Department of Computational Biology and Medical Sciences
Graduate School of Frontier Sciences
The University of Tokyo

2017

Master's Thesis

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

Submitted January 27 2017

Adviser: Professor Suzuki Yutaka

Sereewattanawoot Sarun
セリーワッタナウート　サラン

# Contents

# Index of Figures

# Index of Tables

# Introduction

## Lung Adenocarcinoma

Lung cancer is the most prevalent cancer in the world and is the leading cause of death in both developing and developed countries [Charles S. Dela Cruz 2011]. Many environmental risk factors, including air pollution and the history of smoking have been indicated. However, despite the decrease in the exposure of risk factors, the incidence rate of lung cancer; especially, among patients who have never smoked is proportionally increasing, suggesting complex mechanism underneath.

Lung adenocarcinoma is a major histological subtype of lung cancer for both smokers and non-smokers and contributes half of the overall lung cancer incidents and it has been a subject of intensive studies [Charles S. Dela Cruz 2011] [TCGA 2014]. This is in contrast to the declining rate of squamous cell lung carcinoma, which is associated with patients who are smokers, as the number of smokers continue to decrease. These facts make the cancer research into lung adenocarcinoma worthwhile.

Many of the recurrent driver mutations unique to the adenocarcinoma subtype, such as nucleotide(s) substitutions or deletions in EGFR or KRAS and the gene fusions in ALK or RET has been identified. These advances in scientific research make lung adenocarcinoma one of the better well-

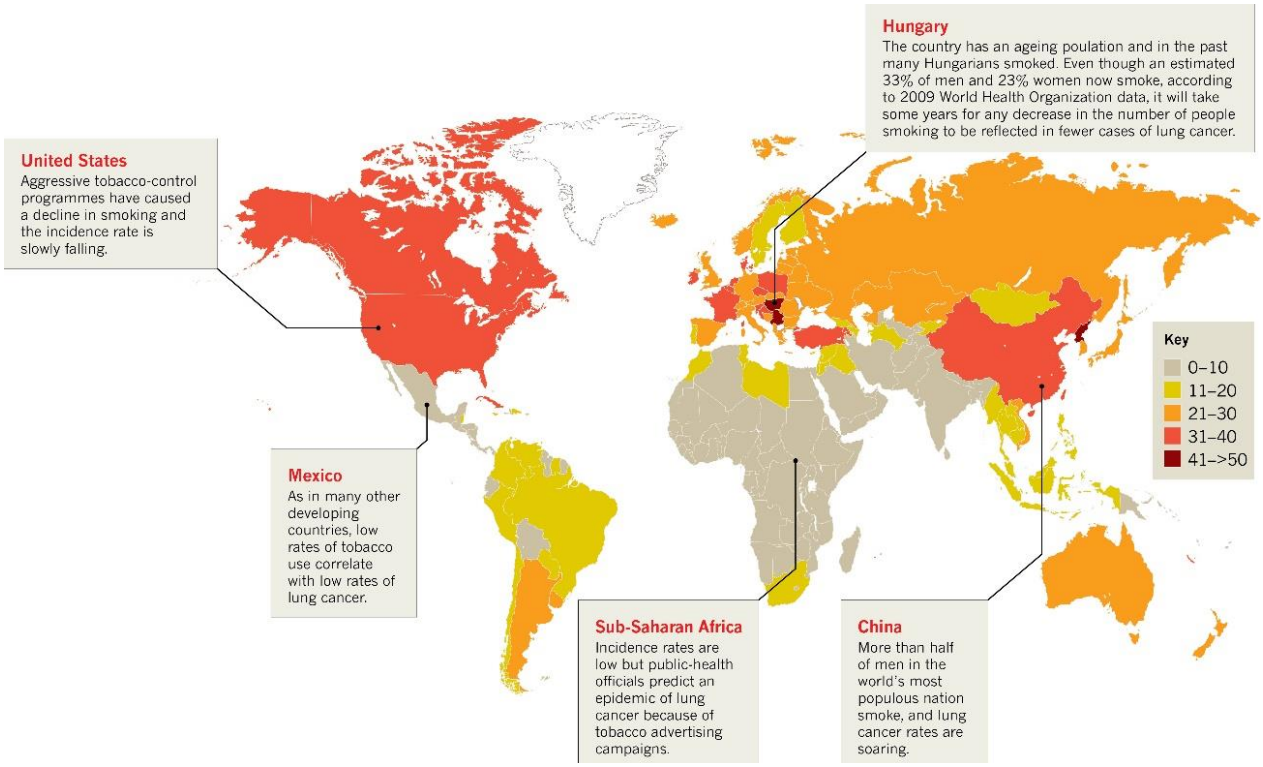


*Figure 1 Map showing prevalence of Lung Cancer of all types in 2012. From Pao and Hutchinson 2012 Nature Medicine*

characterized cancers regarding its causative driver mutations. For lung adenocarcinoma, several successful anti-cancer drugs have been developed [Hughes 2015], by using the causative driver mutations. Despite these successes, the majority of genetic causes of lung adenocarcinoma remain elusive [TCGA 2014]. Moreover, even the most powerful driver mutations are not solely responsible for the carcinogenesis process [Potter 2010] as long-term accumulation of other somatic mutations are required for thorough transformation of the cancer cell.

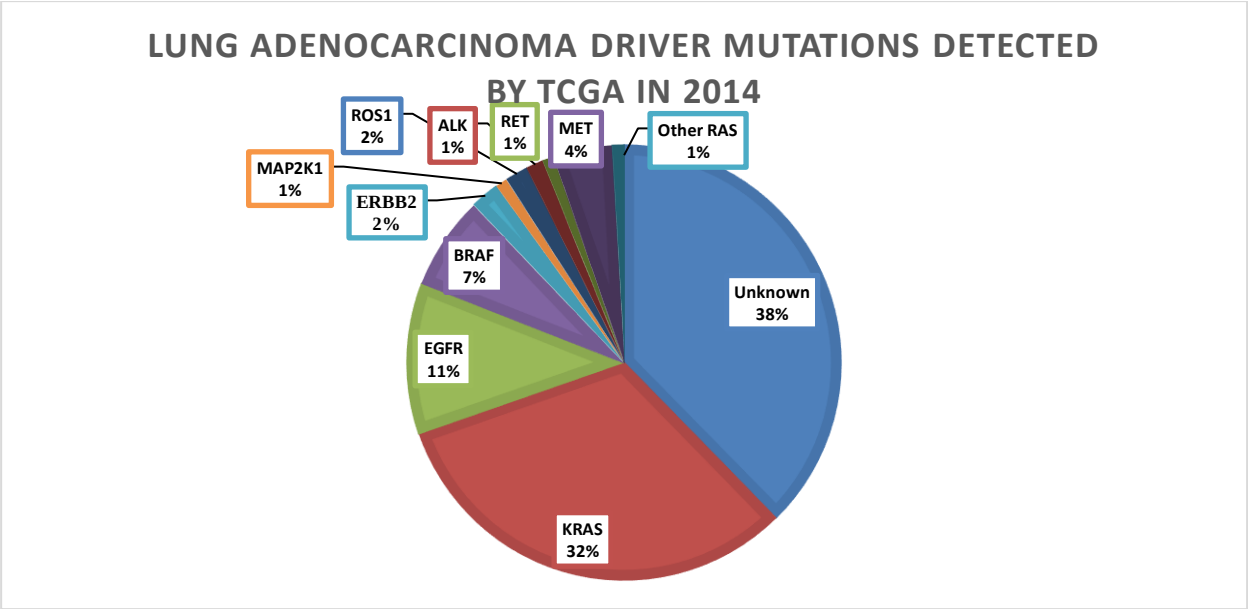


*Figure 2 Recurring Driver mutations detected in Lung Adenocarcinoma worldwide (Adapted from The Cancer Genome Atlas Research Network 2014 Nature).*

Interestingly, the mutation patterns seem to be significantly distinct depending on the ethnic background of patients. Recent studies had shown that, for East Asian patients, including Japanese and Chinese, have higher incidence of mutations in the EGFR gene (11.3% vs up to 50% reported in East Asia [Shiyong Li 2016]) and ALK fusion (1.3% vs 10% reported in East Asia [Shiyong Li 2016]) with lower incidence on KRAS mutations (32.2% vs 10% reported in East Asia [Shiyong Li 2016]) compared to Western counterparts. These studies claimed the necessity in collection of the mutation information from diverse ethnic background groups for comprehensive understanding of lung adenocarcinoma.

Through investigation of driver mutations or other genetic alterations, many carcinogenesis pathways have been identified, with the goal of explaining the molecular mechanism underlying the tumor development, as exemplified in Figure3 (Adapted from Chan BA, Hughes BGM 2015 Translational Lung Cancer Research). Genes with targeted drug therapies in development are shown in green and genes with developed targeted drug therapies are shown in red (EGFR mutations and ALK fusion). In addition to these lung adenocarcinoma specific mutations, other non-specific targets, such as VGFR, are also subjected to the feasibly studies for drug development [Hughes 2015]. These drug developments and lack of understanding of genetic alterations that cause carcinogenesis suggest the benefits of identifying undiscovered driver mutations or functionally relevant mutations to expand the chance for the development of effective cancer treatments.

## Recent Attempts of Cancer Genome Analysis

Recent short read sequencing, or so-called next generation sequencing, has provided us with a high-throughput method to study genomic mutations. Recently, several large-scale projects, including those by The Cancer Genome Atlas (TCGA) (The Cancer Genome Atlas 2014) and International Cancer Genome Consortium (ICGC) (International Cancer Genome Consortium 2010), have been established to characterize the genetic alterations, where many mutations in various cancer types have been identified and catalogued. Those studies focused their attention on the protein coding regions, many of which belong to tyrosine kinase family, since mutations in this family of genes would give pivotal information on the

design of the anti-cancer treatment regimens, one prominent example is the anti-EGFR drug in lung adenocarcinoma treatment. This demonstrate the impacts of large scale somatic mutation studies.

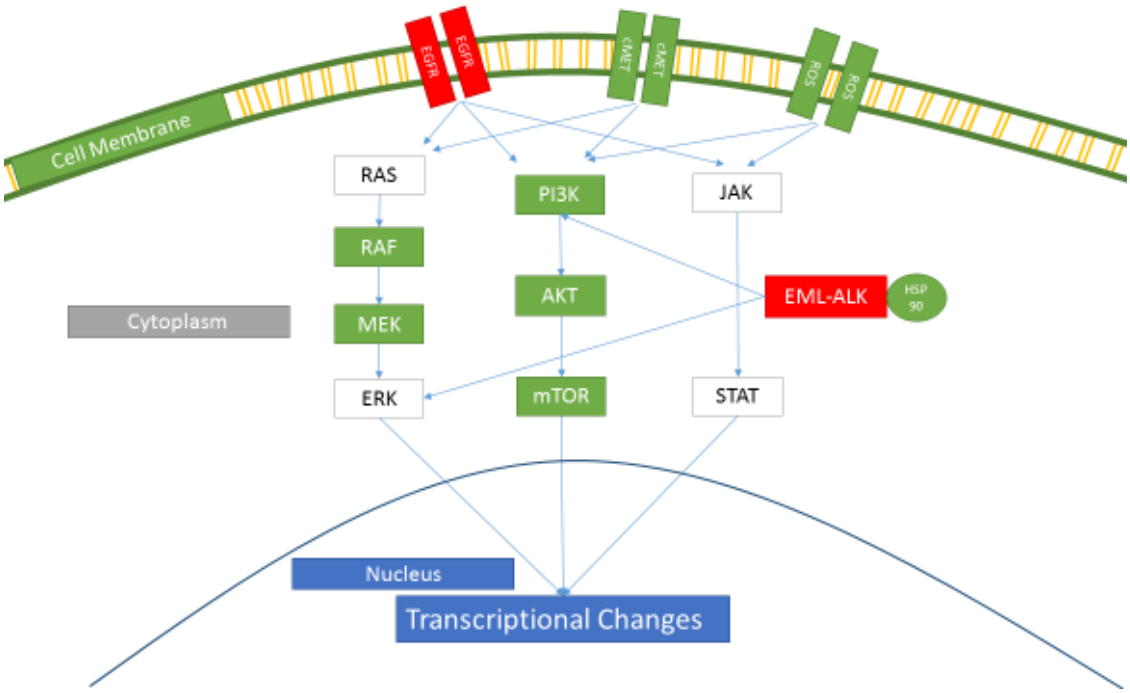


*Figure 3 Purposed cascaded constructed from frequently mutated genes in Lung Adenocarcinoma. Adapted from Chan BA, Hughes BGM 2015 Translational Lung Cancer Research. Potential components for targeted therapy are shown in green. Genes with current Targeted therapy usage are shown in red.*

Previous studies including large international consortia have predominantly focused on protein-coding genes. According to the current reference human genome (UCSC hg38 human genome assembly) only around 2% of the genome are the protein coding regions, consisting of at most 32,958 genes [Matthew L. Speir 2016]. It became clear that sequences and mutations in coding region alone could not explain the numerous observed phenotypes either at cellular level or the complex organismal level. Moreover, it is now widely accepted that nucleotides sequences outside of the coding regions also contain important information crucial to the genome, including the regulatory roles controlling genes expression patterns (Figure4). To document the role of the non-coding regions, many novel techniques have been developed [David S. Johnson 2007] [Nele Gheldof 2011] [Jason D Buenrostro 2013] and employed in projects such as ENCODE [ENCODE 2017] and ROADMAP [ROADMAP PROJECT 2017] . The achievements of these efforts have yielded the data archives, which serves as the pivotal database when any genome-wide researches in the gene expression regulations are conducted.

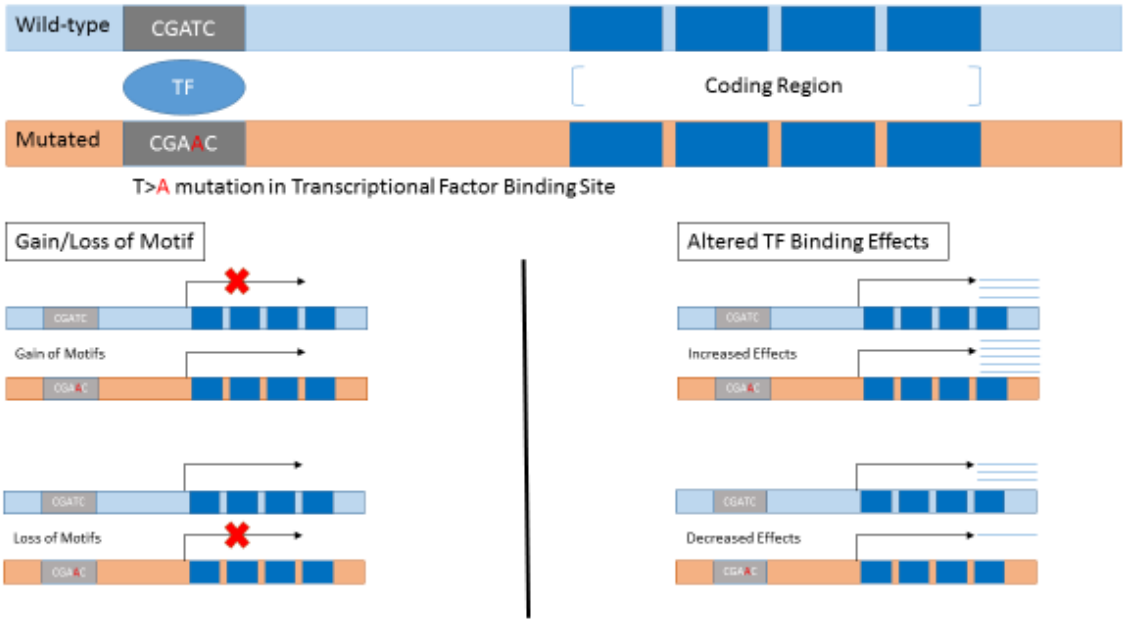


*Figure 4 Theorized effects of non-coding variant on gene regulation, noted that the focus have been put on cis-regulatory elements (Adapted from Ekta Khurana et al. 2016 Nature Review Genetics)*

One of the most unique and useful aspects of those databases is that the epigenome statuses in the region termed promoter and enhancer regions are represented. Various chromatins statuses, which are defined by specific histone modifications that bind those regions, are thought to play a major role in gene expression regulations [Ekta Khurana 2016]. Those specific histone modifications were detected by chromatin immunoprecipitation sequencing (ChIP-seq, see Table1 for details) analysis and form the body of the databases. While somatic mutations in these regulatory regions are widely detected in many cancer genomes, their functional relevancies remain elusive. It is supposed that changes in both sequences [13] and histone modifications [12] of these region could lead to alteration of cellular function and play less important roles in carcinogenesis than the mutations in protein-coding regions themselves.

*Table 1 shows examples of chromatin immunoprecipitation antibodies targets, their function and related regions.*

| MARKINGS | EFFECTS | REGION |
|---|---|---|
| **POLYMERASE-II** | Transcriptional Activations | RNA-Polymerase |
| **H3K4ME1** | Transcriptional Activations | Enhancer |
| **H3K4ME3** | Transcriptional Activations | Promoter |
| **H3K9ME3** | Repression | Heterochromatin and repetitive elements |
| **H3K9_14AC** | Transcriptional Activations | Promoter Preference |
| **H3K27AC** | Transcriptional Activations | Enhancer |
| **H3K27ME3** | Repression | Repressive Domain and Silencing |
| **H3K36ME3** | Transcriptional Elongations | Transcribed Regions |

## Multi-omics Analysis of Cancers

Many works, such as conducted by ENCODE or ROADMAP, have demonstrated that different cancer phenotypes that are associated with different "epigenome" landscapes, such as mutations in the promoter or enhancer regions or changes in histone modifications. However given the large size of non-coding region and variety of possible modifications, it is still difficult to select the cancer relevant of individual mutation and/or modification and conduct the validation analysis.

According to the current model of carcinogenesis, not every somatic mutations detected contribute to the process of carcinogenesis. These so-called "passenger mutations", while can be used as clonal marker, have little or no functional relevance [Potter 2010]. One of the ways to filter out these passenger mutations is to evaluate if the mutation in question has a functional relevance, a task that is still difficult even with current board knowledge of landscapes. Unlike the mutations in coding regions, annotations in non-coding regions are not always straightforward and sometimes not fully reliable.

To reveal functional relevance of mutations in promoter and enhancer regions, I intended to integrate epigenome and transcriptome data with the genomic mutation data in cancers, I expected that such multi-layered omics analysis would enable me to narrow the gap between observed phenotype and non-coding mutations

## Allelic Phasing as another Crucial Information

Another important barrier which potentially prevents the data integration between the genomic mutations with the transcriptome and epigenome data is the lack of allele specific information on the relative position of the promoter and enhancer regions and their regulating genic regions [ENCODE 2017]. Due to the technological limitation, allelic information was lost during conventional short-read sequencing; however, these information were indispensable when examining the biological relevance of the mutations in the regulatory regions. The allelic information is important as somatic mutations are commonly heterozygous, meaning that only gene expression from one allele may be affected, and by losing the allelic information, the gene expression inference is diluted.

Recently, novel technology that allow the synthetic long read to be constructed through short-read sequencers, through 10x Genomics' GemCode (10x Genomics) [Grace X Y Zheng 2016] . This technology complements the drawback that is inherent to the current short-read sequencing technologies through the uses of molecular barcoding technology, which allows allelic phasing of the human genome. In this method, large DNA fragments are confined in oil droplets together with gel-embedded barcodes (GEMs). By hybridization extension, each unique molecular identifier (UMI) is added to the DNA fragments within the droplet. Large-barcoded DNA fragments are, then mixed, sheared and then subjected to sequencing by Illumina short-read sequencer. Long read sequences originated from the large DNA fragments within a single droplet then could be computationally re-assembled based on the each unique molecular identifier (UMI). The read used in anchoring the fragments are called "linked reads" and play a key role in this so-called "synthetic long read sequencing".

Another novel development in long read technology is MinION from Oxford Nanopore Technologies, where instead of indexing and barcoding each read, this new technology physically read the long DNA sequences. In this sequencer, unlike the short-read sequencers, a long strand of DNA, which often reach tens of kilo bases long, passes through a small pore. Distinctive ion current disturbances caused by each of the four DNA bases are then analyzed in time-lapse manner.

