Caractérisation fonctionnelle de clusters de gènes : application à l'étude d'expression différentielle dans le muscle chez le porc

Olivier Dameron
olivier.dameron@univ-rennes1.fr

1 Contexte

La viande de porc, que ce soit sous forme de viande fraîche ou de produits de charcuterie, est la première viande consommée en France. Elle représente 40% des protéines animales consommées. Cependant, la qualité de la viande est souvent décrite comme hétérogène. La qualité des viandes est un caractère complexe encore mal maîtrisé, sous la dépendance de multiples facteurs environnementaux (techniques d'élevage, alimentation, conditions de transport et d'abattage) et génétiques (race, sexe, génotype aux gènes majeurs Halothane et RN). De nombreuses études ont mis en évidence la complexité des interactions existants entre les facteurs environnementaux et génétiques, et régulant la qualité des viandes [1, 2].

Les outils de génomique fonctionnelle offrent la possibilité d'observer les influences de ces différents facteurs au niveau moléculaire et ainsi de permettre une meilleure compréhension des mécanismes moléculaires et génétiques gouvernant les aspects de la qualité. Actuellement, il n'existe pas de critère unique, fiable permettant d'estimer la qualité de la viande in vivo. L'identification de prédicteurs ou biomarqueurs de la qualité de la viande de porc fraîche à l'abattage est donc un point important pour la filière afin d'optimiser l'utilisation de la carcasse. Cette identification passe par une approche globale et par l'intégration des données issues de la post génomique pour aboutir à la détection de gènes responsables de variation phénotypique chez les animaux d'élevage.

La viande provenant de la maturation du muscle, les caractéristiques (composition et structure) et le métabolisme énergétique musculaires influencent fortement les propriétés physicochimiques de la viande [3] et donc ses qualités technologiques ou organoleptiques. Une étude comparative du profil d'expression de 2 muscles « modèles » pour la filière produit transformé : le jambon (Semimembranosus, SM) et la filière viande fraîche : la longe (Longissimus dorsi, LD) a été réalisé dans le cadre du projet GENMASQ (*GENomics enabled Marker Assisted Selection and Certification of Quality in Pork products*) visant à identifier des marqueurs biologiques de variations de la qualité des viandes. Les données transcriptomiques ont été obtenue par hybridation d'une puce "muscle porçin" cDNA 15K. Les données expérimentales ont porté sur 90 animaux. Pour chaque animal, on dispose d'un prélèvement de longe (*Longissimus dorsi, LD*) et d'un prélèvement de jambon (*Semimembranosus, SM*).

L'étude des données transcriptomique a permi de mettre en évidence un

nombre très important de gènes différentiellement exprimés, et ainsi d'identifier des clusters de gènes surexprimés dans la longe (LD) ou dans le jambon (SM).

La base *Gene Ontology Annotation* (GOA) associe à chaque gène une liste de termes de *Gene Ontology* (GO) afin de décrire les processus biologiques dans lesquels il est impliqué, les composants cellulaires où il agit et ses fonctions moléculaires. *Gene Ontology* indique les relations sémantiques entre ces termes (par exemple « Cholesterol biosynthetic process » est une sous-classe (indirecte) de « Lipid biosynthetic process » [4].) L'utilisation de GO pour interpréter les données de GOA permet ainsi de déduire qu'un gène annoté par « Cholesterol biosynthetic process » doit également être considéré comme étant annoté par « Lipid biosynthetic process » et par tous ses ancêtres. Ceci est nécessaire afin de comparer correctement les annotations des gènes en retrouvant leurs plus petits ancêtres communs [5]. Il faut également tenir compte des evidence codes, qui indiquent la façon dont chaque annotation a été associée à un gène [6]

L'approche statistique consiste principalement à déterminer pour chaque cluster les annotations qui sont significativement plus présentes que si le cluster avait été constitué de gènes sélectionnés aléatoirement[7]. De nombreux outils existent pour mener à bien cette tâche. Néanmoins dans notre cas, le nombre de gènes différentiellement exprimés est tel que les résultats de cette approche sont trop généraux, et ne sont donc pas suffisamment informatifs pour être pertinents.

2 Objectif

L'objectif du stage est d'affiner l'étude comparative par une approche exploitant les connaissances de GO.

Afin d'exploiter au mieux les données disponibles, l'enjeux est de combiner les informations sur le niveau d'expression des gènes avec les descriptions fonctionnelles de ces gènes selon GO. Nous souhaitons notamment explorer les techniques utilisant GO [8, 9, 10, 11, 12] ou se basant sur des calculs de similarité sémantique [13, 14, 15] afin d'identifier les fonctions qui sont semblables et qui reviennent souvent dans un cluster de gènes.

