PACIFIC BIOSCIENCES®

# Using the Iso-Seq Application on SMRT Link and BioConda

Elizabeth Tseng, PacBio

@Magdoll

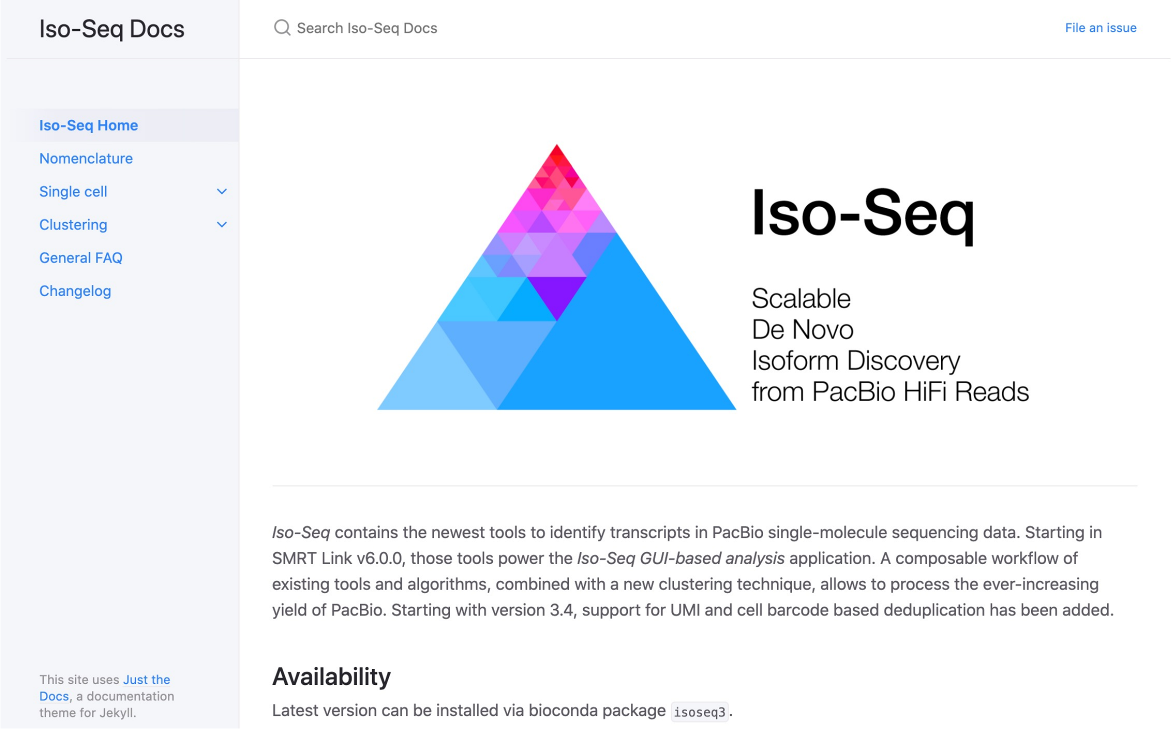
For Research Use Only. Not for use in diagnostic procedures. © Copyright 2020 by Pacific Biosciences of California, Inc. All rights reserved.

# Why use Iso-Seq analysis?

# ISO-SEQ ANALYSIS MAIN FEATURES

- No reference genome required
- No transcriptome assembly required
- Recovers full-length (5’ to 3’) transcripts
- Yields highly accurate (>99%) transcripts

# ISOSEQ.HOW



Iso-Seq Docs

Search Iso-Seq Docs

File an issue

- Iso-Seq Home
- Nomenclature
- Single cell
- Clustering
- General FAQ
- Changelog

This site uses Just the Docs, a documentation theme for Jekyll.

Iso-Seq

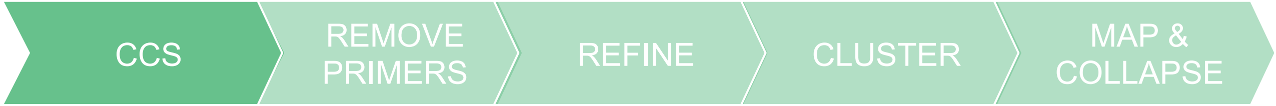
Scalable
De Novo
Isoform Discovery
from PacBio HiFi Reads

*Iso-Seq* contains the newest tools to identify transcripts in PacBio single-molecule sequencing data. Starting in SMRT Link v6.0.0, those tools power the *Iso-Seq GUI-based analysis* application. A composable workflow of existing tools and algorithms, combined with a new clustering technique, allows to process the ever-increasing yield of PacBio. Starting with version 3.4, support for UMI and cell barcode based deduplication has been added.

