Published in final edited form as:
Proteomics Clin Appl. 2015 April ; 9(0): 432–436. doi:10.1002/prca.201400126.

Impact of allele-specific peptides in proteome quantification

Linfeng Wu[1,2] and **Michael Snyder**[1]

[1]Department of Genetics, Stanford University School of Medicine, Stanford, California 94305, USA

[2]Caprion Proteomics US LLC, Menlo Park, California 94025, USA

Abstract

Mass spectrometry-based proteome technologies have greatly improved our ability to detect and quantify proteomes across various biological samples. High throughput bottom-up proteome profiling in combination with targeted mass spectrometry method, e.g. selected reaction monitoring (SRM) assay, is emerging as a powerful approach in the field of biomarker discovery. In the past few years, increasing number of studies have attempted to integrate genomic and proteomic data for biomarker discovery. Here we describe how allele-specific peptide can be applied in biomarker discovery and their impact in protein quantification.

Introduction

Many studies have shown the strength of MS-based technologies in assisting discovery, verification and validation of clinical biomarkers. Due to the high complexity and large dynamic range of proteome in biological samples (e.g. plasma, cerebral spinal fluid, tissues et al.), a two-step strategy is popular for biomarker discovery. In the first discovery phase, quantitative bottom-up proteomics approach is utilized to identify differentially expressed proteins in normal and disease samples, using label-free or stable isotope-labeling methods. In the second phase, the candidate differential proteins are subjected to targeted protein analysis, using a more precise and reproducible MS method, i.e. SRM.

SRM is a mass spectrometry method, mainly performed on triple-quadrupole instruments in which two stages of mass filtering are used to increase selectivity (figure 1). In the first stage, a precursor ion of a particular mass is selected in Q1 (i.e. first quadrupole). The selected precursor undergoes fragmentation by collision with neutral gas in Q2 (second quadrupole). In the second stage, instead of monitoring all the possible fragment ions, only a small number of fragment ions are selected for detection in Q3 (third quadrupole). The specific pair of m/z values associated to the selected precursor and fragment ion is referred to as a “transition”. Multiple SRM transitions can be monitored in the same experiment by rapidly toggling between different precursor/fragment pairs, which is refered to as multiple reaction monitoring, MRM. To improve the precision of peptide quantification, a known

Correspondence should be addressed to, Michael Snyder, mpsnyder@stanford.edu, Dept. of Genetics, MC: 5120, 300 Pasteur Dr., M-344, Stanford, CA 94305-5120.

Conflict of interest statement
Dr. Michael Snyder is on the scientific advisory board of Personalis and GenapSys.

amount of stable isotope labeled standard peptides are frequently spiked into samples, and monitored synchronously. In such a case, signal ratios between endogenous and standard peptides are employed for peptide quantification.

SRM assay starts with selecting target peptides that act as surrogates for the quantification of protein of interest. To date, many efforts have been devoted to select peptides with favorable MS properties, as it affects the sensitivity of SRM assay. Numerous properties can affect the MS signal response of different peptides from the same protein, such as peptide hydrophobicity, charge, post-translational modification and structural properties. In general, information from prior experiments is used to identify peptides that are suitable for SRM assays. Publicly accessible databases that contain such peptides from prior proteomic experiments now exist, for example, PeptideAtlas[1], the Global Proteome Machine Database (GPMDB)[2], Pride[3] and genome annotating proteomic pipeline (GAPP)[4]. There are several excellent reviews on protein biomarker discovery and SRM-based proteomics [5, 6]. Here we will mainly focus on a specific set of peptides—allele-specific peptides—for protein quantification, and their impacts on biomarker detection.

An allele is one of a number of different forms of the same gene or same genetic locus. A dipoid organism contains two alleles on every gene, with one allele inherited from each parent. If both alleles are the same, they are homozygous on that gene; if the two alleles are different, they are heterozygous on that gene. A human population typically includes multiple alleles for each gene among various individuals. Therefore, for the same gene, different individuals might express different isoforms under the same circumstances, resulting from alternative splicing, nonsynonymous single nucleotide polymorphisms (nsSNPs), somatic mutations and other reasons. Peptides specific to the allele that alter the amino acid sequence of this peptides, i.e. allele-specific peptides, can be used to identify the expression of specific isoforms or to identify somatic mutations in biological samples.

