Inference of transcriptional regulation in cancer

X. Shirley Liu*
Dana-Farber Cancer Institute/Harvard University, Boston, MA, USA – xsliu.dfci@gmail.com

Peng Jiang[a] (pjiang@jimmy.harvard.edu),

Matthew L. Freedman[bcd] (mfreedman@partners.org),

Jun S. Liu[e] (jliu@stat.harvard.edu),

X. Shirley Liu[a§]

[a]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute,
Harvard School of Public Health, Boston, MA 02115

[b]Department of Medical Oncology, [c]Center for Cancer Genome Discovery, Dana-Farber
Cancer Institute, Boston, MA 02115

[d]Medical and Population Genetics Program, Broad Institute of Harvard and MIT, Cambridge,
MA 02142

[e]Department of Statistics, Harvard University, Cambridge, MA 02138

Despite the rapid accumulation of tumor profiling data and transcription factor (TF) ChIP-seq profiles, efforts integrating TF binding with the tumor profiling data to gain insight into how TFs regulate tumor gene expression are limited. To systematically search for cancer-associated TFs, we comprehensively integrated 686 ENCODE ChIP-seq profiles representing 147 TFs with 7857 TCGA tumor data in 21 cancer types. For efficient and accurate inference on gene regulatory rules across a large number and variety of datasets, we developed an algorithm Rabit (Regression analysis with background integration). In each tumor sample, Rabit tests whether the TF target genes from ChIP-seq show strong differential regulation after controlling for background effect from copy number alteration (CNA) and DNA methylation. When multiple ChIP-seq profiles are available for a TF, Rabit prioritizes the most relevant ChIP-seq in each tumor. In each cancer type, Rabit further tests whether the TF expression and somatic mutation variations are correlated with differential expression patterns of its target genes across tumors. Our predicted TF impact on tumor gene expression is highly consistent with the knowledge from cancer related gene databases, and reveals many novel aspects of transcriptional regulation in tumor progression. We also applied Rabit on RNA binding protein motifs and found some alternative splicing factors could affect tumor-specific gene expression by binding to target gene 3’UTR regions. Thus, Rabit (http://rabit.dfci.harvard.edu) is a general platform for predicting the oncogenic role of gene expression regulators.

Keywords: First keyword:Frisch-Waugh-Lovell forward feature selection;
Second keyword: Linear regression; Third keyword: Inference on tumor expression
regulators; Fourth keyword: Tumor molecular profiling