TAMPEREEN TEKNILLINEN YLIOPISTO

MATTI ANNALA

DISCOVERY OF A NOVEL FUSION GENE IN GLIOBLASTOMA USING COMPUTATIONAL METHODS

Master of Science thesis

Examiner: Olli Yli-Harja, Prof
Subject approved by the department council 12.01.2011

TIIVISTELMÄ

TAMPEREEN TEKNILLINEN YLIOPISTO
Tietotekniikan koulutusohjelma
ANNALA, MATTI: Uuden fuusiogeenin löytö glioblastoomasta laskennallisin menetelmin
Diplomityö, 59 sivua, 0 liitesivua
Huhtikuu 2013
Pääaine: Signaalinkäsittely
Tarkastaja: Professori Olli Yli-Harja
Avainsanat: syöpägenomiikka, fuusiogeeni, aivosyöpä, laskennallinen biologia

Syöpä on tauti jonka määrittävä piirre on solujen hallitsematon ja invasiivinen kasvu. Syövät saavat alkunsa geneettistä muutoksista jotka muuttavat solun toimintaa ja johtavat haitalliseen fenotyyppiin joka periytyy syöpäsolun jakautuessa. Fuusiogeenit ovat yksi geneettisten muutosten muoto jossa kahden geenin palaset liittyvät yhteen ja muodostavat uudella tavalla käyttäytyvän geenin. Fuusiogeenien on osoitettu olevan tärkeässä roolissa monissa ihmisten syövissä. Tässä työssä käytimme laskennallisia menetelmiä ja koko transkriptomin kattavaa sekvensointia etsiäksemme fuusiogeenejä 40 aivosyöpäpotilaan aineistosta. Löysimme uuden *FGFR3-TACC3* fuusiogeenin, joka määrittää uuden glioblastooman alityypin. Glioblastooma on äärimmäisen tappava ja yleinen aivosyövän muoto ihmisissä. Tutkimalla isompaa potilasaineistoa löysimme 4 / 48 fuusiogeenille positiivista glioblastoomaa, mutta emme yhtään positiivista tapausta 43 matala-asteisen aivosyövän joukosta. Löytämämme fuusiogeeni johtuu tandem-kopioituneesta alueesta kromosomissa 4, ja tuottaa kimeeristä proteiinia joka muuttaa aivosyövän pahalaatuisemmaksi ja voimistaa solukasvua. *FGFR3-TACC3* fuusiogeeni ei koskaan esiintynyt yhdessä *EGFR, PDGFRA* tai *MET* geenien amplifikaation kanssa. On mahdollista, että fuusiogeeniä kantavia potilaita voidaan tulevaisuudessa hoitaa käyttäen olemassaolevia FGFR3 proteiinin toimintaa estäviä lääkkeitä.

ABSTRACT

TAMPERE UNIVERSITY OF TECHNOLOGY
Master's Degree Programme in Information Technology
ANNALA, MATTI: Discovery of a novel fusion gene in glioblastoma using computational methods
Master of Science Thesis, 59 pages, 0 appendix pages
April 2013
Major: Signal Processing
Examiner: Prof Olli Yli-Harja
Keywords: cancer genomics, fusion gene, brain cancer, computational biology

Cancer is a disease characterized by the uncontrolled and invasive growth of cells. All forms of cancer are caused by genomic alterations that alter normal cellular function, leading to a malignant phenotype that is inherited across cell division. Fusion genes are a type of genomic alteration where pieces from two genes are fused together, forming a new gene with altered behaviour. Fusion genes are known to play a role in many human cancers. In this work, we used computational analysis and whole transcriptome sequencing to search for fusion genes in a cohort of 40 brain cancer patients. We discovered a novel fusion gene *FGFR3-TACC3* that characterizes a new subtype of glioblastoma, a highly lethal form of brain cancer. In a larger validation cohort, the fusion gene was found in 4 of 48 glioblastoma patients but not in any of 43 low-grade gliomas tested. The fusion gene is caused by tandem duplication and encodes a chimeric protein that promotes glioma progression and cell growth. The fusion gene was mutually exclusive with the amplification of *EGFR*, *PDGFRA* and *MET*, three oncogenes associated with glioblastoma. The availability of small molecule inhibitors for FGFR3 suggests an effective treatment strategy for glioblastoma patients harboring the fusion.

PREFACE

This work was supported by the Academy of Finland (projects 122973, 132877, 213462), the Tekes Finland Distinguished Professor programme, and the Department of Signal Processing.

The research described in this thesis has been published in the Journal of Clinical Investigation (Parker & Annala et al. 2012). Some of the material in section 2 has been published as a review paper in Cancer Letters (Annala et al. 2012). Independently of our work, the *FGFR3-TACC3* fusion gene was also reported by Singh et al. in *Science* (2012).

I would like to thank the people at the MD Anderson Cancer Center for their effort in producing the data and performing the wet-lab work to validate our hypotheses. Transcriptome sequencing using the SOLiD 3 platform was done by Chang-gong Liu. Sequencing quality control was done by Han Liang. Brittany Parker, David Cogdell, Kirsi Granberg, Yan Sun, Ping Ji, and Xia Li performed wet-lab experiments.

In particular, I would like to thank my colleague and co-author Brittany Parker for her hard and dedicated work in the lab. I would also like to thank professor Wei Zhang from the Cancer Genomics program at MD Anderson, and my thesis advisor professor Matti Nykter for their excellent support and coordination in completing this project.

Tampere, April 19, 2013
Matti Annala

TABLE OF CONTENTS

TERMS AND ABBREVIATIONS

aCGH	Array comparative genomic hybridization. A method where DNA microarrays are used to assess the copy numbers of genomic regions.
BLAST	Web-based sequence alignment tool that can locate a given sequence in an organism's genome or transcriptome.
cDNA	Complementary DNA. DNA produced by reverse transcription of RNA back into DNA.
CDS	Coding sequence of a messenger RNA. The RNA segment that is translated into protein by ribosomes.
Codon	A nucleotide triplet that codes for an amino acid or the start/end of a coding sequence.
Copy number	The number of copies of a gene or genomic region found within a cell.
Cytoplasm	The contents of a cell, excluding the nucleus. The cytoplasm is enclosed by the plasma membrane and includes most of the organelles found in cells.
DNA	Deoxyribonucleic acid. The nucleic acid that acts as a blueprint for the behavior of all living cells.
DNA microarray	A device used to quantify the amounts of thousands of different short DNA sequences within a cell.
ENCODE	Encyclopedia of DNA Elements. A public research consortium that is mapping all functional elements in the human genome.
Exon	A segment of pre-messenger RNA that remains in the processed messenger RNA transcript. See also *intron*.
Eukaryote	A branch of life characterized by cells that contain nuclei.
FDA	Food and Drug Administration. A US agency that promotes public health by supervising food and drug safety.
Frameshift	A change in the reading frame of a protein's coding sequence.
GBM	Glioblastoma. The most common and aggressive type of primary brain cancer in humans.
Glioma	A brain cancer that arises from glial cells.
HTS	High throughput sequencing. A term that describes a number of new DNA sequencing technologies capable of producing millions of short sequence reads per day.
Indel	Insertion or deletion of one or more nucleotides into a DNA or RNA segment.
Intron	A segment of pre-messenger RNA that is spliced out of the transcript to produce the final messenger RNA.

MDACC	The University of Texas MD Anderson Cancer Center. One of the world's leading cancer hospitals and research centers. Located in Houston, Texas.
miRBase	A curated public repository of microRNA annotations for a number of different organisms.
miRNA	MicroRNA. A form of small noncoding RNA that regulates gene expression through RNA interference and other mechanisms.
mRNA	Messenger RNA. Processed RNA transcripts that exit the nucleus and are translated by ribosomes into proteins.
NCBI	National Center for Biotechnology Information.
Nucleus	A membrane-enclosed cellular compartment that contains all of the DNA found in eukaryotic cells (except for mitochondrial DNA).
PCR	Polymerase chain reaction, a wet-lab technique for copying segments of DNA.
Primary cancer	A mass of cancer cells that is situated at the site of origin. Contrast with metastasized cells that have migrated to a new site through the bloodstream or otherwise.
Reading frame	The set of codon locations found in the coding region of a gene.
RefSeq	A curated public repository of RNA and DNA sequence data from multiple biological organisms.
Reverse transcription	Biological process where a complementary DNA strand is produced using an RNA strand as template. The process is performed by reverse transcriptase enzymes.
RNA	Ribonucleic acid.
RNA-seq	RNA sequencing. A technique where high throughput sequencing is used for transcriptomic profiling.
RT-PCR	Polymerase chain reaction preceded by a reverse transcription step where RNA is reverse transcribed into cDNA.
SNP	Single nucleotide polymorphism.
TCGA	The Cancer Genome Atlas. A large-scale collaborative research project that is cataloguing cancer-causative genomic alterations in over 20 different cancer types.
Transcript	A strand of RNA produced by an RNA polymerase enzyme that copies a strand of DNA into RNA.

1 INTRODUCTION

The field of computational biology has advanced rapidly during the last 20 years. Technologies such as DNA microarrays and high-throughput sequencing have provided researchers with an unprecedented amount of biological information (Hawkins et al. 2010), while modern computers have made it possible to analyze the data within practical timescales. We are reaching a stage where organisms can be studied and understood in a holistic manner at all levels of their dynamics. This new field of study has come to be known as systems biology. In practical terms, systems biology studies the complex networks of molecular interactions that govern the functioning of biological organisms (Kitano 2002). As such, it provides a powerful platform for the study of complex and heterogeneous diseases such as cancer. However, in order to fully realize the promise of systems biology, we must first understand the parts that make up the system under study.

This vision led the computational systems biology group at the Tampere University of Technology to initiate a project with the goal of using high throughput sequencing and computational analysis to discover novel features and regulatory mechanisms in human cancers. The project was initiated in cooperation with Prof. Wei Zhang, director of the Cancer Genomics Core Laboratory at the University of Texas M.D. Anderson Cancer Center. The first cancer type chosen for study was brain cancer, with particular emphasis on glioblastoma multiforme, the most common and lethal form of brain cancer in humans (Furnari et al. 2007). Prof. Zhang's group had years of experience in the study of this cancer, and in 2010 they decided to use the newly introduced technique of whole transcriptome sequencing to characterize the RNA content of a large number of brain tumors. Our group was tasked with analyzing the sequencing data and generating biological hypotheses for subsequent functional validation. Particular emphasis was placed on the discovery of novel chromosomal alterations or mutations that drive the malignant behavior of brain cancer.

To fulfill the technical requirements of this project, we implemented a software aimed at identifying fusion genes from whole transcriptome sequencing data. A fusion gene is a chimeric gene that combines pieces from two original genes. They are formed when chromosomes break into pieces and cellular repair mechanisms fail to reassemble the fragments correctly. By combining the growth-inducing potential of one gene with the activating potential of another, fusion genes can single-handedly transform a benign, normal cell into an uncontrollably proliferating cancer cell. Indeed, fusion genes have

been shown to act as drivers of malignant transformation in dozens of human cancers (reviewed in Mitelman et al. 2007). In *BCR-ABL1* fusions found in 95% of chronic myelogenous leukemias (CML), the inclusion of protein domains from BCR renders the growth-inducing ABL1 protein constitutively active (Davis et al. 1985), resulting in cancer even in the absence of other genetic lesions (Daley et al. 1990). After the discovery of *BCR-ABL1* in 1985 (Shtivelman et al. 1985), a drug targeting this fusion protein was successfully tested in 1996 (Druker et al. 1996). This drug, imatinib, received FDA approval in 2001 and single-handedly transformed CML from an invariably lethal cancer into a chronic, manageable condition for 95% of patients (Druker et al. 2006). This example illustrates the clinical impact that targeted molecular therapies can have on cancer treatment. Unfortunately in many cancers the driving mechanisms are still poorly understood, and no suitable molecular targets are available.

In this thesis we discuss the implementation of a software for fusion gene discovery, and then demonstrate how the software was used to identify a novel fusion gene in glioblastoma, the most lethal and common form of primary brain cancer in humans. We also describe the computational analyses and wet-lab experiments that were performed to understand the function, origin, and clinical significance of the fusion gene. In other words, we describe the entire process that goes into the discovery and functional validation of a novel fusion gene.

We start in chapter 2 by providing the reader with the biological background necessary for understanding the biological quantities and entities that make up the subject matter of this thesis. In particular, we give an overview of the study of cancer from the point of view of molecular biology, and discuss the current state of knowledge on brain cancer.

In chapter 3 we describe the experimental methods and computational algorithms used in this thesis. We provide a basic overview of DNA microarrays and high throughput sequencing, and describe the algorithm we used for identifying fusion genes from whole transcriptome sequencing data. We also discuss the other algorithms that were used to translate raw microarray or sequencing measurements into meaningful and quantitative biological phenotypes.

After describing the computational methods, we illustrate their use in chapter 4 through a case study. In the case study we show how the algorithms were used to discover a novel fusion gene in human brain cancer. We also describe how we validated the fusion gene by combining wet-lab experiments with microarray and sequencing data. Finally, we demonstrate how we applied our algorithm to other public datasets and found more patients positive for the fusion gene.

In chapter 5 we conclude the thesis and discuss anticipated future developments in basic research and clinical applications relating to the novel fusion gene.

2 BIOLOGICAL BACKGROUND

2.1 Genes, chromosomes and cellular function

All organisms on our planet are composed of cells, the basic building blocks of life. Cells come in a variety of shapes, sizes, and functions. Simple organisms such as bacteria are unicellular, while more complex organisms such as humans are composed of hundreds of cell types acting in concert to produce our diverse behavior. Cells replicate through cell division. Multicellular organisms begin their life as a single cell, the zygote, which undergoes multiple generations of cell division and produces the billions of cells that make up an organism.

All cells carry within them a set of blueprints that define their function and behavior. This blueprint is encoded in the form of a DNA double helix stored inside the cell. The double helix contains two linear strands of *nucleotides*: the four building blocks of DNA (represented by the letters ACGT). The DNA strands are connected to one another so that the nucleotides form complementary pairs (A-T or C-G). The totality of all DNA within an organism is known as its *genome*. The genome of an organism is subdivided into physically disjoint subunits known as *chromosomes*. Chromosomes are highly condensed structures composed of a long string of DNA wrapped around scaffold proteins. The human genome consists of 46 chromosomes, 23 from each parent, plus the small quantity of DNA found within mitochondria. In eukaryotic cells such as human cells, DNA is tucked away safely in the *nucleus*, a membrane-enclosed compartment inside the cell.

Chromosomes can be subdivided into functional units known as *genes*. Genes are contiguous genomic regions that are transcribed into RNA transcripts in a process known as *transcription*. RNA transcripts are nucleotide chains similar to DNA, with the exception that they are single-stranded and use the nucleotide U instead of T. Another difference is that the deoxyribose sugar found in DNA is replaced with a ribose, rendering RNA molecules shorter-lived than DNA. While DNA never leaves the nucleus, RNA transcripts known as *messenger RNA*s (mRNA) are allowed to pass outside the nucleus into the *cytoplasm*. There they are processed by *ribosomes*, complex molecular machines that *translate* the RNA transcripts into *proteins.* A protein is a chain of *amino acids*, each of which is represented by a nucleotide triplet (a *codon*) in the RNA transcript. The $4^3 = 64$ possible codons are redundant and code for only 20 different amino acids. The ribosome does not translate the entire mRNA transcript, but instead starts translating when it encounters a specific three-nucleotide sequence known as a *start codon*. Trans-

lation stops when the ribosome encounters a nucleotide triplet known as a *stop codon*. The region between (and including) the start and stop codons is known as the *coding region* of a transcript, and the positions of all codons are known as the *frame*. A translated protein folds into a thermodynamically stable conformation and then begins executing its evolved function in the cell. In total, the human genome contains over 20,000 such protein coding genes (ENCODE Project Consortium 2012). Many proteins can combine with other proteins to produce intricate molecular complexes that perform highly sophisticated functions. A simplified view of the information flow from DNA to RNA to protein is shown in Figure 1.

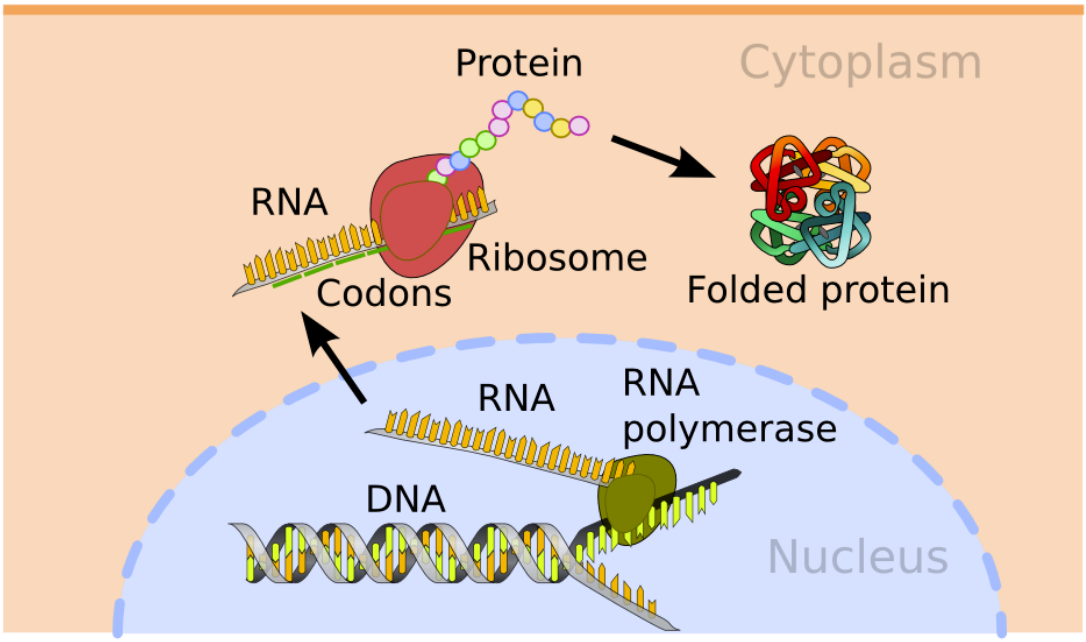

Figure 1. An overview of the canonical mechanism by which information flows from DNA to RNA to protein in eukaryotic cells.

In eukaryotic cells such as human cells, the information flow from DNA to RNA to protein is complicated by a process known as *RNA splicing*. Genes subject to splicing are first transcribed into long transcripts called *pre-messenger RNAs* (pre-mRNA), and these transcripts then undergo splicing, a process where fragments (*introns*) from the middle of the pre-mRNA are cut out, and the remaining fragments (*exons*) are joined back together (Figure 2) (reviewed in Clancy 2008). The pre-mRNAs of some genes can be spliced in multiple alternative ways, leading to different protein structures. Such alternative mature transcripts are known as *splice variants*, and their relative abundance in cells varies in a tissue-specific manner.

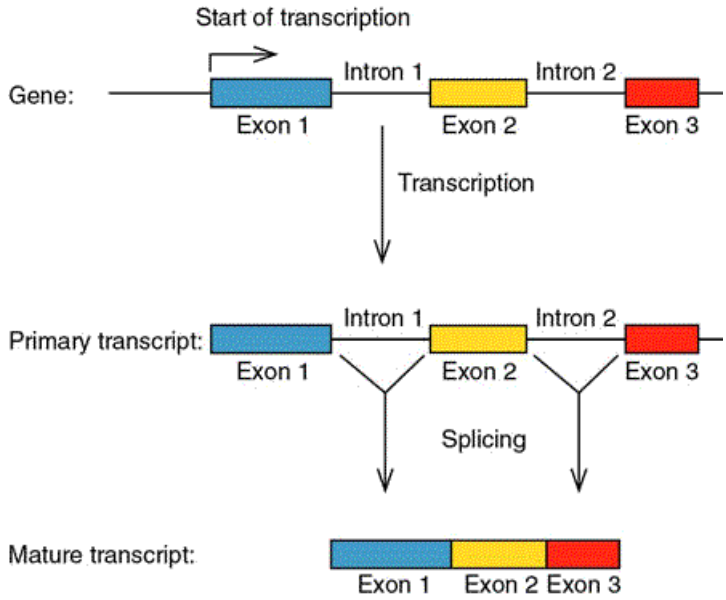

Figure 2. An illustration of mRNA splicing. The gene is first transcribed into the pre-mRNA (primary transcript). The introns are removed and the exons are joined back together to form the mature transcript. Image courtesy of John S. Choinski, University of Central Arkansas.

Proteins are the primary molecules responsible for the majority of functions that take place inside living cells. Yet they are not the only molecules capable of complex function. RNA molecules directly participate in many cellular processes beyond their role as carriers of genetic information between the DNA and ribosomes. Ribosomes themselves are molecular machines composed of equal amounts RNA and protein (Cech 2000). RNA molecules can also form regulatory networks where RNA transcripts target other RNAs for degradation. A classic example is provided by *microRNAs*, short RNA fragments that bind to mRNA transcripts that carry a complementary sequence, and target them for degradation by a protein complex known as the *RNA-induced silencing complex* (reviewed in Sun et al. 2010).

The proteins found within cells vary by cell type. The quantity of a protein inside a cell is determined by multiple factors, including the quantity of mRNA available for translation, the degradation time of the protein, and translation efficiency. Degradation time is affected by a protein's inherent stability and its interactions with other proteins. Translation efficiency is affected by the transcript sequence and the regulatory effects of microRNAs and other molecules. The quantity of mRNA produced by a gene varies widely between genes and cell types. Some genes are only expressed in specific tissue types, while some genes are expressed in all cells. The expression level of a gene is determined by proteins known as *transcription factors*. These proteins enter the nucleus and bind to chromosomal sites harboring a specific DNA sequence. Upon binding, the proteins alter the conformation of the surrounding DNA and cause nearby genes to express at a higher or lower rate.

2.2 Molecular pathology of cancer

Cancers are a heterogeneous class of diseases characterized by the abnormal proliferation of cells. They are the leading cause of death in the developed world, and have proven notoriously resistant against attempts at finding a cure (Jemal et al. 2011). This is largely due to two characteristic features of cancers: resilience and heterogeneity. Cancer cells are *resilient* in that they robustly adapt to external challenges such as drug treatments or changes in their microenvironment. They are *heterogeneous* in the sense that cancers of different tissue or cell type are often driven by different molecular mechanisms. Even histologically identical cancers of the same tissue can be driven by different abnormalities of the cellular machinery, although common themes have been identified (Salk et al. 2010; Visvader 2011). The heterogeneity of cancer makes it difficult to find treatments that are effective for a significant number of patients, while resilience means that even if a treatment is initially effective against a tumor, the tumor will eventually acquire resistance to it.

The currently accepted view is that cancers initially originate from a single cell that acquires a phenotype of uncontrollable proliferation as a result of sporadic genetic changes (Visvader 2011). These genetic changes can range from single nucleotide mutations to large rearrangements that drastically alter the structure of chromosomes. The 46 chromosomes found in the nucleus of every (somatic) human cell constantly acquire cumulative genetic damage, which is why biological organisms have developed repair and backup mechanisms against its effects (Helleday et al. 2008). These backup mechanisms explain why cancers rarely arise due to a single genetic alteration: if one gene starts acting pathologically, compensatory mechanisms will soften the impact. However, if one cell acquires the perfect storm of genetic lesions that causes malignant proliferation, the phenotype will propagate to its offspring across cell divisions.

The genomic alterations that have been implicated in the formation of cancers can be divided into four groups: mutations, copy number alterations, fusion genes, and epigenetic modifications. *Mutations* are changes involving a single nucleotide or few nucleotides in a chromosome. The most common type of mutation is the point mutation, a substitution of one nucleotide with another. Insertion/deletion (indel) mutations are mutations where one or more nucleotides are added to or removed from a genomic locus. *Copy number alterations* are genetic lesions where a large segment of a chromosome is deleted or duplicated. Copy number alterations can also involve entire chromosomes, a phenomenon known as *aneuploidy*. *Epigenetic modifications* are alterations involving nucleosomes, chromatin structure, and DNA methylation. *Fusion genes* are discussed in the next section.

