UNIVERSITY OF SOUTHAMPTON

RNA-SEQUENCING FOR INVESTIGATION OF GENE FUSIONS AND SPLICING ABNORMALITIES IN LEUKAEMIA

by

Marcin Knut

A thesis submitted in partial fulfilment
for the degree of Doctor of Philosophy

in the
Faculty of Medicine
Academic unit of Human Development and Health

2016

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

Faculty of Medicine

Academic unit of Human Development and Health

by Marcin Knut

Myeloid malignancies, a group of neoplasms affecting myeloid blood lineages, are driven by a variety of somatically acquired mutations and gene fusions. Some of these genes, e.g. *U2AF1*, encode factors involved in pre-mRNA splicing but it remains unclear how these abnormalities contribute to malignancy. The presence of gene fusions is typically signalled by the finding of a somatic chromosome translocation but it is unclear to what extent leukaemia might be driven by cryptic abnormalities that have escaped detection by cytogenetic analysis. This study focuses on the use of RNA-Seq by Next Generation Sequencing (NGS) to detect gene fusions and help understand the role of *U2AF1*. Two bioinformatic pipelines were developed. The fusion detection pipeline combined with manual selection of candidates achieved 100% accuracy across 5 samples with previously discovered, cytogenetically visible gene fusions. Across 4 samples, for which gene fusions were suspected, the developed methods allowed for identification and validation of at least 1 gene fusion in each sample, and up to 5 within a single sample, with overall accuracy of 29%. Discovery of *BCR-JAK2* gene fusion in one of the samples explained patient's short-termed remission following application of JAK1/JAK2 inhibitor ruxolitinib which highlights importance of RNA-Seq as a tool potentially leading to targeted therapy. A knockdown was performed on *U2AF1* and both of its isoforms in a cell line. The analysis using second bioinformatic pipeline showed that *U2AF1* depletion modified RNA processing events, disproportionally affecting usage of terminal exons over others. Observed upregulated exons in depleted cells had longer AG exclusion zones (AGEZ) and polypyrimidine tracts (PPTs). A small number of transcripts responded differently to knockdowns of specific isoforms. These results provide new insights into gene fusions as well as *U2AF1* function to better understand myeloid malignancies.

Contents

List of Figures

List of Tables

Abbreviations

A3SS	alternative 3' splice site
A5SS	alternative 5' splice site
AGEZ	AG-exclusion zone
AID	activation-induced deaminase
ALL	acute lymphoblastic leukaemia
AML	acute myeloid leukaemia
APA	alternative polyadenylation site
ASCT	allogeneic stem cell transplantation
BLAST	Basic Local Alignment Search Tool
BLAT	Basic Local Alignment Search Tool - Like Tool (sic.)
BM	bone marrow
bp	base pair
bs	branch site
CCR	complete cytogenetic response
cDNA	complementary deoxyribonucleic acid
CHR	complete hematologic response
CML	chronic myelogenous leukaemia
CNV	copy number variation
DNA	deoxyribonucleic acid
dNTP	deoxyribonucleotide triphosphate
ESE	exonic splicing enhancer

ESS	exonic splicing silencer
ET	essential thrombocythemia
FC	fusion candidate
FDR	False Discovery Rate
FISH	Fluorescent In-Situ Hybridization
FPKM	Fragments per Kilobase of exon model per Million fragments
GWAS	Genome-Wide Association Study
HEK	human embryonic kidney
HES	hypereosinophilic syndrome
HPC	High Performance computing Cluster
ICD-O-3	International Classification of Diseases for Oncology, version 3
indel	insertion/deletion
IUM	initially unmapped read
LD	lactate dehydrogenase
LT-PCR	long-template polymerase chain reaction
MDS	myelodysplastic syndrome
MHEJ	Microhomology End-Joining Pathway
MPN	myeloproliferative neoplasm
MRD	Minimal Residual Disease
mRNA	messenger ribonucleic acid
MXE	mutually exclusive exons
NGS	Next Generation Sequencing
NHEJ	non-homologous end joining pathway
NMD	nonsense-mediated decay
nt	nucleotide
ORF	Open Reading Frame
PB	peripheral blood

polyA	polyadenylation signal
PPT	polypyrimidine tract
pre-mRNA	precursor messenger ribonucleic acid
PSI	percent spliced in
PTC	premature termination codon
PU	probability of unpaired
QC	quality control
RAG	recombination-activating gene
RI	retained intron
RIN	Ribonucleic Acid Integrity Number
RNA	ribonucleic acid
RNA-Seq	ribonucleic acid sequencing
RPL	residual protein level
RT-PCR	Reverse-Transcription Polymerase Chain Reaction
SD	standard deviation
SE	skipped exon
siRNA	short interfering ribonucleic acid
SNP	single nucleotide polymorphism
snRNP	small nuclear ribonucleic protein
SP	supporting pairs
SR	supporting reads
SSO	splice switching oligonucleotide
SVM	support vector machine
TK	tyrosine kinase
UTR	untranslated region
WHO	World Health Organisation
ZF	zinc finger

Declaration of Authorship

I, Marcin Knut, declare that this thesis and the work presented in it are my own. I confirm that: this work was done wholly or mainly while in candidature for a research degree at the University; where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated; where I have consulted the published work of others, this is always clearly attributed; where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work; I have acknowledged all main sources of help; where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself; parts of this work have been published as:

J. Schwaab (i), M. Knut (i), C. Haferlach, G. Metzgeroth, H. P. Horny, A. Chase, W. J. Tapper, J. Score, K. Waghorn, N. Naumann, M. Jawhar, A. Fabarius, W. Hofmann, N. C. P. Cross, A. Reiter. Limited duration of complete remission on ruxolitinib in myeloid neoplasms with PCM1-JAK2 and BCR-JAK2 fusion genes. *Annals of Hematology*, 94(2):233-238, 2015.
(i) Equal contribution

J. Kralovicova (i), M. Knut (i), N. C. P. Cross, I. Vorechovsky. Identification of U2AF(35)-dependent exons by RNA-Seq reveals a link between 3' splice-site organization and activity of U2AF-related proteins. *Nucleic Acids Research*, 43(7):3747-3763, 2015.
(i) Equal contribution

Acknowledgements

My supervisors Will Tapper, Nick Cross, Andy Collins, Sarah Ennis as well as Igor Vorechovsky have provided thorough and dependable first-class supervision and assistance where necessary and I am extraordinarily appreciative of this.

Many fruitful conversations with Andy Chase and Jana Kralovicova yielded further insight into the work, for which I am most grateful.

I acknowledge the use of the IRIDIS High Performance Computing Facility, and associated support services at the University of Southampton, in the completion of this work.

This thesis is dedicated to all the teachers, mentors and supervisors I have had throughout my life. This work is yours as much as it is mine.

Chapter 1

Introduction

An increasing number of gene fusions and mutations are being identified in myeloid malignancies. Some of gene fusions are cytogenetically cryptic, evading detection using conventional cytogenetic techniques, suggesting that other methods, such as Next Generation Sequencing (NGS) may be useful for their detection. Genes that are altered by point mutations fall into several functional classes, one of which are genes encoding components of the spliceosome. Understanding the role of these genes in regulating splicing is also amenable to analysis by NGS. This chapter focuses on characterisation of myeloid malignancies, mutations, gene fusions and their role, as well as contrasting fusion identification methods, with particular attention to NGS technology and related bioinformatics.

1.1 Historical perspective

The first records of cancer were made in the times of ancient Greece. *Corpus Hippocraticum*, a collection of works largely attributed to Hippocrates (460-375 B.C.), contains the first mentions of cancer[1]. As such, he can be considered the first person to investigate pathogenesis of malignant tumours and the author of the term "cancer".

Over the course of millennia after *Corpus Hippocraticum*, the incidence of cancer and investigations into its pathogenesis was recorded in various historical sources, ranging from chronicles to medical treaties. Archaeological findings, together with paleopathological investigations help approximate the incidence of cancer over the course of history. The rarity of cancer occurrence up to the late modern period suggests that cancer is a disease with high

impact on modern society, related to the currently omnipresent carcinogens and increased longevity, however this is still disputed, clouded by historical inadequacies of disease diagnosis techniques[2].

Nowadays, cancer is without doubt one of the most common causes of death. With malignant neoplasms claiming lives of up to 306 per 100,000 people in some industrialized countries (Figure 1.1), they are a group of diseases with undisputed importance for modern medicine. Fortunately, modern medicine is not helpless in the fight with cancer. Armed with advancements in the fields chemistry, engineering, and computer science, it has been battling neoplasms with increasing efficiency.

Those advancements allowed a relatively new field from the boundary of informatics and biology to emerge - bioinformatics. The following chapters focus on development and application of bioinformatic methods to investigate the pathogenesis and improve the diagnosis of myeloid malignancies, a subgroup of leukaemias.

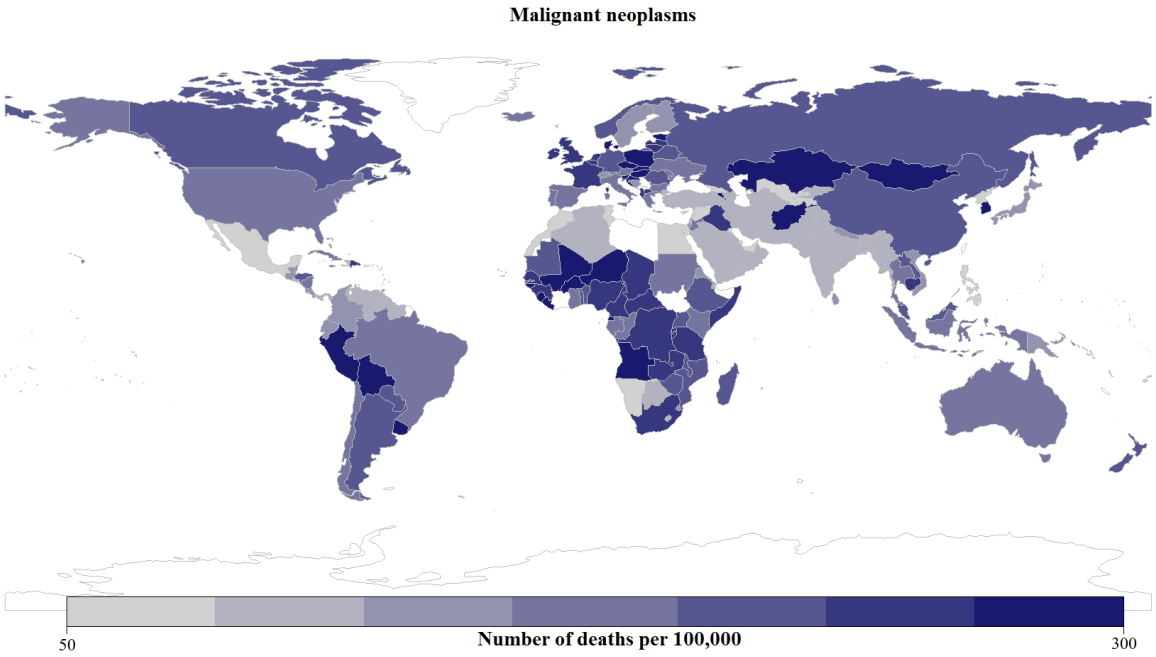

Figure 1.1: Malignant neoplasms, age-corrected number of deaths per 100,000 people, per country, in 2004. Darker colour - more deaths; lighter colour - less deaths; white colour - no data. Data source: [3].

1.2 Myeloid malignancies

1.2.1 Cancer overview

Cancer is not a single disease, but rather a group of diseases, with differing pathology. Most cancers possess the following characteristics: enhanced and potentially limitless replicative potential, tissue invasion and metastasis, sustained angiogenesis, apoptosis evasion, self-sufficiency in growth signals, insensitivity to anti-growth signals[4]. As such, they provide uncontrolled growth and intraorganismal transportation capabilities to cancerous cells.

Cells that do not possess invasive capabilities are considered to be benign, and not classified as cancerous. Their abnormal growth can still be a problem, and can cause a variety of clinical symptoms. Benign neoplasms can also become malignant, acquiring metastatic characteristics.

Cancers can be classified by the originating tissue type, or by primary site of development. Classification by the originating tissue type, also called histological type classification, is commonly used in research. Histological categorisation of myeloid malignancies is shown in Figure 1.2.

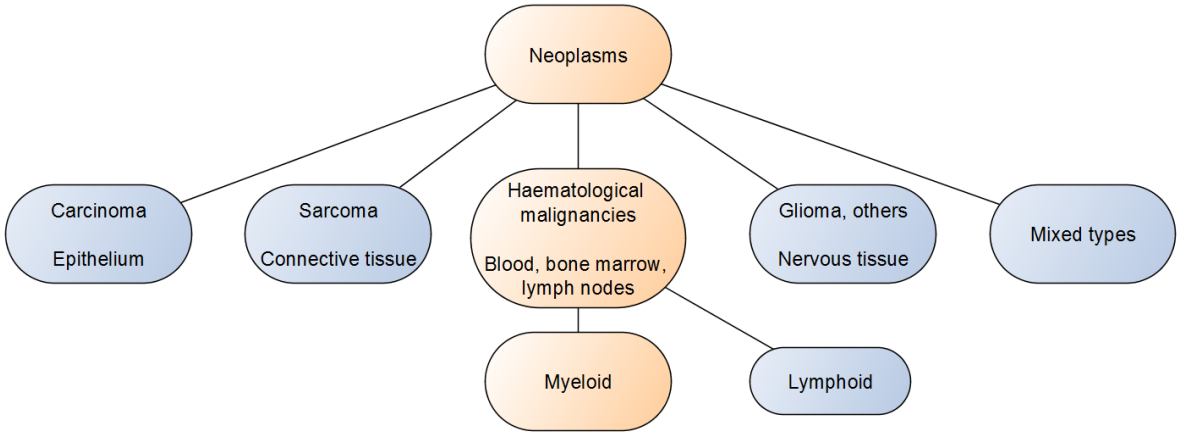

Figure 1.2: Myeloid malignancies histology based group classification. Mixed types examples: teratocarcinoma, carcinosarcoma. Based on International Classification of Diseases for oncology, third edition (ICD-O-3) [5].

Cancers can be divided into five main histological groups, based on frequency of occurrence. Most common cancers affect epithelium, connective tissue, and blood/bone marrow/lymph nodes (haematological malignancies). Cancers originating in different tissues, or in multiple

types of tissues are not as common. Haematological malignancies can be subdivided into two main categories, depending on which cell lineage in haematopoiesis is affected - myeloid or lymphoid.

Myeloid malignancy is an umbrella term for both chronic and acute cancers of myeloid cell lineage. Simplified classification of myeloid malignancies is presented in Table 1.1 with a complete classification presented in Appendix A.

Chronic	Acute
Myelodysplastic syndromes (MDSs)	Acute myeloid leukaemia (AML)
Myeloproliferative neoplasms (MPNs)	
Myelodysplastic/myeloproliferative neoplasms (MDS/MPNs)	

Table 1.1: Classification within myeloid malignancies group [6].

Myeloproliferative syndromes (MPN) are characterised by production of excess cells of myeloid lineage, myelodysplastic syndromes (MDS) are characterised by dysplastic myeloid development. Dysplasia and proliferation in MPN and MDS are typically of a chronic nature, in contrast to acute proliferation of immature blast cells in acute myeloid leukaemia (AML). However, some cases of MPN and MPD progress to AML, often associated with the acquisition of new somatic abnormalities.

Pathology

Myeloid malignancies are a group of clonal diseases driven by transformation of pluripotent stem cells or progenitor cells in a manner that predominantly affects myeloid lineage cells. Presented in Figure 1.3 is a schematic outline of haematopoietic development giving an overview of the myeloid and lymphoid lineages.

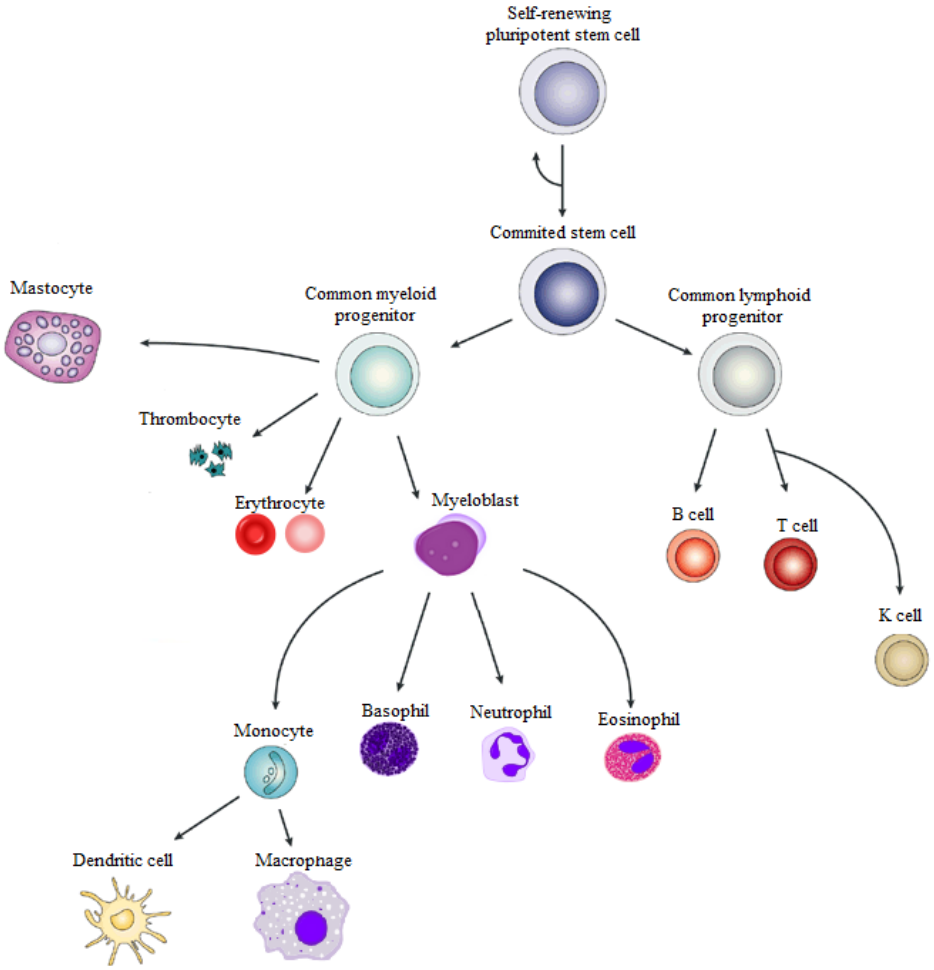

Figure 1.3: Schematic overview of haematopoiesis. Myeloid and lymphoid lineages are shown. Adapted from [7].

Initially, haematopoietic stem cells differentiate into myeloid and lymphoid progenitors, which mature into specific cell types. In myeloid malignancies, somatic mutations are typically acquired at the stem cell level, resulting in clonal expansion of cells down one or more of the myeloid lineages. Although most causal mutations are somatic (nonheritable, arising in non-germ cells), germline predisposition (heritable, arising in germ cells), such as a set of co-inherited point mutations in the region of *JAK2* gene, called the 46/1 haplotype, have

been proven to increase the risk of myeloproliferative neoplasms[8].

As a disease group, there is no one common cause giving rise to all the myeloid malignancies. Instead, there are various mutations in different genes which are often associated with specific disease phenotypes. However, affected genes can be classified into five principal functional groups: signalling pathways components (e.g. *RAS*, *JAK2*, *PDGFRs*), transcription factors (e.g. *CEBPA*, *ETV6*, *RARA*), epigenetic regulators (e.g. *ASXL1*, *IDH1*, *IDH2*, *TET2*), tumour suppressors (e.g. *TP53*), and spliceosome components (e.g. *U2AF1*, *SF3B1*)[6]. Mutations that are implicated in oncogenesis, also called driver mutations, in any of those groups can result in alterations in production of different cells in the myeloid cell lineage. As such, clinical phenotypes such as overproduction of eosinophils (eosinophilia), neutrophils (neutrophilia), basophils (basophilia), mast cells (mastocytosis), erythrocytes (polycythemia), platelets (thrombocytosis) or monocytes (monocytosis) are often observed in patients with myeloid malignancies[9].

In order to diagnose a suspected myeloid malignancy, peripheral blood and bone marrow samples are examined. The number of cells maturing from myeloblasts is assessed by cell counting, and if it exceeds a threshold of 20% (normal being <5%), the disease is diagnosed as an AML. AMLs are further characterised by the specific affected cell (e.g. basophilic, monocytic etc.) and mutations present (e.g. gene fusions such as *PML-RARA*, *AML1-ETO*, or point mutations in genes such as *CEBPA*, *NPM1*). Samples with <20% of blast count are divided into groups using a variety of markers. In order to differentiate between MPNs and MDSs, dysplasia of the cells in bone marrow is investigated, and subgroup assigned on the basis of dysplastic cell type(s) and mutations present. Where there is a lack of dysplasia, the MPN subtype is assigned based on affected cell type and mutation. Those with increased red cell volume along with *JAK2* V617F or similar mutation are categorised as polycythemia vera (PV), a type of MPN. *BCR-ABL1* fusion positive instances are assigned as chronic myelogenous leukaemia, also an MPN subtype. Cases with elevated thrombocyte levels, with normal levels of granulocytes and erythrocytes, along with *JAK2* V617F or other similar mutation, are defined as essential thrombocythemia (ET), another variety of MPN. Another main example of a MPN is primary myelofibrosis (PMF), with presence of collagen fibrosis and megakaryocyte proliferation accompanied by *JAK2* V617F or other clonal marker. It is

possible for a sample to exhibit properties of both MPN and MDS (dysplasia and proliferation), such as in case of chronic myelomonocytic leukaemia [5]. In some cases, elevated levels of specific cells in blood can be the result of an infection, creating a reactive phenotype. Such cases do not present with any mutation, as they are a natural reaction of the organism to a pathogen.

Gene fusions, explained in detail in the next section, are a frequent driver mutation in myeloid malignancies. Overall, 264 different gene fusions are known in haematological disorders [10]. However, the vast majority of them have a very low occurrence rate, with some observed only once. The most widely known and frequently occurring gene fusion, between *BCR* and *ABL1* genes, visible cytogenetically is known as the Philadelphia chromosome. First identified as a recurrent cytogenetic abnormality in 1960 [11], and subsequently characterised at the molecular level in the early 1980s, it was the first gene fusion to be associated with particular type of cancer – chronic myelogenous leukaemia (CML). *ABL1*, which encodes cytoplasmic and nuclear protein tyrosine kinase, is implicated in the processes of cell division and differentiation in healthy cells. *BCR*, encoding a homotetramer scaffolding protein in healthy cells, provides its homotetramer-producing affinities along with regulatory promoter region to the tyrosine kinase ability of *ABL1* in the fusion, creating a chimeric gene with increased tyrosine kinase activity. The importance of this discovery is still clear today, with Imatinib being used as a form of targeted therapy in CML patients. Imatinib prevents *BCR-ABL1* from initiating proliferation signalling pathway by acting as a competitive tyrosine kinase inhibitor on the *ABL1* domain, resulting in apoptosis of the cell [12].

However, *BCR-ABL1* is not the only driver gene fusion in myeloid malignancies. Among the most recurrent ones are *PML-RARA* (AML) and *FIP1L1-PDGFRA* (CEL) [10]. Multiple other chromosomal aberrations, including deletions, inversions, and point mutations all can give rise to myeloid malignancies. Because of this, classification of particular types within the group is problematic, with influx of new information, and updates in classification guidelines released periodically by the World Health Organisation [9], presented in Appendix A.

1.3 Gene fusions

Fusion genes occur when a novel gene or transcript is formed from two previously separate genes or transcripts. In most myeloid cases, the fused gene is composed of exons from both genes.

1.3.1 Importance

Somatically acquired gene fusions are frequent causal mutations in haematological malignancies, and less common in solid cancers. Many leukaemias, including myeloid malignancies, have gene fusions at the root of their pathology [13]. They are most often a gain-of-function type, where a gene fusion produces a working protein or functional RNA with an altered or constitutively activated function when compared to that of its normal type. As such, they are most frequently in-frame, where amino acid coding triplets are not shifted, and the resulting protein often contains functional domains from both proteins. Although loss-of-function gene fusions do occur, their prevalence is lower, and their effect similar to other loss-of-function mutations, i.e. the normal function of the affected gene is lost with no new functional gain. Recent studies[14] suggest that loss-of-function gene fusions occur more frequently than previously expected, although these findings require independent verification.

Due to the potential of targeted drug design for gene fusion products or their partners, fusions are of immense importance to identify and study. With the encouraging results of improved prognosis for some targeted therapies over the past years, it becomes apparent that these therapies offer a step forward in battling neoplasms.

Neoplastic cells with increased replication and diminished apoptosis due to acquisition of 'driver' mutations acquire 'passenger' mutations and it is difficult to distinguish between mutation types [15]. The same applies to gene fusions, after correct identification, it is not necessarily obvious if novel fusions are causal, or just benign passengers. Functional studies using appropriate biological systems are typically required to confirm the causality of any new abnormality.

1.3.2 Overview

In principle, there are two major categories of gene fusions: those arising at the DNA level, and those created at the RNA (transcript) level. The differences are substantial and even though not structurally distinguishable at the protein level, identification of the correct nucleic acid where a fusion originates can influence gene fusion prevalence in cell population, amount of protein product, and resulting morbidity.

DNA-level gene fusions

DNA-level gene fusions are created due to rearrangements within DNA and account for all fusions that are currently known to be pathogenetically important in myeloid neoplasms. Most of these fusions are associated with cytogenetically visible chromosomal rearrangements, particularly reciprocal translocations. Figure 1.4 presents a simplified representation of such fusions, showing their origin.

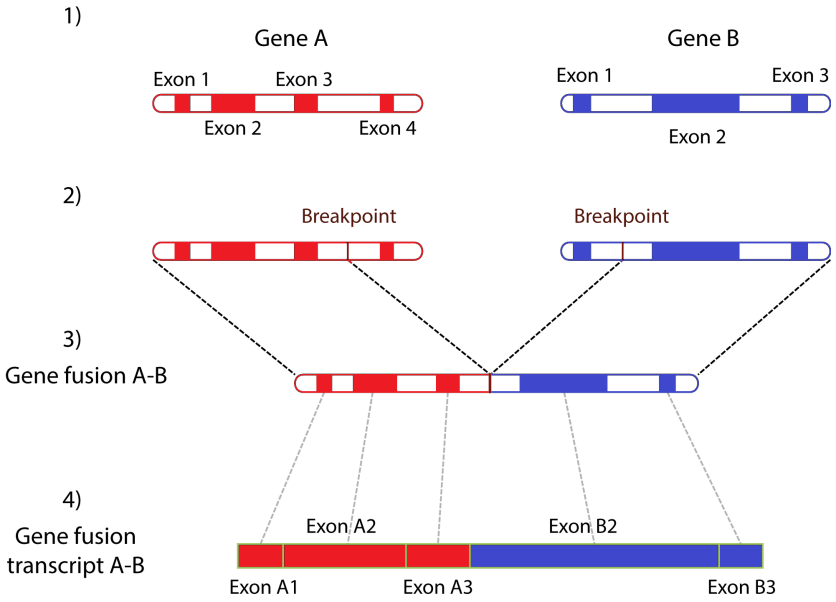

Figure 1.4: Model gene fusion. 1) Wild-type genes A (red) and B (blue); 2) Introduced breakpoints in genes A and B; 3) Gene fusion between genes A and B; 4) Resulting transcript from gene fusion between genes A and B.

In the example shown in Figure 1.4 DNA-level gene breakpoints within intronic regions can be observed. While it is not necessary for the breakpoints to be situated within introns, those arising in exons are less common, presumably in part because exons present much smaller targets for breakage. It is also possible for them to occur within intergenic regions, however

such breakpoints do not create gene fusions. Breakpoints are created due to a variety of different mutations which are described in the next section.

Chromosomal translocations

Gene fusions caused by chromosomal translocations are by far the most frequently observed to date in leukaemia and occur due to the exchange of segments between chromosomes. While chromosomal translocations in lymphoid malignancies are suspected to be mediated mostly by recombination-activating genes (RAGs) complex, and by activation-induced deaminase (AID), the cause of translocations in myeloid malignancies is unclear. Some of them are suspected to occur as a result of Alu repeats recombining with other Alu repeats on different chromosomes [16], but examples of homologous recombination are very rare. It has been also shown that chromosomal rearrangements can be induced by free radicals [17] as well as related to palindrome-mediated genomic instability[18]. Usage of the microhomology end joining pathway (MHEJ), an alternative to the non-homologous end joining (NHEJ) pathway, has also been shown to induce translocations, however the process in not very well understood [19]. An example of a translocation mediated reciprocal gene fusion, whereby one fusion chromosome has a fusion of genes A and B while the other has a fusion of B and A is presented in Figure 1.5. Reciprocal fusions are often created by chromosomal rearrangements.

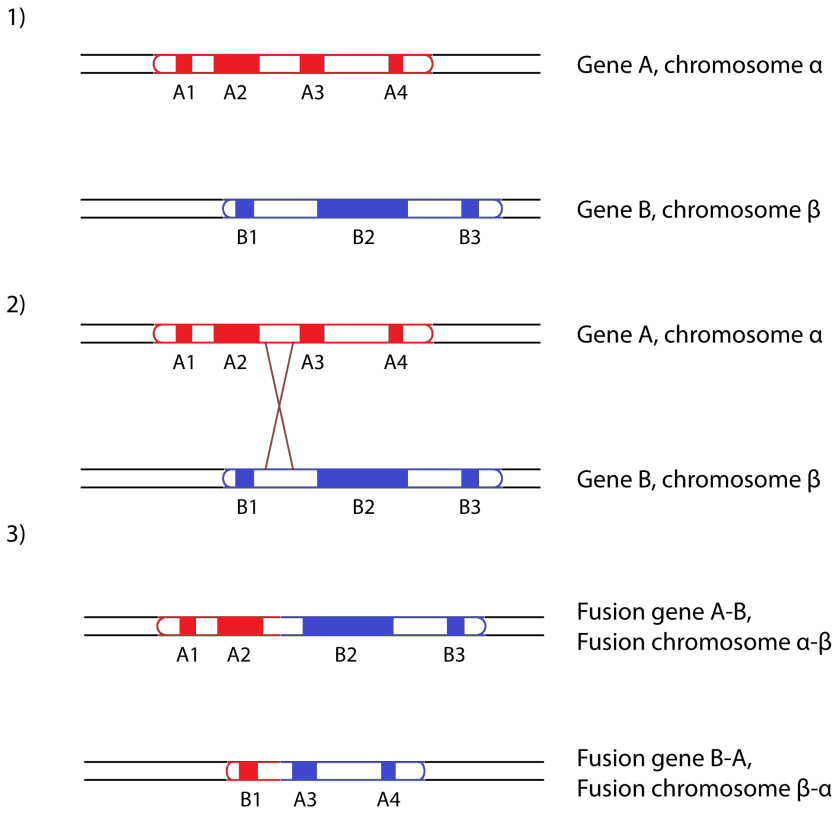

Figure 1.5: Example of reciprocal gene fusion arising due to chromosomal translocation. 1) Wild type genes A (red) and B (blue), located on different chromosomes α and β. 2) Translocation site between chromosomes, situated within genes shown. 3) Post-translocation A-B gene fusion on fusion chromosome $\alpha - \beta$, and B-A gene fusion on chromosome $\beta - \alpha$.

Chromosomal translocation breakpoints can occur anywhere in chromosomes. Only those that create an open reading frame (ORF) containing more than one gene in relatively close proximity within the resulting aberrant chromosome will produce a gene fusion. The breakpoints usually occur within both fusion partners, but this is not necessary. It is possible for a breakpoint to occur within one gene, and if the second breakpoint occurs in close proximity to another gene, a gene fusion may be created by alternative splicing. In such instance, it would be possible for the second partner to be preserved, e.g. the t(12;13) in AML results in an *ETV6-CDX2* fusion and *CDX2* overexpression [20]. Reciprocal translocations in many cases of cancers are expected to generate two products, e.g. *BCR-ABL* and the reciprocal *ABL-BCR*. However, most commonly only one of the fusions is an oncogenesis driver, in this case *BCR-ABL*.

Gene fusions arising due to chromosomal translocations are the most common variety in myeloid neoplasms. There are many notable examples, such as *BCR-ABL* [11], and *PML-RARA*[10]. As such, this type of fusion-creating event is routinely screened for in the clinical environment by a combination of cytogenetics and RT-PCR.

Deletions

Gene deletions are known to arise due to a variety of reasons. In the majority of cases, errors during recombination (unequal crossing over) in mitosis and meiosis are the cause of deletions. It is also possible for a deletion to occur as a result of chromosomal translocation. An example of a gene fusion arising due to a deletion-type mutation is presented in Figure 1.6.

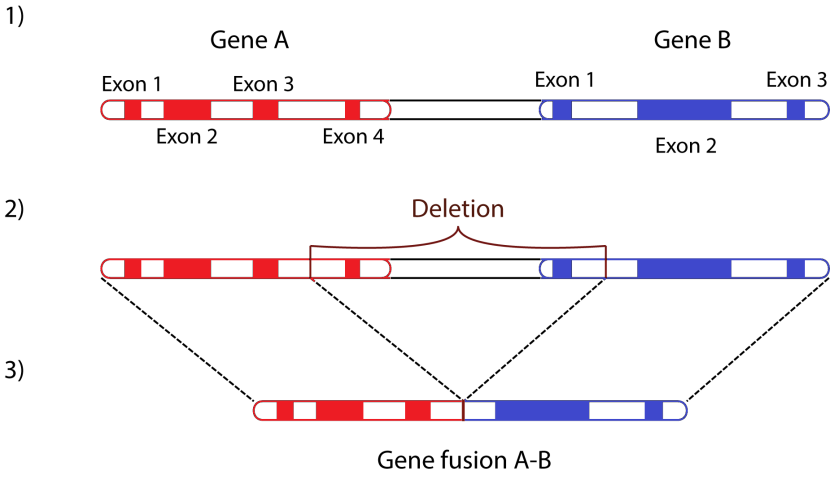

Figure 1.6: Example of gene fusion arising due to a deletion. 1) Wild-type genes A (red) and B (blue) in close proximity on the same chromosome. 2) Deletion covering last exon of Gene A and first exon of Gene B marked. 3) Resulting gene fusion A-B.

Deletions can create gene fusions by removing the stop codon from one gene and the start codon from the other, bringing the reading frame of the second gene within that of the first gene.

However, removal of start codon from the second fusion partner is not necessary. The loss of the stop codon of the first fusion partner may allow it to facilitate transcription of a conjoined product. The absent stop codon will not provide the transcription machinery with a stop signal and transcription may continue into the second fusion partner.

Due to limitations in the operational distance of the transcription machinery, fusion-producing deletions usually result in close juxtaposition of elements that create the fusion gene. Deletions that remove a stop codon from a gene, but do not introduce a potential partner within its vicinity do not usually result in a gene fusion, since the transcription machinery is likely to either encounter a random stop codon or detach from the DNA before reaching the next potential partner.

It is important to underline that when a gene fusion arises due to the deletion of a stop codon without directly affecting its partner, the partner gene can be completely unaffected and produce regular transcripts, as well as be a part of a fusion gene. This can occur under the stipulation that the deletion does not affect regulatory regions of the second partner.

Examples of gene fusions arising from deletions, are *CLR-CLEC2D* [21] and *FIP1L1-PDGFRA* in chronic eosinophilic leukaemia. *CLR* and *CLEC2D* are normally separated by 31kb, whereas *FIP1L1* and *PDGFRA* are separated by 851kb on the same chromosome necessitating a large deletion as the most likely cause of their respective fusions.

Duplications

Gene duplications represent another mechanism by which gene fusions are produced. Similarly to gene deletions, they are mostly caused by errors in recombination. A representation of gene duplication giving rise to gene fusion is presented in Figure 1.7.

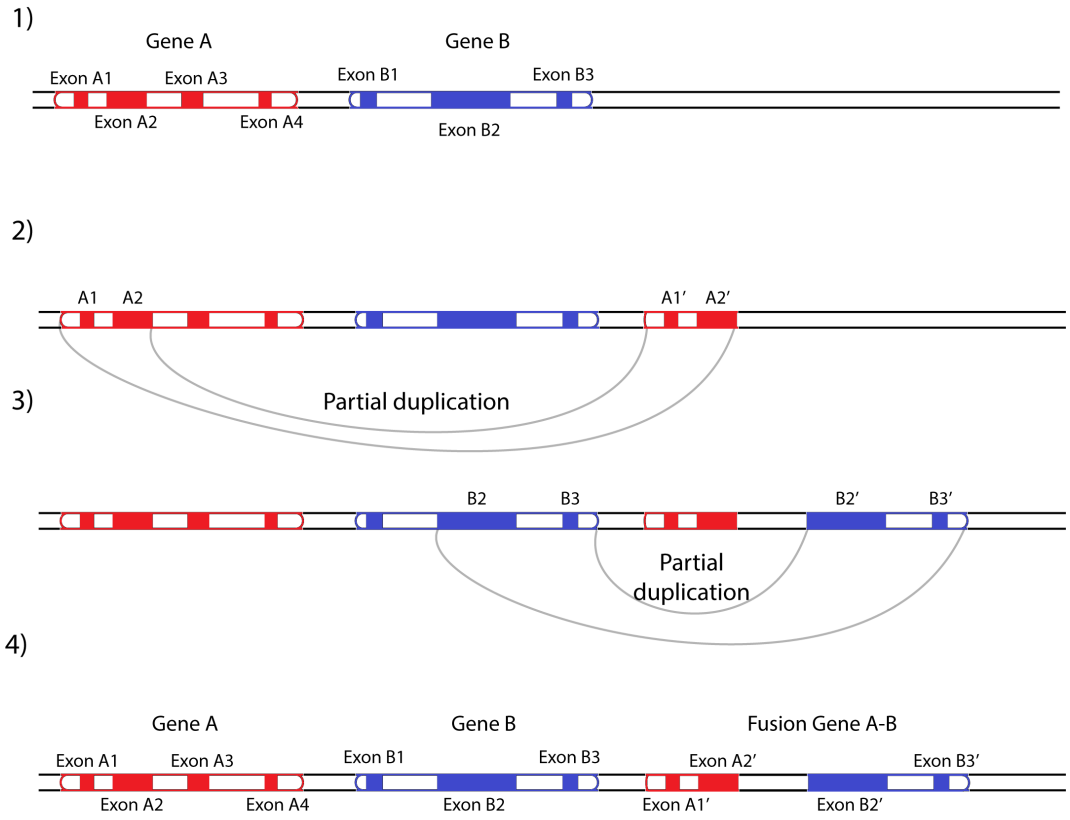

Figure 1.7: Example of gene fusion arising due to duplication; 1) Wild-type genes A (red) and B (blue) located in close proximity on the same chromosome; 2) Partial duplication of Exon 1 and 2 of Gene A; 3) Partial duplication of Exon 2 and 3 of Gene B; 4) Resulting A-B gene fusion with wild-type Gene A and Gene B preserved.

While the example shows double partial duplication and juxtaposition creating a gene fusion, it is not necessary for both genes to undergo incomplete duplication in order to create a gene fusion. Partial duplication can introduce part of a gene that does not possess a stop codon into the proximity of another gene, but not within it, effectively creating a fusion between them. Another type of duplication causing gene fusion would be a duplication event situating a duplicated gene or fragment of one gene within another, creating a fusion event.

As opposed to deletions, where at least one fusion partner gene is dysfunctional, duplications may create a gene fusion with preservation of both partners. Tandem duplications may create

reciprocal fusions, where both A-B and B-A gene fusions are present.

Identification of gene duplication events that give rise to gene fusions can be problematic since they are indistinguishable from normal products in methods such as Sanger sequencing. As such, it is likely that some fusion events described in the literature as deletion-caused, actually originate from duplication event(s). A clear example of a duplication causing a gene fusion is *KIAA1549-BRAF*, the main cause of pilocytic astrocytoma, a brain neoplasm[22]. A more complex example, found in healthy people, is *TFG-GPR128*[23].

Inversions

Another type of mutation that can create gene fusions is inversion. Inversions are usually caused by double breakage in a chromosome, where the chromosome part between breakages is re-inserted into its chromosome in a wrong orientation. An example of gene fusion caused by inversion is shown in Figure 1.8.

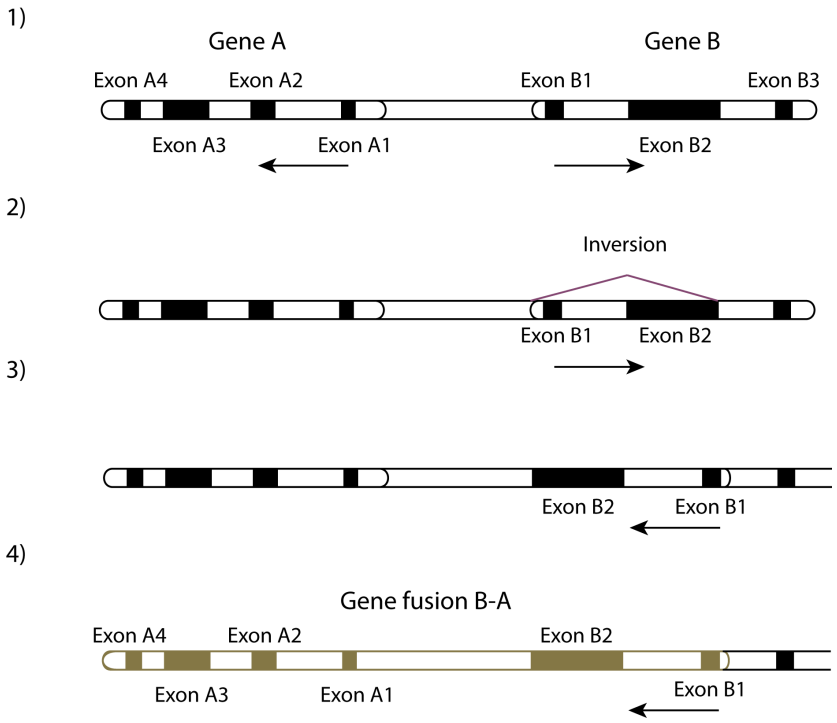

Figure 1.8: Example of gene fusion arising due to inversion. 1) Wild-type Genes A and B in close proximity on the same chromosome with opposite directionality; 2) Inversion area marked in Gene B; 3) Gene B post inversion. Note the directionality change; 4) Highlighted B-A gene fusion.

This example shows a simple inversion where part of one gene has its start codon along with other features inversed, changing its orientation from positive to negative strand and continuing into the second partner. This is not a prerequisite to create an inversion-caused gene fusion, multiple other possibilities exist. Both partners can also be inverted, to produce an ORF that covers both partners. Similarly to deletions, the gene partners have to be close enough to produce a fused transcript. There is a possibility of preservation of one of the partners in inversion-created fusions, but the other will always be affected.

Gene fusions arising due to inversions have been previously observed in myeloid malignancies. One example is the inv(11)(q14q23) causing *MLL-CALM* fusion, reported as a rare, but recurrent event in infant AML[24].

RNA-level gene fusions

In contrast to DNA-level gene fusions, RNA-level gene fusions have no genomic deletions or rearrangements. Changes arise at the transcriptome level, producing transcripts that are combinations of at least two different genes. To date, no RNA-level gene fusions have been identified as causal in myeloid malignancies, but that may be due to the fact that they have not been systematically investigated.

Transcription read-through

One of the most common RNA-level gene fusions is due to transcription read-through, an example of which is shown in Figure 1.9

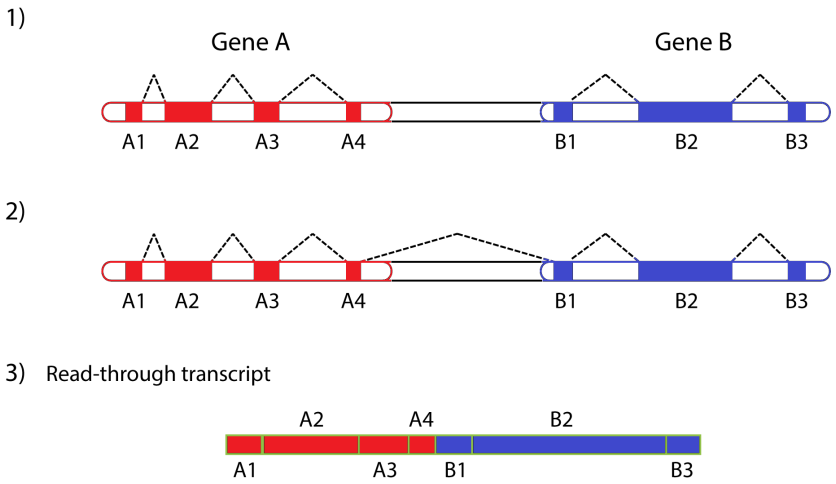

Figure 1.9: Example gene fusion arising due to transcription read-through. 1) Wild-type Genes A (red) and B (blue), located in close proximity on the same chromosome with their regular splicing indicated by dashed line; 2) Indication of read-through transcript splicing across genes A and B; 3) Read-through transcript A-B.

Read-through transcripts arise when the transcription machinery ignores the regular polyadenylation signal (polyA) and continues into a neighbouring gene. This will usually result in a conjunction of regular isoforms of the fusion partners. The occurrence is common, and not necessarily detrimental, as some genes are normally co-regulated in this manner.

Transcription read-through does not normally affect the normal function of either of the genes, but it may be the driver of a malignancy, although this is less likely than gene fusions caused by other mechanisms. This category of fusion-creating event is widely observed in healthy individuals, where 3% of genes have been observed to be involved in the process,

and is thought to be an evolutionary mechanism allowing a larger variety of proteins and functional RNAs to be produced, with some resemblance to alternative splicing[25].

Trans-splicing

Trans-splicing is a mechanism that is largely uninvestigated, but is suspected to have an important role in human embryonic stem cell pluripotency[26]. It is possible for the spliceosome to erroneously introduce a part of a transcript of one gene into that of a different one. An example of such phenomenon is presented in Figure 1.10.

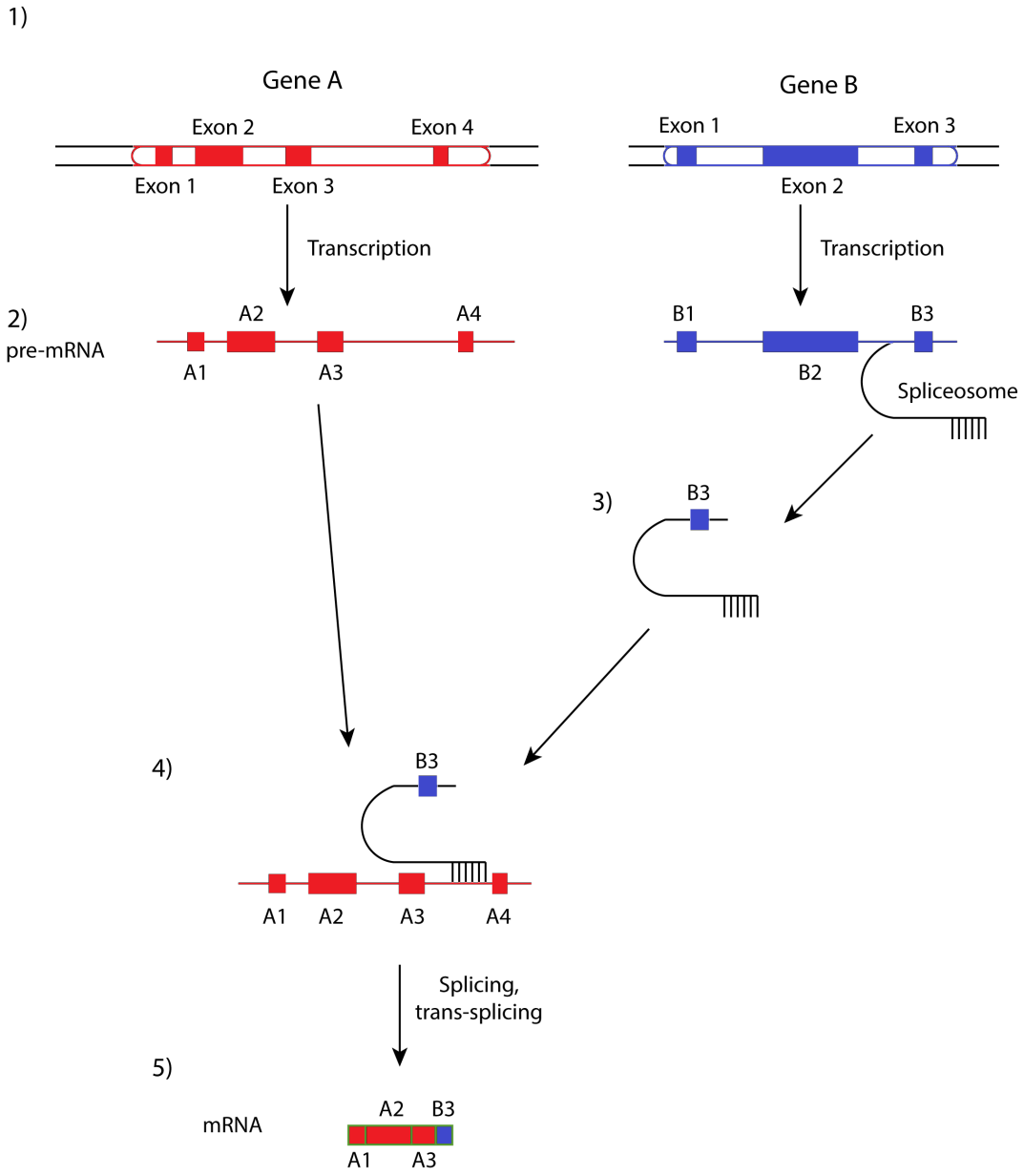

Figure 1.10: Example of gene fusion arising due to trans-splicing. 1) Wild-type Genes A (red) and B (blue); 2) precursor messenger RNA (pre-mRNA) products of Genes A and B; Note spliceosome bound to intron 2-3 in Gene B; 3) Spliceosome carrying exon 3 from Gene B; 4) Spliceosome with exon 3 from Gene B binding to intron 3-4 of Gene A; 5) messenger RNA (mRNA) fusion product of Exons 1,2 and 3 from Gene A and spliceosome-introduced Exon 3 from Gene B.

Trans-splicing occurrences creating gene fusions are relatively rare events, with probably limited significance. It is possible for this mechanism to produce a transcript that contains exons from virtually any gene, without affecting either DNA sequence or genes involved. Fusion-creating trans-splicing is still being investigated, with some evidence that it can create gene fusions identical to those that are known and recurrent oncogenesis drivers, such as *JAZF1-JJAZ1* in endometrial stromal tumours [27].

Other fusion events

While very rare, it is possible for a gene fusion containing not only two, but three fusion partners to occur. It is very unlikely for such event to arise, as it requires two separate gene fusion creating processes to occur within a small region, with the second process fusing a gene fusion to a third partner. An example is *BCR-RALGPS1-ABL1*, an insertion between *BCR-ABL1* fusion site in a patient relapsed to ALL [28].

Any of the described DNA-level gene fusion creating mutations may create a fusion which will not use the pre-existing start or stop codon. Instead, they may create novel codons, although those would be most likely loss of function mutations. This is especially important for any assumptions made when analysing gene fusions. Furthermore, because of creation of those features, partners in a fusion gene can have completely changed directionality, i.e. a mutation may occur where gene partners on the positive strand will create a fusion on a negative strand. Such events add complexity to the study of gene fusions.

1.3.3 Determination methods

Since the discovery of Philadelphia chromosome in 1960 [11], its subsequent molecular characterisation, and with the development of new investigative techniques, it is now possible to search for gene fusions and other abnormalities at a genome wide level and with unprecedented efficiency. Current methods may be divided into two categories : clinical diagnostic, which are used in clinical setting to refine diagnosis, and experimental, which are under development and are not currently used for clinical purposes.

Clinical diagnostic methods

Karyotyping

The oldest method of gene fusion identification is cytogenetic karyotyping, developed in 1970 by Caspersson et al. [29]. It is performed by banding chromosomes in metaphase by application of trypsin and Giemsa staining. An image from karyotyping showing the Philadelphia chromosome is presented in Figure 1.11.

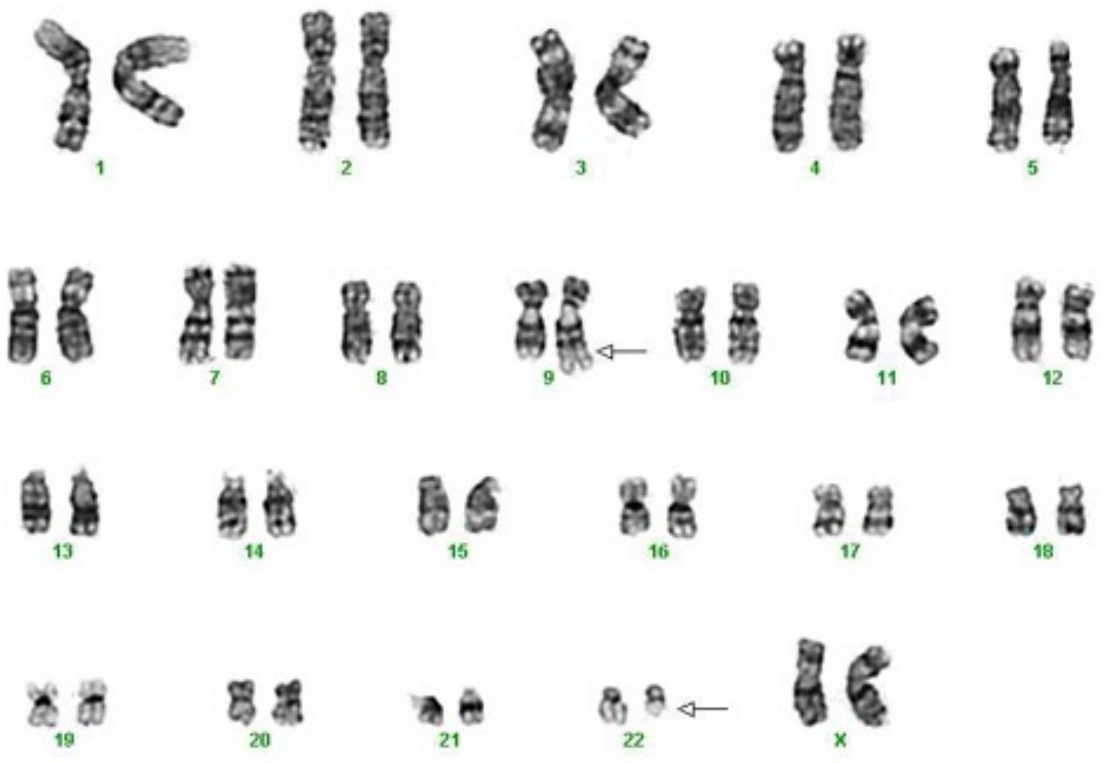

Figure 1.11: Philadelphia chromosome t(9;22)(q34;q11) identified by cytogenetic karyotyping [30]. Arrows indicate breakpoints in chromosomal translocation between chromosome 9 and 22.

Cytogenetic karyotyping, while cost-effective and used routinely, has a set of limitations for identifying gene fusions. Firstly, it can only detect only chromosomal translocations and deletions bigger than 1.5 Mb [31]. Secondly, it is not clear whether a gene fusion has been created and what genes are involved. Lastly, even large abnormalities can be cryptic or missed in imaging due to large similarities between chromosomal regions and human error.

RT-PCR

Reverse-transcription polymerase chain reaction (RT-PCR) is a method of amplifying particular region of RNA after it has been reverse-transcribed to complementary DNA (cDNA). Due to instability of RNA, PCR cannot be performed on it directly. RT-PCR products are subjected to gel electrophoresis in order to distinguish between amplified fragments of different size and to confirm the presence of a product [32]. RT-PCR can be designed to amplify a specific gene fusion junction, and presence of a band corresponding to the product confirms existence of the gene fusion investigated. An example of RT-PCR gene fusion product electrophoresis image is shown in Figure 1.12.

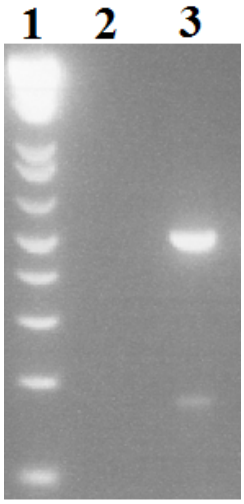

Figure 1.12: Reverse-Transcription Polymerase Chain Reaction (RT-PCR) amplified *BCR-JAK2* fusion junction after gel electrophoresis. 1) 1kb ladder; 2) Negative Control; 3) Amplified *JAK2-BCR* gene fusion.

Since RT-PCR targets specific candidate gene fusion, it can only be used to check for known fusions, or as a confirmatory method for fusion candidates detected by other experimental methods. Although this method can relatively precisely pinpoint fusion breakpoint, it is usually followed by Sanger sequencing (see below) if exact breakpoint identification is required.

RT-PCR was used for confirmation of gene fusion events in Research Chapters 4 and 5.

FISH

Fluorescent in-situ hybridization (FISH) investigates the presence of specific sequences within the genome using fluorescent probes. It is based on a binding of the fluorescently-

labelled probe to its complimentary target within the genome. It is widely applied in gene fusion determination [33]. An example of a gene fusion identified by FISH probes is presented in Figure 1.13

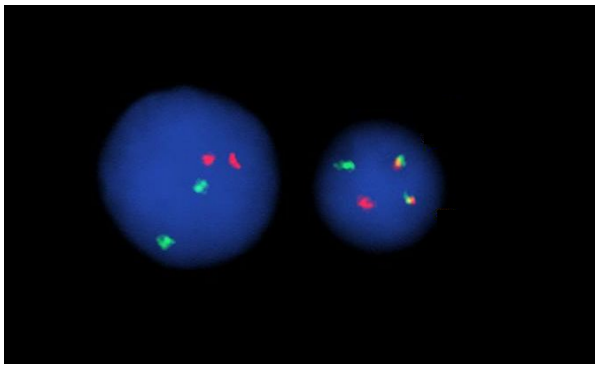

Figure 1.13: Fluorescent In-Situ Hybridizastion (FISH) with dual colour probes showing *BCR-ABL1* gene fusion. Green - *BCR* and its neighbouring gene on chromosome 9. Red - *ABL1* and its neighbouring gene on chromosome 22. Close proximity of red and green probes indicates a fusion between *BCR* and *ABL1* genes. Left - normal, right - fusion. Adapted from [34].

To investigate gene fusions, FISH requires fusion candidates, since probes need to be specifically designed to target genes within neighbourhood of the suspected fusion. This method is therefore used to determine the status of known fusions and to validate candidates that have been discovered by other methods. The resolution of FISH is limited to the gene level, it is therefore usually followed by RT-PCR or Sanger sequencing if there is need to determine exact fusion breakpoint.

Sanger sequencing

Sanger sequencing is historically the first method for DNA sequencing [35]. It is based on the addition of deoxynucleoside triphospates (dNTPs) in a controlled replication reaction. dNTPs are usually fluorescently labelled, which allows for automated, ordered reading of dNTPs incorporated in the reaction, which translates to sequence reading. An example Sanger sequencing readout showing a gene fusion is shown in Figure 1.14

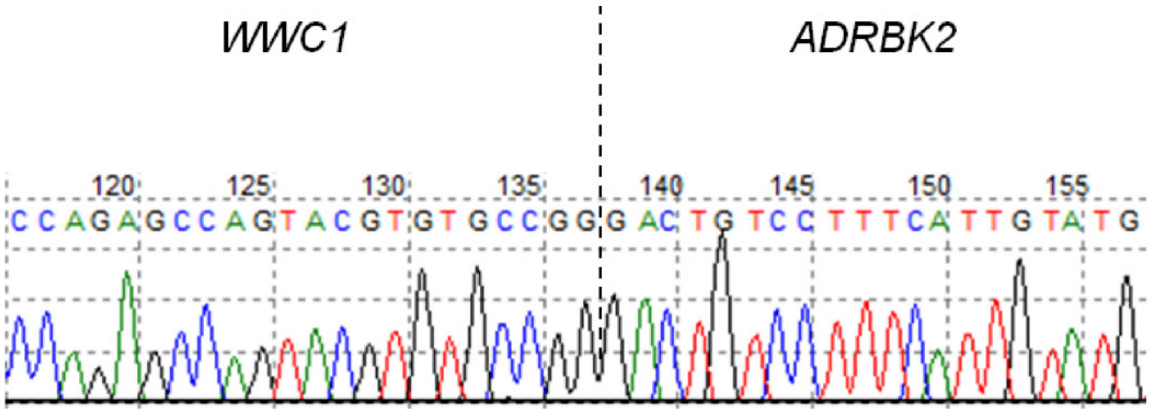

Figure 1.14: Sanger sequencing reads and their intensities indicating a gene fusion between *WWC1* and *ADRBK2* in breast cancer. Read on the left side of the dotted line belongs to *WWC1*, the right side belongs to *ADRBK2*. Observing them together as a single read indicates the presence of a gene fusion between *WWC1* and *ADRBK2*[36].

While being very accurate, Sanger sequencing is not without drawbacks. Fluorescent signals can sometimes be unclear, particularly in long runs of single nucleotide (nt). The method also has to have a precise candidate site, as runs are approximately 800nt long. As such, it is a great method for precise identification of a known fusion breakpoint, or as a confirmatory method for candidates identified by other methods.

Sanger sequencing was used for confirmation of gene fusion events in Research Chapters 4 and 5.

Experimental

Microarrays

Microarrays are arrays of probes designed to bind to sheared, fluorescently labelled genomic DNA or cDNA fragments. Each group of probes on a microarray is programmed to bind specifically to particular fragment. Specific binding of labelled fragments to probes produces a fluorescent signal corresponding to the presence of a fragment. Signal intensities across groups of probes vary according to the amount of labelled fragments that bind to them [37]. Probe design can be adapted for specific needs, with a typical microarray having from a few thousand to millions of probes. An example showing probe intensities in a gene fusion investigated using RNA expression microarrays is shown in Figure 1.15

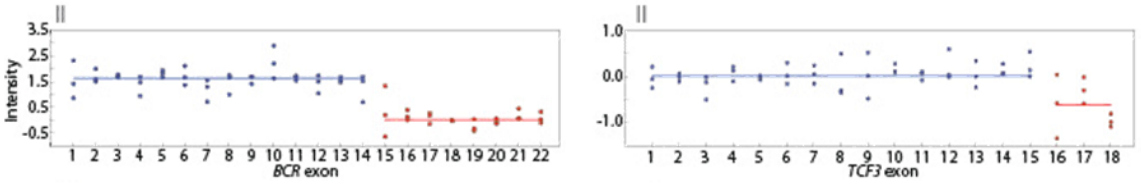

Figure 1.15: Microarray expression probes in *BCR* and *TCF3* exon showing an abrupt drop in intensity which corresponds expression level and is characteristic of a gene fusion between *BCR* & *TCF3*. Note the drop in intensity between exon 14 and 15 of *BCR* as well as between exon 15 and 16 in *TCF3*, which indicates fusion breakpoints between the probes surveying those regions [38].

Microarrays have a variety of different applications. They are widely used as single nucleotide polymorphism (SNP) panels, in genome-wide association studies (GWAS). Although their initial application was restricted to DNA investigation, due to an increase in fluorescent signal sensitivity, they can be used to quantify cDNA abundance for gene expression estimation. Their application for gene fusion identification, is being developed for clinical use. Considering that microarrays can validate the presence of particular sequence and quantify the proportional abundance of genetic material at the same time, they can in principle be applied for fusion determination in two ways. First, probes can be designed to target previously observed fusion breakpoints or candidate breakpoints at the mRNA level. In this case occurrence of probe signal intensity will confirm a fusion. The second option is to design probes to target each exon in cDNA or each exon within a candidate region and generate expression profiles [38]. Abrupt drops of signal intensities between exons within a gene are suggestive of a fusion, but may be caused by alternative splicing. The example presented in Figure 1.15 demonstrates this method and an abrupt drop in the expression of candidate genes suggesting the occurrence of a fusion between them .

Microarrays are very precise when used in a breakpoint-targeted approach, but are susceptible to false-negatives when the breakpoint is shifted compared to what probes are designed for. The accuracy of array based analysis is comparable to RT-PCR and may be followed by Sanger sequencing if greater refinement is required.

NGS

Next Generation Sequencing (NGS), a technique of massively parallel sequencing runs, can be applied to investigate gene fusions [39]. DNA or cDNA is fragmented and pseudo-random primers are used for amplification of genetic material. Different NGS technologies use different approaches to determine the sequence of amplified fragments. A single read, depending on the NGS technology used and the depth of sequencing, can be between 35 and 800 nucleotides, and the total number of reads can vary from tens to hundreds of millions. Results of an example NGS application for gene fusion identification, RNA-Seq, are shown in Figure 1.16.

Figure 1.16: RNA-Sequencing (RNA-Seq) reads covering *BCR-JAK2* fusion junction fusion, as detected by RNA-Seq analysis. Note that high resolution of NGS allows to precisely pinpoint the transcript fusion site with high confidence.

Due to the high number of sequence reads, NGS data output must be analysed using high specification computers. As such, NGS is not without drawbacks, since a consensus universal approach towards analysis of such high quantities of biological data has not yet been achieved. Still, NGS provides much greater flexibility for analysis because candidate regions for gene fusion investigation are not required, and, using data mining, it is possible to identify them with no prior knowledge.

The accuracy of NGS is very high, it can provide base pair resolution of a gene fusion, similar to Sanger sequencing. As an experimental method, gene fusions identified by NGS are usually confirmed using well-established clinical methods, such as RT-PCR and Sanger sequencing.

1.4 Bioinformatics for gene fusion detection

Identification and analysis of gene fusions using NGS technologies requires extensive application of bioinformatic methods. This subsection introduces bioinformatics as a field with particular focus on NGS and gene fusion identification.

1.4.1 Introduction

Bioinformatics is an interdisciplinary field which applies broad sense informatics in biology. The term was coined by Paulien Hogeweg and Ben Hesper in the 1970s, to describe "the study of informatic processes in biotic systems"[40], which is close to its current description, although Hogeweg did not predict such widespread application of computers and associated informatics to study biological information.

The field started to emerge in 1990s, when the race to sequence the human genome began. Two main initiatives to sequence the human genome, public - multinational academic led by International Human Genome Sequencing Consortium, and private - Celera Genomics from USA, led by Craig Venter, announced their successes and published results in the same week in 2001 [41][42]. With such an abundance of data, it was becoming obvious that the lack of computer applications and automated data analysis was hampering DNA studies. Over the following years bioinformatics was growing along with advancing sequencing technologies. With the introduction of NGS in mid-2000s by 454 Pyrosequencing, there was a massive increase in the amount of sequence data. Sequencing was becoming less expensive, and considering how vastly unknown the field was at the time, more people started contributing to the analysis, further developing the field of bioinformatics. With competition-fueled NGS technology development for the elusive "$1,000 genome" [43], an honourable goal which would potentially allow wide-scale clinical application of NGS, bioinformatic applications grew in parallel. The race to rapidly produce high quality and affordable sequence data approached the point when nowadays it is the bioinformatic analysis of data that is the bottleneck in research[44], with the problem coined in the phrase "$1,000 genome, $100,000 analysis"[45].

Analysis of NGS data requires substantial computational power and qualified personnel. The analysis is becoming a problem not only because of shortage of experts, but also due to computational limitations. In 1975, G. Moore made a prediction that the number of the components per chip, which translates to computing capabilities of processors, would double every 12 months (later adjusted to 18 months) [46]. The prediction was accurate, with the trend of transistor number per chip doubling almost every year and a half throughout the last decades. However, in the context of rapid development of NGS technologies even that does

not seem to be enough, as shown in Figure 1.17.

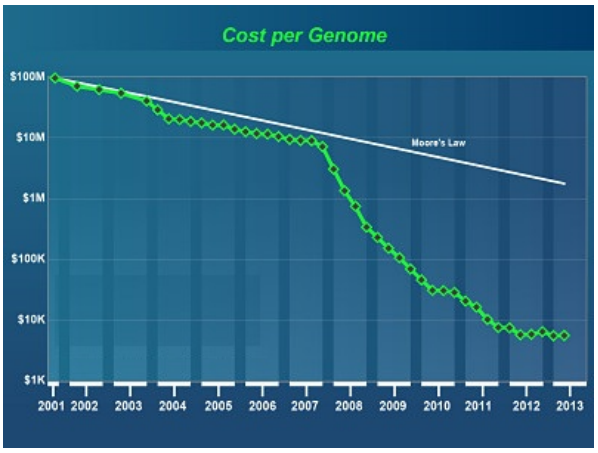

Figure 1.17: Cost of sequencing a genome vs Moore's law. Source: [47]

Even though NGS promoted rapid growth of bioinformatics, the field does not consist of only sequencing data analysis. Other computational applications to biological data are also classified as bioinformatics. Investigation of phylogenetic information, protein structure prediction, and systems biology are all part of the field, somewhat eclipsed by the NGS boom.

1.4.2 Next Generation Sequencing

NGS technologies can be divided into three main categories, depending on the type of signal they record in DNA polymerization reaction. The basic polymerization reaction can be represented by the following equation:

$$(dNMP)_n + dNTP \rightarrow (dNMP)_{n+1} + PPi + H^+$$

$dNMP$: deoxynucleoside monophosphate

$dNTP$: deoxynucleoside triphosphate

PPi : pyrophosphate

H^+ : proton

Currently, the main strategies can be divided according to their ability to record the last nucleotide in obtained $(dNMP)_{n+1}$ (1), the produced pyrophosphate (2), or the emitted proton (3).

1. Detection of the last nucleotide in $(dNMP)_{n+1}$, used by: Illumina (Solexa) in their widely used sequencing-by-synthesis technology [48]; Pacific Biosciences[49] in single molecule real time sequencing used for obtaining long reads of up to 30,000 bases; Complete Genomics in DNA nanoball sequencing, a technology producing short reads of up to 70 bases at decreased cost when compared to other technologies[50]; now bankrupt Helicos Biosciences, in their expensive and inaccurate true single molecule sequencing [51]; Life Sciences in sequencing by oligonucleotide ligation and detection (SOLiD), inexpensive at the cost of producing short reads [52]

2. Pyrophosphate detection, used by: Roche in now discontinued 454 Pyrosequencing, which had quality issues at lower cost[53]

3. Proton detection, used by: Life Technologies in Ion Torrent technology, main competitor of Illumina's sequencing-by-synthesis, with similar cost and efficiency[54]

Although many different NGS technologies exist, all have limitations. Proton detection technologies encounter problems with homopolymer runs, and base calling quality. Pyrophosphate detection technologies are very expensive, which is prohibitive for most experiments. Nucleotide detection methods have some problems with homopolymer runs, but are relatively cheap and have few other issues. As such, they are the most popular NGS methods, with Illumina having 60% of the world's sequencing market[55]. After the conclusion of this project, Pacific Biosciences has improved their platform's error-rate to the point where gene fusion investigation is viable [56]. Their long-read sequencing can be a huge advantage over short-read sequencing to any scientist investigating gene fusions.

