Doctoral Thesis

博士論文

# 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）

セリーワッタナウート　サラン

# Contents

# General Introduction

Cancer is the second leading cause of death worldwide. It is responsible for 1 in 6 documented deaths with an estimated 9.6 million deaths worldwide in 2018 (Bray et al., 2018; Ferlay et al., 2019). Though diverse in tissues of origin and presentations, the processes of cancer development, or carcinogenesis, follow a common multistep transformation from normal cells into cancer cells. These transformations are caused by genetic and epigenetic changes in the cells (Weinberg, 2013). To better understand the biology behind them, many large-scale international studies have been conducted to elucidate these genetic (International Cancer Genome Consortium et al., 2010; The Cancer Genome Atlas Research et al., 2013) and epigenetic changes (Bradley E. Bernstein et al., 2010; Davis et al., 2018) with great success. Highly successful projects, such as lung adenocarcinoma in The Cancer Genome Atlas project (TCGA), have identified genomic driver mutations in genes such as *EGFR* and *ALK* (The Cancer Genome Atlas Research Network, 2014). Discoveries of these mutations have subsequently led to the development of many successful anticancer drugs currently employed in treatment regimens (The American Cancer Society, 2019).

Given the complexities observed in essential and fundamental processes such as DNA replication and the cell cycle (Cooper, 2000) or glycolysis (Berg, Tymoczko, & Stryer, 2002), it is natural to suppose that genes work together in concert to give rise to complex functions and diverse phenotypes. Supporting this notion, pivotal cancer driver genes, such as *EGFR*, have been shown to exert various functions (Sigismund, Avanzato, & Lanzetti, 2018) and act as key regulators in various pathways (Wee & Wang, 2017) in a wide variety of cancers. Additionally, the regulation of genes, both at the genetic level such as by transcription factors or at the epigenetic level by DNA or histone methylations (Klemm, Shipony, & Greenleaf, 2019; Vihervaara, Duarte, & Lis, 2018), and even by noncoding RNAs (Olive, Jiang, & He, 2010; Peng & Croce, 2016), has been shown to also play an important role in influencing gene functions. Indeed, epigenetics projects, such as ENCODE and Roadmap, have highlighted important epigenetic regions and have described

the regulatory landscapes of human genes both in normal and malignant settings, which has opened up the frontier of epigenetic research and has attracted great attention. Despite great efforts, our understanding of a comprehensive picture of genetic-epigenetic interactions and regulations remains far from perfect.

Coding sequences only account for a small number of mutations in cancer genomes (The Cancer Genome Atlas Research et al., 2013). Mutations in cis-regulatory elements, promoters and enhancer regions have also been shown to be as important, if not more, than their coding counterparts (Khurana et al., 2016). This is especially evident in melanoma, where mutations in the *TERT* promoter region have been identified as some of the important driver mutations (Huang et al., 2013; Vinagre et al., 2013). Moreover, the roles of promoter and enhancer regions in cell fate determination and development are becoming clearer (Cantone & Fisher, 2013), especially in hematopoietic cell linages (Cullen, Mayle, Rossi, & Goodell, 2014). These findings solidified the notion that the interactions between genetic and epigenetic elements give rise to specific phenotypes. To fully study these interactions, an integrative network study from both sides proves to be an interesting approach to answer how genetics and epigenetics interact with each other to translate genotypic information to phenotypes.

In this thesis, I intend to elucidate the interaction between genomes and their regulatory epigenomes by multi-omics and integrative network analysis and propose how these insights could help bridge genotypes and phenotypes. This thesis consists of two chapters, where I attempt to I) Elucidate how noncoding regions might regulate their downstream coding counterparts by combining short and long read sequencing and multi-omics analysis in a cancer cell line setting and II) Explore TCGA for large-scale and systemic network detection of both single and multi-omics interactions.

# Chapter I: Identification of Potential Regulatory Elements by Multi-Omics Analysis and Haplotype Phasing in Lung Adenocarcinoma Cell Lines

## Introduction

Lung cancer is one of the most widely studied cancers. The adenocarcinoma subtype comprises half of all lung cancer cases in both smokers and nonsmokers (Collisson et al., 2014; Dela Cruz, Tanoue, & Matthay, 2011). Many environmental and lifestyle risk factors, mainly air pollution and smoking, have been identified. Despite the reduction in risk factor exposure and lifestyle changes, the lung cancer incidence rates are increasing, especially in nonsmokers. This invariably suggests unknown carcinogenic causes (Dela Cruz et al., 2011). This observation is in contrast to the squamous cell subtype, which has been declining along with the reduction in smoking and other risk factors. This unique feature has placed lung adenocarcinoma as the focus of many research groups.

Recurrent genomic driver mutations in *EGFR* and *KRAS* and *ALK-RET* fusions have been documented. Several successful anticancer drugs targeting these genes have been developed (Chan & Hughes, 2014; TheAmericanCancerSociety, 2019). Nevertheless, the driver genes in more than a third (38%) of the cases are yet unknown (Collisson et al., 2014), posing a challenge in curative treatments. The known driver mutations were identified and interpreted only from the coding region of the genome, which accounts for less than 5% of the entire genome. Noncoding regions have not yet been fully investigated, and recent studies have elucidated that these regions have no less importance than coding regions (Vinagre et al., 2013) (Huang et al., 2013) (Chan & Hughes, 2014) and could also harbor biologically relevant mutations. Many novel therapeutic options might be discovered from mutations in noncoding regions.

The benefits of focusing on noncoding regions are not limited to cases with unknown driver mutations (38%). Resistant cancer clones rapidly develop in almost all of the target therapy cases, resulting in remission and relapse. The drug-resistance mutation T790M in *EGFR* (Ma, Wei, & Song, 2011; Yun et al., 2008) or point mutations in *KRAS* or *PIK3CA* in *EGFR*-resistant clones are well-known examples (Del Re et al., 2017; S. Li et al., 2014). Despite substantial efforts, the causative mutations of a large number of relapse cases are still unknown; thus, countering drug resistances has not been widely successful. These difficulties indicate the diversity in tumor responses to each treatment, which could arise from both coding and noncoding mutation backgrounds (Holohan, Van Schaeybroeck, Longley, & Johnston, 2013). A more comprehensive understanding of the interaction between coding and noncoding regions might hold the key in combating drug resistance.

In this chapter, using sequencing data from lung adenocarcinoma cell lines, I will focus on identifying the functional regulatory mutations that are shown to have transcriptional effects that are detectable in their downstream transcripts by multi-omics analysis and haplotype phasing.

## Multi-omics Analysis

Next-generation sequencing (NGS) has enabled detailed studies of genomic mutations by short read high-throughput data generation (Behjati & Tarpey, 2013; Chmielecki & Meyerson, 2014). Whole genome sequencing (WGS) and whole exome sequencing (WES) are the main strategies in outlining the mutation landscapes of cancers. WES focuses exclusively on the coding regions, enabling a detailed and cost-effective approach in studying coding region mutations. WGS provides a more complete landscapes of both coding and noncoding regions, albeit at lower resolution at the same sequencing cost. Both approaches are being utilized extensively according to the expertise and interests of the researchers.

Following the central dogma of DNA-RNA-protein information transfer (Crick, 1970), RNAs and their regulation are indispensable in determining cell functions and

phenotypes. After the postgenomic era, cell type-specific marker genes have been identified and utilized (Redwine & Evans, 2002). More recently, tissue-specific proteins, such as surfactants in the lung epithelium or keratin in the epidermal epithelium, and their RNA expression were found to be correlated in the Genotype-Tissue Expression (GTEx) project (Lonsdale et al., 2013; Melé et al., 2015; Sonawane et al., 2017). Quantitative dynamic responses of mRNA levels to external stress have been shown to dominate the housekeeping functions of the cell as well (Jovanovic et al., 2015). Therefore, an RNA expression level could be used as a quantitative surrogate for cell functions and phenotypes. These quantitative assessments are obtained by NGS-based whole transcriptome sequencing (RNA-seq).

An equally important aspect of understanding mRNA expression is analyzing their regulatory regions. These regions are termed promoters and enhancers. Their presence and functions are represented by histone modifications and DNA methylation. The epigenetic status of histone modifications can be identified by chromatin immunoprecipitation followed by sequencing (ChIP-seq) and targeting the histone modifications of interest (Table 1). This technique has been highly successful at capturing promoter and enhancer regions and has been widely adopted in the ENCODE and Roadmap databases (B. E. Bernstein et al., 2012; Davis et al., 2018).

*Table 1 ChIP-seq markers and their functions*

| MARKER | EFFECTS | REGION |
|---|---|---|
| **POLYMERASE-II** | Transcriptional Activation | RNA-Polymerase |
| **H3K4ME1** | Transcriptional Activation | Enhancer |
| **H3K4ME3** | Transcriptional Activation | Promoter |
| **H3K9ME3** | Repression | Heterochromatin and repetitive elements |
| **H3K9/14AC** | Transcriptional Activation | Promoter Preference |
| **H3K27AC** | Transcriptional Activation | Enhancer |
| **H3K27ME3** | Repression | Repressive Domain and Silencing |
| **H3K36ME3** | Transcriptional Elongation | Transcribed Regions |

## Bridging Regulatory and Coding Regions with Haplotype Phasing

One of the limitations of the current NGS technologies is the lack of allele haplotype information due to the reliance on short-read sequencing. This limitation hinders the integration between the regulatory mutations and their transcripts. Without prior information of the allele configurations, it is not known whether the pairs are in *cis*- or *trans*- (or mixed in regions with copy number aberrations (CNA). The effects of the mutations could not be confidently evaluated. The effects of heterozygous somatic mutations are limited to only one of the alleles. Disregarding allele configurations would lead to incorrect conclusions of the effects of regulatory mutations.

To overcome this limitation, "Long Read Sequencing" (Pollard, Gurdasani, Mentzer, Porter, & Sandhu, 2018) technologies were developed. One of the approaches is called "Synthetic Long Read Sequencing". The GemCode system, developed by 10x Genomics (Zheng et al., 2016), is based on the reconstruction of haplotype alleles from uniquely barcoded short read sequences by a conventional NGS short read sequencer. The reconstructions are achieved by capturing high molecular weight DNA (HMW-DNA) inside confined oil droplets with unique gel-embedded barcodes (GEMs). After hybridization and extension, each unique barcode, collectively called a "molecular identifier" or MI, is attached to the DNA molecule in the same droplets. The barcoded reads from HMW-DNA are sequenced by a conventional Illumina short-read sequencer. The origin of the individual HMW-DNA is identified by computational re-assembly of the reads with the same MIs. These connecting reads are termed "Linked Reads" and play a key role in this "Synthetic Long Read" technology.

Another approach in long read sequencing is nanopore sequencing, which was developed by Oxford Nanopore Technologies (ONT). Instead of relying on NGS for sequencing, the Nanopore-based MinION sequencer conducts direct sequencing of the long DNA strand (Jain, Olsen, Paten, & Akeson, 2016), called "Physical Long Read Sequencing". Read lengths often reach tens of kilobases, although the sequencing accuracy is still far less than that of NGS sequencing.

With the advent of long read sequencing technologies, haplotype phasing between two or more variants, often more than tens of kilobases apart, has now become a reality, enabling the association of regulatory mutations in promoter and enhancer regions with their transcript counterparts.

In this chapter, by combining multi-omics and haplotype phasing analysis, I aimed to document allele-based transcriptional effects of regulatory mutations, elucidate how the interactions between regulatory mutations and their transcriptional counterparts could be investigated and demonstrate their potential roles in cancer biology.

# Material and Methods

## Cell lines

Twenty-three human lung adenocarcinoma cell lines were cultured in RPMI medium (RPMI 1640, Nissui), Dulbecco's modified Eagle's medium (Nissui) or Eagle's minimal essential medium (Nissui) with 10% FBS, MEM nonessential amino acid solution (SIGMA) and antibiotics (antibiotic-antimycotic, Gibco). The cells were cultured at 37°C and 5% CO2. Cell line information and COSMIC reported mutations are shown in Table 2.

*Table 2 Characteristics of the Cell Lines Used in This Study*

| Cell Line | Sex | Ethnicity | Distributor | Catalog Number | Average Ploidy | Mutation Reported by COSMIC |
|---|---|---|---|---|---|---|
| **A427** | Male | Caucasian | ATCC | HTB-53 | 3.13 | KRAS, MSI |
| **A549** | Male | Caucasian | ATCC | CCL-185 | 2.76 | KRAS, SMARCA4 |
| **ABC-1** | Male | Japanese | JCRB | JCRB0815 | 2.39 | TP53, ALK |
| **H322** | Unspecified | Caucasian | ATCC | CRL-5806 | 2.35 | ALK, ERBB2, TP53, BRCA1 |
| **H1299** | Male | Caucasian | ATCC | CRL-5803 | 4.75 | NRAS, SMARCA4, TP53, KMT2D |
| **H1648** | Male | African | ATCC | CRL-5882 | 2.44 | TP53, ARID1A, BRCA2 |
| **H1650** | Male | Caucasian | ATCC | CRL-5883 | 1.99 | EGFR, TP53, SMARCA4 |
| **H1703** | Male | Caucasian | ATCC | CRL-5889 | 2.32 | CDKN2A, TP53, ROS1, BRCA1 |
| **H1819** | Female | Caucasian | ATCC | CRL-5897 | - | - |
| **H1975** | Female | Unspecified | ATCC | CRL-5908 | 2.83 | EGFR, TP53, PIK3CA |
| **H2126** | Male | Caucasian | ATCC | CCL-256 | 3.24 | TP53, SMARCA4 |
| **H2228** | Female | Unspecified | ATCC | CRL-5935 | 3.74 | RET, ALK, KMT2C, TP53 |
| **H2347** | Female | Caucasian | ATCC | CRL-5942 | 3.76 | KRAS, ALK, TP53, NRAS |
| **II-18** | Unspecified | Japanese | RIKEN BRC | RCB2093 | - | - |
| **LC2ad** | Female | Japanese | RIKEN BRC | RCB0440 | 3.37 | RET, TP53, TET2 |
| **PC-9** | Unspecified | Japanese | RIKEN BRC | RCB4455 | - | - |

| | | | | | | |
|---|---|---|---|---|---|---|
| **PC-14** | Unspecified | Japanese | IBL | - | 3.14 | CDKN2A, CCND2, TP53, EGFR, KMT2S |
| **RERF-LC-Ad1** | Male | Japanese | JCRB | JCRB1020 | - | - |
| **RERF-LC-Ad2** | Male | Japanese | JCRB | JCRB1021 | - | - |
| **RERF-LC-KJ** | Male | Japanese | RIKEN BRC | RCB1313 | 2.72 | EGFR, TP53. BRCA2 |
| **RERF-LC-MS** | Unspecified | Japanese | JCRB | JCRB0081 | 4.33 | FGFR2, TP53 |
| **VMRC-LCD** | Male | Japanese | JCRB | JCRB0814 | 2.4 | ARID1A, TP53, KDM5A, MAP2K4 |
| **RERF-LC-OK** | Unspecified | Japanese | JCRB | JCRB0811 | - | - |

## Multi-omics datasets for each cell line

The FASTQ files from whole-genome sequencing; chromatin immunoprecipitation sequencing (ChIP-Seq) for H3K9me, H3K9/14ac, H3K4me3, H3K4me1, H3K36me3, H3K27me3, H3K27ac, RNA polymerase II and input DNA; whole transcriptome sequencing (RNA-seq) and transcriptional start site sequencing (TSS-Seq) for each cell line were retrieved (Suzuki et al., 2014). Annotations of the coding regions were obtained from the KERO database using the UCSC hg38 human genome reference (http://kero.hgc.jp/) (Suzuki et al., 2015). The sequencing and mapping statistics for whole genome sequencing, RNA-seq and input ChIP-seq are shown in Table 3, and each ChIP-seq antibody is shown in Table 4.

## SNPs/SNVs from whole genome sequencing data

The FASTQ files from whole genome sequencing of each cell line were mapped to the UCSC hg38 human genome reference (Speir et al., 2016) using BWA (H. Li & Durbin, 2009) (version 7.15) and the aln algorithm with default setting. PCR duplicates were then removed by SAMtools (H. Li et al., 2009) (version 1.18). SNPs/SNVs were called by

GATK (McKenna et al., 2010) (version 3.3) with default parameters. The SNPs/SNVs called by GATK and those with more than 5 supporting tags and greater than 5% variant frequency were selected. The variant frequencies were calculated by the SAMtools (v1.18) mpileup command with the default setting. (see Table 3 for details)

## Regulatory regions defined by ChIP-seq

ChIP-seq data for 7 histone modifications (H3K9me, H3K9/14Ac, H3K4me3, H3K4me1, H3K36me3, H3K27me3 and H3K27Ac) and polymerase-II were processed. The FASTQ files were remapped to the UCSC hg38 human genome reference using BWA (version 7.15) and the aln algorithm with default settings. PCR duplicates were then removed by SAMtools (version 1.18). Peaks were called by MACS2 (Zhang et al., 2008) broad-peak with default parameters with input DNA as a control. Peaks that were within 150 kb of the transcriptional start site according to TSS-seq data were treated as regulatory regions. If there were multiple transcriptional start sites, the closest transcriptional start site was selected. SNVs that fell within the peaks were then defined as regulatory SNVs. The number of regulatory SNVs was counted collectively. SNVs with multiple markers were counted multiple times and treated separately.

## Whole Transcriptome Sequencing

FASTQ files for RNA-seq were remapped to the UCSC hg38 human genome reference by GSNAP using default parameters. Splice sites and introns were provided by the KERO database. (see Table 3 for details)

## Transcriptional Start Site Sequencing

For transcriptional start site studies, data from 26 lung adenocarcinoma cell lines and 1 small airway epithelium cell line were compared and merged. The TSSs used were generated from the merged dataset. The promoter region for each gene was defined as the region 500 bp upstream to 500 bp downstream of the transcriptional start site clusters. The promoters were treated as regulatory regions.

## Background Germline Variant Filtering

SNPs/SNVs called by GATK in the whole genome sequencing that were located within the regulatory regions were filtered by NCBI's dbSNP v142 for benign germline variants.

## Synthetic long read library preparation by 10x GemCode

From 23 cell lines, high molecular weight DNA was extracted and quantified by the Qiagen MagAttract HMW kit according to the manufacturer's recommendations (10x Genomics, Qiagen #67653).

For each cell line, $1x10^6$ cells were suspended in 200 µl of PBS buffer, 20 µl of proteinase K mixture, 4 µl of RNase A and 150 µl of buffer AL. The samples were then incubated at 25°C for 30 minutes. Fifteen microliters of Qiagen MagAtrract suspension G was added to each sample along with 280 µl of buffer MB. The samples were mixed and incubated at 1400 rpm at 15–25°C for 3 minutes. To wash the beads, samples were placed on a magnetic rack for 1 minute, and the clear supernatant was discarded. The beads were removed from the magnetic rack, suspended in 700 µl of Buffer MW1, mixed and incubated at 1400 rpm at 15–25 °C for 1 minute. The samples were put on to the magnetic rack, and the procedure was repeated once. After Buffer MW1, samples were then washed twice with 700 µl of Buffer PE. Beads with Buffer PE were placed on the magnetic rack for 1 minute. The supernatant was removed on the magnetic rack, 700 µl of nuclease-free water was added and incubated for 60 seconds, the supernatant was discarded, and the processes were repeated once. After the beads were washed with Buffer MW1, PE and nuclease-free water twice, the beads were removed from the magnetic rack, and 150 µl of buffer AE was added to the bead pellets. The samples were mixed and incubated at 1400 rpm at 15–25 °C for 3 minutes. The samples were placed on the magnetic rack and incubated for 1 minute. The supernatant was transferred and stored at 4 °C for DNA quantification by a Qubit dsDNA HS Assay kit (Thermo Fisher Scientific) at a target concentration of 10-20 ng/µl.

For GemCode library preparation, partitioning was performed by GemCode Gel-Beads and Chip (10x Genomics). Indexing and library preparation were performed by the

GemCode library preparation kit (10x Genomics) according to the manufacturer's instructions. In brief, quantified high molecular weight DNA was further diluted by nuclease-free water to a concentration of 1 ng/μl, and 1.2 μl was used. The sample mix was prepared by adding 1.2 μl of diluted genomic DNA to the Master Mix, consisting of nuclease-free water, GemCode Reagent Mix, Primer Release Mix and GemCode Polymerase supplied in the GemCode Reagents Kt. The sample mix, gel beads and Partitioning Oil were applied onto the GemCode Chip. The GemCode Chip was loaded into the GemCode instrument.

Gel beads in emulsions (GEMs) were retrieved from the instrument according to the manufacturer's recommendation and transferred to a 96-well plate for a designated thermal cycling amplification. For post-cycling recovery, 1 μl of Additive 1 and 125 μl of Recovery Agent were added and mixed with each GEM according to the manufacturer's instructions. The aqueous solutions were transferred, and the Recovery Agent and Partitioning Oil were removed. The mixture of Recovery Agent and Partitioning Oil at the bottom was first removed by 135 μl of pipetting. The leftover was removed with DynaBeads MyOne SILANE beads and 0.6X SPRI solution on the GemCode magnetic rack. Beads were washed with Elution Buffer I (Elution Buffer, 10% Tween-20, Additive 2) with SPRI reagent twice and washed with Elution Buffer II (Elution Buffer, Additive 2) once.

The barcoded samples were subjected to library construction by shearing using the Covaris system. Fragmentation was performed with a target peak of 250 bp for whole exome and regulome sequencing and 800 bp for whole genome sequencing. End repair and A-tailing were performed by thermal cycling of the fragmented DNA with the End Repair and A-Tailing Buffer and Enzyme Mix supplied by the GemCode library preparation kit (10x Genomics). Products from end repair and A-tailing were ligated by thermal cycling with Adaptor Mix and DNA Ligase. Post ligation cleanup was performed by 0.8X SPRI solution on the GemCode magnetic rack. Sample indexing PCR with the P5 primer was conducted. The post-PCR cleanup was performed by 1.0X SPRI cleanup on the GemCode magnetic rack.

Target enrichment was performed using the Agilent SureSelectXT protocol with SureSelect V5 plus regulome baits according to the manufacturer's instructions (Agilent, 10x Genomics). See Figure 1 for a summarized workflow.

The FASTQ files were processed using the 10x Genomics LongRanger (version 1.3) pipeline with default setting together with the pre-called SNPs. (see Table 5 for details)

## Non-diploid phasing using 10x Genomics GemCode

The phasing of nondiploid genomes was not supported by the 10x Genomics GemCode LongRanger (version 1.3) pipeline; however, I deemed adaptation of the molecular index (MI), also called unique molecular identifiers (UMIs), in nondiploid genomes phasing a possibility. The approach was based on the exhaustive process of merging UMIs that overlapped at the same nucleotide variant together to reconstruct the extended haplotypes.

First, indexes of WGS-detected SNPs and 10x GemCode UMIs covering each of those SNPs were generated by cross referencing the VCF file of WGS SNPs called by GATK to the bam file of 10x GemCode LongRanger (version 1.3).

From those indexes, I exhaustively merged and extended the overlapping and compatible UMIs into a longer "Pre-Haplotype"; UMIs were deemed compatible if #1 at least one SNP position overlapped and the nucleotide variant matched and #2 there were no different nucleotide variants in any of the overlapped SNP positions. Incompatible UMIs that overlapped were not merged but were designated into their own distinct "Pre-Haplotypes" in the same phase blocks, and a UMI could be a member of more than one "Pre-Haplotype" if the combination allowed it. Only reads with mapping scores >20 and SNPs with scores >20 (in base substitution only, 10x GemCode bam file) were considered. These processes were repeated exhaustively until every UMI was considered.

Due to the random nature of the barcoding and shearing of 10x GemCode library preparations, these "Pre-Haplotypes" did not contain the entire lengths of the alleles. It was often found that in many SNPs, only one variant was covered by a UMI, and the other was

left isolated, thus prematurely stopping the extensions. This resulted in a phase block with multiple short and isolated haplotypes, which were not useful in the SNP-to-SNP linkage analysis. To address this, a second round of merging was performed inside each phase block with the goal of filling the gaps and connecting the "Pre-Haplotypes" so that each final haplotype now spanned the entire phase block.

The second merging was performed in a greedy manner. First, in each phase block, every "Pre-Haplotype" missing a position was determined by checking its SNPs against the full-length allele, and "Pre-Haplotypes" with no missing positions were considered complete and final. For those with missing positions, I searched for the most similar haplotype that could fill in the gaps from the other Pre-Haplotypes. Similarity was determined by the number of compatible SNPs subtracted by the number of incompatible SNPs, and 0 was set for nonoverlapping pairs. This process was repeated until the haplotype was complete with no missing position and no pre-haplotypes remaining. Only the final haplotypes were used in further analysis; see Figure 3 for graphic representation.

## Physical long-read sequencing by MinION

For MinION sequencing, H1975, LC2/ad, and RERF-LC-KJ cells were used.

High molecular weight DNAs were extracted in the same manner as described above. Library preparations were performed according to the manufacturer's instructions (Oxford Nanopore Technologies). In brief, extracted high molecular weight DNA was subjected to end repair and dA-tailing by the NEBNext End repair/dA-tailing module (E7546S, NEB). Purifications were performed using Agencourt AMPure XP beads (Beckman Coulter). Ligation and tethering were performed with NEBNext Blunt/TA Ligase Master Mix (M0367S, NEB) and Ligation Sequencing Kit SQK-LSK208 for 2D, SQK-LSK108 for 1D and SQK-LSK308 for $1D^2$ (Oxford Nanopore Technologies). The obtained libraries were purified by MyOne C1 beads (65001, Thermo Fisher Scientific). Sequencing was performed in 48-hour run mode by MinION Mk 1B with the SpotION Flow Cell (FLO-MIN106, R9.4 version for 2D; FLO-MIN107, R9.5 version for 1D and $1D^2$, Oxford Nanopore Technologies).

Base calling was performed by Metrichor. The FAST5 files were converted into FASTQ format with poretools (Loman and Quinlan 2014). FASTQ files were mapped to the UCSC hg38 human genome reference using BWA-MEM with ont2d settings for 2D reads (H1975, RERF-LC-KJ) and default settings for 1D and $1D^2$ reads (LC2/ad). Conversion to the bam format and sorting were performed by SAMtools (version 1.18). See also Figure 2 for the workflow.

## Validation of 10x GemCode Phase Block by MinION Physical Long Reads

Phased SNPs were checked for coverage with MinION reads with mapping quality scores > 10 and spanned more than one SNP position. Combinations of covered SNP configurations were then referenced with those reads. Because of the lack of single nucleotide resolutions of the MinION reads (90% sequence identity in 2D and 80% in 1D and $1D^2$ combined), phase blocks that had more than twice the number of supporting reads compared with the number of nonsupporting reads were considered evidenced by MinION sequencing.

## Functional Analysis of Regulatory Mutations

Involvements of the mutations in regulatory RNA binding sites were investigated by referencing the location of the regulatory mutations to FANTOM CAT lv3 robust lncRNA region (FANTOM_CAT.lv3_robust.all_lncRNA.bed.gz) (Hon et al., 2017) and FANTOM5 phase 1 and 2 permissive enhancer (human permissive enhancers phase 1 and 2.bed.gz) (Andersson et al., 2014). The regions were downloaded from the RIKEN database and then mapped to the USCS hg38 human genome by liftover (Kent et al., 2002).

Transcriptional factor (TF) binding sites in A549 cells were analyzed using 50 ChIP-seq targeting TFs, chromatin remodeling factors and RNA binding proteins in the A549 cell line deposited in the ENCODE database (Davis et al., 2018). Optical idr threshold peaks in the narraowPeak bed file of GRCh38 were used in the analysis. TF motif analysis was performed by searching reference and alternative sequences ±10 bp around the inquired motifs in the TRANSFAC database (2015.1) (Matys et al., 2006) using MATCH

(Kel et al., 2003). Hits with matrix similarity scores >0.95 were selected, and the results from alternative sequences were compared with the reference sequences. Graphical representation of the positional weight matrix of the biding consensus sites was created from TRANSFAC matrices using seqLogo R Library (Bembom, 2017).

## Luciferase Assay

pNL3.1 (#N1031, Promega) was selected as the vector, and pGL4.53 (#E5011, Promega) was selected as the control. Mutant and wild-type DNA fragments were inserted into the pNL3.1 vector by the Quick Ligation Protocol (M2200, New England Biolabs) using NheI-HF (R3131S, New England Biolabs) and HindII-HF (R3104S, New England Biolabs) according to the manufacturer's instructions (Table 6A). Transformation was performed using the 5 Minute Transformation Protocol (C2987H/C2987I) (New England Biolabs), and plasmids were purified by PureLink™ HiPure Plasmid Kits (K2100, Thermo Fisher Scientific) according to the instructions. Transfection was performed using the ViaFect Transfection Reagent (E4981, Promega) according to the manufacturer's instructions with a medium to final volume ratio of 4:1. Cells were assayed after 24 hours using the Nano-Glo Dual-Luciferase Reporter Assay System (N1610, Promega) according to the manufacturer's instructions with CentroXS3 LB960 (Berthold Technology) and a measurement time of 1 second for both ONE-Glo and NanoDLR.

## ChIP-qPCR

Chromatin immunoprecipitation was performed using 20 µl of ETS-1 (D8O8A) rabbit mAb (#14069, Cell Signaling Technology). After precipitation, quantitative real-time PCR was performed using Power SYBR Green PCR Master Mix (4367659, Applied Biosystems, Thermo Fisher Scientific) with previously reported control primers (RPS26) (Plotnik, Budka, Ferris, & Hollenhorst, 2014) and primers targeting ±100 bp of the motif region (Table 6B) on the 7900HT Fast Real-Time PCR System (Applied Biosystems). The qPCR products of Primer_F_2_123bp and Primer_R_shared were then subjected to Sanger sequencing on a 3730xl DNA Analyzer (Applied Biosystems) with their respective primer sets.

## Survival Analysis

RNA-seq v2 and clinical data of TCGA lung adenocarcinoma (TCGA-LUAD) donors were downloaded from the NCI Genomic Data Commons using TCGA-Assembler v2.0.1 (Zhu, Qiu, & Ji, 2014) (data accessed 2017/03/09). Normalized gene expression counts were log2 transformed and used in the analysis. Overall survival and disease-free survival duration were retrieved from follow-up data in clinical data files. High expression donors were defined as donors with expression z scores > 0.5; likewise, low expression donors were defined as donors with expression z scores < -0.5. Statistical significance was determined using the Kaplan-Meyer estimator with the log-rank test using the survival package in R with each group of donors as cases and others as controls.

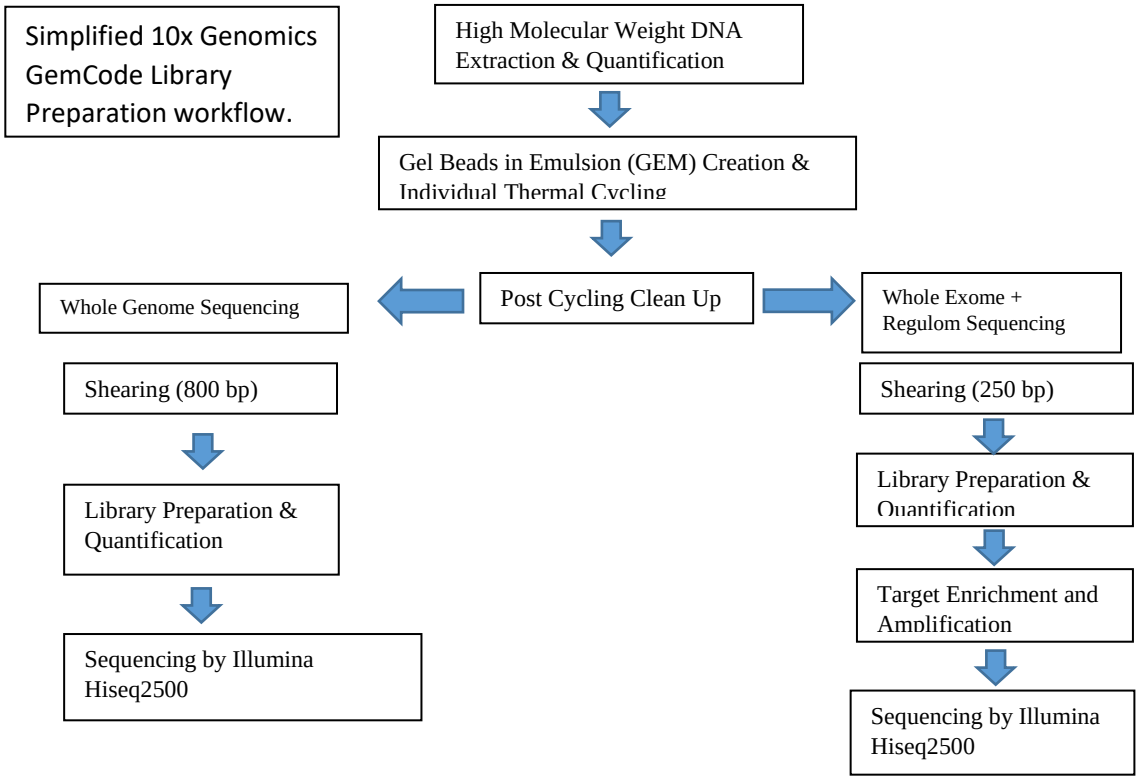



*Figure 1 Simplified workflow for the 10x GemCode Library preparation System (10x Genomics).*

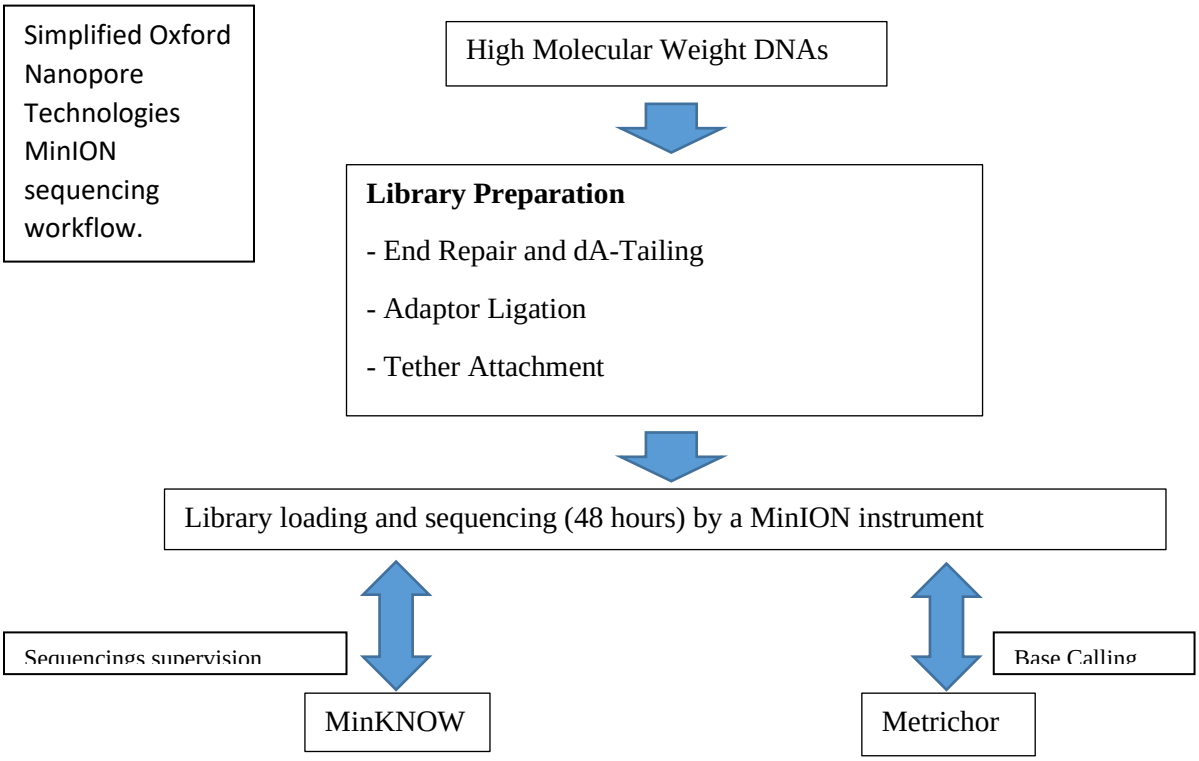



*Figure 2 Simplified workflow for MinION physical long read sequencing (Oxford Nanopore Technologies).*

*Table 3 Basic sequencing characteristics for whole genome sequencing, RNA-seq and Chip-seq background control*

| Cell line | Whole Genome Sequencing | | | Whole Transcriptome Sequencing | | Chip-seq Input Control | |
|---|---|---|---|---|---|---|---|
| | Mapped Read | % Mapped Read | Depth | Mapped Read | % Mapped Read | Mapped Read | % Mapped Read |
| **A427** | 1,084,672,075 | 94.0% | 34.62 | 95,046,694 | 97.0% | 58,870,145 | 97.0% |
| **A549** | 577,537,022 | 71.0% | 15.92 | 51,009,049 | 98.0% | 23,063,615 | 80.0% |
| **ABC-1** | 1,198,942,503 | 94.0% | 38.36 | 89,577,661 | 98.0% | 4,959,932 | 52.0% |
| **H322** | 921,462,662 | 95.0% | 29.13 | 128,407,549 | 97.0% | 5,186,262 | 46.0% |
| **H1299** | 930,092,532 | 95.0% | 29.93 | 121,767,233 | 96.0% | 11,053,640 | 93.0% |
| **H1648** | 1,303,832,736 | 90.0% | 40.78 | 86,409,901 | 98.0% | 18,636,861 | 96.0% |
| **H1650** | 1,093,147,187 | 96.0% | 34.98 | 66,205,127 | 98.0% | 107,477,951 | 96.0% |
| **H1703** | 1,035,232,011 | 87.0% | 31.94 | 190,122,574 | 97.0% | 25,836,885 | 82.0% |
| **H1819** | 1,197,312,856 | 92.0% | 38.13 | 180,743,242 | 98.0% | 47,573,722 | 95.0% |
| **H1975** | 1,056,952,131 | 94.0% | 33.37 | 76,888,082 | 98.0% | 36,642,876 | 97.0% |
| **H2126** | 668,355,912 | 88.0% | 21.31 | 106,874,132 | 98.0% | 11,285,585 | 72.0% |
| **H2228** | 855,605,013 | 90.0% | 27.36 | 129,887,384 | 96.0% | 41,236,999 | 92.0% |
| **H2347** | 983,271,902 | 85.0% | 31.62 | 119,783,099 | 95.0% | 55,967,654 | 97.0% |
| **II-18** | 890,312,525 | 84.0% | 26.75 | 153,260,052 | 96.0% | 10,210,751 | 58.0% |
| **LC2ad** | 1,400,218,662 | 93.0% | 44.78 | 103,957,725 | 97.0% | 2,909,093 | 24.0% |
| **PC-9** | 1,326,079,008 | 94.0% | 42.40 | 121,730,782 | 96.0% | 3,845,359 | 29.0% |
| **PC-14** | 979,278,917 | 97.0% | 31.33 | 82,194,427 | 98.0% | 12,005,835 | 51.0% |
| **RERF-LC-Ad1** | 1,265,604,463 | 95.0% | 40.60 | 128,209,153 | 97.0% | 22,741,126 | 75.0% |
| **RERF-LC-Ad2** | 1,284,008,781 | 95.0% | 41.10 | 103,865,898 | 97.0% | 32,887,224 | 77.0% |
| **RERF-LC-KJ** | 1,113,739,330 | 95.0% | 35.59 | 138,119,858 | 97.0% | 8,693,898 | 59.0% |
| **RERF-LC-MS** | 1,319,743,295 | 93.0% | 42.30 | 119,134,144 | 97.0% | 12,701,625 | 66.0% |
| **VMRC-LCD** | 1,394,724,167 | 93.0% | 44.64 | 109,941,326 | 98.0% | 10,201,434 | 50.0% |
| **RERF-LC-OK** | 684,830,042 | 86.0% | 21.02 | 78,730,703 | 97.0% | 19,353,474 | 97.0% |
| **Average** | 1,068,041,554 | 91.1% | 33.82 | 112,255,035 | 97.1% | 25,362,693 | 73.1% |

