Identification of Neo-antigens for a Cancer Vaccine

by Transcriptome Analysis

by

HoJoon Lee

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved February 2012 by the
Graduate Supervisory Committee:

Stephen Albert Johnston, Chair
Sudhir Kumar
Laurence Miller
Phillip Stafford
Kathryn Sykes

ARIZONA STATE UNIVERSITY

May 2012

ABSTRACT

We propose a novel solution to prevent cancer by developing a prophylactic cancer. Several sources of antigens for cancer vaccines have been published. Among these, antigens that contain a frame-shift (FS) peptide or viral peptide are quite attractive for a variety of reasons. FS sequences, from either mistake in RNA processing or in genomic DNA, may lead to generation of neo-peptides that are foreign to the immune system. Viral peptides presumably would originate from exogenous but integrated viral nucleic acid sequences. Both are non-self, therefore lessen concerns about development of autoimmunity. I have developed a bioinformatical approach to identify these aberrant transcripts in the cancer transcriptome. Their suitability for use in a vaccine is evaluated by establishing their frequencies and predicting possible epitopes along with their population coverage according to the prevalence of major histocompatibility complex (MHC) types.

Viral transcripts and transcripts with FS mutations from gene fusion, insertion/deletion at coding microsatellite DNA, and alternative splicing were identified in NCBI Expressed Sequence Tag (EST) database. 48 FS chimeric transcripts were validated in 50 breast cell lines and 68 primary breast tumor samples with their frequencies from 4% to 98% by RT-PCR and sequencing confirmation. These 48 FS peptides, if translated and presented, could be used to protect more than 90% of the population in Northern America based on the prediction of epitopes derived from them. Furthermore, we synthesized 150 peptides that correspond to FS and viral peptides that we predicted would exist in

tumor patients and we tested over 200 different cancer patient sera. We found a number of serological reactive peptide sequences in cancer patients that had little to no reactivity in healthy controls; strong support for the strength of our bioinformatic approach.

This study describes a process used to identify aberrant transcripts that lead to a new source of antigens that can be tested and used in a prophylactic cancer vaccine. The vast amount of transcriptome data of various cancers from the Cancer Genome Atlas (TCGA) project will enhance our ability to further select better cancer antigen candidates.

ACKNOWLEDGMENTS

I would not have been able to finish my thesis work without the help of many people. My advisor, Dr. Stephen Albert Johnston, has supported me in all respects of my work with his great enthusiasm in the development of cancer vaccines. I would also like to acknowledge my committee members: Dr. Sudhir Kumar, Dr. Laurence Miller, Dr. Phillip Stafford, and Dr. Kathryn Sykes. I really appreciate the advice and time they have given me. I also want to thank Dr. Laurence Miller for obtaining all of the primary tumor samples from the Mayo Clinic. In addition, it has been a privilege to work with the cancer project team in the Center for Innovations in Medicine (CIM) of the Biodesign Institute at ASU: Dr. Tricia Carrigan, Luhui Shen, Kristen Seifert, John-Charles Rodenberry, Felicia Craciunescu, Kari Kotlarczyk, Dr. Jose Cano Buendia, Dr. Jean Chapuis, Daniel Johnson, Dr. Douglas Lake, and Dr. Christine Kuslich. I have also received a great deal of support from the administrators in CIM: Pattie Madjidi, Penny Gwynne, Preston Hunter, Kevin Brown; and talented high school interns: Carrie Lin, Samuel Hooke, Nick Giancola. Last but not least, I thank my friends that have not only been kind enough to help proof read my thesis, but also gave me critical insight in my work; Samantha Chan, Stephanie Touchman, Lucas Restrepo, Kurt Whittemore, Muskan Kukreja and Helen Schwerdt.

TABLE OF CONTENTS

Page

CHAPTER Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

CHAPTER 1

INTRODUCTION

Cancer is one of the leading causes of death in the United States and many other countries. Currently, one in four people will die of cancer in the United States. In total, 1,596,670 new incidents and 571,950 deaths from cancer are projected to occur in the United States in 2011. The chance of being diagnosed with an malignant cancer in a lifetime is 44% and 38% for men and women respectively (1). Cancer is a major threat to public health and is in desperate need for a cure.

1.1 Why a 'Prophylactic' Cancer Vaccine?

One common treatment for cancer is surgery, whose effectiveness is related to how early the cancer is detected. Chemo-therapy may be associated with considerable amount of side-effects. One potential approach is to cure cancer by the development of cancer vaccines. Vaccination is one of the most effective ways to treat infectious diseases in the history of medicine as 26 infectious diseases are preventable through vaccination (2). Vaccination against cancer has multiple advantages over existing treatments, including tumor specificity through personalization, minimal toxicity, and long-term therapeutic effect due to immunological memory (3). Data has been accumulated from human and mouse studies that provide strong evidence that the immune system is involved in tumor rejection (4-10). This suggests a possibility that the immune system could be trained against the tumor.

Over the past two decades, there have been considerable efforts to turn the patient's immune system against pre-existing tumor. These attempts encompass the use of whole cells, peptides, genetically modified tumor cells, heat-shock proteins, or apoptotic tumor cells to elicit the host's immune response to cancer cells (11, 12). However, therapeutic cancer vaccination has not proved strong enough to eradicate malignancies consisting millions of tumor cells. Total tumor burden, immune suppression induced by tumors, immune escaping are all hurdles for a therapeutic vaccines to work (2). Therapeutic tumor vaccines have been extensively examined in animal models and in clinical trials. However, these approaches have not been successful in clinical settings (13). One of the main problems in a therapeutic approach to cancer vaccines include lack of high affinity response, autoimmunity, otherwise, immune tolerance (and even immune escape) of tumor by cancer immunoediting (14), which is largely due to priming the self-antigens.

In the case of prophylactic vaccine, we can avoid the insurmountable obstacles that affect therapeutic cancer vaccines. An enhanced immune system, exposed to a tumor antigen by vaccination, is expected to kill the tumor before it reaches a stage of cancer in which it will suppress and evade the immune system. The cancer vaccines are most effective in protection from tumor challenge based on animal studies (15). Autoimmunity will be a significant concern. Indeed, it makes the development of a cancer vaccine more difficult that tumors largely express 'self' antigens. Nevertheless, several tumor antigens have been identified and cancer vaccines against these antigens have been reported in pre-clinical

studies to induce tumor-specific immune responses and result in long-term memory without autoimmunity (7, 16-20).

Non-self antigens have been shown to generate high-avidity T cell responses more readily than self antigens (21, 22). **The discovery of tumor-specific neo-antigens is crucial for vaccine development** in order to develop effective cancer vaccines. The success of vaccine against infectious diseases also indeed comes from the fact that the causative agents of most infectious diseases have been already reported and isolated. Tumor-specific antigens decrease the risk of autoimmunity and at least systemic tolerance, which is especially critical in prophylactic vaccines. **Tumor-specific neo-antigens in tumors enable us to develop effective prophylactic cancer vaccines as well as possible therapeutic ones.**

1.2 What is the optimal tumor antigen?

Selecting and determining the appropriate antigens that elicit a specific antitumor immune response is one of critical challenges of developing a cancer vaccine. What are tumor antigens? And what makes antigens ideal for vaccine? “Virtually any mutant, over-expressed or abnormally expressed protein in cancer cells, can serve as a target for cancer vaccines and/or T-cell therapy” (23). Basically, any protein that can distinguish tumors from normal have the potential to be tumor antigens. Functionally, tumor antigens may be classified as self or non-self (neo-antigens) (11). Self tumor antigens are derived from non-mutated genes that meet one of following conditions (23); i) expressed limited to only

fetus (oncofetal) or dispensable normal tissues such as prostate or ovary (differentiation), ii) expressed higher in cancer (over-expression), and iii) expressed in cancer with unique post-translational modification. Non-self antigens are derived either exogenously or endogenously. Peptides from cancer associated viruses can be exogenous non-self antigens. Endogenous non-self antigens are generated from mutated proteins that arise as a consequence of genetic alterations in tumors. A diagram is show in Figure 1.1. The analysis of Cheever et al. provided the list of 75 representative cancer antigens under investigation. They also suggested nine criteria to evaluate whether they are appropriate antigens; i) therapeutic function, ii) immunogenicity, iii) oncogenecity, iv) specificity, v) expression level and % positive cells, vi) stem cell expression, vii) number of patients with antigens-positive cancers, viii) number of epitopes and ix) cellular location of expression.

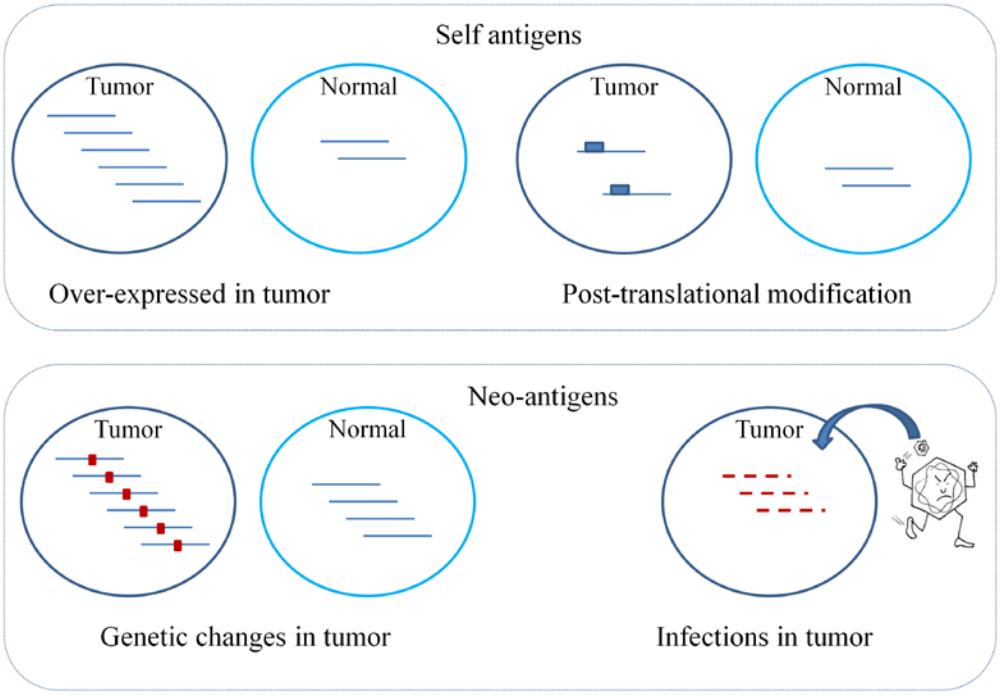


Figure 1.1 Types of tumor antigens There are two types of antigens by their origin. Non-mutated gene can generate self antigens when they are deregulated in cancer. Mutations or infections can generate neo-antigens.

1.3 Why neo-antigens?

Self antigens have two main limitations: autoimmunity and immune tolerance. Use of self antigens in a vaccine may lead to autoimmune toxicities (24). The risk of autoimmune reactions after vaccination has been observed in animal models (25-27), as well as in clinical trials where melanoma patients who have developed vitiligo (loss of pigmentation due to destruction of melanocytes) (28, 29). In addition, self-antigens run the risk of being non-immunogenic thus incapable of breaking immune tolerance (30, 31). “Unlike self-antigens, neo-antigens (non-self antigens) can avoid the risk of autoimmunity and at least systemic tolerance”(23). **Therefore, tumor antigens uniquely represented in**

the tumor and not in normal tissues may be better candidates for a prophylactic vaccine in general.

According to Cheever et al. (23), the majority of the examined antigens are self-antigens while only 11 are non-self antigens (Figure 1.2). In fact, none of them are derived from frame-shifted mutations. The potential of neo-antigens as cancer vaccine antigens, especially frame-shifted, has not been examined intensively. The suggested list of neo-antigen candidates provided by this study would be a good start for testing the potential of neo-antigens as cancer vaccine antigens.

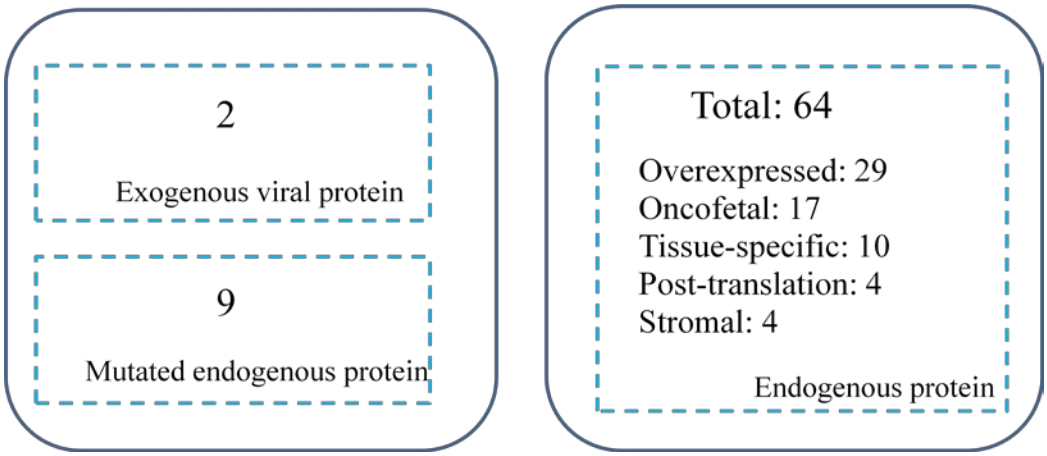


Figure 1.2 Classification of 75 antigens from Cheever et al. This diagram shows the uneven distribution between self and non-self antigens under current investigation.

1.4 Source of neo-antigens

Neo-antigens can originate either exogenously (such as viral proteins from HPV16) or endogenously. The latter ones include un-mutated proteins that have never been exposed to an immune system (such as embryonic antigens) as well as mutated proteins created by genetic changes in tumors (11). The major types of

genetic changes in tumors are: subtle sequence changes in nucleic acids, alterations in chromosome number, rearrangements of chromosome, and gene amplifications/deletions (32). Among these changes, we consider the framesift (FS) as a powerful source of tumor-specific antigens because unique sequences from FS are more likely than point mutations to contain longer sequences and parts of these neo-sequences will be presented on the surface of cancer cells in the context of MHC class I molecules (33). **We propose that neo-antigens derived either from viruses (exogenous) or frame-shifted mutations (endogenous) are ideal antigen candidates.** The immune system should react more strongly against tumors presenting peptides from a viral origin rather than to tumors presenting endogenous non-mutated peptides, which have been exposed to negative selection. Since most cancers are the result of accumulated genetic alterations rather than viral infections, endogenous antigens may still be the predominant tumor targets. Some of these antigens will be neo-antigens arising from mutations. Specifically, alteration of the reading frame caused by genetic changes, frame-shift mutations, may generate 'immunogenic' C-terminally truncated proteins with a neo-peptide tail that stretches beyond the mutation until a stop codon is encountered. The use of FS peptides as tumor antigens was first suggested by Townsend *et al* in 1994 (33). Since then, several studies have shown the potential of FS peptides as novel antigens for cancer treatments by inducing tumor-specific cell-mediated immunity (34-39). There are three types of mutations that induce frame-shifted peptides: chimeric transcripts, insertion/deletions, splicing variants. First, the chimeric transcripts are potential

source of immunogenic tumor-specific antigens derived from new antigenic peptides at a junction or breakpoint: new combination of two peptides or induced frame-shifted peptides. *Imatinib* targets the bcr-abl fusion gene. It has shown a remarkable success in cancer treatment. Furthermore, several studies have shown that fusion peptides can elicit HLA-restricted CTL reactions to lyse tumor cells (40-42). Second, the coding microsatellites (MS) DNAs in genes can also contribute immunogenic tumor-specific antigens by FS mutation due to their propensity for insertion-deletion (Indels) mutations with high mutation rate (43-45). Several initial studies have shown the frequent insertion/deletion at the coding MS DNA of *TGFβ-RII, BAX*, *hMSH3, hMSH6*, and *IGFIIR* genes in the microsatellite instability (MSI) colorectal cancer (15% of colorectal carcinoma) or hereditary nonpolyposis colon cancer (HNPCC). According to Duval and Hamelin (46), the mutations frequency of coding MS DNA for MSI colorectal cancer and HNPCC were 81% and 76% respectively for *TGFβ-RII*, 45% and 49% for *BAX*, 38% and 51% for *hMSH3*, 22% and 24% for *hMSH6*, and 17% and 7% for *IGFIIR*. Currently, about 400 genes with coding repeats were surveyed in the database called "SelTarbase" (47). Besides, there are more than 7,000 unexplored MS DNAs from in the coding sequences of 4,000 genes. Third, alternative splicing variants are also good source of generating FS peptides by skipping exons of which the length is not divisible by three. Recent studies showed that alternative spicing is much more frequent than expected. According to Wang *et al*., 92-94% of human genes have splicing variants, which has been proposal as a major contributor to human phenotype variability given our relative small genome

(48). Furthermore, several studies showed that some splicing event significantly differed in tumors relative to corresponding normal tissues (49, 50).

1.5 Strategies for identifying neo-antigens

The identification of neo-antigens is not an easy task in most cases of sporadic tumors raised from spontaneous genetic alterations since tumor cells are transformed from normal cells. Neo-antigens can be identified by screening immune response of cancer patients. This approach would be lengthy and expensive. We are able to obtain putative neo-antigens by analyzing the transcriptome or proteome in cancer and normal samples. Neo-antigens from this approach have to be confirmed by experimental validation. The advent of high-throughput sequencing technology provides access to massive transcriptome data from various sources. However, high-throughput proteome analysis has not yet been well established. Therefore, we started to analyze caner and normal transcriptomes to identify putative neo-antigens. Analysis of the EST database (51) provides us with the opportunity to define novel tumor-specific changes and their patterns. Several studies have demonstrated that EST analysis facilitated the identification of relevant mutations in tumors, including chimeric transcripts, and mutation pattern (52-57). The same principle can be applied to the recent RNA-seq data (58-62) from tumor and normal tissues. I proposed a study illustrated in Figure 1.3.

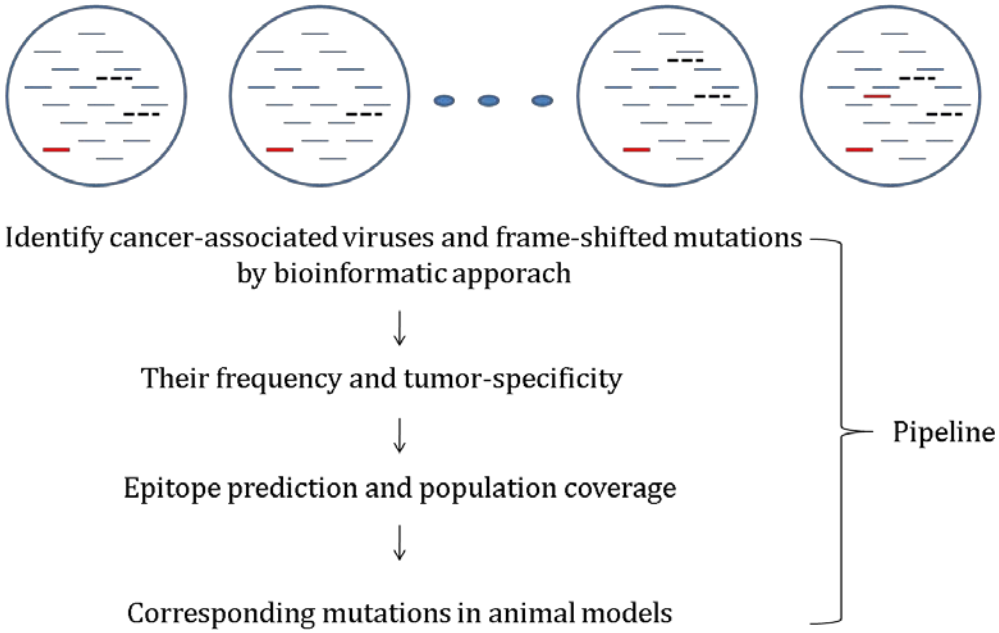


Figure 1.3 Goal of study. We set the pipeline of transcriptome analysis for identifying putative neo-antigens on the behalf of cancer vaccine development. Red bar indicates a frame-shift mutation while dotted line indicates a viral sequence.

In summary, identified frame-shifted and viral peptides from the transcriptome analysis are promising as antigens for inclusion in cancer vaccines, but little is known about their frequency let alone their efficacy for the cancer vaccine. **The systemic screening of these alterations in all different types of tumors is required to establish their frequency (in one type of tumor) and prevalence (among different types of tumors).** This information is crucial in accessing the feasibility and direction of cancer vaccine development as well as in choosing the right antigens. To address this important question, we propose i) to develop a methodological strategy for establish the frequency and prevalence of mutations in human cancers and ii) to identify potential effective tumor-specific antigens for a cancer vaccine. The analyses described in this document such as their frequency, tumor-specificity, and population coverage according epitope

prediction enable us to suggest novel effective neo-antigens for cancer vaccine development. According to the findings from this study, we will be able to impact the direction of cancer vaccine development. Also, the list of tumor-specific genetic alterations derived from this study will give us better a understanding of tumor biology as well as many other applications such as cancer biomarker, diagnosis, prognosis, microarray probes and so on.

CHAPTER 2

VIRAL SEQUENCES

2.1 Introduction

Viruses associated with the development of cancer could be the most obvious and useful tumor-specific markers as they do not originate endogenously. The idea that viruses could be associated with cancer has been around for nearly 100 years, since 1911 when Peyton Rous isolated an avian virus from chicken sarcoma (63, 64). Several infectious agents, especially viruses, are considered to be oncogenic in humans as shown by Javier and Butel and Martin and Gutkind (63, 65). This known list includes human papilloma virus (HPV), hepatitis B virus (HBV), hepatitis C virus (HCV), Epstein-Barr virus (EBV), Human T-lymphotropic virus (HTLV-1), and Kaposi's sarcoma-associated herpesvirus (KSHV). In addition, several studies suggested that Merkel cell polyomavirus, human immunodeficiency virus (HIV), Molluscum contagiosum virus (MCV) and simian 40 (SV40) may have a potential link to development of human cancers. According to Parkin’s study, about 17.8% of global cancer incidents associated with infection in 2002 (66) and the presence of HPV type 18 has been reported to be as high as 20% in cervical cancer (67).

The identification of cancer-associated viruses has two important implications. First, tumor-associated viruses advance our understanding of tumor instigation and development (68). The discovery of oncogenes and tumor suppressors was derived from the study of RNA tumor viruses and DNA tumor

viruses respectively. The function of the *src* gene was discovered and recognized as an oncogene during debates on their cellular or viral origin. The famous tumor suppressor gene, *p53*, was discovered during the study of SV40 large T antigens. Many other molecular mechanisms of cancer were revealed from research of cancer-associated viruses. Second, identification of tumor-associated viruses enabled us to develop vaccines against them which in turn lowers the risk for cancer (69). The HPV and hepatitis vaccines are expected to lower the risk for cervical cancer and hepatocellular carcinoma respectively. Chang reported that the incidents of hepatocellular carcinomas (HCC) dropped from 10-17% to 0.7-1.7% after the HBV vaccination program started in Taiwan twenty years ago (70).

Hausen, who discovered the role of papilloma virus in cervical cancer, promoted the search for additional viruses with a link to malignancy. *In toto*, the list of cancer-associated viruses is quite short compared to their contribution to cancer worldwide. However, the discovery of new causative viruses with diseases including cancers has been a very arduous task (71). The availability of sequences from tumors provides us the opportunity to detect the viral sequences in human cDNA libraries from various sources. Several studies have detected viral sequences in the human transcriptome (72-75). The purpose of our study is to identify viral sequences in existing database in order to provide a list of putative cancer-associated viruses. We have taken a bioinformatic approach to determine the presence of viral sequences in expressed sequence tag (EST) databases from NCBI. We show that some viruses sequences were more prevalent in tumors than normal tissues. Furthermore, we examined the abundance of open reading frames

(ORFs) of viruses in order to show differential expression among viral peptides. This study showed that viral sequences can be reliably detected amidst the abundance of human transcriptome sequences. The suggested list of virus candidates provided by this study would be a good start for an immunological study.

2.2 Bioinformatic analysis

2.2.1 Data Sets

Five different data sets were obtained from the National Center for Biotechnology Information (NCBI); Expressed Sequence Tag (EST) (51), human reference sequences (RefSeqs), set of complete viral genomes, Univec database, non-redundant (nr) nucleotide database. About 8.3 M sequences from 49 different tissues types of tumor and normal had been deposited into EST database of NCBI. 4,004,495 sequences in 2,729 libraries were obtained from normal samples while 3,252,458 sequences in 4,992 libraries were obtained from tumor samples. 3,873 complete viral genome sequences were retrieved from NCBI by querying “viruses [Organism] AND reference sequences”. Human mammary tumor virus (AF243039) was not identified by this query, I manually added it. For identification of ESTs derived from human transcripts, I made use of human Reference Sequences (RefSeq) for identifying human transcripts (76). To eliminate vector sequences in EST database, UniVec database was employed (ftp://ftp.ncbi.nih.gov/pub/UniVec/). UniVec database contains vector sequences as well as sequences of adapters, linkers, and primers for cloning. Last, I used

non-redundant nucleotide database other than human and viruses in order to ensure that possible alignments from other organisms were removed (ftp://ftp.ncbi.nih.gov/blast/db/FASTA/).

2.2.2 Identification of EST sequences derived from viruses

Basically, I used stand-alone BLAST program to identify putative viral sequences in the EST sequences. The EST sequences that aligned with human RefSeq was filtered out. Good alignment implied that length of alignment was ≥100 bp with minimally 85 % sequence similarity or aligned length ranged between 50 to 100bp with ≥90% sequence similarity. Not all EST sequences aligned with human RefSeq using the above filter; these were considered as not originated from human transcripts. These were considered as contaminants or exogenous origin. We used them for further analysis: I aligned these excluded EST sequences with the complete viral genomes. To ensure we captured short alignments, we adjusted our criteria to >50 bp, >90% similarity or >35 bp, >97% similarity. Sequences that matched viral sequences were further filtered by content from the UniVec database (vector sequences). BLAST scores of viral was compared to vector sequences. Those viral ESTs that scored at least 50 or more over vector sequences were kept. To further reduce false positives, we aligned the remaining EST sequences against non-redundant nucleotide sequences excluding viral and human sequences. BLAST scores against viral sequences had to be at least 50 or greater than scores from non-redundant nucleotide sequences. Finally, those ESTs that remained after two filtering steps were presumably of viral origin.

2.2.3 Frequency analysis of virus and their open reading frame (ORF)

Using putative viral EST sequences, we estimated the frequency of their incident in tumor and normal samples. Some sequences traced back to a single viral sequence while others were not resolved clearly due to shared sequence similarity across multiple viral sequences. When scores from the best alignment were higher than scores from any other by at least 50 or more, this EST sequence was considered as a single origin. I used mainly viral EST sequences with unique origin to count the frequencies of corresponding viruses. The same principle was applied to counting the prevalence of ORFs except for highly repetitive viral sequences. In this case, we counted them multiple times for every supporting ORFs.

2.3 Results

2.3.1 Identification of viral EST sequences in EST Db

BLAST program was used to identify putative viral sequences from expressed sequences tag (EST) databases. We identified EST sequences that conservatively would not align with any human reference sequences (RefSeq) or vector sequences (Figure 2.1).

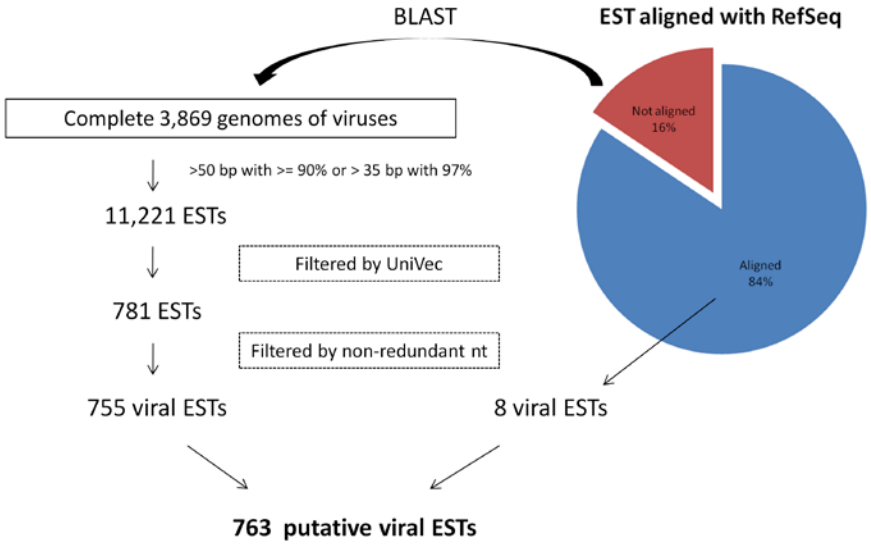


Figure 2.1 The scheme of identification of viral EST sequences. All EST sequences were aligned with human reference RNA sequences (RefSeq) by using BLAST program. The EST sequences that did not aligned with any RefSeq were then aligned with complete viral genomes.

First, all EST sequences (approximately 8.3 M) were aligned to human RefSeq. 16% (1,298,128 sequences) would not align with any RefSeq at all according to our criteria (see Methods). These 'no hit' EST sequences were aligned with known viral sequences from NCBI. There were 11,221 ESTs that aligned with viral sequences according to our criteria (see Methods). These sequences were subject to BLAST analysis against the UniVec database in order to remove vector sequences. 10,440 ESTs were removed. The remaining 781 EST sequences were compared to the non-redundant nucleotide database excluding human and viruses in order to determine whether these sequences originated from sources other than viruses. 26 EST sequences originated from mouse or other

non-viral sources. We obtained 755 putative viral EST sequences. Another set of viral ESTs were detected from removed human sequences that had coincidental viral homology. We blasted these sequences against viral sequences. If this blast score was at least 50 more than human, we regarded these ESTs as viral originated. 8 ESTs were obtained from this process. A total of 763 putative viral ESTs were selected for further analysis.

2.3.2 Identification of putative tumor-associated viruses

Based on the alignments of EST sequences with viral genomes, there are two types of viral EST sequences identified (see Methods). 572 viral EST sequences were evidently traced back to one virus while 183 sequences were not resolved as to their origin as clearly. 22 viruses were supported by 572 viral EST sequences. 15 out 22 viruses were found in at least one tumor library by using viral EST sequences of unique origin (Table 2.1). 6 viruses (squirrel monkey retrovirus, Human papillomavirus type 16, Choristoneura occidentalis granulovirus, Moloney murine leukemia virus, Parainfluenza virus 5, and Mouse mammary tumor virus) were detected only in tumor libraries. Some viruses such as Human papillomavirus 18, Murine type C retrovirus, Enterobacteria phage phiX174 sensu lato were present more in tumor than in normal libraries. Furthermore, we checked seven known cancer-associated viruses and four suspected viruses as proof of concept (Table 2.2). Five were detected by our approach while three (hepatitis C virus, Human T-lymphotropic virus1, and Simian virus 40) were filtered out by one of our criteria. If our filter steps were too strict, we may miss portions of viral sequences. So, I performed the entire

analysis without filtering. Given the higher possibility of false-positives, supplementary Table 1 contains this list of viruses. Beside the 22 viruses, 50 additional were found if the lower stringency filtering is used (supplementary Table A.1).

Table 2.1 The list of putative cancer-associated viruses. From our analysis of EST sequences, 20 viruses were supported by at least one EST sequence from tumor libraries. 6 viruses were found only in tumors. 4 viruses were dominantly found in tumor over normal samples. '?' indicated uncharacterized tissue type. Numbers in ( ) showed the number of EST sequences in each library. For instance, Uterus (1,3,2,1) means that four libraries from uterus and each library has 1,3,2 and 1 supporting viral ESTs respectively.

Virus	Type	# lib (T : N)	Tumor	Normal
Squirrel monkey retrovirus	Retro-transcribing viruses	6 : 0	Breast(1,1,1,1,1,2)	
Human papillomavirus type 16	dsDNA viruses, no RNA stage	2 : 0	Uterus(1,2)	
Choristoneura occidentalis granulovirus	dsDNA viruses, no RNA stage	1 : 0	Colon(1)	
Moloney murine leukemia virus	Retro-transcribing viruses	1 : 0	Bone marrow(1)	
Parainfluenza virus 5	ssRNA negative-strand viruses	1 : 0	Colon(1)	
Mouse mammary tumor virus	Retro-transcribing viruses	1 : 0	Pancreas(1)	
Human papillomavirus 18	dsDNA viruses, no RNA stage	10 : 1	Uterus(1,3,2,1), Lung(1),Cervix(1,4,8,20,1)	Liver(16)
Murine type C retrovirus	Retro-transcribing viruses	9 : 1	Skin(4), Liver(3,9,16), Breast(1,1,1,1),?(1)	Breast(1)
Enterobacteria phage phiX174 sensu lato	ssDNA viruses	22 : 5	Brain(1),Stomach(4),Lung(33,59),Uterus(15,7,74),Thyroid(3,1,3,1),Colon(3,1,2,1,3),Breast(1,2),?(4,13,5,1)	Cerebrum(3),Placenta(5,17),Lung(1),Bone marrow(1)
Human herpesvirus 4	dsDNA viruses, no RNA stage	1 : 1	Lymphoreticular(1)	Pooled tissue(1)
Hepatitis B virus	Retro-transcribing viruses	1 : 1	Liver(3)	Liver(44)
Human herpesvirus 8	dsDNA viruses, no RNA stage	1 : 1	Thyroid(1)	Cerebrum(2)
Shigella phage Sf6	dsDNA viruses, no RNA stage	1 : 2	Prostate(1)	Retina(1,2)
Enterobacteria phage P1	dsDNA viruses, no RNA stage	4 : 9	Colon(1),Cartilage(1),Prostate(1), Lung(1)	Breast(1),Muscle(1),Retina(2), Brain(1,1,1,1,2)

Table 2.2 The list of known cancer-associated viruses This table contains the results for viruses either know to cause or suspect to contribute to human cancers. Five of them were detected in our approach while three viruses were retrieved only after removing some of our filters (see Methods). Three viruses were not present in the whole EST sequences. Overall, 8 out of 11 known cancer associated viruses were detected in EST sequences even though some of them were also found in libraries from normal.

Virus	Type	# lib (T : N)	Tumor	Normal	Comment
Hepatis B virus	Retro-transcribing viruses	1 : 1	Liver(3)	Liver(44)	Detected
Hepatitis C virus*	dsDNA viruses	2 : 0	Uterus(1,2)		Filtered out by human transcripts
Epstein-Barr virus (Human herpesvirus 4)	dsDNA viruses	1 : 1	Lymphoreticular(1)	Pooled tissue(1)	Detected
Human T-lymphotropic virus1	Retro-transcribing viruses	3 : 1	Liver(1^), Testis(1),?(1)	Cerebellum(1)	Filtered out by vector
Kaposi's sarcoma-associated herpesvirus	dsDNA viruses	1 : 1	Thyroid(1)	Cerebrum(2)	Detected
Human papillomavirus type 16	dsDNA viruses	2 : 0	Uterus(1,2)		Detected
Human papillomavirus 18	dsDNA viruses	10 : 1	Uterus(1,3,2,1)-4, Lung(1),Cervix(1,4,8,20,1)	Liver(16)	Detected
Merkel cell polyomavirus	dsDNA viruses	0 : 0			-
Molluscum contagiosum virus	dsDNA viruses	0 : 0			-
Human immunodeficiency virus	Retro-transcribing viruses	0 : 0			-
Simian virus 40	dsDNA viruses	42 : 71	16 tissues	25 tissues	Filtered out by vector

2.3.3 Prevalence of open reading frame (ORF)

There were three viruses with more than 10 supporting viral EST sequences - Human papilloma virus 18 (HPV18), Hepatitis B virus (HBV), and Simian virus 40 (SV40). We enumerate the EST sequences that corresponded to each ORF on a per virus basis. Even distribution was not observed across multiple ORFs in a virus (Figure 2). Among 8 ORFs of HPV18, 5 ORFs had supporting viral EST sequences and most of them originated from only three ORFs; E6, E7, and E1 protein. Viral EST sequences from HBV derived from 3 out 7 ORFs. Most belonged to either X protein or Polymerase. In the case of SV40, viral EST sequences were found in libraries more from normal than from tumor in overall. However, 3 ORFs had more tumor libraries than normal ones among 5 ORF presented in EST sequence database.

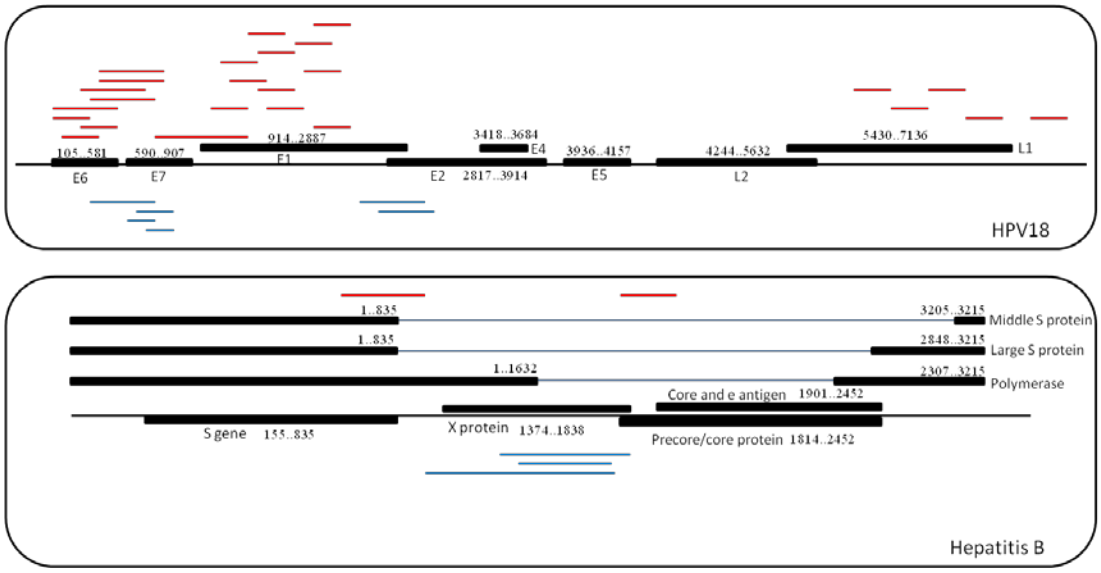


Figure 2.2 The prevalence of open reading frame (ORF) in a virus. The uneven distribution of supporting viral EST sequences among ORFs was observed for most of viruses. For the case of HPV18, 5 out of 8 ORFs had supporting viral ESTs and many of them originated from E6, E7, and E1. In fact, E6 and E7 are known as oncogenic proteins. Among 10 ORFs from Hepatitis virus B, 6 ORFs yielded viral ESTs. 6 viral EST sequences from tumor originated from 6 ORFs while 45 viral EST sequences from normal originated from only 3 ORFs. Viral EST sequences of simian virus 40 were found more in normal than tumor. However, 3 ORFs had more viral ESTs from tumor and normal.

2.3.4 Immune response against viral peptides

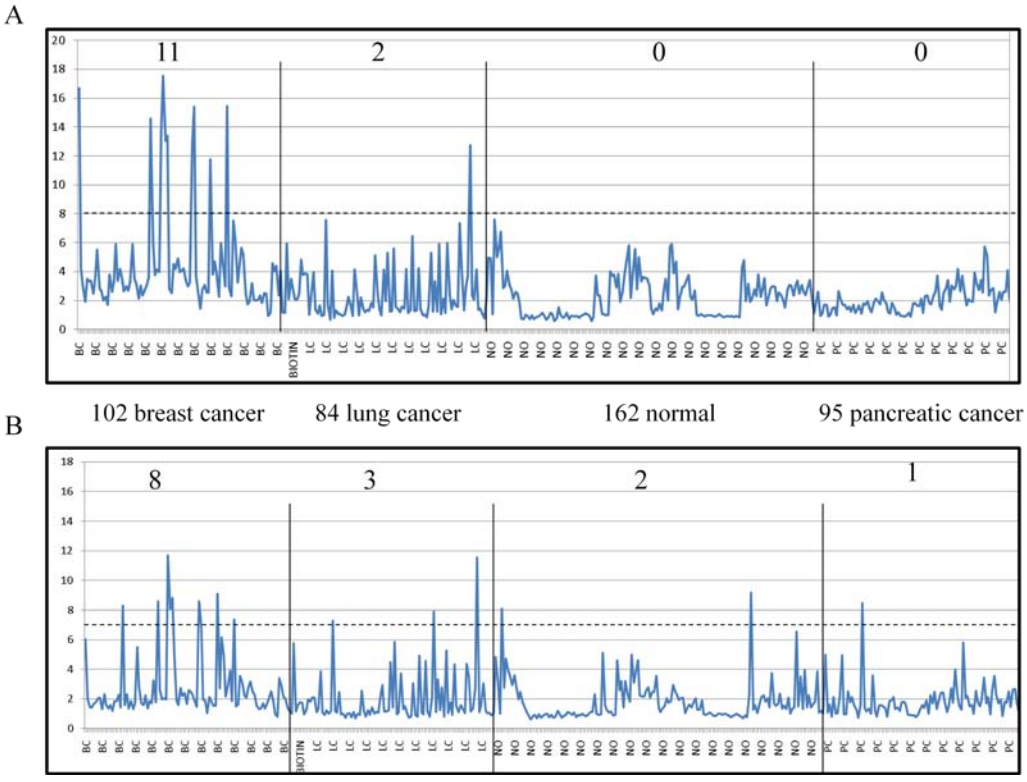
We selected and synthesized 48 predicted B cell epitopes (see supplementary Table A.2) from 30 putative tumor-associated viruses (Table 2.3). Some of peptides were shared by multiple viruses. For instance, APDNDDPNFE is found in Rachiplusia ou MNPV, Plutella xylostella multiple nucleopolyhedrovirus, Bombyx mori NPV, Bombyx mandarina nucleopolyhedrovirus, and Autographa californica nucleopolyhedrovirus.

