire que j'avais déjà croisé à l'occasion de mon stage de master m'ont intégrée avec bienveillance, merci Julien, Mickaël, Simon et d'autres. J'ai rencontré aussi de nombreux doctorants devenus des amis qui se reconnaitront. Je souhaite remercier particulièrement les fidèles des soirées jeux de société où on lâchait prise et où la bienveillance et le rire étaient les maîtres mots, tout cela autour d'une (plusieurs) tasse(s) de thé et des jeux de collaboration jusque tard dans la nuit. Merci Elloise pour ta générosité, tes encouragements et nos discussions livresques. Merci Milan pour ton humour, ton auto-dérision. Merci Guillaume pour ton soutien, tes ramens et tes problèmes de mathématiques. Merci Damien que j'ai rencontré sur la fin de mon séjour canadien, mais qui a pourtant laissé une impression impérissable, merci pour ton enthousiasme et ton rire léger. Si vous passez en France, il y aura toujours une bonne tasse de thé et un jeu de cartes pour vous accueillir.

J'ai été bien entourée dans mes laboratoires d'affectation et certaines de ses personnes sont devenues des amis avec qui l'histoire ne fait que commencer, elles et ils se reconnaîtront, merci à vous.

J'ai aussi eu la chance d'avoir des proches qui m'ont soutenu malgré des périodes intenses, des déplacements internationaux, de longs silences.

Merci à ma bande d'amis infatigables du master, Coralie, Amélie et Vincent. Sans ces soirées jeux, ces escape game, bar à jeux, café céramique, café à chat et autres activités de la vie, je n'aurais pas été aussi résiliente.

Merci à Claire, cette femme formidable rencontrée lors d'un hiver canadien sous COVID en stage de master. Depuis, on ne se lâche plus. De retour en France, ce n'est pas 6h de route ou une thèse qui nous aurait séparés, certainement pas. Merci pour ton soutien, pour tes bons gâteaux, pour nos longues conversations.

Merci à Maxime dit Xima qui a été l'éclaireur de cette thèse, tes pep talks ont eu un impact positif indéniable.

Lors de mon séjour québecois, j'ai eu la chance de retrouver Amélie et Fabio, des amis de la famille qui m'ont accueilli et m'ont offert un soutien bienvenu, ils

m'ont sorti de ma coquille quand le moral n'allait pas fort. Merci Amélie pour ton enthousiasme et ta générosité, merci Fabio pour ces débats profonds pleins d'écoute et d'échanges sincères.

Présents depuis le début, et avant même que je ne sois consciente de mes propres aspirations, je veux remercier mes parents. Maman, Papa, je vous l'ai déjà dit (car pourquoi attendre ?) mais je ne vous remercierai jamais assez pour votre soutien sans failles. Ce soutien qui est passé aussi par une confiance presque aveugle en mes choix, car si vous aviez des doutes sur mes cheminements, vous ne me l'avez jamais fait sentir. J'ai pu assouvir ma curiosité de la vie et des sciences en sachant que j'avais avec moi deux piliers qui me portaient. Merci pour tout, ce doctorat, je vous le dédie.

Enfin, merci à mon compagnon, Brian. Tu as été mon roc, l'optimisme qui submergeait mon pessimisme, la vague calme et à l'écoute dans les bons comme les moins bons moments de cette thèse et de la vie. Tu as la patience d'un cours d'eau tranquille qui m'a porté à mon port. Je ne saurais te remercier assez pour tout ce que tu as fait pour moi.

Tout au long de ce doctorat, que ce soit dans ma vie professionnelle et personnelle, j'ai été entourée de femmes brillantes et persévérantes. C'est aussi grâce à ces beaux exemples que je sais que je peux devenir qui je veux, comme je veux, quand je veux. Merci à elles.

«No woman should be made to fear that she was not enough.»

— Samantha Shannon, *The Priory of the Orange Tree*

Sommaire

Résumé ii

Abstract iii

Remerciements iv

Introduction 1

1 État de l'art 4

1.1 Données omiques et biomarqueurs 4

1.1.1 Données omiques 4

1.1.2 Transcriptomique 5

1.1.3 Biomarqueurs 6

1.2 Bases de données biologiques 7

1.2.1 Bases de données de références d'annotations 7

1.2.2 Bases de données de cohortes de patients atteints de cancer . . 9

1.2.3 Bases de données d'analyses 11

1.3 Apprentissage automatique 12

1.3.1 Classificateurs traditionnels 13

1.3.2 Classificateurs par réseaux de neurones 15

1.4 Méthodes de sélection de variables 15

1.4.1 Filtres, méthodes d'enveloppement et méthodes intégrées aux modèles 16

1.4.2 Méthodes (hybrides) ensemblistes de sélection de variables . . 20

1.4.3 Sélection de variables et AutoML 22

1.5 Perturbations de données par ré-échantillonnage 24

1.6 Évaluation . . . 25
1.6.1 Évaluation de modèles . . . 25
1.6.2 Évaluation qualitative de signatures . . . 27
1.7 Méthodologie pour la conception d'un système de visualisation . . . 29
1.8 Éléments de visualisation . . . 30
1.8.1 Attributs visuels . . . 30
1.8.2 Visualiser des données relationnelles . . . 31

2 Développement d'une approche hybride ensembliste de sélection de variable (HEFS) 33
2.1 Introduction . . . 33
2.2 Vue d'ensemble de la stratégie . . . 34
2.3 Filtrage . . . 36
2.4 Stratégies d'échantillonnages . . . 37
2.5 Entraînement de modèles d'enveloppement . . . 38
2.6 Agrégation . . . 42
2.7 Conclusion . . . 47

3 Évaluation de l'approche HEFS 49
3.1 Introduction . . . 49
3.2 Jeux de données . . . 50
3.2.1 Cancer colorectal . . . 50
3.2.2 Cancer du rein, des poumons et de l'utérus . . . 52
3.3 Évaluation quantitative du mode standard HEFS : performance des modèles d'enveloppement . . . 53
3.4 Évaluation qualitative du mode standard HEFS : caractérisation des signatures . . . 55
3.4.1 Taille variable des signatures des modèles d'enveloppement . . . 55
3.4.2 Agrégation des signatures . . . 56
3.4.3 Caractérisation des signatures stables des 4 scénarios HEFS . . . 59
3.4.4 Évaluation indépendante des signatures stables . . . 60
3.5 Évaluation de la méthode d'agrégation . . . 63
3.6 Généralisation de l'approche HEFS en mode *light* . . . 65

3.6.1 Application à trois cas d'usage . . . 65
3.6.2 Efficacité computationnelle des modes standard et *light* . . . 69
3.7 Synthèse : Trois points critiques de l'approche HEFS . . . 69
3.7.1 Réduction des dimensions . . . 69
3.7.2 Perturbation des données : les stratégies d'échantillonnage . . . 71
3.7.3 Intégration des perturbations : la méthode d'agrégation . . . 71
3.8 Conclusion . . . 72

4 Système de visualisation d'un atlas de biomarqueurs 73
4.1 Introduction . . . 73
4.2 Conception d'un système de visualisation pour exploiter l'atlas de biomarqueurs . . . 75
4.2.1 Vue d'ensemble du système proposé . . . 75
4.2.2 Définition des tâches du système de visualisation . . . 77
4.3 Interfaçage du système de visualisation : THe Biom . . . 81
4.3.1 Aperçu de THe Biom . . . 82
4.3.2 Architecture de l'application . . . 83
4.4 Création de l'atlas . . . 85
4.4.1 Analyses intégrées a posteriori . . . 85
4.4.2 Jeux de données . . . 86
4.4.3 Mode *light* HEFS . . . 87
4.4.4 Données d'annotation de voies biologiques . . . 88
4.5 Conclusion . . . 88

5 Analyses réalisées avec THe Biom 90
5.1 Introduction . . . 90
5.2 Identification et analyse de biomarqueurs spécifiques d'un phénotype 91
5.2.1 Signatures intégrées dans l'atlas, issues de 6 cancers . . . 91
5.2.2 Cas d'usage du cancer du foie de stade II . . . 92
5.3 Biomarqueurs et voies biologiques spécifiques d'un cancer indépendamment du stade . . . 94
5.3.1 Biomarqueurs de l'adénocarcinome des poumons . . . 95

5.4 Biomarqueurs et voies biologiques spécifiques d'un stade d'avancement de cancer indépendamment du type de cancer 98
5.4.1 Analyse globale des similarités par stade 98
5.4.2 Caractérisation du stade III des cancers 99
5.5 Analyse ciblée d'un gène ou d'une voie biologique d'intérêt dans l'atlas 103
5.5.1 Analyse à partir de la sélection d'un gène 103
5.5.2 Analyse à partir de la sélection d'une voie biologique 104
5.6 Conclusion 105

Conclusion et Perspectives **106**

Bibliographie **109**

Annexe A Valeurs d'hyperparamètres des algorithmes d'apprentissage automatique **125**
I Approche HEFS en mode standard 125
II Approche HEFS en mode léger 128

Annexe B Clusters d'échantillons pour la cohorte CRC de TCGA **130**
I Clusters pour échantillons de stade IV en scénario de filtre par variance 131
II Clusters pour échantillons normaux en scénario de filtre par variance 132
III Clusters pour échantillons de stade IV en scénario de filtre par analyse DEG 133
IV Clusters pour échantillons normaux en scénario de filtre par analyse DEG 134

Annexe C Agrégation : les étapes du processus **135**
I Première étape 135
II Deuxième étape 136

Liste des figures

1.1 Développement d'un classificateur depuis la sélection d'un algorithme et son entraînement et test sur des données connues étiquetées jusqu'à sa prédiction de classe sur des données inconnues non étiquetées. . . . 13
1.2 Les trois types principaux de sélection de variables : filtre, méthode d'enveloppement et méthode intégrée. 16
1.3 Deux catégories d'approches ensemblistes de sélection de variables. (A) Perturbation fonctionnelle par application puis concertation de divers algorithmes de sélection de variables à un même jeu de données. (B) Perturbation des données par application d'un unique algorithme de sélection de variables à divers sous-ensembles d'un jeu de données après obtenus après partitionnement. 21
1.4 Matrice de confusion répertoriant les échantillons correctement ou incorrectement classés en classification binaire. 25
1.5 Classification de l'efficacité de treize attributs visuels en fonction du type de données à visualiser. Les attributs en gris sont jugés non pertinent pour la classe de données considérée. Figure extraite de [112] et adaptée de l'étude de Mackinlay [113]. 31

2.1 Vue générale de l'approche proposée de sélection de variable hybride ensembliste. Processus en quatre étapes : filtrage, échantillonnage, entraînement des modèles d'enveloppement, agrégation. Le filtrage appliqué aux données est basé soit sur l'analyse des gènes différentiellement exprimés (DEG), soit sur l'analyse de la variance (Var). L'étape d'échantillonnage est effectuée par séparation répété pour calculer 10 paires différentes d'ensembles d'entraînement et de test pour l'étape de sélection de variable par enveloppement. Cette séparation est basée soit sur stratégie d'échantillonnage stratifié aléatoire (R-S), soit sur un échantillonnage stratifié équilibré par distribution (DB-S). L'approche se poursuit par une étape à large échelle d'entraînement de modèles d'enveloppement et se conclut par une étape d'agrégation. 35

2.2 Processus d'entraînement en apprentissage automatique basé sur le logiciel BioDiscML. Les données d'entrée pour une exécution d'échantillonnage sont composées de patients décrits par F_x variables. Ces données sont divisées en un ensemble de test (1/3)) et un ensemble d'entraînement (2/3). L'ensemble d'entraînement subit un processus de classement des variables basé sur la mesure du gain d'information, où les variables ayant un score de zéro sont éliminées. Les variables résultantes $F_{x'}$ sont ensuite soumises à une boucle d'entraînement par sélection pas à pas ascendante combinée à une élimination pas à pas descendante. Chaque modèle entraîné résultant passe par diverses étapes de validation croisée et pour chacun d'eux, le score de MCC moyen (AVG MCC) et l'écart type associé (STD MCC) sont calculés. Les variables redondantes, écartées pendant le processus d'entraînement, sont recaptées pour une exploitation ultérieure. 39

2.3 Carte de chaleur des scores de stabilité moyenne de Nogueira pour tous les modèles retenus au cours des 10 exécutions d'échantillonnage et pour les 4 scénarios HEFS (DEG DB-S, DEG R-S, Var DB-S, Var R-S). 43

2.4 (A) Agrégation à trois niveaux : (A.1) niveau type de classificateur (entre modèles d'un même type de classificateur et par exécution d'échantillonnage), (A.2) niveau d'échantillonnage (entre types de classificateurs pour une exécution d'échantillonnage donné), (A.3) consensus global (à travers les échantillonnages). (B.1) Premier niveau : processus spécifique à chaque type de classificateur réalisant l'union des signatures des modèles d'enveloppement ayant passé les filtres de performance. (B.2) Deuxième niveau : agrégation par échantillonnage en calculant l'intersection par échantillonnage entre les listes de variables obtenues à partir des différents types de classificateurs au niveau précédent. (B.3) Troisième niveau : agrégation des listes de variables par échantillonnage du niveau précédent pour obtenir une signature de variables stable. Cette signature est composée des variables incluses dans au moins 40 ou 50% des échantillonnages pour le mode standard et *light* respectivement. 46

3.1 Scores de performance calculés pour les 4 scénarios de sélection de variables hybrides ensemblistes incluant les 10 exécutions d'échantillonnage avec une moyenne de 88 800 modèles entraînés chacune (pour huit classificateurs différents). Le panneau supérieur montre la distribution des valeurs du Coefficient de Corrélation de Matthews Moyen (AVG MCC), où une valeur de 1,0 indique des modèles de haute performance. Le panneau inférieur présente l'écart type par rapport à l'AVG MCC (STD MCC), où une valeur de 0,0 est la meilleure, car elle indique une performance égale dans toutes les évaluations. Dans chaque panneau, la ligne pointillée noire met en évidence la valeur seuil appliquée pour filtrer les modèles de faible performance pour les étapes HEFS suivantes. Les seuils sont fixés à un minimum de 0,7 pour l'AVG MCC et à un écart type maximal associé de 0,1 pour le STD MCC. 54

3.2 Variations des longueurs des signatures sans et avec redondance. Chaque boîte à moustaches représente la longueur moyenne des signatures pour chacune des 10 séries d'échantillonnage. 56

3.3 Distribution des quatre signatures HEFS. (1 à 4) Distribution des variables à travers toutes les exécutions d'échantillonnage pour les quatre scénarios HEFS basée sur les listes de variables obtenues après la deuxième étape d'agrégation, c'est-à-dire échantillon par échantillon. Le graphique final (panneau 5) correspond à la comparaison des quatre signatures obtenues à l'étape finale du processus d'agrégation de chaque scénario HEFS (en bleu dans les graphiques 1 à 4). 59

3.4 Importance normalisée des variables dans chacun des quatre modèles d'apprentissage automatique optimisés entraînés sur les données de cancer colorectal TCGA à l'aide de la bibliothèque R h2o. Le jeu de données TCGA a été réduit en fonction des quatre scénarios différents de type HEFS. La variable la plus importante pour un modèle a un score d'un et les autres scores sont normalisés et classés par rapport à celui-ci. Un score supérieur à 0 est indiqué par une barre pleine, un score nul est mis en évidence par un 0 tandis que l'absence de la variable pour le modèle considéré est représentée par un symbole X. . 62

3.5 Comparaison de la composition des signatures entre une stratégie de sélection de variables standard (gris clair) et l'approche HEFS proposée (gris foncé). 64

4.1 Figure extraite de [134]. Distribution du nombre de voies métaboliques (sur un total de 114 voies) KEGG dérégulées dans aucun type de cancer, dans un seul type, dans plusieurs ou dans les 26 types de cancers étudiés. 74

4.2 Schéma de la visualisation choisie pour la vue globale. 78

4.3 Schéma de la visualisation choisie pour la vue multi-signature. . . . 79

4.4 Schéma de la visualisation choisie pour la vue mono-signature. . . . 80

4.5 Vue d'expression génique basée sur des graphiques de boîtes à moustaches. 81

4.6 Aperçu de la page principale de l'application **THe Biom**. 82

4.7 Maquette de **THe Biom** avec sa page principale et ses cinq panneaux de menu et visualisation ainsi que la page secondaire de documentation accessible par le biais d'un bouton. 83

4.8 Architecture de l'application web **THe Biom** et ses composantes Front End et Back End. 84

4.9 Vue d'ensemble des analyses intégrées *a posteriori* et des données d'annotations agrégées à l'application **THe Biom**. (A) Définition des signatures robustes de biomarqueurs de multiples comparaisons de phénotypes oncologiques par notre approche hybride ensembliste de sélection de variables en mode *light*. (B) Mobilisation de la base de données Reactome pour l'obtention des données d'annotations de voies biologiques reliées aux signatures. (C) Intégration de ces données à **THe Biom** (**T**CGA **H**EFS **B**iomarkers). Aperçu de la page principale de l'application, du panneau de menu et des quatre vues d'exploration. 86

5.1 Signatures disponibles par type de cancer (BRCA : sein , HNSC : tête et cou, KIRC : rein, LIHC : foie, LUAD : poumons, UCEC : utérus), comparaisons (N vs Sx désigne une comparaison d'échantillons N normaux contre des écantillons de stade S avec x allant de I à IV) et type de scénarios d'approche HEFS (par DEG ou Var). 91

5.2 Vue mono-signature de THe Biom après sélection de la signature *merge* du cancer LIHC avec une comparaison Normal contre Stade II. 93

5.3 Analyse d'enrichissement d'annotation fonctionnelle des biomarqueurs obtenus à partir de la comparaison du cancer du foie : Normal versus stade II. 94

5.4 Extrait de la vue générale avec focus sur les signatures *merge* de l'adénocarcinome des poumons (LUAD). Les noeuds représentent les signatures, alors que les liens correspondent aux biomarqueurs en commun et leur épaisseur est proportionnelle au nombre de biomarqueurs communs. 95

5.5 Extrait de la vue multi-signature après sélection dans le menu latéral gauche de l'adénocarcinome des poumons (LUAD) en scénario dit *merge* et intégration des 4 comparaisons disponibles. 97

5.6 Vue multi-signature par stade d'avancement des cancers : Stade I (panneau A), Stade II (panneau B), Stade III (panneau C), Stade IV (panneau D). Les différents cancers sont : BRCA (sein), HNSC (tête et cou), KIRC (rein), LIHC (foie), LUAD (poumons), UCEC(utérus). 98

5.7 Visualisation des six signatures de cancers de stade III incluses dans la preuve de concept de **THe Biom**. (A) Vue multi-signature de **THe Biom** des six signatures de cancer de stade III et des voies biologiques qui leur sont associées. (B) Vue multi-signature réduite aux quatre signature de cancers de stade III partageant un gène avec la voie biologique *Platelet degranulation*. Les quatre gènes impliqués sont colorés dans la vue et dans le menu de sélection de gène (en bas à gauche du panneau B). (C) Niveaux d'expression des quatre gènes pour les diverses classes d'échantillons disponibles pour les quatre type de cancer retenus. Les boîtes à moustaches des classes Normal et Stade III sont encadrées de noir suivant l'implication du gène dans la signature qui résulte de la comparaison des échantillons de ces classes. 100

5.8 Capture d'écran de THe Biom à partir de la sélection du gène MMP11 dans le panneau de saisie textuelle à gauche. Les vues multi-signature (en haut à droite) et d'expression de gènes sont mis à jour à partir de la sélection. Le gène est colorié en bleu dans la vue multi-signature. Les boîtes à moustache d'expression de gènes pour les phénotypes inclus dans les comparaisons pour lesquelles le gène fait partie de la signature sont encadrées de noir. 104

5.9 Capture d'écran de **THe Biom** à partir de la sélection de la voie biologique de la cascade des kinases RAF/MAP dans le panneau de saisie textuelle à gauche. Les vues multi-signature (en haut à droite) et d'expression de gènes sont mises à jour. Les boîtes à moustache d'expression de gènes pour les phénotypes inclus dans les comparaisons pour lesquelles les gènes font partie de la signature sont encadrées de noir. 105

B.1 Clustering hiérarchique d'échantillons de stade IV à partir de données CRC TCGA filtrées par variance en utilisant des données d'expression de gènes et en se basant sur la distance euclidienne et la méthode de Ward pour le lien. Les échantillons de patients sont représentés par la vue verticale tandis que les gènes composent la vue horizontale. Deux groupes d'échantillons sont définis. 131

B.2 Clustering hiérarchique d'échantillons normaux à partir de données CRC TCGA filtrées par variance en utilisant des données d'expression de gènes et en se basant sur la distance euclidienne et la méthode de Ward pour le lien. Les échantillons de patients sont représentés par la vue verticale tandis que les gènes composent la vue horizontale. Trois groupes d'échantillons sont définis. 132

B.3 Clustering hiérarchique d'échantillons de stade IV à partir de données CRC TCGA filtrées par analyse de gènes différentiellement exprimés en utilisant des données d'expression de gènes et en se basant sur la distance euclidienne et la méthode de Ward pour le lien. Les échantillons de patients sont représentés par la vue verticale tandis que les gènes composent la vue horizontale. Deux groupes d'échantillons sont définis. 133

B.4 Clustering hiérarchique d'échantillons normaux à partir de données CRC TCGA filtrées par analyse de gènes différentiellement exprimés en utilisant des données d'expression de gènes et en se basant sur la distance euclidienne et la méthode de Ward pour le lien. Les échantillons de patients sont représentés par la vue verticale tandis que les gènes composent la vue horizontale. Deux groupes d'échantillons sont définis. 134

Liste des tableaux

1.1 Les trois catégories d'annotations du système de classification TNM des tumeurs solides. 10
1.3 Outils d'automatisation d'entraînement de modèles d'apprentissage automatique. Caractérisation par deux critères : intégration d'une étape de sélection de variables et fonctionnalité d'extraction de tous les modèles entraînés. 24
1.4 Exemple d'en-tête de tableau pour la mise en place d'un suivi de développement d'une solution visuelle par le concept de Munzner. . . 30

2.1 Exemple de format des données transcriptomiques employées dans notre projet : chaque ligne est un échantillon, chaque colonne présente le nombre de *reads* obtenues pour chaque gène codant ou non. 36
2.2 Comparaison des signatures de variables de trois modèles hautement performants et entraînés par notre stratégie d'entraînement de modèles d'enveloppement. 41

3.1 Nombre total de modèles d'enveloppement entraînés par scénario (lignes) et classificateur (colonnes), ainsi que le nombre total de modèles entraînés par scénario et le nombre total de modèles conservés après application du filtre de performance (MCC moyen de 0,7 ou plus et écart type du MCC de 0,1 ou moins). 53
3.2 Tailles moyennes des signatures transitoires de l'étape 1 d'agrégation à travers les 10 exécutions d'échantillonnage et pour chaque type de classificateur et scénario. 57
3.3 Tailles moyennes des signatures transitoires de l'étape 2 d'agrégation à travers les 10 exécutions d'échantillonnage et pour chaque scénario. 58

3.4 Caractérisation des quatre signatures HEFS par analyse d'enrichissement d'annotation du terme DisGeNet "Colorectal cancer" par l'outil EnrichR. 60

3.5 Évaluation des différentes signatures sur des modèles optimisés et testés sur les données d'origine en validation croisée et sur un jeu de données externe, GSE50760. GBM : *Gradient Boosting Machine*, GLM : *Generalized Linear Model* 61

3.6 Analyse de l'enrichissement des annotations DisGeNet pour caractériser les signatures stables obtenues par le mode *light* HEFS dans les scénarios Var DB-S et DEG DB-S dans le contexte du cancer colorectal de stade IV (CRC), du cancer du rein à cellules claires de stade I (KIRC), de l'adénocarcinome du poumon (LUAD) et du carcinome du corps de l'utérus et de l'endomètre de stade I (UCEC). Les gènes en gras sont ceux qui sont impliqués dans l'annotation associée. . . . 68

4.1 Modèle de Munzner pour la conception de système de visualisation: l'abstraction des tâches et des données en quatre vues distinctes ainsi que les métaphores visuelles proposées. 77

4.2 Nombre d'échantillons normaux et de stade I à IV pour les différents types de cancer dont les données TCGA sont analysées dans le projet **THe Biom**. 87

5.1 Indice de Jaccard pour comparer les quatre signatures de biomarqueurs de l'adénocarcinome des poumons (LUAD) à divers stades (SI à SIV). 96

5.2 Liste des 23 voies biologiques qui incluent au moins un gène provenant d'au minimum deux signatures différentes de cancer de stade III, ainsi que le nombre total de gènes associés, le nombre de signatures et la valeur de l'indice de diversité de Shannon correspondant. 102

A.1 Classificateurs d'apprentissage automatique sélectionnés pour l'approche HEFS en mode standard, leur nom de fonction weka associé et la grille d'hyperparamètres choisie. 128

A.2 Classificateurs d'apprentissage automatique sélectionnés pour l'approche HEFS en mode léger, leur nom de fonction weka associé et la grille d'hyperparamètres choisie. 129

C.1 Taille des signatures transitoires pour chacune des 10 exécutions d'échantillonnage obtenues à l'issue de la première étape d'agrégation pour le scénario DEG DB-S. 135
C.2 Taille des signatures transitoires pour chacune des 10 exécutions d'échantillonnage obtenues à l'issue de la première étape d'agrégation pour le scénario DEG R-S. 135
C.3 Taille des signatures transitoires pour chacune des 10 exécutions d'échantillonnage obtenues à l'issue de la première étape d'agrégation pour le scénario Var DB-S. 136
C.4 Taille des signatures transitoires pour chacune des 10 exécutions d'échantillonnage obtenues à l'issue de la première étape d'agrégation pour le scénario Var R-S. 136
C.5 Taille des signatures transitoires pour chacune des 10 exécutions d'échantillonnage obtenues à l'issue de la deuxième étape d'agrégation pour le scénario DEG DB-S. 136
C.6 Taille des signatures transitoires pour chacune des 10 exécutions d'échantillonnage obtenues à l'issue de la deuxième étape d'agrégation pour le scénario DEG R-S. 137
C.7 Taille des signatures transitoires pour chacune des 10 exécutions d'échantillonnage obtenues à l'issue de la deuxième étape d'agrégation pour le scénario Var DB-S. 137
C.8 Taille des signatures transitoires pour chacune des 10 exécutions d'échantillonnage obtenues à l'issue de la deuxième étape d'agrégation pour le scénario Var R-S. 137

Introduction

Depuis toujours, la médecine cherche à mieux comprendre et soigner les maladies. L'une de ces pathologies est le cancer qui est, avec les maladies cardio-vasculaires, l'une des principales causes dans le monde de décès prématurés [1]. En 2019, on estimait qu'un être humain avait 20% de risque de développer un cancer entre la naissance et 75 ans [2]. L'Organisation Mondiale de la Santé OMS définie le cancer comme suit : "le cancer englobe un vaste groupe de maladies qui peuvent apparaître dans presque tous les organes ou tissus du corps, lorsque des cellules anormales se développent de manière incontrôlée et se répandent au-delà de leurs limites habituelles pour envahir des régions voisines du corps et/ou se propager à d'autres organes". Les causes du cancer peuvent être variées (génétique, environnement, habitudes de vie)[3] et même si la mortalité liée au cancer a baissé depuis une trentaine d'années [4] son diagnostic et son traitement sont encore perfectibles.

La médecine est un domaine scientifique millénaire qui a coévolué avec le développement technologique pour devenir tel que nous le connaissons aujourd'hui [5]. Ces avancées techniques ont amélioré notre compréhension des mécanismes biologiques en santé humaine et ont notamment permis une meilleure prise en charge du cancer. Par exemple, une meilleure compréhension du génome humain et des techniques de clonage de gènes ont abouti à l'identification d'une famille d'oncogènes, la famille des gènes RAS [6]. Un autre exemple peut être cité avec le cancer du sein où le séquençage de biomarqueurs, comme la PAM50, permet aujourd'hui de mieux caractériser les sous-types de cancer du sein et de leur associer une prise en charge thérapeutique adaptée, telle que l'immunothérapie ou la chimiothérapie [7].

À la fin du 20ème siècle, les techniques de séquençage émergent et leur évolution jusqu'à ce jour a permis d'engranger de grands volumes de données à divers niveaux moléculaires et associés à de nombreux phénotypes différents [8]. Cette

importante source d'informations difficilement exploitable a requis la mobilisation de bioinformaticiennes et bioinformaticiens, des scientifiques ayant les compétences computationnelles pour traiter ces données. Ces capacités, mises au service de la recherche biomédicale et en lien étroit avec la recherche en informatique, ont permis d'aboutir à des méthodes poussées de traitement de ces données. Nous sommes maintenant capables d'extraire des variables plus ou moins fiables et explicatives d'un phénotype parmi les milliers fournies à l'origine. On parle d'approches de sélection de variables [9]. Les différentes méthodes de sélection de variables peuvent avoir des avantages et inconvénients selon les algorithmes sur lesquelles elles reposent, mais également en fonction des données sur lesquelles elles sont appliquées. Dans ce contexte, le développement d'une méthode ensembliste pour la sélection de variables présente plusieurs avantages, notamment dans le cas de l'analyse de données complexes en recherche scientifique et biomédicale. Une approche ensembliste vise à combiner les résultats de plusieurs méthodes pour améliorer la fiabilité de la sélection. L'utilisation intégrée de plusieurs méthodes permet alors de réduire les biais individuels de chacune, augmentant ainsi la robustesse des résultats. Pour résumer, les méthodes ensemblistes partent du postulat que la variabilité induite par diverses méthodes et/ou données est une force à exploiter conjointement plutôt qu'une faiblesse à lisser [10]. Les conclusions tirées de telles méthodes sont, au moins d'un point de vue théorique, davantage fiables que des conclusions issues d'un unique composant décisionnaire. Les recherches autour de ces méthodes ont connu un essor certain ces dernières années pour déterminer les meilleures configurations de protocoles.

Malgré l'important volume de données publiques, il existe peu d'efforts visant à rendre accessibles des analyses systématiques de données d'oncologie pour divers phénotypes en utilisant des méthodes avancées de sélection de variables.

Le présent manuscrit décrit les contributions scientifiques que nous avons produites pour apporter des solutions à ce besoin. Ce travail de recherche a été fourni dans le cadre d'une thèse en co-tutelle entre l'Université de Bordeaux en France et l'Université Laval au Québec, Canada.

La problématique de notre projet de recherche était de mettre en place des solutions bioinformatiques pour identifier et rendre accessibles des signatures de biomar-

queurs transcriptomiques robustes en tenant compte de la difficulté de choisir des méthodes de sélection de variables appropriées. Notre projet de recherche s'articule ainsi autour de deux objectifs :

1. Développement d'une méthode bioinformatique pour l'identification de biomarqueurs robustes de maladies complexes comme le cancer,

2. Définition d'un système de visualisation pour la mise à disposition et l'analyse d'un atlas de signatures robustes de biomarqueurs transcriptomiques de types et stades divers de cancers.

Le premier chapitre de ce manuscrit propose un état de l'art des notions scientifiques multi-disciplinaires nécessaires à la réalisation des objectifs de cette thèse. Nous aborderons ainsi la notion de données dites omiques et de biomarqueurs, la définition et l'intérêt des méthodes d'apprentissage automatique ainsi que des méthodes de sélection de variables. Nous poursuivrons avec un point sur les méthodes de ré-échantillonnages au service des procédés de perturbations des données suivi par un état de l'art des méthodes connues d'évaluation de modèles et de signatures. Nous conclurons par une revue des bases de données biologiques à la disposition de tous et un exposé de bonnes pratiques de conception de solutions de visualisation.

Le deuxième et le troisième chapitre expose la méthodologie mise en place puis les résultats obtenus pour répondre à notre premier objectif.

Le quatrième et le cinquième chapitre décrivent l'outil développé puis les résultats préliminaires obtenus dans l'optique de répondre à notre deuxième objectif.

Ce manuscrit s'achèvera sur une conclusion générale de notre projet de recherche ainsi que des perspectives de celui-ci.

Chapitre 1

État de l'art

1.1 Données omiques et biomarqueurs

1.1.1 Données omiques

Pour comprendre le fonctionnement d'un organisme, la communauté scientifique s'intéresse de nos jours à l'étude de ce dernier à divers niveaux comme le systémique, tissulaire ou moléculaire. Le niveau moléculaire implique par exemple divers champs d'étude tels que l'étude de l'ADN, de l'ARN, des protéines ou même des métabolites. Ci-après, un résumé des avancées théoriques et technologiques qui nous permettent aujourd'hui d'étudier le vivant à l'échelle moléculaire.

Depuis des milliers d'années, l'humain s'interroge sur la notion d'hérédité. Cela a commencé, par exemple, par l'observation de traits d'intérêts dans des cultures et leur reproduction sélective qui aboutissait à une conservation des traits souhaités [11]. Il a fallu attendre 1865 et l'établissement des lois de Mendel pour commencer à entrevoir les premiers fondements de la compréhension moderne de l'hérédité [12]. En 1905, Adam Sedwick donne un nom aux sciences biologiques visant à étudier l'hérédité et la variation : la génétique [13]. Ce domaine d'étude, qui est finalement davantage défini par l'étude des gènes, va poursuivre son développement et connaître un accroissement important avec l'amélioration de la technologie. Vers les années 1980, les premières techniques de séquençage d'ADN émergent et la génomique comme domaine de recherche à part entière voit le jour. Il s'agit alors d'étudier les génomes dans leur entièreté et non plus seulement les gènes comme unité unique.

Au début des années 2000, le séquençage nouvelle génération (*Next Generation Sequencing NGS*) se répand à travers la communauté de recherche en génomique et d'autres communautés comme en transcriptomique ou en épigénomique [12, 14]. On peut désormais améliorer notre compréhension sur les relations entre génotype et phénotype.

Les années 1980-90 sont aussi une période de grandes avancées technologiques dans le domaine de l'étude des protéines et des métabolites. L'avènement de la spectrométrie de masse a ainsi permis l'analyse rapide de quantité de protéines [15] et métabolites [16].

De nos jours, ces différents champs d'étude que sont la génomique, la transcriptomique, la protéomique ou encore la métabolomique sont parfois rassemblées sous le terme "omiques".

1.1.2 Transcriptomique

Comme évoqué précédemment, la transcriptomique est l'étude des molécules d'ARN et plus précisément du transcriptome soit l'ensemble des molécules d'ARNs présentes dans une cellule : ARNs messagers et non codants. L'étude du transcriptome permet en général d'approcher l'expression des gènes dans un échantillon et leurs régulations mais aussi des évènements d'épissage alternatif ou l'exploration de phénomènes de polymorphismes mononucléotidiques [17]. Nous avons acquis, durant les vingts dernières années, une meilleure compréhension du transcriptome à travers divers techniques de NGS dont la plus couramment utilisée en laboratoire est le séquençage de l'ARN (*Sequencing RNA RNA-seq*). Son coût et son temps d'exécution a diminué avec les années et ses multiples usages en font une technologie systématiquement utilisée par la communauté de recherche en biologie [18]. Dans l'ensemble, les analyses de transcriptomiques employant des techniques RNA-seq consistent à séquencer les molécules d'ARN extraites des cellules pour générer ce que l'on appelle des *reads*, c'est-à-dire des courtes séquences d'ARN. Ensuite, ces *reads* sont alignées sur un génome de référence dans le but de déterminer leur origine exacte, et d'en évaluer à la fois la qualité et la quantité. Cela permet d'identifier les gènes exprimés et d'obtenir une vision détaillée de l'activité transcriptionnelle d'un échantillon de cellules.

Une méthode couramment employée pour étudier l'expression de gènes dans un échantillon à partir de données RNA-seq est l'analyse de gènes différentiellement exprimés (*differentially expressed genes* DEG) [19]. Cette analyse permet d'estimer la variation statistiquement significative d'expression d'un gène entre deux classes d'échantillons. L'utilisation de cette technique peut permettre d'identifier des gènes présentant une significativité statistique élevée et/ou un ratio d'expression substantiel entre les deux classes.

1.1.3 Biomarqueurs

La recherche biomédicale, comme d'autres domaines de la biologie, bénéficie de l'expansion du nombre de données omiques produites. Ces dernières permettent d'identifier des biomarqueurs. Notons qu'un biomarqueur peut être défini comme "une caractéristique objectivement mesurée et évaluée en tant qu'indicateur de processus biologiques normaux, de processus pathogènes ou de réponses pharmacologiques à une intervention thérapeutique" [20]. Une signature de biomarqueurs est une liste de biomarqueurs pouvant être distincts et/ou reliés entre eux [21]. Ainsi, une signature de biomarqueur peut servir à de multiples usages [22]: prédire le développement d'une maladie, diagnostiquer une maladie, évaluer le pronostic d'un patient, surveiller une maladie, prédire la réponse à un traitement, évaluer la réponse à un traitement, évaluer la sécurité d'un traitement.

Certaines études récentes ont permis l'identification de signatures de biomarqueurs de transcriptomique actuellement utilisés en routine clinique. Parmi ceux-ci, nous pouvons citer, dans le cadre du cancer du sein, la MammaPrint qui est une aide au pronostic, l'Oncotype DX qui permet la surveillance de certains cas du cancer du sein en prédisant une possible récurrence ou encore la PAM50 qui permet de diagnostiquer la maladie en caractérisant son sous-type d'appartenance [23, 24].

Les données transcriptomiques peuvent ainsi être un véritable réservoir de biomarqueurs en recherche biomédicale. Leur exploitation reste un enjeu majeur pour comprendre et combattre des maladies complexes comme le cancer. Néanmoins, ces données peuvent être difficiles à analyser à cause de leurs grandes dimensions : quelques dizaines ou centaines de lignes d'échantillons et des dizaines de milliers de gènes (environ 60 000 pour le nombre total de gènes codants et non codants).

1.2 Bases de données biologiques

Avec le développement des données omiques est venue la nécessité de stocker, organiser et partager autant les données brutes que les analyses qui en découlent. En 2015 on dénombrait déjà plus de 1 500 bases de données recensées alors même que ce chiffre est probablement sous-estimé. En effet, certaines bases ne sont pas publiées dans des revues scientifiques [25]. La prolifération de ces bases rend leur navigation de plus en plus complexe et il est nécessaire d'identifier les besoins d'un projet de recherche pour sélectionner avec discernement les bases de données les plus adéquates. Ainsi, nous évoquerons ci-après certaines catégories d'intérêt de bases de données.

Les bases de données biologiques peuvent concentrer des informations d'annotation afin de caractériser biologiquement des ensembles de gènes comme des biomarqueurs potentiels. À ces bases de données biologiques s'ajoutent les bases de données de cohortes de patients qui mettent à disposition de la communauté de recherche des données de patients ou de modèles biologiques dans des *designs* d'expérimentation contrôlés. Nous pouvons aussi noter l'existence de bases de données d'analyses qui permettent l'exploration et l'extraction d'analyses déjà effectuées ou que l'utilisateur peut réaliser lui-même en ligne par des outils d'analyses intégrés.

Dans la suite de cette section nous faisons l'état de l'art de bases de données de références d'annotations (ontologie de gènes, voies et réseaux biologiques, pathologies), de bases de données de cohortes de patients atteints de cancers (TCGA, GEO) et de bases de données d'analyses.

1.2.1 Bases de données de références d'annotations

Ontologie de gènes

Parmi les bases de données de références d'annotation, l'ontologie de gènes (*gene ontology* GO) est parmi les plus utilisées. Les termes GO centralisent des annotations de trois types [26, 27]:

- processus biologique : procédé auquel des gènes ou leurs produits participent et qui, souvent, impliquent des transformations chimiques ou physiques ;

- fonction moléculaire : décrit l'activité biochimique d'une molécule issue d'un gène, comme la liaison spécifique à des ligands ou structures ;
- composant cellulaire : se réfère à l'endroit dans la cellule où une molécule issue d'un gène est active.

Ces annotations n'incluent pas d'information sur le type de relation entre les différents éléments qu'elles intègrent.

Voies et réseaux biologiques

Les annotations de voies et de réseaux biologiques peuvent être définies comme étant la formalisation des liens entre des gènes, protéines et métabolites qui participent à une fonction biologique [28]. Les voies biologiques définissent ces liens par des réactions chimiques, des évènements de régulation et de signalisation. Les réseaux sont à une échelle plus grande et impliquent parfois une abstraction de processus cellulaires complexes.

Ces types d'annotation font l'objet de nombreuses recherches et de multiples bases de données leur étant consacrées ont vu le jour telles que Reactome [29] ou KEGG [30]. Ces bases de données sont très couramment utilisées lors d'analyses d'annotation et leur contenu tend à se recouper. Cette tendance touche aussi d'autres bases de données de voies et réseaux biologique telles que WikiPathways [31] ou PANTHER [32]. Récemment, pour tirer parti de cette grande masse de données d'annotation et de leur potentiel commun, des efforts tels que Pathway Commons ont vu le jour [33]. Cette base de données permet de collecter et intégrer les annotations de voies et réseaux biologiques de multiples bases de données. Néanmoins, il n'existe pas encore d'outil général d'analyse d'annotations intégrant la base de données Common Pathways.

Pathologies

Pour évaluer l'intérêt clinique d'une signature de biomarqueurs (et notamment de gènes), il est intéressant de savoir si un ou plusieurs d'entre eux sont déjà connus pour être associé(s) à une maladie d'intérêt ou une autre, proche de la maladie cible. Des bases de données commes DisGeNet [34] ou OMIM [35] concentrent des

informations de liens entre gènes et maladies et sont fréquemment intégrées dans les outils généraux d'analyse d'annotation.

1.2.2 Bases de données de cohortes de patients atteints de cancer

La présente section ainsi que le reste du manuscrit se concentre sur la maladie du cancer. Nous exposons ici deux bases de données, TCGA et GEO, qui sont source de données d'oncologie.

TCGA

En oncologie, le consortium *The Cancer Genome Atlas* TCGA est très connu pour son effort considérable de répertorier des données cliniques et génomiques de dizaines de cohortes de patients atteints de types de cancer divers [36]. Au fil des années, le consortium a étendu sa méthodologie à la récolte de données d'autres omiques comme la transcriptomique. C'est à ce jour, une des plus grandes ressources oncologiques diversifiée et standardisé à disposition de la communauté. Une grande partie des données sont publiques comme les données de comptage de *reads* en RNA-seq.

Des ressources programmatiques comme la bibliothèque TCGAbiolinks permettent d'explorer et extraire de façon automatique et stratifiée des données du portail de données Genomic data Commons GDC qui abrite les données de TCGA [37].

GEO

La base de données *Gene Expression Omnibus* GEO centralise des données de microarray et de NGS d'études originales de la communauté de recherche scientifique [38]. GEO contient ainsi les données de milliers d'études, y compris en cancérologie sous des formats brutes et/ou traités. Néanmoins, ces données présentent souvent moins d'échantillons que dans des études comme celles menées dans le cadre de TCGA, mais elles peuvent toute fois être une ressource pertinente, notamment pour la validation de biomarqueurs. Récemment, il a été annoncé que GEO intègrerait un pipeline automatique de génération de comptage de *reads* pour toutes les données

brutes de RNA-seq que la plateforme héberge. Cela faciliterait la réutilisation de ces données [39].