*Table 4 Sequencing statistics for individual Chip-seq antibodies for each cell line.*

| Cell line | Polymerase-II | | H3K4me1 | | H3K4me3 | | H3K9me3 | | H3K9_14Ac | | H3K27Ac | | H3K27me3 | | H3K36me3 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Mapped Read | % Mapped | Mapped Read | % Mapped | Mapped Read | % Mapped | Mapped Read | % Mapped | Mapped Read | % Mapped | Mapped Read | % Mapped | Mapped Read | % Mapped | Mapped Read | % Mapped |
| **A427** | 19,919,326 | 95% | 35,907,915 | 96% | 40,834,430 | 96% | 16,099,060 | 98% | 16,267,399 | 98% | 45,852,658 | 98% | 13,751,977 | 98% | 14,511,308 | 98% |
| **A549** | 42,205,011 | 98% | 24,557,168 | 98% | 28,481,237 | 98% | 16,398,873 | 93% | 25,914,697 | 98% | 13,996,665 | 98% | 24,826,664 | 97% | 33,981,484 | 98% |
| **ABC-1** | 30,106,498 | 97% | 23,448,072 | 96% | 32,348,875 | 97% | 24,035,880 | 95% | 15,643,097 | 98% | 28,957,286 | 96% | 25,120,033 | 96% | 42,806,924 | 97% |
| **H322** | 20,592,481 | 95% | 19,565,669 | 98% | 29,291,795 | 97% | 48,815,268 | 95% | 22,819,233 | 97% | 39,589,006 | 97% | 28,757,036 | 97% | 23,973,241 | 98% |
| **H1299** | 15,517,082 | 92% | 15,500,143 | 91% | 6,845,054 | 89% | 23,174,212 | 94% | 26,347,198 | 98% | 25,902,379 | 98% | 11,715,556 | 92% | 7,919,777 | 93% |
| **H1648** | 42,151,483 | 96% | 29,424,483 | 97% | 26,008,969 | 96% | 31,893,831 | 96% | 20,124,185 | 97% | 32,995,085 | 95% | 16,764,970 | 96% | 34,563,616 | 97% |
| **H1650** | 34,512,016 | 95% | 25,494,598 | 96% | 38,951,570 | 95% | 49,297,255 | 82% | 21,953,905 | 98% | 42,526,121 | 97% | 21,855,198 | 82% | 21,719,937 | 98% |
| **H1703** | 33,931,810 | 91% | 34,798,266 | 98% | 17,985,220 | 91% | 33,066,974 | 97% | 27,913,705 | 98% | 31,111,917 | 98% | 18,727,226 | 98% | 21,500,912 | 98% |
| **H1819** | 14,617,601 | 97% | 35,015,007 | 97% | 17,947,000 | 96% | 38,744,549 | 93% | 22,921,204 | 95% | 23,747,865 | 97% | 19,250,082 | 91% | 27,777,531 | 94% |
| **H1975** | 34,211,588 | 98% | 33,758,149 | 98% | 18,206,422 | 95% | 29,297,788 | 96% | 25,467,485 | 98% | 22,661,866 | 97% | 16,865,773 | 97% | 29,859,308 | 97% |
| **H2126** | 27,096,982 | 96% | 13,390,733 | 98% | 16,108,148 | 96% | 18,365,403 | 95% | 34,921,354 | 98% | 14,662,278 | 97% | 27,126,917 | 97% | 37,864,976 | 97% |
| **H2228** | 34,065,433 | 97% | 40,528,026 | 98% | 18,474,115 | 96% | 45,956,295 | 97% | 26,180,133 | 96% | 33,453,676 | 97% | 26,892,026 | 97% | 24,160,581 | 98% |
| **H2347** | 36,045,314 | 97% | 30,548,297 | 83% | 24,573,340 | 96% | 44,156,118 | 97% | 32,312,697 | 97% | 36,153,407 | 83% | 20,204,256 | 96% | 39,717,531 | 97% |
| **II-18** | 33,022,666 | 96% | 23,130,969 | 95% | 22,114,574 | 97% | 20,440,344 | 93% | 13,650,439 | 98% | 41,775,051 | 97% | 38,796,482 | 98% | 33,065,234 | 96% |
| **LC2ad** | 32,914,384 | 95% | 54,113,092 | 98% | 29,315,441 | 96% | 11,690,048 | 86% | 14,170,753 | 92% | 35,788,989 | 98% | 40,973,180 | 97% | 24,914,911 | 95% |
| **PC-9** | 36,269,970 | 97% | 24,034,872 | 98% | 32,779,453 | 95% | 25,383,329 | 89% | 13,592,966 | 98% | 20,925,733 | 97% | 61,498,760 | 97% | 15,533,061 | 96% |
| **PC-14** | 43,079,306 | 91% | 36,150,087 | 98% | 29,881,364 | 92% | 42,868,733 | 97% | 14,871,279 | 96% | 37,398,511 | 97% | 36,399,198 | 96% | 35,516,283 | 98% |
| **RERF-LC-Ad1** | 31,866,960 | 96% | 42,742,931 | 97% | 29,130,354 | 92% | 29,272,804 | 92% | 25,338,362 | 97% | 26,551,483 | 97% | 13,240,117 | 96% | 25,750,673 | 97% |
| **RERF-LC-Ad2** | 32,740,273 | 94% | 44,180,544 | 98% | 32,501,541 | 93% | 22,862,593 | 87% | 13,817,685 | 98% | 33,367,124 | 96% | 13,408,679 | 95% | 28,652,285 | 96% |
| **RERF-LC-KJ** | 29,962,594 | 94% | 26,907,433 | 97% | 43,186,043 | 95% | 27,254,351 | 93% | 20,979,344 | 97% | 29,811,965 | 98% | 23,546,066 | 93% | 38,833,258 | 95% |
| **RERF-LC-MS** | 21,367,869 | 97% | 20,275,585 | 96% | 33,129,718 | 86% | 23,077,592 | 94% | 12,496,785 | 98% | 17,481,918 | 92% | 20,814,990 | 94% | 16,599,485 | 91% |
| **VMRC-LCD** | 35,513,867 | 97% | 22,012,353 | 97% | 32,101,470 | 94% | 29,637,264 | 96% | 14,310,001 | 97% | 23,498,455 | 97% | 40,330,632 | 97% | 42,317,596 | 98% |
| **RERF-LC-OK** | 23,185,350 | 97% | 38,441,077 | 97% | 64,308,969 | 96% | 19,515,810 | 92% | 25,671,164 | 97% | 27,821,894 | 97% | 19,185,968 | 97% | 65,905,050 | 97% |
| **Average** | 31,376,285 | 96% | 29,950,476 | 96% | 29,389,005 | 94% | 28,752,719 | 93% | 21,125,630 | 97% | 30,103,794 | 96% | 25,490,987 | 96% | 29,984,883 | 97% |

*Table 5 Sequencing and phasing characteristics for 10x GemCode synthetic long read whole exome with regulome sequencing*

| WES+R | Sequencing Statistics | | | | | Phasing Statistics | | |
|---|---|---|---|---|---|---|---|---|
| **Cell Line** | Number of Reads | Mapped Read% | PCR Duplication | Bait Coverage | Depth | Longest Phase Block | N50 Phase Block | SNPs Phased |
| **A427** | 99,593,100 | 99.5% | 3.01% | 99.4% | 59.65 | 835,114 | 116,420 | 11.50% |
| **A549** | 95,848,264 | 99.5% | 3.21% | 99.3% | 56.27 | 729,146 | 76,070 | 11.80% |
| **ABC-1** | 94,462,990 | 99.4% | 17.60% | 99.0% | 52.33 | 1,049,789 | 106,062 | 11.80% |
| **H322** | 88,136,374 | 99.5% | 3.56% | 99.1% | 51.35 | 1,249,705 | 112,172 | 11.50% |
| **H1299** | 103,133,700 | 99.4% | 5.63% | 99.4% | 61.15 | 1,087,437 | 88,677 | 11.50% |
| **H1648** | 85,929,520 | 99.5% | 3.46% | 99.4% | 51.49 | 1,073,574 | 94,214 | 10.70% |
| **H1650** | 85,269,994 | 99.5% | 5.59% | 99.0% | 50.05 | 769,042 | 89,937 | 10.00% |
| **H1703** | 97,084,096 | 99.4% | 5.52% | 99.3% | 54.65 | 781,297 | 104,174 | 11.80% |
| **H1819** | 93,562,794 | 99.3% | 6.80% | 99.2% | 52.51 | 709,032 | 95,635 | 11.50% |
| **H1975** | 83,093,898 | 99.2% | 2.63% | 99.1% | 48.99 | 652,676 | 84,566 | 9.51% |
| **H2126** | 95,109,618 | 99.4% | 7.52% | 99.3% | 53.93 | 918,379 | 125,972 | 11.40% |
| **H2228** | 91,567,448 | 99.2% | 3.15% | 99.4% | 54.40 | 896,157 | 123,272 | 10.20% |
| **H2347** | 93,224,434 | 99.4% | 8.65% | 99.3% | 53.37 | 811,704 | 100,329 | 10.60% |
| **II-18** | 85,938,160 | 99.5% | 1.75% | 99.1% | 50.97 | 468,750 | 78,308 | 10.60% |
| **LC2ad** | 87,391,948 | 99.1% | 3.19% | 99.3% | 51.01 | 1,085,664 | 130,385 | 10.20% |
| **PC-9** | 93,671,674 | 99.1% | 8.50% | 98.9% | 55.43 | 909,689 | 98,398 | 10.80% |
| **PC-14** | 85,912,630 | 99.5% | 2.15% | 99.3% | 51.62 | 559,312 | 88,049 | 9.15% |
| **RERF-LC-Ad1** | 95,459,772 | 99.5% | 3.49% | 99.3% | 55.92 | 773,885 | 98,237 | 11.00% |
| **RERF-LC-Ad2** | 85,929,050 | 99.5% | 3.56% | 99.4% | 51.22 | 781,428 | 97,919 | 10.10% |
| **RERF-LC-KJ** | 102,867,672 | 99.4% | 5.20% | 99.5% | 60.16 | 793,178 | 87,920 | 11.90% |
| **RERF-LC-MS** | 73,659,054 | 99.4% | 4.91% | 99.1% | 41.65 | 748,538 | 103,805 | 9.16% |
| **VMRC-LCD** | 83,375,866 | 99.4% | 5.12% | 99.1% | 47.48 | 876,641 | 89,340 | 10.30% |
| **RERF-LC-OK** | 101,048,218 | 99.5% | 3.86% | 99.4% | 60.36 | 622,497 | 90,476 | 10.50% |
| **Average** | 91,269,545 | 99.4% | 5.0% | 99.2% | 53 | 826,778 | 98,131 | 10.7% |

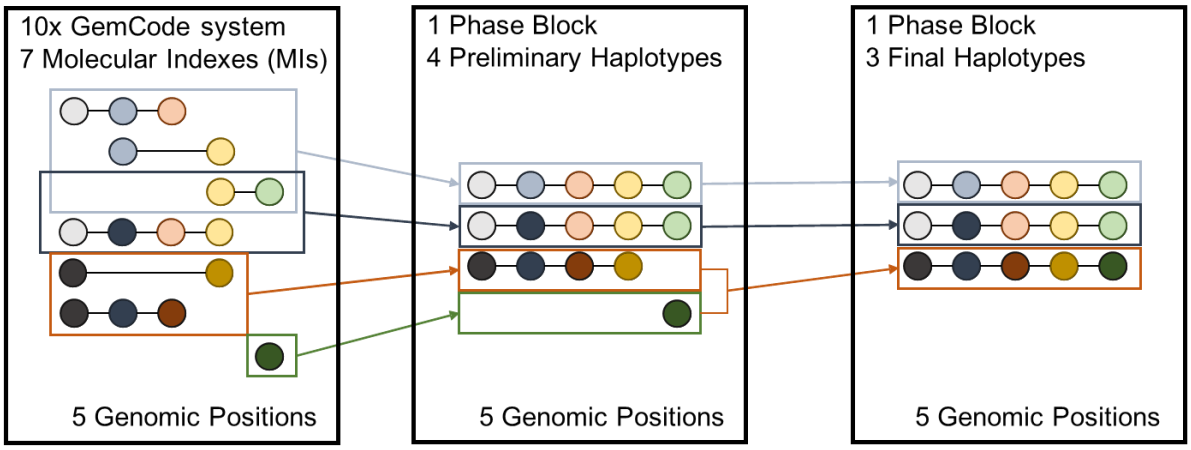



*Figure 3 Graphic Representation of Haplotype Phasing. (Left) Overlapping MIs are retrieved; (Center) Compatible MIs are merged into single "Pre-haplotypes"; (Right) Missing positions are filled to produce final Haplotypes*

*Table 6 DNA fragments and Primers Used for Experimental Validation*

| **(A) DNA Fragments Used in the Luciferase Assay** | |
|---|---|
| **DNA Fragment** | **Sequence** |
| NFATC1 Mutant | GCTAGCTCGATTTATGGTTTCTACACACCAGACACTTTAACCTCCAACCCCCCCCATCCAAA GCCAACAAGAAAATGCGGTGCCGTGTTGGCAGCTGAGCTGCGCCGGAAGAGACGCAGGG AGACGTGAGAGAGGAAAGTGTGAGTGGCCGGGGGGGCCTCCCCCCGTCAGAAGTCGCGCA GTCGCGCCCATAAAACGCCCCCTCCGGAAGCTT |
| NFATC1 Wild type | GCTAGCTCGATTTATGGTTTCTACACACCAGACACTTTAACCTCCAACCCCCCCCATCCAAA GCCAACAAGAAAATGCGGTGCCGTGTTGGCAGCTGAGCTGCGCCCGAAGAGACGCAGGG AGACGTGAGAGAGGAAAGTGTGAGTGGCCGGGGGGGCCTCCCCCCGTCAGAAGTCGCGCA GTCGCGCCCATAAAACGCCCCCTCCGGAAGCTT |
| **(B) Primers Used for qPCR** | |
| **Primer Name** | **Sequence** |
| **Target region** | |
| Primer_F_1_97bp | CCATCCAAAGCCAACAAGAA |
| Primer_F_2_123bp | CCAGACACTTTAACCTCCAACC |
| Primer_F_3_164bp | CACATAAGGGTGTCGTGCAA |
| Primer_R_shared | GGCCACTCACACTTTCCTCT |
| **Positive control** | |
| RPS26_F | CAGCAGAAATGCTGAATGTAAAGG |
| RPS26_R | CATGAGATCCCTACGCGGAC |
| **Negative control** | |
| Negative_control_1_F | CTGCCACTTGAGGGTGAGG |
| Negative_control_1_R | CCATCTTGCATGCAGTTAGCC |

# Results

## Mutations Detected in Lung Adenocarcinoma Cell Lines

Whole genome sequencing data of all 23 lung adenocarcinoma cell lines were retrieved and reanalyzed. On average, 4,017,667 SNVs per cell line were detected. An average of 1,375,802 coding region SNVs were annotated, with 19,086 SNVs in exon regions per cell line. For the regulatory regions, promoters and enhancers plus repressive marks were defined for the individual cell lines by considering ChIP-seq peaks and consensus TSS-seq. The SNVs in those regulatory regions were filtered for benign germline variants by overlapping with dbSNP. As a result, 46,149 potential regulatory SNVs were identified per cell line (Table 7). The functions of these potential regulatory SNVs were interpreted and categorized based on promoter- or enhancer-specific ChIP-seq markers. A summary of the detected variants is shown in Table 7.

*Table 7 Summary of the SNPs/SNVs detected by GATK*

| Cell line | All SNPs/SNVs | Coding SNPs/SNVs | Exon SNPs/SNVs | Regulatory SNVs |
|---|---|---|---|---|
| **A427** | 4,024,063 | 1,397,615 | 18,775 | 70,336 |
| **A549** | 3,762,488 | 1,007,875 | 16,143 | 37,976 |
| **ABC-1** | 3,918,935 | 1,359,715 | 18,666 | 16,068 |
| **H322** | 3,710,129 | 1,273,472 | 17,904 | 20,721 |
| **H1299** | 3,910,954 | 1,343,074 | 18,287 | 49,799 |
| **H1648** | 4,834,699 | 1,701,139 | 24,819 | 55,458 |
| **H1650** | 3,738,924 | 1,272,227 | 17,280 | 68,525 |
| **H1703** | 3,908,849 | 1,340,392 | 18,276 | 48,520 |
| **H1819** | 4,169,230 | 1,441,883 | 19,326 | 61,870 |
| **H1975** | 4,026,746 | 1,333,864 | 19,389 | 36,275 |
| **H2126** | 4,233,027 | 1,457,113 | 19,789 | 76,104 |
| **H2228** | 4,407,002 | 1,512,216 | 19,312 | 80,690 |
| **H2347** | 3,265,345 | 1,316,041 | 18,102 | 37,756 |
| **II-18** | 4,122,525 | 1,428,765 | 20,231 | 37,923 |
| **LC2ad** | 3,955,271 | 1,372,090 | 18,855 | 9,568 |
| **PC-9** | 3,949,215 | 1,368,717 | 18,717 | 43,016 |
| **PC-14** | 3,712,268 | 1,259,609 | 17,977 | 10,717 |
| **RERF-LC-Ad1** | 4,368,425 | 1,514,733 | 20,936 | 68,911 |
| **RERF-LC-Ad2** | 4,213,008 | 1,449,905 | 19,887 | 70,040 |
| **RERF-LC-KJ** | 4,135,667 | 1,426,828 | 19,961 | 33,263 |
| **RERF-LC-MS** | 3,949,142 | 1,348,821 | 17,980 | 48,424 |
| **VMRC-LCD** | 4,078,677 | 1,383,592 | 19,613 | 36,918 |
| **RERF-LC-OK** | 4,011,742 | 1,333,768 | 18,749 | 42,540 |

| | | | | |
|---|---|---|---|---|
| **Average** | 4,017,667 | 1,375,802 | 19,086 | 46,149 |

## Multi-omics Approach in Mutation Analysis

To distinguish between functional regulatory mutations and functionally silent "passenger" mutations (The Cancer Genome Atlas Research et al., 2013; The Cancer Genome Atlas Research Network, 2014), I examined whether the detected SNVs activated or repressed their downstream transcript targets. The activated or repressed status of an allele could be determined from frequencies of the SNVs or SNPs of that particular allele in the transcriptome. Activating regulatory mutations should increase the frequency of the variant sequences in the RNA-seq "tags", while repressive regulatory mutations would decrease it. This followed the so-called "allelic imbalance" detection approach (Figure 4) (Baran et al., 2015; Melé et al., 2015; Sonawane et al., 2017). To implement this, I considered the ratio between variant frequencies of alternative/reference nucleotides of SNPs/SNVs in mRNA transcripts as surrogates for allele expression patterns and the ratio between variant frequencies of alternative/reference nucleotides of regulatory mutations in each individual ChIP-seq as representative of functionality of each mutation.

To address potential problems associated with copy number aberrations (CNAs) common to cancer cell lines (Table 2), I normalized both alternative and reference frequencies of the variants in RNA and ChIP-seq by the corresponding genomic variant frequencies in WGS. Because sequencing depths in WGS could represent ploidy in the genome (Abyzov, Urban, Snyder, & Gerstein, 2011; Roller, Ivakhno, Lee, Royce, & Tanner, 2016), the normalized frequencies should only represent the functional bias of the regulatory modifications and transcripts.

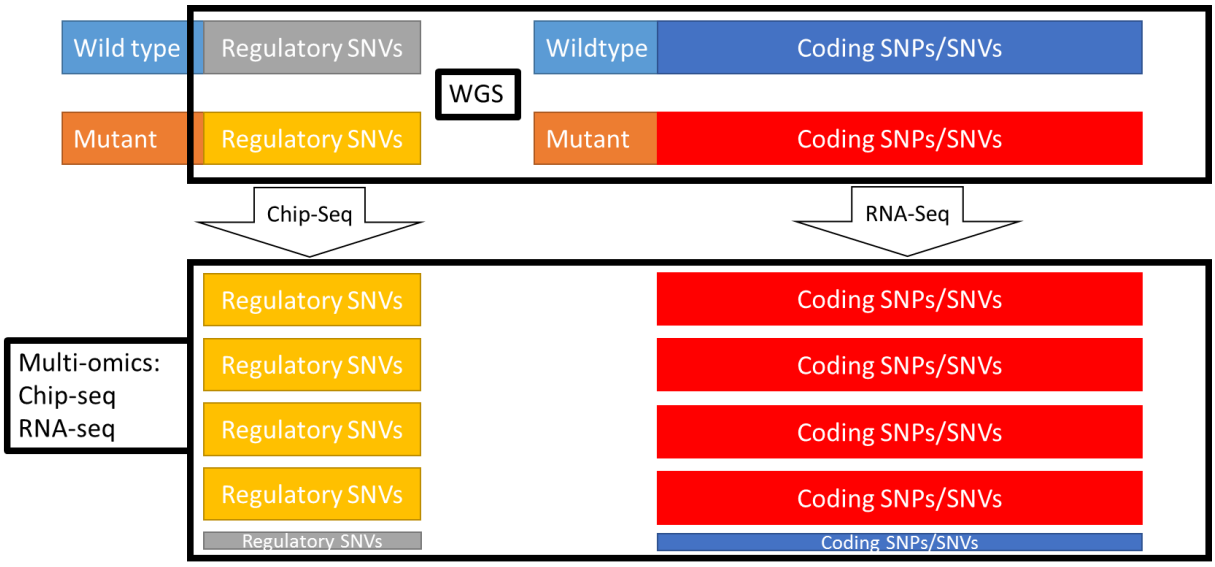
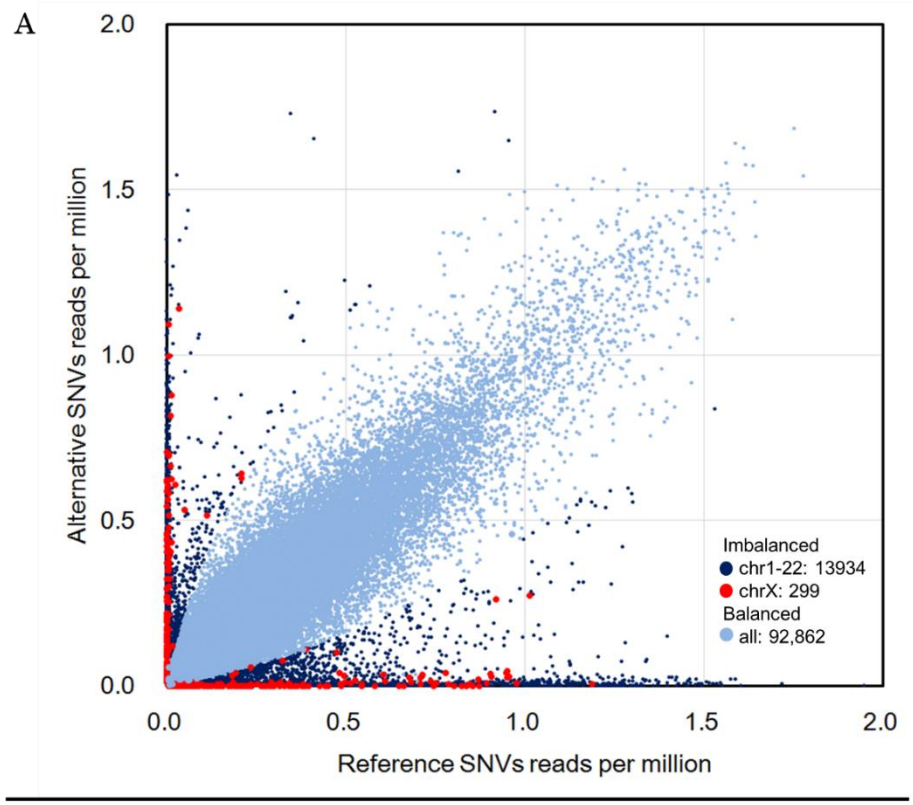


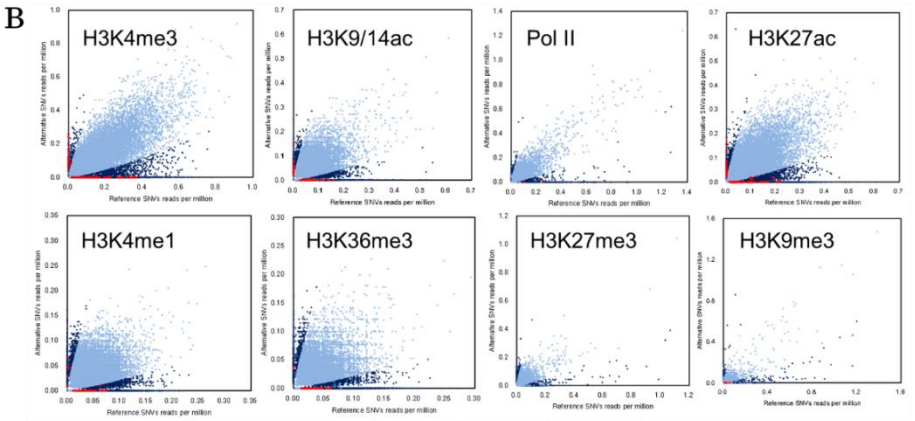

*Figure 4 Approach in detecting allelic imbalance expression. (Top) WGS Variants are classified into regulatory variants or coding variants; (Bottom) Variant frequencies in ChIP-seq (regulatory variants) and RNA-seq (coding variants) are investigated for biases.*

## RNA-seq Reveals Transcript Allelic Imbalance Expression

I examined which RefSeq transcripts exhibited allelic imbalance expression and thus were potentially under mono-allelic transcriptional regulation. Using the ratio of WGS-normalized alternative/reference variants in RNA-seq, I considered variants with the following 2 criteria: #1) the ratio in WGS and RNA-seq was significantly different in 2 by 2 contingency tables at $p<0.01$, Fisher's exact test and #2) the ratio in RNA-seq was at least 5-fold, favoring either alternative or reference variants. From the coding regions of 29,251 transcript counts (averaging 1,271 per cell line), 107,155 coding variants (18,330 per cell

line) were detected. Allelic imbalance expression was detected in 7,915 transcripts (596 per cell line) or 14,233 coding variants (619 per cell line, Figure 5A, Table 8).




| Chromatin mark | | H3K4me3 | H3K9/14ac | Pol II | H3K27ac | H3K4me1 | H3K36me3 | H3K27me3 | H3K9me3 |
|---|---|---|---|---|---|---|---|---|---|
| Imbalanced | chr1-22 | 8964 | 6076 | 2721 | 18056 | 14145 | 12615 | 3667 | 1836 |
| | chrX | 89 | 48 | 2 | 137 | 41 | 35 | 0 | 6 |
| Balanced | all | 14486 | 9719 | 4111 | 28836 | 21881 | 15347 | 4318 | 1875 |

*Figure 5 Allelic imbalance expression plots; X-axis represents reference reads' frequencies; Y-axis represents alternative read's frequencies; Non-X chromosome variants with more than 5 fold bias are in dark blue; X chromosome imbalances are in red. (A) RNA-seq imbalances; (B) ChIP-seq imbalance.*

*Table 8 The number of heterozygous SNVs with imbalanced and balanced transcriptions*

| Sex | Cell line | Autosome + Y | | | | X | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | Heterozygous SNPs | Expression | | | Heterozygous SNPs | Expression | | |
| | | | Balanced | Imbalanced | %imbalance | | Balanced | Imbalanced | %imbalance |
| Female | LC2/ad | 1833572 | 5071 | 422 | 7.68 | 74265 | 13 | 78 | 85.7 |
| | H1819 | 1760876 | 4402 | 679 | 13.4 | 6276 | 28 | 39 | 58.2 |
| | H1975 | 2190422 | 5288 | 591 | 10.1 | 4192 | 0 | 4 | 100.0 |
| | H2228 | 2331330 | 5904 | 671 | 10.2 | 14467 | 7 | 38 | 84.4 |
| | H2347 | 2370936 | 5462 | 585 | 9.7 | 92483 | 15 | 140 | 90.3 |
| Male | A427 | 1856941 | 4147 | 490 | 10.6 | 4796 | 6 | 1 | 14.3 |
| | A549 | 2084037 | 3422 | 479 | 12.3 | 33667 | 0 | 18 | 100.0 |
| | ABC-1 | 1423400 | 2981 | 381 | 11.3 | 4283 | 3 | 8 | 72.7 |
| | H1299 | 1566125 | 3346 | 378 | 10.2 | 7107 | 0 | 10 | 100.0 |
| | H1648 | 2441256 | 5377 | 706 | 11.6 | 3363 | 2 | 21 | 91.3 |
| | H1650 | 1136166 | 2372 | 343 | 12.6 | 2677 | 1 | 21 | 95.5 |
| | H1703 | 1634715 | 3645 | 433 | 10.6 | 4480 | 2 | 8 | 80.0 |
| | H2126 | 1573235 | 3680 | 423 | 10.3 | 5148 | 0 | 6 | 100.0 |
| | RERF-LC-Ad1 | 2275585 | 5607 | 776 | 12.2 | 3557 | 0 | 4 | 100.0 |
| | RERF-LC-Ad2 | 1975614 | 4826 | 541 | 10.1 | 4601 | 2 | 9 | 81.8 |
| | RERF-LC-KJ | 1892689 | 4741 | 546 | 10.3 | 3997 | 0 | 10 | 100.0 |
| | VMRC-LCD | 1873618 | 4595 | 544 | 10.6 | 5253 | 5 | 8 | 61.5 |
| Unknown | PC-14 | 1120101 | 303 | 2304 | 88.4 | 2177 | 1 | 17 | 94.4 |
| | PC-9 | 1538099 | 3825 | 432 | 10.2 | 5973 | 13 | 13 | 50.0 |
| | H322 | 1222734 | 3358 | 433 | 11.4 | 3316 | 2 | 11 | 84.6 |
| | II-18 | 1345778 | 3039 | 354 | 10.4 | 3733 | 1 | 1 | 50.0 |
| | RERF-LC-MS | 1491318 | 3191 | 633 | 16.6 | 4590 | 1 | 4 | 80.0 |
| | RERF-LC-OK | 1855675 | 4147 | 513 | 11.0 | 74904 | 31 | 107 | 77.5 |

To validate the allelic imbalance of RefSeq transcripts, I first inspected whether the imprinting from X inactivation was presented. From 5 cell lines of known female origin and 12 cell lines of known male origin, I observed much larger amounts of heterozygous coding variants in the female cell lines (72 vs 16 on average, Table 8). For the female cell

lines, out of 362 total variants, 299 (83%) were considered to be under mono-allelic regulatory effects. Based on the number of detected variants, the unknown sex origin cell line RERF-LC-OK should be of female origin. These variant imbalances were also observed at the transcript level (Figure 6). Cross-referencing to previously reported X-inactivation imprinted transcripts found 67 transcripts from 17 genes (Morison, Paton, & Cleverley, 2001).

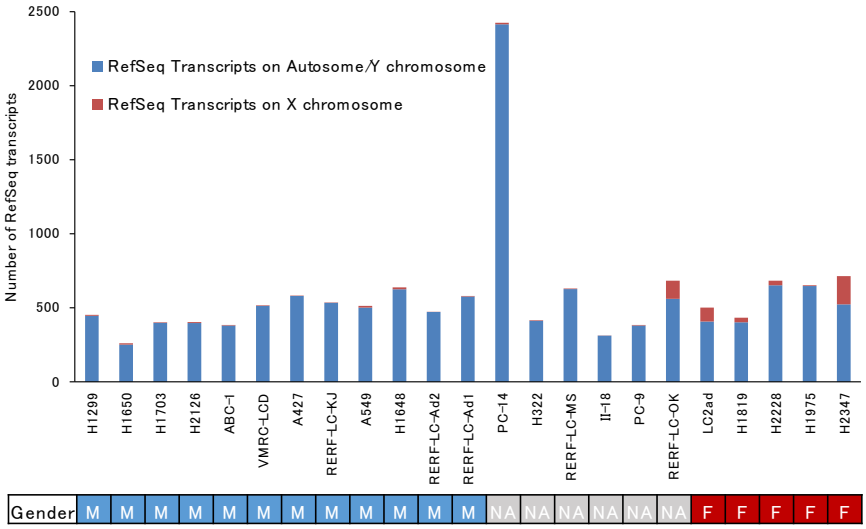
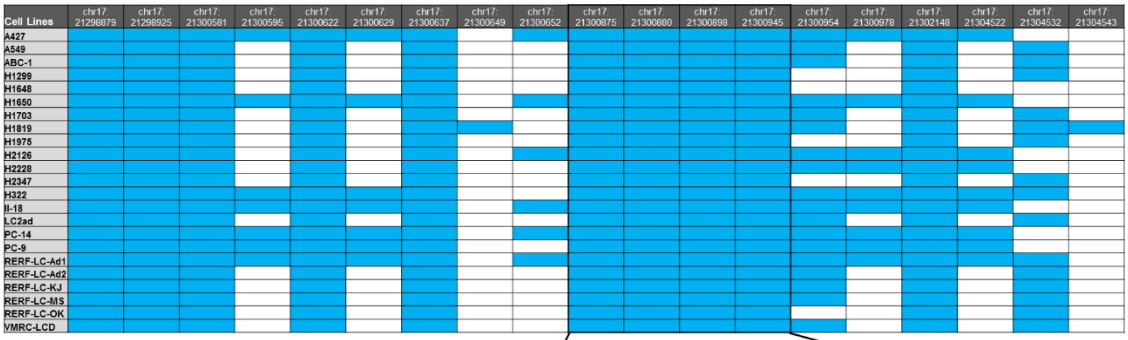


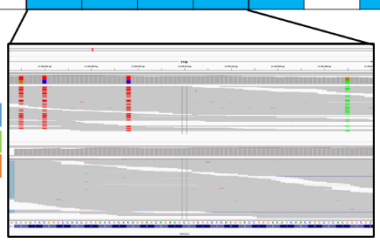

*Figure 6 RefSeq transcripts with mRNA allele imbalance in each cell line. Transcripts on X-chromosome are shown in red, others in blue. Sex of origin is shown on bottom panel. RERF-LC-OK is likely to have a female origin.*

Imprinting is not limited to the X-chromosome. Genes on autosomes also exhibit parental-specific expression via epigenetic controls (Barlow & Bartolomei, 2014). Such imprinting results from linage-specific, sex-specific or developmental-specific processes, which are not the goal of this study. I discarded RefSeq transcripts that were found to be imbalanced in more than 1/3 (7) of the cell lines regardless of the presence of regulatory mutations. A total of 124 transcripts in 76 genes (Table 9) were removed. Two examples of *MAP2K3* and *BCLAF1* imprinting are shown in Figure 7.