2.3 Fusion genes

2.3.1 History

Fusion genes are hybrid genes that combine parts of two or more original genes. They can form as a result of chromosomal rearrangement or abnormal transcription, and have been shown to act as drivers of malignant transformation and progression in many human cancers (reviewed in Mitelman et al. 2007). The first signs of fusion genes in human cancer were identified in 1960 when a reciprocal translocation between the q-arms of chromosomes 9 and 22 was discovered in over 90% of chronic myelogenous leukemia patients (Nowell et al. 1960; Rowley et al. 1973). After two decades the translocation was understood to produce a chimeric *BCR-ABL1* transcript that encodes a constitutively active form of the ABL kinase (Shtivelman et al. 1985). At the same time, Burkitt's lymphoma was found to harbor activating fusions between immunoglobulin genes and *MYC* (Manolov et al. 1972; Zech et al. 1976; Dalla-Favera et al. 1982). These initial findings led to the prompt discovery of many more fusion genes in hematological malignancies and solid cancers (Table 1).

Among hematological malignancies, the identification of *PML-RARA* fusions in acute promyelocytic leukemia paved the way for an effective tretinoin-based molecular therapy (Borrow et al. 1990; Warrell et al. 1991), while a *RUNX1-ETO* chimeric protein was found to characterize a morphologically distinct subtype of acute myeloid leukemia with prolonged median survival (Erickson et al. 1992). Early examples of fusion genes in solid cancers included the discovery of fusions between *EWSR1* and members of the *ETS* transcription factor family in Ewing's sarcoma (Turc-Carel et al. 1983; Aurias et al. 1983), and characteristic *SS18-SSX* fusions in synovial sarcoma (Turc-Carel et al. 1987; Smith et al. 1987; Clark et al. 1994). In myxoid liposarcoma, *FUS-DDIT3* and *EWSR1-DDIT3* fusions were found to be pathognomonic for the disease (Crozat et al. 1993; Rabbitts et al. 1993; Antonescu et al. 2001). A breakthrough happened in 2005 when fusion genes juxtaposing the gene *TMPRSS2* and members of the *ETS* transcription factor family were found in 70% of prostate cancers (Tomlins et al. 2005). Subsequent discoveries in solid cancers included the discovery of *EML4-ALK* fusions and *CHD7* rearrangements in non-small cell lung cancer (Soda et al. 2007; Rikova et al. 2007; Pleasance et al. 2010), *KIAA1549-BRAF* fusions in pediatric glioma (Jones et al. 2008), and R-spondin fusions in colon cancer (Seshagiri et al. 2012).

Some cancers were found to associate with multiple fusion genes that presented in a mutually exclusive manner. For instance, the fusions *TMPRSS2-ERG* and *TMPRSS2-ETV1* are common findings in prostate cancer, but almost never co-occur in a single

tumor (Tomlins et al. 2005). In some cases, fusion genes also exhibit mutual exclusivity or co-occurrence with other types of genomic aberrations, as exemplified by the mutual exclusivity of *ETS* fusions and *SPINK1* overexpression in prostate cancer (Tomlins et al. 2008).

Table 1. Fusion genes in human cancers.

Cancer	Fusion gene	Frequency	Mechanism of formation	Biological impact	References
Acute lymphocytic leukemia	ETV6-RUNX1	25%	Interchromosomal	Oncogenic chimeric	Golub et al. (1995), Romana et al.
Acute myeloid leukemia	RUNX1-ETO	10-15%	Interchromosomal	Oncogenic chimeric	Erickson et al. (1992)
	CBFB-MYH11	10-15%	Inversion	Oncogenic chimeric	Liu et al. (1993)
Acute promyelocytic leukemia	PML-RARA	95%	Interchromosomal	Oncogenic chimeric	Borrow et al. (1990), Warrell et al.
	PLZF-RARA	0-5%	Interchromosomal	Oncogenic chimeric	Chen et al. (1993)
Anaplastic large cell lymphoma	NPM1-ALK	75%	Interchromosomal	Oncogenic chimeric	Morris et al. (1994), Shiota et al. (1994)
	TPM3-ALK	15%	Interchromosomal	Oncogenic chimeric	Lamant et al. (1999)
Burkitt's lymphoma	IG@-MYC	90-100%	Interchromosomal	Promoter exchange	Manolov et al. (1972), Dalla-Favera et
Chronic myelogenous leukemia	BCR-ABL1	95%	Interchromosomal	Oncogenic chimeric	Nowell et al. (1960), Shtivelman et al.
Inflammatory myofibroblastic tumor	TPM3-ALK	50%	Interchromosomal	Oncogenic chimeric	Lawrence et al. (2000)
Adenoid cystic carcinoma	MYB-NFIB	90-100%	Interchromosomal	Loss of microRNA	Persson et al. (2009)
Bladder cancer	FGFR3-TACC3	0-10%	Tandem duplication	Oncogenic chimeric	Williams et al. (2012)
Clear cell sarcoma	EWSR1-ATF1	90-100%	Interchromosomal	Oncogenic chimeric	Bridge et al. (1990), Zucman et al.
Colon cancer	PTPRK-RSPO3	5-10%	Inversion	Promoter exchange	Seshagiri et al. (2012)
	EIF3E3-RSPO2	0-5%	Deletion	Promoter exchange	Seshagiri et al. (2012)
Congenital fibrosarcoma	ETV6-NTRK3	90-100%	Interchromosomal	Oncogenic chimeric	Knezevich et al. (1998)
Ewing sarcoma	EWSR1-FLI1	90%	Interchromosomal	Oncogenic chimeric	Turc-Carel et al. (1983), Aurias et al.
Follicular thyroid carcinoma	PAX8-PPARG	60%	Interchromosomal	Oncogenic chimeric	Kroll et al. (2000)
Glioblastoma	FGFR3-TACC3	0-5%	Tandem duplication	Oncogenic chimeric	Singh et al. (2012), Parker et al. (2012)
Mucoepidermoid carcinoma	MECT1-MAML2	60%	Interchromosomal	Oncogenic chimeric	Tonon et al. (2003)
Myxoid liposarcoma	FUS-DDIT3	90-100%	Interchromosomal	Oncogenic chimeric	Crozat et al. (1993), Rabbits et al.
	EWSR1-DDIT3	0-5%	Interchromosomal	Oncogenic chimeric	Panagopoulos et al. (1996)
Non-small cell lung cancer	EML4-ALK	0-10%	Inversion	Oncogenic chimeric	Soda et al. (2007), Rikova et al. (2007)
NUT midline carcinoma	BRD4-NUT	90-100%	Interchromosomal	Promoter exchange	French et al. (2003)
Papillary thyroid carcinoma	CCDC6-RET	15%	Inversion	Oncogenic chimeric	Grieco et al. (1990)
	NCOA4-RET	15%	Complex rearrangement	Oncogenic chimeric	Santoro et al. (1994)
Pediatric renal cell carcinoma	PRCC-TFE3	20-40%	Interchromosomal	Oncogenic chimeric	Weterman et al. (1996)
Pilocytic astrocytoma	KIAA1549-BRAF	70%	Tandem duplication	Oncogenic chimeric	Jones et al. (2008)
Prostate cancer	TMPRSS2-ERG	60%	Deletion	Promoter exchange	Tomlins et al. (2005)
	TMPRSS2-ETV1	0-5%	Interchromosomal	Promoter exchange	Tomlins et al. (2005)
	TMPRSS2-ETV4	0-5%	Interchromosomal	Promoter exchange	Tomlins et al. (2006)
Secretory breast carcinoma	ETV6-NTRK3	90%	Interchromosomal	Oncogenic chimeric	Tognon et al. (2002)
Serous ovarian cancer	ESRRA-C11orf20	15%	Intrachromosomal	Oncogenic chimeric	Salzman et al. (2011)
Synovial sarcoma	SS18-SSX1	70%	Interchromosomal	Oncogenic chimeric	Turc-Carel et al. (1987), Clark et al.
	SS18-SSX2	30%	Interchromosomal	Oncogenic chimeric	Crew et al. (1995)
	SS18-SSX4	0-5%	Interchromosomal	Oncogenic chimeric	Skytting et al. (1999)

2.3.2 Clinical significance

Traditional cytotoxic drugs used in cancer chemotherapy usually target cells that divide quickly or are DNA repair deficient (both are common hallmarks of cancer). These kinds of therapies have the problem that their molecular targets are not fully specific to cancer cells, often causing the drugs to have strong side effects. Because fusion genes are only found in cancer cells, they provide an excellent target for molecular therapeutics. Indeed, many known fusion genes are already used as FDA approved drug targets. Examples include the treatment of *BCR-ABL1* positive leukemia patients with the ABL kinase inhibitor imatinib (Druker et al. 1996), and the treatment of *EML4-ALK* positive non-small cell lung cancer patients with ALK inhibitor crizotinib (Shaw et al. 2011). However, it must be noted that even the latest drugs have not reached full specificity to fusion proteins, and can have some off-target effects on healthy cells.

Fusion genes have also been employed as diagnostic and prognostic markers. For example, detection of *BCR-ABL1* transcripts is used to confirm chronic myelogenous leukemia diagnoses, and transcript levels are followed throughout treatment to monitor for loss of therapeutic response (Hughes et al. 2006).

2.3.3 Biological impact

Fusion genes can affect cell function through a number of mechanisms. One common mechanism is the overexpression of an oncogene through promoter exchange. For example, the overexpression of *ETS* transcription factors in prostate cancer is caused by their fusion with the androgen regulated *TMPRSS2* promoter (Tomlins et al. 2005). Similarly, B cell lymphomas are characterized by fusion genes where the promoter of an immunoglobulin heavy locus is fused with an oncogene (Croce, 1986). A fusion event can also change the expression level of an oncogene by replacing its 3'-UTR, leading to altered regulation when microRNA binding sites in the 3'-UTR are lost (Persson et al. 2009).

Another mechanism by which fusion genes alter cellular function is through the formation of chimeric proteins. Altered protein structure may render a chimeric protein constitutively active, lead it to activate alternative downstream targets, or sabotage a critical cellular function. For example, *ALK* fusion genes in anaplastic large cell lymphoma involve 5' partner genes that harbor dimerization domains that promote ALK dimerization and autophosphorylation, rendering ALK constitutively active (Chiarle et al. 2008). Another example is provided by the constitutively active BCR-ABL1 kinase in leukemia (Davis et al. 1985).

Not all fusion genes necessarily have biological impact. Cancer genomes are often heavily rearranged and contain pairs of genes that have fused together at random. Therefore, any discovery of a novel fusion gene always requires functional validation to ensure that the fusion actually has biological impact.

2.3.4 Mechanisms of fusion gene formation

The formation of fusion genes in cells can occur through multiple mechanisms. In the most common scenario, a fusion gene is formed via somatic chromosomal rearrangement. The four basic types of chromosomal rearrangement are deletions, translocations, tandem duplications, and inversions (Figure 3).

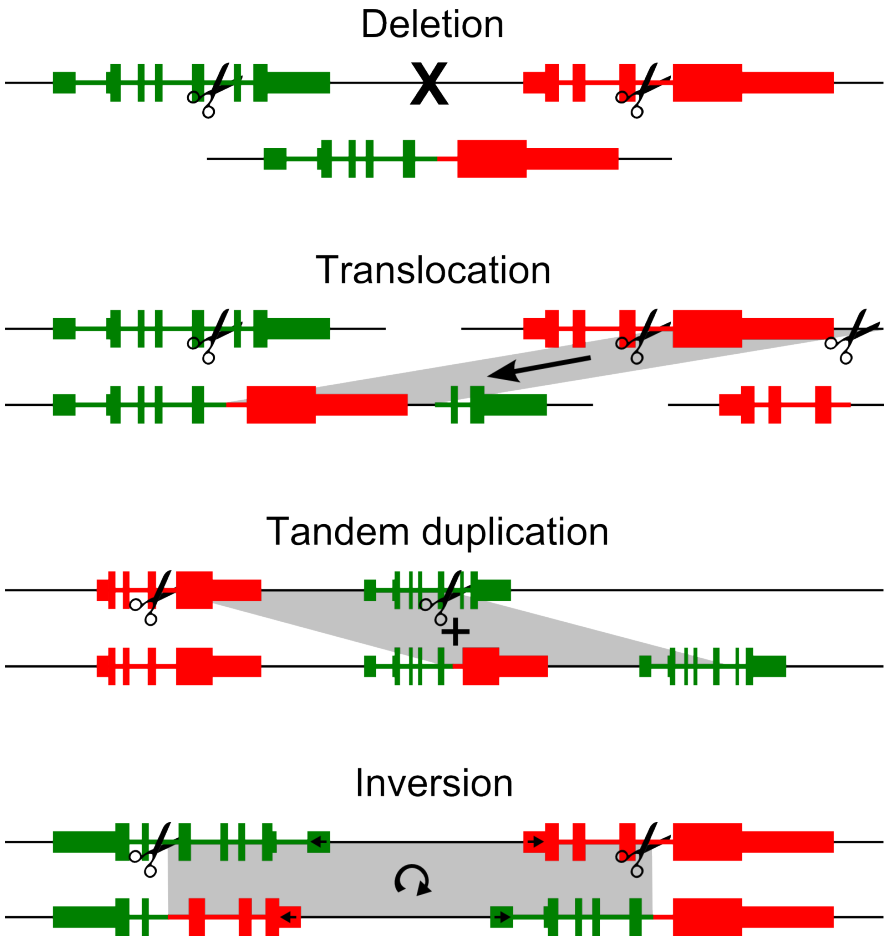

Figure 3. Examples of the different classes of chromosomal rearrangements that can lead to the formation of a fusion gene. The horizontal lines represent chromosomal regions, and the boxes represent gene exons (two genes, red and green). In each scenario, the upper line shows the situation before the rearrangement, and the lower line after the rearrangement.

A fusion gene can arise via deletion when a genomic region between two genes located on the same strand is deleted (Figure 3). The *TMPRSS2-ERG* fusion in prostate cancer is an example of a fusion that results from a 2.7 Mb deletion on chromosome 21 (Perner et al. 2006). Interestingly, fusion genes can also arise from tandem duplication, a type of chromosomal rearrangement where a genomic region is duplicated one or more times, and the copies are tiled next to the original region. When the amplicon breakpoints are situated near existing genes, this can result in the formation of a fusion gene at the junction of the copied and original region (Figure 3). Examples of fusion genes formed through tandem duplication include *KIAA1549-BRAF* fusions in pilocytic astrocytoma

(Jones et al. 2008), and *C2orf44-ALK* fusions in colorectal cancer (Lipson et al. 2012). A tandem duplication or deletion is likely the cause when two genes located on the same chromosomal strand are fused. The order of the two genes in the fusion transcript is also a helpful clue, as tandem duplication creates chimeric transcripts where the genes are in reverse order relative to their positions on the strand.

Occasionally fusion genes arise via inversion events where chromosomal segments are flipped around (Figure 3). For example, the *EML4-ALK* fusion gene in non-small cell lung cancer results from a 12 Mb inversion on chromosome 2 (Soda et al. 2007). If a fusion gene involves two genes located on opposite strands of a chromosome, there is suitable cause to suspect an inversion event. The genes can face inward or outward; an inversion in either scenario can lead to a fusion gene. A characteristic feature of this class of fusion is the formation of reciprocal fusion genes at both ends of the inversion (Ciampi et al. 2005; Soda et al. 2007). However, depending on the properties of the promoters involved, one or both reciprocal fusions may not be transcribed, rendering them impossible to detect through transcriptome sequencing.

In addition to chromosomal rearrangements involving genes on the same chromosome, many fusion genes involve genes located on separate chromosomes. Such fusions are always caused by a translocation of some kind, whether it involves the translocation of a small genomic fragment to a new locus, or a reciprocal translocation involving the swapping of entire chromosome arms (Figure 3). Examples of fusion genes caused by translocations include the *BCR-ABL1* fusion, formed by a reciprocal translocation between 9q and 22q (Shtivelman et al. 1985) More complex rearrangements are also possible but less frequent (Lawson et al. 2011).

2.3.5 Distribution of genomic breakpoints

The genomic breakpoints of fusion genes usually occur in intronic or intergenic regions, and rarely disrupt coding sequences. This phenomenon is partly explained by introns being 35 times longer than exons on average (Zhu et al. 2009). Oncogenic selection may also play a role, as fusions that disrupt an exon have a two-in-three chance of creating a frameshifted protein with little effect on cellular function. Conversely, intronic breakpoints often lead to in-frame chimeric proteins because exons tend to terminate at codon boundaries (Long et al. 1999; Sverdlov et al. 2003; Ruvinsky et al. 2005). Despite the bias for intronic breakpoints, isolated cases of exon disrupting breakpoints have been reported in the literature (Martinelli et al. 2002; Tort et al. 2004).

A characteristic feature of many fusion-generating chromosomal rearrangements is the presence of sequence microhomology at rearrangement breakpoints. A study of 40 *RAF* gene fusions in low-grade glioma found that 85% harbored microhomology at or near the breakpoints (Lawson et al. 2011). The microhomologies ranged in length between 1-

6 bp and were significantly more common than expected by chance. This pattern is characteristic of microhomology-mediated break-induced replication (MMBIR), implying that MMBIR may be a major causative mechanism behind many fusion events (Lawson et al. 2011). Another study that looked at *TMPRSS2-ETS* breakpoints in prostate cancer also found evidence of microhomology, but implicated non-homologous end joining (NHEJ) as the driving mechanism behind the chromosomal rearrangements (Lin et al. 2009).

2.3.6 Read-through and splicing

A particular class of fusion genes known as *read-through* chimeras can arise in the absence of any DNA level alterations. This type of fusion gene forms when an RNA polymerase does not properly terminate transcription at the end of a gene, but instead continues transcribing until the end of the next gene (Figure 4). The chimeric pre-mRNA is spliced to produce a fusion transcript. In almost all cases, the resulting chimeric mRNA will lack the last exon of the upstream gene, and the first exon of the downstream gene. This phenomenon occurs because the last exon of a gene lacks a splicing donor site that is required for spliceosome function. Similarly, the first exon of a gene lacks a splicing acceptor site (Figure 4). Due to the lack of these splicing sites, both exons are spliced out of the mRNA transcript (Akiva et al. 2006). Since the stop codon of a protein-coding gene is usually found in the last exon, the splicing of the last and first exons can lead to the formation of a functional chimeric protein (Figure 4). The reason for the stop codon's preferential localization to the last exon of a gene is the avoidance of non-sense mediated decay, a cellular safety mechanism that degrades mRNAs whose coding sequence terminates prematurely before the last exon (Chang et al. 2007).

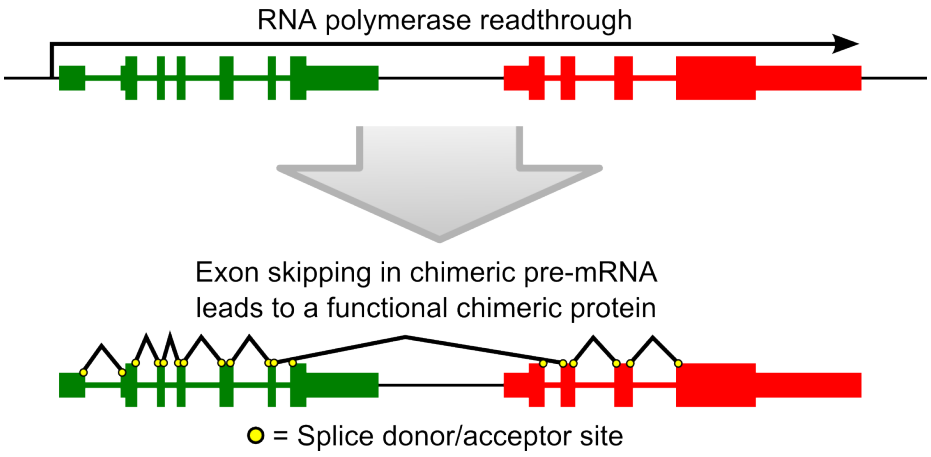

Figure 4. A read-through fusion is formed when an RNA polymerase continues transcribing beyond the end of a gene and transcription continues to an adjacent downstream gene. Exon skipping due to missing splice sites can give rise to a fusion transcript encoding a functional chimeric protein. Boxes indicate exons, thicker boxes indicate coding sequence.

Last and first exon skipping can also occur in fusion genes that arise from chromosomal rearrangements. In this way a rearrangement can produce a functional fusion protein even though one or both genomic breakpoints localize to intergenic regions. Consider a case where two genes A and B are located on the same chromosomal strand, and a deletion event removes the region between the two genes. Further, consider that the breakpoint in the upstream gene A is located in an intron, while the other breakpoint is located 20 kb upstream of gene B. Surprisingly, such a fusion gene can encode a functional chimeric protein, as the first exon of gene B is spliced out of the pre-mRNA (Figure 5). Similar reasoning applies to the case where one breakpoint is located downstream of gene A, and the other breakpoint in an intron of gene B (Figure 5). In fact, a functional fusion protein may arise even if both breakpoints are located in intergenic regions outside genes A and B. Examples of exon skipping in cancer-associated fusion genes are rare, but first exon skipping has been observed in *BCR-ABL1* fusions (Laurent et al. 2001).

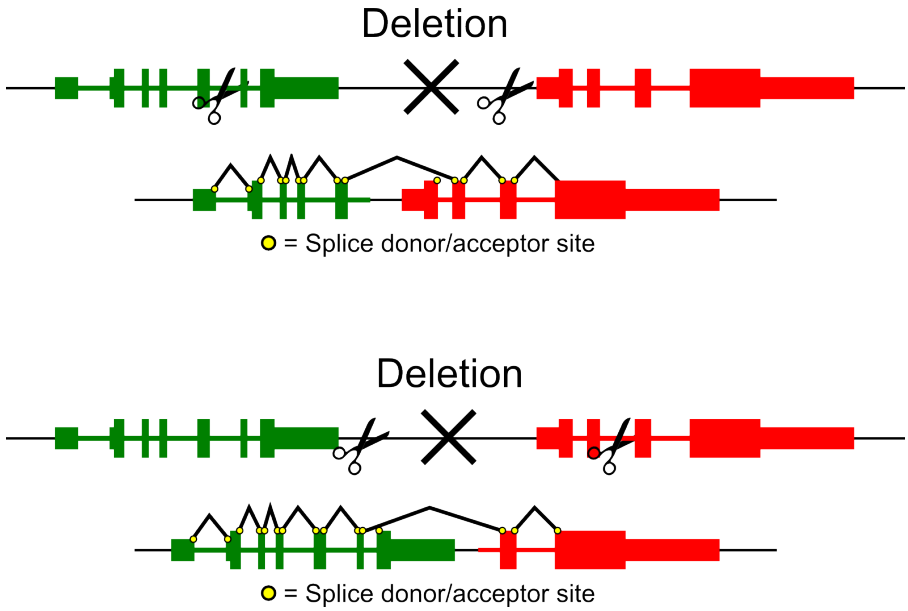

Figure 5. A chromosomal rearrangement with intergenic breakpoints can result in a fusion gene encoding a functional chimeric protein. Illustration depicts two example scenarios. Boxes indicate exons, thicker boxes indicate coding sequence.

2.4 Pathology of brain cancer

Tumors of the brain and central nervous system are rare but difficult diseases with an estimated worldwide mortality of 100,000 people per year (Ferlay et al. 2008). These cancers are difficult to treat because of the vital nature of the involved organs: radical surgery is not possible, and even small tumors can have lethal consequences. Molecular therapy is also more difficult due to the circulatory limitations imposed by the blood-brain barrier.

