論 文 の 内 容 の 要 旨

論文題目 Recognition of Genes and Regulatory Elements Interactions by Multi-omics and Network Analysis

（がん細胞多層オーミクス統合ネットワーク解析による遺伝子発現制御領域塩基変異の機能解析）

氏　　名　セリーワッタナウート　サラン

Introduction (General)

Despite the colossal efforts of several genome-wide studies have identified a large number of genomic regions and mutations associated with hereditary diseases and cancers, in most of the case, how the detected associations collectively realize the phonotypes still remain elusive. There are several potential reasons hindering our understanding but the principal causes are; 1) genes could work in concerts of many complex systems to give rise to even more complex traits. Therefore, focusing investigations on a single gene or in a single organ may not fully explain the function of the gene; 2) in addition to the functions, regulations of the gene also play no less importance roles in the genes' behavior. Indeed, some large-scale studies, such as ENCODE and Roadmap, have described the first overview of the regulatory landscape of human genes, however our understandings on the comprehensive pictures of the regulations, which are diverse and dependent on the tissues and environments, are far from perfect. To address these issues, this thesis consists of two chapters where I attempt to: I) elucidate how non-coding regions might regulate their downstream coding counterparts by combination of short and long read sequencing and multi-omics analysis in cancer cell line setting II). Explore The Cancer Genome Atlas (TCGA) for large scale and systemic detection of both single and multi-omics interactions.

Chapter I: Identification of potential regulatory mutations using multi-omics analysis and haplotyping of lung adenocarcinoma cell lines

Background

Advances in high throughput sequencing technologies have enabled detailed profiling of various cancer genomes, resulted in large scale projects such as those in The Cancer Genome Atlas project (TCGA) or The International Cancer Genome Consortium (ICGC), these projects each contains massive libraries of somatic mutations in both coding and non-coding regions. While interpretation of mutations in coding regions are extensively studied and many key driver mutations have been identified and successfully utilized in anti-cancer treatments, those in non-coding regions remain largely elusive. Recent studies have shown that mutations in non-coding region such as those in *TERT* promoter region found in melanomas could also act as driver mutations and thus proven to be not less important than their coding counterparts. In this chapter, I intend probe into the mode of genes regulatory interactions by elucidating the transcriptional consequences of non-coding somatic variants (SNVs) in 23 lung adenocarcinoma (LUAD) cell lines by combining array of sample matching WGS, RNA-seq and ChIP-seqs with allele-resolution mutation phasing provided by 10x GemCode Synthetic long read platform.

Material and Methods

WGS, RNA-seq, TSS-seq and ChIP-seq, including Pol-II, H3K4me1, H3K4me3, H3K9me3, H3K27me3, H3K36me3, H3K27Ac and H3K9/K14Ac, of 23 LUAD cell lines were mapped to UCSC's human reference genome hg38. Regulatory mutations were SNVs in peaks that are within 50kb up and down stream of TSS. Synthetic long reads from 10x GemCode of 23 LUAD cell lines are available in WES plus regulome bait. 10x GemCode data are handled by 10x LongRanger software for linked-read analysis up to Molecular Identifier (MI) assignment. Haplotype assembly is done based on extending reads with the same MI combinations in the polyploid manner.

Regulatory mutations and transcripts that could be phased and exhibit biases in both allele expressions were considered. Transcript allele expression biases were calculated from RNA-seq and regulatory allele biases were calculated from any of the ChIP-seq, both are normalized by the alleles' relative sequencing depths in WGS (Fig. 1).

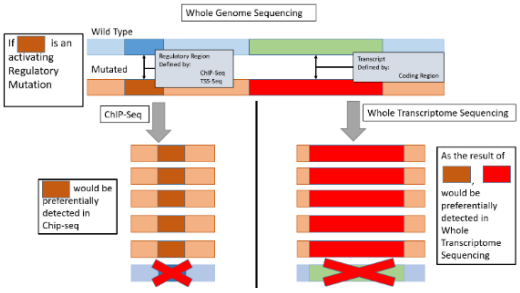

Figure 1: Allele Bias Detection in phased pairs.

Results

Transcripts allele bias were observed in 7,915 transcripts in total (596 per cell line), with 137 regulatory mutations phased to 146 RefSeq transcripts, these drops were from the limitation of phasing with WES and Regulome regions. From 137 SNVs, 104 were predicted overlapped with CpG regions or disrupted the binding motifs of transcription factors by TRANSFAC database or by ENCODE ChIP-seq.

