Title
Integrative Analysis of Genomic and Transcriptomic Data in Taiwanese Lung Adenocarcinomas

Author
Ho, Hao

Publication Date
2017

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Integrative Analysis of Genomic and Transcriptomic Data in Taiwanese Lung Adenocarcinomas

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Statistics

by

Hao Ho

2017

ABSTRACT OF THE DISSERTATION

Integrative Analysis of Genomic and Transcriptomic Data in Taiwanese Lung Adenocarcinomas

by

Hao Ho
Doctor of Philosophy in Statistics
University of California, Los Angeles, 2017
Professor Ker-Chau Li, Chair

In this thesis, we studied genomic and transcriptomic data from over 300 Taiwanese lung cancer patients. For structural variation analysis, we proposed a workflow to detect inter-chromosomal structural variation using whole genome sequencing data and introduced an integrated ESP plot for the visualization. We studied somatic DNA alterations and constructed a comprehensive landscape in Taiwanese lung adenocarcinomas by whole exome sequencing and array CGH data. At the single nucleotide level, we identified non-synonymous recurrent point mutations using a binomial probability model. The potential clinical relevance was demonstrated by a survival analysis of patients' relapse-free survival. Mutation variant allele frequency was integrated for improving prognosis power. When exploring the potential downstream, we identified a miRNA expression correlated with these recurrent point mutations. In the study of differential gene expressions between EGFR mutant and wild-type tumors. We derived a statistical framework that combines differential expression analysis and differential regulation analysis to form an enrichment test for identifying critical regulator on the cis-regulatory network. A modified liquid association was introduced for quantifying the change of co-variations in the differential regulation analysis. By integrating copy number, miRNA expression and gene expression data, several key regulators and their cis-targets were identified and visualized together as a network. For a statistical issue of liquid association, we discussed the effects of ignoring background variables to the liquid association scoring method and proposed adjustment methods to marginalize their influence.

The dissertation of Hao Ho is approved.

Stanley Nelson

Ying Nian Wu

Qing Zhou

Ker-Chau Li, Committee Chair

University of California, Los Angeles

2017

To my parents

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I want to first express my gratitude to my advisor, Dr. Ker-Chau Li. Thank you for all the guidance and support that you gave me. Much more than that, thank you for lifting me up when I was in the deepest valley. I would also like to thank my committee members: Dr. Ying Nian Wu, Dr. Qing Zhou and Dr. Stanley Nelson for offering valuable suggestions to my research. My thanks also go to my officemates and colleagues: Jiashen You, Yi (Jacky) Yi, Nathan James Langholz, Kyle Andrew Hasenstab, Nikhyl Bryon Aragam, Guani Wu, Qian Xiao, Levon Demirdjian, Seunghyun Min, Jiaying Gu, Qiaoling Ye, Medha Uppala, and many others who helped me out when I got stuck in my project. Thank you all for being supportive through out the years.

VITA

2006	B.S. (Mathematics), National Taiwan University, Taipei, Taiwan
2010	M.S. (Mathematics), National Taiwan University, Taipei, Taiwan
2010 - 2017	Research Assistant and Teaching Assistant, Statistics Department, UCLA
2017	Lecturer, Statistics Department, UCLA

PUBLICATIONS

J. F. Chen*, **H. Ho***, J. Lichterman*, Y. T. Lu*, Y. Zhang, M. A. Garcia, S. F. Chen, A. J. Liang, E. Hodara, H. E. Zhau, S. Hou, R. S. Ahmed, D. J. Luthringer, J. Huang, K. C. Li, L. W. K. Chung, Z. Ke, H. R. Tseng, and E. M. Posadas. Subclassification of prostate cancer circulating tumor cells by nuclear size reveals very small nuclear circulating tumor cells in patients with visceral metastases. *Cancer*, 121(18):3240-3251, 2015.

R. Jiang*, Y. T. Lu*, **H. Ho***, B. Li*, J. F. Chen, M. Lin, F. Li, K. Wu, H. Wu, J. Lichterman, H. Wan, C.-L. Lu, W. Ouyang, M. Ni, L. Wang, G. Li, T. Lee, X. Zhang, J. Yang, M. Rettig, L. W. K. Chung, and H. Yang. A comparison of isolated circulating tumor cells and tissue biopsies using whole-genome sequencing in prostate cancer. Oncotarget, 6(42):44781-93, 2015.

H. Y. Chen, S. L. Yu, B. C. Ho, K. Y. Su, Y. C. Hsu, C. S. Chang, Y. C. Li, S. Y. Yang, P. Y. Hsu, **H. Ho**, Y. H. Chang, C. Y. Chen, H. I. Yang, C. P. Hsu, T. Y. Yang, K. C. Chen, K. H. Hsu, J. S. Tseng, J. Y. Hsia, C. Y. Chuang, S. Yuan, M. H. Lee, C. H. Liu, G. I. Wu, C.

A. Hsiung, Y. M. Chen, C. L. Wang, M. S. Huang, C. J. Yu, K. Y. Chen, Y. H. Tsai, W. C. Su, H. W. Chen, J. J. W. Chen, C. J. Chen, G. C. Chang, P. C. Yang, and K. C. Li. R331W missense mutation of oncogene YAP1 is a germline risk allele for lung adenocarcinoma with medical actionability. *Journal of Clinical Oncology*, 33(20):2303-2310, 2015.

CHAPTER 1

Introduction

Lung adenocarcinoma is a heterogeneous disease and shows high rates of somatic DNA alterations. The rapid development of high-throughput technologies allows scientists to study multi-level molecular profiles from biological samples. The study by The Cancer Genome Atlas[11] (TCGA) provided comprehensive molecular profiles from a large cohort of lung adenocarcinomas. This dataset has been wildly used in the studies of driver gene identification, prognostic signature identification, integrative clustering, and other methods for refining in the patient stratification. However, in Asian, lung adenocarcinoma features a higher rate of nonsmokers and the frequencies of driver mutations differ between the East and the West populations. To help better understand the genomic alterations and the potential downstream molecular changes of lung adenocarcinomas in Asian population, an integrative study of multi-level molecular profiles from large cohorts is still in need.

In this thesis, we studied genomic and transcriptomic data from over 300 Taiwanese lung cancer patients. Several high throughput assays including second generation sequencing, array CGH, gene expression microarray, and microRNA microarray were applied in theses specimens. We organized this thesis in the following way. In Chapter 2, we proposed a workflow to detect inter-chromosomal structural variation using whole genome sequencing data. We presented an integrated ESP plot for visualizing the structural variation. This visualization method helped conduct subtype classification, zygosity determination and breakpoints estimation for a complex event.

In Chapter 3, we presented a comprehensive landscape of somatic DNA alterations in Taiwanese lung adenocarcinomas constructed by using whole exome sequencing and array CGH data. At the single nucleotide level, we identified 18 non-synonymous recurrent point

mutations using a binomial probability model. The potential clinical relevance of these recurrent point mutations was demonstrated by a survival analysis of patients' relapse-free survival. Integrating mutation variant allele frequency for improving prognosis power was introduced. When exploring the potential downstream, we identified a miRNA expression correlated with these recurrent point mutations. The miRNA expression also correlated with patients' relapse-free survival.

In Chapter 4, we presented our analysis of the differential gene expressions between EGFR mutant and wild-type tumors. We derived a statistical framework that combines differential expression analysis and differential regulation analysis to form an enrichment test for identifying critical network regulator. A modified liquid association was introduced for quantifying the change of co-variations in the differential regulation analysis. By integrating copy number, miRNA expression and gene expression data, several key regulators and their cis-targets were visualized together as a network.

In Chapter 5, we discussed the effects of ignoring background variables to the liquid association scoring method and proposed adjustment methods to marginalize their influence.

CHAPTER 2

An Inter-chromosomal Structural Variation Study using Second Generation Sequencing Data from a Multiplex Family

2.1 Introduction

Structural variations (SVs) are observed widely in human genomes[44]-[29], and their biological impacts have received increasing attentions in the past decade[42]-[23]. Inter-chromosomal structural variations including duplication, transposition, and translocation are more complex than the basic structural variations such as insertion, deletion (indel) and inversion. They might be composites of multiple types of basic structural variations and involve multiple breakpoint pairs from nonhomologous chromosomes. These complex events can damage the genome in different ways: (1) inserting a transposed segment to a functional gene might disable that gene; (2) missing genomic materials between breakpoints might not be repaired correctly; (3) a translocation joining two genes might create a gene fusion. Some human diseases are caused by these genomic abnormal such as hemophilia A and B, porphyria, severe combined immunodeficiency and cancer.

Historically, to reveal genomic structure has driven the developments of experimental techniques, such as fluorescence in situ hybridization (FISH), comparative genomic hybridization (CGH) and end sequence profiling (ESP) technique. FISH uses fluorescent probes that bind to only the highly similar sequences of the chromosome to localize the specific sequence's presence or absence. This chromosome painting technique directly visualizes labeled chromosomes, but it is laborious and time-consuming. Moreover, the main limitations are low

resolution and lack of copy number changing information. CGH technique measures the copy number changes (gains/losses) by hybridizing different fluorescently labeled test and normal reference samples to thousands of probes and calculating the ratio of the fluorescence intensity of the test to the reference. Current CGH technique offers a resolution typically of 20-80 base pairs (bps), which is much higher than the resolution of 100k bps offered by older bacterial artificial chromosome (BAC) arrays. There are two main limitations in using CGH to reveal structural variations. First, CGH can detect only the unbalanced structural variations. Second, it offers no information about the positions of the duplications. Balanced reciprocal translocation and inversion cannot be detected by CGH technique since they do not change the copy numbers.

End Sequence Profiling (ESP) technique constructs a paired-end library of the sequences of both ends of BACs for the sample genome and maps the end sequences to the reference human genome. After mapping to the reference genome, a BAC corresponds to a pair $(x, y) \in G \times G$ where it is ordered such that $x < y$. Since BACs are set to have target insert size range (l, L), a pair is called valid if x and y are on the same chromosome, have opposite convergent orientation and $l < y - x < L$. Otherwise it is called invalid. Each invalid pair results from a structural variation or experimental error. This technique motivated the ESP sorting problem[41]. ESP technique has drawn a blueprint of the methodology for detecting the structural variations using whole genome re-sequencing approach. However, due to its low resolution, the ability was limited.

Recently, the high throughput second generation sequencing technology like Illumina Genome Analyzer, Roche 454, and ABI SOLiD has replaced the previous sequencing techniques. Millions of reads can be sequenced simultaneously. The dramatically increased amount of reads offers a much higher resolution for detecting the structural variations, but the short read length and the high sequencing error also increased the difficulty of analysis. These second generation sequencing platforms are all able to generate the mate-pair or paired-end data library. The paired-end data is generated by fragmenting sample genome into short fragments (<300 bps) and sequencing both ends of the fragments. The paired-end reads are generated with opposite convergent orientation and tighter insert size (<300 bps)

distribution. The mate-pair data is generated by a different wet-lab technique. First, it fragments sample genome with selected insert size. Second, it circularizes the fragments by an internal adaptor then shears them at random position. Finally, it sequences both sides around the internal adaptor. The mate-pair reads are generated with common strands and larger insert size (1,000-2,000 bps). From the computational point of view, the difference between these two methods is small. However, due to the tighter insert size distribution, the paired-end data provides higher resolution. On the other hand, due to the larger insert size, the mate-pair data has better ability to reveal the large-scale structural variations.

As the second generation sequencing technology appeared, various structural variation prediction methods based on whole genome re-sequencing approach were proposed[36]-[51]. After aligning sample reads to the reference genome, four main features, insert size, order, strand and the depth of coverage, are used for structural variation detection and break-point estimation. We briefly review the published methods for detecting inter-chromosomal structural variations. Most methods used insert size, order, and strands of a pair to classify whether it is valid or invalid (some term invalid pair discordant or aberrant pair). As in the ESP technique, each invalid pair results from a structural variation or experimental error. However, the huge amount of reads and high experimental error rate increased false positive rate.

Campbell *et al.*[2] assumed most of the invalid pairs result from the experimental error. They used the uniqueness mapping score as a filter to reduce the false positive. Only those pairs having two high-score reads were used for structural variation prediction. Although this filter was introduced to reduce the experimental error, it also sacrificed the multiple mapping reads. Since about half of human genome is in the repetitive region, the reads sequenced from the repeated region will have low uniqueness mapping score even there is no experimental error.

Instead of only using the information of individual pair as a filter, more methods considered the relationships between pairs and then clustered the correlated pairs. Lee *et al.*[31] proposed a probabilistic framework to cluster the invalid pairs. Given the insert sizes and aligned positions of two pairs, they evaluated the probability of these two alignments given

that they can be explained by the same structural variation. Then they performed the hierarchical clustering method by using that probability as a similarity measure. This probability depends on the overlapping length between two pairs and the insert size distribution. Some other methods also cluster the invalid pairs based on their overlaps, such as Korbel *et al.*[27], Chen *et al.*[4] and Quinlan *et al.*[40]. One important issue is, after clustering, how many pairs are enough to support a structural variation. Korbel *et al.*[27] proposed a simulation framework to decide the threshold and they also suggested using multiple thresholds.

Unlike the previous methods clustered pairs based on the relationship between two pairs directly, Sindi *et al.*[46] depicted the relation between an invalid pair and the region of corresponding putative breakpoints. The putative breakpoints region of an invalid pair is illustrated as a trapezium on the 2D ESP plot. Invalid pairs with different strands and orders have trapeziums with four different directions. This observation suggests that strands and orders of invalid pairs should be considered for clustering. It also motivated the maximal intersections of breakpoint regions problem.

Instead of clustering invalid pairs, another strategy is to bin reference genome into bins then cluster the bins based on the connections of bins. Clark *et al.*[9] first divided reference genome into bins with length 500-bps sequentially stepped 100-bps apart from their start positions. Each bin, as a source bin, was paired with another destination bin that have most invalid pairs connected to the source bin. A paired bin with the number of invalid pairs more than a threshold will be called a bin-set. The bin-sets were clustered by the positions of their two bins. After filtering out the clusters containing too few or too many bin-sets and the redundant clusters, this approach offers 100-bps resolution breakpoint estimations.

Another important feature is the depth of coverage which can be used for breakpoint estimation and determining zygosity. Chiang *et al.*[6] used local change point analysis techniques to find the proper breakpoints for the windows. Lee *et al.*[31] used the depth of coverage of invalid pairs and valid pairs within the event range to determining zygosity of the event. Chen *et al.*[5] sequenced flow-sorted derivative chromosomes using second generation sequencing technique. They used local change point analysis on the depth of coverage and got single nucleotide resolution breakpoint estimations.

The previous prediction methods can identify the approximate location of breakpoints. Split mapping can be used to find the exact breakpoint positions and reconstruct the sample sequence of the structural variations. If a read is generated across a breakpoint, the prefix and the suffix should be able to align to two different locations. The split mapping problem might be infeasible for the short length reads since there are too many spurious mappings on the whole genome. Ye *et al.*[51] proposed the anchored split mapping algorithm to reduce the search space by considering the mate of a split read should be aligned within a valid region. This algorithm works well for the simple deletion or insertion. However, Campbell *et al.*[2] found that about half of the rearrangements they found have short stretches of microhomology between the two ends fused together. The anchored split mapping does not work for this case.

In this study, we sequenced six blood samples from a multiplex family using ABI-SOLiD system mate-pair library. This family contains six members, one mother, and five children. The mother and four daughters are lung cancer patients, but the son is not. We proposed a strategy to analyze the inter-chromosomal variations of the family members. Our strategy includes variation detection, subtype classification, zygosity determination, breakpoint estimation and variation visualization. Inter-chromosomal variations in cancer patients but not in the normal sample are of interest. We identified two such duplications: (1) a direct duplication from chromosome 7 to chromosome 3, (2) an inverse duplication from chromosome 2 to chromosome 3.

Here we summarize the main contents of this chapter. An introduction of our data generation including reads alignment, pairing, and some summary statistics are given in section 2.2. A two-steps putative inter-chromosomal structural variation region detecting algorithm and the integrated ESP plot will be introduced in section 2.3. This plot is useful for subtype classification, zygosity determination and breakpoints estimation of a complex event. It also suggests the way to perform the subsequent verifications. The results are given in section 2.4. Discussions will be given in section 2.5.

2.2 Materials

2.2.1 Data Generation

In this study, we utilized ABI-SOLiD system to sequence from the samples. The ABI-SOLiD System is a second generation sequencing system using ligation based chemistry with di-base labeled probes, which provides more accurate SNV detection. It queries two adjacent base positions at a time and uses four fluorescent dyes to encode for the sixteen possible two base combinations. The color code is based on the Klein four-group, which is the only abelian subgroup of the permutation group of four elements. Some properties are given as follows:

i. For each di-base, the reverse, the complement, and the reversed complement have the same color.

ii. Two different di-bases with one same base and one different base have different colors.

The first property provides a unique error correcting method that increases the accuracy for SNV detection. However, due to the first property, some de novo assembling methods cannot be directly applied to it.

A mate-pair has forward and reverse two reads, where both reads are sequenced from 5' to 3'. The reverse read is closer to 5' and the forward read is closer to 3'. Therefore, an alignment of a mate-pair is said to be valid (in SOLiD it is called AAA) if two reads have the strand, orientation from R3 to F3 and distant within the target insert size range. Otherwise, it is called invalid.

We utilized the mapping and pairing tools in the BioScope software to align the mate-pair reads. Here we briefly summarize these two algorithms.

Mapping tool in BioScope

i. Convert the sequence of the reference genome into color codes.

ii. For each read, use the first 25 color bases as seed and align the seed to the converted reference sequence. Two mismatched are allowed in seeds alignment. Multiple alignments are also allowed.

iii. For each alignment of a seed, do the local extension to get the maximum aligned length.

- Report all the alignments for all reads.

Pairing tool in BioScope

i. Base on the mate-ID pair two reads

ii. For each mate-pair, if there are valid alignments, compute the pairing score for each valid alignment of a mate-pair. The pairing score is defined as the probability of the alignment given the color codes of the mate-pair

iii. For each mate-pair, if there is no valid alignment, perform the rescue algorithm and score all the alignments

iv. For each mate-pair, report the alignment with the highest score.

We note that, first, the mapping algorithm aligns a read to a position of reference only if the seed of the read can be aligned to that position. Therefore, an alignment of a 50 bps reads might be smaller than 50 bps but no lesser than 25 bps. Second, the pairing algorithm only reports the highest scored pair of alignments. The information of multiple alignments is ignored at this stage. For our sequential analysis, we still use the paired data.

2.2.2 Sequencing Data from a Multiplex Family

We collected six blood samples (symbolized as A to F) from a multiplex family. Five members including the mother (A) and four daughters are cancer patients. The son (F) is the only normal sample. For each sample, we performed two full sequencing runs using the ABI-SOLiD system. Two samples collected earlier (B, C) were sequenced by ABI-SOLiD 3 system. This platform generated 519M and 645M raw 50bp mate-pair reads for each sample. The rest four samples were sequenced by ABI-SOLiD 3 Plus system that provided higher and more stable throughputs. About 800M raw 50 bps mate-pair reads were generated for each sample with this platform.

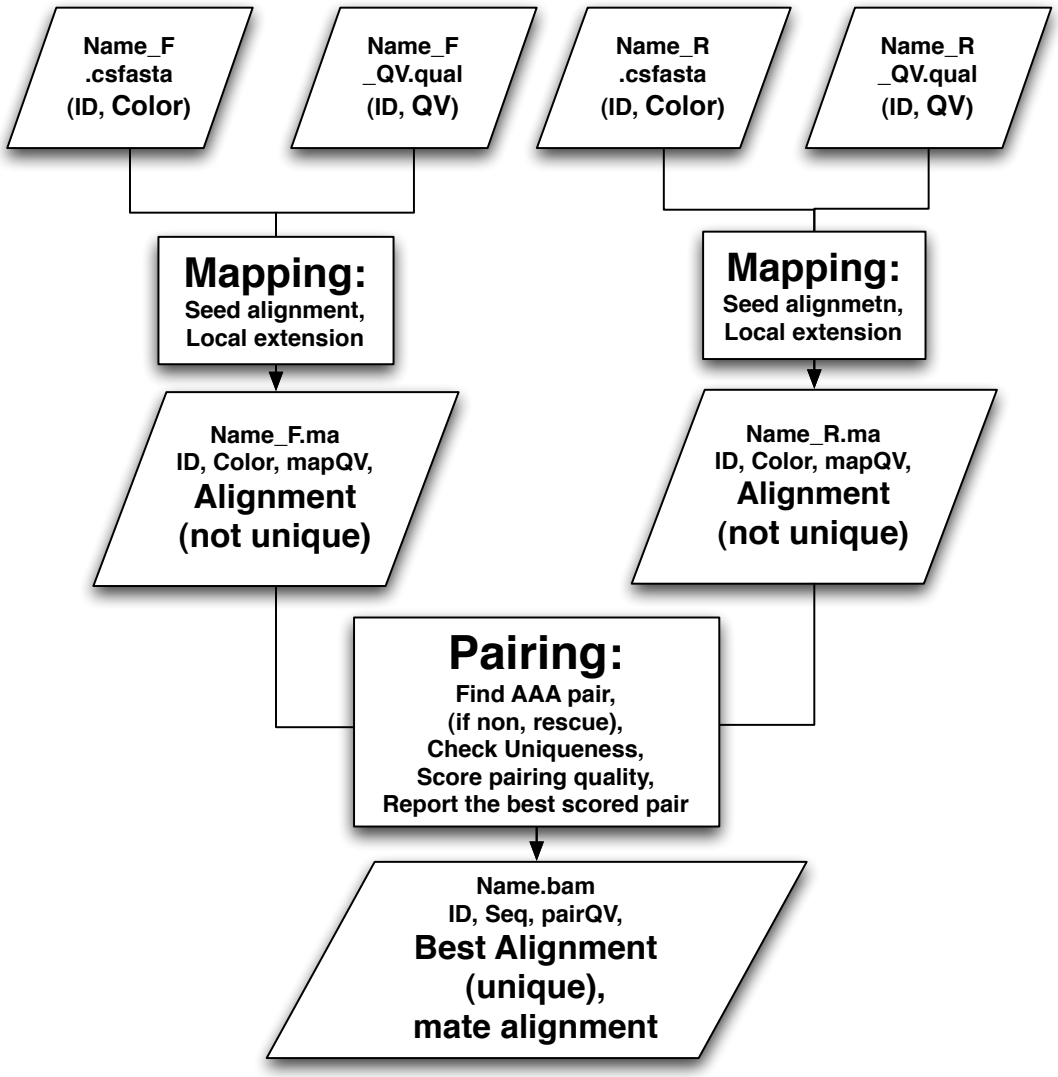

Figure 2.1: A flow chart of BioScope re-sequencing data mapping and pairing tools.

The Bioscope re-sequencing mapping tool and re-sequencing pairing tool were utilized to align the raw data to the reference human genome (hg19). The 1-25 and 16-40 bases of each read were selected for the seed alignment. The criterion of seed alignment was set to be no more than 2 mismatches within the 25bps seeds.

There are about 70% pairs having two reads aligned to the reference, 20% pairs having only one read aligned and 10% having no reads aligned. We notice that the reported alignment for a pair might be the unique alignment or the alignment with the highest score within all multiple alignments. A read having no alignment might be missing or not able to align to the reference. About 10% pairs have two reads aligned to nonhomologous chromosomes. These are the reads we used to detect the inter-chromosomal structural variations.

The pairs having two reads aligned on the same chromosome were used to construct the empirical insert size distribution (fig.2.2) and estimate the quantiles. Since there are only 1% pairs having insert size greater than 3,000, we estimate the truncated mean and standard deviation for the pairs with insert size no greater than 3,000. Five female patients have similar insert size distributions with a median around 1,250, but the normal male sample has a distribution with thinner right tail and larger median.

2.3 Methods

Inter-chromosomal structural variations including duplication, transposition, and translocation are usually complex events and might be composites of basic structural variations such as insertion, deletion, and inversion. However, no matter how complex an inter-chromosomal structural variation is, there must be least one pair of breakpoints from two different chromosomes fused together. These fused breakpoint pairs are our searching targets. Before introducing our procedure, here we describe types of basic inter-chromosomal structural variations (fig.2.3).

Duplication

A sequence of DNA copying and inserting itself to a new position is called duplication. This 'copy-and-paste' event might happen on two homologous or nonhomologous chromosomes.

Table 2.1: Sample status and genome sequencing summary.

ID	A	B	C	D	E	F
Relation	Mother	Daughter	Daughter	Daughter	Daughter	Son
Cancer	Yes	Yes	Yes	Yes	Yes	No
Platform	3+	3	3	3+	3+	3+
Pairs	785M	519M	645M	841M	790M	857M
Insert $\leq$ 3K	59.2%	58.6%	64.5%	59.7%	58.6%	63.1%
Insert $>$ 3K	0.6%	0.6%	0.6%	0.5%	0.6%	0.5%
Inter chr	9.8%	9.4%	9.1%	8.1%	9.5%	7.2%
One read NA	18.8%	20.7%	16.6%	18.2%	18.6%	17.9%
Two read NA	11.5%	10.6%	9.1%	13.5%	12.7%	11.3%
Coverage	19.8x	13.1x	17.0x	20.8x	19.6x	21.8x

An inter-chromosomal duplication involves two fused breakpoint pairs from two nonhomologous chromosomes and an insertion on the inserted chromosome. If the duplicated sequence inserts to the now position inversely, then we call it an inverse duplication.

Transposition

A sequence of DNA transposing itself to a new position is called transposition. This 'cut-and-paste' event might happen on two homologous or nonhomologous chromosomes. An inter-chromosomal transposition involves two fused breakpoint pairs from two nonhomologous chromosomes, one insertion, and one deletion. If the transposed sequence inserts to the now position inversely, then we call it an inverse transposition.

Translocation

There are two main types of translocation, Robertsonian and reciprocal. The Robertsonian involves two acrocentric chromosomes that fuse near the centromere region with loss of the

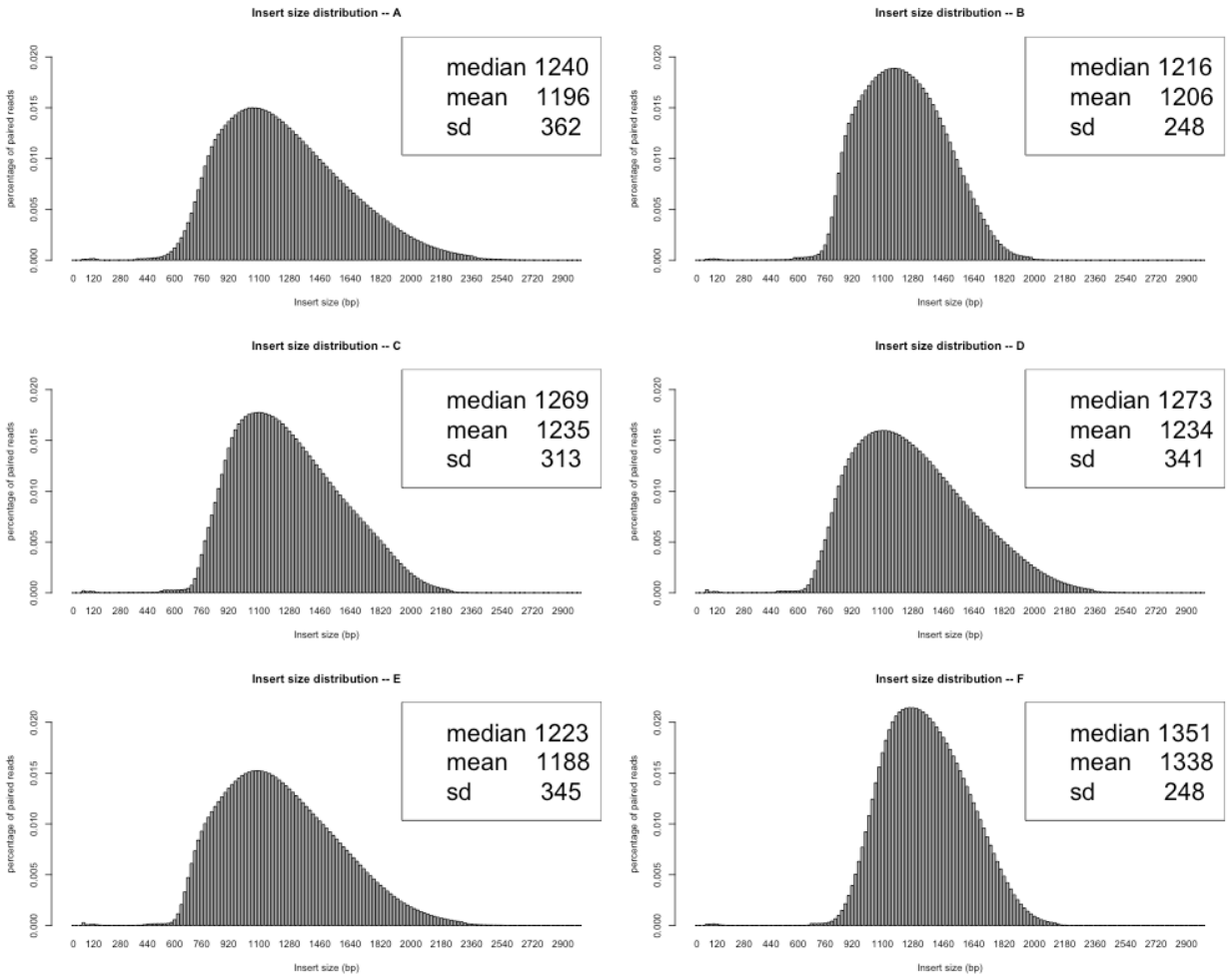

Figure 2.2: Insert size distributions (truncated at 3,000).

short arms. Since one short arm is lost, there is only one fused breakpoint pair from two nonhomologous chromosomes. Exchanging genetic materials on two chromosomes is called reciprocal translocation. It involves two fused breakpoint pairs.

If an inter-chromosomal structural variation has no extra or missing genetic materials, it is called balanced. Otherwise, it is called unbalanced. All these events could be balanced or unbalanced. The analysis of unbalanced structural variations will be more complicated. Our strategy is first identifying the putative region containing a fused breakpoint pair then using our integrated visualization method to classified the event into different types. Finally, we do the sequential analyses based on the information provided from the plot.

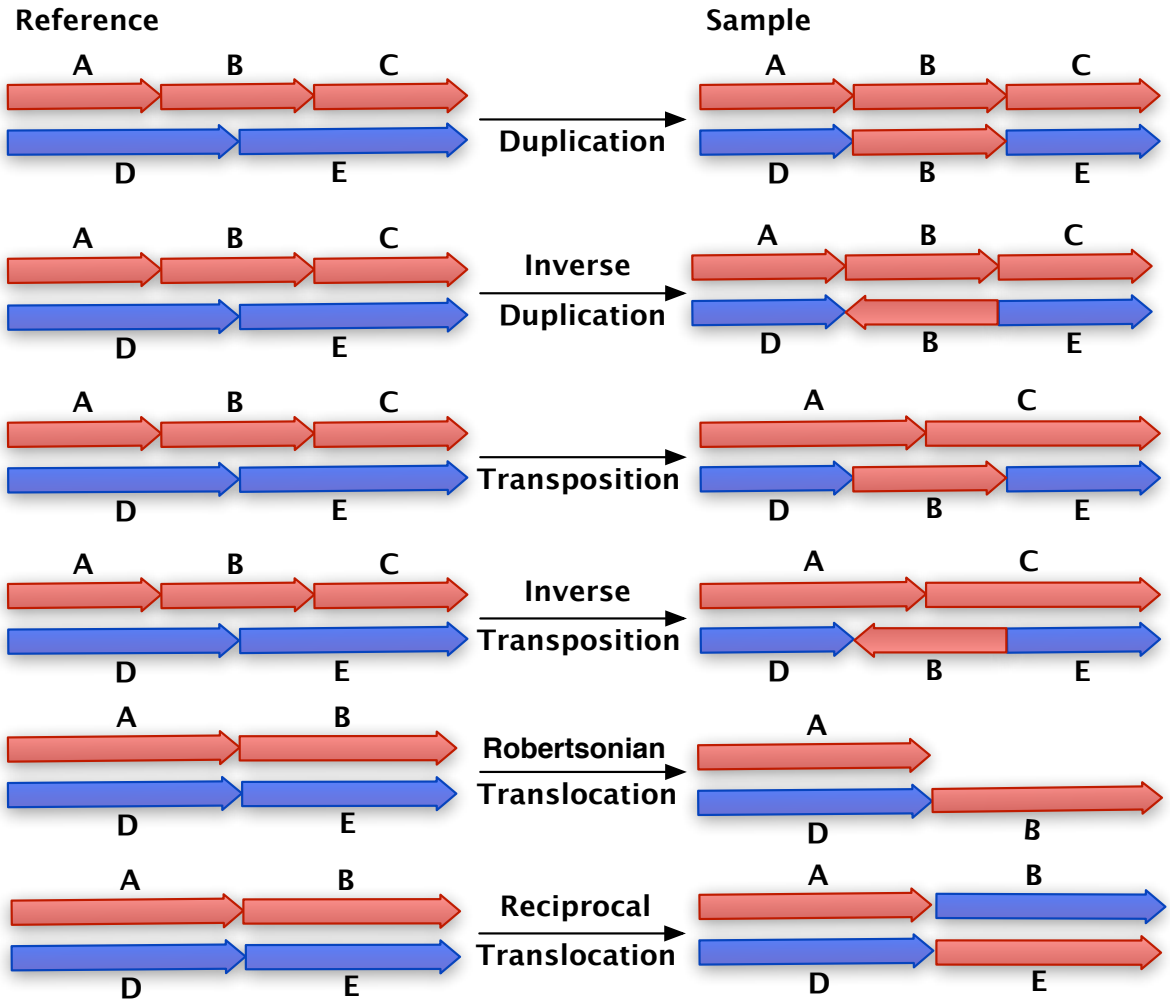

Figure 2.3: An illustration of duplication, transposition and translocation.

