Detektion und Validierung sporadischer Splice-Ereignisse in Transkriptom Sequenzierungs-Daten

Inaugural-Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftliche Fakultät
der Heinrich Heine Universität Düsseldorf

vorgelegt von

Herrn Dipl.-Math. Dr. med. Wolfgang Kaisers
aus
Neuss

Datum der Einreichung: 2018/04/11

Referent: Prof. Dr. Martin Lercher
Koreferent: Prof. Dr. Heiner Schaal

Tag der mündlichen Prüfung: 2018/05/12

Abstract

In general, primary eukaryotic messenger RNA is spliced before translation, a nuclear process resulting in removal of intronic sequence segments. The splicing process is regulated by a complex network consisting of a large variety of interacting factors. For analysis of splicing processes, transcriptome sequencing (RNAseq) has become an established standard. Genomic alignments of RNAseq data contain gapped alignments which are interpreted as consequence of intron removal. Alignment gap locations, possibly shared by multiple alignments, define gap-sites, landmarks representing potential splice-sites. As current alignment algorithms report gap-sites with a considerable false discovery rate, validations are required. Two quality scores, gap quality score (*gqs*) and weighted gap information score (*wgis*), developed for validation of putative splicing events, are described here. While *gqs* solely relies on alignment data, *wgis* additionally considers information on splice-site strength from the genomic sequence. Statistical properties of gap-sites validated by *gqs* and *wgis* are evaluated by their sequence similarity to known exon-intron borders.

splicing, alternative splicing, splice-sites, validation, gqs, wgis

Zusammenfassung

Vor der Translation wird die prä mRNA fast aller eukaryotischer Gene im Zellkern gespleißt. Dabei werden intronische Sequenz-Abschnitte entfernt. Spleißvorgänge werden durch ein komplexes Netzwerk, bestehend aus einer Vielzahl interagierender Faktoren, reguliert. Für die Analyse von Spleißvorgängen ist die Transkriptom Sequenzierung (RNAseq) ein etablierter Standard. Infolge der Entfernung intronischer Sequenzsegmente finden sich Lücken (Alignment Gaps) in genomischen Alignmnents von RNAseq-Daten. Identische Alignment Gaps in (meist mehreren) Alignments werden zu Gap-Sites zusammengefasst und repräsentieren potenzielle Speißstellen. Da ungefilterte Gap-Sites mit einer beträchtlichen False Discovery Rate behaftet sind, ist eine zusätzliche Validierung notwendig. Zwei Gap-Site validierende Scores, der Gap-Quality-Score (*gqs*) und der Gewichtete-Gap-Informations-Score (*wgis*), werden in dieser Arbeit vorgestellt. Während der *gqs* ausschließlich Alignment Daten in die Validierung einbezieht, beurteilt *wgis* auch die Spleißstellenstärke anhand der genomischen DNA Sequenz.

Spleißen, alternatives Spleißen, Spleißstelle, Validierung, gqs, wgis

Inhaltsverzeichnis

Kapitel 1

Vorwort

Diese Arbeit ist eine kumulative Dissertation entsprechend § 6 Abs. 1 der Dissertationsordnung in der Fassung vom 6.12.2013 der Heinrich-Heine Universität Düsseldorf und basiert auf zwei wissenschaftlichen Publikationen:

- *Manuskript 1:* „rbamtools: an R interface to samtools enabling fast accumulative tabulation of splicing events over multiple RNA-seq samples“, erschienen in Bioinformatics (2015), beschreibt einige Grundzüge der Implementation der entwickelten Algorithmen [47]. Das Manuskript ist unter dem DOI `https://doi.org/10.1093/bioinformatics/btu846` abrufbar.

- *Manuskript 2* „Validation of Splicing Events in Transcriptome Sequencing Data “, erschienen im International Journal for Molecular Sciences (2017), enthält die Definition und die Untersuchung der zur Validierung von Splice-Ereignissen verwendeten Scores [46]. Das Manuskript ist unter dem DOI `https://doi.org/10.3390/ijms18061110` abrufbar.

Beide Publikationen sind über die angegeben DOI frei zugänglich. Sie beschreiben Algorithmen für die Detektion und Validierung potenzieller Spleißereignisse (Gap-Sites) in sequenzierten Transkriptomen (RNAseq) und ihrer Implementation in drei R-Paketen.

Die drei R-Pakete sind in CRAN und Bioconductor Paket-Quellen enthalten:

- *rbamtools*: Das Paket enthält eine R-Schnittstelle zur *SAMtools* Bibliothek, die Standard Implementierung für den Zugriff auf das SAM/BAM

Format. In *rbamtools* sind Algorithmen implementiert, mit den Gap-sites (potenzielle Spleißstellen) identifiziert werden können. Das Paket wurde am 12.03.2012 in der Version 2.0 auf CRAN veröffentlicht und enthält neben einer statischen Kopie des SAMtools Interfaces (Version v1.4-r985) 20.787 Zeilen C code und 7.019 Zeilen R code.

- *refGenome*: Das Paket enthält Algorithmen zum Import genomischer Annotationsdaten (beispielsweise aus Ensembl[1]), zur Berechnung genomischer Koordinaten annotierter Spleißstellen und einen spezialisierten Algorithmus für die Annotierung von Gap-sites. Das Paket wurde am 05.08.2013 zuerst auf CRAN veröffentlicht und enthält 2.924 Zeilen C Code und 2.625 Zeilen R Code.

- *spliceSites*: Das Paket entält eine Klassenbibliothek, in der der Analyseprozess für Gap-Sites abstrahiert ist und enthält Funktionen, mit denen biologische Fragestellungen (beispielsweise alternatives Spleißen, Spleißstellenstärke, Spleißstellen-Konsensus-Sequenzen) beantwortet werden können. Das spliceSites Paket ist seit 2013 auf Bioconductor (seit BioC 2.13). Der Quellcode umfasst 1.684 Zeilen C Code und 4.958 Zeilen R Code.

In den Paketen sind zwei Gap-Site bewertende Scores (*gqs* und *wgis*) implementiert, die eine Differenzierung von Sequenzierungs- und Alignment-bedingten Artefakte und zellulären Spleißereignissen ermöglichen sollen.

Dieses Dokument enthält nach dem Vorwort eine Einleitung in der die Inhalte der Arbeiten in einen größeren wissenschaftlichen Zusammenhang eingeordnet werden. Danach werden die Inhalte der Manuskripte skizziert und ihre Kernaussagen wiedergegeben. Im Anhang befindet sich eine Tabelle mit Erklärung verwendeter Abkürzungen.

Dieses Dokument ist in LaTeX und unter der Verwendung der DissOnline-Vorlage der Deutschen Nationalbibliothek[2] erstellt worden.

[1] `https://www.ensembl.org/`
[2] `http://files.dnb.de/dissonline/dissonline_latex_ver_2.1.zip`

Kapitel 2

Einleitung

Im ersten Abschnitt wird ein Überblick über die molekularen Mechanismen des prä-mRNA Spleißens gegeben. Der zweite Abschnitt beschreibt die Entwicklung von Algorithmen zur Detektion und Validierung resultierender Spleißereignisse (Gap-Sites) in Transkriptom-Sequenzierungsdaten.

2.1 Mechanismen des prä-mRNA Spleißens

2.1.1 Das humane Genom

Das menschliche Genom enthält etwa 60.000 Gene. Nur ein Drittel davon, etwa 20.000, sind proteincodierend. Die verbleibenden 40.000 Gene sind nicht codierend. Etwa 16.000 davon werden einer unscharf umschriebenen Gruppe langer nichtcodierender RNAs (lncRNAs), etwa 10.000 einer Gruppe kleiner nichtcodierender RNAs und etwa 14.000 der Gruppe der Pseudogene zugeordnet [76]. Mehr als 90 % der proteincodierenden Gene enthalten Introns, die im Zellkern aus dem Primärtranskript durch Spleißen entfernt werden müssen, um einen korrekten Leserahmen zu generieren [21, 38].

2.1.2 Das Spleißen

Beim Spleißvorgang werden aus der prä-mRNA (auch bezeichnet als hnRNA), noch während der Transkription, intronische Sequenzabschnitte entfernt [17, 40, 19]. Die Exon/Intron-Übergänge, an denen die prä-mRNA geschnitten

wird, werden Spleißstellen genannt; dabei werden 5' Spleißstellen (5'ss) oder auch Donoren von 3' Spleißstellen (3'ss) oder auch Akzeptoren unterschieden. Das Spleißen von prä-mRNA ist ein komplexer Vorgang. Er wird durch einen makromolekularen Komplex, dem Spleißosom, ausgeführt. Darüber hinaus sind zahlreiche andere Komponenten an der Regulation beteiligt. Die Spleißstellenerkennung folgt keinen einfachen Regeln (wie beispielsweise dem genetischen Code), muss aber, um den Leserahmen aufrechtzuerhalten, mit nukleotidgenauer Präzision erfolgen.
Viele Details der Spleißstellenerkennung sind bis heute nicht aufgeklärt. Auch ist über die Fehlerrate des Spleißapparates bis heute nichts Genaues bekannt. Da eine umfängliche Darstellung der Vorgänge beim Spleißen von prä-mRNA den Rahmen dieser Arbeit sprengen würde, sei für weitergehende Information an dieser Stelle auf Übersichtsarbeiten verwiesen [4, 7, 40, 60].

Das Spleißosom Das Spleißosom ist ein makromolekularer Komplex, der in mehreren aufeinanderfolgenden Schritten aus mehreren snRNPs (small nuclear Ribonucleoproteinen) zusammengesetzt wird und der den eigentlichen Spleißvorgang durch zwei Transesterreaktionen katalysiert. Die snRNPs wiederum bestehen aus Uracil-reichen snRNAs (small nuclear RNAs), an die jeweils etwa 6 bis 10 Proteine gebunden sind [70].
Das Major-Spleißosom (U2-abhängig) besteht aus den U1, U2, U4, U5 und U6 snRNPs und katalysiert mehr als 95 % aller Spleißvorgänge; das Minor-Spleißosom (U12-abhängig) besteht aus den U11, U12, U4atac, U5 und U6atac snRNPs [81, 76]. Im Weiteren beziehen sich alle Aussagen auf das Major-Spleißosom[1].
Das (Major-) Spleißosom ist mit mehr als 170 Proteinen assoziiert [6, 16, 44, 51, 60] und wird als eine der komplexesten molekularen Maschinerien eukaryotischer Zellen angesehen [41, 67, 86].

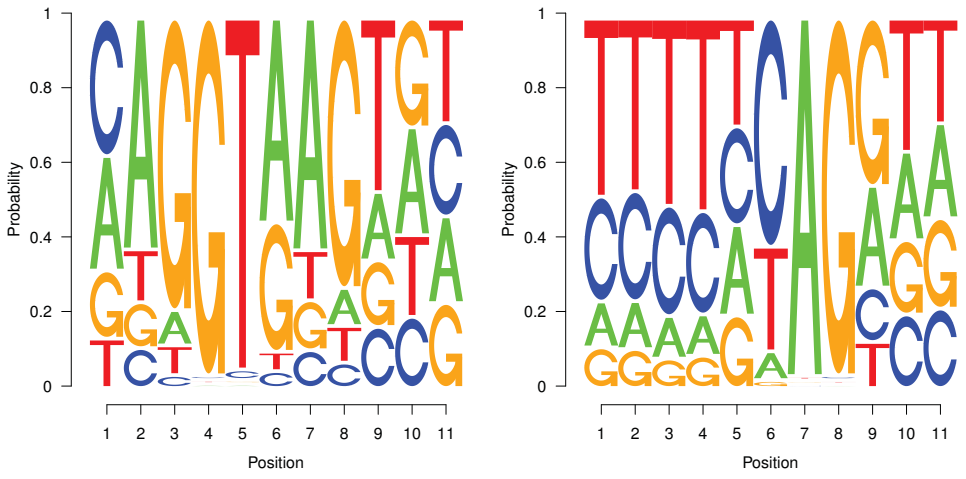
Der Spleißvorgang Die ersten Schritte der Spleißosomassemblierung bestehen aus der Bindung des U1 snRNPs an die 5'ss, einer lockeren Assoziierung des

[1] Beschreibungen zum Minor-Spleißosom finden sich in [81, 11].

Spleißfaktors SF1 an die Verzweigungssequenz stromaufwärts der 3'ss (Branch Point, BP) und dem U2AF (U2 Auxilliary factor; ein Dimer bestehend aus U2AF65 und U2AF35) an den Polypyrimidin Trakt [34, 61].
Der initiale Schritt der E-Komplex Bildung, die Bindung des U1 snRNPs an die 5'ss, wird durch spleißregulatorische Proteine (*trans*-aktivierende Faktoren) reguliert. Die Bindeplätze der *trans*-aktivierenden Faktoren auf der prä-mRNA werden als *cis*-aktive Sequenzen (oder auch spleißregulatorische Elemente, SRE) bezeichnet.
Der E-Komplex geht dann zunächst in den A-Komplex und später in den B-Komplex über. Infolge von Umstrukturierungsvorgängen, bei denen unter anderem das U1 snRNP den B-Komplex wieder verlässt, bildet sich der katalytische C-Komplex, der durch zwei Transesterreaktionen das Intron entfernt.
Die an der E-Komplex Bildung teilnehmenden *trans*-aktivierenden Faktoren wechselwirken über Protein-Protein-Interaktionen mit Proteinen des U1 snRNPs, wodurch die RNA Duplexausbildung zwischen dem freien 5' Ende der U1 snRNA und der 5'ss unterstützt oder behindert wird. Durch den Einfluss ihrer Gegenwart auf die Rekrutierung des U1 snRNPs an die 5'ss regulieren sie die nachfolgenden Schritte der Ausbildung des Spleißosoms [8, 40, 74].

2.1.3 Spleißstellenerkennung

Drei RNA Sequenzbereiche, die 5'ss, die 3'ss und der Branch-Point (BP) sind die zentralen Landmarken (Kern-Signale) bei der Spleißstellenerkennung [33, 86]. Über 98 % aller Introns in Mammalia beginnen mit GU und enden mit AG. Alle übrigen Nukleotide der 5' und 3' Spleißstellen sind deutlich weniger konserviert (siehe Abbildung 2.1) [13, 54]. Der Branchpoint (oder das Verzweigungsnukleotid) ist meist ein Adenosin [34, 64].
Bei der Spleißstellenerkennung spielen in der Regel SREs eine wichtige Rolle, besonders bei nur geringer Komplementarität der mRNA Sequenz an der 5'ss zur U1 snRNA [33, 52].



A: 5' Spleißstellen **B:** 3' Spleißstellen

Abbildung 2.1: Degenerierte humane Spleißstellen Sequenzen

Darstellung der Sequenz-Motiv Verteilung für potenzielle Spleißstellen aus RNAseq Daten mittels Sequenz-Logo [75]. Die Beispiele beruhen auf Exon- und Intron-Sequenzen verifizierter Spleißstellen aus einem RNA-seq Experiment [46] (Abbildung 10, c und d). Man erkennt, daß außer den ersten beiden intronischen Nukleotiden (GT und AG), alle anderen Positionen erhebliche Variation aufweisen. **A:** 5' Spleißstellen. Die Exon-Intron Grenze befindet sich zwischen Position 3 und 4. **B:** 3' Spleißstellen. Die Intron-Exon Grenze befindet sich zwischen Position 8 und 9. Auf den Positionen 1 bis 6 erkennt man den Polypyrimidin-Trakt.

***Trans*-aktivierende Faktoren** Neben dem Spleißosom spielen bei der Erkennung von Spleißstellen spleißregulatorische Proteine eine wesentliche Rolle

.

Mehr als 800 mRNA bindende Proteine (RBB) sind bislang identifiziert worden, die an vielen mRNA prozessierenden Mechanismen, beispielsweise Polyadenylierung, zytosolischer Export, Translation, Nonsense-mediated Decay mitwirken [14, 71]. Aus ihnen rekrutieren sich die *trans*-aktivierenden Faktoren, Proteine, die an prä-mRNA binden und regulierend am Spleißvorgang beteiligt sind, aber nicht direkt dem Spleißosom zugerechnet werden. Die bekanntesten Beispiele für *trans*-aktivierende Faktoren sind SR-Proteine und hnRNPs [7, 86].

SR-Proteine sind eine Protein-Familie, bestehend aus 12 Mitgliedern, die ei-

ne oder zwei RNA-bindende Domäne(n) und eine (Arginin/Serin reiche) RS-Domäne enthalten [7, 32, 77]. Meist binden SR-Proteine an Purin-reiche Sequenzen [33] und unterstützen die Adhäsion von U1 oder U2 snRNP an die prä-mRNA [57] (Ein Modell für die Faktor-Interaktion findet sich in [13], Abbildung 2). Ein klassisches Beispiel für Funktionen von SR-Proteinen ist die Kooperation von U1 snRNP mit ASF/SF2 bei der Bindung an die 5'ss [52, 32]. Die mRNA - Bindungsstellen für SR-Proteine sind in der Regel degeneriert [7, 13, 24, 77].
hnRNPs (heterogenous nuclear Ribonucleoproteins) sind die zweite große Gruppe *trans*-aktivierender Faktoren und fungieren, wenn sie an exonische SREs binden, meist als Spleißrepressoren [43]. hnRNP bestehen aus einem Komplex aus hnRNA (prä-mRNA) und hnRNP-Proteinen [20]. Beispiele für die Aktivität von hnRNP sind die Inhibition der Bindung von U1 snRNP an die 5'ss durch hnRNP A1 oder antagonistische Aktivitäten von SF2 und hnRNP bei der Auswahl alternativer 5'ss [24, 62, 63].

***Cis*-aktive Elemente** Sequenzabschnitte auf der prä-mRNA, an die spleißregulatorische Proteine binden, aber selbst nicht zu den Spleißsignalen gehören, werden *cis*-aktive Elemente (SRE) genannt. Meist handelt es sich dabei um kurze (4-11 Nukleotide lange [9, 56, 73]) Sequenzmotive in der Nähe von Spleißstellen [33], die als Bindungsstellen für *trans*-aktivierende Faktoren dienen. Aus historischen Gründen werden SREs aufgrund ihrer Position und ihrer Funktion in exonische Enhancer (ESE), exonische Silencer (ESS), intronische Enhancer (ISE) und intronische Silencer (ISS) eingeteilt [86].
Aus der Arbeitstruppe um Christopher Burge wurden 238 ESE [28], 133 ESS [88], 102 ISS [85] und 109 ISE [84] identifiziert. Mittels ClustalW wurden daraus für ESE 10, für ESS 6, für ISS 10 und für ISE 6 verschiedene Konsensus Motive extrahiert. Die Konsensus Bereiche sind teilweise degeneriert und die Zuordnung zu potenziell bindenden *trans*-aktivierende Faktoren ist nicht eindeutig (sie gleicht teilweise eher einem Netzwerk [85]). Exonische Spleiß-Enhancer sind in vielen Exons präsent und machen hier einen großen Anteil der Sequenz aus [5, 28, 50, 86].

Einfluss sekundärer Faktoren Neben den Effekten primärer Faktoren, wie dem Spleißosom, den *trans*-aktivierenden und den SREs spielen auch sekundäre Faktoren bei der Spleißstellenerkennung eine Rolle. Ob die Bindung eines *trans*-aktivierenden Faktors an ein SRE die Spleißstellenerkennung fördert oder behindert, kann beispielsweise von der relativen Position des SRE zur Spleißstelle (stromaufwärts oder stromabwärts) abhängen [25, 78, 33]. Weitere potenzielle Faktoren sind Interaktionen zwischen *trans*-aktivierenden Faktoren [55, 33], Phosphorylierung von *trans*-aktivierenden Faktoren [57, 55] oder auch Chromatin-Struktur [40] und Transkriptionsgeschwindigkeit [1].
Die Erkennung einer Spleißstelle hängt also von der Präsenz zentraler Signale (wie der Komplementarität zur U1 snRNA) und den kombinierten Effekten positiv und negativ beeinflussender Faktoren ab [78, 33, 13].

