Université de Montréal

Performances de la puce exon et son application dans l'analyse de l'épissage alternatif associé à la métastase du cancer de sein

Par

Amandine Bemmo

Faculté de médecine

Programme de bio-informatique

Mémoire présenté à la Faculté des études supérieures en vue de l'obtention du grade de maitrise es sciences (M. Sc.) en Bioinformatique

Septembre, 2009

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé :

Performances de la puce exon et son application dans l'analyse de l'épissage alternatif associé à la métastase du cancer de sein

Présenté par :

Amandine Bemmo

a été évalué par un jury composé des personnes suivantes :

Sébastien Lemieux
Président-rapporteur

Miklós Csürös
Directeur de recherche

Jacek Majewski
Codirecteur

Michael Hallett
Membre du jury

Résumé

Nous montrons l'utilisation de la puce exon d'Affymetrix pour l'analyse simultanée de l'expression des gènes et de la variation d'isoformes. Nous avons utilisé les échantillons d'ARN du cerveau et des tissus de référence qui ont été antérieurement utilisés dans l'étude du consortium *MicroArray Quality Control* (MAQC). Nous démontrons une forte concordance de la quantification de l'expression des gènes entre trois plateformes d'expression populaires à savoir la puce exon d'Affymetrix, la puce Illumina et la puce U133A d'Affymetrix. Plus intéressant nous montrons que la majorité des discordances entre les trois plateformes résulterait des positions différentes des sondes à travers les plateformes et que les variations d'isoforme exactes ne peuvent être identifiées que par la puce exon. Nous avons détecté avec succès, entre les tissus de référence et ceux du cerveau, une centaine de cas d'évènements d'épissage alternatif.

La puce exon est requise dans l'analyse de l'épissage alternatif associé aux pathologies telles que les cancers et les troubles neurologiques. Comme application de cette technologie, nous avons analysé les variations d'épissage dans la métastase du cancer de sein développé dans le model de la souris. Nous avons utilisé une gamme bien définie de trois lignées de tumeur mammaire ayant différents potentiels métastatiques. Par des analyses statistiques, nous avons répertorié 2623 transcripts présentant des variations d'expression et d'isoformes entre les types de tumeur. Une analyse du réseau de gènes montre qu'environ la moitié d'entre eux est impliquée dans plusieurs activités cellulaires, ainsi que dans nombreux cancers et désordres génétiques.

Mots clefs : Puce exon, Épissage alternatif, Cancer de sein, Réseau de gènes.

Abstract

We demonstrate how the Affymetrix Exon Array, can be used to simultaneously profile gene expression level, and detect variations at the isoform level. We use a well studied set of brain and reference RNA samples previously used by the MicroArray Quality Control (MAQC) consortium study. We demonstrate a high concordance of gene expression measurements among three popular expression platforms – Affymetrix Exon Array, Illumina, and Affymetrix 3' targeted array (U133A). More interestingly, we show that in many cases of discordant results, the effect can be explained by differential probe placements across platforms, and that the exact isoform change can only be captured by the Exon Array. Finally, we are able to detect hundreds of cases of splicing, transcript initiation, and termination differences between the brain and reference tissue samples.

We propose that the Exon Array is a highly effective tool for transcript isoform profiling, and that it should be used in a variety of systems where such changes are known to be associated with diseases, such as neurological disorders and cancer. As application, we used the Affymetrix Exon Array to identify metastatis-specific alternative splicing in mouse model of breast cancer at the whole genome level. We utilize a well characterized series of three mouse mammary tumor lines exhibiting varying levels of metastatic potential. We catalogued 2623 transcripts which exhibit splicing aberrations during the progression of cancer. A genetic pathway analysis shows the half of them implicated in several cell activities, cancers and genetic disorders.

Key words: Exon Array, Alternative splicing, Breast cancer, Gene pathway.

Table des matières

Résumé iv

Abstract v

Table des matières vi

Liste des tableaux ix

Liste des figures x

Sigles et abréviations xii

Remerciements xiii

Introduction 14

Épissage alternatif 14

Rôle de l'épissage alternatif 17

Épissage alternatif et la pathogenèse 18

Identification de l'épissage alternatif 19

Objectifs et méthodologie 23

Contribution 24

Chapitre I Performances de la puce exon d'Affymetrix dans l'analyse de l'expression des gènes et la variation d'isoformes 25

Abstract 29

Background 29

Results 29

Conclusions 30

Background 31

Results 32

Variability across labs. 32

Variability across summarization methods. 35

Variability across platforms. 35

Alternative Isoform Detection. 36

Differences in Amplification and Labelling Protocols. 39

Using the Exon Array to Profile Alternative Isoforms. 41

Edge Bias Effect. 46

Discussion 51

Conclusions 55

Methods 56

Exon Array Hybridization 56

Data Pre-processing and Analysis 57

Probeset and Gene Mapping 57

Authors' contributions 57

Acknowledgements 58

Additional files 59

Chapitre II Application de la puce exon d'Affymetrix à l'épissage alternatif dans la métastase du cancer de sein 61

Synopsis 62

Abstract 64

Introduction 64

Results 67

Gene expression patterns in tumors 67

Tumor-specific gene expression and isoform variations 69

Pathway analysis of splicing events 80

Discussion 92

Conclusion and perspective 101

Materials and methods 102

Cell Culture and Transfections 102

Mammary gland injection and spontaneous Metastasis Assay 102

RNA extraction and microarray hybridization 103

Data pre-processing and analysis 104

Authors' contributions 108

Acknowledgments 108

Appendix– outlier identification and verification of metastatic potential 109

Supplementary Data 114

Discussion et conclusion générales 115

Bibliographie 123

Liste des tableaux

Table 2.1: List of some alternatively expressed probesets 80

Table 2.2: Top biological functions and diseases retrieved by the gene pathway analysis ... 81

Table 2.3: List of some significant genes that have important implications in normal biological processes and diseases 83

Table 2.4: For the three studies compared (our study against two prior studies), number of common genes differently expressed between the least metastatic sample and the most metastatic sample 94

Table 2.5: Common consistent differently expressed genes between the least metastatic and the most metastatic samples of each study 96

Table 2.6: Common inconsistent differently expressed genes between the least metastatic and the most metastatic samples of each study 98

Table 2.7: Correlation of the gene expression fold-changes of our study with two prior studies 109

Liste des figures

Figure 0.1: Différents types d'épissage alternatif.. 16

Figure 1.1: PCA plots at the probe set level show two main sources of variation among the 20 samples.. 33

Figure 1.2: Comparison of log2(FC) detected between the biological samples for the two labs 34

Figure 1.3: Exon array analysis of the ELAVL1 gene expression differences between brain and reference tissues... 37

Figure 1.4: Visualization of expression patterns of ELAVL1 gene... 38

Figure 1.5: Correlation of fold changes between Affymetrix U133, Illumina, and the Affymetrix Exon Array 40

Figure 1.6 : Examples of Candidates from Splicing Index Analysis 45

Figure 1.7: Edge bias... 48

Figure 1.8: Exon Array average gene expression index as a function of transcript (mRNA) length 50

Figure 2.1: PCA plot at the metaprobeset level shows the clustering of samples.................. 68

Figure 2.2: Visualization of the expression pattern of MED24 gene showing an alternative initiation or termination event 71

Figure 2.3: Visualisation of the expression pattern of CD44 gene showing several internal cassette exons and intron inclusions 72

Figure 2.4: Visualisation of the expression pattern of CDH1 gene showing a whole gene expression change... 73

Figure 2.5: Visualisation of the expression pattern of SRRT gene showing an intron inclusion...74

Figure 2.6: Visualisation of the expression pattern of SLC25A29 gene showing a differential 3' UTR site 75

Figure 2.7: Visualisation of the expression pattern of SLC39A14 gene showing a cassette exon 76

Figure 2.8 : Proportion of gene expressions variation 78

Figure 2.9: Network of molecular interactions containing products of statistically significant genes in breast cancer........ 88

Figure 2.10: Network of molecular interactions containing products of statistically significant genes in breast cancer........ 90

Figure 2.11: Network of molecular interactions containing products of statistically significant genes in breast cancer........ 91

Figure 2.12: workflow of probeset and metaprobeset filtering steps 106

Figure 2.13: protocol flow of gene viualisation and manual curation........ 107

Figure 2.14: Tumor outgrowth of Murine Carcinoma cell lines 111

Sigles et abréviations

ADNc : acide désoxyribonucléique complémentaire

ARN: acide ribonucléique

ARNm: acide ribonucléique messager

AS : alternative splicing

ASE : alternative splicing event

ESE: exonic splicing enhancer

ESS: exonic splicing silencer

EST: expressed sequence tag

FDR: false discovery rate

ISE: intronic splicing enhancer

ISS: intronic splicing silencer

mRNA: messenger ribonucleic acid

PCA: principal component analysis

PCR: polymerase chain reaction

pré-ARNm : acide ribonucléique prémessager

qRT-PCR: quantitative reverse transcription polymerase chain reaction

RNA: ribonucleic acid

RT-PCR: reverse transcriptase polymerase chain reaction

SI : splicing index

Remerciements

L’accomplissement de ce travail de maitrise est le fruit de nombreux échanges et collaborations.

Je tiens tout d’abord à remercier mon codirecteur Dr Jacek Majweski qui m’a fait confiance en m’accueillant dans son laboratoire et en m’offrant l’opportunité de réaliser ce projet. Je lui exprime également ma gratitude pour sa contribution à la mise en forme de mon anglais boiteux.

Je remercie mon directeur Dr Miklos Csuros pour son encadrement et son soutien.

Je remercie tous les membres de mon laboratoire, en particulier David Benovoy, pour leur aide et leur esprit d’équipe.

Je remercie tout les membres du jury de s’être mis au service de l’évaluation de mon mémoire : les professeurs Sébastien Lemieux et Michael Hallett.

Enfin je remercie toute ma famille en particulier mon fiancé Clovis Simo et ma mère Odette Djuidje pour leur soutien moral, leurs encouragements et tous leurs sacrifices pour que je poursuive mes études dans les meilleurs conditions.

Introduction

Épissage alternatif

Chez les eucaryotes, la plupart des gènes sont constitués de parties transcrites (exons) interrompues par des parties non-codantes (introns). L'épissage est un processus qui consiste à l'excision des introns et la ligature des exons. Chez les eucaryotes supérieurs ce processus est effectué lors de la maturation du pré-ARNm à l'intérieur du noyau cellulaire. Le spliceosome qui catalyse cette réaction est constitué d'un ensemble de cinq ribonucléoprotéiques et de plus d'une centaine de protéines [1]. A travers plusieurs interactions protéine-protéine, ARN-ARN ou protéine-ARN, le spliceosome reconnaît la jonction exon-intron et déclenche deux réactions de trans-estérification qui retirent les introns du pré-ARNm et lient les exons. L'ARNm mature sera transporté dans le cytoplasme et traduit en protéine.

Dans la cellule, il existe une variante de l'épissage qui produit différents ARNm d'un même locus génomique. Ce processus, appelé épissage alternatif, est un mécanisme régulateur permettant à un pré-ARNm d'être épissé en plusieurs ARNm matures pouvant coder pour des protéines différentes. Les facteurs cis-régulateurs (ESE, ISE, ESS et ISS) qui dans la plupart des cas sont des sites de liaison des facteurs d'épissage, inhibent (ISS, ESS) ou induisent (ESE, ISE) l'usage d'un site d'épissage (exon ou intron) ou créent une structure secondaire de l'ARN qui affecte la reconnaissance du site d'épissage [2, 3]. La génération de plusieurs isoformes d'un même ARNm précurseur augmente la diversité protéique. Chez l'humain, l'épissage alternatif affecte environ 95% des gènes [4, 5].

Il existe plusieurs types d'épissages alternatifs (Figure 0.1 [6]) : le saut d'exon, l'exclusion mutuelle d'exons, la rétention d'intron, le site d'épissage alternatif en 5' et le site d'épissage alternatif en 3'. Le saut d'exon c'est lorsque pour deux isoformes d'un même gène, un exon est présent dans l'un et absent dans l'autre. On parle d'exclusion mutuelle d'exons lorsque pour deux exons et deux isoformes donnés, dans le premier isoforme un exon est inclus et l'autre absent tandis que celui absent dans le premier isoforme est inclus dans le second isoforme et celui présent dans le premier est absent dans le second isoforme. La rétention d'intron c'est lorsqu'une partie ou la totalité d'une séquence intronique est incluse dans l'ARNm mature. La sélection alternative du site 5' d'épissage de l'exon est lorsque la partialité du côté 5' de l'exon est incluse dans le transcrit; lorsque le même phénomène se produit à l'extrémité 3' on parle de sélection alternative du site 3' d'épissage.

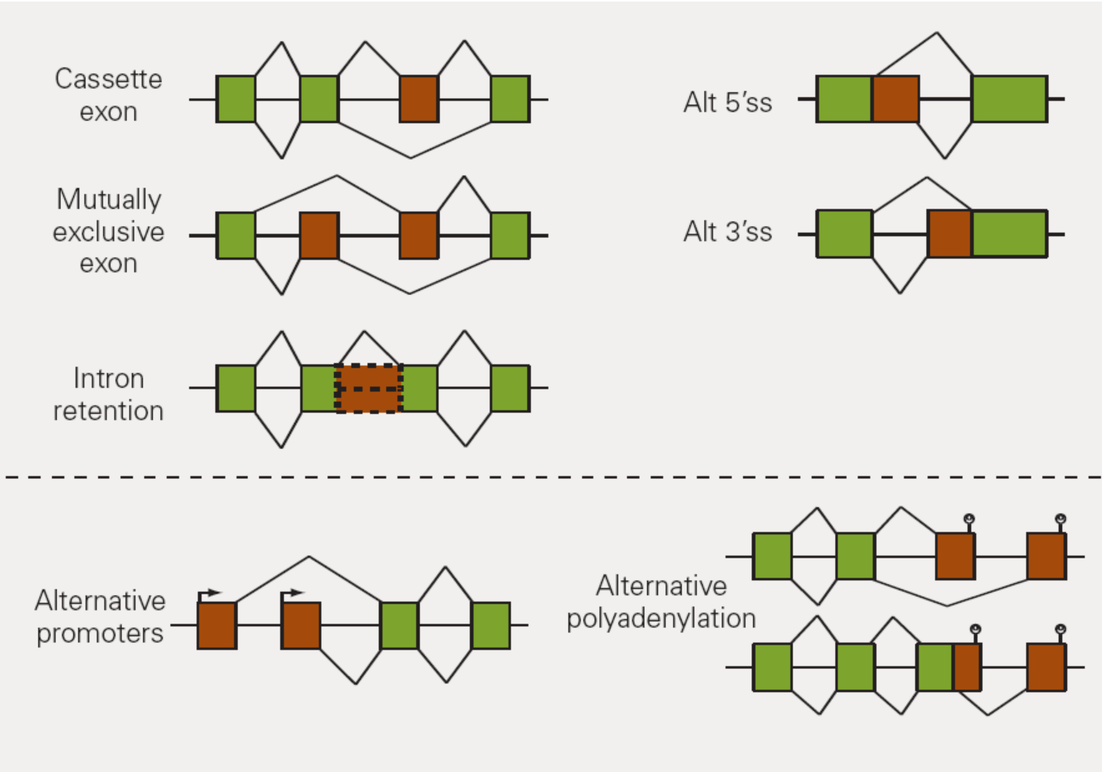

Figure 0.1: Différents types d'épissage alternatif

Les exons sont représentés par des boxes verts (exons constitutifs) et marrons (exons alternatifs) et les introns par des traits noirs. Les différents événements possibles sont le saut d'exon (Cassette exon), l'exclusion mutuelle d'exons (Mutually exclusive exon), la rétention d'intron (Intron rétention), le site d'épissage alternatif en 5' (Alt 5'ss) et le site d'épissage alternatif en 3' (Alt 3'ss). Cependant le choix multiple du promoteur d'initiation à la transcription (Alternative promoters) peut imposer le type d'épissage. L'épissage alternatif peut conduire à l'utilisation de sites de polyadenylation différents (Alternative polyadenylation).

Rôle de l’épissage alternatif

La regulation de l’épissage alternatif est soumise à plusieurs facteurs telles que la spécificité des tissus, le stade de développement, les activités physiologiques, la détermination du sexe et la réponse aux facteurs de stress. L’activité physiologique peut varier d’un isoforme à l’autre; les changements de la séquence de l’ARNm par l’épissage alternatif peuvent se répercuter sur la protéine résultante et modifier ainsi ses propriétés et sa fonction. C’est le cas des gènes Bcl-x, Caspase-9, Ced-4, Caspase-2/Ich-1 et hTid-1 qui encodent à la fois la variante anti-apoptose et la variante pro-apoptose [7, 8].

Près de 9.5% des mutations cataloguées par le *Human Gene Mutation Database* (HGMG) affecteraient l’épissage [9]. Les variations dans les éléments cis-régulateurs pourraient être responsables d’un nombre substantiel d’épissages alternatifs. Une transition de G à A dans l’exon 18 du gène de susceptibilité au cancer de sein BRCA1 altère le ESE et cause ainsi le saut de l’exon 18 [10]. Une insertion d’un T devant le di-nucléotide très conservé GT dans le site donneur de l’intron 4 du gène FAP, qui est un un suppresseur de tumeur, mène au saut de l’exon 4 et conduit à une forme atténuée de la Polypose recto-colique familiale [11]. Plusieurs études ont montré le lien entre le changement du niveau d’expression des facteurs trans-acting et l’épissage alternatif associé aux formations cancéreuses. En utilisant le modèle de la souris du cancer de sein, Stickeler et al [12] ont observé lors du développement du cancer, l’augmentation progressive du niveau d’expression des protéines SR en parallèle avec l’épissage alternatif du gène CD44. Les protéines SR (riches en dipeptides sérine/arginine) forment une famille de facteurs d’épissage trans-acting qui jouent un rôle important dans la régulation de l’épissage alternatif de nombreux gènes incluant CD44.

L’ajout, le retrait ou la variation de la taille des exons ou encore la rétention d’intron peut causer un décalage du cadre de lecture de la traduction introduisant des codons stop prématurés ou absents, ainsi que des substitutions d’acides aminés. La traduction de tels transcrits donnerait des protéines troncaturées ou dysfonctionnelles. Ce type d’ARNm aberrant est en général reconnu et dégradé par les mécanismes de control de qualité de l’ARNm tels que *nonsense-mediated mRNA decay* [9] ou le *nonstop Mediated mRNA Decay* [13, 14]

Épissage alternatif et la pathogenèse

La variation de l’épissage alternatif entre les individus crée une diversité phénotypique et peut également causer des désordres génétiques. Plusieurs maladies, telles que la fibrose kystique et les cancers, ont un lien avec les mutations ou variations dans les facteurs cis-acting ou trans-acting qui conduisent aux transcrits aberrants et à la production de protéines défectueuses. La surexpression du facteur trans-acting HMGA1a (HydroxyMethylGlutaryl coenzyme A1a) qui lie l’exon 5 cause un isoforme aberrant du gène PS2 (presenilin-2) dans lequel l’exon 5 est absent; cet isoforme est une caractéristique de la maladie d’Alzheimer [15].

Nombreux pré-ARNm subissent des variations d’épissage dans plusieurs types de cancer durant les phases de développement, progression et/ou de métastase. Par exemple, le gène CD44 codant pour une protéine d’adhésion, de prolifération et de migration cellulaires montre une variante d’épissage inhabituelle dans le cancer de sein. Il inclut une mixture de dix exons variables; cette forme est associée à l’acquisition du potentiel métastatique [12]. Un autre cas est le saut de l'exon 18, dû à une mutation du gène BRCA1, qui altère l'ESE de cet exon. Ce saut entraîne la suppression de 26 acides aminés dans une région essentielle aux fonctions de réparation de l'ADN, de régulation de la transcription et de suppression de tumeur de BRCA1 [16]. Un grand nombre de transcrits aberrants du gène FHIT, un suppresseur de tumeur, ont été découverts dans plusieurs tumeurs humaines telles que les tumeurs gastriques, cervicales, thyroïdiennes et testiculaires [9]. Ces transcrits résultent des sauts d'exons, des sites d'épissage alternatif en 3' et 5' et des inclusions d'introns. L'impact de ces variantes est l'inactivation du suppresseur de tumeur en réduisant la concentration de l'ARNm de la forme fonctionnelle de FHIT [9]. Pour mieux élucider le mécanisme et la régulation de l'épissage alternatif, une étape cruciale serait l'identification des événements d'épissage alternatif.

Identification de l'épissage alternatif

Jusqu'à présent, la technique la plus fiable d'identification et de validation des événements d'épissage alternatif est le séquençage de l'ADNc. Diverses méthodes telles que la *RT-PCR* et le *Northem blot* sont également très utilisées. Toutefois, la puissance de ces méthodes demeure limitée pour l'analyse des formes d'épissage alternatif dans l'ensemble des tissus ou dans des conditions pathologiques telles que les stades de développement d'un cancer. Plusieurs approches d'identification de l'épissage alternatif dans l'ensemble du génome ont vu le jour notamment les approches bioinformatiques basées sur l'alignement des ARNm et des ESTs de la séquence génomique correspondante, les puces à ADN et les puces à oligonucleotides.

Plusieurs études sur l'épissage alternatif ont utilisé la méthode de l'alignement des ARNm et des ESTs ou l'alignement de l'ARNm, des ESTs et de la séquence génomique correspondante. Le cluster des ESTs correspondant à la séquence génomique est aligné avec l'ARNm, et parfois avec la séquence génomique également, en utilisant les outils d'alignement tels que BLAST [17]. Cependant, cette méthode présente plusieurs inconvénients : le biais de couverture proche de l'extrémité 3' du transcrit, les erreurs de séquençage des ESTs, la faible qualité des bases de données des ESTs due à la contamination par les séquences génomiques ou les séquences d'ARNm partiellement épissées et pouvant introduire de faux positifs d'évènements de rétention d'introns; on note également la faible couverture de plusieurs types de tissus et les différents types de protocoles entre les laboratoires.

Les puces ADN se sont révélées être des outils puissants pour analyser le niveau d'expression de plusieurs gènes en une seule expérimentation. Plusieurs approches utilisant la technologie des puces ADN telles que la puce à jonction d'exons et la puce à fibre optique ont été utilisées avec succès [18, 19] mais présentent toutefois des limites. Plusieurs études [18, 20] ont démontré l'utilisation de la puce à jonction d'exons pour l'analyse de l'épissage alternatif. Les sondes de cette dernière ciblent les jonctions exon-exon puisque les différents isoformes d'un gène ont des jonctions exon-exon différentes. Cette approche est très efficace pour l'analyse des variations d'épissage connues. Cependant elle est en général incapable d'identifier les nouveaux sites d'épissage notamment aux extremités 3' et 5' ainsi que les combinaisons de plusieurs types d'événements d'épissage alternatif dans le même ARNm. Yakley et al [19] ont développé une nouvelle approche combinant la puce à fibre optique et une technique appelée RASL (RNA-mediated annealing, selection, and ligation) pour l'analyse de l'épissage alternatif à grande échelle. Comme la puce à jonction d'exons, cette technique ne permet pas l'identification de nouveaux événements d'épissage alternatif car elle requiert une connaissance préalable des structures exon-intron.

Un autre exemple de plateforme d'analyse de l'épissage alternatif est la puce SpliceArray [21] de la firme Exonhit qui estime l'intensité d'expression des isoformes des transcrits. Cette dernière combine les sondes interrogeant les jonctions exon-exon et celles interrogeant les corps d'exons pour mesurer le niveau d'expression des isoformes. Mais comme la puce à jonctions exon-exon, elle requiert une connaissance préalable de la structure de l'isoforme et ne permet donc pas l'identification de nouveaux isoformes. On note également la puce SpliceExpress [22] de la firme Jivan Biologics, ayant un principe similaire à la puce spliceArray, qui interroge les jonctions exon-exon et les jonctions intron-exon pour tester l'occurrence d'un isoforme; cependant elle est également limitée pour la découverte de nouveaux événements d'épissage alternatif.

A fin d'augmenter la probabilité de découvrir les variations et structure d'épissages inconnues, un nouvel outil est né. Il s'agit de la puce exon, une innovation d'Affymetrix [23]; elle permet l'analyse de l'épissage alternatif à l'échelle des transcrits et également à l'échelle des exons. C'est un outil pouvant analyser individuellement les exons (ou les parties d'exons) comme des objets indépendants et permet ainsi d'observer les sauts ou inclusions d'exons dans les transcrits. Ceci n'est pas possible ou est non optimal avec les puces traditionnelles. Cet outil prend également en charge les événements d'épissage alternatif aux extremités 3' et 5' du gène. Les algorithmes de conception de sondes utilisent une vaste variété de collections d'exons et transcrits prédits ou identifiés telles que ceux des banques *Ensembl* et *GenScan*. La puce exon contient approximativement 5.4 millions de sondes constituant 1.4 millions d'ensemble de sondes (*probeset*) interrogeant individuellement plus d’un million d’exons dans les régions transcrites connues ou prédites. Les *probesets* d’un gène sont regroupés ensemble (en un *metaprobeset*) pour estimer l’intensité d’expression du gène. Les *probesets* sont conçus à partir des PSRs (*probe selection region*) qui sont des séquences génomiques cibles. Dans la majorité des cas, un PSR interroge un exon. La taille moyenne d’un PSR est de 123 paires de base avec une taille minimale de 25 paires de base. Près de 90% des PSRs sont représentés par 4 sondes. L’interrogation de chaque région génique cible par plusieurs sondes (les 4 sondes qui constituent chaque *probeset*) a pour avantage d’améliorer la confiance statistique, de réduire l’impact des sondes inconsistantes et d’améliorer le ratio signal-bruit comparativement aux puces (exemple de SpliceArray et SpliceExpress) qui utilisent une sonde par région génique cible. De plus, à l’echelle du transcrit, la puce exon fournit une estimation robuste du niveau d’expression des gènes car le nombre de sondes par séquence de référence du gène varie entre 30 et 40 et est reparti sur tout l’ensemble du transcrit comparativement aux puces à extremités 3’ ciblées dont les sondes sont situées uniquement vers l’extrémité 3’ du transcrit.

Il serait très couteux pour les puces à jonctions d’exons et/ou d’introns de tester l’expression de tous les isoformes possible d’un même transcrit. Les nouvelles structures de transcrits prédites par la puce exon peuvent être utilisées comme cibles par ces puces, ceci en profilant les isoformes résultants des nouveaux événements d’épissage alternatif identifiés par la puce exon.

Objectifs et méthodologie

La puce exon a été utilisée avec succès dans plusieurs étude [24-26]. Cependant ses performances sont peu connues en particulier comparativement aux puces classiques. Nous voulons démontrer que la puce exon est effectivement un outil puissant et flexible permettant l'analyse des variations d'épissage y compris l'initiation et la terminaison de la transcription. Pour ce faire, nous utilisons l'ARN du jeu de données de l'étude MAQC pour l'hybridation de la puce exon que nous comparons avec deux autres puces très communes à savoir Illumina et Affymetrix U133. Par la suite nous voulons appliquer cette nouvelle technologie à l'investigation des variations d'épissage dans le cancer de sein développé dans le modèle de la souris. Trois lignées cellulaires cancéreuses mammaires humaines (168FARN, 4T07 et 66C14) ayant des potentiels métastatiques différents ont été utilisées. Les lignées seront injectées aux souris et formeront des tumeurs après un certain temps de croissance. L'ARN sera extrait des cellules tumorales et l'expression des gènes sera quantifiée avec la puce exon d'Affymetrix. Par des méthodes statistiques, les gènes différemment exprimés ou présentant des variations d'isoformes dans les tumeurs seront répertoriés et par la suite utilisés dans une analyse de réseau de gènes pour identifier leur pertinence biologique et pathologique.

Contribution

Dans la première partie, j'ai fait les analyses statistiques et la production des figures pour :

- la variation et concordance entre les laboratoires
- la variation due au type de méthode de sommation utilisée pour déterminer le niveau d'expression des gènes
- la variation entre les différentes plateformes
- l'effet du biais du protocole d'amplification de la puce exon aux extrémités du transcrit.

Dans la seconde partie, j'ai effectué la rédaction de l'article et la totalité des analyses informatiques et statistiques à savoir :

- l'estimation du niveau d'expression des exons et des gènes
- les différents étapes de filtrage des données pour obtenir les gènes et exons les plus significatifs
- les analyses d'événements d'épissage alternatif entre les différents types de tumeurs
- l'analyse du réseau des gènes significatifs
- la préparation des figures.

Chapitre I Performances de la puce exon d’Affymetrix dans l’analyse de l’expression des gènes et la variation d’isoformes

Correction

After the publication of the paper presented in this chapter, we were alerted to an error in our manuscript. The x-axis labels for Figure 1.7 were inverted. They should read from left to right: “Distance from the 5’ end” and “Distance from the 3’ end”, respectively. This does not affect our original interpretation of the edge bias effect presented in our original publication in any way.

In this chapter we have taken into account the erratum by replacing the original Figure 7.1 by the figure of the erratum published (BMC Genomics. 2009 Mar 23;10).