Over the past years a sequencing technology based on detection of nucleotides passing through an array of nanopores has been under development, e.g. by Oxford Nanopore. The main difference between NGS technologies and nanopore sequencing is the direct detection of nucleotides in case of nanopore seqencing by measuring the changes in current as different nucleotides pass through [57]. The technology is not without limitations such as high error rate and restricted number of molecules read [58]. Still, it seems promising and might be the next technology to supersede NGS in sequencing. At the beginning of this project, the technology was not yet available, and at the end of the project first publications started to

appear that show single-molecule based sequencing resolution of isoforms in complex mRNA [59]. This can be potentially applied to gene fusion investigation.

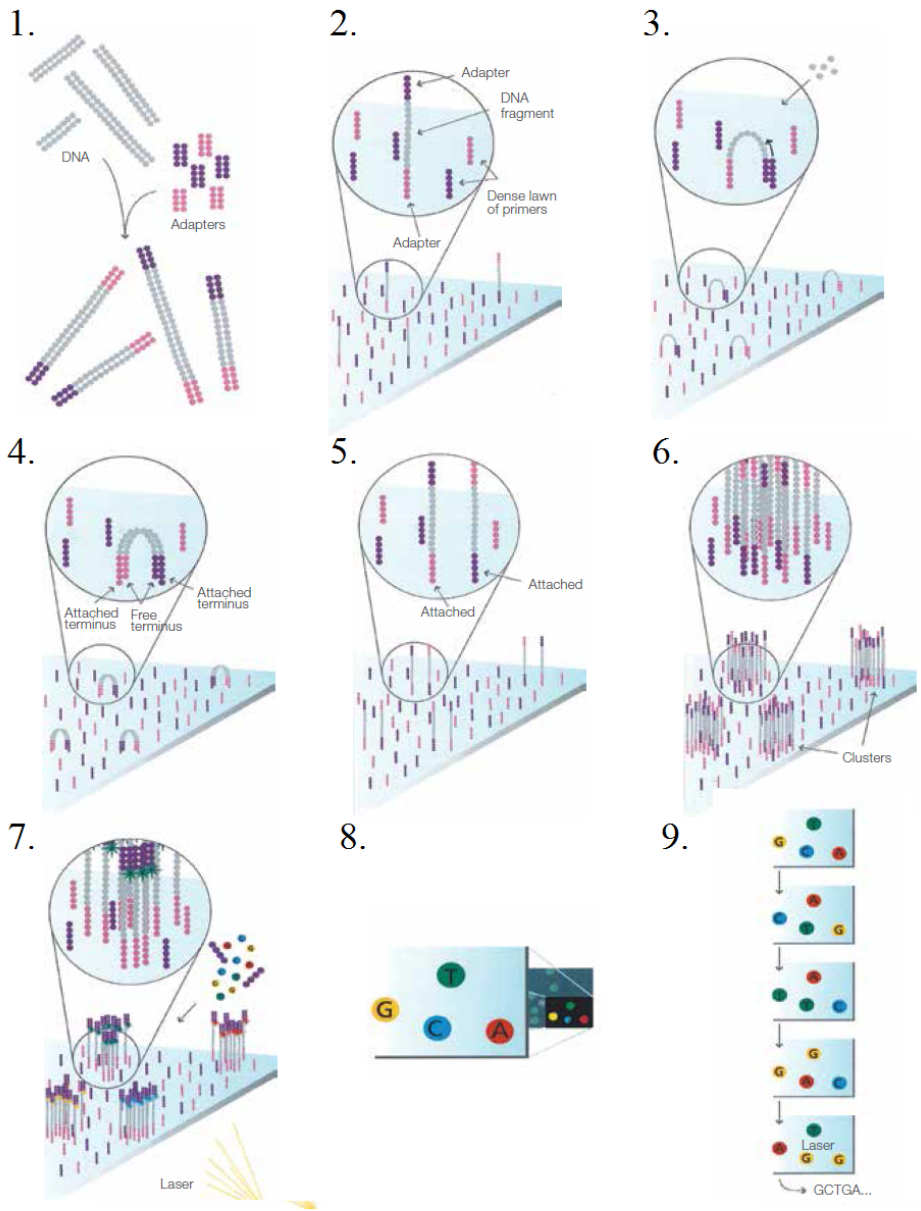

Figure 1.18: Illumina single-end sequencing by synthesis overview. 1. Adapter ligation to randomly fragmented DNA/cDNA. 2. Random binding of single-stranded fragments to the inside surface of the flow cell channels. 3. Addition of unlabeled nucleotides and enzymes for initiation of bridge amplification. 4. Incorporation of nucleotides and double-stranded bridge formation. 5. Bridge denaturation and single-stranded templates anchored to the surface. 6. Upon repetition of steps 3-4, dense clusters are generated. 7. Synthesis by addition of labelled reversible terminators, primers and polymerase. Laser excitation of the labelled terminator. 8. Capture of emitted fluorescence emitted by labelled terminators in clusters. 9. Massively parallel read of emitted fluorescence from millions of clusters. Source: [60]

Illumina's sequencing by synthesis with bridge amplification is presented in Figure 1.18. It starts with pseudo-random fixation of sheared DNA/cDNA fragments onto a surface. Frag-

ments are then amplified, using bridge amplification method, where the unbound end of a fragment is annealed to the neighbouring primer and complementary strand is created. The process is repeated to produce dense clusters of copies of a fragment. Sequencing-by-synthesis is then performed, where labelled nucleotides are bound to reads in clusters. Laser excites the labelled nucleotides, producing a different fluorescent signal for each nucleotide type, which is recorded by a camera. Labelled nucleotides have reverse terminator capabilities to ensure that only one of them can bound to a read at a time. Fluorescent label is then removed from a nucleotide along with its terminator capabilities, and the process of binding of labelled nucleotides is repeated. Fluorescence signals are recorded simultaneously for millions of clusters, producing a sequencing output faster than Sanger sequencing method. Often, after the core steps are finished the clusters are reversed and the process repeated in order to obtain reads from the other end of a sequenced fragment. This is called paired-end sequencing (not shown).

1.4.3 RNA-Seq

RNA-Seq, a variety of NGS, is used to study RNA on different levels - expression, gene, transcript, exon, and splicing. The proportional number of reads originating from a gene is an indicator of its relative expression, while the number of reads covering splice junctions indicate splicing patterns and isoform expression. It can also be used to identify SNPs and small insertions/deletions (indels) [61].

Contrary to what the name suggests, RNA-Seq does not sequence RNA directly, but sequences cDNA created from RNA by reverse transcription. Inability to sequence RNA directly is due to RNA's instability.

RNA-Seq itself, as a technology, has some limitations. Biological tissue samples tend to be heterogeneous, composed of a mixture of different cells. If RNA-Seq is used on a heterogenous sample without specific cell-selection, it may introduce a bias into the data obtained [62], as expression is tissue-specific. Due to the nature of an investigated sample, such selection may be not possible, e.g. selection of fusion-exhibiting cells in whole blood samples for investigation of unknown gene fusions. The problem is partially resolved at the level of study design, where cell selection has to be considered, and the number of biological replicates increased if

a potential for the occurrence of the problem exists. Another challenge of RNA-Seq lies in library preparation. This step introduces non-uniformity of cDNA fragments that represent transcripts [63]. This is normally resolved using normalisation techniques, which correct for the bias, estimating reads lost due to the non-uniformity and incorporating that information in relative abundance estimation. Large datasets tend to contain background, unimportant data, commonly referred to as "noise". When processing millions of reads, background noise, both biological and technical, has to be considered. It is particularly hard to estimate this background noise, as it depends heavily on a sample origin and preparation. However, it can be reduced with usage of biological replicates.

RNA-Seq data is biased by pseudo-random priming of the reads, and other technical issues such as lower read coverage at the ends of transcripts. Normalisation techniques aim to correct for bias while reducing noise. There is currently no single standard for RNA-Seq data normalisation, but each methodology investigating particular component of the data applies its own normalisation technique.

Gene Fusions

Gene fusions and mutations can be investigated using RNA-Seq. Although the method is still experimental and far from being approved for clinical use, its potential is tremendous. At a fraction of a cost per-base when compared to Sanger sequencing, RNA-Seq can be used to investigate the entire transcriptome of a sample. As such, RNA-Seq removes the necessity of selecting gene fusion candidates which is limited by an investigator's subjective assessment.

RNA-Seq provides information on gene fusions with great resolution - down to 1bp if present within an exon or intron length if within an intron. Since RNA-Seq is performed on cDNA, it does not contain information on introns and is therefore blind to any mutations occurring within them. Hence, if a precise fusion breakpoint within DNA is to be discovered, DNA sequencing methods need to be applied to determine the exact breakpoint.

As any method that includes PCR amplification of sample material at some point in preparation, RNA-Seq is susceptible to contamination and PCR artefact presence. It is difficult to assess the significance of these dangers as there are currently no publications on the issue in

the context of fusion investigation in RNA-Seq. Generally, PCR artefacts should be detected during quality control of RNA-Seq data, and can be removed from dataset prior to analysis.

Similar to any other NGS method, RNA-Seq has problems with mapping reads that originate from regions with high concentration of repeats and unspecific regions of transcriptome. The problem is somewhat reduced in RNA-Seq due to read specificity to transcriptome only, where repeats are less common than in intergenic regions.

RNA-Seq data analysis is computationally expensive, and large studies need to be performed using a high performance computing cluster (HPC), commonly referred to as a "supercomputer". They are arrays of multiple high-end computers, joined together allowing for simultaneous use. This is a limiting factor, especially considering that one might want to investigate RNA-Seq data using different methodologies, while HPC usage is expensive. Another available option are cloud-based computing services, which allow utilisation of computing resources on per-need basis. However, they can be costly and suitable only for small experiments. As the information is hosted on a third-party server, security of the data is also questionable.

For the purposes of this thesis, RNA-Seq was performed on Illumina platforms due to its cost efficiency when compared to other platforms as well as relatively low error rate. More details can be found in Chapters 4, 5, and 6.

Software

A typical run involving one lane of sequencing on Illumina HiSeq 2000 with four samples per lane and 100bp paired-end sequencing generates approximately 100,000,000 reads, and during analysis takes approximately 20GB of disk space. The data is in almost random order, and each read has to be mapped to the reference sequence. This poses a complex computational challenge when determining exon-exon boundary, alternative splicing, novel splicing variants, novel exons and gene aberrations. If the fact that every base within a read has assigned different probability score (quality score) is added to this collection of variables, one can appreciate the reason for completely different approaches which bioinformaticians use when confronted with the problem. Indexing strategies are used which are applied either by

a form of hashing or performing a data transform, such as Burrows-Wheeler [64] and scoring potential alignments usually by seed-and-extend method with different scoring matrices [65].

There are two main types of evidence to support a fusion in RNA-Seq data. Supporting pairs (SP), where one read within a pair is mapped to first gene of the fusion and its mate read is mapped to the other gene, and supporting reads (SR), where a read is mapped to both genes, spanning the fusion, as presented in Figure 1.19.

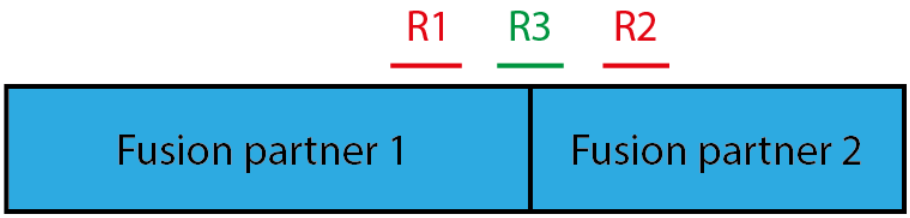

Figure 1.19: Sequencing reads as fusion evidence. R1, R2, red - supporting pair (SP) of two reads where one read is mapped to a fusion partner and its mate to the other fusion partner. R3, green - supporting read (SR) spanning fusion junction.

If single-end RNA-Seq is used, there is no data available for SP, therefore, single-end method is inferior to paired-end in fusion analysis. Since reads originating from regions of high similarity, such as genes with many pseudogenes, conserved motifs, and fragments with high homology are often misidentified as potential fusions, it is necessary to test for their specificity. Different tools approach this problem in variety of ways, with most common being BLAST [66] or BLAT [67] search and score-based cutoff.

One of the major problems with RNA-Seq data analysis for fusion identification is the lack of a single, complete and efficient protocol. Separate tools exist for different steps of analysis and they are often incompatible with each other. Arguably, the most complete software suite is Tuxedo [68] with Tophat-Fusion [66]. However, it still does not encompass quality control and most types of expression level analyses such as exon-level differential expression or alternative splicing events analysis. Another obstacle, directly related to lack of one protocol, is usage of different settings in a variety of software, requiring adjustments whenever different samples are processed and whenever investigating different properties of a sample. The lack of a single, reliable protocol for fusion identification in RNA-Seq often pushes scientists to design their own fusion-detection algorithms [66, 67, 69, 70, 71]. However, most of them are

not used beyond their original publications. Notable exceptions are Tophat-fusion [72] and deFuse [67], that were applied in uncovering novel gene fusions in colorectal cancer (*PRTEN-NOTCH2*) [73] and in lymphoid cancers (*CIITA*-various) [74], respectively. Researchers tend to use a combination of software, e.g. in colorectal cancer research [73] with default settings or focus on a single software package with default settings, e.g. in lymphoid cancer research [74]. For the purpose of this thesis Tophat-Fusion was used to investigate gene fusions. More details on the selection and other available software can be found in Chapter 2.

Databases

Databases are an important source of information in the study of genomics, transcriptomic or other 'omics. With large datasets, grouping of important information and accessibility which allows automation is of utmost significance, as it is impractical for a researcher to perform extensive study on every single datapoint produced by NGS.

From the viewpoint of RNA-Seq application for gene fusion investigation, the Mitelman database [10] is a valuable resource. It contains karyotypic coordinates of all known gene fusions. As such, it enables the fusion partners to be categorized as known or novel.

One of the most commonly used databases in NGS analysis, and for genomics and transcriptomics in general, is UCSC Genome Browser [75], with abundance of information on reported existing features within the genome and transcriptome combined from other databases, it is an invaluable resource. Still, there in no one universal database that combines all the possible resources in one place.

1.5 Gene fusions and treatment

Accurate identification of gene fusions allows to refine diagnosis and can have a great influence on the chosen treatment. While diagnosis classifying a patient to a general group within myeloid malignancies allows for more conventional treatments, only refined, specific diagnosis allows application of more effective targeted chemotherapy, for example ruxolitnib

in *JAK2* gene fusions [76].

1.5.1 Available therapies

Standard chemotherapy

The main goal is either to eradicate cancer cells, or control their numbers if eradication is not possible. This can be achieved by stopping the growth of cancer cells in two ways - either by killing the cells or by stopping them from dividing [77]. Examples of chemotherapeutic drugs used in myeloid malignancies are: Cytarabine (cell cycle inhibitor, used in CML, AML), Daunorubicin (AML), Mitoxantrone (CML), Etoposide (cell cycle inhibitor, AML), Idarubicin (AML), Fludarabine (AML) [78]. Combination chemotherapy, which involves more than one anticancer drug is also applied. An example of such therapy is the combination of arsenic trioxide along with all-trans retinoic acid which is used in a subtype of AML with the *PML-RARA* [79]. Standard chemotherapies have a wide range of side effects, which depend on the particular drugs used, with more than 10% drug-related mortality rate in haematological malignancies reported in 2006 [80].

Targeted chemotherapy

There are three major categories of targeted therapies in myeloid malignancies, all involving a drug agent targeting a specific gene product or a process. Among the first category - epigenetic regulators (epidrugs) are inhibitors of lysine-specific demethylase 1 [81], histone deacetylase inhibitors [82], and hypomethylating agents [83]. The second group includes targeting a specific gene fusion product or a gene involved in a fusion, an example of which is imatinib to target *BCR-ABL1* or *FIP1L1-PDGFRA* [12]. The last category belongs to drugs targeting mutated gene product, e.g. FMS like tyrosine kinase inhibitor [84].

Targeted therapies are often a better solution than chemo-/radio- therapies[12]. They produce less side side effects that can be devastating and a direct cause of death in chemotherapy, and can be used where there is no clear site for radiotherapy. Overall this translates to a better prognosis for patients.

Radiation therapy

The aim of radiation therapy is similar to that of chemotherapy - complete eradication of cancer cells or control of their numbers. This therapy uses radiation often in the form of high-energy x-rays to damage DNA of targeted cells beyond repair [85], effectively killing them. In AML, radiation therapy is usually used only to target metastatic tumours sites or alleviate bone pain occurring due to aggregation of cells.

Stem cell/bone marrow transplant

Transplantation of stem cells or bone marrow is performed after successful chemo-/radio-therapy that killed the bulk of cancerous cells. It reintroduces healthy pluripotent cells into bone marrow from the same patient, taken before chemo-/radio- therapy (autologous), or from a donor (allogeneic). However, transplants are associated with a constant difficulty of donor identification for all patients as well as increased morbidity and mortality in elderly patients, and some countries apply upper age limit for the procedures [86].

The ultimate goal of bioinformatic investigation of gene fusions in the scope of this project is to improve the existing methods by increasing their efficiency and accuracy. Accurate detection of gene fusions is not only essential for the correct diagnosis and targeted treatment possibility, but also has the potential to influence the design of new targeted chemotherapies. What is more, due to the advantages of RNA-Seq, it would also be possible to employ the method to monitor minimal residual disease (MRD) levels throughout the treatment and during remission.

1.6 Splicing

1.6.1 Overview

The flow of information between the basic information-carrying molecules of life - DNA, RNA, and proteins - has been investigated in detail over the past decades. The model of such information flow is described as the central dogma of molecular biology. The dogma, in its least disputed form, was first proposed by Crick in 1958. [87] Shown in Figure 1.20, it provides an overview of the most fundamental information flow in biology.

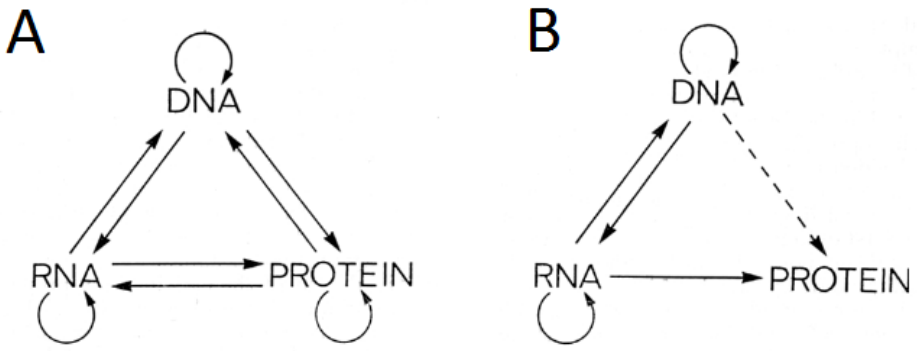

Figure 1.20: Central dogma of molecular biology. A) All possible combinations of information transfer between the polymer groups. B) Current understanding of information flow between the polymer groups. Solid arrows show proved transfer, dotted arrow shows possible transfer. Note absence of the arrows originating at protein group. [88]

Gene expression is a process during which information coded by DNA is processed into a functional product, either from RNA polymer group or from protein polymer group. In both cases, naturally occurring information flow has to undergo processing where DNA information is transformed into polymers from RNA group. Expression is of particular importance in studying functional deficiencies as a result of errors along the path.

In the case of protein expression, information contained within DNA is first processed into pre-mRNA in nucleous in a process called transcription. There, it undergoes a series of modifications, transforming pre-mRNA into mRNA, which is transported out of the nucleous to use as a template for protein production. One of the main focuses of this thesis is splicing, a part of gene expression process that follows transcription. Where end-point of gene expression is a protein, it results in production of mRNA, as presented in Figure 1.21.

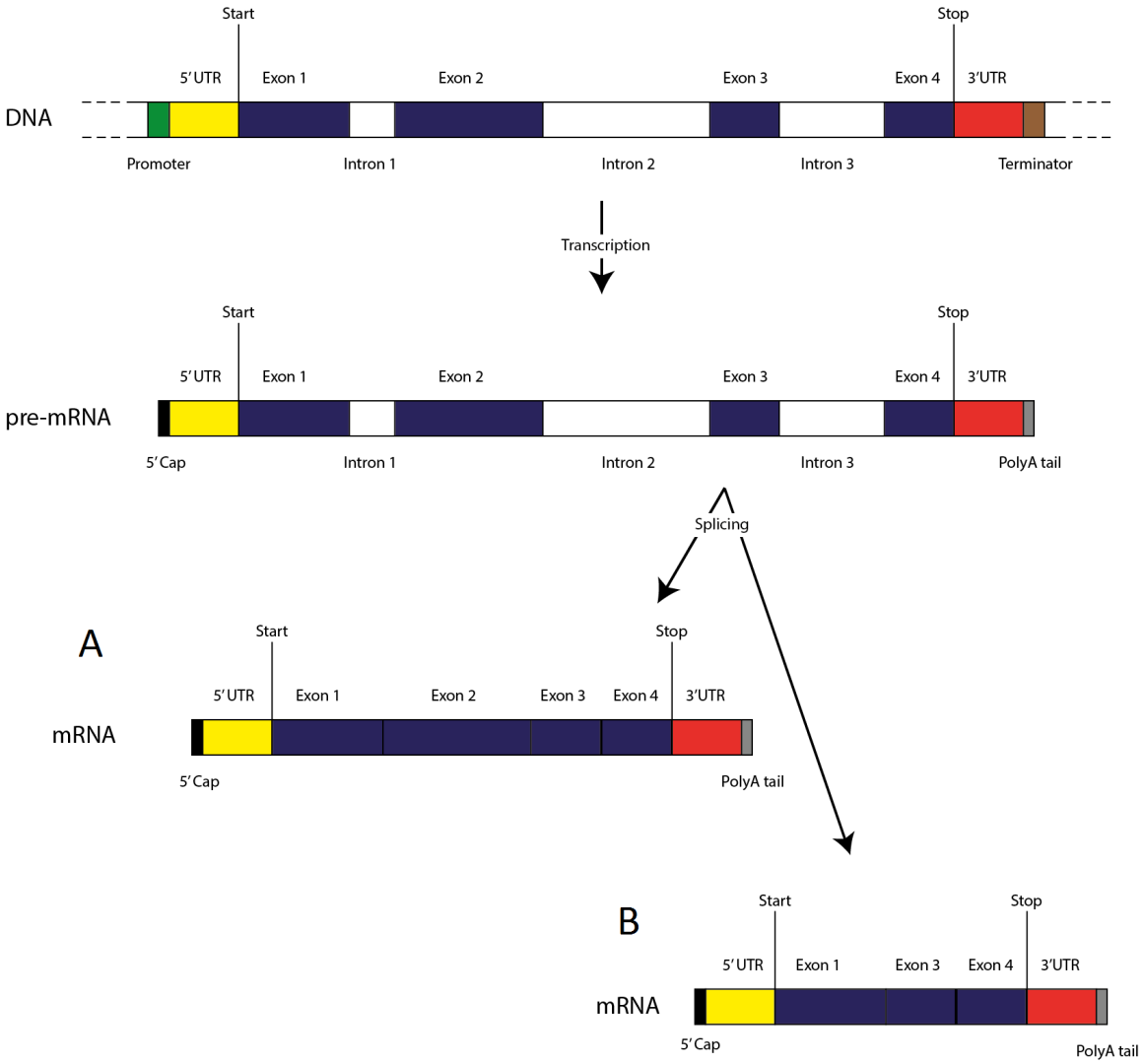

Figure 1.21: DNA to mRNA processing overview. DNA is first transcribed into pre-mRNA where 5'cap and polyA tail are added. Splicing follows, with exons being joined either in order (A) or in alternative configuration (B).

1.6.2 Spliceosome

Spliceosome, a complex of snRNAs and proteins, is responsible for splicing. These two groups of molecules interact with each other, functioning as binding platforms and catalysts. They recognise specific sequences within pre-mRNA and in a process of catalytic interaction promote exon splicing [89]. While part of exons are spliced by removing introns between them, some exons are prone to alternative splicing. It is a process where parts of spliceosome recognise different 5' or 3' end of an exon or miss an exon altogether. This adds to the variety of produced mRNAs, which in turn increases variety of proteins being able to be produced from a single gene, expanding its array of functions. As only 10 % of a typical pre-mRNA is

exonic sequence, spliceosome has to be both precise and flexible in the recognition of specific sites in pre-mRNA [89].

The key spliceosome recognition sites, presented in Figure 1.22, consist of consensus splice sites (AG and GU), polypyrimidine tract, and branch site.

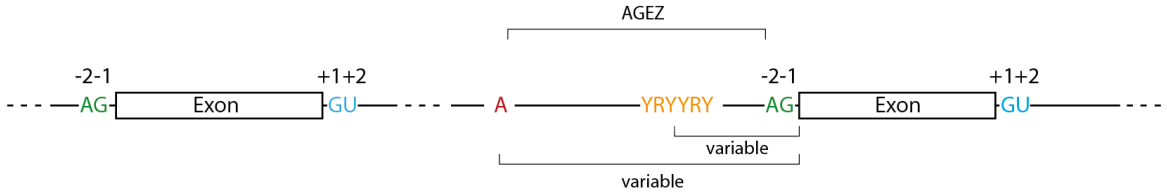

Figure 1.22: Key spliceosome recognition sites. Consensus AG splice sites at position -2, -1 downstream of 3'ss marked in green. Consensus GU splice sites at position +1, +2 upstream of 5'ss marked in blue. Polymyrimidine tract (PPT) marked in yellow. Branch site (BS) marked in red. Distances between PPT and 3'ss as well as between BS and 3'ss are variable. AG-exclusion zone covers the area between the branch site and AG consensus sequence.

The main components of spliceosome are U1, U2, U4/U6 and U5 small nuclear ribonucleic proteins (snRNPs). Spliceosome components bind to pre-mRNA in a specific order, each component recognising its specific site. Formation of spliceosome on an exon during the earliest stage of splicing is shown in Figure 1.23.

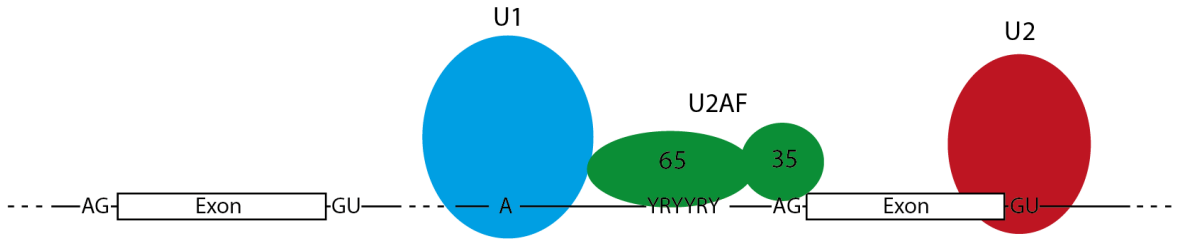

Figure 1.23: Formation of spliceosome on an exon. U1 small nuclear ribonucleic protein (snRNP) recognizes and binds to the 3' end of an exon. U2 snRNP recognizes 5' end of an exon through U2AF. [89]

Throughout the process of splicing, multiple spliceosome components are involved at different steps (Figure 1.24). The most common order of binding is U1, U2, U4/U6.U5. However, there is no requisite for U1 binding before U2. [90]

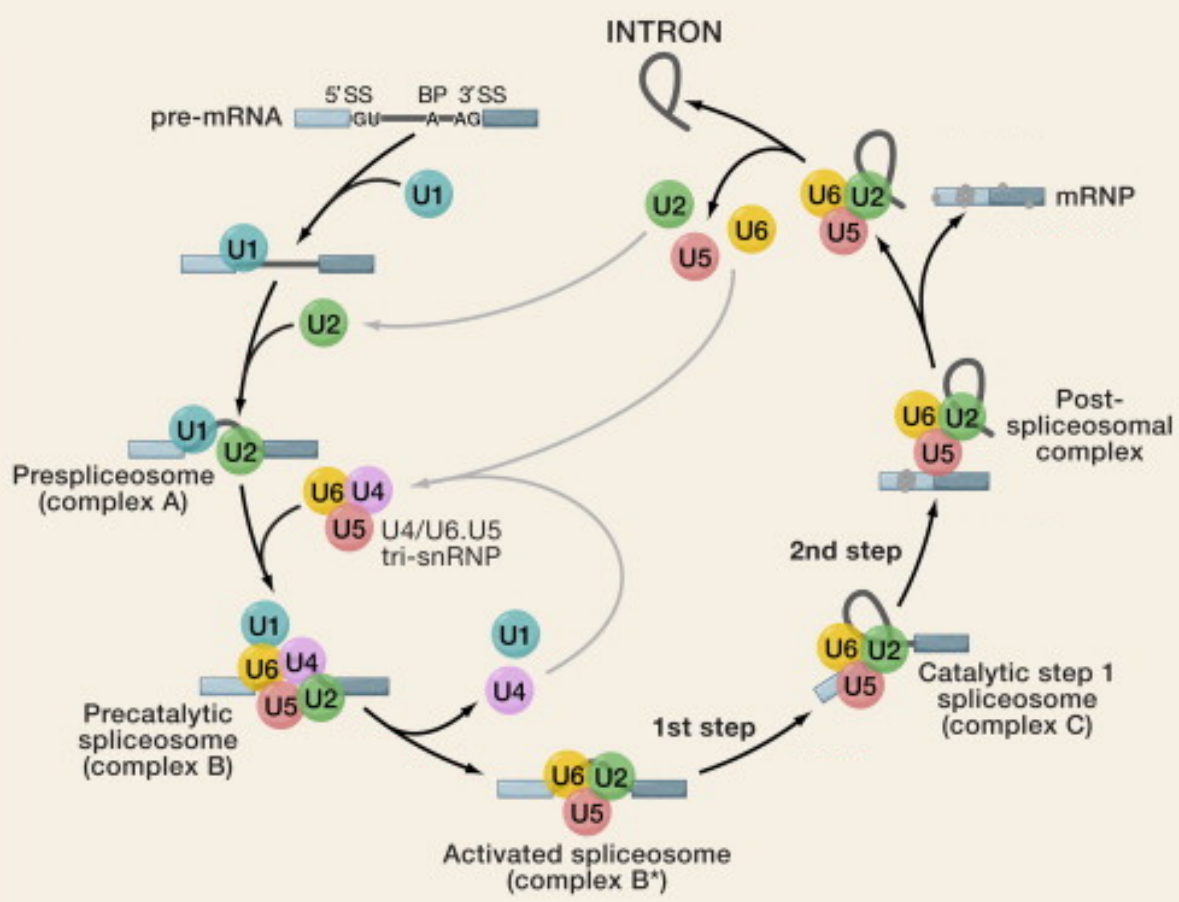

Figure 1.24: Spliceosome components splicing an intron. First, U1 small nuclear ribonucleic protein (snRNP) binds to 3' end of an exon, which is followed by binding of U2 snRNP to 5' end of an exon creating prespliceosome. U4/U6.U5 tri-snRNP complex is recruited to form precatalytic spliceosome. U1 and U4 are then detached, resulting in activated spliceosome. In multi-step rearrangement, lariat is formed by 5'-end of an intron being ligated to A recognised by U2 snRNP. Lariat is then cleaved at the 3'-end of an intron and released, while exons are ligated. Source: [89]

Binding of U1 does not necessarily initiate the splicing process, it is binding of U2 that commits a location to splicing. [91]. As seen in Figure 1.23, U2 interacts with U2AF, which consists of two polypeptides - U2AF35 with weight of 35 kDa, coded by *U2AF1*, and U2AF65 which is 65-kDa and coded by *U2AF2*. U2AF recognises 3' splice site with U2AF65 binding to the polypyrimidine tract, and U2AF35 contacts consensus AG at the end of 3' splice site, with U2AF recruiting U2 [92].

Apart from the key sequences recognised by spliceosome, there are also sites altering the splicing process. Exonic splicing silencers (ESSs) alter splice site choice by inhibiting the splice site process through recruitment of hnRNPs that prevent access of spliceosome to recognition

sites [93]. Exonic splicing enhancers (ESEs) influence splicing by facilitating the process using SR proteins interacting with spliceosome [94].

As site recognition in splicing is precise, mutations altering sequences of recognition sites can have a profound effect on splicing, which results in changes in gene expression. Such mutations can be the direct cause of a disease or act as a factor in severity or susceptibility [95]. Most splicing-affecting mutations disrupt the consensus 3' and 5' splice sites and give rise to different conditions, including breast and ovarian cancer [96]. Mutations within an exon affecting ESEs, or ones outside of an exon that disrupt branch sites or polypyrimidine tract can also cause other cancers such as retinoblastoma [97].

The functionality of spliceosome can also be affected by mutations in genes coding for its elements. Recurrent somatic mutations in the auxiliary splicing factor *U2AF1* have been implicated in multiple cancer cases, including AML, MDS and CLL [98]. Their presence, similarly to the presence of *TP53* mutations, is associated with overall worse overall survival rates, absence of clinical remission, and poor disease-free survival [99]. Application of RNA-Seq to characterize *U2AF1* and its influence on gene expression levels as well as splicing patterns, is likely to be important to understand its function and association with haematological malignancies.

For the purposes of this thesis RNA-Seq obtained on Illumina platform was used to investigate splicing. Expression data analysis was performed using Cufflinks [68] and MISO [100] along with custom scripts, as detailed in Chapters 3 and 6.

Chapter 2

Methods - Gene fusion analysis pipeline

The aim of this chapter is to present the developed analysis pipeline for gene fusion identification. Several programs have been developed for the aim, but an optimal approach is yet to be determined [101]. Rationale for the chosen software and their comparison with publicly available alternatives detail the developed methodology. Presented pipeline is designed to be applied to patient RNA-Seq data in order to identify present gene fusions in myeloproliferative neoplasms (MPNs). The pipeline was applied to RNA-Seq data from patients with MPNs with known gene fusions and optimized with regard to detection of these events. In Chapter 3, the optimized pipeline is applied to RNA-Seq data from MPN patients in which gene fusions were suspected but were unidentified by conventional cytogenetic techniques.

2.1 Introduction

A bioinformatic software pipeline is a combined set of software that automates the analysis of data through independent pieces of software. Analysis pipelines are created for NGS data for three main reasons: to reduce human supervision, to reduce processing time, and to ensure reproducibility.

Processing any NGS data, regardless of sequencing type, entails the following 4 steps [102]:

1. Quality Control (QC) - a routine examination of reads, identifying potential biases.

2. Pre-processing - removal of reads that failed QC and preparation of reads for alignment or assembly.

3. Alignment/Assembly - for organisms with a sequenced and assembled genome reference, the reads are aligned to it in order to identify location of their origin. For organisms with unknown genome reference the reads are matched with each other, assembling bigger constructs that can be analysed. As this thesis investigates human data only, which has a reference genome, the method for assembling reads is not discussed.

4. Analysis - varies, depending on the type of NGS data and aim of the investigation.

The following sections provide a detailed description of the software pipeline built for RNA-Seq data for gene fusion identification, along with an overview of available software that can be used for this purpose.

2.2 Analysis pipeline

2.2.1 Quality control

Quality control of RNA-Seq data entails a variety of assessments and metrics. They are designed to provide a general overview of a sequenced sample quality along with a detailed assessment of reads indicating presence of a bias. The first possible source of bias in RNA-Seq data are contaminants of which two varieties can be distinguished. Samples can become contaminated at any point leading to sequencing of foreign genetic material [103]. This introduces reads that can give skewed representation of the sample content during the analysis. The second type of contamination can be inefficient removal of sequencing adapters [104]. If the sequencing adapters are preserved in the analysed data, observable as reoccurring sequences at the ends of reads, they may lead to inefficient alignment of reads and reduced analytical power. Another source of bias is introduced by pseudo-random priming during library preparation [63]. It appears as non-random distribution of nucleotides at the beginning of reads, and may reduce alignment efficiency in extreme cases. Low base quality may introduce another type of bias, as low quality corresponds to low certainty of nucleotide sequence called [105]. Reads with dubious nucleotide sequences can align to locations in the reference not corresponding to the mRNA molecule they are supposed to reflect, skewing the analysis. GC

bias, PCR overamplification/underamplification are another factors to consider. Sequencing library preparation involves PCR, which is known to have varied efficiency when amplifying fragments of variable GC content [106]. As a result, it is possible for some fragments to become overamplified or underamplified, so that a sample's mRNA molecule composition is misrepresented. QC is used to investigate the occurrence and prevalence of these possible sources of bias and to determine if and what pre-processing needs to be applied. QC software designed to detect these data biases is shown in Table 2.1.

Software	Sequence quality	Sequence content	Overrepresented sequences	Foreign sequences	Visualisation
QC-Chain [103]	✔	✗	✔	✔	✗
FastQC[107]	✔	✔	✔	✗	✔
NGS QC Toolkit[108]	✔	✔	✔	✗	✔
HTQC[109]	✔	✔	✗	✗	✔
PRINSEQ[110]	✔	✔	✔	✗	✔

Table 2.1: Selected Quality Control (QC) software features. Software was selected on the basis of updates or first appearance within the past 2 years. Not maintained software is not listed.

The choice of QC software for pipeline implementation was dictated by multiple factors. Programmes were shortlisted by their ability to perform tests for as many types of biases as possible. Visualisation of QC metrics enabling manual scrutiny of the data was also considered advantageous since rigid cut-offs cannot be applied for some measures, as discussed further in the following sections. Other attributes of the software that cannot be implemented in the pipeline, such as web interface, were not considered. The final criterion for the choice was maintenance of the software. With NGS being a fast developing field, one can only expect QC software to follow its advances to allow for accurate and up-to-date standard of QC. As the software meeting the majority of the criteria, it was chosen to use FastQC [107] for QC. It is important to note that FastQC does not process every read in a dataset to produce its metrics. Instead, it extracts the first 100k reads, which are random and represent the set.

Sequence quality

All NGS platforms estimate the error rate per base of sequence and use a Phred scale [105] (Table 2.2)to encode this information. These Phred scaled error rates provide a measure of sequence quality.

Phred score	Base call accuracy
10	90%
20	99%
30	99.9%
40	99.99%

Table 2.2: Phred quality scores and their corresponding base call accuracy.

Phred scoring system is a metric that encapsulates accuracy of each base pair of sequence. As such, it is invaluable for pre-processing reads, as depending on alignment algorithms used, reads that do not meet sequence quality thresholds need to be filtered. It also provides important information on potential technical problems with the sequencing equipment used and analysed sample integrity.

It has been widely accepted to process reads with Phred score $\geqslant$30 [111], which translates to mean of maximum one incorrect base call per ten 100bp long fragments. However, development of quality-aware aligners, such as implemented Bowtie, reduced the need for filtering of low-quality reads, instead a score penalty is applied to a low quality base call when the read is aligned [112]. Hence, it is possible to salvage reads with Phred score $\geqslant$10, $<$30. The pipeline implements a check halting the processing if the median for any base is less than 20 or lower quartile for any base is less than 5.

Sequence content

The content of the reads is analysed to provide an overview on potential bias. Saturation of each base is considered, giving a valuable overview of the data composition. Deviations from expected ratios might suggest PCR artefacts or another source of bias. This type of check includes K-mer contents, overrepresented sequences, and per-base N contents. Any deviation from expectations might alter the downstream analysis and lead an investigator to incorrect

conclusions. Usually, sequences that contribute to artifacts are removed in pre-processing. The analysis pipeline stops and indicates a QC problem when at any position across all reads the N content exceeds 20%

Since the mean GC content in human genome is 41.7% (computed for hg19) it is generally expected to see GC content of NGS reads around that value. However, since GC content varies between genes greatly, and mRNA composition in a tissue is dynamic, deviations from mean GC content are expected. Hence, no hard cut-offs are implemented in the pipeline based on GC distribution.

Overrepresented sequences

Occurrence of overrepresented sequences in data is another indicator of a quality problem. Recurring identical combinations of reads within a raw NGS data set are not expected and suggest either overamplification of particular fragments or adapter contamination. If there are reads that are found to represent more than 1% of the sample, the analysis pipeline halts, signalling the quality issue and allowing for manual examination.

Visualisation

Visual analysis of data quality is an approachable way of investigation of potential quality issues. In large datasets, such as raw NGS data where millions of reads are considered at a time, it is an effective tool to discover reasons for quality control failure, and an important feature of QC software. The analysis pipeline outputs a graphical report for each sample, allowing for manual scrutiny.

2.2.2 Pre-processing

Quality control is followed by pre-processing of data. Here, reads that did not pass QC are adjusted or removed. One of the core features of pre-processing software is read filtering. Reads that are identified in QC as unusable, such as reads with extremely low score throughout (mean Phred<10), should be removed from the set, as their alignment may introduce issues in downstream analysis. Read trimming is the second core feature. In cases such as adapter contamination or low sequence quality at one of the ends of reads, it is not necessary to

discard the read completely. Instead, only the affected bases can be trimmed, preserving valuable data. Finally, pairing synchronisation has to be considered. Raw paired-end RNA-Seq data contains information on pairing of the reads, which has to be preserved during pre-processing.

There is an abundance of pre-processing software[103][108][109][110][104] with some of them implemented as a functionality of QC applications. There is little difference between them, and anything that incorporates the above-mentioned points can be used. Trimmomatic[104] is the tool of choice for the analysis pipeline. As expected from a pre-processing software, it is able to filter reads, trim them and considers read pairing, preserving synchronisation. It also contains a library of Illumina adapters, so removal of them does not require specific problem identification in QC.

Trimmomatic's workflow is presented in Figure 2.1.

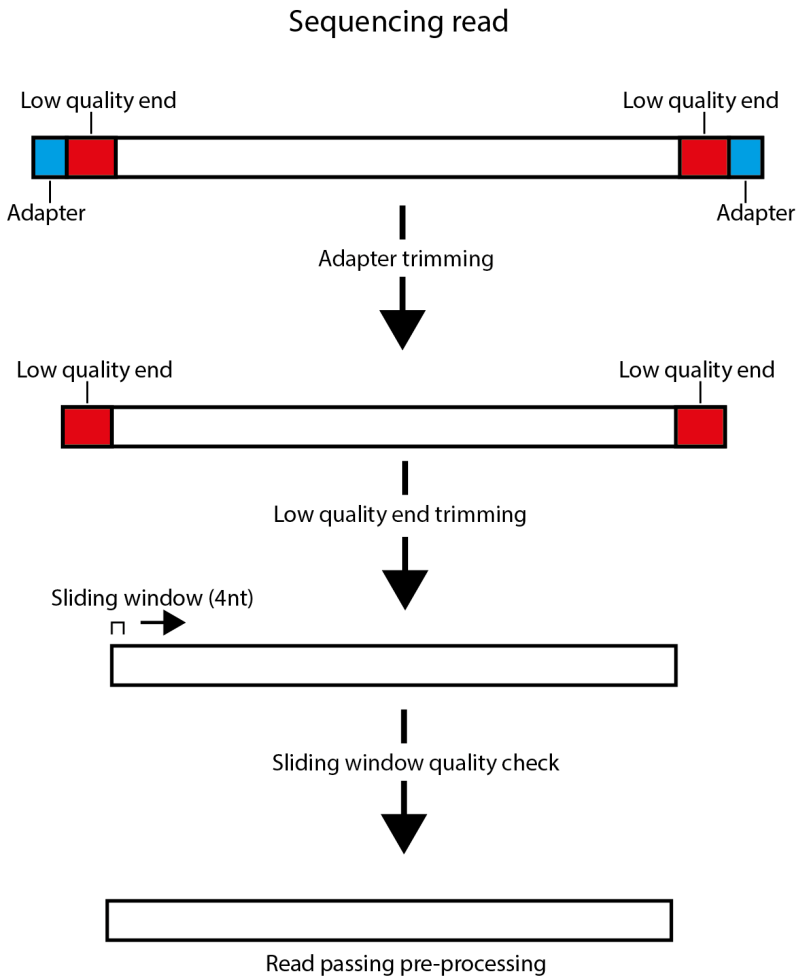

Figure 2.1: Read pre-processing workflow. 1) Adapter trimming - Illumina adapters are trimmed from both ends of a read; 2) Low quality end trimming - nucleotide runs with Phred <30 at the ends of a read are trimmed up to 10nt from an end; 3) Sliding window quality check - every 4 consecutive nucleotides are checked for their mean Phred score from 5' direction. If such run is found, the read is cut, removing the run and following nucleotides. [104]

Pre-processing software was set up to trim the ends of reads with Phred score $\leqslant$10, as bases with such low associated score are too unreliable even for quality-aware aligners. It was also designed to trim Illumina adapters from the ends of reads, as their presence would prevent the correct identification of the read origin location by the aligner. Pre-processing of data in the presence of other biases, has to be done on case-by-case basis, as it is impossible to implement a set of cut-off rules for them that would not potentially remove valid data.

2.2.3 Alignment

Before any analysis, RNA-Seq data needs to be aligned. This process aims to assign reads to the reference in order to find their most likely location of origin. This is perhaps one of the biggest bottlenecks of any NGS analysis, as finding the best match between millions of short reads (30-600nt long) and reference genome (∼3Gbp long) is a computationally complex task. Note that reads can be aligned to either positive or negative strand of the reference genome, which effectively increases reference size to ∼6Gnt for the purpose of computation.

Many different algorithms were designed over the years since the development of NGS to reduce the computational complexity of alignment. It is hard to say precisely how many mappers are available. Seqanswers, a portal for bioinformatics professionals and academics lists 60 [113], Wikipedia lists 58 [114], with the most comprehensive academic review covering 60 of them [115], with the list from the publication constantly updated and currently containing 68 aligners [116].

Aligners utilise genome and transcriptome reference for read mapping. The current version of human genome reference at the time of writing was hg19, build GRCh37 [117], which was used in the analysis. The genome reference represents the common haploid model of nucleotide sequence across all chromosomes, mitochondria, and ribosomes of 13 anonymous individuals from USA, assembled by the Genome Research Consortium [118]. Even though the human reference is the highest quality mammalian reference available, it contains gaps and allele variation errors [117] which are important to consider during analysis. A transcriptome reference annotates the positions of genes, isoforms, and exons in a genome reference. The most comprehensive transcriptome annotation at the time of writing was the UCSC transcriptome annotation for hg19 [119], implemented in the analysis pipeline. Location of transcriptomic features is determined by observed proteins, mRNA, and predictions based on common features characterising a gene [120]. As the transcriptome and proteome are dynamic and tissue-specific, and not all genes follow the same structural pattern, the annotation is constantly updated with new isoforms, more rarely new genes, but represents our current understanding of genetic landscape of the human genome [119].

The majority of aligners use indexing methods, which aim to reduce computation time [112]. Indexing is a process of memory allocation recording unique identifiers from a large set of data, so that data does not have to be parsed during each algorithmic iteration. Some indexing methods may pose problems when inexact matches between reads and reference are considered. However, without inexact matches, it would not be possible to align reads containing a Single Nucleotide Polymorphisms (SNPs) or insertions/deletions (indels).

One of the problems that RNA-Seq data poses over other NGS data is determination of splice sites - this type of sequencing generates mainly mRNA, which has introns spliced out. The genome reference does not include information on exon and intron position necessitating specialised splice alignment. Usually splice alignment is done in two main steps - alignment to genome reference and splice junction resolution with alignment to said junctions [121]. The alignment to genome reference is standard, same tools are used that are applied for other NGS data, with most of them built on the Burrows-Wheeler algorithm [122]. However, it is the splice aligners that are more important for RNA-Seq data, as accurate prediction and expression estimation of isoforms depend on them. Important factors to consider when examining abilities of different alignment software are time, computational requirements, and alignment percentage. Table 2.3 provides an overview on these aspects for some of the most popular aligners that can be used for RNA-Seq.

There are three distinct groups of RNA-Seq aligners [121]. The first group consists of *de novo* splice aligners. This is a group of software that do not need information on exon and intron positions in order to determine splice junctions and align reads skipping intronic sequences. They depend mostly on accurate mapping of exon islands as the first step, and determination of splice junctions between mapped exons. The most popular among them is GSNAP [125]. The main disadvantage of *de novo* splice aligners is their inability to accurately resolve splice junctions in genes with low expression. Annotation-driven aligners are the second group of RNA-Seq aligners. They rely on transcriptome annotation, a collection of information on positions of introns, exons and structure of isoforms. One of the examples is SpliceSeq [130]. Since annotation-driven aligners depend on accurate and complete transcriptome reference, they cannot accurately define splice junctions between exons of unannotated isoforms, invalidating their use for investigation of previously undiscovered mRNA. The last

Aligner	Total time [m]	Preparation time [m]	Aligning time [m]	Memory usage [GB]	Reads Aligned [%]
BFAST[123]	35	17	17	21.4	72.7
BLAT[124]	1741	3	1738	3.8	97.5
GMAP[125]	931	18	913	7.6	99.2
GSNAP[125]	68	20	48	7.6	92.4
Novoalign[126]	90	13	77	7.8	79.5
Smalt[127]	48	5	43	5.2	99.6
SOAPSplice[128]	104	76	29	5.4	87.8
SSAHA2[129]	91	16	76	9.5	100
TopHat[72]	99	71	28	5.1	80.8

Table 2.3: Selected 9 most efficient RNA-Sequencing (RNA-Seq) data aligners. Efficiency measured by lowest total running time, mapping time, memory usage, and number of reads mapped. Data given for alignment of 1 million 30 base pair (bp) single-end reads using default settings on a single processor with 32 GB RAM. Source [115].

group of aligners can be described as *de novo* splice aligners using annotation, an example of which is TopHat2 [131] used in conjunction with Bowtie[112]. They combine strengths of the previous two groups, utilising transcriptome annotation to accurately define splice junctions in genes with low expression, with ability to discover novel junctions in unannotated exons. What is more, the combination of software is quality-aware, meaning that it considers calling quality during mapping process. The choice of aligner for the analysis pipeline was therefore restricted to the software belonging to the last group.

Prohibitively long running time required by algorithms implemented in BLAT or GMAP can limit their applications for RNA-Seq data, as a HPC would be absolutely necessary to process even one sample within reasonable time. Note that the analysis presented in Table 2.3 is given for 1 million of 20bp, unpaired reads, whereas complexity, RAM requirements and running time will be higher for 100 million of 100bp paired-end reads, making BLAT or GMAP even more limiting. When the read alignment efficiency is considered, software with very high percentage of aligned reads can seem as the best. However, in RNA-Seq data, it is not possible for QC and pre-processing to remove all the undesired reads. What is more, sequencing technology is not perfect and part of the reads are of too poor quality to be aligned. Those factors combined show that high percentage of aligned reads may indicate

high number of false alignments. False alignments can affect the analysis greatly, especially when one investigates expression - directly related to abundance of reads aligned to a region [132].

Read alignment percentage indicates what percentage of input reads was successfully aligned to a reference. RNA-Seq is not an error-free method of determining RNA sequence, therefore some reads are expected to contain wrong nucleotides. Just because an aligner has a lower percentage of aligned reads than others does not mean that it is less efficient, it simply indicates more stringency during alignment. In a similar manner, higher percentage of aligned reads does not mean that an aligner is 'wrong', it indicates low stringency. However, as explained later in this Chapter, aligners with lower read alignment can be used more effectively for gene fusion detection.

TopHat2[131] in conjunction with Bowtie[112], was selected as it belongs to the group of *de novo* splice aligners using annotation. Furthermore, as shown in Table 2.3, its previous version provides reasonable mapping efficiency of around 80% indicating that it is less likely to superficially overmap reads and introduce bias into downstream analyses. The software is used in conjunction as Bowtie, which is used for the basic alignment, while TopHat2 handles splice site alignment (Figure 2.2). Bowtie uses an FM index [133] which is a combination of BWA [122] and the suffix array data structure [134] to store the reference genome in a compressed format which enables it to be rapidly searched. As a result, Bowtie aligns reads at a rate of tens of millions per CPU hour. It applies a scoring matrix to assess validity of read alignment to a specific position, where mismatches occur. Reads cannot exceed specific scoring threshold in order to be aligned to a position, allowing mapping of SNPs and small indels. However, Bowtie does not allow for large gaps such as those created by reads covering splice junctions in RNA-Seq data, which is why TopHat2 is used.

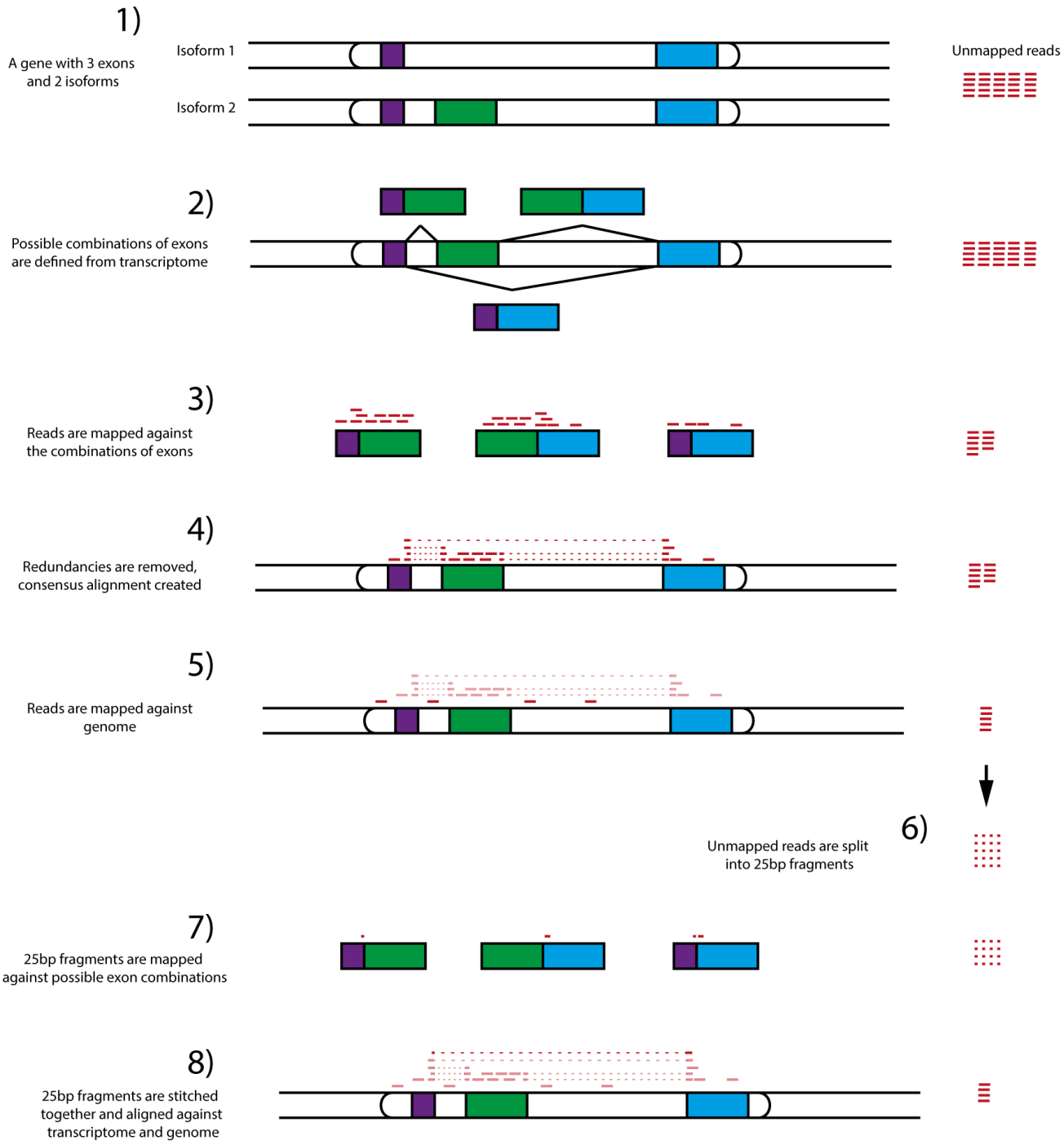

Figure 2.2: RNA-Seq read alignment using TopHat 2 and Bowtie. Sequencing reads are first aligned against transcriptome (1-4), then against genome (5). Unmapped reads are then fragmented (6) and mapped against a combination of genome and transcriptome (7-8), producing final alignment and final unmapped reads. Purple, green, and blue blocks - exons. Read blocks - sequencing reads. Dotted lines indicate spliced read alignment.

In terms of speed, this software combination is very effective. It can be seen in Table 2.3, that it is among the fastest software in terms of aligning time. Preparation time is reduced

by using previously constructed indices for human genome, removing TopHat's disadvantage in this area. It needs to be emphasised that data size of a single sample is usually up to 100x bigger than the one used to produce values in the table. It is then that TopHat's speed in mapping excels.

Since situations where a read can be aligned to multiple sites in the reference equally well occur, the pipeline is set up to accept up to 20 best alignments of equal score, i.e. a single read can be reported to align to up to 20 different places, if it matches equally well in all of them.

2.2.4 Fusion determination

Fusion determination is somewhat related to the splicing issue of RNA-Seq data. Here, a continuous read covering a gene fusion junction, partially matches two different strings in the reference, similarly to a read covering exon-exon junction.

All of the currently available gene fusion detection software packages depend on read-level information to determine gene fusions. In order to achieve the goal, reads that were left unaligned after splice junction resolution and alignment step are re-aligned as 25bp segments to the genome reference. Currently there are no available benchmarks for gene fusion detection software [101]. As such, they are usually chosen on the basis of applicability to the variety of the fusions investigated. Efficiency comparisons between available tools are usually done by a group publishing new software, and due to lack of unified assessment methods, according to most groups of authors their software outperforms the rest. There was no benchmark created for the purpose of this study as any artificially created gene fusion would not reflect the sheer diversity of gene fusion structure from bioinformatic point of view. There is simply too many unknowns at this point to benchmark based on the information currently available, especially if one is focusing on not typical events.

Table 2.4 presents an overview of fusion detection software efficiency, built on the data from the only review published on its own, not as a part of new software publication.

	Real set 1 (27 events)		Real set 2 (12 events)		Mock set (50 events)		Mock set 2 (0 events)
Software	TP	FP	TP	FP	TP	FP	FP
Bellerophontes[135]	NA	NA	NA	NA	12	13	0
ChimeraScan[136]	19	13,327	12	73,353	9	0	∼4,000
deFuse[67]	16	899	7	3,143	32	4	∼1,500
FusionFinder[137]	13	2188	NA	NA	41	10	∼6,000
FusionHunter[138]	8	16	5	6	40	2	0
FusionMap[139]	4	65	2	94	40	3	∼500
MapSplice[140]	NA	NA	NA	NA	40	12	∼6,500
TopHat-Fusion[66]	19	136,621	8	303,100	40	39	∼13,500

Table 2.4: Fusion identification software efficiency overview. Number of true-positive (TP) and false-positive (FP) fusion identification events considered. Not available values (NA) are due to inability of software to finish analysis of a set within 10 days. Approximations for Mock set 2 are due to inaccurate reporting by the authors. Data source: [141].

Fusion detection software determines support of a potential fusion by number of supporting reads (SR) and supporting pairs (SP). Differences between them are presented in Figure 2.3.

Mapped read segments, where one part of a read maps to a different location than the other create potential fusion candidate. At this point potential fusion sites are refined using information from other reads, i.e. other, initially unmapped reads are mapped to the stitched fusion site. This is the underlying method common to all current gene fusion detection packages.

Unfortunately, it seems that the authors that produced the raw data incorporated in the table used flawed methods. They did not seem to use default functionalities of one of the tools, namely TopHat-Fusion, which reduces number of false-positives 100-1000 fold in the experience of the author of this thesis, which is further supported by sensitivity comparison in Table 2.5. What is more, the authors report that Bellerophontes and MapSplice ran for more than 10 days, so they decided to abandon their analysis. It would be considered good practice to execute the problematic software using different hardware configuration and monitor the processes, reporting on the behaviour of the software. This was not done by the authors. It is possible that some other flawed methods were used to produce the data, but they are not

immediately obvious. As such, Table 2.4 has to be analysed particularly critically.

Gene fusion identification software's sensitivity can be assessed by its application to an existing dataset with known and confirmed gene fusions. Such comparison is presented in Table 2.5.

Software	Gold standard genes detected	Sensitivity	Total fusion genes detected
TopHat-Fusion[66]	16	22.54%	59
SOAPfuse[142]	10	14.08%	39
TRUP[143]	22	30.99%	63
FusionMap[139]	20	28.17%	205
deFuse[67]	27	38.03%	196
BreakFusion[144]	15	21.13%	130

Table 2.5: Fusion identification software sensitivity overview. Dataset consists of 183,946,388 101 base pair (bp) Illumina reads from MCF-7 from [145]. Gold standard genes are defined as 71 genes containing fusions previously verified by Polymerase Chain Reaction (PCR) and/or Sanger sequencing in the dataset. Adapted from: [56].

Event though sensitivity assessment is not ideal, as it does not encompass validation of newly found potential gene fusions, it does provide valuable information on software's performance. Seemingly, all fusion detection software packages underperform in this comparison with their sensitivity ranging from 14.08% to 38.03%. This underlines the potential for improvement across the board. Interestingly, the total number of fusion genes detected ranges from 39 to 205. Unfortunately, the comparison does not contain validation of previously unknown gene fusions thus leaving some questions unanswered.

TopHat-Fusion [66] has been implemented in the pipeline, as one of the most effective tools for fusion identification. In Table 2.4 presented in the previous section, it stands out as one of the tools that identified the most true fusion events. As previously mentioned, its false-positive hits presented the table are greatly exaggerated, as post-processing was definitely not applied. In Table 2.5 it is listed as having mid-range sensitivity.

The basis of the software's action lies mostly in unmapped reads. This is the reason why overmapping would introduce potential bias to gene fusion detection. Reads that are un-

mapped in the alignment process by Bowtie and TopHat2 are divided into segments, which are aligned to the reference genome. Potential fusion sites are defined in vicinity of mapped segments, supported by information provided by previously mapped paired-end reads. Detailed, step-by-step processing is shown and described in Figure 2.3.

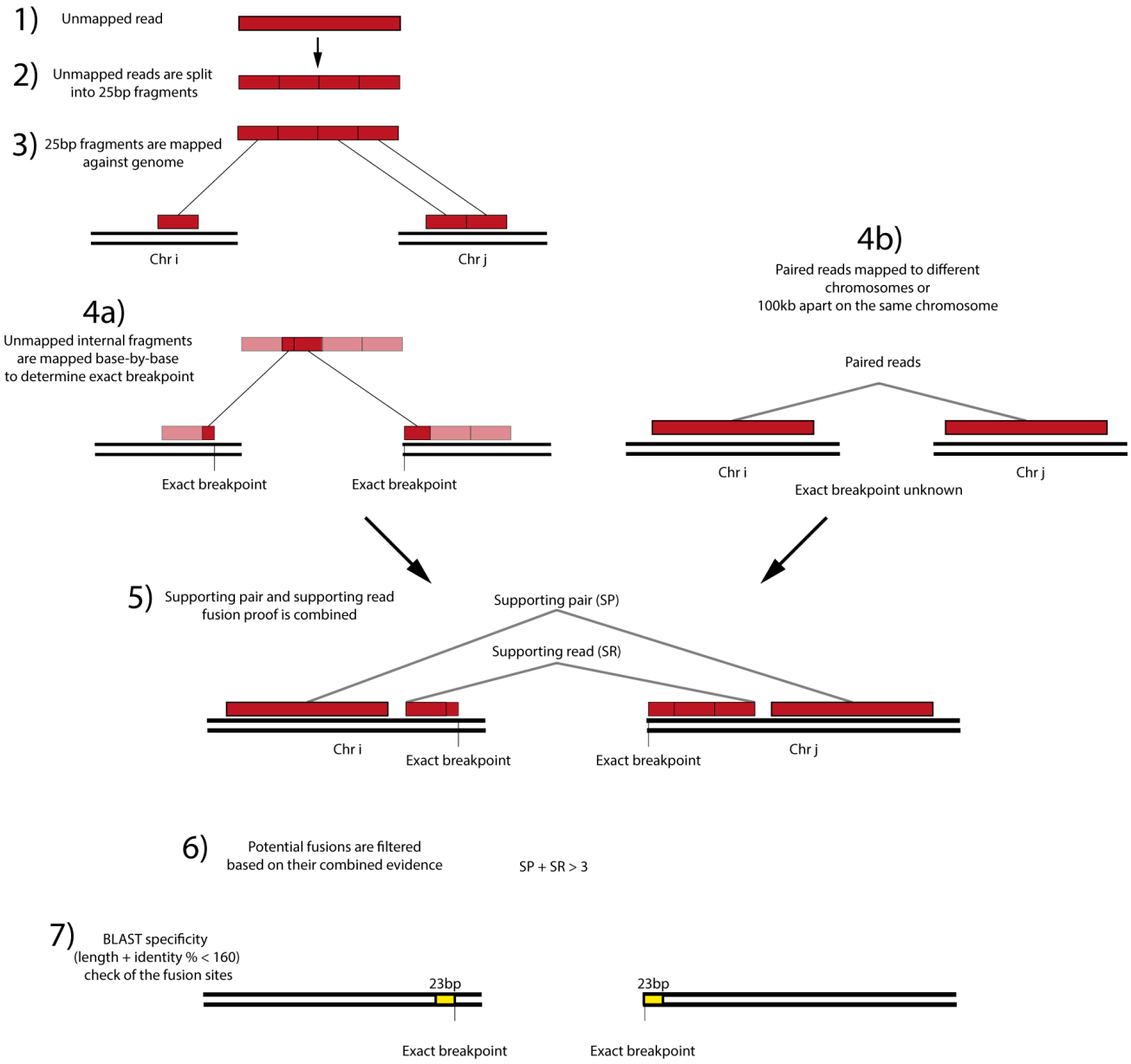

Figure 2.3: Gene fusion determination. Reads that were left as unmapped in alignment are split into fragments (1-2), which are mapped against genome (3). Internal fragments of reads are used to define exact fusion breakpoint (4a). Paired reads that were mapped far apart in alignment (4b) and mapped fragments (4a) are combined (5) providing total evidence and exact breakpoint for each fusion occurrence. In post-processing, fusion candidates are filtered based on their supporting proof (6), and fusion site specificity (7). Red blocks - sequencing reads. Yellow blocks - fragments used for breakpoint specificity check.

TopHat-Fusion not only processes potential fusion sites but, as also presented in Figure 2.3, performs post-processing of the findings. Since after the initial fusion candidate determination the output reaches hundreds of thousands, filtering is necessary to obtain a list of most likely candidates. Post-processing entails filtering based on the amount of proof, i.e. number

of reads and pairs that support the fusion, mapping at and around the fusion site, as well as by investigating uniqueness of the fusion site by performing a BLAST search [146] around the hypothetical breakpoint. Only the fusion breakpoints that have the sum of their length percentage match and identity percentage match at less than 160 are kept, removing breakpoints closely matching to other genomic locations, considering them as unspecific sequence.

The pipeline is expected to produce an output containing some false-positive results. The most efficient method of confirming a breakpoint after *in silico* analysis is RT-PCR [147], which is able to detect a fusion transcript if primers are designed on the sequences from both sides of a breakpoint.

2.3 Positive control analysis and pipeline optimisation

In order to assess the sensitivity and specificity of the pipeline presented for detecting gene fusions, it was used to analyse RNA-Seq data with previously verified fusions. The methods applied allowed for successful identification of fusions in four out of five cases, albeit with high number of false positive findings. This prompted optimisations to be introduced to the developed methods, which are described in detail in this section.

Any developed *in silico* methodology benefits from its verification using real datasets. Without testing it is difficult to predict what complications and exceptions may be encountered, which often necessitate alterations to be implemented. In order to assess the pipeline, it was used to analyse RNA-Seq data from five patients with 4 different gene fusions. To ensure an unbiased assessment, the analysis of these samples was performed with no knowledge as to the specific subtype of myeloid malignancy and specific gene fusions involved.

2.3.1 Methods

Whole blood samples were obtained from patients diagnosed with different myeloid malignancies with gene fusions previously confirmed with a combination of Sanger sequencing and FISH at Salisbury District Hospital, Salisbury, UK. After sample extraction of whole blood from peripheral veins, total RNA was isolated with TRIzol using a TRIzol kit from Life Technologies, Carlsbad, CA, USA. Total RNA quality control was performed by mea-

suring its quantity on Qubit Fluorometer, Life Technologies. All samples had at least the required 1μg of RNA material, normalised to a concentration of 50ng/μL. Total RNA solutes in ultrapure water were then sent to Wellcome Trust Center for Human Genetics at University of Oxford, Oxford, UK, who performed additional quality checks using Bioanalyzer, Agilent Technologies, Santa Clara, CA, USA. RNA integrity must be determined prior to sequencing, as degraded RNA can lead to increased noise in the resulting sequencing library, potentially affecting downstream bioinformatic analysis. All samples achieved RNA Integrity Number (RIN) [148] scores greater than 8/10, indicating little material degradation. Samples were then polyA selected to preserve only polyadenylated RNA species, followed by RNA-Seq library preparation, and sequencing. Paired-end stranded protocol with multiplexing for minimum of 100 million reads per sample, and 100bp per read was used in sequencing of polyA selected RNA, using HiSeq 2000 sequencer (Illumina, San Diego, CA, USA).

RNA-Seq reads provided by Wellcome Trust Center for Human Genetics at University of Oxford were obtained in demultiplexed FASTQ raw format. Bioinformatic analysis, including quality control, pre-processing, alignment, along with gene fusion identification was performed using the pipeline described in the previous chapter using University of Southampton Iridis 4 HPC.

Sample overview

Sample	Present gene fusion
1	*TEL(ETV6)-PDGFRB*
2	*FIP1L1-PDGFRA*
3	*TEL(ETV6)-ABL*
4	*BCR-ABL*
5	*BCR-ABL*

Table 2.6: Positive control samples, their designations and present gene fusions.

2.3.2 Results

Quality control (QC) did not indicate any problems with the samples, with all of them passing quantity and purity checks pre-sequencing, and with >90% of the base calls having Phred score $\geqslant$30, with no contamination or bias post-sequencing. As such no extra pre-

processing of the reads was necessary. As presented in Figure 2.4, sample 1 quality was concordant with high RIN score. A slight dispersion of base pair content at the beginning of the sequences (panel C) is expected and is the result of pseudo-random fragment selection in Next Generation Sequencing (NGS) technology [63]. QC plots for other samples (not shown) followed the same patterns as sample 1, indicating no issues with their quality.