2.1.4 Der Splicing Code

Die Tatsache, dass die Spleißmaschinerie trotz vieler Fehlermöglichkeiten, mit hoher Genauigkeit arbeitet [86] deutet darauf hin, dass es in der prä-mRNA Sequenz Signale geben muss, die die Spleißvorgänge steuern. Die Entschlüsselung dieser Steuervorgänge ist Gegenstand vieler Untersuchungen.
Als Korrelat für die Effizienz der Spleißstellenerkennung wird in der Literatur der Begriff der Spleißstellenstärke verwendet, für den allerdings keine einheitliche Definition existiert. Uneindeutigkeiten entstehen durch das Heranziehen verschiedener Bewertungsmaßstäbe. Beispielsweise kann die Spleißstellenstärke einen berechneten Score oder die tatsächliche Nutzung einer Spleißstelle bezeichnen.
In dieser Arbeit wird der Begriff der „intrinsischen“ Spleißstellenstärke für die Bewertung mit dem HBond-Score oder dem MaxEnt-Score verwendet, die eine enge Umgebung der Spleißstelle analysieren.
Während der HBond-Score [31] beispielsweise nur den Grad der Komplementarität zwischen den Nukleotiden des freien 5’ Endes der U1 snRNA bewertet, gewichtet der MaxEnt Score [91] auch nicht komplementäre Nukleotide. Diese einfachen Modelle sind komplexeren Modellen, die auch die weitere Umgebung der Spleißstellen in die Bewertung mit einbeziehen, beispielsweise bei dem HEXplorer-Score Algorithmus [26], weit unterlegen. Die komplexen Modelle werden auch „Splicing code“ genannt [86]. Von mehreren Arbeitsgruppen

wurden Überlegungen [61, 86] oder Modellentwürfe [5, 73, 90] für die Entwicklung eines Splicing Codes beschrieben. Eine Evaluation von Scores, die Spleißstärken bewerten, findet sich in [37].

Derzeit gibt es allerdings keinen publizierten Algorithmus, der in der Lage wäre, human pathogene Spleißstellenmutationen und deren Auswirkungen auf den Spleißapparat zuverlässig zu diagnostizieren. Dies zeigt, dass die Beschreibung der Spleißstellenstärke auch heute noch unzureichend ist und dass verlässliche Aussagen zur Spleißstellennutzung nur aus experimentellen Daten möglich ist (siehe auch [33]).

2.1.5 Alternatives Spleißen

Praktisch alle Gene werden nicht nur konstitutiv, sondern auch alternativ gespleißt. Die unterschiedliche Nutzung einzelner Spleißstellen führt so zu verschiedenen Transkriptisoformen von nur einer prä-mRNA [27, 74, 68, 82]. Dabei lassen sich in der Regel wenige Haupt-Isoformen identifizieren, die im Median mehr als 30 % der Transkript-Menge ausmachen [17].
Während der Ontogenese (bei der zellulären und gewebespezifischen Differenzierung) laufen regulierte Spleißprogramme ab, die bis zu mehrere hundert Gene einschließen kann. Besonders komplexe alternative Spleißprogramme werden beim Menschen im Zentralnervensystem (ZNS) und in der quergeschreiften Muskulatur (Herz und Skelettmuskulatur) beobachtet [4].

2.1.6 Bedeutung von Spleißereignissen in Biologie und Medizin

Auf den ersten Blick erscheint die Entfernung intronischer Segmente aus prämRNA Sequenzen wenig sinnvoll zu sein, da ausführende Komponenten in der Zelle vorgehalten werden müssen und damit auch ein zusätzlicher Energiebedarf für die Zelle verbunden ist. Andererseits weist die enorme Komplexität des Spleißvorgangs und die evolutionäre Konservierung des Spleißosoms [59] auf eine bedeutende biologische Rolle hin. Die auffallend schwache Determinierung der Regulationsvorgänge scheint hier Freiräume für (ontogenetisch und phylogenetisch) benötigte Komplexität zu schaffen. Sie schränkt gleichzeitig

aber auch Vorhersagbarkeiten ein.

Das bedeutet, dass vermutlich in absehbarer Zeit für die genomweite Untersuchung von Spleißvorgängen keine Alternative zur experimentellen Beobachtung und zur Analyse mit RNAseq verfügbar wird. In diesem Zusammenhang spielt die Validierung potenzieller Spleißereignisse in RNAseq Daten eine wichtige Rolle.

Abberantes Spleißen ist bei vielen Krankheitsbildern ursächlich beteiligt [2, 3, 10, 12, 35, 76, 83, 22, 55, 36]. Bei einigen neuromuskulären Erkrankungen sind bereits spleißmodulierende Therapieansätze in Sicht [18, 72] oder schon in der Klinik verfügbar [30].

2.2 Identifikation von mRNA Spleißereignissen in Transkriptom-Daten

Ausgangspunkt für die Entwicklung von Algorithmen zur Detektion und Validierung von Spleißereignissen war, dass RNAseq sich als kommender Standard für die Transkriptom-Analyse ankündigte [58, 65, 87]. Zum Zeitpunkt des Projektstarts (im November 2010) waren zwar schon Alignment Algorithmen verfügbar [80], es existierte aber noch kein Standardverfahren für die Extraktion von Gap-Sites aus Alignment-Daten. Es war allerdings schon bekannt, dass die Detektion potenzieller Spleißstellen mit einer hohen FDR (False Discovery Rate) behaftet ist.

Die hier ansetzende Zielsetzung war die Bereitstellung und Evaluation von Algorithmen, mit denen (alternative) Spleißereignisse in Transkriptom-Sequenzierungs Daten (RNAseq) zuverlässig identifiziert werden können. Die zuverlässige Identifizierung ist als Vorstufe für einen späteren direkten Nachweis spezifischer Spleißereignisse mittels PCR notwendig, weil die Untersuchung einer großen Anzahl von Spleißereignissen mittels PCR mit einem unverhältnismäßig großen Aufwand an Material und Arbeitszeit verbunden wäre.
Um die Sensitivität des Verfahrens zu erhöhen, sollte die validierende Information für die einzelnen Spleißereignisse aus mehreren Proben kumulativ ausge-

wertet werden. Eine Annotation der Spleißereignisse, das heißt die Zuordnung zu annotierten Spleißstellen sollte möglich, aber nicht Voraussetzung für die Detektion sein.

2.2.1 Experimentelle Daten

Die experimentellen Daten deren Analyse der Konzeption und der Implementation zugrunde lagen waren RNAseq Daten aus einer Studie zur altersabhängigen Veränderungen des Transkriptoms dermaler Fibroblasten. Für diese Untersuchung waren von 30 Probanden 60 Gewebeproben aus gesunder Haut entnommen worden. Die Probanden waren in drei Altersgruppen eingeteilt (Jung: 19 bis 25 Jahre, Mittel: 36 bis 45 Jahre, Alt: 60 bis 66 Jahre). Jeweils die Hälfte der Probanden war männlich oder weiblich. Jedem Probanden wurden zwei Proben entnommen, eine Probe aus UV-exponierter Haut (Schulter) und eine Probe aus einer lichtgeschützten Region (Gesäß). Die gewonnenen Proben sind dann, im Rahmen des BMBF geförderten Gerontosys Verbundes mit verschiedenen Verfahren untersucht worden (beispielsweise auch mittels histologischer Schnitte). Die Proben wurden auf einem Illumina HiSeq 2000 Sequenzierer analysiert. Die Einzelheiten des experimentellen Aufbaus und der Probenverarbeitung sind an anderer Stelle beschrieben [45, 46].
Das Ergebnis der Transkriptom Sequenzierung sind kurze DNA Sequenz Fragmente, die in FASTQ-Dateien [15] ausgegeben werden. Für die Analyse sind 54 FASTQ-Dateien mit jeweils etwa $1-1,5\times10^8$ Reads (Read-Länge 101) verwendet worden. Sechs FASTQ-Dateien sind nach Auswertung diverser Qualitätsmerkmale, vor allem dem HcKmer - Verfahren [49][2], aus der weiteren Analyse ausgeschlossen worden. Die 54 FASTQ-Dateien sind auf ArrayExpress unter der Identifikation E-MTAB-4652 frei zugänglich.
Aus den FASTQ-Dateien wurde ein Alignment der RNAseq-Reads gegen das humane Referenzgenom (GRCh38) mit TopHat (Version 2.0.14) und STAR (Version 2.4.1d modified) berechnet. Das Ausgabe-Format dabei war BAM (siehe 2.3.1). Aus jeder FASTQ-Datei wurde in einem separaten Vorgang eine BAM-Datei erstellt.

[2] `https://arxiv.org/abs/1405.0114`

Analyse der differenziellen Gen Expression Die Identifikation von Differential Expressed Genes (DEG) wurde in R mit den Bioconductor Paketen DESeq2 und edgeR durchgeführt. Dabei wurden Quasi Likelihood F-Tests (Bioconductor edgeR) der Gruppen Jung gegen Alt und eine Korrektur der P-Werte für multiples Testen (Benjamini Hochberg) durchgeführt. Die Tests ergaben, dass bei keinem Gen die FDR (False Discovery Rate) kleiner als 10 % war. Die Ergebnisse dieser Analyse sind in 2017 in PLoS ONE publiziert worden [45][3]. Es wurde deshalb im Weiteren angenommen, dass die Gen-Expression in den sequenzierten Proben homogen ist.

2.2.2 Gap-Sites

Das Herausschneiden intronischer Sequenzen durch den Spleißvorgang verursacht Lücken in Alignments von mRNA gegen genomische DNA (Alignment Gaps). Eine Gap-Site ist durch Alignment-Lücken mit gleichen Koordinaten in einem oder mehreren überdeckenden Alignments definiert (Abbildung 2.2). Diese überdeckenden Alignments werden Gap-Site *definierende* Alignments genannt.
Um Gap-Site definierende Lücken in Alignments zu identifizieren und deren Koordinaten zu berechnen, muss das CIGAR-Segment einer (SAM/BAM) Alignment Struktur ausgewertet werden. Die Codierung des CIGAR Segmentes ist in der SAM-File Format Definition[4] beschrieben. Für weitere Einzelheiten dazu sei auf die Vignette des *rbamtools* Paketes verwiesen (CIGAR-OP Typ N = skipped region from the reference sequence).

Mit dieser Definition läßt sich das Ziel dieser Arbeit als „Identifikation, Annotierung und Validierung von Gap-sites in einem oder mehreren sequenzierten Transkriptomen“ formulieren.

[3] `https://doi.org/10.1371/journal.pone.0175657`
[4] `https://samtools.github.io/hts-specs/SAMv1.pdf`

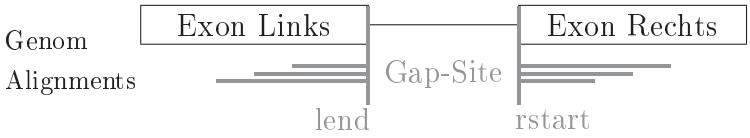


Abbildung 2.2: Gap-Site Definition

Weil die Lese-Richtung der DNA - Sequenz unbekannt ist, wird als Standard die Leserichtung von Links (left) nach Rechts (right) festgelegt. *lend* (left end) ist definiert als die Position des letzten exonischen Nukleotides auf der linken Seite. *rstart* (right start) ist definiert als die Position des ersten exonischen Nukleotides auf der rechten Seite. Die Richtungsbezeichnungen „Links“ und „Rechts“ sind durch niedrigere oder höhere Positions - Koordinaten auf der genomischen Sequenz definiert.

Zur Nomenklatur: Gap-Sites (und Spleißstellen) werden anhand der ersten beiden intronischen Nukleotide (IDIN: intronic dinucleotides) an der 5'ss und der 3'ss bezeichnet (siehe auch Position 4 und 5 in Abbildung 2.1 **A** für eine 5'ss und Position 7 und 8 in Abbildung 2.1 **B** für eine 3'ss). Eine Gap-Site auf einer typischen Spleißstelle des Major Spliceosoms würde ein IDIN-Paar entweder GT-AG (bei Leserichtung von links nach rechts, also bei Ablesung auf dem (+)-Strang) oder CT-AG (bei Leserichtung von rechts nach links, also bei Ablesung auf dem (-)-Strang) sein. Die Strang - korrigierte Schreibweise erfolgt in kursiver Schrift: Die obige Spleißstelle würde also mit *GT-AG* Gap-Site bezeichnet. *GT-AG* Spleißstellen machen mehr als 98 % der humanen Spleißstellen aus [29].

Theoretische Überlegungen zu Validierungs - Scores Mittels Validierungs - Scores sollen Gap-Sites validiert werden. Da die Alignment Information sequenziell eingelesen wird und in der Regel auf mehrere Proben (BAM-Dateien) verteilt ist, fällt die, für eine Gap-Site verfügbare Information, inkrementell an.
Dem Umstand muss neben der Struktur der Datenhaltung auch der Aufbau der validierenden Scores angepasst sein.
Die zur Validierung eingesetzten Scores müssen deshalb so konstruiert sein, dass neu hinzugefügte Information ein einmal erreichtes Niveau nicht wieder

verschlechtern. Vom mathematischen Standpunkt aus gesehen, müssen sich Validierungs-Scores bei Hinzufügen von Information also monoton verhalten. In der Konsequenz wird hierdurch die Verwendung von Mittelwert und Median zur Bewertung in Scores ausgeschlossen.

Für die Validierung selbst kommen verschiedene Vorgehensweisen in Betracht. Im Idealfall, bei dem für jede Gap-Site Gewissheit darüber besteht, ob sie Artefakt oder biologische Realität darstellt, würde man binäre Ergebnisse (0 oder 1) erhalten. Eine Modifikation davon stellen Verfahren dar die Wahrscheinlichkeiten angeben. Dazu wäre beispielsweise eine Art logistischer Regression erforderlich, was bei Gap-Site Zahlen in der Größenordnung von 10^6 mit einigem Rechenaufwand verbunden sein kann.
Der hier verfolgte Ansatz besteht darin, validierende Scores zu berechnen und ihr Verhalten an experimentellen Daten zu untersuchen. Bei einem späteren Einsatz können dann, je nach experimentellen Umfeld, individuelle Schwellenwerte als Filterkriterium gewählt werden. Die Leistungsfähigkeit des Verfahrens wird dann durch die Gap-Site differenzierenden Eigenschaften der Scores determiniert.

Alternative Strategien zur Detektion und Validierung von Gap-Sites in RNAseq Daten Für die Detektion von Spleißereignissen in Transkriptom-Daten sind in der Literatur verschiedene Ansätze beschrieben worden. Die umfassenste Strategie stellt dabei die Rekonstruktion der kompletten mRNA (transcript assembly) dar. Die gegenwärtig existierenden Verfahren (beispielsweise Augustus oder Cufflinks) scheinen allerdings mit der Komplexität eukaryotischer Transkriptome (und der, daraus resultierenen Größe der Lösungsräume) überfordert [39, 79].
Ungefilterte Gap-Sites sind mit einer hohen hohen FDR (False Discovery Rate) behaftet [23]. Beispielsweise lag die theoretische Obergrenze for *GT-AG* Gap-Sites in den 54 analysierten - Transkriptomen bei 89.6 % in TopHat Alignments und bei 42.2 % in STAR Alignments. Vor diesem Hintergrund sind schon mehrfach Strategien zur Reduktion der hohen FDR vorgeschlagen worden:

- Herausfiltern von *GT-AG* Gap-Sites [87]

- Beschränken der Alignment - Ergebnisse auf eine selektierte Teilmenge möglicher IDIN Kombinationen (wie in TopHat realisiert)

- Beschränkung von Alignment-Gaps auf solche, die einen Matching Bereich von minimal 20 Nukleotiden auf beiden Seiten aufweisen [23]

- Herausfiltern von Gap-Sites, die eine minimale Anzahl definierender Alignments aufweisen [23]

- Bewertung von Gap-Sites mit einem Regressions-Modell (basierend auf Nukleotid-Verteilung um bekannte Spleißstellen und Intron - Größe) wie in oLego implementiert [89]

Allerdings versagen diese Strategien gerade bei der Detektion von schwachen Spleißstellen, wo regulative Elemente eine wichtige Rolle spielen und die (möglicherweise) nur sporadisch genutzt werden. Hier werden flexiblere und genauere Verifikations-Mechanismen benötigt.

2.3 Implementation

Die Implementation erfolgte in zwei Schichten: Die rechenintensiven Abschnitte, wie beispielsweise die eingangs der Analyse erforderliche Extraktion von Gap-Site Daten aus BAM-Dateien, die Annotation von Gap-Sites und die Berechnung der MaxEnt und HBond Scores, wurden in C/C++ geschrieben. Die nachgelagerten Teile, eine Klassenbibliothek, in der extrahierte Gap-Site Daten sichtbar sind und die den weiteren Ablauf der Analyse abstrahiert, wurde in R implementiert. Diese Aufteilung ermöglicht eine effiziente Nutzung der Rechnerleistung und gleichzeitig, durch den nahtlosen Übergang auf eine Skript-Platform, Flexibilität bei der weiteren Verarbeitung.
Die Implementation in R-Paketen und die Integration in CRAN und Bioconductor Paketquellen [69, 42] stellt die Verfügbarkeit der Software sicher und ermöglicht Benutzern eine einfache Installationen auf vielen Rechner-Architekturen (Linux, Windows und Apple). Darüber hinaus bot R schon zu Projektbeginn ein leistungsfähiges Umfeld für statistische Berechnungen und

Visualisierung und mit Bioconductor eine umfangreiche bioinformatische Infrastruktur an. Nicht zuletzt spielte bei der Auswahl der Analyse-Platform auch die freie Verfügbarkeit von R eine Rolle.

2.3.1 C/C++ Implementationsebene

SAMtools SAM (Sequence Alignment/Map) Format ist der Standard für die Speicherung von Nukleotid-Sequenz Alignments. BAM speichert die gleiche Information in komprimierter Form unter Verwendung von BGZF (Block GNU Zip Format). Das SAM/BAM File Format wurde 2009 beschrieben [53] und gleichzeitig wurde mit SAMtools eine Open Source Referenzimplementation in C zur Verfügung gestellt[5]. SAMtools ist als eigenständige Software lauffähig, aus den (Linux) Debian Paketquellen direkt installierbar und stellt eine umfangreiche Schnittstelle zu SAM/BAM Dateien bereit. Eine direkte Berechnung von Gap-Site Koordinaten ist damit alleine aber nicht möglich.
Der gesamte Datenzugriff auf BAM-Dateien wird über den Quellcode der SAMtools Bibliothek abgewickelt, der als statische Kopie eingebunden ist.

Datenvolumina Das sequenzierte Transkriptom einer einzelnen Probe ist eine FASTQ Datei von etwa 40 - 50 Gigabyte Größe (10-15 Gigabyte im komprimierten Zustand) und enthält etwa 100 bis 150 $\times 10^6$ Reads. Etwa 30 % der Reads enthielten Alignment-Gaps, also etwa $30 - 50 \times 10^6$ pro Probe. Da Daten von etwa 60 Proben zusammen analysiert werden sollten, bedeutete die Zielvorgabe die Analyse von 3×10^9 Alignment-Gaps aus $6 - 9 \times 10^9$ Reads.

Vorläufige Studien Als problematisch erwies sich die Anforderung, dass alle Alignment-Gaps identifiziert werden sollten, auch wenn sie nicht auf annotierten Positionen lokalisiert sind. Das bedeutete, dass die Anzahl der Alignment-Gaps und der Gap-Sites vorher nicht bekannt ist und gemeinsame Alignment-Gaps nur durch paarweisen Vergleich identifiziert werden können. Gleichzeitig war es aufgrund der Datenvolumina nicht möglich, alle benötigte Information gleichzeitig im Arbeitsspeicher vorzuhalten.