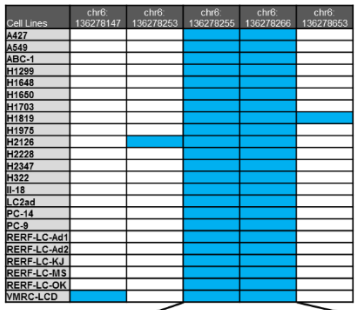
A *MAP2K3* imprinting:
H1975 chr17: 21298879-21304543




| Data Set | chr17:21300875 G>T | | chr17:21300880 C>T | | chr17:21300898 C>T | | chr17:21300945 G>A | |
|---|---|---|---|---|---|---|---|---|
| WGS | G:42 | T:46 | C:42 | T:50 | C:38 | T:51 | G:37 | A:37 |
| mRNA | G:217 | | C:228 | | C:245 | | G:247 | |



B *BCLAF1* imprinting:
H1975 chr6: 136278255-136278266




| Data Set | chr6:136278255 G>C | | chr6:136278266 T>C | |
|---|---|---|---|---|
| WGS | G:36 | C:29 | T:36 | C:27 |
| mRNA | G:172 | | T:182 | |

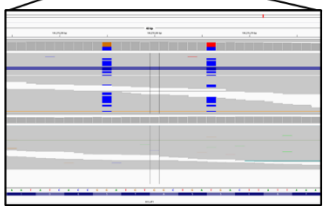


*Figure 7 Two examples of transcripts considered imprinted by RNA-seq imbalance in every cell line. Presences of variants are indicated by blue boxes. Frequencies in H1975 cell line are shown in table and IGV's graphics. (A) MAP2K3 (B) BCLAF1*

*Table 9 List of RefSeq transcripts considered to be imprinted*

| RefSeq | Gene | # Cell Lines | RefSeq | Gene | # Cell Lines | RefSeq | Gene | # Cell Lines | RefSeq | Gene | # Cell Lines | RefSeq | Gene | # Cell Lines | RefSeq | Gene | # Cell Lines |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| NM_002756 | *MAP2K3* | 23 | NM_001012636 | *IL32* | 16 | NM_019610 | *RBMXL1* | 13 | NM_001290208 | *ZNF717* | 11 | NM_133378 | *TTN* | 9 | NM_020957 | *PCDHB16* | 8 |
| NM_014739 | *BCLAF1* | 23 | NM_001012635 | *IL32* | 16 | NM_001037501 | *NBPF8* | 13 | NM_001128223 | *ZNF717* | 11 | NM_003319 | *TTN* | 9 | NM_020445 | *ACTR3B* | 8 |
| NM_145109 | *MAP2K3* | 23 | NM_001012634 | *IL32* | 16 | NM_002568 | *PABPC1* | 13 | NM_014272 | *ADAMTS7* | 11 | NM_182619 | *CLEC18A* | 9 | NM_001040135 | *ACTR3B* | 8 |
| NM_001077441 | *BCLAF1* | 23 | NM_001012633 | *IL32* | 16 | NM_001162536 | *RBMXL1* | 13 | NM_001198832 | *PDE4DIP* | 11 | NM_001193318 | *RNF212* | 9 | NM_001291420 | *GOLGA6L9* | 8 |
| NM_001077440 | *BCLAF1* | 23 | NM_001012632 | *IL32* | 16 | NM_024690 | *MUC16* | 13 | NM_144682 | *SLFN13* | 10 | NM_001277444 | *NBPF9* | 9 | | | |
| NM_170606 | *KMT2C* | 22 | NM_001012631 | *IL32* | 16 | NM_001271223 | *OBSCN* | 13 | NM_001373 | *DNAH14* | 10 | NM_000661 | *RPL9* | 9 | | | |
| NM_001164315 | *ANKRD36* | 22 | NM_001012718 | *IL32* | 16 | NM_052843 | *OBSCN* | 12 | NM_002281 | *KRT81* | 10 | NM_145061 | *SKA3* | 9 | | | |
| NM_001042414 | *PSPC1* | 21 | NM_001286555 | *DUSP22* | 15 | NM_017750 | *RETSAT* | 12 | NM_001290210 | *ZNF717* | 10 | NM_001136214 | *CLEC18A* | 9 | | | |
| NM_022662 | *ANAPC1* | 20 | NM_004399 | *DDX11* | 15 | NM_032926 | *TCEAL3* | 12 | NM_030979 | *PABPC3* | 10 | NM_001271197 | *CLEC18A* | 9 | | | |
| NM_006437 | *PARP4* | 19 | NM_033655 | *CNTNAP3* | 15 | NM_001098623 | *OBSCN* | 12 | NM_001369 | *DNAH5* | 10 | NM_002016 | *FLG* | 8 | | | |
| NM_018264 | *TYW1* | 19 | NM_001005751 | *FAM21A* | 15 | NM_001006933 | *TCEAL3* | 12 | NM_001009931 | *HRNR* | 10 | NM_001201380 | *CNTNAP3B* | 8 | | | |
| NM_001271733 | *MST1L* | 19 | NM_003890 | *FCGBP* | 15 | NM_001278141 | *NBPF12* | 12 | NM_001290209 | *ZNF717* | 10 | NM_001256417 | *NBPF3* | 8 | | | |
| NM_182623 | *FAM131C* | 19 | NM_001080400 | *PLIN4* | 15 | NM_004987 | *LIMS1* | 12 | NM_001099771 | *POTEF* | 10 | NM_001256416 | *NBPF3* | 8 | | | |
| NM_173601 | *GXYLT1* | 18 | NM_001291398 | *FAM21A* | 15 | NM_001193488 | *LIMS1* | 12 | NM_001166017 | *SKA3* | 10 | NM_019120 | *PCDHB8* | 8 | | | |
| NM_017940 | *NBPF1* | 18 | NM_001085457 | *CBWD6* | 15 | NM_001193484 | *LIMS1* | 12 | NM_018937 | *PCDHB3* | 10 | NM_003211 | *TDG* | 8 | | | |
| NM_001284 | *AP3S1* | 18 | NM_001257145 | *DDX11* | 15 | NM_001193485 | *LIMS1* | 12 | NM_138420 | *AHNAK2* | 9 | NM_182588 | *RGPD4* | 8 | | | |
| NM_001099650 | *GXYLT1* | 18 | NM_152438 | *DDX11* | 15 | NM_001193482 | *LIMS1* | 12 | NM_001037675 | *NBPF9* | 9 | NM_001018115 | *FANCD2* | 8 | | | |
| NM_019601 | *SUSD2* | 17 | NM_014675 | *CROCC* | 15 | NM_001193483 | *LIMS1* | 12 | NM_198181 | *GOLGA6L9* | 9 | NM_001243776 | *CEP57* | 8 | | | |
| NM_003174 | *SVIL* | 16 | NM_001257144 | *DDX11* | 15 | NM_001720 | *BMP8B* | 11 | NM_006931 | *SLC2A3* | 9 | NM_032264 | *NBPF3* | 8 | | | |
| NM_004221 | *IL32* | 16 | NM_002139 | *RBMX* | 15 | NM_001079809 | *TMEM183B* | 11 | NM_001024921 | *RPL9* | 9 | NM_015383 | *NBPF14* | 8 | | | |
| NM_023924 | *BRD9* | 16 | NM_030653 | *DDX11* | 15 | NM_021012 | *KCNJ12* | 11 | NM_001256850 | *TTN* | 9 | NM_020185 | *DUSP22* | 8 | | | |
| NM_014696 | *GPRIN2* | 16 | NM_024786 | *ZDHHC11* | 14 | NM_004807 | *HS6ST1* | 11 | NM_133432 | *TTN* | 9 | NM_032144 | *RAB6C* | 8 | | | |
| NM_001009877 | *BRD9* | 16 | NM_182905 | *WASH1* | 14 | NM_001277115 | *DNAH11* | 11 | NM_133437 | *TTN* | 9 | NM_000186 | *CFH* | 8 | | | |
| NM_021738 | *SVIL* | 16 | NM_030930 | *UNC93B1* | 14 | NM_001164586 | *IGFN1* | 11 | NM_001267550 | *TTN* | 9 | NM_033084 | *FANCD2* | 8 | | | |

## ChIP-seq Reveals Allelic Preference Modifications in Regulatory Mutations

Similar to allele expression, the imbalance in histone modifications was determined by the ratio of variant frequencies in ChIP-seq. Instead of relying on a public database, regulatory regions were defined individually for cell lines by histone marks. By processing each ChIP-seq assay in each cell line individually, I detected a total of 100,573 variants in all of the regulatory regions modulating 17,929 transcripts (Figure 5B). A total of 1,794 regulatory SNVs (81 per cell line) were paired with 1,655 coding variants in 730 RefSeq transcripts (38 per cell line, Figure 8).

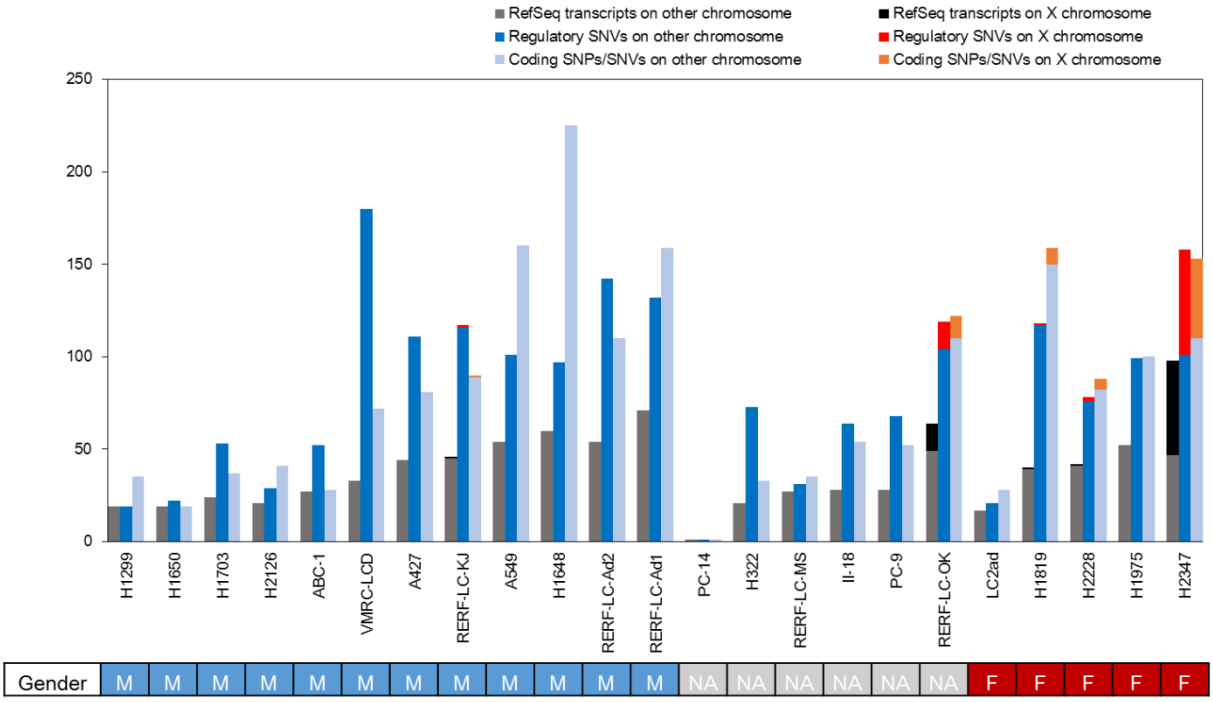



*Figure 8 Breakdown of RefSeq transcripts, coding variants and regulatory SNVs. Number of imbalance transcripts are shown gray/black bars; imbalance regulatory SNVs in blue/red bars; Coding variants in light blue/orange bars.*

## Phasing of Variants Detected in WGS with 10x Genomics GemCode

To further validate the 1,794 regulatory SNVs, their direct associations with the downstream 1,655 coding variants were analyzed. Interactions between regulatory elements and their downstream transcripts could be either *cis*- on the same allele or *trans*- on a different allele. Both of these possibilities represent different regulatory mechanisms. The

phasing of regulatory SNVs and their downstream SNVs/SNPs is indispensable for the identification and interpretation of these interactions.

The 10x Genomics GemCode platform was employed for allele phasing. To expedite the analysis, sequencing was performed using whole exome plus regulome bait (113.7 Mb), which included coding regions and promoters, enhancers and DNA methylated regions deposited in the ENCODE project. An average of 45,679,789 paired-end reads per cell line were sequenced, which averaged 53x coverage in targeted regions. Out of the 4,038,252 variants on average per cell line, 10.8% of the variants were covered and were phased by the default 10x Genomics Long Ranger pipeline. It should be noted that the default pipeline assumes a diploid human genome; thus, the default results were found to be unsuitable for cancer cell line genomes.

To adapt the 10x Genomics GemCode to this study, I modified the analytical pipeline. In brief (see details in the Methods section), for each variant having adequate coverage, UMIs corresponding to the alternative and reference nucleotides were recovered. Allele haplotypes were constructed from each unique combination of the variants supported by the UMIs. I linked the variants in different positions into phased regions called “Phased Blocks”. I obtained 7,004 phased blocks per cell line on average with an average length of 55 kb per block (1.7 Mb maximum, Figure 9A). Anchoring these blocks were WGS variants with an average of 13 variants per block (702 max, Figure 9B). From the combinations of the variants, 3 haplotypes were made on average (Figure 9C). Collectively, in all cell lines, 40,073 blocks (1,742 average) contained 89,333 regulatory SNVs (3,884 average). On average, 2 regulatory SNVs were linked in 33 variant chains (Figure 9D). Statistics for each cell line are shown in Table 10.

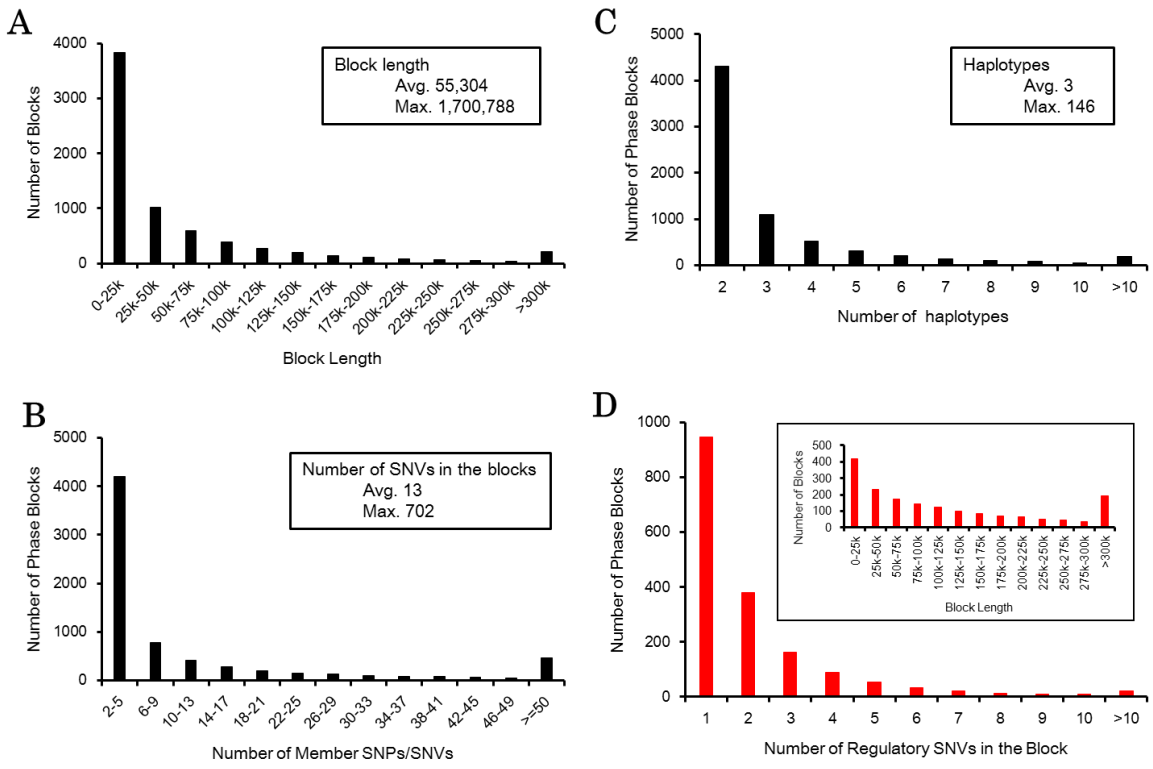



*Figure 9 Phase Block Statistics histograms. (A) Block length distributions. (B) WGS SNV count distributions. (C) Haplotypes from (B) distributions. (D) Stats of blocks containing regulatory SNVs; inset shows length of blocks containing regulatory SNVs.*

*Table 10 Statistics of the Phase Blocks in Each Cell Line*

| Cell line | Phase Block Length | | | # of SNPs/SNVs in a Phase Block | | | # of Haplotypes in a Phase Block | | |
|---|---|---|---|---|---|---|---|---|---|
| | Max | Average | Median | Max | Average | Median | Max | Average | Median |
| **A427** | 1,194,537 | 65,754 | 23,338 | 472 | 16 | 4 | 64 | 3.56 | 2 |
| **A549** | 829,372 | 35,012 | 11,694 | 311 | 10 | 3 | 44 | 2.94 | 2 |
| **ABC-1** | 1,198,524 | 53,875 | 20,988 | 608 | 13 | 3 | 103 | 3.47 | 2 |
| **H322** | 1,138,518 | 53,459 | 18,296 | 387 | 14 | 3 | 43 | 3.41 | 2 |
| **H1299** | 848,580 | 43,843 | 18,352 | 690 | 12 | 4 | 98 | 3.18 | 2 |
| **H1648** | 1,052,406 | 53,067 | 17,726 | 638 | 15 | 4 | 100 | 3.40 | 2 |
| **H1650** | 1,227,114 | 38,179 | 13,549 | 396 | 11 | 3 | 56 | 2.97 | 2 |
| **H1703** | 755,716 | 53,745 | 21,890 | 695 | 14 | 4 | 112 | 3.49 | 2 |
| **H1819** | 745,916 | 46,591 | 17,492 | 441 | 14 | 4 | 44 | 3.38 | 2 |
| **H1975** | 1,025,195 | 38,330 | 12,983 | 573 | 12 | 4 | 94 | 3.03 | 2 |
| **H2126** | 950,576 | 64,867 | 25,444 | 595 | 15 | 4 | 134 | 3.46 | 2 |
| **H2228** | 1,427,558 | 67,865 | 24,282 | 518 | 16 | 4 | 55 | 3.46 | 2 |
| **H2347** | 1,300,851 | 55,205 | 20,739 | 481 | 15 | 4 | 64 | 3.52 | 2 |
| **II-18** | 608,739 | 35,312 | 12,450 | 449 | 11 | 4 | 78 | 3.17 | 2 |
| **LC2ad** | 1,304,146 | 72,936 | 27,677 | 620 | 17 | 4 | 106 | 3.53 | 2 |
| **PC-9** | 945,733 | 50,160 | 22,921 | 372 | 13 | 4 | 32 | 3.05 | 2 |
| **PC-14** | 383,302 | 22,036 | 8,343 | 294 | 5 | 2 | 40 | 2.77 | 2 |
| **RERF-LC-Ad1** | 948,045 | 53,509 | 19,925 | 472 | 15 | 4 | 69 | 3.47 | 2 |
| **RERF-LC-Ad2** | 997,249 | 55,219 | 20,673 | 335 | 14 | 4 | 39 | 3.35 | 2 |
| **RERF-LC-KJ** | 798,407 | 45,667 | 19,387 | 692 | 13 | 4 | 154 | 3.35 | 2 |
| **RERF-LC-MS** | 1,208,661 | 54,802 | 21,687 | 506 | 12 | 4 | 61 | 3.19 | 2 |
| **VMRC-LCD** | 987,944 | 47,138 | 18,998 | 499 | 13 | 4 | 95 | 3.39 | 2 |
| **RERF-LC-OK** | 876,013 | 46,425 | 19,036 | 368 | 13 | 4 | 55 | 3.26 | 2 |
| **Average** | 989,265 | 50,130 | 19,038 | 496 | 13 | 4 | 76 | 3.29 | 2 |

A phase block example in the A549 cell line is shown in Figure 10. Seven WGS variants were linked to the region spanning chr2:38038277-38071060 (32 kb), which overlapped with *CYP1B1*. A variant at chr2:38070511 C>CA fell in a regulatory region, and three variants at chr2:38070996 T>C, chr2:38071007 A>G and chr2: 38071060 C>G fell inside the *CYP1B1* coding region. Seven SNVs, including the above four SNVs, were linked together by UMI (blue lines) of the 10x Genomic GemCode platform, clearly separating the two C-T-A-C (upper) and +A-C-G-G (lower) haplotypes apart.

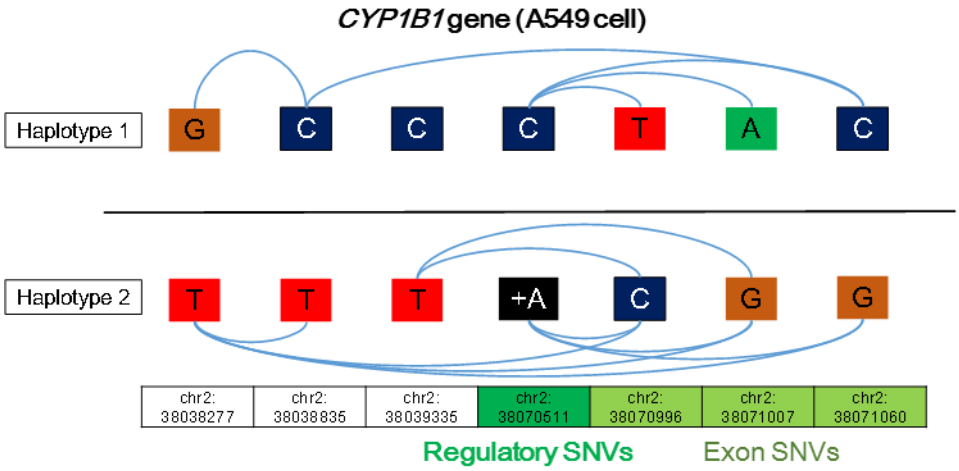



*Figure 10 Example of a Phase Block of CYP1B1 in A549 Cells. Blue lines represents 10x Gemcode MIs connections; regulatory SNV at chr2:38070611 (dark green) is phased to 3 coding variants (light green).*

A practical utilization of phased blocks was illustrated in the phasing of *EGFR* L858R and T790M mutations (Figure 11). These two mutations are known to coexist in the same allele in drug-resistant clones (Liang et al., 2018). The phasing correctly assigned the two mutants (green arrows) together on the same allele in H1975 cells, a cell line known to harbor these two mutations.

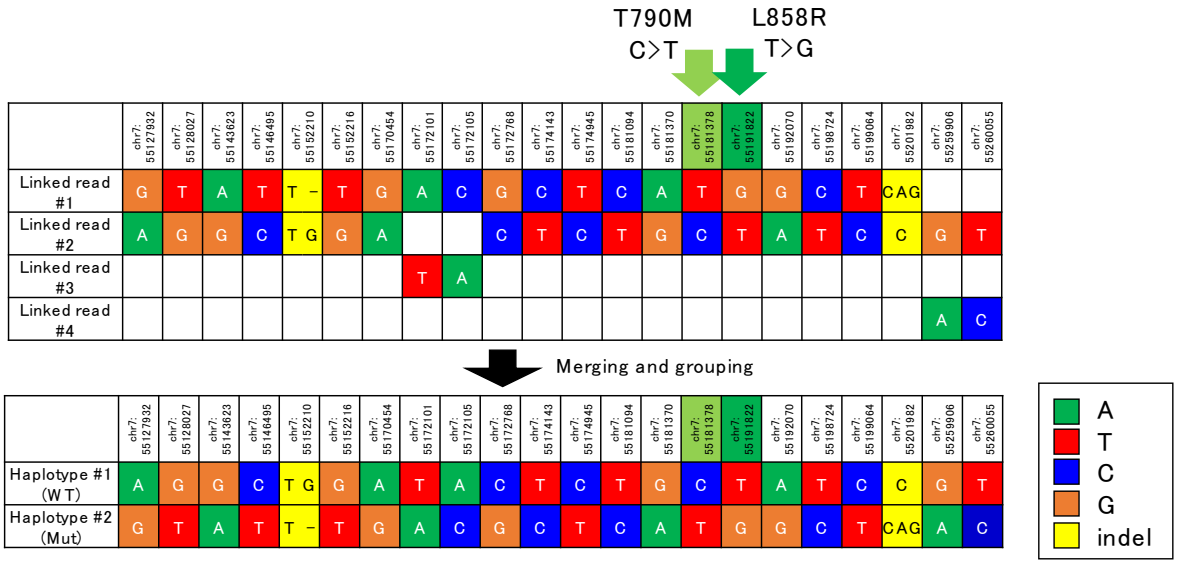



*Figure 11 Phasing of EGFR T790M to L858R. Phase block spanning EGFR in H1975 cell line is shown. T790M and L858R mutations configuration is correctly identified.*

While the majority of the phase block contained two haplotypes, a significant number of the blocks contained more than two haplotypes. To further examine these findings, the reported ploidy of the cell lines in the COSMIC database was referenced (Forbes et al., 2015). Seventeen of the 23 cell lines were reported at an average ploidy of 3.04 (Figure 12). While some cell lines showed large discrepancies with a 1-1.5 ploidy difference, the majority held up comparably well. The most likely reason for the discrepancies was that the haplotype counts in the phase block did not represent the actual copy numbers in those regions but represented the number of unique alleles. One of the most frequently reported regions to undergo amplification is the *ERBB2* gene region, which encodes HER2, a crucial genetic marker in breast cancer. This amplification is the first and most frequently reported one in breast cancer (Kallioniemi et al., 1992) but is also commonly be found in other cancers (Dahlberg et al., 2004; Grob et al., 2012). In the phasing analysis in this study, this region harbored 9 unique haplotype combinations (Figure 13).

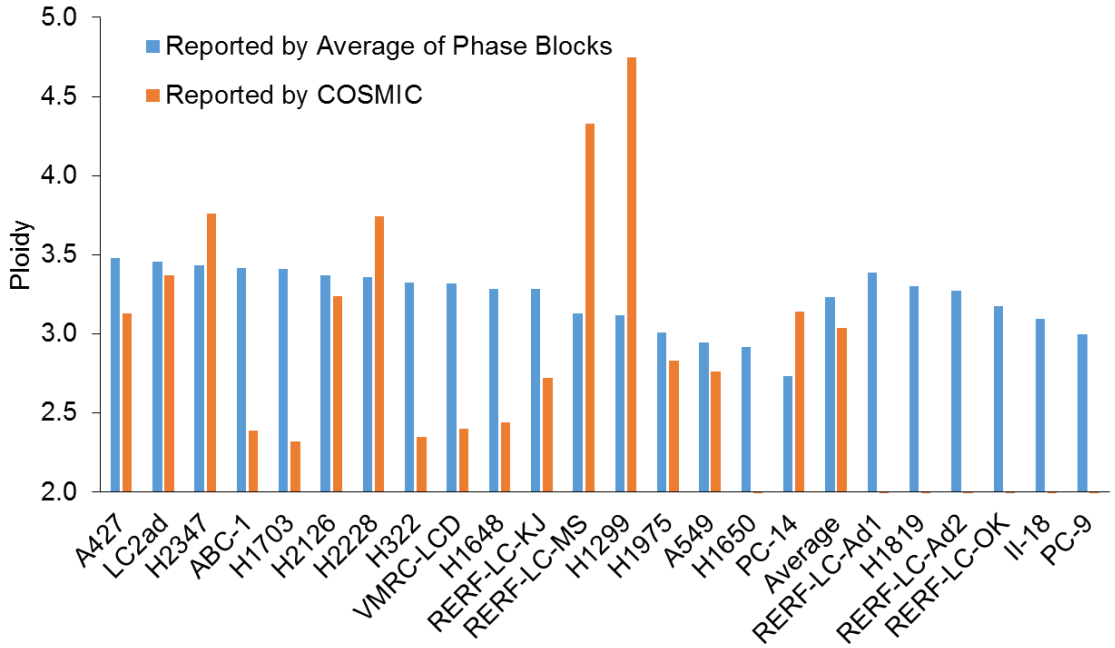



*Figure 12 Phase Block Average Haplotype Compared with the COSMIC Database. Seventeen of 23 cell lines' haplotypes are reported in COSMIC (orange bar).*

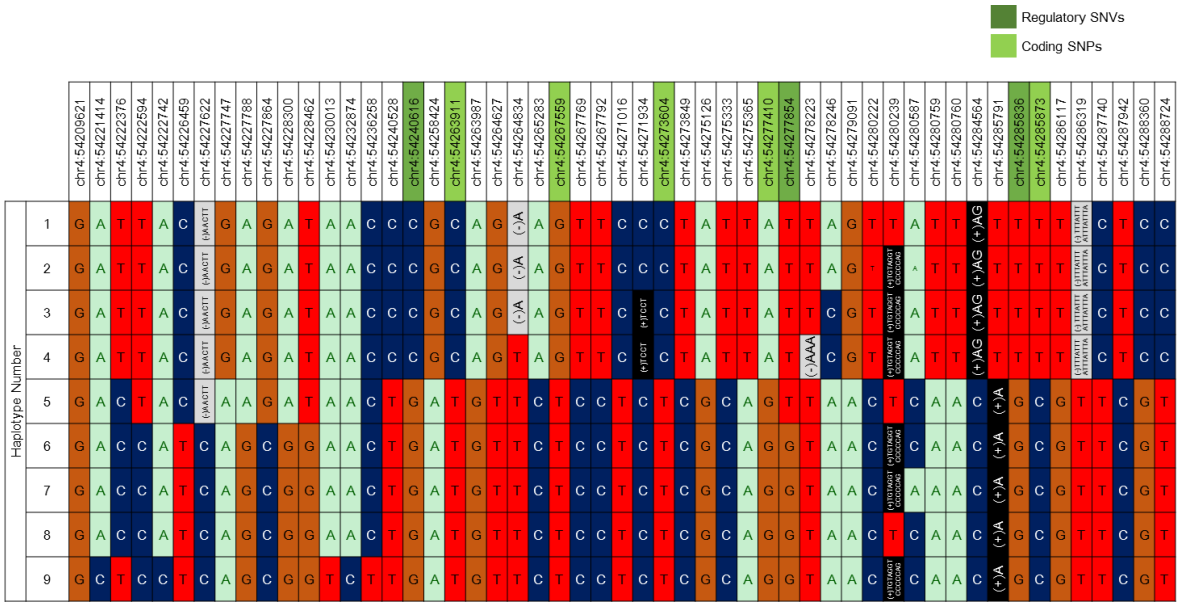



*Figure 13 Phasing of ERBB2 Amplification. Nine unique haplotypes are built from combinations of 10x Gemcode MIs. Several positions overlap with ERBB2 coding region.*

To more directly validate the phasing, whole genome MinION sequencing was conducted for H1975, LC2/ad and RERF-LC-KJ cells. Due to the limitation of this technique in sequencing yield, 674,333 and 511,982 reads from 2D sequencing runs of H1975 and RERF-LC-KJ cells (Table 11) and 5,620,315 reads from 1D and $1D^2$ sequencing runs of LC2/ad cells (Table 12) were generated and mapped to the UCSC hg38 human reference genome.

| Cell line | Run | 1D read | | 2D read | | Total* | Unmapped | Mapped to human genome | Avg. depth | Coverage (≥1×) | Read length | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | pass | fail | pass | fail | | | | | | Mean | Max |
| H1975 | 10 | 42,629 | 291 | 640,277 | 61,363 | 682,209 | 7,876 | 674,333 (98.8%) | 0.7 | 0.46 | 4,815 | 179,616 |
| RERF-LC-KJ | 3 | - | - | 477,280 | 42,680 | 519,960 | 7,978 | 511,982 (98.5%) | 0.58 | 0.36 | 3,627 | 118,237 |

*Table 11 Statistics of H1975 and RERF-LC-KJ cell MinION 2D² sequencing runs*

*Table 12 Statistics of LC2/ad cell MinION 1D and 1D² sequencing runs*

| Cell line | Run | Total* (1D + 1D square) | Unmapped | Mapped to human genome | Avg. depth | Coverage (≥1×) | Read length | |
|---|---|---|---|---|---|---|---|---|
| | | | | | | | Mean | Max |
| LC2/ad | 13 | 6,704,709 | 1,084,394 | 5,620,315 (83.8%) | 6.6 | 0.93 | 6,572 | 2,495,160 |

The physical long reads from MinION covered or partially covered 5,763 (61%) phase blocks in H1975 cells, 4,046 (47%) in RERF-LC-KJ cells and 5,282 (79%) in LC2/ad cells (Table 13). These insufficient coverages were due to the low genome coverage of the MinION sequencing. By examining the combinations of the SNVs in the covered haplotype block, I found that 4,962 (86%), 3,473 (86%) and 4,422 (78%) of H1975, RERF-LC-KJ and LC2/ad cells, respectively, were represented by MinION reads. One of the examples of the validation process is shown in Figure 14, where the phase block covering *SEMA6A* in the H1975 cell line is shown. Taking all the obtained results together, I concluded that the SNV-to-SNV associations inside the phase block were sufficiently accurate for the following analysis.

*Table 13 Statistics of Phase Block Validation by MinION*

| Cell line | H1975 | RERF-LC-KJ | LC2/ad |
|---|---|---|---|
| Flow cell version | R9 + R9.4 | R9.4 | R9.5 |
| Run | 9 (2D passed) + 1 (1D) | 3 (2D) | 3 (1D) + 10 (1D square) |
| Phase block | 9382 | 8623 | 6697 |
| Block covered | 5763 | 4046 | 5282 |
| % block covered | 61.4 | 46.9 | 78.9 |
| SNPs in block | 199,987 | 193,853 | 218,892 |
| SNPs covered | 74,916 | 44,018 | 164,656 |
| % SNPs covered | 37.5 | 22.7 | 75.2 |
| Supported block | 4963 | 3473 | 4422 |
| Not supported block | 800 | 573 | 1260 |
| % supported block | 86.1 | 85.8 | 77.8 |

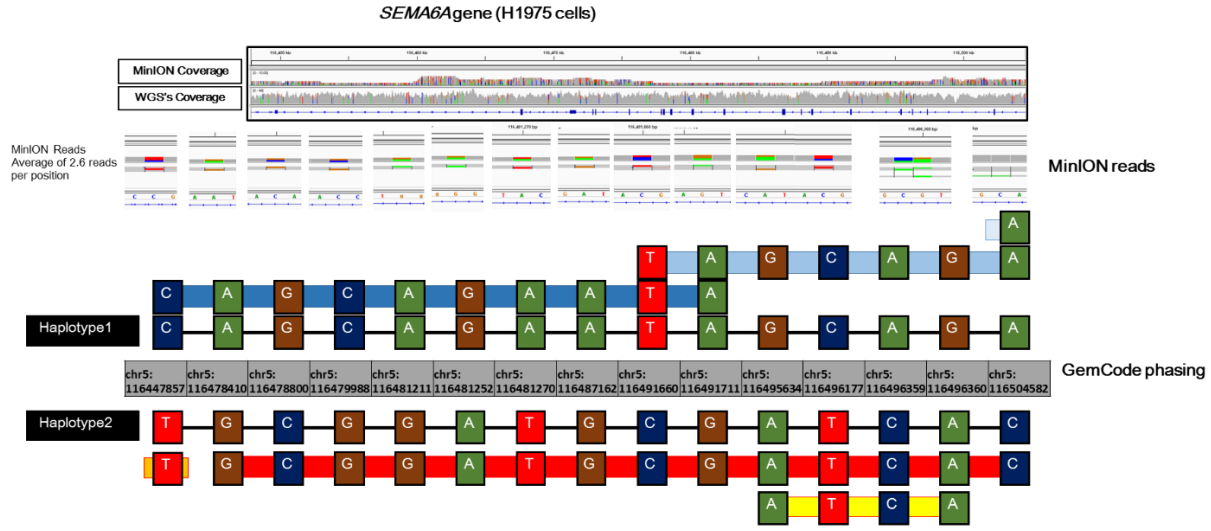



*Figure 14 MinION Validation of 10x GemCode Phase Block Results in SEMA6A in the H1975 Cell Line. 10x Gemcode phasing is shown in center thin black lines; MinION read coverages are shown in blue/light blue for haplotype1 and red/yellow for haplotype2. IGV graphics for each position are shown at the top.*

## Phasing of Regulatory SNVs into Functional Regulatory Mutations

Of the 1,794 phased regulatory SNVs, 137 regulatory SNVs in 146 RefSeq transcripts exhibited ChIP-seq allele imbalances. These SNVs were phased to 166 downstream transcript variants that also exhibited compatible allele imbalance expression patterns (Figure 15, Table 14).

