

Genome and transcriptome analysis for the process of cancer progression

November 2018

Yoshimasa Aoto

A Thesis for the Degree of Ph.D. in Science

Genome and transcriptome analysis for
the process of cancer progression

November 2018

Graduate School of Science and Technology
Keio University

Yoshimasa Aoto

Abstract

Cancer is known as one of the leading causes of death in the world. Recent years have witnessed substantial progress in understanding tumor heterogeneity and the process of tumor progression; however, the entire step-wise process of the transition of tumors from benign to metastatic state remains poorly understood. In this dissertation, I conducted genome- and transcriptome-sequencing analyses for the tumorigenesis process to reveal the entire mechanisms of the tumor progression.

In the genome-sequencing study, I performed a prospective analysis by employing an experimental carcinogenesis mouse model to systematically understand the evolutionary process of tumors. Collaborators surgically collected a part of a lesion of each tumor and followed the progression of these tumors *in vivo* over time. I conducted the comparative time-series analysis of the genomes of tumors with different fates, i.e., those that eventually metastasized and regressed. The results suggested that these tumors acquired and inherited different mutations. Despite the occurrence of an intra-tumor selection event for malignant alteration during the transformation from early- to late-stage papilloma, the fate determination of tumors might be determined at an even earlier stage.

While the genome-sequencing enables us to analyze the genomic mutations, transcriptome-sequencing technology is widely used to comprehensively detect and quantify cellular gene expression. Numerous analytical methods have been proposed for identifying differentially expressed genes (DEGs) between paired samples such as tumor and control specimens, but few studies have reported methods for analyzing differential expression under multiple conditions as my research target, the step-wise process of the tumor progression. In the transcriptome-sequencing study, I propose a novel method, DEclust, for differential expression analysis among more than two matched samples from distinct tissues or conditions. As compared to conventional clustering methods, DEclust more accurately extracts statistically significant gene clusters from multi-conditional transcriptome data. As a result of applying my method to the cancer transcriptome analysis, DEclust extracted 16 gene clusters whose expression patterns were discriminative in the process of the tumor progression; moreover, the genes which were belonging to each of these clusters were significantly enriched for biological functions.

Through the genome and transcriptome studies, I gave insights of the evolutionary process in the tumors over time and the gene clusters whose expression patterns were discriminative in the process of the tumor progression. Examples of the research that directly analyzes the process of tumor progression are rare; hence, my results are meaningful and I consider that it is necessary to advance such a direct approach to reveal the entire process of cancer progression.

Contents

Chapter 1 Introduction	1
1.1 Cancer	2
1.2 High-throughput DNA sequencing.....	4
1.3 Cancer genome analysis.....	5
1.4 Cancer transcriptome analysis.....	8
Chapter 2 Time-series analysis of tumorigenesis in a murine skin carcinogenesis model	10
2.1 Background	10
2.1.1 Intra-tumor heterogeneity.....	10
2.1.2 Research purpose and murine skin carcinogenesis model	11
2.2 Results.....	13
2.2.1 Mutational landscape of two tumor series with different fates	13
2.2.2 Degree of shared mutations between tumors within and among tumor series....	20
2.2.3 Intra-tumor heterogeneity and evolutionary process within tumors.....	22
2.3 Discussion	26
2.4 Materials and Methods.....	27
2.4.1 Mice	27
2.4.2 Skin carcinogenesis and tumor sampling.....	27
2.4.3 Sample preparation and targeted deep sequencing.....	28
2.4.4 Sequence data analysis.....	29
2.4.5 Hierarchical clustering of candidate mutation positions.....	30
2.4.6 Similarity measurements between tumors	30
2.4.7 Evaluation of intra-tumor heterogeneity	30
2.4.8 Estimation of the mutation rate and evolutionary process	31

2.4.9	Data availability.....	32
2.5	Conclusion	32
 Chapter 3 DEclust: A statistical approach for obtaining differential expression profiles of multiple conditions.....		
3.1	Background	33
3.1.1	RNA sequencing and gene expression analysis	33
3.1.2	Cluster analysis.....	34
3.2	Results and Discussion	35
3.2.1	Algorithm	35
3.2.2	Benchmark study of clustering.....	38
3.2.3	Benchmark study of DEG detection	42
3.2.4	Application to biological data	44
3.3	Materials and Methods.....	51
3.3.1	Simulation datasets	51
3.3.2	Analysis for benchmark study of clustering.....	53
3.3.3	Evaluation of clustering accuracy.....	54
3.3.4	Analysis for benchmark study of DEG detection.....	54
3.3.5	Evaluation of DEG detection accuracy	54
3.3.6	Mice	55
3.3.7	Skin carcinogenesis and tumor sampling.....	55
3.3.8	Sample preparation and mRNA-Seq.....	55
3.3.9	RNA-Seq data analysis	56
3.3.10	Data deposition	56
3.4	Conclusion	56
 Chapter 4 Conclusion and future work.....		
Acknowledgment		60

References.....	61
Appendix A – List of publications.....	69
Appendix B – Supplementary information of genome analysis.....	70

Abbreviation

AGPC:	acid guanidinium thiocyanate-phenol-chloroform
AUC:	area under the curve
CGC:	Cancer Gene Census
CNV:	copy number variation
COSMIC:	catalogue of somatic mutations in cancer
DDBJ:	DNA Data Bank of Japan
DEG:	differentially expressed gene
DET:	differential expression test
DMBA:	7,12-dimethylbenz(a)anthracene
DNA:	deoxyribonucleic acid
cDNA:	complementary DNA
FDR:	false discovery rate
FPR:	false-positive rate
GSEA:	gene set enrichment analysis
PCR:	polymerase chain reaction
PPV:	positive predictive value
RNA:	ribonucleic acid
RNA-Seq:	RNA sequencing
ROC:	receiver operating characteristic
SCC:	squamous cell carcinoma
SNP:	single-nucleotide polymorphism
SNV:	single-nucleotide variant
SV:	structural variant
TCGA:	The Cancer Genome Atlas
TPA:	12-O-tetradecanoylphorbol-13-acetone
TPR:	true-positive rate
VAF:	variant allele frequency

Chapter 1

Introduction

Cancer is known as the leading cause of death in the world, with approximately 14 million new cases and more than 8 million deaths occurring in 2012 worldwide (Torre *et al.*, 2015). Cancer is a result of genomic disorders represented by DNA mutations that typically lead to loss of DNA repair function and gain of abnormal proliferation function. Numerous reports on the process of malignant alterations suggest that benign tumors progress in a stepwise fashion while acquiring driver and passenger mutations, which eventually invade surrounding tissues to finally migrate to distant tissues (Fearon and Vogelstein, 1990; Hanahan and Weinberg, 2011; Greaves and Maley, 2012; Vogelstein *et al.*, 2013; McGranahan *et al.*, 2015; Tomczak, Czerwińska and Wiznerowicz, 2015); however, the entire process of tumor progression, from a benign to metastatic state, is still poorly understood.

In this dissertation, I present genome and transcriptome studies for the tumorigenesis process to understand the entire mechanisms of the tumor progression. First, I performed a prospective genome-sequencing analysis by employing an experimental carcinogenesis for mice. My collaborators surgically collected a part of a lesion of each tumor and followed the progression of these tumors *in vivo* over time. I performed a time-series analysis of these tumors to reveal the accumulation process of genomic mutations over time, and also a comparative analysis to reveal the differences between tumors with different fates, e.g., those that eventually metastasized and regressed (Aoto *et al.*, 2018). Second, I performed transcriptome-sequencing to analyze the transcriptome dynamics in the tumorigenesis process. For this purpose, I first developed a novel analytical method to handle multi-conditional, such as time-series, transcriptome data. Next, I applied the proposed method for the tumors obtained from the experimental carcinogenesis to extract gene clusters whose expression patterns were discriminative in the tumorigenesis process (Aoto *et al.*, 2017). I aimed to clarify the tumorigenesis process through understanding each of the omics layers, genome and transcriptome, by conducting above two studies. The genome analysis has a potential to reveal the intra-tumor structure changing with time. The transcriptome analysis also has a potential to identify the genes with characteristic expression patterns in the process of tumorigenesis.

This dissertation is organized as follows. In this chapter, I introduce some basic information about cancer and the originality of my studies briefly. In Chapter 2, I describe the results of genome-sequencing analysis. In Chapter 3, I describe the results of transcriptome-sequencing analysis. Finally, the summary of my studies and also discussion of future prospects are described in Chapter 4.

1.1 Cancer

As described above, cancer is one of the fatal diseases worldwide. Recent trends of cancer disease are summarized in a fact sheet published by the World Health Organization (WHO) in February 2018 (Torre *et al.*, 2015). Cancer is the second leading cause of death globally, and there are several known risk factors such as high body mass index, low fruit and vegetable intake, lack of physical activity, tobacco use, and alcohol use. Cancer can affect any part of the body. The initial event of the development of cancer is genomic disorders such as DNA mutations. The DNA mutations that induce the aberrant growth or the defection in the DNA repair function causing an initial tumor cell. The aberrant growth promotes the tumorigenesis, and the defection in the DNA repair function affects genomic instability of the cancer genome. Thus, the tumor cell proliferates abnormally to form benign tumor, and the benign tumor progresses stepwise accompanied by random acquisition of further somatic mutations owing to genomic instability (Fig. 1.1). Eventually, the tumor invades surrounding tissues to finally metastasizes to distant tissues (Fearon and Vogelstein, 1990; Vogelstein *et al.*, 2013). In most cases except for the familial cancers, somatic mutations cause cancers. Recent studies have reported that the somatic mutations not only cause the initial tumor cell, but also drive tumor progression. The somatic mutations that directly drive tumor progression are called as “driver mutations”, and the somatic mutations that accidentally occur in the process of tumor progression are called as “passenger mutations”.

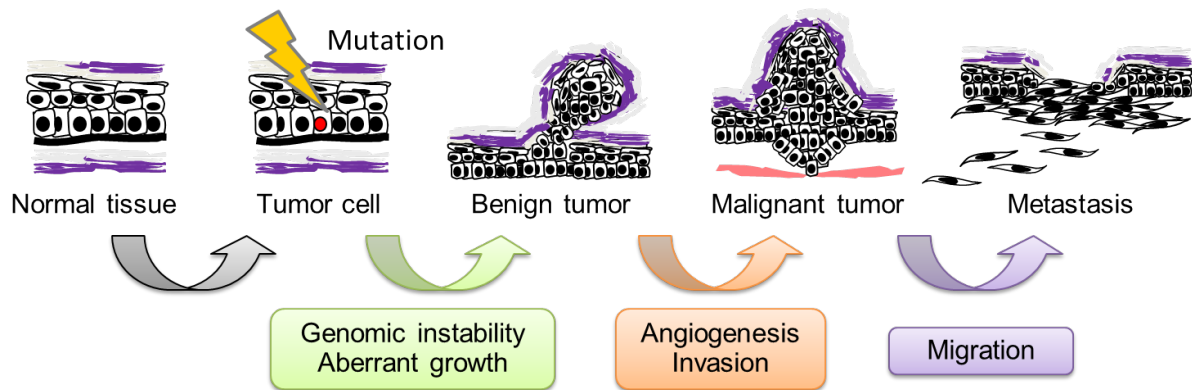


Figure 1.1. Stepwise process of cancer progression. In the process of cancer progression, the tumor acquires abnormal biological functions in a sequential manner. The tumor eventually invades the surrounding tissues and migrates to distant tissues.

Recently, high-throughput sequencing technology has been used for genome-wide cancer analysis for a comprehensive understanding of the genomic mutations, gene expression levels, and epigenetic status (Shendure and Ji, 2008; Metzker, 2010). Cancer genome studies that aim to detect driver gene mutations and the cancer transcriptome studies that detect abnormal gene expression have been actively conducted to understand the cause of malignant alteration (Tomczak, Czerwińska and Wiznerowicz, 2015). These previous researches identified that there are various genes which drive tumor development, while multiple drivers are necessary for the malignant transformation, and each case is caused by different driver events (Martincorena and Campbell, 2015). This difference between tumor cases is known as “inter-tumor heterogeneity”. This inter-tumor heterogeneity increases the difficulty of cancer therapy to which different treatments are required for each case (Burrell *et al.*, 2013; Vogelstein *et al.*, 2013).

While a single original tumor cell proliferates and expands to form a tumor, each tumor cell randomly acquires genomic mutations due to the genomic instability; therefore, there is genetic diversity among tumor cells (Fig. 1.2) (Greaves and Maley, 2012). This genetic diversity between tumor cells is known as “intra-tumor heterogeneity”, and the cell population with the same genetic background is called as a “sub-clone”. The genetic diversity of tumor cells generates the physiological diversity and difference in therapeutic sensitivity among tumor cells. Accordingly, the intra-tumor heterogeneity is considered to be the most critical cause of treatment resistance and the recurrence of cancer (Greaves and Maley, 2012; Swanton, 2012; Burrell *et al.*, 2013). Furthermore, the intra-tumor heterogeneity is also considered to contribute to the cancer progression. The environment inside and outside the tumor is harsh for tumor cells to survive and it is changing continually. At the beginning of the tumorigenesis, the tumor cells have to evade the immune response from surrounding tissue. After establishment of a benign tumor, it is necessary to withstand the nutritional deficiency and the oxidative stress which are the results of the abnormal proliferation (Hanahan and Weinberg, 2011). Only tumor cells that could acquire the traits that enable to overcome these adversities can survive and become malignant; thus, by considering the adversities as selection pressure, evolutionary progression models have been proposed and discussed aggressively (Martincorena and Campbell, 2015; Davis, Gao and Navin, 2017).

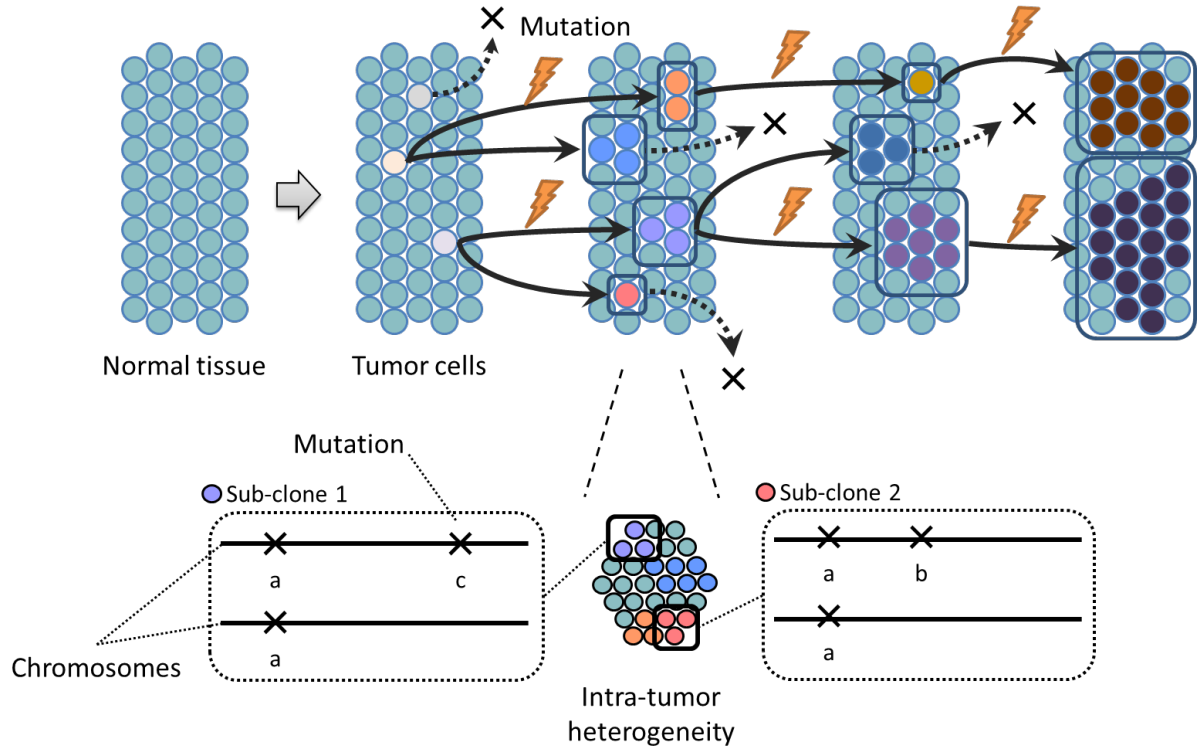


Figure 1.2. Intra-tumor heterogeneity and evolutionary process in tumors. The original tumor cell proliferates and expands to form a tumor. The tumor cells randomly acquire the genomic mutations and form a malignant tumor with genetic diversity.

Through the existing researches to date, the understanding of the inter-/intra-tumor heterogeneity has been boosted; however, the entire process of the tumor progression is still unknown because most of cancer researches use specimens that have already become malignant. Hence, I conducted genome and transcriptome analysis to reveal the whole process of tumor progression, from a benign to metastatic state.

1.2 High-throughput DNA sequencing

High-throughput DNA sequencing technologies have been popularized since the 2000s (Metzker, 2010). While several approaches have been proposed for high-throughput DNA sequencing, either approach has adopted massively parallel sequencing of DNA fragments. In usual cases, an input of DNA sequencer is fragmented DNAs obtained by sonication or enzyme reaction for the genomic DNA. One side (single-end) or both sides (paired-end) of the DNA fragments are sequenced. The cDNA that are obtained from the reverse transcription reaction for the RNA is also available for high-throughput sequencing. The throughputs and

sequencing lengths are different among sequencing platforms. For instance, the HiSeq series from Illumina Inc., the most prevalent technology, allows up to five billion parallel sequencing and ~300bp determination for each DNA fragment (<https://jp.illumina.com/>), and the Illumina MiSeq, I adopted, implements 25 million parallel sequencing reactions and ~600bp read length. The determined sequences are termed as “reads”. These reads are often mapped against a known (reference) genomic sequence to detect DNA mutations or gene expression.

There are several applications to solve various research problems. The whole genome sequencing (WGS) and whole transcriptome sequencing (WTS) technologies provide whole genomic-scale nucleotide sequence information of specimens. Moreover, targeted sequencing technology has been also established which enables to analyze specific genomic regions such as the exons, promoter regions, and so on.

1.3 Cancer genome analysis

As described above, the cause of cancer disease is genomic disorders. In this decade, the cancer genome analysis has been widely conducted by using high-throughput sequencing technologies and that has brought a lot of discoveries. There are several types of genomic mutations ranging from local mutation on the genomic sequence such as single-nucleotide variants (SNVs) and short-insertions / deletions (indels) to chromosomal-scale mutations such as structural variants (SVs) and copy number variations (CNVs). To detect these genomic mutations, the sequenced short-reads are simultaneously mapped against the reference genome sequence (Fig. 1.3). Many analytical methods have been proposed: BWA (Li and Durbin, 2009) and bowtie2 (Langmead and Salzberg, 2012) are the standard tools for mapping, and MuTect2 (Cibulskis *et al.*, 2013), Strelka (Saunders *et al.*, 2012), VarScan2 (Koboldt *et al.*, 2012), SomaticSniper (Larson *et al.*, 2012), and LoFreq (Wilm *et al.*, 2012) are standardly used to detect somatic mutations in cancer.

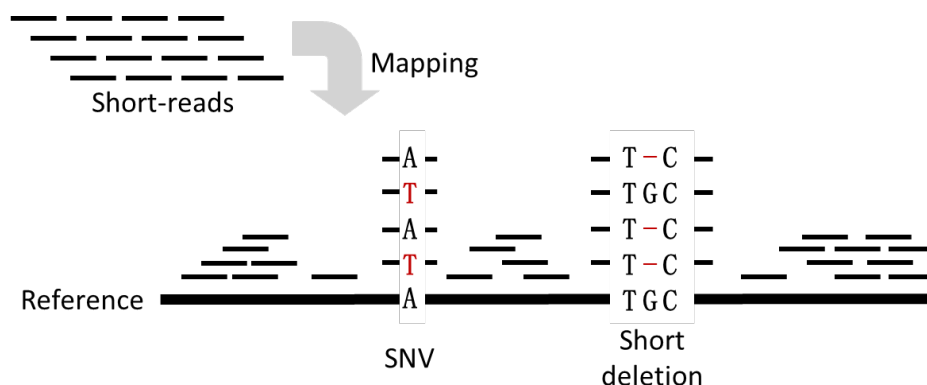


Figure 1.3. Detection of the genomic mutations by using the sequenced short-reads.

Mainly, two methods are adopted for the analysis of intra-tumor heterogeneity (Fig. 1.4). The first method is single-cell sequencing. Single-cell sequencing is appropriate for the cancer genome studies because it enables to analyze each tumor cell separately; however, it is impractical for comprehensive analysis of a large number of tumor cells composing a single tumor mass. The second method is bulk sequencing. The bulk sequencing is a method for comprehensive analysis of a whole tissue. In the process of sample preparation for the bulk sequencing, genomic DNAs are extracted from the tissue without dividing into cells. The extracted genomic DNAs are fragmented and then are randomly chosen to sequence. Although the information on each cell is lost, the bulk sequencing method enables a comprehensive detection of the somatic mutations and an estimation of the prevalence of the mutations in a tumor. For the estimation of the prevalence of each mutation, a variant allele frequency (VAF) is calculated from the mapping results. The VAF is defined as a proportion of the mutation-supporting reads per the entire reads mapped to the genomic position; then, it is considered as a pseudo-probability of the prevalence of the mutant cells in a tumor. The VAFs are widely used for the estimation of the intra-tumor structure, and several methods have also been proposed to predict the intra-tumor sub-clonal structure from the VAFs (Miller *et al.*, 2014; Roth *et al.*, 2014). Moreover, the targeted sequencing technology, which can concentrate on the acquired sequencing reads in specific genomic regions, is used in conjunction with the bulk sequencing to estimate the VAFs more accurately and to detect low frequency mutations.

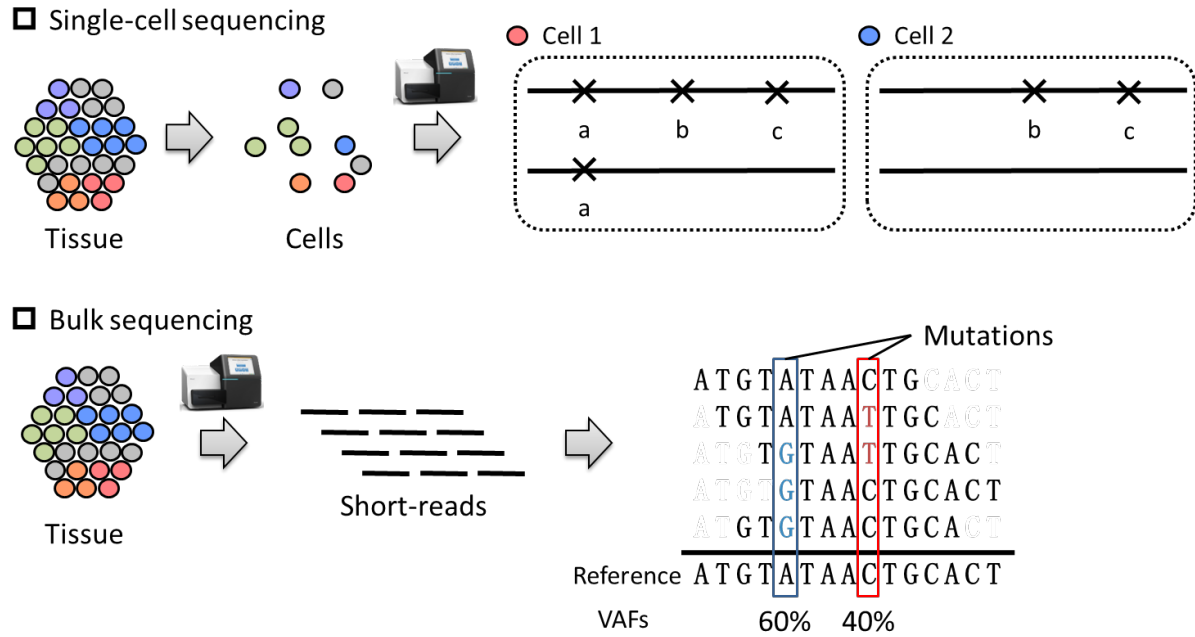


Figure 1.4. Single-cell sequencing and bulk sequencing methods. The single-cell method provides nucleotide sequence information for each cell separately. The bulk method enables to analyze whole or a part of a tissue. When the research target is a whole tissue or a part of the tissue, the bulk sequencing method is usually adopted. Although the sequence information from bulk approach cannot decompose into cell units, the mutated positions and the variant allele frequencies (VAFs) of the mutated positions are available.

Recent cancer genome studies, represented by The Cancer Genome Atlas (TCGA), have gradually revealed the mutational landscapes of various types of tumors. The mutation burden and the driver mutations have been extensively investigated (Tomczak, Czerwińska and Wiznerowicz, 2015). The correlation between the substitution patterns of the somatic mutations and the mutagens has been reported (Alexandrov *et al.*, 2013). For the understanding of the intra-tumor heterogeneity, recent studies adopt a multi-region sequencing. In these studies, single malignant tumor was divided into multiple parts, and sequencing analysis was performed for each part of the same tumor. As the results, they reported an intra-tumor phylogenetic structure and suggested the evolutionary process in tumor (Gerlinger *et al.*, 2012).

Despite these advances in the general understanding of inter-/intra-tumor heterogeneity and the process of tumor progression, the entire evolutionary process of tumors, from a benign to metastatic state, is still poorly understood. In the genome-sequencing study, I conducted a prospective cancer genome study to track the transitions from an early benign tumor to the metastatic tumor by using an experimental carcinogenesis system.

1.4 Cancer transcriptome analysis

While the genome sequencing and subsequent analysis enables identification of genomic mutations, high-throughput RNA sequencing (RNA-Seq) technology facilitates comprehensive detection and quantification of the gene expression in cells and tissues (Wang, Gerstein and Snyder, 2009). For RNA-Seq, reverse transcription reaction is performed for extracted template RNAs to be converted to cDNAs. The sequenced short-reads, yielded by the massively parallel sequencing of the ends of cDNA fragments, are mapped to the reference genome to quantify the gene expression levels. To quantify the gene expression levels, most of the analytical methods count the number of reads mapped to each gene region and normalize the read counts to comparable values. The most naïve normalization method is to divide the number of mapped reads by the total number of reads to correct the difference in the total number of reads between samples. Although, the gene expression is constantly fluctuating, it is considered as a stochastic process (Elowitz *et al.*, 2002); therefore, a statistical treatment is required for the quantification of gene expression levels including the read count normalization. Many analytical methods not only to quantify the gene expression levels but also to statistically compare the expression levels between different experimental conditions have been proposed. DESeq2 (Love, Huber and Anders, 2014), edgeR (Robinson, McCarthy and Smyth, 2010), and Cufflinks (Trapnell *et al.*, 2013) are widely used for gene expression analysis. For cancer transcriptome studies, these tools have been used to detect differentially expressed genes (DEGs) in tumor samples (Rhodes *et al.*, 2004). The DEGs in tumor samples are the candidates of biomarkers. The existing DEG detection tools can handle a pair of experimental conditions such as case and control samples, therefore they are appropriate for most of the previous cancer studies that compare gene expression levels between malignant tumors and normal tissues. In contrast, an analytical method applicable to multi-condition analysis is required for my study to reveal the entire process of tumor progression; however, few methods have been proposed and there is no single standard method. Accordingly, I focused on the clustering analysis that is one of the multivariate analysis methods.