Nomenclature de l'état d'avancement d'un cancer

Dans ces bases de données, les patients et leurs échantillons sont souvent associés à des données cliniques comme l'âge, le sexe, les symptômes, le traitement, mais aussi le stade d'avancement du cancer. Ce dernier est couramment déterminé par le système TNM pour de nombreux types de cancer à tumeur solide différents [40]. Ce système, développé par l'*Union for International Cancer Control* UICC en collaboration avec l'*American Joint Commission on Cancer* AJCC, classifie les cancers en fonction de trois aspects principaux (Tableau 1.1):

Catégorie	Notation
T (Tumeur primaire) Suivi de la taille de la tumeur primaire	TX : pas d'information T0 : pas d'évidence de tumeur primaire Tis : Carcinome in situ T1 à T4 : évalue le degrès d'invasion de la tumeur primaire
N (Ganglions lymphatiques) Migration de cellules cancéreuses à des ganglions lymphatiques	NX : pas d'information N0 : pas d'implication noté de ganglions lymphatiques N1 à N3 : Évidences de régions nodales envahies par du cancer
M (Métastases) Présence ou absence de métastases dans des sites et/ou tissus distants de l'aire locale de la tumeur	M0 : pas d'évidence de métastases distantes M1 : évidences de métastases distantes

Tableau 1.1: Les trois catégories d'annotations du système de classification TNM des tumeurs solides.

Ces trois aspects, que sont T, N et M, décrivent la tumeur d'un patient et leur combinaison permet d'attribuer un stade d'avancement de la tumeur pouvant prendre une valeur de I à IV. Le stade I classe une tumeur limitée au site d'origine, de petite taille, sans propagation aux ganglions lymphatiques ni métastases (T1 ou T2, N0, M0). Le stade II décrit quant à lui une tumeur plus grande ou avec une atteinte limitée des ganglions lymphatiques, mais sans métastases distantes (T2 ou T3, N0 ou N1, M0). Les cancers de Stade III ont une tumeur de taille importante

et/ou atteignant des ganglions lymphatiques régionaux plus éloignés (T3 ou T4, N1 à N3, M0) mais ne présentent toujours pas de métastases distantes. Enfin, le dernier stade d'avancement du cancer, le stade IV, décrit un cancer avec métastases à distance (M1), peu importe la taille de la tumeur ou l'atteinte loco-régionale des ganglions lymphatiques.

1.2.3 Bases de données d'analyses

Avec l'accès facilité à des données d'oncologie comme par le biais de bases de données comme TCGA ou GEO, nombre d'analyses de grande envergure (parfois même de pan-cancer) a été réalisé par la communauté de recherche. Ces analyses sont parfois centralisées dans des applications web qui permettent de les explorer ou de réaliser soi-même certaines analyses. Parmi ces plateformes, nous pouvons citer GEPIA2 [41], TACCO [42] ou encore UALCAN [43].

GEPIA2 est une application web, extension de GEPIA [44] qui intègre des données transcriptomiques de TCGA et d'une base de données d'expression de gène et de génotypage en tissus sains GTEx. GEPIA visait à rendre disponible des analyses comme celles d'expression différentielle de gènes ou de survie. GEPIA2 intègre en plus des analyses d'épissage alternatif et une possibilité d'explorer les données à travers le prisme des sous-types de cancer [41]. Par ailleurs, GEPIA2 permet à l'utilisateur de charger ses propres données d'expression génique pour les comparer à celles de TCGA et GTEx.

TACCO est une application web qui s'inscrit dans la continuité de GEPIA. Les développeurs de TACCO avaient pour ambition d'offrir plus qu'une simple analyse des liens entre l'expression d'un gène unique et la survie des patients. TACCO permet ainsi de créer des modèles complets d'analyse de survie. La plateforme permet aussi d'analyser des liens entre ARN messagers et micro ARNs. TACCO offre la possibilité à l'utilisatrice ou utilisateur de tester si une liste de gènes est pertinente pour différencier et classer divers phénotypes dont des stades précoces et avancés de cancer. Si la liste est trop importante, TACCO propose d'employer une unique méthode de filtre sans préciser laquelle.

UALCAN est une application web mettant à disposition des analyses d'expression de gènes basées sur les données de TCGA notamment et suivant divers critères

comme l'ethnicité, l'âge, le sexe, le type d'échantillon (normal ou tumoral) ou encore le stade du cancer. L'utilisateur doit entrer un ou plusieurs noms de gènes à explorer plus en détail.

Toutes ces bases de données se basent sur des analyses d'expression différentielle pour classer des gènes et parfois en sélectionner comme dans TACCO. En revanche, elles ne proposent pas de processus de sélection de variables avancée pour définir des signatures de gènes biomarqueurs d'un type de cancer ou d'un stade par exemple.

1.3 Apprentissage automatique

L'apprentissage automatique (ou *machine learning* ML) est un domaine de recherche qui suscite un intérêt accru depuis une vingtaine d'années. Cette technologie permet de faire de la prédiction en se basant sur des échantillons connus. Si les données connues ne sont pas étiquetées, l'algorithme tente de reconnaître des motifs cachés et de regrouper des échantillons similaires. Il s'agit alors d'apprentissage non supervisé comme le *clustering*. À l'inverse, si les données connues utilisées pour entraîner le modèle sont étiquetées, il s'agit d'apprentissage supervisé. Dans ce cas, si la variable à prédire est catégorielle, on parle de classification (Figure 1.1) et si la variable à prédire est numérique, alors il s'agit d'une régression. Enfin, si les données connues comportent des échantillons étiquetés et non étiquetés, l'algorithme fait partie de la catégorie de l'apprentissage semi-supervisé [45]. Nous nous concentrerons ici sur la classification.

Un classificateur repose sur un type d'algorithme donné et des hyper-paramètres dont les valeurs peuvent être explicitées avant l'entraînement et donc influencer son comportement d'entraînement. Mais un classificateur n'est pas explicitement programmé pour effectuer une tâche spécifique. En effet, il doit ajuster ses paramètres internes au cours de son entraînement. À noter que le développement de modèles ML nécessite d'obtenir, depuis les données connues, un sous-ensemble de données d'entraînement et un sous-ensemble de données de test pour évaluer les performances du modèle.

Comme mentionné précédemment, un des principaux but des classificateurs est de réaliser de la prédiction sur de nouvelles données. Néanmoins, une autre de ses

applications est la sélection de variables (voir section 1.4).

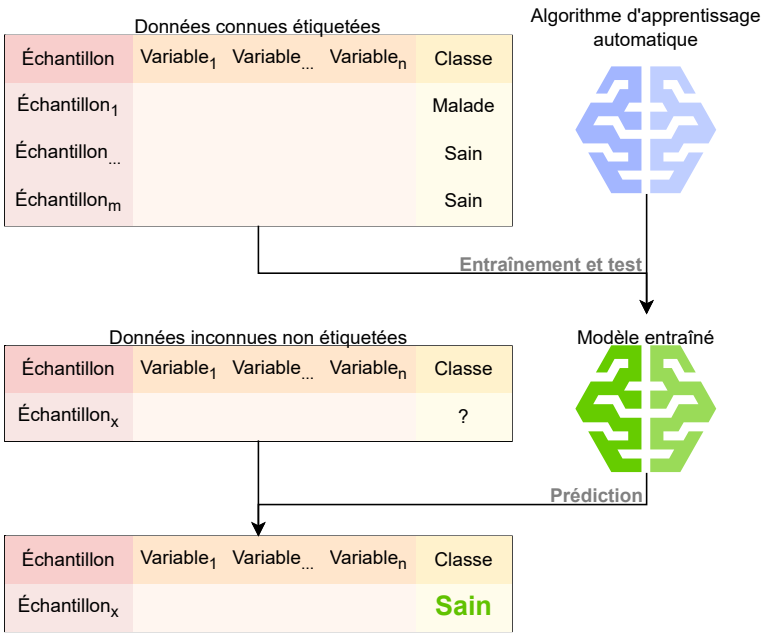

Figure 1.1: Développement d'un classificateur depuis la sélection d'un algorithme et son entraînement et test sur des données connues étiquetées jusqu'à sa prédiction de classe sur des données inconnues non étiquetées.

Ci-après, une description de différents types d'algorithmes traditionnels ou d'apprentissage profond et de leurs avantages et inconvénients.

1.3.1 Classificateurs traditionnels

Les algorithmes ML dits traditionnels ou conventionnels incluent des algorithmes comme les arbres de décision, les modèles bayesiens, les k-plus proches voisins (*k-nearest neighbors* kNN) ou encore les machines à vecteurs de support (*support vector machines* SVM).

Les arbres de décision se basent sur une architecture en arbre pour partitionner de manière répétée l'ensemble de données d'apprentissage en sous-ensembles basés sur les valeurs des variables jusqu'à ce qu'il n'y ait plus de segmentation possible ou jusqu'à ce qu'un critère d'arrêt soit atteint. Lorsque le modèle d'arbre est appliqué à un nouvel ensemble de données, il identifie le chemin dans l'arbre de décision qui correspond le mieux aux caractéristiques de l'instance et applique la classe dominante. Certains algorithmes d'arbres de décision plus sophistiqués sont souvent utilisés dans un contexte clinique, tels que les forêts aléatoires ou les *Extreme Gra-*

dient Boosting XGBoost. Une forêt aléatoire entraîne plusieurs arbres de décision, où chaque segmentation des données est effectuée de manière aléatoire, et détermine finalement la classe sélectionnée comme étant celle retenue par la majorité des arbres [46]. XGBoost est une méthode de boosting (construction séquentielle d'arbres) qui développe un modèle en ajoutant des arbres de décision successifs afin de corriger les erreurs des arbres précédents [47]. Parmi les autres types d'arbres de décisions parfois employés en recherche biomédicale, nous pouvons citer le *Simple Classification And Regression Trees* Simple CART ou le C4.5. Le premier utilise l'impureté de Gini pour discriminer les variables avant de construire le modèle [48], tandis que le second utilise un calcul d'entropie comme mesure de séparation.

Les modèles bayésiens quant à eux se basent sur le théorème de Bayes qui permet de calculer la probabilité d'un événement en mettant à jour une probabilité *a priori* lors de la confrontation avec de nouvelles données. Parmi les classificateurs probabilistes de Bayes, nous pouvons citer le Naive Bayes, le réseau bayésien ou encore les *Averaged 1-Dependence Estimators* (A1DE). Naive Bayes utilise le théorème de Bayes pour classer les échantillons, en appliquant la règle de Bayes qui donne une probabilité préalable aux nouveaux échantillons d'être étiquetés dans une classe donnée, puis choisit la classe avec la probabilité la plus élevée. Notons que Naive Bayes suppose que la valeur d'une variable est indépendante des autres variables [49]. Le réseau bayésien est une méthode probabiliste par graphe qui utilise un graphe acyclique dirigé pour représenter les variables et leurs dépendances [50]. Le troisième classificateur de Bayes cité, A1DE, est moins utilisé dans le domaine biomédical, mais peut être intéressant à étudier, car il a été développé pour résoudre le problème d'indépendance des attributs de la méthode Naive Bayes [51].

Le modèle dit *paresseux* des k-plus proches voisins (kNN) classe un nouvel échantillon inconnu en fonction des classes de ses k plus proches voisins dans l'ensemble de données d'entraînement. La similarité entre les échantillons est mesurée à l'aide d'une métrique de distance, comme la distance euclidienne. La classe majoritaire parmi les k voisins les plus proches est attribuée au nouvel échantillon [52].

Enfin, un modèle SVM est un modèle linéaire qui tente de définir l'hyperplan qui sépare au mieux les classes dans l'espace des variables. Cet hyperplan est choisi de manière à maximiser la marge, c'est-à-dire la distance entre les points les plus

proches de chaque classe (appelés vecteurs de support) et l'hyperplan lui-même. Un modèle SVM peut utiliser une fonction de noyau pour traiter des données non linéairement séparables en les projetant dans un espace de dimension supérieure [53].

1.3.2 Classificateurs par réseaux de neurones

Un classificateur basé sur un réseau de neurones est inspiré par le fonctionnement du cerveau humain : il est composé de couches de neurones artificiels, où chaque neurone reçoit des entrées, effectue une opération mathématique, et transmet le résultat aux neurones de la couche suivante. Cette architecture permet au réseau d'extraire progressivement des caractéristiques de plus en plus abstraites des données, permettant ainsi une meilleure compréhension des informations complexes. En science biomédicale, ces algorithmes peuvent être employés pour la reconnaissance du langage, le diagnostic ou la prédiction d'une maladie par analyse de données textuelles ou d'images [54].

Malgré la visibilité grandissante des algorithmes par réseaux de neurones tant auprès du grand public qu'en recherche appliquée, les classificateurs dits traditionnels sont encore largement utilisés. Ils nécessitent souvent moins de ressources computationnelles, sont plus faciles à interpréter et ont besoin de moins d'échantillons. Leurs performances sont aussi souvent comparables aux modèles par réseaux de neurones sur des données tabulaires comme les données RNA-seq [55]. Dans une optique d'utilisation pour la sélection de variables, les algorithmes d'apprentissage profond sont aussi moins privilégiés. Le domaine de l'explicabilité, qui vise à extraire les variables d'intérêt utiles à la prédiction, est un domaine encore très récent et où il existe peu de consensus sur la méthodologie la plus adaptée [56].

1.4 Méthodes de sélection de variables

Pour identifier des biomarqueurs dans un jeu de données RNA-seq, les bioinformaticiens et bioinformaticiennes se sont intéressés aux approches de détection de motifs et, par extension, aux approches de sélection de variables. Les approches de sélection de variables permettent de sélectionner un sous-ensemble de variables capables de séparer de façon optimale deux classes d'échantillons ou plus [57]. Elles sont ainsi

généralement employées en amont ou en même temps qu'un modèle ML. Dans le présent projet, nous nous intéressons aux approches de sélection de variables qui sont ou peuvent être liées à un procédé ML. Les méthodes de sélection de variables associées à de la classification sont généralement réparties en trois types simples : filtres, méthodes d'enveloppement (connues sous le terme anglais *wrappers*), méthodes intégrées aux modèles (connues sous le terme anglais *embedded methods*) [58] (Figure 1.2). Les récents progrès en recherche en science de l'information permettent aujourd'hui de définir des méta-méthodes de sélection de variables et qui se basent sur ces trois premiers types : méthodes ensemblistes de sélection de variables et les méthodes hybrides ensemblistes de sélection de variables.

Les sections suivantes décrivent ces méthodes ainsi que leurs avantages et inconvénients.

1.4.1 Filtres, méthodes d'enveloppement et méthodes intégrées aux modèles

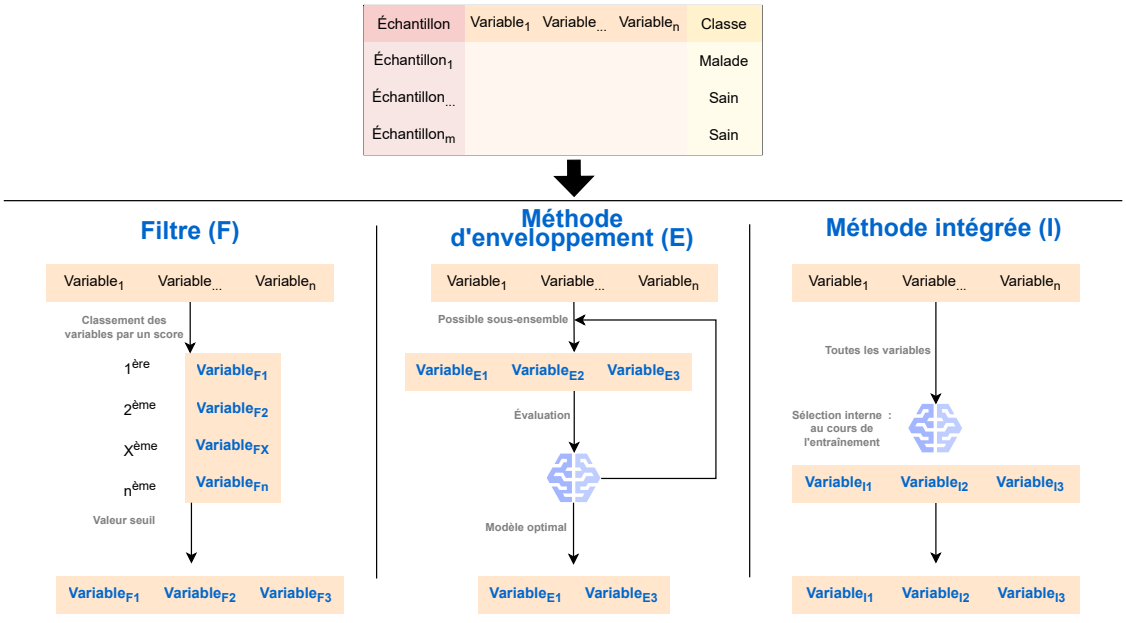

Figure 1.2: Les trois types principaux de sélection de variables : filtre, méthode d'enveloppement et méthode intégrée.

Filtres

Les filtres sont des méthodologies de sélection de variables qui évaluent un score de capacité de séparation sur les données sans tenir compte du modèle envisagé pour la classification subséquente. Les variables sont ensuite ordonnées en fonction des valeurs obtenues et les moins performantes sont mises de côté [59]. Parmi les méthodes de filtres, les plus utilisées sont celles basées sur le gain d'information, le score de Relief ou encore les métriques basées sur la corrélation [58–61]. En contexte biomédical et particulièrement pour l'analyse de données d'expression, nous pouvons rajouter à cette liste la mesure de variance [62] ou l'analyse d'expression différentielle [63–66] (section 1.1.2).

Le gain d'information est une mesure de la variation d'entropie d'une variable après avoir observé une autre variable [67, 68]. Le score de gain d'information entre deux variables X et Y est défini tel que :

$$\text{Gain d'information}(X, Y) = H(Y) - H(Y \mid X)$$

où $H(Y)$ est l'entropie de Y et $H(Y \mid X)$ est l'entropie conditionnelle de Y étant donné X.

Le gain d'information peut prendre comme valeur zéro au minimum et dénote une absence de variation d'entropie, donc aucun gain d'information. Au maximum, le gain d'information est égal à la valeur d'entropie de la variable Y.

Le score de Relief [69] quant à lui, se base sur une mesure de distance plutôt que sur l'entropie. En contexte de classification binaire, le score Relief évalue un poids pour chaque variable X. Dans un premier temps, pour une instance donnée C_i de classe i, une distance est évaluée pour identifier : un ensemble de m instances les plus proches ayant la même classe i que C_i (C_{hit}) et un ensemble de m instances les plus proches de C_i appartenant à une classe différente j (C_{miss}). À noter que m peut prendre la valeur un. Ensuite, pour chaque variable X, un poids $W(X)$ initialement de zéro est ajusté en fonction de leur relation avec les instances de même classe (C_{hit}) et de classe différente (C_{miss}). Ci-après la formule de mise à jour du poids $W(X)$:

$$W(X) = W(X) - \frac{\text{diff}(X, C_i, C_{\text{hit}})}{m} + \frac{\text{diff}(X, C_i, C_{\text{miss}})}{m}$$

où :

$$\text{diff}\left(X, C_1, C_{[\text{hit,miss}]}\right) = \frac{|\text{ valeur }\ (X, C_1) - \text{valeur}\left(X, C_{[\text{hit,miss}]}\right)|}{\max(X) - \min(X)}$$

Un filtre basé sur la variance peut aussi être employé [62]. Il s'agit alors de calculer la variance de chaque variable et d'utiliser la valeur obtenue comme score. Cette méthode n'utilise pas la classe des instances du jeu de données. Cette approche permet de ne retenir que les variables les plus variantes globalement et donc de retenir celles ayant un potentiel de séparation accru sans biais.

L'analyse DEG exposée précédemment (section 1.1.2) peut être considérée comme une méthode de sélection de variables par filtres. Elle est d'ailleurs souvent utilisée préalablement à une autre étape de sélection de variables ou de classification. Cette méthode est spécifique aux données d'expression génique comme les données RNA-seq.

Les méthodes par filtre sont souvent faciles à implémenter, peu coûteuses en temps et ressources computationnelles, mais peuvent donner des résultats bruités et donc moins spécifiques et performants que par d'autres méthodes de sélection de variables. Cela peut être dû à leur processus en une étape qui ne permet pas d'affiner la sélection et de la nécessité de déterminer des seuils sans réelle règle de détermination. Les filtres servent alors davantage d'étape de réduction des dimensions avant une seconde étape de sélection plus spécifique.

Méthodes d'enveloppement

Les méthodes d'enveloppement demandent davantage de temps et de ressources que les méthodes par filtre, mais tendent à obtenir de meilleurs résultats [60]. Les méthodes d'enveloppement exploitent un modèle ML pour évaluer la pertinence d'un sous-ensemble de variables jusqu'à trouver le meilleur [70]

Une méthode par enveloppement très répandue est l'élimination récursive de variables (*recursive feature elimination* RFE) parfois aussi assimilée à la méthode par élimination rétrograde (*backward elimination*) [71]. Celle-ci permet d'éliminer de façon itérative des variables et d'observer quelles variables ont le plus d'impact positif

ou négatif sur la qualité de la classification en calculant un score de performance. Cette méthode tend à mettre de côté des variables redondantes porteuses de la même capacité de séparation des classes que d'autres variables. Ce comportement est à considérer suivant la question de recherche posée.

À l'approche RFE s'oppose l'approche par sélection en marche avant (*forward selection*). Cette méthode consiste à commencer avec un modèle vide et à ajouter les variables une par une, en évaluant à chaque itération leur contribution à la performance du modèle [72]. À chaque étape, un score de performance est calculé, et seules les variables qui améliorent la qualité de la classification sont retenues. Contrairement à la RFE, qui élimine des variables, la sélection en marche avant se concentre sur l'ajout progressif de variables pertinentes.

Les avantages des approches de RFE et de marche avant sont exploités par un autre type de méthode qu'est la sélection par étape (*stepwise selection*). Cette approche permet d'éliminer et d'ajouter des variables dans les deux sens en respectant un système d'itération [73].

Enfin, une autre approche de méthode d'enveloppement nommé top-r vise à analyser des sous-ensembles de variables pour réduire l'espace de recherche [74]. Dans chaque sous-ensemble, les variables jugées non pertinentes sont éliminées, et ce processus se poursuit jusqu'à ce que toutes les variables soient classées par ordre d'importance. Les meilleures variables de chaque sous-ensemble sont ensuite comparées afin d'identifier le sous-ensemble optimal, garantissant une sélection plus efficace des variables pertinentes pour l'analyse.

Méthodes intégrées aux modèles

Le troisième type usuel d'approche de sélection de variables concerne les méthodes intégrées aux modèles. Ces méthodes réalisent la sélection de variables comme étape à part entière du processus d'entraînement des modèles ML et sont ainsi totalement dépendants du type de modèle employé [75]. Elles sont moins coûteuses en temps et ressources que les méthodes d'enveloppement et donnent d'aussi bons résultats.

Les arbres de décision utilisés pour de la sélection de variables sont considérés comme étant des méthodes de sélection de variables intégrées aux modèles [76]. En effet, lors de la construction d'un arbre de décision, l'algorithme évalue chaque

variable pour sélectionner laquelle permet de diviser les instances du jeu de données de manière à maximiser une mesure de pureté (comme l'entropie ou l'indice de Gini). Les modèles de réseaux de neurones peuvent aussi être considérés comme des méthodes de sélection de variables intégrées aux modèles [77]. En effet, leur entraînement nécessite l'ajustement des poids associés aux connexions entre les neurones. Ces poids déterminent l'importance relative des différentes variables d'entrée dans la prédiction de la sortie. Un poids trop faible peut entraîner une non utilisation d'un neurone et donc d'une variable.

1.4.2 Méthodes (hybrides) ensemblistes de sélection de variables

Il est souvent difficile de définir quelle méthode de sélection de variables choisir et c'est pourquoi de récentes études encouragent la communauté de recherche à développer des approches ensemblistes. Celles-ci permettent d'intégrer différentes méthodes de sélection de variables pour produire des signatures dites "robustes", au moins d'un point de vue méthodologique. À cette fin, ces méthodes ensemblistes tirent partie des principes de perturbation des données (principe homogène) ou de perturbation fonctionnelle (hétérogène) (Figure 1.3). La procédure de perturbation des données applique un algorithme de sélection de variable unique sur différents sous-ensembles d'un ensemble de données avant de procéder à une agrégation, tandis que le processus de perturbation fonctionnelle applique différentes méthodes de sélection de variables sur le même ensemble de données avant l'agrégation. L'étape finale d'agrégation permet d'intégrer les diverses variations de sélection pour obtenir un sous-ensemble stabilisé (ou signature) de variables [78].

Dans [79], les auteurs proposent une méthode ensembliste homogène liant une perturbation d'un jeu de données avec une sélection de variables par $L_1-normSVM$. Cette approche a été appliquée à l'identification de biomarqueurs de transcriptomique du cancer du rein via des données RNA-seq. Une autre méthode de sélection ensembliste développée récemment est EnRank [80]. EnRank réalise une approche hétérogène en appliquant divers filtres et méthodes d'enveloppement à un même jeu de données. Ces algorithmes de sélection retournent un classement des variables et les mieux classées sont soumises à des modèles ML dont la performance est éval-

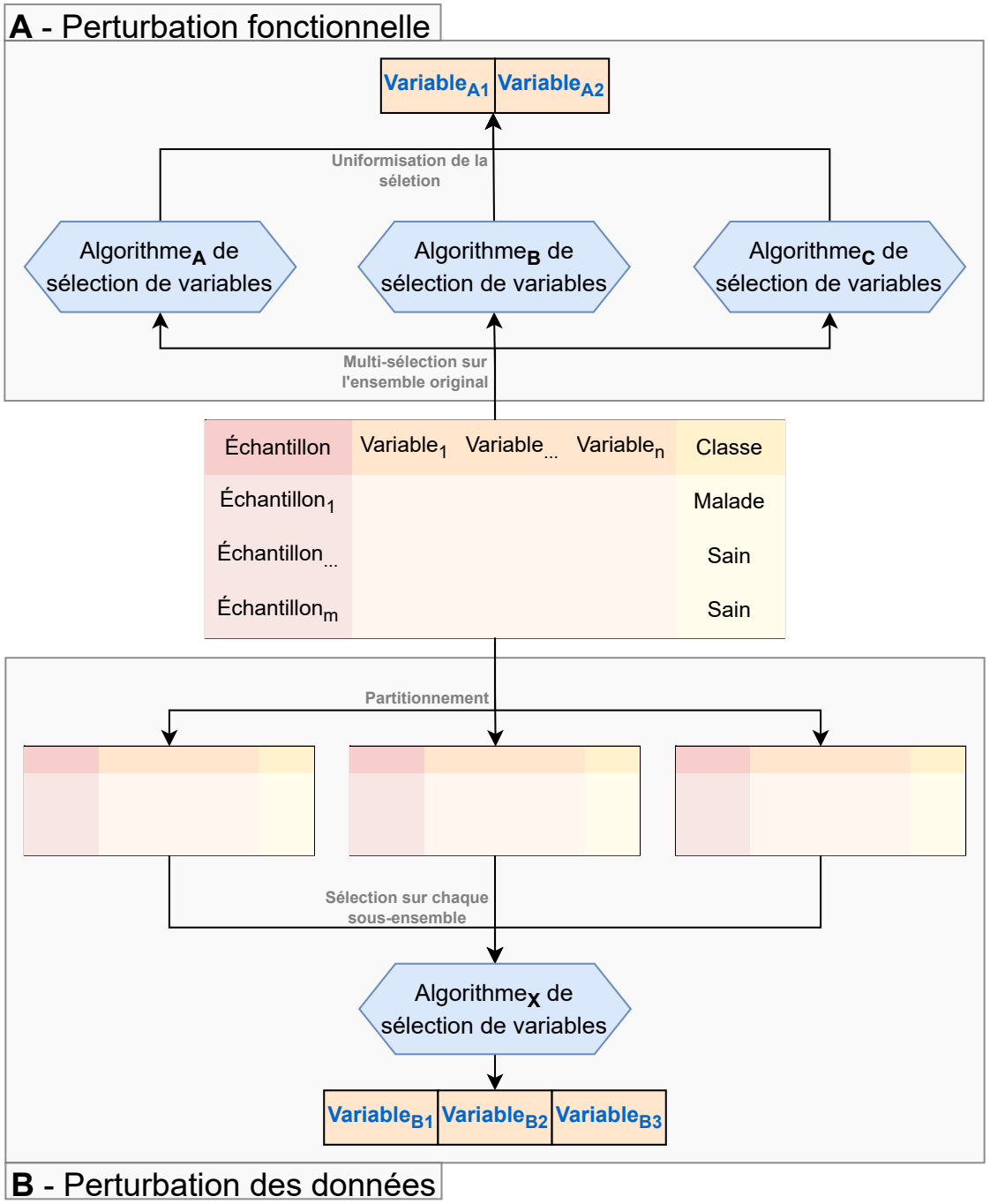

Figure 1.3: Deux catégories d'approches ensemblistes de sélection de variables. (A) Perturbation fonctionnelle par application puis concertation de divers algorithmes de sélection de variables à un même jeu de données. (B) Perturbation des données par application d'un unique algorithme de sélection de variables à divers sous-ensembles d'un jeu de données après obtenus après partitionnement.

uée. Si cette dernière est suffisante, le sous-ensemble de variables testé est conservé.

L'efficacité de EnRank a été évalué dans le cadre de l'identification de biomarqueurs

génomiques d'hypertension pulmonaire.

Dernièrement, les approches hybrides ensemblistes de sélection de variables (*Hybrid ensemble feature selection* HEFS) ont été développées pour combiner divers niveaux de variabilités : au niveau des données et de l'algorithme afin de bénéficier à la fois des notions d'homogénéité et d'hétérogénéité et d'accroître la stabilité du résultat [81, 82]. Le principe théorique de ces approches devrait permettre de produire des conclusions plus robustes, indépendantes d'un sous-ensemble de données d'entrée spécifique ou d'une méthode de sélection de variables. Néanmoins, la conception de solutions de type HEFS reste un défi, car elle nécessite de définir la technique d'échantillonnage, le nombre et les types de méthodes de sélection de variables à utiliser, leur réglage et le processus d'agrégation à appliquer.

1.4.3 Sélection de variables et AutoML

Comme évoqué précédemment, diverses étapes de la construction d'un modèle sont faites manuellement. Dans certains contextes comme des processus de sélection de variables qui font appel à de l'apprentissage automatique, il peut être utile d'utiliser des processus automatisés d'entraînement de modèles (*automated machine learning* autoML). L'autoML est un procédé qui vise à automatiser certaines étapes ML comme l'ingénierie des variables (englobe la création, transformation et la sélection de variables), l'optimisation des hyperparamètres ou plus globalement des pipelines entiers [83]. Cela permet d'accélérer le processus, mais aussi d'avoir une démarche systématique d'entraînement et envisager le développement de méthodes (hybrides) ensemblistes de sélection de variables. L'optimisation de pipelines entiers nous intéresse ici, car nos recherches s'inscrivent tout particulièrement dans le développement d'une approche hybride ensembliste de sélection de variables. Parmi les outils d'autoML ainsi produits par la recherche, nous pouvons citer les deux plus répandus qui sont Auto-WEKA [84] et Auto-Sklearn [85] mais aussi TPOT [86], ATM [87], Automatic Frankensteining [88], ML-Plan [89], Autostacker [90], h2o [91] ou encore BioDiscML [92]. Devant cette liste non exhaustive d'outils, le choix des méthodes est une question importante. Pour faciliter cette étape, la section suivante s'intéresse à caractériser les différents outils selon deux critère principaux (Tableau 1.3) :

- intégration d'une étape de sélection de variables dans l'outil : la présence

de base de cette fonctionnalité peut suggérer la possibilité de faire une présélection de variables et/ou de la moduler pour convenir à l'approche (hybride) ensembliste de sélection de variables qui se basera sur l'outil ;

- extraction de tous les modèles entraînés.

Ces outils d'autoML, ont été originellement développés pour entraîner divers modèles ML. En revanche, BioDiscML est le seul à ne pas avoir eu pour but de réaliser cet entraînement en privilégiant l'optimisation d'hyperparamètres. On constate qu'Auto-Sklearn et BioDiscML sont les seuls capables d'intégrer à la fois une étape de sélection de variables et de permettre l'extraction de tous les modèles entraînés. La philosophie ensembliste de sélection de variables ne nécessite pas forcément l'utilisation que de "meilleurs modèles" mais de modèles suffisamment efficaces pour réaliser une sélection robuste et pertinente. Ainsi, pour développer une approche (hybride) ensembliste de sélection de variables, il est avantageux d'avoir un accès facilité aux détails de tous les modèles entraînés et non seulement le ou les meilleurs. Cette fonctionnalité est peu présente dans les outils d'autoML ou elle est difficilement accessible et les détails obtenus sont souvent parcimonieux. BioDiscML est un des rares outils, avec Auto-Sklearn à offrir cette possibilité. Ajoutons à cela que BioDiscML est un framework modulable qui est développé par le Dr. Mickaël Leclercq, professionnel de recherche au laboratoire d'Arnaud Droit au CRCHUL de Québec. Le Dr. Leclercq fait partie de l'équipe de recherche ayant travaillé sur la présente thèse et cette étroite collaboration a permis l'implémentation facilitée de nouvelles fonctionnalités personnalisées pour le projet. Nous avons donc privilégié BioDiscML comme outil d'autoML dans notre projet de recherche.

Méthode	Sélection de variables	Extraction de tous les modèles entraînés
Auto-WEKA	✓	
Auto-Sklearn	✓	✓
TPOT	✓	
ATM		✓
Automatic Frankensteining		
ML-Plan		✓
Autostacker		
h2o		✓
BioDiscML	✓	✓

Tableau 1.3: Outils d'automatisation d'entraînement de modèles d'apprentissage automatique. Caractérisation par deux critères : intégration d'une étape de sélection de variables et fonctionnalité d'extraction de tous les modèles entraînés.

1.5 Perturbations de données par ré-échantillonnage

Une approche hybride ensembliste de sélection de variables employant notamment des algorithmes ML intègre une étape de perturbation des données. Cette dernière implique de générer, depuis un ensemble des données initial, divers sous-ensembles de données d'entraînement [78]. Cela peut être réalisé par des méthodes conventionnelles de partitionnement des données comme l'échantillonnage aléatoire [93] ou l'échantillonnage stratifié [94].

Ces méthodes permettent de séparer un ensemble de données d'origine en un ensemble de données d'entraînement et de test avec un certain ratio qui est traditionnellement de 2/3 pour l'entraînement et 1/3 pour le test. La méthode classique d'échantillonnage aléatoire permet de prendre aléatoirement 2/3 des échantillons et de les affecter à l'étape d'entrainement et 1/3 pour le test. Cette méthode n'est pas adaptée aux données à classes déséquilibrées, car elle peut négliger la représentation d'une classe de faible taille dans un des échantillonnage.

La méthode d'échantillonnage stratifié l'est davantage, car elle permet de sélectionner aléatoirement 2/3 de chaque classe de l'ensemble de données d'origine pour être affectés à l'ensemble de données d'entraînement. Le même principe s'applique pour créer le sous-ensemble de test avec un ratio de 1/3.

Cette stratégie peut se décliner aussi en échantillonnage stratifié équilibré distribué [95]. Dans un contexte de possible variation intra-classe, cette méthode per-

met de définir des groupes intra-classe et de conserver leur répartition et proportions lors de l'échantillonnage en ensemble d'entraînement et de test. La définition de ces groupes peut se faire par une méthode de partitionnement s'appuyant sur une distance de similarité comme la distance euclidienne.

Pour réaliser la perturbation de données, il est alors possible, depuis un même jeu de données initial, de répéter plusieurs fois en parallèle une de ces stratégies d'échantillonnage et de procéder ainsi à un ré-échantillonnage répété. Les différentes paires de jeux d'entraînement et de test peuvent ainsi être soumises à divers algorithmes de sélection de variables par ML.

1.6 Évaluation

1.6.1 Évaluation de modèles

Un modèle entraîné doit être évalué pour estimer sa capacité à classifier correctement de nouveaux échantillons et indirectement si la signature de variables employée peut être efficace. Pour cela, nous pouvons nous appuyer sur des métriques de qualité telles que l'*accuracy*, la précision, le rappel [96], le score F-1 [97] ou plus récemment, le coefficient de corrélation de Matthews (MCC) [98].

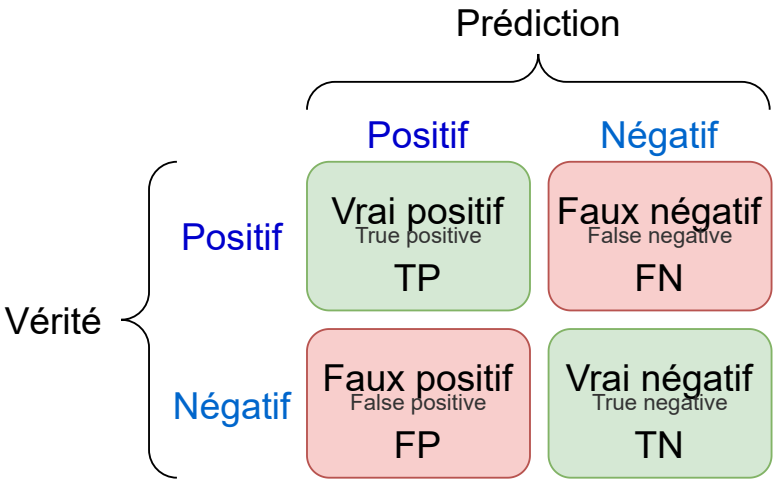

Figure 1.4: Matrice de confusion répertoriant les échantillons correctement ou incorrectement classés en classification binaire.

Ces trois métriques s'appuient sur la matrice de confusion associée à un modèle (Figure 1.4) après que celui-ci a été entraîné puis testé sur le jeu de données de

test. Dans le cas d'une classification binaire, cette dernière répertorie le nombre d'échantillons d'une classe positive (souvent une classe de phénotype anormal en bioinformatique) ayant été bien (vrai positif ou *true positive* TP) ou mal (faux négatif ou *false negative* FN) classés ainsi que les échantillons de la classe négative (souvent une classe de contrôle) bien (vrai négatif ou *true negative* TN) et mal (faux positif ou *false positive* FP) classés.

L'*accuracy* (parfois appelé "précision" en français même si ce terme est déjà consacré pour traduire la métrique de *precision*) permet de calculer la proportion de classification correcte parmi toutes les prédictions faites par le modèle. Elle est définie comme suit :

$$\text{accuracy } = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{FN} + \text{TN}}$$

L'*accuracy* est néanmoins biaisée en cas de proportions de classes déséquilibrées. Par exemple, une classe minoritaire (positive ou négative) entièrement mal classée n'aura que peu d'impact sur la valeur de l'*accuracy* qui restera malgré tout très haute si une grande proportion de la classe majoritaire est bien classée.

Le score F1, qui est la moyenne harmonique de la précision et du rappel, quant à lui est davantage adapté au déséquilibre de classes et est définie comme suit :

$$\text{score F1} = \frac{2 \cdot \text{TP}}{2 \cdot \text{TP} + \text{FP} + \text{FN}} = \frac{2 \times \text{ Precision } \times \text{ Recall}}{\text{Precision } + \text{ Recall}}$$

Le score F1 ne prend pas en compte les vrais négatifs, ce qui implique qu'il évalue mal la capacité du modèle à classifier les négatifs et n'est donc pas une métrique représentative du comportement global du modèle.

Le MCC quant à lui est une métrique adaptée au déséquilibre de classes et qui rend compte du comportement global du modèle vis à vis des deux classes en prenant en compte toutes les valeurs de la matrice de confusion. Le MCC est défini comme suit :

$$\text{MCC} = \frac{\text{TP} \cdot \text{TN} - \text{FP} \cdot \text{FN}}{\sqrt{(\text{TP} + \text{FP}) \cdot (\text{TP} + \text{FN}) \cdot (\text{TN} + \text{FP}) \cdot (\text{TN} + \text{FN})}}$$

Le MCC a récemment gagné un intérêt certain auprès de la communauté de recherche ML, car il comble les lacunes précédemment citées de la très répandue *accuracy* et

apporte un autre regard sur la qualité de la classification que le score F1 [99].

Ces métriques peuvent également être calculées durant l'entraînement afin d'évaluer la qualité de l'ajustement des paramètres du modèle dans un schéma itératif d'entraînement.

1.6.2 Évaluation qualitative de signatures

Évaluation de la stabilité

Dans un contexte de sélection de variables par approche ensembliste, il est intéressant de mesurer la stabilité des signatures obtenues. Il s'agit alors d'estimer à quel point les variables sélectionnées dans diverses signatures sont préservées par les différentes méthodes de sélection de variables utilisées ou depuis les différents ensembles de données d'entraînement [100]. Pour évaluer ceci, divers scores ont été développés et cinq critères théoriques ont été définis pour estimer leur capacité de généralisation [101] :

- Entièrement défini : le score ne doit pas dépendre de la taille des ensembles de variables analysés ;
- Monotonicité stricte : le score est strictement croissant à mesure que les ensembles de variables ont d'éléments en communs ;
- Limites : le score doit être limité par des constantes supérieures et inférieures ne dépendant pas du nombre total de variables ou du nombre de variables sélectionnées ;
- Maximum : le score doit atteindre son maximum si et seulement si les ensembles de variables comparés sont strictement identiques ;
- Correction pour le hasard : le score attendu pour une sélection aléatoire de variables avec des probabilités de sélection égales doit être constante et ne doit donc pas dépendre du nombre de variables sélectionnées.

Quatre scores sont identifiés comme répondant favorablement à ces cinq critères : Bommert and Rahnenführer (2020) [102], Carletta (1996) [103], Nogueira and Brown (2016) [104] et Nogueira (2018) [101]. Le score de Nogueira, parmi les plus employés, correspond à une mesure de stabilité basée sur la fréquence. Elle considère comme

stables les situations dans lesquelles toutes les variables sont choisies pour chaque ensemble ou pour aucun d'eux. Si une variable est sélectionnée uniquement pour certains ensembles, cela réduit le score de stabilité. Le score de Nogueira a une borne supérieure de 1 et une borne inférieure de $-1/(p-1)$ où p est le nombre d'ensembles de variables à comparer. Un score de Nogueira proche de 1 indique une haute stabilité des ensembles de variables comparés tandis qu'un score proche de zéro suggère une faible stabilité.

Caractérisation biologique

Pour limiter davantage les ressources allouées à l'étude de biomarqueurs, il convient d'étudier l'intérêt biologique, voire clinique, de biomarqueurs de maladies. Pour cela, il peut être intéressant d'employer des méthodes d'analyse d'enrichissement d'annotations biologiques telles que l'analyse de sur-représentation (*over-representation analysis* ORA) ou l'analyse d'enrichissement d'ensemble de gènes (*gene sets enrichment analysis* GSEA).