Detection of protein isoforms for diagnosis and prognosis of disease

Most alleles result in little or no observable phenotypes. However, sometimes different alleles encode different protein sequences (i.e., different isoforms), resulting in observable phenotypic traits, such as skin, hair, and eye pigmentation [7]. Occasionally, some alleles have clinical impact in humans. For example, OAS1 (2’-5’-oligoadenylate synthase 1) is an essential protein involved in viral RNA degradation [8]. Mutation at SNP rs10774671 of *OAS1* gene has been shown to alter OAS1 splicing, and to be associated with susceptibility to West Nile virus infection, where A allele compared to G allele results in lower total protein abundance, reduced OAS1 activity and higher virus accumulation in humans [9, 10]. Therefore, detection of total and isoform-specific peptides of OAS1 protein might provide a genetic risk factor for initial infection with West Nile virus in humans.

To date, most identifications of allele expression in biological samples have been carried out at transcript level, or even directly on the genome sequences. It is mainly due to the higher sensitivity and more well established nucleotide detection technologies, e.g. microarray, and high throughput next-generation sequencing technology [11, 12]. However, in terms of disease diagnosis and prognosis, human body fluids, e.g., plasma/serum, urine, saliva, and cerebrospinal fluid, appear to provide several key advantages including low cost, minimum invasiveness, and easy sample collection and processing [13]. Since human body fluids are often lack of DNA and RNA, it makes proteomic technology a better alternative approach to identify and quantify allele expression in these clinical samples.

Several studies have been carried out to identify and quantify allele expression in human body fluids using SRM assay. Overgaard et al. quantified cardiovascular biomarker fibulin-1 and its circulating isoforms in human plasma [14]. They used bioinformatics analysis to predict total and isoform-specific tryptic peptides, and quantified the absolute amount of total fibulin-1, isoform-1C and -1D using SRM assay combined with stable isotope dilution. They found fibulin-1C was the most abundant isoform in plasma, and circulating fibulin-1 isoforms were homo or hetero multimeric complexes.

In another study, Simon et al. analyzed the absolute quantification of ApoE proteins in the plasma of Alzheimer's disease (AD) patients [15]. Allelic polymorphism of the apolipoprotein E (ApoE) gene results in three protein isoforms (ApoE2, ApoE3 and ApoE4) that differ by only 1 or 2 amino acids [16]. One of the alleles, ApoE ε4, is a risk factor for developing Alzheimer's disease (AD) [17]. However, using ApoE protein level in human body fluids as a diagnostic value to distinguish between AD patients and healthy subjects is of great controversy [18–20]. Simon et al. quantified an ApoE4-specific peptide and a peptide common to all ApoE isoforms by SRM assay in a case-control study (n=669); they found that neither total ApoE and ApoE4 levels nor the ApoE/ApoE4 ratio correlated with the diagnosis of AD[15]. Their study reinforced that plasma ApoE levels had no obvious clinical significance.

Detection of mutant proteins in cancer

In addition to inherit from parents, new variation, through DNA mutations, can occur and accumulate in somatic tissues during development and aging, generating genome mosaics. Somatic mutations are generally random, most of which have either no effect or an adverse effect. These events are known to cause disease in humans, especially cancer.

The detection of proteins encoded by mutant genes is often performed using antibody-based immunoassays. However, a large number of disease-causing mutations are missense mutations, which alter the encoded proteins slightly, often by only a single amino acid. To generate antibodies that distinguish a mutant protein from wild type protein can be difficult. Wang et al. use mass spectrometry to detect and quantify peptides expressed from normal and mutant Ras protein in both cultured cancer cell lines and clinical specimens[21]. They showed that the altered Ras protein products resulting from somatic mutations can be identified directly and quantified by SRM-based method. They demonstrated that it is possible to quantify the number and fraction of mutant Ras protein present in cancer cell lines as well as in clinical specimens such as colorectal and pancreatic tumor tissues and premalignant pancreatic cyst fluids[21]. This study showed the advantage of SRM-based method relative to an antibody-based assay to detect mutant proteins, and its potential usage for cancer diagnosis.