For the multivariate analysis of the transcriptome data, clustering analysis according to gene expression patterns is widely used for patient subtype classification (van 't Veer *et al.*, 2002), and for searching the biological functions that are abnormally expressed in lesion sites (Prensner *et al.*, 2011). To cluster the genes whose expression patterns are similar to each other, it is required to measure the similarity between genes or gene clusters. Commonly, Euclidean distance, Pearson's correlation, and cosine similarity are used for the similarity measure between genes, and the average, complete linkage, and single linkage methods are used for the similarity measure between gene clusters (D'haeseleer, 2005). Since these existing similarity measures do not include statistical processing and are based on only the mean of expression levels, existing clustering methods cannot discriminate significant gene

expression differences from noises. Hence, in transcriptome-sequencing study, I first developed a novel clustering method based on statistical tests to extract significant gene clusters from multi-conditional transcriptome data. Second, I performed RNA-Seq for benign, malignant, and metastatic tumors obtained from the experimental carcinogenesis of mice. Finally, I applied the proposed method to the acquired RNA-Seq data to extract discriminative gene clusters in the process of tumor progression.

Chapter 2

Time-series analysis of tumorigenesis in a murine skin carcinogenesis model

In this chapter, I show the results of genome analysis for the tumorigenesis process (Aoto *et al.*, 2018).

Recent years have witnessed substantial progress in understanding tumor heterogeneity and the process of tumor progression; however, the entire process of the transition of tumors from benign to metastatic state remains poorly understood. In the present study, I performed a prospective cancer genome-sequencing analysis by employing an experimental carcinogenesis mouse model to systematically understand the evolutionary process of tumors. The collaborators surgically collected a part of a lesion of each tumor and followed the progression of these tumors *in vivo* over time. I conducted the comparative time-series analysis of the genomes of tumors with different fates, i.e., those that eventually metastasized and regressed. It suggested that these tumors acquired and inherited different mutations. Despite the occurrence of an intra-tumor selection event for malignant alteration during the transformation from early- to late-stage papilloma, the fate determination of tumors might be determined at an even earlier stage.

2.1 Background

2.1.1 Intra-tumor heterogeneity

As described in Chapter 1, cancer is a result of genomic disorders such as DNA mutations. Recent studies focusing on the intra-tumor environment suggest a polyclonal structure of tumors due to genomic instability (Greaves and Maley, 2012; Swanton, 2012; Burrell *et al.*, 2013). Although tumors initially form from a single cell type, as each tumor cell randomly acquires somatic mutations and then proliferates, the polyclonal cell population is formed based on the different genetic background among tumor cells (Greaves and Maley, 2012; Swanton, 2012; Burrell *et al.*, 2013). This genetic diversity of tumor cells generates the physiological diversity and difference in therapeutic sensitivity among tumor cells. Accordingly, the polyclonal structure of tumors is considered to be the most critical cause of treatment resistance and the recurrence of cancer (Greaves and Maley, 2012; Swanton, 2012; Burrell *et al.*, 2013).

More recently, the polyclonal structure of tumors has been addressed under the field termed “intra-tumor heterogeneity”, which has emerged as an essential aspect required for disclosing

the whole landscape of tumor progression and delineating the specific causes of resistance to cancer treatment. To best understand the full spectrum of intra-tumor heterogeneity, evolutionary analysis has been performed using multi-region samples, which obtained from the multiple sites of one malignant tumor, and/or the primary and metastatic tumors from the same animal; accordingly, several models have been proposed to explain the process of tumor progression and the origin of tumors (Gerlinger *et al.*, 2012; Martincorena and Campbell, 2015; Jamal-Hanjani *et al.*, 2017; McGranahan and Swanton, 2017). An adaptive (Darwinian) tumor progression model was suggested by which only certain sub-groups (i.e., sub-clones) that could gain advantageous traits to survive would remain in the tumor environment. In the process of stepwise cancer progression, tumors have to overcome several barriers such as the lack of nutrients, immune response from the surrounding tissues, and lack of the growth space, among others (Hanahan and Weinberg, 2011). These barriers impose a type of selection pressure for tumor cells so that only those that are best adapted to the given tumor environment will survive to proliferate and will acquire various mutations in the process to lead to intra-tumor heterogeneity. Such intra-tumor heterogeneity is considered to contribute to the diversity in treatment sensitivity and the sustainable progression of tumors (Greaves and Maley, 2012; Swanton, 2012; Burrell *et al.*, 2013; McGranahan and Swanton, 2017). In contrast, the neutral evolution theory of tumor progression proposes that tumor cells are derived from an initial malignant cell such as a cancer stem cell that neutrally expands with random mutations, leading to intra-tumor heterogeneity (Williams *et al.*, 2016; McGranahan and Swanton, 2017).

Moreover, several hypotheses have been proposed to explain the origin of tumors, including a founder cancer stem cell that already possesses multiple driver mutations and then rapidly grows and forms a tumor via the acquisition of new driver (trigger) mutations, or an initial driver mutation develops that causes an undetectable tumor, which gradually grows in size owing to the acquisition of new driver mutations (Martincorena and Campbell, 2015). Despite these advances in the general understanding of inter-/intra-tumor heterogeneity and the process of tumor progression, the entire evolutionary process of tumors, from a benign to metastatic state, is still poorly understood. Understanding the process of tumor evolution over time is expected to improve the confidence of early diagnostics and prognostic predictions. However, to date, the majority of cancer studies use specimens that have already transitioned to malignancy, and research based on tracking the transitions from an early benign tumor to the metastatic tumor is relatively limited.

2.1.2 Research purpose and murine skin carcinogenesis model

Given the above background, I have begun to address this issue with a prospective cancer study with the aim of systematically understanding the evolutionary process of tumors. Toward this end, I and my collaborators have employed an experimental carcinogenesis mouse model, which promotes the formation of squamous cell carcinoma (SCC) on the back

skin of the mice. The collaborators performed a classical two-stage carcinogenesis protocol to chemically induce the SCCs for mice using 7,12-dimethylbenz(a)anthracene (DMBA) and 12-O-tetradecanoylphorbol-13-acetone (TPA). In essence, the collaborators surgically collected a part of a lesion of each papilloma that formed, and followed the progression of each of these tumors *in vivo* over time. These identical tumors were repeatedly sampled at different growth stages to implement a time-series analysis from benign to metastatic tumors. During this sampling process, I and collaborators observed various fates of the tumors: some became malignant, others remained benign, whereas others regressed. Therefore, I decided to focus on these intra-tumor changes over time as well as on the inter-tumor differences according to comparison of tumors with different fates. To achieve a mechanistic understanding of these differences, I performed high-throughput deep-targeted genome sequencing for two tumor series that were collected from the same mouse, one that eventually became malignant and another that regressed, and conducted comparative time-series analysis to understand the evolutionary process of tumors and the factors that determine tumor fates.

SCC is one of the most common cancers in Caucasian populations, and the prognosis of metastatic SCC is extremely poor (Huang and Balmain, 2014). The DMBA-induced SCC model is one of the most commonly used *in vivo* models and is widely used for studying the mechanisms of metastasis (Ruddell *et al.*, 2008), which can be applied to both human and mouse cancers (Balmain and Yuspa, 2014; Huang and Balmain, 2014). High-throughput genome sequencing analysis has previously been conducted for DMBA-induced mouse skin SCCs (McCreery *et al.*, 2015; Nassar *et al.*, 2015). Nassar *et al.* (Nassar *et al.*, 2015) investigated the mutational landscape of DMBA-induced SCCs to reveal the substitution patterns of the DMBA-induced somatic mutations and the genes recurrently mutated in this experimental model. McCreery *et al.* (McCreery *et al.*, 2015) performed phylogenetic analysis for the benign, malignant, and metastatic tumors obtained from the same mouse to reveal the origins of metastases and to indirectly predict the process of tumor evolution. Although these studies reported significant results, almost all of the DMBA-induced SCCs were harvested from metastatic tumors and premalignant tumors at the same time when the mice were sacrificed. Although this sampling method still enabled analyzing tumors at different malignant stages and associating the primary SCC samples with metastatic samples, associations of the premalignant samples could not be determined because their origins are different from those of the SCC samples. Therefore, these previous studies could not eliminate the influence of inter-tumor heterogeneity and could not follow the tumorigenesis process of the same tumor over time. In addition, these previous studies performed whole-exome sequencing and obtained approximately 50× mean coverage sequencing data, whereas Shin *et al.* (Shin *et al.*, 2017) indicated that approximately 100× sequencing coverage is required to detect mutations possessed by ~10% of cells in a tumor from clinical samples. Moreover, the known cancer-related gene mutations have been detected with low penetrance in tumor samples due to the influence of intra-tumor heterogeneity (Gerlinger *et*

al., 2012; McGranahan *et al.*, 2015) and low tumor content (Shin *et al.*, 2017). Thus, deep sequencing coverage is necessary for comprehensive mutation analyses; however, the previous data are insufficient to detect low-penetrance variants and to estimate the penetrance in a tumor precisely. With regard to these limitations, my approach, involving partial surgery and targeted deep sequencing, can contribute to gaining a more comprehensive understanding of the tumorigenesis process in different aspects from those obtained in previous research.

2.2 Results

2.2.1 Mutational landscape of two tumor series with different fates

Collaborators established the experimental carcinogenesis model using 35 mice for sample collection. One of the main features of the chemical induction mouse skin cancer model is that a large number of benign tumors are formed in the early term, only a few tumors ultimately become malignant (Schwarz, Münzel and Braeuning, 2013). The collaborators previously reported that DMBA-TPA carcinogenesis induction forms approximately 30 papillomas on the back skin of each mouse (Okumura *et al.*, 2012). In the present study, they selected 5–10 DMBA-induced papillomas for each mouse to follow the progression over time. The size and the position of each papilloma was recorded during the carcinogenesis experiment, and a part of the papilloma lesion was surgically collected to monitor the progress of the same tumor over time. This partial tumor sampling was performed from the early stage of tumor formation at the ninth week after starting the carcinogenesis experiment. By the time all of the mice were sacrificed, a total of 113 tumors were collected at multiple time points: 8 (7.1%) tumors became malignant, 68 (60.2%) tumors were still benign, and the other 37 (32.7%) tumors had eventually regressed. To understand the differences between the tumors that eventually metastasized and those that eventually regressed, I performed a comparative time-series analysis for two prospectively sampled tumor series obtained from the same mouse: one was sampled at four time points and had eventually metastasized (hereafter “malignant series”), and the other was sampled at two time points and eventually regressed (hereafter “regressed series”). Conceptual figure of tumor samples is shown in Fig. 2.1a (Supplementary Fig. S2.1 for photographic images). For simplicity, each sample was named as described in the respective images shown in Fig. 2.1a. The collaborators collected axillary lymph node metastases as metastatic tumor samples, and a matched tail sample was used as a control tissue.

For the sensitive detection of mutant alleles from the heterogeneous tumor samples and for accurate estimation of the penetrance of each mutation in the tumors, a capture sequencing approach was adopted, which can achieve deep-sequencing coverage for targeted genes (see Supplementary Table S2.1 for the list of the targeted genes and see Table 2.1 for sequencing statistics). The single-nucleotide variants (SNVs) were detected using three variant callers

and then filtered according to several criteria to minimize false positives. I defined candidate mutation positions as the positions at which SNVs were detected in at least one tumor sample, and I focused on the 83 candidate mutation positions with $>50\times$ coverage in all seven samples (see Table 2.2 for the sequencing depth of these positions). Next, the variant allele frequencies (VAFs) were calculated for each of the 83 candidate mutation positions in each tumor sample. The VAF is a percentage of variant alleles in the short reads that are mapped to a certain genomic position; therefore, it can be considered as an index of the penetrance of a mutation in a tumor, and can serve as a clue to infer the intra-tumor heterogeneity and sub-clonal structure in tumors (McGranahan *et al.*, 2015; Williams *et al.*, 2016). The VAFs of the positions with less than 3 reads supporting the mutation were filtered out (set to 0), and were used to estimate the tumor purity for each tumor sample.

Table 2.1. Sequencing statistics for each sample.

Sample name	# clusters	# reads	# QC	%QC	%duplicates
Papi9w	29,494,118	58,988,236	35,567,485	60.30	33.99
Papi14w	26,126,652	52,253,304	33,379,356	63.88	28.13
Carc	28,909,899	57,819,798	39,688,768	68.64	23.08
Meta	31,889,353	63,778,706	36,805,260	57.71	34.39
Reg9w	40,090,049	80,180,098	51,276,309	63.95	24.28
Reg14w	31,473,369	62,946,738	32,821,844	52.14	37.64
Tail (ctrl)	31,218,916	62,437,832	30,439,480	48.75	45.13
Ave.	31,314,622			59.34	32.38

The sample names are corresponded to those of Fig. 2.1a. The column #QC (quality control) denotes the number of reads that remained after the pre-processing to exclude the low-quality sequenced reads. The column %QC denotes the percentage of reads remaining after the pre-processing. The column “%duplicates” denotes the percentage of the duplicated reads for each sample, which was calculated from the number of filtered reads in the de-duplicate process.

Table 2.2. Sequencing depth of 83 candidate mutation position for each sample.

Sample name	Max	Min	Average	Median
Papi9w	976	52	349.76	265
Papi14w	2152	52	461.87	332
Carc	1093	72	420.64	319
Meta	1133	65	438.46	309
Reg9w	2165	55	543.08	459
Reg14w	2654	58	527.72	368
Tail (ctrl)	1236	55	407.33	260

The statistics of the coverage of the 83 mutation candidate positions for each sample are shown.

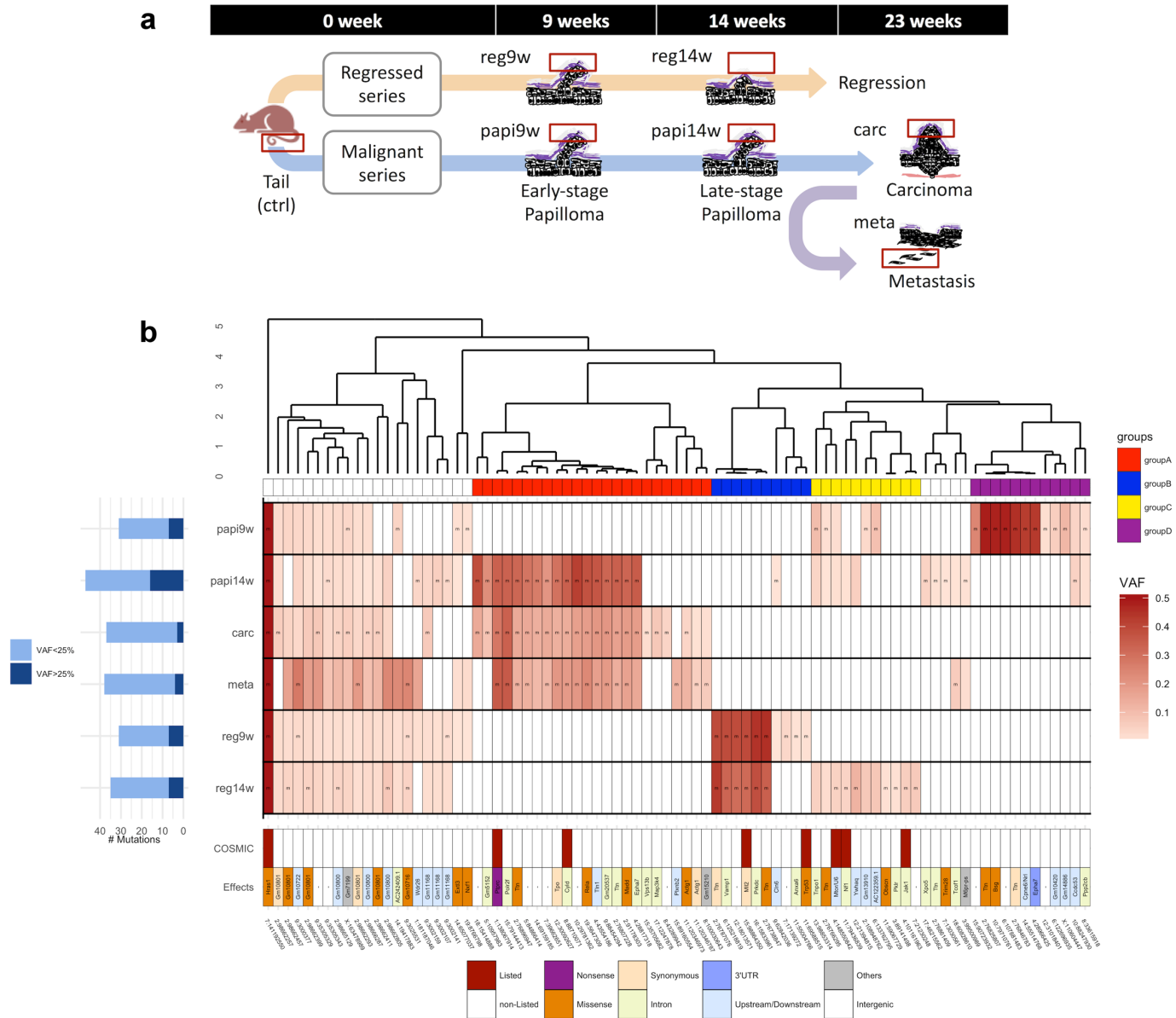


Figure 2.1. Individual mouse of interest and landscape of the mutational status. **(a)** The mouse developed two tumor series: one eventually regressed and the other eventually metastasized, designated “regressed series” and “malignant series”, respectively. For simplicity, each sample was named as shown near the respective images. **(b)** The overall mutational status of 83 genomic positions for each tumor sample. The bar plot shows the number of positions for which mutant alleles were detected. In the tables on the right, each column corresponds to a certain position. The upper table shows the heatmap of the VAFs with the hierarchical tree color-coded by the mutation groups; “m” denotes that the position was detected by the mutation callers as an SNV. The lower table shows the annotation of the mutations: the first line indicates whether the mutated position is included in a gene listed in the COSMIC: Cancer Gene Census (Futreal *et al.*, 2004), and the second line indicates the gene name, including the mutated position and the effect of the mutation.

The heterozygous Q61L *Hras* mutation (chr7:141192550) is known as an initiator of the DMBA-TPA experimental carcinogenesis protocol I employed (Brown, Buchmann and Balmain, 1990; McCreery *et al.*, 2015; Nassar *et al.*, 2015); therefore, the tumor purity was estimated based on the VAF of the *Hras* mutation (Table 2.3). Correspondingly, the raw VAFs were normalized to the estimated tumor purities for proper estimation of the penetrance of the mutations in the tumors while excluding the influence of the normal cell-derived reads. Finally, the candidate mutation positions with a normalized VAF $\neq 0$ were presumed to be the mutated positions for each tumor sample.

Table 2.3. Tumor purities and normalization factors.

Sample name	# Mapped reads	# Variant reads	Raw VAF (%)	95% confidence interval (%)	Tumor purity (%)	NF
Papi9w	912	342	37.500	34.35 – 40.73	68.70 – 81.47	1.3333
Papi14w	2152	974	45.260	43.14 – 47.39	86.28 – 94.78	1.1047
Carc	910	277	30.440	27.46 – 33.54	54.92 – 67.09	1.6426
Meta	836	62	7.4163	5.733 – 9.411	11.47 – 18.81	6.7419
Reg9w	1271	648	50.983	48.20 – 53.77	96.39 – 100.0	0.9807
Reg14w	1553	522	33.612	31.26 – 36.02	62.53 – 72.05	1.4875

The number of mapped reads on the Q61L *Hras* mutation position (chr7:141192550), the number of variant reads which support the *Hras* mutation, the raw VAF of the *Hras* mutation, the 95% confidence interval of the VAF, estimated tumor purity, and normalization factor (NF) for each tumor sample are shown in this table. The heterozygous *Hras* mutation is an initiator of the experimental carcinogenesis protocol we used; therefore, the VAF of the *Hras* mutation should be 50%. Then, I calculated the normalization factor by $0.5 / (\text{raw VAF})$. I used up to the seventh decimal place of the NFs to minimize the effect of rounding errors when I calculated the normalized VAFs.

Figure 2.1b shows the overall mutational status for the 83 candidate mutation positions in each tumor: the annotation of mutations, heatmap of the normalized VAFs with the hierarchical tree, and number of mutated positions are shown (see Supplementary Table S2.2 for more details). First, the number of mutated positions tended to increase during the benign stages in both tumor series; but decreased during the malignant stages. This indicates that the cells in the tumor acquired new mutations during the benign state, whereas the cells with

several private mutations could not survive after malignancy, suggesting a selective process of malignant alteration.

Next, I evaluated the commonality of the mutations detected between the tumor types to better understand the nature of the inter-tumor heterogeneity. From the heatmap with a hierarchical tree shown in Fig. 2.1b, 9/83 (10.84%) mutations were observed in all tumor samples, and the other 74/83 (89.16%) mutations were privately observed or sparsely shared among tumors. There were mutation groups inherited in each tumor series; in particular, two mutation groups were mutually exclusively detected in each tumor series (groups A and B): the mutations belonging to group A were observed after 14 weeks in the malignant tumor series, whereas the mutations belonging to group B were observed only in the regressed tumor series. These mutation groups might reflect the sub-clonal structure in the tumors. Specifically, an intronic mutation of *Cyld* and a nonsense mutation of *Ptprc* were included in mutation group A. The *Cyld* gene is known as a deubiquitinase whose deubiquitination activity negatively regulates the NF- κ B pathway and suppresses epidermal tumor progression (Kovalenko *et al.*, 2003; Alameda *et al.*, 2010). In addition, *Ptprc* encodes CD45, which regulates the phosphorylation of SRC and JAK family members (Porcu *et al.*, 2012), and nonsense mutations in this gene have been recurrently reported in human epithelial and blood cancers (<http://cancer.sanger.ac.uk/cosmic>). Moreover, both of these genes are listed in the COSMIC: Cancer Gene Census (CGC) (Futreal *et al.*, 2004), a catalogue of cancer-related genes. These two genes showed a high penetrance rate in the tumor samples of the malignant tumor series at the later stages (Fig. 2.1b), suggesting that inactivation of these genes caused by mutations could have driven the progression of tumor.

Moreover, a synonymous mutation in *Mll2* and a missense mutation in *Trp53* were included in mutation group B. Mutations of these genes have been reported to promote cancer progression (Futreal *et al.*, 2004), especially the mutation of the *Trp53* tumor suppressor gene, which has been widely reported as a driver mutation in several cancer types (Kandoth *et al.*, 2013; Vogelstein *et al.*, 2013). *Trp53* is a well-known tumor suppressor gene that plays an important role for DNA repair function, cell cycle arrest, and apoptosis (Page *et al.*, 2016). *Mll2* is a histone methyltransferase and its mutants provoke the genomic instability in tumors (Kantidakis *et al.*, 2016). There are reports that the *Trp53* mutations induce not only abnormal proliferation in tumors but also anti-tumor effects (Einspahr *et al.*, 1999; Page *et al.*, 2016), and these reports had also discussed the influence of the *Trp53*-independent apoptosis. Therefore, I considered the possibility that these mutations, in *Trp53* and *Mll2* genes, which are related to genomic instability, could have induced the *Trp53*-independent apoptosis, which in turn resulting in the eventual regression of the tumor.

Another mutation group, group C, showed an opposite pattern to that of the other two tumor series. The tumor of the malignant series acquired group C mutations when it was benign, but did not inherit these mutations after malignant transformation. By contrast, the tumor of the regressed series acquired the mutations at a later stage of the benign term. The mutations of genes listed in the CGC as cancer-related genes, although in non-exonic regions, were

enriched in group C. For instance, downstream mutation of *Mtor* and intronic mutations of *Nf1* and *Jak1* were included in the group C, and these genes have a crosstalk and/or direct regulatory relationship with the RAS signaling pathway (Bollag *et al.*, 1996; Rawlings, Rosler and Harrison, 2004; Laplante and Sabatini, 2013). In comparison, group D is a mutation group that was observed only in the early stages of the malignant tumor series, and there are no gene mutations listed in the CGC. This suggests that the peripheral genomic region of cancer-related genes was more actively mutated in the regressed tumor series. Furthermore, focusing on the malignant tumor series, the mutations showing relatively high VAFs in the 9th week papilloma group (group D) were not detected at the later stages, and the mutations conserved after the 14th week of papilloma development (group A) were not detected in the 9th week; hence, it appears that the 9th to 14th week represents a period of substantial changes in the intra-tumor sub-clonal structure for malignant alteration. Thus, the mutational status among tumors suggests that the sub-clonal structure in the tumors continuously changes over time and also differs between tumors with different fates.

Furthermore, to evaluate the adequacy of these results, I reviewed previous studies using the same mouse carcinogenesis protocol. Nassar *et al.* (Nassar *et al.*, 2015) and McCreery *et al.* (McCreery *et al.*, 2015) performed large-scale analyses of SCC induced in the skin of mice with DMBA-TPA, and identified recurrent mutations among the SCCs, including exclusive driver mutations of the RAS family, and revealed the mutational signature of DMBA enriched with A>T base substitutions. In my results, all of the tumors had the Q61L *Hras* mutation, which was reported as the initial driver mutation for this model (Brown, Buchmann and Balmain, 1990; McCreery *et al.*, 2015; Nassar *et al.*, 2015), and the substitution pattern of the 83 candidate mutation positions showed enrichment for the A>T base substitution (Fig. 2.2); therefore, my results are consistent with previous studies of this model.

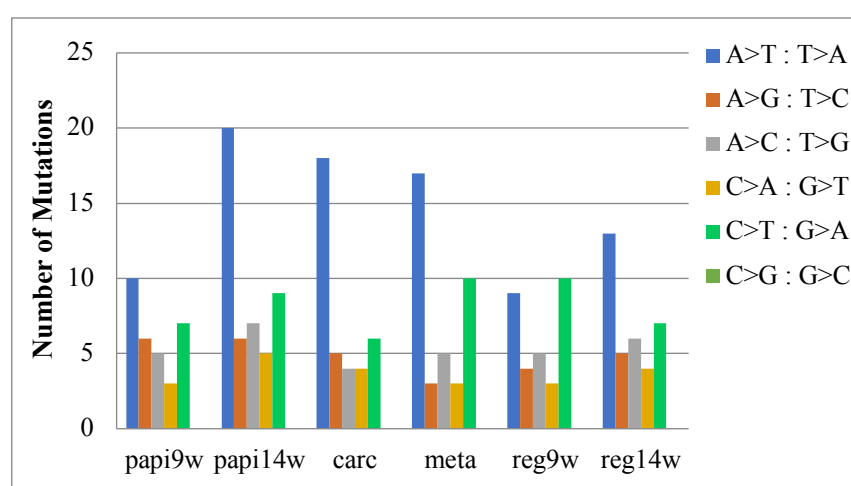


Figure 2.2. Substitution patterns of SNVs. The detected SNVs were categorized by the substitution patterns. The bar-plot denotes the count of each substitution pattern for each tumor sample.