L'ORA consiste à confronter une liste de noms de gènes à tester à une liste de gènes de référence. Cette comparaison vise à identifier les listes (ou termes) de référence qui seraient significativement sur-représentés dans la liste d'intérêt en estimant la probabilité que ce ne soit pas dû au hasard [105].

Quant à la GSEA, elles tirent parti des données d'expression différentielles fournies par l'utilisateur pour identifier des termes de référence significativement représentés. Pour cela, les gènes sont ordonnés en fonction de leur score d'expression différentielle où chaque extrémité concerne les gènes les plus sur- ou sous-exprimés. La fréquence d'enrichissement des termes aux extrémités de la liste ordonnée est finalement évaluée.

Dans ce manuscrit, nous nous concentrerons davantage sur l'utilisation des techniques d'ORA.

Les termes de référence sont centralisés dans diverses bases de données spécialisées de références d'annotations (voir section 1.2.1). Différents outils programmatiques et en ligne sont en capacité de faire des analyses globales d'ORA en interrogeant plusieurs bases de données à la fois. Parmi ces outils, nous pouvons citer les outils en ligne ToppGene, enrichR, DAVID, GSAN ou encore g:profileR. Ces outils

sont très semblables et il est souvent difficile de choisir [106, 107]. Cependant, certains critères peuvent aider à la décision comme le nombre et le type de bases de données reliées (onthologie de gènes, voies et réseaux biologiques, maladies etc.) ou l'accès facilité à une API et/ou à une solution programmatique. L'outil g:profileR, par exemple, a une bibliothèque R associée et la possibilité de générer un lien stable d'analyse ORA pré-configurée. Les solutions de visualisations proposées par ces différents outils peuvent aussi être des critères de sélection suivant les préférences et besoins de l'utilisateur.

1.7 Méthodologie pour la conception d'un système de visualisation

La visualisation de données est parfois perçue comme optionnelle, à visée ostentatoire et facile à réaliser, y compris dans la communauté de recherche [108]. Or la visualisation de données est un domaine à part entière qui peut permettre à la recherche biomédicale d'exploiter avec des solutions interactives la multitude de données disponibles. Les biologistes, bioinformaticiennes et bioinformaticiens mais aussi les cliniciennes et cliniciens sont régulièrement appelés à se former sur les questions de visualisation en prenant en compte divers critères [109]. Ces réflexions interdisciplinaires sont primordiales pour permettre une meilleure représentation des données et donc une meilleure compréhension des phénotypes qu'elles décrivent. Il est donc primordial d'appliquer une méthodologie de conceptualisation pour proposer des métaphores visuelles répondant à des questions biologiques précises. Pour cela, la communauté de visualisation de l'information s'appuie sur le mantra de la recherche d'informations visuelles proposé par Shneiderman en 1996 [110] : ***Overview First, Zoom and Filter, Details on Demand***. Celui-ci préconise de conceptualiser des solutions visuelles pour une vue d'ensemble des données et des interactions pour offrir la possibilité de zoomer et filtrer et enfin, la possibilité pour l'utilisatrice ou utilisateur d'obtenir des détails à la demande sur les données.

Pour encadrer cette conceptualisation de divers niveaux d'études en positionnant l'utilisateur au centre de l'analyse, nous pouvons nous appuyer sur le concept de Munzner [111]. Cette méthodologie vise à décomposer un problème d'analyse

en plusieurs strates consécutives : la question scientifique spécifique du domaine, l'abstraction des données et des tâches, la métaphore visuelle proposée et la conception ou utilisation d'un algorithme pour répondre à la question. Une version simplifiée de ce concept peut être présentée sous la forme d'un tableau tel que le tableau 1.4.

Question biologique	**Abstraction des tâches et des données**	**Métaphore visuelle**

Tableau 1.4: Exemple d'en-tête de tableau pour la mise en place d'un suivi de développement d'une solution visuelle par le concept de Munzner.

1.8 Éléments de visualisation

1.8.1 Attributs visuels

Les différents systèmes de visualisation s'appuient sur des attributs visuels que l'on peut décomposer et dont le choix doit se faire en fonction du type de variables à représenter (Figure 1.5).

Ces recommandations nous invitent à considérer des attributs visuels très différents pour des données quantitatives ou ordinales et nominales. Ces deux dernières classes présentent en effet un classement des attributs similaires et qui est pratiquement l'inverse de celui établi pour les données quantitatives. À titre d'exemple, la visualisation de données d'expression, qui sont des données quantitatives, doit être conceptualisée en portant une attention particulière à des attributs comme la position, la longueur, l'angle ou encore la courbe des données.

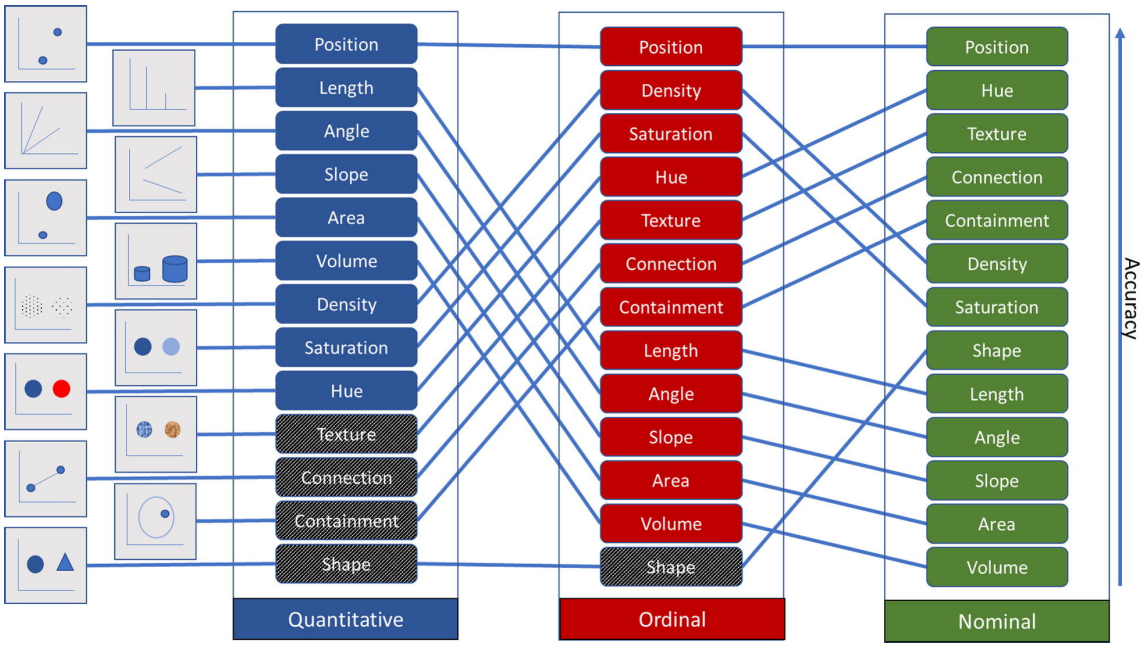

Figure 1.5: Classification de l'efficacité de treize attributs visuels en fonction du type de données à visualiser. Les attributs en gris sont jugés non pertinent pour la classe de données considérée. Figure extraite de [112] et adaptée de l'étude de Mackinlay [113].

1.8.2 Visualiser des données relationnelles

En recherche biomédicale, nous sommes régulièrement amenés à observer des relations entre des entités diverses comme les paires maladie/biomarqueur, gène/annotation ou encore maladie/médicament.

Une métaphore visuelle adéquate pour ce genre de données relationnelles est un diagramme basé sur une matrice [112]. Dans ce cas-là, une entité est représentée en colonne et en ligne et une glyphe représentant la relation entre deux entités est présente dans la cellule associée. Dans ce genre de visualisation, l'ordre des entités est important.

Une autre façon de représenter ces données relationnelles peut être par un diagramme noeuds/liens [112]. Dans ce genre de diagramme, une entité est représentée par une glyphe comme un cercle et une relation entre deux entités est représentée par une ligne. Les diagrammes noeuds/liens, parfois appelés "graphes" par simplification, peuvent être composés d'un grand nombre d'entités et/ou relations. Dans une telle visualisation, la position des entités et des relations est ainsi primordiale. Pour traiter de larges données relationnelles, y compris par graphe, il peut être d'intérêt de s'intéresser au concept de visualisation multi-échelle [112]. Cette démarche réalise

une abstraction des données en regroupant en *méta-entités* des entités partageant des critères communs. En résulte, une diminution du nombre d'entités à visualiser et donc une visualisation plus lisible.

Chapitre 2

Développement d'une approche hybride ensembliste de sélection de variable (HEFS)

2.1 Introduction

Ce deuxième chapitre décrit l'approche proposée pour répondre au premier objectif de la thèse. Il s'agit de permettre l'identification de biomarqueurs robustes de maladies complexes comme le cancer à partir de données transcriptomiques. Pour cela, nous avons développé une approche hybride ensembliste de sélection de variable (HEFS). Comme mentionné précédemment, ces approches sont prometteuses, mais encore récentes et sujettes à de nombreux choix lors de leur conception. Nous décrivons ainsi la méthodologie employée pour explorer cette problématique. Ainsi, ce chapitre présente dans un premier temps les deux méthodes de filtrage employées dans l'étape préalable de réduction des dimensions. Dans un deuxième temps, l'intégration de deux stratégies d'échantillonnage est décrite. Cette étape d'échantillonnage s'inscrit dans un des principes fondateurs des approches HEFS qui visent à intégrer une perturbation des données. Par la suite, ce chapitre décrit le processus de sélection de variable par algorithmes d'apprentissage automatique d'enveloppement. Ce processus-ci vise à permettre l'intégration de perturbations fonctionnelles. Nous présentons de plus dans ce chapitre la méthode d'agrégation qui a pour but d'intégrer les perturbations des données et fonctionnelles pour obtenir

des signatures robustes de biomarqueurs. Enfin, nous présentons deux modes de notre approche HEFS que nous appellerons dans la suite du manuscrit, le mode standard et le mode *light* HEFS. Le premier définit une exécution de l'approche dans un cadre avantageux de ressources computationnelles, tandis que le second permet une exécution dans un cadre plus restreint.

2.2 Vue d'ensemble de la stratégie

Pour étudier l'impact des étapes de filtrage, d'échantillonnage et d'agrégation dans le cadre de l'identification de biomarqueurs à partir de données transcriptomiques, quatre scénarios HEFS en mode standard ont été mis en place (figure 2.1). Pour chaque scénario :

- une première étape de réduction de dimension des variables (ici, des gènes) est réalisée soit par une analyse des gènes différentiellement exprimés (DEG), soit par une analyse de variance (Var) (figure 2.1, bloc de filtrage en jaune et section 2.3) ;
- deuxièmement, une étape de perturbation des données intervient par le biais, soit d'une méthode d'échantillonnage aléatoire stratifié (R-S), soit d'une méthode d'échantillonnage stratifié équilibré selon la distribution (DB-S) (figure 2.1, bloc d'échantillonnage en rouge et section 2.4) ;
- troisièmement, les quatre scénarios ont été soumis à une étape d'entraînement à grande échelle qui entraîne des milliers de modèles d'enveloppement à partir de multiples types de classificateurs largement utilisés ayant différentes configurations d'hyper-paramètres (figure 2.1, bloc d'entraînement en bleu et section 2.5) ;
- enfin, pour tirer profit de la perturbation fonctionnelle et des données, l'approche HEFS se conclut par une étape d'agrégation (figure 2.1, bloc d'agrégation en vert et section 2.6). Cette dernière est régie par des règles génériques qui visent à répondre à plusieurs besoins : évaluer les effets à la fois des étapes de filtrage préliminaires et des deux types d'échantillonnage, ainsi que de tirer parti de

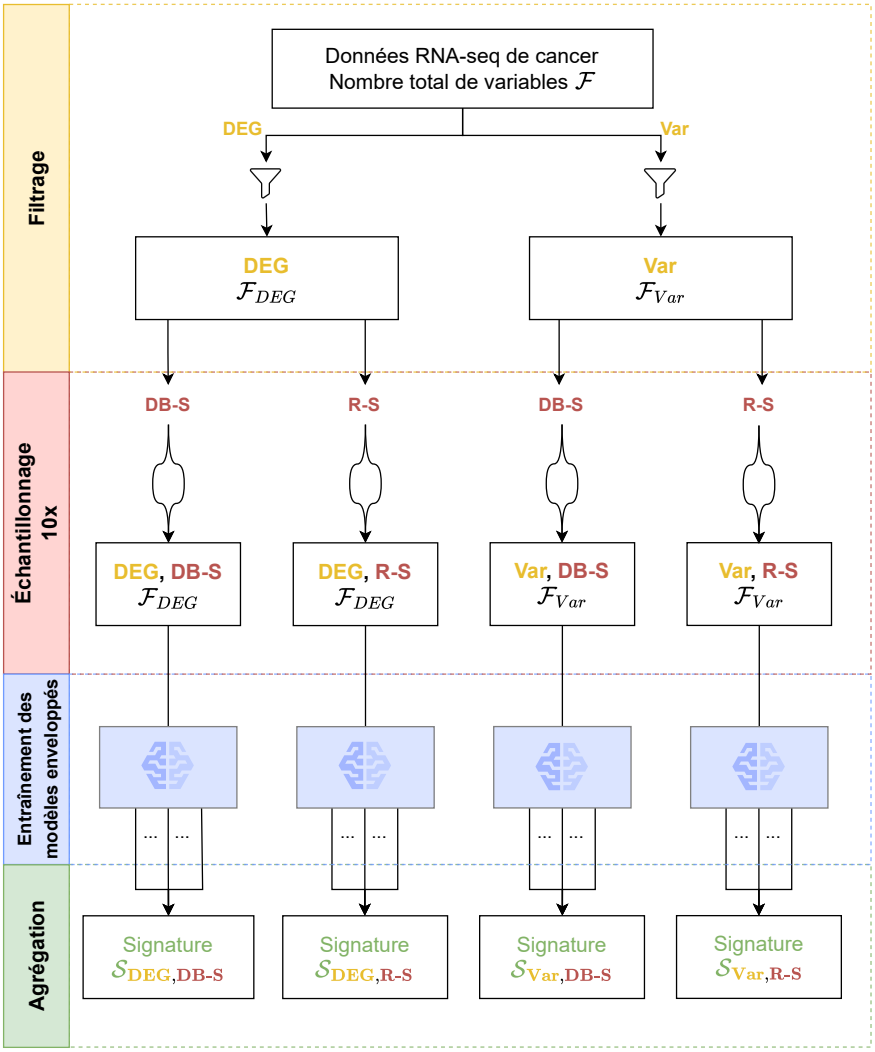

Figure 2.1: Vue générale de l'approche proposée de sélection de variable hybride ensembliste. Processus en quatre étapes : filtrage, échantillonnage, entraînement des modèles d'enveloppement, agrégation. Le filtrage appliqué aux données est basé soit sur l'analyse des gènes différentiellement exprimés (DEG), soit sur l'analyse de la variance (Var). L'étape d'échantillonnage est effectuée par séparation répété pour calculer 10 paires différentes d'ensembles d'entraînement et de test pour l'étape de sélection de variable par enveloppement. Cette séparation est basée soit sur stratégie d'échantillonnage stratifié aléatoire (R-S), soit sur un échantillonnage stratifié équilibré par distribution (DB-S). L'approche se poursuit par une étape à large échelle d'entraînement de modèles d'enveloppement et se conclut par une étape d'agrégation.

la diversité et de la complexité des diverses méthodes de ML en testant de multiples combinaisons d'hyper-paramètres.

L'approche HEFS présentée dans ces travaux a été analysée dans un premier temps dans son mode standard. Puis, sa capacité de généralisation a été éprouvée par l'application d'un mode *light*. Ce mode est plus rapide à l'exécution pour

certains jeux de données et permet une stabilisation plus relaxée des perturbations fonctionnelles et des données jusqu'à la formulation des signatures finales.

2.3 Filtrage

Les jeux de données RNA-seq utilisés avant l'étape de filtrage sont des données tabulaires où chaque ligne est un échantillon et chaque colonne est un gène codant ou non et où le nombre de *reads* obtenues après séquençage est inscrit dans chaque cellule (tableau 2.1). Les gènes sont identifiés par leur identifiant Ensemble.

Échantillon	ENSG00000000003	ENSG00000000005
TCGA-5M-AAT4-01A-11R-A41B-07	9897	89
TCGA-5M-AAT6-01A-11R-A41B-07	1213	4
TCGA-A6-2671-01A-01R-1410-07	2600	11
TCGA-A6-2674-01A-02R-A278-07	4813	5

Tableau 2.1: Exemple de format des données transcriptomiques employées dans notre projet : chaque ligne est un échantillon, chaque colonne présente le nombre de *reads* obtenues pour chaque gène codant ou non.

À partir des jeux de données d'entrée et peu importe le scénario, les gènes codant pour des protéines et les ARNs non codants ayant moins de 10 *reads* alignées à travers tous les échantillons ont été éliminés pour s'affranchir d'une partie du bruit dans les données. Par la suite, un type de filtre parmi deux possibilités a été appliqué : l'un basé sur une analyse des gènes différentiellement exprimés (DEG) et l'autre basé sur une analyse de la variance (Var).

L'analyse DEG, effectuée à l'aide du package R DESeq2, prend en compte les effets de lot possibles et les étiquettes associées à chaque échantillon dans les données (Normal ou Stade IV). Les données ont ensuite été transformées avec la méthode classique de transformation par stabilisation de la variance. Enfin, nous avons conservé les gènes codant pour des protéines et les ARNs non codants avec un changement d'expression d'au moins 2 et une valeur p ajustée inférieure à 0,001 ou 0,05 pour le mode standard ou le mode *light*, respectivement.

Le filtre de variance a été réalisé en utilisant le package edgeR pour normaliser les données en logarithme du nombre de comptes par millions, suivi d'une élimination de l'effet de lot à l'aide du package R limma. Enfin, la fonction adjust_matrix du

package R cola a permis de sélectionner les 10% des variables qui varient le plus à travers tous les échantillons sans prendre en compte les étiquettes de conditions. Les valeurs seuils ci-dessus ont été choisies afin de réduire le nombre de variables et le bruit pour le processus d'enveloppement ultérieur, qui est exigeant en termes de ressources informatiques.

Après avoir appliqué ces filtres sur le jeu de données initial, deux ensembles de données filtrés et étiquetés respectivement par les préfixes DEG et Var sont obtenus.

2.4 Stratégies d'échantillonnages

Deux types d'échantillonnage produisant des paires de données d'entraînement et de test ont été appliqués indépendamment l'un de l'autre. Les deux effectuent des partitionnements de données répétés, car ils divisent l'entrée en n (ici $n = 10$), différentes paires d'ensembles d'entraînement (2/3) et d'ensembles de test (1/3).

La première version de l'échantillonnage est une variante d'un échantillonnage stratifié (R-S) comme introduit dans [114] qui préserve la distribution initiale entre les classes (voir aussi section 1.5).

Une analyse préliminaire de clustering montrait qu'après une analyse DEG et malgré une normalisation des données, deux groupes minimum semblaient se distinguer dans les échantillons de stade IV (Annexe B, figure B.3). Ce même phénomène se remarque dans les échantillons normaux qui sont des échantillons d'une partie dite saine du tissu de patients malades (Annexe B, figure B.4). Cela nous a encouragé à considérer une deuxième stratégie d'échantillonnage prenant en compte une hiérarchie intra-classe des échantillons. Cette dernière est une variante de l'échantillonnage stratifié équilibré distribué comme présentée dans [95, 115] (section 2.4) et est désignée sous le nom de DB-S dans la suite de ce manuscrit. Elle prend en compte la variabilité intra-classe en définissant des groupes intra-classe, ici, dans le but de préserver l'hétérogénéité inter-patients. Les groupes ont été calculés en appliquant un regroupement hiérarchique utilisant la distance euclidienne et la méthode de Ward pour la fusion. Une technique de Silhouette a ensuite été appliquée pour définir automatiquement le nombre approprié de groupes [116]. Cette technique évalue la qualité d'un regroupement en mesurant la distance des échantillons par

rapport à leur propre cluster et leur séparation par rapport aux autres clusters. Elle attribue un score à chaque échantillon, permettant de juger de la pertinence de son assignation. Cela nous a permis d'associer les identifiants d'échantillons à ces groupes nouvellement définis. Un échantillonnage stratifié à deux niveaux a par la suite été appliquée pour garantir que les distributions inter- et intra-classes sont respectées lors de la définition des 10 paires d'ensembles d'entraînement et de test.

Dans le mode *light* HEFS, seul l'échantillonnage par DB-S est considéré.

2.5 Entraînement de modèles d'enveloppement

Pour exploiter la diversité des différents algorithmes d'apprentissage automatique, comme encouragé dans les approches basées sur la perturbation fonctionnelle, nous avons sélectionné huit méthodes d'apprentissage automatique supervisées souvent utilisées dans les études biomédicales pour résoudre les problèmes de classification. Ces classificateurs peuvent être grossièrement organisés en quatre catégories : bayésienne, arbres de décisions, apprentissage *paresseux* et basés sur une fonction (section 1.3.1).

Parmi les classificateurs bayésiens probabilistes, les algorithmes Naive Bayes, Réseau bayésien et Estimateurs 1-Dependence moyennés (A1DE) ont été sélectionnés. En ce qui concerne les classificateurs basés sur les arbres de décision, qui sont des algorithmes de sélection de variables intégrés, nous avons sélectionné C4.5 et Random Forest qui sont couramment utilisés dans un contexte clinique. Nous avons également utilisé Simple Classification And Regression Trees (Simple CART) dont la logique est similaire à la méthode C4.5. Le classificateur *paresseux* du plus proche voisin (kNN) a aussi été considéré. Enfin, nous avons sélectionné les machines à vecteurs de supports (SVM) parmi les classificateurs basés sur une fonction.

Le mode standard HEFS inclut ces huit types d'algorithmes alors que le mode *light* en inclut 5 (Naive Bayes, A1DE, Random Forest, C4.5 et SVM). Cela permet une accélération du temps d'entraînement en éliminant des algorithmes très coûteux en termes de calcul (quantité de ressource et temps alloué) tels que les réseaux bayésiens, les Simple CART ou certains noyaux de SVM. Les kNN ont été mis de côté, car leur concept d'apprentissage paresseux rend leur mise à l'échelle difficile et

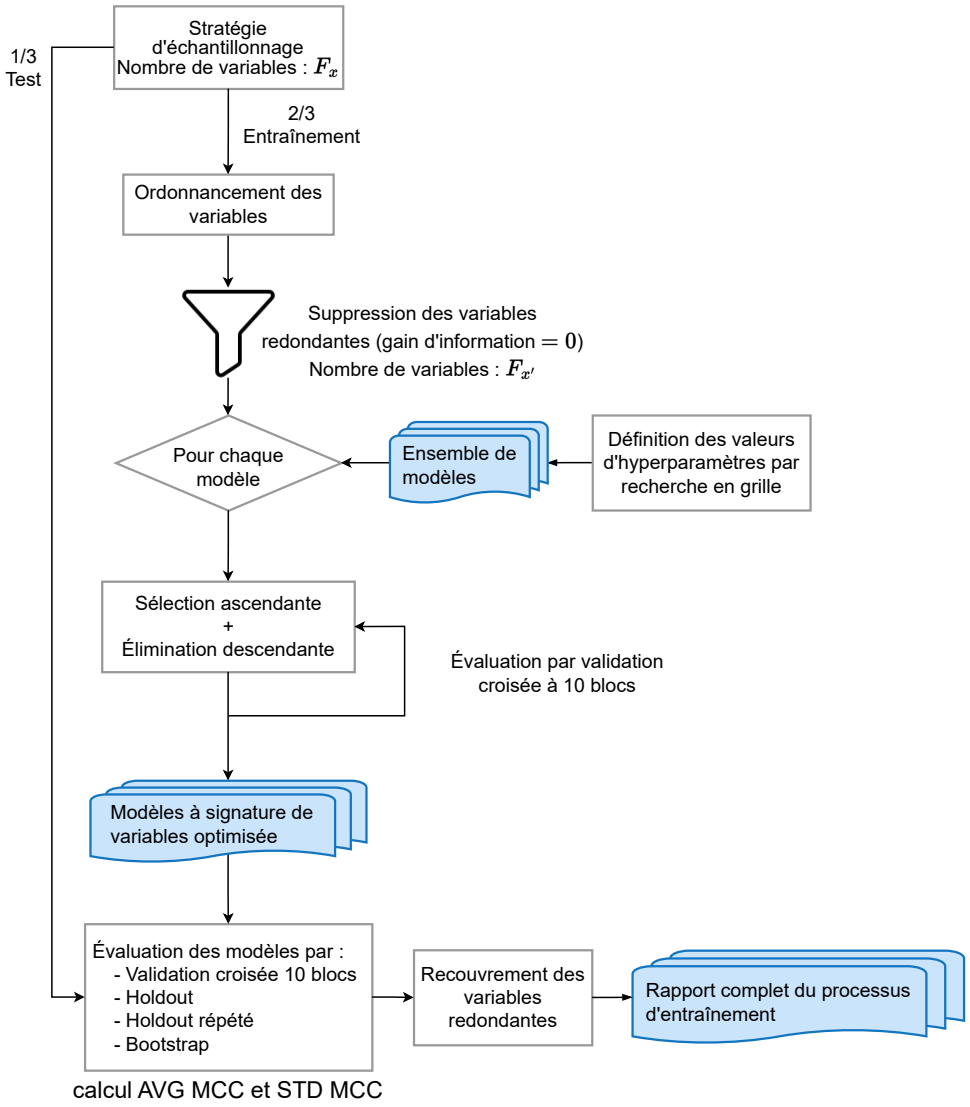

Figure 2.2: Processus d'entraînement en apprentissage automatique basé sur le logiciel BioDiscML. Les données d'entrée pour une exécution d'échantillonnage sont composées de patients décrits par F_x variables. Ces données sont divisées en un ensemble de test (1/3)) et un ensemble d'entraînement (2/3). L'ensemble d'entraînement subit un processus de classement des variables basé sur la mesure du gain d'information, où les variables ayant un score de zéro sont éliminées. Les variables résultantes $F_{x'}$ sont ensuite soumises à une boucle d'entraînement par sélection pas à pas ascendante combinée à une élimination pas à pas descendante. Chaque modèle entraîné résultant passe par diverses étapes de validation croisée et pour chacun d'eux, le score de MCC moyen (AVG MCC) et l'écart type associé (STD MCC) sont calculés. Les variables redondantes, écartées pendant le processus d'entraînement, sont recaptées pour une exploitation ultérieure.

sont sensibles aux données bruitées, ce qui est souvent le cas en contexte de recherche biomédicale.

Le processus d'enveloppement (Figure 2.2) a été réalisé en utilisant le logiciel BioDiscML, qui permet l'exécution d'un pipeline facile d'utilisation, capable d'effectuer un enveloppement d'apprentissage automatique à large échelle tout en conservant l'ensemble des résultats des différentes expérimentations. Les phases d'entraînement des modèles d'enveloppement étaient restreintes en temps et disposaient de ressources informatiques limitées pour chaque exécution d'échantillonnage. Des milliers de modèles ont été entraînés avec plusieurs configurations de leurs nombreux hyper-paramètres. Ces configurations ont été générées à l'aide d'une approche de recherche en grille (Annexe A).

Plusieurs étapes constituent l'entraînement après l'étape d'échantillonnage de l'approche HEFS (Figure 2.2) :

1. **Ordonnancement des variables.** Chaque jeu de données d'entraînement passe par une première étape d'ordonnancement des variables basée sur le score de gain d'information et à la fin de laquelle les variables ayant un score de gain d'information de zéro sont rejetées. Un score de zéro indique une indépendance absolue entre les deux variables testées (ici l'étiquette contre les valeurs d'expression prises par un gène considéré).

2. **Boucle itérative d'entraînement.** Pour chaque modèle, les ensembles de données d'entraînement sont soumis à une boucle d'entraînement de sélection par étape en marche avant et une élimination récursive RFE (voir section 2.3 et [92] pour plus de détails sur la procédure d'entraînement), une forme de stratégie d'élimination récursive de variables largement utilisée par les méthodes de sélection de variables appliquées aux problèmes biomédicaux. Cela permet d'optimiser les performances du modèle en calculant un score MCC à chaque itération et en sélectionnant finalement l'ensemble restreint de variables permettant les meilleures performances des modèles sur les données d'entraînement. Les variables dites redondantes par rapport à celles sélectionnées à cette étape sont mises de côté. La redondance a été évaluée en utilisant les corrélations de Pearson et de Spearman, ainsi que le score de gain d'information. Les variables ayant un score de corrélation de Pearson ou de

Spearman supérieur à 0,985 ont été considérées comme étant redondantes. De plus, compte tenu du contexte de classification binaire, deux variables ayant le même score de gain d'information ont également été considérées comme redondantes, car cela indique une puissance de séparation similaire entre les classes.

3. **Évaluation des modèles à signature de variables optimisée.** Les modèles générés par la boucle d'entraînement ont ensuite été évalués afin d'estimer leur éventuel surapprentissage sur les données d'entraînement. Cela a été fait avec plusieurs procédures de validation croisée sur les données, y compris les ensembles de test : validation croisée à 10 blocs, holdout répété et bootstrap. Pour chacune de ces procédures et pour chaque modèle, un score de MCC est calculé. Par la suite, la moyenne de ces valeurs de MCC est évaluée (AVG MCC), ainsi que l'écart type associé (STD MCC). Une valeur AVG MCC d'un ainsi qu'un STD MCC de zéro indiquent un modèle parfait qui classe tous les échantillons sans erreurs, peu importe le type d'évaluation.

4. **Production d'un rapport complet.** Pour finir, les variables redondantes sont recouvrées et un rapport complet du processus d'entraînement est produit.

Classificateur	Hyper-paramètres	AVG MCC	STD MCC	Signature
SVM	-C 0.001 -N 2 -K "supportVector.PolyKernel -E 4 -C -1"	1	0.001	SLC39A10, IL6R
SVM	-C 0.1 -N 0 -K "supportVector.PolyKernel -E 3 -C -1"	1	0	SLC39A10, IL6R, PMP2, TOMM34
Naive Bayes	-D	1	0	SLC39A10, IL6R, PMP2

Tableau 2.2: Comparaison des signatures de variables de trois modèles hautement performants et entraînés par notre stratégie d'entraînement de modèles d'enveloppement.

Des résultats préliminaires sur des données de cancer colorectal (voir section 3.2.1 pour plus de détails) avec une même exécution d'échantillonnage et avec un filtre DEG ont montré l'intérêt de confronter divers types d'algorithmes et avec différentes

valeurs d'hyper-paramètres (tableau 2.2). Dans les mêmes conditions, deux modèles SVM avec des valeurs d'hyper-paramètres différentes ont obtenu un même score d'AVG MCC et un STD MCC similaire, mais une signature différente. En effet, l'une englobe l'autre. Un autre modèle, un Naive Bayes, présente les mêmes valeurs de performances (AVG MCC de 1 et STD MCC de 0) et une signature similaire, mais pas exactement la même que les deux modèles de SVMs. Ainsi, en contrôlant pour un même jeu de données, deux types d'algorithmes ML au moins peuvent fournir un même niveau d'efficacité, mais des signatures de variables différentes. Dans ces conditions, il n'est pas possible de choisir quelle signature retenir. Il convient alors de ne pas supprimer cette variabilité en faisant un choix arbitraire, mais de la prendre en compte en la stabilisant via une méthode d'agrégation.

2.6 Agrégation

Chaque étape d'une stratégie de sélection de variables a une grande influence sur la composition de la signature. L'étape de filtrage initial, l'échantillonnage, mais aussi les classificateurs choisis, configurés avec diverses valeurs d'hyper-paramètres, conduisent tous à des compositions de signature variables. Il est ainsi judicieux de tirer parti de ces différentes variabilités de scénario en agrégeant les diverses signatures qu'elles produisent dans un résultat de consensus de signatures. Cela devrait ainsi augmenter à terme la stabilité et la généralisation des résultats de sélection de variables.

Une stratégie d'agrégation possible serait de prendre la liste de variables résultant de l'intersection ou de l'union de toutes les signatures des modèles [78]. Mais cette stratégie peut se révéler trop ou pas assez limitante. De plus, étant dans un contexte de sélection de variables par approche hybride ensembliste, il est souhaitable de considérer à la fois la complexité de la perturbation fonctionnelle (c'est-à-dire induite par le type de classificateur et la variabilité des hyper-paramètres) ainsi que la perturbation des données (c'est-à-dire induite par les stratégies d'échantillonnage). Ainsi, les deux possibilités omettent le caractère hiérarchique d'une approche hybride ensembliste. Par ailleurs, notre approche HEFS a l'originalité d'intégrer différents types de classificateurs et pour chaque type, nous faisons intervenir des milliers de

modèles avec des configurations très variables. Il convient alors d'appréhender la variabilité apportée par ces différentes perturbations et envisager une méthodologie d'agrégation qui tire parti de chacune d'elles.

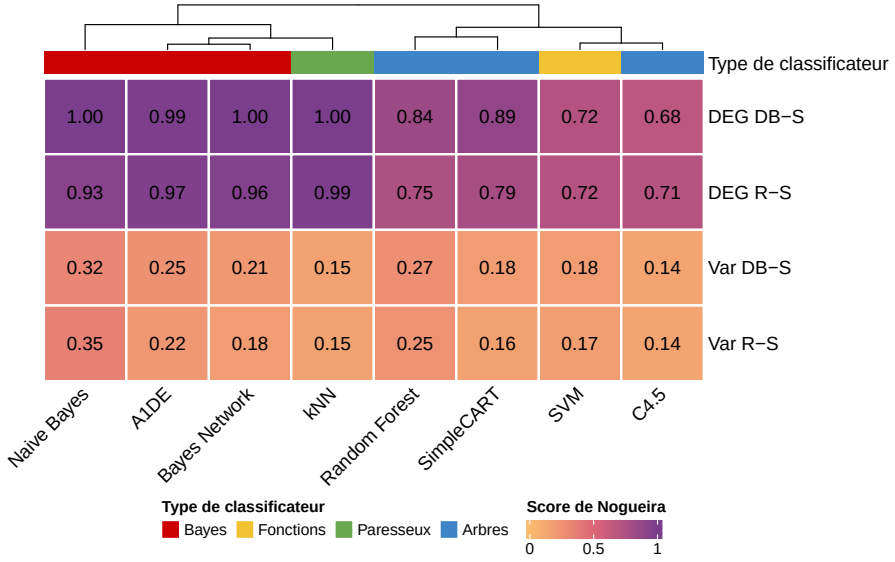

Figure 2.3: Carte de chaleur des scores de stabilité moyenne de Nogueira pour tous les modèles retenus au cours des 10 exécutions d'échantillonnage et pour les 4 scénarios HEFS (DEG DB-S, DEG R-S, Var DB-S, Var R-S).

Une analyse préliminaire de stabilité des signatures a été effectué. Ainsi, le score de Nogueira a permis d'évaluer l'uniformité des signatures à travers des modèles performants par classificateur (AVG MCC d'au moins 0,7 et STD MCC de maximum 0,1), exécution d'échantillonnage et scénario (figure 2.3). Pour un scénario et un classificateur donnés, le score de Nogueira moyen pour les 10 exécutions d'échantillonnage est rapporté dans la carte de chaleur. Les meilleurs scores sont donnés pour les scénarios basés sur DEG, ce qui signifie que les variables composant les signatures sont très stables à travers les modèles du même type de classificateur et pour chacune des 10 exécutions d'échantillonnage. En effet, tous les modèles bayésiens et les kNN ont un score de Nogueira moyen supérieur à 0,93, tandis que les modèles basés sur les arbres de décision et les SVM ont un score entre 0,68 et 0,89. En revanche, la tendance générale pour les scénarios basés sur Var est vers des scores de Nogueira plus bas, ce qui indique des signatures très variables.

Cette instabilité doit être prise en compte et exploitée et pour cela, l'agrégation de l'approche HEFS proposée est une agrégation multi-niveaux (figure 2.4). Pour construire un consensus entre les différentes signatures de modèles obtenus, nous

avons envisagé une stratégie d'intégration à trois niveaux : intégration des signatures de modèles au niveau du classificateur, puis au niveau de l'échantillonnage entre les types de classificateurs, et enfin à travers les échantillonnages pour produire un consensus global. Le premier et le deuxième niveau abordent le problème de la perturbation fonctionnelle et le troisième niveau prend en compte la perturbation des données.

Pour plus de clarté, considérons n comme le nombre d'échantillonnages effectués et m comme le nombre de types de classificateurs. Notez que les résultats présentés dans la suite du document ont été obtenus avec $n = 10$ et $m = 8$ en HEFS mode standard et $n = 10$ et $m = 5$ en mode *light*.

1. **Intégration des signatures de modèles au niveau du classificateur.** Au premier niveau de notre processus d'agrégation, nous avons réalisé une intégration entre les résultats obtenus avec les modèles générés pour chaque combinaison classificateur-échantillonnage. Tout d'abord, les modèles ont été filtrés et seuls ceux ayant un MCC moyen d'au moins 0,7 et un MCC STD de 0,1 ou moins ont été conservés. Les signatures des modèles d'apprentissage automatique restants ont été enrichies avec les variables redondantes de leur échantillonnage correspondant, qui avaient été mises de côté pendant le processus d'entraînement. Ensuite, l'union de ces signatures enrichies a été calculée pour chaque échantillonnage S_i et chaque classificateur C_j, ce qui a donné des listes de variables $n * m$ pour chaque échantillonnage.

2. **Intégration au niveau de l'échantillonnage entre les classificateurs.** Deuxièmement, l'agrégation a été réalisée au niveau de l'échantillonnage, permettant ainsi l'intégration des résultats des classificateurs pour un échantillonnage donné. À cette étape, une intersection des listes de variables obtenues pour chaque classificateur à l'étape précédente a été calculée pour chaque échantillonnage. Cela a donné $n = 10$ listes de variables, une par échantillonnage, c'est-à-dire $F^{S_i} = \bigcap_{j=1}^{m} F_{C_j}^{S_i}$ pour un échantillonnage S_i.

3. **Intégration à travers les échantillonnages pour un consensus global** Enfin, le niveau le plus élevé de l'agrégation a été mis en œuvre afin de produire une signature globale et stable. À partir des listes intermédiaires de

variables par échantillonnage (F^{S_i} calculées à l'étape 2), une signature a été calculée en prenant les variables apparaissant dans au moins 50% de nos listes d'échantillonnage dans le mode standard et au moins 40% dans le mode *light*. Ce relâchement du seuil s'explique par une nécessité de prendre en compte davantage de variabilité alors que la valeur de m (le nombre de types de classificateurs) entre le mode standard et le mode *light* a été diminuée.

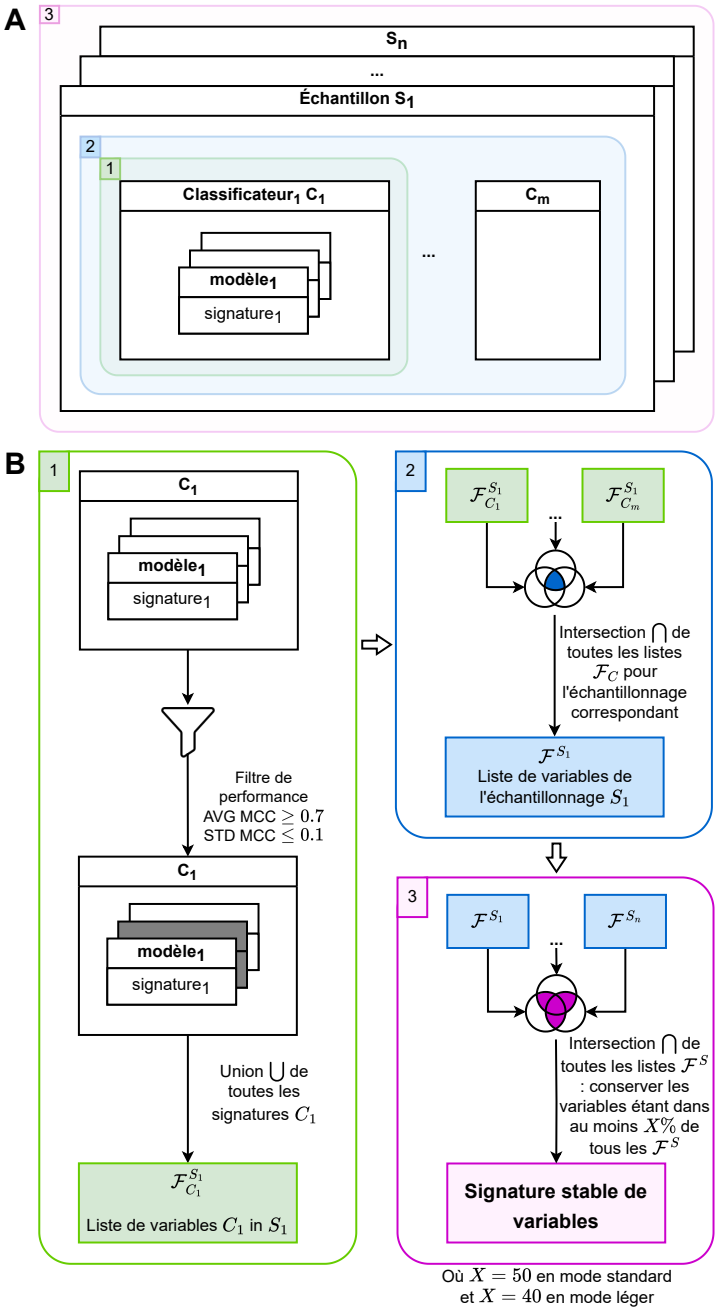

Figure 2.4: (A) Agrégation à trois niveaux : (A.1) niveau type de classificateur (entre modèles d'un même type de classificateur et par exécution d'échantillonnage), (A.2) niveau d'échantillonnage (entre types de classificateurs pour une exécution d'échantillonnage donné), (A.3) consensus global (à travers les échantillonnages). (B.1) Premier niveau : processus spécifique à chaque type de classificateur réalisant l'union des signatures des modèles d'enveloppement ayant passé les filtres de performance. (B.2) Deuxième niveau : agrégation par échantillonnage en calculant l'intersection par échantillonnage entre les listes de variables obtenues à partir des différents types de classificateurs au niveau précédent. (B.3) Troisième niveau : agrégation des listes de variables par échantillonnage du niveau précédent pour obtenir une signature de variables stable. Cette signature est composée des variables incluses dans au moins 40 ou 50% des échantillonnages pour le mode standard et *light* respectivement.

2.7 Conclusion

Ce deuxième chapitre exposait la méthodologie mise en place pour participer à la recherche sur la définition de signatures de biomarqueurs robustes à partir des données transcriptomiques dans un contexte de cancer.

Nous avons abordé cette question par la conception d'une approche hybride ensembliste de sélection de variable (HEFS). Comme évoquée dans le chapitre 1, les approches HEFS présentent un concept robuste, car elles intègrent pleinement des considérations de perturbations de données et fonctionnelles. D'un point de vue conceptuel, leurs résultats sont donc censés être davantage robustes que les méthodes traditionnelles. Dans nos recherches, nous avons cependant souhaité souligner le caractère encore très versatile de ces approches. En effet, lors de l'analyse de données aussi larges que celles de transcriptomique, il est crucial d'explorer diverses étapes préliminaires de réduction des dimensions.

