Detection of allele-specific gene expression on Next Generation Sequencing data

Vladan Mijatovic[1]✉, Luciano Xumerle[1], Alberto Ferrarini[2], Ilaria Iacobucci[3], Chiara Pighi[4], Antonio Mori[1], Chiara Zusi[1], Paola Prandini[1], Elisabetta Trabetti[1], Massimo Delledonne[1], Giovanni Martinelli[3], Albert Zamò[4], Pier F Pignatti[1], Giovanni Malerba[1]

[1]Department of Life and Reproductions Sciences, University of Verona, Verona, Italy
[2]Department of Biotechnology, University of Verona, Verona, Italy
[3]Department of Hematology and Oncological Sciences "L. and A. Seràgnoli", University of Bologna, Bologna, Italy
[4]Department of Pathology and Diagnostics, University of Verona, Verona

Motivation and Objectives

Many genetic variants mediate changes in gene expression. Some studies observed that variation of gene expression between alleles is common, and this variation may contribute to human variability of several traits (Lo et al, 2003, Main et al.,2009).

Next-generation sequencing (NGS) provides robust, comparable and highly informative expression profiling data (Shendure et al., 2008), and is rapidly replacing microarray methods in gene-expression (GE) studies (Wold et al., 2008, Wang et al., 2009). In contrast to microarrays, NGS expression profiling is based on sequencing and counting fragments of mRNA (Feng et al., 2010, van Iterson et al, 2009). The goal of our study is to develop a statistical framework aiming to measure and detect allele-specific GE differences from global GE experiments conducted using NGS technology.

Methods

We developed a statistical method for identification of allele-specific differential expression (ASDE). The method is based on the likelihood estimation of the observed data depending on the parameter Θ. Assuming that each polymorphism biallelic locus presents the alleles A1 and A2 we define Θ as A1/(A1+A2). Therefore Θ ranges from 0 to 1, and the expected value in the case of a fair expression of the two alleles is $\Theta = 0.5$ whilst values departing from 0.5 indicate that one allele is more expressed than the other. The likelihood function (L) is based on the binomial model and depends on the Θ value as follows: $L(\Theta) = k * [(\theta)A1 * (1-\Theta)A2]$ where k is constant of proportionality ($k>0$). The hypothesis of $\Theta \neq 0.5$ (i.e. ASDE) can be easily compared with the null hypothesis of $\Theta = 0.5$ (i.e. the two alleles, A1 and A2, present the same expression value) through a Likelihood Ratio Test (LRT): $LRT= -2 * \ln(L(\Theta=0.5) / L(\text{tested-}\Theta))$. The LRT from different samples can be summed up to a combined-LRT value that expresses the overall support of the model tested at Θ value that maximizes the likelihood function. Therefore the LRT can be used to test different ASDE models on the available data. The arbitrary threshold of LRT > 600 was used to detect ASDE loci.

NGS expression data have been obtained using a pipeline (quality control, alignment, SNP detection and read count) of computer programs that has been developed in our laboratory.

The following software have been used: bowtie (Langmead B et al.,2012) and samtools (Li et al., 2009). The reference sequence of the human genome GRCh37 was used.

Only heterozygous single nucleotide polymorphisms (SNPs) were selected for the following analysis, defined as SNPs with a coverage of at least 10 reads for each allele.

LRT was then applied to the data results of each sample.

Currently we performed the analysis on a total of 7 mantle cell lymphoma libraries (MCL)(Pighi et al., 2011). One-hundred (100) base-pair (bp) sequence paired-end reads were generated for each sample using an Illumina sequencer Hi-seq-1000. The average coverage was 10.4.

Results and Discussion

On average, the MCL cohort (7 samples) contained nearly 70,000 heterozygous sites. Preliminary results suggested 501 ASDE loci of which 470 showed a $0< \Theta <0.05$. We did not observed ASDE loci for $0.20< \Theta <0.80$.

We plan to estimate a reliable threshold of LRT across the entire trancriptome of several samples by simulation studies. We shall also study in more detail if the suggested ASDE loci showing a $\Theta < 0.05$ (or a $\Theta > 0.95$) are true ASDE loci or NGS artifacts.

The method will be extended to compare the Θ values (ASDE status) among groups of individuals (i.e. cases versus controls). The molecular analysis will be extended to additional samples including leukemia (Iacobucci et al., 2012), heart and skeletal muscle cells as well as lymphoblastoid cell libraries of individuals suffering from Autism Spectrum Disorders (ASD)(Prandini et al.,2012) . We developed a method able to detect ASDE loci from global gene expression NSG data. One of the features of this method is that it can easily measure the degree of ASDE through the parameter Θ. The method may also be applied to cancer research because an apparent ASDE locus might underlie the expression of a reduced amount of mutated cancer cells.

In conclusion we are developing a method for integration of information on allele variation with gene expression. This could increase our knowledge of hereditary factors involved in regulatory systems of gene expression.

References

1. Feng L, Liu H, Liu Y, et al. (2010) Power of deep sequencing and agilent microarray for gene expression profiling study. Mol Biotechnol 45:101. doi: 10.1007/s12033-010-9249-6
2. Iacobucci I, Ferrarini A, Sazzini M, et al. (2012) Application of the whole-transcriptome shotgun sequencing approach to the study of Philadelphia-positive acute lymphoblastic leukemia . Blood Cancer J. 2(3): e61. doi: 10.1038/bcj.2012.6
3. Lo HS, Wang Z, Hu Y, et al. (2003) Allelic variation in gene expression is common in the human genome. Genome Res.13(8):1855. doi: 10.1101/gr.1006603
4. Main BJ, Bickel RD, McIntyre LM, et al. (2009) Allele-specific expression assays using Solexa. BMC Genomics. 10:422. doi: 10.1186/1471-2164-10-422
5. Pighi C, Gu TL, Dalai I, et al. (2011) Phospho-proteomic analysis of mantle cell lymphoma cells suggests a pro-survival role of B-cell receptor signaling. Cell Oncol (Dordr). 34(2):141. doi: 10.1007/s13402-011-0019-7
6. Prandini P, Pasquali A, Malerba G, et al. (2012) The association of rs4307059 and rs35678 markers with autism spectrum disorders is replicated in Italian families. Psychiatr Genet. 22(4):177
7. Shendure J, Ji H (2008) Next-generation DNA sequencing. Nat Biotechnol 26:1135. doi:10.1038/nbt1486
8. van Iterson M, 't Hoen PA, Pedotti P, et al. (2009) Relative power and sample size analysis on gene expression profiling data. BMC Genomics 10:439. doi:10.1186/1471-2164-10-439
9. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nature Rev. Genet. 10:57. doi:10.1038/nrg2484
10. Wold B, Myers RM (2008) Sequence census methods for functional genomics. Nature Methods 5:19. doi:10.1038/nmeth1157