Les biomarqueurs émergents : intérêt et limites des études disponibles

Identification of new biomarkers: interest and limitations of available studies

Mots-clés : Biostatistique – Signature transcriptomique – Biais de sélection – Taille de l'échantillon.

Keywords: *Biostatistics – Gene expression profiling – Statistical bias – Sample size.*

P. Roy*, D. Maucort-Boulch*

L'arrivée des techniques d'analyse à haut débit représente une avancée technologique importante pour l'identification de nouveaux biomarqueurs diagnostiques et pronostiques. Cette avancée concerne l'ADN avec l'étude du génome, ses transcrits avec l'analyse du transcriptome, la traduction de ceux-ci avec l'analyse du protéome. L'analyse statistique de ces nouveaux biomarqueurs a permis l'émergence de signatures moléculaires correspondant à la combinaison d'un ensemble de biomarqueurs classiquement sélectionnés sur leurs propriétés discriminantes. Ces signatures permettent de caractériser les patients en termes de diagnostic ou de pronostic selon leur profil d'expression transcriptomique ou protéomique.

Nous nous limitons volontairement à l'étude du transcriptome compte tenu de son implication récente dans l'établissement du pronostic des patientes atteintes d'un cancer du sein, et présentons un certain nombre de notions méthodologiques de base adaptées aux études testant un très grand nombre d'hypothèses. Les références issues de la littérature sont données à titre d'illustration de ce chapitre essentiellement méthodologique.

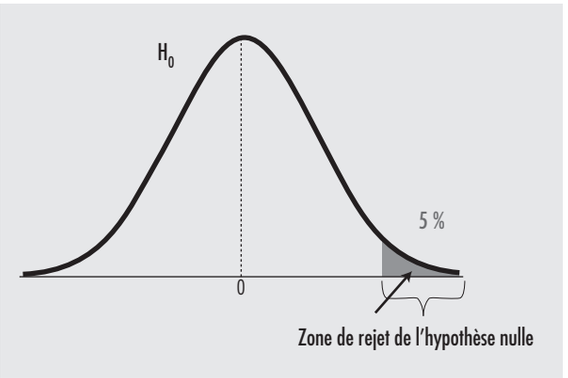
Les études pronostiques visent à identifier les patientes à haut risque de récidive ou de décès, afin de pouvoir proposer des traitements adaptés et identifier des sous-groupes éligibles pour de nouveaux essais thérapeutiques.

Approche classique : test d'une hypothèse unique

La méthodologie classique de la recherche consiste à poser une hypothèse unique, puis à la tester. Il s'agit par exemple de montrer que le niveau d'expression d'un gène candidat est plus élevé au moment du diagnostic chez les patientes qui présenteront une récidive dans les 5 ans, par rapport à celles qui en seront indemnes. Le niveau d'expression du gène candidat reflète son activité biologique, c'est-à-dire un mécanisme présentant une certaine variabilité. Cette variabilité s'observe aussi bien dans la population des femmes qui présenteront une récidive dans les 5 ans que dans celle des femmes qui en seront indemnes. Nous souhaitons pouvoir comparer les niveaux moyens d'expression de ce gène candidat dans ces deux populations, mais nous disposons seulement d'estimations de ces moyennes à partir de celles calculées sur les échantillons de l'étude. Nous disposons donc d'une estimation de la différence des moyennes des niveaux d'expression, le rapport de cette différence à son erreur-type fournissant une estimation de notre grandeur test.

Le test d'hypothèse est toujours celui de l'hypothèse nulle H_0, qui correspond dans le cas présent à celle d'une absence de différence d'expression moyenne du gène candidat entre les deux populations. Sous l'hypothèse nulle, la grandeur test fluctue autour de 0 (la différence des moyennes des deux populations est supposée nulle). L'estimation de la grandeur test calculée à partir des données de l'étude est alors comparée à sa distribution sous l'hypothèse nulle ***(figure 1)***. La distribution classiquement utilisée pour la comparaison ici proposée est une distribution de Student (notons que le choix de la distribution utilisée dépend de la comparaison effectuée). La probabilité d'observer une valeur au moins aussi éloignée de 0 que l'estimation de la grandeur test obtenue à l'issue de l'étude est la valeur du petit "p", le niveau de significativité du test. La valeur de "p" correspond à l'aire sous la courbe au-delà de la valeur estimée (l'aire totale sous la courbe vaut 1). Cette probabilité est comparée à une valeur de référence α,

Figure 1. *Distribution d'échantillonnage de la grandeur test sous H0 (zone de rejet définie dans le cadre d'un test unilatéral).*



* *Équipe Biostatistique-Santé, UMR CNRS 5558, université Lyon-1, service de biostatistique, hospices civils de Lyon, plate-forme d'analyse biostatistique/bio-informatique à haut débit, pôle Rhône-alpin de bio-informatique (PRABI).*

ou risque de première espèce, c'est-à-dire la probabilité de rejeter l'hypothèse nulle sous H_0, fréquemment fixée à 5 %. La valeur d'α permet de définir une valeur seuil de la grandeur test. Une valeur de "p" inférieure à α conduit à rejeter l'hypothèse nulle et à déclarer le test significatif.