2.3.1 Putative Regions Searching

Since all the inter-chromosomal structural variation must involve at least one fused breakpoint pair, the mate-pair reads sampled across this fused breakpoint pair can be used to reveal the structural variation. Sindi *et al.*[46] defined the breakpoint region for a given invalid mate-pair on the 2D ESP plot. For a given invalid mate-pair C, the breakpoints region $B(C)$ is defined to be the set of pairs that can make the given invalid pair valid by admitting any one of them as a fused breakpoint pair. Based on the target insert size range (l, L) and

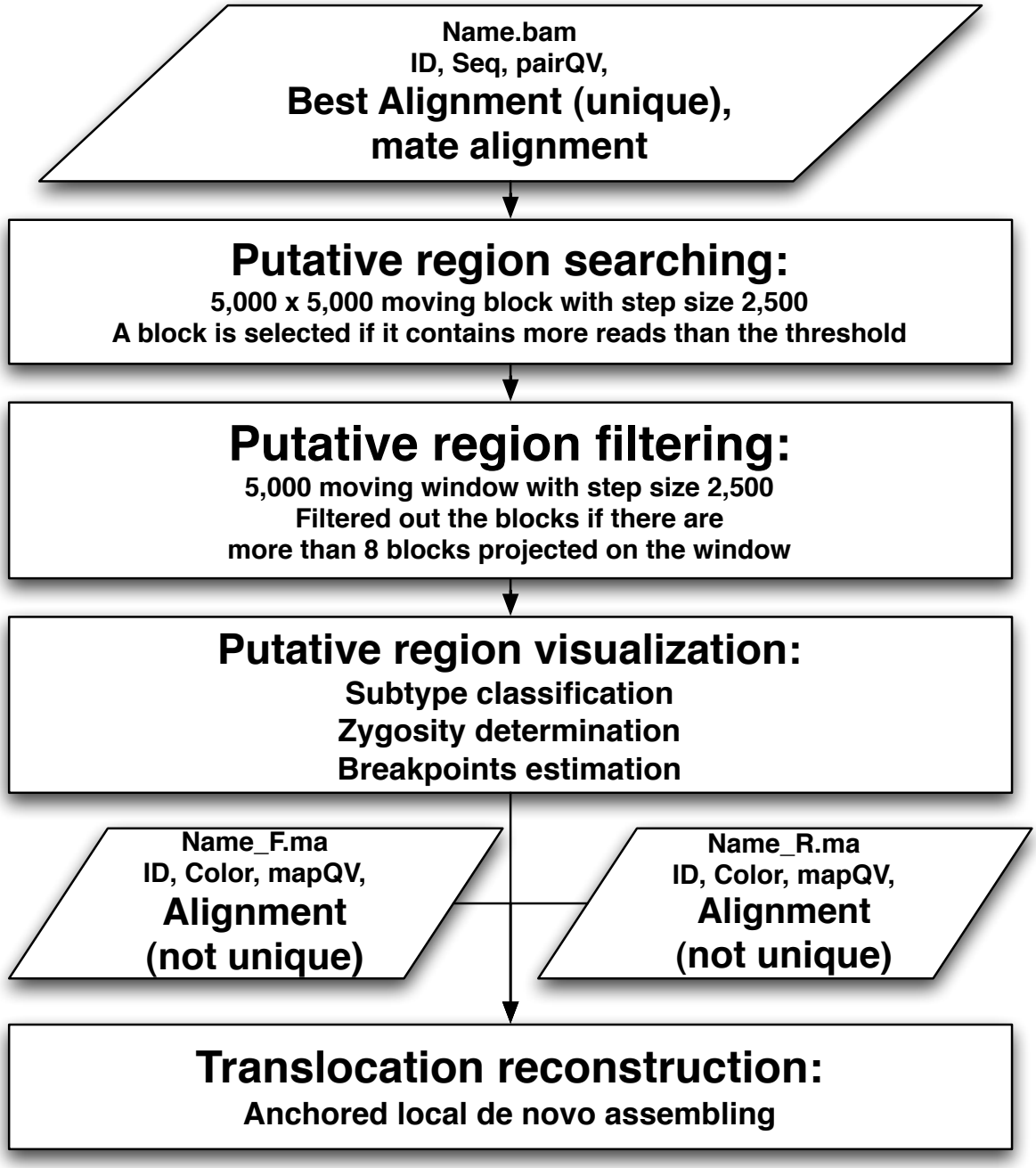

Figure 2.4: Analysis pipeline.

the alignments of an in valid pair, we have this geometric expression of the breakpoint region

$$B(C) = \{(a,b)|l \leq [strand(x_C)a - x_C] + [strand(y_C)b - y_C] \leq L\}.$$

This set forms a trapezium in the ESP plot. They used a plane sweep algorithm to find the maximal intersections of breakpoints. However, strands and order of the way that two chromosomes fused at the breakpoint pair should also be considered. The intersection of trapeziums with different orientations is not suitable.

By exchanging the roles of breakpoint pairs and invalid pairs, we can define the supporting region $S(T)$ of a given breakpoint pair T. The supporting region of a given breakpoint pair T is defined as the set of pairs that an invalid pair aligned to any of them with suitable strands and order could be valid by admitting T as a breakpoint pair. To derive the geometric expression of $S(T)$, first we define $order(T) = +1$ if the segment with smaller position is fused before the segment with larger position, otherwise $order(T) = -1$. Second we define $strand(x_T) = +1$ if the segment with smaller position is fused in positive strand, otherwise $strand(x_T) = -1$. The $strand(y_T)$ is defined in the same way for the segment with larger position. Then we have the geometric expression

$$S(T) = \{(a,b)|l \leq [order(T)strand(y_T)b - y_T] - [order(T)strand(x_T)b - x_T] \leq L\}.$$

The relationships between a breakpoint pair and its supporting invalid mate-pairs are illustrated in figure 2.5.

From the previous derivation; the supporting region of a fused breakpoint pairs can be expressed as a trapezium, and the size of the trapezium only depends on the maximum valid insert size (L). Base on this observation, we use a $2L \times 2L$ block sequentially moving with step size L for each direction to screen the $G \times G$ space. This blocking approach guarantees that any supporting region of a fused breakpoint pair must be completely contained in at least one block. Thus, a block is selected as a putative region if it contains more invalid reads than a threshold.

The threshold depends on the coverage, insert size distribution and the transposed length. Ideally, if we have enough coverage, we could observe invalid pairs filling the whole supporting

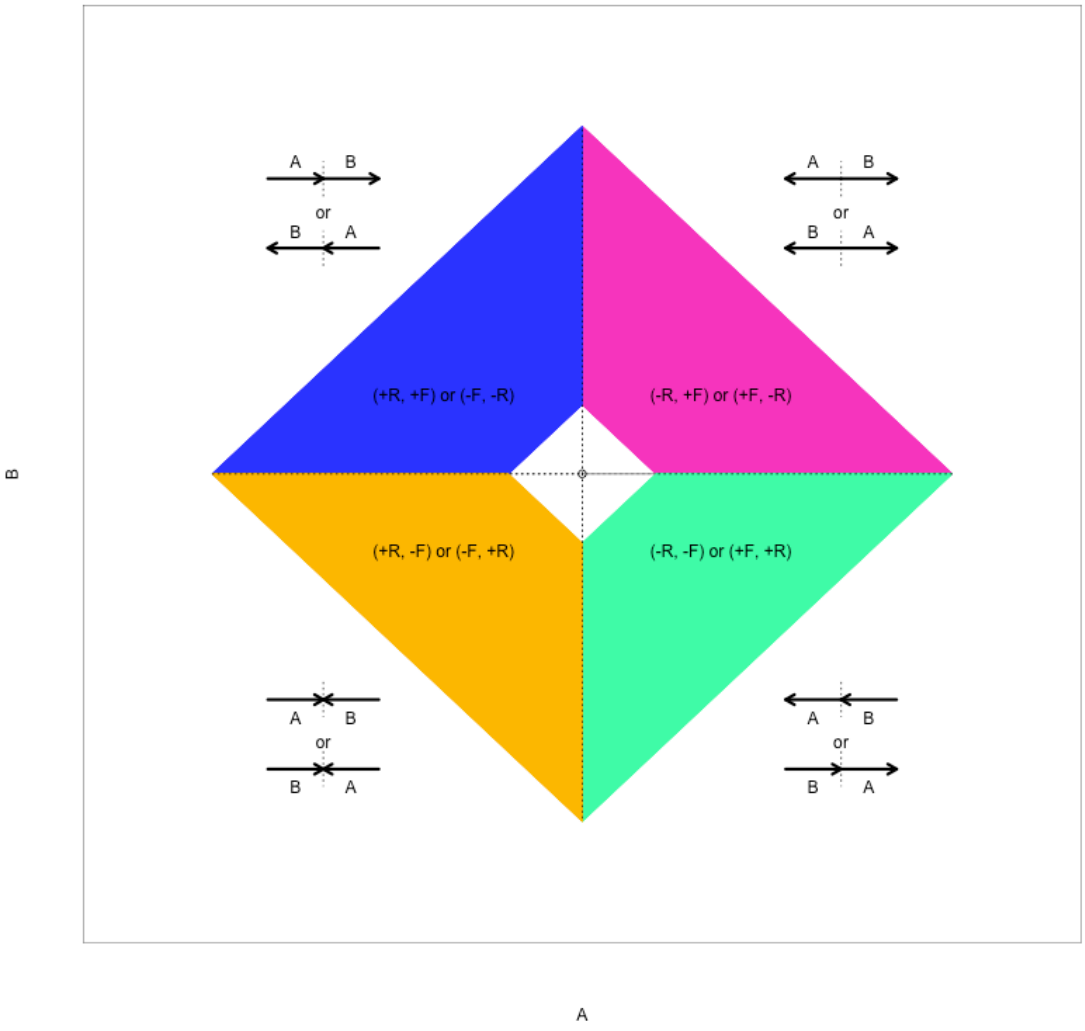

Figure 2.5: The supporting regions of a given breakpoint pair with different ways to fuse.

region of any breakpoint pair. However, we usually do not observe complete trapeziums since structural variation containing multiple breakpoint pairs might put constraints on each other's supporting region. For example, if the length of duplication or transposition is smaller than L, then the supporting regions have more constraints. If the length is too small, then the supporting region will be really small, and we won't observe enough invalid pairs supporting the breakpoint pair. To decide the threshold, we use hypothesis testing for each block. Our the null hypothesis (H_0^B) for a block B is that it fully contains at least a supporting region of a breakpoint pairs of a translocation or duplication/transposition

with length greater than 400 bps. H_0^B is rejected (block B is not selected) if N^B, the number of invalid pairs observed in block B, is smaller than the threshold K. Under the null hypothesis we have $N^B = N_T^B + N_E^B$ where N_T^B is the number of invalid pairs samples across the breakpoint pair within the block and N_E^B is the number of invalid pairs within block B resulting from experimental errors. We assume that N_E^B is negligible comparing to N_T^B. Then we set the threshold (K) satisfying

$$Pr(N_T^B \leq K | H_0^B) < 0.05 \text{ and } Pr(N_T^B \leq K+1 | H_0^B) \geq 0.05.$$

Under the null hypothesis, N_T^B follows a Poisson model with

$$\lambda_T = \frac{N[\sum_{i=1}^{400} Pr(\text{insert size} > i)]}{2G}$$

where N is the total number of pairs, G the length of whole genome. One can use the empirical insert size distribution as the probability in the equation. The two in the denominator stands for two homologous copies of each chromosome.

2.3.2 Putative Regions Filtering

We assume that every $2L$ window could only be involved in at most two inter-chromosomal structural variations (one for each homologous copy). Since each supporting region could only overlap four blocks, we used a $2L$ window moving with step size L to screen the whole genome. Blocks were filtered out if there are more than eight blocks projected to the same window. After the filtering step, we combined the overlapping blocks into super-blocks. These super-blocks were reported as putative regions for the sequential analyses. Each putative region suggests that there is at least one breakpoint in the segment (A_{left}, A_{right}) on chromosome A and one breakpoint in the segment (B_{left}, B_{right}) on chromosome B fused together. Here we follow the original ESP plot and set $A < B$.

2.3.3 Putative Region Visualization

For the putative regions passed the filter, we visualize them using our integrated ESP plot. This integrated ESP plot contains three subplots from the whole 2D ESP plot. In the top

left panel, an inter-chromosome ESP plot of the putative region is given. The horizontal axis stands for the position of chromosome A from A_{left} to A_{right} and the vertical axis stands for the position on chromosome B from B_{left} to B_{right}. All aligned mate-pairs having $x_{start} \in (A_{left}, A_{right})$ and $y_{start} \in (B_{left}, B_{right})$ are drawn as an arrow in this panel. The arrow connects (x_{start}, y_{start}) and (x_{end}, y_{end}) and the direction is decided by $(strand(x), strand(y))$. Two intra-chromosome ESP plots are also given.

The aligned mate-pair having $x_{start} \in (A_{left}, A_{right})$ and $y_{start} \in (A_{left}, A_{right})$ is drawn in the bottom left penal. The aligned mate-pair with $x_{start} \in (B_{left}, B_{right})$ and $y_{start} \in (B_{left}, B_{right})$ are drawn in the top right penal. Some summary information is given in the bottom right panel including the range of the putative region and the total number of pairs in each panel.

Here we introduce all the features in this integrated plot:

i. *Strand and order:* We have shown that different invalid mate-pairs having different strands and order will have different breakpoint regions. The relationship is given in figure 2.5. In the inter-chromosome ESP plot, we use four colors to represent four different invalid mate-pair groups. For example, a group of invalid mate-pairs colored with blue suggest a breakpoint $(\max(x_{end}), \min(y_{start}))$. In the intra-chromosome ESP plot, we only use two colors to represent whether two reads of a mate-pair have common strands or different strands, where green stands for common strands and red stands for different strands. The mate-pairs in red are invalid.

ii. *Insert size:* Some inter-chromosomal structural variations like duplication/transposition also involve structural variations like insertion/deletion on one or two chromosome. The information should be displayed on the intra-chromosome ESP plot. We draw the $y = x + mean(\text{insert size})$ in solid line and $y = x + L$ and $y = x + l$ in two dash lines. Therefore, the mate-pairs aligned outside the two dash lines are invalid which can be used to refine the breakpoint estimate for some cases. To illustrate an insertion or deletion more clearly, we add two more curves on the intra-chromosome ESP plot. The blue curve represents the conditional mean insert size given the smaller reads of

a mate-pair, that is $f_b(x_0) = E[y - x | x \in (x_0, x_0 + 2h)] - h$ where h is a bandwidth parameter. The red curve represents the conditional mean insert size given the larger reads, that is $f_r(y_0) = E[y - x | y \in (y_0, y_0 - 2h)] + h$. If the blue and red curves depart from $y = x + mean$(insert size), there might be a putative insertion/deletion.

iii. *Depth of coverage:* Depth of coverage is an important feature for breakpoint estimation and zygosity determination. However, instead of considering the depth of coverage for all reads, considering the depth of coverage for reads having a smaller position of mate-pairs and the depth of coverage for larger reads separately will give more precise information. We draw the depth of coverage for larger reads and smaller reads separately for each ESP plot.

iv. *Reference genome information:* Reference genome information is drowned on the inter-chromosome ESP plot. We use gray segments to represent the repeated region on the reference genome. Pink segments represent gene regions, and the brown arrows on the pink segments represents the exon regions.

To show how to use this integrated ESP plot to analyze inter-chromosomal structural variation, we simulated the four different types of inter-chromosomal structural variations with several parameters including homozygous/heterozygous, balanced/unbalanced, length and orientation of duplication/transposition. Here we only show three examples and give a brief explanation.

- *Case 1 - Homozygous reciprocal translocation:* In the inter-chromosome plot, the blue group having strictly smaller x and larger y than the cyan group suggest that is a reciprocal translocation. The two trapeziums can be used to estimate the two fused breakpoint pairs. The depth of coverage in the intra-chromosome plot also can be used to estimate the breakpoint. Since there are gaps with no pairs in the intra-chromosome plot, we can easily determine it as a homozygous event.

- *Case 2 - Homozygous duplication with length 1,500 bps:* The colors and the positions of the two groups in the inter-chromosome plot suggest it is a duplication/transposition

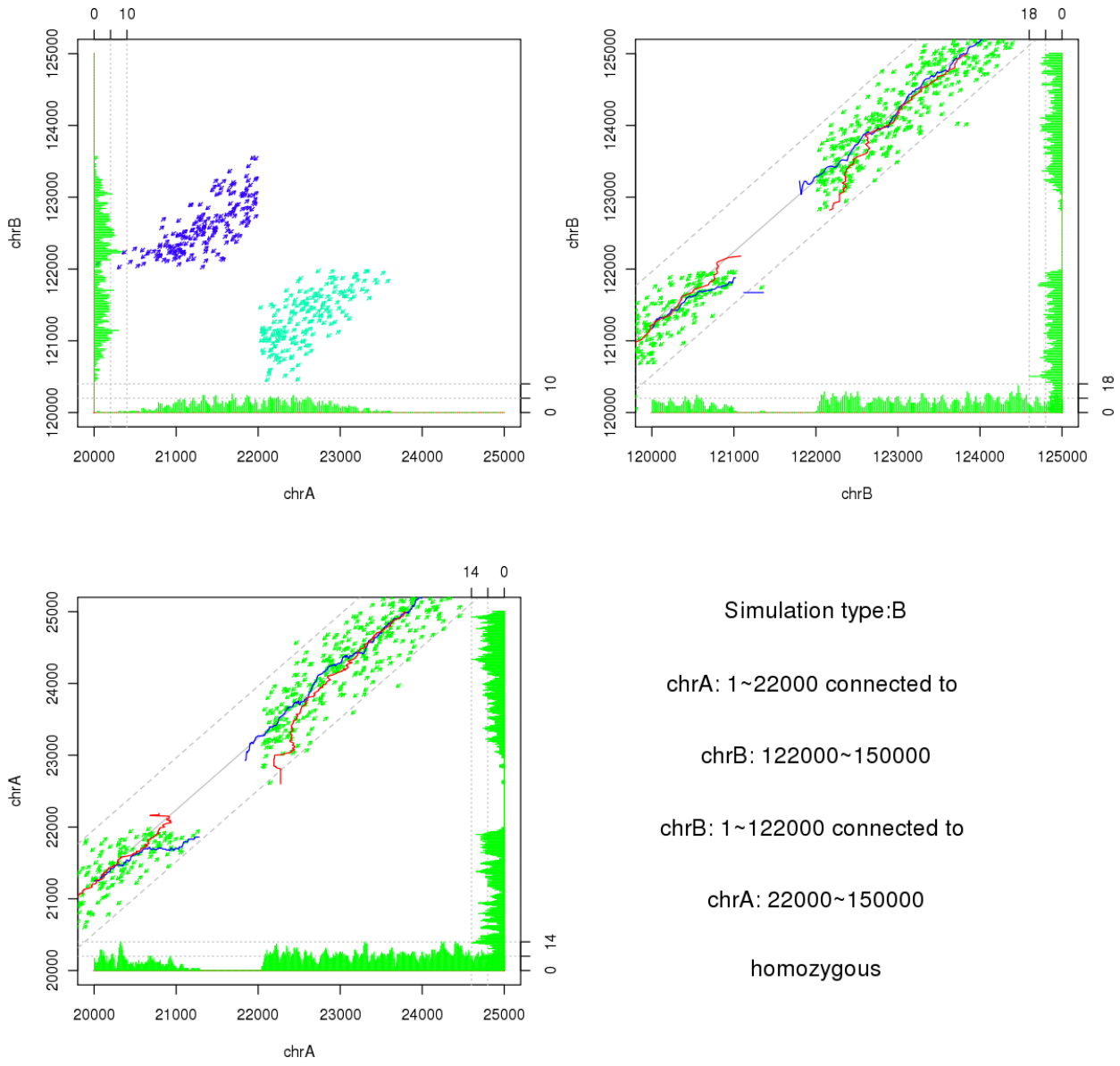

Figure 2.6: An integrated ESP plot of a simulated homozygous reciprocal translocation.

from chromosome A to B. The two trapeziums can be used to estimate the two fused breakpoint pairs. The gap in the intra-chromosome ESP plot for chromosome B suggests a homozygous insertion. The chromosome A plot suggests there is no deletion (not transposition). The depth of coverage in the intra-chromosome ESP plot also can be used to estimate the breakpoints.

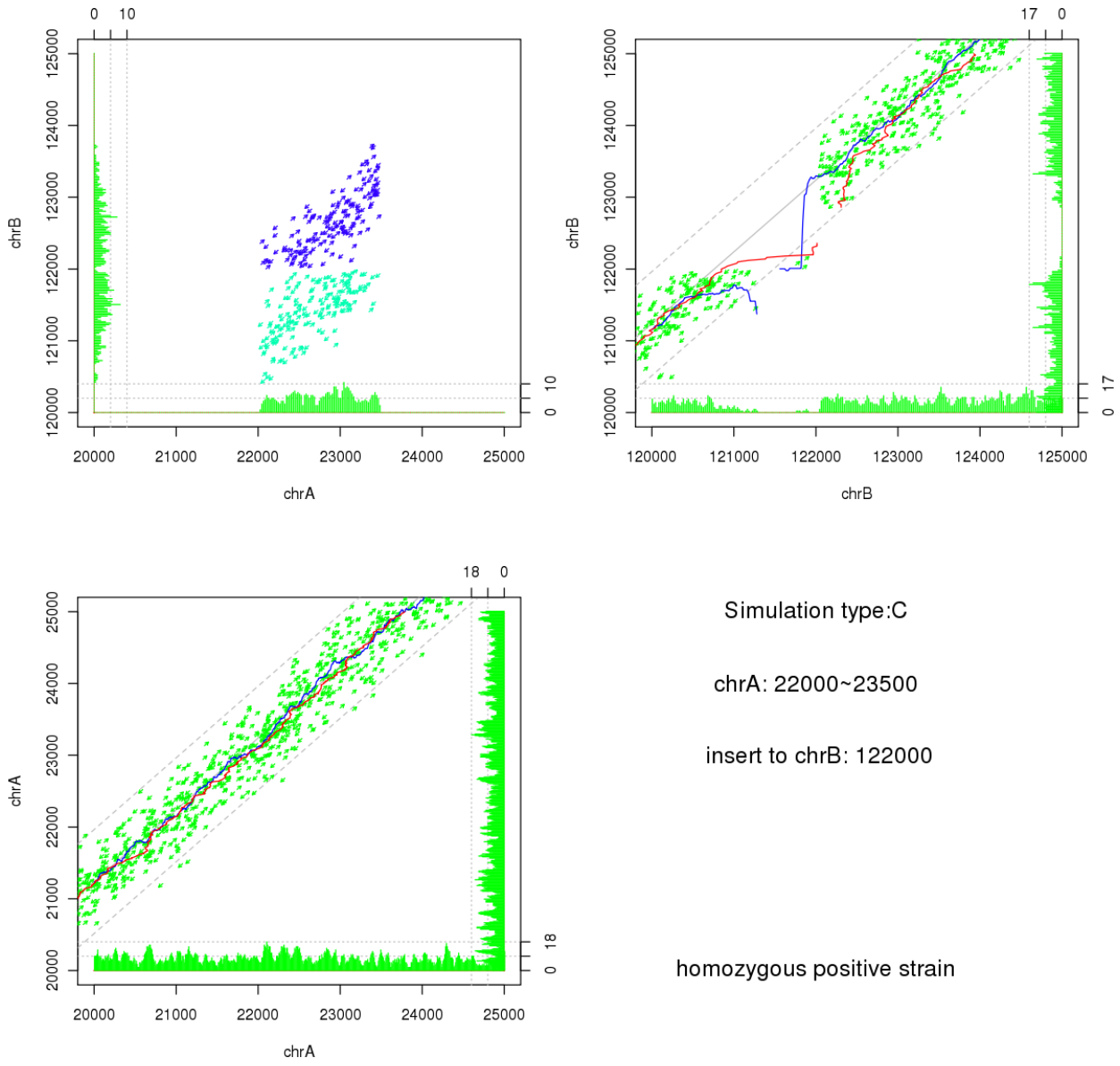

Figure 2.7: An integrated ESP plot of a simulated homozygous duplication with length 1,500 bps.

- *Case 3 - Heterozygous inverse duplication with length 500 bps:* The colors and the positions of the two groups in the inter-chromosome plot suggest it is an inverse duplication/transposition from chromosome A to B. The two trapeziums in the inter-chromosome ESP plot can only be used to estimate the position on chromosome A of the two fused breakpoint pairs. A part the blue line in the intra-chromosome plot for chromosome B suggests an insertion. The invalid meta-pair having too small insert size

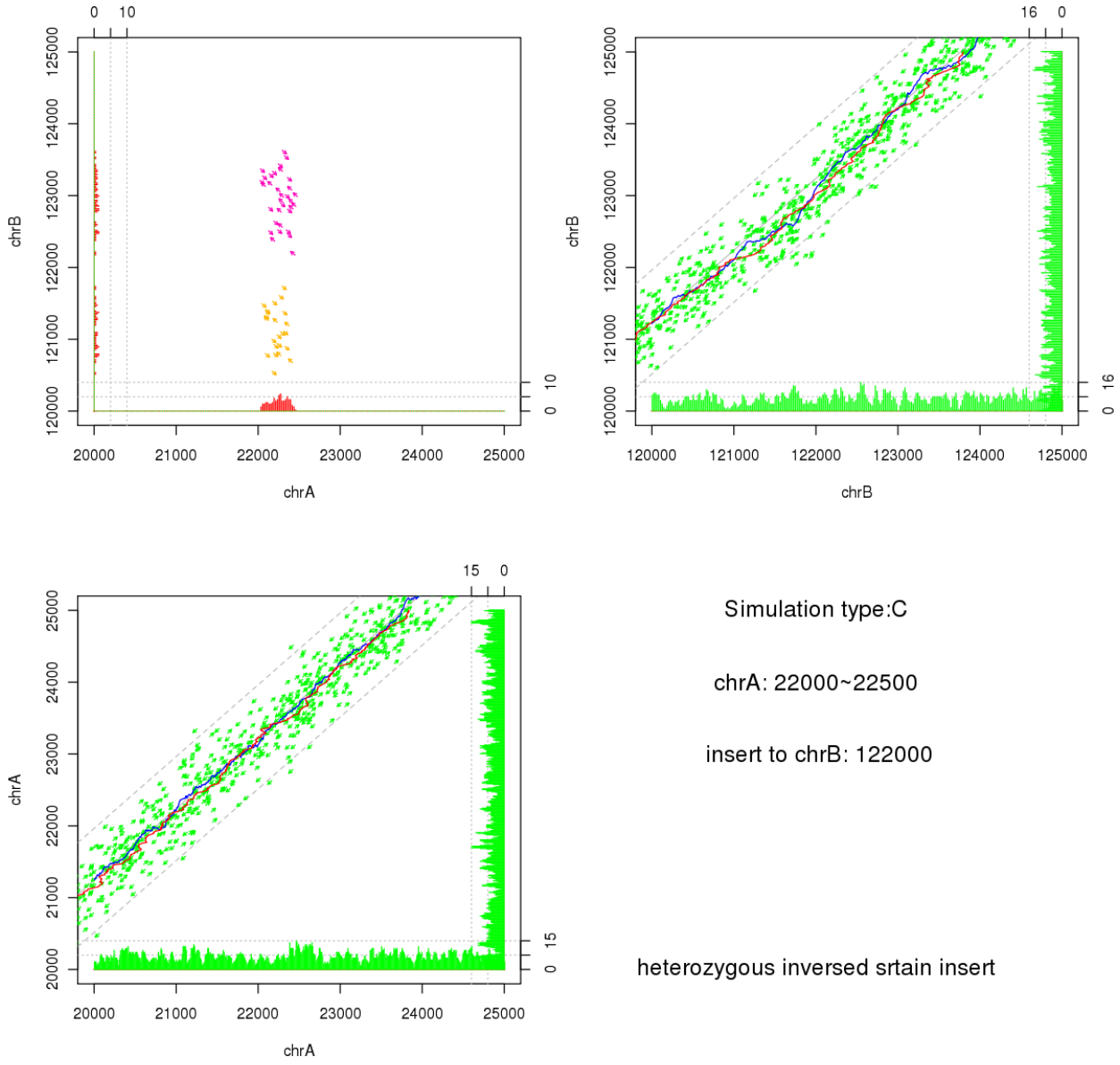

Figure 2.8: An integrated ESP plot of a simulated heterozygous inverse duplication with length 500 bps.

could be used to refine the breakpoint pair estimation. The intra-chromosome ESP plot for chromosome A suggests there is no deletion (not transposition). The depth of coverage in the intra-chromosome ESP plot also can be used to determine the zygosity.

2.3.4 Breakpoint Estimation

Usually, a fused breakpoint pairs can be estimated by the boundary of the trapezium region observed in the inter-chromosome ESP plot. The two edges parallel to the vertical and horizontal axis provide an estimate of the fused breakpoint pairs. Comparing to the estimation method in Sindi *et al.*[46], our estimate must be contained by their maximal intersections region. Furthermore, our estimation is the most conservative estimate within their region.

For a duplication or transposition with length smaller then 1,000 bps, one of the two edges parallel to the vertical and horizontal axis of the trapezium might not be observed. Therefore, the boundaries only provide an interval estimation $(\xi_{left}, \xi_{right})$ for the breakpoints on the inserted chromosome. In this case, we can use the invalid mate-pair supporting the insertion on the inserted chromosome to refine the estimates. The smaller breakpoint on the inserted chromosome can be estimated by $\min\{y | y \in (\xi_{left}, \xi_{right})$ and $y - x < l\}$ and the larger breakpoint can be estimated by $\max\{x | x \in (\xi_{left}, \xi_{right})$ and $y - x < l\}$.

2.3.5 Zygosity Determination

After we estimated the fused breakpoint pairs, we can determine the zygosity by the ratio of local coverage of other inter-chromosome pairs and the valid pairs. For example, if we observed a transposition event from chromosome A to chromosome B and the breakpoint estimates are (ξ_A, ξ_B), (η_A, η_B) where $\xi_A < \eta_A$ and $\xi_B < \eta_B$. Then we can use the ratio of the number of meta-pairs in $(\xi_A, \eta_A) \times (\xi_B, \eta_B)$ and the number of meta-pairs in $(\xi_A - (\eta_B - \xi_B), \xi_A) \times (\xi_A - (\eta_B - \xi_B) + L, \xi_A + L)$ to determine the zygosity. If the ratio is smaller then 0.75, we claim that the event is heterozygous. If the ratio is larger than 0.75 we claim that the event is homozygous.

2.4 Results

We identified 52,000 to 78,000 of blocks in the putative regions searching step across the samples. After the putative region filtering step, around 250 super-blocks were kept. Among

them, there are 11 super-blocks that found in all cancer samples but not in the normal samples. We performed our integrated ESP plot for these 11 regions and identified 2 duplications. These 2 duplications were considered to be the common somatic SVs in the samples from the multiplex family. Both duplications The integrated ESP plot for the two duplications and the estimated breakpoint pairs are given.

Table 2.2: Number of blocks selected in step 1 and 2 for each sample.

ID	A	B	C	D	E	F
Step1	58454	58148	78002	53971	52686	76153
Step2	220	246	259	224	238	260

Table 2.3: Summary of the estimated breakpoint pairs.

A.chr	A.pos	B.chr	B.pos
chr a - q24.2	XXXX933	chr b - p14.2	XXXX437
chr a - q24.2	XXXX192	chr b - p14.2	XXXX475
chr c - p14.3	XXXX637	chr d - q35	XXXX107
chr c - p14.3	XXXX657	chr d - q35	XXXX643

2.5 Discussion

Here we addressed some issues about the current strategy. First, we used an algorithm to search the fused breakpoint pairs and visualize them locally. Their correlations between these fused breakpoint pairs were ignored. To reconstruct the global genome structure, the correlations between the breakpoints should be taken seriously.

Based on the knowledge of known rearrangement mechanisms, we can identify the types of inter-chromosomal structural variations, such as duplication, transposition, and transloca-

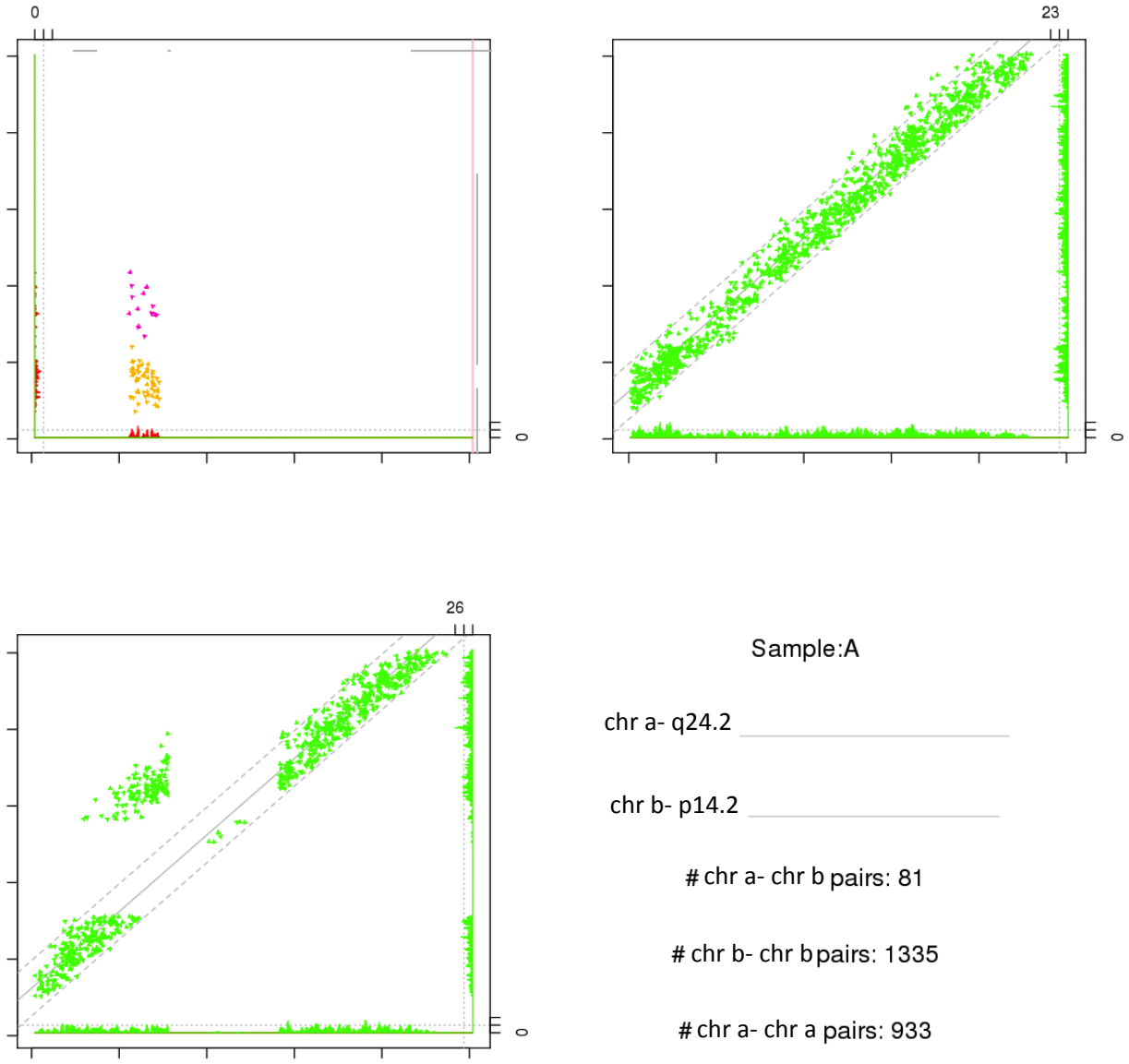

Figure 2.9: A heterozygous inverse duplication from chromosome a to chromosome b.

