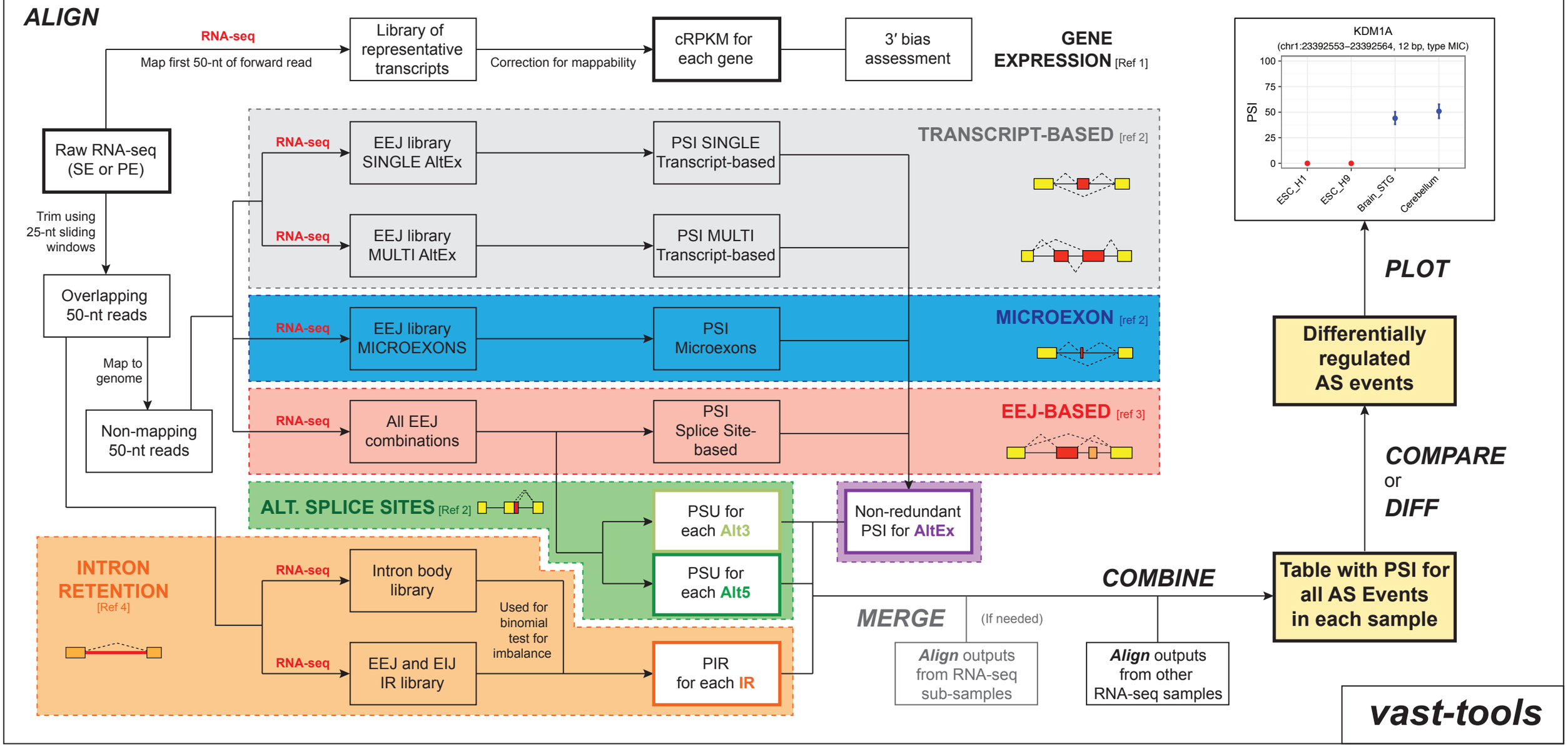
ALIGN
RNA-seq
Map first 50-nt of forward read
Library of representative transcripts
Correction for mappability
cRPKM for each gene
3′ bias assessment
GENE EXPRESSION [Ref 1]
Raw RNA-seq (SE or PE)
Trim using 25-nt sliding windows
Overlapping 50-nt reads
Map to genome
Non-mapping 50-nt reads
TRANSCRIPT-BASED [ref 2]
RNA-seq
EEJ library SINGLE AltEx
PSI SINGLE Transcript-based
RNA-seq
EEJ library MULTI AltEx
PSI MULTI Transcript-based
MICROEXON [ref 2]
RNA-seq
EEJ library MICROEXONS
PSI Microexons
EEJ-BASED [ref 3]
RNA-seq
All EEJ combinations
PSI Splice Site-based
ALT. SPLICE SITES [Ref 2]
PSU for each Alt3
PSU for each Alt5
Non-redundant PSI for AltEx
INTRON RETENTION [Ref 4]
RNA-seq
Intron body library
RNA-seq
EEJ and EIJ IR library
Used for binomial test for imbalance
PIR for each IR
COMBINE
MERGE
(If needed)
Align outputs from RNA-seq sub-samples
Align outputs from other RNA-seq samples
Table with PSI for all AS Events in each sample
COMPARE or DIFF
Differentially regulated AS events
PLOT
KDM1A
(chr1:23392553–23392564, 12 bp, type MIC)
PSI
100
75
50
25
0
ESC_H1
ESC_H9
Brain_STG
Cerebellum
vast-tools

Supplemental Figure S1 – Overview of the *vast-tools* pipeline workflow

vast-tools consists of various modules to quantify and then analyze AltEx, IR, Alt3 and Alt5 events, as well as gene expression levels. *vast-tools* uses raw FASTQ reads (single or paired end, top-left), which are first trimmed to 50-nt reads using 25-nt sliding windows to increase effective exon-exon junction coverage. These sub-reads are mapped to the genomic sequence, and the non-mapping reads will then be mapped to multiple exon-exon junction (EEJ) libraries. To quantify AltEx events (purple), *vast-tools align* uses three complementary modules: (i) the 'Transcript-based module' (gray), which relies on EEJs for single and multiple exon skipping events defined using full transcript