BMC Genomics

Gene Expression and Isoform Variation Analysis using Affymetrix Exon Arrays

BMC Genomics 2008, **9**:529 doi:10.1186/1471-2164-9-529

Amandine Bemmo
David Benovoy
Tony Kwan
Daniel J Gaffney
Roderick V Jensen
Jacek Majewski

ISSN	1471-2164
Article type	Methodology article
Submission date	25 July 2008
Acceptance date	7 November 2008
Publication date	7 November 2008
Article URL	http://www.biomedcentral.com/1471-2164/9/529

Gene Expression and Isoform Variation Analysis using Affymetrix Exon Arrays

Amandine Bemmo[1,3*], David Benovoy[2,3*], Tony Kwan[2,3], Daniel Gaffney[2,3], Roderick V. Jensen,[4] Jacek Majewski[2,3§]

[1] Université de Montréal, Montreal, QC, Canada

[2] Department of Human Genetics, McGill University, Montreal, QC, Canada

[3] McGill University and Genome Quebec Innovation Center, Montreal, QC, Canada

[4] Department of Biological Sciences, Virginia Tech, Blacksburg, Virginia, USA

*These authors contributed equally to this work

§Corresponding author

Abstract

Background

Alternative splicing and isoform level expression profiling is an emerging field of interest within genomics. Splicing sensitive microarrays, with probes targeted to individual exons or exon-junctions, are becoming increasingly popular as a tool capable of both expression profiling and finer scale isoform detection. Despite their intuitive appeal, relatively little is known about the performance of such tools, particularly in comparison with more traditional 3’ targeted microarrays. Here, we use the well studied Microarray Quality Control (MAQC) dataset to benchmark the Affymetrix Exon Array, and compare it to two other popular platforms: Illumina, and Affymetrix U133.

Results

We show that at the gene expression level, the Exon Array performs comparably with the two 3’ targeted platforms. However, the interplatform correlation of the results is slightly lower than between the two 3’ arrays. We show that some of the discrepancies stem from the RNA amplification protocols, e.g. the Exon Array is able to detect expression of non-polyadenylated transcripts. More importantly, we show that many other differences result from the ability of the Exon Array to monitor more detailed isoform-level changes; several examples illustrate that changes detected by the 3’ platforms are actually isoform variations, and that the nature of these variations can be resolved using Exon Array data. Finally, we show how the Exon Array can be used to detect alternative isoform differences, such as alternative splicing, transcript termination, and alternative promoter usage. We discuss the possible pitfalls and false positives resulting from isoform-level analysis.

Conclusions

The Exon Array is a valuable tool that can be used to profile gene expression while providing important additional information regarding the types of gene isoforms that are expressed and variable. However, analysis of alternative splicing requires much more hands on effort and visualization of results in order to correctly interpret the data, and generally results in considerably higher false positive rates than expression analysis. One of the main sources of error in the MAQC dataset is variation in amplification efficiency across transcripts, most likely caused by joint effects of elevated GC content in the 5' ends of genes and reduced likelihood of random-primed first strand synthesis in the 3' ends of genes. These effects are currently not adequately corrected using existing statistical methods. We outline approaches to reduce such errors by filtering out potentially problematic data.

Background

Alternative pre-mRNA splicing is a process that allows for the production of numerous protein variants from a single genomic locus. As researchers are becoming aware of the importance of splicing and mRNA processing in generating transcriptomic diversity, isoform-sensitive microarrays are rapidly gaining popularity in gene expression analysis [27, 28]. In particular, Affymetrix Exon Arrays are becoming a standard for both general and isoform-level expression analysis [25, 29-36]. Briefly, the Exon Array platform relies on 25-mer oligonucleotide probes to target the individual exons of a gene. The expression level of each exon can be detected independently, and summarized to monitor transcript expression levels as well as changes of individual transcript isoforms. The more universal coverage of the "Whole-Transcript" (WT) arrays renders them an attractive alternative to the traditional 3' biased expression microarrays.

We have previously successfully used Exon Arrays to demonstrate variation in isoform level expression in human populations [37] and associate this variation with underlying genetic differences [33]. We showed that the Exon Array is indeed a powerful and flexible tool, allowing for the detection of changes in splicing, transcript initiation, and termination. However, analysis of exon-level data is considerably more complicated than traditional analysis of gene expression. The complexity of the analysis may prevent many researchers from using WT arrays and profiting from associated advances in gene expression profiling.

Here, we use the example of a well studied system to outline the analysis and present results of a typical Exon Array experiment. We use the brain and reference human mRNA samples previously studied by the MicroArray Quality Control (MAQC) consortium [38, 39]. These commercially available samples provide a high quality reference dataset for comparing microarray results across various platforms and laboratories. The human brain has very distinct gene expression signatures, and the comparison with the reference (combined) tissue pool results in detection of numerous genes with differential expression levels. The original MAQC study relied on these samples to demonstrate high concordance between various microarray platforms. Incidentally, the human brain is also rich in specific isoforms, and constitutes a highly suitable system for assessing the performance of the Exon Array as both an expression and isoform-sensitive platform.

Results

Variability across labs.

Five technical replicates of brain and reference were hybridized in two independent labs: McGill University (MU) and Virginia Tech (VT), for a total of 20 samples. Principal component analysis, which is a commonly used method to visualize sources of variability in the data, is shown in Figure 1.1. Our experience with Exon Arrays indicates that in general the ribosomal RNA reduction step is the most inconsistent part of the protocol and is likely to be a major contributor to the differences across labs.

Variability in hybridization intensities, background noise, and random errors across labs may contribute to differences in final conclusions resulting from microarray analyses. In the case of the MAQC data, the final goal was to quantify differences in gene expression levels between the human brain and reference tissues. A relevant metric of such expression difference is the fold change (FC), calculated as FC = Expression(Brain)/Expression(Reference). In Figure 1.2, we show a correlation plot comparing the calculated fold changes in genes expression between the two labs. Despite the inter-lab variability in expression levels shown in the PCA plots, the final results (fold changes) are highly consistent for the two labs, with a correlation coefficient of greater than 0.97.

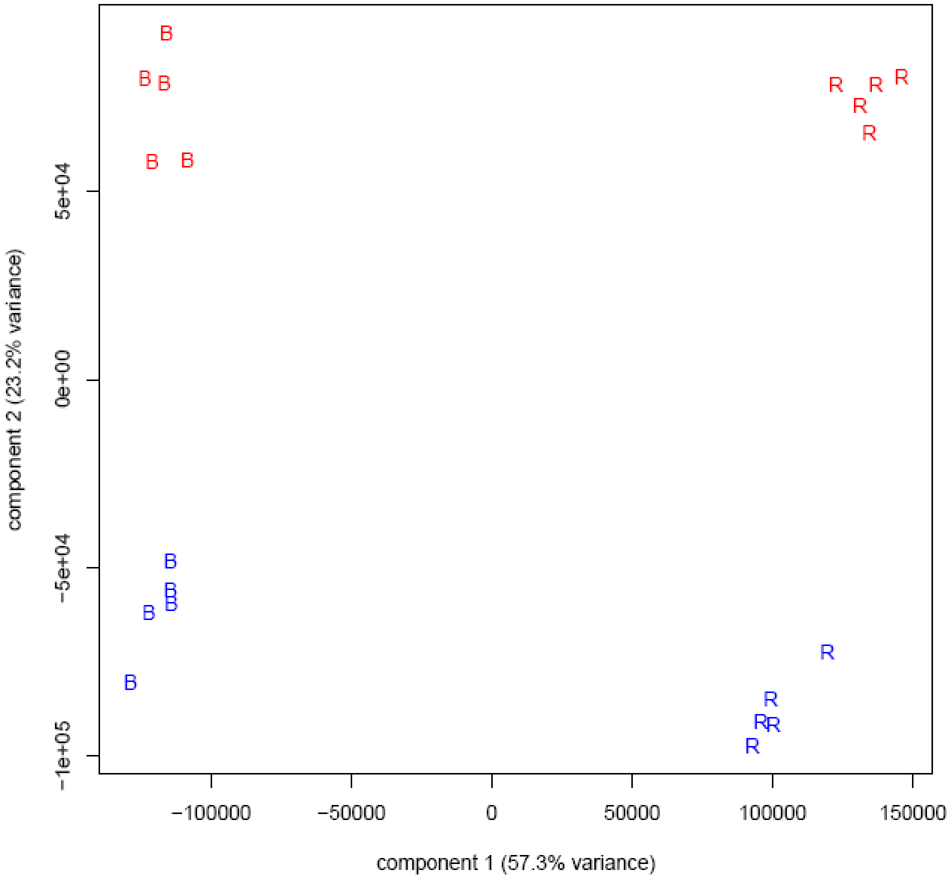

Figure 1.1: PCA plots at the probe set level show two main sources of variation among the 20 samples

PCA plots at the probe set level show two main sources of variation among the 20 samples. The first principal component explains 65% of the variance and corresponds, as expected, to the biological source of the sample: brain (B) vs. reference (R). The second principal component explains 20% of the variance and corresponds to the "lab effect" between VT (blue), and McGill (red) – that is, it illustrates the technical variability across labs.

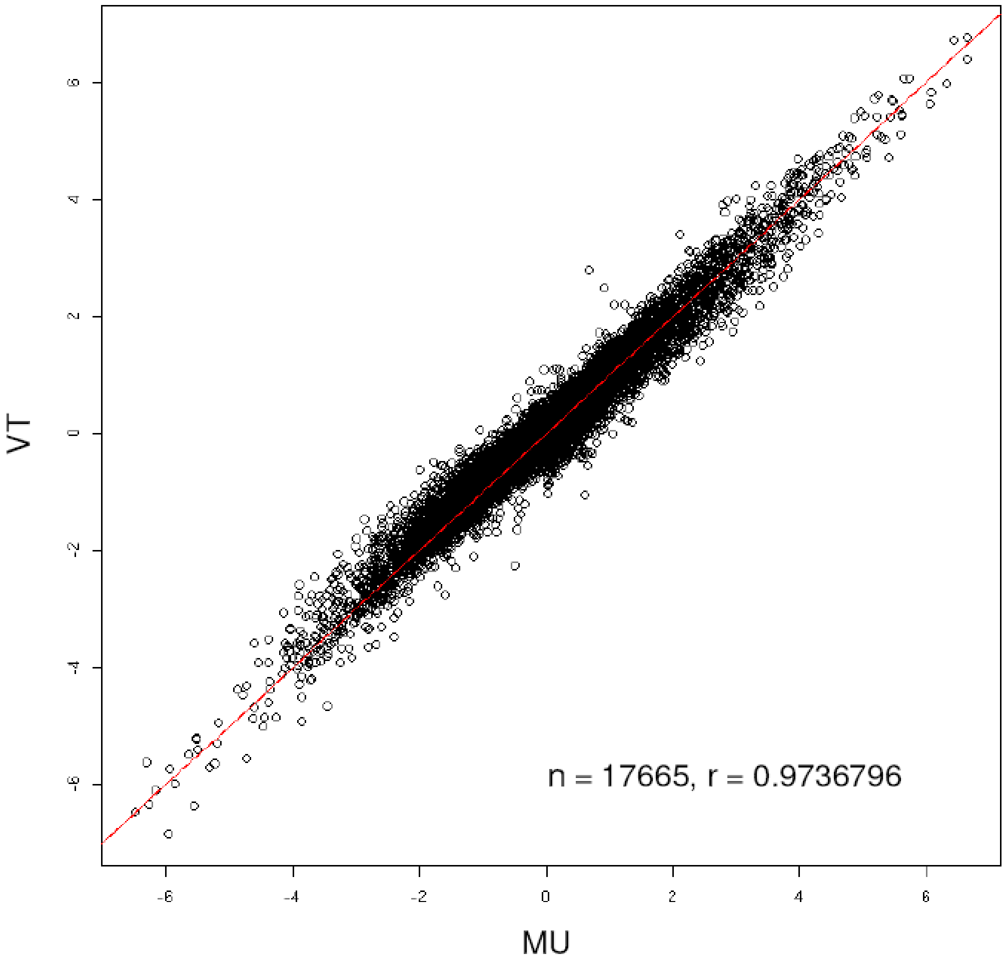

Figure 1.2: Comparison of log2(FC) detected between the biological samples for the two labs

Despite significant variation in expression measure across test sites, the fold change estimates are highly correlated.

Variability across summarization methods.

The aim of the summarization step in microarray analysis is generally to combine signals from multiple probes, which target the same expression unit, into a single expression index. Most of the popular methods strive for robustness against outlier probes (e.g. cross hybridizing, saturated, or non-responsive probes). We used our fold change results to compare two commonly used summarization methods: PLIER and RMA. We noted that RMA does result in a slight compression of fold changes, as has been observed in prior studies using other microarray platforms [38]. However, we find that the correlation of fold changes obtained from the two approaches is very high (r = 0.99).

Variability across platforms.

The original MAQC studies demonstrated that microarray results are highly consistent across different platforms [38]. Here, we compare the performance of the Exon Array in determining gene expression levels with two other popular platforms previously used by MAQC: Illumina Bead Array and Affymetrix U133 Gene Chip. In order to facilitate comparison across labs as well as platforms, we selected a number of genes which are reliably annotated and targeted by a common set of probesets (see Methods).

For the Exon Arrays, the fold changes were calculated by combining the results from the two labs (MU and VT). For the sake of consistency in the comparison, two test sites were chosen at random and combined for each platform within the MAQC dataset. We find that the 3' targeted platforms, Illumina Human-6 BeadChip and Affymetrix U133, produce the most consistent results (R = 0.92). This is not surprising, since the probe selection regions for the two platforms largely coincide, and the amplification protocols are poly-A primed and biased towards the 3’ ends of genes. The correlation with the Exon Array is slightly lower: R = 0.89 for U133 and 0.85 for Illumina. It has been previously shown [40-42], that the Exon Arrays are effective tools for gene expression profiling. Therefore, it is of interest, to examine the main sources of differences between the Exon Arrays and other platforms. Thus, in the analysis below we will concentrate on the genes whose predicted expression patterns are not consistent across platforms. In particular, the Exon Array is able to distinguish between specific isoforms of a given genomic locus, whereas the Illumina and Affymetrix U133 platforms generally target only a single isoform.

Alternative Isoform Detection.

It has previously been pointed out that some discordant results in the original MAQC [38] study were caused by differential isoform expression, and differences in probe placement across platforms. One particular discordant gene, *ELAVL1*, was suspected to express two alternative isoforms, differing in the 3’ UTR region. In Figure 1.3, we use the example of *ELAVL1* to illustrate the advantages of using the Exon Array for profiling individual isoforms.

It is clear that although the Exon Array does not report the entire gene as differentially expressed ($p > 0.02$), individual probesets within the gene reach much higher statistical significance levels ($p < 10^{-9}$). More interestingly, the gene appears to be composed of two “blocks”, with the first block on the 3’ end showing elevated expression in the brain, while the second block has elevated expression in the reference sample. In order to understand the more precise nature of this isoform change, it is advantageous to visualize this data in the context of known gene annotation, EST, and mRNA data. Generally, our lab uses the custom track feature of the UCSC genome browser [43], in order to export our own information and combine it with publicly available data (Figure 1.4).

In Supplementary Figure 1, we present other examples of discordance between the platforms, further illustrating the value of additional information present on the Exon Array in profiling both “whole transcript” and “isoform-level” changes.

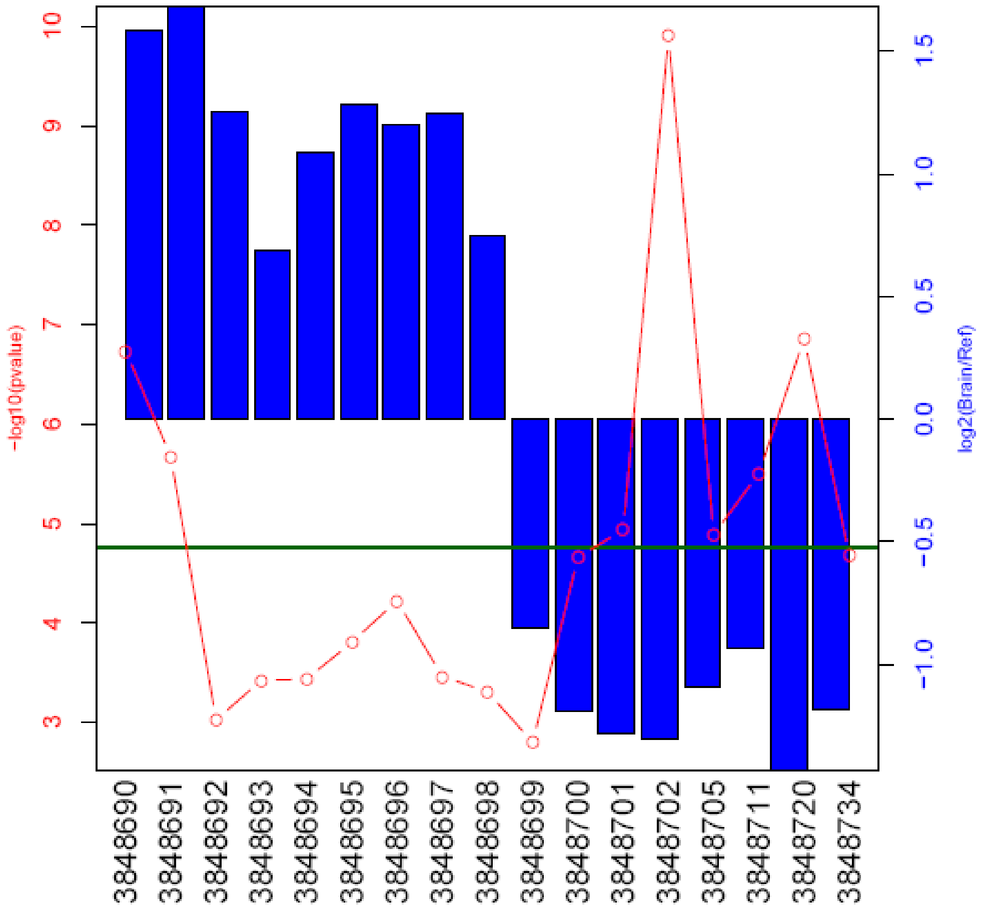

Figure 1.3: Exon array analysis of the ELAVL1 gene expression differences between brain and reference tissues

The horizontal scale corresponds to each probeset within the gene from the 3’ to 5’ ends. The height of the blue bars indicates the $\log_2$(fold change) in expression between the samples. The red line indicates statistical significance, $-\log_{10}$(p-value).

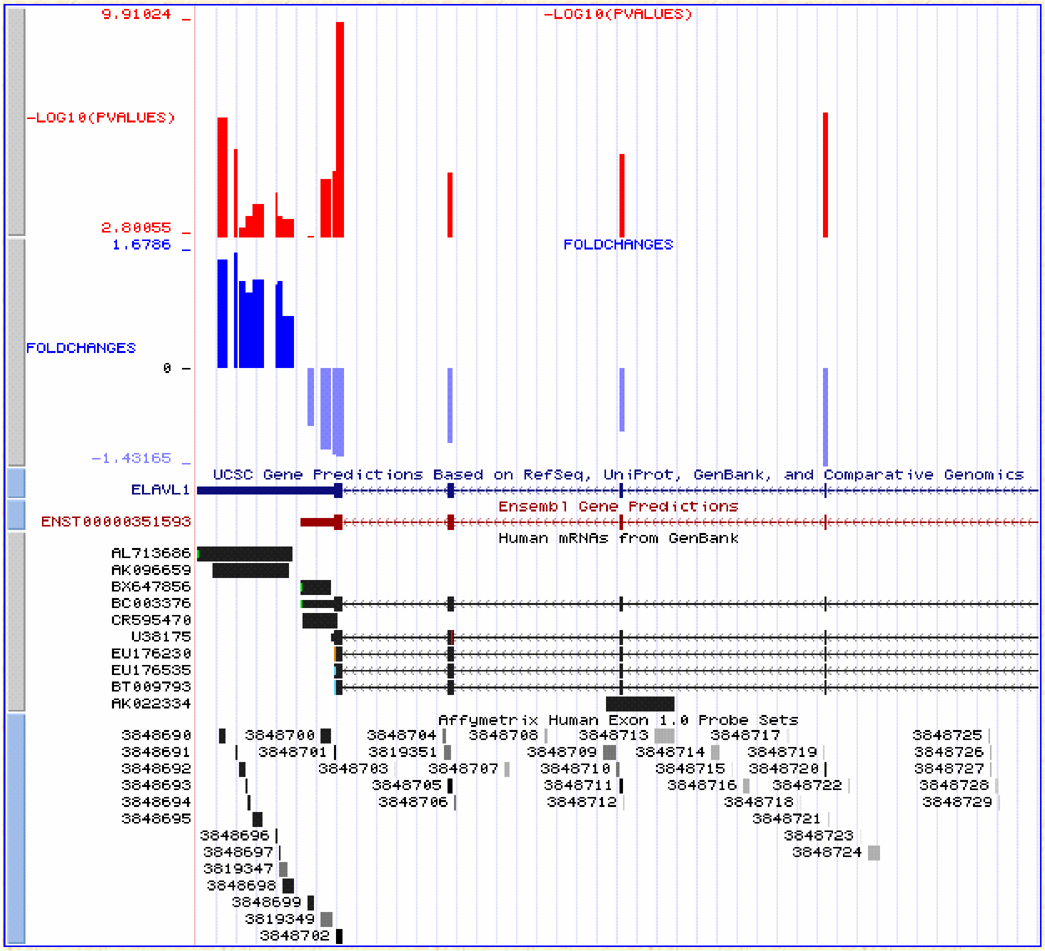

Figure 1.4: Visualization of expression patterns of ELAVL1 gene

The top two custom tracks display the Exon Array information from Figure 1.3: statistical significance and fold change. Note that the two probeset "blocks" correspond to the two isoforms of the gene. The long 3'UTR isoform is predominantly expressed in the brain, whereas the short isoform is more abundant in the reference tissues.

Differences in Amplification and Labelling Protocols.

The four most discordant genes between the 3’ arrays and the WT array (see Figure 1.5) are histone genes: *HIST1H3B, HIST1H1B, HIST1H3C, HIST1H3I,* all of which are part of the histone gene cluster on chromosome 6p21.3. The Exon Array identifies those RNAs as over 50 fold less abundant in the brain than in the reference sample, while the 3’ targeted platforms register no expression differences and very low overall expression levels. It has been shown that most histone genes lack a poly-A tail [44] and that the stability of such non-adenylated transcripts varies greatly with intracellular conditions such as those present in brain tissues [45]. Both Illumina and the Affymetrix U133 arrays use 3’, poly-T primed RNA amplification protocols and do not detect histone gene expression. In contrast, the Exon Array uses WT random primed amplification, which does not necessitate the presence of a poly-A tail. The difference of histone RNA abundance is the most striking example of a result that is specifically detected by the Exon Array, but not the other platforms. However, there are many other such differences within the dataset (see Supplementary Figure 1).

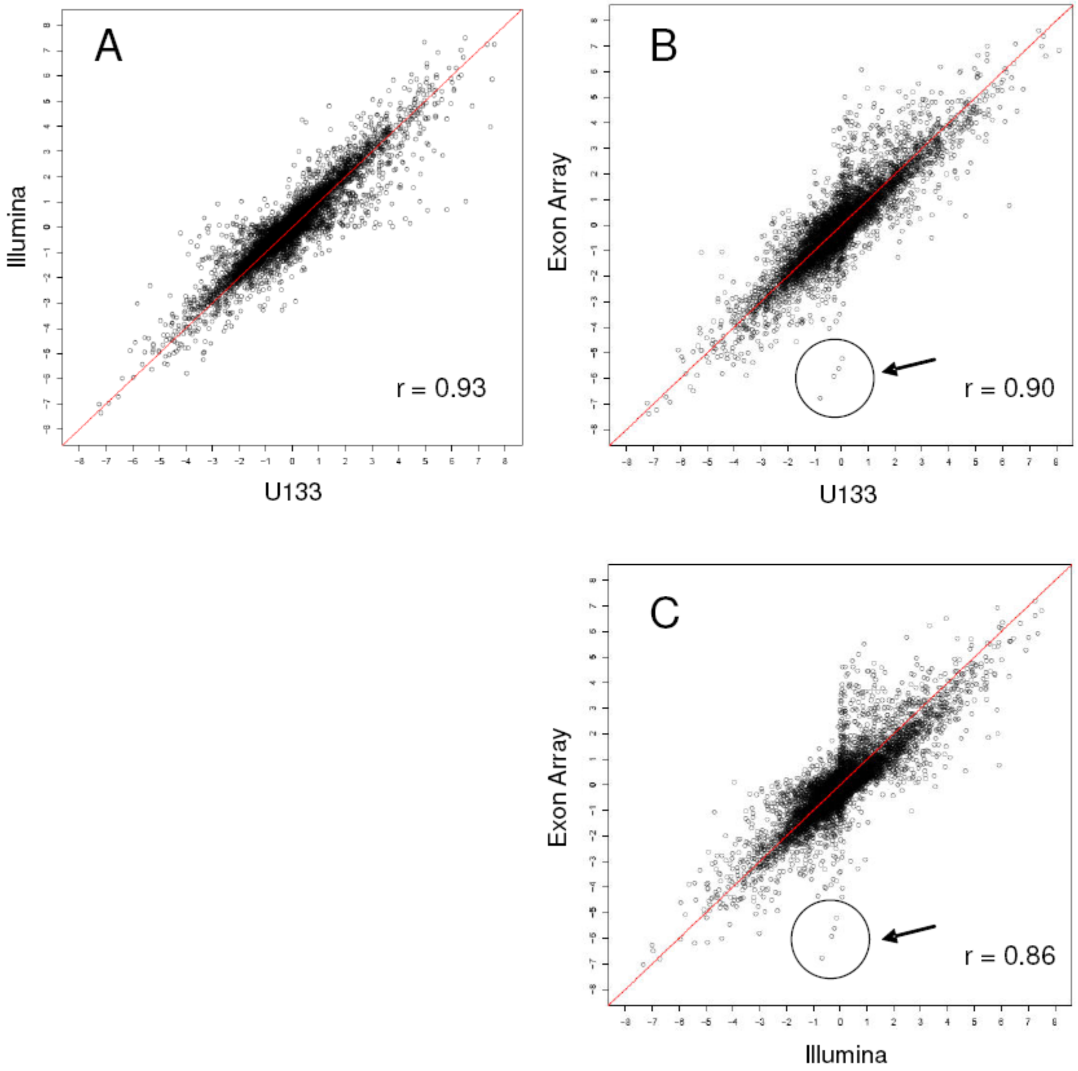

Figure 1.5: Correlation of fold changes between Affymetrix U133, Illumina, and the Affymetrix Exon Array

Fold changes ($\log_2$ transformed) between brain and reference expression levels for 8391 genes common to all three platforms. The arrow points to the highly discordant detection of 4 histone genes: *HIST1H3B, HIST1H1B, HIST1H3C, HIST1H3I*.

Using the Exon Array to Profile Alternative Isoforms.

One of the biggest challenges in profiling alternative isoforms using Exon Arrays is the deconvolution of mRNA processing and transcription. A simple comparison of probeset intensities across samples is not sufficient; if an exon belongs to a transcript that is differentially expressed, the examination of a single exon out of its genomic context will lead to an incorrect conclusion. A very simple and intuitive solution to this problem is the use of the Splicing Index (SI), which is calculated by dividing the probe set intensity by the metaprobeset intensity (i.e. exon expression/gene expression), after the addition of a stabilization constant to both the probeset and meta-probeset scores [6]. This simple procedure normalizes the expression level of each exon and accounts for any possible gene expression differences between samples. However, we find that the splicing index has some undesirable statistical properties (arising from large errors in the estimates in both the numerator and the denominator) as well as being prone to methodological artefacts (see below), and should be used with caution. Thus, we have also used a simpler, but more labour intensive method, of carrying out the entire analysis at the probeset level, and relying on visualization and manual curation (description of the workflow in chapter 2 – materials and methods – data pre-processing and analysis – Filtering signal data) of the results in order to distinguish splicing and expression differences between samples. While more robust statistical approaches are being developed, we strongly advocate visualization of results in the context of genome annotation and EST evidence in order to filter out false positive signals. We have relied on custom scripts and modifications of the UCSC and ENSEMBL genome browsers, but increasingly useful and user-friendly commercial packages for the Exon Arrays are available (e.g. Partek Genomics Suite, Biotique XRay) along with academic BioConductor packages [46-48]. Below, we describe in more detail two approaches to alternative isoform detection. For the case of simplicity, only the core (most confident) subset of Exon Array probesets was considered in this analysis.

Probe set level analysis. At this level of the analysis, each probeset (roughly corresponding to an exon) is used as a unit of expression, instead of a meta probeset (a transcript) as is done in more traditional gene expression analysis. With appropriate statistical significance cut-offs (e.g. a Benjamini-Hochberg [49] False Discovery Rate correction), it is generally possible to select a highly confident set of probesets exhibiting significantly altered expression. However, it is not immediately possible to classify the "hits" as results of alternative isoform expression (e.g. alternative splicing), differential gene expression, or both. The easiest way of factoring out of gene expression is to consider only the genes whose expression does not change across samples or treatments. That is, we can select probesets that are statistically significant, but which belong to genes whose meta-probe set expression does not appear to be significantly altered (nominal $p > 0.05$). For the MAQC samples, we generated a list of the top 100 such genes. The list and links to the UCSC browser are provided in the Supplemental File 2. The top candidates show evidence for differential promoter usage, polyadenylation, and alternative splicing. A few examples appear to be annotation errors, where the Affymetrix annotation combines two distinct genes into a single transcript cluster. In general, we advocate RT-PCR based validation of alternative isoforms. However, cross validation with existing information is also extremely useful. Extensive EST and mRNA based information on tissue specific splicing is available from many sources, e.g. from the ASAPII [50] or Hollywood [51]. Most of the source data can be viewed directly in the UCSC genome browser by displaying the mRNA, spliced EST, or AltEvents tracks.