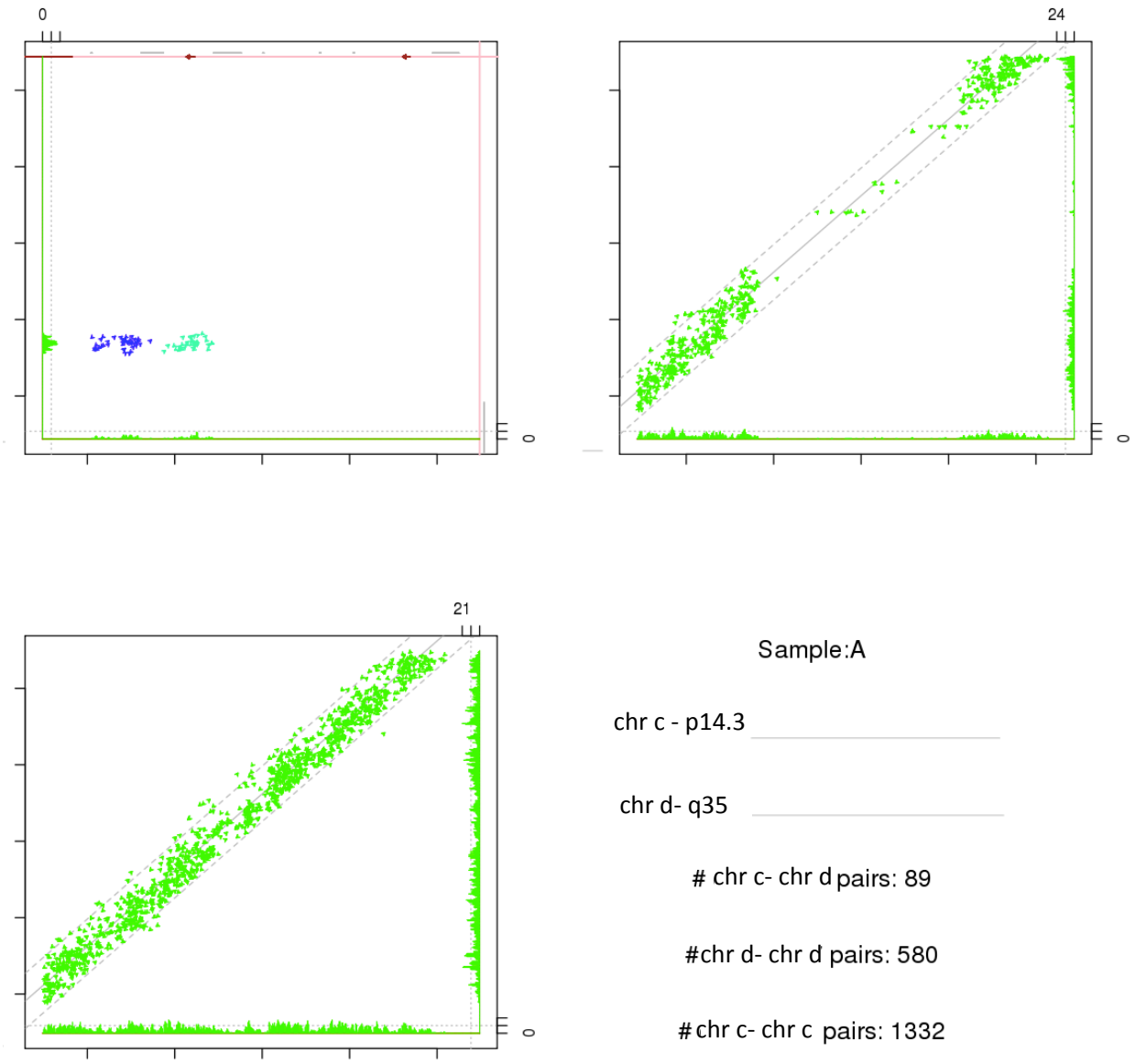

Figure 2.10: A heterozygous duplication from chromosome d to chromosome c.

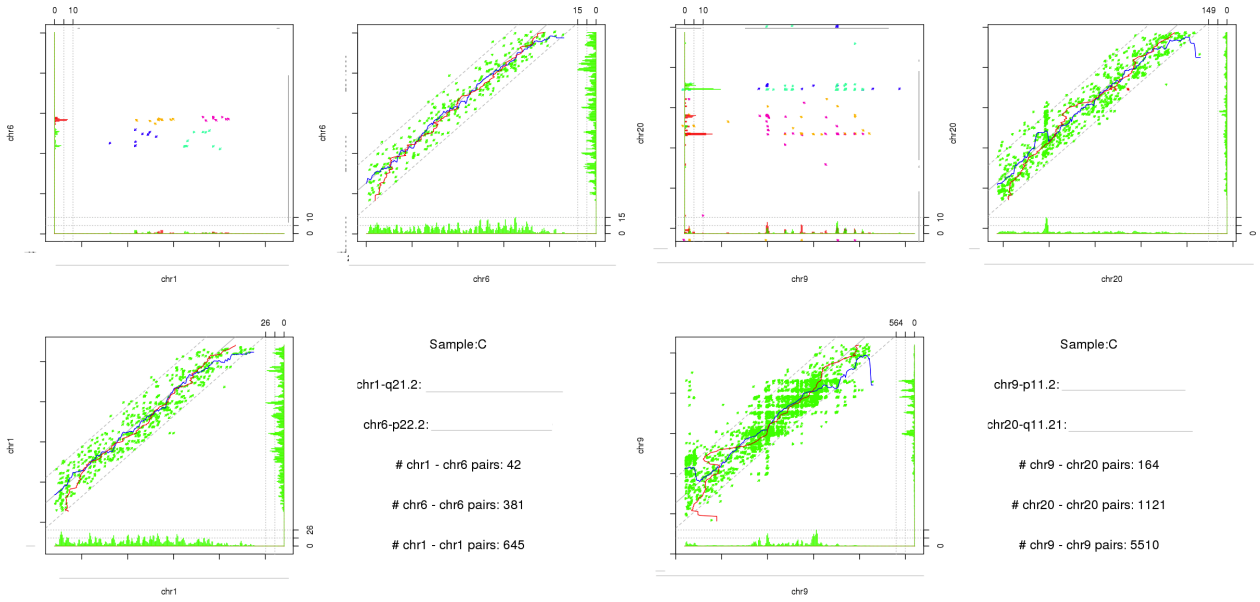

Figure 2.11: Two examples of putative regions containing patterns that cannot be explained by a single structural variation.

tion. However, the real data is often more complicated. There were several putative regions containing patterns that we cannot explain. Two examples were given in figure 2.11. The first plot suggests there might be a duplication from chr6 to chr1. Nevertheless, both positive and inverse strands were found in the duplication. The second case cannot be explained by a single event. These patterns were widely observed at the regions near the centromere which is known as the "hot spots" for transposons.

The assumption that every $2L$ window could only be involved in at most two inter-chromosomal structural variations (one for each homologous copy) for the putative region filtering step might fail in real data analysis. Transposons are known as mobile elements in the genome. They copy and insert themselves to different positions. The source positions may be involved in more than two inter-chromosomal structural variations.

CHAPTER 3

Landscape of Somatic Mutations in Taiwanese Lung Adenocarcinomas

3.1 Introduction

Lung adenocarcinoma (LUAD) is a heterogeneous disease and has a high mutation rate compared to other cancers[30]. With the development of second-generation sequencing technology, several genome-wide studies[10][14][21] had identified frequently mutated driver genes and tumor suppressor genes in LUAD. The study by The Cancer Genome Atlas[10] (TCGA) reported that 75% of the tumors harbored gene abnormality in at least one of 13 driver genes. Specifically, 67% of the tumors carried somatic DNA alterations in KRAS, EGFR, NF1, BRAF, ERBB2, MET, RIT1, HRAS, NRAS, and MAP2K1. Their finding enlarged the potential targetable alterations in MAPK signaling pathway.

In Asian, LUAD has a higher rate of nonsmokers than that for Caucasian. A recent study of Asian LUAD[35] indicated that frequencies of driver mutations differ between the East and the West populations. EGFR mutations were most common in the East, with the rate even higher than KRAS mutations rate in the West. However, genomic data in Asian LUAD is still limited, and many studies only focus on the mutations in the known driver genes. The characterization of mutations beyond the known driver genes in Asian LUAD remains unclear.

Rather than identifying frequently mutated genes, a recent study proposed a new approach to detect significantly mutated residues and defined them as mutational hotspots[3]. This approach was motivated by the following two viewpoints: (i) not all mutations in the

driver genes are driver mutations and (ii) different mutations in the same gene may have different functional impacts. However, for each significantly mutated residue, not all the possible nucleotide variants are observed. Thirty-five percent of hotspots only have one single amino acid variants. Also, different nucleotide variants in the same residue may have different functional impacts. To pin down mutational hotspots to single nucleotide variants, we derived a binomial probability model to evaluate the statistical significance of recurrent point mutations.

In this chapter, we presented a comprehensive and multi-platform analysis on 113 untreated stage I Taiwanese lung adenocarcinomas with attention to potential clinically relevant recurrent point mutations, focal copy number alterations, and the potential downstream molecular changes to further our understanding of lung adenocarcinoma pathobiology beyond the known driver mutations.

3.2 Materials and Methods

3.2.1 Samples and Clinical Data

We analyzed materials from 113 untreated Taiwanese stage I lung adenocarcinomas (LUAD) patients. The fresh frozen primary tumor and adjacent normal tissue samples were collected from 2000 to 2011. The median follow-up was 55.9 months; 51 patients relapsed and 35 died during follow-up. Among the 49 female patients, 98% of them reported never smoking. In contrast, among the 64 male patients, 61% of them reported past or present smoking. Known EGFR and KRAS mutations were examined by nucleotide mass spectrometry. Table 3.1 summarizes demographic characteristics of the cohort.

Table 3.1: Patient characteristics

Characteristics	Patients
Age	year
Median (range)	66 (37-83)
Gender	n (%)
Male	64 (57)
Female	49 (43)
Smoking	n (%)
Ever	40 (35)
Never	73 (65)
Stage	n (%)
IA	62 (55)
IB	51 (45)
EGFR activating mutations	n (%)
Wild type	47 (42)
L858R	35 (31)
Exon19 deletion	31 (27)
Vital and relapse status	n (%)
Dead	35 (31)
Alive and relapsed	17 (14)
Relapse free	62 (55)
Survival	Month
Median months to death (range)	55.9 (0.2-162.4)
Median months to relapse (range)	42.4 (0.2-122.8)

3.2.2 DNA Mutational Analysis

We performed whole exome sequencing (WXS) on paired tumor and adjacent normal samples from all 113 patients in the cohort and also on additional buffy coat samples from 36 patients among them. Whole exome capture was performed using Nextera Rapid Capture Exome and Expanded Exome Kits. Paired-end sequencing was performed using Illumina HiSeq system. Reads were aligned to reference genome hg19 (GRCh37) using BWA[32] and processed by Picard tools and GATK[12] for local realignment, marking duplicates and base recalibration. Somatic point mutations and indels were called by Mutect[8] and Samtools[33]; then filtered by a set of empirical filters including a blacklist derived from WXS data of the additional 36 buffy coat samples.

Somatic mutation signatures were identified by a divergence-based non-negative matrix factorization (NMF) to the mutation probability matrix of the cohort using an R package SomaticSignatures[19]. Signatures were rescaled to have sum equals one.

We obtained significantly mutated genes based on at least two identifications by the following three tools: MutSigCV[30], for identifying genes mutated more frequently than expected by chance given the background mutation processes; OncodriveFM[20], for identifying genes significantly biased toward mutations with high functional impact; OncodriveCLUST[50], for identifying genes with mutations clustered in particular regions.

For clarity, we used the term *recurrent point mutation* (RPM) for the event of mutation at a specific chromosome position with the same nucleotide substitution occurring in an excessing number of patients as expected by chance. The statistical significance and the false discovery rate of RPMs were calculated using a binomial probability model given the size of the cohort and the mutation rate across the cohort. For a point mutation detected in at least k patients in the cohort of size n, we have

$$p\text{-value} = \sum_{x=k}^{n} p^x (1-p)^{n-x} \tag{3.1}$$

where p is the probability of a patient carrying the somatic point mutation. We estimated p by the mean mutation rate across the cohort divided by 3 (using the default of equal

nucleotide substitution), that is

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} \frac{N_i}{3L} \tag{3.2}$$

where N_i is the number of somatic mutations observed in patient i, and L is the total length of the targeted region. With this formulation, RPMs were selected from point mutations detected in at least two patients in the cohort. The false discovery rate was controlled using Benjamini-Hochberg method ($q < 0.05$). In facilitating the calculation false discovery rate, the p-values for all possible point mutations not found in the cohort were set to 1.

The RPMs in EGFR (L858R) and KRAS (codon 12 and 13) were validated by MASS spectrometry, and the rest of the RPMs were validated using Sanger sequencing.

3.2.3 DNA Copy Number Alteration Analysis

We performed array CGH on paired tumor and adjacent normal samples from 111 patients in the cohort. Following the standard protocol, the gDNA was extracted from frozen cancer tissue of each sample with quality checked by agarose electrophorese. The whole genome NimbleGen CGH-array (NimbleGen; NimbleGen Systems Inc, Madison, WI) containing 385,806 probes spacing of about 6,000 base pairs was used. Tumor and adjacent normal samples were both labeled with Cy3 and compared to the reference DNA sample labeled with Cy5 using two separate arrays. The reference DNA sample was extracted from the peripheral blood mononuclear cell of one male and one female from a community cohort. We performed the normalization procedure as in our previous study[52] for each array to obtain the log2 ratio at the probe level. To remove the wave pattern, a widespread technical artifact highly correlated to GC content, we performed a two-step GC correction procedure (Supplementary Materials 3.5.4). Every 10 adjacent probes were aggregated as one probe-block by taking the average of the GC-corrected log2 ratios to reduce the noise. Somatic copy number alteration (SCNA) value at the probe-block level was defined as the difference between the log2 ratios from the tumor sample and the normal sample. Segmentation was performed on the SCNA value using the circular binary segmentation (CBS) algorithm using an R package DNAcopy[39]. GISTIC 2.0[37] was utilized to identify significant focal copy number

alterations from segmented SCNA data.

3.2.4 miRNA Array Analysis

We utilized Affy GeneChip miRNA 3.0 array to quantify the abundance of 1,726 microRNAs for 69 paired tumor and normal samples. The expression profiles were normalized using RMA method across the cohort using Affymetrix Expression Console.

3.2.5 RNA Sequencing Analysis

We performed RNA sequencing (RNAseq) on 107 tumor samples. Candidates of gene fusions were identified using Tophat-funsion[25]; only those gene fusions with two complete exons joined at their splice site in an abnormal order were selected. Due to the poor quality of the data, the RNAseq analysis is only presented in Supplementary Materials 3.5.7.

3.3 Results

3.3.1 Somatic Mutations in 113 Taiwanese Stage I Lung Adenocarcinomas

Whole exome sequencing identified 29,621 somatic mutations in 113 tumor samples. The number of somatic mutations detected in a patient ranged from 5 to 6,675 with mean 262 and median 114. The mean somatic mutation rate was 4.21 per Mb. Note that one tumor was ultra-hypermutated with 6,675 somatic mutations (107.17 per Mb). This patient is a 77-year-old male heavy smoker, one pack per day for 50 years without quitting. Concurrently, a non-synonymous mutation in DNA polymerase gene POLE was identified in this tumor. POLE is related to DNA repair, and its loss of function mutations may cause ultra-hypermutated cancers[22]. Nonetheless, analysis of mutational signatures identified that this tumor had a strong APOBEC signature with predominantly mutations C>T/G at TCW (W = A/T) instead of a POLE signature with predominantly mutations C>A at TCT and C>T at TCG.

3.3.2 Six Genes were Significantly Mutated

Combined analysis of MutSigCV[30], OncodriveFM[20] and OncodriveCLUST[50] identified 6 significantly mutated genes by at least two tools. Consistent with to a previous study of driver mutations in Asian LUAD[35], the most commonly mutated gene is EGFR (62%) including L858R (31%), exon19del (27%), L747P (1%), L861R (1%), L861Q (1%) and exon20ins (1%). Mutations in known LUAD driver genes KRAS (10%), NRAS (3%) and EGFR were found to be mutually exclusive. The tumor suppressor gene TP53 (36%) was the second commonly mutated gene. An RNA splicing gene RBM10 (10%) and a proto-oncogene CTNNB1 (5%) of colorectal cancer, pilomatrixoma, medulloblastoma, and ovarian cancer were also commonly mutated. Eleven out of 41 (27%) non-synonymous mutations in TP53 and 8 out of 11 (73%) in RBM10 were truncating mutations. Mutations in other known LUAD driver genes[10], HRAS (1%), MAP2K1 (1%), BRAF (2%), ERBB2 (2%), NF1 (4%) and RIT1 (1%) were observed but not significant. Note that all 4 non-synonymous mutations in NF1 were co-occurred with TP53 mutations. (Fig. 3.1A shows the co-mutational plot of the cohort).

3.3.3 Three Somatic Mutational Signatures were Identified where 1 Signature was Correlated with Smoking and Mutation Status in EGFR and KRAS

Somatic point mutations were aggregated into 96 types of trinucleotide patterns for each patient. We obtained 3 somatic mutation signatures by performing a divergence-based non-negative matrix factorization (NMF) to the mutation probability matrix of the cohort using an R package SomaticSignatures[19]. The signatures were rescaled to have sum equals one. Note that 11 tumors with less than 30 mutations are excluded from the mutational signature analysis. The APOBEC signature exhibited mutations of C>T/G at TCW (W=A/T) and the Tobacco signature had a predominance of transcriptional strand biased C>A mutations. The scores of these two signatures were correlated with the number of somatic mutations ($p < 0.01$, Spearman's rank-order correlation test). The score of Tobacco signature was significantly higher in smoker patients, EGFR wild type patients, and KRAS mutant patients

(Wilcoxon rank sum test $p < 0.01$, Supplementary Fig. 3.20). The third signature was dominated by mutations C>T at CpG resulting from an endogenous mutational process related to spontaneous deamination of 5-methylcytosine.

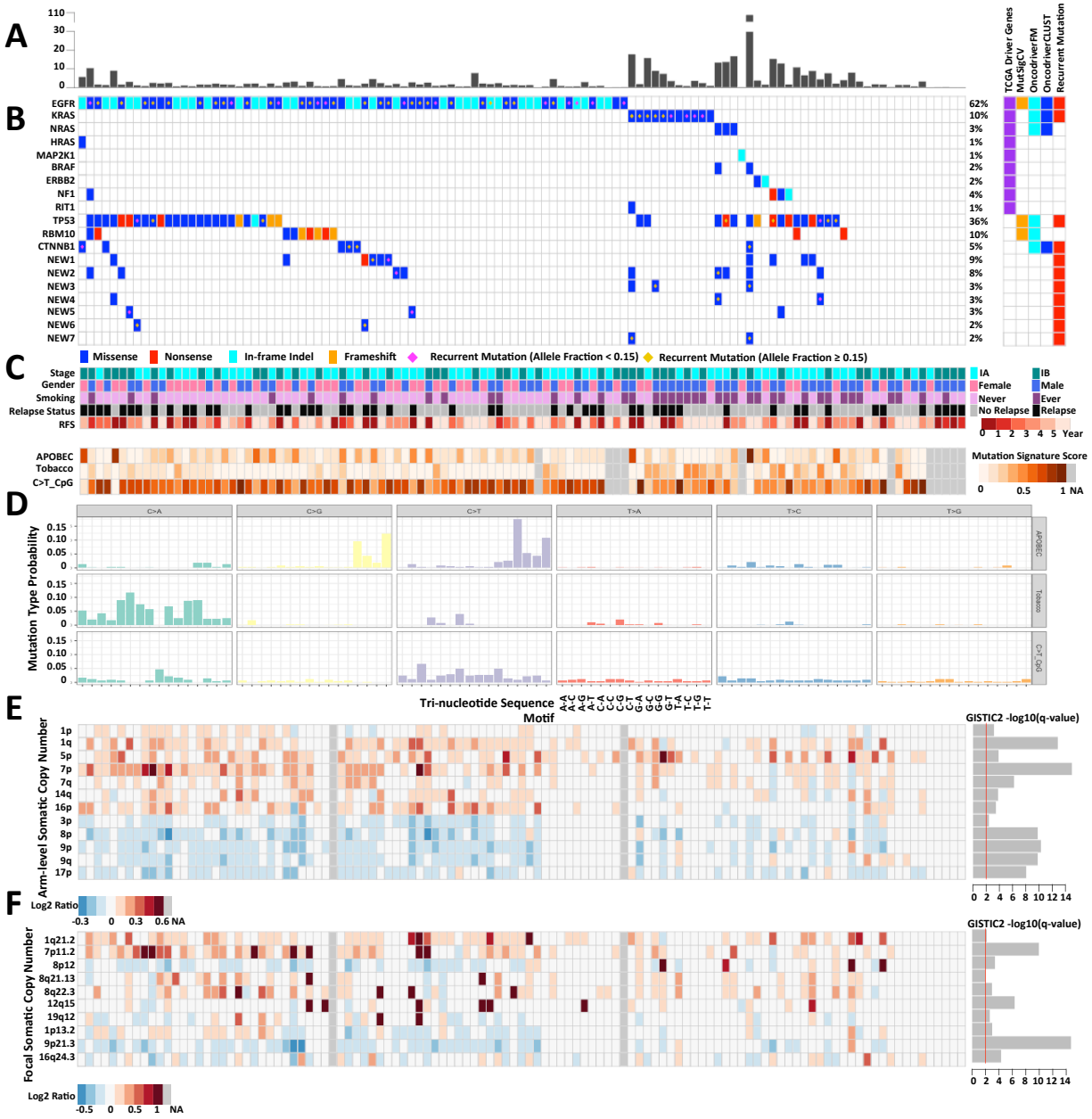

Figure 3.1: Landscape of somatic DNA alterations in 113 Taiwanese stage I lung adenocarcinomas (A) Somatic mutation rate (mutations/Mb) across the cohort. (B) Co-mutation plot of mutations in i. lung cancer driver genes from TCGA, ii. significant genes (Benjamini-Hochberg $q < 0.05$) from at least two of the following algorithms: MutSigCV, OncodriveFM and OncodriveCLUST, iii. genes harboring recurrent point mutations (detected in at least two samples in the cohort, Benjamini-Hochberg $q < 0.05$). Recurrent point mutations are marked by rhombuses. (C) Tumor stage, gender, smoking status, relapse status and relapse-free survival (RFS) across the cohort. (D) Somatic mutational signatures derived in the cohort using a divergence-based non-negative matrix factorization (NMF) to the mutation probability matrix with a constraint every signature with sum equals 1. Eleven samples with too few mutation counts (less than 30) are excluded from the mutational signature analysis. (E) Heatmap of significant arm-level somatic copy number alterations (GISTIC2, Benjamini-Hochberg $q < 0.01$) and significant focal somatic copy number alterations (GISTIC2, Benjamini-Hochberg $q < 0.01$).

3.3.4 Eighteen Recurrent Point Mutations were Identified in 11 Genes Including 4 Significantly Mutated Genes

We defined a recurrent point mutation (RPM) as a mutation at a specific chromosome position with the same nucleotide substitution occurring in an excessing number of patients. To identify the RPMs in the cohort, a binomial probability model was used to assess the statistical significance. In total, we identified 48 statistically significant RPMs occurring in at least two patients in out cohort ($q < 0.05$). Eighteen RPMs are non-synonymous, contained in 11 genes across 60 tumors. Four of the 11 genes were also detected by MutSigCV, OncodriveFM or OncodriveCLUST. The RPM c.T2573G in EGFR which induces the activating mutation p.L858R was the most common RPM (35 tumors). Another RPM c.G2248C (2 tumors) in EGFR inducing the amino acid change p.A750P was co-mutated with an exon19 deletion in both tumors. KRAS harbored three clustered RPMs c.G34T (5), c.G35A (3) and c.G37T (2) in codon 12 and 13; and CTNNB1 had two clustered RPMs c.C98G (2) and c.C110T (2) in codon 33 and 37. TP53 harbored 1 nonsense RPM c.G892T (2) and 3 missense RPMs

c.C844T (2), c.A659G (2), c.C423G (2) within exon 5 to 8. The other 7 RPMs were all detected in different genes and carried by exact two patients including 3 novel RPMs which were not reported in COSMIC database version 81.

The RPMs in EGFR (L858R) and KRAS (codon 12 and 13) were validated as somatic mutations by MASS spectrometry. Other RPMs were validated using Sanger sequencing in three rounds. All 17 out of 17 RPMs with variant allele fraction (VAF) $\geq 15\%$ were validated as somatic mutations. Only 1 out of 7 RPMs with VAF $< 15\%$ were validated as somatic mutations likely due to the limitation of Sanger sequencing (Supplementary Materials 3.5.3).

3.3.5 Somatic Copy Number Alterations in 111 Taiwanese Stage I LUAD

We performed array CGH (NimbleGen) to determine the Somatic copy number alterations (SCNAs) in 111 tumors of the cohort. GISTIC2.0 was used to identify arm level and focal SCNAs. For arm level SCNAs, similar to the previous studies[10][52], the most common amplifications were 7p followed by 1q, 7q, 5p, 14q 16p and 1p; the most common deletions were 8p, 9p, 9q, 17p and 3p ($q < 0.01$). After removing the region near centromere and telomere, we obtained 7 focal amplifications 1q21.2 (PIK4B), 7p11.2 (EGFR), 8p12 (ZNF703), 8q21.13 (ZNF704), 8q22.3 (ZNF706), 12q15 (MDM2), 19q12 (CCNE1) and 3 focal deletions 1p13.2 (NRAS), 9p21.3 (CDKN2A/CDKN2B) and 16q24.3 (CDK10, $q < 0.01$, Fig. 1). We compared the focal SCNA values between the EGFR wild type and mutant tumors. Focal amplifications 7p.11 (EGFR) and focal deletions 9p21.3 (CDKN2A/CDKN2B) was significantly more frequent in EGFR mutant tumors and 8p12 (ZNF703) was significantly more frequent in EGFR wild type tumors ($q < 0.05$, Supplementary Fig. 3.14). ZNF703 gene amplification at 8p12 was reported in previous studies[1][47] as a driver in luminal B breast cancer and was not reported in TCGA LUAD analysis.

3.3.6 Non-EGFR RPMs were Correlated with Poor Relapse Free Survival, and the Impact was Associated with Variant Allele Fraction

We investigated the clinical relevance of the significant somatic DNA alterations obtained (Fig. 3.2). Multivariate Cox regression for relapse free survival (RFS) was performed to determine the impact of the alterations adjusted for clinical characteristics including gender, age, smoking status, and stage. EGFR activating mutations is the standard patient selection criterion for the first-line TKI-inhibitor. We used EGFR activating mutation status as a stratification variable in the survival analysis to examine the differential clinical impact between the wild type (WT) and mutant (MT) groups for better patient management. In Fig. 3.2, we observed that the non-EGFR RPMs were shown to correlate with poor relapse free survival and the variant allele fraction (VAF) impacts the strength of correlation. VAF was defined as the number of variant-harboring reads divided by the depth at the specific position. It represents the relative dosage of mutant sequence in the bulk DNA from the heterogeneous cell proportions. We hypothesize that VAF can be used as a surrogate measure of impact level for the mutation to the tumor.

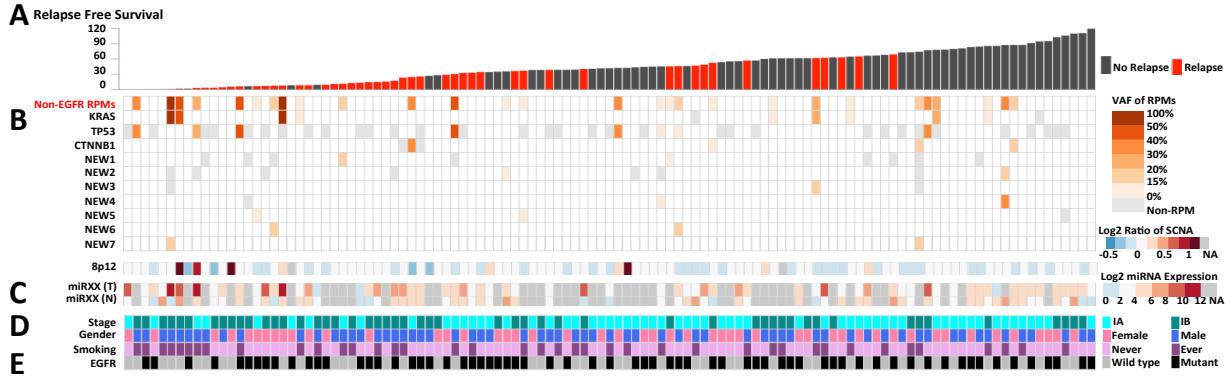

Figure 3.2: Clinical relevance of significant somatic DNA alterations. (A) Relapse free survival across the cohort. Patients were ordered by RFS in an increasing order. (B) Variant allele fraction of the recurrent point mutations. (C) Log2 value of focal somatic copy number alterations at 8p12. (D) Log2 expression of miR-XX in tumor and normal samples. (E) Tumor stage, gender, smoking and EGFR mutation status across the cohort.

The 16 non-EFGR RPMs were bagged as one variable by taking the maximum VAFs of

the non-EGFR RPMs as a continuous score (maxVAF) for each patient due to the small number of carriers for each RPM. Higher maxVAF correlated with poor RFS in the multivariate Cox proportional hazard model for the whole cohort (HR = 15.8, 95% CI = 3.1-81.1, p = 0.0009), for the subset of EGFR wild type tumors only (HR = 31.72, 95% CI = 3.5-289.2, p = 0.002), and for the subset of EGFR mutant tumors only too (HR = 62.7197, 95% CI = 1.5-2688.0, $p = 0.031$). To classify patients into different risk groups, we assigned patients with maxVAF$\geq$ 15% as the high-risk group and the rest of patients as the low-risk group. The15% cutoff was used because of the limitation of Sanger validation. Kaplan-Meier analysis with the univariate log-rank test was performed (Fig. 3.3A). In EGFR mutation group, the difference of RFS between two risk groups was not significant ($p = 0.354$), likely due to the limited number of tumors with EGFR and non-EGFR RPMs co-mutated (Fig. 3.3C-D).

This relationship can be validated in a cohort of 172 Caucasians from TCGA LUAD study. For each patient, we fixed the same 16 non-EGFR RPMs and used the maximum VAFs as a continuous score. The score was shown to correlate with poor RFS in the multivariate Cox proportional hazard model (HR = 5.28, 95% CI = 1.35-20.70, $p = 0.0171$). Kaplan-Meier analysis with univariate log-rank test examined the statistical significance of this risk (Fig. 3.3B). Note that when classifying patients into different risk groups based on the score, we raised the cutoff to 25% since we noticed that in general, TCGA data yield higher VAFs were higher for all mutations than ours (Supplementary Fig. 3.19). In the Caucasian cohort, EGFR activating mutations were not the most common driver and only found in 14 patients. We did not perform stratified survival analysis for EGFR mutation status since all non-EGFR RPMs were only found in EGFR wild type tumors. Limiting tumors to localized (M0, N0) or stage I (M0, N0, and T1-T2a), the correlations were still significant (Supplementary Fig. 3.20).

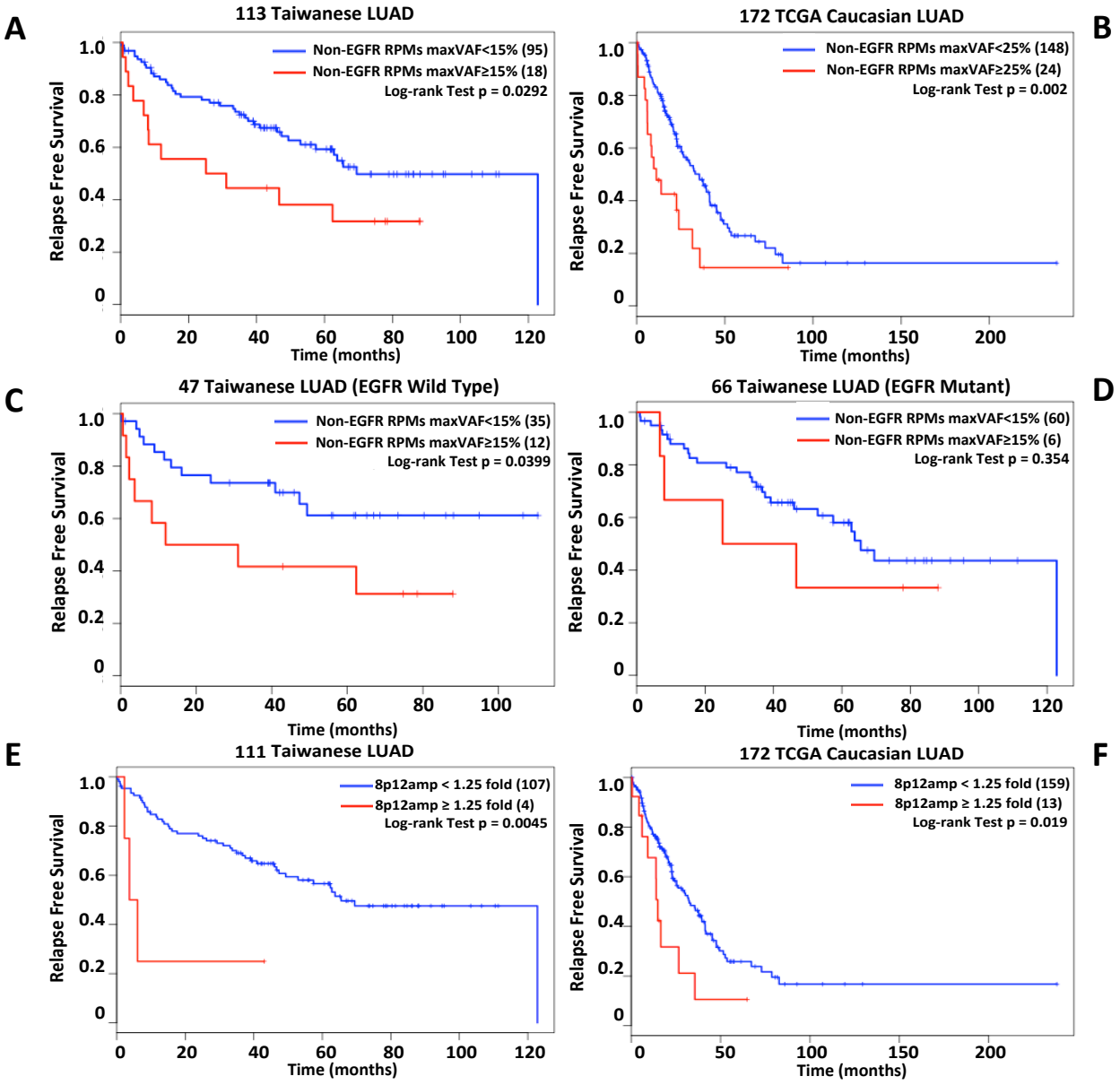
A
113 Taiwanese LUAD
Non-EGFR RPMs maxVAF<15% (95)
Non-EGFR RPMs maxVAF≥15% (18)
Log-rank Test p = 0.0292
Relapse Free Survival
Time (months)
B
172 TCGA Caucasian LUAD
Non-EGFR RPMs maxVAF<25% (148)
Non-EGFR RPMs maxVAF≥25% (24)
Log-rank Test p = 0.002
Relapse Free Survival
Time (months)
C
47 Taiwanese LUAD (EGFR Wild Type)
Non-EGFR RPMs maxVAF<15% (35)
Non-EGFR RPMs maxVAF≥15% (12)
Log-rank Test p = 0.0399
Relapse Free Survival
Time (months)
D
66 Taiwanese LUAD (EGFR Mutant)
Non-EGFR RPMs maxVAF<15% (60)
Non-EGFR RPMs maxVAF≥15% (6)
Log-rank Test p = 0.354
Relapse Free Survival
Time (months)
E
111 Taiwanese LUAD
8p12amp < 1.25 fold (107)
8p12amp ≥ 1.25 fold (4)
Log-rank Test p = 0.0045
Relapse Free Survival
Time (months)
F
172 TCGA Caucasian LUAD
8p12amp < 1.25 fold (159)
8p12amp ≥ 1.25 fold (13)
Log-rank Test p = 0.019
Relapse Free Survival
Time (months)

Figure 3.3: Kaplan-Meier analysis of relapse free survival. (A-D) Relapse free survival by maximum variant allele fraction (maxVAF) of Non-EGFR recurrent point mutations (RPMs): (A) all patients in our cohort,(B) all patients in the validation cohort from TCGA, (C) only patients with EGFR wild type in our cohort, and (D) only patients with EGFR mutant in our cohort. (E-F) RFS by focal amplification at 8p12: (E) all patients in the cohort and (F) all patients in the validation cohort from TCGA. P values were determined by log-rank test.