Table 2.3 The list of viral epitopes on the cancer chip. 48 peptides on peptide array were the putative B cell epitopes derived from 33 distinctive viruses. '# of epitopes' indicates how many epitopes from a virus and 'Viral proteins' indicates their specific origins.

Virus	# epitopes	Viral protein
Human herpesvirus 4 type 2	3	BPLF1 , BALF3 , EBNA-3C
Friend murine leukemia virus	1	gag protein
Human papillomavirus type 16	1	E1
Bovine viral diarrhea virus 1	1	polyprotein
Canine parvovirus	1	polyprotein
Rachiplusia ou MNPV	2	DNA helicase , global transactivator
Parainfluenza virus 5	3	V protein , phosphoprotein , hemagglutinin-neuraminidase protein
Human papillomavirus - 18	2	L1 protein , E1 protein
Squirrel monkey retrovirus	2	protease , gag protein
Plutella xylostella multiple nucleopolyhedrovirus	2	DNA helicase , global transactivator
Pestivirus Giraffe-1	1	polyprotein
Simian virus 40	2	large T antigen , Major capsid protein VP1
Xenotropic MuLV-related virus VP62	2	putative gag-pro-pol polyprotein , putative gag polyprotein
Enterobacteria phage ID18 sensu lato	2	gpB , gpH
Bombyx mori NPV	2	DNA Helicase , GTA
Woolly monkey sarcoma virus	4	Env protein , pre-gag ORF protein , p28sis , hypothetical Gag polyprotein
Beilong virus	2	W protein , nucleocapsid protein
Hepatitis B virus	4	Core and e antigen , precore/core protein , middle S protein , large S protein
Human herpesvirus 4	3	BZLF1 , BPLF1 , BALF3
Moloney murine leukemia virus	2	Pr65 , Pr180
Murine type C retrovirus	1	hypothetical protein MtCrVgp1
Human herpesvirus 5	2	DNA polymerase catalytic subunit , membrane glycoprotein UL18
Bombyx mandarina nucleopolyhedrovirus	2	DNA helicase , GTA
Human adenovirus C	2	single-stranded DNA-binding protein , control protein E4orf6/7
Autographa californica nucleopolyhedrovirus	2	global transactivator-like protein , helicase
Abelson murine leukemia virus	1	p120 Gag-Abl polyprotein
Human herpesvirus 8	2	vIRF-3 , KCP
Rauscher murine leukemia virus	1	gag polyprotein
Moloney murine sarcoma virus	1	Pr65
Spleen focus-forming virus	1	gag polyprotein fragment
Human herpesvirus 1	2	thymidine kinase , DNA replication origin-binding helicase
Murine osteosarcoma virus	1	gag polyprotein
Human immunodeficiency virus 1	2	Vpr , Nef

We used a peptide chip with the 48 spotted peptides to analyze the sera from 443 human samples; 162 normal samples, 102 breast cancer samples, 84 lung cancer samples, and 95 pancreatic cancer samples. After normalization of all intensity in the data to median of 1.0, we set the bar at 7.0 or higher for defining high reactivity. Two peptides showed high reactivity frequently in breast tumor samples relative to normal samples. PYDPEDPGQE was detected in 7 distinctive viruses including Xenotropic MuLV-related virus while PRRRTPSPRRRRSQ

was found only in Hepatitis B virus. Other two peptides showed high reactivity frequently in pancreatic tumor samples relative to normal samples. One of them, TGAESEDSGDEGPSTRH from Kaposi Sarcoma virus, showed also high reactivity frequently in colon tumor samples relative to normal samples (Figure 2.3). Several viral epitopes were highly reactive in many samples from every group including normal samples while 23 viral epitopes did not show any reactivity in any samples.



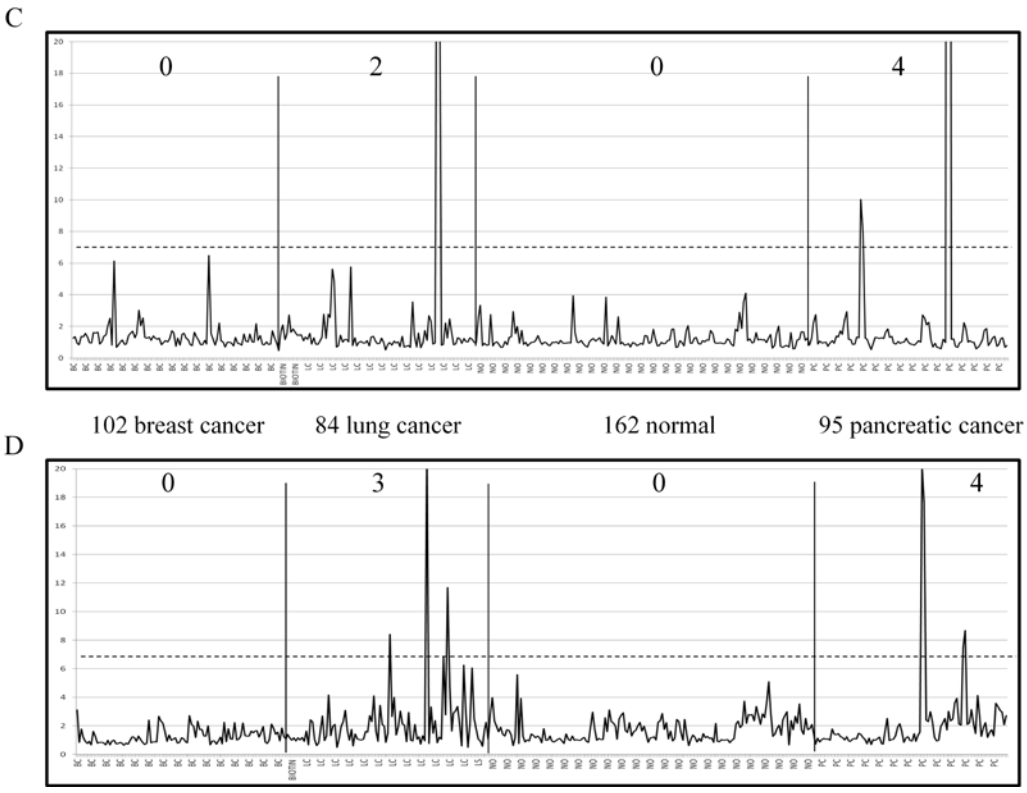


Figure 2.3 High reactive four viral epitopes in cancer samples. Two peptides (A, B) were reactive more frequently in breast cancer samples than normal samples ($p<0.05$, chi-square test). Two peptide (C, D) were reactive more frequently in pancreatic cancer samples than normal samples ($p<0.05$, chi-square test). One peptide (D) was reactive more frequently in lung cancer samples than normal samples ($p<0.05$, chi-square test). A. PYDPEDPGQE in 7 viruses including XMRV showed high reactivity in 10.8%, 2.4%, 0%, and 0% respectively from breast cancer samples, lung cancer samples, normal samples, and pancreatic cancer samples. The cutoff for high reactivity is 8.0. No difference was observed in lung cancer and pancreatic cancer samples. B. PRRRTPSPRRRRSQ from Hepatitis B virus showed high reactivity in 7.8%, 3.6%, 1.2%, and 1.0% respectively from breast cancer samples, lung cancer samples, normal samples, pancreatic cancer samples. Cutoff for high reactivity is 7.0. C. Numbers on the top means the number of highly reactive samples in each group. C. HPKPPPPLPPSAPSL from Ab-MLV showed high reactivity in 0%, 3.6%, 0%, and 4.2% respectively from breast cancer samples, lung cancer samples, normal samples, and pancreatic cancer samples. Cutoff for high reactivity is 7.0. No difference was observed in lung cancer and pancreatic cancer samples. D. TGAESEDSGDEGPSTRH from HHV-8 (or Kaposi Sarcoma virus) showed high reactivity in 0%, 3.6%, 1.2%, and 4.2% respectively from breast cancer samples, lung cancer samples, normal samples, pancreatic cancer samples. Cutoff for high reactivity is 7.0. Numbers on the top means the number of highly reactive samples in each group. X-axis is each sample while Y-axis is reactivity.

Acronyms; XMRV: Xenotropic MuLV-related virus, Ab-MLV: Abelson murine leukemia virus, HHV-8: Human herpesvirus 8.

2.4 Methods

2.4.1 Identification of putative tumor-associated viruses

We took a bioinformatic approach to identify putative tumor-associated viruses by using EST sequences (see 2.2 Bioinformatic analysis).

2.4.2 Selection of peptides for array analysis

Basically, we made an effort to select the most immunogenic part of viral proteins from tumor-associated viruses. We used the B cell epitope program, BepiPred, from Immune Epitope Database (IEDB) supported by National Institute of Allergy and Infectious Diseases (NIAID) in order to select putative epitopes from viral proteins. The strict cutoff value, 1.3, allows us to have 0.96 of specificity and 0.13 of sensitivity.

2.4.3 Samples

Center for Innovations in Medicine, Biodesign Institute, Arizona State University has an existing IRB 0912004625. (i) 102 plasma samples from patients with breast cancer. (ii) 84 plasma samples from patients with lung cancer (iii) 95 plasma samples from patients with pancreatic cancer. (iv) 162 plasma samples for control.

2.4.4 Cancer Peptide Array

The cancer chip is 21-up microarray containing 144 peptides that are 20 amino acids long. 48 of them were our viral epitopes. This is customized microarray printed by AMI.

2.5 Discussion

Previous studies showed that we could detect viral transcripts in the sequencing data from infected cells (71-73). By using the approach of transcriptome analysis, our study is capable of obtaining the list of putative cancer-associated viruses that could be targeted by a vaccine. First, we detected viral transcripts in transcriptome data from cancer samples (Table 2.1). Some of them were found in multiple tumor samples, but not in any normal samples. Some of them were found more frequently in tumor samples than normal samples. Second, we observed the antibody reactions against possible epitopes from selected viral proteins by peptide array approach (Fig 2.3). Four of them showed high reactivity more frequently in cancer than in normal samples while several epitopes showed high reactivity in both tumor and normal samples. The presence of viral transcripts and immunogenicity of viral peptides in cancer samples supported the potential of viral peptides as vaccine antigens.

The same assumption from previous studies that infected cells contain nucleic acid of both host and infectious agent was used in this study. In fact, we could found the all viruses detected in Weber *et al* (74). However, our approach collected more information than any other studies. Basically, we used more sequences; recent EST database of more than 8 million sequences (December 2010), which had been dramatically increased over 3-5 years. Larger data set allowed us to contrast their presence between tumor and normal libraries. Therefore, we were able to select several viruses that were more likely to associate with tumors.

A surprising finding was that there was a differential expression among open reading frames (ORFs) in a virus (Figure 2.2) at least in the EST database. In other words, a particular virus with multiple ORF showed that some ORFs expressed more frequently in tumor than normal while some ORF expressed equally in both tumor and normal. Therefore, we could point out specifically what viral open reading frames (ORFs) were highly expressed. This information can guide us to make better selections of antigens. We suggest two possible explanations about different expression level of ORFs from a virus. First, there may be an intrinsic difference in expression level of each ORF. Critical ORF may have higher expression relative other ORFs. Second, the activity of certain ORF might be associated cancer development due to their functions. Therefore, we observe more mRNA of these critical ORFs in cancer than normal.

To have a more precise estimation, we need to have the four numbers; the numbers of cancer patients with/without a viral infection and the numbers of normal samples with/without a viral infection. If the ratio of a virus over non-virus in cancer is higher than that in normal, that virus may be cancer-associated. This approach eliminates the concerns about contaminations in lab because those contaminations, presumably, will happen to both tumor and normal with same chance. Due to low coverage of EST data, the absence of viral transcripts did not guarantee a negative association. In near future, data from next generation sequencing technology will enable us to conduct this research with higher accuracy.

In summary, we have shown that the predicted epitopes from viruses detected in cancer transcriptome had antibody reactions in tumor samples. Considering the fact that it is often hard to prove causative viruses, if any, for cancer, the approach used in this study provides a good starting list of viruses that we can examine for vaccine antigens by using transcriptome data.

2.6 Conclusion

Dr. Hausen who discovered the causation of cervical cancer by human papilloma virus (HPV) argued that it is worthwhile to search for new cancer-associated viruses (71). What other viruses rather than HPV and hepatitis B virus (HBV) can cause or be associated with cancer? Transcriptome data from cancer patients will give us an opportunity to select putative cancer-associated viruses that we can test them for vaccine target. Our approach will be very useful to get a list of putative viruses when large amount of transcriptome data from cancer and normal samples are available. In addition, we can extend the same approach to search for bacterial or other pathogens in tumor sequence databases. This may be a reasonable pursuit as the infection of *Helicobacter pylori*, or *H. pylori*, shows some association with the incident of gastric cancer (77). Some bacteria sequences such as Erwinia amylovora ATCC 49946and Salmonella enterica subsp. enterica serovar Typhimurium str. LT2 were detected in this study.

CHAPTER 3

FRAMESHITED CHIMERIC TRANSCRIPTS

3.1 Introduction

The use of immunization to prevent disease is one of the most remarkable achievements of modern medicine. According to the Center of Disease Control and Prevention (CDC), there are 26 infectious diseases that are now preventable through prophylactic vaccination. The same principles are now being applied for the treatment of cancer. Despite conceptual promise, cancer vaccines have not been entirely successful, unlike vaccine against infectious diseases. The difficulties in preventing cancer by vaccination strategies are hindered by the selection of the appropriate antigens even though remarkable efforts have been made. Recently, Cheever et al. outlined the suggested criteria for selecting the best antigens to be used in therapeutic vaccines. With these newly defined criteria, 75 cancer antigens were prioritized (23). However, almost all of the examined antigens are classified as self-antigens that may lead to post vaccination side effects such as autoimmunity. In addition, self-antigens run the risk of being non-immunogenic or poorly immunogenic, thus incapable of breaking immune tolerance. One way to avoid the possible side effects associated with using self-antigens as vaccine antigens would be to identify and test tumor specific antigens in a prophylactic setting rather than in a therapeutic setting. In a prophylactic setting, the immune system should not be in a suppressed state thus enabling a more robust and sustained a cellular (CD4+/CD8+ T cell mediated response) and

humoral (the B cell-mediated response to produce antibodies) response to the antigens used in cancer vaccines.

Among the many different types of mutations that occur while tumors develop, we are particularly interested in frame-shifted mutations because of their ability to generate neo-peptides. The use of FS peptides as tumor antigens has been previously mentioned and suggested by Townsend et al in 1994 (33). Since then, several reports have continued to support the use of FS peptides as cancer vaccine antigens since they have the ability to induce tumor-specific cell-mediated immunity (37, 39, 78). The use of gene fusions as a source for generating FS peptides for a cancer antigens has not been extensively studied nor is there ample information regarding the frequency in which gene fusions' chimeric transcripts create frame-shift peptides. Nonetheless, several gene fusions have been reported to play a significant role in malignant hematological disorders, Ewing's sarcoma, and most recently have been shown to be useful as diagnostic and therapeutic targets for drugs such as Imatinib (79). Unlike malignant hematological disorders and Ewing's sarcoma, gene fusions are less prevalent in epithelial-based cancers though this could be strongly contributed to the technical limitations of FISH and SKY cytogenetic analyses and the fact that these methods are not applicable as discovery tools. In order to quickly catalog cytogenetic rearrangements, the genomic coordinates for the genes that are involved in the translocation must be known. Recent non-cytogenetic technological advancements are now currently being employed for the discovery of gene fusions in solid tumors. Through the new technological approaches, cancer

genetics are quickly being analyzed by outlier gene expression patterns, massively parallel paired-end sequencing and 454 transcriptome sequencing. These approaches have identified chimeric transcripts in prostate cancer, and adenocarcinomas of the lung and breast (59, 62, 80-84). Collectively gene fusions have been identified to occur within coding sequences (85) and have been shown to generate frame-shifted mutations (62, 80, 85, 86). Of note, confirmatory changes within the genomics regions that correspond to the newly rearranged fused genes are often not mentioned nor investigated.

With the increasing discovery of chimeric transcripts that results in the fusion of coding sequences from an upstream gene and part of an exon/intron from a downstream gene, it is becoming clear that these transcripts are abundantly present in cancer cells though their role, their importance and whether or not the transcripts get translated has not yet been determined. In this report we have taken a systematic approach to provide a comprehensive analysis of the presence of chimeric transcripts that are relevant to breast cancer. To quickly identify chimeric transcripts that may result in a frameshift neo-peptide we have written and tested an algorithm that is capable of nominating chimeric transcripts from publically available sequence databases and high-throughput sequencing data sets. In addition, we have determined the frequency and have predicted the potential epitope coverage these chimeric transcripts may present. Moreover, we have analyzed the various chimeric transcripts to determine if there are patterns of expression that reflect the stage of tumor differentiation.

3.2 Algorithm to identify FS chimeric transcripts

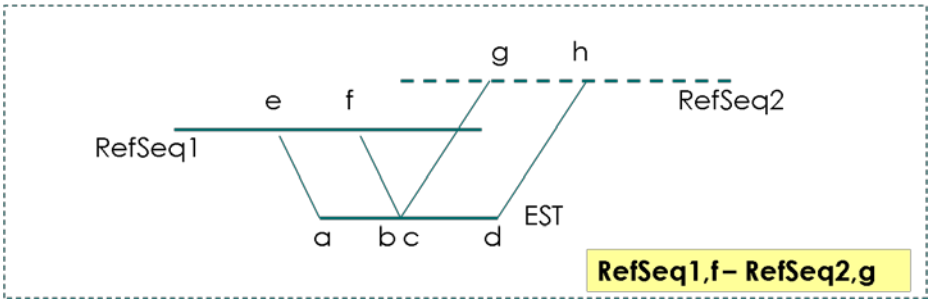


Figure 3.1 Identification of EST derived from chimeric transcripts. Required conditions; i) (f-e) and (h-g) > 80, ii) |c-b| < 10, 3)

Using the stand-alone BLAST program, all EST sequences were aligned to RefSeq. We picked ESTs that aligned with more than 50-85 base pairs and had 95-97% homology to RefSeqs that had been previously annotated by National Center Institute (NCI). We further filtered out our alignment data by eliminating the EST sequences that did not align to multiple RefSeqs or were aligned in the 3'-5' orientation. Lastly, we also eliminated the sequences that aligned with non-coding sequence regions. The remaining EST sequences were then used to identify the chimeric transcripts. Only the ESTs that aligned to two or more distinct RefSeq in consecutive positions were considered to be potential candidates. To be defined as a coding chimeric transcript, the EST sequences had to be at least 100-170 bp long with sequence similarity greater than or equal to 95%- 97% to the RefSeq. Also, the junction point between the two genes had to occur within the coding sequence of the upstream gene and orientation of the upstream gene alignment had to be in the positive (5'-3') orientation. To eliminate false calls, all potential chimeric EST sequences had to be either present in more than one cDNA library or supported by three or more independent EST

sequences. In addition, chimeric transcripts were classified based on the relative position of two genes. Classification of types of chimeric transcript was based on relative position of two fusion genes on the chromosome. Specifically genes found on different chromosomes resulted in inter-chromosomal fusion while genes found in same chromosome were intra-chromosomal or read-through chimeric transcripts. Read-through chimeric transcripts resulted from two neighboring genes on same strand, otherwise intra-chromosomal.

3.3 Results from EST analysis

3.3.1 Putative FS chimeric transcripts

We used our semi-automatic alignment algorithm to identify frame-shifted chimeric transcripts from the available NCBI EST sequence database (Figure 1). Briefly, to support a chimeric transcript, one EST sequence must be able to align to two distinct RefSeqs continuously. Considering the EST database contains approximately 8M EST sequences, we outlined filtering criteria that were applied to eliminate irrelevant sequences. We discarded the EST sequences that did not align properly with annotated RefSeqs and ones that were from untraceable sources. The remaining 7M sequences were then examined for their ability to align with multiple RefSeqs. From this survey, there were 556,989 EST sequences that aligned with multiple RefSeqs. These 556,989 EST sequences supported 2,394 EST chimeric transcripts from tumor and 2,944 EST chimeric transcripts from normal cDNA libraries while 104 EST chimeric transcripts were found in both tumor and normal. Collectively, these supporting EST sequences potentially

represent 5,234 non-redundant putative EST chimeric transcripts that aligned and created a continuous sequence that was composed of two different RefSeqs. Further analysis revealed that 1,133 out of 5,234 EST chimeric transcripts were a product of the reverse strand of the upstream gene combined with the forward strand of the downstream gene. Since this combination is not likely to occur naturally, we excluded these sequences from our analysis. The remaining 4,101 EST chimeric transcripts candidates were then analyzed for the presence of a functional transcriptional coding sequence in the upstream gene. This step removed 1,693 EST chimeric transcripts. Last, we selected putative candidates out of the remaining 2,408 EST chimeric transcripts according to one of the following three criteria; i) the supporting EST sequences were found in two or more independent cDNA libraries, ii) the supporting ESTs were present in multiple copies within one library, or iii) the junction point within the newly identified EST chimeric transcript occurred exactly at the exon boundaries for both genes involved in the combination. Based on these criteria 170 EST chimeric transcripts were supported by two or more representative EST sequences found within multiple libraries, 22 EST chimeric transcripts were supported by three or more EST sequences within one library, and 304 EST chimeric transcripts were joined exactly at the exon boundaries for the two unique fused genes. The selected 496 candidates were then examined for the potential to generate frame-shifted neo-peptides. 321 out of 496 chimeric transcripts from this analysis, if translated, would create a frame-shift peptide while the remaining 175 chimeric transcripts stay in frame.

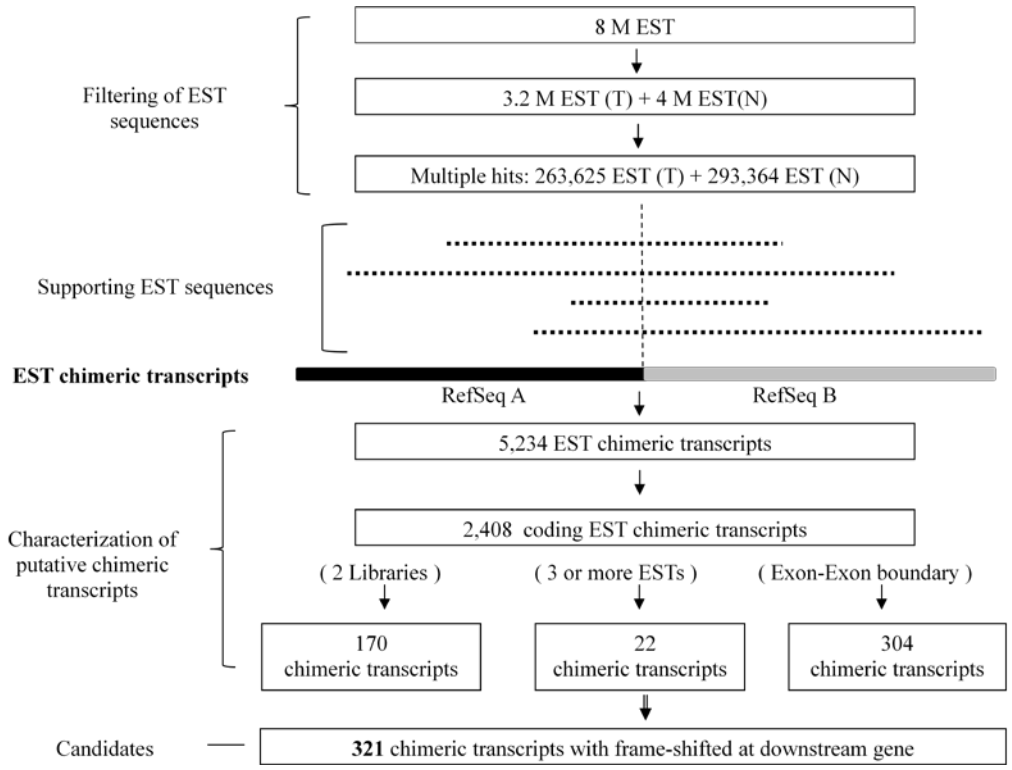


Figure 3.2 Identification of frame-shifted coding gene fusions from EST sequences. Schematic above show the overall selection criteria used to identify frame-shifted EST chimeric transcripts in NCBI EST database. After filtering irrelevant ESTs, EST chimeric transcripts were defined by supporting EST sequences. Supporting ESTs of chimeric transcripts were identified by alignment condition; 85 bp or longer with 95% or more similarity or 50 bp or longer with 97% or more similarity. From this analysis, 321 putative candidates were identified and were predicted to generate a frame-shifted peptide. T indicates tumor and N indicates normal.

3.3.2 Experimental validation in breast cancer

Based on the informatic predictions, 321 out of 496 putative candidates, if translated, would generate frameshift peptides. For 230 out of 321 putative candidates, a neo peptide of 6 or longer amino acids would be generated thus the longer the peptide the more possible epitopes can be present. Additional 13 short FS peptides (from 1 a.a to 5 a.a) were added into the screening list because they were strongly supported by multiple numbers of EST sequences or libraries. 10

candidates out of 243 were removed since the overall length of the transcript was too short to design appropriate primers. To validate the presence of these predicted chimeric transcripts in breast tumors, we screened 50 breast cancer cell lines (see supplementary table) by RT-PCR using 233 different primer pairs. The initial validation was performed with four pools of 10-12 cDNAs that encompass fifty different breast cancer cell lines using standard PCR conditions in order to increase the chances of confirming the predicted candidates. The summary for all 233 PCR reactions is shown in Figure 1B. For 84 primer pairs, no products were amplified though this does not necessary mean that the chimeric transcripts do not exist, rather these transcripts might not be present in Breast Cancer cell lines since the initial informatic analysis utilized sequences from 40 different tissue types. For forty-nine primer pairs a single PCR product that corresponded to the expected size was amplified of which thirty-eight were confirmed by sequencing. For 72 primer pairs, multiple products were amplified however 34 reactions had the expected product size within the various bands amplified and thirty-eight reactions did not contain the correct expect product. Sequence confirmation was obtained for seven out of the 34 reactions that had the correct expect size within various bands. The remaining twenty-eight out of the 321 candidates produced a single PCR product that did not match the expected size. For this group, we sequenced 6 PCR products that were the predominant band and close in predicted size; one additional candidate was confirmed.

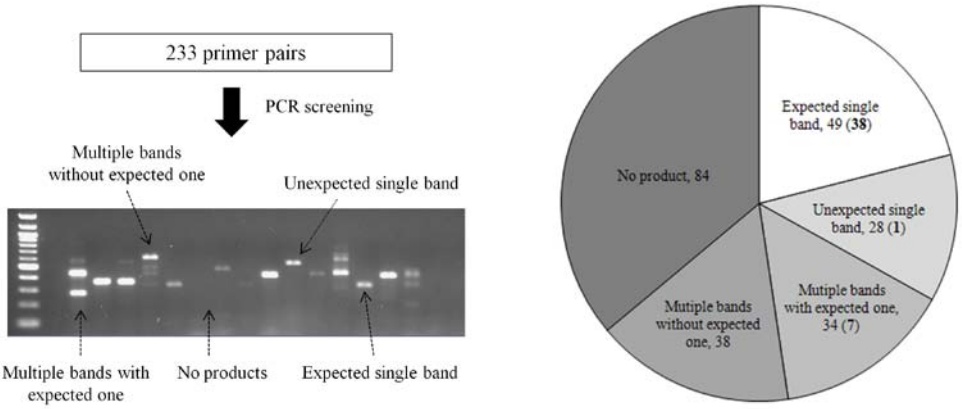


Figure 3.3 Chimeric Transcript PCR Validation Strategy. Shown here is the summary for 233 chimeric transcript PCR reactions. Using Pools of cDNA from various breast cancer cell lines, the fidelity of the primers was determined. Shown by color is the type of PCR product that was observed. In bold within each color is the number of chimeric transcripts that were sequenced confirmed.

Based on all of the sequencing, we identified new combinations of exons that the primers would amplify. For example, exon 13 of NAIP was expected to fuse with exon5 of OCLN. However, our designed primer amplified two predominant bands; the expected product and a smaller product as a result of the forward primer annealing to sequences within exon 12 of NAIP. Sequencing of the unexpected smaller band revealed that exon 12 of NAIP, instead of exon 13, fused with exon 5 of OCLN. By sequencing the unexpected size, but the predominant PCR products, three iso-forms were validated in addition to the original expected chimeric transcripts. Collectively, through this approach, we validated 48 FS chimeric transcripts that when classified by the chromosomal location of the genes involved in the fusions, 13 are intra chromosomal, 34 are read-through, and 1 is inter-chromosomal. Of note, two chimeric transcripts that our analysis identified have also previously been described in the literature; BCAS4-BCAS3, MDS1-EVI1 (Table 3.1).

Table 3.1 Validated Frame-shifted chimeric transcripts. Chimeric transcripts were validated by RT-PCR and confirmed by sequencing. All transcripts have predictive neo-peptides by frame-shifted mutation at downstream genes or by translated from 5'UTR region. The average length of frame-shifted neo-peptide is 32.7 amino acids with range of 1 amino acid to 204 amino acids.

Length of FS peptides	Upstream Gene	Location	Description	Downstream Gene	Location	Description
4	BOLA2	16p11.2	bolA homolog 2 (E. coli)	SMG1	16p12.3	PI-3-kinase-related kinase SMG-1
64	GFOD1	6pter-p22.1	glucose-fructose oxidoreductase domain containing 1	C6orf114	6p23	chromosome 6 open reading frame 114
63	MDS1	3q26	myelodysplasia syndrome 1	EVI1	3q24-q28	ecotropic viral integration site 1
124	C11orf79	11q12.2	chromosome 11 open reading frame 79	C11orf66	11q12.2	chromosome 11 open reading frame 66
35	ABHD14A	3p21.1	abhydrolase domain containing 14A	ACY1	3p21.1	aminoacylase 1
35	RBM14	11q13.1	RNA binding motif protein 14	RBM4	11q13	RNA binding motif protein 4
20	C20orf29	20p13	chromosome 20 open reading frame 29	VISA	20p13	virus-induced signaling adapter
34	RRM2	2p25-p24	ribonucleotide reductase M2 polypeptide	C2orf48	2p25.1	chromosome 2 open reading frame 48
33	ELAC1	18q21	elaC homolog 1 (E. coli)	SMAD4	18q21.1	SMAD family member 4
6	BCAS4	20q13.13	breast carcinoma amplified sequence 4	BCAS3	17q23	breast carcinoma amplified sequence 3
28	C22orf39	22q11.21	chromosome 22 open reading frame 39	HIRA	22q11.21	HIR histone cell cycle regulation defective homolog A
23	PMF1	1q12	polyamine-modulated factor 1	BGLAP	1q25-q31	bone gamma-carboxyglutamate (gla) protein
36	SDHD	11q23	succinate dehydrogenase complex, subunit D, integral	TEX12	11q22	testis expressed 12
8	PRR13	12q12	proline rich 13	PCBP2	12q13.12-q13.13	poly(rC) binding protein 2
3	RMND5A	2p11.2	required for meiotic nuclear division 5 homolog A	ANAPC1	2q12.1	anaphase promoting complex subunit 1
7	TYMP	22q13.33	thymidine phosphorylase	SCO2	22q13.33	SCO cytochrome oxidase deficient homolog 2 (yeast)
1	NAIP	5q13.1	NLR family, apoptosis inhibitory protein	OCLN	5q13.1	occludin
7	C1orf151	1p36.13	chromosome 1 open reading frame 151	NBL1	1p36.13-p36.11	neuroblastoma, suppression of tumorigenicity 1
138	DDIT3	12q13.1-q13.2	DNA-damage-inducible transcript 3	MARS	12q13.2	methionyl-tRNA synthetase
204	RIPK3	14q11.2	receptor-interacting serine-threonine kinase 3	ADCY4	14q12	adenylate cyclase 4
7	COMMD3	10pter-q22.1	COMM domain containing 3	BMI1	10p11.23	BMI1 polycomb ring finger oncogene
55	MED8	1p34.2	mediator complex subunit 8	ELOVL1	1p34.2	elongation of very long chain fatty acids
29	POLR3J3	7q22.1	polymerase (RNA) II (DNA directed) polypeptide J3	LOC100134053	7p13	PREDICTED: similar to POLR2J4 protein
24	BGLAP	1q25-q31	bone gamma-carboxyglutamate (gla) protein	PMF1	1q12	polyamine-modulated factor 1
121	TMEM199	17q11.2	transmembrane protein 199	SARM1	17q11	sterile alpha and TIR motif containing 1
84	C1QTNF6	22q13.1	C1q and tumor necrosis factor related protein 6	IL2RB	22q13.1	interleukin 2 receptor, beta
58	LOC100131434	Xq28	PREDICTED: hypothetical protein LOC100131434	FLJ44451	Xq28	PREDICTED: hypothetical protein FLJ44451
57	COX19	7p22.3	COX19 cytochrome c oxidase assembly homolog	CENTA1	7p22.3	centaurin, alpha 1
35	ACSF2	17q21.33	acyl-CoA synthetase family member 2	CHAD	17q21.33	chondroadherin
17	TIMM23B	10q11.23	PREDICTED: translocase of inner mitochondrial membrane	LOC100132418	10q11.23	PREDICTED: similar to PRO1102
16	NDUFA13	19p13.2	NADH dehydrogenase (ubiquinone) 1 alpha subcomplex, 13	YJEFN3	19p13.11	YjeF N-terminal domain containing 3
3	ADHFE1	8q13.1	alcohol dehydrogenase, iron containing, 1	C8orf46	8q13.1	chromosome 8 open reading frame 46
4	HPS4	22cen-q12.3	Hermansky-Pudlak syndrome 4	ASPHD2	22q12.1	aspartate beta-hydroxylase domain containing 2
6	KIAA1267	17q21.31	KIAA1267	ARL17P1	17q21.32	ADP-ribosylation factor-like 17 pseudogene 1
12	LOC100129406	1p13.2	PREDICTED: hypothetical protein LOC100129406	CTTNBP2NL	1p13.2	CTTNBP2 N-terminal like
15	RNF216	7p22.1	ring finger protein 216	RBAK	7p22.1	RB-associated KRAB zinc finge
13	DEDD	1q23.3	death effector domain containing	NIT1	1q21-q22	nitrilase 1
15	RAD54B	8q21.3-q22	RAD54 homolog B (S. cerevisiae)	LOC100128414	8q22.1	PREDICTED: similar to fibrinogen silencer binding
41	TOPORS	9p21	topoisomerase I binding, arginine/serine-rich	DDX58	9p12	DEAD (Asp-Glu-Ala-Asp) box polypeptide 58
10	NDUFC2	11q14.1	NADH dehydrogenase (ubiquinone) 1, subcomplex unknown	KCTD14	11q14.1	potassium channel tetramerisation domain containing 14
9	LRRC57	15q15.1	leucine rich repeat containing 57	SNAP23	15q15.1	synaptosomal-associated protein, 23kDa
6	IPO11	5q12.1	importin 11 (IPO11), transcript variant 1	SLRN	5q12.1	synleurin
24	SNRPF	12q22	small nuclear ribonucleoprotein polypeptide F	CCDC38	12q22-q23.1	coiled-coil domain containing 38
3	NDUFB8	10q23.2-q23.33	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 8	SEC31B	10q24.31	SEC31 homolog B (S. cerevisiae)
7	MIA	19q13.32-q13.33	melanoma inhibitory activity	RAB4B	19q13.2	RAB4B, member RAS oncogene family
19	THAP2	12q21.1	THAP domain containing, apoptosis associated protein 2	TMEM19	12q21.1	transmembrane protein 19
4	NIT1	1q21-q22	nitrilase 1	DEDD	1q23.3	death effector domain containing
9	RNF139	8q24	ring finger protein 139	NDUFB9	8q13.3	NADH dehydrogenase (ubiquinone) 1 beta subcomplex, 9

3.3.3 Frequency of FS chimeric transripts

In order to establish the frequency for the chimeric transcripts, the 48 FS chimeric transcripts validated in the initial pool screening were screened across all 50 breast cancer cell lines individually by RT-PCR (see Figure 3.4). Frequencies varied from 2% to 98%. From these results, 17 out of 48 were not detected in the non-cancerous cell line, MCF-10A. Due to insufficient amount of cDNA from primary tumor samples, 35 out of 48 FS chimeric transcript were screened in 57 breast tumors. Frequencies in primary samples ranged from 0% to 97%. In addition, only 3 out of the 48 chimeric transcripts that were present in cell lines were absent in primary samples. Among the 57 primary samples, 3 samples consisted of normal breast tissues of which 12 out of 35 chimeras were not detected. Due to low RNA yields and the difficulty in obtaining RNA material from healthy breast tissue only 22 chimeric transcripts were screened in both the MCF-10A and in 3 primary normal breast tissue samples. The remaining 16 chimeric transcripts were only screened in the MCF10A cell line. By chromosomal location classification, read-through transcripts were the most frequent followed by intra-chromosomal and inter-chromosomal. Though the majority of transcripts identified are read-through, some intra-chromosomal chimeras such as GFOD-C6orf114 and inter-chromosomal gene fusion, TRIM61 – FARSB (data not shown due to in-frame mutation) had high frequencies of 86.1% to 71.4% respectively with about 90% precision. In terms of precision of our RT-PCR screenings, we observed some discrepancies in the results of RT-

PCR from samples of same source. cDNA from cell lines yielded more robust RT-PCR results than from primary samples.

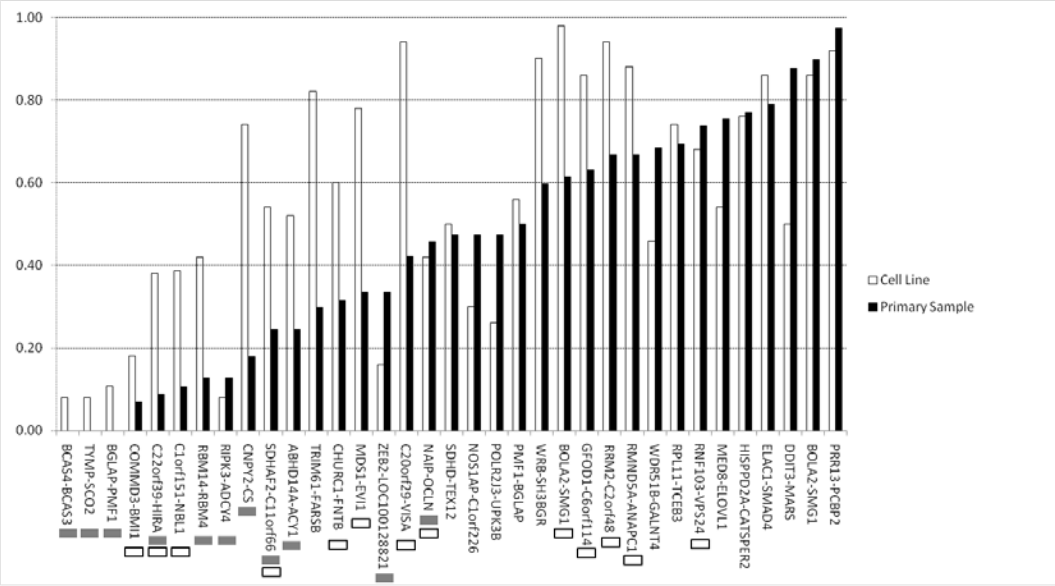


Figure 3.4 The frequency of chimeric transcripts in breast cancer. This graph shows the frequency of 35 FS chimeric transcripts that were examined in both cell lines and patients. The presence of chimeric transcripts in each sample was determined by RT-PCR. The expected size of amplified band was considered as a positive. 10% of them were subject to sequencing to confirm. Gray bar indicates their absence in non-cancerous cell line (MCF-10A) while ▭ indicates their absence in three normal breast tissues. Not all chimeric transcripts were screened in normal samples (see details in supplementary table).