La puissance du test est la probabilité de rejeter l'hypothèse nulle sous une hypothèse alternative H_1 particulière, réaliste, et fixée avant l'étude. Le calcul de puissance n'a de sens qu'a priori. Le risque de deuxième espèce, β, est la probabilité de ne pas rejeter H_0 quand cette hypothèse H_1 est vraie. La puissance vaut donc (1–β). Une analyse rapide de la ***figure 2*** souligne l'évolution opposée bien connue des risques de première et deuxième espèces : l'augmentation de la puissance se fait au prix d'une augmentation du risque de première espèce.

L'aire sous H_1 située à droite de la valeur seuil est la puissance, c'est-à-dire la probabilité de rejeter l'hypothèse nulle lorsque cette hypothèse alternative H_1 particulière est vraie. La puissance augmente lorsque l'hypothèse alternative s'écarte de l'hypothèse nulle. En augmentant la taille des études, les fluctuations des moyennes liées à l'échantillonnage diminuent. Il en découle une diminution des fluctuations de la grandeur test sous H_0 et sous H_1. La valeur seuil du test est déplacée vers la gauche, et l'aire sous la distribution de H_1 à droite du seuil (la puissance) augmente ***(figure 3)***. Le nombre de sujets nécessaires est l'effectif de l'étude permettant, pour un risque de première espèce et sous une hypothèse alternative spécifique fixée à l'avance, de disposer de la puissance nécessaire au rejet de l'hypothèse nulle.

Figure 2. *Fluctuation d'échantillonnage de la grandeur test sous H_0 et H_1.*

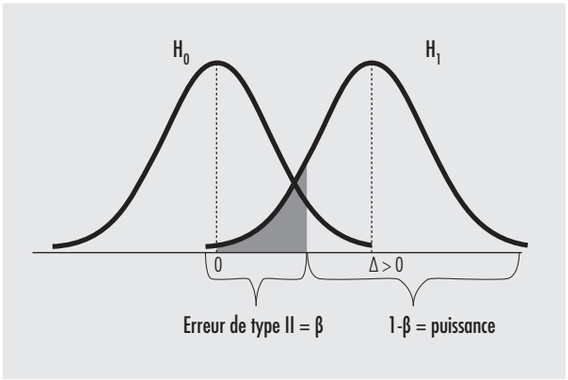
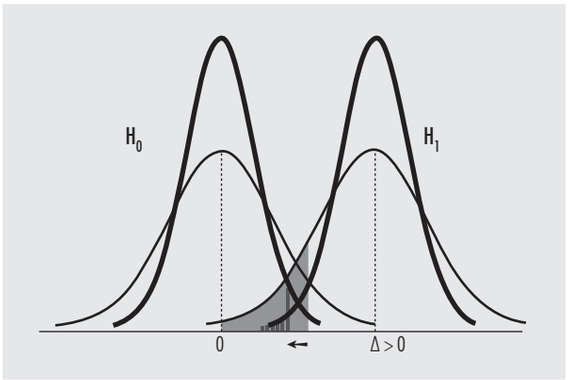


Figure 3. *Influence de la taille de l'échantillon sur la puissance du test.*



Ajustement de modèles et prédictions

La construction d'un modèle, d'un score pronostique ou d'une signature s'appuie sur des critères d'adéquation de celui-ci aux données de l'échantillon de travail sur lequel il a été construit. Le modèle (le score, la signature) retenu à l'issue de l'analyse est un modèle (un score, une signature) moyen(ne). L'application du modèle (du score, de la signature) à des patients qui n'ont pas contribué à sa construction pose la question de l'estimation de sa qualité prédictive. Il est fréquent que les qualités que semblait présenter un modèle sur l'échantillon de "travail" sur lequel il a été construit ne se retrouvent plus sur de nouveaux échantillons "tests" indépendants. On appelle "optimisme", ou "biais", cette surestimation des qualités prédictives du modèle. En l'absence de correction, cet optimisme conduirait à surestimer le pronostic de patients ayant un score favorable et à sous-estimer celui de patients ayant un score défavorable.

