Identification of novel fusion genes and transcript variants in cancer

Thesis for the Philosophiae Doctor (PhD) degree

University of Oslo, 2015

Andreas Midbøe Hoff

Genome Biology Group
Department of Molecular Oncology
Institute for Cancer Research
The Norwegian Radium Hospital
Oslo University Hospital

Centre for Cancer Biomedicine
Norwegian Centre of Excellence
University of Oslo

Faculty of Medicine
University of Oslo

The Norwegian Cancer Society

Table of contents

Acknowledgements

The work presented within this thesis was carried out at the Department of Molecular Oncology, the Norwegian Radium Hospital, Oslo University Hospital. Financially, the project was supported by the Norwegian Cancer Society and the Norwegian Radium Hospital Legacies.

The environment at the Norwegian Radium Hospital and at the Institute for Cancer Research has enabled this work. Here, I would especially like to thank Ragnhild A. Lothe for managing the excellent department of Molecular Oncology, and for allowing me to be a part of this team of magnificent people. Additionally, Ragnhild has been my co-supervisor, and I am grateful for the input you have provided.

Especially warm thanks go out to my main supervisor Rolf I. Skotheim. You are an open door of brilliant ideas and a boost in confidence in times of need.

I also want to thank all of my colleagues who have contributed both academically and socially to this work. Torfinn; thank you for helping create and inspire the foundation of sequencing and bioinformatics analysis at the department, and for being patient with questions from a wet-lab wannabe bioinformatician such as myself. Anne Cathrine, I sincerely appreciate your help in the lab and your many heart-warming laughs. Thank you, Bjarne and Sen, for being patient with my quest for *in silico* knowledge. Thank you, Marthe, for your valuable ideas, and for reading and commenting on this thesis. I also appreciate input from and discussions with Anita, Sigmund, Marianne, Gard, Stian and Jonas, current and former members of the Genome Biology group. I would also like to express my gratitude to the rest of the collaborators and co-authors.

My thoughts and thanks go out to my family and friends, without whom little would have been possible.

Finally, I would like to thank my girlfriend Mayes for keeping up with me in these busy times and for making every day even more joyful.

Oslo, June 28, 2015

Andreas Hoff

List of abbreviations

A, C, G, T	Adenine, Cytosine, Guanine and Thymine
CCLE	Cancer Cell Line Encyclopedia
cDNA	Complementary DNA
CML	Chronic myelogenous leukemia
CRC	Colorectal cancer
CSC	Cancer stem cell
ddPCR	Droplet digital PCR
EC	Embryonal carcinoma
ENCODE	Encyclopedia of DNA Elements
ES cell	Embryonic stem cell
EST	Expressed sequence tag
FAP	Familial adenomatous polyposis
FDA	Food and Drug Administration
FISH	Fluorescence *in situ* hybridization
IGCN	Intratubular germ cell neoplasia
MMR	Mismatch repair
MSI	Microsatellite instable
MSS	Microsatellite stable
NCGC	Norwegian Cancer Genomics Consortium
PCR	Polymerase chain reaction
PGC	Primordial germ cell
PSA	Prostate specific antigen
RA	all-*trans* retinoic acid
RACE	Rapid amplification of cDNA ends
SBS	Sequencing by synthesis
SNP	Single nucleotide polymorphism
TAC	Transit amplifying cell
TCGA	The Cancer Genome Atlas
TGCT	Testicular germ cell tumor
TNM	Tumor node metastasis
UTR	Untranslated region
YST	Yolk sac tumor

Gene nomenclature

ABHD12B	Abhydrolase domain containing 12B
ABL1	ABL proto-oncogene 1, non-receptor tyrosine kinase
AMER1	APC membrane recruitment protein 1
APC	Adenomatous polyposis coli
AXIN2	Axin 2
BCR	Breakpoint cluster region
BPIFA2	BPI fold containing family A, member 2
CASZ1	Castor zinc finger 1
CCND2	Cyclin D2
CIC	Capicua transcriptional repressor
CLEC4D	C-type lectin domain family 4, member D
CLEC6A	C-type lectin domain family 6, member A
CTNNB1	Catenin (Cadherin-Associated Protein), Beta 1, 88kDa
DCC	DCC netrin 1 receptor
DHX35	DEAH (Asp-Glu-Ala-His) box polypeptide 35
DKK1-4	Dickkopf WNT signaling pathway inhibitor (1-4)
DNAJB1	DnaJ (Hsp40) homolog, subfamily B, member 1
DPP3	Dipeptidyl-peptidase 3
DPPA3	Developmental pluripotency associated 3
DUX4	Double homeobox 4
ELK4	ELK4, ETS-domain protein (SRF accessory protein 1)
EPT1	Ethanolaminephosphotransferase 1
ETV6	Ets variant 6
FZD10	Frizzled class receptor 10
GUCY1A3	Guanylate cyclase 1, soluble, alpha 3
HENMT1	HEN1 methyltransferase homolog 1 (Arabidopsis)
JAZF1	JAZF zinc finger 1
JJAZ1(SUZ12)	SUZ12 polycomb repressive complex 2 subunit
KIT	v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog
KITLG	KIT ligand
KLK7	Kallikrein-related peptidase 7
KLK8	Kallikrein-related peptidase 8
KRAS	Kirsten rat sarcoma viral oncogene homolog
LRP5	Low density lipoprotein receptor-related protein 5
MASP2	Mannan-binding lectin serine peptidase 2

MT-CO1	Mitochondrially Encoded Cytochrome C
MT-RNR2	Mitochondrially Encoded 16S RNA
NANOG	Nanog homeobox
PPP6R3	Protein phosphatase 6, regulatory subunit 3
PRKACA	Protein kinase, cAMP-dependent, catalytic, alpha
RCC1	Regulator of chromosome condensation 1
RCC2	Regulator of chromosome condensation 2
RP11-57H14.3	RP11-57H14.3
S100A2	S100 calcium binding protein A2
SLC45A3	Solute carrier family 45, member 3
SOX9	SRY (sex determining region Y)-box 9
TCF7L2	Transcription factor 7-like 2 (T-cell specific, HMG-box)
TMPRSS2	Transmembrane protease, serine 2
TP53	Tumor protein P53
VTI1A	Vesicle transport through interaction with t-SNAREs 1A
VWA2	Von Willebrand factor A domain containing 2

List of papers

I. **High frequency of fusion transcripts involving *TCF7L2* in colorectal cancer: novel fusion partner and splice variants**
Torfinn Nome*, Andreas M. Hoff*, Anne Cathrine Bakken, Torleiv O. Rognum, Arild Nesbakken, Rolf I. Skotheim
PLoS One 2014, 9:e91264 *Equal contribution

II. **Novel RNA variants in colorectal cancers**
Andreas M. Hoff*, Bjarne Johannessen*, Sharmini Alagaratnam, Sen Zhao, Torfinn Nome, Marthe Løvf, Anne C. Bakken, Merete Hektoen, Anita Sveen, Ragnhild A. Lothe, Rolf I. Skotheim
Manuscript *Equal contribution

III. **RNA sequencing reveals fusion genes in testicular germ cell tumors**
Andreas M. Hoff, Sharmini Alagaratnam, Sen Zhao, Jarle Bruun, Peter W. Andrews, Ragnhild A. Lothe, Rolf I. Skotheim
Manuscript

Summary

Cancer cells grow and multiply without normal control of cell growth. They have the ability to invade surrounding tissues, disrupt normal organ functions, metastasize, and if not treated effectively lead to a poor outcome for the patient. These features of the cancer cells reflect dysregulation of extremely complex and balanced cellular machinery encoded by the genome, the blueprint of our cells. The tremendous increase in knowledge of the underlying genetics of cancer development has resulted in better understanding of the biologic processes leading to cancer. On the other hand, the many large data sets also provide a vast amount of false positives and negatives. Furthermore, we are just starting to really investigate in detail the total transcriptome, coding and non-coding molecules. Genetic changes which are shown to be expressed in cancer cells have a clinical potential as biomarkers in cancer detection, diagnostics, prognostics and in targeted therapy. Structural changes of the genome are one type of genetic change found in various malignancies, and these include "fusion genes" which may be expressed as fusion transcripts. Transcript variants specific to cancer may also be produced by other mechanisms such as aberrant splicing.

This thesis includes three papers where applications of high-throughput sequencing of DNA and RNA have been applied to detect novel fusion transcripts and transcript variants relevant to development of colorectal cancer (CRC; papers I and II) or testicular germ cell tumors (TGCTs; paper III).

In paper I, we performed RNA- and whole-genome sequencing of CRC cell lines. We confirmed the presence of *VTI1A-TCF7L2*, the first recurrent fusion gene recently described in CRC and identified *RP11-57H14.3* as a novel fusion gene partner of *TCF7L2,* a transcription factor of the canonical WNT signaling pathway. The original and newly identified fusion transcripts involving *TCF7L2* were found to be expressed in 42 % and 45 % of primary carcinomas from CRC patients (n = 106). Importantly, both fusion transcripts were found to be expressed in several normal colonic mucosa samples. In contrast to cell lines that harbor genomic rearrangements and high level of fusion transcripts, our data suggest that a large fraction of CRC and normal samples express these fusion transcripts at low levels, which is important in regards to using fusion genes as biomarkers.

In paper II, we analyzed exon-level microarray data from 202 CRCs and nominated 25 genes that showed overexpression of their 3' end in one or more CRC. Such deviating 3' expression indicate that these parts of the gene are under the control of an alternative promoter, either within the same gene or from another upstream partner gene as part of a fusion gene. To enable effective characterization of the underlying transcript structures of these 25 genes, we developed a new protocol, RACE-seq, combining 5' rapid amplification of cDNA ends (RACE) of multiple genes, followed by high-level multiplexing and pooling of RACE fragments and samples. We analyzed the combined pool, consisting of 25 candidate genes in 23 CRCs in one single sequencing reaction. From subsequent data analysis, we identified three private fusion transcripts: *VWA2-TCF7L2, DHX35-BPIFA2*, and *CASZ1-MASP2*. We also identified novel transcript junctions supporting a read-through between *KLK8* and *KLK7* and a new 3' splice site in *S100A2*. Both of these were overrepresented in a high number of CRCs as compared to normal colonic mucosa. In addition to the current identified overrepresentation of the novel variants in CRC, both *KLK7* and *S100A2* have previously been implicated in this disease, strengthening their potential role as cancer biomarkers.

In the final study of this thesis, we also used RNA-sequencing to search for driver fusion genes, but this time in TGCTs, a malignancy with few gene (point-) mutations and no previously detected fusion genes. In total, we discovered nine novel transcript breakpoints supporting eight fusion transcripts and one alternative promoter usage. *RCC1-ABHD12B, RCC1-HENMT1, CLEC6A-CLEC4D* and a transcript using an alternative promoter for *ETV6* were recurrently expressed in an extended series of TGCT samples. Additionally, the *RCC1* involving fusions and the alternative promoter usage of *ETV6* was found to be favorably expressed in undifferentiated subtypes of TGCT, and expression was reduced upon induction of *in vitro* differentiation by treating cells with all-*trans* retinoic acid (RA). None of the four recurrent transcripts were detected in normal parenchyma of the testis, and *RCC1-ABHD12B* and *ETV6* were not expressed in any of 20 tested miscellaneous normal tissues. The fusion genes discovered in this study are the first to be described in TGCT.

In conclusion, by state of the art genomics in combination with an efficient experimental protocol we have identified and validated the presence of novel transcript variants in two types of malignancies. We have also through several control experiments shown that care should be taken to suggest transcript variants as biomarkers since low level presence can be found in normal tissues.

Introduction

Cancer – a common enemy

Cancer affects us all in some way, either directly with a personal cancer diagnosis or as bystanders to people we love struck by the disease. With an estimated total of 14.1 million people diagnosed with cancer and 8.2 million deaths from cancer in 2012, cancer is a leading cause of death worldwide [1]. In 2008 it was estimated that 170 million years of healthy life were lost globally due to cancer, making a huge impact on worldwide health and economy [2]. With a steady increase in life expectancy of the worldwide population, it is predicted that there will be a marked increase in worldwide cancer incidence, with an estimation of 22.2 million cases in 2030 [3]. The ever growing impact of cancer on health and economy demands the development of efficient tools, both molecular and others, for improved diagnosis, prognosis, treatment, and follow up of cancer patients. However, cancer is not one disease, but a collection of related diseases with more than 100 types occurring in humans. The four most common cancer types are lung, breast, CRC and prostate cancers, which together account for 4 in 10 of all cancers diagnosed worldwide (Figure 1).

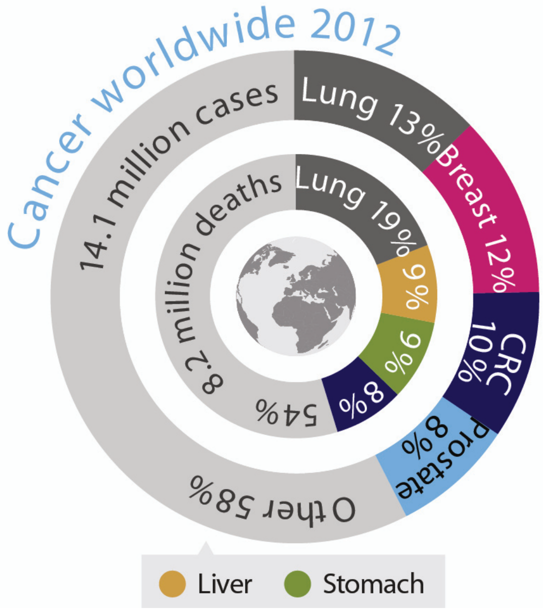


Figure 1: Global cancer incidence (outer ring) and mortality (inner ring) in 2012. Figure adapted from cruk.org/cancerstats [accessed may 2015]

Hallmarks of cancer

The various cancer types are most commonly defined by the cell and tissue type in which they arise. In addition, nearly all cancer types have several subgroups based on histological appearance, and increasingly by their molecular alterations. The latter makes the cancer sub-grouping quite complex, where individual cancers from the same tissue type and histological subgroups exhibit different molecular traits, thus demanding different clinical handling. Finding a single "cure" for cancer is inevitably unlikely. Nevertheless, some common integral components of cancer have been suggested; functional capabilities believed to be necessary for the multistep establishment and maintenance of malignant cancer. Known as the hallmarks of cancer, these include self-sufficiency of growth signals, insensitivity to growth suppressors, resistance to programmed cell death (apoptosis), immortality by limitless replicative potential, sustained angiogenesis, and activation of tissue invasion and metastasis [4]. In 2011, Hanahan and Weinberg reviewed the validity of these original hallmarks in light of recent advances in cancer research. In addition to underlining the continued importance and validity of the original hallmarks, they suggested that reprogramming of energy metabolism and evasion of immune destruction are additional hallmarks of cancer [5]. Although being common hallmarks of cancer, these common traits are often acquired by different underlying molecular mechanisms. The genomes of cancers are often highly unstable, with a higher mutation frequency than normal cells. Genomic instability has been suggested to be an enabling characteristic that facilitates rapid acquisition of hallmark capabilities [5]. An additional enabling characteristic is the interplay of cancer cells with the tumor microenvironment and invading immune cells. Immune cell inflammation can contribute to multiple hallmark capabilities; e.g. by supplying growth factors that sustain proliferative signaling, survival factors that limit cell death, and proangiogenic factors [5].

Cancer – a genetic disease

“We have not slain our enemy, the cancer cell. We have only seen our monster more clearly”. – Harold Varmus, Nobel Banquet, December 10, 1989

These are the words of Harold Varmus upon receiving the Nobel Prize in Physiology or Medicine, for the discovery of human oncogenes. To be able to treat cancer and kill cancer cells, it is of uttermost importance to understand the molecular processes of carcinogenesis. The earliest study of cell division in malignant cells was made in 1890 by David von Hansemann who drew attention to aberrant mitosis in cancer cells [6]. This was followed by the somatic mutation theory of Theodor Boveri, the first who proposed that aberrant chromosomes might be the cause of cancer [7]. Since then, it has been established that cancer is a disease of the molecular machinery encoded in the genome. The central dogma of molecular biology, first stated by Francis Crick in 1958 and re-stated in 1970 [8], describes the flow of information in molecular biology. In essence, although oversimplified, DNA makes RNA, and RNA makes proteins. This flow of information is tightly and complexly regulated at multiple levels and controls the cellular phenotype. Deregulation of this machinery in normal cells may lead to the acquisition of the phenotypic traits considered as the hallmarks of cancer and consequently to malignant transformation [5]. Understanding the genome, transcriptome, proteome, and how it all works together and is deregulated in cancer is essential if we are to understand and “slay our enemy”, the cancer cell.

The cancer genome

The genome of an individual is identical across all different cell types, and consists of 23 pairs of chromosomes which are built up of long DNA helix molecules. In 1953, Watson and Crick discovered that the structure of DNA consists of two strands in a double helix with a phosphate-deoxyribose backbone and the nucleobases; Adenine (A), Cytosine (C), Guanine (G) and Thymine (T) [9]. The nucleobases form hydrogen bonds and have specific binding affinities, where the pyrimidines, C and T, binds to the purines, G and A, respectively. In the early 1990s, researchers set out to characterize the entire sequence of our genome in the human genome project. The first draft of the genome was published in 2001 by the human genome project and by a private effort led by Celera genomics and Craig J. Venter, in parallel issues of Science and Nature [10,11]. A more finished genome with previously missing euchromatin parts was published in 2004 [12]. In total our genome contains 3.3 billion base pairs or “letters of code”. The genome and the chromosomes are further divided into regions of functional or coding DNA; the genes. Upon the completion of the human genome project, the number of genes in the human genome was found to be between 20,000 and 25,000 [12]. This number was substantially lower than the previously estimated number of genes, thought to be closer to 120,000 [13]. In fact, only 2 % of the sequences in our genome are protein-coding.

Mutations in the nucleotide sequences of our genes are a common cause of cancer, and can be accelerated by environmental factors, mutagens, such as tobacco smoke, asbestos and UV-radiation. Some genes are found to be recurrently mutated in cancer and are thought to be drivers of tumorigenesis. These so-called cancer genes are divided into proto-oncogenes and tumor suppressor genes. Mutations of proto-oncogenes are often dominant, activating mutations leading to gain of function and oncogene activation. Mutations of tumor-suppressor genes reduce the activity of the gene and are normally recessive, meaning that loss of both copies of the tumor suppressor genes is usually necessary [14]. In practice, this often occurs as a result of an inactivating truncating mutation followed by loss of heterozygosity [15,16]. Activating or truncating point-mutations, or small nucleotide insertions/deletions, are not the only ways proto-oncogenes and tumor suppressor genes are activated or deactivated. Cancer genomes are frequently aneuploid and have extensive copy number changes [17]. These may cause gain or loss of chromosome material, and consequently more or less copies of proto-oncogenes and tumor suppressor genes. As a result, expression of these genes may be altered.

Epigenetic changes also cause deregulation of cancer genes. These are somatically heritable variations that are not caused by direct changes to the DNA sequence itself, but rather through different modifications [18]. There are three main types of epigenetic inheritance, i.e. DNA methylation, genomic imprinting and histone modification. The cancer genomes are generally hypomethylated, but with frequent aberrant promoter hypermethylation and consequent genetic silencing of tumor suppressor genes [19]. Also, hypomethylation has been shown to activate the expression of cancer proto-oncogenes [18].

Fusion genes and chromosomal rearrangements

In addition to large scale genomic copy number changes resulting in numerical aberrations, chromosomal rearrangements causing structural changes are frequent in cancer cells. These include intrachromosomal rearrangements; duplications, deletions or inversions of parts of single chromosomes, and interchromosomal rearrangements; insertions or translocations involving two chromosomes (Figure 2).

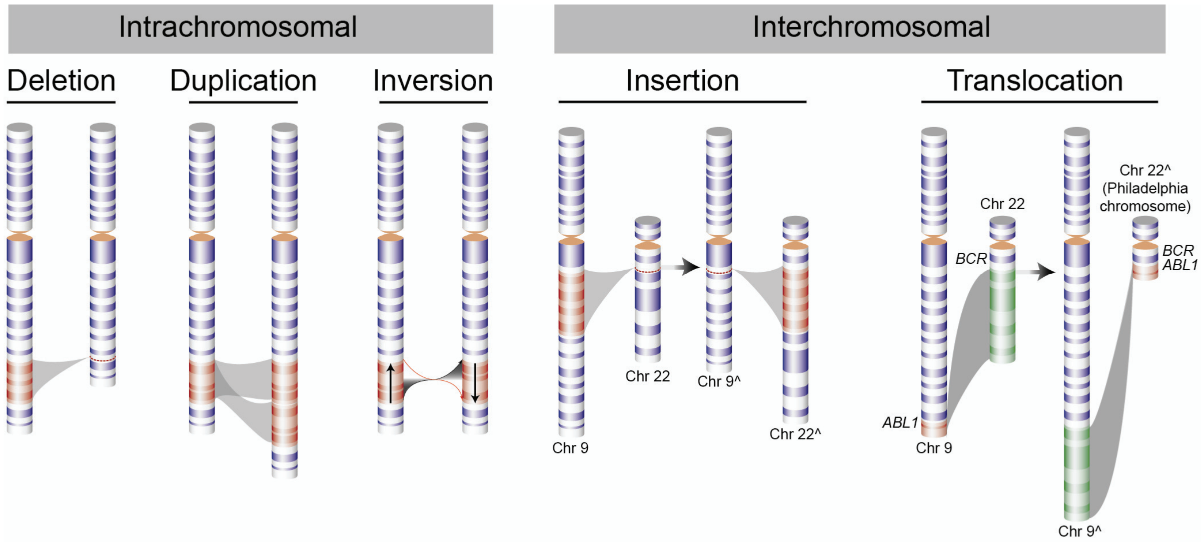


Figure 2: Common types of chromosomal rearrangements. As an example, the interchromosomal rearrangements here depict chromosomes 9 and 22. A translocation t(9;22)(q34;q11) generates the Philadelphia chromosome and the consequent formation of the *BCR-ABL1* fusion gene.

Common for such chromosomal rearrangements is the alteration of genomic context, or the "environment" of genes. Whole genes may be positioned closer or further away from regulatory elements, including promoters, enhancers and silencers, leading to altered regulation and expression of affected genes. Such chromosomal rearrangements can also lead to the coupling of two previously separate genes into a fusion gene. Consequently, fusion transcripts (chimeric RNA from both gene sequences) can be expressed (Figure 3). Fusion genes may put a downstream gene under the regulation of a promoter of an upstream partner gene, resulting in altered expression of the downstream partner gene. Another potential consequence of fusion genes is the expression of fusion transcripts containing coding sequence from both partner genes, potentially encoding a fusion protein. Additionally, a partial in-frame coding sequence of one of the partner genes can be fused to the untranslated region (UTR) or promoter regions of the other partner gene; leading to N- or C-terminally truncated protein products and consequently loss of or altered function (Figure 3). Fusion genes are often highly cancer specific, and some are pathognomonic for subtypes of cancer. The first described recurrent chromosomal rearrangement was the translocation between chromosomes 9 and 22, known as the Philadelphia chromosome, in patients with chronic myelogenous leukemia (CML; Figure 2) [20,21]. Later, a fusion of *BCR* and *ABL1* was found to be the consequence of the Philadelphia chromosome, encoding a hybrid protein with abnormal, oncogenic tyrosine kinase activity [22,23]. By molecular cytogenetic techniques, especially chromosome banding and fluorescence *in situ* hybridization (FISH), several recurrent fusion genes were discovered and characterized in the 1980s and early 90s [24,25]. However, cytogenetic techniques used to detect fusion genes in a guided fashion were biased towards detecting interchromosomal fusion genes in hematological cancers [24]. With the development of high-throughput methods for analyzing gene expression, such as array based technologies in the 1990s and 2000s, recurrent fusion genes were also discovered in common carcinomas, exemplified by the discovery of recurrent fusion genes involving *TMPRSS2* and *ETS* family genes in more than half of prostate cancers [26]. More recently, deep-sequencing technologies have enabled unbiased detection of fusion genes caused by everything from subtle intrachromosomal rearrangements to clear interchromosomal rearrangements. This has resulted in a plethora of new fusion genes being described during the past 5 years. The number of fusion genes described is now close to 10,000, with more than 9,000 of them discovered by deep-sequencing approaches [24]. In one study by The Cancer Genome Atlas (TCGA), close to

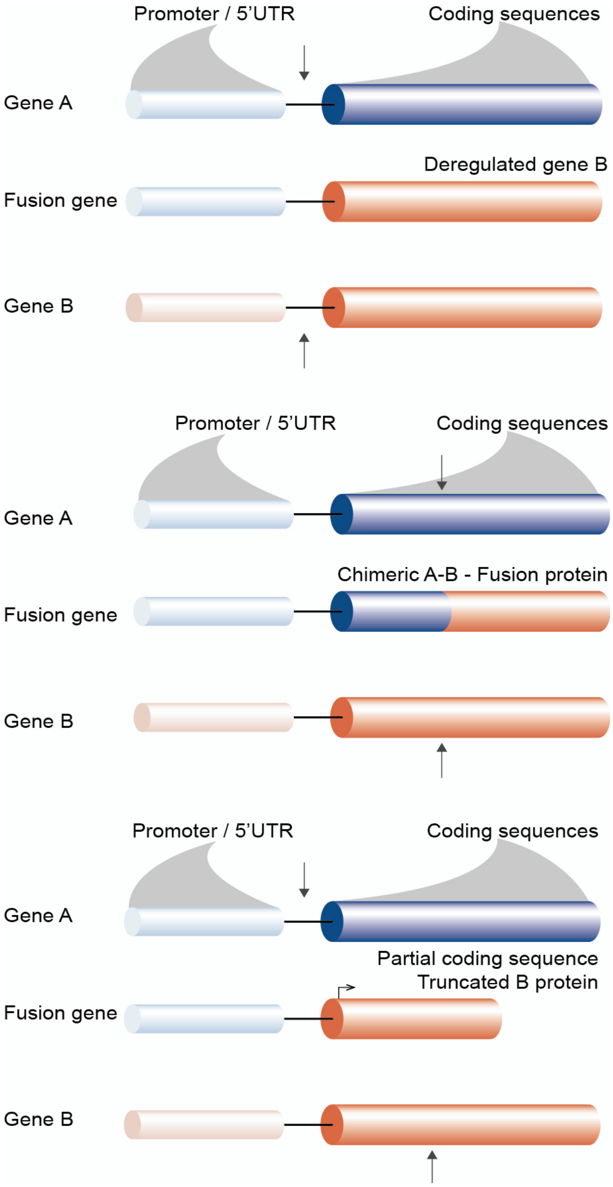


Figure 3 Fusion gene consequences.
Arrows in gene A and B represent breakpoints causing fusion genes with different consequences. Fusion breakpoints commonly occur in intronic regions. Intronic regions are not presented in the figure, but the promoter or 5' UTR is separated from the coding sequences by a line representing a intronic segment.

8,000 fusion transcripts were discovered from transcriptomic data from 4,366 tumors representing 13 different cancer types [27]. Although many of the fusion genes discovered are important drivers in cancer, most of the nearly 10,000 fusion genes are probably passenger mutations. In fact, only 3 % of fusion genes have been recurrently identified, i.e. reported in at least two separate publications [24]. The increased genomic instability in cancer, seen especially in solid carcinomas, is probably contributing heavily to the generation of chromosomal rearrangements and passenger fusion genes detected by deep sequencing [28,29]. Filtering steps to enrich for recurrent and functionally important fusion genes are therefore of importance. Lately, fusion transcripts have also been shown to be expressed without corresponding genomic rearrangement but probably as a result of transcription coupled mechanisms [30].

The transcriptome

The complexity of living organisms and of humans is not reflected in the number of genes. This became evident upon the completion of the human genome project and the discovery of relatively few coding genes. However, at the level of transcription, additional information and complexity is introduced. Although 98 % of the genome is non-protein coding sequences [31], studies have shown that up to three-quarters of the genome is transcribed and involved in gene regulation [32,33]. Most of the transcribed genomic sequences are non-coding RNA, RNA molecules that are not translated into protein sequences [34]. Many of these still have important functional and regulatory roles, such as structural RNAs; including transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs) important in protein translation, and small nuclear RNAs (snRNAs) that are

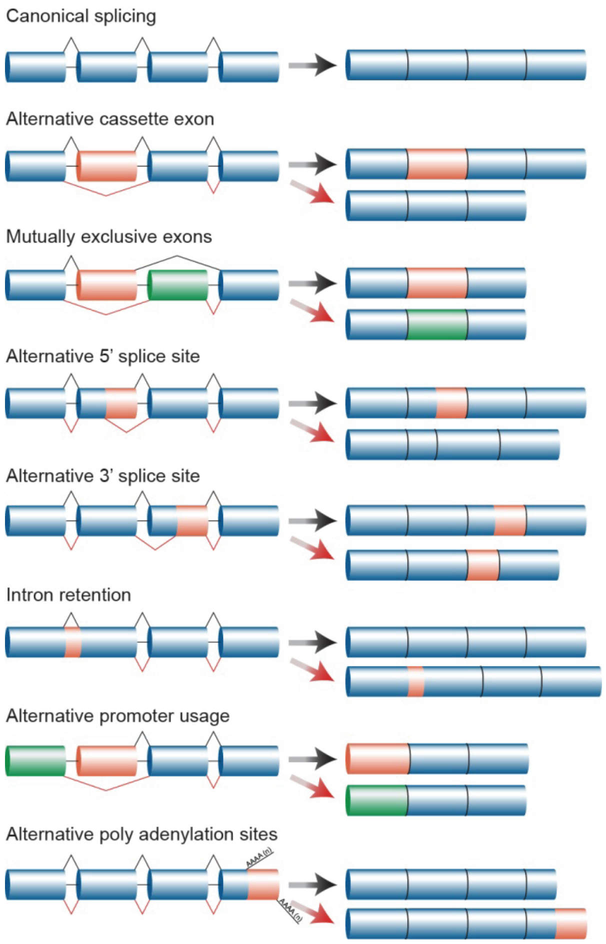

Canonical splicing
Alternative cassette exon
Mutually exclusive exons
Alternative 5’ splice site
Alternative 3’ splice site
Intron retention
Alternative promoter usage
Alternative poly adenylation sites
AAA(n)
AAA(n)

Figure 4: Common mechanisms of alternative mRNA transcript variant generation.

involved in the splicing of pre-mRNA. Regulatory RNAs, including micro RNAs (miRNAs) and long non-coding RNAs (lncRNAs) have several important roles in regulation of gene expression and protein translation. An additional level of complexity is added by alternative splicing of protein-coding pre-mRNAs. In higher eukaryotes, the sequences of genes are composed of coding parts (exons) that are interrupted with long non-coding sequences (introns). Upon transcription, pre-mRNAs are made in the cell nucleus. These are further processed by the spliceosome to excise non-coding intronic sequences and splice together the coding exons into mature mRNAs. It is estimated that 95 % of multi-exon genes in mammalian genomes undergo alternative splicing and can produce several alternative splice variants by combining the exon cassettes in different arrangements [35,36]. Several different splicing and non-splicing mechanisms create different transcript variants as illustrated in Figure 4. The diverse transcript variants may further be translated into protein variants with alternative functions. The process of alternative splicing is important in development and determining cellular phenotype. Analogous to alternative splicing observed in various normal cell types, alternative transcript expression as a result of aberrant splicing has been linked to cancer [37–40]. Alternative transcripts can impact tumorigenesis by having altered functions compared to normal transcript variants of the genes. Fusion transcripts have also been shown to be expressed as a result of splicing events between two or more genes [41]. Genes located in close proximity frequently express chimeric transcripts or fusion transcripts resulting from read-through transcription and subsequent *cis*-splicing [30,42]. Fusion transcripts of two genes that are distant to each other, or on different chromosomes, have also been found to be expressed without evidence of corresponding chromosomal rearrangements, these are likely products of *trans*-splicing [43]. It has been proposed that such fusion transcripts can increase the complexity of the genome by translating into distinct proteins [44,45]. In addition to being expressed in normal cells, fusion transcripts without corresponding genomic rearrangements have been shown to be overexpressed in cancer, and have an impact on cancer biology by regulating both replication and cellular growth [46–48]. One such fusion transcript, *SLC45A3-ELK4* is expressed in both normal prostate tissue and prostate cancer, with high levels of expression in a subset of prostate cancer samples. Only some prostate cancers expressing these fusion transcripts harbor chromosomal rearrangements at the corresponding genomic loci [49,50]. Another fusion transcript between *JAZF1* and *JJAZ1(SUZ12)* is expressed in normal endometrial stromal cells and is translated into JAZF1-JJAZ1, a protein with anti-apoptotic activity. Interestingly, the chimeric RNA and protein is identical to those produced from a t(7;17)(p15;q21) chromosomal rearrangement found only in endometrial stromal tumors [51]. These findings led to the hypothesis that somehow the mechanism of *trans*-splicing, generating such fusion transcripts, guide the chromosomal rearrangements seen in cancer [43].

Colorectal cancer

Cancer of the colon or rectum, CRC, is the third most common cancer worldwide in men and the second in women, with an estimated total of 1.4 million new cases and 694,000 CRC related deaths worldwide in 2012 [1]. Norway is one of the countries in the world with the highest incidence of CRC, with almost 4,000 men and women diagnosed in 2013. The five-year relative survival across all stages is about 60 % in Norway [52]. CRC is associated with a so-called western lifestyle, and the incidence is substantially higher in more developed countries of the world (Figure 5) [1,53]. Countries that have recently adopted the western lifestyle, such as eastern European countries with a booming economical status, also show increasing incidence of CRC [54].

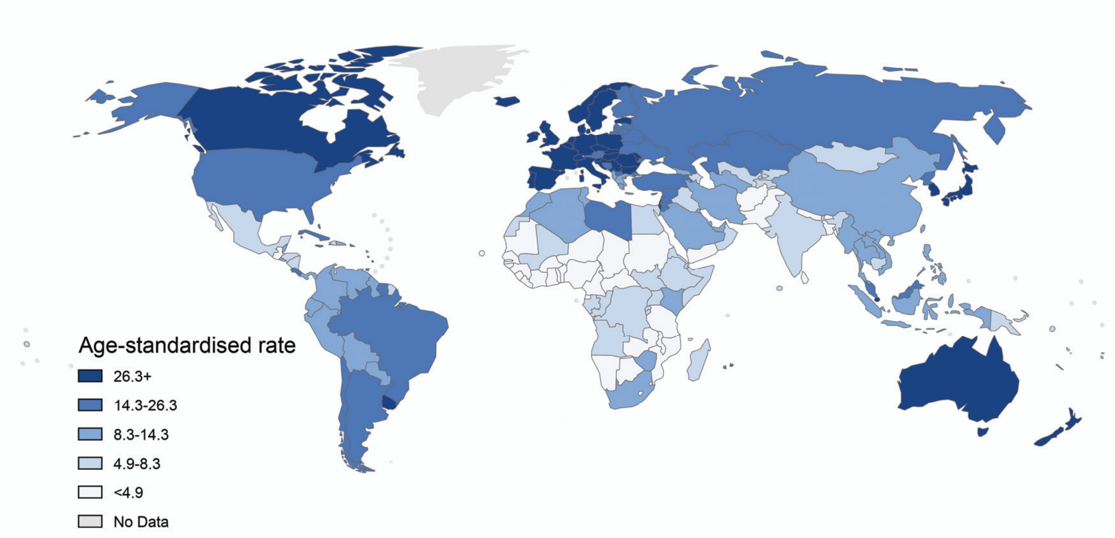


Figure 5: Worldwide age-standardized incidence of colorectal cancer for both sexes. Age-standardized rate is a summary of the rate that a population would have if it had a standard age structure. Source: GLOBOCAN 2012 (IARC) [55]

The strongest risk factor for CRC is age, with over 90 % of patients being older than 50 years, and the average age is about 70 years [56]. Environmental risk factors associated with CRC include intake of alcohol, smoking, intake of processed red meat, obesity and diabetes; all risk factors linked to a western lifestyle [57]. Prognosis of CRC is largely dependent on the cancer stage at the time of diagnosis, which is classified according to the Tumor node metastasis (TNM) system [58]. In the TNM system, CRC is graded according to depth of infiltration of the primary tumor (T1-T4), the extent of involved regional lymph nodes (N0-N2) and the presence of distant metastasis (M0-M1). Stage I and II is considered localized disease with N0 and M0, however with a deeper infiltration of the primary tumor into surrounding tissue in stage II (T3-T4). Stage III is classified as regional disease with detectable cancer cells in lymph nodes (any T stage, N1-N2), whereas in stage IV metastases are found in other organs (any T, any N, M1). Increasing TNM-stage is associated with increasingly poor prognosis, and survival with metastatic disease is low [52]. Treatment of CRC is largely determined according to the TNM system. Main treatment regimes include surgery, and for stage III and IV, adjuvant chemotherapy. Also, for rectal cancer, radiation treatment is common. The TNM criteria for determining treatment regimes are not optimal, exemplified by the paradox that stage IIB/C (T4N0) patients have a worse prognosis than stage IIIA patients (T1-2N1) [59,60]. Additionally, patients within the individual stages are a heterogeneous group, with different clinical outcomes. For instance some stage II patients could benefit from adjuvant chemotherapy since ~25 % of this patients group experience relapse. Most stage III patients receive adjuvant chemotherapy but some of them are over-treated and would have been cured with surgery alone. Biomarkers that could better predict patient prognosis and treatment response are warranted for improved adjusted individualized treatment regime.

About 20-30 % of patients with CRC are believed to have increased hereditary susceptibility to the disease, but only 5 % of the patients have known heritable CRC-associated syndromes caused by single gene defects [61]. Known hereditary CRC-associated syndromes include Lynch syndrome (also known as hereditary nonpolyposis colorectal cancer) and familial adenomatous polyposis (FAP), estimated to account for 3-4 % and 1 % of CRC cases, respectively [61]. Individuals with autosomal dominant Lynch syndrome have an 80 % lifetime risk of developing CRC and often have early onset of the disease. These individuals carry germline mutations in DNA mismatch repair (MMR) genes. As a result, DNA replication errors occur in repeat sequences, causing microsatellite instable (MSI) tumors [61,62]. FAP has been found to be caused by germline mutation in the *APC* gene. Individuals with FAP acquire hundreds to thousands of small adenomatous polyps in the colon, each of which has a small chance of developing into malignant CRC. However, due to the large number of such polyps, development into malignant CRC is nearly inevitable [61,62]. The germ line mutations predisposing for heritable CRC are also seen as somatic mutations in the development of sporadic tumors. However, patients with heritable CRC are often diagnosed at a younger age and often with synchronous or metachronous multiple primary tumors. The development of most sporadic CRCs is thought to be a lengthy process, taking years if not decades. Starting from normal cells located in the crypts of the colon, to dysplastic or hyperplastic aberrant crypt foci developing into benign adenomas and ending up in full blown carcinomas that permeates the basal membrane and infiltrates healthy organs. In 1990, the adenoma to carcinoma sequence was proposed by Fearon and Vogelstein to be a succession of acquired genetic changes priming the cells with hallmark cancer abilities, necessary for malignant growth [63]. The initiating step is activation of the WNT signaling pathway, most commonly through genetic disruption of the *APC* gene [64]. The second event is often activation of the RAS signaling pathway. Inactivation of the tumor suppressor gene *TP53* is associated with later malignant transformation and observed in about half of CRCs [65–67]. Additionally, malignant transformation from adenoma to carcinoma is associated with other genetic and epigenetic changes, the latter leading to aberrant methylation of tumor suppressor genes and consequently gene inactivation. In addition to point mutations, gene rearrangements and gene overexpression also contribute to acquiring hallmarks necessary for malignant transformation.

Box 1 | Stem cells

Stem cells are undifferentiated cells capable of self-renewal by mitosis and able to differentiate and generate other specialized cell types. Stem cells can be divided into **embryonic stem cells** and **adult stem cells**. Embryonic stem cells are derived from the inner cell mass of the blastocyst stage embryo and are **pluripotent**, meaning they can give rise to all cell types of the body derived from the three germ layers. Adult stem cells are found in stem cell niches of tissues throughout the body, and are responsible for replenishing dying cells and regeneration of damaged tissue. These are generally **multipotent**, i.e. able to differentiate into any cell type of a specific lineage. Examples are the hematopoietic stem cells that reside in the bone marrow, which can give rise to all types of blood cells, but not other types of cells. Another type of stem cells is induced pluripotent stem cells. These are created from differentiated somatic cells by inducing them to express genes that are normally expresed in embryonic stem cells. Induced pluripotent stem cells exhibit similar phenotypes as embryonic stem cells and are able to differentiate into cells of all three cell lineages [68].

WNT signaling in colorectal cancer

The colon epithelial lining has narrow invaginations called crypts. Normal crypts of the colon are arranged with a stem cell compartment at the bottom. The concept of stem cells is briefly summarized in **Box 1**. In the colon crypts each stem cell divides asymmetrically and forms a new stem cell and a daughter transit amplifying cell (TAC). The TACs then divide more rapidly and give rise to more differentiated cell types that migrate out of the epithelial crypts as they differentiate. Eventually, the differentiated cells die by apoptosis and are sloughed off and lost in the lumen of the colon. The WNT signaling pathway controls the cell fate along the crypt – lumen axis, with a high gradient of WNT signal molecules in the lower parts of the crypts that activate the WNT signaling cascade. Mesenchymal cells that surround the bottom of the intestinal crypts are thought to produce the WNT signaling ligands that activate the WNT signaling cascade (Figure 6). Briefly, activated cell surface receptors activate an intracellular signaling cascade that hinders phosphorylation and subsequent degradation of β-catenin. As a consequence, levels of cytoplasmic β-catenin increase and migrate into the cell nucleus where it binds to transcription factors TCF4/LEF and activates expression of downstream target genes, important for proliferation and maintaining an undifferentiated cell state. Towards the lumen of the colon, the WNT signal molecule concentration subsides, turning WNT signaling off. Consequently, the progenitor cells stop dividing and differentiate. The plastic organization of the colon’s crypt epithelial tissue, suggests that genetic alterations occurring in more differentiated cell types towards the lumen will be lost due to the short life-span of these cells. However, mutations in the *APC* gene in the stem cell compartment, or relatively undifferentiated TACs or progenitor cells, will activate the WNT signaling cascade even in absence of WNT signal molecules. Consequently, the undifferentiated compartment will expand, priming for further genetic and epigenetic changes that could turn the benign dysplastic or hyperplastic growth into a malignant carcinoma.

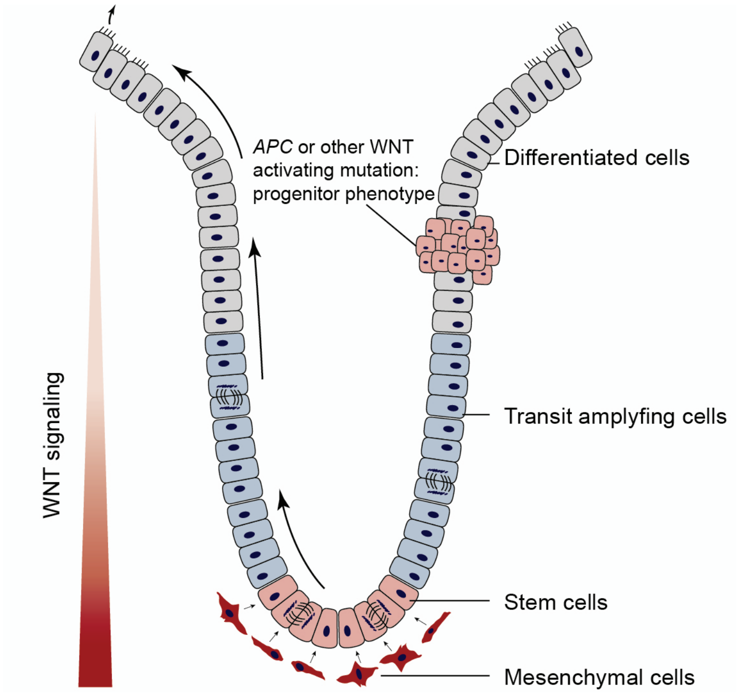


Figure 6: The colon crypt organization

WNT signaling is altered in nearly all CRCs, most commonly through inactivating mutations of the *APC* gene, but also by genetic changes observed in other components of the pathway (Figure 7). Activating point mutations in the *CTNNB1* gene (coding for β-catenin) are seen in some CRCs with intact *APC* genes [69]. These mutations inhibit the binding of β-catenin to the destruction complex and consequently activate WNT signaling. Mutations in *TCF7L2,* the gene coding for the transcription factor TCF4, is also seen in CRCs, with highest frequency in MSI tumors [70]. A fusion gene connecting *VTI1A* to parts of *TCF7L2* has been described, and was the first publication of a recurrent fusion gene in CRC [71]. Although it’s not sure if this fusion activates WNT signaling, knockdown of the expressed fusion transcript leads to loss of anchorage independent growth, a trait necessary for the cancer hallmark invasion and metastasis [5,71]. Fusion genes involving *TCF7L2* are further investigated in this thesis (paper I).

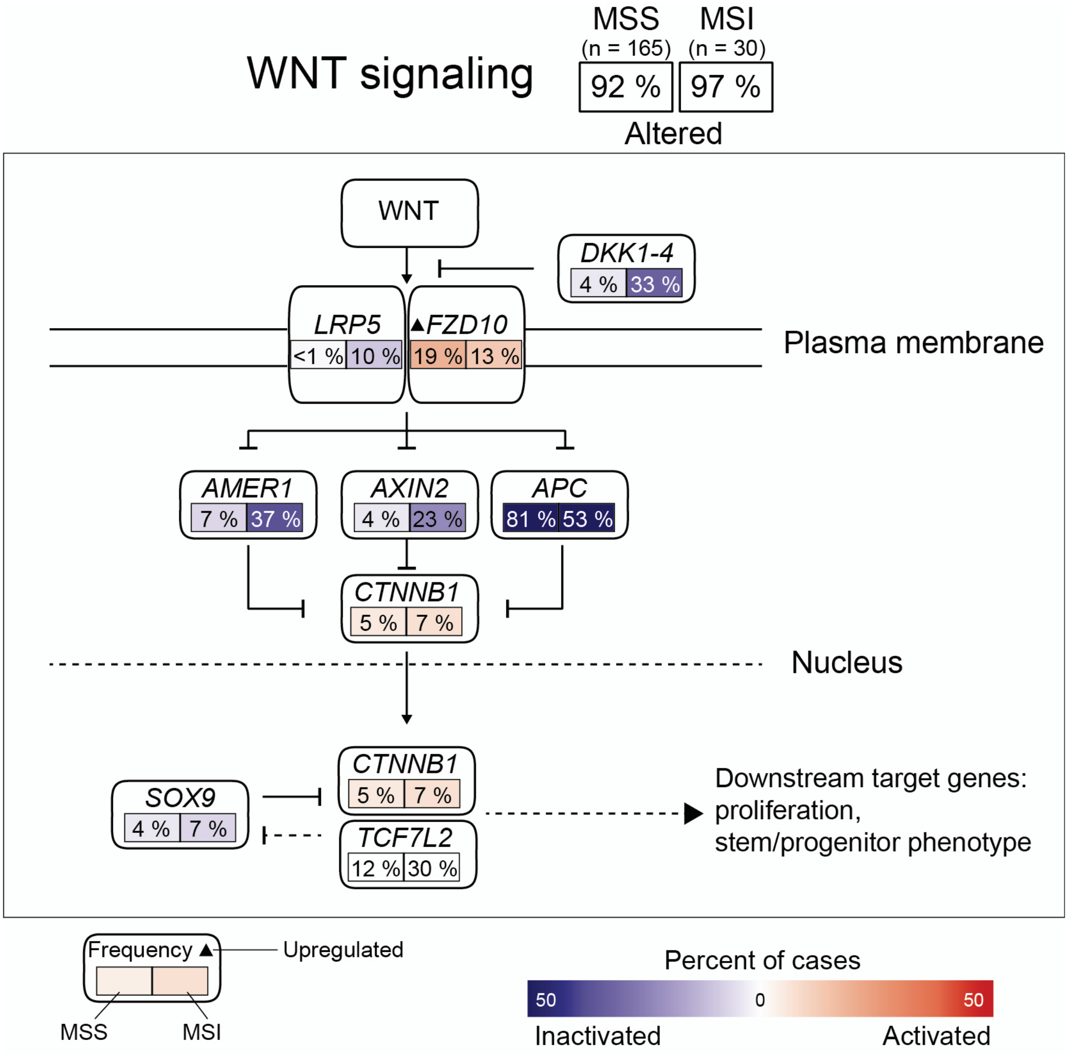


Figure 7: Genetic changes in the WNT signaling pathway. The genetic changes include somatic mutations, homozygous deletions, focal amplifications or deletions, and for the *FZD10* gene significant differential expression. Frequencies are percent of cases altered in the TCGA series of CRCs. Figure adapted from [70].

Cancer stem cells – a model for tumor heterogeneity

Normal tissues of the human body, such as the colonic crypts, are arranged with a stem cell compartment, with stem cells responsible for renewal and maintenance of the tissue's multiple cell types. Similar to normal tissue hierarchies, tumors are frequently heterogeneous; composed of cancer cells displaying different genetic and phenotypic traits [72,73]. These traits include cell morphology, expression of cell surface markers, genetic lesions, cell cycle entry, proliferation and response to therapy. Further reflecting aberrant tissue organization, only a small proportion of the tumor cells are believed to be able to renew and propagate the tumor, while the bulk of the tumor is composed of cells with a more limited growth potential. The small proportion of renewable cancer cells have been proposed to be cancer stem cells (CSCs)[74]. This explanation for tumor heterogeneity contrasts the original stochastic model of tumor formation, where a succession of genetic changes confers a selective advantage for cancer cells, and tumor formation is a process of clonal expansion [75]. In the CSC model, only cells with the capability of self-renewal and ability to differentiate into tumor propagating cells can be the cell components undergoing selective growth. However, subpopulations of cancer cells with CSC properties may arise from any cell type that gains self-renewal and differentiation capabilities, and not necessary develop from normal stem cells. If correct, the CSC hierarchical model for tumor development has broad implications for how cancer is treated. Most standard therapeutic regimes treat tumors as a homogenous mass of cells and targets rapidly dividing cancer cells. This approach has had beneficial advances in several cancer types, but suffers from high relapse rates, treatment resistance and ultimately the deaths of several cancer patients. Some of the CSC attributes help explain these shortcomings. CSCs have in accordance to normal stem cells been observed to have increased activity of multi-drug resistance transporters. Additionally, it has been hypothesized that CSCs divide more slowly than the bulk tumor cells and therefore do not die by treatment targeting rapidly dividing cells. Anti-apoptotic proteins are also found at increased levels in CSCs, making it harder to induce treatment-related apoptosis [76]. Therefore, in traditional treatment, even though the bulk of the tumor consisting of more specialized cancer cells are killed and the patient enters remission, the CSCs may survive and recapitulate the tumor at a later time. One intriguing example is CML, where a translocation resulting in the Philadelphia chromosome and consequent constitutively expression of the fusion kinase BCR-ABL1 occurs in about 95 % of cases [77]. A therapeutic drug, imatinib, has been developed that effectively block the constitutively active kinase [78]. As a result, what seems to be complete remission is almost always observed. However, if the patients are taken off the imatinib, the CML relapses in many of the cases [79,80]. This phenomenon has been attributed to the fact that the CSCs survive and recapitulates the tumor once the treatment is removed [81]. Development of therapies that specifically targets the CSCs is promising if the theory holds true. However, it is of importance that such agents directed against CSCs discriminate between CSCs and normal tissue stem cells [76].

Still, the CSC model remains controversial. Traditional characterization of CSCs from tumors has been cell sorting followed by xenotransplantation into immunodeficient mice. Based on cell surface markers, subpopulations of cancer cells have been identified that are able to repopulate a tumor and its heterogeneous cancer cell populations. This characterization of CSCs has been argued to isolate cancer cells with stemness properties in very specific model systems and only provide a snapshot of the tumor at the time of isolation. It is questionable if the isolated cancer cells with stemness properties are stable CSCs, or if it is possible that cancer cells fluctuate between CSC and non-CSC states. Only if the CSCs are a fixed subpopulation of the cancer cells will therapeutic strategies that target them prove valuable [82,83].

Identification of new markers that specifically distinguish CSCs from normal stem cells is needed to develop new therapeutic strategies that target the CSCs. Isolating and purifying the CSCs and the normal somatic stem cells and subsequently performing comparative studies between the two cell types may identify such markers. However, total purification of CSCs is difficult, as most of current strategies only enrich cells with stemness properties. Consequently, new markers that can be used to isolate and enrich CSCs specifically are also needed. This becomes a chicken and the egg paradox, as new markers are needed for isolation of CSCs to be able to compare them to normal stem cells to subsequently identify CSC specific markers in a stem cell setting. An alternative approach may be the use of model systems for CSC to normal stem cell comparison. Here, utilization of germ cell tumors is promising. Embryonal carcinoma (EC) cells of the germ cell tumors are capable of self-renewal as well as differentiation into the more differentiated subclasses of TGCT non-seminomas. They are remarkably similar, both on the level of gene expression and pluripotency to embryonic stem (ES) cells. The latter are derived from the inner cell mass of the blastocyst stage embryo and are capable of differentiating into any cell type in the body. When cultured *in vitro* through extended passages, ES cells have been shown to acquire genetic changes that are similar to changes found in malignant EC cells, including gain of material from chromosomes 12, 17, and X [84]. Considering this, EC cells are commonly considered a malignant caricature of ES cells, where the progressive adaption of ES cells *in vitro* reflects malignant TGCT transformation *in vivo* [85]. The EC-ES pair thus allows unique comparison of a normal, or at least non-malignancy-derived, stem cell to its malignant counterpart. Genetic markers that specifically segregate EC cells from ES cells may have promise for the understanding of the CSC concept, and also be of value for further therapeutic intervention [86].

Testicular Germ Cell Tumors

Cancer of the male germ cells, located in the testes of men, is the most common cancer type among young males aged 15-44 years (Figure 8A) [87]. TGCT accounts for 95 % of all occurring malignancies of the testes. The other 5 % includes tumors arising from epithelial supporting cells, such as Leydig and Sertoli cells. TGCT is most common amongst Caucasian men, and TGCT incidence has drastically increased in the western developed world in the last 50 years [88–90]. Especially, Scandinavia has a high incidence of TGCT and in Norway the incidence has tripled during the last 50 years (Figure 8B).

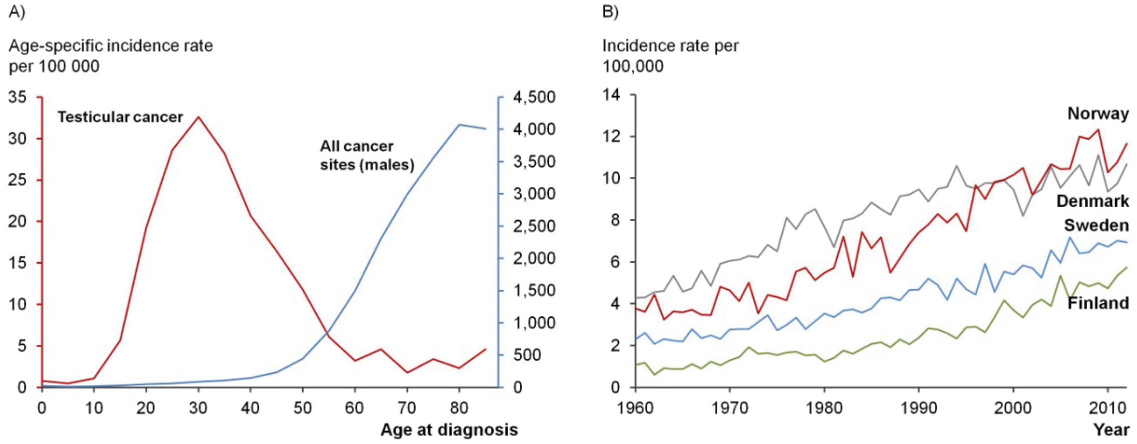


Figure 8: Incidence rates of testicular cancer. (A) Age-specific incidences of testicular cancer compared with other cancer types, 2009-2013. The raw data were obtained from the Norwegian Cancer Registry [52]. **(B)** Age-standardized testicular cancer incidence rates in Nordic countries, 1960-2012. The raw data were obtained from Nordcan [91].

