Tests multiples pour la comparaison des probabilités de survenue d'une infidélité de transcription dans des ARNm sains et cancéreux

Olivier Collignon, Marie Brulliard, Jean-Marie Monnez, Pierre Vallois, Benoit Thouvenot, Sandrine Jacquenet, Virginie Ogier, Olivier Roitel

▶ **To cite this version:**

Olivier Collignon, Marie Brulliard, Jean-Marie Monnez, Pierre Vallois, Benoit Thouvenot, et al.. Tests multiples pour la comparaison des probabilités de survenue d'une infidélité de transcription dans des ARNm sains et cancéreux. 42èmes Journées de Statistique, 2010, Marseille, France, France. inria-00494733

HAL Id: inria-00494733
https://inria.hal.science/inria-00494733

Submitted on 24 Jun 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Tests multiples pour la comparaison des probabilités de survenue d'une infidélité de transcription dans des ARNm sains et cancéreux

Collignon Olivier[1], Brulliard Marie[1], Monnez Jean-Marie[2], Vallois Pierre[2], Thouvenot Benoit[1], Jacquenet Sandrine[1], Ogier Virginie[1], Roitel Olivier[1], Bihain Bernard E[1].

(1) Genclis SAS, 15 rue du bois de la Champelle, 54500 Vandoeuvre-lès-Nancy,
(2) Institut Elie Cartan de Nancy, Université Henri Poincaré, Boulevard des Aiguillettes, 54500 Vandoeuvre-lès-Nancy

RÉSUMÉ

L'implication d'erreurs de transcription dans l'hétérogénéité du cancer avait jusqu'alors été peu considérée. En effet, la transcription est supposée fidèle et contrôlée par un système complexe de vérification. Afin d'étudier l'hétérogénéité des séquences d'ARNm issus de tissus sains et cancéreux de 17 gènes d'intérêt, les probabilités de survenue d'une substitution de base ont été comparées à chaque position des séquences des transcrits à l'aide d'une procédure de tests multiples. Pour cela, les séquences Expressed Sequences Tags, qui sont des copies partielles des ARNm d'un gène, ont été utilisées et un modèle prenant en compte l'erreur de séquençage inhérente à ces données a été proposé. Enfin, l'estimateur Location Based Estimator du nombre moyen de tests faux positifs a été étendu au cas de statistiques de tests discrètes. Cette étude préliminaire a ainsi permis de mettre en évidence les positions des ARNm plus fréquemment sujettes à des substitutions dans les tissus cancéreux que dans les tissus sains et d'introduire la notion d'infidélité de transcription chez l'Homme.
Mots-clés : *procédure de tests multiples, Location Based Estimator, ARNm, Expressed Sequences Tags, transcription, cancer, erreur de séquençage*

ABSTRACT

The possibility that transcription might contribute to molecular heterogeneity of cancer has thus far not been considered. Transcription is indeed supposed to be faithful and controled by a complex error checking system. In order to study normal and cancer mRNA heterogeneity of 17 genes of interest, the probabilities of base substitution were compared along the sequences using a multiple testing procedure. To do this, the Expressed Sequences Tags, which are partial copies of gene mRNA, were analyzed and a model taking sequencing errors into account was proposed. Finaly the Location Based Estimator of the mean number of false positive tests was extended to discrete test statistics. This

preliminary study allowed to highlight positions more frequently subject to base substitutions in cancer than in normal tissues and to introduce the notion of human transcription infidelity.
Keywords : *multiple testing procedure, Location Based Estimator, mRNA, Expressed Sequences Tags, transcription, cancer, sequencing error*

