Isocall: a method for population-scale isoform discovery from long-read RNA sequencing data

Jocelyne Bruand[1], Megan Schertzer[2], Jon Belyeu[1], Tom Mokveld[1], Elizabeth Tseng[1], Zev Kronenberg[1], Khi Pin Chua[1], Ryan Gossart[3], Margaret R. Starostik[4], Jonas Gustafson[5], Rebecca Martin[6], Joy Goffena[5,6], Katherine M. Munson[5], Kaitlyn Sun[5], Evan E. Eichler[5], Stephen B. Montgomery[8], Michael Schatz[4], Rob Patro[9], Michael Love[10], Winston Timp[4], Rajiv McCoy[4], Danny E. Miller[5,7], Michael A Eberle[1], Gloria Sheynkman[2], Yerbol Kurmangaliyev[3], Egor Dolzhenko[1]

[1] Pacific Biosciences, Menlo Park, CA, USA; [2] University of Virginia, Charlottesville, VA; [3] Brandeis University, Waltham, MA; [4] Johns Hopkins University, Baltimore, MD; [5] University of Washington, Seattle, WA; [6] Seattle Children's Research Institute, Seattle, WA; [7] Seattle Children's Hospital, Seattle, WA; [8] Stanford University, Stanford, CA; [9] University of Maryland, College Park, MD; [10] University of North Carolina at Chapel Hill, Chapel Hill, NC

Introduction

Long-read RNA sequencing enables direct observation of full-length RNA molecules, revealing transcript diversity not captured by short-read sequence data.
We present isocall, an efficient multi-sample isoform detection tool, together with workflows to enable population-scale full-length RNA analysis.

Kinnex full-length RNA

By concatenating transcripts into larger fragment libraries, Kinnex increases throughput for full-length RNA transcript sequencing, enabling large-scale studies at high-resolution. New tools and workflows improve and simplify data analysis.

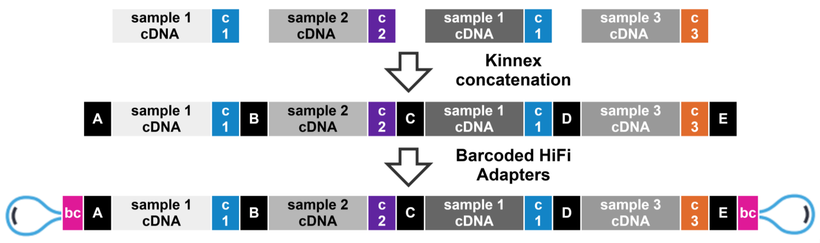


Figure 1. Overview of Kinnex Full-Length RNA. After cDNA generation, Kinnex adapters are attached to the transcripts via PCR. These are then concatenated into a long molecule during array formation. The resulting library can then be cleaned up and sequenced. The library is barcoded at two levels: the original barcode incorporated on the transcript and the HiFi adapter barcode incorporated after concatenation.

Data

We performed isoform joint calling, classification and filtering across two datasets:

- A subset of the 1000 genomes project to assess run time and memory use on large-scale data.
- RNA expression of differentiating stem cell (WTC11) derived primordial endothelial cells from Wissel et. al[1] to look at isocall performance on SIRV data and investigate novel isoforms across time and replicates.

Isocall run time and memory use

Isocall completed isoform joint calling on 250 mapped transcriptomes (~10M S-reads per sample) in ~3 hours using 64 threads and 4 GBs of RAM.

Run time and memory use scale linearly.

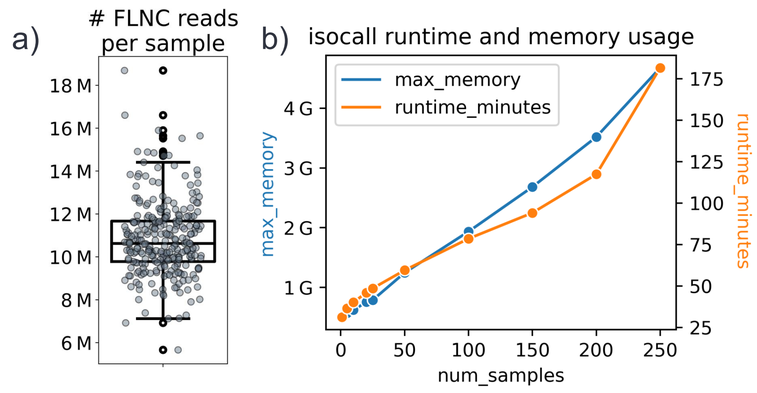


Figure 2. Isocall run time and memory usage on large-scale data. a) samples had ~10M transcript reads each. b) run time and memory on 64 threads, up to 250 samples.