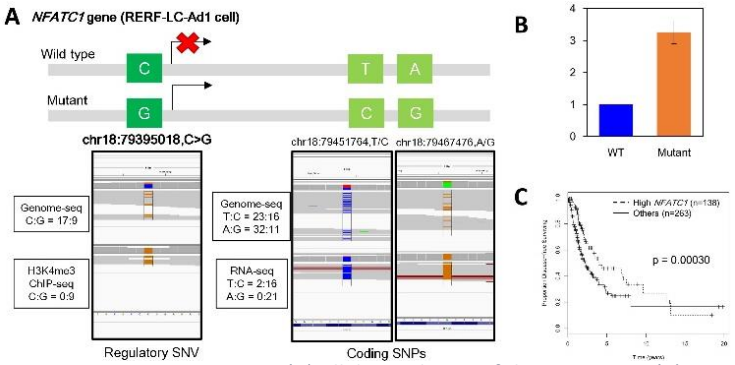

Figure 2: NFATC1 mutations. (A) Allele resolution of the mutations; (B) Luciferase Assay showed 3 times activities for Mutant Motif. (C) High NFATC1 donors have poorer prognosis (p=0.0003).

One of the mutations, chr18:79395018C>G located in regulatory region of *NFATC1*, a transcription factors, (Fig. 2A) in RERF-LC-Ad1 cell line and was predicted to generated novel transcription factor binding site for ETS family gene, this was biologically validated with Luciferase Assay finding the mutant motif to be 3 times more active than the wildtype (Fig. 2B), the mutation was also validated by Sanger Sequencing. Survival analysis in TCGA-LUAD dataset suggested that 31 genes with regulatory mutations, have significant impacts on patient survival, *NFATC1* was associated with better prognosis in overall survival but worse prognosis in disease free survival (Fig. 2C).

These results provided a proof of concept groundwork that integrative studies of multi-omics analysis could be used to probe into the elusive non-coding regions and provided functional annotation of the regulatory elements found in those regions. To further follow-up of these results, I intend to broaden my analysis to large scale public projects in TCGA/ICGC with multi-omics dataset to systemically analyses both single and multi-omics interactions in pan-cancer settings.

Chapter II: Pan-cancer Multi-omics Networks analysis in The Cancer Genome Atlas

Background

In the previous chapter, I have demonstrated that inactions between regulatory elements and their downstream counterparts could be probed by integrative analysis of multi-omics studies. However, cell lines do not

necessary retain the essences of their originating cancers, hindering phenotypes integrations and analysis approach was based on the single gene resolution in small number of samples from single organ of origin, prohibiting system-wide view and thus the power to explore complex or subtle interactions. I intend to address this shortcoming by advancing into the massive libraries of clinical samples deposited in TCGA database, which composes of multiple large projects conducted for diverse cancer species. Indeed, each project contains a large number of samples with matching transcriptome and DNA methylation datasets, which is ideal to reveal the association between the gene expression regulations and their consequential transcriptomes.

To fully utilize such rich datasets, I intend to explore the genes and methylations interactions by means of integrative and uniform network analysis of the two omics. Gene network analysis is a powerful tool frequently used in interpreting and understanding the vast and sophisticated biological systems and many tools have been developed, under various circumstances and assumptions, to fill this crucial role. However, biological networks are complex with various mode of interactions and regulations, thus models that able to integrate features from different omics uniformly and smoothly into the same standard would require too extraordinary parameters to be practical and generalizable. Instead of finding the perfect parameters for any single analysis, I utilized rank analysis based on non-parametric approach to constructed the networks. The genotypic interactions would be visible from the 2-omics networks and by inspecting how the networks behave under various phenotypes, their functional relevancies could be determined.

Material and Methods

Gene Expression and Methylation profiles of 8 TCGA projects with sample matching datasets were retrieve from the ICGC data releases. Total of 4,116 samples have the matching RNA-seq and Methylation array dataset. Duplicated specimens from the same donor were treated separately (Table 1). Genes with low or missing expressions were removed and Methylation sites that were outside of +-10kb windows from any transcript TSS were removed. Total of 15,666 genes and 12,835 methylation sites were picked for analysis.

Projects	All Donors	Applicable Donors	Applicable Specimens
BRCA-US	1,093	1,012	1,130
CESC-US	307	242	246
STAD-US	443	415	415
HNSC-US	528	480	502
LUSC-US	502	424	432
LUAD-US	518	473	496
COAD-US	459	420	464
SKCM-US	470	427	431
Total	4,320	3,893	4,116

Table 1: TCGA Projects donor and specimens.

Results

To uniformly integrate genes and methylation sites, each sample was ranked according to measurements in each feature, the rankings were then treated as homogenized measurements and used in networking of the two omics. The first goal was to group features, regardless of omics, where the samples change their ranking in a synchronized manner. This was done with hierarchical clustering with UPGMA algorithm with co-variance/variance as distances. Clusters were identified by cutting the UPGMA tree with homogenized co-variance distances determined by k-sample Anderson-Darling test (AD test), these clusters were then treated as fundamental functional units. *GAPDH*/Glycolysis unit is shown in Figure 3A as example, with rank synchronization in Figure 3B. The glycolysis pathway enzymes were captured along with *GAPDH* in this unit, supporting the