Dans nos investigations, nous avons intégré des méthodes d'analyse de variance (Var) et d'analyse de gènes différentiellement exprimés (DEG).

Il est aussi important d'adapter la méthode de perturbation des données à la méthode subséquente de sélection de variable et à la structure connue des données étudiées. Ainsi, notre sélection de variables étant basé sur des méthodes d'apprentissage automatique d'enveloppement, nous avons souligné l'importance de conserver les proportions de classes lors de l'élaboration de paires de jeu de données d'entraînement et de test (R-S). Pour aller au-delà et s'agissant de données de cancer qui peuvent avoir des variations intra-classe, nous avons proposé d'explorer en plus une méthode d'échantillonnage prenant en compte de possibles groupes d'échantillons intra-classe (DB-S).

Nous avons choisi une stratégie de sélection de variables appuyée par des algorithmes d'apprentissage automatique d'enveloppement. Ce choix a été dicté par l'absence de consensus de la communauté scientifique quant au moyen de définir un classificateur adapté à un scénario biologique d'intérêt. N'ayant pas de clef de détermination à notre disposition, il est avantageux de s'appuyer sur diverses méthodes et leur variabilité pour observer de forts comportements communs.

Finalement, un autre point critique que nous avons abordé ici est la méthode d'agrégation mise en place pour tirer parti des diverses perturbations réalisées dans le

processus. Celle que nous proposons exploite les variations intrinsèques à un type de classificateur vis-à-vis des multiples configurations possibles de ses hyper-paramètres (1er niveau d'agrégation), le pluralisme des types de classificateurs à notre disposition (2ème niveau) et enfin, l'avantage de l'analyse en multi-échantillonnage (3ème niveau).

Chapitre 3

Évaluation de l'approche HEFS

3.1 Introduction

Ce troisième chapitre expose les investigations menées pour comprendre l'impact de divers choix de concepts (filtrage, échantillonnage, méthode d'agrégation) sur la formulation de signatures robustes de biomarqueurs de cancer à partir d'un même jeu de données. Nous avons donc réalisé une étude approfondie comparant quatre scénarios HEFS en mode standard pour l'identification de biomarqueurs à partir de données de RNA-seq d'adénocarcinome du côlon et du rectum de stade IV, c'est-à-dire de cancer colorectal (CRC), par rapport à des échantillons normaux. Cette approche HEFS, en mode *light*, a aussi été appliquée à l'identification de biomarqueurs du carcinome à cellules claires du rein (KIRC) et de l'adénocarcinome pulmonaire (LUAD) de stade I et du carcinome de l'endomètre du corps utérin (UCEC) de stade III. Dans une première section, nous présentons les cas d'usage étudiés. Ensuite nous exposons les résultats approfondis de comparaisons de formulation des signatures stables pour le CRC par quatre scénarios HEFS utilisant un mode standard de notre protocole. Nous présentons de plus l'analyse de robustesse des résultats obtenus par l'emploi d'une cohorte de CRC d'évaluation externe. Nous avons étudié par ailleurs la capacité de notre approche HEFS à se généraliser à d'autres types de cancer avec un protocole en mode *light* qui impliquait une réduction dimensionnelle précoce moins stricte par l'analyse de l'expression génique différentielle (DEG) et la formation de moins de modèles d'enveloppement. Nous poursuivons par une section de discussion et proposons des solutions viables concernant les nombreux défis qui surviennent

lors du calcul de consensus dans les signatures de biomarqueurs avec des stratégies HEFS. Pour finir, nous exposons nos conclusions.

3.2 Jeux de données

3.2.1 Cancer colorectal

La détection et la compréhension du stade IV du cancer colorectal (CRC) est encore aujourd'hui un motif de recherche majeur, car tandis que les stades précoces ont un pronostic favorable lorsqu'ils sont diagnostiqués, le stade IV qui représente une grande part des patients (environ 21%) présente un pronostic négatif [117]. Le but était donc de développer notre approche et de prouver son efficacité par l'identification de biomarqueurs robustes capables de différencier des échantillons sains d'échantillons de stade IV du CRC.

Jeu de données d'entraînement pour l'analyse en mode standard : TCGA CRC

L'approche HEFS a été expérimentée sur un jeu de données provenant du programme du The Cancer Genome Atlas (TCGA). Les données sur le carcinome adénocarcinome du côlon (identifiant de cohorte COAD) et du rectum (identifiant de cohorte READ) ont été obtenues à partir du portail de données Genomic Data Commons (GDC) en utilisant le package R TCGAbiolinks. Les données brutes de comptages de *reads* RNA-seq des cohortes COAD et READ ont été récupérés, pour un total de 462 et 171 patients respectivement avant le filtrage.

Les données ont été divisées en deux classes d'échantillons étiquetées selon le score de stade d'avancement de la Commission conjointe américaine sur le cancer (AJCC) : étiquette normal pour les échantillons de tissu normal solide et étiquette stade IV pour les échantillons de tumeur principale. Les données cliniques ont été utilisées pour associer les échantillons de tumeur à leur score AJCC correspondant, de telle sorte que tous les échantillons de stade IV (IV, IVa, IVb, IVc) ont été regroupés sous l'étiquette de stade IV. Tous les échantillons de tissu normal solide ont été regroupés dans une classe normale. Les données sur les biospécimens ont été utilisées pour filtrer les échantillons ayant moins de 70% de noyaux tumoraux

ainsi que les échantillons fixés sur formaline (ou FFPE) et pour gérer les doublons de patients et récupérer les métadonnées sur les lots d'analyses.

L'ensemble de données TCGA CRC est alors composé de 138 échantillons, dont 87 ont été identifiés comme stade IV et 51 comme normaux, et associés à environ 40 000 variables. Après l'application en parallèle du filtrage DEG et Var, il demeure respectivement 5 824 et 4 469 variables en mode standard HEFS. Ces deux ensembles de données ont ensuite été échantillonnés en parallèle avec les stratégies R-S et DB-S afin d'obtenir les quatre scénarios qui sont respectivement DEG R-S, DEG DB-S, Var R-S et Var DB-S. Pour DEG DB-S, 2 clusters ont été définis pour les données de stade IV et les données normales (Figures annexes B.3 et B.4). Pour Var DB-S, 2 et 3 clusters ont été respectivement calculés pour les données de stade IV et pour les données normales (Figures annexes B.1 et B.2).

En mode *light* HEFS, le filtrage Var a retenu les mêmes 4 469 variables que dans le mode standard, tandis que le filtrage DEG du mode *light* a retenu 6 825 variables. Ces deux jeux de données ont ensuite été soumis à la méthode d'échantillonnage DB-S.

Jeu de données d'évaluation pour l'analyse en mode standard : GSE50760

Pour évaluer la pertinence des signatures obtenues sur le jeu de données décrit ci-dessus, un jeu de données indépendant additionnel a été sélectionné, à savoir GSE50760 de la base de données GEO [118]. Ce jeu de données regroupe les données RNA-seq de 18 patients atteints de CRC de stade IV et à partir desquels trois types d'échantillons ont été collectés : tissu normal solide, tumeur principale et métastases. Comme nous étions intéressés par la caractérisation de la tumeur principale du CRC de stade IV, nous avons uniquement récupéré le premier et le deuxième type d'échantillons. Les premiers ont été étiquetés comme étant de classe normale et les seconds comme étant de classe de stade IV. Pour obtenir des données de comptages de *reads* comme pour le jeu de données de CRC TCGA, un prétraitement similaire à celui effectué par le portail GDC a été utilisé. Tout d'abord, un contrôle de qualité a été appliqué sur les données brutes en utilisant le logiciel FastQC, puis un alignement des *reads* d'échantillonnage a été effectué contre le génome hg38. Cet alignement a été réalisé à l'aide du logiciel STAR et avec le paramètre -quantMode GeneCounts

pour obtenir le comptage des *reads*.

3.2.2 Cancer du rein, des poumons et de l'utérus

Le carcinome à cellules claires du rein (KIRC) est l'un des cancers les plus mortels avec un taux élevé de patients diagnostiqués tardivement avec des métastases [119]. L'adénocarcinome pulmonaire (LUAD) représente une grande partie des cancers du poumon qui sont encore largement diagnostiqués chaque année [120]. Identifier les biomarqueurs du stade I du KIRC et du LUAD demeure ainsi une tâche importante, car la détection précoce de ces cancers reste difficile. La recherche de biomarqueurs du stade avancé du carcinome de l'endomètre du corps utérin (UCEC) est également d'un grand intérêt, car il s'agit du cancer le plus courant du système reproducteur féminin. Le nombre de diagnostics associé à ce cancer augmentent avec les années, notamment dans les pays développés et s'accompagnent d'un âge au diagnostic de plus en plus jeune [121].

La même procédure de collecte et de traitement des données que celle détaillée précédemment pour le cancer colorectal a été appliquée à trois jeux de données TCGA d'identifiants KIRC, LUAD et UCEC et qui sont associés aux maladies éponymes. Afin de calculer des biomarqueurs pour le stade I par rapport au phénotype normal, pour à la fois KIRC et LUAD, des échantillons de la tumeur principale annotés en stade I AJCC ont été récupérés comme classe stade I et du tissu normal solide de la cohorte, associée comme classe normale. Pour l'analyse de l'UCEC, la classe stade III est composée des échantillons de tumeurs primaires annotées en stade III AJCC et la classe normale est basée sur des échantillons de tissu normal solide. Après le traitement des données, nous avons obtenu 69 échantillons normaux et 262 échantillons de stade I pour le jeu de données KIRC, 45 échantillons normaux et 186 échantillons de stade I pour LUAD, et 19 échantillons normaux et 108 échantillons de stade III pour UCEC.

3.3 Évaluation quantitative du mode standard HEFS : performance des modèles d'enveloppement

Dans chacun des quatre scénarios HEFS en mode standard, environ 88 000 modèles d'enveloppement ont été entraînés (tableau 3.1).

Scénario	A1DE	Bayes Network	Naive Bayes	SVM	kNN	C4.5	Random Forest	Simple CART	Modèles entraînés	Modèles performants
DEG DB-S	24 000	1 672	30	4 521	10 800	8 360	19 800	19 800	88 983	85 855
DEG R-S	24 000	1 656	30	4 564	10 800	8 360	19 800	19 800	89 010	85 808
Var DB-S	24 000	1 629	30	4 280	10 800	8 360	19 800	19 800	88 699	64 492
Var R-S	24 000	1 595	30	4 279	10 800	8 360	19 800	19 800	88 664	78 799

Tableau 3.1: Nombre total de modèles d'enveloppement entraînés par scénario (lignes) et classificateur (colonnes), ainsi que le nombre total de modèles entraînés par scénario et le nombre total de modèles conservés après application du filtre de performance (MCC moyen de 0,7 ou plus et écart type du MCC de 0,1 ou moins).

La figure 3.1 illustre les tendances de performance (avec AVG MCC moyen et STD MCC) des modèles selon un classificateur et un scénario HEFS.

Il est intéressant de noter que les deux scénarios DEG permettent l'obtention de modèles d'enveloppement avec des scores de performance plus favorables que les scénarios Var. Pour chaque échantillonnage, la moyenne des AVG MCC et la moyenne des STD MCC ont été calculées. Une comparaison de ces moyennes entre DEG et Var a été effectué (valeur p ajustée $\leq 10^{-2}$).

En considérant tous les modèles avant le filtre de performance, nous remarquons que l'intervalle de valeurs des scores de performance est le plus grand pour la méthode SVM. En effet, les SVM affichent un score minimal de 0 et maximal de 1 en AVG MCC et score minimal de 0 et maximal d'environ 0.4 en STD MCC là où les autres algorithmes ont plutôt un score minimal de AVG MCC d'environ 0.6 et un score maximal d'environ 0.3 pour le STD MCC (exception faite des Random Forest). Cela montre une sensibilité plus élevée des SVMs aux variations de leurs hyper-paramètres, quel que soit le scénario. Dans une moindre mesure, les modèles basés sur les arbres de décisions montrent un comportement similaire.

Il convient de noter que seuls les modèles avec des scores de 0,7 et plus pour l'AVG MCC et de 0,1 et moins pour le STD MCC ont été conservés pour se concentrer sur les modèles de haute performance. Avec ces seuils, environ 3 000 modèles (3,4% du total des modèles entraînés) n'ont pas été retenus dans les deux scénarios DEG,

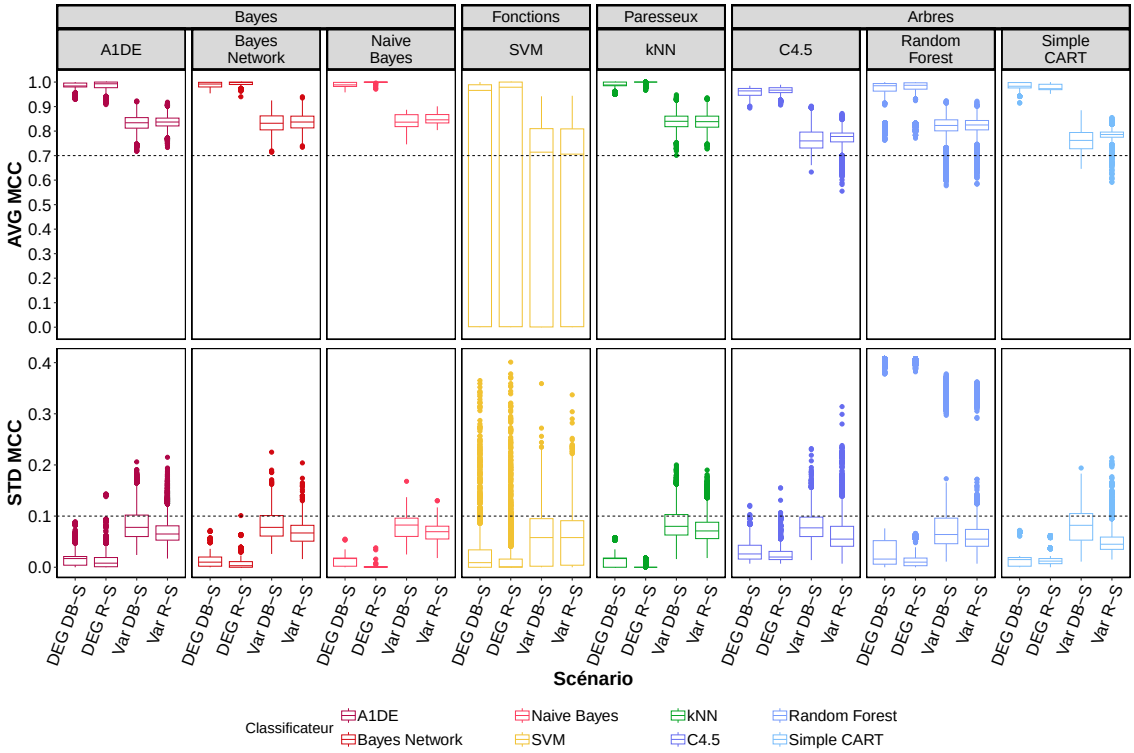

Figure 3.1: Scores de performance calculés pour les 4 scénarios de sélection de variables hybrides ensemblistes incluant les 10 exécutions d'échantillonnage avec une moyenne de 88 800 modèles entraînés chacune (pour huit classificateurs différents). Le panneau supérieur montre la distribution des valeurs du Coefficient de Corrélation de Matthews Moyen (AVG MCC), où une valeur de 1,0 indique des modèles de haute performance. Le panneau inférieur présente l'écart type par rapport à l'AVG MCC (STD MCC), où une valeur de 0,0 est la meilleure, car elle indique une performance égale dans toutes les évaluations. Dans chaque panneau, la ligne pointillée noire met en évidence la valeur seuil appliquée pour filtrer les modèles de faible performance pour les étapes HEFS suivantes. Les seuils sont fixés à un minimum de 0,7 pour l'AVG MCC et à un écart type maximal associé de 0,1 pour le STD MCC.

la plupart d'entre eux ayant été calculés par les méthodes SVM et Random Forest (tableau 3.1). Dans les scénarios Var, l'effet de cette sélection rigoureuse a été plus sévère, avec des pertes de 11,1% et 27,3% pour les modèles Var R-S et Var DB-S, respectivement.

3.4 Évaluation qualitative du mode standard HEFS : caractérisation des signatures

3.4.1 Taille variable des signatures des modèles d'enveloppement

Dans la Figure 3.2, nous analysons la distribution des longueurs de signatures incluant ou non les variables redondantes et en fonction du classificateur et du scénario employés. Chaque boîte à moustaches inclut la valeur moyenne de la taille de tous les modèles entraînés pour chacune des 10 exécutions d'échantillonnage effectuées pour chaque scénario sans redondance (panneau du haut) et avec redondance (panneau du bas). Les signatures finales sont définies à partir des signatures enrichies des variables redondantes. Il est intéressant d'observer qu'avant enrichissement, les signatures des modèles en scénario DEG sont plus courtes que celles en scénario Var (figure 3.2). Cela indique la capacité des modèles à ne nécessiter qu'un nombre restreint de variables en DEG contrairement aux modèles en Var.

Comme attendu, peu importe le scénario, les signatures avec variables redondantes sont plus longues que sans les variables redondantes. Néanmoins, un ratio multiplicatif de treize décrit ce passage de longueur de signature non redondante à redondante dans les scénarios DEG alors que le ratio n'est que d'environ deux pour les scénarios Var. Cela suggère que les données obtenues par le processus étiqueté DEG portent une grande similarité dans leur capacité à séparer les classes des données. À l'inverse, le jeu de données issu du processus étiqueté Var semble inclure davantage d'hétérogénéité et donc moins de variables capables de séparer avec une même force les échantillons des deux classes.

Par la suite, seules les signatures redondantes sont considérées. D'après le panneau du bas de la Figure 3.2, la médiane de la taille moyenne des signatures varie parmi les classificateurs de 24 à 38 pour les deux scénarios DEG, et de 7 à 17 pour les deux scénarios VAR, à l'exception du classificateur kNN. En se concentrant sur le classificateur kNN, la distribution de la taille des signatures est similaire pour les scénarios DEG R-S et Var R-S avec une taille médiane d'environ 24. Dans l'ensemble, nous pouvons observer une plus grande variabilité de taille pour les scénarios DEG par rapport aux scénarios Var. Ce comportement est encore plus prononcé pour les

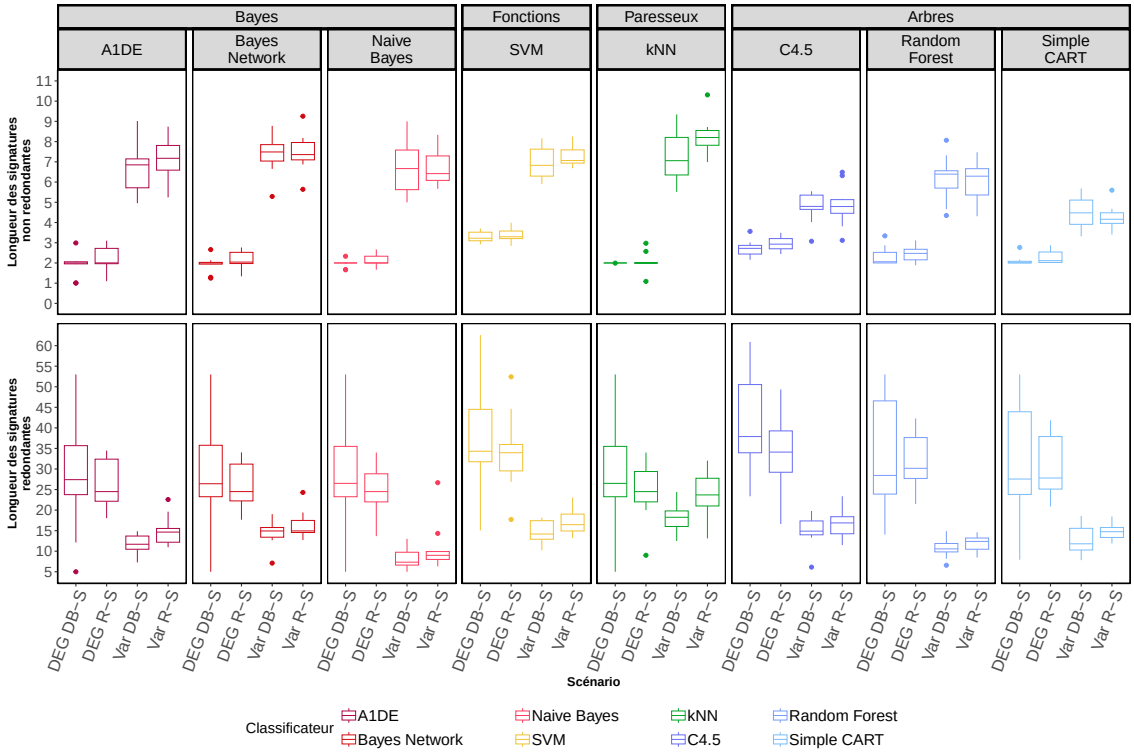

Figure 3.2: Variations des longueurs des signatures sans et avec redondance. Chaque boîte à moustaches représente la longueur moyenne des signatures pour chacune des 10 séries d'échantillonnage.

classificateurs basés sur les arbres, où certaines exécutions d'échantillonnage produisent des signatures de plus de 40 variables. En revanche, pour les deux scénarios Var, les signatures restent courtes, indépendamment du classificateur et de l'itération de l'exécution.

3.4.2 Agrégation des signatures

Notre processus d'agrégation est multi-niveaux (voir section 2.6 pour la méthodologie détaillée).

Le premier niveau vise à obtenir une signature par type d'algorithme pour chaque exécution d'échantillonnage. Il s'agit ici d'une première étape de prise en compte de la perturbation fonctionnelle opérée dans notre approche HEFS. Par exécution d'échantillonnage et par type de classificateur, la moyenne de taille des signatures transitoires à la première étape d'agrégation varie entre 17,60 au minimum avec les Naive Bayes en scénario Var DB-S et 3 938,90 avec les Simple CART en scénario Var R-S (Tableau 3.2, Tableaux de la section I de l'Annexe C). On constate que bien

que les tailles des signatures redondantes employées soient de l'ordre des dizaines, leur union peut excéder la centaine voir le millier. De façon globale, on remarque que les scénarios DEG semblent déjà donner des signatures dites plus stables que les scénarios Var avec des unions moins importantes. Par exemple, les modèles kNN donnent des signatures qui semblent grandement se ressembler en DEG (tailles moyennes de 45,20 et 39,0 pour les scénarios DB-S et R-S respectivement) alors que leur instabilité est forte en Var (tailles moyennes de 2 531,60 et 3 261,00 pour les scénarios DB-S et R-S). À l'inverse, les classificateurs de type arbre de décision semblent produire des signatures instables peu importe le type de scénario. Ces comportement sont corroborés par nos résultats préliminaires d'analyse d'instabilité présentés précédemment (section 2.6, Figure 2.3).

Classificateur	DEG DB-S	DEG R-S	Var DB-S	Var R-S
A1DE	85,50	119,30	2 513,40	3 207,00
Bayes Network	45,20	52,10	920,10	1 251,10
C4.5	2 686,30	2 460,60	2 480,10	2 994,00
kNN	45,20	39,60	2 531,60	3 261,00
Naive Bayes	45,20	39,60	17,60	32,10
Random Forest	1 742,90	2 752,40	2 506,10	3 215,90
Simple Cart	1 244,20	2 859,10	2 730,00	3 938,90
SVM	967,70	878,30	1 146,60	1 487,60

Tableau 3.2: Tailles moyennes des signatures transitoires de l'étape 1 d'agrégation à travers les 10 exécutions d'échantillonnage et pour chaque type de classificateur et scénario.

Cette première étape d'agrégation montre que certains types de classificateurs sont un facteur limitant de variabilité pour les étapes subséquentes d'agrégation alors que d'autres seront source de grande variabilité.

La seconde étape d'agrégation consiste globalement à une intersection des signatures transitoires précédemment exposées pour chaque exécution d'échantillonnage. Cette étape participe à prendre en compte la perturbation fonctionnelle opérée dans notre approche HEFS. On observe, pour chaque scénario, que les tailles moyennes des signatures transitoires par exécution d'échantillonnage varient entre 13,1 et 45,2 (Tableau 3.3, Tableaux de la section II de l'Annexe C). Ces chiffres confirment que les scénarios DEG proposaient des signatures hautement stables très tôt dans le processus. Cette analyse montre aussi que l'agrégation des signatures en scénario Var

peut aboutir à une haute stabilisation des signatures, par l'obtention de signatures transitoires courtes et ce, malgré une variabilité importante observée à la première étape d'agrégation.

Scénario	Taille moyenne des signatures transitoires
DEG DB-S	45,2
DEG R-S	39,6
Var DB-S	13,1
Var R-S	18,4

Tableau 3.3: Tailles moyennes des signatures transitoires de l'étape 2 d'agrégation à travers les 10 exécutions d'échantillonnage et pour chaque scénario.

La troisième étape d'agrégation consiste en une intersection non stricte des signatures transitoires exposées précédemment. Cette dernière étape d'agrégation permet d'intégrer les apports de la perturbation des données impliquée dans notre approche HEFS.

Nous nous sommes alors intéressés à la variabilité de la composition des signatures entre les exécutions d'échantillonnage (figure 3.3, panneaux 1 à 4) et entre les scénarios (figure 3.3, panneau 5). Pour les deux scénarios basés sur DEG, l'intersection entre les signatures est de 20 variables, montrant ainsi une stabilité importante entre les exécutions d'échantillonnage. En revanche, pour les deux scénarios basés sur VAR, l'intersection est extrêmement faible : 1 pour Var R-S et nulle pour Var DB-S. Nous avons ensuite effectué la dernière étape d'agrégation en mode standard (voir section 2.6) qui rassemble pour chaque scénario toutes les variables présentes dans au moins 50% des listes de variables des 10 exécutions d'échantillonnage. Nous avons obtenu respectivement 28, 30, 6 et 4 variables stables pour les scénarios DEG DB-S, DEG R-S, Var DB-S et Var R-S (barres bleues dans la figure 3.3, panneaux 1 à 4). L'intersection de toutes les signatures stables (figure 3.3, panneau 5) des 4 scénarios donne 3 variables communes tandis que 29 sont spécifiques aux scénarios DEG (21 communes aux deux scénarios DEG) et 2 sont spécifiques aux scénarios Var (1 commune aux deux scénarios Var). Le chevauchement global est faible en raison de la taille des quatre signatures incluses. Cependant, le plus petit ensemble de variables des scénarios Var est largement inclus dans les signatures des scénarios DEG.

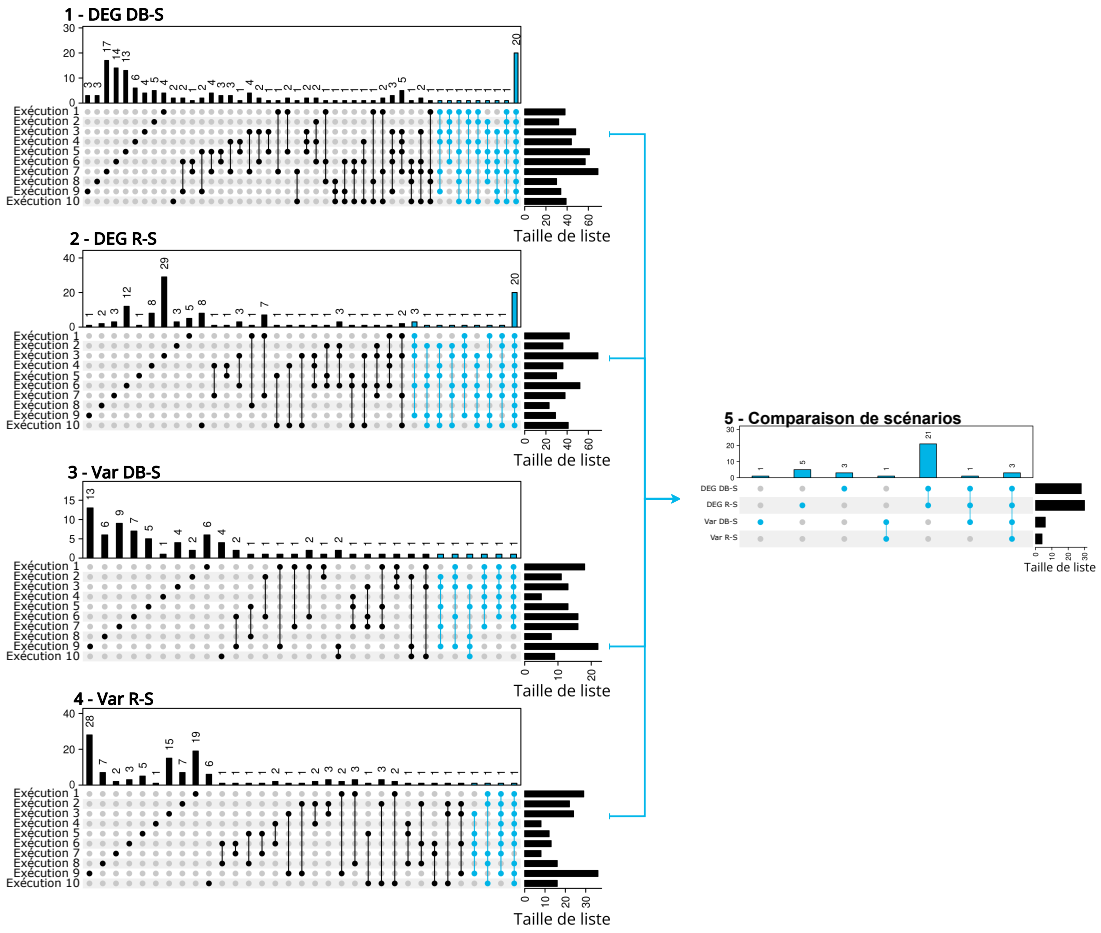

Figure 3.3: Distribution des quatre signatures HEFS. (1 à 4) Distribution des variables à travers toutes les exécutions d'échantillonnage pour les quatre scénarios HEFS basée sur les listes de variables obtenues après la deuxième étape d'agrégation, c'est-à-dire échantillon par échantillon. Le graphique final (panneau 5) correspond à la comparaison des quatre signatures obtenues à l'étape finale du processus d'agrégation de chaque scénario HEFS (en bleu dans les graphiques 1 à 4).

3.4.3 Caractérisation des signatures stables des 4 scénarios HEFS

Les 4 signatures ont été analysées d'un point de vue biologique en utilisant la base de données DisGeNET via l'outil EnrichR. Les résultats sont présentés dans le tableau 3.4 où les gènes liés au CRC sont en gras. Nous pouvons observer que les quatre signatures sont bien caractérisées dans le contexte du CRC. Notons la présence du gène KRT80 (donné par 3 des 4 scénarios) et qui a récemment été proposé comme biomarqueur pronostique indépendant pour le CRC en raison de son rôle dans la promotion de la migration et de l'invasion du CRC via l'activation de la voie

AKT [122].

Scénario	Signature associée	Enrichissement de l'annotation de cancer colorectal : p valeur ajustée
DEG DB-S	**ETV4, ESM1, OTOP2, CDH3, SCGN, SALL4, IL6R, SCARA5, PVT1, ABCA8, MRGBP, BMP3, EIF4E3**, GLP2R, BEST4, FAM135B, CLEC3B, ENPP6, KRT80, CA7, METTL7A, SLC39A10, ELANE, SEMA6A-AS2, SLC51B, SCN9A, MAFG-AS1, TMIGD1	7.8×10^{-2}
DEG R-S	**ETV4, ESM1, OTOP2, CDH3, SCGN, SALL4, IL6R, SCARA5, PVT1, ABCA8, MRGBP, NR3C2, CBLN2**, GLP2R, BEST4, FAM135B, CLEC3B, ENPP6, KRT80, CA7, METTL7A, SLC39A10, ELANE, SEMA6A-AS2, SLC51B, SCN9A, MAFG-AS1, PLPP1, NKRF, UGP2	1.4×10^{-1}
Var DB-S	**ETV4, ESM1, OTOP2, AJUBA**, KRT80, LINC00974	5.3×10^{-2}
Var R-S	**ETV4, ESM1, OTOP2, AJUBA**	3.4×10^{-2}

Tableau 3.4: Caractérisation des quatre signatures HEFS par analyse d'enrichissement d'annotation du terme DisGeNet "Colorectal cancer" par l'outil EnrichR.

3.4.4 Évaluation indépendante des signatures stables

La capacité des quatre signatures HEFS à généraliser à une cohorte externe de patients atteints de CRC de stade IV a été évaluée. Pour cela, nous avons entraîné et optimisé des modèles d'apprentissage automatique sur les données TCGA CRC, puis nous avons évalué la robustesse des modèles, et donc leur signature associée, sur l'ensemble de données GSE50760.

Pour mettre en place une évaluation classique et indépendante de l'apprentissage automatique, nous avons utilisé le package R h2o, qui permet l'entraînement automatique de plusieurs classificateurs d'apprentissage automatique avec une optimisation des hyper-paramètres par recherche aléatoire dans une grille. Tout d'abord, nous avons considéré les deux ensembles de données TCGA obtenus après l'étape de filtrage HEFS effectuée par DEG et Var (figure 2.1). Seules les variables des signatures stables de chacun des 4 scénarios HEFS ont été conservées : pour les scénarios DEG, 28 variables ont été conservées pour DB-S et 30 pour R-S, tandis que pour les

scénarios Var, 6 et 4 variables ont été conservées pour DB-S et R-S respectivement.

Nous avons ensuite effectué l'entraînement de plusieurs modèles d'apprentissage automatique sur ces quatre ensembles de données. Le package h2o a été utilisé pour entraîner des modèles de boosting de gradient (*Gradient Boosting Machine* GBM), de Random Forest et des modèles linéaires (comme des *Generalized Linear Model* GLM). Le meilleur modèle a été sélectionné en comparant les performances du modèle lors d'une évaluation en validation croisée à 10 blocs en utilisant l'ensemble de données TCGA d'entrée associé (tableau 3.5). Il est intéressant de noter que les modèles entraînés en utilisant les signatures basées sur DEG ont affiché des scores MCC parfaits lors d'une validation croisée à 10 blocs, ce qui peut indiquer un surapprentissage. En revanche, les modèles entraînés sur les signatures basées sur Var étaient très performants tout en présentant une importante déviation standard.

Scénario	**Nombre de variables**	**Meilleur modèle optimisé**	**MCC moyen en validation croisée en 10 blocs**	**Score de MCC sur GSE50760**
DEG R-S	30	GBM	1.00 ± 0.00	0.89
DEG DB-S	28	GBM	1.00 ± 0.00	0.95
Var R-S	4	GLM	0.88 ± 0.14	0.89
Var DB-S	6	GBM	0.88 ± 0.15	0.89

Tableau 3.5: Évaluation des différentes signatures sur des modèles optimisés et testés sur les données d'origine en validation croisée et sur un jeu de données externe, GSE50760. GBM : *Gradient Boosting Machine*, GLM : *Generalized Linear Model*

Le meilleur modèle pour chacun des quatre scénarios (des GBM pour tous les scénarios sauf Var R-S pour lequel un GLM a été le meilleur) a ensuite été évalué sur notre ensemble de données externe, c'est-à-dire l'ensemble de données GSE50760 (voir section 3.2.1). Comme pour l'ensemble de données TCGA, nous avons récupéré quatre sous-ensembles issus des données d'évaluation avec différents ensembles de variables correspondant aux quatre signatures HEFS stables. L'évaluation sur les sous-ensembles de GSE50760 a donné des scores MCC élevés, c'est-à-dire 0,95 lors de l'utilisation de la signature DEG DB-S et 0,89 dans les 3 autres cas. Bien que les quatre signatures, Var et DEG (où les Var sont plus courtes que les DEG) aient réussi à séparer avec succès les échantillons de stade IV et les échantillons normaux, les variables n'ont pas la même importance dans le processus de décision. Cette importance est évaluée différemment par h2o pour les modèles GBM et les GLM.

Pour les GBM, l'importance de la variable prend en compte l'impact de l'utilisation de la variable dans le processus de partitionnement sur l'erreur quadratique. Pour cela, la différence de variance de la variable est évaluée entre le noeud et ses noeuds enfants. Pour les modèles GLM, l'importance de la variable est une représentation de l'amplitude des coefficients associés aux variables et utilisées dans la construction du modèle. Dans tous les cas, que ce soit pour un modèle GBM ou GLM, nous accédons aux scores d'importance de variable normalisée où la variable la plus importante a le score maximal de un et les variables avec aucun impact ont un score de zéro. Il est intéressant de noter que les plus informatives semblent être les trois variables communes aux quatre scénarios : OTOP2, ETV4 et ESM1 (figure 3.4).

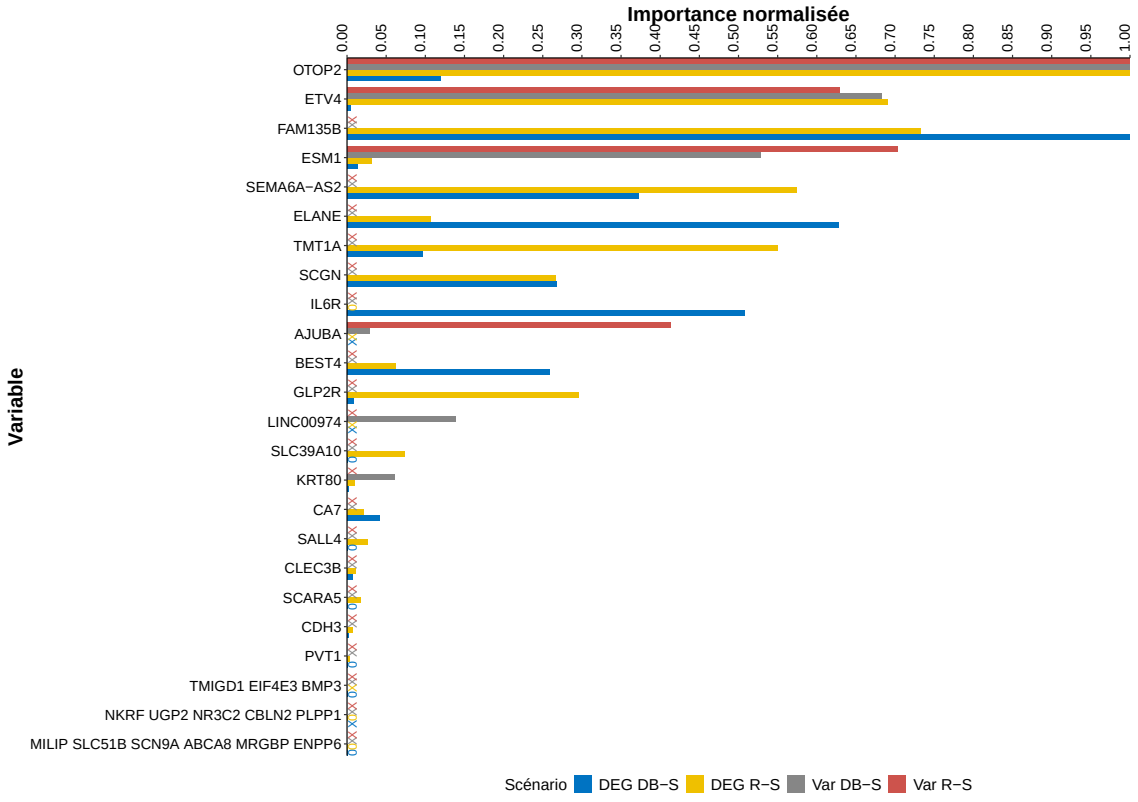

Figure 3.4: Importance normalisée des variables dans chacun des quatre modèles d'apprentissage automatique optimisés entraînés sur les données de cancer colorectal TCGA à l'aide de la bibliothèque R h2o. Le jeu de données TCGA a été réduit en fonction des quatre scénarios différents de type HEFS. La variable la plus importante pour un modèle a un score d'un et les autres scores sont normalisés et classés par rapport à celui-ci. Un score supérieur à 0 est indiqué par une barre pleine, un score nul est mis en évidence par un 0 tandis que l'absence de la variable pour le modèle considéré est représentée par un symbole X.

La signature stable la plus courte produisant une bonne séparation est celle du scénario Var R-S, qui est composée de quatre gènes : OTOP2, AJUBA, ESM1 et

ETV4. AJUBA est le seul gène qui n'a pas été trouvé dans les scénarios HEFS basés sur DEG. Dans [123], OTOP2 a été cité comme un biomarqueur potentiel du CRC. OTOP2 a été prouvé comme étant impliqué dans l'activité des canaux protoniques, ce qui est cohérent avec notre problème biologique d'intérêt puisque la maladie du cancer colorectal est étroitement liée aux changements électrolytiques [124]. ETV4 a été suggéré comme une cible thérapeutique potentielle dans le contexte du CRC dans [125, 126]. ESM1 favorise la progression du CRC, la migration cellulaire et l'invasion [127] et pourrait être d'intérêt en tant que cible thérapeutique [128]. Parmi les 4 gènes soumis, l'outil ToppGene pour l'analyse d'enrichissement des listes de gènes a trouvé que ETV1 et ESM1 étaient significativement associés (valeur p ajustée par Bonferroni est de 3.348×10^{-2}) au médicament PD-0325901 (également connu sous le nom de mirdamétinib). Ce médicament a été étudié dans le contexte du CRC dans deux essais cliniques, NCT00147550 et NCT02510001. Enfin, AJUBA a été trouvé comme étant associé à la migration cellulaire dans le cancer colorectal [129, 130].

3.5 Évaluation de la méthode d'agrégation

Nous avons analysé les avantages de notre méthode d'agrégation multi-niveaux en la comparant à une procédure de sélection de variables classique (figure 3.5). Cette procédure consiste à prendre le meilleur modèle parmi des milliers de modèles entraînés indépendamment de l'exécution d'échantillonnage. Pour chacun de nos quatre scénarios en mode standard, nous avons sélectionné le meilleur modèle avec le meilleur AVG MCC ainsi que le meilleur STD MCC et d'autres métriques de performance, telles que la précision, à travers les méthodes d'apprentissage automatique et les échantillonnages. Nous avons ensuite comparé ces quatre signatures de modèle aux signatures produites par nos quatre scénarios HEFS. Dans l'ensemble, les modèles SVM se sont avérés être les meilleurs, à l'exception du scénario Var DB-S pour lequel un modèle kNN les surpasse.