The most common type of brain cancer in humans are the gliomas. Gliomas are brain cancers that originate from glial cells: a family of non-neuronal cells that perform vital support functions for neurons. Gliomas can be subdivided into ependymomas, astrocytomas, oligodendrogliomas, and mixed gliomas (Louis et al. 2007). The most common form of glioma is a form of high-grade astrocytoma called *glioblastoma*, a highly lethal and aggressive form of brain cancer. The standard-of-care for glioblastoma is surgical resection, followed by radiotherapy and adjuvant temozolomide. Without treatment, life expectancy after glioblastoma diagnosis is 6 months. Modern treatment regimes have increased the median survival time to 14.6 months (Stupp et al. 2005), but the cancer is still invariably lethal.

The genetic mechanisms that drive glioblastoma have been extensively studied, but many open questions still remain. Many glioblastoma cases are known to involve mutually exclusive high-level amplification of the receptor tyrosine kinases *EGFR, PDGFRA,* and *MET*. Other known alterations include deletion of *CDKN2A/B*, amplification of *CDK4*, and deletion of the tumor suppressor *PTEN* (Cancer Genome Atlas Research Network 2008). However, no recurrent fusion genes had ever been discovered in glioblastoma.

3 METHODS

3.1 High throughput measurement

In the past 20 years, many new technologies have become available for the study of the constituents and interactions within biological cells. DNA microarrays and high throughput sequencing in particular have made it possible to comprehensively catalog the genomic and transcriptomic events that occur inside cells. Figure 6 highlights some of the high throughput technologies used in the study of cancer genomics today.

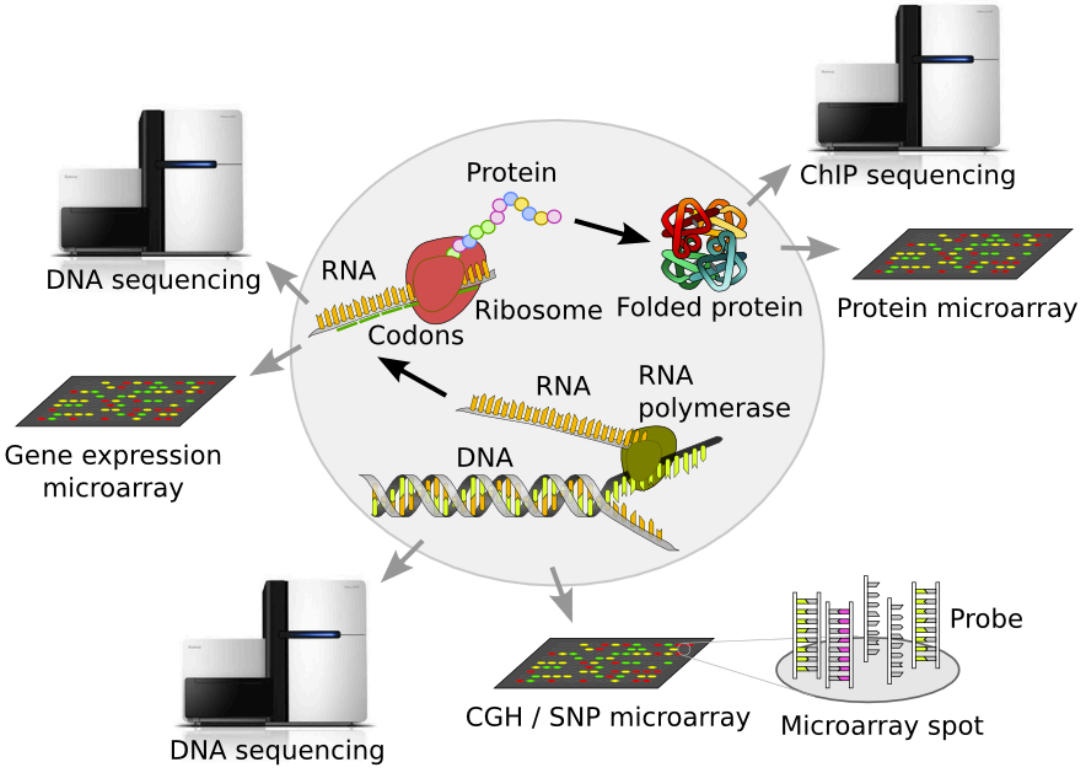

Figure 6. Overview of genome-wide measurement technologies used in the field of cancer genomics today. The middle portion of the figure represents the canonical DNA -> RNA -> protein model of information flow inside cells.

3.1.1 DNA microarrays

Ever since the role of DNA as the blueprint of life was first demonstrated by Avery, MacLeod and McCarty (Avery et al. 1944), people have deviced new strategies for the efficient study of this biopolymer. In 1995, miniaturized DNA microarrays were introduced for the high throughput analysis of DNA fragments with specific sequences (Schena et al. 1995). The basic idea behind DNA microarrays is simple: spots of oligonucleotide probes are printed onto a specially designed surface, and fluorescently labeled DNA fragments from a sample are allowed to base pair with the probes. All oligonucleotide probes in a spot have identical sequences, and so DNA fragments containing a complementary sequence will hybridize to them (Figure 7). Automated fluorescence imaging is used to estimate the number of labeled DNA fragments that have hybridized to the probes in each spot. Modern off-the-shelf microarray platforms can contain hundreds of thousands of spots, enabling the simultaneous interrogation of thousands of different sequences in a single experiment.

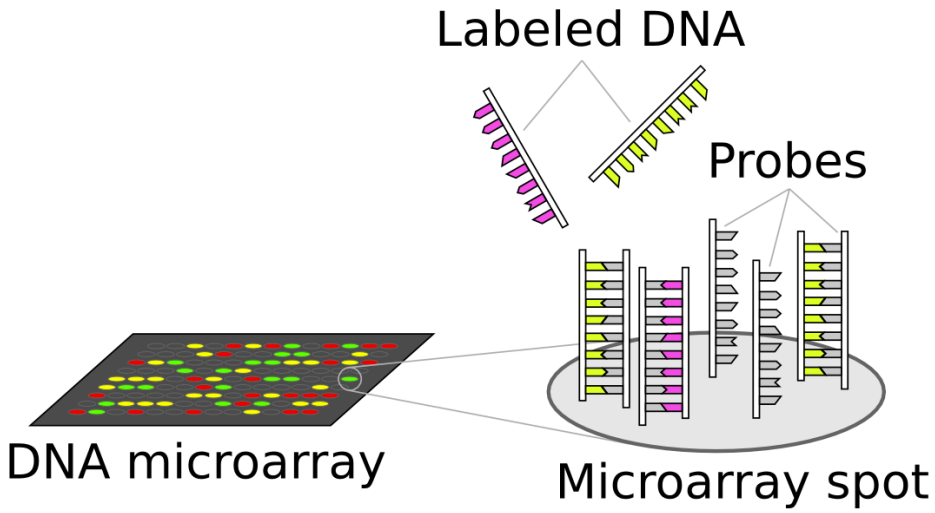

Figure 7. Illustration of the basic principle behind DNA microarrays. Spots of oligonucleotide probes are printed on a surface, and each spot contains multiple DNA probes with an identical sequence. DNA from test and control samples is labeled with different fluorescent dyes and is allowed to hybridize onto spots on the microarray based on sequence complementarity.

By careful probe design, DNA microarrays can be used to probe a number of different genetic features. The main applications of DNA microarrays are:

- Transcriptomic expression profiling, where the enzyme reverse transcriptase (RT) is used to convert RNA into complementary DNA (cDNA), which is then hybridized onto a microarray to calculate expression levels for individual transcripts, exons, or microRNAs (Schena et al. 1995).
- Array comparative genomic hybridization (aCGH), where genomic DNA is hybridized to determine the copy number of different chromosomal loci (Solinas-Toldo et al. 1997; Pinkel et al. 1998).
- Single nucleotide polymorphism (SNP) profiling, where hybridization of genomic DNA is used to identify individual nucleotides at known polymorphic or mutant sites (Mei et al. 2000).
- Methylation profiling, where methyl-immunoprecipitated or bisulfite-treated DNA is hybridized onto an array, and probe intensities are used to determine whether the probed sites are methylated in a test sample (Gitan et al. 2002).
- Chromatin immunoprecipitation profiling (ChIP-chip), where antibodies are used to capture DNA fragments bound by a specific protein, and probes tiling the whole genome are used to determine genomic sites bound by the protein (Blat et al. 1999).

Despite their usefulness, microarrays have a number of limitations that must be taken into account when designing experiments. The first limitation is that probes cannot be changed after an array has been designed or manufactured. Since our knowledge of the human genome has only recently achieved a high standard, old microarray platforms often contain probes that are not actually complementary to their intended targets, or lack probes for genetic features that were discovered after the array was designed. A second limitation is that hybridization does not require perfect complementarity, and hence labeled DNA fragments will also attach to probes with near-match sequences. This non-specific hybridization causes background noise in experiments. Thirdly, microarrays are subject to a number of experimental artifacts, including dye bias (Yang et al. 2002) and spatial artifacts (Wilson et al. 2003) (Figure 8).

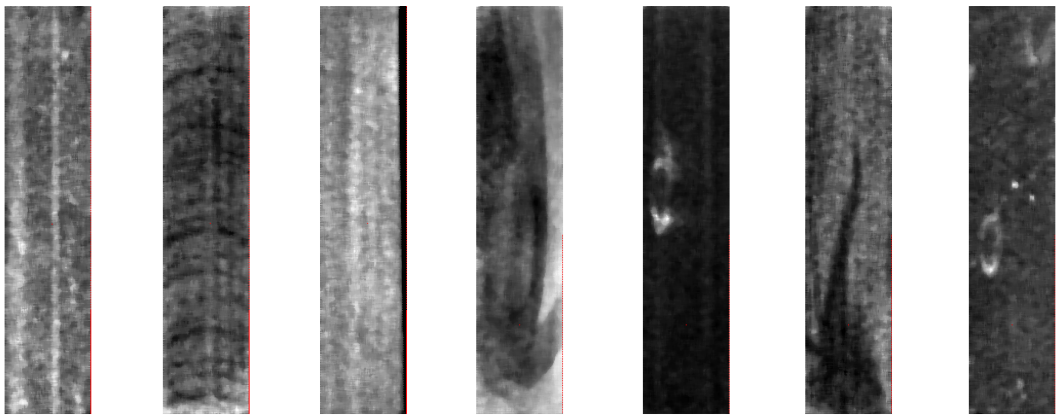

Figure 8. Examples of spatial artifacts in seven microarray hybridization experiments.

3.1.2 High throughput sequencing

The term high throughput sequencing (HTS) describes a family of new technologies aimed at sequencing millions of DNA fragments per day. These technologies are based on the idea of splitting chromosomes or cDNA transcripts into short fragments that are then sequenced in millions of parallel chemical reactions, producing short nucleotide strings or "reads" that are typically between 20-200 bases in length (Mardis, 2008). The current generation of HTS platforms can interrogate tens of gigabases of sequence per day, and sequencing costs are falling rapidly. Indeed, sequencing technologies have recently displaced DNA microarrays in many applications.

A major benefit of HTS platforms over DNA microarrays is that they characterize the total DNA/RNA content found in cells, whereas microarrays only interrogate features selected by the manufacturer. Sequencing technologies also tend to have lower noise levels and less bias, although this depends on the technology and chemistry used. The main sources of bias in sequencing experiments are:

- GC content bias, which causes fragments with high or low GC content to be sequenced to a lower depth.
- Amplification bias, where PCR cycles lead to non-uniform fragment amplification.

All sequencing platforms are subject to sequencing errors, which manifest as sporadic nucleotide substitutions or insertions/deletions (indels) in read sequences. Error rates differ between platforms; some platforms are more subject to indels than substitutions, and vice versa. Error rates can also vary by offset into the read. For instance, the ABI SOLiD platform has higher error rates at the 3' ends of reads.

Once a sequencing run has finished, the sequencing instrument outputs all read sequences (known as *reads)* and associated per-base quality scores. The encoding of the sequence output can vary depending on the technology used, but typically the sequences are represented in either nucleotide space or colorspace. In nucleotide space representation, each sequenced base is denoted by an ACGT symbol or one of the IUPAC nucleotide ambiguity symbols (Cornish-Bowden 1985). Colorspace is a more complex representation that is used by sequencing platforms based on dinucleotide ligation, such as ABI SOLiD. The colorspace alphabet consists of four symbols, each carefully chosen to represent a set of four dinucleotides so that two subsequent colors uniquely specify a nucleotide (Breu 2010). The benefit of colorspace representation is that it makes it possible to differentiate between sequencing artifacts and true single nucleotide mutations in sequencing data (McKernan et al. 2009; Mardis 2008).

To interpret the results of a sequencing experiment, the read sequences produced by the instrument are *aligned* against reference sequences or assembled *de novo* to reconstruct the sequenced chromosomes or transcripts. Read alignment is a process where a computational algorithm takes a short input sequence and tries to find a matching region in a set of larger reference sequences. If reads are aligned against chromosome sequences, for example, the resulting alignments contain information about the relative contribution of different genomic regions to the collection of DNA fragments that were sequenced by the instrument.

Repetitive sequences in the human genome pose a major challenge for sequencing experiments due to the short read lengths of current technologies. This is because short reads originating from repetitive elements cannot be linked to any particular repeat, as the read sequence matches with all of them. This issue has been partially resolved through the introduction of paired end sequencing, a sequencing protocol where both ends of DNA fragments are sequenced. Since DNA is usually fragmented to a length of 200-500 bases, this means that a paired end read pair from a fragment can be uniquely localized by aligning both of the reads against a reference, and then filtering out alignments where the pairs are situated farther than 500 bases apart (Fullwood et al. 2008).

3.2 Wet-lab techniques

3.2.1 Reverse transcription

Reverse transcription is a laboratory technique by which RNA is converted into DNA. This effect is achieved through the use of *reverse transcriptase* enzymes isolated from retroviruses (Myers et al. 1976). The term *complementary DNA* (cDNA) is used when referring to reverse transcribed DNA. The process of reverse transcription is a highly useful technique, as it allows scientists to study RNA molecules using techniques developed for the analysis of DNA. The short-lived nature of RNA would render many experiments difficult, but this problem is circumvented through the use of reverse transcribed RNA as a proxy for the original RNA.

3.2.2 Polymerase chain reaction

Polymerase chain reaction (PCR) is a technique for amplifying (copying) DNA. In this technique, double stranded DNA is repeatedly melted and duplicated using DNA polymerase enzymes, resulting in exponential amplification of DNA (Saiki et al. 1988). The DNA polymerase used in the reaction cannot construct the complementary strand from scratch, but requires the presence of a primer complementary to the template strand, which is extended to create the complementary strand. Through careful design of the primer sequences, DNA can be amplified selectively, so that only desired sequences are amplified.

PCR has many applications in biology, ranging from the global amplification of DNA for sequencing purposes to the validation of the presence of particular DNA/RNA sequences within cells. When PCR is performed on cDNA, the process is referred to as RT-PCR. PCR can be used to measure the levels of specific RNA transcripts within a cell. This technique, known as quantitative RT-PCR (or qPCR), begins with the reverse transcription of RNA into cDNA. PCR with carefully designed primers is then used to amplify cDNA arising from the transcript of interest, and the levels of amplified cDNA are compared against a reference for quantification.

The use of PCR in biological experiments can introduce artifacts. For instance, PCR efficiency is highly dependent on the GC content of sequences. Such factors, combined with the exponential nature of PCR, can easily introduce strong non-uniformities in the amplification of different sequences. PCR chimaeras are another common artifact where the PCR reaction fuses two unrelated DNA fragments together, introducing anomalous sequences in the data (reviewed in Kanagawa 2003). This effect is particularly trouble-

some for high throughput sequencing, although this artifact can be somewhat mitigated through the use of emulsion PCR (Williams et al. 2006).

3.2.3 Immunoblotting

Immunoblotting (also known as Western blotting) is a technique that allows one to estimate the quantity and molecular weight of select proteins of interest. To perform an immunoblot, the cells in a sample are homogenized and the protein content is extracted. The proteins are then placed at one end of a gel containing multiple columns, and a voltage is applied to the gel. This causes the proteins to migrate through the gel at a speed that depends on the protein's size. Voltage is cut before the protein molecules exit the gel, and the proteins are transferred onto a membrane while maintaining the location they had within the gel. A labeled antibody specific to the protein of interest is then used to probe for the protein of interest. Protein from multiple samples can be analyzed simultaneously on an immunoblot by racing the proteins in different columns of the gel. (Burnette, 1981)

3.3 Genome assemblies and annotations

The first nearly complete human reference genome was sequenced and assembled by the Human Genome Project in 2004 (International Human Genome Sequencing Consortium 2004). This reference genome did not represent the genome of any single individual; instead it was an amalgamation of multiple human genomes. Since this time, many genomes of individual humans have been sequenced. All of these genomes are different: in general, no two human genomes are exactly alike. The differences between individual genomes range from single nucleotide polymorphisms (SNPs) to large structural variations. The total inter-individual variation for humans has been conservatively estimated at 0.5% (Levy et al. 2007). Since lack of a common reference makes communication difficult, geneticists have defined reference genomes that represent the most common alleles and structural variants found in the human population. The human reference genome is currently maintained by the Genome Reference Consortium (Church et al. 2011).

A reference genome forms the basis for genomic annotations that denote known functional features of the genome. Examples of such annotations include transcriptome annotations from NCBI and Ensembl, the SNP database dbSNP (Sherry et al. 2001), and the microRNA database miRBase (Griffiths-Jones, 2004). Both the reference genome and annotations are updated at relatively frequent intervals as new knowledge is gathered.

In this thesis, the following reference genomes and annotation were used:

- Human reference genome: GRCh37
- Human transcriptome: NCBI RefSeq release 38
- Human microRNAs: miRBase release 18

3.4 Fusion gene discovery

A number of different strategies have been proposed in the literature for the identification of fusion genes from high throughput sequencing data. One proposed strategy is to perform whole genome DNA sequencing and look for chromosomal breakpoints using specialized algorithms. These algorithms often use a reference genome and look for paired end reads whose ends align to opposite sides of a chromosomal breakpoint (Chen et al. 2009). Another proposed strategy is to look for evidence of fusion transcripts in transcriptome sequencing data (Maher et al. 2009; Maher et al. 2009). The latter approach has significant cost benefits due to reduced sequencing depth, as only a small fraction of the human genome is transcribed at a significant level.

When fusion discovery is done on transcriptome sequencing data, it is possible to make use of the fact that fusion gene breakpoints tend to occur in intronic regions. By this we mean that the breakpoint for the chromosomal rearrangement leading to the fusion is within an intron, so that at the RNA level, two intact exons from separate genes are fused together. This suggests that a simple approach for fusion discovery would be to pick all 200,000 exons in the human exome, and directly align reads against all potential junctions between pairs of those exons. The downside is that this would require the alignment to be performed against 40 billion exon pairs, a task that is not computationally feasible.

To solve the problem, we implemented a fusion discovery algorithm that searched for fusion genes using short anchors extracted from both ends of each read. This approach to fusion gene discovery is not novel; the same technique was used in 2009 by Maher et al. We use the term *anchor-based junction discovery* when referring to algorithms that employ this approach. Our implementation of the algorithm was distinct because our software was designed to work with reads as short as 50 bp, and to support colorspace reads produced by the Applied Biosystems SOLiD series of sequencing instruments. To our knowledge, apart from a commercial service provided by Applied Biosystems, no other software provided these features at the time of our algorithm's implementation. Our software also implements a sophisticated set of filters designed to reduce the number of false positive fusion candidates reported by our tool. Table 2 lists the fusion discovery algorithms that are most widely used today.

Table 2. A comparison of widely used software packages for fusion gene detection.

Software	Installation requirements	Uses DNA-seq to identify genomic breakpoints?	Detects exon disrupting fusions?	Supports colors-pace reads?	References
ChimeraScan	Python, Bowtie	No	No	No	Iyer et al. (2011)
Comrad	Perl, Bowtie, Blat	Yes	Yes	No	McPherson et al. (2011)
Defuse	Perl, Bowtie, Blat	No	Yes	No	McPherson et al. (2011)
Tophat-Fusion	Python, Bowtie	No	Yes	Yes	Kim et al. (2011)
ShortFuse	Python, Bowtie	No	Yes	No	Kinsella et al. (2011)

The implementation of our fusion gene discovery algorithm is shown in Figure 9. The algorithm begins with a filtering step where all reads are aligned against both the reference genome and transcriptome. The Bowtie short read alignment software (Langmead et al. 2009) is used to perform the alignments. All reads that align against either are discarded from further analysis, since their presence can be directly explained through normal transcriptional processes[1]. We are left with anomalous reads that may or may not arise from fusion transcripts. Before flagging a read as evidence of a fusion transcript we need to consider a number of alternative hypotheses:

- The read originates from an as-yet unannotated transcript variant of some gene
- The read originates from a mutated genomic site (either a SNP or indel)
- The read is a result of RNA editing
- The read originates from a PCR chimaera
- The read contains multiple sequencing errors and actually originates from a homologous sequence elsewhere in the genome

To determine the origin of the anomalous reads, we take each anomalous read and split it into two anchor sequences: one from the 5' end and one from the 3' end of the read. The anchor lengths are equal and chosen so that the anchors do not encompass the entire read. If the reads are of varying length, we discard any reads for which the anchors would overlap. Next, we use Bowtie to align the anchors against all annotated exon sequences from the reference transcriptome, while maintaining pairing information between the 5' and 3' anchors. This is akin to paired end read alignment, except that the "paired reads" here are far shorter, and the expected distance between two anchors is fully determined.

In extracting the anchors, colorspace reads require special treatment because the color sequences always begin with a nucleotide symbol that represents the starting base, followed by colors. However, Bowtie ignores the starting nucleotide of a colorspace read, and therefore we simply always place a T as the starting nucleotide for the second anchor.

[1] For the reads that align to the genome but not to the transcriptome, we assume that they arise from unannotated transcriptionally active sites, or are a result of sporadic low-level transcription.

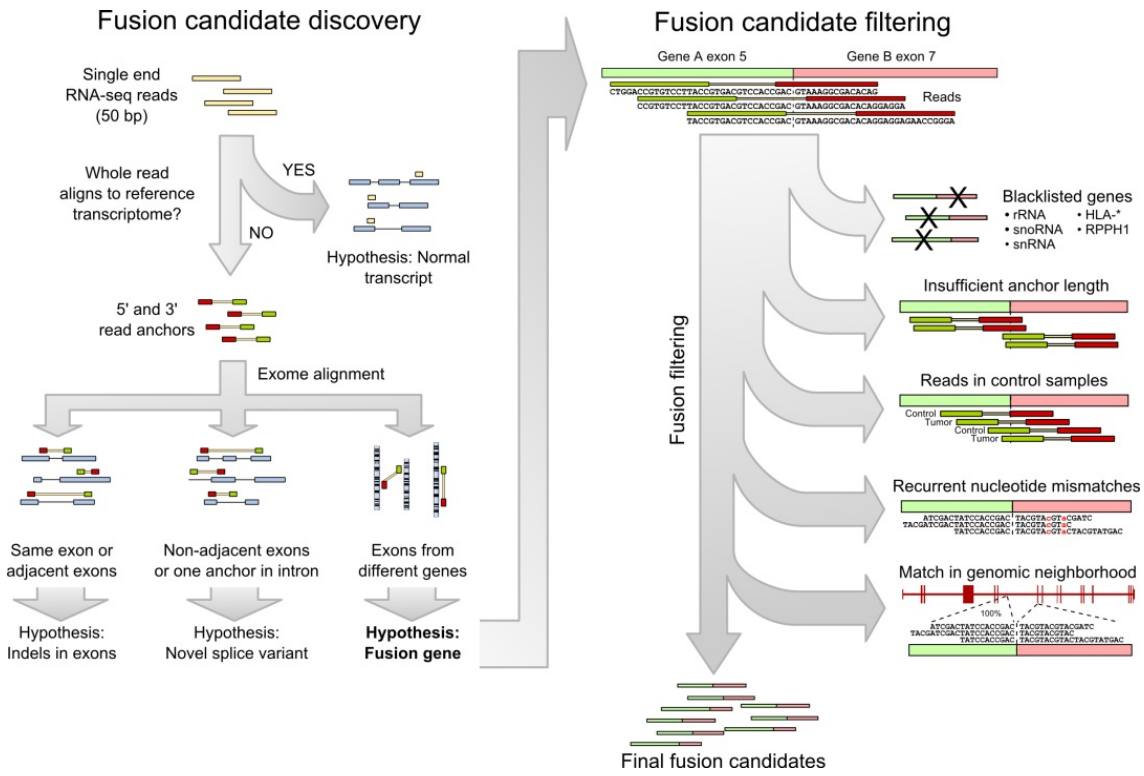

Figure 9. An illustration of the fusion candidate discovery algorithm and the cascade of filters used to discard false positives.

