



VALT Summer School

Long-reads Transcriptomics

Juan Francisco Cervilla Martínez

Earlham Institute, Norwich, UK

Section 3

Single-cell analysis pipeline

Overview of commercial pipelines available

PacBio



Iso-Seq

Scalable
De Novo
Isoform Discovery
from PacBio HiFi Reads

<https://isoseq.how>



 [epi2me-labs / wf-single-cell](https://github.com/epi2me-labs/wf-single-cell) Public

<https://github.com/epi2me-labs/wf-single-cell>

PacBio

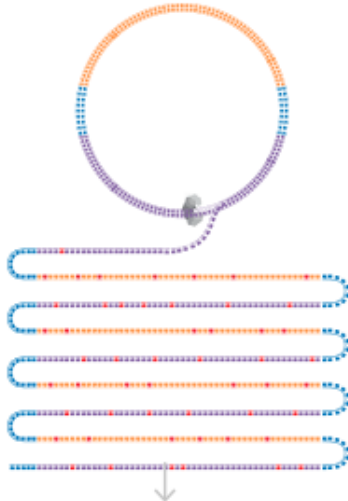


Oxford
NANOPORE
Technologies



CCS <https://ccs.how/>

Generate Highly Accurate
Single-Molecule
Consensus
Reads



HiFi read
(99.9% accuracy)

Adapted from <https://ccs.how/>



SKERA

The PacBio
Concatenated
Read Splitter

<https://skera.how/>

Seq 1 **Seq 2** **Seq 3** **Seq 4**

Adapted from <https://skera.how/read-segments.html>

PacBio
PacBio only

PacBio

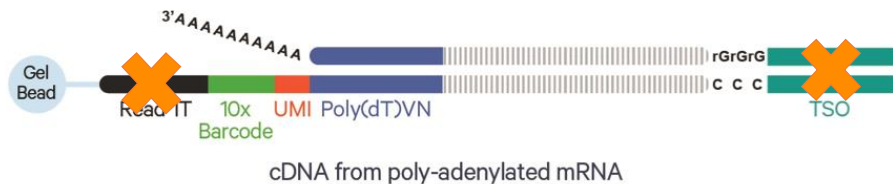




lima

The PacBio
Barcode Demultiplexer &
Primer Remover

<https://lima.how>

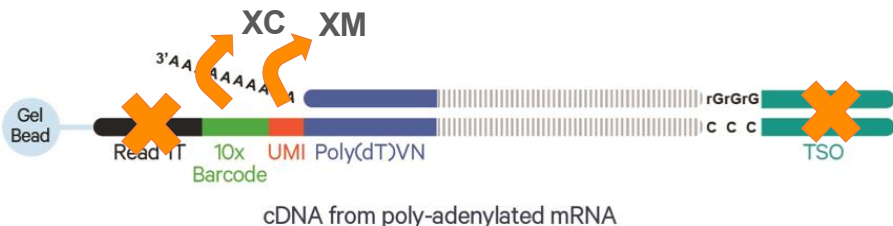


Adapted from https://cdn.10xgenomics.com/image/upload/v1710230393/support-documents/CG000731_ChromiumGEM-X_SingleCell3_ReagentKits_v4_UserGuide_RevA.pdf