Taking advantages of the both long read technologies, I attempted to obtain and validate the positional relationship between the SNVs in regulatory regions and the corresponding transcripts'

SNPs/SNVs. The goal of including phasing information into multi-omic analysis was to provide direct evidence for functional relevance of the mutation and its effects on gene expression regulation.

However, both of the long read technologies are very recent methods and have their own significant technical difficulties; therefore, they required technical optimisation before I could apply these methods for the analyses of cancer omics dataset. The 10x GemCode technology, is an indexing and barcoding-based system which was mainly designed to phase the diploid human genome, its direct adaption for usage in aneuploidy cancer genome was considered error-prone. On the other hand, the MinION sequencer's fidelity and throughput might not be sufficient for application in large cancer genome sequencing. Because of these potential drawbacks, I developed a series of tools for both methods to be used in this study. I generated the first phasing information based on the 10x GemCode technology then referenced and optimized the work flows by utilizing MinION sequencing dataset.

In this study, I intended to elucidate the transcriptional consequences of somatic mutations detected in the regulatory regions. Previous works [Ayako Suzuki 2014] at our laboratory identified a large number of mutations, in both coding and regulatory regions in 26 Lung Adenocarcinoma-derived cell lines. In addition, the multi-omics data from same materials have been collected and they include: histone modifications, transcriptional start sites and transcriptomes and I selected the candidates for the somatic regulatory mutations, by examining biased expression in variant tags between reference and alternative alleles in ChIP-seq and RNA-seq datasets. Then I associated those candidates' regulatory mutations with their regulating transcripts variants by the long read technologies.

# Material and Methods

## Cell lines used in this study

A total of 26 human lung adenocarcinoma cell lines were either cultured in RPMI medium (RPMI 1640, Nissui), Dulbecco's Modified Eagle's medium (Nissui) or Eagle's minimal essential medium (Nissui) along with 10% FBS, MEM Non-essential Amino acid solution (SIGMA) and antibiotics supplementation (Antibiotic-Antimycotic, GIBCO) and were kept at 37◦C and 5% CO2 condition. Three cell lines were tested positive for mycoplasma contamination and were excluded from this study. The basic cell line information and reported mutations [Forbes 2014] are shown in Table2.

*Table 2 summarized the cell lines used in this study. Average ploidy and mutations were retrieved from COSMIC cell line project*

| Cell Line | Sexes | Ethicity | Distributor | Catalogue 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 dataset for each cell line

For each cell line, the FASTQ files for Whole Genome Sequencing, ChIP-seq on H3K9me, H3K9_14Ac, H3K4me3, H3K4me1, H3K36me3, H3K27me3, H3K27Ac, Polymerase-II and input DNA, Whole Transcriptomes Sequencing and Transcriptional Starts Sites Sequencing (TSS-seq) were retrieve from the previous publication [Ayako Suzuki 2014].Statistics for the dataset as reported in this paper is shown in the below Table 3,4 . Annotations for coding regions were obtained from KERO database for UCSC hg38 human genome reference (http://kero.hgc.jp/).

## SNPs/SNVs from Whole genome sequence data

The FASTQ files for whole genome sequencing in each cell line were re-mapped to UCSC hg38 human genome reference [Matthew L. Speir 2016] by using bwa [Durbin 2009] (version 7.15) by aln algorithm with default setting. PCR-duplicates were then removed by samtools [Li H. 2009] (version 1.18). SNPs/SNVs were called by GATK [McKenna A 2010 ] (version 3.3) with default parameters. The SNPs/SNVs called by GATK with more than 5 supporting tags and variant frequency greater than 5% were selected. The variant frequency were calculated by samtools (v1.18) mpileup command with default setting. (see Table 3 for details)

## Regulatory Regions defined by ChIP-seq

ChIP-seq data for 7 histone modification (H3K9me, H3K9_14Ac, H3K4me3, H3K4me1, H3K36me3, H3K27me3 and H3K27Ac) and polymerase-II were processed. The FASTQ files were re-mapped to UCSC hg38 human genome reference using bwa (version 7.15) and aln algorithm with default setting. PCR-duplicates were then removed by samtools (version 1.18). Each dataset peak were calculated by MACS2 [Zhang Y 2008] broad-peak calling with default parameters against input DNA as background control. Peaks that were within 150 kb of transcription start site according to TSS-seq data were treated as regulatory regions. If there was multiple transcriptional start sites, the closest transcriptional start site was selected for the peak. (see Table 3 and 4 for details) SNVs that were within the peaks were then defined as regulatory SNVs. The number of regulatory SNVs were count collectively, if any SNVs were associated with multiple peaks, those SNVs would be counted multiple times and treated as separated SNVs.

## Whole Transcriptome Sequencing

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

## Transcriptional Start Sites Sequencing

For transcriptional start sites, resulted from 26 cell lines and 1 small airway epithelium cell samples were compared and merged. The clusters used were generated from the merged dataset. Promoter region for each gene was defined as region of 500bps upstream to 500bps downstream of the transcriptional start sits clusters. The resulted promoter positions were treated as regulatory regions in ChIP-seq dataset.

## Background Germline Variants Filtering

SNPs/SNVs called by GATK in whole genome sequence that were within the regulatory regions were treated as candidates for regulatory SNVs. These were filtered by NCBI's dbSNP (v142 note that background germline SNPs were not available)

## Synthetic long reads library preparation by 10x GemCode

From 23 cell lines, high molecular weight DNA were extracted and quantify by Qiagen MagAttract HMW kit according to manufacture recommendation (10x Genomics, Qiagen #67653).

For each cell lines, $1x10^6$ cells were suspended in 200 µl of PBS buffer, 20 µl of Proteinase K. Mixture, 4 µl of RNAase A and 150 µl of buffer AL.The samples were then incubate at 25◦C for 30 minutes. 15 µl of Qiagen MagAtrract susupension G were 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, sample were put on the magnetic rack for 1 minute and the clear supernatant were discarded. The beads were removed from the magnetic rack, suspended into 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 by 700 µl of Buffer PE twice. Beads with Buffer PE were put on the magnetic rack for 1 minute. The supernatant were removed on the magnetic rack, 700 µl of Nuclease-free water were added and incubated for 60 seconds, 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 were added to the bead pellets. The samples were mixed and incubated at 1400 rpm at (15–25 °C) for 3 minutes. The samples were put on the magnetic rack and held for 1 minute. The supernatant was transferred and stored at 4 ◦C for DNA Quantification by Qubit dsDNA HS Assay kit (Thermo Fisher Scientific) at target the 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 was performed by GemCode library preparation Kit (10x Genomics) according to the manufacturer's instructions. In brief, quantified High Molecular Weight DNA were further diluted by nuclease-free water to concentration of 1 ng/µl, and 1.2 µl were used. Sample Mix were prepared by adding the 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 GemCode Reagents Kits. The Sample Mix, Gel beads and partitioning Oil were applied onto GemCode Chip. The GemCode Chip was loaded in to the GemCode instrument.

Gel Beads In Emulsions (GEMs) were retrieved from the instrument according to manufacturer's recommendation and transferred to a 96-well plate for a designated thermal cycling amplification. For the post cycling recovery, 1 µl of Additive 1 and 125 µl of Recovery Agent was added and mixed to each GEMs according to manufacturer's instructions. The aqueous solutions were transferred and recovery Agent and Partitioning Oil removed. Mixture of Recovery Agent and Partitioning Oil at the bottom was first remove by 135 µl of pipetting. The leftover were 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 by Covaris system. The fragmentation was performed with target peak of 250 bp for the whole exome and regulome sequencing and 800 bp for the 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

GemCode library preparation Kits (10x Genomics). Products from End repair and A-tailing were ligated by thermal cycling with Adaptor Mix and DNA Ligase. Post ligation cleanups were performed by 0.8X SPRI solution on the GemCode magnetic rack. Sample indexing PCR by P5 primer were conducted. The post PCR cleanups were performed by 1.0XSPRI cleanup on the GemCode magnetic rack.

The obtained products were sequenced by Illumina Hiseq2500 for whole genome sequence. For whole exome and regulome samples, target enrichment were performed using Agilent SureSelectXT protocol with SureSelect V5 plus regulome baits according to the manufacturer's instructions (Agilent, 10x Genomics).See Figure 5 for summarized work flow.

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

## Physical long-read sequencing by MinION

For MinION sequencing H1975, LC2/ad, RERF-LC-KJ and II-18 cell lines were used.

High Molecular Weight DNAs were extracted in the same manner as descried above. Library preparations were performed according to the manufacturer's instructions (Oxford Nanopore Technologies). In brief, extracted high molecular weight DNA were subjected to End repair and dA-tailing by NEBNext End repair/dA-tailing module (E7546S, NEB). Purifications were performed using Agencourt AMPure XP beads (Beckman Coulter). Ligation and Tethering were proceeded with NEBNext Blunt/TA Ligase Master Mix (M0367S, NEB) and Ligation Sequencing Kit 2D (SQK-LSK208, Oxford Nanopore Technologies). The obtained libraries were purified by MyOne C1 beads (65001, Thermo Fisher Scientific). Sequencing was done in 48 hours run mode by MinION Mk 1B with the SpotION Flow Cell (FLO-MIN106, R9.4 version, Oxford Nanopore Technologies). 2D 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 UCSC hg38 human genome reference with LAST aln with parameters optimized for long read sequencing and bwa (version 0.7.15) with mem algorithm and 2dONT optional parameter. The results from LAST aln were converted to sam format by LAST. Conversion to bam format and sorting were done by samtools (version1.18).See also Figure 6 for work flow.

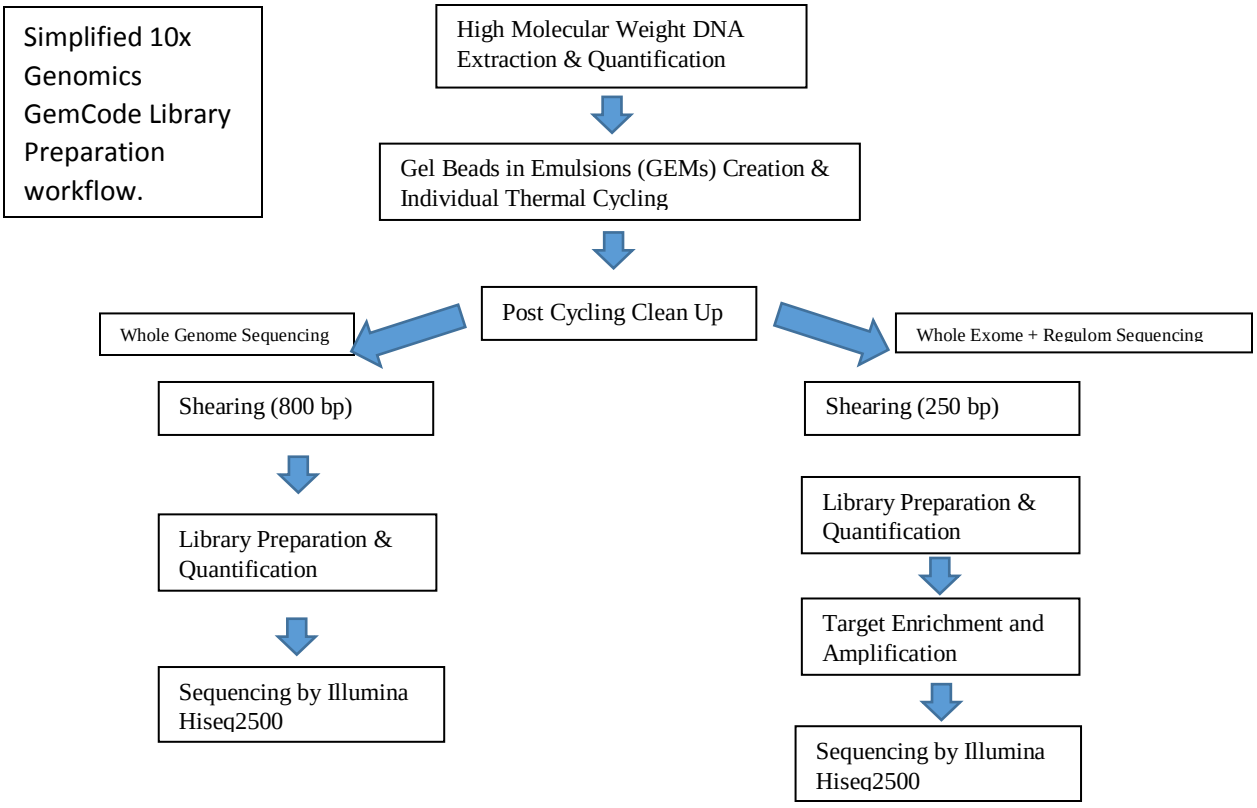


*Figure 5 Simplified work flow for 10x GemCode Library preparation System (10x Genomics).*

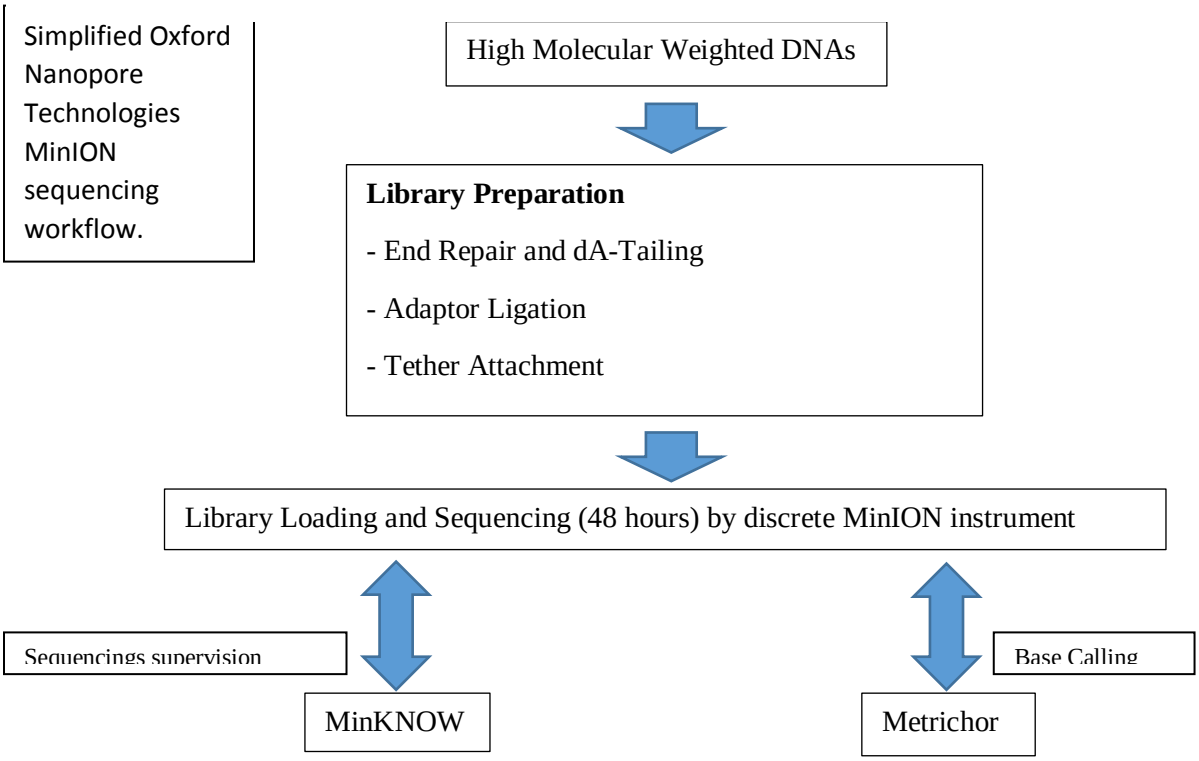


*Figure 6 Simplified work flow for MinION physical long read sequencing (Oxford Nanopore Technologies).*

| Cell line | Whole Genome Sequencing | | | Whole Transcriptome Sequencing | | Chip-seq Input Control | |
|---|---|---|---|---|---|---|---|
| | Mapped Read | % Mapped Read | Depths | 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 3 shows basic sequencing characteristic for whole genome sequencing, RNA-seq and Chip-seq background control*

| Cell lines | 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 4 shows sequencing statistic for individual Chip-seq Antibodies for each cell line.*

| WES+R | Sequencing Statistics | | | | | Phasing Statistics | | |
|---|---|---|---|---|---|---|---|---|
| **Cell Line** | Number of Reads | Mapped Read% | PCR Duplication | Bait Coverage | Depths | 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% |

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

# Results and Discussion

## Mutations Detected in Lung Adenocarcinoma cell lines

For all of the 23 cell lines, the whole genome sequencing data were re-analyzed. On average, I detected 1,375,802 SNPs/SNVs per cell lines. Using the KERO database, published by our laboratory, I identified an average of 1,375,802 SNPs/SNVs in coding region per cell line with an average of 19,086 SNPs/SNVs in exon regions. Regulatory regions were defined by ChIP-seq and TSS-seq. The germline SNPs in regulatory regions were further filtered out by NCBI's dbSNP. The final number of potential regulatory SNVs is 46,149 SNVs on average per cell line (see Table 6 for details). To specify potential function of regulatory SNVs, I considered every SNV with at least one overlapped peak.

Statistic of detected SNPs/SNVs are shown in Table 6, I constantly detected higher amount of SNVs in my current work compared with result for pervious publication using same dataset. I considered this due to changes in SNPs/SNVs calling procedure, the update of reference genome and larger number of ChIP-seq antibodies analyzed.

| 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 |

*Table 6 summarized SNPs/SNVs detected by GATK*

## Multi-omics Analysis reveals imbalance in allele expression

I considered that one of the main indicator of biological relevance of the promoter and enhancer variants, is by examining whether the mutations could activated or repressed their regulating gene's transcripts. One of the methods to observe these effects is to look for changes in expression of variant in the transcripts. Activating mutations in regulatory regions should increase the number of corresponding transcript, while repressive mutations should lower them. I expected such regulatory somatic mutations should have occurred in a heterozygous manner, evidenced by well-known heterozygous driver mutation such as EGFR and KRAS missense substitutions. I considered comparison of the ratio of the regulated transcripts regarding their heterozygous allele expressions to be the most straightforward way to evaluate the potential of regulatory SNVs; under the assumption that heterozygous functional regulatory SNVs, either activating or repressive, should result in detectable uneven allele expression of their corresponding heterozygous transcripts on the same allele. The heterozygosity of the transcripts can be identified by presence of genomic heterozygous SNVs/SNPs within the transcripts. The ratio of transcript variant expression should be detectable, as the ratio of variant frequency of heterozygous SNPs/SNVs in RNA-seq. Similar approach had been taken in several papers aiming at discovering imprinted genes (Baran, et al. 2015).

However, cancer genomes, especially those that have been transformed into cell lines, usually undergo heavy alteration regarding their ploidy. Most (see Table 2) of the cell lines used in this study are known to have aneuploidy genomes. I was concerned that this fact might make the interpretation of the detected allele expression ratio difficult to be used as the indicator of bias in expression levels between the alleles. I intended to circumvent this problem by considering variant tag frequencies in the whole genome sequence dataset. Genomic variant frequency may represent the degree of aneuploidy at that region of the genome and might prove useful in normalizing the expression ratio in ChIP-seq and RNA-seq.

To calculate the frequency in the genomic variants,.I calculated variant tags for both the reference nucleotide and the alternative nucleotide(s) of SNPs/SNVs in whole genome sequencing datasets (see Material and Methods for details). Variant frequency of regulatory SNVs and transcripts SNPs/SNVs in whole genome sequencing could be calculated in 2,874 RefSeq annotated genes on average per cell line, within those genes 10,721 regulatory SNVs from every regulatory regions were detected and total of 4,123 transcript SNPs/SNVs detected (see Table 8 for details). RNA-seq's allele expressions was also calculated using similar methods. For regulatory SNVs, variant frequency were calculated in each ChIP-seq antibodies, bias in at least one marker was considered positive result and the SNVs were counted in a collective manner, if any SNVs were to shown to have more than one activities bias, each bias would be treated separately. I expected the results to represent genome ploidy, bias in transcript expression and bias in any regulatory activity. Figure 7 on the next page visualizes the idea.