Next, we go through the list of anchor alignments and look for cases where both anchors of a read align to at least one known exon. For each such case, we produce a list of putative exon pairs by taking the Cartesian product $E_5 \times E_3 = \{(e_5, e_3)\}$, where E_5 and E_3 are the aligned exons for the 5' and 3' anchor, respectively. To reduce the computational overhead due to anchors with highly abundant sequences, we filter out reads for which either anchor aligns to more than 3 different exons. We also discard all anchor pairs where both anchors aligned to exons of the same gene as such pairs represent novel splice variants, not fusion genes. After these steps, we are left with a list of putative fusion junctions that some reads may potentially overlap with. To find full-length reads that align to the putative fusion junctions, we construct a new Bowtie index out of the fusion junctions, and align all anomalous reads against the index. To ensure that reads overlap both sides of a fusion junction, the putative junction sequences are built so that flanking sequences on both sides of the junction are 5 bp shorter than the read length.

The selection of anchor length is a trade-off between analysis runtime and sensitivity. For our 50 bp colorspace reads, we used an anchor length of 19 bp, leaving a window of 12 bp within which an exon-exon junction would have to fall in order to be detected. We allowed no mismatches in anchor alignments, but two color mismatches were allowed when the full-length reads were aligned against candidate junctions.

3.5 Filtering of fusion candidates

Our initial test runs of the fusion discovery algorithm produced tens of thousands of fusion candidates, the vast majority of which were false positives. To improve the specificity of our software, we implemented a cascade of filters to discard fusion genes that showed clear and automatically detectable signs of being false positives (Figure 9). We will now describe the filters one-by-one. Candidate fusion genes were discarded if they failed even one of the filters.

3.5.1 Blacklisted genes

The construction of a complementary DNA (cDNA) library for transcriptome sequencing is a complex process that involves multiple steps. Some of the steps are known to cause technical artifacts such as chimeric cDNA sequences that combine parts of two unrelated RNA sequences. One source of false chimeras is the reverse transcription step. Reverse transcriptase enzymes are prone to template switching, an event where the enzyme jumps to another template without terminating DNA synthesis (Houseley et al. 2010). Template switching has been proposed as an explanation for the anomalous chimeric transcripts that show up in transcriptome sequencing but are not supported by DNA level alterations (Houseley et al. 2010). Another potential source of false chimeras is the PCR amplification step where cDNA fragments are amplified to increase the amount of DNA available for sequencing. PCR chimeras have been proposed to arise when incomplete elongation occurs during a PCR cycle and the incomplete product partially hybridizes with an unrelated template, followed by chimeric elongation (Figure 10) (reviewed in Kanagawa 2003). False chimeras are enriched among highly transcribed genes such as ribosomal RNA (rRNA), small nuclear RNA (snRNA), and small nucleolar RNA (snoRNA).

In accordance with these observations, we noted that many of the candidate fusions reported by our software involved ribosomal RNA genes or other highly transcribed genes. We therefore chose to discard all fusion genes involving the highly transcribed genes and gene families *RN18S1*, *RN28S1*, *RPPH1*, *SNORD**, *SNORA**, *RNY**, *RN7SL** and *RNU**. Chimeras involving these genes were also observed in normal brain tissue pools, ruling out the possibility of widespread rRNA gene fusions in brain cancer.

We also discarded fusions involving genes located in hypervariable regions of the genome, such as the *HLA* and immunoglobulin loci. These genomic regions undergo exon shuffling during mitosis in order to produce the diverse portfolio of antibodies and immune-related proteins found in human bodies (reviewed in Schatz et al. 2011). This natural exon shuffling resulted in many computationally identified fusion candidates that are not associated with cancer and were therefore discarded.

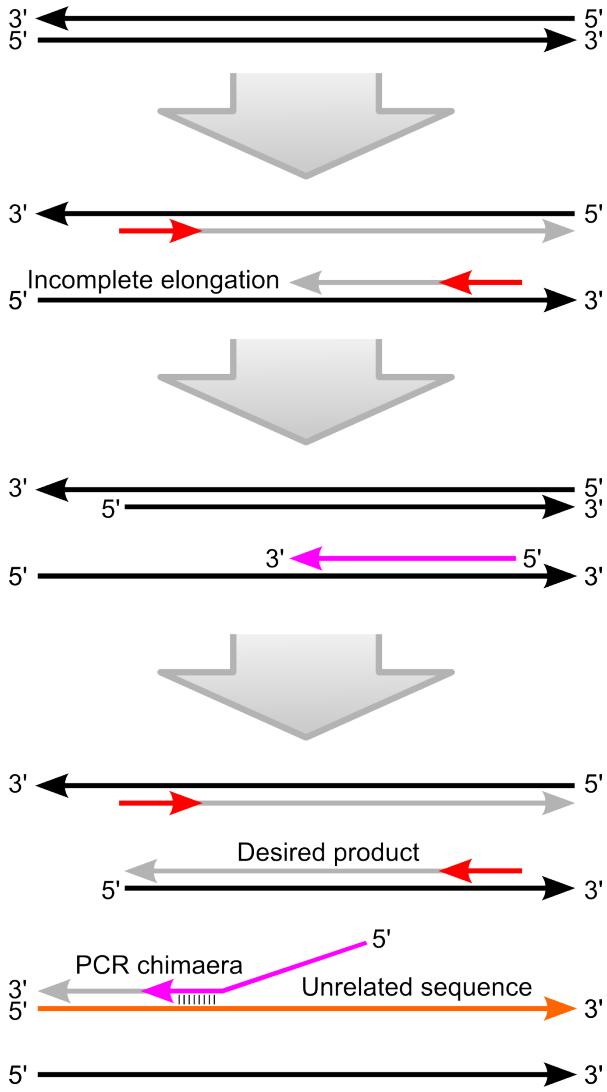

Figure 10. An illustration of the "incomplete elongation" theory for the formation of PCR chimeras. According to this theory, a PCR chimaera is formed when an incomplete elongation product (pink) of a PCR primer (red) hybridizes with an unrelated but partially homologous template (orange), followed by chimeric elongation.

3.5.2 Insufficient anchor overlap

We noted that many fusion candidates did not overlap the junction properly: in some cases the supporting reads only overlapped 5 bases on one side of the junction. We realized that these were cases where the anchor-based junction discovery had discovered a false positive junction, and unrelated reads had then aligned to the junction with only weak support on one side of the junction. We proceeded to discard fusion candidates for which no supporting read contained at least 10 bases on both sides of the junction. Note that 10 bases is shorter than the anchor length used, allowing us to identify more supporting reads in the full read alignment step than were initially identified during anchor-based alignment.

3.5.3 Presence in control samples

In our tests, many fusion candidates showed supporting reads in both tumor and control samples. As we were only interested in fusion genes associated with brain cancer, we discarded all fusion candidates that showed supporting reads in the control samples. It is worth noting that examples of germline fusion genes are known in the literature. For instance, the fusion gene *TFG-GPR128* was initially associated with lymphomas and soft tissue tumors, but was later found to be present in the germlines of healthy individuals (Chase et al. 2010).

3.5.4 Recurrent nucleotide mismatches

We noted that many fusion candidates showed recurrent nucleotide mismatches on one side of the fusion junction. By recurrent nucleotide mismatches we mean a nucleotide at a fixed offset from the fusion junction showing frequent mismatches relative to the reference genome. In many such cases, the location of junction-overlapping reads was also biased towards the other side of the junction, implying that the reads were of different origin. We decided to filter out any fusion candidates for which the average number of sequence mismatches per supporting read was above 0.7. This filter was controversial as it ran the risk of filtering out fusion genes immediately adjacent to a single nucleotide polymorphism (SNP) exhibiting an alternate allele. However, by setting the threshold at 0.7 we ruled out the possibility of a single heterozygous alternate allele leading to a fusion gene being discarded.

3.5.5 Homology in genomic neighborhood

While analyzing some of our fusion candidates using the web-based BLAST alignment software (Altschul et al. 1997), we noticed that some of the fusion candidates actually represented cases where an annotated exon was spliced together with a nearby unannotated exon that happened to start with a similar sequence as an annotated exon elsewhere in the genome. To get rid of this class of false positives, we implemented a filter to check that the 3′ flank of the fusion junction did not have perfect alignments against any genomic sequence within 50kb of the 5′ flank of the fusion junction, and vice versa.

3.6 Prioritization of fusion gene candidates

Since the number of false positives in fusion discovery is often quite high, the ranking of fusions according to their estimated significance becomes very important. The goal is to rank the fusion candidates based on their estimated biological and clinical significance, and the likelihood of the fusion gene being a true positive. The most immediate line of evidence about the biological significance of a fusion gene is provided by the number of reads overlapping the fusion junction. This quantity is important because it reflects the expression level of the fusion gene, and a highly expressed fusion gene has a higher probability of having a significant impact on a cell's phenotype. A fusion gene with many supporting reads also has a reduced likelihood of being a false positive caused by a random sequencing error. Taken together, this suggests ranking fusion candidates in descending order according to the number of supporting reads.

But how should fusion genes be scored when a cohort of samples is searched for fusion genes? Summing the total read evidence across all samples is a good approach, but can run into trouble if all samples are showing supporting reads for a false positive fusion. A better scoring system is achieved by taking into account the heterogeneity of cancer and noting that true cancer-associated fusion genes are usually found only in a subset of patients. We therefore implemented a scoring method where the distribution of supporting reads for each fusion is tested for goodness of fit against a discrete uniform distribution using Pearson's chi-square test. The fusions are then ranked in ascending order according to their p-values calculated with the goodness-of-fit test (Figure 11). This scoring system is particularly useful if the number of control samples is low.

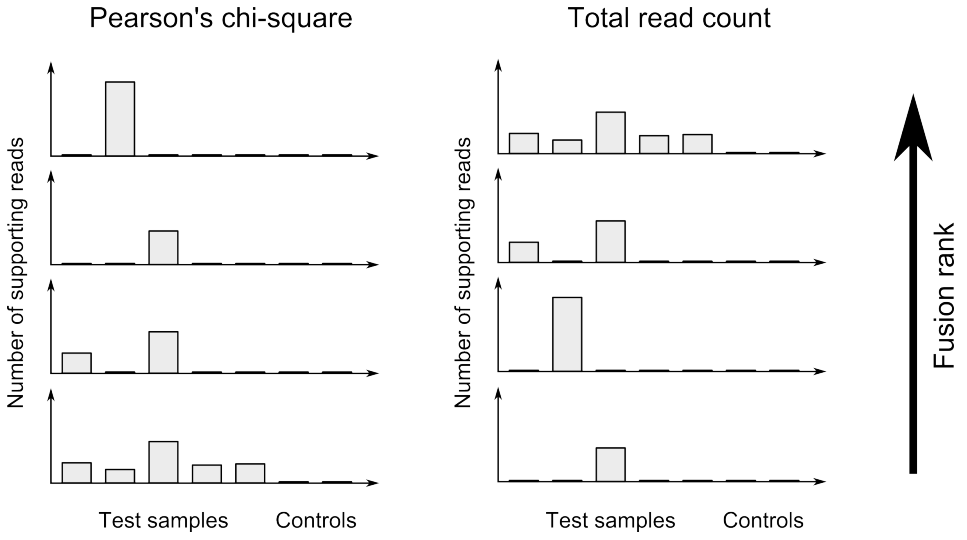

Figure 11. An illustration of the difference between ranking fusion genes based on total read count and Pearson's chi-square statistics for goodness-of-fit.

Another improvement to scoring can be achieved by noting that the library construction phase preceding high throughput sequencing can introduce PCR amplification artifacts where the same cDNA fragment is amplified and subsequently sequenced multiple times. Such *duplicate reads* do not represent the true biological abundance of an RNA fragment, and therefore cannot be directly counted as evidence of a fusion gene's existence. To counteract this artifact, one can count each group of duplicate reads only as a single unique read. To determine which reads are duplicates, two reads can be considered duplicates if they align to the same location relative to the junction. Note that being aligned to the same location does not guarantee that two reads really are duplicates, and therefore this duplicate removal step will reduce the sensitivity of the analysis to a small degree.

3.7 Transcriptomic expression profiling

To quantify the expression levels of all exons in the genome, a custom-built Matlab script was used to extract exon sequences from the transcriptome GBFF files stored on the NCBI RefSeq FTP server. We then used Bowtie to align the whole transcriptome sequencing data against exon sequences parsed from NCBI RefSeq release 38 transcriptome annotations, allowing for a single color mismatch against the reference genome in each alignment. Because the RNA had been reverse transcribed into double-stranded cDNA before sequencing, reads were allowed to align to exon sequences in both forward and reverse-complement orientation. The number of reads aligned to each exon was calculated using a Matlab script that parsed the Bowtie output. To correct for inter-sample sequencing depth bias and exon length bias, we normalized the read counts using the RPKM normalization method:

$$\mathrm{RPKM} = \frac{\text{number of hits}}{\text{exon length in kb} \times (\text{total reads}/10^6)}$$

RPKM stands for *reads per kilobase of transcript length per million reads* and was first introduced by Mortazavi et al. (2008). The normalization method is designed to address two sources of bias in sequencing experiments:

- The number of reads that originate from a specific target (here an exon) depends on the length of the target sequence. This is because the cDNA is sonicated into short fragments before sequencing, and the number of fragments produced depends on the original length of the transcript or exon.
- The number of reads that originate from a specific target depends on the overall amount of reads produced from the entire sample. This is because the total read count in a sample is not a measure of total RNA content, but is instead determined by the protocol and sequencing instrument used.

Because RPKM does inter-sample normalization using total read counts, the method relies heavily on the expression of a handful of highly expressed genes. If these few genes are expressed at a higher level in one sample, RPKM normalization will down-play the expression of other genes in that sample. To resolve this issue, the inter-sample normalization in RPKM can be replaced with *median-of-ratios normalization*. In this approach, each sample is represented as a column vector of gene expression values. Two vectors are normalized by calculating a ratio between each pair of genes (i.e. pointwise ratios between the two vectors), and the median of the ratios is calculated. The median-of-ratios is a robust estimate of the multiplicative bias between two samples. The expression values in the vectors are divided by the median-of-ratios statistic to produce normalized values.

In this study, we normalized expression values using the RPK statistic to correct for gene/exon length, and median-of-ratios normalization for correcting inter-sample multiplicative bias.

3.8 Gene expression analysis using cDNA microarrays

To calculate gene expression levels using DNA microarrays, reverse transcribed RNA is hybridized onto a microarray containing probes complementary to transcript sequences (Schena et al. 1995). Some gene expression microarrays use a dual channel setup where cDNA from a test sample and a control sample are labeled with different fluorescent dyes and simultaneously hybridized onto the microarray. Other microarrays use a single channel setup where only cDNA from a test sample is hybridized onto the array. Sample quantity and quality can affect observed fluorescent intensities, causing systematic differences between samples (hybridization experiments). In a dual channel setup, these differences can be ameliorated by using test/control channel ratios as the quantities that are compared between samples. However, differences in the relative quantities of the two dyes can still cause sample-specific bias. Normalization methods such as quantile normalization and median-of-ratios normalization can be used to combat this source of bias.

In quantile normalization, the goal is to normalize N vectors of data so that their distributions become identical. The vectors must be of the same length M, and are represented as an $M \times N$ matrix. Each column of the matrix is sorted, and means are calculated across the rows of the sorted matrix. Each column of the sorted matrix is then replaced with the mean vector. At this point, all columns of the matrix share the same distribution. In the final step, each column is returned to its original order (Figure 12). Quantile normalization is a powerful technique because it can correct any sample-specific bias that can be represented as a monotonic transformation of the true values. However,

quantile normalization often results in truncated distribution tails, which can be problematic if very highly expressed genes are of interest in a study.

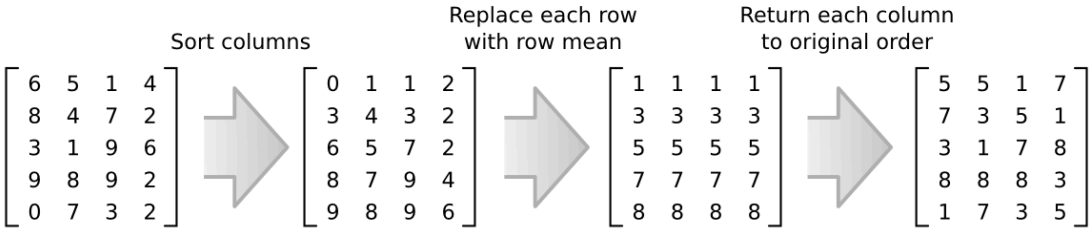

Figure 12. An illustration of the quantile normalization process. After quantile normalization is finished, all column vectors have identical distributions.

Gene expression microarrays are often designed so that each transcript is targeted by multiple probes of different sequences. A set of probes targeting a specific transcript is called a *probeset*. To calculate the expression level of a transcript, probe intensities must be summarized across all probes targeting the same transcript. Multiple summarization methods have been proposed in the literature, ranging from the calculation of a simple mean all the way to algorithms such as *median polish*. Median polish, originally introduced by Tukey (1977), is a robust computational method where the data y_{ij} are represented through an additive model with two variables a and b:

$$y_{ij} = c + a_i + b_j + e_{ij}$$

Or in matrix form:

$$Y = c + \begin{bmatrix} a_1 + b_1 & \cdots & a_M + b_1 \\ \vdots & \ddots & \vdots \\ a_1 + b_N & \cdots & a_M + b_N \end{bmatrix} + E$$

Here c represents the constant background, and e_{ij} represents noise in the original data. In the context of gene expression microarrays, Y is a matrix containing measured log-scaled probe intensities for all probes that target a specific transcript, so that y_{ij} represents the log-scaled measured intensity of probe j in sample i. a_i represents the true transcript expression level in sample i, and b_j represents the *affinity* of probe j to the transcript. The affinity of a probe is a multiplier that relates the true abundance of a transcript to the observed fluorescence signal. To estimate the true transcript expression levels a_i from the data Y, the median polish algorithm performs successive operations known as row and column sweeps. In a row sweep the median of each row is subtracted from that row. In a column sweep the median of each column is subtracted from that column. The medians used for subtraction are tallied in vectors $a = [a_1 \quad \dots \quad a_M]$ and $b = [b_1 \quad \dots \quad b_M]$. Row and column sweeps are applied until the table no longer changes or changes very little. After this, the medians of vectors a and b are subtracted from the vectors and used to define $c = \text{median}(a) + \text{median}(b)$. At this point, the original data Y has been decomposed into a background component c, transcript expression le-

vels a_i, and probe affinities b_j. The median polish algorithm is robust because row and column sweeps are largely unaffected by outlier values in Y.

3.9 Copy number analysis using CGH microarrays

Copy number alterations (CNA) such as deletions and duplications are commonly found in many cancers (Beroukhim et al. 2010). They can be studied using a technique known as array comparative genomic hybridization (aCGH), a technique where test and control DNA are fluorescently labeled with different dyes (typically cyanine-5 and cyanine-3), and then hybridized onto a microarray slide (Figure 13). Once hybridization is complete, an optical reader is used to measure the intensities of the two fluorescent dyes at each spot on the microarray. The data is preprocessed to remove spatial trends and dye bias, and the intensities of the two fluorescent dyes are compared at each microarray spot to calculate the logratios $\log_2(I_{test} - I_{control})$. Each microarray spot contains probes with a specific sequence, and this sequence maps to a unique location in the human genome. The probes (and associated logratios) are ordered based on their position in the genome, and the data is segmented to discover CNA boundaries (Figure 13). Segmentation in this context refers to a process where the data is searched for contiguous regions where the logratios of multiple probes differ significantly from their surroundings, indicating a greater or lesser amount of DNA from that region being present in the cells.

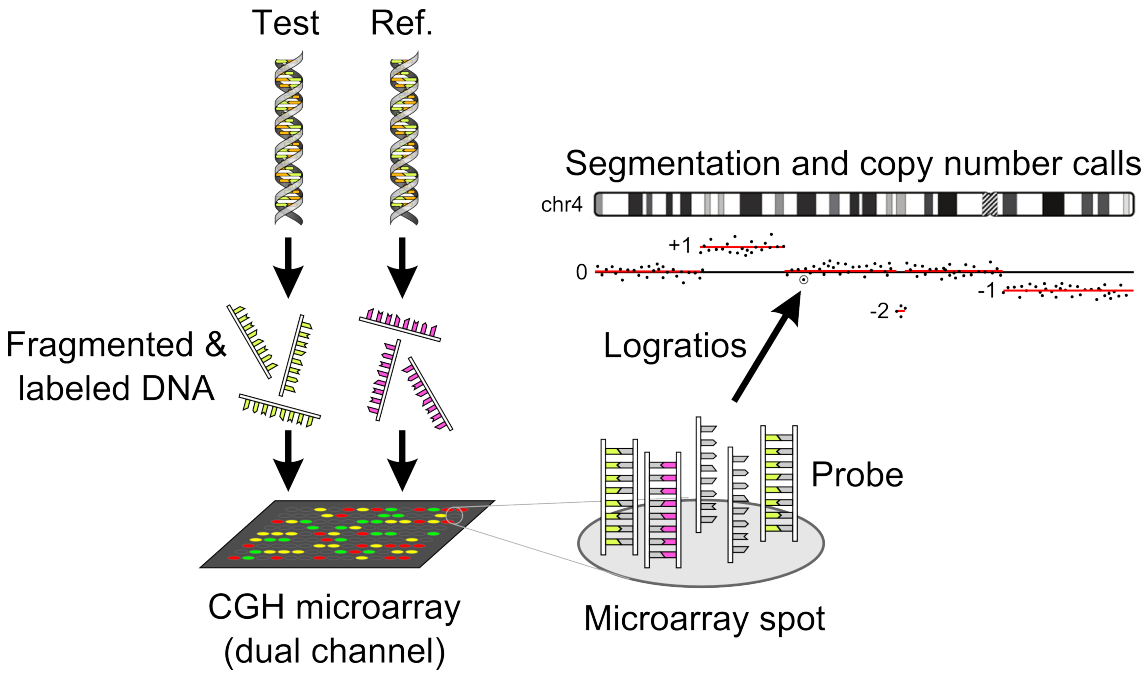

Figure 13. Overview of the different technologies used in studying the molecular biology of cancer today. The middle portion of the figure represents the canonical DNA -> RNA -> protein model of cellular function.

In this study, we used the Agilent Feature Extraction Software to perform spatial detrending and dye bias correction. We extracted all probe sequences from the microarray design files and aligned them against the GRCh37 human reference genome to ensure that the probe locations agreed with the latest genome assembly. Using the updated probe mappings, we used the circular binary segmentation algorithm (Olshen et al. 2004) to segment the probe logratios.

4 RESULTS

4.1 Whole transcriptome sequencing of gliomas

Forty glioma tissue samples were acquired from the Brain Tumor Center tissue bank of the University of Texas MD Anderson Cancer Center. The samples included 20 glioblastoma, 5 anaplastic astrocytome, 6 anaplastic oligodendroglioma, and 9 oligodendroglioma tissues. Additional commercial adult and fetal normal brain tissue was also acquired for use as control.