Iso-Seq

Scalable
De Novo
Isoform Discovery
from PacBio HiFi Reads



PacBio



Oxford NANOPORE Technologies



torognes / vsearch Public

<https://github.com/toroqnes/vsearch>

Identification of adaptors

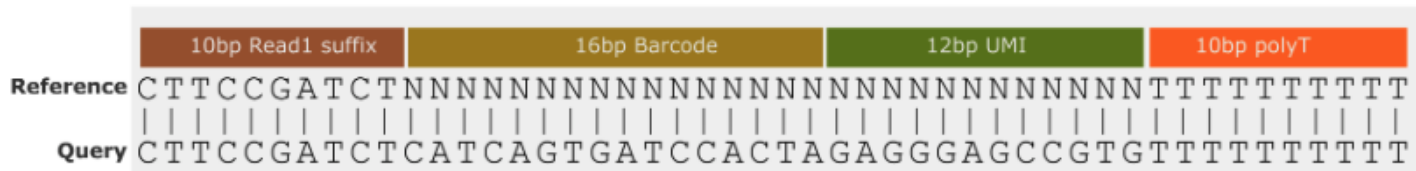


cDNA from poly-adenylated mRNA

jeffdaily / parasail Public

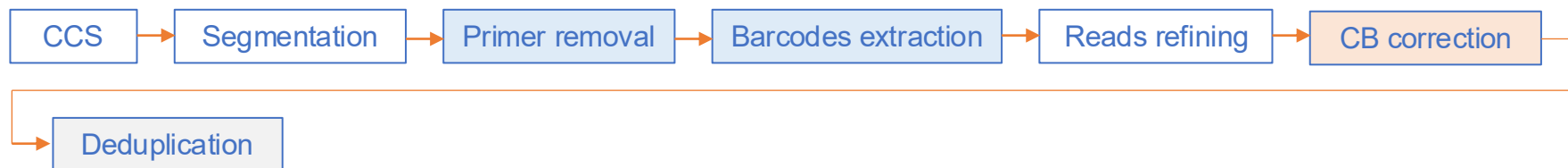
<https://github.com/ieffdaily/parasail>

Probes to extract CBs and UMIs

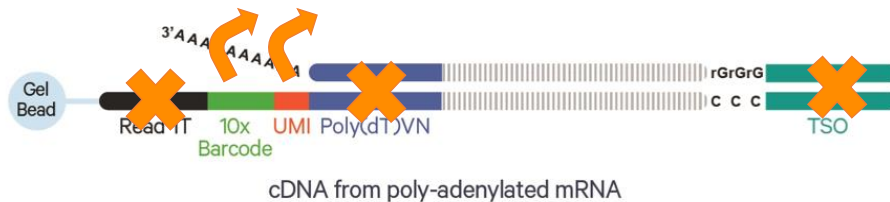


Adapted from <https://github.com/epi2me-labs/wf-single-cell>

PacBio



- Reads refinement – polyA trimming



Iso-Seq

Scalable
De Novo
Isoform Discovery
from PacBio HiFi Reads

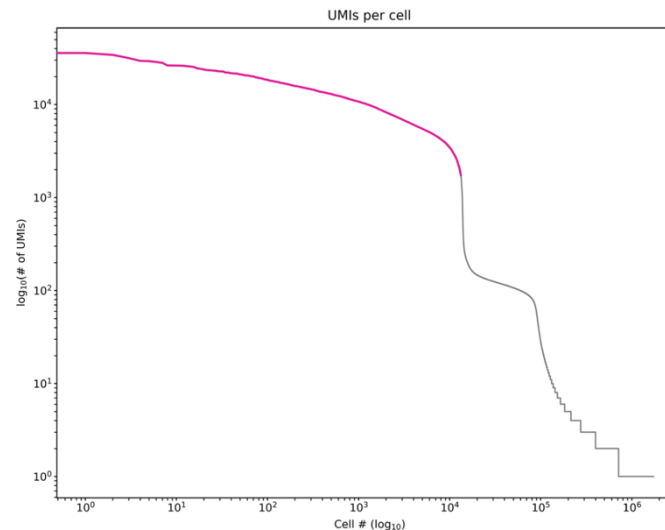
- Reads refinement – polyA trimming
- **Cell barcode correction**

- Cell barcodes are known (10x whitelist)
- Cell barcodes are corrected using Hamming distance ≤ 2 (2 substitutions maximum)
- Low quality cells, empty droplets or ambient RNA are also filtered



Iso-Seq

Scalable
De Novo
Isoform Discovery
from PacBio HiFi Reads



- Reads refinement – polyA trimming
- Cell barcode correction
- **Deduplication – removing PCR artifacts using UMIs**



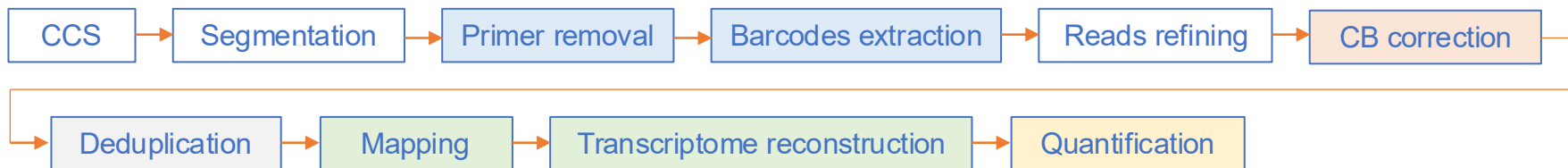
Adapted from <https://isoseq.how/umi/high-level-workflow.html>



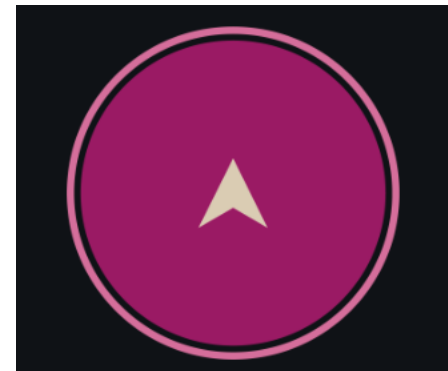
Iso-Seq

Scalable
De Novo
Isoform Discovery
from PacBio HiFi Reads

PacBio



- Reads refinement – polyA trimming
- Cell barcode correction
- Deduplication – removing PCR artifacts with UMIs
- **Mapping to genome reference**