3.3.4 Potential epitopes with population coverage

To see the potential MHC coverage that the chimeric transcripts may have if translated and used as cancer vaccine antigens, we examined all possible epitopes that would be represented by the frame-shifted peptides. The length of the predicted FS peptides ranged from 1 to 204 amino acids with the overall average length being 32 amino acids long. Using 10 amino acids of the upstream genes from the junction plus the full length of downstream FS peptide, all possible MHC I potential epitopes were predicted. For the 48 chimeric transcripts

that generate a FS neo peptide, 1,317 unique epitopes that are able to bind to 35 different MHC alleles were identified. For example the FS peptide, from the chimeric transcript called DDIT – MARS is 138 amino acids in length, is able to generate epitopes for 27 different MHC alleles. However, there were a few chimeric transcripts such as RMND5A - ANAPC1 that did not contain any MHC I binding epitopes. Through this analysis, when we evaluate all of the FS antigens, the coverage regarding the various populations within the United States are as followed: 99.66% for Caucasians, 98.22% for Hispanic, 96.5% for Asian Pacific Islanders, and 92.92% for African American. Allele frequencies were adjusted using the relative frequency of the individual chimeric transcripts. In addition, we made an effort to identify the minimum number of antigens that would have the largest population coverage by including only the top five most frequent HLA alleles found within individual ethnic groups. Based on this analysis, 85.56% of Caucasians from the United States would benefit from a vaccine that contained seven neo-peptides from the following chimeric transcripts; BOLA2-SMG1, GFOD1-C6orf114, ELAC1-SMAD4, TIMM23B-LOC100132418, C22orf39-HIRA, MDS1-EVI1 and DDIT3-MARS whereas 75.93% of Hispanics in the USA would benefit from a vaccine that contained 4 antigens (BOLA2-SMG1, GFOD1-C6orf114, ELAC1-SMAD4, DDIT3-MARS). Other combinations that consisted of 6 antigens (GFOD1-C6orf114, RRM2-C2orf48, TIMM23B-LOC100132418, MDS1-EVI1, DDIT3-MARS, C1QTNF6-IL2RB) or a pool of five antigens (C20orf29-VISA, GFOD1-C6orf114, LOC100129406-CTTNBP2NL, MDS1-EVI1, DDIT3-MARS) would protect 82.5% of Asian

Pacific Islanders and 63.08% of African Americans in the USA respectively. In conclusion, if a cancer vaccine was limited by the number of antigens that could be used at one time, 10 FS antigens would be effective in more than 60% of the population regardless of ethnicity.

Table 3.2 Population coverage of antigens from chimeric transcripts. A. Shows the number of possible MHC-binding epitopes from selected chimeric transcripts. Entire data is provided in Supplementary table 1. Numbers in ( ) means the epitopes derived from junction between upstream 10 amino acids and FS peptide. * indicates the length of FS peptides. B. We project the population coverage based on a group of selected antigens rather than all 48. Overall, about 10 antigens are able to cover more 60% of the population in the USA regardless of ethnicity or HLA types.

A.

Chimeras	FS pep*	B*3501	A*0201	A*0301	B*0702	A*0101	B*4402	B*5301	A*1101	A*2402	B*4001	B*5101
BOLA2-SMG1	4				(1)							
GFOD1-C6orf114	64	2	1	(1)	1			1	(2)	1		
MDS1-EVI1	63	(1)	(1)	2		1		1	4(1)	1		
C20orf29-VISA	20	1			1							
RRM2-C2orf48	34				2						2(1)	
ELAC1-SMAD4	33	3	1	1	1			2	2	2(1)		
C22orf39-HIRA	28	(1)	(1)				(1)			1	(1)	
DDIT3-MARS	138	12	5	5	4	1		5	10	2(1)		1
C1QTNF6-IL2RB	84	3	2		4			1	1	1	2	1
TIMM23B-LOC100132418	17	1	5(1)				1				1	
LOC100129406-CTTNBP2NL	12	1(1)	(1)									

B.

United States	Coverage of all antigens (%)	Top 5 frequent HLA alleles	No. of Antigens	Population Coverage (%)
Caucasians	99.66	A*0201, A*0101, A*0301, B*4402, B*0702	7 - (BOLA2-SMG1),(GFOD1-C6orf114),(ELAC1-SMAD4),(TIMM23B-LOC100132418),(C22orf39-HIRA),(MDS1-EVI1),(DDIT3-MARS)	85.56
Hispanic	98.22	A*0201, A*2402, A*0301, B*3501, B*0702	4 - (BOLA2-SMG1),(GFOD1-C6orf114),(ELAC1-SMAD4),(DDIT3-MARS)	75.93
Asian Pacific Islander	96.5	A*1101, A*2402, B*4001, A*0201, B*5101	6 - (GFOD1-C6orf114), (RRM2-C2orf48), (TIMM23B-LOC100132418), (MDS1-EVI1), (DDIT3-MARS), (C1QTNF6-IL2RB)	82.5
African American	92.92	A*0201, B*5301, A*0301, B*3501, A*0101	5 - (C20orf29-VISA),(GFOD1-C6orf114), (LOC100129406-CTTNBP2NL),(MDS1-EVI1),(DDIT3-MARS)	63.08

3.3.5 FS chimeric transripts in mouse and dog

To see whether these FS chimeric transcripts can be tested in animal models such as mice and dogs, I examined the homologous genes involved in our putative chimeric transcripts, in mouse and dog by using BLAST program. I

assumed that there might be possible corresponding mouse chimeric transcripts corresponding to human ones when both genes in fusion have homologous genes in mouse. The same principle was applied to dog. 64 mouse chimeric transcripts were selected by the homologous gene search for RT-PCR screening. 14 mouse chimeric transcripts were detected and sequence confirmed at least one of 10 mouse cell lines (Table 3.3).

Table 3.3 The list of validated mouse chimeric transcripts. 13 mouse transcripts were validated in 10 mouse cell lines. Peptide marked with * were detected in * sample at mRNA level when multiple peptides were detected.

Gene Fusions	Chimeric peptide	Postive sample	FS Peptide sequnce
Rnf103 - Vps24	In-frame	4T1, Tubo	-
Cnpy2 - Cs	In-frame	4T1	-
Thap2 + Tmem19	FS (19 a.a*, 7 a.a)	Tubor tumor*, B16-F10*, CRL-2116, CCL-51	VTFGLFLRGAGCSPSSFLL, GWWSDSS
Rnf139 + Ndufb9	FS (8 a.a)	Tubor tumor, B16-F10, CRL-2755, CRL-2166, CCL-51, MC4-L5, MC7-2A	TSTGTLLA
Bloc1s1 + Rdh5	No neo-peptide (3'UTR)	4T1, Tubo, CRL-2755, CCL-51	-
Lats2 + Xpo4	FS (6 a.a)* / In-frame	Tubo tumor, CRL-2116, MC4-L2*	TTFHGQ
Tmem170 + Cfdp1	In-frame	4T1, B16-F10, CCL-51	-
Slc35a3 + Hiat1	FS (7 a.a)*/ In-frame	B16-F10*, CCL-51	ASRNRLP
Nos1ap + EG665574	New 5'UTR (4 a.a)	Tubo, Tubo tumor, MC7-2A	VDHS
Samd5 + Sash1	In-frame	B16-F10	-
Rbm14 + Rbm4b	FS (6 a.a)	CRL-2116, CCL-51	GGMCVG
Mia1 + Rab4b	FS(24 a.a*, 7 a.a)	Tubo tumor, MC7-2A*	SNRTPTTLSAWSLDPGWS TLGGRL, TSSSNSW
Pir + Figf	In-frame	CRL-2755, CRL-2116, CCL-51, MC4-L5, MC7-2A	-

24 predicted dog chimeric transcripts were screened in 22 cancer samples from melanomas, osteosarcomas, lymphosarcoma, hemangiosarcoma, breast, mast cell tumor, transitional cell carcinoma, and thyroid adenocarcinoma as well as 13 normal samples from various tissue types. 8 chimeric transcripts were validated in dog cancer samples (Table 3.4).

Table 3.4 The list of validated dog chimeric transcripts.

Gene Fusions	Chimeric peptide	Tumor positive	Normal positive	FS Peptide sequnce
IPO11 - SLRN	Frame-shift	17	11	RLTSKGP
MIA - RAB4B	Frame-shift	19	4	SNRTPTTQSAWSLDLG
MED8 - ELOVL1	Frame-shift	20	1	VP
RNF103 - VPS24	In-frame	20	11	-
CNPY2 - CS	5 'UTR	21	5	-
CHURC1 - FNTB	In-frame	9	1	-
ABHD14A - ACY1	Frame-shift	20	2	LTKRGA
RBM14-RBM4	?	22	8	-

3.4 Methods

3.4.1 Data Sets and Algorithm

To identify potential putative chimeric transcripts, that when translated would result in a frame-shifted neo-peptide; we targeted two publically available datasets and applied an algorithm that was used to identify chimeric transcripts. Specifically, we used the sequences found within the Expressed Sequence Taq (EST) library (51) and the Human RefSeq database (76) from the National Center for Biotechnology Information (NCBI). Using the stand-alone BLAST program, we aligned all EST sequences to RefSeq. We picked ESTs that aligned with more than 50-85 base pairs and had 95-97% homology to RefSeqs that have been previously annotated by National Center Institute (NCI). We further filtered out our alignment data by eliminating the EST sequences that did not align to multiple RefSeqs or were aligned in the 3'-5' orientation. Lastly, we also eliminated the sequences that aligned with non-coding sequence regions. The remaining EST sequences were then used to identify the chimeric transcripts. Only the ESTs that aligned to two or more distinct RefSeq in consecutive positions were considered to be potential candidates. To be defined as a coding

chimeric transcript, the EST sequences had to be at least 100-170 bp long with sequence similarity greater than or equal to 95%- 97% to the RefSeq. Also, the junction point between the two genes had to occur within the coding sequence of the upstream gene and orientation of the upstream gene alignment had to be in the positive (5'-3') orientation. To eliminate false calls, all potential chimeric EST sequences had to be either present in more than one cDNA library or supported by three or more independent EST sequences. In addition, chimeric transcripts were classified based on the relative position of two genes. Classification of types of chimeric transcript was based on relative position of two fusion genes on the chromosome. Specifically genes found on different chromosomes resulted in inter-chromosomal fusion while genes found in same chromosome were intra-chromosomal or read-through chimeric transcripts. Read-through chimeric transcripts resulted from two neighboring genes on same strand, otherwise intra-chromosomal.

3.4.2 Cell lines and tissue samples

The 50 Human Breast cancer cell lines were obtained from the American Type Culture Collection (ATCC) (see supplementary table B.4) and were grown according to recommendations. Human breast cancer tissue specimens were acquired from Mayo Clinic after appropriate patient consent and approval of the Mayo Clinic Institutional Review Board. All specimens were coded and anonymized.

3.4.3 Primer design and RT-PCR validation

Total RNA was extracted from breast cancer cell lines and primary breast tissues using the TRIzol LS reagent (Life Technologies, Carlsbad, CA) following the manufacturers protocol. RNA integrity was determined by gel electrophoresis and concentration was determined by measuring absorbance at 260/280 on the Nano-drop (NanoDrop Products, Wilmington, DE). cDNA was prepared by using the SuperScript^{TM} III First-Strand Synthesis SuperMix (Life Technologies, Carlsbad, CA) that includes random hexamers and oligo dT's following the manufacturer's recommended protocol. cDNA integrity and quality were assessed by performing a β-actin control PCR. End Point PCR primers for each chimeric transcript were designed using Primer3 (87) so that the forward and reverse primer both binds 80bp to 280bp upstream/downstream from the junction point. End-point PCR reactions using approximately 25 ng of cDNA, reagents from (Life Technologies, Carlsbad, CA) and 35 cycles were performed using Mastercycler ep gradient S (Eppendorf, Hamburg, Germany). PCR products were analyzed on 1.5% agarose gels. PCR products were purified and sequence confirmed by Applied Biosystems 3730 (Life Technologies, Carlsbad, CA).

3.4.4 Epitope prediction and population coverage

Predicted frame-shifted peptides including 10 amino acids from the upstream genes were used to analyze all possible epitopes. By using the Immune Epitope Database and Analysis Resource (IEDB) (88) that is provided by the National Institute for Allergy and Infectious Diseases (NIAID), we were able to obtain a list of all possible epitopes that would be produced from validated

chimeric transcripts with their respective population coverage. First, epitopes binding to MHC class I were identified by a prediction algorithm tool from IEDB. We selected artificial neural network (ANN) as a prediction method according to IEDB evaluation. NetMHC (89) was then used in IEDB for ANN implementation. The prediction of peptide-MHC binding was based on artificial network trained on data for 55 MHC alleles and position-specific scoring matrices (PSSMs) for 67 additional HLA alleles. An epitope was considered positive when the $IC_{50} < 500$ because most known epitopes have high ($IC_{50} < 50$) or intermediate ($IC_{50}<500$) affinities. After obtaining all possible epitopes, we then made a hypothetical projection about population coverage by using another analysis tool in IEDB. This tool was designed to calculate the proportion of individuals based on HLA genotypic frequencies (from dbMHC, NCBI) (90). The linkage equilibrium between different HLA alleles was assumed in their calculations. We found that chimeric transcripts were completely independent of each other by chi square test. Both frequencies of HLA genotype and chimeric transcripts were considered for population coverage. Original HLA genotypic frequencies were adjusted by the frequency of chimeric transcript to generate the epitopes to bind the HLA allele. We attempted to find the fewest number of chimeric transcripts which would correspond to an effective FS peptide vaccine for the greatest percentage of the human population. In order to achieve this goal, we needed to figure out the best set of chimeric transcripts which could protect the maximum portion of the population depending on the frequency of these chimeric transcripts and the frequency of MHC alleles which bind to them. There are many possible

combinations of MHC alleles, chimeric transcripts, and cell lines. In order to reduce the complexity of searching through all of these possible combinations to find the optimum number of chimeric transcripts needed for a vaccine, we considered only the 5 most frequent MHC alleles in the human population. Once these frequent MHC alleles were selected, we determined the combination of chimeric transcripts which bind to most of these MHC molecules. The percentage of the population which would be protected by these chimeric transcripts and appropriate MHC alleles was then calculated by using "Population coverage calculation" program (90). In this program, average population coverage by a set of epitopes is generated by the following numbers; projected population coverage, average number of possible epitopes by the population.

3.5 Discussion

Selecting and determining the appropriate antigens to be used in a cancer vaccine is one of the most critical and time consuming steps in the development process of a cancer vaccine. Here we explore the concept of using FS peptides that are generated from gene rearrangements and/or chimeric transcripts as the sources of antigens to be used in a prophylactic cancer vaccine. The screening of publically available sequence data allowed for the rapid identification of chimeric transcript that if translated would produce novel neo-peptides. Through this approach, 48 FS chimeric transcripts out of the called 496 putative candidates derived from the analysis of EST Db were identified and validated in breast cancer cell lines and primary tumors. Out of the 48 confirmed candidates 2

chimeric transcripts, BCAS4-BCAS3 (breast) and RBM14-RBM4 (prostate) have been previously been identified by 454 sequencing transcriptome sequencing and reported by Maher et al. (61, 81). With the increasing availability of high throughput sequencing data, the development of an algorithm screening process may expedite the discovery of neo-peptides that are produced from chimeric transcripts that are generated by either trans-splicing or chromosomal rearrangement mechanisms. Such candidates could be then be used as cancer vaccine antigens and novel therapeutic targets.

For the last several years, the cancer research community has been interested in understanding the role that gene fusions play in leukemias since the identification of the BCR/ABL gene fusions has been so successful for the treatment of chronic myeloid leukemia (CML). For solid tumors (prostate, breast and skin cancers) the search for gene fusions has recently expanded as a result of high profile sequencing projects (58, 80, 81, 85). The number of gene fusions has been doubled in the literature over the past four years, but the overall frequency for each gene fusion has not been described nor evaluated across different solid tumor types. Currently there are over 70 different gene fusions that have been reported for more than 60 different cancer types. Included in this list are 78 gene fusions that have been found in breast of which 33 are considered FS chimeric transcripts (61, 62, 80, 85, 86). In order to make a prophylactic cancer vaccine or to truly understand if the presence of gene fusions is random or a controlled process, identifying the frequencies would help evaluate if chimeric transcripts are a result of a driver or passenger mutations as a result of the combination of

genes involved. Using the criteria that were recently described by to make a prophylactic cancer vaccine or to truly understand if the presence of gene fusions is random or a controlled process, identifying the frequencies would help evaluate if chimeric transcripts are a result of a driver or passenger mutations as a result of the combination of genes involved. Using the criteria that were recently described by Bozic et al. (91), FS chimeric transcripts would be considered as a driver mutation since protein sequences are affected by a frame-shift mutation. However, the data present in this study does not fully support the driver mutation phenomenon because the majority of the chimeric transcripts are present in too high of frequency to be considered a drive mutation. Therefore, by definition, the FS chimeric transcripts in this study could be considered passenger mutations as a result of the overall genomic instability of the tumor. The frequencies for these target antigens will be critical to the development of a prophylactic cancer vaccine. For example, the prevalence of mutation will aid in determining the number of antigens that would be needed to protect at least 70~80% of the population.

Epitopes from antigens will elicit immune response only when it is presented to immune system by major histocompatibility complex (MHC) molecules. Therefore, we can estimate the efficacy of the antigens produced from these chimeric transcripts based on MHC binding epitopes. As we expected, long neo-peptides from frame-shifted mutations, relative to substitution, have a rich pool of epitopes. 46 of HLA-A*0201 epitopes were presented in 48 frame-shifted mutations while 241 of HLA-A*0201 epitopes were presented in 1,307 mutations

according to Segal et al (92). This clearly shows the advantage of using FS peptides as a vaccine antigen in covering a greater proportion of population. Another discover of interest is the observation that several HLA alleles from the 48 FS peptides such as HLA A*0101, HLA A*2601, HLA B*4402 and HLA B*5101 lacked their binding epitopes. In general, frame-shifted chimeric transcripts will be ideal antigens according to nine criteria suggested by Cheever et al.; i) therapeutic function, ii) immunogenicity, iii) oncogenecity, iv) specificity, v) expression level and % positive cells, vi) stem cell expression, vii) No. patients with antigens-positive cancers, viii) No of epitopes and ix) cellular location of expression (23). Frame-shifted chimeric transcripts may be even better antigens in terms of epitope presentation, as we showed in this study, by generating longer neo-peptides. In addition, many of our FS chimeric transcripts were detected in multiple samples and some of them could be tumor-specific even though we need to screen more normal samples to validate it. Furthermore, we validated the corresponding chimeric transcripts in mouse and dog by homology search that could be tested for immune response. Considering no frame-shifted peptides had even been evaluated in their study of 75 antigens, it is worth examining the potential of FS chimeric transcripts as a cancer vaccine antigen.

3.6 Conclusion

Gene fusions in cancer have been proven to be effective diagnostic and therapeutic targets. Here we show the potential of chimeric transcripts as appropriate vaccine antigens. As we studied the transcriptome, we need to take

next step in investigating whether or not the protein is translated from these chimeric transcripts. Several studies have shown the production of a chimeric protein stemming from in-frame gene fusion (93-95). As a next step, we will search corresponding FS peptides from these chimeric transcripts in tumor and normal cells. Finally, detected peptides will be subject to immunological tests by using animal models. As observed in other studies (96), differential expression levels of chimeric transcripts between tumor and normal cells may provide us with clues regarding the presence of chimeric proteins. This study provides an insight into utilizing chimeric transcripts as a first step in a broader effort to develop a cancer vaccine from frame-shifted peptides based on their frequencies and possible epitopes.

CHAPTER 4

PATTERNS IN CHIMERIC TRANSCIRPTS

4.1 Introduction

The list of about 770 gene fusions from the literatures and our study was obtained and stored as a table. One interesting observation was that some genes appeared multiple times in the table. Therefore, I wondered whether these gene fusions are totally random events or not. What patterns could we find from these gene fusions? To search the patters in gene fusions, I collected the information of gene fusions from public data base and our study and analyzed them by using program of complex network analysis, Cytoscape (97). Three patterns were detected in our study; interconnected network, dominant exon combination, and dominant iso-forms. These patterns in gene fusions enable us to do cost-effective targeted sequencing to establish the frequency of aberrant transcripts with a higher accuracy.

4.2 Gene Fusions in the Literatures

The information regarding gene fusions or chimeric transcripts was retrieved from two sources; Mitelman Database (98) and the Catalogue Of Somatic Mutations In Cancer (COSMIC), Sanger Institute (99). 588 distinctive gene fusions were collected from Mitelman Database (July, 2010). 96 distinctive gene fusions with their position of junction between two genes were obtained from COSMIC (CosmicFusionExport_v51). In addition, we retrieved data on

gene fusions that had not been deposited into the two former databases from more recent publications.

4.3 Patterns in Gene Fusions

A total of 770 distinctive genes involved in 698 gene fusions as collected from literatures and our analysis were used for pattern analysis. We used a program called “Cytoscape” (97) to draw connections among genes according to their gene fusions.

4.3.1 Interconnected networks of chimeric transcripts

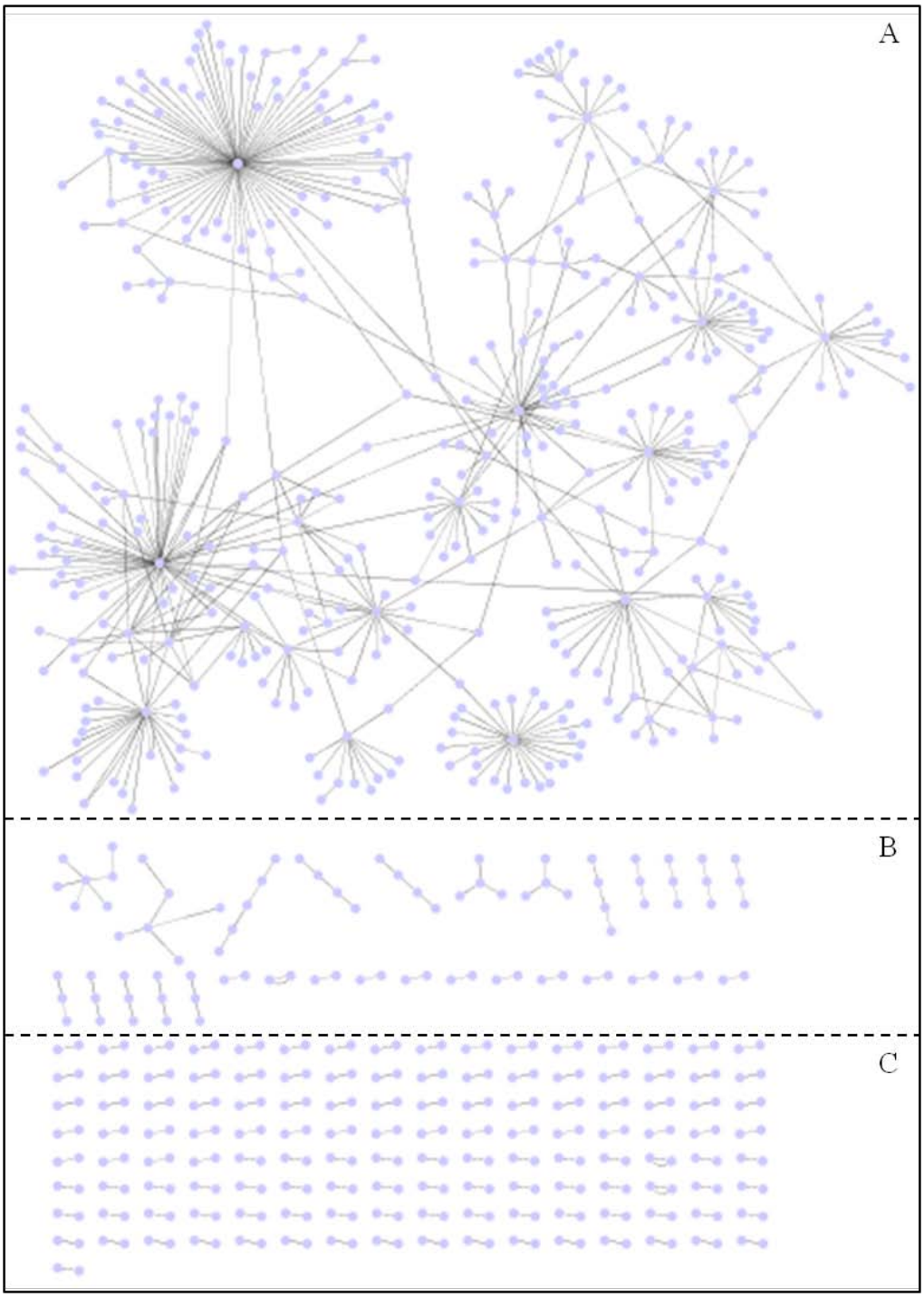


Figure 4.1 Network of gene fusions. Each node indicates a gene and edge connects nodes when two nodes (or genes) form a gene fusion. A. The largest

single cluster was derived from 506 gene fusions. B and C. Small clusters and separate gene fusions.

Interestingly, 506 gene fusions (72.5%) were connected in one large cluster because many of them the shared the same genes as partner (Figure 4.1). How about gene fusions detected in solid tumors? 77 out of 309 (24.9%) gene fusions in solid tumors formed a single clusters. 125 gene fusions (40.5%) belonged to 7 clusters, which comprised of more than 5 members (Figure 4.2).

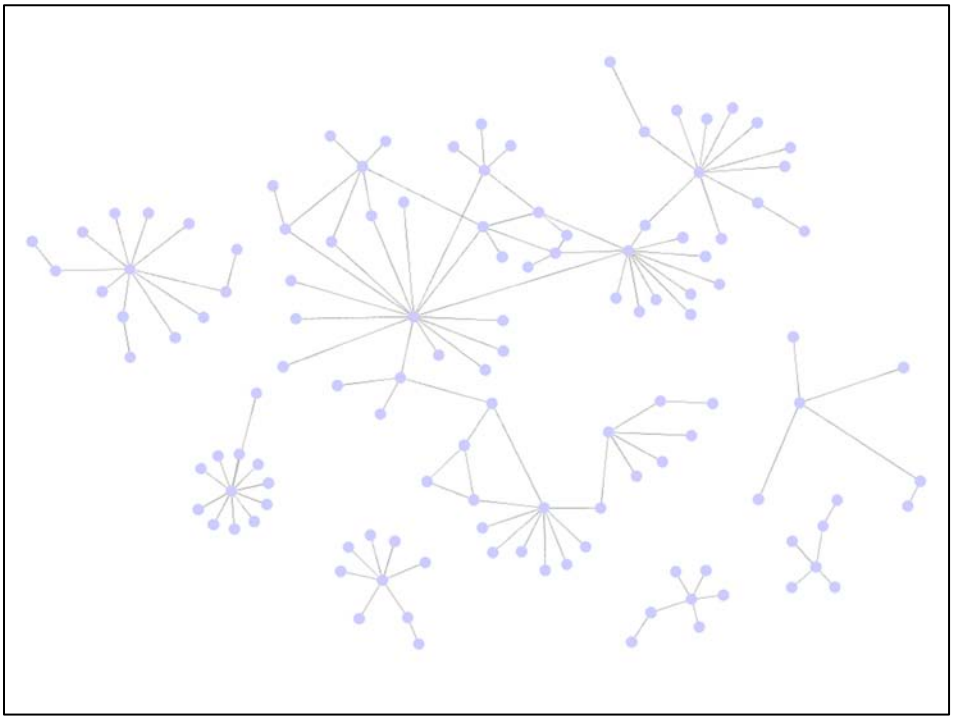


Figure 4.2 Network of gene fusions found in solid tumors. 7 clusters have more than 5 members each. The largest cluster consists of 77 gene fusions.

At an individual level, some genes combined with multiple genes as partners in their gene fusions (see Figure 4.3). By the relative position, 5' or 3', there consisted of two types; anchor upstream and anchor downstream. 43 genes (17% of 247 upstream genes of fusions) at the 5' position of gene fusions combined at least two or more gene at 3' position. 15 of them fused with more

than 5 genes as downstream partners. Myeloid/lymphoid or mixed-lineage leukemia (MLL) joined with 63 different genes at its downstream. 76 genes (20% of 371 downstream genes of fusion) joined more than one gene. B-cell CLL/lymphoma 6 (BCL6) has 22 different upstream gene as a partner. 13 of 76 genes have more than 5 partners.

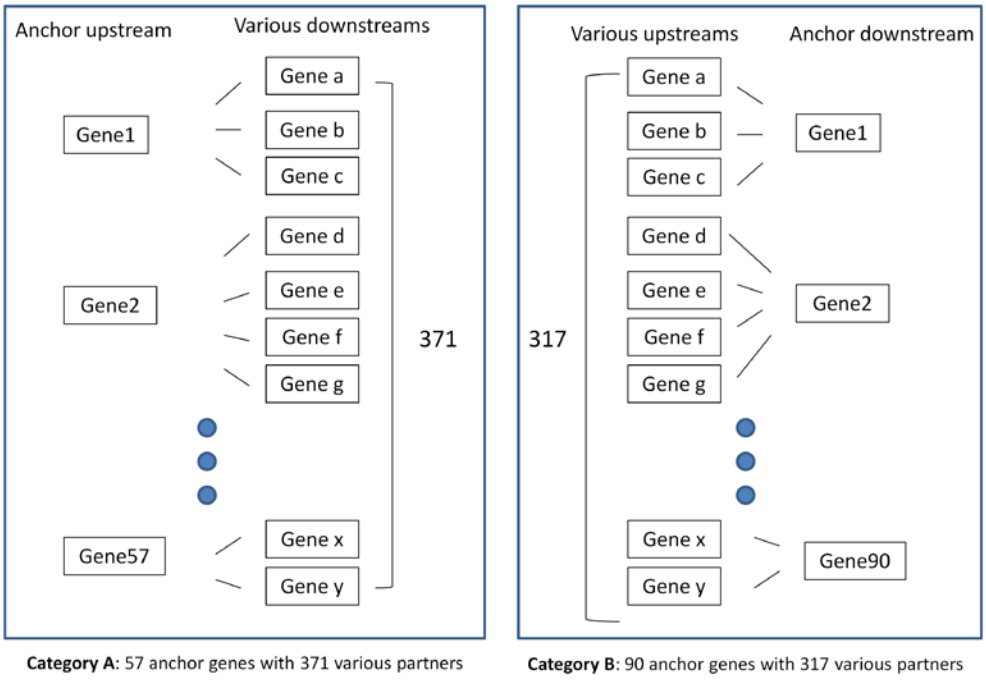


Figure 4.3 The multiple gene fusions with a shared gene. Based on the position of anchor genes, there are two types; anchor upstream and anchor downstream. 42 anchor upstream generated 377 gene fusions while 76 anchor downstream yield 287 gene fusions.

4.3.2 Dominant Exon combination

A total of 16 genes formed gene fusions with 10 or more partners. 12 genes at upstream position joined 10 or more downstream genes to generate gene fusions while 4 genes combined with 10 or more upstream genes to produce chimeras. We were able to retrieve exon information of gene fusions related to EWSR1, ETV1, ALK from COSMIC data. In the case of ETV1, 6 out of 13 exons

combined with 6 different genes. However, EWSR1 and ALK showed that a particular exon dominantly combined with other genes out of 19 and 29 exons respectively from EWSR1 and ALK (Figure 4.4). In most cases, exon 8 of EWSR1 joined to other genes to form gene fusions. Only exon 20 of ALK was involved in the combination with other genes.

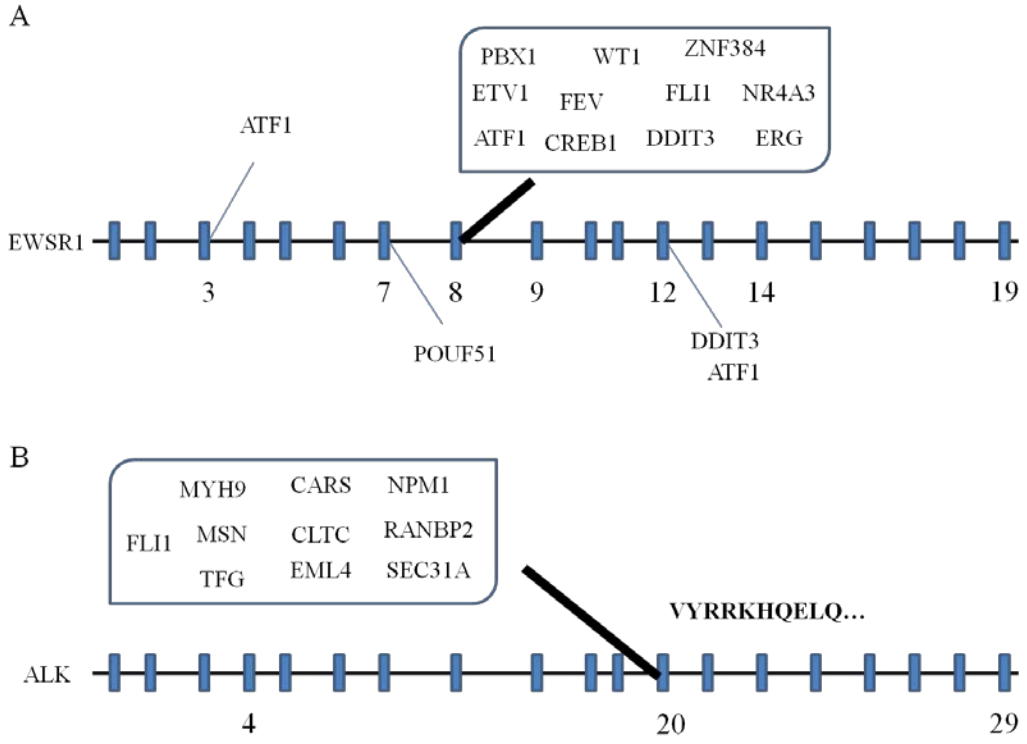


Figure 4.4 Dominant exon combinations. An exon8 combined with most of downstream partners among 19 exons from EWSR1. Out of 29 exons from ALK, only exon20 was involved in gene fusions with 10 different genes.

4.3.3 Dominant iso-forms of chimeric transcripts

42 gene fusions were reported to have iso-forms according to information extracted from the COSMIC database. Therefore, I was interested if there was a bias in the participation of iso-forms in gene fusions. In fact, 18 gene fusions had a dominant iso-form in general. For instance, Table 4.1 showed the dominant iso-

forms of gene fusion between TMPRSS2 and ERG from three studies. In all three studies, the combination between exon 1 of TMPRSS2 and exon 5 of ERG was the most frequent one (100-102). In the case of two gene fusions, COL1A1-PDGFB and ASPSCR1-TFE3, there was a dominant iso-form in each case, but each case had different dominant iso-form. 22 gene fusions did not have any dominant iso-form.

Table 4.1 Iso-forms of TMPRSS2-ERG. TMPRSS2 combined with ERG by means of several different exon combination according to three studies. However, all three studies showed that the combination between exon 1 of TMPRSS2 and exon 5 of ERG was most frequent in comparison to other combinations.

5' Gene	Last Exon	3' Gene	First Exon	Percentage	Total # sample	Sample	Reference
TMPRSS2	1	ERG	6b	0.07	15	prostate, carcinoma-adenocarcinoma	Yoshimoto et al.
TMPRSS2	3	ERG	6b	0.07			
TMPRSS2	**1**	**ERG**	**5**	**0.27**			
TMPRSS2	1	ERG	2	0.07			
TMPRSS2	1	ERG	6b	0.01	67	prostate, carcinoma-NS	Wang et al.
TMPRSS2	4	ERG	5	0.03			
TMPRSS2	1	ERG	3	0.06			
TMPRSS2	3	ERG	6b	0.03			
TMPRSS2	**1**	**ERG**	**5**	**0.45**			
TMPRSS2	3	ERG	5	0.13			
TMPRSS2	1	ERG	2	0.1			
TMPRSS2	4	ERG	2	0.03			
TMPRSS2	3	ERG	2	0.04			
TMPRSS2	3	ERG	6b	0.15	27	prostate, carcinoma-adenocarcinoma	Jhavar et al.
TMPRSS2	1	ERG	3	0.15			
TMPRSS2	**1**	**ERG**	**5**	**0.59**			
TMPRSS2	3	ERG	5	0.15			
TMPRSS2	1	ERG	2	0.07			

4.4 Discussion

In this study, we have shown that some combinations between two genes in chimeric genes are not random events based on observed patterns. First, there are a set of genes that combined with many other genes. Second, a particular single exon among all exons in a gene mainly contributes to generate a gene fusion. Third, a dominant combination of exons between two genes existed in

many gene fusions. These patterns are unlikely to be random based on a statistical test.

What does make these patterns? We suggested two hypothetical mechanisms; 3 dimensional configuration of chromosomes and expression pattern of two genes. It is a reasonable assumption that gene fusion is more likely to happen where two genes located close in space together. Recently published papers support this reasoning. Our second explanation is that the two genes of the fusion are not normally expressed in high level together, but they are highly expressed together in tumor. Therefore, two genes would have a higher probability to combine together by chance. This idea may be tested by using expression data generated by DNA microarray or RNA-seq.

The knowledge about these patterns allows us to perform targeted sequencing/resequencing to identify the putative gene fusions for cancer antigens efficiently. The genes that have many partners will be of primary targets and exon information helps us to design probes. By this approach, we can find new gene fusions derived from the targeted genes as well as accurate frequencies of targeted gene fusions. This information may enable us to select the better candidates as vaccine antigens. The pattern of fusions may also have a tumor bias, so be useful in diagnostics.

4.5 Conclusion

The knowledge about these patterns allows us to perform targeted sequencing/resequencing to identify the putative gene fusions for cancer antigens

efficiently. The genes that have many partners will be of primary targets and exon information helps us to design probes. By this approach, we can find new gene fusions derived from the targeted genes as well as accurate frequencies of targeted gene fusions. The obtained information enables us to select the better candidates as vaccine antigens. The pattern of fusions may also have a tumor bias, so be useful in diagnostics.

CHAPTER 5

CODING MICROSATELLITE DNA

5.1 Introduction

Cancer is a genetic disease, resulting from the sequential accumulation of genetic alterations (103). Common characteristics of tumor (104) are the basis to speculate that there are pivotal genetic alterations to induce a tumor. The targeting of these pivotal mutations brought us remarkable outcomes in cancer treatment like imatinib (105). The advent of high-throughput sequencing technology enables us to detect more tumor specific mutations as new drug targets by systematic sequencing of tumor transcripts. Recent large-scale sequencing studies show that the prevalence and patterns of somatic mutations are substantially different between samples even though there are more mutations involved in tumor than previous estimation (106, 107). These observations may indicate the absence of prevalent and consistent mutations over different cancer types at the DNA level even though there might be the prevalent tumor-specific alternative splicing or fusion transcript from translocations, which could not be detected by the way two studies referenced.

Simple repeat sequences, microsatellite (MS) DNAs, may offer another source of cancer mutations because of their high mutations rate. In addition, genomic instability, the characteristics of cancer, promotes the mutation events at MS DNAs during tumor development. Therefore, we may expect that common Indels occurr at coding MS DNA across different cancer samples. A recent large-

scale sequencing study of cancer genome done by Greenman et al. (107) found 11 mutations in multiple samples and five of them were FS mutations from coding MS DNAs.

Frameshift (FS) mutation from coding MS DNAs by insertion/deltion (Indels) is a potential source to generate tumor specific antigens due to their extensive polymorphism and frequent occurrence in the human genome. Several studies have already shown the potential of FS peptides from coding MS DNAs as novel targets of cancer treatment (35, 39, 46). FS peptides from MS DNAs will be good cancer vaccine antigens because they are likely to be immunogenic unlike one amino acid change from substitutions. These data support the feasibility of this approach in terms of immunogenicity and prevalence of FS peptides.

In this study, we tried to detect the tumor-specific mutations in coding MS DNAs by using the huge amount of EST data and RNA-seq data. Based on our definition of MS DNA, we can count the frequency of Indels in coding MS DNAs in tumor and normal. Through the analysis of transcriptome, we characterized the Indels in coding MS DNAs by several factors; length of repeat, repeat unit, tumor types, and allele. Finally, we selected putative cancer vaccine antigens based on the characteristics of Indels in coding MS DNAs.

5.2 Bioinformatic Approach

5.2.1 Definition of microsatellite DNA

What is a microsatellite DNA? In general, tandem sequence consists of repeating units of 1-6 base pairs in length (i.e. AAAAAAA, ACACACACACAC, and AGTAGTAGTAGTAGT). However, there is no real consensus about what is microsatellite DNA in terms of number of iterations and degeneracy (43).

