Importance de la méthode d'analyse statistique des données du transcriptome

Importance of statistical analysis method in transcriptomics

J.F HOCQUETTE, I. BARNOLA, C. BERNARD, B. MEUNIER, K. SUDRE, A. LISTRAT, I. CASSAR-MALEK
INRA, Unité de Recherches sur les Herbivores, Equipe Croissance et Métabolisme du Muscle, Theix, 63122 Saint-Genès Champanelle, France

INTRODUCTION

L'analyse du transcriptome est basée sur l'hybridation moléculaire des ARN d'un tissu cible (après rétrotranscription en ADNc et marquage) avec des ADNc déposés sur un support (membranes de nylon, lames de verre). De plus, le même échantillon biologique peut être analysé avec plusieurs membranes ou lames différentes pour estimer la variabilité technique de la mesure.
Ce travail, réalisé dans le cadre du programme AGENAE (Analyse du GENome des Animaux d'Elevage), a pour objectif de comparer différentes méthodes de dépouillement ou d'analyse statistique des données du transcriptome.

1. MATERIEL ET METHODES

Le muscle *rectus abdominis* (RA) a été prélevé sur deux groupes de 3 taurillons à potentiel de croissance musculaire faible ou élevé au moment de l'abattage à l'âge de 15 mois. Les ARN extraits du RA ont été mélangés par groupe pour obtenir 2 échantillons représentatifs. Les ARNm des 2 échantillons ont été purifiés. Les ADNc issus de la rétrotranscription ont été marqués par incorporation de [α^{33}P]dCTP. Chaque échantillon d'ADNc a été hybridé sur 4 membranes identiques (Sudre *et al.*, 2003) sur lesquelles avaient été déposés en double sous forme de spots des fragments d'ADNc de muscles de bovins ainsi que des contrôles négatifs (Listrat *et al.*, 2003). Les intensités d'hybridation ont été calculées et normalisées par rapport à la moyenne des intensités de tous les spots pour que les résultats soient comparables d'une membrane à l'autre (Sudre *et al.*, 2003). Nous avons ainsi retenu 318 spots fiables en moyenne dans les deux échantillons sur 1060 spots avec de l'ADNc. Les différentes intensités d'hybridation sur 3 ou 4 membranes pour un même spot ont été considérées comme fiables à condition que les valeurs ne s'éloignent pas plus de 25% de leur moyenne. Les autres données (dont la variance était supérieure à 25% avec 3 résultats sur 4) ont été éliminées considérant que la mesure n'était pas fiable.
Les comparaisons entre échantillons ont été effectuées par la méthode des rapports, par le test *t* de Student, par la combinaison de ces 2 approches et par la méthode SAM ("Statistical Analysis of Microarrays") de façon à comparer les résultats issus de ces différentes méthodes couramment utilisées dans la littérature (Cui et Churchill, 2003).

2. RESULTATS

2.1. METHODE DES RAPPORTS

Nous avons calculé pour chaque spot le rapport des intensités d'hybridation des 2 échantillons analysés. Nous avons considéré comme différentiellement exprimés les spots dont les rapports sont supérieurs à 1,5, ce qui correspond à une différence de 0,5, soit deux écart-types de la variabilité de la mesure (0.25). De plus, il s'agit d'un seuil minimum pour avoir une probable signification biologique. Neuf spots ont ainsi été identifiés dont un contrôle négatif (un puits ne contenant que de l'eau), ce qui pose problème.

2.2. TEST T DE STUDENT AVEC L'ECART-TYPE DE CHAQUE SPOT

Nous avons calculé la valeur de *t* spot par spot, ce qui revient à considérer l'écart-type réel associé à chaque groupe de 3 ou 4 répétitions par spot pour un échantillon d'ARN. Nous avons ainsi pu calculer pour chaque ADNc la plus petite différence significative entre les intensités d'hybridation des 2 échantillons, ce qui nous a permis d'identifier 30 ADNc différentiellement exprimés entre les 2 groupes d'animaux parmi lesquels 2 contrôles négatifs (ADNc végétal), ce qui pose à nouveau un problème.

2.3. TEST *T* DE STUDENT AVEC UN ECART-TYPE MOYEN

Nous avons tout d'abord calculé l'écart-type moyen de la variabilité des intensités d'hybridation sur les 3 ou 4 membranes retenues pour tous les spots. Puis, nous avons utilisé cet écart-type pour calculer la plus petite différence significative entre les intensités d'hybridation obtenues avec les 2 échantillons étudiés ; 28 signaux ont ainsi été identifiés comme significativement différents à 5% dont 8 avec une plus forte probabilité (1%). Parmi ces derniers, on retrouve un contrôle négatif (1 puits avec de l'eau).

2.4. METHODE DES RAPPORTS ASSOCIEE AU TEST *T* DE STUDENT

L'association de la méthode des rapports (> 1,5) et du test *t* utilisant l'écart-type moyen ($P < 0,05$) permet de tenir compte de la variabilité de la mesure tout en ne retenant que les spots dont le différentiel d'expression est suffisant pour avoir une probable réalité biologique. Sept spots ont ainsi été retenus y compris le contrôle négatif cité ci-dessus.

2.5. METHODE *STATISTICAL ANALYSIS OF MICROARRAYS* (SAM)

SAM est basée sur le rapport des intensités et le test t en intégrant la variabilité propre à chaque spot et un terme constant estimant la variabilité moyenne de l'ensemble des spots. Une permutation aléatoire des données permet par ailleurs d'éliminer des faux positifs. Par cette approche, 8 spots ont été identifiés significativement différents entre les 2 échantillons dont aucun contrôle négatif. Cinq sur ces 8 spots sont identifiés par toutes les méthodes précédentes en particulier par la méthode des rapports couplée au test *t*.

CONCLUSION

Les probabilités de détecter des faux positifs (déclarer un gène différentiellement exprimé alors qu'il ne l'est pas) ou des faux négatifs (ne pas être capable de détecter un gène différentiellement exprimé) sont très différentes entre les méthodes statistiques utilisées. Dans le cadre de notre procédure de dépouillement (qui consiste à éliminer les mesures trop variables), la méthode combinée des rapports et du test *t* avec un écart-type moyen donne des résultats très comparables à la méthode SAM. Les autres méthodes utilisées ne sont pas adaptées à nos expérimentations.

Sudre, K., Leroux, C., Piétu, G., Cassar-Malek, I., Petit, E., Listrat, A., Auffray, C., Picard, B., Martin, P., Hocquette, J.F. 2003. J. Biochem., 133, 745 - 756
Cui X., Churchill G.A. 2003. Genome Biology, 4, 210.
Listrat, A., Sudre K., Cassar-Malek, I., Ueda Y., Barnola, I., Rolland, G., Leroux, C., Gentès, G., Renand, G., Martin, P., Hocquette, J.F. 2003. Renc. Rech. Ruminants, 10.