2 Chapter II: Reproducible RNA-seq processing for detection and quantification of microexons

Collaboration note

Most of the work presented in this chapter are results that will be published as a separate manuscript in a peer reviewed journal. While I conceived the core idea of the initial computational analyses with Roberto Munita[1], the development of the software was exclusively performed by me. Close communication and interaction with Ilias Georgakopoulos-Soares[2] was beneficial for this project's development, who also tested the software in collaboration with Veronika Kedlian[3].

2.1 Introduction

The initial report of microexons dates back in 1985 for the *Ubx* gene which in Drosophila was found to contain two 5 nt microexons (Beachy et al., 1985). This discovery was followed by several other reports of constitutive and alternative microexons discovered in various vertebrate genes (Cooper and Ordahl, 1985; Santoni et al., 1989; Small et al., 1988; Ustianenko et al., 2017). Even though some of these microexons were found to be tissue-specific or regulated through brain development (Santoni et al., 1989; Small et al., 1988), systematic analyses of

[1] Former Ph.D. student at Department of Cellular and Molecular Biology, Pontificia Universidad Católica de Chile and current postdoctoral fellow at the Division of Molecular Hematology (DMH), Lund University.
[2] Former Ph.D. student at the Sanger Institute, co-supervised by Serena Nik-Zainal and Martin Hemberg. Current postdoctoral fellow at Department of Bioengineering and Therapeutic Sciences, Institute for Human Genetics, University of California San Francisco.
[3] Current Ph.D. student at the Sanger Institute, supervised by Sarah Teichmann.

microexons were obstructed by technical difficulties associated with their detection in mRNA sequences.

Initial gene annotation of model organisms was extensively carried out by mapping of expressed sequence tags (EST) and other cloned cDNA sequences (Dias Neto et al., 2000; Okazaki et al., 2002). However, the correct alignment of these cDNA sequences was acknowledged to be particularly challenging in the presence of microexons (Florea et al., 1998). The development of an algorithm to correct cDNA alignments, allowed for the detection of 224 previously unknown microexons across human, *Caenorhabditis elegans* and *Drosophila melanogaster* (Volfovsky et al., 2003). Further development of this strategy was directly implemented by GMAP, an EST/cDNA alignment tool, which also incorporated a statistical model to avoid reporting spurious microexons (Wu and Watanabe, 2005).

2.1.1 Computational methods for discovery and quantification of microexons using RNA-seq data

The advent of high throughput RNA sequencing technologies (RNA-seq) provided an unprecedented opportunity to explore the transcriptome. However, widely used RNA-seq platforms, such as Illumina, generate RNA sequencing reads that are shorter (50-150 nt) than the average EST length. Two main strategies have been developed for short RNA-seq read mapping; (1) Exon-first approach, in which reads are first mapped through ungapped alignment, enabling the mapping of reads within exonic regions. Subsequently, only unmapped reads undergo a second round of spliced read mapping. (2) Seed-extend approach, in which the read alignment process is subdivided into units of ungapped alignments, often referred as alignments seeds, and only seeds successfully mapped to the genome are extended (Garber et al., 2011).

However, the alignment of reads that span microexons has been identified as a particularly hard problem, which can prevent the correct alignment of reads unless the aligners have strategies implemented to align to microexons reads (Wu and Watanabe, 2005). Among the RNA-seq aligners that have proven to be more sensitive to microexon detection, there is Olego (Wu et al., 2013), which combines

exon-first and seed-extend approaches to perform RNA-seq alignments. In a first step, exonic reads are mapped using an approach similar to BWA (Li and Durbin, 2009), and during the second step, unmapped reads are split into alignment seeds to discover splice junctions through the seed-extend approach. The feature that makes Olego particularly sensitive to detect microexons is the use of small seeds during this last step to find splice junctions, which enabled Wu and collaborators to identify 1,665 microexons in mouse retina RNA-seq samples, 37.8% of which were not annotated, suggesting great discovery potential of RNA-seq analyses.