L'optimisme découle des fluctuations d'échantillonnage. Les variables dont les estimations ponctuelles des paramètres sont dans les parties extrêmes de leur distribution d'échantillonnage ont plus de chances d'être sélectionnées. Si on applique ce modèle à de nouvelles données et que l'on compare les valeurs prédites par le modèle aux valeurs observées, les prédictions basses sont trop basses et les prédictions élevées sont trop élevées. Bien que des méthodes statistiques sophistiquées de correction a posteriori de cet optimisme aient été proposées, comme les méthodes de "rétrécissement" *(shrinkage)* assimilables à la régression *ridge*, celles-ci sont imparfaites et on comprend aisément que la meilleure façon d'éviter l'optimisme est de limiter l'imprécision des modèles en augmentant la taille des études au moment de leur construction. Le biais d'optimisme découle de la sélection de variable opérée dans les études d'identification. Un échantillon de travail (étude initiale) de grande taille permet de limiter cet optimisme.

Particularités des analyses à haut débit

Identification de gènes différentiels

L'analyse à haut débit du transcriptome permet d'étudier simultanément le niveau d'expression de plusieurs milliers de gènes. Cette approche s'oppose radicalement à la démarche de recherche conduisant à tester une hypothèse unique.

Le ***tableau*** présente les résultats des m tests statistiques effectués correspondant aux m mesures d'expression des gènes analysés à l'aide de la puce à ADN. Parmi ces m tests, m_0 correspond à des gènes non différentiels (dont les moyennes des niveaux d'expression sont égales dans les populations comparées) et m_1 à des gènes

Tableau. *Répartition des résultats des différents tests effectués.*

		Décision		
		H_0 non rejetée	**H_0 rejetée**	
Réalité	H_0 vraie	U	V	m_0
	H_0 fausse	T	S	m_1
		m-R	R	m

différentiels (les moyennes des niveaux d'expression diffèrent entre les populations comparées). Alors que m_0 et m_1 sont inconnus, R correspond au nombre de tests conduisant au rejet de l'hypothèse nulle, c'est-à-dire au nombre de gènes déclarés différentiels à l'issue de l'analyse. La nécessité de contrôler le risque de première espèce en situation de tests multiples apparaît rapidement. Dans le cas simple où seuls dix gènes indépendants sont testés, si le risque de première espèce retenu pour chacun des tests est $\alpha = 5\,\%$, sous l'hypothèse nulle composite H_0^C qu'aucun de ces gènes ne soit associé au pronostic, la probabilité de déclarer à tort au moins l'un de ces dix gènes comme un gène différentiel est déjà de 40 % : $\Pr(V \geq 1 | H_0^C) = 1 - (1-\alpha)^{10} = 1 - (0{,}95)^{10} = 0{,}401$.

Le contrôle de la probabilité d'avoir au moins un faux positif, ou *Family Wise Error Rate* :

$$FWER = p(V \geq 1)$$

est très vite apparu moins intéressant que le contrôle de la proportion d'erreurs de première espèce attendues parmi l'ensemble des hypothèses rejetées, ou *False Discovery Rate* :

$$FDR = E(Q)$$

$$Q = \begin{cases} V/R, \text{ if } R > 0 \\ 0, \text{ if } R = 0 \end{cases}$$

L'analyse de la puissance en situation de test multiple concerne les m_1 gènes différentiels que vise à identifier l'étude. La probabilité de rater au moins un gène différentiel, ou *Family type-II error rate* :

$$\beta_F = p(T \geq 1) = 1 - p(T = 0)$$

apparaît moins intéressante que l'étude de l'espérance de la proportion des gènes différentiels détectés, conduisant à définir différemment l'erreur de seconde espèce et la puissance :

$$\text{Puissance} = \frac{E(S)}{m_1} = 1 - \beta_1$$

Les travaux méthodologiques ont tout d'abord concerné le risque de première espèce, et en particulier le contrôle du FDR *(1-3)*. Les publications sur la puissance présentant des estimations du nombre de sujets à inclure dans les études sont plus récentes *(4-5)*. Ces travaux permettent de retrouver, dans le contexte d'une analyse multitest, la relation entre les risques de première et seconde espèces vue précédemment, l'augmentation de la puissance ayant comme prix celle du FDR. Si on retrouve l'influence classique de l'écart entre l'hypothèse nulle et l'hypothèse alternative, l'influence de la proportion (inconnue) de gènes différentiels (m_1/m) mérite particulièrement d'être soulignée. Pour un nombre total de gènes fixé, lorsque cette proportion diminue, le maintien du contrôle du FDR au niveau prédéfini rend nécessaire d'augmenter la valeur seuil de chaque test. Cette augmentation s'accompagne d'une chute de puissance. De façon similaire, lorsque le nombre de gènes réellement différentiels m_1 est fixe, pour un même niveau de contrôle du FDR, l'utilisation de puces de taille plus importante aura pour principal effet d'entraîner une chute de puissance *(5)*.