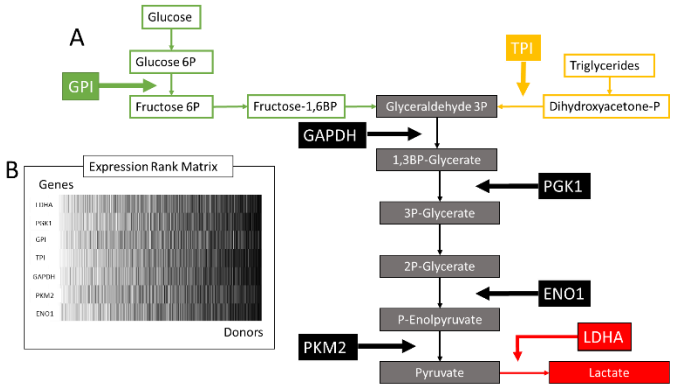

Figure 3: GAPDH unit with related Glycolysis pathway enzyme in (A) and Rank Synchronization of the enzymes in (B)

robustness and notably *LDHA* was also captured with this unit indicating the cancer cells' preferences of lactate productions from glycolysis tern "Warburg Effects".

From 28,501 genes and CpG sites, 4,358 units were made. Activities of each unit in each sample were determined based on co-variance terms of that particular sample and each sample was treated as discrete phenotype but with loose grouping from cancer project of origins. Unit with similar phenotypic activities were then linked together into networks of interactions.

From 4,358 units 654 networks were made, these networks were then undergone manual functional analysis guided by GOs terms and known gene sets enrichment to looked for biologically meaningful interactions. Both known and potentially novel genotypic were captured in these networks.

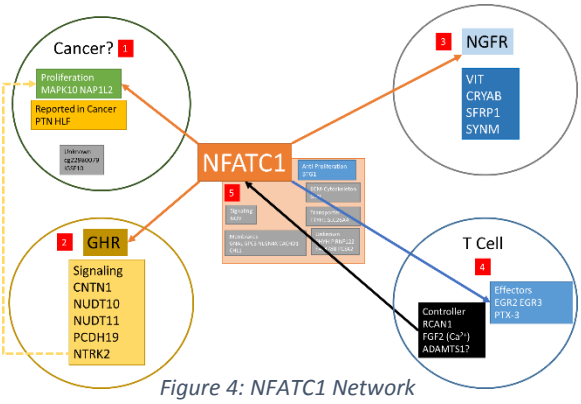

Figure 4: NFATC1 Network

For instance, *NFATC1*, a transcription factors shown to be biologically and clinically significant in previous chapter, was shown to bundled with its known upstream regulators and downstream T Cell activators (Figure 4; T Cell), supporting its association with better overall patient survival. Moreover, pro-proliferative signals outside of T cell such as (Figure 4) *GHR*, *NGFR* were also included, suggestive of its proliferative effects outside of T cell and most importantly, its interaction with various known oncogenic proliferative signals (Figure 4; Cancer) might play a part in driving tumor growth and relapses, reflecting *NFATC1* associations with worse patient disease free survival and thus faster relapse.

These genetic networks were not only effective at single gene resolution level, in functional level, the networks also were able to capture interactions between Cell cycle, DNA Replication, Repairs and Methylations in the chains of closely intertwined networks and most intriguingly in those 654 networks 29 were mix networks containing both genes and CpG sites and one of them was strongly active in melanoma.

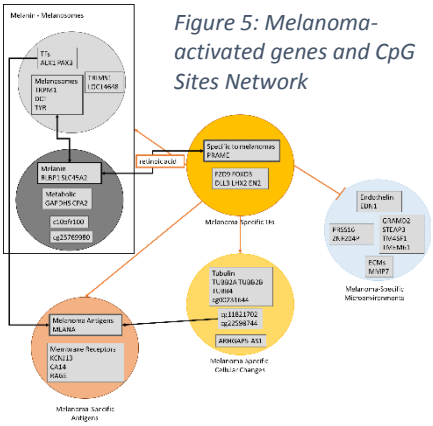

Figure 5: Melanoma-activated genes and CpG Sites Network

In this melanoma-activated network (Figure 5), 4 CpG sites were bundled together with 34 genes which many of them were functionally specific and highly express in melanomas and those 4 CpG sites were annotated to be close to those melanoma specific genes' TSS or bodies and were shallowly methylated in melanoma donors, implying their roles as repressors of those genes and integrating CpG sites and genes together in a single system.

These networks provided novel and comprehensive platforms to the understanding of the biological interactions they were describing and highlight the benefits and plausibility of adopting rank analysis as the bases for integrating and analysis of multi-omics networks.

Summary

In this series of works, I explored the interactions between genes and regulatory elements in both context specific and large-scale database by multi-omics and network analysis. In LUAD, I illustrated *cis*- regulatory effects of non-coding mutations and in TCGA pan-cancer dataset, I bundled the features, regardless of omics, into networks of similar phenotypic behaviors. These networks were then shown to capture both known and novel 2-omics interactions and presenting rank analysis as an approach for multi-omics network analysis.