Systematic discovery and quantitative analyses of microexons using RNA-seq data have been performed by the implementation of pipelines that integrate multiple alignment steps. VAST-TOOLS (Irimia et al., 2014; Tapial et al., 2017) is a multi-module analysis pipeline that can quantify alternative splicing events measured as the “percent spliced-in” (PSI), which corresponds to the percent of the transcript that undergoes a particular splicing event (e.g. cassette exon inclusion). Irimia and colleagues developed a module to discover microexons using RNA-seq data which was based on bowtie alignments to an extensive library of possible exon-microexon-exon junctions (EEEJ) and then many of the discovered microexons were deposited in VASTDB. However, this module to discover microexon is currently unpublished and the public version of VAST-TOOLS is just restricted to quantify alternative splicing events that are annotated on VastDB, a comprehensive and curated database of splice sites (Tapial et al., 2017). Thus, microexon analyses with VAST-TOOLS are not suitable to discover and quantify microexons that are only included under certain experimental conditions (such as disease models) or even perform analyses in genome assemblies that are not included in VAST-TOOLS.

Li and collaborators developed a computational method called Augmented Transcript Mapping, ATMap (Li et al., 2015), which can discover novel microexons using RNA-seq data. ATMap first maps RNA-seq reads to annotated transcripts using Stampy (Lunter and Goodson, 2011). Then, alignments are processed to identify insertions at splice sites, which can be re-aligned into the intronic spaces flanked by canonical dinucleotides. Even though ATMap strategy was shown to be more sensitive for microexon discovery than traditional RNA-seq mappers, this software

has not been released to the public domain. Thus, even though these multi-step computational methods were proven to be very sensitive in the hands of their own developers, no one in the community has been able to use them.

2.1.2 Reproducible bioinformatics analysis using workflow manager platforms

Computational workflows to discover microexons have proven to be an effective way to tailor RNA-seq processing steps in a way that favours sensitive and specific discovery and quantification of microexon alternative splicing events (Irimia et al., 2014; Li et al., 2015). Both, ATMap and VAST-TOOLS microexon module, rely on multiple steps that are performed by software which was developed by third party academic groups. These computational software are often deposited in public repositories, such as GitHub, where multiple versions of a single bioinformatic tool may be released over time. Since a combination of different software versions across the software repositories that a given pipeline needs often leads to different results, reproducibility of workflow based methods is an important challenge.

A diverse range of workflow management systems (WMS) have been developed over time, but only a few have been consistently used by large communities of computational biologists (Di Tommaso et al., 2017; Goecks et al., 2010; Köster and Rahmann, 2012; Larsonneur et al., 2018; Leipzig, 2017; Wang and Peng, 2019). Different WMS have been designed to enhance bioinformatic reproducibility, however their design has been oriented to solve different needs. For example, some WMS are oriented towards enhancing the accessibility of computational tools for biologists with limited experience in bioinformatics. Galaxy (Goecks et al., 2010) and Taverna (Wolstencroft et al., 2013) provide web-based interfaces to build computational workflows without the need of any software installation or command-line execution. On the other hand, command-line based WMS, such as Nextflow (Di Tommaso et al., 2017) and Snakemake (Köster and Rahmann, 2012), enable the design of scalable computational pipelines that can work on a standard laptop as well as high-performance computing systems (HPCS) and cloud

environments. Nextflow and Snakemake enable the implementation of virtual environments and cloud containers that can fully ensure bioinformatic reproducibility.

2.1.3 Computational environments

Since bioinformatic workflows that depend on different combinations of software versions might limit the number of compatible workflows that can be used on a single computational environment, the use of environment managers has become essential for routine use of computational workflows. Conda (https://conda.io) is an open source package repository in which each computational software is available as relocatable binaries. This allows the dynamic building of isolated software without allowing system-wide administrator privileges and enables fine control of package versions. Within the computational biology community, the Bioconda project (Grüning et al., 2018) greatly expanded the bioinformatic tools available as Conda software packages from various language ecosystems such as Python, R, Perl, Java, C/C++ and Julia.

Snakemake enables a direct integration with Conda, which not only allows users to run and develop multiple computational workflows on a single workstation, but also allows the usage of different versions of software for the different steps of a single workflow. Each process within a Snakemake workflow is defined as a *rule* which contains the instruction to process input files and produce specific output files. Each *rule* can be assigned to its own conda environment, thereby enabling the use of software that would otherwise be incompatible. The fine control of the environment together with the extensive documentation have resulted in Snakemake being one of the most extensively used WMSs by the computational biology community.