Recently many studies have been carried out to identify somatic variants across thousands of tumours using the latest genome sequencing and analysis methods [22–24]. These efforts have revealed a wide spectrum of mutations that appear to be specific to various cancer types. Targeted mass spectrometry is capable of monitoring mutant peptides in complex biological samples with high sensitivity and specificity. In the future, it might be possible to combine the known gene mutations of various cancers with the power of multiplex SRM, in order to quantify mutant peptides that are highly specific for cancer.

Impact in protein quantitative trait loci (pQTL) mapping

To date, a number of successful studies have been done to identify genomic loci that are associated with transcript levels, termed expression quantitative trait loci (eQTLs) [25–27]. In these studies, global transcript levels are often monitored to quantify gene expression in a cohort. The transcript levels are then tested for correlation with DNA variants, i.e. genotypes. This systems genetic approach provides useful information into the flow of biological information[28], and is greatly facilitated by the recent development of high throughput next-generation sequencing technology[11]. However, variation in transcript level is not a perfect surrogate for protein expression, as the latter is influenced by an array of post-transcriptional regulatory mechanisms. Because of this, the correlation between mRNA and protein levels is generally modest [29]. Therefore, it is beneficial to integrate proteome profiling with genetic variation in a large cohort to provide novel information for understanding the underlying molecular mechanisms.

Currently, very few pQTL ((i.e., genetic loci impacting protein expression) studies have been performed compared to eQTL studies. Recently our group used isobaric tag-based quantitative mass spectrometry to determine the relative protein levels of 5,953 gene products in sequenced lymphoblastoid cell lines (LCL) from 95 diverse individuals [9]. We observed that a large part of variants controlling protein levels are linked to SNPs which either themselves result in an altered amino acid sequence of the protein or are in high linkage disequilibrium (LD) with a nsSNP (data not published).

One technical concern is that it is not always straightforward to assign peptides to the actual protein. In a conventional bottom-up proteomics, proteins are grouped together following parsimony principle, in which the minimum set of proteins that account for all the observed peptides are reported. For example, if protein A is identified with three peptides, protein B with two of those same peptides, and protein C with the other one, it groups protein A, B and C, assuming that only protein A is present. For protein quantification, only peptides unique to each group are used. Therefore, by this approach, if various isoforms are observed in the same sample or dataset, allele-specific peptides will be selected to distinguish different isoforms and used for protein quantification. When testing these isoform levels for their correlation with genetic variants, these proteins are likely to be significantly linked to the SNPs which alter allele-specific peptide sequence or in LD with the causal SNPs.

For example, in our study, two isoforms of Torsin-1A-interacting protein 1 (TOR1AIP1) were detected. These two isoforms only differ by one amino acid deletion, which is caused by a SNP, rs2245425. If we used allele-specific peptides to quantify these two isoforms, both isoform levels are significantly associated with this causal SNP, and the levels of these two proteins are anti-correlated with each other (figure 2). In this case, we actually detected allele-specific expression from *TOR1AIP1* gene, instead of detecting regulation of total TOR1AIP1 protein level.

Such a phenomenon has also been observed by other groups. Johansson et al identified and quantified the abundance of 1,056 tryptic-digested peptides, representing 163 proteins in the plasma of 1,060 individuals using high throughput mass spectrometry [30]. They performed cis-pQTL analysis, and found that about half (5/11) of the associated SNPs either themselves result in an altered amino acid sequence of the peptide or are in high LD with a nsSNP located in the same peptide region.

Therefore, for pQTL analysis, those allele-specific peptides should be excluded from quantifying protein levels that represent total gene expression. Nonetheless, much information can still be obtained from this type of experiment. For example, allele-specific peptide expression can be tested for correlation with other phenotypes, such as clinical trait, to identify protein isoforms that may contribute to the clinical phenotypes. In addition, association test between allele-specific peptide expression and intro-gene SNPs can be used to validate gene mutation and predicted splicing site. These types of studies complement RNA-seq analysis on discovering novel transcripts.

Conclusions and future prospects

Due to the complexity of proteomes, the current available human proteome database does not cover all the existing isoforms and mutations in human populations. In a clinical cohort study, if researchers are not aware of peptide allele specificity and use them for total protein quantification, it might affect the accuracy of protein abundance measurements. Researchers might observe discordance between the abundance levels of allele-specific and non-specific peptides from the same protein. This impact is especially important for SRM assay, as frequently only one or a few peptides per protein are used as surrogates to quantify the abundance level of target protein. Therefore, it is useful to integrate genomic data with proteomic profiling.