## Availability

Latest version can be installed via bioconda package `isoseq3`.

# HIFI READS FROM CCS



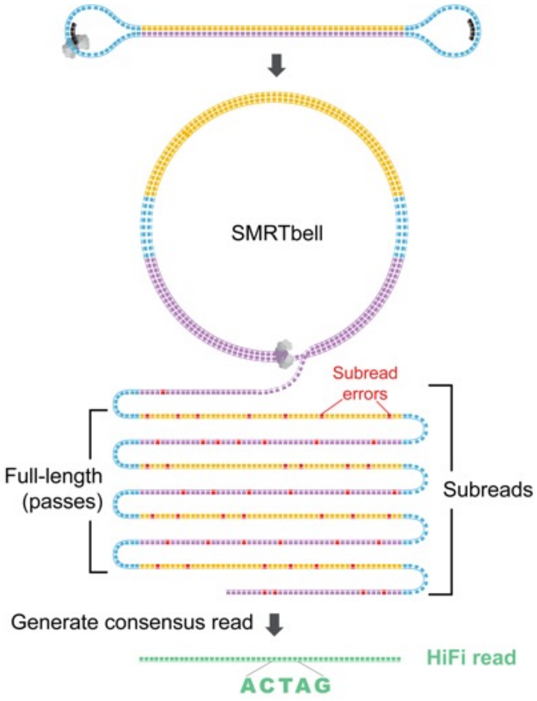
CCS

REMOVE PRIMERS

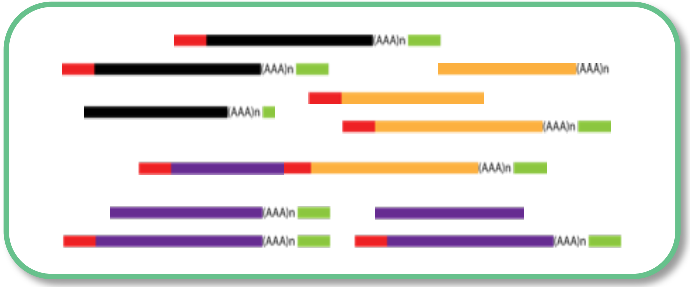
REFINE

CLUSTER

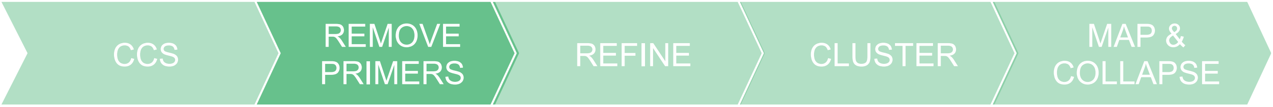
MAP & COLLAPSE

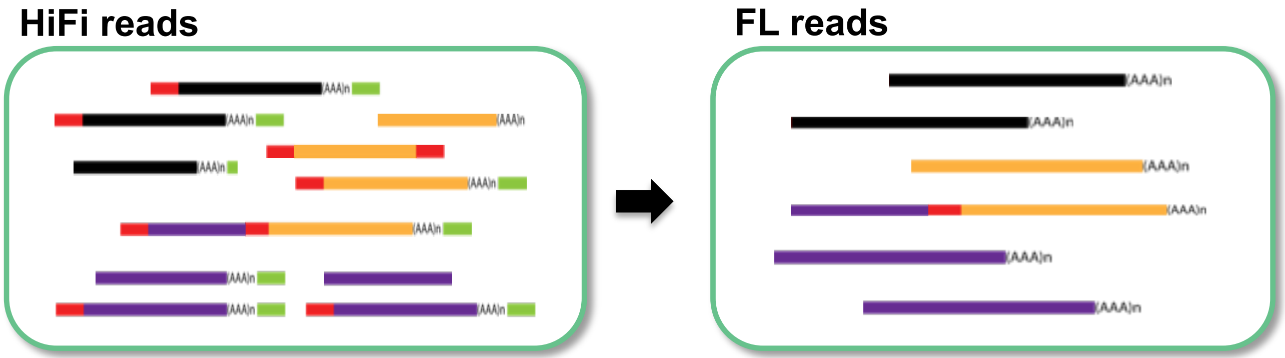


## HiFi reads

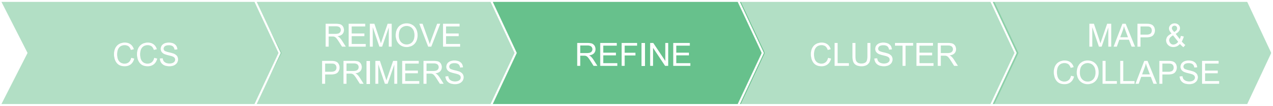


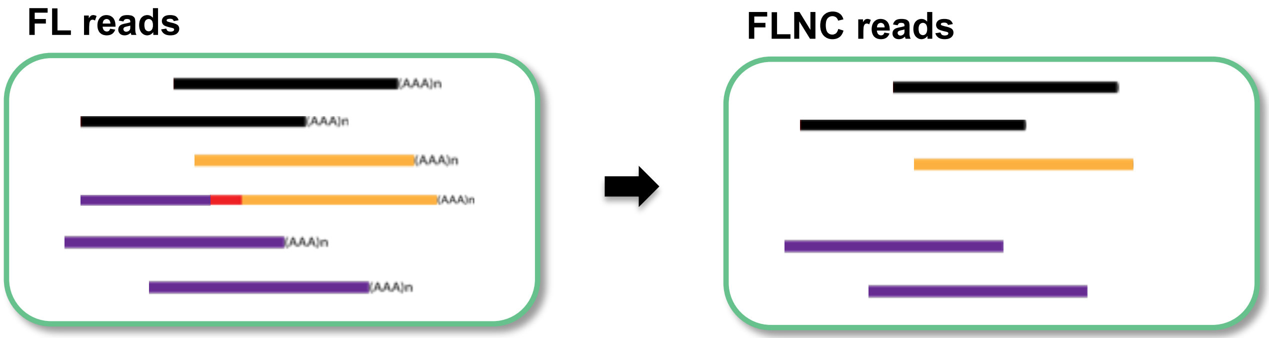
# FULL-LENGTH READS HAVE 5' AND 3' PRIMERS



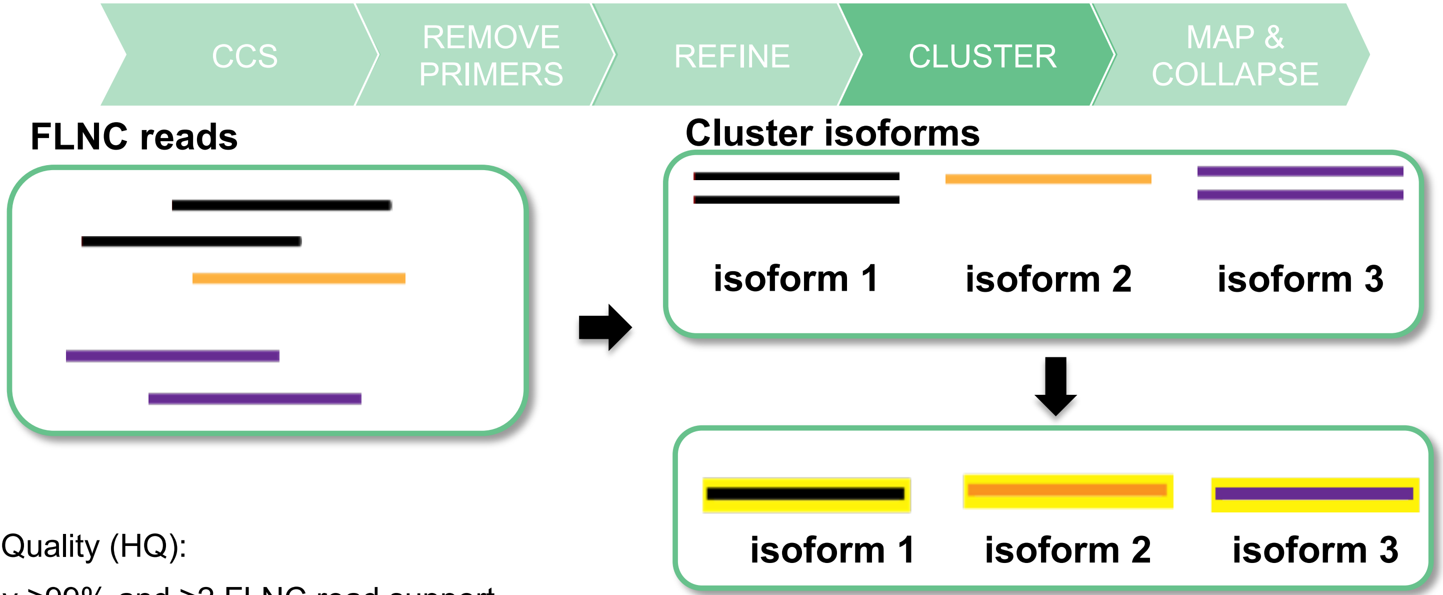