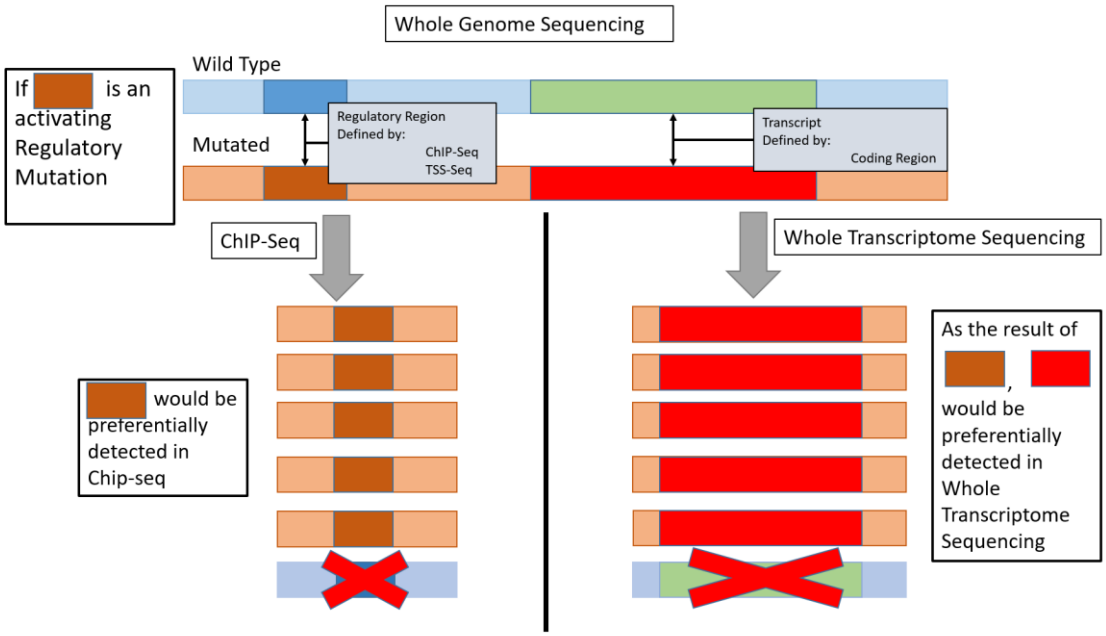


*Figure 7 Strategy in detecting Allelic imbalance expression. Genes with both bias in transcript allele and regulatory SNVs are first selected, then if 1). The difference in variant frequency of the SNPs/SNVs is significant (P<0.05, fisher's exacts test) and 2). The difference is larger than 2 fold changes for both regulatory and transcript regions. I classify the regulatory SNVs potentially functional SNVs. Genes without coverage for all omics were not considered..*

For the genes where the frequencies of the genomic variants were calculated, I intended to evaluate the bias in the allelic expressions and bias in the allele regulatory activities. For the potential regulatory SNVs, variant frequencies detected from the ChIP-seq data were calculated and, normalized against their respective variant frequency in the whole genome sequencing. For the transcript SNPs/SNVs, the variant frequencies of SNPs/SNVs in RNA-seq were calculated and normalized against their variant frequency of the SNPs/SNVs in the whole genome sequencing. For both of the detected biases in the regulatory activities and those in transcript allele expressions, genes for which the variant frequency could not be calculated, in any of the three omics dataset were excluded from this study. As a result, 1,600 RefSeq genes on average per cell line remained, containing 2,360 Regulatory SNVs and 1,946 transcript SNPs/SNVs.

In prior to the normalization, bias in the allelic expression and that in the regulatory activity were evaluated as the observed differences in the based variant tag frequencies in the respective two omics datasets; ChIP-seq dataset for the regulatory SNVs and the RNA-seq dataset for the transcripts, respectively. Statistical significance was evaluated in the 2x2 contingency table between two omics studies by fisher's exact test with the cut off P value of 0.05. Those selected cases (P<0.05) were further screened by minimum coverage of 5 in the number of the variant tags. Also the ratio of the variant tags against the alternative tags in RNA-seq and Chip-seq must change by at least two fold from the whole genome sequencing to be considered valid. By doing so, I expected that the effects from the possible aneuploidy would be canceled.

The genes for which at least one allele-biased expressions and one allele-biased regulatory activity were detected were further considered for the potential candidates having a functionally relevant regulatory mutation. For SNPs/SNVs, which were located both in the transcript and regulatory regions, and those that overlapped with multiple RefSeq (Entrez) gene groups were counted multiple times. For one SNP/SNV could hold more than one function. As a result, the genes of the "allele imbalance expression" were selected.

Table 7 shows a summary of the allele imbalance expression genes. On average, I detected 270 RefSeq genes per cell line, consisting of 524 potentially functional SNVs and 590 transcripts SNPs/SNVs. The cell line harboring the smallest number of such genes was PC-9 at 112 genes with, 176 potentially regulatory SNVs and 208 transcript SNPs/SNVs. The cell line with highest number of imbalance expression genes was H1648 with an un-proportionally large number of genes at 1,341 genes with 1,910 potential regulatory SNVs and 2,647 transcript SNPs/SNVs. Closer inspection revealed an exceptionally high number of coding SNPs/SNVs (24,819 vs average of 19,086, Table 6) was detected and annotated together with also a large number of potentially regulatory SNVs from ChIP-seq markers (55,458 vs average of 46,149, Table 6). Individually, manual inspections of some of the genes detected in this cell line showed no relevant difference compared to others cell lines. In this particular cell line, some unknown event might have taken place, making several *–omic* datasets distinct from other cell lines.

## Genes with detected allelic imbalance expression

Table 8 shows the top 12 autosomal genes with the allele imbalance expression in the most recurrent manner among the cell lines. At the bottom, top 3 genes located in X-chromosome exclusively recurred in female cell lines, thus may be influenced by X chromosome inactivation, the most recurring gene, NBPF1 encodes neuroblastoma break point protein 1. Functions of this protein are associated with development of neural system. Some connections to cancer development had been made but mainly to tumors of neural system. It would be intriguing if similar mechanisms could also take place in lung adenocarcinoma as well. The second most frequently effected gene is TYW1, which encode the protein Wybutosine, a hyper modified guanosine in tRNA processing pathway, little is known in its relation to cancer development in any cancer type.

Despite difficulties in making the biological interpretations of the detected genes, I believe that the merit of these cases lies further in depth understanding of the regulatory mutations and their effects rather than direct functional relevance of encoded proteins. To that end, I intended to further my analysis, focusing on the regulatory SNVs themselves.

| Cell Line | RefSeq Gene with whole genome variant frequency (ploidy known) | | | RefSeq Genes with all three omics' variant frequency calculated | | | RefSeq Genes with Allelic Imbalane Expression | | |
|---|---|---|---|---|---|---|---|---|---|
| | # RefSeq Genes | # Regulatory SNVs | # Coding SNPs/SNVs | # RefSeq Genes | # Regulatory SNVs | # Coding SNPs/SNVs | # RefSeq Genes | # Regulatory SNVs | # Coding SNPs/SNVs |
| **A427** | 3,644 | 15,956 | 4,986 | 1,360 | 1,941 | 1,509 | 221 | 457 | 437 |
| **A549** | 2,289 | 5,737 | 3,005 | 1,102 | 1,336 | 1,199 | 181 | 242 | 412 |
| **ABC-1** | 1,459 | 2,739 | 2,067 | 1,095 | 1,546 | 1,250 | 141 | 299 | 304 |
| **H322** | 1,608 | 3,865 | 2,226 | 1,098 | 1,650 | 1,294 | 141 | 357 | 342 |
| **H1299** | 2,676 | 9,386 | 3,634 | 837 | 888 | 956 | 126 | 143 | 329 |
| **H1648** | 6,101 | 21,455 | 9,776 | 3,222 | 4,182 | 4,384 | 1,341 | 1,910 | 2,647 |
| **H1650** | 1,896 | 9,149 | 2,904 | 943 | 1,390 | 1,186 | 132 | 245 | 275 |
| **H1703** | 2,621 | 10,056 | 3,529 | 1,188 | 1,592 | 1,356 | 130 | 209 | 263 |
| **H1819** | 3,002 | 14,793 | 4,693 | 1,500 | 2,630 | 2,091 | 195 | 822 | 677 |
| **H1975** | 4,184 | 18,914 | 5,646 | 2,114 | 2,996 | 2,535 | 354 | 697 | 774 |
| **H2126** | 2,361 | 7,099 | 3,544 | 1,102 | 1,421 | 1,328 | 171 | 266 | 356 |
| **H2228** | 4,346 | 19,104 | 5,966 | 2,261 | 3,003 | 2,702 | 327 | 558 | 782 |
| **H2347** | 4,425 | 22,059 | 6,202 | 2,631 | 4,236 | 3,196 | 377 | 726 | 829 |
| **II-18** | 2,192 | 7,315 | 2,904 | 1,453 | 2,341 | 1,662 | 193 | 392 | 393 |
| **LC2ad** | 1,280 | 1,700 | 1,975 | 1,165 | 1,341 | 1,433 | 139 | 208 | 253 |
| **PC-9** | 1,252 | 1,718 | 1,678 | 993 | 1,137 | 1,047 | 112 | 176 | 208 |
| **PC-14** | 2,497 | 7,981 | 3,658 | 1,607 | 2,564 | 2,015 | 230 | 481 | 584 |
| **RERF-LC-Ad1** | 4,073 | 17,398 | 5,881 | 2,444 | 3,571 | 2,940 | 318 | 616 | 775 |
| **RERF-LC-Ad2** | 3,411 | 15,828 | 5,035 | 2,000 | 3,200 | 2,523 | 304 | 579 | 641 |
| **RERF-LC-KJ** | 2,786 | 7,596 | 4,150 | 1,908 | 3,149 | 2,438 | 309 | 704 | 620 |
| **RERF-LC-MS** | 2,043 | 7,603 | 2,933 | 767 | 878 | 1,020 | 175 | 220 | 331 |
| **VMRC-LCD** | 2,917 | 9,540 | 4,200 | 1,960 | 3,504 | 2,296 | 249 | 810 | 556 |
| **RERF-LC-OK** | 3,046 | 9,589 | 4,242 | 2,061 | 3,784 | 2,408 | 344 | 924 | 793 |
| **Average** | 2,874 | 10,721 | 4,123 | 1,600 | 2,360 | 1,946 | 270 | 524 | 590 |

*Table 7 shows detailed detected allele imbalance in each cell line, please note that for each cell line a single SNPs/SNVs could be counted multiple times due to associations with multiple markers or RefSeq transctripts.*

| **Autosome** | | | **# of detected Cell lines** | | | **SNPs/SNVs detected** | |
|---|---|---|---|---|---|---|---|
| **RefSeq** | Gene Symbol | All (23) | Female (5) | Male (12) | unknown (6) | # Coding SNPs/SNVs | # Regulatory SNVs |
| **NM_017940** | NBPF1 | 18 | 4 | 10 | 4 | 77 | 173 |
| **NM_018264** | TYW1 | 17 | 3 | 10 | 4 | 11 | 89 |
| **NM_145109** | MAP2K3 | 15 | 4 | 8 | 3 | 19 | 25 |
| **NM_170606** | KMT2C | 15 | 4 | 8 | 3 | 40 | 26 |
| **NM_001005751** | FAM21A | 14 | 4 | 7 | 3 | 13 | 119 |
| **NM_014675** | CROCC | 14 | 2 | 8 | 4 | 31 | 47 |
| **NM_001128223** | ZNF717 | 12 | 5 | 4 | 3 | 188 | 560 |
| **NM_003174** | SVIL | 11 | 4 | 6 | 1 | 24 | 41 |
| **NM_030653** | DDX11 | 10 | 1 | 6 | 3 | 16 | 20 |
| **NM_004399** | DDX11 | 10 | 1 | 6 | 3 | 16 | 20 |
| **NM_152438** | DDX11 | 10 | 1 | 6 | 3 | 17 | 20 |
| **NM_002568** | PABPC1 | 10 | 2 | 6 | 2 | 18 | 29 |
| **Chromosome X** | | | # of detected Cell lines | | | SNPs/SNVs detected | |
| **RefSeq** | Gene Symbol | All (23) | Female (5) | Male (12) | unknown (6) | # Coding SNPs/SNVs | # Regulatory SNVs |
| **NM_002139** | RBMX | 4 | 3 | 0 | 1 | 10 | 3 |
| **NM_001448** | GPC4 | 2 | 1 | 0 | 1 | 2 | 3 |
| **NM_031407** | HUWE1 | 2 | 1 | 0 | 1 | 2 | 5 |

*Table 8 shows top12 autosomal imbalance expression genes and top 3 X-chromosome imbalance expression genes. No male cell line allele imbalance genes were detected on X-Chromosome.*

## Allele Expression imbalance in X-inactivated and imprinted allele.

The most well-known mechanism that would produce allelic imbalance expression is X-inactivation or Lyon hypothesis. For humans, this phenomenon happens exclusively in normal X-chromosome of female cells. This process, by epigenetic control or randomly inactivating one of the alleles of the X-chromosomes, compensates an “extra” copy of X-chromosome presented in the female genomes. I attempted to utilize this phenomenon as the positive control.

Among 23 cell line in this study, 12 were known to came from male, 5 from female and 6 were unknown. From female cell line, H2347 made a good example for the genes located on X chromosome. 56 out of 377 allele imbalance expression genes (Figure 8) were identified on X-chromosome. In another cell line of unpecificed gender, RERF-LC-OK I also detected 18 out of 326 allele imbalance expression genes located on X chromosome. Figure 9 and Figure 10 illustrates the example of allele specific expression that were assiciated with X- inactivation.

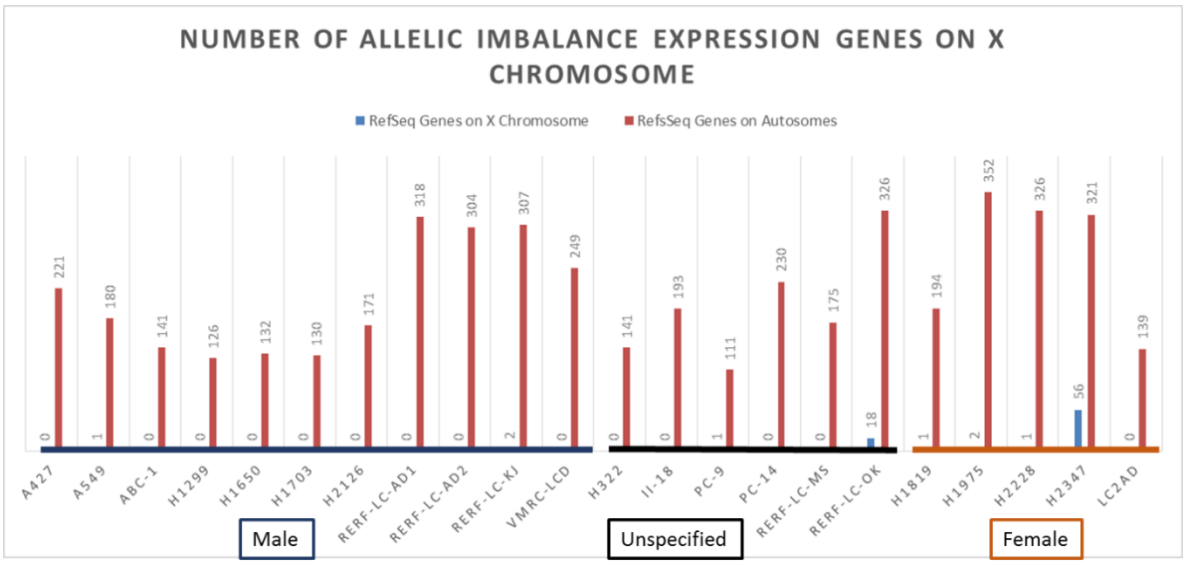


*Figure 8 shows number of allele imbalance expression genes located on X chromosome(blue) compared to autosome(red). Only H2347 and RERF-LC-OK were found to have a good number of genes.*

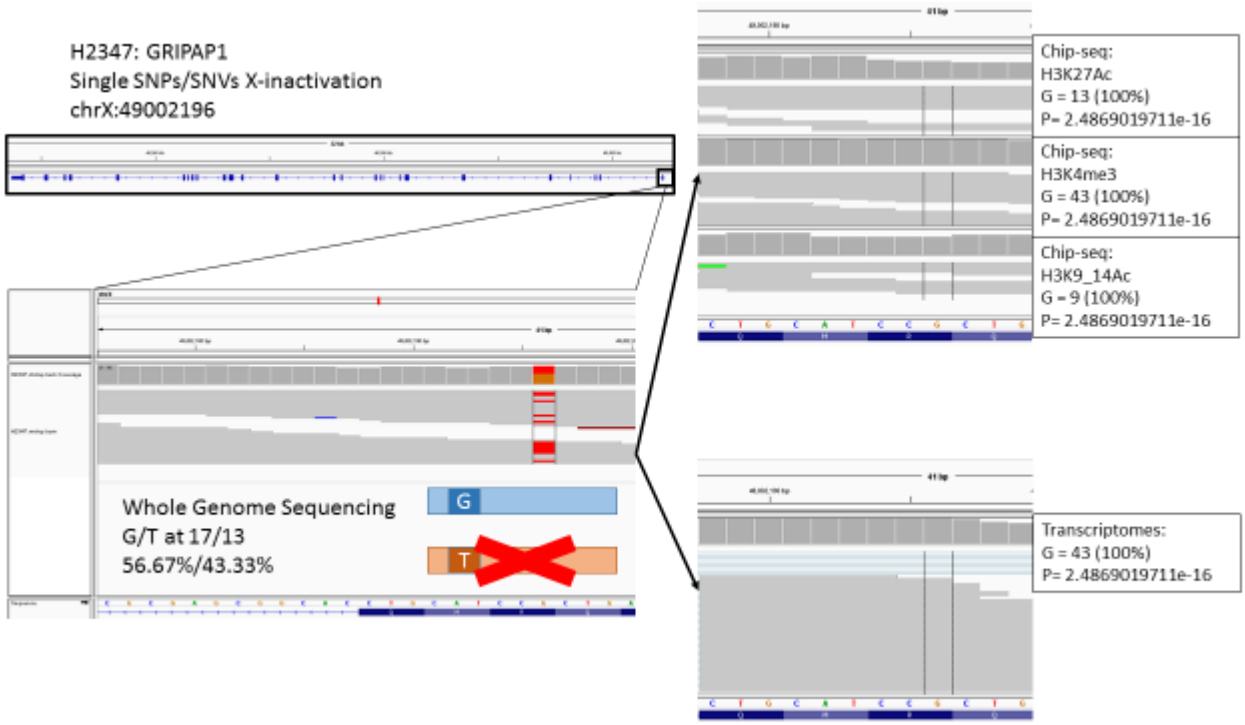


*Figure 9 demonstrate Single SNP/SNV X-inactivation for GRIPAP1 at chrX: 49002196 for H2347 female cell line. Only one allele ("G") is regulatory and transcriptional activated.*