Pbmm2

- PacBio's C++ wrapper of minimap2

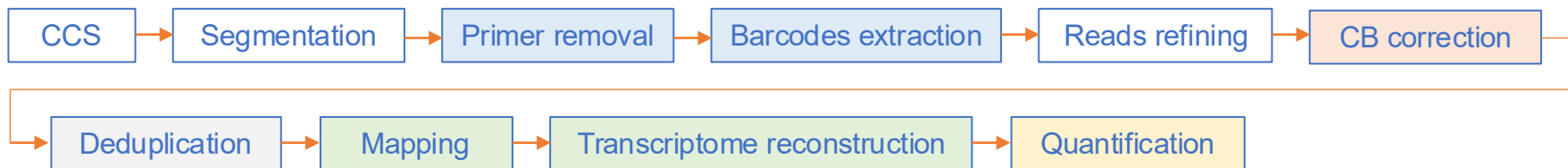
- Reads refinement – polyA trimming
- Cell barcode correction
- Deduplication – removing PCR artifacts with UMIs
- Mapping to genome reference
- **Transcriptome reconstruction and curation**



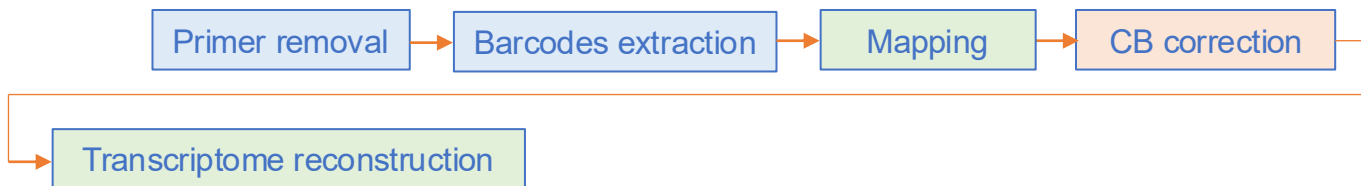
- Reads refinement – polyA trimming
- Cell barcode correction
- Deduplication – removing PCR artifacts with UMIs
- Mapping to genome reference
- Transcriptome reconstruction and curation
- **Expression matrices**



PacBio



Oxford NANOPORE Technologies



- Mapping to genome reference

 lh3 / minimap2 Public

<https://github.com/lh3/minimap2>

- Mapping to genome reference
- **Cell barcode correction**
 - Cell barcodes are known (10x whitelist)
 - Cell barcodes from the sample with high quality are shortlisted
 - Cell barcodes are corrected using Levenshtein distance ≤ 2 (2 indels and/or substitutions maximum)
 - Low quality cells, empty droplets or ambient RNA are also filtered

- Mapping to genome reference
- Cell barcode correction
- **Transcriptome reconstruction**



<https://github.com/gpertea/stringtie>

- Stringtie2
- Using long read mode (-L)
- At least 2 transcripts to call a transcript



<https://github.com/gpertea/gffcompare>

- Resulting transcriptome is annotated with the reference annotation

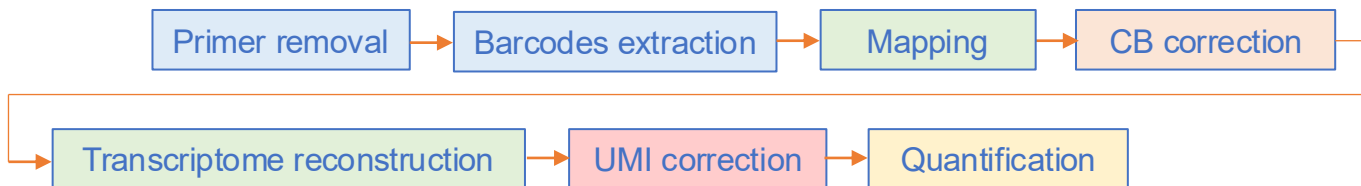
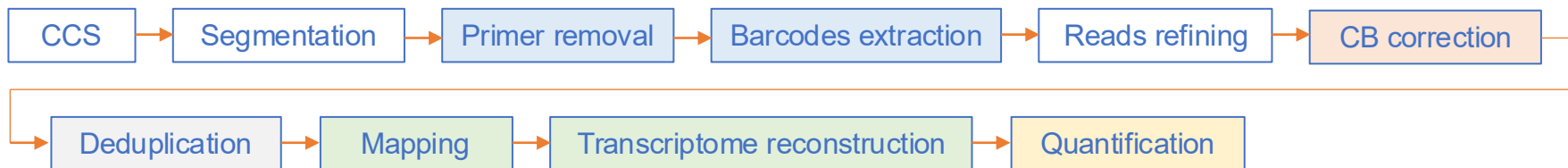
- Mapping to genome reference
- Cell barcode correction
- Transcriptome reconstruction
- Reads are mapped to the resulting transcriptome
- Transcript assignment similar to FLAMES criteria