Of note, the previous studies identified a recurrent mutation in the *Trp53* gene in the malignant SCCs, whereas this mutation was only detected in the early stage of the regressed tumor series in the present study. This indicates that the single *Trp53* mutation does not promote tumor progression. By contrast, the stop-gain mutation of *Ptprc*, which belongs to the protein tyrosine phosphatase (PTP) family, was inherited in the malignant tumor series. Nassar et al. (Nassar *et al.*, 2015) also identified another gene belonging to the PTP family, *Ptpm*, that showed recurrent mutation; hence, the PTP family appears to play a key role in the DMBA-TPA experimental carcinogenesis system.

The results highlighted thus far exhibited an increase in the number of mutations while the tumor is in the benign stage and later a decrease after malignant alteration. The tumor series with different fates, i.e., becoming malignant and eventually regressing, acquired and inherited different mutations; however, not all mutations were inherited, and the VAFs also changed over time. These results suggest that selection pressure was applied to induce the process of malignant alteration, and the sub-clonal structure not only differs between tumor fates but also changes over time.

2.2.2 Degree of shared mutations between tumors within and among tumor series

Subsequently, I evaluated the similarity (or dissimilarity) between tumors by applying the Jaccard similarity coefficient and the generalized Jaccard similarity coefficient (Fig. 2.3). For this analysis, I assumed the positions with a VAF $\neq 0$ of the 83 candidate mutation positions were mutated, and defined a set of mutations for each tumor sample. To determine the commonality of the mutated positions between tumors, I calculated the Jaccard similarity coefficient between the sets of mutations. The heatmap of the Jaccard similarity coefficient (Fig. 2.3a) indicated that the similarity between tumors within the same tumor series was higher than that between tumor series; however, the similarity between the 9th week and later papilloma samples of the malignant tumor series was only moderate. This result supports that different mutations are inherited in each tumor series, and confirms that the intra-tumor structure existing in the malignant tumor series dramatically changed between the 9th to 14th weeks, as proposed from the results of the mutational status analysis (Fig. 2.1b).

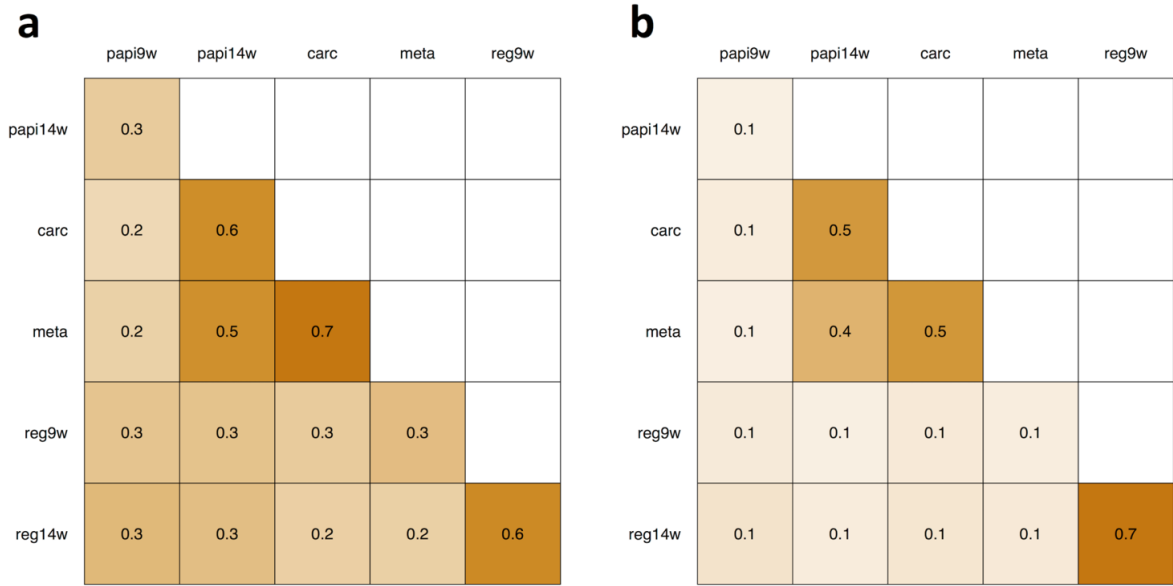


Figure 2.3. Heatmap of similarity between tumors. **(a)** Heatmap of the Jaccard similarity coefficient between tumor samples. **(b)** Heatmap of the generalized Jaccard similarity coefficient between tumor samples.

Moreover, I applied the generalized Jaccard similarity coefficient to quantitatively evaluate the inheritance of the mutations. I used the normalized VAF value to represent the penetrance of a mutation in a tumor, and defined a numeric vector constructed by the normalized VAFs of the 83 candidate mutation positions for each tumor sample. The generalized Jaccard similarity coefficient between these numeric vectors then corresponds to the commonality of the abundances of the mutations between tumors. The heatmap of the generalized Jaccard similarity coefficient between tumors (Fig. 2.3b) not only accentuates the same features represented in Fig. 2.3a but also indicates that the penetrance of the mutations was more highly conserved between tumors of the regressed tumor series than that between tumors of the malignant tumor series. Consistently, the scatter plot of the normalized VAFs between tumors (Fig. 2.4) showed that the plots between tumors with relatively low generalized Jaccard similarity coefficients were clearly biased toward both axes, whereas the plots between tumors with higher similarity coefficients, especially those between tumors within the regressed tumor series, basically lay on the diagonal line. These results suggest that tumors of the regressed series were relatively stable over time, whereas the intra-tumor structure of the malignant tumor series was crucially changed between the 9th to 14th weeks, and then continuously changed gradually after the 14th week.

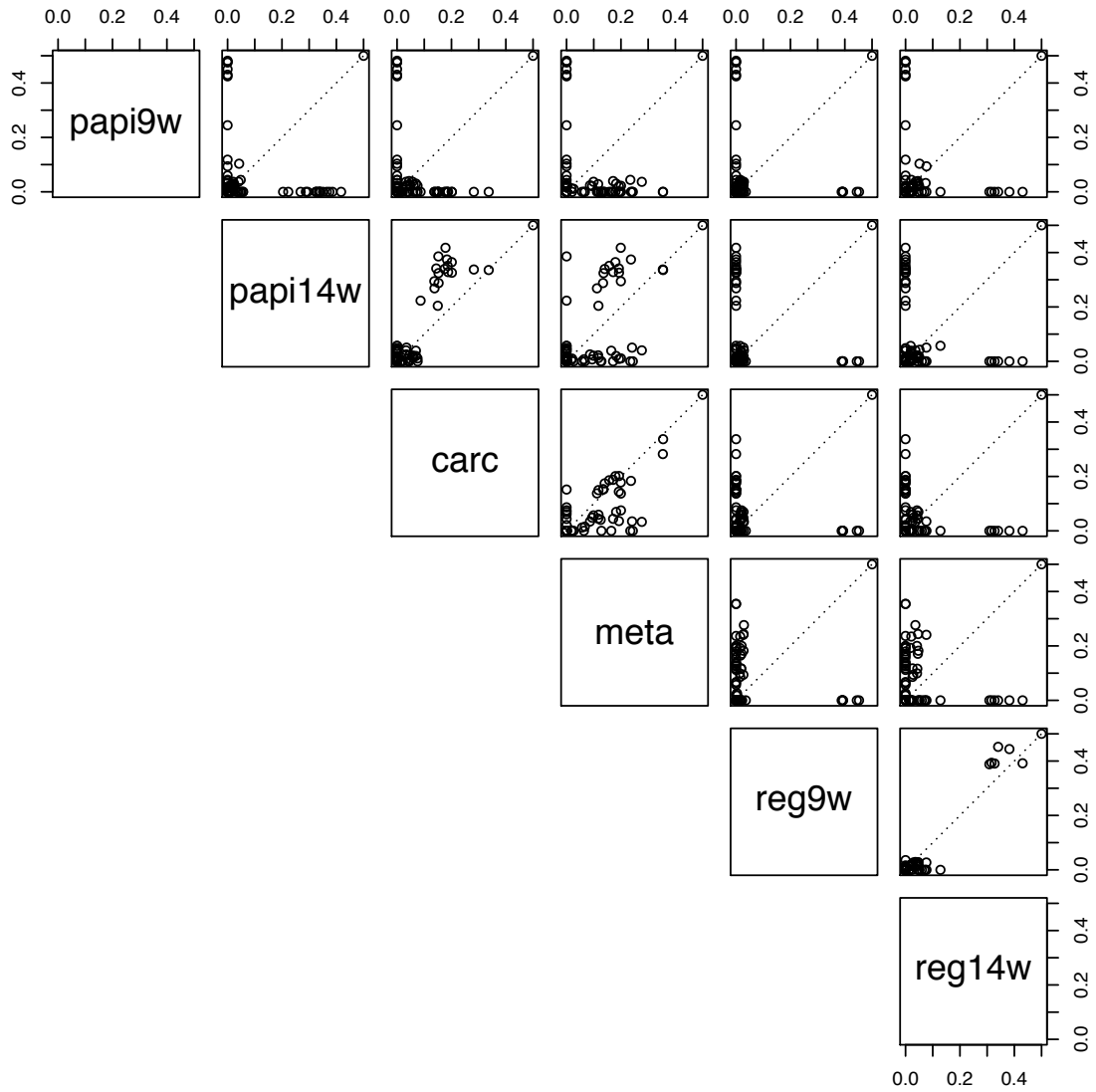


Figure 2.4. Scatter-plot of VAFs of 83 candidate mutation positions. VAFs of the 83 candidate positions were plotted. Each dot corresponds to a certain genomic position so that 83 plots are drawn in each of the boxes. If the VAFs of a certain position was same value between tumors, the corresponded dot should be on the diagonal line.

2.2.3 Intra-tumor heterogeneity and evolutionary process within tumors

Next, I focused on the intra-tumor heterogeneity. To evaluate the degree of intra-tumor heterogeneity, I applied an entropy parameter for the normalized VAFs of the 83 candidate mutation positions. Entropy is often used as an index of the complexity or information

content of certain events. Dr. Shannon proposed an expected value of the information amount as “information entropy” (Shannon, 1948), and this concept has been widely applied in the fields of informatics, statistics, and biology (Newton *et al.*, 2014). In the present study, I regarded the intra-tumor heterogeneity to reflect the complexity of the tumor, which was evaluated by the sum of the complexity calculated for each mutation.

As shown in Fig. 2.5, the spectrum of the normalized VAFs demonstrated that all normalized VAFs were below the 50% line (the plot on the 50% line in each of the tumor samples corresponds to the *Hras* mutation). The genomic positions with a normalized VAF of 0% indicates that the position is homozygous to that of the reference base in all tumor cells, while the positions with a normalized VAF of 50% indicates that the position is heterozygous in all tumor cells; in other words, these positions are homogeneous among tumor cells. Based on this result, I assumed that there was no copy number variation (CNV) at the 83 candidate mutation positions and that these mutations occurred on one side of the diploids. Moreover, the clonal structure of a tumor can be represented by a combination of mutations; therefore, I assumed that the mutations are independent but do not contradict each other. With these assumptions, I defined an entropy parameter as the degree of complexity of a tumor (see Materials and Methods for more details), which was calculated from the normalized VAFs of the heterogeneous positions for each tumor sample (Fig. 2.5, bottom). As a result, the entropy was found to be maximal in the 14th week papilloma sample, and then decreased after becoming malignant in the malignant tumor series. Moreover, the entropy value of the regressed tumor series decreased over time, and was lower than that of the 14th week papilloma sample of the malignant tumor series. These results suggest that the tumors of the malignant series gradually acquired intra-tumor heterogeneity before developing malignancy, whereas the tumors of the regressed series could not sustain the heterogeneous characteristic.

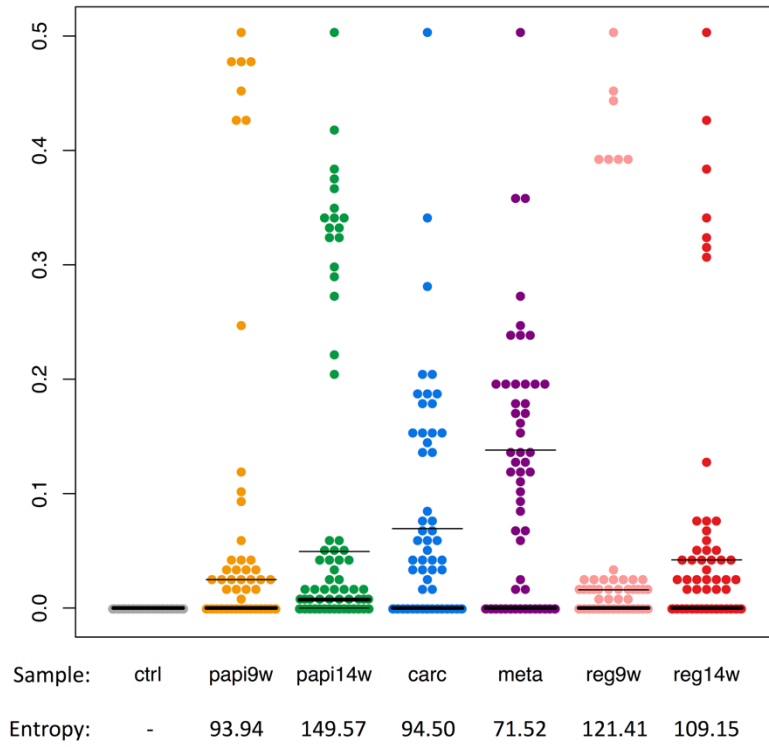


Figure 2.5. Spectrum of normalized VAFs and entropy of each tumor sample. The chart shows the spectrum of the normalized VAFs, where each column represents a certain sample, and the y-axis corresponds to the normalized VAF. Each dot corresponds to a certain candidate mutation position for a total of 83 dots plotted for each column. The values shown at the bottom of each column denote the entropy parameter values.

Furthermore, I estimated the mutation rate to obtain insight into the process of sub-clonal evolution in the tumors. Under the assumption that mutations are neutrally accumulated and penetrate a tumor over time, the number of mutations should proportionally increase with time and the mutation rate of a tumor cell should be constant. Williams et al. (Williams *et al.*, 2016) proposed a regression analysis method to estimate the mutation rate from VAFs. According to their results, I assumed that the position with a normalized VAF of 25% or less in the 83 candidate mutation positions represented a sub-clonal mutation. For each tumor sample, the sub-clonal mutations were sorted in ascending order according to the inverse of the normalized VAFs, and the normalized VAFs were plotted against these sorted ranks, which were used to perform the regression analysis respectively for the ranks (Fig. 2.6). The slope of the regression line denotes the estimated apparent mutation rate (shown at the right bottom of each plot), and the coefficient of determination (R^2) denotes the neutrality of the mutations (shown at the left top of each plot). When $R^2 \geq 0.98$, the tumor was considered to have acquired mutations neutrally; otherwise, the tumor was considered to have acquired mutations in a non-neutral manner (i.e., selectively). First, the mutation rates of the benign tumors were lower than those of the tumors that transitioned to a malignant status; this effect was particularly notable in the 9th week early-stage papilloma sample of the regressed tumor series. For the malignant tumor series (Fig. 2.6a–d), the mutations were neutrally acquired

until the 9th week papilloma, but were then acquired in a biased manner after the 9th week papilloma; notably, an early selection mode (Williams *et al.*, 2016) was suggested for carcinoma and metastasis progression, indicating that the tumor had already gained dominant (selected) mutations (sub-clones) before transitioning to carcinoma. In brief, the tumor was exposed to the selection pressure after 9 weeks before becoming malignant. According to the mutational landscape (Fig. 2.1b) and the similarity between tumors (Fig. 2.3), the intra-tumor structure of the malignant tumor series was dramatically changed between the 9th and 14th weeks, and then was comparatively conserved thereafter. Thus, the selection pressure that had been in effect from the 9th to 14th weeks might have caused the dramatic changes in the tumor. By contrast, for the regressed tumor-series (Fig. 2.6e–f), the tumor had already been acquiring the mutations selectively before the 9th week; however, the process of mutation acquisition after the 9th week was comparatively neutral. These results suggest that different tumor series undergo distinct evolutionary processes, indicating that the tumor fate is determined from a quite early stage.

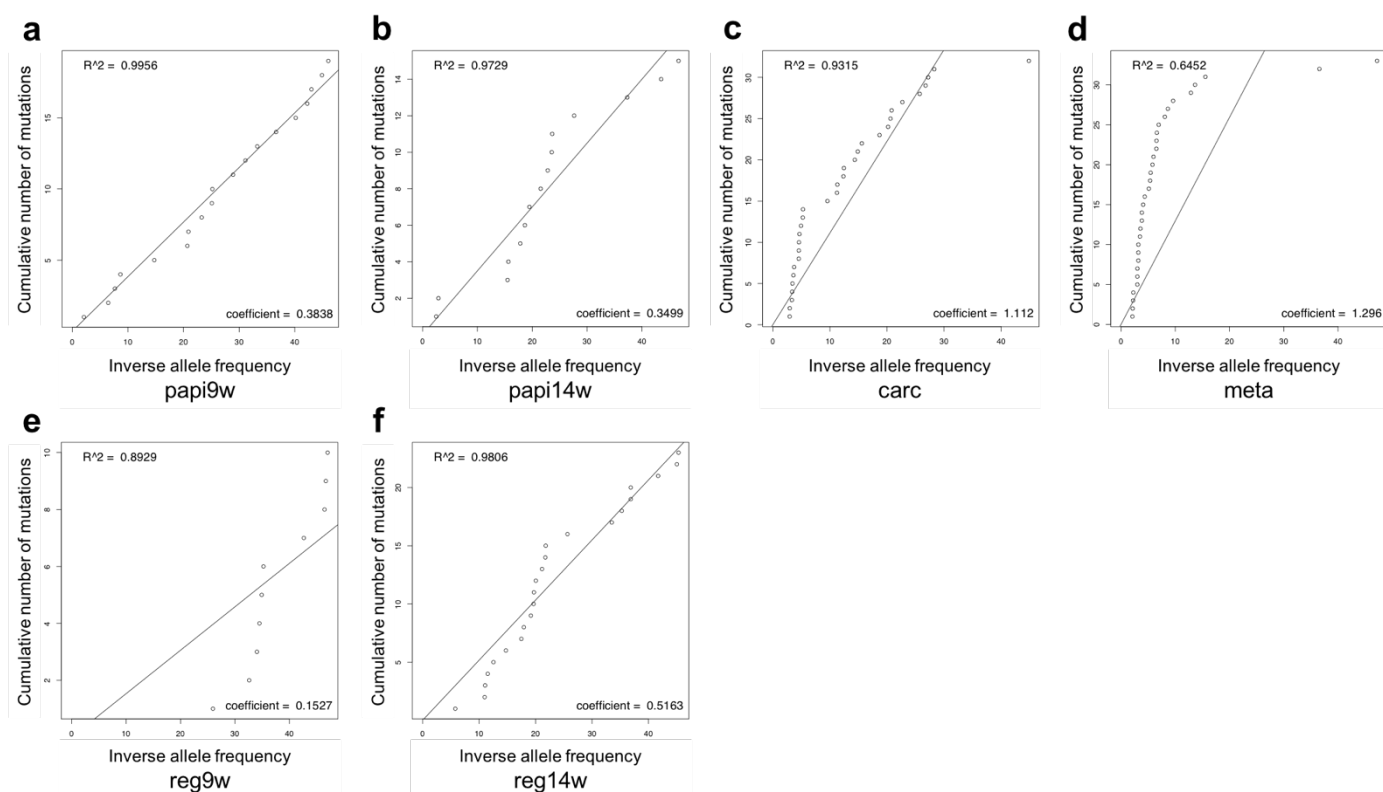


Figure 2.6. Mutation rate analysis for the malignant tumor series (a–d) and the regressed tumor series (e, f). The x-axis corresponds to the inverse allele frequency and the y-axis corresponds to the cumulative number of mutations. The apparent mutation rate is shown at the right bottom of each plot, and the coefficient of determination (R^2) is shown at the left top of each plot.

2.3 Discussion

In this study, I compared tumor series with different fates through analysis of a DMBA-TPA carcinogenesis protocol. Although the DMBA-TPA protocol is limited to the analysis of the tumors initiated by RAS family mutation, it is widely accepted that the mutation landscape, mutation burden, and gene expression profile of the SCCs formed on the mouse back skin through the DMBA-TPA protocol are largely similar to human SCCs such as head and neck, esophageal, lung, and cervical SCCs (Ginos *et al.*, 2004; Huang and Balmain, 2014; McCreery *et al.*, 2015; Nassar *et al.*, 2015; Aoto *et al.*, 2017). In line with these previous studies, my results indicated that although the carcinogenesis protocol using DMBA-TPA provides information on the initial driver mutations and promotes the formation of a number of papillomas on the mouse skin, further drivers are clearly required for the malignant alteration. In most cases, papillomas could not become carcinomas, meaning that further driver events did not occur, while also confirming the general difficulty of malignant transformation. I further presumed that there are differences in the genomic mutations and in the sub-clonal structure among tumors with different fates. To understand the difference between the eventually metastasized and regressed tumors, I performed targeted deep sequencing for these two tumor-series of different destinies sampled from the same mouse. This comparative analysis provided an opportunity to understand the factors and timing of the tumor fate determination.

Previous cancer studies reported that each tumor has a different mutational status and multiple driver mutations are accumulated during tumor progression (Futreal *et al.*, 2004; Vogelstein *et al.*, 2013; Martincorena and Campbell, 2015; Tomczak, Czerwińska and Wiznerowicz, 2015). To understand the evolutionary process of tumors, the temporal changes of the mutational trend, e.g., the order of the occurrence of driver events, have also been discussed (Fearon and Vogelstein, 1990; Vogelstein *et al.*, 2013; McGranahan and Swanton, 2015, 2017; McGranahan *et al.*, 2015; Shain *et al.*, 2015; Jamal-Hanjani *et al.*, 2017). These studies identified that known driver mutations were significantly biased to be clonal; hence, it has been considered that these driver mutations are acquired before malignant transformation, and the highly malignant sub-clones with these driver mutations originate to form malignant tumors (Greaves and Maley, 2012; Martincorena and Campbell, 2015). Consistent with these previous hypotheses, my results suggest that tumors with different fates acquired different mutations, accompanied by changes in the intra-tumor structure over time, and also suggest that selection pressure could have influenced the benign tumors to select highly malignant sub-clones in the tumors.

Moreover, my results provide new insight into the timing of the selective event and fate determination of tumors. The selection event for malignant alteration might have occurred between the 9th and 14th weeks, corresponding to the transformation from early-stage papilloma to late-stage papilloma, and the evolutionary process of these two tumor-series

differed at the time of the 9th week early-stage papilloma. Furthermore, only the tumors that could obtain and sustain sufficient intra-tumor heterogeneity could become malignant. The majority of cancer studies conducted to date have focused on malignant tumors (Tomczak, Czerwińska and Wiznerowicz, 2015). Although several studies analyzed tumors at both the benign and malignant stages, these tumor types are typically collected from different patients (Jamal-Hanjani *et al.*, 2017), or the adjacent lesions of primary tumors were collected as precursor lesions (Shain *et al.*, 2015); therefore, these studies could not follow the temporal changes of the same tumors respectively. Thus, my prospective approach is complementary to existing cancer studies, and offers novel insights.

Adoption of this prospective study model can provide new opportunities for gaining a detailed understanding of the evolutionary process of tumors by conducting statistical analysis with increased numbers of mice and tumor series, and also by analyzing the whole genome rather than targeted regions. Moreover, my study can also be relevant to improving surgical treatment resistance and prognostic prediction (Ludwig and Weinstein, 2005). In particular, my results and further statistical analysis with more mice can help to identify factors that determine the fate of tumors and the prognostic biomarkers for guiding a more precise treatment strategy. However, the gene signatures and the timeline of tumor progression might vary in different tumor samples; therefore, further experiments are required to verify these aspects. As part of my future works, I am planning to perform whole-exome sequencing for more tumor series, and other omics analyses such as transcriptomics and epigenomics.

2.4 Materials and Methods

2.4.1 Mice

This study was conducted in strict accordance with the recommendations in the Guide for the Care and Use of Laboratory Animals of the Ministry of Education, Culture, Sports, Science, and Technology of Japan. The protocol was approved by the Committee on the Ethics of Animal Experiments of Chiba Cancer Center (Permit Number: 17-14). All efforts were made to minimize suffering. FVB/N mice were purchased from CLEA Japan (Tokyo, Japan).

2.4.2 Skin carcinogenesis and tumor sampling

DMBA was used as the carcinogen and TPA was used as a promoter to induce SCC on the back of the mouse skin. Collaborators treated 35 FVB/N mice according to a two-stage carcinogenesis protocol. At 8 weeks of age, the female mice were carefully shaved with an electric clipper, and two days after shaving, a single dose of DMBA (25 µg/mouse in 200 µL of acetone) was applied to the shaved dorsal back skin. One week after initiation, tumor

growth was promoted with TPA (10 μ g/mouse in 200 μ L of acetone) twice weekly for 20 weeks. For each mouse, 5–10 papillomas were selected to follow tumor progression over time. The number, size (diameter in mm), and position of each papilloma were recorded as of 9, 14, and 23 weeks, and carcinoma development was monitored for up to 33 weeks post-TPA treatment. To obtain the prospective tumor samples from the same tumor, a part (approximately half) of the papilloma lesions was surgically collected so that the remaining part could grow toward the later stages. These collected samples were further cut in half; therefore, one quarter of each tumor was used for this research, and the remaining quarter was preserved for subsequent research. Collaborators determined benign tumors (papillomas) and malignant tumors (carcinomas) by visual inspection. Papillomas appear as outgrowths on the mouse back skin. Some of them become flattened on the skin with malignant transformation involving penetration of the deep dermis. In general, these tumors can be easily distinguishable from one another (Okumura *et al.*, 2012; Huang and Balmain, 2014), and I did not use any of the tumors that could not be clearly classified as papilloma or carcinoma, such as an intermediate type. The mice were sacrificed by cervical dislocation when carcinoma formed and axillary lymph node metastases were visually confirmed, or when the tumor volume reached 10% of the mouse body weight. They collected axillary lymph node metastases as metastatic tumor samples. As a control tissue, they collected a tail sample from each mouse before starting the DMBA-TPA experiment. Finally, the time-series tumor samples that eventually became malignant and those that regressed were obtained from the same mouse (Fig. 2.1a and Supplementary Fig. S2.1).