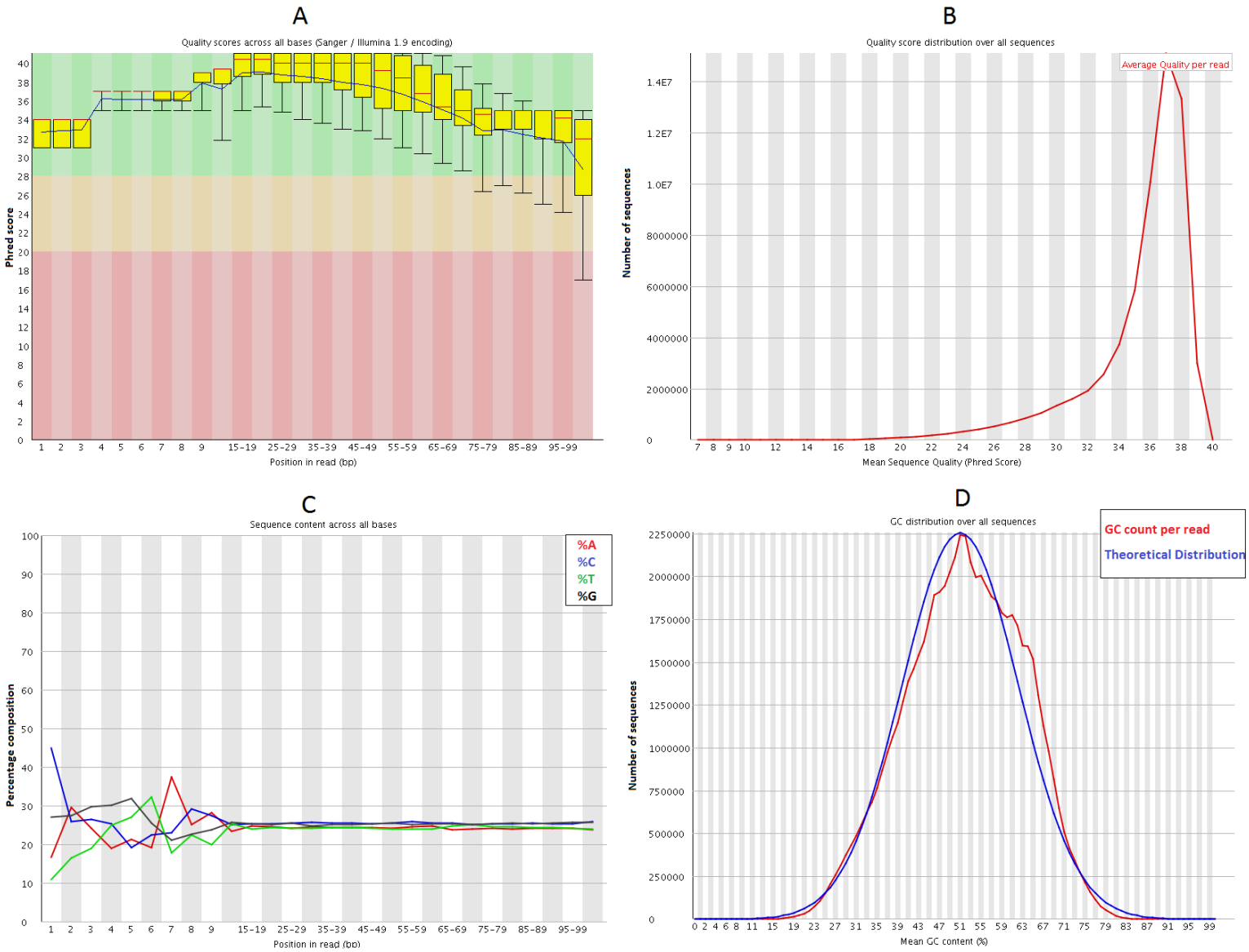

Figure 2.4: Quality control (QC) plots of sample 1. A - Sequence quality across all reads, per position. Red area - low, Yellow area - medium, Green - high; B - Sequence quality distribution; C - Nucleotide distribution per position; D - mean GC content across reads.

Total number of reads per sample is presented in Table 2.7, along with corresponding alignment percentage. On average, the following resources were used per sample: processor time 108h 57m (σ = 10h 3m), RAM = 23.43GB (σ = 0.84GB).

Sample	Total Read Count	Reads Aligned
1	126,594,042	91.5%
2	132,743,620	92.0%
3	149,528,154	92.4%
4	141,063,120	92.8%
5	126,017,040	93.8%

Table 2.7: Total read count and alignment percentage per sample.

High alignment percentage indicates that the quality of the samples was concordant with RIN scores. The initially unmapped reads used in gene fusion analysis revealed varying number of fusion candidates at different steps, as seen in Table 2.8. The initial number of *in silico* fusion candidates (FCs) ranged from 513,419 to 865,667, which is very high and demonstrates the challenges involved in identifying real gene fusions. The high number of initial FCs stems mostly from the very lenient first candidate selection algorithms of TopHat-Fusion. Splitting the reads into 25bp contigs and aligning those increases alignment notably at the cost of very high rate of misalignments, necessitating post-processing. The post-processing by TopHat-Fusion, entailing BLAST specificity check and supporting proof cut-off (used: SR+SP>3, default: SR>3 $\wedge$ SP>2) significantly reduced the number of FCs to the range of 54-97, but there were still too many FCs to permit efficient validation by RT-PCR. Further filtering of FCs to remove false-positives was therefore required.

Sample	Initial Fusion Candidates	FCs post-processing
1	513,419	89
2	676,306	54
3	865,667	97
4	681,029	77
5	545,280	92

Table 2.8: Fusion candidate count at different steps of filtering.

Among the post-filtering FCs, up to two per sample were chosen as the best candidates. Details of them are presented in Table 2.9. The best candidates were chosen on the basis of

their supporting proof (supporting reads and supporting pairs), as well as oncogenesis-related characteristics of the partner genes using COSMIC [149] and GeneCards [150] databases.

Sample	Fusion Genes	Breakpoints	Supporting evidence	Correct
1	*ANKRD27-BRD4*	Chr19:33165976-Chr19:15383943	SR:4 SP:0	No
	BCL7A-VPS37B	Chr12:122481958-Chr19:123353108	SR:4 SP:0	No
2	*FIP1L1-PDGFRA*	Chr4:54289447-Chr4:55141045	SR:3 SP:14	Yes
3	*ABL1-ETV6*	Chr9:133729450-Chr12:12022902	SR:42 SP:52	Yes
4	*ABL1-BCR*	Chr9:133710911-Chr22:23634727	SR:29 SP:12	Yes
5	*ABL1-BCR*	Chr9:133729450-Chr22:23632599	SR:105 SP:69	Yes
	NUP214-XKR3	Chr9:134074401-Chr22:17288972	SR:105 SP:69	Yes

Table 2.9: Details of best fusion candidates from positive control samples. SP - Supporting Pair, SR - Supporting Read.

The methodology allowed to correctly detect fusions in 4/5 samples. The undetected *PDGFRB-ETV6* (SR:58 SP:24) fusion in sample 1 has been initially detected by RNASeq but was filtered out during the automated part of filtering when older version of Tophat 2 (2.0.6) was used at the beginning of the project. This is discussed further in Section 2.3.3.

The identification of the best FCs was accurate, with most correctly identified candidates having high supporting evidence count (among top 3 FCs), and being previously known fusion cases. *FIP1L1-PDGFRA* fusion in sample 2 is an exception, with the lowest (3 SR and 14 SP) supporting evidence.

2.3.3 Pipeline optimisation

RT-PCR, while relatively inexpensive, is inefficient to quickly confirm the existence of a dozen or more potential breakpoints per sample, hence the need for additional filtering criteria of fusion candidates to be added to the pipeline.

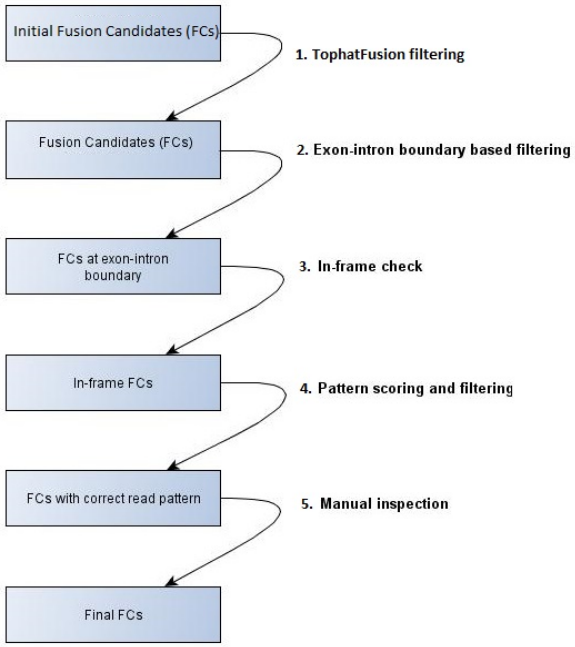

Figure 2.5: Five steps of Fusion Candidate filtering. Starting with Initial Fusion Candidates (FCs) initially determined by TopHat-Fusion, there are five steps leading to final FC determination. Steps 1-4 are automated, and step 5 manually performed. At step 1. TophatFusion filtering is employed with Basic Local Alignment Search Tool (BLAST) specificity check and read requirement check. At step 2. FCs with fusion sites at exon-intron boundaries are selected. At step 3. fusions that create in-frame products are selected, at step 4. a pattern based score is used to select fusion with multiple alignment patterns which are less likely to be artifacts, and at step 5. various criteria, detailed in-text.

The filtering optimisation steps (2-5), as presented in Figure 2.5, were designed to reduce the number of candidates needing verification, while minimizing the probability of excluding true fusions. In total, there are 5 filtering steps:

Step 1: TopHat-Fusion filtering

The first step is TopHat-Fusion filtering detailed in previous sections. It comprises of BLAST specificity check along with supporting proof cut-off (SR+SP>3).

Step 2: Exon-intron boundary based filtering

It is much more frequently observed for gene fusions to have breakpoints within introns rather than exons. Since RNA-Seq mainly sequences mRNA, the exact breakpoint of intronic fusions is unclear. From the mRNA perspective, the breakpoint is somewhere between two exons, and appears as being at the exon-intron boundary during analysis. As a trade-off of missing potential fusions that do occur within exons, this step reduces the list of candidates to those that are more likely to be real occurrences by filtering out the FCs with breakpoints not at the exon-intron boundary with 3bp tolerance.

Step 3: In-frame check

Since RNA-Seq sequences mainly mRNA, it is expected to only observe gain-of-function type of gene fusions, as loss-of-function mRNA undergoes nonsense-mediated decay (NMD). As such, gain-of-function gene fusions are unlikely to arise from an out-of-frame product. This check excludes out-of-frame fusion candidates with 1bp tolerance to accommodate runs of same nucleotides, accommodating shifts in potential breakpoint positioning caused by single nucleotide runs. This filtering step along with exon-intron boundary based filtering are shown to reduce the number of candidates in samples 1-5 by 71% on average, without affecting the correct fusion candidates.

Step 4: Pattern scoring and filtering

As previously observed [151], true positive gene fusions are commonly supported by multiple independent reads which map to slightly different locations surrounding the fusion junction and thereby create a staggered step-like structure. In comparison, false positive fusions tend to consist of one or two groups of duplicated reads which have identical locations (Figure 2.6). These attributes were used to create a pattern based scoring system, applied to select gene fusions that were more likely to be true positives.

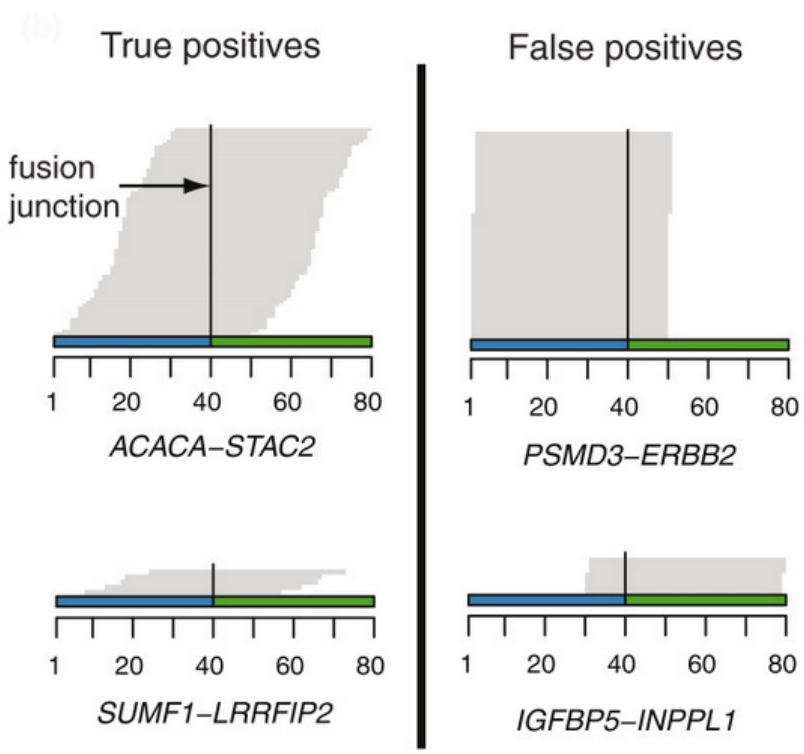

Figure 2.6: Short read alignment patterns across fusion junctions. True positives are characterised by stacking/ladder pattern, whereas false positives are characterised by reads that are mostly identical, and aligned mostly to one of the exons. Source: [151].

The read pattern scoring system, shown in Figure 2.7, was designed based on the number of read types at the same location. Read groups are defined by their locations and FCs are scored according to the number of read groups which span the fusion junction to a maximum of 3 groups. FCs containing a pattern consisting of only one type of reads are scored 1/3 (1) and considered false positive; Patterns consisting of two types of reads score 2/3 (2); Patterns consisting of three or more types of reads have a maximum score of 3/3 (3). Only FCs scoring 2 or higher pass the filter. This filtering step further reduced the number of FCs by 43% on

average, without affecting the correct fusion candidates.

Figure 2.7: Read pattern scoring system. 1 - Read group pattern scoring 1/3 (filtered out); 2 - Read group pattern scoring 2/3 (preserved); 3 - Read group pattern scoring 3/3 (preserved)

Step 5: Manual inspection

There are certain criteria, according to which fusion candidates are inspected, that are either impossible or problematic to automate. Manual examination of potential fusions entails investigation of normal gene function (UCSC [75], GeneCards [150]), gene toxicogenomic characteristics (Comparative Toxicogenomics Database [152]), gene involvement in cancer (COSMIC [149]), as well as copy number variation (CNV) assessment of the regions of both fusion sites. Lastly, candidates are examined to determine if they are a product of transcription read-through, rather than gene fusion - a common occurrence as approximately 4-6% of genes have been implicated to produce read-though transcripts [153].

Summary

The pipeline optimisations were implemented and the positive control sample set was re-ran. The series of modifications applied to the methods were proven to be effective at reducing the number of false candidates, and so far they have not eliminated correct FCs. As presented in Table 2.10 the modified filtering pre-manual inspection managed to remove 87% of FCs on average from samples 1-5 when compared to using TopHat Fusion filtering on its own.

Sample	Initial FCs	FCs (step 1)	FCs at exon-intron boundary(step 2)	In-frame FCs (step 3)	FCs with correct read pattern (step 4)	Final FCs (step 5)
1	513,419	89	60	22	10	2
2	676,306	54	41	15	9	1
3	865,667	97	53	24	13	1
4	681,029	77	51	23	13	1
5	545,280	92	55	10	6	2

Table 2.10: Fusion candidate count at different steps of filtering with added optimisation steps.

Although the filters are effective at removing false positives, they have the potential to also reduce the sensitivity, removing true FCs in future samples. However, the modifications were necessary to reduce the number of candidates, which was as high as 97 (Sample 3) before their introduction. Considering the available resources, confirmation of the reduced number of candidates (up to 2 per sample) is optimal, while confirmation of up to 97 candidates per sample would be completely impractical.

Even though modifications introduced to the methods applied in the analysis were not able to predict gene fusions with perfect certainty and left more than a couple of candidates, manual inspection is able to select good candidates with tolerable effectiveness. With 4/5 correct predictions in the positive control set, there is still possibility of improvement. However, it seems that any additional modifications to the read-based methods will result in potential loss of true positive candidates, especially if one considers that the FIP1L1-PDGFRA in Sample 2 is already on the low end of the filtering boundary.

Unsolved case

Relatively high effectiveness in detection rate of correct fusions in blinded analysis of the positive control dataset shows that the modified pipeline works to an extent. Failure to detect the fusion in 1 out of 5 samples is attributed to an unknown bug in TopHat-Fusion. Log inspection revealed that the FC was initially detected by the software, but it was filtered out during post-processing even though it should have passed all the filtering criteria. The TopHat-Fusion post-processing filters were applied manually one by one to ensure that the FC passes all the criteria. Indeed, during manual filtering, the FC was preserved, and reported in

the final list. In order to replicate the error, an artificial dataset containing reads indicating *ETV6-PDGFRB* fusion along with some background reads was created. However, the created dataset passed all the criteria, with no indication as to why the FC in sample 1 was removed. Inspection of the source code of TopHat-Fusion did not unravel the cause for mis-filtering of the candidate. The error was reported to the authors of TopHat-Fusion at the time. Towards the end of the project, 2 years after the positive control analysis was performed, TopHat was updated to version 2.1.0, and samples re-ran with positive identification of the *ETV6-PDGFRB* fusion.

2.4 Computational requirements

All software that the pipeline consists of is implemented in C, Perl, Python, and Java and the pipeline itself is implemented in bash. There is no graphical user interface to interact with the pipeline. It was designed to run on Linux HPC clusters, and cannot be used on Windows or OSX operating systems. Current implementation uses a node with 16 2.6GHz processors and 64GB of RAM for 30 hours per average sample, with the potential of parallel cluster usage, although with the reasonable run time it has not been the focus in the implementation. Minimum hardware requirements to run a sample in the pipeline within 300 hours are estimated to be 1 2GHz+ processor and 6GB of RAM. The pipeline does require a graphical processing unit.

2.5 Discussion

Construction of analysis pipeline provides ability to analyse RNA-Seq data for gene fusion determination. The power of the pipeline is based on its reproducibility and comprehensive ability to detect relevant events. It allows the user to perform a full gene fusion analysis starting from raw FASTQ sequencing data to gene fusion candidate determination, ready for RT-PCR confirmation. Designed to be run easily, it requires little configuration once it has been initially set up. The efficiency and efficacy of the pipeline was validated on real data from patients with myeloid malignancies, which was done on a set with previously verified gene fusions. This prompted optimisations to be introduced, and the modified pipeline was applied to investigate patients with myeloid malignancies with no known gene fusions in Chapter 4.

At this state, the analysis pipeline has some deficiencies in record keeping. It is planned to alleviate that issue using the methods outlined in Chapter 7.

Relatively read depth such as in Sample 2 (SR: 3, SP:14) can be explained by a couple of factors. Firstly, it is possible that the sample suffered from some degradation which can be reflected by low read depth at locations more prone to damage. Secondly, prevalence of disease-affected cells may simply be low, even though the disease phenotype is present. This is the most likely reason for low read depth and has to be carefully considered and accommodated when analysing gene fusions samples. Presence of low read-depth gene fusions is therefore expected and accommodated by sacrificing specificity through lowering default Tophat-Fusion filtering parameters from default SR>3 $\wedge$ SP>2 to SR+SP>3 and methods described in Section 2.3.3

The QC, pre-processing, and alignment elements of the pipeline are common between RNA-Seq data analysis for gene fusions and analysis of gene expression not necessarily involving gene fusions. These functionalities were adapted for the expression analysis purpose, as detailed in Chapter 3.

Chapter 3

Methods - Gene expression analysis pipeline

In addition to the direct detection of gene fusions as described in previous chapters, investigation of gene expression analysis to help elucidate the pathogenesis of myeloid malignancies by developing an RNA-Seq analysis pipeline for the analysis of effects that myeloid malignancy driver mutations have on gene splicing [154] was performed. This chapter presents the assembled analysis pipeline for gene expression assessment and comparison. Covering the rationale for the chosen workflow and its application, it discusses the current publicly available tools designed for the purpose.

3.1 Introduction

Gene expression investigation was one of the first applications of RNA-Seq [155][132]. While the Next Generation Sequencing (NGS) sequencing methods are not different between RNA-Seq for gene fusion investigation and RNA-Seq for gene expression investigation, their analysis differs greatly. The software for gene fusion investigation focuses primarily on initially unmapped reads [101] presuming that some of them are unaligned because they cover fusion junctions. Expression investigation tools focus on the aligned reads only. The general rationale is that the number of the aligned reads in any region gives a representation of abundance of messenger RNA (mRNA) species from that region [132].

Ever since 2008, RNA-Seq has been competing with microarrays for expression investigation experiments [132]. RNA-Seq's advantages over microarrays, such as greater accuracy with single base pair resolution, lower background noise, wider range of expression quantification, and lack of design-dependent constraints were appreciated early in the history of the method [156]. Since then, NGS methods have rapidly advanced, widening the gap between microarrays and RNA-Seq. Analytical methods of RNA-Seq benefit greatly from the transcriptomics field experience of microarray analysis, with methods such as DEGSeq, a tool for differential gene expression assessment, adjusting methods previously used for microarrays and applying them to RNA-Seq data [157].

As described in Section 2.1, processing any NGS data regardless of sequencing type, usually entails the following 4 steps [102]:

1. Quality Control (QC) - a routine examination of reads, ensuring that they are unbiased.

2. Pre-processing - removal of reads that failed QC and preparation of reads for alignment or assembly.

3. Alignment/Assembly - for organisms with a sequenced and assembled genome reference, the reads are aligned to it in order to identify location of their origin. For organisms with unknown genome reference the reads are matched with each other, assembling bigger constructs that can be analysed. As this thesis investigates human data only, which has a reference genome, the method for assembling reads is not discussed.

4. Analysis - varies, depending on the type of NGS data and aim of the investigation.

The following sections provide a detailed description of the software pipeline built for RNA-Seq data for differential gene expression analysis for the purposes of experiment described in Chapter 6, along with an overview of available software that can be used for this purpose.

3.2 Pipeline

3.2.1 QC, pre-processing, alignment

The process of application of RNA-Seq to a biological sample, including sample preparation, library creation, sequencing, and raw data generation, described in Section 1.4.2 on page 28, applies to all RNA-Seq experiments. As such, QC, pre-processing and alignment of raw data is the same regardless whether the aim of the analysis is gene fusion identification or differential expression investigation. QC, pre-processing, and alignment implemented in the pipelines for both varieties of the analysis is described in Sections 2.2.1 on page 46, 2.2.2 on page 49, 2.2.3 on page 52, respectively.

One caveat, specific to alignment of RNA-Seq for expression estimation, is the issue of duplicate reads. In the library creation process of NGS, fragmented DNA or cDNA is amplified using PCR. At this step, it is possible for PCR to overamplify a fragment, increasing its abundance during sequencing, and increasing the number of sequencing reads originating from it as the result. These artifacts can also arise due to noise in cluster generation, and sequencing artifacts such as poly-A and poly-N reads [158]. Reads originating from these artifacts can be observed as being mapped to exactly the same sequence in the genome/transcriptome. However, it is also possible to obtain a genuine read that aligns to exactly the same location in the genome/transcriptome as another genuine read, with both reads representing valid representation of the abundance of a particular mRNA transcript. This is caused by cDNA shearing at the same location in different molecules [158]. The estimated proportion of duplicate reads, both valid and PCR duplicates, obtained using Illumina paired-end sequencing in RNA-Seq libraries, was observed to be at the level of 42%-50% [159].

Removal of duplicates is a widely performed step in single nucleotide polymorphisms (SNPs) detection in whole genome or exome data, and is shown to improve efficiency of that type of analysis [160]. However, while SNP detection uses the number of supporting reads to determine certainty of a called SNP, RNA-Seq expression analysis relies entirely on the number of reads aligned to a region. With the current state of technology it is impossible to differentiate between a duplicate read arising due to overamplification and a valid duplicate read that by chance has the same sequence as another valid read. For this reason the removal of duplicate

reads would reduce the potential of type I error occurrence but increase the potential of type II error occurrence [159]. As such, it was chosen not to perform any duplicate removal.

3.2.2 Analysis

Differential expression analysis can focus on differences on a particular level of transcriptome. Therefore, software for expression analysis can be divided into four groups:

1. Gene level - the most general approach investigating differences between total gene expression levels.

2. Isoform level - more focused analysis of particular gene isoforms assessing their abundance levels, usually as a proportion of gene level expression.

3. Exon level - narrow approach, where expression of each exon is assessed separately, independently of other exons in the same gene.

4. Splicing - the most specific approach investigating occurrence and differences in utilisation of particular splicing events.

The following sections provide an overview of different approaches for the expression analysis levels, describing how and what tools were utilised in the pipeline.

Gene level and isoform level

Gene level analysis of a transcriptome in RNA-Seq quantifies the abundance of each gene under the assumption that the reads aligned to the annotated region of the gene represent relative abundance of mRNA produced by it [132]. Figure 3.1 provides a schematic overview of a gene and reads aligned in its region.

Figure 3.1: A typical gene with sequencing reads aligned in its region, schematic view. Black blocks - exons; red lines - sequencing reads. Dotted lines indicate reads aligning to splice junctions. Gene level expression analysis software quantifies and normalises the reads to provide a measure of expression for a gene.

Gene level analysis can be performed alongside isoform level analysis, where quantification of each isoform of a gene is executed. Here, different models are applied in order to assess which

reads mapped within a gene originate from particular mRNA molecules differing according to gene isoforms. A schematic overview of isoforms of a gene with reads aligned in its region is presented in Figure 3.2.

Figure 3.2: A typical set of isoforms of the same gene with sequencing reads aligned in their region, schematic view. Black blocks - exons; red lines - sequencing reads. Dotted lines indicate reads aligning to splice junctions. Isoform level expression analysis software applies a statistical model to assign the reads to the most likely isoform they originate from. Expression quantification and normalisation is then performed as in gene level analysis.

Since the beginning of RNA-Seq an abundance of software dedicated to differential expression analysis has been published, with SEQanswers database currently listing 25 [113]. The selection of software for the expression analysis pipeline was done on the following basis:

1. Ability to perform gene level and isoform level analysis

Some software focus either on gene level or isoform level analysis. In order to alleviate potential problems originating from different quantification methods, and to allow direct comparison of quantification between isoforms and genes, an approach that is universal for both levels was considered beneficial, but not required.

2. Expression quantification method allowing comparison with other studies

The most popular method of quantifying expression in RNA-Seq studies is fragments per kilobase of exon model per million fragments (FPKM) [68]. Others, such as proportional expression within a sample or proportional expression of isoforms in a gene would not allow for cross-experiment comparisons and potentially introduce issues with alternative grouping of samples with an experiment.

3. Ability to perform quantification of expression and differential expression analysis

Some software perform the analysis leaving the actual comparisons between samples to the user. Here, since the software was to be implemented in a pipeline, a more comprehensive approach was needed.

4. Effective between and within sample normalisation

It is important to account for both within and between sample variation that can cloud the results. Since the library creation and sequencing is a stochastic process, the total number of reads differ even between technical replicates. It is expected of a differential expression analysis tool to consider that when normalising the data - effective within sample normalisation is required. What is more, the number of reads between samples will differ for the same reasons. Normalisation has to consider them to allow for unbiased comparison of expression between conditions.

Another factor to recognize is the effect of overexpression of a gene, which can lead to incorrect quantification of expression of other genes in the same sample. Consider the following situation: There are two samples with exactly the same number of reads. In one of them, a single gene with normally high expression is overexpressed 100 fold, while other genes are expressed at the same level. A naive normalisation might not consider that, erroneously assuming that lower proportional levels of expression of other genes indicate their lower expression levels in that sample. It is therefore expected of a software to utilise appropriate normalisation for such problems.

5. Utilisation of biological replicates

Biological replicates provide a wealth of information on expected expression variance in a sample group. Any statistical modelling of expected biological variance cannot predict the variation more accurately than utilisation of actual biological replicates. Consideration of biological replicates in a comparison is therefore an important advantage of an expression analysis tool.

6. *De novo* transcript assembly

De novo transcript assembly is performed by reconstructing isoform models without transcriptome annotation, applied usually to organisms with no known or poorly annotated tran-

scriptome. This is not imperative for RNA-Seq data in humans, where the transcriptome annotation is well characterised and available. However, when the transcriptome of human cancer cells is considered, some features might not follow the annotation due to wide genomic changes characteristic of cancer. As such, *de novo* transcript assembly is acknowledged as a potentially useful feature of a software, though not necessary.

Table C.1 presents software packages for expression analysis, providing an overview on their fulfilment of the chosen criteria.

Software	Gene level analysis	Isoform level analysis	Expression quantification	Differential analysis	Sample normalisation	Replicate utilisation	*De novo* assembly
baySeq[161]	✔	✗	None	✔	2/3	✔	✗
BitSeq[162]	✗	✔	FPKM	✔	2/3	✔	✗
Cufflinks[68] (Cuffdiff 2)[163]	✔	✔	FPKM	✔	3/3	✔	✔
DESeq[164]	✔	✔	Size factors	✔	2/3	✔	✗
EBSeq[165]	✔	✔	None	✔	2/3	✔	✗
edgeR[166]	✔	✗	None	✔	2/3	✔	✗
NOIseq[167]	✔	✗	FPKM	✔	2/3	✔	✗

Table 3.1: Selected gene level and isoform level expression analysis software features. Only software that received an update or was first released in the last two years is listed. Sample normalisation scoring scheme - one point for each: within sample normalisation, between sample normalisation, handling of overexpressed genes monopolising reads.

Only part of the software packages investigated can be applied in the analysis pipeline, considering the selected criteria. Cufflinks [163] is the only package that covers all the areas, providing wide functionality. DESeq [164] was a good second choice, combining both isoform and gene level differential analysis. Still, its disadvantage was lack of normalisation in the situation of overexpressed genes monopolising reads, and using DESeq-specific size factors as an expression metric. A combination of BitSeq [162] for isoform level analysis and NOIseq [167] for gene level analysis could potentially be used, but their deficiency in normalisation could prove to be problematic. As such, it was decided to apply Cufflinks in the analysis pipeline.

Cufflinks quantifies gene level expression per gene by calculating FPKM from the reads aligning to a gene region, an example of which is presented in Figure 3.1. Isoform level (Figure

3.2) expression quantification is done by performing maximum-likelihood assignment of reads to isoforms in beta negative binomial model of distribution, followed by FPKM calculation for each isoform. Normalisation of overexpressed genes that can potentially introduce bias to the data is performed by geometric normalisation of the upper quartile of the most expressed genes. Differential expression tests are performed on log-transformed differences between quantified expressions using t-test and applying false-discovery rate (FDR) correction. For each differential expression test, samples are re-normalised in order to account for differences in numbers of reads between the samples [163].

In the analysis pipeline, Cufflinks was implemented to automatically processes the post-alignment output of TopHat 2 (Section 2.2.3, page 52). The processing is performed with no novel transcript discovery - no *de novo* transcript assembly as default. Further processing of Cufflinks output is performed by using CummeRbund package in R environment [68] to obtain lists of genes and isoforms that are significantly differentially expressed ($q<0.05$).

Exon level

Exon level differential expression analysis is done in a similar fashion to gene level analysis. However, in this case, the assessment is restricted to particular exons, investigating their usage, and allowing to assess splicing regulation.

The choice of software was in this case limited, and the criteria used to choose gene level and isoform level packages cannot be applied here. Therefore, the choice of exon level differential expression analysis software was made according to the three most important criteria - expression quantification method allowing comparison with other studies, utilisation of biological replicates, and maintenance status(updated or first published within the past 2 years). Considered software packages and their assessment is presented in Table 3.2

Out of all available software packages for the purpose, only three were maintained: DEXSeq[169], DSGseq[170], and MISO[100]. DSGseq was not used for its lack of exon expression quantification method. MISO was not used mainly because it utilises its own expression quantification metric - percentage spliced-in (PSI), which would not allow cross-experiment comparisons. DEXSeq, as the software meeting all the criteria was implemented in the analysis pipeline.

Software	Expression quantification	Replicate utilisation	Maintained
Alexa-Seq[168]	None	✔	✗
DEXSeq[169]	FPKM	✔	✔
DSGseq[170]	None	✔	✔
GPSeq[171]	Splicing ratio	✔	✗
MISO[100]	PSI	✗	✔
SOLAS[172]	Event inclusion	✔	✗

Table 3.2: Selected exon level expression analysis software features.

When quantifying expression of an exon, one important issue becomes apparent. Isoforms utilise exons differently, with some using alternative splice sites of an exon, which prevents application of absolute boundaries for exon annotation. DEXSeq exon expression quantification does not quantify each and every annotated exon. Instead, it produces a flattened exon annotation, as shown in Figure 3.3.

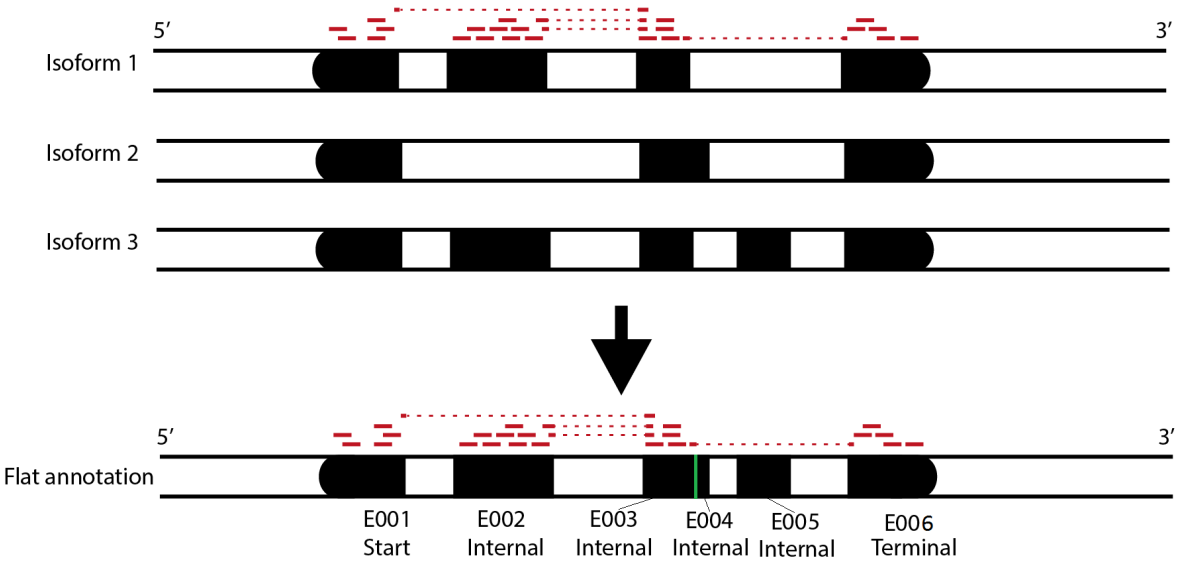

Figure 3.3: A typical set of isoforms of the same gene with sequencing reads aligned in their region with annotation collapsing for exon level analysis, schematic view. Black blocks - exons; red lines - sequencing reads. Dotted lines indicate reads aligning to splice junctions.

This annotation transformation allows for more accurate representation of an exon model for differential expression analysis. DEXSeq performs quantification for each flattened exon bin by calculating FPKM. Differential expression is tested using chi-square test with FDR

correction.

The default implementation of DEXSeq in the pipeline outputs only significantly differentially expressed exons ($q<0.05$). As an extension, the pipeline annotates each exon (specifically, DEXSeq flattened models of exons) as 'start', 'internal' or 'terminal' exons. This approach allows a more robust analysis that can characterise influence of an experimental factor on these three categories of exons.

Alternative polyadenylation (APA) sites are strings of polyadenylation signal that can be alternatively used in the gene expression process and can change mRNA's function by changing untranslated region (UTR) length [173] or, less frequently, shorten its coding length [174]. Their usage can be observed as very high sequence coverage regions in aligned RNA-Seq data. The exon-level analysis pipeline allows to investigate occurrences and usage of APAs by inference from differential exon expression by examination of relations between exon usage patterns and APA sites coverage. Possibility of APA site identification was one of the key requirements for in-depth analysis of splicing activity in Chapter 6.

Splicing

Apart from gene, isoform, and exon level analyses, expression data can also be investigated for differences in splicing events. The problem with the analysis of splicing differences across samples is the lack of their proper annotation. MISO [100], even though it was deemed unfit for exon level analysis, performs splicing event driven analysis well. Its strength lies in utilisation of its own annotation of a wide variety of splicing events, and it is the only available software with such annotation. Even though DEXSeq has previously been used for the investigation of differential splice junction usage and intron retention after its pipeline modifications [175], MISO offers a ready-tool for the purposes of splicing events analysis required for the experiment described in Chapter 6.

The first category of annotated splicing events covers alternative 3' splice sites (A3SS), presented in Figure 3.4. Note that even though it affects 5' end of an exon, A3SS derives its name from 3' end of an intron it affects. Its effect is the alternative usage of an exon, with isoform utilising upstream splice site during splicing of pre-mRNA, effectively splicing out less

of the preceding intron.

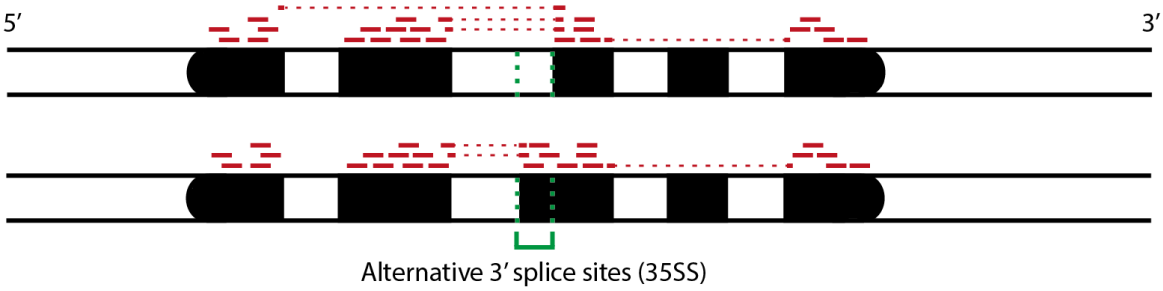

Figure 3.4: A typical example of alternative 3' splice site event (A3SS) with reads aligned in its region. Black blocks - exons; red lines - sequencing reads. Dotted lines indicate reads aligning to splice junctions. Green lines denote A3SSs.

Alternative 5' splice site events (A5SS), similarly to A3SS, utilise a different splice site during splicing of pre-mRNA, as shown in Figure 3.5. In this case however, it is the 5' end of an intron, and 3' end of an exon that is affected. This process effectively shortens the downstream intron, elongating the affected exon.

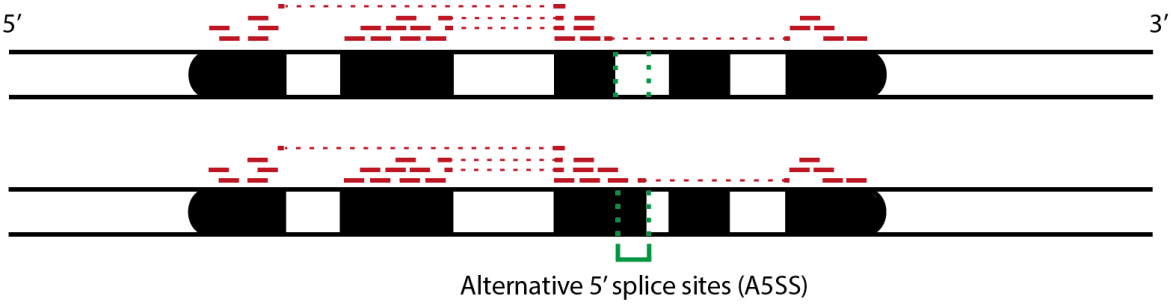

Figure 3.5: A typical example of alternative 5' splice site event (A5SS) with reads aligned in its region. Black blocks - exons; red lines - sequencing reads. Dotted lines indicate reads aligning to splice junctions. Green lines denote A5SSs.

Mutually exclusive exon (MXE) events occur when two exons are never found in the same isoform of a gene, as presented in Figure 3.6. It happens when the splicing machinery can include only one of them.

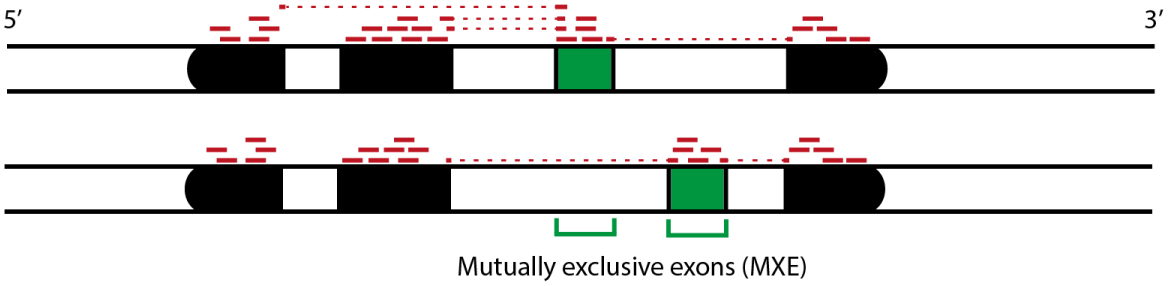

Figure 3.6: A typical example of mutually exclusive exons (MXE) with reads aligned in its region. Black blocks - exons; red lines - sequencing reads. Dotted lines indicate reads aligning to splice junctions. Green blocks denote MXEs.

Retained intron events (RI) are characterised by expression of an intron, as presented in Figure 3.7. Such event occurs when splicing machinery does not splice out an intron, preserving it in mRNA.

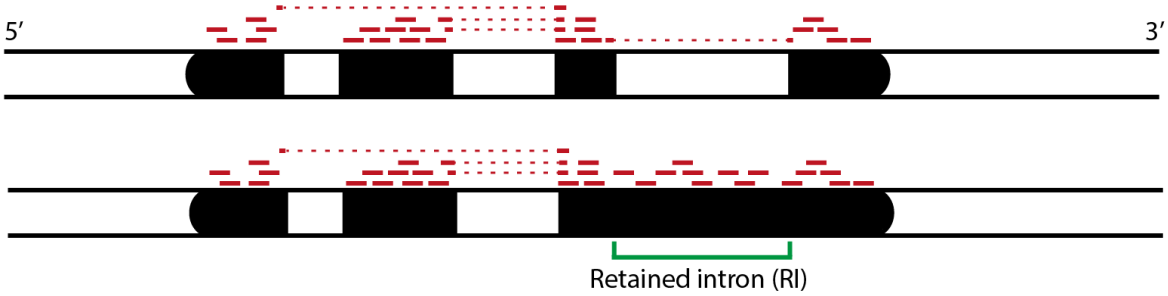

Figure 3.7: A typical example of a retained intron (RI) with reads aligned in its region. Black blocks - exons; red lines - sequencing reads. Dotted lines indicate reads aligning to splice junctions. Green lines indicate the RI.

Skipped exon events (SE) occur when an isoform of a gene does not include an exon, as depicted in Figure 3.8. These cases are caused by the splicing machinery splicing out the skipped exon along with its flanking introns in pre-mRNA processing.

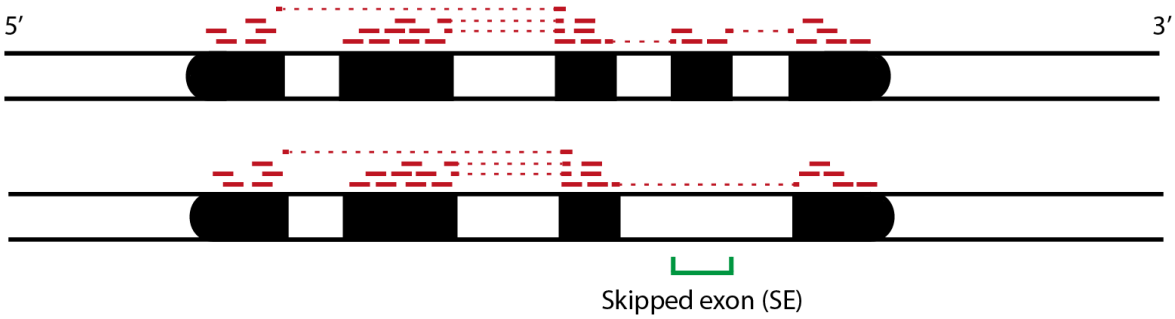

Figure 3.8: A typical example of a skipped exon (SE) with reads aligned in its region. Black blocks - exons; red lines - sequencing reads. Dotted lines indicate reads aligning to splice junctions. Green lines indicate the SE.

MISO provides an annotation for all the listed splicing events. It assesses their occurrence in analysed data by calculating percentage of junction reads giving rise to each possible outcome. The percentage of junction reads is termed percentage spliced in, or psi(Ψ). Test for differential event occurrence between samples is performed using Bayesian factors.

MISO is implemented in the pipeline to output significantly altered occurrences of splicing events using default parameters ($\Delta\Psi$>0.2, K>10).

3.3 Computational requirements

All the software implemented in the pipeline was written in Perl, Python, R, and C. The expression analysis pipeline was implemented in bash, with MISO execution performed in a virtual Python environment. Dependables required (versions used in the pipeline are given in brackets): Python (2.7) Intel Math Kernel Library (intel-mkl/2013.4 and mkl/11.0_4), bedtools (2.17.0), samtools (0.1.19), numpy (1.7.0), scipy(0.12.0), pysam (0.7.6) The pipeline was designed to be used in Linux HPC clusters, and cannot be used on Windows or OSX operating systems. Current implementation utilises a node with 16 2.6GHz processors and 64GB of RAM for 42 hours for the smallest possible 2x2 experiment. Minimum hardware requirements to analyse a 2x2 experiment in the pipeline within 300 hours are estimated to be 1 2GHz+ processor and 12GB of RAM. The pipeline does not require a GPU.

Chapter 4

Research - Analysis of samples with suspected gene fusions

Abstract

The analysis pipeline described in Chapter 2 was applied to RNA-Seq data from patients with myeloid malignancies with suspected cryptic gene fusions. The patients' DNA samples were previously examined using karyotyping and reverse-transcription polymerase chain reaction (RT-PCR), but the results did not indicate any causal gene fusions. This chapter provides a detailed description on how the developed analysis pipeline was applied in three of such cases, showing that its application can lead to the discovery of known and previously unknown gene fusions. Across 4 samples the developed methods allowed for identification and validation of at least 1 gene fusion per sample and up to 5 within a single sample, with overall accuracy of 29%.

4.1 Introduction

Since the discovery of Philadelphia chromosome [11] progress has been made in determining gene fusions as one of the causal genetic abnormalities in myeloid malignancies. Numerous other gene fusions, such as *PML-RARA* or *AML1-ETO* have been found at the roots of the diseases belonging to the group [5]. With the development of new advanced techniques allowing investigation of cytogenetically cryptic gene fusions, new cases, such as *CBFA2T3-GLIS2* have been uncovered [176]. RNA-Seq, as one of the recent methods allowing identification

of cytogenetically cryptic gene fusions, was applied for the analysis of clinical cases with suspected cryptic abnormalities.

Accurate prediction of cryptic gene fusions remains now the central challenge. While the wealth of data provided by RNA-Seq allows investigation of the highly dynamic transcriptome in detail, it remains an experimental method for gene fusion investigation. Verification of the findings using developed, clinically approved method is therefore of great importance, allowing for assessment of RNA-Seq's efficiency, and confirming the findings using approved standards.

4.2 Methods

Sample preparation and sequencing were performed as described in Section 2.3.1 on page 62. Confirmation of gene fusions found *in silico* was performed using RT-PCR with primers listed in Appendix, Table C.2, followed by Sanger sequencing where applicable. RT-PCR and Sanger sequencing were done in a laboratory uncontaminated with fusion constructs involving the genes investigated, maintaining pre- and post-PCR isolation.

Sample overview

Sample 6. Peripheral blood (PB) sample from patient with myeloproliferative neoplasm (MPN) with high *PDGFRB* expression and strong *in vitro* response to imatinib. No known gene fusions involving *PDGFRB* were found using RT-PCR, but a truncated *PDGFRB* transcript was suspected.

Sample 7. PB sample from patient with mastocytosis/eosinophilia and t(4;5)(q12;??) translocation, possibly a fusion involving *PDGFRA at 4q12*.

Sample 8. PB sample from patient with MPN with eosinophilia and a 46,XY,t(9;18)(p24;q12),t(14;18)(q21;q23) karyotype. Fluorescent *in-situ* hybridization (FISH) analysis had indicated a rearrangement of *JAK2* at 9p24 but the partner gene had not been identified. Subsequent work *in vitro* analysis demonstrated that the patient's cells were sensitive to the *JAK2* inhibitor ruxolitinib, supporting the hypothesis that a *JAK2* fusion was driving the disease.

Sample 9. PB sample from patient with hypereosinophilic syndrome (HES) and a t(2;4)(p24;q12) translocation. The patient had an *in vivo* response to imatinib but was *FIP1L1-PDGFRA* negative. Therefore possibly a novel *PDGFRA* fusion gene was suspected. The sample sequenced was taken after the patient went into remission.

4.3 Results and Discussion

4.3.1 Quality Control

The total RNA integrity for samples 6, 7, and 9 was very good, achieving ribonucleic acid integrity number (RIN) scores of 8.1, 7.2, and 8.0, respectively (Figure 4.1). Sample 8 was also sequenced despite its RIN score of 5.8 being below the recommended 7 for RNA-Seq and suggesting some RNA degradation for this sample. However, this had little influence on the RNA-Seq data, since, as shown in Figure 4.2, the sample exhibited only a slight decrease in sequence quality at the very ends of reads with approximately 50% base calls at positions 95-100 having Phred score >30 (Figure 4.2). However, a similar quality drop at the ends of reads was observed to be common across all samples in the set (not shown).

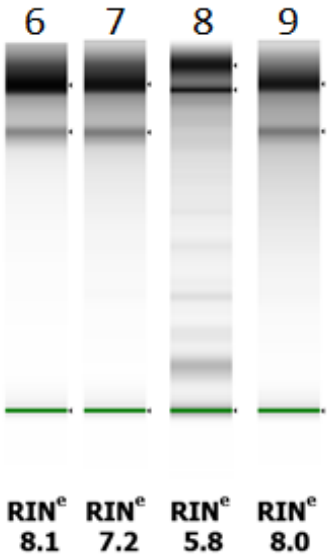

Figure 4.1: Ribonucleic Acid Integrity Number (RIN) profiles of samples 6-9. RIN scoring indicates the level of RNA degradation, lower score - more degradation, visible as signal in the lower part of spectrum.

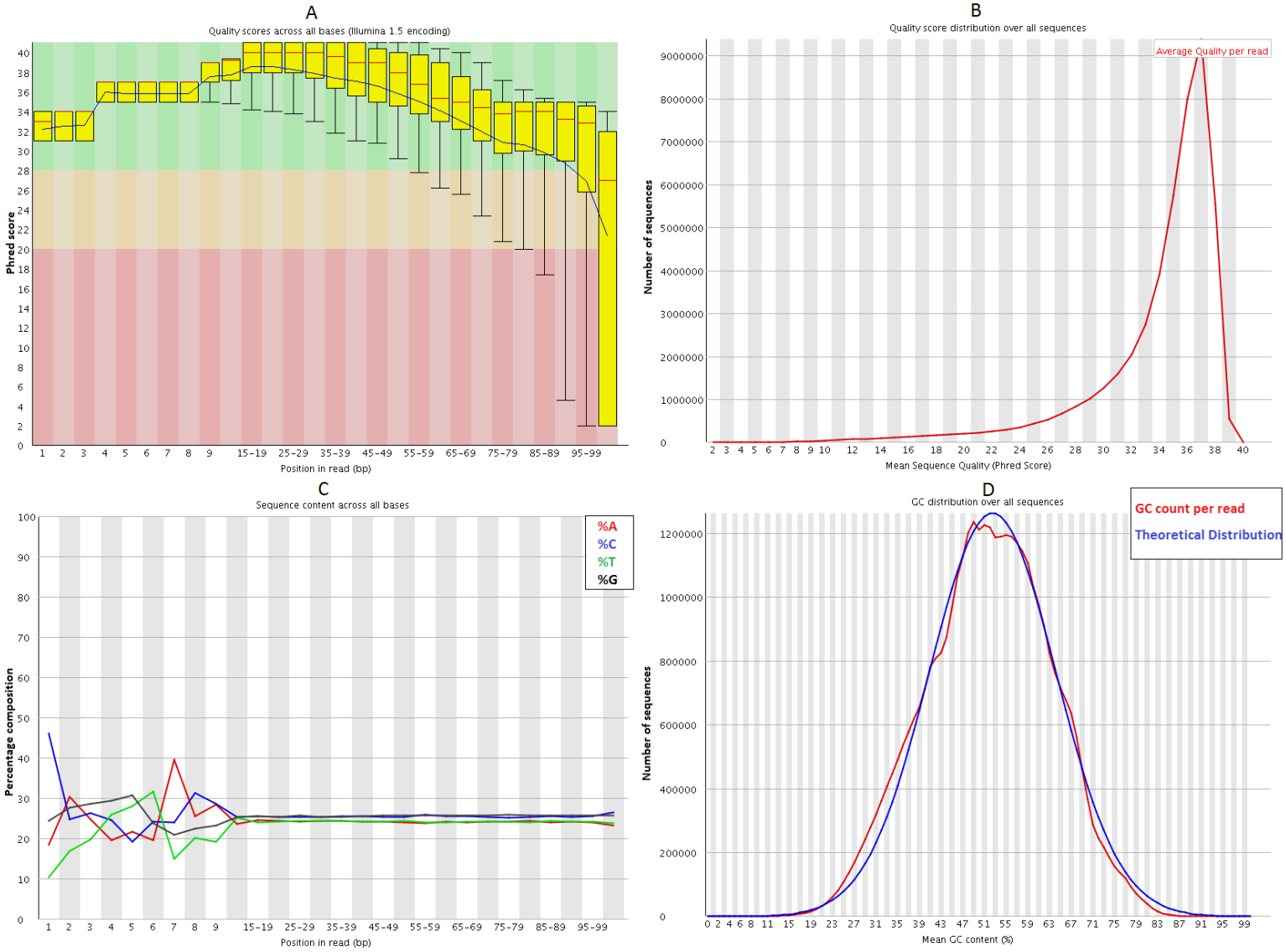

Figure 4.2: Quality control (QC) plots of sample 8. A - Sequence quality across all reads, per position, Red area - low, Yellow area - medium, Green - high; B - Sequence quality distribution; C - Nucleotide distribution per position; D - mean GC content across reads.

Altogether, the quality control did not indicate any problems with the samples, with >90% of the base calls having Phred score ⩾30, with no contamination or observed bias post-sequencing. The GC content of the samples was also in-line with expected distributions (Figure 4.2, C, D), exhibiting only expected nucleotide content variations at the beginning of the reads, which is the result of pseudo-random priming. As such no extra pre-processing of the reads was necessary.

The total number of reads per sample is presented in Table 2.7, along with corresponding alignment percentage. The alignment rate was on average 11% less than for samples 1-5 analysed in Chapter 2 (81.5% vs 92.5%), but similar to the rate reported in other studies [115]. The slightly worse alignment of sample 8, a decrease of 2.4% when compared to other samples, corresponds with its higher RNA degradation - RIN score of 5.8.

Sample	Total Read Count	Reads Aligned
6	73,785,686	82.7%
7	92,338,296	82.3%
8	93,665,732	79.0%
9	94,080,016	82.1%

Table 4.1: Total read count and alignment percentage per sample.

4.3.2 Fusion identification

The initially unmapped reads used in gene fusion analysis revealed varying number of fusion candidates at different steps, as seen in Table 4.2. The initial number of fusion candidates (FCs) ranged from 279,146 to 422,238. Post-processing by TopHat-Fusion, entailing BLAST specificity check and supporting proof cut-off (SR+SP>3) significantly reduced the number of FCs to the range of 30-157. Optimisations introduced in Chapter 2 allowed further reduction of the number of FCs to 0-20 per sample. In order to assess validity of the pattern scoring filtering system, FCs with pattern score of 1 were included.

Sample	Initial FCs	FCs (step 1)	FCs at exon-intron boundary(step 2)	In-frame FCs (step 3)	FCs with correct read pattern (step 4)	Final FCs (step 5)
6	279,146	157	94	13	12	11
7	372,757	95	53	22	21	19
8	385,824	30	21	8	8	2
9	422,238	66	49	21	17	0

Table 4.2: Fusion candidate count at different steps of filtering. Note that the read pattern selection step (step 4) filtered very few Fusion Candidates (FCs) (one in Sample 6, one in Sample 7, four in Sample 9). Since there was so few of those FCs, those from Sample 6 and Sample 7 were included in validation in order to ensure that pattern scoring does not remove valid gene fusions.

4.3.3 Confirmation

The identified FCs were subjected to RT-PCR and Sanger sequencing validation. Details of the results are presented in Table 4.3.

The discrepancy in sample validation between PCR and Sanger sequencing is explained by the fact that Sanger sequencing allows for actual sequencing of an amplified fragment, while PCR provides only approximate information on the length of the amplified fragment. Therefore, PCR, a prerequisite for Sanger sequencing, is much less accurate in fragment identification than Sanger sequencing. This is also the reason why Sanger sequencing was not performed where PCR was proven to give a negative result.

12 identified FCs from sample 6, were subjected to RT-PCR confirmation. It was revealed that in 8 cases there was fragment amplification at the expected size of the fusion junction, suggesting presence of gene fusions (Figure 4.3).

There were 21 FCs identified in sample 7 which were all amplified by RT-PCR. In 11 cases there was observed amplification of a fragment concordant with expected size of an amplified fusion junction (Figure 4.3). In 2 of those cases (*TIPARP-KLHL24*, *RAB20-ING1*) the amplified fragment content was verified by Sanger sequencing as originating from corresponding gene fusions, confirming the findings.

As the donor of sample 8 showed a rearrangement of *JAK2* by FISH and a strong *in vitro* response to *JAK2* inhibitor, fusions involving *JAK2* were investigated. A *BCR-JAK2* gene

Sample	Fusion (partner genes)	RNA-Seq pattern score	RNA-Seq supporting fragments	RT-PCR validation	Sanger sequencing validation
6	*ANPEP-B2M*	1	5	Yes	Negative
	GCA-BAZ2B	2	6	Yes	Positive
	CMTM6-DYNC1LI1	2	4	Yes	Negative
	MAML3-FLOT1	2	4	No	N/A
	ARID4A-PABPC1	2	6	Yes	Negative
	ELOVL5-PTP4A1	2	4	Yes	Positive
	ANP32A-EBC1D14	2	6	Yes	Negative
	USP32-TIMP2	2	4	No	N/A
	FNBP1-ZNF787	2	6	No	N/A
	ATXN1-PTP4A1	3	7	No	N/A
	PXN-UBC	3	9	Yes	Positive
	BRD-HNRNPUL1	2	4	Yes	Positive
7	*PTBP1-BBC3*	1	6	No	N/A
	KLF13-B2M	2	7	Yes	Positive
	PAPSS1-BANK1	2	4	Yes	Negative
	SIN3A-C15orf39	2	8	Yes	Positive
	KLHL2-DOCK2	2	6	Yes	Negative
	RAB3A-FKBP8	2	6	Yes	Negative
	WRD5-NFATC1	2	4	Yes	Negative
	KLF11-PTBP1	2	6	Yes	Negative
	ACTB-PTMA	2	5	No	N/A
	SMEK-SERTAD2	2	4	No	N/A
	GSE1-SCL7A5	2	4	Yes	Positive
	TCF25-ZC3H3	2	4	No	N/A
	TPM4-ZNF516	2	4	No	N/A
	TTYH3-MAD1L1	3	6	No	N/A
	AKAP17A-L3MBTL3	2	4	No	N/A
	LATS2-ING2	2	4	Yes	Negative
	GSE1-SLC7A5	2	4	Yes	Negative
	TIPARP-KLHL24	3	5	Yes	Positive
	DDX5-HNRNPH3	3	5	No	N/A
	RAB20-ING1	3	6	Yes	Positive
8	*BCR-JAK2*	3	8	Yes	Positive
	LOC729852-GLCCI1	3	7	No	N/A

Table 4.3: List of all fusion candidates as identified by RNA-Sequencing (RNA-Seq) along with their verification status. RNA-Seq pattern score refers to pattern scoring described in Section 2.3.3. Full genomic coordinates of fusion breakpoints are found in Table C.1.

fusion was identified among the 2 FCs, and verified by RT-PCR (Figure 4.3) and Sanger sequencing. To ensure that the other detected FCs are false positive, the second best FC was subjected to RT-PCR, and confirmed to be negative.

Sample 9, initially analysed as along with others was proven to be a remission sample due to mislabelling of sample identifiers in a laboratory. As a remission sample, it was expected to be devoid of gene fusions, the 17 FCs identified in it pre-manual inspection are considered to be false-positive or at least pathogenetically irrelevant.

Considering relatively low read depth at the fusion site in analysed FCs, it was expected that a notable part of them was to fail validation. At this level of uncertainty, due to low number of supporting reads and supporting pairs, PCR and Sanger sequencing confirmation is an absolute necessity.

Inclusion of FCs with pattern score of 1 from Sample 6 and Sample 7 confirmed validity of the employed scoring system. Out of 2 of them, one proved to be negative at PCR step, while the other was a confirmed negative at Sanger sequencing step.

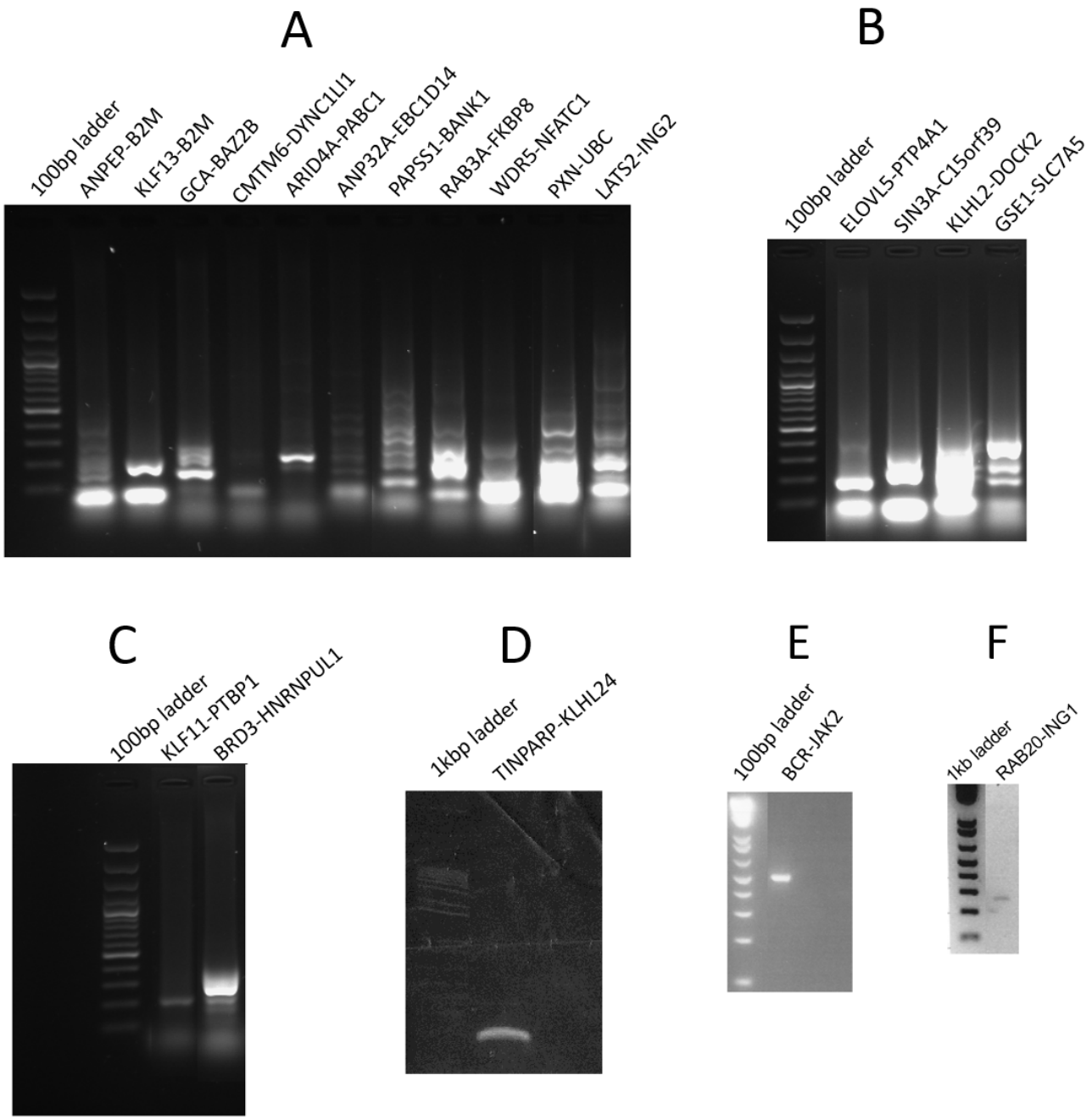

Figure 4.3: Gene fusion validation Reverse-Transcription Polymerase Chain Reaction (RT-PCR) gels. Presence of a fragment concordant with predicted size detailed in Table C.2 indicates amplification of corresponding fusion junction. This is consistent with a fusion being real, but full confirmation requires Sanger sequencing of the products.

4.3.4 Significance of pathological changes

GCA-BAZ2B

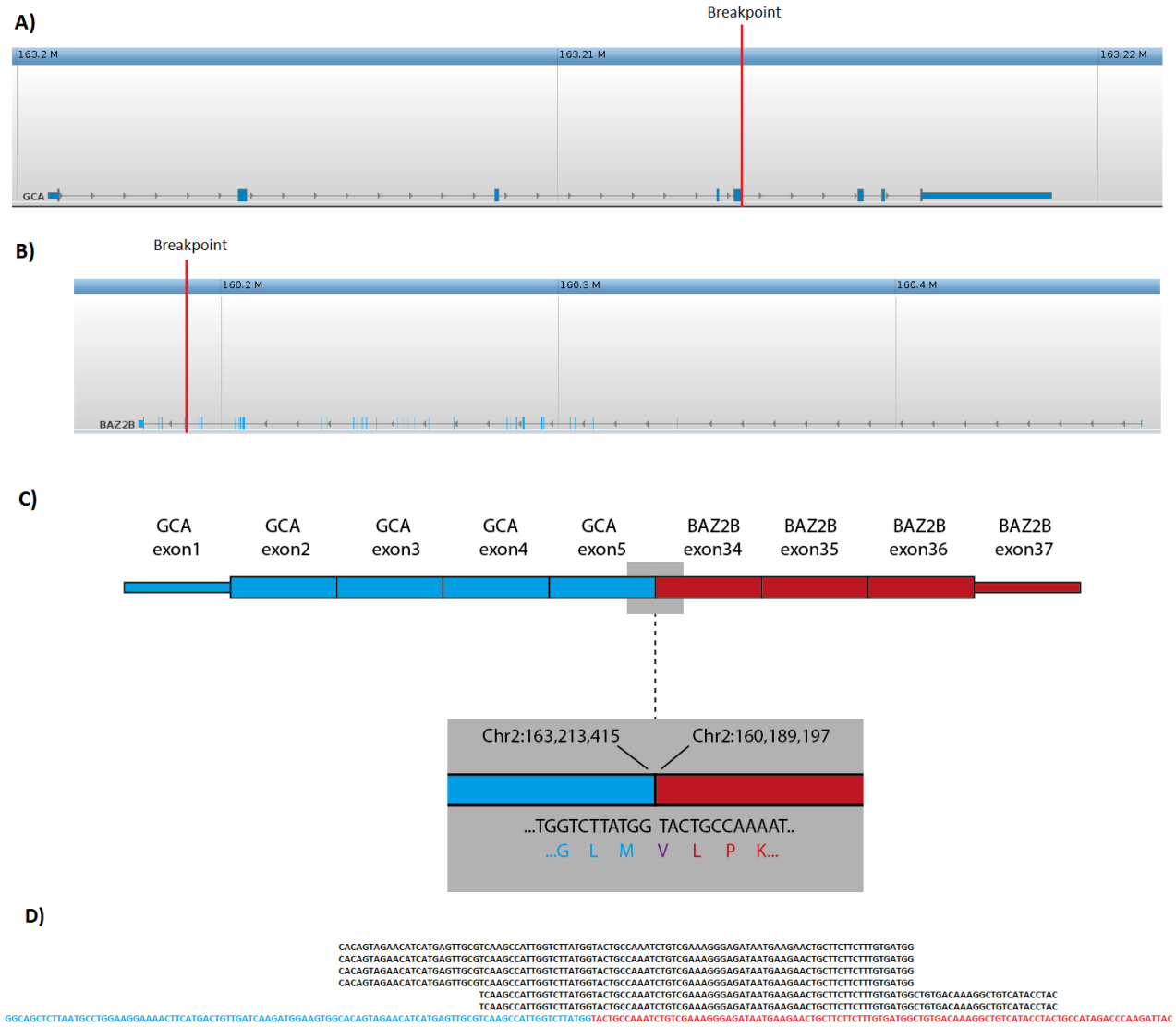

Figure 4.4: *GCA-BAZ2B* gene fusion. Blocks represent exons, thin blocks represent UTR parts of exons. A - transcriptome annotation in the region of *GCA* on chromosome 2 with the breakpoint marked; B - transcriptome annotation in the region of *BAZ2B* on chromosome 2 with the breakpoint marked; C - putative fusion transcript with DNA and predicted amino acid sequence at the breakpoint; D - Read pileup around fusion site.

In sample 6 from a patient with MPN with high *PDGFRB* expression and strong *in vitro* response to imatinib one of the confirmed gene fusions was *GCA-BAZ2B*. The total support for the fusion was 6 reads and pairs as well as 2/3 scoring pattern. The fusion joins last exon of *GCA* with the first exon of *BAZ2B* and is presented in Figure 4.4. Both genes are found on

the same chromosome, approximately 3MB apart, they might be a result of a large inversion or a duplication.

GCA (Grancalcin, EF-Hand Calcium Binding Protein) is a gene coding for protein responsible for granule-membrane fusion and degranulation [150]. There were no reported cases of *GCA* being involved in gene fusions.

BAZ2B (Bromodomain Adjacent To Zinc Finger Domain, 2B) is a protein coding gene suspected to play a role in transcriptional regulation [150]. As with *GCA*, *BAZ2B* was not reported to be involved in gene fusions.

Disruption of regular function of a transcriptional regulator can lead to deregulation of expression of other genes. In this case, a fusion construct that contains part of a transcriptional regulator, *BAZ2B*, may result in a functional protein, deregulating transcriptional machinery and acting as an oncogenesis driver.

ELOVL5-PTP4A1

ELOVL5-PTP4A1 gene fusion from the same sample (6) was supported by a total of 4 fragments, scoring 2/3 for the read pattern. The fusion gene transcript contains regulatory 5' untranslated region (UTR), exon 1, of *ELOVL5*. The rest of the fusion transcript consists of exons 2-6 of *PTP4A1*, as presented in Figure 4.5. Both genes are in relatively close proximity on chromosome 6, approximately 11Mbp apart. Considering the directionality of the genes, the most likely underlying reason for the fusion is either inversion or duplication followed by inversion.