Table 3.2: Multivariate Cox regression analysis of RFS by maxVAF of non-EGFR RPMs

	Taiwanese LUAD (n=113)			Caucasian LUAD (n=172)		
Variable	Hazard ratio	95%CI	p	Hazard ratio	95%CI	p
maxVAF	25.08	4.28-146.98	< 0.0001	5.62	1.42-22.15	0.014
Gender:Male	0.88	0.42-1.87	0.735	1.22	0.79-1.90	0.376
Age	1.04	1.01-1.07	0.003	1.00	0.98-1.02	0.887
Smoking:Ever	1.58	0.70-3.56	0.272	1.20	0.63-2.27	0.576
Stage:IB	1.78	0.99-3.18	0.053	1.66	0.89-3.09	0.110
Stage:II	-	-	-	2.26	1.22-4.19	0.010
Stage:III	-	-	-	2.00	1.00-4.02	0.051
Stage:IV	-	-	-	1.65	0.53-5.12	0.386
EGFR:Mutant	1.81	0.89-3.70	0.104	1.35	0.62-2.95	0.448

We investigated the clinical relevance of somatic copy number alterations. Focal amplification at 8p12 was the only SCNA found to correlate with poor RFS in the multivariate Cox proportional hazard model (HR = 1.73, 95% CI = 1.01-2.96, $p = 0.046$). Based on the distribution of SCNA value at 8p12, we identified four tumors significantly higher than other tumors with a fold change ≥ 1.25. Kaplan-Meier analysis was performed to test the difference of RFS between those four tumors and the other tumors, and it was significant (Fig. 3.3E). This result was validated in the same Caucasian cohort. Both the SCNA value (HR = 2.82, 95% CI = 1.37-5.79, $p = 0.0049$) and dichotomized risk grouping were significantly correlated with RFS (Fig. 3.3F).

Table 3.3: Multivariate Cox regression analysis of RFS by log2 ratio of SCNA at 8p12

	Taiwanese LUAD (n=111)			Caucasian LUAD (n=172)		
Variable	Hazard ratio	95%CI	p	Hazard ratio	95%CI	p
SCNA (log2 ratio)	1.98	1.13-3.46	0.017	1.63	1.02-2.61	0.042
Gender:Male	0.98	0.47-2.04	0.959	1.26	0.81-1.97	0.310
Age	1.03	1.01-1.06	0.014	1.00	0.98-1.03	0.714
Smoking:Ever	1.66	0.76-3.64	0.206	1.18	0.62-2.25	0.617
Stage:IB	1.58	1.09-3.45	0.024	1.66	0.86-2.92	0.144
Stage:II	-	-	-	2.52	1.35-4.71	0.004
Stage:III	-	-	-	1.93	0.97-3.87	0.062
Stage:IV	-	-	-	1.91	0.62-5.93	0.260
EGFR:Mutant	1.76	0.86-3.57	0.120	1.19	0.55-2.58	0.665

3.3.7 Correlation between Significant DNA Alterations and miRNA Expression

We next investigated the miRNA expression in 69 paired tumor and normal samples of the cohort to explore the potential downstream molecular changes for the significant DNA alternations we identified. First, we performed paired t test to select differentially expressed miRNA between tumor and normal samples. In total, 162 out of 1,726 miRNAs were selected based on the criterions: absolute fold change ≥ 2 and $q < 0.05$ using Benjamini-Hochberg method. Then we examined the correlation between these miRNA expression in the tumor samples and the maxVAF score of the Non-EGFR RPMs. A Z-test of correlation based on Fisher transformation found miR-XX expression correlating with the maxVAF score (corr=0.46, $q < 0.05$, Benjamini-Hochberg method). In contrast, the correlation is very weak in the normal sample (corr=0.099, $p = 0.42$, Fig. 3.4A-B).

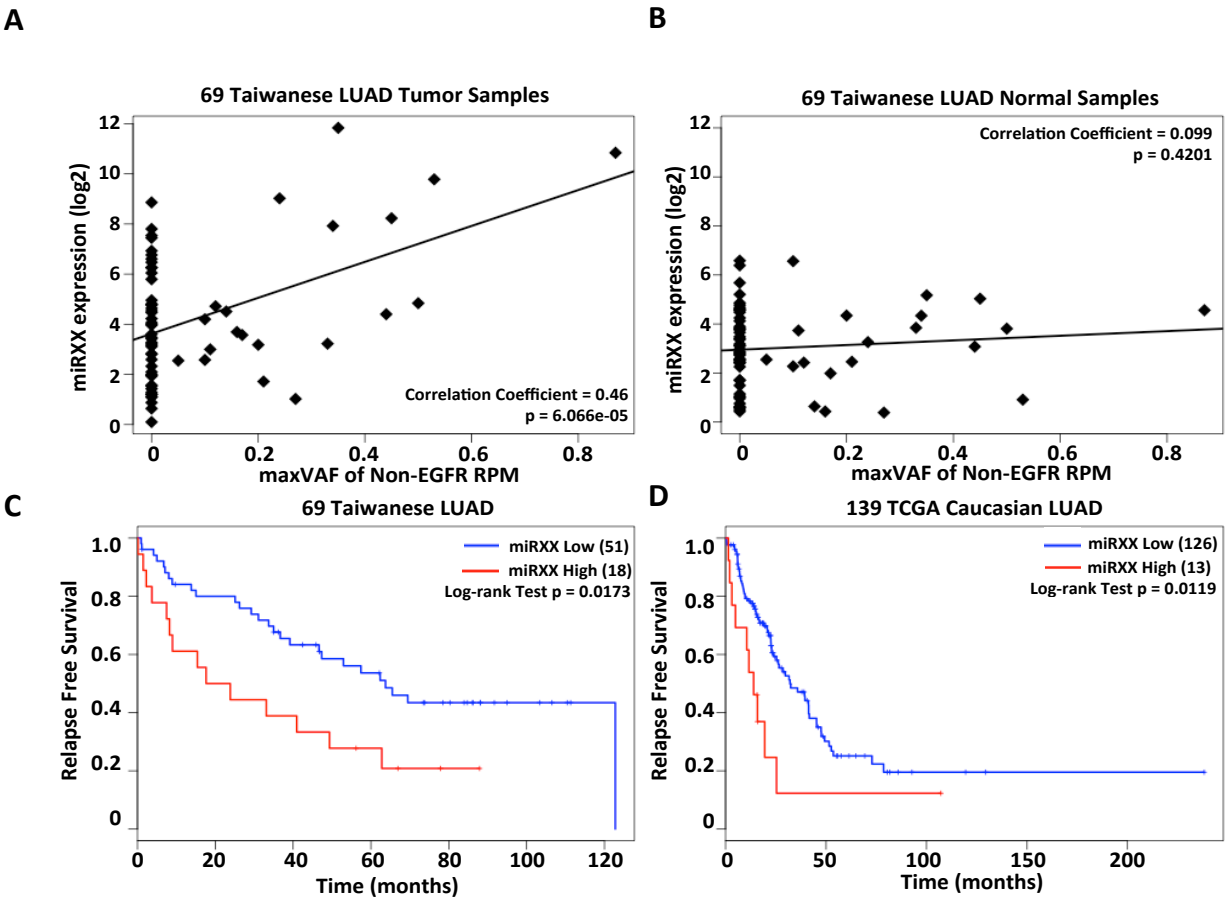

Figure 3.4: Correlations between miR-XX expression, recurrent point mutations and relapse free survival. (A-B) miR-XX Expression against the maximum variant allele fraction (maxVAF) of non-EGFR recurrent point mutations (VAF) in tumor and normal samples. P values were determined by the Z test based on Fisher's transform. (C-D) Relapse free survival by log2 miR-XX expression: (C) patients in our cohort, (D) patients in the validation cohort from TCGA. P values were determined by log-rank test.

We examined the correlation between miR-XX expression and RFS for our cohort. Upon observing that only the tumors with highly expressed miR-XX had poor RFS, we assigned tumors with miR-XX expression greater than the 75th percentile (55.48) to the high-risk group and the rest of tumors as the low-risk group. Kaplan-Meier analysis identified the significant difference of RFS between the two risk groups ($p = 0.0173$, Fig. 3.4C). Using the same cutoff value, this finding was validated in the cohort of 139 Caucasian LUAD patients from TCGA ($p = 0.0119$, Fig. 3.4D).

Table 3.4: Multivariate Cox regression analysis of RFS by miR-XX expression

	Taiwanese LUAD (n=69)			Caucasian LUAD (n=139)		
Variable	Hazard ratio	95%CI	p	Hazard ratio	95%CI	p
miR-XX:High	2.28	1.12-4.65	0.023	2.72	1.21-6.10	0.015
Gender:Male	0.73	0.33-1.63	0.442	1.15	0.69-1.92	0.586
Age	1.04	1.01-1.07	0.010	1.01	0.98-1.03	0.512
Smoking:Ever	1.33	0.58-3.06	0.497	1.26	0.60-2.66	0.546
Stage:IB	2.29	1.15-4.56	0.018	1.44	0.69-2.97	0.330
Stage:II	-	-	-	3.41	1.64-7.07	0.001
Stage:III	-	-	-	1.73	0.76-3.94	0.190
Stage:IV	-	-	-	3.33	1.02-10.94	0.047
EGFR:Mutant	1.95	0.94-4.0	0.072	1.90	0.63-5.69	0.254

3.3.8 Comparison of Mutation Frequency of Significantly Mutated Genes in Lung Adenocarcinomas between the East and the West Populations

To study the differential somatic mutation patterns in lung adenocarcinomas between the East and the West, we compared the mutation frequency of the significantly mutated genes we identified in the Taiwanese cohort with the Caucasian cohort of 172 patients from the TCGA paper[10]. Mutation frequencies of eight key components including three significantly mutated genes were first compared (Fig. 3.5A). The Fisher's exact test identified that the mutation frequency of EGFR was significantly higher in the Taiwanese cohort whereas KRAS and NF1 were more frequently mutated in the Caucasian cohort. Because the proportion of smokers was significantly different between the two cohorts, we stratified patients by the smoking status and performed the comparison within each group (Fig. 3.5B). EGFR mutation was significantly more common in the Taiwanese cohort regardless of the smoking status, which is in agreement with a previous study[35] of a comparison between Asian and Caucasian populations. The mutation frequency of KRAS was higher in the Caucasian cohort. Notably, NRAS was more frequently mutated in the Taiwanese smoker group. No statisti-

cally significant difference of mutation frequency was found for the other three significantly mutated genes TP53, RBM10, and CTNNB1.

The frequencies of 11 RPMs identified in the significantly mutated genes from tumors in our cohort were shown in Fig. 3.5C. Seven out of 11 (64%) RPMs were found in at least one Caucasian patient. Additionally, we used our method to identify RPMs for the 172 Caucasian patients. Based on the binomial probability model, a point mutation carried by at least three patients was identified as an RPM while controlling FDR to be less than 0.05. Six additional RPMs were found in five genes KRAS, U2AF1, PIK3CA, TP53 and BRAF, where four of them were also carried by at least one Taiwanese patient. We note that the frequency of RPMs in codon 12 of KRAS was higher in Caucasian patients, but an RPM c.G37T in codon 13 of KRAS was found in three patients in our cohort only but not in the 172 Caucasian patients.

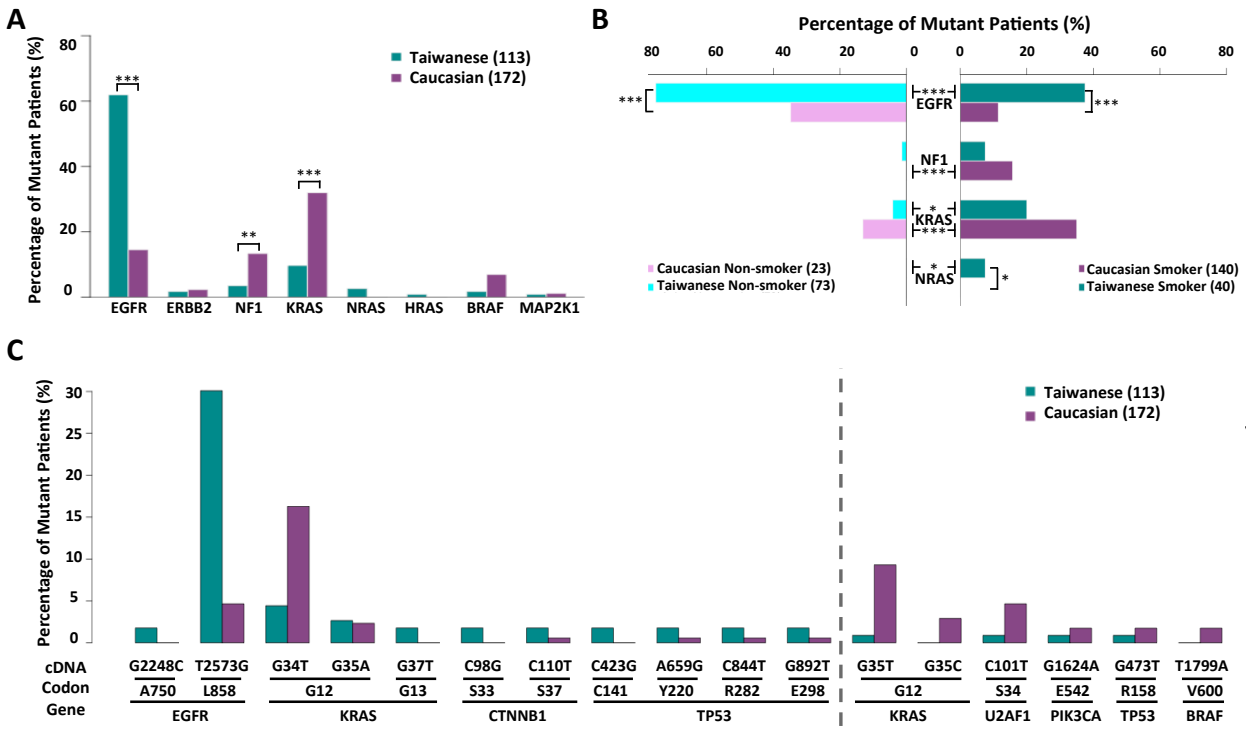

Figure 3.5: Comparisons of somatic mutations between tumors in Taiwanese LUAD cohort and Caucasian LUAD cohort from TCGA. (A) The differential patterns of somatic mutation distribution in the key component of MAPK pathways at gene level for all patients and (B) patients stratified by smoking status. Asterisks indicate statistical significance using the Fisher's exact test (*:$p < 0.05$, **:$p < 0.01$ and ***:$p < 0.001$). (C) Distribution of recurrent point mutations (RPMs) in cDNA changes from our cohort and TCGA cohort. The dashed line separates the RPMs derived from our cohort and TCGA cohort.

3.4 Discussion

In this study, we provided a comprehensive landscape of somatic DNA alterations in 113 Taiwanese lung adenocarcinomas constructing from whole exome sequencing and CGH-array. Our results demonstrated that, at the gene level, 81% of the tumors carries non-synonymous DNA mutations in the known driver genes associated with MAPK signaling and this percentage was higher than Caucasian population. The frequency of EGFR mutation was significantly higher in Taiwanese LUAD regardless of smoking status. Notably, we observed NRAS was more frequently mutated in Taiwanese smokers. A targetable proto-oncogene CTNNB1 on the WNT signaling pathway was also identified because of its clustered high functional impact mutations including two RPMs.

At the single nucleotide level, we utilized a binomial probability model and identified 18 significant non-synonymous RPMs distributed in known driver genes, tumor suppressor genes, and some novel genes including XXXX a subunit of the voltage-dependent calcium channel, YYYY a calcium-dependent cell-adhesion protein and ZZZZ another calcium-dependent cell adhesion molecule. The challenge for identifying RPMs was the large amount of potential false positive somatic mutation calls in the exome sequencing data from two major sources: sequencing/mapping errors in the tumor sample or insufficient depth in the paired normal sample. We derived a set of empirical filters for reducing the potential false positive somatic mutation calls by using multiple control samples (Supplementary Table 3.3). This approach was applied to the list of somatic mutations reported from the TCGA study

while identifying reliable RPMs.

To illustrate the clinical relevance of the significant somatic DNA alterations, we examined the impact to relapse free survival (RFS) of the patients. Because of the limited number of carriers, we did not have the statistical power to test the significance for each RPM. We proposed a new approach using the maximum variant allele fraction to aggregate the 16 non-EGFR RPMs as a score. The high score was significantly correlated with poor relapse free survival in both our and the validation cohorts. The concept of integrating mutation variant allele frequency for improving prognosis power of TP53 mutation was also discussed in a recent study of leukemia[43]. Another focal somatic amplification on 8p.12 was also identified correlating with poor relapse free survival. Noteworthy, we observed four patients carried less than 30 somatic mutations, no known driver mutation or somatic copy number alteration but had poor relapse free survival. More molecular profiling of these patients is needed for better understanding the mechanism of these early relapsed tumors beyond somatic DNA alteration.

This study further explored the potential downstream transcriptomic events through performing RNA sequencing and miRNA expression array. However, quality assessment of read sequence and alignment indicated contamination of human DNA and insufficient mRNA in the library. Thus, we only performed gene fusion detection with our RNAseq data since this analysis only relies on the splice junction reads that were unlikely from the genomic DNA (Supplementary Materials 3.5.7). For miRNA analysis, we correlated the maxVAF score of non-EGFR RPMs to the expression of miRNAs differentially expressed between tumor and normal samples and identified the only significant miRNA miR-XX a driver of lung tumorigenesis reported in a recent study[16]. In summary, this multiplatform genomic study constructed a comprehensive landscape of somatic DNA alterations and revealed clinical relevant molecular events beyond the known driver genes. We believe these findings should help improve patient management and generate hypotheses of new therapeutic targets in Taiwanese lung adenocarcinoma.

3.5 Supplementary Materials

3.5.1 Whole Exome Sequencing Data Analysis

We performed whole exome sequencing on paired tumor and adjacent normal samples from 113 patients. Whole exome sequencing data was processed through the pipeline (Fig. 3.6). Briefly, in the preprocessing stage, we used BWA[32] to map paired-end reads to human reference genome GRCh37/hg19 downloaded from UCSC and generated initial alignments in BAM format. GATK[12] and Picard tools were utilized for local Indel realignment, marking duplicates, base recalibration, and generating cleaned BAM files for subsequent analysis.

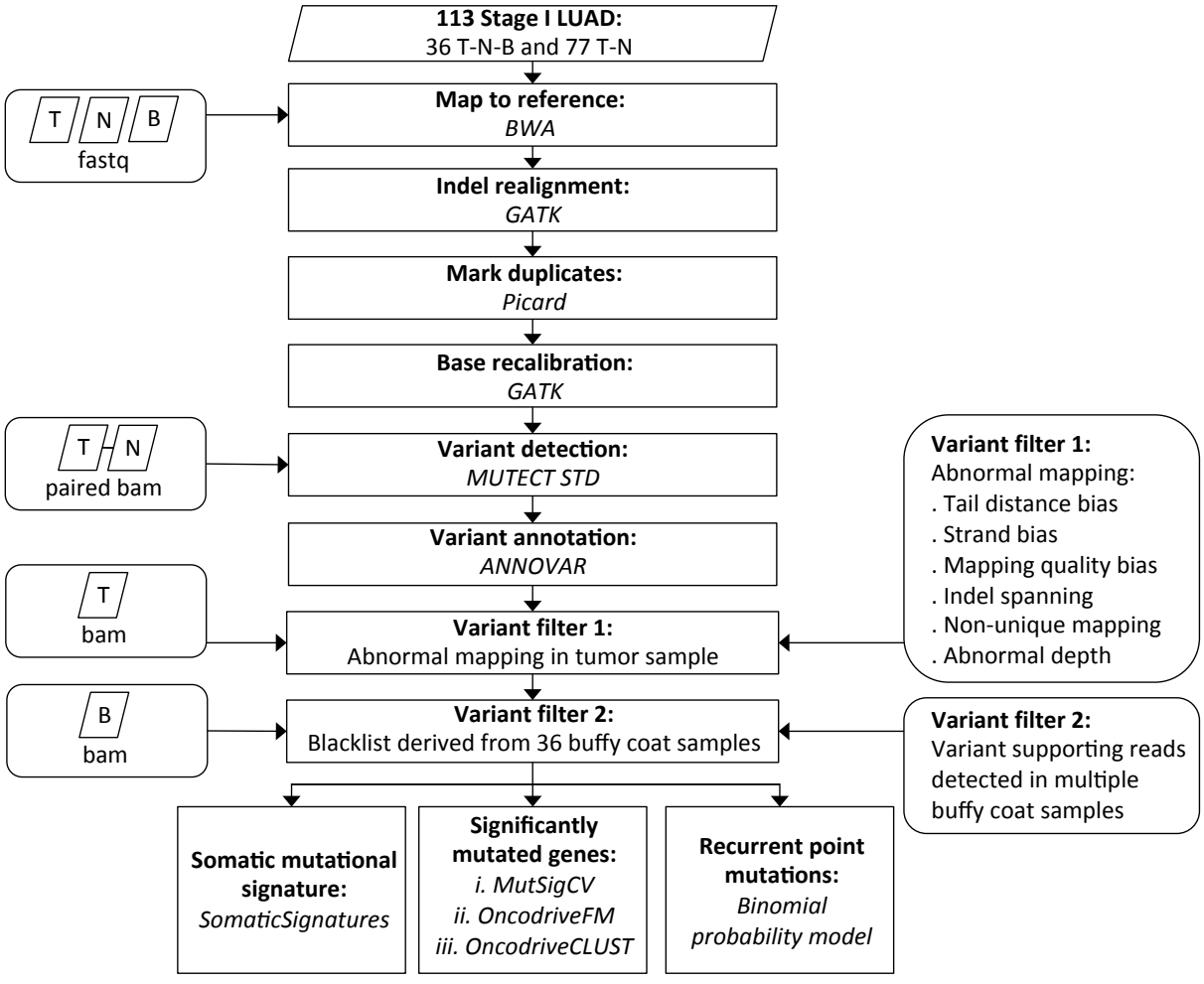

Figure 3.6: Whole exome sequencing analysis pipeline. Samples from tumor, normal and buffy coat are indicated as T, N and B correspondingly.

3.5.2 Somatic Single Nucleotide Variants Calling

Somatic single nucleotide variants (SSNV) were called by a two-step strategy. In the first step, we used Mutect[8] standard mode to detect candidates of somatic mutations. In the second step, we derived a set of empirical filters for removing the potential false positives (Table 3.5).

Table 3.5: Empirical filters to remove potential false positives

Mutect standard mode
Tumor Q5q5 depth ≥ 15
Tumor variant supporting reads ≥ 4
Tumor log odds score (LOD) ≥ 6.3
Normal log odds score (LOD) ≥ 5.5
Abnormal mapping in tumor sample
Tail distance bias: Average tail distance of variant supporting reads between 10 to 90
Strand bias: Fraction of forward stranded variant supporting reads between 1% to 99%
Mapping quality bias: Maximum mapping quality of variant supporting reads ≥ 60
Indel spanning: Number of reads with Indels within 2 base-pairs ≤ 1
Non-unique mapping: Fraction of non-unique mapped reads $< 20\%$
Abnormal depth: Positions with depth > 2 times average require VAF $> 10\%$.
Blacklist derived from 36 buffy coat samples
Less than 5% buffy coat samples have variant supporting reads ≥ 3 or VAF $\geq 5\%$

Mutect is one of the most commonly used software for SSNV calling. It takes paired tumor and normal samples as input and returns a list of SSNV candidates based on the log odds scores (LOD). It also indicates of SSNV calls that passed a post-filtering step as high-confident (HC) calls. We studied the high-confident calls detected in our cohort, and

we noticed a significant amount of skeptical calls. Most skeptical calls were considered to be common artifacts or germline variants in tumor samples with insufficient depth in the paired normal samples. These skeptical calls were found in multiple samples, and this pattern will introduce substantial false positives for subsequent analyses, especially for the recurrent point mutation analysis we highlighted. On the other hand, we also noticed some rejected calls could be as a result of tumor contamination in samples of adjacent normal tissues.

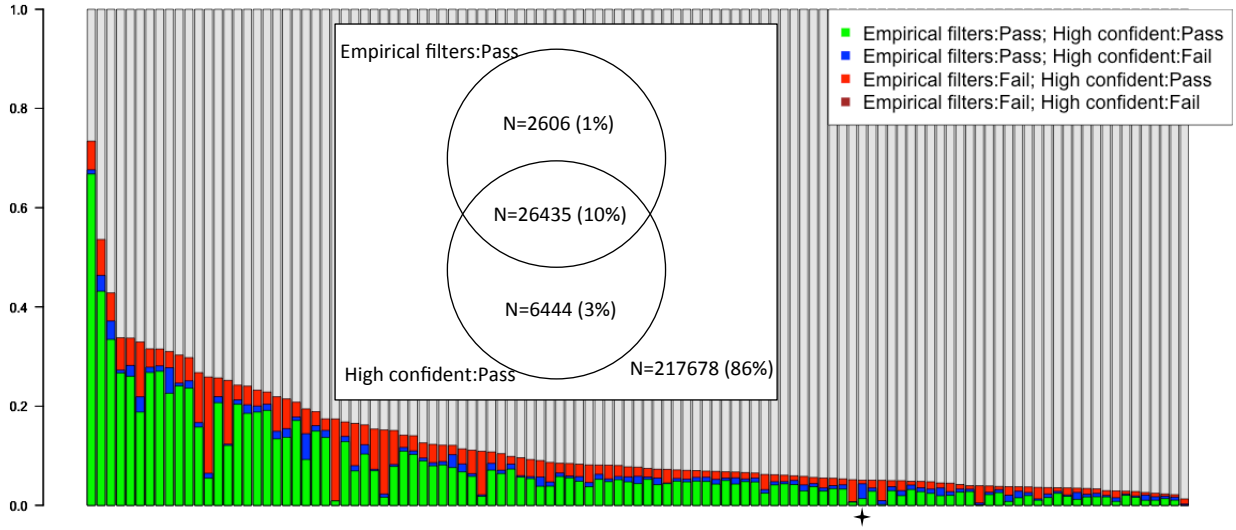

Figure 3.7: The distribution of proportions of SSNV candidates passed our empirical filters vs Mutect high confident mode for each patient. The star indicates patient 256. Amount the SSNVs passed our empirical filters, 68% were rejected by Mutect high confident mode and 63% had variant supporting reads in the normal sample. The van diagram shows the total number of SSNV candidates passed or failed our empirical filters vs Mutect high confident mode.

To address these two issues, instead of using the high-confident calls we utilized its standard mode to generate a list of SSNV candidates and then applied a set of empirically derived filters to remove the skeptical calls. The empirical filters include some criterions for known features of artifacts in whole exome sequencing data and a blacklist derived from the additional 36 whole exome sequencing data of buffy coat samples. This approach removed 98.8% standard calls rejected by Mutect high confident mode and other common artifacts or germline variants. This approach also has a higher tolerance for tumor contaminated normal samples which is useful in practice when using adjacent normal tissues instead of blood as

control samples. (Fig. 3.7).

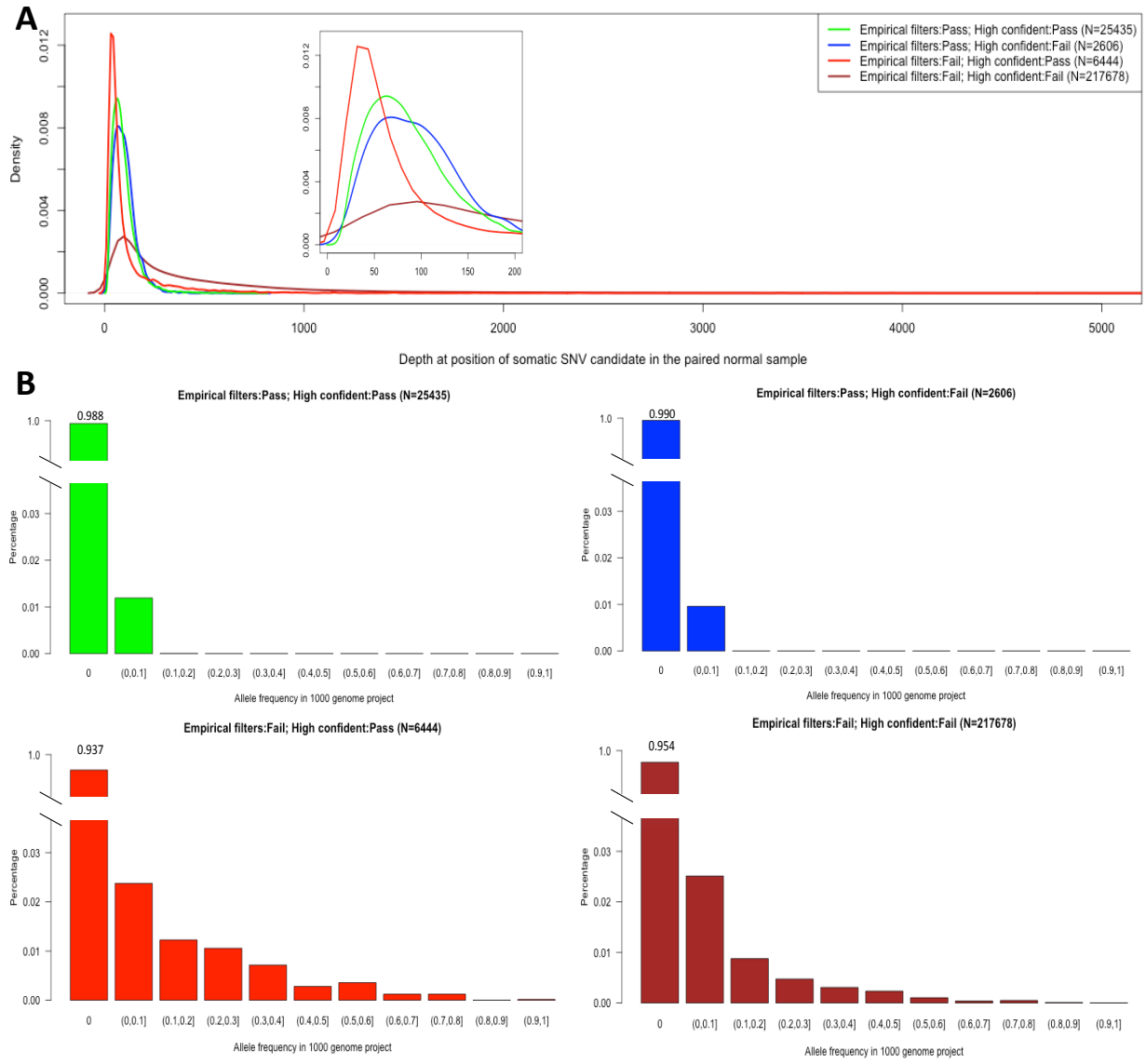

Figure 3.8: A comparison of our empirical filters vs Mutect high confident mode: (A) Density of depth at position of somatic SNV candidate in the paired normal sample. (B) Distributions of Allele frequency in 1000 genome project.

To show the capability of our blacklist to remove common artifacts, we conducted a validation of using Sanger sequencing. The selected 4 SSNV candidates had the following features: i. Rejected by the blacklist, ii. Passed Mutect high confident mode, and iii. Variant allele fraction $\geq 15\%$. All 4 candidates were validated as artifacts with no alternative allele

in both tumor and normal samples. The size of this validation is small since the remaining frozen tumor tissues are limited. Here we presented some comparisons for the SSNVs called by the two approaches. First, figure 3.9A shows that the Mutect high confident calls rejected by our blacklist tended to have a lower depth in the paired normal sample. Second, the SSNV candidates rejected by our empirical filters had a higher proportion of common SNPs in the 1000 genome database. Together, our empirical filters can help remove common germline variants and artifacts even when the paired normal sample had an insufficient depth.

3.5.3 Validation of Recurrent Point Mutations using Sanger Sequencing and MASS Spectrometry

We performed Sanger sequencing and MASS spectrometry techniques to validate the recurrent point mutations. Validations with Sanger sequencing were conducted in a total of 3 rounds. We acknowledged that the limitations of low allele fraction for Sanger sequencing. In the first round, eight recurrent somatic point mutations with allele fraction greater than 15% in genes other than EGFR, KRAS, and TP53 were picked for Sanger sequencing on both tumor and normal samples of the mutant patients. All 8 point mutations were successfully validated as somatic mutations. In the second round, we attempted to validate the remaining 6 recurrent somatic mutations with allele fraction under 15% in genes other than EGFR, KRAS, and TP53. Only 1 mutation was validated as a somatic mutation. The other 5 calls were homozygous reference in the tumor samples. For known mutations in EGFR (L585R) and KRAS (codon 12 and 13), MASS spectrometry technique was performed for validation, where it is known to be more sensitive to low allele variant than Sanger sequencing. All 46 mutations validated as Somatic even for the very low allele fraction cases. Based on previous experience, in the third round, we performed Sanger sequencing to validate mutations in TP53 with allele fraction greater than 15%. All 5 mutations were validated as somatic mutations.

3.5.4 Somatic Copy Number Alteration Analysis

We performed array CGH on 111 paired tumor and normal samples. Following the standard protocol, the genomic DNA was extracted from frozen cancer tissue of each sample with quality checked by agarose electrophorese. The whole genome NimbleGen CGH array (NimbleGen Systems Inc, Madison, WI) containing $385,806$ probes with probe spacing of about $6,000$ base-pair, was used. Instead of hybridizing the tumor and normal samples directly, each sample was labeled with fluorophore Cy3 and cohybridized with a reference DNA sample label with fluorophore Cy5 separately. The reference DNA sample was extracted from the PBMC of one male and one female in a community cohort. This approach provided a common reference for both tumor and normal samples and it allowed us to perform a global normalization across the cohort.

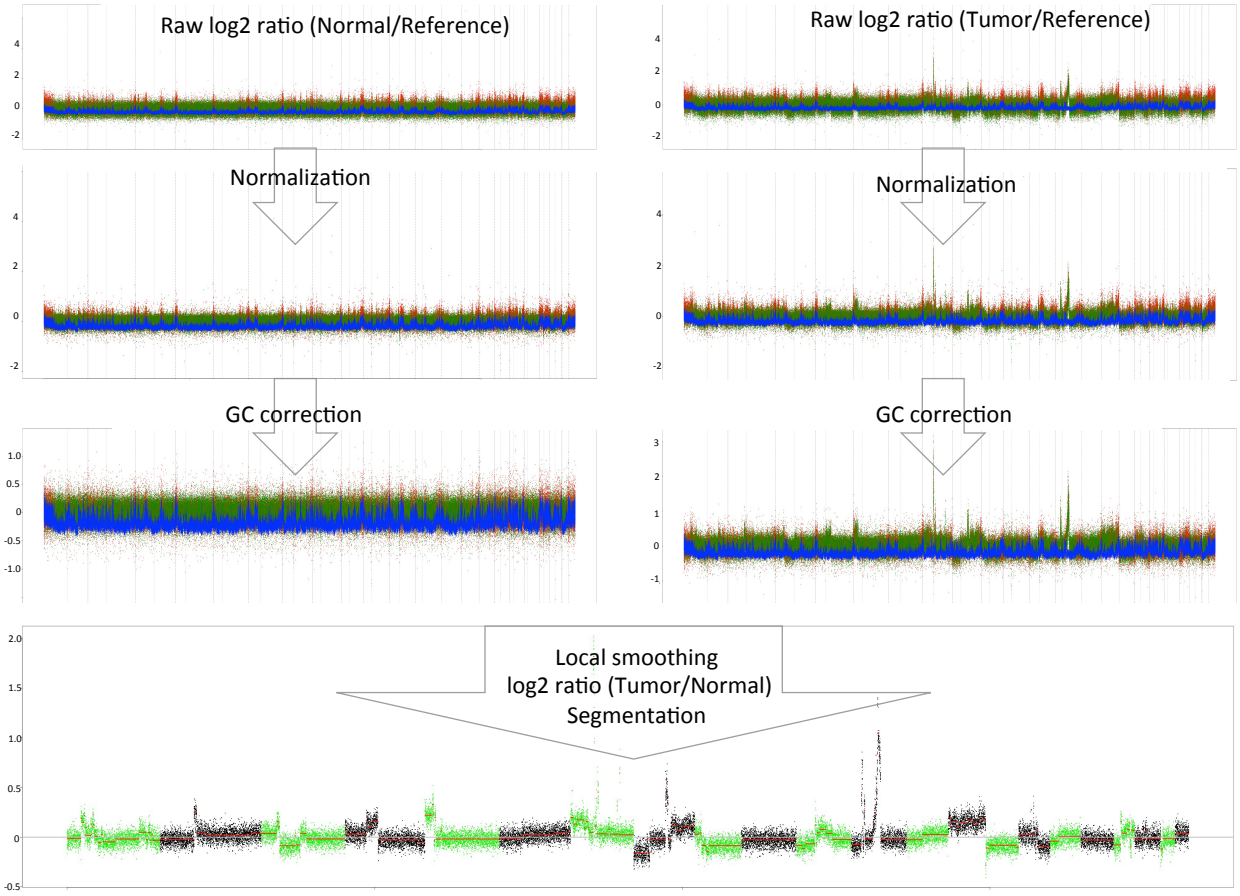

Figure 3.9: Somatic copy number alteration normalization pipeline.