Der erste implementierte Ansatz bestand deshalb aus einer ersten, in C++

[5] `http://samtools.sourceforge.net/`

implementierten Ebene, in der Alignment-Gap Koordinaten berechnet und in SQLite Datenbanken abgelegt wurden. Dazu wurden die Quellcodes von SAMtools und SQLite in ein lauffähiges C++ Programm eingebunden, das Alignment-Gap Koordinaten in einer Datebank auf Festplatten speichert. Mit dieser Architektur konnten etwa 50.000 Alignments pro Sekunde analysiert werden, was Laufzeiten von etwa 30 bis 50 Minuten pro Probe bedeutete. Eine geeignete Indizierung ermöglichte dann die effiziente Zusammenfassung von Alignment-Gaps zu Gap-sites. Die zweite Ebene der Analyse wurde dann in R implementiert und Gap-Site Daten über die SQLite Schnittstelle RSQLite [66] eingelesen.

Die Implementation dieses Lösungsansatzes wurde in drei Versionen (2010 bis 2012) realisiert. SQLite erwies sich dabei als nutzbare Analyse-Platform für nachgelagerte Analysen. Allerdings hatten die Alignment-Gap Datenbanken eine Größe von 40 - 50 Gigabyte pro Probe was sich unhandlich für Kopiervoränge erwies und das Vorhalten verschiedener Testversionen schließlich zu Engpässen bei Festplatten-Speicherplatz führte. Es war außerdem noch nicht absehbar, in wie weit sich die Datenbankgröße als hinderlich für die Routinehandhabung auswirken würde.
Mit den vorhandenen Vorerfahrungen und einigen darüber hinaus gehenden Überlegungen erschien es sinnvoll, einen neuen Ansatz zu implementieren.

Arbeitsspeicher-basierter Ansatz Aufgrund der Vorerfahrung mit datenbankbasierter Analyse konnte die Anzahl der Gap-Sites auf eine Größenordnung von $10^5 - 10^6$ geschätzt werden, eine Menge, die problemlos im Arbeitsspeicher gehalten werden kann. Da BAM-Dateien nach Alignment-Start Position sortiert werden können, ergibt sich die Möglichkeit, Gap-Site Koordinaten in Linked-Lists abzuspeichern und Suchvorgänge für die Insertion immer vom Ende der Liste aus zu beginnen. Auf diese Weise genügt bei den meisten Insertionen ein einziger Vergleich von Gap-Site-Koordinaten. So generierte List-Strukturen sind automatisch nach Koordinaten sortiert, was auch das Zusammenführen (Merge-Vorgang) zweier Listen mittels einer einzigen Iteration erlaubt.
Gap-Site Strukturen Da nur Gap-Site-Strukturen und keine Alignment-Daten

im Arbeitsspeicher gehalten werden, müssen bei der Prozessierung von Alignment-Daten neben den Koordinaten auch die für die Validierung benötigten Informationen mitgeführt werden. Für Gap-Site Strukturen wird außerdem auch ein Merge-Vorgang benötigt.
Gap-Site-Strukturen, Linked-Lists und Merge-Vorgänge wurden in C implementiert. Grundlage dieser Entscheidung war, dass sowohl SAMtools als auch R in C geschrieben sind und in R eine dokumentierte C Schnittstelle existiert. Die C-Schnittstelle von R erlaubt es, über Funktionsaufrufe in C die Datenextraktion aus BAM zu steuern und Ergebnisse der Analyse als native R Objekte (beispielsweise data.frame) sichtbar zu machen. Weil sich eine Implementation von Linked-Lists in C leicht realisieren läßt, wurde dabei auf eine Verwendung von C++ Bibliotheken verzichtet.
Durch diese Implementation konnte der Speicherbedarf im Arbeitsspeicher und auf der Festplatte auf wenige Gigabyte reduziert werden. Gleichzeitig steigerte sich die Verarbeitungsgeschwindigkeit auf etwa 10^6 Alignments pro Sekunde, was gegenüber dem datenbankbasierten Ansatz eine Steigerung um den Faktor 20 bedeutete.

Implementationsstandards auf CRAN und Bioconductor Die Plattformen CRAN und Bioconductor fordern teilweise sehr spezifische Implementationsstandards für alle Pakete ein. Damit das Paket *rbamtools* auf CRAN akzeptiert wurde und im weiteren Verlauf auch dort bleiben konnte, musste der Quellcode von SAMtools umfangreich überarbeitet werden. Aus dem Quellcode mussten beispielsweise abort- und exit- Aufrufe durch Fehlermeldungen in R ersetzt werden, weil sonst bei Fehlern direkt auch die ganze aufrufende R-Instanz terminiert wird. Tiefere Eingriffe in die SAMtools Bibliothek wurden erforderlich, als sich heraus stellte, das der SAMtools Quellcode zu Misalign Errors führt. Wenn Compiler diese Fehler nicht automatisch korrigieren, kann auch daraus ein Programmabbruch resultieren (siehe auch Kapitel 3 in Supplemental Material von *Manuskript 1* [47] hierzu) .

2.3.2 R Implementationsebene

Eine Implementation von Klassenbibliotheken in R ermöglicht die Abstraktion der Datenanalyse in intuitiv und sicher bedienbare Entitäten. Da während einer

Analyse von Gap-Sites viele spezielle Bearbeitungsschritte durchlaufen werden müssen, fächert sich die dafür benötigte Klassenbibliothek in viele Bestandteile auf, die (thematisch getrennt) auf drei R-Pakete aufgeteilt wurden. Das Paket *rbamtools* enthält basale Routinen für den Zugriff auf BAM-Dateien, *refGenome* enthält basale Routinen für die Datenhaltung von Annotationsdaten und *spliceSites* enthält Routinen, die auf biologische Fragestellungen (alternatives Spleißen, MaxEnt-Scores) zentriert sind.
Einzelheiten der Funktionalität können für einzelne Funktionen in Form der Hilfe in R abgerufen werden. Alle drei Pakete enthalten zusätzlich eine Vignette, in der typische Analyseabläufe anhand von Beispieldaten exemplarisch nachvollziehbar sind.

Kapitel 3

Manuskript 1

3.1 Titel und Inhalt

Manuskript 1 wurde unter dem Titel *rbamtools: an R interface to samtools enabling fast accumulative tabulation of splicing events over multiple RNA-seq samples* in Bioinformatics im Jahre 2015 veröffentlicht.
Das Manuskript erschien als „Applications Note“, ein Format in dem kurze Beschreibungen neuartiger Software erscheinen. Dementsprechend beschreibt das Manuskript die Struktur der im *rbamtools* Paket implementierten SAMtools Schnittstelle, die Zugriffe auf BAM-Dateien aus R heraus ermöglicht. Darüber hinaus wird anhand von Beispielen die Analyse von Spleißereignissen in RNA-seq Daten dargestellt und der Begriff der Gap-Sites erklärt.

3.2 Beiträge zum Manuskript

WK ist Autor und Maintainer des *rbamtools* Paketes und hat das Manuskript geschrieben. HScha hat die biologischen Experimente durchgeführt, die Sequenzierungs-Daten zur Verfügung gestellt und die biologische Plausibilität der Ergebnisse geprüft. WK, HScha und HSchw haben das Manuskript revidiert.

3.3 Manuskript

Bioinformatics, 31(10), 2015, 1663–1664
doi: 10.1093/bioinformatics/btu846
Advance Access Publication Date: 5 January 2015
Applications Note

OXFORD

Genome analysis

rbamtools: an R interface to samtools enabling fast accumulative tabulation of splicing events over multiple RNA-seq samples

Wolfgang Kaisers[1,*,†], Heiner Schaal[1,2,†] and Holger Schwender[1,3]

[1]Center for Bioinformatics and Biostatistics, BMFZ, Heinrich Heine University Düsseldorf, [2]Institut für Virologie, and [3]Mathematical Institute, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: John Hancock

Received on July 31, 2014; revised on December 18, 2014; accepted on December 19, 2014

Abstract

Summary: The open source environment R isf the most widely used software to statistically explore biological data sets including sequence alignments. BAM is the de facto standard file format for sequence alignment. With rbamtools, we provide now a full spectrum of accessibility to BAM for R users such as reading, writing, extraction of subsets and plotting of alignment depth where the script syntax closely follows the SAM/BAM format. Additionally, rbamtools enables fast accumulative tabulation of splicing events over multiple BAM files.
Availability and implementation: rbamtools is available on CRAN and on R-Forge.
Contact: kaisers@med.uni-duesseldorf.de
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The samtools format specifies various data slots for sequence alignments many of which are difficult to understand when sequencing experiments are to be analyzed. For analysis of sequencing data, detailed access to contents of BAM files is needed, especially when technical problems arise. *rbamtools* allows R users to investigate alignment results by reading the header section or retrieve and view alignments from regions of interest using basic R structures.

rbamtools provides functions for creation and modification of BAM file header or alignment section contents. *rbamtools* also facilitates writing of BAM files which is not possible in Bioconductor (Gentleman *et al.*, 2004; Morgan *et al.*, 2010).

Additionally *rbamtools* contains a framework for sequential and fast extraction of alignment gap positions (see Table 1) on RNA-seq data which are candidate sites for true splicing events. *rbamtools* is part of an analysis pipeline for analysis of splicing events in RNA-seq data which consists of three R packages: *rbamtools* and refGenome (Kaisers, 2013a) and spliceSites (Kaisers, 2013b).

The identification of splicing inaccuracies is a non trivial task on BAM files, since the positions of alignment gaps must be accounted on billions of reads. With *rbamtools*, processing data from, e.g. 60 RNA-seq samples (containing 8.37×10^9 alignments) can be done in 1.75 h on a standard workstation with minimal working memory demand.

Current versions of the samtools C library contain misalignment (bus) errors (http://en.wikipedia.org/wiki/Bus_error), which may cause program crashes on some architectures (e.g. SPARC). In *rbamtools*, these misalignment errors are corrected (see Supplementary Material).

2 Approach

2.1 Implementation

The package consists of three layers: the samtools C library, C based containers for alignments and alignment gaps as well as an S4 class library in R providing the user interface.

The samtools C library is a static copy of samtools (v1.4-r985). In order to meet CRAN policies, numerous changes had to be introduced into the source code (B.Ripley and K.Hornik, personal communication).

Table 1. Example for a gap site

Exon	Intron	Exon	Position	CIGAR
AG		CCTTGATG	3	2M6N8M
CAG		CCTTGAT	2	3M6N7M
CCAG		CCT	1	4M6N3M
CCCAG	GTCCAG	CCTTGATGTCC	(reference)	

A gap site defined by three alignments which share the same alignment gap site. The position values are 0-based (as described in the SAM file format[a]). The last row represents the (chromosomal) reference sequence.

[a]http://samtools.github.io/hts-specs/SAMv1.pdf

2.2 User interface

The S4 class library closely reflects the internal structures of BAM files. In order to provide detailed access to BAM file content, the API provides 14 classes and numerous functions.

Basic accessors

Basic accessors provide access to all parts of raw file content, header section and alignments for reading and writing. The following example opens the BAM file bam and copies alignments on chromosome 1 into a second BAM file.

```
rd <- bamReader(bam, idx=TRUE)
rg <- bamRange(rd, getRefCoords(rd, "chr1"))
wr <- bamWriter(getHeader(rd, "chr1.bam")
bamSave(wr, rg, refid=0)
```

Inspecting ranges can be useful when downstream analysis indicates regions without any alignments or other technical flaws.

Specialized analysis routines

Specialized analysis routines for visualization of phred score distribution as well as for tabulation of nucleotide content and calculation of GC content and AT/GC ratio are provided.

Alignment depth

This information can be retrieved from a *bamRange* object. Figure 1 shows an example where alignment depth is plotted for Gene CHMP2A (ENSG00000130724).

```
ad <- alignDepth(range)
plotAlignDepth(ad)
```

Gap sites

Gap sites are kept in containers of class `bamGapSite`. The data is gathered from BAM files using the `bamGapList` function which directly operates on `bamReader` objects as shown below.

```
bgl1 <- bamGapList(bamReader(bam1, idx=TRUE)
bgl2 <- bamGapList(bamReader(bam2, idx=TRUE)
bgl  <- merge(bgl1, bgl2)
```

Gap site positions and numbers of crossing read alignments can be obtained from multiple BAM files as data.frame by executing:

```
gap <- readPooledBamGapDf(fileNames)
```

The algorithm processes 1 196 149± 536 alignments per second or 4.3 billion alignments per hour. The data inside `bamGapSites` objects can directly extracted into a `data.frame`. For each gap site, alignment (read) counts are provided which can be used for differential expression analysis and for differential splicing analysis. Gene annotation can be added by using a specialized annotation procedure for gap sites provided by the CRAN refGenome package. Further information on gap sites for example identification of non canonical splice sites, MaxEnt (Yeo and Burge, 2004) and HBond (Freund *et al.*, 2003) scores as well as information on alternative splicing events can be obtained using the Bioconductor spliceSites package.

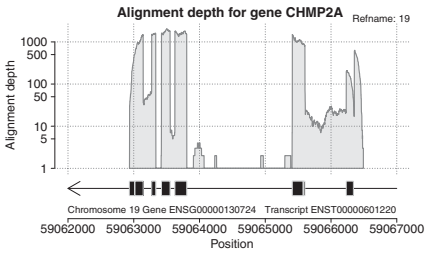


Fig. 1. Number of alignments per genomic position for gene CHMP2A

Application of these rbamtools functions to data from an RNA-seq experiment on 60 human fibroblast samples resulted in 115 968 gap sites which are present in all samples. Thereof, 98.1 % exactly lie on annotated (Ensembl Release 74) splice sites while 1.98 % (2210 gap sites) are located on not yet annotated positions.

Acknowledgements

We thank the R Core Team (R Core Team, 2014), and in particular Profs. Brian Ripley and Kurt Hornik, very much for their fruitful comments and for their valuable help in improving *rbamtools*.

Funding

RNA-seq datasets were obtained from RNA deep sequencing projects which were partly funded by the German Ministry of Research and Education (Network Gerontosys), DFG [SCHA 909/3-1], the Heinz Ansmann Foundation for AIDS Research, Düsseldorf, Germany (He.S), and the Jürgen Manchot Stiftung (H.S.). The financial support of the Deutsche Forschungsgemeinschaft [SCHW 1508/3-1 to H.S.] is gratefully acknowledged.

Conflict of Interest: none declared.

References

Freund,M. *et al.* (2003) A novel approach to describe a u1 snrna binding site. *Nucleic Acids Res.*, **31**, 6963–6975.

Gentleman,R.C. *et al.* (2004). Bioconductor: Open software development for computational biology and bioinformatics. *Genome Biol.*, **5**, R80.

Kaisers,W. (2013a) *refGenome: Gene and Splice Site Annotation Using Annotation Data From Ensembl and UCSC Genome Browsers*. CRAN R package version 1.3.0.

Kaisers,W. (2013b) *spliceSites: A Bioconductor Package for Exploration of Alignment Gap Positions from RNA-Seq Data*. Bioconductor R package version 1.3.3.

Morgan,M. *et al.* (2010) *Rsamtools: Binary alignment (BAM), variant call (BCF), or tabix file import*. R package version 1.16.1.

R Core Team (2014) *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Yeo,G. and Burge,C. (2004). Maximum entropy modeling of short sequence motifs with applications to rna splicing signals. *J. Comput. Biol*, **11**, 377–394.

Kapitel 4

Manuskript 2

4.1 Titel und Inhalt

Manuskript 2 wurde unter dem Titel *Validation of Splicing Events in Transcriptome Sequencing Data* im International Journal of Molecular Sciences im Jahr 2017 veröffentlicht.
Manuskript 2 beschreibt einleitend die Definition von Gap-Sites, die Definition der Gap-Site Validierungs Scores (*gqs* und *wgis*) und den Mechanismus der Annotation von Gap-Sites. In einem zweiten Abschnitt werden dann die Eigenschaften der aus den experimentellen Daten extrahierten Gap-Sites beschrieben. Dazu gehören die Verteilung der Validierungs-Scores, die Festlegung von Filterkriteren für den Validierungsvorgang und die Eigenschaften der validierten Gap-sites.
Es zeigte sich, dass in Alignments von beiden Alignern (TopHat und STAR) sich ein großer Prozentsatz unvalidierter Gap-Sites findet. Für den *gqs* Ansatz (der ausschließlich auf Alignment Daten beruht) zeigte sich, dass eine große Anzahl Alignments (Sequenzier-Tiefe) erforderlich ist, um eine hohe Rate validierter Gap-Sites zu erhalten. Durch Einschluss der Spleißstellenstärke als Validierungs-Kriterium kann diese Zahl deutlich (etwa um den Faktor 10) reduziert werden.

4.2 Beiträge zum Manuskript

WK ist Autor und Maintainer der R-Pakete, hat die bewertenden Scores entwickelt, die Analyse der RNA-seq Daten durchgeführt (Alignment, Gen-Expression), die statistische Analyse für die Score -Validierung durchgeführt und das Manuskript geschrieben. HScha und JP haben die biologischen Experimente durchgeführt. HScha hat die Sequenzierungs-Daten zur Verfügung gestellt und die biologische Plausibilität der Ergebnisse geprüft. WK, HScha, JP und HSchw haben das Manuskript revidiert.

4.3 Manuskript

International Journal of
Molecular Sciences

MDPI

Article

Validation of Splicing Events in Transcriptome Sequencing Data

Wolfgang Kaisers [1,2,*], Johannes Ptok [3], Holger Schwender [2,4] and Heiner Schaal [2,3]

1 Department for Anaesthesiology, University Hospital Düsseldorf, Heinrich Heine University, 40225 Düsseldorf, Germany
2 BMFZ (Biologisch-Medizinisches Forschungszentrum), Heinrich Heine University, 40225 Düsseldorf, Germany; schwender@math.uni-duesseldorf.de (H.S.); schaal@uni-duesseldorf.de (H.S.)
3 Institute of Virology, University Hospital Düsseldorf, Heinrich Heine University, 40225 Düsseldorf, Germany; Johannes.Ptok@uni-duesseldorf.de
4 Mathematical Institute, Heinrich Heine University, 40225 Düsseldorf, Germany
* Correspondence: kaisers@med.uni-duesseldorf.de; Tel.: +49-211-12393

Academic Editor: Li Lin
Received: 5 April 2017; Accepted: 28 April 2017; Published: 23 May 2017

Abstract: Genomic alignments of sequenced cellular messenger RNA contain gapped alignments which are interpreted as consequence of intron removal. The resulting gap-sites, genomic locations of alignment gaps, are landmarks representing potential splice-sites. As alignment algorithms report gap-sites with a considerable false discovery rate, validations are required. We describe two quality scores, gap quality score (*gqs*) and weighted gap information score (*wgis*), developed for validation of putative splicing events: While *gqs* solely relies on alignment data *wgis* additionally considers information from the genomic sequence. FASTQ files obtained from 54 human dermal fibroblast samples were aligned against the human genome (GRCh38) using TopHat and STAR aligner. Statistical properties of gap-sites validated by *gqs* and *wgis* were evaluated by their sequence similarity to known exon-intron borders. Within the 54 samples, TopHat identifies 1,000,380 and STAR reports 6,487,577 gap-sites. Due to the lack of strand information, however, the percentage of identified GT-AG gap-sites is rather low. While gap-sites from TopHat contain ≈89% GT-AG, gap-sites from STAR only contain ≈42% GT-AG dinucleotide pairs in merged data from 54 fibroblast samples. Validation with *gqs* yields 156,251 gap-sites from TopHat alignments and 166,294 from STAR alignments. Validation with *wgis* yields 770,327 gap-sites from TopHat alignments and 1,065,596 from STAR alignments. Both alignment algorithms, TopHat and STAR, report gap-sites with considerable false discovery rate, which can drastically be reduced by validation with *gqs* and *wgis*.