Figure 5.1 An example of deletion in coding MS DNA. The gene called TGFbII has eight As in a raw in the coding region. One deletion of A results in a frame-shifted peptide, TVGRPHISC.

Therefore, I constructed my own functional definition of MS DNA. I focused on mono nucleotide (mono) MS DNA since established genes, *TGFβ-RII*, *BAX*, *hMSH3*, *hMSH6*, and *IGFIIR* , as a maker for the microsatellite instability (MSI) colorectal cancer genes are mono-nucleotides. First, we do not allow any degeneracy, but set seven as minimum number of iterations. The minimum number of iterations was determined based on distribution of the number of MS DNAs (Figure 5.2). Basically, we do not want to investigate either too many MS DNA or few MS DNA.

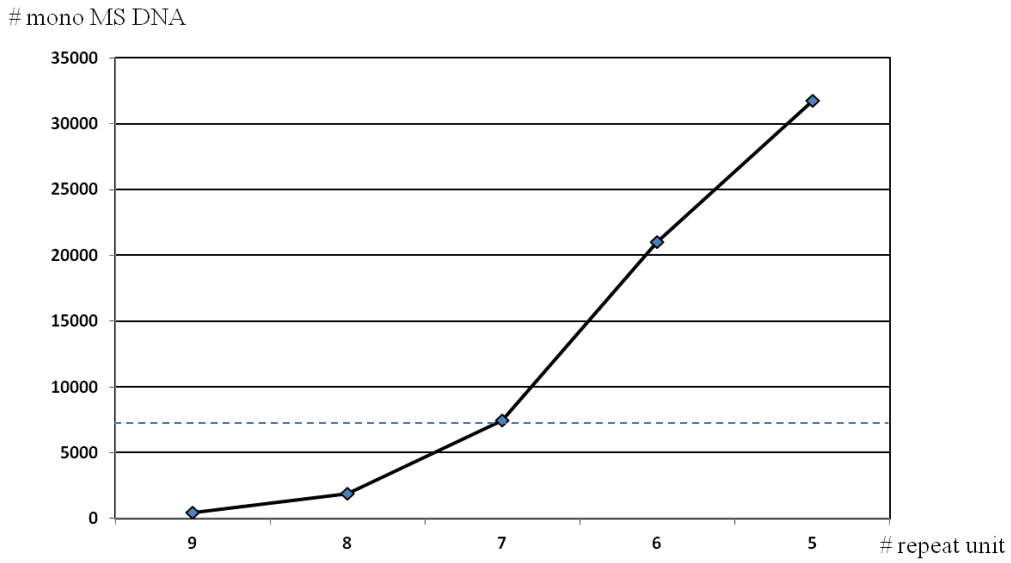


Figure 5.2 The distribution of the number of MS DNA according to the minimum number of repeating unit. We checked the number of coding MS DNA by changing the minimum number of repeat units for being coding MS DNA. 8 as a minimum yielded relatively quite few (1,912) while 6 as a minimum yielded relatively quite many (21,012). Therefore, we selected 7 as minimum repeating for MS DNA, which gave us 7,471 MS DNA.

For the mono MS DNA, there were 4,563 genes (about 23% of total human genes) that contained at least one MS DNA in the coding region. Total number of MS DNA was 10,069 because 1,771 mRNA of 1,203 genes had more than one MS DNA. Seven genes including BRCA2 carry even 10 or more MS DNA in their coding regions (Table 5.1).

Table 5.1 The list of genes that carry 10 or more coding MS DNAs.

Gene	Definition	NCBI RefSeq	Number of coding MS
CCDC168	coiled-coil domain containing 168	NM_001146197.1	14
FSIP2	fibrous sheath interacting protein 2	NM_173651.2	13
CEL	carboxyl ester lipase	NM_001807.3	13
DNAH14	dynein, axonemal, heavy	NM_001373.1	12

	chain 14		
LMTK3	lemur tyrosine kinase 3	NM_001080434.1	10
ANKRD12	ankyrin repeat domain 12	NM_015208.4	10
BRCA2	breast cancer 2, early onset	NM_000059.3	10

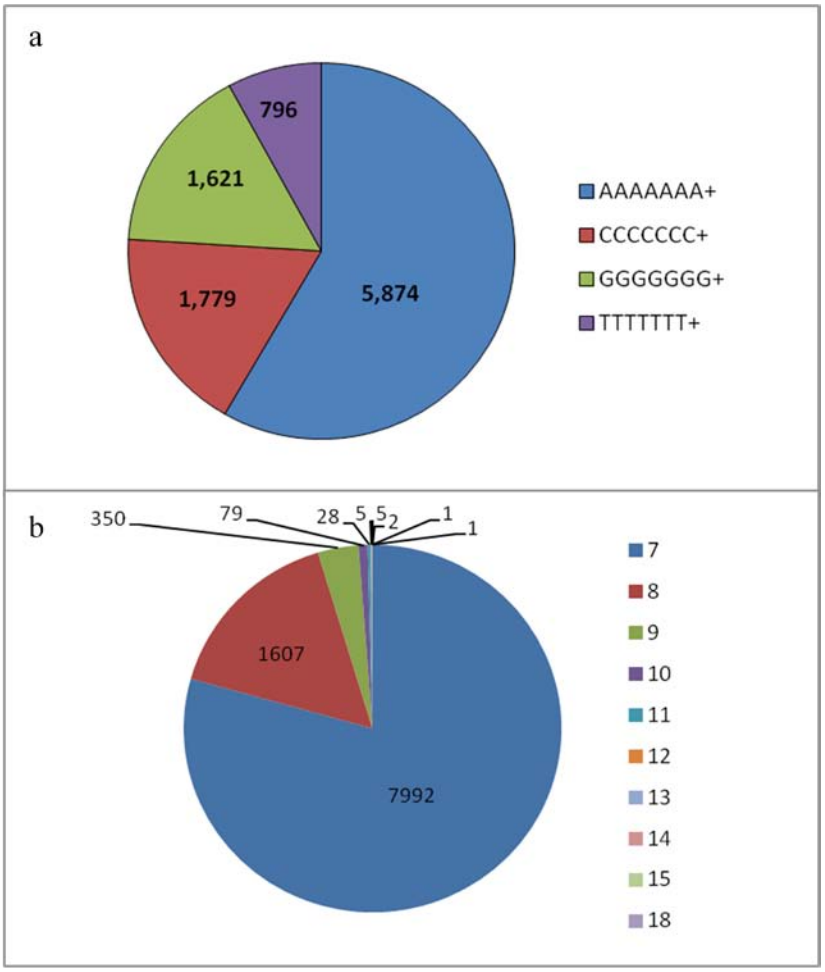


Figure 5.3 The distribution of coding mono MS DNA. (a) The MS DNA with repeating unit of A was the most common while one with T repeating unit is least common. (b) The shorter MS DNAs are more frequent than the longer ones. In fact, 79% were 7 bp and 16% were 8 bp, therefore theses two length covered 95%. The longest is 18 bp.

5.2.2 Algorithm to identify Indels at coding microsatellite DNAs

Basically, we aligned qualified (see method) transcripts with human mRNA reference sequences that have MS DNA in their coding sequences by

using BLASTN without repeat masking option. Therefore, we can align transcript sequences with MS DNAs, which were usually masked due to their low sequence complexity. Among the selected EST sequences, we identified transcript sequences derived from the coding MS DNA of reference sequences according to their coordination. Simply, we identified the alignment of transcript sequences with reference sequences that covered MS DNA with at least 3 bp of both flanking sides by using four numbers; starting and end position of reference sequence in the alignment and starting and end position of reference sequence and end position of coding MS DNA in the reference sequence. Only insertion / deletion (Indels) and substitutions within the MS DNA were counted while any other mutations outside of MS were not counted for this analysis. We counted how many repeat units were added or deleted in the MS DNA according to their alignments. If the observed bases of Indels are different than the bases of repeat unit, we called this as heterogeneous Indels.

Figure 5.4 Selection of ESTs for analysis. ESTs covered entire MS DNA of reference sequences were selected and analyzed. Solid lines indicate qualified ESTs while dotted lines indicate EST that did aligned with entire of MS DNA.

5.3 Results

5.3.1 Identification of putative Indels at coding microsatellite DNA

The alignment between ESTs and RefSeqs were generated by using BLASTN without repeat masking option in order to allow the alignments on MS. The aligned ESTs with the RefSeq were selected when the alignments met our standard (See 5.2.2). 6,078,016 alignments were selected for our analysis because they have clear single origin for the gene. Among these selected alignments, 216,128 EST sequences aligned with MS DNA of genes were derived from mono MS DNA according to their coordination of alignments. Only Indels and substitution in the MS were considered while any other mutations outside of MS were not counted for this analysis. 6,459 coding MS DNA from 2,196 genes (48% of whole genes with coding MS DNA) were aligned with 10 or more supporting EST sequence. A total of 156,244 EST sequences were derived from MS DNA regions and 15,377 of them (9.8%) carried frame-shifted (FS) mutations. For the case of RNA-seq data, we used BLATN program to align qualified reads from three sets of RNA-seq data with their matched mRNA according to BWA alignments. The aligned reads with the RefSeq were selected when the alignments met our standard (See 5.2.2). The average number of coding MS DNA with 10 or more 10 supporting reads were 436.8 for breast cancer data, 395.9 for melanoma data, and 170 for prostate cancer data.

Table 5.2 Indel rate of coding MS DNA. The Indel rate = total bp of Indels / total bp of coding MS DNA in the alignments.

Types	Indel rate	Insertion rate	Deletion rate
Breast	0.00376	0.00287	0.00089
Melanoma	0.00426	0.00226	0.00201
Prostate	0.00147	0.00118	0.00029
EST	0.01331	0.00799	0.00541

Table 5.2 showed the estimated Indel rate from each data set. EST had the highest Indel rate while prostate data had the lowest Indel rate. Most of the Indels (about 97%) observed in RNA-seq data were homogenous Indels while about 49% of Indels counted in EST data were heterogeneous indels. The observed insertions (9,593) outnumbered the observed deletions (5,762) in the EST data. However, homogeneous Indels had no significant difference between insertion and deletion. In most of RNA-seq data from breast and prostate samples, the number of insertion was significantly higher that the number of deletions for both homogenous and heterogeneous. Most of data from melanoma samples did not show any significant difference between insertion and deletion. The size of most Indels is 1 bp (> 90% in all data except deletion (86%) in EST data) In addition, a higher Indel rate was observed in longer microsatellites in general (Figure 5.5).

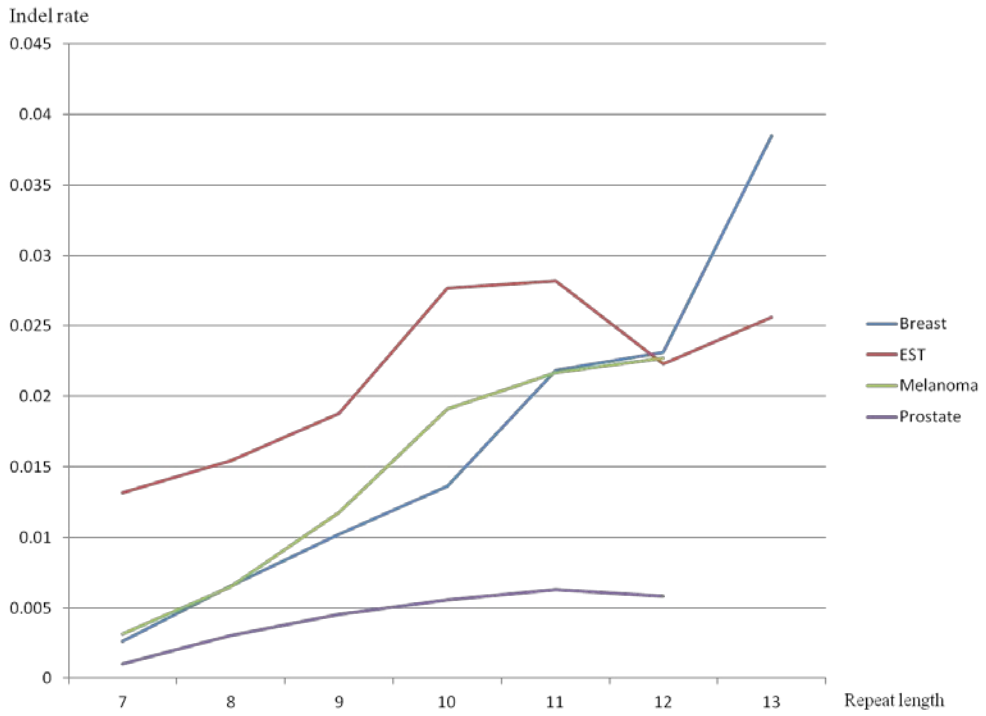


Figure 5.5 The Indel rate by repeat length. In general, a longer microsatellite has higher Indel rate. R^2 values are 0.93, 0.61, 0.95, and 0.86 respectively for breast, EST, melanoma, and prostate.

In EST analysis, each tissue type showed different rate of Indels. Figure 5.5 showed the Indel rate of 11 selected tissue types. The highest Indel rate was observed in bone marrow. Pancreas showed the biggest difference between tumor and normal libraries. Seven tissues types had significantly higher Indel rate in tumor than normal libraries ($p<0.001$); prostate, colon, pancreas, skin, brain, gastrointestinal tract, and stomach. Breast cancer may have a higher Indel occurrence in normal than tumor (significantly by p value = 0.1, but not by 0.05)

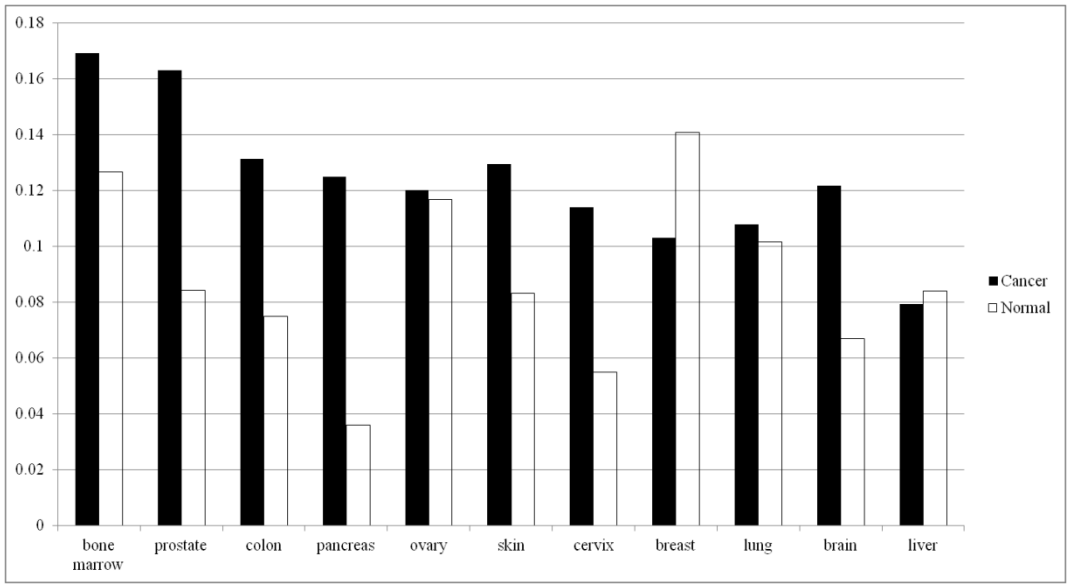


Figure 5.6 Indel rates by tissue types. Each tissue types has different mutation rate. Bone marrow, prostate and colon is in group of the highest mutation rate

5.3.2 Characteristics of Insertion/deletion in coding MS DNA in tumor and normal

Based on the 2x2 contingency table, the occurrence of FS mutations collectively in tumor samples is significantly higher than normal samples ($p < 0.001$, chi-square test) in EST data. Edgren *et al.* screened one normal breast sample and 6 breast cancer cell lines (60). The average number of 6 breast cell lines was used. The occurrence of FS mutations in normal breast was significantly higher than average of 6 cancer cell lines. In Kannan *et al.*, there were matched tumor and normal samples from 10 patients (108). Only 2 out of 10 showed significantly higher rate in tumor than normal.

Table 5.3 The comparison occurrence of frame-shifted mutations in coding microsatellite between tumor and normal. For EST data, we collectively count EST sequences for tumor and normal. The number of reads in tumor was counted from 6 breast cancer cell line and that in normal was counted from one normal breast sample. 10 pair of matched tumor and normal were used for comparison.

Highlighted ones had higher incidents of Indels in tumor than normal significantly ($p<0.01$, chi-square test); EST, breast, prostate #2, and prostate #23.

Data Set	Tumor		Normal	
	WT	FS	WT	FS
EST	60,123	7,461	82,984	6,504
Breast	19,077	446	14,966	827
Prostate #2	4,449	53	3,137	21
Prostate #3	3,434	42	3,057	35
Prostate #6	6,740	78	3,149	32
Prostate #8	3,528	36	1,798	23
Prostate #9	2,928	24	2,024	28
Prostate #11	6,270	70	9,178	83
Prostate #13	13,825	172	1,0574	100
Prostate #15	14,736	139	8,282	63
Prostate #19	14,304	136	7,404	88
Prostate #23	12,511	151	9,243	73

5.3.3 Putative candidates of coding microsatellite DNA

For EST data, we counted tumor and normal collectively since the overall coverage of microsatellite in one library was not high enough. 169 coding microsatellite with more than 5 supporting EST sequences for tumor and normal respectively had a significant difference in occurrence of Indels between tumor and normal. 142 of them had higher Indel rate of tumor than normal while 27 had an opposite trend. 88 frame-shifted peptides derived from Indels at 142 coding microsatellite will be longer than 6 amino acids. The top 10 candidates by chi-square statistic were listed in Table 5.4.

Table 5.4 The list of top 10 coding microsatellite DNA. Among 88 coding MS DNA, this table shows the top 10 MS DNA that has more Indels in tumor than normal libraries according to chi square test.

Gene	Position	MS DNA		FS Len	Chi sqaure	Tumor		Normal	
						WT	FS	WT	FS
TM9SF2	282	A7	FS_del	8	29.67	110	14	253	0
PROL1	558	T7	FS_del	12	24.25	5	2	83	0
VCP	2228	A9	FS_ins	10	17.27	77	31	70	3
EIF3M	519	A7	FS_ins	8	15.65	169	31	197	8
BRD4	982	C7	FS_ins	44	14.75	4	10	13	0
CLDN6	621	G7	FS_ins	23	12.69	15	7	75	4
MFN2	753	T7	FS_del	10	11.79	23	3	149	1
ABR	2630	C7	FS_del	17	11.79	12	4	44	0
DGKZ	355	C7	FS_ins	84	10.47	6	4	23	0
SH3GLB2	1138	C7	FS_del	63	10.37	77	37	29	1

We found six coding MS DNAs that showed an interesting distribution of Indels. When both insertion/deletion at a certain coding MS DNA were observed, most of them are 1 bp of insertion or deletion in general. However, these six MS DNA have a combination of either 1bp insertion and 2 bp deletion or 2 bp insertion or 1 bp deletion. As a result, they will have only one frame-shifted peptide instead of two possible ones (Table 5.5).

Table 5.5 Biased distributions of Indels in terms of their size. Indels at these 6 coding microsatellites had a skewed distribution of size of Indels considering that size of most Indels is 1bp. Highlighted one was generated by Indels among two possible frame-shifted peptides. Longer frame-shifted peptide was selected in the top three MS DNA while shorter frame-shifted peptide was selected in the bottom three MS DNA.

GeneName	Pos MS	MS	# Reads	WT	FS	Ins	List Ins	Del	List del	FS-Del	FS-Ins
ABCF1	313	A_{10}	60	33	23	18	2,2,2,2,1,2,2,1,1,2,2,2,2,1,2,2,2,2	5	1,1,1,1,1	59	13
HELLS	2069	A_7	36	17	11	4	1,1,1,1	7	1,2,2,2,2,1,1	14	24
C1orf144	241	A_8	160	143	12	9	1,1,1,1,1,1,1,1,1	3	2,2,2	1	24
ATF4	1127	A_7	88	65	14	6	1,1,1,1,1,1	8	2,2,2,1,2,1,1,1	35	9
ICA1	798	A_9	7	2	5	1	1	4	2,2,2,2	4	1
SHCBP1L	1057	T_7	6	0	5	0		5	2,2,2,2,1	83	1

We found 14 cases where there was only one microsatellite among multiple microsatellite from the coding region of a gene showed differential frame-shifted occurrence between tumor and normal (Figure 5.7).

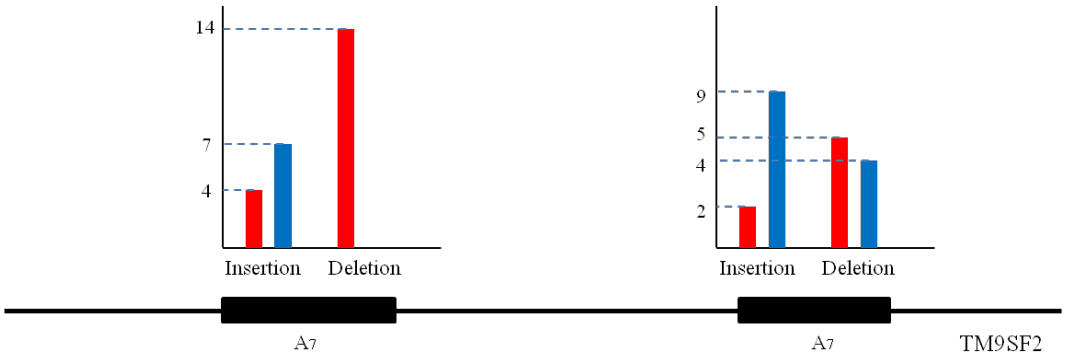


Figure 5.7 Different patterns between multiple microsatellites from a gene. TM9SF2 has 2 microsatellites in the coding sequence. A_7 at position 282 has significantly frequent deletion in tumor than normal while A_7 at position 560 has no significant difference. Red bar indicates tumor while blue bar indicates normal.

For three sets of RNA-seq data, we treated the data set independently since each run had high enough coverage unlike the EST data. 85 coding MS DNAs had significantly higher occurrence than the average at least one of seven breast samples. Some coding MS such as CCT5 had higher incident in 5 out of 7 samples; BT474-1, BT474-2, MCF7, SKBR3-1,SKBR3-2. 144 coding MS DNA had significantly higher occurrence than expected at least one of melanoma cancer samples. 50 of them had higher incidents in multiple samples. In 30 prostate cancer samples, 72 of the coding MS DNAs had significantly higher occurrence than expected at least one of prostate cancer samples. The occurrence of Indels at the 11 coding MS DNA was observed as above the average in all three sets (Table 5.6). We found 4 coding MS DNA that had differential FS

mutation rate between 10 matched tumor and normal from prostate samples (Table 5.7).

Table 5.6 The list of coding MS DNA with high Indel rate. 11 coding MS DNAs showed high Indel rate in all three data sets relative to the average. There were total 40 tumor samples and 11 normal samples from three data sets. In the FS peptides, the first number is the size of a FS peptide from 1bp insertion and second number is the size of a FS peptide from 1 bp deletion. Bold indicates the observed a dominant Indel.

Gene	MS position	MS	# Tumor samples	# Normal samples	FS peptides
RPL22	83	A_8	22	9	**8** / 4
VCP	2228	A_9	20	3	10 / **62**
P4HB	1350	A_8	18	8	**21** / 117
CCT5	891	A_7	12	0	**13** / 25
VEGFB	419	A_8	9	1	**34** / 4
TMBIM4	587	T_{10}	9	0	**2** / 15
PSMA6	648	A_8	8	0	**5** / 0
SEC62	452	A_9	7	1	**8** / 63
HNRNPH1	1038	T_8	6	1	0 / **39**
TCF25	467	A_9	4	1	**26** / 16
SF3B2	2658	A_8	4	1	**15** / 21

Table 5.7 The comparison of the occurrence of Indels at coding MS DNAs between matched tumor and normal. Among the coding MS DNAs with 10 or more supporting reads in both tumor and normal, 4 coding MS DNAs showed differential incidents of Indels between tumor and normal. 2 of them were frequent in tumor while 2 of them were frequent in normal.

Samples	Gene	MS position	MS	Tumor		Normal	
				WT	FS	WT	FS
C03 / N03	RPL22	83	A_8	32	5	36	0
C08 / N08	RPL22	83	A_8	64	5	25	7
C23 / N23	MIF	197	C_7	180	3	414	0
C23 / N23	OR51E2	704	T_8	115	1	16	3

5.4 Methods & Materials

5.4.1 Collection of sequences

Human Reference mRNA sequences were downloaded from NCBI (August 2011 version). This data set contains 32,871 mRNA from 19,763 genes. About 8 million EST sequences were also downloaded from NCBI (December 2010 version). Three sets of RNA-seq data were obtained from Sequence Read Archive (SRA), NCBI. The first set had 7 runs from 6 breast cancer cell lines and one breast normal sample (60). In the second set, there were 14 runs from melanoma patient and cell lines (58). The third set contained 30 runs from 10 matched prostate tumor and normal samples and 10 prostate tumor patients (108).

5.4.2 Selection of qualified sequences

First, all EST sequences were aligned with human mRNA Reference Sequences by BLASTN program. EST sequences that aligned with a single reference sequence or single loci were selected based on similarity and length of alignments and their origin. The similarity and length of alignments has to be ≥50 bp with 97% similarity, ≥85 bp with 95% similarity or ≥100bp with 90% similarity. In addition, they have to be located unambiguously in single reference sequence or loci. The origin of the EST was unequivocal when there was only one alignment with a reference sequences or the blast score of the best alignment was higher than that of second best alignment by at least 50. We excluded non-coding RNA from second best alignment with the only exception when they had the exact same blast score with the best one. After all applied criteria, we selected EST sequences that had the best alignments only from a single gene. The longest

mRNA was selected for further analysis when there were multiple isoforms for a gene.

All reads from RNA-seq data were aligned against human mRNA Reference Sequences by using BWA program version (109). The reads that had mapping quality score is 30 or higher and aligned with coding MS DNA were selected. These reads were aligned again with matched human RNA by using BLASTN.

5.4.3 Selection of coding MS DNA with higher rate of Indels

The average occurrence of frame-shifted mutations from all runs in a set was calculated at first. Simply, we counted all Indels collectively in each study. Afterward, we identified the coding microsatellite DNAs with higher occurrence in each sample by comparing with the average of Indel. The chi-square test was used to assess statistically significance.

5.5 Discussion

The question addressed by this study was whether Indels at coding MS DNA are a good source of antigens for a cancer vaccine. Due to their high mutation rate, we may expect to observe frequent Indels at MS DNA in tumor samples. Therefore, we are interested in identifying what coding MS DNAs have frequently Indels in many tumor samples, but not in normal samples. Our analysis showed that some of coding MS DNA have higher Indel rate in multiple tumor samples, but not or few in normal samples (Table 5.4 and Table 5.6). These coding MS DNA could generate frame-shifted peptides that could be used as antigens for a cancer vaccine.

I presume the chance of having a sequencing error will be same in tumor and normal samples. Therefore, we can tell that Indels at a coding MS DNA more frequent occurred in tumor than normal when more Indels were observed in transcriptome data from tumor samples comparing to normal samples. Since we required the minimum number (6 for EST and 10 for RNA-Seq) of supporting reads respectively from tumor and normal in order to do chi-square test, we may miss some good candidates of coding MS DNA that had no supporting reads from normal samples.

I expect that more Indels at coding MS DNAs will be observed in tumor when these Indels are associated with cancer. Therefore, the coding MS DNA in Table5.6 might be a good candidate since the biased distribution of Indels might indicate their association with cancer development. I speculate that this might be oncogenic function because either insertion or deletion can truncate the protein, so there is no necessity for one dominant way of doing it. However, 3 out of 6 produced a short peptide instead of long one. It is hard to say that 1 amino acid confer a new function.

In summary, this study has shown the potential use of coding MS DNA as vaccine antigens by analyzing transcriptome data and antibody reactions. The pipeline of analyzing transcriptome for coding MS DNA should be easier and more accurate when we have more data generated by next-generation sequencing technology. Therefore, the systematic analysis shown in this study may provide more reliable coding MS DNA that could be tested in animal model when more sequencing data are available.

5.6 Conclusion

Coding MS DNAs may be a good source of FS antigens considering their high Indel rate and functionality in the context of cancer. Furthermore, the genetic instability of cancer elicits more mutations in MS DNAs. Their mutation rate is the highest among spontaneous mutations. The analysis of EST and RNA-seq in this study supports the feasibility of this idea. The low coverage of coding MS DNA in the current transcriptome will be improved by targeted sequencing. Therefore, we can select better coding MS DNA for antigen candidates.

CHAPTER 6

FRAMESHIFTED ALTERNATIVE SPLICING VARIANTS

6.1 Introduction

Alternative splicing, a well-studied event in eukaryotes, increases the diversity of proteins with critical roles in regulation of cells. In fact, alternative splicing is a highly controlled procedure and a critical process that produces significant impact in the regulatory and developmental biology of organisms (110). In humans, 92~94% of total genes have multiple isoforms generated by alternative splicing, and a large number of them are tissue-specific variants (48). As anticipated, abnormal splicing variants derived from mis-regulation in splicing mechanisms are also implicated in cancer. While tumor suppressors are often inactivated by splicing in cancer cells, oncogenes are often activated by this process (111). In fact, several studies reported tumor-specific alternative splicing that had not been detected in normal tissues (112-114).

Abnormal splicing variants in cancer have been tested as a potential source of biomarkers, diagnostic, prognostic, and therapeutic targets for cancer. Many studies have shown splicing variants are either tumor-specific or tumor-associated. For instance, alternatively spliced NF1 in neurofibroma (115), variable CD44 in breast cancer (116), truncated DNMT3B in non-small cell lung cancer (117), aberrant KLF6 in prostate, colon, and lung cancers (118-120), and isoform Ron in breast and colon (121) have been shown to be associated with tumors. As for biomarkers, 41 splicing variants were listed as potential markers

for breast cancer by Venables *et al.*(50), and 48 alternative iso-forms were suggested markers for ovarian cancer by Klinck *et al.*(49). In addition, several splicing variants have been used as prognostic indicators: RHAMM and HAS1 for multiplemyeoloma (122, 123), survivin2B for metastatic gastric, breast and colorectal cancers (124-126), and CD44v6 for prostate cancer (127).

Frame-shifted splicing variants, which generate frame-shifted peptides out of new exon combinations, have not been studied extensively. We speculated that frame-shifted alternative splicing may contribute to cancer development since frame-shifted mutations affect protein sequences dramatically. Truncated proteins by frame-shifted splicing may result in loss of functional domain, which could inactivate the pathways of tumor suppressors. These frame-shifted peptides could be good cancer antigens if they occur frequently in cancer samples. To test the feasibility of this idea, we did a bioinformatics analysis on the transcriptome data to get the frequency of each identified frame-shifted splicing variant. Translated peptides and potential epitopes were then able to be accurately predicted. The bioinformatics analysis in this study provided us a list of putative frame-shifted splicing variants as candidates for cancer antigens. It is anticipated that the vast amount of transcriptome data of various cancers from next-generation sequencing will enhance our ability to further select better cancer antigen candidates derived from splicing variants.

6.2 Bioinformatic Approach

6.2.1 Data Sets

To identify potential putative neo splicing variants, that when translated would result in a frame-shifted neo-peptide, two publically available datasets were applied in an algorithm that was used to identify splicing variants. Specifically, we used the sequences found within the Expressed Sequence Taq (EST) library (51) and the Human RefSeq database (76) from the National Center for Biotechnology Information (NCBI).

6.2.2 Algorithm

Using the stand-alone BLAST program, we aligned all EST sequences to RefSeq. We selected the alignments of EST sequences with RefSeq when they have met one of following conditions: ≥50 bp of length with 97% or more of sequence similarity, ≥85 bp of length with 95% or higher sequence similarity, or ≥100 bp of length with 90% or higher sequence similarity. There were some ESTs that had more than one qualified alignment that were derived from different regions of the EST. Due to local alignment by BLAST, splicing variants will generate separate alignments. As a default setting of BLAST, top alignments will be the best alignments by BLAST score. Any qualified alignments outside of the top alignments by at least 50 bp were identified. Based on this, we were able to count the number of distinctive regions that were aligned in an EST sequence.

To simplify the analysis, we analyzed only ESTs that had two matching alignments from a single RefSeq, instead of analyzing ones with three or more

alignments. Only ESTs of which had all qualified alignments from a single RefSeq were considered to be potential candidates. When both aligned regions in an EST each has multiple matching alignments with different RefSeqs, the common matching RefSeq in both regions was selected. If there were more than one common RefSeqs, the more frequently matched one was selected. There are four combinations of splicing variants in terms of the direction of two alignments; ++, +−, −+, and −−. The subject sequence, or the RefSeq, can be aligned from 5' to 3' (+) or from 3' to 5' (−), relative to the query sequence, or the EST. For the cases of−−, their orientations were reversed to ++. In addition, −+ EST sequences were excluded from further analysis because the leading reverse strands of RefSeqs could not be properly translated into peptides. Furthermore, only NCBI accession numbers that begin with the prefix NM_ were used due to their precise exon boundaries. Finally, any nucleotides that fall within both matched alignments in an EST were not counted more than once. The position of the downstream alignment in the EST sequence, along with its corresponding position in the Refseq, was shifted by the size of the overlapped region accordingly. To eliminate false calls, all identified novel splicing variants had to be supported by EST sequences from more than one cDNA library. Additionally, novel splicing variants that occurred at exon boundaries were counted even though all supported EST sequences were from only one library.

The adjusted alignments of an EST sequence with a RefSeq showed four different types of splicing variants; exon inversion, exon skipping, intron retention and a combination of exon skipping and intron retention (Figure 6.1). We only

considered exon skipping, because the entire configuration of exons or the exact sequences of transcripts for the other three cases could not be computed based on current data. For exon skipping, we used EST sequences only for identifying exon junctions. Sequences of splicing transcripts were then determined by using RefSeq in order to predict peptide translation.

RefSeq 5'		RefSeq 3'		EST 5'		EST 3'	
Start	End	Start	End	Start	End	Start	End
A	B	C	D	E	F	G	H

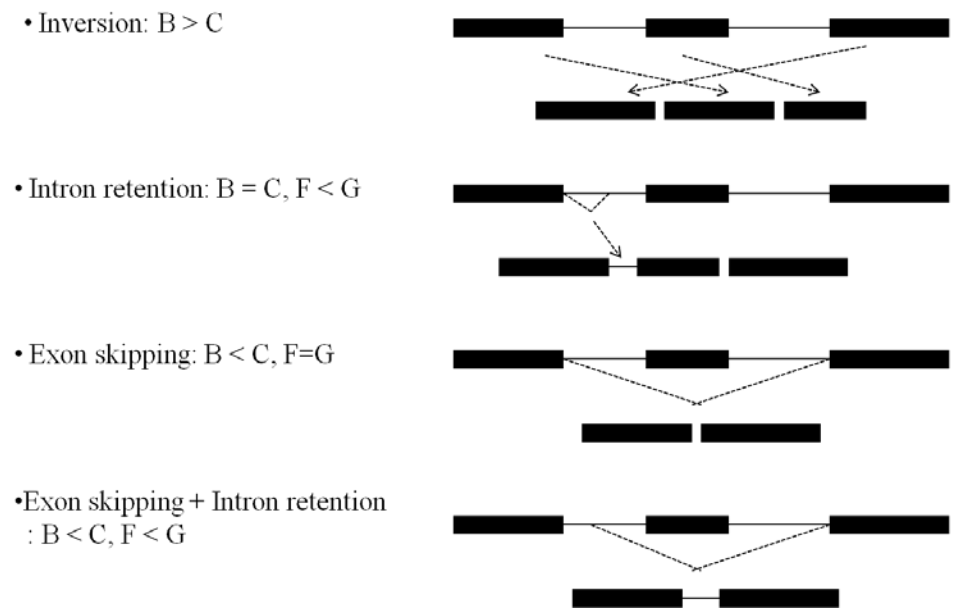


Figure 6.1 Types of splicing variants by aligned positions. Four positions (B, C, F, and G) were used to determine the types of splicing variants.

Among selected splicing variants, we identified those that were a result from frame-shifted mutation. Splicing variants without either a start codon or in-frame splicing variants were excluded. Furthermore, we also excluded frame-shifted splicing variants with neo-peptides consisting of less than 8 amino acids. Finally, we examined whether they were tumor-associated or not by counting the number of tumor and normal libraries in which matching ESTs were detected.

Splicing variants were regarded as tumor-associated only when they meet the following conditions: i) not present in normal libraries, and present in at least 3 Tumor libraries; ii) the occurrence in tumor libraries is at least 3 times higher than that in normal libraries.

6.3 Results

6.3.1 Identification of novel alternative splicing

We used our semi-automatic alignment algorithm to identify frame-shifted alternative splicing variants from the available NCBI EST sequence database (Figure 6.2). Briefly, to support a splicing variant, one EST sequence must have two alignments with a RefSeqs. Considering the EST database contains approximately 8M EST sequences, we outlined filtering criteria that was applied to eliminate irrelevant sequences. Among qualified ESTs, 193,849 EST sequences had multiples alignments at distinctive positions with a RefSeq. To simply analysis, 216,218 EST sequences that had two aligned regions were selected for further analysis. In addition, we discarded the EST sequences that did not align properly with well annotated RefSeqs with accession number that begin with the prefix NM_. After removing 4,179 EST sequences with 3' to 5' of upstream RefSeq and 5' to 3' of downstream RefSeq, we identified 19,121 novel splicing variants supported by these EST sequences.

Novel splicing variants were classified into four types based on positions of two alignments. First, 389 variants had inversed order of exons. Second, 12,456 variants skipped some exons. Third, 6,726 gained addition sequences from

intron with/without exon skipping. To have precisely predicted peptides, only 12,456 exon skipping variants were considered. By our criteria, putative novel splicing variants had to meet two conditions; supporting EST sequences from more than one library or exon-exon combination in one library. Finally, 9,088 variants were supported by EST sequences from two or more libraries. 571 variants occurred at the exact exon boundaries in one library. A total of 9,659 qualified variants were identified as exon skipping variants.

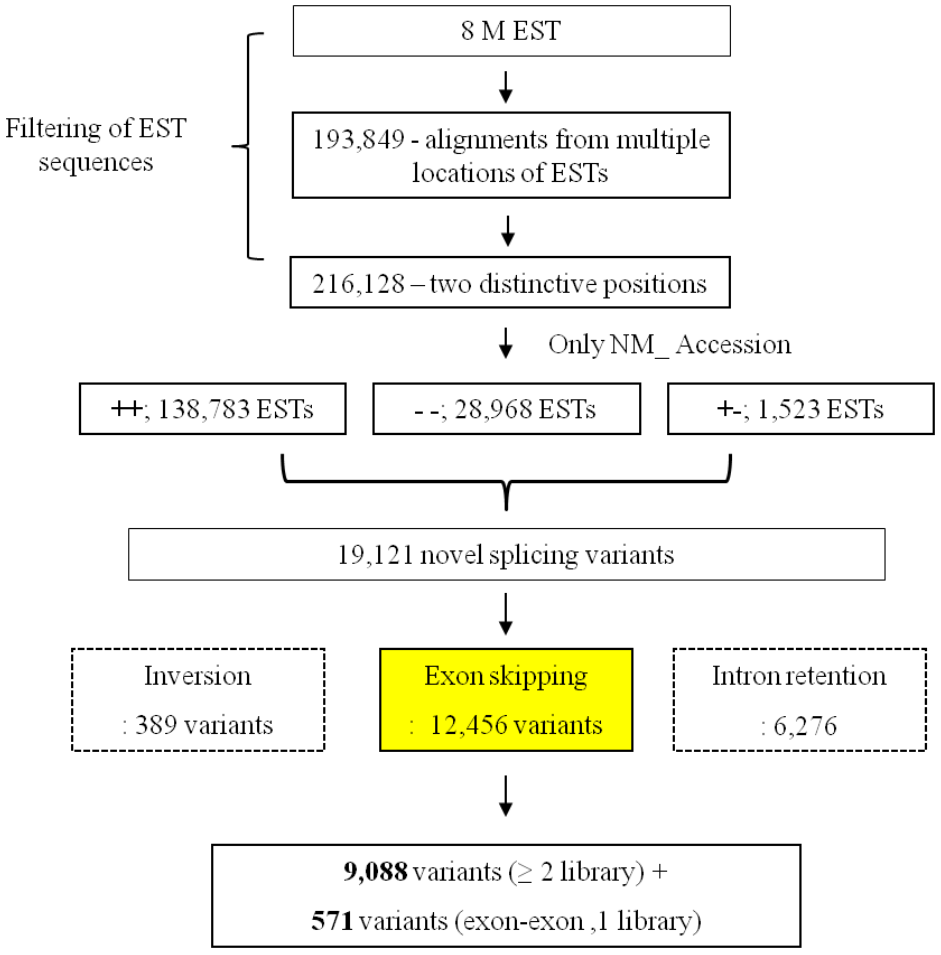


Figure 6.2 Identification of novel splicing variants.