Isocall Performance on SIRV Benchmark

We assessed performance of isocall using the SIRV spike-ins in the iPSC data. Users can control isocall's sensitivity by changing a single parameter which determines the minimum fraction of reads that must be explained by called isoforms at each locus. A higher fraction will yield to a higher sensitivity with a tradeoff on specificity.

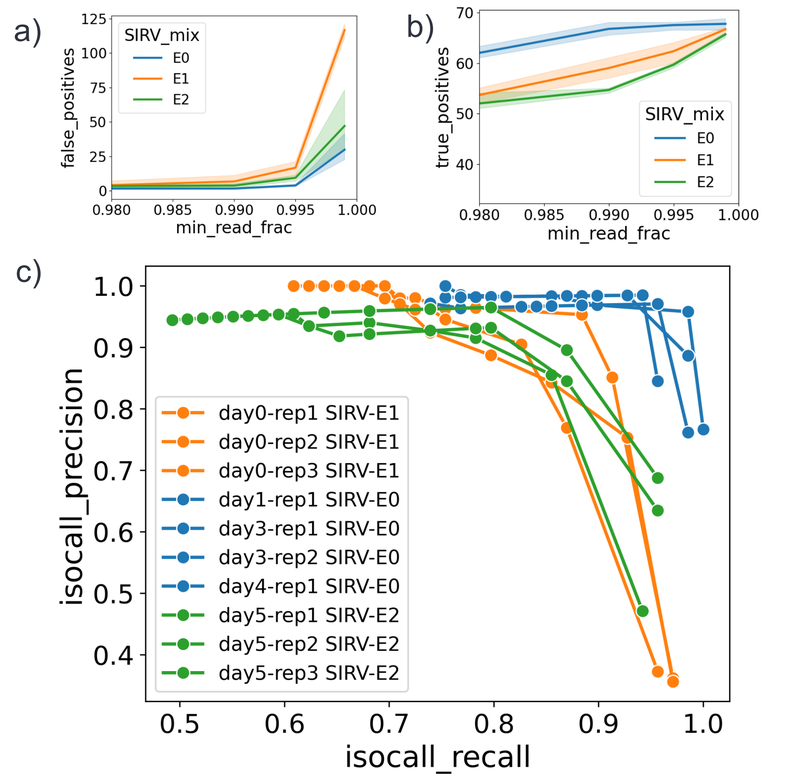


Figure 5. Sensitivity and specificity of isocall on SIRV spike-ins. a) b) increasing the minimum fraction of reads increases both true positives and false positives. c) precision and recall for each sample. Different SIRV mixes have different curves due to the relative abundances of the spike in transcripts.