2.2 Results

2.2.1 Development of a reproducible bioinformatic workflow to discover and quantify microexons in RNA-seq data

MicroExonator is a computational workflow that integrates several existing software packages with custom python and R scripts to perform discovery and quantification of microexons using RNA-seq data. MicroExonator can analyse RNA-seq data stored locally, but it can also fetch any RNA-seq datasets deposited in the NCBI Short Read Archive or other web-based repositories. As microexon annotations remain incomplete and sometimes inconsistent across different transcript annotations, MicroExonator can incorporate prior information from multiple databases such as RefSeq (Pruitt et al., 2014), GENCODE (Harrow et al., 2006), ENSEMBL (Hubbard et al., 2002), UCSC (Hsu et al., 2006) or VastDB (Tapial et al., 2017). To discover putative novel microexons, reads are first mapped using BWA-MEM (Li and Durbin, 2009) to a reference library of splice junction sequences. Misaligned reads are then searched for insertions located at exon-exon junctions. Detected insertions are retained if they can be successfully mapped to the corresponding intronic region with flanking canonical U2-type splicing dinucleotides (Sheth et al., 2006), decreasing the chances of spurious mapping by incrementing the length of the sequence that is aligned inside the intron (Fig 2.1a). To maximise the number of reads that can be assigned to each splice site, annotated and putative novel microexon sequences are integrated as part of the initial splice tags where they were detected. Reads are re-aligned with Bowtie, performing a fast but sensitive mapping of reads which is further processed to quantify PSI microexon values and perform quantitative filters (Fig 2.1b).

MicroExonator employs several filters to remove spurious matches to intronic sequences which may arise due to sequencing errors (Wu and Watanabe, 2005). To illustrate these filters I ran the initial mapping steps over RNA-seq from mouse corresponding to 289 RNA-seq samples from 18 different murine tissues and 1,657 single cells from mice visual cortex (Sloan et al., 2016; Tasic et al., 2016;

Weyn-Vanhentenryck et al., 2018). Given the large amount of spurious intronic matches that can introduce false positive microexon detection, MicroExonator implements a series of filters to provide a high confidence list of microexons (Fig 2.2a-b). As a first filtering step only those insertions that can be detected in a minimum number of independent samples (i.e. technical or biological replicates, three samples is set as default) are considered. Additionally, MicroExonator scores the sequence context of the detected canonical splice sites to measure the strength of their upstream and downstream splice junctions as quantified by a splicing strength score (Parada et al., 2014), and a Gaussian mixture model is used to exclude matches that have low U2 splice-site score values. (Fig 2.1b). Finally, MicroExonator integrates the splicing strength, probability of spurious intronic matching, and genomic conservation scores, in an adaptive filtering function to remove low confidence candidates. This final filtering step leads to a high quality list of microexons, where microexons that have a high probability of spuriously matching (generally microexon of 4 nt or shorter) are excluded (Fig 2.2c, Fig 2.2d). Further technical specifications and usability are included in the Appendix section.