1 Données du problème

L'ARNm est une molécule caractérisée par son instabilité. C'est pourquoi les études reposant sur l'analyse des séquences d'un grand nombre d'ARNm nécessitent la préparation en laboratoire d'une autre molécule plus stable : l'**ADN complémentaire (ADNc)**. Les EST sont des séquences de nucléotides correspondant à de courts fragments d'ADNc. En résumé, ce sont des copies partielles des séquences des ARNm d'un gène présents dans un tissu. La base de données publique **dbEST** réunit aujourd'hui plus de 6,1 millions d'EST d'origine humaine, contenus dans des fichiers informatiques. Par ailleurs, une base d'un ADNc peut être mal lue et remplacée par une autre base lors du séquençage des EST. Ainsi, on peut observer en pratique sur l'EST une base différente de celle existant en réalité sur l'ARNm. Ce phénomène est appelé **erreur de séquençage**. De plus, une autre banque de données de NCBI appelée *RefSeq* fournit pour chaque gène une unique **séquence de référence**, qui est d'excellente qualité, représentant le brin d'ARNm du gène en question.
Le but de cette étude est de tester s'il existe une différence significative entre les fréquences de survenue d'une substitution à une position fixée d'un ARNm cancéreux et d'un ARNm sain. Les EST contenant l'information des ARNm - aux erreurs de séquençage près - , une substitution de certains nucléotides de la séquence d'ARNm d'un gène devrait être détectable dans les EST.

2 Analyse statistique

2.1 Formalisation du problème

Pour chaque gène, l'ensemble des EST de dbEST est aligné sur la séquence de référence à l'aide de l'outil informatique MegaBLAST 2.2.13 [1] ; les EST du gène sont ensuite séparés en deux blocs, en fonction de l'origine du tissu : cancer ou sain. A une position i fixée de la séquence d'ARNm du gène, on souhaite comparer les probabilités de survenue d'une substitution de nucléotides. On considère donc le test :

$$\begin{cases} H_0 : p_{1\bar{B}} = p_{2\bar{B}} \\ H_1 : p_{1\bar{B}} \neq p_{2\bar{B}} \end{cases} \quad (1)$$

où $p_{1\bar{B}}$ (resp. $p_{2\bar{B}}$) est la probabilité qu'un ARNm cancéreux (resp. sain) ait en position i l'une des bases différentes de B qui devrait se trouver à cette position. Les EST étant entachés d'erreurs de séquençage, on ne peut réaliser que le test :

$$\begin{cases} H_0 : q_{1\bar{B}} = q_{2\bar{B}} \\ H_1 : q_{1\bar{B}} \neq q_{2\bar{B}} \end{cases} \tag{2}$$

où $q_{1\bar{B}}$ (resp. $q_{2\bar{B}}$) est la probabilité de ne pas lire la base B sur un EST cancéreux (resp. sain) en position i.
Un modèle est alors proposé afin de prendre en compte ces erreurs de séquençage dans la méthode de comparaison des probabilités de survenue d'infidélité de transcription dans les deux types de tissus. Pour cela trois hypothèses sont émises :

1. (E_1) La probabilité ϵ d'avoir une erreur de séquençage ne dépend pas de l'état sain/cancer,
2. (E_2) La probabilité ϵ d'avoir une erreur de séquençage ne dépend pas de la véritable base B à la position d'intérêt,
3. (E_3) Une base B touchée par une erreur de séquençage est indifféremment remplacée par l'une ou l'autre des trois bases de $\bar{B}$ avec la même probabilité $\frac{\epsilon}{3}$.

Sous ces hypothèses, on montre ainsi que pour $j = 1, 2$:

$$q_{j\bar{B}} = (1 - \frac{4}{3}\epsilon)p_{j\bar{B}} + \epsilon. \tag{3}$$

En conséquence, les tests (1) et (2) sont équivalents. De ce fait, sous les hypothèses $(E_1), (E_2)$ et (E_3), il est équivalent de comparer les fréquences d'occurrence de substitution sur les EST (*i.e* compte tenu des erreurs de séquençage) et sur les ARNm.

2.2 Test de comparaison de probabilités

Soit B la base de la séquence de référence à une position donnée. A cette position sont alignés sur la séquence de référence n_1 EST cancéreux et n_2 EST sains (Figure 1). Soit alors le tableau de contingence :

	B	$\overline{B}$	Somme
Cancéreux	$n_1 - k_1$	k_1	n_1
Sains	$n_2 - k_2$	k_2	n_2
Somme	m_1	m_2	n

Si les conditions $n > 70,\ \frac{n_i m_j}{n} > 5,\ i = 1, 2,\ j = 1, 2$, sont vérifiées, le test (2) peut être réalisé avec un test du χ^2. Sinon, le test exact de Fisher est utilisé [2].

