A Brief Overview of Genome Annotation, With a Focus on the Use of Isoseq

Monica Britton
Bioinformatics Core
July 21, 2020

Genome Annotation vs. Functional Annotation

Genome Annotation:

- **Where are the features (genes, exons, UTRs, etc.) located?**

Functional Annotation:

- **What is the function of the RNA (or the protein encoded by the RNA)?**
- **What is its biological relevance? (Gene Ontology, Pathway)**

How Detailed Does a Genome Annotation Need to be Anyway?

Annotation needs to be complete and detailed enough to answer your research questions:

- **Location of genes (exons, transcripts, etc.) … for expression analyses**
- **Upstream/downstream regions … for motif-finding, ChIP-Seq, ATAC-Seq, etc.**
- **And to satisfy your papers’ reviewers.**

Input to Genome Annotation

A high quality genome assembly (maximum scaffold size, minimum redundancy)

Full length transcript sequences (IsoSeq!)

Protein and transcript sequences from same or closely related organism.

Short-read RNA-Seq data

Predictions of low complexity regions

Ab initio gene/feature predictions

Long Read Transcript Sequencing is Ideal

Previously, PE Illumina reads had to be assembled (Trinity), and/or transcripts “reconstructed” (Stringtie)

- Gene families subject to collapse in assembly
- Chimeric transcripts
- Genes across fragmented scaffolds could not be identified.

Now, full length transcripts can be sequenced, with multiple passes per molecule.

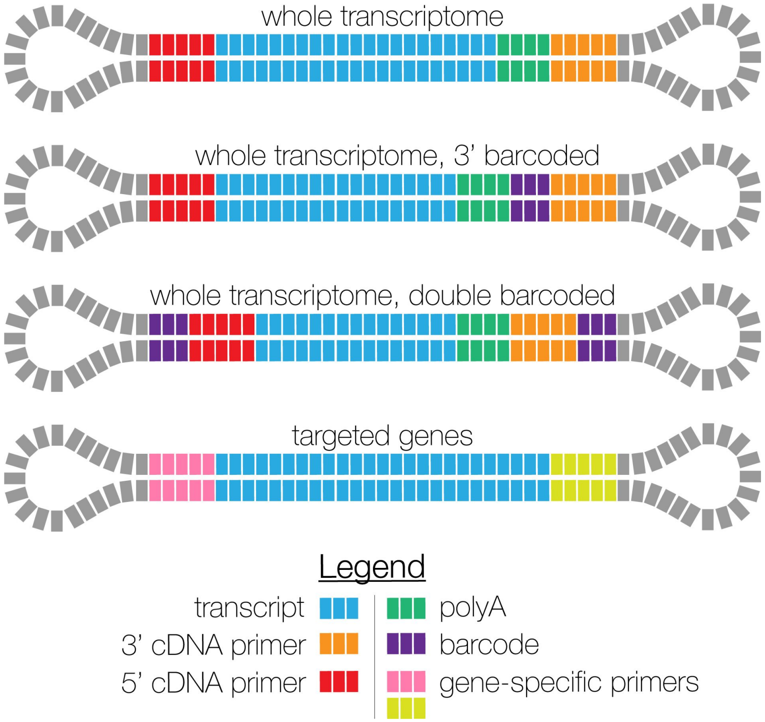
Capturing ALL Possible Transcripts is Impossible (but we can try!)

For full genome annotation, ALL transcripts should be represented.

But, some transcripts are expressed in only one tissue, or at one timepoint.

RNA from multiple tissues, conditions, etc. can be individually barcoded and pooled. This allows for demultiplexing the ccs reads and identification of specific transcripts from each sample.