Pour rappel, les variables sélectionnées au sein de chaque signature HEFS stable doivent être sélectionnées pour au moins 50% des échantillonnages et par au moins un modèle de chaque classificateur. Ainsi, lors de l'analyse des scénarios DEG DB-

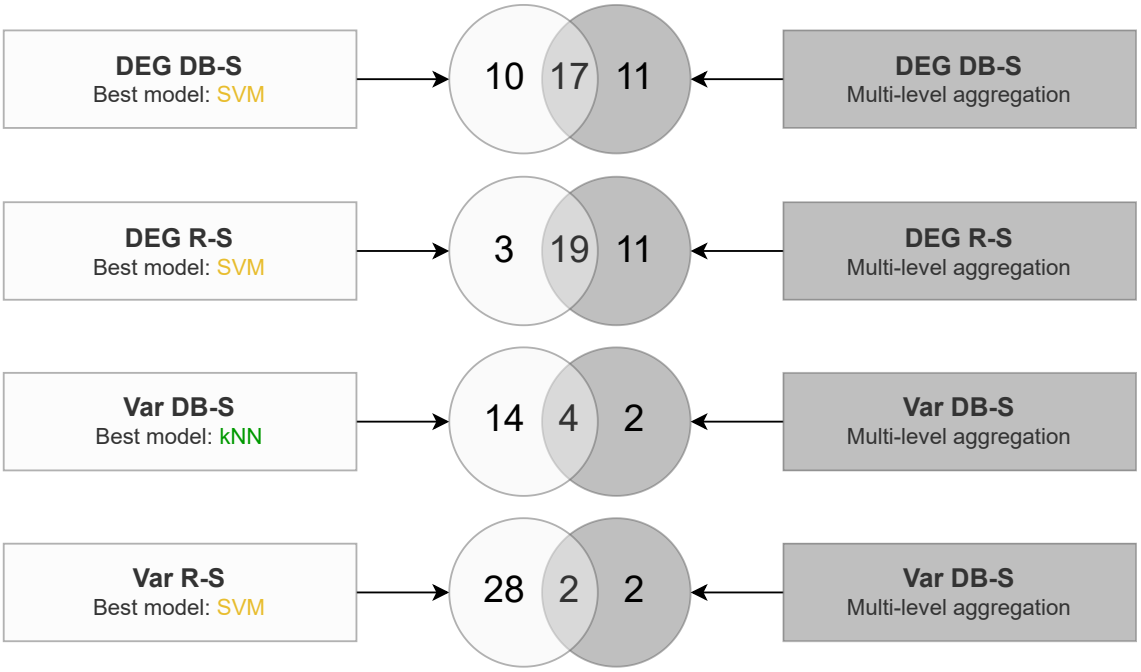

Figure 3.5: Comparaison de la composition des signatures entre une stratégie de sélection de variables standard (gris clair) et l'approche HEFS proposée (gris foncé).

S, le concept même du processus d'agrégation des signatures définit des arguments plus convaincants en faveur de l'ajout de 11 et 17 gènes par opposition aux 10 gènes donnés par le seul modèle SVM.

En ce qui concerne la stabilité de la sélection de variables avec le score de Nogueira (figure 2.3, section 2.6), les modèles SVM sont parmi les classificateurs avec les scores les plus bas pour tous les scénarios et avec une valeur de 0,72, en particulier pour les deux scénarios DEG. Cela met en évidence la grande instabilité des signatures des modèles SVM entre elles par rapport aux signatures de modèle d'autres types de classificateurs. De plus, les longueurs des signatures pour les modèles SVM sont plus grandes, quel que soit le scénario (figure 3.2). Ces résultats démontrent la tendance des SVM à sélectionner des signatures plus longues avec une composition variable de variables. Par rapport aux 10 ou 3 variables spécifiques identifiées par les modèles SVM dans les scénarios DEG, les 11 variables supplémentaires spécifiques aux signatures HEFS offrent la garantie d'avoir été obtenues par au moins 5 modèles SVM sur les 10 échantillonnages en plus d'autres types de modèles d'apprentissage automatique. Ces observations peuvent être étendues aux scénarios Var. En effet, les meilleurs modèles dans ces scénarios calculent de longues signatures, dont une partie importante n'est pas robuste, car elle ne se recoupe pas

avec les résultats HEFS (14 sur 16 pour Var DB-S et 28 sur 30 pour Var R-S). Dans ce cas, il devient encore plus important d'intégrer les perturbations de données et fonctionnelles pour identifier les variables les plus stables tout en filtrant l'ensemble de données en fonction de la variance.

3.6 Généralisation de l'approche HEFS en mode *light*

3.6.1 Application à trois cas d'usage

Nous avons examiné la capacité de notre approche à se généraliser à d'autres cas d'utilisation du cancer pour récupérer des variables stables tout au long du processus d'agrégation. La procédure HEFS en mode *light* a été utilisée pour calculer des signatures stables afin d'identifier des biomarqueurs distinguant les échantillons de KIRC Stage I des échantillons Normaux, les échantillons de LUAD Stage I des échantillons Normaux et les échantillons de UCEC Stage III des échantillons Normaux. Pour assurer la robustesse du mode *light* par rapport au mode standard, l'approche en mode *light* a été validée en ré-analysant le jeu de données sur le CRC. À noter que le mode *light* n'emploie que la stratégie d'échantillonnage DB-S et non R-S, car aucune différence significative n'a été observée dans le jeu de données CRC Stage IV entre ces types échantillonnages.

Les résultats de la procédure HEFS sur le jeu de données CRC sont comparables entre les modes *light* et standard. En mode *light* de l'approche HEFS, 33 et 5 gènes ont respectivement été sélectionnés pour les signatures DEG DB-S et Var DB-S, avec deux variables non chevauchantes (AJUBA et CPNE7) spécifiques au scénario Var (tableau 3.6). AJUBA était déjà le seul gène spécifique à Var comparé à DEG en mode HEFS standard. Il est à noter que 4 des gènes de la signature Var DB-S en mode *light* sont inclus dans la signature de 6 variables de Var DB-S en mode standard décrite dans les sections précédentes. De plus, 25 des 33 gènes identifiés pour le mode DEG DB-S en mode *light* sont communs avec le mode DEG DB-S en mode standard. Trois gènes de la signature DEG DB-S en mode standard n'ont pas été sélectionnés en mode *light*. Cependant, ces gènes étaient parmi les moins stables en mode standard, car ils n'étaient sélectionnés que dans cinq des dix échantillonnages à la dernière étape d'agrégation.

Phénotype	Scénario	Signature	Annotation DisGeNet (valeur p ajustée)
CRC - Normal vs Stade IV	DEG DB-S	**CDH3, IL6R, SCGN, SALL4, ESM1, SCARA5, ETV4, OTOP2, PVT1, CBX8, EIF4E3, ABCA8, MRGBP, LYVE1**, GLP2R, BEST4, FAM135B, CLEC3B, ENPP6, KRT80, CA7, METTL7A, SLC39A10, ELANE, SEMA6A-AS2, MAMDC2, SLC51B, PLPP1, NKRF, VSTM2A, GLIPR2, UGP2, TMIGD1	Colorectal cancer : 1.2*10^-1
	Var DB-S	**OTOP2, AJUBA, ETV4**, KRT80, CPNE7	Colorectal cancer : 1.4*10^-1
KIRC - Normal vs Stade I	DEG DB-S	**ATP6V0A4, TMEM213**, ATP6V0D2, IRX2, SLC9A4	Conventional (Clear Cell) Renal Cell Carcinoma : 5.6*10^-2
	Var DB-S	ADAM18, AC073172.1, SEMG2, LINC01983, LINC02437, PRR35, LINC00864, AC090709.1, LINC02121, AL160286.3	Conventional (Clear Cell) Renal Cell Carcinoma : NA
LUAD - Normal vs Stade I	DEG DB-S	**EMP2, AGER**, PDLIM2, STX11, GYPE	Carcinoma of lung : 1.7*10^-1

	Var DB-S	**EMP2, AGER**, SGCG, PTPN21, STX11, GYPE, LGR4	Carcinoma of lung : 9.3*10^-2
UCEC - Normal vs Stade III	DEG DB-S	**AURKA, DNMT3B, BIRC5, MYBL2, EZH2, CCNB1, CDC25C, PLK1, AURKB, CCNE1, CHEK1, FEN1, UHRF1**, E2F2, TACC3, RAD51, MCM10, PRR11, GTSE1, SPAG5, UBE2T, TPX2, CDC45, CBX7, CDKN3, GINS1, ASF1B, CDCA3, PLSCR4, CENPA, CDC20, NCAPH, ZWINT, HJURP, AUNIP, SGO1, CDCA8, TROAP, ESPL1, CENPO, KIF2C, NUF2, TBC1D7, MTFR2, CDCA5, CENPU, SKA1, CCNB2, TEDC2, CCNF, POC1A, PTTG1, MELK, SKA3, BUB1, SHCBP1, RRM2, UBE2C, KIF18B, JPT1, KIFC1, MEF2C-AS1, LINC02310, AP001528.3, ADH1B, CENPN, AC027449.1, KNSTRN, TICRR, PGD, AC107959.1, SPC24, FAXDC2, ORC1, HAPLN1, MAOB, CDC6, ADM2, RERG, SPC25, PCLAF, FAM83D, KLHDC1, CENPF, EME1, HMGB3, POLQ, ORC6, GPRASP1, PGM5P4, TTK, CKS2, DLGAP5, PIMREG, KIF11, ZNF25, KPNA2, CDCA2, ARHGAP11A, EFNA3, CEP55, AC004554.2	Endometrial carcinoma : 5.8*10^-3

Var DB-S	**CCNB1, BIRC5, MYBL2, CDC25C**, ZWINT, STIL, CKS2, SKA1, TEDC2, UBE2C, IQGAP3, MEF2C-AS1, ASF1B, ZNF300P1, KIF20A, CDC20, HJURP, TCEAL6, TACC3, DEPDC1, TTK, TROAP, KLHL4, ASPA	Endometrial carcinoma : 1.4*10^-1

Tableau 3.6: Analyse de l'enrichissement des annotations DisGeNet pour caractériser les signatures stables obtenues par le mode *light* HEFS dans les scénarios Var DB-S et DEG DB-S dans le contexte du cancer colorectal de stade IV (CRC), du cancer du rein à cellules claires de stade I (KIRC), de l'adénocarcinome du poumon (LUAD) et du carcinome du corps de l'utérus et de l'endomètre de stade I (UCEC). Les gènes en gras sont ceux qui sont impliqués dans l'annotation associée.

Pour le cas d'usage de KIRC, les signatures des deux scénarios HEFS *light* sont respectivement de 5 variables et de 10 variables pour DEG DB-S et Var DB-S (tableau 3.6). Notons qu'après l'étape initiale de filtrage, 5 353 variables ont été conservées pour Var et 7 891 pour DEG, avec un chevauchement de 32%. Les signatures sont distinctes, mais les gènes qui les composent font partie des 32% de variables communes. La signature DEG DB-S contient des gènes déjà connus pour être impliqués dans KIRC, tandis que la signature Var DB-S inclut des longs ARN non codants moins étudiés (tableau 3.6).

Pour le cas d'usage de LUAD, la signature DEG DB-S en mode *light* est longue de 5 variables tandis que la signature Var DB-S en mode *light* inclut 7 variables (Tableau 3.6). Les deux signatures montrent un chevauchement de quatre gènes (EMP2, AGER, STX11, GYPE) dont deux sont connus pour être liés au carcinome pulmonaire.

Enfin, pour le cas d'usage de UCEC, le mode *light* DEG DB-S a abouti à une signature de 102 variables et le mode Var DB-S permet l'obtention d'une signature de 24 variables. Ces deux signatures ont un chevauchement de 16 gènes, ce qui signifie qu'un tiers de la signature Var est inclus dans celle du DEG. Les deux signatures présentent des gènes très intéressants dans l'annotation DisGeNet du carcinome endométrial (tableau 3.6).

3.6.2 Efficacité computationnelle des modes standard et *light*

Nous avons employé, dans nos recherches, deux types de mode HEFS : standard et *light*. Le mode *light* a pu formuler des signatures stables sur le cas d'utilisation du CRC au stade IV en un temps nettement inférieur à celui du mode standard, terminant la tâche en seulement 1 jour contre les 7 jours précédemment nécessaires pour l'entraînement parallèle des différents types de modèles d'enveloppement (utilisant des jobs SLURM limités à 150G de mémoire et 10 CPUs). De plus, sur le jeu de données CRC, les signatures dans les scénarios de mode *light* DEG DB-S et Var DB-S se sont révélées très similaires à celles du mode standard, démontrant l'efficacité et la fiabilité de notre approche en mode *light*. En définitive, une approche complète comme le mode standard est recommandée pour assurer la stabilité à travers un large spectre d'algorithmes. Cependant, le mode *light* pourrait se révéler plus avantageux en temps et en ressources pour calculer une signature minimale et robuste de biomarqueurs.

3.7 Synthèse : Trois points critiques de l'approche HEFS

Les approches HEFS intègrent de multiples perturbations (fonctionnelles et relatives aux données) et sont ainsi intrinsèquement sujettes à des choix aux différentes étapes de leur conception.

3.7.1 Réduction des dimensions

Tout d'abord, une première étape souvent cruciale dans les processus de sélection de variables, y compris en contexte HEFS est de réduire le nombre de dimensions en filtrant les variables les moins susceptibles de porter des informations significatives. Cela permet par la même occasion de réduire drastiquement l'ordre de grandeur du nombre de variables. Dans un contexte de transcriptomique, les utilisatrices et utilisateurs appliquent classiquement un filtre de DEG avant d'utiliser des approches ML. Cependant, il faut être conscient que le filtrage basé sur les DEG introduit un biais puisqu'il prend en compte la classe des échantillons pour réaliser son analyse,

ce qui est contre-intuitif pour l'utilisation ultérieure du ML. En effet, un processus ML nécessite l'emploi séparé d'un jeu de données d'entraînement et d'un jeu de données de test où il est supposé aveugle aux informations de classe pendant la phase d'entraînement. Une autre solution, comme exposée dans Bommert et al. [62], repose sur un filtre Var dans les données de survie d'expression génique. Ici, nous avons appliqué le filtre Var classique sans a priori pour la classe dans le contexte de l'identification de biomarqueurs de diagnostic du cancer à partir des données d'expression, en utilisant les approches HEFS.

Dans nos recherches, nous avons analysé de manière approfondie nos scénarios HEFS standard sur le jeu de données CRC. Nous observons que les modèles ML entraînés sur des ensembles de données filtrés par DEG étaient trop optimistes (score AVG MCC proche de 1 et STD MCC associé proche de 0), avec de longues signatures redondantes (24 à 38 variables dans la plupart des cas), ce qui est une indication classique de surapprentissage. En revanche, les modèles ML entraînés sur des ensembles de données basés sur un filtre Var étaient légèrement moins efficaces, mais donnaient toujours des résultats de haute qualité (AVG MCC autour de 0,8) et avec des signatures plus courtes (de 7 à 17 variables dans la plupart des cas) car le jeu de données d'entrée était davantage bruité. Les variables calculées pour les scénarios Var sur le jeu de données CRC étaient principalement trouvées dans les signatures DEG. Ce schéma a également été observé lors du calcul d'une signature de biomarqueurs pour le stade I du LUAD et le stade III du UCEC en utilisant le mode *light* HEFS.

Il est important de rappeler que dans la sélection de variables par ensemble (EFS) en général, et donc dans l'EFS hybride (HEFS) en particulier, c'est un compromis entre diversité et stabilité qui est recherché. Les signatures basées sur les DEG tendent à être très stables à travers les modèles du même type de classificateur, mais aussi entre les types de classificateurs et entre les séries d'échantillonnage. En effet, au moins 28 variables ont été trouvées d'intérêt après l'étape d'agrégation dans les scénarios standard HEFS basés sur les DEG. Les signatures basées sur les Var, en revanche, sont plus diverses tout en étant capables de s'agréger pour obtenir de courtes signatures stables. Pour le jeu de données CRC, la signature Var de quatre variables obtenue avec le mode standard HEFS a été généralisée à un autre jeu

de données CRC de stade IV avec une efficacité similaire à celle obtenue pour les signatures DEG plus longues.

3.7.2 Perturbation des données : les stratégies d'échantillonnage

Un second paramètre important à considérer dans une procédure HEFS est le choix de la stratégie de perturbation des données. Ici, nous étions intéressés par un schéma de répétition avec un échantillonnage stratifié aléatoire standard (R-S) comparé à un échantillonnage stratifié équilibré en distribution (DB-S) moins utilisé. Ce dernier est destiné à capturer et représenter une éventuelle hétérogénéité intra-classe dans le processus d'entraînement de modèles d'enveloppement de notre approche HEFS. Selon nos résultats, nous n'avons détecté aucun changement significatif entre les modèles d'enveloppement basés sur R-S et DB-S et la qualité de leurs signatures géniques associées. Cela peut suggérer que l'hétérogénéité intra-classe a été prise en compte lors des boucles de sélection de variables intervenant dans l'entraînement des modèles d'enveloppement. Cela peut aussi s'expliquer par un effet intra-classe dans nos ensembles de données trop faible pour avoir un impact. Dans tous les cas, nous estimons que l'hétérogénéité intra-classe dans le contexte de la sélection de variables pour le cancer est quelque chose à considérer et qui devrait être davantage investigué.

3.7.3 Intégration des perturbations : la méthode d'agrégation

La troisième étape que nous considérons ici est celle du processus d'agrégation à grande échelle pour tirer parti de diverses combinaisons de ré-échantillonnages de données et des signatures de modèles d'enveloppement. Que ce soit en contexte de prédiction ou de sélection de variables, il est encore d'usage d'essayer divers algorithmes de ML en optimisant les hyper-paramètres, puis de calculer une métrique de performance et de sélectionner le modèle donnant le score le plus élevé sur un échantillon standard. Cependant, les scores peuvent être extrêmement proches les uns des autres, tandis que les signatures de modèles associées peuvent être différentes. Choisir un modèle unique dans ce contexte peut conduire à la sélection de variables non robustes, gaspillant des ressources précieuses pour leur caractérisation. De plus, cela peut entraîner la négligence d'autres pistes d'investigations de

variables plus robustes, retardant ainsi des recherches cliniques importantes.

3.8 Conclusion

Les travaux de recherche présentés dans ce troisième chapitre ont permis d'évaluer les performances de l'approche HEFS proposée. Plus précisément, nous avons exploré trois étapes cruciales de conception : la réduction de dimension, la perturbation des données par ré-échantillonnage et l'intégration des perturbations des données et fonctionnelles par un processus d'agrégation. Pour étudier ces questions, nous avons mis en œuvre une stratégie multi-niveaux. Premièrement, nous avons intégré une étape particulière de réduction de dimension par l'analyse des gènes différentiellement exprimés (DEG) ou par l'analyse de la variance (Var). Ensuite, nous avons effectué deux types d'échantillonnage : un échantillonnage standard (échantillonnage aléatoire stratifié R-S) et un second visant à capturer la variabilité intra-classe (échantillonnage équilibré en distribution DB-S). Par la suite, nous avons employé un processus d'entraînement de modèles d'enveloppement à grande échelle basé sur de multiples classificateurs de pointe avec de nombreuses configurations d'hyperparamètres. Enfin, un processus d'agrégation multi-niveaux a été conçu pour exploiter toute la variabilité des données et des fonctions, et appliqué afin de définir des signatures de biomarqueurs stables. Cette approche HEFS a donné des résultats prometteurs pour le cancer colorectal de stade IV, les signatures obtenues étant robustes lorsqu'elles ont été confrontées à un jeu de données externe et en comparaison avec un processus de sélection de variables standard. De plus, l'approche a pu se généraliser en mode *light* sur le même jeu de données de CRC et sur trois autres jeux de données qui sont le cancer du rein de stade I, le cancer du poumon de stade I et le cancer de l'endomètre de stade III. En conclusion, nous recommandons notre approche HEFS pour identifier des biomarqueurs de maladies complexes, car ses résultats sont indépendants de toute méthode spécifique de sélection de variables ou de structure d'échantillonnage. Cependant, nous soulignons par nos recherches la nécessité d'apporter une attention particulière à plusieurs étapes clés.

Chapitre 4

Système de visualisation d'un atlas de biomarqueurs

4.1 Introduction

La méthode de sélection de variables que nous avons développée répond au besoin actuel d'identifier de nouveaux biomarqueurs du cancer. Nos cas d'étude présentaient des signatures spécifiques de certains types de cancers et à des stades d'avancement choisis. Dans la suite de nos recherches, nous avons appliqué notre approche de sélection de variables transcriptomiques à un grand nombre de cas de cancer et dans des configurations d'avancement diverses. Les différents types de cancer sont, depuis des décennies, analysées séparément, car les types cellulaires d'origine des tumeurs sont souvent très différents. Cependant, explorer les similarités entre divers types de cancers est une question ouverte [131–133].

Par exemple, analyser les mécanismes communs à la progression du cancer d'une manière globale est une question à laquelle différentes équipes se sont intéressées en analysant en parallèle plusieurs types de cancer (Figure 4.1). Dans ce cadre, des biomarqueurs ont été proposés en lien avec la progression de tumeurs au niveau pan-cancer [135–137].

Dans ce contexte, nous proposons de créer un atlas de signatures de biomarqueurs transcriptomiques de cancers pour faciliter la comparaison de biomarqueurs entre divers types de cancers et/ou par stades d'avancement et avec l'objectif d'identifier des signaux communs et spécifiques à plusieurs phénotypes tumoraux. Pour ce

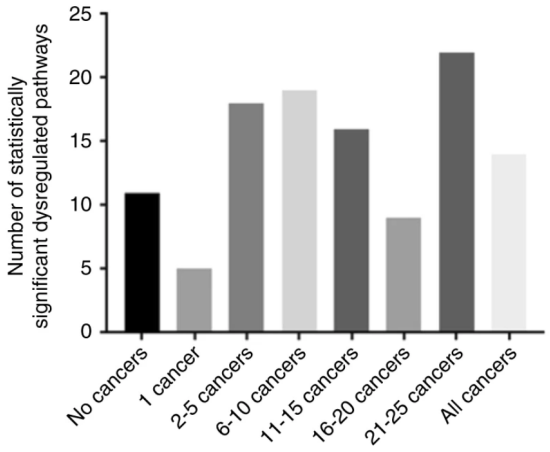

Figure 4.1: Figure extraite de [134]. Distribution du nombre de voies métaboliques (sur un total de 114 voies) KEGG dérégulées dans aucun type de cancer, dans un seul type, dans plusieurs ou dans les 26 types de cancers étudiés.

type d'analyses, nous avons exploité de multiples cohortes de patients atteints de types différents de cancer et à divers stades. Pour conserver cette hiérarchisation d'information dépendante de la cohorte (**type de cancer ⇒ stade de progression**), une analyse horizontale intégrée *a posteriori* peut être employée. Ce concept d'horizontalité se définit par une analyse se concentrant sur un seul type de données (un seul type de données omiques) [138] par opposition à une analyse verticale qui intègre des types de données variés (plusieurs omiques). L'intégration de ces données de plusieurs cohortes a été réalisée *a posteriori* sur un principe similaire à une intégration tardive de données [139]. Ce type d'intégration vise à analyser les ensembles de données séparément (chaque cohorte) pour ensuite intégrer *a posteriori* leurs résultats et proposer un résultat unique standardisé. Cette approche s'oppose à l'intégration précoce des données qui, par des procédés tels que la fusion de données (fusion de l'ensemble des données omiques en un unique tableau de données), permet d'analyser conjointement des ensembles de données différents avant de produire un résultat final. L'emploi d'analyses basées sur le concept d'une intégration horizontale *a posteriori* permet ainsi la production de résultats sur des données de types similaires tout en conservant l'information préliminaire spécifique à chaque ensemble de données (chaque cohorte). Dans le cas présent, il est essentiel de pouvoir conserver l'information d'appartenance de chaque signature de biomarqueurs à un type de cancer et stade de progression. Pour prendre en compte cet objectif, nous avons opté pour le développement d'une méthode de visualisation d'informations

interactive pour réaliser une analyse intégrée des signatures de biomarqueurs issus de plusieurs cohortes selon la hiérarchisation d'information (type de cancer ⇒ stade de progression). En effet, l'exploration d'un atlas composé par l'ensemble des signatures est une tâche difficile du fait de son hétérogénéité intrinsèque. Nous proposons ainsi, dans ce chapitre, une preuve de concept d'un système de visualisation interactif adapté aux problématiques de l'atlas de biomarqueurs et mis à disposition de la communauté scientifique à travers une application web.

Dans la suite de ce chapitre, nous exposerons dans un premier temps les problématiques posées par la visualisation d'un atlas de biomarqueurs suivi par la définition du système de visualisation conceptualisé pour prendre en compte ces défis. Dans un deuxième temps, nous présenterons l'interfaçage réalisé pour intégrer notre solution visuelle à une application web intuitive et publique. Dans une troisième partie, nous exposerons les éléments de matériel et méthodes employées pour construire l'atlas à explorer. Enfin, nous conclurons.

J'ai analysé les données et produit les résultats qui composent l'atlas et j'ai conduit les réflexions de conceptualisation du système de visualisation en collaboration avec Frédéric Lalanne, ingénieur dans l'équipe BKB du LaBRI. La programmation de l'application a été réalisée par ce dernier.

4.2 Conception d'un système de visualisation pour exploiter l'atlas de biomarqueurs

4.2.1 Vue d'ensemble du système proposé

Un atlas de biomarqueurs est composé par un ensemble de biomarqueurs, organisés en signatures et attribués à des phénotypes correspondant aux données de la cohorte à partir de laquelle ils ont été prédits [140].

Il s'agit ainsi de visualiser des données relationnelles où chaque biomarqueur est associé à l'information hiérarchisée : **type de cancer ⇒ stade de progression**. Nous avons modélisé ces relations par des diagrammes noeuds/liens tels que des graphes. La structuration des informations décrivant les biomarqueurs (informations hiérarchisées) est le point central pour guider l'analyse exploratoire des données

en fonction d'une question biologique de départ. Dans ce contexte, pour définir les différentes métaphores visuelles et outils d'interaction du système de visualisation, nous nous sommes appuyés sur le mantra de Shneiderman [110] : ***Overview First, Zoom and Filter, Details on Demand*** (voir 1.7). Ainsi, pour guider une utilisatrice ou un utilisateur dans l'exploration de l'atlas de biomarqueurs, nous sommes partis des questions biologiques pour définir les différentes tâches d'exploration et métaphores visuelles nécessaires à leur réalisation (Tableau 4.1). Concrètement, pour définir le système de visualisation, nous nous sommes basés sur les trois critères essentiels définis par Munzner : les données, les tâches à accomplir et l'encodage ou la métaphore visuelle (voir 1.7).

Les questions biologiques identifiées (Tableau 4.1) coïncident avec plusieurs niveaux d'analyse et ont abouti à la définition de plusieurs métaphores visuelles présentées succinctement:

1. Le premier niveau d'analyse est appelé une vue globale où l'ensemble des signatures (sans détailler les biomarqueurs) pour tous types de **cancer** $\Rightarrow$ **stade de progression** sont représentées. Cette vue globale est le point de départ et compare les biomarqueurs communs ou spécifiques selon l'information relative à un cancer. L'utilisatrice ou l'utilisateur peut ensuite affiner sa recherche par l'utilisation de filtres pour analyser plus spécifiquement des biomarqueurs en fonction des informations.

2. Après application d'un filtre, un focus peut être réalisé pour visualiser une vue multi- ou mono-signature. Une seconde métaphore visuelle détaille les biomarqueurs en donnant davantage de précisions grâce à l'utilisation d'outils d'interaction.

3. Le dernier niveau d'analyse exploite des *box-plot* ou boîtes à moustaches. Cette métaphore visuelle résume pour chaque biomarqueur le niveau d'expression dans la cohorte ciblée.

Question biologique	**Tâches à réaliser**	**Métaphore visuelle**
Analyser l'atlas dans sa globalité	**Vue globale** de l'atlas basée sur un graphe non orienté	Emploi d'un diagramme noeud/liens : - noeuds : signatures - liens : gène(s) en commun entre deux signatures
Décomposer l'atlas	**Vue multi-signature** basée sur un graphe non orienté	Emploi d'un diagramme noeuds/liens: - noeuds : gènes ou annotations - liens : annotation impliquant au moins deux gènes appartenant à au moins deux signatures différentes
	Vue mono-signature basée sur un graphe non orienté	Emploi d'un diagramme noeuds/liens : - noeuds : gènes ou annotations - liens : annotation impliquant au moins un gène de la signature explorée
Obtenir des détails quantitatifs sur des gènes d'intérêt	**Vue d'expression génique**	Emploi de graphiques par boîtes à moustaches

Tableau 4.1: Modèle de Munzner pour la conception de système de visualisation: l'abstraction des tâches et des données en quatre vues distinctes ainsi que les métaphores visuelles proposées.

4.2.2 Définition des tâches du système de visualisation

Vue globale

La vue globale s'appuie sur la modélisation des données par un graphe non orienté (Figure 4.2). Chaque noeud du graphe est associé à une signature de biomarqueurs (groupe de gènes) et les liens entre les noeuds illustrent la présence de biomarqueurs en commun entre deux signatures.

Chaque noeud est associé à des attributs qui sont visualisés avec une étiquette comme le type de cancer et le stade d'avancement. L'épaisseur des liens est proportionnelle aux nombres de gènes en commun entre les deux signatures correspondantes.

Dans cette vue globale, l'utilisatrice ou utilisateur peut également accéder aux détails de la signature de biomarqueurs. Par exemple, par simple clic sur le noeud sig-

nature d'intérêt, nous pouvons visualiser un lien pérenne vers une analyse d'enrichissement d'annotation par ORA réalisée via l'outil g:profileR (voir section 1.6.2). L'information des biomarqueurs communs aux signature est également accessible au niveau des liens.

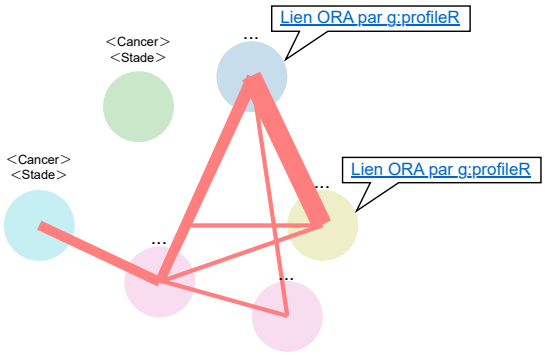

Figure 4.2: Schéma de la visualisation choisie pour la vue globale.

Vue multi-signature

Le niveau de visualisation suivant permet d'analyser en détails la comparaison de plusieurs signatures. La sélection de ces signatures s'opère à partir de filtres qui visent à guider l'analyse approfondie de l'atlas. La métaphore visuelle pour cette vue multi-signature est basée sur un graphe non orienté composé de deux types de noeuds : les gènes et les annotations biologiques. Les noeuds de type gènes sont représentés par des cercles et les noeuds d'annotations par des triangles (Figure 4.3). Un seul type de lien est utilisé et permet d'obtenir l'information "ce gène et cette annotation sont liés". Ce lien est de couleur noire et de taille fixe.

L'appartenance de gènes à des signatures communes pourrait être encodée sur la représentation en cercle des noeuds de gènes directement (couleur, texture) ou par le biais de liens entre eux. Néanmoins, la vue multi-signature peut intégrer un grand nombre de signatures et donc de gènes et la visualisation s'en retrouverait encombrée. Nous avons fait le choix de tirer parti de l'espace entre les noeuds pour visualiser les informations supplémentaires comme le type de cancer et le stade de développement. Ainsi les noeuds de gènes sont entourées d'une ou plusieurs enveloppes correspondant à la ou les signature(s) dont ils font partie (Figure 4.3). Chaque enveloppe est calculée en utilisant un algorithme basé sur le principe des diagrammes Euler [141]. De 2 à n enveloppes sont alors visualisables simultanément

dans la vue multi-signature où n est le nombre de signatures totales incluses dans l'atlas.

Ces enveloppes ont trois caractéristiques visuelles : la couleur qui est dépendante du type de cancer, la texture (sous forme de rayures) pour le stade de cancer et le périmètre de l'enveloppe pour englober les noeuds correspondant aux gènes la composant. Ces trois caractéristiques peuvent se superposer si un ou plusieurs gènes sont en commun entre au moins deux signatures différentes. Les cercles représentant les noeuds de gènes sont généralement de couleur noire et d'une taille fixe, sauf en cas d'appartenance à deux signatures. En effet, pour appuyer l'effet visuel de superposition amené par les enveloppes, les noeuds de gènes chevauchants sont colorés en vert et leur taille est augmentée proportionnellement au nombre de signatures dont ils font partie.

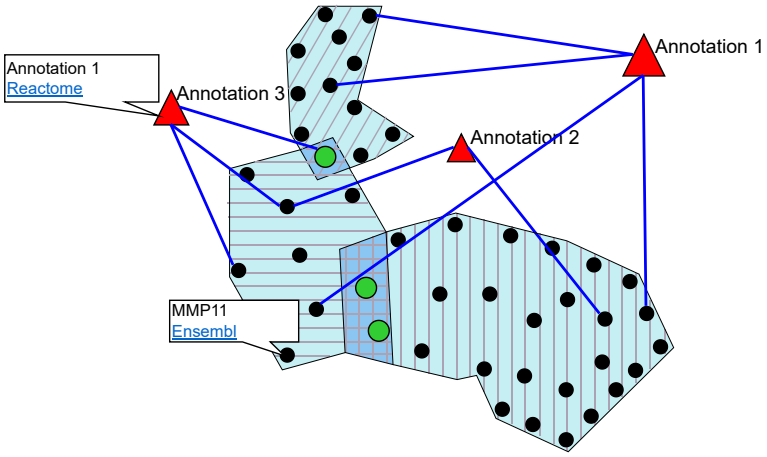

Figure 4.3: Schéma de la visualisation choisie pour la vue multi-signature.

Pour éviter un encombrement de la visualisation et faciliter l'analyse des similarités entre signatures et donc entre phénotypes, les noeuds d'annotation sont affichés uniquement si leur degré (nombre de liens) est d'au moins deux et si leurs noeuds voisins (de type gène), appartiennent à des signatures différentes. Les noeuds d'annotations conservés sont colorés en rouge et leur taille est proportionnelle au nombre de noeuds de gènes auxquels ils sont reliés. Les liens sont de couleur bleue et leur épaisseur est fixe.

Les noeuds et liens qui composent la vue multi-signature ont des attributs que l'utilisateur peut visualiser à la demande en positionnant la souris dessus ou en cliquant. Par exemple, en cliquant sur un noeud modélisant les annotations,

l'utilisateur est directement dirigé vers la page web dédiée à cette annotation dans la base de données de référence. Pour les noeuds de type gènes, ils sont dirigés vers la base de données Ensembl (une base de données génomique de référence notamment pour l'annotation de gènes) [142].

Vue mono-signature

La vue mono-signature partage de multiples caractéristiques avec la vue multi-signature. Elle est en effet basée sur un graphe non orienté intégrant les deux mêmes types de noeuds : des noeuds de gènes et des noeuds d'annotation. Les noeuds de gènes sont à nouveau visualisés par des cercles, mais avec une couleur unique (noir). Les noeuds d'annotation sont à nouveau visualisés par une forme de triangle à la couleur fixe, le gris (Figure 4.4). Contrairement à la vue multi-signature, la vue mono-signature propose d'obtenir davantage de détails sur les biomarqueurs qui la compose. L'enveloppe unique conserve les trois caractéristiques précédemment citées. En revanche, tous les noeuds d'annotation sont représentés quel que soit leur degré.

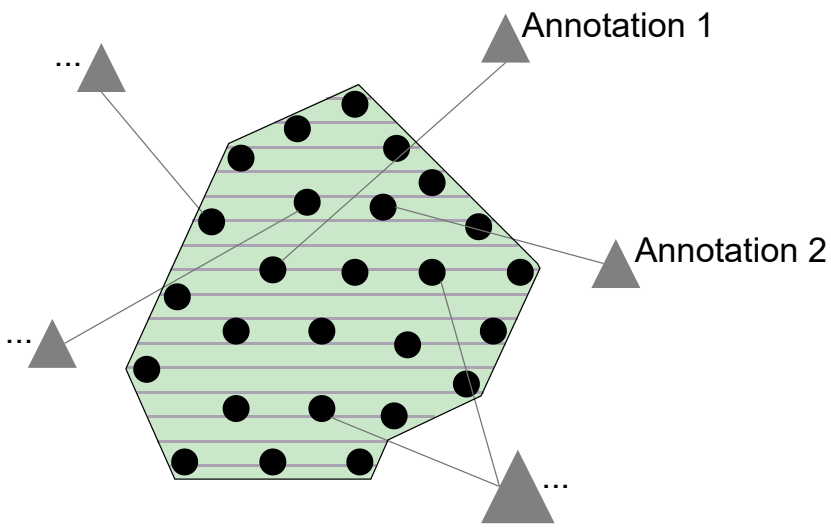

Figure 4.4: Schéma de la visualisation choisie pour la vue mono-signature.

Vue d'expression génique

Cette vue détaillée permet à l'utilisateur d'obtenir des informations détaillées sur les données d'expression génique de gènes sélectionnés pour des conditions spécifiées

(type de cancer et stade d'avancement). Pour un gène d'intérêt donné, une zone rectangulaire lui est dédiée afin d'y représenter un graphique de boîtes à moustaches. L'axe des ordonnées correspond à l'expression génique et l'axe des abscisses définit le type de cancer et stade d'avancement sélectionnés (Figure 4.5). Les boîtes à moustaches permettent de visualiser l'expression du gène d'intérêt parmi tous les échantillons des cohortes utilisées pour formuler les signatures de l'atlas. Cette visualisation offre l'avantage de pouvoir observer les différences de distribution entre diverses conditions en abscisses. Une évaluation statistique de ces différences par paire est par ailleurs réalisée et les différences significatives (test T, seuil de valeur p de 0.05) sont visualisables.

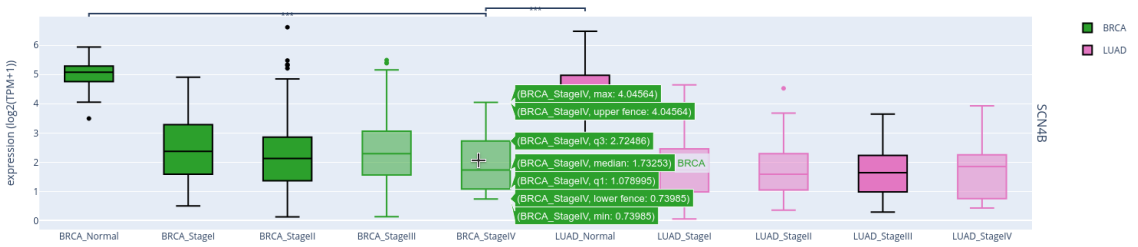

Figure 4.5: Vue d'expression génique basée sur des graphiques de boîtes à moustaches.

Les vues étant inter-connectées, les boîtes à moustaches pour des signatures du type et stade de cancer visualisés dans les différentes vues sont spécifiquement entourées de noir.

Plusieurs zones rectangulaires, chacune correspondant à l'analyse d'expression d'un gène, peuvent être empilées avec le même axe des abscisses pour faciliter la comparaison visuelle d'expression génique de gènes différents pour des phénotypes communs.

4.3 Interfaçage du système de visualisation : THe Biom

Pour mettre en oeuvre une visualisation interactive des vues précédemment décrites, celles-ci ont été intégrées dans une interface déployée dans une application web nommée **THe Biom** pour *TCGA HEFS Biomakers*. Un dépôt GitLab de la version de développement de l'application est disponible : THe Biom.

4.3.1 Aperçu de THe Biom

La figure 4.6 présente un aperçu de la version actuelle de la page principale de notre preuve de concept **THe Biom** et de l'atlas qu'elle donne à explorer au travers des différentes vues conceptualisées.

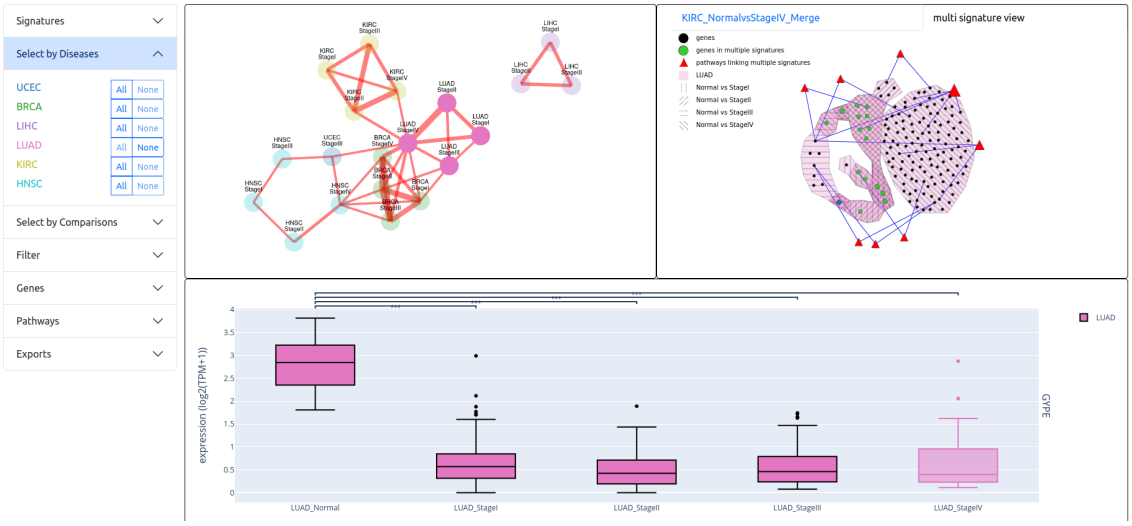

Figure 4.6: Aperçu de la page principale de l'application **THe Biom**.

L'application est destinée à des bioinformaticiennes et bioinformaticiens mais aussi à des biologistes menant des recherches sur le cancer. L'exploration de **THe Biom** et de son atlas doit être intuitive, visuelle et interactive. Nous avons opté pour une interface centrée sur une page principale découpée en plusieurs panneaux et une page secondaire de documentation (Figure 4.7).

Le premier panneau avec lequel l'utilisatrice ou l'utilisateur doit interagir est situé sur la gauche [143]. Il intègre un menu permettant de filtrer les résultats visualisés dans les autres panneaux. Nous pouvons ainsi choisir d'explorer les résultats de signatures de biomarqueurs d'un ou plusieurs types de cancer ou d'un ou plusieurs stades de développement. Si la méthodologie de sélection de variables employée l'intéresse, l'utilisatrice ou utilisateur peut aussi filtrer les résultats en fonction de ce paramètre. Le menu permet aussi de rechercher un gène ou une annotation biologique d'intérêt par une entrée texte à auto-complétion ou une liste déroulante. Enfin, la dernière fonctionnalité du menu concerne l'exportation des données d'exploration en sélectionnant la vue d'intérêt et le format (png ou json).

