Matt: Unix tools for splicing analysis

André Gohr*  Manuel Irimia†
andre.gohr@crg.eu  mirimia@gmail.com

Published in Bioinformatics 2019
DOI: 10.1093/bioinformatics/bty606

*Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Dr. Aiguader 88, 08003
Barcelona, Spain
†Centre for Genomic Regulation (CRG), The Barcelona Institute for Science and Technology, Dr. Aiguader 88, 08003
Barcelona, Spain, and Universitat Pompeu Fabra (UPF), Barcelona, Spain
Summary: Tracking thousands of alternative splicing (AS) events genome-wide makes their downstream analysis computationally challenging and laborious. Here, we present Matt, the first UNIX command-line toolkit with focus on high-level AS analyses. With 50 commands it facilitates computational AS analyses by i) expediting repetitive data-preparation tasks, ii) offering routine high-level analyses, including the extraction of exon/intron features, discriminative feature detection, motif enrichment analysis, and the generation of motif RNA-maps, iii) improving reproducibility by documenting all analysis steps, and iv) accelerating the implementation of own analysis pipelines by offering users to exploit its modular functionality.

Availability and Implementation: matt.crg.eu under GNU LGPLv3, together with comprehensive documentation, and application examples. Matt is implemented in Perl and R, invokes pdfLaTeX, and depends only on Perl Core modules/the R Base package simplifying its installation.

Contact: andre.gohr@crg.eu, mirimia@gmail.com
1 Introduction

Having estimated inclusion levels of AS events (e.g., skipped exons, retained introns) from RNA-seq data genome-wide with dedicated software, subsequent steps of AS analysis pipelines often include the definition and extraction of events differentially or non-differentially spliced between conditions, discriminative feature detection, motif enrichment, and visual data exploration. Obstacles include i) laborious and repetitive data-preparation steps, ii) the need to combine different programs for different high-level analyses including cumbersome conversions between data formats, and iii) partly missing stand-alone programs to be called from within own pipelines; For example, web-servers exist for generating motif RNA-maps (rMAPS, Park et al. 2016, RBPmap, Paz et al. 2014), which visualize motif enrichment in the genomic proximity of AS events, but command-line-based programs are missing. As a consequence, researchers implement many custom scripts for solving highly similar tasks and rely on web services for some analyses, which is inefficient and error prone. Thus, the implementation of pipelines would be greatly facilitated by a Unix toolkit, similar in philosophy to, e.g., SAMtools and Bedtools, but with useful commands for routine high-level AS analyses and handy utility commands for AS data preparation.

Because such a toolkit is lacking, we propose Matt, the first Unix toolkit specifically tailored to AS analyses. All Matt commands are available through a unified POSIX-style command-line interface. Matt is modular; Each command solves a specific task and they are designed to be combined for complex operations. The I/O data format of most Matt commands is tab-separated tables which can be of any column structure, but often rows will correspond to AS events and columns to their features, similar to the established data frames in R. The table I/O format facilitates the application of Matt to output tables of the most common software for inclusion-level estimation of AS events and the extraction of AS events for further analysis with Matt. For analyzing AS events, Matt incorporates MaxEnt splice site models (Yeo and Burge 2004) for the estimation of splice site strength, and the program SVM-BPfinder (Corvelo et al. 2010) for the prediction of branch points and related features. Whenever possible, tables are processed row by row, minimizing memory requirements. The I/O POSIX-style interface and the I/O table standard format make possible piping and combining Matt commands with other scripting languages (e.g., Sed, Awk). Moreover, Matt facilitates reproducibility by automatically saving all Matt calls for later parameter verification. In addition, saving the sequence of performed Matt commands in shell scripts documents each analysis step together with its parameters and can be easily exchanged with collaborators or added as supplement to publications.

Because of its unique collection of 50 commands for routine, high-level AS analyses and data preparation which can be exploited in own Unix pipelines or as back-end of Galaxy, its flexibility to work with tables of any column structure, and data from any species with any preferred genome annotation in GTF format, the Unix toolkit Matt has great potential to facilitate AS research.

2 Overview of selection of Matt commands

Data preparation: Their purpose is to avoid that users repeatedly implement highly similar scripts for data preparation. Interesting commands are def_cats for defining groups of events by user-specified conditions on their features, get__seqs for retrieving sequences from FASTA files of AS events, splice sites, or genes, and the command stratify for the stratification of data sets with respect to potentially confounding features, e.g., sequence GC-content. All commands with details are described on the Matt website.