Options for multiplexing RNA samples within one Isoseq pool



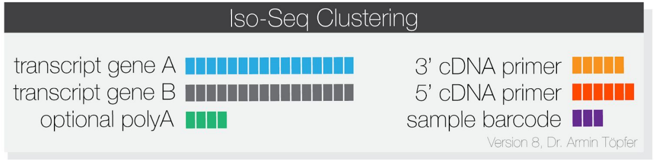
Isoseq3 data processing: from raw bam files to high quality full-length transcripts

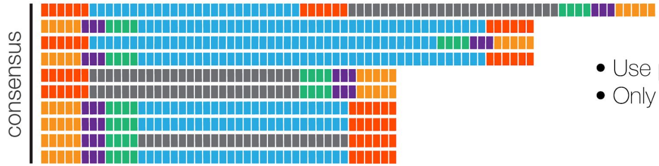
Overview of Multiple Step Process:

- Most steps use bam and xml files as input and output formats
- ccs converts subreads into HiFi circular consensus sequences
- lima demultiplexes barcodes and removes primers
- isoseq3 refine trims polyA tails and remove concatemers
- isoseq3 cluster creates high quality isoforms
- pbmm2 aligns isoforms to genome
- isoseq3 collapse generates gff3 file with exons, genes and transcripts

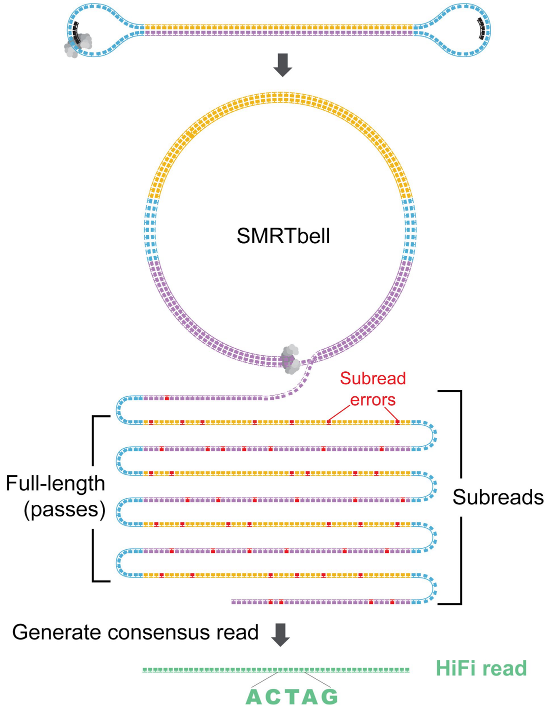
https://github.com/PacificBiosciences/IsoSeq