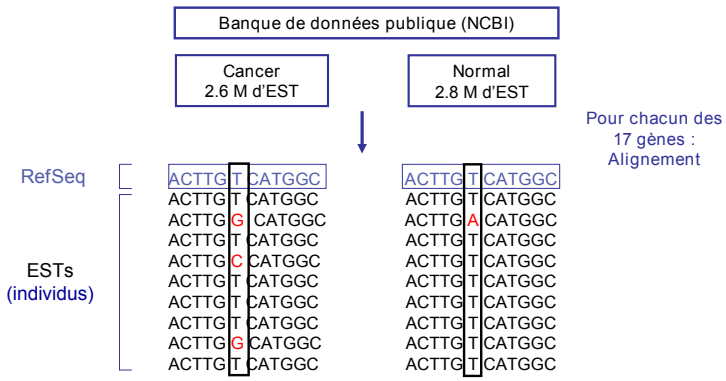

FIG. 1 – *Alignement des EST sur la séquence de référence*

2.3 Test multiples

En utilisant le test adapté, nous sommes donc en mesure de réaliser le test (2) à chacune des positions de la séquence de référence de chaque gène. On dispose pour cela d'une suite de statistiques de tests indépendantes $\{T_i\}_{1,\dots,n}$ avec lesquelles sont réalisés m tests au seuil α. Notons $\{t_i\}_{1,\dots,n}$ leurs réalisations respectives et $p_i = P(T_i \geq t_i \mid H_0)$ les p-values correspondantes. Les p_i sont donc des réalisations de variables aléatoires indépendantes P_i. On note enfin $V(\alpha)$ la variable aléatoire ayant pour réalisation le nombre de tests faux positifs au seuil α, *i.e.* pour lesquels on a rejeté H_0 alors que H_0 est vraie, et m_0 le nombre (inconnu) de tests pour lesquels H_0 est effectivement vraie. La connaissance d'une estimation de $\pi_0 = \frac{m_0}{m}$ nous permettrait d'avoir une estimation du nombre moyen de faux positifs, $E[V(\alpha)] = m_0\alpha$.

2.3.1 Location Based Estimator pour une statistique absolument continue

Supposons que les T_i soient absolument continues. La méthode LBE (Location Based Estimator) permet d'obtenir aisément un estimateur du nombre moyen de faux positifs (avec biais positif). Dalmasso *et al.* [3] ont ainsi montré que $\pi_0 \leq 2E[P]$. Ce résultat s'appuie sur le fait que sous H_0, la $p-$value P suit une loi $\mathcal{U}_{[0,1]}$.
Un estimateur du nombre moyen de faux positifs avec biais positif est donc donné par $2\alpha \sum_{i=1}^{m} P_i$.

2.3.2 Location Based Estimator pour une statistique discrète

Supposons que $T_i = T$ soit une variable aléatoire discrète. Dans ce cas, la $p-$value P correspondante ne suit pas une loi $\mathcal{U}_{[0,1]}$. Soit $I = \{t_i, i \in I\}$, $t_l < t_m$ pour $l < m$ l'ensemble

des modalités de la variable aléatoire discrète T. Calculons $E_0[P] = E[P|H_0]$.

$$\begin{aligned} E[P|H_0] &= 1 - E[F_0(T)|H_0] \\ &= 1 - \sum_{i\in I} F_0(t_i)P(T = t_i|H_0) \\ &= 1 - \sum_{i\in I}\sum_{l<i} P(T = t_l|H_0)P(T = t_i|H_0). \end{aligned} \tag{4}$$

Or

$$(\sum_{i\in I} P(T = t_i|H_0))^2 = 1 = \sum_{i\in I} P^2(T = t_i|H_0) + 2\sum_{i\in I}\sum_{l<i} P(T = t_l|H_0)P(T = t_i|H_0). \tag{5}$$

Donc

$$\begin{aligned} E[P|H_0] &= 1 + \frac{1}{2}(\sum_{i\in I} P^2(T = t_i|H_0) - 1) \\ &= \frac{1}{2} + \frac{1}{2}\sum_{i\in I} P^2(T = t_i|H_0) > \frac{1}{2}. \end{aligned} \tag{6}$$

En écrivant que :

$$E[P] = E[P|H_0]P(H_0) + E[P|H_1]P(H_1) \geq E[P|H_0]P(H_0) > \frac{1}{2}\frac{m_0}{m}, \tag{7}$$

on obtient que $2\alpha\sum_{i=1}^{m} P_i$ est un estimateur avec biais positif de $m_0\alpha$, nombre moyen de faux positifs. Le résultat obtenu dans la section précédente reste donc valable dans le cas d'une statistique de test discrète.

