Generation of gene annotations including UTRs for Bulk and Single-Cell RNASeq analyses

Brunon Salomé, Laurent Jourdren, Sophie Lemoine

INSTITUT DE BIOLOGIE DE L'ÉCOLE NORMALE SUPÉRIEURE
GenomiqueENS – Plateforme génomique

Generation of gene annotations including UTRs for Bulk and Single-Cell RNASeq analyses

Salomé Brunon[1], Laurent Jourdren[1], Sophie Lemoine[1]

1. GenomiqueENS, Institut de biologie de l'ENS (IBENS), Département de biologie, École normale supérieure, CNRS, INSERM, Université PSL, 75005 Paris, France

Background

GenomiqueENS Specialization: Functional genomics projects across diverse organisms, from model species to exotic ones

- **Bulk RNA sequencing: Illumina and Nanopore (ONT)**
- **Single-Cell RNA sequencing: 10X Genomics with Illumina and ONT**

ONT since 2017: Enables transcript quantification without exon discrimination models

In this poster:

- **Impact of poor annotation, especially in single-cell experiments**
- **Exploration of solutions using both short and long-read sequencing technologies**
- **Introduction of a suitable annotation protocol primarily utilizing ONT data**

The necessity of reannotation

In 2021, we conducted a **single-cell project on an invertebrate using the 3' 10X Chromium protocol**, **sequenced** on our **Illumina** platform. We found:

CellRanger quality control report

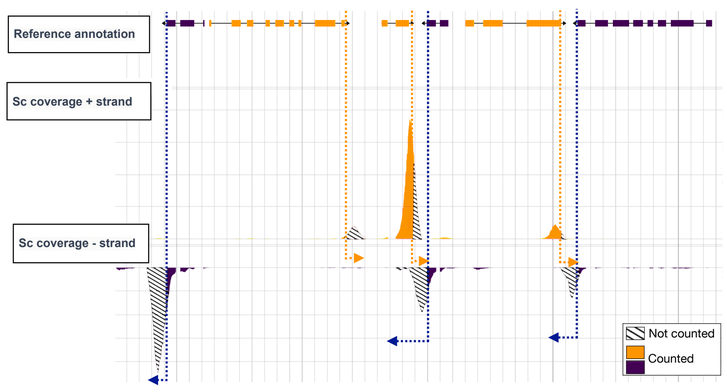
Mapping

Reads Mapped to Genome	82.0%
Reads Mapped Confidently to Genome	79.9%
Reads Mapped Confidently to Intergenic Regions	46.3%
Reads Mapped Confidently to Intronic Regions	3.6%
Reads Mapped Confidently to Exonic Regions	30.0%
Reads Mapped Confidently to Transcriptome	23.3%
Reads Mapped Antisense to Gene	0.5%

- **Only 30% of reads mapped to exons**, unexpected based on prior experiences
- **46% of reads mapped to intergenic regions**
- **82% of reads aligned to the genome**, indicating no data defects

Question: Why such a low percentage of reads mapping to exons?

Visual inspection using a genome viewer reveals:



Majority of 10X single-cell data does not overlap with reference annotations, hindering quantification

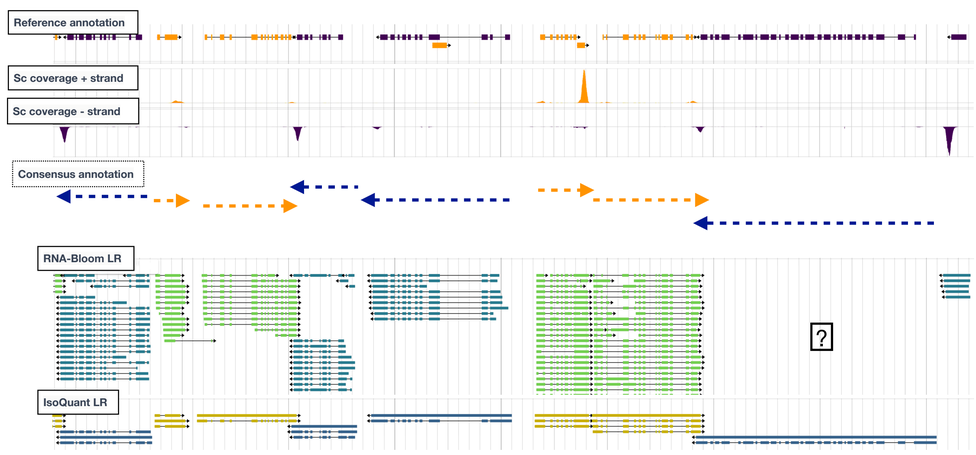
➡ **Annotations depend heavily on coding signals and protein databases** (e.g., GALBA, GeneMark-ETP, Augustus, BRAKER)

➡ **Consequently, non-model organisms often lack complete UTRs**

IsoQuant and RNA-Bloom

Evaluations from the LRGASP[6] Encode challenge led us to test IsoQuant[7] and RNA-Bloom[8].