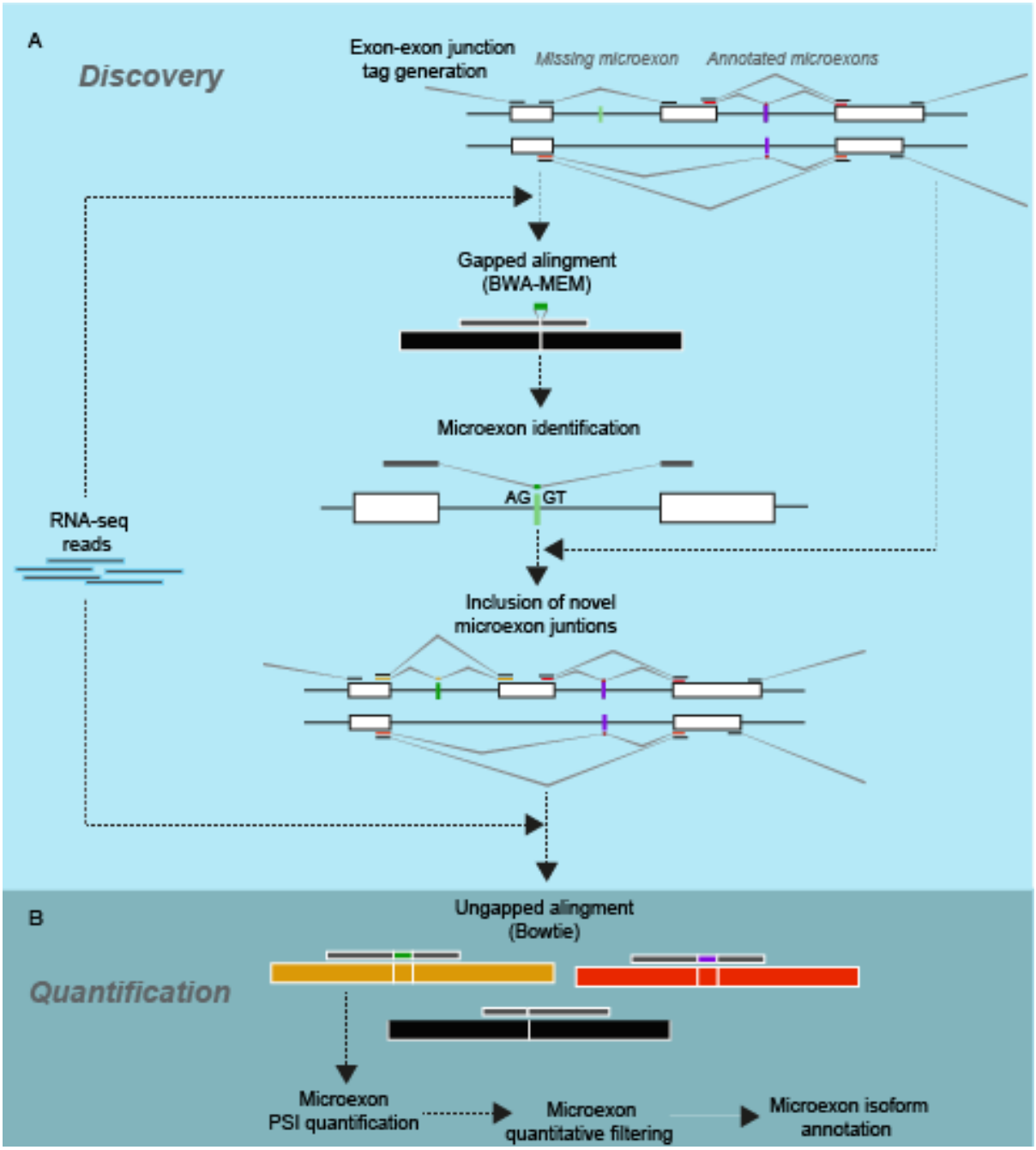

Figure 2.1: Overview of the MicroExonator workflow. A. To discover unannotated microexons, RNA-seq reads are aligned with BWA-MEM to the annotated splice junctions. The resulting alignments are post-processed to identify insertions at splice sites. Inserted sequences are tried to be mapped inside the corresponding introns with flanking GT-AG splice sites. **B.** Both putative novel and annotated microexons are quantified and filtered to produce a final list of microexons into transcript models which can be used for downstream analysis.