Identification des signatures

Une analyse indépendante des données de sept études visant à identifier les signatures pronostiques de différents cancers a mis en évidence les conséquences du manque de puissance de la majorité des études *(6)*. Les auteurs ont utilisé une méthode de validation interne par rééchantillonnage et ont analysé les propriétés prédictives des signatures successivement élaborées en les appliquant chaque fois aux patients qui n'avaient pas contribué à leur définition. Pour une étude donnée, la liste des gènes identifiés pour construire le prédicteur était très instable d'un échantillon à l'autre. Les qualités prédictives de la signature ainsi établie augmentaient avec la taille de l'échantillon de travail. Les résultats initialement publiés des analyses de ces études étaient plus optimistes que ceux retrouvés après application de la procédure de validation interne utilisée. Ainsi, pour 5 des 7 études comparées, les signatures construites ne prédisaient pas mieux le pronostic des patients que le hasard *(6)*. Chez les patientes de moins de 55 ans atteintes d'un cancer du sein pN0, les signatures reconstruites à partir des données de l'étude de L.J. Van't Veer et al. *(7, 8)* prédisaient mieux le risque à 5 ans de métastases à distance que le hasard.

La signature de 70 gènes d'Amsterdam issue de l'étude de L.J. Van't Veer permet d'illustrer la question de l'optimisme des modèles. L'odds-ratio caractérisant le risque à 5 ans de récidive métastatique à distance estimé sur les données à l'origine de la construction de la signature était de 28 (IC_{95} : 7-107 ; $p = 1 \times 10^{-8}$). La mise en œuvre d'une méthode de correction de l'optimisme par validation interne *(leave one out)* permettait de retrouver un odds-ratio de 15 (IC_{95} : 4-56 ; $p = 4{,}1 \times 10^{-4}$), alors que la valeur retrouvée dans l'étude de validation externe de M. Buyse et al. *(9)* était de 2,32 (IC_{95} : 1,35-4).

À partir des données de cette étude, L. Ein-Dor et al. *(10, 11)* montrent que plusieurs signatures ayant des propriétés prédictives similaires à la signature d'Amsterdam peuvent être proposées sans partager de gènes avec cette dernière. Trois explications sont avancées : il existe un nombre très important de gènes dont le niveau d'expression est associé au pronostic ; le niveau de cette association varie peu entre les différents gènes identifiés ; la taille limitée des échantillons de travail explique la magnitude des fluctuations d'échantillonnage. Les auteurs soulignent que seules les études incluant plusieurs milliers de patients permettent d'identifier des signatures ayant plus de 50 % de gènes en commun *(11)*. Cette stabilité des signatures pronostiques est intéressante si on a de bonnes raisons de croire que les signatures les plus stables sont les plus robustes, et donc celles ayant les meilleurs propriétés prédictives.

Discussion

Les techniques d'analyse à haut débit représentent une avancée technologique majeure pour l'identification de nouveaux biomarqueurs diagnostiques et pronostiques. La mise en œuvre de ces techniques nécessite une méthodologie rigoureuse et des plans expérimentaux adaptés. Alors que l'étude de la variabilité technique de ces méthodes a été largement abordée, la prise en compte de la variabilité biologique, omniprésente en médecine, a été sous-estimée.

La question de la dimension des données a fait l'objet de nombreux travaux et concerne autant l'identification des gènes différentiels que celle des signatures. Les premiers prédicteurs de classe visaient à identifier des entités diagnostiques bien distinctes, et donc aisément distinguables *(12)*. On comprend aisément que l'identification de groupes pronostiques soit plus difficile, et l'élaboration de signatures pronostiques plus complexe. Une connaissance approfondie de la structure des données d'expression analysées

devrait permettre d'optimiser le choix des méthodes d'analyse statistique utilisées *(13)*, mais bien plus que le choix de la méthode statistique utilisée, l'une des questions essentielles est celle de la puissance *(4, 5)*.