6.3.2 Putative tumor-associated splicing frame-shifted variants

Figure 6.3 shows the brief scheme to identify tumor-associated frame-shifted splicing variants. Among 9,659 putative novel splicing variants, frame-shifted peptides will be translated from 4,506 variants. 2,996 of them will have frame-shifted peptides with 8 or longer amino acids. According to our criteria, total 96 tumor-associated frame-shifted variants were identified (Table 6.1). The average length of frame-shifted peptide is 29.4 amino acids with the range of 8 amino acids to 167 amino acids. 34 of them had exon-exon combination while new junction of remaining 62 variants occurred in the middle of known exons.

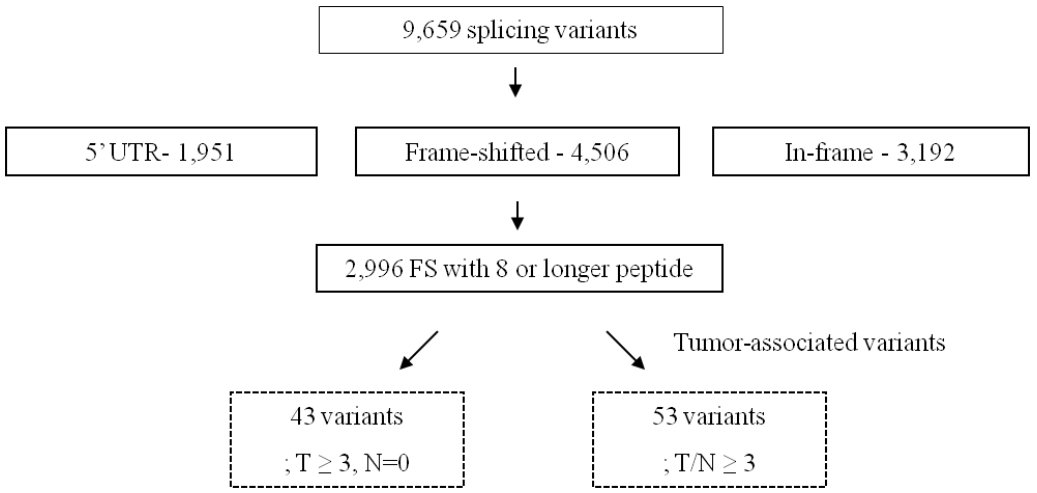


Figure 6.3 Identification of tumor-associated frame-shift splicing variants. Among selected 9,659 splicing variants as qualified ones, 4,506 (46.7%) could generate frame-shift peptides. We only considered the 2,996 splicing variants with 8 or longer amino acids for higher chance of having possible epitopes. 96 splicing were regarded as cancer-associated ones.

Table 6.1 Putative tumor-associated splicing variants. This table shows 20 out of 96 candidates including two cancer genes (indicated by *) by Sanger Inst. RefSeq_ID is NCBI accession number. FS length means the length of frame-

shifted peptides from splicing variants. #Tumor_lib and #Normal_lib indicates the number of libraries to support each splicing variants.

Gene	RefSeq_ID	FS length	# Tumor_lib	Tissue types	# Normal_lib	Tissue types
C11orf2	NM_013265.2	48	4	uncharacterized tissue,cervix,colon	0	
C20orf96	NM_080571.1	9	4	uterus,uncharacterized tissue,lung	0	
CYBASC3	NM_001161452.1	28	4	ovary,lung,colon,placenta	0	
KRT8	NM_002273.3	50	4	ovary,uterus,pancreas	0	
MVK	NM_001114185.1	8	4	uncharacterized tissue,prostate,testis,placenta	0	
NAA10	NM_003491.2	18	4	uncharacterized tissue,skin,cervix,placenta	0	
PDCD2	NM_001199462.1	32	4	ovary,brain,lung,colon	0	
RPS3A	NM_001006.3	9	4	ovary,bone,liver,mammary gland	0	
TFE3*	NM_006521.4	24	3	uterus,skin,placenta	0	
HNRNPA2B1*	NM_031243.2	21	3	uterus,mammary gland,lung	0	
NOL12	NM_024313.2	14	9	uterus,kidney,mammary gland,salivary gland,uncharacterized tissue,skin,lymphoreticular,lung,testis	1	uncharacterized tissue
RPLP0	NM_001002.3	10	7	pancreas,lymphoreticular,prostate,lung,testis,muscle,eye	1	embryonic tissue
DPH2	NM_001384.4	61	6	uterus,kidney,skin,lymphoreticular,brain	1	embryonic tissue
GNB2L1	NM_006098.4	31	5	bone,pancreas,uncharacterized tissue,lung,testis	1	pancreatic islet
RPL8	NM_000973.3	12	5	ovary,lymph node,prostate,muscle,eye	1	skin
IGFLR1	NM_024660.2	22	4	skin,lymphoreticular,lung,eye	1	pooled tissue
KARS	NM_001130089.1	22	4	skin,lung,uterus,cervix	1	brain
MRPS28	NM_014018.2	15	4	parathyroid,uncharacterized tissue,lung,testis	1	mammary gland
HNRNPA2B1*	NM_031243.2	20	3	mammary gland,uncharacterized tissue,prostate	1	pooled tissue
SMC1A	NM_006306.2	17	3	mammary gland,skin,eye	1	liver

6.3.3 Experimental validation

To validate the presence of these predicted splicing variants, we screened several cancer cell lines by RT-PCR. We amplified both wild type and alternative splicing products by using primers we designed (Figure 6.3). RNA samples from 5 different cancer cell lines were used; panc1 (pancreatic cancer), brain (brain cancer), A-459 (lung cancer), SW-480 (colon cancer), and MCF7 (breast cancer). The expected size of band was confirmed by sequencing. Protein phosphatase 4, catalytic subunit (PPP4C) had very faint splicing variants in all cancer types. Expected splicing variant of member RAS oncogene family (RAB34) were detected in pancreatic, breast, and lung cancer cell lines, but not in colon and breast cancer cell lines. Prune homolog (Drosophila) (PRUNE) and mitogen-

activated protein kinase kinase kinase 10 (MAP3K10) were amplified the product of splicing variant in all cancer types that we examined.

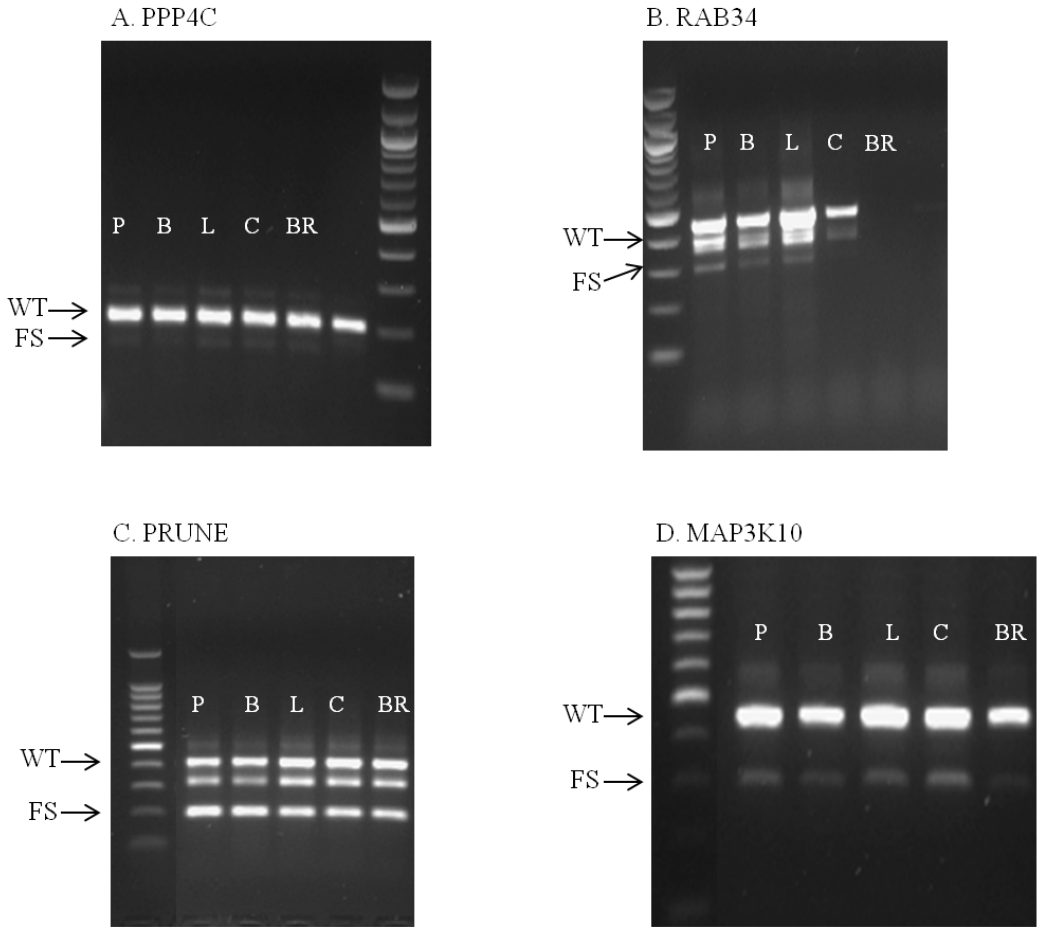


Figure 6.4 Experimental validation using RT-PCR. P, B, L, C, BR denote pancreatic tumor, brain tumor, lung tumor, colon tumor, breast tumor respectively. Wild type and frame-shifted products are indicated. A. PPP4C; 207 bp for wild type, 154 bp for alternative splicing. B. RAB34; 463 bp of wild type, 371 bp for splicing variant. C. PRUNE; 374 bp for wild type, 171 bp for alternative splicing. D. MAP3K10; 412 bp for wild type, 216 bp for splicing variants.

6.3.4 The example case; SMC1

These data were generated by Luhui Shen. SMC1 was one of our putative candidates, which resulted in producing 17 amino acids of neo-peptide. In fact,

our team had preliminary results based on this SMC1. The frame-shifted variants were detected in both tumor and normal samples with differential expression. However, the tumor growth was clearly delayed in the mice vaccinated with the FS vaccine relative to controls (data not shown).

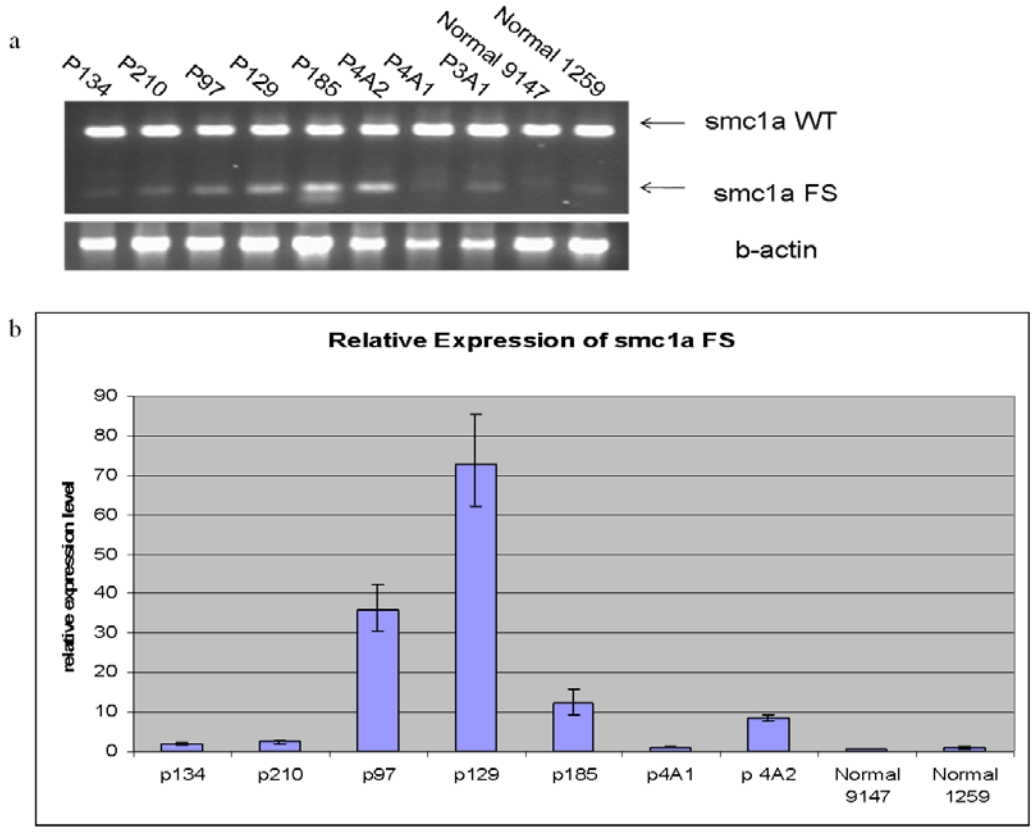


Figure 6.5 The frame-shifted splicing variants of SMC1. a. RT-PCR screening of SMC1 transcripts in primary breast tumor samples. b. The relative expression level of splicing variants. Luhui Shen provided figures.

6.4 Methods

6.4.1 Computational analysis

Refer to section 6.2.

6.4.2 Experimental Validation

Primers were designed by a program, called "Primer3" (87) to amplify both wild type and splicing variants. The sequences of primer pairs were shown in Table 6.2.

Table 6.2 The sequences of primer pairs for RT-PCR. FW, Rev, WT, and AS denote forward primer, reverse primer, product size of wild type, and product size of alternative splicing respectively.

Gene	FW	Rev	WT	AS
RAB34	GCGGTGGTCGTAGCGTCTC	CAGCACCTCAAATCGTTCCATC	463	371
PPP4C	TCATCAAGGAGAGCGAAGTC	AGCCACGGTCCACAAAGTC	207	154
PRUNE	GAAGCCTGTGATTTGGACTC	AGCACAGGACCCCACCAG	374	171
MAP3K10	CACAAGACCACCAAGATGAGC	TGGTCCGAAGGTCATCAAAC	412	216

Total RNA was extracted from cancer cell lines using the TRIzol LS reagent (Life Technologies, Carlsbad, CA) following the manufacturers protocol. cDNA was prepared by using the SuperScript^{TM} III First-Strand Synthesis SuperMix (Life Technologies, Carlsbad, CA) that includes random hexamers and oligo dT's following the manufacturer's recommended protocol. cDNA integrity and quality were assessed by performing a β-actin control PCR. PCR reactions were carried out using approximately 25 ng of cDNA, reagents from (Life Technologies, Carlsbad, CA) and 35 cycles were performed using Mastercycler ep gradient S (Eppendorf, Hamburg, Germany). Amplification conditions were as follows; 95°C for 2 min; 35 cycles of 95°C for 30 sec, 55°C for 20 sec, 72°C for 20 sec. PCR products were analyzed on 1.5% agarose gels. PCR products were purified and sequence confirmed by Applied Biosystems 3730 (Life Technologies, Carlsbad, CA).

6.5 Discussion

The question addressed by this study was whether frame-shifted splicing variants are a good source of antigens for a cancer vaccine. In this study, we identified frame-shifted splicing variants and validated them in cancer cell lines by RT-PCR. Some of frame-shifted transcripts may not be subject to non-sense mediated decay (NMD) as observed in this study, and then may be translated into peptides. At least, the evidence at the mRNA level supports the potential use of frame-shifted splicing variants as vaccine antigens. In fact, frame-shifted splicing

In fact, several frame-shifted splicing variants were identified as cancer specific marker. 9 frame-shifted splicing variants significantly differed in breast tumors compared to normal breast tissues (50). In addition, we may not identify alternative first exon (AFE) and alternative last exon (ALE) since we aligned EST sequences with mRNA that could not provide the information about a neo first/last exon. In addition, several studies reported truncated proteins by splicing in cancers; A-Raf, VEGFR, BCR-ABL, JAK2, and TrkB (128).

Eight different alternative splicing events were suggested by Wang *et al (48)*. However, we used only exon-skipping to simplify the analysis. Splicing events generated from non-coding regions such as 5' UTR or 3' UTR were ignored in this study because we are interested in frame-shifted alternative splicing, which had to occur within an exon.

The function of these truncated proteins has not been clearly understood. However, for vaccine development, we are more interested in their frequency and immunogenicity. We examined the presence of these frame-shifted splicing

variants in normal samples. In fact, most of these splicing variants were detected in normal samples even though the expression level was significantly different between tumor and normal sample. However, this difference in the amount of transcripts may make a difference at the protein level. As a proof of principle, the FS SMC1 that is detected in both tumor and normal tissues showed the delay of tumor growth in mouse model (data not shown) according to our immunology team. Therefore, tumor-specificity of splicing variants could not be determined precisely at the transcript level.

We can expect that RNA-seq data generated by next-generation sequencing technology enable us to identify more slicing variants with higher accuracy. Sequencing biased toward 5' or 3' end in EST sequences will be diminished in RNA-seq data. More splicing variants involved with middle exons will be identified. The systematic bioinformatics approach suggested in this study will guide us to extract the useful information about frame-shifted splicing variants.

6.6 Conclusion

It is worth testing the potential of frame-shifted splicing variants as cancer vaccine antigens. Our studies showed that about half of aberrant splicing in cancer samples was frame-shifted, which is consistent to the result from Venables *et al.* (50) for their maker for breast cancer. Some of them were detected in multiple samples. However, all the frame-shift splicing variants that we tested were found in normal samples as well. The difference in expression level of transcripts may

result in the presence of frame-shifted peptides in tumor, but not in normal. Therefore, we need to have follow-up immunological experiments to validate our candidates. However, our approach can provide the list of probable candidates that could work as antigens by using the information embedded in transcripts.

CHAPTER 7

CONCLUSION

We started by asking whether frame-shifted mutations, if any, can be used as antigens for a prophylactic cancer vaccine because, in general, these neo-peptides from frame-shifted mutations yield more epitopes compared to one new amino acid from point mutations in general. And then we asked what kind of frame-shifted mutations could qualify as suitable antigens. Each chapter in this dissertation explored the possibility of different mutation types that resulted in frame-shifted mutations through the use of cancer transcriptome. The Expressed sequence tag (EST) sequences deposited into National Center for Biotechnology and Information (NCBI) and sequences of transcripts generated from next-generation sequencing technology enables us to retrieve the information about frame-shifted mutations in cancer as well as normal samples. Their frequencies, tumor-specificity, and number of possible epitopes were also obtained by our bioinformatic approach. The potential use of frame-shifted mutations as cancer vaccine antigens was mainly evaluated by means of this acquired data. Amongst the vast extent of all possible frame-shifted peptides derived from coding sequences in the human genome, this evaluation may guide us in narrowing down the possibilities to a specific list of frame-shifted mutations, the targeted probable potential candidates, to be tested in animal models for immunogenicity.

This study has shown that frame-shifted mutations have a high chance of being appropriate antigens for prophylactic cancer vaccine at the mRNA level. Considering the cost of evaluating antigens in animal models, the proposed

pipeline of analysis using transcriptome data in this study readily provides a list of highly probable candidates to be tested in animal models.

Since all candidates were deduced from the mRNA level, follow-up experiments are required. These candidates may, on the other hand, even deem to be unsuitable candidates. Nevertheless, it is worth pursuing the possibility as to whether frame-shifted mutation may be used as vaccine antigens. Therefore, the candidates are ranked on our ranking system. The cancer transcriptome data will enable us to achieve more accurate ranking in the future. Furthermore, any research group can implement this list with the ranking system to select the best candidates for their follow-up experiments.

7.1 Ranking system

Essentially, we use the population coverage of each antigen as the primary contributing score for the ranking. The population coverage of an antigens is determined by two factors; the frequency of mutation and coverage of possible epitopes. Additionally, we provide two more pieces of information along with ranks; normal samples and homologous genes in animal models such as mice and dogs. In fact, the presence of antigens in normal samples is very critical information for vaccine antigens as well as other applications such as diagnostics and drug targets. However, we cannot decide their presence as peptides (antigens) in normal samples by using only transcriptome data. From our studies, most of aberrant transcripts in cancer samples were also found in normal samples even though their expression level was low. It is difficult to resolve whether these low-level transcripts will be translated into peptides or not. Therefore, this

determination is left for further experimentation. The information about homologous genes will be useful for researchers who are working on animal models. Finally, we are able to make a ranking table for all the candidates from viral sequences, chimeric transcripts, coding MS DNA, and splicing variants (Table 7.1).

Table 7.1 The ranking table. The score of each FS mutation will be calculated as follows; h = f * g where f indicates total frequency of mutation and g indicates the epitope coverage. First the total frequency of mutations (f) will be calculated based on the prevalence of each cancer type. The frequency of each cancer type was obtained from "Cancer Statistics 2011" (1). The formula for calculating the total frequency therefore is the sum of the frequency of each cancer type multiplied by the frequency of mutations in that cancer type (a, b, c, d, e): f = 0.14*a + 0.15*b + 0.14*c + 0.03*d + 0.09*e where a, b, c, d, and e refer to breast, prostate, lung, pancreas, and colon cancer types respectively. The frequency of mutation in each tissue type can be determined from transcriptome data by our bioinformatics approach. Second, epitope coverage (g) will be determined by using the algorithm from the Immune Epitope Database and Analysis Resource (IEDB). The rank of each mutation is determined by this score (f). Additional information will be included in the table which does not use a calculated score. The normal column of this additional data indicates whether mutations were detected in normal samples. The columns "Mouse" and "Dog" refer to the presence of homologous genes in each species respectively. Antibody reactivity against predicted peptides may be provided by an immunosignaturing.

Rank	Gene	Peptide	Class	Frequency						Epitope coverage	Score	Normal	Mouse	Dog	Antibody
				Breast	Prostate	Lung	Pancreas	Colon	Total						
1				a	b	c	d	e	f	g	h				

7.2 Future directions

First, the vast amount of transcriptome data from tumor and normal samples will help us to attain greater accuracy of the frequency of aberrant transcripts in cancer as well as normal samples. Currently, a deficit in screening of normal samples hinders us to predict their presence in normal samples. These current limitations will be lifted by the vast amount of data from various tissue

types and normal samples in the near future. Second, immunosignauring technology, which is under development at CIM at the Biodesign Institute, can provide us information regarding whether each antigen is reactive with antibodies from samples of normal and tumor samples. We expect that our ranking table approach will maximize the use of cancer transcriptome data to obtain a relevant list of neo tumor antigens for development of cancer vaccine.

REFERENCES

1. Siegel R, Ward E, Brawley O, Jemal A. Cancer statistics, 2011. CA: A Cancer Journal for Clinicians 2011;61(4):212-36.

2. Finn OJ. Cancer vaccines: between the idea and the reality. Nat Rev Immunol 2003;3(8):630-41.

3. Emens LA, Jaffee EM. Toward a breast cancer vaccine: work in progress. Oncology (Williston Park) 2003;17(9):1200-11; discussion 14, 17-8.

4. Birkeland SA, Storm HH, Lamm LU, *et al.* Cancer risk after renal transplantation in the Nordic countries, 1964-1986. Int J Cancer 1995;60(2):183-9.

5. Dighe AS, Richards E, Old LJ, Schreiber RD. Enhanced in vivo growth and resistance to rejection of tumor cells expressing dominant negative IFN gamma receptors. Immunity 1994;1(6):447-56.

6. Penn I. Sarcomas in organ allograft recipients. Transplantation 1995;60(12):1485-91.

7. Henderson RA, Finn OJ. Human tumor antigens are ready to fly. Adv Immunol 1996;62:217-56.

8. Russell JH, Ley TJ. Lymphocyte-mediated cytotoxicity. Annu Rev Immunol 2002;20:323-70.

9. Smyth MJ, Thia KY, Street SE, *et al.* Differential tumor surveillance by natural killer (NK) and NKT cells. J Exp Med 2000;191(4):661-8.

10. van den Broek ME, Kagi D, Ossendorp F, *et al.* Decreased tumor surveillance in perforin-deficient mice. J Exp Med 1996;184(5):1781-90.

11. Renno T, Lebecque S, Renard N, Saeland S, Vicari A. What's new in the field of cancer vaccines? Cell Mol Life Sci 2003;60(7):1296-310.

12. Sussman HE. Personalized cancer vaccine promises remission. Drug Discov Today 2003;8(15):657-8.

13. Sobol RE. The rationale for prophylactic cancer vaccines and need for a paradigm shift. Cancer Gene Ther 2006;13(8):725-31.

14. Dunn GP, Bruce AT, Ikeda H, Old LJ, Schreiber RD. Cancer immunoediting: from immunosurveillance to tumor escape. Nat Immunol 2002;3(11):991-8.

15. Finn OJ, Forni G. Prophylactic cancer vaccines. Current Opinion in Immunology 2002;14(2):172-7.

16. Heimberger AB, Archer GE, Crotty LE, *et al.* Dendritic cells pulsed with a tumor-specific peptide induce long-lasting immunity and are effective against murine intracerebral melanoma. Neurosurgery 2002;50(1):158-64; discussion 64-6.

17. Pupa SM, Invernizzi AM, Forti S, *et al.* Prevention of spontaneous neu-expressing mammary tumor development in mice transgenic for rat proto-neu by DNA vaccination. Gene Ther 2001;8(1):75-9.

18. Scanlan MJ, Gure AO, Jungbluth AA, Old LJ, Chen YT. Cancer/testis antigens: an expanding family of targets for cancer immunotherapy. Immunol Rev 2002;188:22-32.

19. Soares MM, Mehta V, Finn OJ. Three different vaccines based on the 140-amino acid MUC1 peptide with seven tandemly repeated tumor-specific epitopes elicit distinct immune effector mechanisms in wild-type versus MUC1-transgenic mice with different potential for tumor rejection. J Immunol 2001;166(11):6555-63.

20. Van Der Bruggen P, Zhang Y, Chaux P, *et al.* Tumor-specific shared antigenic peptides recognized by human T cells. Immunol Rev 2002;188:51-64.

21. Bullock TN, Patterson AE, Franlin LL, Notidis E, Eisenlohr LC. Initiation codon scanthrough versus termination codon readthrough demonstrates strong potential for major histocompatibility complex class I-restricted cryptic epitope expression. J Exp Med 1997;186(7):1051-8.

22. Uenaka A, Hirano Y, Hata H, *et al.* Cryptic CTL epitope on a murine sarcoma Meth A generated by exon extension as a novel mechanism. J Immunol 2003;170(9):4862-8.

23. Cheever MA, Allison JP, Ferris AS, *et al.* The prioritization of cancer antigens: a national cancer institute pilot project for the acceleration of translational research. Clin Cancer Res 2009;15(17):5323-37.

24. Cohen AD, Shoenfeld Y. Vaccine-induced autoimmunity. J Autoimmun 1996;9(6):699-703.

25. Ludewig B, Ochsenbein AF, Odermatt B, Paulin D, Hengartner H, Zinkernagel RM. Immunotherapy with dendritic cells directed against tumor antigens shared with normal host cells results in severe autoimmune disease. J Exp Med 2000;191(5):795-804.

26. Overwijk WW, Lee DS, Surman DR, *et al.* Vaccination with a recombinant vaccinia virus encoding a "self" antigen induces autoimmune vitiligo and tumor cell destruction in mice: requirement for CD4(+) T lymphocytes. Proc Natl Acad Sci U S A 1999;96(6):2982-7.

27. Overwijk WW, Theoret MR, Finkelstein SE, *et al.* Tumor Regression and Autoimmunity after Reversal of a Functionally Tolerant State of Self-reactive CD8+ T Cells. The Journal of Experimental Medicine 2003;198(4):569-80.

28. Dudley ME, Wunderlich JR, Robbins PF, *et al.* Cancer regression and autoimmunity in patients after clonal repopulation with antitumor lymphocytes. Science 2002;298(5594):850-4.

29. Okamoto T, Irie RF, Fujii S, *et al.* Anti-tyrosinase-related protein-2 immune response in vitiligo patients and melanoma patients receiving active-specific immunotherapy. J Invest Dermatol 1998;111(6):1034-9.

30. Colella TA, Bullock TNJ, Russell LB, *et al.* Self-Tolerance to the Murine Homologue of a Tyrosinase-Derived Melanoma Antigen. The Journal of Experimental Medicine 2000;191(7):1221-32.

31. Ochsenbein AF, Sierro S, Odermatt B, *et al.* Roles of tumour localization, second signals and cross priming in cytotoxic T-cell induction. Nature 2001;411(6841):1058-64.

32. Lengauer C, Kinzler KW, Vogelstein B. Genetic instabilities in human cancers. Nature 1998;396(6712):643-9.

33. Townsend A, Ohlen C, Rogers M, Edwards J, Mukherjee S, Bastin J. Source of unique tumour antigens. Nature 1994;371(6499):662.

34. Linnebacher M, Gebert J, Rudy W, *et al.* Frameshift peptide-derived T-cell epitopes: a source of novel tumor-specific antigens. Int J Cancer 2001;93(1):6-11.

35. Ripberger E, Linnebacher M, Schwitalle Y, Gebert J, von Knebel Doeberitz M. Identification of an HLA-A0201-restricted CTL epitope generated by a tumor-specific frameshift mutation in a coding microsatellite of the OGT gene. J Clin Immunol 2003;23(5):415-23.

36. Ronsin C, Chung-Scott V, Poullion I, Aknouche N, Gaudin C, Triebel F. A non-AUG-defined alternative open reading frame of the intestinal carboxyl esterase mRNA generates an epitope recognized by renal cell carcinoma-reactive tumor-infiltrating lymphocytes in situ. J Immunol 1999;163(1):483-90.

37. Hunger RE, Brand CU, Streit M, *et al.* Successful induction of immune responses against mutant ras in melanoma patients using intradermal injection of peptides and GM-CSF as adjuvant. Exp Dermatol 2001;10(3):161-7.

38. Schwitalle Y, Kloor M, Eiermann S, *et al.* Immune response against frameshift-induced neopeptides in HNPCC patients and healthy HNPCC mutation carriers. Gastroenterology 2008;134(4):988-97.

39. Koesters R, Linnebacher M, Coy JF, *et al.* WT1 is a tumor-associated antigen in colon cancer that can be recognized by in vitro stimulated cytotoxic T cells. Int J Cancer 2004;109(3):385-92.

40. Cathcart K, Pinilla-Ibarz J, Korontsvit T, *et al.* A multivalent bcr-abl fusion peptide vaccination trial in patients with chronic myeloid leukemia. Blood 2004;103(3):1037-42.

41. van den Broeke LT, Pendleton CD, Mackall C, Helman LJ, Berzofsky JA. Identification and epitope enhancement of a PAX-FKHR fusion protein breakpoint epitope in alveolar rhabdomyosarcoma cells created by a tumorigenic chromosomal translocation inducing CTL capable of lysing human tumors. Cancer Res 2006;66(3):1818-23.

42. Worley BS, van den Broeke LT, Goletz TJ, *et al.* Antigenicity of fusion proteins from sarcoma-associated chromosomal translocations. Cancer Res 2001;61(18):6868-75.

43. Ellegren H. Microsatellites: simple sequences with complex evolution. Nat Rev Genet 2004;5(6):435-45.

44. Pearson CE, Nichol Edamura K, Cleary JD. Repeat instability: mechanisms of dynamic mutations. Nat Rev Genet 2005;6(10):729-42.

45. Shinde D, Lai Y, Sun F, Arnheim N. Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)n and (A/T)n microsatellites. Nucleic Acids Res 2003;31(3):974-80.

46. Duval A, Hamelin R. Mutations at coding repeat sequences in mismatch repair-deficient human cancers: toward a new concept of target genes for instability. Cancer Res 2002;62(9):2447-54.

47. Woerner SM, Yuan YP, Benner A, Korff S, von Knebel Doeberitz M, Bork P. SelTarbase, a database of human mononucleotide-microsatellite mutations and their potential impact to tumorigenesis and immunology. Nucleic Acids Research 2010;38(suppl 1):D682-D9.

48. Wang ET, Sandberg R, Luo S, *et al.* Alternative isoform regulation in human tissue transcriptomes. Nature 2008;456(7221):470-6.

49. Klinck R, Bramard A, Inkel L, *et al.* Multiple alternative splicing markers for ovarian cancer. Cancer Res 2008;68(3):657-63.

50. Venables JP, Klinck R, Bramard A, *et al.* Identification of alternative splicing markers for breast cancer. Cancer Res 2008;68(22):9525-31.

51. Boguski MS, Lowe TM, Tolstoshev CM. dbEST--database for "expressed sequence tags". Nat Genet 1993;4(4):332-3.

52. Babenko VN, Basu MK, Kondrashov FA, Rogozin IB, Koonin EV. Signs of positive selection of somatic mutations in human cancers detected by EST sequence analysis. BMC Cancer 2006;6:36.

53. Brentani H, Caballero OL, Camargo AA, *et al.* The generation and utilization of a cancer-oriented representation of the human transcriptome by using expressed sequence tags. Proc Natl Acad Sci U S A 2003;100(23):13418-23.

54. Brulliard M, Lorphelin D, Collignon O, *et al.* Nonrandom variations in human cancer ESTs indicate that mRNA heterogeneity increases during carcinogenesis. Proc Natl Acad Sci U S A 2007;104(18):7522-7.

55. Hahn Y, Bera TK, Gehlhaus K, Kirsch IR, Pastan IH, Lee B. Finding fusion genes resulting from chromosome rearrangement by analyzing the expressed sequence databases. Proc Natl Acad Sci U S A 2004;101(36):13257-61.

56. Romani A, Guerra E, Trerotola M, Alberti S. Detection and analysis of spliced chimeric mRNAs in sequence databanks. Nucleic Acids Res 2003;31(4):e17.

57. Unneberg P, Claverie JM. Tentative mapping of transcription-induced interchromosomal interaction using chimeric EST and mRNA data. PLoS ONE 2007;2(2):e254.

58. Berger MF, Levin JZ, Vijayendran K, *et al.* Integrative analysis of the melanoma transcriptome. Genome Res 2010;20(4):413-27.

59. Campbell PJ, Stephens PJ, Pleasance ED, *et al.* Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet 2008;40(6):722-9.

60. Edgren H, Murumagi A, Kangaspeska S, *et al.* Identification of fusion genes in breast cancer by paired-end RNA-sequencing. Genome Biol 2011;12(1):R6.

61. Maher CA, Palanisamy N, Brenner JC, *et al.* Chimeric transcript discovery by paired-end transcriptome sequencing. Proc Natl Acad Sci U S A 2009;106(30):12353-8.

62. Zhao Q, Caballero OL, Levy S, *et al.* Transcriptome-guided characterization of genomic rearrangements in a breast cancer cell line. Proc Natl Acad Sci U S A 2009;106(6):1886-91.

63. Javier RT, Butel JS. The history of tumor virology. Cancer Res 2008;68(19):7693-706.

64. Rous P. A Sarcoma of the Fowl Transmissible by an Agent Separable from the Tumor Cells. J Exp Med 1911;13(4):397-411.

65. Martin D, Gutkind JS. Human tumor-associated viruses and new insights into the molecular mechanisms of cancer. Oncogene;27(S2):S31-S42.

66. Parkin DM. The global health burden of infection-associated cancers in the year 2002. International Journal of Cancer 2006;118(12):3030-44.

67. Burger RA, Monk BJ, Kurosaki T, *et al.* Human papillomavirus type 18: association with poor prognosis in early stage cervical cancer. J Natl Cancer Inst 1996;88(19):1361-8.

68. Butel JS. Viral carcinogenesis: revelation of molecular mechanisms and etiology of human disease. Carcinogenesis 2000;21(3):405-26.

69. Finn OJ, Edwards RP. Human papillomavirus vaccine for cancer prevention. N Engl J Med 2009;361(19):1899-901.

70. Chang MH. Cancer prevention by vaccination against hepatitis B. Recent Results Cancer Res 2009;181:85-94.

71. Zur Hausen H. The search for infectious causes of human cancers: where and why. Virology 2009;392(1):1-10.

72. Feng H, Shuda M, Chang Y, Moore PS. Clonal integration of a polyomavirus in human Merkel cell carcinoma. Science 2008;319(5866):1096-100.

73. Feng H, Taylor JL, Benos PV, *et al.* Human Transcriptome Subtraction by Using Short Sequence Tags To Search for Tumor Viruses in Conjunctival Carcinoma. J Virol 2007;81(20):11332-40.

74. Weber G, Shendure J, Tanenbaum DM, Church GM, Meyerson M. Identification of foreign gene sequences by transcript filtering against the human genome. Nat Genet 2002;30(2):141-2.

75. Xu Y, Stange-Thomann N, Weber G, *et al.* Pathogen discovery from human tissue by sequence-based computational subtraction. Genomics 2003;81(3):329-35.

76. Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 2007;35(Database issue):D61-5.

77. Gastric cancer and Helicobacter pylori: a combined analysis of 12 case control studies nested within prospective cohorts. Gut 2001;49(3):347-53.

78. Michael L, Johannes G, Wolfgang R, *et al.* Frameshift peptide-derived T-cell epitopes: A source of novel tumor-specific antigens. International Journal of Cancer 2001;93(1):6-11.

79. Druker BJ, Tamura S, Buchdunger E, *et al.* Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr-Abl positive cells. Nat Med 1996;2(5):561-6.

80. Hampton OA, Den Hollander P, Miller CA, *et al.* A sequence-level map of chromosomal breakpoints in the MCF-7 breast cancer cell line yields insights into the evolution of a cancer genome. Genome Res 2009;19(2):167-77.

81. Maher CA, Kumar-Sinha C, Cao X, *et al.* Transcriptome sequencing to detect gene fusions in cancer. Nature 2009;458(7234):97-101.

82. Ruan Y, Ooi HS, Choo SW, *et al.* Fusion transcripts and transcribed retrotransposed loci discovered through comprehensive transcriptome analysis using Paired-End diTags (PETs). Genome Res 2007;17(6):828-38.

83. Soda M, Choi YL, Enomoto M, *et al.* Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature 2007;448(7153):561-6.

84. Tomlins SA, Rhodes DR, Perner S, *et al.* Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science 2005;310(5748):644-8.

85. Stephens PJ, McBride DJ, Lin ML, *et al.* Complex landscapes of somatic rearrangement in human breast cancer genomes. Nature 2009;462(7276):1005-10.

86. Guffanti A, Iacono M, Pelucchi P, *et al.* A transcriptional sketch of a primary human breast cancer by 454 deep sequencing. BMC Genomics 2009;10:163.

87. Rozen S, Skaletsky H. Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 2000;132:365-86.

88. Peters B, Sidney J, Bourne P, *et al.* The immune epitope database and analysis resource: from vision to blueprint. PLoS Biol 2005;3(3):e91.

89. Lundegaard C, Lamberth K, Harndahl M, Buus S, Lund O, Nielsen M. NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11. Nucleic Acids Res 2008;36(Web Server issue):W509-12.

90. Bui HH, Sidney J, Dinh K, Southwood S, Newman MJ, Sette A. Predicting population coverage of T-cell epitope-based diagnostics and vaccines. BMC Bioinformatics 2006;7:153.

91. Bozic I, Antal T, Ohtsuki H, *et al.* Accumulation of driver and passenger mutations during tumor progression. Proc Natl Acad Sci U S A.

92. Segal NH, Parsons DW, Peggs KS, *et al.* Epitope landscape in breast and colorectal cancer. Cancer Res 2008;68(3):889-92.

93. Konopka JB, Watanabe SM, Witte ON. An alteration of the human c-abl protein in K562 leukemia cells unmasks associated tyrosine kinase activity. Cell 1984;37(3):1035-42.

94. Li H, Wang J, Mor G, Sklar J. A neoplastic gene fusion mimics trans-splicing of RNAs in normal human cells. Science 2008;321(5894):1357-61.

95. Palanisamy N, Ateeq B, Kalyana-Sundaram S, *et al.* Rearrangements of the RAF kinase pathway in prostate cancer, gastric cancer and melanoma. Nat Med 2010;16(7):793-8.

96. Rickman DS, Pflueger D, Moss B, *et al.* SLC45A3-ELK4 is a novel and frequent erythroblast transformation-specific fusion transcript in prostate cancer. Cancer Res 2009;69(7):2734-8.

97. Cline MS, Smoot M, Cerami E, *et al.* Integration of biological networks and gene expression data using Cytoscape. Nat Protocols 2007;2(10):2366-82.

98. Mitelman F JBaMFE. Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer. 2011.

99. Forbes SA, Bindal N, Bamford S, *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic Acids Research 2011;39(suppl 1):D945-D50.