Keywords: splice sites; RNA-seq; TopHat; STAR; MaxEnt

1. Introduction

Analysis of transcriptome sequencing data focuses on differential expression of genes, as well as alternative splicing. Genomic alignments of transcriptome sequencing data contain alignment gaps. Alignment gaps are landmarks indicating potential splice-sites. Thus, due to the complexity of eukaryotic transcriptomes, transcript reconstruction from sequenced mRNA encompasses considerable ambiguities. Low specificity of reported alignment gaps may seriously compromise validity of analysis results. Currently, transcript reconstruction algorithms suffer from inaccuracies [1,2] and even the (much simpler) identification of splice events is associated with a high false discovery rate (FDR) [3]. Additionally, lack of strand information makes assignment of the correct strand difficult.

The FDR can be reduced by additional validations. Early and simplifying "validation" procedures include restriction to (for example) GT-AG-sites [4] and filtering for minimal exonic match length or

filtering for minimal number of supporting alignments (as reported by RGASP, RNA-seq Genome Annotation Assessment Project [3]).

Recently 184 splice sites with non-canonical dinucleotides and U2/U12 like consensus sequences have been reported [5] and therefore, filter based on intronic dinucleotides are likely to be inappropriate. The rationale for usage of quality scores is to differentiate between biology (for example regulated splicing events) and artefacts (for example stochastic noise in splicing, sequencing and alignment). In the following, two approaches for validation of splicing events in transcriptome sequencing data are described and evaluated: An approach solely relying on alignment data (*gqs*) and a second approach which additionally includes information from genomic sequence (*wgis*).

1.1. Genomic Alignments and Gap-Sites

The following section describes the definition of gap-sites and the parameters collected from alignment data. Alignment results are reported in BAM files and structure of genomic alignments is described in BAM file format [6] (see also Appendix A.1). Alignment data contained in BAM files (currently) does not include strand information because it is lost during sequencing [7]. In genomic alignments of sequenced cellular mRNA, reads crossing splice sites cover at least two exons which are separated by an intron and thus result in gapped alignments (Figure 1). Alignment gap locations possibly shared by multiple alignments define a gap-site. Assuming that alignment gaps result from splicing events, gap-site positions are candidates for splice junctions. A gap-site is characterised by two genomic positions: *lend* and *rstart* and the alignments are called "supporting alignments". The number of supporting alignments is called alignment coverage or nAligns-value for the respective gap-site. The number of samples in which a gap-site is identified is called the multiplicity (nProbes) value for this gap-site.

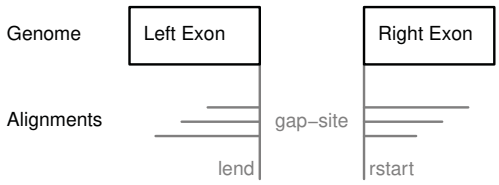


Figure 1. Definition of gap-sites. Since no strand information is provided by alignments, gap-site positions are defined reading direction of genomic sequence (left to right). *lend* is defined as position of last exonic nucleotide on the left side. *rstart* is defined as position of first exonic nucleotide on right side. The "left" and "right" side are directions on genomic reference sequence, defined by lower and higher position coordinates respectively relative to the actual reading position (in accordance with common reading orientation). The gap-site is covered by three alignments (nAligns = 3).

For single alignments covering gap-sites, the minimum length of both (exonic) nucleotide matching regions (the minimum CIGAR length value or *mcl* value, described in Appendix A.1.3) is included into *gqs*. The *mcl* criterion provides a lower limit for support of an alignment gap by subsequent nucleotide matches. Two other alignment derived values, quartet sum of *mcl* (=*qsm*) and number of distinct (left) alignment start positions (*nlstart*) are described in detail in Sections Appendixs A.1.4 and A.1.5 respectively.

On the plus-strand the splice-site donor (or 5′ splice-site) is located at *lend* position and the splice-site acceptor (or 3′ splice-site) is located at *rstart* position (Figure 1). On the minus-strand the opposite rule applies.

1.1.1. Intronic Genome Sequence at Gap-Sites

Statistical properties of gap-sites are evaluated using distribution of intronic dinucleotides (IDIN, for example GT) and IDIN-pairs (for example GT-AG) and sequence logos. Therefore, the two subsequent nucleotides following the *lend* position and preceding the *rstart* position are analysed.

The location of IDIN is called "left" or "right" as long as no strand information is considered or available. Using strand information exon-intron boundaries can further be classified into 5′- and 3′-splice-sites. For "+"-strand, the raw left IDIN are from 5′ splice-junctions (called 5′IDIN) and the raw right IDIN are from 3′ splice-sites (called 3′IDIN). For "−"-strand, the reverse-complement of right IDIN are from 5′ splice-junctions (called 5′IDIN) and the reverse-complement of left IDIN are from 3′ splice-junctions (called 3′IDIN).

Gap-sites are referred to by 5′IDIN (for example GT-sites) or by IDIN-pairs (for example GT-AG-sites or AT-AC-sites). From here on, IDIN and IDIN-pairs in strand-corrected orientation (5′ to 3′) are shown in italic letters. Non cursive printed nucleotides represent uncorrected genomic sequence (left to right).

1.1.2. General Considerations on Quality Scores

Recognition of rare events (for example non-canonical splicing) requires high sensitivity and therefore data is merged from multiple samples and large amounts of alignment data. Merging of data from multiple samples and the structure of BAM files implicate that validating information on gap-sites emerges sequentially. Therefrom, two basic requirements for gap-site scores derive:

- Monotonicity: By adding new alignment information, the score must not decrease and should increase with the number of alignment matches to genomic nucleotides. Due to monotonicity, mean or median values cannot be used. This criterion precludes that an initially high score decreases when low scoring alignments are added.
- Informational limit: Collection of information for each gap-site is restricted to a static size where splicing events are sufficiently confirmed. Using this limit, data from multiple alignments can be packed into integer variables allowing fast processing without necessity of specialised data containers.

1.2. Definition of Gap Quality Score (gqs)

The Gap Quality Score (*gqs*) solely uses information extracted from genomic alignments and therefore can directly be calculated from (multiple) BAM files. As shown in Equation (1), *gqs* essentially consists of two factors: *qsm* and *nlstart*. A calculated example is given in Appendix (Table A1). The *gqs* is distributed between 0 and 1000 (for read length of 100 and on a 64 bit operating system). A score of 1000 requires at least 8 supporting alignments and four maximal *mcl* values (at least 50 matching nucleotides on both sides in at least 4 alignment gaps for *qsm*= 200). Gap-sites attaining a *gqs* of 1000 are called "*gqs*-validated" (see Figure 6b for empirical base of definition).

The structure of *gqs* implies that the probability of being *gqs*-validated increases with higher alignment coverage.

The *gqs* is defined as:

$$gqs = 10 \times \frac{nlstart}{n} \quad \frac{2 \times qsm}{4}. \tag{1}$$

The number of different alignment start positions for a gap-site denoted *nlstart* value (see Appendix A.1.5). *qsm* is the sum of the four largest minor alignment match lengths beside the alignment gap (see Appendix A.1.4). n is the number of bytes in an integer ($n = 4$ on a 32 bit system and $n = 8$ on a 64 bit operating system). The gqs additionally is truncated to integral values.

1.3. Definition of Weighted Gap Information Score (wgis)

The Weighted Gap Information Score (*wgis*) is a successor of *gqs* because of the *gqs* being too insensitive on gap-sites supported by only few alignments. From alignments, the *wgis* basically utilises the same input values as *gqs* (*qsm* and *nlstart*). Additionally, the similarity of the genomic sequence with splice-sites is evaluated using MaxEnt scores: *score5* for 5′ splice-sites and *score3* for 3′ splice-sites [8,9]. All factors are weighted (using log2). Thresholds are applied to *qsm*, *score5* and *score3*. Information on empirical base for thresholds is provided in supplemental material. For gap-sites with any quality below a given threshold, $wgis = 0$ is returned. All sites with $wgis \neq 0$ are called "*wgis*-validated". Strand information is reported by *wgis* via signature. The detailed definition of *wgis* is given in Equation (2).

The *wgis* is defined as:

$$wgis = f_{nls} \times f_{qsm} \times f_{s5} \times f_{s3} \times s_{str}. \quad (2)$$

Factor	Name	Source	Definition	Threshold	Maximal Value
f_{nls}	*nlstart*-factor	BAM-file	$\log_2(\log_2(nlstart)+1)+1$		3
f_{qsm}	*qsm*-factor	BAM-file	$\log_2(\log_2(\max(qsm-13,2)))$	15	2.916
f_{s5}	*score5*-factor	MaxEnt	$\log_2(\max(score5,1))$	1	3.56
f_{s3}	*score3*-factor	MaxEnt	$\log_2(\max(score3,1))$	1	4.01
s_{str}	strand-sign	MaxEnt			1

The BAM-file derived factor f_{nls} is positive (>0). The BAM-file derived factor f_{qsm} and the two MaxEnt score factors (f_{s5}, f_{s3}) are non-negative (≥ 0). Therefore *wgis* = 0 exactly when $qsm \leq 15$ or $score5 \leq 1$ or $score3 \leq 1$ (threshold). The sign of *wgis* is determined by the strand-sign $sgn(wgis) = sgn(s_{str})$. Maximal observed MaxEnt scores on annotated splice-sites are 11.8 for *score5* and 16.1 for *score3* leading to a theoretical maximum of 124.9 for *wgis* (on a 64 bit operating system).

1.3.1. Sequence Similarity between Gap-Sites and Splice-Sites

The probability of a gap-site being a true splice-site is rated using the MaxEnt score developed by Burge et al. The MaxEnt score uses a log-odd ratio (defined as ratio of sequence motif frequency in true splice sites and decoy). Sequence 9-mers and 23-mers are analysed for evaluation of 5′ and 3′ splice-sites respectively (positions −3 to +6 for 5′ sites and −20 to +3 for 3′ sites). The *maximum-entropy* is an estimation procedure for probability distributions originally developed in statistical mechanics [10,11]. In an iterative approach, the maximum entropy distribution has been estimated in advance. Scores are extracted from pre-calculated tables based on sequence related indices. In their paper, Burge et.al. describe that a training data set had been derived from a set of 1821 transcripts spanning 12,715 introns (5′ and 3′ splice sites).

1.3.2. Strand Information in *wgis*

The strand sign (s_{str}) of *wgis* validated gap-sites is determined by evaluation of MaxEnt*score3*. Calculation of *score3* at the left exon-intron boundary (around *lend* position) in original orientation (left to right) yields *score3* in "+"-strand orientation (*score3*(+)). Calculation of *score3* at the right exon-intron boundary (around *rstart* position) in reverse-complement orientation (right to left) yields *score3* in "−"-strand orientation (*score3*(−)). The strand sign is calculated by comparison of these two values as shown in Equation (3).

$$s_{str} = \begin{cases} +1 & \text{when} \quad score3(+) \leq score3(-) \\ -1 & \text{otherwise} \end{cases} \quad (3)$$

The biological interpretation is, that when the right boundary of a gap-site has stronger resemblance to a $3'$ splice-site than the left boundary of the gap-site (read in reversed direction), then "+"-strand is assumed.

1.4. GQL

The distribution of *wgis*, naturally separates gap-sites into four sub-populations. The derived categories consist of quality levels 0 to 3 named "Gap Quality Level" (*gql*): *gql*0 to *gql*3. Gap-sites assigned to *gql*-level i are denoted *gql*i-sites. The set of *gql*0 sites for example is a s "not *wgis*-validated" ($wgis = 0$). Details of *gql* definition are described later on in Section 2.3.1.

1.5. Annotation of Gap-Sites

During annotation of query objects derived from genomic alignments, overlap with annotated genomic features is examined and (if existent) an optimal matching pair is selected. The annotation process for gap-sites is implemented in (CRAN) R-package *refGenome* (see Figure 2). First, intron-connected exon pairs need to be identified because exons are included in GTF file format (the format down-loadable from Ensembl or UCSC) as distinct entities only related by *transcript_id* and *exon_number*.

In a second step, a query list (containing gap-site data) and a reference list (containing positions of connected exons) are traversed simultaneously. During traverse, for each gap-site a region containing possible overlaps is searched for an optimal match (minimising *sod*).

Overlap is defined as the intersection of element-ranges in the query and the reference list. Element-ranges in the query list range from the first nucleotide with an alignment match (leftmost C in query3 in Table A1) to the last nucleotide with an alignment match (rightmost G in query1 in Table A1). An element-range in the reference list range from the start position of the left exon to the end position of the right exon. When no overlap is found (i.e., all intersections are empty), no annotation data is provided for a (query) gap-site (The R implementation in *refGenome* reports missing overlaps as "NA"-values (Not available).). The annotation procedure reports exact matches (sod = 0) as well as inexact matches ($sod > 0$, Figure 2). Exact matches are gap-sites residing on annotated splice sites. Inexact matches may be due to biology, for example alternative splicing events (exon skipping, intron retention or alternative donor or acceptor sites) as well as splicing errors. However, inaccuracies may also be due to technical or bioinformatic issues like sequencing or alignment errors.

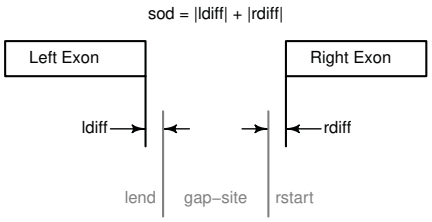


Figure 2. Annotation of gap-sites in R package *refGenome*. Annotation of a gap-sites as implemented in (CRAN) R package *refGenome*. Distance to annotated sites is expressed in *sod* (sum of distances, ≥ 0). Note that in Ensembl annotation (GTF-files) the two exons are represented as two distinct features.

1.6. Validation Strategy for Quality Scores

As no reference method is available for validating gap-sites (the biochemical RT-PCR method allows validation of small numbers of splicing events and is not feasible for genome wide analysis), a systemic approach using global statistical properties is utilised. The values which are statistically

analysed reflect the complementarity to the 5′ end of U1 snRNA or known distribution of nucleotides around splice-sites. In detail, the used criteria are:

- Distribution of IDIN (GT or AG) and IDIN-pairs (GT-AT).
- Sequence logos from 5′ and 3′ exon-intron boundaries.
- The proportion of annotated sites.

2. Results

2.1. Global Statistics

2.1.1. Alignments from TopHat Aligner

TopHat reported in total 3.0×10^9 alignment gaps containing 1,000,380 gap-sites. Thus, gap-sites are in mean covered by 2999 alignments. Therefrom, 273,994 (27.4%) gap-sites are supported by only one alignment (nAligns = 1). TopHat reports 120,434 (12.0%) gap-sites present in all 54 samples and 243,596 (24.3%) annotated splice-sites (sod = 0).

The reported left IDIN (without strand information) are CT, GT, GC and AT and the right IDIN are AG, AC, GC and AT (both decreasingly ordered by abundance). In total, TopHat reports only gap-sites with 6 (from possible 256) different IDIN-pairs (GTAG, CTAC, GCAG, CTGC, ATAC, GTAT; decreasingly ordered by abundance).

In order to provide an upper limit for percentage of *GT-AG* gap-sites, the proportions of GT-AG and CT-AG sites are added. In single samples, the observed percentages vary in the range from 95.21% to 97.34% (Figure 3). The mean proportion is 96.4% (SD = 0.46%). When multiple samples are merged, the proportion of *GT-AG* sites will decline due to accumulation of noisy observations. The proportion in 54 merged transcriptomes is 89.6%. Thus, when multiple samples are merged, the proportion of GT-AG gap-sites can be expected to vary in the range between 89% to 98%.

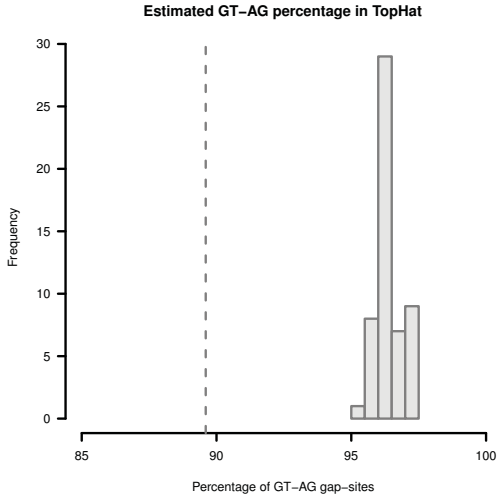


Figure 3. Estimation of *GT-AG*-site percentage in gap-sites reported by TopHat. Upper limit for percentage of *GT-AG* gap-sites (estimated by adding proportions of GT-AG and CT-AG) in TopHat alignments. Gap-sites were collected 54 times from single transcriptomes (vertical bars). Dashed vertical line (89.6%) indicates estimated percentage when gap-sites are collected from the complete batch of 54 transcriptomes.

2.1.2. Alignments from STAR Aligner

STAR reported in total 2.4×10^9 alignment gaps containing 6,487,577 gap-sites. Gap-sites from STAR aligner were supported in mean by 371 alignments. Therefrom 4,437,270 (68.4%) gap-sites are supported by only one alignment (nAligns = 1). STAR reports 129,758 (2.0%) gap-sites are present in all 54 samples and 256,044 (3.94%) annotated splice-sites. STAR reports all possible nucleotide combinations in left and right IDIN (STAR also allows gap-sites with IDIN's containing N, for example AN and NT are present in left IDIN. From left IDIN, 31 were NN and from right IDIN, 74 were NN.) .

The proportion of *GT-AG* gap-sites in single samples (estimated by adding percentages of GT-AT-sites and CT-AG-sites) vary in the range from 57.72% to 82.35% (Figure 4). The mean proportion in single samples is 72.73% (SD = 5.43%). In merged data from 54 samples, 42.2% GT-AG (or CT-AG) sites are present. Thus, when multiple samples are merged, the proportion of *GT-AG* gap-sites can be expected to vary in the range between 42% and 83%.

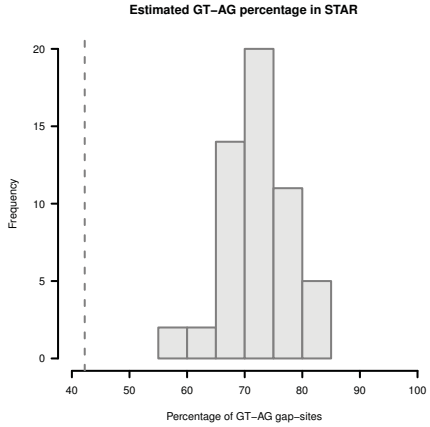


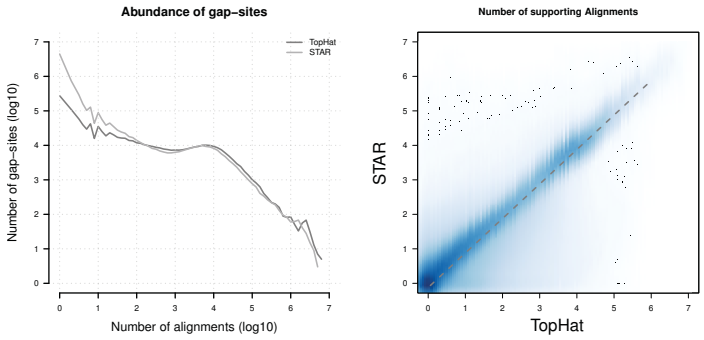
Figure 4. Estimation of *GT-AG*-site percentage in gap-sites reported by STAR. Upper limit for percentage of *GT-AG* gap-sites (added proportions of GT-AG and CT-AG) in STAR alignments. Gap-sites were collected 54 times from single transcriptomes (vertical bars). Dashed vertical line (42.2%) indicates estimated percentage when gap-sites are collected from the complete batch of 54 transcriptomes.