All samples were obtained from surgery and snap frozen. RNA was extracted, then depleted of ribosomal RNA using the Invitrogen Ribominus Eukaryotic Kit. The remaining RNA was reverse transcribed and amplified using the SOLiD Small RNA Expression Kit. cDNA fragments 50-100 bases long were selected using gel electrophoresis followed by cutting the gel. The samples were pooled into 8 pools according to glioma type and sequenced using an Applied Biosystems SOLiD 3 instrument at the Sequencing Core Facility at MD Anderson (Table 3). The sample pools were sequenced to an average depth of 66 million reads. Reads were encoded in colorspace and had a fixed length of 50 colors. The data was received in raw FASTQ format, with associated per-color quality information.

Table 3. Sample pools used for whole transcriptome RNA sequencing.

Pool ID	Pool contents	# of samples	Seq. depth
ZW01	Glioblastoma, survival < 6 months	5	75M reads
ZW02	Glioblastoma, survival 10-15 months	5	75M reads
ZW03	Glioblastoma, survival 15-20 months	5	52M reads
ZW04	Glioblastoma, survival > 20 months	5	67M reads
ZW05	Anaplastic astrocytoma (AA)	5	58M reads
ZW06	Oligodendroglioma	5	66M reads
ZW07	Oligodendroglioma	4	88M reads
ZW08	Anaplastic oligodendroglioma (AO)	6	70M reads
ZW09	Adult normal (prepooled)	23	58M reads
ZW10	Fetal normal (prepooled)	21	55M reads

4.2 Fusion gene discovery

We applied the anchor-based fusion discovery algorithm to whole transcriptome sequencing data from the eight glioma pools and the two normal brain tissue pools. The search produced an initial list of 17564 fusion candidates, but the list was reduced to 52 candidates after applying the cascade of filters described in section 3.5 (Figure 14).

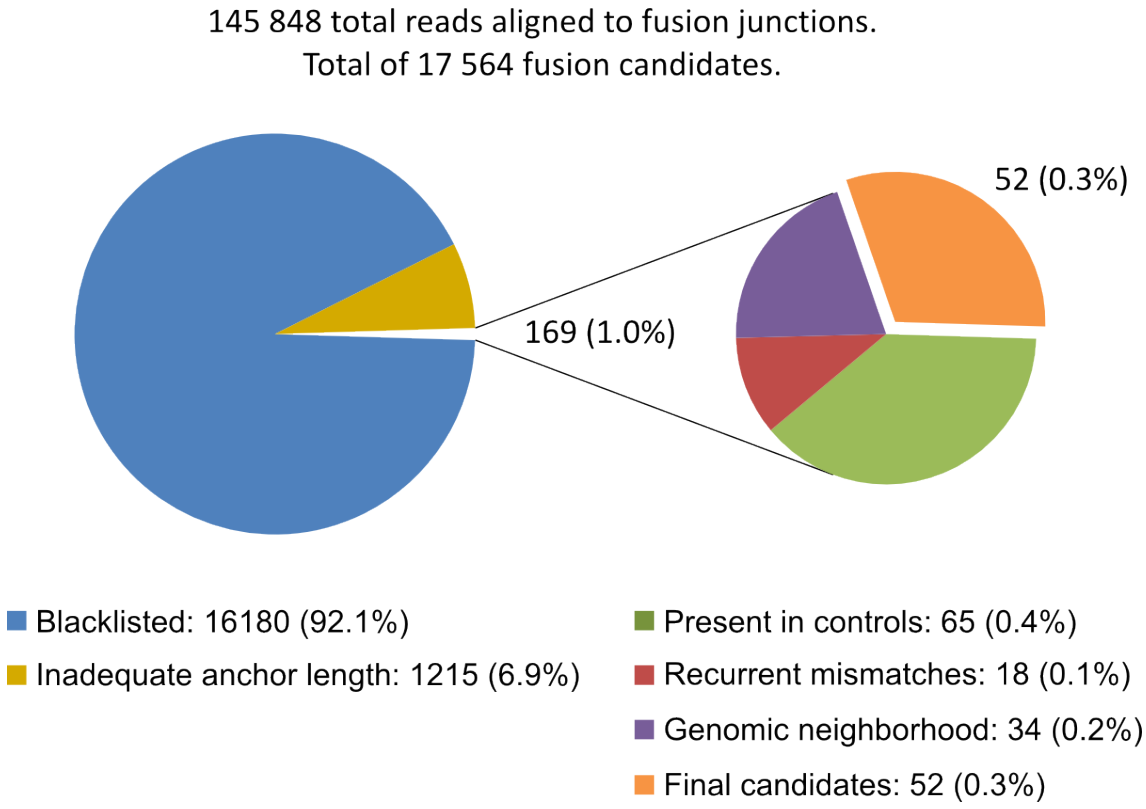

Figure 14. Pie chart showing the number of remaining fusion candidates after each filtering step. A total of 52 fusion candidates remained after applying the full cascade of filters.

The list of 52 putative fusions was scanned for interesting candidates based on each candidate's predicted biological consequence and level of supporting evidence. To determine whether a fusion gene had the potential to be biologically significant, we looked at whether the fusion merges the coding sequences (CDS) of the two genes, or whether it merges the CDS of one gene to the 5' UTR or 3' UTR of the other gene. If coding sequences were fused, we calculated whether the fusion resulted in a frameshift. We also used existing literature to determine whether either of the involved genes had any known associations with glioma or cancer in general. The resulting final list of 7 interesting fusion candidates is shown in Table 4.

Table 4. Top fusion candidates based on the level of evidence and predicted consequences.

Fusion	Exons	Evidence	Putative mechanism
FGFR3-TACC3	e18-e11	16 reads in GBM pool #3.	Tandem duplication
NUP188-SPTAN1	e8-e27	5 reads in GBM pool #4.	Tandem duplication
ZNF713-VSTM2A	e1-e4	4 reads in GBM pool #3.	Tandem duplication
YEATS2-GPBP1	e2-e9	3 reads in oligodendroglioma pool #1.	Interchromosomal translocation
NPAS3-AKAP6	e5-e2	2 reads in GBM pool #3.	Tandem duplication
TADA2B-SORCS2	e1-e2	2 reads in GBM pool #3.	Deletion
CYB5R1-AGXT2L2	e5-e12	2 reads in GBM pool #3.	Interchromosomal translocation

The most striking finding in the data was a putative fusion between fibroblast growth factor receptor 3 (*FGFR3*) and transforming acidic coiled coil protein 3 (*TACC3*) in glioblastoma pool #3. The *FGFR3-TACC3* fusion was supported by a total of 16 reads that overlapped the fusion junction. The reads showed no nucleotide mismatches and were distributed evenly on both sides of the junction (Figure 15). Genomic regions close to *FGFR3* did not contain sequences homologous to the 3' side of the putative fusion junction, and a BLAST alignment found no contiguous genomic regions containing a close match for the junction sequence. Taken together, these findings strongly suggested that the fusion gene was a real biological event and not a technical artifact.

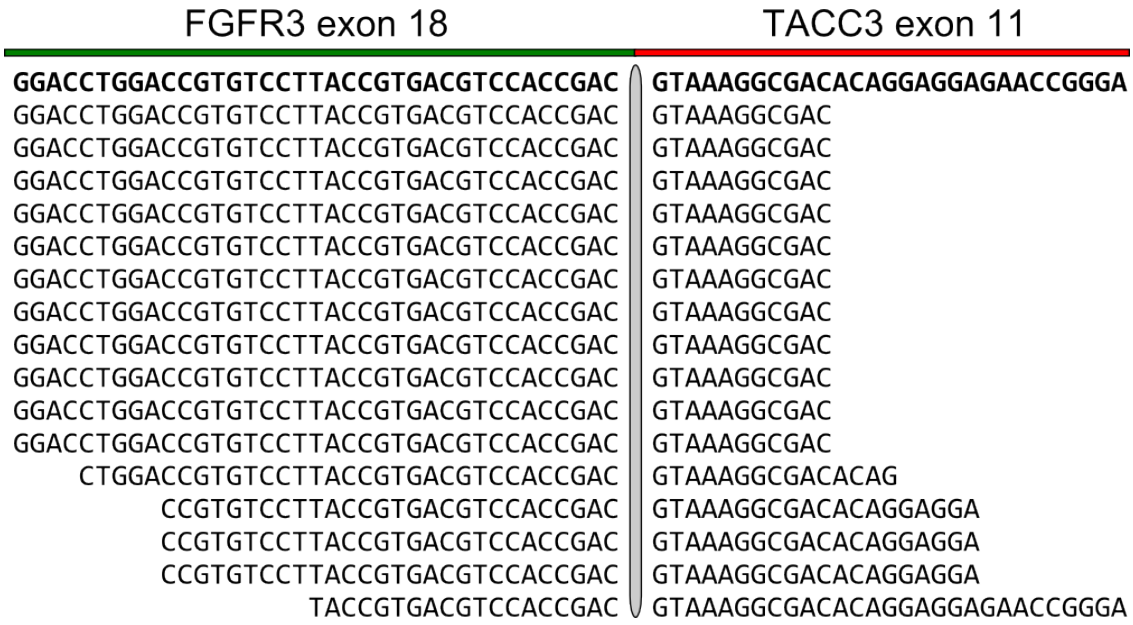

Figure 15. Visualization of the 16 RNA-seq reads that supported the fusion between *FGFR3* and *TACC3* in the third glioblastoma pool. Genomic reference sequence is shown in bold.

4.3 Protein level validation of *FGFR3-TACC3*

To determine the exact number of *FGFR3-TACC3* positive tumors, we acquired an *FGFR3* antibody specific to an epitope found in the N-terminal portion of *FGFR3*. An antibody with a C-terminal epitope would not have detected our fusion protein, as a segment of the *FGFR3* C-terminal had been replaced in the chimeric protein. As a positive control, we transfected SNB19 cells (a widely used glioblastoma cell line) with a vector containing an *FGFR3-TACC3* e18-e11 construct. As a negative control, we transfected SNB19 cells with a vector containing the wildtype *FGFR3* gene. We used the antibody to immunoblot cell lysates from the 40 gliomas in our cohort (plus controls), and observed that 2 of 40 glioma samples produced strong bands indicative of a fusion gene. The first fusion positive tumor (GBM-13) was from glioblastoma pool #3 and yielded bands that matched those from the fusion-transfected SNB19 cells, implying that this sample was the origin of the fusion gene discovered by transcriptome sequencing. The second fusion positive tumor (GBM-07) was from glioblastoma pool #2 and produced weaker bands, implying lower expression of the fusion gene. This weaker expression explained why no fusion-supporting reads were detected from this pool by transcriptome sequencing. We repeated the immunoblot validation with 51 glioma samples acquired from the Tumor Tissue Bank of the Tianjin Medical University Cancer Institute and Hospital. 2 of 51 samples (denoted GBM-T01 and GBM-T02) in the Tianjin cohort were positive for the fusion gene (Figure 16). In total, we observed the *FGFR3-TACC3* fusion in 4 of 48 glioblastomas and in none of 23 lower grade gliomas, implying that the fusion gene was specific to glioblastoma.

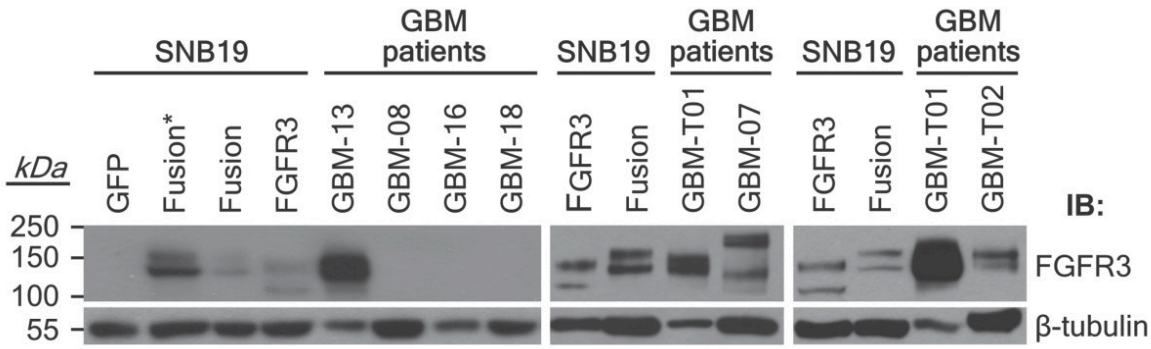

Figure 16. FGFR3 immunoblot of glioblastoma tumors and SNB19 cells transfected with the *FGFR3-TACC3* fusion or wildtype *FGFR3.* β-tubulin was used as loading control. Molecular weight ladder is shown on the left. Asterisk denotes stable transfection.

Fusion negative tumors (represented by GBM-08, GBM-16, and GBM-18) and control SNB19 cells transfected with green fluorescent protein (GFP) produced no immunoblot bands, indicating low endogenous FGFR3 protein level in gliomas (Figure 16). In fusion positive tumors and positive controls, we observed that the antibody always produced two bands. The upper and lower bands corresponded to the fully and partially N-glycosylated forms of FGFR3, respectively. In fusion positive tumors and controls, the bands had migrated a shorter distance in the gel than bands from the *FGFR3*-transfected SNB19 cells, consistent with fusion-induced replacement of the *FGFR3* C-terminal with a longer C-terminal from *TACC3* (Figure 16). Using a molecular weight calculator we calculated a 17 kDa difference between the masses of wildtype FGFR3 and the FGFR3-TACC3 fusion protein (e18-e11 variant), a result that agreed with immunoblot measurements. Intriguingly, we noticed that the bands in GBM-07 had a higher molecular weight than corresponding bands in GBM-13 and GBM-T01, implying that GBM-07 harbored a fusion variant that produced a longer chimeric protein. This prompted us to identify the exact structure of the *FGFR3-TACC3* fusion in all four fusion positive samples.

Immunohistochemical analysis of GBM-T01 and GBM-T02 patient tissues revealed extensive FGFR3 staining that was absent in control tissues (Figure 17), suggestive of a simple diagnostic measure that could be used in clinics to screen for fusion positive patients.

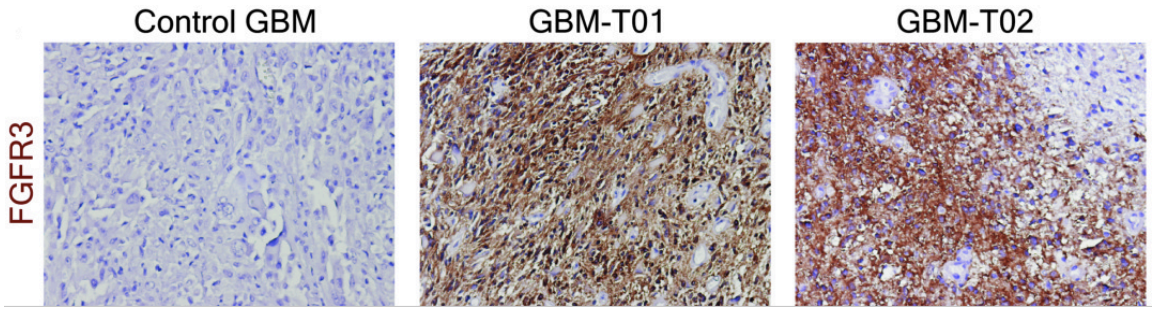

Figure 17. FGFR3 immunostaining of patients GBM-T01 and GBM-T02 and a control fusion-negative tumor. Original magnification, ×200.

4.4 Sanger sequencing of fusion junctions

To determine the exact structures of the fusion transcripts found in the four fusion positive tumors, we designed a set of nested primers for RT-PCR amplification and Sanger sequencing. The outer primers were designed to capture and amplify the RNA sequence surrounding the fusion junction. The inner primers were designed to act as starting points for the primer elongation step in Sanger sequencing. However, available Sanger sequencing instruments could only produce reliable sequences up to a 1000 bases long. This meant that inner primers had to be carefully designed for the correct exon boundaries, otherwise the fusion junction would be missed. For samples GBM-13 and GBM-T01, the primer design was straightforward as we knew the fusion structure of GBM-13 from transcriptome sequencing, and we knew that GBM-T01 likely shared the same structure (based on immunoblot bands).

To determine the fusion structure in GBM-07, we turned to transcriptome sequencing data and calculated the number of sequencing reads that fell within each exon of *TACC3*. We normalized the read counts by exon length, and then plotted the expression level of each of the 16 *TACC3* exons in the glioblastoma pools. As a positive control, we noted that in glioblastoma pool #3, exons 11-16 of *TACC3* were significantly overexpressed relative to exons 1-10 ($p = 0.00025$, Mann-Whitney U test) (Figure 18). This matched with expectations, as the fusion gene's expression is driven by the active *FGFR3* promoter, causing a spike in the expression of *TACC3* exons included in the fusion transcript. Looking at glioblastoma pool #2, we observed weaker but significant overexpression of exons 5-16 of *TACC3*, implying that the fusion gene in GBM-07 fused together *FGFR3* exon 18 with *TACC3* exon 5. This matched with immunoblot results that had shown GBM-07 to have a higher molecular weight than GBM-13.

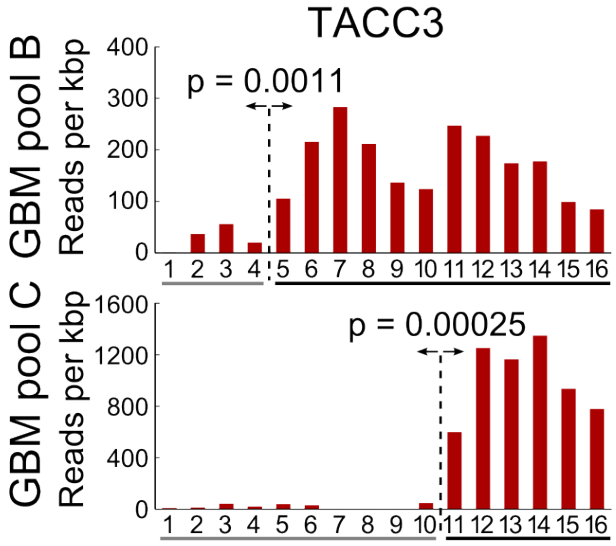

Figure 18. Expression of *TACC3* exons in the two glioblastoma pools harboring a fusion positive tumor. Read counts were quantile normalized and normalized by exon length. Dashed line indicates the location of the fusion junction in *TACC3*. P-values calculated using the Mann-Whitney U test.

Starting with the putative fusion junctions acquired by integrating transcriptome sequencing and immunoblotting, we designed a specific set of primers for each fusion positive tumor and performed reverse transcription and Sanger sequencing on RNA extracted from the tumors. After some trial and error with sample GBM-T02, we were able to sequence all four fusion junctions with single-base accuracy (Figure 19). Samples GBM-13 and GBM-T01 harbored the e18-e11 variant, while GBM-T02 harbored a slightly longer e18-e10 variant (the difference in molecular weight had been too small to be evident in immunoblots). GBM-07 harbored a fascinating fusion variant where the DNA breakpoints occurred inside exons 19 and 4 of *FGFR3* and *TACC3*, respectively. The two exons were disrupted by the fusion and merged together, but the chimeric protein was still in-frame. In fact, all four fusions were in-frame. This was strong evidence for the fact that only in-frame FGFR3-TACC3 proteins are under selective pressure, implying that the FGFR3-TACC3 protein drives tumorigenesis in glioblastoma.

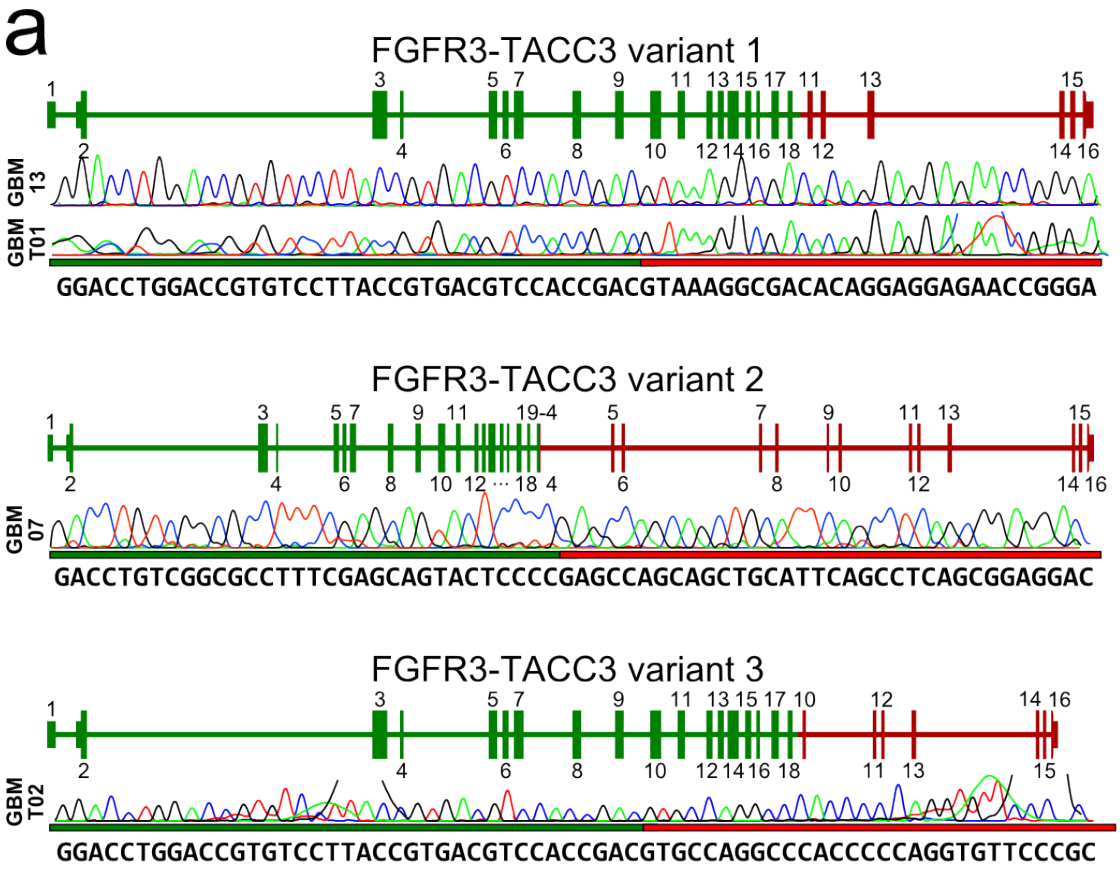

Figure 19. Fusion transcript structures and Sanger electropherograms for the four fusion-positive glioblastomas. GBM-07 and GBM-13 were patients treated at the University of Texas MD Anderson Cancer Center (MDACC); GBM-T01 and GBM-T02 were patients treated at Tianjin Medical University Cancer Institute and Hospital (Tianjin).

4.5 *FGFR3-TACC3* is caused by tandem duplication

After determining the exact structure of the fusion transcripts, we wished to know the nature of the chromosomal rearrangement that had produced the fusion gene. We immediately noted that *TACC3* and *FGFR3* were situated in the same chromosome and locus, separated by a distance of roughly 70 kb. Interestingly, the *TACC3* gene was located upstream of the *FGFR3* gene, while in the fusion gene *FGFR3* exons were found upstream of *TACC3*. This suggested that the fusion was caused by a 70 kb tandem duplication that partially overlapped both genes, as discussed in section 2.3.4 (Figure 20). Similar fusions caused by tandem duplications have been reported earlier in the literature (Jones et al. 2008; Lipson et al. 2012).