100. Jhavar S, Reid A, Clark J, *et al.* Detection of TMPRSS2-ERG Translocations in Human Prostate Cancer by Expression Profiling Using GeneChip Human Exon 1.0 ST Arrays. The Journal of Molecular Diagnostics 2008;10(1):50-7.

101. Wang J, Cai Y, Ren C, Ittmann M. Expression of Variant TMPRSS2/ERG Fusion Messenger RNAs Is Associated with Aggressive Prostate Cancer. Cancer Research 2006;66(17):8347-51.

102. Yoshimoto M, Joshua AM, Chilton-Macneill S, *et al.* Three-color FISH analysis of TMPRSS2/ERG fusions in prostate cancer indicates that genomic microdeletion of chromosome 21 is associated with rearrangement. Neoplasia 2006;8(6):465-9.

103. Vogelstein B, Kinzler KW. Cancer genes and the pathways they control. Nat Med 2004;10(8):789-99.

104. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell 2000;100(1):57-70.

105. Sawyers C. Targeted cancer therapy. Nature 2004;432(7015):294-7.

106. Sjoblom T, Jones S, Wood LD, *et al.* The consensus coding sequences of human breast and colorectal cancers. Science 2006;314(5797):268-74.

107. Greenman C, Stephens P, Smith R, *et al.* Patterns of somatic mutation in human cancer genomes. Nature 2007;446(7132):153-8.

108. Kannan K, Wang L, Wang J, Ittmann MM, Li W, Yen L. Recurrent chimeric RNAs enriched in human prostate cancer identified by deep sequencing. Proc Natl Acad Sci U S A 2011;108(22):9172-7.

109. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25(14):1754-60.

110. Kalsotra A, Cooper TA. Functional consequences of developmentally regulated alternative splicing. Nat Rev Genet;12(10):715-29.

111. Venables JP. Unbalanced alternative splicing and its significance in cancer. Bioessays 2006;28(4):378-86.

112. Ermak G, Jennings T, Boguniewicz A, Figge J. Novel CD44 messenger RNA isoforms in human thyroid and breast tissues feature unusual sequence rearrangements. Clin Cancer Res 1996;2(8):1251-4.

113. Song SW, Fuller GN, Zheng H, Zhang W. Inactivation of the invasion inhibitory gene IIp45 by alternative splicing in gliomas. Cancer Res 2005;65(9):3562-7.

114. Wang L, Duke L, Zhang PS, *et al.* Alternative splicing disrupts a nuclear localization signal in spleen tyrosine kinase that is required for invasion suppression in breast cancer. Cancer Res 2003;63(15):4724-30.

115. Ars E, Serra E, Garcia J, *et al.* Mutations affecting mRNA splicing are the most common molecular defects in patients with neurofibromatosis type 1. Hum Mol Genet 2000;9(2):237-47.

116. Stickeler E, Kittrell F, Medina D, Berget SM. Stage-specific changes in SR splicing factors and alternative splicing in mammary tumorigenesis. Oncogene 1999;18(24):3574-82.

117. Wang L, Wang J, Sun S, *et al.* A novel DNMT3B subfamily, DeltaDNMT3B, is the predominant form of DNMT3B in non-small cell lung cancer. Int J Oncol 2006;29(1):201-7.

118. DiFeo A, Feld L, Rodriguez E, *et al.* A functional role for KLF6-SV1 in lung adenocarcinoma prognosis and chemotherapy response. Cancer Res 2008;68(4):965-70.

119. Narla G, Heath KE, Reeves HL, *et al.* KLF6, a candidate tumor suppressor gene mutated in prostate cancer. Science 2001;294(5551):2563-6.

120. Reeves HL, Narla G, Ogunbiyi O, *et al.* Kruppel-like factor 6 (KLF6) is a tumor-suppressor gene frequently inactivated in colorectal cancer. Gastroenterology 2004;126(4):1090-103.

121. Ghigna C, Giordano S, Shen H, *et al.* Cell motility is controlled by SF2/ASF through alternative splicing of the Ron protooncogene. Mol Cell 2005;20(6):881-90.

122. Adamia S, Reiman T, Crainie M, Mant MJ, Belch AR, Pilarski LM. Intronic splicing of hyaluronan synthase 1 (HAS1): a biologically relevant indicator of poor outcome in multiple myeloma. Blood 2005;105(12):4836-44.

123. Maxwell CA, Rasmussen E, Zhan F, *et al.* RHAMM expression and isoform balance predict aggressive disease and poor survival in multiple myeloma. Blood 2004;104(4):1151-8.

124. Krieg A, Mahotka C, Krieg T, *et al.* Expression of different survivin variants in gastric carcinomas: first clues to a role of survivin-2B in tumour progression. Br J Cancer 2002;86(5):737-43.

125. Ryan B, O'Donovan N, Browne B, *et al.* Expression of survivin and its splice variants survivin-2B and survivin-DeltaEx3 in breast cancer. Br J Cancer 2005;92(1):120-4.

126. Vegran F, Boidot R, Oudin C, Riedinger JM, Lizard-Nacol S. Distinct expression of Survivin splice variants in breast carcinomas. Int J Oncol 2005;27(4):1151-7.

127. Xin Y, Grace A, Gallagher MM, Curran BT, Leader MB, Kay EW. CD44V6 in gastric carcinoma: a marker of tumor progression. Appl Immunohistochem Mol Morphol 2001;9(2):138-42.

128. Druillennec S, Dorard C, Eychene A. Alternative splicing in oncogenic kinases: from physiological functions to cancer. J Nucleic Acids;2012:639062.

APPENDIX A

SUPPLEMENTAL: VIRAL SEQUENCES

Table A.1 The list of viruses without vec-screening

Virus	Virus
Papiine herpesvirus 2	Human herpesvirus 4
Melanoplus sanguinipes entomopoxvirus	Bacillus phage IEBH
Simian virus 40	Acanthamoeba polyphaga mimivirus
Human papillomavirus 18	Goatpox virus Pellor
Ecotropis obliqua NPV	Gallid herpesvirus 3
Human papillomavirus type 16	Enterobacteria phage SP6
Molluscum contagiosum virus subtype 1	Anguillid herpesvirus 1
Suid herpesvirus 1	Saimiriine herpesvirus 2
Saimiriine herpesvirus 1	Enterobacteria phage BP-4795
Oryctes rhinoceros virus	Shrimp white spot syndrome virus
Choristoneura occidentalis granulovirus	Great Island virus
Xylella phage Xfas53	Woolly monkey sarcoma virus
Adoxophyes orana nucleopolyhedrovirus	Murine osteosarcoma virus
Cercopithecine herpesvirus 2	Squirrel monkey retrovirus
Singapore grouper iridovirus	Moloney murine leukemia virus
Lactobacillus phage LP65	Mason-Pfizer monkey virus
Bovine herpesvirus 4	Hepatitis B virus
Caviid herpesvirus 2	Abelson murine leukemia virus
Sinorhizobium phage PBC5	Y73 sarcoma virus
Enterobacteria phage lambda	Ground squirrel hepatitis virus
Acanthocystis turfacea Chlorella virus 1	Fujinami sarcoma virus
Human papillomavirus type 9	Avian myelocytomatosis virus
Human adenovirus 5	Enterobacteria phage phiX174 sensu lato
Ectocarpus siliculosus virus 1	Enterobacteria phage M13
Human herpesvirus 5	Canine parvovirus
Tupaiid herpesvirus 1	Infectious salmon anemia virus
Burkholderia phage phiE125	Beilong virus
Cafeteria roenbergensis virus BV-PW1	Parainfluenza virus 5
Microbacterium phage Min1	Tula virus
Bovine herpesvirus 1	Whitewater Arroyo virus
Enterobacteria phage P1	Physalis mottle virus
Ovine herpesvirus 2	Chiltepin yellow mosaic virus
Shigella phage Sf6	Hepatitis C virus genotype 2
Human herpesvirus 8	Groundnut rosette virus
Enterobacteria phage Min27	Severe acute respiratory syndrome-related coronavirus
Plutella xylostella granulovirus	

Table A.2 The list of 48 viral peptides on the chip. These peptides were predicted B cell epitopes from viral proteins. 'Origin' column indicates which virus and proteins the peptides derived from

Peptide	Origin
PKKPRGPRGPRP	Murine type C retrovirus(hypothetical protein MtCrVgp1),Xenotropic MuLV-related virus VP62(putative gag-pro-pol polyprotein , putative gag polyprotein),Rauscher murine leukemia virus(gag polyprotein),Friend murine leukemia virus(gag protein),Spleen focus-forming virus(gag polyprotein fragment),Moloney murine leukemia virus(Pr180 , Pr65)
PYDPEDPGQE	Murine type C retrovirus(hypothetical protein MtCrVgp1),Xenotropic MuLV-related virus VP62(putative gag-pro-pol polyprotein , putative gag polyprotein),Moloney murine sarcoma virus(Pr65),Rauscher murine leukemia virus(gag polyprotein),Friend murine leukemia virus(gag protein),Moloney murine leukemia virus(Pr180 , Pr65)
DNHSGESNKETSD	Rachiplusia ou MNPV(DNA helicase),Plutella xylostella multiple nucleopolyhedrovirus(DNA helicase),Bombyx mori NPV(DNA Helicase),Bombyx mandarina nucleopolyhedrovirus(DNA helicase),Autographa californica nucleopolyhedrovirus(helicase)
APDNDDPNFE	Rachiplusia ou MNPV(global transactivator),Plutella xylostella multiple nucleopolyhedrovirus(global transactivator),Bombyx mori NPV(GTA),Bombyx mandarina nucleopolyhedrovirus(GTA),Autographa californica nucleopolyhedrovirus(global transactivator-like protein)
SGRGGMPSTTRGSNDGE	Human herpesvirus 4 type 2(BPLF1),Human herpesvirus 4(BPLF1)
RPGGPEEGAVPGPGRPEAE	Human herpesvirus 4 type 2(BALF3),Human herpesvirus 4(BALF3)
GTRPDLTDQPIPD	Xenotropic MuLV-related virus VP62(putative gag-pro-pol polyprotein)
KSKPPKPQVLPD	Xenotropic MuLV-related virus VP62(putative gag-pro-pol polyprotein , putative gag polyprotein)
QTNQAGGEAPQPGDNST	Human herpesvirus 4(BZLF1)
EPDSRDQQSRGQRRGD	Human herpesvirus 4 type 2(EBNA-3C)
ASGMGTPATAEPAPPSN	Abelson murine leukemia virus(p120 Gag-Abl

	polyprotein)
DEDNQDDDDATTSYGKP	Beilong virus(nucleocapsid protein)
TKEEGATKKKTQKP	Bovine viral diarrhea virus 1(polyprotein)
SLDQGEPTNPSDAAAK	Canine parvovirus(polyprotein)
SPRRRTPSPRRRRSQ	Hepatitis B virus(precore/core protein , Core and e antigen)
TSPSPPVEQPQVGQ	Human adenovirus C(control protein E4orf6/7)
RPPGPPSGPSPDASPEA	Human herpesvirus 1(DNA replication origin-binding helicase)
RQRSQPGSAQGSGKRPP	Human herpesvirus 5(DNA polymerase catalytic subunit)
PSTNKPTNSQAKSSTKP	Human herpesvirus 8(KCP)
IGPRKRSAPSATTSSK	Human papillomavirus - 18(L1 protein)
ETETPCSQYSGGSGGGC	Human papillomavirus type 16(E1)
GQKNNNPSFSED	Murine osteosarcoma virus(gag polyprotein)
SVGGGAKPKKPR	Parainfluenza virus 5(phosphoprotein , V protein)
KESEKDSRTKPP	Pestivirus Giraffe-1(polyprotein)
RKGSCPGAAPKKPKEPV	Simian virus 40(Major capsid protein VP1)
PKKIQPPTQLPTQPNAP	Squirrel monkey retrovirus(gag protein)
TEEERQEREKKEAEE	Woolly monkey sarcoma virus(pre-gag ORF protein , hypothetical Gag polyprotein)
QAPEDQGPQREPH	Human immunodeficiency virus 1(Vpr , Vpr)
EPQLRDETTPNDDAD	Enterobacteria phage ID18 sensu lato(gpB)
HPKPPPPLPPSAPSL	Abelson murine leukemia virus(p120 Gag-Abl polyprotein)
QKQPGAVGGPVKKGA	Beilong virus(W protein)
SEKDSKTKPPD	Bovine viral diarrhea virus 1(polyprotein)
AVQPDGGQPAV	Canine parvovirus(polyprotein)
GSSSGTVNPVPTTAS	Hepatitis B virus(large S protein , middle S protein)
KKRPSPKPERPPSP	Human adenovirus C(single-stranded DNA-binding protein)
VPPQGAEPQSNAGPRPH	Human herpesvirus 1(thymidine kinase)
TSPDDSSSGEVPDHPTA	Human herpesvirus 5(membrane glycoprotein UL18)
TGAESEDSGDEGPSTRH	Human herpesvirus 8(vIRF-3)
ADPEGTDGEGT	Human papillomavirus - 18(E1 protein)
AISDDENENDSDTG	Human papillomavirus type 16(E1)
DPEPKPSLE	Murine osteosarcoma virus(gag polyprotein)
TQQVPRPGTGDC	Parainfluenza virus 5(hemagglutinin-neuraminidase protein)
LTEGPPPKE	Pestivirus Giraffe-1(polyprotein)
FNPEEAEET	Simian virus 40(large T antigen)

PIPPANPCPPSNQP	Squirrel monkey retrovirus(protease)
SPGTSQEQRA	Woolly monkey sarcoma virus(Env protein , p28sis)
STEGSNNTEGS	Human immunodeficiency virus 1(Nef , Nef)
QGSNPPNGQQAA	Enterobacteria phage ID18 sensu lato(gpH)

APPENDIX B

SUPPLEMENTAL: CHIMERIC TRANSCRIPTS

Table B.1 The sequences of primers for screening of chimeric transcripts in human

Gene Fusion	FW	Rev	Product Size
BOLA2_Exon2_SMG1_Exon11	CGGAAGTGGCTCCTGTAAG	AGGATCCAAGCTGCCGAGC	370
STAG3L1_Exon6_STAG3L3_Exon5	AGTGCGTGAAGGCTCTGAAAG	AGTGAAGAGCTCCAGGCGTGC	68
RMND5A_Exon2_ANAPC1_Exon25	CGTCTCCGGCATGGATCAG	CAGGCTGCTCACGACAGTG	382
BPTF_Exon9_KPNA2_Exon2	CTATCACCCAATTAGAAAACAACATC	TTTCCCTTGTTCTTGAATCTGTGAAG	284
MSH5_Exon5_AIF1_Exon5	GAGGAGGAGGAAGTCGAGG	AGCCACTGGACACCTCTCC	400
GFOD1_Exon1_C6orf114_Exon2	CGTGTCATCATCCCGCTGC	CAAGAGCAGCATCGTGGGG	355
FPGT_Exon3_TNNI3K_Exon26	AGAGGCAAACTTGTAGCACGTGGAGAATTC	TTCTCCACGTGCTACAAGTTTGCCTCTTAG	386
C11orf79_Exon3_C11orf66_Exon2	CACAGCCTATTGTCTCCTTTGC	AGCTGGTGGCGTAGAATTTCAG	438
GJC3_Exon1_AZGP1_Exon3	CTGGGGTTGCAGTACCACC	TATCTGGGCTGCTGGGTCG	480
SYNJ2BP_Exon3_COX16_Exon2	TTTGGTCACTGAGGAAGAGATC	GGGTCCTCGAATATTCTTCCAG	460
RPL17_Exon6_C18orf32_Exon2	GGTCATTGAGCATATCCAAGTG	CCATATACGACTAACGAAGGGG	303
C20orf29_Exon2_VISA_Exon2	AGGTTACGGAGGCAGTGAC	TGAGGCAGGGCAGGTAAGG	381
MDS1_Exon2_EVI1_Exon4	CATCTACATCCCTGATGATATC	ATGAACAGCAGAAGCTCCTCTC	366
C11orf79_Exon3_C11orf66_Exon5	CACAGCCTATTGTCTCCTTTGC	TTGTACTCCGACGTCATGAAGC	474
PRR13_Exon3b_PCBP2_Exon2	GACTAGGAAGAGCCGAGAC	CGGTGTCCATGTCGAGCAG	416
HISPPD2A_Exon27_CATSPER2_Exon3	GAACCAGACCGGGCATTGC	TGCTCAACAGCCTCTGGGC	430
RRM2_Exon9_C2orf48_Exon2	ACACTGTGATTTTGCTTGCCTG	TTCAAGACGTAAGGCTGGTCAG	359
RNF103_Exon2_VPS24_Exon2	CATTGTGTGGTATGAAACTGGC	CCTTCTTGGCAGCATCTTTCAC	413
WRB_Exon3_SH3BGR_Exon3	ATGCAATGTTCTTAGGATCCTCC	ATCCACCTCCTGTTGTCTTCATC	378
PRR5_Exon6_ARHGAP8_Exon2	CGGGTCACCATGAGGACTC	TCGTGCTCAGCGCAGGATC	433
ABHD14A_Exon3_ACY1_Exon2	TTGGGCCCGACTGTGGTAC	ACGCAGGTACTGGCGGAAG	418
KIAA1984_Exon8_C9orf86_Exon2	GACCTGCTGGATTATCTGAAGAC	CAGACTTCAACCTTCACGATGTC	494
RPL11_Exon4_TCEB3_Exon2	TGGCATCCGGAGAAATGAAAAG	ACATGCTCGTGTTTTCGCAAGC	324
CNPY2_Exon3_CS_Exon3	AGCGCTCTGGAGAATCCCG	CCACCACCGTCTTGCCATG	497
TYMP_Exon9_SCO2_Exon2	CTGCTGGCGCCCGCAGATGGC	GGAGGACCCGAGGCTTGAGCT	240
NAIP_Exon13_OCLN_Exon5	CTCCATTTAAACCACAGCAGAGG	ATTGGAAGAGTATGCCATGGGAC	442
PDLIM2_Exon8_C8orf58_Exon2	AGCCGCAGCTTCCAGAGTC	TCTCCCAGGATGAGCACCC	412
PTPN9_Exon3_BUB3_Exon8b	CGAAGAGATTAACAAGTGGACAG	CCACAGTAACTCTAACACATCCC	320
HMGA2_Exon2_ELAC2_Exon18	CACTTCAGCCCAGGGACAAC	AGGCTTTGAGCTGGTTGGGG	263
CCDC88C_Exon5_MXRA7_Exon5	GCACGTCAACAATGATGTGAAC	ATTTATGGAGGCAGCATGCACC	343
TFG_Exon3_GPR128_Exon2	ACCATGAACGGACAGTTGGATC	CTCAGTCCTTTCCACTCTTCTG	374
NUP214_Exon29_XKR3_Exon2	GGACAGACAACCTTCGGGC	GGAGTCTCTGGCCAAGGAC	456
FTH1_Exon3_SEPT11_Exon6	GGAACATGCTGAGAAACTGATG	TCATTCTCAACCTGCACCACAC	303
IKBKG_Exon2_DAZAP1_Exon9	TTGCCCTGTTGGATGAATAGGC	CGTAGCCAAAACTGAACTGTGG	382
COL6A3_Exon2_PICALM_Exon14	ATCGAGGAGCCCAGGGACAC	AGAAGTCCACCTAGTTCATCAAAGCC	296
EPS8L1_Exon15_TFPT_Exon5	TGCTGCGGGACAACGTCAC	TGCCCTCATCCTCCAGCAC	324
CCDC88C_Exon5_MXRA7_Exon5	GCACGTCAACAATGATGTGAAC	ATTTATGGAGGCAGCATGCACC	343
RBM14_NA_RBM4_Exon2	AAGATATTCGTGGGCAACGTCG	GTCTTCTATGTGCACAAAGCCG	477
LOC100134328_NA_WDR79_^Exon1	AAAACCCCAATCCCATCAACCC	TCACTAGGGGAACCAAACTCTG	613
TRIM61_Exon3_FARSB_Exon3	TAATCAGGTTTTGACTTTCTTCTGTC	TTATACACTGGAGCCTTTATCCTTTC	405
LOC100132296_NA_LOC376475_NA	TCGCAGGCACCATGACTCCTG	GCTCTGGAGCAGGGTACTTGG	300
LOC100131277_NA_TACC1_Exon2	GTGCCCAGAACCGCCTATG	TGGCCTGCTGTGGCACTTC	562
BOLA2_Exon2_SMG1_^Exon11	CGGAAGTGGCTCCTGTAAG	ATGGCAGTTTCGTATTTATCTGTTCAAC	411
LOC729245_NA_OBFC2A_Exon2	GAAGTGGGTGGAGAACGGG	ACTGAAATTTGGCACTTCTGAATAAACC	604
LOC729438_NA_GTF2I_Exon13	TTGCAACTTGGCGGGCCTC	TCGTCACCGGACCTTTGGC	506
LOC100132296_NA_WASH1_Exon3	TCGCAGGCACCATGACTCCTG	TCTGGAGCAGGGTACTTGGCA	300
ZEB2_Exon2_LOC100128821_NA	AGTGGCCGAAAGAGATCAGTTC	ACAACGTGCATGTCTCTGTGAG	576
PMF1_Exon4_BGLAP_Exon4	TATGACAAGTTTATAGCTCAGTTGCAG	TCACACACCTCCCTCCTGG	515
LOC100134445_NA_LOC376475_NA	AGAGGAGGGCACCATGACTC	CTCTGGAGCAGGGTACTTGG	303
OGFOD1_^Exon10_BBS2_^Exon17	CACGAAAGTCTGTGAGGCC	TGTCTTTAATTTGTACAACTCACCAAGG	339
STAG3_^Exon4_PMS2L3_Exon5b	ACTTTGAAGACAGCTTGAATCGC	GATCCACTCCATAGTCCTTAAGC	330

Gene Fusion	FW	Rev	Product Size
LIPT1_Exon1_MRPL30_Exon2	ATGACGCACTTTTCCAGCTC	CTTCAGGTGAGGCCTGAAAC	240
JAZF1_Exon1_CDCA7L_Exon3	CGATGTAGCACCATGACAGG	ACATCATCTCGGAAGCCAAC	201
SDHD_Exon3_TEX12_Exon3	GGTTCTCTGGAGGCTGAGTG	CCTCTGTCGCAGGAACTCTC	585
CHURC1_Exon3_FNTB_Exon2	CTCAGTTCTCGCGAGGTTTC	ACTATCTGGGGGATGGGTTC	536
PLDN_Exon1_SQRDL_Exon2	GGGTGCGACAATCTCTTCTG	CTGCACCCACTTTCCTCTTC	412
UNC5C_Exon1_RAP1GDS1_Exon6	TGCCTTTGGAGAAAGTGGAG	TCTGCCAATGGAGCAAGAAC	527
SAA2_Exon3_SAA4_Exon2	TTGGTCCTGAGTGTCAGCAG	AGTTTAGCAGCCCAGACACC	425
ELAC1_Exon2_SMAD4_Exon2	GGGGTGGAAGATGTCTATGG	ATGCACAATGCTCAGACAGG	357
BCAS4_Exon1_BCAS3_Exon24	CTCCTGATGCTGCTCGTG	GGTGGAGCCTGAGCTGTTAC	317
HSPE1_Exon2_MOBKL3_Exon3	ATGGCAGGACAAGCGTTTAG	TCTGGATGGCATTCACTCTG	392
C22orf39_Exon2_HIRA_Exon2	CCAGGCACTTCCTACACCAC	TTTGTCATCTCCCCCAGAAG	355
C1orf151_Exon1_NBL1_Exon3	GGGGAACATGTCTGAGTCG	AGGGACTCTGTGGACTGTGG	326
BGLAP_Exon1_PMF1_Exon5	TGAGAGCCCTCACACTCCTC	GGTGTATGTCCAACCCCAAG	319
LOC100133907_NA_KIFC3_Exon4	CAGCAAACACTGGCAAACC	GTCGCTCATTCTCCACCATC	203
RBM3_Exon3_LOC729275_NA	TTGAACTGCCATGTCCTCTG	CTGACTGGTCCACATTGCTC	514
HSPBP1_^Exon2_SF3B2_^Exon14	GGGCTTTGACCAGCTACTTG	CCAAAGGAACAGCTCTCAGG	572
LOC100129436_NA_SH3BGRL2_Exon2	ATGGCTGATTCAGGAGGATG	GTTCTGCCTTTGATGCCAAC	414
LOC728531_NA_LOC641977_NA	TGAGCTGGCAGTTCTGTGTC	ACCCTGTTACCAGGCATCTG	450
LOC728987_NA_KCNK1_Exon2	TTCAAACACAAGCACGGAAG	TCAGGAACAGGAGGGTGAAG	323
LDHB_^Exon3_FAM18B_^Exon6	TCCGCACGACTGTTACAGAG	CTTTCTAGCCCCAACTGCTG	472
POLR2J3_Exon2_LOC392713_NA	CCTTCGAGTCGTTCTTGCTC	AATGGCTGCATTCCTTCAAC	323
LOC728844_NA_HNRNPA1_Exon11	TGGGAAAGCATGGGTGATAG	ACCACTGTGCTTGGCTGAG	302
LOC100131446_NA_ARF3_^Exon5	CACAGCAACGACTGTTCACC	GGCTGTCTGGCTTTCACTTC	288
LOC100130093_NA_JMJD4_^Exon6	ACCAAAGTCATGAGGGCAAG	GACTCTCGGTCAAGGACAGG	263
WDR51B_Exon2_GALNT4_^Exon1	CTCTCCTTCCCCATCCTCTC	ATGCAGGGAAATCCTGTCAC	406
TPX2_^Exon4_ERGIC1_^Exon10	GAGGCTGTCGGCTAATAACG	CGGGTCAGAGGTGTCTCTTC	472
CNPY2_^Exon6_UBL5_^Exon5	AGTGGTGGAGGTGCCTTATG	GGGCAGGAAGATGAGGATTC	414
LOC100132296_NA_WASH1_Exon4	TGACTCCTGTGAGGATGCAG	AAGAGCAGCAAGGAGCTGAC	416
CXorf40B_^Exon5_LOC100132460_NA	GAGACTCGGGATGACTCCTG	CGTAGGGAGCGTCTTGTAGC	450
CORT_Exon1_APITD1_^Exon7	GTTACATCCAACCCCAGAGC	GGTCCACTCAAACCACCAAG	416
COPS5_Exon3_HNRNPH3_^Exon10	ATTGGCTCTGCTGAAGATGG	TTCCCATCACAATCTGTTGC	555
TAF15_^Exon15_DKFZP564O0823_^Exon2	AGAGGTGGAGACCCCAAAAG	AGGTGAGGGGTCTGTGTTTG	412
GMPR2_^Exon3_LOC100133713_NA	GCCCTCAGATTCATCGCTAC	ACCGATGGTTACTTGCTTCG	332
SRRM1_^Exon3_DDX17_^Exon4	AATCGGTTCAGCAACAAACAG	CTGGGCAAGCTCTCTGGTAG	323
ANK3_Exon30_SLC25A3_^Exon1a	TGGGTCCATGAGAGGAAAAG	TTCCTCCAAGTCCACAAAGG	595
HMGA2_Exon3_LOC100129940_NA	ACTTCAGCCCAGGGACAAC	GAGGTGGGAAAATGCTTGAG	483
BLOC1S1_Exon3_RDH5_NA	GGTGAGCGTTCCAGCTTCC	ATGTTGATCACCCGGCCCC	508
ANG_^Exon3_RNASE4_Exon3	AGAAGCGGGTGAGAAACAAAAC	GCATCATCAAGTTGCAGTAGCG	366
LOC100128309_NA_LOC100133190_NA	CGCGGGTTTAAGTGGTGGC	CCAGGCACATCCTCTCCTC	495
LOC100132460_NA_CXorf40B_^Exon5	ATCCAGGCTCCCAAGATAACTC	AGGAGGCATCACACAAGGAAAG	543
LOC729264_NA_LOC653550_NA	CCAGCTGGCATCCTAGACC	GGAAAGCAGCAGGCAGGAC	670
PKM2_^Exon6_RPS3_^Exon1	GCTGTGGCTCTAGACACTAAAG	TACAGCAGTCAGTTCCCGAATC	561
CSF1R_^Exon10_GDPD5_^Exon17	AGGTAAGCGTCATATGGACATTC	AGAGGCTGCGGCTGACAAG	446
RFC1_^Exon13_TEK_^Exon1	GAAATACCAAAGGGAGCTGAAAATTG	TTACTAGCCTTTTCTCTCTTCCAAAC	675
KIAA1530_^Exon3_DGCR8_^Exon2	TTCCTTCCGGGTCCTTCGG	GCGCTCTGCTCCTGTGTAC	530
C2orf30_Exon13_SF3A1_^Exon3a	GGAGAAATAAAGAGGGTGTCGG	GCGGAGAAAGTCAAACTGGTAG	556
LASS4_^Exon10_PAX8_^Exon12	GTAATGTGCTGGGACAGGTAC	TCTGTTTTAAGCTCCCTGGGG	535
TBC1D2B_^Exon8_YIPF1_^Exon9	AACTGGACAGGCTGAAAGATAATC	ATTTGTGCCACTCCATCAGTTCTG	422
STAG3L1_Exon6_STAG3L3_Exon5	ACATCGTCATAGCCGGAAAC	TCTCACAGTCCACGTCCATC	478

Gene Fusion	FW	Rev	Product Size
RPL28_Exon4_GLB1_Exon3	ACAGCACTGAGCCCAATAAC	CATGGTCCTCAGAAAACTGG	483
PDE3A_Exon2_VGLL4_Exon7	CCAGGTACGTGGAACAAATC	GAGGGAGGTGTACAGGTGG	375
RAB3GAP1_Exon3_CHMP1A_Exon5	GAGATCACGGACTTCACCAC	ATCTGCATGATGAGGCTGTC	320
APOC1_Exon3_LAMA1_Exon9	TCTCCAGTGCCTTGGATAAG	TTTCCTTACAAACACAGGGC	377
EIF3G_Exon5_TSEN54_Exon9	GACGACAAATGTGTCACCAG	AGATGTCACCATGATCCACC	516
GNB2L1_Exon3_SCARF1_^Exon10	TGGTTATCTCCTCAGATGGC	CCTGGCTAACATGGTGAAAC	393
PRKCI_Exon3_CCDC88B_Exon11	ACAACGAACAGCTCTTCACC	CAGCTCCACATTCTCCTCAG	336
FNTB_Exon1_CAMSAP1L1_Exon2	ACTATTGCCCTCCATCTTCC	TGTGCACTCATCTGTATGGG	387
RPRD2_Exon2_MEF2C_Exon8	ATTCAAGGCTTGTCGTCTTG	AGACCACCTGTGTTACCTGC	409
PHB2_Exon2_SPINT1_Exon9	CAGAACTTGAAGGACTTGGC	TAACCCAAGATGGCTACCAC	495
RNF216_Exon7_XKR8_Exon3	CCAGCCAAGATGAGACAAAG	ATGAAGTCTGTGCCCTGAAG	450
FBXL11_Exon6_TMEM71_Exon3	GGACTCGGAATAAAAATGCC	ATACATAACGCTGGTCTGGG	497
NM_001012636.1_Exon_NM_001012635.1_Exon	TGGTTATCTCCTCAGATGGC	CGGTGAAAACCCATGTCTAC	391
LOC100134209_NA_IQSEC3_Exon2	CGTGACTGACAAGGAGAAGG	TCTTGCTGAGCTGGTATTGG	280
KLC1_Exon2_LOC63920_Exon2	AGTAATTCAGGGGCTGGAAG	GAGATCATGCCCAGCTACAC	410
FAM122C_Exon3_GOSR1_^Exon9	GAGAAGAGTCAACAGTGCCC	ACTGAGTGAACGAGACTCCG	359
MAP3K7IP1_Exon2_MSL2L1_Exon2	ACAGATGACCTGCCTCTCTG	GCTTGTTTTCCTCAAACTGC	260
LOC730908_NA_LOC653458_NA	AAGATGGAAGGACCAGGAAC	GATTAGTAGCCTGCTGTGCC	230
MXD4_Exon3_PSMC4_Exon8	CTGAACTCCCTGCTGATCC	TTCCAAGTCAACCTCCTCAG	430
DLST_Exon5_ELAC2_Exon2	GAACTGCCCTCTAGGGAGAC	AATCATTCCACTTAAGCCCC	339
FBXW8_Exon2_C6orf103_Exon15	CTACAGCCTGGATGAGTTCC	GGCAATATGAACTTGGCTTG	442
STK24_^Exon8_JMJD2B_^Exon23	TGTTTGAATAAGGAGCCGAG	TCAAAGGTTGCTCAGTAGGG	258
NM_001012636.1_Exon_NM_001012635.1_Exon	CAAAGGTACCAGTGGCTCAG	ATGATGCTGAAGGAGCAAAG	280
MTUS1_NA_MALL_^Exon4	CAGCATTTCAGAATGGTTCC	CCTGTAATCCCAGCACTTTG	221
RABEPK_Exon7_LOC100132848_NA	AATGGCAGAAGCTAAATCCC	TTTCCTCTTCAGAGCTGGTG	277
TMED5_Exon1_R3HDM1_Exon19	TTCACACCTTCCCTCGATAG	ATGGGGAATTCCTTGAGAAC	299
RPL9_Exon5_TSNAX_Exon4	AGCTTGTTTCAAATTCAGCG	TAGGCAGTCTCCAAAACAGG	301
GPR98_Exon83_KIAA0831_Exon2	CTTTGCTTGGCTGTTCTTTC	AGTTTGCGATTATGCCTCTG	393
CCDC58_Exon3_CLIP1_Exon21	CTTTGCAGGGAAAATTGATG	CTGAGCTGCCTCTTGTTTTC	457
NM_001012636.1_Exon_NM_001012635.1_Exon	CTGATGACACAGAATGCCTG	AGGCTGAGGTGTATGCTTTG	530
COBLL1_Exon2_SMAD4_Exon5	AGCCAGGAGAAAACCAAAAG	GATTACTTGGTGGATGCTGG	306
IFT81_Exon16_PRKAA1_Exon3	GAAAGGACGAACATTGGATG	TCTATGGACCACCATATGCC	389
C17orf45_Exon1_ITGB3BP_Exon5	CTCATTGGAATCTCCTGTGC	ACGTGCACCTCTTCAAAGTC	362
FBXO38_Exon2_RNF216_Exon5	GCCACGAAAGAAAAGTGTG	ACTGAGCACTTTGAAGTCGG	408
SAMD10_Exon4_ZNF512B_Exon2	AGCACTGTCCCCACAACTAC	CATCCTTCCAGTCGTTCATC	495
LOC100131277_NA_TACC1_^Exon1	TTACCACACAGGATGCACAC	TTCAGAATCCGAGCTGAAAC	455
C14orf153_Exon4_KLC1_Exon2	CAAACCTTCGACCTGTTCAC	GCTCCAACATCTCCAGTGAC	531
ZNF222_Exon4_ZNF223_Exon2	CTTCAGGAACCTGCTCTCAG	GTATCTCGGTGGAATGGTTG	351
COP1_Exon2_CASP1_Exon1	TAAGACCCGAGCTTTGATTG	TCTCTTCCTTGTTCAGCACC	254
HMGCL_Exon5_GALE_Exon3b	CTGGCATCAACTACCCAGTC	CACCTGTTACCAGCACCTTC	306
COMMD3_Exon4_BMI1_Exon2	ACTAGAGGCAGGAAAGCACC	TTCATTTTTGAAAAGCCCTG	430
SPECC1_Exon3_MAP2K3_Exon5	TTCAACTTCCTTGGCTTTTG	CCCAGTTCTGAGATGGTCAC	336
NOS1AP_^Exon10_FLJ13137_^Exon1	AGAACAAGGACATGCTCCAG	TCTCTGTCACACCCACACTG	492
MRPL43_Exon5_SEMA4G_Exon7	GAGAGCATCCACTGCAAGTC	GCCTCTACACAGCCACTAGG	548
RIPK3_Exon9_ADCY4_Exon2	AGCCCTACCTCAACTGGAAC	AAAATAGGACACCTGGTCCC	319
LOC100128892_NA_CCDC80_^Exon1b	ACTGGCACACACAGACACAC	TATCTCCCTTTCCCCCATAG	221
LOC100131484_NA_CDKN1A_Exon2	TCATCACTACTCCCTCCAGC	GGTGTCTCGGTGACAAAGTC	470
DDIT3_^Exon3_MARS_^Exon21	GCTGAGTCATTGCCTTTCTC	ACAAAAGGCAGACAAGAACG	228

Gene Fusion	FW	Rev	Product Size
NM_001012636.1_Exon_NM_001012635.1_Exon	ACAAGATCGAGTTCTGACGC	TTCAAAAGATGGCACTTTCC	449
TMEM199_Exon5_SARM1_Exon2	CAAGACATGGTGGGACTCTC	TGAACATGTGCTCCAAGATG	259
WDR70_Exon3_LOC100132788_NA	GTTGTCAGTTTCTTGGTCGG	TCTTCTTGCTTCACCTCCAC	246
C1QTNF6_Exon2_IL2RB_Exon2	TCCTGCTCTTTCTCCTGATG	AGATGTTGGCTCTCGAGTTG	361
LOC732248_NA_CPLX2_Exon1	AGTCAAGGAAGAGGGGGAG	ACTTATCTCGGATCTGCTGC	463
NM_001012636.1_Exon_NM_001012635.1_Exon	GCAGATATGATGAGGATGGC	ACGGCTGTGTCTCTAACCTG	346
CCDC13_Exon14_HHATL_Exon6	TGGAGTCAGAGAGGAAGCTG	AGATTAGGGGGTCCATCTTG	316
TP53RK_Exon1_SLC13A3_Exon4	GGCGGCCAGAGCTACTAC	TGAGGAAGAGGAAGTTGGTG	450
NM_001012636.1_Exon_NM_001012635.1_Exon	AGAAGTGCAACTGGAGCATC	AGCTGGAGACACCTTCAATG	387
LOC100131434_NA_FLJ44451_NA	CTGATGAGACAAAACGGCTC	CACAAGAATTTTCCCCACAG	410
COX19_Exon2_CENTA1_Exon2	ATGAATTTCGGGACCAAGAG	ACTCCTGTCGCTCGTACTTG	452
MED8_Exon7c_ELOVL1_Exon2	CCTCAGGATTACAGCAGGTG	AGAGTGCCACCAGTGAGAAG	385
NM_001012636.1_Exon_NM_001012635.1_Exon	CATCGAACTCAACCTCAACC	GCAGGAATCAAGACCATCAC	321
SSSCA1_Exon3_FAM89B_Exon2b	GGACTACAAACACCTGTGCC	AAAATAGGACACCTGGTCCC	424
PMF1_Exon3_BGLAP_Exon2	AGGCGCTACCTGTATCAATG	AAAATAGGACACCTGGTCCC	366
RPL7A_^Exon3_DBH_^Exon1	AGAATTTTGGCATTGGACAG	ACGAGATCTGCGTTCTCAAG	284
LOC730060_NA_RRN3_Exon13	GAGAGAGAGGCGGATGAAG	TAAAATGGTCCATGGAGAGC	421
CSNK2B_Exon6_LY6G5B_Exon2	TACCAGCAAGGAGACTTTGG	ATGAAGCGAACCTTGACATC	405
ZNF816A_Exon4_ZNF83_Exon6	CTTGACTTTCAGGGATGTGG	AAAATACCCAGGGAGACCAG	256
LOC642423_NA_LOC100132703_NA	AGGACTCGCAGACGTTACTG	CTCCCCAAACACAGACTCAC	444
CHCHD3_Exon2_EXOC4_Exon2	CGGACGAGAATGAGAACATC	TGTGATGCTCTGGTATGTGC	283
ZNF816A_Exon4_ZNF83_Exon5	CTTGACTTTCAGGGATGTGG	ACCCTGAGGAAGAACCATTC	213
ZNF655_Exon4_LOC100131257_NA	GACATGGAACAGGGACTCAC	TTCTTCTCACTGGTGAAGGC	265
POLR2J3_Exon2_LOC100134053_NA	TCGAGTCGTTCTTGCTCTTC	AGTCAAACAAGCAGAATGGC	332
FAM18B2_Exon5_CDRT4_Exon2	TGTTGTCGTGTGACTTTTGG	CAGGGGTCATGTTTTTCAAG	463
NHLRC3_Exon3_C13orf23_Exon6	GTGTTGCAGTTGACTCCCTC	TGCAATAGTTGCATGTGGTC	462
LOC100128692_NA_MTFR1_Exon8	TTCCAGCTGGCTTTATTCAG	GAGCCTGTATGACGGGAAC	429
NM_001012636.1_Exon_NM_001012635.1_Exon	TTCATTGAGGACCCTGAGAG	GAGCTCTTCACCTTTACCCC	402
VPS36_Exon13_THSD1_Exon2	GGGAATAATGTCACTCACGG	CTCAGTGTCCCATTAGCACC	536
PMF1_Exon4_BGLAP_Exon2	AGGCGCTACCTGTATCAATG	TCGGAACTCTTGAAAAGCAG	332
LOC402160_NA_RNF4_Exon2	AAGAAAAGCGGGATCAAATC	TTTCATCTCCAGCAGTTTCC	571
CENPC1_Exon4_PAICS_Exon11	CCATTTGGTATGGAGCAAAC	TGCAGAAATCTCTTTGTCCC	327
YARS_Exon4_RNF19B_Exon8	AAAGACACAGCCAGTCTTGG	ACCTTAGGTGGTTTGAAGGG	290
CCDC19_Exon11_VSIG8_Exon2	GAACAGATTGAGAAGGAGCG	GGTTCATGAGGTTGATGGAG	512
HBS1L_Exon4_FAM54A_Exon7	GAATTCACATTGGGATCCAG	TTTCCTTACAAACACAGGGC	521
ACSF2_Exon10_CHAD_^Exon4	TTCTGAACCAGCCAGACTTC	CCCAAGTCAATCAGAACCAC	377
THAP2_Exon2_TMEM19_Exon2	TTGAAGCCTCCTGTTTTGAC	AAAGCTGAAATTTGCAATGG	256
RDH11_Exon6_VTI1B_Exon3	GGTGGCTTTTCTCCTTTTTC	TTGCCCTTTGAGACTGTAGC	323
OBSL1_Exon20_CHPF_Exon2	GCCCGGGAGATAAGTATGAG	AGAACCAGTCAAAGTCGTCG	410
MTFR1_Exon7_LOC100128692_NA	TTCCAGCTGGCTTTATTCAG	TGGCGGTATCAGAGATTTTC	488
TIMM23B_NA_LOC100132418_NA	CTGGTCCAAACCAAGAAATG	CTTCACCAGTTGGGATTGAC	287
NDUFA13_Exon4_YJEFN3_Exon2	CCATCGACTACAAACGGAAC	TGTTTCCTCCTCTGGACAAG	353
TMED6_Exon1_COG8_Exon2	ATCTAGTGACGTCTGCCAGG	TCCATGAGCTGAGGAATCTC	284
CBWD3_Exon2_CNTLN_Exon16	GAGGAGGAGGAAAAGTCTGG	ATAGTCCTTTGGGGTCTTGC	338
ZNF167_Exon5_ZNF660_Exon3	CTGTACTTCGGATGGTCAGG	TCCTCATTTTCTCCTGCTTG	340
NM_001012636.1_Exon_NM_001012635.1_Exon	ATACTGGCCCCTCTTTGTTC	TGCATCTATGTCGTGTGGAC	345
EDEM2_Exon3_RBM39_Exon12	CCATGTTCTACCACGCCTAC	TTCAGTCTGCTGGGAAAGTC	265
b-actin	AAATCTGGCACCACACCTTC	AGCACTGTGTTGGCGTACAG	646