Treatment of TGCT is looked upon as one of the success stories of cancer treatment. Mortality rates were high back in 1970, with 95 % of men with metastatic TGCT dying of their disease [92]. With the introduction of combinatory cisplatin chemotherapy the relative 5-year survival rate sky-rocketed, evident in the total survival of 97 % in Norway during 2009-2013 [52]. Still, TGCT affects men in their prime, and can substantially increase treatment related morbidity later in life, including cardiovascular disease, reduced fertility and secondary cancers [93]. All TGCTs are believed to arise from a pre-invasive stage termed intratubular germ cell neoplasia (IGCN), also known as carcinoma *in situ* [94]. IGCN is frequently observed adjacent to invasive TGCTs and further support the development of invasive TGCTs [95]. These pre-invasive lesions are thought to arise during fetal development and involve changes to primordial germ cells (PGC) either during migration to the embryonic genital ridges, or after they have arrived at the gonads. Malignant transformation is then thought to occur later in life, post-puberty [96]. TGCT can further be divided into several subgroups. Seminomas are tumors with cancer cells that have similar features as the PGC. Non-seminomas are further divided into pluripotent ECs and more differentiated subtypes, with either somatic (teratoma) or extra-embryonic differentiation (yolk sac tumors; YST, and choriocarcinomas) [97]. Both seminomas and non-seminomas are thought to arise by the same pathway, as up to 50 % of TGCTs harbor mixed histological subtypes [97].

Genetics of testicular germ cell tumors

The mutation rates (point mutations and small insertions and deletions) of TGCTs are low and similar to those of pediatric cancers [98–100]. Even so, the genomes of TGCTs are generally aneuploid with extensive chromosome instability and have erased paternal imprinting. Non-seminomas are typically hypotriploid, while seminomas are hypertriploid [101]. Gain of the chromosome arm 12p is seen in virtually all cases of TGCT, with 80 % harboring an isochromosome (i12p), while the remaining cases have gain of parts of 12p material and/or extra copies of the whole of chromosome 12 [102,103]. Gain of 12p material is observed in some IGCNs, but only in IGCNs that are adjacent to a germ cell tumor [104]. This suggests that gain of 12p is not a requirement for IGCN development, but that it is associated with malignant transformation [105]. Other chromosome copy number alterations also occur with high frequencies, and genes located within these regions of recurrent chromosomal changes are thought to be important in initiation and development of TGCTs [106]. No clear genetic driver located on 12p has been identified, but some candidates have been presented [105]. Several genes that are located on 12p are overexpressed in TGCTs compared to normal testicular tissue. *CCND2* (encoding cyclin D2) has been shown to be expressed in the G1 phase of the cell cycle in IGCN and most invasive TGCTs. Cells lacking cyclin D has been shown to be less prone to undergo malignant transformation [107]. Other candidate genes located on 12p include *KRAS*, one of the most frequently activated oncogenes, with 17-25 % of all human tumors harboring activating mutations in the gene [108]. Also stem cell specific genes, including *NANOG* and *DPPA3* are located on 12p and are considered as candidate genetic drivers to the cancer development. Even though point mutations are infrequent in TGCTs, some recurrent mutations are observed. These include mutations in the *KIT* gene encoding a receptor tyrosine kinase, observed in about 25 % of seminomas, but not in non-seminomas [109,110] and *KRAS* [100,111]. Genome wide association studies (GWAS) have also identified several loci where particular alleles of single nucleotide polymorphisms (SNPs) are associated with increased risk of developing TGCT. The locus with the strongest odds ratio is located within chromosome band 12q22, close to the *KITLG* gene. KITLG encodes the ligand binding to the KIT receptor [112,113]. KIT-KITLG signaling is important for PGC proliferation, migration to the gonadal ridge during embryogenesis and survival of these cells. The predisposing *KITLG* genotypes and activating mutations of the *KIT* receptor gene may be responsible for alterations in KIT-KITLG signaling, leading to disruption of normal PGC development and subsequent onset of TGCT initiation and development [114]. Mutations in *KIT* have also been indicated to have a higher incidence in bilateral TGCTs, being a potential marker for bilateral disease. Identification of identical *KIT* mutations has also suggested that transformation of PGCs to IGCN occur during or before migration of PGC and separation of cell populations to the genital ridges [96]. However, recent exome-sequencing of bilateral TGCTs did not identify any overlapping mutations between the bilateral TGCTs, suggesting independent development lineages of bilateral TGCT [98].

Cancer biomarker discovery with high-throughput RNA-analyses

A biomarker can be defined as any characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention [115]. Tumor biomarkers that can be used for early detection, diagnosis, prognosis, treatment targets and also predict treatment response are in demand across the field of oncology. Several biomarkers are already implemented in clinical use, however many of these are measurements of characteristics that are present in normal condition but overrepresented in types or subsets of cancer. Often overexpression of these characteristics are also not specific to the cancer subtype but may be observed in other benign and malignant conditions. Therefore many implemented biomarkers suffer from poor sensitivity and specificity. An example of a marker for screening purposes and early detection is the prostate specific antigen (PSA) used to indicate prostate cancer in men. However, an elevated PSA level does not necessarily indicate aggressive prostate cancer, and PSA testing can lead to overtreatment with questionable effects on mortality rates [116]. Fusion genes and also transcript variants can be highly specific cancer molecules and represent a potent class of biomarkers. One of the best cancer biomarkers in clinical use is the Philadelphia chromosome and the consequent expression of a *BCR-ABL1* fusion transcript which in turn is translated into a BCR-ABL1 fusion kinase. Detection of the chromosomal rearrangement t(9;22)(q34;q11) is used as a diagnostic biomarker for CML. Further, expression of the *BCR-ABL1* fusion transcript is used as a biomarker to monitor disease and treatment response [117,118]. Additionally, in 2001 the US Food and Drug Administration (FDA) approved imatinib (Gleevec), a drug blocking the fusion kinase BCR-ABL1 that was shown to be highly effective in treating CML patients [78,119]. Later, several additional new compounds have been developed that block BCR-ABL1 and also other fusion kinases in a similar manner. Examples include the use of ALK-inhibitors crizotinib and ceritinib in treatment of non-small cell lung cancers with rearrangements in *ALK* and the use of imatinib for cancers with *PDGFRA/B* rearrangements [24]. Drugs targeting non-kinase fusions have also been found to be effective, e.g. with the use of Tretinonin (all-*trans* retinoic acid) in treatment of acute promyelocytic leukemia with the *PML-RARA* fusion gene [120]. Recently, a study also found that a small-molecule inhibitor could inhibit a transcription factor fusion in acute myeloid leukemia, providing proof of concept for transcription factor targeted therapy [121]. The splicing process of primary transcripts is frequently altered in cancer compared to normal tissue due to differential expression of known splicing-regulatory genes. Such aberrant splicing can produce cancer specific transcript variants that have biomarker potential [39].

Before high-throughput methodologies, gene detection, SNP characterization, and investigation of structural transcript variation have been carried out by narrow poking of the transcriptome. Specifically, analyses of expressed sequence tags (ESTs) have mostly been used. ESTs are generated by the conversion of mRNA copies into complementary DNAs (cDNAs), which are then cloned to make libraries of transcribed genes from the original cell, tissue or organism. These clones are subsequently sequenced randomly from both directions in a single-pass run. The ESTs usually range in size from 100 to 800 bp. The ESTs are sequenced only once and are prone to sequencing errors, especially at the ends. The amount of ESTs is not directly quantitative and the sensitivity is poor with frequent sampling bias resulting in under-representation of rare transcripts [122]. Comparative hybridization techniques and microarrays have been revolutionary in the field of transcriptomics. The ability to simultaneously interrogate and measure the expression of hundreds of thousands of exons and transcripts has brought about extensive knowledge of the cancer transcriptome. However, microarray technologies are limited by the fact that they are based upon previous knowledge of the transcriptome, and can therefore be considered a guided technology. Still, microarrays are efficient tools for analyzing gene expression and alternative splicing of known transcript variants in cancer. By investigating gene expression in CRC using microarrays, prognostic gene expression signatures have been discovered that promise better risk assessment for stage II and III CRC patients [123]. Also, by using exon microarrays, a novel characteristic of transcriptome instability, characterized by extensive aberrant genome-wide splicing, has been described in CRC and other carcinomas [124,125]. Outlier expression analysis with microarray data has identified recurrent fusion genes in solid epithelial carcinomas, such as the *TMPRSS2-ERG* fusion gene and other ETS rearrangements in prostate cancer [26]. The feasibility to detect known fusion genes has further been demonstrated by development of custom fusion microarrays, with probes spanning potential fusion transcript breakpoints [126–128]. In contrast to microarray technologies, high-throughput, or deep sequencing of cDNA libraries (also termed RNA-sequencing or transcriptome sequencing) determines the cDNA sequence directly and enables complete characterization and quantification of the transcriptome [129]. Here, cDNA is sequenced by deep-sequencing protocols and machines, generating millions of sequencing reads with lengths depending on the applied sequencing technology. Each base of an expressed transcript is covered hundreds or thousands of times, and differences in read densities allow for robust quantification of gene- or transcript-level expression. Sequencing reads can also be mapped across exon-exon junctions, enabling identification and characterization of known as well as novel transcript variants. By sequencing from both ends of cDNA fragments, in paired-end sequencing, mapping information of the two paired sequencing reads can be used for additional information of transcript structures. Fusion gene detection algorithms often rely on the discordant mapping of sequencing reads in a read-pair, followed by detecting reads that cover the fusion breakpoint in itself. This approach enables the unbiased identification of fusion transcripts, involving the fusion of known or previously unknown transcript elements. Additionally, semi-guided RNA-sequencing can uncover the transcriptional complexity of targeted genes, as is explored in paper II of this thesis with the coupling of 5’ RACE and deep sequencing of full length transcripts. The era of deep-sequencing and especially RNA-sequencing have already uncovered that the transcriptome is extensively more complex than originally anticipated, exemplified by the explosive increase in identified fusion transcripts [24], which are common in both healthy and diseased cells and tissues. Careful experimental designs and post-sequencing filtering strategies are required to remove technical artifacts and enrich for functionally relevant transcript variants. The technology will continue to provide insights and discoveries of the transcriptome, which will help uncover mechanisms of human complexity and pathologic disease such as cancer.

Aim

The overall aim of the PhD-project was to reveal novel RNA variants expressed specifically in cancer, which have potential as biomarkers or therapeutic targets and illuminate aspects of the transcriptome in cancer biology. Specifically, we aimed to identify and characterize novel fusion transcripts and other transcript variants expressed in CRCs and TGCTs.

Results in brief

Paper I

‘High frequency of fusion transcripts involving *TCF7L2* in colorectal cancer: Novel fusion partner and splice variants’

In 2011, *VTI1A-TCF7L2* was described as the first recurrent fusion gene in CRC, expressed in a total of 3 out of 97 primary tumor samples, and in the NCI-H508 CRC cell line [71]. In Paper I, we investigated if *VTI1A-TCF7L2*, or other fusion genes involving either of the original partner genes, were detected in RNA-sequencing data from seven CRC cell lines. In addition, we had overlapping whole-genome sequencing data from four of these, namely HCT15, HCT116, HT29 and SW480. By using the fusion finder algorithms deFuse and nFuse we only detected the *VTI1A-TCF7L2* fusion breakpoint in downloaded RNA-sequencing data from the NCI-H508 CRC cell line, as already reported by Bass *et al.* [71]. However, we detected a genomic breakpoint in the HCT116 cell line that spanned from the intronic region of *TCF7L2* to upstream of *RP11-57H14.3* with a corresponding expressed fusion transcript. Further, we investigated the prevalence of both the original *VTI1A-TCF7L2* fusion transcript and the *TCF7L2-RP11-57H14.3* fusion in a series of 106 CRC primary tumors and a panel of CRC cell lines and normal colonic mucosa samples. By nested RT-PCR, we detected expression of these fusion transcripts in 42 % and 45 % of the CRCs respectively. Additionally, both fusion transcripts were seen in 29 % of the normal colonic mucosa samples and in several of the different normal tissue samples from miscellaneous human organs (n = 20). The two fusion transcripts were not expressed in a mutually exclusive manner, and importantly, not exclusively in cancer cells.

Paper II

‘Novel RNA variants in colorectal cancers’

To further characterize fusion genes and transcript variants in CRC, we analyzed exon-level microarray data from 202 CRCs and searched for genes with overexpression of the 3’ end in individual tumors. In total, 25 genes showed overexpression of the 3’ end in at least one cancer sample. To effectively characterize the underlying transcript structures of these candidates, we developed a new sequencing approach, RACE-seq. Here, we combined traditional 5’ RACE of the target genes and additional positive control genes with high-throughput sequencing of pooled RACE products. We identified *VWA2-TCF7L2, DHX35-*

BPIFA2 and *CASZ1-MASP2* as private fusion events in individual tumor samples. In addition, we characterized novel transcript structures for 17 of the 23 other targeted candidate genes, where 13 of these had sequence reads extending upstream from the annotated gene boundary, and 12 had sequence reads supporting new intragenic transcript structures. Eight of the 23 genes had both; reads extending upstream and reads supporting new intragenic transcript structures. Additionally, we found transcript junctions supporting a recurrent read-through fusion transcript between *KLK8* and *KLK7,* and also a novel 3' splice site in *S100A2.* Both were found to be overrepresented in CRC tissue and cell lines from external RNA-seq datasets from TCGA and the Cancer Cell Line Encyclopedia (CCLE).

Paper III

'RNA sequencing reveals fusion genes in testicular germ cell tumors'

To investigate the possibility of driver fusion genes in TGCTs, we performed RNA-sequencing of 3 EC and 2 ES cell lines. We identified and validated 8 fusion transcripts in addition to alternative promoter usage in transcription of the *ETV6* gene. Five of these nine breakpoints had both partner genes located on chromosome arm 12p. By using droplet digital PCR (ddPCR) to assess gene linkage, we found that the private fusion genes; *EPT1-GUCY1A3* and *PPP6R3-DPP3*, are linked on the genetic level in the 833KE and NTERA2 EC cell lines, respectively. *RCC1-ABHD12B, RCC1-HENMT1, CLEC6A-CLEC4D* and the alternative promoter of *ETV6* were found to be recurrently expressed in an extended sample panel, including all subtypes of primary TGCTs and additional EC cell lines. *RCC1-ABHD12B* and the alternative promoter of *ETV6* were found to be expressed more frequently in the less differentiated subtypes of TGCTs. *In vitro* treatment with all-*trans* retinoic acid (RA) induced differentiation of NTERA2 EC cell line which resulted in significant reduced expression of both fusion transcripts involving *RCC1* and the alternative promoter usage in *ETV6.* All 4 recurrent transcripts were undetectable in normal parenchyma of the testis. In addition, expression of *RCC1-ABHD12B* and the alternative *ETV6* promoter were not detected in any of the included normal tissues from organs of the human body (n = 20).

Discussion

The concept of the gene – in need of refinement

In all three papers presented in this thesis we discovered transcripts that have not previously been annotated and that have transcribed elements extending outside of currently annotated gene boundaries. Our findings are in line with reports from the Encyclopedia of DNA Elements (ENCODE) consortium, whose goal is to build a comprehensive parts list of functional elements in the human genome. They have previously reported that up to three-quarters of the genome are transcribed, and that the reduced lengths of intergenic regions lead to a significant overlap of neighboring gene regions [32]. In the pilot project of ENCODE, more than 80 % of 399 protein coding genes had unannotated transcribed fragments upstream or internal to previously annotated gene boundaries [130]. Later, they discovered that 85 % of transcriptional termini, both 5' and 3' ends, of 492 protein coding genes extended beyond the annotated gene boundaries [41]. These findings clearly contradict the notion that a gene is constricted to a defined set of genomic coordinates. ENCODE have suggested a redefinition of a gene as being 'a union of genomic sequences encoding a coherent set of potentially overlapping functional products' [131]. This definition weighs the functional product as the defining factor. However, the extensive complexity of the transcriptome and the increasing overlap between annotated genes are difficult to fit into the definition. Djebali *et al.* have proposed that the transcript should be considered as the unit of inheritance, and that the gene could be considered a higher-order concept intended to capture all transcripts that contribute to a given phenotypic trait [32]. In conclusion, the extensive complexity of the transcriptome revealed by high-throughput techniques in the past decade complicates the concept of a gene. Further understanding of splicing mechanisms and cross-talk in transcription between genetic loci is necessary to understand developmental biology as well as pathological processes which are linked to our genes, such as cancer.

The generation of transcripts containing exons spliced in a non-linear order, as is the case for *TCF7L2-RP1157-H14.3* (paper I) has been described previously by Nigro *et al.* as a process of exon-scrambling [132]. The process was first described when investigating the tumor suppressor gene *DCC,* where scrambled transcripts were found to be expressed in both normal and neoplastic cells. With RNA-sequencing such scrambled transcripts from hundreds of genes have recently been described [133]. These scrambled transcripts were suggested to be results of expressed circular RNAs, many of them expressed at a level comparable to their canonical linear counterpart.

Both *VTI1A-TCF7L2* and *TCF7L2-RP11-57H14.3* fusion transcripts investigated in paper I were found to be expressed, although often at low levels, in a high frequency of normal and malignant samples without corresponding genomic rearrangements. The observation that fusion transcripts expressed in normal cells without genomic aberrations are overexpressed with genomic aberrations in cancer cells is intriguing. The chance of recurrent genomic aberrations in cancer resulting in the expression of fusion transcripts that are identical to those that are also expressed at low levels in normal cells without corresponding chromosomal aberrations is miniscule. This suggests that the generation of fusion transcripts and introduction of corresponding genomic breaks is somehow a linked mechanism [43]. Probably, the spatial organization of the genome in the nucleus plays an important role in this process, as elements far apart in linear sequence may not necessarily be far apart in the three-dimensional nucleus space. The organization of the genome in the nucleus has been shown to be nonrandom, and follow patterns for different tissue types [134–136]. This may explain the observations of tissue-specific and cancer-specific recurrent chromosomal rearrangements, e.g. the expression of the fusion transcript *JJAZ1-JAZF1* in endometrial cells and the consequent chromosomal rearrangement seen in endometrial stromal tumors [51]. Investigating the spatial orientation of the genome may uncover more secrets about the long-distance interactions of the genome, *trans*-splicing and corresponding genomic rearrangements. Techniques for investigating the genomic spatial arrangement have been rapidly developing in recent years. Genomic chromosome confirmation capture followed by sequencing (HI-C sequencing) allows investigation of chromosomal interaction patterns across the whole-genome. This approach has been used to show that translocation breakpoints have higher contact frequencies than random parts of the genome [137,138].

Fusion transcripts and transcript variants as cancer biomarkers

To identify new cancer biomarkers, it is possible to look at all levels of biological molecules; from metabolites found in serum, proteins, DNA mutations, epigenetic alterations, non-coding RNAs, coding mRNAs, etc. However, mRNAs represent an especially potent source for biomarker discovery. In contrast to proteins and metabolites, all mRNA can be characterized and quantified simultaneously using high-throughput RNA-sequencing. In contrast to alterations at the DNA level, changes in mRNAs are alterations that are actually expressed and can have a direct impact on the cell. Additionally, mRNA molecules that are overexpressed or specific in cancer often have thousands of copies, enabling high sensitivity detection of such markers compared to cancer specific mutations at the DNA level that are present in one or two of the DNA alleles per cancer cell. In addition, sequencing analyses of the entire genome requires considerably more power and sequencing-reads, as the mRNA, or protein coding part of the transcriptome only covers about 2 % of the genome.

No matter how fusion transcripts or transcript variants are generated, or if and how they have functional consequences for cancer, as long as they are accurate in their task with high specificity and sensitivity, these molecules may be potent cancer biomarkers. With the advent of high-throughput RNA analyses methods, we can rapidly and efficiently screen the transcriptome for cancer-specific transcript variants and fusion genes. However, detection of fusion transcripts and transcript variants that are specific and highly recurrent in cancer is challenging due to the high degree of instability seen both in the genome and transcriptome of solid cancers. These phenomena generate a myriad of fusion genes and transcript variants that are private to individual cancers. Still, findings of highly recurrent fusion genes occur, exemplified by the finding of a 400 kb deletion on chromosome 19 in fibrolamellar hepatocellular carcinoma resulting in a fusion transcript, *DNAJB1-PRKACA,* that is expressed in all investigated cases of the disease [139,140]. The fusion transcript is also not expressed in normal samples or other hepatocellular carcinomas making it a highly potent diagnostic biomarker [140].

Fusion transcripts and transcript variants in CRC and TGCT

By utilizing high-throughput RNA analyses approaches we discovered several novel transcript variants and fusion transcripts expressed in CRC and TGCT. In paper I, we discovered that both *VTI1A-TCF7L2* and *TCF7L2-RP11-57H14.3* are expressed in normal samples both from the colon and other miscellaneous tissue types, diminishing their potential as cancer detection biomarkers. However, genomic rearrangements involving *TCF7L2* found in a small subset of CRCs and CRC cell lines can have functional consequences as demonstrated by Bass *et al.* for the original *VTI1A-TCF7L2* fusion [71]. Furthermore, we identified a read-through between *KLK8* and *KLK7* and also alternative 3' splice site usage in *S100A2* that were both found to be overrepresented in CRC *vs.* normal (paper II). This is of interest since both these genes have previously been implicated in CRC [141–145]. However, although our results point to potential biomarker value, validation studies in a larger series and correlation with clinical data is needed to establish if these transcripts can be used as biomarkers, either in aiding diagnostics, or stratification and prognostics. Several fusion transcripts in addition to alternative promoter usage of *ETV6* were identified in TGCT (paper III). The read-through *CLEC6A-CLEC4D* was seen to be expressed in a TGCT specific manner, with the exception of normal placenta. Additionally, the fusion transcript *RCC1-ABHD12B* and the alternative promoter usage in *ETV6* transcription were TGCT specific, but mostly expressed in the undifferentiated seminoma and EC subtypes. These identified recurrent transcripts could have functional roles in TGCTs and among them are the first fusion genes to be described in this malignancy. The specificity of these transcripts could prove to be important as clinical biomarkers for TGCT, as also novel biomarkers are warranted in TGCT [146].

Methodological considerations

Several RNA characterization methods have been used in this thesis. For genome-scale analyses, high-throughput RNA-sequencing (Papers I, II, and III) and exon microarrays (Paper II) have been used. These methods and associated wet-lab protocols used in the present thesis have been developed by commercial vendors, and are not discussed further here. Two wet-lab methods were developed more specifically for our projects, pooled/multiplexed RACE-sequencing for upstream fusion partner identification and digital PCR for establishment of RNA-transcripts which are associated with DNA-level linkage. Additionally, computational considerations and challenges of the applied bioinformatic pipelines, aspects of different sequencing technologies and machines, and validation of findings in external sequencing data will be discussed in the following chapters.

RACE-sequencing

RACE is a modified PCR technique used to specifically amplify and characterize the sequence of the full length ends of a transcript of interest [147]. Traditionally, characterization of RACE fragments has been carried out by cloning, plasmid isolation and subsequent sequencing of a few clones by classic Sanger sequencing. This approach is time-consuming, has a high cost, suffers from low sensitivity and is often not very successful. Presented in this thesis, we developed a protocol for high-throughput characterization of RACE fragments, where pools from several samples with RACE-fragments from multiple assays are sequenced in a single high-throughput sequencing run. In the approach of paper II, we used RACE-sequencing to characterize the underlying 5' parts of transcripts of genes with overexpressed 3' parts in CRC, as nominated by exon microarray data. However, the RACE-sequencing approach also has other potential applications, such as testing for known fusion breakpoints within a clinical cohort or across tumor types. The method enables ultra-deep read coverage of the targeted sequences and should be sensitive in detecting any transcript breakpoints. In hindsight, some improvements to the pipeline used in paper II can be suggested. First, when designing primers for the RACE assays, they should be located at sufficient distance downstream (or upstream if using 3' RACE) of the suspected transcript breakpoint. In our setup the primers were designed too close to expected breakpoints for several of the assays, resulting in no read coverage of the downstream part of suspected breakpoints. For the identified novel fusion *CASZ1-MASP2*, the target assay for *MASP2* was designed only 25 bp downstream of the identified fusion breakpoint. Consequently, no full-length read (150 bp) aligned to the *MASP2* part of the fusion. The fusion was therefore not nominated by the fusion detection algorithm deFuse, but was found by using the Unix command-line utility grep for the primer sequence. Although we discovered the *CASZ1-MASP2* fusion despite of this shortcoming, we may still have missed other underlying transcript breakpoints. Second, isolation or enrichment of mRNA should be performed prior to RACE cDNA synthesis. Several of the genes with the highest number of reads in paper II were clearly untargeted genes, including several mitochondrial RNA genes such as *MT-RNR2* and *MT-CO1* (see supplementary figure 2 in paper II). Probably these genes are highly expressed in the cells and are sequenced even if they are not enriched for.

During development of the RACE-sequencing protocol, a few other commercial solutions have entered the market that uses some similar principles to combining RACE and sequencing. However, both the Ovation® fusion panel and the FusionPlex™ provided by NuGEN technologies (San Carlos, CA, USA) and ArcherDX (Boulder, CO, USA), respectively, target already fragmented cDNA targets and amplify and sequence short stretches around suspected breakpoints. Therefore, multiple assays are needed if there are several potential breakpoints in a target gene. However, this approach may be better suited for detecting known breakpoints of fusion gene partners, but will not be able to characterize full-length transcripts and transcript-variants.

Droplet digital PCR for establishing gene linkage

To establish if a fusion transcript is caused by a corresponding genomic rearrangement, investigation at the DNA level is necessary. Since the genomic breakpoints creating fusion genes most often occur in intronic or intergenic segments, predicting the exact location of the genomic breakpoint is difficult. The use of cytogenetic methods or FISH is a long lasting standard for detecting chromosomal rearrangement that causes a suspected fusion gene. In FISH, fluorescent probes that bind DNA can be designed to the two partner genes that participate in a fusion transcript. The co-localization of the two fluorescent signals is an indication of a chromosomal rearrangement. Long-range PCR is also a possibility with high-fidelity DNA polymerases that are able to amplify up to 15 kb or longer genomic DNA. However, getting long-range PCR to amplify such long stretches requires tedious optimization of reaction conditions and primers. In paper III, we utilized ddPCR linkage analysis to investigate if the two partners of a fusion gene are linked on the same DNA molecules. This approach takes advantage of the oil/water emulsion of PCR reagents and DNA template into thousands of nano-liter sized droplets. By designing duplex PCR assays with different fluorescent probes (FAM and VIC/HEX), it is possible to interrogate the genomic loci upstream and downstream of the fusion transcript breakpoint simultaneously. Since template molecules distribute randomly into droplets, droplets can be expected to contain one or the other, both or none of the target molecules by chance in a multiplexed assay. However, if the two template targets are located in close proximity, and located on the same DNA molecule as a result of a chromosomal rearrangement, these would distribute together in a non-random fashion with a higher number of double positive droplets (Figure 9).

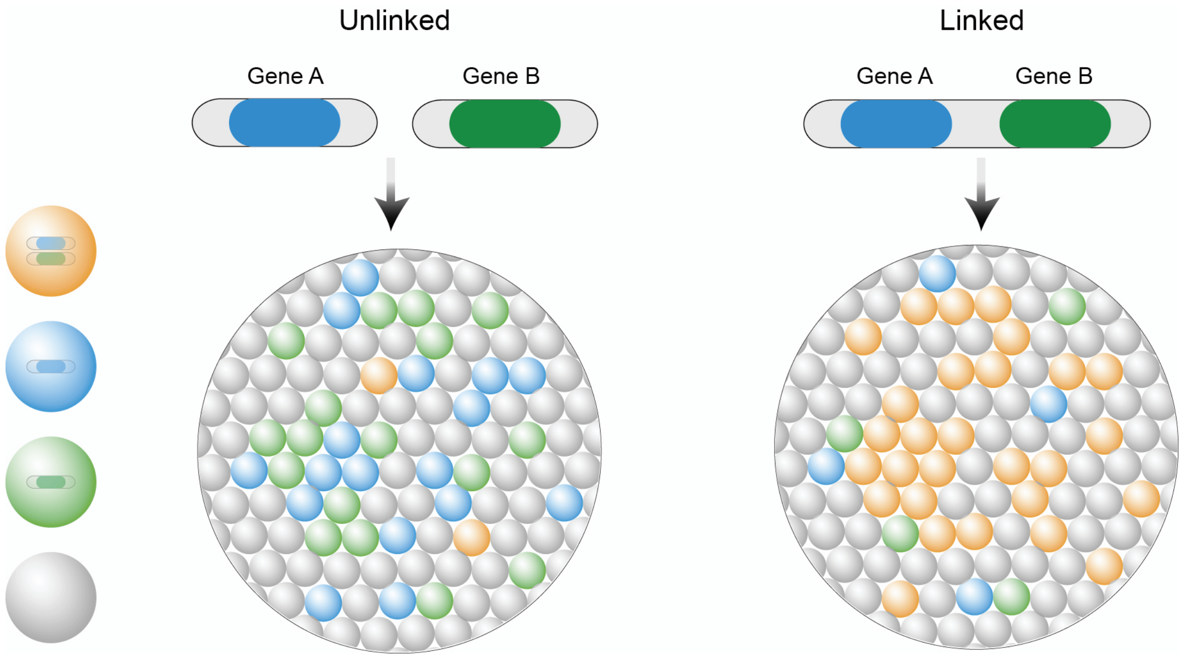


Figure 9: ddPCR linkage analysis. By interrogating two genes or genomic loci with duplex ddPCR assays with different fluorescent probes it is possible to establish if the two genes are linked at the DNA level. If the two genes are unlinked their template distribute randomly into droplets resulting in a high fraction of single-template positive droplets (green and blue) and few double-positive droplets (orange). However, if the two genes are linked and located on the same template they distribute together non-randomly into droplets, resulting in a higher fraction of double-positive droplets. The fraction of double-positive droplets will depend on the distance between the genes on the templates and the amount of DNA degradation.

To our knowledge, this approach has not been used previously to detect rearrangements of genes resulting in fusion genes. However, linkage analysis with ddPCR has been proven successful in showing the arrangement of the Killer-cell immunoglobulin-like receptor gene complex and in chromosomal phasing [148,149]. In paper III, we used the technique to identify genetic linkage between the partner genes of *EPT1-GUCY1A3* and *PPP6R3-DPP3* indicating genomic rearrangements causing the fusion genes in their respective EC cell lines, 833KE and NTERA2. However, we found no indication of genetic linkage for the partner genes of the highly recurrent *RCC1* fusions. These may therefore be expressed as a result of *trans*-splicing or some other splicing related mechanism. However, if large introns (e.g. introns larger than 50 kb) are involved at the fusion breakpoint, we cannot rule out chromosomal rearrangements as an underlying cause of the *RCC1* fusion transcripts. This is because the input DNA in our study contains only a low fraction of DNA fragments longer than 50 kb and none detected with 100 kb length (as shown by the "milepost" ddPCR control assay in paper III – supplementary figure 2). A chromosomal rearrangement resulting in a fusion gene may include intronic regions that are longer than these DNA fragments. Such rearrangements will therefore be missed by ddPCR linkage analysis. As an improvement to the ddPCR linkage approach, separate DNA analysis protocols should be established that enable the isolation of intact long DNA molecules. Regan *et al.* investigated chromosomal phasing using a protocol for gentle DNA isolation with a polysaccharide precipitation-based chemistry (PrepFiler Forensic DNA Extraction Kit, Life technologies; Carlsbad, CA, USA) [149]. This approach enabled detection of gene linkage of assays up to 210 kb apart, which should be satisfactory for detection of chromosomal rearrangements causing fusion genes. In conclusion the ddPCR linkage method is a technique requiring little optimization and represents a fast alternative to establish if two originally distant genomic loci are linked on the chromosome level. Still, whole-genome sequencing is the gold-standard for investigating if a fusion transcript is caused by a genomic rearrangement as it is used not only for detecting the genomic rearrangement, but also for characterizing the exact location and sequence of the breakpoint. This was demonstrated by detecting *TCF7L2-RP11-57H14.3* in paper I by a combination of RNA-sequencing and whole-genome sequencing data. However, whole-genome sequencing requires extensive library prep and subsequent data analysis, and is an expensive way to investigate the presence of a single chromosomal rearrangement for a candidate fusion transcript, in particular when many samples are to be investigated.

Bioinformatics and sequence analysis pipelines

Sample preparation and sequencing of DNA or RNA libraries on deep sequencing machines have become routine processes and are rarely the most time-consuming parts of the project. However, the bottle-neck for such high-throughput experiments is often to establish a pipeline for analyzing the resulting millions of sequencing reads. This is especially true for fusion transcript and transcript variant detection as there is no point-andclick way of getting out the data you are interested in. In general, analysis pipelines of sequencing reads starts with quality control followed by trimming or removal of bad-quality reads if necessary. In our studies we used the FastQC software for quality control [150]. In paper II, we discovered that several of the sequencing reads generated by the RACE-sequencing approach contained the SMARTer II adapter oligos that were adhered during RACE cDNA synthesis. These adapters were removed using the cutadapt software that also removed low quality reads [151]. Following quality control, sequencing reads are aligned to the reference human genome. Since the transcriptome is incompletely characterized, RNA sequences are aligned to the genome but allowing for spliced alignments since the reference genome includes intronic and intergenic sequences. This is a challenging task and there are several algorithms that map sequencing reads in such fashion. In these studies we mostly utilized the TopHat2 algorithm that uses the short-read mapper Bowtie to initially map reads to the genome [152,153]. Based on initial mapping, splice junctions and exon boundaries are nominated and guide the final alignment. TopHat2 was shown to perform reasonably well when benchmarked together with 10 other alignment programs and outperformed the others on detecting true positive splice junctions [154]. In paper II, we used the TopHat2 nominated splice junctions from the RACE-sequencing data and filtered them against already annotated junctions and junctions found in normal tissue samples to identify novel junctions and novel transcript variants in CRC. These junctions were further used as input to generate a gene annotation for alignment of the validation series using the STAR aligner, which is one of the fastest and most accurate aligners when using a guiding gene annotation [154,155].

Detection of fusion genes using RNA-sequencing was first described to be feasible by Maher *et al.* in 2009 [50]. Since then a plethora of computational algorithms that use mapping information to nominate candidate fusion transcripts have been developed. Currently, the OMICtools database lists 25 programs under gene fusion detection with RNA-sequencing analysis [156,157]. In the projects included in this thesis we mainly used the deFuse fusion detection algorithm which uses the TopHat and Bowtie mapping information to nominate fusion transcript breakpoints [158]. However, in paper I, we used nFuse which utilize a combination of both RNA and whole-genome sequencing data to detect fusion transcripts with matched genomic rearrangements [159]. In paper III, we also applied SOAPfuse, in addition to deFuse, to identify high-confidence fusion transcripts which are detected by two unrelated softwares [160]. Fusion breakpoint discovery suffers from a high false positive rate. Reasons for this include the ambiguous mapping of several sequencing reads to multiple locations of the genome due to a high amount of repetitive sequences and homologues sequences from conserved domains and pseudogenes. Additionally, cancer may have such a degree of genomic and transcriptional chaos that correct alignment of reads is rendered difficult. Several of the fusion detection programs have been developed using specific data sets and their algorithms for fusion detection exhibit striking differences in sensitivity and specificity. Since RNA-seq data sets are different, both technically and biologically from different tissue types and disease types, it's hard to foresee a one-size-fits-all approach. Thus, most original papers publishing fusion finder programs describe their algorithm as superior to its competitors [160–165]. No matter which program is used, careful functional and technical filtering is necessary for detecting true positive fusion genes. In general, data sets with more reads and longer reads favor the sensitivity and specificity for detecting true fusion transcripts. With the development of techniques and sequencing chemistry that supports longer-read sequencing, fusion gene detection is anticipated to provide increasingly more trustworthy results.

DNA sequencing machines and deep sequencing technology

With the release of the HiSeq X Ten system, Illumina (San Diego, CA, USA) recently claimed that they have reached the much awaited US $1,000 genome mark. Yet, this price tag is based on the purchase of the HiSeq X Ten system and the sequencing of 18,000 genomes per year [166]. Few laboratories have the volume of samples justifying such capacity. More importantly, however, the price of genome and transcriptome sequencing has plummeted with the explosive technological development and competitive commercial market in next generation sequencing seen over the last decade. In comparison to the now reasonably priced sequencing approaching $1,000, the price-tag of the original human genome was closer to $3-billion. With the advent of cheaper sequencing of the genomes of individuals, the promise of personalized medicine comes ever closer. In addition to sequencing costs, another expensive factor is the computational analysis and interpretation of the sequencing results, requiring extensive computational power, large amounts of data storage and the involvement of an interdisciplinary team of professionals. Illumina is currently the leading provider of high-throughput DNA sequencing machines. The company provides easy to use benchtop sequencers; MiSeq and also recently the NextSeq 500 which are ideal for small sequencing experiments. On the other side of the scale they also provide the big work-horses of producing high-throughput sequences of whole-genomes, namely the HiSeq 2500, 3000, 4000 and the aforementioned HiSeq X ten bundle system. Illumina sequencing technologies are all based on sequencing by synthesis (SBS) where DNA or cDNA molecules are attached to a glass slide, called the flow cell, and amplified to form clusters. Reversible terminator nucleotides that are fluorescently labeled are then added and pictures are taken per cycle of added nucleotide/wash. The SBS technology generally produces short reads, however recent development in sequencing chemistry have enabled at least semi-long reads (2 x 300 bp on the MiSeq).

Other companies are also trying to compete for a share of the market, and several competing systems have been developed and released. Pacific Biosciences' (Menlo Park, CA, USA) RS II enables single molecule real-time sequencing with read lengths longer than 10 kb. The single molecule sequencing avoids the need for amplification which improves coverage uniformity and avoids PCR artifacts. The long reads are ideal for *de novo* assembly of small genomes or for sequencing difficult parts of larger genomes. Additionally, as previously mentioned, long reads are favored when doing full length cDNA characterization and fusion gene detection. However, the RS II's advantages come at a cost of low sequencing output, generating a maximum of 1 Gb per run distributed on approximately 50,000 reads. In comparison, the Illumina HiSeq X ten generates 1.8 Tb of sequencing data per run.

BGI is one of the leading genome sequencing centers in the world with its headquarters in Shenzhen, China. Currently, the sequencing facility in Shenzhen harbors over one hundred HiSeq 2000 machines and also several other sequencing platforms. In 2013, BGI acquired Complete Genomics, a company with their own genome sequencing technology based on combinatorial probe-anchor ligation technology [167,168]. Recently, in June 2015, they announced the Revolocity sequencing system based on the Complete Genomics sequencing technology [169]. The Revolocity currently enables sequencing and semi-automated sample preparation of 10,000 genomes per year at 50x coverage and will expand to 30,000, making it comparable to the HiSeq X Ten system from Illumina. However, with a price tag of US $12,000,000 and short sequencing reads, this system is also aimed at large sequencing factories such as BGI itself.

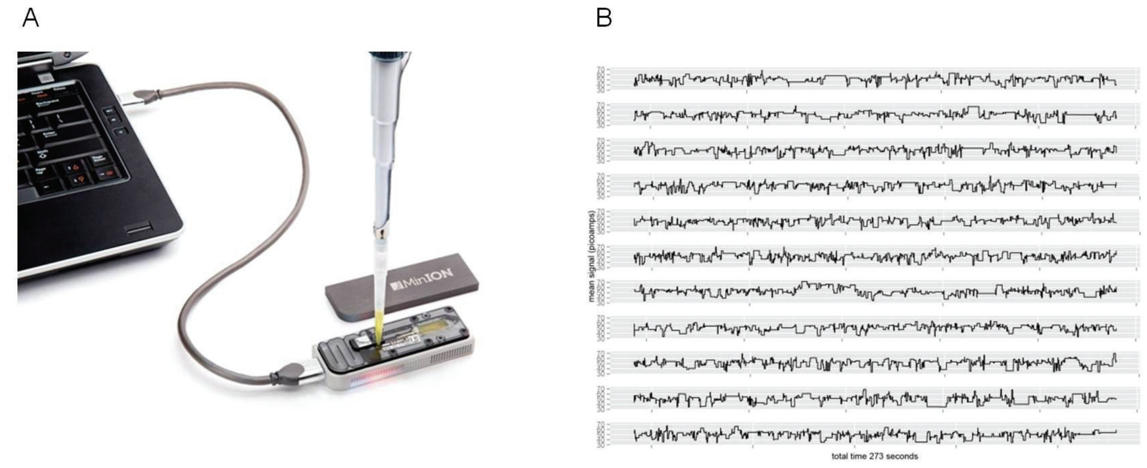


Figure 10: The MinIon nanopore sequencer. (A) The MinIon is about the size of a large USB stick and plugs into the USB port for direct real-time sequencing. **(B)** The first publicly available “raw” read from the MinIon in the form of a wiggle-plot showing the change in current as DNA passes through the nanopore. Distinct current changes are correlated with nucleotide compositions for base calling [170].

One of the most interesting developments in sequencing technologies announced in recent years is nanopore sequencing developed by Oxford Nanopore Technologies (Oxford, UK). This technology also enables sequencing of single molecules, DNA or RNA, without the need for amplification or conversion to cDNA. Nanopore sequencing is based on a protein nanopore embedded in an electrical resistant membrane. A voltage is applied across the membrane and the ionic current passing through the nanopore is measured. When a DNA or RNA molecule passes through the membrane, there is a distinct fluctuation in the current, enabling identification of the particular base, and even modifications such as methylation. In 2012 Oxford nanopore announced the development of the MinIon, a US $1,000 device on the size of a large USB-stick that plugs into the USB-port of a computer and does direct sequencing of DNA or RNA with limited sample preparation and generating longer than ever before read lengths (Figure 10A) [171]. After a 2-year long silence, Oxford nanopore started an early access programme in the spring of 2014, shipping MinIon devices to laboratories with suitable research questions. Although hampered by sequencing errors, the first MinIon sequencing read was shared on twitter by Nick Loman (twitter handle @pathogenomenick) the 11th of June 2014. It is an 8,476 bp long sequence with parts aligning with 68 % identity to *Pseudomonas aeruginosa*, proving its usefulness in diagnostics [170](Figure 10B). Later, application of the MinIon to rapidly provide information on bacterial and viral outbreaks has been demonstrated [172,173]. Although the MinIon presents an exciting development of single molecule sequencing, a drastic decrease in error rates is required to improve its range of use.

Validation of own discoveries in external sequencing data sets

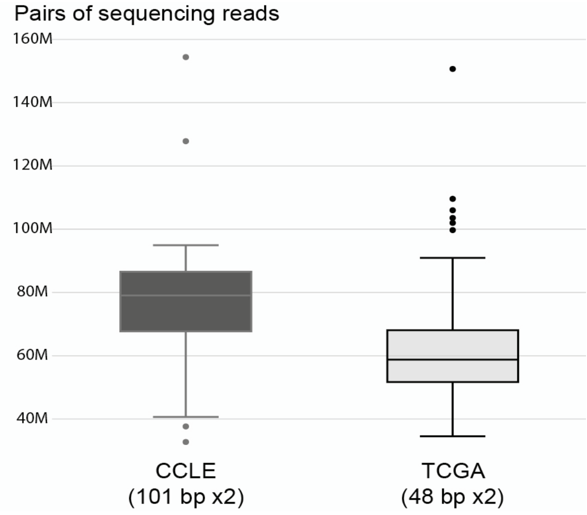


Figure 11: Distribution of the numbers of pairs of sequencing reads from the external RNA-sequencing data used from the CCLE and TCGA in paper II.

To validate the fusion genes and splice junctions found in Paper II, we utilized external RNA-sequencing data from the TCGA and the CCLE from primary CRC samples and CRC cell lines, respectively. However, the three fusion genes that we identified from RACE-sequencing of CRC samples were not found to be recurrent in these data sets. Additionally, the novel splice junctions detected in CRC were mostly supported by few reads compared to that of the RACE-sequencing data. The TCGA data generally have fewer and shorter read sequences (2 x 48 bp) compared with data from the CCLE (2 x 101 bp; Figure 11). This was reflected in a greater number of reads covering the exon junctions investigated in paper II in CCLE compared to TCGA samples. Recently, the TCGA also have deposited RNA sequencing data from 150 TGCT samples, although yet without an associated scientific article. The data can be accessed by approved researchers, and we downloaded these data to validate our findings of the PCR-validated recurrent fusion transcripts in TGCT (paper III). By using deFuse on these RNA sequences we were only able to validate *RCC1-ABHD12B* in 1/102 samples. However, when specifically searching for the fusion breakpoint sequence in the raw read files using grep with a query seed of 30 bp, we detected the breakpoint sequence in 14/102 samples. Still grep is bound to have limited sensitivity as it allows no flexibility from alternate SNP alleles or sequencing errors and requires 100 % match of the 30 bp queried seed to a 48 bp read. Indeed, of the 14 samples with the detected *RCC1-ABHD12B* breakpoint sequence, most had only one read matching the grep query sequence. We therefore realize that using deFuse or grep in combination with RNA sequencing data from TCGA to validate confirmed fusion breakpoints have a low sensitivity, and we have thus not included these results in the manuscript. Compared to real-time PCR analyses of the recurrent fusion transcripts in our cohort of samples one could hypothesize a certain threshold level of detection for deFuse and grep in the TCGA data set (Figure 12). In a similar approach, Panagopoulos *et al.* could not discover the already known *CIC-DUX4* fusion transcript using FusionMap, FusionFinder or ChimeraScan, but were able to detect the fusion transcript by using the grep utility [174]. One can argue that deFuse and other fusion detection programs are useful tools for nomination of the most significant fusion breakpoints and not so much for targeted detection of a single fusion candidate. Higher sensitivity comes at a cost of lower specificity and more false positive fusions. However, better bioinformatic tools to validate known breakpoints, be it fusion transcripts or transcript variants, should improve the validation strategies using external data from large consortiums.

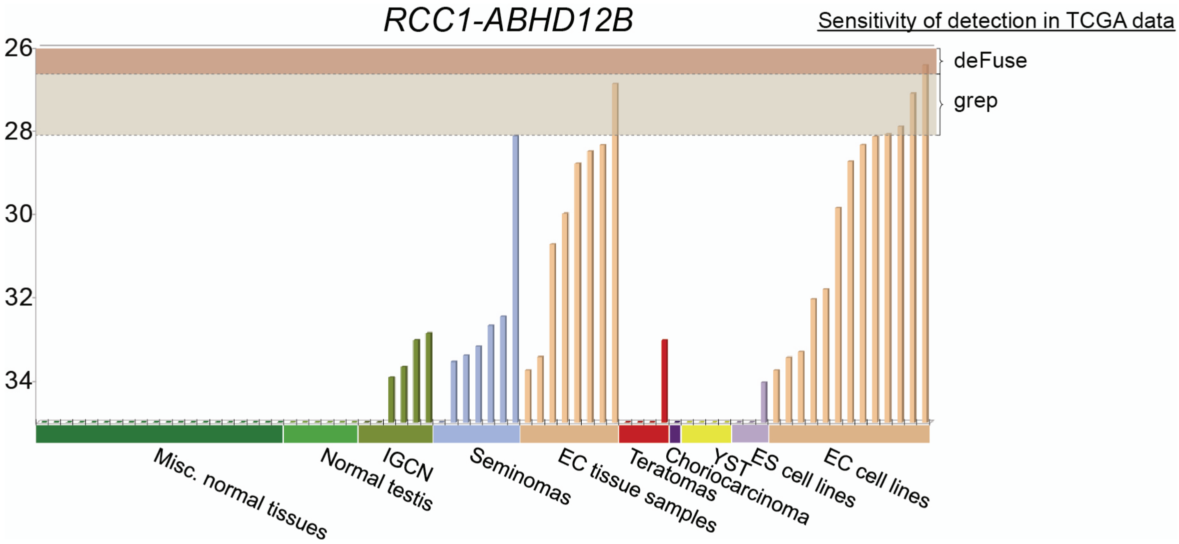


Figure 12: qRT-PCR results for *RCC1-ABHD12B* from paper III with compared detection rate of *RCC1-ABHD12B* breakpoint sequences in external TCGA data using deFuse or grep. The ratio of detected *RCC1-ABHD12B* breakpoint sequence in 102 TCGA samples varied from deFuse to grep. However, grep in RNA-seq data probably still have sub-optimal sensitivity when compared to the frequency of samples expressing *RCC1-ABHD12B* established by qRT-PCR in our series of TGCT samples and EC cell lines.

Open access DNA and RNA sequencing data

Sharing of the rapidly increasing amounts of DNA and RNA sequencing data is paramount to the progress of genomics. Lack of sharing prohibits other scientists from reproducing the results and to use the data for additional projects. Further, it may contribute to over-sequencing and redundant research projects. Several big consortiums have in the recent years set out to characterize and sequence thousands of individuals and cancer samples to help contribute enough power to understand the complexities in the human genome in health and disease. In this thesis we have implemented data from the TCGA and the CCLE which have become invaluable resources in cancer research. The TCGA has set out to comprehensively characterize the molecular basis of cancer through application of genome analysis techniques on thousands of samples. Currently, data for 33 cancer types are available, amounting to over 1.1 petabytes of data stored at the Cancer Genomics Hub (CGHub) [175]. Also housed at the CGHub is the data from the CCLE (26 terabytes) which provides genomic data, analysis and visualization for about 1000 cell lines [176]. The genomic data on cancer cell lines is especially useful, as cancer cell lines are frequently used as model systems for cancer and can be used for functional *in vitro* studies. Several countries have also established programs dedicated to prospectively characterize the genetics of disease and that of cancer patients. The Norwegian Cancer Genomics Consortium (NCGC; www.cancergenomics.no) is a collaborative effort in Norway for sequencing of cancer patients and to establish a protocol and infrastructure for using this information in future personalized medicine. In the UK, Genomics England together with the department of health launched a similar project with a goal to sequence 100,000 genomes, mainly from patients with rare diseases and cancer.

Even though sharing of sequencing data is essential, there are several ethical considerations and logistics that need to be addressed. First, sequencing data are not mere intensity values; the million of sequencing reads are all unique identifier sequences of the very personal genome. The sequence of the genome is much more of a personal fingerprint than any other biological trait. Therefore it should be considered that genetic information deposited online can be used to retrospectively identify individuals. The genetic information of the genome is not only a unique personal identifier, it is also a source for discovering future genetic diseases or increased disease risks for the individual in question and his or her family. If this information falls into the wrong hands, one can imagine the severe violations of privacy. In combination with large amounts of social information available through the internet, researchers have shown that the identity of research subjects can be inferred retroactively from sequencing data sets [177]. In 2013, the genome and the transcriptome of the HeLa cell line was sequenced and published by a German group [178]. The HeLa cancer cell line is the most used cell line in cancer research history, and information on the genome and transcriptome is invaluable for researchers world-wide that do work on the cell line. However, when the cell line was established from Henrietta Lacks, who was treated for cervical cancer at John Hopkins hospital more than 60 years ago, little or no information was given to herself or her next of kin. The lack of consent and information has been the topic of debate about the HeLa cell line for a long time. However, the release of the genome and transcriptome of HeLa added new fuel to the debate. The DNA sequencing data was retracted, but after a meeting between professionals and Henrietta Lacks' family, an agreement was made to re-publish the data under control of an approved access-system [179]. Here scientists can apply to access the data for use in biomedical research. The applications are evaluated by a committee involving amongst others members of Henrietta Lacks' family. Sufficient information to the patients and informed consent is required for ensuring privacy when doing sequencing. Additionally, controlled access systems should be continued for restricting use of the data to biomedical research and conserving the identity of persons involved. In Norway, there are strict legislations regarding person sensitive information, including DNA sequencing data. Any genetic testing is under regulation of the biotechnology law and any research project involving genetic testing has to be approved by a regional ethics committee. Storage and analysis of DNA sequencing data from patients have to follow strict guidelines for handling sensitive personal information on secure computer servers and networks. However, with the increasing amount of DNA sequencing data generated from cancer patients in Norway, from establishments such as the NCGC, efforts to enable controlled access data sharing should also be pursued. World-wide collaborative research and sharing of genomic data and clinical parameters will undoubtedly advance the field of cancer research in the decades to come.

Conclusions

The present thesis contributes with novel information on the identification of transcript variants and fusion genes in colorectal and testicular cancers. The development of experimental and bioinformatics approaches presented within this thesis facilitates the identification of novel cancer specific transcript variants. The fusion transcript and transcript variants identified in this manner may play a role in cancer and have potential as cancer biomarkers.

By using a combination of RNA sequencing and whole-genome sequencing, we were able to confirm the first fusion gene recently described in CRC by others [71]. In addition, we discovered a new fusion gene, where both fusion products involved *TCF7L2* and have corresponding genomic breakpoints in individual CRC cell lines. Importantly, we show that both the original and new fusions are expressed in a high number of CRC samples as well as in normal samples, suggesting that these fusion transcripts are not cancer specific.

We established a RACE-seq protocol that represents a powerful tool to discover either known or unknown transcript variants or fusion transcripts in malignancy. We used it and detected three non-recurrent fusion transcripts in CRC. Also, we used the approach to characterize new exon-exon junctions for a high fraction of genes we interrogated. Of these, the read-through *KLK8-KLK7* and alternative 3' splice site usage in *S100A2* were overrepresented in CRC compared to normal samples, but the value of their overexpression is not determined.

With RNA-sequencing followed by a defined bioinformatics pipeline to detect fusion transcripts, we were able to identify and validate the first fusion genes described in TGCT. Some of these were shown to be malignancy specific and recurrently expressed. Furthermore, three of the recurrent transcript variants were repressed *in vitro* upon differentiation induced with RA, indicating that these novel transcripts are related to the pluripotent state of TGCTs. We also demonstrated the power of ddPCR to establish whether fusion partner genes are linked at the DNA level.

Future perspectives

Methodological implementations

We are planning to implement the new RACE-seq protocol from Paper II in an effort to characterize the fusion genes status in 52 prostate cancer samples. By designing RACE-assays for each known downstream partner gene, such as ETS transcription factors, we plan to characterize all known fusion breakpoints and potentially discover new fusion partner genes in a single experimental set up. This set of prostate cancer samples contains multiple samples from different tumor foci in 13 prostate cancer patients. By establishing fusion gene status, we aim to elucidate aspects of heterogeneity within prostate tumors and the involvement of fusion genes in clonal development.

Further, we plan to establish a DNA isolation protocol for gentle isolation of intact long DNA molecules for use with ddPCR linkage analysis and validation of gene fusions at the DNA level. Although it is not feasible to apply this DNA isolation protocol for all samples of the large biobanks housed at the department, it can be used for evaluating DNA rearrangements underlying identified transcript variants in particularly interesting samples.

To detect and validate fusion transcripts in external RNA-sequencing data sets, such as those from TCGA, we have uncovered the need for a tool that can detect a given fusion transcript with high sensitivity. Application of traditional fusion detection tools, such as deFuse to detect a specific validated fusion gene probably suffers from a high amount of false negatives due to the different purpose of these algorithms. Fusion detection software are in general designed to nominate fusion genes with a high specificity out of thousands of potential candidates, naturally affecting the sensitivity of such approaches.

Clinical and functional aspects of discovered cancer specific variants

The relevance of the transcript variants discovered in paper II to be overrepresented in CRC, including the read-through *KLK8-KLK7* and the alternative 3' splice site usage in *S100A2* should be further explored. By designing specific real-time PCR assays for these transcript variants, we plan to investigate the expression of these in our clinical biobank of CRC samples. This approach would establish if these variants are expressed specifically or at a higher level in CRC compared to normal tissue and whether they are associated with particular clinical data or relevant other molecular data, and as such should be further pursued as potentially useful CRC biomarkers.

Some of the fusion transcripts discovered in TGCT were found to be frequently expressed in cancer samples and not in normal testis or other normal tissue samples from other anatomical sites. Also, the *RCC1* involving fusion transcripts and the alternative promoter of *ETV6* were found to be associated with the pluripotent state of the malignancies. Their potential roles in TGCT should be further explored both functionally and clinically. Sensitive detection of these transcripts in excreted body fluids or in blood of patients with TGCT could fill the need for biomarkers enabling better clinical management of TGCT patients [146]. Of interest to the role of *RCC1* in cancer, another member of the *RCC* family, *RCC2*, was recently published by us as a clinically interesting biomarker in CRC [180].

Biological questions – fusion transcripts and fusion genes

In paper I, we discovered that the *TCF7L2* involving fusion genes are expressed in a large fraction of CRCs and also in normal tissues, although often at low levels. Corresponding genomic rearrangements of these fusion genes are present in individual CRC cell lines, which in return have high expression of the fusion transcripts. This phenomenon has also been observed by others in prostate and endometrial cancers [49–51]. It is an intriguing fact that fusion transcripts are expressed normally, with corresponding genomic rearrangements seen in cancer, resulting in stronger and stable expression of these fusion transcripts. This suggests a coupled mechanism between the expression of fusion transcripts and generation of corresponding chromosomal rearrangements [43]. If we are able to reveal this mechanism and the dynamic regulation of the transcriptome, this will for sure be of important value for understanding malignancy and cancer. We plan to pursue this biological question in collaboration with others. With the aid of continuous improvement of techniques for investigating the three-dimensional folding of the genome and association with fusion transcript expression, long-read sequencing techniques and sensitive detection of fusion transcripts with ddPCR, we aim to uncover secrets of the cancer transcriptome.