3 Résultats

Dans le tableau 1 sont présentés les résultats obtenus en appliquant la procédure de tests multiples aux 17 gènes d'intérêt, en utilisant le test χ^2 si ses conditions de validité sont vérifiées ou le test exact de Fisher sinon. Sur l'ensemble des 23463 positions des 17 gènes étudiés, 3112 tests sont déclarés positifs, pour un nombre moyen de faux positifs estimé à 1221. Autrement dit, environ 2000 positions de la séquence d'ARNm des gènes étudiés seraient sujettes à des probabilités d'occurence de substitutions qui sont différentes dans les tissus cancéreux et dans les tissus sains. En considérant les équivalents unilatéraux du test (2), 2725 (resp. 1285) positions connaissant une plus grande (resp. une plus faible) probabilité d'être sujette à une infidélité de transcription sont mises en évidence.

Gènes	# de tests réalisés	# de tests positifs	% de tests positifs	LBE
ALB	2182	230	10.54	126
ALDOA	2272	131	5.77	164
ATP5A1	1931	175	9.06	114
CALM2	1096	99	9.03	60
ENO1	1776	213	11.99	87
FTH1	1195	271	22.68	58
FTL	811	152	18.74	32
GAPDH	1267	365	28.81	45
HSPA8	2239	200	8.93	121
LDHA	1649	120	7.28	94
RPL7A	882	146	16.55	40
RPS4X	928	143	15.41	45
RPS6	810	115	14.20	38
TMSB4X	606	126	20.79	25
TPI1	1206	151	12.52	58
TPT1	794	133	16.75	33
VIM	1819	342	18.80	81
Ens des gènes	23463	3112	13.26	1221

TAB. 1 – Résultats du test (2) pour les 17 gènes

4 Conclusion

Des positions d'ARNm plus fréquemment sujettes à des substitutions dans les tissus cancéreux que dans les tissus sains sont mises en évidence par la procédure de tests multiples. D'après notre modèle, les erreurs de séquençage ne peuvent contribuer aux différences de probabilités observées. Ces différences sont donc imputées à des erreurs de transcription. Ces changements de nucléotides conduiraient ainsi à la production de protéines dites aberrantes, dont la séquence d'acides aminés ne correspond pas à celle définie par l'ADN et dont la mesure permettrait par la suite de détecter les patients atteints de certains cancers. De plus, l'estimateur LBE a été étendu au cas de statistiques discrètes. Enfin, prendre en compte la dépendance pouvant exister entre les tests semble être une perspective intéressante [4].

Bibliographie

[1] Zhang, Z. and Schwartz, S. and Wagner, L. and Miller, W. (2000), A greedy algorithm for aligning DNA sequences, Journal of Computational Biology, 7, 1-2, 203–214.

[2] Conover, W.J. (1980) Practical Nonparametric Statistics, Wiley & Sons.

[3] Dalmasso, C. and Broët, P. (2005) Journal de la Société française de statistique, 146, 1-2, 63–75.

[4] Friguet, C. and Kloareg, M. and Causeur, D. (2009) A factor model approach to multiple testing under dependence, Journal of the American Statistical Association, 104, 488, 1406–1415.

[5] Brulliard, M. and Lorphelin, D. and Collignon, O. and others (2007) Nonrandom variations in human cancer ESTs indicate that mRNA heterogeneity increases during carcinogenesis, Proceedings of the National Academy of Sciences, 104, 18, 7522.

[6] Collignon, O. (2009), Recherche statistique de biomarqueurs du cancer et de l'allergie à l'arachide, Thèse de Doctorat.