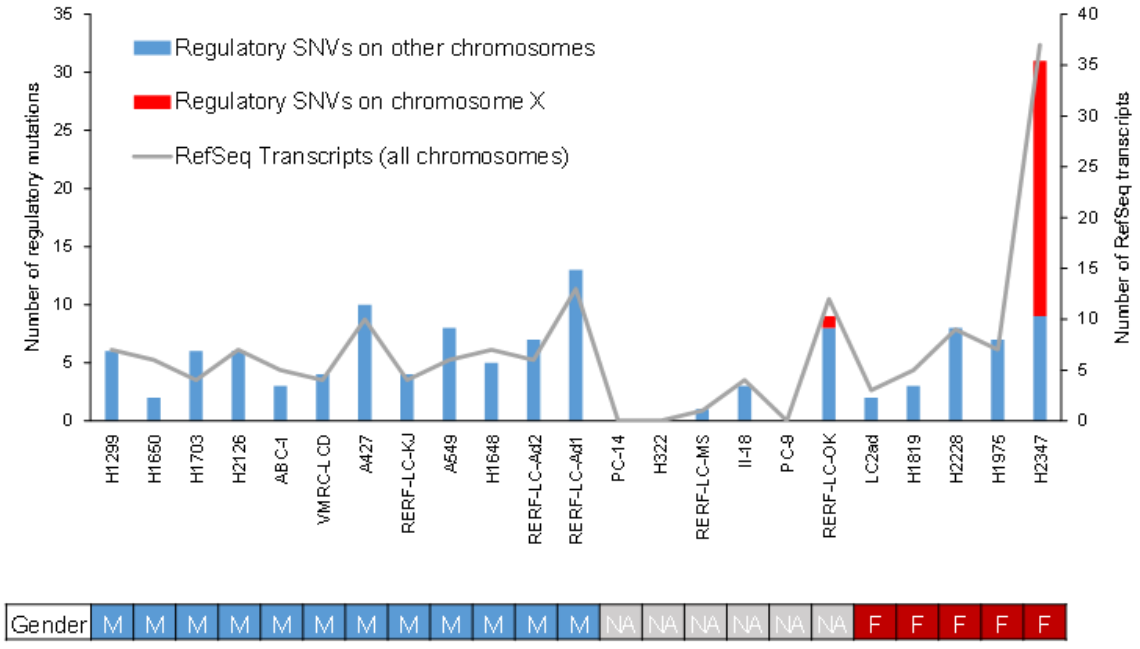



*Figure 15 Imbalanced and Phased RefSeq Transcript Counts in each Cell Line. Numbers of transcripts are shown in grey. Number of regulatory SNVs are shown in red for X-chromosome, blue for others.*

*Table 14 Summary of Imbalanced and Phased Regulatory Mutations and Their Annotations*

| Cell line | Regulatory mutation | | | | Imbalanced ChIP | Gene | RefSeq | TF ChIP-seq (A549; ENCODE) | Transfac (2015.1) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **Chr** | **Position** | **Ref** | **Alt** | | | | | **Loss of TFBS** | **Gain of TFBS** |
| **A427** | chr12 | 54016223 | C | CCCCTAG | H3K4me1 | *HOXC4,HOXC6,* | NM_014620,NM_153693, | - | - | - |
| **A427** | chr15 | 40282532 | C | T | H3K4me3,H3K9/14ac,H3K27ac | *ANKRD63,* | NM_001190479, | *CTCF* | - | - |
| **A427** | chr16 | 1959861 | G | GC | H3K4me3,H3K9/14ac | *RPS2,* | NM_002952, | *TAF1,TCF12,PHF8,* | *AP2ALPHA,SP1SP3* | *BCL6B,EGR1,OSX,WT1,ZN451* |
| **A427** | chr17 | 5191978 | C | G | H3K4me3,H3K27ac | *ZNF594,* | NM_032530, | *SIN3A,ETS1,SP1,CTCF,POLR2A,GABPA,SMC3,ELF1,NR3C1,SIX5,TCF12,MYC,TAF1,RAD21,ZBTB33* | - | *TBX5* |
| **A427** | chr17 | 42702516 | G | GT | H3K36me3 | *CNTNAP1,* | NM_003632, | - | *CPBP,KID3,TCFAP2C* | *GABPA,VMYB* |
| **A427** | chr2 | 85418648 | A | ACG | PolII,H3K27ac | *CAPG,* | NM_001256139, | *SIN3A,ETS1,SP1,POLR2A,PHF8,GABPA,SMC3,ELF1,CEBPB,SIX5,TCF12,MYC,TAF1* | - | *AHRHIF,AHR,EGR1,HES1* |
| **A427** | chr20 | 43658427 | G | GT | H3K9/14ac | *MYBL2,* | NM_002466, | - | - | - |
| **A427** | chr20 | 51803238 | A | AT | H3K9/14ac | *SALL4,* | NM_020436, | - | - | - |
| **A427** | chr3 | 149375499 | T | C | H3K4me3,H3K9/14ac,H3K27ac | *TM4SF1,* | NM_014220, | *SIN3A,ETS1,SP1,ELF1,POLR2A,YY1,TCF12,ZBTB33* | - | - |
| **A427** | chr6 | 3260207 | C | CTT | H3K4me3 | *PSMG4,* | NM_001128592,NM_001128591,NM_001135750, | - | - | - |
| **A549** | chr15 | 23440915 | T | C | H3K27me3 | *MKRN3,* | NM_005664, | - | *GATA1,ZFP64* | *GCMA,GCMB* |
| **A549** | chr15 | 23566678 | T | A | H3K4me3 | *MKRN3,* | NM_005664, | - | *IRF5* | - |
| **A549** | chr16 | 25257835 | C | CG | H3K4me3 | *ZKSCAN2,* | NM_001012981, | *SIN3A,POLR2A,PHF8,REST,GABPA,SREBF1* | *AP2ALPHA* | *WT1,ZN451* |
| **A549** | chr17 | 80355364 | G | T | H3K36me3 | *ENDOV,* | NM_173627,NM_001164638,NM_001164637, | - | *GAF,TEF1* | - |
| **A549** | chr17 | 80355526 | G | A | H3K36me3 | *ENDOV,* | NM_173627,NM_001164638,NM_001164637, | - | *CPBP,ZAC* | *ELF1,SPI1* |
| **A549** | chr2 | 38070511 | C | CA | H3K9/14ac | *CYP1B1,* | NM_000104, | - | - | - |
| **A549** | chr2 | 85418650 | A | ACGCG | H3K27ac | *CAPG,* | NM_001256139, | *SIN3A,ETS1,SP1,POLR2A,PHF8,GABPA,SMC3,ELF1,CEBPB,SIX5,TCF12,MYC,TAF1* | - | - |
| **A549** | chr4 | 188109160 | A | ACG | H3K4me3 | *TRIML2,* | NM_173553, | *CEBPB,* | - | *AHR* |
| **ABC-1** | chr1 | 226679940 | C | CG | H3K27ac | *ITPKB,* | NM_002221, | - | - | *E2F1,GKLF* |
| **ABC-1** | chr10 | 27100911 | C | CA | H3K9/14ac | *ANKRD26,* | NM_014915,NM_001256053, | *ETS1,TAF1* | - | - |
| **ABC-1** | chr19 | 57476910 | C | T | H3K4me3 | *ZNF772,* | NM_001144068,NM_001024596, | - | - | *IRF4,IRF6* |
| **H1299** | chr1 | 66301190 | A | AT | H3K27ac | *PDE4B,* | NM_001037339, | - | - | - |
| **H1299** | chr1 | 113811755 | CGTTTTCCTGCTT | C | H3K4me3,H3K9/14ac,H3K27ac | *RSBN1,* | NM_018364, | - | *EHF,NFAT1,SOX17* | - |
| **H1299** | chr10 | 68334075 | A | AAAC | H3K4me3 | *PBLD,* | NM_001033083,NM_022129, | - | - | *CMYB,FOXA2,FOXD2,FOXD3,FOXG1,FOXI1,FOXJ2,FOXJ3,FOXK1,FOXL1,FOXO1A,FOXO1,FOXO3A,FOXO3,FOXO4,FOXO6,FOXP3,FREAC2,MYB,SOX5,SRY* |
| **H1299** | chr14 | 21069184 | G | A | H3K9/14ac | *ARHGEF40,* | NM_018071, | - | - | - |
| **H1299** | chr14 | 23953737 | A | ACT | H3K27ac | *DHRS4,* | NM_021004, | *POLR2A,TAF1* | - | - |
| **H1299** | chr19 | 36548272 | T | A | H3K36me3 | *ZNF529,* | NM_020951, | - | *GATA3,ZNF333* | *BCL6,MEF2C,TEF3,TEF5* |
| **H1648** | chr12 | 114684071 | G | GGAGA | H3K27ac | *TBX3,* | NM_005996,NM_016569, | *SIN3A,POLR2A,REST,PHF8,SIN3A,YY1,SIX5,NR3C1,CHD1,TAF1,ZBTB33* | - | - |
| **H1648** | chr2 | 131528390 | G | A | H3K4me3 | *CCDC74A,* | NM_138770, | - | - | - |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **H1648** | chr5 | 149002770 | C | CA | H3K27ac | *SH3TC2,* | NM_024577, | - | *FOXC1* | *BRN1* |
| **H1648** | chr5 | 149004604 | G | A | H3K27ac,H3K4me1 | *SH3TC2,* | NM_024577, | - | *CPBP* | - |
| **H1648,H1819** | chr5 | 178941464 | C | CCAAA | H3K4me3,H3K9/14ac | *ZNF454,* | NM_001178089,NM_001178090,NM_182594, | - | - | - |
| **H1650** | chr1 | 110389770 | G | GGA | H3K4me3 | *SLC16A4,* | NM_001201548,NM_001201547,NM_004696,NM_001201549,NM_001201546, | - | - | *AP4* |
| **H1650** | chr18 | 12955468 | T | C | H3K36me3 | *SEH1L,* | NM_031216,NM_001013437, | - | - | - |
| **H1703** | chr11 | 78040678 | T | TA | H3K27ac | *KCTD14,* | NM_023930, | - | - | *CDX1,HOXA13,SATB1* |
| **H1703** | chr4 | 54240616 | C | G | H3K36me3 | *PDGFRA,* | NM_006206, | - | - | *HSF4* |
| **H1703** | chr4 | 54277854 | G | T | H3K4me1 | *PDGFRA,* | NM_006206, | - | *ERG,FLI1,GABPBETA,OSR1* | - |
| **H1703** | chr4 | 54285836 | T | G | H3K36me3 | *PDGFRA,* | NM_006206, | - | - | *STAT3* |
| **H1703** | chr6 | 73394947 | C | G | H3K4me3,H3K9/14ac,H3K27ac | *DDX43,* | NM_018665, | - | - | *GLI2,GLI3,GLI,ZBTB7C* |
| **H1703** | chr7 | 100627013 | C | CTG | H3K4me1 | *TFR2,* | NM_001206855, | *RNF2* | - | *HSF4,KAISO,NF1B* |
| **H1819** | chr1 | 20612675 | G | A | H3K4me1 | *CDA,* | NM_001785, | - | *ING4,KID3* | - |
| **H1819** | chr6 | 27867700 | G | C | PolII,H3K9/14ac,H3K27ac | *HIST1H2AL,* | NM_003511, | - | - | - |
| **H1975** | chr1 | 8374283 | A | C | H3K4me1 | *SLC45A1,* | NM_001080397, | - | *MEF2D,ZFP800* | *HMBOX1,RFX4* |
| **H1975** | chr16 | 56638440 | C | CG | H3K27ac | *MT1A,* | NM_005946, | *POLR2A* | *ZAC* | - |
| **H1975** | chr17 | 64082851 | T | TTA | H3K27ac | *ERN1,* | NM_001433, | *SREBF1* | *GFI1* | *FOXD3,HNF3B,MEF2C,PMX1,ZNF333* |
| **H1975** | chr19 | 9471504 | A | AT | H3K36me3 | *ZNF560,* | NM_152476, | - | - | *PRX2* |
| **H1975** | chr5 | 96877423 | C | CA | H3K9/14ac | *ERAP2,* | NM_022350,NM_001130140, | - | *EGR1,RREB1,WT1* | *PUR1,SMAD5* |
| **H1975** | chr5 | 96896546 | G | GAAA | H3K36me3 | *ERAP2,* | NM_022350,NM_001130140, | - | - | *PARP,SPIB* |
| **H1975** | chr7 | 24830481 | A | G | H3K36me3 | *DFNA5,* | NM_001127453, | - | - | - |
| **H2126** | chr1 | 66265697 | T | TGTGAA | H3K4me1 | *PDE4B,* | NM_001037339, | *TCF12* | - | *SOX10* |
| **H2126** | chr17 | 39056981 | C | CT | H3K27ac | *PLXDC1,* | NM_020405, | *ETS1,SP1,POLR2A,ZBTB33* | - | - |
| **H2126** | chr19 | 48696618 | G | C | H3K9/14ac,H3K27ac | *FUT2,* | NM_001097638,NM_000511, | - | - | - |
| **H2126** | chr19 | 53881131 | C | CAG | H3K27me3 | *PRKCG,* | NM_002739, | *POLR2A* | - | - |
| **H2126** | chr2 | 29114657 | G | C | H3K4me1 | *CLIP4,* | NM_024692, | *REST,ATF3* | - | *BEN* |
| **H2126** | chr2 | 206274726 | C | T | H3K9/14ac,PolII | *ZDBF2,* | NM_020923, | *SIN3A,POLR2A,PHF8,GABPA,TAF1* | - | - |
| **H2228** | chr1 | 27935147 | T | TG | PolII | *SMPDL3B,* | NM_014474,NM_001009568, | - | - | *CREL,SPIB* |
| **H2228** | chr12 | 4518175 | G | A | H3K4me1 | *C12orf4,* | NM_020374, | - | - | *NKX25,NKX28,NKX2B* |
| **H2228** | chr16 | 3443336 | C | T | PolII | *ZNF597,* | NM_152457, | *SIN3A,ETS1,YY1,TAF1* | - | - |
| **H2228** | chr16 | 73070710 | G | GTC | PolII | *ZFHX3,* | NM_001164766, | *SIN3A,SP1,MAZ,POLR2A,NR3C1,ATF3,JUN,GABPA,ELF1,SIN3A,TCF12,MYC,FOSL2,TAF1,JUND,RAD21* | - | - |
| **H2228** | chr17 | 50533812 | C | CT | H3K4me1 | *EPN3,* | NM_017957, | - | *CREB* | - |
| **H2228** | chr20 | 31948814 | C | A | H3K36me3 | *PDRG1,* | NM_030815, | - | *PBX1* | - |
| **H2228** | chr4 | 48173432 | C | CAGGTTGTTTGGT | H3K27ac | *TXK,* | NM_003328, | - | - | *FOXO1A,SRY* |
| **H2228** | chr4 | 87975555 | T | TG | PolII,H3K27ac | *SPP1,* | NM_001251830,NM_000582,NM_001040058,NM_001040060,NM_001251829, | *SIN3A,SP1,TAF1* | *FOXC1,FOXO1A,FOXO1,SOX4,SRY* | - |
| **H2347** | chr11 | 62557963 | T | TG | H3K36me3 | *ROM1,* | NM_000327, | - | *ZNF302* | *HES1,ING4,KID3,USF* |
| **H2347** | chr12 | 47758954 | G | T | H3K9/14ac | *RAPGEF3,* | NM_001098531, | *SIN3A,NR3C1* | *BCL6B,CPBP,EGR1,OSX,SP1SP3,SP1,SP2,SP4,WT1* | *EGR2,KLF17,LKLF* |
| **H2347** | chr2 | 26299918 | G | GA | H3K9/14ac | *GPR113,* | NM_153835, | - | *CIZ* | - |
| **H2347** | chr22 | 37187692 | G | GC | H3K4me3,H3K9/14ac,H3K27ac | *C1QTNF6,* | NM_031910,NM_182486, | - | *BEN* | *ZFP516* |
| **H2347** | chr5 | 157552395 | T | TA | H3K36me3 | *ADAM19,* | NM_033274, | - | *GTF2IRD1* | - |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **H2347** | chr5 | 179023703 | C | CCA | H3K4me3,H3K27ac | *ZNF454,* | NM_001178089,NM_001178090,NM_182594, | *ETS1,SP1,POLR2A,PHF8,ELF1,NR3C1,SIX5,ZBTB33* | - | *ARNT,KID3* |
| **H2347** | chr5 | 179023705 | G | C | H3K4me3,H3K27ac | *ZNF454,* | NM_001178089,NM_001178090,NM_182594, | *ETS1,SP1,POLR2A,PHF8,ELF1,NR3C1,SIX5,ZBTB33* | - | *ISL2,NKX31,NKX32* |
| **H2347** | chr5 | 179023789 | C | T | H3K27ac | *ZNF454,* | NM_001178089,NM_001178090,NM_182594, | *ETS1,SP1,POLR2A,PHF8,ELF1,NR3C1,SIX5,ZBTB33* | - | *ELK1,ER81,ETV7,ZSCAN2* |
| **H2347** | chr6 | 28399927 | C | CT | H3K4me3 | *ZSCAN12,* | NM_001163391, | *SIN3A,ETS1,SP1,POLR2A,ATF3,PHF8,GABPA,YY1,SIX5,TAF1,ZBTB33* | - | - |
| **H2347** | chrX | 11115164 | G | T | H3K36me3 | *HCCS,* | NM_005333,NM_001122608,NM_001171991, | - | - | *P50,SALL3,STAT6* |
| **H2347** | chrX | 40049931 | A | G | H3K36me3 | *BCOR,* | NM_017745,NM_001123385, | - | - | - |
| **H2347** | chrX | 46574956 | C | T | H3K4me3,H3K27ac | *CHST7,* | NM_019886, | - | - | - |
| **H2347** | chrX | 47483521 | C | CA | H3K4me3,H3K9/14ac,H3K27ac | *ZNF41,* | NM_153380,NM_007130, | *REST,TAF1* | *BCL6B,GKLF,WT1,ZFP740* | - |
| **H2347** | chrX | 48528539 | G | T | H3K36me3 | *EBP,* | NM_006579, | - | - | - |
| **H2347** | chrX | 48539949 | G | A | H3K4me3,H3K27ac | *TBC1D25,* | NM_002536, | *SIN3A,CTCF,PHF8,* | - | *KID3* |
| **H2347** | chrX | 48559236 | C | A | H3K36me3 | *TBC1D25,* | NM_002536, | - | *BCL6B,PAX2* | - |
| **H2347** | chrX | 49002196 | G | T | H3K4me3,H3K9/14ac,H3K27ac | *GRIPAP1,* | NM_207672,NM_020137, | *SIN3A,ETS1,SP1,POLR2A,REST,GABPA,CEBPB,YY1,SIX5,TCF12,CHD1,TAF1* | *BEN* | - |
| **H2347** | chrX | 65534259 | A | T | H3K4me3,H3K27ac | *LAS1L,* | NM_001170649,NM_031206,NM_001170650, | - | *GATA3,GATA5,HOXA10,ZBTB44* | - |
| **H2347** | chrX | 65534494 | C | A | H3K4me3,H3K27ac | *LAS1L,* | NM_001170649,NM_031206,NM_001170650, | *REST,KDM5A,* | - | - |
| **H2347** | chrX | 71256219 | C | T | H3K27ac | *ZMYM3,* | NM_201599, | - | - | *PMX1* |
| **H2347** | chrX | 100665623 | C | A | H3K36me3 | *SRPX2,* | NM_014467, | - | - | - |
| **H2347** | chrX | 100665625 | G | T | H3K36me3 | *SRPX2,* | NM_014467, | - | - | - |
| **H2347** | chrX | 100822341 | C | A | H3K36me3 | *CSTF2,* | NM_001325, | - | - | - |
| **H2347** | chrX | 103586064 | G | T | H3K4me3,H3K9/14ac,H3K27ac | *TCEAL4,* | NM_001006937,NM_001006935,NM_024863, | *PHF8,YY1,TAF1* | - | *SOX4* |
| **H2347** | chrX | 119399107 | C | CT | H3K4me3,H3K9/14ac | *SLC25A43,* | NM_145305, | *REST* | - | - |
| **H2347** | chrX | 119399109 | A | T | H3K4me3,H3K9/14ac,H3K27ac | *SLC25A43,* | NM_145305, | *REST* | - | *HSF4* |
| **H2347** | chrX | 120603552 | C | T | H3K27ac | *C1GALT1C1,* | NM_001011551,NM_152692, | - | - | - |
| **H2347** | chrX | 132957182 | C | A | H3K4me3 | *HS6ST2,* | NM_147175,NM_001077188, | - | - | *NR1B2* |
| **H2347** | chrX | 132957183 | G | GA | H3K4me3 | *HS6ST2,* | NM_147175,NM_001077188, | - | - | - |
| **H2347** | chrX | 136504229 | G | A | H3K36me3 | *HTATSF1,* | NM_001163280, | - | *PLZFB* | - |
| **H2347** | chrX | 155216488 | G | C | H3K4me3,H3K27ac | *VBP1,* | NM_003372, | *SP1,POLR2A,REST,PHF8,YY1,TAF1* | *REX1,YY1,YY2* | *NFMUE1* |
| **II-18** | chr1 | 220733283 | C | A | H3K27ac | *MARC2,* | NM_017898, | *SP1,CTCF,MAZ,GABPA,SMC3,ELF1,TCF12,RAD21* | *RELA* | *HSF4,P50RELAP65* |
| **II-18** | chr11 | 64203784 | C | CTTTG | H3K36me3 | *FERMT3,* | NM_178443,NM_031471, | - | - | *FOXM1,FOXO1A,HFH2,HNF3G,SRY* |
| **II-18** | chr12 | 105084230 | T | TTTTG | H3K4me3,H3K9/14ac,H3K27ac | *ALDH1L2,* | NM_001034173, | *SIN3A,YY1* | - | *FOXD3,HFH3,HNF3* |
| **LC2ad** | chr7 | 94655777 | G | GAAA | H3K4me3 | *PEG10,* | NM_001172437,NM_001172438, | *SIN3A,PHF8* | - | - |
| **LC2ad** | chr7 | 122303951 | G | C | H3K27ac | *FEZF1,* | NM_001024613,NM_001160264, | - | *OSR1,OSR2* | *ARNT* |
| **RERF-LC-Ad1** | chr1 | 8422756 | T | G | H3K9/14ac,H3K27ac | *SLC45A1,* | NM_001080397, | - | *PRRX2* | - |
| **RERF-LC-Ad1** | chr10 | 69222380 | T | TCA | H3K4me1 | *HKDC1,* | NM_025130, | *CTCF,CEBPB* | - | *ZBTB44* |
| **RERF-LC-Ad1** | chr11 | 75403270 | A | T | H3K36me3 | *KLHL35,* | NM_001039548, | - | - | - |
| **RERF-LC-Ad1** | chr18 | 79395018 | C | G | H3K4me3 | *NFATC1,* | NM_172390,NM_006162,NM_172388, | *RNF2* | - | *CETS1P54,CETS1,EHF,ELK1,ER71,ER81,ERM,ETV7,FLI1,GABP* |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | | | | *AGABPB,GABPA,GADP,PEA3,PEA3* |
| **RERF-LC-Ad1** | chr19 | 20545301 | G | GT | H3K9me3 | *ZNF737,* | NM_001159293, | - | - | - |
| **RERF-LC-Ad1** | chr19 | 20565417 | G | GCA | H3K27ac | *ZNF737,* | NM_001159293, | - | - | - |
| **RERF-LC-Ad1** | chr19 | 35003143 | C | CGT | H3K4me1 | *ZNF792,* | NM_175872, | *SIN3A,ETS1,SP1,ATF3,SIX5* | - | - |
| **RERF-LC-Ad1** | chr2 | 225582441 | G | C | H3K36me3 | *NYAP2,* | NM_020864, | - | *BTEB2,E2F3,EGR1,MOVOB,SP1,SP2,WT1* | *GKLF,MAZR,MAZ,SP1SP3,WT1,ZFP281,ZFP740,ZN451* |
| **RERF-LC-Ad1** | chr20 | 22582148 | G | A | H3K36me3 | *FOXA2,* | NM_021784, | *POLR2A* | - | *TFAP2A* |
| **RERF-LC-Ad1** | chr3 | 179244703 | G | GA | H3K36me3 | *KCNMB3,* | NM_171829, | - | - | - |
| **RERF-LC-Ad1** | chr6 | 26197055 | G | A | H3K4me3,PolII,H3K9/14ac,H3K27ac | *HIST1H3D,* | NM_003530, | - | *MYB* | - |
| **RERF-LC-Ad1** | chr6 | 26199147 | G | C | H3K9/14ac | *HIST1H3D,* | NM_003530, | - | - | - |
| **RERF-LC-Ad1** | chr6 | 158031620 | G | GTC | H3K4me1 | *SYNJ2,* | NM_001178088, | *SIN3A,SP1,NR3C1,TCF12,FOSL2* | - | - |
| **RERF-LC-Ad2** | chr1 | 17303761 | C | CTCCTCTGAG | H3K4me1 | *PADI4,* | NM_012387, | - | - | - |
| **RERF-LC-Ad2** | chr11 | 308885 | G | GC | H3K4me1 | *IFITM2,* | NM_006435, | - | - | *LFA1* |
| **RERF-LC-Ad2** | chr15 | 89785543 | A | AAC | H3K27ac | *ANPEP,* | NM_001150, | *NR3C1* | - | - |
| **RERF-LC-Ad2** | chr16 | 81096914 | G | C | H3K4me1 | *GCSH,* | NM_004483, | *EHMT2* | *KID3* | - |
| **RERF-LC-Ad2** | chr16 | 81097354 | G | C | H3K4me1 | *GCSH,* | NM_004483, | *EHMT2* | - | - |
| **RERF-LC-Ad2** | chr19 | 57709211 | G | GC | H3K4me3 | *ZNF154,* | NM_001085384, | - | - | - |
| **RERF-LC-Ad2** | chr5 | 150139698 | A | AC | H3K4me1 | *PDGFRB,* | NM_002609, | - | - | *AML1,PEBP2B* |
| **RERF-LC-KJ** | chr1 | 155932383 | G | GA | H3K36me3 | *RXFP4,* | NM_181885, | - | *SIX1* | - |
| **RERF-LC-KJ** | chr10 | 46130177 | G | T | H3K36me3 | *AGAP7,* | NM_001077685, | - | *NKX25,RBPJK,ZNF860* | *BCL6* |
| **RERF-LC-KJ** | chr19 | 52602057 | G | GAA | H3K9/14ac | *ZNF701,* | NM_018260, | - | - | *BARBIE* |
| **RERF-LC-KJ** | chr19 | 53408011 | C | T | H3K36me3 | *ZNF765,* | NM_001040185, | - | - | - |
| **RERF-LC-MS** | chr1 | 182586389 | C | A | H3K4me3 | *RNASEL,* | NM_021133, | - | *LEF1,TCF3* | *HELIOSA* |
| **RERF-LC-OK** | chr11 | 68772512 | T | TG | H3K36me3 | *MTL5,* | NM_001039656,NM_004923, | - | *RFX* | - |
| **RERF-LC-OK** | chr15 | 89781340 | C | G | H3K27ac | *MESP1,* | NM_018670, | - | - | *CETS1,ELF1,ELF5,ELK1,ESE1,ETS1,ETS,PEA3,PU1,SPI1,SPIB,SPIC* |
| **RERF-LC-OK** | chr5 | 96884962 | T | TG | H3K36me3 | *ERAP2,* | NM_022350,NM_001130140, | - | - | - |
| **RERF-LC-OK** | chr5 | 96895154 | T | TA | H3K36me3 | *ERAP2,* | NM_022350,NM_001130140, | - | - | - |
| **RERF-LC-OK** | chr5 | 149446233 | G | GT | PolII | *PCYOX1L,* | NM_024028, | *SP1,POLR2A,NR3C1,FOXA1* | - | - |
| **RERF-LC-OK** | chr7 | 151345978 | G | T | H3K4me1 | *NUB1,* | NM_001243351,NM_016118, | - | - | - |
| **RERF-LC-OK** | chr9 | 131396736 | T | C | H3K4me1 | *PRRC2B,* | NM_013318, | - | - | *EGR3,KID3* |
| **RERF-LC-OK** | chr9 | 131528708 | G | A | H3K36me3 | *UCK1,* | NM_001135954,NM_031432, | - | - | - |
| **RERF-LC-OK** | chrX | 51743749 | C | T | H3K4me3 | *GSPT2,* | NM_018094, | *PHF8,GABPA* | *ZBED6* | - |
| **VMRC-LCD** | chr11 | 94768533 | C | T | H3K4me3 | *AMOTL1,* | NM_130847, | - | - | *RELA* |
| **VMRC-LCD** | chr12 | 94149889 | G | A | H3K27ac | *PLXNC1,* | NM_005761, | - | *KLF17,LKLF* | - |
| **VMRC-LCD** | chr15 | 74853540 | C | G | H3K36me3 | *ULK3,* | NM_001099436, | - | *BEN* | *GKLF* |
| **VMRC-LCD** | chr7 | 73827962 | T | TTAGTCACTTCTG | PolII,H3K27ac | *WBSCR27,* | NM_152559, | *POLR2A,NR3C1,MYC,FOSL2* | - | *AP1FJ,AP1* |

To illustrate functional interpretations, a putative regulatory mutation in the *ZDBF2* region in the H2126 cell line is shown in Figure 16. The regulatory mutation C>T at chr2:206274726 exhibited a ChIP-seq imbalance in RNA-polymerase II and H3K9/14ac assays. This SNV was phased to the coding variant A>G at chr2:206305007 for which RNA-seq imbalance expression was also found.

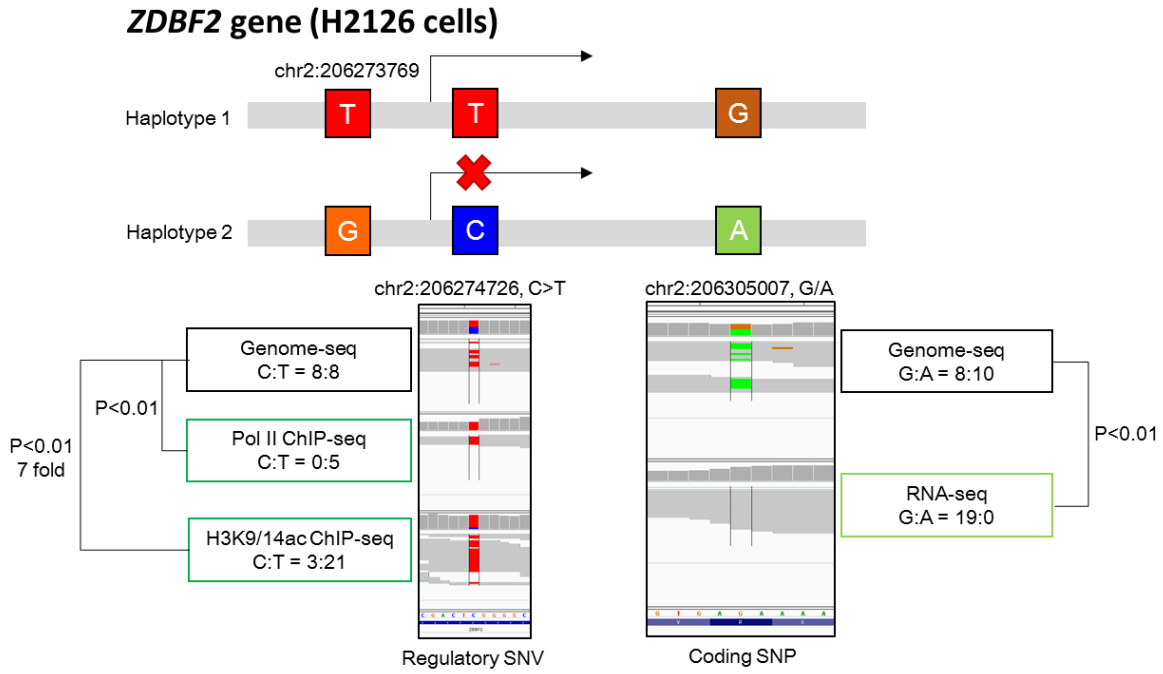



*Figure 16 Effects of Regulatory Mutations on ZDBF2 in H2126 Cells. (Top) Two haplotypes are made from phasing. (Bottom) Allele imbalances reveal that only haplotype1 (T-T-G) is active.*

I next examined whether the dysregulation caused by these putative regulatory mutations directly affect oncogenes. *PDGFRA* is a known mutation and amplification target in lung cancer. In H1703 cells, both gene amplification and regulatory mutations in *PDGFRA* were found (Figure 17). Phasing analysis suggested that the variants found were organized into two haplotypes: (1) G-T-T-T-G-G-G-C and (2) C-C-G-C-A-C-C-T. WGS variant frequencies also suggested gene amplification of the (2) allele. In agreement with copy number changes, allelic imbalance analysis revealed a heavy bias towards (2) in both ChIP and RNA-seq, suggesting regulatory effects of the mutations acting on the (2) allele. These two observations suggested regulatory mutations with gain of function in this gene.

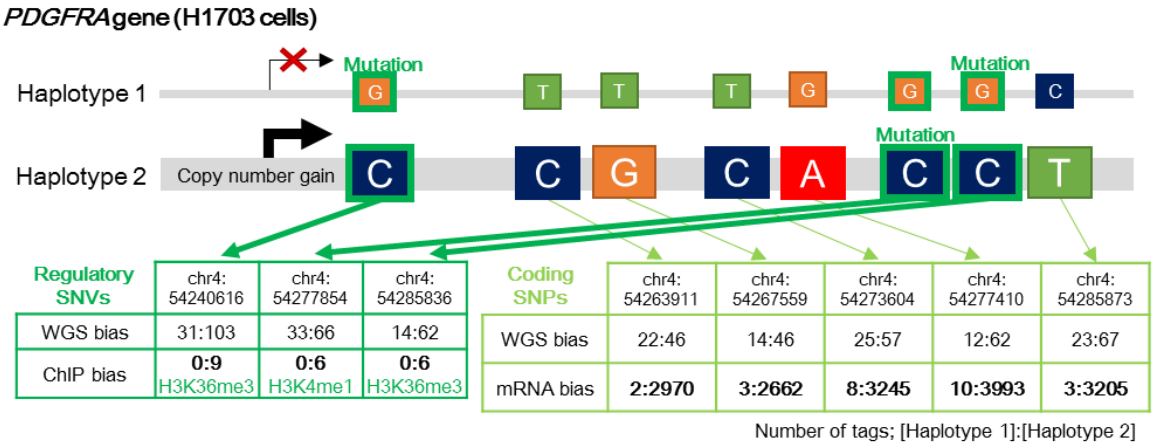



*Figure 17 Gene Amplifications and Regulatory Mutation in PDGFRA in the H1703 Cell Line. (Top) Two haplotypes are made from phasing with both regulatory and coding variants identified. (Bottom, arrows) WGS biases reveal copy number gain in haplotype 2; ChIP-seq and RNA-seq biases also reveal mono-allele activation of haplotype 2.*

To further investigate the 137 regulatory mutations for potential biological functions, I cross-referenced them with the FANTOM CAT database (v3 robust release). Thirty-one of the mutations overlapped with lncRNA-associated regions, and 5 overlapped with FANTOM5 permissive enhancer regions (Table 15).

*Table 15 Regulatory Mutations in the FANTOM CAT database*

| CELL LINE | CHROMOSOME | POSITION (UCSC HG38) | REFERENCE | ALTERNATIVE | HISTONE MARKS |
|---|---|---|---|---|---|
| **A427** | chr6 | 3260207 | C | CTT | H3K4me3 |
| **A427** | chr3 | 149375499 | T | C | H3K27Ac, H3K4me3, H3K914Ac |
| **A427** | chr12 | 54016223 | C | CCCCTAG | H3K4me1 |
| **A549** | chr16 | 25257835 | C | CG | H3K4me3 |
| **A549** | chr17 | 80355364 | G | T | H3K36me3 |
| **A549** | chr17 | 80355526 | G | A | H3K36me3 |
| **ABC-1** | chr10 | 27100911 | C | CA | H3K914Ac |
| **ABC-1** | chr1 | 226679940 | C | CG | H3K27Ac |
| **H1299** | chr1 | 66301190 | A | AT | H3K27Ac |
| **H1299** | chr14 | 23953737 | A | ACT | H3K27Ac |
| **H1648** | chr12 | 114684071 | G | GGAGA | H3K27Ac |
| **H1648** | chr2 | 131528390 | G | A | H3K4me3 |
| **H1703** | chr7 | 100627013 | C | CTG | H3K4me1 |
| **H1703** | chr11 | 78040678 | T | TA | H3K27Ac |
| **H2126** | chr1 | 66265697 | T | TGTGAA | H3K4me1 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **H2126** | chr19 | 53881131 | C | CAG | H3K27me3 | |
| **H2126** | chr2 | 206274726 | C | T | H3K914Ac, Pol II | |
| **H2228** | chr1 | 27935147 | T | TG | Pol II | |
| **H2228** | chr17 | 50533812 | C | CT | H3K4me1 | |
| **H2347** | chr11 | 62557963 | T | TG | H3K36me3 | |
| **H2347** | chr6 | 28399927 | C | CT | H3K4me3 | |
| **H2347** | chr2 | 26299918 | G | GA | H3K914Ac | |
| **II-18** | chr12 | 105084230 | T | TTTTG | H3K27Ac, H3K4me3, H3K914Ac | |
| **LC2AD** | chr7 | 122303951 | G | C | H3K27Ac | |
| **RERF-LC-AD1** | chr19 | 20565417 | G | GCA | H3K27Ac | |
| **RERF-LC-AD1** | chr6 | 26197055 | G | A | H3K27Ac, H3K4me3, H3K914Ac, Pol II | |
| **RERF-LC-AD1** | chr6 | 26199147 | G | C | H3K914Ac | |
| **RERF-LC-AD1** | chr18 | 79395018 | C | G | H3K4me3 | |
| **RERF-LC-AD1** | chr10 | 69222380 | T | TCA | H3K4me1 | |
| **RERF-LC-AD2** | chr5 | 150139698 | A | AC | H3K4me1 | |
| **RERF-LC-KJ** | chr19 | 53408011 | C | T | H3K36me3 | |

| **CELL LINE** | Chromosome | Position (UCSC hg38) | Reference | Alternative | Histone Marks | FANTOM5 enhancer regions |
|---|---|---|---|---|---|---|
| **ABC-1** | chr1 | 226679940 | C | CG | H3K27Ac | chr1:226867580-226868182; |
| **H2228** | chr16 | 73070710 | G | GTC | Pol II | chr16:73104505-73104770; |
| **H2347** | chr2 | 26299918 | G | GA | H3K914Ac | chr2:26522632-26522929; |
| **RERF-LC-AD1** | chr6 | 158031620 | G | GTC | H3K4me1 | chr6:158452388-158452885; |
| **RERF-LC-OK** | chr15 | 89781340 | C | G | H3K27Ac | chr15:90324413-90324614; |

## *Cis*-regulatory mutations causing transcriptional dysregulations

Another indication for functionally relevant regulatory mutations is alterations of sequence motifs, creating and removing transcriptional factor (TF) binding sites or CpG methylation sites.

To characterize the involvement of CpG sites, the locations of the 137 regulatory mutations were referenced with CpG site regions. For transcriptional factor binding sites, data on the anti-TF ChIP-seq in ENCODE for the A549 cell line were referenced. Twenty-nine of the mutations were found within CpG sites, possibly disrupting DNA methylation modifications (Figure 18A). Forty-nine of the regulatory mutations (Figure 18B) were

found to be located within the supposed TF binding sites. For example, the A427 cell line had a mutation in the promoter region of *ZNF594* (C>G; chr17:5191978). According to the TF binding site data in ENCODE, this mutation occurred in the binding site of three functionally important TFs, *POLR2A*, *TAF1* and *MYC*. The binding consensus sequence of these TFs may be disrupted by this mutation.

To systematically analyze the effect on TF binding sites, disruptions or generations of TF binding motifs within ±10 bp of the regulatory mutations were investigated by using the TRANSFAC database. Eighty-four of the 137 regulatory mutations were predicted, with 24 causing loss of TF binding sites, 40 generating novel TF binding sites and 20 resulting in replacements (Figure 18C, Table 14). Such alterations in TF binding influence downstream transcripts. For example, repressor activity was found in the promoter mutation of the *SLC16A4* gene in the H1650 cell line. This mutation (G>GAA; chr1:110389770) was predicted to generate a novel TF binding site for AP-4. Allele expression analysis suggested that the mutant allele is silenced; thus, it was likely that this novel TF binding site acted as a repressor (Figure 19A). As an example of an activator effect, a regulatory mutation in the *NFATC1* gene in the RERF-LC-Ad1 cell line (C>G; chr18:79395018) was predicted to create a novel TF binding site for ETS family transcriptional factors. Allele expression analysis showed that the mutant allele was fully expressed (Figure 19B). A total of 104 such regulatory mutations were identified.