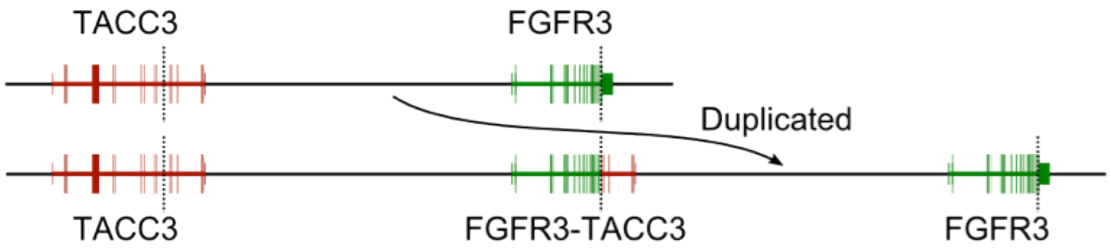

Figure 20. Based on the relative locations of *FGFR3* and *TACC3*, we hypothesized that the fusion gene was caused by tandem duplication of the region between the two genes.

To validate the presence of a tandem duplication, we hybridized genomic DNA from the four fusion positive glioblastomas onto custom-designed CGH microarrays with dense probe coverage at 4p16.3. The microarray slides were imaged using a laser-based Agilent scanner and the Cy5/Cy3 fluorescence ratios were segmented computationally to reconstruct the copy number landscape of the tumor cells. In all four fusion positive tumor samples, we obtained a result showing a clear duplication in the region between *TACC3* and *FGFR3* (Figure 21). This result supported our hypothesis of a tandem duplication causing the fusion gene.

Interestingly, in many of the samples the tandem duplication boundaries stretched beyond the 3' end of *FGFR3*. We initially found this problematic, because this implied that the fusion transcripts should have joined *FGFR3* exon 19 (the last exon) with *TACC3*. However, due to the lack of a splice donor site in the last exon, the exon is skipped by the RNA splicing machinery as described earlier in section 2.3.6.

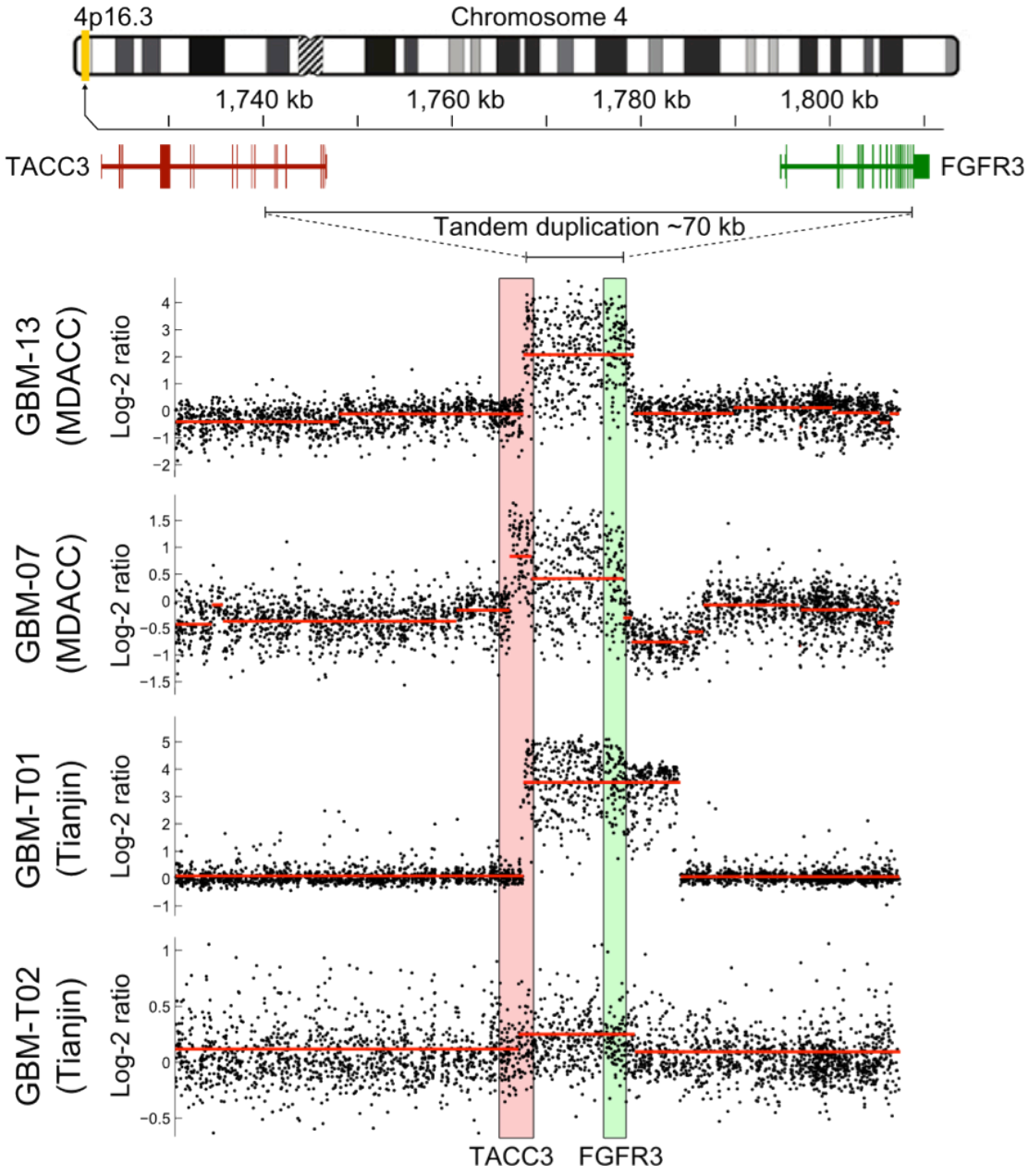

Figure 21. Validation of the tandem duplication in fusion positive glioblastomas using customized Agilent CGH microarrays with dense coverage for the fusion locus. Each dot represents a microarray probe. Probe signal logratios are measured relative to the reference channel containing commercial reference DNA.

The CGH microarray results clearly proved the presence of DNA duplication between *TACC3* and *FGFR3*, but did not conclusively prove the presence of tandem duplication, as the duplicated region might have been copied to a different region in the genome. To prove tandem duplication, we designed primers for Sanger sequencing the genomic breakpoints, using the rough breakpoints calculated from CGH microarrays as a guide. After lots of trial and error, the exact breakpoints were successfully sequenced and were found to confirm the tandem duplication hypothesis (Figure 22).

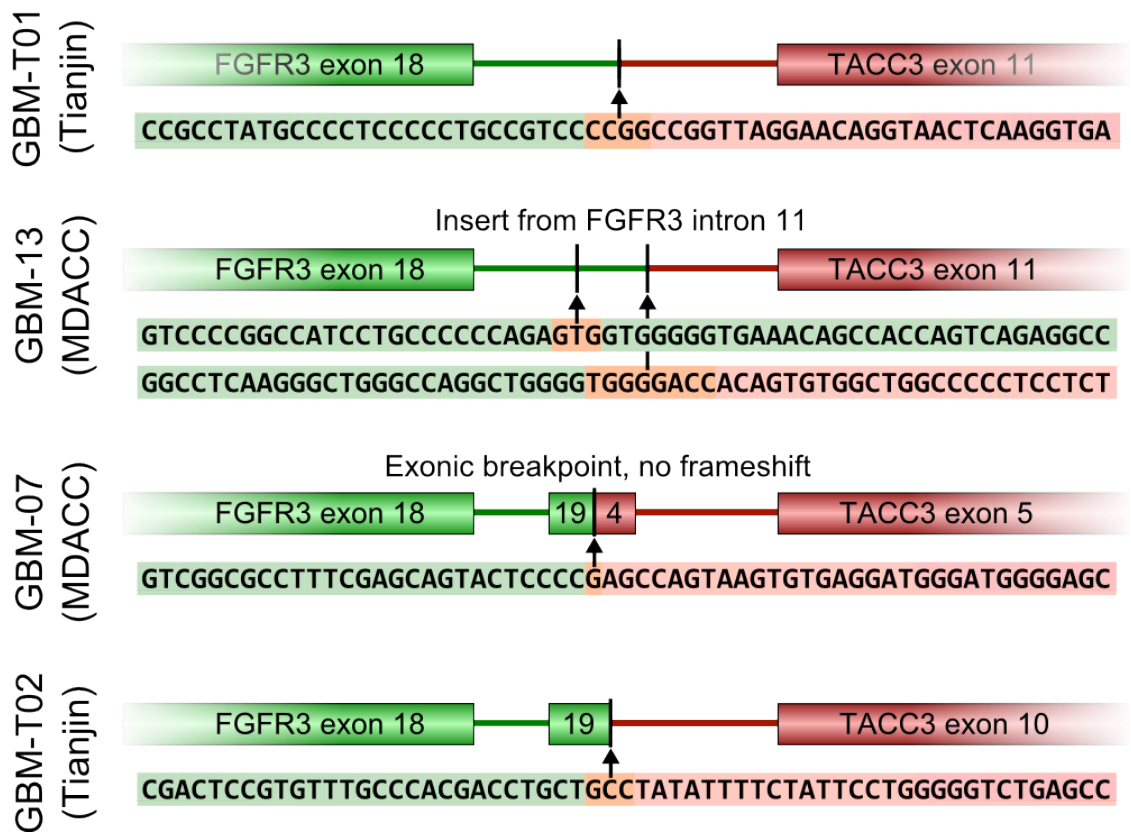

Figure 22. Genomic structures of the four fusion genes found in our cohort. Homologous sequence at the breakpoint is shown in orange colour.

4.6 Biological function of *FGFR3-TACC3*

To understand the biological significance of *FGFR3-TACC3* fusions in glioblastoma, we first looked at literature. FGF receptor 3 (FGFR3) is a membrane-bound growth factor receptor that is activated by its ligand, fibroblast growth factor (FGF). After activation, FGFR3 dimerizes and functions as a tyrosine kinase that activates the MAPK and PI3K pathways (reviewed in Eswarakumar et al. 2005; Turner et al. 2010). *FGFR3* is frequently mutated in bladder and cervical cancers (Cappellen et al. 1999), and the mutation leads to constitutive dimerization and auto-phosphorylation of the protein. Transforming acidic coiled-coil containing protein 3 (*TACC3*) encodes a centrosomal protein that is involved in mitosis (Gergely et al. 2000) and is overexpressed in lung and colon carcinomas and in multiple myeloma (Still et al. 1999). *TACC3* overactivity has been previously associated with a number of cancers, including GBM (Duncan et al. 2010).

Based on the fusion transcripts we inferred that all four in-frame fusion proteins in our cohort contained the extracellular Ig-like domains, the transmembrane domain, and most of the tyrosine kinase domain of *FGFR3*, fused to the transforming acidic coiled-coil domain of *TACC3* (Figure 23).

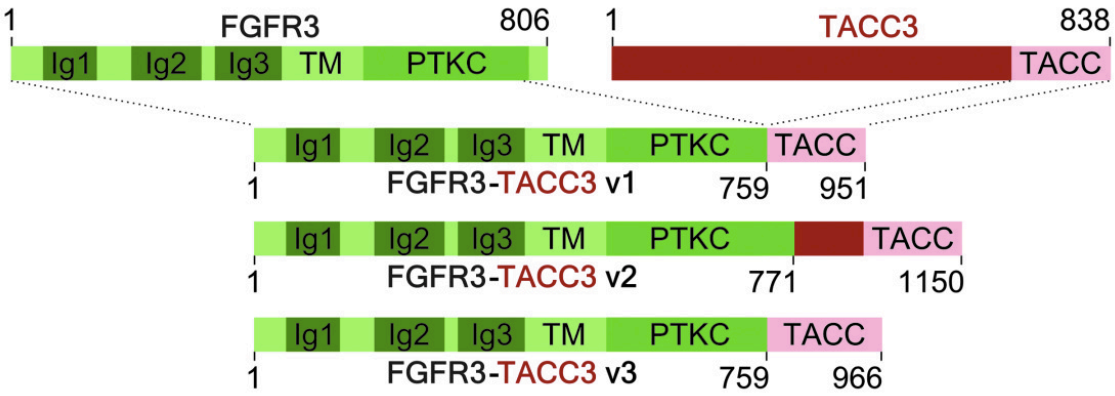

Figure 23. Schematic of protein domains — Ig, transmembrane (TM), and protein kinase (PTKC) — contained within the FGFR3-TACC3 fusion protein.

To determine how much the fusion affected the expression of *TACC3*, we performed qRT-PCR with primers designed to capture a region in the 3' end of *TACC3* transcripts (as only the 3' end was included in the fusion transcript). We found that the fusion gene had increased the expression of *TACC3* at least 20-fold when compared with fusion negative GBM samples (Figure 24).

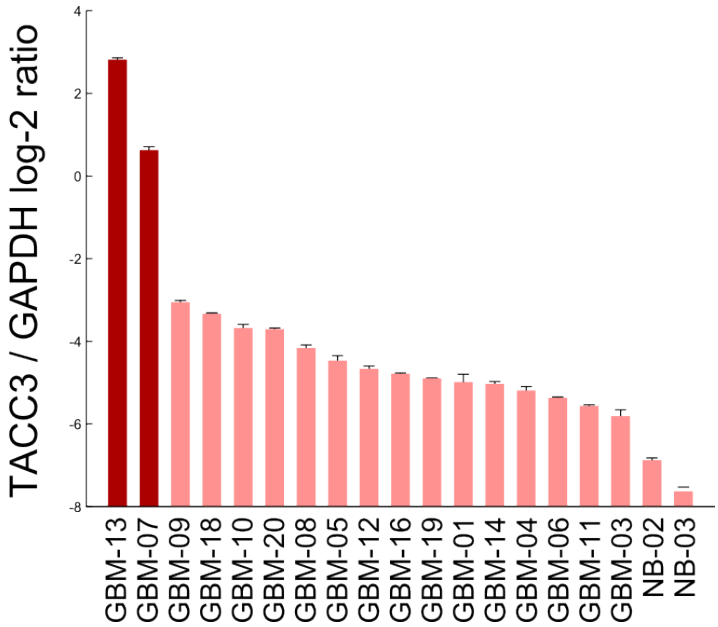

Figure 24. Log-2 ratio between the expression of *TACC3* and *GAPDH* as observed using qRT-PCR with primers designed to capture the 3' end of *TACC3*.

To interrogate the impact of the *FGFR3-TACC3* fusion on the molecular pathways within glioblastoma cells, we transfected SNB19 cells with a stable *FGFR3-TACC3* vector or an empty vector, then hybridized reverse transcribed cDNA from the two transfected cell lines onto Agilent gene expression microarrays. The probe intensities were quantile normalized and summarized into gene expression values using median polish. Logratios between each gene's expression in fusion-transfected and empty vector cells were calculated, and the Ingenuity Pathway Analysis software (Ingenuity® Systems, www.ingenuity.com) was then used to calculate the enrichment of differentially expressed genes in different biological pathways. The software calculated enrichment P-values using Fisher's exact test, and also calculated activation Z-scores by making use of a proprietary database where each gene participating in a biological function is annotated either as an inhibitor or activator of the biological function. The results of the analysis suggested that transfection of the fusion gene into SNB19 glioblastoma cells further increased the activity of biological pathways related to tumorigenesis and cellular mobility (Figure 25).

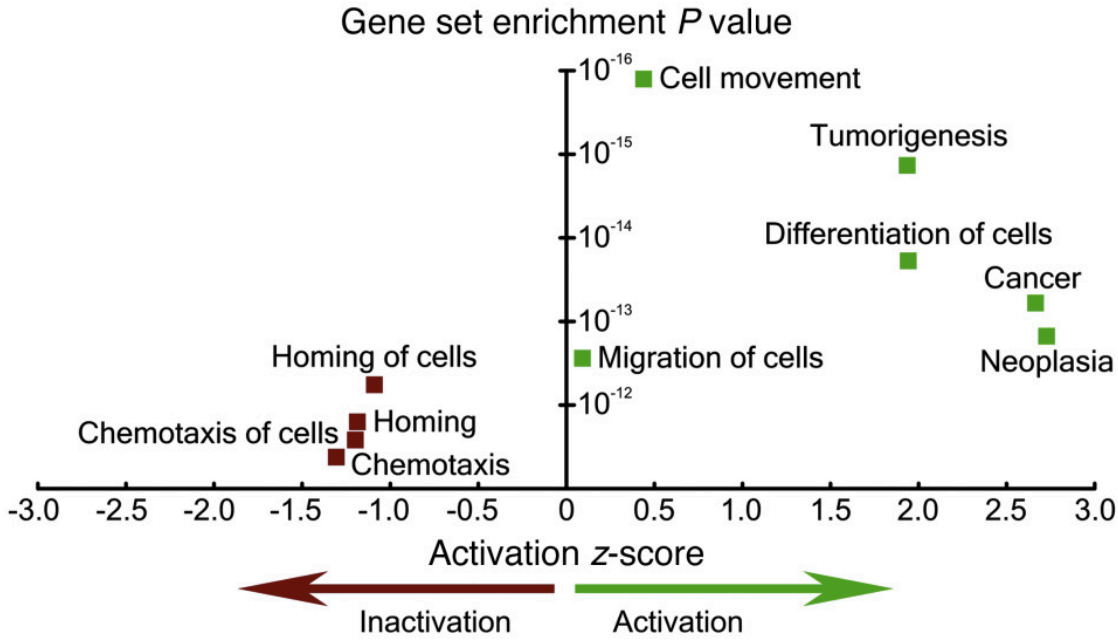

Figure 25. Scatter plot showing the gene set enrichment of different biological functions, calculated based on the genes differentially expressed after *FGFR3-TACC3* transfection into SNB19 cells. A gene set's enrichment P-value is indicated by its position along the y-axis. Activation or inactivation of a gene set is indicated by its position along the x-axis.

Finally, to show that the fusion gene affects tumor growth *in vivo*, we implanted *FGFR3-TACC3*, wildtype *FGFR3*, and empty vector SNB19 cells into the brains of immunocompromised mice and compared survival patterns (Figure 26). One million cells from each line were injected into the brains of a total of 35 nude mice (n = 5 per group). 5 tumor-free mice died of diarrhea and were censored, while the rest of the mice developed large tumors by the time of their termination. Mice implanted with the *FGFR3-TACC3* fusion died significantly earlier than mice implanted with empty vector (within 70–80 days of implantation compared with 110–175 days; P = 0.007, log-rank test). There was no statistical difference in survival observed between mice implanted with empty vector or wildtype *FGFR3*, implying that the oncogenic function of the fusion gene is not explained by overexpression of *FGFR3* alone.

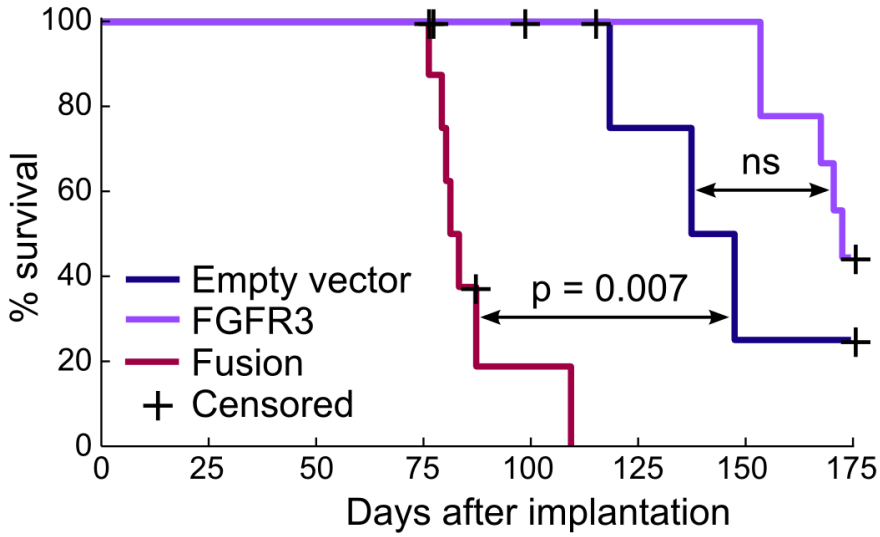

Figure 26. Kaplan-Meier survival plot showing the difference in survival between immunocompromised mice implanted with *FGFR3-TACC3*, wildtype *FGFR3*, and empty vector SNB19 cells. The curves indicate the fraction of uncensored mice alive at specific timepoints after implantation. Five mice died of diarrhea (not due to cancer) and were censored from the study (times of death shown with black crosses).

4.7 *FGFR3-TACC3* escapes microRNA regulation

Previously we noted that the protein level expression of FGFR3 was significantly higher in fusion positive glioblastoma than in fusion negative ones. In fact, the difference was so dramatic that it was not sufficiently explained by the presence of a few extra copies of *FGFR3*. To understand this phenomenon, we set out to determine whether the tandem duplication was altering the post-transcriptional regulation of *FGFR3* transcripts. We realized that even though the fusion gene's expression was driven by the *FGFR3* promoter, the fusion transcript lacked the 3'-UTR of *FGFR3*. This led us to look at regulation by microRNAs, a class of small RNA that bind to the 3'-UTRs of specific genes and cause transcript degration or inhibition of protein translation (reviewed in Sun et al. 2010). Using the TargetScan microRNA target prediction database (Lewis et al. 2005), we determined that the *FGFR3* 3'-UTR was targeted by eight different miRNA families. We then used our small RNA sequencing data to calculate the expression levels of all human microRNAs in our sample pools. One of the *FGFR3*-targeting microRNAs, miR-99a, was found to be the most highly expressed microRNA in the glioblastoma pools, and the fourth highest expressed microRNA in normal brain tissue (Figure 27).

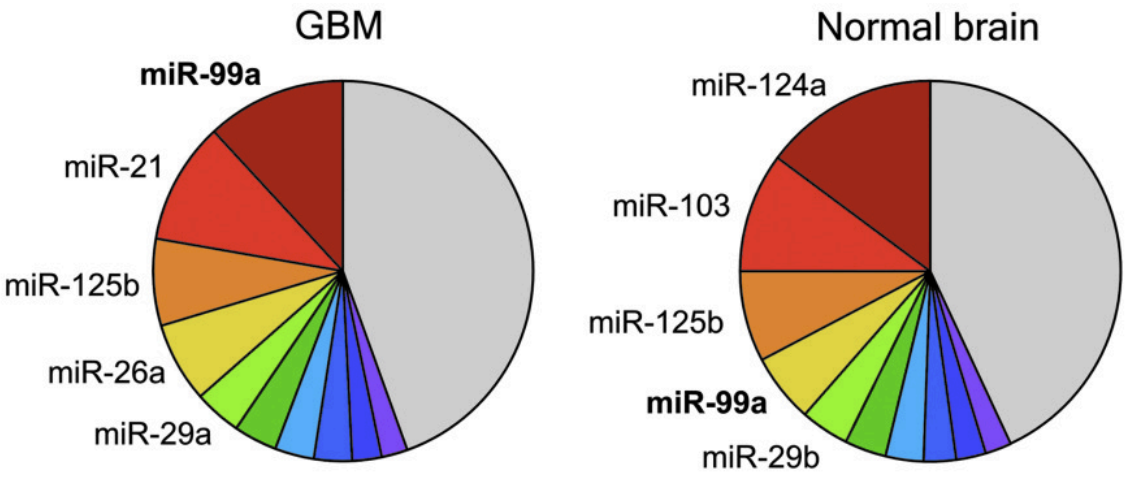

Figure 27. Pie charts showing the fraction of total small RNA sequencing reads arising from different microRNAs. Reads were normalized by the total number of aligned reads. Only the top 10 microRNA are shown, and the top 5 are labeled. The gray area represents the rest of the microRNAs.

To determine whether the loss of miR-99a regulation might have a significant effect on *FGFR3-TACC3* protein levels, we immunoblotted three glioblastoma cell lines and three bladder cancer cell lines for *FGFR3*, then performed qRT-PCR to determine the expression level of miR-99a in the same cell lines. *FGFR3* expression is known to be high in bladder cancer but relatively low in brain cancer. Through the combined immunoblot and qRT-PCR analysis, we could show a clear inverse correlation between *FGFR3* and miR-99a expression (Figure 28).