# REMOVE CONCATEMERS AND POLY(A) TAILS





# CLUSTER TO GET ISOFORMS



CCS

REMOVE PRIMERS

REFINE

CLUSTER

MAP & COLLAPSE

FLNC reads

Cluster isoforms

isoform 1 isoform 2 isoform 3

isoform 1 isoform 2 isoform 3

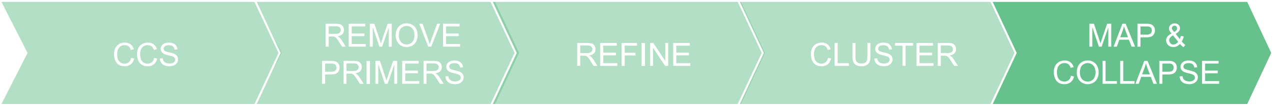
- High Quality (HQ):

accuracy ≥99% and ≥2 FLNC read support

- Low Quality (LQ):

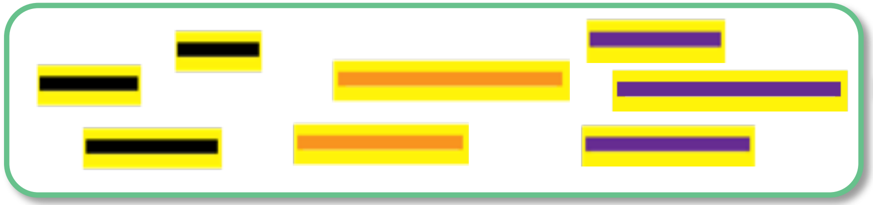
accuracy <99% and ≥2 FLNC read support

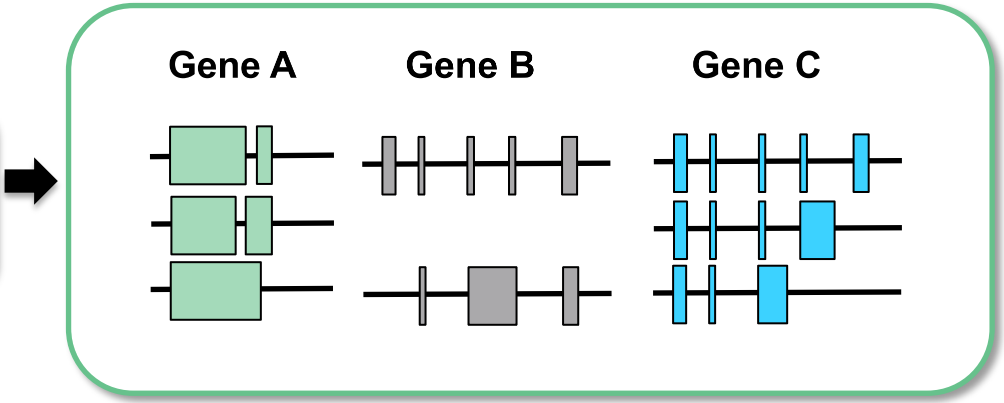