Here we briefly summarize our normalization pipeline.

i. Normalization[52]: First we established a common reference profile by taking the average of the probe intensities of the Cy5 channels in all array. Then for each array, we performed LOWESS adjustment for each channel by the common reference. We used the residuals from the linear regression of adjusted Cy3 against adjusted Cy5 as the normalized CNA value for each probeset.

ii. GC correction: We performed LOWESS adjustment for the normalized CNA value by the percentage of GC-content in the 10K window spanning each probeset (GC10K). Second, the same LOWESS adjustment was performed for the percentage of GC-content in the sequence of each probeset (GCprobe) by GC10K. Lastly, we performed LOWESS adjustment for adjusted CNA value by adjusted GCprobe and got this GC-corrected CNA value.

iii. Local smoothing: We aggregated each 10 consecutive probesets into a probe-block by taking the average of the GC-corrected CNA values.

iv. Somatic CNA value: For each patient, the log2 ratio of local smoothed tumor CNA value to normal CNA value.

v. Segmentation: We performed the circular binary segmentation (CBS) algorithm on the Somatic CNA (SCNA) value using an R package DNAcopy[39].

The performance of the normalization was evaluated by the following criterions: i. Derivative log2 ratio standard deviation (DLRSD) - a measure of probe-to-probe log ratio noise of an array, ii. Correlation to the common reference - a measure of intensity-dependent ratio iii. Correlation to GC10K - a measure of the effect of GC-content local genome region, and iv. Correlation to GCprobe - a measure the effect of GC-content in the probe sequence (Fig. 3.10). The normalization step reduced DLRSD and the correlation to the common reference, whereas the GC correction reduced the correlation to GC10K and GCprobe. Combining these two steps reduced the values for all these 4 criterions. The local smoothing and

segmentation steps further reduced the random noise. The segmented log2 ratio was used as SCNA value in all subsequent analyses.

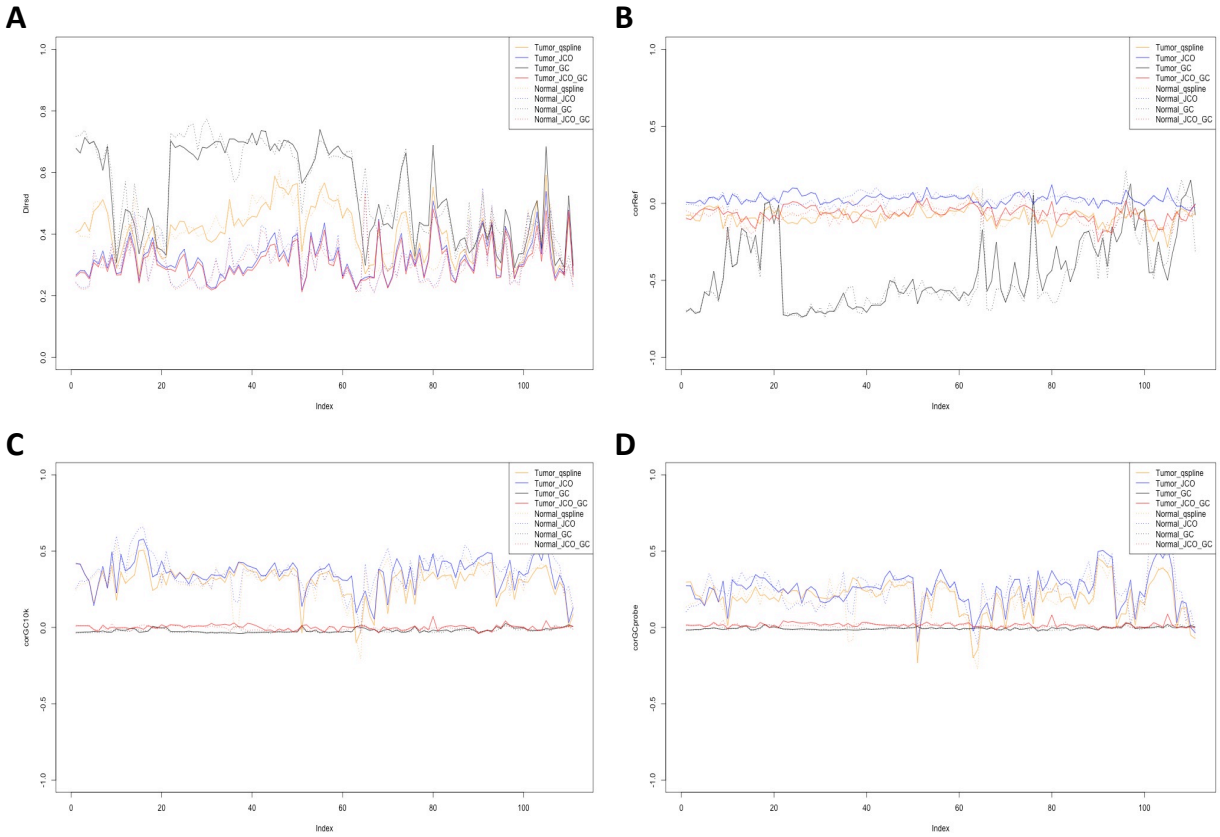

Figure 3.10: Evaluation of normalization approaches: (A) Derivative log2 ratio standard deviation, (B) Correlation to the common reference, (C) GC10K and (D) GCprobe

3.5.5 Concordance of SCNA between Two Platforms: Array CGH vs Whole Exome Sequencing

Here we compared the SCNA values quantified by array CGH versus whole exome sequencing data. For WXS, we performed VarScan2.0[26] software to get the $\log 2$ tumor to normal depth ratio in each 100bp interval adjusted by the GC-content. For this comparison, these 100bp intervals were mapped to the corresponding probe-blocks based on the chromosomal coordinate on reference genome hg19. We took the average of adjusted $\log 2$ ratio in each probe-block (Fig.3.11). The CBS algorithm was utilized for segmentation.

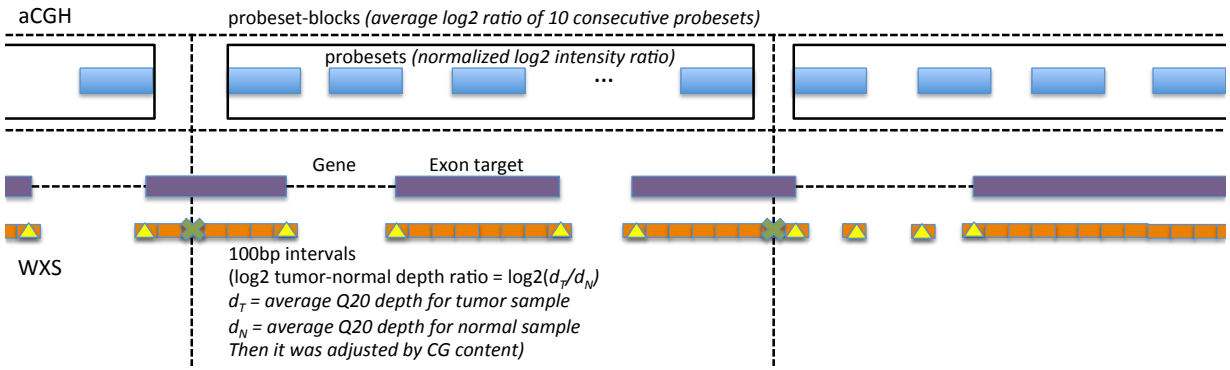

Figure 3.11: Mapping 100bp intervals from WXS to the corresponding aCGH probe-block. The green crosses indicate the ambiguous intervals and the yellow triangles indicate the partially on target intervals. Both intervals were excluded before taking the average.

Figure 3.12A-B shows two scatter plots of SCNA value derived from WXS against aCGH data in a sample. The two profiles shared a strong linear relationship. We also noticed that the amplitude of SCNA values from WXS was larger than aCGH. However, there was not evident to conclude that as a result of an underestimate in aCGH or an overestimate in WXS. Using Pearson's correlation coefficient as an measure of concordance and mean absolute copy number value as a measure of global copy number alternations, we further investigated the relationship between SCNA value (log2 ratio of tumor to normal) from whole exome sequencing data and relative copy number value for tumor to reference and normal to reference from array CGH data separately. In tumor samples, as the mean absolute CN value greater than 0.04, Pearson's correlation coefficients were greater than 0.5; whereas for some patients had no significant SCNA, the concordance between two platforms was low. No patient had mean absolute CN value greater than 0.04 in normal samples, but there was only one patient, 256, who's copy number profile in the normal sample was correlated with his somatic copy number profile (Fig. 3.12D-F). This again suggested a possibility of tumor contamination in the paired adjacent normal sample for patient 256.

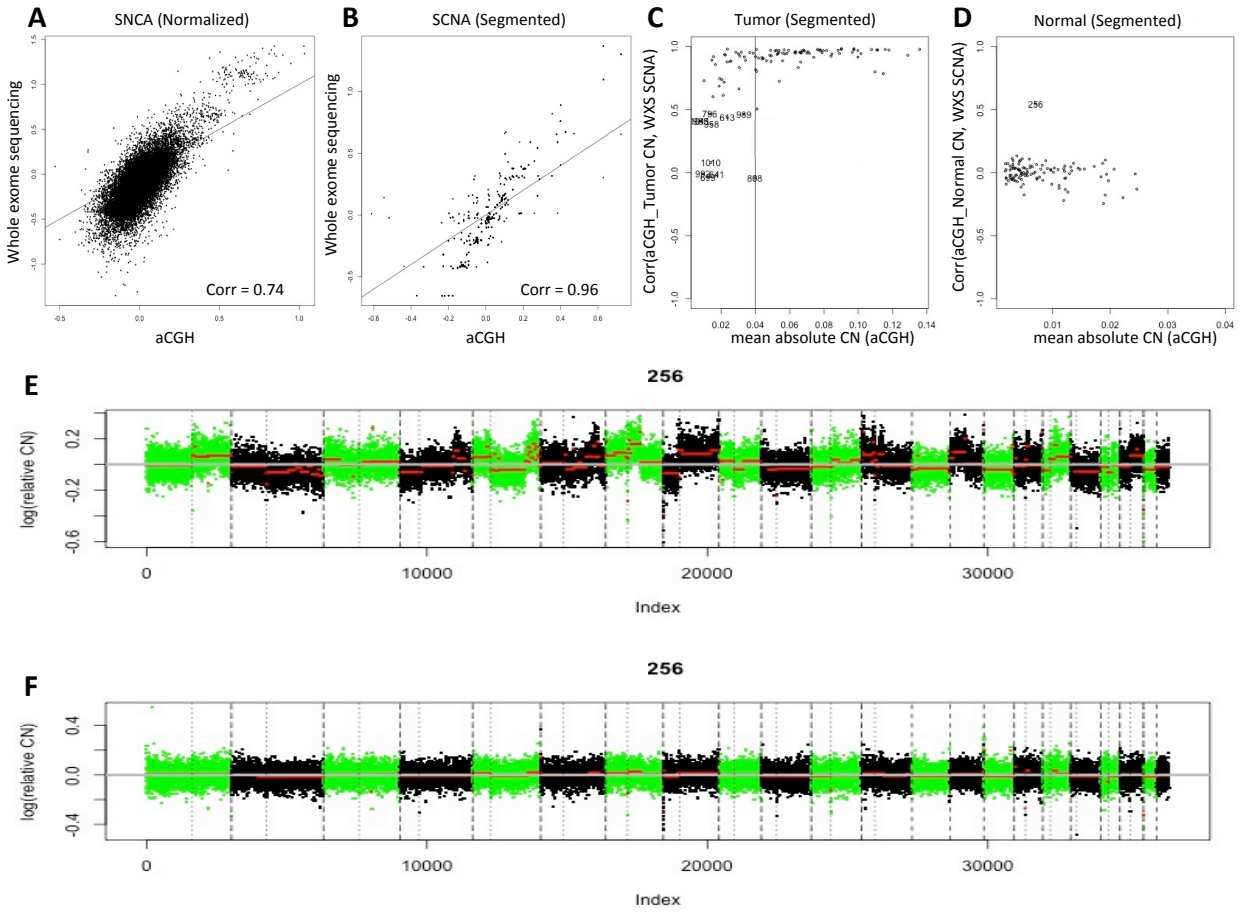

Figure 3.12: Comparison of SCNA values between two platforms: aCGH vs WXS. (A-B) Two scatter plots of SCNA value derived from WXS against aCGH data in sample where the solid line indicated the 45-degree line. (C-D) (E-F) The global SCNA profiles in tumor sample and normal sample.

3.5.6 Significant Focal Somatic Copy Number Alterations

Significant focal SCNAs were identified using GISTIC2.05[37] (Fig. 3.13). We compared the focal SCNA value between the EGFR wild type and mutant tumors and identified 3 focal SCNAs correlated with EGFR mutation status including focal amplifications at 7p.11 (EGFR), focal deletions at 9p21.3 (CDKN2A/CDKN2B), and focal amplification at 8p12 (ZNF703, Fig. 3.14).

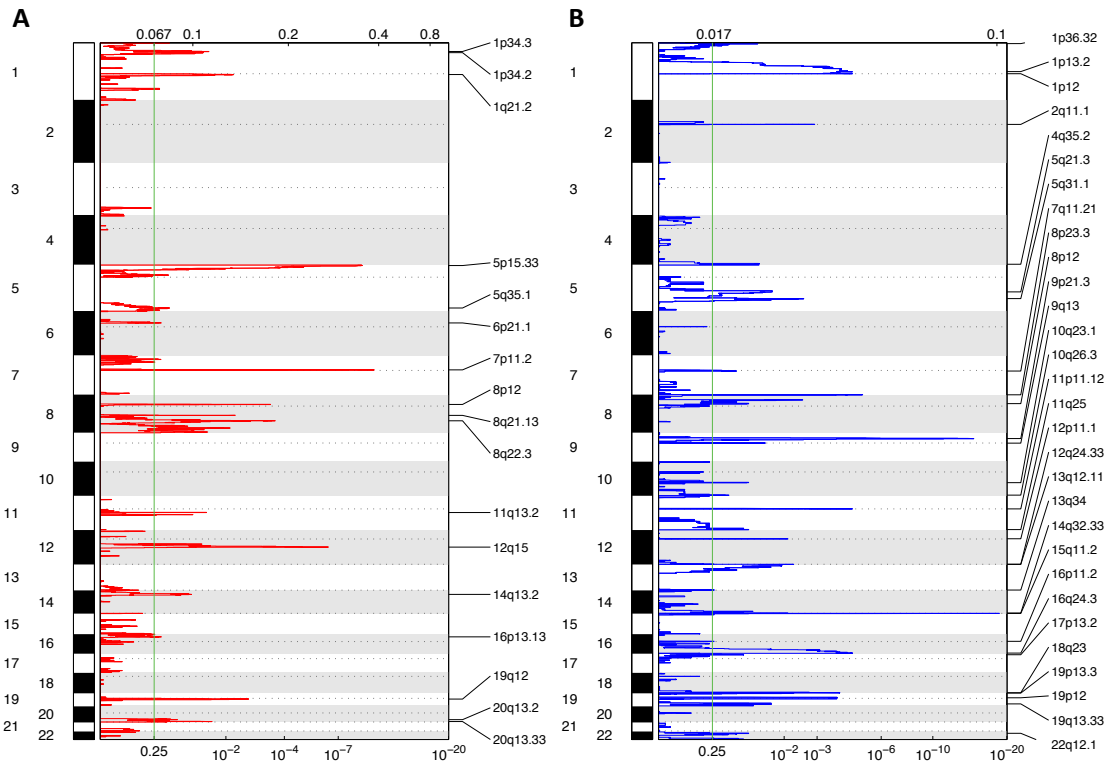

Figure 3.13: Significant focal amplification (red) and deletion (blue) plotted along the genome.

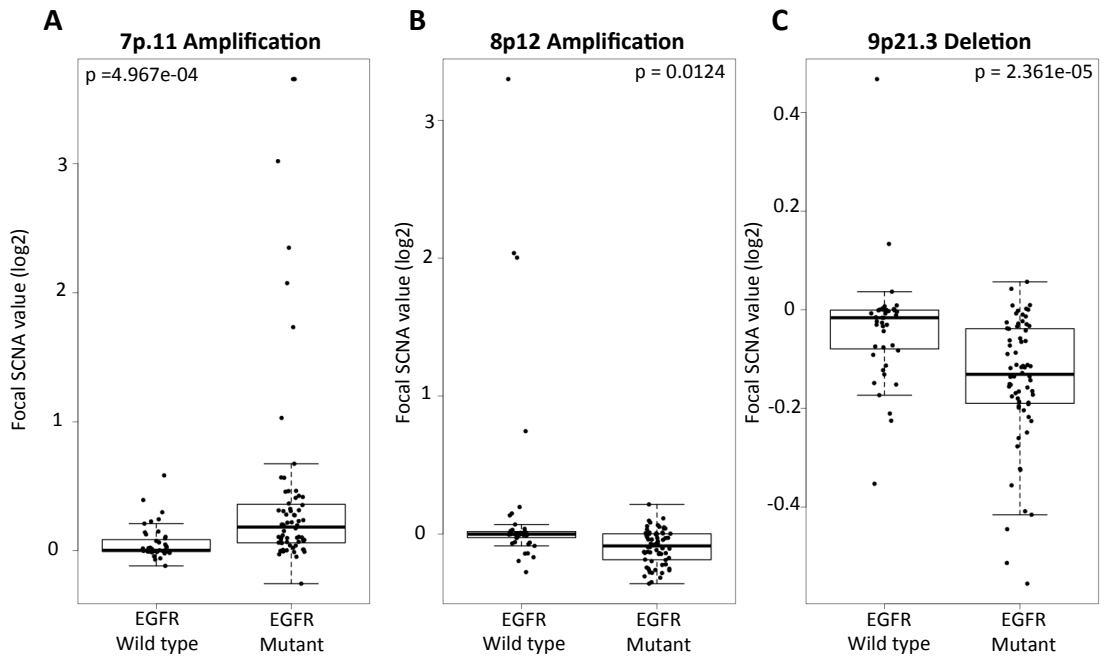

Figure 3.14: Boxplots of SCNA values stratified by EGFR mutation status: (A) focal amplifications at 7p.11, (B) focal deletions at 9p21.3, and focal amplification at 8p12.

3.5.7 RNA Sequencing Data Analysis

We performed RNA sequencing on the tumor sample from 107 patients. After QC and read trimming, we utilized Tophat2[24] to map paired-end reads to the human reference genome hg19.

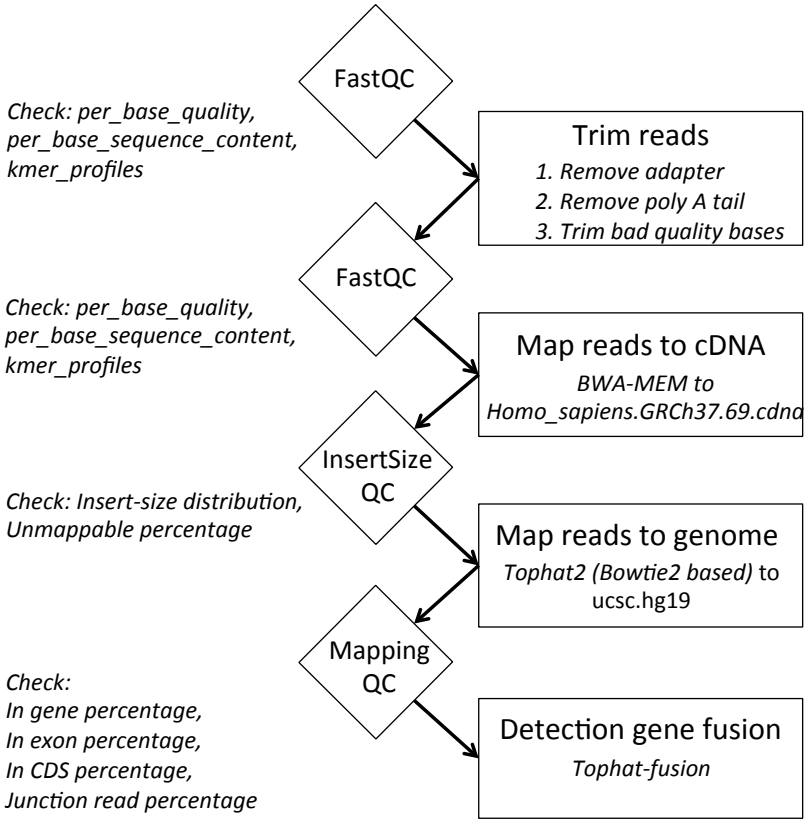

Figure 3.15: RNA sequencing data analysis pipeline.

The percentage of uniquely mapped reads to different regions for each sample is shown in figure 3.16A. The rate of reads mapped to exon regions of protein coding genes was low and had a median 4.89% compared to results from other literature[53] (about 35% for Ribo-Zero-Seq and 75% for mRNA-Seq). Also, we observed that some samples had many reads mapped to exons of non-coding genes in mitochondria (Fig. 3.16A-B). Together, due to the insufficient RNA-content, we did not perform the quantification of gene expression on this RNAseq data. For the issue of somatic mutation confirmation in expressed RNA transcripts, using a depth of 5 as a cutoff, 658 somatically mutated sites were expressed, and

30% mutations were confirmed by at least 1 variant supporting read in the RNAseq data (Fig. 3.16C).

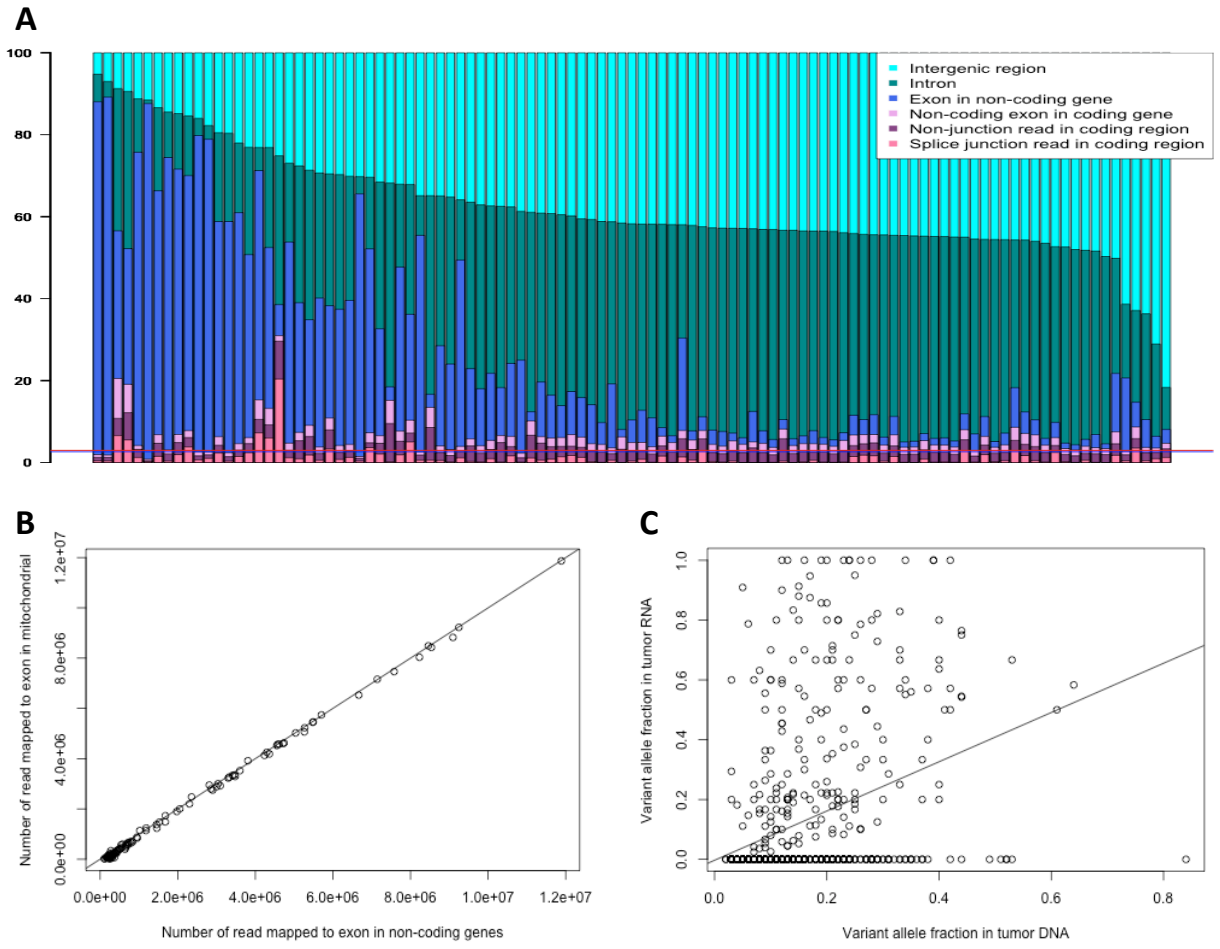

Figure 3.16: (A) Percentage of uniquely mapped reads in different regions for each sample. (B) Number of reads mapped to exons in mitochondria against number of reads mapped to exons in non-coding genes. (C) Variant allele fraction in tumor RNA samples against tumor DNA samples for only sites with a depth ≥ 5.

3.5.8 Gene Fusion Analysis

We performed gene fusion analysis using our RNAseq data since it only relies on the splice junction reads that were unlikely from the genomic DNA. We utilized Tophat-fusion[25] to perform split mapping for each sample. To be conservative, we only kept the fusions

of two complete exons jointed at the splice sites in an abnormal order. In total, 21 gene fusions were identified. Some cancer-related genes were involved in the fusions, including EGFR, JAK2, and MAX (Table 3.6). However, we did not observe the common driver fusions involving ALK, ROS, and RET reported in 10% Asian lung adenocarcinomas from the previous study[35]. No fusion identified in our cohort was found in the TCGA cohort.

Table 3.6: Summary of gene fusions.

Patient	Gene1	Chromosome1	Position1	Gene2	Chromosome2	Position2
552	HIPK1	chr1	114506143	RP4-663N10.1	chr1	115908291
624	TARBP1	chr1	234536599	SGIP1	chr1	67184977
624	TARBP1	chr1	234536927	SGIP1	chr1	67184977
768	RANBP9	chr6	13632602	RANBP9	chr6	13641539
226	RANBP9	chr6	13632602	RANBP9	chr6	13644961
651	RANBP9	chr6	13632602	RANBP9	chr6	13644961
704	RANBP9	chr6	13632602	RANBP9	chr6	13644961
803	RANBP9	chr6	13632602	RANBP9	chr6	13644961
A1099	RANBP9	chr6	13632602	RANBP9	chr6	13644961
803	RANBP9	chr6	13632602	RANBP9	chr6	13659064
973	RANBP9	chr6	13632602	RANBP9	chr6	13697128
245	RP1-240B8.3	chr6	62362160	KHDRBS2	chr6	62442669
768	RP1-240B8.3	chr6	62362160	KHDRBS2	chr6	62442669
802	TULP4	chr6	158735300	RP11-732M18.3	chr6	158703295
546	EGFR	chr7	55270318	DPYD	chr1	98187227
1014	JAK2	chr9	4986022	FAM120A	chr9	96233423
506	TSSC2	chr11	3405514	IFT46	chr11	118430579
A1676	AHNAK	chr11	62259204	XIAP	chrX	123019481
712	FNTB	chr14	65453815	MAX	chr14	65560533
628	GDF15	chr19	18493257	SH3GLB2	chr9	131777183
1016	USP9X	chrX	40945362	MACC1	chr7	20201493

3.5.9 Analysis of 172 Caucasian Lung Adenocarcinomas from TCGA

To compare the somatic molecular events between Taiwanese and Caucasian lung adenocarcinomas, we studied 172 Caucasian Lung Adenocarcinomas from the TCGA study[10]. We observed some inconsistencies between the data released from their paper and the website. To be consistent with their paper, only the 230 patients selected by their inclusion and exclusion criteria were considered. Amount these 230 patients; we selected 172 patients with *race* indicated as "WHITE" in the clinical information downloaded from their website. To get the complete information for relapse free survival analysis, information from the three files: i. Clinical information `nationwidechildrens.org_clinical_patient_luad.txt`, ii. Follow-up information `nationwidechildrens.org_clinical_follow_up_v1.0_luad.txt`, and iii. New tumor event `nationwidechildrens.org_clinical_nte_luad.txt` were combined. The relapse free survival time and status were calculated following the guidance[17].

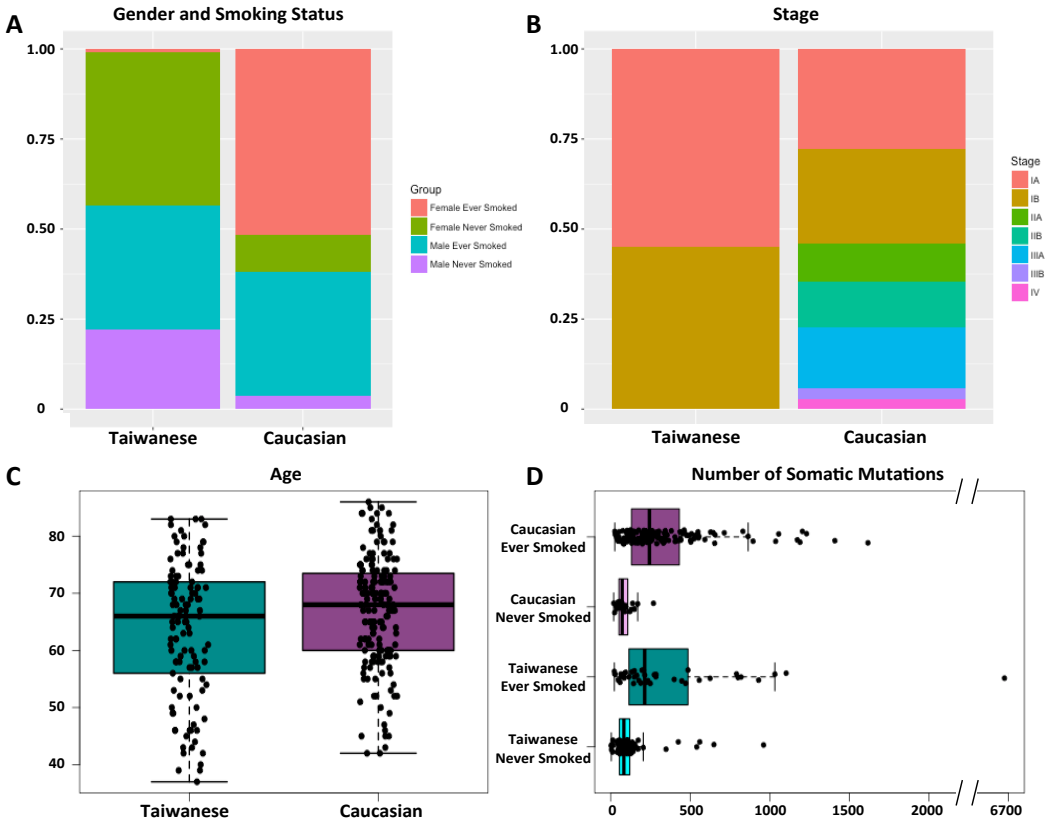

Figure 3.17: Comparison of background information between the Taiwanese and Caucasian cohorts.

For somatic mutations analysis, the mutation list from their paper was used. When identifying the recurrent point mutations, we observed the similar skeptical calls on the list. These skeptical calls were found in multiple normal samples from their cohort and also our cohort. Therefore, we applied 2 empirical filters in our method to reduce the false positives: i. Less than 5% normal blood samples have variant supporting reads ≥ 3 or $VAF \geq 5\%$, and ii. Mapping quality bias: Maximum mapping quality of variant supporting reads ≥ 60. Note that, there were 145 blood samples performed whole exome sequencing from the TCGA cohort. For SCNA analysis, the level-3 segmented hg19 SCNA profile from SNP6.0 was used. For miRNA expression analysis, the level-3 normalized miRNA expression from miRNAseq was used.

3.5.10 Supplementary Figures

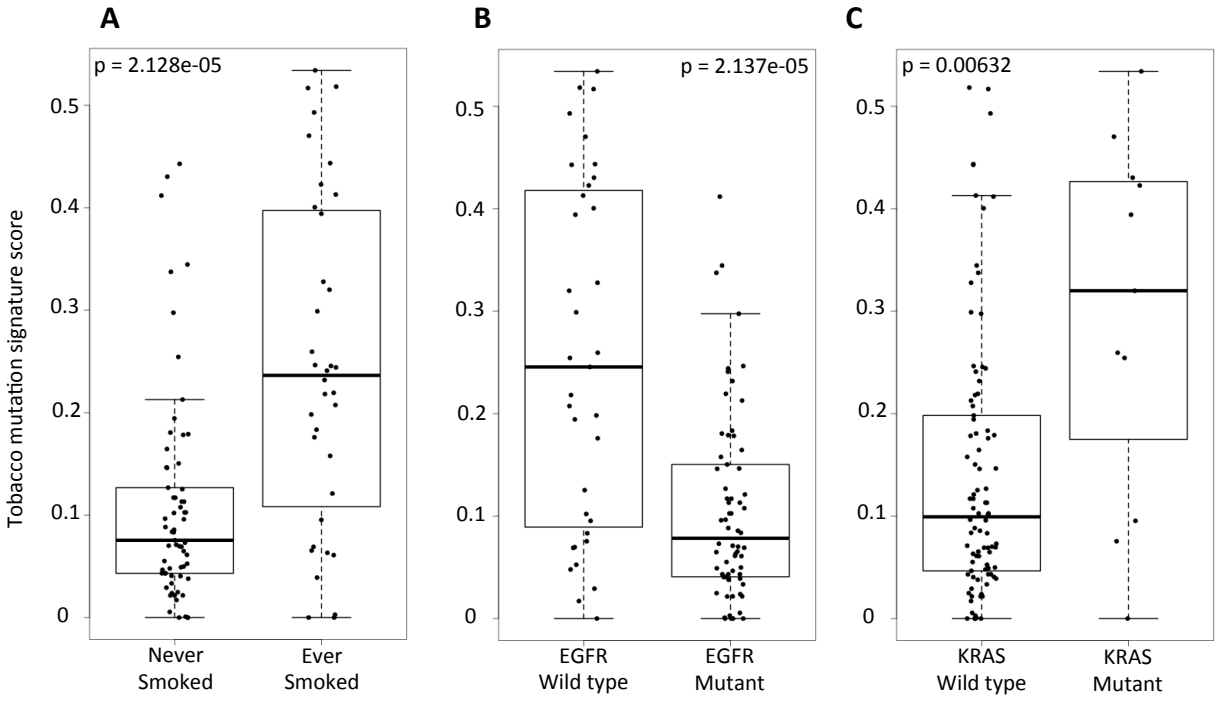

Figure 3.18: Tobacco mutation signature score by (A) smoking status (B) EGFR mutation status (C) KRAS mutation status.