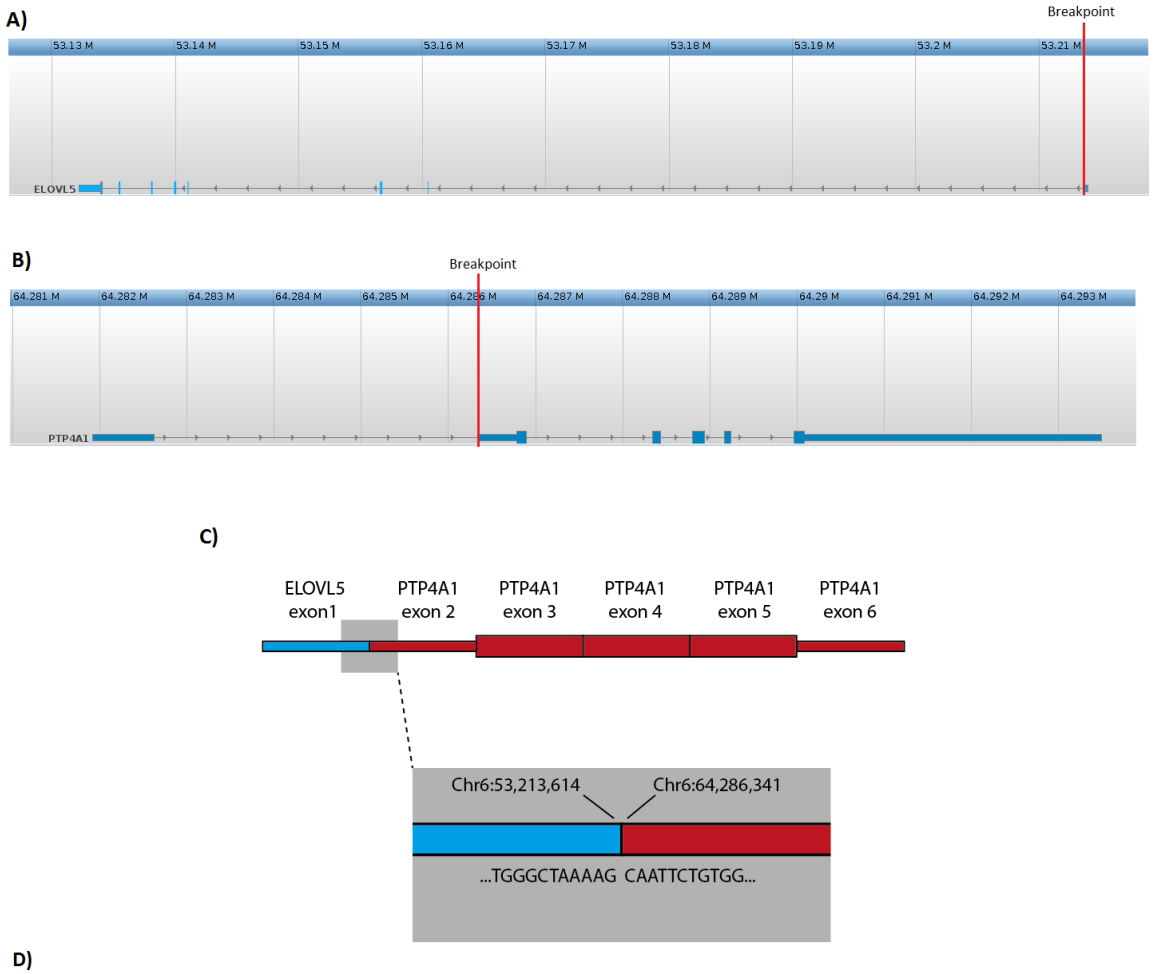

Figure 4.5: *ELOVL5-PTP4A1* gene fusion. Blocks represent exons, thin blocks represent Untranslated Region (UTR) parts of exons. A - transcriptome annotation in the region of *ELOVL5* on chromosome 6 with the breakpoint marked; B - transcriptome annotation in the region of *PTP4A1* on chromosome 6 with the breakpoint marked; C - putative fusion transcript with DNA and predicted amino acid sequence at the breakpoint; D - Read pileup around fusion site.

ELOVL5 (ELOVL family member, Fatty Acid Elongase 5), a protein coding gene, is associated with metabolic disorders [150]. It has not been reported to be a fusion partner.

PTP4A1 (Protein Tyrosine Phosphatase Type IVA, Member 1) codes for a protein that plays a role in regulatory cell processes [150]. There are no publications indicating that it was observed as a fusion partner.

Little is known about involvement of protein tyrosine phosphatases in myeloid malignancies, but it has been reported that genes belonging to the group are overexpressed in AML [177]. In the observed case of gene fusion, *PTP4A1*, a protein tyrosine phosphatase is affected by regulatory region of *ELOVL5*. As such, it is possible that this particular fusion is the oncogeneic driver in sample 6.

PXN-UBC

The third detected fusion in sample 6 was *PXN-UBC*. The total number of reads supporting the fusion was 9, with 3/3 scoring pattern. The fusion transcript is made of coding exon 1 of *PXN* and last, coding, exon 2 of *UBC*, as presented in Figure 4.6. The genes involved in the fusion are approximately 5Mbp apart in the same directionality. Therefore, it is likely that the fusion arose due to a deletion.

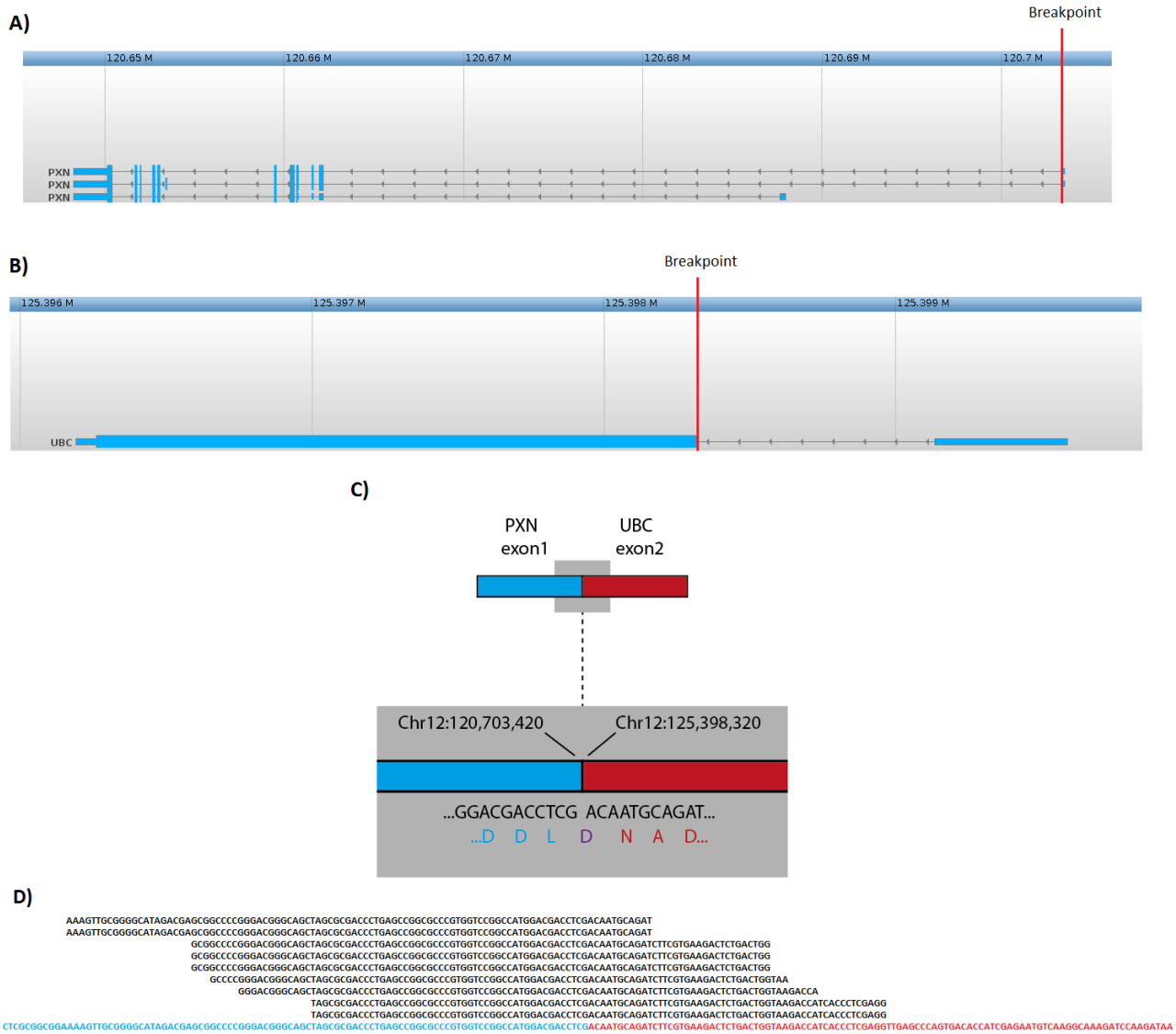

Figure 4.6: *PXN-UBC* gene fusion. Blocks represent exons, thin blocks represent Untranslated Region (UTR) parts of exons. A - transcriptome annotation in the region of *PXN* on chromosome 12 with the breakpoint marked; B - transcriptome annotation in the region of *UBC* on chromosome 12 with the breakpoint marked; C - putative fusion transcript with DNA and predicted amino acid sequence at the breakpoint; D - Read pileup around fusion site.

PXN (Paxillin) codes for a protein involved in cell adhesion [150]. There are no reported gene fusions involving this gene.

UBC (Ubiquitin C) is a protein coding gene, with the product being involved in the process of ubiquitination, an important process associated with a range of cell regulatory mechanisms [150]. It was not been previously reported to be a fusion partner.

This particular gene fusion is most likely of a passenger type. The genes involved have no known connection to oncogenesis. Disruption of their regular functions is more likely to result in other maladies rather than cancer. Similarly, the fusion construct created is very unlikely to be of gain-of-function type in the context of cancerogenous properties.

BRD3-HNRNPUL1

The last gene fusion detected in sample 6 was *BRD3-HNRNPUL1*. It involves exon 1 of *BRD3* which consists of 5'UTR. The remainder of the fusion transcript includes all exons of *HNRNPUL1* starting from exon 2, which is shown in Figure 4.7. The genes are situated on chromosomes 9 and 19, respectively. This is an example of interchromosomal gene fusion.

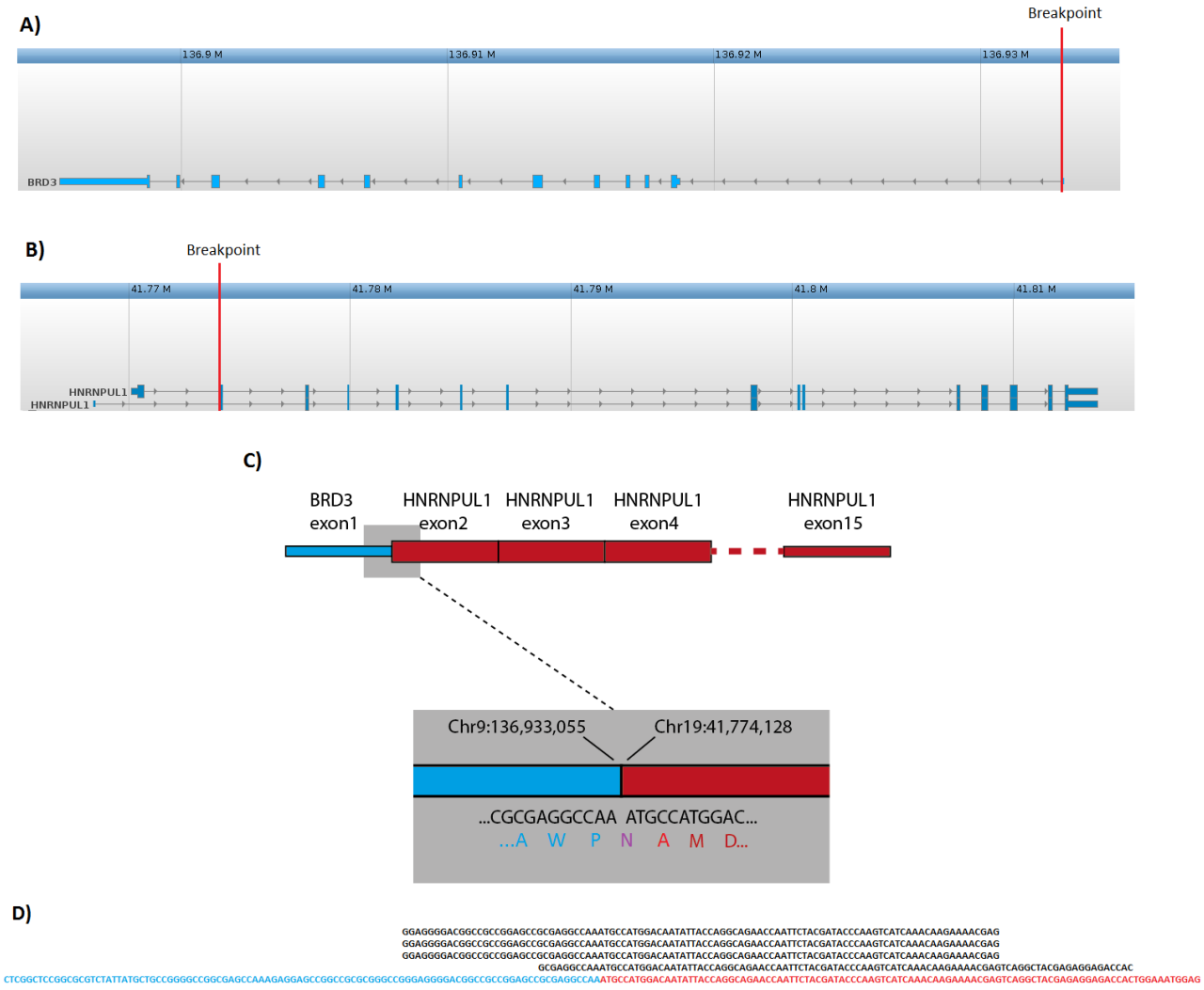

Figure 4.7: *BRD3-HNRNPUL1* gene fusion. Blocks represent exons, thin blocks represent Untranslated Region (UTR) parts of exons. A - transcriptome annotation in the region of *BRD3* on chromosome 9 with the breakpoint marked; B - transcriptome annotation in the region of *HNRNPUL1* on chromosome 19 with the breakpoint marked; C - putative fusion transcript with DNA and predicted amino acid sequence at the breakpoint; D - Read pileup around fusion site.

BRD3 (Bromodomain Containing 3) is a gene of unknown function, classified based on its homology to *RING3* a kinase [150]. The gene has been previously observed in a fusion

with *NUTM1* in NUT midline carcinoma [178]. However, the previously observed fusion was thought to be of importance because of disruption of *NUTM1*, rather than *BRD3*.

HNRNPUL1 (Heterogeneous Nuclear Ribonucleoprotein U-Like 1) codes for a protein involved in RNA transport [150]. There are currently no gene fusions including *HNRNPUL1* reported.

Even though *BRD3* has been previously observed as a fusion partner in a cancer patient, it is unlikely that *BRD3* specifically was connected to oncogenesis. Rather, it was its partner's disruption that likely led to the development of cancer. However, the fact that *HNRNPUL1*, which is involved in RNA transport, is regulated by *BRD3*'s regulatory region, may have a detrimental effect on gene expression and may be the oncogenesis driver. However, this is not very likely and the fusion is more likely to be of a passenger type.

KLF13-B2M

The first detected fusion in sample 7 from a patient with mastocytosis and eosinophilia was *KLF13-B2M*. It was supported by a total of 7 RNA-Seq reads and scored 2/3 for its read distribution pattern. The resulting gene fusion transcript involves exon 1 of *KLF13* followed by exons 3 and 4 of *B2M*, as presented in Figure 4.8. Both genes are found on chromosome 15, approximately 13Mbp apart. With the directionality of the genes being the same, the most likely cause for the fusion is either deletion or partial duplication.

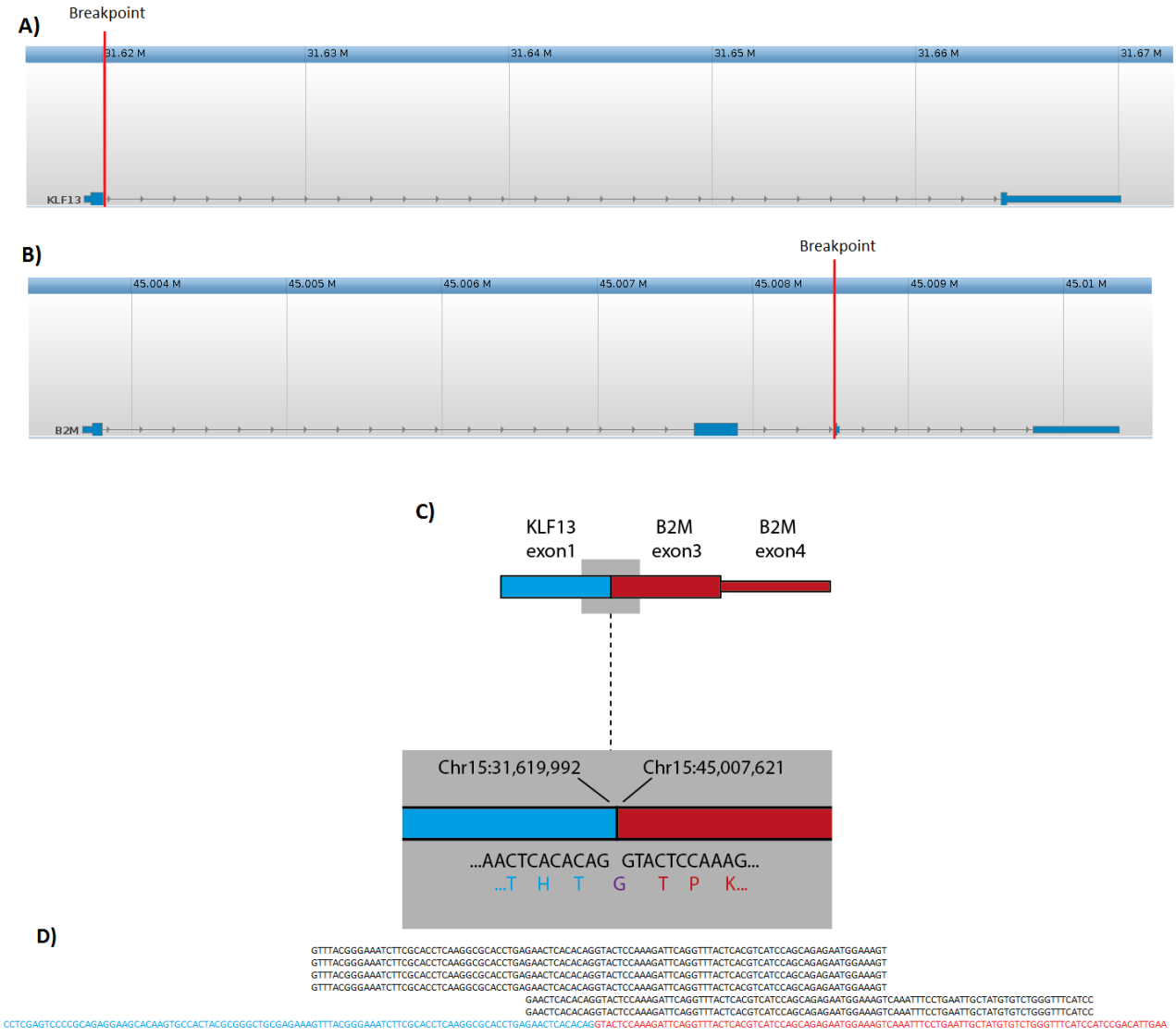

Figure 4.8: *KLF13-B2M* gene fusion. Blocks represent exons, thin blocks represent Untranslated Region (UTR) parts of exons. A - transcriptome annotation in the region of *KLF13* on chromosome 15 with the breakpoint marked; B - transcriptome annotation in the region of *B2M* on chromosome 15 with the breakpoint marked; C - putative fusion transcript with DNA and predicted amino acid sequence at the breakpoint; D - Read pileup around fusion site.

KLF13 (Kruppel-Like Factor 13) is a transcription factor with zinc-finger domains, involved in transcription repression [150]. It has not been previously observed in gene fusions.

B2M (Beta-2-Microglobulin) is a protein coding gene. The protein product is involved in antibacterial activity in serum [150]. There are no reports of the gene being involved in a gene fusion.

Disruption of transcription regulation mechanisms may be the driver of oncogenesis. In the case of detected *KLF13-B2M* gene fusion, *KLF13* is an affected transcription factor. As it contributes the first exon to the fusion construct, its regular function is removed and may lead to a loss-of-function driven oncogenesis.

SIN3A-C15orf39

The second gene fusion in sample 7 was *SIN3A-C15orf39*. Supported by a total of 8 reads and scoring 2/3 for its read pattern, the fusion involved exon 1 of *SIN3A* along with exons 2 and 3 of *C15orf39*. The overall structure of the fusion transcript is shown in Figure 4.9. Both genes are situated on chromosome 15 in a relatively close proximity, with the breakpoints being approximately 250kbp apart. Opposite directionality of the genes suggests that the fusion arose due to either large inversion or a duplication event followed by inversion.

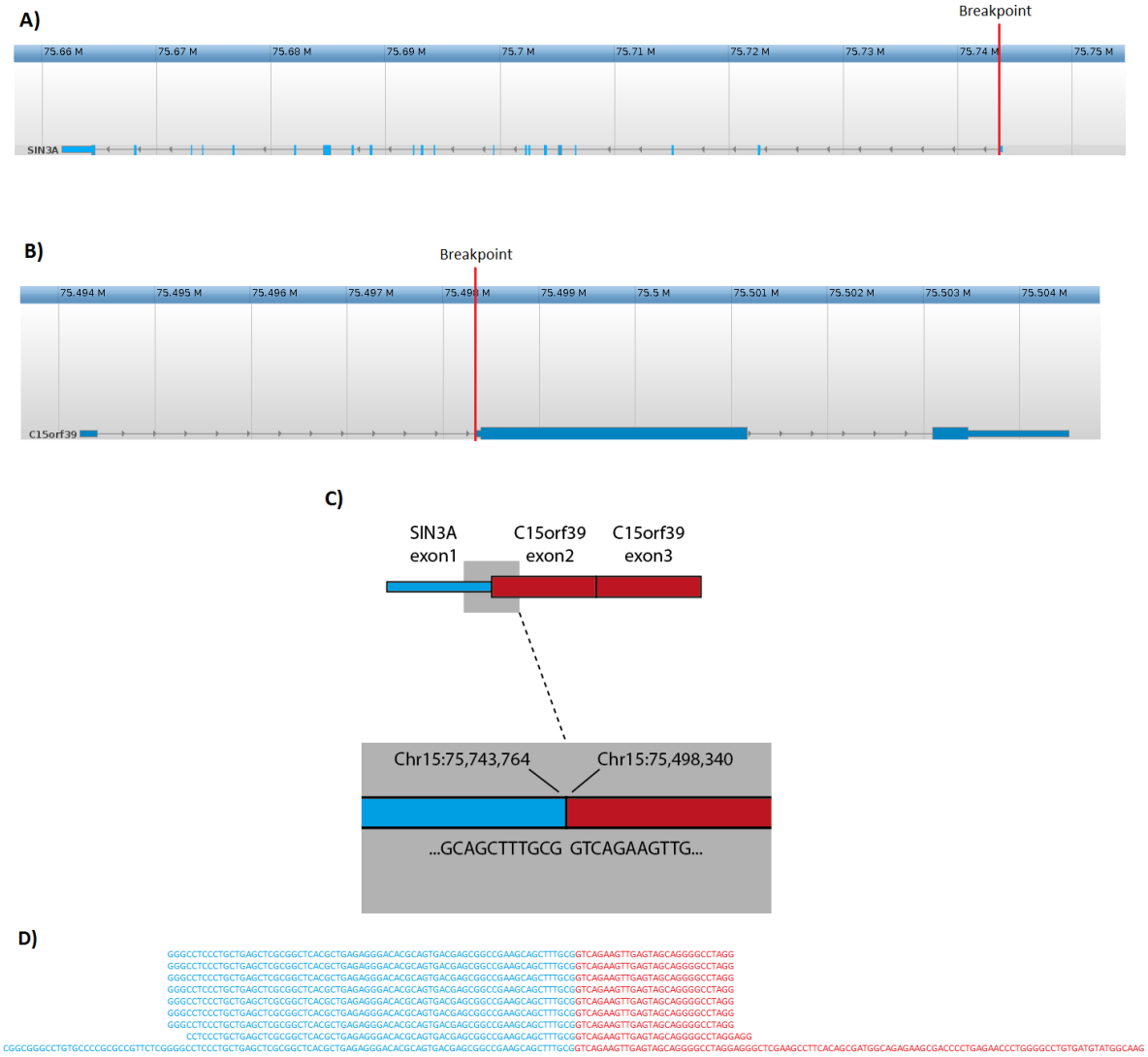

Figure 4.9: *SIN3A-C15orf39* gene fusion. Blocks represent exons, thin blocks represent Untranslated Region (UTR) parts of exons. A - transcriptome annotation in the region of *SIN3A* on chromosome 15 with the breakpoint marked; B - transcriptome annotation in the region of *C15orf39* on chromosome 15 with the breakpoint marked; C - putative fusion transcript with DNA and predicted amino acid sequence at the breakpoint; D - Read pileup around fusion site.

SIN3A (SIN3 Transcription Regulator Family Member A) codes for a protein involved in trascriptional regulation [150]. There are no previously reported fusions involving the gene.

C15orf39 is a gene of unknown function [150]. It was not previously observed as a gene fusion partner.

Similarily to the case of *GCA-BAZ2B*, a gene involved in transcription regulation is disrupted here. *SIN3A*, contributing its first exon to the fusion construct, is likely to have its function lost if the fusion event did not arise due to a duplication and it affected the original copy of the gene. It is possible that hindering *SIN3A*'s function lead to oncogenesis.

GSE1-SCL7A5

Another gene fusion detected in sample 7 was *GSE1-SCL7A5*. It scored 2/3 for the pattern scoring and had total support of 4 reads. the fusion transcript includes exon 1 of *GSE1*, followed by exons 3-10 of *SCL7A5*, as presented in Figure 4.10. The genes involved are situated on the same chromosome, chromosome 16, approximately 2.2Mbp apart. Since the genes have opposite directionality, it is most likely that the fusion was due to an inversion or a duplication followed by inversion.

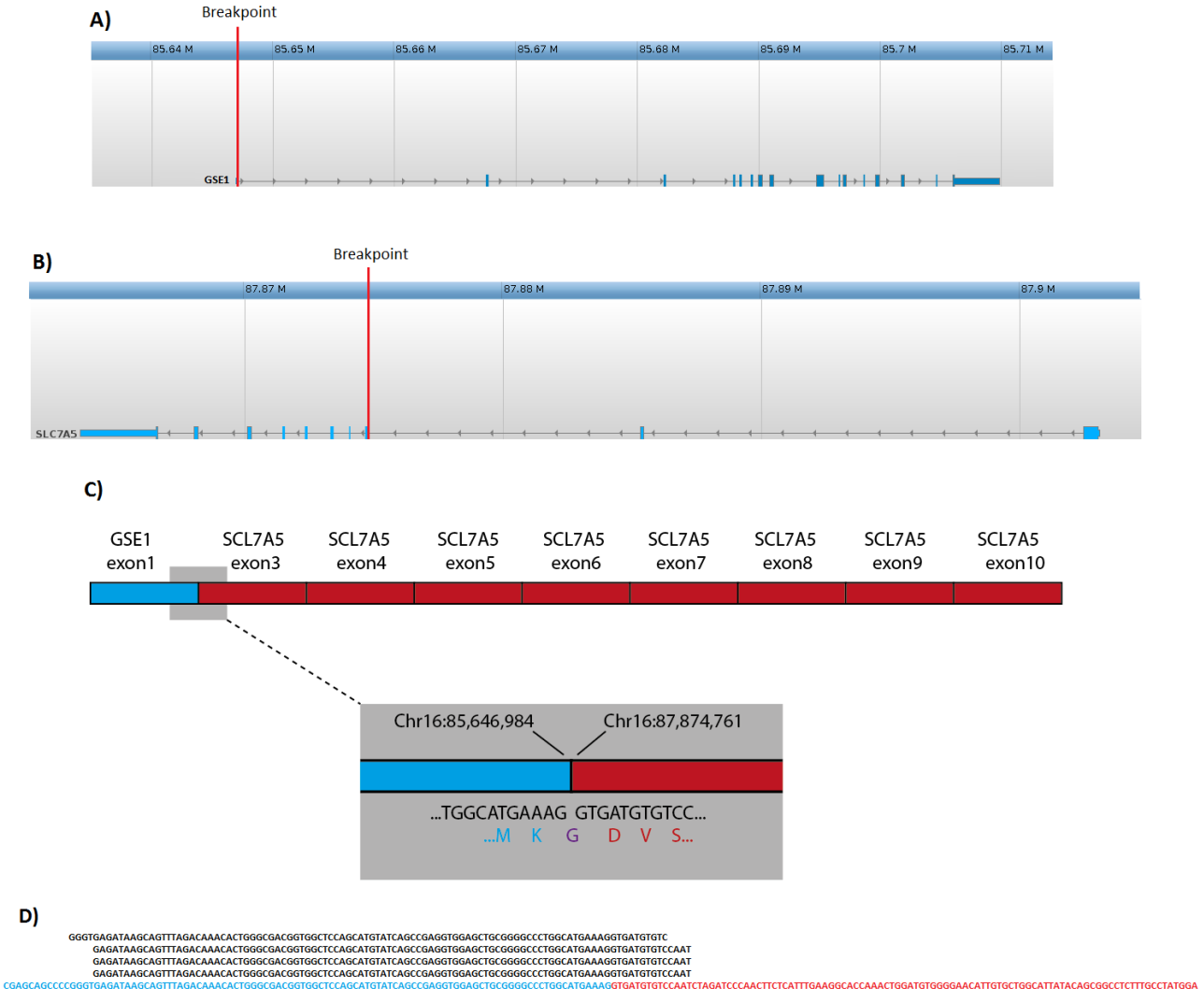

Figure 4.10: *GSE1-SCL7A5* gene fusion. Blocks represent exons, thin blocks represent Untranslated Region (UTR) parts of exons. A - transcriptome annotation in the region of *GSE1* on chromosome 16 with the breakpoint marked; B - transcriptome annotation in the region of *SCL7A5* on chromosome 16 with the breakpoint marked; C - putative fusion transcript with DNA and predicted amino acid sequence at the breakpoint; D - Read pileup around fusion site.

GSE1 (Gse1 Coiled-Coil Protein) is a protein coding gene of unknown function [150]. It was not reported to be involved in gene fusions.

SCL7A5 (Solute Carrier Family 7 (Amino Acid Transporter Light Chain, L System), Member 5) is a protein coding gene. Its product is involved in cellular transport [150]. *SCL7A5* has not been reported as a fusion partner.

In this case it is most likely that the gene fusion is of passenger type. The genes involved have functions unrelated to potential oncogenesis and the fusion construct is unlikely to be of gain-of-function type leading to oncogenesis.

TIPARP-KLHL24

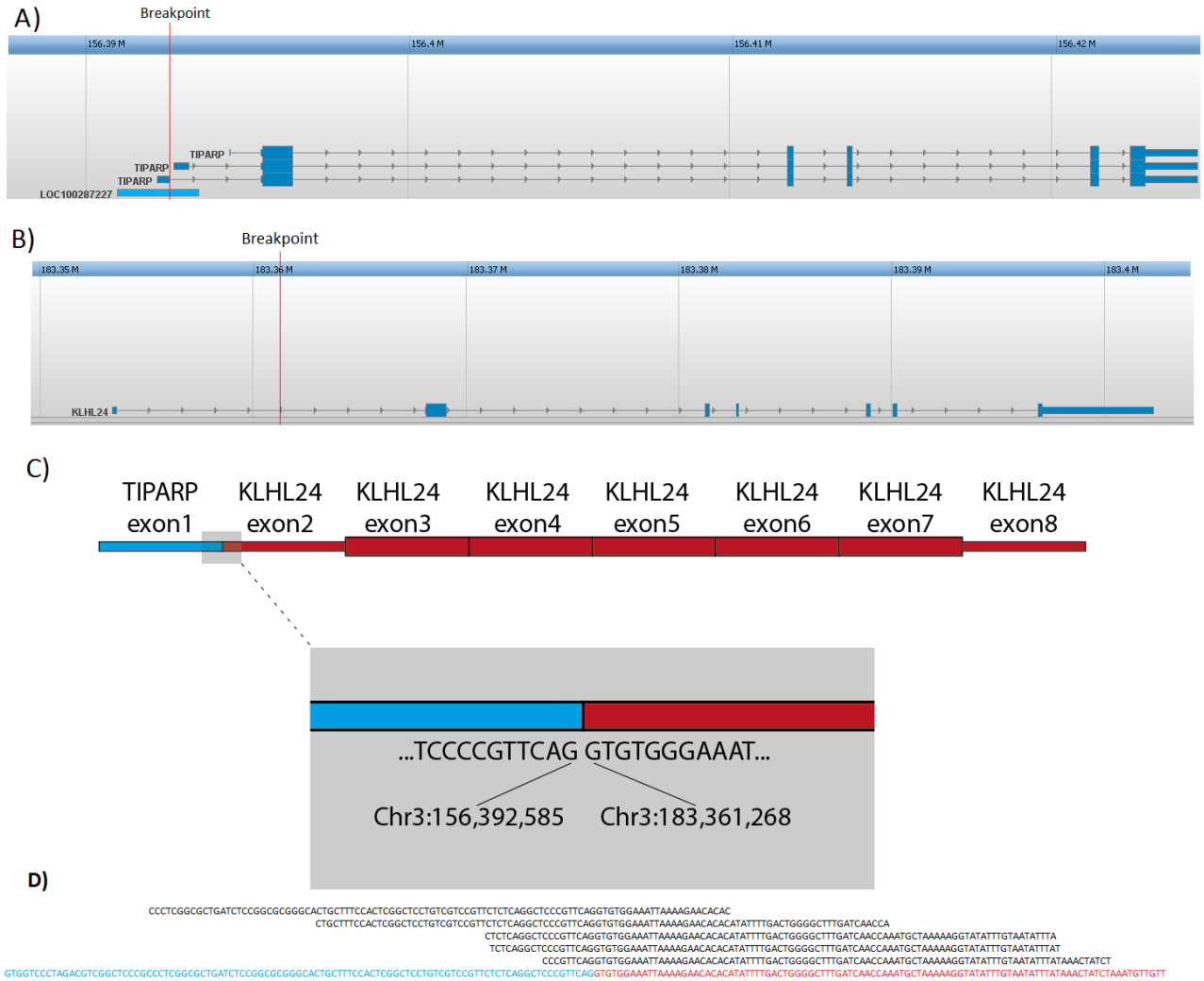

Figure 4.11: *TIPARP-KLHL24* gene fusion. Blocks represent exons, thin blocks represent Untranslated Region (UTR) parts of exons. A - transcriptome annotation in the region of *TIPARP* on chromosome 3 with the breakpoint marked; B - transcriptome annotation in the region of *KLHL24* on chromosome 3 with the breakpoint marked; C - putative fusion transcript with DNA sequence at the breakpoint; D - Read pileup around fusion site.

The next confirmed gene fusion in sample 7 was *TIPARP-KLHL24*. The fusion joins exon 1 of *TIPARP* and exon 2 of *KLHL24*, as presented in Figure 4.11. Since exon 1 of *TIPARP* consists only of 5'UTR, and exon 2 of *KLHL24* contains mostly 5' UTR, the putative fusion protein is expected to be identical to regular protein coded by *KLHL24*. Since both genes are found on the same chromosome, separated by approximately 27Mbp, it is possible that the fusion was created by a large deletion spanning the region or, more likely, a segmental

duplication inserting the genes in close proximity.

TIPARP (TCDD-Inducible Poly(ADP-Ribose)) is a gene coding for protein catalyzing histone poly(ADP-ribosyl)ation and may be involved in T-cell function [150]. A truncated transcript of the gene was previously observed in head-and-neck squamous cell carcinoma. There was one reported case of *TIPARP* being mutated in acute lymphocytic leukaemia (ALL), with c.1781G>A [179].

KLHL24 (Kelch-Like Family, Member 24) was found to reduce kainate receptor-mediated currents in hippocampal neurons, most probably by modulating channel properties [150]. There are no reported cases of the gene mutation in haematopoietic and lymphoid cancers

The importance of *TIPARP-KLHL24* in oncogenesis is suspected to be minimal, it is more likely that the fusion is of passenger type. None of the genes were indicated to be strongly related to oncogenesis. Especially when the effect of the fusion is considered, deregulation of *KLHL24* expression with no structural changes is unlikely to be the driver, giving rise to mastocytosis and eosinophilia symptoms.

RAB20-ING1

A gene fusion involving *RAB20* and *ING1* was identified and verified in the same sample. The fusion was supported by a total of 6 RNA-Seq reads, and obtained a maximum scoring pattern of 3/3. Putative gene fusion transcript contains exon 1 of *RAB20* and exon 4 of *ING1* (Figure 4.12). Since both *RAB20* and *ING1* are found on the same chromosome, on the opposite strands, separated by approximately 150kbp, it is suspected that the fusion arose due to an inversion.

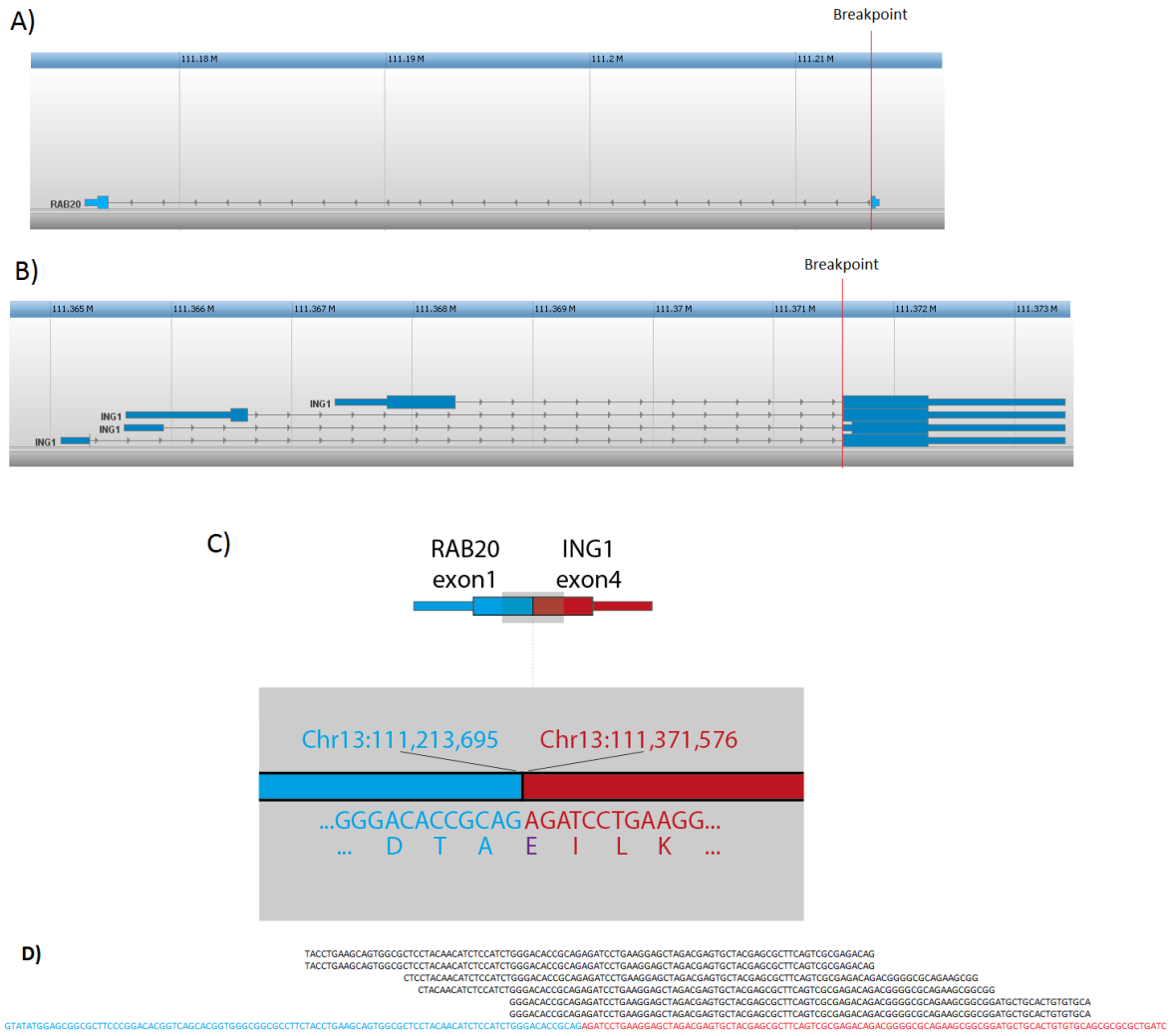

Figure 4.12: *RAB20-ING1* gene fusion. Blocks represent exons, thin blocks represent Untranslated Region (UTR) parts of exons. A - transcriptome annotation in the region of *RAB20* on chromosome 13 with the breakpoint marked; B - transcriptome annotation in the region of *ING1* on chromosome 13 with the breakpoint marked; C - putative fusion transcript with DNA and predicted amino acid sequence at the breakpoint; D - Read pileup around fusion site.

RAB20 (RAS oncogene family, Member 20) is a member of RAS oncogene family, and it is involved in endocytosis/recycling [150]. There are no previously reported cases of *RAB20* mutation in haematopoietic and lymphoid cancers. However, mutations in members of RAS oncogene family are frequently found in human tumours, and are known to disrupt RAS signalling pathways leading to neoplasm development [180].

ING1 (Inhibitor Of Growth Family, Member 1) belongs to a family of growth inhibitors, and encodes a protein that functions as a tumour and growth suppressor [150]. One case of

c.202C>T previously observed in ALL. ING family is known for its involvement in oncogenesis of solid cancers, though the mutations observed are usually of a loss-of-function type [181].

Clinical presentation of mastocytosis and eosinophilia is typically attributed to presence of *PDGFRA* or *PDGFRB* fusion or D816V *KIT* mutation [5]. Seeing how *ING1* is a tumour suppressor, it is possible that loss-of-function may contribute to loss of growth control. Another possibility is that the fusion is a gain-of-function type with *RAB20-ING1* putative protein serving as an oncogene, as RAS family members exhibit gain-of-function mutations in cancer. Finally, the possibility of *RAB20-ING1* being only a passenger gene fusion with little importance cannot be dismissed.

BCR-JAK2

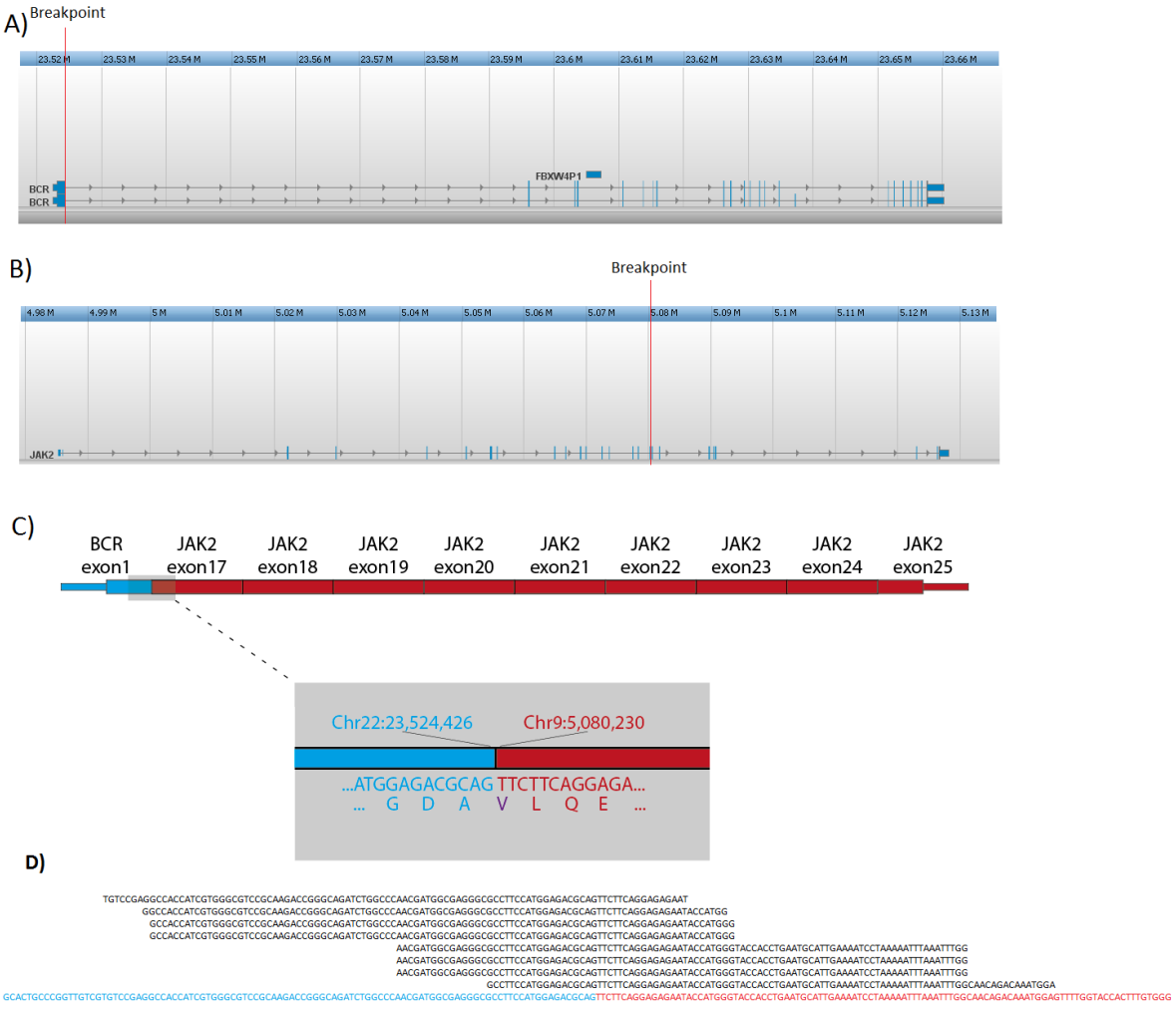

Figure 4.13: *BCR-JAK2* gene fusion. Blocks represent exons, thin blocks represent Untranslated Region (UTR) parts of exons. A - transcriptome annotation in the region of *BCR* on chromosome 22 with the breakpoint marked; B - transcriptome annotation in the region of *ING1* on chromosome 9 with the breakpoint marked; C - putative fusion transcript with DNA and predicted amino acid sequence at the breakpoint; D - Read pileup around fusion site.

In sample 8, taken from a patient with MPN, eosinophilia, and strong *in vitro* response to *JAK2* inhibitor, *BCR-JAK2* fusion was identified and validated. The abnormality was supported by a total of 8 RNA-Seq reads obtaining a maximum scoring pattern of 3/3. The putative product of the fusion contains exon 1 of *BCR* and exons 17-onwards of *JAK2*, as presented in Figure 4.13. Since *BCR* and *JAK2* are found on chromosome 22 and 9, respectively, the abnormality was most likely created as a result of chromosomal translocation. Interestingly, despite the fusion arising due to a chromosomal translocation, it was cytogenetically

masked as a t(9;18)(p24;q12).

BCR (Breakpoint Cluster Region) protein product has serine/threonine kinase activity and is a GTPase-activating protein [150]. The first exon of *BCR* is coding for a domain allowing tetramerization of the molecule [182]. *BCR* is involved in different gene fusions, most notably in *BCR-ABL*, abnormality characteristic of CML [11].

JAK2 (Janus Kinase 2 (Protein Tyrosine Kinase)) is a member of family of tyrosine kinases that associated with cytokine receptors [150]. Upon receptor activation JAKs phosphorylate the transcription factors known as STATs and initiate the JAK-STAT signaling pathway. Mutations in *JAK2* are inked to a range of myeloid malignancies, including polycythemia vera, and essential thrombocythaemia [183].

The *BCR-JAK2* gene fusion has previously been reported as causal in myeloid malignancies [184]. In the putative protein product of the abnormality, *BCR* part has the tetramerization domain conserved, whereas *JAK2* part has kinase domain conserved. It is therefore predicted that the tetramerization domain alters behaviour of *JAK2*, affecting JAK-STAT signalling pathway, effectively being the oncogenesis driver. The patient's *in vitro* reaction to a tyrosine kinase inhibitor confirms the finding. Further details can be found in Chapter 5.

4.3.5 Methods assessment

The validation of gene fusions by Sanger sequencing was proven to be crucial for confirmation purposes. Initially, 34 FCs were chosen and 21 of them were confirmed by PCR. Only 10 out of the 21 were validated by Sanger sequencing, underlining how important additional validation is. However, due to the nature of Sanger sequencing, and laboratory methods being inherently non-deterministic, validation can provide proof for existence of a gene fusion, it is unreliable as proof of absence. There were multiple fusions confirmed and validated per sample, with sample 7 harbouring five of them. The high number of gene fusions is concordant with some previous reports of up to 13 verified gene fusions in a myeloid malignancy case [185].

10 out of 34 FCs were validated which translates to 29% efficiency of the analysis methods. Considering that the supporting reads and pairs cut-off was set to be very low (SR+SP>3), there is less noise than previously expected from the analysis of the samples with previously known gene fusions described in Chapter 2, where the percentage of false positive calls was 92%. However, the current results, indicate that the actual number of true positive gene fusions may be higher than assumed in the analysis of samples with known gene fusions, as it indicates that there are more passenger fusions than expected, however the number of fusions is much higher than that anticipated by traditional analysis and particularly by cytogenetics. Although some of the genes that are involved have a connection with malignancy, as described above, *BCR-JAK2* is the only fusion that has been described previously and the only fusion to involve a tyrosine kinase. Deregulated tyrosine kinases are strongly associated with a myeloproliferative phenotype and overall it is questionable whether any of the other fusions that were identified and verified are of any pathogenetic relevance. It would certainly be of interest to perform fluorescent in situ hybridization analysis with specific probes to determine if any of the predicted cytogenetic rearrangements might be present and a high throughput alternative to this would be whole genome paired end sequencing. Individual fusions could be analysed for oncogenic activity using cell line or animal models, but this would be a significant amount of work. Probably the most important evidence for the relevance of any fusion is the finding that it is recurrent in patient samples and absent in healthy controls. Although it would have been possible to screen the fusions above in patient cohorts, this was not pursued because (i) there were no novel fusions that were considered to be very strong candidates and (ii) it would have meant using up significant amounts of precious stored patient material. Instead it was decided that the strategy should be to perform RNAseq on further cases in an attempt to try and identify recurrencies. Since patients with an overt myeloproliferative neoplasm and eosinophilia without a known abnormality are uncommon, this work will extend beyond the time available for study.

As all currently available software for gene fusion detection is based on the fusion junctions themselves, focusing on supporting reads and supporting read pairs, it is an indication that this is the level of noise expected when the determination methods rely only on reads directly indicative of gene fusions, which is further supported by a high number of false positive FCs

reported for other software [141].

In summary, the developed pipeline is effective at determining gene fusions in low read depth cases, however the identified fusions need to be verified by PCR and Sanger sequencing.

Chapter 5

Research - Limited duration of complete remission on ruxolitinib in myeloid neoplasms with *PCM1-JAK2* and *BCR-JAK2* fusion genes

Acknowledgements

This chapter has been previously published as:

J. Schwaab (i), M. Knut (i), C. Haferlach, G. Metzgeroth, H. P. Horny, A. Chase, W. J. Tapper, J. Score, K. Waghorn, N. Naumann, M. Jawhar, A. Fabarius, W. Hofmann, N. C. P. Cross, A. Reiter. Limited duration of complete remission on ruxolitinib in myeloid neoplasms with PCM1-JAK2 and BCR-JAK2 fusion genes. *Annals of Hematology*, 94(2):233-238, 2015.

(i) Equal contribution

Declaration of Authorship

The author of this thesis acknowledges that his own work on this chapter covers bioinformatic analysis of sample originating from Patient 2 and subsequent discovery of BCR-JAK2

gene fusion in said sample.

Abstract

Rearrangements of chromosome band 9p24 are known to be associated with *JAK2* fusion genes, e.g., t(8;9)(p22;p24) with a *PCM1-JAK2* and t(9;22)(p24;q11) with a *BCR-JAK2* fusion gene, respectively. In association with myeloid neoplasms, the clinical course is aggressive, and in absence of effective conventional treatment options, long term remission is usually only observed after allogeneic stem cell transplantation (ASCT). With the discovery of inhibitors of the JAK2 tyrosine kinase and based on encouraging *in vitro* and *in vivo* data, we treated two male patients with myeloid neoplasms and a *PCM1-JAK2* or a *BCR-JAK2* fusion gene, respectively, with the JAK1/JAK2 inhibitor ruxolitinib. After 12 months of treatment, both patients achieved a complete clinical, hematologic, and cytogenetic response. Nonhematologic toxicity was only grade 1 while no hematologic toxicity was observed. However, remission in both patients was only short-term, with relapse occurring after 18 and 24 months, respectively, making ASCT indispensable in both cases. This data highlight [186] the ongoing importance of cytogenetic analysis for the diagnostic work-up of myeloid neoplasms as it may guide targeted therapy and [187] remission under ruxolitinib may only be short-termed in *JAK2* fusion genes but it may be an important bridging therapy prior to ASCT.

5.1 Introduction

The JAK1/JAK2 inhibitor ruxolitinib was developed following the identification of constitutively dysregulated JAK/STAT signaling in myeloproliferative neoplasms (MPN) frequently caused by mutations in *JAK2* and *MPL* [186]. It was recently approved for the treatment of symptomatic splenomegaly or according to the individual risk status in patients with primary or secondary myelofibrosis independently of their JAK2 mutation status [187]. More than ten different JAK inhibitors are currently being tested in clinical trials. At least so far, these inhibitors all produce a marked and sustained reduction of the enlarged spleen and associated clinical symptoms in a significant proportion of patients while complete hematologic (e.g., normalization of blood counts), morphologic (e.g., disappearance of fibrosis), or molecular (e.g., negativity for *JAK2* V617F) remissions have not yet been reported [187].

JAK2 fusion genes are rare as compared to *JAK2* mutations. *JAK2* fusions are associated with diverse hematologic malignancies including acute and chronic leukemias of myeloid and lymphoid phenotypes. The most frequent subtype is a myeloid neoplasm, e.g., MPN or myelodysplastic/myeloproliferative neoplasm (MDS/MPN) in chronic or blast phase. To date, five different fusion partners have been identified with *PCM1*, *BCR*, and *ETV6* being the most important in myeloid neoplasms [188][189][190]. The prognosis is poor, and in most cases, long-term remissions have only been achieved after allogeneic stem cell transplantation (ASCT).

Following the rapid and sustained complete remissions on imatinib in patients with *ABL1* and *PDGFR* fusion genes, it was suggested that JAK inhibitors may be capable of inducing similar remissions in patients with *JAK2* fusions [191][192]. We recently reported encouraging in vitro data to support the use of ruxolitinib for such patients: Ba/F3 cells transformed to IL3 independence by *ETV6-JAK2* showed reduced proliferation and survival compared with cells transformed with *ABL1*, *FLT3*, or *FGFR1* fusion genes that were associated with reduced phosphorylation of ETV6-JAK2, ERK, and STAT5. The growth of primary cells from two patients with *JAK2* rearrangements showed reduced colony growth in culture treated with ruxolitinib compared with healthy controls, and fluorescence in situ hybridization (FISH) demonstrated a reduction in *JAK2* rearranged colonies in ruxolitinib-treated cultures [193]. Meanwhile, the first two *PCM1-JAK2*-positive patients who achieved a complete hematologic and cytogenetic response on ruxolitinib were reported [194][195]. Here, we report on rapid response but also early relapse of two patients with disparate *JAK2* fusion genes during treatment with ruxolitinib.

5.2 Patients and Methods

The clinical characteristics of both patients were recently reported and are summarized in Table 5.1.

Bone marrow histology revealed a hypercellular marrow with signs of dysplasia in both cases. In both patients, the erythroid lineage was marked, and in the *PCM1-JAK2*-positive patient, giant erythrons were present (Figure 5.1).

	Patient 1: *PCM1-JAK2*			Patient 2: *BCR-JAK2*		
Treatment time (months)	0	12	24	0	12	18
Hemoglobin (g/dL)	12.3	14.9	14.3	11.2	14.3	9.1
Leukocytes (x10e9/L)	57.0	3.3	3.6	46.3	4.7	4.7
Platelets (x10e9/L)	60	200	153	1,458	272	452
Myeloid precursors (%)	10	0	9	22	0	6
Eosinophils (x10e9/L)	570	60	324	3,430	30	46
Lactate Dehydrogenase (<245 U/L)	739	245	340	838	277	316
Aberrant metaphases	[17/20]	[0/20]	[6/25]	[24/24]	[0/21]	[9/25]

Table 5.1: Patient Characteristics.

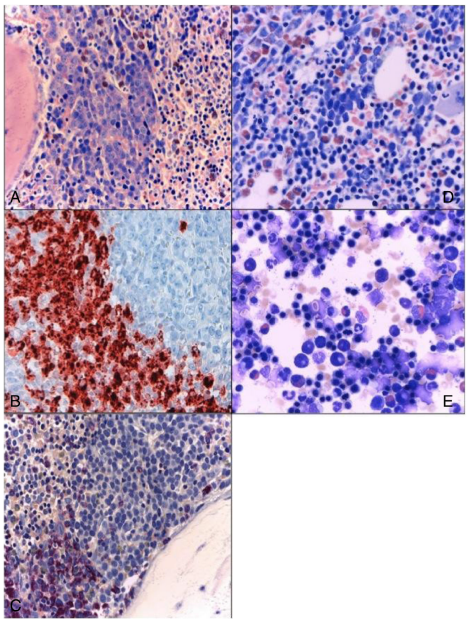

Figure 5.1: Morphologic features of two patients with *PCM1-JAK2* and *BCR-JAK2* fusion gene. **A-C (*PCM1-JAK2* patient); A,** Giemsa stain: Extremely hypercellular bone marrow (BM) with markedly altered microarchitecture and atypically localized peritrabecular giant erythron. Note the absence of megakaryocytes and a slight diffuse increase in eosinophils. **B,** ABC stain: marked increase in basophilic granulocytes forming large clusters (a finding not seen in other myeloid neoplasms with the exception of rare basophilic leukemias). **C,** Naphthol AS-D chloroacelate esterase stain: marked alteration of BM microarchitecture with an (unstained) giant erythron in an unusual peritrabecular localisation and also abnormally localized clusters of immature neutrophilic cells (red). Note the absence of megakaryocytes. **D-E (*BCR-JAK2* patient); D,** Giemsa stain: Hypercellular BM with dominating erythropoiesis but without giant erythrons. **E,** BM smear with marked erythropoiesis.

Cytogenetic analyses and reverse transcription polymerase chain reaction (RT-PCR) following RNA extraction and cDNA synthesis were performed according to standard protocols. Primers are shown in the supplement.

Transcriptome sequencing (RNA sequencing)

The transcriptome of patient #2 was sequenced using an Illumina HiSeq 2000 at the Wellcome Trust Centre for Human Genetics at Oxford, UK. Bowtie [196] was used to align read sequences to the human genome (version hg19) and transcriptome (version UCSC hg19). TopHat [197] was used to resolve splice junctions, and potential fusions were identified and filtered by TopHat-Fusion [198]. Considering the possibility of low coverage as the result of the sample origin, default settings were altered to only allow potential fusions with at least four supporting reads and/or pairs (–num-fusion-reads 0 –num-fusion-pairs 0 –num-fusion-both 4), obtaining 26 potential fusions. Out of these, one was situated in a region of interest — potential *BCR-JAK2* fusion between BCR exon 1 and *JAK2* exon 17, supported by limited evidence of four spanning mate pairs and was selected for validation by RT-PCR.

Long template PCR

For amplification of the genomic *BCR-JAK2* breakpoint, long template PCR (LT-PCR) was used. All primers were designed using Primer3 (http://frodo.wi.mit.edu/primer3/), checked to be free of single-nucleotide polymorphisms (http://genome.ucsc.edu/) and tested on normal control DNA with a suitable primer (see Table C.3 for primer sequences). To amplify breakpoint bands up to 12 kb in size, the Expand LTPCR system 2 (Roche, Burgess Hill, UK) was used with an annealing temperature of 64 °C and an elongation time of 8 min. The BCR 1AF gave a patient-specific band with both JAK2 Int 16R JAK2 Exon 17R primers, which were then Sanger sequenced to identify the precise location of the breakpoint (Figure 5.2).

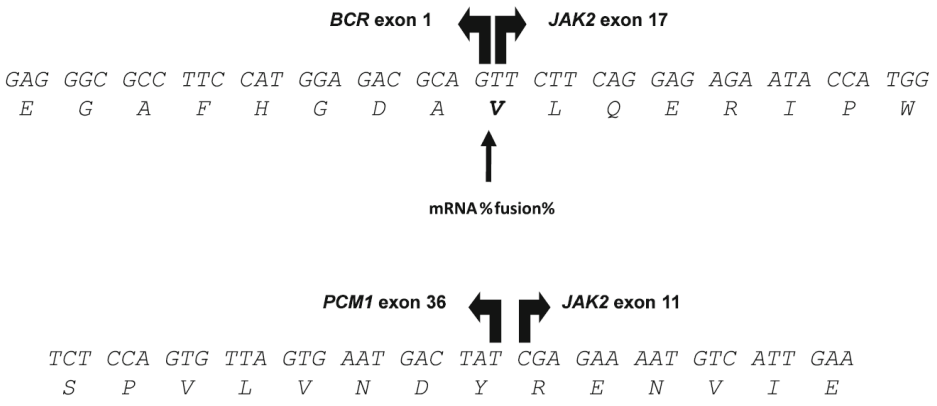

Figure 5.2: Schematic description of the genomic breakpoints in the two patients. Sequences and amino acids of the respective fusion genes *BCR-JAK2* and *PCM1-JAK2*. In the first case, a new valine (V) residue has been created at the fusion junction.

Mutation analysis

At the time of cytogenetic relapse, mutational analysis was performed sequencing the region coding for the *JAK2* kinase domain by conventional Sanger sequencing. For the *BCR-JAK2* case, the fusion was amplified by heminested PCR using primers BCR1F plus JAK25-R in the first round and BCR-C plus JAK25-R in the second round. The product was sequenced with BCR-C and JAK25-R. For the *PCM1-JAK2* case, the fusion was amplified by heminested PCR using primers PCM34-F and JAK25-R1 (first step) and PCM35-F and JAK25-2R (second step). Sequencing was performed using primers from the second step (Table C.3).

Chromosome banding analysis and FISH analysis

Chromosome banding analyses and FISH analyses were performed according to standard procedures. For the detection of *JAK2* rearrangements, a break-apart probe was used (Kreatech, Amsterdam, The Netherlands). In addition, metaphase FISH with whole chromosome painting probes for chromosome 9, 18, and 22 as well as loci-specific probes for *BCR* and *ABL* were performed.

5.3 Results

The morphologic phenotype of patient #1 in association with a t(8;9) (p22;p24) suggested the presence of a *PCM1-JAK2* fusion gene. Indeed, amplification by RT-PCR and sequence analysis revealed a fusion between *PCM1* exon 36 with *JAK2* exon 11. The molecular char-

acterization of the t(9;18) in patient #2 was more complicated. The t(9;18)(p24;q12) seen in this case has not been reported previously and did not suggest the formation of any known fusion gene. The break at 9p24, however, suggested the involvement of *JAK2* which was confirmed by split apart FISH. RACE-PCR from *JAK2* was uninformative but RNA sequencing indicated the presence of a *BCR-JAK2* fusion gene with joining of *BCR* exon 1 to *JAK2* exon 17. This was unexpected since karyotype analysis had not indicated the involvement of chromosome 22. Further FISH analysis with whole chromosome painting probes for chromosomes 9 and 18 in addition to a probe flanking BCR did not confirm a rearrangement of this gene. To resolve this discrepancy, we employed LT-PCR on genomic DNA using multiple primers, which confirmed a fusion between *BCR* intron 1 and *JAK2* intron 16. The *BCR-JAK2* fusion was subsequently confirmed by RT-PCR and PCR from genomic DNA. *BCR-JAK2* in this case is therefore likely to be the result of a small insertion of *BCR* into the *JAK2* locus on the der(18).

After final approval and patients consent for participation in an individual supply program, both patients were treated with ruxolitinib. Dose was adapted to platelet counts (15 mg BID for platelets between 100×109/L and 200×109/L, 20 mg BID for platelets >200×109/L). Markers for clinical response were the following: lactate dehydrogenase (LDH) and leukocyte count in addition to reduced (patient #1) or elevated platelet counts (patient #2), respectively. No grade II– IV hematologic or non-hematologic toxicities were observed. Patient #1 achieved a complete normalization of leukocytes, platelets, and LDH after 12 months, with 20/20 metaphases having a normal karyotype, indicating complete cytogenetic response (CCR). The complete hematologic response (CHR) occurred earlier in patient #2 (6 months); CCR in 14/14 metaphases was also observed for the first time after 12 months on ruxolitinib. Both patients remained positive for their individual fusion genes by RT-PCR analysis. Mild reduction of platelets and elevated LDH in combination with cytogenetic relapse occurred in patient #1 after 24 months on ruxolitinib (Table 5.1). Relapse in patient #2 after 18 months on ruxolitinib became obvious through elevated platelets and decrease of hemoglobin level in addition to cytogenetic relapse (Table 5.1). Sequencing analysis of the *JAK2* kinase domain did not reveal any secondary mutations in either patient.

5.4 Discussion

In chronic myeloid neoplasms, cytogenetic analysis is of utmost importance for identification and targeted treatment of underlying pathogenetic tyrosine kinase (TK) fusion genes. With the exception of the cytogenetically invisible *FIP1L1-PDGFRA* fusion, all out of the 50+ currently known TK fusion genes have been identified through cytogenetically visible rearrangements, most commonly involving chromosome bands 4q12 (*PDGFRA*), 5q31-33 (*PDGFRB*), 8p11-12 (*FGFR1*), or 9p24 (*JAK2*) [191][199]. These rearrangements are typically balanced reciprocal translocations, but insertions or more complex translocations are seen occasionally. *PDGFRA* and *PDGFRB* fusion genes are exquisitely sensitive to treatment with imatinib. Patients achieve rapid and sustained complete hematologic, cytogenetic, and molecular remissions. Primary resistance is very uncommon and secondary resistance, predominantly through point mutations, is rare. The 5- year probabilities for progression-free and overall survival are over 90 % independently whether treatment was initiated in chronic or blast phase [200][201].

Until recently, the prognosis and therapeutic options for patients with *FGFR1* and *JAK2* fusion genes were rather different. Both fusion genes are seen in cases with an aggressive clinical course with rapid progression to blast phase, usually within the first 2 years after diagnosis. In the majority of patients, long-term survival frequently has only been achieved after ASCT. Until recently, no clinically effective inhibitors for these two TKs were available but in vitro and murine studies demonstrated the potential of treatment with ruxolitinib for *JAK2* fusions and ponatinib for *FGFR1* fusions. We here report the third *PCM1-JAK2*-positive patient and the first *BCR-JAK2*-positive patient to be treated with ruxolitinib. Initial responses were very good for both cases. However, relapse occurred after 18 and 24 months, respectively, making ASCT indispensable in both patients. The two previously described *PCM1-JAK2*-positive patients were also in complete response approximately 1 year after start of ruxolitinib. Our data however suggest that the efficacy of ruxolitinib in the treatment of *JAK2*-driven fusion genes is valuable but potentially only limited. Because of the dismal prognosis, ruxolitinib may therefore be most useful as a bridging therapy before ASCT in eligible patients.

On the basis of this recent development, concerns need to be raised regarding the current WHO classification and the subcategory of 'myeloid/lymphoid neoplasms with eosinophilia and rearrangement of *PDGFRA*, *PDGFRB*, and *FGFR1'*. In particular, (i) eosinophilia is not present in all patients with respective fusion genes, e.g., *FOP-FGFR1* in t(6;8)(q27;p11) is often associated with a phenotype resembling polycythemia vera while *PRKG2-PDGFRB* in t(4;5)(q21;q33) was associated with basophilia; (ii) the involvement of the myeloid and lymphoid lineages is featuring a single stem cell disorder with disparate morphologic phenotypes in bone marrow (MPN) and lymph node (lymphatic blast phase); and (iii) rearrangements of *JAK2* and other kinases such as *RET*, *FLT3*, etc., associated with eosinophilia in the majority but not all cases, should be included in this subcategory.

Chapter 6

Research - Identification of U2AF(35)-dependent exons by RNA-Seq - a link between 3' splice-site organization and activity of U2AF-related proteins

Acknowledgements

This chapter has been previously published as:

J. Kralovicova (i), M. Knut (i), N. C. P. Cross, I. Vorechovsky. Identification of U2AF(35)-dependent exons by RNA-Seq reveals a link between 3' splice-site organization and activity of U2AF-related proteins. *Nucleic Acids Research*, 43(7):3747-3763, 2015.

(i) Equal contribution

Declaration of Authorship

The author of this thesis acknowledges that his own work on this chapter covers bioinformatic analysis of the RNA-Seq data along with statistical analysis done under supervision of I. Vorechovsky.

Abstract

The auxiliary factor of U2 small nuclear RNA (U2AF) is a heterodimer consisting of 65- and 35-kD proteins that bind the polypyrimidine tract (PPT) and AG dinucleotides at the 3' splice site (3'ss). The gene encoding U2AF35 (*U2AF1*) is alternatively spliced, giving rise to two isoforms U2AF35a and U2AF35b. Here, a knock down U2AF35 and each isoform and characterized transcriptomes of HEK293 cells with varying U2AF35/U2AF65 and U2AF35a/b ratios using 2x2 experimental design was performed. Depletion of both isoforms preferentially modified alternative RNA processing events without widespread failure to recognize 3'ss or constitutive exons. Over a third of differentially used exons were terminal, resulting largely from the use of known alternative polyadenylation (APA) sites. Intronic APA sites activated in depleted cultures were mostly proximal whereas tandem 3'UTR APA was biased toward distal sites. Exons upregulated in depleted cells were preceded by longer AG exclusion zones and polypyrimidine tracts (PPTs) than downregulated or control exons and were largely activated by PUF60 and repressed by CAPERα. The U2AF(35) repression and activation was associated with a significant interchange in the average probabilities to form single-stranded RNA in the optimal PPT and branch site locations and sequences further upstream. Although most differentially used exons were responsive to both U2AF subunits and their inclusion correlated with U2AF levels, a small number of transcripts exhibited distinct responses to U2AF35a and U2AF35b, supporting the existence of isoform-specific interactions. These results provide new insights into function of U2AF and U2AF35 in alternative RNA processing.

6.1 Introduction

Eukaryotic genes contain intervening sequences or introns that are removed from mRNA precursors by a large and highly dynamic RNA–protein complex, termed the spliceosome [89]. The spliceosome consists of small nuclear ribonucleoproteins (snRNPs), including the U1, U2, U4/U5/U6 of the major U2 spliceosome and the U11, U12, U4atac / U6atac/U5 of the less abundant U12 spliceosome, and a large number of non-snRNP proteins [89]. Spliceosomes assemble on each intron in an ordered manner, starting with recognition of the 5' splice site (5'ss) by U1 snRNP or the 3'ss by the U2 pathway [90],which involves binding of the U2 auxiliary factor (U2AF) to the 3'ss region to facilitate U2 snRNP recruitment to the branch

point (BP) [202][91]. U2AF is a stable heterodimer composed of a *U2AF2*-encoded 65-kD subunit (U2AF65), which contacts the polypyrimidine tract (PPT), and a *U2AF1*-encoded 35-kD subunit (U2AF35). U2AF35 interacts with almost invariant AG dinucleotides at 3'ss and stabilizes U2AF65 binding to RNA [203][204][205][92].