ccs converts subreads into HiFi circular consensus sequences



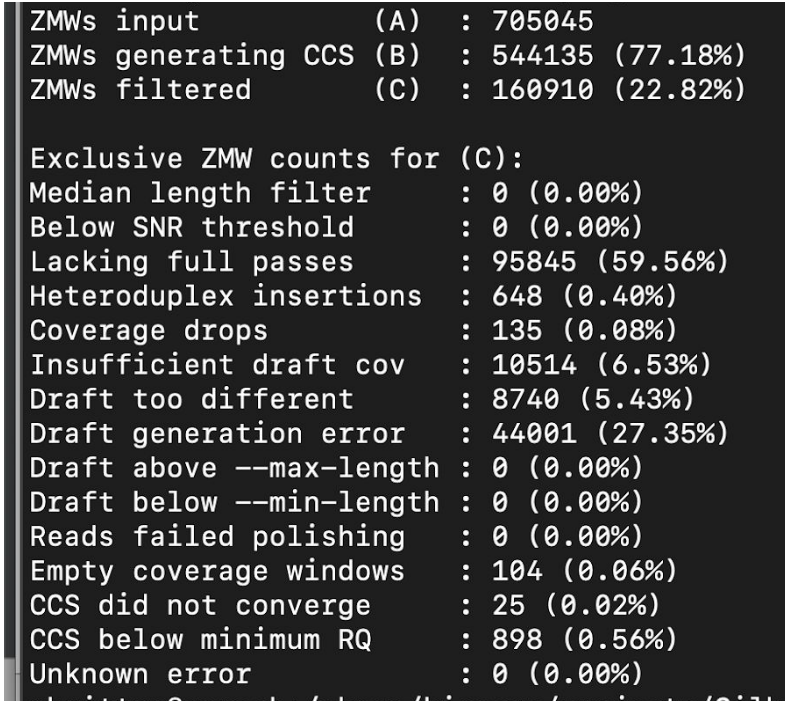


- Use **polished** CCS reads
- Only full-pass ZMWs



https://github.com/PacificBiosciences/ccs

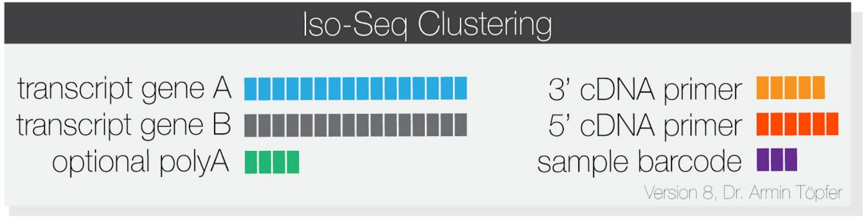
Example ccs report

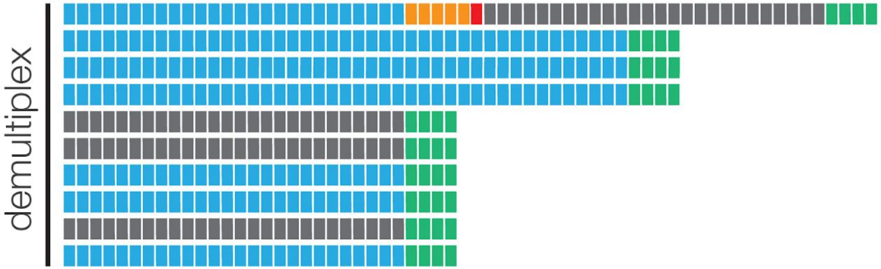


```
ZMWs input          (A)  : 705045
ZMWs generating CCS (B)  : 544135 (77.18%)
ZMWs filtered       (C)  : 160910 (22.82%)

Exclusive ZMW counts for (C):
Median length filter     : 0 (0.00%)
Below SNR threshold      : 0 (0.00%)
Lacking full passes      : 95845 (59.56%)
Heteroduplex insertions  : 648 (0.40%)
Coverage drops           : 135 (0.08%)
Insufficient draft cov   : 10514 (6.53%)
Draft too different      : 8740 (5.43%)
Draft generation error   : 44001 (27.35%)
Draft above --max-length : 0 (0.00%)
Draft below --min-length : 0 (0.00%)
Reads failed polishing   : 0 (0.00%)
Empty coverage windows   : 104 (0.06%)
CCS did not converge     : 25 (0.02%)
CCS below minimum RQ     : 898 (0.56%)
Unknown error            : 0 (0.00%)
```

lima demultiplexes barcodes and removes primers to produce full-length (FL) reads

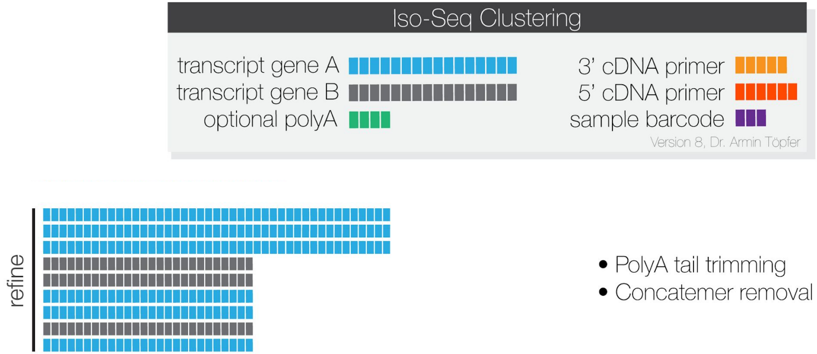




- Barcoded and unbarcoded cDNA primer removal
- Orientation
- Unwanted primer combination removal

https://github.com/PacificBiosciences/barcoding

isoseq3 refine



- PolyA tail trimming
- Concatemer removal

Trims polyA tails and removes concatemers to produce FLNC (full-length non-concatemer) reads