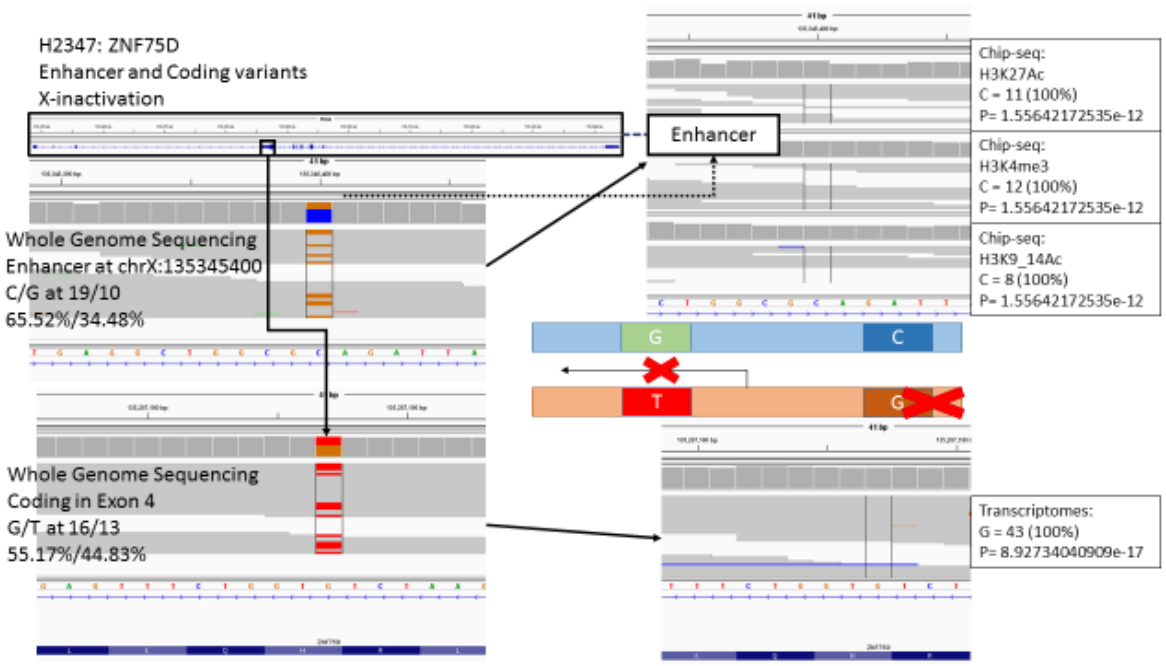


*Figure 10 demonstrate Enhancer and Coding SNPs/SNVs pair X-inactivation for ZNF75D at chrX:135345400 (Enhancer) and chrX:135287187 (Coding) in H2347 female cell line. Only one allele is active.*

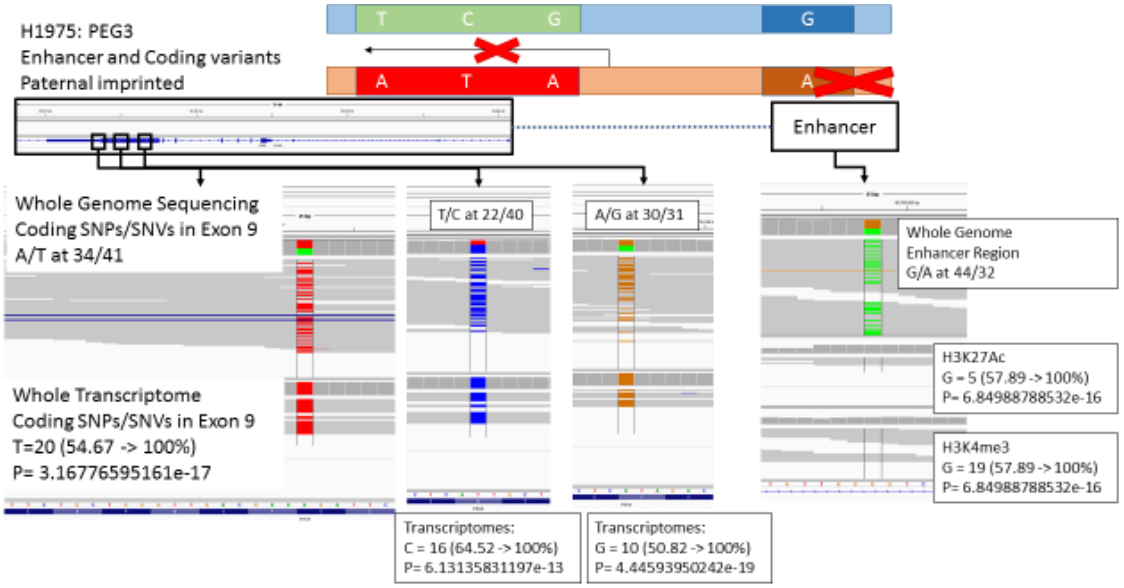


*Figure 11 demonstrate detection of Paternal Imprinting of PEG3 in H1975 cell line, its enhancer SNVs chr19:56765618 is marked by H3K4me3 and H3K27Ac and its coding variant is are detected in exon number 9 at chr19: 56814572, chr19: 56815602 and chr19: 56816135. Even though located on autosome, this genes also shows imprinting effects.*

Autosome could also exhibit allele specific expression by imprinting in the wild type contexts. Similar to the X-activation, only one of these alleles would be active. One of the well-known imprinted genes is the paternally imprinted PEG3, encoded a zinc protein family transcriptional factor, its imbalance expression was detected in H1975 cell line and shown in Figure 11.

These cases, if detected, would provide solid evidence that functional regulatory elements could be identified in the proposed manner. By further excluding the natively imprinted genes, I assumed that genes with allele imbalance expression could also be explained by regulatory elements with somatic regulatory SNVs as one of the candidates. I inspected each of the cases and found some intriguing SNV, which might directly influenced oncogenic development in their respective cell lines. Two of these were the KMT2C and MAP2K3 genes, which will be exemplified in the next section.

## Potential Functional Relevance of Regulatory SNVs in imbalanced genes

The first example is KMT2C. KMT2C encodes lysine methyltransferase 2C which main function is the methylation of Lysine 4 of Histone 3. Disruption of this gene had been documented as oncogenic (Dou 2015). I detected imbalance of this gene in 15 out of the 23 cell lines, figure 12 shows the visualization in H1975 cell line. It is well known that histone modifications by this protein are important factors of epigenetic control in many genes in various cellular circumstances. Knock-down of genes in this family were reported to have resulted in changes in methylation level, attenuated growth in cell lines (Changcun Guo 2013) and deletion of KMT2C/D were reported to had a more favorable outcomes in pancreatic ductal adenocarcinoma (Joshua BN Dawkins 2016).

Another example was MAP2K3. RAS pathway is one of the most recurrently impaired pathways in lung adenocarcinomas. The function of this pathway is related to a broad biological events and activation of this pathway is mediated by many factors. One of the factors is active form of MAP2K3, a member of kinase family, which is activated during stressful or mitogenic events (NCBI 2017). From analysis of 23 cell lines, I detected 15 cell line with allele imbalance expression of this gene. Figure 13 shows imbalance expression of this gene in H322 cell line. This result implied that aberrant accumulation of the MAP2K3 invoked by impaired transcriptional regulation might trigger the over activation of RAS pathway. However, in MAP2K3, discrepancies were encountered in both potential regulatory SNVs and transcripts SNPs/SNVs. For the regulatory SNPs/SNVs, a portion of assumed inactivated variant was left over in ChIP-seq dataset, while most the transcripts' un-transcribed allele were not expressed. In one of the transcript SNPs/SNVs, the allele imbalance expression was not observed, even though others SNPs/SNVs, which assumed to be on the same transcript were all exhibit allele imbalance expression. These finding raised my concern that conclusion regarding SNPs/SNVs allele configuration could not be correctly made without implementation of SNPs/SNVs phasing and these mistake could lead to misinterpretation of the results.

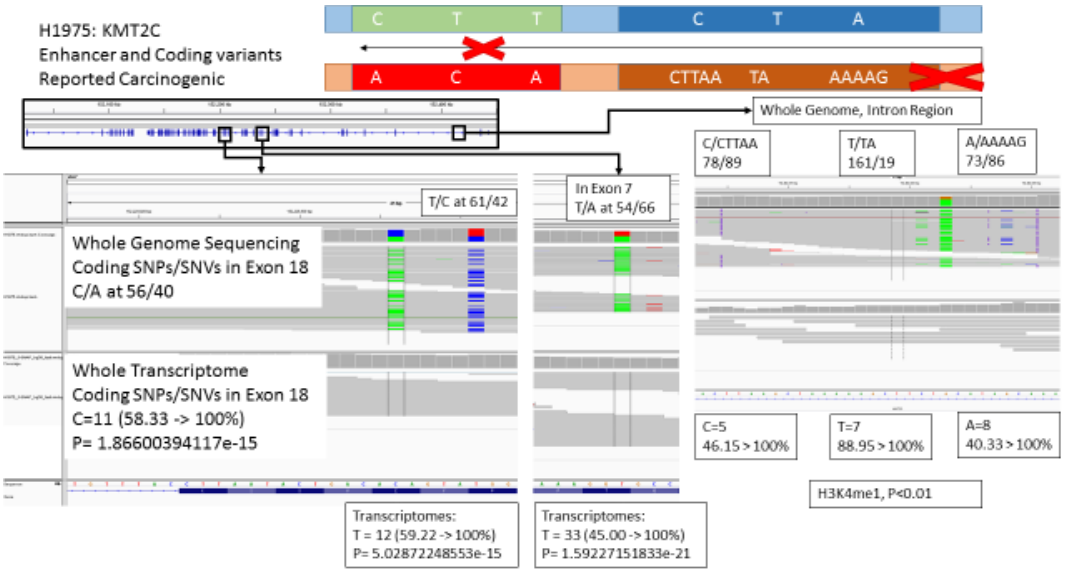


*Figure 12 shows allele imbalance expression in KMT2C in H1975. The regulatory SNVs were 3 insertion, defined as enhancer by H3K4me1 marker at chr7: 152402604, chr7: 152402626 and chr7: 152402630. Multiple coding SNPs/SNVs were detected, shown here are exon18's chr7:152229936 and chr7: 152229941 and chr7: 152273771 on exon7.*

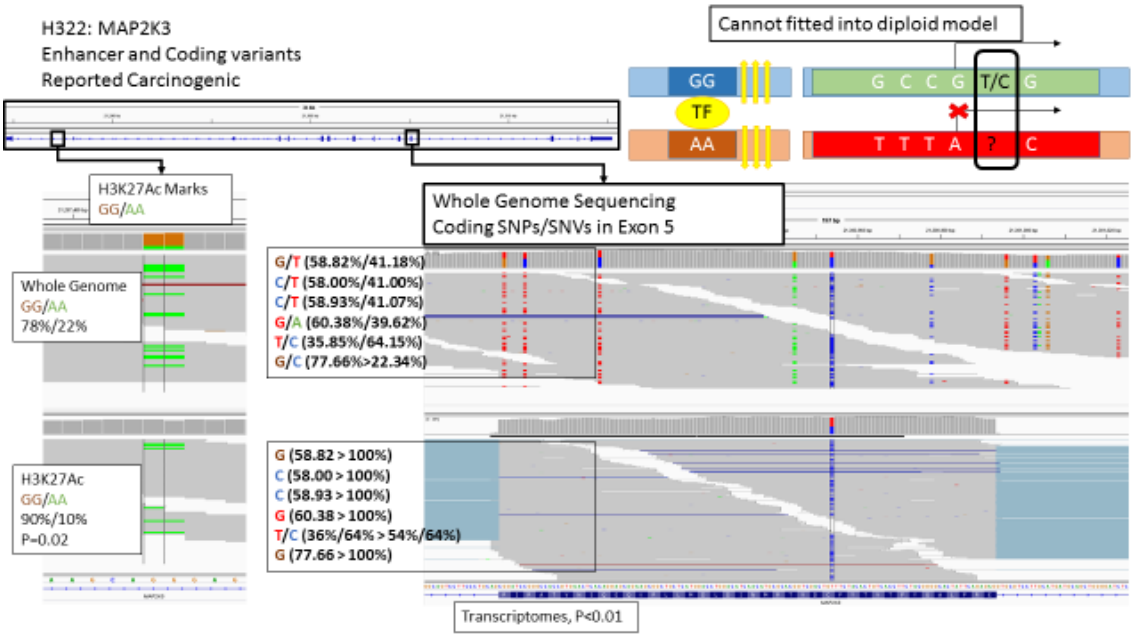


*Figure 13 shows allele imbalance expression in MAP2K3 in H332 cell line, the relation between the regulatory SNVs and Coding SNPs/SNVs cannot be fitted into diploid model. The regulatory SNVs were marked by H3K27Ac at chr17: 21287464- 21287465 and the transcripts were identified by 16 coding SNPs/SNVs shown here are SNPs/SNVs in exon5 at chr17:21300875, chr17: 21300880, chr17: 21300898, chr17: 21300945, chr17: 21300954 and chr17: 21300978.*

Indeed, in all of the above examples, every potential candidate's were not selected from their somatic mutation in coding regions and these candidates were not known before. I considered that others genes might also are important to the cancer development even though their involvements were not yet documented. To further explore the function of a single regulatory SNV and provide answer to discrepancies in MAP2K3 case, I planned to directly associate the regulatory SNVs to the transcripts by mean of phasing.

## Phase Block Construction and Phasing of SNPs/SNVs

I attempted to obtain the phasing information as the essential information for associating regulatory SNVs to their regulating transcripts. I considered that, with the phasing information, identification of functional relevant regulatory SNVs would be firstly enabled. Using the indexing and barcoding technologies, the 10x GemCode system is able to recover these phasing information. However, its default pipeline only provides the phasing information assuming the diploid genome. Indeed, the analysis for the cancer genome is not supported in its original publication; therefore, I intend to circumvent this problem by developing new pipeline, which enables the assembly of the phasing information from the molecular indexes (tagged with "MI"). These molecular indexes are the collection of the variants, which were automatically constructed from the reads with the same barcodes or the group of barcodes, which were determined by the LongRanger software. MIs are supposed to retain information representing the original high molecular weight DNAs, whose sequences should be originating from the adjacent genomic regions. Therefore, by further careful inspection of those MIs would give clues for improving the default LongRanger results, particularly for the presumed polypoid genomes.

The methods I propose are based on the exhaustive approach, in merging multiple molecular indexes. I examined whether if the molecular indexes were mutually consistent, and discrepant in regards to their member SNPs/SNVs. If the overlapped SNPs/SNVs position of any pair of molecular indexes all held the same variants, those pair of molecular indexes would be considered compatible. If any overlapped SNPs/SNVs held different variants, the pair would be considered incompatible. The molecular indexes without any overlapping SNPs/SNVs would not be considered

I created the working database for each cell line by listing all of the SNPs/SNVs detected in the whole genome sequencing. I retrieved the molecular index(es), which were associated with each of the SNPs/SNVs from the output of LongRanger (tag "MI"). Only SNPs/SNVs from 10x GemCode read with mapping quality over 20 and base quality, if applicable, over 20 were included.

For each SNPs/SNVs, I assembled the associated molecular indexes into a particular haplotype. The molecular index having the largest number of SNPs/SNVs was chosen as a starting point. The starting molecular index was merged with all of the available compatible molecular indexes. This finished product was termed at the "haplotype". Other overlapping but incompatible molecular indexes were further used as starting point for next "haplotype". This step of phasing was repeated until every available molecular index was processed and every possible "haplotype" was considered in an exhaustive manner. (see Figure 14 for graphic explanation)

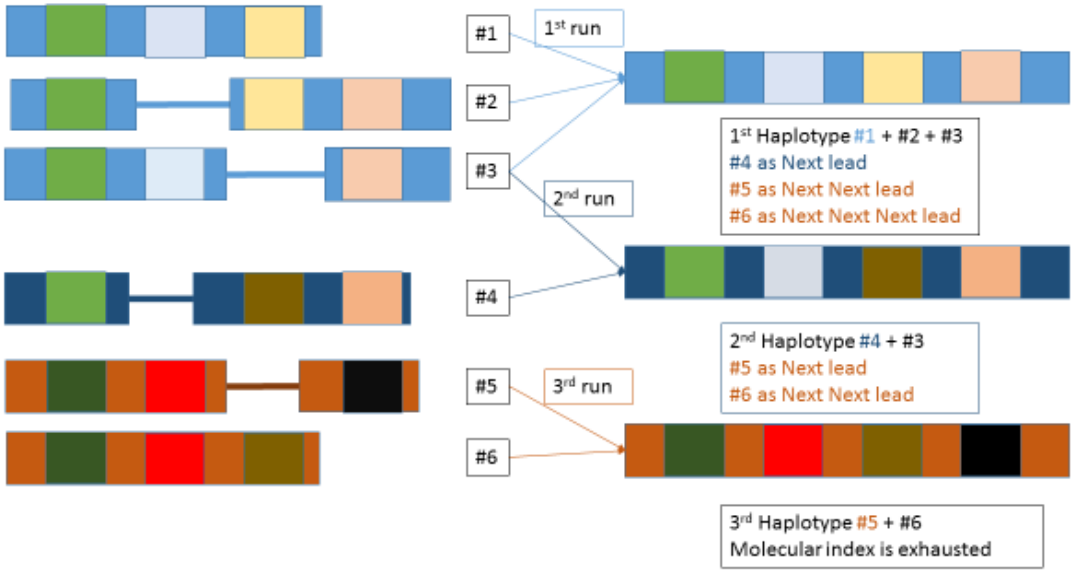


*Figure 14 Phasing scheme: greedy merging and extending the molecular indexes if the SNPs/SNVs are compatible, considered new haplotype otherwise. Each "haplotype" was exhaustively checked against every possible overlapped Molecular indexes.*

By following the method as outlined above, I obtained a collection of the "haplotypes", which I termed as the "phased block". These "phase blocks" were the genomic partitions spanning a certain region, anchored by the SNPs/SNVs . Each individual haplotype holds a list of nucleotides that are linked together by the "compatible" molecular indexes. One "phase block" could hold any number of "haplotype" as long as the physical connections were supported by molecular indexes. (see Figure 15 for the graphical scheme)

Due to the randomness in the original barcoding, it was possible that two adjacent SNPs/SNVs would not share the same molecular index and do not result in the same haplotype in the above process. Indeed, this was true for many cases, especially in the lowly covered regions. Practically, these cases could be identified as the case where the "phase block" contained multiple, short and isolated "haplotypes". These short and isolated associations would not be particularly useful in the phasing of regulatory SNVs the later analysis. Therefore, I attempted to solve this problem by merging the short "haplotypes" together with each other to create a more complete "phase block". (Figure 15, Right)

This second merging was done internally for a given one "Phase Block" by greedy approach. I first determined which genomic positions are missing from each haplotypes by comparing the region covered by "Haplotypes" and "Phase Block". Then for the haplotype with fewest missing position, greedily look for the most similar haplotypes that could fill those gap. Similarity for any pair of haplotypes with determined by number of compatible SNPs/SNVs subtracted by number of un-compatible SNPs/SNVs. Haplotypes that were considered in this step

must all be the member of the same "Phase Block" This process would then done exhaustively until every missing positions were filled. (Figure 16 and Figure15 Right to Center)

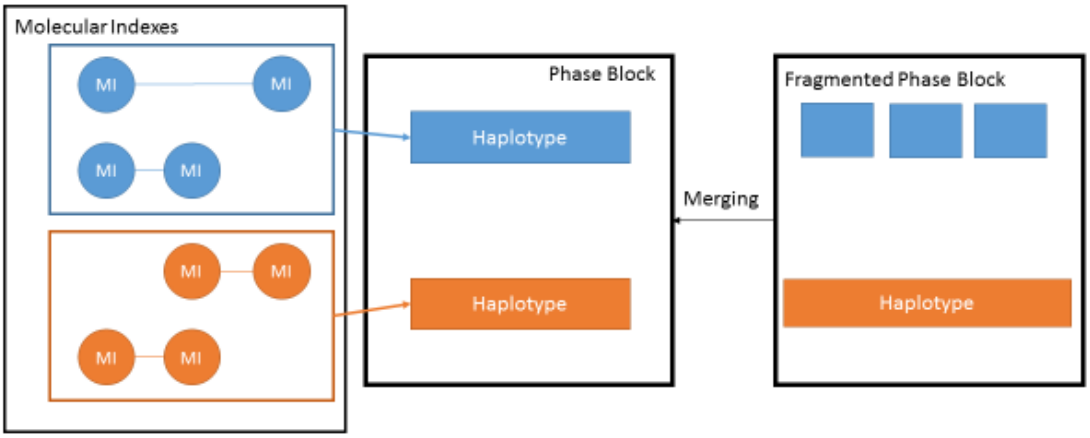


*Figure 15 Relations of component in phasing. Left).10x GemCode molecular index. Center). Phase Block consisting of 2 Haplotypes. Right). Phasing Block with multiple, short and isolated Haplotype.*

The relations between the SNPs/SNVs were determined by examining co-localization of the SNPs/SNVs in the finally merged haplotypes (Figure 15 Center). The SNPs/SNVs on the same haplotype were treated as cis-related. Relations of SNPs/SNVs from different phased block were unknown.