Kinnex RNA pre-processing workflow generates transcripts bams for biosamples

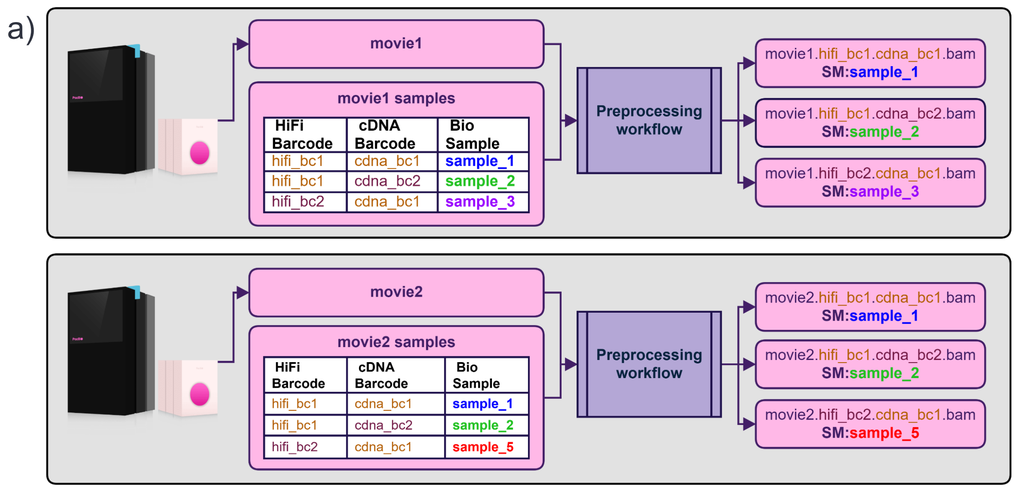
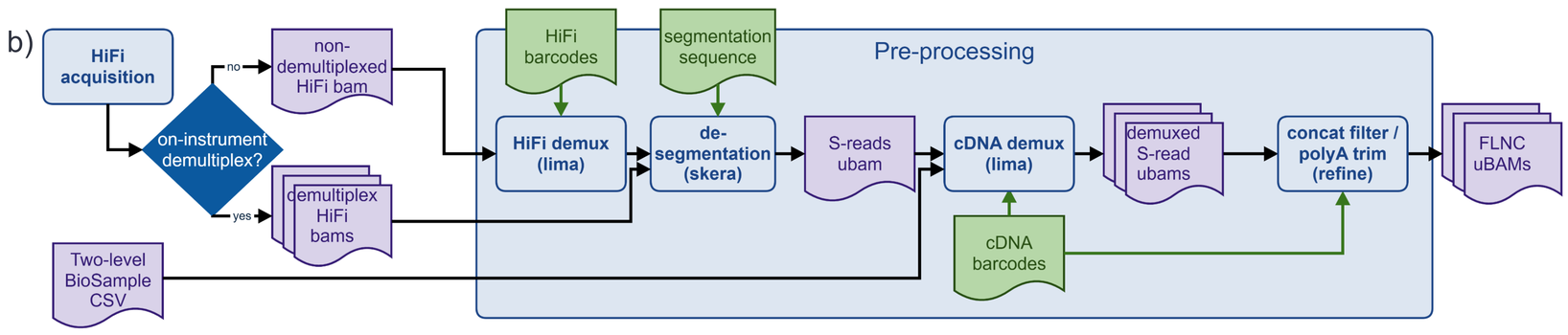


Figure 3: Kinnex RNA preprocessing workflow. The pre-processing workflow allows users to seamlessly obtain transcript BAM files from PacBio long-read RNA sequencing data. a) Users define a two-level biosample CSV specifying both the HiFi and cDNA barcodes for each sample. The workflow generates sample annotated BAM files. b) The preprocessing workflow starts from instrument BAM files. It desegments, demultiplexes, filters concatemers, and trims the poly-A tail to generate per-sample transcript BAM files.



Kinnex RNA secondary analysis workflow generates results for each biosample across acquisitions

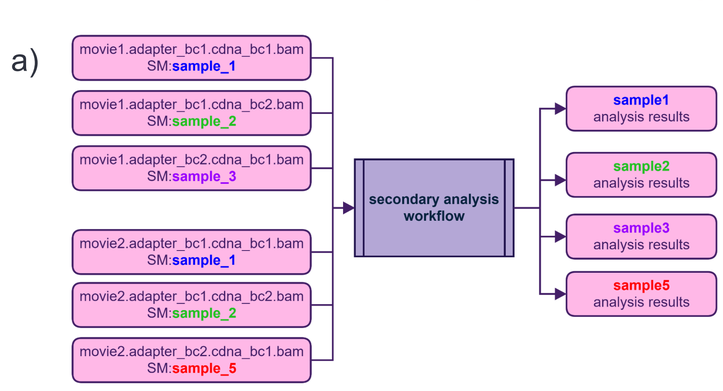
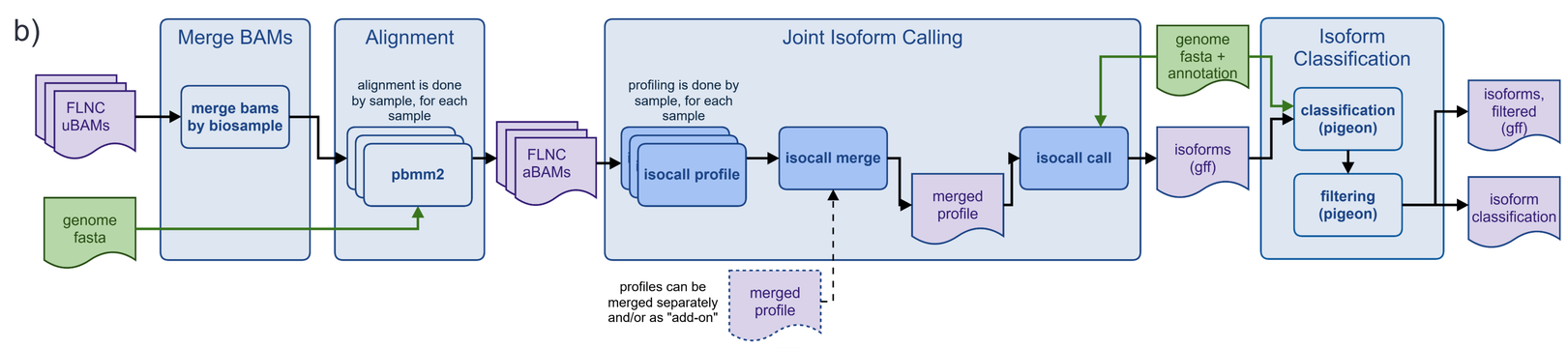