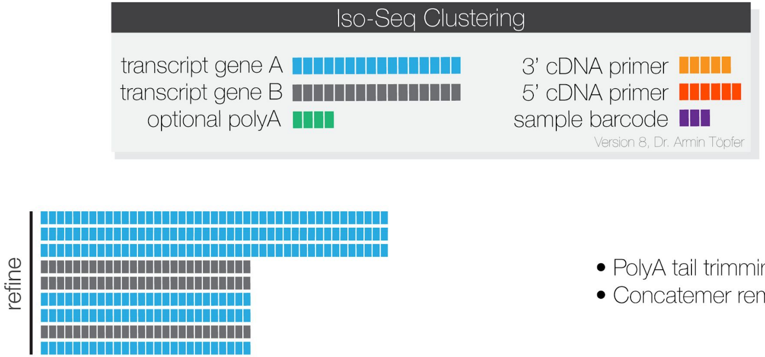
This is also the step where files from multiple SMRT cells would be merged

isoseq3 cluster:
Generation of transcriptome fastas

Output (in addition to bams) are putative isoforms:

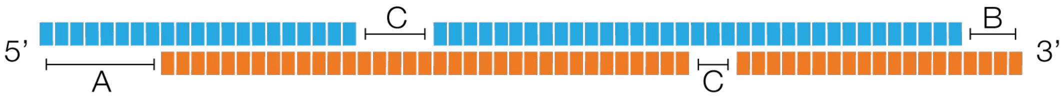
<prefix>.hq.fasta.gz with predicted accuracy ≥ 0.99

<prefix>.lq.fasta.gz with predicted accuracy < 0.99



- PolyA tail trimming
- Concatemer removal

Similar transcripts:



A) <100 bp 5’ overhang
B) <30 bp 3’ overhang
C) <10 bp gaps

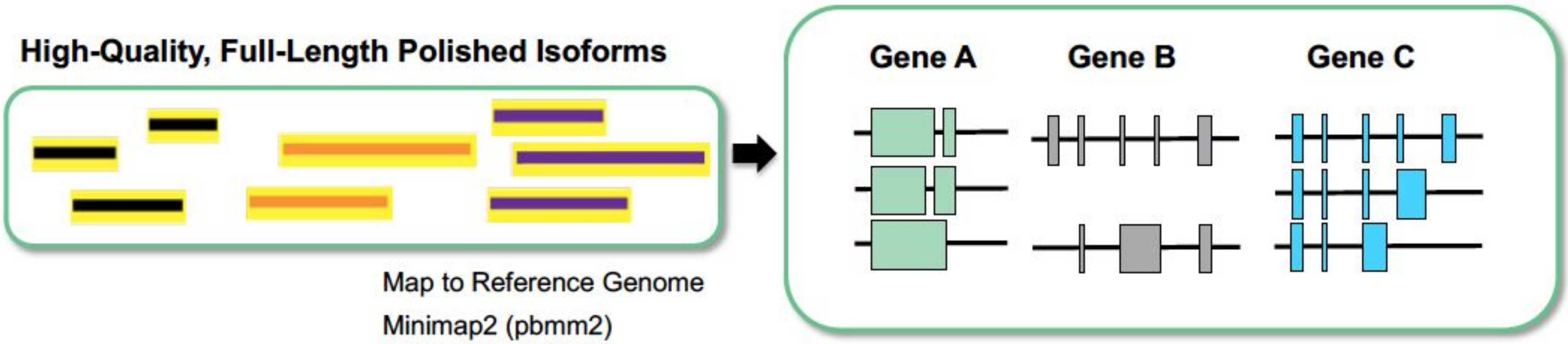
Pbmm2 (PacBio Minimap2)

Align transcriptome to genome with minimap2, using PacBio specific wrapper

Has preset options for alignment of PacBio-generated data.

https://github.com/PacificBiosciences/pbmm2

isoseq3 collapse



Generate gff annotation file from pbmm2 alignments

Some other software with similar functions

StringTie2 (http://ccb.jhu.edu/software/stringtie/)

- Revamped version of StringTie for using long reads
- Can be used with uncorrected long reads (Isoseq subreads, ONT)
- Generates gff/gtf compatible with other JH software

TAMA (https://github.com/GenomeRIK/tama/wiki)

- Transcription Annotation by Modular Algorithms
- Replaces last steps of Isoseq3 pipeline
- Can merge multiple transcriptomes together
- Multiple accessory tools, including ORF-finding, protein prediction

When is Genome Annotation Finished? (Probably Never)

The gff/gtf file generated by the software described here, may be sufficient …

if your goal is to generate “enough” annotation to assign RNA-Seq reads to specific genes.