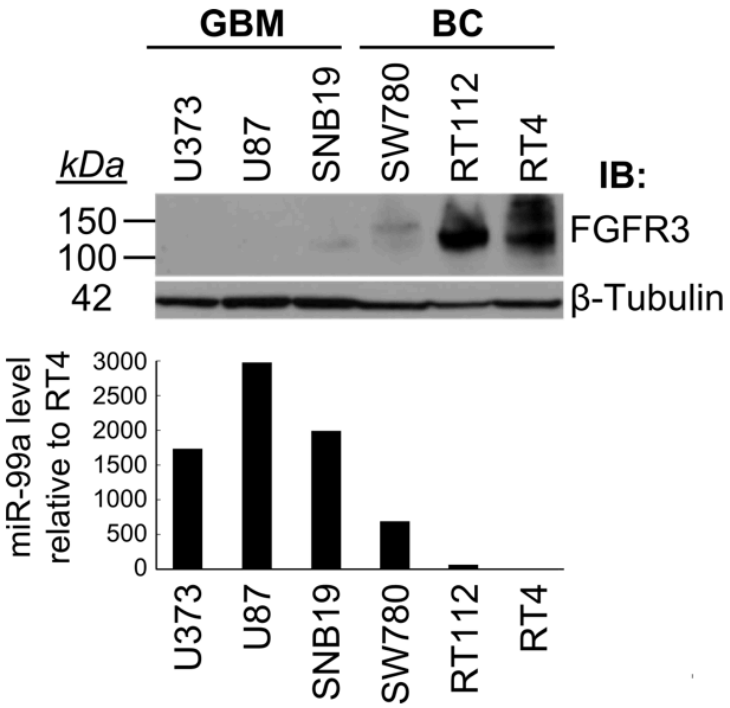

Figure 28. Combined results of an FGFR3 immunoblot and a miR-99a qRT-PCR experiment. The data suggests an inverse correlation between FGFR3 and miR-99a levels.

To show the effect of miR-99a on FGFR3 protein levels more directly, we constructed a vector containing wild-type *FGFR3*, then mutated the vector to remove the miR-99a binding site from the 3'-UTR of *FGFR3*. The binding site removal was achieved by deleting six bases from the 3'-UTR (Figure 29). We then constructed two luciferase reporter assays, one with the wildtype *FGFR3* 3'-UTR, and one with the mutated 3'-UTR. Upon transfection of miR-99a, the luciferase assay with the wildtype 3'-UTR reported a strong decrease in luciferase activity, while the assay with the mutated 3'-UTR was unaffected. This shows that miR-99a binding to the *FGFR3* 3'-UTR strongly inhibits protein translation.

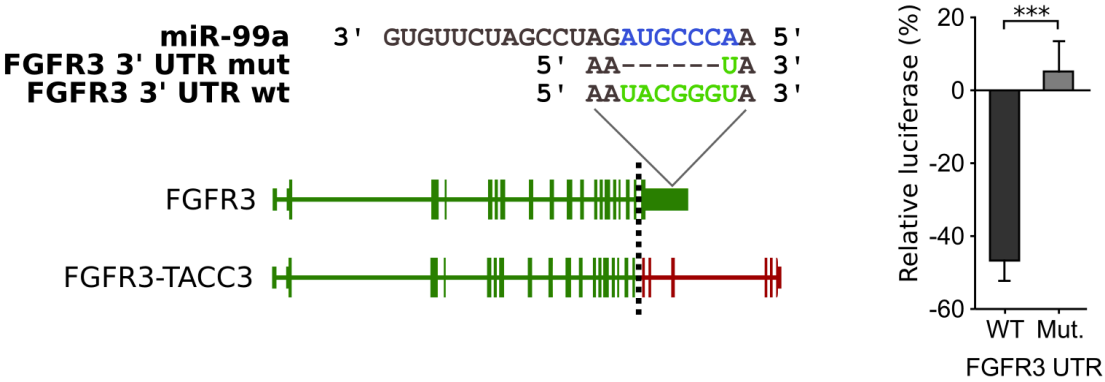

Figure 29. On the left, a diagram showing the mutated and wildtype *FGFR3* 3'-UTRs, and the sequence of miR-99a. The miR-99a binding site was deleted from the mutant through a six base deletion in the 3'-UTR. On the right, a luciferase reporter assay shows that translation inhibiting effect of miR-99a on *FGFR3* is dependent on its binding site in the 3'-UTR.

4.8 Search for *FGFR3-TACC3* in TCGA samples

Near the end of our project, the Cancer Genome Atlas (TCGA) glioblastoma working group released a new dataset containing whole transcriptome sequencing data for 154 glioblastoma patients. We used this opportunity to search for *FGFR3-TACC3* fusions in this new dataset. By running our fusion gene discovery algorithm on the 154 samples, we initially identified 22 samples positive for the fusion. Other members of the TCGA working group also identified dozens of *FGFR3-TACC3* fusions among the 154 samples. However, we noticed that 21 samples shared an identical fusion structure (e18-e11), an observation at odds with the heterogenous nature of the fusion in our cohort. This raised our suspicion, and we decided to look for artifacts in the data. We noted that when the samples were ordered according to the numerical sample identifiers, all but one of the fusion positive cases clustered together. Subsequent analysis revealed that the 154 samples had been sequenced in four batches. In the second batch, nearly all samples showed weak evidence for the fusion gene, while one sample showed stronger evidence. All samples in batch #2 also shared the fusion structure e18-e11. This strongly suggested that inter-sample contamination had occurred in batch #2, so that a low amount of nucleotide material from the true fusion positive sample had contaminated the other samples (Figure 30). This theory was further supported by the observation that overexpression of *FGFR3* and *TACC3* was only observed in the two true fusion positive samples (Figure 30). We therefore concluded that the TCGA cohort actually contained only two fusion positive cases. After reporting the issue to TCGA, the laboratory responsible for batch #2 reported that they had discovered the root of the problem in one of the sample preparation protocols they used. The number of *FGFR3-TACC3* positive patients in the TCGA manuscript was downgraded to two.

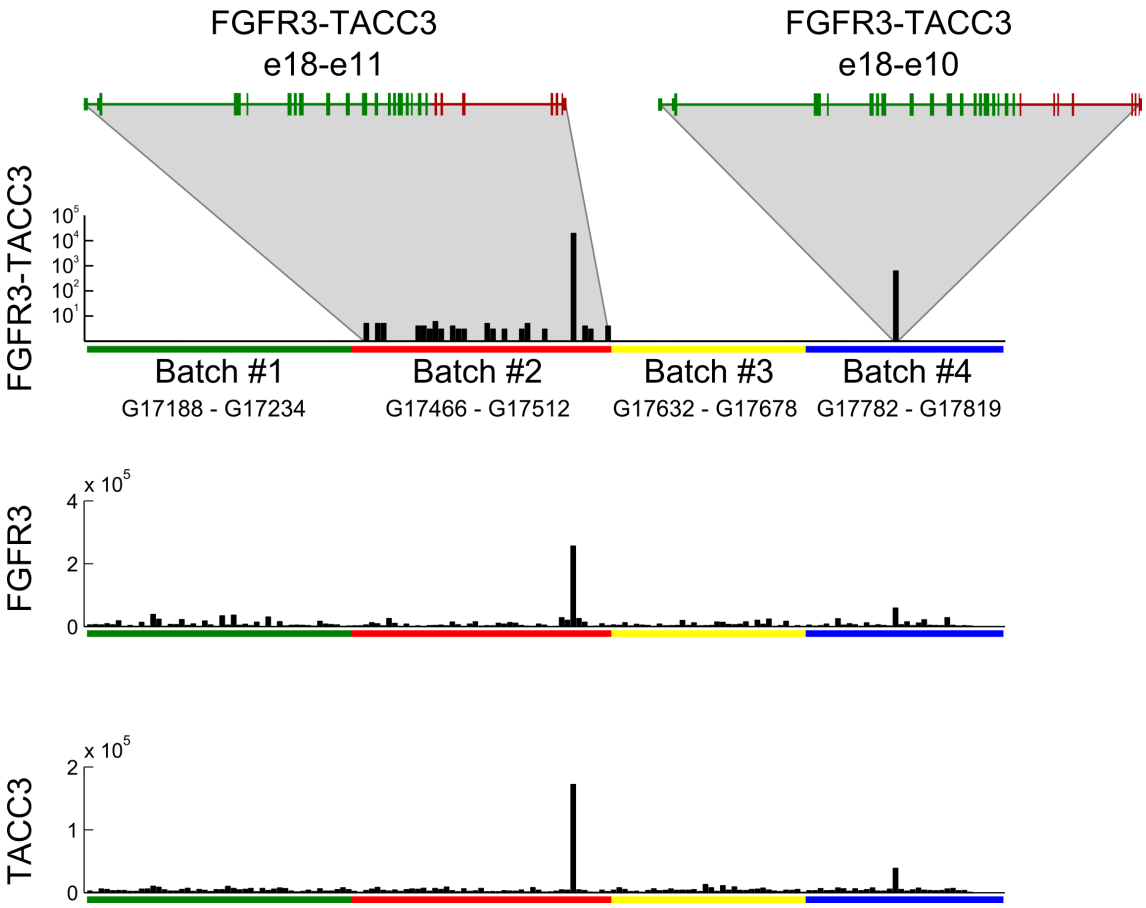

Figure 30. Transcriptome sequencing data from the Cancer Genome Atlas GBM project reveals two fusion positive tumors, and inter-sample contamination in batch #2. Overexpression of *FGFR3* or *TACC3* is not observed in contaminated samples. Structures of the fusion genes are shown at the top.

4.9 Other fusion genes

In addition to the *FGFR3-TACC3* fusion, we also studied two other fusion genes identified by our transcriptome sequencing. A *ZNF713-VSTM2A* fusion was validated in one glioblastoma tumor, but did not occur in other tumors. Analysis of the locations of the two genes revealed that both genes were located in chromosome 7, flanking the *EGFR* gene. *EGFR* encodes for an epidermal growth factor receptor and is a well-known oncogene amplified in 70% of glioblastomas. We hypothesized that this fusion gene was a side effect of *EGFR* amplification, caused by a tandem duplication whose boundaries happened to overlap the two genes. This hypothesis was later supported by evidence from the Cancer Genome Atlas glioblastoma sequencing project, where we discovered numerous non-recurrent fusion genes involving pairs of genes flanking the *EGFR* locus.

The third fusion, *NUP188-SPTAN1*, was also validated by RT-PCR in one glioblastoma tumor, but did not occur in other tumors. We did not pursue this fusion gene further as we could not show it was anything more than a one-time event.

5 CONCLUSIONS

In this thesis, we have shown how the combination of high throughput measurements and computational algorithms can yield novel biological insights into complex diseases such as human cancers. The identification of the first recurrent fusion gene ever reported in glioblastoma provides new hope for a treatment through targeted molecular therapy. The availability of multiple small molecule inhibitors of FGFR (Pardo et al. 2009; Lamont et al. 2011; Gavine et al. 2012; Gozgit et al. 2012) suggests that some of these molecules might be successfully adapted for clinical use. Our collaborators at the M.D. Anderson Cancer Center are currently studying the effect that different FGFR3 inhibitors have on cells transfected with *FGFR3-TACC3*. The fusion gene also has the potential to act as a prognostic marker that could be used in a clinical setting to determine the types of treatments most effective for an individual patient.

After the initial publication of *FGFR3-TACC3* in glioblastoma by us and Singh et al. (2012), Williams et al. (2012) reported their discovery of *FGFR3-TACC3* fusions in bladder cancer. They also reported a novel *FGFR3-BAIAP2L1* fusion gene that is caused by interchromosomal translocation and involves a different 3' partner. Intriguingly, both *TACC3* and *BAIAP2L1* contain a coiled coil domain in their 3' end. This may imply that the oncogenicity of the *FGFR3-** fusion genes may result from the introduction of oligomerization domains to *FGFR3*. Coiled coil domains are known to function as oligomerization domains in many human proteins. The constitutive activation of an oncogene through fusion-induced oligomerization is a well-characterized phenomenon known to occur in multiple human cancers (Davis et al. 1985; Chiarle et al. 2008). This theory is further supported by the fact that many bladder cancers harbor *FGFR3* mutations that are known to cause constitutive oligomerization of FGFR3 (Cappellen et al. 1999).

Recently, another group reported their discovery of FGFR fusions in over 10 different cancer types (Wu et al. 2013). This group also noted that all fusion genes involved an FGFR kinase domain fused with an oligomerization domain from another gene. This implies that *FGFR3-TACC3* fusions represent the most widely distributedone of the few fusions known to occur in multiple cancer types. The development of targeted molecular therapies for patients with *FGFR3-TACC3* positive cancer could therefore provide broad-spectrum relief in the struggle against cancer.

REFERENCES

Akiva, P., Toporik, A., Edelheit, S., Peretz, Y., Diber, A., Shemesh, R., Novik, A. & Sorek, R. 2006. Transcription-mediated gene fusion in the human genome. Genome Research 16, 1, pp. 30-36.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25, 17, pp. 3389-3402.

Annala, M.J., Parker, B.C., Zhang, W. & Nykter, M. 2013. Fusion genes and their discovery using high throughput sequencing. Cancer Letters (in press).

Antonescu, C.R., Tschernyavsky, S.J., Decuseara, R., Leung, D.H., Woodruff, J.M., Brennan, M.F., Bridge, J.A., Neff, J.R., Goldblum, J.R. & Ladanyi, M. 2001. Prognostic impact of P53 status, TLS-CHOP fusion transcript structure, and histological grade in myxoid liposarcoma: a molecular and clinicopathologic study of 82 cases. Clinical Cancer Research 7, 12, pp. 3977-3987.

Aurias, A., Rimbaut, C., Buffe, D., Dubousset, J. & Mazabraud, A. 1983. Chromosomal translocations in Ewing's sarcoma. New England Journal of Medicine 309, pp. 496-497.

Avery, O.T., MacLeod, C.M. & McCarty, M. 1944. Studies on the chemical nature of the substance inducing transformation of pneumococcal types. The Journal of Experimental Medicine 79, 2, pp. 137-158.

Beroukhim, R., Mermel, C.H., Porter, D., Wei, G., Raychaudhuri, S., Donovan, J., Barretina, J., Boehm, J.S., Dobson, J., Urashima, M., McHenry K.T., Pinchback, R.M., Ligon, A.H., Cho, Y.J., Haery, L., Greulich, H., Reich, M., Winckler, W., Lawrence, M.S., Weir, B.A., Tanaka, K.E., Chiang, D.Y., Bass, A.J., Loo, A., Hoffman, C., Prensjer, J., Liefeld, T., Gao, Q., Yecies, D., Signoretti, S., Maher, E., Kaye, F.J., Sasaki, H., Tepper, J.E., Fletcher, J.A., Tabernero, J., Baselga, J., Tsao, M.S., Demichelis, F., Rubin, M.A., Jänne, P.A., Daly, M.J., Nucera, C., Levine, R.L., Ebert, B.L., Gabriel, S., Rustgi, A.K., Antonescu, C.R., Ladanyi, M., Letai, A., Garraway, L.A., Loda, M., Beer, D.G., True, L.D., Okamoto, A., Pomeroy, S.L., Singer, S., Golub, T.R., Lander, E.S., Getz, G., Sellers, W.R. & Meyerson, M. 2010. The landscape of somatic copy-number alteration across human cancers. Nature 463, 7283, pp. 899-905.

Blat, Y. & Kleckner, N. 1999. Cohesins bind to preferential sites along yeast chromosome III, with differential regulation along arms versus the centric region. Cell 98, 2, pp. 249-259.

Borrow, J., Goddard, A.D., Sheer, D. & Solomon, E. 1990. Molecular analysis of acute promyelocytic leukemia breakpoint cluster region on chromosome 17. Science 249, 4976, pp. 1577-1580.

Breu, H. 2010. A theoretical understanding of 2 base color codes and its application to annotation, error detection, and error correction. [WWW]. [Cited 17/4/2013]. Available at: http://www3.appliedbiosystems.com/cms/groups/mcb_marketing/documents/generaldocuments/cms_058265.pdf (in English).

Burnette, W.N. 1981. "Western blotting": electrophoretic transfer of proteins from sodium dodecyl sulfate-polyacrylamide gels to unmodified nitrocellulose and radiographic detection with antibody and radioiodinated protein A. Analytical Biochemistry 112, 2, pp. 195-203.

Cancer Genome Atlas Research Network. 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455, 7216, pp. 1061-1068.

Cappellen, D., de Oliveira, C., Ricol, D., de Medina, S., Bourdin, J., Sastre-Garau, X., Chopin, D., Thiery, J.P. & Radvanyi, F. 1999. Frequent activating mutations of FGFR3 in human bladder and cervix carcinomas. Nature Genetics 23, 1, pp. 18-20.

Cech, T.R. 2000. The ribosome is a ribozyme. Science 289, 5481, pp. 878-879.

Chang, Y.F., Imam, J.S. & Wilkinson, M.F. 2007. The nonsense-mediated decay RNA surveillance pathway. Annual Reviews of Biochemistry 76, pp. 51-74.

Chase, A., Ernst, T., Fiebig, A., Collins, A., Grand, F., Erben, P., Reiter, A., Schreiber, S. & Cross, N.C. TFG, a target of chromosome translocations in lymphoma and soft tissue tumors, fuses to GPR128 in healthy individuals. Haematologica 95, 1, pp. 20-26.

Chen, K., Wallis, J.W., McLellan, M.D., Larson, D.E., Kalicki, J.M., Pohl, C.S., McGrath, S.D., Wendl, M.C., Zhang, Q., Locke, D.P., Shi, X., Fulton, R.S., Ley, T.J., Wilson, R.K., Ding, L. & Mardis, E.R. 2009. BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nature Methods 6**,** 9, pp. 677-681.

Chiarle, R., Voena, C., Ambrogio, C., Piva, R. & Inghirami, G. 2008. The anaplastic lymphoma kinase in the pathogenesis of cancer. Nature Reviews Cancer 8, 1, pp. 11-23.

Church, D.M., Schneider, V.A., Graves, T., Auger, K., Cunningham, F., Bouk, N., Chen, H-C., Agarwala, R., McLaren, W.M., Ritchie, G.R.S., Albracht, D., Kremitzki, M., Rock, S., Kotkiewicz, H., Kremitzki, C., Wollam, A., Trani, L., Fulton, L., Fulton, R., Matthews, L., Whitehead, S., Chow, W., Torrance, J., Dunn, M., Harden, G., Threadgold, G., Wood, J., Collins, J., Heath, P., Griffiths, G., Pelan, S., Grafham, D., Eichler, E.E., Weinstock, G., Mardis, E.R., Wilson, R.K., Howe, K., Flicek, P. & Hubbard, T. 2011. Modernizing reference genome assemblies. PLoS Biology 9, 7, e1001091.

Ciampi, R., Knauf, J.A., Kerler, R., Gandhi, M., Zhu, Z., Nikiforova, M.N., Rabes, H.M., Fagin, J.A. & Nikiforov, Y.E. 2005. Oncogenic AKAP9-BRAF fusion is a novel mechanism of MAPK pathway activation in thyroid cancer. Journal of Clinical Investigation 115, 1, pp. 94-101.

Clancy, S. 2008. RNA splicing: introns, exons, and spliceosome. Nature Education 1, 1.

Clark, J., Rocques, P.J., Crew, A.J., Gill, S., Shipley, J., Chan, A.M-L., Gusterson, B.A. & Cooper, C.S. 1994. Identification of novel genes, SYT and SSX, involved in the t(X;18)(p11.2;q11.2) translocation found in human synovial sarcoma. Nature Genetics 7, 4, pp. 502-508.

Cornish-Bowden, A. 1985. Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. Nucleic Acids Research 13, 9, pp. 3021-3030.

Croce, C.M. 1986. Chromosome translocations and human cancer. Cancer Research 46, pp. 6019-6023.

Crozat, A., Aman, P., Mandahl, N. & Ron, D. 1993. Fusion of CHOP to a novel RNA-binding protein in human myxoid liposarcoma. Nature 363, 6430, pp. 640-644.

Daley, G.Q., Van Etten, R.A. & Baltimore, D. 1990. Induction of chronic myelogenous leukemia in mice by the P210 bcr/abl gene of the Philadelphia chromosome. Science 247**,** 4944, pp. 824-830.

Dalla-Favera, R., Bregni, M., Erikson, J., Patterson, D., Gallo, R.C. & Croce, C.M. 1982. Human c-myc onc gene is located on the region of chromosome 8 that is translocated in Burkitt's lumphoma cells. Proceedings of the National Academy of Sciences of the USA 79, 24, pp. 7824-7827.

Davis, R.L., Konopka, J.B. & Witte, O.N. 1985. Activation of the c-abl oncogene by viral transduction or chromosomal translocation generates altered c-abl proteins with similar in vitro kinase properties. Molecular Cell Biology 5, 1, pp. 204-213.

Druker, B.J., Tamura, S., Buchdunger, E., Ohno, S., Segal, G.M., Fanning, S., Zimmermann, J. & Lydon, N.B. 1996. Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr-Abl positive cells. Nature Medicine 2, 5, pp. 561-566.

Druker, B.J., Guilhot, F., O'Brien, S.G., Gathmann, I., Kantarjian, H., Gattermann, N., Deininger, M.W., Silver, R.T., Goldman, J.M., Stone, R.M., Cervantes, F., Hochhaus, A., Powell, B.L., Gabrilove, J.L., Rousselot, P., Reiffers, J., Cornelissen, J.J., Hughes, T., Agis, H., Fischer, T., Verhoef, G., Shepherd, J., Saglio, G., Gratwohl, A., Nielsen, J.L., Radich, J.P., Simonsson, B., Taylor, K., Baccarani, M., So, C., Letvak, L., Larson, R.A. & IRIS investigators. 2006. Five-year followup of patients receiving imatinib for chronic myeloid leukemia. The New England Journal of Medicine 355, 23, pp. 2408-2417.

Duncan, C.G., Killela, P.J., Payne, C.A., Lampson, B., Chen, W.C., Liu, J., Solomon, D., Waldman, T., Towers, A.J., Gregory, S.G., McDonald, K.L., McLendon, R.E., Bigner, D.D. & Yan, H. 2010. Integrated genomic analyses identify ERRFI1 and TACC3 as glioblastoma-targeted genes. Oncotarget 1, 4, pp. 265-277.

ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489**,** 7414, pp. 57-74.

Erickson, P., Gao, J., Chang, K.S., Look, T., Whisenant, E., Raimondi, S., Lasher, R., Trujillo, J., Rowley, J. & Drabkin, H. 1992. Identification of breakpoints in t(8;21) acute myelogenous leukemia and isolation of a fusion transcript, AML1/ETO, with similarity to Drosophila segmentation gene, runt. Blood 80, 7, pp. 1825-1831.

Eswarakumar, V.P., Lax, I. & Schlessinger, J. 2005. Cellular signaling by fibroblast growth factor receptors. Cytokine & Growth Factor Reviews 16, 2, pp. 139-149.

Ferlay, J., Shin, H.R., Bray, F., Forman, D., Mathers, C. & Parkin, D.M. 2008. GLOBOCAN 2008 v2.0, cancer incidence and mortality worldwide: IARC CancerBase No. 10. [WWW]. [Cited 17/4/2013]. Available at: http://globocan.iarc.fr. (in English).

Fullwood, M.J., Wei, C-L., Liu, E.T. & Ruan, Y. 2009. Next-generation DNA sequencing of paired-end tags (PET) for transcriptome and genome analyses. Genome Research 19, 4, pp. 521-532.