References

1. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A: **Global cancer statistics, 2012**. CA Cancer J Clin 2015, 65:87–108.

2. Soerjomataram I, Lortet-Tieulent J, Parkin DM, Ferlay J, Mathers C, Forman D, *et al.*: **Global burden of cancer in 2008: a systematic analysis of disability-adjusted life-years in 12 world regions**. Lancet 2012, 380:1840–1850.

3. Bray F, Jemal A, Grey N, Ferlay J, Forman D: **Global cancer transitions according to the Human Development Index (2008-2030): a population-based study**. Lancet Oncol 2012, 13:790–801.

4. Hanahan D, Weinberg RA: **The hallmarks of cancer**. Cell 2000, 100:57–70.

5. Hanahan D, Weinberg RA: **Hallmarks of cancer: the next generation**. Cell 2011, 144:646–674.

6. Hansemann D: **Ueber asymmetrische Zelltheilung in Epithelkrebsen und deren biologische Bedeutung**. Arch Pathol Anat Physiol Klin Medicin 1890:299–326.

7. Boveri T: **Zur Frage der Entstehung Maligner Tumoren**. Fischer, Jena 1914.

8. Crick F: **Central dogma of molecular biology**. Nature 1970, 227:561–563.

9. Watson JD, Crick FH: **Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid**. Nature 1953, 171:737–738.

10. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, *et al.*: **Initial sequencing and analysis of the human genome**. Nature 2001, 409:860–921.

11. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, *et al.*: **The Sequence of the Human Genome**. Science 2001, 291:1304–1351.

12. International Human Genome Sequencing Consortium: **Finishing the euchromatic sequence of the human genome**. Nature 2004, 431:931–945.

13. Liang F, Holt I, Pertea G, Karamycheva S, Salzberg SL, Quackenbush J: **Gene index analysis of the human genome estimates approximately 120,000 genes**. Nat Genet 2000, 25:239–240.

14. Vogelstein B, Kinzler KW: **Cancer genes and the pathways they control**. Nat Med 2004, 10:789–799.

15. Knudson AG: **Antioncogenes and human cancer**. Proc Natl Acad Sci U S A 1993, 90:10914–10921.

16. Balmain A, Gray J, Ponder B: **The genetics and genomics of cancer**. Nat Genet 2003, 33:238–244.

17. Gordon DJ, Resio B, Pellman D: **Causes and consequences of aneuploidy in cancer**. Nat Rev Genet 2012, 13:189–203.

18. Feinberg AP, Tycko B: **The history of cancer epigenetics**. Nat Rev Cancer 2004, 4:143–153.

19. Esteller M: **Epigenetics in Cancer**. N Engl J Med 2008, 358:1148–1159.

20. Nowell P, Hungerford D: **Minute Chromosome in Human Chronic Granulocytic Leukemia**. Science 1960, 132:1497–1497.

21. Rowley JD: **Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining**. Nature 1973, 243:290–293.

22. Shtivelman E, Lifshitz B, Gale RP, Canaani E: **Fused transcript of abl and bcr genes in chronic myelogenous leukaemia**. Nature 1985, 315:550–554.

23. Stam K, Heisterkamp N, Grosveld G, de Klein A, Verma RS, Coleman M, *et al.*: **Evidence of a new chimeric bcr/c-abl mRNA in patients with chronic myelocytic leukemia and the Philadelphia chromosome**. N Engl J Med 1985, 313:1429–1433.

24. Mertens F, Johansson B, Fioretos T, Mitelman F: **The emerging complexity of gene fusions in cancer**. Nat Rev Cancer 2015, 15:371–381.

25. Speicher MR, Carter NP: **The new cytogenetics: blurring the boundaries with molecular biology**. Nat Rev Genet 2005, 6:782–792.

26. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, *et al.*: **Recurrent fusion of *TMPRSS2* and ETS transcription factor genes in prostate cancer**. Science 2005, 310:644–648.

27. Yoshihara K, Wang Q, Torres-Garcia W, Zheng S, Vegesna R, Kim H, *et al.*: **The landscape and therapeutic relevance of cancer-associated transcript fusions**. Oncogene 2014.

28. Negrini S, Gorgoulis VG, Halazonetis TD: **Genomic instability–an evolving hallmark of cancer**. Nat Rev Mol Cell Biol 2010, 11:220–228.

29. Lengauer C, Kinzler KW, Vogelstein B: **Genetic instabilities in human cancers**. Nature 1998, 396:643–649.

30. Akiva P, Toporik A, Edelheit S, Peretz Y, Diber A, Shemesh R, *et al.*: **Transcription-mediated gene fusion in the human genome**. Genome Res 2006, 16:30–36.

31. Elgar G, Vavouri T: **Tuning in to the signals: noncoding sequence conservation in vertebrate genomes**. Trends Genet 2008, 24:344–352.

32. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, *et al.*: **Landscape of transcription in human cells**. Nature 2012, 489:101–108.

33. The ENCODE Project Consortium: **An integrated encyclopedia of DNA elements in the human genome**. Nature 2012, 489:57–74.

34. Morris KV, Mattick JS: **The rise of regulatory RNA**. Nat Rev Genet 2014, 15:423–437.

35. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ: **Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing**. Nat Genet 2008, 40:1413–1415.

36. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, *et al.*: **Alternative isoform regulation in human tissue transcriptomes**. Nature 2008, 456:470–476.

37. Venables JP: **Aberrant and Alternative Splicing in Cancer**. Cancer Res 2004, 64:7647–7654.

38. Blencowe BJ: **Alternative splicing: new insights from global analyses**. Cell 2006, 126:37–47.

39. Skotheim RI, Nees M: **Alternative splicing in cancer: noise, functional, or systematic?** Int J Biochem Cell Biol 2007, 39:1432–1449.

40. Danan-Gotthold M, Golan-Gerstl R, Eisenberg E, Meir K, Karni R, Levanon EY: **Identification of recurrent regulated alternative splicing events across human solid tumors**. Nucleic Acids Res 2015, 43:5130–5144.

41. Djebali S, Lagarde J, Kapranov P, Lacroix V, Borel C, Mudge JM, *et al.*: **Evidence for Transcript Networks Composed of Chimeric RNAs in Human Cells**. PLoS ONE 2012, 7:e28213.

42. Prakash T, Sharma VK, Adati N, Ozawa R, Kumar N, Nishida Y, *et al.*: **Expression of conjoined genes: another mechanism for gene regulation in eukaryotes**. PLoS One 2010, 5:e13284.

43. Li H, Wang J, Ma X, Sklar J: **Gene fusions and RNA trans-splicing in normal and neoplastic human cells**. Cell Cycle 2009, 8:218–222.

44. Frenkel-Morgenstern M, Lacroix V, Ezkurdia I, Levin Y, Gabashvili A, Prilusky J, *et al.*: **Chimeras taking shape: Potential functions of proteins encoded by chimeric RNA transcripts**. Genome Res 2012, 7:1231–1242.

45. Parra G, Reymond A, Dabbouseh N, Dermitzakis ET, Castelo R, Thomson TM, *et al.*: **Tandem chimerism as a means to increase protein complexity in the human genome**. Genome Res 2006, 16:37–44.

46. Plebani R, Oliver GR, Trerotola M, Guerra E, Cantanelli P, Apicella L, *et al.*: **Long-range transcriptome sequencing reveals cancer cell growth regulatory chimeric mRNA**. Neoplasia 2012, 14:1087–1096.

47. Zhang Y, Gong M, Yuan H, Park HG, Frierson HF, Li H: **Chimeric Transcript Generated by cis-Splicing of Adjacent Genes Regulates Prostate Cancer Cell Proliferation**. Cancer Discov 2012, 7:598–607.

48. Yun SM, Yoon K, Lee S, Kim E, Kong S-H, Choe J, *et al.*: ***PPP1R1B-STARD3* chimeric fusion transcript in human gastric cancer promotes tumorigenesis through activation of PI3K/AKT signaling**. Oncogene 2014, 33:5341–5347.

49. Rickman DS, Pflueger D, Moss B, VanDoren VE, Chen CX, de la Taille A, *et al.*: ***SLC45A3-ELK4* is a novel and frequent erythroblast transformation-specific fusion transcript in prostate cancer**. Cancer Res 2009, 69:2734–2738.

50. Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, *et al.*: **Transcriptome sequencing to detect gene fusions in cancer**. Nature 2009, 458:97–101.

51. Li H, Wang J, Mor G, Sklar J: **A neoplastic gene fusion mimics trans-splicing of RNAs in normal human cells**. Science 2008, 321:1357–1361.

52. Cancer registry of Norway, Oslo: **Cancer in Norway 2012; Cancer incidence, mortality, survival and prevalence in Norway**. [Available: http://kreftregisteret.no/], Accessed 1 June 2015.

53. Center MM, Jemal A, Smith RA, Ward E: **Worldwide variations in colorectal cancer**. CA Cancer J Clin 2009, 59:366–378.

54. Center MM, Jemal A, Ward E: **International trends in colorectal cancer incidence rates**. Cancer Epidemiol Biomarkers Prev 2009, 18:1688–1694.

55. Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C, *et al.*: **GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11**. Int Agency Res Cancer. [Available: http://globocan.iarc.fr/Default.aspx], Accessed 9 June 2015.

56. Haggar FA, Boushey RP: **Colorectal Cancer Epidemiology: Incidence, Mortality, Survival, and Risk Factors**. Clin Colon Rectal Surg 2009, 22:191–197.

57. Huxley RR, Ansary-Moghaddam A, Clifton P, Czernichow S, Parr CL, Woodward M: **The impact of dietary and lifestyle risk factors on risk of colorectal cancer: a quantitative overview of the epidemiological evidence**. Int J Cancer 2009, 125:171–180.

58. Edge SB, American Joint Committee on Cancer, editors: **AJCC cancer staging manual**. 7. ed. New York: Springer. 648 p.

59. O'Connell JB, Maggard MA, Ko CY: **Colon Cancer Survival Rates With the New American Joint Committee on Cancer Sixth Edition Staging**. J Natl Cancer Inst 2004, 96:1420–1425.

60. Kim MJ, Jeong S-Y, Choi S-J, Ryoo S-B, Park JW, Park KJ, *et al.*: **Survival paradox between stage IIB/C (T4N0) and stage IIIA (T1-2N1) colon cancer**. Ann Surg Oncol 2015, 22:505–512.

61. Rustgi AK: **The genetics of hereditary colon cancer**. Genes Dev 2007, 21:2525–2538.

62. Kinzler KW, Vogelstein B: **Lessons from Hereditary Colorectal Cancer**. Cell 1996, 87:159–170.

63. Fearon ER, Vogelstein B: **A genetic model for colorectal tumorigenesis**. Cell 1990, 61:759–767.

64. Powell SM, Zilz N, Beazer-Barclay Y, Bryan TM, Hamilton SR, Thibodeau SN, *et al.*: ***APC* mutations occur early during colorectal tumorigenesis**. Nature 1992, 359:235–237.

65. Al-Sohaily S, Biankin A, Leong R, Kohonen-Corish M, Warusavitarne J: **Molecular pathways in colorectal cancer**. J Gastroenterol Hepatol 2012, 27:1423–1431.

66. Vogelstein B, Fearon ER, Hamilton SR, Kern SE, Preisinger AC, Leppert M, *et al.*: **Genetic Alterations during Colorectal-Tumor Development**. N Engl J Med 1988, 319:525–532.

67. Iacopetta B: ***TP53* mutation in colorectal cancer**. Hum Mutat 2003, 21:271–276.

68. Takahashi K, Yamanaka S: **Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors**. Cell 2006, 126:663–676.

69. Morin PJ, Sparks AB, Korinek V, Barker N, Clevers H, Vogelstein B, *et al.*: **Activation of β-Catenin-Tcf Signaling in Colon Cancer by Mutations in β-Catenin or APC**. Science 1997, 275:1787–1790.

70. Network TCGA: **Comprehensive molecular characterization of human colon and rectal cancer**. Nature 2012, 487:330–337.

71. Bass AJ, Lawrence MS, Brace LE, Ramos AH, Drier Y, Cibulskis K, *et al.*: **Genomic sequencing of colorectal adenocarcinomas identifies a recurrent *VTI1A-TCF7L2* fusion**. Nat Genet 2011, 43:964–968.

72. Heppner GH: **Tumor heterogeneity**. Cancer Res 1984, 44:2259–2265.

73. Marusyk A, Polyak K: **Tumor heterogeneity: causes and consequences**. Biochim Biophys Acta 2010, 1805:105–117.

74. Reya T, Morrison SJ, Clarke MF, Weissman IL: **Stem cells, cancer, and cancer stem cells**. Nature 2001, 414:105–111.

75. Nowell PC: **The clonal evolution of tumor cell populations**. Science 1976, 194:23–28.

76. Soltanian S, Matin MM: **Cancer stem cells and cancer therapy**. Tumor Biol 2011, 32:425–440.

77. Sawyers CL: **Chronic myeloid leukemia**. N Engl J Med 1999, 340:1330–1340.

78. Druker BJ, Tamura S, Buchdunger E, Ohno S, Segal GM, Fanning S, *et al.*: **Effects of a selective inhibitor of the Abl tyrosine kinase on the growth of Bcr-Abl positive cells**. Nat Med 1996, 2:561–566.

79. Mahon F-X, Réa D, Guilhot J, Guilhot F, Huguet F, Nicolini F, *et al.*: **Discontinuation of imatinib in patients with chronic myeloid leukaemia who have maintained complete molecular remission for at least 2 years: the prospective, multicentre Stop Imatinib (STIM) trial**. Lancet Oncol 2010, 11:1029–1035.

80. Ross DM, Branford S, Seymour JF, Schwarer AP, Arthur C, Yeung DT, *et al.*: **Safety and efficacy of imatinib cessation for CML patients with stable undetectable minimal residual disease: results from the TWISTER study**. Blood 2013, 122:515–522.

81. Ross DM, Hughes TP, Melo JV: **Do we have to kill the last CML cell?** Leukemia 2011, 25:193–200.

82. Gupta PB, Chaffer CL, Weinberg RA: **Cancer stem cells: mirage or reality?** Nat Med 2009, 15:1010–1012.

83. Clevers H: **The cancer stem cell: premises, promises and challenges**. Nat Med 2011, 17:313–319.

84. Baker DE, Harrison NJ, Maltby E, Smith K, Moore HD, Shaw PJ, *et al.*: **Adaptation to culture of human embryonic stem cells and oncogenesis in vivo**. Nat Biotechnol 2007, 25:207–215.

85. Andrews PW, Matin MM, Bahrami AR, Damjanov I, Gokhale P, Draper JS: **Embryonic stem (ES) cells and embryonal carcinoma (EC) cells: opposite sides of the same coin**. Biochem Soc Trans 2005, 33:1526–1530.

86. Alagaratnam S, Harrison N, Bakken AC, Hoff AM, Jones M, Sveen A, *et al.*: **Transforming pluripotency: an exon-level study of malignancy-specific transcripts in human embryonal carcinoma and embryonic stem cells**. Stem Cells Dev 2013, 22:1136–1146.

87. Znaor A, Lortet-Tieulent J, Jemal A, Bray F: **International Variations and Trends in Testicular Cancer Incidence and Mortality**. Eur Urol 2014, 65:1095–1106.

88. Huyghe E, Plante P, Thonneau PF: **Testicular cancer variations in time and space in Europe**. Eur Urol 2007, 51:621–628.

89. Huyghe E, Matsuda T, Thonneau P: **Increasing incidence of testicular cancer worldwide: a review**. J Urol 2003, 170:5–11.

90. Horwich A, Shipley J, Huddart R: **Testicular germ-cell cancer**. Lancet 2006, 367:754–765.

91. Engholm G, Ferlay J, Christensen N, Kejs AMT, Johannesen TB, Khan S, *et al.*: **NORDCAN: Cancer Incidence, Mortality, Prevalence and Survival in the Nordic Countries**. [Available: http://www.ancr.nu], Accessed 9 June 2015.

92. Masters JRW, Köberle B: **Curing metastatic cancer: lessons from testicular germ-cell tumours**. Nat Rev Cancer 2003, 3:517–525.

93. Haugnes HS, Bosl GJ, Boer H, Gietema JA, Brydøy M, Oldenburg J, *et al.*: **Long-Term and Late Effects of Germ Cell Testicular Cancer Treatment and Implications for Follow-Up**. J Clin Oncol 2012, 30:3752–3763.

94. Skakkebæk NE: **Possible carcinoma-in-situ of the testis**. Lancet 1972, 2:516–517.

95. Jacobsen GK, Henriksen OB, von der Maase H: **Carcinoma in situ of testicular tissue adjacent to malignant germ-cell tumors: a study of 105 cases**. Cancer 1981, 47:2660–2662.

96. Hussain SA, Ma YT, Palmer DH, Hutton P, Cullen MH: **Biology of testicular germ cell tumors**. Expert Anticancer Ther 2008, 8:1659–1673.

97. Woodward, PJ, Heidenreich, A, Looijenga, LHJ: **Germ cell tumours**. In: Eble JN, International Agency for Research on Cancer, editors. Pathology and genetics of tumours of the urinary system and male genital organs. World Health Organization classification of tumours. Lyon: IARC Press. pp. 221–249.

98. Brabrand S, Johannessen B, Axcrona U, Kraggerud SM, Berg KG, Bakken AC, *et al.*: **Exome sequencing of bilateral testicular germ cell tumors suggests independent development lineages**. Neoplasia 2015, 17:167–174.

99. Cutcutache I, Suzuki Y, Tan IB, Ramgopal S, Zhang S, Ramnarayanan K, *et al.*: **Exome-wide Sequencing Shows Low Mutation Rates and Identifies Novel Mutated Genes in Seminomas**. Eur Urol 2015, 68:77–83.

100. Litchfield K, Summersgill B, Yost S, Sultana R, Labreche K, Dudakia D, *et al.*: **Whole-exome sequencing reveals the mutational spectrum of testicular germ cell tumours**. Nat Commun 2015, 6:5973.

101. Gilbert D, Rapley E, Shipley J: **Testicular germ cell tumours: predisposition genes and the male germ cell niche**. Nat Rev Cancer 2011, 11:278–288.

102. Atkin NB, Baker MC: **i(12p): specific chromosomal marker in seminoma and malignant teratoma of the testis?** Cancer Genet Cytogenet 1983, 10:199–204.

103. Kraggerud SM, Skotheim RI, Szymanska J, Eknæs M, Fosså SD, Stenwig AE, *et al.*: **Genome profiles of familial/bilateral and sporadic testicular germ cell tumors**. Genes Chromosomes Cancer 2002, 34:168–174.

104. Ottesen AM, Skakkebæk NE, Lundsteen C, Leffers H, Larsen J, Rajpert-De Meyts E: **High-resolution comparative genomic hybridization detects extra chromosome arm 12p material in most cases of carcinoma in situ adjacent to overt germ cell tumors, but not before the invasive tumor development**. Genes Chromosomes Cancer 2003, 38:117–125.

105. Sheikine Y, Genega E, Melamed J, Lee P, Reuter VE, Ye H: **Molecular genetics of testicular germ cell tumors**. Am J Cancer Res 2012, 2:153–167.

106. Skotheim RI, Lothe RA: **The testicular germ cell tumour genome**. APMIS 2003, 111:136–151.

107. Houldsworth J, Reuter V, Bosl GJ, Chaganti RS: **Aberrant expression of cyclin D2 is an early event in human male germ cell tumorigenesis**. Cell Growth Differ 1997, 8:293–299.

108. Kranenburg O: **The *KRAS* oncogene: past, present, and future**. Biochim Biophys Acta 2005, 1756:81–82.

109. Kemmer K, Corless CL, Fletcher JA, McGreevey L, Haley A, Griffith D, *et al.*: ***KIT* Mutations Are Common in Testicular Seminomas**. Am J Pathol 2004, 164:305–313.

110. McIntyre A, Summersgill B, Grygalewicz B, Gillis AJM, Stoop J, Gurp RJHLM van, *et al.*: **Amplification and Overexpression of the *KIT* Gene Is Associated with Progression in the Seminoma Subtype of Testicular Germ Cell Tumors of Adolescents and Adults**. Cancer Res 2005, 65:8085–8089.

111. McIntyre A, Summersgill B, Spendlove HE, Huddart R, Houlston R, Shipley J: **Activating mutations and/or expression levels of tyrosine kinase receptors *GRB7*, *RAS*, and *BRAF* in testicular germ cell tumors**. Neoplasia 2005, 7:1047–1052.

112. Kanetsky PA, Mitra N, Vardhanabhuti S, Li M, Vaughn DJ, Letrero R, *et al.*: **Common variation in *KITLG* and at 5q31.3 predisposes to testicular germ cell cancer**. Nat Genet 2009, 41:811–815.

113. Rapley EA, Turnbull C, Al Olama AA, Dermitzakis ET, Linger R, Huddart RA, *et al.*: **A genome-wide association study of testicular germ cell tumor**. Nat Genet 2009, 41:807–810.

114. Looijenga LHJ, Gillis AJM, Stoop H, Biermann K, Oosterhuis JW: **Dissecting the molecular pathways of (testicular) germ cell tumour pathogenesis; from initiation to treatment-resistance**. Int J Androl 2011, 34:e234–e251.

115. Biomarkers Definitions Working Group.: **Biomarkers and surrogate endpoints: preferred definitions and conceptual framework**. Clin Pharmacol Ther 2001, 69:89–95.

116. Basch E, Oliver TK, Vickers A, Thompson I, Kantoff P, Parnes H, *et al.*: **Screening for Prostate Cancer With Prostate-Specific Antigen Testing: American Society of Clinical Oncology Provisional Clinical Opinion**. J Clin Oncol 2012, 30:3020–3025.

117. Tumor Markers. Natl Cancer Inst. [Available: http://www.cancer.gov/about-cancer/diagnosis-staging/diagnosis/tumor-markers-fact-sheet], Accessed 18 June 2015.

118. Hughes T, Deininger M, Hochhaus A, Branford S, Radich J, Kaeda J, *et al.*: **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 2006, 108:28–37.

119. Druker BJ, Guilhot F, O'Brien SG, Gathmann I, Kantarjian H, Gattermann N, *et al.*: **Five-Year Follow-up of Patients Receiving Imatinib for Chronic Myeloid Leukemia**. N Engl J Med 2006, 355:2408–2417.

120. Warrell RP, Frankel SR, Miller WH, Scheinberg DA, Itri LM, Hittelman WN, *et al.*: **Differentiation Therapy of Acute Promyelocytic Leukemia with Tretinoin (All-trans-Retinoic Acid)**. N Engl J Med 1991, 324:1385–1393.

121. Illendula A, Pulikkan JA, Zong H, Grembecka J, Xue L, Sen S, *et al.*: **Chemical biology. A small-molecule inhibitor of the aberrant transcription factor CBFβ-SMMHC delays leukemia in mice**. Science 2015, 347:779–784.

122. Nagaraj SH, Gasser RB, Ranganathan S: **A hitchhiker's guide to expressed sequence tag (EST) analysis**. Brief Bioinform 2007, 8:6–21.

123. Sveen A, Nesbakken A, Ågesen TH, Guren MG, Tveit KM, Skotheim RI, *et al.*: **Anticipating the Clinical Use of Prognostic Gene Expression–Based Tests for Colon Cancer Stage II and III: Is Godot Finally Arriving?** Clin Cancer Res 2013, 19:6669–6677.

124. Sveen A, Ågesen TH, Nesbakken A, Rognum TO, Lothe RA, Skotheim RI: **Transcriptome instability in colorectal cancer identified by exon microarray analyses: Associations with splicing factor expression levels and patient survival**. Genome Med 2011, 3:32.

125. Sveen A, Johannessen B, Teixeira MR, Lothe RA, Skotheim RI: **Transcriptome instability as a molecular pan-cancer characteristic of carcinomas**. BMC Genomics 2014, 15:672.

126. Skotheim RI, Thomassen GOS, Eken M, Lind GE, Micci F, Ribeiro FR, *et al.*: **A universal assay for detection of oncogenic fusion transcripts by oligo microarray analysis**. Mol Cancer 2009, 8:5.

127. Løvf M, Thomassen G, Bakken AC, others: **Fusion gene microarray reveals cancer type-specificity among fusion genes**. Genes 2011.

128. Løvf M, Thomassen GOS, Mertens F, Cerveira N, Teixeira MR, Lothe RA, *et al.*: **Assessment of Fusion Gene Status in Sarcomas Using a Custom Made Fusion Gene Microarray**. PLoS ONE 2013, 8:e70649.

129. Ozsolak F, Milos PM: **RNA sequencing: advances, challenges and opportunities**. Nat Rev Genet 2011, 12:87–98.

130. Denoeud F, Kapranov P, Ucla C, Frankish A, Castelo R, Drenkow J, *et al.*: **Prominent use of distal 5' transcription start sites and discovery of a large number of additional exons in ENCODE regions**. Genome Res 2007, 17:746–759.

131. Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, *et al.*: **What is a gene, post-ENCODE? History and updated definition**. Genome Res 2007, 17:669–681.

132. Nigro JM, Cho KR, Fearon ER, Kern SE, Ruppert JM, Oliner JD, *et al.*: **Scrambled exons**. Cell 1991, 64:607–613.

133. Salzman J, Gawad C, Wang PL, Lacayo N, Brown PO: **Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types**. PLoS One 2012, 7:e30733.

134. Parada LA, McQueen PG, Misteli T: **Tissue-specific spatial organization of genomes**. Genome Biol 2004, 5:R44.

135. Misteli T: **Spatial positioning; a new dimension in genome function**. Cell 2004, 119:153–156.

136. Meaburn KJ, Misteli T, Soutoglou E: **Spatial genome organization in the formation of chromosomal translocations**. Semin Cancer Biol 2007, 17:80–90.

137. Belton J-M, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J: **Hi–C: A comprehensive technique to capture the conformation of genomes**. Methods 2012, 58:268–276.

138. Engreitz JM, Agarwala V, Mirny LA: **Three-dimensional genome architecture influences partner selection for chromosomal translocations in human disease**. PloS One 2012, 7:e44196.

139. Honeyman JN, Simon EP, Robine N, Chiaroni-Clarke R, Darcy DG, Lim II, *et al.*: **Detection of a recurrent *DNAJB1-PRKACA* chimeric transcript in fibrolamellar hepatocellular carcinoma**. Science 2014, 343:1010–1014.

140. Graham RP, Jin L, Knutson DL, Kloft-Nelson SM, Greipp PT, Waldburger N, *et al.*: ***DNAJB1-PRKACA* is specific for fibrolamellar carcinoma**. Mod Pathol 2015, 28:822–829.

141. Walker F, Nicole P, Jallane A, Soosaipillai A, Mosbach V, Oikonomopoulou K, *et al.*: **Kallikrein-related peptidase 7 (*KLK7*) is a proliferative factor that is aberrantly expressed in human colon cancer**. Biol Chem 2014, 395:1075–1086.

142. Talieri M, Mathioudaki K, Prezas P, Alexopoulou DK, Diamandis EP, Xynopoulos D, *et al.*: **Clinical significance of kallikrein-related peptidase 7 (*KLK7*) in colorectal cancer**. Thromb Haemost 2009, 104:741–747.

143. Talieri M, Li L, Zheng Y, Alexopoulou DK, Soosaipillai A, Scorilas A, *et al.*: **The use of kallikrein-related peptidases as adjuvant prognostic markers in colorectal cancer**. Br J Cancer 2009, 100:1659–1665.

144. Salama I, Malone PS, Mihaimeed F, Jones JL: **A review of the S100 proteins in cancer**. Eur J Surg Oncol 2008, 34:357–364.

145. Giráldez MD, Lozano JJ, Cuatrecasas M, Alonso-Espinaco V, Maurel J, Mármol M, *et al.*: **Gene-expression signature of tumor recurrence in patients with stage II and III colon cancer treated with 5'fluoruracil-based adjuvant chemotherapy**. Int J Cancer J Int Cancer 2013, 132:1090–1097.

146. Favilla V, Cimino S, Madonia M, Morgia G: **New advances in clinical biomarkers in testis cancer**. Front Biosci 2010, 2:456–477.

147. Frohman MA, Dush MK, Martin GR: **Rapid production of full-length cDNAs from rare transcripts: amplification using a single gene-specific oligonucleotide primer**. Proc Natl Acad Sci 1988, 85:8998–9002.

148. Roberts C, Jiang W, Jayaraman J, Trowsdale J, Holland MJ, Traherne JA: **Killer-cell Immunoglobulin-like Receptor gene linkage and copy number variation analysis by droplet digital PCR**. Genome Med 2014, 6:20.

149. Regan JF, Kamitaki N, Legler T, Cooper S, Klitgord N, Karlin-Neumann G, *et al.*: **A Rapid Molecular Approach for Chromosomal Phasing**. PLoS ONE 2015, 10:e0118270.

150. Andrews S: **FastQC a quality-control tool for high-throughput sequence data**. [Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/].

151. Martin M: **Cutadapt removes adapter sequences from high-throughput sequencing reads**. EMBnet J 2011, 17.

152. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq**. Bioinformatics 2009, 25:1105–1111.

153. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. Genome Biol 2009, 10:R25.

154. Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, The RGASP Consortium, *et al.*: **Systematic evaluation of spliced alignment programs for RNA-seq data**. Nat Methods 2013, 10:1185–1191.

155. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, *et al.*: **STAR: ultrafast universal RNA-seq aligner**. Bioinformatics 2013, 29:15–21.

156. Desfeux A, Henry V, Gonzalez B, Bandrowski A: **OMICtools: Gene fusion detection**. [Available: http://omictools.com/gene-fusion-detection-c141-p1.html], Accessed 3 June 2015.

157. Henry VJ, Bandrowski AE, Pepin A-S, Gonzalez BJ, Desfeux A: **OMICtools: an informative directory for multi-omic data analysis**. Database 2014, 2014:bau069.

158. McPherson A, Hormozdiari F, Zayed A, Giuliany R, Ha G, Sun MGF, *et al.*: **deFuse: An Algorithm for Gene Fusion Discovery in Tumor RNA-Seq Data**. PLoS Comput Biol 2011, 7:e1001138.

159. McPherson A, Wu C, Wyatt AW, Shah S, Collins C, Sahinalp SC: **nFuse: Discovery of complex genomic rearrangements in cancer using high-throughput sequencing**. Genome Res 2012, 22:2250–2261.

160. Jia W, Qiu K, He M, Song P, Zhou Q, Zhou F, *et al.*: **SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data**. Genome Biol 2013, 14:R12.

161. Francis RW, Thompson-Wicking K, Carter KW, Anderson D, Kees UR, Beesley AH: **FusionFinder: A Software Tool to Identify Expressed Gene Fusion Candidates from RNA-Seq Data**. PLoS ONE 2012, 7:e39987.

162. Iyer MK, Chinnaiyan AM, Maher CA: **ChimeraScan: a tool for identifying chimeric transcription in sequencing data**. Bioinformatics 2011, 27:2903–2904.

163. Abate F, Acquaviva A, Paciello G, Foti C, Ficarra E, Ferrarini A, *et al.*: **Bellerophontes: an RNA-Seq data analysis framework for chimeric transcripts discovery based on accurate fusion model**. Bioinformatics 2012, 28:2114–2121.

164. Davidson NM, Majewski IJ, Oshlack A: **JAFFA: High sensitivity transcriptome-focused fusion gene detection**. Genome Med 2015, 7:43.

165. Kim D, Salzberg SL: **TopHat-Fusion: an algorithm for discovery of novel fusion transcripts**. Genome Biol 2011, 12:R72.

166. Check Hayden E: **Is the $1,000 genome for real?** Nature 2014. [Available: http://www.nature.com/doifinder/10.1038/nature.2014.14530], Accessed 4 June 2015.

167. **BGI-Shenzhen Completes Acquisition of Complete Genomics.** Complete Genomics. [Available: http://www.completegenomics.com/news-events/press-releases/bgi-shenzhen-completes-acquisition-of-complete-genomics-198854331/], Accessed 23 June 2015.

168. Drmanac R, Sparks AB, Callow MJ, Halpern AL, Burns NL, Kermani BG, *et al.*: **Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays**. Science 2010, 327:78–81.

169. **Complete Genomics Previews Revolocity Sequencing System at European Human Genetics Conference 2015.** Complete Genomics. [Available: http://www.completegenomics.com/news-events/press-releases/complete-genomics-previews-revolocity-sequencing-system-at-european-human-genetics-conference-2015/], Accessed 23 June 2015.

170. Loman N: **Wiggle plot showing Oxford Nanopore signal data for a P. aeruginosa read**. 2014. [Available: http://figshare.com/articles/Wiggle_plot_showing_Oxford_Nanopore_signal_data_for_a_P_aeruginosa_read/1053026], Accessed 14 June 2015.

171. Eisenstein M: **Oxford Nanopore announcement sets sequencing sector abuzz**. Nat Biotechnol 2012, 30:295–296.

172. Kilianski A, Haas JL, Corriveau EJ, Liem AT, Willis KL, Kadavy DR, *et al.*: **Bacterial and viral identification and differentiation by amplicon sequencing on the MinION nanopore sequencer**. GigaScience 2015, 4:12.

173. Quick J, Ashton P, Calus S, Chatt C, Gossain S, Hawker J, *et al.*: **Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of Salmonella**. Genome Biol 2015, 16:114.

174. Panagopoulos I, Gorunova L, Bjerkehagen B, Heim S: **The "grep" command but not FusionMap, FusionFinder or ChimeraScan captures the *CIC-DUX4* fusion gene from whole transcriptome sequencing data on a small round cell tumor with t(4;19)(q35;q13)**. PloS One 2014, 9:e99439.

175. Wilks C, Cline MS, Weiler E, Diehkans M, Craft B, Martin C, *et al.*: **The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data**. Database 2014, 2014. [Available: http://database.oxfordjournals.org/content/2014/bau093], Accessed 5 June 2015.

176. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, *et al.*: **The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity**. Nature 2012, 483:603–607.

177. Gymrek M, McGuire AL, Golan D, Halperin E, Erlich Y: **Identifying personal genomes by surname inference**. Science 2013, 339:321–324.

178. Landry JJM, Pyl PT, Rausch T, Zichner T, Tekkedil MM, Stütz AM, *et al.*: **The Genomic and Transcriptomic Landscape of a HeLa Cell Line**. Genes Genomes Genet 2013,:g3.113.005777.

179. Callaway E: **Deal done over HeLa cell line**. Nature 2013, 500:132–133.

180. Bruun J, Kolberg M, Ahlquist TC, Røyrvik E, Nome T, Leithe E, *et al.*: **Regulator of chromosome condensation 2 identifies high-risk patients within both major phenotypes of colorectal cancer**. Clin Cancer Res 2015.

Paper I

High frequency of fusion transcripts involving *TCF7L2* in colorectal cancer: novel fusion partner and splice variants

Torfinn Nome*, Andreas M. Hoff*, Anne Cathrine Bakken, Torleiv O. Rognum, Arild Nesbakken, Rolf I. Skotheim

PLoS One 2014, 9:e91264

*Equal contribution

PLOS | ONE

High Frequency of Fusion Transcripts Involving *TCF7L2* in Colorectal Cancer: Novel Fusion Partner and Splice Variants

Torfinn Nome[1,2]⁹, Andreas M. Hoff[1,2]⁹, Anne Cathrine Bakken[1,2], Torleiv O. Rognum[3,4], Arild Nesbakken[2,5], Rolf I. Skotheim[1,2]*

1 Department of Cancer Prevention, Institute for Cancer Research, Norwegian Radium Hospital - Oslo University Hospital, Oslo, Norway, **2** Centre for Cancer Biomedicine, Faculty of Medicine, University of Oslo, Oslo, Norway, **3** University of Oslo, Oslo, Norway, **4** Department of forensic pathology and clinical forensic medicine, Division for forensic medicine, The Norwegian Institute of Public Health, Oslo, Norway, **5** Department of Gastrointestinal Surgery, Oslo University Hospital-Aker, Oslo, Norway

Abstract

VTI1A-TCF7L2 was reported as a recurrent fusion gene in colorectal cancer (CRC), found to be expressed in three out of 97 primary cancers, and one cell line, NCI-H508, where a genomic deletion joins the two genes [1]. To investigate this fusion further, we analyzed high-throughput DNA and RNA sequencing data from seven CRC cell lines, and identified the gene *RP11-57H14.3* (ENSG00000225292) as a novel fusion partner for *TCF7L2*. The fusion was discovered from both genome and transcriptome data in the HCT116 cell line. By triplicate nested RT-PCR, we tested both the novel fusion transcript and *VTI1A-TCF7L2* for expression in a series of 106 CRC tissues, 21 CRC cell lines, 14 normal colonic mucosa, and 20 normal tissues from miscellaneous anatomical sites. Altogether, 42% and 45% of the CRC samples expressed *VTI1A-TCF7L2* and *TCF7L2-RP11-57H14.3* fusion transcripts, respectively. The fusion transcripts were both seen in 29% of the normal colonic mucosa samples, and in 25% and 75% of the tested normal tissues from other organs, revealing that the *TCF7L2* fusion transcripts are neither specific to cancer nor to the colon and rectum. Seven different splice variants were detected for the *VTI1A-TCF7L2* fusion, of which three are novel. Four different splice variants were detected for the *TCF7L2-RP11-57H14.3* fusion. In conclusion, we have identified novel variants of *VTI1A-TCF7L2* fusion transcripts, including a novel fusion partner gene, *RP11-57H14*.3, and demonstrated detectable levels in a large fraction of CRC samples, as well as in normal colonic mucosa and other tissue types. We suggest that the fusion transcripts observed in a high frequency of samples are transcription induced chimeras that are expressed at low levels in most samples. The similar fusion transcripts induced by genomic rearrangements observed in individual cancer cell lines may yet have oncogenic potential as suggested in the original study by Bass et al.

Citation: Nome T, Hoff AM, Bakken AC, Rognum TO, Nesbakken A, et al. (2014) High Frequency of Fusion Transcripts Involving *TCF7L2* in Colorectal Cancer: Novel Fusion Partner and Splice Variants. PLoS ONE 9(3): e91264. doi:10.1371/journal.pone.0091264

Editor: Nathan A. Ellis, University of Illinois at Chicago, United States of America

Received October 16, 2013; **Accepted** February 10, 2014; **Published** March 7, 2014

Funding: The study was funded by grants from the Norwegian Cancer Society (TN and AMH were financed as PhD-students from the PR-2007-0166 grant to RIS), NorStore (for storage of computer files; project NS9013K), and partly supported by the Research Council of Norway through its Centres of Excellence funding scheme, project number 179571. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: rolf.i.skotheim@rr-research.no

⁹ These authors contributed equally to this work.

Introduction

Colorectal cancer (CRC) is the second most common and deadly cancer disease world-wide [2], and there is high demand of good biomarkers for both early detection and to stratify patients according to prognosis and predicted treatment responses. However, CRC is a heterogeneous disease, and few biomarkers have yet made it to routine clinical use.

Fusion genes represent one class of cancer genes with promising biomarker potential, and when caused by chromosomal rearrangements such as translocations, deletions or duplications, they are commonly highly cancer specific. So far, few highly recurrent fusion genes have been identified in CRC. Examples from other cancer types includes the well-known *BCR-ABL1* fusion (Philadelphia chromosome), present in 95% of chronic myelogenous leukemias [3], and *TMPRSS2-ERG* which is present in around 50% of prostate cancers [4]. In some cases, the fusion gene can be a therapeutic target, demonstrated by Imatinib binding to and blocking the kinase domain of the BCR-ABL1 fusion protein.

The first fusion gene reported to be recurrent in CRC, fusing sequences of *VTI1A* and *TCF7L2*, was reported in 2011 [1]. *VTI1A-TCF7L2* fusion transcripts were detected in three out of 97 (3%) CRCs, as well as the colon cancer cell line NCI-H508. In NCI-H508, the fusion is caused by a ~540 kb deletion between the genes *VTI1A* and *TCF7L2*. *TCF7L2* encodes the TCF4 transcription factor, which dimerizes with β-catenin. β-catenin is involved in regulation of the WNT-signaling pathway, which is commonly altered in the majority, if not all, CRCs [5]. Depletion of the *VTI1A-TCF7L2* fusion transcript resulted in significant loss of anchorage independent growth in the fusion positive CRC cell line NCI-H508 [1].

Frequent mutations in a polyadenine tract within the *TCF7L2* gene have previously been reported for CRC, especially in microsatellite instable tumors [6]. Microsatellite stable tumors have, however, also recently been shown to harbor frequent mutations in the *TCF7L2* gene, indicating a possible important function in CRC biology [7].

In this study, we aimed to investigate further the presence and variants of *VTI1A-TCF7L2* fusions in CRC. We have analyzed deep sequencing whole-genome and transcriptome data from CRC for related fusions involving one of the fusion partners, and for novel fusion breakpoints. The recurrence of the *VTI1A-TCF7L2* fusion, and a new fusion transcript between *TCF7L2* and the novel partner gene *RP11-57H14.3*, was analyzed in a series of CRCs and normal samples.

Materials and Methods

Material

A total of 106 CRC tissue samples and 14 paired normal samples from two independent patient series, Series 1 and Series 2, were screened for presence of the fusion transcripts. All CRCs are from patients treated surgically at hospitals in the Oslo region, Norway. The two patient series included both microsatellite stable (n = 85) and instable (n = 20) CRCs (one sample not scored), as assessed by previous studies [8–10]. The series were enriched for clinical stages II and III CRC (52 stage II, 53 stage III, and 1 stage IV). Staging was in accordance to the American Joint Cancer Committee/Union for International Cancer Control (AJCC/UICC). The 14 paired normal samples from Series 1 were collected from visually disease free areas of colonic mucosa. The research biobanks have been registered according to national legislation, and the study has been approved by the Regional Committee for Medical Research Ethics (numbers 2781 and 236-2005-16141).

A total of 21 colorectal cell lines were included [11]. The cell lines HCT116, HCT15, LoVo, NCI-H508, RKO, SW48, and SW620 were purchased from the American Type Culture Collection (ATCC, Manassas, VA, USA). The cell lines Co115, Colo320, EB, FRI, HT29, IS1, IS2, IS3, LS1034, LS174T, SW480, TC7, TC71, and V9P were kindly provided by Dr. Richard Hamelin, Inserm, France. Identities of the cell lines were verified by the AmpFLSTR Identifiler PCR Amplification Kit (Applied Biosystems by Life Technologies, Carlsbad, CA, USA).

Additionally, we screened twenty normal tissues from miscellaneous sites of the body for the *TCF7L2* involving fusion transcripts (adipose, bladder, brain, cervix, colon, esophagus, heart, kidney, liver, lung, ovary, placenta, prostate, skeletal muscle, spleen, stomach, testes, thymus, thyroid and trachea; FirstChoice Human Normal Tissue Total RNA, each a pool of RNA from at least three individuals, with the exception of an individual sample from the stomach; Ambion, Applied Biosystems by Life Technologies, Carlsbad, CA, USA).

Identification of Fusions from Whole-transcriptome and Whole-genome Sequencing Data

Paired-end RNA-sequencing data from seven colorectal cancer cell lines (HCT15, HCT116, HT29, LS1034, RKO, SW48, and SW480) and from 16 normal tissues of miscellaneous origins were included in the study. These were all sequenced using the Illumina GAIIx (cell lines) or HiSeq 2000 (normal tissues; both sequencers from Illumina Inc., San Diego, CA, USA). The colon cancer RNA-seq data included 220 million 76 bp sequence reads (European Nucleotide Archive study accession ID ERP002049) [12] and the normal samples included between 73 and 80 million 50 bp sequence reads per sample (ArrayExpress accession id [E-MTAB-513] and European Nucleotide Archive study [EMBL:ERP000546]). Additionally, we obtained paired-end whole-genome sequences of the cell lines HCT15, HCT116, HT29 and SW480 to an average coverage of about 30×[12] (Data may be made available upon request to researchers).

A list of potential fusion transcripts was produced by deFuse version 0.6.0 [13] on the seven previously mentioned RNA-sequenced cell lines in addition to the CRC cell line NCI-H508 (analysis id 0c7a79cc-bbaf-4c6d-93e1-866f5f5f3d0d) downloaded from Cancer Genomics Hub (hosted by the University of California Santa Cruz, CA, USA), data provided by the Cancer Cell Line Encyclopedia [14] and 16 tissues from the Illumina Human Body Map v2. For the cell lines with paired whole-genome sequence (HCT116, HCT15, HT29 and SW480), RNA fusion transcripts and DNA breakpoints were identified by nFuse version 0.2.0 [15]. A fusion nomination required three spanning read pairs and two split reads from the RNA-seq data.

Detection of Fusion Transcripts by Reverse-transcriptase PCR

For the *VTI1A-TCF7L2* fusion, the same RT-PCR primers and nested primers were used as in the original publication [1]. For the *TCF7L2-RP11-57H14.3* fusion, primers and nested primers were designed to the fusion transcript breakpoint sequence, as identified by deFuse, by utilizing the Primer3 web software [16]. The optimal primer sequences (Table S1) were further ordered from BioNordika Norway AS (Oslo, Norway) and synthesized by Eurogentec (Liège, Belgium). We performed sensitive nested RT-PCR with 20+30 cycles for both assays using 50 ng of cDNA from each sample as input into first round PCR. The products of the first PCR were diluted to a final concentration of 1/200 in the nested PCR. The following cycling protocol was used for all PCR reactions: 15 minutes of HotStarTaq DNA polymerase activation at 95°C, a three-step cycle of denaturation for 30 seconds at 95°C, primer annealing for one minute and 15 seconds at optimal primer melting temperatures (Table S1), and extension for one minute at 72°C. After the last cycle, a final extension step was performed at 72°C for 6 minutes. The nested-PCR products were separated by electrophoresis at 200 V for 30 minutes on a 2% agarose gel and visualized using ethidium bromide and UV light. Triplicate RT-PCR runs were performed for both assays, using identical parameters, for all samples. No template negative controls were included from each cDNA synthesis, first-round PCR and nested-PCR reactions.

Sanger Sequencing of Fusion Transcript Breakpoints

Samples that were positive for the second replicate RT-PCR assays, and showed a single nucleotide band on the agarose gel, were sequenced directly from both sides using both forward and reverse nested primers. Prior to sequencing, the nested PCR products were purified using Illustra ExoStar 1-step cleanup (GE Healthcare, Little Chalfront, UK). The cycle sequencing reactions were performed using the BigDye Terminator v.3.1 cycle sequencing kit (Applied Biosystems, Foster City, CA, USA) following supplier's recommendation. Further, the sequencing products were cleaned and purified using BigDye Xterminator (Applied Biosystems), before they were analyzed by capillary electrophoresis using the ABI 3730 DNA Analyzer (Applied Biosystems). The sequences were analyzed using the Sequencing Analysis v.5.3.1 software.

Assessment of the Identified Genomic Breakpoint *TCF7L2-RP11-57H14.3* by PCR

The genomic breakpoint identified from whole-genome sequencing data, connecting *TCF7L2* to *RP11-57H14.3* in the colon cancer cell line HCT116, was validated using PCR.

Primers for genomic PCR (Table S1) were designed to the genomic breakpoint sequence, as identified by nFuse, using the same approach as described above for the RT-PCR assays. Twenty-one CRC cell lines, including HCT116, were tested for the identified genomic breakpoint using 100 ng of DNA as input for each reaction. The following cycling protocol was used for the genomic PCR: 15 minutes of HotStarTaq DNA polymerase activation at 95°C, a three-step cycle (repeated 35 times) of denaturation for 30 seconds at 95°C, primer annealing for one minute and 15 seconds at 60°C and extension for one minute at 72°C. After the last cycle, a final extension step was performed at 72°C for 6 minutes. The PCR products were separated by electrophoresis at 200 V for 30 minutes on a 2% agarose gel, and visualized using ethidium bromide and UV light.

Results

RP11-57H14.3 is a Novel Fusion Partner in *TCF7L2* Containing Fusion Transcripts

To search for the *VTI1A-TCF7L2* fusion and novel fusion partners, we analyzed paired-end RNA-sequencing data from eight colon cancer cell lines and sixteen normal tissue samples from miscellaneous anatomical sites. By applying the software deFuse, we identified between 16 and 1050 potential fusion transcripts per sample. From whole genome sequencing data of four colon cancer cell lines, the nFuse software identified between 70 and 126 potential genomic breakpoints. Of all fusions identified, only NCI-H508 harbored the *VTI1A-TCF7L2* fusion transcript, with a single breakpoint spanning from exon two in *VTI1A* to exon six in *TCF7L2*, as already reported for this cell line by Bass et al. [1]. However, we identified a genomic breakpoint (chr10:114,850,371-114,640,318 (GRCh37)) in the CRC cell line HCT116 that spanned from the intronic region of *TCF7L2* to upstream of *RP11-57H14.3* (ENSG00000225292) with a corresponding RNA fusion (Table 1). *RP11-57H14.3* is located in the intergenic region between *VTI1A* and *TCF7L2* approximately 44 kb upstream of *TCF7L2*. The predicted genomic breakpoints correlate with increased genomic coverage in the region between them, suggesting a genomic duplication causing the fusion (Figure 1). Both the genomic breakpoint and fusion transcript between *TCF7L2* and *RP11-57H14.3* were verified by PCR and RT-PCR. The identified genomic breakpoint was verified in the HCT116 cell line, but not detected in any of the 20 other cell lines tested, thereby reducing the likelihood that the genomic breakpoint reflects a common DNA copy number polymorphism.

High Prevalence of *TCF7L2*-involving Fusion Transcripts, Both Involving *VTI1A* and the *RP11-57H14.3* Genes

Using the same primers as described by Bass et al. [1] (Figure 2), we detected *VTI1A-TCF7L2* fusion transcripts with different exon-exon combinations in 45 out of 106 CRC samples (42%) (Table 2). Fusion transcripts were as well detected from 4 of 14 normal colonic mucosa samples. Out of the 14 paired tumor-normal samples, eight pairs were positive exclusively in tumor, three pairs were positive in both tumor and normal, and one pair positive exclusively in normal (Table 3). Prevalence of fusion transcripts was similar in microsatellite stable and instable tumors. Ten out of 21 cell lines, including the NCI-H508, harbored fusion transcripts of different sizes. Finally, five normal tissue samples from different anatomical sites of the body expressed the fusion transcripts (Table S2). Because the RT-PCR results revealed inconsistent results (Figure S1 and Figure S2), we performed all RT-PCRs in triplicate, with identical parameters, to investigate recurrence of fusion transcripts (Table S3). Fusion transcripts were only detected consistently in all three runs in four samples; three tumor samples and the NCI-H508 cell line (3.1% of all CRC samples and cell lines). This frequency is similar to the originally identified frequency of *VTI1A-TCF7L2* transcripts (3%) by Bass et al. [1].

We detected *TCF7L2-RP11-57H14.3* fusion transcripts with different exon-exon combinations in 48 out of 106 CRC samples (45%; Table 2). Out of the 14 paired tumor-normal samples, five pairs were positive exclusively in tumor, and four pairs were positive in both tumor and normal (Table 3). 19 out of 21 cell lines, harbored fusion transcripts of different sizes. Also, 15 out of 20 normal tissue samples were positive for fusion transcripts (Table S2). As with the *VTI1A-TCF7L2* fusion, we performed all RT-PCRs in triplicate to investigate the recurrence of fusion transcripts (Table S3). In total there were 20 samples that repeatedly tested positive for fusion transcripts; six tumor samples, five samples from normal tissues and nine cell lines (including the HCT116 cell line). Interestingly, the NCI-H508 cell line was negative for *TCF7L2-RP11-57H14.3* fusion transcripts in all three replicates.

We found no correlation between CRCs positive for *TCF7L2* containing fusion transcripts involving *VTI1A* and *RP11-57H14.3* ($p = 0.33$; Fisher's Exact test). Further, the frequencies of fusion transcripts were not significantly different between microsatellite instable *vs.* stable tumors, nor between clinical stages (data not shown).

All nested-PCR replicates for both assays contained negative no template controls. These negative control reactions never produced detectable PCR-products (Figure S1 and S2).

Chimeric Sequences Generally Covered Intact Exonic Splice Sites from the Partner Genes

From one of the RT-PCR runs, all samples that were positive for the fusions and had a single PCR product were selected for Sanger-sequencing of the chimeric RNA-sequences to identify the exact breakpoints. For *VTI1A-TCF7L2* (n = 25), we obtained sequences from all 25 such isolated fusion transcript RT-PCR products, where 24 out of 25 had sequences connecting upstream sequences of the exons 1, 2, 3, 5 or 7 of *VTI1A* to downstream sequences of the exons 4 or 6 of *TCF7L2*, with preservation of the same exon-exon boundaries as in the already annotated gene structures (ENST00000393077 in *VTI1A* and ENST00000369395 in *TCF7L2*). One sequenced product did not contain clear exon-exon boundaries, but connected the two transcripts in middle of exon 7 in *VTI1A* and 6 in *TCF7L2*. In total, seven different fusion breakpoints were identified (Table S3) including the original breakpoint between exon 2 of *VTI1A* and exon 6 of *TCF7L2* in NCI-H508 (Figure 3) [1]. This exact breakpoint was not found in any of the other samples, but most of the intact fusion transcripts consisted of the same part of *TCF7L2* (n = 17) with some having exon 4 of *TCF7L2* spliced to exon 6 (n = 7). The *VTI1A* upstream contribution varied more, having five different combinations connecting to downstream *TCF7L2* parts.

For *TCF7L2-RP11-57H14.3* (n = 27), we obtained sequences from all 27 isolated fusion transcript RT-PCR products, where 26 out of 27 had sequences connecting sequences of exon 4 of *TCF7L2* to sequences of exons 1, 2, or 3 of *RP11-57H14.3*, also with preservation of the same exon-exon boundaries as in the already annotated gene structures (exon numbering is the same as

Table 1. nFuse and defuse: Verification of the original *VTI1A-TCF7L2* fusion transcript and identification of a fusion transcript and genomic breakpoint involving *TCF7L2* and *RP11-57H14.3*.