The *U2AF1* gene is alternatively spliced, giving rise to conserved mRNA isoforms termed U2AF1a, U2AF1b and U2AF1c [206]. In mice, *U2AF1a* is more abundant than *U2AF1b* and contains a highly conserved 67-bp exon 3 in the mRNA whereas *U2AF1b* incorporates exon Ab of the same size [206]. The *U2AF1c* isoform includes both exons that introduce a premature termination codon (PTC) in the mRNA, which is targeted by RNA surveillance [206]. Both U2AF35a and U2AF35b contain a central U2AF65 recognition domain of the UHM (U2AF homology motif) type flanked by two C3H-type zinc finger (ZF) domains and a C-terminal arginine/serine-rich (RS) region [207][208][209][210]. Both U2AF35a and U2AF35b interact with U2AF65 and could stimulate its binding to a PPT [206]. During evolution, the two U2AF35 proteins have been under high selection pressure, suggesting that they may play specific functions in vertebrates and plants [206][211][212], but their putative functional differences in 3'ss selection are not known.

Although U2AF35 is essential for viability of both yeast and higher eukaryotes [213][214][215][216], the 3'ss AG was found to be dispensable for the *in vitro* splicing of pre-mRNAs with strong PPTs [217]. Binding of U2AF65 and U2AF35 to weak 3'ss was promoted by splicing activators under splicing conditions [218][219], however, several splicing events assumed to depend critically on U2AF35 did not show any defect under conditions of limited U2AF35 availability *in vivo* [211][220] and some alternative 3'ss responsive to U2AF35 depletion were intrinsically stronger than their nonresponsive counterparts [221]. Thus, the distinction between U2AF35-dependent and -independent introns has remained obscure. In addition, overexpression of U2AF65 and depletion of U2AF35 resulted in activation of the same cryptic 3'ss, suggesting that their balance is important for 3'ss selection [221]. However, global RNA processing changes in response to U2AF35 depletion or varying ratios of U2AF35/U2AF65 have not been examined.

In this study, the transcriptome of human embryonic kidney (HEK) 293 cells lacking U2AF35 and each U2AF35 isoform is characterized.

6.2 Materials and Methods

Cell cultures, transfections and library preparations

HEK293 cells were grown under standard conditions in DMEM supplemented with 10% (v/v) bovine calf serum (Life Technologies). For depletion Experiments (FigureB.1), the cells were treated with small interfering RNAs (siRNAs) or splice-switching oligonucleotides (SSOs) targeting splice sites of mutually exclusive *U2AF1* exons 24 h after seeding. Transfections were carried out in 6- or 12-well plates using jetPRIME (Polyplus) according to manufacturer's recommendations. The cells were harvested after 24 and 48 h or received the second hit after 48 h when splitting the cells into new plates. The remaining cultures were harvested 24 and 48 h after the second hit. For RNA sequencing (RNA-Seq), total RNA was extracted using RNeasy Plus (Qiagen) from cells harvested 72 h after the first hit. The NEBNext poly(A) mRNA magnetic isolation module (E7490L) and the Human/mouse/rat Ribo- ZeroTM rRNA Removal Kit (Cambio/Epicentre) for RNA were employed according to manufacturers' recommendations. The libraries were prepared using theNEBNext Ultra DNA Library Prep Kit for Illumina (E7370L), sizeselected and multiplexed before paired-end sequencing on the HiSeq 2500 Ultra-High-Throughput Sequencing System (Illumina).

siRNAs to downregulate the remaining proteins were as previously described [221]; the siRNA duplex to heterogeneous nuclear ribonucleoprotein C (hn-RNP C) was reported earlier [222].

Detection of Spliced products

Total RNA was transcribed using the Moloney murine leukemia virus reverse transcriptase (Promega) and oligo(dT) primers according to the manufacturer's recommendations. Alternatively spliced U2AF35 exons were visualized by complete *HinfI* digests of amplified polymerase chain reaction (PCR) products, as described previously [206]. Signal intensities of amplified fragments were measured as described [221].

Immunoblotting

Western blot analyses were carried out as described [221] using antibodies against U2AF35 (10334–1-AP, Protein Tech Group), U2AF65 (U4758, Sigma), actin (ab37063, Abcam), tubulin (ab56676, Abcam) and CAPERα (PA5– 31103, Thermo Fisher Scientific). AntiXpress antibodies were purchased from LifeTechnologies (R910–25). Antibodies against PUF60, hnRNP C and hnRNP I (PTB) were a generous gift of Adrian Krainer, Gideon Dreyfuss and Christopher Smith, respectively.

RNA-Seq analysis

Apart from the knockdown of U2AF35 and its isoforms (Figure B.1), analysis of RNA-Seq data of previously published knockdown experiments with 11 proteins was performed: heterogenous nuclear ribonucleoprotein (hnRNP) C (Illumina HiSeq 2000) [222], hnRNP A1, hnRNP A2B1, hn- RNPH1, hnRNP F, hnRNP M, hnRNPU(Illumina GAII) [223],HOXA1(Illumina HiSeq 2000) [224], AFF2,AFF3 and AFF4 (Illumina GAIIx) [225]. The raw FASTQ data were aligned against the human genome and transcriptome reference using TopHat (v. 2.0.9) [226] and Bowtie (v. 2.1.0) [227] using default stringencies and parameters, except for modification of the UCSC reference (hg19) [228] by introducing the *U2AF1* isoform s, which lacks exons Ab and 3. Sequences recognized as originating from mtRNA, rRNAand tRNA were subsequently removed. Analysis of differential exon usage was performed using DEXSeq (1.12.1) (32) and MISO (mixture-of-isoforms) [223]. DEXSeq-detected exons were selected based on statistical significance of differential usage ($q < 0.05$). Unlike DEXSeq, MISO (v. 0.4.9; hg19) [223] computes percentage of spliced in (psi, Ψ) splice junction-spanning reads and examines their significance using Bayesian factors. The filtering cut-offs were set to default parameters, on the basis of Ψ difference and event significance ($\Delta\Psi > 0.2$ and $K > 10$). Statistically significant events were individually verified in genome browsers to exclude false positives as a result of misannotated transcripts, low expression, overlapping transcripts and apparent misclassifications.

Differential gene and isoform expression between sample sets was analyzed with Cufflinks (v. 2.1.1) [229], which normalizes the reads using a fragments per kilobase of exon model per million reads (FPKM) measure. Gene and isoform expression assessment was aided by the transcriptome reference (hg19, UCSC) with no novel transcript discovery and was followed by

CummeRbund (v. 2.2.0) [229] analysis of differential gene and isoform expression (in R environment; v. 3.0.2). Selection of significantly differentially expressed genes was made on the basis of FDR-adjusted P-values ($q < 0.05$). RNA-Seq data for U2AF35 depletion experiments are available at ArrayExpress under the accession number E-MTAB-2682. Finally, gene- and exon-level functional enrichment analyses of differentially expressed events were performed using DAVID [230][231].

Validation of U2Af35-dependent events

RNA was extracted using TRI reagent, treated with DNase I (LifeTechnologies) and reverse-transcribed as described above. Target transcripts were chosen based on P- and FPKM-values that are shown in full in Tables C.5, C.6, C.7, and Supplementary Materials in [296]. PCR primers (Table C.9) were designed to amplify two or more isoforms with different sizes. Exogeneous transcripts were amplified using RT-PCR with vector primers PL3 and PL4 [232] or their combinations with transcript-specific primers.

Plasmid constructs

Splicing reporter minigenes were cloned into pCR3.1 (Invitrogen) using primers shown in Table C.9. IgM minigenes were a generous gift of Martha Peterson, University of Kentucky. Plasmid DNA was extracted using Wizard R Plus SV Minipreps (Promega). Expression constructs of U2AF35 isoforms were described previously [221]. PUF60 was subcloned into pCI-neo (Promega) with the Xpress tag at the N terminus, employing the pET28a- PUF60-His construct [221] as a template. All constructs were sequenced prior to transfections to exclude undesired mutations.

Sequence features of U2AF(35)-dependent exons and introns

Sequences individually validated in genome browsers were examined using algorithms that predict both traditional and auxiliary splice-site recognition motifs. Intrinsic splice site strength was computed using both maximum entropy and frequency matrix scores [233][234][235][236]. Prediction of BPs, PPTs and AG exclusion zones (AGEZs) was carried out using a support vector machine (SVM) algorithm [237]. De novo motif discovery, motif enrichment and motif location analyses were performed using the MEME suite of programmes [238] with sequences flanking differentially used 3'ss, 5'ss and internal exons as the input.

Measurements of single-strandedness across 3'ss

Computation of PU (probability of unpaired) values for all substrings of high-confidence upregulated and downregulated sequences in their natural context, extending input sequences by 30 nt in each direction was done. The PU values employ the equilibrium partition function of RNAfold [239] and were as defined previously [240]. Input sequences were fixed relative to the position of upregulated and downregulated 3'ss. Their PU values were averaged for each intron position. The means were compared by the Wilcoxon–Mann–Whitney test. Delta PU values are defined here as the difference between mean PU values of upregulated and downregulated exons at the indicated positions.

6.3 Results

Identification of U2AF(35)-dependent exons

Knock down of U2AF35 and each alternatively spliced U2AF35 isoform in HEK293 cells (Figure 6.1 and 6.2) along with examination of genome-wide exon/transcript usage by RNASeq was done.

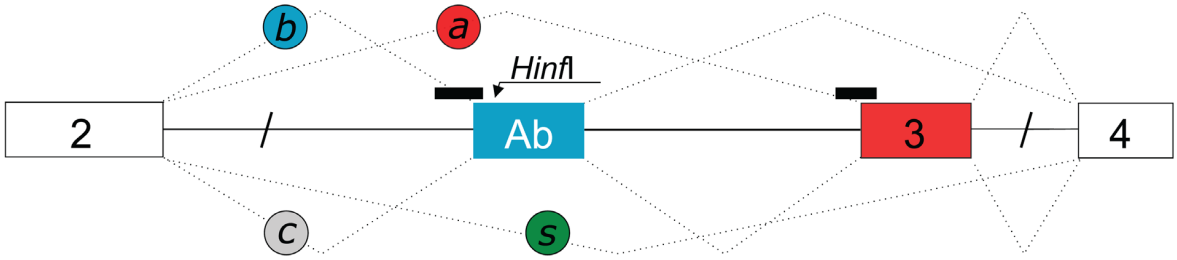

Figure 6.1: Alternative splicing of U2AF1 and location of Splice Switching Oligonucleotides (SSOs). Exons are shown as boxes, introns as horizontal lines, SSOs as black rectangles across 3'ss and spliced products (*a, b, c, d*)as dotted lines.

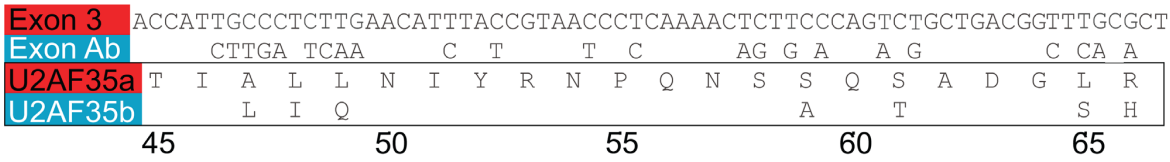

Figure 6.2: Nucleotide (upper panel) and amino acid (lower panel) sequences of alternatively spliced *U2AF1* exons. Amino acids are numbered at the bottom.

~90% depletion with siRNAs targeting both isoforms and a reversal in the relative abundance of U2AF35a and U2AF35b using isoform-specific siRNAs was achieved, with a maxi-

mum at 72 and 96 h post-transfection (Figure 6.3 and Figure B.1).

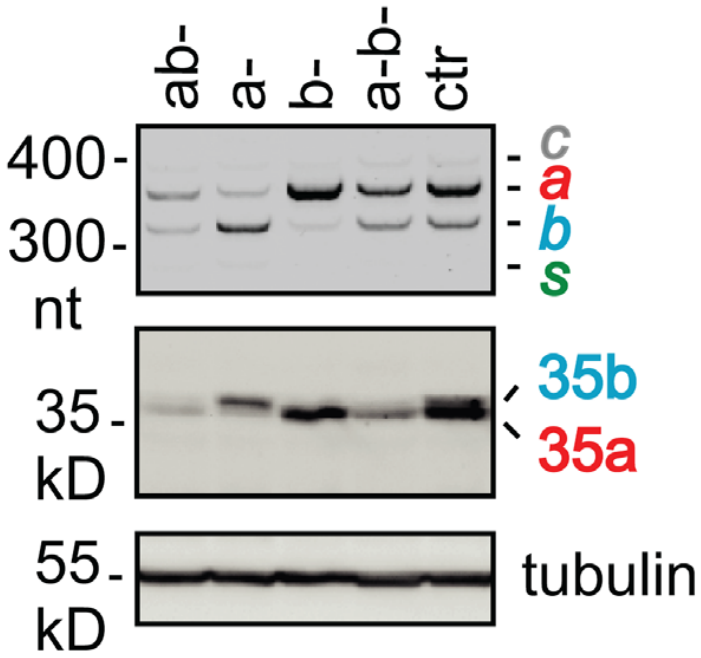

Figure 6.3: *HinfI* digested Reverse-Transcription Polymerase Chain Reaction (RT-PCR) products showing the relative abundance of *U2AF1* isoforms in depleted samples (upper panel) and immunoblot with antibodies against U2AF35 and tubulin (lower panels). ab-, depletion of both isoforms using siRNA U2AF35ab (30nM); a-, depletion of U2AF35a (60nM); b-, depletion of U2AF35b (60nM); a-b-, depletion of U2AF35 using equimolar mixtures of isoform-specific short interfering ribnonucleic acids (siRNAs); siRNAs were as described [211][220]. Ctrl, a scrambled control.

In addition to siRNAs, SSOs targeting 3' splice sites of alternatively spliced *U2AF1* exons were employed, resulting in a less robust and more delayed response in knockdown levels (Figure B.1). Using the Illumina HiSeq 2500 platform, a total of 546 390 339 reads mapped to the annotated human transcriptome and genome was obtained, averaging to 68M per RNA sample (Table C.4). The fraction of *U2AF1* mRNAs ranged between 0.001% for depleted samples and 0.01% for controls, with isoformspecific knockdowns giving intermediate levels. The range of *U2AF1*/*U2AF2* ratios varied between 0.1 and 2.3 (Figure 6.4 and 6.5).

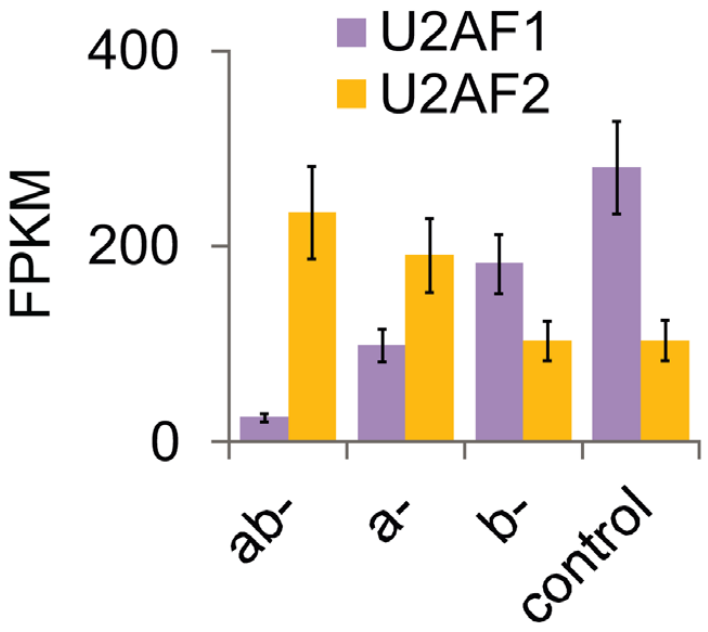

Figure 6.4: Normalized expression of *U2AF1* and *U2AF2* genes in depleted cultures and controls. FPKM, fragments per kilobase of exon model per million reads. Error bars are 95% confidence intervals. *U2AF1/U2AF2* ratios in ab-, a-, b- and control cultures were 0.09, 0.43, 1.47, and 2.25, respectively.

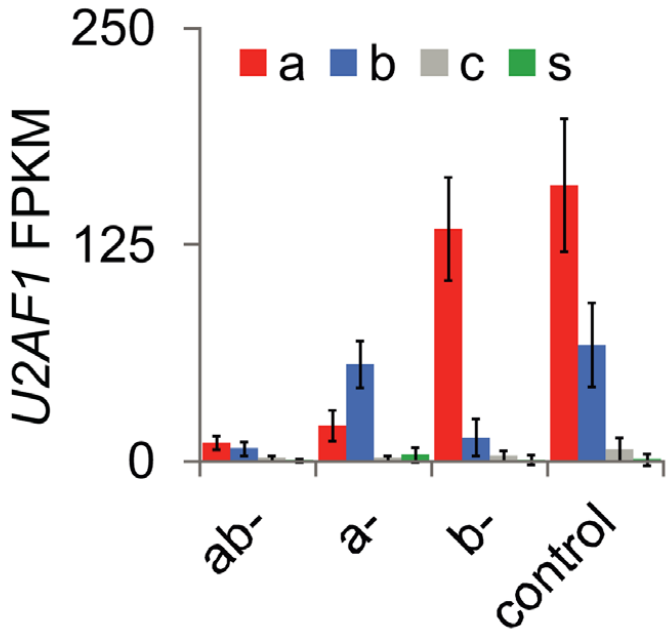

Figure 6.5: Normalized expression of *U2AF1* isoforms. FPKM, fragments per kilobase of exon model per million reads. Error bars are 95% confidence intervals.

Surprisingly, the *U2AF1* depletion was associated with a 2-fold increase of each alternatively spliced *U2AF2* isoform (Figure 6.6).

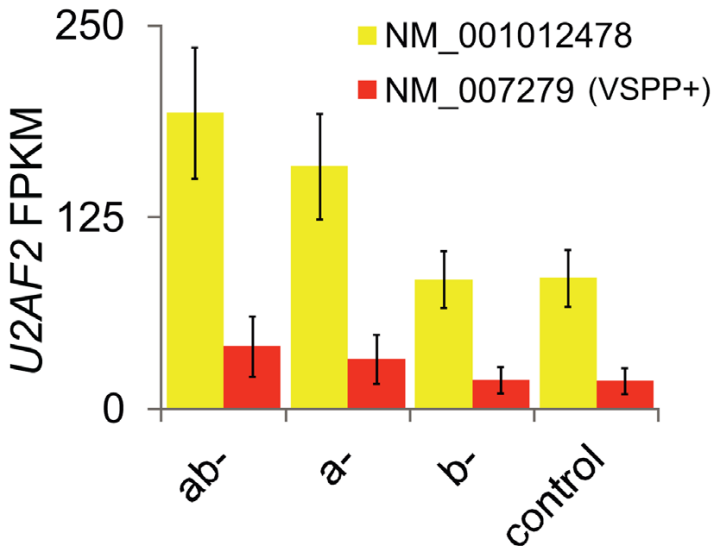

Figure 6.6: Normalized expression of *U2AF2* isoforms. Extra amino acids included in U2AF65 as a result of alternative GC 5'ss usage are in parentheses. FPKM, fragments per kilobase of exon model per million reads. Error bars are 95% confidence intervals.

The *U2AF1a*-specific depletion led to a lower overall *U2AF1* expression than the *U2AF1b* knockdown while expression of isoforms recognized by RNA surveillance remained low (Figure 6.5). The relative abundance of *U2AF1b* slightly increased in cells treated with siRNAs targeting both isoforms compared to untreated cells (Figure 6.5 and 6.7).

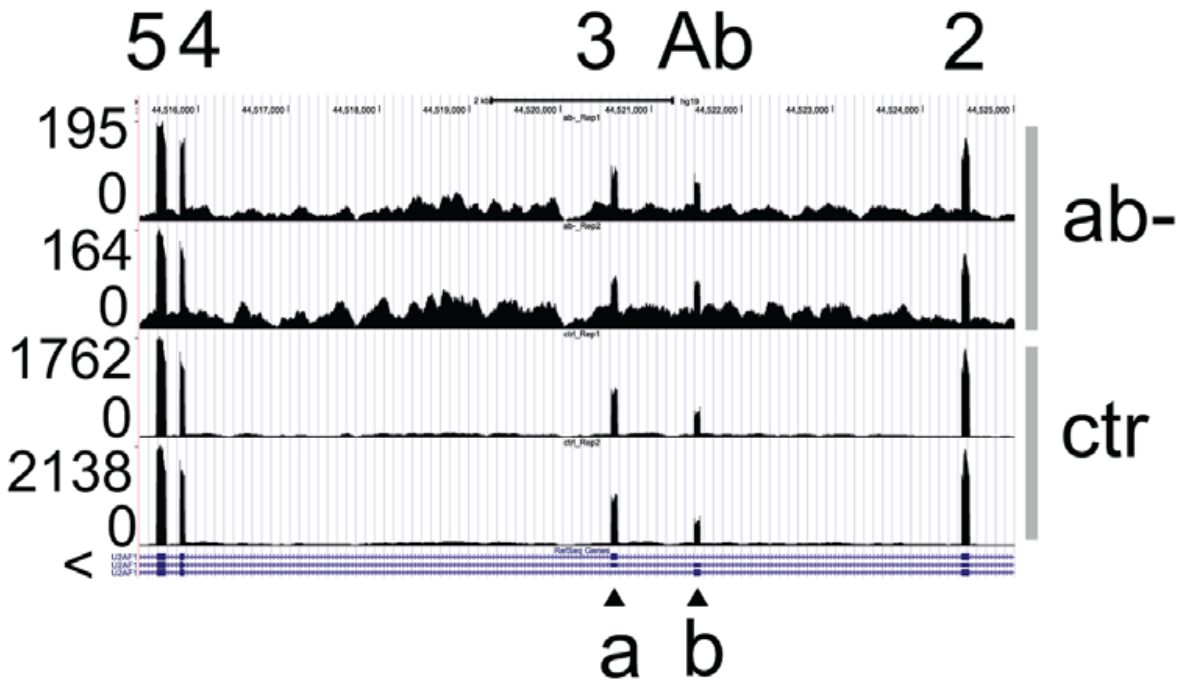

Figure 6.7: A genome browser view of exon Ab- and 3-containing isoforms in depleted cells (ab-) and controls (ctr). Exons are numbered at the top and corresponding isoforms are shown at the bottom. Browser views are in the native gene orientation throughout; the 5'>3' transcriptional orientation is denoted by the > sign.

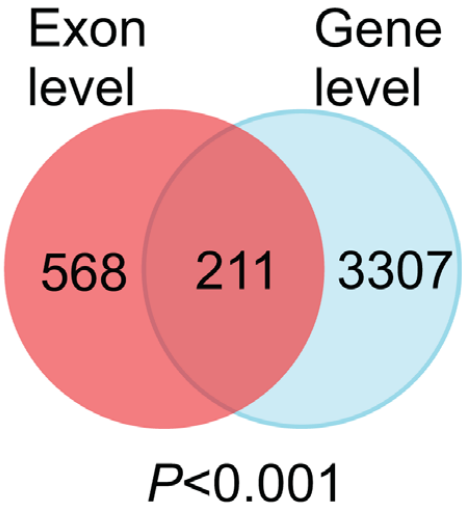

Figure 6.8: Significant sharing of genes identified by DEXSeq (exon-level) and Cufflinks (gene-level) as differentially expressed in ab- cultures versus controls.

As U2AF35 contacts the 3'ss AG dinucleotide as a part of complexes assembled *ad hoc* on each intron [203][204][205], DEXSeq and MISO algorithms were used to identify exons differentially used in depleted cells. DEXSeq is based on generalized linear models and relies on

biological controls to identify differential exon usage [241], whereas MISO employs a Bayesian approach and splice junction-spanning reads to detect specific alternative splicing events [223]. Altogether, DEXSeq identified a total of 484 upregulated and 575 downregulated exons in siRNA U2AF35ab-depleted cultures (termed ab-), with no bias toward either ($P > 0.05$, binomial test), whereas the number of MISO-detected events was ∼60% higher (Table C.5, C.6, C.7 and data not shown). Gene-level expression analyses with Cufflinks [229] revealed 1507 upregulated and 2011 downregulated genes (Table C.8).

Overlap of genes with differentially used exons and Cufflinks was highly significant (Figure 6.8), suggesting that these exons may contribute to overall gene expression alterations observed in depleted cultures. It is a possibility that a single exon may have altered usage indicated by much lower or higher expression when compared to control to the point where this change influences expression values for gene level analysis. This is most likely the reason for what is observed in the group of 211 genes overlapping between exon-level and gene-level analysis. Where a highly expressed exon, common for major isoforms, is observed to have much lower expression due to alternative initiation codons being used or other reasons such as usage of alternative stop codons or exon skippage, this may be reflected in the overall gene level expression if the expression value changes are considerable enough to alter gene-level expression measures.

Characterization of global alternative polyadenylation changes induced by U2AF35 depletion

Among differentially used exons, start and terminal exons were more frequent than expected while internal exons were significantly more common among downregulated than upregulated events (Figure 6.9).

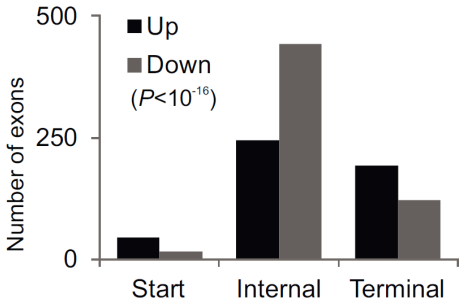

Figure 6.9: Distribution of start, internal, and terminal exons upregulated and downregulated in cells depleted of U2AF35. P-value was derived from a χ^2-test for the 3x2 contingency table. P-values for the first exons versus the pool of internal and terminal exons was $< 10^{-7}$.

There was a similar bias observed for start/end exons for published RNA-Seq data of cell cultures depleted of other RNA-binding proteins, including hnRNP C, a U2AF65 competitor [222], but not for cells depleted of a DNA-binding factor (Table C.10). Almost a half of differentially used exons were represented more than once per transcript, consistent with the presence of multiexon segments upregulated or downregulated in depleted cells (Figure 6.10).

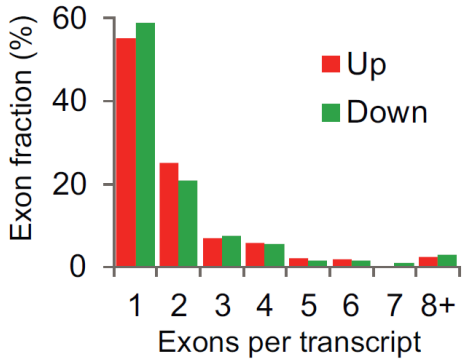

Figure 6.10: Proportion of transcripts with $\geq$2 differentially expressed exons following U2AF35 depletion.

Browser verification of individual events revealed that activation or repression of terminal exons was largely due to the altered usage of previously annotated alternative polyadenylation (APA) sites [242], with intronic APA sites as the most common APA category (Figure 6.11, 6.12 and Table C.11).

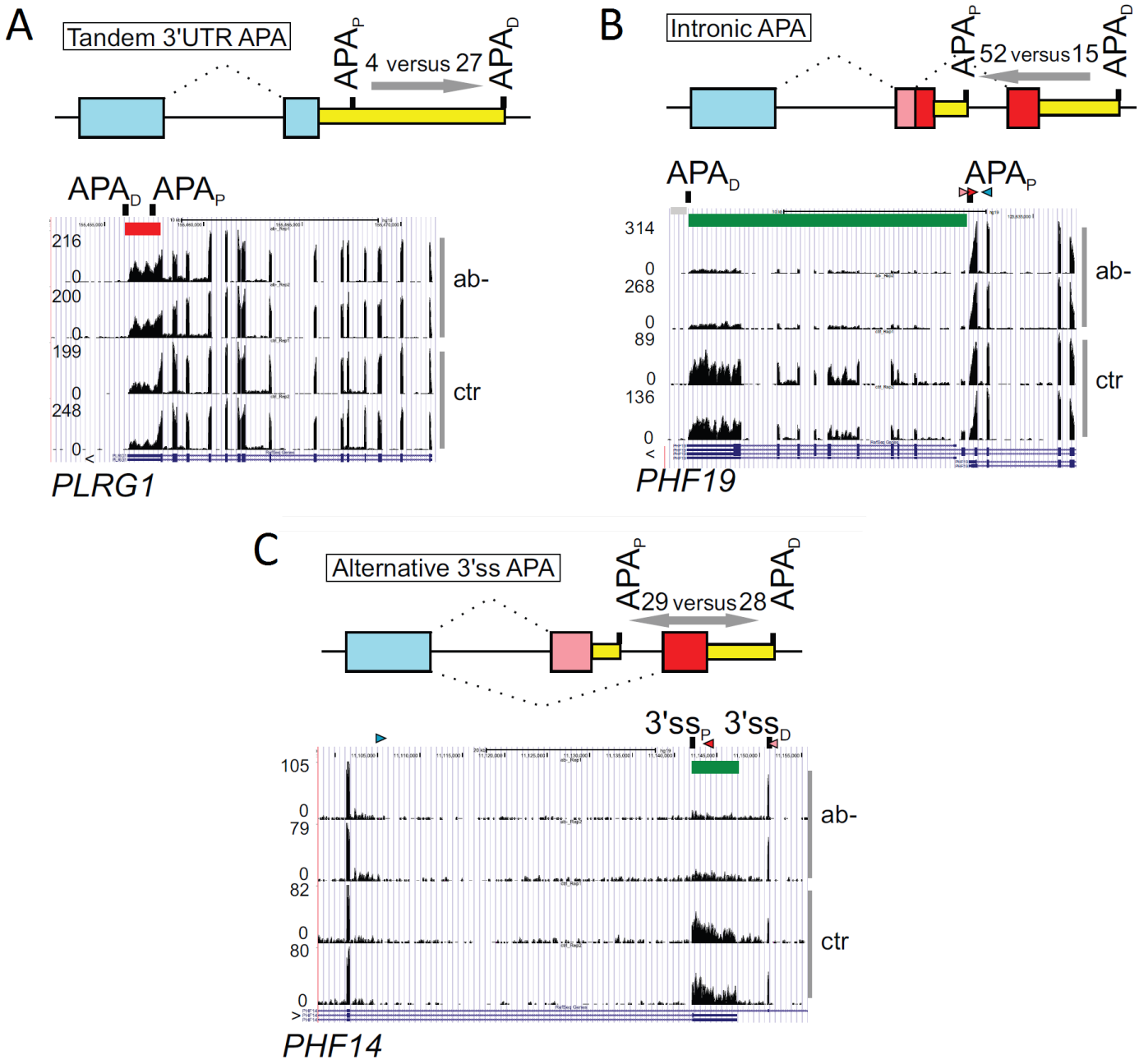

Figure 6.11: Alternative polyadenylation (APA) site usage in the indicated APA categories. Number of proximal and distal APA sites altered in depleted cells is shown above arrows that indicate shifts in APA site usage. APA_P, APA_D, proximal and distal polyadenylation sites. Upregulated and downregulated exons are indicated by red and green rectangles throughout. Each category is schematically shown at the top; yellow rectangles are 3' untranslated regions (UTRs), blue boxes are constitutive exons, red and pink boxes are alternative exons. Splicing is denoted by dotted lines.

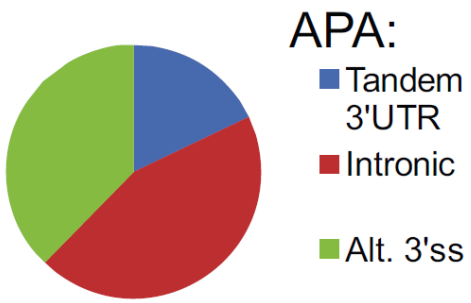

Figure 6.12: Frequency distribution of alternative polyadenylation site (APA) categories altered in ab- cells.

Categorization of 155 APA sites individually confirmed in a genome browser as affected by U2AF(35) depletion and verified against APA repositories revealed that while intron-proximal and -distal APA sites were about equally represented when APA was associated with alternative 3'ss, intronic APA sites promoted in ab- cultures were largely proximal whereas tandem 3'UTR APA sites were biased toward distal sites (Figure 6.11). Unexpectedly, breakdown of 211 DEXSeq-positive exons in genes that were either upregulated or downregulated in depleted cultures (i.e. Cufflinks-positive) showed preferential involvement of the first exons while terminal and internal exons were about equally represented (Figure 6.13).

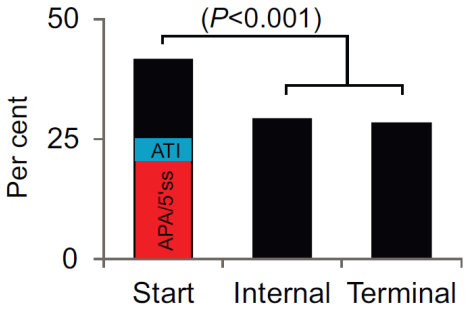

Figure 6.13: Breakdown of start, internal and terminal exons for Cufflinks-positive genes. APA/5'ss, altered alternative polyadenylation site (APA) or 5' splice site (5'ss) of the first intron in ab- cells; ATI, annotated alternative transcription initiation sites altered in ab- cultures. P-value was computed as in panel A.

Their individual browser inspection revealed that the excess was attributable to APA and 5'ss of the first introns while altered usage of annotated alternative transcription initiation sites was rare, further supporting a prominent impact of U2AF35 depletion on APA.

Together, these data indicated that APA was a major contributor to the differential exon usage in depleted cells and revealed APA category-dependent shifts of proximal and distal APA

sites conferred by a lack of U2AF(35) and/or an increase of U2AF65. They also suggested that a simple distribution of differentially used start, internal and end exons in RNA-Seq depletion experiments could be indicative of the relative importance of a depleted factor for eachRNA processing step.

A high validation rate of DEXSeq-detected alternative RNA processing events

Extensive experimental validation of 82 high-confidence DEXSeq-detected events from the same and independent depletion experiments with the same cell line confirmed 76 exons (Table C.9, Figures B.2, B.3), including alternative 3'ss APA in PHF14 and intronic APA in PHF19 (Figure 6.14).

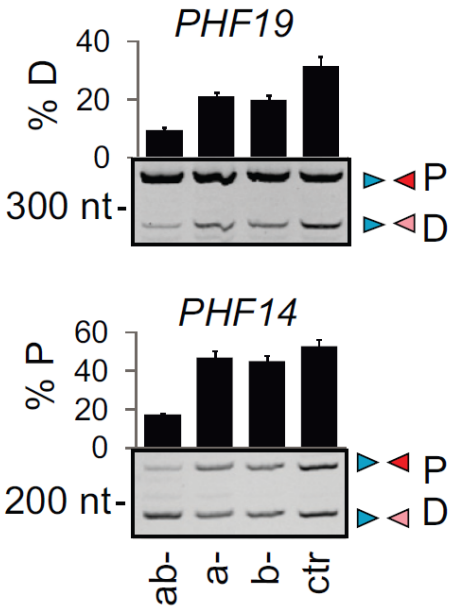

Figure 6.14: Validation of intronic/alternative 3' splice site (3'ss) alternative polyadenylation (APA) site usage in two plant homology domain-encoding genes shown in Figure 6.11, panels B and C. Polymerase Chain Reaction (PCR) primers are in Table C.9.

Apart from endogenous RNAs, altered exon inclusion as a result of U2AF(35) depletion was found also for exogenous transcripts (see below), including murine IgM transcripts without the RNA polymerase II (polII) pause site located between proximal and distal APA sites (Figure 2I).

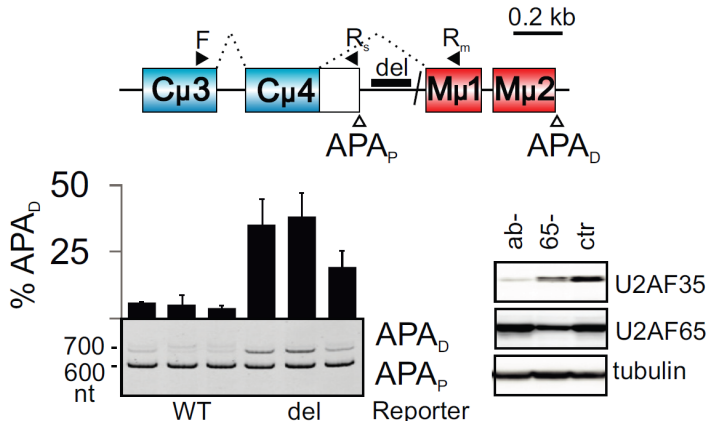

Figure 6.15: Control of mouse intronic alternative polyadenylation (APA) site usage by human U2AF35. Schematics of the mouse IgM minigene with APA sites giving rise to membrane (APA_D) and soluble (APA_P) immunoglobulins M (upper panel). Del, deletion of the RNA polymerase II (polII) pausing site (100). The lower panel shows RNA products of the wild-type and deletion-containing IgM minigenes transfected into HEK293 cells individually depleted of each U2AF subunit. Final concentration of U2AF65 short interfering RNA (siRNA) was 40 nM. Immunoblot is to the right.

As browser verification of MISO-detected events showed a higher false positivity than for DEXSeq (data not shown), their systematic validation was not carried out. However, a high stringency of default DEXSeq settings and a high sensitivity of MISO may provide a higher accuracy in identifying genuine alternative RNA processing events affected by U2AF35 depletion, complementing each other.

U2AF35 dependency can be explained largely by a lack of the U2AF heterodimer

Apart from altered exon usage seen only in ab- cultures, most transcripts showed a gradient in RNA processing defects, with a hierarchy ab- > a- > b- > controls, mirroring total levels of U2AF35 or U2AF (Figure B.2 and Figures 6.3, 6.4, 6.5). Individual depletion of each U2AF subunit in HEK293 cells showed that U2AF65 depletion, which reduces U2AF35 levels [211] (Figure 6.15), shifted usage of most exons in the same direction as U2AF35 depletion (Figure B.3). In MAPK8IP3 transcripts, however, depletion of U2AF35 promoted inclusion of an 18-nt exon; in contrast, U2AF65 depletion and depletion of U2AF35a failed to activate this exon and led to skipping of the preceding 12-nt exon instead. To understand this phenomenon, these pre-mRNAs in a dose-dependent transfection experiment shown in Figure 6.16 were examined. Interestingly, U2AF65 depletion increased the relative abundance of U2AF35b, suggesting that skipping of the 12-nt MAPK8IP3 exon in a- and U2AF65- cells was due to the

excess of U2AF35b. As U2AF65 depletion reduced U2AF65 more than U2AF35, potentially limiting the amount of the available heterodimer, residual levels of U2AF in each sample were estimated by measuring signal intensity from immunoblots from the same transfection (Supplementary Figure S3A). U2AF levels correlated significantly with the usage of most exons, particularly with those excluded from pre-mRNAs in depleted cells. In contrast,many exons upregulated in U2AF35 depleted cells were not activated in cells lacking U2AF65, and several very small exons, including a 12-nt MAPK8IP3 exon, did not correlate with U2AF (Supplementary Figures B.5, B.6, B.7). Thus, most but not all differential exon usage induced by U2AF35 depletion could be attributed to a lack of the U2AF heterodimer and sequences of thee exons should therefore reveal binding signatures of both U2AF subunits.

Identification of 3' splice sites altered by U2AF35 depletion

Combined MISO [223] and DEXSeq [241] analyses identified a total of 231 differentially used alternative 3'ss pairs. Their individual inspection in genome browsers confirmed 138 pairs of 3'ss, with 93 intron-proximal sites promoted and 45 repressed in ab- cultures (Table C.12). Only 51/138 sites (37%) promoted in ab- cultures were intrinsically stronger than their competing counterparts (binomial test, two-tail P = 0.003; Table C.14). A significant lack of canonical cytosine at position −3 and guanine at position +1 relative to upregulated proximal 3'ss (Figure B.9) was observed. Both positions contribute substantially to the 3'ss strength (37).

Since over a third of APA sites altered in depleted cells were associated with annotated alternative 3'ss usage (Figures 6.11 panel C, and 6.12, Table C.11), a test for their intrinsic strength contributes to APA selection was performed. Analysis of browser-verified 57 pairs of alternative APA 3'ss showed that unlike alternative 3'ss not associated with APA, intrinsically weaker sites were not preferred (28 stronger versus 29 weaker in ab- cultures). Likewise, enrichment of weaker sites when comparing proximal upregulated and downregulated APA 3'ss with APA 3'ss of terminal exons was not observed. Although the number of APA-associated alternative 3'ss was smaller than the number of non-APA 3'ss pairs, this result is consistent with an additional function of U2AF in APA control that is independent of interactions with alternative 3'ss of internal exons.

U2AF(35) depletion can influence 5' splice-site usage

U2AF35 contacts 3'ss AG [203][204][205], but at least 32 alternative 5'ss pairs influenced by U2AF35 depletion were identified. Eleven of them were tested using RT-PCR and 10 were confirmed (Tables C.9 and C.13; Figure B.10 panels A-G). Over a third of these 5'ss (n = 11) were located in the first introns. As for 3'ss, proximal (26 versus 6) and weaker (19 versus 13) 5'ss were activated more often in depleted cultures than their distal and stronger counterparts. Activation of cryptic 3'ss was occasionally accompanied by cryptic 5'ss activation of the same exon (Figure B.10 panel H). Because intronic APA sites compete with upstream 5'ss [243], the intrinsic strength of the first 5'ss upstream of 67 intronic APA sites affected by U2AF35 depletion was also determined. Their average scores tended to be lower (P = 0.12; Table C.14) than those of authentic counterparts of aberrant 5'ss, which were used as controls [236].

U2AF35 depletion activates exons with longer AGEZs and PPTs

Examination of 204 browser-verified internal exons that excluded multiple-exon regions associated with APA sites revealed that U2AF35 depletion promoted both exon inclusion and skipping as well as intron splicing and, sporadically, intronization within large exons (Figure B.11 panels A and B). Interestingly, exon activation in depleted cells often occurred in weak, incompletely removed introns (Figure B.11 panel C). U2AF(35)-sensitive exons were largely alternatively spliced and had sequence features typical of alternative exons (Figure B.12 and Table C.15). Importantly, upregulated exons had a significantly longer AGEZ than downregulated or control exons (Figure 6.16 panel A), but did not show a comparable decrease of other purine dinucleotides upstream of 3'ss (Figure B.13 panel A).

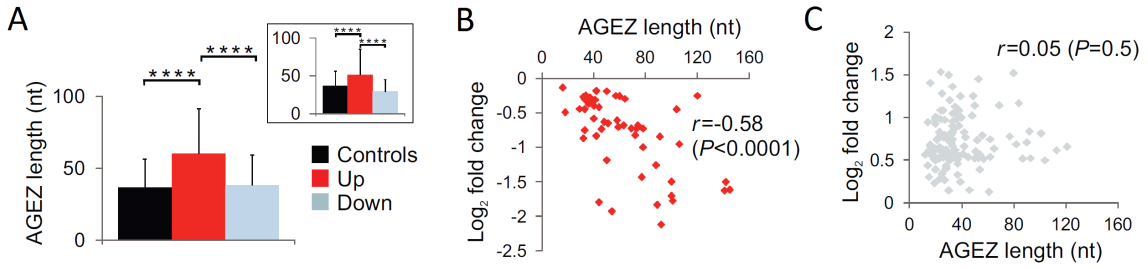

Figure 6.16: AG-exclusion zone (AGEZ) length of alternative 3' splice site (3'ss) (inset) and internal exons affected by U2AF35 depletion. Columns show means, error bars denote Standard Deviations (SDs). The number of each event is in the legend to Figure B.13. P-values were derived from t tests; $****, P < 0.00005$. AGEZ length correlated with the expression change of upregulated (B) but not downregulated (C) exons; r, Pearson correlation coefficient.

The AGEZ length correlated with the expression change of upregulated but not downregulated exons (Figure 6.16 panels B and C). Upregulated exons had also longer PPTs for the best predicted BP although the associated P values were lower than for the AGEZ (Figure 6.17 panel A).

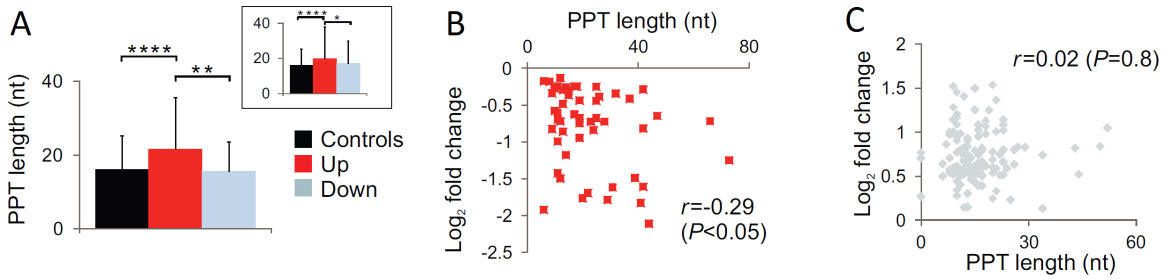

Figure 6.17: Polypyrimidine Tract (PPT) length of alternative 3' splice site (3'ss) (inset) and internal exons affected by U2AF35 depletion. Columns show means, error bars denote Standard Deviations (SDs). The number of each event is in the legend to Figure B.13. PPT length was computed for Branch Points (BPs) with the highest Support Vector Machine (SVM) scores [237]. P-values were derived from t-tests; $***, P < 0.00005$; $**, P < 0.005$. PPT length correlated with the expression change of upregulated (B) but not downregulated (C) exons; r, Pearson correlation coefficient.

The PPT length also correlated with the change in usage of upregulated exons (Figure 6.17 panels B and C). A specific lack of AGs and longer AGEZs/PPTs was found also upstream of alternative 3'ss upregulated in ab- cultures (insets in Figures 6.16 panel A, 6.17 panel D, and B.13 panel B) and this tendency was observed also for as few as 57 alternative 3'ss leading to APA (Figure B.14, Figure 6.11 panel C).

Interestingly, the lack of AGs upstream of upregulated exons was associated with adenine depletion between position -17 and -38, which approximately corresponds to the optimal BP location [217][244], whereas guanine depletion was more widely distributed (Figure B.15 panels A and B). Instead, upregulated exons showed enrichment for adenine at positions -3 to -9 while uridine tended to show the opposite, with enrichment between -17 and -38 and depletion closer to the 3'ss. Conversely, downregulated exons showed adenine enrichment in the optimal BP region, particularly in smaller exons (Figure B.15 panels C-E), and uridine enrichment closer to 3'ss. Additional known and unknown motifs in sequences flanking U2AF(35)-sensitive junctions using the MEME suite of programs [238] were also searched for, however, no significant hits were found.

The conclusion is that 3'ss upregulated in ab- cells have longer AGEZs/PPTs, adenine depletion in the optimal BP location and enrichment closer to 3'ss, with uridines showing the opposite pattern. This arrangement moves PPTs further upstream of upregulated exons as compared to their downregulated counterparts. The observed widespread repression by U2AF(35) could thus reflect a lack of AG dinucleotides upstream of 3'ss, which are generally repressive when located downstream of BPs [217][245][245], and/or longer, more upstream PPTs, whichmay bind exon-repressive PPT binding proteins [222][246]. Because these sequence characteristics are likely to influence secondary structure formation across 3'ss, this group of exons should provide a powerful tool to study regulation of 3'ss by RNA-binding proteins and RNA folding.

Exons repressed by U2AF(35) are stimulated by PUF60 and inhibited by CAPERα

To examine the role of U2AF-related proteins in usage of U2AF-dependent exons, inclusion of 44 exons in HEK293 cells depleted of PUF60, CAPERα (RBM39), CAPERα (RBM23) and two other Y-binding proteins (Figure 6.18) was measured.

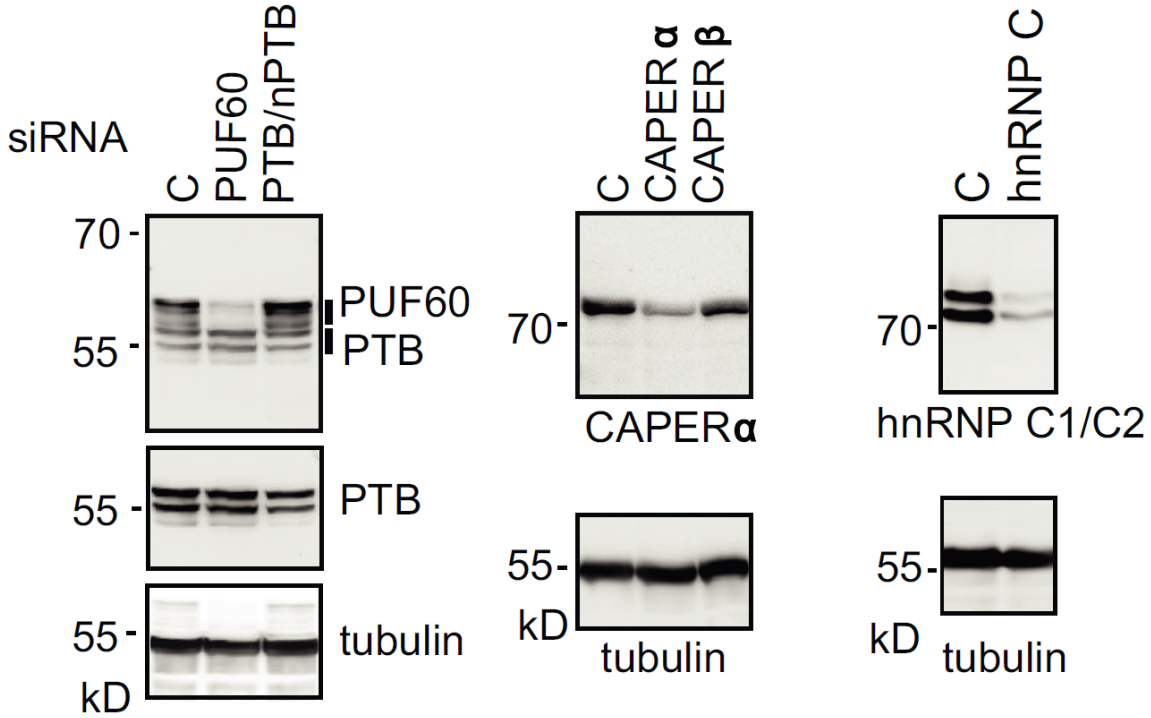

Figure 6.18: Immunoblots prepared from lysates from HEK293 cells depleted of poly(Y)-binding proteins (indicated at the top). Antibodies are shown at the bottom or to the right.

PUF60 and U2AF were individually capable of protecting the 3'ss AG in footprinting experiments [247], but the exact function of CAPERα/β in recognition of 3'ss or U2AF-dependent exons is unknown. In addition, depletion of hnRNP A1 was performed, which allows U2AF to discriminate between pyrimidine-rich RNA sequences followed or not by a 3' splice-site AG [248], and DEK, which facilitates the U2AF35-AG interaction and prevents binding of U2AF65 to pyrimidine tracts not followed by AG [249] (data not shown). Remarkably, the majority of exons upregulated in cells depleted of U2AF35 were downregulated in cells depleted of PUF60 (Figures 6.19, B.16, B.17, and B.18).

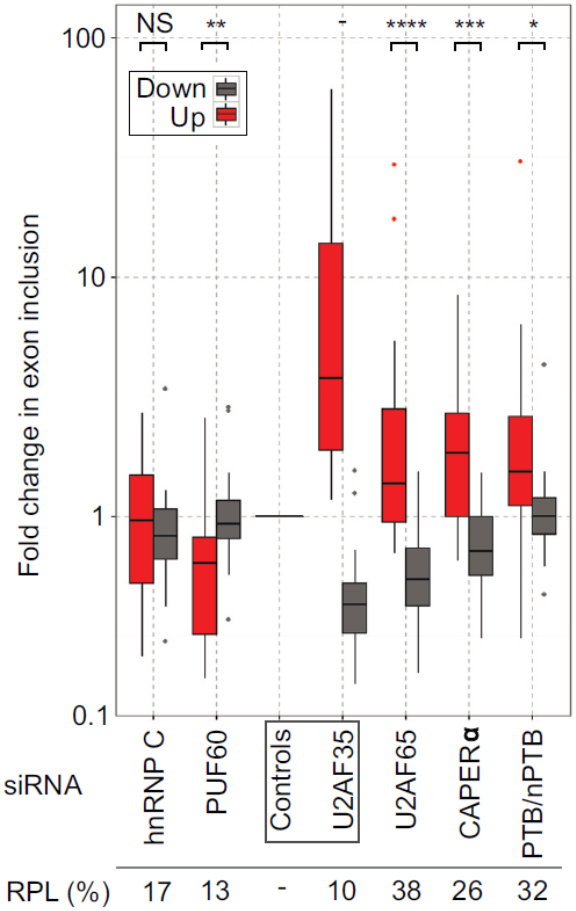

Figure 6.19: Functional antagonism and synergism of Y-binding proteins and U2AF. Exon inclusion levels of each transcript are in Figures B.16, B.17, and B.18; Reverse-Transcription Polymerase Chain Reaction (RT-PCR) primers are in Table C.9. Residual protein levels (RPL) were estimated from immunoblots shown in Figure 6.18 . The Y-axis is on a $\log_{10}$ scale. Average changes between inclusion levels of upregulated and downregulated exons were compared by the Wilcoxon-Mann-Whitney test ($*, P < 0.05; **, P < 0.005; ***, P < 0.0005; ****, P < 0.00005$; NS: not significant; -: not tested).

In contrast, CAPERα, and to a lesser degree PTB, showed synergism with U2AF(35) for this group of exons (Figure 6.19). A significant directionality of hnRNP C, DEK and hnRNP A1 could not be established with this sample size (Figure 6.19 and data not shown). To confirm that PUF60 stimulates exons repressed by U2AF, exon inclusion of a minigene reporter transfected into HEK293 cells lacking or overexpressing PUF60 was measured (Figure 6.20).

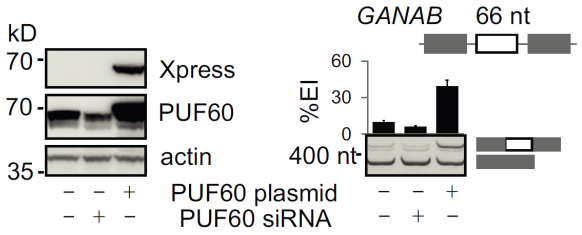

Figure 6.20: Opposite effects of PUF60 depletion and overexpression on a *GANAB* exon. Immunoblots with the indicated antibodies are to the left and Reverse-Transcription Polymerase Chain Reaction (RT-PCR) to the right. The *GANAB* minigene is schematically shown at the top; alternative exon is denoted by a white rectangle. Exogeneous RNA products were amplified by primers PL3 and PL4. Error bars are Standard Deviations (SDs) of duplicate transfections.

It was found that cells lacking PUF60 showed increased skipping of this exon, whereas the PUF60 overexpression increased its inclusion in the mRNA. Thus, U2AF(35)-induced exon usage was predictive of responses to other Y-binding proteins, revealing the connection between antagonism and synergism of U2AF-related proteins and characteristic 3'ss organization described above.

Unpaired regions upstream of U2AF(35)-activated and - repressed exons

Because characteristic changes in nucleotide frequencies upstream of U2AF(35)-dependent exons (Figure B.15) are likely to affect formation of RNA secondary structure, which can influence 3'ss utilization [250][251], computation was performed for position-specific probabilities of RNA single-strandedness [240] for high-confidence upregulated and downregulated internal exons. Remarkably, upregulated exons showed on average significantly higher unpaired probabilities at most positions between -25 and -50 than downregulated exons while a lower single-strandedness was observed closer to their 3'ss (Figure 6.21).

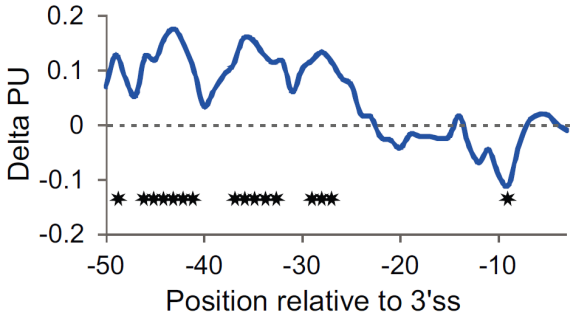

Figure 6.21: Positional differences in unpaired probabilities upstream of U2AF(35)-activated and repressed exons. Positive delta PU values signify a higher average single-strandedness of upregulated exons in the optimal Branch Point (BP) location and further upstream as compared to downregulated exons whereas negative values reveal their tendency to engage in local base-pairing interactions closer to 3' splice site (3'ss). Stars denote positions with P-values <0.05.

This finding suggests that intramolecular base-pairing interactions over relatively long distances upstream of 3'ss control exon repression and activation by U2AF and, most likely, by U2AF-related proteins that showed functional antagonism and synergism with U2AF and bind single-stranded RNA (Figure 6.19).

U2AF(35) preferentially regulates nuclear proteins involved in RNA binding

Functional enrichment analysis [231] of exons/genes differentially used in U2AF35 depleted cells showed that they were enriched in proteins involved in RNA/nucleotide binding, respectively (Figure 6.22).

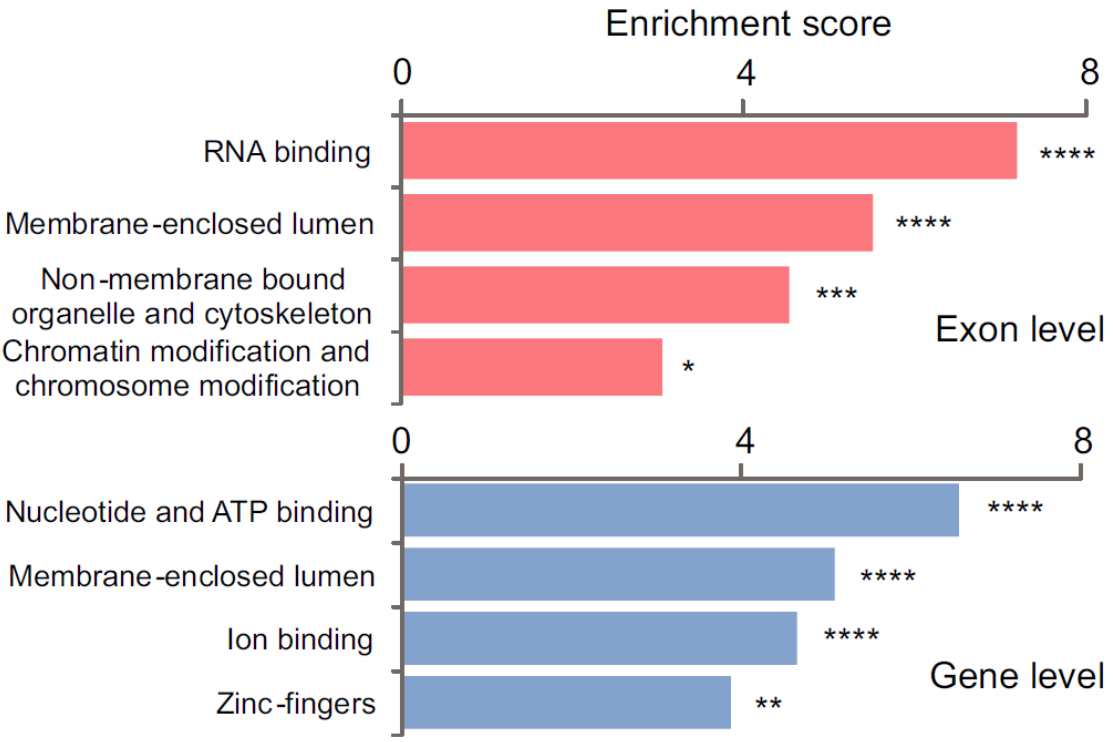

Figure 6.22: Functional enrichment analysis using DAVID [231]. Asterisks denote the False Discovery Rate (FDR) significance ($*, P < 0.05; **, P < 0.005; ***, P < 0.0005; ****, P < 0.00005$).

Figure 6.23 panel A shows a single example of alternative 3'ss in a known RNA-binding factor SF1. These 3'ss are responsible for production of BP-binding SF1 proteins with variable C-termini [252] and are associated with tissue-specific APA sites (Figure B.19). These proline-rich regions interact with PRPF40A [253], a component of the splicesomal E complex [254]. Depletion of U2AF35 was associated with upregulation of SF1 mRNA and promotion of the distal 3'ss (Figure 6.23).

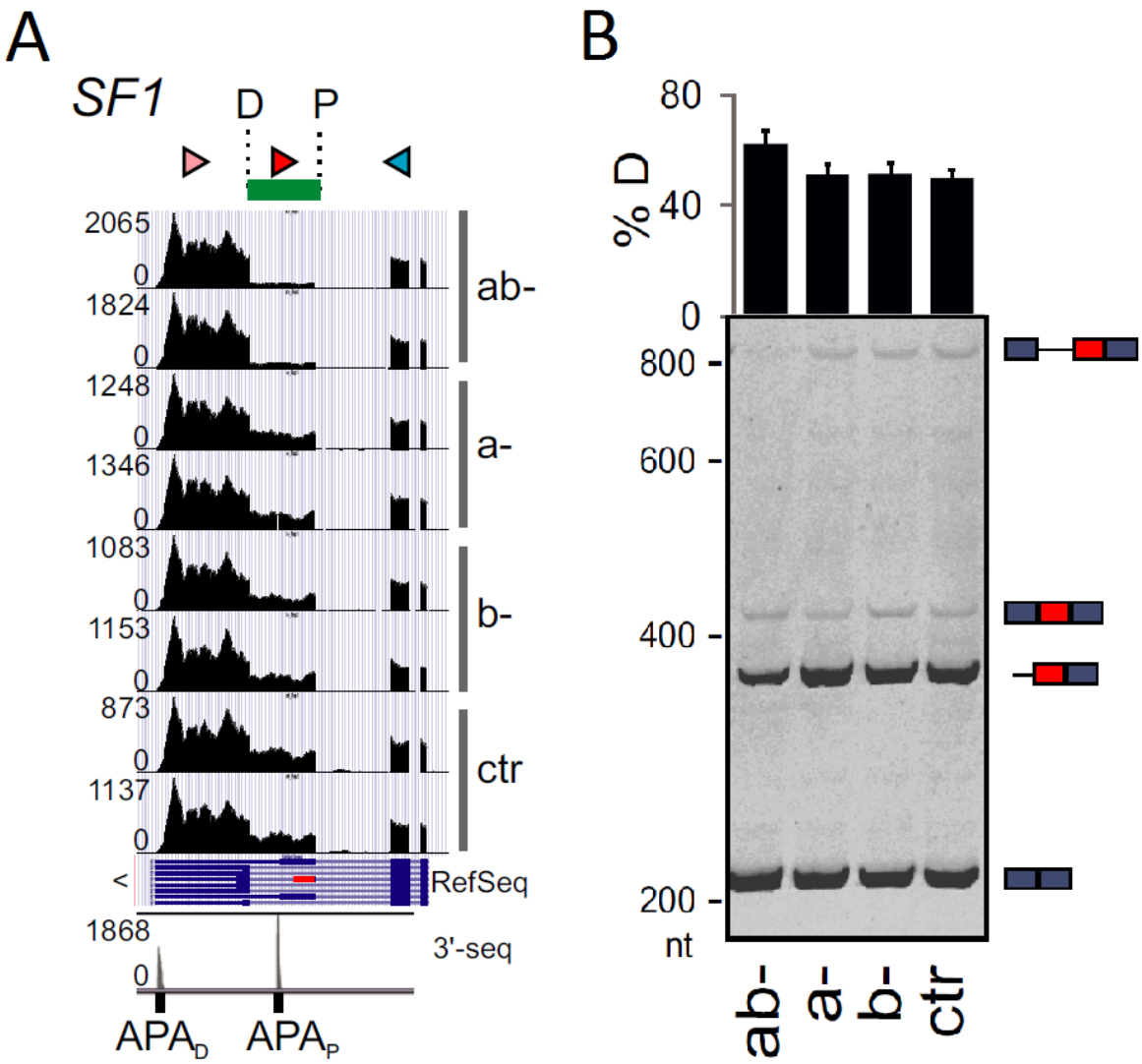

Figure 6.23: Regulation of alternative 3' splice site (3'ss) site usage of *SF1* by U2AF(35). (**A**) P, D, proximal and distal 3'ss of the last *SF1* intron. Arrowheads denote Reverse-Transcription Polymerase Chain Reaction (RT-PCR) primers (Table C.9) used in panel B. The last track shows unified 3'-seq coverage from multiple tissues with the location of two APA sites [242]. Figure B.19 shows their usage in various cell types; the proximal alternative polyadenylation (APA) site is used only weakly in HEK293 cells. (**B**) Activation of distal 3'ss *SF1* in depleted cells. RNA products are schematically shown to the right. Error bars denote Standard Deviations (SDs.)

Transfection of the SF1 splicing reporter constructs into ab- cells confirmed repression of the proximal site (Figure 6.24).

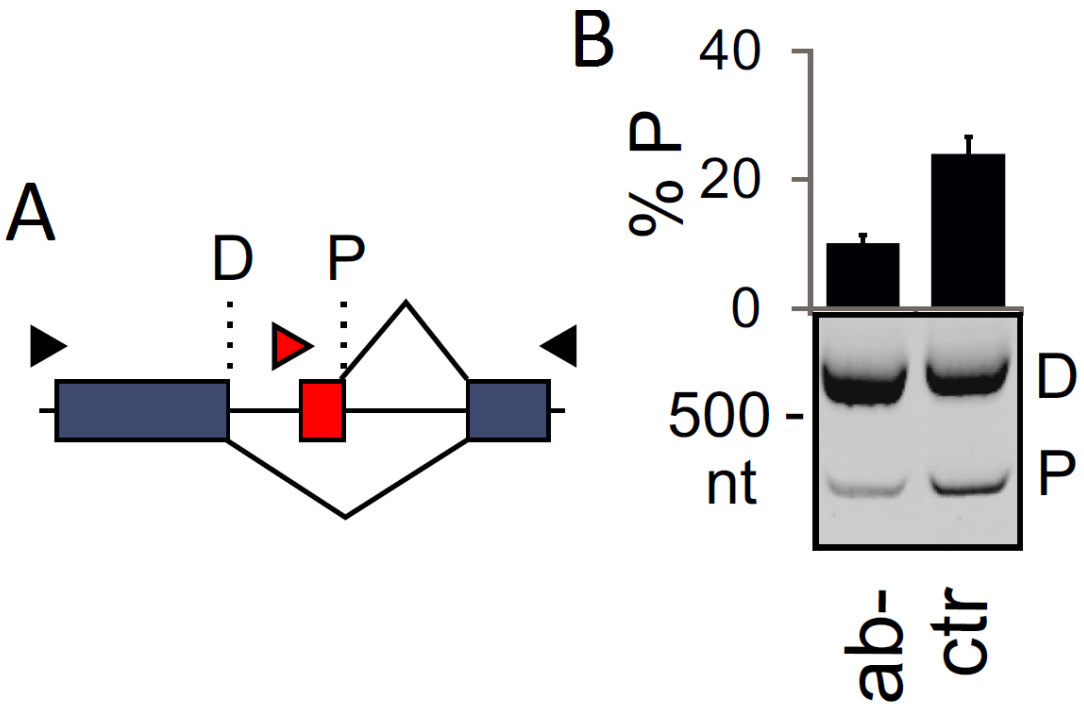

Figure 6.24: Distal and proximal 3'splice site (3'ss) usage in *SF1* influenced by U2AF(35). (**A**) Schematics of the *SF1* minigene. Arrowheads show primers used for Reverse-Transcription Polymerase Chain Reaction (RT-PCR) in panel B. (**B**) RNA products of the *SF1* minigene. Transient transfections were into HEK293 cells (mock)-depleted of U2AF35.

Thus, U2AF(35) regulates the length of SF1 3'UTR and, potentially, its PRPF40A interactions.

In addition to RNA binding, differentially expressed exons/genes were enriched in proteins involved in cytoskeleton organization, chromatin modification and proteins found in the organelle lumen (Figure 6.22). Figure 6.25 gives a summary of exons in transcripts involved in actin dynamics.

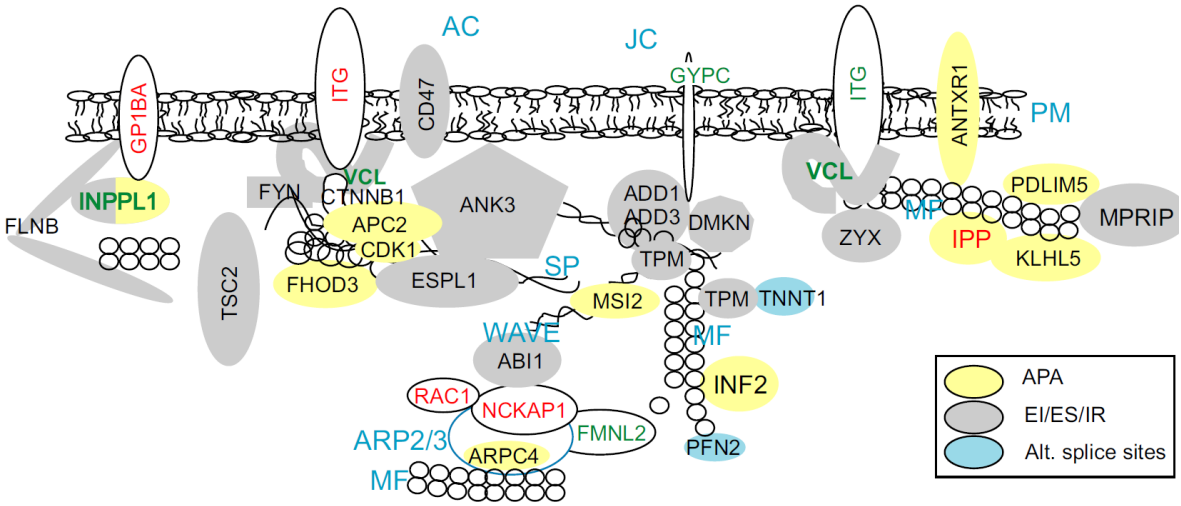

Figure 6.25: Exon-centric regulation of actin dynamics. Protein products are drawn as coloured shapes in a cellular context. Red and green text shows genes that were upregulated and downregulated in ab- cells, respectively; blue text denote protein complexes or subcellular structures: PM, plasma membrane; MF, microfilaments; SP, spectrin; AC, ankyrin complex; JC, junctional complex. EI, exon inclusion; ES, exon skipping; IR, intron retention. Actin monomers are schematically shown as small circles. ITG shapes denote multiple integrins upregulated and downregulated in ab- cultures (Table C.8).

Tropomyosin genes (TPMs), which control function of actin filaments in a tissue-specific manner [255], serve as the most prominent examples. In ab- cultures, exon 6a of TPM1 and TPM2 was repressed and exon 6b, which has a long PPT in both genes, was activated (Figure 6.26).