Splicing Index (SI) analysis. SI is calculated by dividing the probe set intensity by the metaprobeset intensity. This simple procedure normalizes the expression level of each exon and should account for any possible gene expression differences between samples. An example of a successful use of SI analysis is illustrated in Figure 1.6A. Intuitively, the splicing index may be viewed as an approximate fractional inclusion level of a probeset within a transcript. However, we find many statistical and methodological problems arising from the use of the SI metric. Specifically, comparing SI values across samples makes the assumption that all probesets within a gene have comparable response (linear or log-linear) to changes in RNA levels. This assumption is generally violated, and hence SI comparisons result in high false positive rates. The most severe non-linearities in response are exhibited by probesets that are expressed close to the background levels, and probesets within highly expressed genes whose detection range is saturated. One of the most common methodological artefacts is illustrated in Figure 1.6B; probesets that are close to the 3' ends of genes are not amplified as efficiently as interior probesets while probesets close to the 5' end have elevated GC content and reduced specificity (see below). In addition, probesets that belong to skipped exons, which are included at low levels in both samples – i.e. these are actually alternatively spliced exons, but are NOT differentially spliced across samples. It should be noted that such artefacts are not limited to the use of the splicing index, and also applied to other commonly employed methods that attempt to correct for expression differences, such as the two-way ANOVA method implemented by Partek and Biotique XRay software. Some of the arising problems may be avoided by various filtering approaches; e.g. removing probesets with extremely high or low SI values, or probesets with extremely low coefficients of variation (possibly saturated). A more detailed discussion of these effects is presented at the Affymetrix website [6] and methods are being developed to enable these filtering criteria in an automated fashion [52]. Such approaches are likely to reduce false positive rates, at a cost of a reduced coverage of the genome. In Supplemental File 3 we present the top 100 candidates resulting from the SI analysis of the MAQC data, after filtering out all probesets expressed below background (average detection above background [DABG] p-value > 0.05).

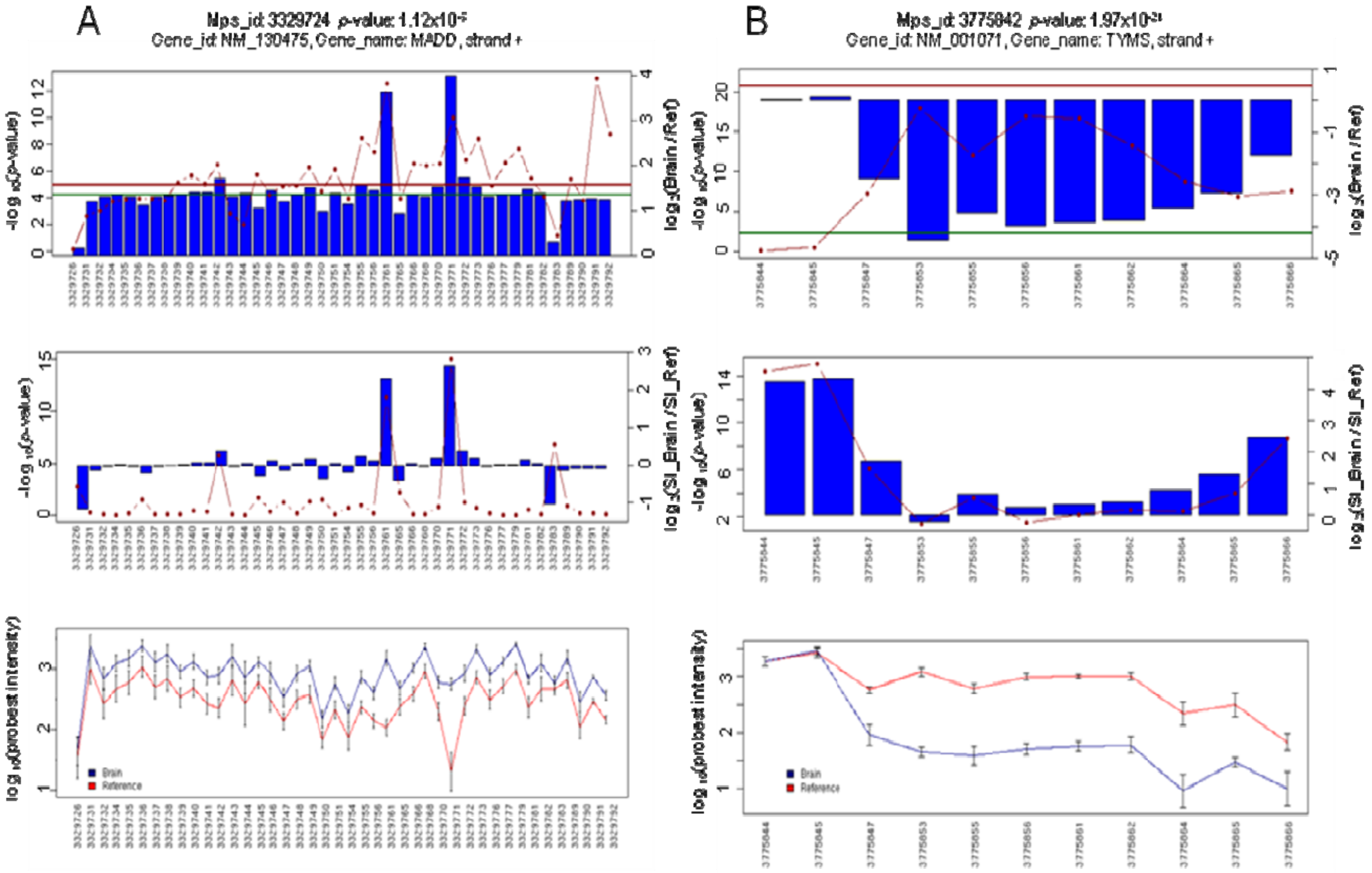

Figure 1.6 : Examples of Candidates from Splicing Index Analysis
(See the legend on the next page)

Figure 1.6: Examples of Candidates from Splicing Index Analysis

Top panels show the p-values (dotted line) and fold-changes (blue bars) for the expression of individual probesets. The centre panels show the values normalized for overall difference in gene expression (SI). Bottom panels show the raw hybridization levels of each probeset. A) MADD - successful use of the splicing index. In this example, in the presence of an overall 3-fold gene expression difference between the samples, the SI factors out the expression difference and indicates three alternatively spliced probesets – 3329761, 3329771, and 33291783 – all of which have strong supporting RefSeq annotation evidence for alternative splicing. B) TYMS - a typical false positive, where differences in probe response levels close to the edges of the transcript suggest alternative isoform usage. Such results are often erroneous, resulting from non-uniform response of individual probesets to large (in this case ~ 20 fold) changes in gene expression. Note the elevated signal intensity (bottom panel) at the 5' end of the gene, suggesting saturation, and a reduced intensity at the 3' terminus, possibly to reduced amplification efficiency.

Edge Bias Effect.

In the course of the splicing index analysis described above, we noted an excess of "hits" occurring in the 3' and 5' regions of transcripts. We hypothesized that this effect could arise partly due to a bias during the first strand synthesis in the random primed amplification step used in Exon Array processing. Briefly, first strand synthesis proceeds from the 3' end to the 5' end of each transcript, initiating at random points along the mRNA molecule. Each probeset in the interior of the mRNA is likely to be represented by multiple randomly primed initiation events. However, probesets towards the 3' end of the mRNA have a lower chance of coverage – simply because the molecule ends and priming cannot occur at any point downstream of the 3' end. In order to test this hypothesis and quantify the possible biases, we calculated mean probeset hybridization intensities as a function of distance from the 3’ and 5’ edge of the targeted mRNA molecule. The results are shown in Figure 1.7. It is evident that the intensity of the signal increases depending on the distance from the polyA site. No such effect is seen for the distance from transcription start site (5’). This effect is further illustrated in Figure 1.8, which shows that Exon Array gene expression levels are highly correlated with gene length, i.e. short genes appear to be expressed at lower levels than long genes, which is most likely caused by relatively lower efficiency in amplifying short mRNA molecules.

We also noted that the ability of the Exon Array to detect hybridization above background noise levels is not uniform across transcripts. The Exon Array allows the calculation of DABG p-values, which signify the probability that signal originates from the background noise distribution, rather than true gene expression. In general, probesets with DABG values lower than 0.05 can be accepted to represent true signal. Average DABG values are least significant at both 3’ and 5’ ends of the gene. The reduction at the 3’ end results from the reduced signal intensity levels described above. The reduction at the 5’ end is more puzzling, in the absence of a corresponding reduction in signal. We hypothesize that the 5’ effect is most likely the result of an elevated GC content of probes located close to promoter regions which are generally unmethylated, GC-rich and enriched in CpG islands [53, 54]. In fact the DABG trend at the 5’ end inversely mirrors the GC content of the probesets (data not shown).

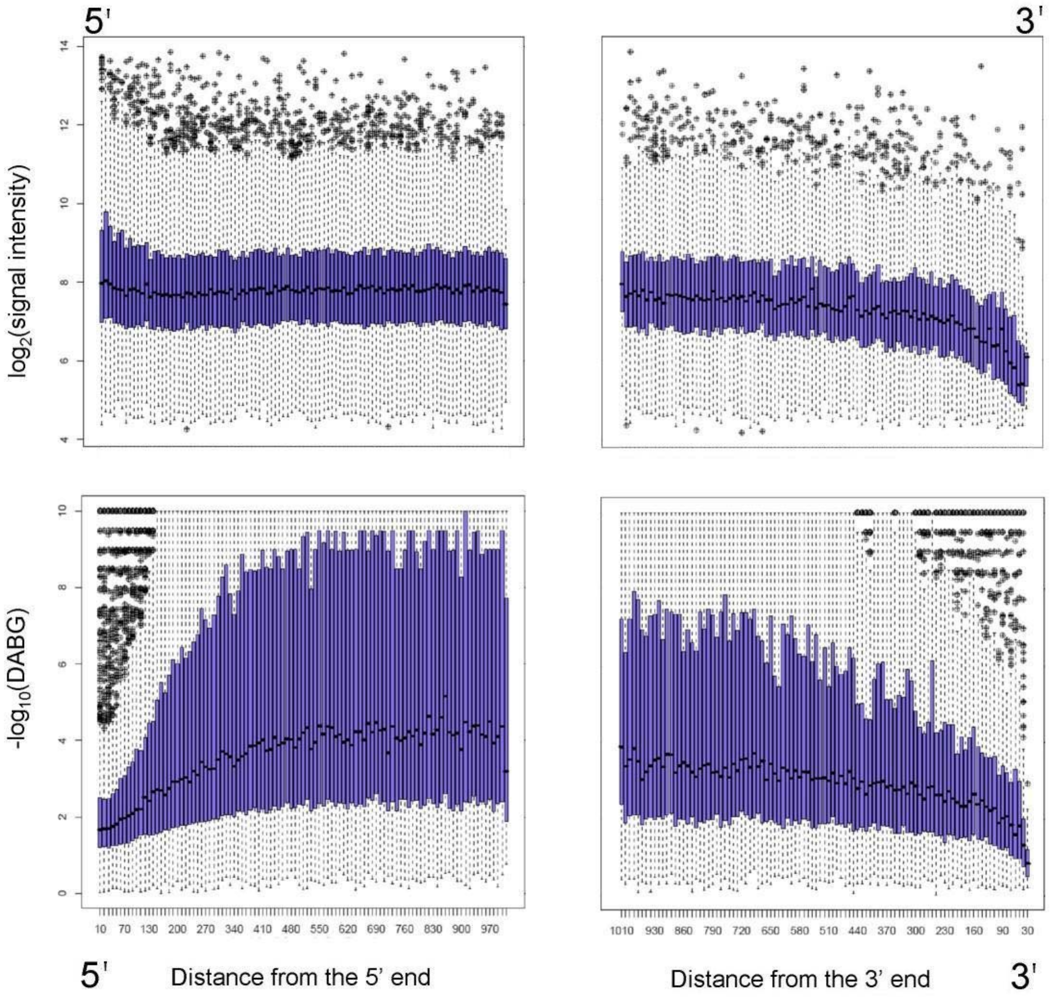

Figure 1.7: Edge bias

This figure illustrates variation of hybridization intensity across transcripts. For each probeset expressed above background levels, we determined the average hybridization intensity as a function of distance from the 5' and 3' ends of the mRNA molecule. Top panels show the average signal intensity as a function of probeset distance from the 5' and 3' ends of transcripts. A significant decrease in signal strength is seen at the 3' end, while a slight increase occurs at the 5' end. Bottom panels illustrate the ability of the array to detect the hybridization signal above background levels. Mean DABG values decrease at both 5' and 3' extremities of genes. The 3' effect results directly from the reduction in hybridization intensity. The 5' effect is most likely the result of increased GC content of the 5’ probes located close to unmethylated gene promoters and CpG islands. Both effects cause false positive results in Splicing Index and Splicing ANOVA analyses in the presence of changes in expression of the whole transcript. Only genes with detectable expression (average DABG p-value < 0.05) and total mRNA length greater than 1000 nucleotides were included in this analysis. The values were calculated as log-averages of core probeset intensity across all samples. Each point on the plot corresponds to all probeset ending within a bin of length 10 bp, at the indicated distance from mRNA termini.

In effect, probesets that are close to the ends of a gene are likely to exhibit response properties different from the rest of the transcript, and hence produce excess false positive results. Such artefacts are difficult to correct using filtering methods, because the terminal probesets in question are usually detected as expressed above background, but do not respond to expression changes as well as those in the remainder of the gene. In the future, it may be possible to correct for the edge bias by improving the amplification protocol, or computational adjustments. However, at this point interesting Exon Array results in the 3’ and 5’ ends of genes, particularly those obtained from SI or two-way ANOVA analyses, should be treated with extra caution.

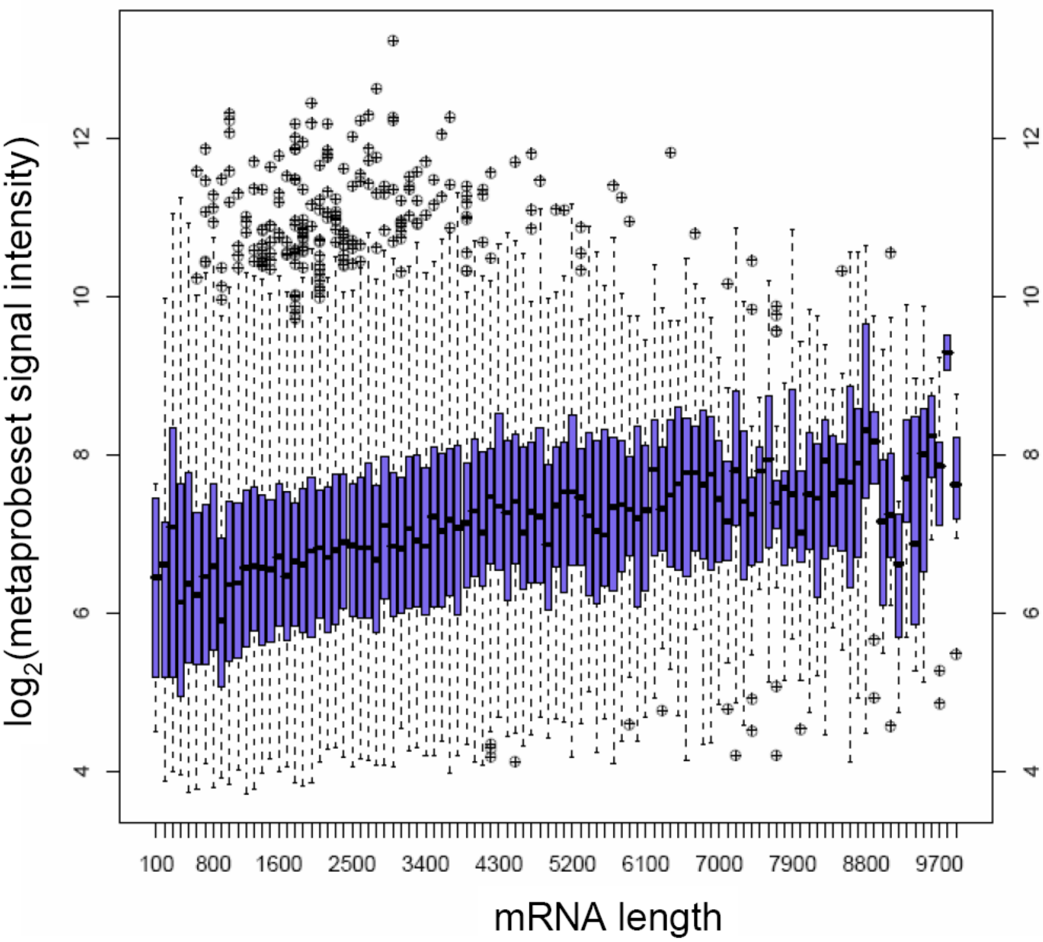

Figure 1.8: Exon Array average gene expression index as a function of transcript (mRNA) length

There is a highly significant positive correlation of expression and length (R = 0.18, $p < 10^{-20}$). This effect is most likely an artefact of the edge bias illustrated in Figure 1.7; short transcripts have a lower overall efficiency of first strand synthesis and appear to be expressed at lower levels. The effect is not observed in the 3' amplified U133 (R = 0.05) and Illumina (R = -0.03) results.

Discussion

The recognition of alternative splicing and alternative isoform expression as an important component in gene expression analysis has prompted the introduction of isoform sensitive microarray platforms. By targeting individual exons, exon junctions, and annotated isoform variants, such platforms possess the ability to profile not only the expression levels of the entire transcript, but also variations in the types of expressed isoforms. The Affymetrix Exon Array 1.0 ST is one of such commercially available platforms. To date, it has been shown that the Exon Array produces gene expression measurements that are comparable with the previous generation 3' targeted arrays. However, little is known about the in-depth level of similarities and particularly differences among WT and 3' based technologies.

This comparison utilizes the well studied brain and reference samples previously used in the MAQC study to determine sources of variability in profiling gene expression using microarrays. These samples are particularly valuable for the purposes of benchmarking the performance of the Exon Array for two reasons: 1) they allow easy comparison of gene expression level measurements with other platforms that have already been tested, and 2) they allow detection of alternative splicing and isoform difference, since neural tissues are known to be particularly prone to alternative splicing.

Our first conclusions concern the utility of the Exon Array as an expression profiling tool. We note that although the Exon Array results are very consistent with 3' profiling methods, the level of agreement between the Exon Array and 3' targeted platforms (Illumina and Affymetrix U133) is slightly lower than the agreement between the 3' platforms. There are at least two reasons for the decreased concordance.

Firstly, the Exon Array uses a whole transcript, randomly primed amplification protocol, while the two other platforms rely on polyA tail priming. As a result, the two approaches amplify a slightly different RNA pool. This is illustrated very well by the example of several histone genes (known to lack a polyA tail), which the Exon Array indicates are expressed at a much lower level in the brain than in the reference, while the other two platforms indicate a uniform very low level of expression of histone transcripts. As far as we know, differences in expression of histone genes across tissues and treatments have not previously been detected by microarray analysis, and this result is only detectable using the WT approach.

Secondly, many of the outliers in the correlation plot (Figure 1.5) are due to the presence of real variations in the expression of specific isoforms. This is illustrated using a previously noted example of the *ELAVL1* gene, which showed discordance across platforms in the original MAQC study, as well as in additional new examples (Supplementary Figure 1). The detected expression differences of transcript variants may have important biological significance. For example the longer 3' UTR in the dominant *ELAVL1* transcript in brain has a different set of putative micro RNA binding sites than the shorter 3' UTR in the reference RNA.

It should also be noted that discordant results will often be obtained because of differences in the annotation provided by microarray manufacturers. We circumvented most of such problems here by re-mapping the probes and selecting only a subset of genes that we were confident were correctly targeted by all three platforms, but researchers should keep in mind that the annotations and gene assignments provided by manufacturers contain numerous errors [55]. In the case of the Exon Array, we found that the most common annotation error resulted from joining together distinct transcripts into single meta-probesets, particularly in the case of transcripts that partially overlap. Thus, we recommend that lists of candidates from individual experiments should be carefully curated.

We also outline how the Exon Array can be used to detect alternative splicing and alternative mRNA processing events. Although are analysis methods are not in themselves novel, and most of them have been briefly described elsewhere [6, 37], our goal is to convey to the potential users their intuitive appeal and potential pitfalls. The most challenging step remains the decoupling of whole transcript expression, and individual probeset inclusion. The simplest solution to this problem is to consider only the genes that do not change overall expression levels, but contain probesets that exhibit individual variations. Although this approach produces a highly confident set of alternative events, it can result in a huge reduction of the dataset, particularly in case of comparisons across samples with highly heterogeneous gene expression levels. In the case of MAQC dataset, which has been chosen for the exact reason of it's extreme gene expression variability, imposing the restriction of expression fold change of less than 2 reduces the total number of genes considered by 31% (from 17665 to 12198).

A more inclusive approach is to attempt to correct for gene expression differences that may occur concurrently to splicing differences. We discuss two such approaches: 1) the splicing index, which compares probeset inclusion across samples after normalizing by gene expression levels, and 2) two-way ANOVA, where the interaction term between sample type and probeset can be used to indicate differential inclusion of probesets within transcripts. Both approaches suffer from similar systematic biases; they assume a uniform (linear or log-linear) response of each probeset within a meta-probeset. This assumption is violated in many cases, particularly for probesets that hybridize at very high levels (saturated response) or probesets with hybridization levels close to background (poorly or non-responsive). As a result, in the presence of significant gene expression changes, such analyses predominantly indicate three types of events: dead probesets, saturated probesets, and probesets that may be predominantly skipped (alternative), but not necessarily differentially included across samples. All three types of results constitute false positives, and contribute to the high false positive rates of such analyses.

We also point out two major systematic errors. First, we show that hybridization intensity decreases for probesets close to the 3' mRNA ends, an effect that we believe stems from the random amplification protocol used by the Exon Array. We argue that this is not an annotation artefact, but most likely results from the end of template and reduced random priming potential in the first strand synthesis step amplification. As a result, 3' regions of genes are detected at near background levels, and frequently indicate alternative isoform presence using the SI or ANOVA approaches. A similar problem exists at the 5' end of transcripts, where we hypothesize that the reduction in DABG levels is caused by the elevated GC content of the probesets. These problems are particularly troubling, since many cases of alternative polyadenylation and promoter usage may in fact be associated with changes in transcript expression. This may be due to different promoter strength, or microRNA mediated regulation in 3' UTR (as is likely to be the case in the *ELAVL1* example shown in Figures 1.3 and 1.4). Such real and potentially extremely interesting cases may be difficult to distinguish from differences in probe hybridization potential.

Many of the above systematic errors can be avoided by filtering out potentially troublesome subsets of the data: probesets with extremely low variability (saturated), probeset with low inclusion levels (close to background), and genes with extremely high differences in expression levels across samples. However, such filtering decreases the false positive rates at the cost of reduced genomic coverage.

In our earlier studies, we have also pointed out that in many experimental designs, particularly when samples originate from different genetic backgrounds (e.g. different individuals), the presence of sequence variants within probe target sequences may be a very significant source of errors [33, 37]. This effect can be especially prominent in eQTL association studies, where we have shown that it can be responsible for a false positive rate > 80% in alternative splicing analysis [56]. Thus, unless all tested samples are isogenic, we highly recommend additionally “masking” all probes containing known polymorphisms before performing the analysis.

Conclusions

In summary, the WT profiling provides a wealth of valuable information, which is either not available or misrepresented in traditional 3’ gene expression arrays. However, it should be noted that the isoform-level analysis of Exon Arrays is significantly more complicated, suffers from higher false positive rates, and requires more manual intervention than traditional gene expression analysis. We strongly advocate visualization of candidate isoform changes in the context of available genome annotation as a means to both reduce false positive rates and interpret the nature of detected variants.

Methods

Exon Array Hybridization

The Universal Human Reference RNA (catalogue no. 740000) and Human Brain Reference RNA (catalogue no. 6050) were obtained from Stratagene and Ambion, respectively. The RNA quality was assessed using RNA 6000 NanoChips with the Agilent 2100 Bioanalyzer (Agilent, Palo Alto, USA). Five technical replicates of each sample were hybridized independently at two test sites: McGill University and Genome Quebec Innovation Centre (Montreal, Quebec, Canada) and Virginia Tech (Blacksburg, Virginia, USA). Biotin-labelled target for the microarray experiment were prepared using 1μg of total RNA. The RNA was subjected to an rRNA removal procedure with the RiboMinus Human/Mouse Transcriptome Isolation Kit (Invitrogen) and cDNA was synthesized using the GeneChip® WT (Whole Transcript) Sense Target Labelling and Control Reagents kit as described by the manufacturer (Affymetrix). The sense cDNA was then fragmented by UDG (uracil DNA glycosylase) and APE 1 (apurinic/apyrimidic endonuclease 1) and biotin-labelled with TdT (terminal deoxynucleotidyl transferase) using the GeneChip® WT Terminal labelling kit (Affymetrix, Santa Clara, USA). Hybridization was performed using 5 micrograms of biotinylated target, which was incubated with the GeneChip® Human Exon 1.0 ST array (Affymetrix) at 45˚C for 16-20 hours. Following hybridization, non-specifically bound material was removed by washing and detection of specifically bound target was performed using the GeneChip® Hybridization, Wash and Stain kit, and the GeneChip® Fluidics Station 450 (Affymetrix). The arrays were scanned using the GeneChip® Scanner 3000 7G (Affymetrix) and raw data was extracted from the scanned images and analyzed with the Affymetrix Power Tools software package (Affymetrix). The microarray data has been deposited in the Gene Expression Omnibus Database (GEO: GSE13072).

Data Pre-processing and Analysis

The Affymetrix Power Tools software package (Affymetrix) was used to quantile normalize the probe fluorescence intensities and to summarize the probe set (representing exon expression) and meta-probe set (representing gene expression) intensities using a probe logarithmic intensity error model (PLIER, [57]) or robust multichip analysis (RMA, [58]). The above procedures were carried out separately for the two test sites (McGill University and Virginia Tech). The raw data (.cel files) was downloaded from the MAQC website for the Illumina and U133 arrays. In order to keep the number of replicates and test sites consistent across platforms, we only used two of the MAQC test sites (a total of 10 technical replicates of each sample). For the probeset-level analysis and alternative isoform detection, we only used the most confident subset of core probesets from the Exon Array.

Probeset and Gene Mapping

To determine a subset of genes common to the three platforms, we used the mapping provided by the MAQC study [39] to select 12091 probesets common Illumina and Affymetrix U133 arrays. Subsequently, we used the Exon Array probeset annotation and retained only the genes where the Exon Array meta-probeset coordinates contained both the Illumina and U133 probesets. This procedure resulted in 8391 genes with a high confidence concordant mapping across the three platforms.

Authors' contributions

A.B. and D.B. performed the statistical and computational analysis, and prepared the figures. T.K and D.G. carried out parts of the alternative splicing analysis. R.J. and J.M.

conceived of the study and supervised the hybridization of the microarrays. J.M. wrote the manuscript.

Acknowledgements

This work was supported by grants from CIHR and Genome Canada/Genome Quebec (to J.M.). J.M. is a Canada Research Chair recipient. D.G. is supported by a FRSQ post-doctoral fellowship. We would like to thank the staff of the Genome Quebec Microarray Platform as well as Clive Evans and the technical staff of the Core Laboratory Facility at the Virginia Bioinformatics Institute for expert help and microarray processing."

Additional files

Additional file 1

Examples of discordance between platforms. The data is visualized using custom tracks within the UCSC genome browser. We also show the location of U133 and Illumina probes for each gene. The table gives the fold change and significance levels for each platform. A. KISS1R, probable polyadenylation difference. WT profiling indicates that the expression change of the coding sequence of the gene is actually in the opposite direction to the change detected by 3’ profiling. B. CRTAC1. A whole transcript change which is only detected by the Exon Array, most likely because the 3’ methods target a non-variable UTR region. C. PSD3. Expression change detected by all three platforms, but the Exon Array identifies the nature of the isoform change – annotated alternative promoter usage. D. BCAS1. A putative alternative promoter (not annotated) indicated by the Exon Array.

File location:

http://www.biomedcentral.com/content/supplementary/1471-2164-9-529-S1.ppt

Additional file 2

UCSC browser links illustrating probeset level expression differences (fold-change and p-values) for the top 100 isoforms differentially expressed between the brain and reference samples, obtained from the probeset level analysis.

File location:

http://www.biomedcentral.com/content/supplementary/1471-2164-9-529-S2.html

Additional file 3

UCSC browser links illustrating the probeset level expression differences (fold-change and p-values) as well as the normalized (SI) differences for the top 100 isoforms differentially expressed between the brain and reference samples, obtained from the Splicing Index analysis.

File location:

http://www.biomedcentral.com/content/supplementary/1471-2164-9-529-S3.html

Chapitre II Application de la puce exon d’Affymetrix à l’épissage alternatif dans la métastase du cancer de sein

Synopsis

Dans le chapitre précédent, nous avons utilisé un amalgame de tissus, biologiquement non significatif, comme jeu de données pour nous familiariser avec la puce exon et optimiser ses méthodes d'analyses. Nous avons également mis en évidence les forces et les faiblesses de cette plate forme. Dans ce chapitre, ayant pour acquis les mises en garde de la puce exon, nous appliquons les techniques précédemment acquises à un réel système biologique qui est l'épissage alternatif dans différents stades métastatiques du cancer de sein. C'est un système peu exploré et mal compris qui occupe une grande importance dans le domaine biomédical.

Genome-wide investigation of changes in pre-mRNA splicing associated with metastasis of breast cancer

Amandine Bemmo [1,6], Cristel Dias [2], April A.N. Rose [3], Caterina Russo[3] , Peter Siegel [3,4,5] Jacek Majewski [2,6]

[1] Université de Montréal, Montreal, Quebec, Canada

Departments of [2] Human Genetics, [3] Medicine, [4] Biochemistry and [5] Anatomy and Cell biology, McGill University, Montreal, Quebec, Canada

[6] McGill University and Genome Quebec Innovation Center, Montreal, Quebec, Canada

Abstract

To identify metastatis-specific alternative splicing events (ASE), we used the Affymetrix Exon Array to profile mRNA isoform variations at the whole genome level in a breast cancer mouse model by using non-metastatic (168FARN and 4T07) and metastatic (4T1) mouse mammary tumor cell lines. Statistical analysis identified significant expression changes in 10744 out of 493710 (2%) exon probesets belonging to 2623 out of 16654 (16%) genes, corresponding to putative alternative isoforms that are differentially expressed across tumors of varying metastatic potential. A gene pathway analysis showed that 1224 of these genes have been reported to be involved in diseases and have biological functions predominantly related to cancer, cell interactions, cell proliferation, cell migration and cell death. Our analysis suggests that a large number of genes that exhibit alternative splicing or other isoform changes are associated with metastasis and that these changes may be functionally involved in the progression of cancer.