*Figure 16 shows filling of the haplotypes with missing position into complete ones.*

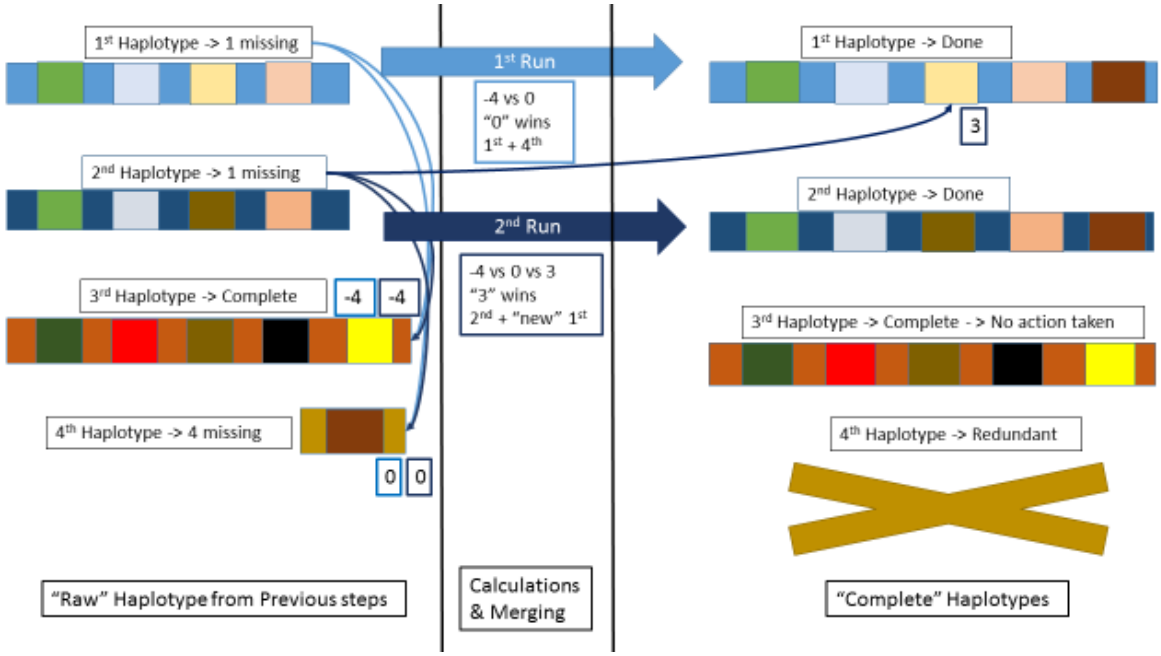

Table 9 details the statistics of the phase block. Majority of the phase block were diploid (Table 9, Median, also see Figure 24, Figure 25). The aneuploidy nature of cancer genome cell lines were shown, on average, approximately 40% (2,744 from 6,975) of the phase block contained more than two haplotypes. The maximum number of haplotypes was 154. Given such a high number, I was concerned that many might be errors, which were derived from either the flaw of developed phasing strategy or error in input data. They include: sequencing errors, SNPs call errors or barcoding errors in 10x GemCode. Extremely high number of haplotype were few, 2.8% for >10 (Table14 and Table 15) and 0.4% for >20 (Data not shown). Therefore, I used all the constructed phase block in further analysis, assuming that the contributions from possible errors were minor.

Another important characteristic of the phase blocks is the length of region of the genome covered by them. "Length" of the phase block was defined as the distance between the most 5' SNPs/SNVs and the most 3' SNPs/SNVs member of the block. In many paper, the phasing quality is measured as the N50, defined the genomic distance where a half of the SNPs/SNVs are phased by shorter than this given length. This variable is common used and provided a good scale to compare the performances of different benchmark tests. However, I focus on a relation between pairs (or more) of SNPs/SNVs, not the region in the genome itself, I felt that the length of the phase block is a better indicator to access the performance. Table 9 shows details on phase blocks length, and number of SNPs/SNVs embedded within them, on average the phase blocks were 50 kilobases long and contained 13 SNPs/SNVs (with maximum of 989 kilobases in length with 496 SNPs/SNVs.)

To check reliability of utilizing length, I examine the possibility that "long" phase blocks might be deceiving and contain few, but far away SNPs/SNVs which might came from error decision in my phasing strategy. To look for those dubious cases, the relation between length and number of SNPs/SNVs were plotted and inspected (Figure 17). I found good correlation between them ($R^2$=0.76 by least square method, see Table 9) and concluded that long phase blocks are mostly informative.

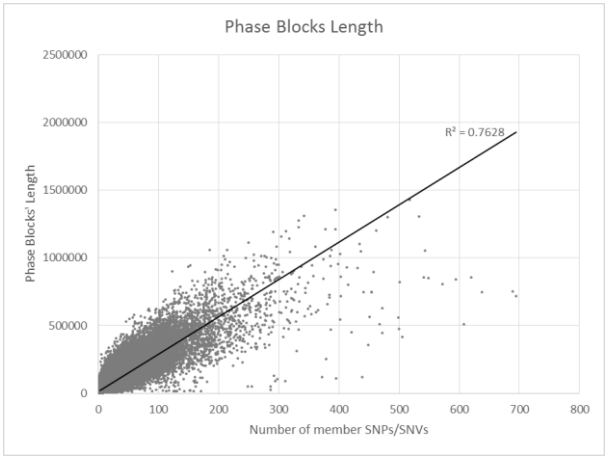


*Figure 17 shows a graph of number of SNPs/SNVs (x axis) and phase blocks' length (y axis). $R^2$ = 0.76 calculated by least square method.*

| Cell lines | Phase Blocks' Length | | | # of SNPs/SNVs in Phase Block | | | # of Haplotypes in 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 |

*Table 9 shows phase block characteristic, including phase blocks' haploid genomic length, number of member SNPs/SNVs and number of detected haplotypes. Phase block genomic Length is calculated by the most 3' SNPs/SNVs position subtract by the most 5' SNPs/SNVs position, calculation was done on only 1 ploid.*

.

| Phase Block Length | H1975 | H2347 | RERF-LC-Ad1 | RERF-LC-KJ | H1648 | RERF-LC-OK | H2228 | VMRC-LCD | RERF-LC-Ad2 | H1819 | H1299 | H1703 | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0-25k** | 5,668 | 4,959 | 4,933 | 4,751 | 4,757 | 4,369 | 3,984 | 4,334 | 4,073 | 4,241 | 4,051 | 3,670 | 3,836 |
| **25k-50k** | 1,426 | 1,304 | 1,379 | 1,392 | 1,213 | 1,323 | 1,023 | 1,257 | 1,087 | 1,149 | 1,236 | 1,135 | 1,053 |
| **50k-75k** | 774 | 855 | 785 | 780 | 717 | 689 | 757 | 712 | 686 | 633 | 622 | 659 | 614 |
| **75k-100k** | 478 | 573 | 542 | 503 | 448 | 455 | 467 | 431 | 445 | 394 | 383 | 384 | 382 |
| **100k-125k** | 292 | 364 | 371 | 340 | 300 | 282 | 323 | 297 | 297 | 316 | 273 | 279 | 263 |
| **125k-150k** | 204 | 281 | 275 | 227 | 213 | 218 | 265 | 189 | 228 | 193 | 169 | 212 | 189 |
| **150k-175k** | 146 | 207 | 213 | 148 | 160 | 149 | 211 | 165 | 170 | 149 | 136 | 140 | 139 |
| **175k-200k** | 110 | 137 | 149 | 114 | 148 | 112 | 147 | 105 | 115 | 90 | 99 | 110 | 103 |
| **200k-225k** | 63 | 119 | 121 | 78 | 103 | 88 | 116 | 80 | 99 | 87 | 79 | 83 | 81 |
| **225k-250k** | 46 | 100 | 97 | 65 | 82 | 57 | 96 | 55 | 68 | 66 | 55 | 65 | 62 |
| **250k-275k** | 41 | 84 | 69 | 60 | 52 | 49 | 81 | 51 | 81 | 55 | 33 | 69 | 51 |
| **275k-300k** | 37 | 58 | 60 | 39 | 45 | 31 | 69 | 33 | 53 | 35 | 33 | 47 | 39 |
| **300k-325k** | 20 | 51 | 47 | 20 | 43 | 23 | 53 | 35 | 39 | 21 | 27 | 32 | 30 |
| **325k-350k** | 22 | 46 | 35 | 23 | 34 | 27 | 54 | 24 | 37 | 22 | 13 | 36 | 26 |
| **>350k** | 55 | 167 | 153 | 80 | 182 | 81 | 250 | 91 | 141 | 86 | 53 | 115 | 114 |
| **Total** | 9,382 | 9,305 | 9,229 | 8,620 | 8,497 | 7,953 | 7,896 | 7,859 | 7,619 | 7,537 | 7,262 | 7,036 | 6,981 |

*Table 10 shows number of phase block distribution by length of cell line with above average block length*

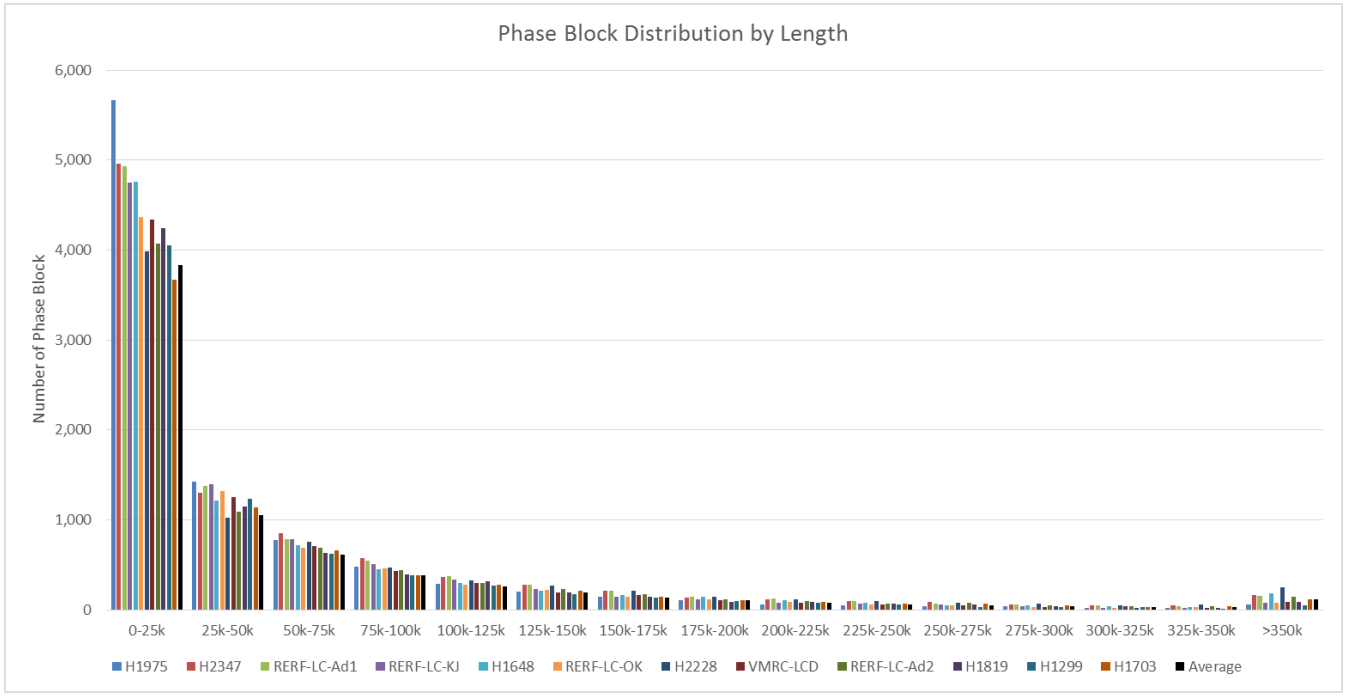


*Figure 18 shows graph of phase block distribution by length of cell line with above average block length*

| Phase Block Length | Average | A427 | LC2ad | II-18 | A549 | ABC-1 | PC-9 | RERF-LC-MS | H2126 | H322 | H1650 | PC-14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0-25k** | 3,836 | 3,549 | 3,268 | 4,145 | 4,009 | 3,392 | 3,287 | 2,956 | 2,680 | 2,968 | 3,159 | 1,021 |
| **25k-50k** | 1,053 | 913 | 876 | 991 | 980 | 999 | 1,155 | 802 | 768 | 825 | 781 | 195 |
| **50k-75k** | 614 | 633 | 651 | 530 | 522 | 565 | 570 | 562 | 522 | 441 | 390 | 68 |
| **75k-100k** | 382 | 410 | 419 | 299 | 279 | 338 | 377 | 330 | 312 | 255 | 218 | 56 |
| **100k-125k** | 263 | 326 | 279 | 199 | 171 | 259 | 241 | 235 | 231 | 180 | 167 | 24 |
| **125k-150k** | 189 | 196 | 231 | 140 | 123 | 173 | 183 | 171 | 175 | 163 | 110 | 10 |
| **150k-175k** | 139 | 168 | 163 | 108 | 71 | 152 | 123 | 119 | 132 | 98 | 73 | 4 |
| **175k-200k** | 103 | 106 | 161 | 53 | 66 | 117 | 90 | 104 | 98 | 77 | 46 | 9 |
| **200k-225k** | 81 | 123 | 109 | 39 | 49 | 76 | 77 | 76 | 91 | 59 | 36 | 2 |
| **225k-250k** | 62 | 74 | 99 | 39 | 30 | 58 | 70 | 54 | 65 | 55 | 31 | 1 |
| **250k-275k** | 51 | 83 | 65 | 25 | 23 | 50 | 39 | 44 | 59 | 32 | 21 | 2 |
| **275k-300k** | 39 | 59 | 54 | 17 | 19 | 36 | 28 | 39 | 41 | 40 | 17 | 2 |
| **300k-325k** | 30 | 54 | 52 | 11 | 13 | 28 | 28 | 30 | 25 | 35 | 9 | 0 |
| **325k-350k** | 26 | 39 | 47 | 9 | 9 | 32 | 16 | 14 | 22 | 22 | 8 | 0 |
| **>350k** | 114 | 222 | 259 | 32 | 34 | 119 | 75 | 95 | 163 | 117 | 45 | 1 |
| **Total** | 6,981 | 6,955 | 6,733 | 6,637 | 6,398 | 6,394 | 6,359 | 5,631 | 5,384 | 5,367 | 5,111 | 1,395 |

*Table 11 shows number of phase block distribution by length of cell line with below average block length*

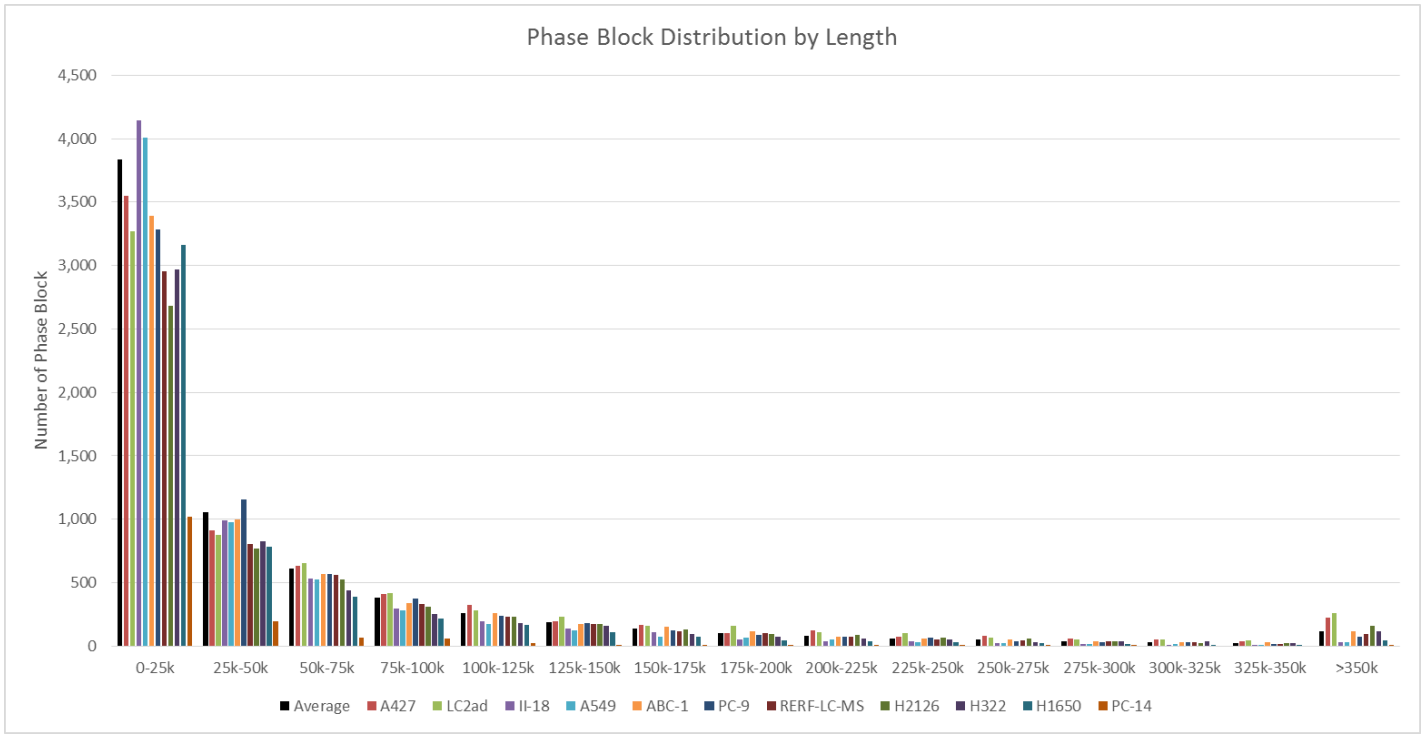


*Figure 19 shows graph of phase block distribution by length of cell line with below average block length*

| Number of SNPs/SNVs | H1975 | H2347 | RERF-LC-Ad1 | RERF-LC-KJ | H1648 | RERF-LC-OK | H2228 | VMRC-LCD | RERF-LC-Ad2 | H1819 | H1299 | H1703 | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0-10** | 6,977 | 6,586 | 6,469 | 6,152 | 6,202 | 5,790 | 5,480 | 5,707 | 5,472 | 5,436 | 5,368 | 5,127 | 5,090 |
| **11-20** | 1,030 | 1,037 | 1,053 | 971 | 876 | 884 | 907 | 833 | 810 | 800 | 787 | 685 | 734 |
| **21-30** | 443 | 509 | 494 | 503 | 397 | 413 | 430 | 424 | 393 | 382 | 381 | 356 | 350 |
| **31-40** | 293 | 291 | 326 | 307 | 261 | 263 | 248 | 236 | 227 | 233 | 237 | 212 | 213 |
| **41-50** | 191 | 188 | 224 | 176 | 172 | 191 | 192 | 162 | 166 | 185 | 117 | 149 | 146 |
| **51-60** | 126 | 155 | 142 | 116 | 116 | 82 | 123 | 122 | 117 | 110 | 94 | 127 | 102 |
| **61-70** | 73 | 102 | 120 | 99 | 97 | 74 | 109 | 92 | 105 | 78 | 75 | 69 | 74 |
| **71-80** | 43 | 95 | 76 | 67 | 68 | 46 | 72 | 59 | 66 | 71 | 49 | 54 | 56 |
| **81-90** | 45 | 67 | 57 | 51 | 55 | 45 | 57 | 48 | 54 | 40 | 37 | 50 | 42 |
| **91-100** | 36 | 60 | 63 | 29 | 42 | 41 | 51 | 40 | 36 | 40 | 29 | 41 | 35 |
| **>100** | 125 | 215 | 205 | 149 | 211 | 124 | 227 | 136 | 173 | 162 | 88 | 166 | 138 |
| **Total** | 9,382 | 9,305 | 9,229 | 8,620 | 8,497 | 7,953 | 7,896 | 7,859 | 7,619 | 7,537 | 7,262 | 7,036 | 6,981 |

*Table 12 shows number of phase block distribution by number of member SNPs/SNVs of cell line with above average member SNPs/SNVs*

