Workshop: Inference of allele specific expression levels from RNA-Seq data

Sahar Al Seesi and Ion Măndoiu
Computer Science and Engineering
University of Connecticut
Storrs, CT, USA
{*sahar, ion*}*@engr.uconn.edu*

Most current methods for estimating gene/isoform expression levels from high-throughput whole transcriptome sequencing (RNA-Seq) data rely on mapping the reads to a reference genome and/or transcriptome and do not consider the difference between the two parental alleles (diploid transcriptome). The diploid transcriptome can be easily inferred when a diploid genome is available, as in recent studies of cis- and trans-regulation [8] and parent-of-origin effects [5] that use hybrids of inbred species or strains. However, reconstructing the diploid genome of human subjects remains a difficult task [3]. Hence, existing studies of allele-specific gene expression rely on simple alleles coverage analysis for heterozygous Single Nucleotide Polymorphic (SNP) sites within transcripts. Such approaches typically do not allow inference of allele-specific expression of individual gene isoforms, result in less robust estimates since they use only RNA-Seq reads that overlap heterozygous SNP sites, and are affected by systematic read mapping biases toward reference alleles [1][6].

In this work, we integrate a recent method for SNV detection and genotyping from RNA-Seq data [4] with the scalable haplotype reconstruction algorithm of [2] and a diploid version of the Expectation Maximization (EM) algorithm for isoform expression estimation of [9] into a pipeline for estimation of allele-specific isoform expression levels. Our pipeline does not require genome sequencing data, but can incorporate such data when available. Inferring the two haplotypes and re-mapping the reads against the diploid transcriptome resolves the above mentioned bias towards reference alleles, while the EM model improves inference accuracy by using all reads, including those that map to more than one isoform, incorporating additional sources of disambiguation information such as the distribution of RNA-Seq fragment lengths, and correcting biases introduced by library preparation and sequencing protocols.

Preliminary results show the ability of the proposed pipeline to accurately infer allele specific isoform expression levels for synthetic hybrids with varying levels of heterozygosity, generated by pooling whole brain RNA-Seq reads of different mouse strains studied as part of the Sanger Institute Mouse Genome Project [7].

ACKNOWLEDGMENT

This project was supported by in part by awards IIS-0546457 and IIS-0916948 from NSF, and Agriculture and Food Research Initiative Competitive Grant no. 2011-67016-30331 from the USDA National Institute of Food and Agriculture.

REFERENCES

[1] J.F. Degner, J.C. Marioni, A.A. Pai, J.K. Pickrell, E. Nkadori, Y. Gilad and J.K. Pritchard, Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data. Bioinformatics, 25(24):3207-3212, 2009.

[2] J. Duitama, T. Huebsch, G. McEwen, E. Suk, and M.R. Hoehe, ReFHap: A Reliable and fast algorithm for Single Individual Haplotyping, BCB '10: Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology, 160-169, 2010.

[3] J. Duitama, G.K. McEwen, T. Huebsch, S. Palczewski, S. Schulz, K. Verstrepen, E-K Suk and M.R. Hoehe, Fosmid-based whole genome haplotyping of a HapMap trio child: evaluation of Single Individual Haplotyping techniques, Nucleic Acids Research, to appear, 2012.

[4] J. Duitama and P.K. Srivastava and I.I. Mandoiu, Towards accurate detection and genotyping of expressed variants from Whole Transcriptome Sequencing data, BMC Genomics, to appear, 2012.

[5] C. Gregg, J. Zhang, J.E. Butler, D. Haig, and C. Dulac, Sex-specific parent-of-origin allelic expression in the mouse brain. Science 239:682-685, 2010.

[6] G.A. Heap, J.H.M. Yang, K. Downes, B.C. Healy, et al. Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. Human Molecular Genetics, 19(1):122134, 2010.

[7] T.M. Keane, L. Goodstadt, P. Danecek, et al. Mouse genomic variation and its effect on phenotypes and gene regulation, Nature, 477(7364):289-294, 2011.

[8] C.J. McManus, J.D. Coolon, M.O. Duff, J. Eipper-Mains, B.R. Graveley, and P.J. Wittkopp, Regulatory divergence in Drosophila revealed by mRNA-seq. Genome Research, 20:816-825, 2010.

[9] M. Nicolae, S. Mangul, I.I. Mandoiu, A. Zelikovsky, Estimation of alternative splicing isoform frequencies from RNA-Seq data, Algorithms for Molecular Biology 6:9, 2011.

978-1-4673-1321-6/12/$31.00