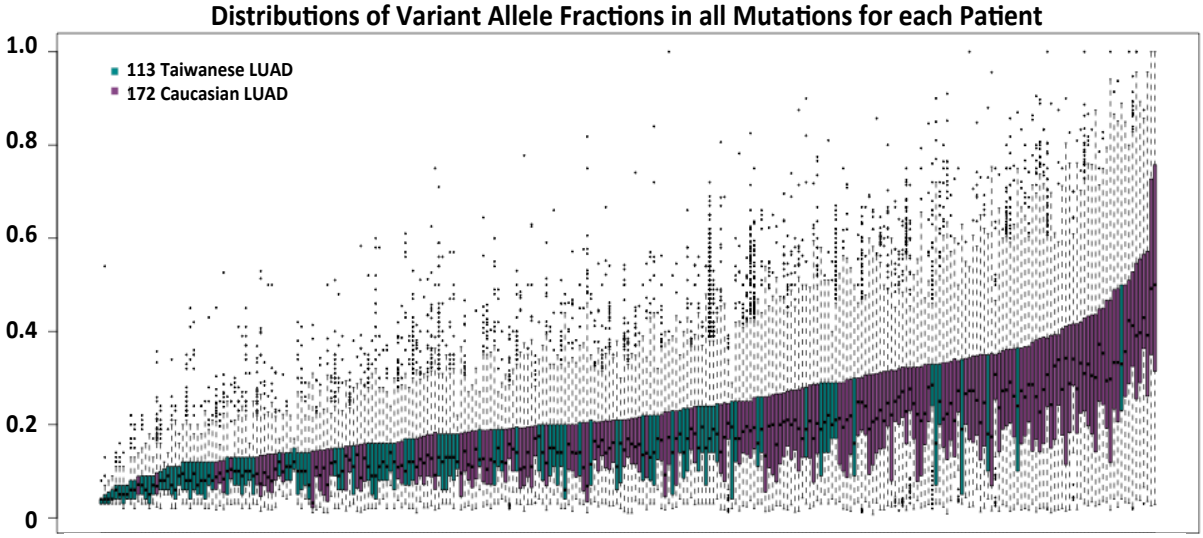

Figure 3.19: Distributions of variant allele fraction from patients across two cohorts.

Figure 3.20: Kaplan-Meier analysis of relapse free survival. Relapse free survival by maximum variant allele fraction (maxVAF) of Non-EGFR recurrent point mutations (RPMs): (A) patients with stage M0 and N0 in the Caucasian cohort and (B) patients with stage M0, N0 and from T1 to T2a in the Caucasian cohort

CHAPTER 4

Identification of Key Cis-Regulators of Differential Gene Expression between EGFR Mutant and Wild-Type Lung Adenocarcinomas

4.1 Introduction

Non-small cell lung cancer (NSCLC), the most common cause of cancer deaths worldwide, had a predominant subtype, Lung adenocarcinoma (LUAD). EGFR activating mutations were commonly found in LUAD patients and had a higher rate in Asians than in Caucasians[35]. Our previous study (Chapter 3) showed more than 70% of female never-smoker adenocarcinomas in Taiwan had EGFR activating mutations. EGFR mutation activates downstream anti-apoptotic signal transduction via Akt pathway through mediated phosphorylation or proliferative signals via MAPK/ERK pathway. From the clinical viewpoint, patients with these mutations showed a good response to EGFR-Tyrosine Kinase Inhibitors (TKIs) in phase II and III trials[28]. Several studies have established that EGFR-TKIs were, in general, more effective for patients with EGFR-activating mutation than EGFR wild-type. Based on the IPASS study[18], 71% of patients with EGFR activating mutation responded well to EGFR-TKIs. The molecular cause of response heterogeneity remained unclear.

One study from our research group[52] addressed this issue of response heterogeneity and identified a cluster of copy number alteration (CNA) on chromosome 7p associated with EGFR mutation status and response of EGFR-TKIs. In this study, we performed miRNA and gene expression profiling on the tumor samples from the same cohort. This information

allows us to study multilevel molecular events that differentiate between EGFR mutant and wild-type patients.

Differential expression (DE) analysis is widely used for identifying gene expressions with significant changes between two groups at the mean level. When identifying genes associated with complex diseases, like cancer, it is more insightful to investigate the differential co-regulations in additional to differential expression analysis[11]. We hypothesized that some differentially expressed genes were driven by the regulations of common cis-regulators where the activation/disruption of the regulations depended on EGFR mutation status. To identify these key cis-regulators, we proposed a modified liquid association score to quantify the strength of the changes in regulations.We also derived a statistical framework that combines differential expression analysis and differential regulation analysis to form an enrichment test for identifying critical network regulator.

Copy number and miRNA play important roles in transcriptional and post-transcriptional gene regulation. By integrating copy number, miRNA expression and gene expression profiles, our analysis on 177 lung adenocarcinomas shed light on a critical regulatory disparity between EGFR-mutated and wild-type lung adenocarcinomas.

4.2 Materials

4.2.1 Samples and Clinical Data

The investigation was approved by the Institutional Review Boards of both Nation Taiwan University Hospital (NTUH) and Taichung Veterans General Hospital (TCVGH) with written ethical consents from all patients. In total, frozen tumor tissues from 177 lung adenocarcinomas were collected by the lung cancer tissue lab with donors from both NTUH and TCVGH between 2000 to 2009. The median follow-up was 34.8 months; 57 patients died during follow-up. Table 4.1 summarizes demographic characteristics of the cohort.

Table 4.1: Patient characteristics

Characteristics	Patients
Age	year
Median (range)	64 (35-86)
Gender	n (%)
Male	86 (49)
Female	91 (51)
Smoking	n (%)
Ever	25 (14)
Never	118 (67)
Unknown	34 (19)
Stage	n (%)
I	87 (49)
II	31 (17.5)
III	46 (26)
IV	12 (7)
Unknown	1 (0.5)
EGFR activating mutations	n (%)
Wild type	75 (42)
L858R	67 (38)
Exon19 deletion	33 (19)
Both	2 (1)
Vital status	n (%)
Dead	57 (32)
Alive	120 (68)
Survival	Month
Median months to death (range)	34.8 (0-169.7)

4.2.2 EGFR Mutation Status

EGFR mutation status was identified by nucleotide mass spectrometry for all the samples. We utilized Matrix Assisted Laser Desorption Ionization Time-of-Flight Mass Spectrometry (MALDI-TOF MS) by using MassARRAY system (SEQUENOM, San Diego, CA). Biochemistry experiments were performed according to standard manual.

4.2.3 Gene Expression Array Analysis

Genome-wide messenger RNA expression was quantified by the Affymetrix U133plus2 array on 123 patients. MAS5.0 normalization was performed to scale the 2% trimmed-mean of the intensity to 500. Meanwhile, the scaling factor was used as a quality filter for samples with insufficient input. We excluded 13 samples with log2 scaling factor greater than 3.5 (Supplementary Fig. 4.8A). To reduce the chance of false positive from unexpressed genes, we removed the probesets whose average MAS5.0 normalized signals were less than 150 (the median value). The log2 MAS5.0 normalized data were used in the subsequent analyses.

4.2.4 Array CGH Analysis

DNA Copy number profiling by array CGH was performed on 145 samples. Following the standard protocol, the genomic DNA was extracted from frozen cancer tissue of each sample with quality checked by agarose electrophorese. The NimbleGen CGH array (NimbleGen Systems Inc, Madison, WI) containing 385, 806 probes spacing of about 6, 000 bp, was used. The normal DNA was extracted from the PBMC of one male and one female in a community cohort.

The log2 ratio of intensities from test channel (Cy3, 532) and normal reference channel (Cy5, 635) is designed to reflect relative copy number aberration (CNA) levels. Since the blood from the same two normal people was used as the reference, we calculated Pearson's correlation coefficient for each pair of the log2 reference intensities as a quality filter. We excluded seven samples which showed unusual low expression similarity (Supplementary Fig.

4.8B).

To remove the wave pattern, a widespread technical artifact highly thought to be GC content correlated, we performed a two-step GC correction procedure (Supplementary Materials 3.5.4, Supplementary Fig. 4.8C-D). To further reduce the random noise, we performed a local smoothing step by aggregate each 10 probesets as a probe-block and use the average value of the log2 ratio between two GC corrected intensities as the CNA level in the subsequent analyses.

4.2.5 miRNA Expression Real-Time PCR Analysis

We performed Applied Biosystems TaqMan assays (A-MEXP-1871), a real-time PCR assay, to quantify 365 human microRNA expressions on 107 samples. We excluded miRNAs with low and unreliable expression. By regarding the raw CT value to be lower than 35 in 99% or more samples across the cohort, we obtained 122 miRNAs to proceed with further analysis.

We used the CT value of the control probe RNU6B as the reference to calculate a relative expression (ΔCT). A systematic bias was observed in the distribution of pairwise correlations between miRNAs (Supplementary Fig. 4.9E). The overall correlations between all 122 microRNA probes were positively correlated with less than 0.5% pairs negatively correlated. To remove the systematic bias, for each sample, we performed a mean centering approach instead of using the control probe RNU6B for calculating the relative expressions. The distribution became normal now, and the right skewed has disappeared (Supplementary Fig. 4.9F).

4.3 Results

4.3.1 Differential Expression Analysis between EGFR Mutant and Wild-type Lung Adenocarcinomas

In the previous study[52], Yuan et. al. identified the differential copy number alteration (CNA) regions between EGFR mutant (MT) and wild-type (WT) patients. Chromosome 7p

was most enriched in containing sites of differential CNAs. To further explore the molecular variation between EGFR MT and WT patients at the RNA level, we performed a differential expression analysis. For each probeset, we utilized a two-sided student's t-test to examine the statistical significance of the mean difference in log2 expressions between two groups. Using $p < 0.05$ and absolute fold change > 1.5 as cutoffs, 693 probesets were differentially expressed (DE), where 558 were higher in mutant patients, and 135 were higher in wild-type patients. Chromosome 7p was the most enriched region of genes expressed significantly higher in EGFR mutant patients (Fig. 4.1). We also performed differential CNA analysis on this dataset. Among the 36,274 probe-blocks, using $p < 0.05$ as a cutoff, 3,577 probe-blocks were significant, where $2,714$ probe-blocks were higher in mutant patients, and 863 were higher in wild-type patients.

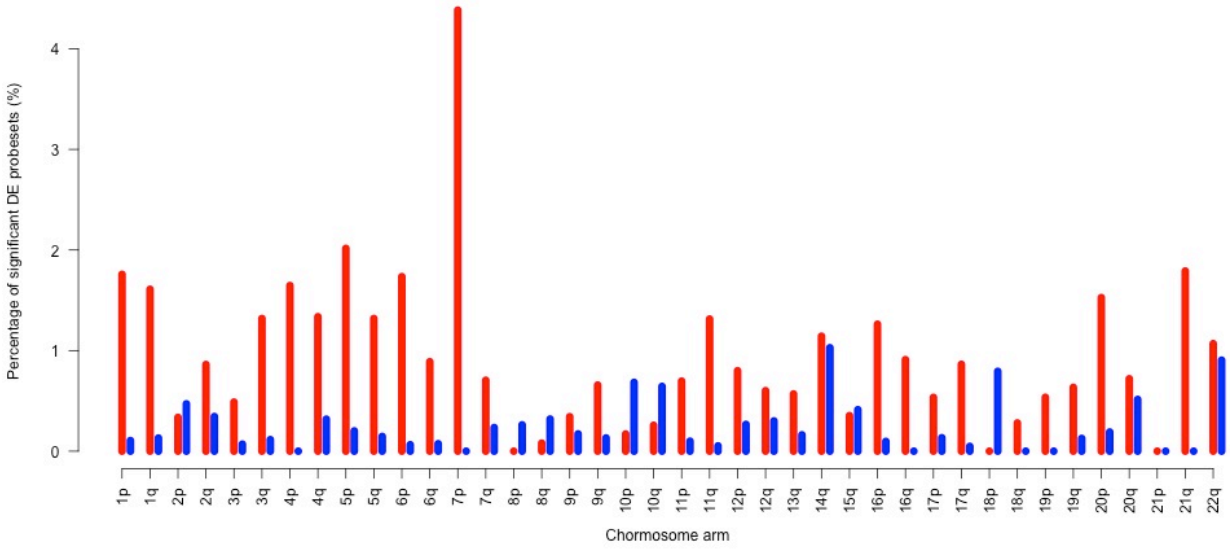

Figure 4.1: Chromosomal distributions of significant differentially expressed probesets.

4.3.2 Ternary Relationship of EGFR Mutation, Copy Number and Expression

Three molecular profiles: mutation, copy number, and gene expression of EGFR were performed in this cohort. Both copy number and gene expression were significantly higher for the EGFR mutant group (Fig. 4.2A-B). Going beyond the group mean comparison, we investigated how the covariation pattern between its copy number and gene expression may

vary. We found a significant positive correlation (Pearson's correlation coefficient = 0.41) in the mutant group but not in the wild-type group (Fig. 4.2C). This observation motivated the analysis we conducted in the next section.

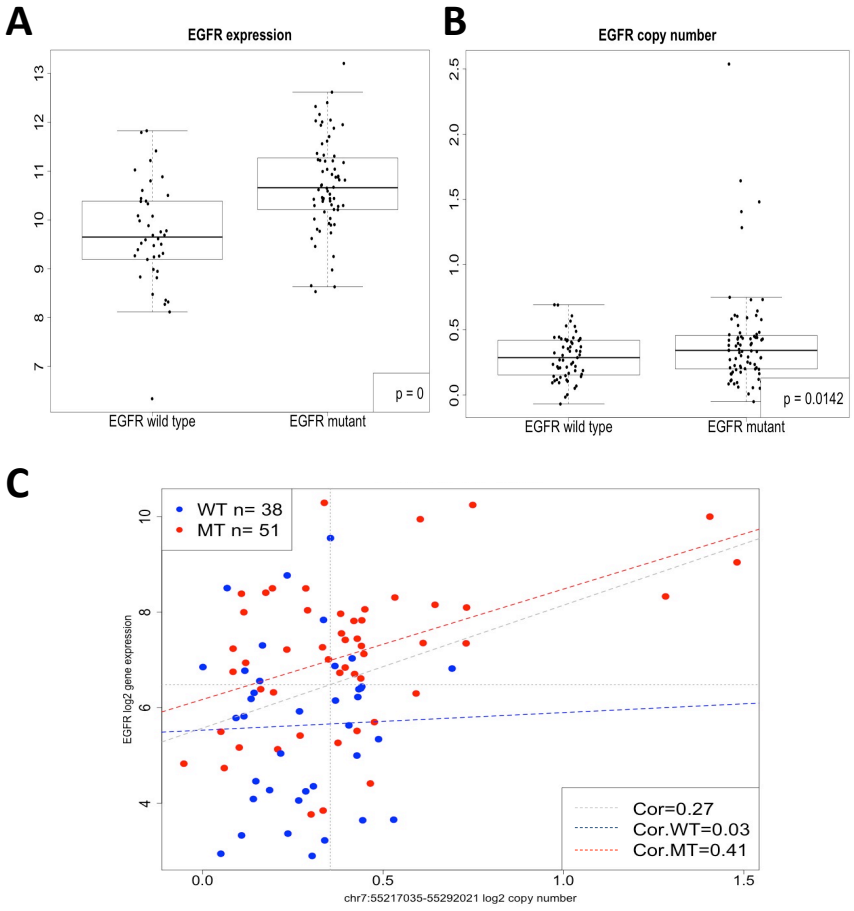

Figure 4.2: Ternary relationship of EGFR mutation, copy number and expression. The p values were calculated by the Welch two sample t-test.

4.3.3 Identification of Genes Differentially Regulated by the Cis-Copy Number Alteration between EGFR Mutant and Wild-Type Patients

Copy number alteration (CNA) plays an important role in gene regulation. It may affect the gene expression through several ways: changes in gene dosage, disruption of the gene, or alteration of the regulatory regions[48]. Due to the limited resolution of the array CGH data, we considered each probe-block as a regulator and the overlapped genes as its targets. We

define a gene to be differentially regulated by a cis-CNA between two groups if the change of covariation between groups is significant. We utilized a modified liquid association (LA) score to quantify the strength of changes, and the statistical significances were examined using a permutation test (Section 4.5.1).

Since the LA score only quantified the changes of covariation, to avoid selecting pairs with weak correlation in each group, we further required the absolute value of Pearson's correlation coefficients must be greater than a cutoff in at least one group. The cutoff was set to be $\tanh(\frac{1}{\sqrt{n_i-3}}z_{0.975})$ which resulted from converting the 0.05 significance level back to the correlation level using the inverse Fisher transformation, where $z_{0.975}$ is the 0.975 quantile of a standard normal distribution and n_i is the sample size of the ith group. Together, we called a gene to be differentially regulated (DR) by the cis-CNA regulator, if

i. the modified LA score is significant ($p < 0.05$, permutation test) and

ii. $|r_i| > \tanh(\frac{1}{\sqrt{n_i-3}}z_{0.975})$ in at least one group, where r_i is the sample Pearson's correlation coefficient in the ith group.

We aim to identify the differentially expressed/altered regulators that differentially regulated their target genes, and their target genes were also differentially expressed between groups. We only performed the differential regulation analysis on the DE genes that had differential cis-CNAs. By taking the intersection of both lists, 119 pairs were selected to proceed with differentially regulation analysis. Using criteria i and ii, we identified 4 significant pairs. On chromosome 7p, three regions with higher CNA in mutant patients positively regulated their target genes in mutant patients, and the target genes had higher expression in mutant patients. The fourth significant region was on chromosome 10q. It had a higher CNA in wild-type patients and positively regulated its target gene in wild-type patients. The target gene also had higher expression in wild-type patients (Fig. 4.3).

To investigate the clinical relevance of these four genes, we performed Kaplan-Meier analysis using their gene expressions. The impact on overall survival (OS) was examined. For each gene, patients were stratified into two groups by the median expression of that gene across the cohort. The over-expression of CXXX was found significantly correlated with

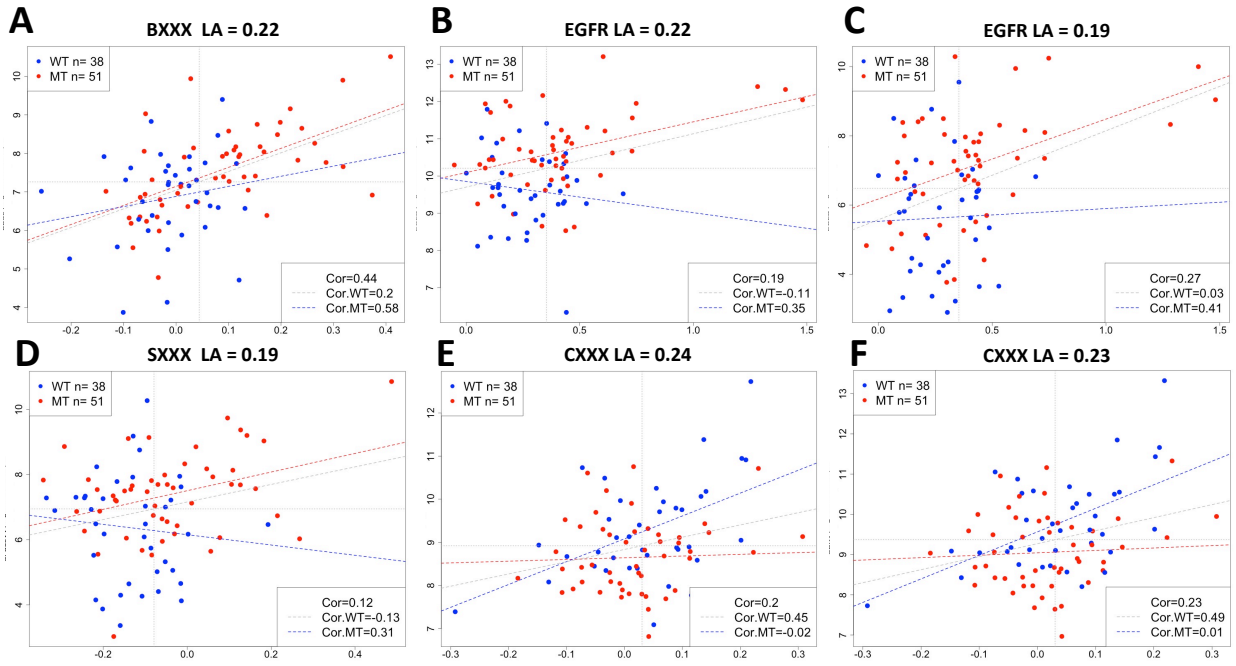

Figure 4.3: Differential regulation patterns of copy number alteration and the target gene.

poor overall survival. After adjusting for clinical covariates stage, gender, smoking status and age using a multivariate Cox regression model, it remained significant (Supplementary Table 4.2). This correlation was validated in other four independent validation cohorts and all validations were significant. This result showed the robust prognostic power of CXXX expression (Fig. 4.4). We did not find a significant correlation of overall survival with the other three gene expressions, including EGFR's expression. The workflow of differential regulation analysis on cis-copy numbers and the target genes is given in Supplementary Fig. 4.7.

4.3.4 Identification of Hub miRNAs Differential Regulating the Target Gene Expressions between EGFR Mutated and Wild-Type Patients

We took a similar approach to identify genes that differentially regulated by miRNAs. To proceed, we collected the miRNA-target links from the public database miRwalk2.0[15] to construct a background network. Both predicted and curated links were considered as candidates.

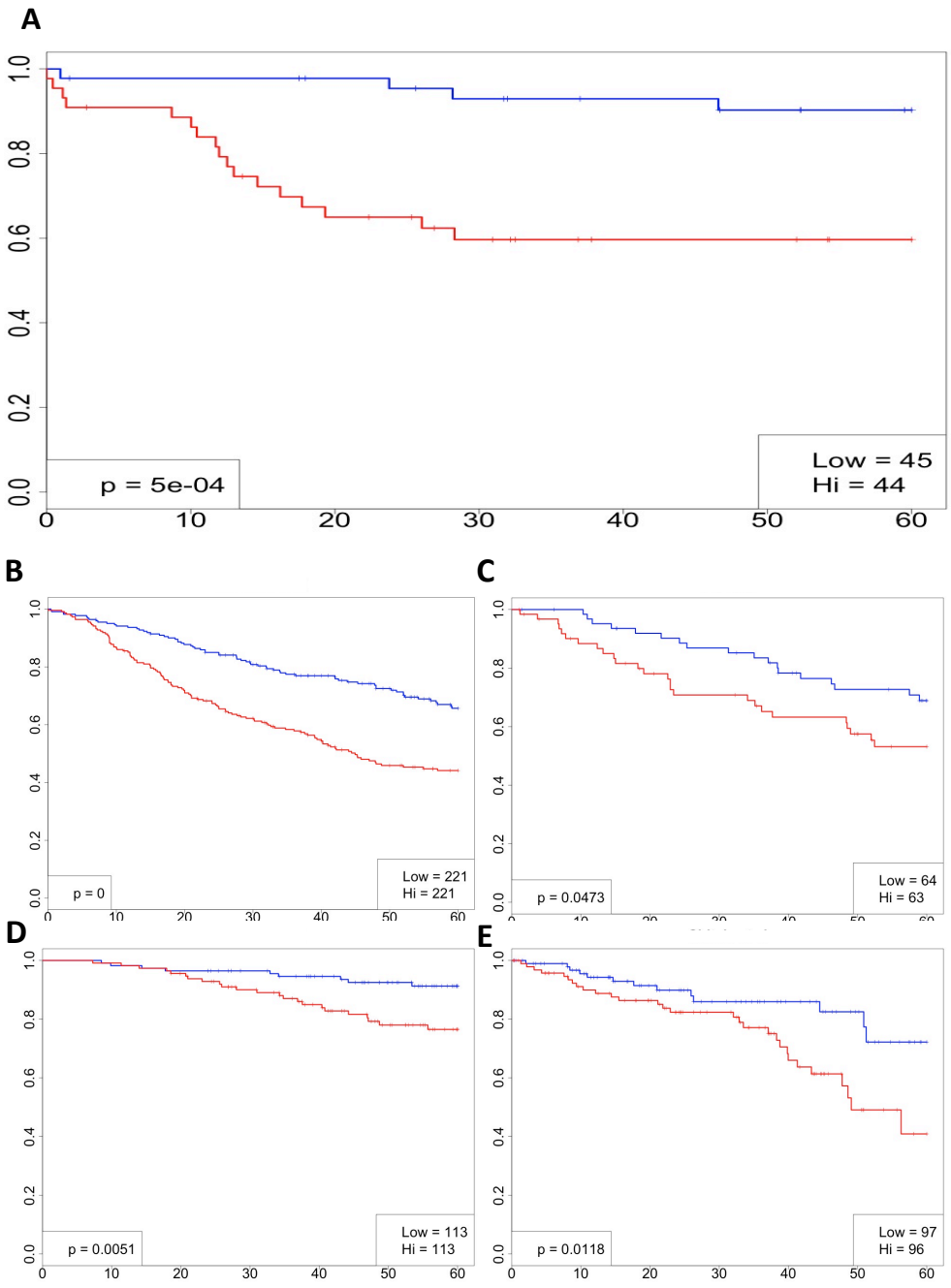

Figure 4.4: Kaplan-Meier analysis of OS by CXXX expression. (A) Our training cohort. (B) Independent validation cohort from Der et. al. 2014[13] (C) Independent validation cohort from Shedden et. al. 2008[45] (D)Independent validation cohort from Okayama et. al. 2012[38] (E) Independent validation cohort from Chitale et. al. 2009[7].

Differential expression analysis identified 29 DE miRNAs among 122 miRNAs, where 16 miRNAs expressed higher in mutant patients and 13 were higher in wild-type patients. Selecting only the pairs of DE miRNA and DE target in the background network, 10,035 miRNA-target pairs (in 27 miRNAs and 466 targets) were selected to perform differentially regulation analysis. Using criteria i and ii, we identified 732 significant miRNA-target pairs (in 27 miRNAs and 293 targets).

Because the number of DE target genes for each miRNA varies, we proposed a Fisher exact test to if there is an enrichment of differential regulated targets. Specifically, given a regulator, we tested whether miRNA differentially regulated targets have a higher chance to be differentially expressed. A 2 by 2 contingency table of the number of target genes differentially regulated against differentially expressed was created for applying one-sided Fisher's exact test. A cartoon example is given in Supplementary Fig. 4.12. We defined a regulator that passed the test of enrichment to be a hub regulator. We identified 7 miRNAs as hub regulators ($p < 0.05$), including miRXX3a (OR:6.36, p <0.0001), miRXX2 (OR:4.52, p <0.0001), miRX5a (OR:2.56 p =0,0084), miRXX5a (OR:2.40, p =0.0174), and three miRNAs in the miRXX cluster: miRX7 (OR:1.93, $p = 0.0016$), miRX8a (OR:2.30, p =0.0002) and miRX9b (OR:2.231 p =0.0041). The target gene expressions differentially regulated by miRXX3a was visualized by a heat map (Fig. 4.13). The hub regulators with all their differentially regulated DE targets were collected and visualized as a network (Fig. 4.5). This network contained 112 (22%) genes out of 507 DE genes. All the genes in this network can be linked to at least one key regulators including 2 copy number alteration regions, and 7 miRNAs.

We again performed Kaplan-Meier analysis to examine the impact to overall survival for each hub regulators. Three hub miRNA, miR-XX3a, miR-X7, and miR-X9b were significantly correlated with poor overall survival (Fig. 4.6). After adjusting for clinical covariates stage, gender, smoking status and age using a multivariate Cox regression model, they remained significant (Supplementary Table 4.3-4.5). The full workflow of differential regulation analysis is given in Supplementary Fig. 4.11.

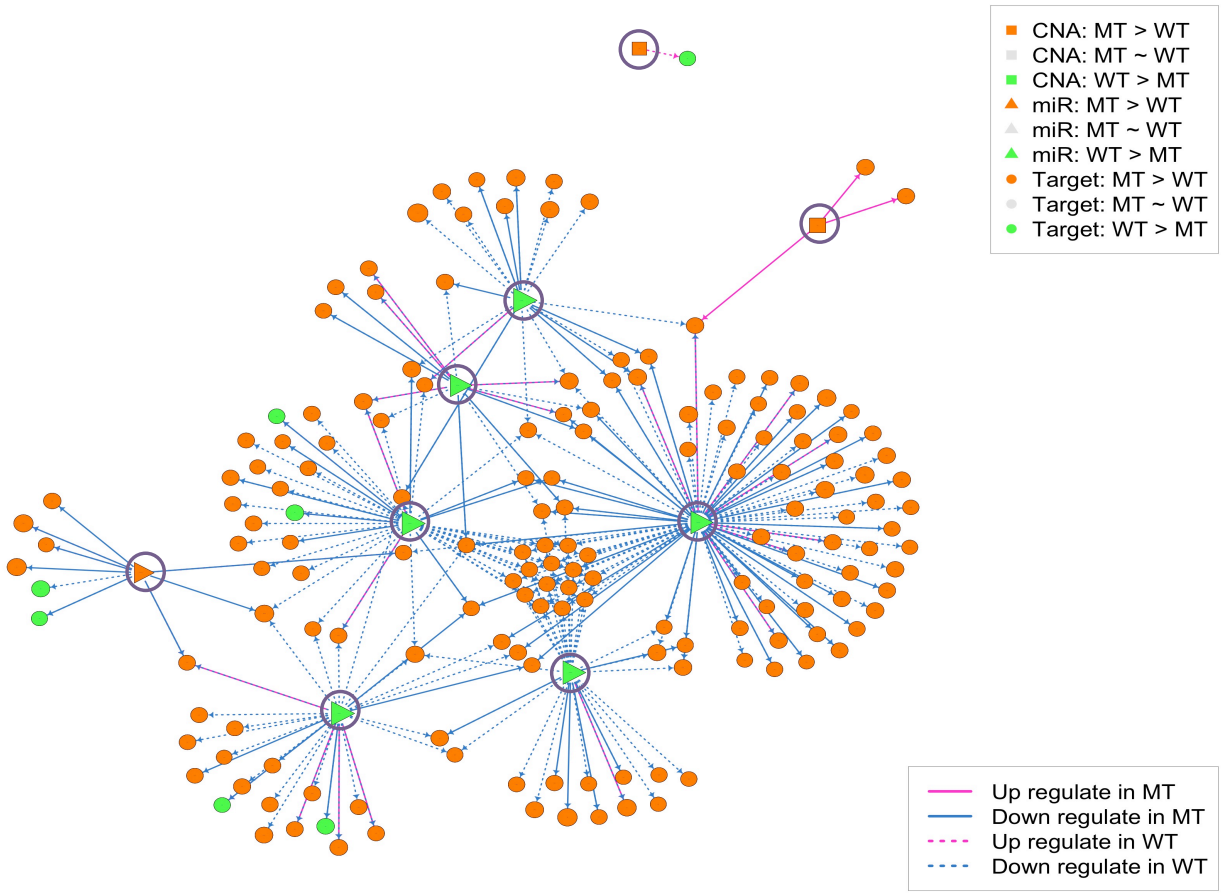

Figure 4.5: Differential regulatory network in lung adenocarcinomas between EGFR mutant and wild-type patients. The hub regulators were circled including 2 copy number alteration regions, 1 transcription factor, and 6 miRNAs.

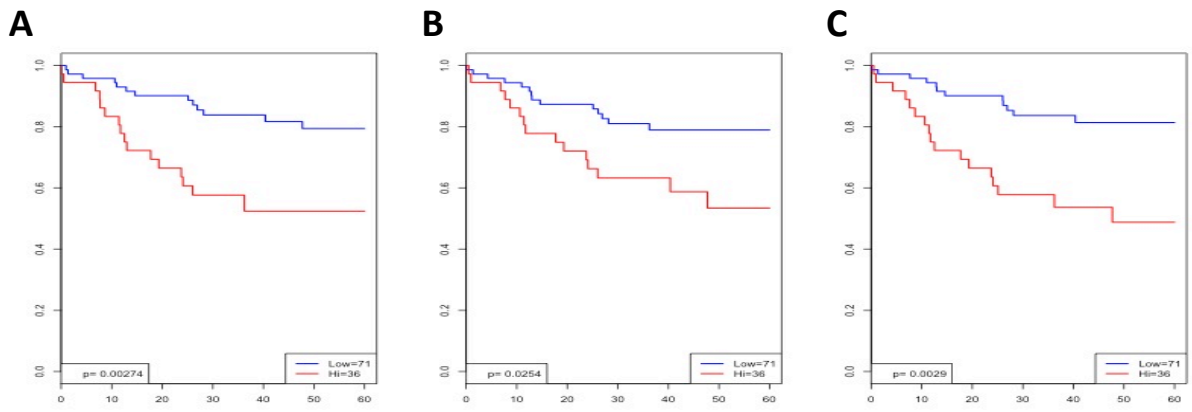

Figure 4.6: Kaplan-Meier analysis of OS by (A) miRXX3a, (B) miRX7, and (C) miRX9b.

4.4 Discussion

In this integrative study of copy number, miRNA expression, and gene expression profiles, we identified 507 differentially expressed genes between EGFR mutated and wild-type patients. We derived a statistical framework to combine differential expression analysis and differential regulation analysis. By utilizing this approach, we identified 4 DE genes that were differentially regulated by their cis-CNAs and 7 DE miRNAs that differentially regulated 109 DE genes. In total, 22% of the DE genes were linked to at least one of these key regulators. Among these 9 key regulators, 1 gene and 4 miRNA were found to correlate patients' overall survival.

It is noteworthy that due to the nature of statistical association, great caution has to be taken in any causality interpretation. The directions of our reported regulatory network were inherited directly from the links of the regulators and their cis-target. We note that our differential regulation analysis did not address the issue of whether the activation/disruption was driven by EGFR mutation. We only demonstrated the different patterns of regulations between two groups. Nonetheless, these findings may provide better understanding of the TKI-response heterogeneity between EGFR-mutant and wild-type patients and help improve the patient management in lung adenocarcinoma.

4.5 Methods

4.5.1 Pre-Centered Liquid Association Score

Liquid association (LA) is a novel notion of statistical association between variables, initially conceptualized for the purpose of uncovering how the strength and the sign of association between two variables may be varying and driven dynamically via unknown sources (Li 2002). In the original derivation, the liquid association score of a pair of variables (X, Y) with $E[X] = E[Y] = 0$ and $Var(X) = Var(Y) = 1$ mediated by a third variable Z (called the scouting variable) is defined as $LA(XY|Z) = E[g'(Z)]$, where $g(z) = E[XY|Z = z]$. Suppose Z follows standard normal distribution, by using Stein's lemma, we have $LA(XY|Z) =$

$E[XYZ]$.