<https://github.com/LuyiTian/FLAMES>

- Intronic transcripts set to unknown
- Uniquely mapped reads are assigned to transcripts
- Ambiguously mapped reads are assigned or not depending on alignment score and coverage

PacBio



- Mapping to genome reference

- Cell barcode correction

- Transcriptome reconstruction

- **UMI correction**

- UMI barcodes are NOT known (random sequences)

- To reduce searching space, reads are clustered to assigned genes (or genomic interval if not assigned)

UMI-tools

<https://github.com/CGATOxford/UMI-tools>

- UMIs are corrected using Levenshtein distance ≤ 2 (2 indels and/or substitutions maximum)

- Mapping to genome reference
- Cell barcode correction
- Transcriptome reconstruction
- UMI correction
- **Expression matrices**
- Corrected UMIs are collapsed and feature counts are summed
- Expression matrices are further processed
 - Cell with counts below 200 genes/transcripts (default)
 - Genes expressed in fewer than 3 cells (default)
 - Cells with more than 20% mitochondrial expression (default)
- Counts are normalized to 10.000 reads/cell (default) and log10 transformed

Registered Report | [Open access](#) | Published: 07 June 2024

Systematic assessment of long-read RNA-seq methods for transcript identification and quantification

[Francisco J. Pardo-Palacios](#), [Dingjie Wang](#), [Fairlie Reese](#), [Mark Diekhans](#), [Sílvia Carbonell-Sala](#), [Brian Williams](#), [Jane E. Loveland](#), [Maite De María](#), [Matthew S. Adams](#), [Gabriela Balderrama-Gutierrez](#), [Amit K. Behera](#), [Jose M. Gonzalez Martinez](#), [Toby Hunt](#), [Julien Lagarde](#), [Cindy E. Liang](#), [Haoran Li](#), [Marcus Jerryd Meade](#), [David A. Moraga Amador](#), [Andrey D. Pribelski](#), [Inanc Birol](#), [Hamed Bostan](#), [Ashley M. Brooks](#), [Muhammed Hasan Çelik](#), [Ying Chen](#), ... [Angela N. Brooks](#)  [+ Show authors](#)

[Nature Methods](#) **21**, 1349–1363 (2024) | [Cite this article](#)

50k Accesses | **103** Citations | **111** Altmetric | [Metrics](#)

- **For transcriptome reconstruction**

SQANTI3 

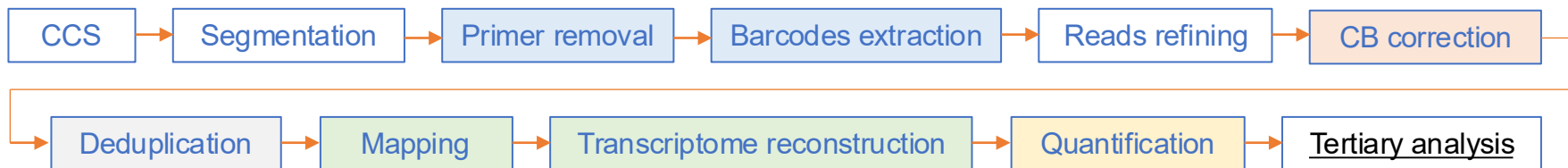
SQANTI 
single-cell

Under development

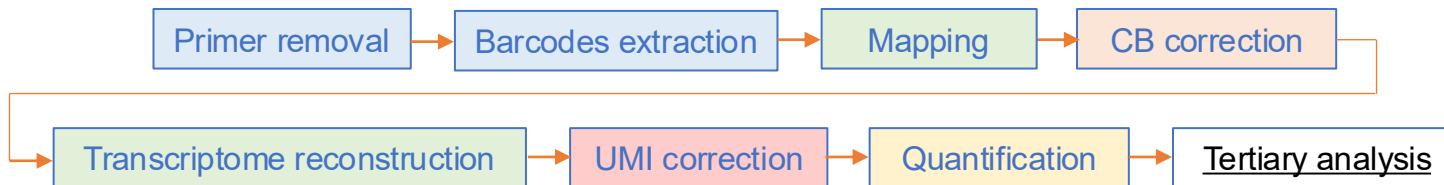
- **For transcriptome curation**

- ***Which one is the unique step of the wf-single-cell pipeline?***

PacBio



Oxford NANOPORE Technologies



Thank You!



For more information about the LongTREC Summer School:

<https://longtrec.eu>