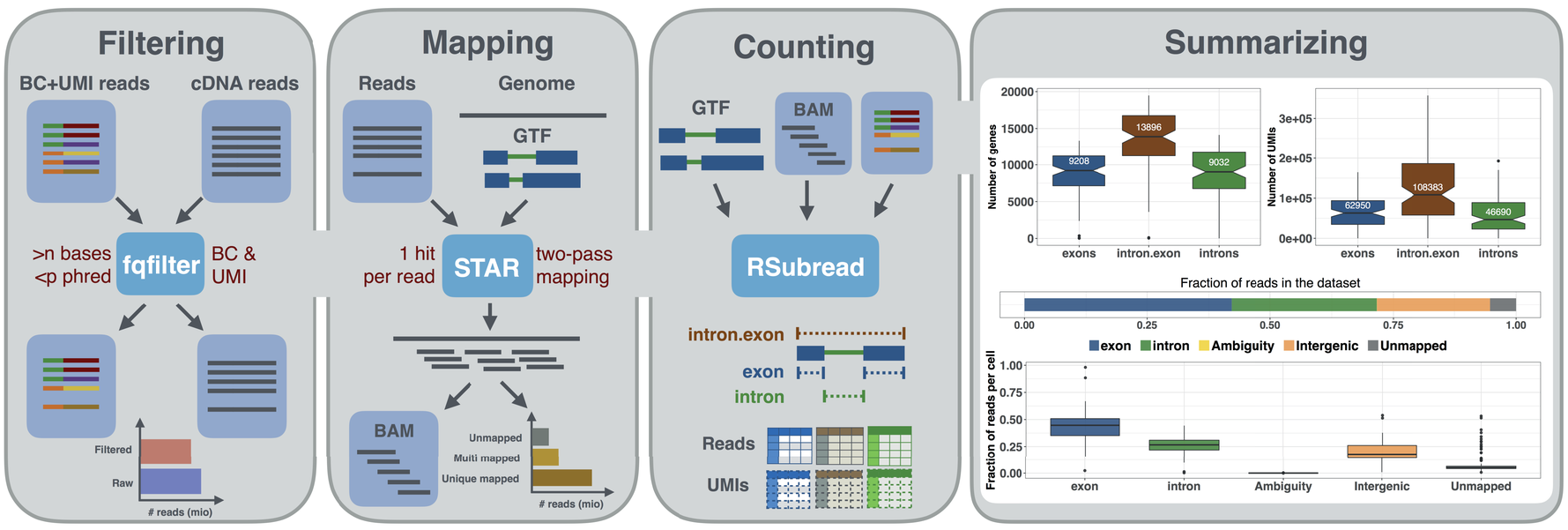
zUMIs: A fast and flexible pipeline to process RNA sequencing data with UMIs

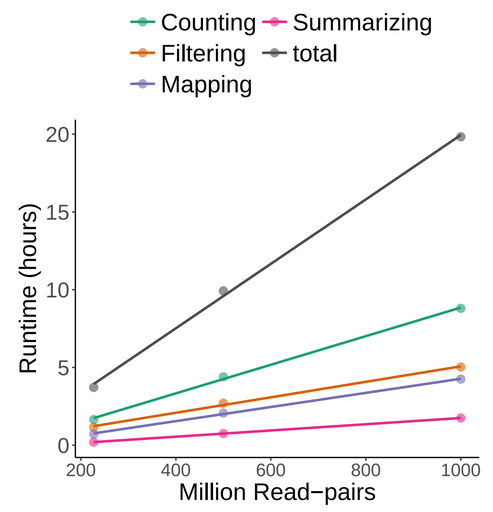
Swati Parekh, Christoph Ziegenhain, Beate Vieth, Wolfgang Enard & Ines Hellmann