information from ESTs, cDNAs, genome annotation and RNA-seq-based transcript annotations (e.g. cufflinks) [Ref 2 in figure: (Irimia et al. 2014)]; (ii) the 'Microexon module' (blue), which uses EEJs and exon-microexon-exon junctions to specifically quantify short exons (3-15 nucleotides) [Ref 2: (Irimia et al. 2014)]; and (iii) the 'splice site-based module' (red), whose EEJ library contains all possible forward combinations of exons for each gene [Ref 3: (Han et al. 2013)]. The outputs from the three modules are then merged to obtain a non-redundant PSI estimate for each alternative exon (purple box). In addition, the output of the splice site-based module is used to quantify the use of alternative 5' and 3' splice site choices (Alt3 and Alt5 events; green)[Ref 2: (Irimia et al. 2014)]. Finally, an independent module defines and quantifies percent of intron retention (orange), by mapping raw sub-reads to: (i) a library of EEJ and exon-intron junctions (EIJs), which are used to quantify retention levels; and (ii) a library of intron body fragments (200-nt in the middle of the intron) which are taken into account for the binomial test for read imbalance between the two EIJs [Ref 4: (Braunschweig et al. 2014)]. In parallel, the first 50 nucleotides of the reads (only the forward end, when paired) are mapped to a library of one representative isoform per gene to derive gene expression quantifications using the cRPKM metric [Ref 1: (Labbe et al. 2012)]. The PSIs for all types of AS events are then combined (together with any other RNA-seq samples analyzed using *vast-tools align*) into a unique table using *vast-tools combine*. This table can then be used to run *vast-tools compare* or *vast-tools diff* to identify differentially spliced AS events, and *vast-tools plot* to create figures displaying the PSIs across samples (further details and availability in https://github.com/vastgroup/vast-tools).