Cell line	GeneA	GeneB	Software	Split[†]	Spanning[†]	Score
HCT116	*RP11-57H14.3*	*TCF7L2*	deFuse	3	4	0,84[a]
			nFuse	NA	27	5,5[b]
NCI-H508	*VTI1A*	*TCF7L2*	deFuse	64	27	0,93[a]

RP11-57H14.3 is located 44 kb upstream of *TCF7L2* on chromosome 10.

a)deFuse probability score,

b)nFuse path score. †) Split reads contain the fusion boundary in the read itself, while spanning reads are paired ends that harbor the fusion boundary within the insert sequence.

doi:10.1371/journal.pone.0091264.t001

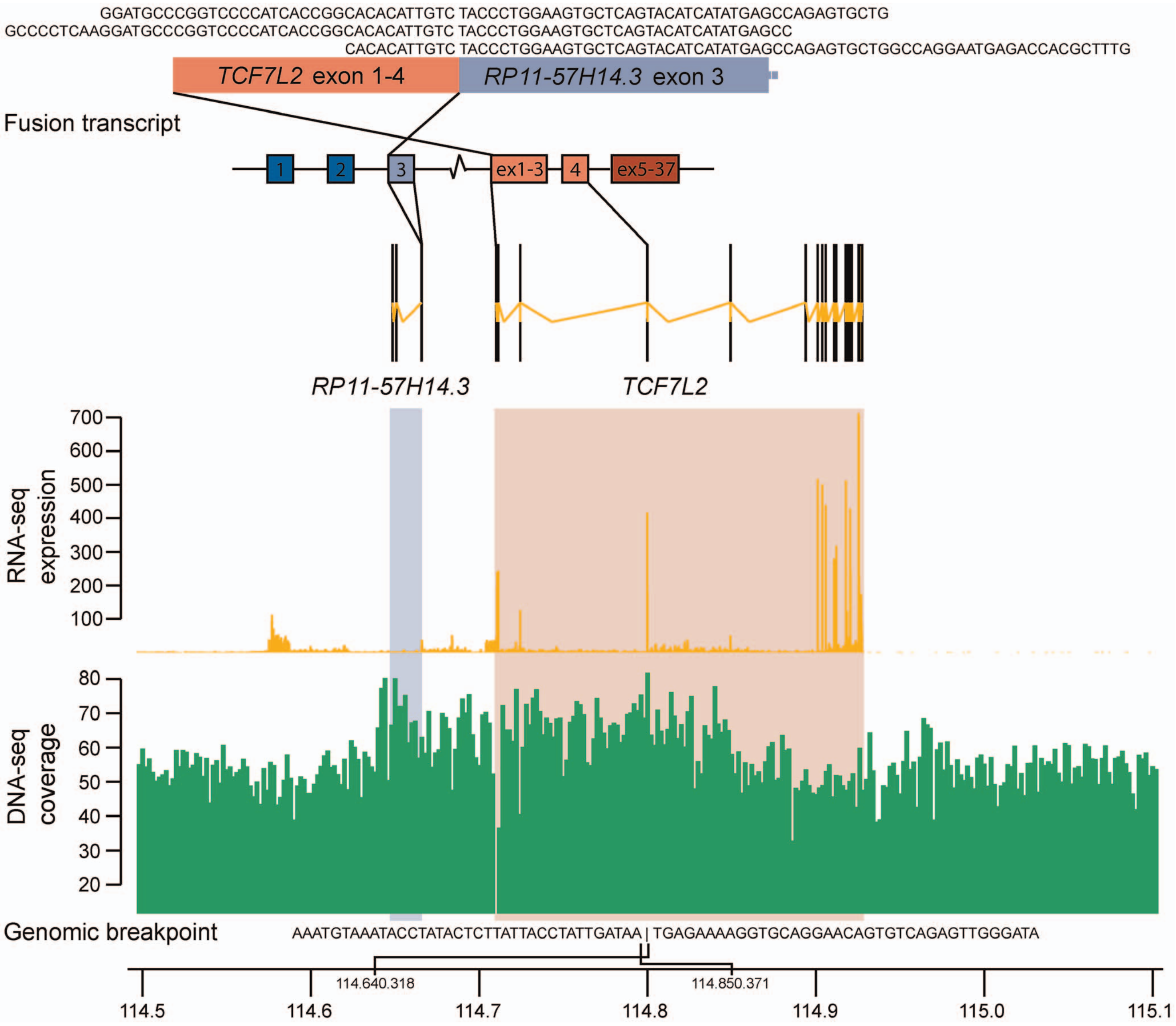


Figure 1. The *TCF7L2-RP11-57H14.3* fusion transcript, identified in the HCT116 cell line, harbors a rearranged genomic locus. Three chimeric RNA sequence-reads spanned the fusion transcript breakpoint, passing from exon 4 of *TCF7L2* (ENST00000369395) to exon 3 of *RP11-57H14.3* (ENST00000428766), on chromosome 10. Dark colors indicate exons not part of the fusion transcript. RNA-seq expression and DNA-seq coverage levels are based on sequencing data of the HCT116 cell line. The two gene loci are marked in blue and red boxes. The genomic breakpoint sequence as identified by nFUSE in the CRC cell line HCT116 is given; spanning from the intronic region of *TCF7L2* to upstream of *RP11-57H14.3*. The coordinates of the breakpoint (chr10:114,850,371-114,640,318 (GRCh37)) are marked on the chromosome position axis. The location of the breakpoint correlates well with the increased genomic coverage seen from the genome sequencing data.

doi:10.1371/journal.pone.0091264.g001

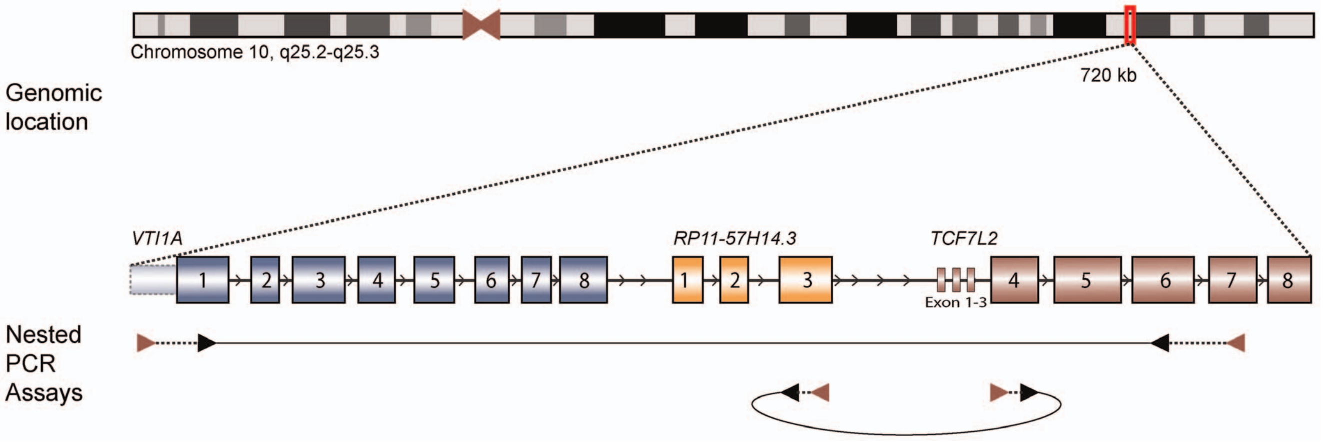


Figure 2. Schematic presentation of the genomic location of *VTI1A, RP11-57H14.3* and *TCF7L2*. All three genes are located within 720 kb on the same strand on the long arm of chromosome 10. The exon numbering refers to exons annotated in transcripts ENST00000393077 in *VTI1A*, ENST00000428766 in *RP11-57H14.3* and ENST00000369395 in *TCF7L2*. Also, the nested PCR assays used for detection of both fusion transcripts are shown. The red and black arrows represent the first round and second round primers used, respectively.
doi:10.1371/journal.pone.0091264.g002

above for *TCF7L2* and according to ENST00000428766 for *RP11-57H14.3*). One curious case of chimeric sequence, identified in the CRC cell line FRI, was a fusion transcript spanning three genes in a non-canonical genomic order, joining *TCF7L2* exon 4 with *VTI1A* exons 5, 6, and 7, and further extending into *RP11-57H14.3* exons 1, 2, and 3. Also in this unique case the same exon-exon boundaries were used as in the already annotated gene structures described above. In total, four different fusion transcripts were identified involving *TCF7L2-RP11-57H14.3,* all occurring in several samples each.

Discussion

We have in the present report identified the gene *RP11-57H14.3* as a novel fusion partner for *TCF7L2* in CRC. By sensitive nested RT-PCR, we have revealed that both the previously reported *VTI1A-TCF7L2* fusion transcripts, and the herein identified *TCF7L2-RP11-57H14.3*, are highly frequent among CRCs, although expressed at low levels. We also detected expression of the fusion transcripts in normal colonic mucosa, as well as in normal tissues from other anatomical sites. Triplicate nested-PCRs to investigate the presence of *TCF7L2* involving fusion transcripts showed variable results, with some samples initially testing positive for a fusion transcript, but negative when performing a consecutive run, or *vice-versa*. However, Sanger sequencing resulted in confirmation of clean exon to exon breakpoint junctions, reducing the likelihood that PCR artifacts, such as polymerase template switching [17], as a cause for the inconsistency. Negative controls were also performed at all RT-PCR steps, including cDNA synthesis, first-round PCR and nested-PCR. None of the negative controls resulted in detectable products, supporting that the high frequency of fusion transcripts observed is not a result of PCR contamination (Figure S1 and Figure S2). Although the fusion transcripts were found at high frequency within the tested biobank materials, the identification of *TCF7L2*-containing fusions from whole-transcriptome sequencing data was only successful from the two cell lines with matching genomic breakpoints. These two cell lines did as well have consistently strong expression of the fusion transcripts, as they produced consistent and clear RT-PCR results in all replicates. Based on these results, we suggest that the fusion transcripts produced between *TCF7L2* and either *VTI1A* or *RP11-57H14.3* are expressed at low levels in tumor samples, and some normal samples. The presence of fusion transcripts expressed at low levels are consistent with three fusion transcripts we recently identified in CRC, where *AKAP13-PDE8A, COMMD10-AP3S1,* and *CTB-35F21.1-PSD2* were identified in 17–58% of 106 primary cancer tissues [12].

All three partner genes are located on the long arm of chromosome 10, within 721 kbp, and are all read from the same strand in the order *VTI1A, RP11-57H14.3, TCF7L2* (Figure 2). This suggests RNA polymerase read-through as a potential

Table 2. The number of positive fusion transcript PCRs for the samples, run in triplicates.

	n =	Strong Positives^a: *VTI1A-TCF7L2*	Positives^b: *VTI1A-TCF7L2*	Strong Positives^a: *TCF7L2-RP11-57H14.3*	Positives^b: *TCF7L2-RP11-57H14.3*
Series 1-tumor	14	1 (7.1%)	11 (79%)	4 (29%)	9 (64%)
Series 1-normal	14	0	4 (29%)	0	4 (29%)
Series 2	92	2 (2.2%)	34 (37%)	2 (2.2%)	39 (42%)
Cell lines	21	1 (4.8%)	11 (52%)	9 (43%)	19 (90%)
Normals	20	0	5 (25%)	5 (25%)	15 (75%)

a)Strong positives are defined as testing positive in all RT-PCR replicates.
b)Samples noted as positives have tested positive for the fusion transcript(s) in one or more of the three RT-PCR replicates.
doi:10.1371/journal.pone.0091264.t002

Table 3. Matched tumor and normal colonic mucosa from series 1.

	VTI1A-TCF7L2		*TCF7L2-RP11-57H14.3*	
Pair #	**Tumor**	**Normal**	**Tumor**	**Normal**
1	Y	Y	N	N
2	Y	N	N	N
3	Y	N	Y	Y
4	Y	N	Y	N
5	Y	N	Y	Y
6	Y	N	Y	N
7	N	N	N	N
8	N	N	Y	N
9	Y	Y	Y	Y
10	N	Y	Y	Y
11	Y	N	N	N
12	Y	N	N	N
13	Y	N	Y	N
14	Y	Y	Y	N

For both fusion transcripts the tumor, or both the tumor and normal samples were frequently positive. In one pair the *VTI1A-TCF7L2* fusion transcript was positive only in the normal sample and not the matched tumor.
doi:10.1371/journal.pone.0091264.t003

mechanism for generating the *VTI1A-TCF7L2* transcripts in the absence of a corresponding genomic breakpoint. Several reports have shown that genes in close proximity in the human genome are expressed as conjoined genes, also called tandem chimeras, transcripts that are combined of at least part of one exon from two or more distinct genes that lie on the same chromosome [18–20]. It has been suggested that the expression of conjoined genes increase the complexity of the human genome by translating into distinct proteins, or that these transcripts play a role in regulation of canonical transcript levels. One suggested mechanism for their generation is that the transcription machinery avoids the termination signal of the upstream gene and continues transcribing the downstream gene before terminating at the downstream termination signal [19,21]. The majority of the conjoined genes are believed to be expressed at low levels, as they are often supported by only a single expressed sequence tag or mRNA sequence in genome databases, which is in line with our observations of weakly expressed *TCF7L2*-containing fusion transcripts in CRC.

The generation of *VTI1A-TCF7L2* transcripts may well be explained as a product of polymerase read-through, but this cannot be the mechanism of operation for the *TCF7L2-RP11-57H14.3* fusion transcripts. In this case, the exons of *TCF7L2* and *RP11-57H14.3* are spliced together in a non-canonical genomic order using consensus splice-sites. This transcriptional mechanism has previously been described by Nigro et al. as exon scrambling; a process where exons are joined accurately at consensus splice sites, but in an order different from that present in the primary transcript [22]. They discovered this phenomenon when investigating a candidate tumor suppressor gene (*DCC*), and identified that the resulting scrambled transcripts are expressed and found at relatively low levels in both normal and neoplastic cells. Recently,

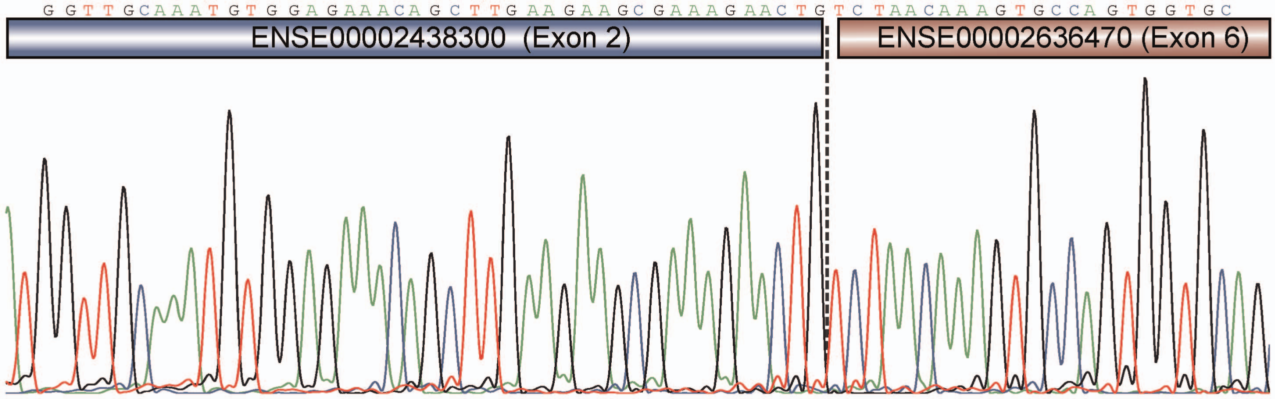


Figure 3. Confirmation of the original *VTI1A-TCF7L2* fusion transcript breakpoint in the cell line NCI-H508. Sanger sequencing confirmed the original fusion transcript discovered in NCI-H508, showing the breakpoint sequence spanning exon-exon junctions between exon 2 in *VTI1A* and exon 6 in *TCF7L2*. Bass et al. also discovered three other fusion transcripts by nested-PCR. However, as transcript annotation was not sufficiently described, we are not able to say if these fusions are identical to some of the transcripts we have identified.
doi:10.1371/journal.pone.0091264.g003

based on deep sequencing of RNA, such scrambled transcripts from hundreds of genes were identified [23]. This group also suggested that a substantial fraction of the scrambled transcripts are circular RNAs, which explain the joining of exons in a non-canonical linear order. They also found that many of the circular isoforms were present at levels comparable to their canonical linear counterparts.

Altogether, these added levels of transcriptional and genomic complexity is in line with the recent report of the ENCODE project, reporting substantial reduction in the lengths of intergenic regions, and increasingly overlapping of genes previously assumed to be distinct genetic loci, altogether prompting a redefinition of the concept of a gene [24].

The identification of genomic breakpoints in individual cell lines for *TCF7L2* fusions both involving *VTI1A* and *RP11-57H14.3*, is intriguing. The genomic breakpoints identified coincide well with the frequently observed fusion transcripts discovered in other samples which do not have such genomic rearrangements. For the cell lines with identified genomic breakpoints, the fusion transcripts were detected at strong levels in all RT-PCR replicates, suggesting that these cells express the fusion transcripts at higher levels. Furthermore, the cell line with *VTI1A-TCF7L2* genomic fusion (NCI-H508) was negative for the *TCF7L2-RP11-57H14.3* fusion transcripts in all replicates, supporting the ~540 kb genomic deletion of the intergenic region between *VTI1A* and *TCF7L2* originally identified by Bass et al.

The presence of the fusion transcripts in both normal cells and CRC samples together with identification of genomic breakpoints from individual cancer cell lines, joining the same two genes on the genome level, are in line with the report of the fusion transcript *JAZF1-JJAZ1* identified in both normal endometrial stromal cells and endometrial stromal tumors [25]. *JAZF1-JJAZ1* is detectable at the transcript level in normal endometrial stromal cells but genomic rearrangements are found only in the neoplastic cells. Li et al. hypothesize that trans-splicing generating these fusion transcripts, as well as other fusion transcripts, may occur regularly in normal cells and tissues. Further, they suggest that there may be a link between trans-splicing generating these fusion transcripts and the generation of the genomic rearrangements [26]. The genomic rearrangements or other mechanisms may lead to overexpression of these fusion transcripts, which may have oncogenic potential.

The importance of the gene *TCF7L2* in CRC development is favored by its function. *TCF7L2* encodes the TCF4 transcription factor, which is a key down-stream transcription factor in the WNT/β-catenin-signaling pathway, altered in the majority of CRCs [5]. The original report of *VTI1A-TCF7L2* found that depletion of the fusion transcript resulted in significant loss of anchorage independent growth in the fusion positive CRC cell line NCI-H508 [1]. Frequent mutations in a polyadenine tract within the *TCF7L2* gene have previously been reported for CRC, especially in microsatellite instable tumors [6]. Furthermore, microsatellite stable tumors have recently been shown to harbor frequent mutations in *TCF7L2* [7]. The observation that *TCF7L2* involving fusion transcripts are detectable in such a large fraction of CRC samples, but also normal colonic mucosa and other normal tissue types, reveals that the *TCF7L2* fusion transcripts are neither specific to cancer nor to the colon and rectum. Hence, they do not have the potential as cancer detection biomarkers as originally expected. When that is said, the similar fusion transcripts induced by genomic rearrangements observed in individual cancer cell lines may yet have oncogenic potential as suggested in the original study by Bass et al. The phenomenon of genomic rearrangements observed in cancer cells that correlate with transcription induced chimeras observed in most normal cells is in itself intriguing and needs to be explored further.

In conclusion, we have identified the gene *RP11-57H14.3* as a novel fusion partner of *TCF7L2*. Both this fusion transcript and the previously reported *VTI1A-TCF7L2* are processed into several different splice variants, and *TCF7L2* involving fusion transcripts are expressed at detectable levels, in a high proportion of CRCs, and also in normal tissues from both colonic mucosa and from other anatomical sites. We suggest that these fusion transcripts are transcription induced chimeras, but that individual cancer cells have genomic rearrangements that lead to expression of highly similar fusion transcripts that potentially have a role in cancer development or progression.

Supporting Information

Figure S1 *VTI1A-TCF7L2*: Nested-PCR products in tumor samples, matched normals and CRC cell lines run in triplicate and analyzed on 2% agarose gels. The results show a much higher degree of fusion-transcript positives than what has previously been reported for *VTI1A-TCF7L2*. However, the results diverge somewhat from run to run. A) Nested-PCR results from patient series 1 and 2. B) Nested-PCR results from the 21 CRC cell lines and some negative controls. The nested-PCR product of NCI-H508 seems more abundant compared to the other products based on the band luminescence. C) Additional negative controls, including no template controls from cDNA synthesis, first-round and second-round PCR.
(TIF)

Figure S2 *TCF7L2-RP11-57H14.3*: Nested-PCR products in tumor samples, matched normals and CRC cell lines run in triplicate and analyzed on 2% agarose gels. The results diverge somewhat for each replicate. A) Nested-PCR results from patient series 1 and 2. B) Nested-PCR results from the 21 CRC cell lines and some negative controls. There are at least three PCR-products from the cell line HCT116, which are present and identical in all replicates. C) Additional negative controls, including no template controls from cDNA synthesis, first-round and second-round PCR.
(TIF)

Table S1 Primers used in this study.
(DOCX)

Table S2 Both fusion transcripts involving *TCF7L2* were detected, by RT-PCR, in several human tissues.
(DOCX)

Table S3 Total overview of nested RT-PCR and Sanger sequencing confirmation of fusion transcripts.
(DOCX)

Acknowledgments

We are grateful for all scientific input as well as resources shared from Professor Ragnhild A. Lothe and her laboratory.

Author Contributions

Conceived and designed the experiments: TN AMH RIS. Performed the experiments: TN AMH ACB. Analyzed the data: TN AMH. Contributed reagents/materials/analysis tools: TN AMH TOR AN RIS. Wrote the paper: TN AMH RIS. Read and approved the final version of the manuscript: TN AMH ACB TOR AN RIS.

References

1. Bass AJ, Lawrence MS, Brace LE, Ramos AH, Drier Y, et al. (2011) Genomic sequencing of colorectal adenocarcinomas identifies a recurrent VTI1A-TCF7L2 fusion. Nat Genet 43: 964–968. doi:10.1038/ng.936.
2. Ferlay J, Shin H-R, Bray F, Forman D, Mathers C, et al. (2010) Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int J Cancer 127: 2893–2917. doi:10.1002/ijc.25516.
3. Sawyers CL (1999) Chronic myeloid leukemia. N Engl J Med 340: 1330–1340. doi:10.1056/NEJM199904293401706.
4. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, et al. (2005) Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science 310: 644–648. doi:10.1126/science.1117679.
5. Segditsas S, Tomlinson I (2006) Colorectal cancer and genetic alterations in the Wnt pathway. Oncogene 25: 7531–7537. doi:10.1038/sj.onc.1210059.
6. Thorstensen L, Lind GE, Løvig T, Diep CB, Meling GI, et al. (2005) Genetic and epigenetic changes of components affecting the WNT pathway in colorectal carcinomas stratified by microsatellite instability. Neoplasia 7: 99–108. doi:10.1593/neo.04448.
7. Cancer Genome Atlas Network (2012) Comprehensive molecular characterization of human colon and rectal cancer. Nature 487: 330–337. doi:10.1038/nature11252.
8. Lothe RA, Peltomäki P, Meling GI, Aaltonen LA, Nyström-Lahti M, et al. (1993) Genomic instability in colorectal cancer: relationship to clinicopathological variables and family history. Cancer Res 53: 5849–5852.
9. Diep CB, Thorstensen L, Meling GI, Skovlund E, Rognum TO, et al. (2003) Genetic tumor markers with prognostic impact in Dukes' stages B and C colorectal cancer patients. J Clin Oncol 21: 820–829.
10. Berg M, Danielsen SA, Ahlquist T, Merok MA, Ågesen TH, et al. (2010) DNA sequence profiles of the colorectal cancer critical gene set KRAS-BRAF-PIK3CA-PTEN-TP53 related to age at disease onset. PLoS ONE 5: e13978. doi:10.1371/journal.pone.0013978.
11. Ahmed D, Eide PW, Eilertsen IA, Danielsen SA, Eknæs M, et al. (2013) Epigenetic and genetic features of 24 colon cancer cell lines. Oncogenesis 2: e71. doi:10.1038/oncsis.2013.35.
12. Nome T, Thomassen GO, Bruun J, Ahlquist T, Bakken AC, et al. (2013) Common fusion transcripts identified in colorectal cancer cell lines by high-throughput RNA sequencing. Transl Oncol 6: 546–553.
13. McPherson A, Hormozdiari F, Zayed A, Giuliany R, Ha G, et al. (2011) deFuse: an algorithm for gene fusion discovery in tumor RNA-Seq data. PLoS Comput Biol 7: e1001138. doi:10.1371/journal.pcbi.1001138.
14. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, et al. (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483: 603–607. doi:10.1038/nature11003.
15. McPherson A, Wu C, Wyatt AW, Shah S, Collins C, et al. (2012) nFuse: discovery of complex genomic rearrangements in cancer using high-throughput sequencing. Genome Res 22: 2250–2261. doi:10.1101/gr.136572.111.
16. Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132: 365–386.
17. Odelberg SJ, Weiss RB, Hata A, White R (1995) Template-switching during DNA synthesis by Thermus aquaticus DNA polymerase I. Nucleic Acids Res 23: 2049–2057.
18. Prakash T, Sharma VK, Adati N, Ozawa R, Kumar N, et al. (2010) Expression of conjoined genes: another mechanism for gene regulation in eukaryotes. PLoS ONE 5: e13284. doi:10.1371/journal.pone.0013284.
19. Parra G, Reymond A, Dabbouseh N, Dermitzakis ET, Castelo R, et al. (2006) Tandem chimerism as a means to increase protein complexity in the human genome. Genome Res 16: 37–44. doi:10.1101/gr.4145906.
20. Akiva P, Toporik A, Edelheit S, Peretz Y, Diber A, et al. (2006) Transcription-mediated gene fusion in the human genome. Genome Res 16: 30–36. doi:10.1101/gr.4137606.
21. Kim RN, Kim A, Choi S-H, Kim D-S, Nam S-H, et al. (2012) Novel mechanism of conjoined gene formation in the human genome. Funct Integr Genomics 12: 45–61. doi:10.1007/s10142-011-0260-1.
22. Nigro JM, Cho KR, Fearon ER, Kern SE, Ruppert JM, et al. (1991) Scrambled exons. Cell 64: 607–613.
23. Salzman J, Gawad C, Wang PL, Lacayo N, Brown PO (2012) Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. PLoS ONE 7: e30733. doi:10.1371/journal.pone.0030733.
24. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, et al. (2012) Landscape of transcription in human cells. Nature 489: 101–108. doi:10.1038/nature11233.
25. Li H, Wang J, Mor G, Sklar J (2008) A neoplastic gene fusion mimics trans-splicing of RNAs in normal human cells. Science 321: 1357–1361. doi:10.1126/science.1156725.
26. Li H, Wang J, Ma X, Sklar J (2009) Gene fusions and RNA trans-splicing in normal and neoplastic human cells. Cell Cycle 8: 218–222.

Supporting information

Supporting Figures

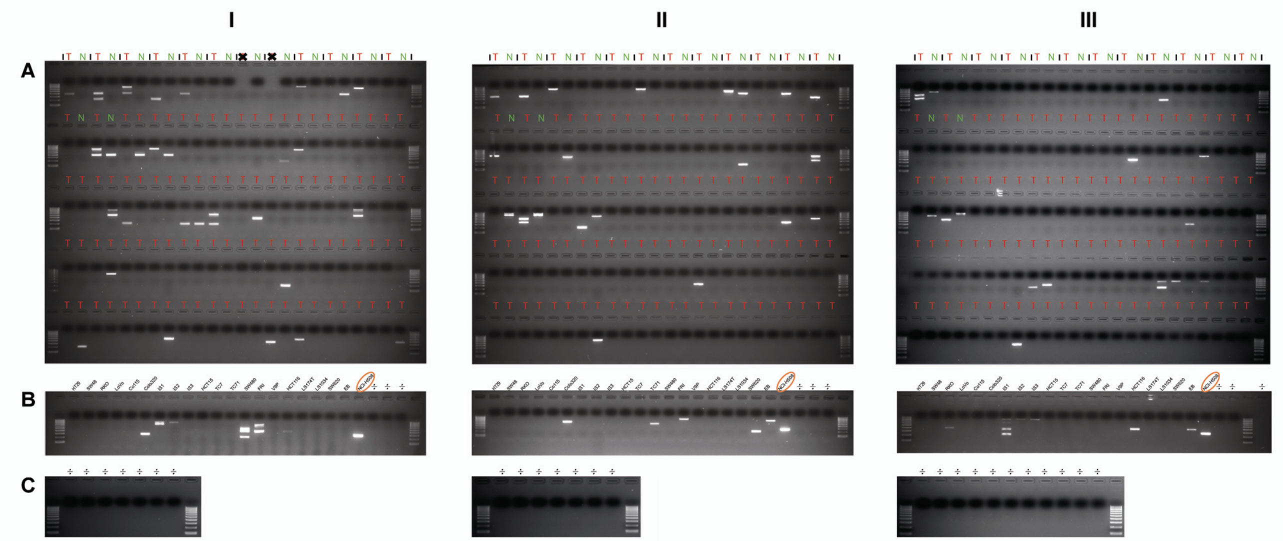


Figure S1: *VTI1A-TCF7L2*: Nested-PCR products in tumor samples, matched normals and CRC cell lines run in triplicate and analyzed on 2% agarose gels. The results show a much higher degree of fusion-transcript positives than what has previously reported for *VTI1A-TCF7L2*. However, the results diverge somewhat from run to run. **A)** Nested-PCR results from patient series 1 and 2. **B)** Nested-PCR results from the 21 CRC cell lines and some negative controls. The nested-PCR product of NCI-H508 seems more abundant compared to the other products based on the band luminescence. **C)** Additional negative controls, including no template controls from cDNA synthesis, first-round and second-round PCR.

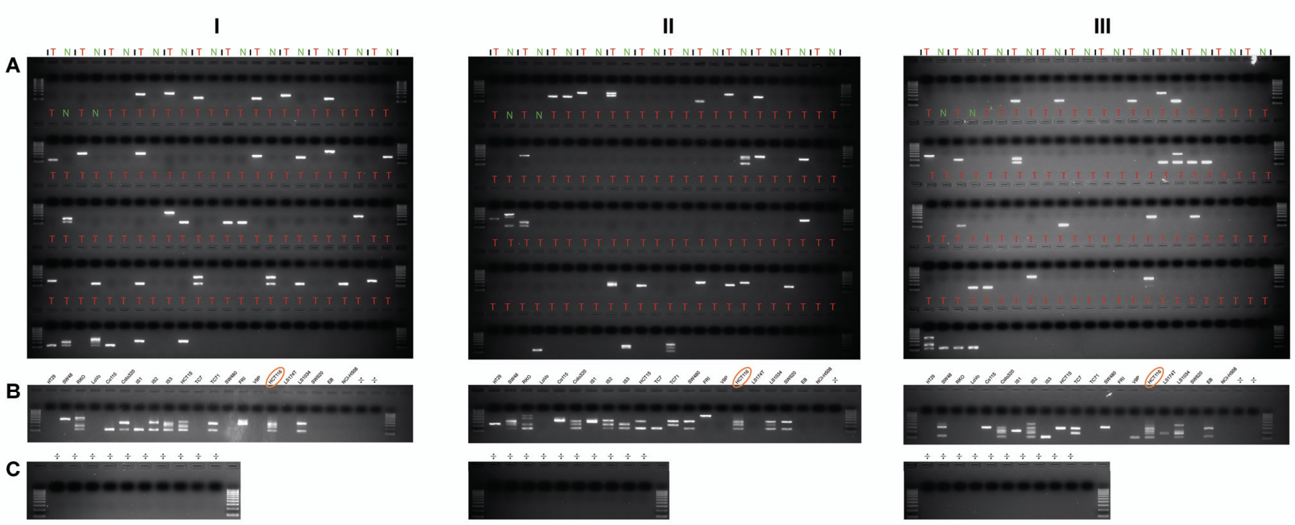


Figure S2: ***TCF7L2-RP11-57H14.3*****: Nested-PCR products in tumor samples, matched normals and CRC cell lines run in triplicate and analyzed on 2% agarose gels.** The results diverge somewhat for each replicate. **A)** Nested-PCR results from patient series 1 and 2. **B)** Nested-PCR results from the 21 CRC cell lines and some negative controls. There are at least three PCR-products from the cell line HCT116, which are present and identical in all replicates. **C)** Additional negative controls, including no template controls from cDNA synthesis, first-round and second-round PCR.

Supporting Tables

Table S1: Primers used in this study.

Primer name:	5’-3’ primer sequence:	Melting temperature (Tm[a])
VTI1A_5’UTR_F (first round)	TTTCCCTGACCTAGGCTTTG	62°C
TCF7L2_Ex6_R (first round)	GGATGGGGGATTTGTCCTAC	62°C
VTI1A_Ex1_F (second round)	CCGACTTCGAAGGTTACGAG	62°C
TCF7L2_ex5_R (second round)	TACGTCGGCTGGTAAGTGTG	62°C
RP11-57H14.3_F1 (first round)	TCCTGGAGATGCCTCTGAGT	58°C
TCF7L2_R1 (first round)	CTACCTCCCCAACGGATCG	58°C
RP11-57H14.3_F2 (second round)	CAAAGCGTGGTCTCATTCCT	57°C
TCF7L2_R2 (second round)	CAGGGAGCCTCCAGAGTAGA	57°C
TCF7L2_DNA_F (genomic breakpoint)	TGGGTGCTGTGCTATGTGTT	60°C
RP11_DNA_R (genomic breakpoint)	GGTAGAGGTTGGCTGCAGTT	60°C

a) Melting temperature (Tm) used for optimal primer annealing step during PCR. Annealing temperature was set to the average Tm of the participating primer pair.

Table S2: Both fusion transcripts involving *TCF7L2* were detected, by RT-PCR, in several human tissues.

Tissue	*VTI1A-TCF7L2*	*TCF7L2-RP11-57H14.3*
Adipose	N	Y
Bladder	N	Y
Brain	Y	Y
Cervix	N	Y
Colon	Y	Y
Esophagus	N	Y
Heart	N	Y
Kidney	Y	Y
Liver	N	N
Lung	Y	Y
Ovary	N	Y
Placenta	Y	Y
Prostate	N	N
Skeletal muscle	N	N
Spleen	N	N
Stomach	N	N
Testes	N	Y
Thymus	N	Y
Thyroid	N	Y
Trachea	N	Y

Table S3: Total overview of nested RT-PCR and Sanger sequencing confirmation of fusion transcripts.

Sample	Type	MSI-Status	Patient Series	*VTI1A-TCF7L2* #1	*VTI1A-TCF7L2* #2	*VTI1A-TCF7L2* #3	SUM *VTI1A-TCF7L2* Status	Confirmed exon-exon breakpoint *VTI1A-TCF7L2*	*TCF7L2-RP11-57H14.3* #1	*TCF7L2-RP11-57H14.3* #2	*TCF7L2-RP11-57H14.3* #3	SUM *TCF7L2-RP11-57H14.3*	Confirmed exon-exon breakpoint TCF7L2-RP11-57H14.3
012	tumor	MSI-H	1	1	1	1	**3**	EX3(V)-EX6(T)	0	0	0	**0**	
012_norm	normal		1	0	0	1	**1**		0	0	0	**0**	
029	tumor	MSI-L	1	1	1	0	**2**	EX3(V)-EX6(T)	0	0	0	**0**	
029_norm	normal		1	0	0	0	**0**		0	0	0	**0**	
041	tumor	MSS	1	1	1	0	**2**	EX7(V)-EX4(T)-EX6(T)	0	1	0	**1**	EX4(T)-EX1(RP)-EX3(RP)
041_norm	normal		1	0	0	0	**0**		0	1	0	**1**	EX4(T)-EX1(RP)-EX3(RP)
042	tumor	MSI-L	1	1	0	0	**1**		1	1	1	**3**	EX4(T)-EX1(RP)-EX2(RP)-EX3(RP)
042_norm	normal		1	0	0	0	**0**		0	0	0	**0**	
045	tumor	MSI-H	1	1	0	0	**1**		1	1	0	**2**	
045_norm	normal		1	0	0	0	**0**		0	0	1	**1**	
065	tumor	MSS	1	0	1	0	**1**	EX7(V)-EX4(T)-EX6(T)	1	0	0	**1**	
065_norm	normal		1	0	0	0	**0**		0	0	0	**0**	
067	tumor	MSS	1	NA	0	0	**0**		0	0	0	**0**	
067_norm	normal		1	0	0	0	**0**		0	0	0	**0**	
070	tumor	MSS	1	NA	0	0	**0**		1	1	1	**3**	EX4(T)-EX3(RP)
070_norm	normal		1	0	0	0	**0**		0	0	0	**0**	
072	tumor	MSI-H	1	1	1	0	**2**	EX5(V)-EX6(T)	1	1	1	**3**	EX4(T)-EX1(RP)-EX2(RP)-EX3(RP)
072_norm	normal		1	0	1	1	**2**	EX5(V)-EX6(T)	0	0	1	**1**	
074	tumor	MSS	1	0	0	0	**0**		0	1	0	**1**	EX4(T)-EX2(RP)-EX3(RP)
074_norm	normal		1	1	0	0	**1**		1	0	0	**1**	
086	tumor	MSI-H	1	1	1	0	**2**	no*	0	0	0	**0**	
086_norm	normal		1	0	0	0	**0**		0	0	0	**0**	
092	tumor	MSS	1	0	1	0	**1**	EX3(V)-EX6(T)	0	0	0	**0**	
092_norm	normal		1	0	0	0	**0**		0	0	0	**0**	
096	tumor	MSS	1	0	1	0	**1**	EX5(V)-EX6(T)	1	0	1	**2**	
096_norm	normal		1	0	0	0	**0**		0	0	0	**0**	
098	tumor	MSS	1	1	0	0	**1**		1	1	1	**3**	EX4(T)-EX1(RP)-EX2(RP)-EX3(RP)
098_norm	normal		1	1	0	0	**1**		0	0	0	**0**	
C1011II	tumor	MSS	2	1	1	1	**3**	EX7(V)-EX4(T)-EX6(T)	0	0	0	**0**	
C1022II	tumor	MSI-H	2	1	0	0	**1**		0	0	0	**0**	
C1030II	tumor	MSI-L	2	0	0	0	**0**		0	0	0	**0**	
C1033III	tumor	MSS	2	0	1	0	**1**	EX1(V)-EX6(T)	0	0	0	**0**	
C1034III	tumor	MSS	2	0	1	0	**1**	EX7(V)-EX6(T)	0	0	0	**0**	
C1043II	tumor	MSS	2	1	0	0	**1**		1	0	0	**1**	

Sample	Type	MSI-Status	Patient Series	*VTI1A-TCF7L2* #1	*VTI1A-TCF7L2* #2	*VTI1A-TCF7L2* #3	SUM *VTI1A-TCF7L2* Status	Confirmed exon-exon breakpoint *VTI1A-TCF7L2*	*TCF7L2-RP11-57H14.3* #1	*TCF7L2-RP11-57H14.3* #2	*TCF7L2-RP11-57H14.3* #3	SUM *TCF7L2-RP11-57H14.3*	Confirmed exon-exon breakpoint TCF7L2-RP11-57H14.3
C1049II	tumor	MSS	2	0	0	1	**1**		1	1	0	**2**	EX4(T)-EX1(RP)-EX3(RP)
C1068III	tumor	MSS	2	1	0	0	**1**		1	0	1	**2**	
C1077III	tumor	MSS	2	1	0	0	**1**		0	0	0	**0**	
C1079II	tumor	MSS	2	0	0	0	**0**		0	0	0	**0**	
C1085II	tumor	MSS	2	0	0	0	**0**		1	0	0	**1**	
C1089II	tumor	MSS	2	0	0	0	**0**		0	0	0	**0**	
C1091II	tumor	MSI-L	2	1	0	0	**1**		1	0	0	**1**	
C1102II	tumor	MSS	2	0	0	0	**0**		1	0	0	**1**	
C1103III	tumor	MSI-L	2	0	0	0	**0**		0	0	0	**0**	
C1112III	tumor	MSS	2	0	0	0	**0**		0	0	1	**1**	
C1118III	tumor	MSS	2	0	0	0	**0**		0	0	0	**0**	
C1122II	tumor	MSS	2	0	0	0	**0**		0	0	0	**0**	
C1135II	tumor	MSS	2	0	0	0	**0**		0	0	1	**1**	
C1142III	tumor	MSS	2	0	0	1	**1**		0	0	0	**0**	
C1144III	tumor	MSS	2	1	1	0	**2**	EX3(V)-EX6(T)	0	0	0	**0**	
C1145III	tumor	MSS	2	0	0	0	**0**		1	1	0	**2**	EX4(T)-EX1(RP)-EX2(RP)-EX3(RP)
C1152III	tumor	MSS	2	0	0	0	**0**		0	0	0	**0**	
C1156II	tumor	MSS	2	0	1	0	**1**	EX3(V)-EX6(T)	0	1	0	**1**	EX4(T)-EX1(RP)-EX2(RP)-EX3(RP)
C1159III	tumor	MSS	2	0	0	0	**0**		0	0	0	**0**	
C1165III	tumor	MSS	2	0	1	0	**1**	EX5(V)-EX6(T)	0	0	0	**0**	
C1190II	tumor	MSI-H	2	0	0	0	**0**		0	0	0	**0**	
C1198III	tumor	MSS	2	0	0	0	**0**		1	0	1	**2**	
C1251III	tumor	MSS	2	1	0	0	**1**		1	0	1	**2**	
C1257II	tumor	NA	2	0	0	0	**0**		0	0	1	**1**	
C1263II	tumor	MSS	2	0	0	0	**0**		1	0	0	**1**	
C1264II	tumor	MSS	2	0	0	0	**0**		1	1	1	**3**	EX4(T)-EX3(RP)
C1267III	tumor	MSS	2	0	0	0	**0**		1	0	0	**1**	
C1271III	tumor	MSS	2	0	0	0	**0**		0	0	0	**0**	
C1273II	tumor	MSI-H	2	0	0	0	**0**		0	0	0	**0**	
C1275III	tumor	MSS	2	0	0	0	**0**		1	0	0	**1**	
C1280III	tumor	MSS	2	0	0	0	**0**		0	0	0	**0**	
C1283II	tumor	MSI-H	2	1	0	0	**1**		1	0	1	**2**	
C1284III	tumor	MSS	2	1	1	1	**3**	EX7(V)-EX6(T)	0	0	0	**0**	
C1285III	tumor	MSS	2	0	0	0	**0**		0	0	1	**1**	
C1286III	tumor	MSS	2	0	0	0	**0**		0	0	0	**0**	

Sample	Type	MSI-Status	Patient Series	*VTI1A-TCF7L2* #1	*VTI1A-TCF7L2* #2	*VTI1A-TCF7L2* #3	SUM *VTI1A-TCF7L2* Status	Confirmed exon-exon breakpoint *VTI1A-TCF7L2*	*TCF7L2-RP11-57H14.3* #1	*TCF7L2-RP11-57H14.3* #2	*TCF7L2-RP11-57H14.3* #3	SUM *TCF7L2-RP11-57H14.3*	Confirmed exon-exon breakpoint TCF7L2-RP11-57H14.3
C1291II	tumor	MSS	2	0	0	0	**0**		0	0	0	**0**	
C1292III	tumor	MSS	2	0	0	0	**0**		1	1	0	**2**	
C1294II	tumor	MSS	2	0	0	0	**0**		0	0	0	**0**	
C1296II	tumor	MSS	2	0	0	0	**0**		1	0	0	**1**	
C1301III	tumor	MSS	2	0	0	0	**0**		0	0	0	**0**	
C1314III	tumor	MSI-H	2	0	0	0	**0**		0	1	0	**1**	
C1321II	tumor	MSS	2	0	0	0	**0**		0	0	0	**0**	
C1323III	tumor	MSS	2	1	0	0	**1**		0	0	0	**0**	
C1330III	tumor	MSS	2	0	0	0	**0**		0	0	0	**0**	
C1333III	tumor	MSS	2	1	0	0	**1**		0	0	0	**0**	
C1334III	tumor	MSS	2	0	0	0	**0**		0	0	1	**1**	
C1338III	tumor	MSS	2	0	0	0	**0**		0	0	0	**0**	
C1340III	tumor	MSI-L	2	0	0	0	**0**		0	1	0	**1**	EX4(T)-EX2(RP)-EX3(RP)
C1350II	tumor	MSS	2	0	0	0	**0**		0	0	0	**0**	
C1355III	tumor	MSS	2	0	0	0	**0**		0	0	0	**0**	
C1356III	tumor	MSS	2	0	0	0	**0**		0	0	0	**0**	
C1357II	tumor	MSI-L	2	0	0	1	**1**		0	1	0	**1**	
C1364II	tumor	MSS	2	0	0	0	**0**		0	0	0	**0**	
C1379II	tumor	MSS	2	0	0	1	**1**		0	0	0	**0**	
C1380III	tumor	MSS	2	0	0	0	**0**		0	0	0	**0**	
C1389III	tumor	MSS	2	0	0	0	**0**		1	1	0	**2**	EX4(T)-EX1(RP)-EX3(RP)
C1391II	tumor	MSS	2	0	0	0	**0**		0	0	0	**0**	
C1393III	tumor	MSS	2	1	0	0	**1**		0	0	0	**0**	
C1395II	tumor	MSS	2	0	0	1	**1**		1	1	0	**2**	EX4(T)-EX1(RP)-EX2(RP)-EX3(RP)
C1402III	tumor	MSS	2	0	0	0	**0**		0	0	0	**0**	
C844II	tumor	MSS	2	0	0	0	**0**		0	0	0	**0**	
C861y	tumor	MSS	2	0	0	1	**1**		0	0	0	**0**	
C874III	tumor	MSS	2	1	1	0	**2**	EX3(V)-EX4(T)-EX6(T)	0	0	0	**0**	
C891II	tumor	MSI-L	2	1	0	0	**1**		1	0	1	**2**	
C895II	tumor	MSS	2	1	0	0	**1**		0	0	0	**0**	
C896III	tumor	MSS	2	0	0	0	**0**		0	0	0	**0**	
C903II	tumor	MSS	2	0	0	0	**0**		0	0	0	**0**	
C914III	tumor	MSS	2	0	0	0	**0**		0	0	0	**0**	
C932III	tumor	MSS	2	0	0	0	**0**		0	0	0	**0**	
C935II	tumor	MSS	2	0	0	0	**0**		0	0	0	**0**	

Sample	Type	MSI-Status	Patient Series	*VTI1A-TCF7L2* #1	*VTI1A-TCF7L2* #2	*VTI1A-TCF7L2* #3	SUM *VTI1A-TCF7L2* Status	Confirmed exon-exon breakpoint *VTI1A-TCF7L2*	*TCF7L2-RP11-57H14.3* #1	*TCF7L2-RP11-57H14.3* #2	*TCF7L2-RP11-57H14.3* #3	SUM *TCF7L2-RP11-57H14.3*	Confirmed exon-exon breakpoint TCF7L2-RP11-57H14.3
C937III	tumor	MSS	2	0	0	0	**0**		0	0	0	**0**	
C938III	tumor	MSS	2	0	0	0	**0**		1	0	0	**1**	
C940III	tumor	MSS	2	1	0	0	**1**		1	0	1	**2**	
C950II	tumor	MSS	2	1	0	1	**2**		0	0	0	**0**	
C963III	tumor	MSS	2	1	0	0	**1**		0	0	1	**1**	
C964II	tumor	MSS	2	0	1	0	**1**	EX1(V)-EX6(T)	1	1	1	**3**	
C965II	tumor	MSI-H	2	0	0	0	**0**		0	1	1	**2**	EX4(T)-EX1(RP)-EX2(RP)-EX3(RP)
C970II	tumor	MSS	2	0	0	0	**0**		1	0	1	**2**	
C975III	tumor	MSS	2	0	0	1	**1**		0	0	0	**0**	
C976II	tumor	MSS	2	0	0	0	**0**		0	1	0	**1**	EX4(T)-EX2(RP)-EX3(RP)
C978II	tumor	MSS	2	0	1	0	**1**		0	0	0	**0**	
C980II	tumor	MSI-H	2	0	0	0	**0**		1	0	0	**1**	
C981III	tumor	MSS	2	0	0	0	**0**		0	0	0	**0**	
C982III	tumor	MSI-L	2	0	0	0	**0**		0	1	0	**1**	EX4(T)-EX1(RP)-EX2(RP)-EX3(RP)
C983III	tumor	MSS	2	0	1	1	**2**	EX7(V)-EX4(T)-EX6(T)	1	1	0	**2**	
C985III	tumor	MSS	2	0	1	1	**2**		0	0	1	**1**	
Co115	Cell line			0	0	0	**0**		1	1	1	**3**	
Colo320	Cell line			1	1	0	**2**	EX5(V)-EX6(T)	1	1	1	**3**	
EB	Cell line			0	1	1	**2**	EX7(V)-EX6(T)	0	0	1	**1**	
FRI	Cell line			1	1	0	**2**	EX7(V) EX4(T)-EX6(T)	1	1	0	**2**	EX4(T)-EX5,6,7(V)-EX1,2,3(RP)
HCT116	Cell line			1	0	1	**2**		1	1	1	**3**	Several†
HCT15	Cell line			0	0	0	**0**		1	1	1	**3**	
HT29	Cell line			0	0	0	**0**		0	1	0	**1**	EX4(T)-EX2(RP)-EX3(RP)
IS1	Cell line			1	0	1	**2**		1	1	1	**3**	
IS2	Cell line			1	0	1	**2**		1	1	1	**3**	
IS3	Cell line			0	0	1	**1**		1	1	1	**3**	
LoVo	Cell line			0	0	0	**0**		0	0	0	**0**	
LS1034	Cell line			0	0	0	**0**		1	1	1	**3**	
LS174T	Cell line			0	0	0	**0**		0	0	1	**1**	
NCIH508	Cell line			1	1	1	**3**	EX2(V)-EX6(T)	0	0	0	**0**	
RKO	Cell line			0	0	0	**0**		0	1	1	**2**	
SW48	Cell line			0	0	0	**0**		1	1	1	**3**	
SW480	Cell line			1	0	1	**2**		1	1	0	**2**	
SW620	Cell line			0	1	0	**1**	EX1(V)-EX6(T)	0	1	0	**1**	
TC7	Cell line			0	0	0	**0**		0	1	1	**2**	EX4(T)-EX3(RP)

Sample	Type	MSI-Status	Patient Series	*VTI1A-TCF7L2* #1	*VTI1A-TCF7L2* #2	*VTI1A-TCF7L2* #3	SUM *VTI1A-TCF7L2* Status	Confirmed exon-exon breakpoint *VTI1A-TCF7L2*	*TCF7L2-RP11-57H14.3* #1	*TCF7L2-RP11-57H14.3* #2	*TCF7L2-RP11-57H14.3* #3	SUM *TCF7L2-RP11-57H14.3*	Confirmed exon-exon breakpoint TCF7L2-RP11-57H14.3
TC71	Cell line			0	1	0	**1**	EX3(V)-EX4(T)-EX6(T)	1	1	0	**2**	
V9P	Cell line			0	0	0	**0**		0	0	1	**1**	
Adipose	Normal tissue			0	0	0	**0**		1	0	1	**2**	
Bladder	Normal tissue			0	0	0	**0**		1	1	0	**2**	
Brain	Normal tissue			0	0	1	**1**		1	1	1	**3**	
Cervix	Normal tissue			0	0	0	**0**		1	1	0	**2**	EX4(T)-EX1(RP)-EX2(RP)-EX3(RP)
Colon	Normal tissue			0	0	1	**1**		1	1	1	**3**	
Esophagus	Normal tissue			0	0	0	**0**		0	1	0	**1**	EX4(T)-EX1(RP)-EX2(RP)-EX3(RP)
Heart	Normal tissue			0	0	0	**0**		1	1	0	**2**	EX4(T)-EX1(RP)-EX3(RP)
Kidney	Normal tissue			1	1	0	**2**		1	1	1	**3**	
Liver	Normal tissue			0	0	0	**0**		0	0	0	**0**	
Lung	Normal tissue			1	1	0	**2**		1	0	1	**2**	
Ovary	Normal tissue			0	0	0	**0**		1	1	0	**2**	EX4(T)-EX3(RP)
Placenta	Normal tissue			0	1	0	**1**		1	1	0	**2**	EX4(T)-EX2(RP)-EX3(RP)
Prostate	Normal tissue			0	0	0	**0**		0	0	0	**0**	
Skeletal_muscle	Normal tissue			0	0	0	**0**		0	0	0	**0**	
Spleen	Normal tissue			0	0	0	**0**		0	0	0	**0**	
Stomach	Normal tissue			0	0	0	**0**		0	0	0	**0**	
Testes	Normal tissue			0	0	0	**0**		1	1	1	**3**	EX4(T)-EX3(RP)
Thymus	Normal tissue			0	0	0	**0**		1	0	0	**1**	
Thyroid	Normal tissue			0	0	0	**0**		1	1	1	**3**	EX4(T)-EX1(RP)-EX2(RP)-EX3(RP)
Trachea	Normal tissue			0	0	0	**0**		1	0	0	**1**	

* The nested-PCR product from tumor sample 086 (from series 1) was Sanger sequenced. However, no clear exon-exon breakpoint was found, as the sequence spanned from one codon on exon 7 of *VTI1A* to the middle of exon 6 of *TCF7L2*.

† Based on gel band lengths, all four fusion transcript variants of *TCF7L2-RP11-57H14.3* were expressed in the HCT116 cell line.

Paper II

Novel RNA variants in colorectal cancers

Andreas M. Hoff*, Bjarne Johannessen*, Sharmini Alagaratnam, Sen Zhao, Torfinn Nome, Marthe Løvf, Anne C. Bakken, Merete Hektoen, Anita Sveen, Ragnhild A. Lothe, Rolf I. Skotheim

Manuscript

*Equal contribution

II

Novel RNA variants in colorectal cancers

Andreas M. Hoff*[,1,2,3], Bjarne Johannessen*[,1,2,3], Sharmini Alagaratnam[1,2,3], Sen Zhao[1,2,3], Torfinn Nome[1,2,3], Marthe Løvf[1,2,3], Anne C. Bakken[1,2,3], Merete Hektoen[1,2,3], Anita Sveen[1,2,3], Ragnhild A. Lothe[1,2,3], Rolf I. Skotheim†[,1,2,3]

*These authors contributed equally to this work

†Corresponding author: Rolf.I.Skotheim@rr-research.no

[1]Department of Molecular Oncology, Institute for Cancer Research, Oslo University Hospital-Norwegian Radium Hospital, Oslo, Norway

[2]KG Jebsen Colorectal Cancer Research Centre, Oslo University Hospital, Oslo, Norway

[3]Centre for Cancer Biomedicine, University of Oslo, Oslo, Norway

E-mail addresses: AMH, Andreas.M.Hoff@rr-research.no; BJ, Bjarne.Johannessen@rr-research.no; SA, Sharm@rr-research.no; SZ, Sen.Zhao@rr-research.no; TN, Torfinn.nome@nmbu.no; ML, Marthe.Lovf@rr-research.no; ACB, Anne.Cathrine.Bakken@rr-research.no; MH, Merete.Hektoen@rr-research.no; AS, Anita.Sveen@rr-research.no; RAL, Ragnhild.A.Lothe@rr-research.no; RIS, Rolf.I.Skotheim@rr-research.no;

Abstract

Background: With an annual worldwide incidence close to 1.4 million, and a five-year survival rate of about 60 %, colorectal cancer (CRC) is a major clinical burden. The patient group is heterogeneous, and biomarkers for patient stratification are in high demand. Genome-scale RNA data represents a rich source of genetic information that can be used to study aberrant gene expression and transcript variants, which may be exclusive in particular cancer subpopulations. To identify novel RNA variants in CRC, we analyzed exon-level microarray expression data from 202 CRCs and searched for genes with overexpression of the 3' end in individual tumors. **Results:** We nominated 25 genes in which at least one cancer sample had increased one-sided expression. To efficiently investigate underlying transcript structures, we used a novel approach of rapid amplification of cDNA ends followed by high throughput sequencing (RACE-seq). RACE products from the targeted genes in a total of 23 CRC samples were pooled together and sequenced. We identified *VWA2-TCF7L2*, *DHX35-BPIFA2* and *CASZ1-MASP2* as private fusion events, and novel transcript structures for 17 of the 23 other candidate genes. Additionally, we found junctions supporting a recurrent read-through fusion transcript between *KLK8* and *KLK7*, and a novel 3' splice site in *S100A2*, both of which were overrepresented in CRC tissue and cell lines from external RNA-seq datasets. **Conclusion:** Novel and recurrent cancer-specific RNA variants in CRC are identified using a high-throughput method for characterization of multiple gene transcripts in several samples in the same experimental setup.

Keywords

Colorectal cancer, fusion genes, transcript variants, RACE-seq

Background

Colorectal cancer (CRC) alone accounts for close to 10 % of all cancer cases worldwide, and is a heavy burden on human health and economy. It has been estimated that in total 694,000 people died from CRC in 2012 [1]. Developing through several molecular mechanisms, CRC is a heterogeneous disease with an urgent need for biomarkers carrying diagnostic, prognostic and predictive information. The cancer transcriptome represents a complex collection of RNA molecules which reflect the expression program of cancer cells at a given time. Aberrant gene expression signatures have been successfully identified for subclasses of cancer [2,3] and for prognostication [4]. However, few gene expression signatures have yet been implemented in the clinic, and there is a high demand for additional tools to stratify the heterogeneous patient population.

Alternative pre-mRNA splicing and core promoter usage may create cancer-specific transcripts [5,6]. In addition, fusion transcripts, chimeric RNA joined together from two individual genes as a consequence of chromosomal rearrangements or complex post-transcriptional processes can be highly cancer-specific. Fusion genes which result from chromosomal rearrangements have been shown to be pathognomonic for certain cancer types and are used routinely as diagnostic markers in hematological cancers and childhood sarcomas [7,8]. More recently, with the advent of genome-scale technologies, fusion genes have been identified also in adult epithelial cancers. Although most of them are present only in small subsets of carcinomas, some are found to be highly recurrent, such as rearrangement of *TMPRSS2* and *ETS* transcription factor genes in more than half of all prostate cancers [9]. Fusion transcripts may also be expressed as a result of transcriptional mechanisms like trans-splicing and read-through events for adjacent genes [10]. Proof of non-random expression of such fusion transcripts in normal tissue types with translation into chimeric proteins have also been described [11]. Several reports have shown that such fusion transcripts have an impact on cancer biology, by regulating both replication and cell growth in cancer [12–15]. Chimeric mRNAs expressed in normal cells are sometimes overexpressed in cancer cells. This is the case for *SLC45A3-ELK4,* found to be expressed in both normal prostate tissue and prostate cancer, with high levels of expression in a subset of prostate cancer samples. Only some prostate cancers expressing these fusion transcripts harbor chromosomal rearrangements at the corresponding genomic loci [16,17]. The first recurrent fusion gene identified in CRC, *VTI1A-TCF7L2,* was originally detected in three of 97 CRCs, and found to be caused by a genomic deletion in the NCI-H508 CRC cell line [18]. However, when probed for with sensitive PCR, expression of *VTI1A-TCF7L2* fusion transcript was seen in a higher frequency of both normal and malignant tissue, probably as a result of read-through splicing [19]. The presence of splicing-generated fusion transcripts in normal cells and corresponding chromosome rearrangements followed by overexpression in cancer has been proposed to be a linked mechanism [20]. Intragenic deviating expression patterns can be caused by different promoter strengths of two fusion partner genes, usage of alternative core promoters or differential splicing. Exon-level microarrays, with probe sets in each annotated exon, as well as RNA sequencing technologies, enable investigation of complicated structural transcription events in cancer.