Human Genomics & Anthropology, LMU, Germany
parekh@bio.lmu.de

Schematic of the zUMIs pipeline

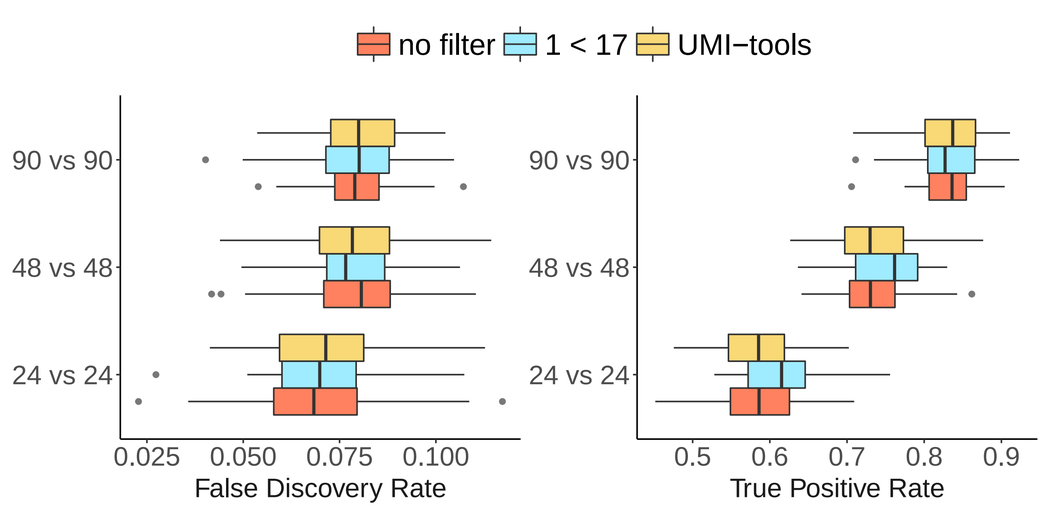


zUMIs is fast



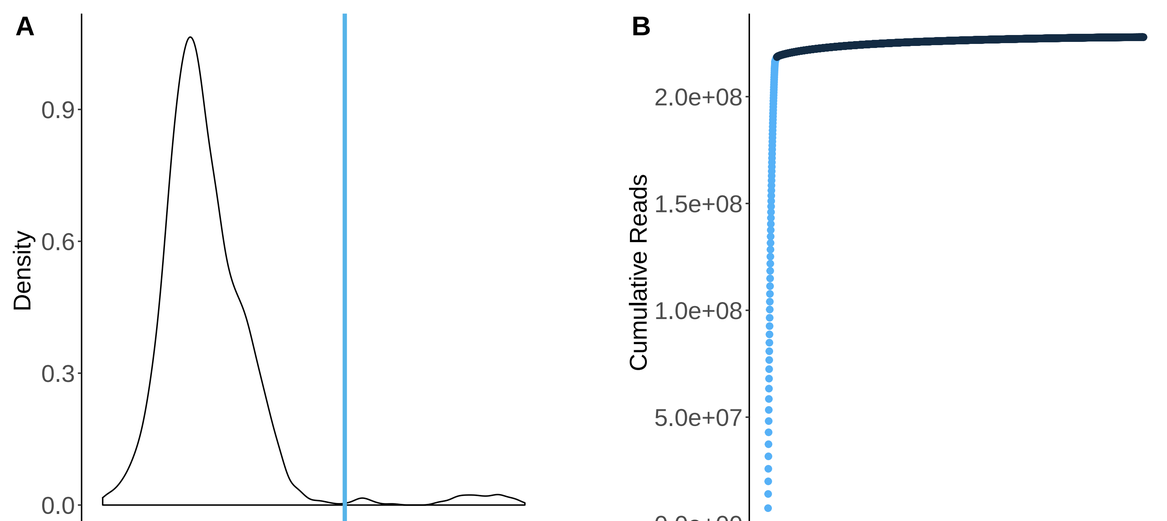
- Runtime at each step in zUMIs[1] with 227, 500 and 1000 million reads.

Impact of UMI quality filtering on DGE



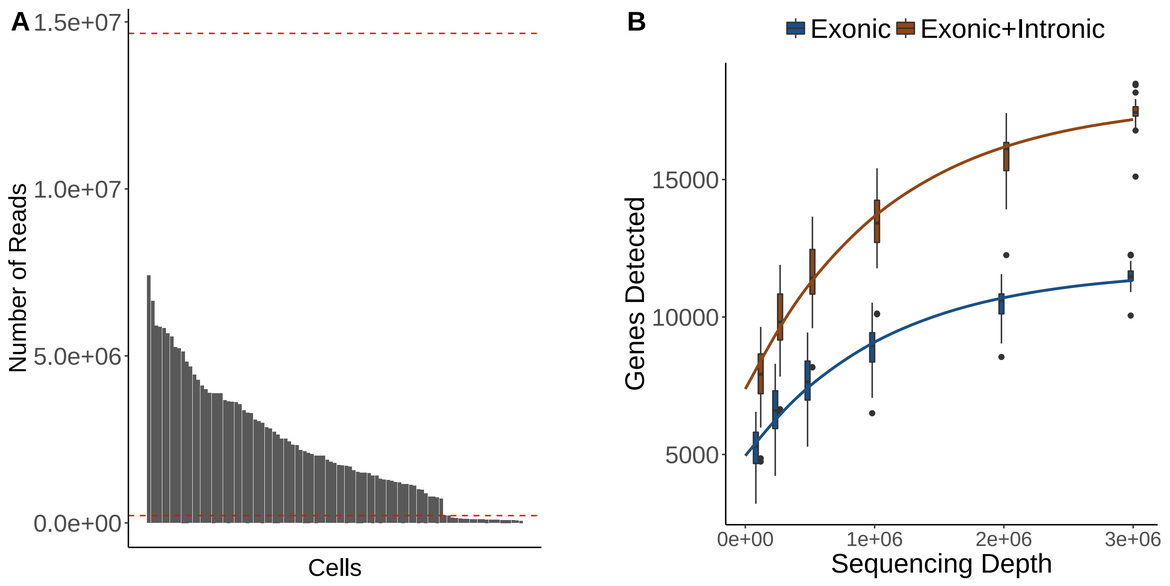
- TPR and FDR to detect differential gene expression (DGE) for each filtering criterion using powsimR [4].
- Distance filtering has no significant impact on DGE for higher sample size and decent sequencing quality.