Then the next step is functional annotation (which will be discussed soon…)

NCBI RefSeq Annotation

You don’t need extensive annotation to submit a genome to NCBI.

NCBI will (eventually) run the Gnomon annotation pipeline and generate a gff3 file.

https://www.ncbi.nlm.nih.gov/genome/annotation_euk/gnomon/

Publication-Quality Annotation

Labor-intensive, requires far more hands-on time than just generating a gff/gtf file.

Some specialized annotation software are:

- **Maker** (https://www.yandell-lab.org/software/maker.html)
- **Braker2** (https://github.com/Gaius-Augustus/BRAKER)
- **PASA** (https://github.com/PASApipeline/PASApipeline/wiki) **and Evidence Modeler** (https://evidencemodeler.github.io/)

Functional IsoTranscriptomics (“FIT”) (https://tappas.org/)

The FIT software suite is written to specifically be a downstream part of Isoseq analyses.

Combines expression and annotation to obtain differences in function

Three useful tools to enhance annotation and to explore differential isoform expression within your datasets.

- **SQANTI3 (https://github.com/ConesaLab/SQANTI3)**
- **IsoAnnotLite (https://isoannot.tappas.org/isoannot-lite/)**
- **tappAS (https://app.tappas.org/)**

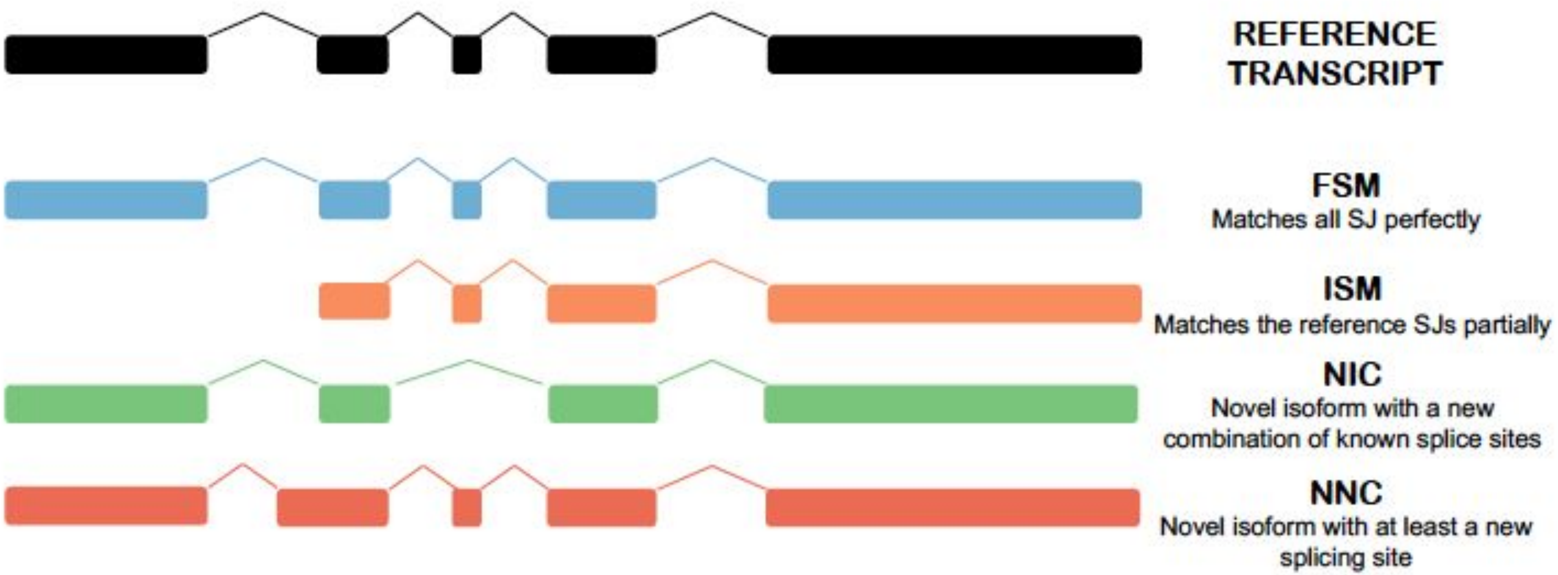
SQANTI3

Structural and Quality Annotation of Novel Transcript Isoforms (https://github.com/ConesaLab/SQANTI3)

Uses genome annotation (gff3) and reads (including short reads) to “QC” isoform predictions, including splicing.