In this study, we have used exon-level expression data from a series of CRC as a screening tool to identify genes with differential internal expression, which can be indicative of their involvement as partner genes in fusion transcripts or being transcribed from different promoters. The transcript structures of nominated candidate genes were investigated by a combination of traditional rapid amplification of cDNA ends (RACE) and high-throughput RNA sequencing. This combination of methods facilitated the identification of fusion transcripts and transcript variants overrepresented in CRC.

Results

From exon-level genome-wide microarray data of 202 CRCs, we selected 25 genes with differential expression levels between the 5’ and 3’ parts in at least one cancer sample (Table 1). These abnormal expression profiles typically reflect that the gene is transcribed by an alternative and stronger promoter, either within the gene itself, or from a separate gene. Twenty-four of the top-scoring genes were targeted by more than one probe set at both sides of the respective breakpoints, while one candidate (*FABP7*) obtained a very high expression break score (EBS) and was included even if it was targeted only by a single probe set 5’ to the putative breakpoint. Exon-level expression profiles for all candidate genes can be found in supplementary figures S1A-S1Y (Additional file 1). One candidate gene (*S100A2*), has previously been nominated from RNA sequencing data as downstream partner of three fusion transcripts in the CRC cell line RKO, with *ZNF833*, *RP1-28O10.1*, and *AMPD3* as 5' fusion partners [21]. An overview of the pipeline used to identify and characterize novel RNA variants in CRC is provided in Figure 1.

Table 1: Top 25 genes with elevated 3’ expression levels in individual CRCs.

The top 25 genes with elevated 3’ expression, identified by analysis of exon microarray data, were selected for follow-up transcript characterization by 5’ RACE and sequencing. In addition to the 25 top-scoring genes, three genes with previously known transcriptional changes were included as positive controls.

Gene symbol	Ensembl ID	Chromosome	Strand	Deviating sample	Expression Break score (EBS)	Type
ACY3	ENSG00000132744	11	-1	Sample12_T	2.60	5' RACE candidate
ASPRV1	ENSG00000244617	2	-1	Sample5_T	4.00	5' RACE candidate
BAAT	ENSG00000136881	9	-1	Sample13_T	3.67	5' RACE candidate
BPIFA2	ENSG00000131050	20	1	Sample16_T	2.65	5' RACE candidate
CA6	ENSG00000131686	1	1	Sample9_T	4.16	5' RACE candidate
COLGALT2	ENSG00000198756	1	-1	Sample19_T	4.67	5' RACE candidate
FABP7	ENSG00000164434	6	1	Sample10_T	5.67	5' RACE candidate
FGF12	ENSG00000114279	3	-1	Sample12_T	3.37	5' RACE candidate
GUCY1A2	ENSG00000152402	11	-1	Sample1_T	3.68	5' RACE candidate
HOXC12	ENSG00000123407	12	1	Sample15_T	4.61	5' RACE candidate
IL11	ENSG00000095752	19	-1	Sample20_T	3.47	5' RACE candidate
INA	ENSG00000148798	10	1	Sample14_T	5.73	5' RACE candidate
KLK7	ENSG00000169035	19	-1	Sample2_T	3.91	5' RACE candidate
KRT24	ENSG00000167916	17	-1	Sample5_T	3.45	5' RACE candidate
LY6D	ENSG00000167656	8	-1	Sample5_T	3.15	5' RACE candidate

					Expression	
					Break score	
Gene symbol	**Ensembl ID**	**Chromosome**	**Strand**	**Deviating sample**	**(EBS)**	**Type**
LYPD3	ENSG00000124466	19	-1	Sample5_T	5.06	5’ RACE candidate
MASP2	ENSG00000009724	1	-1	Sample4_T	3.25	5' RACE candidate
MOGAT1	ENSG00000124003	2	1	Sample6_T	2.29	5' RACE candidate
MUC15	ENSG00000169550	11	-1	Sample5_T	3.81	5' RACE candidate
NINJ2	ENSG00000171840	12	-1	Sample8_T	3.40	5' RACE candidate
SOHLH2	ENSG00000120669	13	-1	Sample3_T	2.75	5' RACE candidate
S100A2	ENSG00000196754	1	-1	Sample5_T	5.85	5' RACE candidate
SLC22A2	ENSG00000112499	6	-1	Sample17_T	4.33	5' RACE candidate
SLC38A11	ENSG00000169507	2	-1	Sample19_T	4.54	5' RACE candidate
SNAP25	ENSG00000132639	20	1	Sample7_T	3.08	5' RACE candidate
RP11-57H14.3	ENSG00000225292	10	1	HCT116	NA	Positive control
TCF7L2	ENSG00000148737	10	1	NCI-H508	NA	Positive control
VNN1	ENSG00000112299	6	-1	HT29	NA	Positive control

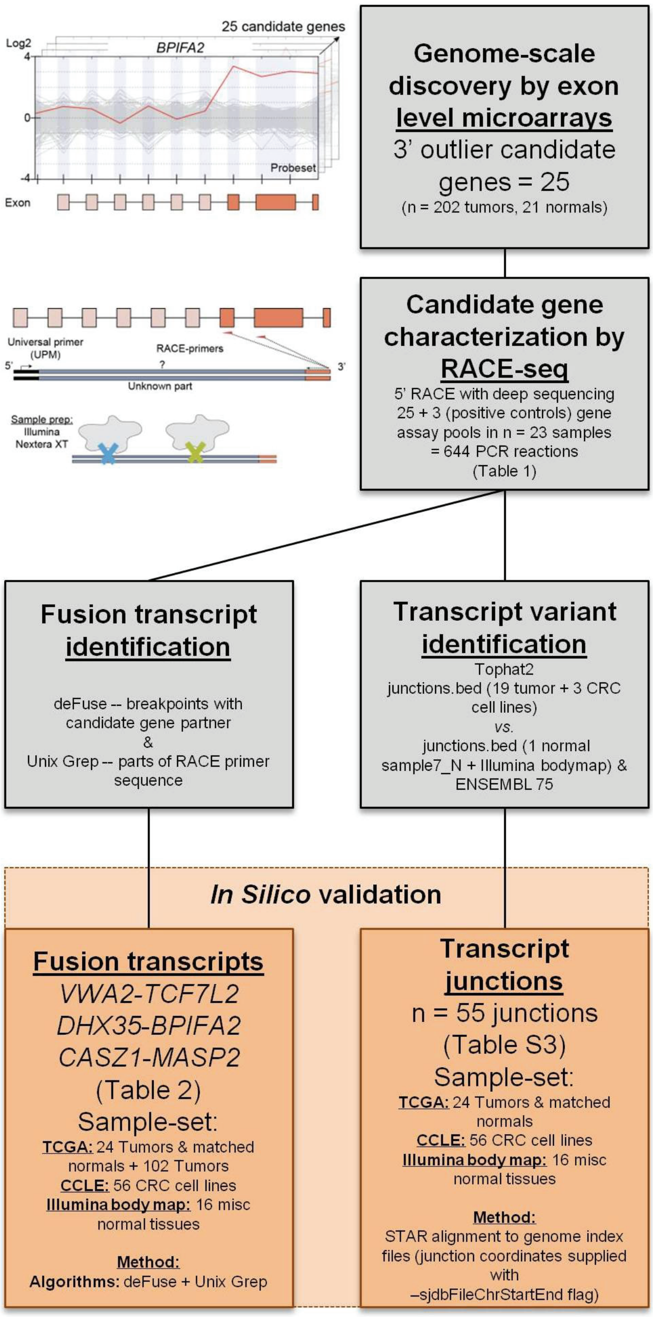

25 candidate genes
Log2
BPIFA2
4
0
-4
Probeset
Exon
Genome-scale discovery by exon level microarrays
3' outlier candidate genes = 25
(n = 202 tumors, 21 normals)
Universal primer (UPM)
RACE-primers
5'
?
3'
Unknown part
Sample prep: Illumina Nextera XT
Candidate gene characterization by RACE-seq
5' RACE with deep sequencing
25 + 3 (positive controls) gene assay pools in n = 23 samples
= 644 PCR reactions
(Table 1)
Fusion transcript identification
deFuse -- breakpoints with candidate gene partner
&
Unix Grep -- parts of RACE primer sequence
Transcript variant identification
Tophat2
junctions.bed (19 tumor + 3 CRC cell lines)
vs.
junctions.bed (1 normal sample7_N + Illumina bodymap) & ENSEMBL 75
In Silico validation
Fusion transcripts
VWA2-TCF7L2
DHX35-BPIFA2
CASZ1-MASP2
(Table 2)
Sample-set:
TCGA: 24 Tumors & matched normals + 102 Tumors
CCLE: 56 CRC cell lines
Illumina body map: 16 misc normal tissues
Method:
Algorithms: deFuse + Unix Grep
Transcript junctions
n = 55 junctions
(Table S3)
Sample-set:
TCGA: 24 Tumors & matched normals
CCLE: 56 CRC cell lines
Illumina body map: 16 misc normal tissues
Method:
STAR alignment to genome index files (junction coordinates supplied with –sjdbFileChrStartEnd flag)

Figure 1: Pipeline to identify and characterize novel RNA variants in CRC

Analysis of genome-scale exon level microarray data revealed 25 candidate genes with overexpression of their 3' end, here exemplified with the *BPIFA2* gene. The candidate genes were characterized with RACE-seq, a combination of 5' RACE and deep sequencing. For the 25 candidate genes and also 3 positive control genes, nested RACE-primers were designed downstream of the suspected breakpoints (orange arrows). The resulting pools of RACE fragments (28 assays per sample) were prepared for sequencing with the Nextera XT protocol (Illumina), using tagmentation to simultaneously fragment and tag RACE-amplicons with adapters for sequencing. The fusion transcripts *VWA2-TCF7L2, DHX35-BPIFA2* and *CASZ1-MASP2,* and also 55 transcript junctions were identified by two separate computational approaches from the RACE-seq data. These were probed for in an external data series from the TCGA, CCLE and the Illumina body map.

Novel fusion transcripts in CRC

The 25 nominated candidate genes and three positive control genes were tested by 5' RACE in all samples with highest EBS (n = 23; *i.e.* a total of 644 RACE reactions). The RACE-products from all gene assays for each sample were then pooled and sequenced, generating 15.5 million pairs of sequencing reads. After demultiplexing, trimming away 5' RACE adapter sequences, and quality assurance, 12.6 million pairs remained (Table S1 in additional file 2). Trimming 5' RACE adapter sequences increased the number of paired-end reads aligned by up to 79 % for each individual sample (Table S1 in additional file 2).

As positive controls for fusion gene detection by RACE-seq, we included *TCF7L2* and *RP11-57H14.3* as 5' RACE targets in the NCI-H508 and HCT-116 cell lines, respectively. Both previously described fusions with these two genes as downstream partners were among the top nominated fusion breakpoints [18,19] (Table 2). In addition to the two positive control fusions involving *TCF7L2,* another *TCF7L2* fusion with *VWA2* as a novel upstream fusion partner was identified in one tumor sample (Figure 2; Table 2). This sample also had a high number of reads covering the first four exons of *VWA2*. Several other genes, not directly targeted by 5' RACE, showed similarly high coverage in individual tumor samples, indicating that they are indirectly amplified by the 5' RACE assays (Figure S2 in additional file 3). By evaluating sequence alignments for these genes, we identified two genes, *CASZ1* and *DHX35,* with high sequence coverage of the first two and 11 exons in individual samples, respectively. *DHX35* was identified as an upstream fusion partner of *BPIFA2* in the same sample that had high sequence coverage of the first 11 exons of *DHX35* (Table 2; Figure 3). This sample was selected due to an elevated 3' expression profile of *BPIFA2,* which matched the identified fusion breakpoint. Similarly, the sample with elevated 3' expression of the *MASP2* gene had high sample-specific coverage of the two first exons of *CASZ1* in the RACE-seq data (Figure 4). No fusion candidate with *CASZ1* as an upstream partner gene was nominated by the fusion detection software. However, the RACE assay for *MASP2* was designed to be in close proximity to the expression breakpoint observed in the exon microarray data. Thus, we used the Unix command-line utility grep to search for parts of the nested gene-specific primer (NGSP) sequence targeting *MASP2*. Using grep, several reads containing sequence from the *MASP2* NGSP mapping to both exon 9 of this gene and exon 2 of *CASZ1* were retrieved, indicating a fusion between these genes. After realigning raw reads to a fusion scaffold, we identified 107 split reads crossing the junction between *CASZ1* exon 2 (ENST00000344008) and *MASP2* exon 9 (ENST00000400897).

Fusion transcript validation

The three novel fusion transcripts were successfully validated using RT-PCR assays spanning the respective fusion boundaries followed by Sanger sequencing (Figure S3 in additional file 4). All three fusion transcripts were generated by using intact splicing sites from their respective partner genes. The reading frames of the parental genes were retained for *VWA2-TCF7L2* and *DHX35-BPIFA2*. Both fusion transcripts potentially encode fusion proteins, as all four fusion partners have breakpoints within their coding sequences. The 5' UTR of *CASZ1* is joined together with the 3' part of the coding sequence of *MASP2*. Here, a start codon located 119 bp downstream of the fusion breakpoint encodes an open reading frame (ORF) that is in-frame with the reference *MASP2* ORF. As a consequence, the *CASZ1-MASP2* fusion transcript may encode an N-terminal truncated MASP2 protein under the control of the *CASZ1* promoter and 5' UTR. None of the fusion transcripts were detected in external data sets from The Cancer Genome Atlas (TCGA) and the Cancer Cell Line Encyclopedia (CCLE), indicating that the fusions are private events.

Table 2: Fusion transcripts detected by the RACE-seq approach.

Both previously known as well as novel fusion transcripts were detected by the RACE-seq approach and subsequent analysis of paired-end sequencing reads. Only the top-scoring nominated candidates are shown for each fusion pair. The known as well as the novel fusion transcripts *VWA2-TCF7L2* and *DHX35-BPIFA2* were nominated by the fusion detection software, whereas the novel *CASZ1-MASP2* was detected by the Unix utility grep and scaffold realignment.

Gene A	**Gene B**	**Break position A**	**Break position B**	**Split reads**[*]	**Spanning reads**[*]	**In Frame**	**Probability score**[†]	**Validated**[§]	**Sample**
VTI1A	*TCF7L2*	Chr10: 114,220,341	Chr10: 114,900,943	1234	163	Yes	0.96	Previously reported	NCI-H508
TCF7L2	*RP11-57H14.3*	Chr10: 114,799,885	Chr10: 114,648,494	114	41	No	0.95	Previously reported	HCT-116
VWA2	*TCF7L2*	Chr10: 116,014,807	Chr10: 114,900,943	206	67	Yes	0.98	Yes	Sample17_T
DHX35	*BPIFA2*	Chr20: 37,632,550	Chr20: 31,767,410	862	4	Yes	0.23	Yes	Sample16_T
CASZ1	*MASP2*	Chr1: 10,820,757	Chr1: 11,090,938	107	NA	Yes	NA	Yes	Sample4_T

*) The number of split reads and spanning reads supporting the fusion junction, determined from the fusion detection software or scaffold realignment for the *CASZ1-MASP2* fusion (split reads contain the fusion boundary in the read itself, whereas spanning reads are paired ends that harbor the fusion boundary within the insert sequence). †) Probability score determined by the fusion detection software. §) Two fusion genes involving *TCF7L2* have previously been reported and served as positive control fusion targets for the RACE-seq. The novel fusions were validated by Sanger sequencing.

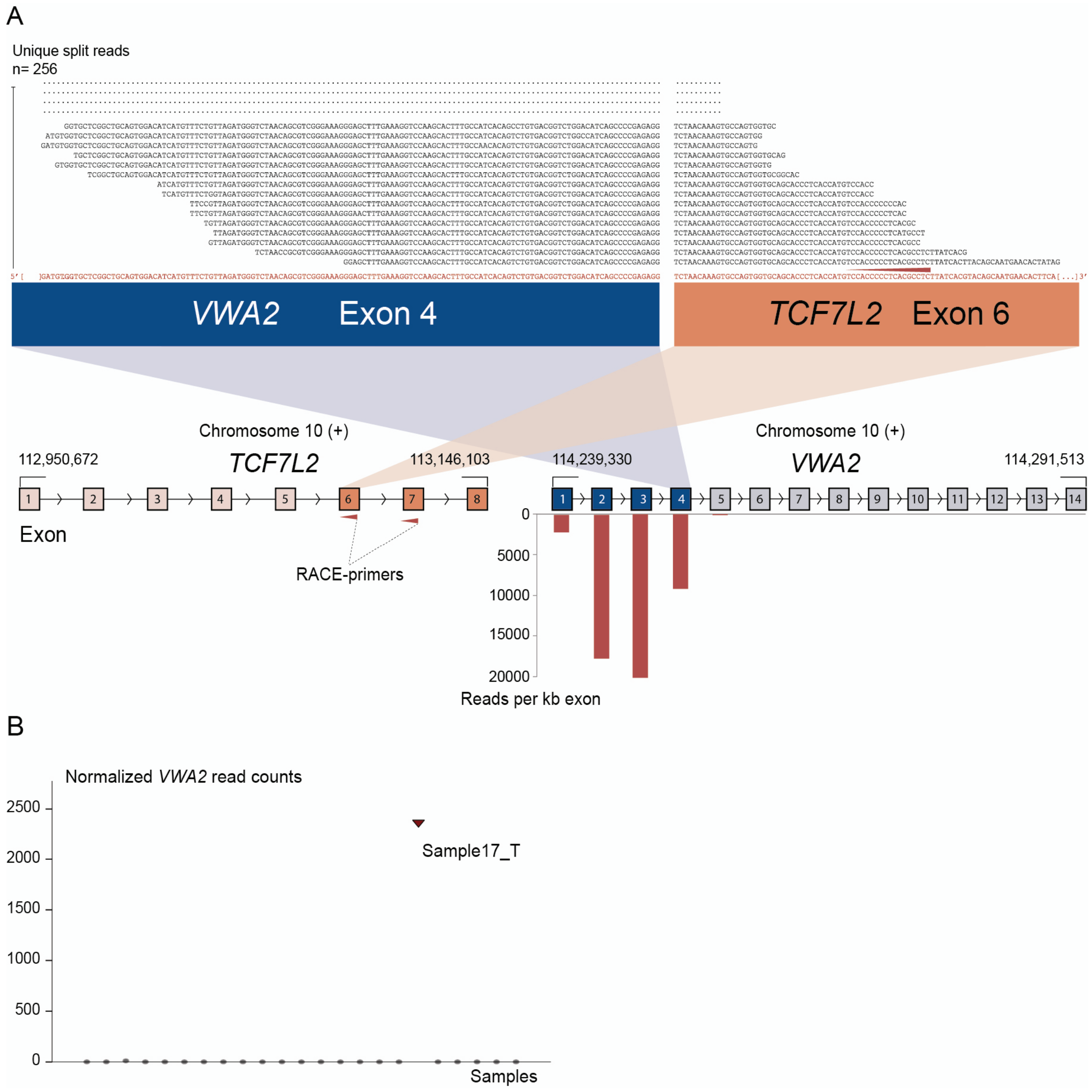


Figure 2 The novel fusion transcript *VWA2-TCF7L2*.

A. As indicated by red arrows, a RACE assay with first and second round primers targeting exon 7 and 6 of *TCF7L2* respectively was included in the sequencing setup as a positive control. In addition to enabling identification of previously known fusions involving *TCF7L2*, a new fusion was discovered that connects exon 4 of *VWA2* to exon 6 of *TCF7L2.* The exon numbers refer to transcript structures ENST00000392982 for *VWA2* and ENST00000369395 for *TCF7L2*. The 5' partner gene, *VWA2,* is located 1.1 Mbp downstream of *TCF7L2*. The four exons upstream of the fusion breakpoint in *VWA2* show high read coverage, measured in reads per kilobase exon sequence (RPK). On the contrary, the exons after the breakpoint in *VWA2* were not picked up by the assay, indicating that the RACE assay targeting *TCF7L2* specifically amplified the upstream *VWA2* part in the tumor sample harboring the *VWA2-TCF7L2* fusion. **B.** Normalized read counts mapping to the *VWA2* gene for all samples included in the RACE-seq experimental set up. Only the sample with the identified *VWA2-TCF7L2* fusion transcript had sequencing reads covering the *VWA2* genes.

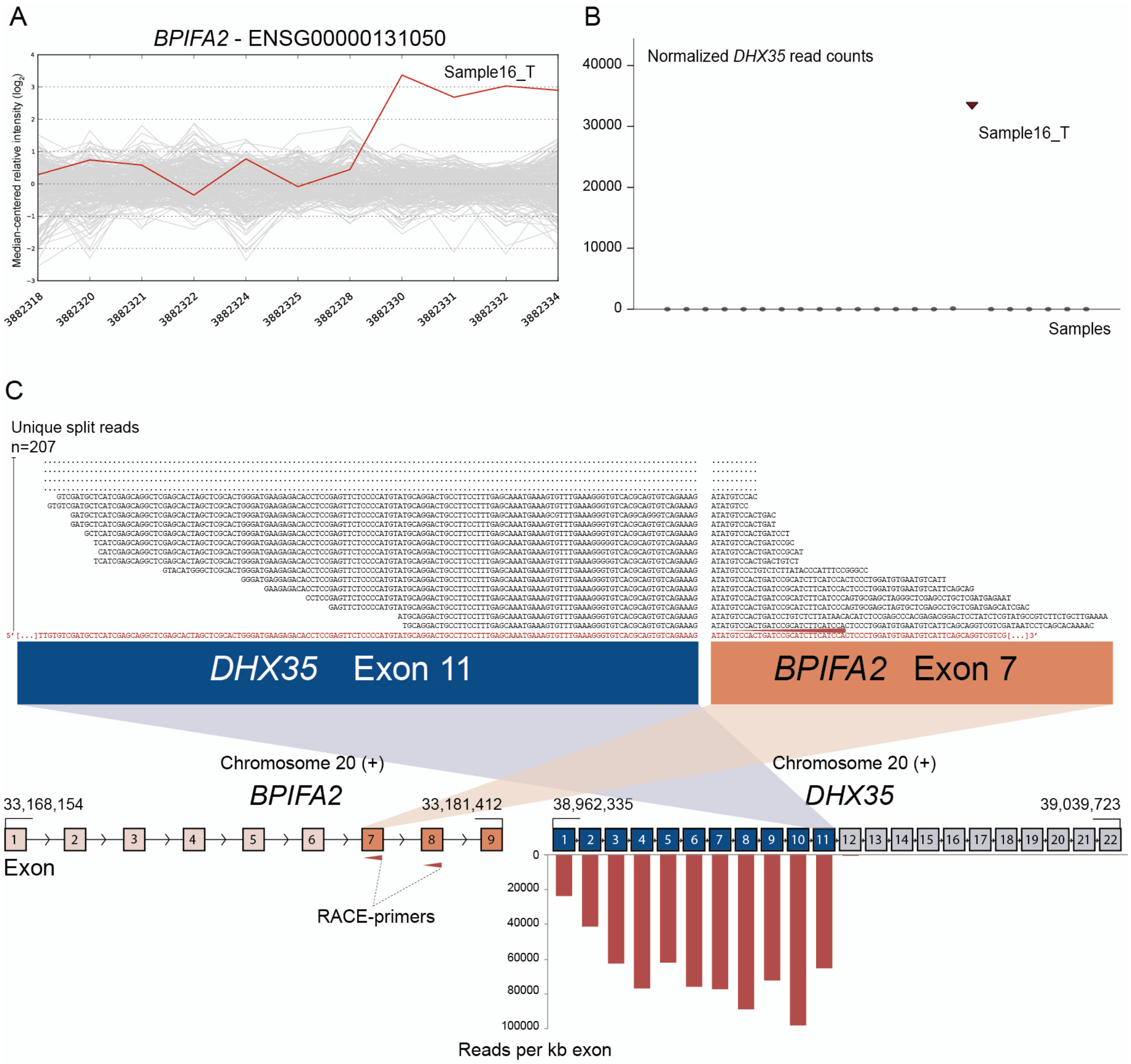


Figure 3: Novel fusion between *DHX35* and *BPIFA2* is in concordance with 3’ overexpression of *BPIFA2.*

A. The exon expression profile of *BPIFA2* shows that one CRC sample has a 3-fold increase in expression levels in the 3’ part of the gene compared to the median of the cohort. The last four probe sets showed increased intensity levels and target exon 7, 8 and 9 of *BPIFA2.* B. Normalized read counts mapping to *DHX35* were high only in sample16_T, the same sample which exhibits increased 3’ expression of *BPIFA2*. C. From RACE-seq we identified a fusion between exon 11 of *DHX35* and exon 7 of *BPIFA2*. Exons 1-11 of *DHX35* show high read coverage, measured in RPK. Exons 12-22, located downstream of the breakpoint in *DHX35*, were not covered by any reads, indicating that the RACE assay that targets *BPIFA2* specifically amplifies the upstream *DHX35* part of the fusion. By realigning sequencing reads from the sample in question to a fusion scaffold, we found that 207 unique split reads align across the fusion boundary.

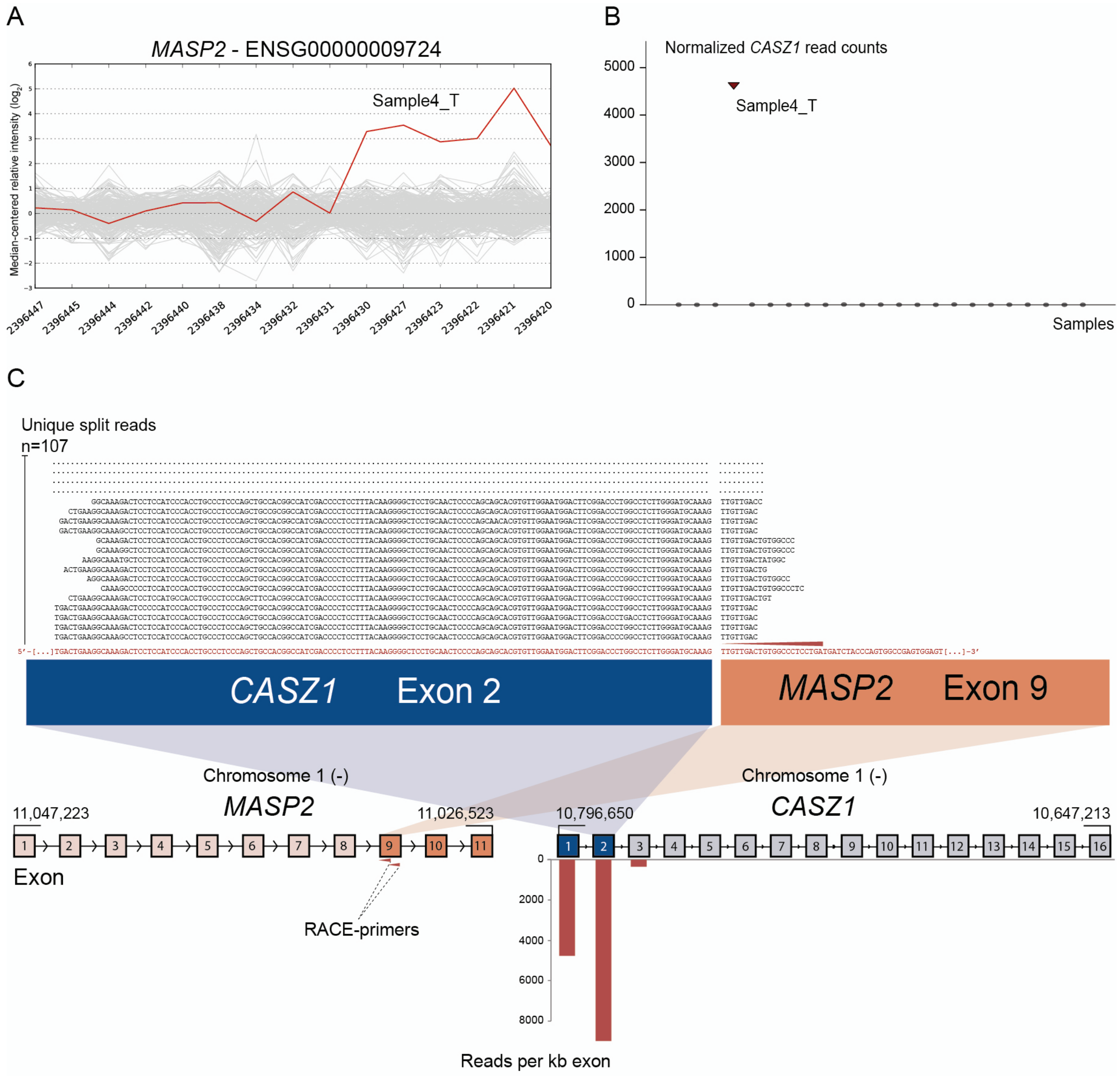


Figure 4 Novel fusion between *CASZ1* and *MASP2* is in concordance with 3’ overexpression of *MASP2.*

A. The exon expression profile of *MASP2* shows that one CRC sample has a 3- to 5-fold increase in expression of the 3’ part of the gene compared to the median of the cohort. The last six probe sets showed increased intensity and targeted exon 9, 10 and 11 of *MASP2.* **B.** Normalized read counts mapping to *CASZ1* were high only in sample 4_T, the same sample which exhibit increased 3’ expression of *MASP2*. **C.** From RACE-seq we identified that the gene *CASZ1* had high read coverage of its first two exons in the same sample that exhibited the deviating exon expression profile for *MASP2.* No fusions involving *MASP2* were nominated by deFuse, but when using the grep command for parts of the *MASP2* NGSP sequence (Red arrow), we identified several reads containing a fusion boundary between *CASZ1* exon 2 and *MASP2* exon 9. Upon realigning sequencing reads, we found that 107 unique reads split aligns across the fusion boundary.

Novel transcript splice junctions in CRC

From the RACE-seq data we identified 147 novel splice junctions that had read coverage greater than 100 in at least one sample each (Table S2 in additional file 2). From this list and by visual inspection of splicing junctions in the Integrated Genome Viewer (IGV), 55 junctions were selected that could potentially explain the deviating gene expression profiles in 20 of the genes (Table S3 in additional file 2). In total, the RACE-seq data supported new transcript structures for 17 of the 23 targeted genes, not including the genes with validated fusion transcripts. Thirteen of the genes had sequence reads extending upstream from the annotated gene boundary, whereas 12 had sequence reads supporting new intragenic transcript structures including internal promoters and eight of the 23 genes had both.

We included the gene *VNN1* as a positive control for junction detection. We recently reported a novel transcript variant of this gene as being expressed in a large proportion of CRCs, including the HT29 cell line [22]. In addition to expression in HT29 , 15 of the 20 primary carcinoma samples included in the RACE-seq expressed this transcript variant (Figure S4 in additional file 5; Table S4 in additional file 2), confirming previous results [22].

New transcript variants of *S100A2* and *KLK7* are overrepresented in CRC

RACE-seq data from the candidate genes *S100A2*, *KLK7*, *FGF12* and *BAAT* had reads supporting junctions of read-through fusion transcripts from upstream adjacent genes *S100A16*, *KLK8*, *MB21D2* and *MRPL50* respectively. The read-through *KLK8-KLK7* joins two members of the Kallikrein-related peptidase gene family that are 12 kb apart on chromosome 19. The exon microarray profile for *KLK7* showed elevated expression levels at the 3’ end of the gene in several samples (Figure 5). The two samples with the highest EBS score were selected for RACE-seq. RACE-seq reads from both samples aligned to a junction between exon 2 of *KLK8* (ENST00000291726) and exon 3 of *KLK7* (ENST00000595820). The resulting read-through fusion transcript has an in-frame ORF, potentially encoding a fusion protein with coding sequences from both *KLK8* and *KLK7*. In total, eight of 19 CRC tumor samples and two of three CRC cell lines expressed this junction (Table 3 and Table S4 in additional file 2). In the external datasets, six of 24 CRCs from the TCGA had at least one sequence read covering the *KLK8-KLK7* junction and 15 of 56 CRC cell lines had multiple sequence reads covering the junction. In contrast, the junction was not detected by sequence reads from any of the matching normal colonic mucosa samples, although a single read covered the junction in normal breast tissue in the Illumina human body map (Table S4 in additional file 2).

S100A2 was included based on elevated expression levels at the 3’ end for two samples (Figure 6). Our approach to identify new junctions revealed alternative splicing between exon 1 and 2, with a new 3’ splice site only six bp downstream of the canonical splice site (ENST00000368708). Although this junction alone could not explain the deviating exon expression profiles, it was found to be overrepresented in CRC. In total, the splice site was covered by sequencing reads in 13 of 19 CRCs and in all three CRC cell lines (Table 3). Furthermore, 14 of 24 (58 %) of the TCGA tumor samples had at least one read covering the junction, while 33 of 56 (59 %) CRC cell lines had multiple reads covering the splice junction. Importantly, none of the matched normal samples from the TCGA had any reads covering the junction. In the Illumina human body map dataset, only one sample from normal lung tissue was found to have 12 reads covering the same splice junction.

Table 3 Validation frequencies of *KLK8-KLK7* and *S100A2* junctions

The novel junctions covering the read-through of *KLK8* and *KLK7* and the alternative 3’ splice site of *S100A2* were discovered from the RACE-seq data. The junctions were validated by aligning all RACE-seq sequences and external data sets from the TCGA and the CCLE.

Gene	Chromosome	Pos 1	Pos 2	Distance	RACE-Seq*	TCGA_tumor*	TCGA_normal*	CCLE*	Bodymap
KLK8-KLK7	19	51485170	51504353	19183	10/22 (45 %)	6/24 (25 %)	0/24	15/56 (27 %)	1/16†
S100A2	1	153536357	153537981	1624	16/22 (73 %)	14/24 (58 %)	0/24	33/56 (59 %)	1/16†

*) The number of reads for the junctions to be considered positive were >10, >=2 and >= 1 for the RACE-seq, CCLE and TCGA tumor/normal data sets, respectively. †) From the Illumina Human body map data set, one read covering the *KLK8-KLK7* read-through was identified in normal breast tissue, and 12 reads were found to cover the *S100A2* alternative splice site in normal lung tissue.

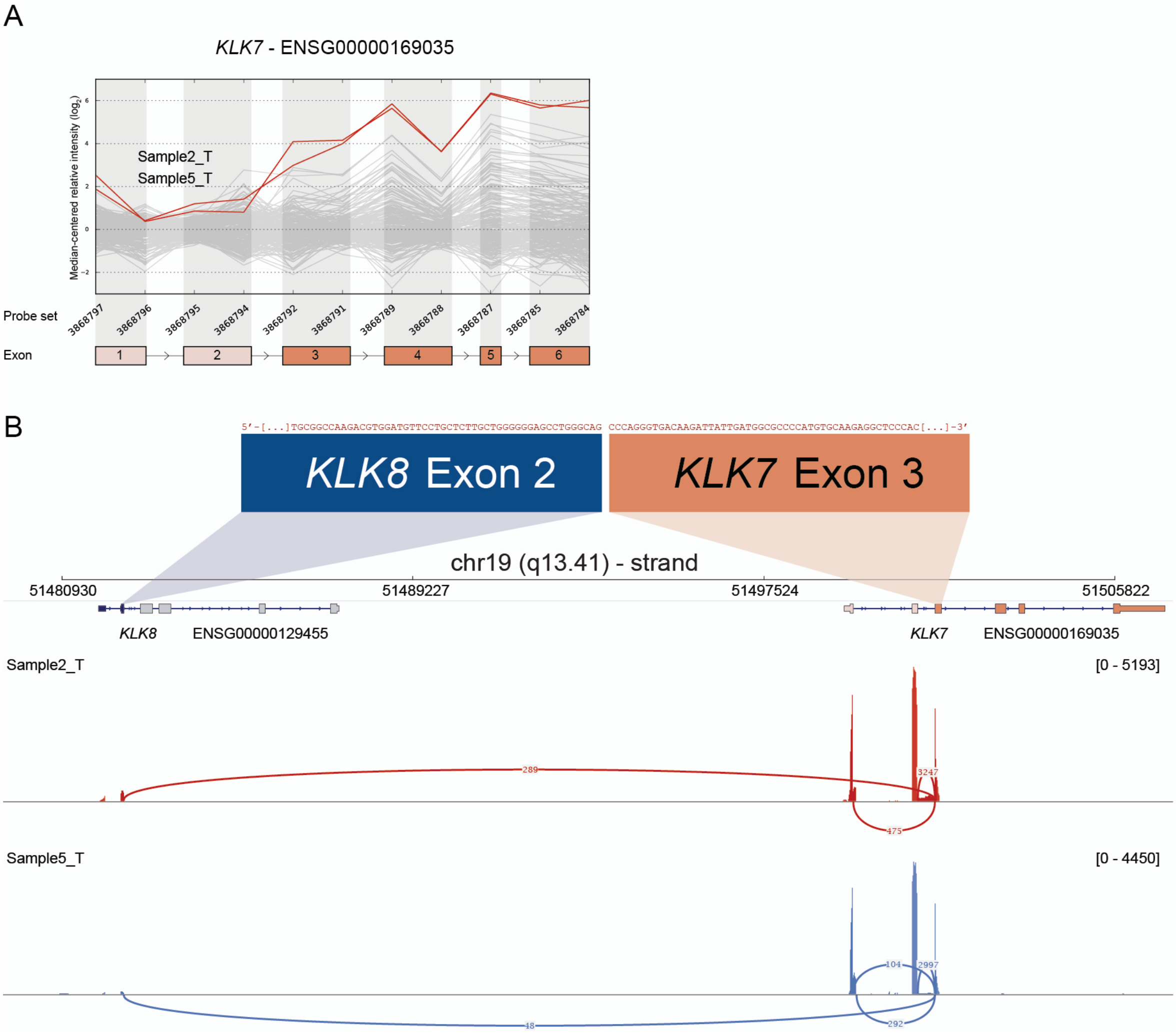


Figure 5: Read-through from upstream *KLK8* to *KLK7* in samples with deviating 3' expression of *KLK7*. **A.** The exon expression profile of *KLK7* shows that several samples have increased expression of probe sets targeting exon 3 to 6. The top two samples (2_T & 5_T) were selected for RACE-seq to identify underlying transcript structural changes. **B.** A sashimi plot from IGV shows the alignment of sequencing reads from the two nominated samples for *KLK7*. The height of the bars represents the number of aligned reads, while arcs represent junctions connected to exon 3 of *KLK7* and the coverage of these junctions, as determined by Tophat2 alignment and the sashimi plot package, are shown as numbers on the arcs.

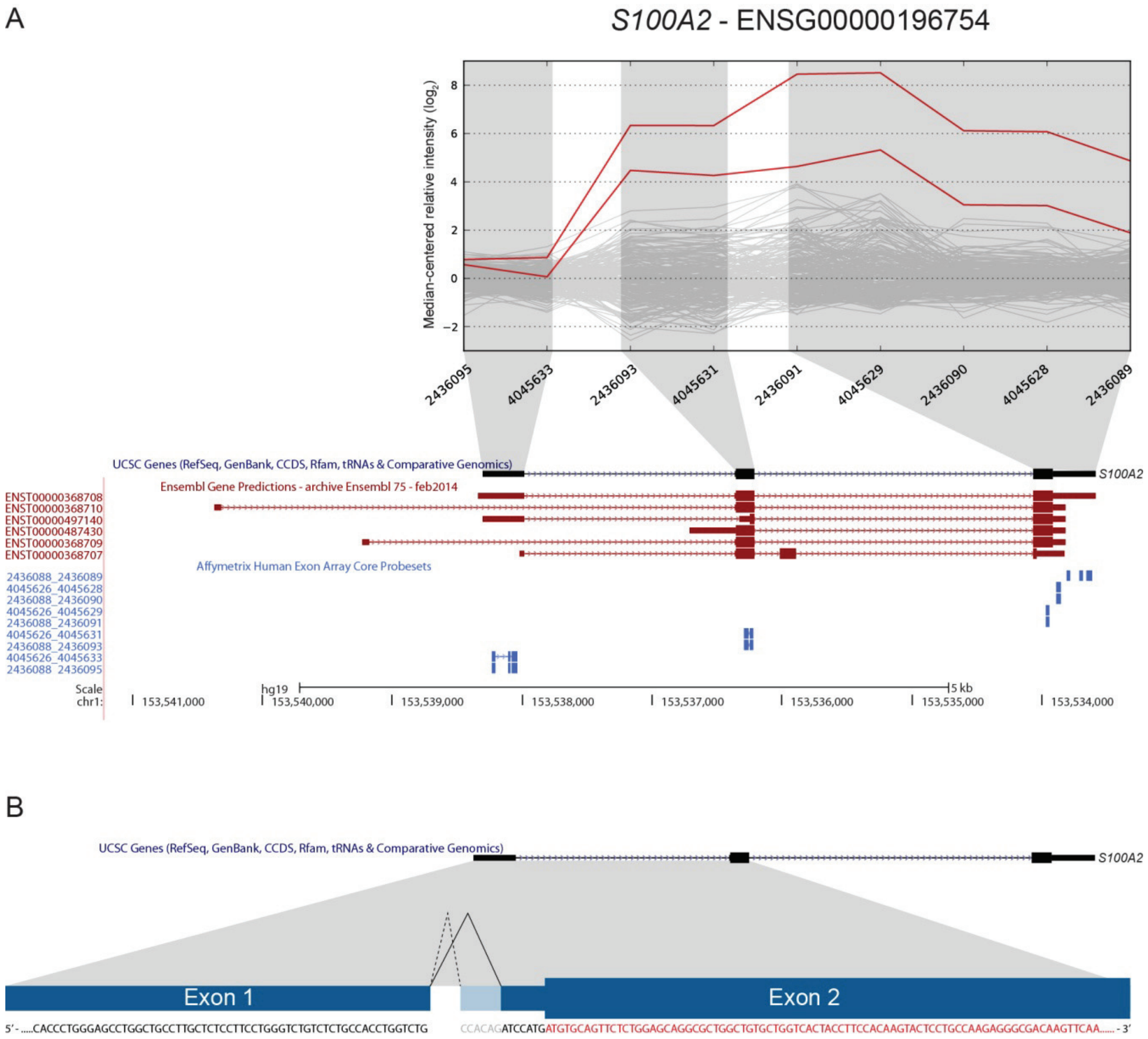


Figure 6: A novel 3' splice-site in *S100A2* is found to be overrepresented in CRC.

A. Two samples had increased expression levels of probe sets targeting exons 2 and 3 of *S100A2* (UCSC annotation). Ensembl release 75 has annotated additional transcript variants (in red; ENST00000368710, ENST00000368709) and upstream 5' UTR exons that are not targeted by exon microarray probe sets. **B.** We identified the use of a novel 3' splice-site when splicing exon 1 to exon 2. The splice-site is located 6 bp downstream of the canonical splice-site. Use of this alternative splice-site was found to be overrepresented in CRC samples (see Table 3). The splice-site occurs in the 5'UTR (coding sequence in red), and according to our analysis does not affect the annotated ORF.

Discussion

We have identified a set of novel transcript variants in CRC by a novel RACE-seq approach following an exon-level transcriptomics screen enabling simultaneous detection of variants from multiple genes and samples. Proof-of-concept was demonstrated by detection of known fusion transcripts involving *TCF7L2*, and alternative promoter usage in the *VNN1* gene.

Among the three novel private fusions transcripts, *VWA2* was detected as a new and previously unknown fusion partner of *TCF7L2*, strengthening the hypothesis that this WNT-effector transcription factor is involved in CRC. Interestingly, the *VWA2-TCF7L2* fusion transcript shares the same breakpoint of *TCF7L2* as the known *VTI1A-TCF7L2* fusion. Knock-down of *VTI1A-TCF7L2* in the NCI-H508 CRC cell line was previously shown to inhibit anchorage-independent growth [18]. The *DHX35* gene has previously been identified as a 5' fusion partner to the *ITCH* gene in the SK-BR-3 breast cancer cell line [23], indicating that it may be a 5' partner in multiple fusion genes. All three fusions were predicted to encode intact ORFs, implying a potential functional role.

In a similar approach to ours, the Encyclopedia of DNA elements (ENCODE) performed 5' RACE combined with high-density resolution tiling microarrays to annotate transcript products from 399 known protein-coding loci. The results revealed that >80 % of the tested genes had unannotated transcribed fragments both upstream and internal to previously known gene boundaries [24]. The ENCODE project has suggested that up to three-quarters of the human genome is capable of being transcribed, and that the current concept of a gene is in need of refinement [25]. Our data are in line with this, inasmuch as 17 of 23 candidate genes have reads supporting new transcript structures, and 13 have reads extending upstream of the currently annotated gene boundaries.

Among the identified transcript variants, we found a recurrent read-through fusion transcript between *KLK8* and *KLK7* and a novel 3' splice site in *S100A2*. These were both found to be overrepresented in CRC tumor samples and cell lines when investigated in external RNA-seq datasets from TCGA, CCLE and the Illumina Human body map. *KLK7* encodes a serine protease of the kallikrein-related peptidases, and has previously been shown to be overexpressed in CRC. *KLK7* overexpression has also been found to be associated with increased cell proliferation *in vitro* and increased tumor growth in nude mice [26]. Furthermore, overexpression of *KLK7* mRNA and protein product have been related to poor clinical prognosis of CRC patients [27,28], and the use of alternative promoters has been indicated in tissue or disease-specific *KLK* regulation [29]. The *KLK8-KLK7* read-through, with an intact ORF and overrepresentation in CRC, has biomarker potential and may have functional consequences for the disease which warrant further studies. The observed increase in expression levels of 3' exons of *S100A2* is likely explained by the additional transcript variants annotated by Ensembl (Figure 6). However, the identification of a new 3' splice site, detected in more than 50 % of both primary CRC tumors and CRC cell lines but none of the matched normal samples makes it a potential cancer-specific splice variant. *S100A2* is a member of the S100 gene family, where S100 proteins are Ca^{2+} binding proteins with broad implications in cancer [30]. Moreover, *S100A2* expression has been proposed as a prognostic biomarker for tumor recurrence in CRC patients treated with adjuvant chemotherapy after surgery [31]. Although theoretically not affecting the ORF and the subsequent translated protein, the alternative 3' splice usage in *S100A2* identified here may have other regulatory or functional roles in CRC.

In terms of methodology, the idea of combining 5' RACE with transcript variant characterization using paired-end sequencing shows potential for the identification of tissue- or disease-specific transcriptional structural changes. High-throughput sequencing characterization of RACE amplicons is highly time-efficient, more sensitive and technically feasible compared to traditional characterization of RACE fragments with cloning, plasmid isolation and Sanger sequencing. In our setup, we computationally nominated candidate genes with differential 3' expression levels in CRC samples from exon microarray data to identify novel cancer-specific markers. However, the combination of 5' RACE with paired-end sequencing also has other potential applications, such as testing for known fusion genes across tumor types or within different clinical cohorts. A future improvement would be to design RACE assays sufficiently downstream of suspected breakpoints to ensure sequencing reads on both ends, giving traditional fusion detection algorithms enough power to detect all fusions.

Conclusions

Using a novel high-throughput approach, we identified three previously unknown private fusion transcripts, and new transcript structures in 17 of the 23 other candidate genes. The novel transcripts included a recurrent read-through transcript between *KLK8* and *KLK7* and use of an alternative 3’ splice site in *S100A2* that are overrepresented in CRC. Both *KLK7* and *S100A2* have previously been implicated in CRC, making these transcript variants interesting as candidate markers in CRC.

Methods

Patient samples and cell lines

A consecutive series of 202 primary colorectal carcinomas (stage I-IV) and normal colonic mucosa samples from 21 of the patients were included in the analysis. The CRC series was collected between 2005 and 2009 at Aker University Hospital, Oslo, Norway. Three CRC cell lines HT29, HCT116 and NCI-H508, were used in the study. They were obtained from American Type Culture Collection (Manassas, VA, USA), and were authenticated by in-lab STR analysis [32].

Research on the biomaterial, including with use of deep sequencing technology, was approved by the Regional Ethics Committee of South-Eastern Norway (2010/1805/REK south-east C).

Exon level gene expression analysis

We previously performed genome-scale exon-level analyses of gene expression for 125 CRC and 21 normal mucosa samples [33–35] that have been deposited in the NCBI's Gene Expression Omnibus [GEO; accession numbers GSE24550, GSE29638, GSE42690]. Here an additional 77 CRC from the same series are analyzed using the Affymetrix HuEx-1_0-st-v2 arrays. The samples were prepared and hybridized onto the arrays according to the manufacturer's instructions, as previously described [33]. From scanned images of the microarrays, cell intensity (CEL) files were generated by the Affymetrix GeneChip Command Console software (version 1.0). Using the Affymetrix Expression Console software (version 1.1), raw data was preprocessed by background correction of individual probes based on GC-content, inter-chip quantile normalization, and eventually summarized on the exon (core probe set) level by the robust multi-array average approach [36]. Raw and processed data are deposited in GEO under accession number GSE69182. For each probe set, the log2 expression signal of each sample relative to the median of that probe set across all samples was calculated. For the RACE analyses of candidate genes, we included 19 of the 202 CRCs with significant deviating exon-level expression profiles for at least one of the candidate genes each, and one normal colonic mucosa sample, as well as the three CRC cell lines (HT29, HCT116 and NCI-H508).

Computational selection of fusion transcript candidates

An algorithm was developed to detect samples that possess abnormal expression profiles from the exon microarray data. The microarray data consists of intensities measured by a total number of 287,329 probe sets with an average of four probes per exon annotated by RefSeq [37] and full-length mRNA GenBank records [38]. To detect samples that diverge from the rest of the set, normalization by subtraction of the median signal intensity was applied for each individual probe set. Normalization was followed by division of the standard deviation at each probe set. Sample *i* is assigned a normalized relative expression value at probe set *j* that equals

$$s_{i,j} = \frac{p_{i,j} - \mu_j}{\sigma_j}$$

where $p_{i,j}$ is the log signal intensity, and μ_j and σ_j are the median value and the standard deviation across all samples at probe set *j*, respectively. The EBS for sample *i* in combination with a particular gene *g* that is targeted by *k* probe sets equals the maximum magnitude of the difference of means from 5' to 3' along the gene:

$$\mathrm{EBS}_{ig} = \max_{j=1,\ldots,k-1} \left|\mathrm{mean}\left(s_{i,j+1}, \ldots, s_{i,k}\right) - \mathrm{mean}\left(s_{i,1}, \ldots, s_{i,j}\right)\right|,$$

where *k* is the position of the putative breakpoint along the 5' to 3' axis of the gene. Large EBS_{ig} is therefore indicative of candidates where the expression level of sample *i* deviates from the rest of the set in either end of gene *g*. In most cases, fusion transcripts are regulated by the promoter of the 5' fusion partner [39]. Thus, the 3' fusion partners often show increased signal intensities downstream of the breakpoint, due to the influence of a more active promoter. Only candidates with increased expression at the 3' end, and thus being potential 3' fusion partners, were nominated for further analysis. Genes with elevated EBS in any of the normal samples were discarded from further analyses. Top-scoring gene candidates were manually curated by inspecting the probe sets underlying the deviating exon expression profiles in the Annmap database [40]. Gene candidates with deviating exon expression profiles caused by probe sets mapping to several paralogous gene copies in the genome were filtered out, as well as profiles that were likely to be caused by alternative transcription of already annotated transcript variants. For robustness, the algorithm selected only fusion partner candidates where more than one probe set had different intensity levels.

5’ Rapid Amplification of cDNA ends (RACE)

We designed nested RACE-PCR assays for each of the candidate genes and three genes used as positive controls (*TCF7L2*, *RP11-57H14.3*, and *VNN1*). Based on the exon-level expression profiles, a reverse gene-specific primer (GSP) and a reverse NGSP were designed against sequences downstream of the expected breakpoints. A forward internal control primer (ICP) was designed against sequences upstream of the same gene, to be used as a positive control and for primer optimization. To enable touchdown PCR, all primers were designed using primer 3 software [41] with theoretical T_m > 70°C, GC content 50-70% and length 23-28 nucleotides. Primer sequences for all gene assays are listed in Table S5 (in additional file 2). To obtain the full-length 5’ end sequence of the mRNA transcripts, 5’ RACE-ready cDNA was synthesized from 1 µg total RNA for each of the 24 samples using the SMARTer™ RACE cDNA Amplification kit according to protocol (Clontech, Mountain View, CA, USA). This technology incorporates the use of an oligo (dT) primer in combination with a SMARTer II A oligonucleotide adapter and the SMARTScribe Reverse Transcriptase, which in effect generates modified cDNAs containing 5’ adapter sequences from all transcribed mRNAs. Products were diluted in 100 µl milli-Q water. Synthesis of 5’ RACE-ready cDNA was confirmed by a control PCR assay detecting the housekeeping gene *GUSB*. Additionally, negative no-template controls were included in each set-up to control for contamination.

First-round RACE-PCR was performed for all candidate genes in a 10 µl reaction volume. 5’ RACE-ready cDNA from the 24 samples was diluted 4x before 2 µl of this dilution was used as input. In total, 672 first round RACE-PCR reactions were performed separately. Touchdown PCR was used, which started with 5 cycles of denaturation at 94°C for 30 seconds, annealing and extension at 72°C for 3 minutes, followed by 5 cycles of denaturation at 94°C for 30 seconds, annealing at 70°C for 30 seconds and extension at 72°C for 3 minutes, and finally, 25 cycles with denaturation at 94°C for 30 seconds, annealing at 68°C for 30 seconds and extension at 72°C for 3 minutes. The first-round PCR products were diluted 50x and 2.5 µl of each was used as template input into the nested RACE-PCR. The nested RACE-PCR reactions were carried out in a 25 µl reaction volume for each assay in each of the samples. To normalize the amounts of nested RACE products from each reaction we used a quantitative DNA binding approach (SequalPrep™ Normalization Plate Kit, 96-well; Applied Biosystems® by Life Technologies, Carlsbad, CA, USA). We added 20 µl nested RACE products to each well in the normalization plates and continued according to the manufacturer’s protocol. For each sample, equal volumes of normalized amplicons from the 28 different assays were pooled together. The pools were further quantified using the Qubit® 2.0 Fluorometer and Qubit® dsDNA HS Assay Kit. One sample was found to have insufficient amount of pooled amplicons, leaving 23 samples for sequencing. For each sample, 1 ng of the nested RACE product pools was used as input to the Nextera XT sample preparation protocol.

Sample preparation for MiSeq sequencing

The Nextera XT sample preparation kit from Illumina (San Diego, CA, USA) was used. We performed tagmentation of the input nested RACE products, before performing 12 cycles amplification of the fragments, both according to the manufacturer's protocol. During amplification, dual indexes were added to enable multiplexing of the nested RACE product pools. The amplicon pools were further cleaned and size selected using 0.6x Ampure XP beads. A subset of cleaned libraries was quality controlled using Bioanalyzer HS DNA chips. After clean up, we normalized the pools using the provided library normalization beads. To pool the samples, 5 µl of each library was combined in an Eppendorf tube and 24 µl of the pool was mixed with 576 µl of hybridization buffer. The library pool was then loaded onto a thawed MiSeq reagent cartridge version v2 before 2 x 150 nucleotides sequencing in the MiSeq instrument.

Identification of upstream fusion partners

After demultiplexing the paired-end reads, fastq files from the individual samples were subjected to quality control and inspection by using the FastQC software [42]. The sequencing reads were further trimmed to remove bad quality ends, and also to remove SMARTer II adapter oligos that were adhered during RACE cDNA synthesis. Trimming and filtering was done using the cutadapt software [43]. Trimmed reads from each sample were aligned to the Ensembl GRCh37 (iGenomes May 14 2014) reference using Tophat2 v.2.0.11 [44], Bowtie2 v.2.1.0 [45] and Samtools v.0.1.19 [46]. Normalized read counts were generated for the top 100 expressed genes using HTseq-count [47] and the R package DESeq2 [48].