# MAP AND COLLAPSE ISOFORMS



CCS → REMOVE PRIMERS → REFINE → CLUSTER → MAP & COLLAPSE

High-Quality, Full-Length Polished Isoforms





Map to Reference Genome

Minimap2 (pbmm2)

# BENEFITS OF ISO-SEQ ANALYSIS APPLICATION

- High-quality transcripts
- Full-Length Non-concatemer reads
- Mapped & collapsed isoforms
- Removes artifacts
- Removes poly(A) tails

PACBIO®

# Iso-Seq Analysis Using pbBioConda

# INSTRUCTIONS TUTORIAL

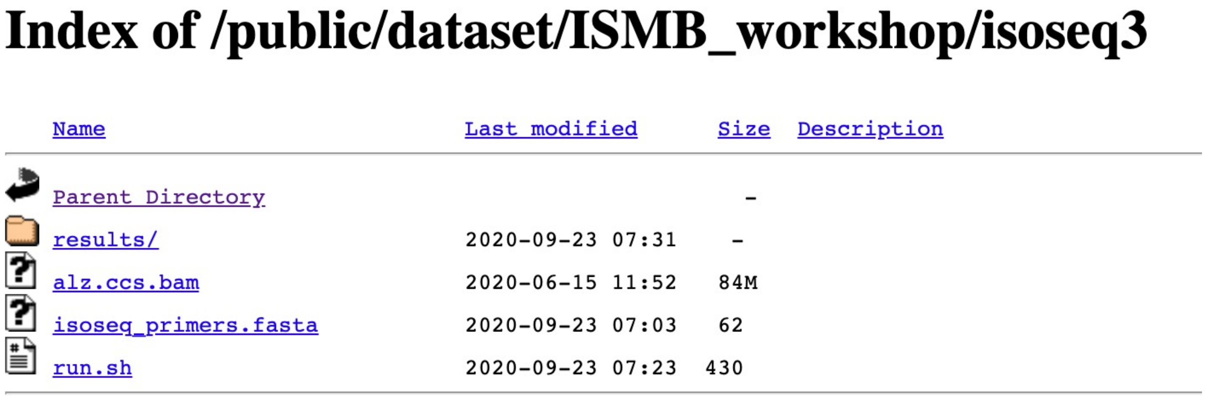
Follow the instructions tutorial for installing all the software needed.