2.4.3 Sample preparation and targeted deep sequencing

For sensitive detection of the low-frequent mutant alleles and for precise estimation of the VAFs, I adopted a targeted capture-based deep-sequencing approach. I and collaborators compiled a list of the target genes by reference to the COSMIC database (<http://cancer.sanger.ac.uk>) (Forbes *et al.*, 2017) to select the most frequently mutated genes in the SCC samples, and used previous reports of the housekeeping genes (Eisenberg and Levanon, 2003; Vilella *et al.*, 2009) to select the genes deemed to be most crucial for cell survival. Finally, we selected 500 genes and designed the target capture bait library using the Agilent SureDesign program (<https://earray.chem.agilent.com/suredesign>; see Supplementary Table S2.1 for details).

For sample preparation, the genomic DNA was extracted from the tumors and normal tissues (tail of each mouse) using DNAiso Reagent (9770A; Takara, Otsu, Japan). The extracts were treated according to the SureSelect^{XT} Target Enrichment System for Illumina Paired-End Sequencing Protocol for 200-ng samples (Agilent Technology, Santa Clara, CA, USA). In this step, pre-capture polymerase chain reaction (PCR) was performed for 10 cycles, the incubation for library hybridization was performed for 24 h, and post-capture PCR was performed for 12 cycles. The products were confirmed using the Agilent 2100 bioanalyzer.

The paired-end sequencing (2×75 bp) was performed using Illumina MiSeq System with MiSeq Reagent Kit v3 (150 cycle; Illumina, San Diego, CA, USA).

2.4.4 Sequence data analysis

The sequencing adapters and low-quality ends of raw sequenced reads were trimmed using Trimmomatic (version-0.36) (Bolger, Lohse and Usadel, 2014) with the options “ILLUMINACLIP:\${adapter.fa}:2:30:10 LEADING:30 TRAILING:30 SLIDINGWINDOW:4:25 MINLEN:1”, and the low-quality reads were filtered using FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html) with the options “-q 30 and -p 80”. Next, the remaining paired reads were mapped to the mouse reference genome (GRCm38-release71.fa) using the BWA-MEM algorithm (Li and Durbin, 2009) (Burrows-Wheeler Aligner; BWA; version-0.7.15-r1140). The reads with a mapping quality under 25 were filtered, and the others were treated according to the Genome Analysis Toolkit (GATK) best practices (de-duplicates, base recalibration, and local realignment) using GATK (McKenna *et al.*, 2010) (version 3.6), picard tools (<http://broadinstitute.github.io/picard>; version 2.6.0), and samtools (Li *et al.*, 2009) (version 1.3.1). The SNVs were detected by MuTect2 (Cibulskis *et al.*, 2013) (within GATK 3.6; default settings), Strelka (Saunders *et al.*, 2012) (version 1.0.14; skip depth filters to follow the recommendation), and LoFreq (Wilm *et al.*, 2012) (version 2.1.2; default settings with 0.01 significance level and Bonferroni correction). Among these unions of detected SNVs, the SNVs corresponding to any of the following cases were excluded: (a) registered in the dbSNP (Sherry, 2001) (release version 3), (b) supported by <3 reads, (c) reads supporting mutant alleles mapped in the control sample, and (d) other SNVs detected within 10 bp on either side. I defined candidate mutation positions as the positions detected to be SNVs in at least one tumor sample, which clearly passed above the filters, and I focused on 83 candidate mutation positions whose sequence coverage was $>50\times$ in all 7 samples. Next, I calculated the percentage of variant alleles among the mapped short reads for each candidate mutation position from the mapped read bases whose Phred quality score was >13 . The VAFs of the positions for which the number of reads supporting the mutation was less than 3 were filtered out (set to 0). Moreover, I estimated the tumor purity of the tumor samples according to the VAF of the Q61L *Hras* mutation (chr7:141192550), because this heterozygous mutation is known as an initiator of DMBA-TPA experimental carcinogenesis (Brown, Buchmann and Balmain, 1990; McCreery *et al.*, 2015; Nassar *et al.*, 2015). Accordingly, I estimated the proper penetrance rates of the mutations in the tumors while excluding the influence of the normal cell-derived reads by normalizing the raw VAFs with the estimated tumor purities. The normalized VAFs were used for the subsequent analyses. The sequencing, read mapping, and tumor purities are summarized in Tables 2.1–2.3.

2.4.5 Hierarchical clustering of candidate mutation positions

I assigned a 6-dimensional vector constructed by the normalized VAFs of 6 tumor samples for each of the 83 candidate mutation positions, and performed agglomerative hierarchical clustering. I adopted a weighted Manhattan distance (Canberra distance) approach as the distance measure between the assigned numeric vectors and used the group mean method for clustering. The Canberra distance between numeric vectors $\mathbf{x}$ and $\mathbf{y}$ can be calculated as follows:

$$d(\mathbf{x}, \mathbf{y}) = \sum_i \frac{|x_i - y_i|}{|x_i| + |y_i|} \quad (2.1).$$

2.4.6 Similarity measurements between tumors

To evaluate the similarity between tumors, I applied the Jaccard similarity coefficient and generalized Jaccard similarity coefficient. The Jaccard similarity coefficient is a similarity measure representing the degree of commonality between two sets. The Jaccard similarity coefficient between sets A and B can be calculated as follows:

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \quad (2.2).$$

In my study, I defined a set of mutations for each tumor sample and calculated the Jaccard similarity coefficient between the sets of mutations to represent the degree of similarity between tumors (Makohon-Moore *et al.*, 2017).

In addition, I defined an 83-dimensional numeric vector constructed by the VAFs of the 83 candidate mutation positions, and assigned a numeric vector for each tumor sample. I calculated the generalized Jaccard similarity coefficient between tumor samples using the defined numeric vectors as the similarity between the tumors. The generalized Jaccard similarity coefficient between numeric vectors $\mathbf{x}$ and $\mathbf{y}$ can be calculated as follows:

$$gJ(\mathbf{x}, \mathbf{y}) = \frac{\sum_i \min(x_i, y_i)}{\sum_i \max(x_i, y_i)} \quad (2.3).$$

2.4.7 Evaluation of intra-tumor heterogeneity

To evaluate the intra-tumor heterogeneity, I applied an entropy parameter. I regarded the intra-tumor heterogeneity as the complexity of the tumor, which was evaluated by the sum of the complexity provided by each mutation. From the spectrum of the normalized VAFs (Fig. 3), I assumed that there was no CNV at the 83 candidate mutation positions and that these

mutations occurred on one side of the diploids. With this assumption, I defined the probability that a cell randomly selected from a tumor i has a mutation m as $P(2 \times \text{nVAF}_{m,i})$, and defined its self-information amount as $-\log P(2 \times \text{nVAF}_{m,i})$, where $\text{nVAF}_{m,i}$ denotes a normalized VAF of a mutation m in a tumor T_i . Moreover, the clonal structure of a tumor can be represented by a combination of mutations; thus, I assumed that the mutations are independent but do not contradict each other. Under these assumptions, I defined an entropy parameter as the degree of the complexity of a tumor according to the following equation:

$$H(T_i) = - \sum_{m \in M_i} \log P(2 \times \text{nVAF}_{m,i}) \quad (2.4),$$

where M_i denotes a set of candidate mutations in a tumor T_i .

2.4.8 Estimation of the mutation rate and evolutionary process

According to a previously reported method (Williams *et al.*, 2016), the mutation rate and evolutionary process of the tumors were estimated from the normalized VAFs. On the assumption that the sub-clonal mutations were neutrally accumulated and penetrated a tumor over time, the number of mutations should proportionally increase with time, and thus the mutation rate of a given tumor should be constant. Therefore, Williams *et al.* (Williams *et al.*, 2016) proposed a fitting model to estimate the mutation rate of a tumor and to infer the neutrality of the sub-clonal mutations in a tumor as follows:

$$M(f) = \mu_e \left(\frac{1}{f} - \frac{1}{f_{\max}} \right) \quad (2.5),$$

where f denotes the relative fraction of a mutation in a tumor, $M(f)$ denotes the cumulative number of mutations, and μ_e denotes the apparent mutation rate. The f value corresponds to the VAF of a mutation, and $f_{\max} = 0.5$ when assuming a diploid tumor. Under the same assumptions described above, the VAF of a mutation should increase for mutations that occur earlier; hence, the cumulative number of mutations $M(f)$ can be estimated by sorting the mutations according to the VAFs. Thus, I assumed that the position with a normalized VAF that was less than 25% in the 83 candidate mutation positions was a sub-clonal mutation. To exclude false positives, I eliminated the mutations whose VAFs were below 2% in this analysis. For each tumor sample, the sub-clonal mutations were sorted in ascending order by the inverse of the normalized VAFs, the normalized VAFs were plotted against these ranks, and regression analysis was performed based on equation (2.5). Finally, the slope of the regression line and the coefficient of determination were calculated, respectively.

2.4.9 Data availability

All sequencing data used in this work are available from the DNA Data Bank of Japan (DDBJ) Sequence Read Archive (DRA) under the accession number DRA007132.

2.5 Conclusion

In this chapter, I have described the results of prospective cancer genome analysis for two tumor series, with one ending as a metastasized tumor and the other as a regressed tumor. The mutational landscape shows that these tumor series had acquired and inherited different mutations compared to each other. These results suggested that it is important for malignant transformation to increase and maintain intra-tumor heterogeneity. Moreover, selection pressure in tumor evolution could affect during the formation of late-stage benign tumors; while on the other hand, the tumor fates might be determined at a considerably earlier stage.

Chapter 3

DEclust: A statistical approach for obtaining differential expression profiles of multiple conditions

In this chapter, I propose a novel analytical method for multi-conditional transcriptome data to extract gene clusters whose expression patterns are statistically overrepresented among multiple conditions. Afterwards, I show the results of cancer transcriptome analysis of applying the proposing method to the tumors obtained by the experimental carcinogenesis (Aoto *et al.*, 2017).

High-throughput RNA sequencing technology is widely used to comprehensively detect and quantify cellular gene expression. Thus, numerous analytical methods have been proposed for identifying differentially expressed genes (DEGs) between paired samples such as tumor and control specimens, but few studies have reported methods for analyzing differential expression under multiple conditions. I propose a novel method, DEclust, for differential expression analysis among more than two matched samples from distinct tissues or conditions. As compared to conventional clustering methods, DEclust more accurately extracts statistically significant gene clusters from multi-conditional transcriptome data, particularly when replicates of quantitative experiments are available.

As a result of applying my method to the cancer transcriptome analysis, DEclust extracted 16 gene clusters whose expression patterns were discriminative in the process of the tumor progression; moreover, the genes which were belonging to each of these clusters were significantly enriched for biological functions.

DEclust can be used for any multi-conditional transcriptome data, as well as for extending any DEG detection tool for paired samples to multiple samples. Accordingly, DEclust can be used for a wide range of applications for transcriptome data analysis. DEclust is freely available at <http://www.dna.bio.keio.ac.jp/software/DEclust>.

3.1 Background

3.1.1 RNA sequencing and gene expression analysis

As mentioned in Chapter 1, high-throughput RNA sequencing (RNA-Seq) technology has been used to comprehensively detect and quantify cellular gene expression (Wang, Gerstein and Snyder, 2009). Most transcriptomic studies, such as those used in cancer research, are focused on genes whose expression levels differ in a statistically significant manner between experimental conditions, and numerous analytical methods for identifying differentially

expressed genes (DEGs) between a pair of matched samples—such as tumor and control specimens—based on RNA-Seq data. edgeR (Robinson, McCarthy and Smyth, 2010), DESeq2 (Love, Huber and Anders, 2014), and cuffdiff2 (Trapnell *et al.*, 2013) are widely used for comprehensively quantifying gene expression levels and detecting DEGs between a pair of samples or a pair of experimental conditions by using statistical tests; however, to date, few studies have proposed analytical methods for detecting differential expression among multiple conditions.

3.1.2 Cluster analysis

Cluster analysis according to gene expression patterns is a simple analysis that is performed on multi-conditional transcriptome data for predicting gene functions, searching for biomarkers, and summarizing large datasets; in cluster analysis, sets of objects are grouped in such a manner that objects in the same group (“cluster”) are similar to each other but are different from objects in other groups. For clustering based on gene-expression patterns, Euclidean distance, Pearson’s correlation, and cosine distance are commonly used for measures of distance or similarity between genes (D’haeseleer, 2005). Although the conventional definitions of distance appear to be useful (D’haeseleer, 2005), they do not consider any statistical significance based on a dispersion of gene expression: Experimentally quantified gene expression levels possess dispersion due to stochastic biological noise, experimental error, and other factors (Elowitz *et al.*, 2002; Trapnell *et al.*, 2013), but the conventional definitions of distance consider only the means of expression levels among replicates. Therefore, the existing clustering methods cannot distinguish significant differences in gene expression levels from stochastic biological noise, and, consequently, these methods cannot accurately generate statistically overrepresented patterns of gene expression profiles among multiple conditions from the expression profiles of the genes whose expression levels are significantly changed among the experimental conditions.

To overcome the aforementioned drawbacks associated with analysis methods, I propose here a novel approach for differential expression analysis. My method, termed DEclust, provides a scoring system for a hierarchical clustering based on differential expression profiles among multiple conditions. I define a collection of statistical test results obtained for all combinatorial pairs of conditions as a pairwise *differential expression test (DET) profile* and assign it to each gene. DEclust searches for a set of statistically overrepresented patterns on DET profiles among multiple conditions by clustering the pairwise DET profiles.

3.2 Results and Discussion

3.2.1 Algorithm

For applying a statistical approach to multi-conditional transcriptome data, DEClust adopts a novel distance score between a pair of clusters of genes based on their pairwise DET profiles. The pairwise DET profile is an integer vector constructed from statistical test results for all pairs of conditions. In accordance with common analytical pipelines, RNA-Seq reads are mapped to a reference genome, and an expression level for each gene per condition is estimated. Based on these expression values, a statistical test is applied to all pairs of conditions by using an existing DEG detection tool. The result of the statistical test of differential expression for a pair of conditions is labeled as +1 (upregulated pair), -1 (downregulated pair) or 0 (insignificantly altered pair). Subsequently, these integer values for all pairs of conditions are formed in the vector as a pairwise DET profile. Therefore, the pairwise DET profile features $_NC_2$ dimensions (N is the number of conditions). The expression level of a gene g in condition k is denoted as e_k^g . If the expression level, e_k^g , is statistically different from e_l^g , then an element of the corresponding dimension in the pairwise DET profile is defined to be 1 ($e_k^g < e_l^g$) or -1 ($e_k^g > e_l^g$); otherwise, it is 0 (Fig. 3.1a). The pairwise DET profile assigned to a gene g is defined as follows:

$$v(g) = (d_{1,2}^g, d_{1,3}^g, \dots, d_{N-1,N}^g) \quad (3.1),$$
$$d_{k,l}^g = \begin{cases} +1 & (e_k^g < e_l^g) \\ 0 & (\text{no - significance}), \\ -1 & (e_k^g > e_l^g) \end{cases}$$

where $d_{k,l}^g$ represents the statistical test result for the comparison between a pair of certain conditions k and l .

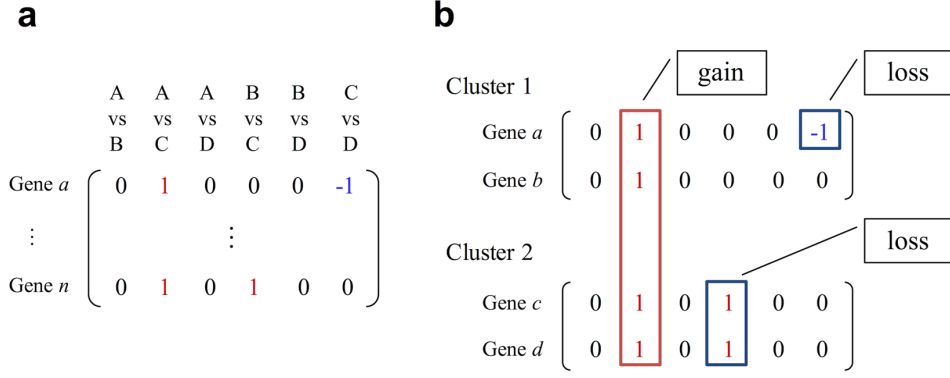


Figure 3.1. Pairwise DET profile and definition of gain and loss. **(a)** An example of a pairwise DET profile for genes. All pairwise combinations of four conditions (A, B, C, and D) constitute the six-dimensional pairwise DET profile. **(b)** An example of the calculation of gains and losses. The pairwise DET profile of Cluster 1 is (0 1 0 0 0 0), and the pairwise DET profile of Cluster 2 is (0 1 0 1 0 0); thus, the pairwise DET profile $v(C_1 \cup C_2)$ is (0 1 0 0 0 0). Accordingly, the number of elements that are 1 or -1 in $v(C_1 \cup C_2)$ is 1 ($s = 1$), and that in $|C_1 \cup C_2|$ is 4; therefore, $\text{gain}_{1,2} = 4$. The total number of elements that are 1 or -1 in $v\{\forall g \in C_n \cup C_m\}$ is 7 ($t = 7$). Hence, $\text{loss}_{1,2} = 3$ and $D(C_1, C_2) = 0.857$.

For inducing statistically overrepresented patterns on DET profiles among multiple conditions, inter-cluster distance is defined to be shorter when a large number of the same statistical test results are shared between the given clusters of genes. DEclust calculates the rate of the statistical test results shared among pairwise DET profiles of genes in the clusters to be merged. First, the definition of the pairwise DET profile is extended to a set (cluster) of genes as follows. An element that corresponds to a statistical test result between a pair of conditions k and l in the pairwise DET profile for a cluster C_n of genes is defined to be 1 if $d_{k,l}^g = 1$ for $\forall g \in C_n$ and -1 if $d_{k,l}^g = -1$ for $\forall g \in C_n$; otherwise, it is 0. Thus, the pairwise DET profile assigned for a cluster C_n is defined as follows:

$$v(C_n) = (d_{1,2}^{C_n}, d_{1,3}^{C_n}, \dots, d_{|C|-1,|C|}^{C_n}) \quad (3.2)$$

Second, the distance between two clusters C_n and C_m based on the pairwise DET profile is defined as follows:

$$D(C_n, C_m) = 1 - \frac{(\text{gain}_{n,m} - \text{loss}_{n,m})}{(\text{gain}_{n,m} + \text{loss}_{n,m})} \quad (3.3),$$

where $\text{gain}_{n,m}$ is defined as $s \times |C_n \cup C_m|$, for s denoting the number of elements of +1 or -1 in the pairwise DET profile $v(C_n \cup C_m)$ for the union of the clusters C_n and C_m ; and $\text{loss}_{n,m}$ is defined to be $(t - \text{gain}_{n,m})$, for t denoting the total number of elements of +1 or -1 in the pairwise DET profiles $v(g)$ for all genes g in the union of the clusters C_n and C_m .

Fig. 3.1b shows an example of the gain and loss; the gain and loss are non-negative integer values and if $\text{gain}_{n,m} = 0$ and $\text{loss}_{n,m} = 0$, then $D(C_n, C_m) = 0$. If all pairwise DET profiles of genes in a union of clusters are identical, $\text{gain}_{n,m} = t$ and $\text{loss}_{n,m} = 0$, and thus $D(C_n, C_m) = 0$. By contrast, if no element has the same value of +1 or -1 among all pairwise DET profiles in a union of clusters, $\text{gain}_{n,m} = 0$ and $\text{loss}_{n,m} = t$, and therefore $D(C_n, C_m) = 2$. Thus, $D(C_n, C_m)$ corresponds to an indicator of the shared (or unshared) statistical test result rate in the union of the clusters C_n and C_m . If any pairs of clusters have the same distance under the definition of equation (3.3), DEclust uses a conventional distance measure to search for the closest pair of clusters.

By using the aforementioned definition of distance, DEclust adopts the agglomerative-hierarchical-type clustering method to identify statistically overrepresented patterns on DET profiles among multiple conditions. DEclust initially merges genes that feature the same statistical test results (i.e., the same pairwise DET profile), and subsequently merges two gene clusters C_n and C_m with minimum distance $D(C_n, C_m)$ so as to conserve the common statistical test results as much as possible. In this manner, DEclust can not only be applied to multi-conditional transcriptome data but also be used to extract statistically significant gene clusters (Algorithm 3.1).

Algorithm 3.1. Pseudo code of DEclust.

Input: $G = \{g_1, g_2, \dots, g_n\}$ (set of genes to be clustered).
Output: N (Hierarchical tree written in Newick tree format).
Procedure:
Assign one gene to each cluster.
Calculate the distance D according to equation (3.3) for all pairs of clusters.
while there is more than one cluster **do**
 Search for a pair of clusters with the minimum distance.
 If Any pairs of clusters have the same distance **then**
 Compare conventional distance of expression levels.
 end if
 Merge the clusters.
 Store the clusters and the distance D in Newick format to N .
 Calculate the distance D using the equation (3.3) between a new cluster and other clusters.
end while
return N

3.2.2 Benchmark study of clustering

For demonstrating the superiority of my method over other methods, I evaluated the accuracy of DEclust and of hierarchical clustering performed with conventional distance measures (hereafter called *existing methods*) by using simulated datasets based on assuming RNA-Seq analysis. In this benchmark study, I used 12 distance measures as the existing methods: the combinations of 4 inter-cluster distances (the group average method, single-linkage method, complete-linkage method, and Ward's method (Ward, 1963)) and 3 inter-gene distances (Euclidean distance, Pearson's correlation, and cosine distance). In the text that follows, I denote a DEG as a gene whose expression levels differ in a statistically significant manner between at least one pair of conditions. In the benchmark study, I assumed four artificial conditions, and for evaluation, I established ten pairwise DET profiles as class labels of DEGs (Fig. 3.2) and one pairwise DET profile for non-DEGs (and thus prepared 11 correct class labels in total). Moreover, 20% of the genes were randomly selected and a class label for DEGs was randomly assigned for each selected gene. Next, I designed the probabilistic distributions as controls for each gene to generate simulated datasets of RNA-Seq read counts. The RNA-Seq read count data of 69 lymphoblastoid cell lines (LCLs) derived from unrelated Nigerian individuals (Pickrell *et al.*, 2010) was used as the reference. The read counts for each condition were sampled from the control distribution as many times as the number of

replicates, and these were treated as the biological replicates for each condition. Further, the read counts for differentially expressed conditions were relative to a certain fold-change in concordance with the class label (i.e., with the DET profile). According to the read counts, I generated simulated short reads to obtain gene expression profiles for four artificial conditions with a certain number of replicates (entire description about simulation datasets is presented in section 3.3.1). The simulated reads were mapped to a reference genome, and expression levels were estimated for each gene per condition. The statistical tests were applied to all pairs of conditions to detect significant differences in gene expression. I calculated the pairwise DET profile from the results of the statistical tests and assigned it to each gene. Lastly, I implemented DEClust and the existing methods for all of the genes expressed in one or more conditions. The estimation of gene expression levels and the statistical tests were performed using edgeR (Robinson, McCarthy and Smyth, 2010), DESeq (Anders and Huber, 2010), DESeq2 (Love, Huber and Anders, 2014), and cuffdiff2 (Trapnell *et al.*, 2013). The inputs of DEClust were the pairwise DET profiles and estimated normalized expression levels, and the inputs of the existing methods were the estimated normalized expression levels. In this manner, I performed the simulation thrice for each parameter set. As a result, I generated three simulation datasets in which the genes belonging to the cluster with the same class label and the expression profiles of each gene were different among simulations for each parameter set.

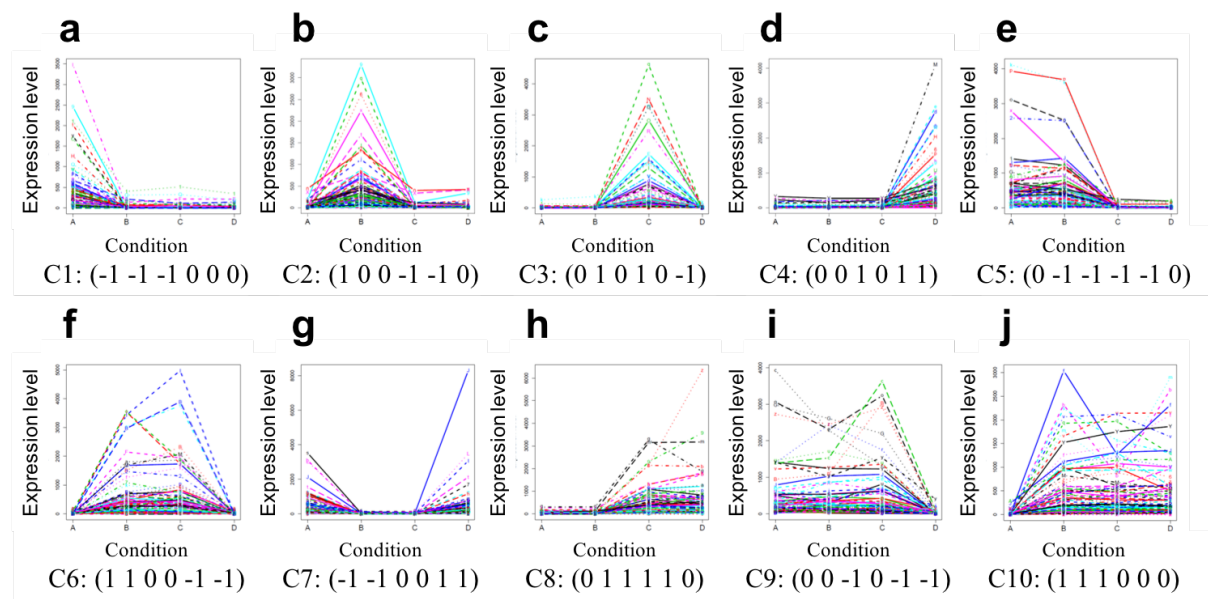


Figure 3.2. Ten correct labels of benchmark study. Line plots with corresponding pairwise DET profiles. The handling of each dimension is same as Fig. 3.1a. Four of the correct labels represent genes that are overexpressed in any one condition (a-d), while four other correct labels represent genes that are overexpressed in any two conditions (e-h). The rests of the labels represent genes that are overexpressed in any three conditions (i, j).

For all pairs of genes featuring identical class labels, I assessed whether they belonged to the same resultant cluster after clustering. I converted the clustering problem into a binary classification problem, and as a measure of the evaluation, I used the area under the curve (AUC) of the receiver operating characteristic (ROC). I calculated the AUCs for each parameter set of inter-gene distances, inter-cluster distances, and number of replicates, and compared the AUCs between DEclust and the existing methods (Fig. 3.3). Not only did DEclust show higher accuracy than the existing methods for almost all parameter sets, the AUCs of DEclust were also improved with an increase in the number of replicates. Specifically, DEclust showed superior accuracy in all the datasets containing two or more replicates when DESeq2 was used for statistical tests. Moreover, the AUCs of DEclust were higher than those of the existing methods even when edgeR, DESeq or cuffdiff2 was used with sufficient replicates. Thus, DEclust succeeded in identifying the correct class labels of differentially expressed patterns from multi-conditional transcriptome data more accurately than the existing clustering methods. Conversely, the existing methods showed lower accuracy than DEclust in my benchmark dataset. The existing methods depend solely on the means of the expression levels; therefore, the existing methods could not distinguish significant differences in expression levels from stochastic noise. Consequently, the existing methods could not extract statistically overrepresented and differentially expressed gene clusters. In either method, the variance of AUCs among the three simulation datasets for each parameter set was not so large.