BUT, works best with an existing reference annotation

SQANTI3 -- types of transcripts


REFERENCE TRANSCRIPT
FSM
Matches all SJ perfectly
ISM
Matches the reference SJs partially
NIC
Novel isoform with a new combination of known splice sites
NNC
Novel isoform with at least a new splicing site

SQANTI3

Helps to answer questions including:

- **How similar are the isoforms compared to the reference transcriptome?**
- **Have we found known…**
  - **Isoforms?**
  - **Transcription Starting or Terminating Sites?**
  - **Splice-junctions?**
- **Have we found novel isoforms?**
- **Are there any artifacts due to library preparation or sequencing issues?**
- **Can we use complementary data to accept or discard isoforms?**

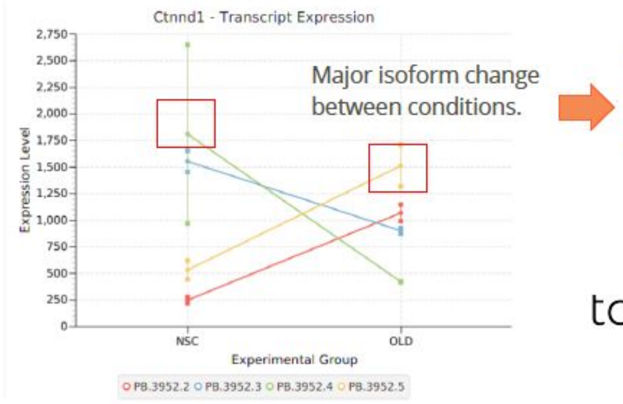
IsoAnnotLite

Functional annotation framework using existing annotations from a relatively small number of reference species to annotate the transcripts in the SQANTI-generated gff3.

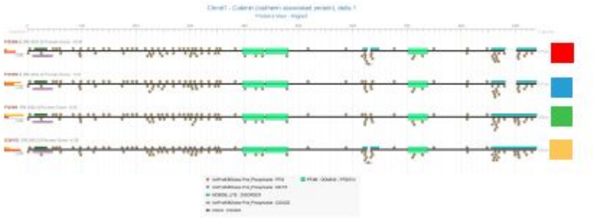
https://isoannot.tappas.org/isoannot-lite/

tappAS

A java GUI application that allows the user to explore and analyze alternative splicing and alternative UTR processing from a FUNCTIONAL perspective.



Major isoform change between conditions.



- Combines expression with functional annotations.
- Provides interactive maps to visualize these functional elements in the sequence.

https://app.tappas.org/

Functional Annotation

Functional annotation is typically performed on a transcriptome and/or protein (amino acid) dataset.

At a minimum, the annotation should include description, orthologs, Gene Ontology terms, Pathway identifiers.

The accuracy and usefulness of functional annotation is only as good as database that is used!

Trinotate (https://github.com/Trinotate/Trinotate.github.io/wiki)

Originally developed to annotate Trinity assemblies, can be used with most transcriptome fastas

Can take a while to run blasts for large datasets, mostly supports swissprot/uniprot

Also includes protein domains and other database searches

Free and has good support

Output can be parsed.

Blast2GO (https://www.blast2go.com/)

GUI-based and Command-Line (CLI) versions available

Marketed as user-friendly “Software for Biologists”

Free version has minimal functionality

Paid version also runs analyses, generates plots, etc.

Most annotation (GO mapping) is part of a proprietary database.

ENTAP (https://entap.readthedocs.io/en/latest/)

Eukaryotic Non-Model Transcriptome Annotation Pipeline

Designed specifically to address fragmentation and assembly issues that result in inflated transcript estimates and poor annotation rates.

Runs relatively fast, since uses Diamond instead of blast

Databases can be customized

Output can be parsed

A Brief Overview of Genome Annotation, With a Focus on the Use of Isoseq

Monica Britton
Bioinformatics Core
July 21, 2020