2.1.3. Comparison of Alignment Numbers

In general, the number of gap-sites decreases with higher alignment coverage (Figure 5a). For gap-sites with nAligns >100, the gap-site numbers essentially are distributed equally in TopHat and STAR alignments.

The STAR aligner reports substantially more gap-sites with low coverage (nAligns < 100): While TopHat reports 582,264 (58.2% of all alignments) with nAligns <10, STAR reports 5,939,689 (91.6% of all alignments) with nAligns <10, over 10 times more than TopHat (see Table 1).

There are 906,219 gap-sites which are reported by both aligners (TopHat and STAR). On these gap-sites, the nAligns numbers from STAR are approximately 76% of numbers from TopHat (Figure 5b). Thus, the in mean 10 times lower nAligns numbers in STAR alignments result only in a 24% decrease on single gap-sites and is mainly caused by a 10 times larger number of gap-sites with nAligns <100 (which are only partially also reported by TopHat).



(**a**) Distribution of nAligns in gap-sites

(**b**) Comparison of nAligns ($\log_{10}$) in gap-sites reported by both aligners (TopHat and STAR)

Figure 5. Distribution of nAligns values in gap-sites present in TopHat and STAR alignments. (**a**) Distribution of alignment coverage on gap-sites reported by TopHat and STAR; (**b**) Alignment coverage on gap-sites reported by both aligners (TopHat and STAR). The dashed line represents data from a linear regression model: $\text{nAligns}_{\text{STAR}} = 0.756 \times \text{nAligns}_{\text{TopHat}}$. Coordinates of both axes in both sub-figures are logarithmised ($\log_{10}$).

Table 1. Global statistics for TopHat and STAR aligner.

		Gap-Sites			
Aligner	**Alignment Gaps**	**Number**	**Coverage**	**Single Align**	**All Samples**
TopHat	2,999,472,708	1,000,380	2999	273,994	120,434
STAR	2,410,424,541	6,487,577	371	4,437,270	129,758

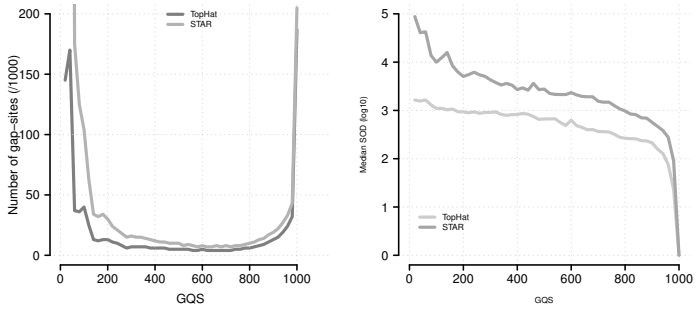
Global statistics for TopHat and STAR aligner: Alignment gaps (Total number in alignments), Gap-sites (Total number of unique gap-sites identified in 54 samples), Coverage (Alignment gaps/gap-sites = mean alignment coverage), Single align (Total number of gap-sites covered by one single alignment), All samples (Number of gap-sites present in all 54 samples).

2.2. Validation of *gqs*

2.2.1. Distribution of *gqs*

The *gqs* values distribute in a characteristic U-like shape (Figure 6a). Alignments from STAR contain more gap-sites with *gqs* <200 than alignments from TopHat.

The median *gqs* value was 195 from TopHat alignments and 12 from STAR alignments. In order to achieve a median *sod* of 0 (equivalent to: >50% of gap-sites are annotated) via the *gqs*-based filter, a *gqs* of ≥990 is needed in TopHat-alignments and *gqs* = 1000 in STAR-alignments. Thus, a threshold of *gqs* = 1000 is used as threshold for *gqs*-validation.

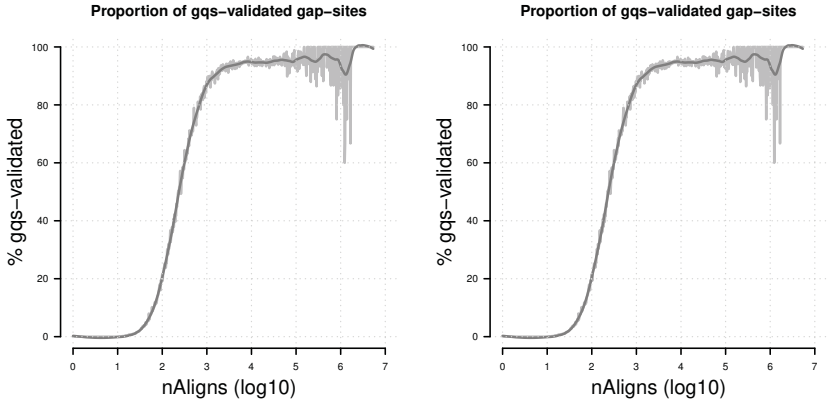


(**a**) Distribution of *gqs* values (**b**) Distribution of median *sod* values

Figure 6. Distribution of *gqs* and *sod* values. (**a**) The distribution of *gqs* values follows a characteristic U-shaped pattern. This pattern is similar to the multiplicity of events (Supplemental Data Figure 9); (**b**) Median *sod* (distance of gap-site to annotated site) values for *gqs* categories.

2.2.2. Number of *gqs*-Validated Gap-Sites

In alignments from TopHat, 156,251 gap-sites were validated by *gqs* (18.5% of all reported gap-sites). In alignments from STAR, 166.294 gap-sites were validated by *gqs* (2.6% of all reported gap-sites). The proportion of *gqs*-validated gap-sites increases uniformly with alignment depth, producing a sigmoidal increasing line on log_{10}-transformed nAligns numbers (Figure 7). For validation of more than 50% gap-sites with a coverage of more than 468 alignments TopHat alignments and more than 240 alignments STAR alignments are required.



(**a**) Alignment with TopHat (**b**) Alignment with STAR

Figure 7. Percentage of *gqs*-validated gap-sites. Proportion of *gqs*-validated gap-sites for different alignment coverage's. (**a**) Alignments from TopHat aligner; (**b**) Alignments from STAR aligner. The rising proportions of *gqs*-validated gap-sites reflects the fact that *gqs*-validation is more likely for higher alignment coverage.

2.2.3. Distribution of IDIN on *gqs*-Validated Gap-Sites

In *gqs*-validated gap-sites, the upper limit for the percentage of *GT-AG*-sites is 98.61% (49.54% + 49.07%) from TopHat alignments and 97.59% (49.13% + 48.46%) from STAR alignments (Table 2). The analogue calculation for splice-sites from the minor spliceosome (*AT-AC* sites) yields estimates of 0.25% for TopHat and 0.08% for STAR.

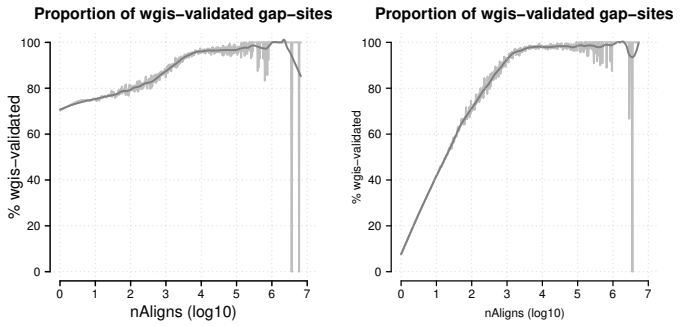
Table 2. Percentage of intronic dinucleotides on *gqs*-validated gap-sites.

Left IDIN			Right IDIN			Combined		
IDIN	**TopHat**	**STAR**	**IDIN**	**TopHat**	**STAR**	**IDIN**	**TopHat**	**STAR**
GT	49.65	49.26	AG	50.10	49.68	GT-AG	49.54	49.13
CT	49.65	49.09	AC	49.20	48.60	CT-AC	49.07	48.46
GC	0.56	0.61	GC	0.59	0.60	CT-GC	0.59	0.54
TG		0.06	GG		0.16	GC-AG	0.56	0.48
CA		0.09	CC		0.14	AT-AC	0.14	0.04
AT	0.14	0.08	AT	0.11	0.11	GT-AT	0.11	0.04

Percentage of intronic dinucleotides (IDIN) and intronic dinucleotide pairs (IDIN-pairs) on *gqs*-validated gap-sites. The shown IDIN and IDIN-pairs are not corrected for strand orientation. IDIN are counted separately on data from each aligner (TopHat and STAR). Nucleotides represent untransformed genomic sequence (no correction for strand orientation). The numbers in each column sum up to 100 (TopHat) or ≈100 (STAR).

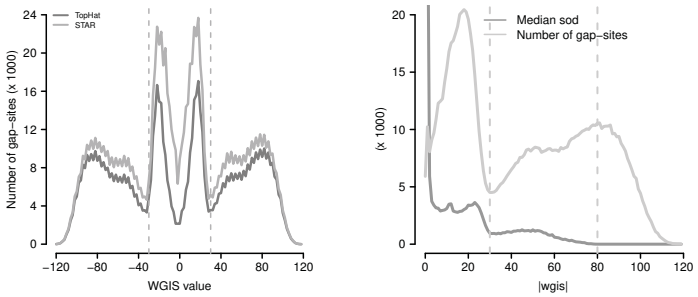
2.3. Validation of wgis

Observed values for *wgis* ranged from −119.2 to +118.7. In total, 1,083,629 gap-sites were validated by *wgis* in either TopHat or STAR alignments. In alignments from TopHat, 770,327 (77.0%) gap-sites were validated by *wgis*. In alignments reported by STAR, 1,065,596 (16.43%) gap-sites were validated by *wgis*. In general, the proportion of *wgis*-validated gap-sites increases with number of supporting alignments (nAligns, Figure 8). Using *wgis* provided strand information, 50.77% and 50.49% of *wgis* validated gap-sites were assigned to "+"-strand in TopHat and STAR alignments respectively. The *wgis*-distribution is almost identical in gap-sites assigned to "+"-strand and to "−"-strand (Figure 9).



(**a**) Alignment with TopHat (**b**) Alignment with STAR

Figure 8. Percentage of *wgis*-validated gap-sites. Proportion of *wgis*-validated gap-sites for different alignment coverage's (nAligns). Gap-sites were categorised by $\log_{10}$(nAligns) value (rounded by one digit). For each category, the proportion of *wgis* validated gap-sites were tabled. The raw proportions were smoothed using a *loess* model. (**a**) Alignments from TopHat aligner. The proportion of *wgis*-validated gap-sites is >70% throughout the whole range; (**b**) Alignments from STAR aligner. For nAligns >19 ($\log_{10}$(nAligns) >1.28), the majority of gap-sites (>50%) are validated by *wgis*.



(**a**) Distribution of nonzero *wgis* scores (in TopHat and STAR alignments).

(**b**) Distribution of median *sod* values (in STAR alignments).

Figure 9. Distribution of *wgis* and median *sod* values. (**a**) Distribution of non zero *wgis* (*wgis* values were cut into ranges of width 2 for reduction of scattering); (**b**) Distribution of median *sod* and number of gap-sites from STAR with respect to absolute wgis values ($|wgis|$). Range limits are drawn at $|wgis| = 30$ (local minimum of number of gap-sites) and at $|wgis| = 80$ (where median *sod* drops to <10).

2.3.1. Definition of GQL Limits

Parallel evaluation of median *sod* values and number of gap-sites (Figure 9b) in STAR alignments shows that *wgis*-validated gap-sites appear to be a heterogeneous population separated by two natural limits:

- A limit at $|wgis| = 30$, where number of gap-sites and *sod* have a local minimum.
- A second limit at $|wgis| = 80$, where median *sod* drops to <10 in STAR alignments (Median *sod* = 0 for $|wgis| > 75$ in TopHat alignments).

Together with the limit $|wgis| > 0$ (separating *wgis*-validated from not validated gap-sites), four different groups, called *gql* (gap-site quality level) can be separated. Definition of *gql* and proportions of assigned gap-sites are shown in Table 3.

Table 3. *gql*: Definition and distribution.

							TopHat			STAR	
gql			***wgis***			N_{total}	P_{total} (%)	P_{val} (%)	N_{total}	P_{total} (%)	P_{val} (%)
0			$\|wgis\|$	=	0	230,053	23.0		5,421,981	83.6	
1	0	<	$\|wgis\|$	≤	30	256,950	25.7	33.4	446,316	6.9	41.9
2	30	<	$\|wgis\|$	≤	80	331.955	33.2	43.1	416,316	6.4	39.1
3	80	<	$\|wgis\|$			181,422	18.1	23.6	202,964	3.1	19.1

Definition of *gql* levels and assignment of gap-sites (collected from 54 fibroblast samples using STAR aligner). N_{total}: Absolute number of gap-sites. P_{total}: Proportion of all gap sites in percent. P_{val}: Proportion of validated gap-sites in percent.

2.3.2. Distribution of Intronic Dinucleotide Pairs

Distribution of intronic dinucleotide pairs in gap-sites with different *gql* levels are shown in Table 4. Intronic dinucleotide pairs associated with the minor spliceosome (*AT-AC*) constitute <1% of *wgis*-validated gap-sites.

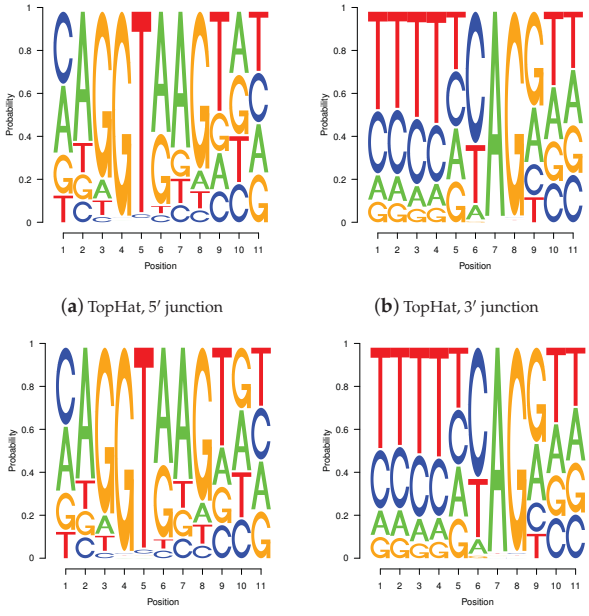
Table 4. Strand corrected intronic dinucleotide pairs of *wgis*-validated gap-sites.

IDIN-Pair	TopHat			STAR			Literature
	*gql*1	*gql*2	*gql*3	*gql*1	*gql*2	*gql*3	
GT-AG	95.70	98.71	100.00	94.76	98.66	100.00	99.24 %
GC-AG	3.83	1.24		3.38	1.15		0.7 %
GT-AT	0.23	0.03		0.16	0.02		
AT-AC	0.04			0.02			0.05%
CT-AC	0.20	0.02		0.54	0.02		

Percentage of intronic dinucleotide pairs in alignments from TopHat and STAR (IDIN-pairs not reported by TopHat not shown). Strand assignment solely bases on *wgis* (on MaxEnt *score3*). Percentage values from literature are taken from [18]. Empty spaces indicate proportion <0.005%. Obviously *GT-AT* sites and *CT-AC* sites are assigned to the wrong direction.

2.4. Sequence Logos of Validated Gap-Sites

Sequence logos of *wgis* validated gap-sites for 5′-junctions (Figure 10a,c) and for 3′-junctions (Figure 10b,d) show high similarity between TopHat and STAR alignments. The sequence logos show presence of the second GT at intronic position 5 and 6 in 5′ splice-junctions (positions 8 and 9 in Figure 10a,c) as well as pyrimidine rich 3′ terminal intronic regions (positions 1 to 6 in Figure 10b,d). The sequence logos are also highly similar to those calculated on annotated splice sites (shown in supplemental material).



(**a**) TopHat, 5′ junction (**b**) TopHat, 3′ junction

(**c**) STAR, 5′ junction (**d**) STAR, 3′ junction

Figure 10. Sequence logos for *wgis* verified gap-sites. Sequence logos for *wgis* verified gap-sites (GQL > 0). Tabled nucleotides are corrected for strand orientation reported by *wgis*. (**a**,**c**) Splice-junction is located between position 3 and 4 (gap-site *lend* is located at position 3); (**b**,**d**) Splice-junction is located between position 8 and 9 (gap-site *rstart* is located at position 9). Sequence logos for *wgis* validated gap-sites show a high degree of similarity in alignments from TopHat and STAR.

2.5. Comparison of TopHat and STAR Alignments and gqs and wgis Validation

For comparison of validated gap-sites from TopHat and STAR aligner, tables with (*gqs* and *wgis*) validated gap-sites from TopHat and STAR were merged (using genomic coordinates as identity criterion).

2.5.1. Global Statistics

In total, 6,581,738 gap-sites were identified by either TopHat or STAR in the 54 fibroblast samples. Therefrom, 5,581,358 (84.8%) are only present in STAR alignments, 906,219 (13.8%) are identified by both aligners and 94,161 (1.43%) are solely present in TopHat alignments (Table 5). From alignments reported exclusively by STAR, exclusively by TopHat or by both aligners, 97.6%, 82.9% and 65.0% of gap-sites are not validated by either *gqs* or *wgis*.

In result, STAR as well as TopHat do report gap-sites not seen by the other aligner. Some of these may even be validated by both scores.

Table 5. Validation numbers of gap-sites.

Aligner	Not Validated	GQS	WGIS	GQS & WGIS	Sum
STAR	5,449,996	5548	125,217	597	5,581,358
STAR & TopHat	587,008	22,798	161,127	135,286	906,219
TopHat	78,044	680	15,164	273	94,161
Sum	6,115,048	29,026	301,508	136,156	6,581,738

STAR: Gap-sites identified only by STAR, TopHat: Gap-sites identified only by TopHat, STAR & TopHat: Gap-sites identified by both aligners. GQS: Gap-sites validated only by *gqs*, WGIS: Gap-sites validated only by *wgis*, GQS & WGIS: Gap-sites validated by *gqs* and *wgis*. Number of gap-sites for different validation types and presence in alignments. For gap-sites reported by both aligners, validation numbers were taken from STAR alignments. For these gap-sites, nAligns values are in mean ≈25% larger in TopHat alignments than in STAR alignments. (Larger *nlstart* and *qsm* values result in more validated gap-sites. Due to identical genomic coordinates, the MaxEnt scores are equal in both alignments).

2.5.2. Relation between Score Value and Gap-Site Multiplicity

The proportion of samples in which a gap-site is observed can be related to (absolute) score values (for *gqs* and $|wgis|$) which describe to which extent one criterion reflects the other one. Figure 11 shows that, in general, increasing score values (*gqs* and $|wgis|$) are associated with a higher proportion of samples in which a gap-site is identified.

There are gap-sites with low *gqs* which are present in a substantial fraction of samples. A high *gqs* is required (980 for TopHat and 975 for STAR alignments) in order to achieve presence in the majority (>50%) of samples.

Relations for $|wgis|$ indicate a clearer relationship than observed for *gqs*. Gap-sites are present in the majority of samples (>50%) when their $|wgis|$ value is >82. The median multiplicity becomes >1 when $|wgis|$ is >26. Thus the gap-site multiplicities also support the *gql* classification criteria (In detail: The 50% limit is exceeded at $|wgis|$ values of 81 in TopHat alignments and 82 in STAR alignments. The nProbes = 1 limit is exceeded at $|wgis|$ values of 28 in TopHat alignments and 27 in STAR alignments.). Additionally, the straight and sigmoidal shape of the curve for $|wgis|$ shows that *wgis* has a much better capability to predict gap-site multiplicity than *gqs*.