For all samples, the gene fusion discovery software deFuse v.0.6.0 [49] was used to identify fusion transcripts that included one of the nominated candidate genes. In addition, normalized read counts were used to identify sample-specific expressed genes which were amplified using the 5’ RACE, but not among the original candidate genes. To determine which of these genes had high 5’ read coverage, visual inspection of aligned reads was done using the Integrative Genomics Viewer (IGV) [50,51]. Genes identified in this manner were specifically searched for in the deFuse candidate lists. If not nominated by deFuse, the Unix command grep was used to find reads containing parts of the NGSP primer for the candidate gene in the specific sample. These reads were further aligned with BLAT [52] to identify if parts of the reads connected to the sample-specific expressed gene that was not targeted by the RACE gene-specific primers.

In silico validation of fusion transcripts in external datasets

For *in silico* validation of fusion transcripts, raw RNA sequencing data (paired-end fastq files; 48 nucleotides read-length) from 126 CRCs and 24 matched normal colonic mucosa from TCGA (Table S6 in additional file 2) was downloaded from The Cancer Genomics Hub (CGHub; https://browser.cghub.ucsc.edu/). As additional normal control samples, paired-end RNA sequencing data from the Illumina Human Body Map v2 dataset consisting of 16 non-malignant miscellaneous tissue types was downloaded (ArrayExpress accession ID E-MTAB-513 and European Nucleotide Archive study accession ID ERP000546; paired-end fastq files; 50 nucleotides read-length). Moreover, aligned BAM files were obtained for 56 colon adenocarcinoma cell lines from CCLE and converted back to paired end fastq files (101 nucleotides read-length; Table S7 in additional file 2). The deFuse software was applied to detect genome-wide fusion candidates in the TCGA tumor and normal colonic mucosa samples, as well as in the CRC cell lines. Potential fusions were filtered against the Illumina Human Body Map v2 dataset. To specifically examine candidate fusion genes *DHX35-BPIFA2*, *CASZ1-MASP2*, *VWA2-TCF7L2*, and *VTI1A-TCF7L2*, the Unix command grep was used to search for scaffold sequences spanning 8, 10, 15, 20, and 24 bases at each side of the putative breakpoints.

Identification of novel and cancer-specific transcript variants

Transcript splice junctions (*i.e.* possible cancer specific alternative splicing variants) were first identified from the junctions.bed files generated at the alignment step with Tophat2. The junctions.bed files from 19 CRC tumor samples and three CRC cell lines were merged. To keep only novel junctions, those previously annotated in Ensembl release 75 were removed. To select for cancer specificity, junctions that also were detected from the normal sample included in the RACE-seq as well as junctions from any of the 16 Illumina human body map tissues were filtered out. Finally, only junctions supported by a minimum of 100 split reads were kept for downstream analysis.

Transcript junctions that corresponded to intragenic differential exon-level expression for a candidate gene were selected from this list for *in silico* validation. Additional junctions were added for validation based on manual inspection of the corresponding sample's read alignment in IGV. Selected transcript junctions were used as input when generating genome index files (using the --sjdbFileChrStartEnd flag) from Ensembl release 75 with the STAR aligner v.2.4.0d [53]. Fastq files from the downloaded validation series were aligned to the generated genome index. In total, paired-end data from 24 matched tumor and normal pairs from the TCGA colon adenocarcinoma series, 56 colon adenocarcinoma cell lines from the CCLE and 16 miscellaneous tissues from the Illumina human body map were aligned to determine the presence of the selected transcript junctions.

Availability of supporting data

The data sets supporting the results of this article are available in the NCBI's Gene Expression Omnibus repository,
GSE24550 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE24550),
GSE29638 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE29638),
GSE42690 (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE42690),
GSE69182
(http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?token=qxobkqiyhvefbkn&acc=GSE69182).

Abbreviations

CRC: Colorectal cancer
RACE: Rapid amplification of cDNA ends
EBS: Expression break score
NGSP: Nested gene specific primer
GSP: Gene specific primer
ORF: Open reading frame
TCGA: The Cancer Genome Atlas
CCLE: Cancer Cell Line Encyclopedia
IGV: Integrated Genome Viewer
ENCODE: Encyclopedia of DNA Elements

Competing interests

The author(s) declare that they have no competing interests.

Authors' contributions

BJ, SA and AS conducted the exon microarray experiments and associated data analyses. AMH and ML performed the RACE-seq experiments. AMH, BJ, SZ, and TN analyzed the RACE-seq data. AMH, ACB, and MH processed biological material, including performing PCR and Sanger sequencing. AMH and BJ drafted the manuscript. RAL and RIS conceived and directed the project. All authors participated in discussion of results, writing and have read and approved the final manuscript.

Acknowledgements

The study was funded by grants from the Norwegian Cancer Society (grant numbers PR-2006-0442 to R.A.L., PR-2007-0166 to R. I. S., which also financed A. M. H. as a PhD-student), South-Eastern Norway Regional Health Authority (B. J. was financed as PhD-student from the Project #2012067 grant), and partly supported by the Research Council of Norway through its Centres of Excellence funding scheme, project number 179571. We also acknowledge granted storage of computer files by NorStore (project NS9013K) and processing power from Notur (project NN9313K). We are grateful for the contribution from our clinical collaborator Professor Arild Nesbakken and his team.

References

1. Torre LA, Bray F, Siegel RL, Ferlay J, Lortet-Tieulent J, Jemal A: **Global cancer statistics, 2012**. CA Cancer J Clin 2015, 65:87–108.

2. Sotiriou C, Pusztai L: **Gene-Expression Signatures in Breast Cancer**. N Engl J Med 2009, 360:790–800.

3. Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, *et al.*: **Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling**. Nature 2000, 403:503–511.

4. Sveen A, Nesbakken A, Ågesen TH, Guren MG, Tveit KM, Skotheim RI, *et al.*: **Anticipating the Clinical Use of Prognostic Gene Expression–Based Tests for Colon Cancer Stage II and III: Is Godot Finally Arriving?** Clin Cancer Res 2013, 19:6669–6677.

5. Venables JP: **Aberrant and Alternative Splicing in Cancer**. Cancer Res 2004, 64:7647–7654.

6. Skotheim RI, Nees M: **Alternative splicing in cancer: noise, functional, or systematic?** Int J Biochem Cell Biol 2007, 39:1432–1449.

7. Mitelman F, Johansson B, Mertens F: **Fusion genes and rearranged genes as a linear function of chromosome aberrations in cancer**. Nat Genet 2004, 36:331–334.

8. Mitelman F, Johansson B, Mertens F: **The impact of translocations and gene fusions on cancer causation**. Nat Rev Cancer 2007, 7:233–245.

9. Tomlins SA, Rhodes DR, Perner S, Dhanasekaran SM, Mehra R, Sun XW, *et al.*: **Recurrent fusion of *TMPRSS2* and ETS transcription factor genes in prostate cancer**. Science 2005, 310:644–648.

10. Djebali S, Lagarde J, Kapranov P, Lacroix V, Borel C, Mudge JM, *et al.*: **Evidence for Transcript Networks Composed of Chimeric RNAs in Human Cells**. PLoS ONE 2012, 7:e28213.

11. Frenkel-Morgenstern M, Lacroix V, Ezkurdia I, Levin Y, Gabashvili A, Prilusky J, *et al.*: **Chimeras taking shape: Potential functions of proteins encoded by chimeric RNA transcripts**. Genome Res 2012, 7:1231–1242.

12. Plebani R, Oliver GR, Trerotola M, Guerra E, Cantanelli P, Apicella L, *et al.*: **Long-range transcriptome sequencing reveals cancer cell growth regulatory chimeric mRNA**. Neoplasia 2012, 14:1087–1096.

13. Zhang Y, Gong M, Yuan H, Park HG, Frierson HF, Li H: **Chimeric Transcript Generated by cis-Splicing of Adjacent Genes Regulates Prostate Cancer Cell Proliferation**. Cancer Discov 2012, 7:598–607.

14. Yun SM, Yoon K, Lee S, Kim E, Kong S-H, Choe J, *et al.*: ***PPP1R1B-STARD3*** **chimeric fusion transcript in human gastric cancer promotes tumorigenesis through activation of PI3K/AKT signaling**. Oncogene 2014, 33:5341–5347.

15. Frenkel-Morgenstern M, Gorohovski A, Vucenovic D, Maestre L, Valencia A: **ChiTaRS 2.1-an improved database of the chimeric transcripts and RNA-seq data with novel sense-antisense chimeric RNA transcripts**. Nucleic Acids Res 2014, 43:D68–D75.

16. Rickman DS, Pflueger D, Moss B, VanDoren VE, Chen CX, de la Taille A, *et al.*: ***SLC45A3-ELK4*** **is a novel and frequent erythroblast transformation-specific fusion transcript in prostate cancer**. Cancer Res 2009, 69:2734–2738.

17. Maher CA, Kumar-Sinha C, Cao X, Kalyana-Sundaram S, Han B, Jing X, *et al.*: **Transcriptome sequencing to detect gene fusions in cancer**. Nature 2009, 458:97–101.

18. Bass AJ, Lawrence MS, Brace LE, Ramos AH, Drier Y, Cibulskis K, *et al.*: **Genomic sequencing of colorectal adenocarcinomas identifies a recurrent** ***VTI1A-TCF7L2*** **fusion**. Nat Genet 2011, advance online publication.

19. Nome T, Hoff AM, Bakken AC, Rognum TO, Nesbakken A, Skotheim RI: **High Frequency of Fusion Transcripts Involving** ***TCF7L2*** **in Colorectal Cancer: Novel Fusion Partner and Splice Variants**. PLoS One 2014, 9:e91264.

20. Li H, Wang J, Ma X, Sklar J: **Gene fusions and RNA trans-splicing in normal and neoplastic human cells**. Cell Cycle 2009, 8:218–222.

21. Nome T, Thomassen GO, Bruun J, Ahlquist T, Bakken AC, Hoff AM, *et al.*: **Common fusion transcripts identified in colorectal cancer cell lines by high-throughput RNA sequencing**. Transl Oncol 2013, 6:546–553.

22. Løvf M, Nome T, Bruun J, Eknæs M, Bakken AC, Mpindi JP, *et al.*: **A novel transcript,** ***VNN1-AB*****, as a biomarker for colorectal cancer**. Int J Cancer 2014, 135:2077–2084.

23. Edgren H, Murumagi A, Kangaspeska S, Nicorici D, Hongisto V, Kleivi K, *et al.*: **Identification of fusion genes in breast cancer by paired-end RNA-sequencing**. Genome Biol 2011, 12:R6.

24. Denoeud F, Kapranov P, Ucla C, Frankish A, Castelo R, Drenkow J, *et al.*: **Prominent use of distal 5' transcription start sites and discovery of a large number of additional exons in ENCODE regions**. Genome Res 2007, 17:746–759.

25. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, *et al.*: **Landscape of transcription in human cells**. Nature 2012, 489:101–108.

26. Walker F, Nicole P, Jallane A, Soosaipillai A, Mosbach V, Oikonomopoulou K, *et al.*: **Kallikrein-related peptidase 7 (*****KLK7*****) is a proliferative factor that is aberrantly expressed in human colon cancer**. Biol Chem 2014, 395:1075–1086.

27. Talieri M, Mathioudaki K, Prezas P, Alexopoulou DK, Diamandis EP, Xynopoulos D, *et al.*: **Clinical significance of kallikrein-related peptidase 7 (*KLK7*) in colorectal cancer**. Thromb Haemost 2009, 104:741–747.

28. Talieri M, Li L, Zheng Y, Alexopoulou DK, Soosaipillai A, Scorilas A, *et al.*: **The use of kallikrein-related peptidases as adjuvant prognostic markers in colorectal cancer**. Br J Cancer 2009, 100:1659–1665.

29. Christophi GP, Isackson PJ, Blaber S, Blaber M, Rodriguez M, Scarisbrick IA: **Distinct promoters regulate tissue-specific and differential expression of kallikrein 6 in CNS demyelinating disease**. J Neurochem 2004, 91:1439–1449.

30. Salama I, Malone PS, Mihaimeed F, Jones JL: **A review of the S100 proteins in cancer**. Eur J Surg Oncol 2008, 34:357–364.

31. Giráldez MD, Lozano JJ, Cuatrecasas M, Alonso-Espinaco V, Maurel J, Mármol M, *et al.*: **Gene-expression signature of tumor recurrence in patients with stage II and III colon cancer treated with 5'fluoruracil-based adjuvant chemotherapy**. Int J Cancer J Int Cancer 2013, 132:1090–1097.

32. Ahmed D, Eide PW, Eilertsen IA, Danielsen SA, Eknæs M, Hektoen M, *et al.*: **Epigenetic and genetic features of 24 colon cancer cell lines**. Oncogenesis 2013, 2:e71.

33. Sveen A, Ågesen TH, Nesbakken A, Rognum TO, Lothe RA, Skotheim RI: **Transcriptome instability in colorectal cancer identified by exon microarray analyses: Associations with splicing factor expression levels and patient survival**. Genome Med 2011, 3:32.

34. Agesen TH, Sveen A, Merok MA, Lind GE, Nesbakken A, Skotheim RI, *et al.*: **ColoGuideEx: a robust gene classifier specific for stage II colorectal cancer prognosis**. Gut 2012, 61:1560–1567.

35. Sveen A, Johannessen B, Teixeira MR, Lothe RA, Skotheim RI: **Transcriptome instability as a molecular pan-cancer characteristic of carcinomas**. BMC Genomics 2014, 15:672.

36. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, *et al.*: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data**. Biostatistics 2003, 4:249–264.

37. Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, *et al.*: **RefSeq: an update on mammalian reference sequences**. Nucleic Acids Res 2014, 42:D756–D763.

38. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank**. Nucleic Acids Res 2006, 34:D16–D20.

39. Aman P: **Fusion genes in solid tumors**. Semin Cancer Biol 1999, 9:303–318.

40. Yates T, Okoniewski MJ, Miller CJ: **X:Map: annotation and visualization of genome structure for Affymetrix exon array analysis**. Nucleic Acids Res 2008, 36:D780–D786.

41. Rozen S, Skaletsky H: **Primer3 on the WWW for general users and for biologist programmers**. Methods MolBiol 2000, 132:365–386.

42. Andrews S: **FastQC a quality-control tool for high-throughput sequence data**. [Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/].

43. Martin M: **Cutadapt removes adapter sequences from high-throughput sequencing reads**. EMBnet J 2011, 17.

44. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq**. Bioinformatics 2009, 25:1105–1111.

45. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. Genome Biol 2009, 10:R25.

46. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, *et al.*: **The Sequence Alignment/Map format and SAMtools**. Bioinformatics 2009, 25:2078–2079.

47. Anders S, Pyl PT, Huber W: **HTSeq—a Python framework to work with high-throughput sequencing data**. Bioinformatics 2015, 31:166–169.

48. Love MI, Huber W, Anders S: **Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2**. Genome Biol 2014, 15:550.

49. McPherson A, Hormozdiari F, Zayed A, Giuliany R, Ha G, Sun MGF, *et al.*: **deFuse: An Algorithm for Gene Fusion Discovery in Tumor RNA-Seq Data**. PLoS Comput Biol 2011, 7:e1001138.

50. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, *et al.*: **Integrative genomics viewer**. Nat Biotechnol 2011, 29:24–26.

51. Thorvaldsdóttir H, Robinson JT, Mesirov JP: **Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration**. Brief Bioinform 2013, 14:178–192.

52. Kent WJ: **BLAT-The BLAST-Like Alignment Tool**. Genome Res 2002, 12:656–664.

53. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, *et al.*: **STAR: ultrafast universal RNA-seq aligner**. Bioinformatics 2013, 29:15–21.

Additional files

Additional file 1: Contains supplementary figures S1A to S1Y

Figure S1A

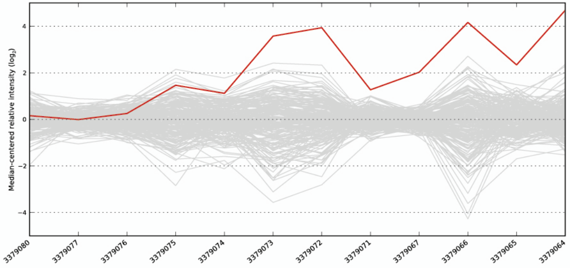


Figure S1B

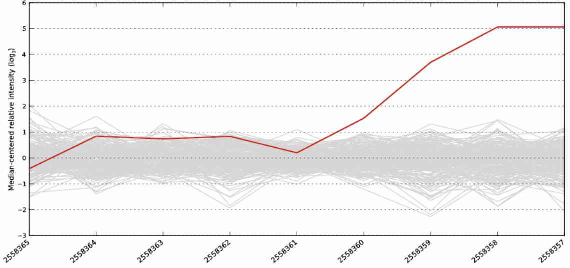


Figure S1C

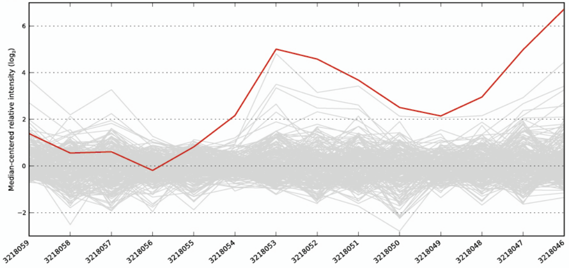


Figure S1D

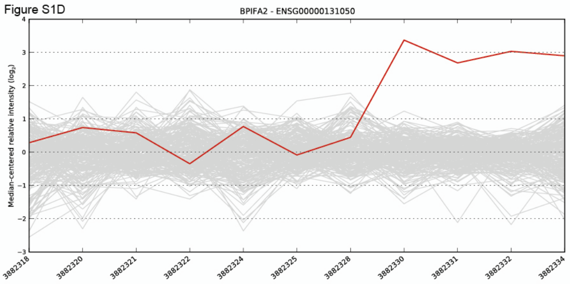


Figure S1E

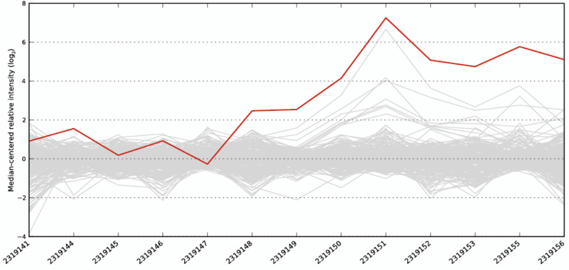


Figure S1F

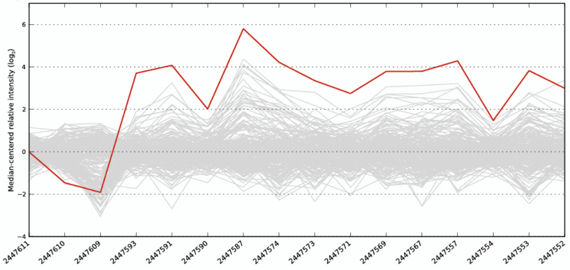


Figure S1G

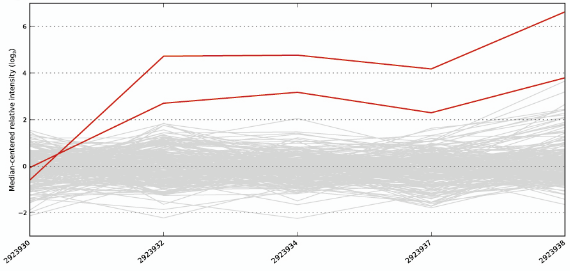


Figure S1H

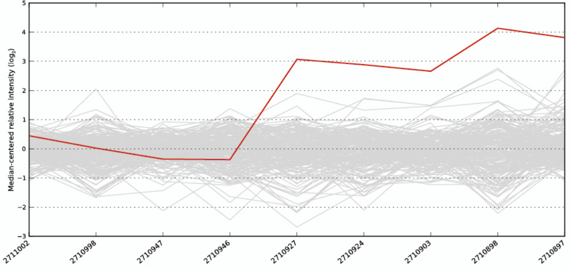


Figure S1I

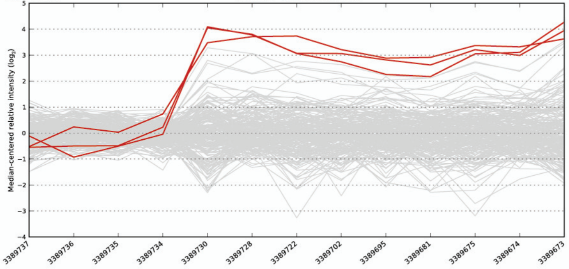


Figure S1J

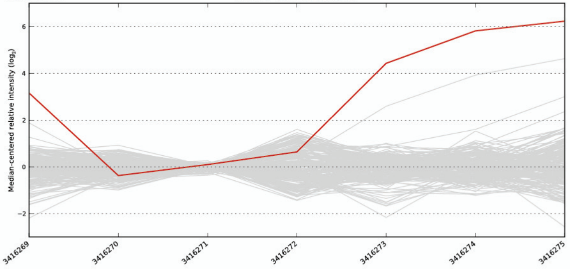


Figure S1K

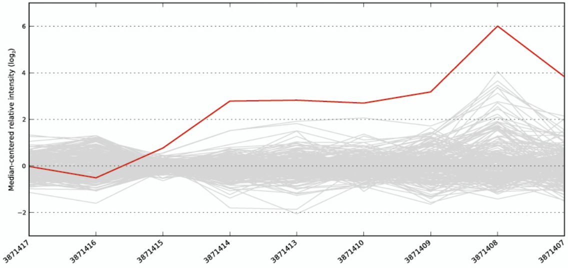


Figure S1L

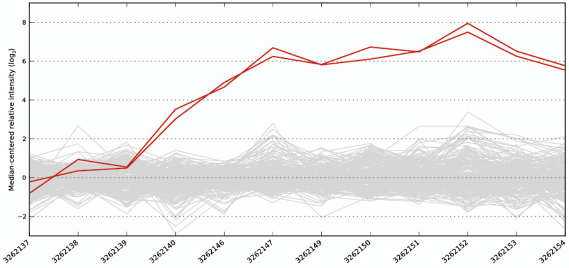


Figure S1M

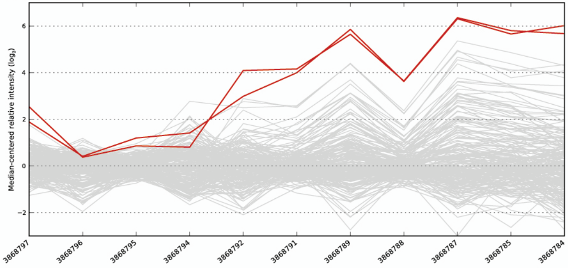


Figure S1N

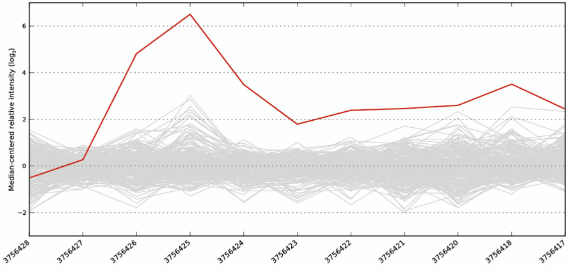


Figure S1O

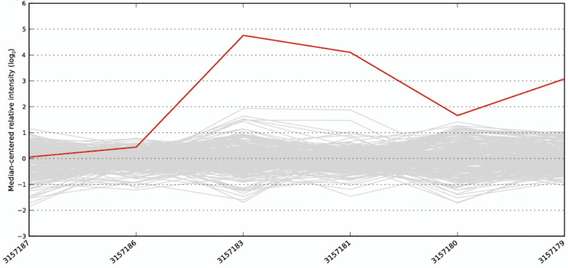


Figure S1P

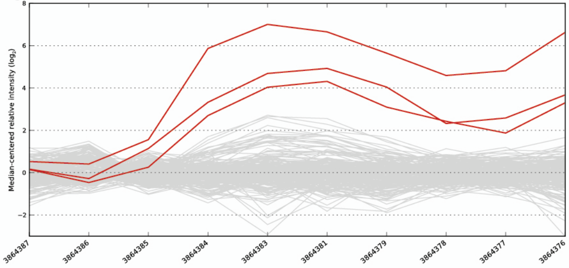


Figure S1Q

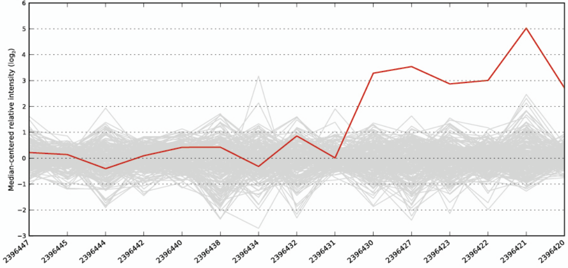


Figure S1R

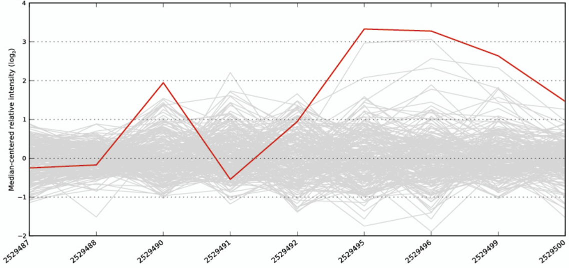


Figure S1S

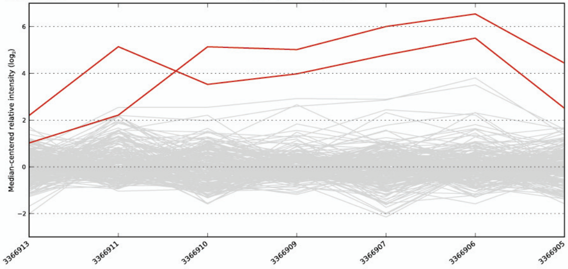
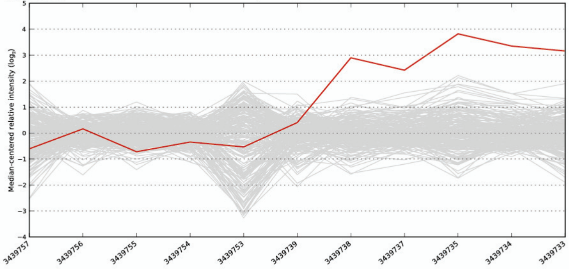


Figure S1T



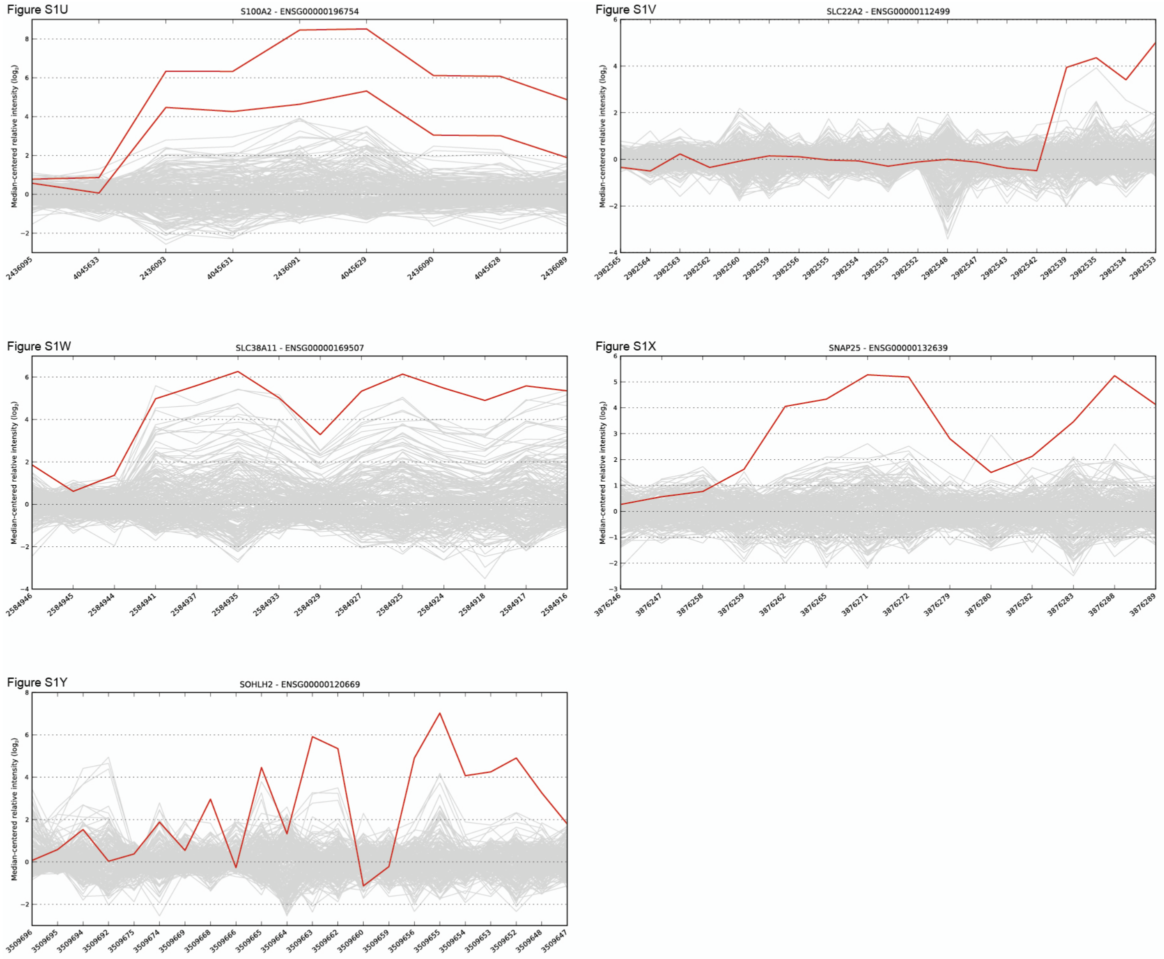


Figure S1: Exon-level expression profiles for all 25 selected candidate genes. Each line represents an individual tumor sample, with the expression values median-centered and the most significantly deviating sample(s) marked in red. X-axis numbering refers to probe sets on the Affymetrix HuEx-1_0-st-v2 microarray.

Additional file 2: Contains supplementary Tables S1 to S7

Table S1: Sequence reads produced by high-throughput sequencing of RACE products, and alignment numbers before and after trimming.

	WITHOUT TRIMMING						AFTER TRIMMING					
Samples	**Left reads input**	**Left reads mapped**	**Right reads input**	**Right reads mapped**	**Aligned pairs**	**Overall read mapping rate**	**Left reads input**	**Left reads mapped**	**Right reads input**	**Right reads mapped**	**Aligned pairs**	**Overall read mapping rate**
Sample1_T	337490	134495 (39.9%)	337490	185811 (55.1%)	111444	47.5%	290733	218985 (75.3%)	290733	220021 (75.7%)	184302	75.5%
Sample2_T	405183	126071 (31.1%)	405183	147152 (36.3%)	106969	33.7%	286407	170232 (59.4%)	286407	166211 (58.0%)	143493	58.7%
Sample3_T	490151	240423 (49.1%)	490151	227980 (46.5%)	184732	47.8%	429720	341214 (79.4%)	429720	333629 (77.6%)	296755	78.5%
Sample4_T	92558	30186 (32.6%)	92558	50897 (55.0%)	26168	43.8%	69061	53984 (78.2%)	69061	56101 (81.2%)	46894	79.7%
Sample5_T	259254	91045 (35.1%)	259254	112686 (43.5%)	82250	39.3%	172709	133414 (77.2%)	172709	131292 (76.0%)	115564	76.6%
Sample6_T	221512	93812 (42.4%)	221512	121830 (55.0%)	81132	48.7%	172586	134683 (78.0%)	172586	135963 (78.8%)	116717	78.4%
Sample7_N	949370	641862 (67.6%)	949370	615850 (64.9%)	547377	66.2%	832470	709983 (85.3%)	832470	695880 (83.6%)	646394	84.4%
Sample7_T	937297	474458 (50.6%)	937297	458915 (49.0%)	394124	49.8%	737717	573080 (77.7%)	737717	558646 (75.7%)	498866	76.7%
Sample8_T	537716	281907 (52.4%)	537716	251030 (46.7%)	223229	49.6%	438703	374265 (85.3%)	438703	361323 (82.4%)	332897	83.8%
Sample9_T	721347	348366 (48.3%)	721347	348316 (48.3%)	296311	48.3%	554354	426417 (76.9%)	554354	415051 (74.9%)	367984	75.9%
Sample10_T	938279	607220 (64.7%)	938279	611892 (65.2%)	547282	65.0%	820110	700296 (85.4%)	820110	687235 (83.8%)	639160	84.6%
Sample11_T	828108	534222 (64.5%)	828108	532212 (64.3%)	473642	64.4%	713173	603405 (84.6%)	713173	594964 (83.4%)	548878	84.0%
Sample12_T	891642	484725 (54.4%)	891642	469390 (52.6%)	395853	53.5%	721834	581393 (80.5%)	721834	570057 (79.0%)	509679	79.8%
Sample13_T	954285	544282 (57.0%)	954285	518540 (54.3%)	452876	55.7%	797104	662655 (83.1%)	797104	643236 (80.7%)	590462	81.9%
Sample14_T	913931	422375 (46.2%)	913931	371684 (40.7%)	329406	43.4%	694342	559643 (80.6%)	694342	538048 (77.5%)	488309	79.0%
Sample15_T	1003262	557686 (55.6%)	1003262	562527 (56.1%)	483618	55.8%	818000	689581 (84.3%)	818000	679539 (83.1%)	625629	83.7%
Sample16_T	955003	574108 (60.1%)	955003	591286 (61.9%)	514575	61.0%	824191	700553 (85.0%)	824191	695146 (84.3%)	651450	84.7%
Sample17_T	1042165	589295 (56.5%)	1042165	586722 (56.3%)	507243	56.4%	821846	688397 (83.8%)	821846	678353 (82.5%)	629075	83.2%
Sample18_T	1246399	672611 (54.0%)	1246399	575291 (46.2%)	498652	50.1%	993465	795391 (80.1%)	993465	766917 (77.2%)	687878	78.6%
Sample19_T	460095	246339 (53.5%)	460095	207988 (45.2%)	179428	49.4%	348581	272972 (78.3%)	348581	262408 (75.3%)	231736	76.8%
HCT-116	287685	190561 (66.2%)	287685	178717 (62.1%)	156036	64.2%	240168	204653 (85.2%)	240168	200354 (83.4%)	185791	84.3%
HT29	782188	553726 (70.8%)	782188	408693 (52.2%)	370633	61.5%	614865	541431 (88.1%)	614865	511567 (83.2%)	479949	85.6%
NCI-H508	209624	139209 (66.4%)	209624	129023 (61.5%)	113157	64.0%	173914	149068 (85.7%)	173914	145104 (83.4%)	134019	84.6%

Table S2: Novel transcript splice junctions covered by a minimum of 100 reads from the RACE-seq data.

Junctions				Samples																						
Chromosome	**Pos1**	**Pos2**	**Distance**	**1_T**	**2_T**	**3_T**	**4_T**	**5_T**	**6_T**	**7_T**	**8_T**	**9_T**	**10_T**	**11_T**	**12_T**	**13_T**	**14_T**	**15_T**	**16_T**	**17_T**	**18_T**	**19_T**	**HCT-116**	**HT29**	**NCI-H508**	**Total**
chr1	11091304	11095305	4001	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
chr1	11091320	11095321	4001	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
chr1	153536281	153538129	1848	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
chr1	153536306	153539204	2898	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
chr1	153536329	153540992	4663	0	0	0	0	0	0	0	1	0	0	0	1	1	1	0	0	0	0	0	0	0	0	4
chr1	153536329	153538641	2312	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	0	3
chr1	153536329	153538850	2521	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
chr1	153536329	153539864	3535	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
chr1	153536342	153540992	4650	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
chr1	153536357	153537982	1625	0	1	0	1	1	1	1	0	1	0	1	1	1	1	1	1	0	0	1	1	1	0	15
chr1	153536357	153540992	4635	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0	0	0	0	0	0	3
chr1	153536363	153538641	2278	0	0	1	0	0	0	0	1	0	1	0	1	1	1	0	0	1	1	0	0	0	0	8
chr1	153536363	153540244	3881	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	2
chr1	153536363	153537100	737	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
chr1	153536363	153539864	3501	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
chr1	153536363	153540861	4498	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
chr1	153536378	153537990	1612	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
chr1	153539339	153540992	1653	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	2
chr1	153539339	153540861	1522	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
chr1	183947654	184021555	73901	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	3
chr1	183947654	184006715	59061	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2
chr1	183947654	184020729	73075	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
chr1	183966636	184020729	54093	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
chr2	70188831	70190497	1666	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
chr2	70188831	70191636	2805	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
chr2	70190632	70215578	24946	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
chr2	70215637	70223818	8181	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
chr2	70224020	70242580	18560	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
chr2	165811255	165812729	1474	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
chr2	165811255	165812168	913	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
chr2	223521230	223553063	31833	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
chr2	223553241	223558204	4963	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2
chr3	171362813	171371700	8887	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
chr3	192078327	192362055	283728	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
chr3	192120485	192362055	241570	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
chr6	123068284	123101436	33152	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
chr6	123070479	123101436	30957	1	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	3
chr6	123073436	123101436	28000	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
chr6	133013676	133014130	454	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
chr6	133015321	133024025	8704	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1

Junctions				Samples																						
Chromosome	**Pos1**	**Pos2**	**Distance**	**1_T**	**2_T**	**3_T**	**4_T**	**5_T**	**6_T**	**7_T**	**8_T**	**9_T**	**10_T**	**11_T**	**12_T**	**13_T**	**14_T**	**15_T**	**16_T**	**17_T**	**18_T**	**19_T**	**HCT-116**	**HT29**	**NCI-H508**	**Total**
chr6	133032922	133034965	2043	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	2
chr7	100636026	100639276	3250	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
chr7	100636920	100641592	4672	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
chr7	100636920	100640170	3250	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
chr7	100637697	100642369	4672	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
chr7	100637697	100645618	7921	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
chr7	100637697	100640947	3250	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
chr7	100638402	100646323	7921	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
chr7	100638654	100642573	3919	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
chr7	100638988	100643660	4672	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
chr7	100638988	100646909	7921	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
chr7	100639308	100643980	4672	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
chr7	100639483	100647404	7921	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
chr7	100640152	100644824	4672	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
chr7	100640649	100642072	1423	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
chr7	100640649	100645321	4672	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
chr7	100640912	100642335	1423	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
chr7	100640912	100645584	4672	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
chr7	100641574	100644824	3250	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
chr7	100642334	100645584	3250	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
chr7	100643578	100646828	3250	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
chr7	100643959	100647788	3829	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
chr7	100643959	100647209	3250	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
chr8	143867103	143870816	3713	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
chr9	104130599	104133596	2997	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
chr9	104130599	104130830	231	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
chr9	104130604	104133596	2992	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
chr9	104133459	104133596	137	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
chr9	104133745	104146025	12280	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
chr9	104147200	104152859	5659	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	2
chr10	114428768	114648494	219726	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	2
chr10	114708157	114710525	2368	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
chr10	114711032	114711250	218	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	0	1	1	0	0	0	5
chr10	114711645	114799784	88139	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
chr10	114711885	114799784	87899	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
chr10	114724383	114781575	57192	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
chr10	114724424	114799784	75360	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2
chr10	114728153	114799784	71631	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	2
chr10	114799885	114849217	49332	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
chr10	114816757	114900943	84186	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
chr10	114818199	114900943	82744	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Junctions				Samples																						
Chromosome	Pos1	Pos2	Distance	1_T	2_T	3_T	4_T	5_T	6_T	7_T	8_T	9_T	10_T	11_T	12_T	13_T	14_T	15_T	16_T	17_T	18_T	19_T	HCT-116	HT29	NCI-H508	Total
chr10	114833700	114900943	67243	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2
chr10	114833787	114900943	67156	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
chr10	114845992	114900943	54951	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	1	0	0	0	0	0	4
chr10	114845992	114849159	3167	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2
chr10	114847225	114849159	1934	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2
chr10	114852046	114900943	48897	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
chr10	114886640	114900943	14303	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
chr10	114886812	114900943	14131	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
chr11	26587443	26618666	31223	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2
chr11	26588686	26618666	29980	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
chr11	26588686	26604673	15987	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
chr11	26593653	26604673	11020	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	3
chr11	26593653	26618666	25013	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
chr11	26593653	26612818	19165	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
chr11	67412590	67414367	1777	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
chr11	67412623	67413912	1289	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
chr11	67412845	67413163	318	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	0	0	0	4
chr11	67413352	67415006	1654	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
chr11	67413352	67415061	1709	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
chr11	67414009	67418112	4103	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
chr11	67414388	67414987	599	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
chr11	106856857	107047772	190915	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
chr11	106856857	106962905	106048	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
chr12	675291	719533	44242	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
chr12	675344	753140	77796	0	0	0	0	1	0	1	1	1	0	0	1	0	0	0	0	1	0	1	0	1	0	8
chr12	675344	750526	75182	1	0	0	0	0	0	1	1	1	0	0	1	0	0	0	1	0	1	1	0	0	0	8
chr12	675344	707605	32261	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
chr12	675344	727730	52386	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
chr12	675728	752677	76949	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
chr12	698975	752677	53702	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	2
chr12	719625	752677	33052	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
chr12	730296	750526	20230	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
chr12	750656	752677	2021	1	0	0	0	0	0	1	1	1	0	0	1	1	0	0	1	0	1	1	0	0	0	9
chr12	52824454	53293815	469361	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
chr12	52845415	52913619	68204	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
chr12	52845451	52867111	21660	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
chr12	52867074	52913619	46545	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
chr12	52884482	53293815	409333	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2
chr12	52884493	53293826	409333	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2
chr12	53044337	53048689	4352	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
chr12	53186152	53293815	107663	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2

Junctions				Samples																						
Chromosome	**Pos1**	**Pos2**	**Distance**	**1_T**	**2_T**	**3_T**	**4_T**	**5_T**	**6_T**	**7_T**	**8_T**	**9_T**	**10_T**	**11_T**	**12_T**	**13_T**	**14_T**	**15_T**	**16_T**	**17_T**	**18_T**	**19_T**	**HCT-116**	**HT29**	**NCI-H508**	**Total**
chr12	53294428	53295695	1267	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
chr12	53294440	53298442	4002	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
chr12	53294440	53298733	4293	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1
chr12	53295801	53298773	2972	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
chr12	53298592	53298678	86	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
chr12	54343920	54350112	6192	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
chr12	54348862	54350112	1250	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
chr13	36767850	36767943	93	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
chr17	38858185	38859846	1661	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
chr17	38859490	38859844	354	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
chr17	38859490	38878083	18593	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
chr17	38859593	38978450	118857	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
chr17	38859905	38878145	18240	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
chr19	51451974	51501086	49112	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
chr19	51485680	51504354	18674	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1
chr19	51485704	51487107	1403	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
chr19	51487155	51504358	17203	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
chr19	51501060	51504354	3294	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1
chr19	55880298	55884648	4350	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1
chr20	10199625	10204734	5109	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
chr20	10199625	10218621	18996	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
chr20	10199625	10207326	7701	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
chr20	10199625	10208026	8401	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
chr20	10204758	10214859	10101	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
chr20	10208086	10256077	47991	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Table S3: Transcript splice junctions used as input to the STAR aligner

Gene	Ensembl_ID	Chromosome	strand	Pos1	Pos2	Distance	Note
CA6	ENSG00000131686	1	+	9006024	9009322	3298	alt_splice
CA6	ENSG00000131686	1	+	9006024	9017196	11172	alt_splice
CA6	ENSG00000131686	1	+	9006024	9018969	12945	alt_splice
CA6	ENSG00000131686	1	+	9009501	9017196	7695	alt_splice
CA6	ENSG00000131686	1	+	9017344	9018969	1625	alt_splice
S100A2	ENSG00000196754	1	-	153536357	153537982	1625	alt_splice
S100A2	ENSG00000196754	1	-	153536363	153539178	2815	alt_promo
S100A2	ENSG00000196754	1	-	153536363	153540320	3957	alt_promo
S100A2	ENSG00000196754	1	-	153536363	153580475	44112	Read-through
COLGALT2	ENSG00000198756	1	-	183947654	184020729	73075	alt_promo
COLGALT2	ENSG00000198756	1	-	183947654	184021555	73901	alt_promo
SLC38A11	ENSG00000169507	2	-	165811255	165812168	913	alt_promo
SLC38A11	ENSG00000169507	2	-	165811255	165812729	1474	alt_promo
MOGAT1	ENSG00000124003	2	+	223521230	223553063	31833	alt_promo
MOGAT1	ENSG00000124003	2	+	223553241	223558204	4963	alt_splice
MOGAT1	ENSG00000124003	2	+	223558520	223559081	561	alt_splice
FGF12	ENSG00000114279	3	-	192078327	192232456	154129	alt_promo
FGF12	ENSG00000114279	3	-	192078327	192362055	283728	alt_promo
FGF12	ENSG00000114279	3	-	192078327	192635419	557092	Fusion
FABP7	ENSG00000164434	6	+	123068284	123101436	33152	alt_promo
FABP7	ENSG00000164434	6	+	123070479	123101436	30957	alt_promo
FABP7	ENSG00000164434	6	+	123073436	123101436	28000	alt_promo
FABP7	ENSG00000164434	6	+	123100864	123101436	572	alt_promo
VNN1	ENSG00000112299	6	-	133005644	133006394	750	alt_splice
VNN1	ENSG00000112299	6	-	133006542	133007738	1196	alt_splice
SLC22A2	ENSG00000112499	6	-	160645836	160655093	9257	alt_promo
SLC22A2	ENSG00000112499	6	-	160645836	160655314	9478	alt_promo
LY6D	ENSG00000167656	8	-	143867103	143870816	3713	alt_promo
BAAT	ENSG00000136881	9	-	104130599	104130830	231	alt_promo
BAAT	ENSG00000136881	9	-	104130604	104130837	233	alt_promo
BAAT	ENSG00000136881	9	-	104130604	104133600	2996	alt_splice
BAAT	ENSG00000136881	9	-	104147200	104152859	5659	Read-through
MUC15	ENSG00000169550	11	-	26587443	26618666	31223	New_exon
MUC15	ENSG00000169550	11	-	26588686	26604673	15987	New_exon
MUC15	ENSG00000169550	11	-	26593653	26604673	11020	New_exon
MUC15	ENSG00000169550	11	-	26593653	26612818	19165	New_exon
ACY3	ENSG00000132744	11	-	67412623	67413912	1289	alt_promo
GUCY1A2	ENSG00000152402	11	-	106856857	106962905	106048	alt_promo
GUCY1A2	ENSG00000152402	11	-	106856857	107047772	190915	alt_promo
NINJ2	ENSG00000171840	12	-	675344	707605	32261	New_exon
NINJ2	ENSG00000171840	12	-	675344	750526	75182	New_exon
NINJ2	ENSG00000171840	12	-	675344	753140	77796	New_exon
NINJ2	ENSG00000171840	12	-	750656	752677	2021	New_exon
HOXC12	ENSG00000123407	12	+	54343920	54350112	6192	alt_promo
HOXC12	ENSG00000123407	12	+	54348862	54350112	1250	alt_promo
SOHLH2	ENSG00000120669	13	-	36767850	36767943	93	alt_splice
KLK7	ENSG00000169035	19	-	51485170	51504354	19184	Read-through
IL11	ENSG00000095752	19	-	55880298	55884648	4350	alt_promo
SNAP25	ENSG00000132639	20	+	10199625	10204734	5109	alt_splice
SNAP25	ENSG00000132639	20	+	10199625	10207326	7701	alt_splice
SNAP25	ENSG00000132639	20	+	10199625	10208026	8401	alt_splice
SNAP25	ENSG00000132639	20	+	10199625	10218621	18996	alt_splice
SNAP25	ENSG00000132639	20	+	10204758	10214859	10101	alt_splice
SNAP25	ENSG00000132639	20	+	10208086	10256077	47991	alt_splice
BPIFA2	ENSG00000131050	20	+	37632550	31767410	-5865140	Fusion

Table S4: Number of aligned reads from CCLE, TCGA and RACE-seq samples to input junctions, as determined by the STAR aligner. Due to the size of Table S4 it is not included here, but will be made available online with the future publication. Interested readers can request a digital copy of the table from the corresponding author.

Table S5: Primers used in the RACE amplification of target genes.

Ensembl_ID	Gene	Type	RACE Type	GSP	NGSP	ICP
ENSG00000132744	*ACY3*	Candidate	5'	GAACTGTGGCCACCAGGGTCCTCAT	TTGAGAAAATGTCAGCCCGCAGCAC	CCTGGACTCTGTGGCCAAAAATGGA
ENSG00000244617	*ASPRV1*	Candidate	5'	CCCTGGTCCTGGGGACTGAGCCTAT	GCAGCCAGAGGTTTGGGACGACATT	GGCTGGTGTTGGTCCAAGGATGGAT
ENSG00000136881	*BAAT*	Candidate	5'	TTACTTCTGGTTTGCGGGGCAGGTC	GTAAGCCAAGGCCAAGGAGGCGAAG	GGGTCTCTTCCCAGGGGTAATTGATTTG
ENSG00000131050	*BPIFA2*	Candidate	5'	TGAGGGTTTGCAGCTGGGTTTTGTG	TGGATGAAGATGCGGATCAGTGGACA	CGTGATCAACACGCTGAAAAGCACTG
ENSG00000131686	*CA6*	Candidate	5'	AACGAAGGCTGCCAGTACAGCCAAA	ATCCGGCGCATCTTGGGCTATATCA	AGATCAGCCTGCCCTCCACCATGC
ENSG00000198756	*COLGALT2*	Candidate	5'	CGAAGGGCTGCCTGTCGTAGTTTCA	TGGTTCATCCATAGGCCTCCACTCCA	CTGGAGCGGCTGGACTACCCCAAG
ENSG00000164434	*FABP7*	Candidate	5'	TCCGTGTTCTTGAATGTGCTGAGAGTCC	CGTTGGTTTGGTCACATTTCCCACCT	AGAAGGCAGGAGCTGCTTGCTGAGG
ENSG00000114279	*FGF12*	Candidate	5'	CGCTGTTTTCGTCCTTGGTCCCATC	TGCAGGAAGTATCCCTGCTGGCTGA	AGGGAGTCCAACAGCGACCGAGTGT
ENSG00000152402	*GUCY1A2*	Candidate	5'	TGCAAAGTGCCACCTACAGCTCGAA	TTGGAGTGGTCTGCATAGGAGCATCTG	AACCTGGACTCGCTGGGCGAGAG
ENSG00000123407	*HOXC12*	Candidate	5'	TGACCAGAAACTCGCCCTCCAGCTC	GCTTCTTCCGAGAGCGGCTGTTGAT	GGCAGCTTGGTATCGCCGTTGAA
ENSG00000095752	*IL11*	Candidate	5'	AGGGTGGGCAGGGAATCCAGGTTGT	GGTCCCCGTCAGCTGGGAATTTGT	GTGCTCCTGACCCGCTCTCTCCTG
ENSG00000148798	*INA*	Candidate	5'	GCTCCAGGATCTGCCTCTCCAAGGAC	CTGGCGCCGATACTCGTGGATCTC	AGGGAGATCCGCGCCCAGTATGAGT
ENSG00000169035	*KLK7*	Candidate	5'	CACCCAGCGCTCATTGACCAGGAC	GCACATGGGGCGCCATCAATAATCT	CCCTTCTCCTGCCCCTGCAGATCTTA
ENSG00000167916	*KRT24*	Candidate	5'	GTCATCAGCAGCCAATCTGGCATTGT	CCCAGCATTTTCAACAGTGGCAGCA	TGGAGACGGTGGATCGGGAAGAGAT
ENSG00000167656	*LY6D*	Candidate	5'	CGTGGTCTTGCAGAAGCGAGAGCTG	CGGGCAGACCACAGAATGCTTGC	CTCCTTGCAGCCCTGGCTGTGGCTA
ENSG00000124466	*LYPD3*	Candidate	5'	GGCGCGCACTTCACTGTCTTCATCT	TGCTTTCTGCACGCAGCTGTAGCACT	GAAAGCAGGTGCCCAGGCCATGA
ENSG00000009724	*MASP2*	Candidate	5'	ACTCCACTCGGCCACTGGGTAGATCA	TCATCAGGAGGGCCACAGTCAACAA	AAGATGGATCTTGGGACCGGCCAAT
ENSG00000124003	*MOGAT1*	Candidate	5'	CTTTCCGCTGGCGGATGAACAGAGT	TGCACCCCCAAGGACAATGACAGAG	CCCCCATGGAATAATGGCAGTTGGA
ENSG00000169550	*MUC15*	Candidate	5'	ATGGATCAAGGGAGGGCTTGTGGAA	TCCATGGCTCCCCGATAGAAGTGAA	TGTTGGCCTTAGCCAAAATTCTGTTGA
ENSG00000171840	*NINJ2*	Candidate	5'	CGATGACCACCTGCAGGAGCAGAGA	CTGATGAGGGTGACCAGGGTGGTGT	CCTGGAAGCTCCGACCCCAGGAG
ENSG00000196754	*S100A2*	Candidate	5'	TCTGGGCAGCCCTGGAAGAAGTCAT	CATCCAGGCTGCCCATCAGCTTCTT	CGCTGGCTGTGCTGGTCACTACCTT
ENSG00000112499	*SLC22A2*	Candidate	5'	TTTCGGCTTCCTCGATGGTCTCAGG	CAGCACCAGACCTCCAGCAACCAAG	GCCATTCCTGGTCTACCGGCTCACT
ENSG00000169507	*SLC38A11*	Candidate	5'	GGAGCAGATACCCTGGAAAGCCGAAA	CCCAAAGGAAACCCAGCTTGCTTCA	AATCCGTACCCAGCCCCAGCATCTT
ENSG00000132639	*SNAP25*	Candidate	5'	CCTTGTTCATCCAACATAACCAAAGTCC	GTTGCAGCATACGACGGGTGCTTTC	AGATGCAGCGAAGGGCTGACCAGTT
ENSG00000120669	*SOHLH2*	Candidate	5'	GTAGGACCTCTGCAGGGGCAATCCA	TCTTCAGGCCCAGTTGCGTTTTCAG	TCAGGGCATGGAATGGATATTGCTTT
ENSG00000169919	*GUSB*	cDNA_CTRL	N/A	TCTCTCGCAAAAGGAACGCTGCACT	N/A	TGCTAGAGCAGTACCATCTGGGTCTGGA
ENSG00000225292	RP11-57H14.3	Positive_CTRL	5'	TCTGAGTTCCCTGGAGATGCGGTGA	TTCCTGGCCAGCACTCTGGCTCATA	TGGCGTCAGCAGGAAAGATCAGCTC
ENSG00000148737	TCF7L2	Positive_CTRL	5'	TCCTAGCGGATGGGGGATTTGTCCT	GTGATAAGAGGCGTGAGGGGGTGGA	CCTCCCTTGCTGCACTCAGGGACAT
ENSG00000112299	VNN1	Positive_CTRL	5'	GGCTTCAGACTAAACAAGCGTCCGTCA	CTGGGTTCCGAAAGTGCCACTGAGG	TGCACACTGTGGAAGGGCGCTATTA

Table S6: RNA-seq data from The Cancer Genome Atlas, used for validation of fusion transcripts and splice junctions.