Introduction

Alternative splicing (AS) of pre-mRNA is a key post-transcriptional mechanism allowing the production of distinct proteins from a single gene. It has been estimated that over 90% of human genes undergo AS [4, 5]- creating isoforms that can have different properties and functions. A well-studied example is the gene BCL-X whose two major protein isoforms have antagonist functions [59] the short form promotes apoptosis while the long one is antiapoptotic. AS may also lead to aberrant transcripts that are targeted for degradation. Several gene instances with premature termination codons introduced by missplicing events are degraded by the Nonsense-Mediated Decay pathway [12].

Numerous pre-mRNA splicing events undergo changes or become aberrant in various types of cancer during development, progression, and/or metastasis. Some well known examples are genes CD44, MDM2, and FHIT which are implicated in tumor progression, as well as genes BRCA1 and APC implicated in breast cancer susceptibility [9]. The serine-arginine-rich (SR) proteins, a family of trans-acting splicing factors, have a key role in alternative splicing regulation of several genes, including CD44, a cell adhesion, and proliferation and migration protein. There exists evidence that SR protein quantity increases during breast cancer tumorigenesis, suggesting that this could lead to changes in AS [12]. CD44 shows an unusual splice variant in mammary tumorigenesis: a mixture of 10 internal variable exons is present in metastases whereas preneoplasias show a more restricted exon inclusion pattern [12]. In BRCA1, a skipping of exon 18 by a G-to-A transition mutation in exon 18 leads to predisposition to breast and ovarian cancer. This exon, composed of 26 amino acids, is a part of an important region for the tumor suppressor function of BRCA1 [16].

Changes in splicing during cancer progression appear to affect transformation, mobility and metastatic ability of cancer cells by altering cell-cell adhesion and cell-matrix interaction that could result in increase of cell migration and invasion [60]. A recent study revealed that several genes may influence metastatic properties in breast cancer [24]. Among these genes CD24, CA9 and EpBH2 are highly over-expressed in metastatic cells compared to non-metastatic cells [24]. The expression of the two former genes is associated with a low breast cancer survival rate of the patient [61-63]. Cancer-specific or metastasis-specific splice variants may serve as potential biomarkers for developing therapeutic drug targets.

Traditional genome-wide tools of expression analysis (DNA microarrays) consist of probes targeted to a single region of each gene and thus are limited in profiling isoform-level changes. Custom designed exon junction microarrays do target alternative splice sites, but are not efficient for the analysis of unknown alternative splicing events (ASE) and the discovery of novel isoforms [9]. The Affymetrix GeneChip Mouse Exon 1.0 ST (Exon Array), a recent tool to investigate AS has been applied successfully in many studies [25, 26, 30]. It allows the expression profiling of over a million individual – known and predicted - exons. In this study, we used the Exon Array to identify metastasis-specific AS and isoform variations in a breast cancer mouse model. We analyzed both exon and gene expression in five tumor tissues derived from cell lines ranging from non-metastatic to highly metastatic: 67NR, 168FARN, 4T07, 665C14 and 4T1 [64]. The criterion used for the metastatic classification is based on the level of metastatic nodule growth in lung. The non-metastatic tumor cell line 67NR forms primary tumors but does not proliferate to distant tissues. 168FARN is weakly detected in lymph node but also fails to cause extravasation. Cells of the 4T07 cell line reach the lung via the blood but are unable to develop metastatic nodules. 66CI4 cells reach the lung and form visible metastatic nodules. Lastly, 4T1, the most metastatic, spontaneously metastasizes to distant sites, namely lung, bone and liver, by the formation of visible nodules in these organs. Because of technical reasons (see Appendix), tumors originating from the supposed 67NR and 66Cl4 cell lines were excluded from the final analysis presented here. Using the remaining three cell lines, metastasis-specific isoform variants were identified using statistical analysis of the Exon Array data. A gene pathway profiling was performed on candidate genes using the Ingenuity Pathway Analysis (IPA version 6.0) software package (Ingenuity Systems, Mountain View, CA). Nearly 16% of genes display expression or isoform variations across tumors. Half of those genes are known to be associated with cancer pathogenesis, while the remaining candidates may constitute novel candidate isoforms involved in metastasis.

Results

To identify AS associated with metastasis, we measured global exon expressions in five tumors ranging from non-metastatic to metastatic: 67NR, 168FARN, 4T07, 66C14 and 4T1. We obtained five biological replicates for 67NR, 4T07 and 66CI4, and four biological replicates for samples 168FARN and 4T1. We used the Exon Array to profile the expression at two levels: the transcript-level and exon-level. The latter of which consists of over a million of known and predicted exons.

Gene expression patterns in tumors

At the first stage of the analysis, we carried out a Principal Component Analysis (PCA) (Figure 2.1) on the gene expression values to identify the variables that best explain the variance within our dataset. Several observations could be made from this analysis. As expected, in a two dimensional plot showing the first two PCA components accounting for most of the variance in the data, the biological replicates within each tumor type cluster together. Two of the replicates did not cluster with their respective tumor types, and were removed from further analysis, as they are indicative of anomalous behaviour within their class. The tumor samples 168FARN and 4T07 cluster closely together, which is expected, since they have similar levels of metastatic potential. Tumors from 66C14 are the most divergent from the rest. Unexpectedly, tumors from 67NR and 4T1 appear quite similar in their expression profiles and cluster closely together. Since these two types have the most difference in metastatic potential, our initial expectation was that they would have the most divergent gene expression profiles.

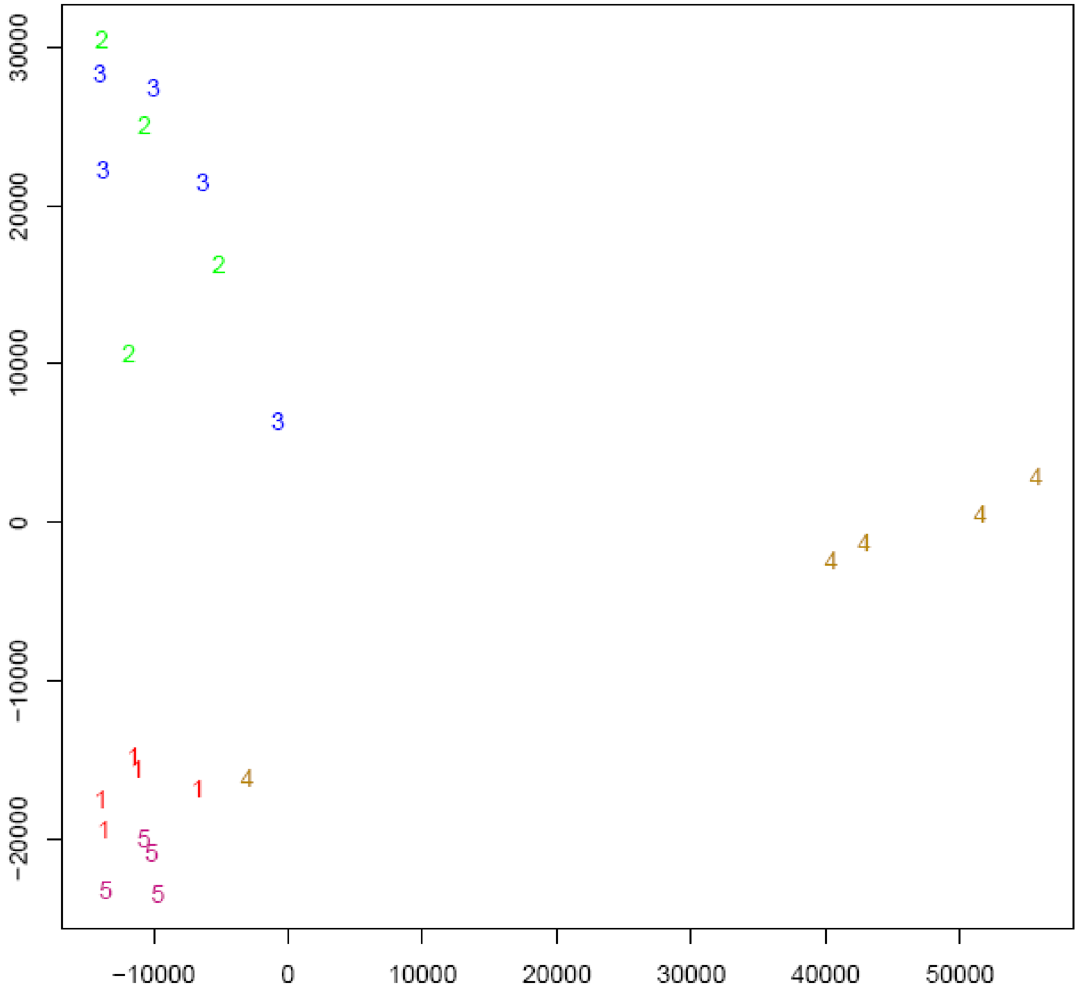

Figure 2.1: PCA plot at the metaprobeset level shows the clustering of samples

Each number corresponds to a tumor sample type (1:67NR, 2:168FARN, 3:4T07, 4:66C14 and 5:4T1), and the frequency of a number represents the number of biological replicates of the corresponding sample. Biological replicate outliers are circled.

In view of the unexpected similarity of expression profiles of samples 67NR and 4T1, we re-examined their metastatic potential using spontaneous metastasis assay in the lung. A spontaneous metastasis assay is a diagnostic test where tumor cells are injected into the orthotopic site (mammary fat pad) where they must form a primary tumor before they can metastasize to either the lung or the bone. This assay revealed that 67NR is in fact highly metastatic. On the other hand, the 66Cl4 cells are very poorly metastatic in this experiment. Based on the results of this test, we suspected that either a spontaneous mutation event, or a sample mislabelling at the source (the cell lines were obtained from different sources, see Materials and Methods) has occurred. This problem is in fact, and unfortunately, a common occurrence in cell line culture collections [65]. Hence, for further analysis we excluded samples 67NR and 66C14 and included only the tumors derived from cell lines with confirmed correct metastatic character (168FARN, 4T07, and 4T1).

Tumor-specific gene expression and isoform variations

We analyzed the full exon annotation (including predicted, non-core exons), and the core gene annotation. The gene-level intensity is estimated by combining the exons that belong to annotated transcript clusters (genes). Consequently, probesets outside of a transcript cluster is discarded from the analysis. The excluded probesets proportion constitutes more than the half of total number of probsets on the Exon Array. Therefore, we investigated 493710 probesets (roughly corresponding to exons) belonging to 16654 core metaprobesets (corresponding to transcripts), respectively. We applied several filtering steps to discard genes and exons with expression close to the background to minimize the false positive rate (see Methods). We obtained 183610 expressed probesets belonging to 11082 genes. To determine the statistical significance (p-value) of expression changes across samples, we carried out a one-way ANOVA test on probeset and metaprobeset expressions. At the probeset level, we performed two concurrent analyses: a probeset expression-intensity analysis and a probeset gene-level normalized intensity analysis. The gene-level normalized intensity is the ratio of the probeset expression intensity to the expression intensity of the metaprobeset that the probeset belongs to. The splicing index (SI) for a probeset is then defined as the ratio of gene-level normalized intensities in one sample relative to another. Subsequently to ANOVA tests, we applied a 0.05-level FDR (False discovery rate) correction to determine the cutoff p-value to retrieve significant transcripts. Significant variations where located by performing pairwise T-test comparisons (168FARN and 4T07 against 4T1). The 2-logarithmic expression fold-changes (4T07/4T1 and 168FARN/4T1) between paired samples comparisons were also computed.

The p-value cutoff of 6.36×10^{-4} (for the probeset expression analysis) and of 7.10×10^{-4} (for the SI analysis), corresponding to the 5% FDR, both yielded 10744 (2.18%) probesets showing significant expression changes belonging to 2623 (15.75%) metaprobesets. At the metaprobeset level, the statistical significance of the variation was determined by a p-value cutoff of 8.46×10^{-3} corresponding to the 5% FDR. We identified 1772 (10.64%) metaprobesets that show expression changes at the whole transcript level and 851 (5.11%) showing transcript-isoform changes without corresponding whole gene expression changes.

To visualize ASE in the context of EST/mRNA or genome annotation, we uploaded our data onto the UCSC Genome Browser [66]. For each gene, we plotted the T-test p-values and the expression fold-changes of each individual exon (examples shown in Figures 2.2-2.7).

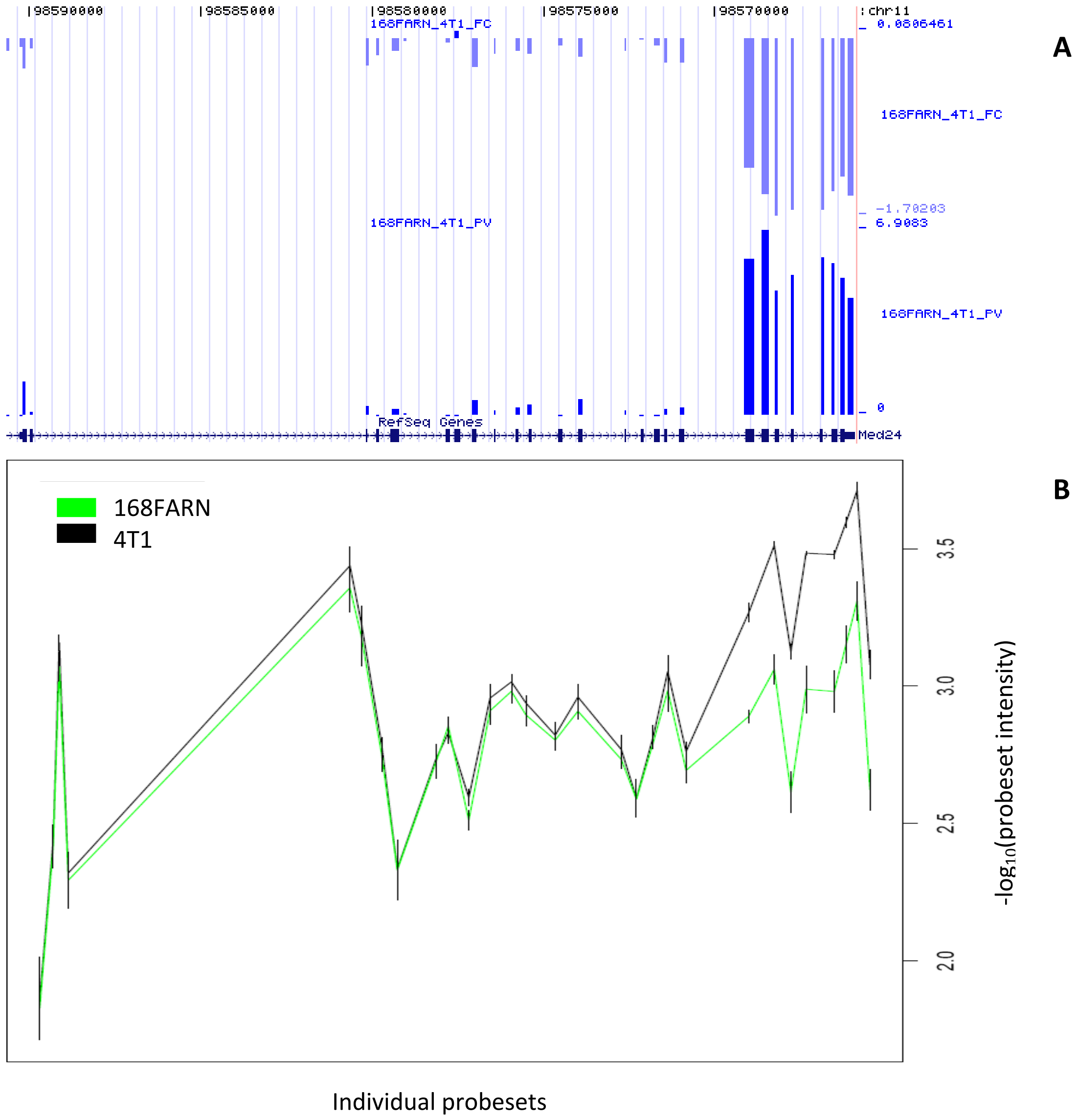

Figure 2.2: Visualization of the expression pattern of MED24 gene showing an alternative initiation or termination event

A - The horizontal scale corresponds to each probeset within the gene from the 5' to 3' ends. The blue bars indicate the comparison between 168FARN and 4T1 samples. From top to bottom we plotted the $\log_2$(fold-change) in expression, between the samples compared, and the statistical significance, $-\log_{10}$(p-value). **B -** $\log_{10}$(expression intensity) of individual probesets (from panel A) in samples 168FARN and 4T1. In this example, the seven last probesets and the 3' UTR region over-expressed in 4T1 indicate an alternative end or an alternative initiation in the gene isoform.

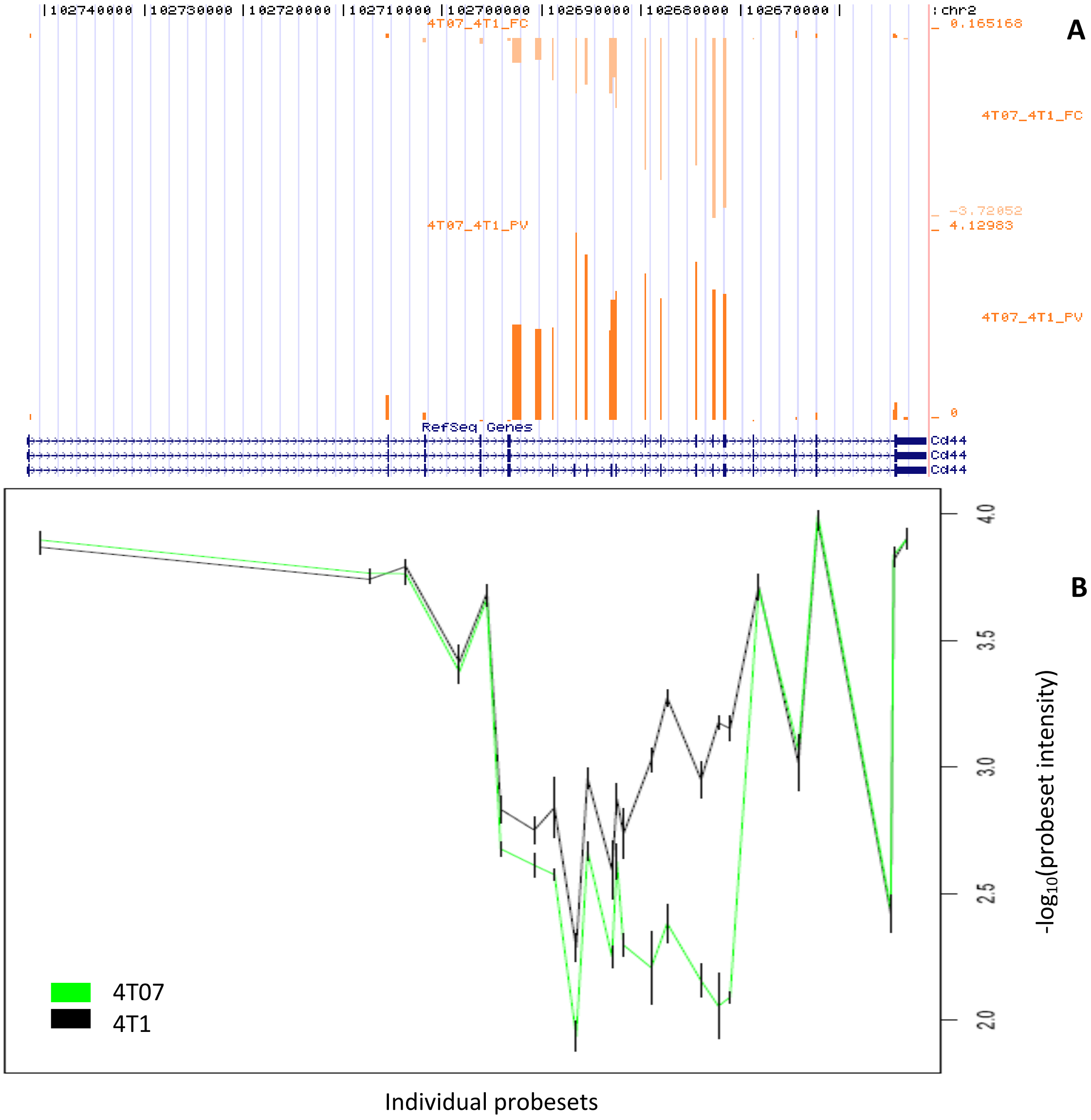

Figure 2.3: Visualisation of the expression pattern of CD44 gene showing several internal cassette exons and intron inclusions

A - The horizontal scale corresponds to each probeset within the gene from the 5' to 3' ends. The orange bars indicate the comparison between 4T07 and 4T1. From top to bottom we plotted the $\log_2$(fold-change) in expression, between the samples compared, and the statistical significance, -$\log_{10}$(p-value). **B –** $\log_{10}$(expression intensity) of individual probesets (from panel A) in samples 4T07 and 4T1. In this example, exons 8, 11 and 13, two intronic sequences between E6 and E5, and one intronic sequence between E8 and E9 are over-expressed in 4T1 sample.

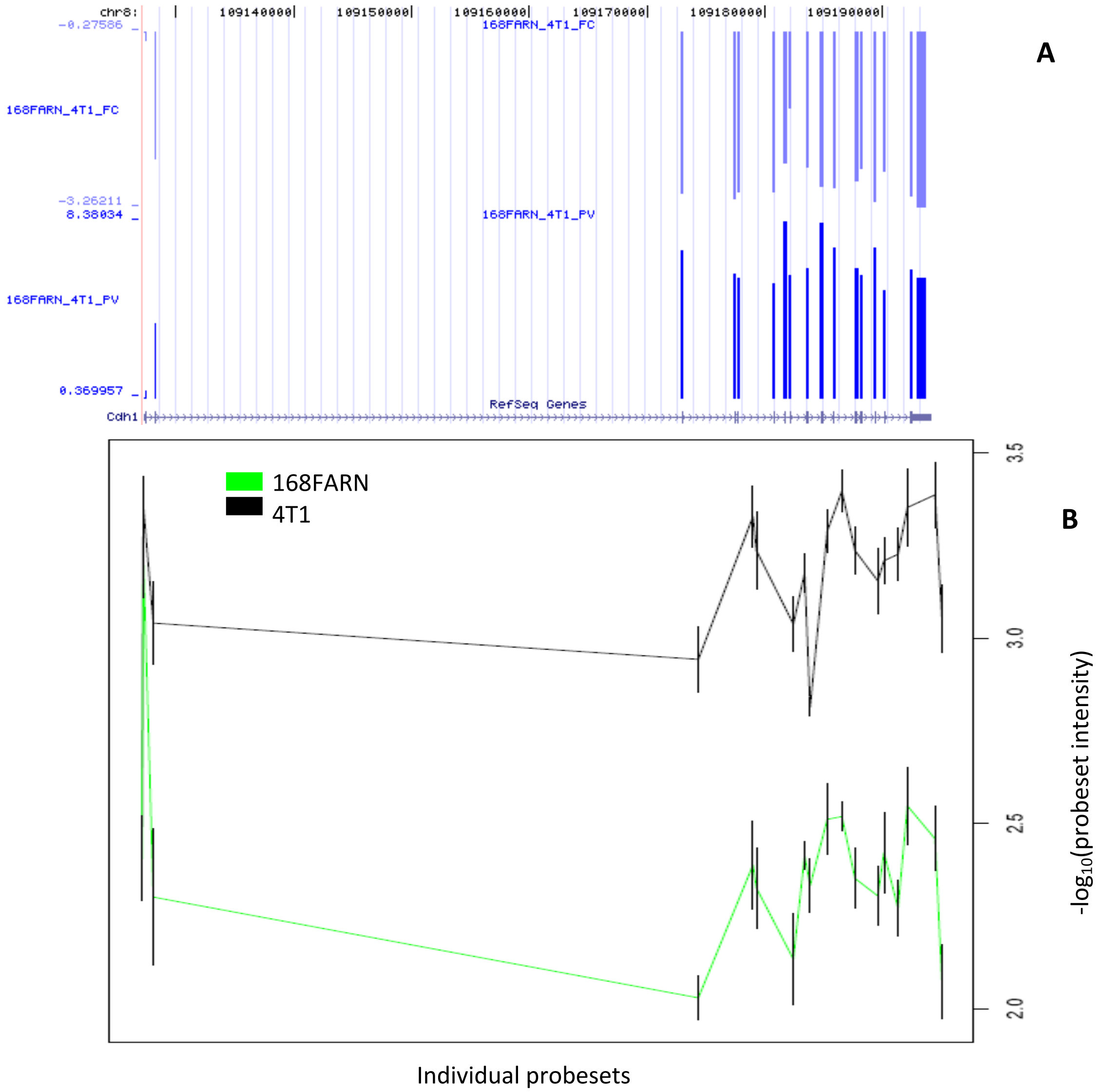

Figure 2.4: Visualisation of the expression pattern of CDH1 gene showing a whole gene expression change

A - The horizontal scale corresponds to each probeset within the gene from the 5' to 3' ends. The blue bars indicate the comparison between 168FARN and 4T1. From top to bottom we plotted the $\log_2$(fold-change) in expression, between the samples compared, and the statistical significance, -$\log_{10}$(p-value). **B** - $\log_{10}$(expression intensity) of individual probesets (from panel A) in samples 168FARN and 4T1. CDH1 is predominantly expressed in 4T1 sample comparatively to 168FARN and 4T07.

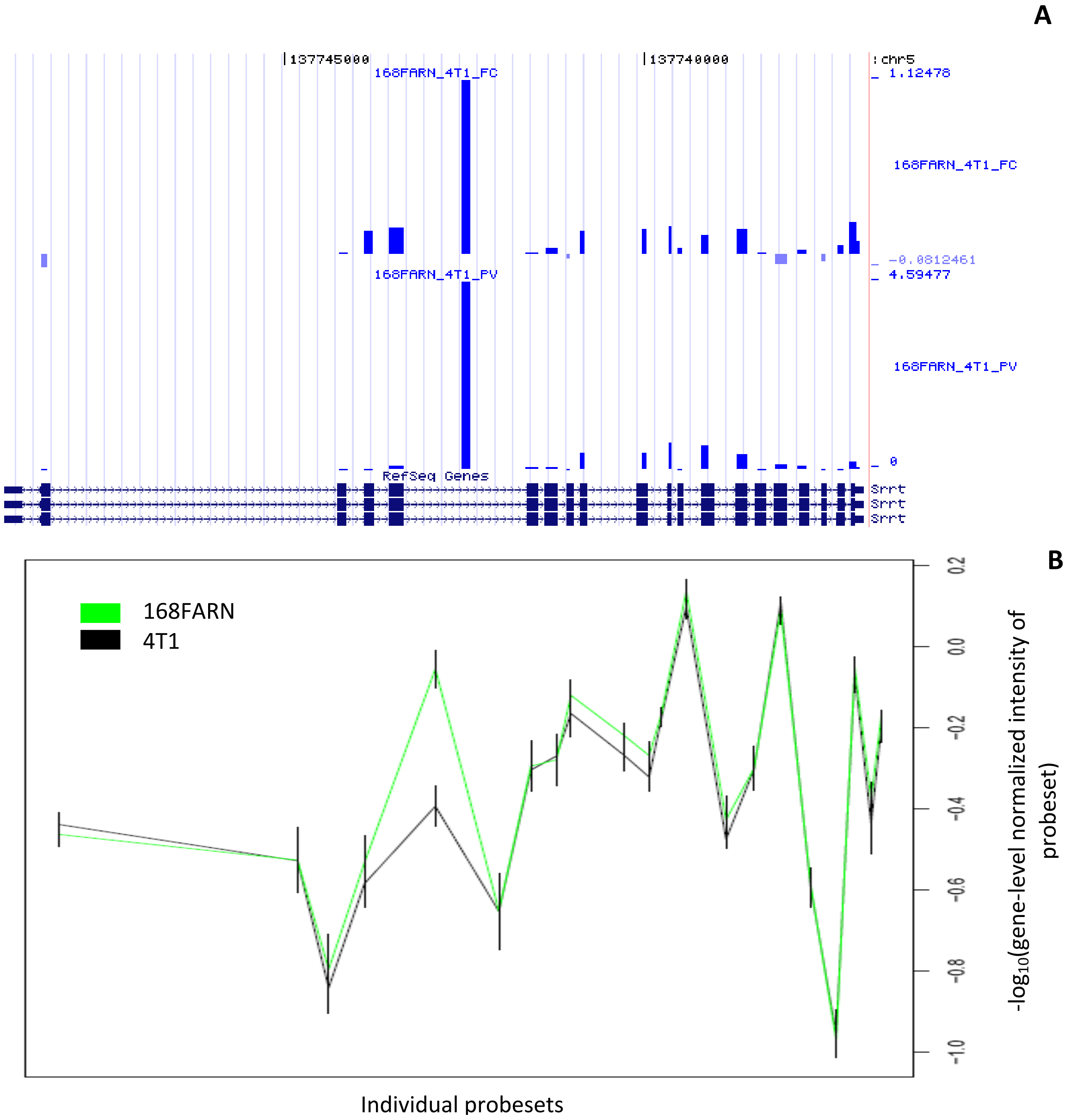

Figure 2.5: Visualisation of the expression pattern of SRRT gene showing an intron inclusion

A - The horizontal scale corresponds to each probeset within the gene from the 5’ to 3’ ends. The blue bars indicate the comparison between 168FARN and 4T1 samples. From top to bottom we plotted the the $\log_2$(fold-change) in gene-level normalized intensity between the samples compared and the statistical significance, $-\log_{10}$(p-value). **B -** $\log_{10}$(gene-level normalized intensity) of individual probesets (from panel A) in samples 168FARN and 4T1. We note an intron inclusion between exons 4 and 5 in samples 168FARN and 4T07.