Mots-clefs : données d'expression de gènes, Gene Ontology, similarité sémantique

3 Déroulement du stage

Nous proposons d'adopter la démarche suivante :

1. étude bibliographique des techniques le caractérisation fonctionnelles de clusters de gènes et notamment des solutions tenant compte des ontologies et des mesures de similarité sémantique ;
2. pour l'ensemble des gènes différentiellement exprimés :
   - identifier les principales fonctions ;
   - regrouper ces fonctions selon leur similarité, les organiser en dendrogramme et déterminer le seuil optimal de regroupement ;
   - réaliser une analyse différentielle avec les autres clusters en tenant compte des données d'expression.

3. inversement, réaliser la clusterisation selon le niveau d'expression, puis procéder à une caractérisation fonctionnelle afin de comparer les deux approches.

Le stage aura lieu au sein de l'U936 INSERM, à la faculté de médecine de Rennes. Il sera encadré par Olivier Dameron (maître de conférences U936) et Frédéric Hérault (INRA U598)

La gratification mensuelle de stage est de 425€ nets. FIXME METTRE A JOUR

4 Candidature

Le profil de candidat recherché correspond à un stage de master 2 en bioinformatique. Le candidat devra être capable d'utiliser les outils statistiques et de mener à bien la mise en œuvre de l'approche ontologique.

Une connaissance des ontologies sera appréciée mais ne constitue pas un pré-requis.

Pour candidater, merci de contacter Olivier Dameron [1] et Frédéric Hérault [2] en joignant un CV et une lettre de motivation.

Références

[1] V. Olsson and J. Pickova. The influence of production systems on meat quality, with emphasis on pork. *Ambio*, 34 :338–343, 2005.

[2] K. Rosenvold and H.J. Andersen. Factors of significance for pork quality–a review. *Meat science*, 64 :219–237, 2003.

[3] J. F. Hocquette, I. Ortigues-Marty, M. Damon, P. Herpin, and Y. Geay. Métabolisme énergétique des muscles squelettiques chez les animaux producteurs de viande. *Productions Animales*, 13 :185–200, 2000.

[4] Louis du Plessis, Nives Skunca, and Christophe Dessimoz. The what, where, how and why of gene ontology–a primer for bioinformaticians. *Briefings in bioinformatics*, 2011. In press.

[5] Seung Yon Rhee, Valerie Wood, Kara Dolinski, and Sorin Draghici. Use and misuse of the gene ontology annotations. *Nature Reviews Genetics*, 9(7) :509–515, 2008.

[6] Mark F Rogers and Asa Ben-Hur. The use of gene ontology evidence codes in preventing classifier assessment bias. *Bioinformatics (Oxford, England)*, 25(9) :1173–1177, 2009.

[7] Joseph S Verducci, Vincent F Melfi, Shili Lin, Zailong Wang, Sashwati Roy, and Chandan K Sen. Microarray analysis of gene expression : considerations in data mining and statistical treatment. *Physiological genomics*, 25(3) :355–363, 2006.

[8] K. Wolstencroft, P. Lord, L. Tabernero, A. Brass, and R. Stevens. Protein classification using ontology classification. *Bioinformatics*, 22(14) :e530–e538, 2006.

1. olivier.dameron@univ-rennes1.fr
2. frederic.herault@rennes.inra.fr

[9] Brendan Sheehan, Aaron Quigley, Benoit Gaudin, and Simon Dobson. A relation based measure of semantic similarity for gene ontology annotations. *BMC Bioinformatics*, 9(1) :468, 2008.

[10] Andreas Schlicker, Francisco S Domingues, Jörg Rahnenführer, and Thomas Lengauer. A new measure for functional similarity of gene products based on gene ontology. *BMC bioinformatics*, 7 :302, 2006.

[11] Brenton Louie, Silas Bergen, Roger Higdon, and Eugene Kolker. Quantifying protein function specificity in the gene ontology. *Standards in genomic sciences*, 2(2) :238–244, 2010.

[12] Barry R Zeeberg, Hongfang Liu, Ari B Kahn, Martin Ehler, Vinodh N Rajapakse, Robert F Bonner, Jacob D Brown, Brian P Brooks, Vladimir L Larionov, William Reinhold, John N Weinstein, and Yves G Pommier. Redundancyminer : De-replication of redundant go categories in microarray and proteomics analysis. *BMC bioinformatics*, 12(1) :52, 2011.

[13] James Z Wang, Zhidian Du, Rapeeporn Payattakool, Philip S Yu, and Chin-Fu Chen. A new method to measure the semantic similarity of go terms. *Bioinformatics (Oxford, England)*, 23(10) :1274–1281, 2007.

[14] Jing Wang, Xianxiao Zhou, Jing Zhu, Chenggui Zhou, and Zheng Guo. Revealing and avoiding bias in semantic similarity scores for protein pairs. *BMC bioinformatics*, 11(1) :290, 2010.

[15] Guangchuang Yu, Fei Li, Yide Qin, Xiaochen Bo, Yibo Wu, and Shengqi Wang. GOSemSim : an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*, 26(7) :976–978, 2010.