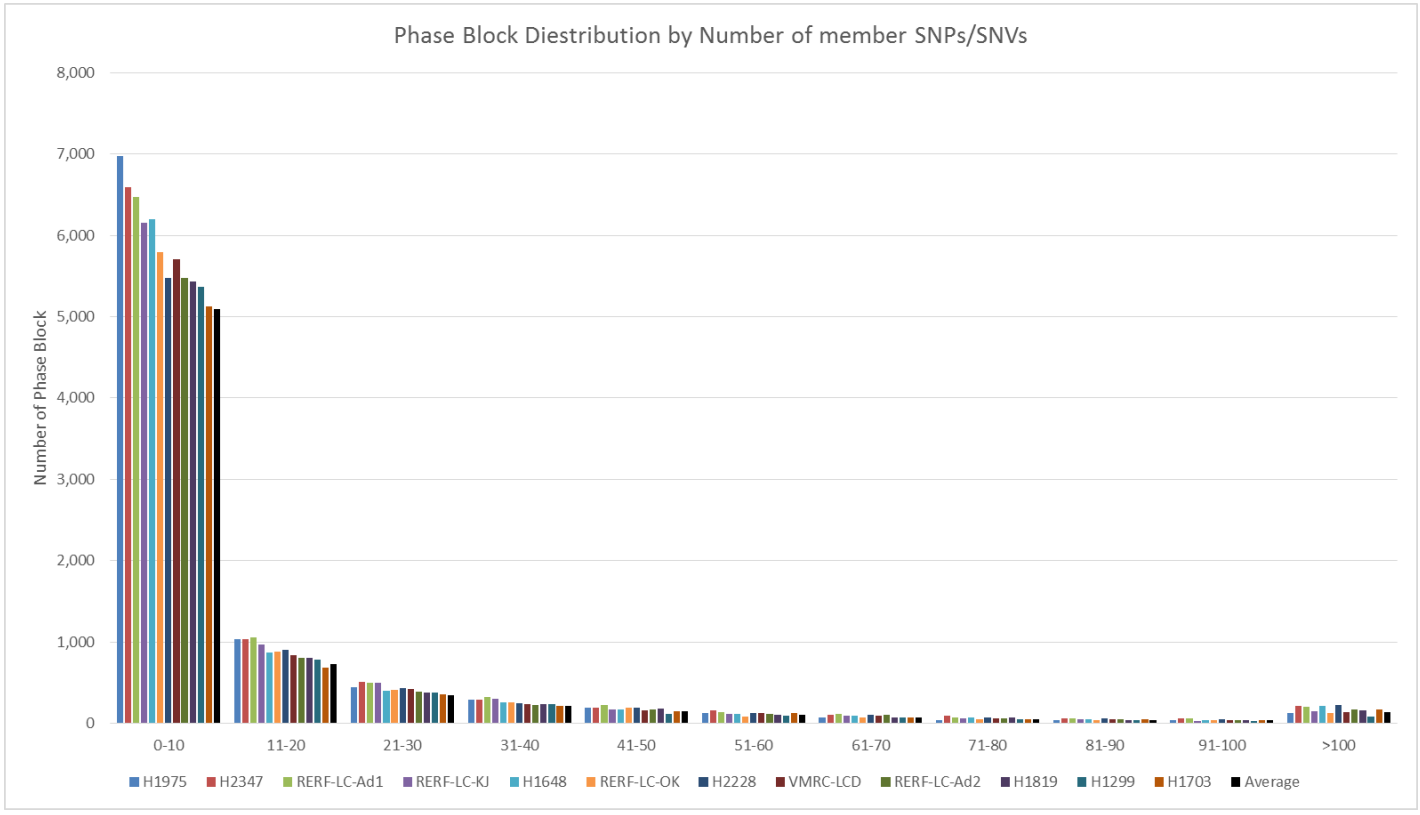


*Figure 20 shows graph of phase block distribution by number of member SNPs/SNVs of cell line with above average member SNPs/SNVs*

| Number of SNPs/SNVs | Average | A427 | LC2ad | II-18 | A549 | ABC-1 | PC-9 | RERF-LC-MS | H2126 | H322 | H1650 | PC-14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **0-10** | 5,090 | 4,863 | 4,708 | 5,058 | 5,001 | 4,760 | 4,667 | 4,179 | 3,838 | 4,003 | 3,959 | 1,262 |
| **11-20** | 734 | 736 | 713 | 660 | 654 | 601 | 658 | 576 | 584 | 484 | 481 | 68 |
| **21-30** | 350 | 379 | 356 | 287 | 248 | 288 | 301 | 270 | 281 | 261 | 221 | 32 |
| **31-40** | 213 | 240 | 227 | 189 | 134 | 179 | 201 | 182 | 170 | 138 | 106 | 9 |
| **41-50** | 146 | 153 | 147 | 145 | 110 | 129 | 129 | 112 | 113 | 97 | 99 | 11 |
| **51-60** | 102 | 130 | 129 | 74 | 64 | 127 | 96 | 79 | 81 | 77 | 54 | 2 |
| **61-70** | 74 | 72 | 91 | 45 | 37 | 66 | 62 | 68 | 68 | 58 | 38 | 2 |
| **71-80** | 56 | 75 | 59 | 48 | 40 | 54 | 62 | 43 | 55 | 40 | 34 | 5 |
| **81-90** | 42 | 60 | 59 | 26 | 24 | 36 | 38 | 25 | 34 | 50 | 18 | 1 |
| **91-100** | 35 | 41 | 32 | 24 | 17 | 37 | 29 | 25 | 28 | 37 | 27 | 0 |
| **>100** | 138 | 206 | 212 | 81 | 69 | 117 | 116 | 72 | 132 | 122 | 74 | 3 |
| **Total** | 6,981 | 6,955 | 6,733 | 6,637 | 6,398 | 6,394 | 6,359 | 5,631 | 5,384 | 5,367 | 5,111 | 1,395 |

*Table 13 shows number of phase block distribution by number of member SNPs/SNVs of cell line with below average member SNPs/SNVs*

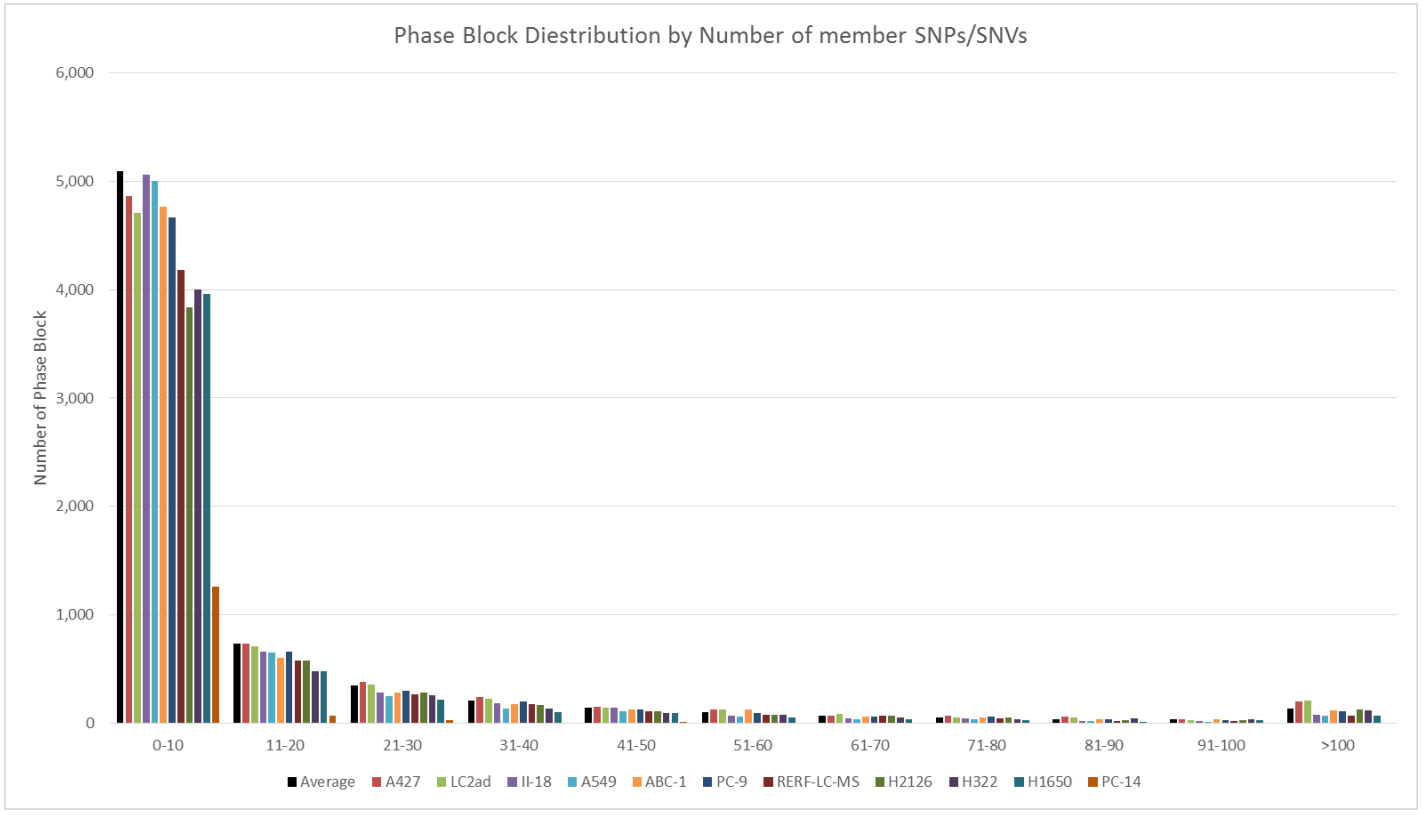


*Figure 21 shows graph of phase block distribution by number of member SNPs/SNVs of cell line with below average member SNPs/SNVs*

| Number of Haplotypes | H1975 | H2347 | RERF-LC-Ad1 | RERF-LC-KJ | H1648 | RERF-LC-OK | H2228 | VMRC-LCD | RERF-LC-Ad2 | H1819 | H1299 | H1703 | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **2** | 6,228 | 5,464 | 5,437 | 5,081 | 5,135 | 4,826 | 4,826 | 4,544 | 4,525 | 4,573 | 4,367 | 4,093 | 4,232 |
| **3** | 1,319 | 1,420 | 1,314 | 1,328 | 1,282 | 1,198 | 1,145 | 1,326 | 1,166 | 1,106 | 1,175 | 1,127 | 1,075 |
| **4** | 625 | 689 | 747 | 718 | 646 | 649 | 534 | 649 | 624 | 534 | 602 | 560 | 528 |
| **5** | 349 | 417 | 449 | 453 | 390 | 401 | 341 | 343 | 360 | 344 | 339 | 336 | 317 |
| **6** | 227 | 298 | 326 | 295 | 256 | 240 | 245 | 264 | 233 | 270 | 214 | 220 | 210 |
| **7** | 183 | 228 | 240 | 203 | 168 | 152 | 192 | 166 | 172 | 160 | 167 | 158 | 151 |
| **8** | 134 | 189 | 166 | 128 | 141 | 105 | 150 | 138 | 123 | 133 | 118 | 126 | 113 |
| **9** | 90 | 144 | 145 | 100 | 121 | 125 | 97 | 97 | 113 | 96 | 81 | 83 | 86 |
| **10** | 63 | 104 | 100 | 95 | 90 | 57 | 67 | 75 | 70 | 87 | 62 | 79 | 66 |
| **>10** | 162 | 345 | 301 | 216 | 263 | 196 | 292 | 246 | 223 | 228 | 134 | 252 | 198 |
| **Total** | 9,380 | 9,298 | 9,225 | 8,617 | 8,492 | 7,949 | 7,889 | 7,848 | 7,609 | 7,531 | 7,259 | 7,034 | 6,975 |

*Table 14 shows number of phase block distribution by number of haplotypes of cell line with above average member number of haplotypes*

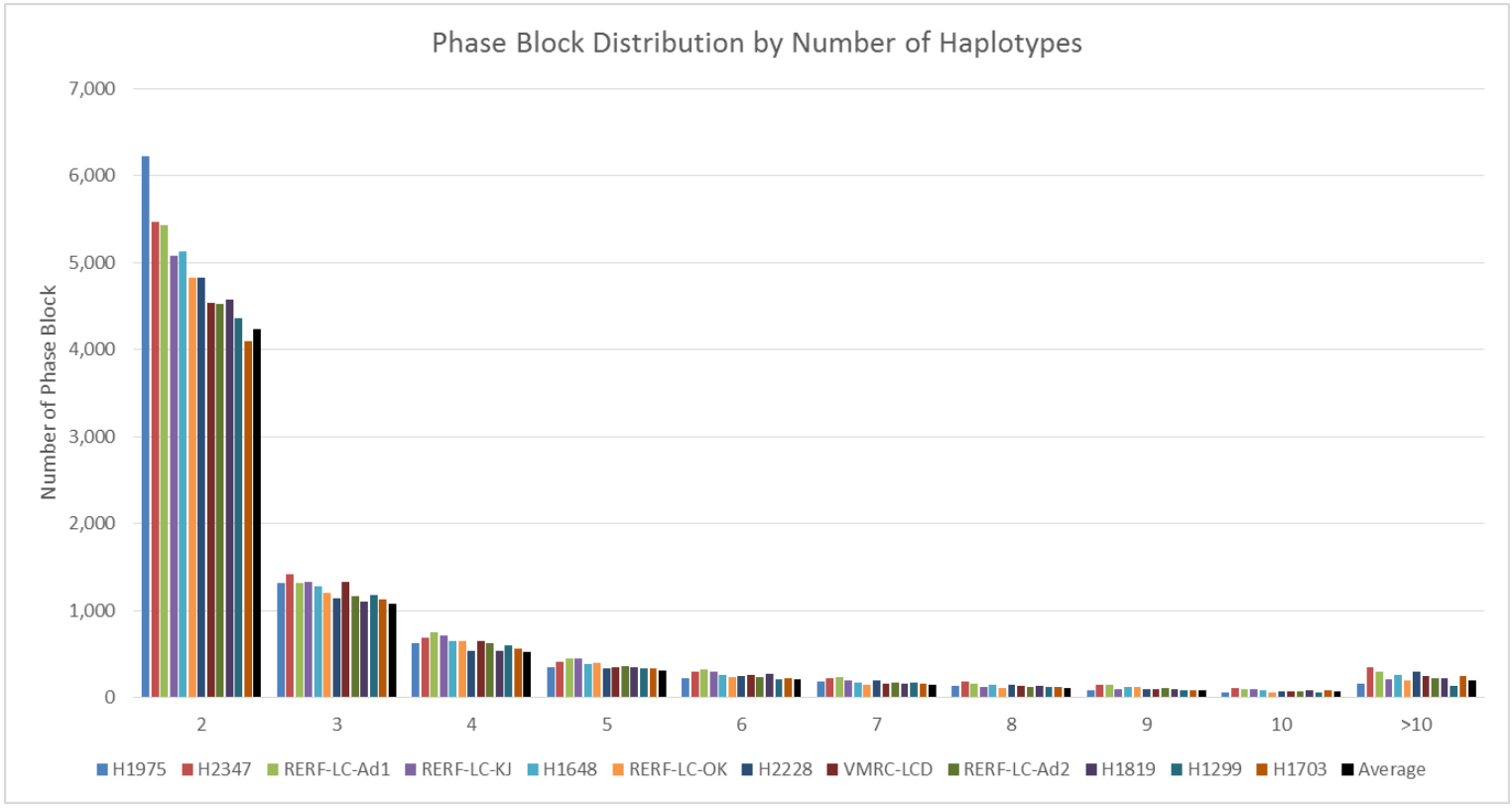


*Figure 22 shows graph of phase block distribution by number of haplotypes of cell line with above average number of haplotypes*

| Number of Haplotypes | Average | A427 | LC2ad | II-18 | A549 | ABC-1 | PC-9 | RERF-LC-MS | H2126 | H322 | H1650 | PC-14 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **2** | 4,232 | 4,074 | 4,063 | 4,078 | 4,318 | 3,727 | 3,939 | 3,377 | 3,184 | 3,290 | 3,300 | 879 |
| **3** | 1,075 | 1,024 | 934 | 1,082 | 915 | 1,038 | 1,085 | 973 | 843 | 777 | 865 | 290 |
| **4** | 528 | 530 | 475 | 500 | 429 | 505 | 459 | 404 | 404 | 393 | 349 | 110 |
| **5** | 317 | 359 | 323 | 275 | 233 | 308 | 276 | 269 | 261 | 228 | 179 | 50 |
| **6** | 210 | 229 | 232 | 170 | 122 | 180 | 182 | 187 | 151 | 154 | 106 | 24 |
| **7** | 151 | 157 | 176 | 155 | 91 | 132 | 111 | 117 | 132 | 118 | 79 | 18 |
| **8** | 113 | 120 | 113 | 106 | 79 | 120 | 76 | 76 | 95 | 82 | 73 | 7 |
| **9** | 86 | 90 | 85 | 58 | 58 | 83 | 70 | 58 | 70 | 71 | 43 | 5 |
| **10** | 66 | 85 | 71 | 63 | 43 | 77 | 41 | 43 | 52 | 56 | 31 | 4 |
| **>10** | 198 | 282 | 256 | 144 | 104 | 218 | 104 | 123 | 189 | 191 | 78 | 8 |
| **Total** | 6,975 | 6,950 | 6,728 | 6,631 | 6,392 | 6,388 | 6,343 | 5,627 | 5,381 | 5,360 | 5,103 | 1,395 |

*Table 15 shows number of phase block distribution by number of haplotypes of cell line with below average number of haplotypes*

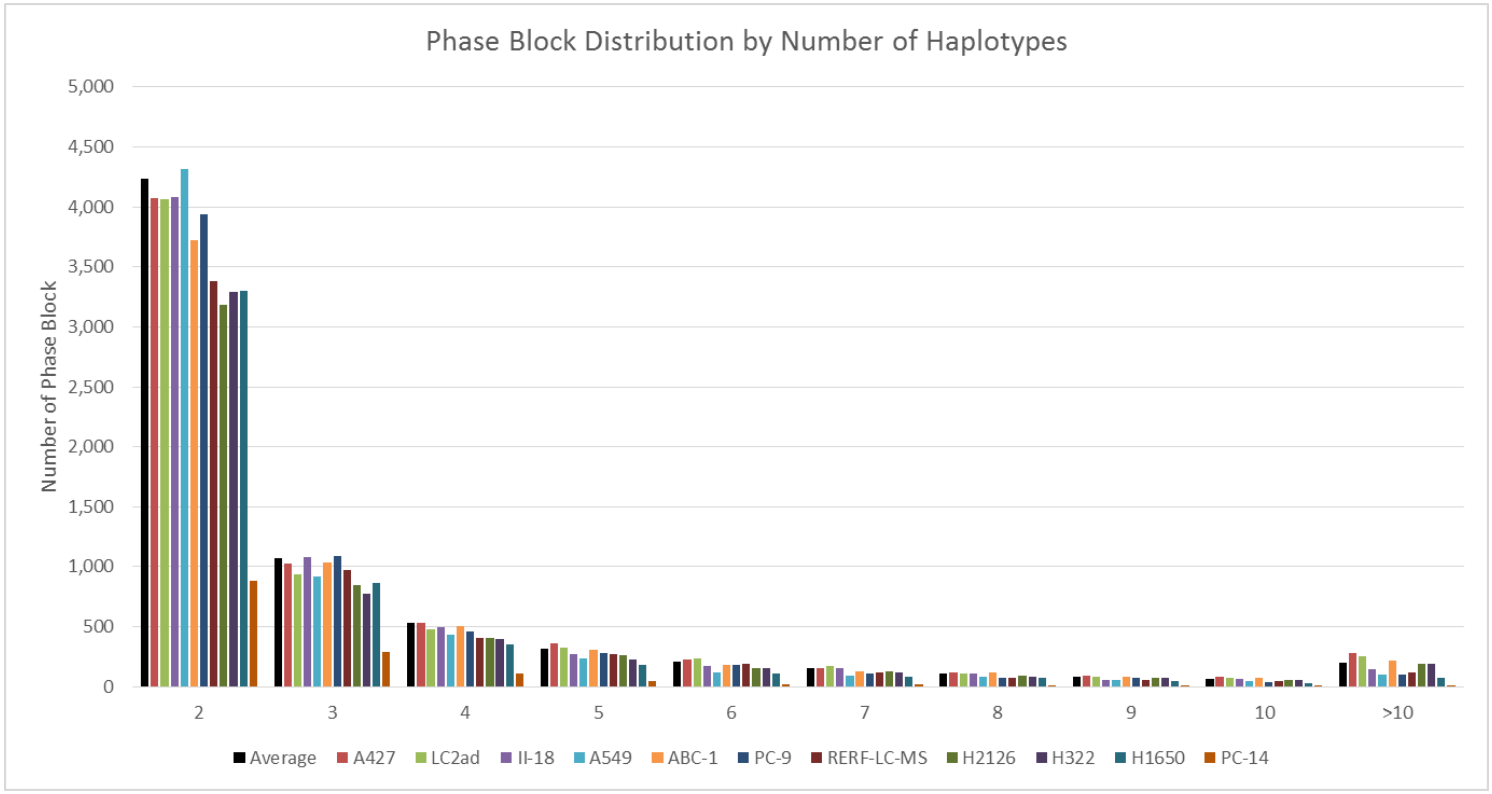


*Figure 23 shows number of phase block distribution by number of haplotypes of cell line with below average number of haplotypes*

## Phasing of Known Somatic Mutation

To validate my phasing, I selected a known case in driver mutation in the EGFR coding region in H1975 cell line as the example (Figure24). The substitution of G to T at codon number 2573 causing L858R mutation has been detected in various clinical non-squamous cell lung cancers, including lung adenocarcinoma. The tumors harboring this mutation are sensitive to anti-EGFR therapy, such as gefitinib. Despite rapid initial responce, many patients quickly develop drug resistance to the therapy. Further investigations have identified secondary mutation in the EGFR binding site of the drugs. A hotspot of the secondary mutation is substitution of T to C at codon number 2369 causing T790M mutation. I assume that, in H1975, this secondary mutation should present on the same allele. Others neighboring variant spanning 132kb genomic regions were also phased.