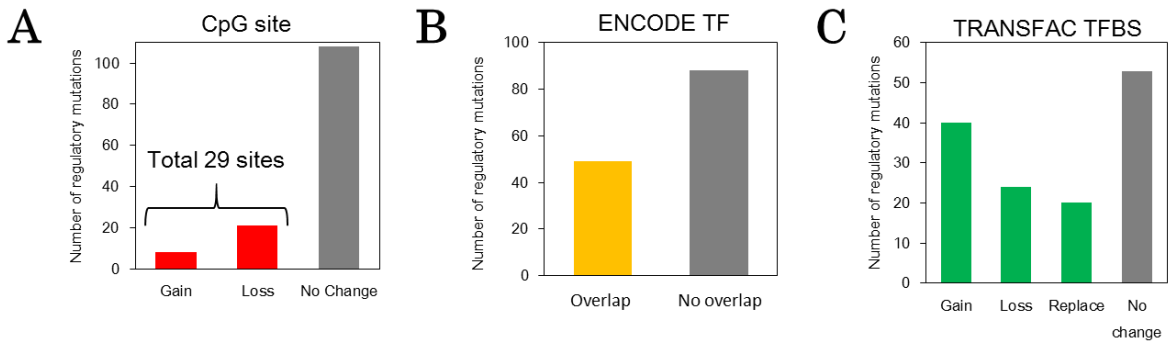



*Figure 18 Functional Analysis of Regulatory Mutations. (A) twenty-nine of the mutations overlap with CpG sites. (B) Forty-nine fall within TF binding sites. (C) Eighty-four are predicted to disrupt or create TF motifs.*

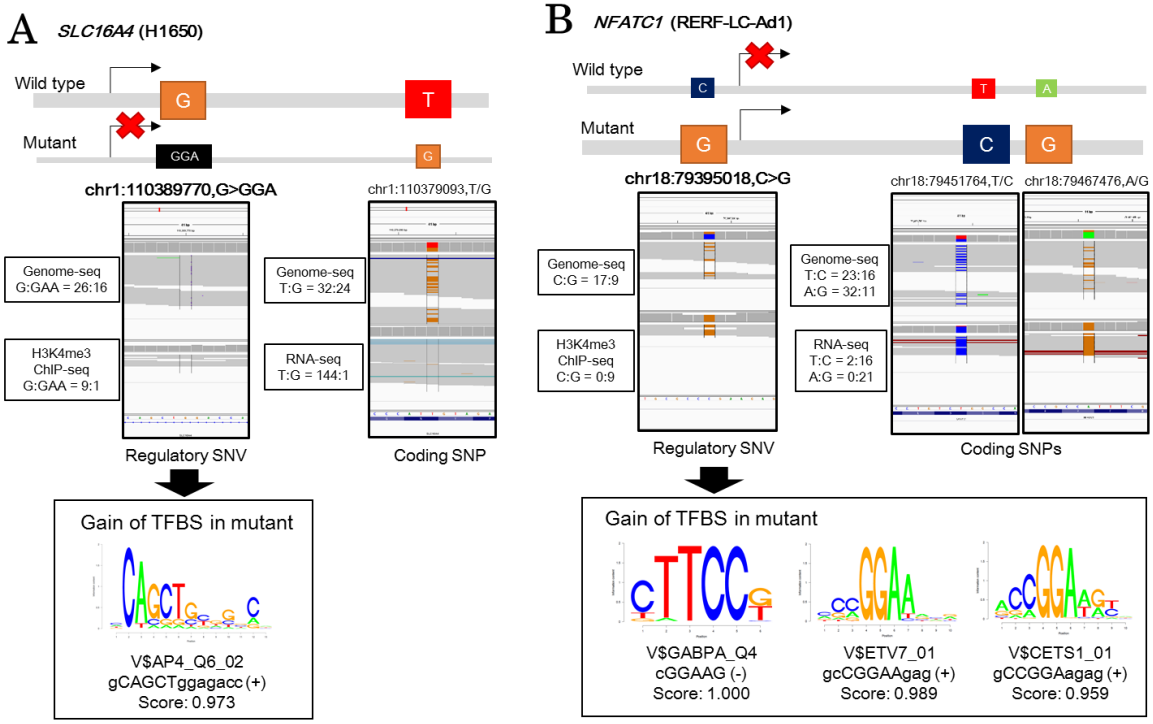



*Figure 19 Effects of TFBS Creation. (A) Repressor in SLC16A4 in H1650 cells; (B) Activator in NFATC1 in RERF-LC-Ad1 cells.*

To biologically confirm the novel TF binding sites predicted by TRANSFAC, I focused on the *NFATC1* promoter mutation (C>G; chr18:79395018) in RERF-LC-Ad1 cells. The SNV was predicted to create a novel ETS family TF binding site. Using a luciferase assay, I experimentally validated the change in promoter activity of the motifs. I compared the abilities of the mutant motif (GCCGGAA) and the wild-type motif (GCCCGAA) to drive reporter gene expression. The mutant allele exhibited a 3-fold increase in reporter activities compared to the wild-type allele (Figure 20 A, B). To further verify this result, I performed Sanger sequencing of ETS1 ChIP-enriched DNA fragments to compare the affinities of the alleles to ETS1, a major TF in the ETS family. Using 3 primers to target ±100 bp regions from the mutation (Figure 20A), 4- to 6-fold enrichments in ETS1 binding compared to the input DNA control were observed. This value was comparable to that obtained for the *RPS26* gene, a positive control (Figure 20C, D).

*NFATC1* is a known transcription factor that plays a major role in the immune response and T-cell activation against cancer cells (Heim et al., 2018). In addition,

*NFATC1* has been reported to function as both an oncogene and a tumor suppressor gene in cancer (Mancini & Toker, 2009; Robbs, Cruz, Werneck, Mognol, & Viola, 2008; S. Xu et al., 2018; W. Xu et al., 2016). The important roles of *NFATC1* may be reflected in survival analysis in TCGA-LUAD projects. From a dataset of 506 patients with overall survival data and 401 patients with disease-free survival data, high expression of *NFATC1* was associated with better overall survival (Figure 21A, Kaplan-Meier; p=0.0049) but worse disease-free survival (Figure 21B, Kaplan-Meier; p=0.0003). Further analysis of these observations will be discussed in Chapter II.

From 146 RefSeq transcripts under the influence of regulatory mutations, 31 were found to be significantly associated with survival outcomes in the TCGA-LUAD project (Figure 22). These findings reinforced the importance of genes under aberrant regulation and the notion that mutations in regulatory regions of the genome could play biologically and clinically significant roles in cancer cells.

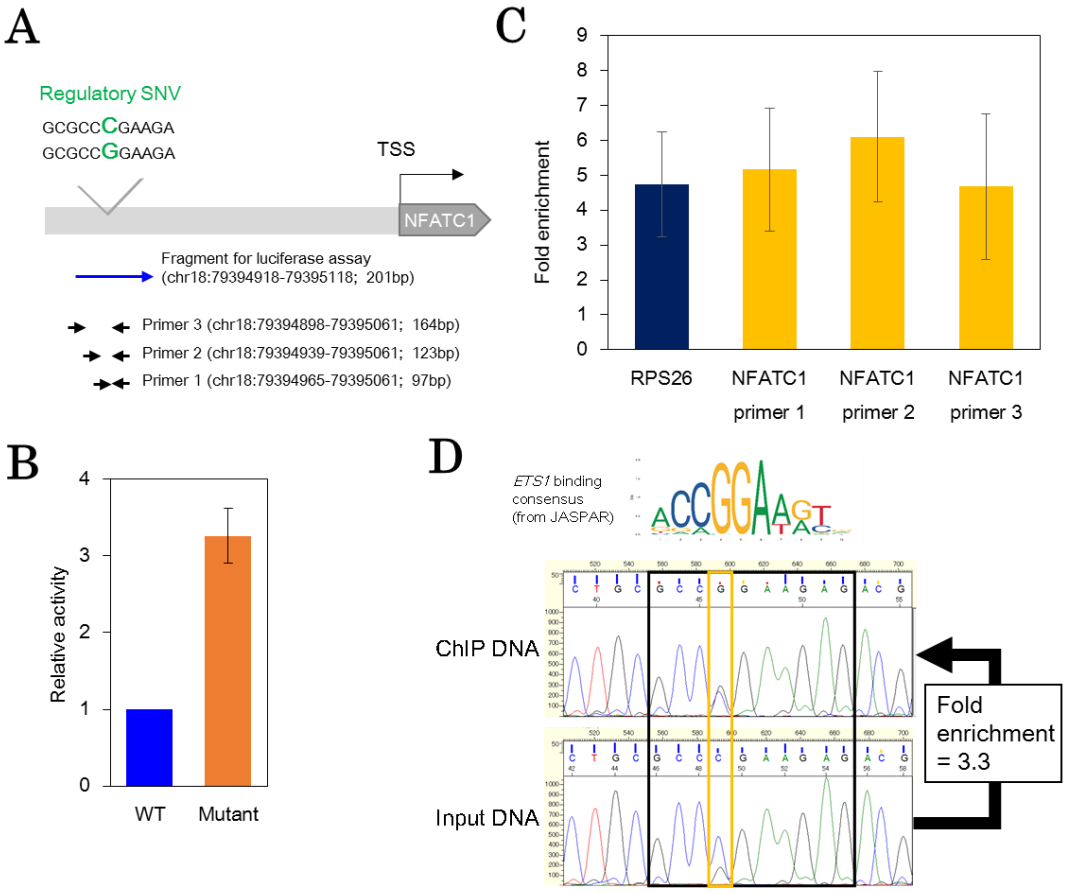



*Figure 20 Biological validation of the NFATC1 regulatory mutation. (A) DNA primer and fragments used for validation. (B) Luciferase assay reveal 3 times higher activities for mutant fragment. (C) qPCR of ETS1-ChIP products show comparable enrichments from NFATC1 primers to RPS26 (positive control). (D) Sanger sequencing of ETS1 ChIP-seq products reveal 3 fold enrichments of mutant motifs compare to control*

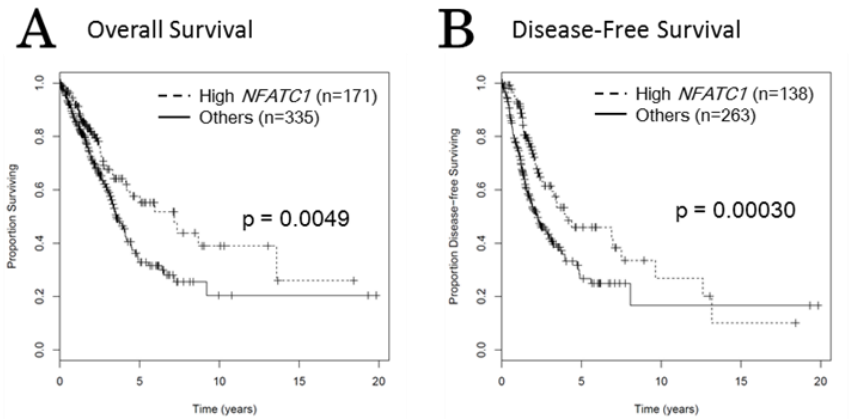
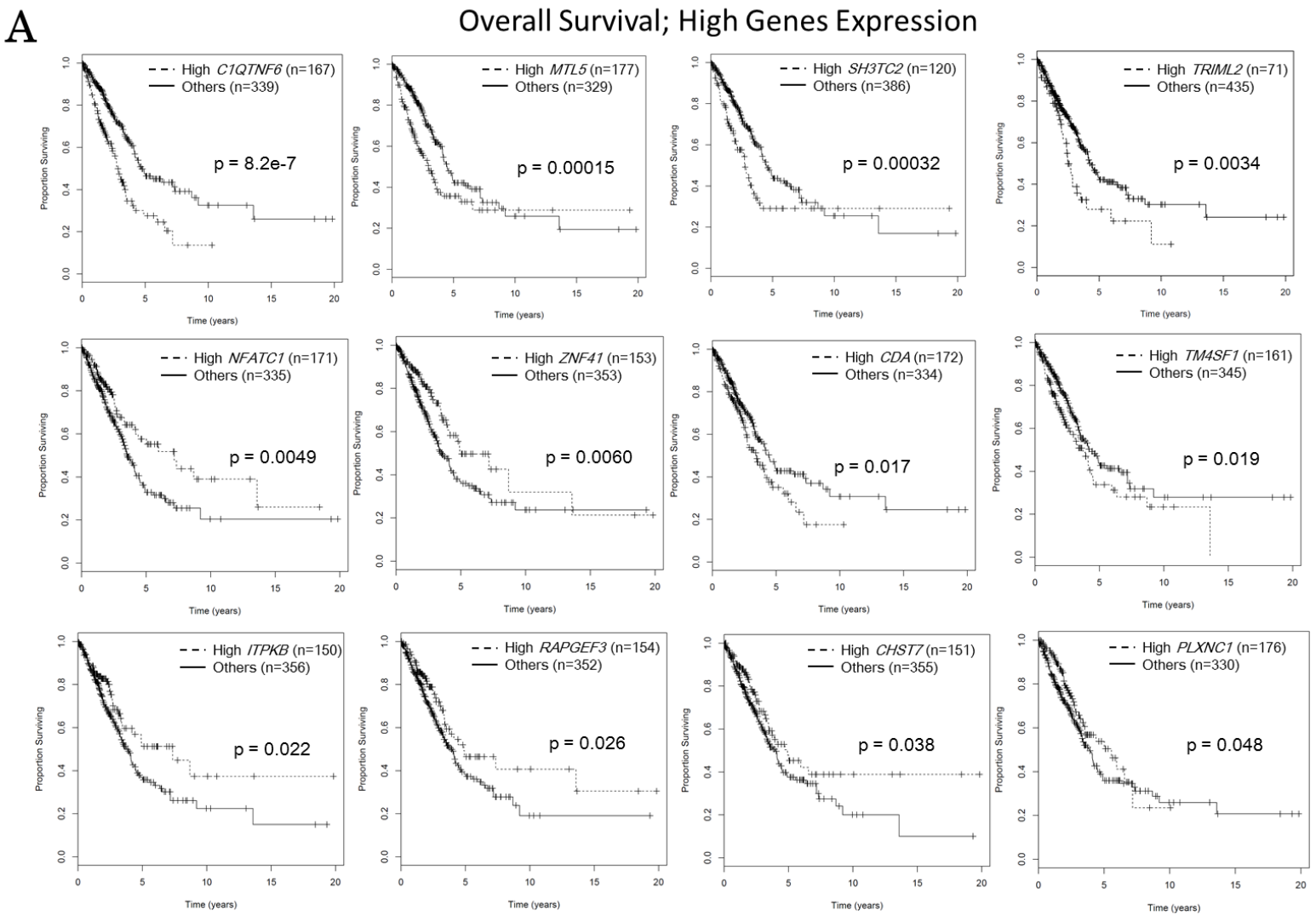


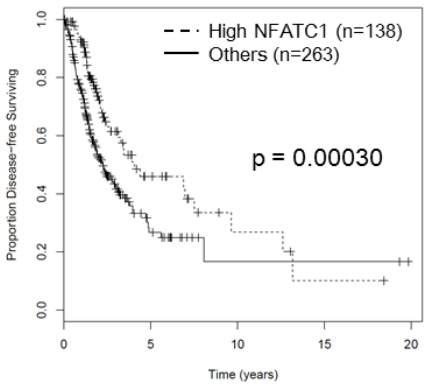

*Figure 21 Kaplan-Meier Plot of NFATC1 expression and Patients Survival in TCGA-LUAD. (A) Overall survival. (B) Disease-Free Survival.*

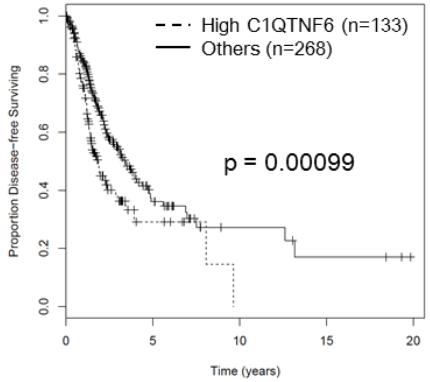
*Figure 22 Associations of Patient Survival Outcomes with 31 Genes with Regulatory Mutations in TCGA-LUAD. (A) Overall survival with higher expression (12 genes). (B) Overall survival with lower expression (10 genes). (C) Disease free survival with higher expression (15 genes). (D) Disease free survival with lower expression (10 genes).*

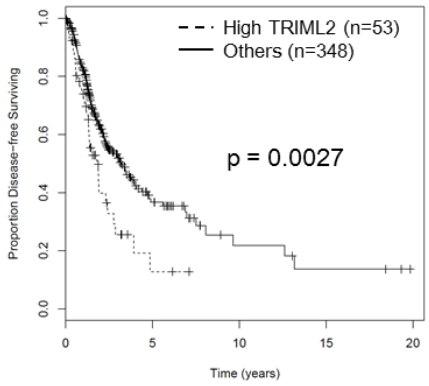


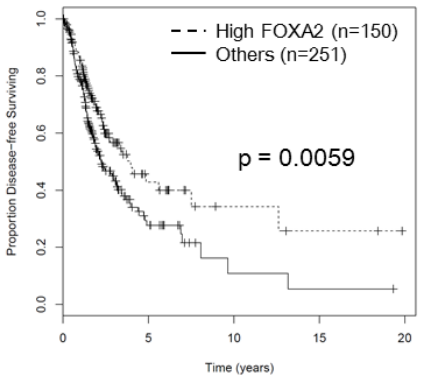
B

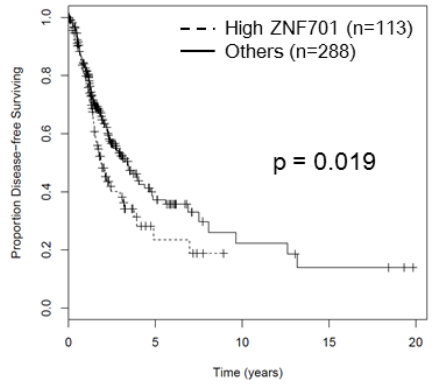
# Disease-Free Survival; High Genes Expression

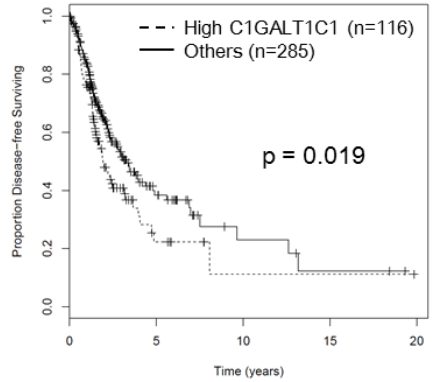


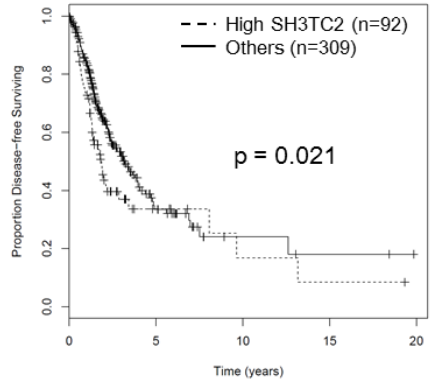



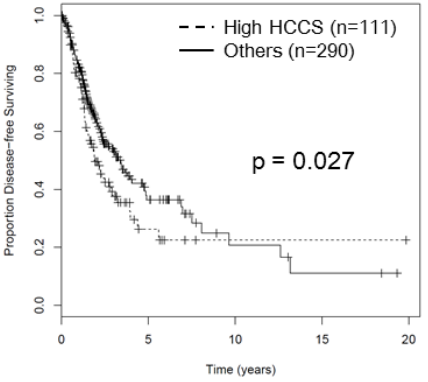



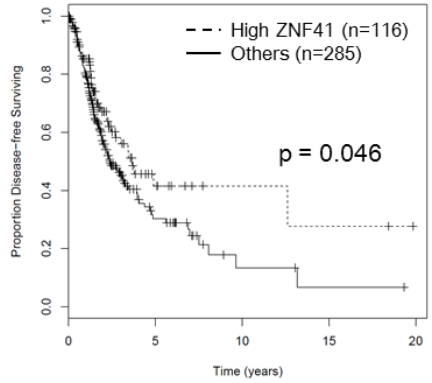



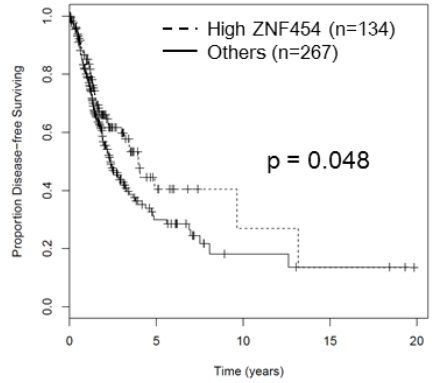



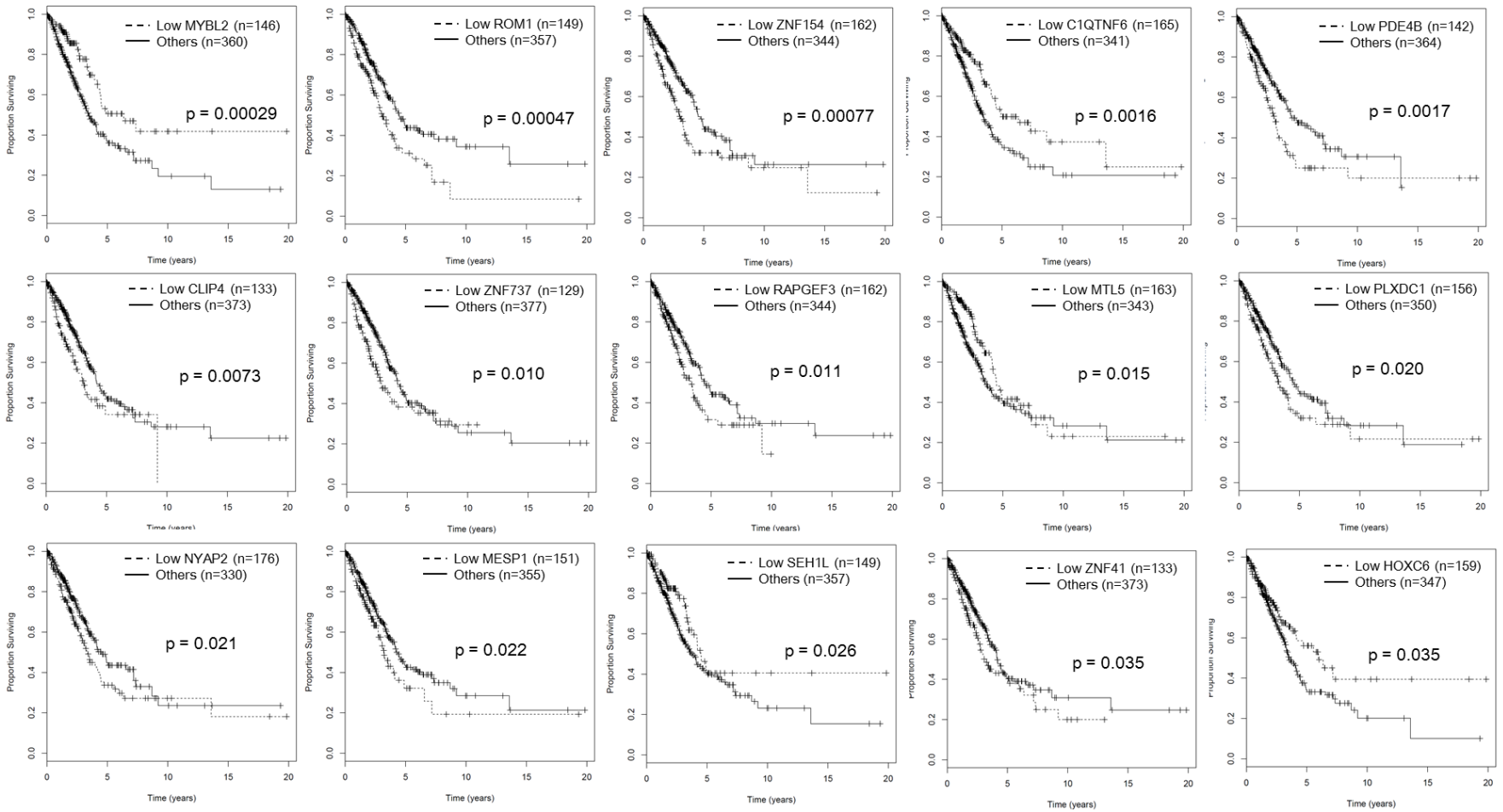



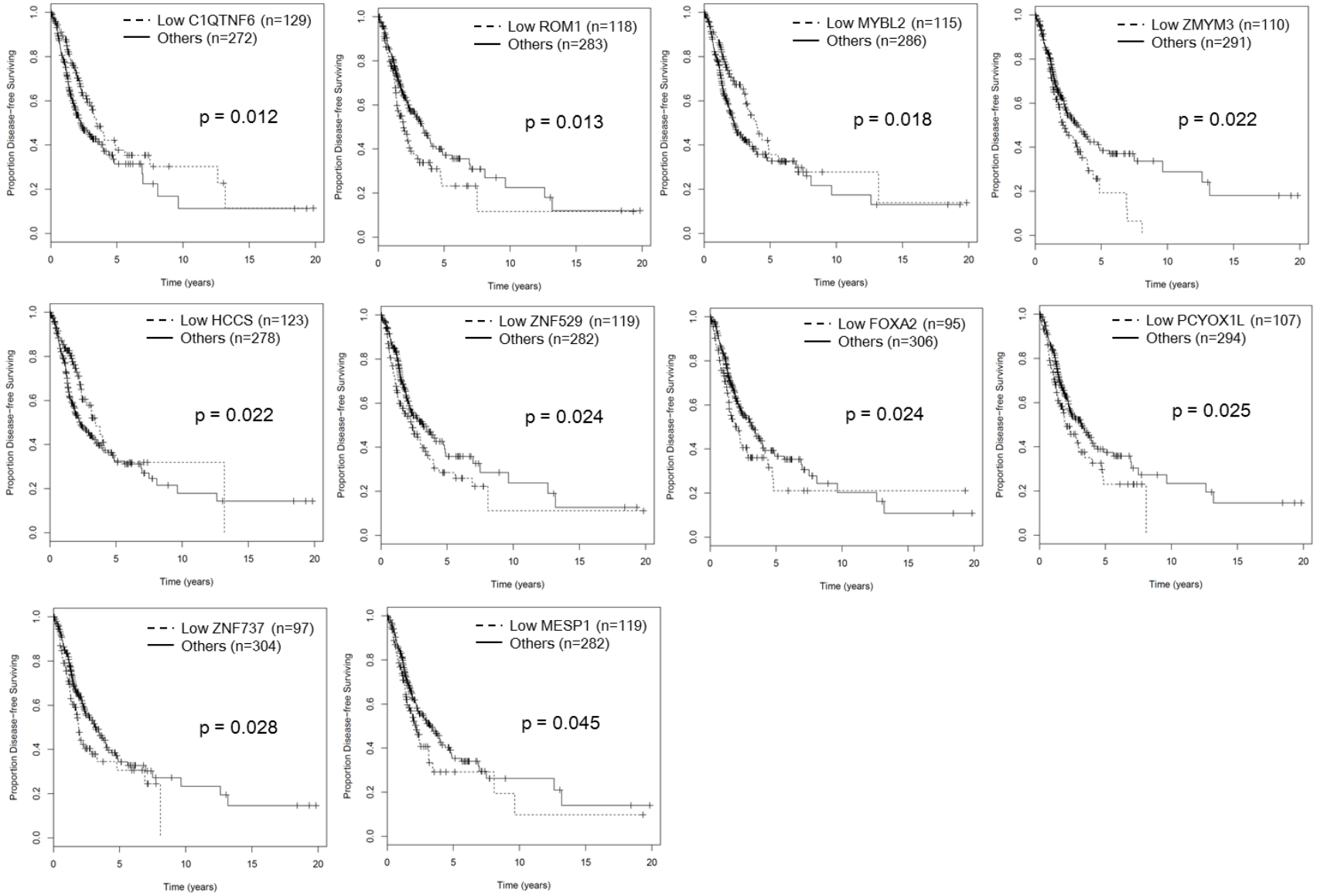



C


Overall Survival; Low Genes Expression
Low MYBL2 (n=146)
Others (n=360)
p = 0.00029
Low ROM1 (n=149)
Others (n=357)
p = 0.00047
Low ZNF154 (n=162)
Others (n=344)
p = 0.00077
Low C1QTNF6 (n=165)
Others (n=341)
p = 0.0016
Low PDE4B (n=142)
Others (n=364)
p = 0.0017
Low CLIP4 (n=133)
Others (n=373)
p = 0.0073
Low ZNF737 (n=129)
Others (n=377)
p = 0.010
Low RAPGEF3 (n=162)
Others (n=344)
p = 0.011
Low MTL5 (n=163)
Others (n=343)
p = 0.015
Low PLXDC1 (n=156)
Others (n=350)
p = 0.020
Low NYAP2 (n=176)
Others (n=330)
p = 0.021
Low MESP1 (n=151)
Others (n=355)
p = 0.022
Low SEH1L (n=149)
Others (n=357)
p = 0.026
Low ZNF41 (n=133)
Others (n=373)
p = 0.035
Low HOXC6 (n=159)
Others (n=347)
p = 0.035
Proportion Surviving
Time (years)

**D**

# Disease-Free Survival; Low Genes Expression



## Discussion

In this chapter, I identified a total of 137 regulatory mutations that potentially have functional consequences in 146 downstream RefSeq transcripts. Various functional aspects of these mutations were explored. In addition, 31 potential involvement sites in lncRNA-associated regions, disruption of 29 CpG sites and 49 TF binding sites were identified. The motif analysis revealed 24 losses of motifs, 40 gains of motifs and 20 replacements. Thirty-one genes were also associated with patient prognosis. Biological and clinical significance sets these regulatory mutations apart from the numerous but not functional "Passenger Mutations" (Vineis, Schatzkin, & Potter, 2010). However, because it is possible that germline variants in the cell lines are not registered in dbSNP database, functional germline variants might be remaining in the final results. Further filtering could be done by cross-referencing with mutation hotspot regions reported in TCGA, however the power of this approach is still limited due to small sample size and focuses on the coding regions. Further analysis with more complete hotspot regions would prove to be interesting in the future. These results were based on the systemic interrogations of multi-omics datasets and long read sequencing results. This approach could be adapted easily with the advent of new technologies and could serve as a stepping stone in further larger and more comprehensive studies in the same direction.

This study has some limitations. The first obvious drawback is the discrepancy between the number of allelic imbalance variants and phased variants. Here, only 137 out of 1,794 SNVs (7.6%) were phased. This was most likely due to lack of coverage by WES plus the regulome bait used in 10x GemCode synthetic long read sequencing. While the bait was designed to widely capture regulatory regions from various cell linages and phenotypes, it was becoming clearer that regulatory elements were organized in a highly tissue specific or sometimes even in a cell type specific manner (Heinz, Romanoski, Benner, & Glass, 2015). This produces a major challenge in designing a single universal bait to capture the entire regulatory landscape.

Moreover, due to the novelty of long read technologies, difficulties in applications were encountered. The 10x GemCode system was originally designed to handle diploid genomes. It was observed that allele phasing regions with CNA lesions were problematic. CNA lesions are prominent in cancer and usually hold important onco- and tumor suppressor genes. Analyzing phasing from 10x GemCode is less useful in cancer settings. Phasing of the whole genome with further developments in overcoming focal and large copy number aberrations should be considered. CNA could be interrogated by MinION sequencers, but their lower accuracy, precision and throughput limit their use in large-scale mutation studies. Further developments in physical long read technologies and usage in tandem with conventional sequencing methods are promising combinations.

During interpretations, annotations had to be done with some prior knowledge of the mutations or the region around the mutations themselves. This included information such as the landscape of TF binding sites and sequence motifs. In this work, only 50 ENCODE ChIP-seq datasets, which account for only 41 types of TFs in one cell line, were utilized. With over 1,600 TFs in human cells (Lambert et al., 2018), these numbers represented only a meager fraction of the TF library. Sequence motif analysis could provide extra clues, but 3D structure, DNA conformation or the expression of TFs could not be considered. The lack of accurate predictions and the need for experimental validations limit analyses on a larger scale. Although 33 out of 137 mutations (24%) were left uncharacterized, these results were considered satisfactory.

I consider the first chapter as a proof of concept study, answering how regulatory mutations could be identified in a systemic manner with multi-omics allele imbalance analysis and haplotype phasing.

Beyond cell lines, the next challenge is to identify and analyze the regulatory actions in the clinical setting. As the cell lines were heavily transformed during repeated rounds of culturing, their current phenotypes were difficult to pinpoint (Y. Liu et al., 2019), which increased the difficulty in interpretation and hindered network-wide level interaction analyses.

To address this lack of a systematic in vivo analysis, which is beyond the scope of model cell culture systems, I aimed to conduct a follow-up study on the concept outlined in this chapter. In Chapter II, I shift the focus from a small set of specialized datasets to the more general and large-scale TCGA database.

# Chapter II: Pan-cancer Multi-omics Network Analysis in The Cancer Genome Atlas

## Introduction

In the previous chapter, I demonstrated that the interactions between regulatory elements and their downstream counterparts could be investigated and elucidated by integrative analysis of multi-omics studies. However, such detailed studies require multiple sample matching specialized assays. To detect interactions at the network scale, large and diverse genotypic information is crucial. TCGA projects publish genotypic information for thousands of donors and include matching RNA-seq and DNA methylation arrays, thus enabling investigations of these 2 omics interactions. Each TCGA project also focuses on distinct cancer origins and morphologies, and this phenotypic information could also directly contribute to network construction and annotation, providing one extra layer in the identification and interpretation of the interactions. In this chapter, I will explore the 2-omics network interactions in the TCGA datasets.

Gene network analysis is a powerful tool that is frequently used in the interpretation and comprehension of vast and sophisticated biological systems. Many tools have been developed to fulfill this crucial role (Delgado & Gómez-Vela, 2019) with many approaches proposed. The most traditional and less computationally demanding is the coexpression network analysis. This approach is relatively cheap and simple, but it has drawbacks in generalizing and imputing the results into different settings. To address these shortcomings, many methods have been developed based on modeling, such as the ordinary differential model, Bayesian probabilistic model and neural network model. However, no single model has demonstrated robustness in large-scale network analysis with potentially multiple modes of regulation mixed in. Without depending on general prior knowledge of the biological system, designing any one model to encompass the entire gene network is unlikely to be successful; thus, I chose the simpler and more flexible coexpression network approach as the starting point.

Many software programs have been developed based on coexpression networks, with Weight Gene Co-expression Network Analysis (WGCNA) (Langfelder & Horvath, 2008) being the most widely used. For multi-omics inputs, iCluster (Shen, Olshen, & Ladanyi, 2009) is frequently used. However, there were a number of reasons for each that rendered them unsuitable.

WGCNA is one of the most commonly used software suites for gene module network analysis. It is based on fitting the data into a scale-free network model and using hierarchical clustering based on Pearson correlation to cluster similar genes. Then, the clustering data are classified into modules by its unique tree-cutting algorithm. The modules are then optimized by remerging and re-cutting. While WGCNA could be modified to cluster 2-omics networks, the scale-free network nature of the genes and the methylation sites were not necessarily presented. Moreover, while tree cutting provides a flexible framework, the exact nature of the cut is still set globally. It is not always possible to achieve optimal parameters, especially with 2-omics as inputs. For this reason, integration of 2-omics would be hindered.

iCluster focuses on classifying samples by multi-omics assay. It is based on the joint latent variable model, which aims to classify and cluster the samples using multi-omics data. For iCluster, each omics is modeled separately and treated as a single enclose aspect in sample classification. Finally, the samples are clustered in a joint variance-based model not focusing on the genotypic interactions within. With none of these approaches suitable for my analysis, I decided to devise an approach to uniformly integrate and construct the 2-omics networks.

To start off, I decided to adapt my multi-omics analysis approach to use with the TCGA dataset. This is not straightforward, as accurate allele resolution information is not available due to the lack of “Long Read Sequencing” or phasing information. However, collections of multiple individual genotypes also enable another analysis approach. In the same manner as the allele imbalance expression of a single sample, when viewed across multiple samples, genotypic features that interact with each other should exhibit

synchronized expressions across the samples. These interactions could then be linked into networks based on their phenotypic activities. To capture these synchronizations and activities, I based my approach on rank analysis of the expression level in RNA-seq and beta-value in methylation array.

# Material and Methods

## TCGA projects used in this study

Eight TCGA projects with a total of 4,116 matching RNA-seq (Exp-S) and Methylation Array (Meth-A) donors were retrieved from the ICGC data portal site (https://dcc.icgc.org/releases). The projects were picked based on 2-omics data availability, project sizes and varieties of tumor histology (squamous carcinoma, adenocarcinoma and melanoma, Table 16). Methylation arrays were obtained from reported beta-values of Infinium HumanMethylation450K files.

*Table 16 TCGA projects used to construct 2-omics networks*

| Project | Aberrations | Total Donors | Donors with EXP-S | Donors with Meth-A | Donors with Both | Specimens with Both |
|---|---|---|---|---|---|---|
| **Breast Cancer - Ductal & lobular USA** | BRCA-US | 1,093 | 1,041 | 1,013 | 1,012 | 1,130 |
| **Cervical Cancer - Cervical squamous cell carcinoma USA** | CESC-US | 307 | 259 | 243 | 242 | 246 |
| **Gastric Cancer – Adenocarcinoma USA** | STAD-US | 443 | 418 | 443 | 415 | 415 |
| **Head And Neck Cancer - Squamous cell carcinoma USA** | HNSC-US | 528 | 481 | 494 | 480 | 502 |
| **Lung Cancer - Squamous cell carcinoma USA** | LUSC-US | 502 | 428 | 427 | 424 | 432 |
| **Lung Cancer – Adenocarcinoma USA** | LUAD-US | 518 | 478 | 481 | 473 | 496 |
| **Colon Cancer – Adenocarcinoma USA** | COAD-US | 459 | 428 | 424 | 420 | 464 |
| **Skin Cancer - Cutaneous melanoma USA** | SKCM-US | 470 | 430 | 430 | 427 | 431 |
| **Total** | | 4,320 | 3,963 | 3,955 | 3,893 | 4,116 |

## RNA expression data

mRNA expression levels were calculated from the reported RSEM value of each gene in TCGA level 3 data. Missing genes across the different projects were treated as genes with 0 expression. Genes with an average RSEM of less than $10^{-6}$ were removed with the intention of removing missing data.

## Methylation Level

A total of 12,835 CpG methylation sites in Infinium HumanMethylation450K chips were selected based on site locations within ±10 kb of the TSS of any of the transcripts

annotated by Illumina. TSS positions were retrieved from the KERO database using the UCSC hg38 human genome reference (http://kero.hgc.jp/). Beta values were viewed as continuous values representing the fraction of methylated alleles or chances of finding methylated alleles without considering actual allele configurations.

## Rank covariance-based distance

The RSEM and beta-values for 15,666 genes and 12,835 CpG sites were combined to create a 2-omics matrix with 28,501 features with 4,116 sample elements. For each feature row, the measurement of each sample was then ranked from lowest (rank 1) to highest (rank 4,116), and this matrix was then treated as a 2-omics uniform measurement system.