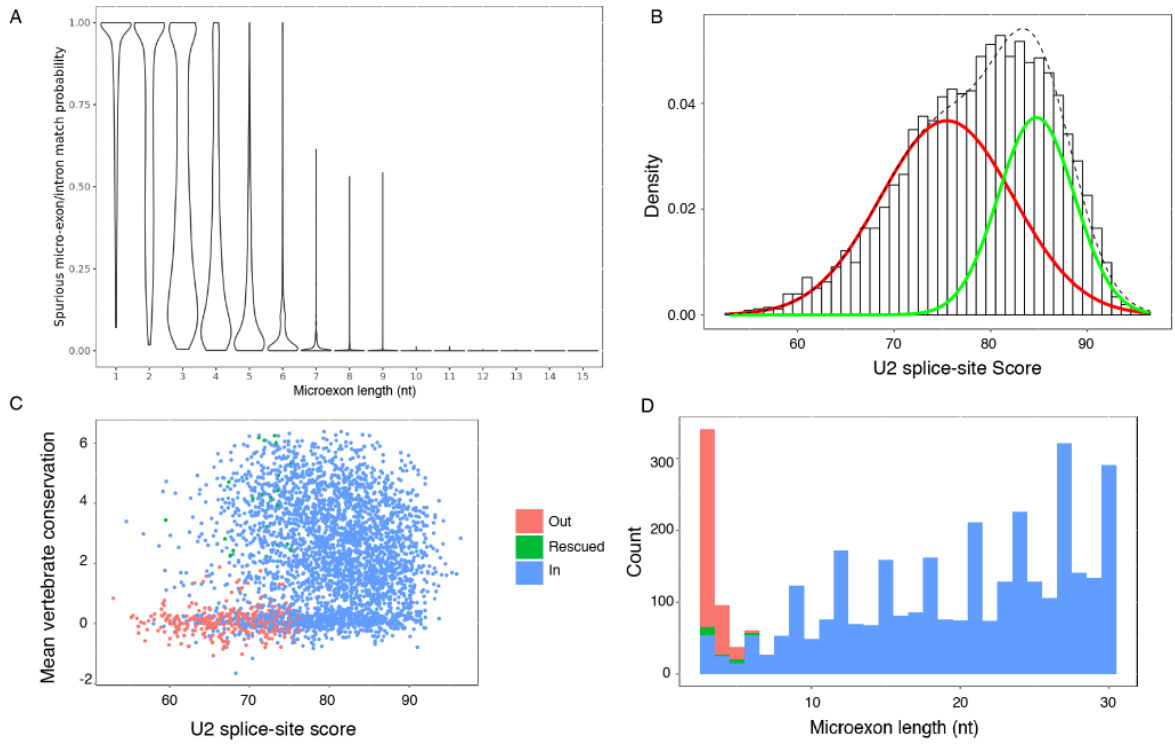

Figure 2.2: Quantitative microexon exon filtering. A. Probability of microexon spurious matches was calculated taking into account microexon length, splice site canonical dinucleotides and the size of the introns in which each microexon was discovered (see 2.3 Methods section) . **B.** A two component Gaussian mixture is used to fit the U2 consensus splicing score distribution. Lower U2 splice-site score gaussian curve (red line) is assumed to fit the distribution of spurious microexons, whereas true microexon distribution of U2 splice-site scores should be represented by a gaussian curve with higher U2 splice-site score (green line). **C.** Distribution of U2 splice-site score and mean vertebrate conservation values (phyloP score over microexons and their dinucleotides) for the total amount of candidate microexons before the final filters. Red dots represent microexons that were filtered out, blue dots microexons that were kept in the final high confident list of microexons and green dots microexons that were initially filtered out, but were rescued due to high conservation values (phyloP score ≥ 2 is used as the default value). **D.** Proportion of microexons that were filtered out, kept or rescued across different microexon sizes.

2.2.2 Benchmarking of computational methods for microexon discovery

To compare MicroExonator with other methods I incorporated a set of synthetic microexons into the GENCODE gene annotation (Fig 2.3a-b). The microexon sizes were drawn from the previously reported distributions (Irimia et al., 2014; Li et al., 2015) with greater abundance of in-frame microexons (Fig 2.3c). Moreover, I modified a copy of the mouse reference genome to replace the flanking intronic region of simulated microexons with sequences extracted from annotated splice sites. To simulate spurious microexons and evaluate their impact over the specificity of microexon discovery protocols, I randomly incorporated insertions across splice junctions. As these inserted sequences have the potential to map to intronic spaces, only microexon discovery protocols that have modules to statistically differentiate microexons from spurious matches are expected to perform well in my simulation.

