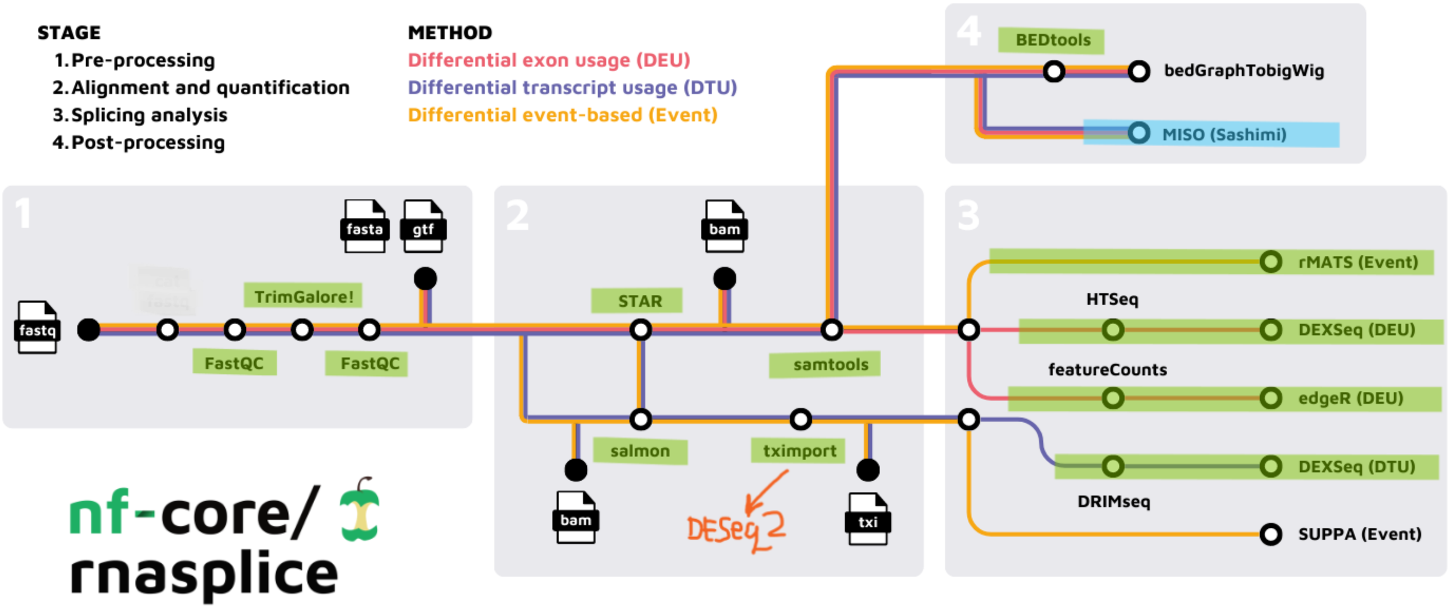
Progress until Aug 21st



Left to do

After many technical hurdles on the way, I managed to run the vast majority of branches of the pipeline. What I still need to run is:

1. fix and run SUPPA (lowermost branch), an alternative splicing event detector (is this exon included or skipped? is a larger or smaller chunk of it included?). I think the issue stems from some prefixes that appear in front of gene and transcript identifiers, similar to what I struggled with with tximport before it. I solved that, so should be able to solve this as well. When I do, the run takes ~2h.
2. DESeq2 is not part of this pipeline per se, but it would be nice to also have an idea of differential gene expression (DEG) analysis between the two conditions, not just alternative splicing. This would be a good control if the results are similar to those in the original publication. DESeq is an R package, and we did a DESeq2 analysis with Ioana as well, so I have it up and running. Once the data is set up, the actual run takes a few minutes.
3. bedGraphTobigWig is just a small script that creates BigWig files which are easier to visualize in the genome browser than the giant BAM files that are normally output from the alignment with STAR and Salmon. Not critically necessary, but it'd be nice to make this available to the public.

Why delayed

I'm also attaching the Markdown and PDF files where I document my progress (I'm at 127 pages now). This took ages longer than expected - when I started the pipeline Monday last week, it BSODed my entire computer, not just the virtual machine. I spent a day and a half just making extra extra backups and progressively changing settings to make sure the virtual machine is truly contained and can't overwhelm my Windows host. Then, it was at least one day to fix the issue with tximport, and another day because I also tried running branches of the pipeline separately, and the one with rMATS worked but, then, one of the basic underlying tools, samtools, started misbehaving, so I had to abandon the remainder of that run, and started the pipeline again from upstream, from before the samtools point to get the DEXseq and edgeR branches running.