In many applications, the scouting variable Z might be a binary variable, for example, the EGFR mutation status in this case. Sun et. al. [49] defined the binary LA score as

$$LA(XY|Z) = \frac{E[XY|Z=a] - E[XY|Z=b]}{a-b}. \tag{4.1}$$

To scale the scouting variable properly, they set $a = \sqrt{q/p}$ and $b = -\sqrt{p/q}$ where $p = Pr(Z = a)$ and $q = 1 - p$. This standardized the scouting variable, such that $E[Z] = 0$ and $Var(Z) = 1$. The binary LA score for a standardized binary scouting variable can be written as

$$LA(XY|Z) = \sqrt{pq}[E[XY|Z=a] - E[XY|Z=b]] = E[XYZ]. \tag{4.2}$$

LA was motivated originally by triplets where their pairwise correlations are not significant. However, in this study, both differential copy number alteration and gene expression are required before proceeding the LA analysis. Thus, the conditional means of the variables X and Y were affected by Z. To adjust for the additional prerequisites, we pre-centered variables X and Y within each group separately before computing LA score, ie

$$LA(XY|Z) \;\; = \;\; E[(X - E[X|Z])(Y - E[Y|Z])Z]. \tag{4.3}$$

For implementation, we first centered variables X and Y in both groups separately then performed the original liquid association score. Re-centering adjusted to LA is illustrated by a simulated example (Supplementary Fig. 4.10).

The statistical significance was examined by a permutation test for each regulator-target pair. Because the purpose is to test whether the change of variation depends on the scouting variable only the index of the scouting variable needs to be permuted, keeping the indexes the regulator and it targets. After N times of permutation, the p value was estimated by finding the proportion of permutations yielding absolute LA score larger than observed.

4.6 Supplementary Tables and Figures

Table 4.2: Multivariate Cox regression analysis of OS by CXXX expression.

Variable	Hazard ratio	95%CI	p
CXXX:High	4.86	1.53-15.40	0.007
Gender:Male	4.25	1.25-14.48	0.021
Age	1.00	0.96-1.05	0.901
Smoking:Ever	0.40	0.12-1.34	0.137
Stage:II	1.76	0.37-8.25	0.476
Stage:III	2.83	0.88-9.16	0.082
Stage:IV	5.40	1.26-23.10	0.023
EGFR:Mutant	0.74	0.28-1.93	0.540

Table 4.3: Multivariate Cox regression analysis of OS by miR-XX3a expression.

Variable	Hazard ratio	95%CI	p
miR-XX3a:High	3.50	1.15-10.60	0.027
Gender:Male	1.51	0.46-4.96	0.495
Age	1.00	0.95-1.05	0.921
Smoking:Ever	0.98	0.33-2.94	0.978
Stage:II	3.21	0.82-12.67	0.095
Stage:III	1.47	0.46-4.78	0.517
Stage:IV	2.93	0.64-13.40	0.167
EGFR:Mutant	1.54	0.47-5.01	0.477

Table 4.4: Multivariate Cox regression analysis of OS by miR-X7 expression.

Variable	Hazard ratio	95%CI	p
miR-X7:High	2.95	1.10-7.88	0.031
Gender:Male	1.28	0.36-4.5	0.704
Age	1.00	0.95-1.05	0.899
Smoking:Ever	1.07	0.33-3.47	0.906
Stage:II	2.34	0.59-9.25	0.225
Stage:III	2.00	0.56-7.10	0.286
Stage:IV	5.46	1.27-23.48	0.023
EGFR:Mutant	1.23	0.38-3.97	0.731

Table 4.5: Multivariate Cox regression analysis of OS by miR-X9b expression.

Variable	Hazard ratio	95%CI	p
miR-X9b:High	3.40	1.29-8.96	0.013
Gender:Male	1.10	0.32-3.74	0.877
Age	0.99	0.94-1.05	0.762
Smoking:Ever	1.36	0.46-4.03	0.582
Stage:II	2.84	0.73-11.00	0.132
Stage:III	1.67	0.49-5.70	0.411
Stage:IV	3.62	0.82-16.00	0.090
EGFR:Mutant	1.29	0.42-4.01	0.660

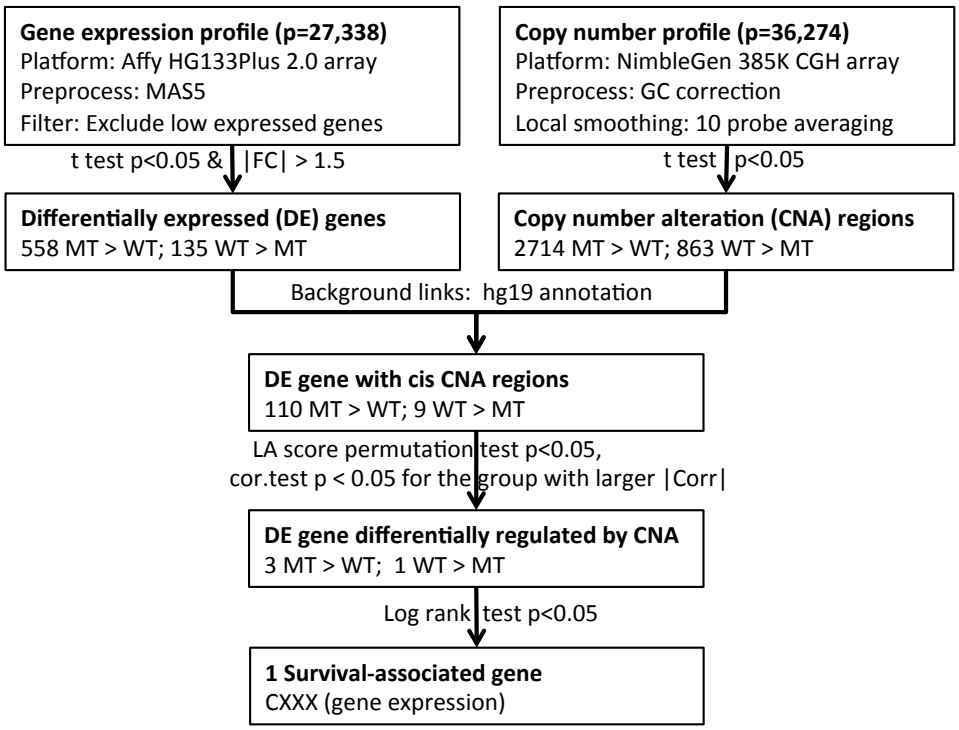

Figure 4.7: Workflow of identifying genes differentially regulated by the cis-copy number alteration between EGFR mutant and wild-type patients

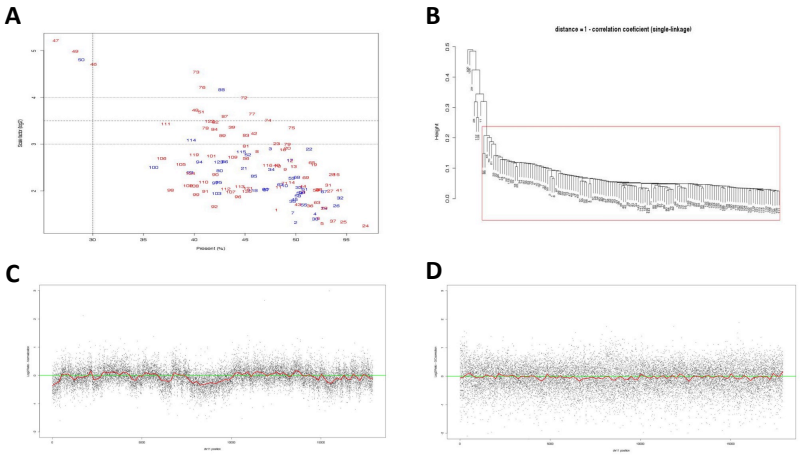

Figure 4.8: (A) A scatter plot of two quality control statistics: log2 scale factor against percentage of present. Thirteen sample with log2 scale factor > 3.5 were excluded. (B) A single-linkage hierarchical clustering based on the distance = 1-Pearson's correlation coefficient between the intensities of the common reference channel in different CGH arrays. Seven samples were considered as outliers and excluded. (C)Visualization of the waved bias pattern in real data, and (D) the null pattern after performing the GC correction.

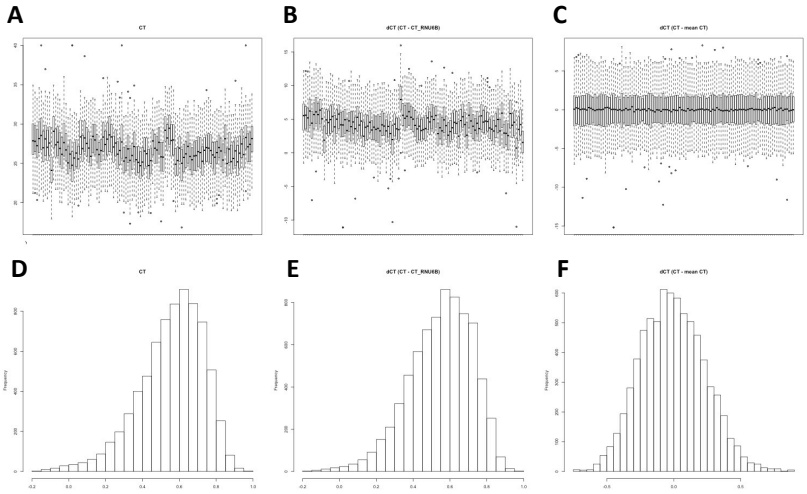

Figure 4.9: Mean centering approach removed the systematic bias. (A-C) Distributions of miRNA abundances for each sample: (A) raw CT value, (B) ΔCT value using RNU6B as the reference (C) ΔCT value using the average CT in each assay as the reference. (D-F) Histograms of pairwise correlations between miRNAs: (D) raw CT value, (E) ΔCT value using RNU6B as the reference (F) ΔCT value using the average CT in each assay as the reference.

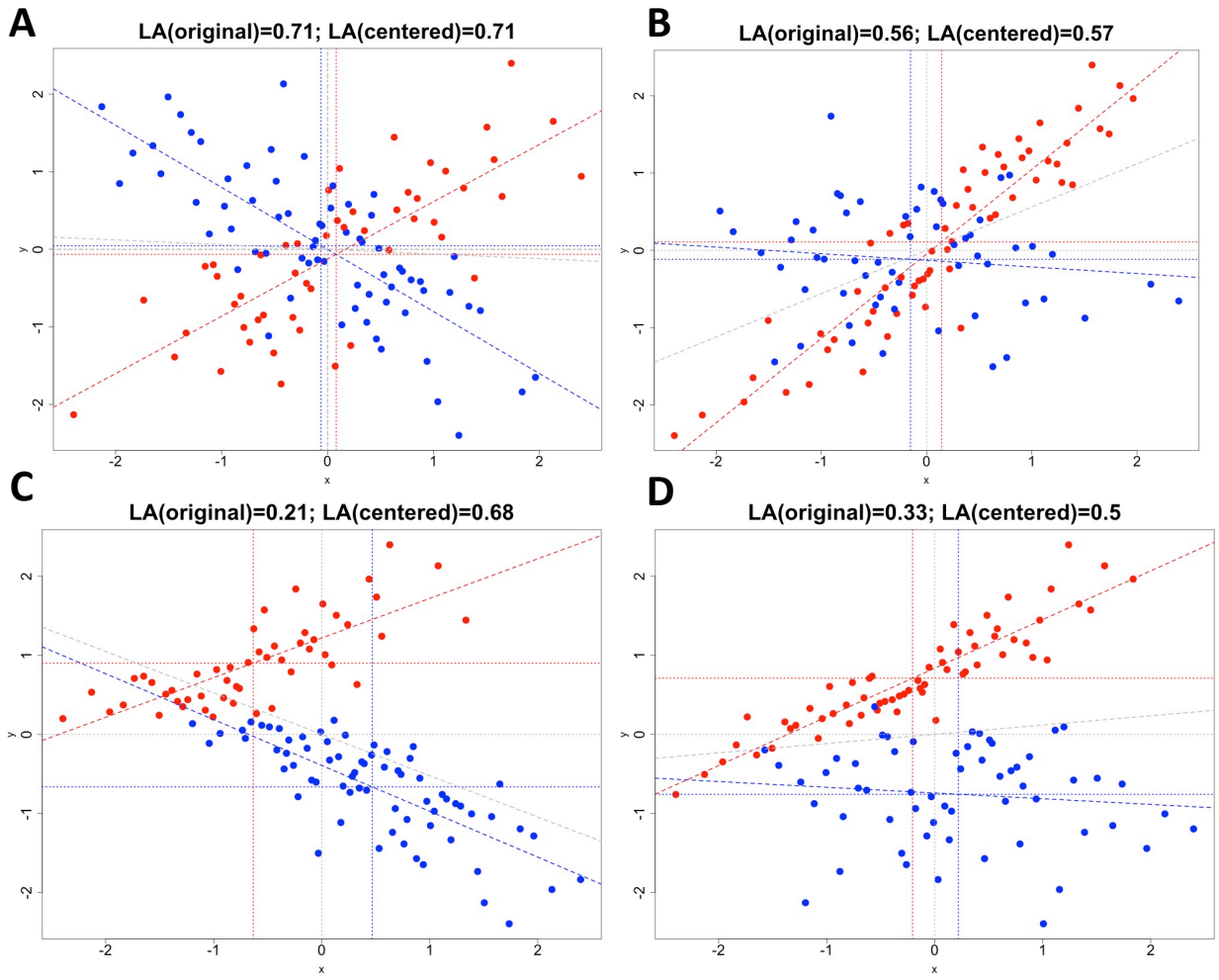

Figure 4.10: Visualizations of the original and pre-centered liquid association in simulated examples. (A) (X, Y) follows a mixture bivariate normal model where $Cov(X, Y|Z = 1) > 0$, $Cov(X, Y|Z = 0) < 0$, and $E[X|Z = z] = E[Y|Z = z] = 0$ for $z = 0, 1$. (B) (X, Y) follows a mixture bivariate normal model where $Cov(X, Y|Z = 1) > 0$, $Cov(X, Y|Z = 0) = 0$, and $E[X|Z = z] = E[Y|Z = z] = 0$ for $z = 0, 1$. In (C) and (D), (X, Y) was generated from the same distribution as (A) and (B) with a mean shift (μ_{1x}, μ_{1y}) for $Z = 1$ and another (μ_{0x}, μ_{0y}) for $Z = 0$.

,

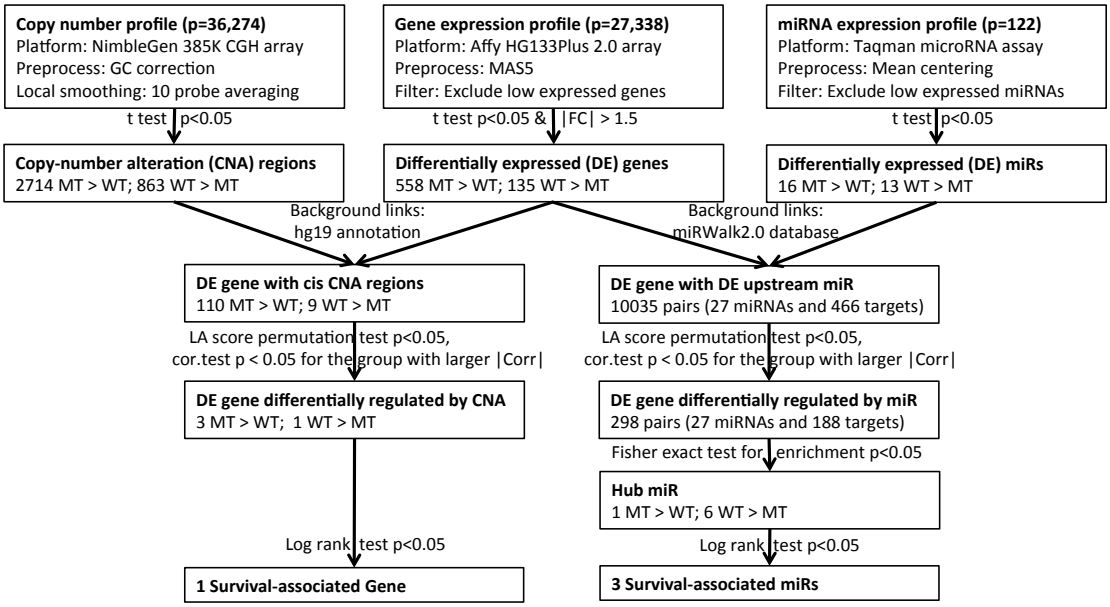

Figure 4.11: Full workflow of identifying the hub miRNAs and copy number alterations which differentially regulated the target gene expressions between EGFR mutant and wild-type patients

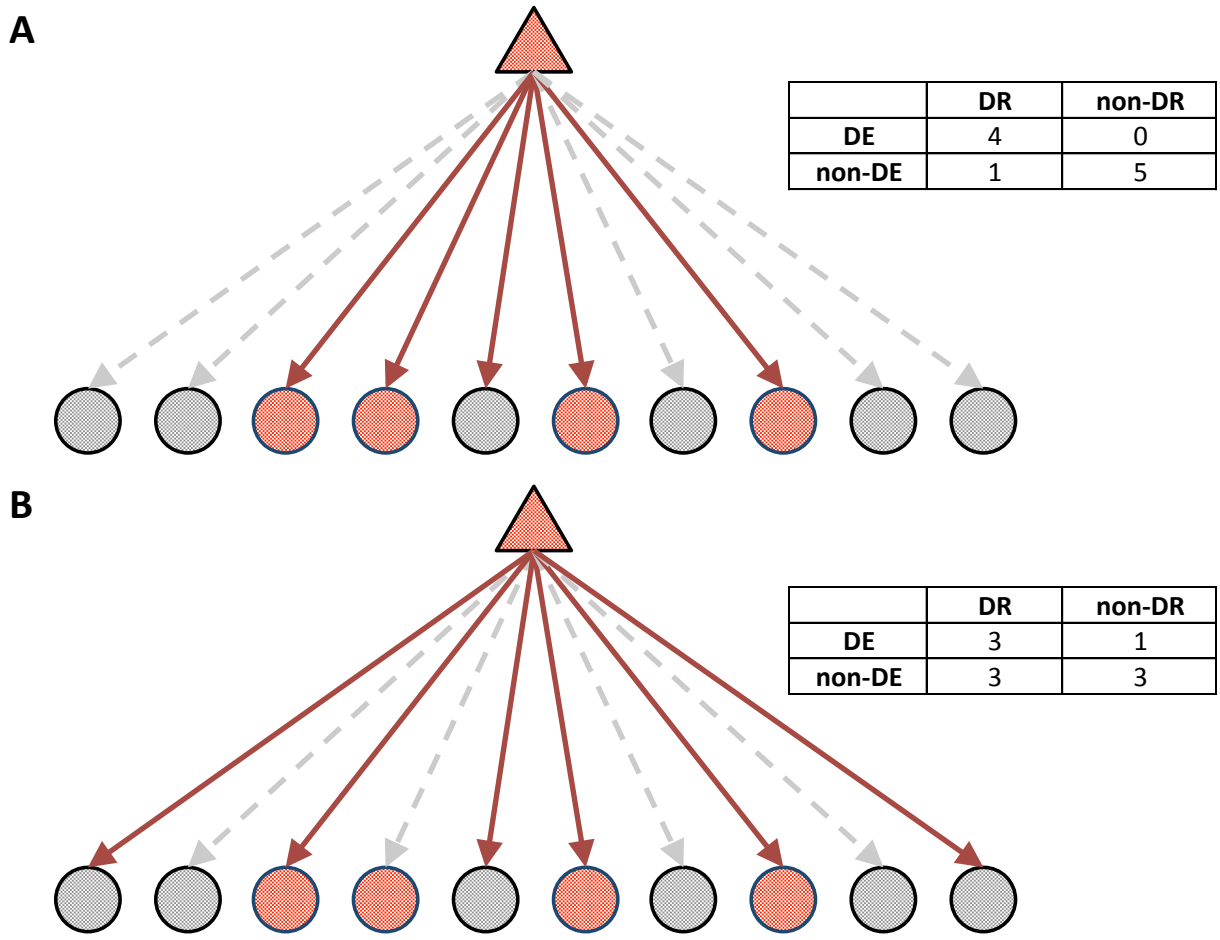

Figure 4.12: Two cartoon examples: (A) a significant master regulator and (B) an insignificant regulator. A red circle indicates the target gene was differentially expressed where a gray arrow indicates not. A red arrow indicates the target gene was differentially regulated by the regulator where a gray arrow indicates not. One-sided Fisher's exact test was used to test for enrichment using the two by two contingency table where the null hypothesis is H_0 : whether a target was differentially expressed does not depend on whether a target was differentially regulate by the regulator.

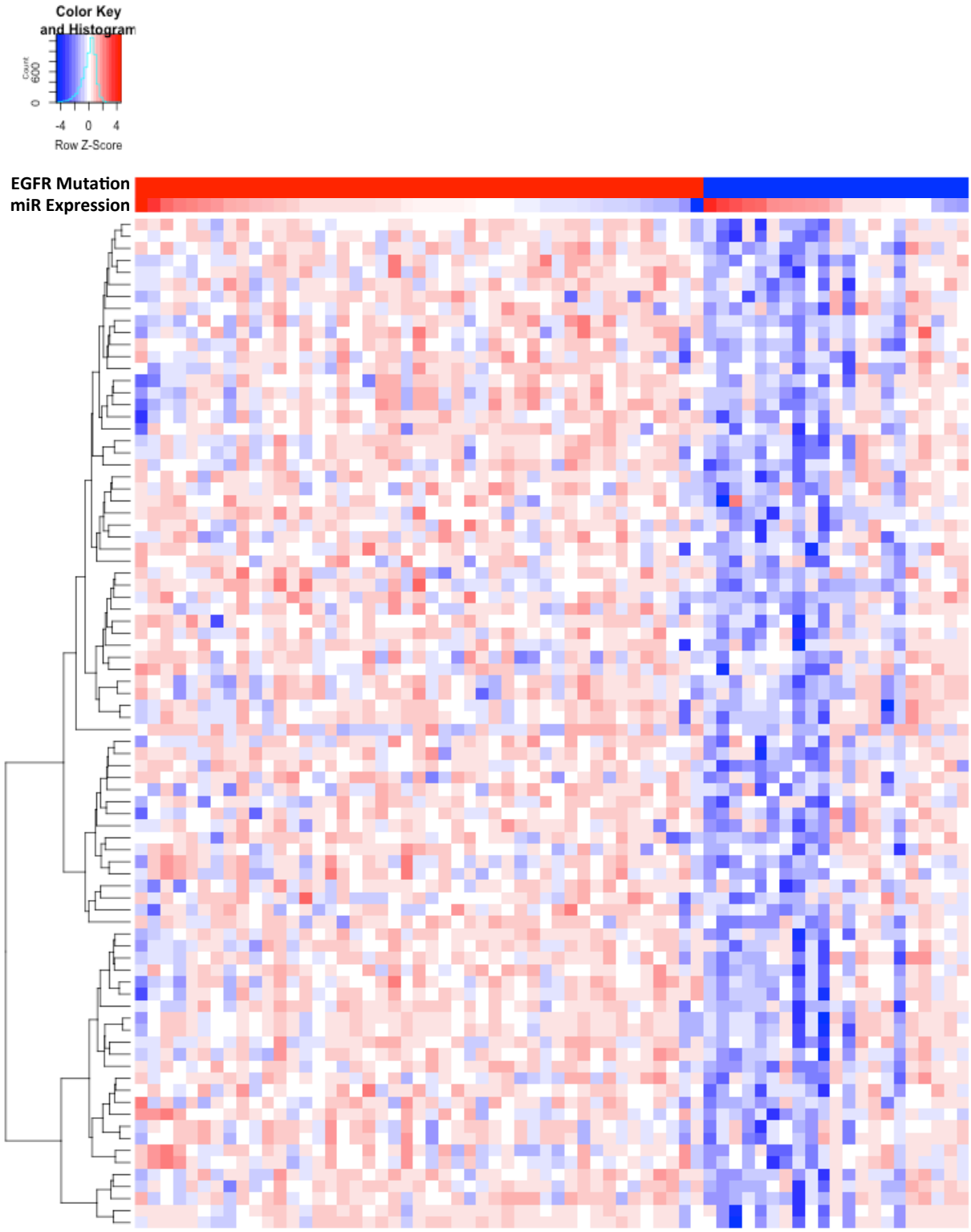

Figure 4.13: A heat map of the target gene expressions differentially regulated by miRXX3a.

CHAPTER 5

Liquid Association Adjusted by Background Variables

5.1 Introduction

Liquid association (LA) was introduced to study dynamic patterns of co-expression between genes (Li 2002, PNAS[34]). LA depicts the change in the covariation of two variables X and Y as a third variable Z varies. A positive LA score indicates an increasing trend of covariation between X and Y as Z increases while a negative LA score indicates the opposite trend. The formula of computing LA scores is simple. In a system with p variables of interest, the computation of $\binom{p}{3}$ LA scores, one for each triplet of variables, can be parallelized. However, in many applications, there often exist additional background variables such as the demographical variables in cancer genomic studies. Their presence may complicate the study of the relationship between the main variables of interest. In this chapter, effects of ignoring background variables will be discussed, and some adjustment methods to marginalize their influence will be presented.

5.2 Methodology

The original liquid association score of a pair of variables (X, Y) mediated by another variable Z is defined as

$$LA(X, Y|Z) = E[g'(Z)] \tag{5.1}$$

where $g(z) = E[XY|Z = z]$. Suppose Z follows standard normal distribution. By using Stein's lemma, we have

$$LA(X, Y|Z) = E[g'(Z)] = E[g(Z)Z] = E[E[XY|Z]Z] = E[XYZ]. \tag{5.2}$$

When there is a background variable U affecting the system, ignoring U might be inappropriate. We define the adjusted liquid association score as

$$LA(X,Y|Z\|U) = E[\frac{\partial}{\partial z}h(Z,U)] \tag{5.3}$$

where $h(z,u) = E[XY|Z=z,U=u]$.

Lemma 5.2.1 *$LA(X,Y|Z\|U) = LA(X,Y|Z) - E[XYs(Z,U)]$, where $s(z,u) = \frac{\partial}{\partial z}\log f_{U|Z}(u|z)$ and $f_{U|Z}(u|z)$ is the conditional density function of U given $Z=z$.*

Corollary 5.2.1.1 *Suppose Z follows a standard normal distribution. By using Stein's lemma, we have*

$$LA(X,Y|Z\|U) = E[XYZ] - E[XYs(Z,U)] = E\{XY[Z - s(Z,U)]\}. \tag{5.4}$$

The definition of the adjusted LA score and Lemma 5.2.1 provides two situation where adjustment is not needed: (A) (X,Y) and U are conditionally independent given Z. The equations 5.1 and 5.3 are identical because $h(z,u)$ equals $g(z)$. (B) Z and U are independent. The original and adjusted LA scores are identical because the second term, $E[XYs(Z,U)]$, in Lemma 5.2.1 vanishes. In practice, it may not be easy to check the validity of (A). However, (B) suggests that we do not need to perform the adjustment for all the background variables except for those correlated with Z. Thus, a pre-screening step to eliminate the background variables independent of Z can help reduce the number of background variables to proceed with the LA adjustment.

One critical situation requiring adjustment is when Z and (X,Y) are conditional independent given U. In this case, $h(z,u) = E[XY|Z=z,U=u] = E[XY|U=u]$ does not depend on z. We have

$$0 = LA(X,Y|Z\|U) = LA(X,Y|Z) - E[XY\frac{\partial}{\partial z}\log f_{U|Z}(U|Z)]. \tag{5.5}$$

This implies $LA(X,Y|Z) = E[XY\frac{\partial}{\partial z}\log f_{U|Z}(U|Z)]$. It shows that even the covariation of X and Y only depend on the background variable U, we might still observe high LA score

$LA(X, Y|Z)$ due to the dependency between Z and U. The proposed adjustment corrects the LA score to zero rightfully.

When the covariation of X and Y depends on both Z and U, the original LA scoring method may obscure the true correlation pattern changes. For example, suppose Z and U are negative correlated and the covariation of (X, Y) depends on an increasing function of both Z and U marginally. We expect to observe a positive LA score, but the original LA score might turn out close to 0 or even be negative. The new adjusted LA score will capture the correct direction.

5.3 Estimation

In general, to compute the adjusted LA score, nonparametric methods can be applied to estimate the partial derivative of the log conditional score function. However, with two common parametric distribution assumptions on the background variables, we can derive closed form formulas for estimating $LA(X, Y|Z\|U)$.

5.3.1 Bivariate Normal Model

Suppose (Z, U) follows the bivariate normal distribution with $EZ = EU = 0$, $Var[Z] = Var[U] = 1$ and $Cov(Z, U) = \rho$. We have

$$\frac{\partial}{\partial z} \log f_{U|Z}(U|Z) = \frac{\partial}{\partial z} \log\{\frac{1}{\sqrt{2\pi(1-\rho^2)}} \exp[\frac{-(u-\rho z)^2}{2(1-\rho^2)}]\} = \frac{\rho(u-\rho z)}{1-\rho^2}$$

and

$$LA(X, Y|Z\|U) = E[XY(\frac{Z-\rho U}{1-\rho^2})]. \tag{5.6}$$

Therefore, we can estimate $LA(X, Y|Z\|U)$ by

$$\widehat{LA}(X, Y|Z\|U) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i(\frac{z_i - \hat{\rho} u_i}{1-\hat{\rho}^2}) \tag{5.7}$$

where $\hat{\rho}$ is the Pearson's sample correlation coefficient between Z and U.

5.3.2 Logistic Regression Model

When U is a binary variable, the conditional distribution of U given Z is often modeled by logistic regression. We then have

$$\frac{\partial}{\partial z}\log f_{U|Z}(U|Z) = \frac{\partial}{\partial z}\log\{(\frac{e^{\alpha+\beta z}}{1+e^{\alpha+\beta z}})^u(\frac{1}{1+e^{\alpha+\beta z}})^{1-u}\} = \beta(u - \frac{e^{\alpha+\beta z}}{1+e^{\alpha+\beta z}})$$

and

$$LA(X,Y|Z\|U) = E[XY(Z - \beta(U - \frac{e^{\alpha+\beta Z}}{1+e^{\alpha+\beta Z}}))]. \tag{5.8}$$

Therefore, we can estimate $LA(X,Y|Z\|U)$ by

$$\widehat{LA}(X,Y|Z\|U) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i (z_i - \hat{\beta}(u_i - \hat{p}_i)) \tag{5.9}$$

where $\hat{\beta}$ is the estimate of the slope and $\hat{p}_i$ is the fitted value.

5.4 Simulation

We conducted a comprehensive simulation study to demonstrate how this approach adjusts the effect of the background variable to the LA score in various situations. First, we generated $\{Z_i\}_{i=1}^n$ independently from the standard normal distribution. Parametric models were utilized for generating background variables. To generate continuous $\{U_i\}_{i=1}^n$, a bivariate normal model was used, where

$$U_i = \beta Z_i + \sqrt{1-\beta^2}\epsilon_i \text{ and } \epsilon_i \sim N(0,1). \tag{5.10}$$

To generate binary $\{U_i\}_{i=1}^n$, a logistic regression model was used, where

$$U_i|Z_i = z_i \sim \text{Ber}(p_i) \text{ and } p_i = \frac{e^{\beta z_i}}{1+e^{\beta z_i}}. \tag{5.11}$$

To illustrate the effect of correlation between Z and U, three levels of β were used for each model. Lastly, we generated $\{(X,Y)_i\}_{i=1}^n$ from a conditional bivariate normal model with $E[X_i] = E[Y_i] = 0$, $Var[X_i] = Var[Y_i] = 1$ and the covariance took three values under three

different states depending on a third variable W_i

$$Cov[X_i, Y_i|W_i] = \begin{cases} 0.8, & \text{if } W_i \geq 1 \\ 0, & \text{if } -1 \leq W_i < 1 \\ -0.8, & \text{if } W_i < -1. \end{cases}$$

The variable W_i was set to be the linear combination of Z_i and U_i and three sets of coefficients were considered. The specific settings were summarized in Table 5.1.

Bivariate normal model				Logistic regression model			
β	W	LA score	Adjusted	β	W	LA score	Adjusted
0	Z	0.386(0.012)	0.386(0.012)	0	Z	0.386(0.011)	0.386(0.011)
0	U	0.001(0.011)	0.001(0.011)	0	$2U-1$	-0.002(0.016)	-0.001(0.013)
0	$Z+2U$	0.255(0.014)	0.256(0.013)	0	$Z-4U+2$	0.196(0.012)	0.195(0.012)
0.6	Z	0.387(0.011)	0.388(0.014)	1	Z	0.387(0.011)	0.386(0.011)
0.6	U	0.232(0.013)	0(0.014)	1	$2U-1$	0.331(0.014)	0(0.014)
-0.6	$Z+2U$	-0.064(0.013)	0.327(0.015)	1	$Z-4U+2$	-0.004(0.012)	0.29(0.012)
0.9	Z	0.387(0.011)	0.387(0.025)	2	Z	0.386(0.011)	0.386(0.014)
0.9	U	0.347(0.012)	0.003(0.024)	2	$2U-1$	0.483(0.012)	0.001(0.017)
-0.9	$Z+2U$	-0.301(0.012)	0.376(0.025)	2	$Z-4U+2$	-0.107(0.012)	0.35(0.016)

Table 5.1: Settings of models and parameters for the simulations.

For each setting, we simulated 10,000 observations (n) and computed the original and the adjusted LA scores for 100 times (N). The mean and standard deviation were summarized in Table 5.1. We observed the following.

i. When $\beta = 0$, Z and U were independent. The original and adjusted LA score were nearly identical.

ii. When $W = Z$, (X, Y) and U are conditionally independent given Z. The original LA score and adjusted LA score are also identical.

iii. When $W = U$, the covariation of X and Y only depends on U. Large positive original LA scores were still observed due to the dependency between Z and U. The adjustment eliminated the background variable effect and had scores close to 0.

iv. When Z and U had a strongly negative correlation, even $Cov(X, Y|Z, U)$ was an increasing function for both Z and U marginally, we still observed a zero or even negative original LA score. The adjusted LA score captured the positive direction.

To understand the patterns beyond the LA scores, we visualized some typical examples (Fig. 5.1). Each setting had three scatter plots of $\{(x, y)_i\}_{i=1}^n$ colored by $\{z_i\}_{i=1}^n$ in all observations and $\{z_i - s(z_i, u_i)\}_{i=1}^n$ in observations with $u_i = 1$ or $u_i = 0$ only. In figure 5.1A, all three scatter plots had similar patterns. In figure 5.1B, the left panel showed a typical positive liquid association pattern where the red points had a positive correlation, and the green points had a negative correlation. However, the two right panels showed that the positive and negative covariation depended on the background variable U. Moreover, the covariation did not depend on the level of $\{z_i - s(z_i, u_i)\}_{i=1}^n$ within each state of U. In figure 5.1C, the left panel showed no liquid association pattern. However, the two right panels showed that the positive and negative covariation depended on the background variable U but the covariation depended on the level of $z_i - s(z_i, u_i)$ within each state of U. Specifically, for $U = 1$, a negative covariation was found only when $z_i - s(z_i, u_i)$ was low. Similarly, for $U = 0$, a positive covariation was found only when $z_i - s(z_i, u_i)$ was high.

5.5 Application: EGFR Expression as a Scouting Variable Adjusted by EGFR Mutation Status as a Background Variable

We performed the adjusted LA analysis on a dataset of 122 Taiwanese lung adenocarcinomas (Chapter 4). Gene expression profiles were measured by Affy U133plus2 array containing $26,291$ probesets after the data preprocessing and screening steps. EGFR mutation status was identified by nucleotide mass spectrometry.