For each feature pair, distances between the pairs were based on treating variances as the best possible achievable covariance (variance of x is the covariance of x against itself). Covariance represent how well the rank permutation between the pairs aligned. The distance between the two measures represents how far away the pair alignment is from the perfect permutation (and also represents a portion of unexplained variances between the pair). To normalize for ties, the distances between each pairing are then the difference of variance and covariance normalized by the variance.

$$\text{Distance}(x, y) = \frac{Variance(x) * Variance(y) - Covariance(x, y)^2}{Variance(x) * Variance(y)} \quad \ldots\ldots\ldots (1)$$

$$\text{Distance}(x, y) = 1 - Spearman\ R^2 (= \text{ Unexplained Variances of Ranks})$$

## Clustering of synchronized features into units

Features with either similar or opposite rank permutations were clustered together into a tree structure based on a rank correlation-based distance matrix using hierarchical clustering with the unweighted pair group method with arithmetic mean (UPGMA) algorithm (Figure 23A, B).

Functional units, defined as groups of features with synchronized changes in their measurements, were picked up by cutting the UPGMA tree into groups of features that

every pairwise distance determined to come from the same distributions using the k-samples Anderson-Darling test. At each linkage, distances between groups of features that were in the left and right leaves were checked for heterogeneities in 3 groups of distances, represented by areas under the curve (AUC) between cumulative distribution function (CDF) in 3-sample Anderson-Darling (AD) test. The 3 groups of distances were #1, the internal distances between features in a group in the left leaf; #2, the internal distances between features of a group in the right leaf; and #3, every pairwise distance between features in the left and right leaves. Pairing with a smaller AUC, which resulted in a p-value greater than 0.01, was treated as a homogenized unit (Figure 23C). Pairing with a sufficiently large AUC between the CDFs that resulted in a p-value less than 0.01 was considered heterogeneous in origin and treated as a separated unit (Figure 23D), and both were carried over to the next level of linkage in the same arm.

If the left and/or right leaves contained more than 1 cluster, every pairing between the left and right would be checked and prioritized based on a smaller AUC (larger p-value) pairing, and if one of the left or right leaves contained only 1 feature, then the Anderson-Darling test would only be done with 2 applicable groups, but if both the left and right leaves contained only 1 feature, both would be considered homogeneous by default.

## Functional Unit Phenotype Activity Analysis

Phenotype activities of each functional unit were determined from the contribution of each donor (phenotype) rank to that unit. Because functional units were constructed from homogenizations of the coefficient of determination ($R^2$) and $R^2$ itself could be expressed by a portion of covariances over variances, each donor's term during covariance calculation could be viewed as that donor's contribution to the correlation. By averaging the contributions in every pair of correlations in the unit, each donor contribution to the unit could be determined (Figure 24). Consistently high or low ranking donors in the unit would produce high donor contributions, and thus donor contributions would align with extreme donors, representing phenotype activities. To account for the overall correlation ($R^2$), the

donor contributions were viewed as vectors for each unit, with the vector size representing the overall correlation in that unit.

## Linking units with similar phenotypes activities into networks

Units with similar phenotype activities were linked into networks based on an angular distance matrix derived from donor contribution vectors by hierarchical clustering with the UPGMA algorithm and Anderson-Darling test ($p<0.01$) as tree cutting methods as described earlier. The final results were the networks of units with similar donor contributions and thus similar extremes and phenotype activities.

An angular distance matrix was constructed by calculating every pairwise unit angular distance between the sample contribution vectors as follows:

$$cosine\ similarity(\mathrm{x},\mathrm{y}) = \frac{Donor\ Contribution_x * Donor\ Contribution_y}{\|Donor\ Contribution_x\| * \|Donor\ Contribution_y\|}$$

$$\text{angular distance}(\mathrm{x},\mathrm{y}) = \frac{\cos^{-1}(cosine\ similarity(x,y))}{\pi}$$

By using angular distance, the influences of the unit overall correlation ($R^2$, vector magnitude) in each vector were normalized, enabling comparisons of donor contributions in different units (vector direction).

## Database cross referencing

Networks were cross referenced with MSigDB gene sets (Liberzon et al., 2011) for characterizations (Accessed October 2019, http://software.broadinstitute.org/gsea/msigdb/index.jsp). Collections of hallmark gene sets (H), curated gene sets (C2), computational gene sets (C4), GO gene sets (C5) and oncogenic gene sets (C6) were used. For each network, gene sets in each collection were checked for overrepresentation in a 2×2 contingency table with Fisher's exact test and FDR < 0.05 by the Benjamini-Hochberg procedure. Individual manual curation of each network was then performed by using gene sets and correlation strength as guides.

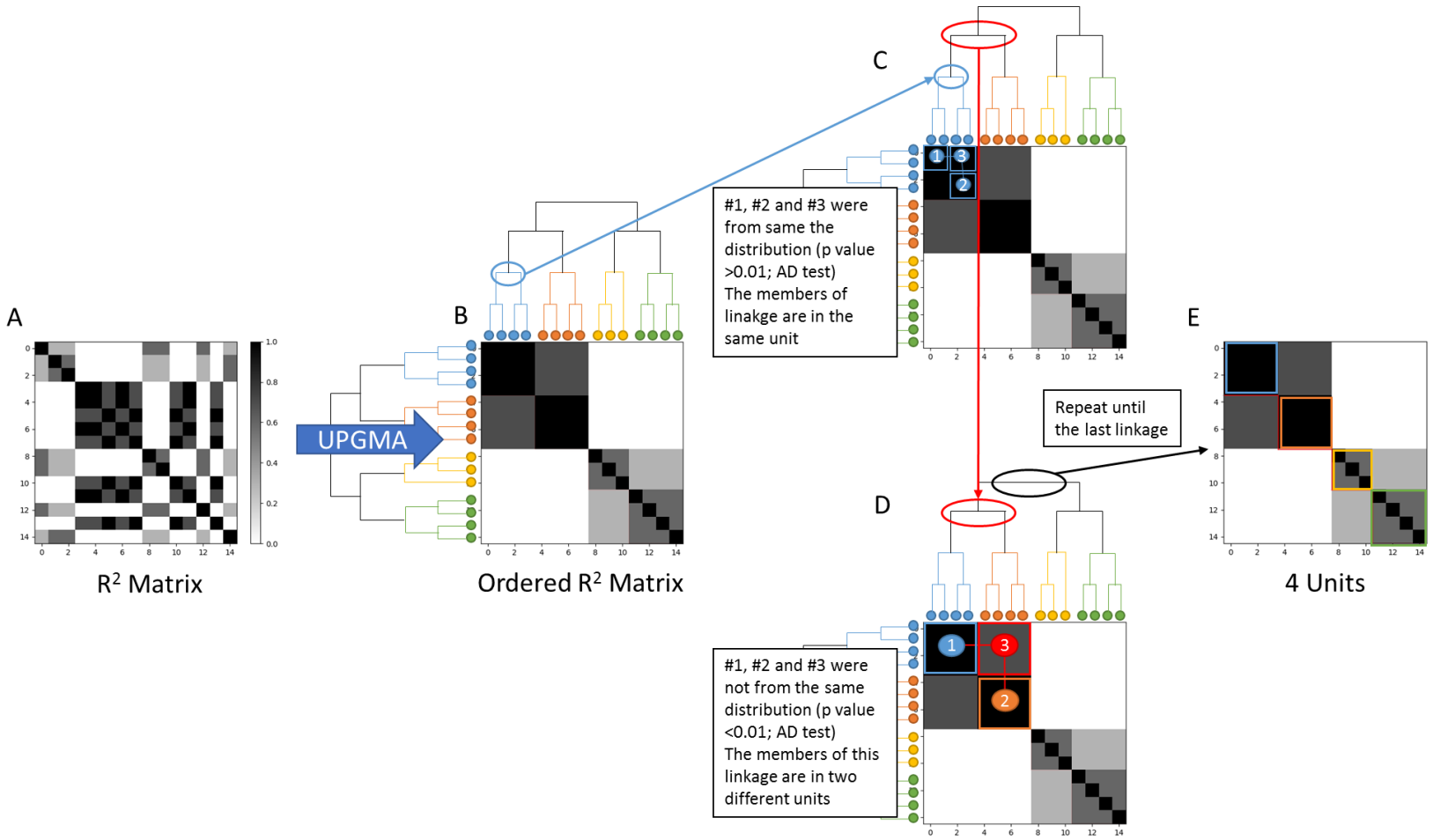



*Figure 23 2-omics Clustering Methods. (A) Pre-cluster Example $R^2$ matrix. (B) $R^2$ matrix after clustering with UPGMA. (C) Linkages with left, right and between distances from the same distribution (k-sample Anderson-Darling test; p>0.01) are merged. (D) Linkages with left, right and between distances from different distribution (k-sample Anderson-Darling test; p<0.01) are not merged with both arms carry-over to the next level. (E) Final grouping is retrieved at the last linkage.*

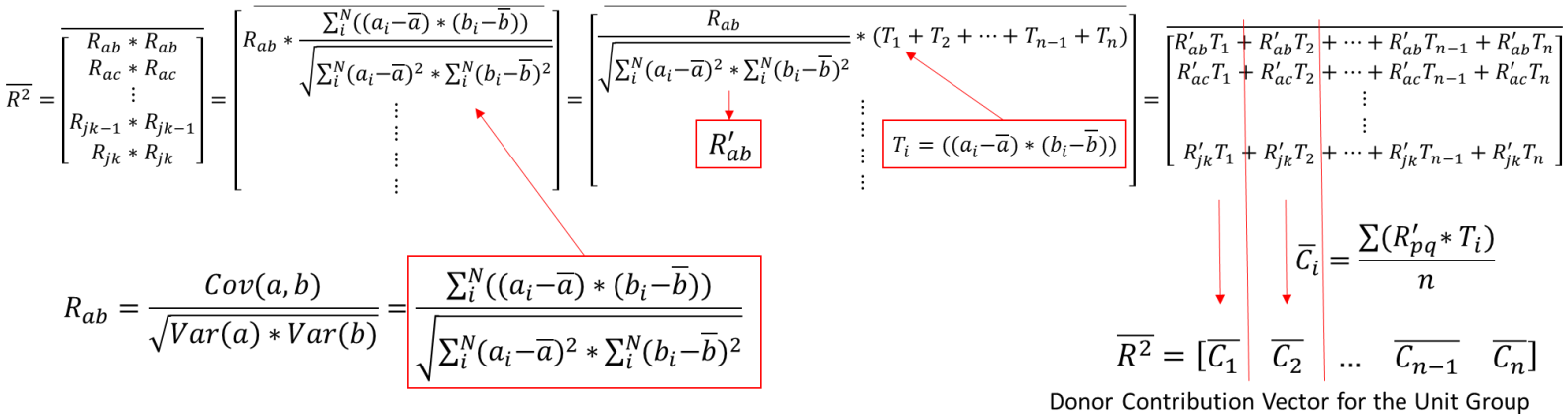



*Figure 24 Single donor contribution in the unit group calculation. (Leftmost) Average $R^2$ is used as surrogate for each unit correlation strength. (Bottom left) Spearman R could be viewed as fraction between covariance and variances of ranks. (Middles) Substitutions of covariance calculation yield connection between each feature ranks and the average $R^2$. (Rightmost) Each feature fractions are combined to make donor contribution vector for each unit.*

# Results

## Genes, methylation sites and phenotypes selected for networking

The RNA-seq (EXP-S) and methylation array (Meth-A) data files of 8 TCGA projects were retrieved from ICGC data portal sites. A total of 4,116 donors with both matching RNA-seq and methylation array data were selected. The cancer phenotypes included squamous cell carcinomas, adenocarcinomas from various origins and skin melanoma.

Gene expression and DNA methylation information from all 4,116 donors were then integrated into a single dataset. Missing values were treated as zero. Features with average expression lower than $10^{-6}$ were removed. A total of 15,666 genes were used. For the methylation data in the Infinium HumanMethylation450K chip, 12,835 CpG sites were selected based on their proximities within ±10 kb of annotated transcript TSSs.

## Clustering of 2-omics feature units

I aimed to group the genes and CpG sites into functional units of synchronized features (Figure 25A) both with the same and opposite expression. The lack of modeling or assumptions of in rank analysis enabled the uniformity of genes and CpG sites in a single step (Figure 25B). Moreover, phenotype activities of each unit could be easily retrieved by looking at the phenotypes with extreme ranks (Figure 25C). Ranking also limited artifacts, such as technical noise or batch effects in the input, but no further attempt at removing noises or batch effects was made. This approach could be viewed as uniform coexpression rank analysis of a multi-omics dataset.

For each feature in both omics, measurements for each donor were ranked from the lowest (1) to the highest values (4,116). This resulted in the uniformly ranked 28,501 features in the 4,116 donor matrix. I designed and conducted the nonparametric analysis in every step from the input integration, clustering and unit identifications by distribution-based tree cutting (see Material and Methods for details).


*Figure 25 Approach to Construct a 2-Omics Network in TCGA; (A) an example unit of 3 genes. A,B and C's changes in expression synchronize across the donors, thus presumed to be working together; (B) genes and CpG sites are unified by rank analysis; (C) an example unit of 2 genes and 1 CpG site and the usage of extreme phenotypes as annotators (Donor1, Donor5).*

To look for the units, UPGMA hierarchical clustering of the 2-omics ranked matrix was conducted based on the rank variance-covariance as distance metrics (see Material and Methods for details). A Spearman correlation ($R^2$) matrix was used to visualize the clustering (Figure 26A pre-cluster; Figure 26B post-cluster). In total, 28,501 features were clustered into 4,358 units; 2,315 units were purely genes, 2,010 units were purely CpG

sites, and 33 units were mixes. Each cluster had an average of 6.5 features. The correlations in these units were determined to be homogenized (Figure 26C-E, areas are outlined by red lines). These units were viewed as building blocks of biological interactions.

For instance, glycolysis is a ubiquitous and crucial energy production pathway. A unit group containing *GAPDH* (Figure 26E, arrow) also contained 6 genes important in glycolysis (Figure 27A). High degrees of positive rank correlations were observed in every gene (Figure 27B). All of these genes encode enzymes involved in glycolysis, which glucose-6-phosphate (glucose) or dihydroxyacetone-P (triglyceride) is converted to both pyruvate (aerobic terminal) and lactate (anaerobic glycolysis terminal). These observations may indicate that cancer cells make use of anaerobic glycolysis, producing lactate from pyruvate by lactate dehydrogenase enzyme (encoded by *LDHA*) even in abundance of oxygen, termed the "Warburg effect" (Liberti & Locasale, 2016).

To further demonstrate the presence of regulatory actions, I focused on a unit group containing the *NKX2-1* gene (Figure 26C, arrow; Figure 28A), which is a transcriptional factor. *NKX2-1* is known to play a central role in various tissues, including lung alveolar type II development and is necessary for surfactant production (Minoo, 2000). This unit was found to contain several genes encoding surfactant-related proteins. Their expression ranks were strongly correlated and significantly higher in lung tissues than in other tissues (average rank 3592 vs 1612; p-value $\simeq$ 0, t test of ranks; Figure 28 B-E).

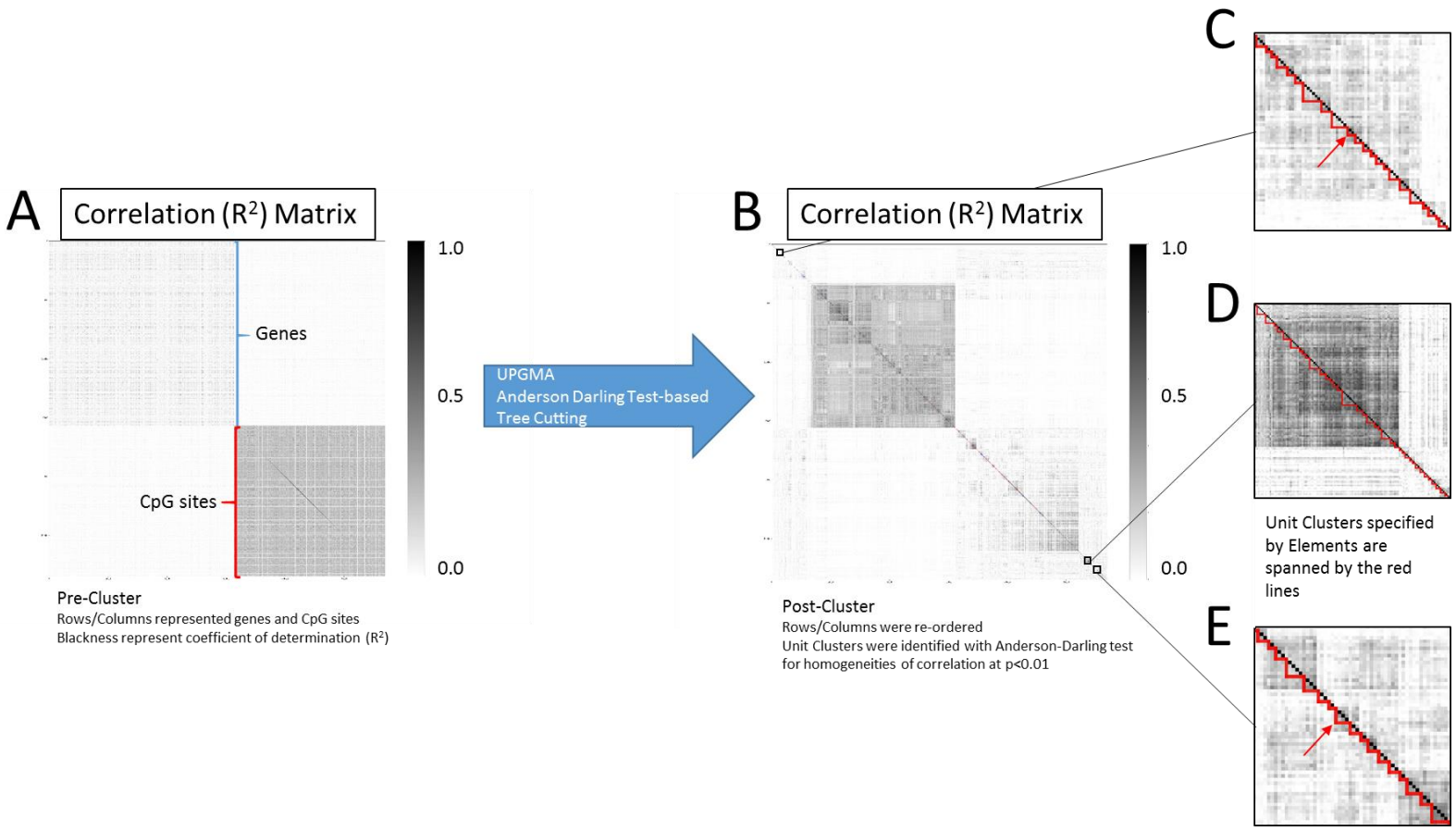



*Figure 26 Correlation ($R^2$) Matrix of Unit Clusters; (A) Pre-cluster $R^2$ matrix. Both row and column represent both genes and CpG sites. Dots represent $R^2$ of each row/column pair. (B) Post-unit clustering $R^2$ matrix. Clustering are done with UPGMA and Tree Cutting are done with Anderson-Darling test ($p<0.01$). (C-E) Close up of $R^2$ inside example units (areas under red lines). (C, arrow) NKX2-1 lung surfactant unit. (E, arrow) GAPDH glycolysis unit.*

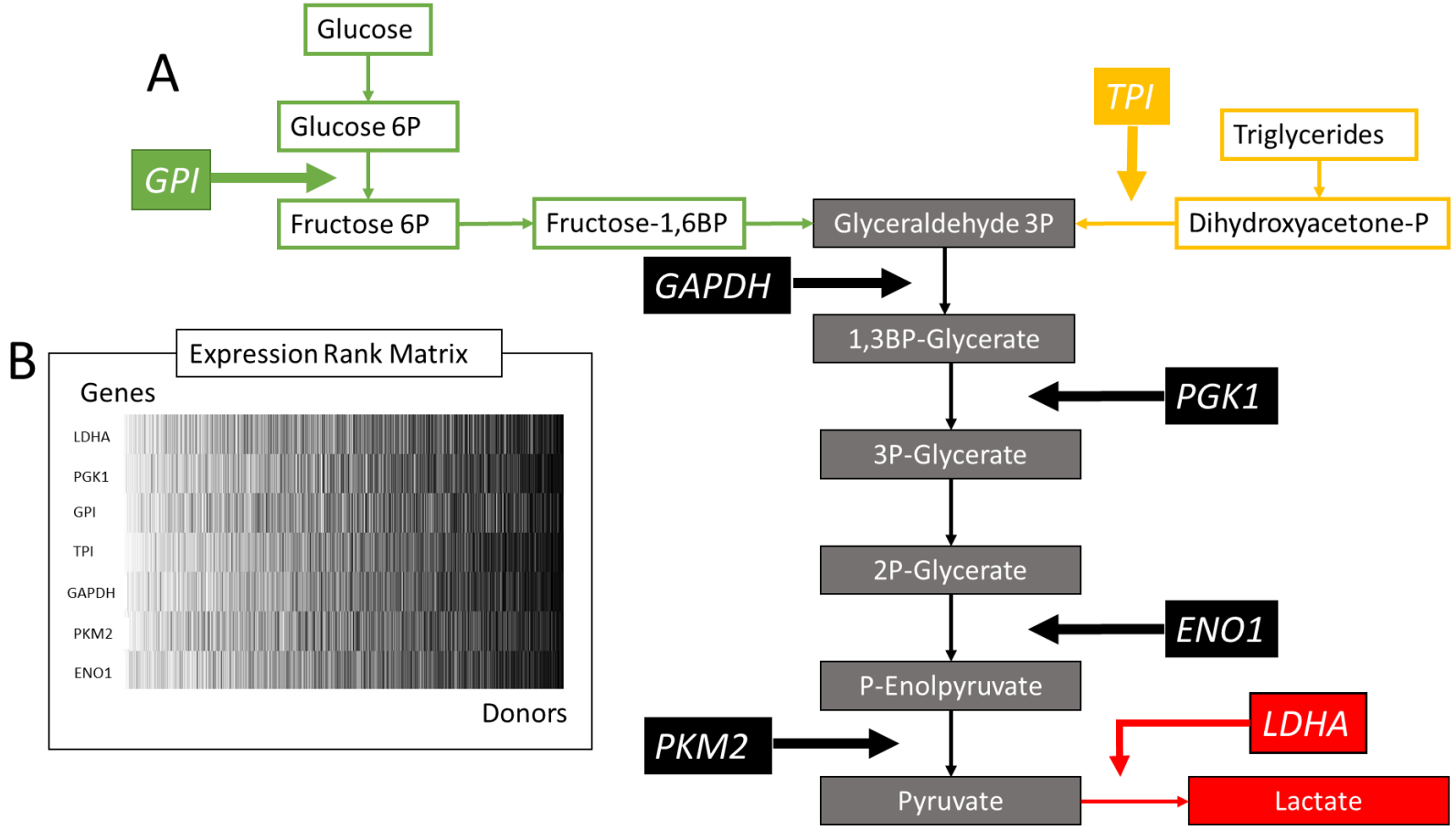



*Figure 27 Glycolysis pathway captured by the GAPDH unit. (A) Every genes encode enzymes related to glycolysis pathway. (B) Rank synchronization of the genes.*

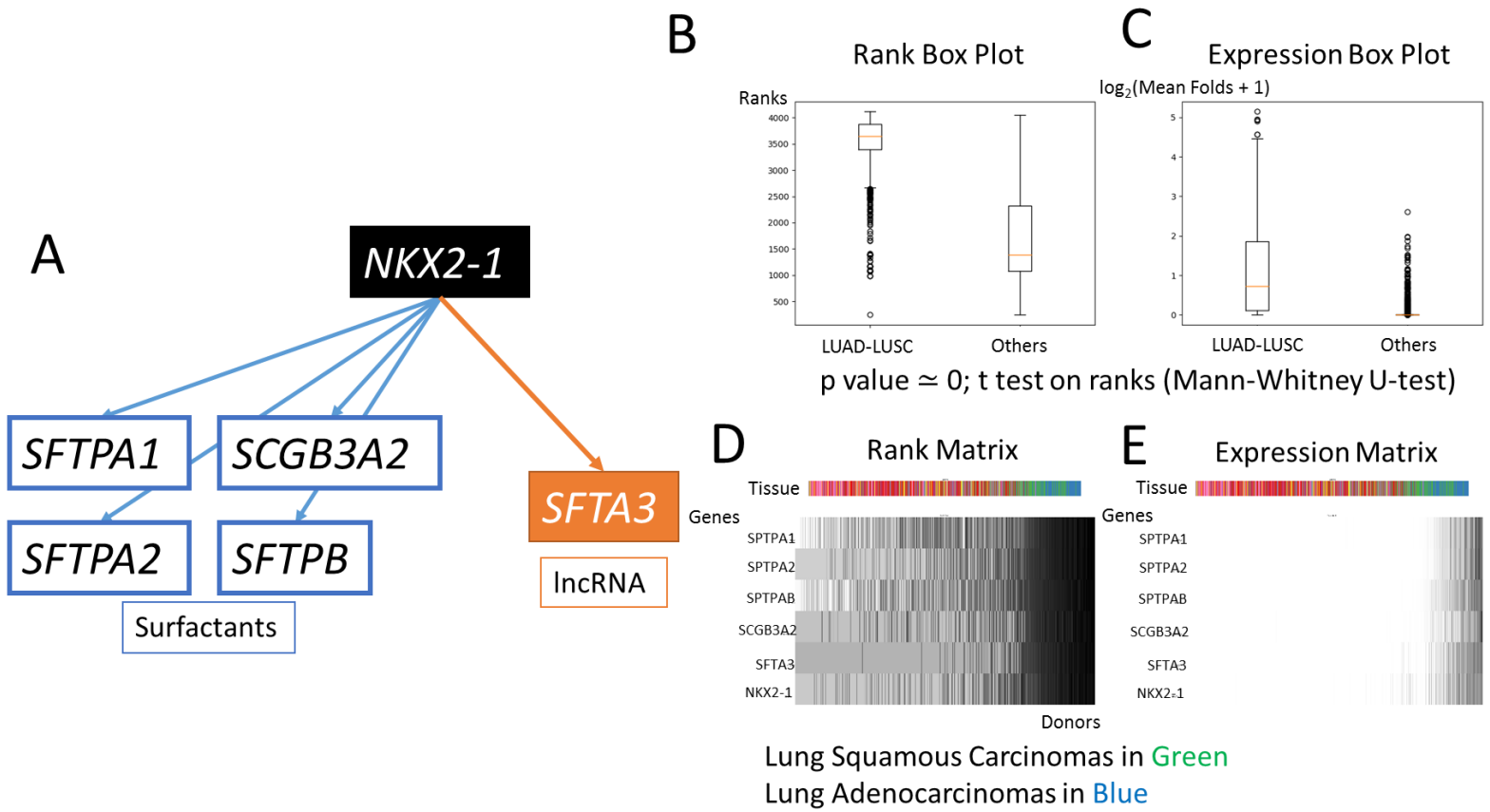



*Figure 28 Regulation of surfactant genes by NKX2-1. (A) Genes in surfactant production are captured. (B,C) Higher expression of this unit genes in lung tissue (LUAD-LUSC) by (B) rank and (C) expression in log scale. (D,E) Genes' (D) rank and (E) RSEM synchronization.*

Average pairwise $R^2$ values were used as surrogates for the degree of synchronization in each unit. Strongly synchronized units with a high average $R^2$ represented units with uniform biological interactions. On the other hand, weakly synchronized units with low average $R^2$ represented units of less biological significance. Some artifacts or network noise may be included among the latter units.

To address the concern with the less correlated units, I examined the degree of inherently existing correlations to estimate the expected average $R^2$. I employed Monte Carlo simulation analysis with 100,000,000 combinations using the same unit size distribution (Figure 29A) as the original clustering. The random combinations produced an average strength of the correlations with the average $R^2$ value of 0.06 (Max: 0.56, Min: $3.6 \times 10^{-6}$) (Figure 29D). This suggests that random combinations with high correlations should be rare and occur only at a limited frequency in smaller units (Figure 29E). This random noise was compared with the obtained units, which produced an average $R^2$ of 0.44 (Max: 1, Min: 0.015) (Figure 29B) and with different distributions (p value $\simeq$ 0, KS test). Nevertheless, stronger correlations were found in smaller units (Figure 29C), albeit with different distributions and intensities (Figure 29B). The weaker average $R^2$ in nonrandom larger units (Figure 29C) might be because highly specialized gene complexes tended to form smaller units compared to the more generalized genes. For the latter, the 1-to-1 correlations were weaker, lowering the average $R^2$. Although I am fully aware of this drawback, I had to leave it to be resolved in future research, since no single solution has been shown to handle a large-scale system and it did not hinder the interpretation of the detected units.

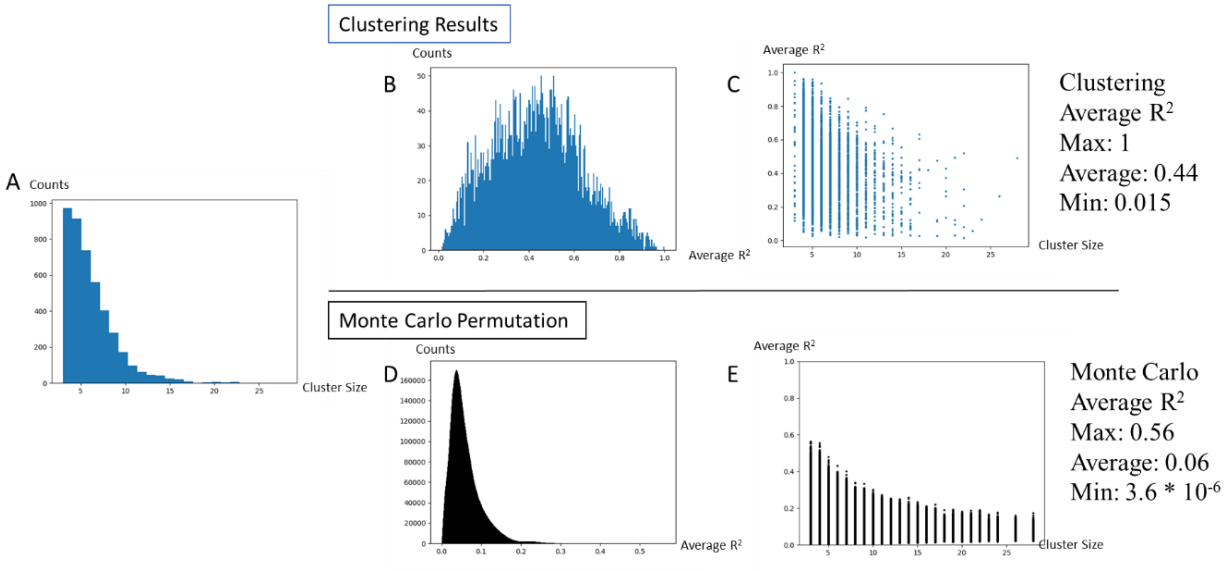



*Figure 29 Cluster strength compared to Monte Carlo simulations. (A) Histogram of unit sizes' distribution; (B) Histogram of unit average $R^2$ distribution; (C) Scatter plot of relation between unit size and average $R^2$; (D) Simulated average $R^2$ distribution with (A) distribution; (E) Relation between simulated size and average $R^2$.*

Individual units, though they could precisely identify basic functional units, could not represent the whole picture of biological systems. As seen in the previous examples, even the simplest systems could not be entirely captured by any single unit. In the *GAPDH* unit, while some part of the metabolic reaction chain could be identified, a wider view related to general energy production or glycolysis was not captured. It was likely that each phenotype regulated and utilized each functional unit separately to serve their various metabolic needs. Each unit may also be under different controls, thereby having varied expressions. A similar narrowed view was also observed for *NKX2-1*. In this case, only the most upstream regulator, the transcriptional factor *NKX2-1*, and the most downstream effectors, the surfactant-related proteins, were represented, omitting everything else in-between.

## Linking Units into Networks with Phenotype Activities

In the covariance and correlation analysis (see Material and Methods for details), the consistently high- and low-ranking donors made big contributions in each unit. High contributions represented high activities (Figure 30B; black dots). Utilizing this view, the units with similar phenotypic identities were networked together (see Material and Methods

for details) by angular distance between their donor contribution vectors. Networks of similar donor identities were determined using UPGMA and Anderson-Darling test-based tree cutting in the same manner as previously described (Figures 30C and D; blue lines covering red lines). As a result, the 4,358 units were constructed into 654 networks with an average of 6.66 units per network (Figure 31A) or 43.58 features per cluster (Figure 31B). Twenty-nine networks were found to be mixed networks of both genes and CpG sites.

These networks, however, did not appear as rigid as the more basic unit groups. This was due to the difference in interpretation regarding synchronization and homogeneities of the rank in the expression and phenotype activities. Synchronization of the expression ranks was examined as crucial and direct evidence for grouping features with homogeneous functions and interactions. However, the goal of the networks was to capture interactions between each unit group. This part proved to be difficult since each phenotype could regulate the activities of each unit to suit its needs, independent of the unit-unit interactions. Phenotype activities of each unit, in turn, did not need to be homogenized for the units to have biologically meaningful interactions; in other words, biologically significant interactions could present across networks as well.

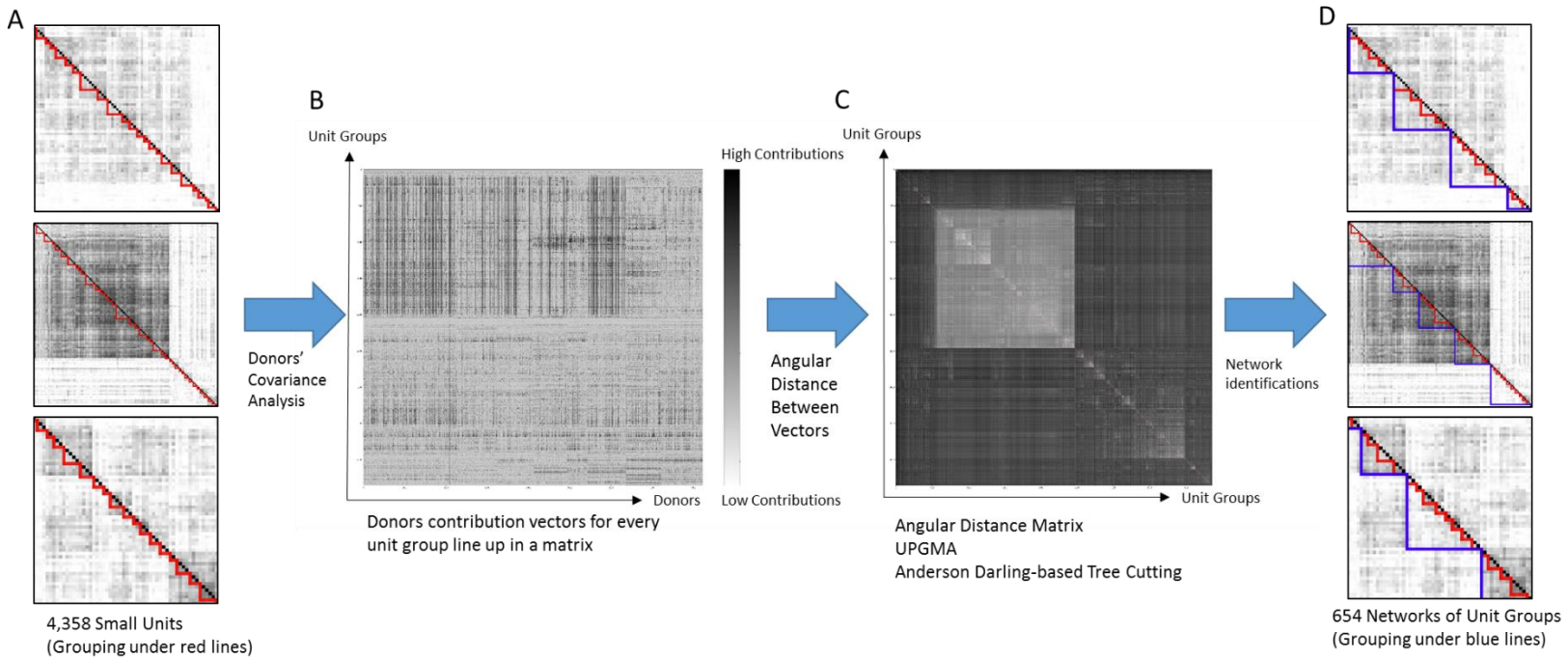



*Figure 30 Phenotype-based Network Construction. (A) Close up views of unit correlations; (B) Every units' donor contribution vectors line up in a matrix' (C) Angular distance matrix between each unit vector; (D) Networks of units (under blue lines) construct from UPGMA and Tree Cutting of (C).*

In the same manner as functional unit construction, networks with homogeneously high angular distances (poor similarities) were constructed (Figure 31C). An inverse correlation was observed between the mean angular distance and the mean pairwise $R^2$ of the networks (Pearson R=-0.90; Figure 31D), showing that they originated from units with poor average $R^2$. These networks were less likely to hold biologically meaningful interactions.