Facilitated by the decreasing cost of producing large volumes of DNA sequence data, it has become possible to perform genomic and proteomic analysis from the same samples. Therefore, making genetic data broadly accessible to everyday bench scientists who perform proteomic study is useful. Currently, the genome locations of MS identified peptides have been incorporated in SRM experiment libraries, e.g. PeptideAtlas. It will be desirable to also include the annotated genetic variants and their population distributions for the same peptide region in the libraries, which can be display on browsers and quickly scanned for allele specificity. Such a database should be useful for the proteomics community to select and/or exclude allele-specific peptides for SRM assay.

Acknowledgements

This work was supported by NIH grants to Dr. Michael Snyder (Microbiome grant: 3U54DK10255602S2- 02S2 and GTEx grant: 3U01HG00761101S1- 01S1).

Abbreviations

nsSNP	nonsynonymous single nucleotide polymorphisms
pQTL	protein quantitative trait loci
eQTL	expression quantitative trait loci
LCL	lymphoblastoid cell lines
LD	linkage disequilibrium

References

1. Deutsch EW, Lam H, Aebersold R. PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows. EMBO Rep. 2008; 9:429–434. [PubMed: 18451766]
2. Craig R, Cortens JP, Beavis RC. Open source system for analyzing, validating, and storing protein identification data. J Proteome Res. 2004; 3:1234–1242. [PubMed: 15595733]
3. Vizcaino JA, Cote R, Reisinger F, Foster JM, et al. A guide to the Proteomics Identifications Database proteomics data repository. Proteomics. 2009; 9:4276–4283. [PubMed: 19662629]
4. Shadforth I, Xu W, Crowther D, Bessant C. GAPP: a fully automated software for the confident identification of human peptides from tandem mass spectra. J Proteome Res. 2006; 5:2849–2852. [PubMed: 17022656]
5. Picotti P, Aebersold R. Selected reaction monitoring-based proteomics: workflows, potential, pitfalls and future directions. Nat Methods. 2012; 9:555–566. [PubMed: 22669653]
6. Schiess R, Wollscheid B, Aebersold R. Targeted proteomic strategy for clinical biomarker discovery. Mol Oncol. 2009; 3:33–44. [PubMed: 19383365]
7. Candille SI, Absher DM, Beleza S, Bauchet M, et al. Genome-wide association studies of quantitatively measured skin, hair, and eye pigmentation in four European populations. PLoS One. 2012; 7:e48294. [PubMed: 23118974]
8. Hovanessian AG, Justesen J. The human 2'-5' oligoadenylate synthetase family: unique interferon-inducible enzymes catalyzing 2'-5' instead of 3'-5' phosphodiester bond formation. Biochimie. 2007; 89:779–788. [PubMed: 17408844]
9. Wu L, Candille SI, Choi Y, Xie D, et al. Variation and genetic control of protein abundance in humans. Nature. 2013; 499:79–82. [PubMed: 23676674]
10. Lim JK, Lisco A, McDermott DH, Huynh L, et al. Genetic variation in OAS1 is a risk factor for initial infection with West Nile virus in man. PLoS Pathog. 2009; 5:e1000321. [PubMed: 19247438]
11. Metzker ML. Sequencing technologies - the next generation. Nat Rev Genet. 2010; 11:31–46. [PubMed: 19997069]
12. Hacia JG, Fan JB, Ryder O, Jin L, et al. Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. Nat Genet. 1999; 22:164–167. [PubMed: 10369258]
13. Veenstra TD, Conrads TP, Hood BL, Avellino AM, et al. Biomarkers: mining the biofluid proteome. Mol Cell Proteomics. 2005; 4:409–418. [PubMed: 15684407]
14. Overgaard M, Cangemi C, Jensen ML, Argraves WS, et al. Total and isoform-specific quantitative assessment of circulating Fibulin-1 using selected reaction monitoring mass spectrometry and time-resolved immunofluorometry. Proteomics Clin Appl. 2014
15. Simon R, Girod M, Fonbonne C, Salvador A, et al. Total ApoE and ApoE4 isoform assays in an Alzheimer's disease case-control study by targeted mass spectrometry (n=669): a pilot assay for methionine-containing proteotypic peptides. Mol Cell Proteomics. 2012; 11:1389–1403. [PubMed: 22918225]
16. Hatters DM, Peters-Libeu CA, Weisgraber KH. Apolipoprotein E structure: insights into function. Trends Biochem Sci. 2006; 31:445–454. [PubMed: 16820298]

17. Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, et al. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science. 1993; 261:921–923. [PubMed: 8346443]
18. Slooter AJ, de Knijff P, Hofman A, Cruts M, et al. Serum apolipoprotein E level is not increased in Alzheimer's disease: the Rotterdam study. Neurosci Lett. 1998; 248:21–24. [PubMed: 9665654]
19. Lehtimaki T, Pirttila T, Mehta PD, Wisniewski HM, et al. Apolipoprotein E (apoE) polymorphism and its influence on ApoE concentrations in the cerebrospinal fluid in Finnish patients with Alzheimer's disease. Hum Genet. 1995; 95:39–42. [PubMed: 7814023]
20. Panza F, Solfrizzi V, Colacicco AM, Basile AM, et al. Apolipoprotein E (APOE) polymorphism influences serum APOE levels in Alzheimer's disease patients and centenarians. Neuroreport. 2003; 14:605–608. [PubMed: 12657895]
21. Wang Q, Chaerkady R, Wu J, Hwang HJ, et al. Mutant proteins as cancer-specific biomarkers. Proc Natl Acad Sci U S A. 2011; 108:2444–2449. [PubMed: 21248225]
22. Kandoth C, McLellan MD, Vandin F, Ye K, et al. Mutational landscape and significance across 12 major cancer types. Nature. 2013; 502:333–339. [PubMed: 24132290]
23. Comprehensive molecular portraits of human breast tumours. Nature. 2012; 490:61–70. [PubMed: 23000897]
24. Khurana E, Fu Y, Colonna V, Mu XJ, et al. Integrative annotation of variants from 1092 humans: application to cancer genomics. Science. 2013; 342:1235587. [PubMed: 24092746]
25. Montgomery SB, Sammeth M, Gutierrez-Arcelus M, Lach RP, et al. Transcriptome genetics using second generation sequencing in a Caucasian population. Nature. 2010; 464:773–777. [PubMed: 20220756]
26. Pickrell JK, Marioni JC, Pai AA, Degner JF, et al. Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature. 2010; 464:768–772. [PubMed: 20220758]
27. Lappalainen T, Sammeth M, Friedlander MR, t Hoen PA, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013; 501:506–511. [PubMed: 24037378]
28. Civelek M, Lusis AJ. Systems genetics approaches to understand complex traits. Nat Rev Genet. 2014; 15:34–48. [PubMed: 24296534]
29. Schwanhausser B, Busse D, Li N, Dittmar G, et al. Global quantification of mammalian gene expression control. Nature. 2011; 473:337–342. [PubMed: 21593866]
30. Johansson A, Enroth S, Palmblad M, Deelder AM, et al. Identification of genetic variants influencing the human plasma proteome. Proc Natl Acad Sci U S A. 2013; 110:4673–4678. [PubMed: 23487758]