Table B.2 The sequences of primers for screening gene fusions in mouse

Gene Fusion	FW	Rev	Product Size
Pdcd10+Arhgap15	aaaggtgggaagtgaagtcc	ggaaaaggaggaaattgagc	419
Pdia5+Sec22a	aagacctgtgtcagcaggag	tccttctgaagctcatccag	341
Pir+Figf	aaggtttacactcgcacacc	tacgcatgtctctctagggc	499
Dedd+Nit1	accatgaacgtggactcatc	cagcaagggatcagtagtgg	430
Rnf216+Xkr8	agaccggcttattatccacc	gtacagccactcacctttgg	471
Map3k7ip1+Msl2l1	agatgacctgccactctgtc	gaaggtttctcggtctcctc	366
Nos1ap+EG665574	agcacatctctctgctggtc	tggactgacctttcatcctc	506
Mxd4+Psmc4	agtacctggagcgtagggac	ccacatagtcttccaggtcg	356
Ddx17+Srrm1	agtttggtaatcctggggag	ttcagagacgccaatttttc	468
Leprotl1+Dctn6	agtttgtcctttggaggagc	gacagcccacttcaaatacg	465
Leprotl1+Mboat4	agtttgtcctttggaggagc	atctcttttcgggctctgac	537
Commd3+Bmi1	atggagctctcggagtctg	gggtgagctgcataaaaatc	407
Cflar+Cep68	attacacaggcagaggcaag	gtcagggactgatgttcgtc	276
Sgce+Scaper	attattctggtggtgggagc	acattgattgcattgcagag	392
LOC639541+Camsap1l1	attgtcctccatcttcctcc	tgtgcactcatctgtatggg	334
Alad+Rbm35b	cactccatccagcagacttc	tgtggtgtttgtgtctctgc	389
Rbm14+Rbm4b	cacttggaagattttcgtgg	ttgctcttattcttgctggc	293
Rbm14+Rbm4	cacttggaagattttcgtgg	ttctcaacaaggtcactccc	384
Shfm1+Tsen54	caggaacctcaatcatggac	cccagagctgtattcagtgg	451
Fam152a+Tbc1d1	caggtttctgacagcgtctc	ttttggatcttcctggagtg	491
Thap2+Tmem19	catcagcttccacaggtttc	cgatggttaagatgaatccg	315
0610038F07Rik+4930579J09Rik	catgattgaaatccctttgc	taagcaagagatctgggtgg	396
Dtna+Rprd1a	catgcaagctgagatttgtg	gactggccatcagtcttcac	359
Mospd2+Snd1	ccctcccttcttgatctacc	agcgactctgcaatgttttc	445
Rprd2+Mef2c	ctaccattggatgaagtggc	cctgcacttggaggtctatg	289
Med8+Elovl1	ctcagaagcagatccagagc	atcatgaagcctcgaagttg	399
Pde3a+Vgll4	ctgcagtggcaagtctcac	tgggacagtgagagaggttg	315
Bloc1s1+Rdh5	ctggaccatgaggtgaagac	tctccagactctccaggttg	400
Commd3+Bmi1	cttgaaacagatcgacccag	ctggttttgtgaacctggac	361
Adamts19+Slc27a6	gacctgtacctgctgctcc	tcttaaaaacagtgggcagg	392
Sssca1+Mtvr2	gactgaagctgagacgaagg	tcaaaggctgatttggaaag	433
Cops5+Hnrnph3	gaggcaacttggaagtgatg	gaaggcagcacatctttagg	560
Nit1+Dedd	gcaatctgttatgacatgcg	accatgaacgtggactcatc	279
Yars+Ibrdc3	gtaccgactgtcctctgtgg	cctcttcttccttctctgcc	323
LOC432548+Rhoa	gttgttcttgattcccatcg	ggctgtcaatggaaaaacac	397
Lmo7+Tgfbi	taataagctcaaacctggcg	tgatctgctggatgttgttg	402
Slc7a5+2610207I05Rik	tacatgctggaggtctacgg	caaggatggattgtaggctg	380
Mia1+Rab4b	tactatggagacctggcagc	agaggatgaccacgatgttg	384
Ift81+Prkaa1	ttacacccaagaggagcttg	atcaagcaggacgttctcag	427
Tmed5+R3hdm1	ttcacaccctctttggacag	attactttgttggccaggtg	363
Arid3b+Clk3	aagatcagaatcaacggcag	agtcctctccaaaagatggg	368
Slc35a3+Hiat1	acctttcaacaggctcacag	ttcatgcaataccaccaatg	365

Gene Fusion	FW	Rev	Product Size
Itpr1+2900073G15Rik	agaaaccttgattggttccc	aacatggtgaaattgatggg	518
Atp5k+Sash1	agaggagaatagcagcggag	acagccgtcttctgagtttg	365
Polr2j+Slc17a7	cagaccaccccagactacag	caacatgtttagggtggagg	258
Ecsit+Zfp653	ccacgtggacttcatctacc	acgttcttcaggcagttcac	382
Tnrc6a+Nedd4	ccaggaacacaaacagcttc	gcattttcatcttctgagcc	355
Tgfbr1+Rnf20	cctaattcctcgagacaggc	tcaatatcccactgcaggtc	324
Rnf139+Ndufb9	cctgcctctacatcatcgac	ctcagcttcttccactgctc	381
Nktr+Itsn2	cttcgacatcgagatcaacc	tgctagcattcccaaaattc	376
Med20+Usp49	gcttatggtgaagctcaagg	agtccttgcacaggtagcag	417
Aven+Nbea	ggaagaagacagcgattcag	acggtttgtacagctggaag	422
Fxyd2+Pck1	ggacagagaatcccttcgag	atggccaagttagtcttccc	373
Lats2+Xpo4	ggtggactcacaattccaag	aaaatatgcctgcaaactgc	320
D17Wsu92e+Ephb2	gtcaataccccctgataccc	atgtctaaggggtccaggtc	396
Tmem170+Cfdp1	gtgttctttcatgtccctgc	ccttccttcttttccctttg	229
Ap2m1+Bank1	tattcgaaccgaagctgaac	ttctccagagcttcatccac	291
Mybl1+D030016E14Rik	tcacccacaaagttcctagc	ttcagagttgaagctcgtcc	255
Samd5+Sash1	tggtgagctaccctaagctg	acagccgtcttctgagtttg	362
1810020G14Rik+2900073G15Rik	tgttggaggttcttttggtc	aaacatggtgaggaacatgg	529
Sec23ip+Heatr6	ttactcttctcctcctcggc	ggttcttgctgtcattcagg	380
Pcca+Exosc4	ttaggcctgatgtggctaag	tgcacgtagatgtcgatctg	451
Ylpm1+Cfdp1	ttcagctccctgacgactac	ccttccttcttttccctttg	299
Bcas2+Cluap1	ttttgaccaaggctatgagg	cagaagggatgtcagtctgc	274

Table B.3 The sequences of primers for screening gene fusions in dog

Gene Fusion	FW	Rev	Product Size
BOLA2_Exon2_SMG1_Exon11	GACTGCTGCAATGGAACTCA	TCTGGCAGTTCTCCAGTTGA	347
RNF103_Exon2_VPS24_Exon2	GAGCTGGTGGAAAAGTCAGG	TGGCCTTCATCACTTCTGTG	396
ABHD14A_Exon3_ACY1_Exon2	ACAATCCTGGCTGGTCTCAC	TTGAAGACAGGCACCACATC	448
RBM14_Exon1_RBM4_Exon2	CACGGCCTCTTAACACTTGG	TACGCACTGCTCCATATTGC	728
WRB_Exon3_SH3BGR_Exon3	CCAGATACGCCAGGCTAGAG	CAAAGGGATGCCATTTTGAG	277
CNPY2_Exon3_CS_Exon3	ACCCCATCACGCTAAAGATG	ACTGTCTTGCCATGTTGCTG	356
ELAC1_Exon2_SMAD4_Exon2	TGCACCAGTTGCAGAAGAAG	AGCTTGCTTGCACTCTCCTC	217
CHURC1_Exon3_FNTB_Exon2	AACATCTGCCTGGAGAATGG	TGTGCAGGATCCAGTAGCAG	383
HISPPD2A_Exon27_CATSPER2_Exon3	CGTAATCGGAAAGCTGGTTC	ACACTCAAGGACCCATCCAG	424
SDHD_Exon3_TEX12_Exon3	GTTCCAAGGCTGCATCTCTC	TCAAATCTTTCTCCAGGGATTC	249
RMND5A_Exon2_ANAPC1_Exon25	GAAGGTGCTGCACAAGTTCTC	GAGCACAGATTTCCCTTTGG	440
COMMD3_Exon4_BMI1_Exon2	CATACTAGAGGCGGGAAAGC	TTTTTGAAAAGCCCTGGAAC	430
MED8_Exon7c_ELOVL1_Exon2	TGCTGGAGCCTCAGGATTAC	GCCCAAGTGAGAGAACGAAG	257
TMEM199_Exon5_SARM1_Exon2	GACCTGGGAAAGCAAGTGAG	GGAAGCTCTCGCAGTAGTGG	340
ADHFE1_Exon13_C8orf46_NA (intron gain)	GGCAGAAATATTGGGAGCTG	TATGCTCGATTCCCACACAG	428
KIAA1267_Exon2_ARL17P1_Exon3	GACCAGAGCTGATCCTGAGC	ATGTAGGCCGAGCTTGTCTG	470
RNF216_Exon7_RBAK_Exon2	TTGTTTCATCCAAGCTGCTG	GGTTTGGTGCTGTCATACCC	318
DEDD_^Exon4_NIT1_Exon6	CGTCATTGATGACCATGAGC	CCCAGATTGCCATAGAGGTC	424
LRRC57_Exon5_SNAP23_Intron+Exon8	TCGGAGCATACCTGACACAG	AAGAGCATGGAACCAAAACG	277
IPO11_NA_SLRN_NA (IPO11_Exon28_SLRN_^Exon2)	TGGAGTTATGGGTCGAGTCC	TTTGCAGTTACATTCCCAAGG	323
SNRPF_Exon2_CCDC38_^Exon12	TTTGCCTCTCAATCCGAAAC	CCAATTCTGCTGCTTTTTCC	251
MIA_Exon3_RAB4B_Exon2	GCCGTTTCCTGACCATACAC	TTCCCACCCACGTTGACTAC	344
NIT1_Exon6_DEDD_^Exon4 (2bp off from exon boundary)	GCCCAGTCTTGAGTCTCCTG	TCCTTGGGGTCTGGTGTTAC	407
Rnf139_Exon1_Ndufb9_Exon2	ATCGACGCCATCTTCAACTC	GAGGCAAATCACCTTCCTTTC	467

Table B.4 The 50 human breast cancer cell lines.

No.	Cell Line	ATCC_Name	Tissue
1	MCF-10A	CRL-10317	breast
2	BT-474	HTB-20	breast
3	Hs 319.T	CRL-7236	breast
4	HCC1428	CRL-2327	breast
5	HCC1599	CRL-2331	breast
6	Hs 605.T	CRL-7365	breast
7	Hs 362.T	CRL-7253	breast
8	ZR-75-1	CRL-1500	breast
9	MCF-7	HTB-22	breast
10	Hs 281.T	CRL-7227	breast
11	HCC1500	CRL-2329	breast
12	BT-20	HTB-19	breast
13	HCC1143	CRL-2321	breast
14	UACC-812	CRL-1897	breast
15	SW527	CRL-7940	breast
16	MDA-MB-453	HTB-131	breast
17	ZR-75-30	CRL-1504	breast
18	MDA-MB-468	HTB-132	breast
19	HCC1187	CRL-2322	breast
20	SK-BR-3	HTB-30	breast
21	MDA-MB-175-VII	HTB-25	breast
22	Hs 574.T	CRL-7345	breast
23	HCC 1008	CRL-2320	breast
24	Hs 742.T	CRL-7482	breast
25	Hs 748.T	CRL-7486	breast
26	BT-483	HTB-121	breast
27	HCC202	CRL-2316	breast
28	HCC 2157	CRL-2340	breast
29	BT-549	HTB-122	breast
30	MDA-MB-415	HTB-128	breast
31	HCC1395	CRL-2324	breast
32		HTB-127	breast
33	MDA-MB-231	HTB-26	breast
34	CAMA-1	HTB-21	breast
35	MDA-MB-134-VI	HTB-23	breast
36	Hs 606.T	CRL-7368	breast
37	HCC1806	CRL-2335	breast
38	HCC1419	CRL-2326	breast

39	AU565	CRL-2351	breast
40	HCC1937	CRL-2336	breast
41	Hs 578T	HTB-126	breast
42	Hs 739.T	CRL-7477	breast
43	DU4475	HTB-123	breast
44	HCC70	CRL-2315	breast
45	HCC38	CRL-2314	breast
46	HCC1954	CRL-2338	breast
47	MB 157	CRL-7721	breast
48	HCC2218	CRL-2343	breast
49	Hs 343.T	CRL-7245	breast
50	UACC-893	CRL-1902	breast

Table B.5 Dog samples for screening

Histology	Tissue Types	ID	Sample
Cancer	Melanomas	1	Parks
		2	Jones
		3	17CM98
		4	CML82-10C2
	Osteosarcomas	5	Abrams
		6	MacKinley
		7	Vogel
		8	D17
		9	Yamada
		10	Gracie
		11	Moresco
	Lymphosarcoma	12	Oswald
		13	1771
	Hemangiosarcoma	14	Denny
		15	Fitz
	Mammary	16	CMT27
		17	CMT12
	Mast Cell Tumor	18	C2 from CL
		19	C2 from Ascites
	Transitional Cell Carcinoma	20	K9TCC
		21	Bliley
	Thyroid Adenocarcinoma	22	CTAC
Normal tissue	Cerebellum	23	
	Spleen	24	
	Mammary	25	
	Ovary	26	
	Pancreas	27	
	Thyroid	28	
	Lung	29	
	Salivary gland	30	
	Small Intestine	31	
	Stomach	32	
	Tonsil	33	
	Heart	34	
	Liver	35	

APPENDIX C

SUPPLEMENTAL: PATTERNS OF GENE FUSIONS

Table C.1 The list of gene fusions used in pattern analysis

Gene Fusion		Gene Fusion		Gene Fusion		Gene Fusion	
5' Gene	**3' Gene**	**5' Gene**	**3' Gene**	**5' Gene**	**3' Gene**	**5' Gene**	**3' Gene**
SEPT8	AFF4	DEDD	NIT1	HMGA1	LAMA4	ITPR2	ETV6
ABHD14A	ACY1	DEK	NUP214	HMGA2	CCNB1IP1	JAZF1	PHF1
AC141586	CCNF	DEPDC1B	ELOVL7	HMGA2	COX6C	JAZF1	SUZ12
ACAD10	ALDH2	DLEU2	PSPC1	HMGA2	CXCR7	KCNQ5	RIMS1
ACBD6	RRP15	DMRT1	BCL6	HMGA2	EBF1	KCTD2	ARHGEF12
ACSF2	CHAD	DSCAML1	MLL	HMGA2	FHIT	KIAA1267	ARL17P1
ACSL3	ETV1	DTX2	PMS2L5	HMGA2	LHFP	KIAA1549	BRAF
ACTB	GLI1	DUSP10	PRDM16	HMGA2	LPP	KIAA1618	ALK
ADHFE1	C8orf46	EBF1	LOC204010	HMGA2	NFIB	KIF5B	PDGFRA
AFF1	DSCAML1	EFTUD2	KIF18B	HMGA2	RAD51L1	KLK2	ETV4
AFF1	ELF2	EIF3K	CYP39A1	HMGA2	WIF1	LCP1	BCL6
AFF1	FXYD6	EIF4A2	BCL6	HMGXB3	PPARGC1B	LDHC	SERGEF
AFF1	PBX1	ELAC1	SMAD4	HN1	USH1G	LEO1	SLC12A1
AFF1	RABGAP1L	ELF2	MLL	HNRPA2B1	ETV1	LIFR	PLAG1
AFF3	BCL2	ELF4	ERG	HOOK3	RET	LMAN2	AP3S1
AGPAT5	MCPH1	EML1	ABL1	HPS4	ASPHD2	LOC100129406	CTTNBP2NL
AHCYL1	RAD51C	EML4	ALK	HSP90AA1	BCL6	LOC100131434	FLJ44451
AKAP9	BRAF	EPC1	PHF1	HSPH1	PREI3	LPP	BCL6
AMD1	GAPDH	ERC1	PDGFRB	IFNGR2	RUNX1	LPP	C12ORF9
ANKHD1	C5orf32	ERO1L	FERMT2	IGH@	BCL10	LRMP	BCL6
ANKRD28	NUP98	EST14	ETV1	IGH@	BCL11A	LRRC57	SNAP23
ARFGEF2	SULF2	ETV6	ABL1	IGH@	BCL2	MACROD1	RUNX1
ARHGAP19	DRG1	ETV6	ABL2	IGH@	BCL3	MALAT1	TFEB
ASPSCR1	TFE3	ETV6	ACSL6	IGH@	BCL6	MALT1	MAP4
ASTN2	PTPRG	ETV6	ARNT	IGH@	BCL8	MBNL1	BCL6
ASTN2	TBC1D16	ETV6	BAZ2A	IGH@	BCL9	MBOAT2	PRKCE
ATIC	ALK	ETV6	CDX2	IGH@	CCND1	MBTPS2	YY2
AX747630	ETV1	ETV6	EVI1	IGH@	CCND2	MDS1	EVI1
BC017255	TMEM49	ETV6	FGFR3	IGH@	CCND3	MED8	ELOVL1
BCAS4	BCAS3	ETV6	FLT3	IGH@	CCNE1	MEF2D	DAZAP1
BCAS4	PRKCBP1	ETV6	FRK	IGH@	CD44	MIA	RAB4B

Gene Fusion		Gene Fusion		Gene Fusion		Gene Fusion	
5' Gene	3' Gene	5' Gene	3' Gene	5' Gene	3' Gene	5' Gene	3' Gene
BCL11B	NKX2	ETV6	GOT1	IGH@	CDK6	MIPOL1	DGKB
BCL11B	TLX3	ETV6	ITPR2	IGH@	CEBPA	MKL1	RBM15
BCL11B	TRD@	ETV6	JAK2	IGH@	CEBPB	MLL	SEPT2
BCL3	MYC	ETV6	MDS1	IGH@	CEBPD	MLL	SEPT5
BCL6	CIITA	ETV6	MDS2	IGH@	CEBPE	MLL	SEPT6
BCL6	IKZF1	ETV6	NCOA2	IGH@	CEBPG	MLL	SEPT9
BCL6	IL21R	ETV6	NTRK3	IGH@	CHST11	MLL	SEPT11
BCL6	PIM1	ETV6	PDGFRA	IGH@	CNN3	MLL	ABI1
BCR	ABL1	ETV6	PDGFRB	IGH@	CRLF2	MLL	ACACA
BCR	FGFR1	ETV6	PER1	IGH@	DDX6	MLL	ACTN4
BCR	JAK2	ETV6	PTPRR	IGH@	EPOR	MLL	AFF1
BCR	PDGFRA	ETV6	RUNX1	IGH@	ERVWE1	MLL	AFF3
BGLAP	PMF1	ETV6	STL	IGH@	ETV6	MLL	AFF4
BIRC3	MALT1	ETV6	SYK	IGH@	FCGR2B	MLL	ARHGAP26
BOLA2	SMG1	EWSR1	ATF1	IGH@	FCRL4	MLL	ARHGEF12
BRCC3	FUNDC2	EWSR1	CREB1	IGH@	FGFR3	MLL	ARHGEF17
BRD3	C15orf55	EWSR1	DDIT3	IGH@	FOXP1	MLL	BCL9L
BRD4	C15orf55	EWSR1	ERG	IGH@	ID4	MLL	C2CD3
BRD4	C15ORF55	EWSR1	ETV1	IGH@	IGL@	MLL	CASC5
BTG1	MYC	EWSR1	ETV4	IGH@	IL3	MLL	CASP8AP2
C15ORF21	ETV1	EWSR1	FEV	IGH@	IRF4	MLL	CBL
C19ORF25	APC2	EWSR1	FLI1	IGH@	KDM4C	MLL	CIP29
C1orf151	NBL1	EWSR1	NFATC2	IGH@	LHX4	MLL	CREBBP
C1QTNF6	IL2RB	EWSR1	NR4A3	IGH@	MAF	MLL	DAB2IP
C20orf29	VISA	EWSR1	PATZ1	IGH@	MAFB	MLL	DCP1A
C22orf39	HIRA	EWSR1	PBX1	IGH@	MALT1	MLL	DCPS
C3ORF27	EVI1	EWSR1	POU5F1	IGH@	MUC1	MLL	EEFSEC
CACNA2D4	WDR43	EWSR1	SP3	IGH@	MYC	MLL	ELL
CANT1	ETV4	EWSR1	WT1	IGH@	MYCN	MLL	EP300
CAPRIN1	PDGFRB	EWSR1	ZNF384	IGH@	NFKB2	MLL	EPS15
CARS	ALK	EWSR1	ZNF444	IGH@	ODZ2	MLL	FLNA
CBFB	MYH11	FBXL18	RNF216	IGH@	PAFAH1B2	MLL	FNBP1
CCDC6	PDGFRB	FCHSD1	BRAF	IGH@	PAX5	MLL	FOXO3
CCDC88C	PDGFRB	FGFR1	PLAG1	IGH@	PCSK7	MLL	FOXO4

Gene Fusion		Gene Fusion		Gene Fusion		Gene Fusion	
5' Gene	3' Gene	5' Gene	3' Gene	5' Gene	3' Gene	5' Gene	3' Gene
CCDC94	MLL	FGFR1	ZNF703	IGH@	RHOH	MLL	FRYL
CCND1	FSTL3	FGFR1OP	FGFR1	IGH@	SPIB	MLL	GAS7
CCND1	TACSTD2	FGFR1OP2	FGFR1	IGH@	TRA@	MLL	GMPS
CCT3	C1orf61	FIP1L1	PDGFRA	IGH@	TRD@	MLL	GPHN
CD74	ROS1	FIP1L1	RARA	IGH@	WHSC1	MLL	KIAA0284
CDH11	USP6	FLJ35294	ETV1	IGK@	BCL10	MLL	LAMC3
CDK5RAP2	PDGFRA	FOXP1	ETV1	IGK@	BCL2	MLL	LASP1
CDK6	MLL	FPGT	TNNI3K	IGK@	BCL3	MLL	LOC100128568
CENPK	MLL	FUS	ATF1	IGK@	BCL6	MLL	LPP
CEP110	FGFR1	FUS	CREB3L1	IGK@	CCND1	MLL	MAML2
CHCHD7	PLAG1	FUS	CREB3L2	IGK@	CCND2	MLL	MAPRE1
CHIC2	ETV6	FUS	DDIT3	IGK@	CDK6	MLL	MLLT1
CHURC1	FNTB	FUS	ERG	IGK@	KDSR	MLL	MLLT10
CIC	DUX4	FUS	FEV	IGK@	MYC	MLL	MLLT11
CIITA	BCL6	FXYD6	MLL	IGK@	PVT1	MLL	MLLT3
CLPTM1L	PVT1	GAPDH	BCL6	IGK@	ZC3H12D	MLL	MLLT4
CLTC	ALK	GAS5	BCL6	IGL@	BCL2	MLL	MLLT6
CLTC	TFE3	GCN1L1	PLA2G1B	IGL@	BCL3	MLL	MYO1F
CLTCL1	ALK	GFOD1	C6orf114	IGL@	BCL6	MLL	NCKIPSD
CNBP	USP6	GIT2	PDGFRB	IGL@	BCL9	MLL	NEBL
CNPY2	CS	GNA12	SHANK2	IGL@	CCND1	MLL	NRIP3
COL1A1	PDGFB	GOPC	ROS1	IGL@	CCND2	MLL	PICALM
COL1A1	USP6	GRB7	PERLD1	IGL@	CCND3	MLL	SH3GL1
COL1A2	PLAG1	GRHPR	BCL6	IGL@	CDK6	MLL	SMAP1
COL6A3	CSF1	HAS2	PLAG1	IGL@	MAF	MLL	SORBS2
COMMD3	BMI1	HCMOGT1	PDGFRB	IGL@	MYC	MLL	TET1
COX19	ADAP1	HDAC11	FBLN2	IGL@	PVT1	MLL	TIRAP
CPSF6	FGFR1	HERPUD1	ERG	IGL@	REL	MLL	TNRC18
CRTC1	MAML2	HERVK	FGFR1	IKZF1	BCL6	MLL	UBE4A
CRTC3	MAML2	HERVK17	ETV1	IL2	TNFRSF17	MLL	VAV1
CTAGE5	SIP1	HERVK22Q11	ETV1	IL6R	ATP8B2	MLL	ZFYVE19
CTNNB1	PLAG1	HIP1	PDGFRB	INPP4A	HJURP	MLLT10	CLP1
CYTH1	PRPSAP1	HISPPD2A	CATSPER2	INTS4	GAB2	MN1	ETV6
DDIT3	MARS	HIST1H4I	BCL6	IPO11	SLRN	MNX1	ETV6
DDX5	ETV4	HJURP	EIF4E2	ITK	SYK	MPO	ZNF296

Gene Fusion		Gene Fusion		Gene Fusion	
5' Gene	**3' Gene**	**5' Gene**	**3' Gene**	**5' Gene**	**3' Gene**
MRPS10	HPR	POU2AF1	BCL6	SSH2	SUZ12
MSI2	HOXA9	PPP2R2A	CHEK2	STAT5B	RARA
MSN	ALK	PRCC	TFE3	STIL	TAL1
MYB	MNX1	PRKAR1A	RARA	STRADB	NOP58
MYB	NFIB	PRKG2	PDGFRB	STRN	PDGFRA
MYC	BCL7A	PRR13	PCBP2	STRN4	TECR
MYC	ZBTB5	PVT1	CHD7	SULF2	PRICKLE2
MYC	ZCCHC7	R3HDM2	NFE2	SUSD1	ROD1
MYH9	ALK	RABEP1	PDGFRB	TAF15	NR4A3
MYO18A	FGFR1	RABGAP1L	MLL	TAF15	ZNF384
MYO18A	PDGFRB	RAD51C	ATXN7	TAX1BP1	AHCY
MYO9B	FCHO1	RAD54B	LOC100128414	TBL1XR1	RGS17
MYST3	CREBBP	RAF1	DAZL	TCEA1	PLAG1
MYST3	EP300	RANBP2	ALK	TCF12	NR4A3
MYST3	NCOA2	RASA2	ACPL2	TCF3	HLF
MYST3	NCOA3	RB1	ITM2B	TCF3	PBX1
MYST4	CREBBP	RBM14	PACS1	TCF3	TFPT
NAIP	OCLN	RBM14	RBM4	TCF3	ZNF384
NAPA	BCL6	RBM15	MKL1	TCTA	TAL1
NDE1	PDGFRB	RBM6	CSF1R	TEX14	PTPRG
NDUFA13	YJEFN3	RC3H2	RGS3	TFG	ALK
NDUFB8	SEC31B	RECK	ALX3	TFG	NR4A3
NDUFC2	KCTD14	RERE	PIK3CD	TFG	NTRK1
NFIA	EHF	RET	CCDC6	TFRC	BCL6
NFKB1	MLL	RET	ERC1	THAP2	TMEM19
NIN	PDGFRB	RET	GOLGA5	THRAP3	USP6
NIT1	DEDD	RET	KTN1	TIA1	DIRC2
NME1	NME2	RET	NCOA4	TIMM23B	LOC100132418
NONO	TFE3	RET	PCM1	TMEM199	SARM1
NOP2	TCF3	RET	PRKAR1A	TMEM88	TLN1
NOS1AP	C1orf226	RET	RFG9	TMPIT	STYXL1

Gene Fusion		Gene Fusion		Gene Fusion	
5' Gene	3' Gene	5' Gene	3' Gene	5' Gene	3' Gene
NPEPPS	USP32	RET	TRIM24	TMPRSS2	ERG
NPM1	ALK	RET	TRIM33	TMPRSS2	ETV1
NPM1	MLF1	RGS22	SYCP1	TMPRSS2	ETV4
NPM1	RARA	RHOH	BCL6	TMPRSS2	ETV5
NUMA1	RARA	RIF1	PKD1L1	TOPORS	DDX58
NUP214	ABL1	RIPK3	ADCY4	TP53BP1	PDGFRB
NUP214	XKR3	RMND5A	ANAPC1	TPM3	ALK
NUP98	ADD3	RNF103	VPS24	TPM3	NTRK1
NUP98	CCDC28A	Rnf139	Ndufb9	TPM3	PDGFRB
NUP98	DDX10	RNF216	RBAK	TPM3	TPR
NUP98	HHEX	RPL11	TCEB3	TPM4	ALK
NUP98	HOXA11	RPN1	EVI1	TPR	NTRK1
NUP98	HOXA13	RPN1	PRDM16	TRA@	CDKN2A
NUP98	HOXA9	RPS10	HPR	TRA@	IRF4
NUP98	HOXC11	RPS6KB1	TMEM49	TRA@	MTCP1
NUP98	HOXC13	RRM2	C2orf48	TRA@	MYC
NUP98	HOXD11	RSBN1	BCAS3	TRA@	NOTCH1
NUP98	HOXD13	RUNX1	AFF3	TRA@	OLIG2
NUP98	IQCG	RUNX1	CBFA2T3	TRA@	PVRL2
NUP98	KDM5A	RUNX1	CPNE8	TRA@	TCL1A
NUP98	LNP1	RUNX1	EVI1	TRA@	TRB@
NUP98	NSD1	RUNX1	FGA7	TRB@	CCND2
NUP98	PHF23	RUNX1	LPXN	TRB@	EVI1
NUP98	PRRX1	RUNX1	MDS1	TRB@	HOXA@
NUP98	PRRX2	RUNX1	PRDM16	TRB@	HOXA10
NUP98	PSIP1	RUNX1	PRDX4	TRB@	HOXA11
NUP98	RAP1GDS1	RUNX1	RPL22P1	TRB@	IRS4
NUP98	SETBP1	RUNX1	RUNX1T1	TRB@	LCK
NUP98	TOP1	RUNX1	SH3D19	TRB@	LMO1
NUP98	TOP2B	RUNX1	TRPS1	TRB@	LMO2
NUP98	WHSC1L1	RUNX1	USP42	TRB@	LYL1
ODZ4	NRG1	RUNX1	YTHDF2	TRB@	MTCP1
OMD	USP6	RUNX1	ZFPM2	TRB@	MYB
P2RY8	CRLF2	RUNX1	ZNF687	TRB@	NOTCH1

Gene Fusion		Gene Fusion		Gene Fusion	
5' Gene	3' Gene	5' Gene	3' Gene	5' Gene	3' Gene
PAPOLA	AK7	RYK	ATP5O	TRB@	TAL1
PARP1	MIXL1	SAMD12	PHF20L1	TRB@	TAL2
PAX3	FOXO1	SAMD12	PVT1	TRB@	TLX1
PAX3	FOXO4	SAMD5	SASH1	TRB@	TRG@
PAX3	NCOA1	SCAMP2	WDR72	TRD@	LMO1
PAX3	NCOA2	SDHAF2(C11orf79)	C11orf66	TRD@	LMO2
PAX5	ASXL1	SDHD	TEX12	TRD@	NKX2
PAX5	BRD1	SEC31A	ALK	TRD@	PVT1
PAX5	C20ORF112	SET	NUP214	TRD@	RANBP17
PAX5	DACH1	SFPQ	ABL1	TRD@	TAL1
PAX5	ELN	SFPQ	EIF5A	TRD@	TLX1
PAX5	ETV6	SFPQ	TFE3	TRD@	TLX3
PAX5	FOXP1	SFRS3	BCL6	TRG@	IGH@
PAX5	HIPK1	SLC12A7	C11orf67	TRG@	TRB@
PAX5	JAK2	SLC20A2	DBX2	TRIM24	FGFR1
PAX5	KIF3B	SLC26A6	PRKAR2A	TRIM61	FARSB
PAX5	LOC392027	SLC34A2	ROS1	TRIP11	PDGFRB
PAX5	PML	SLC45A3	ELK4	TTL	ETV6
PAX5	POM121	SLC45A3	ERG	TXLNG	SYAP1
PAX5	SLCO1B3	SLC45A3	ETV1	TYMP	SCO2
PAX5	ZNF521	SLC45A3	ETV5	UBR4	GLB1
PAX7	FOXO1	SMYD3	ZNF695	USP10	ZDHHC7
PAX8	PPARG	SNHG5	BCL6	USP16	RUNX1
PBX1	MLL	SNRPF	CCDC38	WDR51B	GALNT4
PCM1	JAK2	SPOCK1	TBC1D9B	WDR55	DND1
PDCD1LG2	C18orf10	SPTBN1	FLT3	WRB	SH3BGR
PDE4DIP	PDGFRB	SPTBN1	PDGFRB	ZBTB16	RARA
PICALM	MLLT10	SRGAP3	RAF1	ZDHHC7	ABCB9
PIK3C2A	TEAD1	SRP9	RPS8	ZEB2	LOC100128821
PLA2R1	RBMS1	SS18	SSX	ZMIZ1	ABL1
PLCXD2	PHLDB2	SS18	SSX1	ZMYM2	FGFR1
PLXND1	TMCC1	SS18	SSX2	ZNF294	TIAM1
PMF1	BGLAP	SS18	SSX4	ZNF649	ZNF577
PML	RARA	SS18L1	SSX1		
POLR2J3	UPK3B	SSBP2	JAK2		

Gene Fusion	Class
TMPRSS2+ERG	Dominant
TAF15+NR4A3	Dominant
SSX2+SS18	Dominant
SSX1+SS18	Dominant
SS18+SSX2	Dominant
SS18+SSX1	Dominant
NPM1+ALK	Dominant
KIAA1549+BRAF	Dominant
JAZF1+SUZ12	Dominant
FUS+DDIT3	Dominant
EWSR1+WT1	Dominant
EWSR1+NR4A3	Dominant
EWSR1+FLI1	Dominant
EWSR1+ERG	Dominant
EWSR1+CREB1	Dominant
EWSR1+ATF1	Dominant
ETV6+NTRK3	Dominant
ATIC+ALK	Dominant
COL1A1+PDGFB	Multiple dominant
ASPSCR1+TFE3	Multiple dominant
TMPRSS2+ETV1	No dominant
TFG+ALK	No dominant
SRGAP3+RAF1	No dominant
SEC31A+ALK	No dominant
RANBP2+ALK	No dominant
RAF1+ESRP1	No dominant
PRCC+TFE3	No dominant
MYB+NFIB	No dominant
MSN+ALK	No dominant
FUS+ERG	No dominant
FUS+CREB3L2	No dominant
FUS+CREB3L1	No dominant
EWSR1+ZNF384	No dominant
EWSR1+SP3	No dominant
EWSR1+POU5F1	No dominant
EWSR1+PBX1	No dominant
EWSR1+PATZ1	No dominant
EWSR1+ETV4	No dominant
EWSR1+DDIT3	No dominant
EML4+ALK	No dominant
CREB3L2+FUS	No dominant
CLTC+ALK	No dominant

Table C.2 The list of gene fusions with iso-forms

APPENDIX D

SUPPLEMENTAL: CODING MICROSATELLITE DNA

Table D.1 The sequences of frame-shifted peptides derived from Indels

Accession #	Gene	MS position	MS	FS pep - Insertion	FS pep - Deletion
NM_000983.3	RPL22	83	A_8	EASSEVHS	RSKF
NM_007126.3	VCP	2228	A_9	CVHHWRYQPA	MCSSLALPTGLTSLILPSSDLAVLISSSTSH FLMRSPVLPSSRLTCASPQLPRMWTWSS WLK
NM_000918.3	P4HB	1350	A_8	RLCGVLCPMVWSLQTVGSHLG	TSLWSSMPHGVVTANSWLPFGINWERRT RTMRTSSSPRWTRLPTRWRPSKCTASPH SSSFLPVPTGRSLITTGNARWMVLRNSW RAVARMGQGMMTISRTWKKQRSQTWR KTMIRKL
NM_012073.3	CCT5	891	A_7	SGRCEDCNSHMSI	WKMRRLQFSHVHLNHPNQKQSISWM
NM_003377.3	VEGFB	419	A_8	GQCCEARQGCHSPPPSPAPFCS GLGLCPRSTLPS	RTVL
NM_016056.2	TMBIM4	587	T_{10}	IL	YTLSLYGHLYMRVLP
NM_002791.1	PSMA6	648	A_8	SEEEI	*
NM_003262.3	SEC62	452	A_9	RKREKKRW	KKRKKKMVKRKNPKRRKLQELLKRRK LRKNSNLSHMMIRFFWMEMRCMYGSMT QFTLKHLSWD
NM_005520.2	HNRNPH1	1038	T_8	*	LTAKFKMGLKVFVSSTPEKADQVARLLL NLNQKMKSNWP
NM_014972.2	TCF25	467	A_9	TEKQEKQHGRSIGKRTRRYRS HPRED	NRKTRKAAREKHRKTD
NM_006842.2	SF3B2	2658	A_8	TESSAPGQPWGQQEI	NGKLSPRTAVGAARNIRSSSF