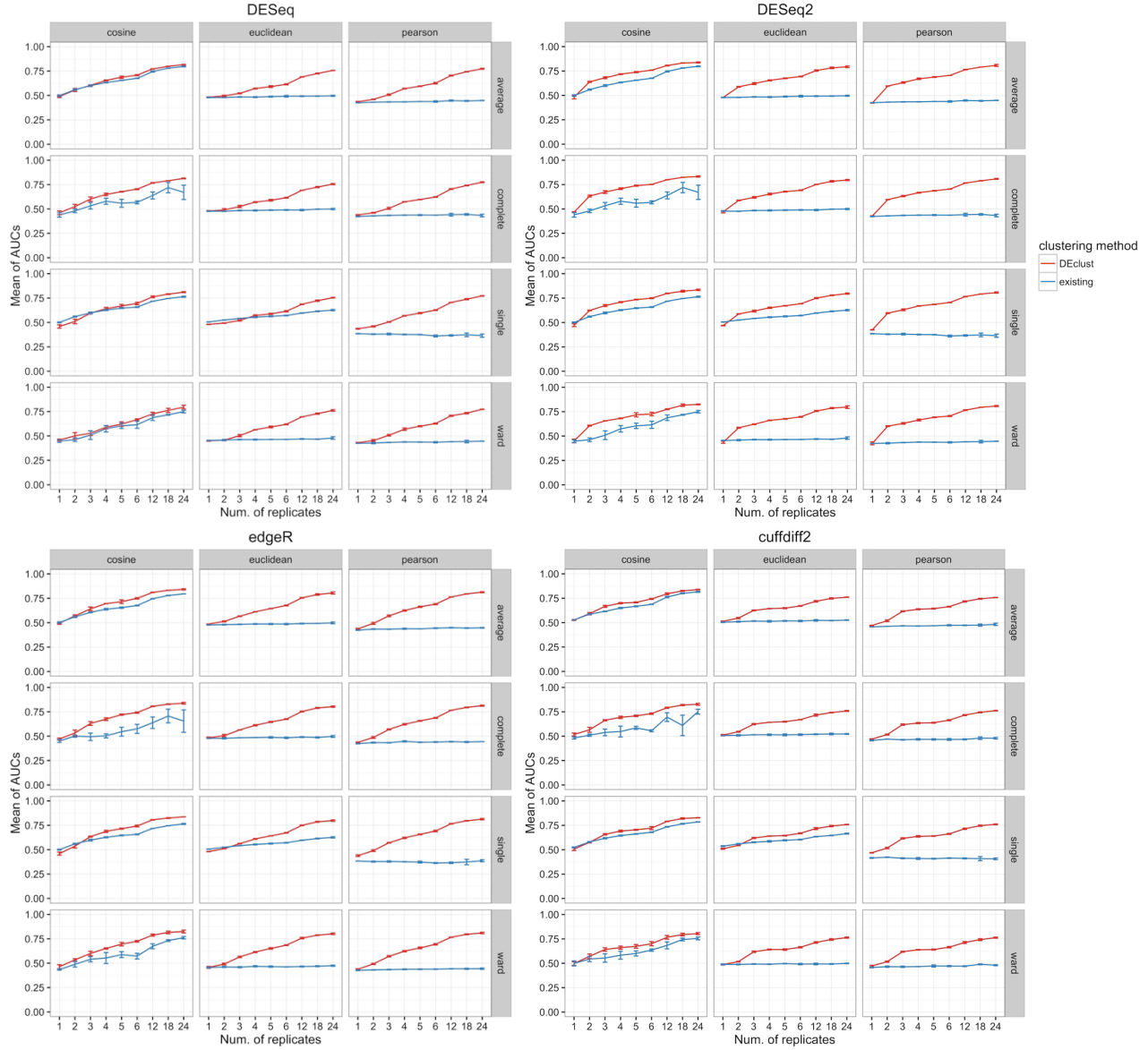


Figure 3.3. Results of clustering evaluation. *DEclust* is my method and *existing* is conventional hierarchical clustering method. For the quantification of the gene expression and statistical test to identify the differential expression, DESeq, DESeq2, edgeR, and cuffdiff2 were used. For existing methods and secondary distance measure of DEclust, the group average, complete-linkage, single-linkage, and Ward’s method (Ward, 1963) were used for the inter-cluster distance measure, and the cosine distance, Euclidean distance, and Pearson’s correlation were used for the inter-gene distance measure.

To summarize this benchmark study, when any biological replicates were available, the highest accuracy was achieved by DEclust with the use of the statistical test results of

DESeq2, and when the number of biological replicates was sufficient, DEclust with edgeR was most accurate. Furthermore, if optimal statistical test results are obtained, $AUC = 1.00$ with my method.

3.2.3 Benchmark study of DEG detection

Although DEclust uses the results of statistical tests for all combinatorial pairs of conditions, multiDE (Kang, Du and Zhang, 2016) has been proposed to detect DEGs among multiple conditions. In addition, edgeR and DESeq2 have an extended function, the likelihood ratio test, to detect DEGs among multiple conditions. Therefore, I evaluated the power of DEG detection using multiDE, edgeR, DESeq2, and DEclust. For evaluation of the DEG detection capability of DEclust, I defined a “DEG cluster” and “non-DEG cluster” as follows. If there was at least one element of a pairwise DET profile of a cluster with a value of 1 or -1 , the cluster was a DEG cluster; otherwise (that is, all elements were 0), it was a non-DEG cluster. In this manner, DEclust can be applied to the classification problem of whether a gene is a DEG or not. I used the same simulation datasets that were used for the benchmark study to evaluate the performance of DEclust and the conventional clustering methods (the details for the evaluation method are described in the Materials and Methods).

The performance of edgeR, DESeq2, multiDE, and DEclust was evaluated according to the true-positive rate (TPR), positive predictive value (PPV), accuracy, and F-measure (harmonic mean of TPR and PPV). DEclust uses the pairwise based statistical test results obtained from edgeR, DESeq, DESeq2, or cuffdiff2; the evaluation results using each tool are shown separately (Fig. 3.4a). Overall, DEclust, based on statistical testing between all pairs of conditions, had higher TPR, and edgeR, DESeq2, and multiDE, based on the statistical testing for multiple conditions, had higher PPV. Focusing on the F-measure (the harmonic mean of TPR and PPV), DEclust with DESeq2, DEclust with edgeR, sole use of DESeq2, and sole use of edgeR were advantageous relative to the other methods. Moreover, not only was the DEG detection power of DEclust with DESeq2 and sole use of DESeq2 sufficiently close, but also DEclust with DESeq2 was advantageous when only a few biological replicates were available (Fig. 3.4b).

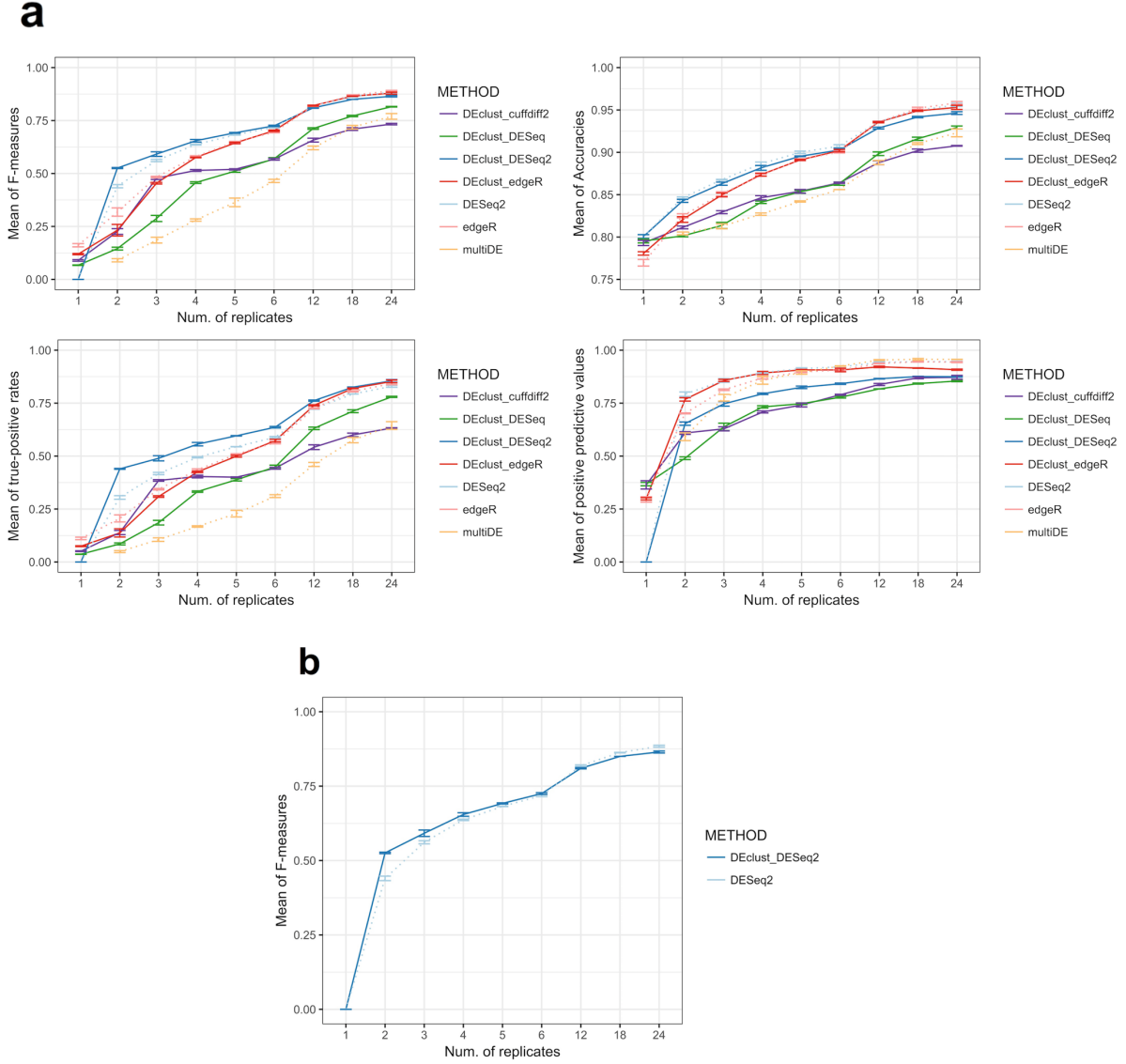


Figure 3.4. Results of DEG detection evaluation. The F-measures, accuracies, TPRs, and PPVs for each method for each number of replicates are plotted. The vertical axis shows the mean values of corresponding indicator and the error bars are drawn in accordance with the corrected sample standard deviation of three simulations for each parameter set. *DEclust*, my method, uses the statistical test results obtained from edgeR, DESeq, DESeq2, or cuffdiff2, and the evaluation results using each of these tools are separately shown as “DEclust_[DEGs detection tool]”. For the secondary distance of DEclust, the group average method with cosine distance was used. The complete results for edgeR, DESeq2, multiDE, and DEclust are shown in (a), and only F-measure of DESeq2 and DEclust with DESeq2 are shown in (b).

Although the single statistical test for multiple conditions using edgeR, DESeq2, and multiDE can more accurately identify DEGs, statistical tests for all combinatorial pairs of

conditions can identify DEGs with higher sensitivity and determine which pairs of conditions are significantly different. In the present study, the DEG detection ability of DEclust, which was in accordance with the statistical test results for all combinatorial pairs of conditions, was comparable with that of tools that detect DEGs among multiple conditions using a single test. According to the definition of the pairwise DET profile assigned for each cluster (equation (3.2)) and the distance between clusters (equation (3.3)), the distance between DEG clusters and non-DEG clusters should be 2 (maximum value) because none of the significant statistical test results are shared; therefore, DEclust can accurately cluster the DEGs and non-DEGs if quitting the merge step of hierarchical clustering when the distance between any pair of clusters n, m satisfies $D(C_n, C_m) \geq 1$. Thus, DEclust has high compatibility for multi-conditional transcriptome analysis.

3.2.4 Application to biological data

To demonstrate the practical usefulness of DEclust, I performed mRNA-Seq analysis on mouse tumor samples and applied DEclust to extract statistically overrepresented and significant gene clusters among multiple stages in the process of cancer progression. I and collaborators obtained normal skin, papilloma, carcinoma, and metastatic tumors from each of the same two mice (i.e., two replicates were obtained for each stage), and performed mRNA-Seq on these tumor samples. I obtained an average of 40.10 M reads by performing mRNA-Seq with an Illumina Genome Analyzer IIX, and I mapped the reads to a mouse reference genome (Table 3.1). Subsequently, I identified the genes that were differentially expressed between a pair of the four tumorigenesis stages by using DESeq2. In total, 6,584 DEGs were detected (Table 3.2).

Next, I applied DEclust to the detected 6,584 DEGs between a pair of stages using the results of DESeq2 to extract statistically overrepresented gene clusters among the four tumorigenesis stages. The hierarchical clustering tree of DEGs was calculated, and the DEGs were divided into 16 clusters by quitting the merge step if the distance between any pair of clusters n, m satisfied $D(C_n, C_m) \geq 1$ (Fig. 3.5a). Fig. 3.5b shows the 16 clusters together with the corresponding expression patterns and pairwise DET profiles. According to the pairwise DET profiles of the 16 clusters, the genes that belong to one of 16 clusters shared at least one statistical test result of a pair of stages. Thus, DEclust can extract statistically significant gene clusters from multi-conditional transcriptome data. Furthermore, I performed a gene set enrichment analysis (GSEA) for each cluster (the top three hallmarks are shown in Table 3.3). The genes of each cluster were annotated with hallmarks such as the *KRAS* signaling pathway, mammalian target of rapamycin (*mTOR*) signaling pathway, and checkpoints of the cell cycle, which are typical cancer-related pathways (Vermeulen, Van Bockstaele and Berneman, 2003; Vogelstein *et al.*, 2013).

Table 3.1. Statistical summary of mouse mRNA-Seq data.

Samples	Normal skin		Papillomas		Carcinomas		Metastasis	
	Replicate 1	Replicate 2	Replicate 1	Replicate 2	Replicate 1	Replicate 2	Replicate 1	Replicate 2
# reads	40,133,148	41,321,764	39,146,671	40,790,238	40,500,492	41,275,570	37,819,465	39,833,125
# passed	35,446,518	37,002,801	35,360,696	36,024,516	35,651,129	36,937,481	34,235,323	36,688,396
% passed	88.32	89.55	90.33	88.32	88.27	89.49	90.52	92.11
# mapped	29,260,581	31,015,376	28,866,123	27,802,980	29,631,879	31,901,997	28,255,455	31,387,689
% mapped	82.54	83.82	81.63	77.18	82.88	86.37	82.53	85.55

An average of 40.10 M reads was obtained from mRNA-Seq with Illumina Genome Analyzer IIX. An average of 89.61% of reads passed a quality filter, and 82.81% of those reads were uniquely mapped to mouse reference genome.

Table 3.2. Result of DEG detection for the mouse mRNA-Seq data.

		Up-regulated			
		Normal skin	Papillomas	Carcinomas	Metastasis
Down-regulated	Normal skin		1,464	1,535	1,405
	Papillomas	1,668		886	1,428
	Carcinomas	1,596	798		766
	Metastasis	1,358	1139	579	

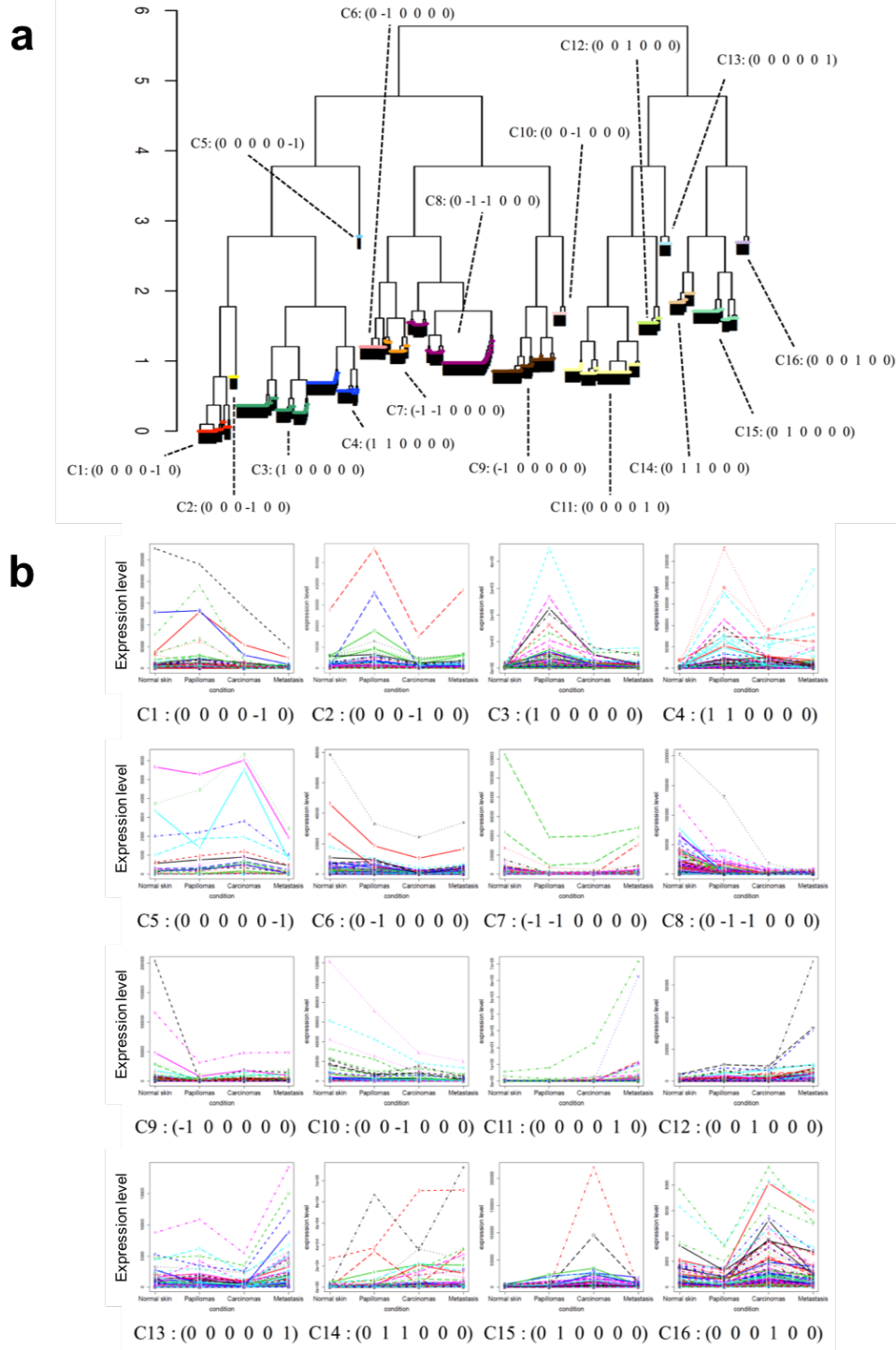


Figure 3.5. Cancer transcriptome analysis using DEClust. **(a)** Hierarchical tree of clustering result from DEClust for the transcriptome data. The vertical axis indicates inter-cluster distance (equation (3.3)). The DEGs are divided into 16 clusters (which are color-coded). The cluster numbers correspond to the numbers in Table 3.3, and each cluster is assigned a pairwise DET profile of six dimensions. **(b)** Line plots of the expression patterns for each gene in each.

Table 3.3. Result of GSEA for each cluster obtained by using DEclust.

Cluster	# genes	6 dim pattern						Note	GSEA Hallmarks
		NP	NC	NM	PC	PM	CM		
1	373	(0	0	0	0	-1	0)	P > M	MTORC1_SIGNALING, E2F_TARGETS, MYC_TARGETS_V1
2	96	(0	0	0	-1	0	0)	P > C	ADIPOGENESIS, MITOTIC_SPINDLE, ESTROGEN_RESPONSE_EARLY
3	117	(0	0	0	0	0	1)	C < M	HEME_METABOLISM, XENOBIOTIC_METABOLISM
4	727	(1	0	0	0	0	0)	N < P	E2F_TARGETS, MTORC1_SIGNALING, MYC_TARGETS_V1
5	345	(1	0	0	0	-1	0)	N < P P > M	E2F_TARGETS, MYC_TARGETS_V1, MTORC1_SIGNALING
6	965	(0	-1	-1	0	0	0)	N > C N > M	MYOGENESIS, HYPOXIA, ESTROGEN_RESPONSE_LATE
7	328	(0	-1	0	0	0	0)	N > C	ADIPOGENESIS, COMPLEMENT, PI3K_AKT_MTOR_SIGNALING
8	721	(-1	0	0	0	0	0)	N > P	MYOGENESIS, INTERFERON_GAMMA_RESPONSE, EPITHELIAL_MESENCHYMAL_TRANSITION
9	143	(0	0	-1	0	0	0)	N > M	G2M_CHECKPOINT, MYOGENESIS, HEME_METABOLISM
10	490	(0	0	1	0	1	0)	N < M P < M	COMPLEMENT, COAGULATION, ESTROGEN_RESPONSE_EARLY
11	742	(0	0	0	0	1	0)	P < M	BILE_ACID_METABOLISM, UV_RESPONSE_DN, INTERFERON_GAMMA_RESPONSE
12	238	(0	0	1	0	0	0)	N < M	XENOBIOTIC_METABOLISM, ADIPOGENESIS, REACTIVE_OXIGEN_SPECIES_PATHWAY
13	140	(0	0	0	1	0	0)	P < C	INFLAMMATORY_RESPONSE, COMPLEMENT, IL6_JAK_STAT3_SIGNALING
14	46	(0	0	0	0	0	-1)	C > M	EPITHELIAL_MESENCHYMAL_TRANSITION
15	577	(0	1	1	0	0	0)	N < C N < M	EPITHELIAL_MESENCHYMAL_TRANSITION, KRAS_SIGNALING_UP, INFLAMMATORY_RESPONSE
16	536	(0	1	0	0	0	0)	N < C	EPITHELIAL_MESENCHYMAL_TRANSITION, KRAS_SIGNALING_UP, INFLAMMATORY_RESPONSE

In this table, the ‘N’ means normal skin, ‘P’ means papilloma, ‘C’ means carcinoma, and ‘M’ means metastatic tumor. The six-dimensional patterns are corresponding to the pairwise DET profiles for each cluster. The column “Note” represents an interpretation of the pairwise DET profile, and the column “GSEA Hallmarks” shows the top three hallmarks for each cluster.

To compare DEclust and an existing method with the use of the real dataset, I performed hierarchical clustering using the group average method with the cosine distance for the 6,584 DEGs and plotted the hierarchical tree (Fig. 3.6a). The hierarchical tree was cut into 16 clusters to obtain the same number of clusters as in the analysis performed with DEclust. As a result of the assignment of a pairwise DET profile to each cluster, all of the elements in the assigned pairwise DET profiles were zero (Fig. 3.6b). This means that when the existing method was used, the genes showing significant differences and the genes showing non-significant differences in expression levels at the same pair of stages were mixed in the same cluster. Notably, the existing method could not distinguish significant expression differences from noise. Moreover, I performed the GSEA for each cluster generated by the existing method and obtained the top three hallmarks for each cluster (Table 3.4). A comparison of the results obtained by DEclust and the existing method revealed that any cluster obtained by DEclust was annotated but several clusters obtained by the existing method could not be annotated because the genes belonging to these clusters were not significantly enriched for biological functions. In addition, interleukin-6/Janus kinase/signal transducer and activator of transcription-3 (*IL6/JAK/STAT3*) and phosphatidylinositol 3-kinase/protein kinase B/mammalian target of rapamycin (*PI3K/AKT/mTOR*) signaling pathways were detected only by DEclust. These pathways not only are frequently activated in cancer cells, but are also involved in cancer progression by promoting tumor cell growth, angiogenesis, and metastasis (Porta, Paglino and Mosca, 2014; Yu *et al.*, 2014). Thus, DEclust can extract biologically significant gene clusters.

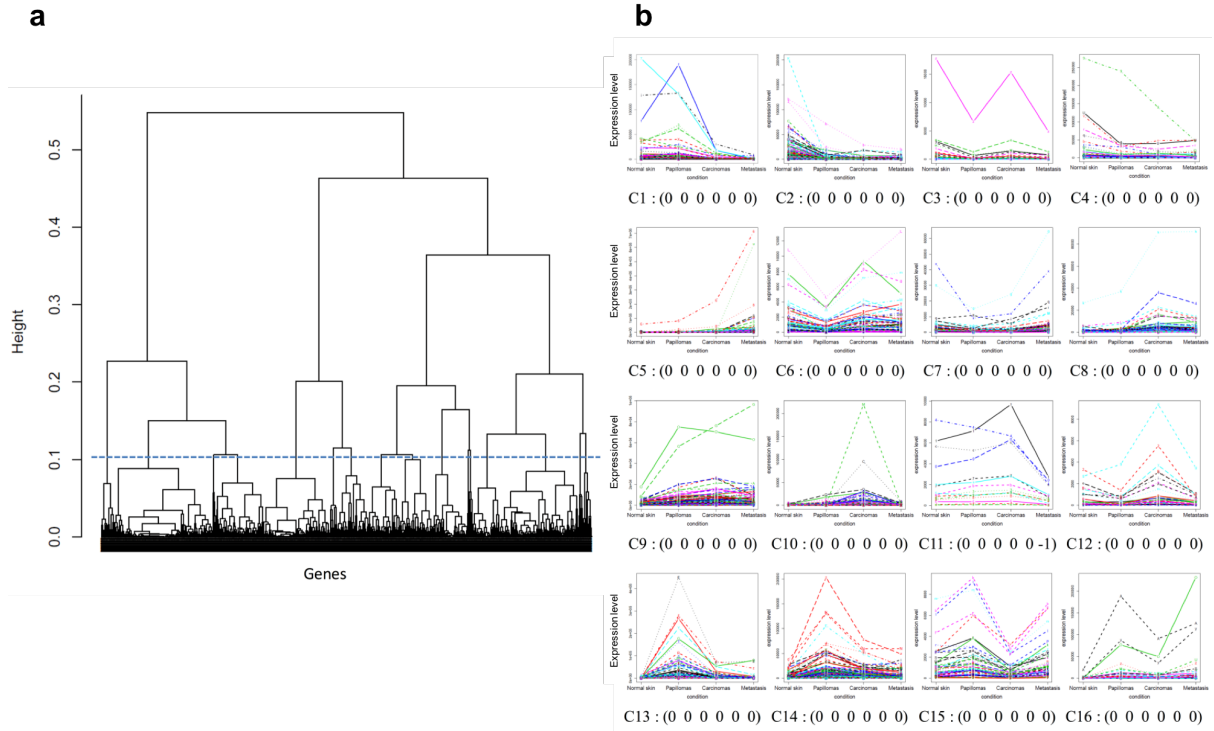


Figure 3.6. Result of hierarchical clustering by using existing clustering method. **(a)** Hierarchical tree generated by the existing clustering method with the group average method for inter-cluster distance measure and the cosine distance for inter-gene distance measure. The clusters were divided at the highlighted in blue dotted lines. The cluster numbers are assigned to left to right cluster. **(b)** Line plots of the expression patterns for each gene in each cluster. The annotations in the bottom of figures are a cluster number associated with S11 Table and their pairwise DET profiles.