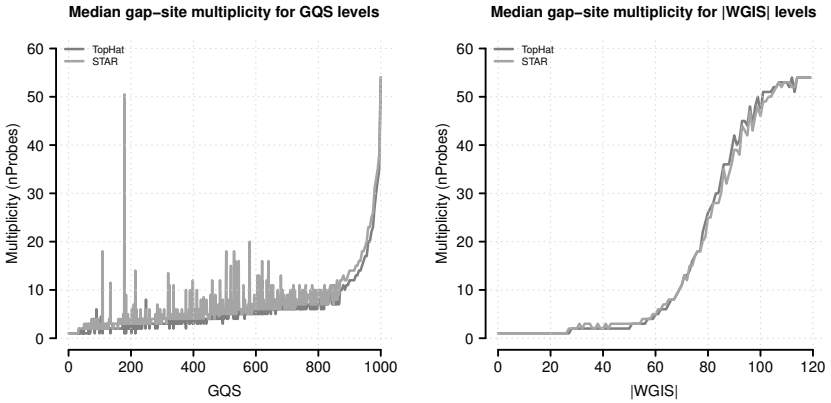


Figure 11. Median gap-site multiplicities for different score levels. Gap-site multiplicity (nProbes, the number of samples in which a gap-site is identified) are used as category (absolute values of *wgis* are rounded to integral numbers). For each score category, the median nProbes value is calculated and displayed in the figure. (**Left**) Gap-site multiplicities for different *gqs* values; (**Right**) Gap-site multiplicities for different $|wgis|$ values.

2.5.3. Dependence of Gap-Site Validation on Gap-Site Coverage

The distribution of gap-site numbers in TopHat and STAR alignments verified by *gqs* or *wgis* for different levels of alignment coverage is shown in Figure 12. Globally, STAR reports 38.3% more *wgis* validated gap-sites than TopHat and 6.4% more *gqs*-validated gap-sites (More details are provided in supplemental material.). A (local) maximum of gap-site numbers is present at $\approx 10^4$ alignments coverage, a value essentially determined by the total sequencing depth (on 54 fibroblast samples).

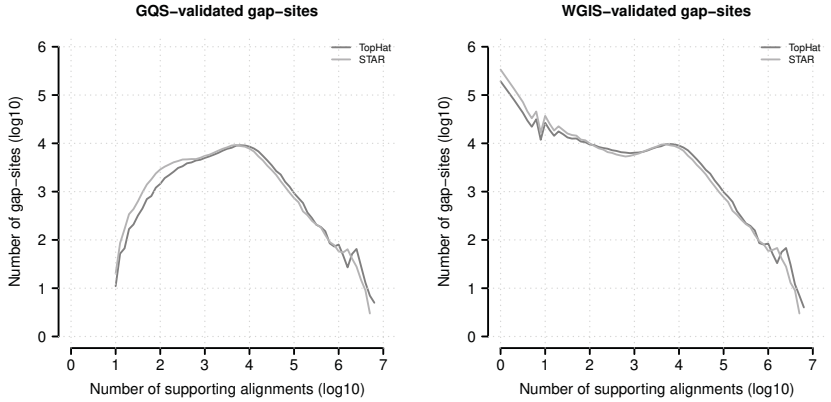


Figure 12. Absolute numbers of verified gap-sites for different alignment depth's. Gap-sites were categorised according to number of supporting alignments (nAligns): The $\log_{10}$(nAligns) value rounded to one digit was used as category (x-axis). The lines display the $\log_{10}$ of tabled validated gap-sites (*gqs*-validated left, *wgis*-validated right).

The majority of *wgis* validated and only a small minority of *gqs* validated gap-sites are supported by <100 alignments: For *gqs*-validation in 3.0% and 5.7% of gap-sites in TopHat and STAR alignments respectively. For *wgis*-validation in 74.1% and 71.4% of gap-sites in TopHat and STAR alignments respectively.

2.5.4. Relation between *gqs* and *wgis* Validation

Globally, the majority of verified gap-sites are verified by *wgis* alone (Table 5) but there may be differences in gap-sites with high alignment coverage. Therefore the verification with *gqs* and *wgis* is compared for different nAligns ranges (Figure 13). The percentage of gap-sites verified only by *gqs* is <3.2% in TopHat alignments and <6% in STAR alignments and thus very low.

The vast majority of gap-sites is verified by *wgis* alone or by both scores with the proportion of *gqs* validated sites increasing when alignment coverage is larger than 100. Proportions including gap-sites not validated by both scores (*gqs* and *wgis*) are shown in Supplemental material.

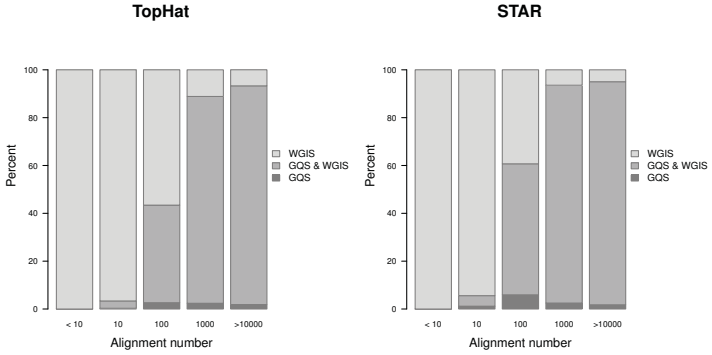
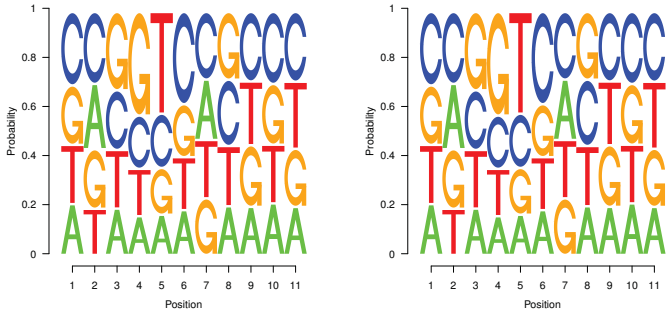


Figure 13. Proportion of verified gap-sites for different alignment depth's.

2.6. Unvalidated Gap-Sites

A gap-site is not *gqs* validated when $nlstart < 8$ or $qsm < 200$. A gap-site is not *wgis* validated when $qsm \leq 15$ or when MaxEnt $score5 \leq 1$ or $score3 \leq 1$ in either strand direction. Low read coverage of gap-sites reduces likelihood of *gqs* validation while *wgis* validation still relies on MaxEnt scores as long as $qsm > 15$. While for *gqs*-validation, a coverage at least 8 alignments is required, *wgis*-validation can already be achieved with one single alignment. From *wgis*-unvalidated gap-sites 8431 (1.00%) in TopHat and 9803 (0.16%) in STAR alignments are found in all 54 samples. Also, TopHat and STAR report 8689 and 10,590 Ensembl annotated splice-sites respectively not validated by *wgis* as well as 7330 and 8186 splice-sites respectively not validated by *gqs*. Sequence logos for gap-sites reported by STAR aligner and not validated by *wgis* are shown in Figure 14. The sequence logos largely do not match splice-site consensus sequences demonstrating that validation is a critical prerequisite for further analysis.



(**a**) STAR, left junction (**b**) STAR, right junction

Figure 14. Sequence logos for gap-sites not validated by *wgis*. Sequence logos for (5,421,981) gap-sites not validated by *wgis* (GQL = 0) from STAR aligner. Tabled nucleotides are corrected for strand orientation reported by *wgis*. (**a**) Splice-junction is located between position 3 and 4 (gap-site *lend* is located at position 3); (**b**) Splice-junction is located between position 8 and 9 (gap-site *rstart* is located at position 9).

2.6.1. Maximal Alignment Coverage on Unvalidated Gap-Sites

In gap-sites not validated by *gqs*, the maximal alignment coverage is 2,994,298 in alignments from TopHat and 1,664,581 in alignments from STAR comprising 45.18% (TopHat) and 30.67% (STAR) of the maximal number of supporting alignments.

In gap-sites not validated by *wgis*, the maximal alignment coverage is 5,880,547 in alignments from TopHat and 3,545,564 in alignments from STAR comprising 88.73% (TopHat) and 65.34% of the maximal number of supporting alignments.

3. Discussion

3.1. Performance of Quality Scores

3.1.1. *gqs*

Gap-sites collected from 54 fibroblast samples and validated by *gqs* contain a reasonable high proportion (≈98% in both aligners) of GT-AG sites. Using *gqs*, ≈156,000 gap-sites are validated in TopHat alignments and ≈ 166,000 in STAR alignments (6% more) in merged data from 54 samples. It may be advantageous that no additional information from genomic sequence is required for calculation of *gqs*. But for the validation of >50% of gap-sites, alignment coverage of more than ≈250–500 is required, a high number indicating limitations in sensitivity. This number is a static value independent of number of samples and depends solely on total alignment depth (on merged samples together). Also, *gqs* cannot provide strand information. Thus, validation of gap-sites with *gqs* shows, that suitable quality criteria can be derived from pure alignment data (BAM file content), but sensitivity is low and strand information must be provided from a second source which might be a challenging task.

3.1.2. WGIS

The *wgis* uses the similarity of a gap-site with known splice-sites as quality criterion additionally to the alignment data. Using *wgis*, ≈770,000 gap-sites are validated in TopHat alignments and ≈1,066,000 in STAR alignments (38% more) in merged data from 54 samples. Thus, *wgis* validates 4.9 times (in TopHat alignments) and 6.4 times (in STAR alignments) as much as *gqs*. Approximately

$\frac{3}{4}$ of gap-sites reported by TopHat are validated by *wgis*. Thus, the validation sensitivity is considerably higher in *wgis* than in *gqs*. Also, the strand information provided by *wgis* (without requirement of annotation) is a potentially valuable feature. Gap-sites validated by *wgis* can further be categorised into quality levels *gql*1 to *gql*3. Thus, usage of features in genomic sequence appears to be a valuable source of information for validation of gap-sites which may substantially increase sensitivity and also may provide strand information.

3.2. Comparison of TopHat and STAR Aligner

3.2.1. Sensitivity and FDR

TopHat is known to have a low mapping yield while STAR has been shown to report many alignments with a high base-wise accuracy [3]. RGASP further found the rate of correctly mapped gapped-reads in the range of 96.3% to 98.4% but also notes at the same time, that many erroneously reported gap-sites are a major obstacle in STAR alignments [3].

For detection of occasionally observed splice sites and low abundant events, merging of data from multiple samples is necessary in order to increase sensitivity [19]. But usage of this approach increases sensitivity as well as FDR, possibly leading to *GT-AG* proportions of $<$50% in alignments from STAR. Also, the fact that 68.4% of unique gap-site positions ($\approx 4.4 \times 10^6$ in our 54 samples) are only supported by one single alignment indicates a low specificity (high FDR) for STAR aligner.

An FDR in this magnitude diminishes reliability of downstream analysis. It is easily conceivable how high FDR for gap-sites deteriorate results of dependent functions like intron detection. Thus, we propose that for transcript reconstruction, a high FDR for gap-sites may be an impeding factor additionally to the difficulties imposed by the complexity of the human genome [2].

3.2.2. Distribution of Gapped Alignments

In merged data from 54 RNA-seq samples, we found that although STAR aligner reports much more gap-sites than TopHat, the number of gapped alignments in STAR alignments actually is lower than in TopHat alignments. Thus, the higher mapping numbers reported by STAR (which partly are due to the ability to report truncated alignments [3]) seem to be not present in gapped alignments. Also, the higher number of gap-sites reported by STAR is mainly due to a different distribution of gapped alignments leading to a reduced alignment coverage on gap-sites (in mean 75.6% coverage in TopHat alignments). STAR reports 6.4% more *gqs*-validated gap-sites and 38.3% more *wgis*-validated gap-sites than TopHat. Thus the already reported higher sensitivity for STAR in general [3] is also present in (*gqs* or *wgis*) validated gap-sites.

3.3. Validation Strategies

As validation strategies for gap-sites may greatly enhance accuracy of the detection process, alternative strategies have already been proposed or implemented:

- Restrict alignment gaps to a small subset of possible intronic dinucleotides (as implemented in TopHat)
- Filter on alignment gaps supported by minimal 20 matching nucleotides on each side of the gap [3]
- Filter junctions on number of supporting alignments [3]
- Filter out alignments assessed as low confidence by regression model (basing on nucleotide distributions around splice junctions and intron size, as implemented in oLego) [20].

As the first filter bases on look-up of genomic sequence not present in BAM files and the second filter requires examination of CIGAR items adjacent to alignment gaps, both filters cannot be implemented with a simple algorithm. A filter simply basing on number of supporting alignments requires consideration of alignment depth (and a normalisation step). Also, in order to provide a consistent quality criterion, very high threshold values are required (at least 1.5×10^6 alignments;

see Section 2.6.1). Thus, using this criterion alone aggravates the low sensitivity of *gqs* for gap-sites with low coverage.

Based on considerations that the first three strategies "as is" are insufficient, we combined alignment based criteria into *gqs* and included sequenced based information using a complex criterion (MaxEnt) into *wgis*. The results indicate that criteria solely based on BAM-file content seem to lack sensitivity, but this handicap may be overcome by assessment of splice-site similarity. Criteria which are more sophisticated than simply filtering on IDIN-pairs even have the potential to detect rare splice-events possibly blurred by statistical noise.

Usage of MaxEnt Score

Filtering gap-sites based on MaxEnt scores may greatly enhance percentage of GT-AG-sites although this method might have limitations. First, evaluation of genomic sequence is not absolutely reliable for prediction of splice-site utilisation [21]. Also, MaxEnt in its current form assigns low scores to splice-sites recognised by minor spliceosome and ignores positions 10 and 11 of the 5′ splice-site. A correction cannot be introduced by shifting thresholds because this would lead to increased validation of gap-sites not similar to splice-sites recognised by the major spliceosome. Due to optimality properties of the MaxEnt estimation process, the only way for possible improvement strategies are:

- Calculation of new MaxEnt score tables basing on larger (or modified) "standard" samples.
- Creation of a second score solely basing on recognition by the minor spliceosome and using the maximum of both.

3.4. Limitations

Although it can be supposed, that reasonable proportions of nucleotide distributions (IDIN, IDIN-pairs and sequence logos) indicate that algorithms and decision rules act correctly in a majority of cases, these numbers should be interpreted with caution. As (potential) splice-sites are counted by occurrence and not by abundance, merging of data from multiple sources (for example by merging samples or by collecting data in an annotation data base) inevitably leads to rising numbers for rare events and thus may lead to overestimated proportions. The size of this effect is affected by the extensiveness of included information.

4. Materials and Methods

4.1. Fibroblast Samples

The transcriptome data analysed in this study originated from an investigation where effects of age, gender and UV exposition were studied in samples of dermal fibroblasts obtained from healthy human donors. A main result of the study is, that no consistent differential expression of genes is observable. The gene expression hence is assumed to be essentially homogeneous in these samples.

Details of sample preparation and the results differential expression analysis recently have been described elsewhere [12]. In short, the mRNAs of 54 samples were sequenced on an Illumina HiSeq 2000 sequencer yielding in total 8,785,501,333 reads. Subsequent alignments were calculated on unpreprocessed FASTQ files. Alignments were calculated using bowtie2 (2.2.5) [13], tophat (2.0.14) [14] and STAR (2.4.1d modified) [15]. Collection and processing of dermal samples from donors was approved by the Ethical Committee of the Medical Faculty of the University of Düsseldorf (# 3361) in 2011.

4.1.1. Software

All described algorithms are implemented in R and are available from CRAN or Bioconductor. The software interface to samtools, the container and extraction algorithms for gap-sites are written in C/C++ and contained in CRAN package *rbamtools* [16]. Implementations for import of annotation

data (GTF) and the annotation procedure for gap-sites are written in C/C++ and contained in CRAN package *refGenome*. Implementations for calculation of MaxEnt scores were translated into C and are publicly available in Bioconductor package *spliceSites* (version $\geq$ 1.23.5). Algorithms for extraction of DNA sequence for exon/intron boundaries use Bioconductor framework (Biostrings). The source code for R (for analysis and package development) as well as the latex source code for this document was developed using RStudio [17].

4.1.2. Statistical Evaluation

For *wgis*-validated gap-sites, IDIN are reported after correction for strand orientation. Strand assignments are calculated as described in Section 1.3.2 (the *wgis*-reported strand). In sequence logos, letter height corresponds to nucleotide frequencies at given positions. Nucleotides are decreasingly ordered according to their frequency from top to bottom. All shown sequence logo's are calculated after correction for strand orientation.

5. Conclusions

The aligners TopHat and STAR report a high rate of unvalidated gap-sites emphasising the necessity for further validation. Validation of gap-sites without using genomic sequence data (*gqs*) requires high alignment coverages which can greatly be reduced when the similarity to known splice-sites is evaluated.

Supplementary Materials: Supplementary materials can be found at www.mdpi.com/1422-0067/18/6/1110/s1.

Acknowledgments: This study was supported in part by the Deutsche Forschungsgemeinschaft (DFG, SCHW 1508/3-1) to H. Schw. and (DFG, SCHA 909/401) and the German Ministry of Research and Education (Network Gerontosys, Stromal Aging to Heiner Schaal WP3, part C) to H. Scha. We thank Lisa Müller for editing the manuscript.

Author Contributions: Wolfgang Kaisers developed the software, analysed the data and wrote the paper. Johannes Ptok designed and performed experiments for alternative validation methods for gap-sites with low read number. Holger Schwender reviewed and edited the paper. Heiner Schaal conceived and designed the experiments and reviewed and edited the paper.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

FDR	False discovery rate
nAligns	Number of supporting alignments (for a gap-site)
gqs	Gap quality score
wgis	Weighted gap information score
IDIN	Intronic dinucleotides
mcl	Minimum CIGAR length
qsm	Quartet sum of MCL
lend	Left end
$\log_2$	Logarithm to the base 2
$\log_{10}$	Logarithm to the base 10
rstart	Right start
nlstart	Number of left start (positions)
RGASP	RNA-seq Genome Annotation Assessment Project
SD	Standard deviation

Appendix A. Information Collected for Gap-Sites

Appendix A.1. BAM File Format

Gap-sites are identified using alignment data provided in BAM (Binary Alignment/Map) format (see also https://samtools.github.io/hts-specs/SAMv1.pdf). In a BAM file, each alignment is

represented by a set of data items, mainly the reference sequence (chromosome number), the position and CIGAR items.

Appendix A.1.1. Gap-Site Positions Extracted from BAM File Format

Multiple gapped alignments define a single gap-site when they contain identical located alignment gaps. The genomic position of the rightmost matching nucleotide on the left side of the alignment gap is named *lend* (left end). The position of the leftmost matching position on the right side of the alignment gap is named *rstart* (right start). In the following example from a single genomic alignment

```
01234567890123456
CCCTACGTCCCAGTCAC (reference)
   TAC       TCAC (query)
```

the significant genomic positions are lend = 5 and rstart = 13. *lend* and *rstart* are the defining positions of a gap-site. The *gap-length* is defined as the number of nucleotides which are not covered by alignment matches. In our example, the gap-length is 7 (=*rstart* − *lend* − 1).