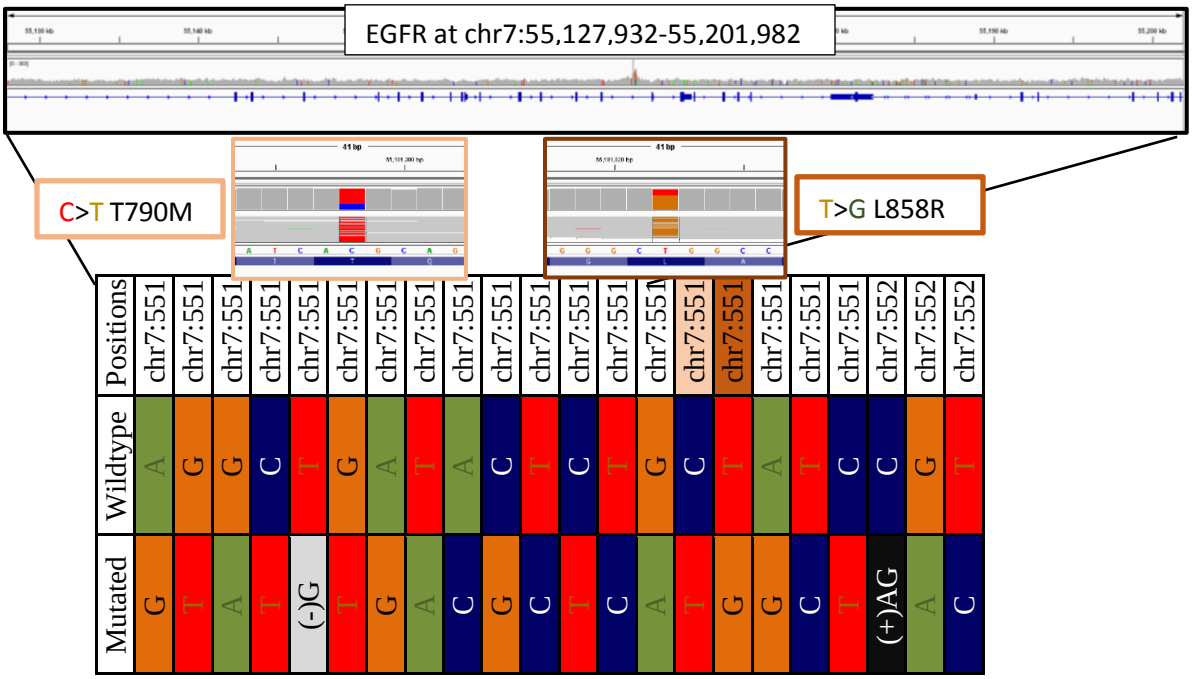


*Figure 24 shows Phase Blocks and SNVs phasing of known T790M (Light Brown) and L858R (Dark Brown) EGFR driver mutations in H1975 cell line.consisting of 22 SNPs/SNVs with 132kb length*

In addition to small somatic mutation, I also detected larger genomic aberrations such as copy number alteration. Copy number alterations in the ERBB2 genes is one example. It has been reported that amplification of this gene is frequently identified in lung adenocarcinoma and copy number alternations of this gene has been proposed as one of the driver mutations. In Figure25 extending from original diploid phasing, I identified multiple haplotypes in ERBB2 coding region in H1975 cell line along with other smaller somatic mutations.

## Phasing of Genes with detected Allele Imbalance Expression

All of the previous examples of allelic imbalance genes, mediated by X-inactivation, imprinting and potential regulatory SNVs were speculated without direct evidence of their relation between regulatory elements and the transcripts, In every connection diploid genome was assumed, which is not always the case in cancer cell lines. Indeed the allele imbalance profile of MAP2K3 in H322 cell line indicated that this assumption did not hold true and further evaluations of SNPs/SNVs phasing were needed.

To this end, I merged the phasing information in "Phase Blocks" constructed earlier with the list of SNPs/SNVs in genes with allele imbalance expression, each gene was considered "Phased" if at least one regulatory SNV and one coding SNPs/SNVs were presented in the same phase block.

Table 16 shows the number of successfully phased genes with RefSeq annotation and the SNPs/SNVs within them. On average 59 from 265 (22%) candidates genes were phased, and those genes covered 58 out of 516 (11%) regulatory SNVs candidates and 128 out of 582 (23%) their regulating candidate transcripts. The coverages of whole exome with regulome bait, which I used in the 10x GemCode phasing did not cover the genes of interested, especially in regulatory elements. Among those regulatory elements 83% (423 from 516) remained unphased. Nevertheless, a total of 1,146 allele imbalance expression genes, in which 1,071 regulatory SNVs and 2,269 transcripts SNPs/SNVs were phased, were detected from 2. Cell lines taken together. I considered that a fairly adequate fraction of genes were phased and deserved the further analysis as the first dataset for this purpose.

Figure 26, exemplifies an allele imbalance of gene CDKN1A detected in LC2ad. CDKN1A is an important regulator of the cell cycle regulation and have a well-reported role in carcinogenesis. In this case, the allele imbalance in the transcriptome showed "Leftover" allele "A" (19 tags, 16%), too large to be error but still much smaller than genomic portion (16% vs 33%). The ChIP-seq variant tags suggested the "total" imbalance without any leftover allele. This complicated the interpretation of the functional potential of the regulatory SNVs. In the case, further haplotype configuration was necessary to draw a sound conclusion. With the haplotype configurations acquired from my phasing analysis, it became clear that in this region there were at least triploid haplotype, in which the regulatory variant "A" was in cis-configuration with both coding variant "C" and "A" on separate haplotype. This illuminated the reason behind the "Leftover" in transcriptome but a "total" imbalance in ChIP-seq and supported the conclusion that variant "A" in this position had a transcriptional activating effects.

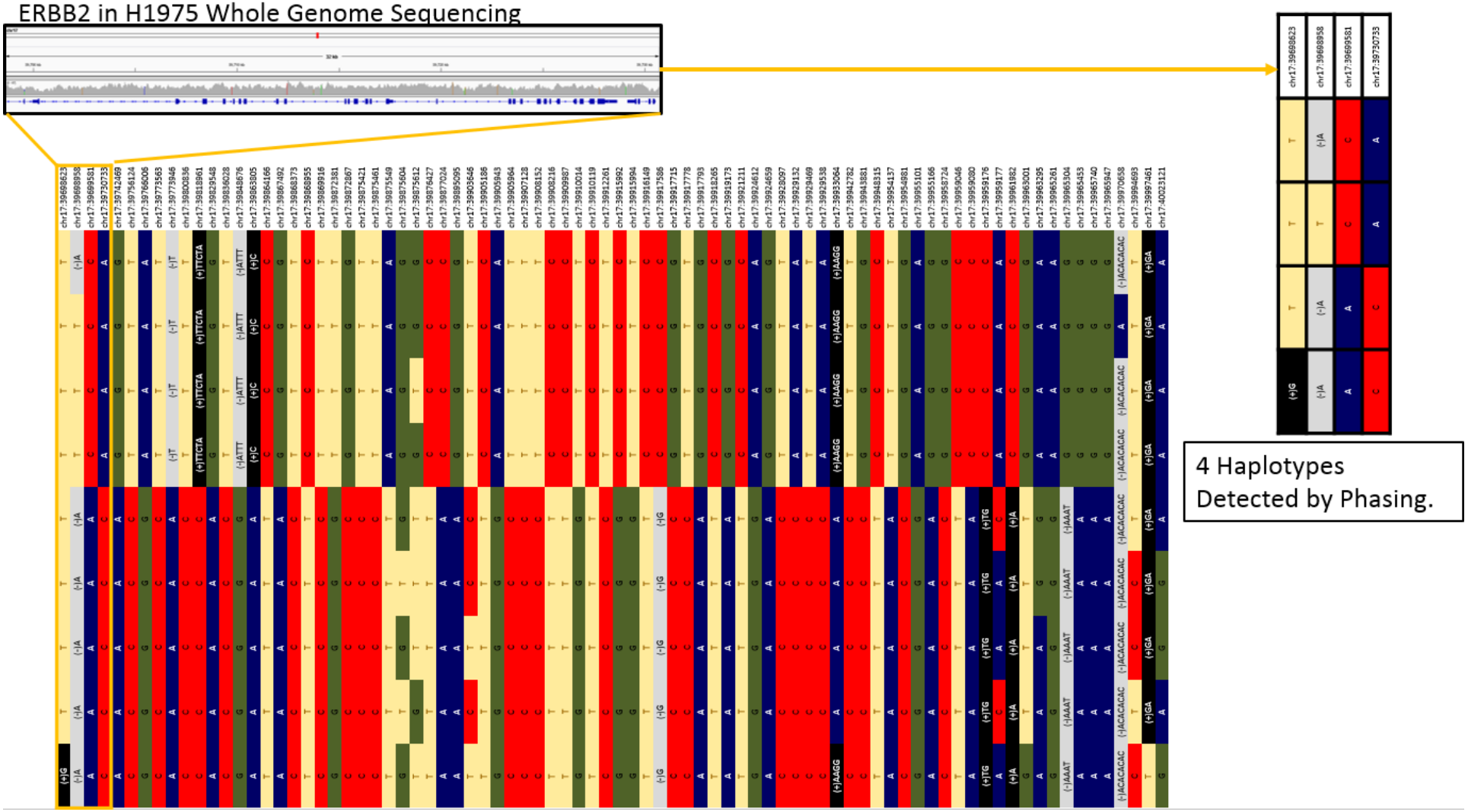


*Figure 25shows 4 haplotypes of ERBB2 gene detected by phasing SNPs/SNVs in coding region*

*n.*

| Cell Line | Pre-Phasing | | | Successfully Phased Genes | | | Unphased Genes | | |
|---|---|---|---|---|---|---|---|---|---|
| | RefSeq Gene | Regulatory SNVs | Coding SNPs/SNVs | RefSeq Gene | Regulatory SNVs | Coding SNPs/SNVs | RefSeq Gene | Regulatory SNVs | Coding SNPs/SNVs |
| **A427** | 221 | 457 | 437 | 70 | 74 | 110 | 151 | 362 | 77 |
| **A549** | 181 | 242 | 412 | 40 | 36 | 50 | 141 | 180 | 88 |
| **ABC-1** | 141 | 299 | 304 | 39 | 47 | 102 | 102 | 236 | 71 |
| **H322** | 141 | 357 | 342 | 32 | 27 | 86 | 109 | 310 | 74 |
| **H1299** | 126 | 143 | 329 | 29 | 31 | 80 | 97 | 98 | 72 |
| **H1648** | 1,341 | 1,910 | 2,647 | 112 | 95 | 199 | 1,229 | 1610 | 1674 |
| **H1650** | 132 | 245 | 275 | 31 | 31 | 56 | 101 | 200 | 106 |
| **H1703** | 130 | 209 | 263 | 28 | 36 | 104 | 102 | 162 | 45 |
| **H1819** | 195 | 822 | 677 | 49 | 52 | 159 | 146 | 749 | 245 |
| **H1975** | 354 | 697 | 774 | 85 | 75 | 134 | 269 | 568 | 165 |
| **H2126** | 171 | 266 | 356 | 39 | 27 | 46 | 132 | 215 | 133 |
| **H2228** | 327 | 558 | 782 | 91 | 84 | 157 | 236 | 439 | 112 |
| **H2347** | 377 | 726 | 829 | 137 | 136 | 329 | 240 | 544 | 151 |
| **II-18** | 193 | 392 | 393 | 39 | 40 | 101 | 154 | 324 | 77 |
| **LC2ad** | 139 | 208 | 253 | 42 | 31 | 61 | 97 | 163 | 57 |
| **PC-9** | 230 | 481 | 584 | 51 | 44 | 126 | 179 | 409 | 128 |
| **PC-14** | 4 | 4 | 5 | 0 | 0 | 0 | 4 | 4 | 5 |
| **RERF-LC-Ad1** | 318 | 616 | 775 | 71 | 77 | 251 | 247 | 496 | 147 |
| **RERF-LC-Ad2** | 304 | 579 | 641 | 80 | 81 | 152 | 224 | 456 | 125 |
| **RERF-LC-KJ** | 309 | 704 | 620 | 71 | 72 | 128 | 238 | 585 | 112 |
| **RERF-LC-MS** | 175 | 220 | 331 | 50 | 39 | 93 | 125 | 171 | 60 |
| **VMRC-LCD** | 249 | 810 | 556 | 68 | 80 | 150 | 181 | 682 | 142 |
| **RERF-LC-OK** | 344 | 924 | 793 | 98 | 109 | 263 | 246 | 763 | 166 |
| **Average** | 265 | 516 | 582 | 59 | 58 | 128 | 207 | 423 | 175 |

*Table 16 shows number of successfully phased and unphased allele imbalance genes*

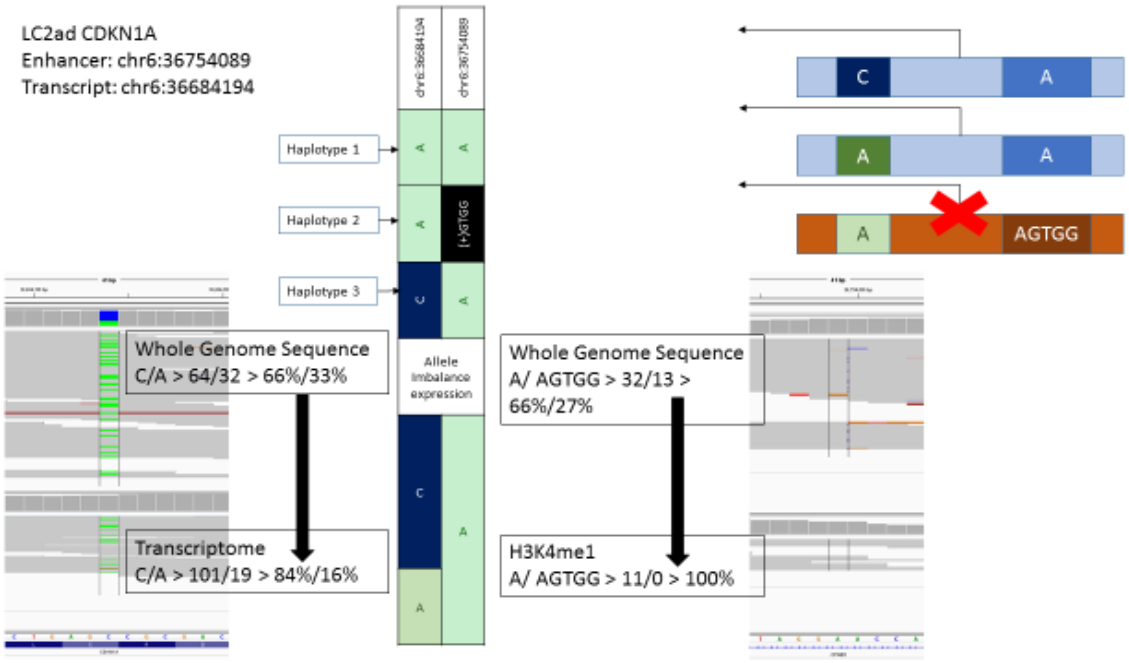


*Figure 26 shows phasing of allele imbalance expression positive gene CDKN1A in LC2ad. The regulatory SNVs in marked by H3K4me1 at chr6:36754089 and Coding SNP/SNV is found at chr6:36684194. The variant "A" in regulatory SNVs was found to be in cis-configuration with both coding variants on the different haplotypes. Explaining the incomplete imbalance found in transcriptomes, supporting the functional relevance of the regulatory SNVs.*

In Table 17, I summarized the list of allele imbalance expression genes with phasing information, which were detected in more than 3 cell lines. CROCC was the most frequent gene. This gene encodes Rootletin a protein crucial in centromere cohesion although its role in cancer development is still unknown. One of the most interesting genes in this list was BRCA1, a tumor suppressor gene. This gene is one of the most important genes in breast cancers. Its mutation was also reported in Lung adenocarcinoma cell lines (Table 2). In my data set the mutation were found in A549, RERF-LC-Ad2 and RERF-LC-Ad1 cell lines.

The relevance of the recurring genes should be interpreted in a careful manner. Further independent validation analysis may be need regarding the possible technical errors, such as bait coverage, SNPs/SNVs density and the presence of transcript SNPs/SNVs.

During the allele imbalance analysis, I encountered the problem of polyploidy, which was exemplified above in the cases of the CDKN1A gene in LC2ad. Therefore, I considered that phasing should be the most direct approach in solving the problem. The case of H322's MAP2K3 (Figure 13) provided me with a good starting point to consider this approach. By following the same approach as I employed in CDKN1A gene, I retrieved the haplotypes from the phase block (see Figure 27 for full phase block). Due to a large number of the corresponding SNPs/SNVs, I removed any irrelevant SNPs/SNVs from the constructed haplotypes retaining only the regulatory SNVs and transcripts of interested. (see Figure 28 ). Using allele imbalance expression, I examined on which transcripts were active and if those activation were concordant with the detected ChIP-seq

variants. For the coding SNPs/SNVs, which expression profile could not be consistently explained by the diploid model (Figure 28, Top, in Black Box). However this could be explained by also considering the phasing information. The unique localization of the variants in the haplotypes were revealed from there. The regulatory SNVs "GG" was first thought to be more active than "AA" from allele-bias regulatory activity alone (genomic 78% tags vs H3K27Ac 90% tags), and after considering the haplotypes, I realized that many of the "GG" variants were also on the same alleles with the inactive transcripts. For the "GG" variants to be detected as "active", the frequency transcripts from inactive allele are required to be low, rendering the imbalance detectable for the transcripts. In many coding SNPs/SNVs in transcriptome, however, this is very unlikely. Frequency of all of the coding SNPs/SNVs in the whole genome sequencing data suggested a 50% heterozygous frequency for this gene (Figure 13). While functional relevance of "GG" over "AA" allele may be present, the imbalance observed in ChIP-seq data was more likely to be derived from a much stronger, undetected mechanism.

