RESEARCH PROJECT PROPOSAL

Benchmark Analysis of Algorithms for Determining Full-Length Isoforms by Nanopore RNA-Seq Data

Emiliano Navarro Garre, BSc in Genetics (2021-2022), Universitat Autónoma de Barcelona

BACKGROUND

Transcriptome study presents a great importance, since **errors in splicing processes** are associated with diseases such as **cancer**, **dystrophies** or **Alzheimer's** disease, giving rise to different transcripts. Their characterisation is the first step to develop an effective gene therapy.

Traditionally, Illumina technology has been used to identify and quantify these transcripts, however, with the emergence of long-read methods, such as the **Oxford Nanopore Technology** (ONT), and programs that take these data as input, this **new approach** can be **much more accurate**.

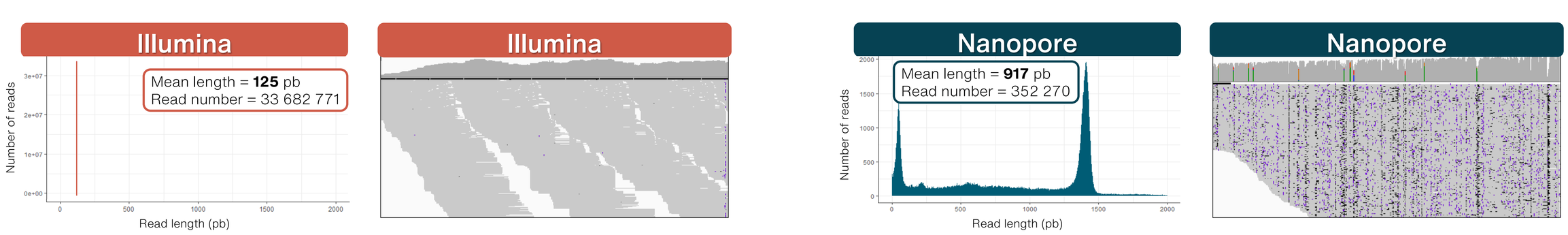

Fig. 1. Differences in lenght distribution and coverage between short-read (Illumina) and long-read (Nanopore) sequencing methods.

HYPOTHESES

- **ONT** has **great advantages in isoform detection and characterisation**.
- These **programs** can have **similar or higher accuracies** than methods using short-read sequencing data.
- It is unknown whether the published programs work correctly with all datasets or just with the program developers' own datasets.

OBJECTIVES

- To understand **how algorithms** capable of characterising isoforms **using long-read** RNA-seq data **work**.
- To determine **which parameters** of the algorithms **should be modified** according to the characteristics of the sample and **optimise** them: **BENCHMARK ANALYSIS.**
- To determine **which algorithms** and their parameters **work best** with different sample types to **assist** in the **choice** of algorithm based on **sample characteristics**.