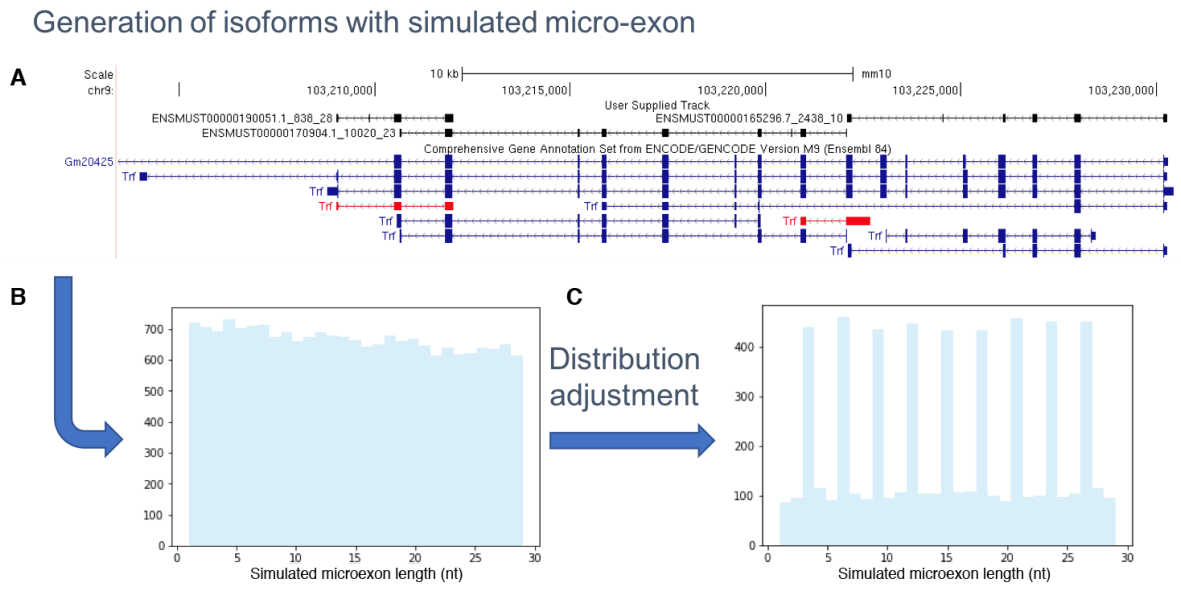

Figure 2.3: Ground truth generation for the assessment of microexon exon discovery modules. A. UCSC image showing the new isoforms generated by the insertion of simulated microexons. **B.** Raw size distribution of simulated microexons. **C.** Distribution adjustment to expected symmetric/asymmetric microexon proportions.

I used Polyester (Frazee et al., 2015) to simulate reads with a standard Illumina sequencing error rate and processed them using either MicroExonator, HISAT2 (Kim et al., 2015), STAR (Dobin et al., 2013), or Olego (Wu et al., 2013). The results show that the microexon filtering steps allow MicroExonator to distinguish simulated microexons from spurious microexons with a sensitivity >80% for all microexon lengths (Fig 2.4a-c). Even though all four aligners could detect a significant fraction of the simulated microexons, they are all limited in their ability to discover very short microexons; STAR's sensitivity drastically declines for microexons <10 nt, while the sensitivity of HISAT2 and Olego drops for microexons <8 nt. Moreover, the direct output of STAR and HISAT2's do not represent a reliable source of microexons, as they have low specificity. Using the default parameters results in a false discovery rate (FDR) of 43.0% and 33.3%, respectively. Olego had the highest specificity (FDR = 13.0%) of the other mappers, while MicroExonator achieves an FDR of 9.8%. Since MicroExonator's false discovery events are concentrated in the shortest microexons, discarding microexons <3 nt or <4 nt reduces the FDR to 2.4% and 0.75%, respectively.

The simulations also allow us to calculate the ground truth percent spliced in (PSI) values for the microexons, and to quantity how frequently a splice junction is incorporated in a transcript. MicroExonator is the only method that has low PSI errors for microexons <10 nt (Fig 2.4d). Even though MicroExonator's error rates are slightly higher for microexons >10 nt, they are still comparable to other methods. Taken together, these results show that MicroExonator is more accurate for annotating and quantifying microexons from RNA-seq data compared to conventional RNA-seq aligners.