Here we used EGFR gene expression as the mediating variable Z and EGFR mutation

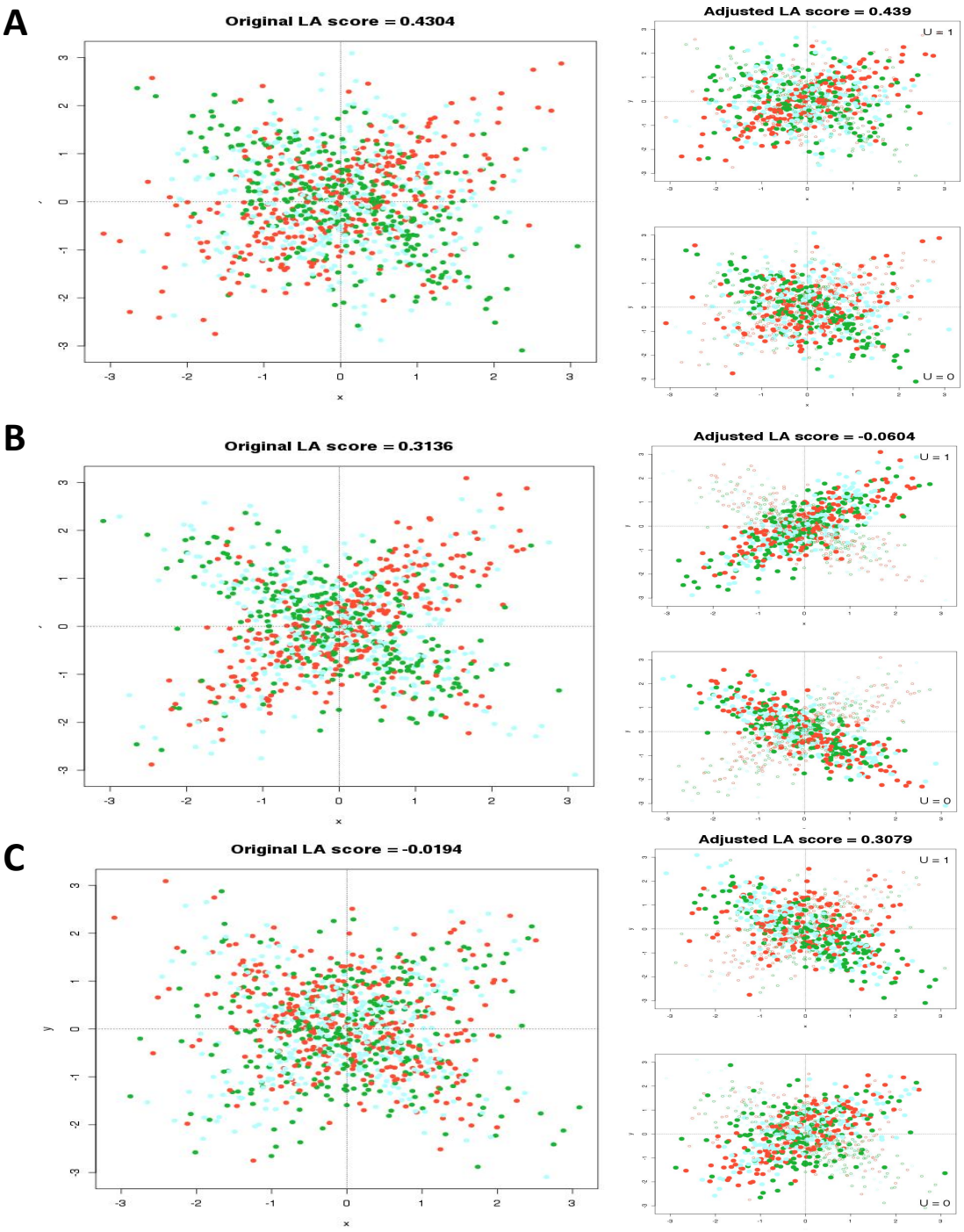

Figure 5.1: Visualization of original and adjusted LA score in some typical examples (A) $\beta = 1$ and $W = Z$, (B) $\beta = 1$ and $W = 2U - 1$, (C) $\beta = 1$ and $W = Z - 4U + 2$. For each setting, three scatter plots were shown. The left one shows all observations and colored by z_i (green: less than Q1, cyan: between Q1 and Q3, red: greater than Q3). The right smaller two shows observations with $U = 1$ and $U = 0$ and colored by $z_i - s(z_i, u_i)$.

status as the background variable U (binary) to search the top and bottom 100 pairs (X, Y) with largest positive or negative LA score. When adjusting background variable effect, the logistic regression approach was used. All LA scores were significant under a permutation test.

We observed the top LA score increased about 0.05 with the adjustment. In practice, the ranking is more useful for selecting hub genes. Using the original LA scores, about 42% top/bottom pairs contain ZNF12, and the percentage increased to 57% when using the adjusted LA score. The genes appeared in top/bottom 100 pairs at least 5 times are given in Table 5.2. ZNF12, BAG1, and PIGO are the genes appeared at least 5 times in both original and adjusted lists. We also found that ATP5G3 appeared 10 times in original top LA list but never appeared after adjusted. It suggests that the liquid association of ATP5G3 and its paired genes might also be mediated by EGFR mutation status. In contrast, LAPTM4B appeared 5 times in the adjusted list but never appeared before adjusted.

	Original LA analysis	Adjusted LA analysis
ZNF12	83	114
BAG1	25	23
ATP5G3	10	-
UQCRFS1	6	3
PIGO	5	6
SYNGR2	5	4
C7orf30	1	6
CCDC104	1	6
CLTA	3	6
LAPTM4B	-	5
MAP4K3	4	5

Table 5.2: Times appeared in the list of top or bottom 100 LA pairs.

5.6 Background Adjustment for Liquid Association Using a Binary Scouting Variable

In many applications the scouting variable Z itself might be a binary variable. For example, we might want to find the covariation pattern between the expression profile of a pair of genes under different disease status or driver mutation status. In Sun 2008[49], given a well standardized binary variable Z with $Pr(Z = a) = p$ and $Pr(Z = b) = q = 1 - p$ where $a = \sqrt{q/p}$ and $b = -\sqrt{p/q}$, they defined the binary LA score to be

$$LA(X, Y|Z) = \frac{E[XY|Z = a] - E[XY|Z = b]}{a - b}. \tag{5.12}$$

This also can be written as

$$LA(X, Y|Z) = \sqrt{pq}[E[XY|Z = a] - E[XY|Z = b]] = E[XYZ]. \tag{5.13}$$

It simply follows the fact that $(a - b)^{-1} = \sqrt{pq} = ap = -bq$.

Here we derived the LA score for a binary scouting variable adjusted by another background variable U. We have

$$\begin{aligned} LA(X, Y|Z\|U) &= \int \{\frac{E[XY|Z = a, U = u] - E[XY|Z = b, U = u]}{a - b}\} f_U(u)du \\ &= \int \{apE[XY|Z = a, U = u] + bqE[XY|Z = b, U = u]\} f_U(u)du \\ &= \int \sum_{z\in\{a,b\}} E[XY \frac{zf_Z(z)}{f_{Z|U}(z|u)}|Z = z, U = u] f_{Z,U}(z, u)du \\ &= E[XYZt(Z, U)]. \end{aligned}$$

where $t(z, u) = \frac{f_Z(z)}{f_{Z|U}(z|u)}$ is the ratio of marginal probability to conditional probability. For U is also a binary variable, we can estimate it using

$$\widehat{LA}(X, Y|Z\|U) = \frac{1}{n}\sum_{i=1}^{n} x_i y_i z_i \frac{n_{+u_i} \cdot n_{z_i+}}{n_{z_i u_i} \cdot n}$$

where $n_{+u_i} = \sum_{j=1}^{n} \mathbf{1}_{\{u_j=u_i\}}$, $n_{z_i+} = \sum_{j=1}^{n} \mathbf{1}_{\{z_j=z_i\}}$ and $n_{z_i u_i} = \sum_{j=1}^{n} \mathbf{1}_{\{z_j=z_i\}} \mathbf{1}_{\{u_j=u_i\}}$.

5.7 Appendix

Proof of Lemma 1

Using integration by part and the chain rule we have

$$\begin{aligned} & LA(X,Y|Z\|U) \\ = & \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{\partial}{\partial z}h(z,u)f_{ZU}(z,u)dzdu \\ = & \int_{-\infty}^{\infty}\{h(z,u)f_{ZU}(z,u)|_{z=-\infty}^{\infty}-\int_{-\infty}^{\infty}h(z,u)\frac{\partial}{\partial z}f_{ZU}(z,u)dz\}du \\ = & -\int_{-\infty}^{\infty}[\int_{-\infty}^{\infty}h(z,u)f_{U|Z}(u|z)du]\frac{\partial}{\partial z}f_Z(z)dz \\ & -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}h(z,u)f_Z(z)\frac{\partial}{\partial z}f_{U|Z}(u|z)dzdu. \end{aligned}$$

The first term can be written as

$$\begin{aligned} & -\int_{-\infty}^{\infty}g(z)f_Z'(z)dz \\ = & -g(z)f_Z(z)|_{-\infty}^{\infty}+\int_{-\infty}^{\infty}g'(z)f_Z(z)dz \\ = & LA(X,Y|Z) \end{aligned}$$

and the second term can be written as

$$\begin{aligned} & -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}h(z,u)f_Z(z)\frac{\partial}{\partial z}[\log f_{U|Z}(u|z)]f_{U|Z}(u|z)dzdu \\ = & -\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}E[XY\frac{\partial}{\partial z}\log f_{U|Z}(u|z)|Z=z,U=u]f_{ZU}(z,u)dzdu \\ = & -E[XYs(Z,U)]. \end{aligned}$$

We finished the proof by combining these two terms.

REFERENCES

[1] A. V. Bazarov and P. Yaswen. Who is in the driver's seat in 8p12 amplifications? ZNF703 in luminal B breast tumors. *Breast cancer research : BCR*, 13(3):308, 2011.

[2] P. J. Campbell, P. J. Stephens, E. D. Pleasance, S. O'Meara, H. Li, T. Santarius, L. A. Stebbings, C. Leroy, S. Edkins, C. Hardy, J. W. Teague, A. Menzies, I. Goodhead, D. J. Turner, C. M. Clee, M. A. Quail, A. Cox, C. Brown, R. Durbin, M. E. Hurles, P. A. W. Edwards, G. R. Bignell, M. R. Stratton, and P. A. Futreal. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. *Nature Genetics*, 40(6):722–729, 2008.

[3] M. T. Chang, S. Asthana, S. P. Gao, B. H. Lee, J. S. Chapman, C. Kandoth, J. Gao, N. D. Socci, D. B. Solit, A. B. Olshen, N. Schultz, and B. S. Taylor. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nature biotechnology*, 34(2):1–11, 2015.

[4] B.-J. Chen, H. C. Causton, D. Mancenido, N. L. Goddard, E. O. Perlstein, and D. Pe'er. Harnessing gene expression to identify the genetic basis of drug resistance. *Molecular systems biology*, 5(310):310, 2009.

[5] X. Chen, H. Xu, P. Yuan, F. Fang, M. Huss, V. B. Vega, E. Wong, Y. L. Orlov, W. Zhang, J. Jiang, Y. H. Loh, H. C. Yeo, Z. X. Yeo, V. Narang, K. R. Govindarajan, B. Leong, A. Shahab, Y. Ruan, G. Bourque, W. K. Sung, N. D. Clarke, C. L. Wei, and H. H. Ng. Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells. *Cell*, 133(6):1106–1117, 2008.

[6] D. Y. Chiang, G. Getz, D. B. Jaffe, M. J. T. O. Kelly, X. Zhao, S. L. Carter, C. Russ, C. Nusbaum, M. Meyerson, and S. Eric. Massively Parallel Sequencing. *Nature*, 6(1):99–103, 2009.

[7] D. Chitale, Y. Gong, B. S. Taylor, S. Broderick, C. Brennan, R. Somwar, B. Golas, L. Wang, N. Motoi, J. Szoke, J. M. Reinersman, J. Major, C. Sander, V. E. Seshan, M. F. Zakowski, V. Rusch, W. Pao, W. Gerald, and M. Ladanyi. An integrated genomic analysis of lung cancer reveals loss of DUSP4 in EGFR-mutant tumors. *Oncogene*, 28(31):2773–2783, 2009.

[8] K. Cibulskis, M. S. Lawrence, S. L. Carter, A. Sivachenko, D. Jaffe, C. Sougnez, S. Gabriel, M. Meyerson, E. S. Lander, and G. Getz. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology*, 31(3):213–9, 2013.

[9] M. J. Clark, N. Homer, B. D. O'Connor, Z. Chen, A. Eskin, H. Lee, B. Merriman, and S. F. Nelson. U87MG decoded: The genomic sequence of a cytogenetically aberrant human cancer cell line. *PLoS Genetics*, 6(1), 2010.

[10] E. a. Collisson, J. D. Campbell, A. N. Brooks, A. H. Berger, W. Lee, J. Chmielecki, D. G. Beer, L. Cope, C. J. Creighton, L. Danilova, L. Ding, G. Getz, P. S. Hammerman, D. Neil Hayes, B. Hernandez, J. G. Herman, J. V. Heymach, I. Jurisica, R. Kucherlapati, D. Kwiatkowski, M. Ladanyi, G. Robertson, N. Schultz, R. Shen, R. Sinha, C. Sougnez, M.-S. Tsao, W. D. Travis, J. N. Weinstein, D. a. Wigle, M. D. Wilkerson, A. Chu, A. D. Cherniack, A. Hadjipanayis, M. Rosenberg, D. J. Weisenberger, P. W. Laird, A. Radenbaugh, S. Ma, J. M. Stuart, L. Averett Byers, S. B. Baylin, R. Govindan, M. Meyerson, S. B. Gabriel, K. Cibulskis, J. Kim, C. Stewart, L. Lichtenstein, E. S. Lander, M. S. Lawrence, C. Kandoth, R. Fulton, L. L. Fulton, M. D. McLellan, R. K. Wilson, K. Ye, C. C. Fronick, C. a. Maher, C. a. Miller, M. C. Wendl, C. Cabanski, E. Mardis, D. Wheeler, M. Balasundaram, Y. S. N. Butterfield, R. Carlsen, E. Chuah, N. Dhalla, R. Guin, C. Hirst, D. Lee, H. I. Li, M. Mayo, R. a. Moore, A. J. Mungall, J. E. Schein, P. Sipahimalani, A. Tam, R. Varhol, A. Gordon Robertson, N. Wye, N. Thiessen, R. a. Holt, S. J. M. Jones, M. a. Marra, M. Imielinski, R. C. Onofrio, E. Hodis, T. Zack, E. Helman, C. Sekhar Pedamallu, J. Mesirov, G. Saksena, S. E. Schumacher, S. L. Carter, L. Garraway, R. Beroukhim, S. Lee, H. S. Mahadeshwar, A. Pantazi, A. Protopopov, X. Ren, S. Seth, X. Song, J. Tang, L. Yang, J. Zhang, P.-C. Chen, M. Parfenov, A. Wei Xu, N. Santoso, L. Chin, P. J. Park, K. a. Hoadley, J. Todd Auman, S. Meng, Y. Shi, E. Buda, S. Waring, U. Veluvolu, D. Tan, P. a. Mieczkowski, C. D. Jones, J. V. Simons, M. G. Soloway, T. Bodenheimer, S. R. Jefferys, J. Roach, A. P. Hoyle, J. Wu, S. Balu, D. Singh, J. F. Prins, J. Marron, J. S. Parker, C. M. Perou, J. Liu, D. T. Maglinte, P. H. Lai, M. S. Bootwalla, D. J. Van Den Berg, T. Triche Jr, J. Cho, D. DiCara, D. Heiman, P. Lin, W. Mallard, D. Voet, H. Zhang, L. Zou, M. S. Noble, N. Gehlenborg, H. Thorvaldsdottir, M.-D. Nazaire, J. Robinson, B. Arman Aksoy, G. Ciriello, B. S. Taylor, G. Dresdner, J. Gao, B. Gross, V. E. Seshan, B. Reva, S. Onur Sumer, N. Weinhold, C. Sander, S. Ng, J. Zhu, C. C. Benz, C. Yau, D. Haussler, P. T. Spellman, P. K. Kimes, B. M. Broom, J. Wang, Y. Lu, P. Kwok Shing Ng, L. Diao, W. Liu, C. I. Amos, R. Akbani, G. B. Mills, E. Curley, J. Paulauskis, K. Lau, S. Morris, T. Shelton, D. Mallery, J. Gardner, R. Penny, C. Saller, K. Tarvin, W. G. Richards, R. Cerfolio, A. Bryant, . Daniel P. Raymond, N. a. Pennell, C. Farver, C. Czerwinski, L. Huelsenbeck-Dill, M. Iacocca, N. Petrelli, B. Rabeno, J. Brown, T. Bauer, O. Dolzhanskiy, O. Potapova, D. Rotin, O. Voronina, E. Nemirovich-Danchenko, K. V. Fedosenko, A. Gal, M. Behera, S. S. Ramalingam, G. Sica, D. Flieder, J. Boyd, J. Weaver, B. Kohl, D. Huy Quoc Thinh, G. Sandusky, H. Juhl, E. Duhig, P. Illei, E. Gabrielson, J. Shin, B. Lee, K. Rogers, D. Trusty, M. V. Brock, C. Williamson, E. Burks, K. Rieger-Christ, A. Holway, T. Sullivan, M. K. Asiedu, F. Kosari, N. Rekhtman, M. Zakowski, V. W. Rusch, P. Zippile, J. Suh, H. Pass, C. Goparaju, Y. Owusu-Sarpong, J. M. S. Bartlett, S. Kodeeswaran, J. Parfitt, H. Sekhon, M. Albert, J. Eckman, J. B. Myers, R. Cheney, C. Morrison, C. Gaudioso, J. a. Borgia, P. Bonomi, M. Pool, M. J. Liptay, F. Moiseenko, I. Zaytseva, H. Dienemann, M. Meister, P. a. Schnabel, T. R. Muley, M. Peifer, C. Gomez-Fernandez, L. Herbert, S. Egea, M. Huang, L. B. Thorne, L. Boice, A. Hill Salazar, W. K. Funkhouser, W. Kimryn Rathmell, R. Dhir, S. a. Yousem, S. Dacic, F. Schneider, J. M. Siegfried, R. Hajek, M. a. Watson, S. McDonald, B. Meyers, B. Clarke, I. a. Yang, K. M. Fong, L. Hunter, M. Windsor, R. V. Bowman, S. Peters, I. Letovanec,

K. Z. Khan, M. a. Jensen, E. E. Snyder, D. Srinivasan, A. B. Kahn, J. Baboud, D. a. Pot, K. R. Mills Shaw, M. Sheth, T. Davidsen, J. a. Demchok, L. Yang, Z. Wang, R. Tarnuzzer, J. Claude Zenklusen, B. a. Ozenberger, and H. J. Sofia. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511):543–50, 2014.

[11] A. de la Fuente. From 'differential expression' to 'differential networking' - identification of dysfunctional regulatory networks in diseases. *Trends in Genetics*, 26(7):326–333, 2010.

[12] M. A. DePristo, E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. del Angel, M. A. Rivas, M. Hanna, A. McKenna, T. J. Fennell, A. M. Kernytsky, A. Y. Sivachenko, K. Cibulskis, S. B. Gabriel, D. Altshuler, and M. J. Daly. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, 43(5):491–8, 2011.

[13] S. D. Der, J. Sykes, M. Pintilie, C.-Q. Zhu, D. Strumpf, N. Liu, I. Jurisica, F. a. Shepherd, and M.-S. Tsao. Validation of a Histology-Independent Prognostic Gene Signature for Early-Stage, NonSmall-Cell Lung Cancer Including Stage IA Patients. *Journal of Thoracic Oncology*, 9(1):59–64, 2014.

[14] L. Ding, G. Getz, D. a. Wheeler, E. R. Mardis, M. D. McLellan, K. Cibulskis, C. Sougnez, H. Greulich, D. M. Muzny, M. B. Morgan, L. Fulton, R. S. Fulton, Q. Zhang, M. C. Wendl, M. S. Lawrence, D. E. Larson, K. Chen, D. J. Dooling, A. Sabo, A. C. Hawes, H. Shen, S. N. Jhangiani, L. R. Lewis, O. Hall, Y. Zhu, T. Mathew, Y. Ren, J. Yao, S. E. Scherer, K. Clerc, G. a. Metcalf, B. Ng, A. Milosavljevic, M. L. Gonzalez-Garay, J. R. Osborne, R. Meyer, X. Shi, Y. Tang, D. C. Koboldt, L. Lin, R. Abbott, T. L. Miner, C. Pohl, G. Fewell, C. Haipek, H. Schmidt, B. H. Dunford-Shore, A. Kraja, S. D. Crosby, C. S. Sawyer, T. Vickery, S. Sander, J. Robinson, W. Winckler, J. Baldwin, L. R. Chirieac, A. Dutt, T. Fennell, M. Hanna, B. E. Johnson, R. C. Onofrio, R. K. Thomas, G. Tonon, B. a. Weir, X. Zhao, L. Ziaugra, M. C. Zody, T. Giordano, M. B. Orringer, J. a. Roth, M. R. Spitz, I. I. Wistuba, B. Ozenberger, P. J. Good, A. C. Chang, D. G. Beer, M. a. Watson, M. Ladanyi, S. Broderick, A. Yoshizawa, W. D. Travis, W. Pao, M. a. Province, G. M. Weinstock, H. E. Varmus, S. B. Gabriel, E. S. Lander, R. a. Gibbs, M. Meyerson, and R. K. Wilson. Somatic mutations affect key pathways in lung adenocarcinoma. *Nature*, 455(7216):1069–1075, 2008.

[15] H. Dweep and N. Gretz. miRWalk2.0: a comprehensive atlas of microRNA-target interactions. *Nature Methods*, 12(8):697–697, 2015.

[16] M. D. Edmonds, K. L. Boyd, T. Moyo, R. Mitra, R. Duszynski, M. P. Arrate, X. Chen, Z. Zhao, T. S. Blackwell, T. Andl, and C. M. Eischen. MicroRNA-31 initiates lung tumorigenesis and promotes mutant KRAS-driven lung cancer. *Journal of Clinical Investigation*, 126(1):349–364, 2016.

[17] C. T. Endpoints. Guidance for Industry: Clinical Trial Endpoints for the Approval of Cancer Drugs and Biologics. *Biotechnology Law Report*, 26(4):375–386, 2007.

[18] M. Fukuoka, Y. L. Wu, S. Thongprasert, P. Sunpaweravong, S. S. Leong, V. Sriuranpong, T. Y. Chao, K. Nakagawa, D. T. Chu, N. Saijo, E. L. Duffield, Y. Rukazenkov, G. Speake, H. Jiang, A. A. Armour, K. F. To, J. C. H. Yang, and T. S. K. Mok. Biomarker analyses and final overall survival results from a phase III, randomized, open-label, first-line study of gefitinib versus carboplatin/paclitaxel in clinically selected patients with advanced non - small-cell lung cancer in Asia (IPASS). *Journal of Clinical Oncology*, 29(21):2866–2874, 2011.

[19] J. S. Gehring, B. Fischer, M. Lawrence, and W. Huber. SomaticSignatures: Inferring mutational signatures from single-nucleotide variants. *Bioinformatics*, 31(22):3673–3675, 2015.

[20] A. Gonzalez-Perez and N. Lopez-Bigas. Functional impact bias reveals cancer drivers. *Nucleic Acids Research*, 40(21):1–10, 2012.

[21] M. Imielinski, A. H. Berger, P. S. Hammerman, B. Hernandez, T. J. Pugh, E. Hodis, J. Cho, J. Suh, M. Capelletti, A. Sivachenko, C. Sougnez, D. Auclair, M. S. Lawrence, P. Stojanov, K. Cibulskis, K. Choi, L. De Waal, T. Sharifnia, A. Brooks, H. Greulich, S. Banerji, T. Zander, D. Seidel, F. Leenders, S. Ansén, C. Ludwig, W. Engel-Riedel, E. Stoelben, J. Wolf, C. Goparju, K. Thompson, W. Winckler, D. Kwiatkowski, B. E. Johnson, P. a. Jänne, V. a. Miller, W. Pao, W. D. Travis, H. I. Pass, S. B. Gabriel, E. S. Lander, R. K. Thomas, L. a. Garraway, G. Getz, and M. Meyerson. Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell*, 150(6):1107–1120, 2012.

[22] H. Is, A. N. Activating, A. L. K. Ligand, C. Loss, O. F. Replication, and R. Drives. Complete Loss of Replication Repair Drives Ultra-Hypermutated Cancers. *Cancer discovery*, 5(3):227, 2015.

[23] J. M. Kidd, G. M. Cooper, W. F. Donahue, H. S. Hayden, N. Sampas, T. Graves, N. Hansen, B. Teague, C. Alkan, F. Antonacci, E. Haugen, T. Zerr, N. A. Yamada, P. Tsang, T. L. Newman, E. Tuzun, Z. Cheng, H. M. Ebling, N. Tusneem, R. David, W. Gillett, K. A. Phelps, M. Weaver, D. Saranga, A. Brand, W. Tao, E. Gustafson, K. McKernan, L. Chen, M. Malig, J. D. Smith, J. M. Korn, S. A. McCarroll, D. A. Altshuler, D. A. Peiffer, M. Dorschner, J. Stamatoyannopoulos, D. Schwartz, D. A. Nickerson, J. C. Mullikin, R. K. Wilson, L. Bruhn, M. V. Olson, R. Kaul, D. R. Smith, and E. E. Eichler. Mapping and sequencing of structural variation from eight human genomes. *Nature*, 453(7191):56–64, 2008.

[24] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome biology*, 14(4):R36, 2013.

[25] D. Kim and S. L. Salzberg. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biol*, 12(8):R72, 2011.

[26] D. C. Koboldt, Q. Zhang, D. E. Larson, D. Shen, M. D. Mclellan, L. Lin, C. a. Miller, E. R. Mardis, L. Ding, and R. K. Wilson. VarScan 2 : Somatic mutation and copy

number alteration discovery in cancer by exome sequencing VarScan 2 : Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, 22(3):568–576, 2012.

[27] J. O. Korbel, A. Abyzov, X. J. Mu, N. Carriero, P. Cayting, Z. Zhang, M. Snyder, and M. B. Gerstein. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome biology*, 10(2):R23, 2009.

[28] F.-C. Kuan, L.-T. Kuo, M.-C. Chen, C.-T. Yang, C.-S. Shi, D. Teng, and K.-D. Lee. Overall survival benefits of first-line EGFR tyrosine kinase inhibitors in EGFR-mutated non-small-cell lung cancers: a systematic review and meta-analysis. *British Journal of Cancer*, 113(10):1519–1528, 2015.

[29] H. Y. K. Lam, X. J. Mu, A. M. Stütz, A. Tanzer, P. D. Cayting, M. Snyder, P. M. Kim, J. O. Korbel, and M. B. Gerstein. Nucleotide-resolution analysis of structural variants using BreakSeq and a breakpoint library. *Nature Biotechnology*, 28(1):47–55, 2010.

[30] M. S. Lawrence, P. Stojanov, P. Polak, G. V. Kryukov, K. Cibulskis, A. Sivachenko, S. L. Carter, C. Stewart, C. H. Mermel, S. a. Roberts, A. Kiezun, P. S. Hammerman, A. McKenna, Y. Drier, L. Zou, A. H. Ramos, T. J. Pugh, N. Stransky, E. Helman, J. Kim, C. Sougnez, L. Ambrogio, E. Nickerson, E. Shefler, M. L. Cortés, D. Auclair, G. Saksena, D. Voet, M. Noble, D. DiCara, P. Lin, L. Lichtenstein, D. I. Heiman, T. Fennell, M. Imielinski, B. Hernandez, E. Hodis, S. Baca, A. M. Dulak, J. Lohr, D.-A. Landau, C. J. Wu, J. Melendez-Zajgla, A. Hidalgo-Miranda, A. Koren, S. a. McCarroll, J. Mora, R. S. Lee, B. Crompton, R. Onofrio, M. Parkin, W. Winckler, K. Ardlie, S. B. Gabriel, C. W. M. Roberts, J. a. Biegel, K. Stegmaier, A. J. Bass, L. a. Garraway, M. Meyerson, T. R. Golub, D. a. Gordenin, S. Sunyaev, E. S. Lander, and G. Getz. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, 499(7457):214–8, 2013.

[31] J. K. Lee, D. M. Havaleshko, H. Cho, J. N. Weinstein, E. P. Kaldjian, J. Karpovich, a. Grimshaw, and D. Theodorescu. A strategy for predicting the chemosensitivity of human cancers and its application to drug discovery. *Journal of Urology*, 179(2):787, 2008.

[32] H. Li and R. Durbin. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

[33] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, and R. Durbin. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[34] K.-C. Li. Genome-wide coexpression dynamics: Theory and application. *Proceedings of the National Academy of Sciences*, 99(26):16875–16880, 2002.

[35] S. Li, Y.-L. Choi, Z. Gong, X. Liu, M. Lira, Z. Kan, E. Oh, J. Wang, J. C. Ting, X. Ye, C. Reinhart, X. Liu, Y. Pei, W. Zhou, R. Chen, S. Fu, G. Jin, A. Jiang, J. Fernandez,

J. Hardwick, M. W. Kang, H. I, H. Zheng, J. Kim, and M. Mao. Comprehensive characterization of oncogenic drivers in Asian lung adenocarcinoma. *Journal of thoracic oncology : official publication of the International Association for the Study of Lung Cancer*, 11(12):2129–2140, 2016.

[36] P. Medvedev, M. Stanciu, and M. Brudno. Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods*, 6(11s):S13–S20, 2009.

[37] C. H. Mermel, S. E. Schumacher, B. Hill, M. L. Meyerson, R. Beroukhim, and G. Getz. GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology*, 12(4):R41, 2011.

[38] H. Okayama, T. Kohno, Y. Ishii, Y. Shimada, K. Shiraishi, R. Iwakawa, K. Furuta, K. Tsuta, T. Shibata, S. Yamamoto, S. I. Watanabe, H. Sakamoto, K. Kumamoto, S. Takenoshita, N. Gotoh, H. Mizuno, A. Sarai, S. Kawano, R. Yamaguchi, S. Miyano, and J. Yokota. Identification of genes upregulated in ALK-positive and EGFR/KRAS/ALK-negative lung adenocarcinomas. *Cancer Research*, 72(1):100–111, 2012.

[39] A. B. Olshen, E. S. Venkatraman, R. Lucito, and M. Wigler. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572, 2004.

[40] A. R. Quinlan, R. A. Clark, S. Sokolova, M. L. Leibowitz, Y. Zhang, M. E. Hurles, J. C. Mell, and I. M. Hall. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome. *Genome Research*, 20(5):623–635, 2010.

[41] B. J. Raphael, S. Volik, C. Collins, and P. A. Pevzner. Reconstructing tumor genome architectures. *Bioinformatics*, 19(SUPPL. 2), 2003.

[42] R. Redon, S. Ishikawa, K. R. Fitch, L. Feuk, G. H. Perry, T. D. Andrews, H. Fiegler, M. H. Shapero, A. R. Carson, W. Chen, E. K. Cho, S. Dallaire, J. L. Freeman, J. R. González, M. Gratacòs, J. Huang, D. Kalaitzopoulos, D. Komura, J. R. MacDonald, C. R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M. J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, J. Zhang, L. Armengol, D. F. Conrad, X. Estivill, C. Tyler-Smith, N. P. Carter, H. Aburatani, C. Lee, K. W. Jones, S. W. Scherer, and M. E. Hurles. Global variation in copy number in the human genome. *Nature*, 444(7118):444–54, 2006.

[43] D. A. Sallman and E. Padron. Integrating mutation variant allele frequency into clinical practice in myeloid malignancies. *Hematology/ Oncology and Stem Cell Therapy*, 9(3):89–95, 2016.

[44] J. Sebat, B. Lakshmi, J. Troge, J. Alexander, J. Young, P. Lundin, S. Månér, H. Massa, M. Walker, M. Chi, N. Navin, R. Lucito, J. Healy, J. Hicks, K. Ye, A. Reiner, T. C. Gilliam, B. Trask, N. Patterson, A. Zetterberg, and M. Wigler. Large-Scale Copy Number Polymorphism in the Human Genome. *Science*, 305(2004):525–528, 2004.

[45] K. Shedden, J. M. G. Taylor, S. a. Enkemann, M.-S. Tsao, T. J. Yeatman, W. L. Gerald, S. Eschrich, I. Jurisica, T. J. Giordano, D. E. Misek, A. C. Chang, C. Q. Zhu, D. Strumpf, S. Hanash, F. a. Shepherd, K. Ding, L. Seymour, K. Naoki, N. Pennell, B. Weir, R. Verhaak, C. Ladd-Acosta, T. Golub, M. Gruidl, A. Sharma, J. Szoke, M. Zakowski, V. Rusch, M. Kris, A. Viale, N. Motoi, W. Travis, B. Conley, V. E. Seshan, M. Meyerson, R. Kuick, K. K. Dobbin, T. Lively, J. W. Jacobson, and D. G. Beer. Gene expression-based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nature medicine*, 14(8):822–827, 2008.

[46] S. Sindi, E. Helman, A. Bashir, and B. J. Raphael. A geometric approach for classification and comparison of structural variants. *Bioinformatics*, 25(12):222–230, 2009.

[47] F. Sircoulomb, N. Nicolas, A. Ferrari, P. Finetti, I. Bekhouche, E. Rousselet, A. Lonigro, J. Adélaïde, E. Baudelet, S. Esteyriès, J. Wicinski, S. Audebert, E. Charafe-Jauffret, J. Jacquemier, M. Lopez, J. P. Borg, C. Sotiriou, C. Popovici, F. Bertucci, D. Birnbaum, M. Chaffanet, and C. Gines tier. ZNF703 gene amplification at 8p12 specifies luminal B breast cancer. *EMBO Molecular Medicine*, 3(3):153–166, 2011.

[48] B. E. Stranger, M. S. Forrest, M. Dunning, C. E. Ingle, C. Beazley, N. Thorne, R. Redon, C. P. Bird, A. D. Grassi, C. Lee, C. Tyler-smith, N. Carter, S. W. Scherer, S. Tavaré, P. Deloukas, M. E. Hurles, and E. T. Dermitzakis. Expression Phenotypes. *Recherche*, 315(February):848–853, 2007.

[49] W. Sun, S. Yuan, and K.-C. Li. Trait-trait dynamic interaction: 2D-trait eQTL mapping for genetic variation study. *BMC Genomics*, 9(1):242, 2008.

[50] D. Tamborero, A. Gonzalez-Perez, and N. Lopez-Bigas. OncodriveCLUST: Exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics*, 29(18):2238–2244, 2013.

[51] K. Ye, M. H. Schulz, Q. Long, R. Apweiler, and Z. Ning. Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, 25(21):2865–2871, 2009.

[52] S. Yuan, S. L. Yu, H. Y. Chen, Y. C. Hsu, K. Y. Su, H. W. Chen, C. Y. Chen, C. J. Yu, J. Y. Shih, Y. L. Chang, C. L. Cheng, C. P. Hsu, J. Y. Hsia, C. Y. Lin, G. Wu, C. H. Liu, C. D. Wang, K. C. Yang, Y. W. Chen, Y. L. Lai, C. C. Hsu, T. C. Lin, T. Y. Yang, K. C. Chen, K. H. Hsu, J. J. W. Chen, G. C. Chang, K. C. Li, and P. C. Yang. Clustered genomic alterations in chromosome 7p dictate outcomes and targeted treatment responses of lung adenocarcinoma with EGFR-activating mutations. *Journal of Clinical Oncology*, 29(25):3435–3442, 2011.

[53] W. Zhao, X. He, K. A. Hoadley, J. S. Parker, D. N. Hayes, and C. M. Perou. Comparison of RNA-Seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC genomics*, 15(1):419, 2014.