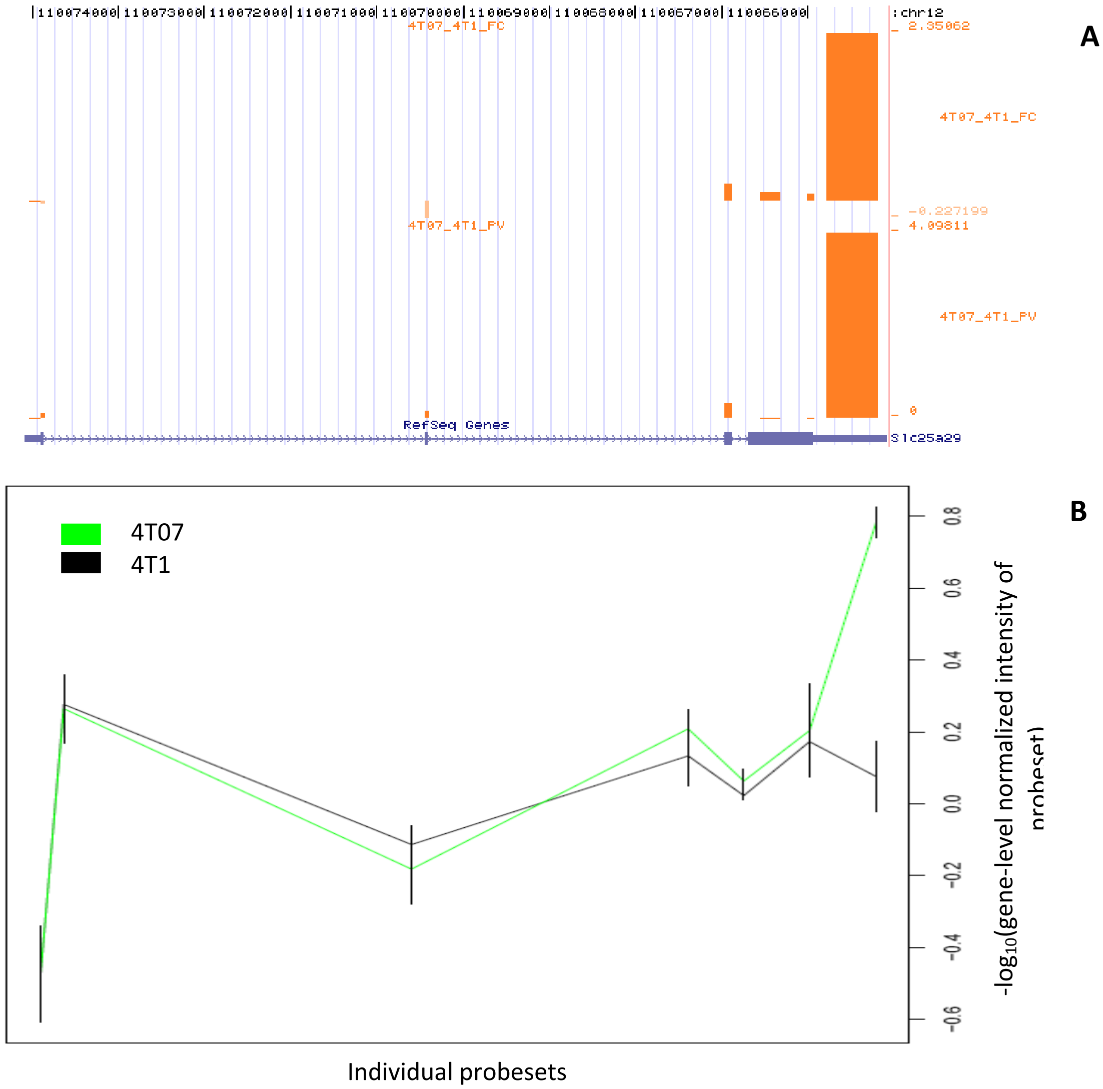

Figure 2.6: Visualisation of the expression pattern of SLC25A29 gene showing a differential 3’ UTR site

A - The horizontal scale corresponds to each probeset within the gene from the 5’ to 3’ ends. The orange bars indicate the comparison between 4T07 and 4T1 samples. From top to bottom we plotted the $\log_2$(fold-change) in gene-level normalized intensity between the samples compared and the statistical significance, $-\log_{10}$(p-value). **B -** $\log_{10}$(gene-level normalized intensity) of individual probesets (from panel A) in samples 4T07 and 4T1. The 3’UTR site is over-expressed in 4T07 and 4T1 whereas it is weakly detected in 4T1.

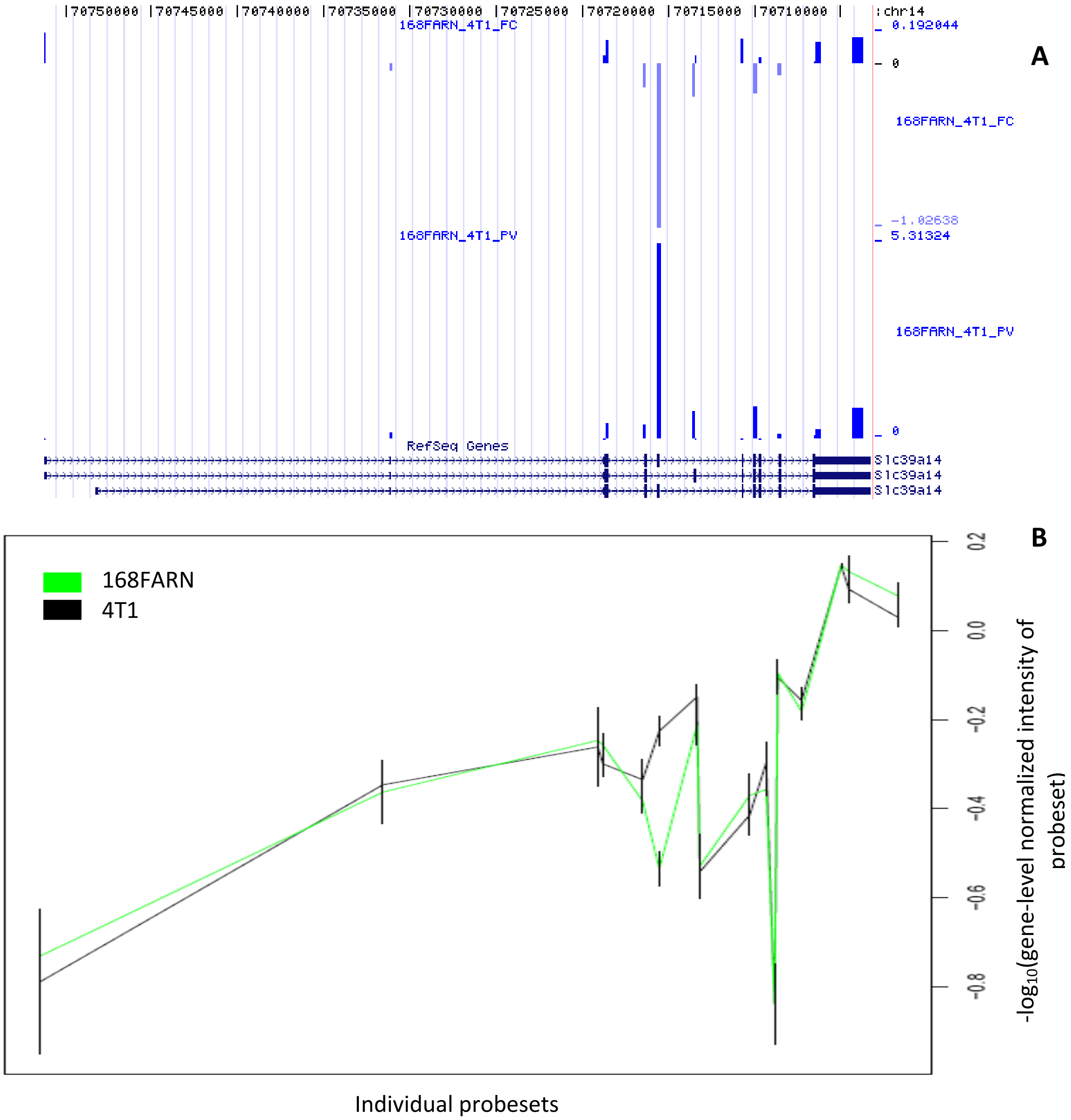

Figure 2.7: Visualisation of the expression pattern of SLC39A14 gene showing a cassette exon

A - The horizontal scale corresponds to each probeset within the gene from the 5' to 3' ends. The blue bars indicate the comparison 168FARN and 4T1. From top to bottom we plotted the $\log_2$(fold-change) in gene-level normalized intensity between the samples compared and the statistical significance, $-\log_{10}$(p-value). **B -** $\log_{10}$(gene-level normalised intensity) of individual probesets (from panel A) in samples 168FARN and 4T1. Exon 5 is predominantly expressed in 4T1 sample.

Using this visualization, we manually curated the results to generate a list of 203 top candidates (143 and 60 from the expression intensities analysis and SI analysis, respectively. Details reported in Additional File 1 and Additional File 2) exhibiting changes that could be confidently classified into isoform variation categories, while the rest of genes (92.3%) present complex expression patterns that are difficult to interpret (Figure 2.8). The top candidates show evidence for differential promoter usage, polyadenylation, ASE and whole gene expression changes. We calculated the proportion of each isoform variation type among our classified candidate genes (Figure 2.8): 26.1% of genes showed whole gene expression changes with some of them showing additional splicing changes. A large proportion of genes showed only isoform changes (examples in Table 2.1), namely intron inclusion or inclusion of cryptic, unannotated exons (46.4%) and cassette exon usage (13.5%). 7.2% of isoform changes occurred at the level of transcript initiation or transcript termination. Thirteen genes show changes within the UTR regions: three genes have differential 5’ UTR changes and 10 present 3’ UTR changes. We found only one gene showing an alternative 5’ splice site.

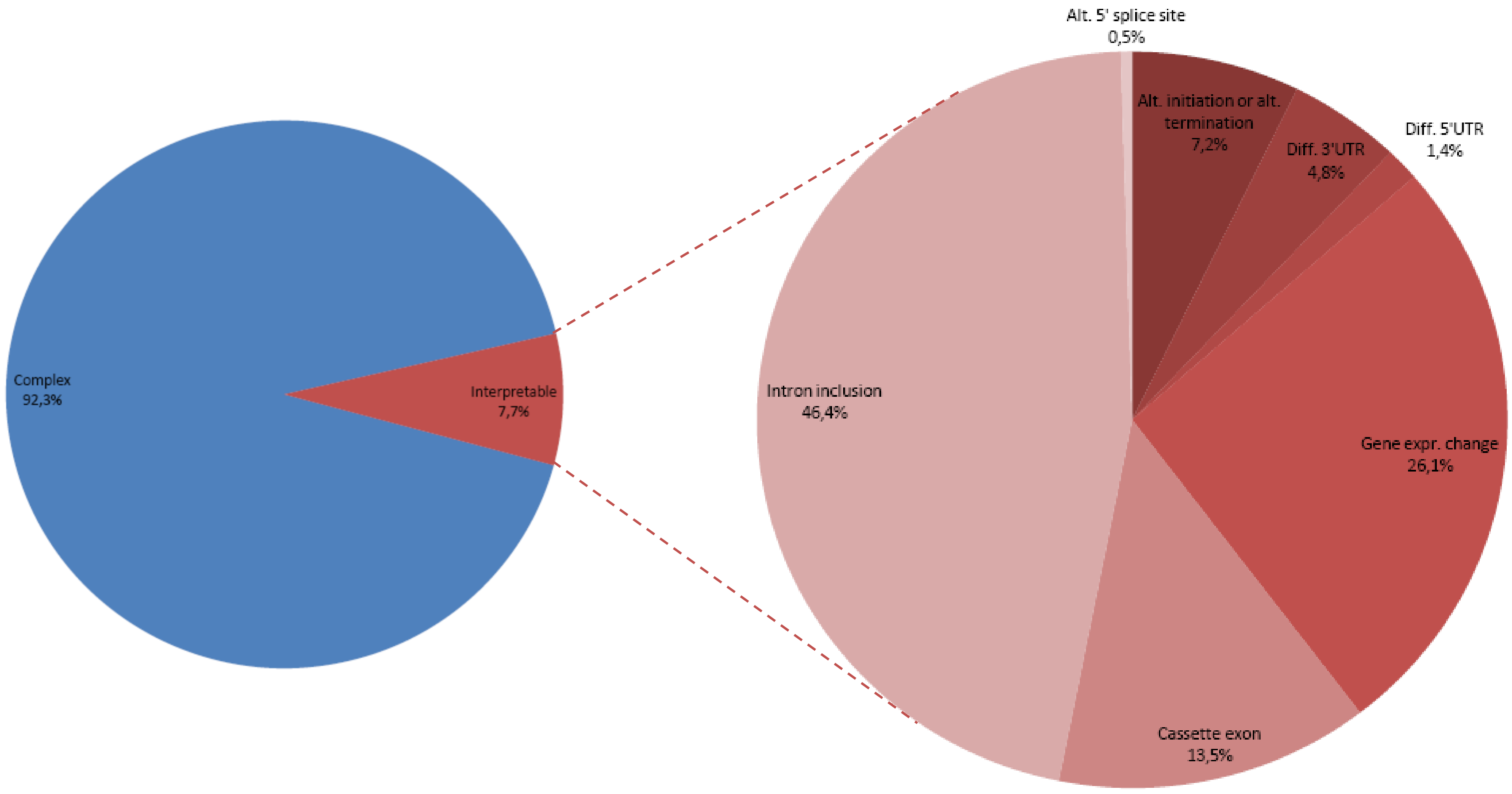

Figure 2.8 : Proportion of gene expressions variation
(See the legend on the next page)

Figure 2.8: Proportion of gene expressions variation

The left pie chart represents the total significant genes: 7.7% of them are interpretable while 92.3% are not. The pie chart on the rigth shows the percentage of interpretable genes splitted according to the nature of gene variations: whole gene expression change (Gene expr. changes), alternative initiation or alternative termination (Alt. Initiation or alt. termination), cassette exon, alternative 5' splice site (Alt 5' ss), differential 3' UTR (Diff. 3' UTR), differential 5' UTR (Diff. 5' UTR) and intron inclusion.

Our analysis identified several interesting examples of isoform variants within our samples. For example, we found two trans-acting splicing regulator factors that present isoform changes: HNRNPH1 and CLK1. The first example, HNRNPH1, is a member of HNRNP protein family and retains an intronic sequence between exons 9 and 10 in tumor samples 168FARN and 4T07. The HNRNP proteins are required for pre-mRNA processing and maturation. They bind to newly synthesized RNA in the nucleus until they are exported to the cytoplasm. Interestingly, a frameshift mutation in HNRNPH1 has been previously identified in gastric cancer [67]. The second example, CLK1, shows an intron between exons 5 and 6 that contains a highly expressed sequence in 4T01. This gene codes for a member of the CDC2-like family. Expressed in the nucleus, this protein phosphorylates other serine/arginine-rich proteins whose concentration is involved in the regulation of the splices sites selection in pre-mRNA maturation [68].

Table 2.1: List of some alternatively expressed probesets

The gene name[1], the probeset ID[2] and the relative probeset location[3] in the gene are indicated. For each pairwise comparison, the T-test P-value[4] and the $\log_2$(fold-change)[5] are given. The nature of the isoform change[6] is shown (CE: cassette exon, II: intronic sequence inclusion, 3' UTR: differential 3' UTR). An existing Refseq, mRNA, or EST supporting the event is mentioned[7].

I(Ex-Ey): Intron between exon x and exon y.

Gene name[1]	PS[2]	PS location[3]	168FARN vs. 4T1		4T07 vs. 4T1		ASE[6]	Evidence[7]
			P-value[4]	FC[5]	P-value	FC		
CD44	4534496	E13	3.14×10^{-04}	2.81	3.38×10^{-04}	2.64	CE	Yes
	4740112	E11	6.39×10^{-04}	2.77	6.39×10^{-04}	2.72	CE	Yes
	4461784	I (E9-E10)	1.94×10^{-04}	2.06	1.55×10^{-03}	1.46	CE	No
	5425762	E8	1.02×10^{-05}	1.50	2.42×10^{-04}	0.96	CE	Yes
	4423264	I (E5-E6)	3.61×10^{-05}	1.51	2.28×10^{-03}	0.82	II	No
	4622064	I (E5-E6)	9.18×10^{-05}	1.18	8.16×10^{-03}	0.52	II	No
Itgb1	5044002	I (E8-E9)	2.00×10^{-05}	1.58	1.12×10^{-03}	0.89	II	Yes
Slc25a29	4968317	3' UTR	9.56×10^{-05}	-2.17	9.56×10^{-05}	-2.15	3' UTR	Yes
MAPK14	4487560	I (E1-E2)	1.47×10^{-01}	-0.18	2.05×10^{-04}	-0.79	II	No
Msx1	4993066	3' UTR	2.81×10^{-03}	0.88	5.22×10^{-04}	1.19	3' UTR	No
Srrt	5382632	I (E4-E5)	2.54×10^{-05}	-1.12	7.90×10^{-06}	-1.36	II	No
MFi2	5508279	E13	1.25×10^{-04}	-1.88	1.54×10^{-06}	-3.37	CE	No

Pathway analysis of splicing events

We performed a pathway analysis on all statistically significant genes by using the IPA software to identify pathways enriched for particular cellular functions and relevance to disease. Among the 2623 differentially expressed genes and isoforms, 1224 genes have been previously reported to have well-annotated functions in normal biological processes as well as human diseases. Our analysis identified pathways involved in cellular growth and proliferation, cellular death, tissue development, cell to cell signalling and interaction, cellular movement, genetic disorder and cancer (Table 2.2 and Table 2.3).

We also noted in the pathway analysis some genes, such as IFT172, ACSBG1, MED24, AGRN and CPXM2, that have not yet been distinctly associated with any disease, but belong to pathways that are implicated in cancer pathogenesis. For example in MED24 (mediator complex subunit 24), the seven last exons and the 3' UTR region of the transcript are predominantly expressed in 4T1 sample compared to samples 168FARN and 4T07 (Figure 2.2). This expression pattern could potentially create a protein with unknown activities. The constitutive isoform encodes a subunit of the mediator complex TRAP (a transcriptional coactivator complex necessary for the expression of almost all genes). This gene acts indirectly on VDR [69], a 4T1-upregulated gene involved in the decreasing of tumor cell death by inhibiting p38 activities that induce tumor cells in colon cancer (Table 2.3). Hence, pathway analysis may help us identify new genes associated with cancer and metastasis.

Table 2.2: Top biological functions and diseases retrieved by the gene pathway analysis

For each function or disease, the number of significant genes involved is mentioned. A gene could be involved in more than one function or disease.

Function or disease	# of genes
Cancer	611
Genetic disorder	567
Cellular growth and proliferation	467
Cellular death	423
Cellular development	314
Tissue development	282
Cellular movement	254
Cell-to-cell signalling and interaction	239

The pathway analysis shows many significant genes and complex interactions that regulate cell growth, cell interactions, cell death and cell movement. Some of the most interesting candidates are described in detail below. The CD44 gene, a cell adhesion, proliferation and migration protein, has been previously reported to present an inclusion of 10 internal variable exons (from exons 6 to 15) in mammary tumorgenesis [12]. CD44 isoform variations in cancer are associated with the metastasic potential of tumor cells, being involved in numerous processes, including cell proliferation, adhesion and invasion [70]. This protein increases the adhesion and invasion of tumor cell lines in breast cancer, and decreases cell death and apoptosis of tumor cell lines in colon cancer (Table 2.3). We have identified a novel isoform of CD44 showing retention of intronic sequences (Figure 2.3); in this variant, two introns between exons 5 and 6, and one intron between exons 9 and 10 contain highly expressed sequences in the most metastatic tumor sample 4T1. We also note in this isoform a high inclusion rate of exons 8, 11 and 13 in sample 4T1. In the pathway, CD44 binds to MAPk1, a member of the ERK (extracellular signal-regulated kinase) complex, and also to a complex of collagen proteins (Figure 2.9). Collagen is the major constituent of the extracellular matrix and basement membrane where it plays an essential function in the organization of cells.

Table 2.3: List of some significant genes that have important implications in normal biological processes and diseases

For each gene, the symbol[1], the ENTREZ gene name[2], the biological function[3] and the type of splicing event are given[4]. Diseases[5,6] where the gene has been implicated are also indicated.

Gene symbol[1]	Entrez gene Name[2]	Biological function[3]	Expression pattern[4]	Diseases and biological process[5]	Diseases implication[6]
CD44	CD44 molecule (Indian blood group)	regulation of cell growth; cell adhesion; cell-matrix adhesion; cell-cell adhesion	High inclusion of introns and variable exons in 4T1	metastasis, infection, atherosclerosis, tumorigenesis, renal cancer, papillary renal cell carcinoma, gastric cancer, neoplasia	Increases adhesion [71] and invasion [72] of breast cancer cell lines. Decreases cell death and apoptosis of tumor cell lines [73]. Increases cell death of normal cell lines [74] Increase migration [75], movement [76] and binding [77] of tumor cell lines.

BDNF	brain-derived neurotrophic factor	anti-apoptosis; nerve development; negative regulation of neuron apoptosis; positive regulation of neuron differentiation; regulation of retinal cell programmed cell death; regulation of synaptic plasticity; inner ear development	Complex	Alzheimer's disease, hypertrophy, amyotrophic lateral sclerosis, depressive disorder, hemorrhage, tremor, type A insulin resistance, somnipathy, motor dysfunction, digestive system disorder	Decreases cell death [78, 79] of tumor cell lines. Increase developmental process of tumor cell lines [80]. Increase apoptose of normal cell [81].
CDCP1	CUB domain containing protein 1		Whole gene expression change: overexpressed in 4T1		Decreases cell death of tumor cell lines [82]
ITGB1	integrin, beta 1 (fibronectin receptor, beta polypeptide, antigen CD29 includes MDF2, MSK12)	G1/S transition of mitotic cell cycle; cellular defense response; cell adhesion; positive regulation of cell proliferation; germ cell migration; cell-cell adhesion mediated by integrin;	Intron inclusion in 4T1 between exons 8 and 9	fibrosis, tumorigenesis, dedifferentiation, delayed hypersensitive reaction, chondrodysplasia, neuropathy, heart failure, experimentally-induced adenomyosis, neoplasia	Decreases cell death [83, 84] of tumor cell lines. Increases cell death [85, 86] of normal cell lines Increase migration [87], cell adhesion [87] and cell binding [81] of tumor cells lines.
NUAK1	NUAK family, SNF1-like kinase, 1	protein amino acid phosphorylation	Whole gene expression change: overexpressed in 4T1	colorectal cancer, tumorigenesis	Decreases cell death of tumor cell lines [88]. Increase invasion of lymphoma cell [89].

TNFRSF11B	tumor necrosis factor receptor superfamily, member 11b	skeletal system development; apoptosis; signal transduction; negative regulation of odontogenesis of dentine-containing tooth	Whole gene expression change: overexpressed in 168FARN and 4T07	splenomegaly, aortic dissection, colorectal cancer, chronic obstructive pulmonary disease	Decreases cell death of tumor cell lines [90]. Genetic disorder [91].
VDR	vitamin D (1,25-dihydroxyvitamin D3) receptor	skeletal system development; regulation of transcription, DNA-dependent; signal transduction; multicellular organismal development; organ morphogenesis;	Whole gene expression change: overexpressed in 4T1.	Crohn's disease, end stage renal disease, non-small cell lung cancer, ovarian cancer, lung cancer, leukemia, lung neoplasm, prostate cancer	Decreases cell death of tumor cell lines [92] Increase developmental process of tumor cell lines [93].
ANGPT2	angiopoietin 2	angiogenesis; signal transduction; multicellular organismal development; cell differentiation	Alternative termination and more ASE	fibrosis, head and neck cancer, thyroid cancer, brain cancer, lymphoid cancer, angioimmunoblastic t-cell lymphoma, gastric cancer, intestinal cancer,	Increases cell death of normal cell lines [94].

CDH1	cadherin 1, type 1, E-cadherin (epithelial)	cell adhesion; homophilic cell adhesion; protein metabolic process; positive regulation of transcription factor import into nucleus; regulation of caspase activity; protein homooligomerization	Whole gene expression change: overexpressed in 4T1	gastric cancer, skin cancer, skin squamous cell carcinoma, serous ovarian carcinoma, prostate cancer, renal cancer,colon cancer, colon carcinoma, cervical cancer, cervical carcinoma, ovarian cancer	Increases cell death of normal cell lines [95]. Increase growth process of tumor cell lines [96]. Increase invasion of tumor cell lines [97].
HPRT1	hypoxanthine phosphoribosyltransferase 1	purine nucleotide biosynthetic process; nucleoside metabolic process; protein homotetramerization	Cassette exon: exon 1 is overexpressed in 168FARN and 4T07	rheumatoid arthritis, acute lymphoblastic leukemia, Crohn's disease, schizophrenia, neurological disorder	Increases cell death of normal cell lines [98].
LGALS7	l ectin, galactoside-binding, soluble, 7		Whole gene expression change: Overexpressed in 168FARN and 4T07	lymphoid cancer, B-cell non-hodgkin lymphoma, neoplasia, cancer	Increases cell death of normal cell lines [99].
MAPK14	mitogen-activated protein kinase 14	protein amino acid phosphorylation; cell motion; chemotaxis; response to stress; cell surface receptor linked signal transduction; protein kinase cascade;	Intron inclusion in 168FARN between exon 1 and 2	head and neck cancer, brain cancer, glioblastoma, breast cancer, colon cancer, colorectal cancer	Increases cell death [100, 101] of normal cell lines. Increase developmental process of tumor cell lines [102].
SLK	STE20-like kinase (yeast)	nucleotide-excision repair; protein amino acid phosphorylation; apoptosis	Cassette exon: exon 13 is highly expressed in 4T1		Increases cell death of normal cell lines [103].

LAMC2	laminin, gamma 2	cell adhesion; epidermis development	Whole gene expression change: overexpressed in 4T1	epidermolysis bullosa letalis, lymphoid cancer, angioimmunoblastic t-cell lymphoma, neoplasia	Increase adhesion of tumor cells line [104].
ITGB4	integrin, beta 4	cell communication; cell adhesion; cell-matrix adhesion; integrin-mediated signaling pathway	Whole gene expression change: overexpressed in 4T1	prostatic intraepithelial neoplasm, pancreatic cancer, pancreatic adenocarcinoma	Increase invasion of breast cancer tumor cells [105]. Increase growth of tumor cell lines [106].
HOXA10	homeobox A10	transcription; regulation of transcription, DNA-dependent; multicellular organismal development; spermatogenesis		bilateral cryptorchidism, unilateral cryptorchidism, colorectal cancer	Increase developmental process of tumor cell lines [107].

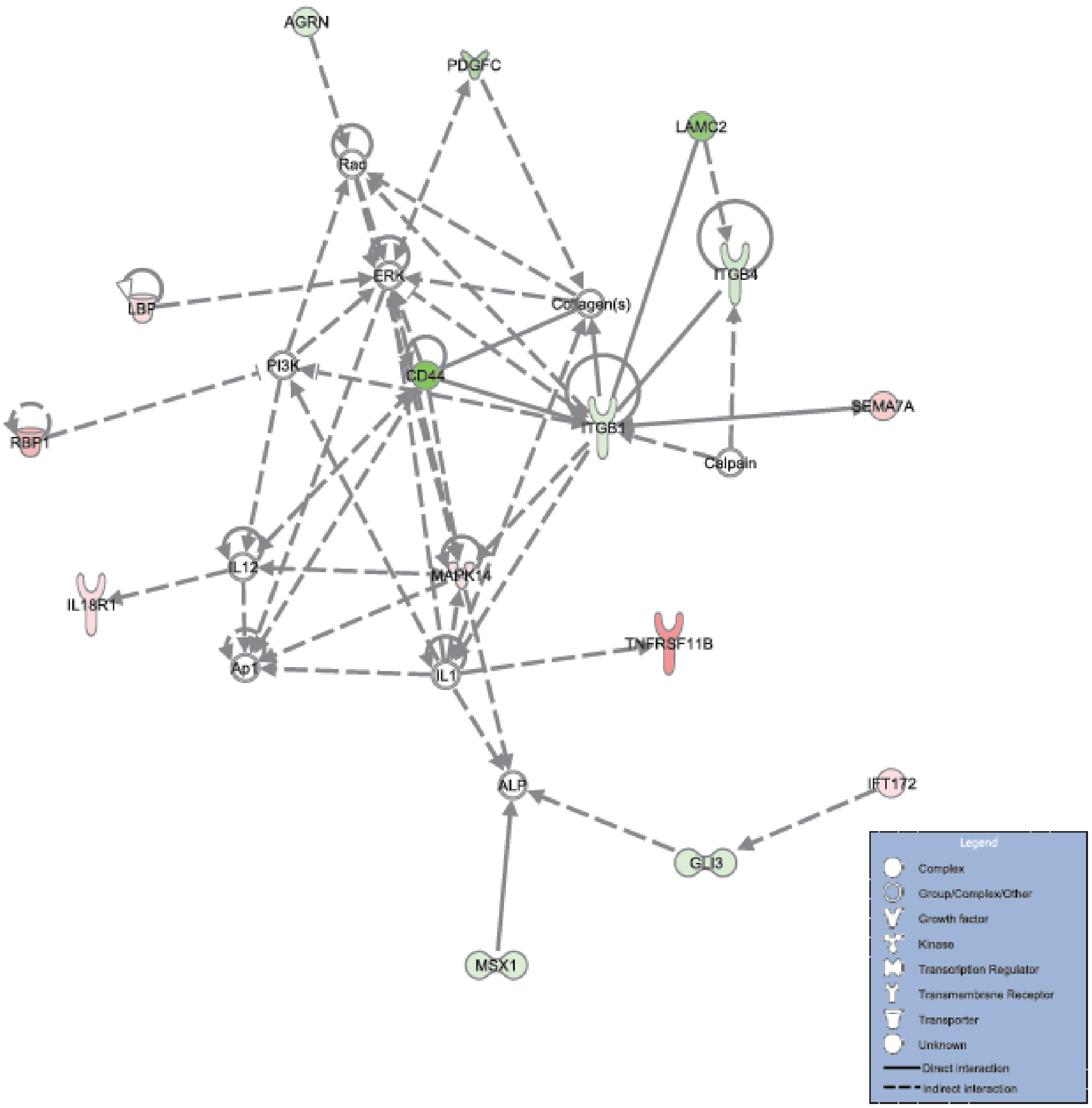

Figure 2.9: Network of molecular interactions containing products of statistically significant genes in breast cancer

Over-expressed genes in 4T1 and under-expressed genes in 4T1 are respectively indicated by green and red colors. The rate of over-expression or under-expression is proportional to the color intensity. The top functions or diseases where the proteins are involved are cancer, tissue development, cell-to-cell signaling and interaction.