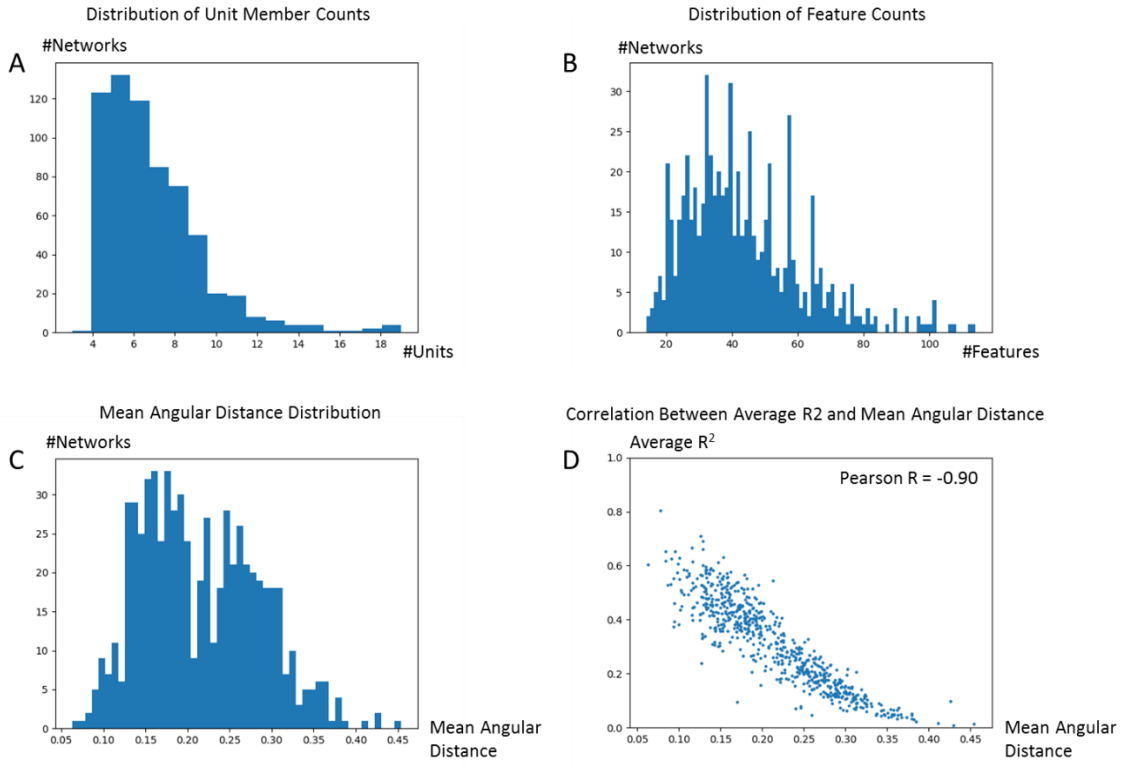



*Figure 31 Overview of Networking Statistics. (A) Histogram of network size distribution by units; (B) Histogram of network size distribution by genes or CpG sites; (C) Histogram of networks' internal mean angular distance; (D) Scatter plot between networks' internal mean angular distance and mean average $R^2$.*

To interpret the networks, I first focused on the known and curated gene sets deposited in MSigDB for similarities. I utilized 5 gene sets, cancer hallmark gene sets (H), curated gene sets (C2), computational gene sets (C4), GO gene sets (C5) and oncogenic gene sets (C6). Due to differences in the designs and methods in each gene set, the sets themselves were not always consistent with each other, including with mine. However, with a significant number of network hits in each gene set (Table 17), these results indicate

the biological significance of the networks. I manually inspected the networks of substantial biological interest feature-by-feature. While this would not be scalable to a comprehensive analysis, I considered it necessary to biologically uncover the meaningful interactions in the networks.

*Table 17 Cross-referencing Networks to MSigDB*

| **Gene Set** | **Aberration** | **Network Hits** | | **Gene Set Hits** | | **Average Hits** |
|---|---|---|---|---|---|---|
| | | Counts | % of Total | Counts | % of Total | **per Network** |
| **Cancer Hallmark** | H | 68 | 10.40% | 41 | 82.00% | 2.07 |
| **Curated Gene Sets** | C2 | 234 | 35.93% | 1754 | 31.89% | 17.03 |
| **Computational Gene Sets** | C4 | 95 | 14.53% | 404 | 47.09% | 11.77 |
| **GO Gene Sets** | C5 | 117 | 17.89% | 1497 | 14.98% | 23.26 |
| **Oncogenic Gene Sets** | C6 | 540 | 82.60% | 83 | 43.92% | 2.85 |

To demonstrate the networks' abilities in capturing known interactions, a network containing genes involved in mitosis were investigated. This network was arbitrary selected from a group of networks which overlapped with mitotic related GO terms. The selected network contains 7 units with 50 genes, all of which are known to be involved in mitosis and DNA replications, including *AURKA* and *AURKB* which are conserved mitotic and cell cycle regulators with many known interactions. One unit contains *AURKA* and *AURKB* (Figure 32) and 3 units strongly connected to them by literal evidence (Figure 32, Black arrows). *AURKA* and *AURKB* are known to interact with each other. *AURKA* are also reported to interact with *UBE2C* and phosphorylate *PLK1* with the effects of driving the cell cycle. *AURKB* regulate the alignment and segregation of chromosomes by targeting many centromeres-related substrates including *HASPIN*, *CENPA, HJURP* and *CENPN*. It is also reported to interact with *NCAPD2* in condensin complex. These results show that the networks were able to capture proven biological significant interactions.

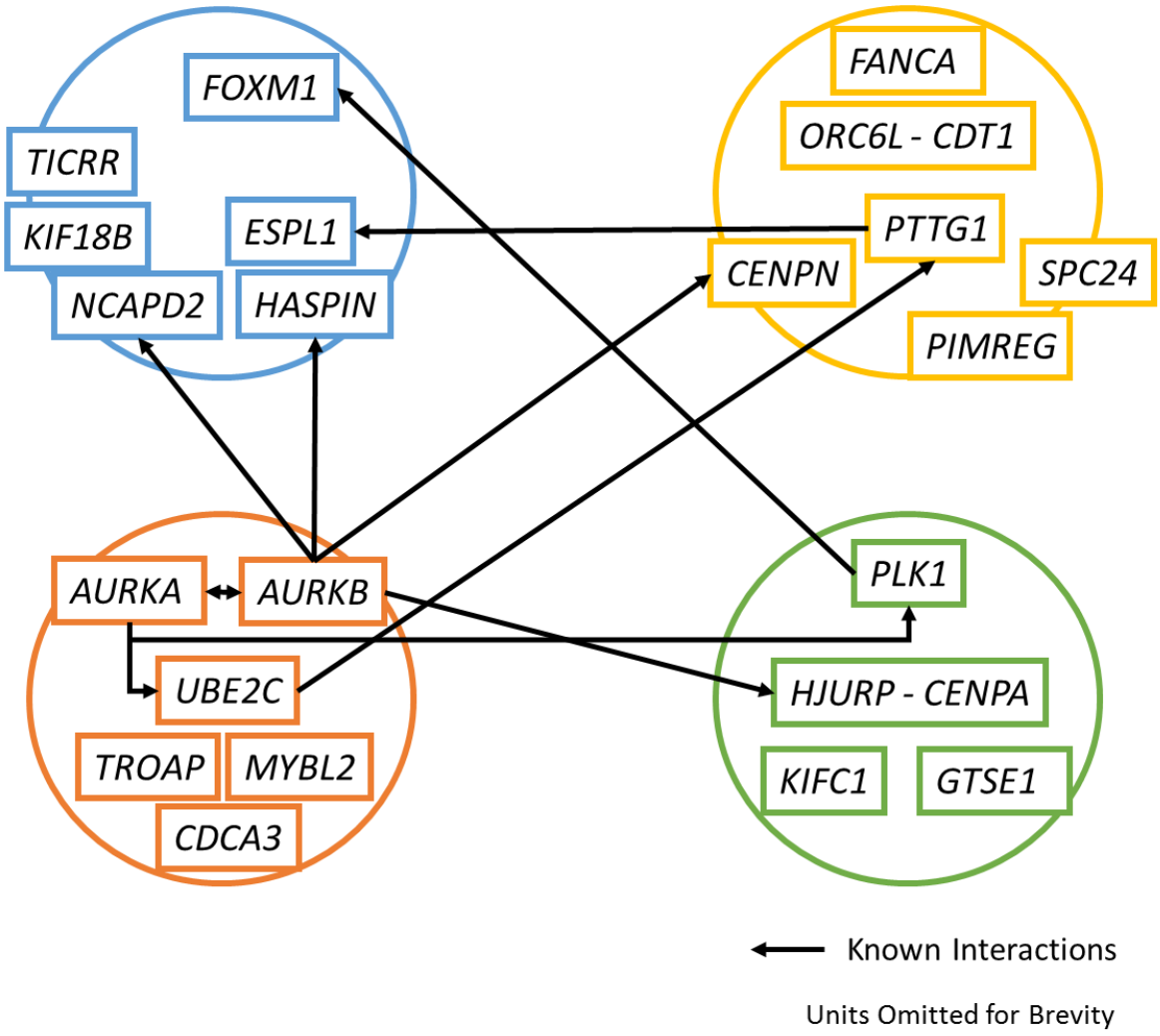



*Figure 32 Network containing known interactions of AURKA and AURKB (Orange). AURKA is known to interact with UBE2C and PLK1, driving mitosis forward. AURKB is known to interact with centromeres=related genes to regulate chromosomes alignment and segregation during mitosis*

To exemplify the networks, units that were linked to GAPDH/glycolysis are shown in Figure 33. In this network, 8 units with a total of 49 genes were grouped. Gene Ontology analysis revealed that the member genes of this network were highly enriched with the GO terms (MSigDB C5) related to glycolysis and carbohydrate metabolism, as expected (Table 18). The networking linked *ALDOA*, *PGAM1*, *PGAM4* and *PC* (Figure 33A) to the *GAPDH* unit, completing the reaction for glyceraldehyde 3P for entry into the TCA cycle or production of lactic acid.

This network revealed that the genes in other pathways also interact with the glycolysis pathway. For example, the *ALODA/PGAM1* unit (Figure 33A, blue) and cell membrane-ER related unit (Figure 33A, yellow) hinted at a novel interaction between membrane transport proteins, the receptor system and the glycolysis pathway. Moreover, the *PC* unit included *ALDH4A1*, which encodes aldehyde dehydrogenase and is localized in mitochondria. This enzyme produces glutamate, providing an alternative substrate for the TCA cycle. Interestingly, high *ALDH* activities are associated with malignancies in some cancer species. This network also included *BLCAP* and *CTNNB1*, which directly influenced cell proliferation.

*Table 18 GO Terms for the GAPDH Network*

| MSigDB C5 GO term | Odds Ratio | Raw P Value | B-H correction | Gene Hits |
|---|---|---|---|---|
| GO_GLYCOLYTIC_PROCESS_THROUGH_FRUCTOSE_6_PHOSPHATE | 113.4 | 3.53E-12 | 3.53E-08 | 7 |
| GO_GLYCOLYTIC_PROCESS | 39.8 | 1.49E-11 | 3.53E-08 | 9 |
| GO_MONOSACCHARIDE_BIOSYNTHETIC_PROCESS | 39.5 | 1.97E-10 | 7.46E-08 | 8 |
| GO_CARBOHYDRATE_CATABOLIC_PROCESS | 23.0 | 1.44E-09 | 6.57E-07 | 9 |
| GO_GENERATION_OF_PRECURSOR_METABOLITES_AND_ENERGY | 11.0 | 4.40E-08 | 3.59E-06 | 11 |
| GO_CARBOHYDRATE_BIOSYNTHETIC_PROCESS | 17.3 | 8.70E-08 | 8.79E-05 | 8 |
| GO_CARBOHYDRATE_METABOLIC_PROCESS | 8.6 | 1.39E-06 | 1.45E-04 | 10 |

The rank synchronizations in every unit in the networks were not expected to be perfect. The whole network was deactivated (Figure 33B; blue arrows) or activated in similar phenotypes (Figure 33B; red arrows), representing interactions between the members. However, each unit in the network did not exhibit exact patterns of synchronization in ranking, as in the unit level (Figure 33B Lower). Many functionally unknown genes (Figure 33C) were also assigned to this network. These interactions may be worth subjecting to in-depth analysis to uncover the links between cancer cell metabolism, the Warburg effects and cell proliferation.

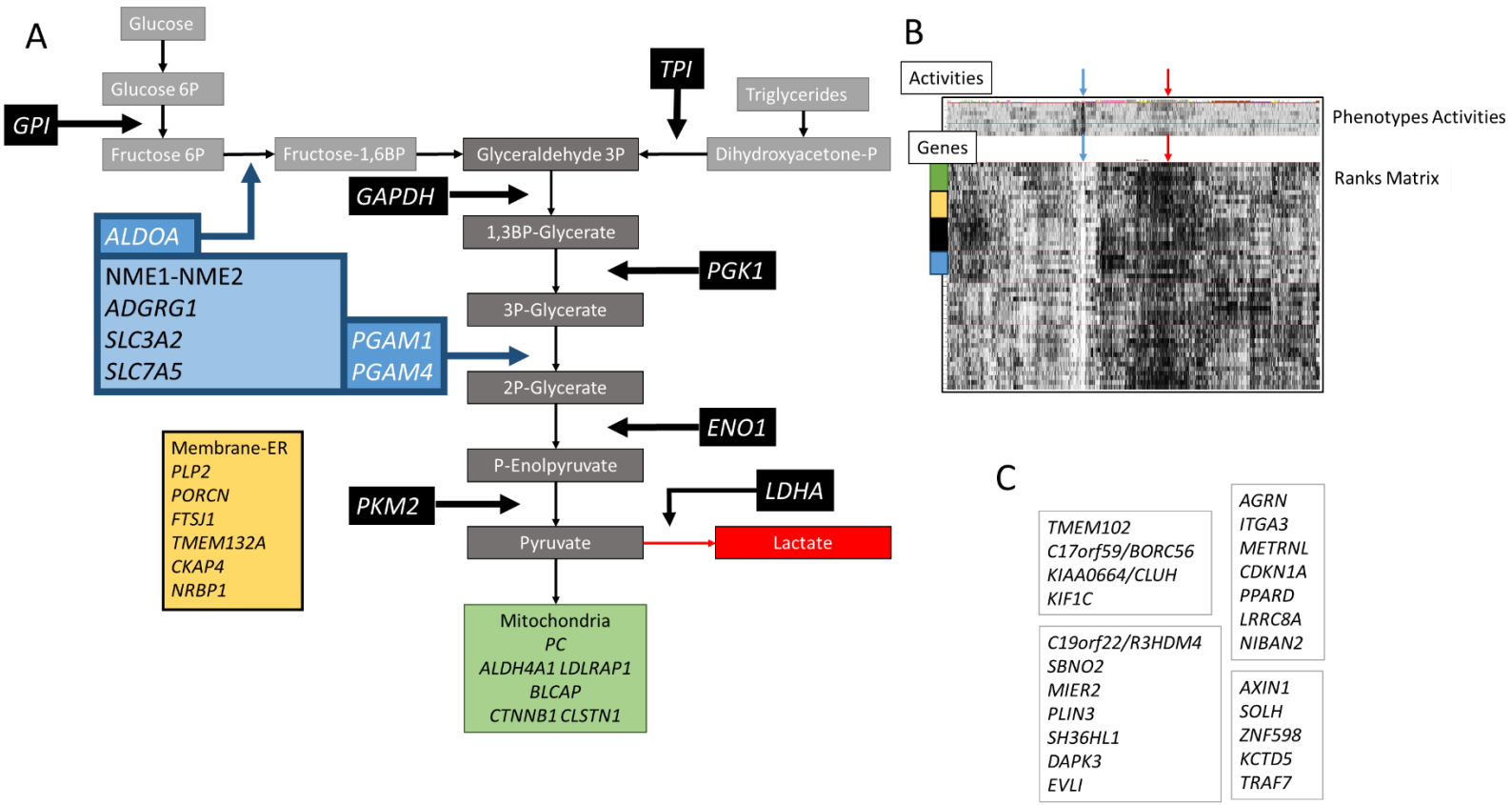



*Figure 33 Network of units surrounding the GAPDH/Glycolysis Pathway. (A) Genes in network relate or potentially relate to glycolysis pathway; (B) Rank matrix of the network (Lower). Rows represent genes, column represent samples. (Green) Rows of green unit in (A). (Yellow) Rows of yellow unit in (A). (Black) Rows of black unit (GAPDH) in (A). (Blue) Rows of blue unit in (A). (Unlabeled) Rows of units of ambiguous functions (C). Donor contribution vectors of each unit (Upper). (Blue arrow) samples of low activities. (Red arrow) samples of high activities; (C) 4 units of ambiguous functions.*

## Network analysis captures the regulators and effects of *NFATC1*

In the previous chapter, aberrant regulation of *NFATC1* was detected and validated in lung adenocarcinoma cell lines. Its expression level was shown to be significantly associated with the prognosis of clinical cases in the TCGA LUAD dataset. To further investigate the roles of *NFATC1* that might be involved in carcinogenesis, I explored the networks of genes and CpG sites surrounding *NFATC1*.

Similar to *GAPDH* (glycolysis) and *NKX2-1* (lung surfactant production), the unit group containing *NFATC1* was relatively small. Limited information regarding its function was available. However, after expanding the analysis to the phenotypic network level, the presumed interactions were revealed.

One of the most well-known aspects of *NFATC1* function is in T cells, where it is primarily activated by calcineurin under the regulation of *RCAN1*. After activation, *NFATC1* is localized to the nucleus, where it functions as a T cell activator mainly by activating *EGR2* transcriptional factors. T cells are then differentiated to their effector variants and provide cell-mediated immune responses. This activation process was reflected in units containing *RCAN1* (Figure 34A; 4).

More intriguingly, further analysis of this network revealed potentially novel downstream effects of *NFATC1* in 3 separate functions that were potentially not related to the immune system. The first was a signaling pathway centered around *GHR* (Growth Hormone Receptor) and *NTRK2* (Neurotrophic Tyrosine Receptor Kinase 2) (Figure 34A; 2). Their downstream signaling could consequently influence cell differentiation and proliferation. The second was also a signaling pathway based on *NGFR* (Nerve Growth Factor Receptor) (Figure 34A; 3). Like the first, this pathway similarly played a role in cell differentiation and proliferation but possibly in different contexts. The last pathway consisted of *MAPK10*, *NAP1L2*, *PTN*, *HLF*, and *IGSF10* with 1 CpG site (cg22980079). The first two genes *MAPK10* and *NAP1L2* are known to have proliferative capacity. *PTN* and *HLF* have been reported to be dysregulated in various cancers. These interactions indicated the possible direct involvement of *NFATC1* in cancer cells, most likely promoting

cancer cell proliferation. The included CpG site (cg22980079) was annotated to *FAM193A*, but its roles are still unknown.

Considering the entire network, in active disease, the expression of *NFATC1* could indicate both cancer cell proliferation and tumor infiltrating T cell responses. Indeed, higher disease activities from higher tumor proliferation could lead to stronger responses. In this view, both oncogenic and antitumor effects of *NFATC1* could coexist and cooperate in representing disease activities. This partially explained the conflicting survival analysis results, in which higher NFATC1 was associated with worse disease-free outcome and better overall survival. Worse disease-free outcomes could be attributed to higher relapses from higher pretreatment disease activities. Because most of the cancer relapsed, better overall survival could be attributed to stronger immune responses after relapse.

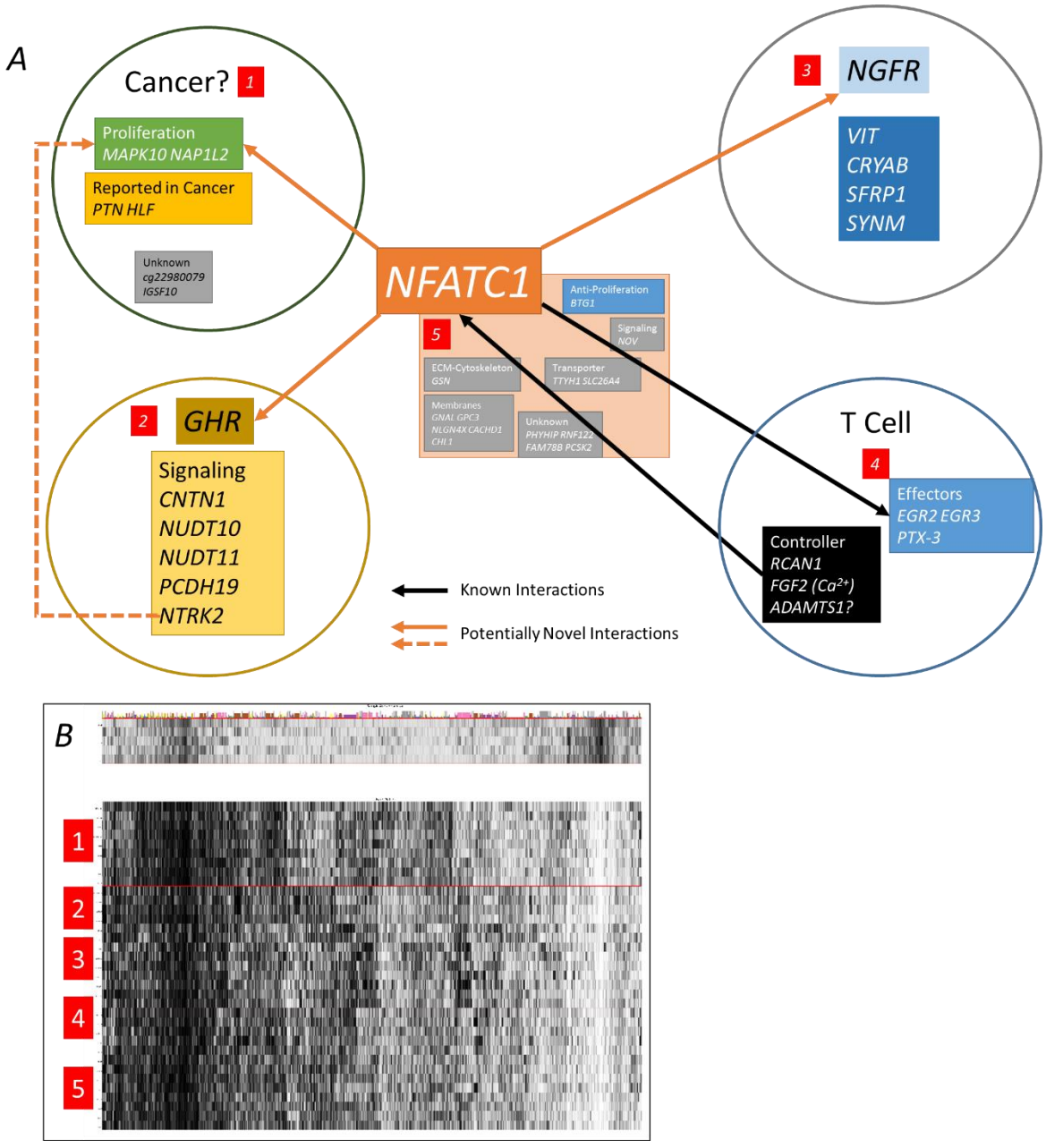



*Figure 34 Network Surrounding NFATC1. (A) 5 units in the network. (1-3) consist of genes relate to cell proliferations. (4) consists of a known NFACT1 regulation and activation in T Cells. (5) contains NFATC1; (B) Donor contribution vectors (Upper) and rank matrix (Lower) of 5 units.*

.

## Networks of Interactions Involving DNA Replication, Repair and Methylation

During DNA replication, the newly synthesized DNA strand is unmethylated. The hemi-methylated DNA strands are methylated mainly by *DNMT1* in the process term "Methylation Maintenance" (Greenberg & Bourc'his, 2019; Law & Jacobsen, 2010). This process is crucial in DNA replication and is required for normal regulation of the cell transcriptomes. DNA damage is not only limited to genetic elements but also to losses of epigenetic information (Dabin, Fortuny, & Polo, 2016), including DNA methylation. Every cell needs to actively maintain its methylation status via *DNMT1*. Knockdown of or mutations in the *DNMT* family of genes has also been shown to cause major disruption in genome integrity and cell survival (Liao et al., 2015).

The roles of DNA methyltransferases are not limited to maintaining methylation patterns. "De novo" methylations of unmethylated sites were also described in both normal and pathological circumstances. De novo methylation is thought to be mainly carried out by *DNMT3A* and *DNMT3B*. Under normal circumstances, this process occurs in stem cells or during embryogenic development. However, this event is also frequently observed in cancer cells (Kulis & Esteller, 2010). Dysregulation of methylation patterns is believed to be one of the major driving events of carcinogenesis. Clinical trials studying the usage of DNMT inhibitors as a strategy in the treatment of various cancers, mainly leukemia, are currently underway (Gnyszka, JastrzĘBski, & Flis, 2013; Wong, Lawrie, & Green, 2019). Moreover, oncogenic driver genes, such as *BRCA1* or *TERT*, were shown to host aberrant DNA methylation at their promoter regions. Unlike methylation maintenance, de novo methyltransferases were less studied in their activation conditions and controls, hindering associations between the detected methylation patterns and carcinogenesis.

Interactions representing DNA methylation were identified from 6 networks (Figure 35). GO term analysis showed enrichments related to the mitotic cell cycle and DNA replication control (Table 19). Closer inspections of genes in each network revealed that 5 networks were mainly associated with DNA repair, which included 4 networks headed by

*BRCA1*, *BRCA2-FANCB*, *FANCA* and *FANCL* and one network that was characterized by cell cycle regulatory units, including *CDK1* and *CHEK1*, known for their functions in DNA double strand break repairs (Figure 35A). These 5 networks were related to the network associated with the DNA replication complexes, including DNA polymerase family B complex genes (*POLA2*, *POLD1*, *POLE*, *POLE2*) (Figure 35B). Additional cell cycle controllers (*FANCD2*, *CHEK2*), replication-related catalysts (*LIG1*, *PRIM1*, *PRIM2*), and chromatin structure control complexes (Kinetochore, Spindle fibers) were also presented. Most interestingly, the DNA methyltransferases *DNMT1* and *DNMT3B* and genes encoding members of the histone methyltransferase complexes *CBX2* and *EZH2* were also included in the DNA replication complex network.

Unlike the *NFATC1* network, these processes were separated into 6 networks. However, their phenotypic activities remained largely similar between them. Minor differences were observed between each network, and the majority of the phenotypes showed strong correlations. Despite the strong correlations, the minor differences were considered heterogeneous, thus splitting the networks into 5 smaller networks (Figure 35C).

I further searched for literal evidence on the interactions between complexes in the 6 networks. In DNA replication and the methylation networks (Figure 35B), the interaction between the DNA polymerase complex and *DNMT1* was bridged by *UHRF1*. *UHRF1* has been shown to recruit *DNMT1* to the site of DNA methylation during DNA replication (X. Liu et al., 2013). For the internetwork interactions, DNA repair complexes such as *BRCA1-BRCA2-HMMR* and *BRCA1-FANCA* were identified. Regulations of DNA replication complexes via E2F-family transcriptional factors and cell division cycle-associated genes (CDCA family) in the 4 DNA repair networks (Figure 35A) were also detected.

*Table 19 Top 5 GO terms for Cell Cycle and DNA Replication Networks*

| **NETWORK #614; 45 GENES** | | | | |
|---|---|---|---|---|
| **GO TERMS** | Odds Ratio | Raw P Value | BH correction | # Hits |
| **GO_CELL_CYCLE** | 7.5 | 1.40E-09 | 1.40E-05 | 20 |
| **GO_CELL_CYCLE_PROCESS** | 8.1 | 4.29E-09 | 1.40E-05 | 17 |
| **GO_MITOTIC_CELL_CYCLE** | 8.6 | 9.33E-09 | 2.14E-05 | 15 |
| **GO_CELL_CYCLE_G1_S_PHASE_TRANSITION** | 17.4 | 1.54E-08 | 3.11E-05 | 9 |
| **GO_REGULATION_OF_DNA_DEPENDENT_DNA_REPLICATION** | 42.4 | 3.26E-07 | 3.84E-05 | 5 |
| **NETWORK #615; 25 GENES** | | | | |
| **GO TERMS** | Odds Ratio | Raw P Value | BH correction | # Hits |
| **GO_DNA_REPLICATION** | 31.7 | 2.09E-09 | 2.09E-05 | 8 |
| **GO_CELL_CYCLE_PROCESS** | 14.4 | 2.45E-09 | 1.22E-05 | 13 |
| **GO_REGULATION_OF_MITOTIC_CELL_CYCLE** | 17.9 | 7.90E-09 | 1.22E-05 | 10 |
| **GO_CELL_CYCLE** | 12.0 | 9.40E-09 | 2.35E-05 | 14 |
| **GO_REGULATION_OF_CELL_CYCLE_PROCESS** | 15.4 | 2.99E-08 | 2.35E-05 | 10 |
| **NETWORK #616; 79 GENES** | | | | |
| **GO TERMS** | Odds Ratio | Raw P Value | BH correction | # Hits |
| **GO_CELL_CYCLE** | 13.8 | 1.43E-27 | 1.43E-23 | 47 |
| **GO_CHROMOSOME** | 15.0 | 2.72E-25 | 1.43E-23 | 37 |
| **GO_DNA_REPLICATION** | 26.0 | 9.91E-22 | 1.36E-21 | 22 |
| **GO_DNA_METABOLIC_PROCESS** | 14.2 | 1.28E-21 | 3.20E-18 | 31 |
| **GO_DNA_DEPENDENT_DNA_REPLICATION** | 38.6 | 4.56E-21 | 3.20E-18 | 18 |
| **NETWORK #617; 38 GENES** | | | | |
| **GO TERMS** | Odds Ratio | Raw P Value | BH correction | # Hits |
| **GO_CELL_CYCLE_PROCESS** | 20.3 | 1.90E-17 | 1.90E-13 | 23 |
| **GO_MITOTIC_CELL_CYCLE** | 21.3 | 4.59E-17 | 1.90E-13 | 21 |
| **GO_CELL_CYCLE** | 18.1 | 6.51E-17 | 2.17E-13 | 25 |
| **GO_CELL_DIVISION** | 22.2 | 1.58E-14 | 2.17E-13 | 16 |
| **GO_DNA_REPLICATION** | 31.1 | 2.47E-13 | 3.94E-11 | 12 |
| **NETWORK #618; 52 GENES** | | | | |
| **GO TERMS** | Odds Ratio | Raw P Value | BH correction | # Hits |
| **GO_MITOTIC_CELL_CYCLE** | 42.6 | 8.29E-35 | 8.28E-31 | 37 |
| **GO_CELL_CYCLE_PROCESS** | 29.9 | 1.57E-29 | 8.28E-31 | 36 |
| **GO_CELL_CYCLE** | 28.3 | 5.43E-29 | 7.82E-26 | 39 |
| **GO_ORGANELLE_FISSION** | 39.8 | 4.92E-27 | 1.81E-25 | 25 |
| **GO_MITOTIC_NUCLEAR_DIVISION** | 46.0 | 9.02E-26 | 1.23E-23 | 22 |
| **NETWORK #619; 50 GENES** | | | | |
| **GO TERMS** | Odds Ratio | Raw P Value | BH correction | # Hits |
| **GO_CELL_CYCLE** | 26.8 | 3.16E-27 | 3.16E-23 | 37 |
| **GO_MITOTIC_CELL_CYCLE** | 22.0 | 1.86E-22 | 3.16E-23 | 28 |
| **GO_CELL_CYCLE_PROCESS** | 18.3 | 7.16E-21 | 9.31E-19 | 29 |
| **GO_CELL_DIVISION** | 23.9 | 6.56E-20 | 2.38E-17 | 22 |
| **GO_CHROMOSOMAL_REGION** | 30.5 | 7.98E-19 | 1.64E-16 | 18 |

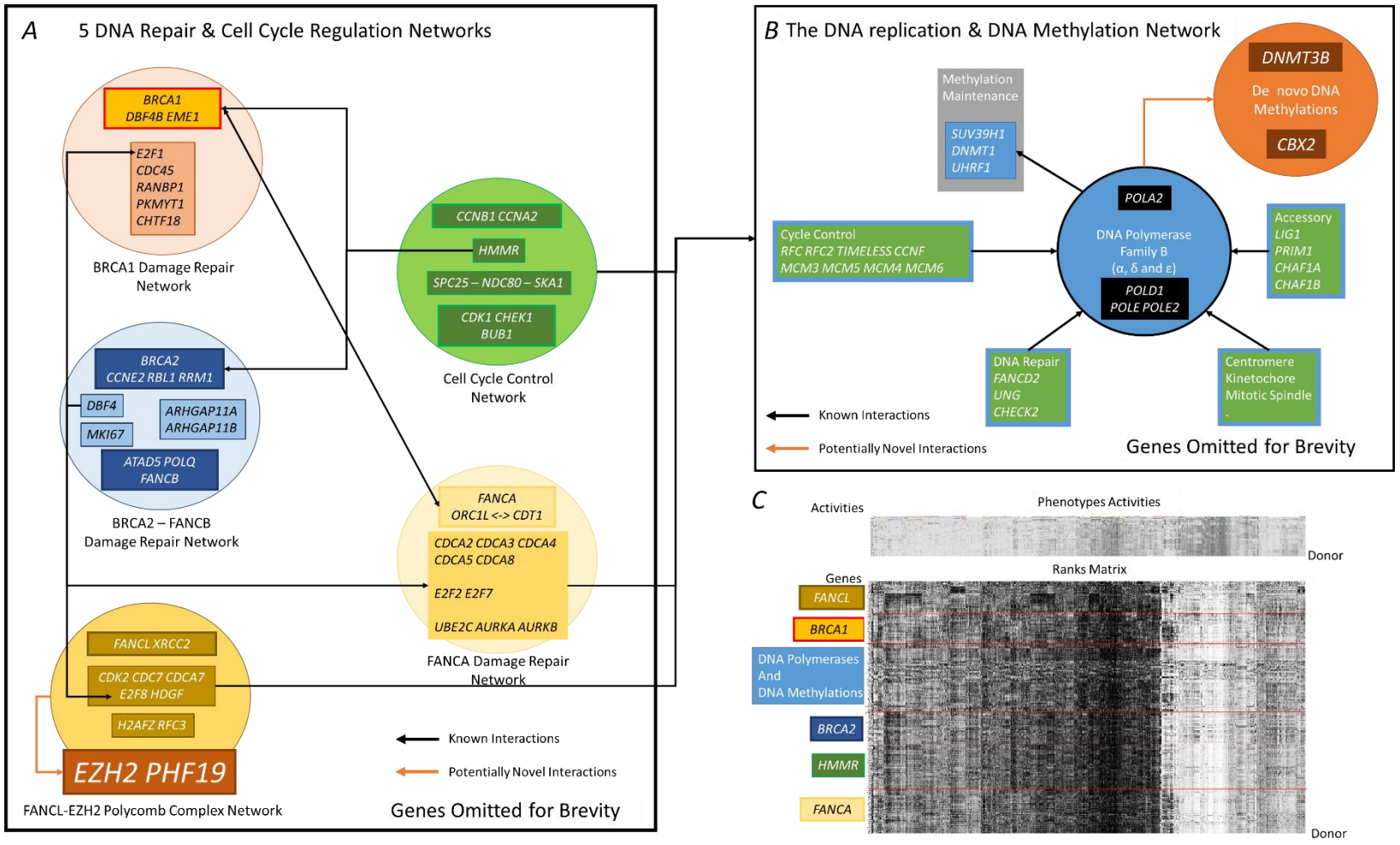



*Figure 35 6 Networks describing DNA Replication, Repair and Methylation. (A) 4 DNA repair and 1 cell cycle regulators Networks headed by BRCA1, BRCA2-FANCB, FANCL, FANCA and CDK1. Known interactions between the networks are shown in black arrows. Relation between histone methylation complex (EZH2) and DNA repairs are suggested. (B) Network containing DNA replication and methylation genes. Both maintenance (DNMT1) and De novo (DNMT3B) methyltransferases are included. Known interactions within (B) and from (A) are shown in black arrows. Interactions between DNMT3B and DNA repairs are suggested. (C) Donor contribution vectors (Upper) and rank matrix (lower) of the networks.*

Taken together, these 6 networks suggested that the DNA polymerase complex was closely monitored by various cell cycle regulatory elements. The inputs of these regulatory factors included a cell cycle rhythm controller such as the *TIMELESS*, RFC or MCM complexes under normal conditions. In the presence of DNA damage, the cells may employ DNA repair mechanisms, including *BRCA1*, *BRCA2*, *FANCA*, and *FANCB*, in coordination, and these mechanisms influence the replication complex by various transcription factors. The exact roles of each gene in the E2F family of transcription factors are not yet clear. However, by analyzing the DNA repair networks, further annotations were possible. *E2F1* shared a network with *BRCA1* and might interact with *BRCA1* more closely than other genes. Additionally, *E2F2* and *E2F7* might be more closely related to *FANCA* than previously known. To restore epigenetic status, *UHRF1* recruited *DNMT1* to maintain DNA methylation patterns. *SUV39H1* and the Polycomb group restored histone methylation and appropriated chromatin structures. Both in coordination with DNA replications and repairs complexes. Interestingly, activation of *DNMT3B* and de novo methylation in tandem with DNA damage repair might provide a clue as to how de novo methylations are activated in cancer cells and explain the roles of dysregulation and aberrant methylations found in various cancer types.

## 2-Omics Melanoma Specific Network

The above networks were mostly concerned with gene-gene interactions. Networks having both omics were investigated to demonstrate the integration between genes and the CpG sites. From the 29 “mixed” networks, one network was found to represent interactions in melanoma. This network contained 4 CpG sites and 34 genes (Figure 36A). Their activities were highly pronounced in melanoma samples (TCGA-SKCM, Figure 36B; red arrows, pink phenotypes). Closer inspections of this network revealed the genes associated with melanoma or melanocyte functions in all units (Figure 36A). These included units containing genes in melanosomes and melanin production, melanoma-specific antigen (*MLANA)* and melanoma-specific transcriptional factor (*PRAME*). Three CpG sites were

located in the tubulin unit. cg00231644 was annotated to *TUBB4* and cg11821702 and cg22598744 were annotated to *MLANA*. Capturing melanoma-specific epigenetic interactions. Last, a microenvironment unit of *EDN1* (endothelin 1) and *MMP7* (degrading specific ECM) had prominently lower expression in melanoma (Figure 36B). These results collectively suggest specific genetic and methylation changes in melanoma or skin samples.

Interestingly, this network was not enriched in any GO term but overlapped with gene sets involving downregulation of P53 (MSigDB; c6: P53 DN. V1 DN), melanin production (MSigDB; c2: REACTOME MELANIN BIOSYNTHESIS) and breast cancers (MSigDB; c2: SMID BREAST CANCER RELAPSE IN BONE DN, SMID BREAST CANCER BASAL UP). These overlaps indicated oncogenic potential in the network, coupled with normal melanocyte functions. It is possible that melanoma might hijack normal melanocyte genetic and epigenetic machinery to turn malignant.

