Identification of Potential Regulatory Elements by Multi-omics Analysis and Haplotype Phasing in Multiple Lung Adenocarcinoma Cell Lines

Graduate School of Frontier Sciences

Department of Computational Biology and Medical Sciences

Laboratory of Systems Genomics (Suzuki-Laboratory)

Sereewattanawoot Sarun ID 47-156432 Master's degree 2017 March

Keywords: Lung Cancer, Adenocarcinoma, Haplotype Phasing, Multi-omics

Academic Adviser: Professor Suzuki Yutaka

Background

In this study, I intend to elucidate the transcriptional consequences of the somatic mutations (SNVs), which are frequently identified in potential regulatory regions in cancer genomes. While frequently identified, their functional relevance still remains elusive. For this purpose, I first attempted to identify the allelic background of SNVs in potential regulatory regions with regards to SNPs/SNVs of their corresponding transcripts. Then, with the presence of SNVs in the regulatory regions, I selected the genes which showed allelic bias in their transcript levels. In this study, I used a series of cell lines derived from Lung Adenocarcinoma as the model cases. Lung Adenocarcinoma is one of the most prominent and extensively studied cancer both in clinical specimen and cell lines. The most important driver genes, which are now being used or developed into therapeutics targets, such as EGFR, ALK fusion or KRAS, are known. However, even for this intensively studied cancer, genetic background in carcinogenesis of a large population (24.4%) of cases remains unknown. Moreover, the carcinogenic effects of even the most powerful driver genes themselves could not be solely responsible for entire carcinogenesis process. Previous work at our lab (Suzuki et al. 2014 NUCLEIC ACID RES) identified a large number of mutations, both in coding and regulatory regions in 26 Lung Adenocarcinoma-derived cell lines. In addition, the multi-omics data, such as histone modifications and transcriptomes, from the same material have been collected. With these dataset, I attempt to identify SNVs which lead to aberrant transcriptional regulations, thereby contributing to carcinogenesis.

Material and Methods

The multi-omics dataset I processed and used in this study for each cell line are as follows:

Whole genome sequencing, RNA-seq, TSS-seq and ChIP-seq data sets, ChIP-seq data include Pol-II, H3K4me1, H3K4me3, H3K9me3, H3K27me3, H3K36me3, H3K27Ac and H3K9/K14Ac, all available from previous work (Suzuki et al 2014). Whole genome sequence, RNA-seq and ChIP-seq were re-mapped to UCSC's human reference genome hg38. Regulatory regions were defined by TSS-seq and ChIP-seq. Bias in allele expression were calculated from changes in variant frequency across two-omics domains.

Synthetic long reads from 10x GemCode were available in Whole genome sequence for 2 cell lines and Agilent SureSelect V5 with regulatome regions for 23 others. The 10x GemCode data were handled by 10x LongRanger software for linked-read analysis. Final Phasing was done with my own phasing schemes.

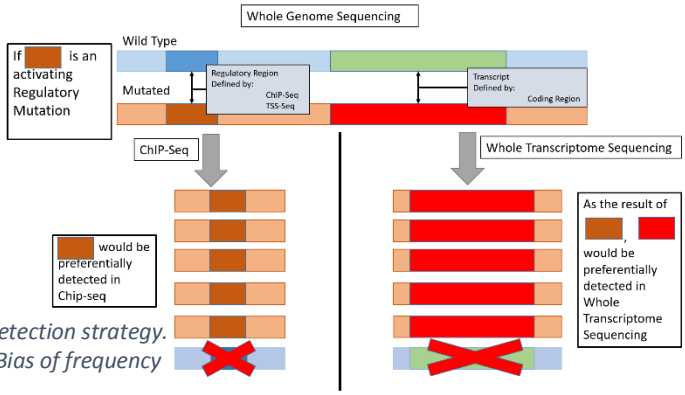

Figure 1 show allele expression/regulatory activity bias detection strategy. Variant frequency were calculated for all omics dataset. Bias of frequency in Chip to WGS and RNA to WGS were noted

Results

With multi-omics datasets, on average I identified a total of 265 RefSeq genes for which the SNVs in the regulatory genes are likely to invoke the aberrant transcriptional regulation. Those genes were identified by examining the bias towards one variant of the SNPs/SNVs in both regulatory activity and transcript expression per cell line. Average number of the observed cases for the SNVs of the biased regulatory activity as represented by the Chip-seq data per cell line was 516. On the other hand biased transcriptions of their regulating genes were also identified by 582 SNPs/SNVs in the transcript region per cell line. The most frequently recurring gene among cell lines was NBPF1 in19 out of the 23 cell line, although biological relevance of this gene to lung adenocarcinoma is still unknown. The most biologically interesting genes were the KMT2C and MAP2K3 genes. Both of the genes were aberrant in 15 cell lines. The KMT2C gene encodes a Lysine Methyltransferase 2C, a crucial enzyme in epigenetic control. Deleterious mutations in this gene were reported to decrease cancer growth rate [Dawkins JB1, 2016]. The MAP2K3 gene is a kinase family gene. Accumulation of its active form was reported to activate RAS pathway [NCBI, 2017], the most commonly mutated pathway in lung adenocarcinoma.