APPENDIX E

SUPPLEMENTAL: ALTERNATIVE SPLICING

Table E.1 The list of reaming 76 putative tumor-associated splicing variants

Gene	RefSeq_ID	FS length	# Tumor_lib	Tissue types	# Normal_lib	Tissue types
NRM	NM_007243.1	20	3	uncharacterized tissue,brain,prostate	0	
PRSS27	NM_031948.3	15	3	uncharacterized tissue,cervix,esophagus	0	
TXN2	NM_012473.3	9	3	salivary gland,prostate,cervix	0	
RDH11	NM_016026.3	36	3	uncharacterized tissue,lung,testis	0	
BORA	NM_024808.2	30	3	bone marrow,uncharacterized tissue,brain	0	
RPS3	NM_001005.3	30	3	lung,eye	0	
SAAL1	NM_138421.2	28	3	stomach,skin,testis	0	
SEMA3B	NM_001005914.1	62	3	stomach,pancreas,uncharacterized tissue	0	
FPGS	NM_001018078.1	101	3	uncharacterized tissue,lung	0	
SLC13A3	NM_001193342.1	35	3	kidney,uncharacterized tissue,testis	0	
ARHGEF1	NM_199002.1	28	3	cartilage,bone,uncharacterized tissue	0	
FANCI	NM_001113378.1	39	3	uncharacterized tissue	0	
SARS2	NM_017827.3	23	4	bone,uterus,brain	0	
CAPN3	NM_000070.2	15	3	uncharacterized tissue,thyroid	0	
SPAG5	NM_006461.3	13	3	uncharacterized tissue	0	
ZNF263	NM_005741.4	42	3	skin,brain,placenta	0	
DFFA	NM_213566.1	12	3	mammary gland,adrenal cortex,muscle	0	
NSL1	NM_001042549.1	12	3	lung,cervix	0	
C17orf85	NM_018553.3	35	3	uterus,uncharacterized tissue,brain	0	
CIRH1A	NM_032830.2	15	3	skin,cervix	0	
APEH	NM_001640.3	11	3	skin,lymphoreticular,prostate	0	
DPP3	NM_130443.2	22	3	cartilage,uterus,lung	0	
EEF1A1	NM_001402.5	44	3	stomach,mammary gland,eye	0	
ARMC8	NM_014154.2	40	3	uncharacterized tissue,pancreatic islet	0	
TMEM179	NM_207379.1	36	3	brain,lung	0	
VASP	NM_003370.3	41	3	uncharacterized tissue,cervix,muscle	0	
MRPL43	NM_032112.2	19	3	uterus	0	
MRPL43	NM_032112.2	9	3	skin,ovary,pancreatic islet	0	
DEDD	NM_001039712.1	26	3	pancreas,uncharacterized tissue,brain	0	
AURKB	NM_004217.2	38	3	uterus,pancreas,brain	0	
CRCP	NM_001040648.1	12	3	stomach,brain,colon	0	
WTAP	NM_152858.1	19	3	skin,brain,adrenal cortex	0	
NUP43	NM_198887.1	37	3	bone,parathyroid,uncharacterized tissue	0	
SRSF5	NM_006925.3	15	5	bone,stomach,parathyroid	1	eye
HSPH1	NM_006644.2	9	4	kidney,skin,testis	1	testis
IGFLR1	NM_024660.2	22	4	skin,lymphoreticular,lung,eye	1	pooled tissue
NUDT8	NM_181843.1	100	4	uterus,uncharacterized tissue,lung	1	eye
STK25	NM_006374.3	20	4	ovary,lymph node,bone marrow,muscle	1	spleen
TNFAIP2	NM_006291.2	64	4	brain,prostate,lung,colon	1	pooled tissue
TTLL12	NM_015140.3	14	4	bone,uncharacterized tissue	1	liver
UQCC	NM_001184977.1	13	4	brain	1	cerebrum
WDR34	NM_052844.3	72	4	liver,brain,cervix,placenta	1	brain
KRT18	NM_199187.1	90	8	pancreas,liver,mammary gland,skin,lymphoreticular,colon	2	embryonic tissue,prostate
NOP16	NM_016391.4	17	8	ovary,cartilage,lymph node,bone,uterus,parathyroid,lung	2	heart,placenta
SNX27	NM_030918.5	39	7	cartilage,bone,uterus,lung,colon	2	kidney,lung
RAB25	NM_020387.2	14	10	ovary,lymph node,stomach,mammary gland,uncharacterized tissue,brain,prostate,colon,placenta	3	pooled tissue,embryonic tissue,lung
ATP5B	NM_001686.3	22	3	bone marrow,uncharacterized tissue,prostate	1	pooled tissue
BFAR	NM_016561.2	53	3	liver,uncharacterized tissue	1	brain
C16orf62	NM_020314.5	10	3	mammary gland,lung,eye	1	brain
C19orf40	NM_152266.3	34	3	ovary,skin,brain	1	uncharacterized tissue
C9orf140	NM_178448.3	167	3	kidney,brain,lung	1	embryonic tissue
DERA	NM_015954.2	16	3	uncharacterized tissue,eye	1	cerebrum
EXOSC2	NM_014285.5	10	3	uncharacterized tissue,brain,lung	1	colon
GTPBP5	NM_015666.3	22	3	kidney,testis	1	brain
HSPH1	NM_006644.2	9	3	kidney,uncharacterized tissue	1	lung
IST1	NM_014761.2	64	3	uncharacterized tissue	1	brain
MAGED2	NM_177433.1	36	3	ovary,pancreatic islet,brain	1	lung
MED19	NM_153450.1	34	3	kidney,brain	1	eye
MRPL2	NM_015950.3	14	3	bone,mammary gland,parathyroid	1	lung
MTFR1	NM_001145839.1	13	3	kidney,uncharacterized tissue,brain	1	skin
MVK	NM_000431.2	8	3	liver,brain,prostate	1	embryonic tissue

Gene	RefSeq_ID	FS length	# Tumor_lib	Tissue types	# Normal_lib	Tissue types
NUPL2	NM_007342.2	16	3	uncharacterized tissue,brain	1	cerebellum
PEX13	NM_002618.3	49	3	uncharacterized tissue,brain,prostate	1	testis
PSPH	NM_004577.3	29	3	liver,uncharacterized tissue,placenta	1	embryonic tissue
REM3	NM_006743.4	88	3	lymph node,lymphoreticular,prostate	1	testis
RNF217	NM_152553.2	11	3	pancreas,uncharacterized tissue,skin	1	kidney
RPL7L1	NM_198486.2	12	3	uncharacterized tissue,brain	1	pancreatic islet
SENP2	NM_021627.2	8	3	liver,uncharacterized tissue,brain	1	brain
SLC29A2	NM_001532.2	12	3	ovary,uterus,cervix	1	colon
SLC35B2	NM_178148.2	25	3	bone marrow,uncharacterized tissue	1	testis
TH	NM_199293.2	13	3	uncharacterized tissue,brain	1	placenta
WIPI2	NM_001033519.1	22	3	uterus,brain,placenta	1	endocrine
XRRA1	NM_182969.1	22	3	parathyroid,uncharacterized tissue,lung	1	testis
CIRH1A	NM_032830.2	18	6	bone,uncharacterized tissue,cervix,colon	2	kidney,prostate
TATDN2	NM_014760.3	41	6	liver,uncharacterized tissue,testis,gastrointestinal tract	2	embryonic tissue,uncharacterized tissue
GTSE1	NM_016426.6	25	9	ovary,bone marrow,skin,brain,esophagus,muscle,testis	3	liver,skin,brain
C19orf2	NM_134447.1	28	12	cartilage,ovary,bone,uterus,kidney,uncharacterized tissue,cervix,thyroid,colon	4	skin,lung,testis,placenta

Table E.2 The sequences of frame-shifted peptides from splicing variants

Gene	RefSeq ID	FS sequences
C11orf2	NM_013265.2	PCTGLSLHPMAPRIWSRWSFPAGRCQDRPNKHVWPPQKKKKKK KKKKK
C20orf96	NM_080571.1	CFTSSPLRW
CYBASC3	NM_001161452.1	LLLQLRPGSRPFPVTYVSVTGRQPYKSW
KRT8	NM_002273.3	LLRSRHSTRILPTAAGLRLRACTRSSMRSCRAWLGSTGMTCGAQR LRSLR
MVK	NM_001114185.1	GGPRRIWS
NAA10	NM_003491.2	RSVKWSPNTMQMGRTPMP
PDCD2	NM_001199462.1	GLWLFRPQNVLQMPQSILLQQGASDPRLEIGT
RPS3A	NM_001006.3	FGKAHGASW
TFE3	NM_006521.4	CSAQARNRSEDETQPLPLGTLLAF
HNRNPA2B1	NM_031243.2	KEGVLLQVTNEEVVNHRVFKK
NOL12	NM_024313.2	VPTACCRCCFCWDV
RPLP0	NM_001002.3	GVRQWQHLQP
DPH2	NM_001384.4	LPCSSLTSYWEMLWLWLHDWRRRQGQRCSFWVTQPTAAAAWM CWVLSKLELRLSYILALPA
GNB2L1	NM_006098.4	GWPGHVMGSQRRQTPLHARWWGHHQRPVLQP
RPL8	NM_000973.3	IRELCHRYLPQP
IGFLR1	NM_024660.2	NCPVWRHNPCLASWMSWRCWKS
KARS	NM_001130089.1	VGSMPKELLGESSSSMIFEERG
MRPS28	NM_014018.2	EIPERNQGPVAAIRS
HNRNPA2B1	NM_031243.2	EGVLLQVTNEEVVNHRVFKK
SMC1A	NM_006306.2	CCGIYCHEEPQREDSSI
NRM	NM_007243.1	AGDAVLGAHTQRPCVVGGSG
PRSS27	NM_031948.3	PLRRPCTRSCWGQGS
TXN2	NM_012473.3	CQRCPLCWP
RDH11	NM_016026.3	SLPPNPSAARETKGISPIKDSKCVFPRTSPGKDPLP
BORA	NM_024808.2	FSLKMSSYPLLGLIMKGNSFHNVIPVNALT
RPS3	NM_001005.3	GLLWCAAVHHGEWGQRLRGCGVWETPRTEG
SAAL1	NM_138421.2	GDGGSGSKGRPVEQTEVFLCISKPSSFL
SEMA3B	NM_001005914.1	LPQQDLWHLQFHQGLPRRCHPVCAEPPPHVQLCPAHWGAPSFPT SWSQLHLHSNCRGPGCSR

Gene	RefSeq ID	FS sequences
FPGS	NM_001018078.1	VLGSQRHPGQGSCGSCPWHLCSSPHPTCGSGFGTRSGRAGRRCCG AGPSPGTWTVRTPPAARRPACAGSARRCRAARGRAVAPRFESCS SMLPGTGTRRPC
SLC13A3	NM_001193342.1	GIGAVCMDWWAAAPPGECAPRPGCAAHHCGHRLLH
ARHGEF1	NM_199002.1	GVGGGILPPETPPVSAWGELCPPAWLHL
FANCI	NM_001113378.1	VSPGVSELRRNSKKYGKAGEAVWFSSDPPVLFFHFLRTE
SARS2	NM_017827.3	LHARAPGPRGPPLLCPCCLRVSH
CAPN3	NM_000070.2	CLQKHLPVALSTSLC
SPAG5	NM_006461.3	ISVSIMWTQRRKL
ZNF263	NM_005741.4	SHSQSGGPRHPGGTRRKAMGSQCPELQGGPEPQRPSSRRREI
DFFA	NM_213566.1	SPKLPLVRRWMQ
NSL1	NM_001042549.1	GAKPGGLALGAV
C17orf85	NM_018553.3	CYQHPFPKKSQFPGAYWTSFEGEEEGSGQLTLPGP
CIRH1A	NM_032830.2	LLSSHHPLKRRNLEP
APEH	NM_001640.3	SPSQAMWATRM
DPP3	NM_130443.2	HFPACQLLPLCDLISSALPYVE
EEF1A1	NM_001402.5	CLQNWWYWYCSCWPSGDWCSQTRYGGHLCSSQRYNGSKICRN AP
ARMC8	NM_014154.2	RHEKCCNWKQQAESQSHCFRSCSKIVVLASARNLKHRAEN
TMEM179	NM_207379.1	QFRTPGWPLKALAGRGWPEDASPGQEPSKGAGRGWA
VASP	NM_003370.3	WPQLLLEPNSGKSASRRRPQGGPQPPKLRVVEAEVGDSWKR
MRPL43	NM_032112.2	PASGGSDLVNHSFLCKWHP
MRPL43	NM_032112.2	CLLLGAVTL
DEDD	NM_001039712.1	AAAAAHHHSPRPAALRHPQEETGCVP
AURKB	NM_004217.2	DHGGVGRCSNVLPWEEGDSQRHKARKSALRAQGRAEDC
CRCP	NM_001040648.1	TSASQIQAILVP
WTAP	NM_152858.1	GLMASDYSEEVATSEKFPF
NUP43	NM_198887.1	QENCSNPGGRGCSDPRSCHFTPAWAKEQNAISKNIHI
SRSF5	NM_006925.3	VKGVLHSLTAAGQTH
HSPH1	NM_006644.2	DSCGIVNSY
IGFLR1	NM_024660.2	NCPVWRHNPCLASWMSWRCWKS
NUDT8	NM_181843.1	DEVFALPLAHLLQTQNQGYTHFCRGGHFRYTLPVFLHGPHRVWG LTAVITEFALQLLAPGTYQPRLAGLTCSGAEGLARPKQPLASPCQ ASSTPGLNKGL
STK25	NM_006374.3	KHQAMDHHGVPGRRLSTGLA
TNFAIP2	NM_006291.2	PRAAVSGIQQWWNGRQNWKRKKEKMSSRLAGAFRVLWRAVST ASIRRHIQVAPRPLQAGPAMGP
TTLL12	NM_015140.3	LIVGGGAPDRKGFQ
UQCC	NM_001184977.1	GVRCLIHSIHGFL
WDR34	NM_052844.3	VAARAWAQPPLPGAECGHRREGATLAGHRGRPAAAHRGLRPGH AAAATEHQAQEASPRGDRGGRHGSGLLQL
KRT18	NM_199187.1	HRDSRGSGRNGRHPEREGDHAKPERPPGLLPGQSEEPGDREPEAG EQNPGALGEEGTPGQRLEPLLQDHRGPEGSDLRKYCGQCPHRSA D
NOP16	NM_016391.4	SGKTSSILCRRGRWRWS
SNX27	NM_030918.5	HFPDGEVTAERCGHLAFPYPLPFPSPPSSYSFHVPFQTE
RAB25	NM_020387.2	GTIVVQWGPSWCLT
ATP5B	NM_001686.3	TTNPSRISLPSWVWMNFLRKTS
BFAR	NM_016561.2	WSCSSITGAAGNLNTTSWSTRLWPNGRRKKLSSGWSSWALGHLF TGKGFYLNE

Gene	RefSeq ID	FS sequences
C16orf62	NM_020314.5	GSADRDDGKV
C19orf40	NM_152266.3	DAAFFMSPKLIWWQEMATERGLFGLEIPIILKEL
C9orf140	NM_178448.3	RVQGTLVHCPTRHLSQRRGPGRQRGNSLPEPSSMLTCPQQPHRAT FPAAPGLQGCPRTGPSQPSMQLPSYPEDGSGLSRGHKDVRPGPPG QERVQVLRACAPQPQHQVDCSAVGGPVAAREKPPVSRLGSAHQ GLPTSAFEGACHALGDPGIFTGLEAGDRTVSVPG
DERA	NM_015954.2	LLQPPFVFIPPGCVML
EXOSC2	NM_014285.5	GFWSRFPPPW
GTPBP5	NM_015666.3	GPRGHAGEGGRQSCGRPVLRGR
HSPH1	NM_006644.2	DSCGIVNSY
IST1	NM_014761.2	IVGPGPKPEASAKLPSRPADNYDNFVLPELPSVPDTLPTASAGAST SASEDIDFDDLSRRFEEL
MAGED2	NM_177433.1	RCQPDRHSHIWALRWPWWSWCQHQWQLWCLWFLLQV
MED19	NM_153450.1	ETPSDSDHKKKKKKKEEDPERKRKKKEKKKKKVE
MRPL2	NM_015950.3	AGNVRSNSRPSIQR
MTFR1	NM_001145839.1	LHWGSTKVHLLLI
MVK	NM_000431.2	GGPRRIWS
NUPL2	NM_007342.2	AKFCPTFNKSMEEQGK
PEX13	NM_002618.3	DYRRLPPGPANFFCIFSRDGVSPCYPGWSPSPDLVMSPLRSPKVLG LQA
PSPH	NM_004577.3	CDLNSLCIFVAIFHTKCFKCGESIKHLYS
RBM3	NM_006743.4	GLWMVVRSVWIMQASLLGEPEEVALGPMGVVAATLEVVGTRAM GVAGIMTVDLEGMDMDMDMDVPETIMAETRVVMTATQEEITETIMT T
RNF217	NM_152553.2	GLFVFPIYCLC
RPL7L1	NM_198486.2	EVWRHLLGRPHS
SENP2	NM_021627.2	GIFELFIL
SLC29A2	NM_001532.2	SPCPSSPPSQPW
SLC35B2	NM_178148.2	VLSDLGCAAGKSDDPQLWGHSHITG
TH	NM_199293.2	HQALGAVPSCEGV
WIPI2	NM_001033519.1	RYGRCVHCREIVLQQPSGHRQP
XRRA1	NM_182969.1	DRKRGCCPTSSSLPISLRVRLS
CIRH1A	NM_032830.2	MTSLLSSHHPLKRRNLEP
TATDN2	NM_014760.3	GDQQPDRTQAGLKSVSQVEDVFRELIGTQKTRTGCFPPSGS
GTSE1	NM_016426.6	VQMKMMKSSSDPLDIKKDVLLPAWN
C19orf2	NM_134447.1	GFAASWLFKKPRPSECHTVIFKEESYMN

APPENDIX F

SUPPLEMENTAL: SEQUENCE DATABASES

F.1 EST sequence database

EST sequences are downloaded from ftp://ftp.ncbi.nlm.nih.gov/blast/db/FASTA. The information about the libraries in EST Db were obtained from the file called Hs_LibData.dat. This file was downloaded from CGAP under National Cancer Institute (NCI); cgap.nci.nih.gov/Info/CGAPDownload. There were 8,627 libraries from normal and cancer samples in this file (Table F.1).

Table F.1 The number of libraries in EST database by their origin of sample.

Type	No. of EST library
Normal	2284
Neoplasia	4410
Preneoplasia	10
Uncharacterized	1923

Table F.2 Tumor EST libraries 41 tissue types were presented in tumor libraries.

Tissue type	No. of libraries	Tissue type	No. of libraries
bone marrow	5	cerebellum*	5
placenta	5	liver	27
peripheral nervous system	1	skin	13
cervix	10	adrenal medulla	1
head and neck	641	soft tissue	4
stomach	248	brain	275
prostate	157	bone	9
pancreas	18	kidney	136
esophagus	4	ovary	158
colon	783	adrenal cortex	1
nervous	11	germ cell	6
endocrine	3	pancreatic islet*	1
pituitary gland	1	adipose	1
eye	3	mammary gland	731
thymus	4	lymphoreticular	18
lung	191	synovium	1

lymph node	8	genitourinary	81
salivary gland	2	cartilage	13
testis	16	uterus	112
gastrointestinal tract	2	muscle	6
thyroid	5		

Table F.3 Normal EST libraries 48 tissue types were presented in normal libraries.

Tissue type	No. of libraries	Tissue type	No. of libraries
ear	2	thyroid	6
bone marrow	11	cerebellum	4
placenta*	353	liver	24
peripheral nervous system	5	vascular	16
cervix	1	pineal gland	3
head and neck	45	skin	11
uncharacterized tissue	102	soft tissue	4
stomach	75	brain	66
prostate	143	bone	4
pancreas	10	kidney	14
esophagus	1	spleen	6
colon	138	ovary	9
nervous	5	pancreatic islet	9
endocrine	8	adipose	6
pituitary gland	5	whole body	15
eye	25	mammary gland	337
thymus	7	lymphoreticular	17
lung	103	synovium	2
lymph node	10	retina	17
pooled tissue	99	genitourinary	13
salivary gland	4	cartilage	4
testis	156	uterus	6
heart	15	muscle	9
gastrointestinal tract	4	cerebrum	355

F.2 RNA-seq data

Table F.4 The table of RNA-Seq data

PubMed ID	Accession	Sample ID	Sample Description
21247443	SRX025833	SRR064437	normal_breast
	SRX025832	SRR064287	KPL-4
	SRX025831	SRR064441	SKBR3-2
	SRX025830	SRR064440	SKBR3-1
	SRX025829	SRR064439	BT474-2
	SRX025828	SRR064438	BT474-1
	SRX025827	SRR064286	MCF7
21571633	SRX022089	SRR057658	normal_N23
	SRX022088	SRR057657	normal_N19
	SRX022087	SRR057656	normal_N15
	SRX022086	SRR057655	normal_N13
	SRX022085	SRR057654	normal_N11
	SRX022084	SRR057653	normal_N09
	SRX022083	SRR057652	normal_N08
	SRX022082	SRR057651	normal_N06
	SRX022081	SRR057650	normal_N03
	SRX022080	SRR057649	normal_N02
	SRX022079	SRR057648	Prostate_carcinoma_C40
	SRX022078	SRR057647	Prostate_carcinoma_C39
	SRX022077	SRR057646	Prostate_carcinoma_C37
	SRX022076	SRR057645	Prostate_carcinoma_C33
	SRX022075	SRR057644	Prostate_carcinoma_C29
	SRX022074	SRR057643	Prostate_carcinoma_C27
	SRX022073	SRR057642	Prostate_carcinoma_C23
	SRX022072	SRR057641	Prostate_carcinoma_C19
	SRX022071	SRR057640	Prostate_carcinoma_C18
	SRX022070	SRR057639	Prostate_carcinoma_C16
	SRX022069	SRR057638	Prostate_carcinoma_C15
	SRX022068	SRR057637	Prostate_carcinoma_C13
	SRX022067	SRR057636	Prostate_carcinoma_C11
	SRX022066	SRR057635	Prostate_carcinoma_C09
	SRX022065	SRR057634	Prostate_carcinoma_C08
	SRX022064	SRR057633	Prostate_carcinoma_C07
	SRX022063	SRR057632	Prostate_carcinoma_C06
	SRX022062	SRR057631	Prostate_carcinoma_C05
	SRX022061	SRR057630	Prostate_carcinoma_C03
	SRX022060	SRR057629	Prostate_carcinoma_C02

20179022	SRX006135	SRR018269	leukemia cell line K-562
	SRX006134	SRR018268	leukemia cell line K-562
	SRX006133	SRR018267	melanoma patient-derived short-term culture
	SRX006132	SRR018266	melanoma cell line 501 Mel
	SRX006131	SRR018265	melanoma patient-derived short-term culture
	SRX006130	SRR018264	melanoma cell line MeWo
	SRX006129	SRR018263	melanoma cell line MeWo
	SRX006128	SRR018262	melanoma patient-derived short-term culture
	SRX006127	SRR018261	melanoma patient-derived short-term culture
	SRX006126	SRR018260	melanoma patient-derived short-term culture
	SRX006125	SRR018259	melanoma patient-derived short-term culture
	SRX006124	SRR018258	melanoma patient-derived short-term culture
	SRX006123	SRR018257	melanoma patient-derived short-term culture
	SRX006122	SRR018256	melanoma cell line MeWo

APPENDIX G

ALL POSSIBLE FRAME-SHIFTED PEPTIDES

I examined all possible frame-shifted (FS) peptides from coding sequences. Basically, all coding sequences were translated from second and third nucleotide. The translated sequences were splited into each piece by stop codons in the middle. The second reading frame of all mRNA generated 803,872 peptides while the third reading frame yielded 1,049,355 peptides. Figure G.1 shows the distribution of their lengths. Frame-shifted mutations are prone to have more than one premature stop codon at the end according to this distrituion. However, short frame-shifted peptides wer not made by the excessive number of stop codons. The number of stop codins in the second and third reading frames (see Figure G.2) were not more or less than expected number by the frequency in the codon table (3 stop codons out of all 64 codons). Therefore, the relative positions of stop codons actually generated the shorter frame-shifted peptides.

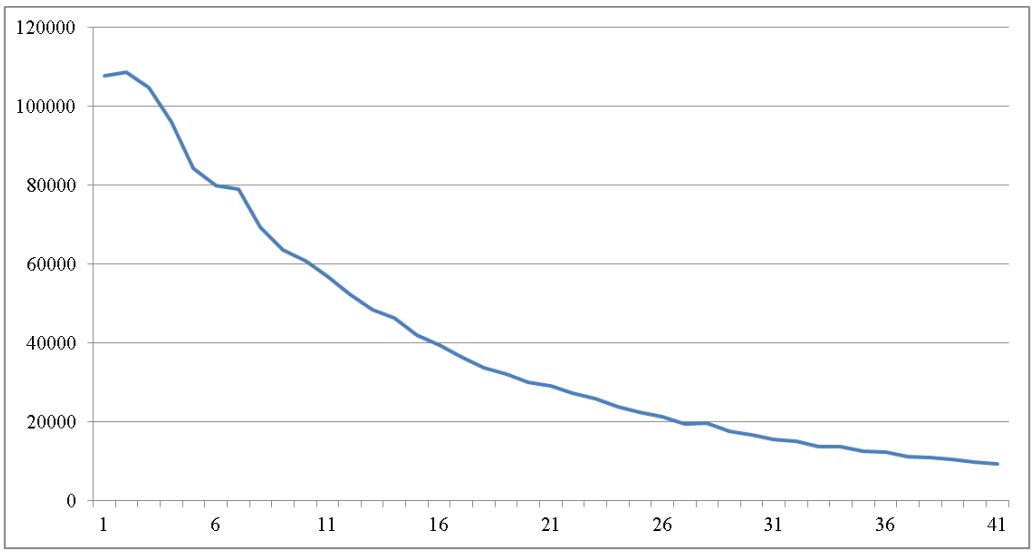


Figure G.1 The length of frame-shifted peptides from coding sequences. Interestingly, the distribution of length of peptides was totally skewed to 0 or 1 amino acid long. About 50% of frame-shifted peptides were 10 amino acids long

or less. X-axis indicated the length of frame-shifted peptide while y-axis indicated the number of peptides of that length.

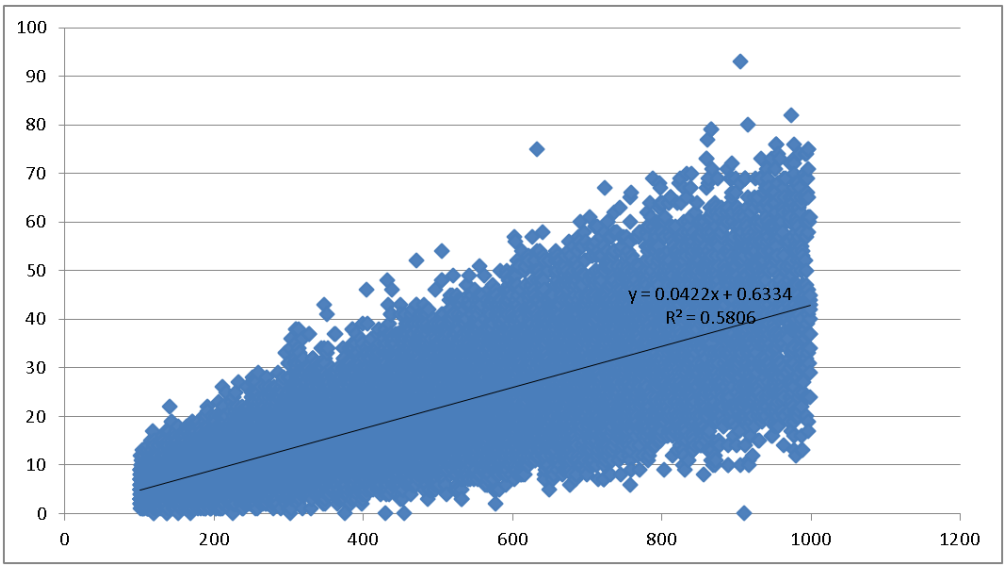


Figure G.2 The number of stop codons. This figure shows the number of stop codons from sequence of which lengths ranged from 100 amino acids to 1,000 amino acids. Most of translated peptides from second and third frame of coding sequences have an expected number of stop codons.

APPENDIX H

INSTITUTIONAL REVIEW BOARD (IRB)

ASU Knowledge Enterprise Development

Office of Research Integrity and Assurance

To:	Stephen Johnston BDB
From:	Carol Johnston, Chair Biosci IRB
Date:	10/24/2011
Committee Action:	**Renewal**
Renewal Date:	10/24/2011
Review Type:	Expedited F5
IRB Protocol #:	0912004625
Study Title:	Profiling Human Sera for Unique Antibody Signatures
Expiration Date:	11/21/2012

The above-referenced protocol was given renewed approval following Expedited Review by the Institutional Review Board.

It is the Principal Investigator's responsibility to obtain review and continued approval of ongoing research before the expiration noted above. Please allow sufficient time for reapproval. Research activity of any sort may not continue beyond the expiration date without committee approval. Failure to receive approval for continuation before the expiration date will result in the automatic suspension of the approval of this protocol on the expiration date. Information collected following suspension is unapproved research and cannot be reported or published as research data. If you do not wish continued approval, please notify the Committee of the study termination.

This approval by the Biosci IRB does not replace or supersede any departmental or oversight committee review that may be required by institutional policy.

Adverse Reactions: If any untoward incidents or severe reactions should develop as a result of this study, you are required to notify the Biosci IRB immediately. If necessary a member of the IRB will be assigned to look into the matter. If the problem is serious, approval may be withdrawn pending IRB review.

Amendments: If you wish to change any aspect of this study, such as the procedures, the consent forms, or the investigators, please communicate your requested changes to the Biosci IRB. The new procedure is not to be initiated until the IRB approval has been given.

ASU Knowledge Enterprise Development

Office of Research Integrity and Assurance

Biosci IRB

To: Johnston,Stephen Albert **Date:** 09 / 26 / 2011

From: Biosci IRB **Expiration Date:** 11 / 22 / 2011

Re: **Protocol #** 0912004625: Profiling Human Sera for Unique Antibody Signatures

This letter serves as a IRB notification reminder by the Biosci IRB. It is the primary responsibility of the Principal Investigator to ensure that the re-approval status for lapsed protocols is achieved. All protocols must be re-approved annually by the IRB unless shorter intervals have been specified.

Please note that the level of review given to the continuing review process is the same as that of any new protocol. All requests for re-approval must be reviewed at a convened IRB meeting, except for those protocols that meet the criteria for expedited review.

Please submit the following documents at least three weeks prior to the expiration date to allow for full committee review:

1) A completed Continuing Review Form.

2) Two (2) copies of each consent form(s) used in the study (If data collection is ongoing).

Please note that you can obtain a copy of the Continuing Review Form through our web site: http://researchintegrity.asu.edu/humans.

As of July 1, 2003, all personnel involved in human subjects research must complete the Human Subjects training course. It is the responsibility of the Principal Investigator to make sure all personnel associated with this study have completed the human subjects training course (see the Office of Research Integrity and Assurance website for a link to the NIH training).

It is a violation of Arizona State University policy and federal regulations to continue research activities after the approval period has expired. If the IRB has not reviewed and re-approved this research by its current expiration date, all enrollment, research activities and intervention on previously enrolled subjects must stop. If you believe that the health and welfare of the subjects will be jeopardized if the study treatment is discontinued, you may submit a written request to the IRB to continue treatment activities with currently enrolled subjects.

Your assistance and cooperation in ensuring that the above-mentioned protocol is received for re-approval evaluation at the Office of Research Integrity and Assurance before the lapse date is greatly appreciated.

Sent to Office of Research Integrity 10/18/11

ARIZONA STATE UNIVERSITY
Office of Research Integrity and Assurance
660 S. Mill Avenue Suite 315
Arizona State University
Tempe AZ 85287-6111
(Mail Code 6111)
Phone: 480-965-6788
Fax: (480) 965-7772

CONTINUING REVIEW FORM- IRB

- In accordance with Federal Regulations 45CFR46, the IRB must review nonexempt protocols at least annually, or more frequently if warranted.
- Please type your responses in the boxes provided. Use as much space as necessary (the boxes will expand). Please answer each question – if a question is not applicable, please put N/A in the box.
- Studies that are in the data analysis phase are considered open, researchers must complete this form.

1. Principal Investigator	
Principal Investigator: Stephen Albert Johnston	
ASU department address: Center for Innovations in Medicine, Biodesign Institute B230 MC5901	
E-mail address: Stephen.johnston@asu.edu	
Phone number: 480-727-0792	Fax Number: 480-727-0782
Co-Investigator(s) Name(s) and Contact Information: Phillip Stafford, Phillip.stafford@asu.edu; Kathryn Sykes, Kathryn.sykes@asu.edu; Lucas Restrepo, lucas.restrepo@asu.edu; Muskan Kukreja, muskan.kukreja@asu.edu	

2. Protocol Information
2a) Title of protocol: Profiling Human Sera for Unique Antibody signatures
2b) HS #: 0912004625
2c) If project is funded or funding is being sought, provide list of all sponsors and grant numbers: DTRA HDTRA1-11-1-0010; DoD BCRP W81XWHO710549; Please indicate the grant status for each source of funding: ☒ Active ☐ Pending
2d) ASU account number/project number: FQS0052, FQS0030
2e) Location(s) of research activity: Biodesign Center for Innovations in Medicine B225, B229, B233, B237
2f) IRB approval dates from additional institutions: All samples are provided to us from institutions that are current with their IRB approval. We currently do not have that information but can obtain if needed. **Please note that copies of current IRB approvals from additional institutions are required.*

3. Protocol Status
3a) Active: ☒ Yes ☐ No (If no, submit a close out report: http://researchintegrity.asu.edu/humans/forms
3b) Please indicate remaining duration of the study: 9 years
4. Participant Information
4a) Is this study closed to enrollment of new subjects: ☒ Yes ☐ No

4b) Total number of participants approved for the study (to be enrolled): N/A
4c) Number of participants enrolled (e.g. signed a consent form) during the **past approval period**: N/A
4d) Total number of participants enrolled **since study began**: N/A
4e) Total number of individuals screened (e.g. individuals that responded to study advertisements or other recruitment tractices and were questioned by investigators) in the past approval period (if applicable): N/A (this includes the number that was later enrolled)
4f) Of the total number of individuals screened in the past approval period, what percentage has been ineligible to participate in the study (if applicable)? N/A
4g) Number of enrolled participants who withdrew from the study: N/A Please state the reason(s) the participant(s) withdrew.
4h) Number of participants still to be enrolled: N/A (If this brings the sample to greater than what is listed in 4b, submit a request for modification see 7d).
4i) Participant enrollment breakdown by gender, age and ethnicity: (This information is required for all studies that are NIH-sponsored. It is recommended, but not required, that other researchers provide this information). N/A

5. Data Sources Check all categories that apply to your protocol	
☐	Human subjects intervention with use of informed consent form
☐	Discarded, identified pathological materials, no intervention
☐	Genetic analysis
☐	Interviews or questionnaires
☐	Medical records or other records from human subjects
☒	Other please specify: **We have clinical diagnosis and treatment status, genotype information and maybe smoking history. All information though is provided to us.**

6. Adverse Events or Unexpected Problems

6a) Have there been any complaints from subjects in the past approval period?
☐ Yes If yes, describe ☒ No

6b) Have there been any ***adverse events*** or unexpected problems in the past approval period?
☐ Yes ☒ No
If yes, please explain in detail and indicate when the IRB was notified of the event or problem. If the IRB was not notified, please explain why this was not done.

6c) Does the study have a Data Safety Monitoring Board (DSMB)? ☐ Yes ☒ No
If yes, please indicate the date of the last DSMB review:

Please note that investigators are required to submit DSMB reports to the ASU IRB at the time they are made available to the investigator.

7. Protocol Modifications or Revisions

7a) Have there been any modifications or revisions to the protocol in the past approval period?
☐ Yes ☒ No
If yes, please indicate the date of the approval from the Committee for the modification or revision and provide a brief description.

7b) Have there been any deviations from the approved protocol? ☐ Yes ☒ No
If yes, please describe to self-report the protocol violation.
7c) Do you want to add any new co-investigators to the study? ☐ Yes ☒ No
If yes, submit their names and copies of the human subjects training required by the IRB:
http://researchintegrity.asu.edu/training/humans
7d) Do you wish to submit a modification at this time? ☒ Yes ☐ No
If yes, please describe the modification request and rationale for the changes: Please remove Dr. Patricia Carrigan from the protocol. She has left ASU.

8. Current Consent Form

8a) Please attach a copy of your current consent form for renewal if you are enrolling new subjects. N/A

8b) Is this the original consent form or a revised form? ☐ Original ☐ Revised (If revised, please provide date of ASU IRB approval for the revision. Attach a copy of the stamped form and unstamped form)

9. Protocol Progress Report

9) Please submit a ***detailed*** progress report. The progress report must be substantive and complete, and include the goal(s) of the study, findings to-date, how data is being stored, and plans for the next year/review period. If this project is funded, please send a copy of the most recent progress report that was sent to the funding agency:

The last year, our team has optimized the immunosignature microarray and has contracted with Applied Microarrays (AMI) in Tempe, AZ to print our arrays. We obtained a new set of 10,000 different random peptides, as the last set had been depleted. We ensured that the new peptides were carefully diluted in a new buffer/organic mix that is compatible with AMI's printing process. The added precision of commercial printing has allowed us to obtain higher reproducibility across patients, and find much more subtle changes in antibody responses. We have completed the Valley Fever project by printing a set of 100-peptide 'diagnostic arrays' to do the test-training sample sets. We have obtained 65 look-back blinded samples from John Galgiani at U of A in Tucson that were all false negative samples from his clinic. We classified these samples with 0% error (after excluding problematic samples that were inherently high-background or had been subject to degradation effects). We are in the process of writing these data up in a manuscript.

We completed a project on glioblastoma multiformae, using blinded samples from Barrow Neurological Institute (BNI) in which we were able to identify brain cancer grade as well as presence or absence of an important methylation enzyme, MGMT. This enzyme's status has been shown to be an effective predictor of response to Temozolamide. We have submitted this manuscript to NeuroOncology.

We have completed a project on Esophageal Cancer, using blinded samples obtained from Mayo Clinic, in which we were able to distinguish presence or absence of Esophageal cancer in patients. We are currently examining samples from patients with Barrett's Esophagus, to determine whether we can detect early cancer predisposition.

We have built a pathogen microarray, in which 5K peptides from human pathogens were tiled on a standard glass slide. We are currently optimizing this platform to distinguish patients who are convalescent from one or another infectious agent. We have found that printing methods that enhance the immunosignaturing effect are deleterious to the discrimination of our pathogen epitope arrays. Thus we are altering the printing characteristics for these arrays, and are using a slide surface that spaces the peptides out much further than our aminosilane slides allow.

10. Publications, Presentations and Recent Findings

10a) Have there been any presentations or publications resulting from this study during the past approval

period? ☒ Yes ☐ No If yes, please submit a copy of the abstract, or the publication, with this application.

"Immunosignaturing can detect products from molecular markers in brain cancer" – submitted to NeuroOncology
"Physical parameters Affecting Antibody Profiles as Biomarkers of Health Status" – revision resubmitted to Molecular and Cellular Proteomics
"Sample Preparation for Immunosignaturing" – revision resubmitted to Vaccine

Presentations:
BRP FY11 Vision Setting Meeting, November 2010 – Panel Member
3rd Annual Oncology Biomarker Conference, January 2011 – Invited speaker
Leading Innovation and Knowledge Sharing (LINKS) BCMRP meeting, February, 2011 - Panel member
BCRP FY10 Programmatic Review Meeting, March 2011 – Panel member
Canary Foundation, March 2011 – Invited participant
Era of Hope Abstract Placement Meeting, April 2011 – Committee member
NBCC Artemis Project, April 2011 – Workshop participant
Era of Hope Meeting, Orlando, August 2011 – Invited speaker, Organizing committee member

10b) Have there been any recent findings either from this study, or a related study (through a literature review for example), that would have an effect on this study's risk/benefit analysis? ☐ Yes ☒ No
If yes, please describe and cite references:

11.Conflicts of Interest and Commercialization

11. Does any member of the research team have a potential conflict of interest with this study that could affect study participants and/or study outcome? For more information about examples of conflicts of interests, please visit the ASU objectivity website: http://researchintegrity.asu.edu/coi
☐ Yes (If yes, please describe and disclose in the consent form) ☒ No

b) Does the PI or Co-I have a current conflict disclosure form on file at the ASU Office of Research Integrity and Assurance?
☐ Yes ☐ No

c) If there are conflicts of interests, please describe the ways in which you have and will minimize harm to research subjects and/or the objectivity of research. No prospective human trials have been proposed, only blinded retrospective samples are currently being run on the immunosignaturing platform.

12. Training

12.The research team must verify completion of human subjects training within the last 3 years. (http://researchintegrity.asu.edu/training/humans)

CITI training – Provide the date that the PI and Co-I's completed the training:
If you completed **NIH** training prior to 9/15/10 this will be accepted. Provide a copy of the certificate.

13. Required Signatures

Revised 8/11

Principal Investigator:  Date: 10/18/11

FOR IRB USE
Chair or Committee member name:

Signature: Date:

Revised 8/11