Appendix A.1.2. CIGAR Items in BAM File Format and Derived Values

Details of alignments, for example match length or gap-size, are provided in the CIGAR string for each alignment. A CIGAR string consists a sequence of CIGAR items. Each CIGAR item consists of number (the length or number of affected nucleotides) and a single character (the type of "CIGAR operation"). The CIGAR item 50 M represents a subsequent match of 50 nucleotides (between query and reference sequence). The CIGAR item 1000 N indicates an alignment gap of length 1000. An alignment gap is represented by triple sequence of CIGAR items xMyNzM where y is the gap-length. Examples are shown in Table A1. Multiple alignment gaps sharing the same *lend* and *rstart* position (on the same reference sequence) constitute a gap-site. The gap-site sharing alignments are called "supporting alignments". The values *nlstart* and *qsm*, used in *gqs* and *wgis*, are directly calculated from CIGAR strings in all supporting alignments.

Table A1. Gapped alignments and gap-sites.

Name	Match	Gap	Match	Position	CIGAR	*mcl*
query1	AG		CCTTGATG	3	2M6N8M	2
query2	CAG		CCTTGAT	2	3M6N7M	3
query3	CCAG		CCT	1	4M6N3M	3
reference	CCCAG	GTCCAG	CCTTGATGTCC			

In the gap-site defining alignments, three different match lengths (=CIGAR length for M-item) on the left side of the alignment gap are present (values 2, 3 and 4). Therefore, the *nlstart* value for the shown gap-site is 3. The *mcl* values for the three alignments are 2, 3, 3 thus resulting in a *qsm* of 8 (the missing *mcl* value is set to 0). The *gqs* resulting from the shown alignments is 15.

Appendix A.1.3. The *mcl* Value of a Gap-Site

The minimum CIGAR length (*mcl*) value is defined as the minimum of adjacent match lengths (M item in CIGAR segment) on each side of the alignment gap (N item in CIGAR segment). Thus, an *mcl* value of 10 for an alignment gap defined by CIGAR items *x*M*y*NzM ensures that $x \geq 10$ and $y \geq 10$ ensuring, that the gap at least is supported by 20 subsequent nucleotide matches in query (read) sequence.

Appendix A.1.4. The *qsm* Value of a Gap-Site

From alignments defining a gap-site, the quartet sum of mcl (*qsm*) is calculated as the sum of the four largest *mcl* (see Table A1) values from alignments defining the gap-site. Assuming, a gap-site is defined by n alignments with *mcl* values $(mcl_i)_{i=1\ldots n}$ and that the *mcl* values are ordered decreasingly (i.e., $mcl_i \geq mcl_{i+1}$) the definition can be stated as

$$qsm := \sum_{i=1}^{4} mcl_i. \tag{A1}$$

For read lengths of 101, the maximal value of *qsm* from one alignment can be 50 and the maximal *qsm* value of 200 can be achieved with at least four alignments. The *qsm* provides a measure for the amount of coverage based information supporting the presence of a splice site. As maximal four alignments account for *qsm*, information from long continuous alignments is emphasised.

Appendix A.1.5. The *nlstart* Value of a Gap-Site

The *nlstart* value of a gap-site is the number of different match lengths in the left adjacent matching segment of alignments defining a gap-site. When a gap-site is defined by n alignments $(A_i)_{i=1,\ldots,n}$ containing CIGAR items x_iMy_iNz_iM, the *nlstart* value is defined as the number of different x_i values. In Table A1, the x_i values are {2,3,4}, so *nlstart*= 3.

Regarding the x_i values as realisations of a random variable where all observed values occur with equal probability, and using the definition of Shannon entropy, it is known that $\log 2(n)$ provides a measure of information content of the values x_i (i.e., the number of bits necessary for encoding the x_i values).

In a discrete probability space $\boldsymbol{A} = (A, p)$ where A is a finite set ($A = \{a_1, \ldots, a_n\}$) and p is a probability measure ($p(a_i) =: p_i$), the *Shannon entropy* of A is defined as

$$H(A) := -\sum_{i=1}^{n} p_i \log_2(p_i). \tag{A2}$$

In case of equal probabilities ($p_1 = \ldots = p_n$) the entropy reduces to $H(A) = \log_2(n)$.

Appendix B. Weight Functions for *qsm* and *nlstart*

The *wgis* applies weight functions to *nlstart*, *qsm* and MaxEnt values. From the monotonicity criterion follows, that only monotone increasing weight functions may be used. In order to emphasise information contained in few high quality alignments (which increases sensitivity), the weight functions must be concave (having monotonic decreasing slope) advocating the logarithm function.

The logarithm for base *a* (log_a) solely adds a constant factor (log(a)) to the natural logarithm (See Supplemental Material). As the shape of the function is unchanged, the decisive properties of a score are equal for all base values. Thus, the logarithm base can be chosen arbitrarily (>0). For ease of calculation, $\log_2$ is utilised for all weight functions.

Appendix B.1. Weighting of nlstart value

The logarithmic weight for *nlstart* values emphasises gap-sites constellations with more then one alignment start position. *nlstart* values of 2 and 8 are translated into *nlstart*-factor values of 2 and 8 respectively.

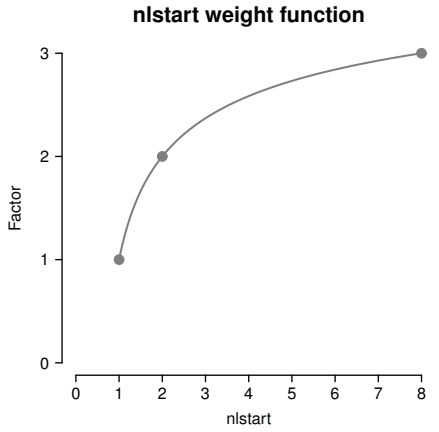


Figure A1. Weight function for *nlstart*. Occurrent values for *nlstart* are values in $\{1,\ldots,8\}$. The weighted value for *nlstart* is defined as $\log_2(\log_2(nlstart)+1)+1)$. Points are added at *nlstart* values of 1, 2 and 8 (indicating weighted values of 1, 2 and 3 respectively).

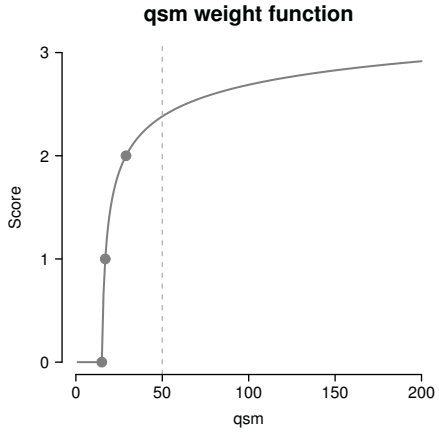


Figure A2. Weight function for *qsm*. Collected *qsm* values are integral values between 1 and ($2\times$ read-length). The weighted value for *qsm* is defined as $\log_2(\log_2(\max(qsm-13,2)))$. Points are added at *qsm* values of 15, 17, and 29 (indicating weighted values of 0, 1 and 2 respectively).

Appendix B.2. Weighting of qsm Value

A major obstacle for alignment based validation of splicing events is a too small number of aligned nucleotides (on either side of the gap-site). Therefore, a lower limit for the number of aligned nucleotides on either adjacent region is introduced, implemented as lower limit for *qsm*. The weight function for *qsm* is constructed so that *qsm* values of 17 and 29 result in a *qsm*-factor of 1 and 2 respectively. The maximal possible *qsm*-factor (for reads of length of 101) is 2.91. The *qsm* weight function describing the relation between *qsm* and *qsm*-factor is shown in Figure A2.

References

1. Hayer, K.E.; Pizarro, A.; Lahens, N.F.; Hogenesch, J.B.; Grant, G.R. Benchmark analysis of algorithms for determining and quantifying full-length mRNA splice forms from RNA-seq data. *Bioinformatics* **2015**, *31*, 3938–3945.
2. Steijger, T.; Abril, J.F.; Engstrom, P.G.; Kokocinski, F.; Hubbard, T.J.; Guigo, R.; Harrow, J.; Bertone, P.; Abril, J.F.; Akerman, M.; et al. Assessment of transcript reconstruction methods for RNA-seq. *Nat. Methods* **2013**, *10*, 1177–1184.
3. Engstrom, P.G.; Steijger, T.; Sipos, B.; Grant, G.R.; Kahles, A.; Ratsch, G.; Goldman, N.; Hubbard, T.J.; Harrow, J.; Guigo, R.; et al. Systematic evaluation of spliced alignment programs for RNA-seq data. *Nat. Methods* **2013**, *10*, 1185–1191.
4. Wang, Z.; Gerstein, M.; Snyder, M. RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **2009**, *10*, 57–63.
5. Parada, G.E.; Munita, R.; Cerda, C.A.; Gysling, K. A comprehensive survey of non-canonical splice sites in the human transcriptome. *Nucleic Acids Res.* **2014**, *42*, 10564–10578.
6. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078–2079.
7. Zhao, S.; Zhang, Y.; Gordon, W.; Quan, J.; Xi, H.; Du, S.; von Schack, D.; Zhang, B. Comparison of stranded and non-stranded RNA-seq transcriptome profiling and investigation of gene overlap. *BMC Genom.* **2015**, *16*, 675.
8. Yeo, G.; Burge, C.B. Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.* **2004**, *11*, 377–394.
9. Eng, L.; Coutinho, G.; Nahas, S.; Yeo, G.; Tanouye, R.; Babaei, M.; Dork, T.; Burge, C.; Gatti, R.A. Nonclassical splicing mutations in the coding and noncoding regions of the ATM Gene: Maximum entropy estimates of splice junction strengths. *Hum. Mutat.* **2004**, *23*, 67–76.
10. Jaynes, E.T. Information Theory and Statistical Mechanics. *Phys. Rev.* **1957**, *106*, 620–630.
11. Jaynes, E.T. Information Theory and Statistical Mechanics. II. *Phys. Rev.* **1957**, *108*, 171–190.
12. Kaisers, W.; Boukamp, P.; Stark, H.J.; Schwender, H.; Tigges, J.; Krutmann, J.; Schaal, H. Age, gender and UV-exposition related effects on gene expression in in vivo aged short term cultivated human dermal fibroblasts. *PLoS ONE* **2017**, *12*, e0175657.
13. Langmead, B.; Salzberg, S.L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **2012**, *9*, 357–359.
14. Trapnell, C.; Pachter, L.; Salzberg, S.L. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* **2009**, *25*, 1105–1111.
15. Dobin, A.; Davis, C.A.; Schlesinger, F.; Drenkow, J.; Zaleski, C.; Jha, S.; Batut, P.; Chaisson, M.; Gingeras, T.R. STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **2013**, *29*, 15–21.
16. Kaisers, W.; Schaal, H.; Schwender, H. rbamtools: An R interface to samtools enabling fast accumulative tabulation of splicing events over multiple RNA-seq samples. *Bioinformatics* **2015**, *31*, 1663–1664.
17. RStudio Team. *RStudio: Integrated Development Environment for R;* RStudio, Inc.: Boston, MA, USA, 2015.
18. Farrer, T.; Roller, A.B.; Kent, W.J.; Zahler, A.M. Analysis of the role of Caenorhabditis elegans GC-AG introns in regulated splicing. *Nucleic Acids Res.* **2002**, *30*, 3360–3367.
19. Hooper, J.E. A survey of software for genome-wide discovery of differential splicing in RNA-Seq data. *Hum. Genom.* **2014**, *8*, 3.
20. Wu, J.; Anczukow, O.; Krainer, A.R.; Zhang, M.Q.; Zhang, C. OLego: Fast and sensitive mapping of spliced mRNA-Seq reads using small seeds. *Nucleic Acids Res.* **2013**, *41*, 5149–5163.
21. Mount, S.M. Genomic sequence, splicing, and gene annotation. *Am. J. Hum. Genet.* **2000**, *67*, 788–792.

Sample Availability: The raw FASTQ files are available under ArrayExpress accession E-MTAB-4652 (ENA study ERP015294). The software is available in R packages: *rbamtools* and *refGenome* on CRAN and *spliceSites* (including algorithm for *wgis*) on Bioconductor.

Kapitel 5

Ausblick

5.1 Manuskript 1: rbamtools: an R interface to samtools enabling fast accumulative tabulation of splicing events over multiple RNA-seq samples

Das rbamtools Paket enthält eine praktisch vollständige Schnittstelle, um aus R Alignment Daten aus BAM-Format zu lesen und nach BAM-Format zu schreiben. Das Datenformat und die Schnittstelle hat sich seit Implementation deutlich weiter entwickelt und ist heute Teil der HTSlib Bibliothek (`http://www.htslib.org/`). Da HTSlib auch in C geschrieben ist, könnte eine Weiterentwicklung darin bestehen, für die HTSlib Bibliothek eine C++ Schnittstelle zu implementieren, die den Zugriff auf die von HTSlib unterstützten Datenformate für Programmierer vereinfacht. Beispielsweise könnte auf dieser Ebene schon Multithreading implement werden. Dies Implementation könnte dann auch in Schnittstellen zu anderen Programmier -platformen verwendet werden (beispielsweise in R).
Eine weitere lohnenswerte Weiterentwicklung könnte eine Implementation des BGZF Formats als C++ Bibliothek sein. Die SAMtools Implementation zeigt, dass BGZF die Vorteile der Datenkompression mit dem des Random Access vereinen kann. Dieses Prinzip hätte das Potenzial, sich auch in anderen Anwendungsbereichen zu etablieren.

5.2 Manuskript 2: Validation of Splicing Events in Transcriptome Sequencing Data

Die Ergebnisse aus der Analyse von Gap-Sites zeigen, dass bei der Identifikation von Spleißereignissen, gerade wenn sproradisch genutzte Spleißstellen in multiplen Transkriptomen identifiziert werden sollen, eine Gap-Site Validierung unverzichtbar ist. Diese Situation könnte sich ändern, wenn beim Alignment Prozess ein stärkeres Augenmerk auf biologische Aspekte des Speißvorganges gelegt wird.
Mit den Scores *gqs* und *wgis* sind zwei Ansätze für die Validierung von Gap-Sites entwickelt und ihre Eigenschaften untersucht worden. Es stellte sich heraus, dass eine Validierung, die alleine auf Alignment-Daten beruht, eine hohe Sequenzierungs-Dichte erfordert. Durch deutlich längere Reads könnten sich die Erfordernisse hier ändern (und möglicherweise auch die Transkript-Rekonstruktion vereinfachen).
Die Alternative, das Hinzuziehen von Prädiktoren, die auf der Analyse genomischer DNA Sequenz beruhen, kann theoretisch die Sensibilität der Gap-Site Validierung erhöhen, aber der Wert der Prädiktion bleibt angesichts der Komplexität der Aktion spleißregulatorischer Faktoren ungewiss.

Bei der Evaluation der beiden Scores stellte sich heraus, dass sich die Anzahl der beobachtbaren Gap-Sites in mehreren Proben (nach Merge-Vorgang) durch ein einfaches wahrscheinlichkeitstheoretisches Modell beschreiben läßt. Dieses Modell und eine sich daraus ableitende Möglichkeit der Fallzahlberechnung ist in einem weiteren Manuskript beschrieben, das im International Journal for Molecular Sciences (2017) erschienen ist [48].

Literaturverzeichnis

[1] Alpert, T. ; Herzel, L. ; Neugebauer, K. M.: Perfect timing: splicing and transcription rates in living cells. In: *Wiley Interdiscip Rev RNA* 8 (2017), Mar, Nr. 2. https://doi.org/10.1002/wrna.1401

[2] Baralle, D. ; Buratti, E.: RNA splicing in human disease and in the clinic. In: *Clin. Sci.* 131 (2017), Mar, Nr. 5, S. 355–368

[3] Baralle, F. E. ; Buratti, E.: RNA and splicing regulation in neurodegeneration. In: *Mol. Cell. Neurosci.* 56 (2013), Sep, S. 404–405

[4] Baralle, F. E. ; Giudice, J.: Alternative splicing as a regulator of development and tissue identity. In: *Nat. Rev. Mol. Cell Biol.* 18 (2017), Jul, Nr. 7, S. 437–451

[5] Barash, Y. ; Calarco, J. A. ; Gao, W. ; Pan, Q. ; Wang, X. ; Shai, O. ; Blencowe, B. J. ; Frey, B. J.: Deciphering the splicing code. In: *Nature* 465 (2010), May, Nr. 7294, S. 53–59

[6] Behzadnia, N. ; Golas, M. M. ; Hartmuth, K. ; Sander, B. ; Kastner, B. ; Deckert, J. ; Dube, P. ; Will, C. L. ; Urlaub, H. ; Stark, H. ; Luhrmann, R.: Composition and three-dimensional EM structure of double affinity-purified, human prespliceosomal A complexes. In: *EMBO J.* 26 (2007), Mar, Nr. 6, S. 1737–1748

[7] Black, D. L.: Mechanisms of alternative pre-messenger RNA splicing. In: *Annu. Rev. Biochem.* 72 (2003), S. 291–336

[8] Blencowe, B. J.: Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases. In: *Trends Biochem. Sci.* 25 (2000), Mar, Nr. 3, S. 106–110

[9] Blencowe, B. J.: Alternative splicing: new insights from global analyses. In: *Cell* 126 (2006), Jul, Nr. 1, S. 37–47

[10] Buratti, E. ; Baralle, M. ; Baralle, F. E.: Defective splicing, disease and therapy: searching for master checkpoints in exon definition. In: *Nucleic Acids Res.* 34 (2006), Nr. 12, S. 3494–3510

[11] Burge, C. B. ; Padgett, R. A. ; Sharp, P. A.: Evolutionary fates and origins of U12-type introns. In: *Mol. Cell* 2 (1998), Dec, Nr. 6, S. 773–785

[12] Caceres, J. F. ; Kornblihtt, A. R.: Alternative splicing: multiple control mechanisms and involvement in human disease. In: *Trends Genet.* 18 (2002), Apr, Nr. 4, S. 186–193

[13] Cartegni, L. ; Chew, S. L. ; Krainer, A. R.: Listening to silence and understanding nonsense: exonic mutations that affect splicing. In: *Nat. Rev. Genet.* 3 (2002), Apr, Nr. 4, S. 285–298

[14] Castello, A. ; Fischer, B. ; Eichelbaum, K. ; Horos, R. ; Beckmann, B. M. ; Strein, C. ; Davey, N. E. ; Humphreys, D. T. ; Preiss, T. ; Steinmetz, L. M. ; Krijgsveld, J. ; Hentze, M. W.: Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. In: *Cell* 149 (2012), Jun, Nr. 6, S. 1393–1406

[15] Cock, P. J. ; Fields, C. J. ; Goto, N. ; Heuer, M. L. ; Rice, P. M.: The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. In: *Nucleic Acids Res.* 38 (2010), Apr, Nr. 6, S. 1767–1771

[16] Collins, L. ; Penny, D.: Complex spliceosomal organization ancestral to extant eukaryotes. In: *Mol. Biol. Evol.* 22 (2005), Apr, Nr. 4, S. 1053–1066

[17] Djebali, S. ; Davis, C. A. ; Merkel, A. ; Dobin, A. ; Lassmann, T. ; Mortazavi, A. ; Tanzer, A. ; Lagarde, J. ; Lin, W. ; Schlesinger, F. u. a.: Landscape of transcription in human cells. In: *Nature* 489 (2012), Sep, Nr. 7414, S. 101–108

[18] Douglas, A. G. ; Wood, M. J.: Splicing therapy for neuromuscular disease. In: *Mol. Cell. Neurosci.* 56 (2013), Sep, S. 169–185

[19] Dreyfuss, G. ; Kim, V. N. ; Kataoka, N.: Messenger-RNA-binding proteins and the messages they carry. In: *Nat. Rev. Mol. Cell Biol.* 3 (2002), Mar, Nr. 3, S. 195–205

[20] Dreyfuss, G. ; Matunis, M. J. ; Pinol-Roma, S. ; Burd, C. G.: hnRNP proteins and the biogenesis of mRNA. In: *Annu. Rev. Biochem.* 62 (1993), S. 289–321