Une page secondaire de documentation est accessible par un bouton situé dans la partie supérieure droite de la page principale. À terme, cette page contiendra un

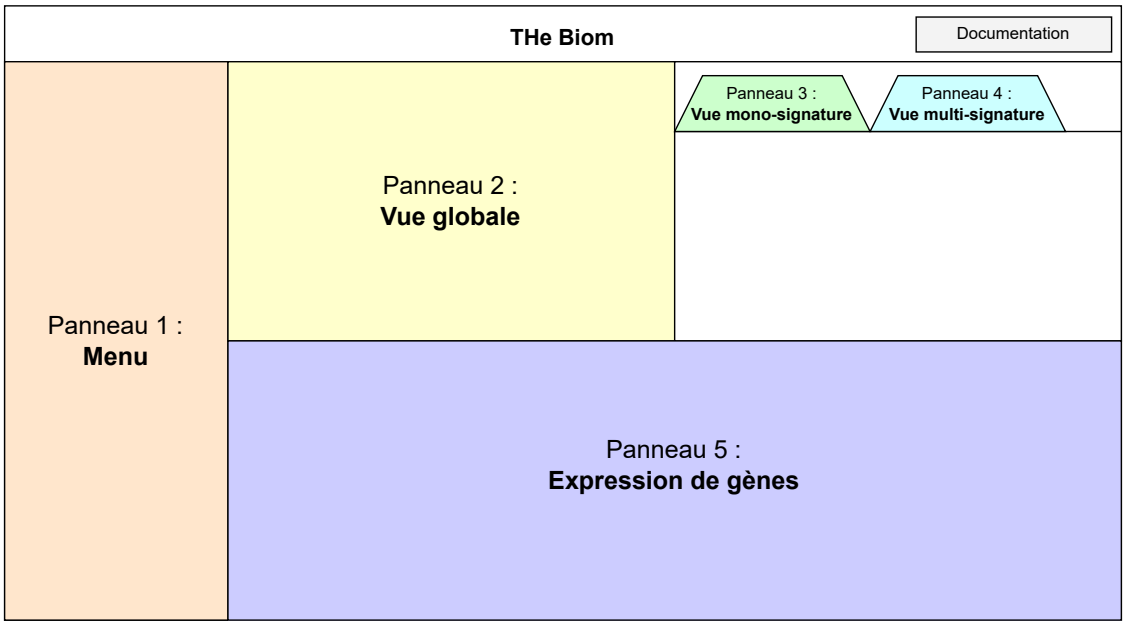

Figure 4.7: Maquette de **THe Biom** avec sa page principale et ses cinq panneaux de menu et visualisation ainsi que la page secondaire de documentation accessible par le biais d'un bouton.

tutoriel d'utilisation, une "Foire Aux Questions", un lien vers l'article scientifique qui sera associé à la mise en ligne de l'application ainsi qu'un lien vers l'article scientifique déjà publié qui décrit l'approche de sélection de variables employée pour définir les signatures de biomarqueurs de l'atlas [144].

4.3.2 Architecture de l'application

L'application web est organisée en deux modules : le Front End et le Back End (Figure 4.8). Cette architecture repose sur le framework dash en Python. La partie Front End, qui est la partie visible par l'utilisateur, intègre des éléments de HTML, javascript et CSS (bootstrap, cytoscape, plotly). Le Back End est la partie non visible par l'utilisateur. Il gère les pré-traitements des données et traite les requêtes envoyées par le Front End, en y répondant de manière appropriée. Dans **THe Biom**, le Back End s'appuie sur des fichiers générés par un traitement en amont et externe à **THe Biom** :

- Un fichier de signatures : toutes les listes de gènes associés à chaque signature et un lien internet vers une analyse ORA par g:Profiler pour chaque signature ;

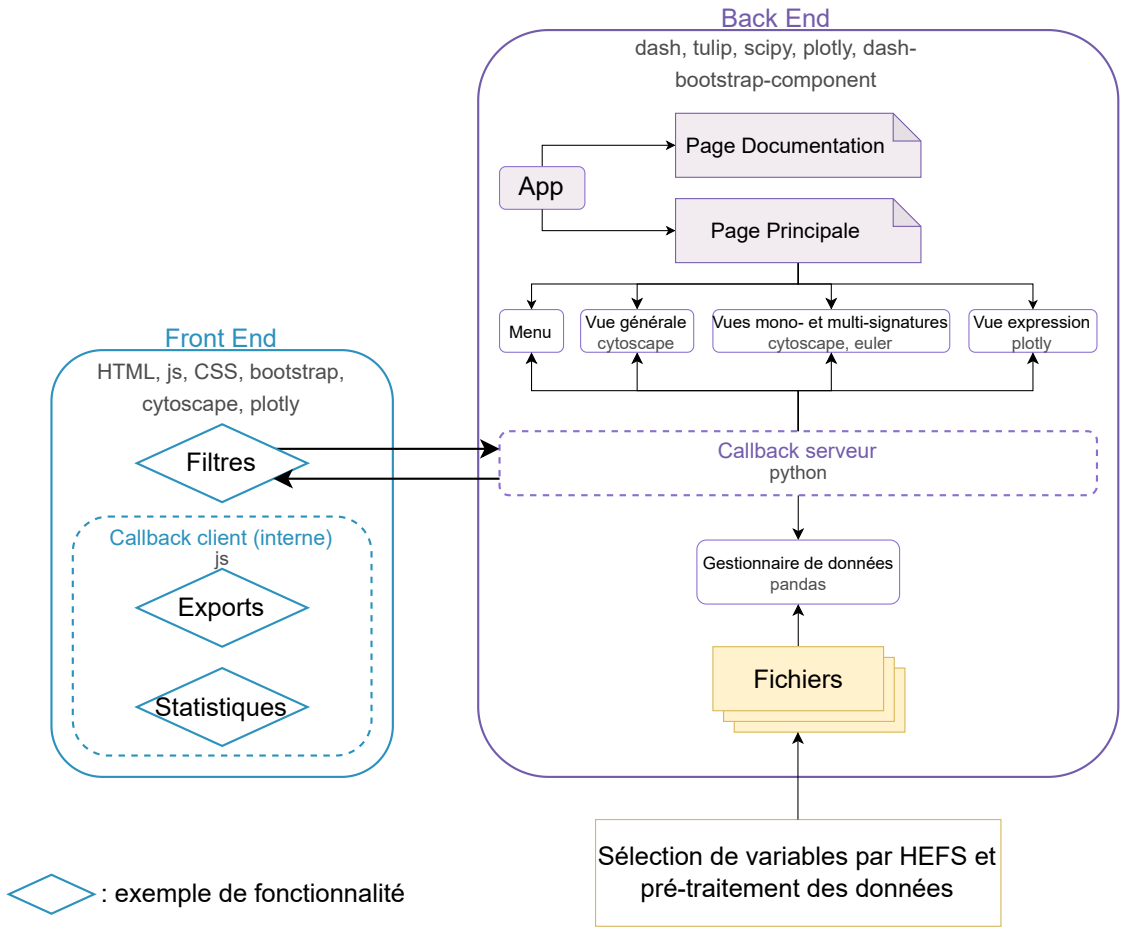

Figure 4.8: Architecture de l'application web **THe Biom** et ses composantes Front End et Back End.

- Un fichier d'annotations biologiques : liste de tous les gènes uniques présents dans les signatures et les noms et identifiants de toute annotation biologique leur étant associée ;
- Un fichier d'expression : valeur d'expression en $log2(TPM+1)$ de chaque gène appartenant à une signature.

Ces fichiers sont convertis en *dataframes* pandas et pris en charge par un gestionnaire de données. Ces données servent à un module de gestion des requêtes (ou *callback*) serveur qui adapte les réponses et modifie les différentes vues présentes notamment dans la page principale. Les vues globale et multi- et mono-signature utilisent la bibliothèque *cytoscape* pour visualiser la représentation des graphes. La vue d'expression génique emploie quant à elle la bibliothèque *plotly* pour générer les visualisation par boîtes à moustaches dans la vue d'expression génique.

4.4 Création de l'atlas

4.4.1 Analyses intégrées a posteriori

Comme mentionné dans l'introduction du présent chapitre (section 4.1), pour conserver la hiérarchisation des données, nous employons des analyses intégrées *a posteriori* grâce à la mise en oeuvre d'un système de visualisation. La Figure 4.9 décrit les étapes d'analyses et d'intégration des résultats et de données d'annotation à l'application **THe Biom**. L'encadré A de la Figure 4.9 expose le type de données utilisées et les méthodes de sélection de variables employées (voir section suivante 4.4.2 et 4.4.3). Les signatures obtenues à la suite de la sélection de variables sont alors intégrées à **THe Biom** pour former l'atlas. L'encadré B de la Figure 4.9 définit les ressources mobilisées pour intégrer des informations d'annotations à **THe Biom** (voir section suivante 4.4.4). Enfin, l'encadré C de la Figure 4.9 est un aperçu de l'application web **THe Biom** et des quatre vues intégrées à sa page principale (voir sections précédentes 4.2 et 4.3).

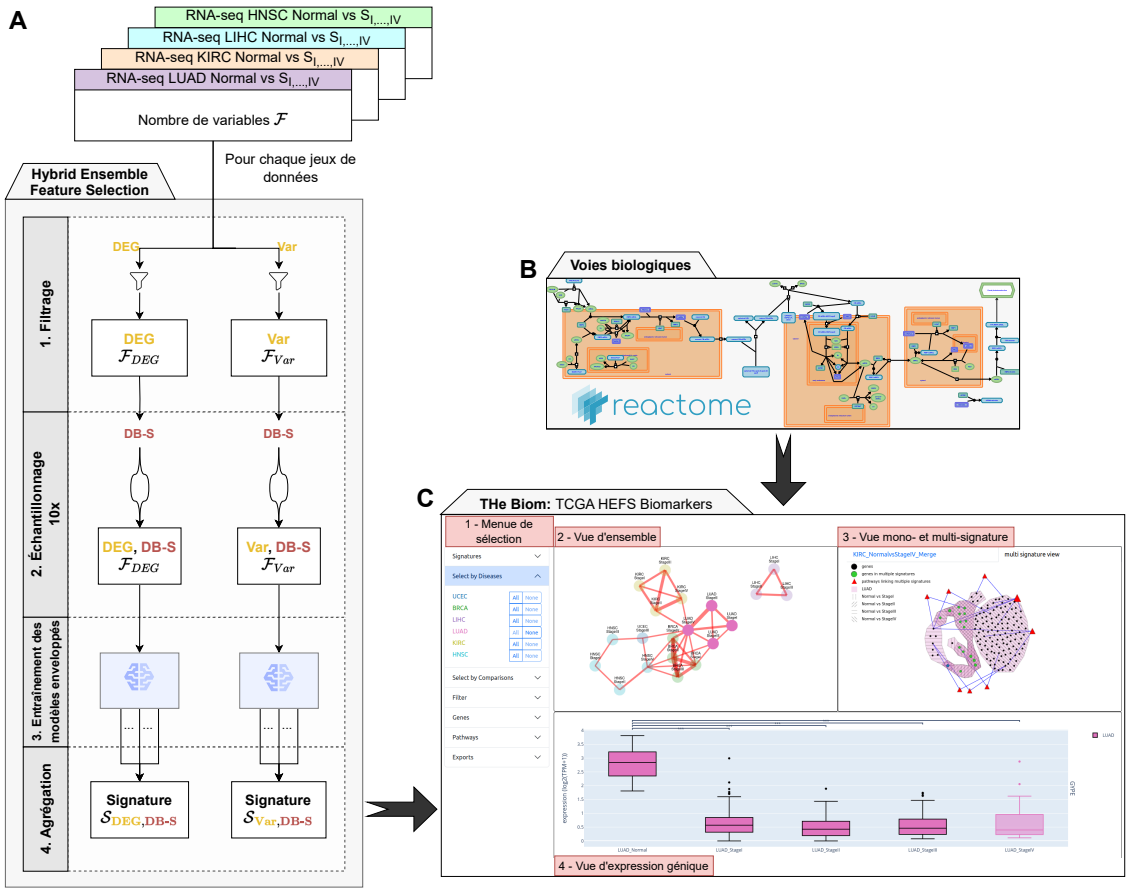

Figure 4.9: Vue d'ensemble des analyses intégrées *a posteriori* et des données d'annotations agrégées à l'application **THe Biom**. (A) Définition des signatures robustes de biomarqueurs de multiples comparaisons de phénotypes oncologiques par notre approche hybride ensembliste de sélection de variables en mode *light*. (B) Mobilisation de la base de données Reactome pour l'obtention des données d'annotations de voies biologiques reliées aux signatures. (C) Intégration de ces données à **THe Biom** (**T**CGA **H**EFS **Biom**arkers). Aperçu de la page principale de l'application, du panneau de menu et des quatre vues d'exploration.

4.4.2 Jeux de données

La création de l'atlas s'est appuyée sur l'exploitation de données publiques de transcriptomiques de cancers. Nous avons tiré parti des cohortes de patients du consortium TCGA. Pour notre preuve de concept, six cohortes ont été sélectionnées et concernent des cancers de l'utérus (UCEC), du sein (BRCA), du foie (LIHC), des poumons (LUAD), des reins (KIRC) et de la tête et du cou (HNSC) (Tableau 4.2). Des échantillons Normaux et de divers stades d'avancement I à IV (tels que déterminés par l'UICC et AJCC, voir section 1.2.2) ont été analysés. À noter que pour la cohorte LIHC, qui concerne un type de cancer du foie, les échantillons de stade

Type de cancer	Normal	Stade I	Stade II	Stade III	Stade IV
UCEC	19	310	48	108	28
BRCA	105	170	574	227	19
LIHC	48	157	75	79	-
LUAD	45	186	81	58	18
KIRC	69	262	57	122	83
HNSC	42	24	62	71	228

Tableau 4.2: Nombre d'échantillons normaux et de stade I à IV pour les différents types de cancer dont les données TCGA sont analysées dans le projet **THe Biom**.

IV n'ont pas été considérés dans le processus de sélection de variables en raison de leur faible quantité (seulement quatre échantillons).

Le traitement préliminaire des données a été réalisé en utilisant le pipeline présenté dans le chapitre précédent (Section 3.2). Les mêmes filtres ont été ré-appliqués : minimum de 70% de noyaux tumoraux dans les échantillons cancéreux, absence d'échantillons FFPE et pas de données manquantes dans les données cliniques permettant l'identification du stade d'avancement.

Ces données sont utilisées pour définir des signatures robustes de biomarqueurs par un processus de sélection de variables. À l'issue de ce dernier, les données d'expression en transcripts par million (TPM) de tous les gènes faisant partie d'au moins une signature sont extraits depuis les données publiques de TCGA. Ces données d'expression sont ensuite transformées en $log2(TPM + 1)$ pour permettre leur analyse dans le système de visualisation de **THe Biom**.

4.4.3 Mode *light* HEFS

La définition de signatures robustes de biomarqueurs a été réalisée à partir de l'approche HEFS mode *light*. Ces dernières étaient initialement centrées sur l'identification de biomarqueurs spécifiques des quatre principaux stades d'avancement de cancer par rapport au phénotype normal du tissu originel ou de biomarqueurs de progression entre des stades d'avancement différents.

Les signatures pour des comparaisons d'échantillons à des stades d'avancements différents n'ont finalement pas été inclus dans l'atlas. En effet, une analyse préalable des données sur quatre types de cancer (LUAD, HNSC, KIRC, BLCA) n'a pas permis d'obtenir de signatures pour ce type de comparaisons. Elles n'ont donc pas été davantage explorées. Ainsi, seules les comparaisons d'échantillons Normaux

contre des échantillons d'un stade spécifique ont été effectuées.

L'application de cette approche HEFS permet l'obtention de signatures robustes de biomarqueurs provenant de deux types de scénarios incluant un filtre par DEG ou par Var. Comme exposé dans le chapitre précédent, ces deux méthodes présentent chacune des avantages et inconvénients. Il est ainsi pertinent de pouvoir analyser conjointement leurs résultats. Les signatures sont disponibles à l'analyse dans leur version DEG, Var ou *merge* où *merge* est une union des signatures DEG et Var pour un même phénotype.

4.4.4 Données d'annotation de voies biologiques

Avec l'objectif de caractériser les signatures et faciliter l'exploration de l'atlas, nous avons agrégé des données d'annotations et spécifiquement de voies biologiques à **THe Biom**. Nous avons choisi d'utiliser la base de données Reactome dont l'exportation des données est facilitée notamment par son API Rest. Les identifiants Ensembl de tous les gènes ont été mis en correspondance avec les voies biologiques les incluant sans règle de topologie particulière.

4.5 Conclusion

L'application **THe Biom** propose des fonctionnalités et métaphores visuelles dédiées à l'exploration d'un atlas de signatures de biomarqueurs transcriptomiques de cancers. Le système de visualisation a l'avantage de permettre des comparaisons à deux niveaux interconnectés : au niveau du type de cancer et du stade d'avancement. Pour cela, **THe Biom** s'est basé sur le mantra de Shneiderman pour offrir différents niveaux d'analyse à l'utilisateur tout en s'appuyant sur le Modèle de Munzner au travers des quatre vues : la vue globale, multi-signature, mono-signature et la vue expression.

Pour étudier les gènes qui composent les signatures et leurs interactions, nous nous sommes appuyés sur des modélisations par graphe et des attributs de visualisation personnalisés. Ces analyses qualitatives sont supportées par des méthodes de confrontations à des bases de données d'annotations biologiques comme par l'outil g:Profiler qui réalise une analyse d'enrichissement d'annotations sur les signatures

de l'atlas ou par présence de gènes dans une voie biologique Reactome.

L'atlas est élaboré à partir de données transcriptomiques RNA-seq d'un total de six cohortes de cancers de type différent et pour quatre stades d'avancement via des comparaisons de données Normal contre Stade I à IV. Nous avons employé une approche d'HEFS en mode *light* pour déterminer des signatures de biomarqueurs robustes pour chaque phénotype cancéreux.

En définitive, les possibilités d'exploration multi-niveaux de l'atlas permises par **THe Biom** peuvent répondre à des problématiques qui suscitent toujours un grand intérêt dans la communauté de recherche biomédicale. En effet, dans la continuité des analyses de phénotypes de cancers spécifiques ou de pan-cancer, plusieurs problématiques biologiques d'intérêt peuvent être envisagées : existe-t-il des biomarqueurs transcriptomiques et des voies biologiques spécifiques et communes à un type ou stade de cancer ?

Chapitre 5

Analyses réalisées avec THe Biom

5.1 Introduction

THe Biom, combinant un atlas de signatures de cancer avec un système de visualisation interactive, vise à adresser des problématiques biologiques d'ordre pan-cancer et/ou mono-cancer. Pour illustrer son intérêt, nous avons identifié trois analyses biologiques à réaliser :

1. Identification et analyse d'une liste de biomarqueurs transcriptomiques spécifiques d'un type de cancer à un stade donné ?

2. Identification et analyse d'une liste de biomarqueurs transcriptomiques et de voies biologiques spécifiques d'un cancer indépendamment du stade d'avancement ?

3. Identification et analyse d'une liste de biomarqueurs transcriptomiques et de voies biologiques spécifiques d'un stade d'avancement de cancer indépendamment du type de cancer ?

Le chapitre est constitué de trois sections présentant des clefs de résolution de ces trois problématiques accompagnées à chaque fois d'un cas d'usage. À la suite de ces sections, une autre est dédiée à l'exploitation de THe Biom pour explorer le comportement d'un gène ou d'une voie biologique d'intérêt. Le chapitre est clos par une section de conclusion.

5.2 Identification et analyse de biomarqueurs spécifiques d'un phénotype

5.2.1 Signatures intégrées dans l'atlas, issues de 6 cancers

Notre preuve de concept intègre 37 signatures de biomarqueurs obtenues après l'analyse des données RNA-seq de six cohortes de cancer TCGA par notre méthode HEFS en deux scénarios (DEG et Var) à travers quatre comparaisons d'échantillons Normaux contre ceux de stade I, II, III ou IV (Figure 5.1).

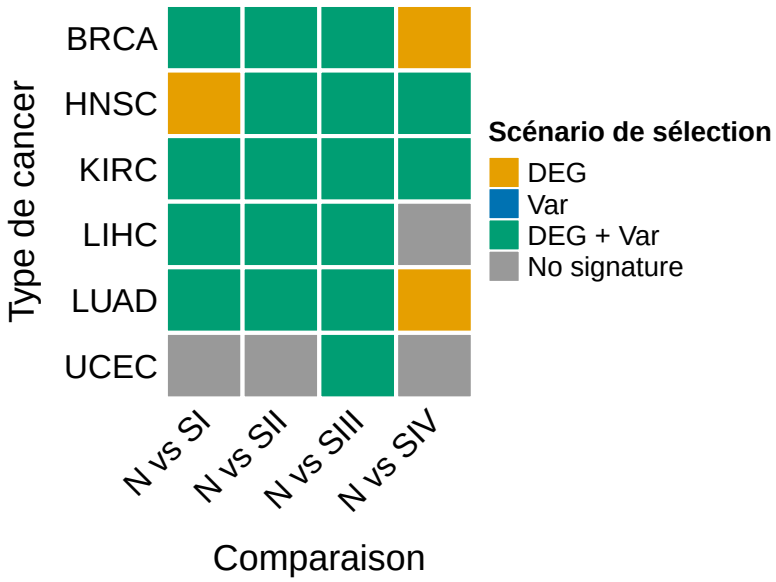

Figure 5.1: Signatures disponibles par type de cancer (BRCA : sein , HNSC : tête et cou, KIRC : rein, LIHC : foie, LUAD : poumons, UCEC : utérus), comparaisons (N vs Sx désigne une comparaison d'échantillons N normaux contre des écantillons de stade S avec x allant de I à IV) et type de scénarios d'approche HEFS (par DEG ou Var).

Pour chaque cancer, toutes les comparaisons de stade d'avancement ont été réalisées pour obtenir des signatures. La comparaison de classe Normale contre Stade III a donné des signatures par scénarios DEG et Var pour les six types de cancer. En revanche, avec le cancer de l'utérus (UCEC) notre méthode n'a pas pu obtenir de signatures robustes exceptée pour le stade III. D'une manière générale, on observe que la méthode DEG a facilité l'obtention d'une signature par rapport à la méthode Var (BRCA, LUAD et HNSC). Cela peut s'expliquer par le caractère moins bruité du filtre DEG appliqué dans les scénarios DEG comparé au scénario par Var. Pour certaines comparaisons de stade, des signatures n'ont pas pu dépasser les seuils de

filtre de la méthode HEFS *light*.

Pour chaque comparaison disposant à la fois d'une signature DEG et Var, une signature étendue dite *merge* est formulée par union des deux. Ainsi en plus des 37 signatures DEG ou Var, nous disposons de 17 signatures *merge*. Ces 54 signatures ont été intégrées dans l'atlas de **THe Biom**.

Pour la suite de ce chapitre, nous nous sommes focalisés sur les signatures de chaque stade à partir de la comparaison avec les échantillons dits normaux. L'objectif ici a été d'analyser les spécificités de chaque stade en fonction d'une référence commune neutre (Normal). Aussi, par extension, une signature de stade II d'un cancer correspond à la signature de biomarqueurs de la comparaison d'échantillons Normaux à des échantillons de Stade II.

5.2.2 Cas d'usage du cancer du foie de stade II

La cohorte LIHC de TCGA concerne l'étude du cancer du foie et plus précisément du carcinome hépatocellulaire du foie. Les cancers du foie sont la quatrième cause mondiale de décès par cancer et le carcinome hépatocellulaire représente 90% des cas de cancer du foie [145, 146]. De multiples facteurs de risques ont été identifiés, mais la compréhension de ce cancer demeure encore incomplète et davantage d'hypothèses de travail doivent être fournies pour investiguer les mécanismes de cette maladie. Nous proposons ici une liste de biomarqueurs possibles du stade II de ce cancer, à un stade de développement encore précoce.

Une analyse exploratoire réalisée dans **THe Biom**, permet d'obtenir une signature *merge* issue de la comparaison Normal contre Stade II du cancer du foie. Onze gènes codant pour des protéines sont ainsi identifiés (figure 5.2) : HOXD10, CLEC4M, COLEC10, ISLR, GDF2, BMPER, BMP10, CYP4F22, ANGPTL6, VIPR1, RSPO3.

Une analyse ORA (lien g:Profiler fourni par notre application), réalisée à partir de ces onze gènes, illustre son intérêt biologique (figure 5.3) (paramètres : correction du multi-test par la méthode de Bonferroni, taille minimale d'ensemble de gènes d'annotation de 5 et taille maximale de 5 000). Nous avons complété cette analyse par une analyse ORA d'annotations de maladies par le biais de l'outil en ligne ToppGene (Figure 5.2). ToppGene a en effet l'avantage, comparé à g:profileR, de

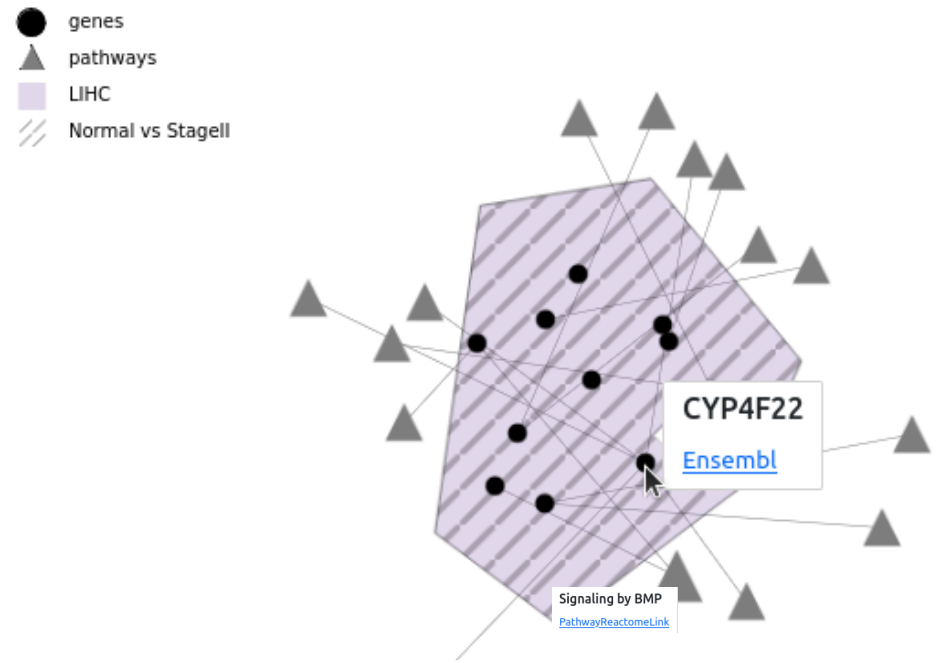

Figure 5.2: Vue mono-signature de THe Biom après sélection de la signature *merge* du cancer LIHC avec une comparaison Normal contre Stade II.

proposer des annotations de la base de données de maladies, DisGeNET. Cette analyse ORA complète présente trois types d'annotation d'intérêt :

- Informations obtenues grâce à la **Gene Ontology GO** : La régulation de la transduction du signal de la protéine SMAD ainsi que BMP sont des annotations visibles de multiples fois à différents niveaux d'annotation dans notre analyse ORA. Il s'agit de voies de régulation importantes dans la carcinogenèse et qui ont été reconnues comme ayant des interactions fortes, notamment dans le cadre du carcinome hépatocellulaire [147–149].

- Informations obtenues grâce à **DisGeNet** : Notre analyse montre également un enrichissement significatif du terme d'annotation *Liver carcinoma* et conforte la pertinence de la signature pour caractériser le carcinome du foie.

- Informations obtenues grâce à **Reactome** : Une voie biologique *Signaling by BMP* provenant de la base de données Reactome a été enrichi significativement.

Concernant les annotations Reactome, notre application **THe Biom** permet aussi d'observer l'annotation *Signaling by BMP* et les gènes qui y sont directement liés (Figure 5.2). Elle permet aussi d'investiguer d'autres voies biologiques de la base

de données Reactome qui ne sont pas nécessairement significativement enrichies. Enfin, nous pouvons identifier des gènes hautement annotés, comme le gène CYP4F22 qui est connecté à cinq voies biologiques.

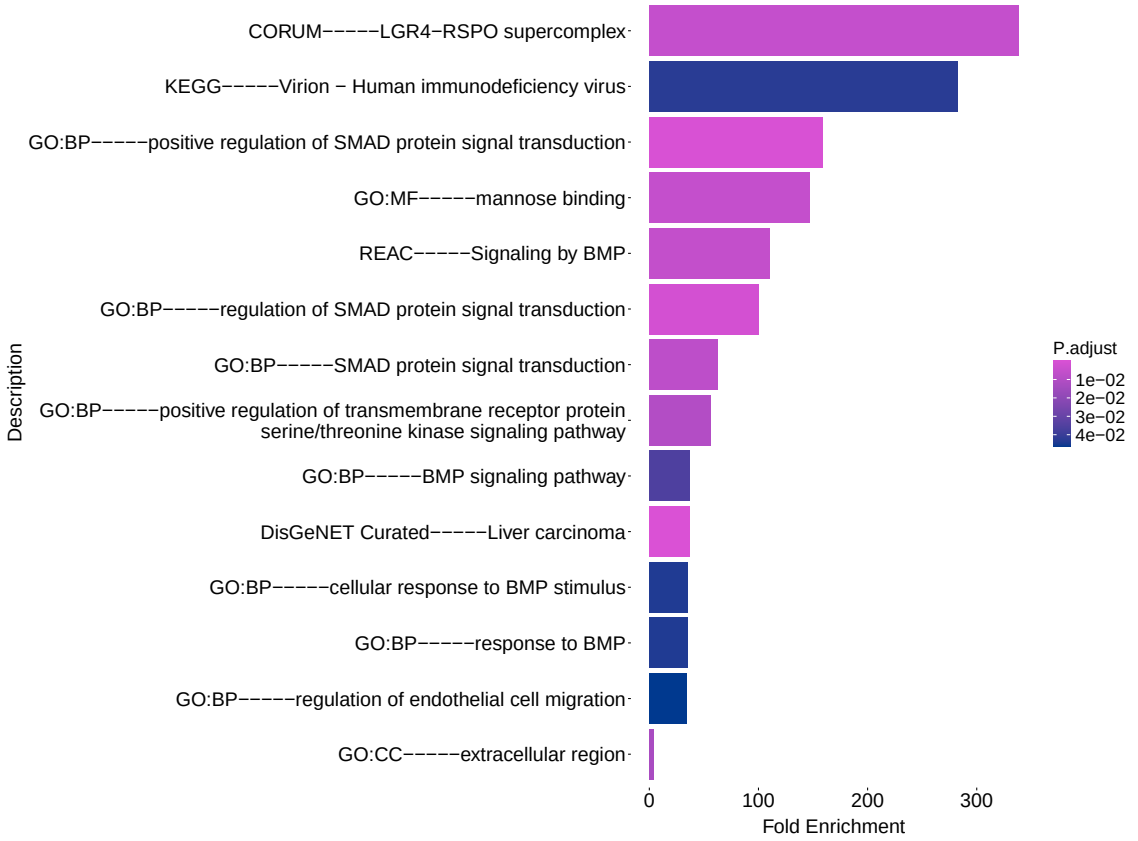

Figure 5.3: Analyse d'enrichissement d'annotation fonctionnelle des biomarqueurs obtenus à partir de la comparaison du cancer du foie : Normal versus stade II.

5.3 Biomarqueurs et voies biologiques spécifiques d'un cancer indépendamment du stade

Un utilisateur ou une utilisatrice souhaitant identifier des biomarqueurs et voies biologiques spécifiques, sans sélection d'un stade de cancer en particulier, peut sélectionner les signatures dans le menu de sélection latéral gauche et naviguer dans les vues globale et multi-signatures de l'application.

5.3.1 Biomarqueurs de l'adénocarcinome des poumons

Quatre signatures *merge* ont été formulées en comparant les échantillons de classe Normal avec ceux de chaque stade de développement de ce cancer des poumons. Ces quatre signatures sont composées de 8 (Stade I), 18 (Stade II), 10 (Stade III) et 136 (Stade IV) gènes. La vue globale présente les liens (si biomarqueur(s) en commun) entre chaque paire de signatures (Figure 5.4). La paire Stade II/Stade IV a un lien plus épais pour représenter les onze biomarqueurs communs.

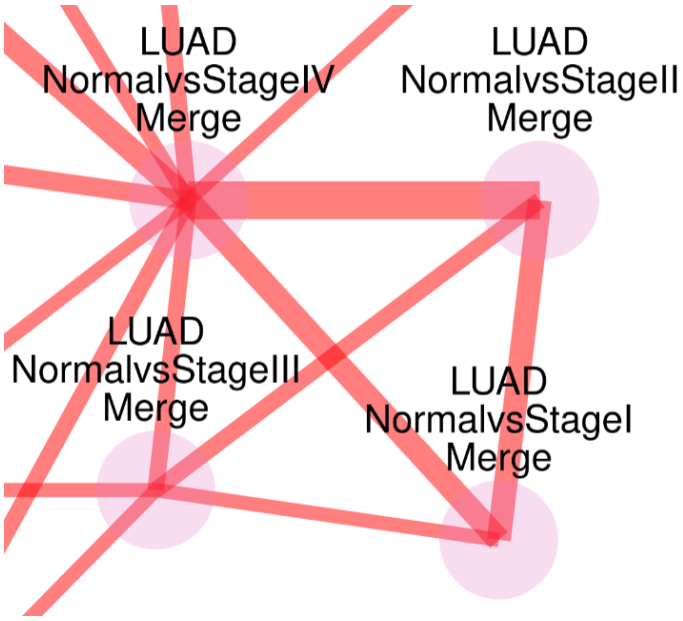

Figure 5.4: Extrait de la vue générale avec focus sur les signatures *merge* de l'adénocarcinome des poumons (LUAD). Les noeuds représentent les signatures, alors que les liens correspondent aux biomarqueurs en commun et leur épaisseur est proportionnelle au nombre de biomarqueurs communs.

Pour affiner notre analyse, nous avons pris en compte les proportions de biomarqueurs partagés entre chaque paire de comparaison (ratio de biomarqueurs communs entre deux signature). Pour cela, nous avons utilisé l'indice de Jaccard (Tableau 5.1) qui mesure la similarité entre deux ensembles en quantifiant la proportion d'éléments partagés. Cela permet ici d'identifier des biomarqueurs communs potentiellement pertinents pour évaluer la progression du cancer.

Formule de Jaccard:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

où A et B représentent deux signatures composées de biomarqueurs, $A \cap B$ est l'intersection de ces signatures (les biomarqueurs communs), et $A \cup B$ est leur union

(l'ensemble des biomarqueurs présents dans l'une ou l'autre des signatures).

En prenant en compte les ratios, la comparaison de la signature du Stade I *versus* le Stade II donne la meilleure valeur avec l'indice de Jaccard (Tableau 5.1). Cette paire de comparaison confronte le stade le plus précoce du cancer et le stade suivant. Cinq gènes composent le chevauchement de leur liste de biomarqueurs (SGCG, STX11, GYPE, AGER, EMP2) et 21 forment l'union des deux.

Signature	LUAD SI	LUAD SII	LUAD SIII	LUAD SIV
LUAD SI	1			
LUAD SII	0,24	1		
LUAD SIII	0,06	0,08	1	
LUAD SIV	0,04	0,08	0,01	1

Tableau 5.1: Indice de Jaccard pour comparer les quatre signatures de biomarqueurs de l'adénocarcinome des poumons (LUAD) à divers stades (SI à SIV).

Pour analyser plus en profondeur les informations sur le cancer LUAD, et comparer les quatre stades entre eux, l'utilisateur peut zoomer en sélectionnant la vue multi-signature (Figure 5.5). Les enveloppes délimitent différentes régions pour regrouper les biomarqueurs (noeuds ronds) de chaque signature. Le nombre de rayures à orientations différentes est indicatif du nombre d'enveloppes ou signatures chevauchantes. Quatre zones montrent un double chevauchement (Figure 5.5 - zones à double sens de rayures), trois zones à chevauchement triple (Figure 5.5 - zones à triple sens de rayures) et aucune zone à quadruple chevauchement.

La vue multi-signature (Figure 5.5) permet d'identifier 16 biomarqueurs (noeuds en vert) appartenant à au moins deux signatures. La caractérisation biologique de l'ensemble de ces 16 biomarqueurs par analyse d'enrichissement d'annotation de type ORA ne présente néanmoins pas d'information directe. En effet, les analyses par g:Profiler et ToppGene avec les mêmes paramètres que précédemment ne donnent qu'un enrichissement significatif de l'annotation *Sarcoglycan-sarcospan complex SG-SPN* de la base de données Corum spécialisée en complexe protéique des mammifères (valeur p ajustée de 0.047 via le gène SGCG).

On peut néanmoins observer dans la vue multi-signature (Figure 5.5) que dans le cadre de l'adénocarcinome des poumons, une voie biologique qui est la voie de la dégranulation des plaquettes, est commune à cinq gènes. Ces derniers sont soit spécifiques du stade III (ACTN2), du stade IV (VEGFD, A2M, GAS6) ou communs

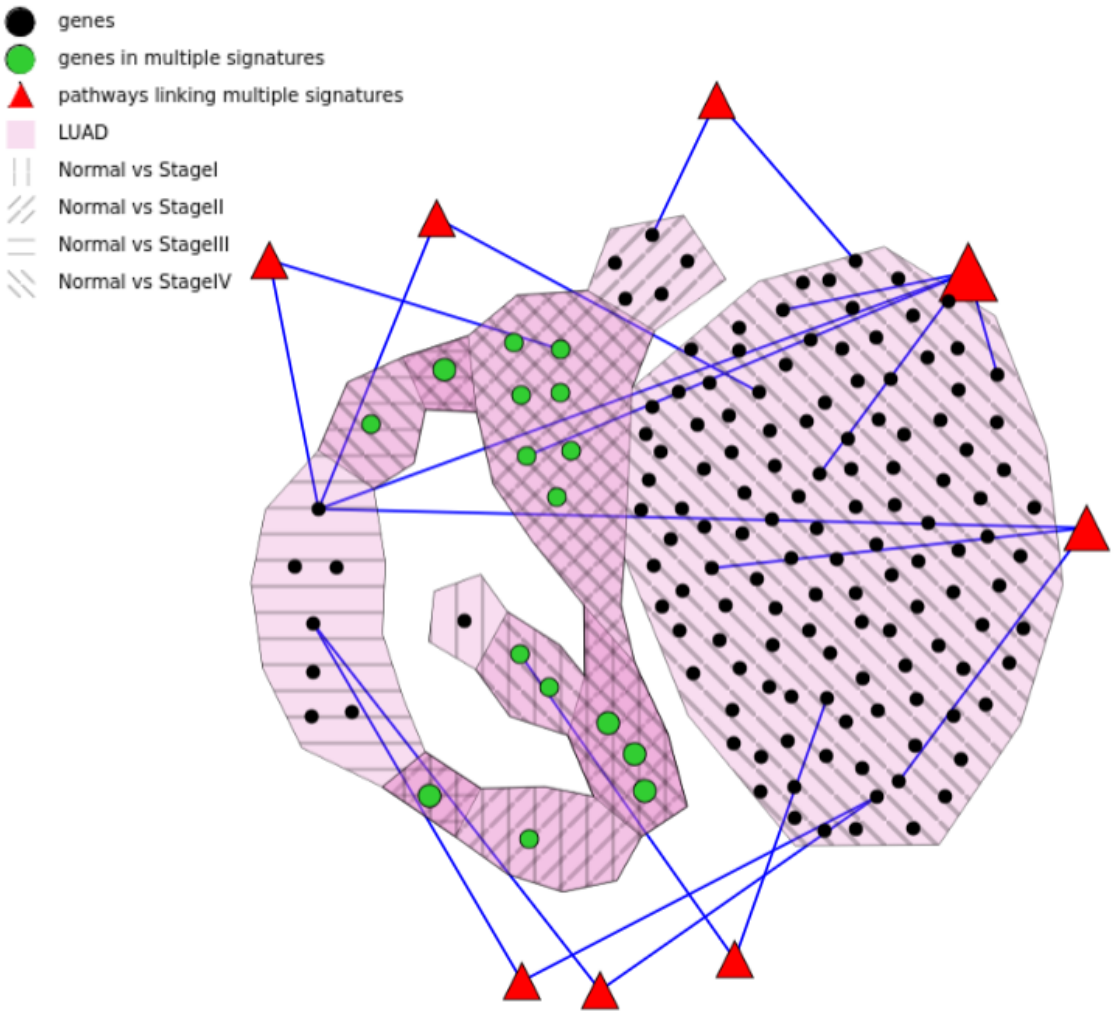

Figure 5.5: Extrait de la vue multi-signature après sélection dans le menu latéral gauche de l'adénocarcinome des poumons (LUAD) en scénario dit *merge* et intégration des 4 comparaisons disponibles.

aux stade II et IV (CLEC3B). Le rôle des plaquettes dans l'adénocarcinome des poumons fait l'objet de recherches récentes et il semblerait qu'elles puissent avoir un impact dans la dissémination de métastases [150]. Cela semble cohérent avec la présence de trois gènes dans la signature du stade le plus avancé du cancer des poumons.

5.4 Biomarqueurs et voies biologiques spécifiques d'un stade d'avancement de cancer indépendamment du type de cancer

5.4.1 Analyse globale des similarités par stade

La figure 5.6 démontre rapidement l'intérêt de la vue multi-signature pour comparer plusieurs types de cancer (enveloppes colorées en fonction) à un stade spécifique.

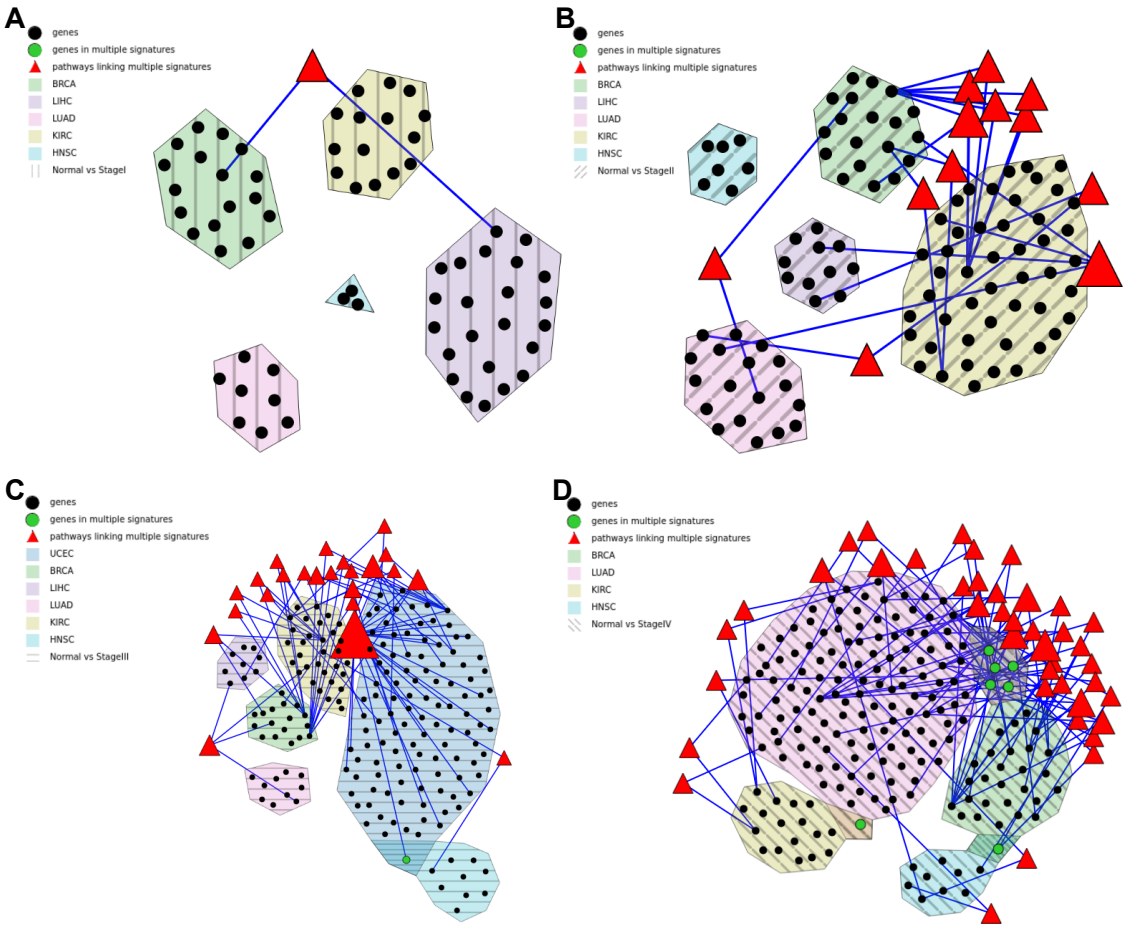

Figure 5.6: Vue multi-signature par stade d'avancement des cancers : Stade I (panneau A), Stade II (panneau B), Stade III (panneau C), Stade IV (panneau D). Les différents cancers sont : BRCA (sein), HNSC (tête et cou), KIRC (rein), LIHC (foie), LUAD (poumons), UCEC(utérus).