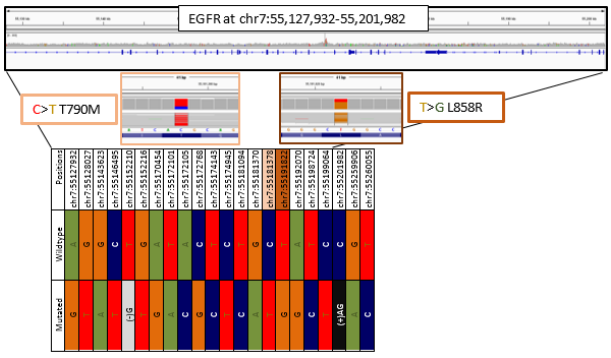

Figure 2 shows phasing of EGFR T790M and L858R mutation along with their 20 Neighboring SNPs.

For further connecting the regulatory SNVs with their regulating transcripts, that are represented by the SNPs and SNVs within them, I developed the phasing scheme by modifying the default LongRanger pipeline. The average, phased continuous region was 50kb (max 989kb) containing 13 SNPs (max 496) including previously reported phased somatic mutations such as the EGFR L858R and T790M mutations. The average detected ploidy was 3.3 comparable with pervious report [COSMIC, 2017]. The genes which were known to have copy number alteration or amplification such as ERBB2 were also detected to have multiple distinct haplotypes, providing a supporting evidence for the precise detection in my scheme. Phasing of the allele imbalance expression genes revealed regulatory aberrations in the 58 in 59 RefSeq genes on average per cell line. In this list the CDKN1A and BRCA1genes were included, which are well-known genes for their cancerous alternations but those mutations are mostly reported as those causing the gene loss.

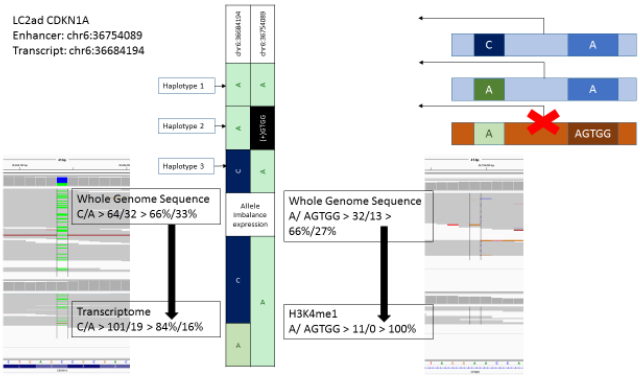

Figure 3 shows incorporation of allele imbalance analysis and phasing analysis of CDKN1A in LC2ad cell line. Regulatory variant "A" was shown to had complete dominance over its insertion counterpart

Discussion

The results of this study provide a unique approach in studying cancer genome and mutation, shedding light on the importance of the mutation in the regulatory elements. Many novel potential functionally relevant regulatory SNVs were detected. Those genes included two genes which are known to relate to carcinogenesis and the gene which was firstly suggested in the present study. The biological validation of these detected SNVs is still needed. Regarding the methodology, I believe that difficulties in evaluating the effects which the regulatory elements have on gene expression could be overcome as described in the present study. By utilizing bias in allele expression detected by RNA-seq as the consequence, I identified the genes having the regulatory aberrations. However, not every functional regulatory element could be studied in this way. The results from the phasing also showed that the lack of information in allele configuration could lead to wrong interpretation of the imbalance variants. Nevertheless, I believe this strategy could have paved the way for further identification of much more important regulatory elements and somatic mutation in regulatory region. Moreover, the results from phasing still lack supporting evidences and must be validate by other means such as by physical long read (such as MinION).

With more cross validation and control data, I believed that major improvement in phasing efficiency and accuracy should be achieved. Indeed, I found that phasing provides many crucial and unique information for SNPs/SNVs phasing and copy number alterations detection. Continuous improvement in this new field would one day turn phasing into new standard in future cancer genome sequencing.