[21] Dunham, I. ; Kundaje, A. ; Aldred, S. F. ; Collins, P. J. ; Davis, C. A. ; Doyle, F. ; Epstein, C. B. ; Frietze, S. ; Harrow, J. ; Kaul, R. u. a.: An integrated encyclopedia of DNA elements in the human genome. In: *Nature* 489 (2012), Sep, Nr. 7414, S. 57–74

[22] Dvinge, H. ; Kim, E. ; Abdel-Wahab, O. ; Bradley, R. K.: RNA splicing factors as oncoproteins and tumour suppressors. In: *Nat. Rev. Cancer* 16 (2016), 07, Nr. 7, S. 413–430

[23] Engstrom, P. G. ; Steijger, T. ; Sipos, B. ; Grant, G. R. ; Kahles, A. ; Ratsch, G. ; Goldman, N. ; Hubbard, T. J. ; Harrow, J. ; Guigo, R. u. a.: Systematic evaluation of spliced alignment programs for RNA-seq data. In: *Nat. Methods* 10 (2013), Dec, Nr. 12, S. 1185–1191

[24] Eperon, I. C. ; Makarova, O. V. ; Mayeda, A. ; Munroe, S. H. ; Caceres, J. F. ; Hayward, D. G. ; Krainer, A. R.: Selection of alternative 5' splice sites: role of U1 snRNP and models for the antagonistic effects of SF2/ASF and hnRNP A1. In: *Mol. Cell. Biol.* 20 (2000), Nov, Nr. 22, S. 8303–8318

[25] Erkelenz, S. ; Mueller, W. F. ; Evans, M. S. ; Busch, A. ; Schoneweis, K. ; Hertel, K. J. ; Schaal, H.: Position-dependent splicing activation and repression by SR and hnRNP proteins rely on common mechanisms. In: *RNA* 19 (2013), Jan, Nr. 1, S. 96–102

[26] Erkelenz, S. ; Theiss, S. ; Otte, M. ; Widera, M. ; Peter, J. O. ; Schaal, H.: Genomic HEXploring allows landscaping of novel potential splicing regulatory elements. In: *Nucleic Acids Res.* 42 (2014), Nr. 16, S. 10681–10697

[27] Ezkurdia, I. ; Rodriguez, J. M. ; Santa Pau, E. Carrillo-de ; Vazquez, J. ; Valencia, A. ; Tress, M. L.: Most highly expressed protein-coding genes have a single dominant isoform. In: *J. Proteome Res.* 14 (2015), Apr, Nr. 4, S. 1880–1887

[28] Fairbrother, W. G. ; Yeh, R. F. ; Sharp, P. A. ; Burge, C. B.: Predictive identification of exonic splicing enhancers in human genes. In: *Science* 297 (2002), Aug, Nr. 5583, S. 1007–1013

[29] Farrer, T. ; Roller, A. B. ; Kent, W. J. ; Zahler, A. M.: Analysis of the role of Caenorhabditis elegans GC-AG introns in regulated splicing. In: *Nucleic Acids Res.* 30 (2002), Aug, Nr. 15, S. 3360–3367

[30] Finkel, R. S. ; Mercuri, E. ; Darras, B. T. ; Connolly, A. M. ; Kuntz, N. L. ; Kirschner, J. ; Chiriboga, C. A. ; Saito, K. ; Servais, L. ; Tizzano, E. u. a.: Nusinersen versus Sham Control in Infantile-Onset Spinal Muscular Atrophy. In: *N. Engl. J. Med.* 377 (2017), 11, Nr. 18, S. 1723–1732

[31] Freund, M. ; Asang, C. ; Kammler, S. ; Konermann, C. ; Krummheuer, J. ; Hipp, M. ; Meyer, I. ; Gierling, W. ; Theiss, S. ; Preuss, T. ; Schindler, D. ; Kjems, J. ; Schaal, H.: A novel approach to describe a U1 snRNA binding site. In: *Nucleic Acids Res.* 31 (2003), Dec, Nr. 23, S. 6963–6975

[32] Fu, X. D.: The superfamily of arginine/serine-rich splicing factors. In: *RNA* 1 (1995), Sep, Nr. 7, S. 663–680

[33] Fu, X. D. ; Ares, M.: Context-dependent control of alternative splicing by RNA-binding proteins. In: *Nat. Rev. Genet.* 15 (2014), Oct, Nr. 10, S. 689–701

[34] Gao, K. ; Masuda, A. ; Matsuura, T. ; Ohno, K.: Human branch point consensus sequence is yUnAy. In: *Nucleic Acids Res.* 36 (2008), Apr, Nr. 7, S. 2257–2267

[35] Garcia-Blanco, M. A. ; Baraniak, A. P. ; Lasda, E. L.: Alternative splicing in disease and therapy. In: *Nat. Biotechnol.* 22 (2004), May, Nr. 5, S. 535–546

[36] Geuens, T. ; Bouhy, D. ; Timmerman, V.: The hnRNP family: insights into their role in health and disease. In: *Hum. Genet.* 135 (2016), Aug, Nr. 8, S. 851–867

[37] Grodecka, L. ; Buratti, E. ; Freiberger, T.: Mutations of Pre-mRNA Splicing Regulatory Elements: Are Predictions Moving Forward to Clinical Diagnostics? In: *Int J Mol Sci* 18 (2017), Jul, Nr. 8. `https://doi.org/10.3390/ijms18081668`

[38] Harrow, J. ; Frankish, A. ; Gonzalez, J. M. ; Tapanari, E. ; Diekhans, M. ; Kokocinski, F. ; Aken, B. L. ; Barrell, D. ; Zadissa, A. ; Searle, S. u. a.: GENCODE: the reference human genome annotation for The ENCODE Project. In: *Genome Res.* 22 (2012), Sep, Nr. 9, S. 1760–1774

[39] Hayer, K. E. ; Pizarro, A. ; Lahens, N. F. ; Hogenesch, J. B. ; Grant, G. R.: Benchmark analysis of algorithms for determining and quantifying full-length mRNA splice forms from RNA-seq data. In: *Bioinformatics* 31 (2015), Dec, Nr. 24, S. 3938–3945

[40] Herzel, L. ; Ottoz, D. S. M. ; Alpert, T. ; Neugebauer, K. M.: Splicing and transcription touch base: co-transcriptional spliceosome assembly and function. In: *Nat. Rev. Mol. Cell Biol.* 18 (2017), Oct, Nr. 10, S. 637–650

[41] Hoskins, A. A. ; Moore, M. J.: The spliceosome: a flexible, reversible macromolecular machine. In: *Trends Biochem. Sci.* 37 (2012), May, Nr. 5, S. 179–188

[42] Huber, W. ; Carey, V. J. ; Gentleman, R. ; Anders, S. ; Carlson, M. ; Carvalho, B. S. ; Bravo, H. C. ; Davis, S. ; Gatto, L. ; Girke, T. u. a.: Orchestrating high-throughput genomic analysis with Bioconductor. In: *Nat. Methods* 12 (2015), Feb, Nr. 2, S. 115–121

[43] Irimia, M. ; Blencowe, B. J.: Alternative splicing: decoding an expansive regulatory layer. In: *Curr. Opin. Cell Biol.* 24 (2012), Jun, Nr. 3, S. 323–332

[44] Jurica, M. S. ; Moore, M. J.: Pre-mRNA splicing: awash in a sea of proteins. In: *Mol. Cell* 12 (2003), Jul, Nr. 1, S. 5–14

[45] Kaisers, W. ; Boukamp, P. ; Stark, H. J. ; Schwender, H. ; Tigges, J. ; Krutmann, J. ; Schaal, H.: Age, gender and UV-exposition related effects on gene expression in in vivo aged short term cultivated human dermal fibroblasts. In: *PLoS ONE* 12 (2017), Nr. 5, S. e0175657

[46] Kaisers, W. ; Ptok, J. ; Schwender, H. ; Schaal, H.: Validation of Splicing Events in Transcriptome Sequencing Data. In: *Int J Mol Sci* 18 (2017), May, Nr. 6. https://doi.org/10.3390/ijms18061110

[47] Kaisers, W. ; Schaal, H. ; Schwender, H.: rbamtools: an R interface to samtools enabling fast accumulative tabulation of splicing events over multiple RNA-seq samples. In: *Bioinformatics* 31 (2015), May, Nr. 10, S. 1663–1664

[48] Kaisers, W. ; Schwender, H. ; Schaal, H.: Sample Size Estimation for Detection of Splicing Events in Transcriptome Sequencing Data. In: *Int J Mol Sci* 18 (2017), Sep, Nr. 9. https://doi.org/10.3390/ijms18091900

[49] Kaisers, Wolfgang ; Schwender, Holger ; Schaal, Heiner: Hierarchical clustering of DNA k-mer counts in RNA-seq fastq files reveals batch effects. In: *arXiv* arXiv:1405.0114 (2014)

[50] Ke, S. ; Shang, S. ; Kalachikov, S. M. ; Morozova, I. ; Yu, L. ; Russo, J. J. ; Ju, J. ; Chasin, L. A.: Quantitative evaluation of all hexamers as exonic splicing elements. In: *Genome Res.* 21 (2011), Aug, Nr. 8, S. 1360–1374

[51] Kelemen, O. ; Convertini, P. ; Zhang, Z. ; Wen, Y. ; Shen, M. ; Falaleeva, M. ; Stamm, S.: Function of alternative splicing. In: *Gene* 514 (2013), Feb, Nr. 1, S. 1–30

[52] Kohtz, J. D. ; Jamison, S. F. ; Will, C. L. ; Zuo, P. ; Luhrmann, R. ; Garcia-Blanco, M. A. ; Manley, J. L.: Protein-protein interactions and 5'-splice-site recognition in mammalian mRNA precursors. In: *Nature* 368 (1994), Mar, Nr. 6467, S. 119–124

[53] Li, H. ; Handsaker, B. ; Wysoker, A. ; Fennell, T. ; Ruan, J. ; Homer, N. ; Marth, G. ; Abecasis, G. ; Durbin, R.: The Sequence Alignment/Map format and SAMtools. In: *Bioinformatics* 25 (2009), Aug, Nr. 16, S. 2078–2079

[54] Lim, L. P. ; Burge, C. B.: A computational analysis of sequence features involved in recognition of short introns. In: *Proc. Natl. Acad. Sci. U.S.A.* 98 (2001), Sep, Nr. 20, S. 11193–11198

[55] Long, J. C. ; Caceres, J. F.: The SR protein family of splicing factors: master regulators of gene expression. In: *Biochem. J.* 417 (2009), Jan, Nr. 1, S. 15–27

[56] Lunde, B. M. ; Moore, C. ; Varani, G.: RNA-binding proteins: modular design for efficient function. In: *Nat. Rev. Mol. Cell Biol.* 8 (2007), Jun, Nr. 6, S. 479–490

[57] Manley, J. L. ; Tacke, R.: SR proteins and splicing control. In: *Genes Dev.* 10 (1996), Jul, Nr. 13, S. 1569–1579

[58] Marioni, J. C. ; Mason, C. E. ; Mane, S. M. ; Stephens, M. ; Gilad, Y.: RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. In: *Genome Res.* 18 (2008), Sep, Nr. 9, S. 1509–1517

[59] Martin, W. ; Koonin, E. V.: Introns and the origin of nucleus-cytosol compartmentalization. In: *Nature* 440 (2006), Mar, Nr. 7080, S. 41–45

[60] Matera, A. G. ; Wang, Z.: A day in the life of the spliceosome. In: *Nat. Rev. Mol. Cell Biol.* 15 (2014), Feb, Nr. 2, S. 108–121

[61] Matlin, A. J. ; Clark, F. ; Smith, C. W.: Understanding alternative splicing: towards a cellular code. In: *Nat. Rev. Mol. Cell Biol.* 6 (2005), May, Nr. 5, S. 386–398

[62] Mayeda, A. ; Krainer, A. R.: Regulation of alternative pre-mRNA splicing by hnRNP A1 and splicing factor SF2. In: *Cell* 68 (1992), Jan, Nr. 2, S. 365–375

[63] Mayeda, A. ; Munroe, S. H. ; Caceres, J. F. ; Krainer, A. R.: Function of conserved domains of hnRNP A1 and other hnRNP A/B proteins. In: *EMBO J.* 13 (1994), Nov, Nr. 22, S. 5483–5495

[64] Mercer, T. R. ; Clark, M. B. ; Andersen, S. B. ; Brunck, M. E. ; Haerty, W. ; Crawford, J. ; Taft, R. J. ; Nielsen, L. K. ; Dinger, M. E. ; Mattick, J. S.: Genome-wide discovery of human splicing branchpoints. In: *Genome Res.* 25 (2015), Feb, Nr. 2, S. 290–303

[65] Mortazavi, A. ; Williams, B. A. ; McCue, K. ; Schaeffer, L. ; Wold, B.: Mapping and quantifying mammalian transcriptomes by RNA-Seq. In: *Nat. Methods* 5 (2008), Jul, Nr. 7, S. 621–628

[66] MÃ $\frac{1}{4}$ller, Kirill ; Wickham, Hadley ; James, David A. ; Falcon, Seth: *RSQLite: 'SQLite' Interface for R*, 2017. `https://CRAN.R-project.org/package=RSQLite`. – R package version 2.0

[67] Nilsen, T. W.: The spliceosome: the most complex macromolecular machine in the cell? In: *Bioessays* 25 (2003), Dec, Nr. 12, S. 1147–1149

[68] Pan, Q. ; Shai, O. ; Lee, L. J. ; Frey, B. J. ; Blencowe, B. J.: Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. In: *Nat. Genet.* 40 (2008), Dec, Nr. 12, S. 1413–1415

[69] R Core Team: *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing, 2017. `https://www.R-project.org/`

[70] Raker, V. A. ; Plessel, G. ; Luhrmann, R.: The snRNP core assembly pathway: identification of stable core protein heteromeric complexes and an snRNP subcore particle in vitro. In: *EMBO J.* 15 (1996), May, Nr. 9, S. 2256–2269

[71] Ray, D. ; Kazan, H. ; Cook, K. B. ; Weirauch, M. T. ; Najafabadi, H. S. ; Li, X. ; Gueroussov, S. ; Albu, M. ; Zheng, H. ; Yang, A. u. a.: A compendium of RNA-binding motifs for decoding gene regulation. In: *Nature* 499 (2013), Jul, Nr. 7457, S. 172–177

[72] Romano, M. ; Buratti, E.: Targeting RNA binding proteins involved in neurodegeneration. In: *J Biomol Screen* 18 (2013), Oct, Nr. 9, S. 967–983

[73] Rosenberg, A. B. ; Patwardhan, R. P. ; Shendure, J. ; Seelig, G.: Learning the sequence determinants of alternative splicing from millions of random sequences. In: *Cell* 163 (2015), Oct, Nr. 3, S. 698–711

[74] Schaub, A. ; Glasmacher, E.: Splicing in immune cells-mechanistic insights and emerging topics. In: *Int. Immunol.* 29 (2017), Apr, Nr. 4, S. 173–181

[75] Schneider, T. D. ; Stephens, R. M.: Sequence logos: a new way to display consensus sequences. In: *Nucleic Acids Res.* 18 (1990), Oct, Nr. 20, S. 6097–6100

[76] Scotti, M. M. ; Swanson, M. S.: RNA mis-splicing in disease. In: *Nat. Rev. Genet.* 17 (2016), Jan, Nr. 1, S. 19–32

[77] Shepard, P. J. ; Hertel, K. J.: The SR protein family. In: *Genome Biol.* 10 (2009), Nr. 10, S. 242

[78] Smith, C. W. ; Valcarcel, J.: Alternative pre-mRNA splicing: the logic of combinatorial control. In: *Trends Biochem. Sci.* 25 (2000), Aug, Nr. 8, S. 381–388

[79] Steijger, T. ; Abril, J. F. ; Engstrom, P. G. ; Kokocinski, F. ; Hubbard, T. J. ; Guigo, R. ; Harrow, J. ; Bertone, P. ; Abril, J. F. ; Akerman, M. u. a.: Assessment of transcript reconstruction methods for RNA-seq. In: *Nat. Methods* 10 (2013), Dec, Nr. 12, S. 1177–1184

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

[81] Turunen, J. J. ; Niemela, E. H. ; Verma, B. ; Frilander, M. J.: The significant other: splicing by the minor spliceosome. In: *Wiley Interdiscip Rev RNA* 4 (2013), Nr. 1, S. 61–76

[82] Wang, E. T. ; Sandberg, R. ; Luo, S. ; Khrebtukova, I. ; Zhang, L. ; Mayr, C. ; Kingsmore, S. F. ; Schroth, G. P. ; Burge, C. B.: Alternative isoform regulation in human tissue transcriptomes. In: *Nature* 456 (2008), Nov, Nr. 7221, S. 470–476

[83] Wang, G. S. ; Cooper, T. A.: Splicing in disease: disruption of the splicing code and the decoding machinery. In: *Nat. Rev. Genet.* 8 (2007), Oct, Nr. 10, S. 749–761

[84] Wang, Y. ; Ma, M. ; Xiao, X. ; Wang, Z.: Intronic splicing enhancers, cognate splicing factors and context-dependent regulation rules. In: *Nat. Struct. Mol. Biol.* 19 (2012), Oct, Nr. 10, S. 1044–1052

[85] Wang, Y. ; Xiao, X. ; Zhang, J. ; Choudhury, R. ; Robertson, A. ; Li, K. ; Ma, M. ; Burge, C. B. ; Wang, Z.: A complex network of factors with overlapping affinities represses splicing through intronic elements. In: *Nat. Struct. Mol. Biol.* 20 (2013), Jan, Nr. 1, S. 36–45

[86] Wang, Z. ; Burge, C. B.: Splicing regulation: from a parts list of regulatory elements to an integrated splicing code. In: *RNA* 14 (2008), May, Nr. 5, S. 802–813

[87] Wang, Z. ; Gerstein, M. ; Snyder, M.: RNA-Seq: a revolutionary tool for transcriptomics. In: *Nat. Rev. Genet.* 10 (2009), Jan, Nr. 1, S. 57–63

[88] Wang, Z. ; Rolish, M. E. ; Yeo, G. ; Tung, V. ; Mawson, M. ; Burge, C. B.: Systematic identification and analysis of exonic splicing silencers. In: *Cell* 119 (2004), Dec, Nr. 6, S. 831–845

[89] Wu, J. ; Anczukow, O. ; Krainer, A. R. ; Zhang, M. Q. ; Zhang, C.: OLego: fast and sensitive mapping of spliced mRNA-Seq reads using small seeds. In: *Nucleic Acids Res.* 41 (2013), May, Nr. 10, S. 5149–5163

[90] Xiong, H. Y. ; Alipanahi, B. ; Lee, L. J. ; Bretschneider, H. ; Merico, D. ; Yuen, R. K. ; Hua, Y. ; Gueroussov, S. ; Najafabadi, H. S. ; Hughes, T. R. u. a.: RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. In: *Science* 347 (2015), Jan, Nr. 6218. https://doi.org/10.1126/science.1254806

[91] Yeo, G. ; Burge, C. B.: Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. In: *J. Comput. Biol.* 11 (2004), Nr. 2-3, S. 377–394

Anhang A

Abkürzungen

Abkürzung	Englischer Text	Deutscher Text
RNAseq	Transcriptome Sequencing	Transkriptom Sequenzierung
SAM	Sequence Alignment Format	Sequenz Alignment Format
BAM	Binary Alignment Format	Binäres Alignment Format
CIGAR	Concise Idiosyncratic Gapped Alignment Report	
5'Ss	5' Spleißstelle	5' splice-site
3'Ss	3' Spleißstelle	3' splice-site
IDIN	Intronic dinucleotides	Intronische Dinukleotide
gqs	Gap quality score	Gap Quality Score
wgis	Weighted gap information score	Gewichteter Gap Informations Score