Il est intéressant d'identifier rapidement les biomarqueurs communs à plusieurs types de cancer pour un stade donné (noeuds colorés en vert: panneau C et D) ou leur absence. À noter que le stade IV et celui pour lequel on dénombre le plus de biomarqueurs communs à plusieurs types de cancer pour un même stade de

développement. Une analyse plus détaillée pour le stade III est présentée dans la section suivante.

5.4.2 Caractérisation du stade III des cancers

En partant des six signatures des différents types de cancer, nous proposons de nous focaliser sur le stade III (Figure 5.7-A). L'analyse de ce stade d'avancement du cancer est pertinente pour appréhender les mécanismes précédant la progression en stade métastatique (stade IV).

La vue multi-signature de **THe Biom** (Figure 5.7-A) permet d'identifier des liens entre les signatures de stade III des six cohortes de patients. Il existe un gène directement chevauchant entre les deux signatures (noeud gène vert) du cancer de l'utérus UCEC et de la tête et du cou HNSC. Il s'agit du gène ESPL1 qui est un gène peu étudié en oncologie.

La voie biologique de séparation des chromatides soeurs (*Separation of Sister Chromatids*) est représentée par un noeud de forme triangle de taille très supérieure aux autres noeuds d'annotation. Cette voie est donc liée à de nombreux gènes, mais presque tous (22 parmi les 23 reliés) sont spécifiques du cancer de l'utérus.

On observe par ailleurs que la voie biologique de dégranulation des plaquettes (*Platelet degranulation*) semble d'intérêt commun à de multiples cancer, car elle est partagée par quatre signatures différentes de cancers (foie, rein, sein, poumons). La Figure 5.7-B présente une vue multi-signature concentrée sur ces quatre signatures. Les quatre gènes concernés sont mis en valeur dans la vue multi-signature et dans le menu de sélection de gènes (ECM1 en bleu, ACTN2 en orange, KNG1 en vert et VEGFD en rouge). Les niveaux d'expression de ces gènes obtenus à la demande (Figure 5.7-C) montre que les gènes ECM1 et ACTN2 ont des niveaux d'expression proches entre les classes d'échantillons utilisés lors de la sélection de variables (foie Normal contre cancer du foie LIHC de stade III pour ECM1 et poumons Normaux contre cancer des poumons LUAD de stade III pour ACTN2). Une analyse statistique du test T de Student ne révèle d'ailleurs pas de différences d'expression significatives entre ces distributions.

Nous avons évalué la diversité des biomarqueurs dans chaque voie biologique observée dans la vue multi-signature avec les six signatures de cancer de stade III

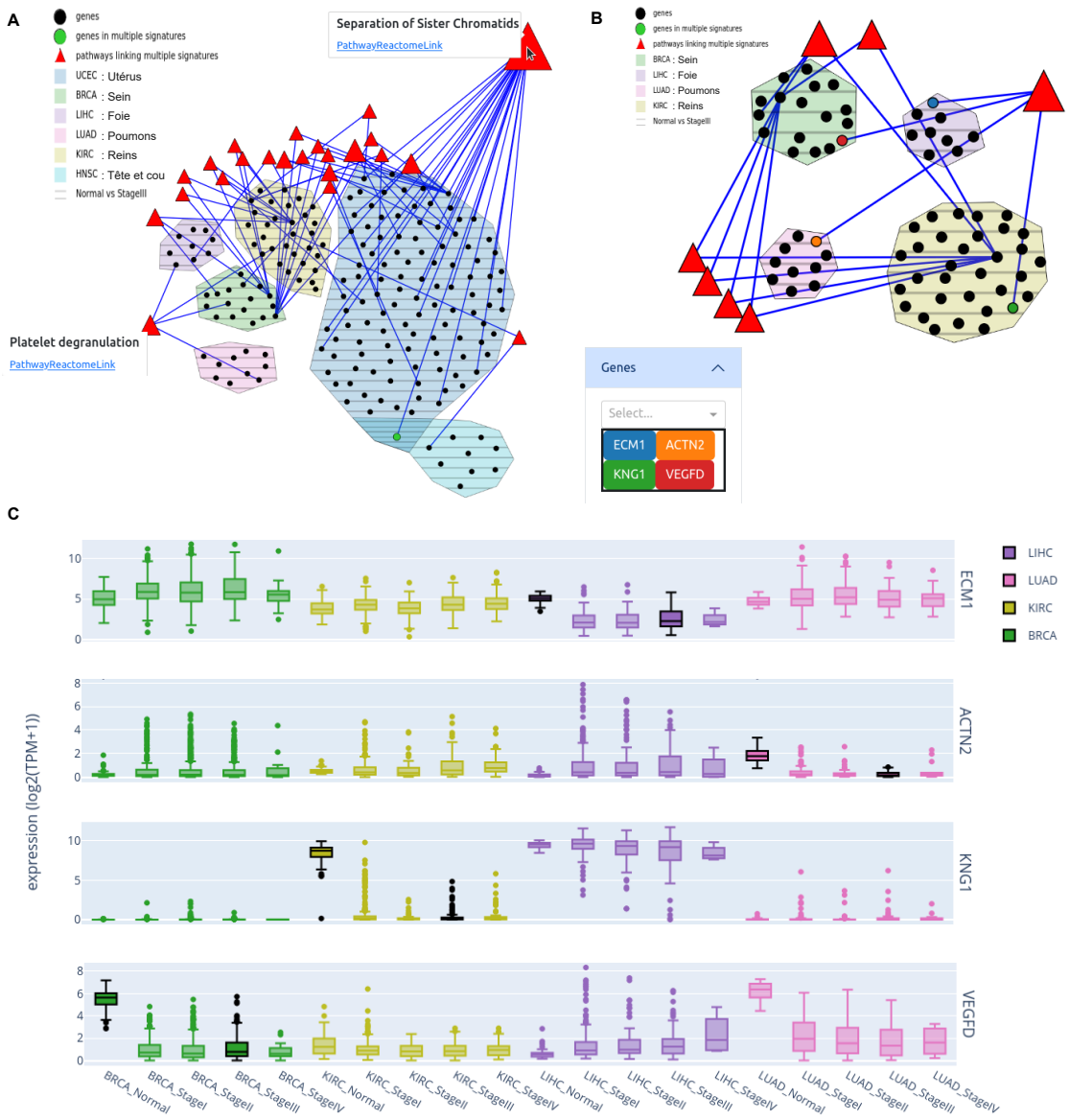

Figure 5.7: Visualisation des six signatures de cancers de stade III incluses dans la preuve de concept de **THe Biom**. (A) Vue multi-signature de **THe Biom** des six signatures de cancer de stade III et des voies biologiques qui leur sont associées. (B) Vue multi-signature réduite aux quatre signature de cancers de stade III partageant un gène avec la voie biologique *Platelet degranulation*. Les quatre gènes impliqués sont colorés dans la vue et dans le menu de sélection de gène (en bas à gauche du panneau B). (C) Niveaux d'expression des quatre gènes pour les diverses classes d'échantillons disponibles pour les quatre type de cancer retenus. Les boîtes à moustaches des classes Normal et Stade III sont encadrées de noir suivant l'implication du gène dans la signature qui résulte de la comparaison des échantillons de ces classes.

(Figure 5.7-A). Pour cela, nous avons calculé l'indice de Shannon (Tableau 5.2). Il s'agit plus précisément d'évaluer la spécificité de chaque terme d'annotation en prenant en compte l'origine de signature des biomarqueurs annotés par le terme.

Formule de l'indice de Shannon:

$$H' = -\sum_{i=1}^{S} p_i \ln p_i$$

où $p_i = n_i/N$, S est le nombre de signatures associées à la voie biologique, p_i la proportion de gènes de la $i$ème signature, n_i le nombre de gènes de la signature i et N le nombre total de gènes parmi toutes les signatures S. Plus la valeur de l'indice de Shannon est élevée, plus l'annotation intègre des gènes venant de signatures différentes.

Les résultats du tableau 5.2 permettent de confirmer que la voie biologique *Platelet degranulation* est davantage commune aux divers cancers analysés (indice de Shannon de 1,39) que la voie biologique *Separation of Sister Chromatids* (indice de Shannon de 0,18).

La voie biologique de séparation des chromatides soeurs a déjà été étudiée dans le cadre du cancer de l'utérus [151, 152]. Le gène UBE2C, que nous identifions dans **THe Biom** comme étant un des biomarqueurs du stade III du cancer endométrial de l'utérus, a par ailleurs été récemment proposé comme biomarqueur de pronostic pour cette pathologie [153].

Quant au rôle des plaquettes et leur dégranulation, il a été mis en lumière dans cette dernière décennie pour leur rôle important dans la progression du cancer [154, 155].

Voie biologique (nom Reactome)	Nombre de signature incluses	Nombre total de gènes inclus	Indice de diversité de Shannon
Platelet degranulation	4	4	1.39
Anchoring of the basal body to the plasma membrane	2	2	0.69
Assembly of collagen fibrils and other multimeric structures	2	2	0.69
Collagen biosynthesis and modifying enzymes	2	2	0.69
Collagen chain trimerization	2	2	0.69
ECM proteoglycans	2	2	0.69
G alpha (s) signalling events	2	2	0.69
Integrin cell surface interactions	2	2	0.69
Loss of Nlp from mitotic centrosomes	2	2	0.69
Loss of proteins required for interphase microtubule organization from the centrosome	2	2	0.69
Nephron development	2	2	0.69
Non-integrin membrane-ECM interactions	2	2	0.69
RHO GTPases activate PKNs	2	2	0.69
Recruitment of NuMA to mitotic centrosomes	2	2	0.69
Recruitment of mitotic centrosome proteins and complexes	2	2	0.69
TP53 Regulates Transcription of DNA Repair Genes	2	2	0.69
Ub-specific processing proteases	2	2	0.69
APC-Cdc20 mediated degradation of Nek2A	2	3	0.64
Collagen degradation	2	3	0.64
Neddylation	2	3	0.64
AURKA Activation by TPX2	2	4	0.56
Regulation of PLK1 Activity at G2/M Transition	2	5	0.50
Separation of Sister Chromatids	2	23	0.18

Tableau 5.2: Liste des 23 voies biologiques qui incluent au moins un gène provenant d'au minimum deux signatures différentes de cancer de stade III, ainsi que le nombre total de gènes associés, le nombre de signatures et la valeur de l'indice de diversité de Shannon correspondant.

5.5 Analyse ciblée d'un gène ou d'une voie biologique d'intérêt dans l'atlas

Un autre type d'analyse peut être réalisé en s'intéressant spécifiquement à un gène ou à une voie biologique d'intérêt. Pour réaliser ce type de sélection, le panneau latéral gauche propose un champs textuel pour saisir le nom d'un ou plusieurs gènes (par identifiant Ensembl ou symbole) et/ou le nom Reactome d'une voie biologique. Une seconde possibilité pour se focaliser sur un gène, consiste à cliquer directement sur un noeud.

À partir de la sélection d'un ou plusieurs gènes et/ou d'une ou plusieurs voies biologiques, les vues multi-signature, mono-signature et vue d'expression de gènes sont mises à jour en lien avec la sélection.

5.5.1 Analyse à partir de la sélection d'un gène

Pour illustrer ce type d'analyse, nous nous sommes intéressés à l'étude du gène MMP11 (Figure 5.8). Le panneau de vue multi-signature met le noeud gène MMP11 en valeur avec une couleur bleue. Il est présent dans les signatures du cancer de la tête et du cou (HNSC) de stade IV ainsi que dans les quatre signatures (Stade I à IV) du cancer du sein (BRCA). L'aperçu des boîtes à moustaches nous permet de comparer son expression pour ces deux cancers et montre une différence d'expression entre les phénotypes dits Normaux et les échantillons de stade I à IV du cancer du sein et le stade IV du cancer de la tête et du cou.

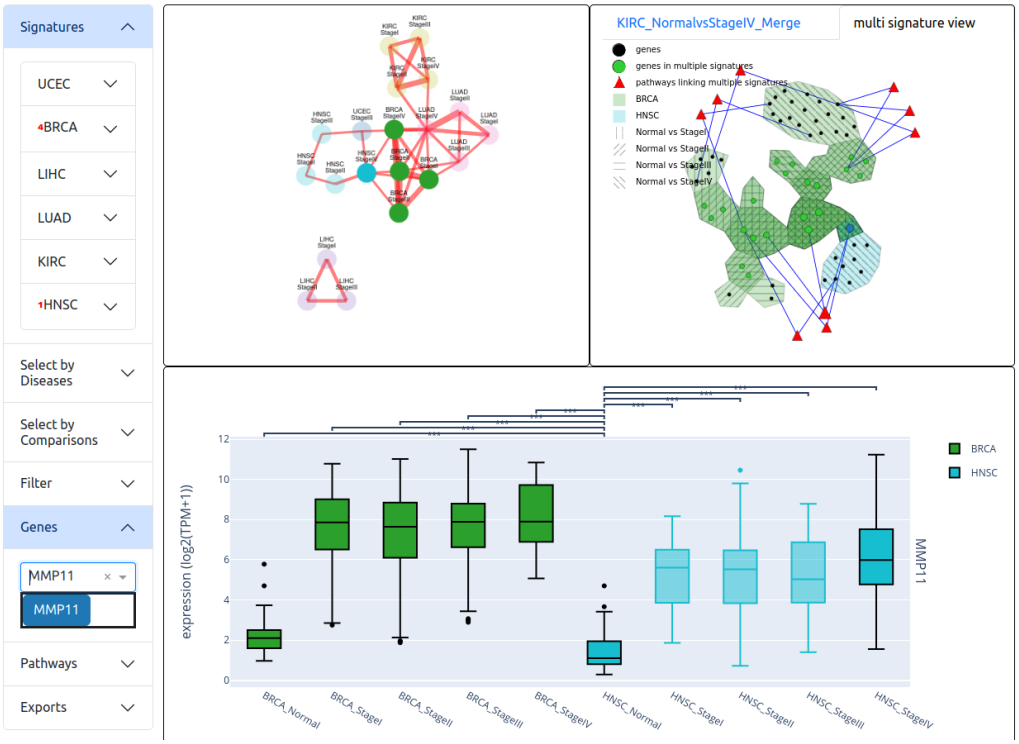

Figure 5.8: Capture d'écran de THe Biom à partir de la sélection du gène MMP11 dans le panneau de saisie textuelle à gauche. Les vues multi-signature (en haut à droite) et d'expression de gènes sont mis à jour à partir de la sélection. Le gène est colorié en bleu dans la vue multi-signature. Les boîtes à moustache d'expression de gènes pour les phénotypes inclus dans les comparaisons pour lesquelles le gène fait partie de la signature sont encadrées de noir.

5.5.2 Analyse à partir de la sélection d'une voie biologique

À partir de la sélection d'une voie biologique, les gènes associés à cette annotation sont mis à jour différemment dans les vues multi- et mono-signature selon si elle concerne un ou plusieurs gènes.

Pour illustrer, nous avons sélectionné la voie biologique de la cascade des kinases RAF/MAP (Figure 5.9). Trois gènes (ANGPT1, DLG2, ACTN2) de trois signatures différentes lui sont associés.

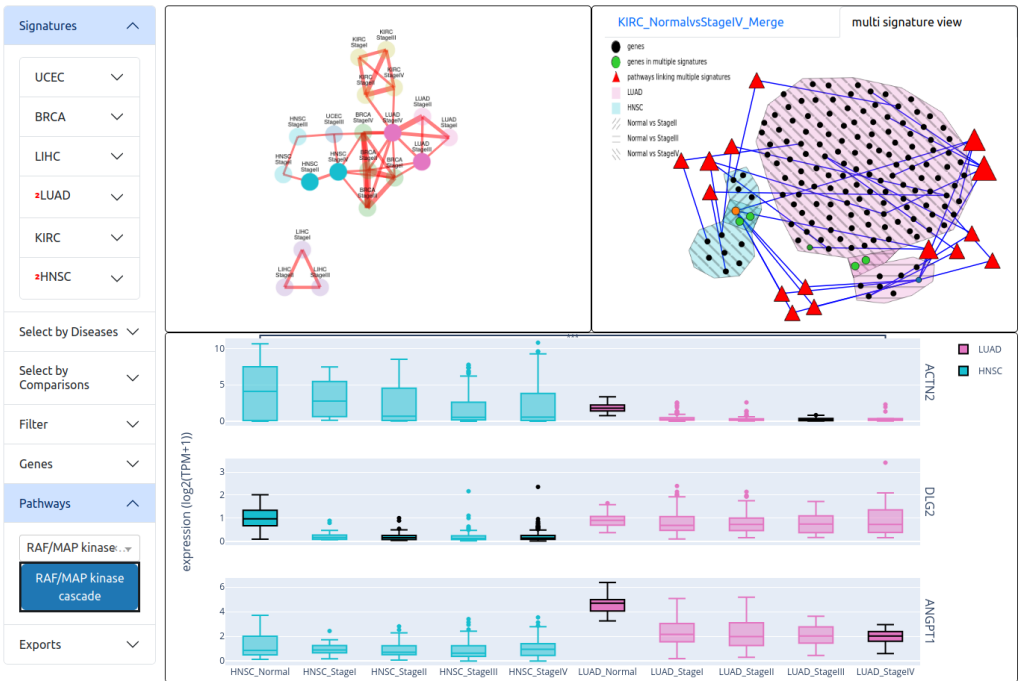

Figure 5.9: Capture d'écran de **THe Biom** à partir de la sélection de la voie biologique de la cascade des kinases RAF/MAP dans le panneau de saisie textuelle à gauche. Les vues multi-signature (en haut à droite) et d'expression de gènes sont mises à jour. Les boîtes à moustache d'expression de gènes pour les phénotypes inclus dans les comparaisons pour lesquelles les gènes font partie de la signature sont encadrées de noir.

5.6 Conclusion

Ce chapitre illustre plusieurs scénarios d'analyse correspondant à différentes question biologiques. Par l'utilisation de **THe Biom**, l'utilisatrice ou utilisateur peut comparer des biomarqueurs en se focalisant sur un ou plusieurs types ou stades de cancer. De plus, différentes analyses lui facilitent l'exploration des signatures allant de la comparaison globale des signatures jusqu'à la visualisation de l'expression des gènes.

Conclusion et Perspectives

Cette thèse avait pour objectif de développer des approches pour l'analyse de maladies complexes tels que le cancer. Plus précisément, elle adressait les deux objectifs suivants :

1. Développement d'une méthode bioinformatique pour l'identification de biomarqueurs robustes de maladies complexes comme le cancer (Chapitre 2 et 3)

2. Définition d'un système de visualisation pour la mise à disposition et l'analyse d'un atlas de signatures robustes de biomarqueurs transcriptomiques de types et stades divers de cancers (Chapitre 4 et 5)

Le premier objectif s'est concrétisé par le développement d'une nouvelle approche de sélection de variables hybride ensembliste HEFS. Une méthode ensembliste présente l'avantage de palier aux différentes variations méthodologiques et biologiques des données d'entrées en recherche biomédicale. Trois étapes critiques à la mise en place d'une approche hybride ensembliste ont été explorées et prises en compte. La première étape de réduction de dimensions par filtre peut être décisive et doit éviter d'intégrer des biais de décisions lors de son application. Une solution pour prendre en compte ce problème repose sur la mise en place d'approches de perturbation des données. Peu évoqué dans les méthodes ensemblistes de sélection de variable, ajouter une étape de perturbation permet d'éviter d'améliorer la séparation des échantillons. L'étape finale d'agrégation est également une étape clé pour le développement d'une méthode ensembliste et doit permettre d'intégrer une grande variabilité (ex : des modèles de ML aussi performants les uns que les autres, mais avec des signatures différentes) et la stabiliser pour proposer une sélection de variable robuste. Cette approche a été appliquée sur différents cancers, y compris avec une cohorte externe de validation. En définitive, cette nouvelle méthode répond au be-

soin croissant de fiabilité dans l'identification des biomarqueurs. Elle offre à la communauté de recherche biomédicale la possibilité d'explorer de vastes ensembles de données, tout en favorisant l'émergence de pistes de recherche solides. Cela permet d'optimiser les ressources humaines, temporelles et matérielles. Ces développements ont été valorisés par une publication dans le journal *Nucleic Acids Research* [144]. Dans un contexte de ressources informatiques illimitées, une première perspective de ces travaux serait d'ajouter davantage d'algorithmes d'apprentissage automatique et une grille de valeurs d'hyper-paramètres plus étendue à la méthodologie HEFS que nous avons développé. Cela participerait à accroître sa robustesse théorique. Par ailleurs, nous disposons d'un nombre croissant de données omiques et pour de multiples cohortes de patients. L'exploitation de ces données permettrait de représenter et analyser au mieux les maladies correspondantes. Pour cela, une seconde perspective de ces travaux serait d'ajouter un niveau de perturbation des données supplémentaire comme la perturbation de données par intégration de données multi-cohortes.

La solution de visualisation proposée pour répondre au deuxième objectif est en cours de développement, mais les résultats préliminaires sont encourageants. Une version en mode *light* HEFS a été employée pour définir des signatures pour six cancers à divers stades d'avancement. L'intégration de celles-ci dans **THe Biom**, notre application web, a permis une exploration en profondeur de ces signatures, de leur intérêt individuel ainsi que de leur relations directes et indirectes et donc des phénotypes qu'elles définissent. La poursuite de ce projet devrait aboutir à un large atlas de signatures de biomarqueurs de transcriptomiques de cancers. À terme, cet atlas pourra permettre aux équipes de recherche en oncologie d'avoir un accès facilité à des données transcriptomiques de cancers déjà analysées et pré-caractérisées à large échelle et suivant diverses stratifications. Son application pourrait conduire à l'identification de nouvelles cibles thérapeutiques et à la mise au point de nouvelles molécules adaptées.

Ce travail a fait l'objet d'une présentation orale associée à un poster lors de la 32ème édition de la conférence *Intelligent Systems for Molecular Biology* qui a eu lieu du 12 au 16 juillet 2024 à Montréal, QC, Canada.

La preuve de concept présentée dans ces travaux peut être adaptée pour don-

ner l'opportunité à l'utilisatrice ou utilisateur d'explorer de nouvelles questions biologiques. Par exemple, il serait possible de tester une hypothèse de gènes biomarqueurs en important ses propres données de listes de gènes d'intérêt pour un phénotype inclus ou non dans l'atlas. De plus, bien que **THe Biom** ait été originellement développée pour exploiter les données de la base de données TCGA, l'application pourrait s'étendre à d'autres cohortes pour mieux représenter les phénotypes cancéreux. Cette proposition peut aussi s'appliquer à un contexte multi-omique. Le cancer étant une maladie complexe ayant des causes et conséquences à de multiples niveaux moléculaires, il serait pertinent d'explorer les liens de biomarqueurs potentiels dans cette organisation moléculaire. Le système de visualisation développé serait alors amené à être adapté pour permettre d'exploiter ce niveau d'information supplémentaire.

Bibliographie

1. Bray, F., Laversanne, M., Weiderpass, E. & Soerjomataram, I. The ever-increasing importance of cancer as a leading cause of premature death worldwide. *Cancer* **127,** 3029–3030. ISSN: 1097-0142 (2021).

2. Mattiuzzi, C. & Lippi, G. Current Cancer Epidemiology. *Journal of Epidemiology and Global Health* **9,** 217–222. ISSN: 2210-6014 (Dec. 1, 2019).

3. Anand, P. *et al.* Cancer is a Preventable Disease that Requires Major Lifestyle Changes. *Pharmaceutical Research* **25,** 2097–2116. ISSN: 1573-904X (Sept. 1, 2008).

4. Santucci, C. *et al.* Progress in cancer mortality, incidence, and survival: a global overview. *European Journal of Cancer Prevention* **29,** 367. ISSN: 0959-8278 (Sept. 2020).

5. Thimbleby, H. Technology and the Future of Healthcare. *Journal of Public Health Research* **2,** jphr.2013.e28. ISSN: 2279-9036 (Dec. 1, 2013).

6. Malumbres, M. & Barbacid, M. RAS oncogenes: the first 30 years. *Nature Reviews Cancer* **3,** 459–465. ISSN: 1474-1768 (June 2003).

7. Parker, J. S. *et al.* Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *Journal of Clinical Oncology* **27,** 1160–1167. ISSN: 0732-183X (Mar. 10, 2009).

8. Schuster, S. C. Next-generation sequencing transforms today's biology. *Nature Methods* **5,** 16–18. ISSN: 1548-7105 (Jan. 2008).

9. Saeys, Y., Inza, I. & Larrañaga, P. A review of feature selection techniques in bioinformatics. *Bioinformatics* **23,** 2507–2517. ISSN: 1367-4803 (Oct. 1, 2007).

10. Yang, P., Hwa Yang, Y., B. Zhou, B. & Y. Zomaya, A. A Review of Ensemble Methods in Bioinformatics. *Current Bioinformatics* **5,** 296–308 (Dec. 1, 2010).

11. Zeder, M. A. Domestication and early agriculture in the Mediterranean Basin: Origins, diffusion, and impact. *Proceedings of the National Academy of Sciences* **105,** 11597–11604 (Aug. 19, 2008).

12. Kulski, J. in *Next Generation Sequencing - Advances, Applications and Challenges* (ed Kulski, J. K.) 3–60 (InTech, Croatia, Jan. 14, 2016). ISBN: 978-953-51-2240-1.

13. Gayon, J. From Mendel to epigenetics: History of genetics. *Comptes Rendus Biologies. Trajectories of genetics, 150 years after Mendel / Trajectoire de la génétique, 150 après Mendel Guest Editors / Rédacteurs en chef invités : Bernard Dujon, Georges Pelletier* **339,** 225–230. ISSN: 1631-0691 (July 1, 2016).

14. Schuster, S. C. Next-generation sequencing transforms today's biology. *Nature Methods* **5,** 16–18. ISSN: 1548-7105 (Jan. 2008).

15. Bai, B. *et al.* Proteomic landscape of Alzheimer's Disease: novel insights into pathogenesis and biomarker discovery. *Molecular Neurodegeneration* **16,** 55. ISSN: 1750-1326 (Aug. 12, 2021).

16. Danzi, F. *et al.* To metabolomics and beyond: a technological portfolio to investigate cancer metabolism. *Signal Transduction and Targeted Therapy* **8,** 1–22. ISSN: 2059-3635 (Mar. 22, 2023).

17. Lowe, R., Shirley, N., Bleackley, M., Dolan, S. & Shafee, T. Transcriptomics technologies. *PLOS Computational Biology* **13,** e1005457. ISSN: 1553-7358 (May 18, 2017).

18. Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K. & Liu, X. Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cells. *PLoS ONE* **9,** e78644. ISSN: 1932-6203 (Jan. 16, 2014).

19. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15,** 550. ISSN: 1474-760X (Dec. 5, 2014).

20. Strimbu, K. & Tavel, J. A. What are biomarkers? *Current Opinion in HIV and AIDS* **5,** 463. ISSN: 1746-630X (Nov. 2010).

21. Molinski, J., Tadimety, A., Burklund, A. & Zhang, J. X. J. Scalable Signature-Based Molecular Diagnostics Through On-chip Biomarker Profiling Coupled with Machine Learning. *Annals of Biomedical Engineering* **48,** 2377–2399. ISSN: 1573-9686 (Oct. 1, 2020).

22. Califf, R. M. Biomarker definitions and their applications. *Experimental Biology and Medicine* **243,** 213–221. ISSN: 1535-3702 (Feb. 2018).

23. Sotiriou, C. & Pusztai, L. Gene-Expression Signatures in Breast Cancer. *New England Journal of Medicine* **360,** 790–800. ISSN: 0028-4793 (Feb. 19, 2009).

24. Li, H. *et al.* MR Imaging Radiomics Signatures for Predicting the Risk of Breast Cancer Recurrence as Given by Research Versions of MammaPrint, Oncotype DX, and PAM50 Gene Assays. *Radiology* **281,** 382–391. ISSN: 0033-8419 (Nov. 2016).

25. Zou, D., Ma, L., Yu, J. & Zhang, Z. Biological Databases for Human Research. *Genomics, Proteomics & Bioinformatics* **13,** 55–63. ISSN: 1672-0229 (Feb. 1, 2015).

26. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nature genetics* **25,** 25–29. ISSN: 1061-4036 (May 2000).

27. The Gene Ontology Consortium *et al.* The Gene Ontology knowledgebase in 2023. *Genetics* **224,** iyad031. ISSN: 1943-2631 (May 2, 2023).

28. Creixell, P. *et al.* Pathway and Network Analysis of Cancer Genomes. *Nature methods* **12,** 615–621. ISSN: 1548-7091 (July 2015).

29. Fabregat, A. *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Research* **46,** D649–D655. ISSN: 0305-1048 (D1 Jan. 4, 2018).

30. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* **28,** 27–30. ISSN: 0305-1048 (Jan. 1, 2000).

31. Martens, M. *et al.* WikiPathways: connecting communities. *Nucleic Acids Research* **49,** D613–D621. ISSN: 0305-1048 (D1 Jan. 8, 2021).

32. Mi, H. *et al.* The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Research* **33,** D284–D288. ISSN: 0305-1048 (suppl_1 Jan. 1, 2005).

33. Rodchenkov, I. *et al.* Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. *Nucleic Acids Research* **48,** D489–D497. ISSN: 0305-1048 (D1 Jan. 8, 2020).

34. Piñero, J. *et al.* The DisGeNET knowledge platform for disease genomics: 2019 update. *Nucleic Acids Research* **48,** D845–D855. ISSN: 0305-1048 (D1 Jan. 8, 2020).

35. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Research* **43,** D789–D798. ISSN: 0305-1048 (D1 Jan. 28, 2015).

36. Tomczak, K., Czerwińska, P. & Wiznerowicz, M. Review
The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia* **2015,** 68–77. ISSN: 1428-2526, 1897-4309 (2015).

37. Colaprico, A. *et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research* **44,** e71. ISSN: 0305-1048 (May 5, 2016).

38. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Research* **41,** D991–D995. ISSN: 0305-1048 (D1 Jan. 1, 2013).

39. Clough, E. *et al.* NCBI GEO: archive for gene expression and epigenomics data sets: 23-year update. *Nucleic Acids Research* **52,** D138–D144. ISSN: 0305-1048 (D1 Nov. 2, 2023).

40. *AJCC cancer staging manual* (eds Edge, S. B. & American Joint Committee on Cancer) 7th ed (Springer, New York, NY, 2010). 648 pp. ISBN: 978-0-387-88440-0.

41. Tang, Z., Kang, B., Li, C., Chen, T. & Zhang, Z. GEPIA2: an enhanced web server for large-scale expression profiling and interactive analysis. *Nucleic Acids Research* **47,** W556–W560. ISSN: 0305-1048 (W1 July 2, 2019).

42. Chou, P.-H. *et al.* TACCO, a Database Connecting Transcriptome Alterations, Pathway Alterations and Clinical Outcomes in Cancers. *Scientific Reports* **9,** 3877. ISSN: 2045-2322 (Mar. 7, 2019).

43. Chandrashekar, D. S. *et al.* UALCAN: A Portal for Facilitating Tumor Subgroup Gene Expression and Survival Analyses. *Neoplasia* **19,** 649–658. ISSN: 1476-5586 (Aug. 1, 2017).

44. Tang, Z. *et al.* GEPIA: a web server for cancer and normal gene expression profiling and interactive analyses. *Nucleic Acids Research* **45,** W98–W102. ISSN: 0305-1048 (W1 July 3, 2017).

45. Maglogiannis, I. G. *Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in EHealth, HCI, Information Retrieval and Pervasive Technologies* 420 pp. ISBN: 978-1-58603-780-2 (IOS Press, 2007).

46. Breiman, L. Random Forests. *Machine Learning* **45,** 5–32. ISSN: 1573-0565 (Oct. 1, 2001).

47. Chen, T. & Guestrin, C. *XGBoost: A Scalable Tree Boosting System* in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Association for Computing Machinery, New York, NY, USA, Aug. 13, 2016), 785–794. ISBN: 978-1-4503-4232-2.

48. Breiman, L., Friedman, J. H., Olshen, R. A. & Stone, C. J. *Classification And Regression Trees* 368 pp. ISBN: 978-1-315-13947-0 (Routledge, New York, Oct. 25, 2017).

49. Lewis, D. D. *Naive (Bayes) at forty: The independence assumption in information retrieval* in *Machine Learning: ECML-98* (eds Nédellec, C. & Rouveirol, C.) (Springer, Berlin, Heidelberg, 1998), 4–15. ISBN: 978-3-540-69781-7.

50. Puga, J. L., Krzywinski, M. & Altman, N. Bayesian networks. *Nature Methods* **12,** 799–800. ISSN: 1548-7105 (Sept. 1, 2015).

51. Webb, G. I., Boughton, J. R. & Wang, Z. Not So Naive Bayes: Aggregating One-Dependence Estimators. *Machine Learning* **58,** 5–24. ISSN: 1573-0565 (Jan. 1, 2005).

52. Cover, T. & Hart, P. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* **13,** 21–27. ISSN: 1557-9654 (Jan. 1967).

53. Noble, W. S. What is a support vector machine? *Nature Biotechnology* **24,** 1565–1567. ISSN: 1546-1696 (Dec. 2006).

54. Shahid, N., Rappon, T. & Berta, W. Applications of artificial neural networks in health care organizational decision-making: A scoping review. *PLOS ONE* **14,** e0212356. ISSN: 1932-6203 (Feb. 19, 2019).

55. Shwartz-Ziv, R. & Armon, A. Tabular data: Deep learning is not all you need. *Information Fusion* **81,** 84–90. ISSN: 1566-2535 (May 1, 2022).

56. Santorsola, M. & Lescai, F. The promise of explainable deep learning for omics data analysis: Adding new discovery tools to AI. *New Biotechnology* **77,** 1–11. ISSN: 1871-6784 (Nov. 25, 2023).

57. Torres, R. & Judson-Torres, R. L. Research Techniques Made Simple: Feature Selection for Biomarker Discovery. *Journal of Investigative Dermatology* **139,** 2068–2074.e1. ISSN: 0022-202X (Oct. 1, 2019).

58. Tang, J., Aleyani, S. & Liu, H. in *Data Classification* 37–64 (CRC Press, Jan. 1, 2014). ISBN: 978-1-4665-8674-1.

59. Jović, A., Brkić, K. & Bogunović, N. *A review of feature selection methods with applications* in *2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* 2015 38th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO) (May 2015), 1200–1205.

60. Remeseiro, B. & Bolon-Canedo, V. A review of feature selection methods in medical applications. *Computers in Biology and Medicine* **112,** 103375. ISSN: 0010-4825 (Sept. 1, 2019).

61. Cherrington, M., Thabtah, F., Lu, J. & Xu, Q. *Feature Selection: Filter Methods Performance Challenges* in *2019 International Conference on Computer and Information Sciences (ICCIS)* 2019 International Conference on Computer and Information Sciences (ICCIS) (Apr. 2019), 1–4.

62. Bommert, A., Welchowski, T., Schmid, M. & Rahnenführer, J. Benchmark of filter methods for feature selection in high-dimensional gene expression survival data. *Briefings in Bioinformatics* **23,** bbab354. ISSN: 1477-4054 (Jan. 1, 2022).

63. Li, L., Ching, W.-K. & Liu, Z.-P. Robust biomarker screening from gene expression data by stable machine learning-recursive feature elimination methods. *Computational Biology and Chemistry* **100,** 107747. ISSN: 1476-9271 (Oct. 1, 2022).

64. Petrini, I., Cecchini, R. L., Mascaró, M., Ponzoni, I. & Carballido, J. A. Papillary Thyroid Carcinoma: A thorough Bioinformatic Analysis of Gene Expression and Clinical Data. *Genes* **14,** 1250. ISSN: 2073-4425 (June 11, 2023).

65. Zhang, Z. & Liu, Z.-P. *Identifying Cancer Biomarkers from High-Throughput RNA Sequencing Data by Machine Learning* in *Intelligent Computing Theories and Application* (eds Huang, D.-S., Jo, K.-H. & Huang, Z.-K.) (Springer International Publishing, Cham, 2019), 517–528. ISBN: 978-3-030-26969-2.

66. Abbas, M. & EL-Manzalawy, Y. Machine learning based refined differential gene expression analysis of pediatric sepsis. *BMC Medical Genomics* **13,** 122. ISSN: 1755-8794 (Aug. 28, 2020).

67. Chormunge, S. & Jena, S. Efficient Feature Subset Selection Algorithm for High Dimensional Data. *International Journal of Electrical and Computer Engineering (IJECE)* **6,** 1880–1888. ISSN: 2722-2578 (Aug. 1, 2016).

68. Li, J. *et al.* Feature Selection: A Data Perspective. *ACM Comput. Surv.* **50,** 94:1–94:45. ISSN: 0360-0300 (Dec. 6, 2017).

69. Robnik-Šikonja, M. & Kononenko, I. Theoretical and Empirical Analysis of ReliefF and RReliefF. *Machine Learning* **53,** 23–69. ISSN: 1573-0565 (Oct. 1, 2003).

70. Chen, G. & Chen, J. A novel wrapper method for feature selection and its applications. *Neurocomputing* **159,** 219–226. ISSN: 0925-2312 (July 2, 2015).

71. Duan, K.-B., Rajapakse, J., Wang, H. & Azuaje, F. Multiple SVM-RFE for gene selection in cancer classification with expression data. *IEEE Transactions on NanoBioscience* **4,** 228–234. ISSN: 1558-2639 (Sept. 2005).

72. Vergara, J. R. & Estévez, P. A. A review of feature selection methods based on mutual information. *Neural Computing and Applications* **24,** 175–186. ISSN: 1433-3058 (Jan. 1, 2014).

73. Chowdhury, M. Z. I. & Turin, T. C. Variable selection strategies and its importance in clinical prediction modelling. *Family Medicine and Community Health* **8,** e000262. ISSN: 2305-6983 (Feb. 16, 2020).

74. Sharma, A., Imoto, S. & Miyano, S. A Top-r Feature Selection Algorithm for Microarray Gene Expression Data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **9,** 754–764. ISSN: 1557-9964 (May 2012).

75. Guyon, I. & Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **3,** 1157–1182. ISSN: 1532-4435 (null Mar. 1, 2003).

76. Liu, H., Zhou, M. & Liu, Q. An embedded feature selection method for imbalanced data classification. *IEEE/CAA Journal of Automatica Sinica* **6,** 703–715. ISSN: 2329-9274 (May 2019).

77. Chandrashekar, G. & Sahin, F. A survey on feature selection methods. *Computers & Electrical Engineering. 40th-year commemorative issue* **40,** 16–28. ISSN: 0045-7906 (Jan. 1, 2014).

78. Bolón-Canedo, V. & Alonso-Betanzos, A. Ensembles for feature selection: A review and future trends. *Information Fusion* **52,** 1–12. ISSN: 1566-2535 (Dec. 1, 2019).

79. Moon, M. & Nakai, K. Stable feature selection based on the ensemble L1-norm support vector machine for biomarker discovery. *BMC Genomics* **17,** 1026. ISSN: 1471-2164 (Dec. 22, 2016).

80. Liu, X., Zhang, Y., Fu, C., Zhang, R. & Zhou, F. EnRank: An Ensemble Method to Detect Pulmonary Hypertension Biomarkers Based on Feature Selection and Machine Learning Models. *Frontiers in Genetics* **12.** ISSN: 1664-8021 (Apr. 27, 2021).

81. Colombelli, F., Kowalski, T. W. & Recamonde-Mendoza, M. A hybrid ensemble feature selection design for candidate biomarkers discovery from transcriptome profiles. *Knowledge-Based Systems* **254,** 109655. ISSN: 0950-7051 (Oct. 27, 2022).

82. Cheng, L.-H., Hsu, T.-C. & Lin, C. Integrating ensemble systems biology feature selection and bimodal deep neural network for breast cancer prognosis prediction. *Scientific Reports* **11,** 14914. ISSN: 2045-2322 (July 21, 2021).

83. Waring, J., Lindvall, C. & Umeton, R. Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. *Artificial Intelligence in Medicine* **104,** 101822. ISSN: 0933-3657 (Apr. 1, 2020).

84. Kotthoff, L., Thornton, C., Hoos, H. H., Hutter, F. & Leyton-Brown, K. in *Automated Machine Learning: Methods, Systems, Challenges* (eds Hutter, F., Kotthoff, L. & Vanschoren, J.) 81–95 (Springer International Publishing, Cham, 2019). ISBN: 978-3-030-05318-5.

85. Feurer, M. *et al.* in *Automated Machine Learning: Methods, Systems, Challenges* (eds Hutter, F., Kotthoff, L. & Vanschoren, J.) 113–134 (Springer International Publishing, Cham, 2019). ISBN: 978-3-030-05318-5.

86. Olson, R. S. *et al. Automating Biomedical Data Science Through Tree-Based Pipeline Optimization* in *Applications of Evolutionary Computation* (eds Squillero, G. & Burelli, P.) (Springer International Publishing, Cham, 2016), 123–137. ISBN: 978-3-319-31204-0.

87. Swearingen, T. *et al. ATM: A distributed, collaborative, scalable system for automated machine learning* in *2017 IEEE International Conference on Big Data (Big Data)* 2017 IEEE International Conference on Big Data (Big Data) (Dec. 2017), 151–162.

88. Wistuba, M., Schilling, N. & Schmidt-Thieme, L. in *Proceedings of the 2017 SIAM International Conference on Data Mining (SDM)* 741–749 (Society for Industrial and Applied Mathematics, June 9, 2017).