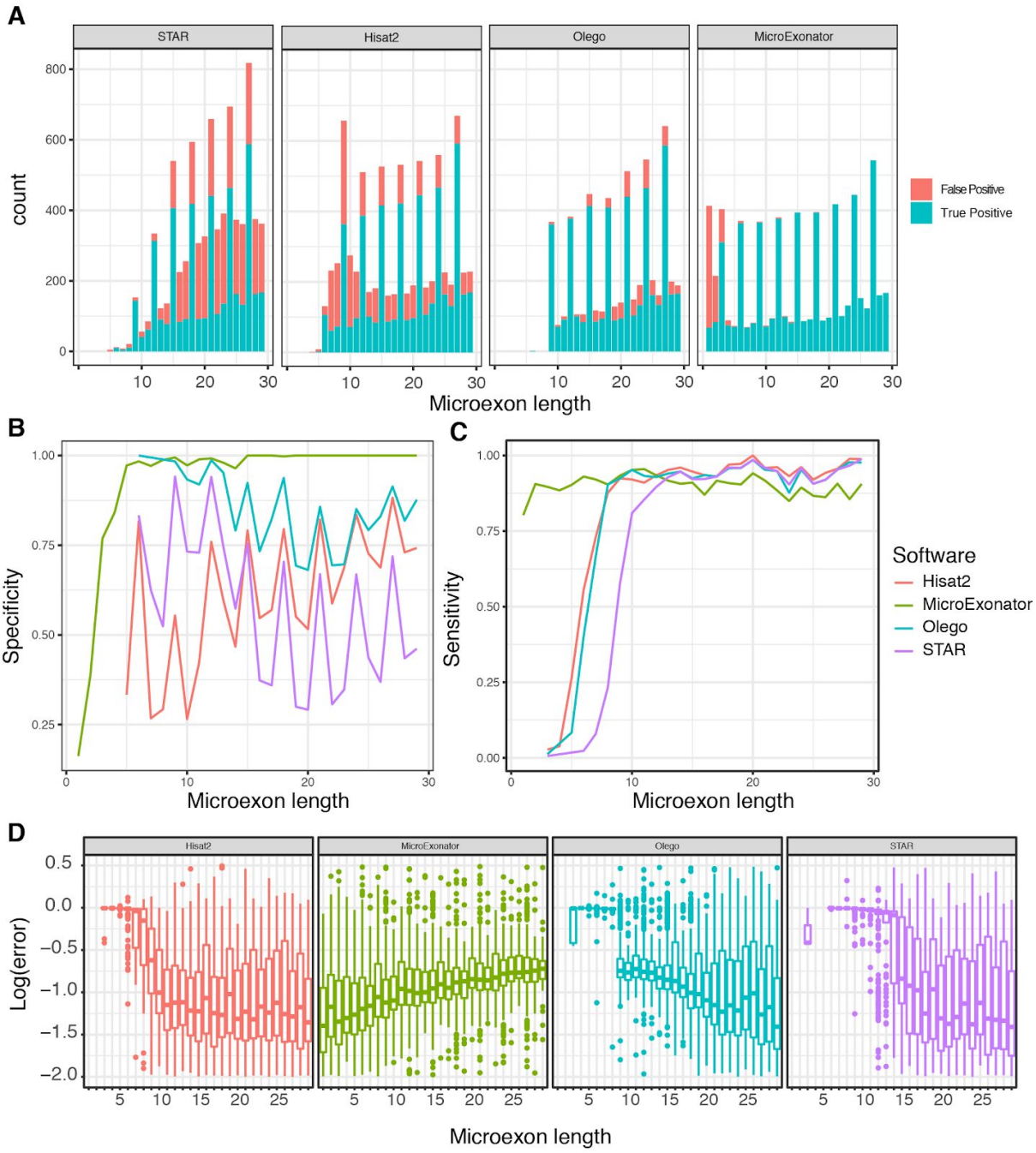

Figure 2.4: Evaluation of microexon discovery performance of RNA-seq aligners and MicroExonator using synthetic data. A. Size distribution of simulated microexons that were detected by the different software. **B-C.** Specificity and sensitivity of detected simulated microexons using multiple available tools for evaluation. **D.** Log10 error PSI values show the accuracy of the microexon quantification.

2.3 Methods

2.3.1 Annotation guided microexon discovery using RNA-seq data

MicroExonator was implemented over the Snakemake workflow engine (Köster and Rahmann, 2012), to facilitate a reproducible processing of a large number of RNA-seq samples. During an initial discovery module, MicroExonator uses annotated splice junctions supplied by the user (a gene model annotation file can be provided in GTF or BED format) to find novel microexons. RNA-seq reads are first mapped to a library of reference splice junction tags using BWA-MEM (Li and Durbin, 2009) with a configuration that enhances deletion detection (bwa mem -O 6,2 -L 25). The library of splice junction tags consists of annotated splice junctions between exons ≥ 30 nt and spanning introns ≥80 nt. For each splice junction, a reference sequence tag is generated by taking 100 nt upstream and downstream from the corresponding transcript sequence. Splice junction alignments are processed to extract read insertions with anchors ≥8 nt that map to exon-exon junction coordinates. Inserted sequences are then re-aligned inside the corresponding intronic sequence, but only matches flanked by canonical splice site dinucleotides (GT-AG) are retained (Fig 2.1a). The obtained reads are re-mapped to the reference genome using HISAT2 (Kim et al., 2017). A preliminary list of microexon candidates is generated based on reads whose insertions are aligned to the intronic spaces with no mismatches (soft clipping alignments are ignored). To further avoid misalignment artifacts, reads containing putative microexon sequences are mapped to the genome using HISAT2. Reads that map with higher mapping scores to the genome than the microexon junctions are discarded.