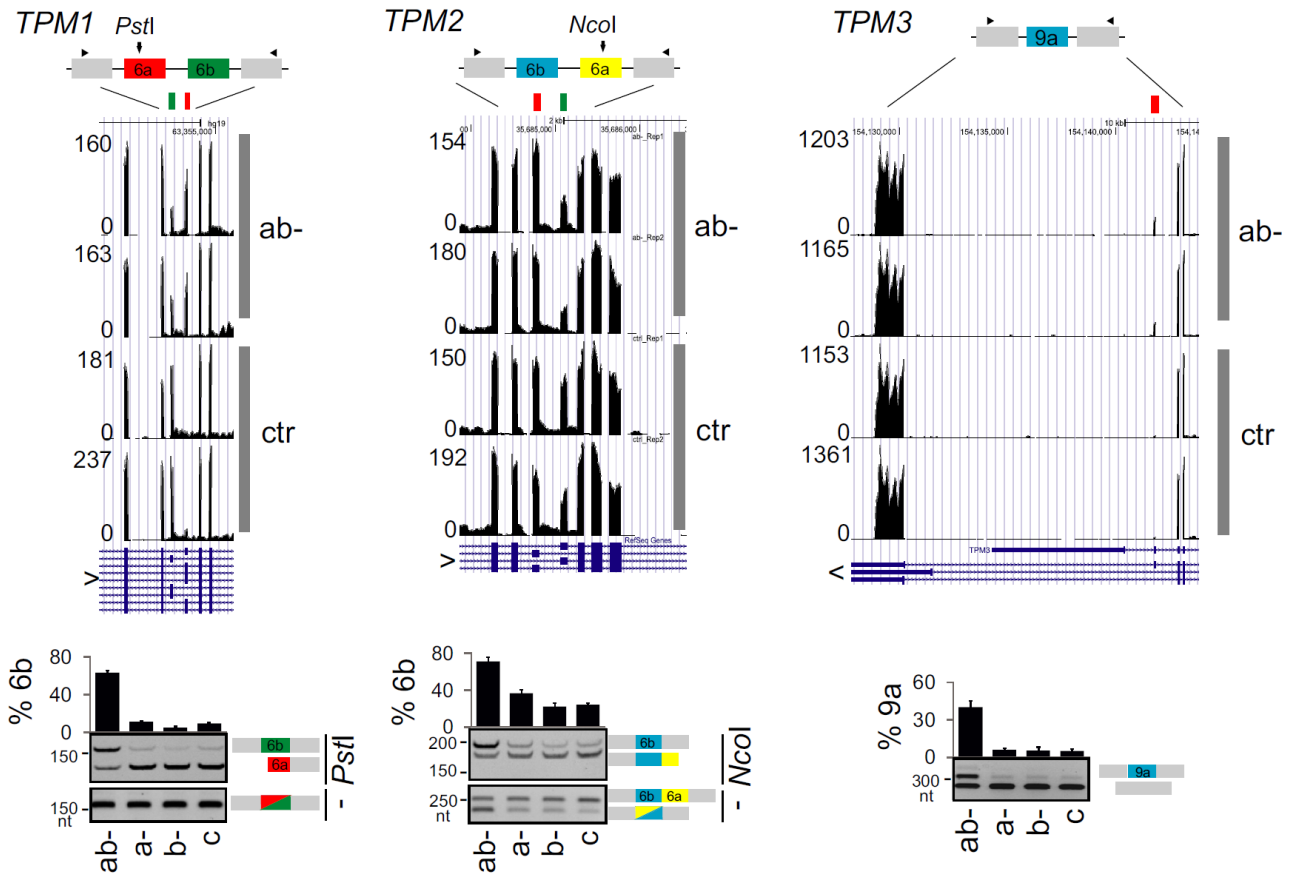

Figure 6.26: Identification of U2AF(35)-sensitive exons in the tropomyosin genes (upper panels) and their validation by Reverse Transcription Polymerase Chain Reaction (RT-PCR) (lower panels). Restriction enzymes to establish the identity of mutually exclusive exons are indicated to the right; small fragments of digested products are not shown. Alternative exons are coloured, arrowheads denote PCR primers.

Isoforms containing exon 6b have lower calcium sensitivity than isoforms with exon 6a, which may be required for a specific interaction with troponin [256]. TNNT1, a gene coding for a slow skeletal muscle troponin T, sustained a cryptic 3'ss activation upon U2AF35 depletion (data not shown). Besides TPM1/2, exon 9a of TPM3 was upregulated in depleted cells as well as the TPM4 transcripts (Figure 6.26 and Table C.8). TPM3 isoforms expressing exon 9a are more widely distributed in tissues than isoforms containing exon 9c [257], suggesting that U2AF(35) restricts tissue expression of -tropomyosin.

Finally, Figure 6.27 shows U2AF(35)-induced alterations of exon usage in eight genes coding for the histone modifying SAGA (Spt-Ada-Gcn5 Acetyltransferase) complex components, including KAT2A and KAT2B.

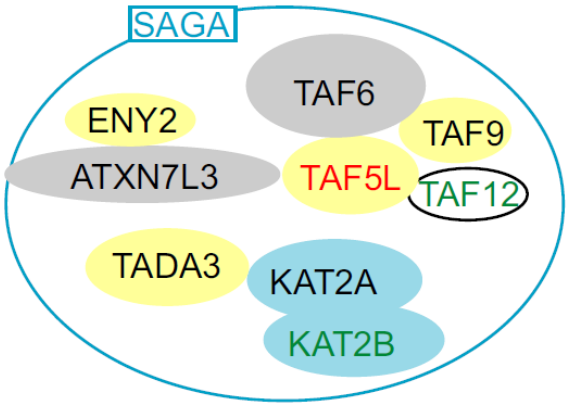

Figure 6.27: Components of the SAGA complex influenced by U2AF35 depletion. Red and green text shows genes that were upregulated and downregulated in ab- cells, respectively; blue text denote protein complexes or subcellular structures.

The U2AF35 depletion promoted the use of a distal 5'ss in KAT2A (Figure B.10 panel F), modifying the balance of alternatively spliced GCN5 isoforms and, most likely, histone acetyltransferase activity. 5'ss selection was influenced also in a GCN5 paralog KAT2B (PCAF; Figure B.10 panel G). The TADA3 gene, which encodes the GCN5 interaction partner, sustained activation of the proximal APA site upon depletion. In contrast, a proximal APA was repressed in TAF5L, a component of the SAGA architecture module (Figure 6.28).

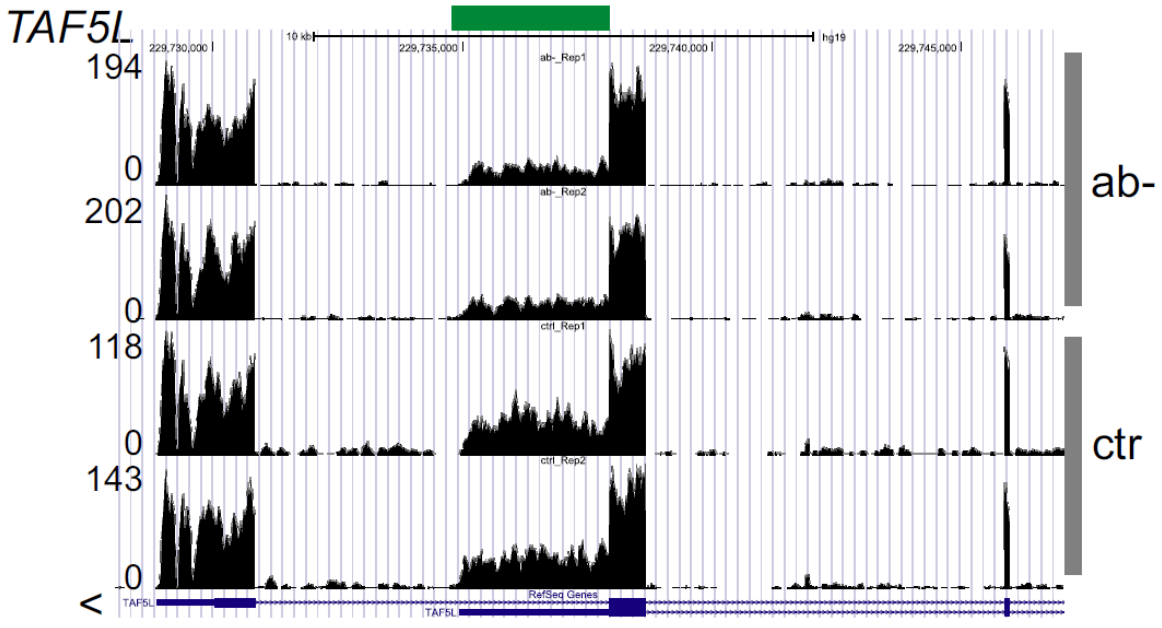

Figure 6.28: A genome browser view of a differentially used alternative polyadenylation sites (APA) in a representative SAGA transcript. Additional SAGA transcripts are shown in Figure B.10 panels F and G.

Activation of proximal APA was found also for TAF9 while a proximal 3'ss was promoted in ATXN7L3, a component of the deubiquitination module.

Exon-, isoform- and gene-level control of U2AF35 interaction partners

Over 25% (14/51) genes encoding high-confidence interaction partners of U2AF35 [258] were significantly upregulated or downregulated in depleted cells, which was more than expected (hypergeometric test, $P < 0.05$, Table 6.1).

Partner gene	Gene-level	Exon-level
U2AF2	-1.17	NS
SRSF1	-0.43	Longer 3'UTR as a result of promotion of distal APA
SF3B3	0.54	Longer 3'UTR as a result of promotion of distal APA
SF3B14	-0.65	NS
SF3B1	NS	Promotion of distal APA, retention of intron containing proximal APA
CDC5L	-0.71	NS
PLRG1	NS	Longer 3'UTR as a result of promotion of distal APA
ZCCHC8	-0.51	NS
U2AF26	1.43	NS
SNRPA1	0.97	NS
SAP18	-0.53	NS
SON	-0.58	Promotion of putative proximal APA site
MCM5	-0.55	NS
NHP2L1	-0.47	Promotion of proximal alternative transcription initiation site

Table 6.1: Gene- and exon-level alterations of high-confidence interaction partners of U2AF35 in ab- cells. Negative log2-fold values indicate upregulation in ab- cells; positive values indicate downregulation. NS, not significantly altered by U2AF35 depletion.

U2AF2 showed the highest increase, followed by *CDC5L*, which encodes a key component of the PRPF19-CDC5L complex required for the catalytic step of splicing [259]. Upon U2AF35 depletion, the CDC5L interaction partner *PLRG1* showed 3'UTR lengthening but CTNNBL1, CCAP1, CCAP3 and CCAP6 mRNAs were not noticeably altered. SNRPA1, which binds U2 snRNA [260], had reduced mRNA levels in depleted cultures. Transcripts encoding U2AF35-related protein U2AF26, which can interact with U2AF65 and functionally substitute U2AF35 in constitutive and enhancer-dependent splicing [261], were also downregulated while SF3

components *SF3B1* and *SF3B3* showed alterations in APA selection.

Figure B.20 shows examples of exon usage dependencies of high-confidence U2AF35 interaction partners. In depleted cells, PTC-containing cryptic exons in *U2AF2* and *CAPERα* were downregulated and both genes were upregulated. Expression of *CAPERα* exon 4 was also increased while *PUF60* exon 5 was downregulated. Alternative 5'ss of *U2AF2* exon 10, which controls the inclusion of four amino acids in U2AF65, was not altered by U2AF35 depletion and the two *U2AF2* isoforms were upregulated in depleted cells to the same extent (Figure 6.6). Together, these data identify high-confidence interaction partners of U2AF35 whose expression was altered upon U2AF(35) depletion and reveal exon-centric regulation of closely related U2AF genes.

Evidence for a distinct function of U2AF35 isoforms

U2AF35a and b differ from each other at seven aminoacid positions located in the RNP2 motif of the UHM, a proximal part of a long α-helix A and a disordered segment between the two folded regions [210]. Isoform-specific depletions identified transcripts exhibiting a gradient in RNA processing defects reflecting total levels of U2AF(35) but also exons activated only upon ab- depletion where U2AF levels were the lowest (Figures B.2, B.3, B.4, and B.5). However, events that occurred only in b- (Figure B.21) and a- cultures(Figure 6.29) were also found, in which the less abundant *U2AF1b* was in excess (Figure 6.5 and Figure B.4 lane 6).

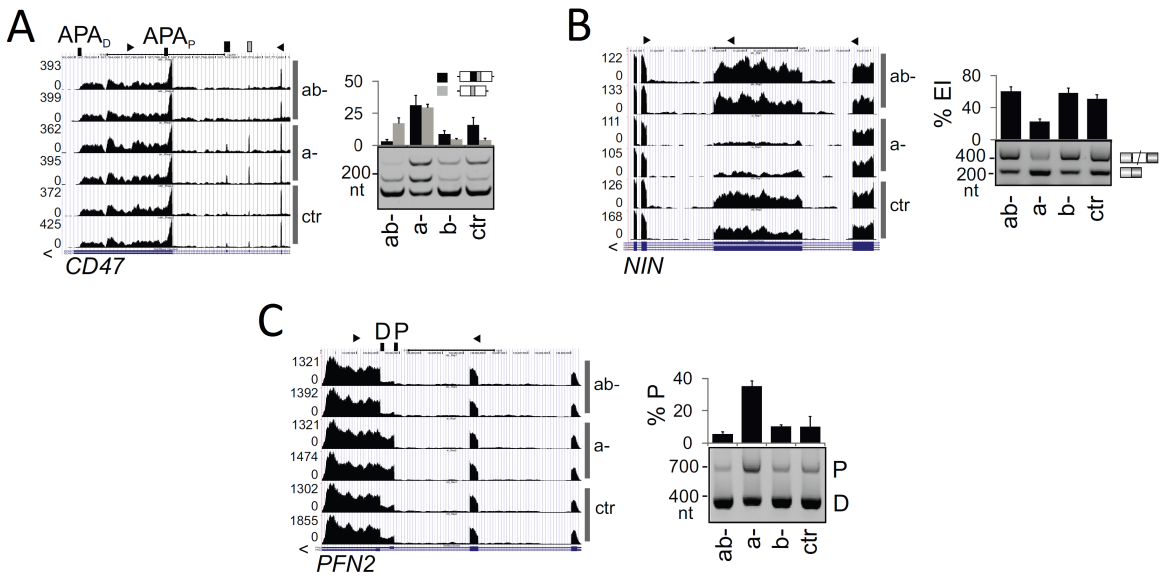

Figure 6.29: Genome browser views of endogenous transcripts showing isoform-specific responses to U2AF35 depletion (left panels) and their validation using Reverse-Transcription Polymerase Chain Reaction (RT-PCR) (right panels).

They were confirmed in independent transfections using varying siRNA concentrations (Figure 6.30) and with exogenous transcripts (Figures 6.31 and 6.32).

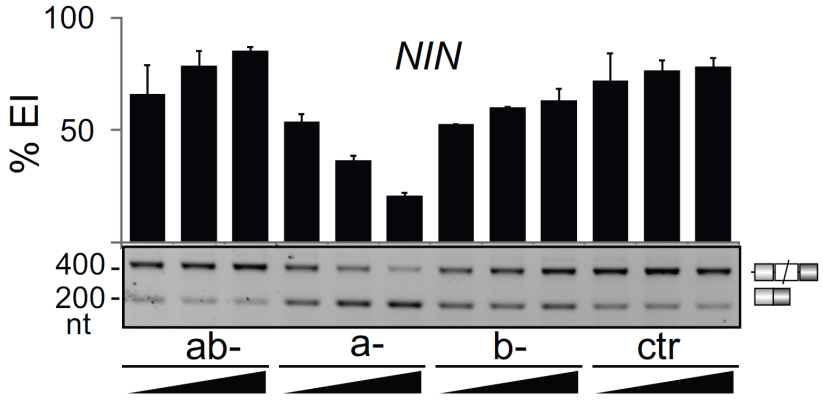

Figure 6.30: *NIN* exon inclusion levels in the indicated depletions. Final concentration of short interfering RNAs (siRNAs) was 6.7, 20 and 60 nM. Error bars are Standard Deviations (SDs).

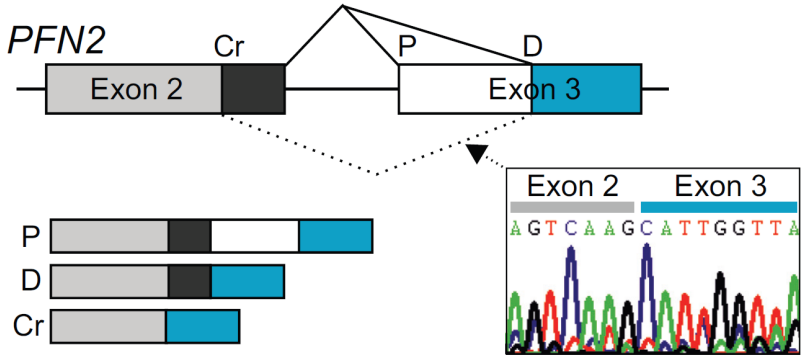

Figure 6.31: Schematics of the *PFN2* minigene. Chromatogram illustrates transcripts spliced to the cryptic 5'splice site (5'ss) of intron 2 (Cr); P, D, proximal and distal 3'ss.

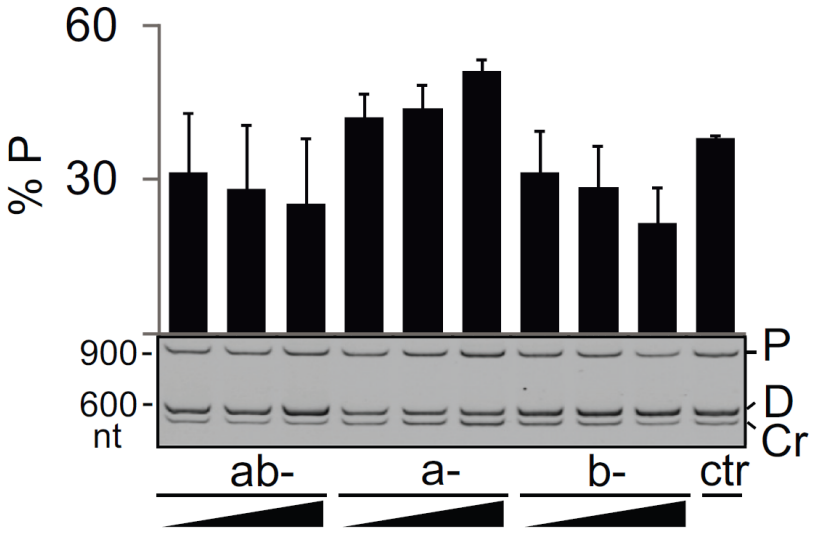

Figure 6.32: Opposite effects of U2AF35a and U2AF35b on splice site selection in exogenous *PFN2* transcripts. Spliced products are shown in Figure 6.31

.

For example,U2AF35a depletion activated a proximal 3'ss of *PFN2* intron 2 and a distal cryptic 5'ss of the same intron whereas U2AF35b depletion was associated with the opposite effect in a dose dependent manner (Figure 6.32). The proximal 3'ss was promoted by U2AF35b and repressed by U2AF35a also in reconstitution experiments in which individual addition of plasmids expressing each isoform to ab- cells was done(Figure 6G 6.33).

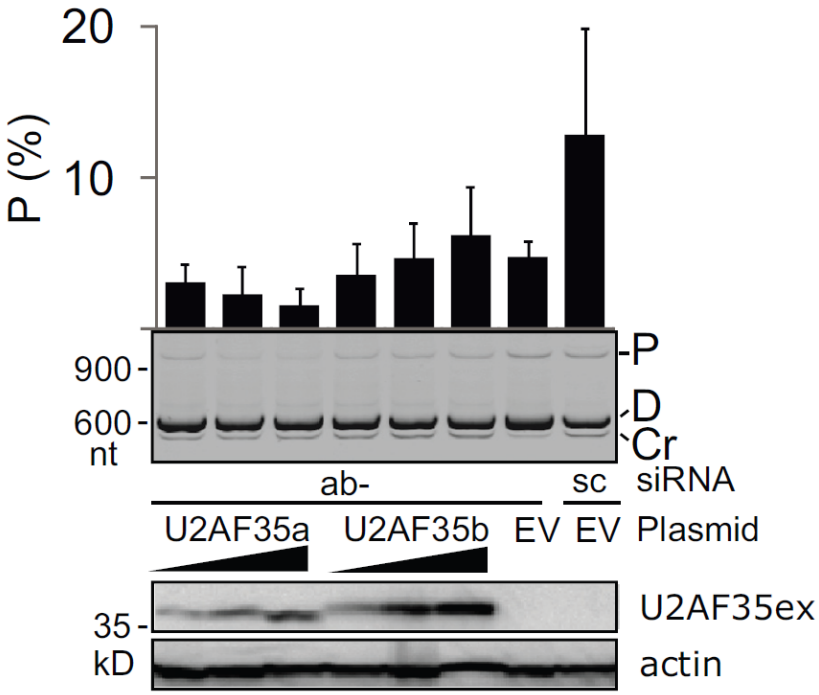

Figure 6.33: Isoform-specific rescue of 3'splice site (3'ss) of PFN2 intron 2. The amount of rescue plasmid DNA was 20, 65 and 200 ng. Immunoblot with Xpress (U2AF35ex) and β-actin antibodies is shown in the lower panel.

Repression and activation of the *PFN2* 3'ss was confirmed in cultures depleted with SSOa and SSOb (Figure B.1 panel B and data not shown). Alternative 3'ss of *PFN2* generate isoforms with distinct C-termini of profilin 2, a key actin-monomer binding protein, that have distinct binding affinities for proline-rich sequences and show tissue-specific expression [262][263]. Interestingly, exogenous U2AF35b expression was higher compared to U2AF35a (Figure 6.33). The existence of isoform-specific effects was also supported by the number of differentially expressed genes/exons in isoform-specific depletions, with a significant overlap in each category and a low ratio of exon-level versus gene level events in b- samples (Figure 6.34).

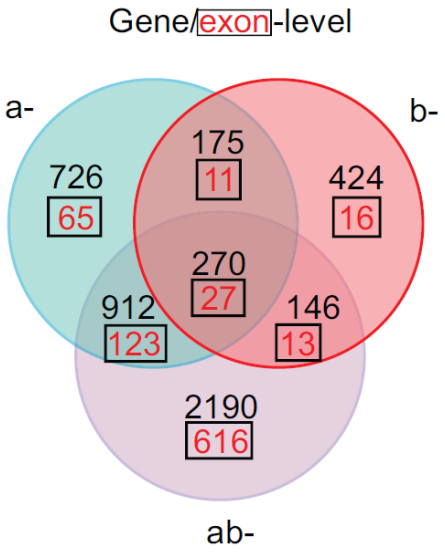

Figure 6.34: Three-way Venn diagram showing overlaps of differentially expressed genes/exons ($q < 0.05$) in ab-, a-, and b- depletions versus controls. Gene lists are in Tables C.5, C.6, C.7 and C.8.

In contrast to most transcripts, correlation between U2AF levels and exon usage was absent or decreased for genes with isoform-specific responses (Figures 6.35, B.4, B.5, B.6, and B.7).

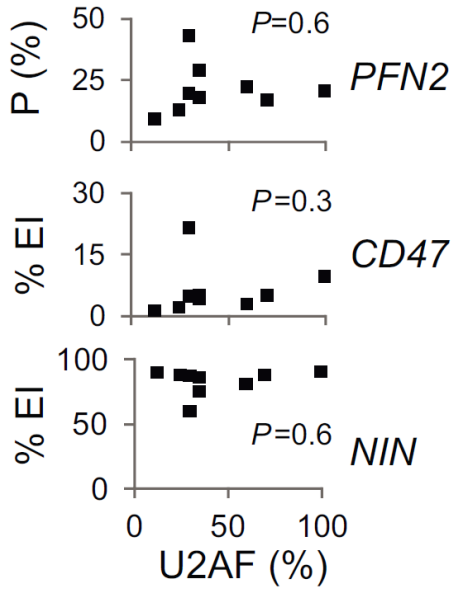

Figure 6.35: Exon/proximal 3'splice site (3'ss) usage in the indicated transcripts (y axis) and residual U2AF heterodimer levels (x axis) estimated from a transfection experiment shown in Supplementary Figures B.4, B.5, B.6, and B.7.

To reinforce these findings further, independent isoform-specific depletion experiments with total RNA depleted of rRNA were carried out. These samples contain a higher fraction of

unprocessed RNA than polyA-selected RNA, giving more information about intron splicing [264]. DEXSeq analysis followed by browser-assisted verification revealed that the bias toward start and terminal exons was even greater than for poly(A) samples and reconfirmed isoform-specific effects validated experimentally (Table C.10, Figures 6.29, 6.30, 6.31, 6.32, 6.33, 6.34, 6.35, and B.21).

Taken together, identification of transcripts with distinct responses to U2AF35a and U2AF35b argues for the existence of isoform-specific interactions that may confer even opposite effects on splice-site selection and provide additional level of exon regulation.

6.4 Discussion

The work shows the first global characterization of RNA processing alterations associated with depletion of U2AF35 isoforms. Our data reveal that U2AF function is not limited to 3'ss recognition but involves extensive APA control (Figures 6.9, 6.10, 6.11, 6.12, 6.13, 6.14, and 6.15), particularly through intronic APA sites, suggesting that U2AF contributes to the tight regulation of tissue specific expression. Characteristic 3'ss organization of U2AF(35)-repressed and -activated exons and functional antagonism and synergism of U2AF related proteins PUF60 and CAPERα was also described (Figures 6.16, 6.17, 6.18, 6.19, 6.20, and 6.21). The exon repression and activation was associated with significant shifts of average unpaired probabilities in their canonical BP and PPT regions and further upstream. Finally, exon-centric regulations of genes encoding U2AF35 interaction partners were described and transcripts with distinct responses to U2AF35a and U2AF35b were identified.

Our results indicate that most but not all changes in APA and exon usage in cells depleted of U2AF35 were attributable to the lack of U2AF heterodimer (Figures B.4, B.5, B.6, B.7, and 6.35). They were replicated in cells lacking U2AF65 (Figures 6.19, B.2, B.3, B.16, B.17, and B.18), in agreement with RASL-Seq data submitted during review of this manuscript and showing significant overlap of U2AF35- and U2AF65-induced events [265]. As U2AF65 interacts with the 3' end processing complex, the association of U2AF with the phosphorylated C-terminal domain (CTD) of polII and PRPF19 [266][267] could be important for the U2AF-dependent APA control. U2AF35 depletion was associated with upregulation of

CDC5L (Table 6.1), a key component of the PRPF19-CDC5L complex required for promotion of co-transcriptional splicing [268]. The impaired balance between post- and cotranscriptional splicing in cells lacking U2AF is supported by frequent alterations of weak introns that contained cryptic or alternative exons, often near 3' gene ends (Figures B.10 panel C, and B.20). These 'detained' introns are spliced post-transcriptionally in humans [269]. In *Drosophila* and the mouse, alternative introns are less efficiently spliced co-transcriptionally than constitutive introns and co-transcriptional splicing is less efficient toward 3' gene ends than at upstream gene locations [270]. The link between U2AF and transcription and polII elongation rate [271], which can alter APA choice [272], is further supported by a biased distribution of differentially used start and terminal exons in cells depleted of elongation factors (Table C.10) and significant overlap of genes differentially expressed in U2AF35-, hnRNP C- and AFF4-depleted cells (Figure B.22). AFF4 is a key elongation factor while hnRNP C directly competes with U2AF65 at authentic and cryptic 3'ss [222], in agreement with the observed tendency to antagonize U2AF(35)- repressed exons (Figure 6.19). Altered elongation rates influenced inclusion of ∼15–40% cassette exons on a genome-wide scale and the slow elongation has been associated with shorter introns [273], which were prevalent among U2AF-sensitive events (Figure B.12).

The APA selection was influenced by U2AF in an APA category-dependent manner (Figure 6.11). U2AF appears to activate proximal APA sites if there is no intron between competing APA sites and distal APA sites if recognition of the 5'ss is productive (Figure 6.11 panels A and C). U2AF65 contacts CFIm59 but not CFIm68 [274], which promoted distal APA sites [275]. PolII CTD also interacts with the CFIm subunit PCF11 [276] and several splicing factors, including PRP40, CA150 and PSF [277][278][279]. Significant shifts in cleavage site usage, largely toward proximal sites, were observed for a knockdown of PABPN1 [280], a nuclear protein with high affinity to poly(A) tails, with 43 genes shared between PABPN1- and U2AF35-depleted samples ($P < 10^{-8}$, data not shown). Finally, U2AF-dependent intronic APA sites represented the most frequent APA category (Figure 6.11 panel C), whereas in normal cells intronic APA sites are less frequent than tandem 3'UTR and alternative terminal exons [281], which could be due to widespread binding of U2AF(65) to introns, with >80% tags in intronic sequences [265](69).

Similar to other RNA-binding proteins [246][247][282][283], U2AF(35) can both inhibit and activate splicing (Supplementary Figures S2 B.2, B.3, B.4, B.5, B.6, B.7). Activated and repressed exons had a distinct 3'ss organization and functional regulation by Y-binding proteins (Figures 6.16, 6.17, 6.18, 6.19, 6.20, 6.21, B.15, B.16, B.17, B.18, and B.19). The U2AF(35)-repressed exons were largely stimulated by PUF60 (Figure B.16) and inhibited by CAPERα and also by PTB, in line with longer AGEZs/PPTs previously found for PTB-repressed cassettes [246]. Because PUF60 preferentially contacts uridines [284], the PUF60-U2AF antagonism might be explained by competition for binding to uridine-rich sequences. However, uridine was enriched upstream of both upregulated and downregulated exons (Figure B.15) and also within these exons (Figure B.12 panel B), arguing for the importance of adenine/uridine frequency shifts between optimally located BP and PPT regions (Figure B.15 panel A) and between regions showing higher single-strandedness (Figure 6.21). These changes are likely to alter not only protein binding but also tertiary contacts and folding transitions by helicases and other RNA chaperones. Structural requirements across 3'ss may also help explain atypical responses of some transcripts, such as GSK3B to PUF60 depletion (Figure B.16), and position-dependent alternative splicing activity of a growing number of proteins and their targets [246][285][286][287].

PUF60 contains two RNA recognition motifs (RRMs) and a C-terminal UHM [284]. This UHM binds the N-terminus of SF3b155 at the UHM-ligand motif (ULM) around W200 [288]. The functional PUF60-U2AF antagonism might also result from competition of U2AF35 and PUF60 for the U2AF65 ULM because binding of PUF60 to the U2AF65 ULM can only occur if this motif is not already bound by the U2AF35 UHM [288]. In contrast, PUF60 and U2AF65 can bind to the N terminus of SF3b155 simultaneously and noncompetitively [288]. In addition, U2AF and PUF60 had the opposite effect on BP accessibility and U2AF was not strictly required for splicing of some pre-mRNAs *in vitro* when PUF60 was present [247]. Finally, anti-PUF60 antibodies co-precipitated polII CTD and three components of the general transcription factor TFIIH [289], linking PUF60 to transcription. Unlike PUF60 on U2AF-repressed exons (Figure 6.19), however, slow and fast polII elongation do not usually have opposite effects on the inclusion of a given alternative exon [273].

Our study revealed activation and inhibition of many cryptic exons, both in U2AF-related and unrelated genes. The elevated *U2AF2* mRNA (Figure 6.4) can be explained by repression of a PTC-containing exon in intron 5 in depleted cells (Figure B.20). This exon is highly conserved between mouse and man, has a GC 5'ss and is surrounded by Y-rich sequences. Such cryptic exon activation appeared to be common in other U2AF35 partners (Figure B.20, Table 6.1). Similar to *U2AF2*, *CAPERα* was upregulated upon U2AF depletion, most likely through elimination of a PTC-introducing cryptic exon (Figure B.20). CAPERα showed a strong synergism with U2AF on upregulated exons (Figure 6.19) and has the same domain structure as U2AF65, except for the lack ofULM [290]. In contrast, a 48-nt *CAPERβ* exon activated in ab- cells (Figure B.20) does not present PTCs but is translated only if upstream alternative exons are included in the mRNA. The mRNA expression of *PUF60*, which lacks both the ULM and the RS domain [290], was unaffected, although PUF60 exon 5 appeared to respond positively to the excess of U2AF35b (Figure B.20). This 51-nt cassette exon introduces extra 17 amino acids close to the first RRM of PUF60 and encodes two serines that are phosphorylated [291].

Although positions −3 and +1 relative to U2AF(35)- dependent 3'ss may participate in U2AF35-RNA interactions, with binding preferences favoring cytosine at position −3 and guanine at position +1 (Figure B.9), possibly via ZFs [221], experimental evidence for this interaction is missing. The higher dependency of weaker and proximal alternative 3'ss on U2AF35 was finally seen at the genome-wide level (Figure B.9, Table C.14), resolving previous uncertainties [211][221]. However, U2AF(35)-dependent exons and 3'ss were largely alternatively spliced, which makes it difficult to distinguish characteristic sequence features of cassette exons from direct effects of U2AF35 depletion. For example, alternative splicing and exon skipping was found to correlate positively with the BP-3'ss distance [237]. The distance between 3'ss and the best predicted BP of exons/3'ss upregulated in ab- cells was marginally higher than in controls and these 3'ss had also a higher number of BPs with positive SVM scores (data not shown). Thus, it remains to be tested how BP choice is affected by a lack of U2AF(35) and by the observed positional changes in single-strandedness (Figure 6.21). Interestingly, cooperative interactions of U2AF and SF1 increased the SF1 binding repertoire and SF1 binding was significantly biased toward terminal exons [283].

The higher exogenous expression of U2AF35b over U2AF35a (Figure 6.33) and elevated relative abundance of U2AF35b upon U2AF65 depletion (Figures B.4, B.5, B.6, and B.7) suggest that interactions of each isoform with U2AF65 could be important for U2AF stability, thus contributing to tight regulation of U2AF levels *in vivo* and accurate exon/APA usage. Description of exons with isoform-specific responses to U2AF35 should facilitate characterization of physical interactions of highly conserved U2AF35a and U2AF35b isoforms in future studies. These interactions may involve the extraordinarily long α helix A [292] and are likely to be affected by post-translational modifications of residues encoded by exon 3/Ab as the size difference between U2AF35b and U2AF35a appears to be larger than predicted (Figure 6.3 and Figure B.1).

Although U2AF35 is believed to be a 3'ss recognition factor, many examples of 5'ss usage alterations upon depletion were found (Figure B.10). This may be explained by altered U2AF65-promoted recruitment of U1 snRNP to weak 5'ss [293] but also elongation kinetics of polII. For example, Rsd1, a yeast homologue of CAPERα, bridges interactions between U1 and U2 snRNPs through the RS domain of Prp5 (DDX46) [294], which was repressed in ab- cultures (Tables C.5, C.6, and C.7). Prp5 is required for a transcription elongation checkpoint and a release of stalled polII [295]. Differentially used alternative 5'ss were also found in cells lacking other RNA-binding proteins, including PTB [246], nevertheless a large excess of U2AF(35)- dependent 3'ss over 5'ss in our data set is consistent with the predominant role of this factor in 3'ss recognition.

Finally, genome-wide identification of U2AF(35)- dependent events and our validation panel will provide an important resource for more detailed biochemical and structural studies of 3'ss and APA sites, expanding not only the number of U2AF35-dependent exons but also exons sensitive to other factors involved in 3'ss/BP/APA selection.

Chapter 7

Conclusion

Application of novel techniques such as RNA-Seq for the purpose of enhancing efficacy and efficiency of gene fusion detection in myeloid malignancies is a step forward. While well-developed tools such as karyotyping, RT-PCR, FISH, and Sanger sequencing can be a cost-effective method of initial investigation for the cancer-related gene fusions, it is evident that they fail in some cases. Employing NGS as an additional tool for diagnostics seems to be within our reach, closing the gap in effective fusion identification. Of course, this allows for more appropriate choice of therapy, making drug agent targeting specific gene products such as epidrugs - e.g. inhibitors of lysine-specific demethylase 1 [81] - more viable, and provides patients with better recovery perspectives.

There are still challenges that need to be overcome before RNA-Seq can be considered as a diagnostic tool. Being consistently effective is probably the main issue at the moment. The problem is multi-layered. Firstly, the number of reads obtained per sample seems to increase the efficiency of fusion detection. There are no studies done to date to investigate the relation. Secondly, one, robust analysis pipeline has to be used. While there are multiple ways in which the currently available software can be combined and modified to suit one's needs, the current progress is very much at the research stage, where we are discovering which approach is the most effective one. Until a consensus is reached on which analytical method is the best one, this will persist as an issue not only for RNA-Seq to be used in diagnostics but also for cancer-unrelated research purposes. Thirdly, lack of consistent framework for identification of causal gene fusion amongst detected ones subtracts from the method's ability

to be consistent. Currently, causal fusions can be identified based on previous reports in literature and databases, by investigating function and topography of genes involved, as well as toxicogenomic characteristics if drug reaction information is available. All that is little more than an educated guess, and hampers identification of truly novel events.

Practicality of RNA-Seq application for routine diagnostics is currently not without issues either. The cost of sequencing, while dropping notably as the method is being improved and becoming more popular, is still considerable when compared to other, routine techniques such as karyotyping, RT-PCR, FISH, and Sanger sequencing. However, looking at the trend where the costs have been decreasing from tens of thousands of pounds to over a thousand pounds per sample within the last six years, the future prospects look optimistic. Application of RNA-Seq requires not only sequencing itself, but also computational power. Necessity for NGS-capable computational infrastructure is an interesting issue in modern healthcare, with no previous technology considered for diagnostic use having such high requirements. If RNA-Seq is to be employed as a routine diagnostic tool, a considerable computational power, one that greatly exceeds the current capacity of any healthcare provider in the world, has to be employed. However, with the current development of computing, especially in the direction of utilisation of multiple networked computers such as cloud computing or distributed computing, the perspective is encouraging.

Although challenges undeniably exist, even with its current limitations RNA-Seq is a valuable tool. The development and application of an analysis pipeline for the detection of gene fusions allowed confirmation of previously known events, as described in Chapter 2.3. It proved that RNA-Seq combined with a bioinformatic approach can achieve efficiency equal to and even exceeding that of cytogenetic methods. It functions as a proof that RNA-Seq can be an effective tool when used on its own, that it can function to detect any gene fusions, not just those missed by current techniques. This can mean that in the future, when RNA-Seq is sufficiently developed as a diagnostic tool it can completely supersede current methods, as it offers not only the same but even more.

The same approach to RNA-Seq analysis allowed to uncover multiple previously unknown, novel gene fusions as detailed in Chapter 4. The core of the advantage of RNA-Seq over

other methods - being able to detect events that are missed by cytogenetic techniques. It provides a tremendous prospects for the future development, proving that the technique is worth further research, potentially yielding benefits for the patients. In the cases where the cancerogenous gene fusion was failed to be identified, RNA-Seq can be employed to locate it, providing invaluable information useful for appropriate treatment choice, reducing mortality and morbidity rates.

In the design and construction of RNA-Seq analysis pipeline for gene fusion detection mutliple software packages were considered. From QC through pre-processing, alignment, and finally to fusion determination, there were are multiple options available. As RNA-Seq is still very much a research technique, all the software offered different approaches to tackle similar issues and achieve similar goals. None of the software considered is inherently incorrect in its approach, which can obscure the choice of the most effective tools. Implementation methods as well as optimisation differs between packages, but they are of very little consequence. More important points, such as efficiency and efficacy were considered and, as proven by the results, the chosen packages showed to be functional and able to achieve the goals. Alterations, detailed in Chapter 2.3, were implemented during the work on the pipeline that took into account exon-intron boundary based filtering, in-frame check, read pattern scoring and filtering as well as manual inspection. The changes were proven to raise efficiency of the pipeline achieving the desired results. However, this was done not without issues. The primary one being low read-depth of potential gene fusions necessitating multiple verifications. This was proven to be worthwhile since multiple low-depth events were confirmed by PCR and Sanger sequencing. If the pipeline was optimised in a different manner or default software stringencies were used, those important findings would be missed.

RNA-Seq was employed not only to identify gene fusions but also to investigate expression patterns. Originally, when RNA-Seq was first used, it was with expression investigation as the aim, its applicability for fusion investigation became apparent years later. As such, methodology for the investigation of expression is much better developed. Multiple software packages were implemented to allow investigation of expression from different angles, namely, from the point of differential expression on the scale of a gene and an exon as well as differential splicing event observation and differential usage of splicing alternatives.

Combining multiple software for deep analysis of expression patterns and splicing changes following knockdown of *U2AF1* allowed to uncover novel characteristics of the auxiliary splicing factor. An analysis approach allowing to investigate multiple varieties of alternative splicing events combined with expression investigation at different levels of transcriptome proved to be an effective way to study characteristics of a gene involved in the process of splicing. The importance of the findings in the context of myeloid malignancies is yet to be determined and future experiments can be built upon the work presented in Chapters 3 and 6. The experiment itself can be improved upon by expanding its design from 2x2 and using more replicates. Statistical power of such improvement could potentially alter significance of observed events. Also, controls in the experiment could be improved by introducing negative controls such as scrambled sequence which would allow to control for off-target events.

Further work can be built upon contents of this thesis. While the implementation of gene fusion analysis pipeline along with the alterations made to increase its efficiency is ready to be used for analysis of new samples, more can be done to expand it. While new advances are likely to be made in over the next years, selective implementation of them can only increase efficiency of the analysis. What is more, there are some avenues that one might want to pursue now, such as selection of different software at different steps of analysis to experimentally assess their performance.

In particular, it would be beneficial to perform experiments using the newly-emerging platform, developed by Oxford Nanopore. Since it was proven to be effective in resolving isoforms of a complex gene [59], it opens a window for a very cost-effective analysis of RNA without the need for PCR or material fragmentation. Firstly, the platform would have to be compared in terms of efficiency with the platforms currently dominating in research, such as Illumina platforms. In order to do this, one could re-analyse samples with confirmed gene fusions, such as those described in Chapter 4. Such methodology would not only assess the platform's applicability for the purpose but also potentially uncover new genetic aberrations. This could potentially provide new insights into gene fusion mechanics while raising standards in their investigation.

Not only gene fusion analysis can be expanded, new insight into splicing mechanism opens new routes to pursue the link between splicing and gene fusions. While there is currently no solid proof that these events are linked, preliminary data suggests interesting correlations, encouraging to expand upon.

In conclusion, the work described in this thesis encompasses discovery of new gene fusions and provides valuable new insights into intricate splicing machinery, all of which was achieved using cutting-edge bioinformatic technologies.

Bibliography

[1] A. Karpozilos and N. Pavildis. The treatment of cancer in Greek antiquity. *European Journal of Cancer*, 40(14):2033–2040, 2004.

[2] A. R. David and M. R. Zimmerman. Cancer: an old disease, a new disease or something in between? *Nature Reviews Cancer*, 10:728–733, 2010.

[3] World Health Organisation. Death and DALY estimates for 2004 by cause for WHO Member State. *Online, available at* `http://www.who.int/entity/healthinfo/global\_burden\_disease/` *Accessed on 21/04/2014*, 2004.

[4] D. Hanahan and R. A. Weinberg. The Hallmarks of Cancer. *Cell*, 100:57–70, 2000.

[5] World Health Organisation. *International Classification of Diseases for Oncology, 3rd revision (ICD-O-3)*, volume World Health Organisation, Geneva. 2000.

[6] A. Murati, M. Brecqueville, R. Devillier, M. J. Mozziconacci, V. Gelsi-Boyer, and D. Birnbaum. Myeloid malignancies: mutations, models and management. *BMC Cancer*, 12:304, 2012.

[7] D. L. Stirewalt and J. P. Radich. The role of FLT3 in haematopoietic malignancies. *Nature Reviews Cancer*, 3:650–665, 2003.

[8] A. V. Jones, P. J. Campbell, P. A. Beer, S. Schnittger, A. M. Vannucchi, K. Zoi, and M. J. Percy et al. The JAK2 46/1 haplotype predisposes to MPL-mutated myeloproliferative neoplasms. *Blood*, 115(22):4517–4523, 2010.

[9] J. W. Vardiman, J. Thiele, D. A. Arber, R. D. Brunning, M. J. Borowitz, A. Porwit, and N. L. Harris. The 2008 revision of the World Health Organisation (WHO) classification

of myeloid neoplasms and acute leukaemia: rationale and important changes. *Blood*, 114:937–951, 2009.

[10] F. Mertens F. Mitelman, B. Johansson. The impact of translocations and gene fusions on cancer causation. *Nature Reviews Cancer*, 7(4):233–245, 2007.

[11] P. C. Nowell and D. Hungerford. A minute chromosome in chronic granulocytic leukemia. *Science*, 132(3438):1497, 1960.

[12] J. Gora-Tybor and T. Robak. Targeted drugs in chronic myeloid leukemia. *Current Medicinal Chemistry*, 15(29):3036–3051, 2008.

[13] F. Mitelman, F. Mertens, and B. Johansson. Prevalence estimates of recurrent balanced cytogenetic aberrations and gene fusions in unselected patients with neoplastic disorders. *Genes Chromosomes Cancer*, 43:350–366, 2005.

[14] Cancer Genome Atlas Research Network. Genomic and Epigenomic Landscapes of Adult De Novo Acute Myeloid Leukemia. *New England Journal of Medicine*, 368(22):2059–2074, 2013.

[15] R. C. Lindsey and B. J. Ebert. The biology and clinical impact of genetic lesions in myeloid malignancies. *Blood*, 122(23):3741–3748, 2013.

[16] J. L. Hess. MLL: a histone methyltransferase disrupted in leukemia. *Trends in Molecular Medicine*, 10(10):500–507, 2004.

[17] Z. E. Karanjawala, N. Murphy, D. R. Hinton, C.-L. Hsieh, and M. R. Lieber. Oxygen Metabolism Causes Chromosome Breaks and Is Associated with the Neuronal Apoptosis Observed in DNA Double-Strand Break Repair Mutants. *Current Biology*, 12(5):397–402, 2002.

[18] H. Kurahashi, H. Inagaki, T. Ohye, H. Kogo, T. Kato, and B. S. Emanuel. Palindrome-mediated chromosomal translocations in humans. *DNA Repair*, 5(9-10):1136–1145, 2006.

[19] M. Nambiar and S. C. Raghavan. How does DNA break during chromosomal translocations? *Nucleic Acids Research*, 39(14):5813–5825, 2011.

[20] A. Chase, A. Reiter, L. Burci, G. Cazzaniga, A. Biondi, J. Pickard, I. A. G. Roberts, J. M. Goldman, and N. C. P. Cross. TFG, a target of chromosome translocations in lymphoma and soft tissue tumors, fuses to GPR128 in healthy individuals. *Haematologica*, 95(1):20–26, 2010.

[21] H. Wen, Y. Li, S. N. Malek, Y. C. Kim, J. Xu, P. Chen, and F. Xiao et al. New Fusion Transcripts Identified in Normal Karyotype Acute Myeloid Leukemia. *PLOS ONE*, 7(12):e51203, 2012.

[22] D. T. W. Jones, S. Kocialkowski, L. Liu, D. M. Pearson, L. M. Backlund, K. Ichimura, and V. P. Collins. Tandem Duplication Producing a Novel Oncogenic BRAF Fusion Gene Defines the Majority of Pilocytic Astrocytomas. *Cancer Research*, 68:8673–8677, 2008.

[23] A. Chase, T. Ernst, A. Fiebig, A. Collins, F. Grand, P. Erben, A. Reiter, S. Schreiber, and N. C. P. Cross. TFG, a target of chromosome translocations in lymphoma and soft tissue tumors, fuses to GPR128 in healthy individuals. *Haematologica*, 95(1):20–26, 2010.

[24] D. S. Wechsler, L. D. Engstrom, B. M. Alexander, D. G. Motto, and D. Roulston. A novel chromosomal inversion at 11q23 in infant acute myeloid leukemia fuses MLL to CALM, a gene that encodes a clathrin assembly protein. *Genes Chromosomes Cancer*, 36(1):26–36, 2003.

[25] P. Akiva, A. Toporik, S. Edelheit, Y. Peretz, A. Diber, R. Shemesh, and R. Sorek. Transcription-mediated gene fusion in the human genome. *Genome Research*, 16(1):30–36, 2006.

[26] C. S. Wu, C. Y. Yu, C. Y. Chuang, M. Hsiao, C. F. Kao, H. C. Kuo, and T. J. Chuang. Integrative transcriptome sequencing identifies trans-splicing events with important roles in human embryonic stem cell pluripotency. *Genome Research*, 24:25–36, 2014.

[27] H. Li, J. Wang, G. Mor, and J. Sklar. A neoplastic gene fusion mimics trans-splicing of RNAs in normal human cells. *Science*, 321(5894):1357–1361, 2008.

[28] S. L. McCarron, J. Kelly, N. Coen, S. McCabe, M. Fay, M. O'Dwyer, and P. J. Hayden et al. A novel e8a2 BCR–ABL1 fusion with insertion of RALGPS1 exon 8 in a patient with relapsed Philadelphia chromosome-positive acute lymphoblastic leukemia. *Leukaemia and lymphoma*, 52(5):919–921, 2011.

[29] J Johansson T. Caspersson, L. Zech. Differential banding of alkylating fluorochromes in human chromosomes. *Experimental Cell Research*, 60:315–319, 1970.

[30] L. Pray. Gleevec: the Breakthrough in Cancer Treatment. *Nature Education*, 1(1):37, 2008.

[31] W. A. Bickmore. Karyotype Analysis and Chromosome Banding. *eLS*, 2001.

[32] W. C. Gause and J. Adamovicz. The use of the PCR to quantitate gene expression. *PCR Methods Appliation*, 3(6):S123–S135, 1994.

[33] M. Werner, M. Ewig, A. Nasarek, L. Wilkens, R. von Wasilewski, J. Tchinda, and M. Nolte. Value of fluorescence in situ hybridization for detecting the bcr/abl gene fusion in interphase cells of routine bone marrow specimens. *Diagnostic Molecular Pathology*, 6(5):282–287, 1997.

[34] University of Wisconsin-Madison. Fluorescent in situ hybridisation (FISH). *Online, available at:* `http://www.slh.wisc.edu/clinical/cytogenetics/fish/Accessedon21/04/2014`, Year unknown.

[35] F. Sanger and A. R. Coulson. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of Molecular Biology*, 94(3):441–446, 1975.

[36] K. C. H. Ha, E. Lalonde, L. Li, L. Cavallone, R. Natrajan, M. B. Lambros, and C. Mitsopoulos. Identification of gene fusion transcripts by transcriptome sequencing in BRCA1-mutated breast cancers and cell lines. *BMC Medical Genomics*, 4:75, 2011.

[37] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown. Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science*, 270(5235):467–470, 1995.

[38] R. I. Skotheim, G. O. Thomassen, M. Eken, G. E. Lind, F. Micci, F. R. Ribeiro, and N. Cerveira et al. A universal assay for detection of oncogenic fusion transcripts by oligo microarray analysis. *Molecular Cancer*, 8(DOI:10.1186/1476-4598-8-5), 2009.

[39] C. A. Maher, C. Kumar-Sinha, X. Cao, S. Kalyana-Sunderam, B. Han, X. Jing, and L. Sam et al. Transcriptome sequencing to detect gene fusions in cancer. *Nature*, 458:97–101, 2009.

[40] P. Hogeweg. The roots of bioinformatics in theoretical biology. *PLOS Computational Biology*, 7(DOI: 10.1371/journal.pcbi.1002021), 2011.

[41] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 15(409(6822)):860–921, 2001.

[42] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, and H. O. Smith et al. The sequence of the human genome. *Science*, 16(291(5507)):1304–1351, 2001.

[43] R. F. Service. GENE SEQUENCING: The Race for the $1000 Genome. *Science*, 311(5767):1544–1546, 2006.

[44] E. Pennisi. Will Computers Crash Genomics? *Science*, 331(6018):666–668, 2011.

[45] E. R. Mardis. The $1,000 genome, the $100,000 analysis? *Genome Medicine*, 2:84, 2010.

[46] G. E. Moore. Progress in digital integrated electronics . *International Electron Devices Meeting Procedings*, 21:11–13, 1975.

[47] K. A. Wetterstrand. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). *Online, Available at* `https://www.genome.gov/sequencingcosts/Accessedon21/04/2014`, 2014.

[48] D. R. Bentley, S. Balasubramanian, H. P. Swerdlow, G. P. Smith, J. Milton, C. G. Brown, and K. P. Hall. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature*, 456:53–59, 2008.

[49] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, and P. Peluso et al. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, 323(5910):133–138, 2009.

[50] R. Drmanac, A. B. Spark, M. J. Callow, A. L. Harpen, N. L. Burns, B. G. Kermani, and P. Carnevali et al. Human Genome Sequencing Using Unchained Base Reads on Self-Assembling DNA Nanoarrays. *Science*, 327(5961):78–81, 2009.

[51] J. F. Thompson and K. E. Steinman. Single Molecule Sequencing with a HeliScope Genetic Analysis System. *Current Protocols in Molecular Biology*, 92:7.10.1–7.10.14., 2010.

[52] D. R. Smith, A. R. Quinlan, H. E. Peckham, K. Makowsky, W. Tao, B. Woolf, and L. Shen et al. Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Research*, 18:1638–1642, 2008.

[53] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, and J. Berka et al. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437:376–380, 2005.

[54] N. Rusk. Torrents of sequence. *Nature Methods*, 8:14, 2011.

[55] Max Nisen. How Illumina's Gene Sequencing Technology Could Transform Health Care. *Online, Available at* `http://www.businessinsider.com/illumina-genome-sequencing-growth-2013-10` *Accessed on 21/04/2014*, 2013.

[56] J. L. Weirather, P. T. Afshar, T. A. Clark, E. Tseng, L. S. Powers, J. G. Underwood, and J. Zabner et al. Characterization of fusion genes and the significantly expressed fusion isoforms in breast cancer by hybrid sequencing. *Nucleic Acids Research*, 43(18):e116, 2015.

[57] J. J. Kasianowicz, E. Brandin, D. Branton, and D. W. Deamer. Characterization of individual polynucleotide molecules using a membrane channel. *Proceeding of the National Academy of Sciences*, 93(24):13770–13773, 1996.

[58] K. Judge, S. R. Harris, S. Reuter, J. Parkhill, and S. J. Peacock. Early insights into the potential of the Oxford Nanopore MinION for the detection of antimicrobial resistance genes. *Journal of Antimicrobial Chemotherapy*, 70(9), 2015.

[59] M. T. Bolisetty, G. Rajadinakaran, and B. R. Graveley. Determining exon connectivity in complex mRNAs by nanopore sequencing. *Genome Biology*, 16(204):doi:10.1186/s13059–015–0777–z, 2015.

[60] Illumina. Illumina Sequencing Technology. *Online, Available at: http://res.illumina.com/documents/products/techspotlights/techspotlight_sequencing.pdf Accessed on 21/04/2014*, 2014.

[61] J. B. Li R. Piskol, G.Ramaswami. Reliable identification of genomic variants from RNA-seq data. *American Journal of Human Genetics*, 93(4):641–651, 2013.

[62] F. Ozsolak and P. M. Milos. RNA sequencing: advances, challenges and opportunities. *Nature Reviews Genetics*, 12:87–98, 2011.

[63] A. Roberts, C. Trapnell, J. Donaghey, J. L. Rinn, and L. Pachter. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biology*, 12:R22, 2011.

[64] G. R. Grant, M. H. Farkas, A. D. Pizarro, N. F. Lahens, J. Schug, B. P. Brunk, and C. J. Stoeckert et al. Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM). *Bioinformatics*, 27(18):2518–2528, 2011.

[65] H. Li and N. Homer. A survey of sequence alignment algorithms for next-generation sequencing. *Briefings in Bioinformatics*, 11(5):473–483, 2010.

[66] D. Kim and S. L. Salzber. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biology*, 12:R72, 2011.

[67] A. McPherson, F. Hormozidiari, A. Zayed, R. Giuliany, G. Ha, M. G. Sun, and M. Griffith et al. deFuse: An Algorithm for Gene Fusion Discovery in Tumor RNA-Seq Data. *PLOS Computational Biology*, 7(5), 2011.

[68] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, and H. Pimentel et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, 7:562–578, 2012.

[69] R. W. Francis, K. Thompson-Wicking, K. W. Carter, D. Anderson, U. R. Kees, and A. H. Beesley. FusionFinder: A Software Tool to Identify Expressed Gene Fusion Candidates from RNA-Seq Data. *PLOS One*, 7(8), 2012.

[70] Y. Li, J. Chien, D. I. Smith, and Jian Ma. FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq. *Bioinformatics*, 27(12):1708–1710, 2011.

[71] O. Sakarya, H. Breu, M. Radovich, Y. Chen, Y. N. Wang, C. Barbacioru, and S. Utiramerur et al. RNA-Seq Mapping and Detection of Gene Fusions with a Suffix Array Algorithm. *PLOS Computational Biology*, 8(DOI: 10.1371/journal.pcbi.1002464), 2012.

[72] C. Trapnell, L. Pachter, and S. L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, 2009.

[73] Y. Wu, X. Wang, F. Wu, R. Huang, F. Xue, G. Liang, and M. Tao et al. Transcriptome Profiling of the Cancer, Adjecent Non-Tumor and Distant Normal Tissues from a Colorectal Cancer Patient by Deep Sequencing. *PLOS One*, 7(DOI: 10.1371/journal.pone.0041001), 2011.

[74] C. Steidl, S. P. Shah, B. W. Woolcock, L. Rui, M. Kawahara, P. Farinha, and N. A. Johnson et al. MHC class II transactivator CIITA is a recurrent gene fusion partner in lymphoid cancers. *Nature*, 471:377–381, 2011.

[75] W. J. Kent, C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle, A. M. Zahler, and D. Haussler. The human genome browser at UCSC. *Genome Research*, 12(6):996–1006, 2002.

[76] J. Schwaab, M. Knut, C. Haferlach, G. Metzgeroth, H. P. Horny, A. Chase, and W. J. Tapper. Limited duration of complete remission on ruxolitinib in myeloid neoplasms with PCM1-JAK2 and BCR-JAK2 fusion genes. *Annals of Hematology*, 94(2):233–238, 2015.

[77] R. A. Harvey, P. C. Champe, R. Finkel, L. X. Cubeddu, and M. A. Clark. Lippincott's Illustrated Review: Pharmacology. Lippincott Williams and Wilkins:Chapter 2: Principles of Cancer Chemotherapy, 2008.

[78] C. Knox, V. Law, T. Jewison, P. Liu, S. Ly, A. Frolkis, and A. Pon et al. DrugBank 3.0: a comprehensive resource for 'omics' research on drugs. *Nucleic Acids Research*, 39:D1035–41, 2011.

[79] G. Quezada, L. Kopp, E. Estey, and R. J. Wells. All-trans-retinoic acid and arsenic trioxide as initial therapy for acute promyelocytic leukemia. *Pediatric Blood and Cancer*, 51(1):133–135, 2008.

[80] M. E. R. O'Brien, A. Borthwick, A. Rigg, A. Leary, L. Assersohn, K. Last, and S. Tan et al. Mortality within 30 days of chemotherapy: a clinical governance benchmarking issue for oncology patients. *British Journal of Cancer*, 95(12):1632–1636, 2006.

[81] T. Schenk, W. C. Chen, S. Gollner, L. Howell, L. Jin, K. Hebestreit, and H. U. Klein et al. Inhibition of the LSD1 (KDM1A) demethylase reactivates the all-trans-retinoic acid differentiation pathway in acute myeloid leukemia. *Nature Medicine*, 18(4):605–611, 2012.

[82] H. Fredly, B. T. Gjertsen, and O. Bruserud. Histone deacetylase inhibition in the treatment of acute myeloid leukemia: the effects of valproic acid on leukemic cells, and the clinical and experimental evidence for combining valproic acid with other antileukemic agents. *Clinical Epigenetics*, 5(12), 2013.

[83] X. Thomas. Histone deacetylase inhibition in the treatment of acute myeloid leukemia: the effects of valproic acid on leukemic cells, and the clinical and experimental evidence for combining valproic acid with other antileukemic agents. *Expert Opinion on Drug Discovery*, 7(11):1039–1051, 2012.

[84] A. Y. Leung, C. H. Man, and Y. L. Kwong. FLT3 inhibition: a moving and evolving target in acute myeloid leukaemia. *Leukemia*, 27(2):260–268, 2013.

[85] M. K. Bucci, A. Bevan, and M. Roach. Advances in Radiation Therapy: Conventional to 3D, to IMRT, to 4D, and Beyond. *CA: A Cancer Journal for Clinicians*, 55(2):117–134, 2005.

[86] L. L. Popplewell and S. J. Forman. Is there an upper age limit for bone marrow transplantation? *Nature*, 29(4):277–284, 2002.

[87] F. Crick. On protein synthesis. *Symp. Soc. Exp. Biol. XII*, page 138–163, 1958.

[88] F. Crick. Central Dogma of Molecular Biology. *Nature*, 227:561–563, 1970.

[89] M. C. Wahl, C. L. Will, and R. Luhrmann. The spliceosome:design principles of a dynamic RNP machine. *Cell*, 136:701–718, 2009.

[90] I. Shcherbakova, A. A. Hoskins, L. J. Friedman, V. Serebrov, I. R. Correa Jr, M. Q. Xu, and J. Gelles et al. Alternative spliceosome assembly pathways revealed by single-molecule fluorescence microscopy. *Cell Reports*, 5:151–165, 2013.

[91] P. D. Zamore and M. R. Green. Identification, purification, and biochemical characterization of U2 small nuclear ribonucleoprotein auxiliary factor. *Proceedings of the National Academy of Science*, 86:9243–9247, 1989.

[92] S. Guth, T. O. Tange, E. Kellenberger, and J. Valcarcel. Dual function for U2AF(35) in AG-dependent pre-mRNA splicing. *Molecular and Cellular Biology*, 21:7673–7681, 2001.

[93] Z. Wang, X. Xiao, E. V. Nostrand, and B. Burge. General and Specific Functions of Exonic Splicing Silencers in Splicing Control. *Molecular Cell*, (1):61–70, 2006.

[94] J. Zhu, A. Mayeda, and A. R. Krainer. Exon Identity Established through Differential Antagonism between Exonic Splicing Silencer-Bound hnRNP A1 and Enhancer-Bound SR Proteins. *Molecular Cell*, (6):1351–1361, 2001.

[95] A. J. Ward and T. A. Cooper. The pathology of Splicing. *Journal of Pathology*, 220(2):152–163, 2010.

[96] L. S. Friedman, E. A. Ostermeyer, C. I. Szabo, P. Dowd, E. D. Lynch, S. E. Rowell, and M.-C. King. Confirmation of BRCA1 by analysis of germline mutations linked to breast and ovarian cancer in ten families. *Nature Genetics*, 8:399–404, 1994.

[97] S. H. Lafevre, L. Chauveinc, S. Stoppa-Lyonnet, J. Michon, L. Lumbroso, P. Berthet, and D. Frappaz et al. A T to C mutation in the polypyrimidine tract of the exon 9 splicing site of the RB1 gene responsible for low penetrance hereditary retinoblastoma . *Journal of Medical Genetics*, 39:e21, 2002.

[98] A. N. Brooks, P. S. Choi, L. de Wall, T. Sharifnia, M. Imielinski, G. Saksena, and C. S. Pademallu et al. A Pan-Cancer Analysis of Transcriptome Changes Associated

with Somatic Mutations in U2AF1 Reveals Commonly Altered Splicing Events . *PLOS One*, 9(4):e96437, 2014.

[99] J. qian, D. M. Yao, J. Lin, W. Qian, C. Z. Wang, H. Y. Chai, and J. Yang et al. U2AF1 mutations in Chinese patients with acute myeloid leukemia and myelodysplastic syndrome. *PLoS One*, 7(9):e45760, 2012.

[100] Y. Katz, E. T. Wang, and E. M. Airoldiand C. B. Burge. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*, 7:1009–1015, 2010.

[101] M. Becutti, M. Carrara, F. Cordero, S. Donatelli, and R. A. Calogero. The structure of state-of-art gene fusion-finder algorithms. *Genome Bioinformatics*, 1(1):2, 2013.

[102] Naiara Rodriguez-Ezpeleta, Michael Hackenberg, and Ana M. Aransay. *Bioinformatics for High Throughput Sequencing*. Springer, 2012.

[103] Q. Zhou, X. Su, A. Wang, and K. Ning. QC-Chain: Fast and Holistic Quality Control Method for Next-Generation Sequencing Data. *PLoS ONE*, 8(4):e60234, 2013.

[104] A. M. Bolger na M. Lohse and B. Usadel. Trimmomatic: A flexible read trimming tool for Illumina NGS data. *Bioinformatics*, doi: 10.1093/bioinformatics/btu170, 2014.

[105] B. Ewing and P. Green. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, 8(3):186–194, 1988.

[106] M. F. Polz and C. M. Cavanaugh. Bias in Template-to-Product Ratios in Multitemplate PCR. *Applied and Environmental Microbiology*, 64(10):3724–3730, 1988.

[107] Simon Anders. FastQC. *Online, available at* `http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`, *Accessed on 21/04/2014*, 2010.

[108] R. K. Patel and M. Jain. NGS QC Toolkit: A Toolkit for Quality Control of Next Generation Sequencing Data. *PLoS ONE*, 7(2):e30619, 2012.

[109] X. Yang, D. Liu, J. Wu, J. Zou, X. Xiao, F. Zhao, and B. Zhu. HTQC: a fast quality control toolkit for Illumina sequencing data. *BMC Bioinformatics*, 14:33, 2013.

[110] R. Shmieder and R. Edwards. Quality control and preprocessing of metagenomic datasets. *Bioinformatics*, 27:863–864, 2011.

[111] Illumina. Quality Scores for Next-Generation Sequencing, Technical Report. (Available at: http://res.illumina.com/documents/products/technotes/technote_q-scores.pdf, Accessed on 21/04/2014), 2011.

[112] Ben Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10:R25, 2009.

[113] J.-W. Li, K. Robinson, M. Martin, A. Sjodin, B. Usadel, M. Young, E. C. Olivares, and D. M. Bolser. The SEQanswers wiki: a wiki database of tools for high-throughput sequencing analysis. *Nucleic Acids Res.*, 40:D1313–1317, 2012.

[114] Wikipedia contributors multiple unknown. List of sequence alignment software. 2014. `http://en.wikipedia.org/wiki/List_of_sequence_alignment_software`.

[115] N. A. Fonesca, J. Rung, A. Brazma, and J. C. Marioni. Tools for mapping high-throughput sequencing data. *Bioinformatics*, 28(24):3169–3177, 2012.

[116] N. A. Fonesca, J. Rung, A. Brazma, and J. C. Marioni. HTS mappers. *Online: http://wwwdev.ebi.ac.uk/fg/hts_mappers/*, Accessed on 01/07/2014.

[117] D. M. Church, V. A. Schneider, T. Graves, K. Auger, F. Cunningham, N. Bouk, and H.-C. Chen et al. Modernizing Reference Genome Assemblies. *PLOS Biology*, 9(7):e1001091, 2011.

[118] Editorial. E pluribus unum. *Nature Methods*, 7(5):331, 2010.

[119] L. R. Meyer, Ann. S. Zweig, A. S. Hinrichs, D. Karolchik, R. M. Kuhn, M. Wong, and C. A. Sload et al. The UCSC Genome Browser database: extensions and updates 2013. *Nucleid Acids Research*, 41(D1):D64–D69, 2013.

[120] F. Hsu, W. J. Kent, H. Clawson, R. M. Kuhn, M. Diekhans, and D. Haussler. The UCSC Known Genes. *Bioinformatics*, 22(9):1036–1046, 2006.

[121] Jeffrey A. Martin and Zhong Wang. Next-generation transcriptome assembly. *Nature Reviews Genetics*, 12:671–682, 2011.

[122] Michael Burrows and David J. Wheeler. A block sorting lossless data compression algorithm, Technical Report. *Online, Available at* `http://www.hpl.hp.com/techreports/Compaq-DEC/SRC-RR-124.pdf` *Accessed on 21/04/2014*, 1994.

[123] N. Homer, B. Merriman, and S. F. Nelson. BFAST: An alignment tool for large scale genome resequencing. *PLoS ONE*, 4(11):e7767, 2009.

[124] W. J. Kent. BLAT - the BLAST-like alignment tool. *Genome Res.*, 12(4):656–664, 2002.

[125] Thomas D. Wu and Colin K. Watanabe. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21:1859–1875, 2005.

[126] Novocraft Technologies. Novoalign. *Online:* `http://www.novocraft.com`, Accessed on 01/07/2014.

[127] Wellcome trust Sanger Institute. Smalt. *Online:* `http://www.sanger.ac.uk/resources/software/smalt/`, Accessed on 01/07/2014.

[128] S. Huang, J. Zhang, R. Li, W. Zhang, Z. He, T.-W. Lam, Z. Peng, and S.-M. Yiu. SOAPsplice: Genome-Wide ab initio Detection of Splice Junctions from RNA-Seq Data. *Frontiers in Genetics*, 2:46, 2011.

[129] Z. Ning, A. J. Cox, and J. C. Mullikin. SSAHA: a fast search method for large DNA databases. *Genome Research*, 11(10):1725–1729, 2001.

[130] M. C. Ryan, J. Cleland, R. Kim, W. C. Wong, and J. N. Weinstein. SpliceSeq: A Resource for Analysis and Visualization of RNA-Seq Data on Alternative Splicing and Its Functional Impacts. *Bioinformatics*, 28(18):2385–2387, 2012.

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

[132] J. C. Marioni, C. E. Mason, S. M. Mane, M. Stephens, and Y. Gilad. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research*, 18(9):1509–1517, 2008.

[133] P. Ferragina and G. Manzini. An experimental study of a compressed index. *Information Sciences*, 1-2:13–28, 2001.

[134] U. Manber and G. Myers. Suffix Arrays: A New Method for On-Line String Searches. *SIAM Journal on Computing*, 22(5):935–948, 1993.

[135] F. Abate, A. Acquaviva, G. Paciello, C. Foti, E. Ficarra, A. Ferrarini, and M. Delledonne et al. Bellerophontes: an RNA-Seq data analysis framework for chimeric transcripts discovery based on accurate fusion model. *Bioinformatics*, 28(16):2114–2121, 2012.

[136] M. K. Iyer, A. M. Chinnaiyan, and C. A. Maher. ChimeraScan: a tool for identifying chimeric transcription in sequencing data. *Bioinformatics*, 27(20):2903–2904, 2011.

[137] R. W. Francis, K. Thompson-Wicking, K. W. Carter, D. Anderson, U. R. Kees, and A. H. Beesley. FusionFinder: a software tool to identify expressed gene fusion candidates from RNA-Seq data. *PloS One*, 7(6):e39987, 2012.

[138] Y. Li, J. Chien, D. I. Smith, and J. Ma. FusionHunter: identifying fusion transcripts in cancer using paired-end RNA-seq. *Bioinformatics*, 27(12):1708–1710, 2011.

[139] H. Ge, K. Liu, T. Juan, F. Fang, M. Newman, and W. Hoeck. FusionMap: Detecting fusion genes from next-generation sequencing data at base-pair resolution. *Bioinformatics*, 27(14):1922–1928, 2011.

[140] K. Wang, D. Singh, Z. Zeng, S. J. Coleman, Y. Huang, G. L. Savich, and X. He et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. *Nucleic Acids Research*, 38(18):e178, 2010.

[141] M. Carrara, M. Beccuti, F. Lazzareto, F. Cavallo, F. Cordero, S. Donatelli, and R. A. Calogero. State-of-the-art fusion-finder algorithms sensitivity and specificity. *BioMed Research International*, page 340620, 2013.

[142] J. Wenlong, Q. Kunlong, H. Minghui, S. Pengfei, Z. Quan, Z. Feng, and Y. Yuan et al. SOAPfuse: an algorithm for identifying fusion transcripts from paired-end RNA-Seq data. *Genome Biology*, 14:R12, 2013.