89. Mohr, F., Wever, M. & Hüllermeier, E. ML-Plan: Automated machine learning via hierarchical planning. *Machine Learning* **107,** 1495–1515. ISSN: 1573-0565 (Sept. 1, 2018).

90. Chen, B., Wu, H., Mo, W., Chattopadhyay, I. & Lipson, H. *Autostacker: a compositional evolutionary learning system* in *Proceedings of the Genetic and Evolutionary Computation Conference* (Association for Computing Machinery, New York, NY, USA, July 2, 2018), 402–409. ISBN: 978-1-4503-5618-3.

91. LeDell, E. & Poirier, S. H2O AutoML: Scalable Automatic Machine Learning. *7th ICML Workshop on Automated Machine Learning (AutoML)* (July 2020).

92. Leclercq, M. *et al.* Large-Scale Automatic Feature Selection for Biomarker Discovery in High-Dimensional OMICs Data. *Frontiers in Genetics* **10,** 452. ISSN: 1664-8021 (2019).

93. Joseph, V. R. & Vakayil, A. SPlit: An Optimal Method for Data Splitting. *Technometrics* **64,** 166–176. ISSN: 0040-1706, 1537-2723 (Apr. 3, 2022).

94. Kohavi, R. *A study of cross-validation and bootstrap for accuracy estimation and model selection* in *Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2* (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, Aug. 20, 1995), 1137–1143. ISBN: 978-1-55860-363-9.

95. Zeng, X. & Martinez, T. R. Distribution-balanced stratified cross-validation for accuracy estimation. *Journal of Experimental & Theoretical Artificial Intelligence* **12,** 1–12. ISSN: 0952-813X (Jan. 1, 2000).

96. Sokolova, M., Japkowicz, N. & Szpakowicz, S. in *AI 2006: Advances in Artificial Intelligence* (eds Sattar, A. & Kang, B.-h.) red. by Hutchison, D. *et al.,* 1015–1021 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2006). ISBN: 978-3-540-49787-5 978-3-540-49788-2.

97. Hripcsak, G. & Rothschild, A. S. Agreement, the F-Measure, and Reliability in Information Retrieval. *Journal of the American Medical Informatics Association* **12,** 296–298. ISSN: 1067-5027 (May 1, 2005).

98. Baldi, P., Brunak, S., Chauvin, Y., Andersen, C. A. F. & Nielsen, H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **16,** 412–424. ISSN: 1367-4803 (May 1, 2000).

99. Chicco, D. & Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **21,** 6. ISSN: 1471-2164 (Jan. 2, 2020).

100. Bommert, A. M. *Integration of feature selection stability in model fitting* PhD thesis (Technische Universität Dortmund, 2020).

101. Nogueira, S., Sechidis, K. & Brown, G. On the Stability of Feature Selection Algorithms. *Journal of Machine Learning Research* **18,** 1–54 (2018).

102. Bommert, A. & Rahnenführer, J. *Adjusted Measures for Feature Selection Stability for Data Sets with Similar Features* in *Machine Learning, Optimization, and Data Science* (eds Nicosia, G. *et al.*) (Springer International Publishing, Cham, 2020), 203–214. ISBN: 978-3-030-64583-0.

103. Carletta, J. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics* **22** (ed Hirschberg, J.) 249–254 (1996).

104. Nogueira, S. & Brown, G. *Measuring the Stability of Feature Selection* in *Machine Learning and Knowledge Discovery in Databases* (eds Frasconi, P., Landwehr, N., Manco, G. & Vreeken, J.) (Springer International Publishing, Cham, 2016), 442–457. ISBN: 978-3-319-46227-1.

105. Backes, C. *et al.* GeneTrail—advanced gene set enrichment analysis. *Nucleic Acids Research* **35,** W186–W192. ISSN: 0305-1048 (suppl_2 July 1, 2007).

106. Geistlinger, L. *et al.* Toward a gold standard for benchmarking gene set enrichment analysis. *Briefings in Bioinformatics* **22,** 545–556. ISSN: 1477-4054 (Jan. 1, 2021).

107. Ayllon-Benitez, A., Bourqui, R., Thébault, P. & Mougin, F. GSAn: an alternative to enrichment analysis for annotating gene sets. *NAR Genomics and Bioinformatics* **2,** lqaa017. ISSN: 2631-9268 (June 1, 2020).

108. O'Donoghue, S. I. *et al.* Visualization of Biomedical Data. *Annual Review of Biomedical Data Science* **1,** 275–304. ISSN: 2574-3414 (Volume 1, 2018 July 20, 2018).

109. Midway, S. R. Principles of Effective Data Visualization. *Patterns* **1.** ISSN: 2666-3899 (Dec. 11, 2020).

110. Shneiderman, B. *The eyes have it: a task by data type taxonomy for information visualizations* in *Proceedings 1996 IEEE Symposium on Visual Languages* Proceedings 1996 IEEE Symposium on Visual Languages (Sept. 1996), 336–343.

111. Munzner, T. A Nested Model for Visualization Design and Validation. *IEEE Transactions on Visualization and Computer Graphics* **15,** 921–928. ISSN: 1941-0506 (Nov. 2009).

112. Mougin, F. *et al.* Visualizing omics and clinical data: Which challenges for dealing with their variety? *Methods. Comparison and Visualization Methods for High-Dimensional Biological Data* **132,** 3–18. ISSN: 1046-2023 (Jan. 1, 2018).

113. Mackinlay, J. Automating the design of graphical presentations of relational information. *ACM Trans. Graph.* **5,** 110–141. ISSN: 0730-0301 (Apr. 1, 1986).

114. Acharya, A. S., Prakash, A., Saxena, P. & Nigam, A. Sampling: why and how of it? *Indian Journal of Medical Specialities* **4.** ISSN: 09762892 (July 7, 2013).

115. Hepburn, A. C. *et al.* Identification of CNGB1 as a Predictor of Response to Neoadjuvant Chemotherapy in Muscle-Invasive Bladder Cancer. *Cancers* **13,** 3903. ISSN: 2072-6694 (Jan. 2021).

116. Zhao, S., Sun, J., Shimizu, K. & Kadota, K. Silhouette Scores for Arbitrary Defined Groups in Gene Expression Data and Insights into Differential Expression Results. *Biological Procedures Online* **20,** 5. ISSN: 1480-9222 (Mar. 1, 2018).

117. Sudo, M. *et al.* Long-term outcomes after surgical resection in patients with stage IV colorectal cancer: a retrospective study of 129 patients at a single institution. *World Journal of Surgical Oncology* **17,** 56. ISSN: 1477-7819 (Mar. 23, 2019).

118. Kim, S.-K. *et al.* A nineteen gene-based risk score classifier predicts prognosis of colorectal cancer patients. *Molecular Oncology* **8,** 1653–1666. ISSN: 1878-0261 (Dec. 2014).

119. Cui, H. *et al.* Identification of the key genes and pathways involved in the tumorigenesis and prognosis of kidney renal clear cell carcinoma. *Scientific Reports* **10,** 4271. ISSN: 2045-2322 (Mar. 6, 2020).

120. Wang, L. *et al.* Integrative Serum Metabolic Fingerprints Based Multi-Modal Platforms for Lung Adenocarcinoma Early Detection and Pulmonary Nodule Classification. *Advanced Science* **9,** 2203786. ISSN: 2198-3844 (2022).

121. Wang, M., Yue, S. & Yang, Z. Downregulation of PSAT1 inhibits cell proliferation and migration in uterine corpus endometrial carcinoma. *Scientific Reports* **13,** 4081. ISSN: 2045-2322 (Mar. 11, 2023).

122. Li, C. *et al.* Keratin 80 promotes migration and invasion of colorectal carcinoma by interacting with PRKDC via activating the AKT pathway. *Cell Death & Disease* **9,** 1–12. ISSN: 2041-4889 (Sept. 27, 2018).

123. Parikh, K. *et al.* Colonic epithelial cell diversity in health and inflammatory bowel disease. *Nature* **567,** 49–55. ISSN: 1476-4687 (Mar. 2019).

124. Zhang, M., Li, T., Zhu, J., Tuo, B. & Liu, X. Physiological and pathophysiological role of ion channels and transporters in the colorectum and colorectal cancer. *Journal of Cellular and Molecular Medicine* **24,** 9486–9494. ISSN: 1582-4934 (2020).

125. Jung, Y. *et al.* Clinical validation of colorectal cancer biomarkers identified from bioinformatics analysis of public expression data. *Clinical Cancer Research: An Official Journal of the American Association for Cancer Research* **17,** 700–709. ISSN: 1557-3265 (Feb. 15, 2011).

126. Fonseca, A. S. *et al.* ETV4 plays a role on the primary events during the adenoma-adenocarcinoma progression in colorectal cancer. *BMC Cancer* **21,** 207. ISSN: 1471-2407 (Mar. 1, 2021).

127. Kang, Y. H. *et al.* ESM-1 regulates cell growth and metastatic process through activation of NF-KB in colorectal cancer. *Cellular Signalling* **24,** 1940–1949. ISSN: 0898-6568 (Oct. 1, 2012).

128. Zhang, H. *et al.* Targeting Endothelial Cell-Specific Molecule 1 Protein in Cancer: A Promising Therapeutic Approach. *Frontiers in Oncology* **11.** ISSN: 2234-943X (2021).

129. Jia, H. *et al.* The LIM protein AJUBA promotes colorectal cancer cell survival through suppression of JAK1/STAT1/IFIT2 network. *Oncogene* **36,** 2655–2666. ISSN: 1476-5594 (May 2017).

130. Yang, D. *et al.* Smad1 promotes colorectal cancer cell migration through Ajuba transactivation. *Oncotarget* **8,** 110415–110425. ISSN: 1949-2553 (Nov. 30, 2017).

131. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics* **45,** 1113–1120. ISSN: 1546-1718 (Oct. 2013).

132. Gobin, E. *et al.* A pan-cancer perspective of matrix metalloproteases (MMP) gene expression profile and their diagnostic/prognostic potential. *BMC Cancer* **19,** 581. ISSN: 1471-2407 (June 14, 2019).

133. Fredriksson, N. J., Ny, L., Nilsson, J. A. & Larsson, E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nature Genetics* **46,** 1258–1263. ISSN: 1546-1718 (Dec. 2014).

134. Rosario, S. R. *et al.* Pan-cancer analysis of transcriptional metabolic dysregulation using The Cancer Genome Atlas. *Nature Communications* **9,** 5330. ISSN: 2041-1723 (Dec. 14, 2018).

135. Rokavec, M., Kaller, M., Horst, D. & Hermeking, H. Pan-cancer EMT-signature identifies RBM47 down-regulation during colorectal cancer progression. *Scientific Reports* **7,** 4687. ISSN: 2045-2322 (July 5, 2017).

136. Martínez-Jiménez, F. *et al.* Pan-cancer whole-genome comparison of primary and metastatic solid tumours. *Nature* **618,** 333–341. ISSN: 1476-4687 (June 2023).

137. Liot, S. *et al.* Loss of Tenascin-X expression during tumor progression: A new pan-cancer marker. *Matrix Biology Plus. Complexity of Matrix Phenotypes* **6-7,** 100021. ISSN: 2590-0285 (May 1, 2020).

138. Mihaylov, I., Kańduła, M., Krachunov, M. & Vassilev, D. A novel framework for horizontal and vertical data integration in cancer studies with application to survival time prediction models. *Biology Direct* **14,** 22. ISSN: 1745-6150 (Nov. 21, 2019).

139. Sharifi-Noghabi, H., Zolotareva, O., Collins, C. C. & Ester, M. MOLI: multi-omics late integration with deep neural networks for drug response prediction. *Bioinformatics* **35,** i501–i509. ISSN: 1367-4803 (July 15, 2019).

140. Julkunen, H. *et al.* Atlas of plasma NMR biomarkers for health and disease in 118,461 individuals from the UK Biobank. *Nature Communications* **14,** 604. ISSN: 2041-1723 (Feb. 3, 2023).

141. Simonetto, P., Auber, D. & Archambault, D. Fully Automatic Visualisation of Overlapping Sets. *Computer Graphics Forum* **28,** 967–974. ISSN: 1467-8659 (2009).

142. Harrison, P. W. *et al.* Ensembl 2024. *Nucleic Acids Research* **52,** D891–D899. ISSN: 0305-1048 (D1 Jan. 5, 2024).

143. Pietro, D. & J., T. Menu Positioning on Web Pages. Does it Matter? *International Journal of Advanced Computer Science and Applications* **6.** ISSN: 21565570, 2158107X (2015).

144. Claude, E., Leclercq, M., Thébault, P., Droit, A. & Uricaru, R. Optimizing hybrid ensemble feature selection strategies for transcriptomic biomarker discovery in complex diseases. *NAR Genomics and Bioinformatics* **6,** lqae079. ISSN: 2631-9268 (Sept. 1, 2024).

145. Donne, R. & Lujambio, A. The liver cancer immune microenvironment: Therapeutic implications for hepatocellular carcinoma. *Hepatology* **77,** 1773. ISSN: 0270-9139 (May 2023).

146. Llovet, J. M. *et al.* Hepatocellular carcinoma. *Nature Reviews Disease Primers* **7,** 1–28. ISSN: 2056-676X (Jan. 21, 2021).

147. Hernanda, P. Y. *et al.* SMAD4 exerts a tumor-promoting role in hepatocellular carcinoma. *Oncogene* **34,** 5055–5068. ISSN: 1476-5594 (Sept. 2015).

148. Zhang, J., Zhu, Y., Hu, L., Yan, F. & Chen, J. miR-494 induces EndMT and promotes the development of HCC (Hepatocellular Carcinoma) by targeting SIRT3/TGF-B/SMAD signaling pathway. *Scientific Reports* **9,** 7213. ISSN: 2045-2322 (May 10, 2019).

149. Zhang, L. *et al.* BMP signaling and its paradoxical effects in tumorigenesis and dissemination. *Oncotarget* **7,** 78206–78218. ISSN: 1949-2553 (Sept. 20, 2016).

150. Hyslop, S. R. *et al.* Targeting platelets for improved outcome in KRAS-driven lung adenocarcinoma. *Oncogene* **39,** 5177–5186. ISSN: 1476-5594 (July 2020).

151. Supernat, A. *et al.* Deregulation of RAD21 and RUNX1 expression in endometrial cancer. *Oncology Letters* **4,** 727–732. ISSN: 1792-1074 (Oct. 1, 2012).

152. Price, J. C. *et al.* Sequencing of Candidate Chromosome Instability Genes in Endometrial Cancers Reveals Somatic Mutations in ESCO1, CHTF18, and MRE11A. *PLOS ONE* **8,** e63313. ISSN: 1932-6203 (June 3, 2013).

153. Ma, S., Chen, Q., Li, X., Fu, J. & Zhao, L. UBE2C serves as a prognosis biomarker of uterine corpus endometrial carcinoma via promoting tumor migration and invasion. *Scientific Reports* **13,** 16899. ISSN: 2045-2322 (Oct. 6, 2023).

154. Plantureux, L., Crescence, L., Dignat-George, F., Panicot-Dubois, L. & Dubois, C. Effects of platelets on cancer progression. *Thrombosis Research. Papers and Abstracts of the 9th International Conference on Thrombosis and Hemostasis Issues in Cancer, April 13-15, 2018, Bergamo, Italy* **164,** S40–S47. ISSN: 0049-3848 (Apr. 1, 2018).

155. Egan, K. *et al.* Platelet Adhesion and Degranulation Induce Pro-Survival and Pro-Angiogenic Signalling in Ovarian Cancer Cells. *PLOS ONE* **6,** e26125. ISSN: 1932-6203 (Oct. 12, 2011).

Annexe A

Valeurs d'hyperparamètres des algorithmes d'apprentissage automatique

I Approche HEFS en mode standard

Catégorie	**Algorithme**	**Options**	**Grille de valeurs**
Bayes	A1DE	-F	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
	weka : Bayes.Av	-M	[0.1, 0.2, 0.5, 1, 2, 3, 4, 5]
	eragedNDepend	-W	True/False
	enceEstimators.	-S	[10, 20, 30, 40, 50, 60, 70, 80, 90, 100,
	A1DE		110, 120, 130, 140, 150]

	Bayes Network weka: bayes.BayesNet	-D -Q -E	True/False [default TAN, HillClimber [-P 1, -P 2, -P 3, -P 100000, -P 1 -R, -P 2 -R, -P 3 -R, -P 100000 -R], SimulatedAnnealing [-A 10 -D 0.001, -A 10 -D 0.5, -A 10 -D 0.999, -A 50 -D 0.001, -A 50 -D 0.5, -A 50 -D 0.999, -A 100 -D 0.001, -A 100 -D 0.5, -A 100 -D 0.999, -A 200 -D 0.001, -A 200 -D 0.5, -A 200 -D 0.999], K2 [-P 1, -P 2, -P 3, -P 100000, -P 1 -R, -P 2 -R, -P 3 -R, -P 100000 -R]] [BayesNetEstimator [-A 0, -A 0.5, -A 1], BMAEstimator [-A 0, -A 0.5, -A 1, -A 0 -k2, -A 0.5 -k2, -A 1 -k2], MultiNomialBMAEstimator [-A 0, -A 0.5, -A 1, -A 0 -k2, -A 0.5 -k2, -A 1 -k2], SimpleEstimator [-A 0, -A 0.5, -A 1]]
	Naive Bayes weka: NaiveBayes	-K -D	True/False True/False
Fonction	SVM weka: functions.SMO	-C -N -K	[0.001, 0.01, 0.1, 1, 10, 100] [0, 1, 2] [PolyKernel [-E 2, -E 3, -E 4, -E 5, -E10, -E20], NormalizedPolyKernel [-E 2, -E 3, -E 4, -E 5, -E 10, -E 20], RBFKernel [-G 0.01, -G 0.1, -G 1, -G 5, -G 10, -G 20, -G 30], Puk [-S 0.01, -S 0.1, -S 1, -S 5, -S 5, -S 10, -S 20, -S 30]]

Paresseux	kNN weka: lazy.IBk	-K	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]
		-I	True/False
		-F	True/False
		-E	True/False
		-X	True/False
		-A	[default LinearNNSearch, default KDTree, default FilteredNeighbourSearch, default CoverTree, default BallTree]
Arbre	C4.5 weka: trees.J48	-U	True/False
		-O	True/False
		-C	[0.25, 0.50, 0.75, False]
		-M	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
		-R	True/False
		-N	[2, 3, 4, 5, 6, 7, 8, 9, 10, False]
		-S	True/False
		-Q	[1, False]
	Random Forest weka: RandomForest	-P	[25, 50, 75, 100]
		-I	[100, 300, 500, 700, 900]
		-K	[0, 1, 2, 3, 4]
		-M	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
		-S	1
		-N	[0, 2, 3, 4, 5, 6, 7, 8, 9, 10]

	Simple CART	-S	1
	weka:	-M	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
	trees.SimpleCart	-N	[2, 3, 4, 5, 6, 7, 8, 9, 10]
		-U	True/False
		-H	True/False
		-A	True/False
		-C	[0.1, 0.2, 0.3, 0.4, 0.5]

Tableau A.1: Classificateurs d'apprentissage automatique sélectionnés pour l'approche HEFS en mode standard, leur nom de fonction weka associé et la grille d'hyperparamètres choisie.

II Approche HEFS en mode léger

Catégorie	**Algorithme**	**Options**	**Grille de valeurs**
Bayes	A1DE	-F	[1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
	weka : Bayes.Av	-M	[0.1, 0.2, 0.5, 1, 2, 3, 4, 5]
	eragedNDepend	-W	True/False
	enceEstimators.	-S	[10, 20, 30]
	A1DE		
	Naive Bayes	-K	True/False
	weka: Naive-	-D	True/False
	Bayes		
Fonction	SVM	-C	[0.001, 0.01, 0.1, 1, 10, 100]
	weka: func-	-N	[0, 1, 2]
	tions.SMO	-K	[NormalizedPolyKernel [-E 2, -E 3, -E 4, -E 5, -E 10, -E 20], RBFKernel [-G 0.01, -G 0.1, -G 1, -G 5, -G 10, -G 20, -G 30], Puk [-S 0.01, -S 0.1, -S 1, -S 5, -S 5, -S 10, -S 20, -S 30]]

Arbre	C4.5	-U	True/False
	weka: trees.J48	-O	True/False
		-C	[0.25, 0.50, 0.75, False]
		-M	[0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10]
		-R	True/False
		-N	[2, 3, 4, 5, 6, 7, 8, 9, 10, False]
		-S	True/False
		-Q	[1, False]
	Random Forest	-P	[50, 100]
	weka: Random-Forest	-I	[100, 300, 500]
		-K	[0, 1, 2, 3, 4]
		-M	[0, 2, 4, 6, 8, 10]
		-S	1
		-N	[0, 2, 4, 6, 8, 10]

Tableau A.2: Classificateurs d'apprentissage automatique sélectionnés pour l'approche HEFS en mode léger, leur nom de fonction weka associé et la grille d'hyperparamètres choisie.

Annexe B

Clusters d'échantillons pour la cohorte CRC de TCGA

I Clusters pour échantillons de stade IV en scénario de filtre par variance

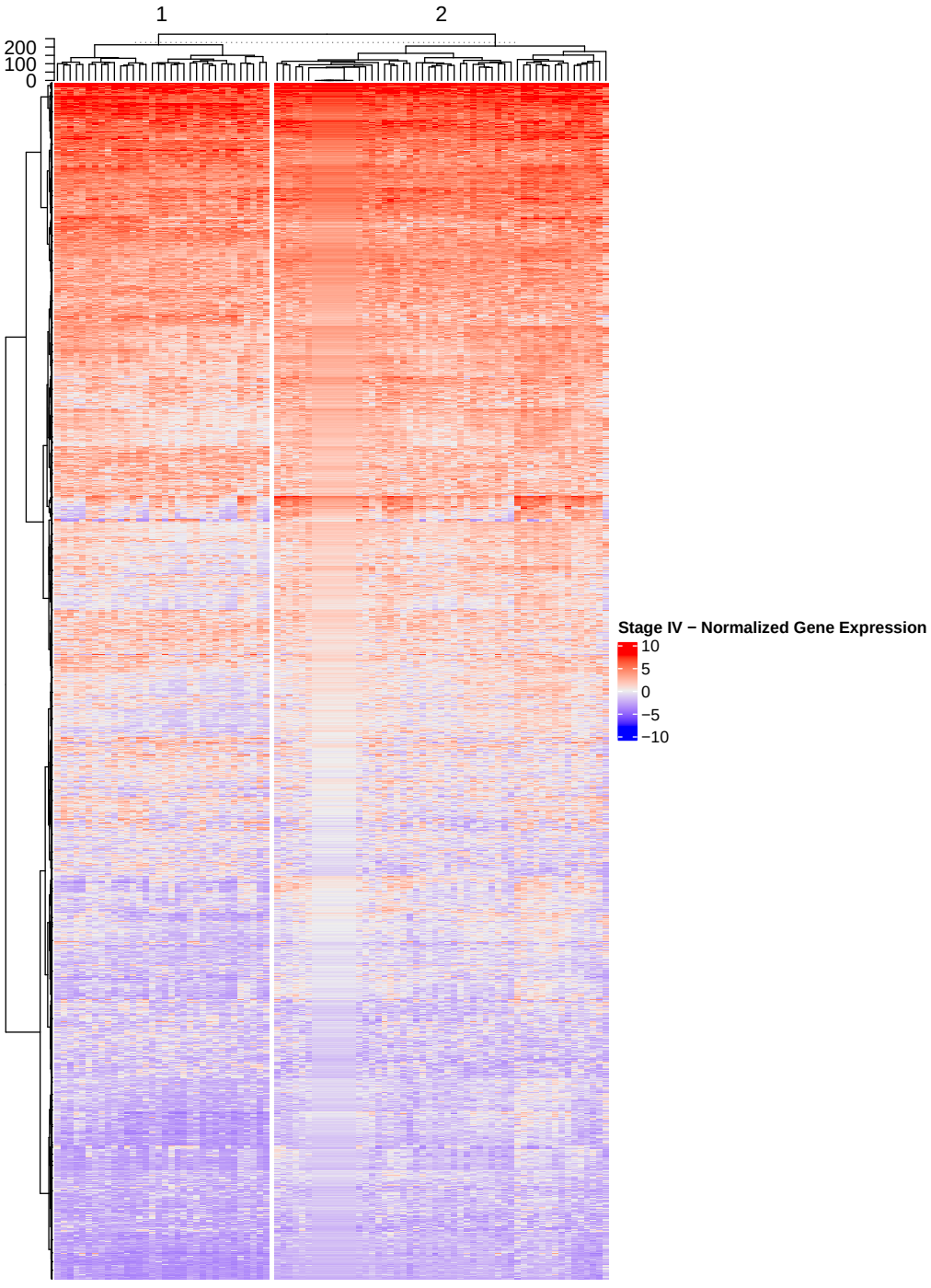

Figure B.1: Clustering hiérarchique d'échantillons de stade IV à partir de données CRC TCGA filtrées par variance en utilisant des données d'expression de gènes et en se basant sur la distance euclidienne et la méthode de Ward pour le lien. Les échantillons de patients sont représentés par la vue verticale tandis que les gènes composent la vue horizontale. Deux groupes d'échantillons sont définis.

II Clusters pour échantillons normaux en scénario de filtre par variance

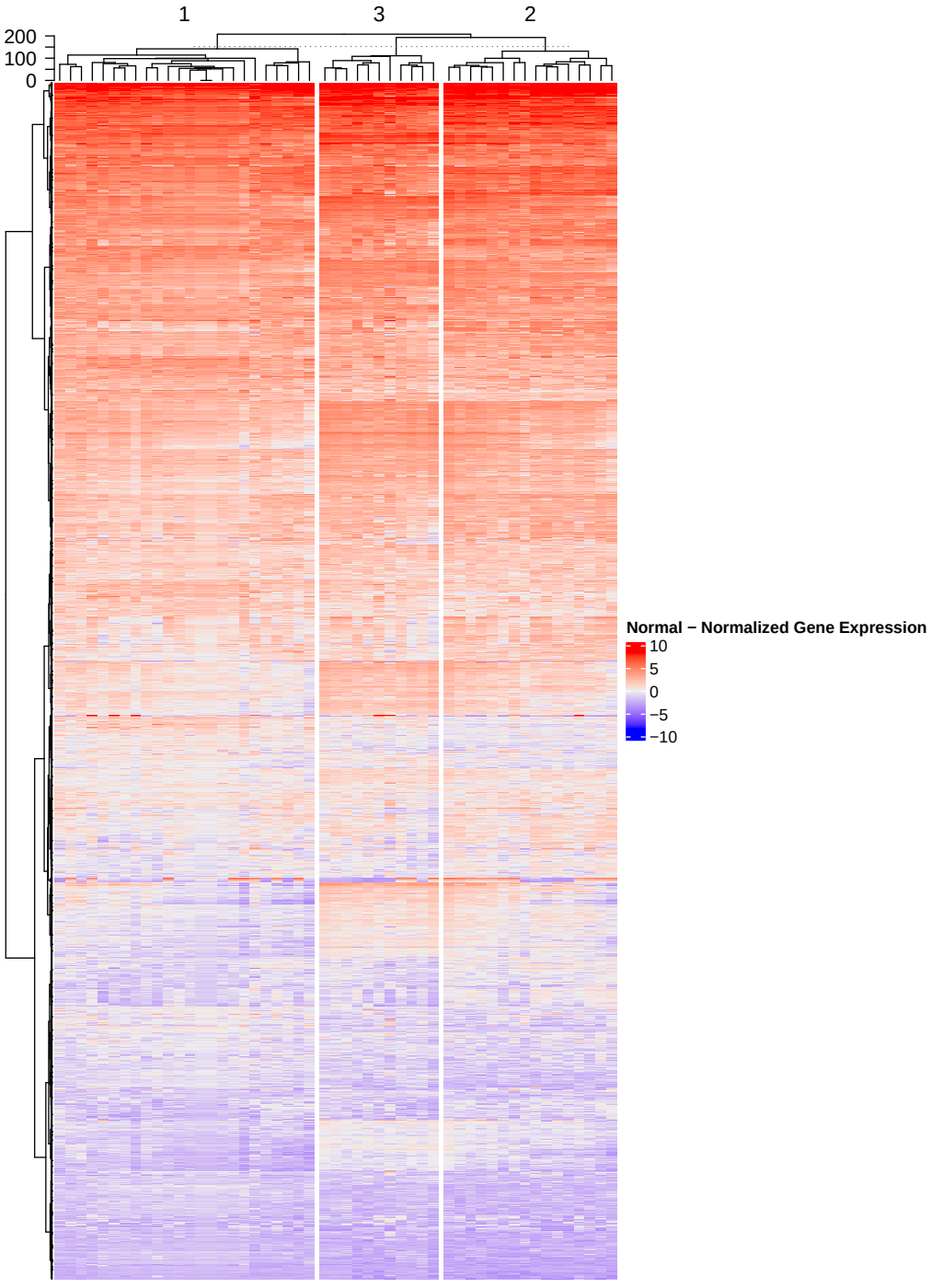

Figure B.2: Clustering hiérarchique d'échantillons normaux à partir de données CRC TCGA filtrées par variance en utilisant des données d'expression de gènes et en se basant sur la distance euclidienne et la méthode de Ward pour le lien. Les échantillons de patients sont représentés par la vue verticale tandis que les gènes composent la vue horizontale. Trois groupes d'échantillons sont définis.

III Clusters pour échantillons de stade IV en scénario de filtre par analyse DEG

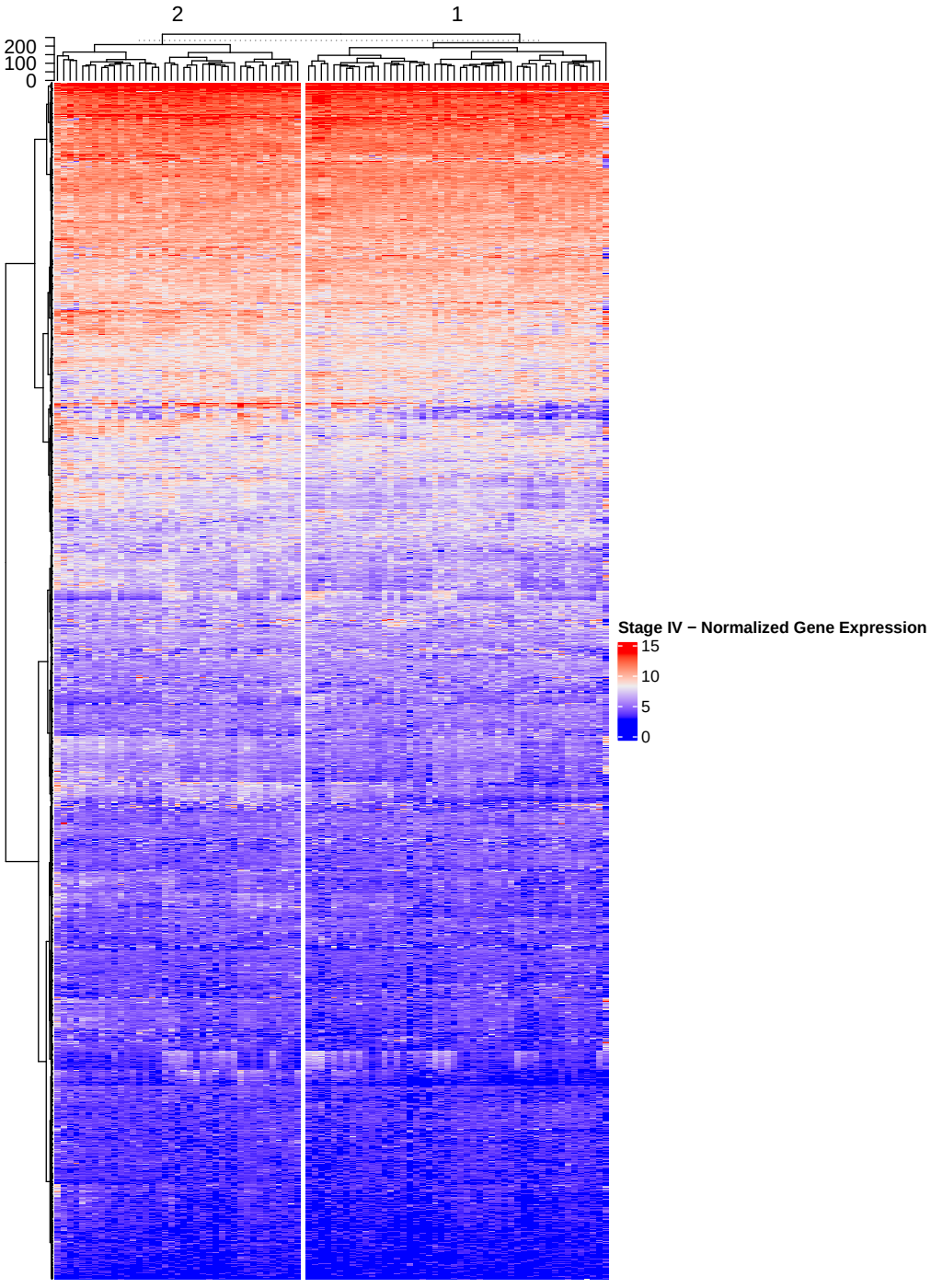

Figure B.3: Clustering hiérarchique d'échantillons de stade IV à partir de données CRC TCGA filtrées par analyse de gènes différentiellement exprimés en utilisant des données d'expression de gènes et en se basant sur la distance euclidienne et la méthode de Ward pour le lien. Les échantillons de patients sont représentés par la vue verticale tandis que les gènes composent la vue horizontale. Deux groupes d'échantillons sont définis.

IV Clusters pour échantillons normaux en scénario de filtre par analyse DEG

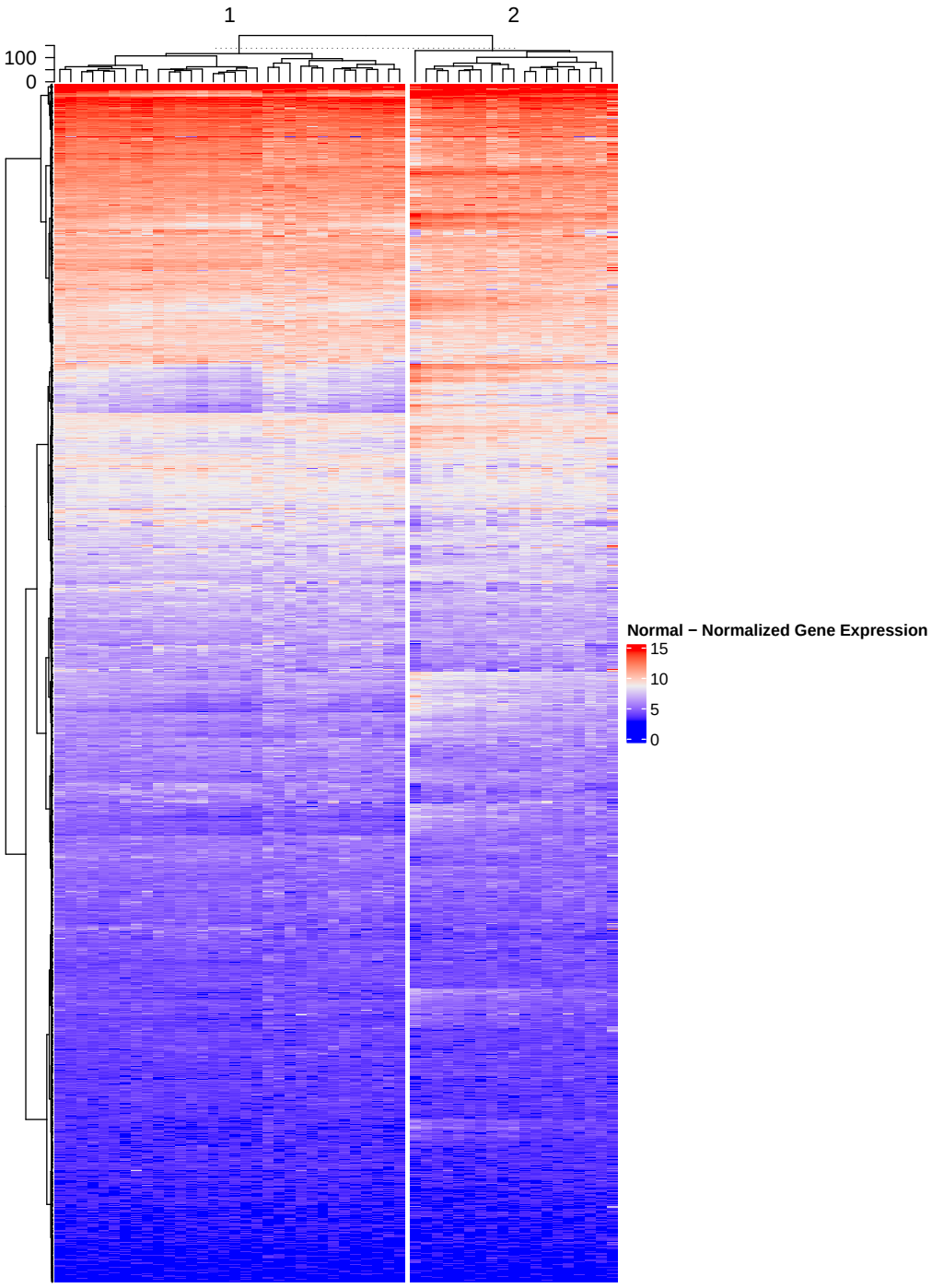

Figure B.4: Clustering hiérarchique d'échantillons normaux à partir de données CRC TCGA filtrées par analyse de gènes différentiellement exprimés en utilisant des données d'expression de gènes et en se basant sur la distance euclidienne et la méthode de Ward pour le lien. Les échantillons de patients sont représentés par la vue verticale tandis que les gènes composent la vue horizontale. Deux groupes d'échantillons sont définis.

Annexe C

Agrégation : les étapes du processus

I Première étape

Classificateur	Exéc. 1	Exéc. 2	Exéc. 3	Exéc. 4	Exéc. 5	Exéc. 6	Exéc. 7	Exéc. 8	Exéc. 9	Exéc. 10
A1DE	38	32	48	132	61	219	69	126	91	39
Bayes Network	38	32	48	44	61	57	69	30	34	39
C4.5	2291	1819	3466	3525	2851	3858	914	3009	2511	2619
kNN	38	32	48	44	61	57	69	30	34	39
Naive Bayes	38	32	48	44	61	57	69	30	34	39
Random Forest	1688	32	48	4621	148	4301	69	1384	778	4360
Simple Cart	264	32	48	2598	170	5093	69	1720	524	1924
SVM	1259	960	776	1448	778	880	863	945	792	976

Tableau C.1: Taille des signatures transitoires pour chacune des 10 exécutions d'échantillonnage obtenues à l'issue de la première étape d'agrégation pour le scénario DEG DB-S.

Classificateur	Exéc. 1	Exéc. 2	Exéc. 3	Exéc. 4	Exéc. 5	Exéc. 6	Exéc. 7	Exéc. 8	Exéc. 9	Exéc. 10
A1DE	302	36	69	165	30	54	103	117	29	288
Bayes Network	113	36	69	36	30	52	38	23	29	95
C4.5	3149	2806	3529	2530	2218	2149	2653	2779	439	2354
kNN	42	36	69	36	30	52	38	23	29	41
Naive Bayes	42	36	69	36	30	52	38	23	29	41
Random Forest	4569	137	1126	3916	3276	3544	3143	3473	54	4286
Simple Cart	5309	206	420	3929	3561	1872	2972	5007	73	5242
SVM	1030	512	1131	658	696	1012	1281	443	1126	894

Tableau C.2: Taille des signatures transitoires pour chacune des 10 exécutions d'échantillonnage obtenues à l'issue de la première étape d'agrégation pour le scénario DEG R-S.

Classificateur	Exéc. 1	Exéc. 2	Exéc. 3	Exéc. 4	Exéc. 5	Exéc. 6	Exéc. 7	Exéc. 8	Exéc. 9	Exéc. 10
A1DE	2554	3738	2553	1818	2666	2519	3725	1719	2591	1251
Bayes Network	1108	790	888	878	1407	1338	1426	409	814	143
C4.5	2660	2703	2782	2223	2799	2548	2909	2080	2727	1370
kNN	2903	2449	2251	2635	2502	2646	3026	1795	3290	1819
Naive Bayes	37	12	15	7	14	16	16	18	22	19
Random Forest	2591	2666	2634	1860	4110	2158	2818	1395	2959	1870
Simple Cart	3207	2132	3025	2037	2945	2696	3236	903	4120	2999
SVM	1732	1207	775	1033	880	1334	1793	738	1301	673

Tableau C.3: Taille des signatures transitoires pour chacune des 10 exécutions d'échantillonnage obtenues à l'issue de la première étape d'agrégation pour le scénario Var DB-S.

Classificateur	Exéc. 1	Exéc. 2	Exéc. 3	Exéc. 4	Exéc. 5	Exéc. 6	Exéc. 7	Exéc. 8	Exéc. 9	Exéc. 10
A1DE	4253	4012	2615	3541	2509	2901	2554	2752	4255	2678
Bayes Network	1691	1588	1007	1102	1185	1304	1299	745	1690	900
C4.5	4062	2776	2979	2626	4006	2459	2652	2674	2857	2849
kNN	3963	4144	2804	2487	4042	2854	1945	3485	4160	2726
Naive Bayes	48	91	29	31	16	13	9	25	43	16
Random Forest	3006	2756	2568	3543	3922	2618	3977	2914	4083	2772
Simple Cart	3336	4310	3211	4258	4205	4024	4285	4169	4389	3202
SVM	2024	1844	1337	1050	1697	1303	1130	1022	2016	1453

Tableau C.4: Taille des signatures transitoires pour chacune des 10 exécutions d'échantillonnage obtenues à l'issue de la première étape d'agrégation pour le scénario Var R-S.

II Deuxième étape

Exécution	Taille des signatures
1	38
2	32
3	48
4	44
5	61
6	57
7	69
8	30
9	34
10	39

Tableau C.5: Taille des signatures transitoires pour chacune des 10 exécutions d'échantillonnage obtenues à l'issue de la deuxième étape d'agrégation pour le scénario DEG DB-S.

Exécution	Taille des signatures
1	41
2	42
3	36
4	69
5	36
6	30
7	52
8	38
9	23
10	29

Tableau C.6: Taille des signatures transitoires pour chacune des 10 exécutions d'échantillonnage obtenues à l'issue de la deuxième étape d'agrégation pour le scénario DEG R-S.

Exécution	Taille des signatures
1	9
2	18
3	11
4	13
5	5
6	13
7	16
8	16
9	8
10	22

Tableau C.7: Taille des signatures transitoires pour chacune des 10 exécutions d'échantillonnage obtenues à l'issue de la deuxième étape d'agrégation pour le scénario Var DB-S.

Exécution	Taille des signatures
1	16
2	29
3	22
4	24
5	8
6	12
7	13
8	8
9	16
10	36

Tableau C.8: Taille des signatures transitoires pour chacune des 10 exécutions d'échantillonnage obtenues à l'issue de la deuxième étape d'agrégation pour le scénario Var R-S.