BITS :: Call for Abstracts 2022 - Oral communication

Type	Oral communication
Session	Algorithms for Bioinformatics
Title	IRescue: single cell uncertainty-aware quantification of transposable elements expression
All Authors	Polimeni B(1,2), Marasca F(1,3), Ranzani V(1), Bodega B(1,4).

Affiliation

(1) INGM, Istituto Nazionale di Genetica Molecolare 'Romeo ed Enrica Invernizzi', Milan, Italy.
(2) Ph.D. Program in Translational and Molecular Medicine, DIMET, University of Milan-Bicocca, Monza, Italy.
(3) Department of Clinical Sciences and Community Health, University of Milan, Milan, Italy.
(4) Department of Biosciences, University of Milan, Milan, Italy.

Motivation

Transposable elements (TEs) are mobile DNA sequences that contribute to the evolution of species, genome plasticity and transcription regulation, and their transcription in human cells have been associated with both detrimental effects and physiological functions.
Due to their extremely repetitive and interspersed nature, the transcriptome analysis of TEs is notoriously challenging. Taking precautions during library design choices regarding read length and pairing, and using specialized tools, can improve both the read mappability and expression estimate of TEs.
While a plethora of software is available for bulk RNA-Seq, few attempts have been made for measuring the expression of TEs in scRNA-seq data so far. In this context, the most common available library types are droplet-based, usually characterized by short single-end reads, poorly mappable on interspersed repeats.
Here, we present IRescue (Interspersed Repeats single-cell quantifier), a software for the error-correction, deduplication and quantification of scRNA-seq Unique Molecule Identifiers (UMIs) mapping on TEs. IRescue is currently the only software that, in case of UMIs mapping multiple times on different TE features (e.g. TE subfamilies), takes into account all mapped features to estimate the correct one, rather than excluding multi-mapping UMIs or picking one randomly.

Methods

IRescue implements a novel algorithm based on building UMI-TE equivalence classes to solve ambiguous mappings. Briefly, scRNA-seq reads aligned on a reference genome using a splicing-aware aligner, are mapped to the corresponding TE. For each cell, UMIs mapping on the same set of TEs are grouped into equivalence classes (ECs) to solve ambiguous mappings and duplicated UMIs. UMI deduplication is achieved by storing UMI sequences in an undirected graph, connecting UMIs diverging for up to one mismatch. The deduplicated UMI count is inferred by calculating the number of unique neighborhoods in the graph, and assigned to the TE feature showing the highest number of alignment events in the EC. Then, the final TE counts of the cell is calculated by the running sum of the TE counts obtained from every EC of the cell, and stored in a TE x Cell sparse matrix, compatible with most toolkits for scRNA-seq downstream analysis.
To test IRescue's performance, data simulations were done using a 3'-end PBMC dataset as a template to reproduce the positioning bias of droplet-based scRNA-seq. Simulated TE counts were obtained by mapping UMIs on a reference, while simulated scRNA-seq reads were obtained by adding a 0.5% mismatch error rate to the aligned reference sequence to emulate Illumina sequencing errors.