High-level analyses: Matt offers the extraction of 75/50 features of interest for studying exons/introns (get__efeatures/get__ifeatures), including GC content, length, splice
Figure 1: **Top:** Flow chart of example pipeline. *1, *2: here, users may *stratify* data sets to make similar their distributions of, e.g., exon length or sequence GC-content. **Bottom (result of *3):** Motif RNA-map for Nova2 with binding motif described by Perl regular expression \([CT]CA[CT]\{0,3\}[CT]\{0,1\}CA[CT]\{4,23\}[CT]CA[CT]\}\{2\}, which is an example of describing a clustered motif with variable spaces by regular expressions. Regions with significant enrichment (FDR ≤ 0.01, permutation test) are indicated by thick lines.

site strengths, and predicted branch points. With the commands `cmpr_exons/cmpr_introns`, these features get extracted and the Mann-Whitney U test is applied to the comparison of feature distributions between exon/intron groups automatically, allowing users to detect discriminative features. The commands `get_pwm_hits/get_regexp_hits` allow the detection of hits of binding motifs of RNA binding proteins (RBP). Motifs may be specified either by Position Weight Matrix models (PWM), or Perl REGular EXPressions (REGEXP) being capable of describing motifs with variable spacers or motif clusters shown to be important for RBPs (Cereda et al. 2014). Users can test motif enrichment (with a permutation test) comparing two sequence sets (`test_pwm_enrich/test_regexp_enrich`), and perform this analysis with all 297 RBP binding motifs (PWMs) from CISBP-RNA (Ray et al. 2013) with `test_cisbp_enrich`. Complementing the motif analysis, PDF graphics of motif RNA-maps (e.g., Fig. 1, bottom) can be generated for user-specified binding motifs (`rna_maps`), or for all 338 IUPAC consensus binding motifs of CISBP-RNA with `rna_maps_cisbp`.
3 Example analysis: Nova2 splicing factor

A flow chart of this analysis is depicted in Figure 1 top; A shell script with all steps and details is available on the Matt website, Section Examples. Having applied VAST-TOOLS (Tapial et al. 2017) to RNA-seq data from mouse brain of Nova2 knockouts (KO) and wild type controls (WT) (Saito et al. 2016), we obtained a table with rows corresponding to AS events (skipped exons, retained introns, alternative 3' and 5' splice sites) and columns to their inclusion levels across samples (called Percent Spliced In: PSI). With the command get_vast, specifically designed to extract AS events of interest from vast-tools tables, we obtained all skipped exons, their PSI values across samples, ΔPSI values (KO vs. WT), and gene IDs in form of a new exon table. Next, we applied def_cats to this exon table and obtained, according to their ΔPSI value, 152 silenced exons (ΔPSI ≥ 25), 167 enhanced exons (ΔPSI ≤ -25), 322 unregulated exons (abs(ΔPSI) ≤ 1), and all the rest of exons. The exon table was extended by a column GROUP with group IDs silenced, enhanced, unregulated, NA, accordingly. Next, with the command cmpr_exons, we compared the three groups and obtained i) a table with extracted features for the exons, and ii) a PDF report with box plots visualizing the distributions of features for all groups and information on features with significantly different distributions between groups. We find that the median exon length of enhanced exons is significantly shorter than for silenced and unregulated exons (page 15 of report), and upstream introns of enhanced exons have significantly stronger branch points than those introns of unregulated exons (page 44). Next, we obtained a PDF report with motif RNA-maps for all 338 CISBP-RNA IUPAC binding motifs with rna_maps_cisbp, and a specific Nova2 RNA-map for a REGEXP (motif repeats & variable spacers) derived from known binding sites (Ule et al. 2006) with rna_maps. The latter (Fig. 1, bottom) reveals a Nova2 motif enrichment upstream of silenced exons and downstream of enhanced exons, confirming previous findings (Ule et al. 2006). Last, we obtained 50 nt long sequences upstream of the exons with get_seqs and applied test_cisbp_enrich to them. The results, comprehensively documented as a table, reveal significant (p value ≤ 0.001) motif depletion for RBPs Srsf2, Srsf12, Eif4b, HnrnpA2/B1 comparing silenced vs. unregulated exons, and a significant motif enrichment for Nova2.

This work has been supported by the European Research Council (ERC-StG-LS2-637591, ERC-AdG-MASCP-670146), the Spanish Ministry of Economy and Competitiveness (BFU2014-55076-P, BFU2014-55058-P), and the CRG (SEV-2012-0208), a Centro de Excelencia Severo Ochoa 2013-2017. We thank GK. Sandve, N. Morais, S. Bonnin, M. Ferreira, N. Agostinho, K. Ha., and S. Bajew for valuable comments on the manuscript.

References