Barcode	Tumor Analysis Id	Pairs of sequencing reads	Normal Analysis Id	Pairs of sequencing reads
TCGA-AZ-6605	00fb8fe0-38e4-492f-aa94-2e121e1f2e91	55486229	47143116-4037-47f6-98a3-a0e55f922afd	58063148
TCGA-AZ-6603	0c62b6c5-1db7-49f9-9828-76dc103a3065	70729980	19e6bfe6-2f95-41c1-8b09-ae43f2a835dd	52846099
TCGA-AA-3655	1258626b-b0e5-466c-8197-c778cbc482c4	51719615	29c3c123-6520-4393-9035-dc4be3e0dfa4	55633014
TCGA-AZ-6598	156b0548-74d6-4bad-9bed-de16f471575e	48902613	747a946f-b12d-4076-9bb0-d9c84435e49f	67432558
TCGA-A6-2684	33f581db-84dc-40a9-a895-452c97cef7aa	99717188	71d55417-894f-4110-863a-e7426bd9ddbd	64873328
TCGA-A6-5667	38005f35-d336-4672-96db-87afca996ed4	56642649	c181130f-d3d8-453e-b160-0d723e51e54f	58639709
TCGA-AA-3662	383ff035-5d93-4a32-8ffa-49d5e5380181	57716339	b3e00f7a-fcd1-4413-941e-698181ebebfb	48553863
TCGA-AA-3697	3b26b219-772f-416a-8b86-a89c4ec35107	42161491	9df1484d-53b6-4818-8a65-d5e96b21939f	65781655
TCGA-A6-5662	3bd039ee-e395-42b1-9daf-6bea630453db	58797213	cebe69b0-8cd8-41f1-b6b9-c400ff24931b	51831724
TCGA-AA-3713	53c7b323-91f6-4407-b526-eb99e0b7ce44	65100348	ea9b0a79-f224-4c19-bfda-b7326185873c	50238563
TCGA-AA-3663	54f21772-c706-4c4e-af67-9f50a98e0482	47675488	81d38d2f-4450-43c9-abf3-5828469a16fd	48915108
TCGA-AA-3712	587773ba-a79a-4f00-afd3-b5368d0801c1	61135422	98c7246c-1c1e-42c7-b46a-5b4309f0417a	58387107
TCGA-A6-2675	62ad4eb1-df30-4c54-9c4c-770178ddfcf2	46976625	b99adacb-4869-4365-bced-03a956895832	52034724
TCGA-F4-6704	93270b30-7984-49f9-8f11-0847e36c00e9	105993154	10473b13-0bb1-46b9-90c6-4c48dd85f185	71228880
TCGA-AA-3489	955a9415-95a0-4456-9c46-87a57d84cef3	74644442	15b36e3b-6f6e-4c14-add2-9342cd7b040f	51087384
TCGA-A6-5659	97119d3c-54c3-42fa-b02e-77debda6db6a	58728680	49221035-c3cc-4164-b801-2e9da549c003	32477709
TCGA-AA-3511	990cfee2-e76d-41c7-a2de-60e6e4d861e4	65542491	dc84d72c-4555-48ec-9e42-747957182193	60962272
TCGA-A6-5665	9d175977-68a6-47a0-a907-e645744e4616	54894682	b65e0223-684f-43b4-9db5-8cc91d88f6a6	54269419
TCGA-AA-3496	a2ab3b6b-f8f0-48ec-8017-67cf4483edd9	66057024	31b1146e-3afa-4256-b4f4-fee5310de10e	37158143
TCGA-AA-3660	a6760a77-cb86-4213-8821-21a6a72f9d46	52884618	e80fce59-b1b3-4c32-bea0-a4e0df6c0a6b	45505737
TCGA-AZ-6600	af07eb30-0f05-4e58-80cf-ddc1d3ed0664	65093175	3125a583-c2de-4206-bb01-1542c38379e3	52183233
TCGA-AZ-6599	ce858bcb-6a50-4138-8e8a-41981cb4a421	53804797	bee9eed1-7449-4889-b3dc-98e68783b577	64579051
TCGA-AZ-6601	e97d0b3a-d101-44eb-b157-64b6d63c7e3f	58468332	15bdbacd-bf63-41be-9784-f2fa6392df91	58346263
TCGA-A6-2686	f1eaed0e-809f-4b0b-abc7-bce20facea24	62394237	88aea277-2c70-4ea8-9b98-9e308071a8f9	78991595
TCGA-A6-3810	01450db8-8b81-45c8-90a7-a571cff67922	87999234		
TCGA-G4-6320	01852cd7-f3f6-49c8-a9ce-59f16378db41	48909286		
TCGA-D5-6533	0712ef8b-cb13-461b-b5b9-e754ca821b22	52081176		
TCGA-AA-3526	07b48edc-8f77-42d0-a876-b742f1acedc5	58599278		
TCGA-G4-6302	0a2be50d-003f-4194-be56-ee3efd63ac71	50345345		
TCGA-CM-6167	0eb69c65-9839-4d47-90a5-90a887b1ef68	49027312		
TCGA-NH-A6GA	139a9ce5-fd10-4d22-9186-d6d4cce25af3	74258259		
TCGA-D5-5538	13d31511-789c-4bf6-9527-c05ccc59128b	53367485		
TCGA-NH-A50T	1562f565-a000-437c-a9c8-f058639023b4	63656977		
TCGA-CA-5255	15f1af0d-f1fe-4049-b1ab-9595050c0faa	52204337		
TCGA-DM-A1DA	17a14091-a6cd-4bfc-aefd-451813d0adc0	84145812		
TCGA-AZ-6606	1a751365-2fc5-4388-868e-380c43c4c6cf	61895177		
TCGA-A6-6137	1dc6366b-24d1-4aa1-af1d-43d5c0050762	56864749		
TCGA-A6-3809	1e7486af-f1dd-42e2-a9cc-6283714f2fce	81721978		
TCGA-AY-5543	1f5970e9-e134-45d7-b1de-c098b806a9ef	70877135		
TCGA-D5-6539	20de162c-d261-4e27-8ce8-6e90ffe01c54	54365134		
TCGA-CM-5863	259394bc-1129-40e7-865b-e1aaade37a58	59990303		
TCGA-G4-6625	25a0f263-40fa-4027-8389-cdb03e4ded57	56697130		
TCGA-F4-6463	2a9ba30f-1564-457d-b4a1-c85923a1c858	51784556		
TCGA-F4-6808	30f2ef1a-0dbc-4e48-a341-2ee82ee20bf7	44451025		
TCGA-CK-5913	325f837c-58a6-4e61-8df5-90767d5747ae	62378561		
TCGA-AD-6965	393db15f-9f5f-4ddf-a035-d63d993b3106	45988934		
TCGA-DM-A28C	3c65aeda-3597-4cad-8ab7-b8b66534a6c8	67816052		
TCGA-CK-4947	3c91fc88-e835-41c8-b210-04dfdcbeaf28	59492848		
TCGA-CM-6168	41a92221-81b3-414b-aa2e-63c2070d1401	59309838		
TCGA-F4-6856	42b408ac-0aea-45db-ae3f-033f3d3448a3	34509686		
TCGA-RU-A8FL	43aa202c-004f-4292-9c98-8a2e87e05dc4	60139908		
TCGA-A6-6781	453ed81e-2f7c-41d1-b0a0-e6b2e3d6735b	150674345		
TCGA-CM-6676	48a058cb-5520-4bd2-a588-ac71301646b7	58055247		
TCGA-A6-6141	49dcf9dd-a058-47de-bf9e-9e90f4cff877	46529119		

Barcode	Tumor Analysis Id	Pairs of sequencing reads	Normal Analysis Id	Pairs of sequencing reads
TCGA-G4-6310	49f20084-c3f5-40a3-a2a7-71a527608606	35059072		
TCGA-DM-A1D6	4cfe69b5-0e9e-4b45-8ac1-0f481e6f4c6c	103556678		
TCGA-A6-6651	53b37cdd-358f-450c-8293-309c2c9cb56a	68082194		
TCGA-CM-5349	543b938f-e8a8-4972-968c-ddcd3f3151c6	48356111		
TCGA-D5-6922	55b73313-fd5b-4a1a-8b64-6be4446bf457	66350024		
TCGA-G4-6315	591cd1bb-772b-4595-97c7-0dd4d5749dd2	47519287		
TCGA-A6-5656	60defced-1166-460c-8e8c-b6653e51715a	101987746		
TCGA-AD-6899	6438a93a-8807-4bd4-82fc-badaec9f36c0	72843788		
TCGA-A6-6138	64a194fe-03d6-4eba-a44a-97b779640645	55452132		
TCGA-AD-6889	687e691d-0545-4408-a267-412357506a51	56648348		
TCGA-AY-A54L	68a0aeca-6541-40cb-8777-159658573e65	64160497		
TCGA-CK-6747	6a040134-0ae6-458b-be76-655bc0a62114	71828759		
TCGA-CM-5864	6dc3e8b2-26e5-4757-9d80-53a13976245d	41687528		
TCGA-NH-A50U	7218bddc-c3a6-48bf-8875-28476e5e2451	84592874		
TCGA-D5-5539	727527dd-a98e-4953-bd2a-436127479796	61666427		
TCGA-F4-6569	7607bfb5-e253-42ab-a857-3f342b217d13	65347476		
TCGA-A6-A566	76775244-8b7e-4ad8-b02a-831d24df2928	43890062		
TCGA-F4-6570	772e5153-a09d-4dcd-a2c7-a50ef03fa936	59121343		
TCGA-G4-6299	7b2c8113-8006-4fc7-a3e3-8065d9f4854b	54741505		
TCGA-AZ-4682	7c3c572e-e822-455a-9c57-eed7d8f6eceb	71094823		
TCGA-CA-6716	807da051-49ab-4307-9d41-5ec2104af9cb	60628668		
TCGA-CK-4950	81cb39c2-5ba4-478a-bfbd-4e026a4d657b	84151814		
TCGA-AM-5820	86d78b6b-710b-4804-970e-672b761d93f4	57071530		
TCGA-CA-5796	87014eff-96d3-4c69-ae51-7cb0159117e6	39122904		
TCGA-NH-A50V	88c66027-f9bb-43f3-af2a-37acff4c3c03	65410005		
TCGA-DM-A288	8a930fdc-b6bc-4cd0-a3c6-e77f0ea44946	48990434		
TCGA-F4-6855	8bdac7ae-da03-4533-a277-8d9b3f19e245	42174910		
TCGA-D5-5540	8cd19b31-de1f-44d1-b361-f2d9452b6e2e	49056195		
TCGA-T9-A92H	8cee552d-6908-478e-ba62-93187e71003c	63196329		
TCGA-A6-A565	8ebd29b8-06ae-451a-a5e3-6a764c9e55f4	81477212		
TCGA-D5-6541	8ecfcb26-89b1-4ef8-85e6-710dfe6835da	81713291		
TCGA-G4-6307	967ac7cd-a4f1-4426-a49b-1d29a621e62b	90957923		
TCGA-G4-6306	9a64ad22-2466-4042-b539-15d194f14446	37606423		
TCGA-AD-6890	9b736c02-0193-48fb-aab0-d1c6f2bf4f90	44415017		
TCGA-A6-2677	9c630191-e339-41ff-8b74-d8102d433f02	109602293		
TCGA-DM-A0XD	a0361fa0-dade-4885-8647-71f4c3397a4e	73542403		
TCGA-A6-6780	a248a7c8-353a-4890-8453-0398dae660e1	72542996		
TCGA-AD-5900	a7617ecb-2d15-4a68-b5b3-48b6b69aa563	50348627		
TCGA-DM-A28H	a7780211-a135-4d23-b81f-5beb8fabe234	68519169		
TCGA-AZ-5407	ab796e6a-b246-427f-8a43-76e8b6784c4e	45425942		
TCGA-D5-6926	abd0fc1d-48b4-4fb6-aeaa-1277bc6f72d0	57746537		
TCGA-CK-5915	ac0d5d31-ba19-44d7-821b-dc2b84ac1a51	51700463		
TCGA-AZ-4616	aca1afda-ffa3-4d98-82a8-88c7d59c7ab9	59252537		
TCGA-AA-A02K	ade9cb51-d46c-47cc-8992-2f282ca9fbc2	69205871		
TCGA-CM-6679	ae0f3660-d820-435d-b030-3cdd70433709	67109244		
TCGA-D5-5537	b2cd3c40-ca79-4939-bcb2-803a88a37d20	52334890		
TCGA-QG-A5YV	b3e63ab6-760c-4e1b-9f14-71b31e73530c	63090685		
TCGA-D5-6924	b7550578-77eb-4329-8bd6-fa6e66d3b836	38453173		
TCGA-DM-A28K	b7b22f88-9dee-4fc9-a547-873ed286bd9a	56852825		
TCGA-F4-6461	bac248f6-35cf-4105-8cf1-ecd312dce2e7	50886538		
TCGA-DM-A28M	bb8568b4-f52f-4979-9be1-66be301d3ddd	78422255		
TCGA-D5-6536	c9bc8c42-c0c8-44ce-a3a6-65c36d0ca071	51480916		
TCGA-CM-4744	d0608c54-f725-41b7-9f40-2213e1d761d7	78962811		
TCGA-A6-3809	d18f1c43-d818-48af-9389-1afbcf43356e	74146836		
TCGA-G4-6311	d247d359-5aa0-46ae-b0ba-a17f7ce06fd6	47629841		
TCGA-G4-6627	d504efae-2105-4e13-8a62-12cc1d204aae	54189450		
TCGA-G4-6295	d70fad4c-8f6f-4a07-ade8-982cf4e489c8	48773554		

Barcode	Tumor Analysis Id	Pairs of sequencing reads	Normal Analysis Id	Pairs of sequencing reads
TCGA-D5-6530	d71ce075-aa5e-4a59-b77c-96a9b5f472a3	53730637		
TCGA-AY-A71X	d7a74f28-cff5-4285-9381-017e91430950	76435054		
TCGA-F4-6854	d7ab1cc2-6582-40da-a845-92c074f0ae2e	71010710		
TCGA-CM-6674	dceaf1f2-d773-45a2-b23d-8ebf55ceb369	45860603		
TCGA-A6-6653	de9341f3-c957-405d-9b36-a9cf0e901096	51684966		
TCGA-DM-A28E	e32fd179-5ca2-4801-b123-80d75455a9a4	62666313		
TCGA-A6-6650	e7241925-40af-4aec-8afc-6405eb435671	63488443		
TCGA-NH-A5IV	ec6b3979-de3c-4047-ad54-f5b5a7c2c67c	68030709		
TCGA-AZ-6607	ef6ae243-ed5d-4e44-95ad-387504275524	46346736		
TCGA-D5-6529	efe76f19-a282-41ac-ac73-23499ec2f88b	65660651		
TCGA-4N-A93T	efed4ca9-32ee-49fa-9373-f7a91601a841	65163355		
TCGA-D5-6535	f0118de9-5679-47dc-a5f8-794bd5544472	53080259		
TCGA-AU-3779	f352289e-f836-4f61-aa45-dc9b9ab8cf96	52481791		
TCGA-CM-6172	f42ad7bf-a199-4419-a2c4-2d0d97a3eeda	53690814		
TCGA-CK-4948	f7c407c3-f840-4c5d-acea-efc8c6d99c21	78181183		

Table S7: RNA-seq data from the Cancer Cell Line Encyclopedia used for validation of fusion transcripts and splice junctions.

Cell line	Barcode	Analysis Id	Pairs of sequencing reads
NCI-H508	CCLE-NCI-H508-RNA-08	0c7a79cc-bbaf-4c6d-93e1-866f5f5f3d0d	81974061
SNU-81	CCLE-SNU-81-RNA-08	0e30aa7c-8411-4108-8dd0-226288b32327	40670713
HCT-15	CCLE-HCT-15-RNA-08	127d1510-a54b-4a55-b62d-741fd11fa781	79236767
HT-29	CCLE-HT-29-RNA-08	13442d3a-26b1-41b0-b45a-94b2576b2aab	51321817
SNU-61	CCLE-SNU-61-RNA-08	187c4576-27cf-46ab-8d8a-e882d547df3e	90231671
LoVo	CCLE-LoVo-RNA-08	188d620d-1838-43d1-a278-3b5d7decf91d	94792165
RKO	CCLE-RKO-RNA-08	19f57bd9-e629-4ca6-970e-a34ab31001c0	65043608
HCC-56	CCLE-HCC-56-RNA-08	1d95e7cc-a3e0-4770-9e8a-163e154a051f	83139954
SW837	CCLE-SW837-RNA-08	230c98ac-9f16-4a52-8311-a7434bfda57f	43521953
Hs 698.T	CCLE-Hs 698.T-RNA-08	27835524-c72a-48e6-8121-6448bb11219e	69093386
SK-CO-1	CCLE-SK-CO-1-RNA-08	29280869-87f6-408a-af86-94375a36d7aa	73747960
SNU-C5	CCLE-SNU-C5-RNA-08	2dd2783e-29e7-46fd-9d13-3fd0e8ff15d6	42230353
RCM-1	CCLE-RCM-1-RNA-08	315a776d-aee2-4960-b6ed-bda8e4395eb1	87974359
SNU-1033	CCLE-SNU-1033-RNA-08	376d30aa-1739-4932-b552-5172aa89b785	89278008
LS123	CCLE-LS123-RNA-08	3ab53c52-8a97-4600-913c-99b15f5d71fa	62275140
SW948	CCLE-SW948-RNA-08	495f6fed-34fb-4e8a-8b4c-abd9653790a6	67146059
SNU-1040	CCLE-SNU-1040-RNA-08	4e188ecc-db0b-4fdf-a19d-1ee952fffcf6	94960459
MDST8	CCLE-MDST8-RNA-08	55bfe3c2-ca8f-4173-b379-20afec0525df	66633712
SW1116	CCLE-SW1116-RNA-08	57e8df2b-7ba5-4395-b21d-6723abfb802f	47208882
CL-34	CCLE-CL-34-RNA-08	5ad6322b-51bb-434c-a1d5-2e52d5c27bbf	92964046
NCI-H747	CCLE-NCI-H747-RNA-08	66825466-fc65-4847-89b6-9a4f0e1f875d	68328206
COLO-678	CCLE-COLO-678-RNA-08	6ca39c8b-95e1-4603-bb03-b3ad33c542e8	90101894
SNU-175	CCLE-SNU-175-RNA-08	6d4c972d-f702-4193-9251-93cceb6a22bf	78392082
LS513	CCLE-LS513-RNA-08	71abf291-ae60-4668-8786-bf1327df0616	83382388
OUMS-23	CCLE-OUMS-23-RNA-08	75b57b0c-df5b-429b-a024-3777b6291bf0	93938459
SNU-C2A	CCLE-SNU-C2A-RNA-08	7c1f331a-a9fd-4d84-a41d-b87179f8bd06	76807492
Hs 255.T	CCLE-Hs 255.T-RNA-08	7dd73e3d-e30a-49d7-8137-4569cd2dfb9f	85354793
SW620	CCLE-SW620-RNA-08	7fd7ebc4-473e-420f-accd-8b2732c6a2ca	81665453
SNU-C4	CCLE-SNU-C4-RNA-08	8752a43c-6320-497e-80ae-1a3e482aabc2	74960644
SW48	CCLE-SW48-RNA-08	88089c91-2054-4fb9-9bdd-624443d141bf	37642250
COLO 741	CCLE-COLO 741-RNA-08	8b2a3d47-9518-43dd-9074-1d48f51a40ad	79950665
HCT 116	CCLE-HCT 116-RNA-08	8f17979c-77dd-4743-bafa-a436fce0d2fd	78820314
T84	CCLE-T84-RNA-08	925b98c1-40ab-4050-a17d-cf56200818a6	127827846
NCI-H716	CCLE-NCI-H716-RNA-08	95e1c652-a7f2-41f6-9a30-503b8c7c37a5	76694776
CL-14	CCLE-CL-14-RNA-08	984cd7e5-5e8b-4498-9899-22310069d7d1	66660769
SW1463	CCLE-SW1463-RNA-08	9a98973b-40f4-4fca-960d-f2d1878a4601	32724967
GP2d	CCLE-GP2d-RNA-08	afa83f66-1b9e-440f-8a18-7c1f34a1411a	83640184
SNU-503	CCLE-SNU-503-RNA-08	b0d39203-22f1-4f20-8748-54439652d876	77379971
C2BBe1	CCLE-C2BBe1-RNA-08	b39b60cd-ed66-4824-9548-6e1396da753c	81888642
SW480	CCLE-SW480-RNA-08	b944955c-0c3c-41a9-8f17-e4c484fe24cf	74971267
LS1034	CCLE-LS1034-RNA-08	bac80f8b-66ce-436a-bc37-2a6bb05884aa	70748555
SW1417	CCLE-SW1417-RNA-08	bbd3b9f5-7e8f-41eb-a3fe-135b7acd35bd	57064664
SNU-1197	CCLE-SNU-1197-RNA-08	bbdfc980-47fe-4b4c-892f-6daddac199a6	90611970
KM12	CCLE-KM12-RNA-08	bde66db2-40f5-4a47-92b4-151050895755	87680166
Hs 675.T	CCLE-Hs 675.T-RNA-08	c602930d-a4bf-4acb-8c64-db73f99bd26d	90492603
LS 180	CCLE-LS 180-RNA-08	ca646e08-060f-402c-aaa0-abfdcdf3b4a4	71419500
HT115	CCLE-HT115-RNA-08	cbfd3584-ca62-49cd-ac0c-421773331ce7	154416450
COLO 201	CCLE-COLO 201-RNA-08	cff7ddf1-d404-45e2-9bac-ed18b9638a56	81798161
LS411N	CCLE-LS411N-RNA-08	d0cb3f54-1fcf-4ca6-8119-3b0737c8a69c	77680067
SNU-C1	CCLE-SNU-C1-RNA-08	d62c141e-ffb9-494a-81d8-dd601412bcc2	76858944
CW-2	CCLE-CW-2-RNA-08	dfc229b0-a48d-48e4-8d16-420cfa02580f	81363514
SW403	CCLE-SW403-RNA-08	dfcd7b32-4b80-456b-b20c-98761581d73e	42857669
SNU-407	CCLE-SNU-407-RNA-08	e67c993c-c98f-4089-8d36-b5c37ad727eb	82230291
CL-40	CCLE-CL-40-RNA-08	e6b5d8f8-76ac-4598-954a-aadbf4306afa	88499234
CL-11	CCLE-CL-11-RNA-08	e8659530-7b1d-43fd-a080-6a87ff015aa6	80908345
COLO-320	CCLE-COLO-320-RNA-08	ecfd1349-f493-4f96-b3a1-7b27cc434640	84061201

Additional file 3: Figure S2

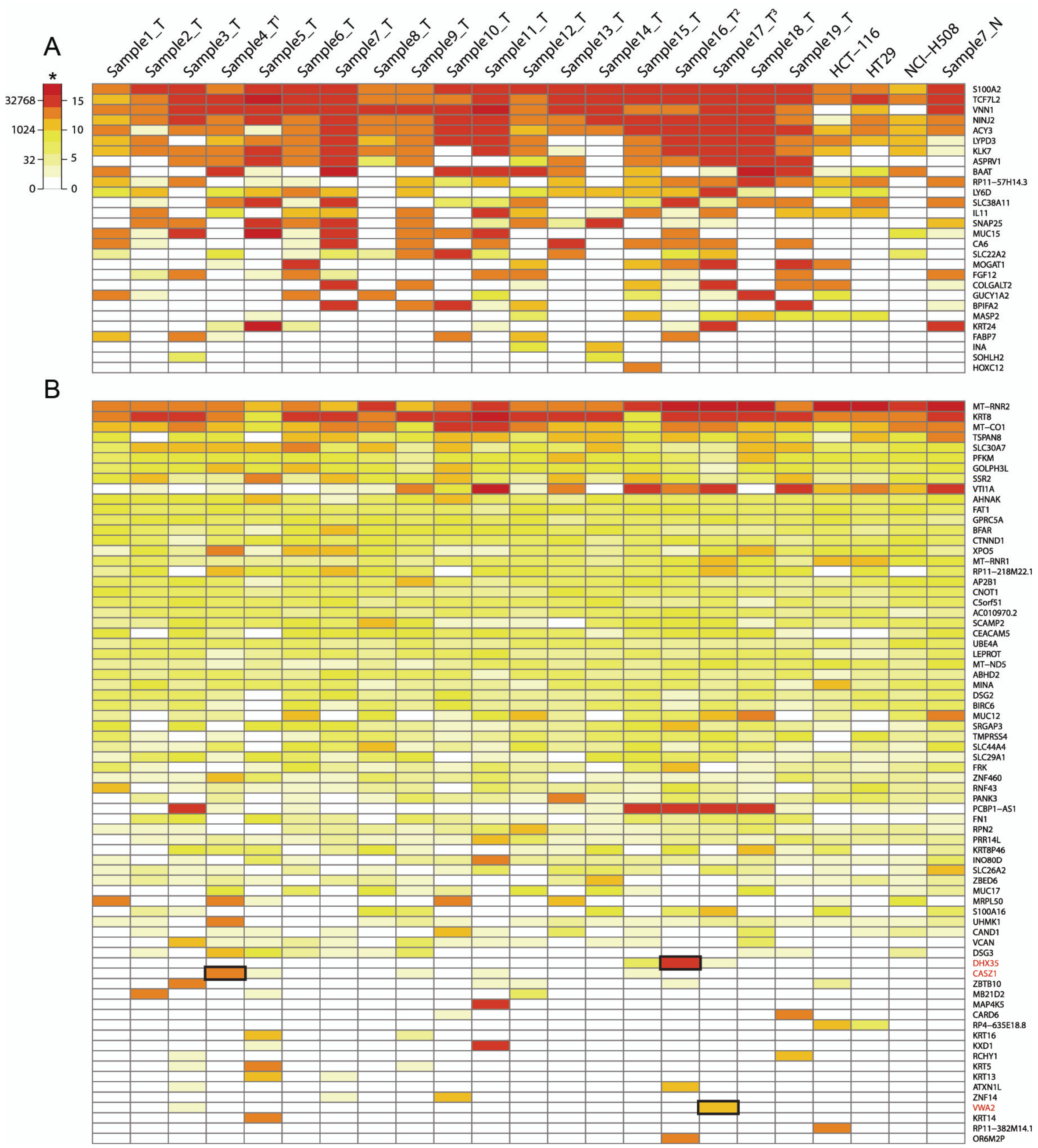


Figure S2: Heat map showing $\log_2$ values of normalized read counts of the top 100 covered genes from the RACE-seq experiment. *Heat colors represents log2 values of normalized read counts. Corresponding read count values are also indicated. [1]Sample4_T was the index sample with elevated expression at the 3' end of *MASP2*, as seen from the exon microarray data. [2]Sample16_T was the index sample with elevated expression at the 3' end of *BPIFA2*, as seen from the exon microarray data. [3]A fusion between *VWA2* and *TCF7L2* was identified by defuse in sample17_T. The heat map is divided into **A:** Candidate genes targeted by RACE amplification, and **B:** top non-target genes. Upstream partner genes of the identified fusion transcripts *VWA2-TCF7L2, DHX35-BPIFA2* and *CASZ1-MASP2* are indicated with red font. Black boxes indicate the normalized read counts values in the samples expressing the identified fusion transcripts. For *DHX35* and *CASZ1*, the samples expressing the fusion transcripts were also the samples identified to have elevated 3' expression as seen from the exon microarray data.

Additional file 4: Figure S3

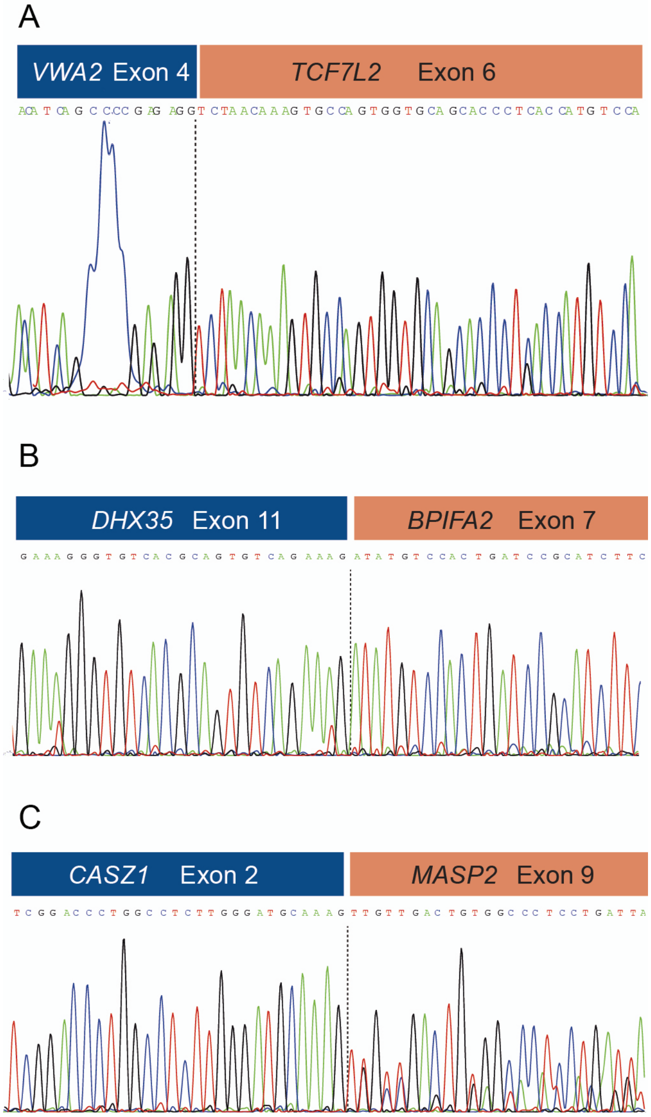


Figure S3: Electropherograms from Sanger sequencing showing sequences covering intact exon to exon breakpoint boundaries for **A:** *VWA2-TCF7L2*, **B:** *DHX35-BPIFA2* and **C:** *CASZ1-MASP2*.

Additional file 5: Figure S4

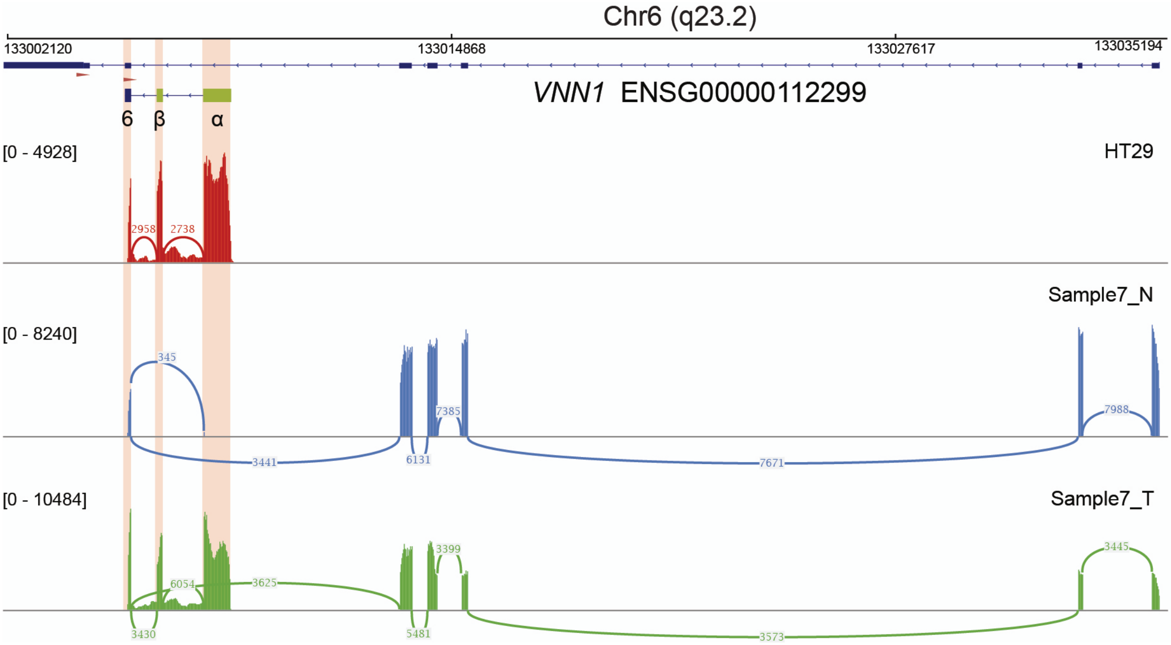


Figure S4: Sashimi plot that shows the RACE-seq read coverage of the *VNN1* gene in the HT29 CRC cell line and the tumor/normal pair included in the experimental set up. Canonical UCSC gene annotation is shown on the top track, with red arrows indicating the location of the *VNN1* RACE assay. The alternative start exons α and β of the recently described transcript [22] are shown in green. Bars show coverage values at genomic locations, while arcs depict splicing junctions. The numbers of reads crossing the splicing junctions are annotated on the arcs, here determined by Tophat2 alignment and the sashimi plot package in IGV.

Paper III

RNA sequencing reveals fusion genes in testicular germ cell tumors

Andreas M. Hoff, Sharmini Alagaratnam, Sen Zhao, Jarle Bruun, Peter W. Andrews, Ragnhild A. Lothe, Rolf I. Skotheim

Manuscript

RNA sequencing reveals fusion genes in testicular germ cell tumors

Andreas M. Hoff[1,2], Sharmini Alagaratnam[1,2], Sen Zhao[1,2], Jarle Bruun[1 ,2], Peter W. Andrews[3,4], Ragnhild A. Lothe[1,2], Rolf I. Skotheim[1,2†]

[1]Department of Molecular Oncology, Institute for Cancer Research, Oslo University Hospital-Norwegian Radium Hospital, Oslo, Norway
[2]Centre for Cancer Biomedicine, University of Oslo, Oslo, Norway
[3]Department of Biomedical Science, University of Sheffield, Western Bank, Sheffield, United Kingdom.
[4]Centre for Stem Cell Biology, University of Sheffield, Western Bank, Sheffield, United Kingdom

[†]**Corresponding author:** Rolf I. Skotheim, Department of Molecular Oncology, Institute for Cancer Research, Oslo University Hospital-Norwegian Radium Hospital, Ullernschausseen 70, 0310 Oslo, Norway. E-mail address: rolf.i.skotheim@rr-research.no, phone: +47 2278 1727

Financial support

The study was funded by grants from the Norwegian Cancer Society (AMH was financed as PhD student from the PR-2007-0166 grant to RIS), NorStore (for storage of computer files; project NS9013K), the Research Council of Norway through its Centers of Excellence funding scheme (project number 179571), and from the KG Jebsen Foundation.

Conflict of interest

The author(s) declare that they have no conflict of interest.

Keywords

TGCT, RNA-seq, ddPCR, fusion genes, stem cells

Abstract

To elucidate the existence of malignancy specific fusion genes and aberrant fusion transcript expression in testicular germ cell tumors (TGCT), we performed RNA sequencing of embryonal carcinoma (EC) and embryonic stem cell lines, and used the latter as a non-malignant pluripotent comparison. By applying the fusion detection algorithms deFuse and SOAPfuse and a filtering pipeline, we identified eight novel fusion transcripts and one gene with alternative promoter usage, *ETV6*. Four of these nine transcripts were found recurrently expressed in an extended panel of primary TGCTs and additional EC cell lines. *RCC1* was found to form fusion transcripts with *HENMT1* and *ABHD12B,* located 80 Mbp downstream on chromosome 1 and on chromosome 14, respectively. *RCC1-ABHD12B* and the transcript variant of *ETV6* using an alternative promoter were found to be preferentially expressed in the more undifferentiated TGCT subtypes. *In vitro*-induced differentiation of the NTERA2 cell line resulted in significantly reduced expression of both fusion transcripts involving *RCC1* and the *ETV6* transcript variant*,* indicating that they are markers for pluripotency in a malignant setting. None of the four recurrent transcripts were expressed in normal parenchyma of the testis, implying malignancy specific expression. By droplet digital PCR linkage analysis we found that the private fusion genes *EPT1-GUCY1A3* and *PPP6R3-DPP3* were fused on the DNA-level in the 833KE and NTERA2 EC cell lines respectively. In conclusion, we combined RNA sequencing with a defined filtering strategy to identify eight novel fusion transcripts, to our knowledge the first fusion genes described in TGCT.

Introduction

Testicular germ cell tumors (TGCTs) are the most common cancer in young males aged 15-44 years [1]. Although it is a highly treatable cancer type, exemplified by a 10-year net survival rate of 98 % in England and Wales [2], the disease affects men in their prime and treatment can lead to substantially increased morbidity, including cardiovascular disease, reduced fertility and secondary cancers [3]. Histologically, there are two main subtypes of TGCTs, seminomas and non-seminomas. Both are thought to develop from the pre-invasive stage termed intratubular germ cell neoplasia (IGCN; also known as carcinoma *in situ).* Non-seminomas are further divided into the pluripotent embryonal carcinomas (EC) and more differentiated subtypes, with either somatic (teratoma) or extra-embryonic differentiation (yolk sac tumors, YST, and choriocarcinomas)[4].

EC cells are highly similar to embryonic stem (ES) cells, derived from the inner cell mass of the blastocyst stage embryo [5]. Both cell types exhibit pluripotent characteristics phenotypically and in gene expression profiles [6,7]. Upon extended passaging *in vitro*, ES cells have been shown to acquire genetic changes similar to those seen in malignant transformation *in vivo* of TGCTs and EC, including gain of genetic material from chromosomes 12, 17, and X [8]. Gain of chromosome arm 12p, often as an isochromosome, i(12p), is found in virtually all cases of TGCT [9,10]. Crucially, despite these similarities, EC cells are malignant in character, whereas ES cells are not. Comparative studies between the two cell types may therefore be useful for characterization of cancer-specific differences in a pluripotent context [5,11]. One such study revealed that several transcription factors located on 12p are overexpressed in EC cells as compared to ES cells [6]. Although 12p material is gained in virtually all cases of TGCT, no clear genetic driver for TGCT malignant transformation has been pinpointed [12,13].

Recently, whole-exome sequencing studies have revealed that the number of non-synonymous mutations in coding regions of the TGCT genome are few, on a scale similar to that of pediatric cancers [14–16]. A number of pediatric cancers with a low mutational load are frequently found to harbor fusion genes with oncogenic properties. Examples are *MLL* rearrangements in acute lymphoblastic leukemia [17], and subtypes of sarcomas classified by distinct chromosomal translocations [18]. In Ewing sarcoma, fusions involving *EWSR1* are pathognomonic, while the mutation rate is low, estimated at 0.15/Mb of coding sequence [19]. In this study, we have performed RNA sequencing of EC cell lines and their non-malignant counterpart, ES cell lines. Application of a fusion gene analysis pipeline led to the identification of nine novel fusion genes and transcripts, to our knowledge the first described in TGCT.

Material and Methods

Cell lines and patient samples

Three EC cell lines (2102Ep, 833KE, and NTERA2) and 2 ES cell lines (H9 and Shef3) were subjected to RNA sequencing. All EC and ES cell lines were sorted for expression of the pluripotency marker SSEA3 as previously described [11]. The extended experimental validation panel consisted of four categories of samples: 1) 2 additional EC (Tera 1 and NCCIT) and 2 additional ES (Shef6 and Shef7) cell lines (n=4), 2) NTERA2 and 2102Ep cells treated with all-*trans* retinoic acid (RA) for 0, 3 and 7 days to induce differentiation, as previously described (n=6) [20,21]. 3) Thirty-five testicular tissue samples including 5 normal testicular parenchyma, 6 premalignant IGCN and 24 primary TGCTs, all with only one histological subtype each; EC (n=8), seminoma (n=7), choriocarcinoma (n=1), YST (n=4), and teratoma (n=4). 4) Twenty normal tissues from miscellaneous sites of the body were used for exploration of cancer-specificity of the novel transcripts (adipose, bladder, brain, cervix, colon, esophagus, heart, kidney, liver, lung, ovary, placenta, prostate, skeletal muscle, spleen, stomach, testes, thymus, thyroid and trachea; FirstChoice Human Normal Tissue Total RNA). These were each a pool of RNA from at least three individuals, with the exception of one individual sample from the stomach (Ambion, Applied Biosystems by Life Technologies, Carlsbad, CA, USA).

DNA isolated from cell pellets was STR fingerprinted using the AmpFLSTR Identifiler PCR Amplification Kit (Applied Biosystems). Profiles positively matched with those reported in the literature for 2102Ep [7], and obtained from EACC/HPACC (for 833KE), ATCC (for NTERA2, NCCIT, and TERA1), the Wisconsin International Stem Cell Bank (H9), and the UK Stem Cell Bank (Shef3, Shef6, and Shef7). The biobank is registered according to Norwegian legislation (no. 953; Biobank Registry of Norway) and the project has been approved by the National Committee for Medical and Health Research Ethics (S-05368 and S-07453b).

External data for *in silico* validation

Paired-end RNA sequencing data from the Illumina Human Body Map v2 dataset, consisting of 16 non-malignant miscellaneous tissue types, was analyzed as an additional source of normal controls (ArrayExpress accession ID E-MTAB-513 and European Nucleotide Archive study accession ID ERP000546).

Paired end RNA-sequencing of EC and ES cell lines

Library construction was performed using the standard Illumina mRNA library preparation protocol (Illumina Inc, San Diego, CA, USA), including poly-A mRNA isolation, fragmentation, and gel-based size selection. Shearing to about 250 bp fragments was achieved using the Covaris S2 focused ultrasonicator (Covaris Inc, Woburn, MA, USA). Paired-end sequencing, 76 bp from each end, was performed according to protocol on a Genome Analyzer IIx (Illumina Inc.).

Fusion transcript identification

To identify fusion transcripts specific for EC, we used the fusion detection algorithm deFuse v. 0.6.1 [22] with hg19 sequence reference from UCSC and Ensembl release 69 annotation. To enrich for true positive fusion transcripts specifically expressed in EC cells, several heuristic filtering steps of the initial fusion breakpoint candidates were performed, some adapted from the recommended procedures of the original publication of deFuse [22], namely: 1) Only the nominated fusion breakpoint candidates with a probability score greater than 0.5 were considered. 2) Breakpoints nominated in EC that were also found in ES cell lines or tissues of the Human Body Map v2 were removed to enrich for malignancy-specific candidates and remove systematic technical artifacts. 3) We removed candidates that had more than 5 multi-mapping spanning reads, or a ratio of multi-mapping spanning reads greater than 25 %. 4) To filter out candidates nominated due to homologous or repeat sequences, we removed candidates that had a deFuse homology score greater than 10 and candidates having either of the following three criteria: cDNA adjusted-, genome adjusted- or EST adjusted percent identity greater than 0.1. In an effort to enrich for functionally interesting fusion candidates, we applied 4 functional filters of which candidates needed to pass 3 in order to proceed. 1) Both gene partners are annotated as protein-coding genes. 2) The number of split reads and spanning reads supporting the fusion breakpoint sequence is greater than or equal to 3 or 5, respectively. 3) The fusion breakpoint includes 5’ UTR or coding parts of at least one of the partner genes. 4) The distance between the two partner genes is greater than 30 kb. We also used an additional fusion finder algorithm, SOAPfuse v.1.26 [23], and included all fusion breakpoints that were picked up by both deFuse and SOAPfuse independently. For the remaining fusion transcripts, breakpoint alignments were evaluated in the UCSC genome browser and the Integrative Genomic Viewer (IGV). In general, we removed chimeric breakpoint sequences likely to derive from read-through transcripts, and those not aligning to conserved exon to exon boundaries.

Validation of fusion transcript breakpoints by reverse-transcriptase PCR and Sanger sequencing

Selected fusion transcript candidates were validated with reverse transcription PCR (RT-PCR) in the RNA-sequenced cell lines and in an extended validation panel. Primers were designed to the fusion transcript breakpoint sequences as detected by deFuse by using the Primer3 web application [24]. All primer sequences used in this study are listed in Additional file 1, Table S1. Briefly, reverse transcription was performed using the high-capacity reverse transcription kit according to protocol (Applied Biosystems by Life Technologies, CA, USA). From 50 ng of starting cDNA template, a PCR protocol was initiated with 15 minutes of HotStarTaq DNA polymerase activation at 95°C, followed by 30 thermal cycles of denaturation for 30 seconds at 95°C, primer annealing for 1 minute at optimal primer melting temperatures (Additional file 1, Table S1), and extension for one minute at 72°C. After the last cycle, a final extension step was performed at 72°C for 10 minutes. The PCR products were separated by electrophoresis at 200 V for 30 minutes on a 2 % agarose gel and visualized using ethidium bromide and UV light.

To ensure specific amplification of the breakpoint sequences, PCR products from the cell lines that were nominated by RNA-seq to harbor the individual fusion transcripts were sequenced by Sanger sequencing. PCR products that showed a single nucleotide band on the agarose gel were sequenced directly from both sides using forward and reverse primers. Prior to sequencing, the PCR products were purified using Illustra ExoStar 1-step cleanup (GE Healthcare, Little Chalfront, UK). The cycle sequencing reactions were performed using the BigDye Terminator v.1.1 cycle sequencing kit (Applied Biosystems, Foster City, CA, USA) following manufacturer's recommendations. The sequencing products were purified using BigDye Xterminator (Applied Biosystems) before being analyzed by capillary electrophoresis using the ABI 3730 DNA Analyzer (Applied Biosystems). The resulting sequences were analyzed using the Sequencing Analysis v.5.3.1 software.

The quantity of fusion transcripts assessed by TaqMan real-time PCR

Several fusion transcripts confirmed by regular RT-PCR were recurrent, however with varying nucleotide band intensities between samples, as observed by agarose gel electrophoresis. We performed TaqMan quantitative RT-PCR (qRT-PCR) to quantify the relative expression of these fusion transcripts. Primers and MGB-probes were designed with the Primer Express v.3.0 software (Applied Biosystems) to cross the fusion transcript boundaries (Additional file 1, Table S1). Two endogenous control assays targeting *ACTB* and *GUSB* were analyzed in all samples to normalize for input template amounts. The qRT-PCR reactions were performed in reaction volumes of 10 µl, with 15 ng of template cDNA, TaqMan universal mastermix II with uracil-N-glycosylase (Applied Biosystems) and final primer and probe concentrations of 0.9 µM and 0.2 µM, respectively. The PCR reactions were run in triplicate on an ABI 7900HT fast real-time PCR system (Applied Biosystems). Expression levels were reported as the median cycle threshold (C_T) of the triplicates and normalized to median C_T values of the endogenous controls. A threshold value at $C_T = 35$ was set for all assays as positive expression.

Assessment of DNA-level fusions with multiplexed droplet digital PCR

Droplet digital PCR (ddPCR) takes advantage of oil/water emulsion, separating PCR reagents and template into thousands of nano-liter sized droplets. Subsequent thermal cycling by traditional fluorescent PCR specifically amplifies target templates in the droplets. The number of target molecules in a reaction mixture is inferred by counting the number of droplets with and without amplified fluorescent signal. It is also possible to measure two target molecules simultaneously, by multiplexing 2 PCR assays with different fluorescent dyes (FAM and VIC/HEX). Since template molecules distribute randomly into droplets, droplets can be expected to contain one, the other, both or none of the target molecules by chance in a multiplexed assay. However, if two template targets are located in close proximity on the same DNA molecule and thereby linked, these would distribute together in a non-random fashion with a higher number of double positive droplets than expected by chance.

Here, we performed ddPCR to investigate DNA-level linkage of the partner genes of 2 recurrent fusion transcripts, as well as 2 fusion transcripts that each were expressed only in one EC cell line. As proof of concept, we included duplex linkage assays for the known fusion *VTI1A-TCF7L2*, which in the NCI-H508 cell line is known to be formed by a genomic deletion [25]. To evaluate the integrity of DNA fragments, and as an additional positive control of the ddPCR linkage approach, we used a custom milepost experiment which measures linkage with assays 1 kb, 10 kb, 50 kb and 100 kb apart. In all experiments, a reference assay with FAM fluorescence was multiplexed with one of the milepost assays with VIC fluorescence. FAM and VIC assays were also designed for the two partner genes of the fusion transcripts. All assays used in the ddPCR linkage experiments are listed in Additional file 1, Table S2. As control experiments for the linkage assays we performed fragmentation of genomic DNA with the NspI restriction endonuclease. We ensured that none of the assay's target sequences overlapped with the restriction enzyme target sequence. Each ddPCR experiment was carried out in 22 µl reaction volumes, with final concentrations of 0.9 µM of each primer and 0.25 µM probe, 1x ddPCR supermix (Bio-Rad) and 25 to 50 ng genomic DNA. Droplet generation was performed with 20 µL of the reaction mix, according to the manufacturer's protocol. Droplets were then transferred to a 96-well plate and PCR performed with the following thermal cycling profile: initial enzyme activation at 95°C for 10 minutes, followed by 40 cycles of denaturation at 94°C for 30 seconds and annealing/extension at 60°C for 1 minute. As a final step, the enzyme was deactivated at 98°C for 10 min. The droplets were read using the QX200 droplet reader according to manufacturer's protocol. The data was analyzed using the QuantaSoft software (1.7.4.0917; Bio-Rad). Crosshair gating was used to set a threshold for the four quadrants of droplet populations: double-negative, FAM-positive, VIC-positive and double-positive. QuantaSoft outputs the concentration in molecules/µl for each of the assays. Additionally, a linkage concentration is calculated based on the ratio of double-positive droplets, given in linked molecules/µl. We calculated percent linkage as the concentration of linked molecules divided by the mean concentration of the individual assays transformed to percentage.

Results

Identification of fusion transcripts in EC cell lines from paired-end RNA-seq data

RNA sequencing of the three EC (2102Ep, 833KE, and NTERA2) and two ES (H9 and Shef3) cell lines generated a total of 199 pairs of 76 bp sequencing reads that passed filtering (Additional file 1, Table S3).

Fusion transcript analyses of the RNA-seq data resulted in an initial list of 1210 unique fusion breakpoints with a probability score above 0.5. Subsequent heuristic filtering nominated nine fusion transcripts which were considered strong enough for further experimental validation (Figure 1). Briefly, 283 candidate fusions were first removed as they were also detected in ES cell lines or normal human tissues (i.e. external data from the Illumina Human Body Map v2 data set). Further technical filtering of fusion breakpoints with a high ratio of multi-mapping spanning reads and breakpoints with a high degree of homology or breakpoint sequence identity, removed 621 and 130 candidates respectively. To enrich for functionally important breakpoints, we removed breakpoints that did not pass at least three out of four functional filters. As an additional step, we used the SOAPfuse fusion finder to identify a list of 85 potential EC fusion breakpoint candidates. Of these, only 11 overlapped with the initial list of EC specific breakpoints generated by deFuse, where five of these were already kept through the filtering process. The remaining six overlapping candidates were retrieved for evaluation in the final candidate list. After filtering steps, a list of 65 unique candidate fusion breakpoints remained, and was manually curated by viewing alignments in IGV and the UCSC genome browser. Fusion transcripts likely to be generated by polymerase read-through were filtered out, except for a read-through between *CLEC6A* and *CLEC4D* located on chromosome arm 12p found to be recurrent in all three EC cell lines. Fusion candidates where the breakpoint did not match intact conserved exon – exon boundaries were filtered out, except for a breakpoint, *ETV6- RP11-434C1.1,* also located on chromosome arm 12p, which was not strictly a fusion transcript but a transcript produced from an unannotated alternative promoter. This alternative promoter of *ETV6* was specially considered based on the known oncogenic relevance of the ETS family of transcription factors [26]. The final list of transcripts selected for experimental validation consisted of two inter-chromosomal and seven intra-chromosomal fusion transcript candidates (Figure 2, Table 1). Of the seven intrachromosomal fusions, five included breakpoints with both partner genes located on chromosome arm 12p.

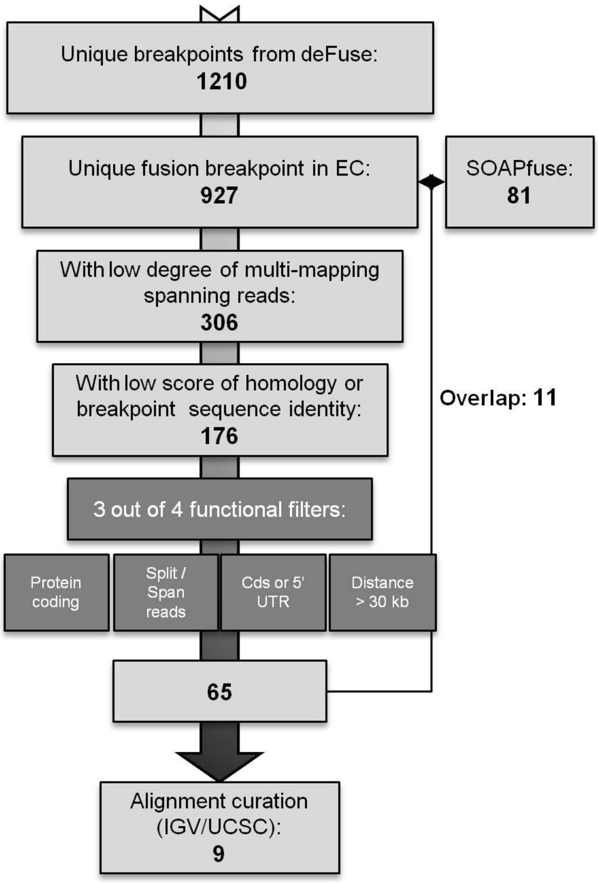

Unique breakpoints from deFuse:
1210
Unique fusion breakpoint in EC:
927
SOAPfuse:
81
With low degree of multi-mapping spanning reads:
306
With low score of homology or breakpoint sequence identity:
176
Overlap: 11
3 out of 4 functional filters:
Protein coding
Split / Span reads
Cds or 5' UTR
Distance > 30 kb
65
Alignment curation (IGV/UCSC):
9

Figure 1: Filtering pipeline of nominated fusion transcripts.

The identified fusion transcripts were filtered in a successive manner, resulting in nine final fusion transcripts that were experimentally validated.

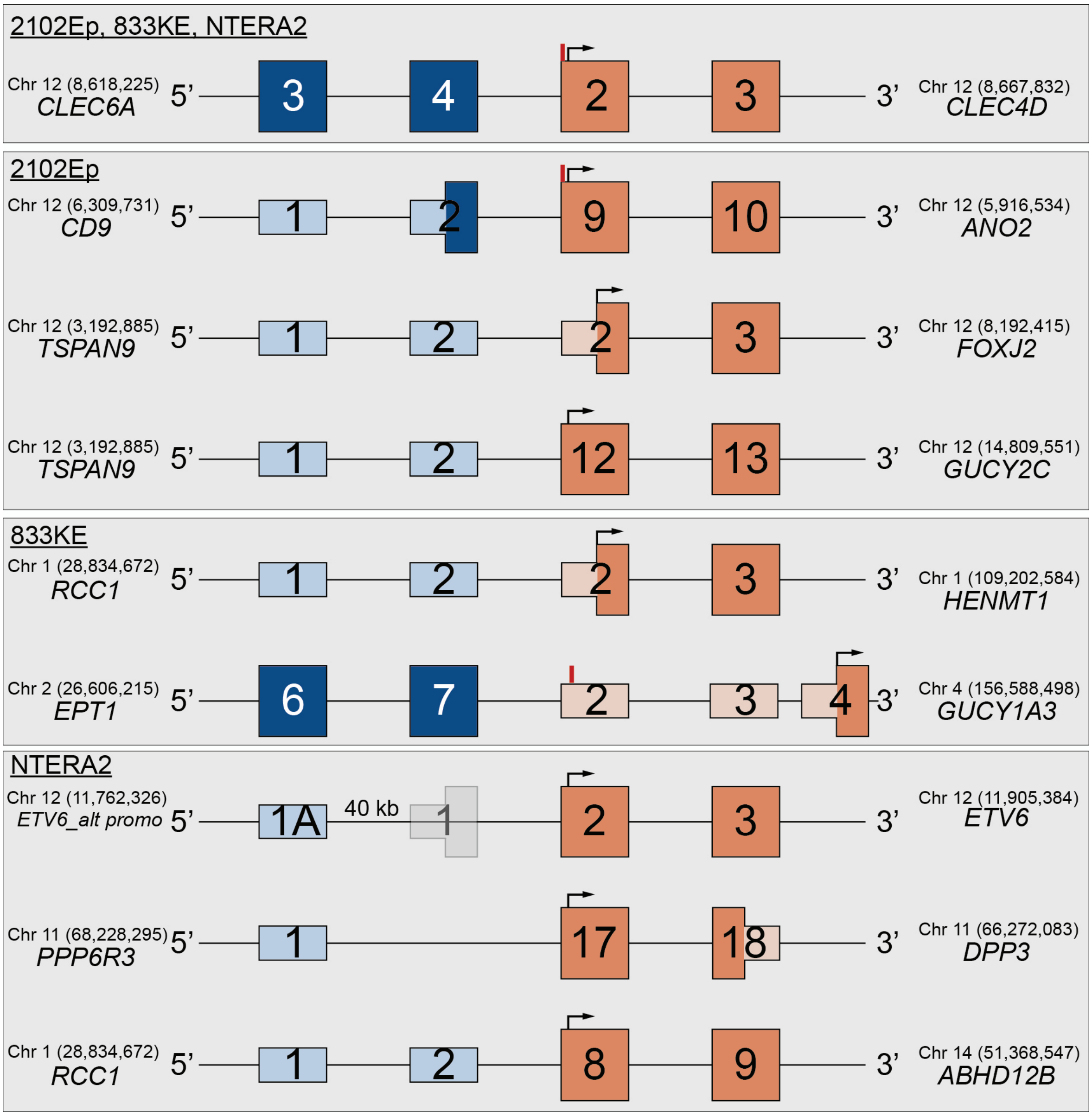

2102Ep, 833KE, NTERA2
Chr 12 (8,618,225) CLEC6A 5' 3 4 2 3 3' Chr 12 (8,667,832) CLEC4D
2102Ep
Chr 12 (6,309,731) CD9 5' 1 2 9 10 3' Chr 12 (5,916,534) ANO2
Chr 12 (3,192,885) TSPAN9 5' 1 2 2 3 3' Chr 12 (8,192,415) FOXJ2
Chr 12 (3,192,885) TSPAN9 5' 1 2 12 13 3' Chr 12 (14,809,551) GUCY2C
833KE
Chr 1 (28,834,672) RCC1 5' 1 2 2 3 3' Chr 1 (109,202,584) HENMT1
Chr 2 (26,606,215) EPT1 5' 6 7 2 3 4 3' Chr 4 (156,588,498) GUCY1A3
NTERA2
Chr 12 (11,762,326) ETV6_alt promo 5' 1A 40 kb 1 2 3 3' Chr 12 (11,905,384) ETV6
Chr 11 (68,228,295) PPP6R3 5' 1 17 18 3' Chr 11 (66,272,083) DPP3
Chr 1 (28,834,672) RCC1 5' 1 2 8 9 3' Chr 14 (51,368,547) ABHD12B

Figure 2: TGCT fusion transcripts identified by RNA-sequencing.