Automatic barcode detection in zUMIs



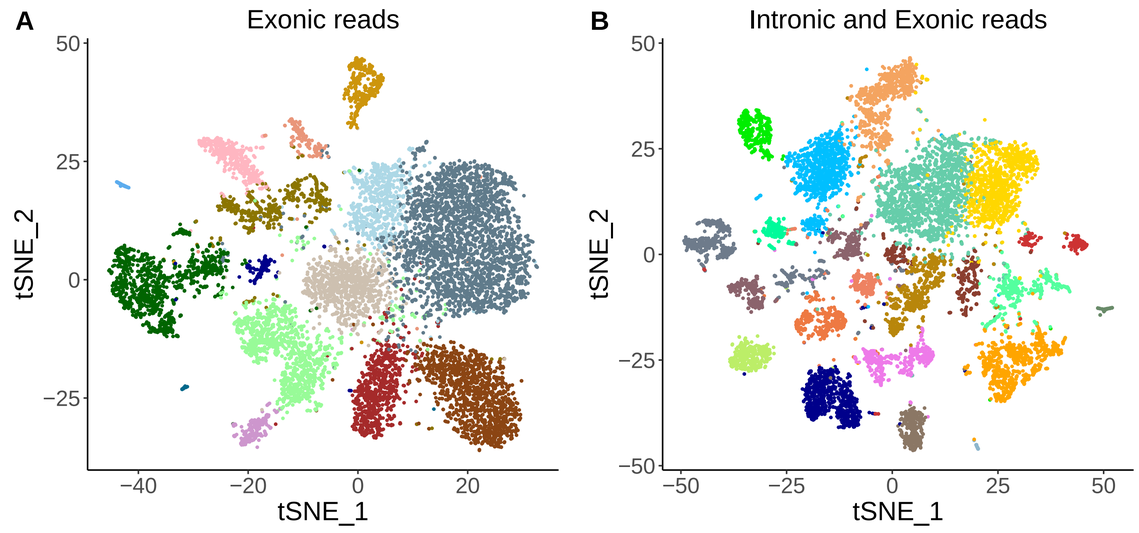
- zUMIs infers which barcodes mark intact cells from the data.
- successfully identified reasonable number of cells from 10x genomics and DroNc-seq data [2]

Adaptive downsampling in zUMIs



- zUMIs downsamples all selected barcodes to be within three absolute deviations from the median number of reads per barcode.
- Various downsampling ranges allow users to evaluate if the library is sequenced to saturation.

Contribution of intron reads in scRNA-seq



- t-SNE plots colored by identified clusters using only exonic reads and inclusive of intronic reads on droplet-based single nuclei RNA-seq (DroNc-seq) data [2].
- Gene level quantification inclusive of intronic reads identifies significantly more genes and improves the cluster resolution compared to only exonic reads.

Key features of zUMIs

- zUMIs can process raw reads to count matrix reported with exon, intron and intron+exon reads with one command.
- zUMIs flexibility allows to accommodate data generated with most of the major scRNA-seq protocols that use BCs and UMIs.
- Model based clustering to select intact cells with the most number of reads makes zUMIs compatible with droplet-based methods.
- Adaptive downsampling function in zUMIs facilitates dealing with hugely varying library sizes.
- zUMIs can count intronic reads which improves the quantification of nascent mRNAs required in various applications [2, 3].

zUMIs

`https://github.com/sdparekh/zUMIs`
Contact: parekh@bio.lmu.de

References

[1] Parekh S, Ziegenhain C, Vieth B, Enard W, Hellmann I. zUMIs: A fast and flexible pipeline to process RNA sequencing data with UMIs. bioRxiv. 2017. p. 153940. doi:10.1101/153940
[2] Habib N, Avraham-Davidi I, Basu A, Burks T, Shekhar K, Hofree M, et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. Nat Methods. 2017;14: 955–958.
[3] La Manno G, Soldatov R, Hochgerner H, Zeisel A, Petukhov V, Kastriti M, et al. RNA velocity in single cells. bioRxiv. 2017. p. 206052. doi:10.1101/206052
[4] Vieth B, Ziegenhain C, Parekh S, Enard W, Hellmann I. powsimR: Power analysis for bulk and single cell RNA-seq experiments. Bioinformatics. 2017; doi:10.1093/bioinformatics/btx435