MicroExonator is currently available at GitHub (https://github.com/hemberg-lab/MicroExonator), where all the code and instructions on how to use it are available. Additional technical specifications and usability can be found in the appendix.

2.3.2 Quantification of microexon inclusion

In a subsequent quantification module, novel microexon candidates are integrated with the provided gene annotation to generate a second library of splice junctions tags, where putative novel loci from the discovery phase and annotated microexons are integrated at the middle of the tag sequences (Fig 2.1b). Reads are aligned again to this expanded library of splice junction tags using Bowtie (Langmead et al., 2009), which performs a fast ungapped alignment allowing for 2 mismatches (bowtie -v 2 -S). Reads that map to splice junction tags are also mapped to the reference genome using Bowtie also allowing two mismatches. Reads that could only fully map to a single splice junction tag but no other location are counted towards novel or annotated microexons.

2.3.3 Filtering of spurious intronic matches

MicroExonator uses a series of filters to distinguish real splicing events from spurious matches. For a random sequence of length L_s, where all four nucleotides have the same frequency, the probability of at least one spurious match inside an intron with flanking GT-AG dinucleotides can be calculated as:

Equation I. $$P_s = 1 - (1 - \frac{1}{4^{L_s+4}})^K$$

Where K is the number of k-mers of length $L_s + 4$ that are possible to extract from an intron of length L_i. K can be calculated as K = L_i - ($L_S + 4$). As only intronic matches that are flanked by canonical dinucleotides (4 nt) are allowed, the length of the sequence that is searched inside the intron corresponds to $L_s + 4$. Additionally, splice site signals are evaluated by measuring how well they match the canonical splicing motif as defined by the U2 position frequency matrices (Sheth et al., 2006). I call this U2-splice score or splice strength (normalized to range from 1 to 100), and it is used to build a two component Gaussian mixture model (Figure 2.2b).

Microexons shorter than 3nt cannot be identified with high specificity, and thus they are reported as a separate list. Microexons that are 3nt or longer, are prioritised according to a score (M_s) that is determined from the Gaussian mixture model probability and other parameters that are relevant for distinguishing real microexons from sequencing errors and other artefacts. The score is computed as:

Equation II. $$M_S = 1 - \frac{1 - P_S P_{U2}}{n}$$

Where P_{U2} is the probability of an intronic match, given a U2 Score, to belong to the component with higher U2 Score from the resultant gaussian mixture model and n is the number of intronic matches. During the final filter, microexons are prioritised according to M_s values.

An adaptive threshold to filter microexons by M_s values is calculated after every MicroExonator run. For this purpose, a linear model is used to fit the number of detected microexons as a function of their length, using different M_s values ranging between 0 and 1. MicroExonator suggests the M_s threshold under which the minimal residual standard deviation sum is obtained. A html report file is automatically generated at the end of every MicroExonator run, and it contains a plot of the variation of the sum of residual standard deviation values under different M_s thresholds.

By default, MicroExonator uses the suggested M_s score to filter out low scoring microexons, but the threshold can be set manually by the user. If conservation data (e.g. phyloP/PhastCons) is provided, then all low scored microexons that exceed a user-defined conservation threshold (default value = 2) are also included in the high confidence list of microexons and flagged as “rescued”.

2.3.4 RNA-seq simulation

I used simulations to evaluate the performance of different methods for microexon discovery and quantification. I used Polyester (Frazee et al., 2015) to simulate

RNA-seq reads from modified mouse GENCODE gene models (V11). To generate true positive microexons, I inserted a set of randomly selected sequences with a length of 1 to 30 nucleotides inside annotated introns longer than 80 nts (Fig 2.3a-b). At the same time, to simulate the splice site sequence distribution, I replaced splice site sequences from the simulated microexons with annotated mouse splice site sequences. In addition, to simulate spurious microexon matching (false positive microexons), I randomly included a set of insertions corresponding to intronic sequences at exon-exon junctions that were not flanked by canonical splicing sequences. The insertion rates and lengths were simulated parameters extracted from real RNA-seq experiments from postnatal forebrain samples. Taken together, our simulations provide a realistic set of false positive microexons that emulates real RNA-seq experiment condition as closely as possible.