Figure 4. Kinnex RNA Secondary Analysis Workflow. The secondary analysis workflow generates results per biosample regardless of acquisition. a) The secondary analysis workflow takes biosample-annotated transcripts BAMs from any number of acquisitions to generate biosample-level analysis results. In the provided example, sample_1 and sample_2 are sequenced across two acquisitions while sample_3 and sample_5 are only on a single acquisition. b) The secondary analysis workflow first merges the BAM files based on their biosample annotation. Transcript reads are mapped to the reference genome. Joint isoform calling and isoform classification are then performed at the sample level.



Novel isoform on HEY2-AS

When looking at reads that support intron junctions, normalized to TPMs, we found novel isoforms with divergent abundance over time. One example is shown here.

a)

	day0-rep1	day0-rep2	day0-rep3	day1-rep1	day3-rep1	day3-rep2	day4-rep1	day5-rep1	day5-rep2	day5-rep3
ENST00000606334.7	0.172	0.026	0.057	0.195	0.000	0.000	0.000	0.000	0.000	0.000
ENST00000834219.1	0.287	0.128	0.113	0.358	0.036	0.193	0.054	0.000	0.027	0.062
ENST00000834227.1	0.057	0.205	0.000	0.130	0.000	0.000	0.000	0.000	0.000	0.000
ENST00000834229.1	0.057	0.308	0.113	0.000	0.000	0.000	0.000	0.000	0.000	0.000
ENST00000834217.1	0.172	0.026	0.085	0.065	0.000	0.000	0.000	0.000	0.000	0.000
ENST00000834228.1	0.230	0.257	0.085	0.130	0.000	0.032	0.000	0.070	0.000	0.000
ENSG00000237742.9.novel1	30.681	35.810	49.170	21.413	0.929	0.770	0.163	0.035	0.000	0.031
ENSG00000237742.9.novel2	6.199	7.111	9.726	3.873	0.179	0.160	0.054	0.000	0.000	0.124

b)

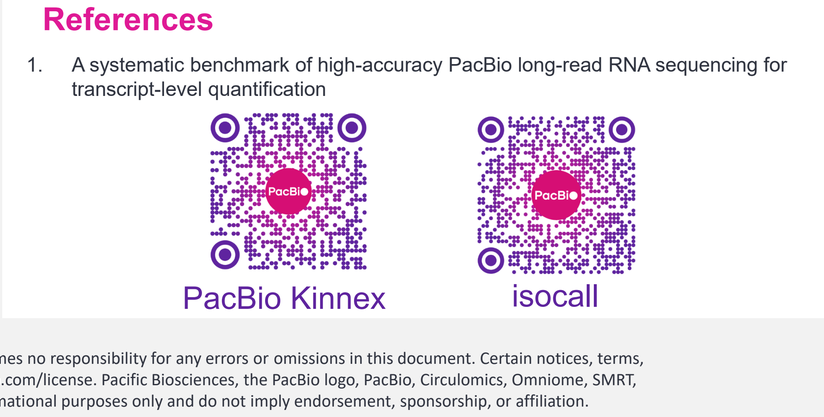


Figure 6. Novel isoform on HEY2-AS. a) CPMs for all known isoforms in gencode v48 on HEY2-AS and two most abundant novel isoforms. b) IGV screenshot of the novel isoforms and supporting reads. Highly expressed exons are not present in the reference annotation transcripts; many reads support the novel intron chains.

Conclusions

- Isocall enables population-scale joint isoform calling by analyzing hundreds of samples.
- Isocall allows for the discovery of novel isoforms.
- RNA workflows streamline analysis

References

1. A systematic benchmark of high-accuracy PacBio long-read RNA sequencing for transcript-level quantification

PacBio Kinnex

isocall