Si les possibilités exploratoires des techniques à haut débit sont prometteuses, elles comprennent un certain nombre de limites. La première concerne la puissance, c'est-à-dire l'aptitude à identifier des gènes à expression différentielle, qui diminue avec le nombre d'hypothèses testées. La taille de l'étude initiale est déterminante. Les résultats disponibles à ce jour sont issus d'études de petite taille qui, bien qu'apportant des informations essentielles, n'en restent pas moins des études pilotes. Imaginons que l'expression de 300 gènes soit liée au pronostic du cancer, et que ces gènes soient classés selon leurs contributions pronostiques respectives. Deux études de petite taille n'identifieront pas systématiquement les gènes de premier rang, et le niveau de recouvrement des signatures identifiées dépendra de la taille respective des études *(11)*. Il est illusoire de mettre en place des études de validation coûteuses si l'étude initiale n'a pas la puissance nécessaire à l'identification des biomarqueurs les plus pertinents.

Une autre limite est celle de l'optimisme des études, qui est également lié à la taille de l'étude initiale. Cet optimisme pourrait conduire à surestimer la contribution pronostique des marqueurs issus des techniques à haut débit. Les marqueurs pronostiques clinicobiologiques usuels sont largement validés, souvent par de nombreuses études de taille importante. Leur valeur pronostique est connue avec précision. Les simulations de C. Truntzer et al. *(14)* ont montré que l'analyse conjointe de variables pronostiques clinicobiologiques validées et l'analyse exploratoire du transcriptome pouvaient conduire à surestimer la contribution des variables transciptomiques au pronostic lorsque les études initiales étaient de taille insuffisante.

Conclusion

Les techniques d'analyse à haut débit sont prometteuses pour identifier de nouveaux biomarqueurs pronostiques. Le potentiel d'investigation disponible ne doit pas pour autant nous faire oublier un concept statistique de base : la combinaison des niveaux d'expression d'un grand nombre de gènes d'une patiente donnée a de grandes chances d'être unique ou presque, justifiant l'utilisation du terme "signature". Mais le pronostic de cette patiente sera prédit en comparant sa signature individuelle aux signatures moyennes des groupes de patientes de bon et mauvais pronostic, ou en utilisant une fonction de classification construite pour distinguer des signatures moyennes.

Les travaux déjà publiés sont très prometteurs. Ils soulignent également les progrès à faire pour bénéficier davantage encore des retombées potentielles des méthodes à haut débit. La planification d'études d'identification de taille importante permettant de limiter l'optimisme des modèles (des scores, des signatures) et de disposer d'une puissance adéquate pour identifier les gènes les plus pertinents constitue la prochaine étape. ■

Références bibliographiques

[1] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Statist Soc Ser B 1995;5:289-300.

[2] Yekutieli D, Benjamini Y. A resampling based false discovery rate controlling multiple test procedure. J Statist Plann Inference 1999;82:171-96.

[3] Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann Stat 2002;29:1165-88.

[4] Mei-Ling Ting Lee, Whitmore GA. Power and sample size for DNA microarray studies. Stat Med 2002;21:3543-70.

[5] Pawitan Y, Michiels ST, Koscielny S, Gusnanto A, Ploner A. False discovery rate, sensitivity and sample size for microarray studies. Bioinformatics 2006;21:3017-24.

[6] Michiels S, Koscielny S, Hill C. Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet 2005;365:488-92.

[7] Van't Veer LJ, Dai H, van de Vijver MJ et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415:530-6.

[8] Van de Vijver MJ, He YD, van't Veer LJ et al. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med 2002;347:1999-2009.

[9] Buyse M, Loi S, van't Veer L et al. Validation and clinical utility of a 70-gene prognostic signature for Women With Node-Negative Breast cancer. J Natl Cancer Inst 2006;98:1183-92.

[10] Ein-Dor L, Kela I, Getz G, Givol D, Domany E. Outcome signature genes in breast cancer: is there a unique set? Bioinformatics 2005;21:171-8.

[11] Ein-Dor L, Zuk O, Domany E. Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer. PNAS 2006;103:5923-8.

[12] Golub T, Slonim D, Tamayo P. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science 1999;286:531-7.

[13] Truntzer C, Mercier C, Estève J, Gautier C, Roy P. Importance of data structure in comparing two dimension reduction methods for classification of microarray gene expression data. BMC Bioinformatics 2007;8:90.

[14] Truntzer C, Maucort-Boulch D, Roy P. Comparative optimism in models involving both classical clinical and gene expression information. BMC Bioinformatics 2008;9:434.