- If you do not have an HPC server to install pbbioconda, you should have already:
  - Create an AWS account
  - Create an AWS Linux Instance to run Iso-Seq 3 Analysis Pipeline
  - Connect to your AWS Instance

- Upgrades and Install Software

# DOWNLOAD THE DATA

https://downloads.pacbcloud.com/public/dataset/ISMB_workshop/



## Index of /public/dataset/ISMB_workshop/isoseq3

| Name | Last modified | Size | Description |
|---|---|---|---|
| Parent Directory | | - | |
| results/ | 2020-09-23 07:31 | - | |
| alz.ccs.bam | 2020-06-15 11:52 | 84M | |
| isoseq_primers.fasta | 2020-09-23 07:03 | 62 | |
| run.sh | 2020-09-23 07:23 | 430 | |

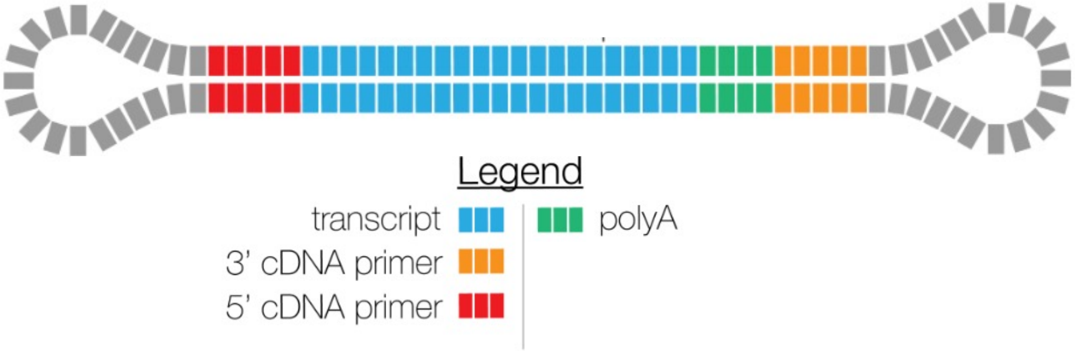
Example:

```
$ wget -nv https://downloads.pacbcloud.com/public/dataset/ISMB_workshop/isoseq3/alz.ccs.bam
```

# SPECIFY ISO-SEQ PRIMERS

```
$ more primers.fasta

>5p
GCAATGAAGTCGCAGGGTTGGG
>3p
GTACTCTGCGTTGATACCACTGCTT
```



# INPUT CCS BAM FILE

```
$ samtools view -h alz.ccs.bam
```

```
m141008_060349_42194_c100704972550000001823137703241586_s1_p0/63/ccs4*0255
**00CCCGGGATCCTCTAGAATGC~~~~~~~~~~~~~~~~~~~~~~RG:Z:83ba013f np:i:35
rq:f:0.999682 sn:B:f,11.3175,6.64119,11.6261,14.5199 zm:i:63
```

# REFERENCE GENOME

```
$ grep '>’ hg38.fa # to list the headers per chromosome
```

```
>chr1  AC:CM000663.2  gi:568336023  LN:248956422  rl:Chromosome  M5:6aef897c3d6ff0c78a
ff06ac189178dd  AS:GRCh38
>chr2  AC:CM000664.2  gi:568336022  LN:242193529  rl:Chromosome  M5:f98db672eb0993dcfd
abafe2a882905c  AS:GRCh38
>chr3  AC:CM000665.2  gi:568336021  LN:198295559  rl:Chromosome  M5:76635a41ea913a405d
ed820447d067b0  AS:GRCh38
>chr4  AC:CM000666.2  gi:568336020  LN:190214555  rl:Chromosome  M5:3210fecf1eb92d5489
da4346b3fddc6e  AS:GRCh38
>chr5  AC:CM000667.2  gi:568336019  LN:181538259  rl:Chromosome  M5:a811b3dc9fe66af729
dc0dddf7fa4f13  AS:GRCh38  hm:47309185-49591369
…
```