Many significant genes including collagen subunit genes interact with ERK (Figure 2.10) and almost all of them are upregulated in 4T1. ERK is a complex consisting of MAP kinase proteins. It plays a role in cell division, growth and proliferation. ERK phosphorylates many cytoplasmic and nuclear substrates required for the transcription of several genes to pass from G1 stage to S stage in the cellular division process [108]. In breast cancer, the inhibition of ERK enhances the anti-estrogenic treatment [109].

Another interesting candidate is CDH1 (cadherin 1, type 1, E-cadherin), a tumor suppressor gene [110] from the cadherin superfamily, that encodes an epithelial cell-cell adhesion protein. It implements calcium-dependent homophilic interactions at sites of cell-cell contacts. In 4T1, it is highly expressed compared to 168FARN or 4T07 tumors. We observed a whole gene expression change with the exception of the 5’ UTR and the first exon (Figure 2.4). Mutations in this gene are related to gastric, thyroid, colorectal, and ovarian cancers (Table 2.3). The loss of its function is thought to contribute to cancer progression by increasing proliferation, invasion, and metastasis [96, 97]. CDH1 acts on the NfKb, F-Actin and Mapk complexes (Figure 2.11), all of them implicated in cell interactions, cell development or cell movement.

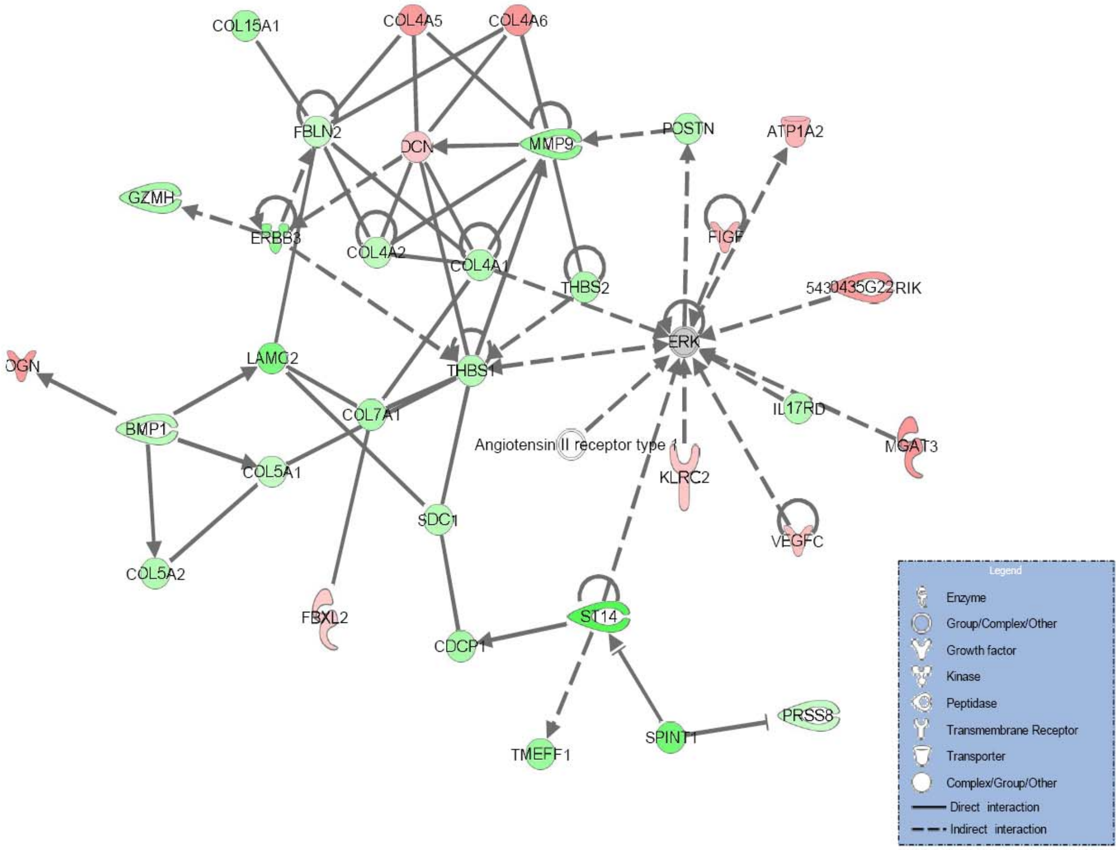

Figure 2.10: Network of molecular interactions containing products of statistically significant genes in breast cancer

Over-expressed genes in 4T1 and under-expressed genes in 4T1 are respectively indicated by green and red colors. The rate of over-expression or under-expression is proportional to the color intensity. The top functions or deseases where the proteins are involved are cancer, cell cycle and cell death.

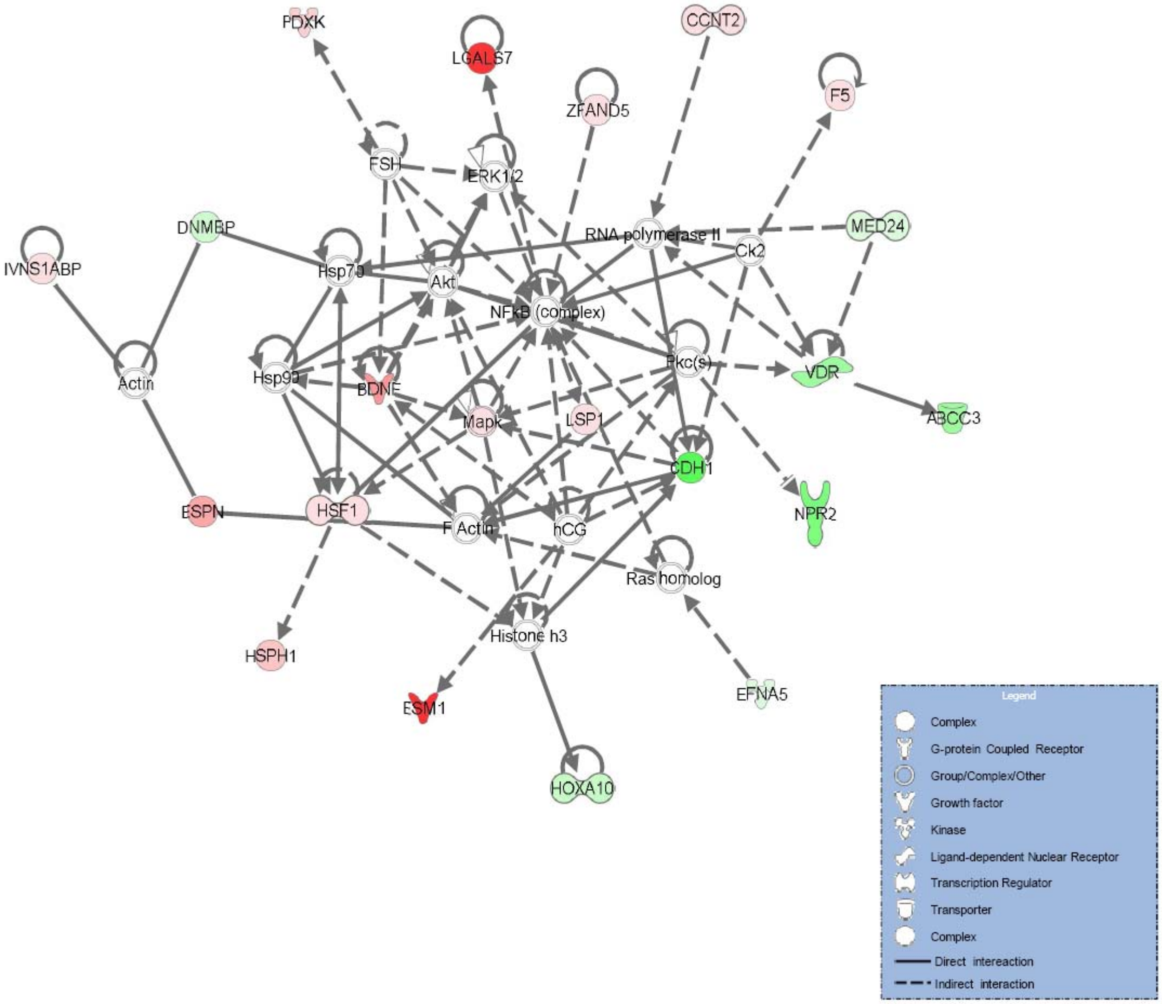

Figure 2.11: Network of molecular interactions containing products of statistically significant genes in breast cancer

Over-expressed genes in 4T1 and under-expressed genes in 4T1 are respectively indicated by green and red colors. The rate of over-expression or under-expression is proportional to the color intensity. The top functions or diseases where the proteins are involved are cancer, cell morphology, cell-to-cell signaling and interaction.

Discussion

In this work, we used a splicing-sensitive microarray technology to investigate gene isoform differences across breast tumors with varying levels of metastatic potential. Using several stringent statistical selection criteria and filtering steps on probesets (exons) and metaprobesets (transcripts), we obtained a confident set of 2623 candidate genes that undergo isoform or whole expression variations associated with metastatic potential. A large number of the detected differences are represented by non-core probesets; that is, those that are supported by EST and predictive evidence not present within RefSeq and full-length mRNA GenBank records. Besides the expression variations of known coding regions, cancer cells are susceptible to express such predominantly non-coding regions because of general misregulation of gene expression. Therefore, the inclusion of non-core probesets in our analysis was appropriate. This enabled us to enrich the novel ASE proportion especially the intron retention/cryptic category. 277346 of the 493710 (56.2%) probesets analysed are non-core. 49588 of them satisfied the expression filtering criteria including 2037 (4.1%) showing expression variation across tumor samples. This proportion represents 19% of the total statistically significant probesets obtained (10744). However, probesets outside of annotated transcript clusters (known genes) are excluded from our Exon Array analysis. We note that some of these excluded probesets (518024), representing about half of total probesets on the array, may be differentially expressed or alternatively spliced. They may form new genes or produce new isoform variants by elongating the ends of known transcripts, however this was not investigated in our analysis.

Because of the inherent difficulty of unambiguously interpreting statistically significant expression changes in Exon Array data [111], only 7.7% of probeset variations were classified into known isoform change categories. The remainder of the changes (92.3%) are difficult to interpret and may reflect the complexity of gene expression variation in cancer. For example, a single isoform may arise from multiple ASEs which make its expression pattern less obvious to detect. Expression differences occurring in cancer cells genes are not always crystalline and explained by standard known changes. Although the Exon Array is a powerful tool, some ASEs may be missed or misinterpreted by Exon Array. For example a fundamental limitation of Exon Array and any other approach to measure the gene expressions is splicings that introduce aberrations in the transcript such as premature stop codons. If these transcripts are degraded at a high rate by the RNA surveillance mechanisms (nonsense-mediated decay), their expression levels may be detectable only partially or not at all since their concentration in the cell is unstable. Hence, the analysis presented here uncovers only a part of the transcriptional and post-transcriptional aberrations that distinguish metastatic tumors. With the implementation of mRNA sequencing technologies in the near future, such analysis will become much more complete.

Table 2.4: For the three studies compared (our study against two prior studies), number of common genes differently expressed between the least metastatic sample and the most metastatic sample

168FARN against 4T1 for the Exon Array and Yang J studies, and 67NR against 4T1 for Lou Y study. The study name is mentioned at the first column. The second column contains the total number of common genes retrieved by ANOVA p-value thresholds of 8.46×10^{-3} (for the Exon Array) , 1.16×10^{-2} (for Lou Y) and 9.16×10^{-3} (for Yang J). The two last columns indicate respectively the number of significant common hits with fold-changes going in same directions and the number of significant common hits with fold-changes going in an opposite directions.

Lab	Total hits with pv<pv-cutoff	# of hits with FC going in same direction	# of hits with FC going in different direction
Lou Y et al.	584	419	165
Yang J et al.	270	228	42
Lou Y et al. and Yang J et al.	58	51	7

We also compared our data with two prior studies (Yang et al [112] and Lou et al [24]) that used tumors derived from the same cell lines, but assayed on conventional microarray platforms, to profile gene expression in breast cancer. We looked for the common statistically significant genes which are differentially expressed between the least metastatic sample and the most metastatic sample: 168FARN against 4T1 for our Exon Array analysis and Yang et al studies, and 67NR against 4T1 for Lou et al (Table 2.4). For each prior study we performed an ANOVA-test on gene expression intensities followed by a 0.05-level FDR (False Discovery Rate) correction to determine the cutoff p-value for identifying significant differentially expressed genes. A comparison of the three studies revealed 58 hits common to the three studies. Of these, 51 showed consistent expression patterns across the three groups (Table 2.5) while the remaining seven had discordant expression behaviour

(Table 2.6). Several of the consistent genes have been previously described to be related to cancer. An example is the MAPK6 gene that encodes a kinase protein required in cell growth, proliferation, migration and death activities. Up-regulated in breast cancer, it causes the over-expression of MCF7 cells proliferation [113]. Another example is TGFA, a protein also involved in cell growth, apoptosis, differentiation, and migration, that decreases apoptosis of PE01 cells and LNCaP cells in ovarian cancer [114] and prostate cancer [115], respectively. The large number of common candidate genes showing concordant behaviour across three independent studies and three different microarray platforms suggests that they may have important roles in metastasis and are independent of experimental conditions and measurement techniques. The seven discordant hits insinuate that there exist some differences between our gene expression results and the two other studies we found in the literature. We hypothesise that one of the confounding factors that affect tumor behaviours across labs may be the age, volume, the growth rate of the primary tumors at the time of removal, experiment conditions, the genetic instability of the cell lines, the injection site. These factors alone or combined could influence the expression profile of gene samples.

Table 2.5: Common consistent differently expressed genes between the least metastatic and the most metastatic samples of each study

Common genes whose fold-changes going in same directions across the three studies (Exon Array, Lou Y et al and Yang J et al) are given. For each gene, the gene id, the ANOVA p-value and the fold-change is mentioned in each study.

	Exon Array (168FARN vs. 4T1)			Lou Y et al (67NR vs. 4T1)			Yang J et al (168FARN vs. 4T1)		
Gene Id	Id	P-value	Fold-change	Id	P-value	Fold-change	Id	P-value	Fold-change
Pltp	6892964	2.31×10^{-07}	1.66	1417963_at	2.95×10^{-03}	1.25	100927_at	1.54×10^{-08}	3.44
Tob1	6783642	3.18×10^{-07}	1.43	1423176_at	5.12×10^{-05}	2.47	99532_at	1.44×10^{-07}	3.56
Cxcr7	6751469	1.25×10^{-06}	1.79	1417625_s_at	4.22×10^{-03}	0.75	93430_at	3.18×10^{-07}	0.53
Aqp1	6946406	2.71×10^{-06}	1.01	1416203_at	2.39×10^{-04}	2.20	93330_at	2.19×10^{-07}	3.74
Abcb1a	6928740	3.55×10^{-06}	1.70	1419758_at	1.99×10^{-05}	0.62	102910_at	7.43×10^{-05}	3.49
Abcb1a	6928740	3.55×10^{-06}	1.70	1419759_at	7.80×10^{-04}	0.74	102910_at	7.43×10^{-05}	3.49
Epb4.9	6825713	8.08×10^{-06}	-1.34	1460223_a_at	2.29×10^{-05}	-1.92	103600_at	1.04×10^{-05}	-4.03
Pgcp	6829612	8.24×10^{-06}	1.82	1416441_at	1.83×10^{-05}	2.25	93039_at	3.67×10^{-03}	4.14
Hspa4l	6896997	1.46×10^{-05}	1.38	1418253_a_at	5.38×10^{-04}	1.34	99489_at	7.28×10^{-03}	0.50
Hspa4l	6896997	1.46×10^{-05}	1.38	1449010_at	1.58×10^{-03}	1.36	99489_at	7.28×10^{-03}	0.50
Ak1	6876211	1.84×10^{-05}	1.06	1422184_a_at	6.69×10^{-04}	1.19	96801_at	6.03×10^{-06}	5.10
Junb	6983894	2.88×10^{-05}	-1.38	1415899_at	1.21×10^{-05}	-2.41	102362_i_at	5.78×10^{-07}	-1.63
Shroom3	6932510	3.61×10^{-05}	-1.69	1422629_s_at	1.34×10^{-03}	-1.16	100024_at	9.79×10^{-04}	-3.26
Naglu	6784237	3.93×10^{-05}	0.66	1417706_at	5.66×10^{-05}	0.72	93373_at	3.32×10^{-05}	2.09
Sfn	6925916	4.99×10^{-05}	-1.67	1448612_at	2.04×10^{-05}	-3.16	96704_at	3.16×10^{-03}	-3.37
Xdh	6857183	6.16×10^{-05}	0.59	1451006_at	4.03×10^{-03}	0.60	97950_at	2.11×10^{-06}	5.82
Sod3	6930799	8.65×10^{-05}	1.58	1417634_at	1.02×10^{-03}	1.57	94902_at	1.86×10^{-05}	0.03
Tgfa	6947596	8.68×10^{-05}	-1.72	1421942_s_at	1.08×10^{-02}	-0.73	92369_at	3.82×10^{-06}	-3.29
Tgfa	6947596	8.68×10^{-05}	-1.72	1421943_at	3.93×10^{-07}	-4.73	92369_at	3.82×10^{-06}	-3.29
Spint1	6880508	9.34×10^{-05}	-2.62	1416627_at	1.80×10^{-06}	-4.22	97206_at	1.17×10^{-05}	-5.73

Klf5	6821304	9.66×10^{-05}	-1.08	1451021_a_at	4.69×10^{-05}	-2.78	97937_at	2.54×10^{-04}	-4.18
Ncoa3	6883210	1.06×10^{-04}	0.47	1422737_at	7.62×10^{-05}	0.29	102024_at	7.56×10^{-03}	0.43
Col7a1	6992378	1.08×10^{-04}	-1.54	1419613_at	1.40×10^{-04}	-3.35	93383_at	2.29×10^{-03}	-1.29
Mapk6	6996935	1.45×10^{-04}	0.41	1419169_at	8.59×10^{-03}	0.10	103416_at	2.77×10^{-03}	0.06
Usp10	6979519	1.99×10^{-04}	-0.57	1448230_at	6.36×10^{-05}	-1.44	99639_at	5.53×10^{-03}	-1.24
Dtx2	6934962	2.10×10^{-04}	0.60	1421720_a_at	8.47×10^{-03}	0.70	96818_at	2.32×10^{-04}	0.95
Pdgfra	6931740	2.72×10^{-04}	1.02	1421916_at	1.28×10^{-03}	1.58	95079_at	3.39×10^{-05}	2.51
Pdgfra	6931740	2.72×10^{-04}	1.02	1421917_at	6.58×10^{-04}	1.65	95079_at	3.39×10^{-05}	2.51
Mesdc2	6962179	3.54×10^{-04}	0.48	1416181_at	2.04×10^{-04}	0.11	95405_at	1.75×10^{-05}	1.59
Prdx2	6977778	4.25×10^{-04}	-0.53	1418506_a_at	3.21×10^{-04}	-0.72	99608_at	5.25×10^{-04}	-1.02
Fzd6	6829952	5.44×10^{-04}	-0.88	1417301_at	3.93×10^{-04}	-3.36	101142_at	6.25×10^{-03}	-2.38
Fzd6	6829952	5.44×10^{-04}	-0.88	1448662_at	9.34×10^{-04}	-1.62	101142_at	6.25×10^{-03}	-2.38
Chek1	6994666	1.00×10^{-03}	-1.03	1450677_at	2.69×10^{-03}	-1.40	103064_at	7.50×10^{-05}	-1.32
Cd24a	6767537	1.09×10^{-03}	-0.70	1416034_at	1.96×10^{-06}	-5.62	100600_at	3.52×10^{-09}	-4.84
Cd24a	6767537	1.09×10^{-03}	-0.70	1448182_a_at	1.99×10^{-06}	-5.95	100600_at	3.52×10^{-09}	-4.84
Slc11a2	6838557	1.19×10^{-03}	0.61	1417584_at	1.31×10^{-03}	0.53	104451_at	3.59×10^{-06}	1.41
1110003E01Rik	6938710	1.21×10^{-03}	0.54	1416767_a_at	2.88×10^{-03}	0.76	96716_at	6.80×10^{-03}	0.62
1110003E01Rik	6938710	1.21×10^{-03}	0.54	1416768_at	4.40×10^{-03}	0.48	96716_at	6.80×10^{-03}	0.62
Timp2	6792649	1.39×10^{-03}	0.17	1420924_at	8.16×10^{-03}	1.03	93507_at	3.57×10^{-04}	1.87
Timp2	6792649	1.39×10^{-03}	0.17	1450040_at	2.05×10^{-05}	1.17	93507_at	3.57×10^{-04}	1.87
Timp2	6792649	1.39×10^{-03}	0.17	1460287_at	8.50×10^{-05}	1.30	93507_at	3.57×10^{-04}	1.87
Cxcl5	6932364	1.88×10^{-03}	2.52	1419728_at	1.49×10^{-03}	1.03	98772_at	4.01×10^{-04}	1.78
Ltc4s	6788017	1.98×10^{-03}	-0.79	1419692_a_at	4.10×10^{-03}	-0.11	92401_at	1.19×10^{-03}	-0.63
Spsb2	6949811	3.59×10^{-03}	0.45	1422106_a_at	1.01×10^{-02}	0.27	103993_at	7.68×10^{-04}	1.67
Plscr2	6991362	5.25×10^{-03}	1.49	1448961_at	1.71×10^{-03}	0.91	102053_at	3.93×10^{-03}	2.86
Fgl2	6929119	5.99×10^{-03}	0.83	1421854_at	2.26×10^{-03}	0.70	97949_at	8.18×10^{-04}	0.24
Fgl2	6929119	5.99×10^{-03}	0.83	1421855_at	2.02×10^{-03}	0.50	97949_at	8.18×10^{-04}	0.24
Syt8	6965262	6.38×10^{-03}	-0.69	1450800_at	1.61×10^{-04}	-2.55	92682_at	2.34×10^{-03}	-4.72

Smarce1	6791316	7.39×10^{-03}	-0.25	1422676_at	1.73×10^{-03}	-0.47	96651_at	6.85×10^{-03}	-0.31
Slc30a4	6890453	7.58×10^{-03}	0.65	1418843_at	2.99×10^{-03}	1.02	95571_at	6.55×10^{-04}	2.01
Aco2	6832142	8.34×10^{-03}	0.18	1451002_at	3.22×10^{-04}	0.52	96870_at	3.01×10^{-04}	0.89

Table 2.6: Common inconsistent differently expressed genes between the least metastatic and the most metastatic samples of each study

Common genes whose fold-changes going in opposite directions across the three studies (Exon Array, Lou Y et al and Yang J et al) are given. For each gene, the gene id, the ANOVA p-value and the fold-change is mentioned in each study.

	Exon Array (168FARN vs. 4T1)			Lou Y et al (67NR vs. 4T1)			Yang J et al (168FARN vs. 4T1)		
Gene Id	Id	P-value	Fold-change	Id	P-value	Fold-change	Id	P-value	Fold-change
Mrc2	6784564	2.02×10^{-05}	-1.22	1421044_at	9.10×10^{-04}	1.66	100759_at	1.42×10^{-05}	3.11
Mrc2	6784564	2.02×10^{-05}	-1.22	1421045_at	2.04×10^{-03}	2.62	100759_at	1.42×10^{-05}	3.11
Atp6v0a1	6784236	1.65×10^{-04}	-0.41	1417632_at	1.52×10^{-03}	0.30	103275_at	2.48×10^{-04}	0.34
Sparc	6788410	2.51×10^{-03}	-0.29	1416589_at	7.33×10^{-03}	0.40	97160_at	2.11×10^{-03}	0.87
Sparc	6788410	2.51×10^{-03}	-0.29	1448392_at	7.87×10^{-03}	0.54	97160_at	2.11×10^{-03}	0.87
Decr2	6854462	5.17×10^{-03}	0.42	1423495_at	2.38×10^{-04}	-1.26	102677_at	4.59×10^{-03}	-2.21
Hk1	6774391	6.24×10^{-03}	-0.56	1420901_a_at	3.95×10^{-05}	0.64	99335_at	1.96×10^{-04}	0.18

To understand the functional significance of alternatively spliced genes, we carried out a global gene pathway analysis. We found that nearly half of the significant genes have been reported to be involved in the major pathways related to cancer, which include cell cellular growth, cellular proliferation, cellular migration, cell interactions and cellular death. The majority of them have antecedents in various cancers and genetic disorders. Genes with significant expression and isoform differences that have not previously been reported to be associated with any disease may represent novel cancer genes or genes specific to metastatic breast cancer. Meanwhile, the other half of candidates in our analysis, currently have unknown or poorly annotated functions. A large number of differentially expressed genes encode proteins, form complexes or interact with proteins/complexes involved in biological processes whose incorrect regulations are implicated in cancer processes and genetic disorder (Figures 2.9 – 2.11, Table 2.3).

One example of these types of complexes is represented by collagen genes. Collagen plays important roles in many pathological states, including tumor progression and metastasis. In our analysis, collagen subunit genes (COL15A1, COLA41, COLA42, COLA45, COLA46 COL5A1, COL5A2 and COL7A1) are over-expressed in 4T1. Collagen is an extracellular matrix protein required in the regulation of tumor growth, invasiveness, and angiogenesis. Up-regulation of collagen has been associated with the promotion of invasion and metastasis in different types of human cancers [116-119]. Collagen type-IV enhances tumor cell invasion and migration [120] while collagen type-V regulates cell proliferation and migration [121]. High collagen IV levels have been associated with metastatic lung, colorectal, breast and gastric cancers [122, 123]. CD44, a gene that we and other study groups have found to be alternatively spliced in metastatic cancers, is known to be involved in cancer hallmarks; it binds to collagen proteins (type I) [124], however little is known about the regulation of this interaction.

Another example is the interaction of NF-kB with many genes (LGALS7, ZFAND5, LSP1, CDH1, MAPK14 and HSF1) that undergo isoform variation or whole gene expression change (Figure 2.9). NF-kB, or eukaryotic nuclear factor kappa-light-chain-enhancer of activated B cells, is involved in autoimmune response, inflammation, cell proliferation and cell death by controlling the expression of genes implicated in these processes [125]. Misregulation of NF-κB has been associated to cancer, autoimmune diseases and inflammatory responses [126, 127]. MAPK14, showing an intron inclusion in 4T07, is a member of the MAPK complex in which controlled regulation plays a part in cell proliferation and differentiation, whereas an uncontrolled activation can lead to oncogenesis [128].

Our results suggest that in breast cancer numerous genes present expression variations or splicing defects whose protein products could disturb normal biological functions. Since most of our candidate genes are previously implicated in cancer, we have the confidence that these splicing events are real biological events, and are relevant to breast cancer. Moreover, the finding of novel differentially spliced or expressed genes could extend the list of breast cancer genes and be candidates for innovative treatments and diagnostics. Compared to other approaches based on DNA microarrays that interrogate single whole genes, studying genome wide analysis of AS in breast cancer at the exon level adds more knowledge about the type of variations occurring in genes; this can easily lead to improved and more specific diagnostics or treatment methods. For example HMGA1, one of our statistically significant genes, has been reported as a breast cancer marker by

Venables et al [129]. HMGA1 is a small DNA binding protein that is involved in metastatic progression of cancer cells [130]. In Venables's study, HMGA1 lacks the exon 7 in normal breast cells compared to breast cancer cells. This cassette exon takes out an AT-hook DNA binding domain. In our study, HMGA1's exon 7, that is statistically significant, is expressed in all tumor samples but with different rates. It is lightly overregulated in 168FARN and 4T07 comparatively to 4T1 sample.

Conclusion and perspective

The connection between cancer and AS is becoming increasingly compelling, based both on prior data, and the results of our work presented here. We identified 2623 genes that are differently spliced and/or express different isoforms between tumor types. Most of the metastasis-specific spliced genes are involved in key pathogenic processes in cancer. The detection of ASEs, especially novel ASEs, are of a great importance for breast-cancer studies. By establishing which genes actively participate in different cancer stages, tumor-specific alternatively spliced mRNAs and proteins can be used as breast cancer biomarkers [129]. For example, the gene SRRT intron inclusion in 168FARN and 4T07 (Figure 2.5) could be a potential breast cancer marker, and could help early or progression diagnosis. ASEs identified but not yet implicated in any disease can be breast cancer-specific and potentially serve as drug target studies [131, 132]. For example a therapeutic small interfering RNA (siRNA) can be used to neutralize a specific gene isoform that disrupts a normal biological function [133, 134]: a siRNA can be designed to specifically and potently target and silence a gene isoform (with little or no effect on the others genes) and introduced artificially in cells to stop the production of the corresponding proteins. Further protein domain characterizations of sequences alternatively spliced have the potential to improve the understanding of the complicated biological processes connecting isoform variations and cancer.