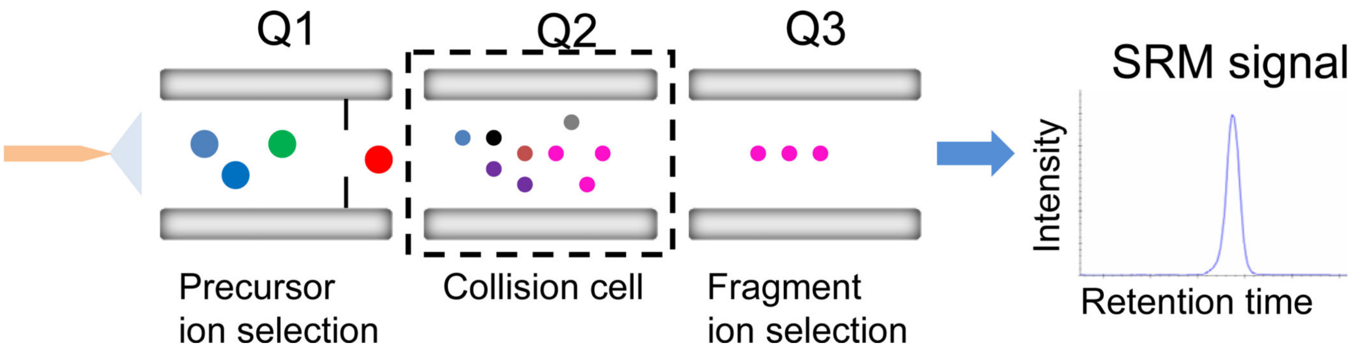

Figure 1. Schematic diagram of LC-SRM assay in a triple quadruple mass spectrometer

Peptides were separated by liquid chromatography. The eluent from LC column went into a triple quadruple mass spectrometer for on-line SRM monitoring. Precursor ion of a particular mass is selected in Q1. The selected precursor undergoes fragmentation by collision with neutral gas in Q2. Fragment ions with a specific m/z value are selected for detection in Q3. The intensity value (i.e. peak area under curve) represents quantitative peptide signal.

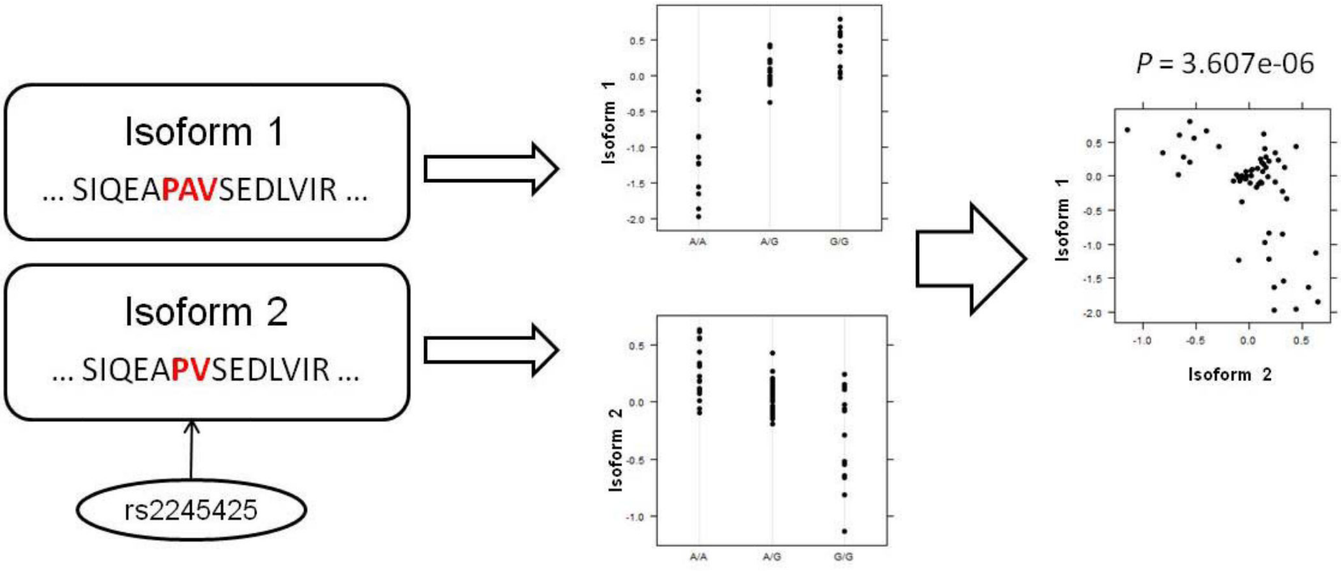

Figure 2. Correlation between TOR1AIP1 isoform levels and their association with rs2245425

Two TOR1AIP1 isoforms were identified and quantified by isobaric tag-based quantitative mass spectrometry method in 95 LCLs. Two isoforms are differed by only one amino acid deletion, which is altered by single nucleotide polymorphism at rs2245425. Peptides unique to each isoform are used to quantify the abundance levels of these two isoforms, respectively. Then the isoform levels are tested for SNP association, respectively. Both isoforms are significantly associated with rs2245425. In addition, they are anti-correlated with each other.