

Control error model for longer reads and polyA-tales in bed alignments.

Dear Flux authors/maintainers/developpers,

We (I plus Nicolas Philipe & Mikaël Salson) have been using FluxSimulator quite a lot recently in order to create a fair benchmark for RNA-Seq mapping softwares, and we must say that it is a really good piece of software!

However we have faced some difficulties in the design of our simulated libraries :

The first difficulty concerns the error model that Flux uses and generalizes for longer reads. In our simulations we want to produce RNA-Seq data with longer reads, as the trend seems to go that way (200bp or greater). But when we do that, the scaled error model generates a lots of errors, and I can understand that. But I think that the errors rate will get lower as technology will evolves. What I am asking you is how could we tune your errors model, to make fewer mistakes (say 1%) on longer reads? I have seen that we can generate custom errors models, but since that data we want to simulate does not already exists this is not a possible alternative. What are your thoughts about that?

The second difficulty concerns the polyA-tales simulations and the alignments reported in the bed file. Some reads does have a part of the polyA-tale and a part that comes from the genome, but the alignment does not mention the polyA-tale part. It is then difficult for us to assess mapping position of those reads. Have I missed something?

Thank you in advance for your answers,

Best regards,

Jérôme Audoux.