[143] L. Fernandez-Cuesta, R. Sun, R. Menon, J. George, S. Lorenz, L. A. Meza-Zepeda, and M. Peifer et al. Identification of novel fusion genes in lung cancer using breakpoint assembly of transcriptome sequencing data. *Genome Biology*, 16(7), 2015.

[144] K. Chen, J. W. Wallis, C. Kandoth, J. M. Kalicki-Veizer, K. L. Mungall, A. J. Mungall, and S. J. Jones et al. BreakFusion: targeted assembly-based identification of gene fusions in whole transcriptome paired-end sequencing data. *Bioinformatics*, 28(14):1923–1924, 2012.

[145] M. Schueler, M. Munschauer, L. H. Gregersen, A. Finzel, A. Loewer, W. Chen, M. Landthaler, and C. Dietrich. Differential protein occupancy profiling of the mRNA transcriptome. *Genome Biology*, 13(15):R15, 2014.

[146] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[147] J. Z. Levin, M. F. Berger, X. Adiconis, P. Rogov, A. Melnikov, T. Fennell, and C. Nusbaum et al. Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. *Genome Biology*, 10:R115, 2009.

[148] A. Schroeder, O. Mueller, S. Stocker, R. Salowsky, M. Leiber, M. Gassmann, and S. Lightfoot et al. The RIN: an RNA integrity number for assigning integrity values to RNA measurements. *BMC Molecular Biology*, 7:3, 2006.

[149] S. A. Forbes, N. Bindal, S. Bamford, C. Cole, C. Y. Kok, D. Beare, and M. Jia et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Research*, 39:D945–D950, 2010.

[150] G. Stelzer, I. Dalah, T. I. Stein, Y. Satanhower, N. Rosen, N. Nativ, and D. Oz-Levi et al. In-silico Human Genomics with GeneCards. *Human Genomics*, 5(6):709–717, 2011.

[151] H. Edgren, A. Murumagi, S. Kasgapeska, D. Nicorici, V. Hongisto, K. Kleivi, and I. H. Rye et al. Identification of fusion genes in breast cancer by paired-end RNA-sequencing. *Genome Biology*, 12:R6, 2011.

[152] A. P. Davies, C. G. Murphy, R. Johnson, J. M. Lay, K. Lennon-Hopkins, C. Saraceni-Richards, and D. Sciaky et al. The Comparative Toxicogenomics Database: update 2013. *Nucleic Acids Research*, 41(D1):D1104–D1114, 2013.

[153] S. Nacu, W. Yuan, Z. Kan, D. Bhatt, C. S. Rivers, J. Stinson, and B. A. Peters et al. Deep RNA sequencing analysis of readthrough gene fusions in human prostate adenocarcinoma and reference samples. *BMC Medical Genomics*, 4:11, 2011.

[154] K. Yoshida, M. Sanada, Y. Shiraishi, D. Nowak, Y. Nagata, R. Yamamoto, and Y. Sato et al. Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature*, 478:64–69, 2011.

[155] U. Nagalakshimi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder. The Transcriptional Landscape of the Yeast Genome Defined by RNA Sequencing. *Science*, 320(5881):1344–1349, 2008.

[156] Z. Wang and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

[157] L. Wang, Z. Feng, X. Wang, and X. Zhang. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. *Bioinformatics*, 26(1):136–138, 2009.

[158] A. Ratan, Y. Zhang, V. M. Hayes, S. C. Schuster, and W. Miller. Calling SNPs without a reference sequence. *BMC Bioinformatics*, 11:130, 2010.

[159] U. H. Trivedi, T. Cezard, S. Bridgett, A. Montazam, J. Nichols, M. Blaxter, and K. Gharbi. Quality control of next-generation sequencing data without a reference. *Frontiers in Genetics*, 5:111, 2014.

[160] F. M. You, N. Huo, K. R. Deal, Y. Q. Gu, M.-C. Luo, P. E. McGuire, J. Dvorak, and O. D. Anderson. Annotation-based genome-wide SNP discovery in the large and complex Aegilops tauschii genome using next-generation sequencing without a reference genome sequence. *BMC Genomics*, page 59, 2011.

[161] T. J. Hardcastle and K. A. Kelly. baySeq: Empirical Bayesian methods for identifying differential expression in sequence count data. *BMC Bioinformatics*, 11:422, 2010.

[162] P. Glaus, A. Honkela, and M. Rattray. Identifying differentially expressed transcripts from RNA-seq data with biological variation. *Bioinformatics*, 28(13):1721–1728, 2012.

[163] C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, and L. Pachter. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology*, 31:46–53, 2013.

[164] S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biology*, 11:R106, 2010.

[165] N. leng, J. A. Dawson, J. A. Thomson, V. Ruotti, A. I. Rissman, and B. M. G. Smiths et al. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics*, 29(8):1035–1043, 2013.

[166] M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 2010.

[167] S. Tarazona, F. Garcia-Alcalde, J. Dopazo, A. Ferrer, and A. Conesa. Differential expression in RNA-seq: A matter of depth. *Genome Research*, 21:2213–2223, 2011.

[168] M. Griffith, O. L. Griffith, J. Mwenifumbo, R. Goya, A. S. Morrissy, R. D. Morin, and R. Corbett et al. Alternative expression analysis by RNA sequencing. *Nature Methods*, 7:843–847, 2010.

[169] S. Anders, A. Reyes, and W. Huber. Detecting differential usage of exons from RNA-seq data. *Genome Research*, 22(10):2008–2017, 2012.

[170] W. Wang, Z. Qin, Z. Feng, X. Wang, and X. Zhang. Identifying differentially spliced genes from two groups of RNA-seq samples. *Gene*, 518(1):164–170, 2013.

[171] S. Srivastava and L. Chen. A two-parameter generalized Poisson model to improve the analysis of RNA-seq data. *Nucleic Acids Research*, 38(17):e170, 2010.

[172] H. Richard, M. H. Schulz, M. Sultan, A. Nurnberger, S. Schrinner, D. Balzereit, and E. Dagand et al. Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments. *Nucleic Acids Research*, 38(10):e112, 2010.

[173] B. Tian, J. Hu, H. Zhang, and C. S. Lutz. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Research*, 33(1):201–212, 2005.

[174] B. Tian, Z. Pan, and J. Y. Lee. Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing. *Genome Research*, 17(2):156–165, 2007.

[175] Y. Li, X. Rao, W. W. Mattox, C. I. Amos, and B. Liu. RNA-Seq Analysis of Differential Splice Junction Usage and Intron Retentions by DEXSeq. *PLoS One*, 10(9):e0136653, 2015.

[176] T. A. Gruber, A. L. Gedman, J. Zhang, C. S. Koss, S. Marada, H. Q. Ta, and S.-C. Chen et al. An Inv(16)(p13.3q24.3)-Encoded CBFA2T3-GLIS2 Fusion Protein Defines an Aggressive Subtype of Pediatric Acute Megakaryoblastic Leukemia. *Cancer Cell*, 22(5):683–697, 2012.

[177] D. Arora, S. Kothe, M. van den Eijnden, R. H. van Huijsuijnen, F. Heidel, T. Fischer, and S. Scholl et al. Expression of protein-tyrosine phosphatases in Acute Myeloid Leukemia cells: FLT3 ITD sustains high levels of DUSP6 expression. *Cell Communication Signalling*, 10(19), 2012.

[178] Y. Gokmen-Polar, O. D. Cano, K. A. Kesler, P. J. Loehrer, and S. Badve. NUT midline carcinomas in the thymic region. *Modern Pathology*, 27:1649–1656, 2014.

[179] M. Katoh and M. Katoh. Identification and characterization of human TIPARP gene within the CCNL amplicon at human chromosome 3q25.31. *International Journal of Oncology*, 23(2):541–547, 2003.

[180] A. Fernandez-Medarde and E. Santos. Ras in Cancer and Developmental Diseases. *Genes and Cancer*, 2(3):334–358, 2011.

[181] E. Gunduz, L. B. Beder, R. Tamamura, H. Nagatsuka, and N. Nagai. Inhibitor of Growth (ING) Family: An Emerging Molecular Target for Cancer Therapy. *Journal of Hard Tissue Biology*, 17(1):1–10, 2008.

[182] T. Tauchi, K. Miyazawa, G. S. Feng, H. E. Broxmeyer, and K. Toyama. A coiled-coil tetramerization domain of BCR-ABL is essential for the interactions of SH2-containing signal transduction molecules. *Journal of Biological Chemistry*, 272(2):1389–1394, 1997.

[183] E. J. Baxter, L. M. Scott, P. J. Cambell, C. East, N. Fourcoulas, S. Swanton, and G. S. Vassiliou et al. Acquired mutation of the tyrosine kinase JAK2 in human myeloproliferative disorders . *The Lancet*, 365(9464):1054–1061, 2005.

[184] M. M. Elnaggar, S. Agersborg, T. Sahoo, A. Girgin, W. Ma, R. Rakkhit, I. Zorilla, and A. Leal. BCR-JAK2 fusion as a result of a translocation (9;22)(p24;q11.2) in a patient with CML-like myeloproliferative disease. *Molecular Cytogenetics*, 5(1):23, 2012.

[185] H. Wen, Y. Li, S. N. Malek, Y. C. Kim, J. Xu, P. Chen, and F. Xiao et al. New Fusion Transcripts Identified in Normal Karyotype Acute Myeloid Leukemia . *PLOS One*, 7(12):e51203, 2012.

[186] A. V. Jones, S. Kreil, K. Zoi, K. Waghorn, C. Curtis, L. Zhang, and J. Score et al. Widespread occurrence of the JAK2 V617F mutation in chronic myeloproliferative disorders. *Blood*, 106(6):2162–2168, 2005.

[187] S. Verstovsek, R. A. Mesa, J. Gotlib, R. S. Levy, V. Gupta, J. F. DiPersio, and J. V. Catalano et al. A double-blind, placebo-controlled trial of ruxolitinib for myelofibrosis. *The New England Journal of Medicine*, 366(9):799–807, 2012.

[188] A. Reiter, C. Walz, A. Watmore, C. Schoch, I. Blau, B. Schlegelberger, and U. Berger et al. The t(8;9)(p22;p24) is a recurrent abnormality in chronic and acute leukemia that fuses PCM1 to JAK2. *Cancer Research*, 65(7):2662–2667, 2005.

[189] F. Griesinger, H. Henning, F. Hillmer, M. Podleschny, R. Steffens, A. Pies, and B. Wormann et al. A BCR-JAK2 fusion gene as the result of a t(9;22)(p24;q11.2) translocation in a patient with a clinically typical chronic myeloid leukemia. *Genes Chromosomes Cancer*, 44(3):329–333, 2005.

[190] P. Peeters, S. D. Raynaud, J. Cools, I. Wlodarska, J. Grosgeorge, P. Philip, and F. Monpoux et al. Fusion of TEL, the ETS-variant gene 6 (ETV6), to the receptor-associated kinase JAK2 as a result of t(9;12) in a lymphoid and t(9;15;12) in a myeloid leukemia. *Blood*, 90(7):2535–2540, 1997.

[191] C. Walz, N. C. Cross, R. A. Van Etten, and A. Reiter. Comparison of mutated ABL1 and JAK2 as oncogenes and drug targets in myeloproliferative disorders. *Leukemia*, 22(7):1320–1334, 2008.

[192] P. Valent, G. J. Gleich, A. Reiter, F. Roufosse, P. F. Weller, A. Hellmann, and G. Metzgeroth et al. Pathogenesis and classification of eosinophil disorders: a review of recent developments in the field. *Expert Review of Hematology*, 5(2):157–176, 2012.

[193] A. Chase, C. Bryant, J. Score, C. Haferlach, V. Grossmann, J. Schwabb, and W. K. Hofmann et al. Ruxolitinib as potential targeted therapy for patients with JAK2 rearrangements. *Haematologica*, 98(3):404–408, 2013.

[194] E. Lierman, D. Selleslag, S. Smits, J. Billiet, and P. Vandenberghe. Ruxolitinib inhibits transforming JAK2 fusion proteins in vitro and induces complete cytogenetic remission in t(8;9)(p22;p24)/PCM1-JAK2–positive chronic eosinophilic leukemia. *Blood*, 120(7):1529–1531, 2012.

[195] E. Rumi, J. D. Milosevic, I. Casetti, I. Dambruoso, D. Pietra, E. Boveri, and M. Boni. Efficacy of ruxolitinib in chronic eosinophilic leukemia associated with a PCM1-JAK2 fusion gene. *Journal of Clinical Oncology*, 31(17):e269–e271, 2013.

[196] B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology*, 10(3):R25, 2009.

[197] C. Trapnell, L. Pachter, and S. L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, 2009.

[198] K. Daehwan and S. L. Salzberg. TopHat-Fusion: an algorithm for discovery of novel fusion transcripts. *Genome Biology*, 12:R72, 2011.

[199] N. Savage, T. I. George, and J. Gotlib. Myeloid neoplasms associated with eosinophilia and rearrangement of PDGFRA, PDGFRB, and FGFR1: a review. *International Journal of Laboratory Hematology*, 35(5):491–500, 2013.

[200] G. Metzgeroth, J. Schwaab, D. Gosenca, A. Fabarius, C. Haferlach, A. Hochhaus, and N. C. P. Cross et al. Long-term follow-up of treatment with imatinib in eosinophilia-associated myeloid/lymphoid neoplasms with PDGFR rearrangements in blast phase. *Leukemia*, page doi:10.1038/leu.2013.129, 2013.

[201] J. V. Jovanovic, J. Score, K. Waghorn, D. Cilloni, E. Gottardi, G. Metzgeroth, and P. Erben et al. Low-dose imatinib mesylate leads to rapid induction of major molecular responses and achievement of complete molecular remission in FIP1L1-PDGFRA-positive chronic eosinophilic leukemia. *Blood*, 109(11):4635–4640, 2007.

[202] B. Ruskin, P. D. Zamore, and M. R. Green. A factor, U2AF, is required for U2 snRNP binding and splicing complex assembly. *Cell*, 52:207–219, 1988.

[203] D. A. Zorio and T. Blumenthal. Both subunits of U2AF recognize the 3' splice site in Caenorhabditis elegans. *Nature*, 402:835–838, 1999.

[204] L. Merendino, S. Guth, D. Bilbao, C. Martinez, and J. Valcarcel. Inhibition of msl-2 splicing by Sex-lethal reveals interaction between U2AF35 and the 3' splice site AG. *Nature*, 402:8838–841, 1999.

[205] S. Wu, C. M. Romfo, T. W. Nilsen, and M. R. Green. Functional recognition of the 3' splice site AG by the splicing factor U2AF35. *Nature*, 402:832–835, 1999.

[206] T. R. Pacheco, A. Q. Gomes, N. L. Barbosa-Morais, V. Benes, W. Ansorge, M. Wollerton, and C. W. Smith et al. Diversity of vertebrate splicing factor U2AF35: identification of alternatively spliced U2AF1 mRNAs. *Journal of Biological Chemistry*, 279:27039–27049, 2004.

[207] M. Zhang, P. D. Zamore, M. Carmo-Fonesca, A. I. Lamond, and M. R. Green. Cloning and intracellular localization of the U2 small nuclear ribonucleoprotein auxiliary factor small subunit. *Proceedings of the National Academy of Science*, 89:8769–8773, 1992.

[208] E. Birney, S. Kumar, and A. R. Krainer. Analysis of the RNA-recognition motif and RS and RGG domains: conservation in metazoan pre-mRNA splicing factors. *Nucleic Acids Research*, 21:5803–5816, 1993.

[209] D. Z. Rudner, K. S. Breger, and D. C. Rio. Molecular genetic analysis of the heterodimeric splicing factor U2AF: the RS domain on either the large or small Drosophila subunit is dispensable in vivo. *Genes & Development*, 12:1010–1021, 1988.

[210] C. L. Kielkopf, S. Lucke, and M. R. Green. U2AF homology motifs: protein recognition in the RRM world. *Genes & Development*, 18:1513–1526, 2004.

[211] T. R. Pacheco, M. B. Coelho, J. M. Desterro, I. Mollet, and M. Carmo-Fonesca. In vivo requirement of the small subunit of U2AF for recognition of a weak 3' splice site. *Molecular and Cellular Biology*, 26:8183–8190, 2006.

[212] B. B. Wang and V. Brendel. Molecular characterization and phylogeny of U2AF35 homologs in plants. *Plant Physiology*, 140:624–636, 2006.

[213] D. Rudner, K. Kannar, S. Breger, and D. Rio. Mutations in the small subunit of the Drosophila U2AF splicing factors cause lethality and developmental defects. *Proceedings of the National Academy of Science*, 93:10333–10337, 1996.

[214] D. A. Zorio and T. Blumenthal. U2AF35 is encoded by an essential gene clustered in an operon with RRM/cyclophilin in Caenorhabditis elegans. *RNA*, 5:487–494, 1999.

[215] G. Golling, A. Amsterdam, Z. Sun, M. Antonelli, E. Maldonado, W. Chen, and S. Burgess et al. Insertional mutagenesis in zebrafish rapidly identifies genes essential for early vertebrate development. *Nature Genetics*, 31:135–140, 2002.

[216] C. J. Webb, S. Lakhe-Reddy, C. M. Romfo, and J. A. Wise. Analysis of mutant phenotypes and splicing defects demonstrates functional collaboration between the large and small subunits of the essential splicing factor U2AF in vivo. *Molecular and Cellular Biology*, 16:584–596, 2005.

[217] R. Reed. The organization of 3's splice-site sequences in mammalian introns. *Genes & Development*, 3:2113–2123, 1989.

[218] P. Zuo and T. Maniatis. The splicing factor U2AF35 mediates critical protein-protein interactions in constitutive and enhancer-dependent splicing. *Genes & Development*, 10:1356–1368, 1996.

[219] B. R. Graveley, K. J. Hertel, and T. Maniatis. The role of U2AF35 and U2AF65 in enhancer-dependent splicing. *RNA*, 7:806–818, 2001.

[220] T. R. Pacheco, L. F. Moita, A. Q. Gomes, N. Hacohen, and M. Carmo-Fonesca. RNA interference knockdown of hU2AF35 impairs cell cycle progression and modulates alternative splicing of Cdc25 transcripts. *Molecular Biology of the Cell*, 17:4187–4199, 2006.

[221] J. Kralovicova and I. Vorechovsky. Allele-dependent recognition of the 3' splice site of INS intron 1. *Human Genetics*, 128:383–400, 2010.

[222] K. Zarnack, J. Konig, M. Tajnik, I. Martincorena, S. Eustermann, I. Stevant, and A. Reyes et al. Direct competition between hnRNP C and U2AF65 protects the transcriptome from the exonization of Alu elements. *Cell*, 152:453–466, 2013.

[223] Y. Katz, E. T. Wang, E. M. Airoldi, and C. B. Burge. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nature Methods*, 7:1009–1015, 2010.

[224] C. Trapnell, D. G. Hendrickson, M. Sauvageau, L. Goff, J. L. Rinn, and L. Pachter. Differential analysis of gene regulation at transcript resolution with RNA-seq. *Nature Biotechnology*, 31:46–53, 2012.

[225] Z. Luo, C. Lin, E. Guest, A. S. Garrett, N. Mohaghegh, S. Swanson, and S. Marshall et al. The super elongation complex family of RNA polymerase II elongation factors: gene target specificity and transcriptional output. *Molecular and Cellular Biology*, 32:2608–2617, 2012.

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

[227] B. Langmead and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nature Methods*, 9:357–359, 2012.

[228] D. Karolchik, G. P. Barber, J. Casper, H. Clawson, M. S. Cline, M. Diekhans, and T. R. Dreszer et al. The UCSC Genome Browser database: 2014 update. *Nucleic Acids Research*, 42:D764–D770, 2014.

[229] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D. R. Kelley, and H. Pimentel et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols*, 7:562–578, 2012.

[230] W. Huang da, B. T. Sherman, and R. A. Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37:1–13, 2009.

[231] W. Huang da, B. T. Sherman, and R. A. Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4:44–57, 2009.

[232] J. Kralovicova, M. Knut, N. C. P. Cross, and I. Vorechovsky. Identification of U2AF(35)-dependent exons by RNA-Seq reveals a link between $3'$ splice-site organization and activity of U2AF-related proteins . *Nucleic Acids Research*, 43(7):3747–3763, 2015.

[233] J. Kralovicova, T. R. Gaunt, S. Rodriguez, P. J. Wood, I. N. M. Day, and I. Vorechovsky. Variants in the human insulin gene that affect pre-mRNA splicing: is -23HphI a functional single nucleotidepolymorphism at IDDM2? *Diabetes*, 55:260–264, 2006.

[234] M. B. Shapiro and P. Senapathy. RNA splice junctions of different classes of eukaryotes: sequence statistics and functional implications in gene expression. *Nucleic Acids Research*, 15:7155–7174, 1987.

[235] G. Yeo and C. B. Burge. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *Journal of Computational Biology*, 11:377–394, 2004.

[236] I. Vorechovsky. Aberrant 3' splice sites in human disease genes: mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization. *Nucleic Acids Research*, 34:4630–4641, 2006.

[237] E. Buratti, M. C. Chivers, J. Kralovicova, M. Romano, M. Baralle, A. R. Krainer, and I. Vorechovsky. Aberrant 5' splice sites in human disease genes: mutation pattern, nucleotide structure and comparison of computational tools that predict their utilization. *Nucleic Acids Research*, 35:4250–4263, 2007.

[238] A. Corvelo, M. Hallegger, C. W. Smith, and E. Eyras. Genome-wide association between branch point properties and alternative splicing. *PLoS Computational Biology*, 6:e1001016, 2010.

[239] T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, and J. Ren et al. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Research*, 37:W202–W208, 2009.

[240] U. Muchstein, H. Tafer, J. Hackermuller, S. H. Bernhart, P. F. Stadler, and I. L. Hofacker. Thermodynamics of RNA-RNA binding. *Bioinformatics*, 22:1177–1182, 2006.

[241] M. Hiller, R. Pudimat, A. Busch, and R. Backofen. Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. *Nucleic Acids Research*, 34:e117, 2006.

[242] S. Anders, A. Reyes, and W. Huber. Detecting differential usage of exons from RNA-seq data. *Genome Research*, 22:2008–2017, 2012.

[243] S. Lianoglou, V. Garg, J. L. Yang, C. S. Leslie, and C. Mayr. Ubiquitously transcribed genes use alternative polyadenylation to achieve tissue-specific expression. *Genes & Development*, 27:2380–2396, 2013.

[244] H. G. Martinson. An active role for splicing in 3'-end formation. *Wiley Interdisciplinary Reviews: RNA*, 2:459–470, 2011.

[245] A. J. Taggart, A. M. DeSimone, J. S. Shih, M. E. Filloux, and W. G. Fairbrother. Large-scale mapping of branchpoints in human pre-mRNA transcripts in vivo. *Nature Structural & Molecular Biology*, 19:719–721, 2012.

[246] C. W. Smith, T. T. Chu, and B. Nadal-Ginard. Scanning and competition between AGs are involved in 3' splice site selection in mammalian introns. *Molecular & Cellular Biology*, 13:4939–4952, 1993.

[247] M. Llorian, S. Schwartz, T. A. Clark, D. Hollander, L. Y. Tan, R. Spellman, and A. Gordon et al. Position-dependent alternative splicing activity revealed by global profiling of alternative splicing events regulated by PTB. *Nature Structural & Molecular Biology*, 17:1114–1123, 2010.

[248] M. L. Hastings, E. Allemand, D. M. Duelli, M. P. Myers, and A. R. Krainer. Control of pre-mRNA splicing by the general splicing factors PUF60 and U2AF. *PLoS ONE*, 2:e538, 2007.

[249] J. P. Tavanez, T. Madl, H. Kooshapur, M. Sattler, and J. Varcarcel. hnRNP A1 proofreads 3' splice site recognition by U2AF. *Molecular Cell*, 45:314–329, 2012.

[250] L. M. Soares, K. Zanier, C. Mackereth, M. Sattler, and J. Valcarcel. Intron removal requires proofreading of U2AF/3' splice site recognition by DEK. *Science*, 312:1961–1965, 2006.

[251] A. Watakabe, K. Inoue, H. Sakamoto, and Y. Shimura. A secondary structure at the 3' splice site affects the in vitro splicing reaction of mouse immunoglobulin mu chain pre-mRNAs. *Nucleic Acids Research*, 17:8159–8169, 1989.

[252] S. Jacquenet, D. Ropers, P. S. Bilodeau, L. Damier, A. Mougin, C. M. Stoltzfus, and C. Branlant. Conserved stem-loop structures in the HIV-1 RNA region containing the A3 3' splice site and its cis-regulatory element: possible involvement in RNA splicing. *Nucleic Acids Research*, 29:464–478, 2001.

[253] S. Arning, P. Gruter, G. Bilbe, and A. Kramer. Mammalian splicing factor SF1 is encoded by variant cDNAs and binds to RNA. *RNA*, 2:794–810, 1996.

[254] FBP WW domains and the Abl SH3 domain bind to a specific class of proline-rich ligands. *EMBO Journal*, 16:2376–2383, 1997.

[255] E. M. Makarov, N. Owen, A. Bottrill, and O. V. Mararova. Functional mammalian spliceosomal complex E contains SMN complex proteins in addition to U1 and U2 snRNPs. *Nucleic Acids Research*, 40:2639–2652, 2011.

[256] C. Gooding and C. W. Smith. Tropomyosin exons as models for alternative splicing. *Advances in Experimental Medicine and Biology*, 644:27–42, 2008.

[257] R. Maytum, F. Bathe, M. Konrad, and M. A. Geeves. Tropomyosin exon 6b is troponin-specific and required for correct acto-myosin regulation. *Journal of Biological Chemistry*, 279:18203–18209, 2004.

[258] B. Vrhovski, G. Schevzov, S. Dingle, J. L. Lessard, P. Gunning, and R. P. Weinberger. Tropomyosin isoforms from the gamma gene differing at the C-terminus are spatially and developmentally regulated in the brain. *Journal of Neuroscience Research*, 72:373–383, 2003.

[259] A. Hegele, A. Kamburov, A. Grossman, C. Sourlis, S. Wowro, M. Weimann, and C. L. Will et al. Dynamic protein-protein interaction wiring of the human spliceosome. *Molecular Cell*, 45:567–580, 2012.

[260] P. Ajuh, B. Kuster, K. Panov, J. C. Zomerdijk, M. Mann, and A. I. Lamond. Functional analysis of the human CDC5L complex and identification of its components by mass spectrometry. *EMBO Journal*, 19:6569–6581, 2000.

[261] F. Caspary and B. Seraphin. The yeast U2A'/U2B complex is required for pre-spliceosome formation. *EMBO Journal*, 17:6348–6358, 1998.

[262] J. Shepard, M. Reick, S. Olson, and B. R. Graveley. Characterization of U2AF(26), a splicing factor related to U2AF(35). *Molecular & Cellular Biology*, 22:221–230, 2002.

[263] A. Di Nardo, R. Gareus, D. Kwiatkowski, and W. Witke. Alternative splicing of the mouse profilin II gene generates functionally different profilin isoforms. *Journal of Cell Science*, 113:3795–3803, 2000.

[264] A. Lambrechts, A. Braun, V. Jonckheere, A. Aszodi, L. M. Lanier, J. Robbens, and U. Van Colen et al. Profilin II is alternatively spliced, resulting in profilin isoforms

that are differentially expressed and have distinct biochemical properties. *Molecular & Cellular Biology*, 20:8209–8219, 2000.

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

[266] C. Shao, B. Yang, T. Wu, J. Huang, P. Tang, Y. Zhou, and J. Zhou et al. Mechanisms for U2AF to define 3' splice sites and regulate alternative splicing in the human genome. *Nature Structural & Molecular Biology*, 21:997–1005, 2014.

[267] S. Millevoi, F. Geraghty, B. Idowu, J. L. Tam, M. Antoniou, and S. Vagner. A novel function for the U2AF 65 splicing factor in promoting pre-mRNA 3'-end processing. *EMBO Reports*, 3:869–874, 2002.

[268] C. J. David, A. R. Boyne, S. R. Millhouse, and J. L. Manley. The RNA polymerase II C-terminal domain promotes splicing activation through recruitment of a U2AF65-Prp19 complex. *Genes & Development*, 25:972–983, 2011.

[269] C. Girard, C. L. Will, J. Peng, E. M. Makarov, B. Kastner, I. Lemm, and H. Urlaub et al. Post-transcriptional spliceosomes are retained in nuclear speckles until splicing completion. *Nature Communications*, 3:994, 2012.

[270] P. L. Boutz, A. Bhutkar, and P. A. Sharp. Detained introns are novel, widespread class of post-transcriptionally spliced introns. *Genes & Development*, 29:63–80, 2015.

[271] Y. L. Khodor, J. S. Menet, M. Tolan, and M. Rosbash. Cotranscriptional splicing efficiency differs dramatically between Drosophila and mouse. *RNA*, 18:2174–2186, 2012.

[272] A. Ujvari and D. S. Luse. Newly Initiated RNA encounters a factor involved in splicing immediately upon emerging from within RNA polymerase II. *Journal of Biological Chemistry*, 279:49773–49779, 2004.

[273] P. A. Pinto, T. Henriques, M. O. Freitas, T. Martins, R. G. Domingues, P. S. Wyrzykowska, and P. A. Coelho et al. RNA polymerase II kinetics in polo polyadenylation signal selection. *EMBO Journal*, 30:2431–2444, 2011.

[274] N. Fong, H. Kim, Y. Zhou, J. Xiong, J. Qiu, T. Saldi, and K. Diener et al. Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. *Genes & Development*, 28:2663–2676, 2014.

[275] S. Millevoi, C. Loulergue, S. Dettwiler, S. Z. Karaa, W. Keller, M. Antoniou, and S. Vagner. An interaction between U2AF65 and CF Im links the slicing and 3' end processing machineries. *EMBO Journal*, 25:4854–4864, 2006.

[276] G. Martin, A. R. Gruber, W. Keller, and M. Zavolan. Genome-wide analysis of pre-mRNA 3' end processing reveals a decisive role of human cleavage factor I in the regulation of 3' UTR length. *Cell Reports*, 1:753–763, 2012.

[277] D. D. Licatalosi, G. Geiger, M. Minet, S. Schroeder, K. Cilli, J. B. McNeil, and D. L. Bentley. Functional interaction of yeast pre-mRNA 3' end processing factors with RNA polymerase II. *Molecular Cell*, 9:1101–1111, 2002.

[278] D. P. Morris and A. L. Greenleaf. The splicing factor, Prp40, binds the phosphorylated carboxyl-terminal domain of RNA polymerase II. *Journal of Biological Chemistry*, 275:39935–39943, 2000.

[279] A. C. Goldstrohm, T. R. Albrecht, C. Sune, M. T. Bedford, and M. A. Garcia-Blanco. The transcription elongation factor CA150 interacts with RNA polymerase II and the pre-mRNA splicing factor SF1. *Molecular and Cellular Biology*, 21:7617–7628, 2001.

[280] A. Emili, M. Shales, S. McCracken, W. Xie, P. W. Tucker, R. Kobayashi, and B. J. Blencowe et al. Splicing and transcription-associated proteins PSF and p54nrb/nonO bind to the RNA polymerase II CTD. *RNA*, 8:1102–1111, 2002.

[281] M. Jenal, R. Elkon, F. Loayza-Puch, G. van Haaften, U. Kuhn, F. M. Menzies, and J. A. Oude Vrielink et al. The poly(A)-binding protein nuclear 1 suppresses alternative cleavage and polyadenylation sites. *Cell*, 149:538–553, 2012.

[282] R. Elkon, A. P. Ugalde, and R. Agami. Alternative cleavage and polyadenylation: extent, regulation and function. *Nature Reviews: Genetics*, 14:496–506, 2013.

[283] A. Kanopka, O. Muhlemann, and G. Akusjarvi. Inhibition by SR proteins of splicing of a regulated adenovirus pre-mRNA. *Nature*, 381:535–538, 1996.

[284] M. Corioni, N. Antih, G. Tanackovic, M. Zavolan, and A. Kramer. Analysis of in situ pre-mRNA targets of human splicing factor SF1 reveals a function in alternative splicing. *Nucleic Acids Research*, 39:1868–1879, 2010.

[285] P. S. Page-McCaw, K. Amonlirdviman, and P. A. Sharp. PUF60: a novel U2AF65-related splicing activity. *RNA*, 5:1548–1560, 1999.

[286] J. R. Tollervey, T. Curk, B. Rogelj, M. Briese, M. Cerada, M. Kayikci, and J. Konig et al. Characterizing the RNA targets and position-dependent splicing regulation by TDP-43. *Nature Neuroscience*, 14:452–458, 2011.

[287] J. Kralovicova and I. Vorechovsky. Position-dependent repression and promotion of DQB1 intron 3 splicing by GGGG motifs. *Journal of Immunology*, 176:2381–2388, 2006.

[288] X. Zhou, W. Wu, H. Li, Y. Cheng, N. Wei, J. Zong, and X. Feng et al. Transcriptome analysis of alternative splicing events regulated by SRSF10 reveals position-dependent splicing modulation. *Nucleic Acids Research*, 42:4019–4030, 2014.

[289] L. Corsini, M. Hothorn, G. Stier, V. Rybin, K. Scheffzek, T. J. Gibson, and M. Sattler. Dimerization and protein binding specificity of the U2AF homology motif of the splicing factor PUF60. *Journal of Biological Chemistry*, 284:630–639, 2009.

[290] J. Liu, L. He, I. Collins, H. Ge, D. Libutti, J. Li, and J.-M. Egly et al. The FGB interacting repressor targets TFIIH to inhibits activated transcription. *Molecular Cell*, 5:331–341, 2000.

[291] D. H. Dowhan, E. P. Hong, D. Auboeuf, A. P. Dennis, M. M. Wilson, S. M. Berget, and B. W. O'Malley. Steroid hormone receptor coactivation and alternative RNA splicing by U2AF65-related proteins CAPERalpha and CAPERbeta. *Molecular Cell*, 17:429–439, 2005.

[292] J. V. Olsen, M. Vermeulen, A. Santamaria, C. Kumar, M. L. Miller, L. J. Jensen, and F. Gnad et al. Quantitative phosphoproteomics reveals widespread full phosphorylation site occupancy during mitosis. *Science Signalling*, 3:ra3, 2010.

[293] C. L. Kielkopf, N. A. Rodionova, M. R. Green, and S. K. Burley. A novel peptide recognition mode revealed by the X-ray structure of a core U2AF35/U2AF65 heterodimer. *Cell*, 106:595–605, 2001.

[294] P. Forch, L. Merendino, C. Martinez, and J. Valcarcel. U2 small nuclear ribonucleoprotein particle (snRNP) auxiliary factor of 65 kDa, U2AF65, can promote U1 snRNP recruitment to 5' splice sites. *Biochemical Journal*, 372:235–240, 2003.

[295] W. Shao, H.-S. Kim, Y. Cao, Y.-Z. Xu, and C. C. Query. A U1-U2 snRNP interaction network during intron definition. *Molecular and Cellular Biology*, 32:470–478, 2012.

[296] K. T. Chathoth, J. D. Barrass5, S. Webb, and J. D. Beggs. A splicing-dependent transcriptional checkpoint associated with prespliceosome formation. *Molecular Cell*, 53:779–790, 2014.

Appendices

Appendix A

Classification of myeloid malignancies

Myeloproliferative neoplasms (MPN)

Chronic myelogenous leukemia, *BCR-ABL1*–positive

Chronic neutrophilic leukemia

Polycythemia vera

Primary myelofibrosis

Essential thrombocythemia

Chronic eosinophilic leukemia, not otherwise specified

Mastocytosis

Myeloproliferative neoplasms, unclassifiable

Myeloid and lymphoid neoplasms associated with eosinophilia and abnormalities of *PDGFRA*, *PDGFRB* or *FGFR1*

Myeloid and lymphoid neoplasms associated with *PDGFRA* rearrangement

Myeloid neoplasms associated with *PDGFRB* rearrangement

Myeloid and lymphoid neoplasms associated with *FGFR1* abnormalities

Myelodysplastic/myeloproliferative neoplasms (MDS/MPN)

Chronic myelomonocytic leukemia

Atypical chronic myeloid leukemia, *BCR-ABL1*–negative

Juvenile myelomonocytic leukemia

Myelodysplastic/myeloproliferative neoplasm, unclassifiable

Provisional entity: refractory anemia with ring sideroblasts and thrombocytosis

Myelodysplastic syndrome (MDS)

Refractory cytopenia with unilineage dysplasia

Refractory anemia

Refractory neutropenia

Refractory thrombocytopenia

Refractory anemia with ring sideroblasts

Refractory cytopenia with multilineage dysplasia

Refractory anemia with excess blasts

Myelodysplastic syndrome with isolated del(5q)

Myelodysplastic syndrome, unclassifiable

Childhood myelodysplastic syndrome

Provisional entity: refractory cytopenia of childhood

Acute myeloid leukemia and related neoplasms

Acute myeloid leukemia with recurrent genetic abnormalities

AML with t(8;21)(q22;q22); *RUNX1-RUNX1T1*

AML with inv(16)(p13.1q22) or t(16;16)(p13.1;q22); *CBFB-MYH11*

APL with t(15;17)(q22;q12); *PML-RARA*

AML with t(9;11)(p22;q23); *MLLT3-MLL*

AML with t(6;9)(p23;q34); *DEK-NUP214*

AML with inv(3)(q21q26.2) or t(3;3)(q21;q26.2); *RPN1-EVI1*

AML (megakaryoblastic) with t(1;22)(p13;q13); *RBM15-MKL1*

Provisional entity: AML with mutated *NPM1*

Provisional entity: AML with mutated *CEBPA*

Acute myeloid leukemia with myelodysplasia-related changes

Therapy-related myeloid neoplasms

Acute myeloid leukemia, not otherwise specified

AML with minimal differentiation

AML without maturation

AML with maturation

Acute myelomonocytic leukemia

Acute monoblastic/monocytic leukemia

Acute erythroid leukemia

Pure erythroid leukemia

Erythroleukemia, erythroid/myeloid

Acute megakaryoblastic leukemia

Acute basophilic leukemia

Acute panmyelosis with myelofibrosis

Myeloid sarcoma

Myeloid proliferations related to Down syndrome

Transient abnormal myelopoiesis

Myeloid leukemia associated with Down syndrome

Blastic plasmacytoid dendritic cell neoplasm

Guidelines for using the revised WHO classification of myeloid neoplasms

Specimen requirements

Peripheral blood (PB) and bone marrow (BM) specimens collected prior to any definitive therapy.

PB and cellular BM aspirate smears and/or touch preparations stained with Wright-Giemsa or similar stain.

BM biopsy, at least 1.5 cm in length and at right angles to the cortical bone, is recommended for all cases if feasible.

BM specimens for complete cytogenetic analysis and, when indicated, for flow cytometry, with an additional specimen cryopreserved for molecular genetic studies. The latter studies should be performed based on initial karyotypic, clinical, morphologic, and immunophenotypic findings.

Assessment of blasts

Blast percentage in PB and BM is determined by visual inspection.

Myeloblasts, monoblasts, promonocytes, megakaryoblasts (but not dysplastic megakaryocytes) are counted as blasts when summing blast percentage for diagnosis of AML or blast transformation; count abnormal promyelocytes as blast equivalents in APL.

Proerythroblasts are not counted as blasts except in rare instances of pure acute erythroleukemia.

Flow cytometric assessment of CD34 cells is not recommended as a substitute for visual inspection; not all blasts express CD34, and artifacts introduced by specimen processing may result in erroneous estimates.

If the aspirate is poor and/or marrow fibrosis is present, IHC on biopsy sections for CD34 may be informative if blasts are CD34 .

Assessment of blast lineage

Multiparameter flow cytometry (at least 3 colors) is recommended; panel should be sufficient to determine lineage as well as aberrant antigen profile of neoplastic population.

Cytochemistry, such as myeloperoxidase or nonspecific esterase, may be helpful, particularly in AML, NOS, but it is not essential in all cases.

IHC on biopsy may be helpful; many antibodies are now available for recognition of myeloid and lymphoid antigens.

Assessment of genetic features

Complete cytogenetic analysis from BM at initial diagnosis when possible.

Additional studies, such as FISH, RT-PCR, mutational status, should be guided by clinical, laboratory, and morphologic information.

Mutational studies for mutated *NPM1* , *CEBPA* , and *FLT3* are recommended in all cytogenetically normal AML; mutated *JAK2* should be sought in *BCR-ABL1* –negative MPN, and mutational analysis for *KIT* , *NRAS* , *PTNP11* , etc, should be performed as clinically indicated.

Appendix B

Supplementary Figures

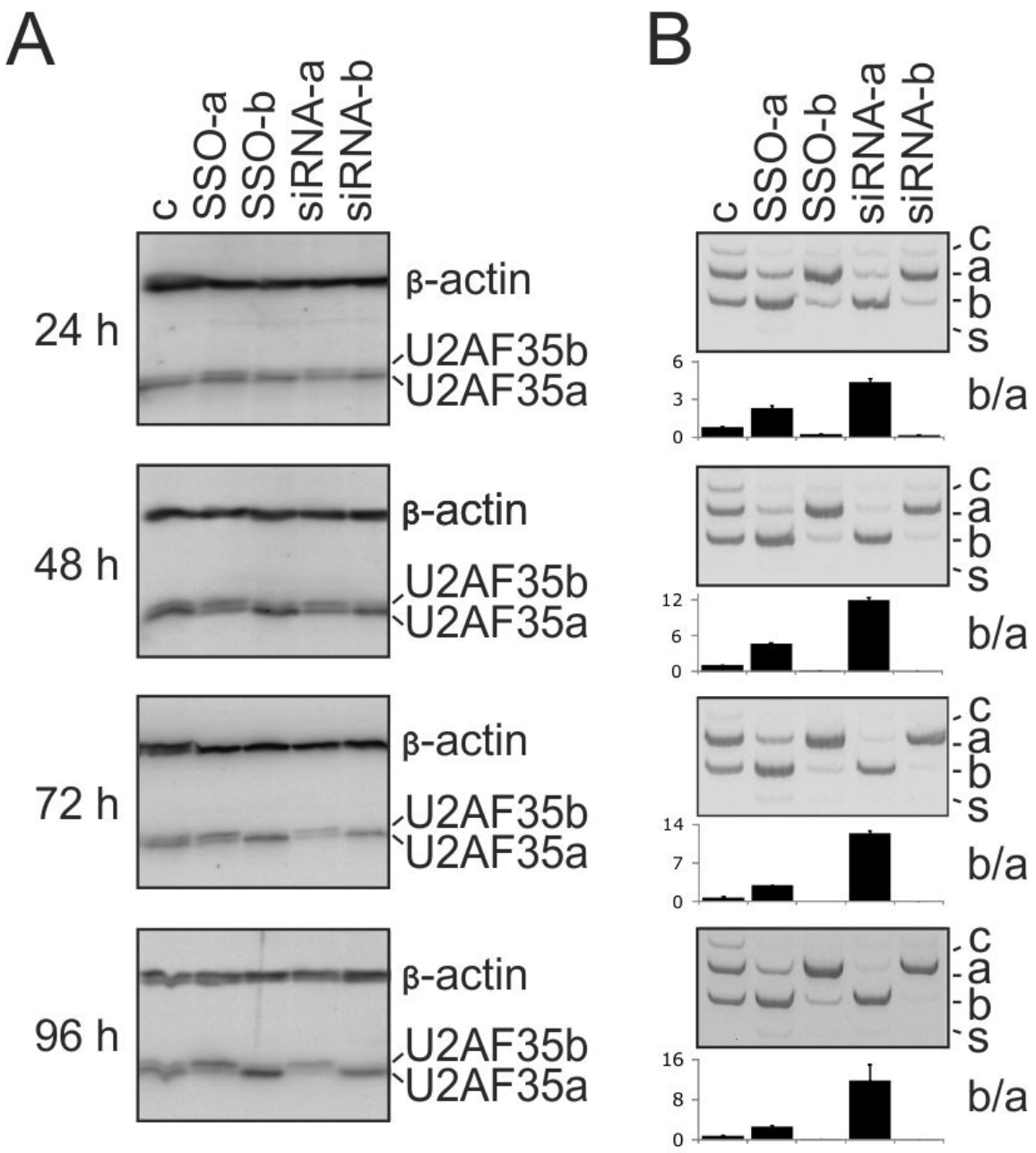

Figure B.1: Time-course transfection experiment with splice-switching oligonuclotides (SSOs) and siRNAs targeting U2AF35a and U2AF35b isoforms (top). Time between the first hit and cell lysis is shown to the left; c, control. **A**, Immunoblots with antibodies against U2AF35 and β-actin. Final concentrations of SSOs and siRNAs were 60 nM. siRNAs were described previously [220][211][221]; sequences of SSOs are in Supplementary Materials in [296]. **B**, *U2AF1* splice products detected by RT-PCR; b/a ratio of *U1AF1b* and *U2AF1a*; c, inclusion of both alternative exons; s, skipping of both exons. Error bars denote SDs of duplicate transfections.

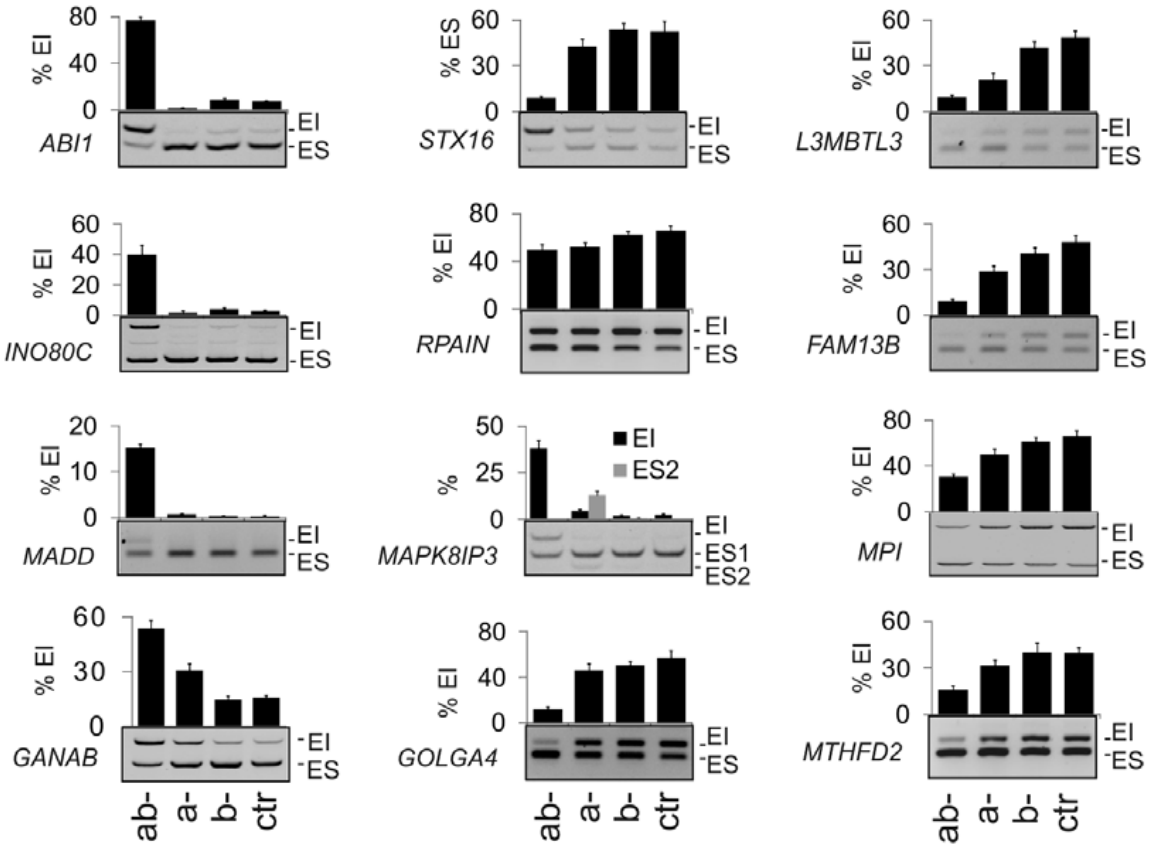

Figure B.2: Examples of RNA processing defects detected by DEXSeq in cultures depleted of U2AF35 and its isoforms. Gene symbols are shown to the left, RNA products to the right. EI, exon inclusion, ES, exon skipping. Error bars are SDs of two replicates. PCR primers are in Supplementary Materials in [296]. Corresponding immunoblots are shown in Figures 6.3, 6.15, and FPKMs in Figures 6.4, 6.5, and 6.6.

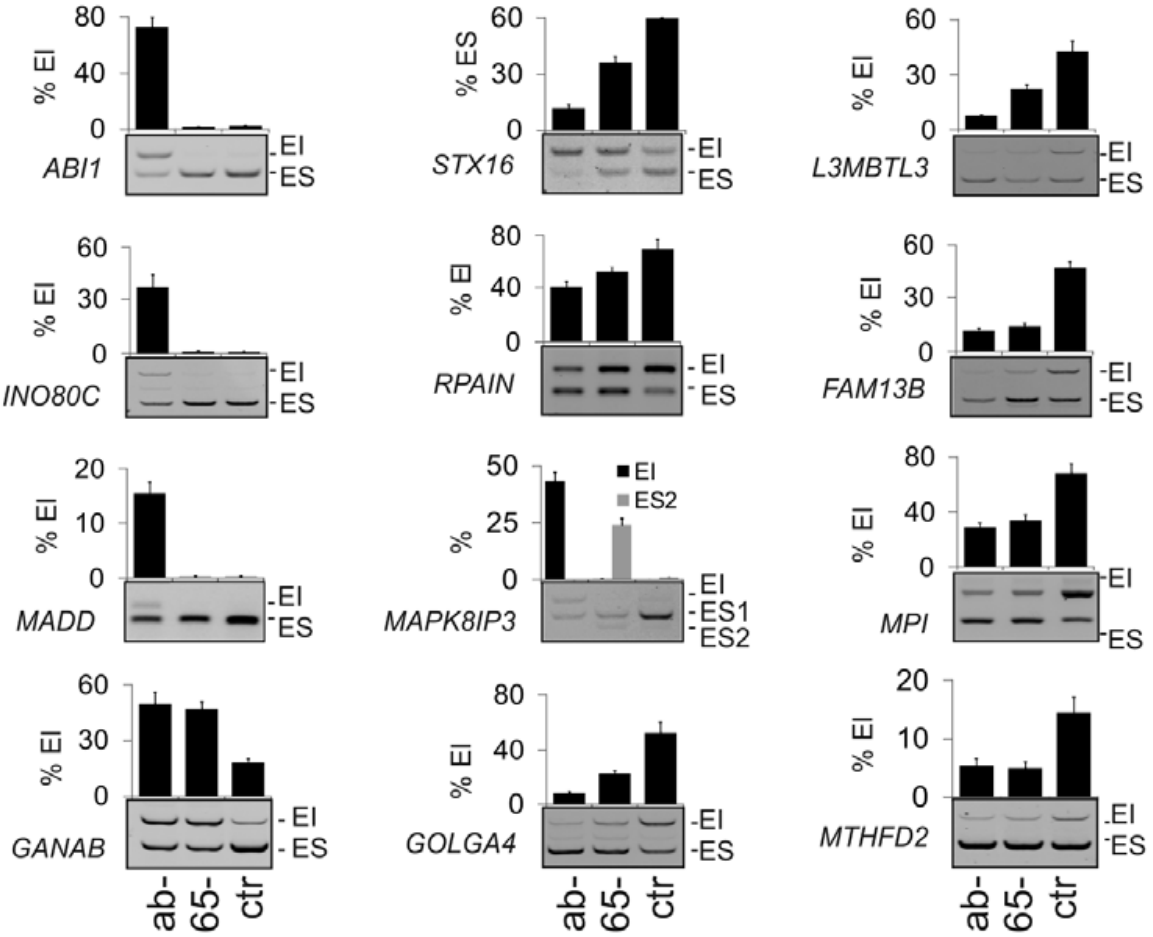

Figure B.3: Examples of RNA processing defects detected by DEXSeq in cultures depleted of U2AF35 and of U2AF65. Gene symbols are shown to the left, RNA products to the right. EI, exon inclusion, ES, exon skipping. Error bars are SDs of two replicates. PCR primers are in Supplementary Materials in [296]. Corresponding immunoblots are shown in Figures 6.3, 6.15, and FPKMs in Figures 6.4, 6.5, and 6.6.

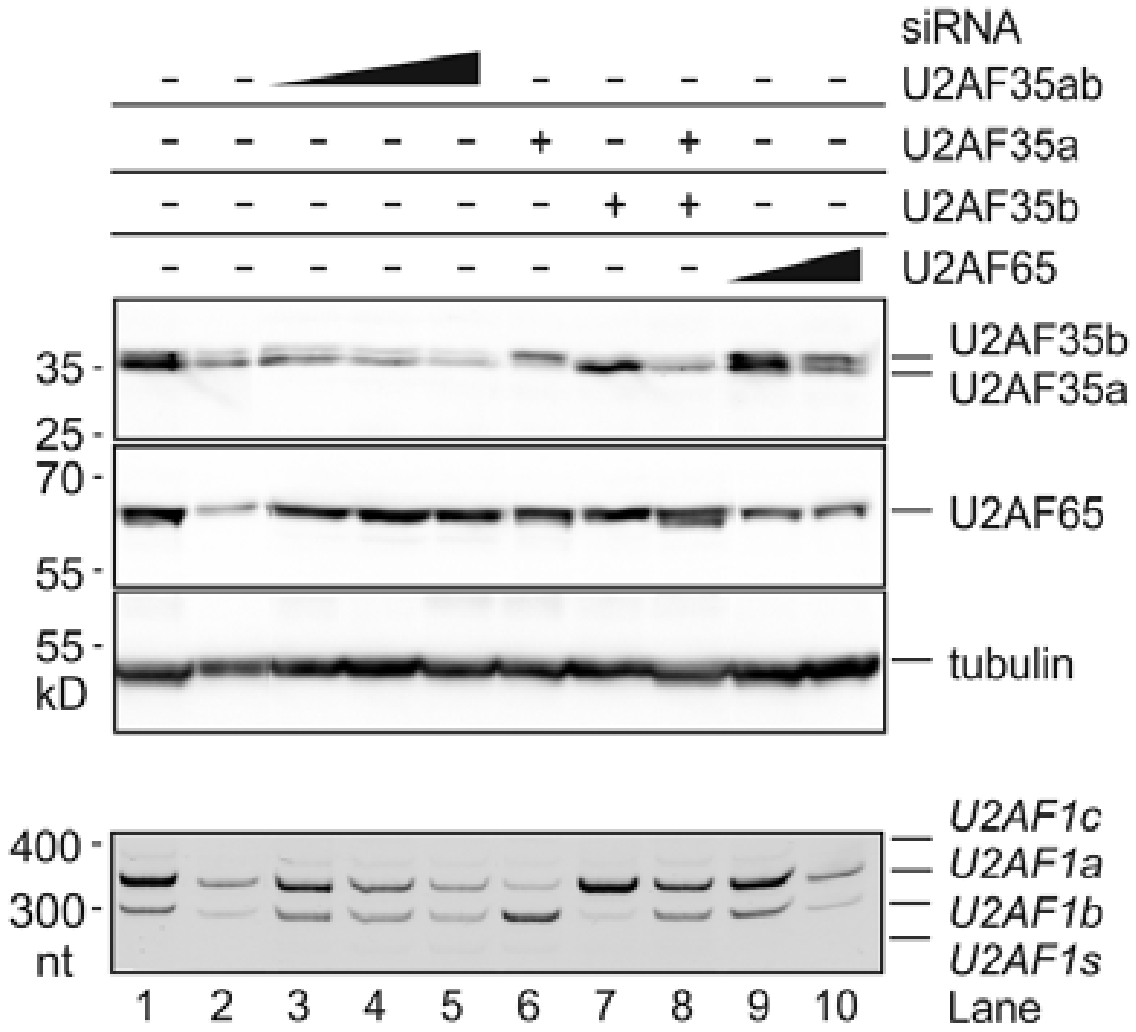

Figure B.4: Immunoblots of a transfection experiment with the indicated siRNAs (upper panel) and *HinfI*-digested RT-PCR of *U2AF1* transcripts (lower panel). Final concentration of U2AF35ab- siRNA was 3, 10 and 30 nM. Final concentration of U2AF65 siRNA was 30 and 50 nM. Isoform-specific siRNAs were added to a final concentration of 50nM. Lane 2 contains 25% of the control lysate/PCR product.

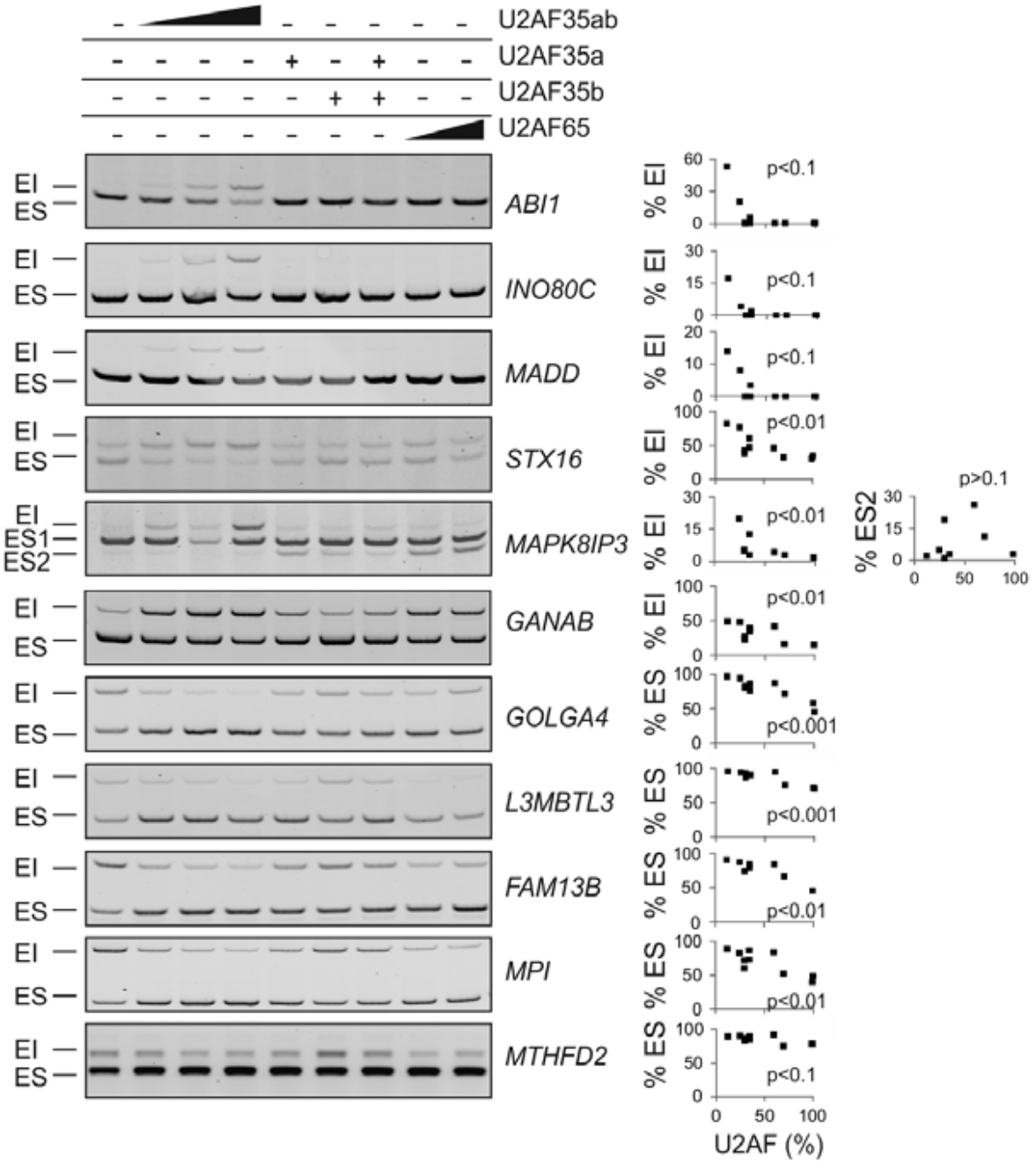

Figure B.5: U2AF(35)-dependent transcripts (left panel) and correlation between U2AF heterodimer levels and exon inclusion (EI) or skipping (ES). ES1/ES2, exon skipping of one/two exons. P-values for Pearson correlation coefficients were derived from t-tests.

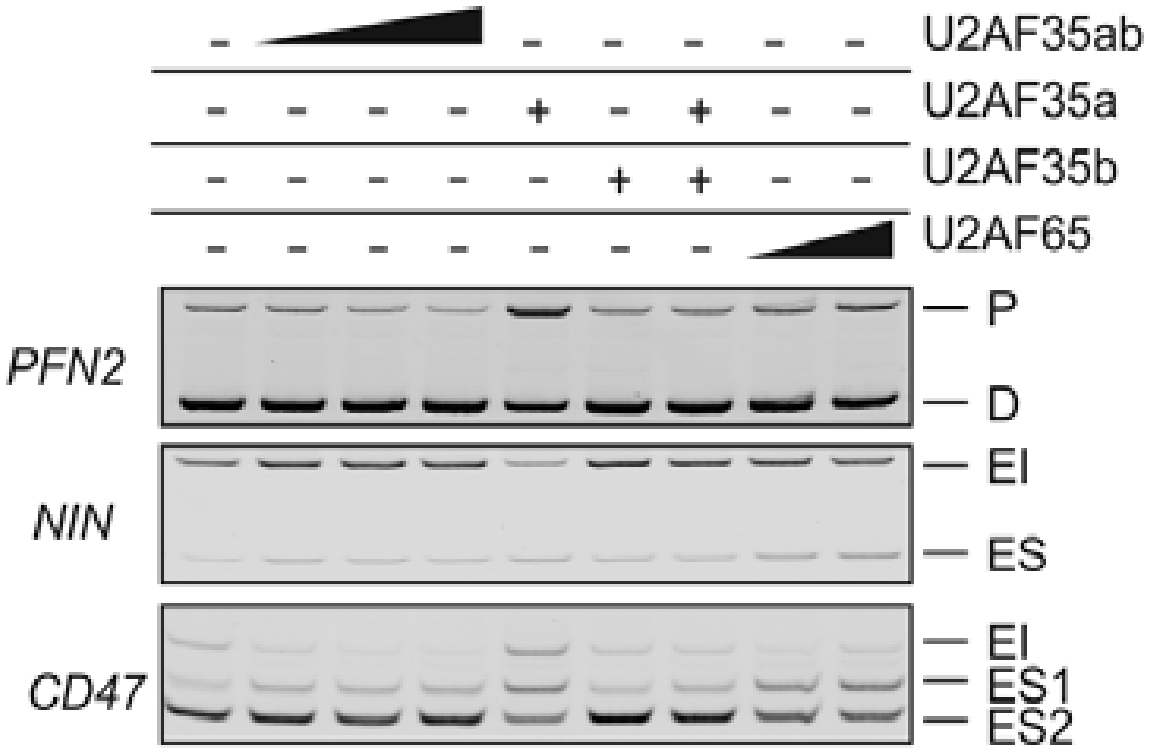

Figure B.6: Transcripts with isoform-specific responses amplified by RT-PCR from the same transfection experiment. Correlation of their exon usage with U2AF is shown in Figure 6.35.

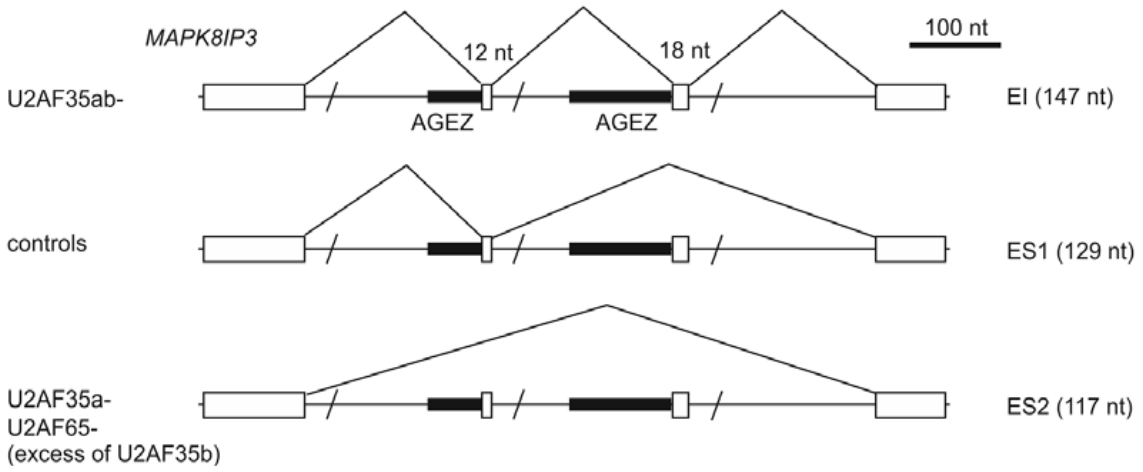

Figure B.7: Summary of *MAPK8IP3* exon usage. The size of RNA products is to the left. AGEZ, AG exclusion zone.

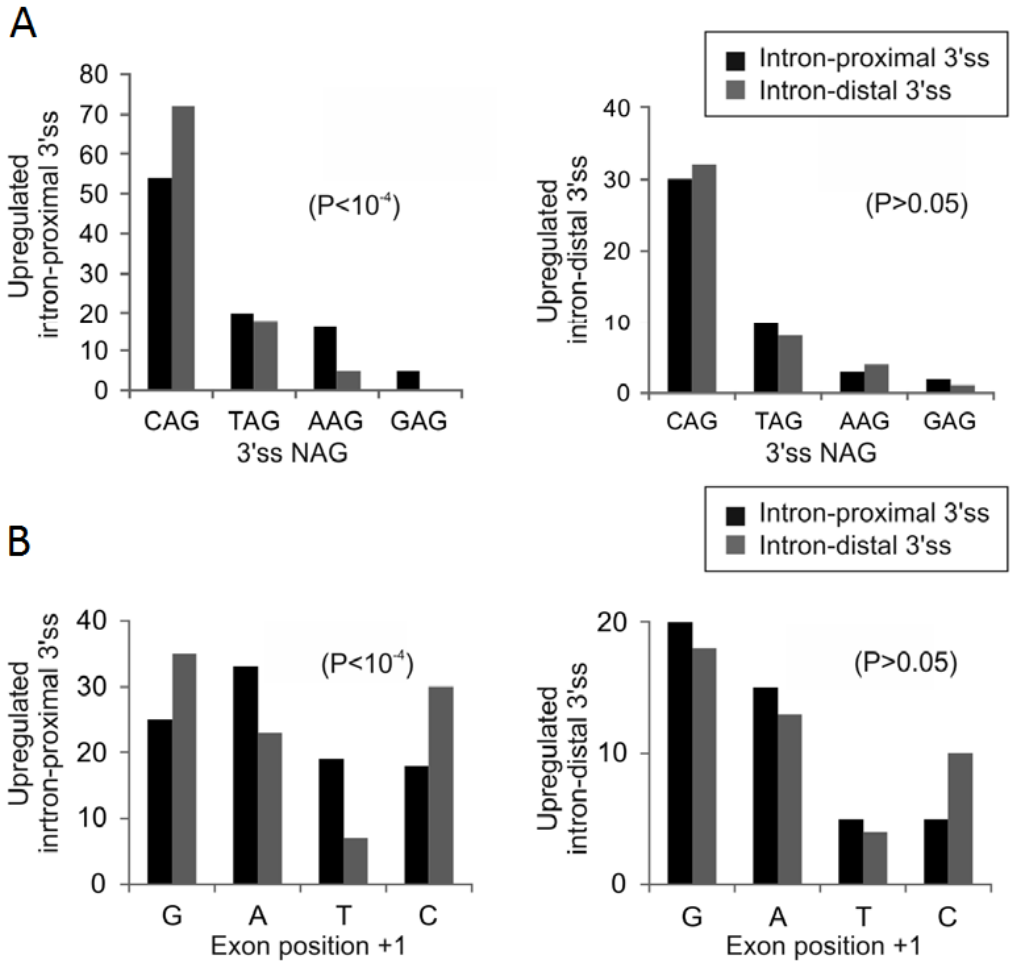

Figure B.8: Nucleotides at position -3 **(A)** and -1 **(B)** relative to U2AF(35)-dependent alternative 3' splice sites. Number of 3'ss with the indicated nucleotides at position -3. The number of upregulated proximal 3'ss was 93. The total number of upregulated distal 3'ss was 45. P-values were derived from χ^2 tests for 4x2 contingency tables.

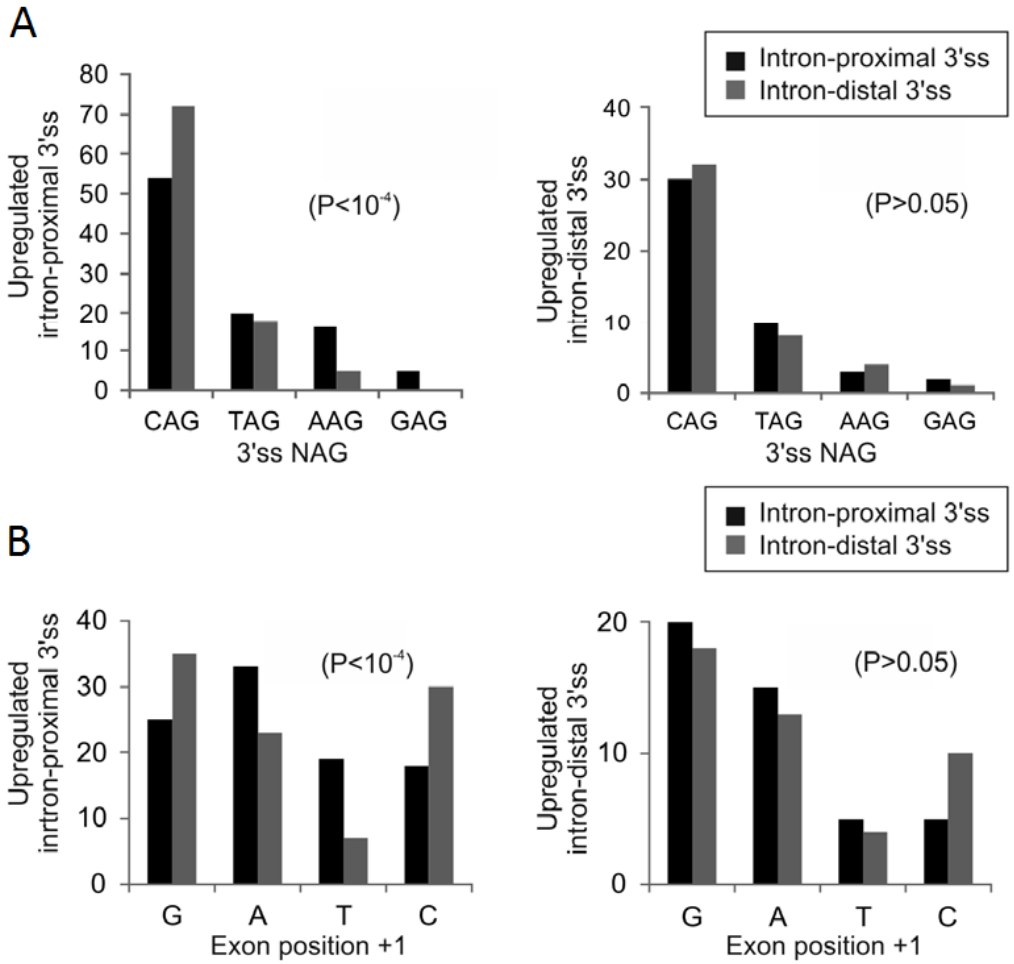

Figure B.9: Nucleotides at position -3 **(A)** and -1 **(B)** relative to U2AF(35)-dependent alternative 3' splice sites. Number of 3'ss with the indicated nucleotides at position -3. The number of upregulated proximal 3'ss was 93. The total number of upregulated distal 3'ss was 45. P-values were derived from χ^2 tests for 4x2 contingency tables.