Results

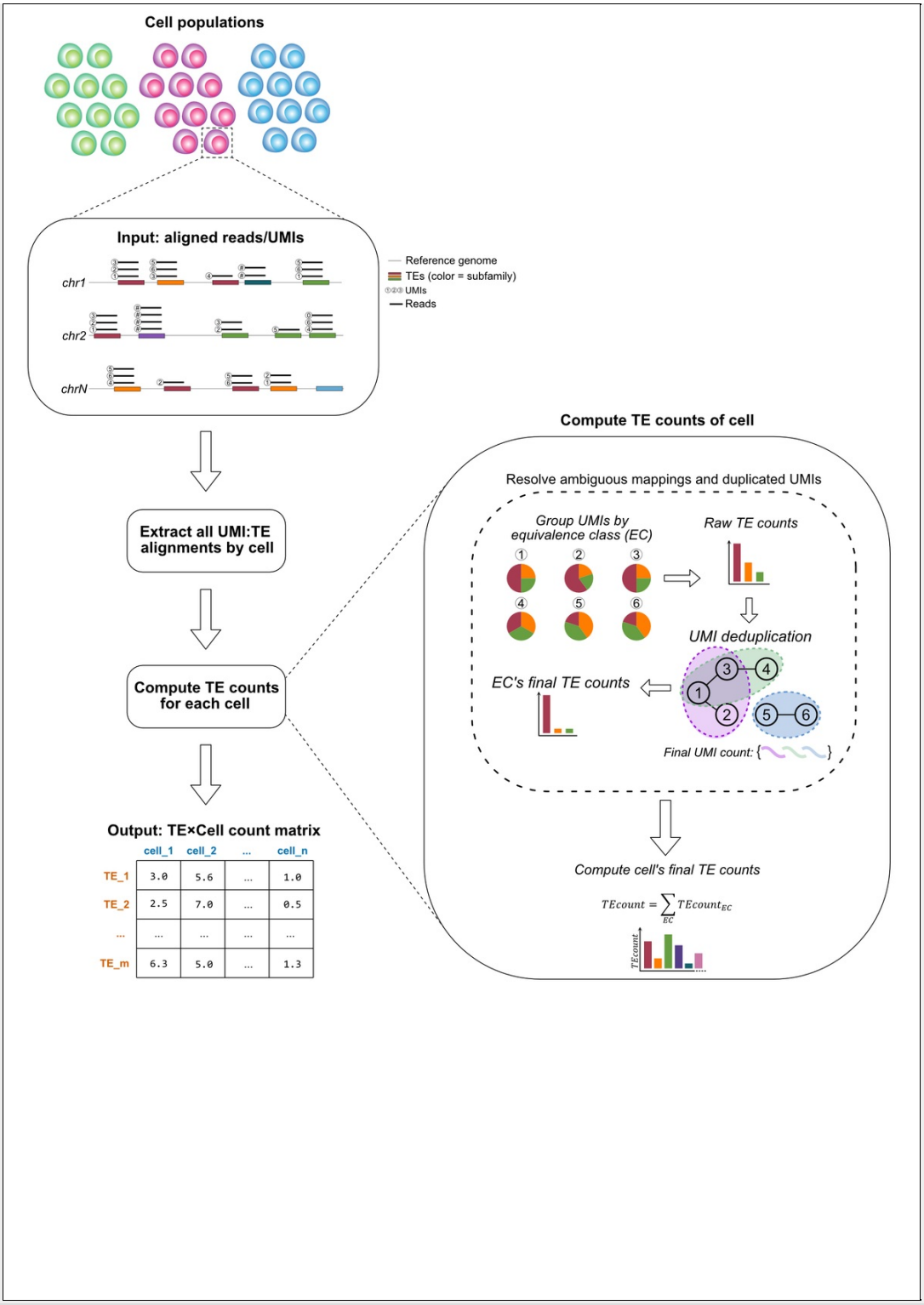
We demonstrate the precision of IRescue with simulated data, identify TE expression signatures associated with colorectal cancer (CRC) and show the heterogeneity and expression dynamics of CRC TE markers previously characterized only at bulk-level .
IRescue counts showed a high correlation score (R=0.81) with simulations, despite about 40% of TE-associated UMIs mapping on different TE subfamilies. We tested the ability to infer the correct cell identity by applying a cell clustering algorithm on both measured and simulated counts, showing that IRescue correctly predicts the identity of most cells, and performs better than other published software.
As a case study for IRescue, we analyzed the TE expression dynamics in CRC, a condition in which some TE subfamilies have been shown to be overexpressed in bulk RNA-seq. We discovered the TE expression signature in CRC and found that it is enriched in evolutionarily young TEs, compared to its normal counterpart. Furthermore, we show that known CRC TE markers are heterogeneously expressed in different CRC cell clusters, and, by analyzing the spliced read alignments, that the expression of such markers can be explained by the transcription of tumor-specific alternative isoforms of human oncogenes.

Info

The source code of IRescue will be freely available on Github soon, upon manuscript submission. As of today, the source code is available upon request.

filename	fig1.png

Figure



Availability	-

Corresponding Author

Name, Surname	Benedetto, Polimeni
Email	polimeni@ingm.org

Submitted on	28.04.2022