A quantitative framework of integrating multi-modal cancer genomic data

C.H. Yeang
Institute of Statistical Science, Academia Sinica, 128 Academia Road, Section 2, Taipei, Taiwan, chyeang@stat.sinica.edu.tw

Cancer cells harbor simultaneous alterations at genomic, transcriptomic, proteomic, and epigenomic levels. To gain a comprehensive picture, multi-modal measurements have been applied to tumors. We developed a quantitative method to integrate cancer molecular alteration data in the same modeling framework. The goal is to identify the statistical/causal associations that link molecular alterations on DNA (sequence mutations or variations, copy number variations, DNA methylations) to variations of gene (mRNA, microRNA and protein) expressions. The outputs are “association modules” consisting of the upstream molecular alterations (effectors), downstream gene expressions (targets), and regulators that mediate the associations from effectors to targets. We applied the method to the integrated datasets of NCI-60 cell lines and TCGA GBM and successfully identified and verified the putative causal links that have strong prognostic effects. Furthermore, we reconstructed the association modules from the lung adenocarsinoma dataset of a cohort of Taiwanese non-smoking females, and verified those modules with external datasets of East Asian and Western patients. Strikingly, the prognostic effects of the inferred modules are pronounced in East Asian or East Asian female datasets alone. The results echo ethnic specificity of molecular mechanisms in cancer and question the validity of Caucasian-dominated datasets (such as TCGA) in other ethnic groups.

An integrated analysis of molecular aberrations in NCI-60 cell lines [1]

Cancer cells typically harbor a large number of aberrations at genetic and/or epigenetic levels, such as sequence mutations, DNA copy number alterations, DNA methylations and and nucleosome modifications. Yet only a small fraction of these molecular aberrations can cause malignancy, metastasis or other clinical phenotypes. A central question in cancer genomics is to separate these ``drivers'' from a large number of ``passenger'' molecular aberrations. This problem is generally difficult for the mechanisms connecting the drivers and the consequential phenotypes are complicated and remain at best partially known. Therefore, we substituted the clinical phenotypes with molecular phenotypes of gene

expressions and proposed a “layered modeling approach” to identify the molecular aberrations that modulate mRNA, protein or microRNA expressions. Resembling step-wise forward regression, it sequentially incorporated covariates (molecular aberrations) into the model to explain mRNA expressions. However, since direct associations between DNA molecular aberrations and gene expressions may over-fit the data and produce non biologically sensical outcomes, we prioritized the selection of variables according to their levels of complexity and uncertainty regarding the possible mechanisms. Layer 1 models associate gene expressions with molecular aberrations on the same loci. Layer 2 models associate expressions with aberrations on different loci but have known mechanistic links. Layer 3 models associate expressions with nonlocal aberrations which have unknown mechanistic links. We first applied this modeling framework to the integrated datasets of NCI-60 cancer cell lines including de novo sequence mutations on selected genes, copy number variations, DNA methylations, mRNA, microRNA and protein expressions. The analysis outcomes discovered/reaffirmed several prominent links: (1)Protein expressions are generally consistent with mRNA expressions. (2)Several gene expressions are modulated by composite local aberrations. For instance, CDKN2A expressions are repressed by either frame-shift mutations or DNA methylations. (3)Amplification of chromosome 6q in leukemia elevates the expression of MYB, and the downstream targets of MYB on other chromosomes are up-regulated accordingly (4)Amplification of chromosome 3p and hypomethylation of PAX3 together elevate MITF expression in melanoma, which up-regulates the downstream targets of MITF. (5)Mutations of TP53 are negatively associated with its direct target genes.

Deciphering causal and statistical relations of molecular aberrations and gene expressions in NCI-60 cell lines [2]

The layered approach establishes logistic regression models for each gene expression profile separately. In a biological system, a driver molecular aberration may modulate the expressions of multiple genes together. Therefore, we extended the layered models to establish ``association modules'' covering the upstream DNA molecular aberrations and downstream effects on gene expressions. Each association module consists of three components: (1)effector DNA molecular aberrations, (2)downstream target genes whose

expression profiles are associated with effector molecular aberrations, (3)regulators (transcription factors or signaling proteins) that mediate the effects from effectors to targets. We proposed an algorithm to construct association modules from the layered models of individual genes. By applying the module-finding algorithm to the integrated datasets of NCI-60 cancer cell lines, we found that gene expressions were driven by diverse molecular aberrations including chromosomal segments' copy number variations, gene mutations and DNA methylations, microRNA expressions, and the expressions of transcription factors. We applied a series of in-silico validations to examine the enrichment of certain functional categories/pathways and regulatory binding motifs of passenger genes, as well as co-citations between effectors and targets according to prior publications. Furthermore, we identified the putative target genes of MYB amplifications in leukemia, silenced MYB expressions, and found 6 of 11 predicted MYB targets were down-regulated in an MYB-siRNA treated leukemia cell line.