# SOFTWARE INSTALLATION CHECK

## Access to your conda environment

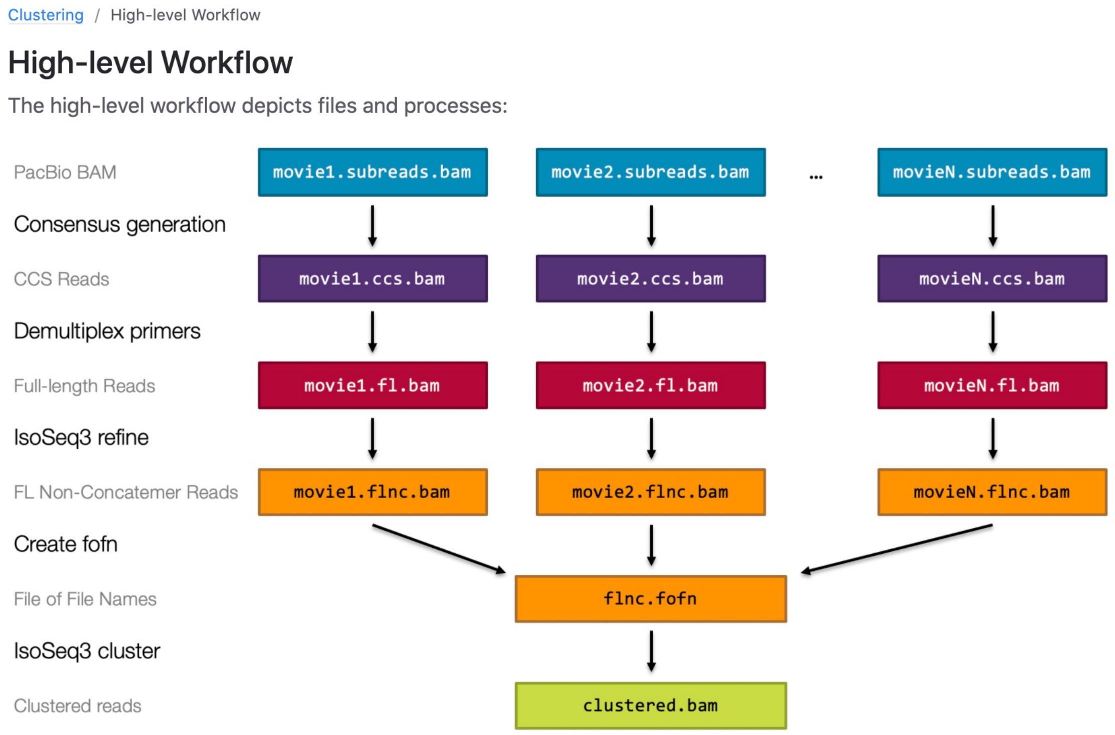
```
$ source activate <name of your environment>
```

## Check your installation

```
$ isoseq3 --version
isoseq3 3.4.x
$ lima --version
lima 1.11.0
$ pbmm2 --version
pbmm2 1.3.0
```

PACBIO®

# ISO-SEQ WORKFLOW



Clustering / High-level Workflow

## High-level Workflow

The high-level workflow depicts files and processes:

PacBio BAM: movie1.subreads.bam, movie2.subreads.bam ... movieN.subreads.bam

Consensus generation

CCS Reads: movie1.ccs.bam, movie2.ccs.bam, movieN.ccs.bam

Demultiplex primers

Full-length Reads: movie1.fl.bam, movie2.fl.bam, movieN.fl.bam

IsoSeq3 refine

FL Non-Concatemer Reads: movie1.flnc.bam, movie2.flnc.bam, movieN.flnc.bam

Create fofn

File of File Names: flnc.fofn

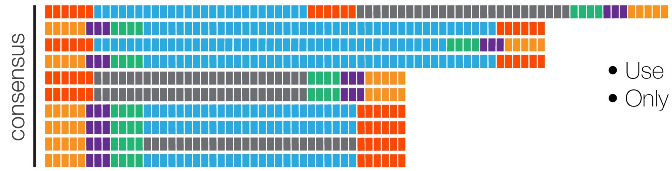
IsoSeq3 cluster

Clustered reads: clustered.bam

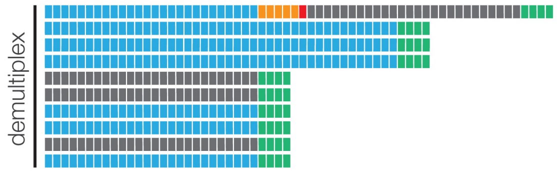
← Most SQ2 projects will directly give you ccs.bam now