Furnari, F.B., Fenton, T., Bachoo, R.M., Mukasa, A., Stommel, J.M., Stegh, A., Hahn, W.C., Ligon, K.L., Louis, D.N., Brennan, C., Chin, L., DePinho, R.A. & Cavenee, W.K. 2007. Malignant astrocytic glioma: genetics, biology, and paths to treatment. Genes & Development 21, 21, pp. 2683-2710.

Gavine, P.R., Mooney, L., Kilgour, E., Thomas, A.P., Al-Kadhimi, K., Beck, S., Rooney, C., Coleman, T., Baker, D., Mellor, M.J., Brooks, A.N. & Klinowska, T. 2012. AZD4547: an orally bioavailable, potent, and selective inhibitor of the fibroblast growth factor receptor tyrosine kinase family. Cancer Research 72, 8, pp. 2045-2056.

Gergely, F., Karlsson, C., Still, I., Cowell, J., Kilmartin, J. & Raff, J.W. 2000. The TACC domain identifies a family of centrosomal proteins that can interact with microtubules. Proceedings of the National Academy of Sciences of the USA 97, 26, pp. 14352-14357.

Gitan, R.S., Shi, H., Chen, C.M., Yan, P.S. & Huang, T.H.M. 2002. Methylation-specific oligonucleotide microarray: a new potential for high-throughput methylation analysis. Genome Research 12, 1, pp. 158-164.

Gozgit, J.M., Wong, M.J., Moran, L., Wardwell, S., Mohemmad, Q.K., Narasimhan, N.I., Shakespeare, W.C., Wang, F., Clackson, T. & Rivera, V.M. 2012. Ponatinib (AP24534), a multitargeted pan-FGFR inhibitor with activity in multiple FGFR-amplified or mutated cancer models. Molecular Cancer Therapy 11, 3, pp. 690-699.

Griffiths-Jones, S. 2004. The microRNA registry. Nucleic Acids Research 32 (database issue), D109-D111.

Hawkins, R.D., Hon, G.C. & Ren, B. 2010. Next-generation genomics: an integrative approach. Nature Reviews Genetics 11, 7, pp. 476-486.

Helleday, T., Petermann, E., Lundin, C., Hodgson, B. & Sharma, R.A. 2008. DNA repair pathways as targets for cancer therapy. Nature Reviews Cancer 8, 3, pp. 193-204.

Houseley, J. & Tollervey, D. 2010. Apparent non-canonical trans-splicing is generated by reverse transcriptase in vitro. PLoS ONE 5, e12271.

Hughes, T., Deininger, M., Hochhaus, A., Branford, S., Radich, J., Kaeda, J., Baccarani, M., Cortes, J., Cross, N.C.P., Druker, B.J., Gabert, J., Grimwade, D., Hehlmann, R., Kamel-Reid, S., Lipton, J.H., Longtine, J., Martinelli, G., Saglio, G., Soverini, S., Stock, W. & Goldman, J.M. 2006. Monitoring CML patients responding to treatment with tyrosine kinase inhibitors: review and recommendations for harmonizing current methodology for detecting BCR-ABL transcripts and kinase domain mutations and for expressing results. Blood 108, 1, pp. 28-37.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. Nature 431, 7011, pp. 931-945.

Jemal, A., Bray, F., Center, M.M., Ferlay, J., Ward, E. & Forman, D. 2011. Global cancer statistics. CA: A Cancer Journal for Clinicians 61, 2, pp. 69-90.

Jones, D.T.W., Kocialkowski, S., Liu, L., Pearson, D.M., Bäcklund, L.M., Ichimura, K. & Collins, V.P. 2008. Tandem duplication producing a novel oncogenic BRAF fusion gene defines the majority of pilocytic astrocytomas. Cancer Research 68, 21, pp. 8673-8677.

Kanagawa, T. 2003. Bias and artifacts in multitemplate polymerase chain reactions (PCR). Journal of Bioscience and Bioengineering 96, 4, pp. 317-323.

Kitano, H. 2002. Systems biology: a brief overview. Science 295, 5560, pp. 1662-1664.

Lamont, F.R., Tomlinson, D.C., Cooper, P.A., Shnyder, S.D., Chester, J.D. & Knowles, M.A. 2011. Small molecule FGF receptor inhibitors block FGFR-dependent urothelial carcinoma growth in vitro and in vivo. British Journal of Cancer 104, 1, pp. 75-82.

Langmead, B., Trapnell, C., Pop, M. & Salzberg, S.L. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10, 3, R25.

Laurent, E., Talpaz, M., Kantarjian, H. & Kurzrock, R. 2001. The BCR gene and Philadelphia chromosome-positive leukemogenesis. Cancer Research 61, 6, pp. 2343-2355.

Lawson, A.R., Hindley, G.F., Forshew, T., Tatevossian, R.G., Jamie, G.A., Kelly, G.P., Neale, G.A., Ma, J., Jones, T.A., Ellison, D.W. & Sheer, D. 2011. RAF gene fusion breakpoints in pediatric brain tumors are characterized by significant enrichment of sequence microhomology. Genome Research 21, 4, pp. 505-514.

Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G., Lin, Y., MacDonald, J.R., Pang, A.W., Shago, M., Stockwell, T.B., Tsiamouri, A., Bafna, V., Bansal, V., Kravitz, S.A., Busam, D.A., Beeson, K.Y., McIntosh, T.C., Remington, K.A., Abril, J.F., Gill, J., Borman, J., Rogers, Y.H., Frazier, M.E., Scherer, S.W., Strausberg, R.L. & Venter, J.C. 2007. The diploid genome sequence of an individual human. PLoS Biology 5, 10, e254.

Lewis, B.P., Burge, C.B. & Bartel, D.P. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 120, 1, pp. 15-20.

Lin, C., Yang, L., Tanasa, B., Hutt, K., Ju, B.G., Ohgi, K., Zhang, J., Rose, D.W., Fu, X.D., Glass, C.K. & Rosenfeld, M.G. 2009. Nuclear receptor-induced chromosomal proximity and DNA breaks underlie specific translocations in cancer. Cell 139, 6, pp. 1069-1083.

Lipson, D., Capelletti, M., Yelensky, R., Otto, G., Parker, A., Jarosz, M., Curran, J.A., Balasubramanian, S., Bloom, T., Brennan, K.W., Donahue, A., Downing, S.R., Frampton, G.M., Garcia, L., Juhn, F., Mitchell, K.C., White, E., White, J., Zwirko, Z., Peretz, T., Nechushtan, H., Soussan-Gutman, L., Kim, J., Sasaki, H., Kim, H.R., Park, S.I., Ercan, D., Sheehan, C.E., Ross, J.S., Cronin, M.T., Jänne, P.A. & Stephens, P.J. 2012. Identification of new ALK and RET gene fusions from colorectal and lung cancer biopsies. Nature Medicine 18, 3, pp. 382-384.

Long, M. & Deutsch, M. 1999. Association of intron phases with conservation at splice site sequences and evolution of spliceosomal introns. Molecular Biology and Evolution 16, 11, pp. 1528-1534.

Louis, D.N., Ohgaki, H., Wiestler, O.D., Cavenee, W.K., Burger, P.C., Jouvet, A., Scheithauer, B.W. & Kleihues, P. 2007. The 2007 WHO classification of tumours of the central nervous system. Acta Neuropathology 114, 2, pp. 97-109.

Maher, C.A., Kumar-Sinha, C., Cao, X., Kalayna-Sundaram, S., Han, B., Jing, X., Sam, L., Barrette, T., Palanisamy, N. & Chinnaiyan, A.M. 2009. Transcriptome sequencing to detect gene fusions in cancer. Nature 458, 7234, pp. 97-101.

Maher, C.A., Palanisamy, N., Brenner, J.C., Cao, X., Kalyana-Sundaram, S., Luo, S., Khrebtukova, I., Barrette, T.R., Grasso, C., Yu, J., Lonigro, R.J., Schroth, G., Kumar-Sinha, C. & Chinnaiyan, A.M. 2009. Chimeric transcript discovery by paired-end transcriptome sequencing. Proceedings of the National Academy of Sciences of the USA 106, 30, pp. 12353-12358.

Manolov, G. & Manolova, Y. 1972. Marker band in one chromosome 14 from Burkitt lymphomas. Nature 237, 5349, pp. 33-34.

Mardis, E.R. 2008. Next-generation DNA sequencing methods. Annual Review of Genomics and Human Genetics 9, pp. 387-402.

Martinelli, G., Amabile, M., Giannini, B., Terragna, C., Ottaviani, E., Soverini, S., Saglio, G., Rosti, G. & Baccarani, M. 2002. Novel types of bcr-abl transcript with breakpoints in BCR exon 8 found in Philadelphia positive patients with typical chronic myeloid leukemia retain the sequence encoding for the DBL- and CDC24 homology domains but not the pleckstrin homology one. Haematologica 87, 7, pp. 688-694.

McKernan, K.J., Peckham, H.E., Costa, G.L., McLaughlin, S.F., Fu, Y., Tsung, E.F., Clouser, C.R., Duncan, C., Ichikawa, J.K., Lee, C.C., Zhang, Z., Ranade, S.S., Dimalanta, E.T., Hyland, F.C., Sokolsky, T.D., Zhang, L., Sheridan, A., Fu, H., Hendrickson, C.L., Li, B., Kotler, L., Stuart, J.R., Malek, J.A., Manning, J.M., Antipova, A.A., Perez, D.S., Moore, M.P., Hayashibara, K.C., Lyons, M.R., Beaudoin, R.E., Coleman, B.E., Laptewicz, M.W., Sannicandro, A.E., Rhodes, M.D., Gottimukkala, R.K., Yang, S., Bafna, V., Bashir, A., MacBride, A., Alkan, C., Kidd, J.M., Eichler, E.E., Reese, M.G., De La Vega, F.M. & Blanchard, A.P. 2009. Sequence and structural variation in a human genome uncovered by short-read, massively parallel ligation sequencing using two-base encoding. Genome Research 19, 9, pp. 1527-1541.

Mei, R., Galipeau, P.C., Prass, C., Berno, A., Ghandour, G., Patil, N., Wolff, R.K., Chee, M.S., Reid, B.J. & Lockhart, D.J. 2000. Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays. Genome Research 10, 8, pp. 1125-1137.

Mitelman, F., Johansson, B. & Mertens, F. 2007. The impact of translocations and gene fusions on cancer causation. Nature Reviews Cancer 7, 4, pp. 233-245.

Mortazavi, A., Williams, B.A., McCue, K., Schaeffer, L. & Wold, B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods 5, 7, pp. 621-628.

Myers, J.C., Spiegelman, S. & Kacian, D.L. 1977. Synthesis of full-length DNA copies of avian myeloblastosis virus RNA in high yields. Proceedings of the National Academy of Sciences of the USA 74, 7, pp. 2840-2843.

Nowell, P.C. & Hungerford, D.A. 1960. A minute chromosome in human chronic granulocytic leukemia. Science 132, pp. 1497-1500.

Pardo, O.E., Latigo, J., Jeffery, R.E., Nye, E., Poulson, R., Spencer-Dene, B., Lemoine, N.R., Stamp, G.W., Aboagye, E.O. & Seckl, M.J. 2009. Fibroblast growth factor receptor inhibitor PD173074 blocks small cell lung cancer growth in vitro and in vivo. Cancer Research 69, 22, pp. 8645-8651.

Parker, B.C., Annala, M.J., Cogdell, D.E., Granberg, K.J., Sun, Y., Ji, P., Li, X., Gumin, J., Zheng, H., Hu, L., Yli-Harja, O., Haapasalo, H., Visakorpi, T., Liu, X., Liu, C-G., Sawaya, R., Fuller, G.N., Chen, K., Lang, F.F., Nykter, M. & Zhang, W. 2012. The tumorigenic FGFR3-TACC3 gene fusion escapes miR-99a regulation in glioblastoma. Journal of Clinical Investigation 123, 2, pp. 855-865.

Perner, S., Demichelis, F., Beroukhim, R., Schmidt, F.H., Mosquera, J-M., Setlur, S., Tchinda, J., Tomlins, S.A., Hofer, M.D., Pienta, K.G., Kuefer, R., Vessella, R., Sun, X-W., Meyerson, M., Lee, C., Sellers, W.R., Chinnaiyan, A.M. & Rubin, M.A. 2006. TMPRSS2:ERG fusion-associated deletions provide insight into the heterogeneity of prostate cancer. Cancer Research 66, 17, pp. 8337-8341.

Persson, M., Andren, Y., Mark, J., Horlings, H.M., Persson, F. & Stenman, G. 2009. Recurrent fusion of MYB and NFIB transcription factor genes in carcinomas of the breast and head and neck. Proceedings of the National Academy of Sciences of the USA 106, 44, pp. 18740-18744.

Pinkel, D., Segraves, R., Sudar, D., Clark, S., Poole, I., Kowbel, D., Collins, C., Kuo, W.L., Chen, C., Zhai, Y., Dairkee, S.H., Ljung, B.M., Gray, J.W. & Albertson, D.G. 1998. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nature Genetics 20**,** 2, pp. 207-211.

Pleasance, E.D., Stephens, P.J., O'Meara, S., McBride, D.J., Meynert, A., Jones, D., Lin, M.L., Beare, D., Lau, K.W., Greenman, C., Varela, I., Nik-Zainal, S., Davies, H.R., Ordonez, G.R., Mudie, L.J., Latimer, C., Edkins, S., Stebbings, L., Chen, L., Jia, M., Leroy, C., Marshall, J., Menzies, A., Butler, A., Teague, J.W., Mangion, J., Sun, Y.A., McLaughlin, S.F., Peckham, H.E., Tsung, E.F., Costa, G.L., Lee, C.C., Minna, J.D., Gazdar, A., Birney, E., Rhodes, M.D., McKernan, K.J., Stratton, M.R., Futreal, P.A. & Campbell, P.J. A small-cell lung cancer genome with complex signatures of tobacco exposure. Nature 463, 7278, pp. 184-190.

Rabbitts, T.H., Forster, A., Larson, R. & Nathan, P. 1993. Fusion of the dominant negative transcription regulator CHOP with a novel gene FUS by translocation t(12;16) in malignant liposarcoma. Nature Genetics 4, 2, pp. 175-180.

Rikova, K., Guo, A., Zeng, Q., Possemato, A., Yu, J., Haack, H., Nardone, J., Lee, K., Reeves, C., Li, Y., Hu, Y., Tan, Z., Stokes, M., Sullivan, L., Mitchell, J., Wetzel, R., Macneill, J., Ren, J.M., Yuan, J., Bakalarski, C.E., Villen, J., Kornhauser, J.M., Smith, B., Li, D., Zhou, X., Gygi, S.P., Gu, T.L., Polakiewicz, R.D., Rush, J. & Comb, M.J. 2007. Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. Cell 131, 6, pp. 1190-1203.

Rowley, J.D. 1973. A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and giemsa staining. Nature 243, pp. 290-293.

Ruvinsky, A., Eskesen, S.T., Eskesen, F.N. & Hurst, L.D. 2005. Can codon usage bias explain intron phase distributions and exon symmetry? Journal of Molecular Evolution 60, 1, pp. 99-104.

Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B. & Erlich, H.A. 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 239, 4839, pp. 487-491.

Salk, J., Fox, E. & Loeb, L. 2010. Mutational heterogeneity in human cancers: origin and consequences. Annual Review of Pathology: Mechanisms of Disease 5, pp. 51-75.

Schatz, D.G. & Swanson, P.C. 2011. V(D)J recombination: mechanisms of initiation. Annual Review of Genetics 45, pp. 167-202.

Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270**,** 5235, pp. 467-470.

Seshagiri, S., Stawiski, E.W., Durinck, S., Modrusan, Z., Storm, E.E., Conboy, C.B., Chaudhuri, S., Guan, Y., Janakiraman, V., Jaiswal, B.S., Guillory, J., Ha, C., Dijkgraaf, G.J., Stinson, J., Gnad, F., Huntley, M.A., Degenhardt, J.D., Haverty, P.M., Bourgon, R., Wang, W., Koeppen, H., Gentleman, R., Starr, T.K., Zhang, Z., Largaespada, D.A., Wu, T.D. & de Sauvage, F.J. 2012. Recurrent R-spondin fusions in colon cancer. Nature 488, 7413, pp. 660-664.

Shaw, A.T., Yeap, B.Y., Solomon, B.J., Riely, G.J., Gainor, J., Engelman, J.A., Shapiro, G.I., Costa, D.B., Ou, S.H., Butaney, M., Salgia, R., Maki, R.G., Varella-Garcia, M., Doebele, R.C., Bang, Y.J., Kulig, K., Selaru, P., Tang, Y., Wilner, K.D., Kwak, E.L., Clark, J.W., Iafrate, A.J. & Camidge, D.R. 2011. Effect of crizotinib on overall survival in patients with advanced non-small-cell lung cancer harbouring ALK gene rearrangement: a retrospective analysis. Lancet Oncology 12, 11, pp. 1004-1012.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M. & Sirotkin, K. 2001. dbSNP: the NCBI database of genetic variation. Nucleic Acids Research 29, 1, pp. 308-311.

Shtivelman, E., Lifshitz, B., Gale, R.P. & Canaani, E. 1985. Fused transcript of abl and bcr genes in chronic myelogenous leukaemia. Nature 315, pp. 550-554.

Smith, S., Reeves, B.R., Wong, L. & Fisher, C. 1987. A consistent chromosome translocation in synovial sarcoma. Cancer Genetics and Cytogenetics 26, 1, pp. 179-180.

Soda, M., Choi, Y.L., Enomoto, M., Takada, S., Yamashita, Y., Ishikawa, S., Fujiwara, S., Watanabe, H., Kurashina, K., Hatanaka, H., Bando, M., Ohno, S., Ishikawa, Y., Aburatani, H., Niki, T., Sohara, Y., Sugiyama, Y. & Mano, H. 2007. Identification of the transforming EML4-ALK fusion gene in non-small-cell lung cancer. Nature 448, 7153, pp. 561-566.

Solinas-Toldo, S., Lampel, S., Stilgenbauer, S., Nickolenko, J., Benner, A., Dohner, H., Cremer, T. & Licher, P. 1997. Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. Genes, chromosomes and cancer 20, 4, pp. 399-407.

Still, I.H., Vince, P. & Cowell, J.K. 1999. The third member of the transforming acidic coiled coil-containing gene family, TACC3, maps in 4p16, close to translocation breakpoints in multiple myeloma, and is upregulated in various cancer cell lines. Genomics 58, 2, pp. 165-170.

Stupp, R., Mason, W.P., van den Bent, M.J., Weller, M., Fisher, B., Taphoorn, M.J., Belanger, K., Brandes, A.A., Marosi, C., Bogdahn, U., Curschmann, J., Janzer, R.C., Ludwin, S.K., Gorlia, T., Allgeier, A., Lacombe, D., Cairncross, J.G., Eisenhauer, E., Mirimanoff, R.O., European Organisation for Research and Treatment of Cancer Brain Tumor and Radiotherapy Groups & National Cancer Institute of Canada Clinical Trials Group. 2005. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. New England Journal of Medicine 352, 10, pp. 987-996.

Sun, W., Li, Y-S.J., Huang, H-D., Shyy, J.Y-J. & Chien, S. 2010. MicroRNA: a master regulator of cellular processes for bioengineering systems. The Annual Review of Biomedical Engineering 12, pp. 1-27.

Sverdlov, A.V., Rogozin, I.B., Babenko, V.N. & Koonin, E.V. 2003. Evidence of splice signal migration from exon to intron during intron evolution. Current Biology 13, 24, pp. 2170-2174.

Tomlins, S.A., Rhodes, D.R., Perner, S., Dhanasekaran, S.M., Mehra, R., Sun, X.W., Varambally, S., Cao, X., Tchinda, J., Kuefer, R., Lee, C., Montie, J.E., Shah, R.B., Pienta, K.J., Rubin, M.A. & Chinnaiyan, A.M. 2005. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science 310, 5748, pp. 644-648.

Tomlins, S.A., Rhodes, D.R., Yu, J., Varambally, S., Mehra, R., Perner, S., Demichelis, F., Helgeson, B.E., Laxman, B., Morris, D.S., Cao, Q., Cao, X., Andren, O., Fall, K., Johnson, L., Wei, J.T., Shah, R.B., Al-Ahmadie, H., Eastham, J.A., Eggener, S.E., Fine, S.W., Hotakainen, K., Stenman, U.H., Tsodikov, A., Gerald, W.L., Lilja, H., Reuter, V.E., Kantoff, P.W., Scardino, P.T., Rubin, M.A., Bjartell, A.S. & Chinnaiyan, A.M. 2008. The role of SPINK1 in ETS rearrangement-negative prostate cancers. Cancer Cell 13, 6, pp. 519-528.

Tort, F., Campo, E., Pohlman, B. & Hsi, E. 2004. Heterogeneity of genomic breakpoints in MSN-ALK translocations in anaplastic large cell lymphoma. Human Pathology 35, 8, pp. 1038-1041.

Turc-Carel, C., Philip, I., Berger, M-P., Philip, T. & Lenoir, G.M. 1983. Chromosomal translocation in Ewing's sarcoma. New England Journal of Medicine 309, pp. 497-498.

Turc-Carel, C., Dal Cin, P., Limon, J., Rao, U., Li, F.P., Corson, J.M., Zimmerman, R., Parry, D.M., Cowan, J.M. & Sandberg, A.A. 1987. Involvement of chromosome X in primary cytogenetic change in human neoplasia: nonrandom translocation in synovial sarcoma. Proceedings of the National Academy of the USA 84, 7, pp. 1981-1985.

Turner, N. & Grose, R. 2010. Fibroblast growth factor signalling: from development to cancer. Nature Reviews Cancer 10**,** 2, pp. 116-129.

Visvader, J.E. 2011. Cells of origin in cancer. Nature 469, 7330, pp. 314-322.

Warrell, R.P. Jr, Frankel, S.R., Miller W.H. Jr, Scheinberg, D.A., Itri, L.M., Hittelman, W.N., Vyas, R., Andreeff, M., Tafuri, A., Jakubowski, A., Gabrilove, J., Gordon, M.S. & Dmitrovsky, E. 1991. Differentiation therapy of acute promyelocytic leukemia with tretinoin (all-trans-retinoid acid). New England Journal of Medicine 324, 20, pp. 1385-1393.

Williams, R., Peisajovich, S.G., Miller, O.J., Magdassi, S., Tawfik, D.S. & Griffiths, A.D. 2006. Amplification of complex gene libraries by emulsion PCR. Nature Methods 3, 7, pp. 545-550.

Wilson, D.L., Buckley, M.J., Helliwell, C.A. & Wilson, I.W. 2003. New normalization methods for cDNA microarray data. Bioinformatics 19, 11, pp. 1325-1332.

Wu, Y-M., Su, F., Kayana-Sundaram, S., Khazanov, N., Ateeq, B., Cao, X., Lonigro, R.J., Vats, P., Wang, R., Lin, S.F., Cheng, A.J., Kunju, L.P., Siddiqui, J., Tomlins, S.A., Wyngaard, P., Sadis, S., Roychowdhury, S., Hussain, M., Feng, F.Y., Zalupski, M.M., Talpaz, M., Pienta, K.J., Rhodes, D.R., Robinson, D.R. & Chinnaiyan, A.M. 2013. *Cancer Discovery* (in press).

Yang, Y.H., Dudoit, S., Luu, P., Lin, D.M., Peng, V., Ngai, J. & Speed, T.P. 2002. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systemic variation. Nucleic Acids Research 30, 4, e15.

Zech, L., Haglund, U., Nilsson, K. & Klein, G. 1976. Characteristic chromosomal abnormalities in biopsies and lymphoid-cell lines from patients with Burkitt and non-Burkitt lymphomas. International Journal of Cancer 17, 1, pp. 47-56.

Zhu, L., Zhang, Y., Zhang, W., Yang, S., Chen, J-Q. & Tian, D. 2009. Patterns of exon-intron architecture variation of genes in eukaryotic genomes. BMC Genomics 10, p. 47.