Table 3.4. Result of GSEA for each cluster obtained by using existing method.

Cluster	# genes	6 dim pattern						Note	GSEA Hallmarks
		NP	NC	NM	PC	PM	CM		
1	1250	(0	0	0	0	0	0)	N.A.	MYOGENESIS, ESTROGEN_RESPONSE_EARLY, ESTROGEN_RESPONSE_LATE
2	34	(0	0	0	0	0	0)	N.A.	--
3	95	(0	0	0	0	0	0)	N.A.	UV_RESPONSE_DN, INFLAMMATORY_RESPONSE, EPITHELIAL_MESENCHYMAL_TRANSITION
4	125	(0	0	0	0	0	0)	N.A.	FATTY_ACID_METABOLISM, ADIPOGENESIS KRAS_SIGNALING_DN
5	47	(0	0	0	0	0	0)	N.A.	--
6	675	(0	0	0	0	0	0)	N.A.	HEME_METABOLISM, UV_RESPONSE_DN, TNFA_SIGNALING_VIA_NFKB
7	634	(0	0	0	0	0	0)	N.A.	XENOBIOTIC_METABOLISM, ESTROGEN_RESPONSE_EARLY, COAGULATION
8	376	(0	0	0	0	0	0)	N.A.	KRAS_SIGNALING_UP, INTERFERON_ALPHA_RESPONSE, INTERFERON_GAMMA_RESPONSE
9	97	(0	0	0	0	0	0)	N.A.	TGF_BETA_SIGNALING, UV_RESPONSE_DN INFLAMMATORY_RESPONSE
10	346	(0	0	0	0	0	0)	N.A.	FATTY_ACID_METABOLISM, BILE_ACID_METABOLISM, INTERFERON_ALPHA_RESPONSE
11	508	(0	0	0	0	0	0)	N.A.	EPITHELIAL_MESENCHYMAL_TRANSITION INFLAMMATORY_RESPONSE, KRAS_SIGNALING_UP
12	534	(0	0	0	0	0	0)	N.A.	EPITHELIAL_MESENCHYMAL_TRANSITION KRAS_SIGNALING_UP, INFLAMMATORY_RESPONSE
13	157	(0	0	0	0	0	0)	N.A.	CHOLESTEROL_HOMEOSTASIS, APICAL_JUNCTION, XENOBIOTIC_METABOLISM
14	303	(0	0	0	0	0	0)	N.A.	XENOBIOTIC_METABOLISM TGF_BETA_SIGNALING, ADIPOGENESIS
15	315	(0	0	0	0	0	0)	N.A.	G2M_CHECKPOINT, E2F_TARGETS, MITOTIC_SPINDLE
16	1088	(0	0	0	0	0	0)	N.A.	E2F_TARGETS, MYC_TARGETS_V1, MTORC1_SIGNALING

3.3 Materials and Methods

3.3.1 Simulation datasets

To evaluate DEClust and conventional clustering methods, I generated simulated short reads from mouse reference cDNA sequences so that gene expression profiles for four artificial conditions were obtained. I set up ten pairwise DET profiles as correct labels for DEGs (Fig. 3.2). Four of the correct labels represent genes that are overexpressed in any one condition (clusters 1-4), while four other correct labels represent genes that are overexpressed in any two conditions (clusters 5-8). The rests of the correct labels represent genes that are overexpressed in any three conditions (clusters 9 and 10). Second, 20% of genes were randomly selected as DEGs and the correct label for DEGs was randomly assigned for each selected gene. Other genes were classified in cluster 0 whose pairwise DET profile is defined as (0 0 0 0 0 0), indicating that the expression levels of any pairs of conditions are not significantly different. Third, a control distribution that follows the negative binomial distribution was generated for each gene. The RNA-Seq read counts for each condition were sampled as many times as the number of replicates from the control distribution, and the read counts for differentially expressed conditions were biased by a randomly selected fold-change. The fold-change was selected from a pool of ($\times 1.5$, $\times 2.0$, $\times 2.5$, $\times 3.0$).

I generated the control distribution for each gene as follows. First, I downloaded a mouse reference genome, mouse reference cDNA, and mouse gene annotation file from Ensembl (<http://www.ensembl.org/index.html>, GRCm38-release71), and RNA-Seq read count data of 69 Nigerian lymphoblastoid cell lines (LCLs) from Pickrell et al. (Pickrell *et al.*, 2010) to yield practical datasets. I extracted genes whose total exon lengths are greater than 200bp from the Pickrell et al. data. I calculated a fragments per kilobase of exons per million mapped fragments (FPKM) (Trapnell *et al.*, 2010) for each gene as follows:

$$\text{FPKM}_{i,j} = \frac{10^9 x_{i,j}}{M_j L_i} \quad (3.4).$$

The FPKM is commonly used to convert the RNA-Seq read counts to comparable normalized expression values. Where $x_{i,j}$ represents the number of reads mapped to a gene i of a sample j . M_j is a total number of reads for a sample j , and L_i is a total exon length of a gene i . 10^9 represents a normalization coefficient for the exon length and the total number of reads. From the FPKM of each sample, the mean and variance of the FPKM values for each gene were calculated and paired as FPKM statistics data. The FPKM statistics data were filtered if the means of the FPKM were less than one. Second, the FPKM statistics data and mouse cDNA sequence were matched so that the total exon length of the original gene of the FPKM statistics data and the total length of the mouse cDNA sequence are nearest. As a

result, 14,907 combinations of the FPKM statistics data and mouse cDNA sequence were obtained.

The control distributions were generated based on the FPKM statistics data, which contain the mean and variance of the FPKM values. I calculated the mean and variance of read counts from the mean and variance of FPKM values as follows (the detail of an equation conversion is shown at the end of this section):

$$E[x_{i,j \in k}] = \frac{M_k L_i}{10^9} E[\text{FPKM}_{i,j \in k}] \quad (3.5)$$

$$\text{Var}[x_{i,j \in k}] = \left(\frac{M_k L_i}{10^9} \right)^2 \text{Var}[\text{FPKM}_{i,j \in k}] \quad (3.6),$$

where M_k represents the number of total reads for any samples of a condition k . In this benchmark study, the numbers of the total reads were assumed to be evenly distributed among samples. According to the mean and variance of the read counts, I generated a probabilistic distribution as the control distribution, which follows the negative binomial distribution. Finally, the read counts were sampled from the control distribution as many times as the number of replicates, and the RNA-Seq reads were generated from the matched mouse cDNA sequence as mentioned above. Thus, I simulated 14,907 mouse transcripts (genes) RNA-Seq datasets for my benchmark study.

As to generate datasets, the total number of reads for each sample was simulated approximately 1.0×10^7 , the single-end RNA-Seq reads with 120 bp of length were generated in accordance with the read counts. For each condition, the sampling of the read count was repeated with the number of replicates progressively ranging from 1-6, 12, 18, and 24. Moreover, I simulated three times for each parameter set.

At the end of this section, the equation conversion from the mean and variance of the FPKM (equation 3.4) to the mean and variance of read counts can be performed as follows:

$$\begin{aligned} E[\text{FPKM}_{i,j \in k}] &= \frac{1}{N} \sum_{a \in k} \left(\frac{10^9 x_{i,a}}{M_k L_i} \right) \\ &= \frac{10^9}{M_k L_i} \frac{1}{N} \sum_{a \in k} x_{i,a} \\ &= \frac{10^9}{M_k L_i} E[x_{i,j \in k}] \\ \therefore E[x_{i,j \in k}] &= \frac{M_k L_i}{10^9} E[\text{FPKM}_{i,j \in k}] \end{aligned}$$

$$\begin{aligned}
Var[FPKM_{i,j \in k}] &= \frac{1}{N-1} \sum_{a \in k} (E[FPKM_{i,j \in k}] - FPKM_{i,a})^2 \\
&= \frac{(10^9/M_k L_i)^2}{N-1} \sum_{a \in k} \left(\frac{1}{N} \sum_{l \in k} x_{i,l} - x_{i,a} \right)^2 \\
&= (10^9/M_k L_i)^2 Var[x_{i,j \in k}] \\
\therefore Var[x_{i,j \in k}] &= \left(\frac{M_k L_i}{10^9} \right)^2 Var[FPKM_{i,j \in k}]
\end{aligned}$$

3.3.2 Analysis for benchmark study of clustering

For the benchmark study of clustering, simulated reads were mapped to a mouse reference genome (GRCm38-release71) with a gene annotation file by using TopHat2 (Kim *et al.*, 2013) (version-2.0.8) with Bowtie2 (Langmead and Salzberg, 2012) (version-2.1.0). The gene annotations were confined to 14,907 simulated gene annotations. In this benchmark study, edgeR (Robinson, McCarthy and Smyth, 2010), DESeq (Anders and Huber, 2010), DESeq2 (Love, Huber and Anders, 2014), and cuffdiff2 (Trapnell *et al.*, 2013) were used as statistical tools for detecting significant differential expression between a pair of conditions. At the start of the differential expression analysis, HTSeq (Anders, Pyl and Huber, 2015) (version-0.6.1) was used with the aforementioned confined gene annotation file to count mapped reads for each gene per condition. Subsequently, edgeR (version-3.6.8), DESeq (version-1.16.0), and DESeq2 (version-1.4.5) were used to apply statistical tests for all pairs of conditions to detect significant differential expression. Correspondingly, cuffdiff2 (version-2.2.1) was used with the confined gene annotation file to estimate expression levels and to apply statistical tests. For correcting the multiple testing of differential expression using four tools, I controlled the false discovery rate (FDR). The null hypothesis for multiple conditions was “the expression level of a gene is not different in any of the experimental conditions”; therefore, I estimated the FDR from (total number of genes) $\times$ (total number of pairs of conditions) null hypotheses, and controlled it to under 5% by using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995). The adjusted p-values were calculated by using R package p.adjust. Consequently, normalized expression values and statistical test results for each gene were obtained from each differential expression analysis tool. Lastly, DEclust and conventional hierarchical clustering algorithms (existing methods) were applied to genes whose normalized expression values were non-zero in at least one condition. For applying DEclust, the statistical test results for each gene were discretized and vectorized to pairwise DET profiles. The group average method, single-linkage method, complete-linkage method, and Ward’s method (Ward, 1963) were used for inter-cluster distance measures, and the Euclidean distance, Pearson’s correlation, and cosine distance

were used for inter-gene distance (or similarity) measures for the existing method and for DEclust as secondary distance measures.

3.3.3 Evaluation of clustering accuracy

For all pairs of genes featuring identical class labels, I assessed whether they belonged to the same cluster after clustering. The pairs of genes that featured identical class labels and belonged to the same cluster were true positives. The pairs of genes that featured distinct labels and belonged to different clusters were true negatives. The pairs of genes that featured the same labels but belonged to different clusters were false negatives. The pairs of genes that featured distinct labels but belonged to the same clusters were false positives. Thus, I converted the clustering problem into a binary classification problem. The true-positive rate (TPR; true positive / condition positive) and the false-positive rate (FPR; false positive / condition negative) were calculated at each merging step in hierarchical clustering. An ROC curve was created by plotting the TPR against the FPR, and the AUC of the ROC was used as a measure for clustering evaluation.

3.3.4 Analysis for benchmark study of DEG detection

For the benchmark study of DEG detection, I used the same simulation datasets that were used for the benchmark study of clustering. In the same manner as the analysis for the benchmark study of clustering, the simulated reads were mapped to a mouse reference genome, the read counts of each gene per condition were obtained by using HTSeq, and edgeR, DESeq, DESeq2, and cuffdiff2 were applied for statistical testing for all pairs of conditions and to calculate the normalized expression values for each condition. Subsequently, DEclust was applied to divide 14,907 simulated genes into clusters by quitting the merge step if the distance between any pair of clusters n, m satisfied $D(C_n, C_m) \geq 1$. Separately, the statistical tests for multiple conditions were performed by using edgeR, DESeq2, or multiDE (Kang, Du and Zhang, 2016) (version-1.0).

3.3.5 Evaluation of DEG detection accuracy

For evaluation of the DEG detection power of DEclust, I defined a “DEG cluster” and “non-DEG cluster”. If at least one element of a pairwise DET profile of a cluster was 1 or -1 , the cluster was a DEG cluster; otherwise (that is, all elements were 0), it was a non-DEG cluster. I defined whether a gene was selected as a DEG in a simulation as the correct label, and tested whether a gene belonged to a DEG cluster or non-DEG cluster obtained by using DEclust; therefore, if a gene was selected as a DEG and it belonged to any of the DEG clusters, it was a true positive. If a gene was not selected as a DEG and it belonged to any of the non-DEG clusters, it was a true negative. If a gene was not selected as a DEG but belonged to any of the DEG clusters, it was a false positive, and a gene selected as a DEG but

that belonged to any of the non-DEG clusters was a false negative. For evaluation of the tools for detecting DEGs among multiple conditions, I defined whether a gene was selected as a DEG as the correct label, and tested whether a gene was detected as a DEG. Therefore, if a gene was simulated as a DEG and was detected as a DEG by a tool, it was true positive. If a gene was not simulated as a DEG and was not detected as a DEG, it was true negative. If a gene was not simulated as a DEG but was detected as a DEG, it was false-positive. If a gene was simulated as a DEG but was not detected as a DEG, it was false-negative. I evaluated edgeR, DESeq2, multiDE, and DEClust according to the true-positive rate (TPR), positive predictive rate (PPV), accuracy, and F-measure (harmonic mean of TPR and PPV).

3.3.6 Mice

This study was conducted in strict accordance with the recommendations in the Guide for the Care and Use of Laboratory Animals of the Ministry of Education, Culture, Sports, Science, and Technology of Japan. The protocol was approved by the Committee on the Ethics of Animal Experiments of Chiba Cancer Center (Permit Number: 16–15). All efforts were made to minimize suffering. FVB/N mice were purchased from CLEA Japan, Tokyo, Japan.

3.3.7 Skin carcinogenesis and tumor sampling

DMBA (7,12-dimethylbenz(a)anthracene) was used as a carcinogen and TPA (12-O-tetradecanoylphorbol-13-acetone) was used as a promoter. Collaborators treated 17 FVB/N mice according to a two-stage carcinogenesis protocol. At 8 weeks of age, the female mice were carefully shaved with electric clippers, and two days after shaving, a single dose of DMBA (25 µg/mouse in 200 µL of acetone) was applied to the shaved dorsal back skin. One week after initiation, tumors were promoted with TPA (10 µg/mouse in 200 µL of acetone) twice weekly for 20 weeks. The number and size (diameter, in mm) of each papilloma were recorded from 8 until 20 weeks, and carcinoma development was monitored for up to 40 weeks post-TPA treatment. The health of the mice was monitored every day by laboratory animal technicians. The mice were sacrificed by cervical dislocation when the tumor volume reaches 10% of the individual's body weight. Lastly, normal skin, papilloma, carcinoma, and metastatic tumor samples were collected from each of two mice (two biological replicates, four stages).

3.3.8 Sample preparation and mRNA-Seq

Total RNA was extracted from the papillomas, carcinomas, metastatic tumors, and normal tissues by using the AGPC (acid guanidinium thiocyanate-phenol-chloroform) method (Chomczynski, 1993), and the total RNA samples were treated with DNase to eliminate genomic DNA. Each RNA sample was then reverse-transcribed using the SMARTER[®] Ultra[™] Low RNA Kit (634935; Takara, Otsu, Japan); 10 ng of total RNA was used and PCR

was performed for 12 cycles. The products were treated according to the TruSeq™ DNA Sample Preparation v2 Guide for Illumina sequencing. In this step, PCR was performed for 15 cycles. Lastly, mRNA-Seq was performed using the Illumina Genome Analyzer IIX for eight samples (two replicates, four stages) according to the cBot workflow and Genome Analyzer IIX Paired End Run workflow. For each sample, 84 bp single-end reads were obtained.

3.3.9 RNA-Seq data analysis

Low-quality mRNA-Seq reads were filtered by using FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html) with the options -q 20 and -p 80. The remaining reads were mapped to the mouse reference genome (GRCm38-release71.fa) with gene annotation (Mus musculus.GRCm38.71.gtf) by using TopHat2 (version-2.0.8) with Bowtie2 (version-2.1.0). To identify significant differential expression, HTSeq (version-0.6.1) and DESeq2 (version-1.4.5) were used with the FDR controlled to <5%. DEClust was applied to the results of DESeq2, and a hierarchical tree of genes was calculated. The group average method with cosine distance was used for a secondary distance measure if any pairs of clusters featured the same distance under the definition of equation (3.3) (S1 Text). The final clusters were determined by quitting the merge step if the distance between any pair of clusters n , m satisfied $D(C_n, C_m) \geq 1$. The discriminative genes among tumorigenesis stages were investigated using a GSEA software (Subramanian *et al.*, 2005) for each stage and cluster. Hallmark gene sets were used and the gene sets with an FDR q-value below 5% were identified; these were coherently expressed signatures derived by aggregating several Molecular Signatures Database gene sets. Moreover, the conventional hierarchical clustering method (the group average method with cosine distance) was applied to the detected DEGs for a comparison with the results obtained using DEClust. The hierarchical tree was cut into 16 clusters to obtain the same number of clusters as in the analysis with DEClust, and the GSEA was performed on each gene cluster.

3.3.10 Data deposition

The source code of DEClust and other related data are deposited in the web site (<http://www.dna.bio.keio.ac.jp/software/DEClust>) and the Mendeley Data web site (<http://dx.doi.org/10.17632/wdfb5w7vbb.1>, DOI: 10.17632/wdfb5w7vbb.1). All mRNA-Seq reads have been deposited in the Read Archive at DDBJ (Accession number DRA005289).

3.4 Conclusion

In this chapter, I proposed a novel transcriptome analysis method, DEClust, to search for differentially expressed and statistically overrepresented patterns of gene expression profiles

among multiple experimental conditions. The results of the benchmark study showed not only that DEClust can extract statistically significant gene clusters from multi-conditional transcriptome data more accurately than conventional clustering methods, but also that the accuracy of DEClust increases if two or more replicates of the quantitative experiments are available. Moreover, the results of the application to biological data (mRNA-Seq analysis of mouse tumorigenesis stages) suggested that DEClust can potentially yield biologically significant gene clusters.

The ready availability of high-throughput sequencing technologies has facilitated large-scale quantitative analyses in all fields of life science. As a result, I expect increased demand for software that is applicable to multi-conditional studies. DEClust can be applied to any multi-conditional transcriptome data, and to the results of any DEG detection tool given an appropriate input format; thus, DEClust can be used for a wide range of applications for transcriptome data analysis into the future. The computational complexity of DEClust is $O(N^3)$, where N is the number of objects (genes). The distances between all pairs of clusters (equation (3.3)) must be updated in every iteration of hierarchical clustering. As a reference, when I performed the mouse transcriptome analysis, the computational time of DEClust was approximately ten minutes for 6,584 DEGs with 8 threads of Opteron 1.4 GHz CPUs. The source code of DEClust is freely available at <http://www.dna.bio.keio.ac.jp/software/DEClust> or <http://dx.doi.org/10.17632/wdfb5w7vbb.1>. DEClust is implemented using C++ with boost C++ library (<http://www.boost.org/>) and OpenMP (<http://openmp.org/wp/>).

Chapter 4

Conclusion and future work

In this dissertation, I have presented the studies analyzing the process of cancer progression. My aim is to understand the entire process of tumorigenesis by a longitudinal approach that is complementary to existing cross-cancer-sectional approach of tumors. I conducted genome and transcriptome sequencing analysis for the tumors obtained by the experimental carcinogenesis of mice.

I have described the results of cancer genome study in Chapter 2 and cancer transcriptome study in Chapter 3. Through these studies, I gave insights of the evolutionary process in the tumors over time and the gene clusters whose expression patterns were discriminative in the process of the tumor progression (Fig. 4.1). The results of the genome analysis (Chapter 2) suggested that each of the tumor-series acquired and inherited different mutations, and the intra-tumor structure of the malignant series had been dramatically changed between 9th and 14th weeks. The results of the transcriptome analysis (Chapter 3) suggested that each tumor progression stage has different gene expression profile, meaning that different genes were activated in each of the stages.

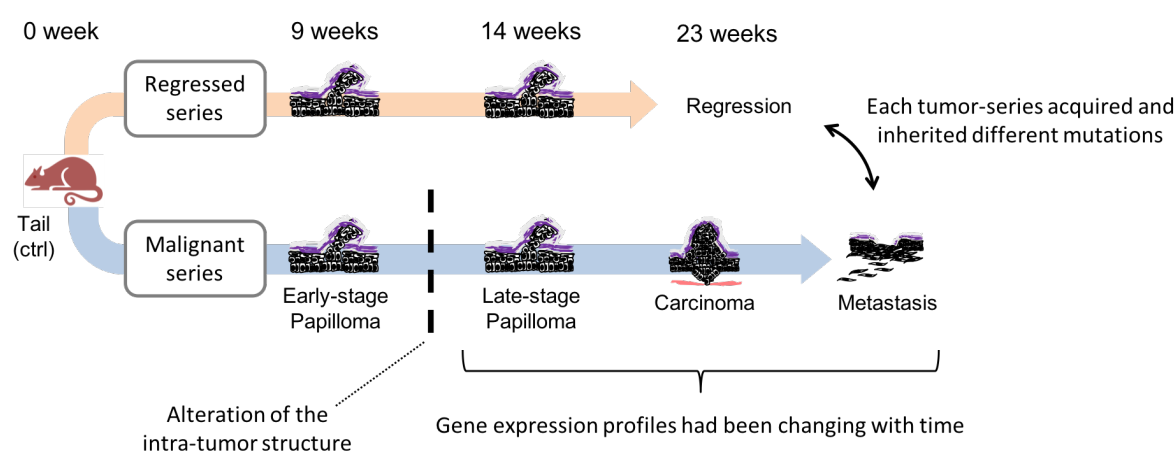


Figure 4.1. Summary of the genome and transcriptome studies. The results of the genome analysis suggested that each of the tumor-series acquired and inherited different mutations, and the intra-tumor structure of the malignant series had been dramatically changed between 9th and 14th weeks. The results of the transcriptome analysis suggested that each tumor progression stage has different gene expression profile.

In the light of the initial purpose of this dissertation, I and my collaborators were able to advance understanding the entire process of tumor progression through conducting each of the omics studies. An example of the research that directly analyzes the process of tumor progression is few; hence, my results are meaningful and I consider that it is necessary to advance such a direct approach to reveal the entire process of cancer progression.

As I described in Chapter 1, the inter- and intra-tumor heterogeneity make it difficult to understand cancer disease. Since each tumor evolves with the accumulation of different mutations, it is indispensable to investigate numerous tumors with diverse fates for full understanding of the entire evolutionary process in tumors. In this point, my studies are only case studies, so I am now planning to scale up by increasing the number of mouse individuals to obtain >50 tumor-series. Moreover, the extension of the targeted genome sequencing to WGS or WES can be considered. Recent improvement of sequencing technologies has provided a reduction in cost to the amounts of short-reads. Although the targeted sequencing technology I adopted allows deep sequencing for the targeted genomic region at low cost, I predict that equivalent read coverage on a whole genome scale will be practically obtainable in the near future. Furthermore, integrative analysis of genomic mutations and transcriptome dynamics can be considered for more precise detection of the cancer drivers, selective events, and fate determination of tumors. These scale-up of the analysis objects and analysis methods should lead to a comprehensive understanding of tumorigenesis process.

To be more specific, I am planning to infer intra-tumor evolutionary trees based on genomic mutations for various tumors by WGS or WES analysis. From the changes of the evolutionary trees with time differences between tumors, I can know when and what kind of sub-clones produce malignant tumors. Furthermore, with reference to the transcriptome dynamics, I can clarify the cause and the timing of the cancer driver events and the fate determination of tumors at the genetic level. I consider that these future results should contribute not only to reveal the entire process of tumor progression, but also to provide biomarkers for early diagnosis and early treatment.

Acknowledgment

First of all, I would like to express my sincere thanks to Professor Yasubumi Sakakibara who has provided the comprehensive support for my study as a supervisor of my bachelor, master, and doctor courses.

I also would like to express my cordial gratitude to Full-time lecturer Kengo Sato who has provided a number of technical advices for my study and maintenance of the research environment of Sakakibara laboratory. I also would like to thank the collaborators, Dr. Tsuyoshi Hachiya, Dr. Sumitaka Hase, Dr. Fuyuki Ishikawa, Dr. Yuichi Wakabayashi, and Dr. Kazuhiro Okumura. Especially, Dr. Tsuyoshi Hachiya taught me the fundamental knowledges of bioinformatics and the basic attitude towards research.

I am very grateful to my colleagues in Sakakibara Laboratory, for making a good academic environment. I would like to thank Dr. Kris Popendorf, Dr. Yutaka Saito, Dr. Mamiko Kamada, and Mr. Yohei Sugawara for giving me critical comments on my study as the senior members in the laboratory. My thanks also go to the other colleagues, Dr. Vasanthan Jayakumar, Mr. Yugaku Tanaka, Ms. Sae Shirakizawa, Ms. Misato Seki, Mr. Manato Akiyama, and Mr. Tatsuya Mizuguchi, who helped me a lot to feel comfort studying and living in the laboratory.

Lastly, I would like to express my sincere thanks to Professor Yasubumi Sakakibara, Professor Kotaro Oka, Professor Masaya Imoto, and Associate Professor Akira Funahashi for examining and judging my doctoral dissertation.