https://isoseq.how/clustering/high-level-workflow.html

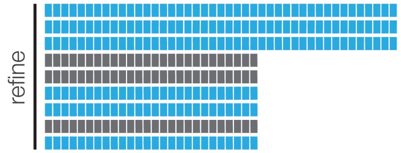
# ISO-SEQ WORKFLOW



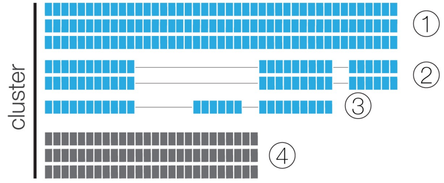
- Use **polished** CCS reads
- Only full-pass ZMWs



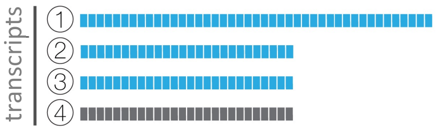
- Barcoded and unbarcoded cDNA primer removal
- Orientation
- Unwanted primer combination removal



- PolyA tail trimming
- Concatemer removal

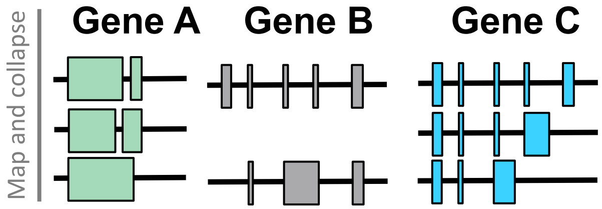


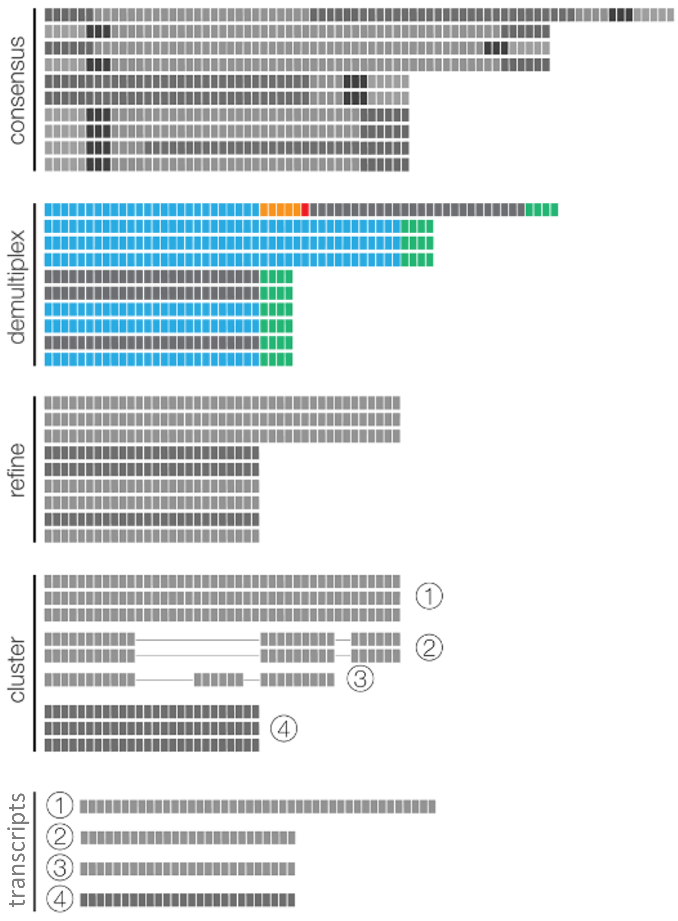
- Hierarchical, n*log(n) clustering, alignment of shorter to longer sequences
- Iterative cluster merging
- Generate consensus for each read cluster using QV guided PoA