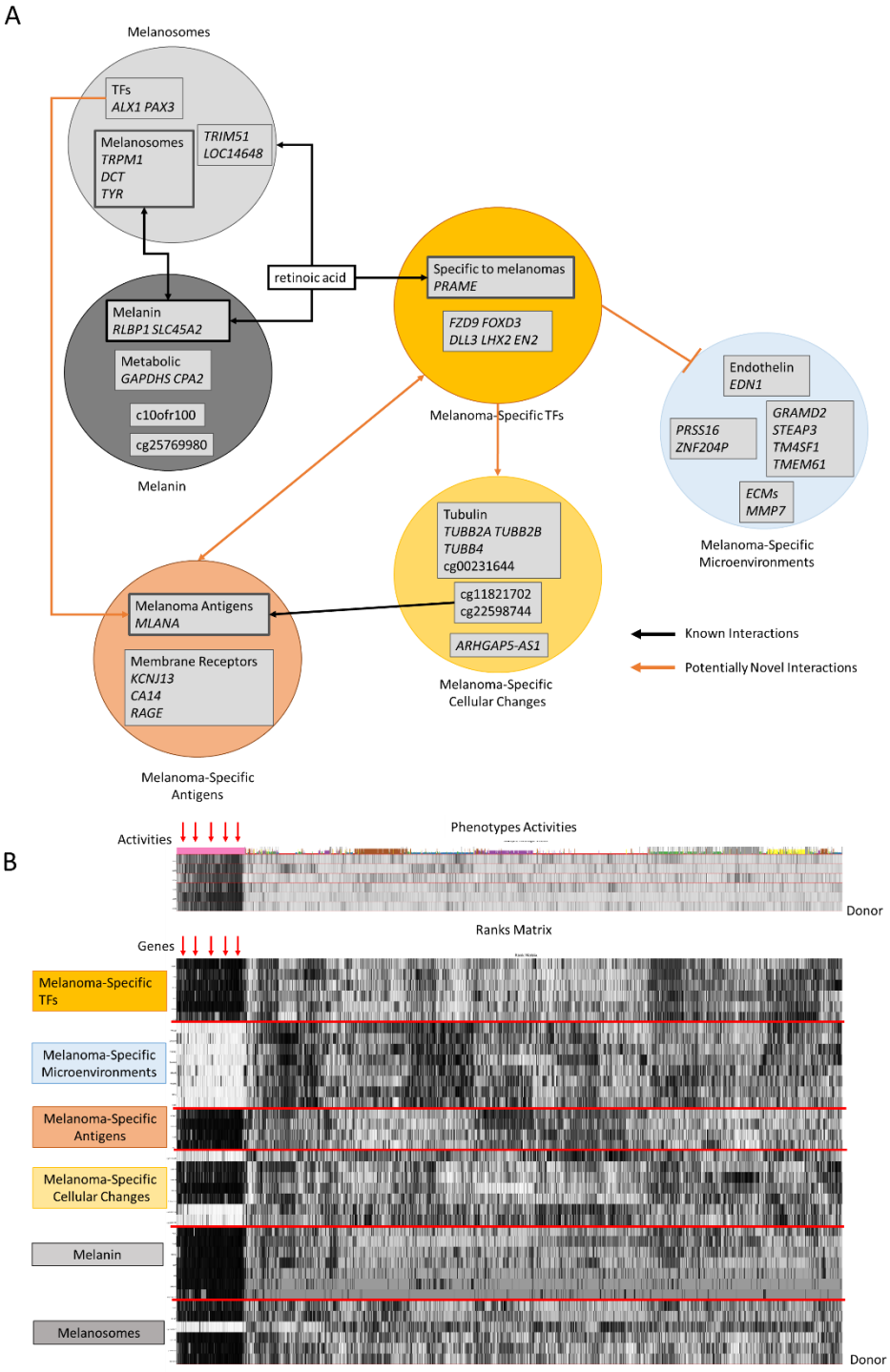



*Figure 36 2-omics Networks in Melanoma. (A) 6 units containing genes and CpG sites in melanoma (SKCM) network. Interactions between MLANA, a melanoma specific antigen, and 2 CpG sites annotated to MLANA are captured. (B) Donor contribution vectors (Upper) and rank matrix (Lower) of the network. (Red arrows). Columns of Melanoma samples.*

## Ineffective Wnt Pathway Negative Feedback in COAD

The Wnt signaling pathway is an important pathway that governs cell fate and development and is frequently perturbed in cancer cells, especially in colorectal cancer. This pathway is also a target of drug interventions (Zhan, Rindtorff, & Boutros, 2017) (Schatoff, Leach, & Dow, 2017) (Novellasdemunt, Antas, & Li, 2015). The pathway consists of two routes after the binding of Wnt protein to the Fizzled family of membrane receptors, canonical and noncanonical pathways. These signals are then relayed and amplified and are under the regulation of many factors, forming complicated positive and negative feedback loops.

Reported Wnt pathway negative feedback loop regulators include *AXIN2*, *NKD1*, *NKD2*, *NOTUM* and DKK family genes. *AXIN2* is shown to be directly upregulated by Wnt activation. *AXIN2* stops Wnt signaling by destabilizing β-catenin (Jho et al., 2002). Another characterized gene is *NKD1*. *NKD1* interacts with *DVL2* to negatively regulate Wnt signaling (Larraguibel et al., 2015).

In colon cancer (TCGA COAD), the Wnt pathway is very frequently activated. One might expect the negative feedback loop to be inactivated, either by repressed expressions or mutations. This turned out not to be the case; in the TCGA COAD project no recurrent mutations in the above genes were reported. More intriguingly, the expression of these genes was significantly upregulated in COAD specimens (Figure 37; right side; box plot). These negative feedback loops were mainly represented in 3 units (Figure 37; center circles). *DVL2* was located on its own (Figure 37; Left Circle; *DVL2*). *AXIN2* was associated with *FGF18*, a known downstream target of the Wnt pathway, and *BMP4-SAMD6*, genes involved in TGF-β signaling pathways. Interactions between the Wnt and TGF-β pathways have been reported in normal (Attisano & Labbé, 2004; Attisano & Wrana, 2013) and cancer cells (Vallée, Lecarpentier, Guillevin, & Vallée, 2017; Warner, Greene, & Pisano, 2005). *NKD1*, *NKD2* and *NOTUM* were associated with *KIAA1199*, a proliferative signal molecule located downstream of Wnt, and *SLC6A6*, a transporter associated with increased survival of colorectal cancer cells. *DKK4* was associated with

functionally unclear genes that could mediate transcriptional regulation (*POU5F1B*), signal transduction (*LY6G6D*) and metabolic effects (*CEL*, *CELP*, *TG*). Collectively, these observations suggested that the previously identified negative feedback loops in the Wnt pathway work together. However, they were not effective at shutting down the Wnt pathway in COAD donors. In the *NKD1*, COAD donor ranks of the *DVL2* unit were significantly lower than the other cases (Figure 37; Left; Boxplot p value=$1.2*10^{-117}$, t test on ranks), rendering *NKD1* ineffective. However, the other regulators need further studies. These findings highlight the strength of the network analysis in providing a comprehensive view and revealing biologically significant phenomena.

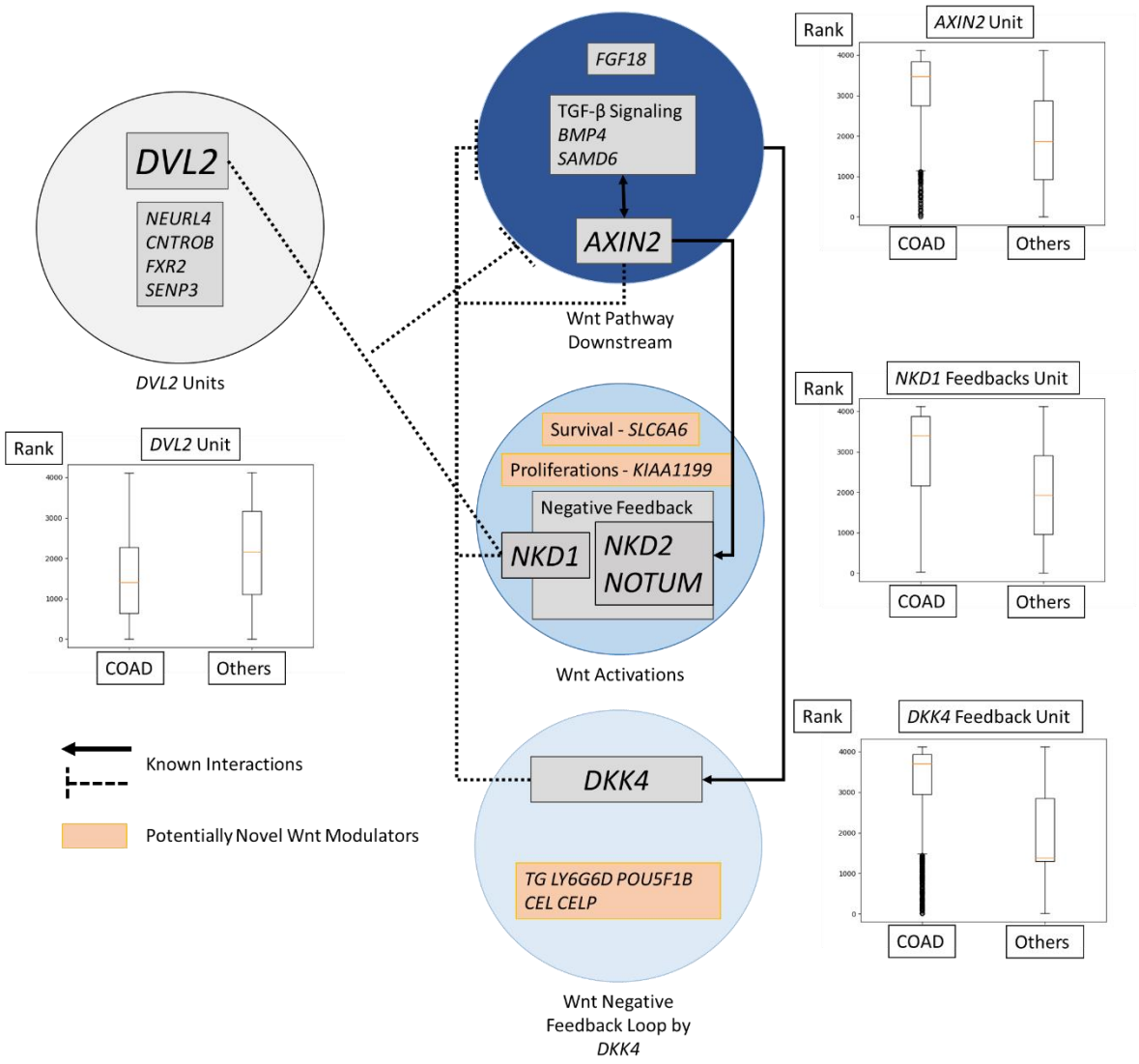



*Figure 37 Collection of Wnt Negative Feedback Loop Units. (Upper blue circle) AXIN2 and other Wnt downstream targets. (Middle blue circle) NKD1, NKD2 and NOTUM unit. (Lower blue circle) DKK4 unit. (Grey circle) DVL2 unit. All negative feedback genes (Blue circles) have higher expression in COAD (Right boxplots). NKD1 might be rendered ineffective in COAD by lower expression of DVL2 (Left boxplot).*

## Discussion

In this chapter, I expanded the one-to-one regulatory interaction conducted in Chapter I into networks of 2-omics interacting features by rank-based multi-omics network analysis in 8 pan-cancer TCGA projects. From the total of 15,666 genes and 12,835 CpG sites of 4,116 donors, I constructed 4,358 functional units of strongly synchronized features. These were linked to form 654 networks by phenotype activities. Twenty-nine of

these networks contained mixed interactions between genes and CpG sites. Not all of the units and networks were found to be biologically meaningful due to inherent complexities in the biological system and the simplicities in rank analysis. Cross-referencing with known gene sets, Gene Ontologies and literal searches revealed that a significant number of the networks and units were biologically significant and suggestive of novel biological interactions. A network containing the known interactions of *AURKA* and *AURKB* in cell cycle regulations was illustrated. Due to the interpretation approach, not every potentially meaningful network could be analyzed, and it is possible that the networks not reported here could still hold significant interactions.

While I concluded that a nonparametric rank analysis approach would yield the most appropriate methods in this work, it still had a number of drawbacks. Due to its root in coexpression networks, complex, multifactor interactions would not be apparent. The lack of overall detection power of the rank analysis also hindered the detection of such interactions. This was circumvented by increasing the detection power with a larger input size. Another disadvantage was its oversimplification, and any detected interactions would need to be interpreted based on their known biological function alone. Rank analysis also helped suppress artifacts from noise and batch effects. Overall, I believe that the benefits of rank analysis outweigh the drawbacks.

Many improvements could be made on the rank analysis methods. Effects of zeros and outliners, in particular, could be removed by trimming off rankings from zeros and extreme measurements and focusing on rankings of more continuous measurements in the middle. Tie correction is also another area that could be improved to make the rankings better represent the data. Lastly, parameters from already proven gene networks could be incorporated to increase the detection power and the accuracy of the results.

I consider these results not as a complete multi-omics networks atlas on any scale but as a collection of biologically relevant and potentially functionally important pan-cancer networks of detected genetic and epigenetic interactions. These networks and the

analysis employed might serve to improve our understanding of the cancer genomes and transcription regulations.

# Conclusion

In this thesis, I explored and identified interactions of genetic and epigenetic elements, both from the same and different omics, by a combination of multi-omics and network analysis. By first integrating genomics mutations, mRNA expression, histone modifications and long read allele configurations of 23 lung adenocarcinoma cell line datasets, I identified and validated the regulatory elements, their importance in cancer genomes and their long-range *cis*-interactions with their transcriptome counterparts, providing a solid platform and proof-of-concept evidence of the role of the regulatory elements in cancer and established integrative analysis of multi-omics dataset as an approach to studying the interaction between omics.

To follow up on those results, I moved on to the pan-cancer multi-omics network level by an integrative expression and methylations analysis of 8 TCGA projects. I was able to identify and characterize both known and novel interactions of genes-to-genes and genes-to-CpG sites. These interactions range from single gene or CpG site resolution to functional and biological process level resolution. Establishing rank analysis as an approach to integrate features from different omics uniformly and its synergistic effects with network analysis to produce comprehensive views of biological systems.

Overall, this work described regulatory mutations genetic and epigenetic interactions and their potential roles in the development of cancer phenotypes. While not intended to be an complete atlas of interactions, the results could be applied directly and conceptually to advance our insights on how genotypes interact and translate into phenotypes.

# Reference


Abyzov, A., Urban, A. E., Snyder, M., & Gerstein, M. (2011). CNVnator: An approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Research, 21*(6), 974-984.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., . . . Sandelin, A. (2014). An atlas of active enhancers across human cell types and tissues. *Nature, 507*, 455.

Attisano, L., & Labbé, E. (2004). TGFβ and Wnt pathway cross-talk. *Cancer and Metastasis Reviews, 23*(1), 53-61. doi:10.1023/A:1025811012690

Attisano, L., & Wrana, J. L. (2013). Signal integration in TGF-β, WNT, and Hippo pathways. *F1000prime reports, 5*, 17-17. doi:10.12703/P5-17

Baran, Y., Subramaniam, M., Biton, A., Tukiainen, T., Tsang, E. K., Rivas, M. A., . . . Lappalainen, T. (2015). The landscape of genomic imprinting across diverse adult human tissues. *Genome Research, 25*(7), 927-936.

Barlow, D. P., & Bartolomei, M. S. (2014). Genomic Imprinting in Mammals. *Cold Spring Harbor Perspectives in Biology, 6*(2).

Behjati, S., & Tarpey, P. S. (2013). What is next generation sequencing? *Archives of disease in childhood. Education and practice edition, 98*(6), 236-238. doi:10.1136/archdischild-2013-304340

Bembom, O. (2017). seqLogo: Sequence logos for DNA sequence alignments.

Berg, J., Tymoczko, J., & Stryer, L. (2002). Biochemistry. 5th edition. In. New York: W H Freeman.

Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C., Snyder, M., & Consortium, E. P. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature, 489*(7414), 57-74. doi:10.1038/nature11247

Bernstein, B. E., Stamatoyannopoulos, J. A., Costello, J. F., Ren, B., Milosavljevic, A., Meissner, A., . . . Thomson, J. A. (2010). The NIH Roadmap Epigenomics Mapping Consortium. *Nature biotechnology, 28*(10), 1045-1048. doi:10.1038/nbt1010-1045

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., & Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians, 68*(6), 394-424. doi:10.3322/caac.21492

Cantone, I., & Fisher, A. G. (2013). Epigenetic programming and reprogramming during development. *Nature Structural &Amp; Molecular Biology, 20*, 282. doi:10.1038/nsmb.2489

Chan, B. A., & Hughes, B. G. M. (2014). Targeted therapy for non-small cell lung cancer: current standards and the promise of the future. *Translational Lung Cancer Research, 4*(1), 36-54.

Chmielecki, J., & Meyerson, M. (2014). DNA Sequencing of Cancer: What Have We Learned? *Annual Review of Medicine, 65*(1), 63-79. doi:10.1146/annurev-med-060712-200152

Collisson, E. A., Campbell, J. D., Brooks, A. N., Berger, A. H., Lee, W., Chmielecki, J., . . . John Flynn, H. (2014). Comprehensive molecular profiling of lung adenocarcinoma. *Nature, 511*(7511), 543-550. doi:10.1038/nature13385

Cooper, G. (2000). *The Cell: A Molecular Approach. 2nd Edition.* Sunderland ( MA): Sinauer Associates.

Crick, F. (1970). Central Dogma of Molecular Biology. *Nature, 227*(5258), 561-563. doi:10.1038/227561a0

Cullen, S. M., Mayle, A., Rossi, L., & Goodell, M. A. (2014). Chapter Two - Hematopoietic Stem Cell Development: An Epigenetic Journey. In M. Rendl (Ed.), *Current Topics in Developmental Biology* (Vol. 107, pp. 39-75): Academic Press.

Dabin, J., Fortuny, A., & Polo, S. E. (2016). Epigenome Maintenance in Response to DNA Damage. *Molecular cell, 62*(5), 712-727. doi:10.1016/j.molcel.2016.04.006

Dahlberg, P. S., Jacobson, B. A., Dahal, G., Fink, J. M., Kratzke, R. A., Maddaus, M. A., & Ferrin, L. J. (2004). ERBB2 Amplifications in Esophageal Adenocarcinoma. *The Annals of Thoracic Surgery, 78*(5), 1790-1800. doi:https://doi.org/10.1016/j.athoracsur.2004.05.037

Davis, C. A., Hitz, B. C., Sloan, C. A., Chan, E. T., Davidson, J. M., Gabdank, I., . . . Cherry, J. M. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic acids research, 46*(D1), D794-D801. doi:10.1093/nar/gkx1081

Del Re, M., Tiseo, M., Bordi, P., D'Incecco, A., Camerini, A., Petrini, I., . . . Danesi, R. (2017). Contribution of KRAS mutations and c.2369C > T (p.T790M) EGFR to acquired resistance to EGFR-TKIs in EGFR mutant NSCLC: a study on circulating tumor DNA. *Oncotarget, 8*(8), 13611-13619. doi:10.18632/oncotarget.6957

Dela Cruz, C. S., Tanoue, L. T., & Matthay, R. A. (2011). Lung cancer: epidemiology, etiology, and prevention. *Clinics in chest medicine, 32*(4), 605-644. doi:10.1016/j.ccm.2011.09.001

Delgado, F. M., & Gómez-Vela, F. (2019). Computational methods for Gene Regulatory Networks reconstruction and analysis: A review. *Artificial Intelligence in Medicine, 95*, 133-145. doi:https://doi.org/10.1016/j.artmed.2018.10.006

Ferlay, J., Colombet, M., Soerjomataram, I., Mathers, C., Parkin, D. M., Piñeros, M., . . . Bray, F. (2019). Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *International Journal of Cancer, 144*(8), 1941-1953. doi:10.1002/ijc.31937

Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., . . . Campbell, P. J. (2015). COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res, 43*(Database issue), D805-811. doi:10.1093/nar/gku1075

Gnyszka, A., JastrzĘBski, Z., & Flis, S. (2013). DNA Methyltransferase Inhibitors and Their Emerging Role in Epigenetic Therapy of Cancer. *Anticancer Research, 33*(8), 2989-2996.

Greenberg, M. V. C., & Bourc'his, D. (2019). The diverse roles of DNA methylation in mammalian development and disease. *Nature Reviews Molecular Cell Biology, 20*(10), 590-607. doi:10.1038/s41580-019-0159-6

Grob, T. J., Kannengiesser, I., Tsourlakis, M. C., Atanackovic, D., Koenig, A. M., Vashist, Y. K., . . . Wilczak, W. (2012). Heterogeneity of ERBB2 amplification in adenocarcinoma, squamous cell carcinoma and large cell undifferentiated carcinoma of the lung. *Modern Pathology, 25*(12), 1566-1573. doi:10.1038/modpathol.2012.125

Heim, L., Friedrich, J., Engelhardt, M., Trufa, D. I., Geppert, C. I., Rieker, R. J., . . . Finotto, S. (2018). NFATc1 Promotes Antitumoral Effector Functions and Memory CD8<sup>+</sup> T-cell Differentiation during Non–Small Cell Lung Cancer Development. *Cancer Research, 78*(13), 3619. doi:10.1158/0008-5472.CAN-17-3297

Heinz, S., Romanoski, C. E., Benner, C., & Glass, C. K. (2015). The selection and function of cell type-specific enhancers. *Nat Rev Mol Cell Biol, 16*(3), 144-154. doi:10.1038/nrm3949

Holohan, C., Van Schaeybroeck, S., Longley, D. B., & Johnston, P. G. (2013). Cancer drug resistance: an evolving paradigm. *Nature Reviews Cancer, 13*(10), 714-726. doi:10.1038/nrc3599

Hon, C.-C., Ramilowski, J. A., Harshbarger, J., Bertin, N., Rackham, O. J. L., Gough, J., . . . Forrest, A. R. R. (2017). An atlas of human long non-coding RNAs with accurate 5′ ends. *Nature, 543*, 199.

Huang, F. W., Hodis, E., Xu, M. J., Kryukov, G. V., Chin, L., & Garraway, L. A. (2013). Highly recurrent TERT promoter mutations in human melanoma. *Science, 339*(6122), 957-959. doi:10.1126/science.1229259

International Cancer Genome Consortium, Hudson, T. J., Anderson, W., Artez, A., Barker, A. D., Bell, C., . . . Yang, H. (2010). International network of cancer genome projects. *Nature, 464*(7291), 993-998. doi:10.1038/nature08987

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology, 17*(1), 239. doi:10.1186/s13059-016-1103-0

Jho, E.-h., Zhang, T., Domon, C., Joo, C.-K., Freund, J.-N., & Costantini, F. (2002). Wnt/β-Catenin/Tcf Signaling Induces the Transcription of Axin2, a Negative Regulator of the Signaling Pathway. *Molecular and Cellular Biology, 22*(4), 1172. doi:10.1128/MCB.22.4.1172-1183.2002

Jovanovic, M., Rooney, M. S., Mertins, P., Przybylski, D., Chevrier, N., Satija, R., . . . Regev, A. (2015). Dynamic profiling of the protein life cycle in response to pathogens. *Science, 347*(6226), 1259038. doi:10.1126/science.1259038

Kallioniemi, O. P., Kallioniemi, A., Kurisu, W., Thor, A., Chen, L. C., Smith, H. S., . . . Gray, J. W. (1992). ERBB2 amplification in breast cancer analyzed by fluorescence in situ hybridization. *Proceedings of the National Academy of Sciences of the United States of America, 89*(12), 5321-5325. doi:10.1073/pnas.89.12.5321

Kel, A. E., Gossling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O. V., & Wingender, E. (2003). MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res, 31*(13), 3576-3579.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., & Haussler, D. (2002). The human genome browser at UCSC. *Genome Res, 12*(6), 996-1006. doi:10.1101/gr.229102. Article published online before print in May 2002

Khurana, E., Fu, Y., Chakravarty, D., Demichelis, F., Rubin, M. A., & Gerstein, M. (2016). Role of non-coding sequence variants in cancer. *Nature Reviews Genetics, 17*, 93. doi:10.1038/nrg.2015.17

Klemm, S. L., Shipony, Z., & Greenleaf, W. J. (2019). Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics, 20*(4), 207-220. doi:10.1038/s41576-018-0089-8

Kulis, M., & Esteller, M. (2010). 2 - DNA Methylation and Cancer. In Z. Herceg & T. Ushijima (Eds.), *Advances in Genetics* (Vol. 70, pp. 27-56): Academic Press.

Lambert, S. A., Jolma, A., Campitelli, L. F., Das, P. K., Yin, Y., Albu, M., . . . Weirauch, M. T. (2018). The Human Transcription Factors. *Cell, 172*(4), 650-665. doi:https://doi.org/10.1016/j.cell.2018.01.029

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics, 9*(1), 559. doi:10.1186/1471-2105-9-559

Larraguibel, J., Weiss, A. R. E., Pasula, D. J., Dhaliwal, R. S., Kondra, R., & Van Raay, T. J. (2015). Wnt ligand-dependent activation of the negative feedback regulator Nkd1. *Molecular biology of the cell, 26*(12), 2375-2384. doi:10.1091/mbc.E14-12-1648

Law, J. A., & Jacobsen, S. E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature reviews. Genetics, 11*(3), 204-220. doi:10.1038/nrg2719

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics, 25*(14), 1754-1760. doi:10.1093/bioinformatics/btp324

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., . . . Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics, 25*(16), 2078-2079. doi:10.1093/bioinformatics/btp352

Li, S., Li, L., Zhu, Y., Huang, C., Qin, Y., Liu, H., . . . Liang, N. (2014). Coexistence of EGFR with KRAS, or BRAF, or PIK3CA somatic mutations in lung cancer: a comprehensive mutation profiling from 5125 Chinese cohorts. *British Journal of Cancer, 110*(11), 2812-2820. doi:10.1038/bjc.2014.210

Liang, H., Pan, Z., Wang, W., Guo, C., Chen, D., Zhang, J., . . . written on behalf of, A. M. E. L. C. C. G. (2018). The alteration of T790M between 19 del and L858R in NSCLC in the course of EGFR-TKIs therapy: a literature-based pooled analysis. *Journal of thoracic disease, 10*(4), 2311-2320. doi:10.21037/jtd.2018.03.150

Liao, J., Karnik, R., Gu, H., Ziller, M. J., Clement, K., Tsankov, A. M., . . . Meissner, A. (2015). Targeted disruption of DNMT1, DNMT3A and DNMT3B in human embryonic stem cells. *Nature genetics, 47*(5), 469-478. doi:10.1038/ng.3258

Liberti, M. V., & Locasale, J. W. (2016). The Warburg Effect: How Does it Benefit Cancer Cells? *Trends in biochemical sciences, 41*(3), 211-218. doi:10.1016/j.tibs.2015.12.001

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., & Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics (Oxford, England), 27*(12), 1739-1740. doi:10.1093/bioinformatics/btr260

Liu, X., Gao, Q., Li, P., Zhao, Q., Zhang, J., Li, J., . . . Wong, J. (2013). UHRF1 targets DNMT1 for DNA methylation through cooperative binding of hemi-methylated DNA and methylated H3K9. *Nature Communications, 4*(1), 1563. doi:10.1038/ncomms2562

Liu, Y., Mi, Y., Mueller, T., Kreibich, S., Williams, E. G., Van Drogen, A., . . . Aebersold, R. (2019). Multi-omic measurements of heterogeneity in HeLa cells across laboratories. *Nature Biotechnology, 37*(3), 314-322. doi:10.1038/s41587-019-0037-y

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., . . . Moore, H. F. (2013). The Genotype-Tissue Expression (GTEx) project. *Nature Genetics, 45*(6), 580-585. doi:10.1038/ng.2653

Ma, C., Wei, S., & Song, Y. (2011). T790M and acquired resistance of EGFR TKI: a literature review of clinical reports. *Journal of thoracic disease, 3*(1), 10-18. doi:10.3978/j.issn.2072-1439.2010.12.02

Mancini, M., & Toker, A. (2009). NFAT proteins: emerging roles in cancer progression. *Nat Rev Cancer, 9*(11), 810-820. doi:10.1038/nrc2735

Matys, V., Kel-Margoulis, O. V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., . . . Wingender, E. (2006). TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res, 34*(Database issue), D108-110. doi:10.1093/nar/gkj143

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., . . . DePristo, M. A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome research, 20*(9), 1297-1303. doi:10.1101/gr.107524.110

Melé, M., Ferreira, P. G., Reverter, F., DeLuca, D. S., Monlong, J., Sammeth, M., . . . Guigó, R. (2015). The human transcriptome across tissues and individuals. *Science, 348*(6235), 660. doi:10.1126/science.aaa0355

Minoo, P. (2000). Transcriptional regulation of lung development: emergence of specificity. *Respiratory research, 1*(2), 109-115. doi:10.1186/rr20

Morison, I. M., Paton, C. J., & Cleverley, S. D. (2001). The imprinted gene and parent-of-origin effect database. *Nucleic Acids Res, 29*(1), 275-276.

Novellasdemunt, L., Antas, P., & Li, V. S. W. (2015). Targeting Wnt signaling in colorectal cancer. A Review in the Theme: Cell Signaling: Proteins, Pathways and Mechanisms. *American Journal of Physiology-Cell Physiology, 309*(8), C511-C521. doi:10.1152/ajpcell.00117.2015

Olive, V., Jiang, I., & He, L. (2010). mir-17-92, a cluster of miRNAs in the midst of the cancer network. *The international journal of biochemistry & cell biology, 42*(8), 1348-1354. doi:10.1016/j.biocel.2010.03.004

Peng, Y., & Croce, C. M. (2016). The role of MicroRNAs in human cancer. *Signal Transduction and Targeted Therapy, 1*(1), 15004. doi:10.1038/sigtrans.2015.4

Plotnik, J. P., Budka, J. A., Ferris, M. W., & Hollenhorst, P. C. (2014). ETS1 is a genome-wide effector of RAS/ERK signaling in epithelial cells. *Nucleic Acids Research, 42*(19), 11928-11940.

Pollard, M. O., Gurdasani, D., Mentzer, A. J., Porter, T., & Sandhu, M. S. (2018). Long reads: their purpose and place. *Human Molecular Genetics, 27*(R2), R234-R241. doi:10.1093/hmg/ddy177

Redwine, J. M., & Evans, C. F. (2002). Markers of Central Nervous System Glia and Neurons In Vivo During Normal and Pathological Conditions. In B. Dietzschold & J. A. Richt (Eds.), *Protective and Pathological Immune Responses in the CNS* (pp. 119-140). Berlin, Heidelberg: Springer Berlin Heidelberg.

Robbs, B. K., Cruz, A. L. S., Werneck, M. B. F., Mognol, G. P., & Viola, J. P. B. (2008). Dual Roles for NFAT Transcription Factor Genes as Oncogenes and Tumor Suppressors. *Molecular and Cellular Biology, 28*(23), 7168. doi:10.1128/MCB.00256-08

Roller, E., Ivakhno, S., Lee, S., Royce, T., & Tanner, S. (2016). Canvas: versatile and scalable detection of copy number variants. *Bioinformatics, 32*(15), 2375-2377. doi:10.1093/bioinformatics/btw163

Schatoff, E. M., Leach, B. I., & Dow, L. E. (2017). Wnt Signaling and Colorectal Cancer. *Current colorectal cancer reports, 13*(2), 101-110. doi:10.1007/s11888-017-0354-9

Shen, R., Olshen, A. B., & Ladanyi, M. (2009). Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics (Oxford, England), 25*(22), 2906-2912. doi:10.1093/bioinformatics/btp543

Sigismund, S., Avanzato, D., & Lanzetti, L. (2018). Emerging functions of the EGFR in cancer. *Molecular oncology, 12*(1), 3-20. doi:10.1002/1878-0261.12155

Sonawane, A. R., Platig, J., Fagny, M., Chen, C.-Y., Paulson, J. N., Lopes-Ramos, C. M., . . . Kuijjer, M. L. (2017). Understanding Tissue-Specific Gene Regulation. *Cell reports, 21*(4), 1077-1088. doi:10.1016/j.celrep.2017.10.001

Speir, M. L., Zweig, A. S., Rosenbloom, K. R., Raney, B. J., Paten, B., Nejad, P., . . . Kent, W. J. (2016). The UCSC Genome Browser database: 2016 update. *Nucleic Acids Res, 44*(D1), D717-725. doi:10.1093/nar/gkv1275

Suzuki, A., Makinoshima, H., Wakaguri, H., Esumi, H., Sugano, S., Kohno, T., . . . Suzuki, Y. (2014). Aberrant transcriptional regulations in cancers: genome, transcriptome and epigenome analysis of lung adenocarcinoma cell lines. *Nucleic Acids Res, 42*(22), 13557-13572. doi:10.1093/nar/gku885

Suzuki, A., Wakaguri, H., Yamashita, R., Kawano, S., Tsuchihara, K., Sugano, S., . . . Nakai, K. (2015). DBTSS as an integrative platform for transcriptome, epigenome and genome sequence variation data. *Nucleic Acids Res, 43*(Database issue), D87-91. doi:10.1093/nar/gku1080

The Cancer Genome Atlas Research, N., Chang, K., Creighton, C. J., Davis, C., Donehower, L., Drummond, J., . . . Shaw, K. R. M. (2013). The Cancer Genome Atlas Pan-Cancer analysis project. *Nature Genetics, 45*, 1113. doi:10.1038/ng.2764

The Cancer Genome Atlas Research Network. (2014). Comprehensive molecular profiling of lung adenocarcinoma. *Nature, 511*(7511), 543-550. doi:10.1038/nature13385

TheAmericanCancerSociety. (2019). Targeted Therapy for Non-Small Cell Lung Cancer.

Vallée, A., Lecarpentier, Y., Guillevin, R., & Vallée, J.-N. (2017). Interactions between TGF-β1, canonical WNT/β-catenin pathway and PPAR γ in radiation-induced fibrosis. *Oncotarget, 8*(52), 90579-90604. doi:10.18632/oncotarget.21234

Vihervaara, A., Duarte, F. M., & Lis, J. T. (2018). Molecular mechanisms driving transcriptional stress responses. *Nature Reviews Genetics, 19*(6), 385-397. doi:10.1038/s41576-018-0001-6

Vinagre, J., Almeida, A., Pópulo, H., Batista, R., Lyra, J., Pinto, V., . . . Soares, P. (2013). Frequency of TERT promoter mutations in human cancers. *Nature Communications, 4*(1), 2185. doi:10.1038/ncomms3185

Vineis, P., Schatzkin, A., & Potter, J. D. (2010). Models of carcinogenesis: an overview. *Carcinogenesis, 31*(10), 1703-1709. doi:10.1093/carcin/bgq087

Warner, D. R., Greene, R. M., & Pisano, M. M. (2005). Cross-talk between the TGFβ and Wnt signaling pathways in murine embryonic maxillary mesenchymal cells. *FEBS Letters, 579*(17), 3539-3546. doi:10.1016/j.febslet.2005.05.024

Wee, P., & Wang, Z. (2017). Epidermal Growth Factor Receptor Cell Proliferation Signaling Pathways. *Cancers, 9*(5), 52. doi:10.3390/cancers9050052

Weinberg, R. A. (2013). *The Biology of Cancer, Second Edition*: Garland Science.

Wong, K. K., Lawrie, C. H., & Green, T. M. (2019). Oncogenic Roles and Inhibitors of DNMT1, DNMT3A, and DNMT3B in Acute Myeloid Leukaemia. *Biomarker insights, 14*, 1177271919846454-1177271919846454. doi:10.1177/1177271919846454

Xu, S., Shu, P., Zou, S., Shen, X., Qu, Y., Zhang, Y., . . . Zhang, J. (2018). NFATc1 is a tumor suppressor in hepatocellular carcinoma and induces tumor cell apoptosis by activating the FasL-mediated extrinsic signaling pathway. *Cancer Medicine, 7*(9), 4701-4717. doi:10.1002/cam4.1716

Xu, W., Gu, J., Ren, Q., Shi, Y., Xia, Q., Wang, J., . . . Wang, J. (2016). NFATC1 promotes cell growth and tumorigenesis in ovarian cancer up-regulating c-Myc through ERK1/2/p38 MAPK signal pathway. *Tumor Biology, 37*(4), 4493-4500. doi:10.1007/s13277-015-4245-x

Yun, C.-H., Mengwasser, K. E., Toms, A. V., Woo, M. S., Greulich, H., Wong, K.-K., . . . Eck, M. J. (2008). The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP. *Proceedings of the National Academy of Sciences, 105*(6), 2070. doi:10.1073/pnas.0709662105

Zhan, T., Rindtorff, N., & Boutros, M. (2017). Wnt signaling in cancer. *Oncogene, 36*(11), 1461-1473. doi:10.1038/onc.2016.304

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., . . . Liu, X. S. (2008). Model-based Analysis of ChIP-Seq (MACS). *Genome Biology, 9*(9), R137. doi:10.1186/gb-2008-9-9-r137

Zheng, G. X., Lau, B. T., Schnall-Levin, M., Jarosz, M., Bell, J. M., Hindson, C. M., . . . Ji, H. P. (2016). Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat Biotechnol, 34*(3), 303-311. doi:10.1038/nbt.3432

Zhu, Y., Qiu, P., & Ji, Y. (2014). TCGA-assembler: open-source software for retrieving and processing TCGA data. *Nat Methods, 11*(6), 599-600. doi:10.1038/nmeth.2956

# Acknowledgements

I would like to express my gratitude to my advisor, Professor Suzuki Yutaka, for all his kind and thoughtful guidance and teaching which made this thesis possible and to him I own all the experiences and opportunities I had during my 5 years' time.

I would like to express my gratitude to Professor Suzuki Ayako for her always kind and helpful teaching and advice and her central role in formulating the researches in this thesis and though those teaching and example, making this thesis possible.

I would like to express my gratitude to Professor Seki Masahide for his compassionate and kind teaching and advices and without his experimental expertise and guidance this thesis would not be possible.

I would like to express my gratitude to Professor Katsuya Tsuchihara and Professor Sugano Sumio, for their advices and supports had a strong impact on this thesis and empowered me to many valuable opportunities.

I would like to express my thanks and gratitude to everyone at Suzuki laboratory both currently and alumni, to whom I owned my laugh and joy during my 5 years stay, especially to Mr. Kunigo Keisuke and Dr. Maekawa Sho during my Masters years, Mr. Sakamoto Yoshitaka for his both his friendship and expertise in sequencing technologies and Dr. Runtuwene Ronald Lucky for his support and friendship.

I would like to express my thanks to my family, who were always supportive of my decisions and providing encouragements to always go forward.

I also would like to express my sincere thanks to The Japanese Ministry of Education, Culture, Sports, Science, and Technology (MEXT) for the financial support of MEXT scholarship from the very beginning of my Masters year and Embassy of Japan in Bangkok for the recommendation opportunity and I would also like to extend my thanks to Japanese tax-payers for funding this admirable scholarship program.