METHODS

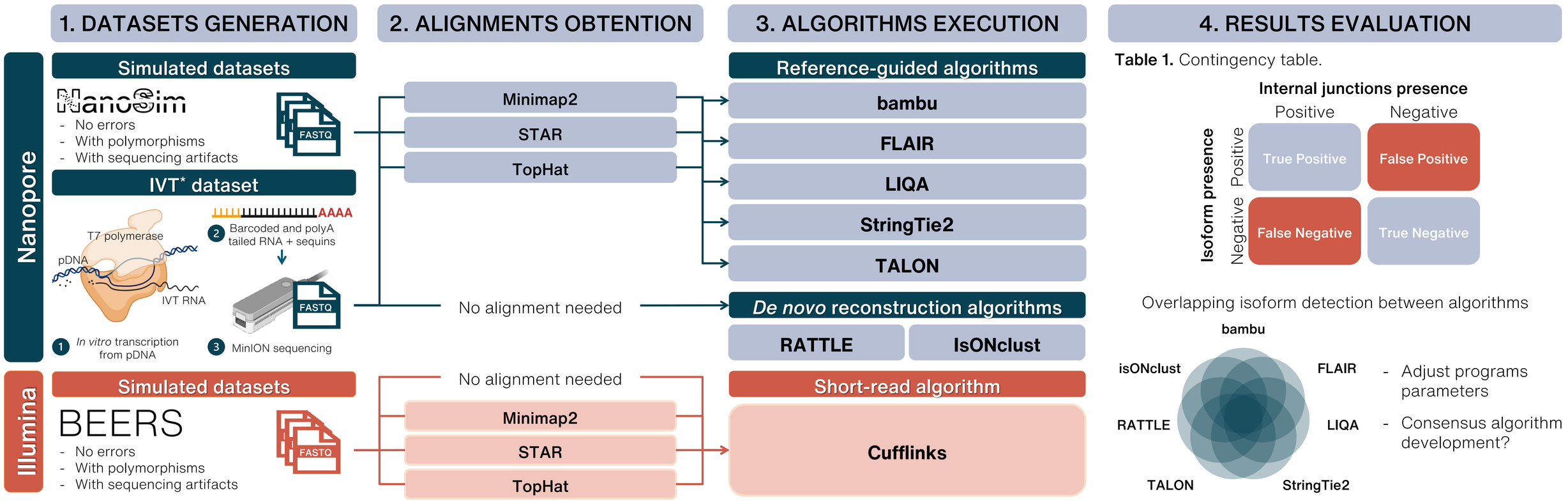

Table 1. Contingency table.

Isoform presence \ Internal junctions presence	Positive	Negative
Positive	True Positive	False Positive
Negative	False Negative	True Negative

Overlapping isoform detection between algorithms

- Adjust programs parameters
- Consensus algorithm development?

Fig. 2. Steps to carry out the project. ***IVT**: *In vitro* transcription.

EXPECTED RESULTS

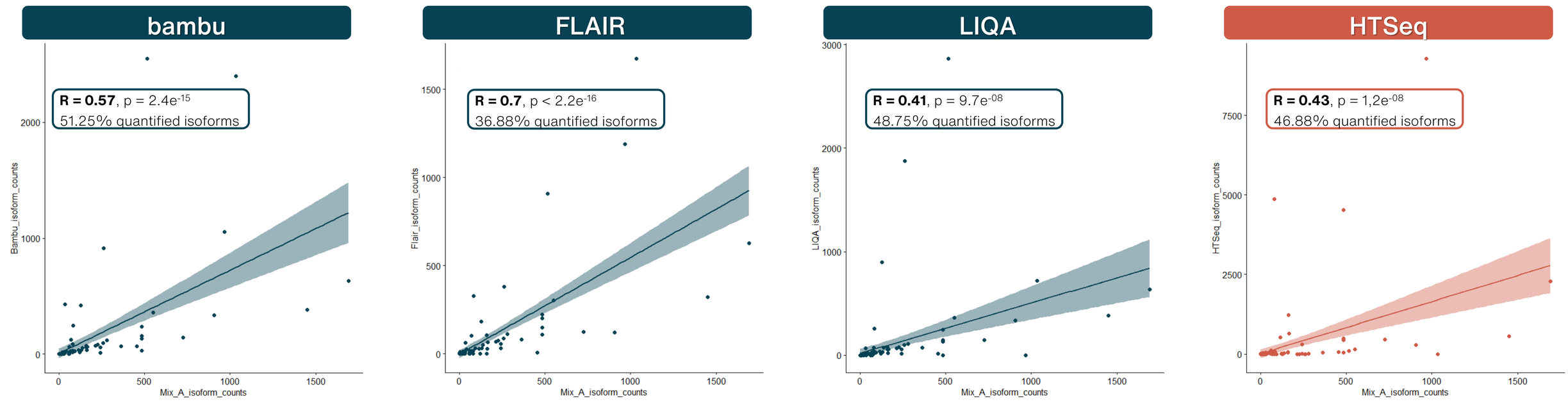

Fig. 3. Correlation between the isoform quantification performed by each tested algorithm and the expected quantification in mixture A of the sequins.

RELEVANT REFERENCES

- Hayer, K., et al. "Benchmark analysis of algorithms for determining and quantifying full-length mRNA splice forms from RNA-seq data." *Bioinformatics* 31.24 (2015): 3938-3945.
- Chen, Y., et al. "bambu: Reference-guided isoform reconstruction and quantification for long read RNA-Seq data." *Preprint publication*. *R package* version 2.2.0 (2022).
- Tang, A., et al. "Full-length transcript characterization of SF3B1 mutation in chronic lymphocytic leukemia reveals downregulation of retained introns." Nature communications 11.1 (2020): 1-12.
- Hu, Y., et al. "LIQA: long-read isoform quantification and analysis." Genome biology 22.1 (2021): 1-21.
- Putri, G., et al. "Analysing high-throughput sequencing data in Python with HTSeq 2.0." Bioinformatics 38.10 (2022): 2943-2945.

Scan the QR code to find out the steps to obtain the expected results!