References

- van 't Veer, L. J. *et al.* (2002) 'Gene expression profiling predicts clinical outcome of breast cancer', *Nature*, 415(6871), pp. 530–536. doi: 10.1038/415530a.
- Alameda, J. P. *et al.* (2010) 'An inactivating CYLD mutation promotes skin tumor progression by conferring enhanced proliferative, survival and angiogenic properties to epidermal cancer cells', *Oncogene*, 29(50), pp. 6522–6532. doi: 10.1038/onc.2010.378.
- Alexandrov, L. B. *et al.* (2013) 'Signatures of mutational processes in human cancer', *Nature*, 500(7463), pp. 415–421. doi: 10.1038/nature12477.
- Anders, S. and Huber, W. (2010) 'Differential expression analysis for sequence count data.', *Genome biology*, 11(10), p. R106. doi: 10.1186/gb-2010-11-10-r106.
- Anders, S., Pyl, P. T. and Huber, W. (2015) 'HTSeq-A Python framework to work with high-throughput sequencing data', *Bioinformatics*, 31(2), pp. 166–169. doi: 10.1093/bioinformatics/btu638.
- Aoto, Y. *et al.* (2017) 'DEclust: A statistical approach for obtaining differential expression profiles of multiple conditions', *PLoS ONE*, 12(11). doi: 10.1371/journal.pone.0188285.
- Aoto, Y. *et al.* (2018) 'Time-Series Analysis of Tumorigenesis in a Murine Skin Carcinogenesis Model', *Scientific Reports*, 8(1), p. 12994. doi: 10.1038/s41598-018-31349-x.
- Balmain, A. and Yuspa, S. H. (2014) 'Milestones in skin carcinogenesis: the biology of multistage carcinogenesis', *The Journal of investigative dermatology*, pp. E2–E7. doi: 10.1038/skinbio.2014.2.
- Benjamini, Y. and Hochberg, Y. (1995) 'Controlling the False Discovery Rate : A Practical and Powerful Approach to Multiple Testing', *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1), pp. 289–300. doi: 10.2307/2346101.
- Bolger, A. M., Lohse, M. and Usadel, B. (2014) 'Trimmomatic: A flexible trimmer for Illumina sequence data', *Bioinformatics*, 30(15), pp. 2114–2120. doi: 10.1093/bioinformatics/btu170.
- Bollag, G. *et al.* (1996) 'Loss of NF1 results in activation of the Ras signaling pathway and leads to aberrant growth in haematopoietic cells', *Nature Genetics*, 12(2), pp. 144–148. doi: 10.1038/ng0296-144.

Brown, K., Buchmann, A. and Balmain, A. (1990) 'Carcinogen-induced mutations in the mouse c-Ha-ras gene provide evidence of multiple pathways for tumor progression.', *Proceedings of the National Academy of Sciences of the United States of America*, 87(2), pp. 538–42. doi: 10.1073/pnas.87.2.538.

Burrell, R. A. *et al.* (2013) 'The causes and consequences of genetic heterogeneity in cancer evolution', *Nature*, 501(7467), pp. 338–345. doi: 10.1038/nature12625.

Chomczynski, P. (1993) 'A reagent for the single-step simultaneous isolation of RNA, DNA and proteins from cell and tissue samples', *BioTechniques*, 15(3), pp. 532–537. doi: 10.1038/news.2010.498.

Cibulskis, K. *et al.* (2013) 'Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples', *Nature Biotechnology*, 31(3), pp. 213–219. doi: 10.1038/nbt.2514.

D'haeseleer, P. (2005) 'How does gene expression clustering work?', *Nature biotechnology*, 23(12), pp. 1499–1501. doi: 10.1038/nbt1205-1499.

Davis, A., Gao, R. and Navin, N. (2017) 'Tumor evolution: Linear, branching, neutral or punctuated?', *Biochimica et Biophysica Acta - Reviews on Cancer*, 1867(2), pp. 151–161. doi: 10.1016/j.bbcan.2017.01.003.

Einspahr, J. G. *et al.* (1999) 'Relationship of p53 mutations to epidermal cell proliferation and apoptosis in human UV-induced skin carcinogenesis', *Neoplasia.*, 1(5), pp. 468–475.

Eisenberg, E. and Levanon, E. Y. (2003) 'Human housekeeping genes are compact', *Trends in Genetics*, pp. 362–365. doi: 10.1016/S0168-9525(03)00140-9.

Elowitz, M. B. *et al.* (2002) 'Stochastic gene expression in a single cell', *Science*, 297(5584), pp. 1183–1186. doi: 10.1126/science.1070919.

Fearon, E. R. and Vogelstein, B. (1990) 'A genetic model for colorectal tumorigenesis', *Cell*, pp. 759–767. doi: 10.1016/0092-8674(90)90186-I.

Forbes, S. A. *et al.* (2017) 'COSMIC: Somatic cancer genetics at high-resolution', *Nucleic Acids Research*, 45(D1), pp. D777–D783. doi: 10.1093/nar/gkw1121.

Futreal, P. A. *et al.* (2004) 'A census of human cancer genes', *Nature Reviews Cancer*, 4(3), pp. 177–183. doi: 10.1038/nrc1299.

Gerlinger, M. *et al.* (2012) 'Intratumor Heterogeneity and Branched Evolution Revealed by

- Multiregion Sequencing', *New England Journal of Medicine*, 366(10), pp. 883–892. doi: 10.1056/NEJMoa1113205.
- Ginos, M. A. *et al.* (2004) 'Identification of a Gene Expression Signature Associated with Recurrent Disease in Squamous Cell Carcinoma of the Head and Neck', *Cancer Research*, 64(1), pp. 55–63. doi: 10.1158/0008-5472.CAN-03-2144.
- Greaves, M. and Maley, C. C. (2012) 'Clonal evolution in cancer.', *Nature*, 481(7381), pp. 306–13. doi: 10.1038/nature10762.
- Hanahan, D. and Weinberg, R. A. (2011) 'Hallmarks of cancer: The next generation', *Cell*, pp. 646–674.
- Huang, P. Y. and Balmain, A. (2014) 'Modeling cutaneous squamous carcinoma development in the mouse', *Cold Spring Harbor Perspectives in Medicine*, 4(9). doi: 10.1101/cshperspect.a013623.
- Jamal-Hanjani, M. *et al.* (2017) 'Tracking the Evolution of Non–Small-Cell Lung Cancer', *New England Journal of Medicine*, 376(22), pp. 2109–2121. doi: 10.1056/NEJMoa1616288.
- Kandoth, C. *et al.* (2013) 'Mutational landscape and significance across 12 major cancer types', *Nature*, 502(7471), pp. 333–339. doi: 10.1038/nature12634.
- Kang, G., Du, L. and Zhang, H. (2016) 'multiDE: a dimension reduced model based statistical method for differential expression analysis using RNA-sequencing data with multiple treatment conditions.', *BMC bioinformatics*, 17(1), p. 248. doi: 10.1186/s12859-016-1111-9.
- Kantidakis, T. *et al.* (2016) 'Mutation of cancer driver MLL2 results in transcription stress and genome instability', *Genes and Development*, 30(4), pp. 408–420. doi: 10.1101/gad.275453.115.
- Kim, D. *et al.* (2013) 'TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions.', *Genome biology*, 14(4), p. R36. doi: 10.1186/gb-2013-14-4-r36.
- Koboldt, D. C. *et al.* (2012) 'VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing', *Genome Research*, 22(3), pp. 568–576. doi: 10.1101/gr.129684.111.
- Kovalenko, A. *et al.* (2003) 'The tumour suppressor CYLD negatively regulates NF-kappaB

signalling by deubiquitination.’, *Nature*, 424(6950), pp. 801–5. doi: 10.1038/nature01802.

Langmead, B. and Salzberg, S. L. (2012) ‘Fast gapped-read alignment with Bowtie 2’, *Nat Methods*, 9(4), pp. 357–359. doi: 10.1038/nmeth.1923.

Laplane, M. and Sabatini, D. M. (2013) ‘mTOR signaling in growth control and disease’, *Cell*, 149(2), pp. 274–293. doi: 10.1016/j.cell.2012.03.017.mTOR.

Larson, D. E. *et al.* (2012) ‘Somaticsniper: Identification of somatic point mutations in whole genome sequencing data’, *Bioinformatics*, 28(3), pp. 311–317. doi: 10.1093/bioinformatics/btr665.

Li, H. *et al.* (2009) ‘The Sequence Alignment/Map format and SAMtools’, *Bioinformatics*, 25(16), pp. 2078–2079. doi: 10.1093/bioinformatics/btp352.

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

Love, M. I., Huber, W. and Anders, S. (2014) ‘Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.’, *Genome biology*, 15(12), p. 550. doi: 10.1186/s13059-014-0550-8.

Ludwig, J. A. and Weinstein, J. N. (2005) ‘Biomarkers in cancer staging, prognosis and treatment selection’, *Nature Reviews Cancer*, 5(11), pp. 845–856. doi: nrc1739 [pii]10.1038/nrc1739.

Makohon-Moore, A. P. *et al.* (2017) ‘Limited heterogeneity of known driver gene mutations among the metastases of individual patients with pancreatic cancer’, *Nature Genetics*, 49(3), pp. 358–366. doi: 10.1038/ng.3764.

Martincorena, I. and Campbell, P. J. (2015) ‘Somatic mutation in cancer and normal cells’, *Science*, 349(6255), pp. 1483–1489. doi: 10.1126/science.aab4082.

McCreery, M. Q. *et al.* (2015) ‘Evolution of metastasis revealed by mutational landscapes of chemically induced skin cancers’, *Nature Medicine*, 21(12), pp. 1514–1520. doi: 10.1038/nm.3979.

McGranahan, N. *et al.* (2015) ‘Clonal status of actionable driver events and the timing of mutational processes in cancer evolution’, *Science Translational Medicine*, 7(283), p. 283ra54-283ra54. doi: 10.1126/scitranslmed.aaa1408.

McGranahan, N. and Swanton, C. (2015) ‘Biological and therapeutic impact of intratumor

heterogeneity in cancer evolution', *Cancer Cell*, 27(1), pp. 15–26. doi: 10.1016/j.ccell.2014.12.001.

McGranahan, N. and Swanton, C. (2017) 'Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future', *Cell*, pp. 613–628. doi: 10.1016/j.cell.2017.01.018.

McKenna, A. H. *et al.* (2010) 'The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data', *Genome Research*, 20(9), pp. 1297–1303. doi: 10.1101/gr.107524.110.

Metzker, M. L. (2010) 'Sequencing technologies - the next generation.', *Nature reviews. Genetics*, 11(1), pp. 31–46.

Miller, C. A. *et al.* (2014) 'SciClone: Inferring Clonal Architecture and Tracking the Spatial and Temporal Patterns of Tumor Evolution', *PLoS Computational Biology*, 10(8). doi: 10.1371/journal.pcbi.1003665.

Nassar, D. *et al.* (2015) 'Genomic landscape of carcinogen-induced and genetically induced mouse skin squamous cell carcinoma', *Nature Medicine*, 21(8), pp. 946–954. doi: 10.1038/nm.3878.

Newton, P. K. *et al.* (2014) 'Entropy, complexity, and Markov diagrams for random walk cancer models', *Scientific Reports*, 4. doi: 10.1038/srep07558.

Okumura, K. *et al.* (2012) 'Independent genetic control of early and late stages of chemically induced skin tumors in a cross of a Japanese wild-derived inbred mouse strain, MSM/MS', *Carcinogenesis*, 33(11), pp. 2260–2268. doi: 10.1093/carcin/bgs250.

Page, A. *et al.* (2016) 'Protective role of p53 in skin cancer: Carcinogenesis studies in mice lacking epidermal p53', *Oncotarget*, 7(15). doi: 10.18632/oncotarget.7897.

Pickrell, J. *et al.* (2010) 'Understanding mechanisms underlying human gene expression variation with RNA sequencing', *Nature*, 464(7289), pp. 768–772. doi: 10.1038/nature08872.

Porcu, M. *et al.* (2012) 'Mutation of the receptor tyrosine phosphatase PTPRC (CD45) in T-cell acute lymphoblastic leukemia', *Blood*, 119(19), pp. 4476–4479. doi: 10.1182/blood-2011-09-379958.

Porta, C., Paglino, C. and Mosca, A. (2014) 'Targeting PI3K/Akt/mTOR Signaling in Cancer.', *Frontiers in oncology*, 4(April), p. 64. doi: 10.3389/fonc.2014.00064.

Prensner, J. R. *et al.* (2011) 'Transcriptome sequencing across a prostate cancer cohort

- identifies PCAT-1, an unannotated lincRNA implicated in disease progression', *Nature Biotechnology*, 29(8), pp. 742–749. doi: 10.1038/nbt.1914.
- Rawlings, J. S., Rosler, K. M. and Harrison, D. A. (2004) 'The JAK/STAT signaling pathway', *Journal of Cell Science*, 117(8), pp. 1281–1283. doi: 10.1242/jcs.00963.
- Rhodes, D. R. *et al.* (2004) 'ONCOMINE: A Cancer Microarray Database and Integrated Data-Mining Platform', *Neoplasia*, 6(1), pp. 1–6. doi: 10.1016/S1476-5586(04)80047-2.
- Robinson, M. D., McCarthy, D. J. and Smyth, G. K. (2010) 'edgeR: a Bioconductor package for differential expression analysis of digital gene expression data', *Bioinformatics*, 26(1), pp. 139–140. doi: 10.1093/bioinformatics/btp616.
- Roth, A. *et al.* (2014) 'PyClone: Statistical inference of clonal population structure in cancer', *Nature Methods*, 11(4), pp. 396–398. doi: 10.1038/nmeth.2883.
- Ruddell, A. *et al.* (2008) 'p19/Arf and p53 suppress sentinel lymph node lymphangiogenesis and carcinoma metastasis', *Oncogene*, 27(22), pp. 3145–3155. doi: 10.1038/sj.onc.1210973.
- Saunders, C. T. *et al.* (2012) 'Strelka: Accurate somatic small-variant calling from sequenced tumor-normal sample pairs', *Bioinformatics*, 28(14), pp. 1811–1817. doi: 10.1093/bioinformatics/bts271.
- Schwarz, M., Münzel, P. A. and Braeuning, A. (2013) 'Non-melanoma skin cancer in mouse and man', *Archives of Toxicology*, pp. 783–798. doi: 10.1007/s00204-012-0998-9.
- Shain, A. H. *et al.* (2015) 'The Genetic Evolution of Melanoma from Precursor Lesions', *New England Journal of Medicine*, 373(20), pp. 1926–1936. doi: 10.1056/NEJMoa1502583.
- Shannon, C. E. (1948) 'A mathematical theory of communication', *The Bell System Technical Journal*, 27(4), pp. 379–423, 623–656. doi: 10.1002/j.1538-7305.1948.tb00917.x.
- Shendure, J. and Ji, H. (2008) 'Next-generation DNA sequencing', *Nature biotechnology*, 26(10), pp. 1135–1145. doi: 10.1038/nbt1486.
- Sherry, S. T. (2001) 'dbSNP: the NCBI database of genetic variation', *Nucleic Acids Research*, 29(1), pp. 308–311. doi: 10.1093/nar/29.1.308.
- Shin, H. T. *et al.* (2017) 'Prevalence and detection of low-allele-fraction variants in clinical cancer samples', *Nature Communications*, 8(1). doi: 10.1038/s41467-017-01470-y.
- Subramanian, A. *et al.* (2005) 'Gene set enrichment analysis: a knowledge-based approach

for interpreting genome-wide expression profiles', *Proc Natl Acad Sci U S A*, 102(43), pp. 15545–15550. doi: 10.1073/pnas.0506580102.

Swanton, C. (2012) 'Intratumor heterogeneity: Evolution through space and time', *Cancer Research*, pp. 4875–4882. doi: 10.1158/0008-5472.CAN-12-2217.

Tomczak, K., Czerwińska, P. and Wiznerowicz, M. (2015) 'The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge.', *Contemporary oncology (Poznań, Poland)*, 19(1A), pp. A68-77. doi: 10.5114/wo.2014.47136.

Torre, L. A. *et al.* (2015) 'Global Cancer Statistics, 2012', *CA: a cancer journal of clinicians.*, 65(2), pp. 87–108. doi: 10.3322/caac.21262.

Trapnell, C. *et al.* (2010) 'Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation.', *Nature biotechnology*, 28(5), pp. 511–515. doi: 10.1038/nbt.1621.

Trapnell, C. *et al.* (2013) 'Differential analysis of gene regulation at transcript resolution with RNA-seq.', *Nature biotechnology*, 31(1), pp. 46–53. doi: 10.1038/nbt.2450.

Vermeulen, K., Van Bockstaele, D. R. and Berneman, Z. N. (2003) 'The cell cycle: A review of regulation, deregulation and therapeutic targets in cancer', *Cell Proliferation*, pp. 131–149. doi: 10.1046/j.1365-2184.2003.00266.x.

Vilella, A. J. *et al.* (2009) 'EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates', *Genome Research*, 19(2), pp. 327–335. doi: 10.1101/gr.073585.107.

Vogelstein, B. *et al.* (2013) 'Cancer Genome Landscapes', *Science*, 339(6127), pp. 1546–1558. doi: 10.1126/science.1235122.

Wang, Z., Gerstein, M. and Snyder, M. (2009) 'RNA-Seq: a revolutionary tool for transcriptomics.', *Nature reviews. Genetics*, 10(1), pp. 57–63. doi: 10.1038/nrg2484.

Ward, J. H. (1963) 'Hierarchical grouping to optimize an objective function', *Journal of the American Statistical Association*, pp. 236–244. doi: 10.1080/01621459.1963.10500845.

Williams, M. J. *et al.* (2016) 'Identification of neutral tumor evolution across cancer types', *Nature Genetics*. Nature Publishing Group, 48(3), pp. 1–9. doi: 10.1038/ng.3489.

Wilm, A. *et al.* (2012) 'LoFreq: A sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets', *Nucleic*

Acids Research, 40(22), pp. 11189–11201. doi: 10.1093/nar/gks918.

Yu, H. *et al.* (2014) ‘Revisiting STAT3 signalling in cancer: new and unexpected biological functions’, *Nat Rev Cancer*, 14(11), pp. 736–746. doi: 10.1038/nrc3818.

Appendix A – List of publications

Journal papers (related to this dissertation)

1. Yoshimasa Aoto, Kazuhiro Okumura, Tsuyoshi Hachiya, Sumitaka Hase, Yuichi Wakabayashi, Fuyuki Ishikawa, Yasubumi Sakakibara, Time-Series Analysis of Tumorigenesis in a Murine Skin Carcinogenesis Model, Scientific Reports, Aug. 29, 2018, doi: 10.1038/s41598-018-31349-x.
2. Yoshimasa Aoto, Tsuyoshi Hachiya, Kazuhiro Okumura, Sumitaka Hase, Kengo Sato, Yuichi Wakabayashi, Yasubumi Sakakibara, DEclust: A statistical approach for obtaining differential expression profiles of multiple conditions, PLoS ONE, Nov. 21, 2017, doi: 10.1371/journal.pone.0188285.

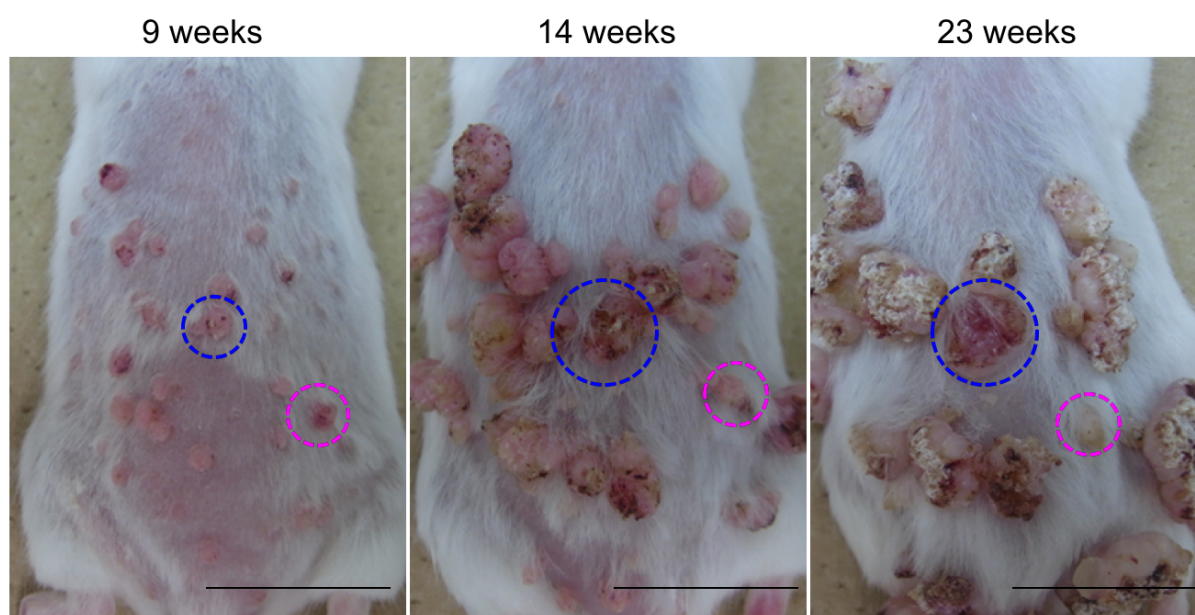
Journal papers (others)

1. Kazuhiro Okumura, Megumi Saito, Eriko Isogai, Yoshimasa Aoto, Tsuyoshi Hachiya, Yasubumi Sakakibara, Yoshinori Katsuragi, Satoshi Hirose, Ryo Kominami, Ryo Goitsuka, Takuro Nakamura, Yuichi Wakabayashi, Meis1 regulates epidermal stem cells and is required for skin tumorigenesis, PLoS ONE, 9(7): e102111, doi:10.1371/journal.pone.0102111.

International conferences

1. Yoshimasa Aoto*, Tsuyoshi Hachiya, Kazuhiro Okumura, Sumitaka Hase, Kengo Sato, Yuichi Wakabayashi, Yasubumi Sakakibara, Digital Clustering: An accurate clustering algorithm based on statistical test results, The fourteenth Asia Pacific Bioinformatics Conference, San Francisco, United States, January 2016 (poster presentation)
2. Sota Takagi, Yuki Kouda, Hanae Nodono*, Yoshimasa Aoto, Kazutoshi Yoshitake, Takashi Gojyobori, Kengo Sato, Yasubumi Sakakibara, Midori Matsumoto, Comprehensive screening of peptide hormones that induce planarian sexualization, 1st Asian Planaria Meeting, Hong Kong, China, May 2014 (poster presentation)

Appendix B – Supplementary information of genome analysis



Supplementary Figure S2.1. Photographic images of papillomas and carcinoma in the two-stage skin carcinogenesis by DMBA-TPA protocol. Papillomas appear as outgrowths on the dorsal back skin of mouse. The blue circles indicate malignant series tumor. The pink circles indicate regressed series tumors. Scale bar = 20 mm.

Supplementary Table S2.1. List of target genes.

ENSEMBL GENE ID	GENE SYMBOL	ENSEMBL GENE ID	GENE SYMBOL
ENSMUSG00000000088	Cox5a	ENSMUSG00000028465	Tln1
ENSMUSG00000000326	Comt	ENSMUSG00000028478	Clta
ENSMUSG00000000346	Dazap2	ENSMUSG00000028530	Jak1
ENSMUSG00000000530	Acvrl1	ENSMUSG00000028648	Ndufs5
ENSMUSG00000000568	Hnrnpd	ENSMUSG00000028670	Lypla2
ENSMUSG00000000581	C1d	ENSMUSG00000028691	Prdx1
ENSMUSG00000000594	Gm2a	ENSMUSG00000028692	Akr1a1
ENSMUSG00000001016	Ilf2	ENSMUSG00000028745	Capzb
ENSMUSG00000001082	Mfsd10	ENSMUSG00000028755	Cda
ENSMUSG00000001119	Col6a1	ENSMUSG00000028757	Ddost
ENSMUSG00000001127	Araf	ENSMUSG00000028798	Eif3i
ENSMUSG00000001150	Mcm3ap	ENSMUSG00000028809	Srrm1
ENSMUSG00000001270	Ckb	ENSMUSG00000028811	Yars
ENSMUSG00000001289	Pfdn5	ENSMUSG00000028837	Psmb2
ENSMUSG00000001416	Cct3	ENSMUSG00000028893	Sesn2
ENSMUSG00000001472	Tcf25	ENSMUSG00000028955	Vamp3
ENSMUSG00000001517	Foxm1	ENSMUSG00000028964	Park7
ENSMUSG00000001761	Smo	ENSMUSG00000028980	H6pd
ENSMUSG00000001786	Fbxo7	ENSMUSG00000028991	Mtor
ENSMUSG00000001833	Sept7	ENSMUSG00000028998	Tomm7
ENSMUSG00000001847	Rac1	ENSMUSG00000029020	Mfn2
ENSMUSG00000001870	Ltbp1	ENSMUSG00000029062	Cdk11b
ENSMUSG00000002015	Bcap31	ENSMUSG00000029070	Mxra8
ENSMUSG00000002413	Braf	ENSMUSG00000029106	Add1
ENSMUSG00000002504	Slc9a3r2	ENSMUSG00000029131	Dnajb6
ENSMUSG00000002524	Puf60	ENSMUSG00000029231	Pdgfra
ENSMUSG00000002768	Mea1	ENSMUSG00000029465	Arpc3
ENSMUSG00000002778	Kdelr1	ENSMUSG00000029472	Anapc5
ENSMUSG00000002814	Top3a	ENSMUSG00000029538	Srsf9
ENSMUSG00000002949	Timm44	ENSMUSG00000029580	Actb
ENSMUSG00000003037	Rab8a	ENSMUSG00000029616	Erp29
ENSMUSG00000003068	Stk11	ENSMUSG00000029623	Pdap1
ENSMUSG00000003072	Atp5d	ENSMUSG00000029713	Gnb2
ENSMUSG00000003344	Btbd2	ENSMUSG00000030007	Cct7
ENSMUSG00000003378	Grik5	ENSMUSG00000030120	Mlf2
ENSMUSG00000003380	Rabac1	ENSMUSG00000030180	Kdm5a