Figures

I did look into the figures from the presentation, and identified a total of 28 figures that could use a touch-up or two (see attachment and below). I will start out on the ones that I already have the data for or where I can simply simulate it, but there are quite a few where I need the data from you, if you'd like me to re-make them (annotated in red). Also, for some of them, I realized I don't recall the reasoning behind their existence in detail (green boxes) - I realize you might not have the time to answer me in this regard, but, should you find it, I'd be spared going through the course videos and could be a lot faster about it.

Intro slides (23.5.23 version)

(1) S13 Random sampling schools IQ need data
what was point of this? (→ videos)

(2) S21 Sample size estimation
what was point of gr. 1? (→ videos) need data
(3) gr. 2? need data
make graphs that fit w/o covering one another

(4) S 22 swarm vs box plot need data
- make with nicer labels, labels for elements
- larger font

(5) S 24 bar plots diameter... of something need data
- needs translating to English
- what's happening with resolution?
- larger font

(6) probability density plot need data
- larger font

(7) S25 genotype⁺ trained vs untrained → muscle fibre necrosis
- what was point? (videos!) need data
- title for plot
- better separation on X axis of conditions, maybe soft, faint vertical zebra pattern for eye to follow which points belong to each category
- brackets for significance indicators to clearly see what was tested (instead of lines)

(8) S33 Test-retest reliability need data
[- any way to also depict validity?]
? Bland-Altman plots - what are they showing?
- video
- Internet
- better axis labels & titles?
- need translating

(9) S41 - "Area Ivot" simulate data?
- what is it about?
- translate to English

(10) S43 Poisson distribution: better resolution simulate data

(11) S44 binomial distribution: better resolution simulate data

(12) S52 unpaired vs paired t-test: simulate data?
- version of graph where dots from one pair same color btw the two existent graphs

(13) S56 - 1 - normal distribution: translation
(14) S56 - 2 - chi square distr.: translation
+ better reso

(15) S63 - linear regression: translation

(16) S66 - 1
- what was ACE? need data
- why the "1%" indicator instead of p-val?

(17) S66-2:
- what is PD sys RR? need data
- needs translation
- what's depicted? treatment of same patient after 1st and second dose?
- better title

(18) S70: variance / SumSq: better color than neon green on gray bg. Maybe make 3 samples in diff. colors & show SumSq almost overlapping but not fully simulate data

(19) S71: ibidem simulate data

(20) S72: ANOVA post-hoc p val vs repeated comparison p-val
- significance brackets indiv. comp. + 1 bracket w/ 3 prongs for ANOVA p-val have data?
- what was on y axis? 95% ci?

(21) S74: univariate ANOVA
- translation have data
- remake graph (we have cell data available!)
- change name of treatment "Whatever" to smth like "Treatment 2"

(22) S77: repeated measures ANOVA need data
- translation

(23) S84: odds ratios infertility
- what's odd diagonal line?
- make lines thicker overall, hard to see

(24) S85 - passage vs growth box plots fact & unfact - translate y axis have data
(S86 - replace with one on S74)

(25) S88 - biopsies + tissues + animal replicates
- change "T" to "A" for clarity
- do we have data? simulated LMM2024 script

(26) S91 Life expectancy
- translate need data
- table has typo

(27) S93 - cluster analysis
- what is graph on bottom? need data
- green points & lines on gray bg - change

(28) S97 regression tree need data
- could use higher reso, nicer font + some color

Working plan

My current working plan is:

Weeks 34/35
Remainder of this week and until end of next week
- read up a bit more on edgeR, rMATS, DEXSeq, and SUPPA to make sure I understand all of the expected outcomes
- check all of the results files for basic plausibility - do formats look correct?
- in parallel, fix SUPPA, run it and DESeq2
- fix the 9 figures I already have data for or can simulate it

Week 36
- import results files into tibbles & cleanup for differential gene expression results (DESeq2)
- import results files into tibbles & cleanup for alternative splicing results (rMATS, DEXSeq, SUPPA, edgeR)
- figures with descriptive stats, pointing out unusual events
- heat map DEGs
- volcano plot DEGs
- GO term enrichment analysis DEGs
- differentially regulated exons
- transcripts with differential abundance control vs pre-eclampsia
- any other interesting alternative splicing events other than exons
- figures 11-20

Week 37
- decide which parts of figures should be interactive
- how can I make interactive? Do I need to build a small SQL database that can be accessed by R Shiny app?
- start making interactive result presentation
- figures 20-30

Week 39
- interactive result presentation
- split scripts into blocks

Week 40
- buffer
- split scripts into blocks
- if there is time: train some model, e.g. find gene co-expression and exon co-splicing networks.