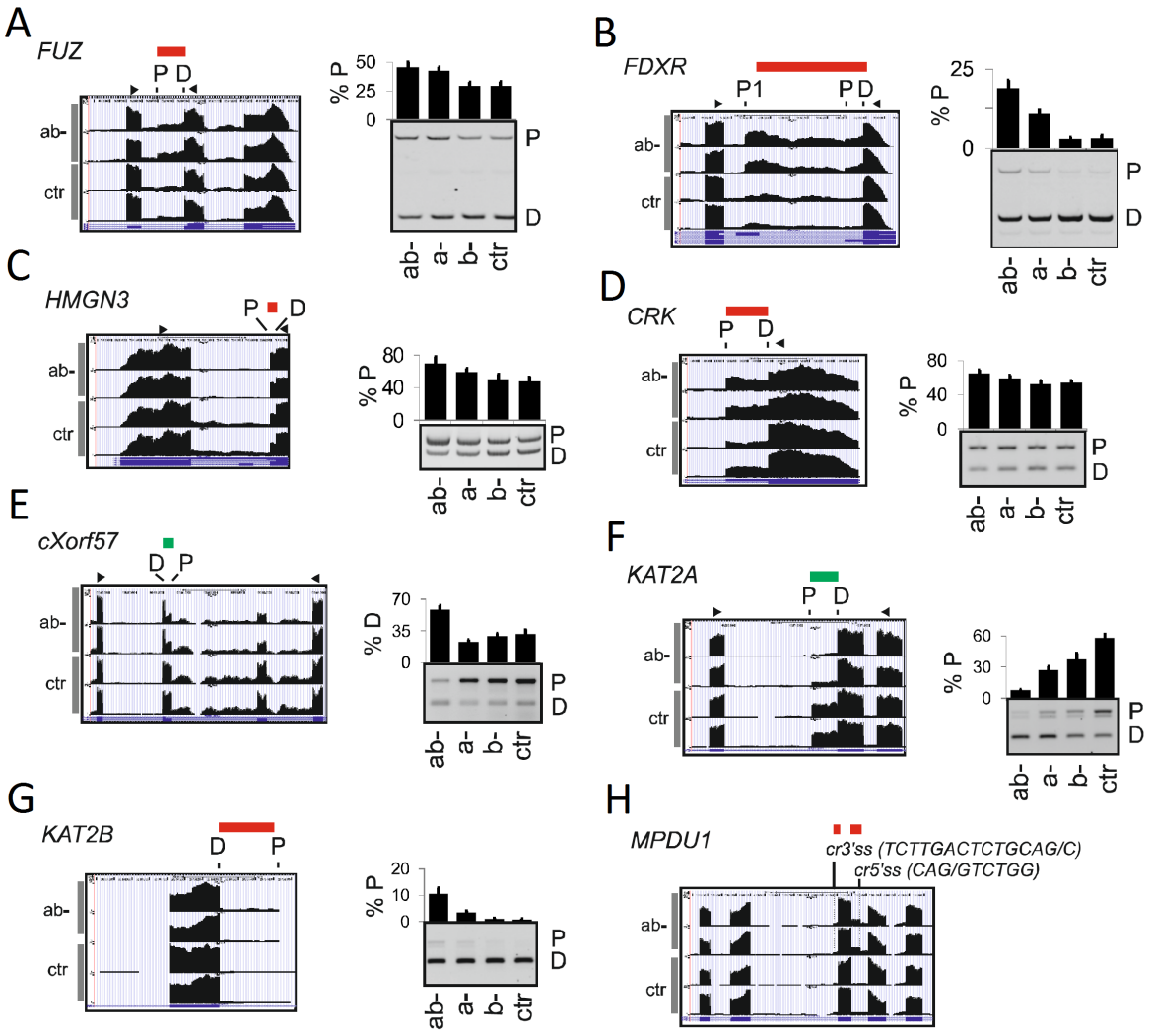

Figure B.10: Activation and repression of alternative 5' splice sites in depleted cultures. **A**, *FUZ*, **B**, *FDXR*, **C**, *HMGN3*, **D**, *CRK*, **E**, *cXorf57*, **F**, *KAT2A*, **G**, *KAT2B*, **H**, *MPDU1*. P, proximal 5'ss; D, distal 5'ss. Genome browwser views of RNA-Seq data from ab- and control cultures are shown to the left. Primers (Supplementary Materials in [296]) are denoted by arrowheads. Depletions are indicated at the bottom. Exonic segments up-/down-regulated in depleted cells are denoted by red/green rectangles.

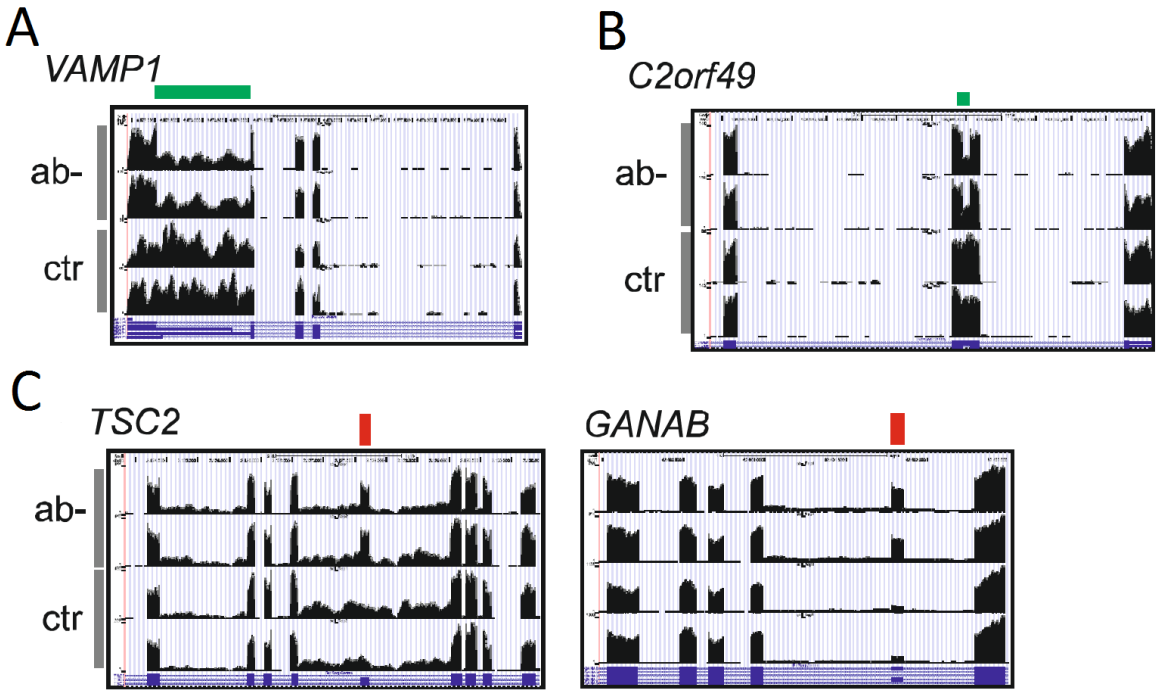

Figure B.11: U2AF(35) depletion can promote intron splicing and exon inclusion. **A**, Reduction of intron retention upon U2AF35 depletion. DEXSeq-detected downregulation of a 3'UTR intron (green rectangle, resulting in shorter 3'UTR, which was associated with *VAMP1* upregulation (Supplementary Materials in [296]). **B**, Intronization of a mid-portion of *C2orf49* exon 3 (green rectangle), employing a GC 4'ss (CAG/GCAAGC, where / is the exon-intro junction). **C**, Examples of upregulated exons (red rectangles) within intronic sequences incompletely eliminated from the indicated pre-mRNAs in untreated cells.

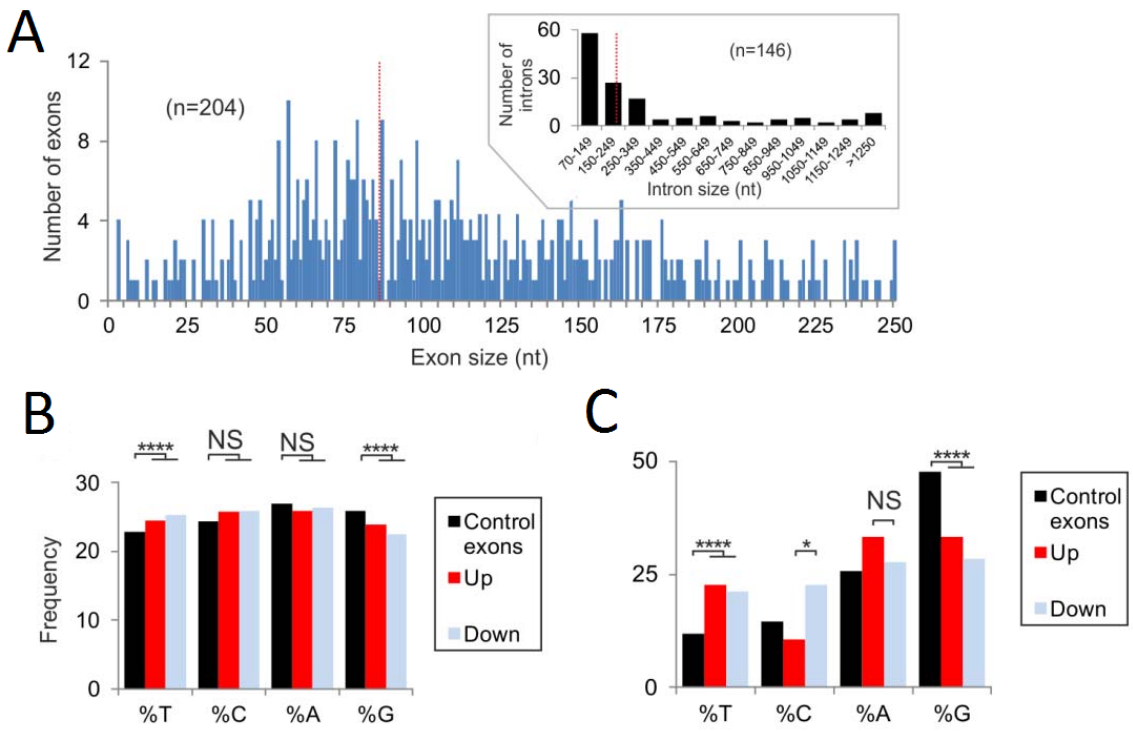

Figure B.12: U2AF(35)-dependent exons are smaller and are depleted of guanine and enriched in uridine as compared to average human exons.**A**, Size distribution of U2AF(35)-dependent internal exons. Median is denoted by a dotted vertical line. Size distribution of DEXSeq-detected pre-mRNA segments annotated as introns is shown in the inset. **B**, Nucleotide frequencies at position +1 of U2AF(35)-dependent exons, $****$, P<0.00005; NS, not significant. **C**, Nucleotide frequencies at position -1 of U2AF(35)-dependent exons. $*$, P<0.05. The first position of U2AF35-dependent exons was also depleted of guanine and showed the same hierarchy in splicing efficiency ($G > A > C > T$) as a distal cryptic 3'ss induced by a lack of U2AF35 or overexpression of U2AF65 [221]. Organization of U2AF(35)-dependent 3'ss is shown in Figures 6.16, 6.17, 6.18, 6.19, 6.20, 6.21, B.13, B.14, and B.15.

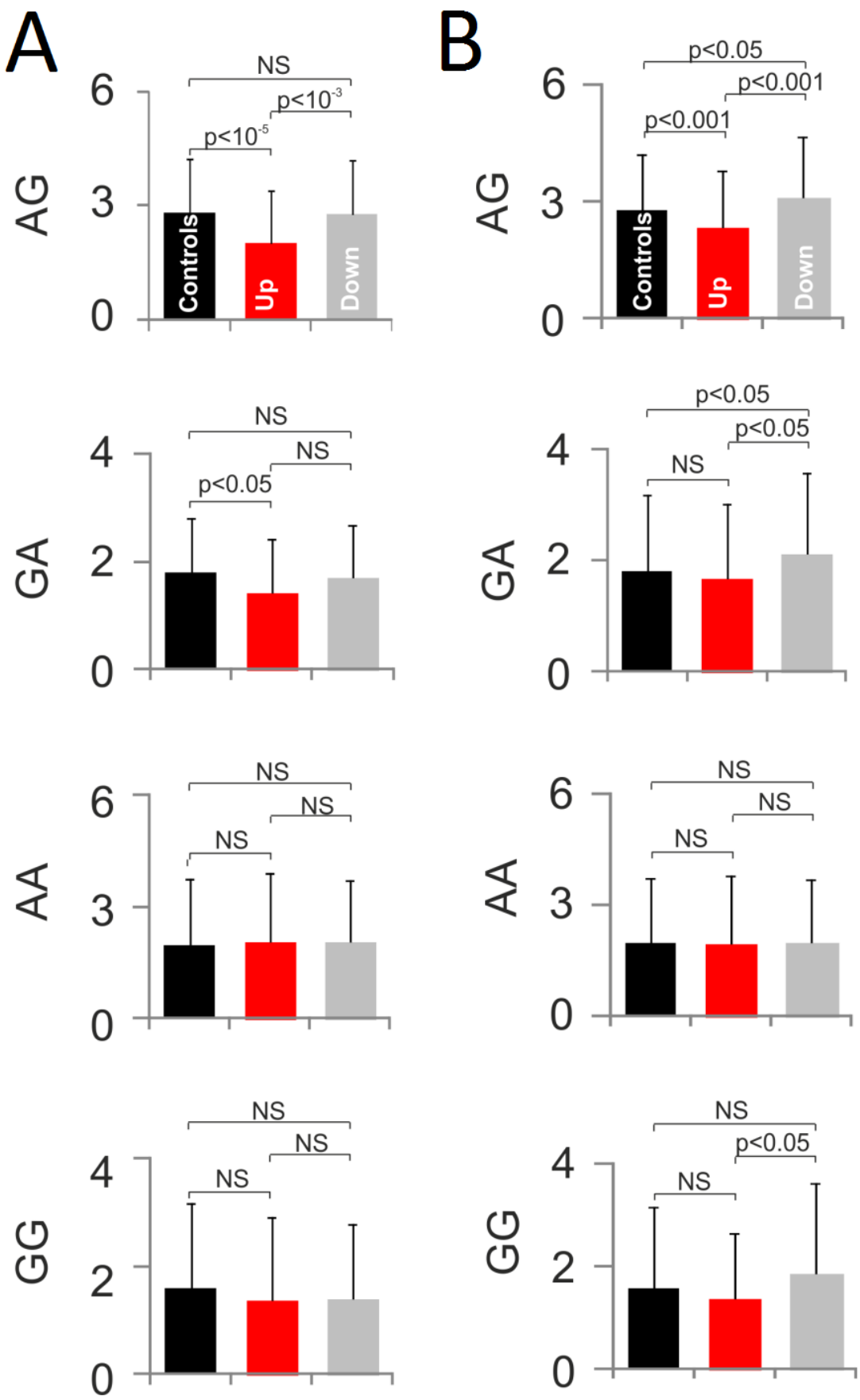

Figure B.13: Lack of AG dinucleotides in the last 50 nt of introns upstream of internal exons (A) and differentially used alternative 3' splice sites (B) in cells depleted of U2AF35. Columns represent means, error bars denote SDs. P-values were derived from two-tail t-tests. The number of up- and down-regulated internal exons was 67 and 137, respectively. The number of alternative 3' ss pairs was 138. The number of control exons was 177,290.

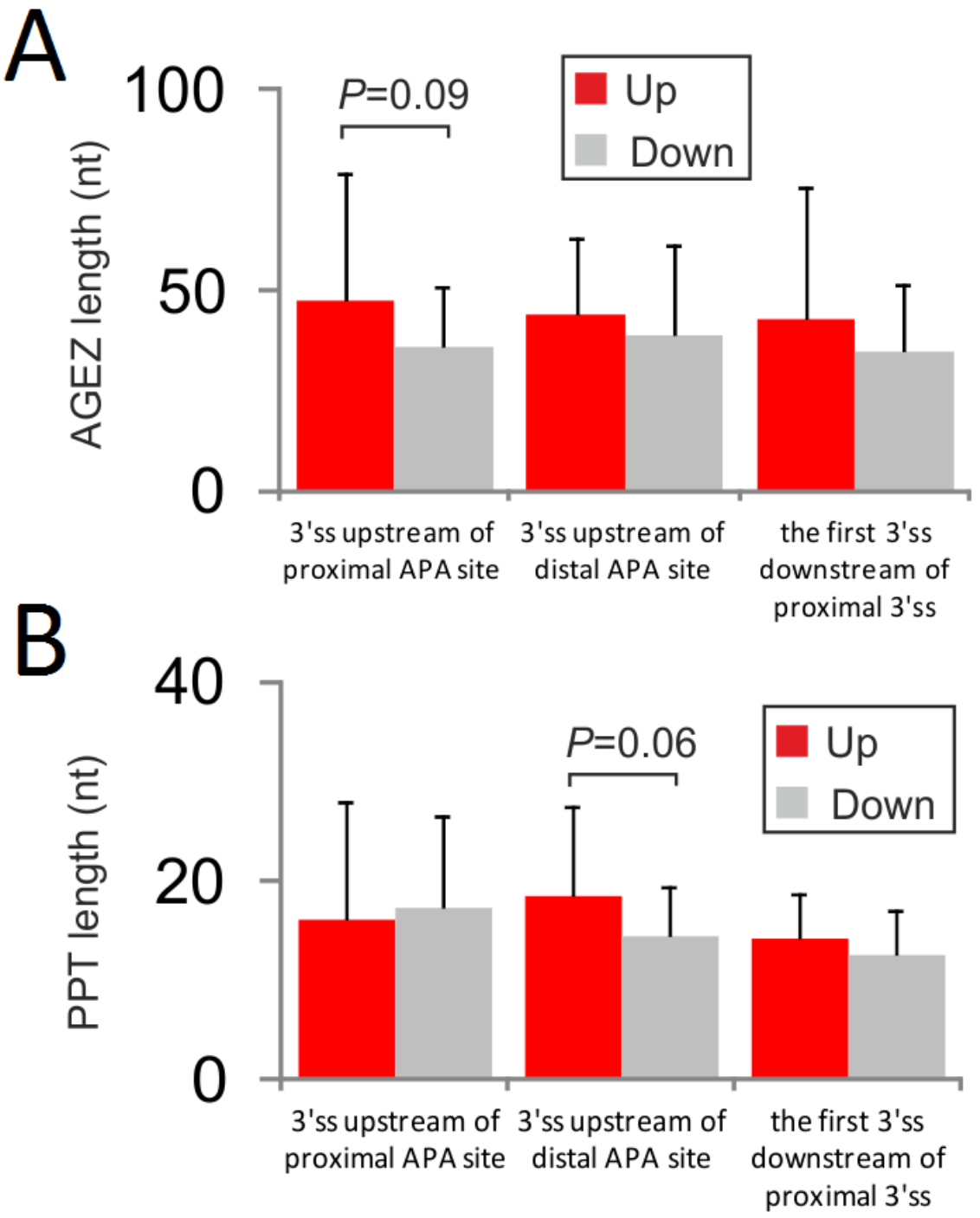

Figure B.14: AGEZ/PPT length of alternative 3'ss leading to differentially used APA sites. **A**, AGEZ length, **B**, PPT length. Each APA 3'ss (n=57, Figure 6.11 panel E) is shown in Supplementary Materials in [296]. Error bars are SDs; P-values were derived from t-tests.

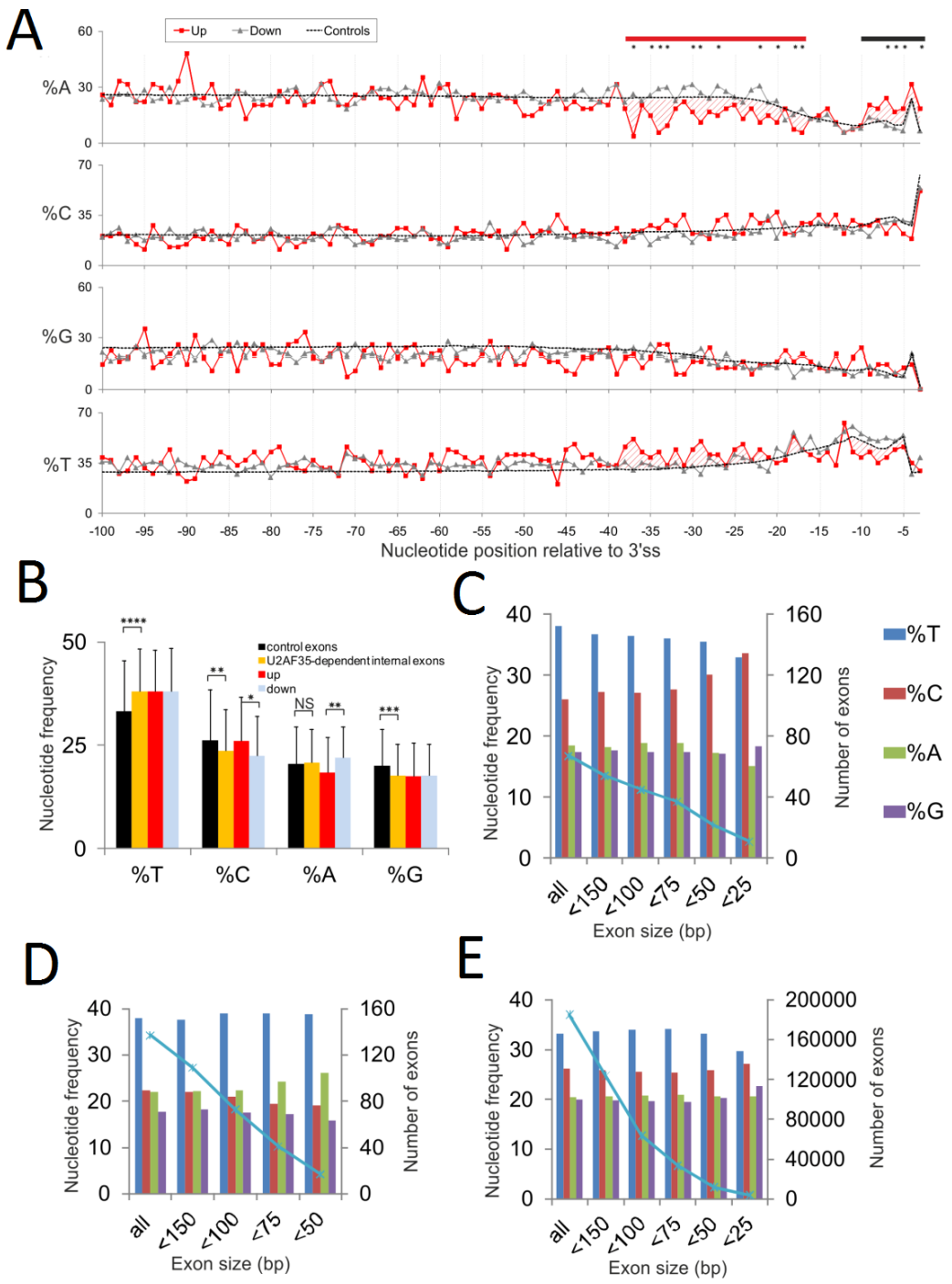

Figure B.15: Organisation of 3' splice sites of U2AF(35)-dependent exons. **A**, Nucleotide frequencies upstream of 3' ss of internal exons up-/down-regulated in cells depleted of U2AF35 (n=203) and control exons (n=177,290). Depletion/excess of adenine/uridine in upregulated exons in the canonical BP (red rectangle) loactions is highlighted; asterisks denote positions exhibiting significant differences (P<0.05) in adenince frequencies between up- and down-regulated exons in these regions. **B**, Systematic comparisons of nucleotide frequencies in 50-nt sequences upstream of U2AF(35)-sensitive internal exons. Error bars denote SDs. $*, P < 0.05, **, P < 0.005, ***, P < 0.0005, ****, P < 0.00005$; t-tests. **C-E**, The same frequencies by size of exons up- **(C)** and down- **(D)** regulated in ab- cultures and control exons **(E)**. Number of analyzed exons is shown to the right.

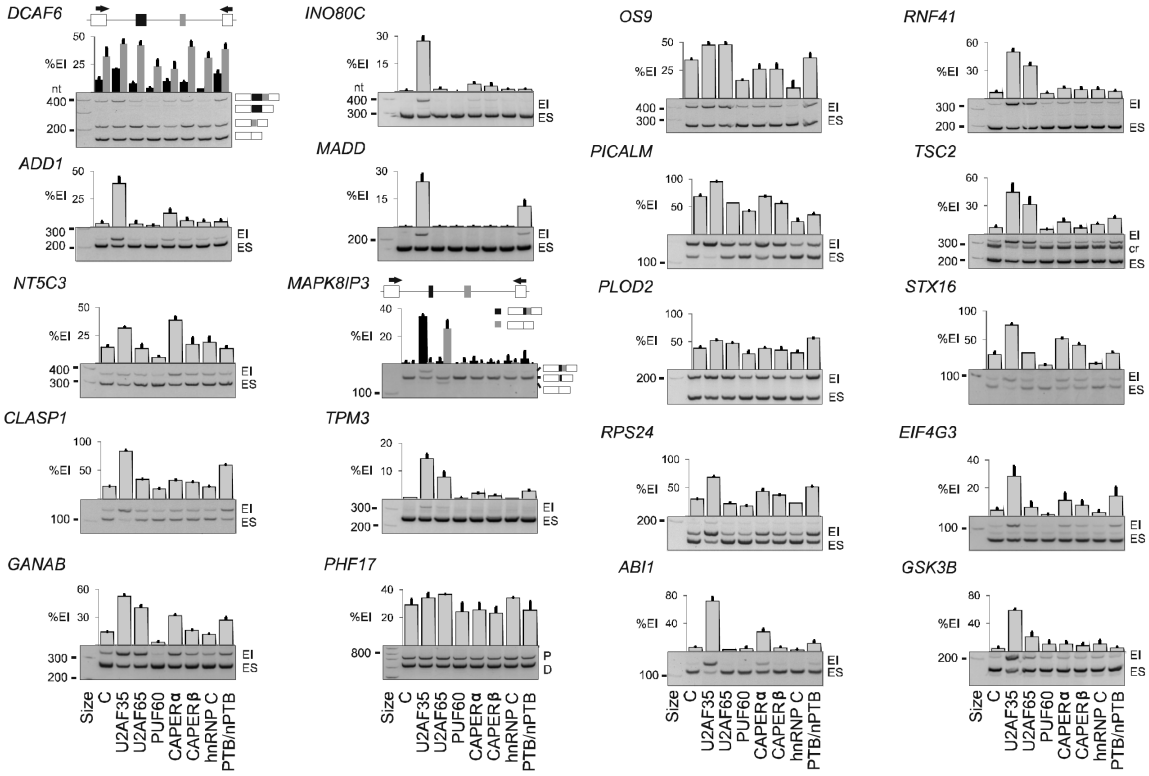

Figure B.16: Antagonism and synergism of U2AF-related proteins in U2AF(35)-dependent exons, exons upregulated in cells depleted of U2AF35. ES, exon skipping; EI, exon inclusion; P, D, proximal and distal 3' ss; cr, cryptic splice-site. Lane 1, size marker. Columns represent mean %EI or %ES ; error bars are SDs of two transfections into HEK293 cells. Amplifications primers are in Supplementary Materials in [296]. Pre-mRNAs containing >3 exons are schematically shown at the top. Immunoblots in Fig. 3G 6.18 and estimated of residual protein levels in Figure 6.19. Depletion levels of CAPERβ were estimated only at the RNA level (data not shown).

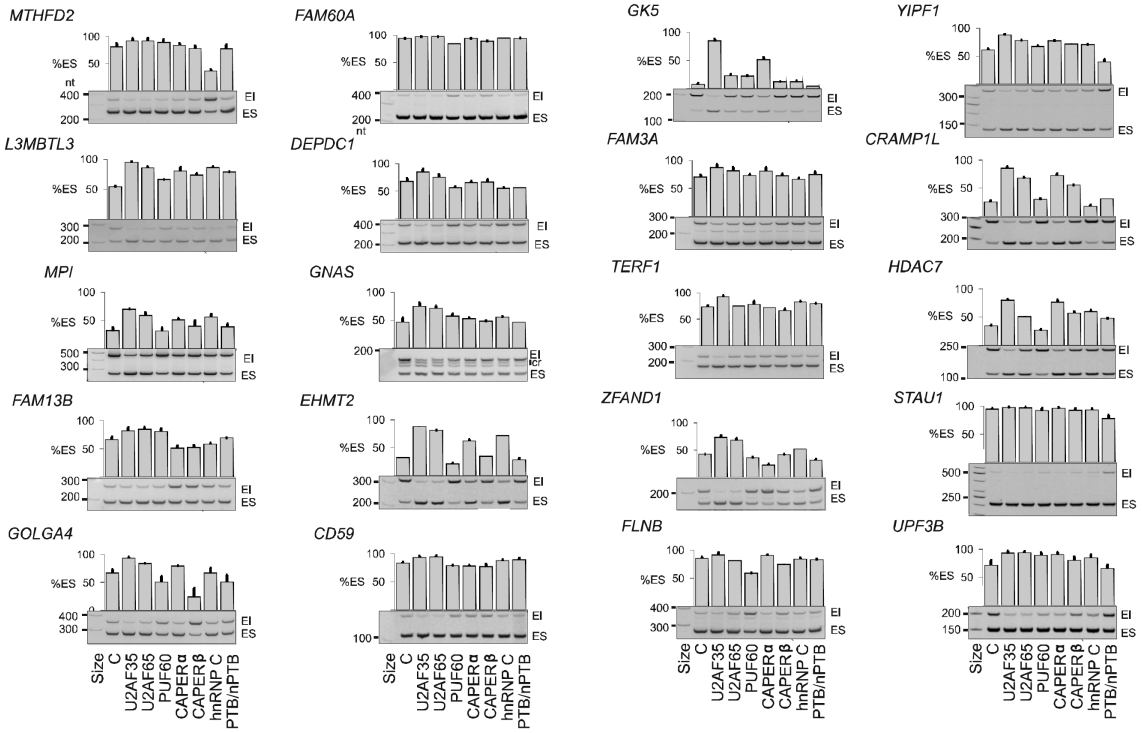

Figure B.17: Antagonism and synergism of U2AF-related proteins in U2AF(35)-dependent exons, exons downregulated in cells depleted of U2AF35. ES, exon skipping; EI, exon inclusion; P, D, proximal and distal 3' ss; cr, cryptic splice-site. Lane 1, size marker. Columns represent mean %EI or %ES ; error bars are SDs of two transfections into HEK293 cells. Amplifications primers are in Supplementary Materials in [296]. Pre-mRNAs containing >3 exons are schematically shown at the top. Immunoblots in Figure6.18 and estimated of residual protein levels in Figure 6.19. Depletion levels of CAPERβ were estimated only at the RNA level (data not shown).

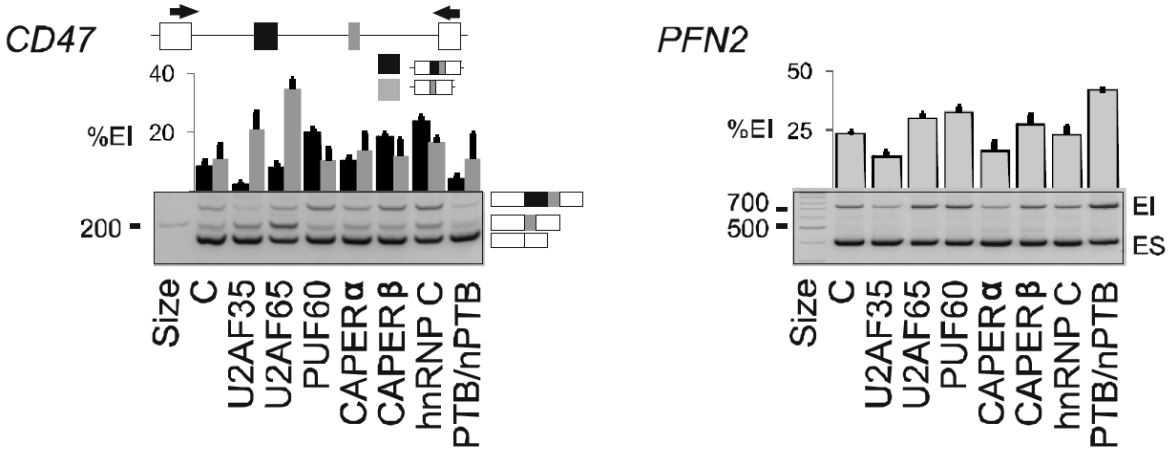

Figure B.18: Antagonism and synergism of U2AF-related proteins in U2AF(35)-dependent exons, exons with isoform-specific responses. ES, exon skipping; EI, exon inclusion; P, D, proximal and distal 3' ss; cr, cryptic splice-site. Lane 1, size marker. Columns represent mean %EI or %ES ; error bars are SDs of two transfections into HEK293 cells. Amplifications primers are in Supplementary Materials in [296]. Pre-mRNAs containing >3 exons are schematically shown at the top. Immunoblots in Figure 6.18 and estimated of residual protein levels in Figure6.19. Depletion levels of CAPERβ were estimated only at the RNA level (data not shown).

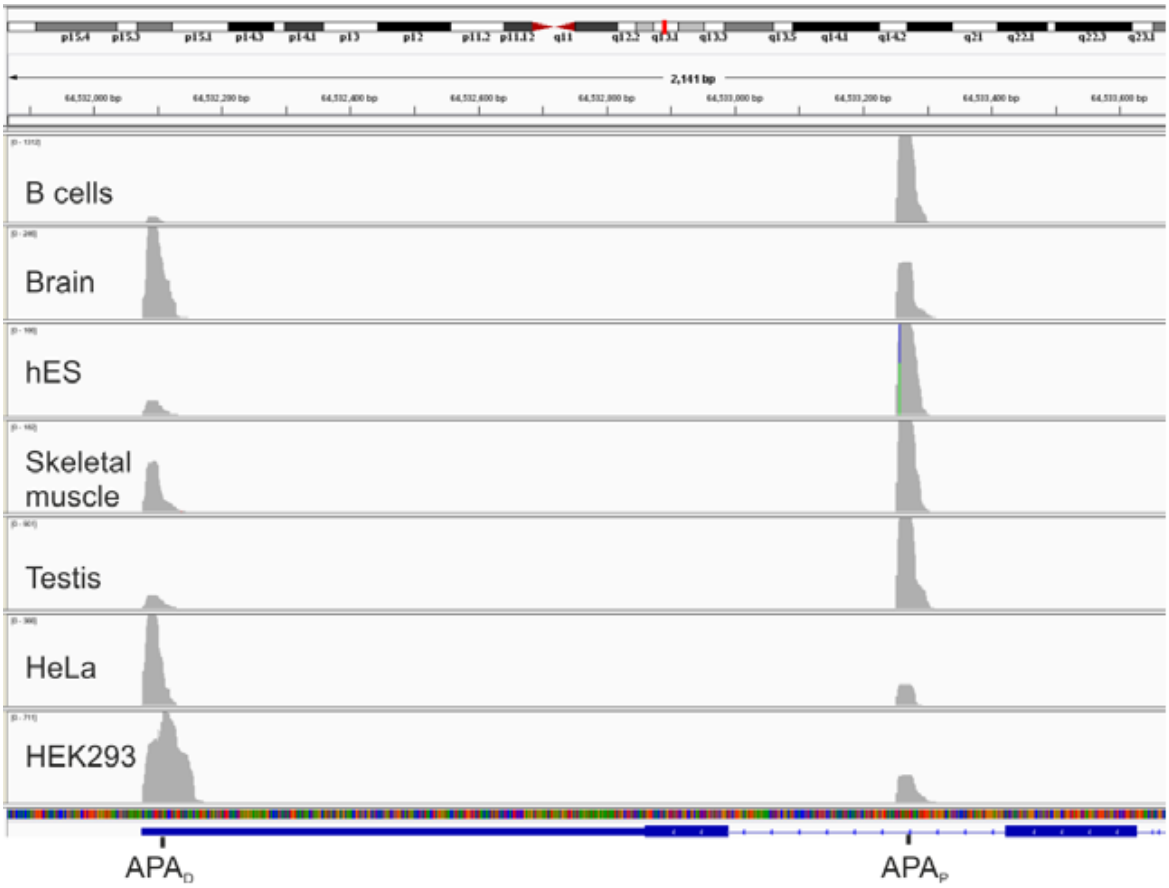

Figure B.19: Tissue-specificity of alternative 3'-end processing of human SF1. Distal and proximal APA (APA$_D$ and APA$_P$)site usage is shown for 3'-seq data described previously [242]. The proximal APA site is preferred in B cells, skeletal muscles and embryonic stem cells (hES).

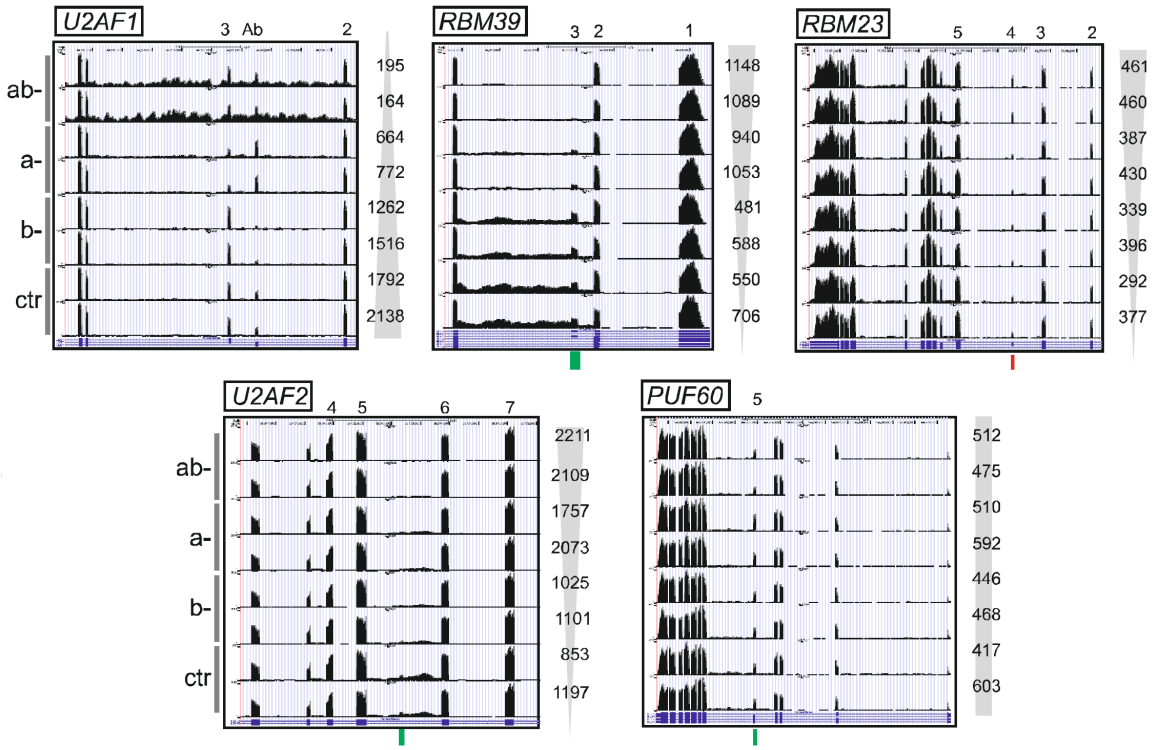

Figure B.20: Exon usage dependencies of U2AF35 binding partners. U2AF35 depletion (ab-) was associated with downregulation of a 60-bp cryptic exon in U2AF2 intron 5, *RBM39* (CAPERα) exon 3, and *PUF60* exon 5, and with upregulation of *RBM23* (CAPERβ) exon 4; a- and b- are isoform-specific depletions of U2AF35. Exon numbers are at the top; vertical arrows denote direction of gene-level expression changes expressed as read numbers (y-axis). Genomic coordinates are at the top and transcript annotations at the bottom. A functional alternative 3'ss of *RBM39* exon 13 was U2AF(35)-insensitive (data not shown).

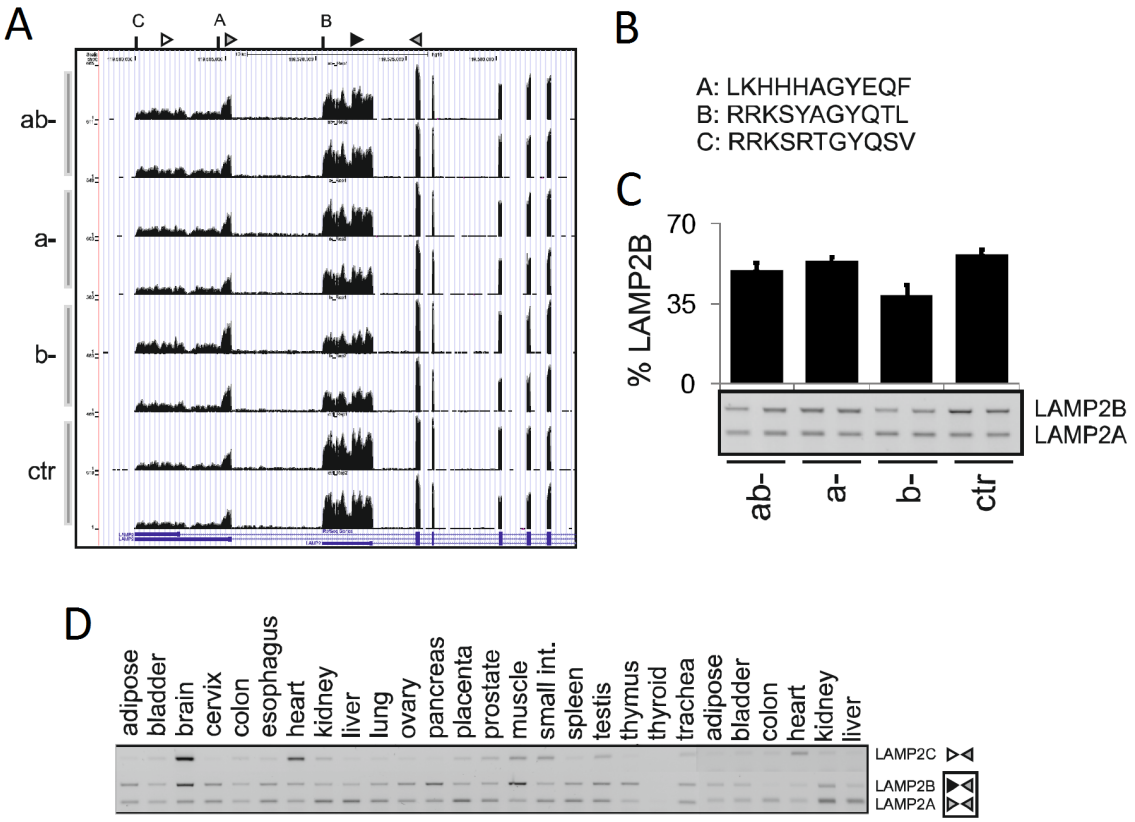

Figure B.21: Putative U2AF1b-specific interactions in regulation of *LAMP2* APA. **A**, Genome browser view of alternative RNA processing of LAMP2 transcripts. Annotated APA sites [242] are shown as vertical bars; primers (Supplementary Materials in [296]) are denoted by arrowheads, mRNA isoforms (top) encode distinct C-termini; their peptide sequences are in panel **B**. **C**, RT-PCR validation showing the relative abundance of isoform B in depleted cells and controls. **D**, Relative abundance of *LAMP2* isoforms in the indicated human tissues. Primers shown to the right correspond to arrowheads in panel A; box denotes multiplex reactions run for 28 cycles; isoform C was amplified in separate reactions for 34 cycles.

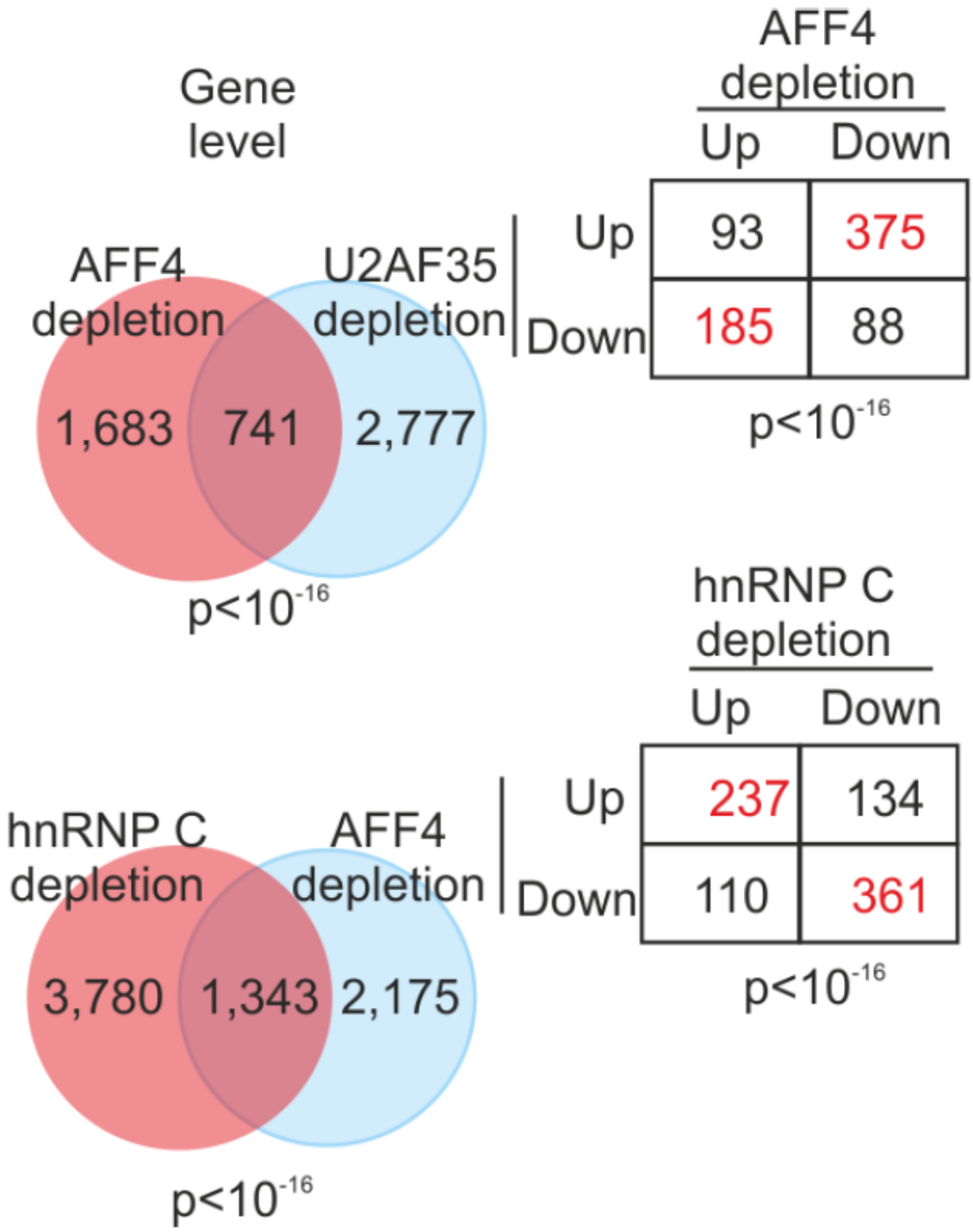

Figure B.22: Transcripts upregulated upon U2AF35 depletion tend to be downregulated in AFF4-depleted cells and vice versa. P-values were derived from hypergeometric and χ^2-tests for the indicated Venn diagrams and contingency tables, respectively, RNA-Seq data were analyzed using Cufflinks from this (U2AF35) and previous studies (hnRNP C, AFF4) [222][225].

Appendix C

Supplementary Tables

Sample	Gene fusion	Breakpoint 1	Breakpoint 2
6	*ANPEP-B2M*	chr15:90358003	chr15:45007620
6	*GCA-BAZ2B*	chr2:163213414	chr2:160189196
6	*CMTM6-DYNC1LI1*	chr3:32529429	chr3:32578595
6	*MAML3-FLOT1*	chr4:141074013	chr17:27210248
6	*ARID4A-PABPC1*	chr14:58765417	chr8:101721450
6	*ELOVL5-PTP4A1*	chr6:53213614	chr6:64286340
6	*ANP32A-EBC1D14*	chr15:69113036	chr4:7032053
6	*USP32-TIMP2*	chr17:58422841	chr17:76870000
6	*FNBP1-ZNF787*	chr9:132805230	chr19:56614595
6	*ATXN1-PTP4A1*	chr6:16761528	chr6:64286340
6	*PXN-UBC*	chr12:120703419	chr12:125398319
6	*BRD3-HNRNPUL1*	chr9:136933066	chr19:41774127
7	*PTBP1-BBC3*	chr19:797504	chr19:47725174
7	*KLF13-B2M*	chr15:31619991	chr15:45007620
7	*PAPSS1-BANK1*	chr4:108603170	chr4:102942670
7	*SIN3A-C15orf39*	chr15:75743763	chr15:75498339
7	*KLHL2-DOCK2*	chr4:166129053	chr5:169230061
7	*RAB3A-FKBP8*	chr19:18314705	chr19:18652804
7	*WRD5-NFATC1*	chr9:137023014	chr18:77211815
7	*KLF11-PTBP1*	chr2:10186276	chr19:797504
7	*ACTB-PTMA*	chr7:5570154	chr2:232576057
7	*SMEK2-SERTAD2*	chr2:55785910	chr2:64864008
7	*GSE1-SCL7A5*	chr16:85647003	chr16:87874760
7	*TCF25-ZC3H3*	chr16:89940266	chr8:144523289
7	*TPM4-ZNF516*	chr19:16187506	chr18:74155166
7	*TTYH3-MAD1L1*	chr7:2671911	chr7:1855863
7	*AKAP17A-L3MBTL3*	chr6:130343371	chrX:1710661
7	*LATS2-ING2*	chr13:21635484	chr4:184431434
7	*GSE1-SLC7A5*	chr16:85647003	chr16:87874760
7	*TIPARP-KLHL24*	chr3:156392584	chr3:183361267
7	*DDX5-HNRNPH3*	chr17:62502193	chr10:70096955
7	*RAB20-ING1*	chr13:111213695	chr13:111371576
8	*BCR-JAK2*	chr22:23526487	chr9:5080157
8	*LOC729852-GLCCI1*	chr7:7841373	chr7:8043537

Table C.1: Genomic locations of FCs. Locations are given for hg19 build GRCh37.

Sample	Gene fusion	Forward primer	Reverse primer	Size [bp]
6	*ANPEP-B2M*	CAGGGTCCAGGTTCCAGC	AGATAGAAAGACCAGTCCTTGCT	216
6	*GCA-BAZ2B*	TGTTGATCAAGATGGAAGTGGC	GGCAGTAGGTATGACAGCCT	150
6	*CMTM6-DYNC1LI1*	TGACAGGACTTCAGCTGAGA	AACTTCCGGATATGTGACTGAA	114
6	*MAML3-FLOT1*	GAGCAAACCCCAGCAAGATG	AGGATGGAGCGCAGATGTC	186
6	*ARID4A-PABPC1*	GCGTCATGAGCAGCCAATAG	AGGATAGTATGCAGCACGGT	238
6	*ELOVL5-PTP4A1*	TGTCTCCTTCTACATCCGCC	GCCTGGCAGTAATCTCCACT	171
6	*ANP32A-EBC1D14*	CGTGGGTTCGGGGTTTATTG	CCATTTCCCGGCTGTCTTTC	159
6	*USP32-TIMP2*	AGGATGCTTTCAAGAGGACCT	TGTCGTTTCCAGAGTCCACT	152
6	*FNBP1-ZNF787*	CTCGGGCCATTTTGCTGTG	TTCCCGCAGCTCCATGTC	249
6	*ATXN1-PTP4A1*	GATCCAAAACAAGCCCCGTG	GGCTTCTTGGTGGAGCAGTA	245
6	*PXN-UBC*	GCGCCCGTGGTCCGGCCA	GTAGTCAGACAGGGTGCGCC	210
6	*BRD-HNRNPUL1*	GCCGCCGGAGCCGCGAG	CGACTCTGGAATTGCTGCCTGT	116
7	*PTBP1-BBC3*	TTGGGTCGGTTCCTGCTATT	CTGGGTAAGGGCAGGAGTC	154
7	*KLF13-B2M*	CGGGCTGCGAGAAAGTTTAC	GACAAGTCTGAATGCTCCACT	220
7	*PAPSS1-BANK1*	TGGTGCTGAAAACAGACTCC	AAGGAATATATGTGGAAAATGAGGC	150
7	*SIN3A-C15orf39*	CTGTGACCGCTTCGTTAGTG	GCTTGCCATACATCACAGGC	246
7	*KLHL2-DOCK2*	GAATGGTGCTGGCTGTGTTG	CGCATCCTGGTAGCTAAGGA	232
7	*RAB3A-FKBP8*	ATCTACATGTGAGGCTCCGC	GCTCAGAGGGTTCAGCACA	230
7	*WRD5-NFATC1*	CTGTC(T/C)GGCCACAACTTCCT	AACTTCCCGACAGTCTCTCG	198
7	*KLF11-PTBP1*	TTGGGTCGGTTCCTGCTATT	CAGGTCGCGACGCTTAGG	184
7	*ACTB-PTMA*	CACAGAGCCTCGCCTTTG	CGATACTGCCACTGTGCAAA	250
7	*SMEK-SERTAD2*	ACATCTCCTGGTGGCTTCAA	CTTCCAGCCCATCTTCATGC	215
7	*GSE1-SCL7A5*	CCCGGGTGAGATAAGCAGTT	AATGCCAGCACAATGTTCCC	169
7	*TCF25-ZC3H3*	CGATGACGCGGAAGAAGAAG	TTTCTGGGTACGGTGGAGC	196
7	*TPM4-ZNF516*	CCCTGGAGGCGGTGAAAC	CCGTCCTATCTCTCCATGGT	213
7	*TTYH3-MAD1L1*	GTCGACGGGTCCCTGAAG	GTAGCTGCATCTTGGAACCC	243
7	*AKAP17A-L3MBTL3*	CTTTCGGCGGTGATGAAACA	TGGGACGCAGGCGGAGCC	105
7	*LATS2-ING2*	AGCCACCAGTGCCCGGTCTC	TCTGGAGAAGCTGCTGTAGACG	176
7	*GSE1-SLC7A5*	CCCGGGTGAGATAAGCAGTT	AATGCCAGCACAATGTTCCC	169
7	*TIPARP-KLHL24*	CGGCTCTGTGGTCCCTAG	GAATCACGCACCCCAAGATC	288
7	*DDX5-HNRNPH3*	GTCATCGAGGCCATTTCCAG	TCCTCTTTGCTGCAACCAAA	238
7	*RAB20-ING1*	CGCCTTCTACCTGAAGCAGT	AGTGACGCCTGTCCTTCTTG	223
8	*BCR-JAK2*	CAACAGTCCTTCGACAGCAG	TATGAGGATAGGTGCCCTGG	500
8	*LOC729852-GLCCI1*	TTGCTTTGAAAGGTAGGCGG	TGAACATGAGGATCCCGTGG	240

Table C.2: Primers for RT-PCR confirmation of gene fusions

cDNA primers	Sequence (5'-3')
BCR-C	ACCGCATGTTCCGGGACAAAAG
BCR-1F	CAACAGTCCTTCGACAGCAG
JAK2-19R	CCAGGGCACCTATCCTCATA
JAK25-R	ACCAGCCCTCATGTGTGAA
JAK25-R1	CACCTGCTTATAATGCTGGCC
JAK25-R2	TGAACACCAGCCCTCATGTG
PCM34-F	ACTTCCCTCCAGGCTAACAC
PCM35-F	ATGTCCCATTGGAACGAGAAGC
Genomic DNA primers	
JAK2 Int 16 R	TTCCATCTTCACTTCCGATTCCA
JAK2 Exon 17R	CTCCACTGCAGATTTCCCACAAA
...	...

Table C.3: Primer sequences, excerpt. For full table see Supplementary Materials in [296].

Sample number	Depletion	Replicate	Library preparation	Total reads	Mapped reads
201	ab-	1	poly(A) selection	71,997,266	67,746,694
202	ab-	2	poly(A) selection	73,066,622	68,791,493
203	a-	1	poly(A) selection	68,710,962	64,373,203
204	a-	2	poly(A) selection	71,148,044	66,826,986
205	b-	1	poly(A) selection	65,551,902	62,062,511
206	b-	2	poly(A) selection	70,707,318	65,797,232
207	ctrl	1	poly(A) selection	68,382,454	64,114,620
208	ctrl	2	poly(A) selection	76,839,596	71,038,554
241	a-	1	rRNA depletion	99,400,914	64,176,831
242	b-	1	rRNA depletion	95,775,200	61,335,955
243	ctrl	1	rRNA depletion	93,637,670	58,560,947
244	SSOa-	1	rRNA depletion	102,495,970	56,577,790
245	SSOb-	1	rRNA depletion	94,439,338	48,017,027
246	a-	2	rRNA depletion	89,937,626	57,316,617
247	b-	2	rRNA depletion	88,724,794	52,651,059
248	ctrl	2	rRNA depletion	98,663,920	55,433,043

Table C.4: Description of samples for RNA-Seq.

DEXSeq	Total
ab- vs a-	515
ab- vs b-	1388
ab- vs ctrl	1057
a- vs b-	375
a- vs ctrl	260
b- vs ctrl	84

MISO	Total	Alternative 3'ss	Alternative 5'ss	Mutually exclusive exons	Retained introns	Skipped exons
ab- vs a-	937	137	46	67	127	560
ab- vs b-	1797	196	86	142	344	1029
ab- vs ctrl	1705	202	104	127	255	1017
a- vs b-	589	0	37	68	144	340
a- vs ctrl	558	57	50	43	78	330
b- vs ctrl	198	19	17	34	65	63

Table C.5: Summary of DEXSeq- and MISO-detected events. For full table see Supplementary Materials in [296].

Gene	Chromosome	Strand	ExonID	Location (hg19)	Size (nt)	Dispersion	P-Value	Q-value	log2 -fold	Treatment effect
ABCA7	chr19	+	E035	chr19:1057313-1057428	116	1.48E-2	7.64E-05	1.61E-2	0.49	Down
ABCC10	chr6	+	E003	chr6:43399489-43399879	391	1.30E-2	0	0	-1.39	Up
ABCC5	chr3	-	E026	chr3:183701541-183702682	1142	2.60E-3	2.33E-14	3.97E-11	-0.30	Up
ABI1	chr10	-	E010	chr10:27060004-27060018	15	4.07E-2	6.78E-11	6.79E-08	-1.79	Up
ABR	chr17	-	E022	chr17:1012174-1012324	151	5.37E-2	1.25E-4	2.40E-2	-0.90	Up
ACCN2	chr12	+	E010	chr12:50474373-50474510	138	6.02E-2	2.44E-4	4.08E-2	-0.92	Up
ACIN1	chr14	-	E014	chr14:23540334-23540395	62	8.02E-3	3.78E-06	1.27E-3	0.40	Down
ACO2	chr22	+	E018	chr22:41924483-41924993	511	2.39E-3	4.12E-05	9.74E-3	-0.16	Up
ACOX1	chr17	-	E013	chr17:73969706-73969866	161	4.65E-2	1.74E-6	6.48E-4	1.09	Down
ADA	chr20	-	E004	chr20:43251229-43251293	65	3.72E-2	4.32E-7	1.88E-4	1.21	Down
...	...	...	...	...	...	...	...	...	...	...

Table C.6: Differentially used exons in ab- cultures versus controls detected by DEXSeq, excerpt. For full table see Supplementary Materials in [296].

ab- versus controls	a- versus controls	b- versus controls	a- versus b-
ABCA7	*ABCA11P*	*ABCC5*	*ABCA11P*
ABCC10	*ABCC10*	*AK2*	*ABCC9*
ABCC5	*ABCC5*	*ARPC4*	*ACP1*
ABI1	*ABCC9*	*ASCC3*	*ACTR3C*
ABR	*ADAT3*	*ATP5E*	*ADAMTS2*
ACCN2	*ADD3*	*BMP2T*	*ADAT3*
ACIN1	*AGAP3*	*BRD2*	*AGAP3*
ACO2	*AK2*	*C1orf182*	*AHRR*
ACOX1	*ALG9*	*CDC42*	*ALDOA*
ADA	*ALS2*	*CDK1*	*ALG9*
...	...	...	...

Table C.7: Pairwise comparisons of DEXSeq-detected exons in a-, b-, and ab- cultures and in controls, excerpt. For full table see Supplementary Materials in [296].

experiment	gene	log2 fold change	q-value	treatment effect
a- versus controls	*AAAS*	0.89	6.00E-4	down
a- versus controls	*AADAT*	1.29	6.00E-4	down
a- versus controls	*AARS*	-0.46	2.04E-2	up
a- versus controls	*AASS*	0.90	6.00E-4	down
a- versus controls	*AATK*	1.25	4.14E-3	down
a- versus controls	*ABAT*	1.084	6.00E-4	down
a- versus controls	*ABCA2*	0.46	1.63E-2	down
a- versus controls	*ABCA3*	0.53	6.06E-3	down
a- versus controls	*ABCA7*	0.70	6.00E-4	down
a- versus controls	*ABCB1*	-0.53	2.56E-2	up
...	...	...	...	...

Table C.8: Cufflinks-detected differential gene expression, excerpt. For full table see Supplementary Materials in [296].

Primer set	Primer or gene designation	Primer sequence (5'-3')
Validation set of RT-PCR primers for U2AF35-dependent exons	*PKM2*	CAGTGATGTGGCCAATGCAG
		TACCAGTGCCACGTTACAGC
	RBM39	CTCTTCCCGAACACGAGCAC
		TACGTTCTTCATGGCCGTTG
	ITGB3BP	GAGGGCAGTAGAGAGCTTGA
		CCTCTCCATGTTGGCTTACA
	ZFAND1	GAGAGATGGCGGAGTTGG
		TTGAAAGAGCATGGGTAAGA
	CD46	CTGTGATTGTTATTGCCATAG
		ATACCCAAATTCATACAAGTT
	MPI	TCTTTGGGGAGCTTTTGCTAC
		GTGGGTGTGCTGGCTATTACT
...	...	...

Table C.9: PCR primers. For full table see Supplementary Materials in [296].

Experiment	mRNA	Exon usage	Start exons	Terminal exons	Internal exons	χ^2 (p-value)	Data source
hnRNP C (-) versus untreated HeLa cells (R1)	Poly(A)-selected/RNASeq	down	2929	316	5707	4,367 (<0.0001)	[222]
		up	94	2522	4453		
hnRNP C (-) versus untreated HeLa cells (R2)	Poly(A)-selected/RNASeq	down	792	127	876	448(<0.0001)	[222]
		up	34	208	403		
HOXA1 (-) versus controls (3 replicas each)	Poly(A)-selected/RNASeq	down	11	54	112	1.5 (0.5)	[224]
		up	18	60	115		
hnRNP I (-) versus controls	Affymetrix microarrays 32/219 total	down	ND	21	146	2.5 (0.1)	[246]
		up	ND	10	32		
AFF3 versus untreated (3 replicas each)	Poly(A)-selected/RNASeq	down	1	5	16	14.95(<0.001)	[225]
		up	1	32	11		
AFF4 (-)	Poly(A)-selected/RNASeq	down	4	7	20	6.1 (0.05)	[225]
		up	0	14	25		
U2AF35 (-)	rRNA depletion	down	110	4	23	294 ($< 10^{-16}$)	This study
		up	1	200	14		

Table C.10: Start, terminal, and internal exons in depletion experiments. For full table see Supplementary Materials in [296].

APA type	Gene	APA site upregulated in ab-	competing 5'ss	Maximum entropy score	3'ss upstream of proximal APA site	Shapiro and Senapathy score	3'ss upstream of of distal APA site	Shapiro and Senapathy score	first 3'ss downstream proximal 3'ss APA site	Shapiro and Senapathy score
...	...	...	...	...	...	...	...	...	...	...
III	*LAMP2*	Distal	N/A	N/A	TTTCTCACCTACAGC	85.87	TTCTCCACATCTAGC	76.34	TTCTCCACATCTAGC	76.34
III	*PAFAH1B2*	Distal	N/A	N/A	CCACTGTGCCCCAGG	77.43	TCTTAATGTTTCAGA	86.1	TCTTAATGTTTCAGA	86.1
III	*GLS*	Distal	N/A	N/A	GCTTGAACAACTAGC	59.45	TGCTACGTGTTTAGG	75.83	TCTTTTCTTCACAGG	96.15
III	*LEPROTL1*	Distal	N/A	N/A	TCTTATTTCCATAGA	82.96	TTTCTGTTTTCTAGA	88.55	TTTCTGTTTTCTAGA	88.55
III	*ZNF226*	Distal	N/A	N/A	TTTTTGTATTTCAGA	95.45	CTTTGTCCTTACAGG	95.09	CTTTGTCCTTACAGG	95.09
III	*MTERFD2*	Distal	N/A	N/A	TTTGTGTCTTCCAGT	90.12	CTTTATTTTTTTAGG	93.2	TTTGTTTTCTCCAGC	92.14
III	*DTNA*	Distal	N/A	N/A	TTATATCATTTCAGC	85.4	CAATCTTTCTGTAGG	84.32	CATTGTCTCTCCAGA	87.29
III	*RPL28*	Distal	N/A	N/A	TCCCCCGCCCCCAGG	78.08	TCCTCTGTTCACAGG	88.94	TCCTCTGTTCACAGG	88.94
III	*PVRL3*	Distal	N/A	N/A	ACTGTTTCCATTAGA	75.08	GTTTGAATTTTTAGA	78.62	TTGTTACTTTACAGA	92.64
III	*AGAP3*	Distal	N/A	N/A	TTCGATATTTGCAGA	85.18	GTTTTCTCCTACAGT	88.76	GTTTTCTCTTCCAGT	91.08
...	...	...	...	...	...	...	...	...	...	...

Table C.11: List of APA sites influenced by U2AF35 depletion by APA category, excerpt. Type I: U2AF35-sensitive tandem 3'UTR APAs; Type II: U2AF35-sensitive transcripts with intronic APA sites; Type III: U2AF35-sensitive alternative 3'ss APA. For full table see Supplementary Materials in [296].

Transcript	Coordinates (hg19)
ALG9	chr11:111709101-111709122
ANKHD1	chr5:139916923-139916974
ANP32E	chr1:150195566-150195620
APTX	chr9:32987729-32987844
ARFIP2	chr11:6501653-6501693
ARHGAP8	chr22:45255560-45255611
ARMC8	chr3:137907243-137907297
ARMCX5-GPRASP2	chrX:101854634-101854640
ATP10D	chr4:47538703-47538748
ATPIF1	chr1:28564272-28564348
...	...

Table C.12: Alternative 3' splice sites influenced by U2AF35 depletion, excerpt. For full table see Supplementary Materials in [296].

Transcript	Coordinates (hg19)	Validated by RT-PCR
AKT1S1	chr19:50380282-50380494	
ARFIP1	chr4:153701243-153701378	
BANF1	chr11:65769984-65770041	
CLN6	chr15:68503894-68503987	
CRK	chr17:1339914-1340084	Y
CUL7	chr6:43019851-43019947	
CWC25	chr17:36966529-36966721	
CXORF57	chrx:105881025-105881154	Y
DMKN	chr19:36001273-36001279	
FDXR	chr17:72868894-72868991	Y
...	...	...

Table C.13: Alternative 5' splice sites influenced by U2AF35 depletion, excerpt. For full table see Supplementary Materials in [296].

Upregulated alternative 3'ss	Number of pairs	Mean score of proximal 3'ss	Mean score of distal 3'ss	Upregulated alternative 5'ss	Number of pairs	Mean score of proximal 5'ss	Mean score of distal 5'ss
Intron-distal	45	83.58	77.86	Intron-distal	6	78.6	76.82
Intron-proximal	93	77.76	80.37	Intron-proximal	25	77.51	75.07

Table C.14: Shapiro and Senapathy scores of alternative 3' and 5' splice sites influenced by U2AF35 depletion. For full table see Supplementary Materials in [296].

Gene	Strand	Exon ID	Exon size	p-value	q-value	Treatment effect
MADD	+	E031	63	7.64E-05	1.61E-2	up
TFIP11	-	E014	77	2.99E-10	2.62E-7	up
ANK3	-	E009	7812	0	0	up
INO80C	-	E005	54	5.12E-12	6.11E-97	up
INO80C	-	E006	54	2.54E-11	2.74E-8	up
ABI1	-	E010	15	6.78E-11	6.79E-8	up
MAPK8IP3	+	E012	18	1.09E-06	4.21791E-4	up
TPM3	-	E004	79	0	0	up
NPTN	-	E009	348	1.11E-16	2.49E-13	up
EIF4G3	-	E034	33	1.29E-9	9.55E-7	up
...	...	...	...	...	...	...

Table C.15: Internal exons influenced by U2AF35 depletion, excerpt. For full table see Supplementary Materials in [296].

Gene	Strand	Exon ID	Location	Dispersion	p-value	q-value	log2-fold	Treatment (b-) effect
ABCA11P	-	E002	chr4:433779-438221	3.71E-3	2.28E-14	6.11E-10	-0.22	down
ABCA11P	-	E001	chr4:419224-420762	6.45E-3	1.25E-9	8.78E-06	0.36	up
ACBD6	-	E008	chr1:180471180-180472022	3.61E-3	2.81E-7	6.85E-4	0.20	up
ACTB	-	E001	chr7:5566779-5567522	2.72E-3	4.88E-5	2.90E-2	-0.08	down
ALKBH5	+	E004	chr17:18111533-18113267	1.87E-3	5.51E-5	0.03	-0.06	down
ALYREF	-	E006	chr17:79849199-79849462	3.61E-3	8.36E-5	0.04	0.16	up
AMOT	-	E001	chrX:112018105-112021892	1.64E-3	1.86E-5	1.43E-2	-0.07	down
AMZ2	+	E003	chr17:66244228-66244780	6.47E-3	1.87E-5	0.01	0.27	up
ANKRD13	-	E013	chr1:70819662-70820417	3.39E-3	2.8E-11	2.24E-7	0.26	up
ANKRD17	-	E001	chr4:73940502-73942007	2.73E-3	8.28E-5	0.04	-0.13	down
...	...	...	...	...	...	...	...	...

Table C.16: List of significantly differentially expressed exons between a- and b- cultures in samples depleted of rRNA, excerpt. For full table see Supplementary Materials in [296].