Materials and methods

Cell Culture and Transfections

The 4T1 murine mammary carcinoma cell line was obtained from the American Type Culture Collection. Non-metastatic 67NR, 168FRNA, 4T07 and lung-metastatic 66cl4 murine mammary carcinoma cell lines were kindly provided by Dr. Fred Miller (Barbara Ann Karmanos Cancer Institute, Detroit, MI). All cell lines were grown in DMEM supplemented with 10% fetal bovine serum, 10 mmol/L HEPES, 1 mmol/L sodium pyruvate, 1.5 g/L sodium bicarbonate, penicillin/streptomycin, and fungizone.

Mammary gland injection and spontaneous Metastasis Assay

Female BALB/c mice (4-6 weeks) were purchased from Charles River Laboratories. The mice were housed in facilities managed by the McGill University Animal Resources Centre, and all animal experiments were conducted under a McGill University–approved Animal Use Protocol in accordance with guidelines established by the Canadian Council on Animal Care. In the spontaneous metastasis studies, 4T1 mammary carcinoma cells were harvested from subconfluent plates, washed once with PBS, and re-suspended (10^5 cells) in 50 µL of a 50:50 solution of Matrigel (BD Biosciences) and PBS. This cell suspension was injected into the right abdominal mammary fat pad of BALB/c mice and measurements were taken beginning on day 7 postinjection for the time periods indicated. For each cell lines 67NR, 4T07 and 66Cl4, five tumors were grown individually in five mice; for each cell lines 168FARN and 4T1 four tumors were grown individually in four mice. Tumor volumes were calculated using the following formula: $\pi LW^2/6$, where L is the length and W is the width of the tumor. Tumors were surgically removed, using a cautery unit, once they reached a volume between 100 and 125 mm3.

RNA extraction and microarray hybridization

Total RNA was purified using RNeasy Mini Kit Columns following the manufacturer's instructions. We assessed The RNA quality using RNA 6000 NanoChips with the Agilent 2100 Bioanalyzer (Agilent, Palo Alto, USA). Tumors were hybridized independently at the functional genomics facility of McGill University and Genome Quebec Innovation Centre (Montreal, Quebec, Canada). Biotin-labelled target for the microarray experiment were prepared using 1µg of total RNA. We subjected the RNA to a ribosomal RNA removal process with the RiboMinus Human/Mouse Transcriptome Isolation Kit (Invitrogen). cDNA was synthesized using the GeneChip® WT (Whole Transcript) Sense Target Labelling and Control Reagents kit as described by the manufacturer (Affymetrix). Then, the sense cDNA was fragmented by UDG (uracil DNA glycosylase) and APE 1 (apurinic/apyrimidic endonuclease 1) and biotin-labelled with TdT (terminal deoxynucleotidyl transferase) using the GeneChip® WT Terminal labelling kit (Affymetrix, Santa Clara, USA). Hybridization was performed using 5 micrograms of biotinylated target, which was incubated with the GeneChip® Mouse Exon 1.0 ST array (Affymetrix) at 45°C for 16-20 hours. Subsequently to hybridization, non-specifically bound material was removed by washing and detection of specifically bound target was performed using the GeneChip® Hybridization, Wash and Stain kit, and the GeneChip® Fluidics Station 450 (Affymetrix). The arrays were scanned using the GeneChip® Scanner 3000 7G (Affymetrix) and raw data was extracted from the scanned images and analyzed with the Affymetrix Power Tools software package (Affymetrix).

Data pre-processing and analysis

Signal estimation

Signal estimates were derived from the CEL files of the 23 arrays. The Affymetrix Power Tools software package (Affymetrix) was used to quantile normalize the probe fluorescence intensities and to summarize the probe set (representing exon expression) and meta-probe set (representing gene expression) intensities using a probe logarithmic intensity error model (PLIER[57]) for probe set and ITER-PLIER for meta-probe set. Presence or absence of probe set expression was determined by the Detection Above background (DABG) statistics. For the probeset-level analysis we used the full set of probesets from the Exon Array including core and non-core probesets.

Filtering signal data

The filtering steps and parameters described in this paragraph come from the Affymetrix technical note for the identification of ASE [6]. Two outlier biological replicates (from 4T07 and 66C14 samples), which don't cluster with the replicates within the tumor type they belong to, were identified and removed following the Principal Component

Analysis (PCA). In order to be considered as expressed and included in the analysis each exon had to satisfy the following four criteria (Figure 2.12): (1) the exon is called as Present in at least 50% of the samples of at least one tumor type. An exon is called as “present” if its probeset detection above background (DABG) p-value is less than 0.05; (2) the probeset must have a low cross-hybridization potential (equal to 1) to discard false positives. The signal intensities of probe sets having a high cross-hybridization potential may come from a different gene sequence; (3) the probeset must have a gene-level normalized intensity lower than 5 (very large gene-level normalized intensity may also implicate cross-hybridization to other gene sequences); (4) the probeset must have a gene-level normalized intensity greater than 0.20 (very low gene-level normalized intensity probesets were removed to discard features that may have non-linear signal response.) For each gene containing the previously filtered exons, two filtering criteria were used: (1) the gene had at least 50% of core exons called as “present” in at least 50% of samples in at least two groups; (2) the IterPLIER gene intensity is greater than a threshold of 30.

We performed two concurrent AS analyses: AS analysis with the probeset intensities and AS analysis with the gene-level normalized intensities. There is no optimal method to analyze isoform level data, and the relative merits of each approach are described in some detail by Bemmo et al. [111]. For each analysis, a one-way ANOVA-test was done on probeset scores to retrieve probesets that have a statistically significant change of expression or inclusion rates between groups. We selected probesets having an ANOVA p-value lower than the p-value cutoff (6.36×10^{-4} for probeset intensity analysis and 7.10×10^{-4} for SI analysis) established by the Benjamini-Hochberg FDR (False discovery rate) correction [135] at a 0.05 level. 168FARN and 4T07 were compared against 4T1 by pairwise Student’s t-tests on probesets scores. Logarithmic fold-changes were computed between groups

(168FARN/4T1 and 4T07/4T1). The genes expression intensities of metaprobesets were analysed by the same way as probesets. The statistical significance of a gene was determinated by a FDR p-value cutoff of 8.46 × 10^{-3} computed from Anova p-values. Since the SI analysis performs best when a gene has a large number of constitutive exons comparatively to alternative exons, we restricted the SI analysis to genes whose overall gene expression not change. Fold-changes and p-values of exons within each gene have been uploaded and visualized in the UCSC Genome Browser environment.

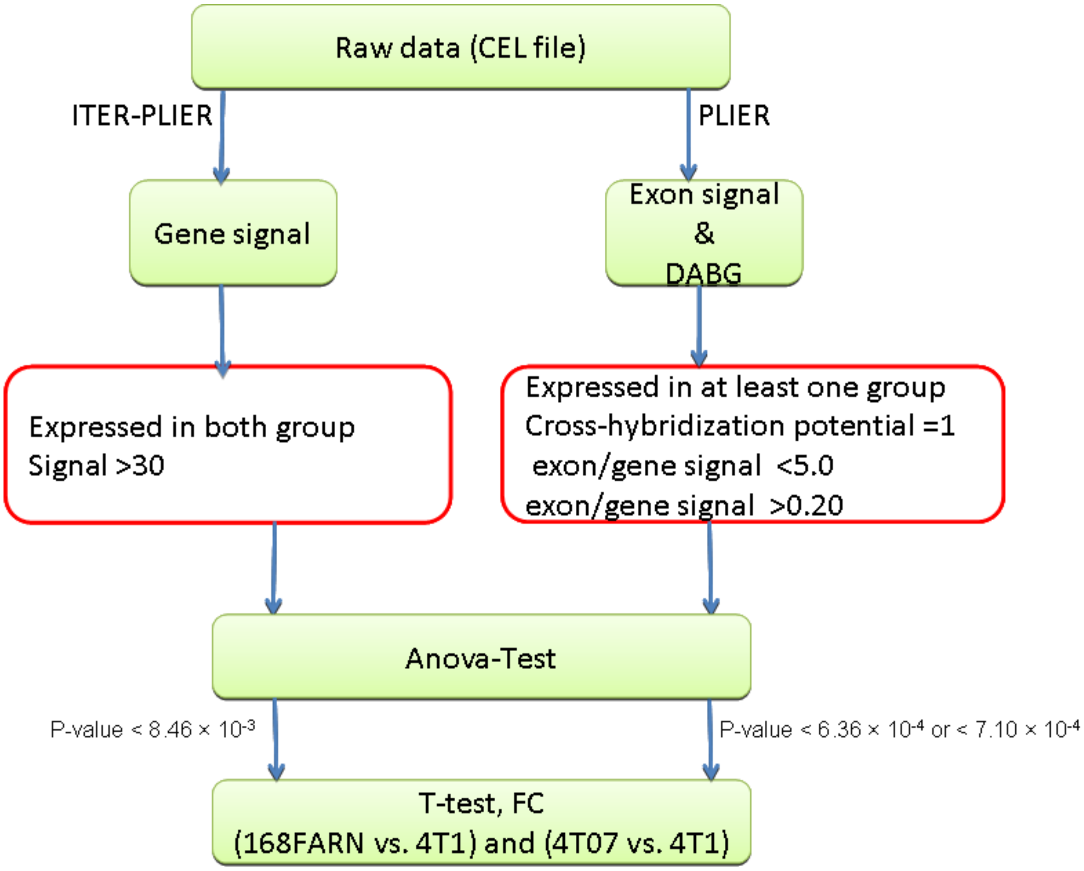

Figure 2.12: workflow of probeset and metaprobeset filtering steps

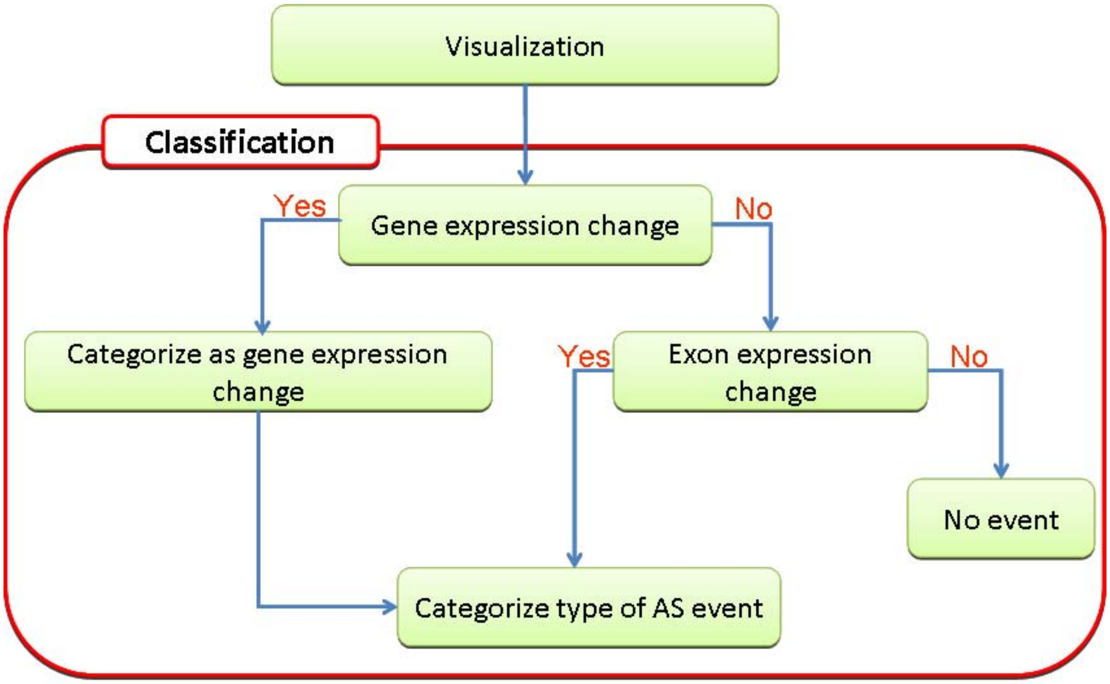

Figure 2.13: protocol flow of **gene viualisation and manual curation**

The exon T-test p-values and the exon fold-changes of pairwise comparisons are visualised in the context of gene belongings.

The visualization enabled us to classify ASEs (Figure 2.13). We examined the exon expression fold-changes within the gene: if the whole expression changes (ANOVA p-value <8.46 × 10^{-3}), we categorise it as gene expression change and determine the ASE. If the whole gene expression doesn't change (ANOVA p-value≥8.46 × 10^{-3}), we look at the exon-level: if the exon expression change (ANOVA p-value<6.36 × 10^{-4} for probeset intensity analysis or 7.10 × 10^{-4} for SI analysis), we categorize the ASE.

Authors' contributions

A.B performed the statistical and computational analysis, carried out the AS analysis, prepared the figures and wrote the manuscript. C.D extracted the RNA from tumors. A.A.N.R and C.R did the mammary gland injection and the spontaneous metastasis essay. P.S and J.M. conceived of the study and supervised the hybridization of the microarrays. J.M supervised the writing of the manuscript.

Acknowledgments

This work was supported by grants from CIHR and Genome Canada/Genome Quebec (to J.M.). J.M. is a Canada Research Chair recipient. We would like to thank the staff of the Genome Quebec Microarray Platform and the technical staff for expert help and microarray processing.

Appendix– outlier identification and verification of metastatic potential

We used the Affymetrix Mouse Exon 1.0 ST Array to measure the genome-wide expression of genes in five tumor samples: 67NR, 168FARN, 4T07, 66C14 and 4T1 using four or five biological replicates. The probeset and metaprobeset intensities were normalized and summarized with PLIER and ITER-PLIER. Several observations from the PCA analysis (Figure 2.1) concerning the behaviour of samples 67NR and 66C14 brought us to suspect that somewhere along the process of the analysis, the two samples may have been misidentified.

Table 2.7: Correlation of the gene expression fold-changes of our study with two prior studies

Three p-value thresholds (0.01, 0.001 and 0.0001) were used to derive common significant genes across the three studies. For each p-value block, the last row represents the comparaison between the three studies.

[1] Studies against which our Exon Array analysis is compared
[2] Number of genes commons to the three studies and having a p-value lower than the p-value cutoff (Pv-cutoff)
[3] Number of common genes having fold-changes going in the same direction (same sign)
[4] Number of common genes having fold-changes going in an opposite direction
A : No switch between samples 67NR and 66C14 was done (comparison between 67NR and 4T1)
B : A switch between samples 67NR and 66C14 was done (comparison between 66C14 and 4T1)

Lab[1]	Total hits with pv<pv-cutoff[2]		# of EA hits with FC in same direction[3]		# of EA hits with FC in different direction[4]	
	A	B	A	B	A	B
		Pv-cutoff = 0.01				
Lou Y et al.	61	448	46	409	15	39
Yang J et al.	20	93	15	67	5	26
Lou Y et al.and Yang J et al.	2	24	2	21	0	0
		Pv-cutoff = 0.001				
Lou Y et al.	9	192	8	184	1	8
Yang J et al.	4	32	4	25	0	7
Lou Y et al.and Yang J et al.	0	6	0	5	0	0
		Pv-cutoff = 0.0001				
Lou Y et al.	5	70	5	70	0	0
Yang J et al.	3	14	3	13	0	1
Lou Y et al.and Yang J et al.	0	3	0	3	0	0

We carried out a correlation analysis at the gene expression level between our results and the two prior studies [24, 112] discussed in the main section of this paper. For each study we performed a comparison between the most divergent samples 67NR and 4T1 with a T-test and logarithmic fold-change ratios. We established three p-value cut-offs (0.01, 0.001 and 0.0001) to retrieve genes common to the prior studies and ours. For each p-value level, we recovered the number of common genes with an expression pattern going in the same direction across studies and the number of genes with an expression pattern going in the opposite direction. Suspecting a possible sample mislabelling or switch, we preformed the same comparison with the sample labels of 67NR and 66C14 switched and obtained a significantly more correlated result. Although only some genes are consistently significant across the three studies, most support a permutation between samples 67NR and 66C14 (Table 2.7). Since the correlation results are not fully consistent across the three independent studies, we felt that this comparative analysis was not sufficient to conclusively determine sample misidentification. Moreover, we questioned whether the tumor growth rate (Figure 2.14) could explain the clustering of our tumor samples. In our analysis, we let the tumor grow to equal volume, and then harvested them at different times. The main observation is that the expression grouping of the tumors does not depend entirely on the tumor growth rate. We note that 168FARN and 4T07 grow very differently, but produce virtually identical expression patterns. However, 4T01 and 4T07 grow quite similarly, but their expression profiles do not group together.

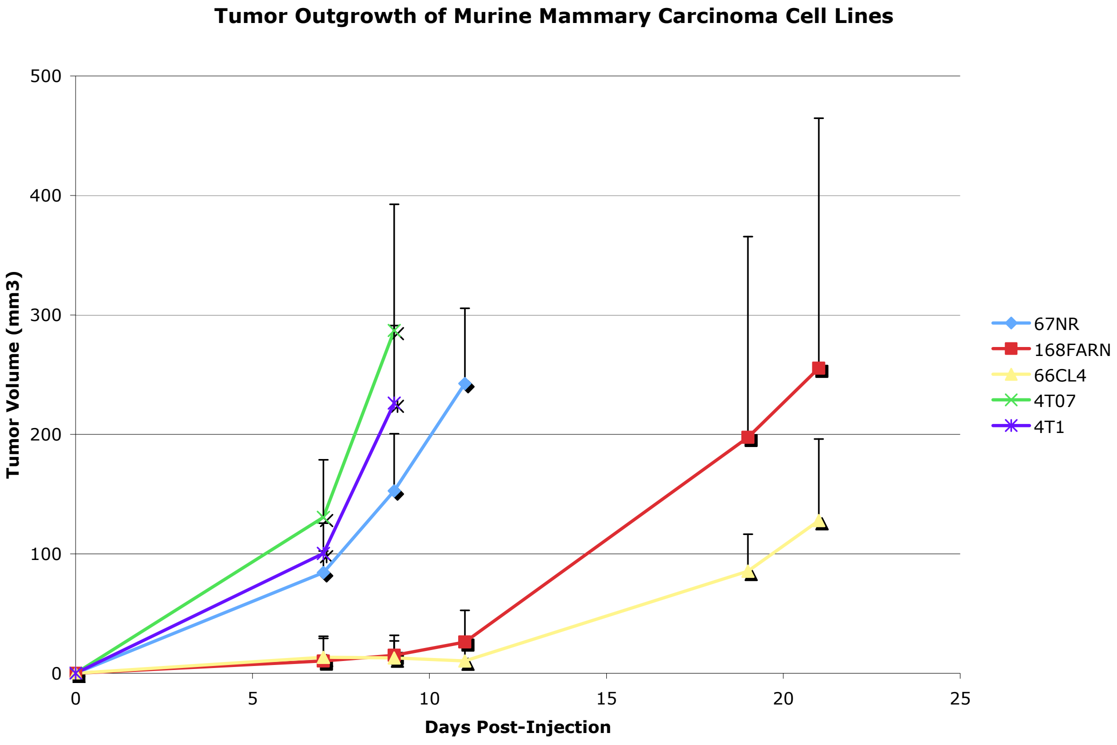

Figure 2.14: Tumor outgrowth of Murine Carcinoma cell lines

The tumor volume obtained in function of days after the tumor injection

In order to investigate more substantially our sample mislabelling hypothesis, we selected five genes whose expressions are variable across the five tumor samples and that should allow us to verify the sample labels: Tmprss6 (NM_027902), Cdh1 (NM_009864), Tacstd1 (NM_008532), Adh7 (NM_009626) and Twist1 NM_011658). We performed real time qRT-PCR at three RNA levels corresponding to three process points between the cell lines injection stage and the tumor RNA extraction stage. (1) RNA samples derived from new tumor RNA extractions of the original mice injected; (2) new cell line injections of mice were done; after the tumors have grown to an equal volume, we extracted the derived RNA samples; (3) the original tumor RNA extracted used for the array hybridizations. We tested the expression level of the five genes at the three process points by real time qRT-PCR. For each gene we compared the expression levels of the corresponding tumor samples in the three levels. Since the expression patterns of tested genes are almost identical at the three process points, this definitively discarded the hypothesis that RNA samples had been mislabeled somewhere in the course of analysis in our laboratories.

As the expression profile of the non-metastatic tumor sample 67NR appears quite similar to 4T1, we looked for evidence of further metastases using a spontaneous metastasis assay in the lung. A spontaneous metastasis assay is one where tumor cells are injected into the orthotopic site (mammary fat pad) where they must form a primary tumor before they can metastasize to either the lung or the bone. Spontaneous metastases revealed that 67NR is in fact highly metastatic. The unexplained high metastatic potential of 67NR revealed by this assay forced us to discard this sample from analysis. Concerning the 66Cl4 cells, they were very poorly metastatic in the spontaneous metastasis assay. These cells have differential growth characteristics at the primary site. When injected, only around 50% to 70% of them have a take rate comparable to other cell lines having a 100% take rate. The tumors themselves grow much slower; therefore, we do not know if any expected gene expression changes we see are due to the tumor outgrowth or the metastatic capacity. As a result, to avoid any confusion, 66C14 was removed from further analysis.

Although we were never able to conclusively determine why the two problematic samples did not exhibit the appropriate metastatic behaviours, the most likely hypothesis is that a sample mixup has occurred at the laboratories from which we had received the cell lines. A less likely hypothesis is a spontaneous mutation having occurred during the cell culture process. As a result, the troubleshooting and retracing our steps has taken several months of work, and has become a substantial part of this Master’s thesis. While perhaps not valuable from the point of advancement of science, for me this exercise has been a valuable experience in troubleshooting and laboratory experimentation.

Supplementary Data

Additional file 1

UCSC browser links illustrating probeset level expression differences (fold-change and p-values) for the top 143 isoforms differentially expressed between the samples, obtained from the probeset level analysis.

File location :

http://www.genome.mcgill.ca/majewski/mouse_UCSC_wiggle_track_files/EXPR_TOP_CANDIDATES_FC_PV.doc

Additional file 2

UCSC browser links illustrating the probeset level expression differences (fold-change and p-values) as well as the normalized (SI) differences for the top 60 isoforms differentially expressed between the samples, obtained from the Splicing Index analysis.

File location:

http://www.genome.mcgill.ca/majewski/mouse_UCSC_wiggle_track_files/SI_TOP_CANDIDATES_FC_PV.doc

Discussion et conclusion générales

Discussion

La puce exon d'Affymetrix est une plateforme très sensible aux variations de l'expression des gènes qui constituent un mécanisme capital dans la régulation de nombreux processus biologiques. Nous avons montré que la quantification de l'expression génique par la puce exon est analogue à celle des puces des générations antérieures à savoir les puces ADN à extrémité 3' ciblée. Cette analogie a été établie en utilisant les échantillons de tissus de référence et ceux des tissus du cerveau de l'étude du consortium MAQC. Ces échantillons offrent une comparaison facile de la quantification de l'expression des gènes avec d'autres plateformes préalablement testées. Ils permettent également la détection de l'épissage alternatif dans les tissus neuraux qui sont connus être très sujets à l'épissage alternatif.

Cependant la concordance entre la puce exon et les deux puces ADN à extrémité 3' ciblée (Illumina et Affymetrix U133) est légèrement moins élevée que celle entre les deux puces ADN à extrémité 3' ciblée. Ceci est dû entre autre au fait que dans le protocole d'amplification de la puce exon, les amorces sont placées de façon aléatoire sur tout l'ensemble du transcrit alors que dans les puces ADN à extrémité 3' ciblée, les amorces sont spécifiques et placées sur la queue polyadénylation de l'extrémité 3'. Or tous les gènes ne possèdent pas de queue polyadénylation; c'est le cas de nombreux gènes codant pour les histones. De plus, une variation d'expression dans la queue polyadénylation pourrait biaiser la quantification du gène par les puces classiques. Les différences d'annotation des gènes entre les manufacturiers des plateformes pourraient également constituer une source de discordance. Pour ce faire, nous avons restreint l'analyse à l'ensemble de gènes consistants entre les trois plateformes.

L’étape la plus délicate dans l’analyse des puces exons est le découplage du taux d’inclusion du gène et celui de ses exons individuels dans les tissus. Les deux méthodes (l’index d’épissage et l’intensité du *probeset*) que nous avons principalement étudiées ont pour lacune de présumer une relation linéaire entre le *probeset* et le *metaprobeset*. Mais plusieurs cas transgressent cette linéarité; c’est l’exemple des sondes hautement exprimées (saturant leur site lors de l’hybridation) ou des sondes dont le niveau d’expression est proche du bruit de fond. De tels cas augmentent le taux de faux positifs dans les analyses. De plus, la puce exon présente des biais qui rendent son analyse plus difficile. L’intensité d’hybridation est faible à l’ extrémités 3’ de l’ARNm; ceci est du au fait que le protocole d’amplification utilisé par la puce exon couvre très peu les extrémités 3’ entraînant ainsi un faible niveau d’expression (près du bruit de fond) des sondes situées à cette extrémité ; par consequent les exons à cette extrémité sont fréquemment identifiés comme ayant subi des événements d’épissage alternatif de saut d’exons. On note également à l’extrémité 5’ un biais du signal d’hybridation des sondes situées à cette extrémité. Ce biais est dû au bruit de fond produit par le contenu élevé en bases GC de ces sondes (localisées proche des promoteurs qui sont généralement non-méthylés et riches en bases GC et en îlots CpG) qui cause des hybridations non spécifiques. De telles erreurs peuvent être réduites en introduisant des étapes supplémentaires de filtrage qui en échange restreindraient l’ensemble de gènes recouvrant le génome.

Dans la seconde partie de nos travaux, nous avons effectué une utilisation avertie de la puce exon pour l’analyse des variations d’épissage dans cinq différents types de tumeurs du cancer de sein non métastatiques (67NR), faiblement métastatiques (168FARN, 4T07), et très métastatiques (66C14 et 4T1). L’expression des gènes, à l’échelle des exons, dans les cinq tumeurs cancéreuses mammaires a été quantifiée avec la puce exon. Une analyse préliminaire du profil d'expression des tumeurs a été faite par une analyse en composantes principales (ACP) qui à partir de l'expression des gènes, donne une représentation synthétique et graphique de la distribution des tumeurs dans l'espace. Nous nous attendions à ce que les tumeurs soient groupées par similitude de potentiel métastatique. A la lumière des résultats des différentes qRT-PCR faites aux étapes clefs de l'expérimentation, nous avons pu écarter l'hypothèse d'une erreur technique au cours de l'expérience. De plus l'expérience de *spontaneous metastasis essay* effectuée montre que NR67, la lignée supposée être la moins métastatique, est en effet très métastatique et que 66Cl4 est plutôt faiblement métastatique. Cette dernière lorsqu'injectée, seulement 50% à 70% des cellules commencent à croitre comparativement aux autres dont 100% des cellules amorcent la croissance. Nous ne pouvons donc pas savoir si une variation d'expression est due au potentiel métastatique ou reflète le faible taux de croissance. L'élimination des lignées 67NR et 66C14 des analyses nous a permis d'éviter toute confusion d'interprétation des résultats.

Les analyses statistiques et manuelles rigoureuses révèlent un ensemble établi de 2623 gènes exprimés différemment ou presentant des isoformes alternatifs entre les tumeurs cancéreuses. La gestion du taux de faux positifs étant un problème majeur de la puce exon, nous avons mis sur pieds plusieurs filtres très efficaces dont les plus cruciaux sont l'élimination des *probesets* présentant une non-linéarité avec leur *metaprobeset*, et la suppression des *probesets* ayant une très grande expression normalisée (par celle du *metaprobeset*) celle-ci pouvant refléter une cross-hybridation à d'autres séquences. Nous notons également qu'au cours de la pathogenèse du cancer, plusieurs régions non codantes dans les conditions biologiques normales sont susceptibles d' être exprimées à cause du désordre de l'expression génique. C'est le cas de plusieurs retentions d'introns que nous avons identifiées, ceux-ci constituant une grande proportion des événements d'épissage alternatifs interprétables. Cependant, 92.3% des gènes significatifs présentent des patterns d'expression nébuleux; ceci pourrait refléter la complexité de l'expression des gènes dans le cancer. Plusieurs types d'événements d'épissage alternatifs peuvent se produire au sein d'un même isoforme de gène et rendre son pattern d'expression difficile à interpréter. Les variations d'expression génique se produisant dans le cancer ne sont pas toujours catégoriques et explicables par les types standards existants. L'élaboration des gènes dans le cancer est tortueuse à élucider par n'importe quelle technologie jusqu'à cette date. Bien que la puce exon soit un outil robuste, il n'est cependant pas parfait. Hors mis le biais possible aux extrémités 5' et 3' des gènes, la puce exon peut manquer ou mal interpréter certains événements d'épissages. Par exemple une limitation fondamentale à la puce exon et également à toute autre approche de quantification de l'expression des gènes est l'épissage introduisant des aberrations dans les transcrits tels que les codons stop prématurés ou absents. Si ces transcrits sont rapidement dégradés par les mécanismes de surveillance de l'ARN qui assurent la qualité et la fidélité des molécules d'ARNm, leur niveau d'expression pourrait être partiellement ou non détectable à cause de l'instabilité de leur concentration dans la cellule.