Taken together, I concluded that phasing should provide a critical and unique piece of information for precise interpretations of the cancer genome in many aspects, including functional relevance of single nucleotide substitutions, copy number alterations and more complex polyploidy. In cancer genome analysis, genomic regions or genes where the polyploidy and copy number aberrations take place may be more likely to have functional relevance. As phasing technology is rapidly advancing, I believe that incorporating phasing information would become the standard practice in future sequencing pipeline.

*Table 17 shows list of phased allele imbalance expression genes with more than 3 supporting cell lines. In total there were 35 RefSeq transcript and 25 Gene Symbol.*

| RefSeq | GeneSymbol | # Cell Lines | # Coding SNPs/SNVs | # Regulatory SNVs |
|---|---|---|---|---|
| **NM_014675** | CROCC | 10 | 26 | 43 |
| **NM_001128223** | ZNF717 | 7 | 180 | 498 |
| **NM_001256139** | CAPG | 6 | 5 | 4 |
| **NM_002180** | IGHMBP2 | 6 | 13 | 10 |
| **NM_001128592** | PSMG4 | 5 | 4 | 8 |
| **NM_201380** | PLEC | 5 | 31 | 12 |
| **NM_001178090** | ZNF454 | 4 | 2 | 19 |
| **NM_022350** | ERAP2 | 4 | 6 | 6 |
| **NM_001008892** | NIPA2 | 4 | 4 | 5 |
| **NM_001008894** | NIPA2 | 4 | 4 | 5 |
| **NM_001008860** | NIPA2 | 4 | 4 | 5 |
| **NM_030922** | NIPA2 | 4 | 4 | 5 |
| **NM_001184888** | NIPA2 | 4 | 4 | 5 |
| **NM_001184889** | NIPA2 | 4 | 4 | 5 |
| **NM_001130140** | ERAP2 | 4 | 6 | 6 |
| **NM_001178089** | ZNF454 | 4 | 2 | 19 |
| **NM_182594** | ZNF454 | 4 | 2 | 19 |
| **NM_001047** | SRD5A1 | 3 | 4 | 11 |
| **NM_001031665** | ZNF816 | 3 | 7 | 11 |
| **NM_001202456** | ZNF816 | 3 | 7 | 11 |
| **NM_001202457** | ZNF816 | 3 | 7 | 11 |
| **NM_002610** | PDK1 | 3 | 2 | 6 |
| **NM_005742** | PDIA6 | 3 | 3 | 5 |
| **NM_001142645** | TMEM194B | 3 | 2 | 3 |
| **NM_015311** | OBSL1 | 3 | 10 | 2 |
| **NM_018260** | ZNF701 | 3 | 5 | 12 |
| **NM_152559** | WBSCR27 | 3 | 3 | 5 |
| **NM_001135924** | VWDE | 3 | 18 | 4 |
| **NM_001163391** | ZSCAN12 | 3 | 4 | 3 |
| **NM_181453** | GCC2 | 3 | 2 | 3 |
| **NM_005649** | ZNF354A | 3 | 3 | 14 |
| **NM_007297** | BRCA1 | 3 | 7 | 1 |
| **NM_002568** | PABPC1 | 3 | 13 | 15 |
| **NM_015914** | TXNDC11 | 3 | 3 | 9 |
| **NM_001134647** | AFAP1 | 3 | 6 | 11 |

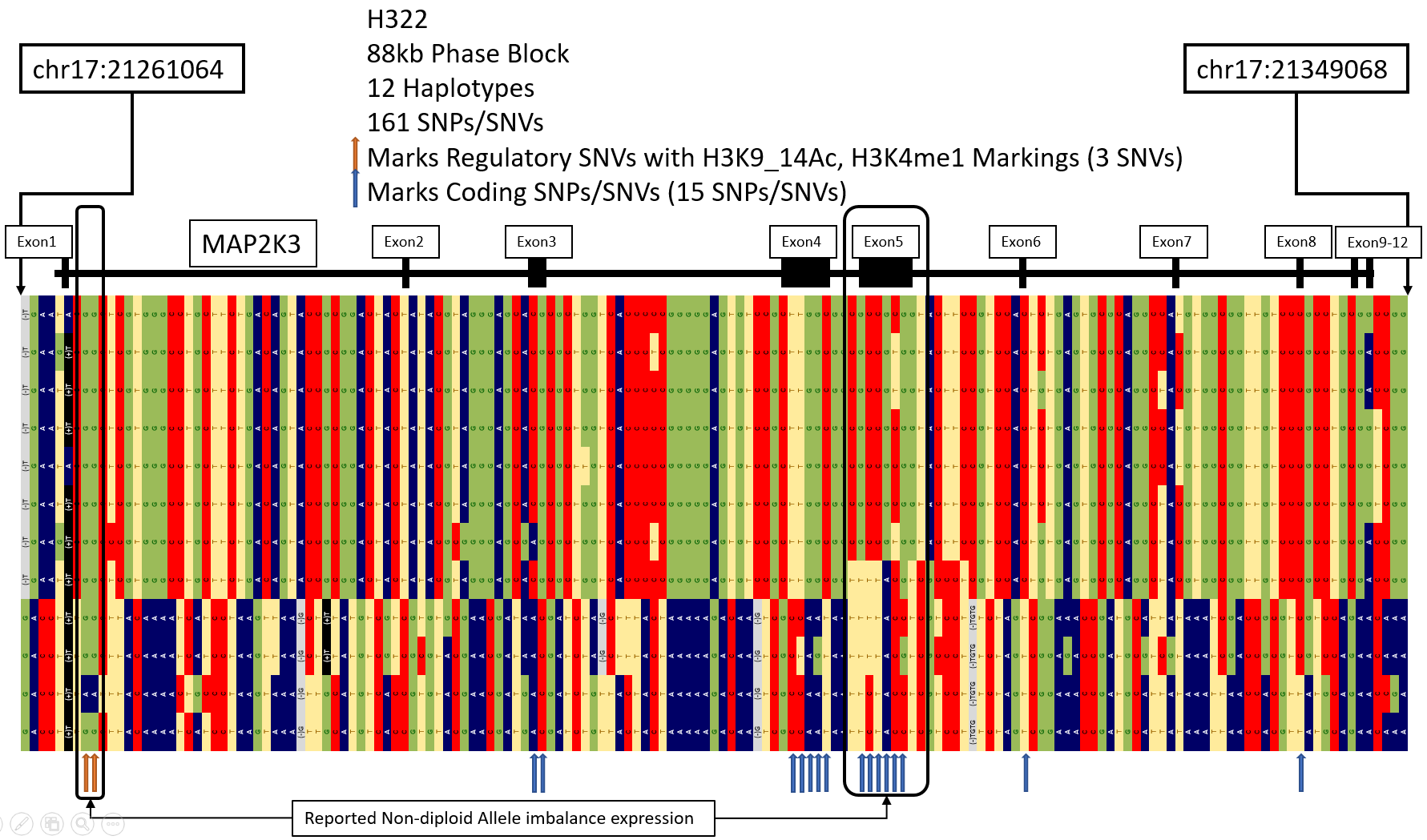


*Figure 27 shows a full phase block of H322 containing MAP2K3 gene. The Block start from chr17:21261064 and end at chr17:21349068, the length of the phase block was 88kb with 161 SNVs and 12 Haplotypes.*

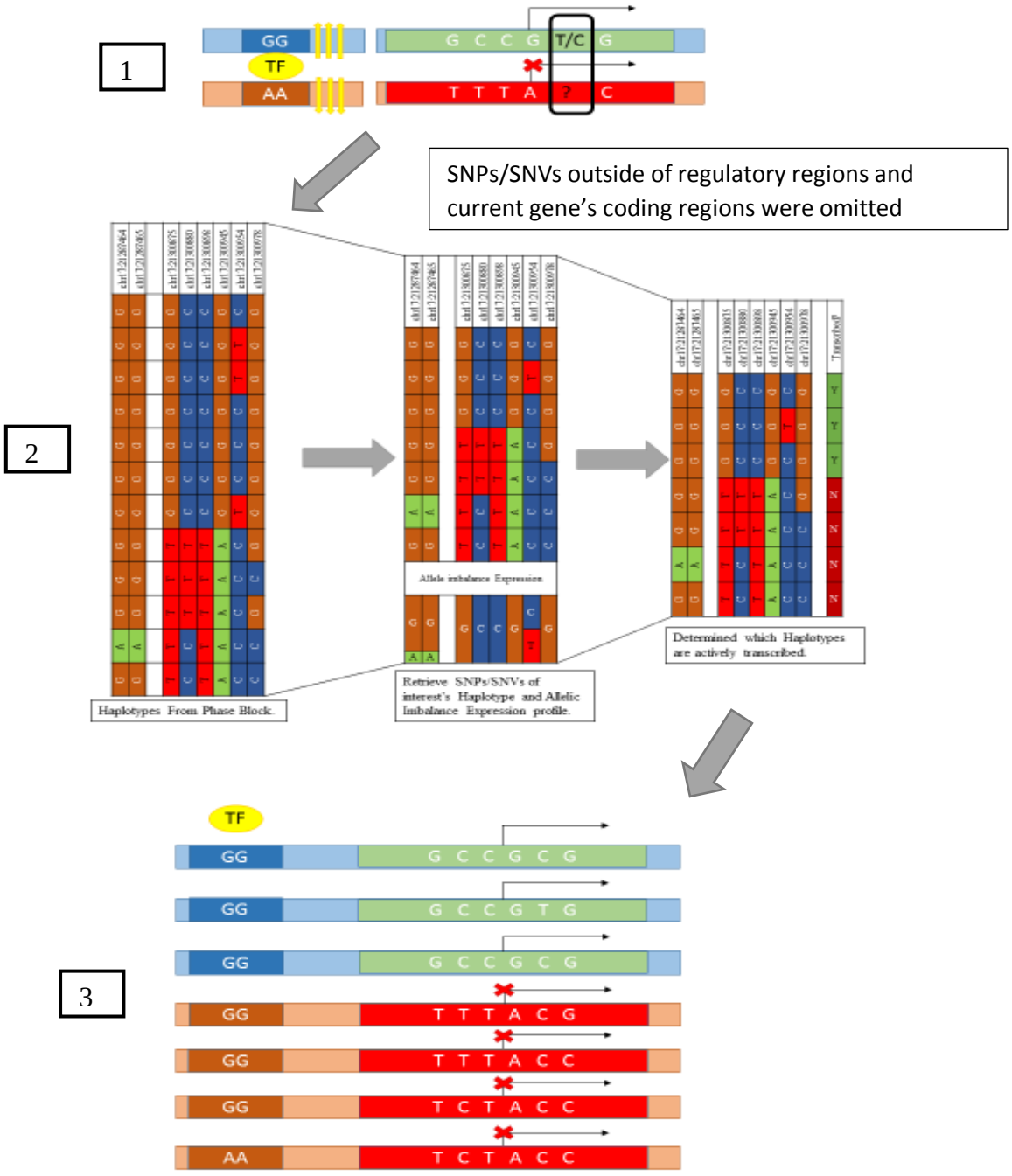


*Figure 28 shows phasing and allele imbalance analysis of H322's MAP2K3 gene. The regulatory SNVs (chr17: 21287464- 21287465) and coding SNPs/SNVs in exon5 (chr17:21300875, chr17: 21300880, chr17: 21300898, chr17: 21300945, chr17: 21300954 and chr17: 21300978) as the SNPs/SNVs of interested. 2). Haplotypes configuration were constructed from phase block with only SNPs/SNVs of interested and transcriptional statuses were decieded. 3). Final regulatory SNVs-transcripts configurations and their effects were completed.*

# Conclusion and Future Plans

In this study, I have attempted to elucidate the function of the ubiquitous but unknown somatic mutations in the regulatory region of the cancer genome, in hope of gaining new insight in both the sophisticate process of carcinogenesis, and the even complex gene regulation systems. These insights would improve our understanding of cancer genomics, which in turn would provide us with better tools in cancer preventions and treatments.

By performing allele imbalance expression analysis by utilizing variant frequency in multi-omics dataset covering genome, transcriptome and regulatory elements and the no less importantly adaptation of the most recent synthetic and physical long read sequencing, in study of 23 lung adenocarcinoma cell lines. I identified, on average, 516 potential functional regulatory SNVs in 256 genes per cell line, and 58 of these SNVs were further delineated by the Haplotypes phasing. These genes included both previously established oncogenic or tumor suppressor genes such as KMT2C, CDKN1A or BRCA1 and novel potentially functional genes such as MAP2K3.

The functional regulatory variants detected in this study were only a small fraction of perhaps larger and more complex system of epigenetic regulations and its outputs. I believe with increasing number of new findings from different approaches combined, the more comprehensive view of such fabricated molecular regulatory network will be revealed. I believe that this work, although small, has paved the first step towards that goal.

In addition to the results, I believe that the strategies and methods developed in this study, especially for those used for the polypoid haplotype phasing from synthetic long read sequencing and allele imbalance expression analysis will help researchers to utilize those latest omics analytical methods in various circumstances. However, as present methods are new and are still lack of supporting evidences generally, fine-tuning and improvement would be necessary. The phasing schemes, in particular, need to be evaluated by physical long read method.

Nevertheless, I believe that this study had provided a solid groundwork for analyzing regulatory elements, starting from data collections, methodological development and finally to biological interpretations; thus would enable us further explore the mysteries of aberrant gene expression regulations in cancers.

# References


Ayako Suzuki, Hideki Makinoshima, Hiroyuki Wakaguri, Hiroyasu Esumi,Sumio Sugano, Takashi Kohno, Katsuya Tsuchihara and Yutaka Suzuki. 2014. "Aberrant transcriptional regulations in cancers: genome, transcriptome and epigenome analysis of." *Nucleic Acids Research* 13557-13572.

Baran, Yael, Meena Subramaniam, Anne Biton, Tukiainen Taru, Emily K. Tsang, Rivas A. Manuel, Matti Pirien, et al. 2015. "The landscape of genomic imprinting across diverse adult human tissues." *Genome Research* 25: 927-936.

Changcun Guo, Lee H. Chen, Yafen Huang, Chun-Chi Chang, Ping Wang, Christopher J. Pirozzi, Xiaoxia Qin, Xuhui Bao, Paula K. Greer, Roger E. McLendon, Hai Yan, Stephen T. Keir, Darell D. Bigner, and Yiping He. 2013. "KMT2D maintains neoplastic cell proliferation and global histone H3 lysine 4 monomethylation." *Oncotarget* 4(11): 2144-2153.

Charles S. Dela Cruz, Lynn T. Tanoue,and Richard A. Matthay. 2011. "Lung Cancer: Epidemiology, Etiology, and Prevention." *Clinics in Chest Medicine* 605-644.

David S. Johnson, Ali Mortazavi, Richard M. Myers and Barbara Wold. 2007. "Genome-Wide Mapping of in Vivo Protein-DNA Interactions." *Science* 1497-1502.

Dou, Rajesh C. Rao&Yali. 2015. "Hijacked in cancer: the KMT2 (MLL) family of methyltransferases." *Nature Reviews Cancer* 15, 334-346.

Durbin, Heng Li Richard. 2009. "Fast and accurate short read alignment with Burrows–Wheeler transform ." *Bioinformatics* 1754-1760.

Ekta Khurana, Yao Fu,Dimple Chakravarty,Francesca Demichelis,Mark A. Rubin & Mark Gerstein. 2016. "Role of non-coding sequence variants in cancer." *Nature Reviews Genetics* 17, 93-108.

ENCODE. 2017. *ENCODE: Encyclopedia of DNA Elements.* 1 25. https://www.encodeproject.org/.

Forbes. 2014. "COSMIC: exploring the world's knowledge of somatic mutations in human cancer." *Nucleic Acids Rearch.*

Grace X Y Zheng, Billy T Lau, Michael Schnall-Levin, Mirna Jarosz, John M Bell, Christopher M Hindson, Sofia Kyriazopoulou-Panagiotopoulou, Donald A Masquelier, Landon Merrill, Jessica M Terry, Patrice A Mudivarti, Paul W Wyatt, Rajiv Bharadwaj, Anthony J. 2016. "Haplotyping germline and cancer genomes with high-throughput linked-read sequencing." *Nature Biotechnology* 303-311.

Hughes, Bryan A. Chan Brett G.M. 2015. "Targeted therapy for non-small cell lung cancer: current standards and the promise of the future." *Translational Lung Cancer Research* 4(1): 36-54.

Jason D Buenrostro, Paul G Giresi, Lisa C Zaba, Howard Y Chang & William J Greenleaf. 2013. "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position." *Nature Methods* 1213-1218.

Joshua BN Dawkins, Jun Wang, Eleni Maniati, James A Heward, Lola Koniali, Hemant M kocher, Sarah A. Martin, Claude Chelala, Frances R Balkwill, Jude Fitzgibbon and Richard P Grose. 2016. "Reduced Expression of Histone Methyltransferases KMT2C and KMT2D Correlates with Improved Outcome in Pancreatic Ductal Adenocarcinoma." *Cancer Research* 76(16): 4861-71.

Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup. 2009. "The Sequence Alignment/Map format and SAMtools." *Bioinformatics* 20780-9.

Matthew L. Speir, Ann S. Zweig, Kate R. Rosenbloom, Brian J. Raney, Benedict Paten, Parisa Nejad, Brian T. Lee, Katrina Learned, Donna Karolchik, Angie S. Hinrichs, Steve Heitner, Rachel A. Harte, Maximilian Haeussler and Luvina Guruvadoo. 2016. "The UCSC Genome Browser database: 2016 update." *Nucleic Acids Res* D717-D725.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010 . "The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data." *GENOME* 20:1297-303.

NCBI. 2017. *MAP2K3 mitogen-activated protein kinase kinase 3 [ Homo sapiens (human) ].* 01 27. https://www.ncbi.nlm.nih.gov/gene/5606.

Nele Gheldof, Marion Leleu, Daan Noordermeer, Jacques Rougemont, Alexandre Reymond. 2011. "Detecting Long-Range Chromatin Interactions Using the Chromosome Conformation Capture Sequencing (4C-seq) Method." *Methods in Molecular Biology* 211-225.

Potter, Paolo Vineis Arthur Schatzkin and John D. 2010. "Models of carcinogenesis: an overview." *Carcinogenesis* 31(10): 1703–1709.

ROADMAP PROJECT. 2017. *ROADMAP epigenomics Project.* 1 25. http://www.roadmapepigenomics.org/.

Shiyong Li, BSa, Yoon-La Choi, MD, PhDb, Zhuolin Gong, PhDa, Xiao Liu, PhDa, Maruja Lira, BSc, Zhengyan Kan, PhDc, Ensel Oh, PhDb, Jian Wang, PhDd, Jason C. Ting, PhDd, Xiangsheng Ye, PhDd, Christoph Reinhart, PhDd, Xiaoqiao Liu, PhDe, Yunfei Pei, PhDe, W. 2016. "Comprehensive Characterization of Oncogenic Drivers in Asian Lung Adenocarcinoma." *Journal of Thoracic Oncology* 2129–2140.

Susanne Horn, Adina Figl, P. Sivaramakrishna Rachakonda, Christine Fischer, Antje Sucker, Andreas Gast, Stephanie Kadel, Iris Moll, Eduardo Nagore, Kari Hemminki, Dirk Schadendorf and Rajiv Kumar. 2013. "TERT Promoter Mutations in Familial and Sporadic Melanoma." *Science* 959-961.

TCGA. 2014. "Comprehensive molecular profiling of lung adenocarcinoma." *Nature* 511 543-550.

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS. 2008. "Model-based analysis of ChIP-Seq (MACS)." *Genome Biol.* 9(9):R137.

ACKNOWLEDGEMENTS

I would like to express my heartfelt gratitude to Dr. Suzuki Y., my academic advisor whose expertise, knowledge and guidance made this work possible. I especially grateful to Dr. Suzuki A. for her vital teaching and invaluable advices.

I am also would like to extend my thanks to my friends and colleagues at my laboratory for their continue support, advice and friendships though out my study and especially technical assistance both experimentally by Y. Ishikawa and computationally by Y. Kuze .and in equal importance the human genome center's Shirokanedai3 supercomputer staff and maintenance team.

I am thankful for financial support from Monbukagakusho (MEXT) scholarship from The Ministry of Education, Culture, Sports, Science, and Technology and Embassy of Japan in Bangkok for this priceless, once in the life time opportunity.