ENSMUSG00000003402	Prkcsb	ENSMUSG000000030201	Lrp6
ENSMUSG00000003410	Elavl3	ENSMUSG000000030246	Ldhd
ENSMUSG00000003528	Slc25a1	ENSMUSG000000030265	Kras
ENSMUSG00000003813	Rad23a	ENSMUSG000000030337	Vamp1
ENSMUSG00000003868	Ruvbl2	ENSMUSG000000030591	Psmd8
ENSMUSG00000003970	Rpl8	ENSMUSG000000030602	Pak4
ENSMUSG00000004054	Map3k11	ENSMUSG000000030678	Maz
ENSMUSG00000004264	Phb2	ENSMUSG000000030795	Fus
ENSMUSG00000004937	Sgta	ENSMUSG000000030849	Fgfr2
ENSMUSG00000005043	Sgsh	ENSMUSG000000030890	Ilk
ENSMUSG00000005054	Cstb	ENSMUSG000000030954	Gp2
ENSMUSG00000005103	Wdr1	ENSMUSG000000031024	St5
ENSMUSG00000005198	Polr2a	ENSMUSG000000031029	Eif3f
ENSMUSG00000005370	Msh6	ENSMUSG000000031066	Usp11
ENSMUSG00000005483	Dnajb1	ENSMUSG000000031167	Rbm3
ENSMUSG00000005540	Fcer2a	ENSMUSG000000031207	Msn
ENSMUSG00000005566	Trim28	ENSMUSG000000031311	Nono
ENSMUSG00000005575	Ube2m	ENSMUSG000000031447	Lamp1
ENSMUSG00000005610	Eif4g2	ENSMUSG000000031511	Arhgef7
ENSMUSG00000005621	Zfp592	ENSMUSG000000031535	Dkk4
ENSMUSG00000005672	Kit	ENSMUSG000000031586	Rbpms
ENSMUSG00000005779	Psmc4	ENSMUSG000000031672	Got2
ENSMUSG00000005871	Apc	ENSMUSG000000031729	Ist1
ENSMUSG00000006057	Atp5g1	ENSMUSG000000031760	Mt3
ENSMUSG00000006095	Tbcb	ENSMUSG000000031776	Arl2bp
ENSMUSG00000006304	Arcp2	ENSMUSG000000031785	Gpr56
ENSMUSG00000006315	Tmem147	ENSMUSG000000031788	Kifc3
ENSMUSG00000006333	Rps9	ENSMUSG000000031818	Cox4i1
ENSMUSG00000006344	Ggt5	ENSMUSG000000031839	Hsbp1
ENSMUSG00000006498	Ptbp1	ENSMUSG000000031948	Kars
ENSMUSG00000006932	Ctnnb1	ENSMUSG000000031950	Gabarapl2
ENSMUSG00000007458	M6pr	ENSMUSG000000031990	Jam3
ENSMUSG00000007564	Ppp2r1a	ENSMUSG000000031996	Aplp2
ENSMUSG00000007850	Hnrnp1	ENSMUSG000000032085	Tagln
ENSMUSG00000007872	Id3	ENSMUSG000000032115	Hyou1
ENSMUSG00000007891	Ctsd	ENSMUSG000000032231	Anxa2
ENSMUSG00000008036	Ap2s1	ENSMUSG000000032294	Pkm
ENSMUSG00000008958	Vps72	ENSMUSG000000032312	Csk

ENSMUSG00000009073	Nf2	ENSMUSG000000032330	Cox7a2
ENSMUSG00000009090	Ap1b1	ENSMUSG000000032356	Rasgrf1
ENSMUSG00000009291	Pttg1ip	ENSMUSG000000032423	Syncrip
ENSMUSG00000009470	Tnpo1	ENSMUSG000000032479	Mtap4
ENSMUSG00000009549	Srp14	ENSMUSG000000032480	Dhx30
ENSMUSG00000009630	Ppp2cb	ENSMUSG000000032498	Mlh1
ENSMUSG00000009927	Rps25	ENSMUSG000000032562	Gnai2
ENSMUSG00000010097	Nxf1	ENSMUSG000000032637	Atxn2l
ENSMUSG00000010376	Nedd8	ENSMUSG000000032737	Inpp1l
ENSMUSG00000012405	Rpl15	ENSMUSG000000032845	Alpk2
ENSMUSG00000012535	Tnpo3	ENSMUSG000000033020	Polr2f
ENSMUSG00000012848	Rps5	ENSMUSG000000033068	Entpd6
ENSMUSG00000013663	Pten	ENSMUSG000000033159	Cnppd1
ENSMUSG00000014294	Ndufa2	ENSMUSG000000033379	Atp6v0b
ENSMUSG00000014426	Map3k4	ENSMUSG000000033430	Terf2ip
ENSMUSG00000014599	Csf1	ENSMUSG000000033569	Bai3
ENSMUSG00000014606	Slc25a11	ENSMUSG000000033809	Alg3
ENSMUSG00000014748	Tex261	ENSMUSG000000033826	Dnahc8
ENSMUSG00000014769	Psmbl	ENSMUSG000000033916	Chmp2a
ENSMUSG00000014859	E2f4	ENSMUSG000000033938	Ndufb7
ENSMUSG00000015120	Ube2i	ENSMUSG000000034165	Cend3
ENSMUSG00000015291	Gdi1	ENSMUSG000000034210	4732418C07Rik
ENSMUSG00000015605	Srf	ENSMUSG000000034659	Tmem109
ENSMUSG00000015721	Nlrp5	ENSMUSG000000034681	Rnps1
ENSMUSG00000015937	H2afy	ENSMUSG000000034868	Myl12b
ENSMUSG00000016427	Ndufa1	ENSMUSG000000034928	Rnf44
ENSMUSG00000016528	Mapkapk2	ENSMUSG000000035027	Map2k2
ENSMUSG00000016554	Eif3d	ENSMUSG000000035086	Becn1
ENSMUSG00000017390	Aldoc	ENSMUSG000000035112	Wnk4
ENSMUSG00000017404	Rpl19	ENSMUSG000000035354	Uvrag
ENSMUSG00000017428	Psmd11	ENSMUSG000000035885	Cox8a
ENSMUSG00000017652	Cd40	ENSMUSG000000036111	Lmo1
ENSMUSG00000018287	Spag7	ENSMUSG000000036606	Plxnb2
ENSMUSG00000018326	Ywhab	ENSMUSG000000036712	Cyld
ENSMUSG00000018340	Anxa6	ENSMUSG000000036721	Zscan12
ENSMUSG00000018537	Pcgf2	ENSMUSG000000036751	Cox6b1
ENSMUSG00000018559	Ctdnep1	ENSMUSG000000037152	Ndufc1
ENSMUSG00000018567	Gabarap	ENSMUSG000000037206	Islr

ENSMUSG00000018697	Aatf	ENSMUSG00000037337	Map4k1
ENSMUSG00000018770	Atp5g3	ENSMUSG00000037362	Nov
ENSMUSG00000018965	Ywhah	ENSMUSG00000037373	Ctbp1
ENSMUSG00000019087	Atp6ap1	ENSMUSG00000037563	Rps16
ENSMUSG00000019210	Atp6v1e1	ENSMUSG00000037646	Vps13b
ENSMUSG00000019461	Plscr3	ENSMUSG00000037706	Cd81
ENSMUSG00000019494	Cops6	ENSMUSG00000037788	Vopp1
ENSMUSG00000019804	Snx3	ENSMUSG00000037916	Ndufv1
ENSMUSG00000019889	Ptpk	ENSMUSG00000038346	Zfp384
ENSMUSG00000020122	Egfr	ENSMUSG00000038366	Laspl
ENSMUSG00000020149	Rab1	ENSMUSG00000038462	Uqcrfs1
ENSMUSG00000020163	Uqcr11	ENSMUSG00000038489	Polr2l
ENSMUSG00000020267	Hint1	ENSMUSG00000038502	Ptov1
ENSMUSG00000020287	Mpg	ENSMUSG00000038612	Mcl1
ENSMUSG00000020321	Mdh1	ENSMUSG00000038618	Rassf7
ENSMUSG00000020358	Hnrnpab	ENSMUSG00000038619	Ensa
ENSMUSG00000020368	Canx	ENSMUSG00000038650	Rnh1
ENSMUSG00000020372	Gnb2l1	ENSMUSG00000038690	Atp5j2
ENSMUSG00000020395	Itk	ENSMUSG00000038909	Myst2
ENSMUSG00000020440	Arf5	ENSMUSG00000038970	Lmtk2
ENSMUSG00000020444	Guk1	ENSMUSG00000039105	Atp6v1g1
ENSMUSG00000020458	Rtn4	ENSMUSG00000039195	111008P14Rik
ENSMUSG00000020484	Xbp1	ENSMUSG00000039205	Ciz1
ENSMUSG00000020538	Srebfl	ENSMUSG00000039262	Prrc2b
ENSMUSG00000020580	Rock2	ENSMUSG00000039452	Snx22
ENSMUSG00000020673	Tpo	ENSMUSG00000039488	Cntn5
ENSMUSG00000020695	Mrc2	ENSMUSG00000039831	Arhgap29
ENSMUSG00000020716	Nf1	ENSMUSG00000039852	Rere
ENSMUSG00000020738	Sumo2	ENSMUSG00000039953	Clstn1
ENSMUSG00000020821	Kif1c	ENSMUSG00000040158	Tax1bp3
ENSMUSG00000020850	Prpf8	ENSMUSG00000040488	Ltbp4
ENSMUSG00000021025	Nfkb1a	ENSMUSG00000040521	Tsfm
ENSMUSG00000021037	Ahsa1	ENSMUSG00000040537	Adam22
ENSMUSG00000021127	Zfp361l	ENSMUSG00000040614	Nlrp9c
ENSMUSG00000021144	Mta1	ENSMUSG00000040687	Madd
ENSMUSG00000021218	Gdi2	ENSMUSG00000040836	Gpr161
ENSMUSG00000021376	Tpmt	ENSMUSG00000041075	Fzd7
ENSMUSG00000021466	Ptch1	ENSMUSG00000041126	H2afv

ENSMUSG00000021518	Ptdss1	ENSMUSG00000041237	Pklr
ENSMUSG00000021576	Pdcd6	ENSMUSG00000041278	Ttc1
ENSMUSG00000021577	Sdha	ENSMUSG00000041355	Ssr2
ENSMUSG00000021614	Vcan	ENSMUSG00000041596	Vmn1r90
ENSMUSG00000021877	Arf4	ENSMUSG00000041881	Ndufa7
ENSMUSG00000021978	Extl3	ENSMUSG00000041939	Mvk
ENSMUSG00000022100	Xpo7	ENSMUSG00000042406	Atf4
ENSMUSG00000022105	Rb1	ENSMUSG00000043079	Synpo
ENSMUSG00000022174	Dad1	ENSMUSG00000045284	Dcaf12l1
ENSMUSG00000022194	Pabpn1	ENSMUSG00000045569	Mc2r
ENSMUSG00000022212	Cpne6	ENSMUSG00000046449	C77370
ENSMUSG00000022223	Sdr39u1	ENSMUSG00000046865	Fbl
ENSMUSG00000022285	Ywhaz	ENSMUSG00000046982	Tshz1
ENSMUSG00000022311	Csmc3	ENSMUSG00000047547	Cltb
ENSMUSG00000022433	Csnk1e	ENSMUSG00000047649	Cd3eap
ENSMUSG00000022443	Myh9	ENSMUSG00000047945	Marcks1l
ENSMUSG00000022471	Xrcc6	ENSMUSG00000048076	Arf1
ENSMUSG00000022521	Crebbp	ENSMUSG00000048154	MI12
ENSMUSG00000022561	Gpaa1	ENSMUSG00000048578	Mlec
ENSMUSG00000022570	Tsta3	ENSMUSG00000048930	Tada3
ENSMUSG00000022672	Prkdc	ENSMUSG00000048960	Prex2
ENSMUSG00000022841	Ap2m1	ENSMUSG00000050310	Rictor
ENSMUSG00000022884	Eif4a2	ENSMUSG00000050824	Sstr5
ENSMUSG00000022982	Sod1	ENSMUSG00000050856	Atp5k
ENSMUSG00000023010	Tmbim6	ENSMUSG00000050926	Dcaf12l2
ENSMUSG00000023175	Bsg	ENSMUSG00000051306	Usp42
ENSMUSG00000023903	Mmp25	ENSMUSG00000051747	Ttn
ENSMUSG00000023944	Hsp90ab1	ENSMUSG00000051853	Arf3
ENSMUSG00000024073	Birc6	ENSMUSG00000052026	Slc6a7
ENSMUSG00000024121	Atp6v0c	ENSMUSG00000052397	Ezr
ENSMUSG00000024213	Nudt3	ENSMUSG00000052423	B4galt3
ENSMUSG00000024248	Cox7a2l	ENSMUSG00000052429	Prmt1
ENSMUSG00000024346	Pfdn1	ENSMUSG00000052681	Rap1b
ENSMUSG00000024456	Diap1	ENSMUSG00000052713	Zfp608
ENSMUSG00000024483	Ankhd1	ENSMUSG00000052957	Gas1
ENSMUSG00000024525	Impa2	ENSMUSG00000053141	Ptprt
ENSMUSG00000024527	Afg3l2	ENSMUSG00000053399	Adamts18
ENSMUSG00000024613	Tcof1	ENSMUSG00000053560	Ier2

ENSMUSG00000024639	Gnaq	ENSMUSG00000053565	Eif3k
ENSMUSG00000024646	Cyb5	ENSMUSG00000054102	Nlrp9a
ENSMUSG00000024767	Otub1	ENSMUSG00000054252	Fgfr3
ENSMUSG00000024792	Zfp11	ENSMUSG00000054452	Aes
ENSMUSG00000024824	Rad9	ENSMUSG00000055022	Cntn1
ENSMUSG00000024826	Dpf2	ENSMUSG00000055681	Cope
ENSMUSG00000024830	Rps6kb2	ENSMUSG00000057177	Gsk3a
ENSMUSG00000024851	Pitpnm1	ENSMUSG00000057278	Snrpg
ENSMUSG00000024858	Adrbk1	ENSMUSG00000058454	Dhcr7
ENSMUSG00000024914	Drap1	ENSMUSG00000058558	Rpl5
ENSMUSG00000024927	Rela	ENSMUSG00000058569	Tmed9
ENSMUSG00000024944	Arl2	ENSMUSG00000058655	Eif4b
ENSMUSG00000024962	Vegfb	ENSMUSG00000059070	Rpl18
ENSMUSG00000025130	P4hb	ENSMUSG00000059447	Hadhb
ENSMUSG00000025132	Arhgdia	ENSMUSG00000059518	Znhit1
ENSMUSG00000025231	Sufu	ENSMUSG00000059552	Trp53
ENSMUSG00000025321	Itgb8	ENSMUSG00000059708	Akap17b
ENSMUSG00000025366	Esyt1	ENSMUSG00000060429	Sntb1
ENSMUSG00000025381	Cnpy2	ENSMUSG00000060508	Nlrp9b
ENSMUSG00000025428	Atp5a1	ENSMUSG00000060680	Gm8894
ENSMUSG00000025474	Tubgcp2	ENSMUSG00000060802	B2m
ENSMUSG00000025499	Hras1	ENSMUSG00000061315	Naca
ENSMUSG00000025503	Taldo1	ENSMUSG00000061462	Obscn
ENSMUSG00000025508	Rplp2	ENSMUSG00000061904	Slc25a3
ENSMUSG00000025651	Uqcrc1	ENSMUSG00000062070	Pgk1
ENSMUSG00000025743	Sdc3	ENSMUSG00000062284	Gm6030
ENSMUSG00000025745	Hadha	ENSMUSG00000062825	Actg1
ENSMUSG00000025793	Hgs	ENSMUSG00000062867	Impdh2
ENSMUSG00000025892	Gria4	ENSMUSG00000063229	Ldha
ENSMUSG00000026024	Als2	ENSMUSG00000063239	Grm4
ENSMUSG00000026209	Dnpep	ENSMUSG00000063410	Stk24
ENSMUSG00000026276	Sept2	ENSMUSG00000063457	Rps15
ENSMUSG00000026395	Ptpre	ENSMUSG00000063511	Snrnp70
ENSMUSG00000026450	Chit1	ENSMUSG00000063870	Chd4
ENSMUSG00000026750	Psbm7	ENSMUSG00000063882	Uqcrh
ENSMUSG00000026786	Apbb1ip	ENSMUSG00000064068	Mtx1
ENSMUSG00000026842	Abi1	ENSMUSG00000065954	Tacc1
ENSMUSG00000026864	Hspa5	ENSMUSG00000066306	Numa1

ENSMUSG00000026976	Pax8	ENSMUSG00000067150	Xpo5
ENSMUSG00000027193	Api5	ENSMUSG00000067657	AC124479.2
ENSMUSG00000027223	Mapk8ip1	ENSMUSG00000067713	Prkag1
ENSMUSG00000027224	Duoxa1	ENSMUSG00000068206	Pick1
ENSMUSG00000027247	Arhgap1	ENSMUSG00000068267	Cenpb
ENSMUSG00000027276	Jag1	ENSMUSG00000070319	Eif3g
ENSMUSG00000027367	Stard7	ENSMUSG00000071054	Safb
ENSMUSG00000027404	Snrbp	ENSMUSG00000071076	Jund
ENSMUSG00000027406	Idh3b	ENSMUSG00000071531	Gprn2
ENSMUSG00000027422	Rrbp1	ENSMUSG00000071645	Tut1
ENSMUSG00000027523	Gnas	ENSMUSG00000071650	Ganab
ENSMUSG00000027566	Psma7	ENSMUSG00000073226	Gm10482
ENSMUSG00000027665	Pik3ca	ENSMUSG00000073640	Rpl27-ps3
ENSMUSG00000027852	Nras	ENSMUSG00000073802	Cdkn2b
ENSMUSG00000027937	Jtb	ENSMUSG00000074034	Gm5921
ENSMUSG00000027944	Hax1	ENSMUSG00000074305	C230081A13Rik
ENSMUSG00000027951	Adar	ENSMUSG00000074643	Cpne1
ENSMUSG00000028039	Efna3	ENSMUSG00000075706	Gpx4
ENSMUSG00000028041	Adam15	ENSMUSG00000076432	Ywhaq
ENSMUSG00000028049	Scamp3	ENSMUSG00000078676	Casc3
ENSMUSG00000028140	Mrpl9	ENSMUSG00000079426	Arpc4
ENSMUSG00000028289	Epha7	ENSMUSG00000079523	Tmsb10
ENSMUSG00000028312	Smc2	ENSMUSG00000089986	Ankhd1-Eif4ebp3
ENSMUSG00000028367	Txn1	ENSMUSG00000090247	Bloc1s1

I designed the custom bait library by using Agilent SureDesign. The whole gene regions of these 500 genes, including the both exon and intron, were targeted. The genes which listed in the COSMIC: Cancer Gene Census were highlighted in red.

Supplementary Table S2.2. Details of 83 mutation candidate positions.

position	Normalized VAF						Annotation	
	Papi 9w	Papi 14w	Carc	Meta	Reg 9w	Reg 14w	Effect	GeneID
7:141192550	0.4999	0.5001	0.5001	0.5000	0.5001	0.5001	Missense	Hras1
2:98662257	0.0259	0.0197	0.0611	0.0000	0.0224	0.0268	Synonymous	Gm10801
2:98662457	0.0396	0.0000	0.0442	0.1711	0.0205	0.0453	Missense	Gm10801
9:3000377	0.0368	0.0403	0.0330	0.2766	0.0289	0.0362	Upstream	Gm10722
2:98662399	0.0302	0.0197	0.0696	0.1822	0.0274	0.0472	Missense	Gm10801
9:35305329	0.0222	0.0108	0.0752	0.2000	0.0204	0.0420	Intergenic	-
9:35305343	0.0237	0.0254	0.0342	0.0862	0.0154	0.0257	Intergenic	-
2:98665129	0.0284	0.0220	0.0451	0.1162	0.0191	0.0432	Downstream	Gm10800
X:163478989	0.0369	0.0205	0.0568	0.0992	0.0138	0.0421	Others	Gm7199
2:98662293	0.0199	0.0090	0.0360	0.1926	0.0138	0.0183	Synonymous	Gm10801
2:98666287	0.0213	0.0071	0.0483	0.0940	0.0269	0.0282	Downstream	Gm10800
2:98662411	0.0000	0.0126	0.0592	0.1174	0.0186	0.0257	Missense	Gm10801
2:98662805	0.0000	0.0505	0.0348	0.2408	0.0271	0.0765	Downstream	Gm10800
14:19417883	0.0441	0.0000	0.0000	0.2345	0.0152	0.0213	Intron	AC242409.1
9:3025531	0.0000	0.0000	0.0000	0.2437	0.0277	0.0461	Missense	Gm10716
1:181187046	0.0000	0.0390	0.0000	0.1644	0.0164	0.0000	Upstream	Wdr26
9:3002159	0.0000	0.0391	0.0692	0.0000	0.0177	0.0198	Upstream	Gm11168
9:3002107	0.0000	0.0567	0.0000	0.0000	0.0149	0.0188	Upstream	Gm11168
9:3002141	0.0000	0.0337	0.0000	0.0000	0.0206	0.0211	Upstream	Gm11168
14:65077037	0.0130	0.0040	0.0000	0.0179	0.0042	0.0000	Missense	Extl3
19:8766798	0.0109	0.0000	0.0000	0.0259	0.0036	0.0000	Missense	Nxf1
18:15414888	0.0000	0.3858	0.1519	0.0000	0.0000	0.0000	Intergenic	-
5:10957983	0.0000	0.2231	0.0865	0.0000	0.0000	0.0000	Intron	Gm5152
1:138067914	0.0000	0.3376	0.2823	0.3540	0.0000	0.0000	Nonsense	Ptpcr
15:79144413	0.0000	0.3357	0.3370	0.3548	0.0000	0.0000	Intron	Polr2f
2:76896947	0.0000	0.3243	0.1524	0.1359	0.0000	0.0000	Missense	Ttn
5:84866414	0.0000	0.2879	0.1527	0.1341	0.0000	0.0000	Intergenic	-
14:69159185	0.0000	0.2684	0.1377	0.1117	0.0000	0.0000	Intergenic	-
7:39662651	0.0000	0.2042	0.1498	0.1171	0.0000	0.0000	Intergenic	-
12:30092627	0.0000	0.2944	0.1371	0.1996	0.0000	0.0000	Synonymous	Tpo
8:88710071	0.0000	0.3413	0.1442	0.1922	0.0000	0.0000	Intron	Cyld

10:29781362	0.0000	0.4175	0.1781	0.1993	0.0000	0.0000	Intergenic	-
19:5647309	0.0000	0.3744	0.1837	0.2367	0.0000	0.0000	Missense	Rela
4:43554186	0.0000	0.3647	0.2011	0.1795	0.0000	0.0000	Upstream	Tln1
9:88454294	0.0000	0.3259	0.2016	0.1928	0.0000	0.0000	Intron	Gm20537
2:76807228	0.0000	0.3395	0.1745	0.1402	0.0000	0.0000	Intron	Ttn
2:91178303	0.0000	0.3282	0.1876	0.1712	0.0000	0.0000	Missense	Madd
4:28817318	0.0000	0.3509	0.1860	0.1568	0.0000	0.0000	Intron	Epha7
15:35705682	0.0000	0.0000	0.0213	0.0000	0.0000	0.0000	Intron	Vps13b
17:12247878	0.0000	0.0000	0.0755	0.0000	0.0000	0.0000	Intron	Map3k4
8:43289942	0.0000	0.0000	0.0438	0.0000	0.0000	0.0000	Intergenic	-
15:89160554	0.0000	0.0000	0.0000	0.1282	0.0000	0.0000	Upstream	Plxnb2
11:120346973	0.0000	0.0000	0.0405	0.1253	0.0000	0.0000	Missense	Actg1
11:120346787	0.0000	0.0000	0.0123	0.0570	0.0000	0.0000	Synonymous	Actg1
8:100030643	0.0000	0.0000	0.0153	0.0672	0.0000	0.0000	Others	Gm15210
2:76787076	0.0000	0.0000	0.0000	0.0000	0.3917	0.4303	Synonymous	Ttn
6:125218819	0.0000	0.0000	0.0000	0.0000	0.3908	0.3274	Intron	Vamp1
12:18013571	0.0000	0.0000	0.0000	0.0000	0.3883	0.3083	Intergenic	-
15:98864350	0.0000	0.0000	0.0000	0.0000	0.3936	0.3162	Synonymous	Mll2
16:15833961	0.0000	0.0000	0.0000	0.0000	0.4440	0.3818	Intron	Prkdc
2:76738847	0.0000	0.0000	0.0000	0.0000	0.4521	0.3405	Missense	Ttn
9:62842036	0.0000	0.0068	0.0000	0.0000	0.0046	0.0000	Upstream	Cln6
7:17139272	0.0000	0.0000	0.0000	0.0000	0.0357	0.0000	Intergenic	-
11:55004766	0.0000	0.0000	0.0000	0.0000	0.0083	0.0000	Intron	Anxa6
11:69588515	0.0000	0.0000	0.0000	0.0000	0.0127	0.0000	Missense	Trp53
13:98853314	0.1033	0.0425	0.0000	0.0000	0.0000	0.0513	Intron	Tnpo1
2:76766295	0.0158	0.0062	0.0000	0.0000	0.0000	0.0229	Missense	Ttn
4:148550842	0.0208	0.0166	0.0000	0.0000	0.0000	0.0502	Downstream/ Upstream	Mtor/ U6
11:79436839	0.0000	0.0119	0.0000	0.0000	0.0000	0.0462	Intron	Nf1
12:21394815	0.0000	0.0572	0.0000	0.0000	0.0000	0.1283	Downstream	Ywhaq
2:108948762	0.0324	0.0000	0.0000	0.0000	0.0000	0.0739	Upstream	Gm13910
6:133762795	0.0939	0.0000	0.0000	0.0000	0.0000	0.0770	Upstream	AC122359.1
11:59067729	0.0000	0.0000	0.0000	0.0000	0.0000	0.0132	Missense	Obscn
3:89141498	0.0000	0.0000	0.0000	0.0000	0.0000	0.0131	Intron	Pklr
4:101161963	0.0000	0.0000	0.0000	0.0000	0.0000	0.0597	Intron	Jak1

7:21230248	0.0000	0.0000	0.0000	0.0000	0.0000	0.0687	Intergenic	-
17:46215562	0.0000	0.0465	0.0000	0.0000	0.0000	0.0000	Intron	Xpo5
2:76861409	0.0000	0.0121	0.0000	0.0000	0.0000	0.0000	Intron	Ttn
7:13030561	0.0000	0.0064	0.0000	0.0000	0.0000	0.0000	Missense	Trim28
18:60828610	0.0000	0.0074	0.0000	0.0639	0.0000	0.0000	Intron	Tcof1
3:99500989	0.0000	0.0101	0.0000	0.0203	0.0000	0.0000	Others	M6pr-ps
15:90723932	0.2446	0.0000	0.0000	0.0000	0.0000	0.0000	Intergenic	-
2:76825807	0.4816	0.0000	0.0000	0.0000	0.0000	0.0000	Missense	Ttn
10:79710761	0.4756	0.0000	0.0000	0.0000	0.0000	0.0000	Missense	Bsg
6:107881483	0.4772	0.0000	0.0000	0.0000	0.0000	0.0000	Intergenic	-
2:76846783	0.4504	0.0000	0.0000	0.0000	0.0000	0.0000	Synonymous	Ttn
14:55514768	0.4244	0.0000	0.0000	0.0000	0.0000	0.0000	Downstream	Cpne6/Nrl
4:28966425	0.4303	0.0000	0.0000	0.0000	0.0000	0.0000	3'UTR	Epha7
12:31018401	0.0187	0.0000	0.0000	0.0000	0.0000	0.0000	Intergenic	-
6:122696935	0.0597	0.0000	0.0000	0.0000	0.0000	0.0000	Downstream	Gm10420
X:110604447	0.1181	0.0000	0.0000	0.0000	0.0000	0.0000	Downstream	Gm14886
10:88247930	0.0437	0.0484	0.0000	0.0000	0.0000	0.0000	Downstream	Ccdc53
8:33615918	0.0226	0.0154	0.0000	0.0000	0.0000	0.0000	Intron	Ppp2cb

The normalized VAFs of each tumor sample, the effect of the mutation, and the genes including the position are shown. The order corresponds to Fig. 2.1b.