Nous avons effectué la corrélation de nos données avec deux études anterieures ayant utilisé les mêmes lignées cellulaires pour l'analyse des variations d'expression génique dans le cancer de sein. La quasi totalité des gènes significatifs communs aux trois études sont concordants, ceci réitérant que la puce exon est un quantificateur d'expression génique au même titre que les autres plates formes traditionnelles. Cependant, quelques gènes significatifs présentent des ratios d'expression discordants entre les tumeurs. Notre hypothèse est l'occurrence d'un ou de plusieurs facteurs biologiques ou techniques pouvant affecter l'expression des gènes : le taux de croissance des tumeurs, les conditions expérimentales, l'instabilité génétique des lignées cancéreuses ou le site d'injection des cellules cancéreuses.

Pour comprendre les enjeux biologiques derrière ces variations d'expressions et d'isoformes, nous avons fait une analyse du réseau des gènes statistiquement significatifs. Près de la moitié des gènes répertoriés ont des antécédents dans le cancer et les désordres génétiques. La majorité de ces gènes, dans les conditions biologiques normales, jouent un rôle dans la croissance, l'adhésion, l'interaction et la prolifération cellulaire. L'identification des gènes n'étant pas encore associés à une quelconque maladie, mais impliqués dans les réseaux de gènes liés au cancer pourrait être de nouveaux isoformes de gènes spécifiques au cancer de sein. Le réseau de gènes nous montre également un grand nombre de gènes, subissant des variations d'expression ou d'isoforme, interagissant avec des complexes protéiques qui sont impliqués dans des processus biologiques et dont la dérégulation joue un rôle dans la pathogénèse du cancer et les désordres génétiques. C'est l'exemple des interactions entre plusieurs sous unités du collagène qui sont surexprimées dans la tumeur la plus métastatique 4T1. La surexpression des protéines du collagène joue un rôle dans l'invasion, la migration et la prolifération des cellules cancéreuses dans plusieurs types de cancers [116-123].

À la lumière de nos résultats, nous énonçons que dans le cancer de sein, plusieurs gènes subissent des variations d'expression. Les protéines encodées par de tels gènes pourraient déstabiliser les fonctions normales biologiques. Puisque la plupart de nos gènes candidats ont été rapportés dans nombreux cancers, nous avons l'assurance que les événements d'épissage observés sont biologiquement réels et pertinent dans le mécanisme du cancer de sein. Les variations observées concernent tout l'ensemble du transcrit ou peuvent être très subtiles et n'affecter qu'une partie du transcrit tel qu'un exon. Comparativement à d'autres approches basées sur les puces ADN interrogeant tout l'ensemble du transcrit, l'étude de l'épissage alternatif dans le cancer de sein à l'échelle exonique offre une meilleure précision sur la nature de l'événement d'épissage affectant les transcrits. Ceci pourrait conduire à des diagnostiques ou à des traitements plus précis.

Conclusion

La puce exon fournit des informations absentes ou mal représentées dans les puces ADN classiques. Cependant, l'analyse des isoformes des gènes par la puce exon est très complexe et souffre du taux élevé de faux positifs. Ceci requiert plus d'analyses manuelles comparativement aux puces traditionnelles. Néanmoins la puce exon demeure un outil robuste pour une analyse détaillée de l'épissage alternatif et dont l'efficacité a été mise en évidence dans plusieurs études de variation d'épissage dans divers systèmes biologiques. Nous avons montré que les gènes dans les cellules cancéreuses mammaires présentent des variations d'épissage entre les différents stades pathologiques, certaines régions géniques s'exprimant différemment entre les types de tumeurs. L'investigation du réseau des gènes montre que ces perturbations affectent les gènes associés aux fonctions cellulaires dont le dysfonctionnement favorise la prolifération, l'invasion et la progression des cellules tumorales. L'identification des événements d'épissage, en particulier les nouveaux événements, sont d'une grande importance en cancérologie. En établissant les gènes qui participent activement dans les différents stades de développement du cancer, les produits des gènes ayant un patron d'expression spécifique à un stade pathologique peuvent être utilisés comme des biomarqueurs. Les isoformes des gènes identifiés peuvent éventuellement servir de cible thérapeutique. Ceci peut être accompli en ciblant et inhibant l'expression d'un isoforme spécifique qui altère une fonction normale biologique. La caractérisation des régions géniques subissant l'épissage alternatif et leurs fonctions biologiques pourrait contribuer à l'éclaircissement du mécanisme liant l'épissage alternatif et le cancer.

Bibliographie

1. Faustino, N.A. and T.A. Cooper, *Pre-mRNA splicing and human disease.* Genes Dev, 2003. **17**(4): p. 419-37.

2. Maniatis, T. and B. Tasic, *Alternative pre-mRNA splicing and proteome expansion in metazoans.* Nature, 2002. **418**(6894): p. 236-43.

3. Black, D.L., *Mechanisms of alternative pre-messenger RNA splicing.* Annu Rev Biochem, 2003. **72**: p. 291-336.

4. Pan, Q., et al., *Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing.* Nat Genet, 2008. **40**(12): p. 1413-5.

5. Wang, E.T., et al., *Alternative isoform regulation in human tissue transcriptomes.* Nature, 2008. **456**(7221): p. 470-6.

6. Affymetrix. *Identifying and Validating Alternative Splicing Events*. 2006; Available from: http://www.affymetrix.com/support/technical/technotes/id_altsplicingevents_technote.pdf.

7. Wu, J.Y., H. Tang, and N. Havlioglu, *Alternative pre-mRNA splicing and regulation of programmed cell death.* Prog Mol Subcell Biol, 2003. **31**: p. 153-85.

8. Syken, J., T. De-Medina, and K. Munger, *TID1, a human homolog of the Drosophila tumor suppressor l(2)tid, encodes two mitochondrial modulators of apoptosis with opposing functions.* Proc Natl Acad Sci U S A, 1999. **96**(15): p. 8499-504.

9. Kalnina, Z., et al., *Alterations of pre-mRNA splicing in cancer.* Genes Chromosomes Cancer, 2005. **42**(4): p. 342-57.

10. Liu, H.X., et al., *A mechanism for exon skipping caused by nonsense or missense mutations in BRCA1 and other genes.* Nat Genet, 2001. **27**(1): p. 55-8.

11. Neklason, D.W., et al., *Intron 4 mutation in APC gene results in splice defect and attenuated FAP phenotype.* Fam Cancer, 2004. **3**(1): p. 35-40.

12. Stickeler, E., et al., *Stage-specific changes in SR splicing factors and alternative splicing in mammary tumorigenesis.* Oncogene, 1999. **18**(24): p. 3574-82.

13. van Hoof, A., et al., *Exosome-mediated recognition and degradation of mRNAs lacking a termination codon.* Science, 2002. **295**(5563): p. 2262-4.

14. Frischmeyer, P.A., et al., *An mRNA surveillance mechanism that eliminates transcripts lacking termination codons.* Science, 2002. **295**(5563): p. 2258-61.

15. Manabe, T., et al., *Induced HMGA1a expression causes aberrant splicing of Presenilin-2 pre-mRNA in sporadic Alzheimer's disease.* Cell Death Differ, 2003. **10**(6): p. 698-708.

16. Mazoyer, S., et al., *A BRCA1 nonsense mutation causes exon skipping.* Am J Hum Genet, 1998. **62**(3): p. 713-5.

17. Kan, Z., et al., *Gene structure prediction and alternative splicing analysis using genomically aligned ESTs.* Genome Res, 2001. **11**(5): p. 889-900.

18. Li, C., et al., *Cell type and culture condition-dependent alternative splicing in human breast cancer cells revealed by splicing-sensitive microarrays.* Cancer Res, 2006. **66**(4): p. 1990-9.

19. Yeakley, J.M., et al., *Profiling alternative splicing on fiber-optic arrays.* Nat Biotechnol, 2002. **20**(4): p. 353-8.

20. Johnson, J.M., et al., *Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays.* Science, 2003. **302**(5653): p. 2141-4.

21. Lee, J.H., et al., *Alterations in Gemin5 expression contribute to alternative mRNA splicing patterns and tumor cell motility.* Cancer Res, 2008. **68**(3): p. 639-44.

22. *Alternative Splicing Arrays of Jivan compagny*. Available from: http://www.jivanbio.com/products/Genome-Wide_Alternative_Splicing_Array.php.

23. *GeneChip ExonArray design.* Affymetrix Technical Note.

24. Lou, Y., et al., *Epithelial-mesenchymal transition (EMT) is not sufficient for spontaneous murine breast cancer metastasis.* Dev Dyn, 2008. **237**(10): p. 2755-68.

25. Kapur, K., et al., *Exon arrays provide accurate assessments of gene expression.* Genome Biol, 2007. **8**(5): p. R82.

26. Thorsen, K., et al., *Alternative splicing in colon, bladder, and prostate cancer identified by exon array analysis.* Mol Cell Proteomics, 2008. **7**(7): p. 1214-24.

27. Frey, B.J., et al., *Genome-wide analysis of mouse transcripts using exon microarrays and factor graphs.* Nat Genet, 2005. **37**(9): p. 991-6.

28. Lee, C. and M. Roy, *Analysis of alternative splicing with microarrays: successes and challenges.* Genome Biol, 2004. **5**(7): p. 231.

29. Clark, T.A., et al., *Discovery of tissue-specific exons using comprehensive human exon microarrays.* Genome Biol, 2007. **8**(4): p. R64.

30. Gardina, P.J., et al., *Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array.* BMC Genomics, 2006. **7**: p. 325.

31. Hung, L.H., et al., *Diverse roles of hnRNP L in mammalian mRNA processing: a combined microarray and RNAi analysis.* RNA, 2008. **14**(2): p. 284-96.

32. Kim, E., A. Goren, and G. Ast, *Insights into the connection between cancer and alternative splicing.* Trends Genet, 2008. **24**(1): p. 7-10.

33. Kwan, T., et al., *Genome-wide analysis of transcript isoform variation in humans.* Nat Genet, 2008. **40**(2): p. 225-31.

34. McKee, A.E., et al., *Exon expression profiling reveals stimulus-mediated exon use in neural cells.* Genome Biol, 2007. **8**(8): p. R159.

35. Thorsen, K., et al., *Alternative splicing in colon, bladder, and prostate cancer identified by exon-array analysis.* Mol Cell Proteomics, 2008.

36. Yeo, G.W., et al., *Alternative splicing events identified in human embryonic stem cells and neural progenitors.* PLoS Comput Biol, 2007. **3**(10): p. 1951-67.

37. Kwan, T., et al., *Heritability of alternative splicing in the human genome.* Genome Res, 2007. **17**(8): p. 1210-8.

38. Canales, R.D., et al., *Evaluation of DNA microarray results with quantitative gene expression platforms.* Nat Biotechnol, 2006. **24**(9): p. 1115-22.

39. Shi, L., et al., *The MicroArray Quality Control (MAQC) project shows inter- and intraplatform reproducibility of gene expression measurements.* Nat Biotechnol, 2006. **24**(9): p. 1151-61.

40. Okoniewski, M.J., et al., *High correspondence between Affymetrix exon and standard expression arrays.* Biotechniques, 2007. **42**(2): p. 181-5.

41. Robinson, M.D. and T.P. Speed, *A comparison of Affymetrix gene expression arrays.* BMC Bioinformatics, 2007. **8**: p. 449.

42. Xing, Y., et al., *Assessing the conservation of mammalian gene expression using high-density exon arrays.* Mol Biol Evol, 2007. **24**(6): p. 1283-5.

43. Karolchik, D., et al., *The UCSC Genome Browser Database: 2008 update.* Nucleic Acids Res, 2008. **36**(Database issue): p. D773-9.

44. Adesnik, M., et al., *Evidence that all messenger RNA molecules (except histone messenger RNA) contain Poly (A) sequences and that the Poly(A) has a nuclear function.* J Mol Biol, 1972. **71**(1): p. 21-30.

45. Snider, B.J. and M. Morrison-Bogorad, *Brain non-adenylated mRNAs.* Brain Res Brain Res Rev, 1992. **17**(3): p. 263-82.

46. Okoniewski, M.J., et al., *An annotation infrastructure for the analysis and interpretation of Affymetrix exon array data.* Genome Biol, 2007. **8**(5): p. R79.

47. Purdom, E., et al., *FIRMA: a method for detection of alternative splicing from exon array data.* Bioinformatics, 2008.

48. Okoniewski, M.J. and C.J. Miller, *Comprehensive analysis of affymetrix exon arrays using BioConductor.* PLoS Comput Biol, 2008. **4**(2): p. e6.

49. Benjamini, Y., et al., *Controlling the false discovery rate in behavior genetics research.* Behav Brain Res, 2001. **125**(1-2): p. 279-84.

50. Kim, N., et al., *The ASAP II database: analysis and comparative genomics of alternative splicing in 15 animal species.* Nucleic Acids Res, 2007. **35**(Database issue): p. D93-8.

51. Holste, D., et al., *HOLLYWOOD: a comparative relational database of alternative splicing.* Nucleic Acids Res, 2006. **34**(Database issue): p. D56-62.

52. Xing, Y., et al., *MADS: A new and improved method for analysis of differential alternative splicing by exon-tiling microarrays.* RNA, 2008.

53. Louie, E., J. Ott, and J. Majewski, *Nucleotide frequency variation across human genes.* Genome Res, 2003. **13**(12): p. 2594-601.

54. Majewski, J. and J. Ott, *Distribution and characterization of regulatory elements in the human genome.* Genome Res, 2002. **12**(12): p. 1827-36.

55. Dai, M., et al., *Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data.* Nucleic Acids Res, 2005. **33**(20): p. e175.

56. Benovoy, D., T. Kwan, and J. Majewski, *Effect of polymorphisms within probe-target sequences on olignonucleotide microarray experiments.* Nucleic Acids Res, 2008. **36**(13): p. 4417-23.

57. Affymetrix. *Guide to Probe Logarithmic Intensity Error (PLIER) Estimation*. 2005; Available from: www.affymetrix.com/support/technical/technotes/plier_technote.pdf.

58. Irizarry, R.A., et al., *Exploration, normalization, and summaries of high density oligonucleotide array probe level data.* Biostatistics, 2003. **4**(2): p. 249-64.

59. Boise, L.H., et al., *bcl-x, a bcl-2-related gene that functions as a dominant regulator of apoptotic cell death.* Cell, 1993. **74**(4): p. 597-608.

60. Venables, J.P., *Unbalanced alternative splicing and its significance in cancer.* Bioessays, 2006. **28**(4): p. 378-86.

61. Brennan, D.J., et al., *CA IX is an independent prognostic marker in premenopausal breast cancer patients with one to three positive lymph nodes and a putative marker of radiation resistance.* Clin Cancer Res, 2006. **12**(21): p. 6421-31.

62. Trastour, C., et al., *HIF-1alpha and CA IX staining in invasive breast carcinomas: prognosis and treatment outcome.* Int J Cancer, 2007. **120**(7): p. 1451-8.

63. Hussain, S.A., et al., *Hypoxia-regulated carbonic anhydrase IX expression is associated with poor survival in patients with invasive breast cancer.* Br J Cancer, 2007. **96**(1): p. 104-9.

64. Aslakson, C.J. and F.R. Miller, *Selective events in the metastatic process defined by analysis of the sequential dissemination of subpopulations of a mouse mammary tumor.* Cancer Res, 1992. **52**(6): p. 1399-405.

65. Nardone, R.M., *Curbing rampant cross-contamination and misidentification of cell lines.* Biotechniques, 2008. **45**(3): p. 221-7.

66. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Res, 2002. **12**(6): p. 996-1006.

67. Mori, Y., et al., *Instabilotyping reveals unique mutational spectra in microsatellite-unstable gastric cancers.* Cancer Res, 2002. **62**(13): p. 3641-5.

68. Johnson, K.W. and K.A. Smith, *Molecular cloning of a novel human cdc2/CDC28-like protein kinase.* J Biol Chem, 1991. **266**(6): p. 3402-7.

69. Chiba, N., et al., *Binding of liganded vitamin D receptor to the vitamin D receptor interacting protein coactivator complex induces interaction with RNA polymerase II holoenzyme.* J Biol Chem, 2000. **275**(15): p. 10719-22.

70. Marhaba, R. and M. Zoller, *CD44 in cancer progression: adhesion, migration and growth regulation.* J Mol Histol, 2004. **35**(3): p. 211-31.

71. Draffin, J.E., et al., *CD44 potentiates the adherence of metastatic prostate and breast cancer cells to bone marrow endothelial cells.* Cancer Res, 2004. **64**(16): p. 5702-11.

72. Bourguignon, L.Y., et al., *CD44 interaction with Na+-H+ exchanger (NHE1) creates acidic microenvironments leading to hyaluronidase-2 and cathepsin B activation and breast tumor cell invasion.* J Biol Chem, 2004. **279**(26): p. 26991-7007.

73. Bates, R.C., et al., *A CD44 survival pathway triggers chemoresistance via lyn kinase and phosphoinositide 3-kinase/Akt in colon carcinoma cells.* Cancer Res, 2001. **61**(13): p. 5275-83.

74. Fujii, K., et al., *CD44 is the physiological trigger of Fas up-regulation on rheumatoid synovial cells.* J Immunol, 2001. **167**(3): p. 1198-203.

75. Bourguignon, L.Y., et al., *Hyaluronan-CD44 interaction with IQGAP1 promotes Cdc42 and ERK signaling, leading to actin binding, Elk-1/estrogen receptor transcriptional activation, and ovarian cancer progression.* J Biol Chem, 2005. **280**(12): p. 11961-72.

76. Aziz, K.A., et al., *Involvement of CD44-hyaluronan interaction in malignant cell homing and fibronectin synthesis in hairy cell leukemia.* Blood, 2000. **96**(9): p. 3161-7.

77. Barbour, A.P., et al., *Expression of the CD44v2-10 isoform confers a metastatic phenotype: importance of the heparan sulfate attachment site CD44v3.* Cancer Res, 2003. **63**(4): p. 887-92.

78. Middlemas, D.S., et al., *Brain-derived neurotrophic factor promotes survival and chemoprotection of human neuroblastoma cells.* J Biol Chem, 1999. **274**(23): p. 16451-60.

79. McAlhany, R.E., Jr., J.R. West, and R.C. Miranda, *Glial-derived neurotrophic factor (GDNF) prevents ethanol-induced apoptosis and JUN kinase phosphorylation.* Brain Res Dev Brain Res, 2000. **119**(2): p. 209-16.

80. Biton, S., et al., *Nuclear ataxia-telangiectasia mutated (ATM) mediates the cellular response to DNA double strand breaks in human neuron-like cells.* J Biol Chem, 2006. **281**(25): p. 17482-91.

81. Takahashi, H., et al., *Tissue transglutaminase, coagulation factor XIII, and the pro-polypeptide of von Willebrand factor are all ligands for the integrins alpha 9beta 1 and alpha 4beta 1.* J Biol Chem, 2000. **275**(31): p. 23589-95.

82. Uekita, T., et al., *CUB domain-containing protein 1 is a novel regulator of anoikis resistance in lung adenocarcinoma.* Mol Cell Biol, 2007. **27**(21): p. 7649-60.

83. van Golen, C.M., et al., *N-Myc overexpression leads to decreased beta1 integrin expression and increased apoptosis in human neuroblastoma cells.* Oncogene, 2003. **22**(17): p. 2664-73.

84. Gendron, S., J. Couture, and F. Aoudjit, *Integrin alpha2beta1 inhibits Fas-mediated apoptosis in T lymphocytes by protein phosphatase 2A-dependent activation of the MAPK/ERK pathway.* J Biol Chem, 2003. **278**(49): p. 48633-43.

85. Kim, S., et al., *Inhibition of endothelial cell survival and angiogenesis by protein kinase A.* J Clin Invest, 2002. **110**(7): p. 933-41.

86. Faraldo, M.M., et al., *Perturbation of beta1-integrin function alters the development of murine mammary gland.* EMBO J, 1998. **17**(8): p. 2139-47.

87. Vlahakis, N.E., et al., *Integrin alpha9beta1 directly binds to vascular endothelial growth factor (VEGF)-A and contributes to VEGF-A-induced angiogenesis.* J Biol Chem, 2007. **282**(20): p. 15187-96.

88. Suzuki, A., et al., *ARK5 suppresses the cell death induced by nutrient starvation and death receptors via inhibition of caspase 8 activation, but not by chemotherapeutic agents or UV irradiation.* Oncogene, 2003. **22**(40): p. 6177-82.

89. Suzuki, A., et al., *ARK5 is transcriptionally regulated by the Large-MAF family and mediates IGF-1-induced cell invasion in multiple myeloma: ARK5 as a new molecular determinant of malignant multiple myeloma.* Oncogene, 2005. **24**(46): p. 6936-44.

90. Vitovski, S., et al., *Investigating the interaction between osteoprotegerin and receptor activator of NF-kappaB or tumor necrosis factor-related apoptosis-inducing ligand: evidence for a pivotal role for osteoprotegerin in regulating two distinct pathways.* J Biol Chem, 2007. **282**(43): p. 31601-9.

91. Hofbauer, L.C., et al., *The roles of osteoprotegerin and osteoprotegerin ligand in the paracrine regulation of bone resorption.* J Bone Miner Res, 2000. **15**(1): p. 2-12.

92. Qi, X., et al., *p38 MAPK activation selectively induces cell death in K-ras-mutated human colon cancer cells through regulation of vitamin D receptor.* J Biol Chem, 2004. **279**(21): p. 22138-44.

93. Puccetti, E., et al., *AML-associated translocation products block vitamin D(3)-induced differentiation by sequestering the vitamin D(3) receptor.* Cancer Res, 2002. **62**(23): p. 7050-8.

94. Lobov, I.B., P.C. Brooks, and R.A. Lang, *Angiopoietin-2 displays VEGF-dependent modulation of capillary structure and endothelial cell survival in vivo.* Proc Natl Acad Sci U S A, 2002. **99**(17): p. 11205-10.

95. Stockinger, A., et al., *E-cadherin regulates cell growth by modulating proliferation-dependent beta-catenin transcriptional activity.* J Cell Biol, 2001. **154**(6): p. 1185-96.

96. Bindels, E.M., et al., *E-cadherin promotes intraepithelial expansion of bladder carcinoma cells in an in vitro model of carcinoma in situ.* Cancer Res, 2000. **60**(1): p. 177-83.

97. Luo, J., D.M. Lubaroff, and M.J. Hendrix, *Suppression of prostate cancer invasive potential and matrix metalloproteinase activity by E-cadherin transfection.* Cancer Res, 1999. **59**(15): p. 3552-6.

98. Evert, B.O., U. Wullner, and T. Klockgether, *Cell death in polyglutamine diseases.* Cell Tissue Res, 2000. **301**(1): p. 189-204.

99. Bernerd, F., A. Sarasin, and T. Magnaldo, *Galectin-7 overexpression is associated with the apoptotic process in UVB-induced sunburn keratinocytes.* Proc Natl Acad Sci U S A, 1999. **96**(20): p. 11329-34.

100. Kaiser, R.A., et al., *Targeted inhibition of p38 mitogen-activated protein kinase antagonizes cardiac injury and cell death following ischemia-reperfusion in vivo.* J Biol Chem, 2004. **279**(15): p. 15524-30.

101. Silva, G., et al., *The antiapoptotic effect of heme oxygenase-1 in endothelial cells involves the degradation of p38 alpha MAPK isoform.* J Immunol, 2006. **177**(3): p. 1894-903.

102. Nagata, Y. and K. Todokoro, *Requirement of activation of JNK and p38 for environmental stress-induced erythroid differentiation and apoptosis and of inhibition of ERK for apoptosis.* Blood, 1999. **94**(3): p. 853-63.

103. Sabourin, L.A. and M.A. Rudnicki, *Induction of apoptosis by SLK, a Ste20-related kinase.* Oncogene, 1999. **18**(52): p. 7566-75.

104. Ogawa, T., et al., *The short arm of laminin gamma2 chain of laminin-5 (laminin-332) binds syndecan-1 and regulates cellular adhesion and migration by suppressing phosphorylation of integrin beta4 chain.* Mol Biol Cell, 2007. **18**(5): p. 1621-33.

105. Chung, J., et al., *The Met receptor and alpha 6 beta 4 integrin can function independently to promote carcinoma invasion.* J Biol Chem, 2004. **279**(31): p. 32287-93.

106. Bertotti, A., P.M. Comoglio, and L. Trusolino, *Beta4 integrin is a transforming molecule that unleashes Met tyrosine kinase tumorigenesis.* Cancer Res, 2005. **65**(23): p. 10674-9.

107. Rots, N.Y., et al., *A differential screen for ligand-regulated genes: identification of HoxA10 as a target of vitamin D3 induction in myeloid leukemic cells.* Mol Cell Biol, 1998. **18**(4): p. 1911-8.

108. Meloche, S. and J. Pouyssegur, *The ERK1/2 mitogen-activated protein kinase pathway as a master regulator of the G1- to S-phase transition.* Oncogene, 2007. **26**(22): p. 3227-39.

109. Kronblad, A., et al., *ERK1/2 inhibition increases antiestrogen treatment efficacy by interfering with hypoxia-induced downregulation of ERalpha: a combination therapy*

potentially targeting hypoxic and dormant tumor cells. Oncogene, 2005. **24**(45): p. 6835-41.

110. Wong, A.S. and B.M. Gumbiner, *Adhesion-independent mechanism for suppression of tumor cell invasion by E-cadherin.* J Cell Biol, 2003. **161**(6): p. 1191-203.

111. Bemmo, A., et al., *Gene expression and isoform variation analysis using Affymetrix Exon Arrays.* BMC Genomics, 2008. **9**: p. 529.

112. Yang, J., et al., *Twist, a master regulator of morphogenesis, plays an essential role in tumor metastasis.* Cell, 2004. **117**(7): p. 927-39.

113. Zimmermann, J., et al., *Proteasome- and p38-dependent regulation of ERK3 expression.* J Biol Chem, 2001. **276**(14): p. 10759-66.

114. Macleod, K., et al., *Altered ErbB receptor signaling and gene expression in cisplatin-resistant ovarian cancer.* Cancer Res, 2005. **65**(15): p. 6789-800.

115. Lin, J., et al., *The phosphatidylinositol 3'-kinase pathway is a dominant growth factor-activated cell survival pathway in LNCaP human prostate carcinoma cells.* Cancer Res, 1999. **59**(12): p. 2891-7.

116. Piccaluga, P.P., et al., *Gene expression analysis of peripheral T cell lymphoma, unspecified, reveals distinct profiles and new potential therapeutic targets.* J Clin Invest, 2007. **117**(3): p. 823-34.

117. Nakagawa, H., et al., *Role of cancer-associated stromal fibroblasts in metastatic colon cancer to the liver and their expression profiles.* Oncogene, 2004. **23**(44): p. 7366-77.

118. Donninger, H., et al., *Whole genome expression profiling of advance stage papillary serous ovarian cancer reveals activated pathways.* Oncogene, 2004. **23**(49): p. 8065-77.

119. Sado, Y., et al., *Organization and expression of basement membrane collagen IV genes and their roles in human disorders.* J Biochem, 1998. **123**(5): p. 767-76.

120. Mukhopadhyay, N.K., et al., *Integrin-dependent protein tyrosine phosphorylation is a key regulatory event in collagen-IV-mediated adhesion and proliferation of human lung tumor cell line, Calu-1.* Ann Thorac Surg, 2004. **78**(2): p. 450-7.

121. Pucci-Minafra, I., et al., *Type V collagen induces apoptosis of 8701-BC breast cancer cells and enhances m-calpain expression.* Breast Cancer Res 2000. **2:E008**.

122. Ambiru, S., et al., *Increased serum type IV collagen 7-S levels in patients with hepatic metastasis.* Am J Gastroenterol, 1995. **90**(5): p. 783-7.

123. Hong, W.S., et al., *Elevation of serum type IV collagen in liver cancer as well as liver cirrhosis.* Anticancer Res, 1995. **15**(6B): p. 2777-80.

124. Cichy, J. and E. Pure, *The liberation of CD44.* J Cell Biol, 2003. **161**(5): p. 839-43.

125. Gilmore, T.D., *Introduction to NF-kappaB: players, pathways, perspectives.* Oncogene, 2006. **25**(51): p. 6680-4.

126. Brasier, A.R., *The NF-kappaB regulatory network.* Cardiovasc Toxicol, 2006. **6**(2): p. 111-30.

127. Perkins, N.D., *Integrating cell-signalling pathways with NF-kappaB and IKK function.* Nat Rev Mol Cell Biol, 2007. **8**(1): p. 49-62.

128. Pearson, G., et al., *Mitogen-activated protein (MAP) kinase pathways: regulation and physiological functions.* Endocr Rev, 2001. **22**(2): p. 153-83.

129. Venables, J.P., et al., *Identification of alternative splicing markers for breast cancer.* Cancer Res, 2008. **68**(22): p. 9525-31.

130. Liau, S.S., A. Jazag, and E.E. Whang, *HMGA1 is a determinant of cellular invasiveness and in vivo metastatic potential in pancreatic adenocarcinoma.* Cancer Res, 2006. **66**(24): p. 11613-22.

131. Karlgren, M. and M. Ingelman-Sundberg, *Tumour-specific expression of CYP2W1: its potential as a drug target in cancer therapy.* Expert Opin Ther Targets, 2007. **11**(1): p. 61-7.

132. McKenzie, E.A., *Heparanase: a target for drug discovery in cancer and inflammation.* Br J Pharmacol, 2007. **151**(1): p. 1-14.

133. Zhang, L., et al., *RNA interference: a potential strategy for isoform-specific phosphatidylinositol 3-kinase targeted therapy in ovarian cancer.* Cancer Biol Ther, 2004. **3**(12): p. 1283-9.

134. Duchaine, T.F. and F.J. Slack, *rna interference and micro rna -oriented therapy in cancer: rationales, promises, and challenges.* Curr Oncol, 2009. **16**(4): p. 61-6.

135. Benjamini, Y. and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* J. R. Stat. Soc. Ser. B Methodol, 1995. **57**: p. 289-300.