- **IsoQuant**: a tool designed for **the genome-based analysis of long RNA reads**
- **RNA-Bloom**: a fast and memory-efficient **de novo transcript sequence assembler**.



- **RNA-Bloom captures a comprehensive set of transcripts, identifying nearly all potential transcripts**
- **When RNA-Bloom fails, IsoQuant serves as a reliable alternative**
- However, **RNA-Bloom produces an excessive number of transcripts**, complicating annotation
- **IsoQuant helps by providing consensus**, simplifying RNA-Bloom's complex output **at the isoform level**, and making annotation more manageable and interpretable

CellRanger quality control report

Mapping	IsoQuant	RNA-Bloom
Reads Mapped to Genome	82.0%	85.2%
Reads Mapped Confidently to Genome	80.2%	83.7%
Reads Mapped Confidently to Intergenic Regions	11.4%	0.6%
Reads Mapped Confidently to Intronic Regions	1.4%	1.0%
Reads Mapped Confidently to Exonic Regions	67.4%	82.1%
Reads Mapped Confidently to Transcriptome	66.1%	82.1%
Reads Mapped Antisense to Gene	2.0%	1.0%

RNA-Bloom counts the entire SingleCell signal

IsoQuant is less exhaustive than RNA-Bloom

In conclusion, a consensus between IsoQuant and RNA-Bloom strategies offers a promising and balanced approach, leveraging the strengths of both tools.

StringTie2 annotation: potential inaccuracies

StringTie2[1] is a good candidate for building a new annotation:

Widely used for its flexibility with short and long reads, with or without annotations

Known for being user-friendly and convenient for annotation tasks

Available inputs

- **Reference Annotation**: Existing annotations needing improvement (**Annot**)
- **Illumina Bulk RNA-Seq Data**: Sequenced in 2019, providing stranded short reads (**SR**)
- **ONT Bulk RNA-Seq Data**: Sequenced in 2021, providing long reads and restranded data (**LR**)

Impact of the new annotation on gene expression counts

CellRanger quality control report

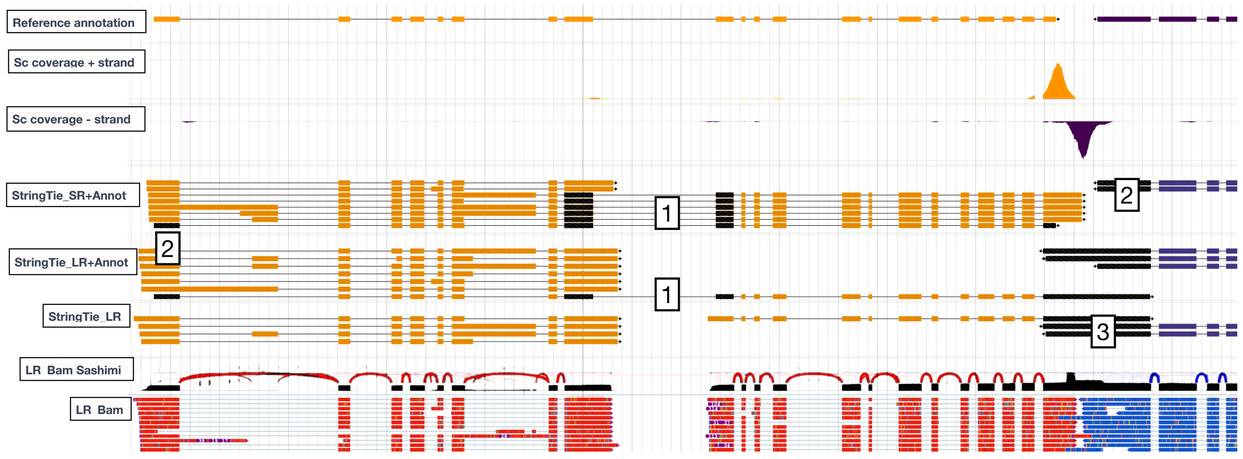
SingleCell counts on Reference annotation

Mapping	
Reads Mapped to Genome	82.0%
Reads Mapped Confidently to Genome	79.9%
Reads Mapped Confidently to Intergenic Regions	46.3%
Reads Mapped Confidently to Intronic Regions	3.6%
Reads Mapped Confidently to Exonic Regions	30.0%
Reads Mapped Confidently to Transcriptome	23.3%
Reads Mapped Antisense to Gene	0.5%

SingleCell counts on StringTie2 annotation (SR+annot)

Mapping	
Reads Mapped to Genome	82.0%
Reads Mapped Confidently to Genome	80.3%
Reads Mapped Confidently to Intergenic Regions	21.3%
Reads Mapped Confidently to Intronic Regions	2.3%
Reads Mapped Confidently to Exonic Regions	56.7%
Reads Mapped Confidently to Transcriptome	50.5%
Reads Mapped Antisense to Gene	1.9%

Validity of the new annotation is visually verified:



Provided annotation can bias the model against the data

1 Annotation retains fusion between genes despite evidence of separation

2 Proposes exons lacking evidence.

Without annotation, results may not consider input data

3 Annotation shows UTR extensions without clear links to input data.

Integrated into many annotation pipelines (Funannotate[2], PASA[3], BRAKER3[4], nf-core/nanoseq[5]…)

➡**Pipelines supplement RNA-Seq data** with protein data from related species and additional evidence

➡**Expect results to align more closely with experimental observations**

CellRanger quality control report

Mapping

Reads Mapped to Genome	82.0%
Reads Mapped Confidently to Genome	78.9%
Reads Mapped Confidently to Intergenic Regions	20.0%
Reads Mapped Confidently to Intronic Regions	3.0%
Reads Mapped Confidently to Exonic Regions	55.9%
Reads Mapped Confidently to Transcriptome	56.4%
Reads Mapped Antisense to Gene	2.5%

Testing PASA with Annot + SR + LR:

Results similar to StringTie2, shows improved but unreliable counts upon visual inspection

PASA is computationally intensive, complex to install and run, and delivers mediocre results despite varied input data

Similar conclusions for Funannotate and BRAKER.

Egzotek : a nextflow pipeline to automate annotation

We are developing a pipeline, Egzotek[9], to automate the annotation process from FASTQ to GTF as described in our protocol[10].

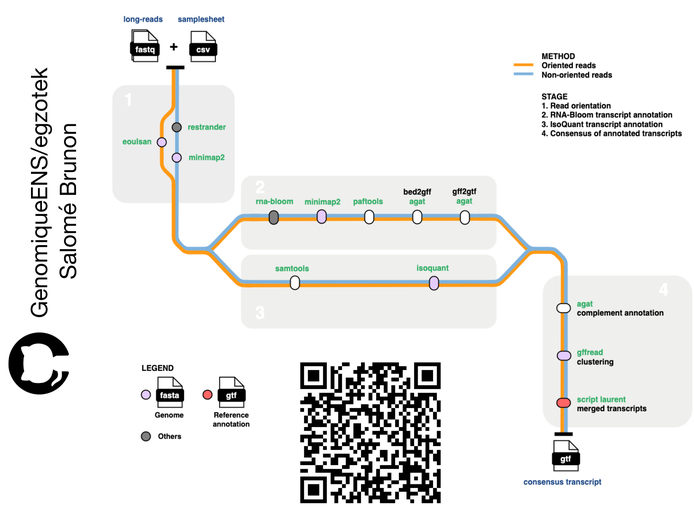
This pipeline:

- Integrates IsoQuant and RNA-Bloom
- Incorporates experimental metadata upstream
- Includes data transformation steps to produce a well-formatted annotation file

We are working on the final consensus step to:

- Reduce redundant transcripts
- Incorporate existing reference annotations, if available
- Group transcripts under meaningfully named genes for user-friendly annotation

GenomiqueENS/egzotek
Salomé Brunon



1. Read orientation
   i. Oriented protocol (eoulsan)
   ii. Non-oriented protocol (restrander)
2. Transcript annotation with RNA-Bloom
   i. Transcript annotation (rna-bloom) (with optional short-read polishing)
   ii. Genome mapping (minimap2)
   iii. Bam to bed file conversion (minimap2-paftools)
   iv. Bed to gff file conversion (agat)
   v. Gff to gtf file conversion (agat)
3. Transcript annotation with Isoquant
   i. Genome mapping (minimap2)
   ii. Sam to bam file conversion (samtools)
   iii. Transcript annotation (isoquant)
4. Complement annotation (agat)
5. Clusterisation (gffread)
6. Merge annotation

References

[1] Shumate A, Wong B, Pertea G, Pertea M Improved transcriptome assembly using a hybrid of long and short reads with StringTie, PLOS Computational Biology 18, 6 (2022)
[2] https://doi.org/10.5281/zenodo.1134477
[3] https://github.com/PASApipeline
4] Bruna, T., Gabriel, L. & Hoff, K. J. (2024). Navigating Eukaryotic Genome Annotation Pipelines: A Route Map to BRAKER, Galba, and TSEBRA. arXiv
[5] https://github.com/nf-core/nanoseq
[6] Pardo-Palacios, F.J., Wang, D., Reese, F. *et al.* Systematic assessment of long-read RNA-seq methods for transcript identification and quantification. *Nat Methods* **21**, 1349–1363 (2024)
[7] Prjibelski, A.D., Mikheenko, A., Joglekar, A. *et al.* Accurate isoform discovery with IsoQuant using long reads. *Nat Biotechnol* **41**, 915–918 (2023)
[8] Ka Ming Nip, Saber Hafezqorani, Kristina K. Gagalova, Readman Chiu, Chen Yang, René L. Warren, and Inanc Birol. Reference-free assembly of long-read transcriptome sequencing data with RNA-Bloom2. Nature Communications. 2023 May 22;14(1):2940
[9] https://github.com/GenomiqueENS/egzotek
[10] dx.doi.org/10.17504/protocols.io.36wgqd5qyvk5/v1