Integrative characterization of molecular aberrations in glioblastoma genomes [3]

The ultimate purpose of integrative analysis is to facilitate diagnosis and treatments of tumors. To justify this purpose we applied the layered models and association module finding algorithm to the integrated genomic data of glioblastoma multiforme (GBM). GBM is the most common and malignant primary brain tumor in adults. We identified the association modules from the integrated GBM data generated by the Cancer Genome Atlas (TCGA). Furthermore, the inferred association modules were validated by six tests using external information and datasets of central nervous system tumors: (1)indication of prognostic effects among patients, (2)coherence of target gene expressions, (3)retention of effector-target associations in external datasets, (4)recurrence of effector molecular aberrations in GBM, (5)functional enrichment of target genes, and (6)co-citations between effectors and targets. Modules associated with well-known molecular aberrations of GBM -- such as chromosome 7 amplifications, chromosome 10 deletions, EGFR and NF1 mutations -- pass the majority of the validation tests. Furthermore, several modules associated with less well-reported molecular aberrations -- such as chromosome 11 CNVs, CD40 and PLXNB1 methylations -- are also validated by external information. In particular, modules constituting

trans-acting effects with chromosome 11 CNVs and cis-acting effects with chromosome 10 CNVs manifest strong negative and positive associations with survival times in brain tumors. Functional and survival analyses indicate that immune/inflammatory responses and epithelial-mesenchymal transitions are among the most important determining processes of prognosis. Finally, certain molecular aberrations uniquely recur in GBM but are relatively rare in non-GBM glioma cells. These results justify the utility of an integrative analysis on cancer genomes and provide testable characterizations of driver aberration events in GBM.

Putative prognosis effectors in lung adenocarcinoma are ethnic and gender specific [4]

Lung adenocarcinoma possesses distinct patterns of *EGFR/KRAS* mutations between East Asian and Western, male and female patients. However, beyond the well-known *EGFR/KRAS* distinction, gender and ethnic specific molecular aberrations and their effects on prognosis remain largely unexplored. Association modules capture the dependency of an effector molecular aberration and target gene expressions. We established association modules from the copy number variation (CNV), DNA methylation and mRNA expression data of a Taiwanese female cohort. The inferred modules were validated in four external datasets of East Asian and Caucasian patients by examining the coherence of the target gene expressions and their associations with prognostic outcomes. Modules 1 (*cis*-acting effects with chromosome 7 CNV) and 3 (DNA methylations of *UBIAD1* and *VAV1*) possessed significantly negative associations with survival times among two East Asian patient cohorts. Module 2 (*cis*-acting effects with chromosome 18 CNV) possessed significantly negative associations with survival times among the East Asian female subpopulation alone. By examining the genomic locations and functions of the target genes, we identified several putative effectors of the two *cis*-acting CNV modules: *RAC1*, *EGFR*, *CDK5* and *RALBP1*. Furthermore, module 3 targets were enriched with genes involved in cell proliferation and division and hence were consistent with the negative associations with survival times. We demonstrated that association modules in lung adenocarcinoma with significant links of prognostic outcomes were ethnic and/or gender specific. This discovery has profound implications in diagnosis and treatment of lung adenocarcinoma and echoes the fundamental principles of the personalized medicine paradigm.

1. C.H. Yeang. (2010) An integrated analysis of molecular aberrations in NCI-60 cell lines. *BMC Bioinformatics* 11:495.

2. S.D. Li, T. Tagami, Y.F. Ho, and C.H. Yeang. (2011) Deciphering causal and statistical relations of molecular aberrations and gene expressions in NCI-60 cell lines. *BMC Systems Biology* 5:186.

3. N. Sintupisut, P.L. Liu, C.H. Yeang. (2013) An integrative characterization of recurrent molecular aberrations in glioblastoma genomes. *Nucleic Acids Research* 41(19):8803-8821.

4. A. Woolston, N. Sintupisut, T.P. Lu, L.C. Lai, M.H. Tsai, E.Y. Chuang, C.H. Yeang. (2015) Putative effectors for prognosis in lung adenocarcinoma are ethnic and gender specific. *Oncotarget* 6(23):19483-19499.