All the nine novel transcripts have fusion breakpoints at intact exon-exon boundaries, except for the *ETV6* gene, where a new alternative promoter (exon 1A) was connected to exon 2. The breakpoint boundaries are indicated between upstream partner gene (blue) and downstream partner gene (orange). Full height of boxes of solid color represent predicted coding regions of original partner genes. Arrows mark the start codons of fusion transcript ORFs identified by the ORF finder at the National Centre for Biotechnology Information. Red lines mark the stop codons of upstream partner gene ORFs. The genomic coordinates indicate the exact coordinate of the fusion breakpoint in the specific partner gene.

Table 1: Nominated breakpoints from deFuse analysis of RNA-sequencing data.

Nine breakpoints remained after heuristic filtering steps of initial candidates. Of these, *CLEC6A-CLEC4D* was nominated in all three EC cell lines. Breakpoints are listed according to the cell lines in which they were identified and with ascending genomic distance between the two partner genes. Presence of ORFs was determined using the ORF finder at the National Centre for Biotechnology Information.

Cell line	Gene A	Gene B	Chromosome bands		Distance (kb)	deFuse score	ORF
2102Ep	*CLEC6A*	*CLEC4D*	12p13.31	12p13.31	31	0.99	Y
	CD9	*ANO2*	12p13.31	12p13.31	253	0.97	Y
	TSPAN9	*FOXJ2*	12p13.33-p13.32	12p13.31	4,790	0.97	Y
	TSPAN9	*GUCY2C*	12p13.33-p13.32	12p13.1-p12.3	11,370	0.94	Y
833KE	*CLEC6A*	*CLEC4D*	12p13.31	12p13.31	31	0.99	Y
	RCC1	*HENMT1*	1p35.3	1p13.3	80,325	0.92	Y
	EPT1	*GUCY1A3*	2p23.3	4q32.1	Interchromosomal	0.97	Y
NTERA2	*CLEC6A*	*CLEC4D*	12p13.31	12p13.31	31	0.83	Y
	ETV6	*RP11-434C1.1**	12p13.2	12p13.2	59	0.81	Y
	PPP6R3	*DPP3*	11q13.2-13.3	11q13.2	1,951	0.82	Y
	RCC1	*ABHD12B*	1p35.3	14q22.1	Interchromosomal	0.98	Y

* *RP11-434C1.1* was nominated as a partner to *ETV6*, located 85kb downstream. However, visual inspection revealed that the breakpoint localized to non-coding regions between these two genes and reflects an alternative promoter of *ETV6*.

Technical and clinical validation of the fusion transcripts

We performed RT-PCR to validate the presence of the nine nominated fusion transcript breakpoints in the EC and ES cell lines investigated, and for further clinical evaluation in a series of IGCN and TGCTs. All nine nominated fusion transcripts were confirmed by RT-PCR spanning the breakpoint. Successful Sanger sequences were produced from eight of these, confirming breakpoint sequences between the two gene pairs, and all found to use intact exon – exon boundaries (Additional file 2, Figure S1). The eight fusion transcripts as well as the alternative promoter usage of *ETV6* all had intact open reading frames (ORFs), theoretically encoding functional proteins. The ORFs of five out of nine encode N-terminally truncated proteins of the downstream partner gene, while three out of nine encode the full length coding sequence of the downstream partner. None of the fusion transcripts encode potential hybrid proteins with in-frame coding sequence from both intact partner proteins. However, one of the fusion transcripts, *PPP6R3-DPP3,* encodes an out-of-frame ORF encoding 198 amino acids.

Five of the fusion transcripts were only expressed in the originally nominated EC cell lines, and were thus considered private fusion events. The remaining four candidates were however found to be recurrently expressed in TGCT (Additional file 1, Table S4), and crucially, not in normal testicular parenchyma. The four recurrent candidates included the read-through between *CLEC6A* and *CLEC4D*, alternative promoter usage of *ETV6,* and two fusion transcripts both involving the first two exons of *RCC1* as an upstream partner gene connected to *HENMT1* and *ABHD12B,* located 80 Mbp apart on chromosome 1 and on chromosome 14 respectively. Expression of these recurrent transcripts was variable, with both strongly positive and weaker bands detected by agarose gel electrophoresis. For more accurate assessment of expression, we used qRT-PCR to quantify the expression level of each fusion transcript. All custom TaqMan assays were found to have efficiencies between 80-90 %. Here, a total of 73 tissue samples and cell lines were tested. None of the recurrent transcripts were expressed in normal testicular parenchyma (n = 6). The read-through between *CLEC6A* and *CLEC4D* was found to be expressed in all subtypes of TGCT, as well as pre-malignant IGCN (6/6) and ES cell lines (3/3). However, only two of the four teratoma tissue samples showed expression of the read-through. The read-through was only detected in one (the placenta) of 20 normal tissues from the human body (Figure 3A). Alternative promoter usage of the *ETV6* gene was predominantly detected in EC tissue samples and cell lines (75 % and 92 %, respectively; 6/8 and 12/13). Alternative promoter usage was also observed in ES cell lines (100 %; 3/3), seminoma (14 %; 1/7), Cc (100 %; 1/1) and YST (50 %; 2/4). None of the 20 included normal tissues expressed the alternative promoter of *ETV6* (Figure 3B). The intrachromosomal fusion transcript *RCC1-HENMT1* was widely expressed in all subtypes of TGCT, IGCN, ES cell lines, and in 6/20 normal tissue types (spleen, esophagus, trachea, thyroid, thymus and skeletal muscle; Figure 3C). By contrast, the interchromosomal fusion transcript *RCC1-ABHD12B* was found expressed predominantly in EC tissue samples and cell lines (100 %) and in seminomas (86 %; 6/7). *RCC1-ABHD12B* was further detected in 67 % (4/6) IGCN, in one of four teratomas and in one of three ES cell lines. None of the tested tissue samples from sites of the human body showed expression of this fusion transcript (Figure 3D).

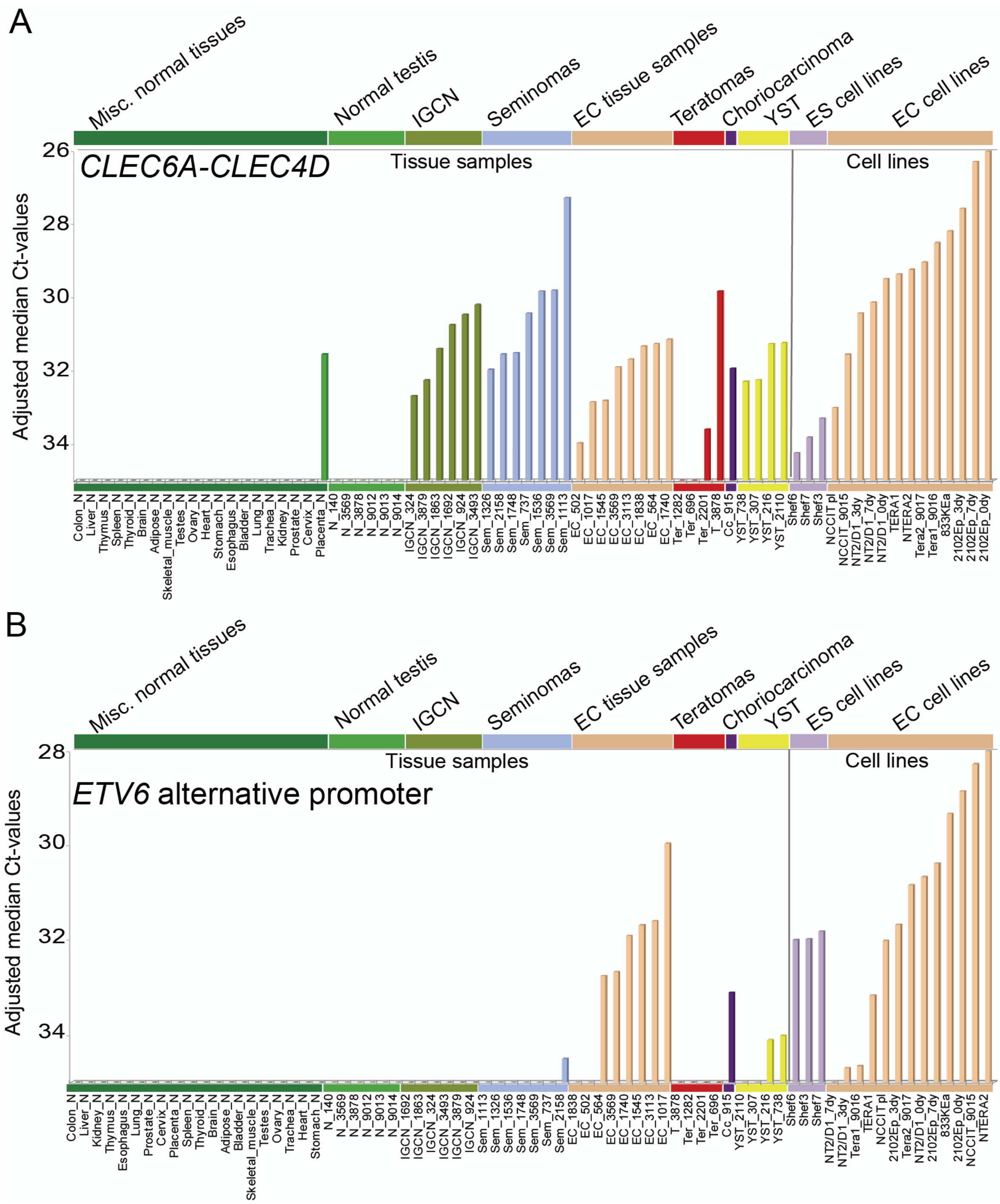


Figure 3: Quantitative RT-PCR of recurrent fusion transcripts. Expression values are reported in C_T values normalized to median C_T values of endogenous controls, with higher expression corresponding to lower C_T values. Transcripts were considered absent for $C_T > 35$. Samples are grouped together according to histological subtype and ordered with increasing expression. **A.** The read-through *CLEC6A-CLEC4D* was expressed in all subtypes of TGCTs and also in the pre-malignant IGCN samples, but not in tissue from normal testis. From normal tissue samples from 20 different human organs only placenta expressed *CLEC6A-CLEC4D*. **B.** The *ETV6* transcript with an alternative exon 1 was expressed mainly in undifferentiated EC tumor samples and cell lines. It was also found to be expressed in the ES cell lines, in 1/6 seminomas, 1 choriocarcinoma and 2/3 YSTs. None of the samples from normal testis or normal human organs expressed this novel transcript.

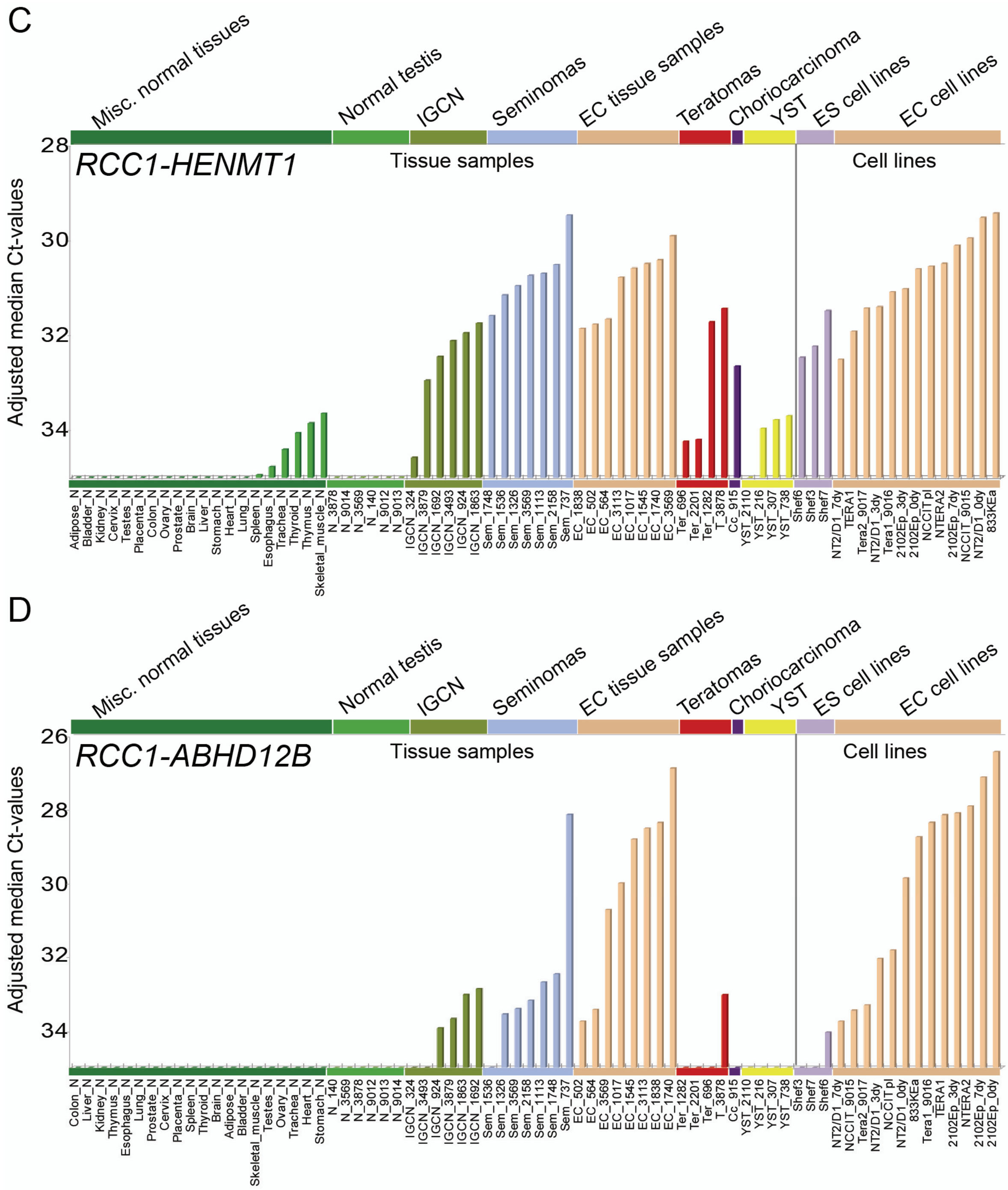


Figure 3: Quantitative RT-PCR of recurrent fusion transcripts, continued. C. The intrachromosomal fusion transcript *RCC1-HENMT1* was expressed in all IGCN and TGCT samples, except for 1 YST. Also, the fusion transcript was detected in 3/3 ES cell lines and in 5/20 samples from normal human tissue. **D.** The interchromosomal fusion transcript *RCC1-ABHD12B* was expressed in 4/6 IGCN samples, undifferentiated subtypes of TGCTs including 5/6 seminomas, and all EC cell lines and samples. Also, 1/3 ES cell lines and 1/4 teratomas expressed the fusion transcript. None of the normal testis samples or normal human tissues expressed the fusion transcript.

***RCC1* involving fusion transcripts and alternative promoter usage in *ETV6* are associated with undifferentiated subtypes of TGCT**

Total RNA isolated from NTERA2 and 2102Ep cell lines treated with RA for 0, 3 and 7 days were included in the validation panel. The NTERA2 cell line has previously been shown to differentiate in culture when treated with morphogens such as RA [21]. 2102Ep, on the other hand, does not differentiate upon *in vitro* treatment with RA and remains pluripotent [27]. Intriguingly, qRT-PCR analyses of both fusion transcripts involving *RCC1* revealed that expression decreased upon treatment with RA in NTERA2, but not in 2102Ep (Figure 4). The measured $\Delta\Delta C_T$ was -3.9 and -3.0 between 0 and 7 days of RA treatment for *RCC1-ABHD12B* and *RCC1-HENMT1* respectively. Additionally, expression of the *ETV6* transcript involving the alternative promoter was silenced after 7 days of RA treatment ($C_T > 35$; $\Delta\Delta C_T$ of -6.8). Expression of the *CLEC6A-CLEC4D* read-through did not change significantly upon RA treatment in NTERA2. The expression of all fusion transcripts remained unchanged in the 2102Ep cell line upon treatment with RA (Figure 4), in line with this cell line's previously reported nonresponsiveness to RA treatment.

Linkage assays by ddPCR to identify coupling of fusion genes on the DNA-level

Because the two fusion transcripts involving *RCC1* were recurrently expressed, and the two partner genes were not located close to *RCC1*, the possibility of genome-level rearrangements as a mechanism resulting in gene fusion was tested by ddPCR linkage analysis. For another two fusion genes *EPT1-GUCY1A3* and *PPP6R3-DPP3*, expressed in the 833KE and NTERA2 cell lines respectively, DNA copy number data indicated a shift in the vicinity of all four genetic loci (data not shown).

To establish the integrity of DNA to be included in the linkage analyses, and as proof of concept, we performed a ddPCR milepost experiment. Here, multiplexed fluorescent TaqMan assays measured linkage 1 kb, 10 kb, 50 kb and 100 kb apart. DNA isolated by the AllPrep method from the NTERA2 cell line showed highest integrity, with 90 % linkage at 1 kb, 52 % at 10 kb and 11 % at 50 kb (Additional file 3, Figure S2). No evidence of linkage was seen at 100 kb. DNA from the 833KE cell line and DNA isolated by phenol-chloroform from NTERA2 showed slightly lower linkage scores, indicating more highly degraded DNA (Additional file 3, Figure S2). After DNA fragmentation with the NspI restriction endonuclease, all linkage was substantially reduced, but some background levels remained (0 – 3.5 %; Additional file 3 - Figure S2). Fusion gene linkage analysis of *VTI1A-TCF7L2*, previously reported to be caused by a deletion in the NCI-H508 cell line, showed that 18.5 % of molecules are linked and contain both fusion partner targets (Figure 5). The interchromosomal fusion *EPT1-GUCY1A3* and the intrachromosomal fusion *PPP6R3-DPP3* showed evidence of DNA-level linkage, with 15 and 18 % linkage rates respectively (Figure 5). However, we found no evidence for DNA-level linkage for the recurrently expressed fusions *RCC1-ABHD12B* and *RCC1-HENMT* (Figure 5). As a control experiment, we found that the DNA-level linkage in all tested samples was lost upon digesting the DNA with the restriction enzyme NspI.

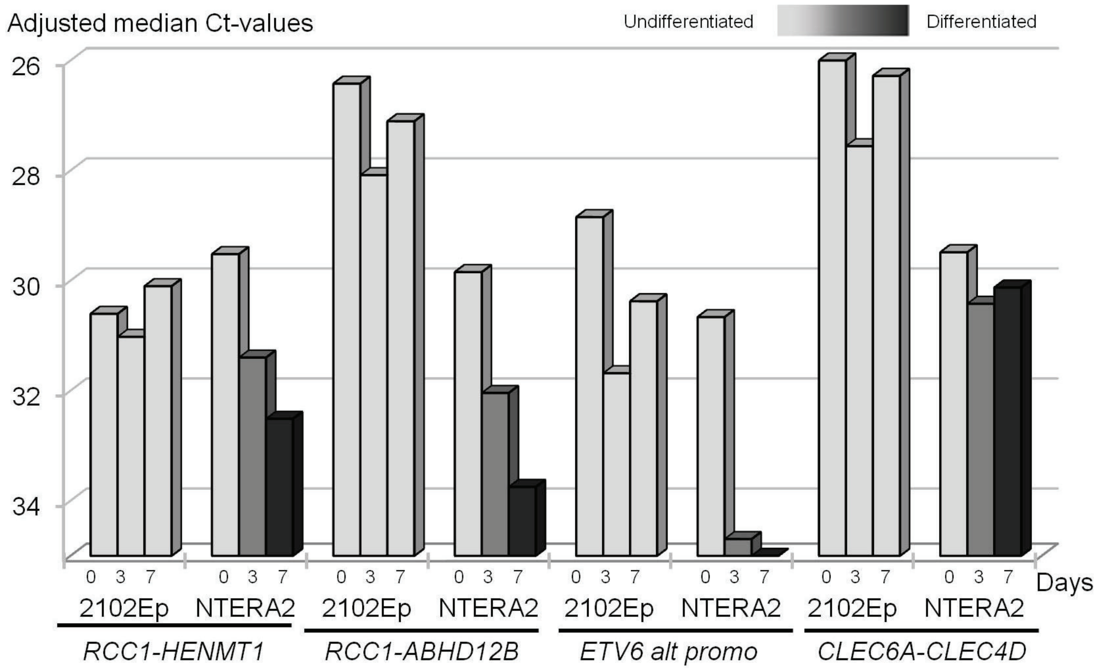


Figure 4: Fusions involving *RCC1* and the *ETV6* alternative promoter transcripts are down-regulated upon treatment with RA.

Quantitative RT-PCR results for the fusion transcripts involving *RCC1*, the *ETV6* alternative promoter and *CLEC6A-CLEC4D*, in 2102Ep and NTERA2 cells treated with RA for 0, 3 and 7 days. Expression is reported as C_T values normalized to median C_T values of endogenous controls. The lighter to darker color gradient represents an *in vitro* undifferentiated to differentiated state. No clear patterns of expression are seen in the 2102Ep cell line, except for a general lower level of expression at 3 days of RA treatment. NTERA2 has reduced expression of *RCC1-HENMT1, RCC1-ABHD12B* and *ETV6* alternative promoter after 3 and 7 days.

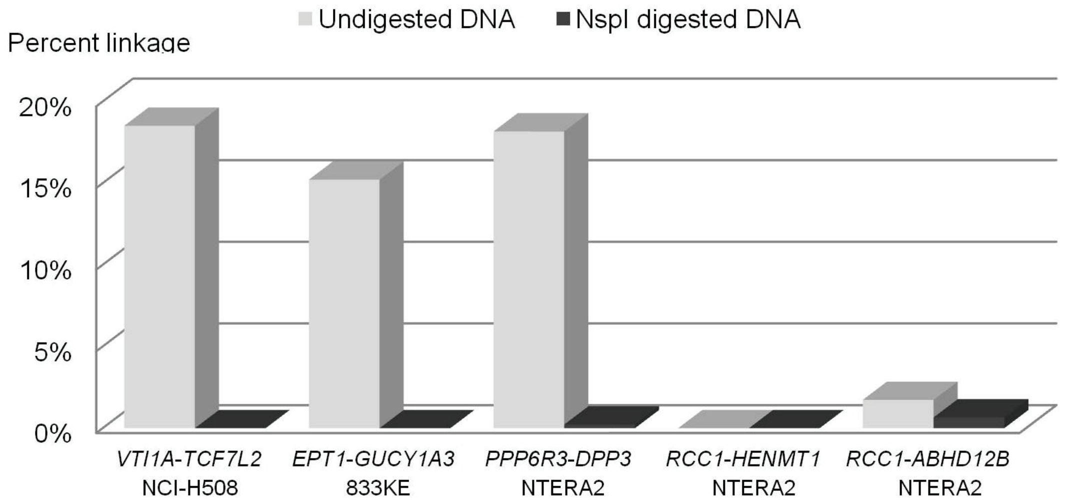


Figure 5: *EPT1-GUCY1A3* and *PPP6R3-DPP3* are chromosomally rearranged.

DNA-level linkage of fusion partner genes, reported in percent linkage from ddPCR analysis, confirmed a genomic rearrangement underlying the known *VTI1A-TCF7L2* fusion. The fusions *EPT1-GUCY1A3* and *PPP6R3-DPP3* were shown to be linked on the DNA-level in 833KE and NTERA2, respectively. No DNA-level linkage was detected for the partner genes involved in the *RCC1* fusion transcripts. Linkage was undetected after fragmentation of DNA with the NspI endonuclease.

Discussion

In this study, we have identified the presence of fusion genes and transcripts in TGCT. Nine novel fusion genes and transcripts are reported, of which *RCC1-HENMT1, RCC1-ABHD12B, CLEC6A-CLEC4D* and alternative promoter usage of *ETV6* are recurrently expressed in a significant number of clinical TGCT samples. Whereas these were detected only on the transcript level, two DNA-level fusions were identified, *EPT1-GUCY1A3* and *PPP6R3-DPP3*, although these were privately expressed by individual EC cell lines.

The non-synonymous mutation rate in TGCTs has recently been found to be low, on a scale similar to pediatric cancers [14–16]. The fact that few genes are recurrently mutated indicates that other molecular mechanisms are responsible for the development of TGCT. The genomes of TGCTs are generally aneuploid, with several recurrent gains and losses of chromosomal material [10,28]. Structural rearrangements in these aneuploid tumors are known from cytogenetic banding analyses, and i(12p) is found in the majority of TGCT [9,29,30]. Cases without i(12p) often have gain of parts of 12p material , and/or extra copies of the whole of chromosome 12 [31]. Gain or loss of chromosomal material not only leads to increase or loss of gene copies and subsequent expression changes, but can also introduce genomic rearrangements that form fusion genes. The nine novel fusion genes and transcripts found in this study, all consisting of intact ORFs which potentially encode full-length or N-terminally truncated proteins, suggest that fusion genes play an important role, and may be drivers for the malignant development of TGCTs. *CLEC6A* and *CLEC4D,* found to be involved in a read-through expressed in a high number of clinical TGCT samples, and *ETV6* with a novel alternative promoter, are all located on chromosome arm 12p. In addition, all genes involved in the *CD9-ANO2*, *TSPAN9-FOXJ2* and *TSPAN9-GUCY2C* fusion transcripts, expressed in the 2102Ep EC cell line, are located on 12p. These findings indicate that 12p is a dynamic region in TGCTs, and that gain of 12p material may be associated with expression of recurrent fusion transcripts and transcript variants.

Both fusion transcripts that involved *RCC1* and the alternative promoter usage of *ETV6* are overrepresented in undifferentiated histological subtypes of TGCT, and show substantially reduced expression in the NTERA2 EC cell line upon RA-induced differentiation *in vitro*. This indicates that they are all associated with the pluripotent phenotype. The fusion transcripts involving *RCC1* both consisted of the two first exons in the 5' UTR of *RCC1* connected to either *HENMT1* or *ABHD12B* located 80 Mb downstream on chromosome 1 and on chromosome 14, respectively. The long genomic distance between the partner genes suggests that these fusion transcripts are not expressed as a result of a read-through mechanism [32]. To investigate if the fusion transcripts involving *RCC1* were caused by genomic rearrangements, we applied linkage analysis using multiplexed fluorescent PCR assays with ddPCR. To our knowledge, this approach has not been used previously to detect rearrangements of genes resulting in fusion genes. However, linkage analysis with ddPCR has been proven successful in showing the arrangement of the Killer-cell immunoglobulin-like receptor gene complex and in chromosomal phasing [33,34]. We found no indication of DNA-level linkage for the partner genes of the *RCC1* fusion transcripts. However, the partner genes of *EPT1-GUCY1A3* and *PPP6R3-DPP3*, which also had indications of chromosomal breakpoints from DNA copy number data, were found to be linked at the DNA-level, indicating bona-fide genomic rearrangements in their respective cell lines, 833KE and NTERA2. The absence of linked partner genes of the *RCC1* fusions, implies that these fusion transcripts are expressed as the result of post-transcriptional mechanisms, such as *trans*-splicing [35]. However, we cannot rule out chromosomal rearrangements as an underlying cause of the *RCC1* fusion transcripts, since only a proportion of the input DNA in our assay consisted of DNA fragments longer than 50 kb, and none more than 100 kb in length. A chromosomal rearrangement resulting in a fusion gene may include intronic regions that are longer than these DNA fragments. Such rearrangements will be missed by ddPCR linkage analysis. For optimal sensitivity, linkage analysis should be carried out on DNA samples isolated by protocols that maintain long DNA molecules intact.

Expression of the interchromosomal *RCC1-ABHD12B* fusion transcript and transcripts involving the alternative promoter of *ETV6* was not detected in normal testis or other normal tissue samples from 20 different human organs. Also, expression of *CLEC6A-CLEC4D* was only observed in normal tissue from the placenta, indicating that it may be specifically expressed in adult male TGCT. These molecules are therefore highly specific for TGCTs in a stemness setting, and could prove to have important roles for TGCT malignant transformation, as well as biomarkers for TGCT disease. Diagnosis of TGCT through sensitive detection of these molecules in excreted body fluids such as seminal fluid or serum could have clinical potential [36].

In conclusion, to our knowledge, we present here the first fusion genes to be described in TGCT, including recurrent expression of *RCC1* involving fusions and alternative promoter usage of the *ETV6* gene, both associated with the pluripotency phenotype. These transcript variants may be important drivers of malignancy, and could potentially serve as diagnostic markers in the clinic.

References

1. Znaor A, Lortet-Tieulent J, Jemal A, Bray F: **International Variations and Trends in Testicular Cancer Incidence and Mortality**. Eur Urol 2014, 65:1095–1106.

2. Quaresma M, Coleman MP, Rachet B: **40-year trends in an index of survival for all cancers combined and survival adjusted for age and sex for each cancer in England and Wales, 1971-2011: a population-based study**. Lancet 2015, 385:1206–1218.

3. Haugnes HS, Bosl GJ, Boer H, Gietema JA, Brydøy M, Oldenburg J, *et al.*: **Long-Term and Late Effects of Germ Cell Testicular Cancer Treatment and Implications for Follow-Up**. J Clin Oncol 2012, 30:3752–3763.

4. Woodward, PJ, Heidenreich, A, Looijenga, LHJ: **Germ cell tumours**. In: Eble JN, International Agency for Research on Cancer, editors. Pathology and genetics of tumours of the urinary system and male genital organs. World Health Organization classification of tumours. Lyon: IARC Press. pp. 221–249.

5. Andrews PW, Matin MM, Bahrami AR, Damjanov I, Gokhale P, Draper JS: **Embryonic stem (ES) cells and embryonal carcinoma (EC) cells: opposite sides of the same coin**. Biochem Soc Trans 2005, 33:1526–1530.

6. Sperger JM, Chen X, Draper JS, Antosiewicz JE, Chon CH, Jones SB, *et al.*: **Gene expression patterns in human embryonic stem cells and human pluripotent germ cell tumors**. Proc Natl Acad Sci 2003, 100:13350–13355.

7. Josephson R, Ording CJ, Liu Y, Shin S, Lakshmipathy U, Toumadje A, *et al.*: **Qualification of Embryonal Carcinoma 2102Ep As a Reference for Human Embryonic Stem Cell Research**. Stem Cells 2007, 25:437–446.

8. Baker DE, Harrison NJ, Maltby E, Smith K, Moore HD, Shaw PJ, *et al.*: **Adaptation to culture of human embryonic stem cells and oncogenesis in vivo**. Nat Biotechnol 2007, 25:207–215.

9. Atkin NB, Baker MC: **i(12p): specific chromosomal marker in seminoma and malignant teratoma of the testis?**. Cancer Genet Cytogenet 1983, 10:199–204.

10. Skotheim RI, Lothe RA: **The testicular germ cell tumour genome**. APMIS 2003, 111:136–151.

11. Alagaratnam S, Harrison N, Bakken AC, Hoff AM, Jones M, Sveen A, *et al.*: **Transforming pluripotency: an exon-level study of malignancy-specific transcripts in human embryonal carcinoma and embryonic stem cells**. Stem Cells Dev 2013, 22:1136–1146.

12. Looijenga LHJ, Zafarana G, Grygalewicz B, Summersgill B, Debiec-Rychter M, Veltman J, *et al.*: **Role of gain of 12p in germ cell tumour development**. APMIS 2003, 111:161–171.

13. Sheikine Y, Genega E, Melamed J, Lee P, Reuter VE, Ye H: **Molecular genetics of testicular germ cell tumors**. Am J Cancer Res 2012, 2:153–167.

14. Brabrand S, Johannessen B, Axcrona U, Kraggerud SM, Berg KG, Bakken AC, *et al.*: **Exome sequencing of bilateral testicular germ cell tumors suggests independent development lineages**. Neoplasia 2015, 17:167–174.

15. Litchfield K, Summersgill B, Yost S, Sultana R, Labreche K, Dudakia D, *et al.*: **Whole-exome sequencing reveals the mutational spectrum of testicular germ cell tumours**. Nat Commun 2015, 6:5973.

16. Cutcutache I, Suzuki Y, Tan IB, Ramgopal S, Zhang S, Ramnarayanan K, *et al.*: **Exome-wide Sequencing Shows Low Mutation Rates and Identifies Novel Mutated Genes in Seminomas**. Eur Urol 2015, 68:77–83.

17. Andersson AK, Ma J, Wang J, Chen X, Gedman AL, Dang J, *et al.*: **The landscape of somatic mutations in infant MLL-rearranged acute lymphoblastic leukemias**. Nat Genet 2015, 47:330–337.

18. Osuna D, de Alava E: **Molecular pathology of sarcomas**. Rev Recent Clin Trials 2009, 4:12–26.

19. Brohl AS, Solomon DA, Chang W, Wang J, Song Y, Sindiri S, *et al.*: **The Genomic Landscape of the Ewing Sarcoma Family of Tumors Reveals Recurrent *STAG2* Mutation**. PLoS Genet 2014, 10.

20. Skotheim RI, Lind GE, Monni O, Nesland JM, Abeler VM, Fosså SD, *et al.*: **Differentiation of human embryonal carcinomas in vitro and in vivo reveals expression profiles relevant to normal development**. Cancer Res 2005, 65:5588–5598.

21. Andrews PW: **Retinoic acid induces neuronal differentiation of a cloned human embryonal carcinoma cell line in vitro**. Dev Biol 1984, 103:285–293.

22. McPherson A, Hormozdiari F, Zayed A, Giuliany R, Ha G, Sun MGF, *et al.*: **deFuse: An Algorithm for Gene Fusion Discovery in Tumor RNA-Seq Data**. PLoS Comput Biol 2011, 7:e1001138.

23. Jia W, Qiu K, He M, Song P, Zhou Q, Zhou F, *et al.*: **SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data**. Genome Biol 2013, 14:R12.

24. Rozen S, Skaletsky H: **Primer3 on the WWW for general users and for biologist programmers**. Methods MolBiol 2000, 132:365–386.

25. Bass AJ, Lawrence MS, Brace LE, Ramos AH, Drier Y, Cibulskis K, *et al.*: **Genomic sequencing of colorectal adenocarcinomas identifies a recurrent *VTI1A-TCF7L2* fusion**. Nat Genet 2011, 43:964–968.

26. Sharrocks AD: **The ETS-domain transcription factor family**. Nat Rev MolCell Biol 2001, 2:827–837.

27. Matthaei KI, Andrews PW, Bronson DL: **Retinoic acid fails to induce differentiation in human teratocarcinoma cell lines that express high levels of a cellular receptor protein**. Exp Cell Res 1983, 143:471–474.

28. Gilbert D, Rapley E, Shipley J: **Testicular germ cell tumours: predisposition genes and the male germ cell niche**. Nat Rev Cancer 2011, 11:278–288.

29. Castedo SMMJ, Jong B de, Oosterhuis JW, Seruca R, Idenburg VJS, Dam A, *et al.*: **Chromosomal Changes in Human Primary Testicular Nonseminomatous Germ Cell Tumors**. Cancer Res 1989, 49:5696–5701.

30. Castedo SMMJ, Jong B de, Oosterhuis JW, Seruca R, Meerman GJ te, Dam A, *et al.*: **Cytogenetic Analysis of Ten Human Seminomas**. Cancer Res 1989, 49:439–443.

31. Kraggerud SM, Skotheim RI, Szymanska J, Eknæs M, Fosså SD, Stenwig AE, *et al.*: **Genome profiles of familial/bilateral and sporadic testicular germ cell tumors**. Genes Chromosomes Cancer 2002, 34:168–174.

32. Frenkel-Morgenstern M, Lacroix V, Ezkurdia I, Levin Y, Gabashvili A, Prilusky J, *et al.*: **Chimeras taking shape: Potential functions of proteins encoded by chimeric RNA transcripts**. Genome Res 2012, 7:1231–1242.

33. Roberts C h, Jiang W, Jayaraman J, Trowsdale J, Holland MJ, Traherne JA: **Killer-cell Immunoglobulin-like Receptor gene linkage and copy number variation analysis by droplet digital PCR**. Genome Med 2014, 6:20.

34. Regan JF, Kamitaki N, Legler T, Cooper S, Klitgord N, Karlin-Neumann G, *et al.*: **A Rapid Molecular Approach for Chromosomal Phasing**. PLoS ONE 2015, 10:e0118270.

35. Gingeras TR: **Implications of chimaeric non-co-linear transcripts**. Nature 2009, 461:206–211.

36. Favilla V, Cimino S, Madonia M, Morgia G: **New advances in clinical biomarkers in testis cancer**. Front Biosci 2010, 2:456–477.

Additional files:

Additional file 1: Contains Supplementary Tables S1 to S4:

Table S1: primers used in RT-PCR and qRT-PCR validation and quantification of fusion genes and transcripts

Gene_A	Gene_B	Assay	Forward_primer	TaqMan_probe	Reverse_primer	Tm[a]
TSPAN9	*GUCY2C*	RT-PCR_30x	GTCAGCAGCCAAGGTGTTTC	NA	TTTGTCGTATTTGCACTGTCG	54°C
TSPAN9	*FOXJ2*	RT-PCR_30x	GTCAGCAGCCAAGGTGTTTC	NA	GGTGGCATTAGGATCTGTGG	60°C
CLEC6A	*CLEC4D*	RT-PCR_30x	ATGGGAGCACATTTGGTTGT	NA	TTGATGCATTTGAGCTTTGC	57°C
EPT1	*GUCY1A3*	RT-PCR_30x	TTGTGCATTATGTGTGACTCTTC	NA	CTCACGCTTTCTCCAAGAGC	54°C
RCC1	*HENMT1*	RT-PCR_30x	CAGTGGTCGCTTCTTCTCCT	NA	ACGAACTGGTACCGCTGTCT	60°C
RCC1	*ABHD12B*	RT-PCR_30x	CAGTGGTCGCTTCTTCTCCT	NA	AAGGCACTGTCCTGTCATCC	60°C
CD9	*ANO2*	RT-PCR_30x	GGCACCAAGTGCATCAAATA	NA	GAGGGTAGGCAGCCTCATAG	60°C
Upstream_ETV6	*ETV6*	RT-PCR_30x	ATCTGGCTGCCACTTTGAGT	NA	TCGAGTCTTCCTCCATCCTG	60°C
PPP6R3	*DPP3*	RT-PCR_30x	CTTGATACGTCCGCCATTTT	NA	GACTGTTGCATACCCCTCGT	60°C
CLEC6A	*CLEC4D*	TaqMan qRT-PCR	GAGCACATTTGGTTGTGTTCAAC	CAGAGCAGTGGAAGGA	TGAGAAGTAAGATGAAAACTACAGCAATAA	60°C
Upstream_ETV6	*ETV6*	TaqMan qRT-PCR	TTGAGTCTTAGTTCTTTGTGGAAACTTC	CTGCATAGCAGGAACG	GGGCTCTCTGGAGGTGTATATGA	60°C
RCC1	*HENMT1*	TaqMan qRT-PCR	TCGCAGTGGTCGCTTCTTCT	CTTGGATTTGTTAAGGATTC	AGCACTGACTCAGCAAATAAGAGTTACT	60°C
RCC1	*ABHD12B*	TaqMan qRT-PCR	TCGCAGTGGTCGCTTCTTC	AACTCTTATTTGATTTACCGGAAC	GGCATCCATAAGTGTACGTAAAAATC	60°C

[a] Annealing temperature used in PCR cycling.

Table S2: Primers and probes for assays used in ddPCR linkage experiments.

Assay_name	Assay_location	Amp_length	Dye	Forward_primer	TaqMan_probe	Reverse_primer
Milepost_Ref	chr10:114,220,376-114,220,451	75	6-FAM	CCAGAGCCGTTTAAGCTAATGTC	TCCTCTAGCCTGTTTTAT	TGCTAATCAGCAATGCTTCACA
Milepost_1kb	chr10:114,219,282-114,219,357	76	VIC	TCACTGTGACTGAAGAAGCAGAGTCT	AGATAGCATTTTCATTGTCC	AAAGGCCTTCTTCCTCACCAA
Milepost_10kb	chr10:114,208,464-114,208,533	70	VIC	GGATTGGTGGGCACTTCACT	CCCACTTGTCCTGGTAC	GACAACTTAGGTCCGGGAGAAA
Milepost_50kb	chr10:114,170,396-114,170,468	73	VIC	CTGTGCCCATGAAACTGAACCT	AGACCCCTGGTCAGC	CCTACCAGTGGCCAAGAATTG
Milepost_100kb	chr10:114,120,346-114,120,416	71	VIC	CCAGCCTCCTGCCCTAAGATA	CCTCTGGTGGATTCC	GCCCCCAGGAATACTGTCTTC
VTI1A-TCF7L2	chr10:114,220,376-114,220,451	75	6-FAM	CCAGAGCCGTTTAAGCTAATGTC	TCCTCTAGCCTGTTTTAT	TGCTAATCAGCAATGCTTCACA
VTI1A-TCF7L2	chr10:114,859,726-114,859,795	70	VIC	CACGCAAGTACCCCAGCTTT	CTGTCCCAAGTTCCCA	TCCTCCAGCCTGCTAAGAGAGA
EPT1-GUCY1A3	chr2:26,605,907-26,605,981	75	6-FAM	TGCCTGAAGTCCCGTAATCAG	TACAACTGAAATTGAGGCCCA	TCAGAGGGCAGACTCAAAAAGC
EPT1-GUCY1A3	chr4:156,588,568-156,588,677	110	VIC	CCAAGTAAGTGGTCGCTGCAT	CAGGCGGTCCCTC	AAGGTTTCTCAGGAGAGGAAGGA
PPP6R3-DPP3	chr11:68,236,074-68,236,147	73	6-FAM	TCATAGAGAGGCCTCGAGGTCTA	CCAGCTGCTCTGCGGA	CCTCCCCAACACTGAAACCA
PPP6R3-DPP3	chr11:66,272,189-66,272,287	99	VIC	TGCTGCTGCGTAAGGAATCTC	AAGCTCATTGTTCAGCCCA	GCTCATCAAGGGAGCTCAAGA
RCC1-HENMT1	chr1:28,833,161-28,833,249	89	6-FAM	CCACTGTTTATCTTACTGCCTCATAGTAG	CACATTGTCGTTCTCAATAT	AAAGGCAAATGAGGATCCAGAA
RCC1-HENMT1	chr1:109,202,369-109,202,476	108	VIC	AACACCCCGGGAAATGCT	TTTGTTTGCAATTATCTCAC	CCCTTCTTAGCTGCGTCTTTGA
RCC1-ABHD12B	chr1:28,833,161-28,833,249	89	6-FAM	CCACTGTTTATCTTACTGCCTCATAGTAG	CACATTGTCGTTCTCAATAT	AAAGGCAAATGAGGATCCAGAA
RCC1-ABHD12B	chr14:51,368,772-51,368,869	98	VIC	GGGCACTAAGATTTGGGTAAAGTATC	AGTATTTGCTATTTTCCTCATGTC	TCAGAGAGAGTACTGAATTGGTGAAGTC

Table S3: Pairs of sequencing reads passing filter

Cell line	Filtered clusters (M)
2102Ep	29.3
H9	27.6
833KE	46.3
Shef3	48.1
NTERA2	47.5

Table S4: Candidate fusion transcript breakpoints validated by RT-PCR

Sample	Histological Subgroup	*TSPAN9-GUCY2C*	*TSPAN9-FOXJ2*	*CLEC6A-CLEC4D*	*EPT1-GUCY1A3*	*RCC1-HENMT1*	*CD9-ANO2*	*ETV6-RP11-434C1.1*	*PPP6R3-DPP3*	*RCC1-ABHD12B*
N_140	Normal testis	0	0	0	0	0	0	0	0	0
N_3569	Normal testis	0	0	0	0	0	0	0	0	0
N_3878	Normal testis	0	0	0	0	0	0	0	0	0
N_9012	Normal testis	0	0	0	0	0	0	0	0	0
N_9013	Normal testis	0	0	0	0	0	0	0	0	0
N_9014	Normal testis	0	0	0	0	0	0	0	0	0
IGCN_324	IGCN	0	0	0	0	0	0	0	0	0
IGCN_924	IGCN	0	0	2	0	0	0	0	0	0
IGCN_1692	IGCN	0	0	2	0	0	0	0	0	0
IGCN_1863	IGCN	0	0	4	0	3	0	0	0	0
IGCN_3493	IGCN	9	9	9	9	9	9	9	9	9
IGCN_3879	IGCN	0	0	2	0	0	0	0	0	0
Sem_737	Seminomas	0	0	2	0	1	0	0	0	1
Sem_1113	Seminomas	0	0	2	0	1	0	0	0	3
Sem_1326	Seminomas	0	0	4	0	3	0	0	0	0
Sem_1536	Seminomas	0	0	2	0	0	0	0	0	0
Sem_1748	Seminomas	0	0	2	0	0	0	0	0	3
Sem_2158	Seminomas	0	0	2	0	3	0	0	0	3
Sem_3569	Seminomas	0	0	2	0	0	0	0	0	0
EC_502	EC tissue samples	9	9	9	9	9	9	9	9	9
EC_564	EC tissue samples	0	0	2	0	1	0	1	0	0
EC_1017	EC tissue samples	0	0	2	0	1	0	1	0	1
EC_1545	EC tissue samples	0	0	0	0	0	0	0	0	0
EC_1740	EC tissue samples	0	0	2	0	3	0	1	0	1
EC_1838	EC tissue samples	0	0	2	0	1	0	0	0	1
EC_3113	EC tissue samples	0	0	2	0	1	0	2	0	1

Sample	Histological Subgroup	*TSPAN9-GUCY2C*	*TSPAN9-FOXJ2*	*CLEC6A-CLEC4D*	*EPT1-GUCY1A3*	*RCC1-HENMT1*	*CD9-ANO2*	*ETV6-RP11-434C1.1*	*PPP6R3-DPP3*	*RCC1-ABHD12B*
EC_3569	EC tissue samples	0	0	4	0	1	0	4	0	3
NCCIT_9015	EC cell lines	0	0	2	0	1	0	2	0	3
Tera1_9016	EC cell lines	0	0	2	0	0	0	3	0	1
Tera2_9017	EC cell lines	0	0	2	0	0	0	1	0	3
NT2/D1_0dy_RA	EC cell lines	0	0	2	0	1	0	1	1	1
NT2/D1_3dy_RA	EC cell lines	0	0	2	0	0	0	3	1	1
NT2/D1_7dy_RA	EC cell lines	0	0	2	0	0	0	0	1	0
2102Ep_0dy_RA	EC cell lines	1	2	2	0	0	1	1	0	1
2102Ep_3dy_RA	EC cell lines	1	2	2	0	0	1	1	0	1
2102Ep_7dy_RA	EC cell lines	1	2	2	0	0	1	1	0	1
833KEa SSEA3	EC cell lines	0	0	2	2	1	0	1	0	1
Cc_915	Choriocarcinoma	0	0	4	0	0	0	3	0	0
YST_216	YST	0	0	4	0	0	0	0	0	0
YST_307	YST	0	0	2	0	0	0	0	0	0
YST_738	YST	0	0	2	0	0	0	0	0	0
YST_2110	YST	0	0	2	0	0	0	0	0	0
Ter_696	Teratoma	0	0	4	0	0	0	0	0	0
Ter_1282	Teratoma	0	0	0	0	0	0	0	0	0
Ter_2201	Teratoma	0	0	0	0	0	0	0	0	0
T_3878	Teratoma	0	0	2	0	0	0	0	0	0
Shef3	ES	0	0	4	0	0	0	1	0	0
Shef6	ES	0	0	4	0	0	0	0	0	0
Shef7	ES	0	0	4	0	0	0	1	0	0

Coding:

0	Negative	1	Clear positive - single band	2	Clear positive - multiple bands	3	Weak positive - single band	4	Weak positive - multiple bands
9	Noise								

Additional file 2: Supplementary Figure S1

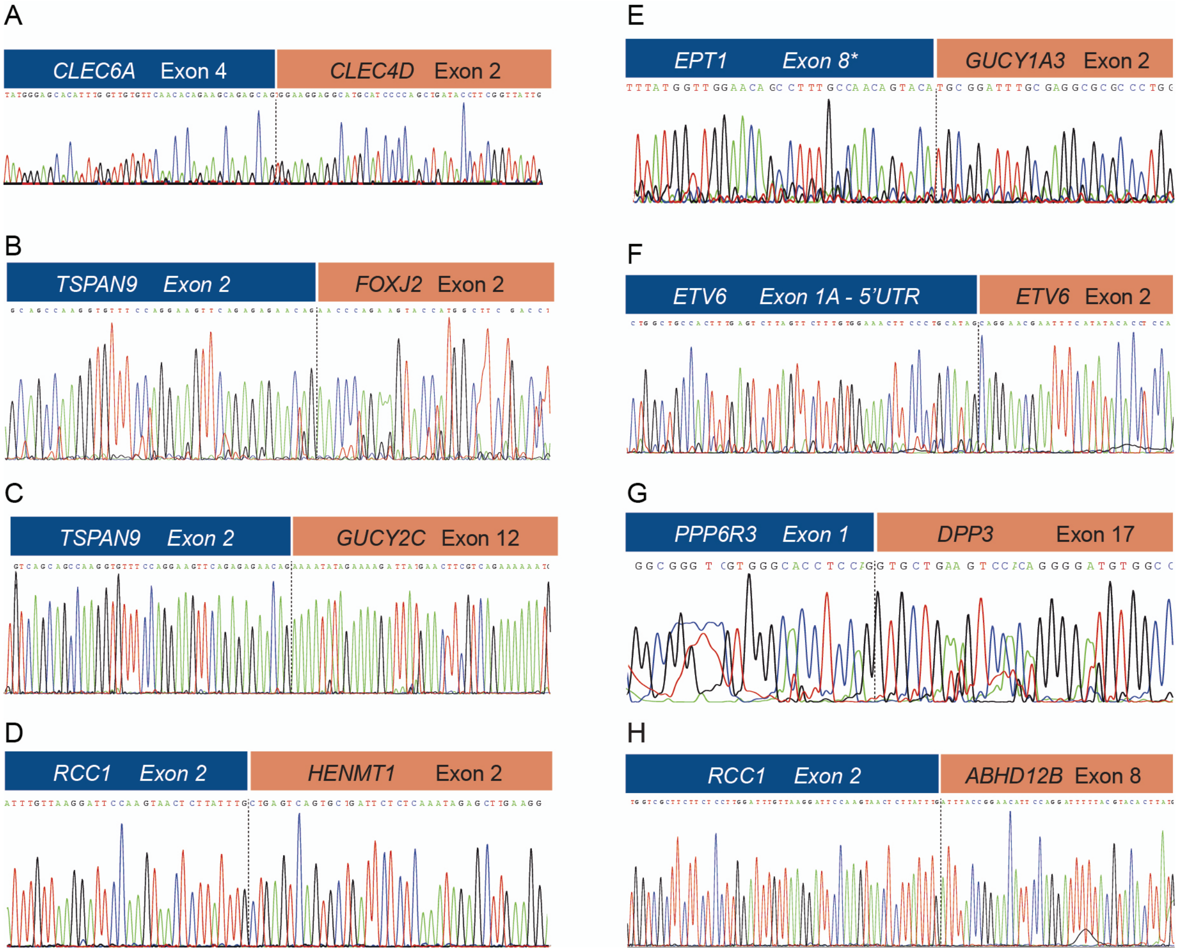


Figure S1: Sanger sequencing electropherograms validating fusion transcript breakpoints.

Electropherograms from Sanger sequencing confirming breakpoint sequences of the nominated novel transcript breakpoints from RNA sequencing. **A.** *CLEC6A-CLEC4D*, 2102Ep **B.** *TSPAN9-FOXJ2*, 2102Ep **C.** *TSPAN9-GUCY2C*, 2102Ep **D.** *RCC1-HENMT1*, 833KE **E.** *EPT1-GUCY1A3*, 833KE. *For *EPT1-GUCY1A3*, two fusion transcript variants detected by gel electrophoresis, the Sanger sequencing electropherogram shows the longest product spanning *EPT1* exon 8 to *GUCY1A3* exon 2. **F.** *ETV6* alternative promoter, NTERA2 **G.** *PPP6R3-DPP3*, NTERA2 **H.** *RCC1-ABHD12B*, NTERA2.

Additional file 3: Supplementary Figure S2

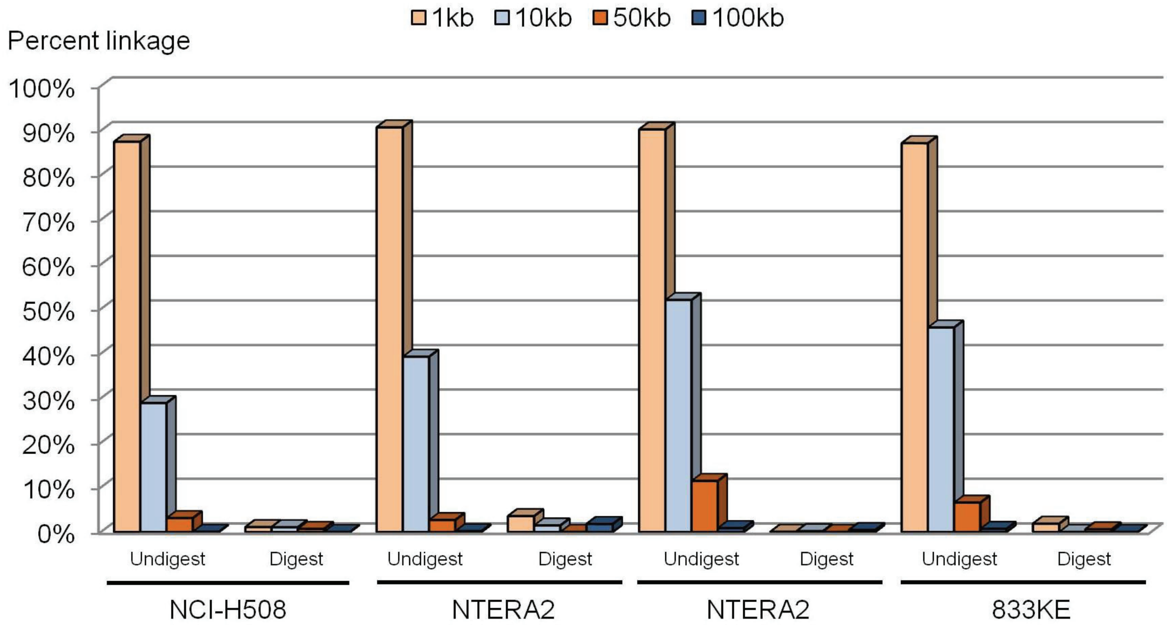


Figure S2: Milepost ddPCR linkage results of DNA from the NCI-H508, NTERA2 and 833KE cell lines. From the NTERA2 cell line, we included two different DNA samples, isolated with different protocols. Results are reported as percent linked DNA molecules, calculated with concentration of linked molecules divided by the mean concentration of the individual target sequences. A reference assay with FAM fluorescence was multiplexed with VIC assays located 1 kb, 10 kb, 50 kb and 100 kb upstream. DNA from the NTERA2 cell line, isolated with the AllPrep protocol, has longer/more intact DNA fragments. Linkage is depleted in all DNA samples upon treatment with the NspI restriction endonuclease. No linkage was detected at 100 kb for any of the DNA samples.