- One consensus per read cluster

- Align to reference genome
- Remove redundancy





# PRIMER REMOVAL & DEMULTIPLEXING

Command line：

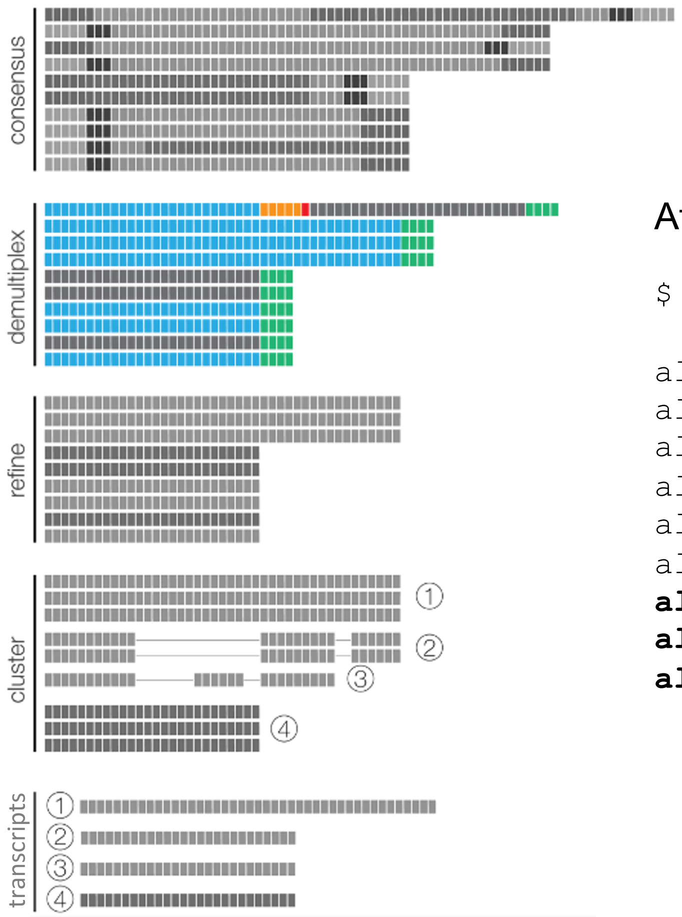
```
lima --isoseq --dump-clips --peek-guess -j 24\
alz.ccs.bam isoseq_primers.fasta alz.demult.bam
```

```
Input files:
     alz.ccs.bam  #HiFi reads
     isoseq_primers.fasta #Iso-Seq primers

Output files:
     alz.demult.bam

Options:
     --isoseq #specialized isoseq option for lima
     --dump-clips # show the clipped primers
     --peek-guess # remove spurious false positive signal
     -j 24 # Number of threads to use
```

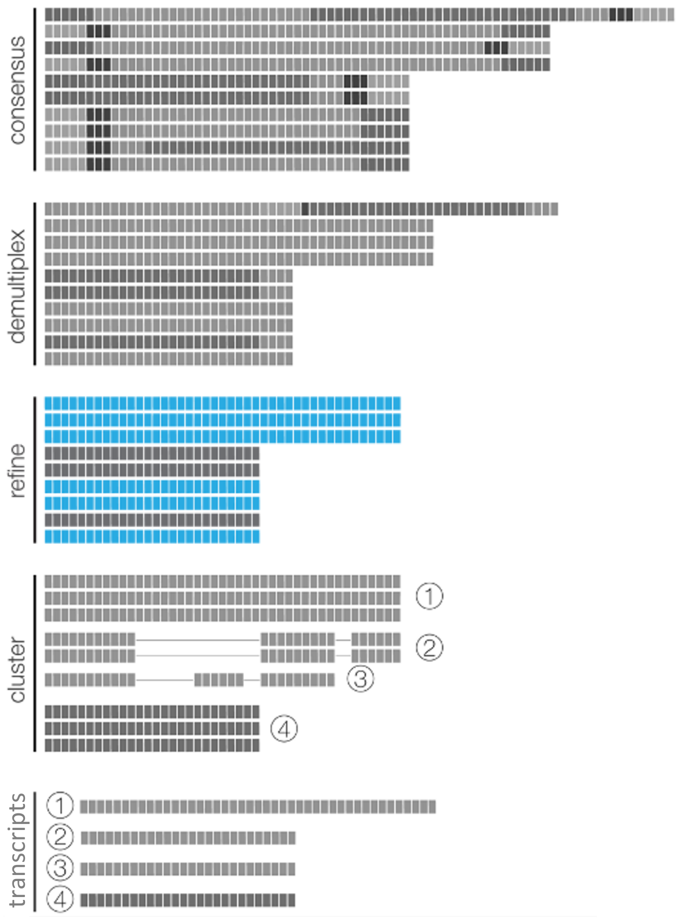
# PRIMER REMOVAL & DEMULTIPLEXING



After completion, you will see the following files:

```
$ ls -ltrh

alz.demult.json
alz.demult.lima.clips
alz.demult.lima.counts
alz.demult.lima.guess          #lima reports
alz.demult.lima.report
alz.demult.lima.summary
alz.demult.5p--3p.bam
alz.demult.5p--3p.bam.pbi
alz.demult.5p--3p.subreadset.xml
```



# TRIMMING POLY(A) TAILS AND CONCATEMER REMOVAL

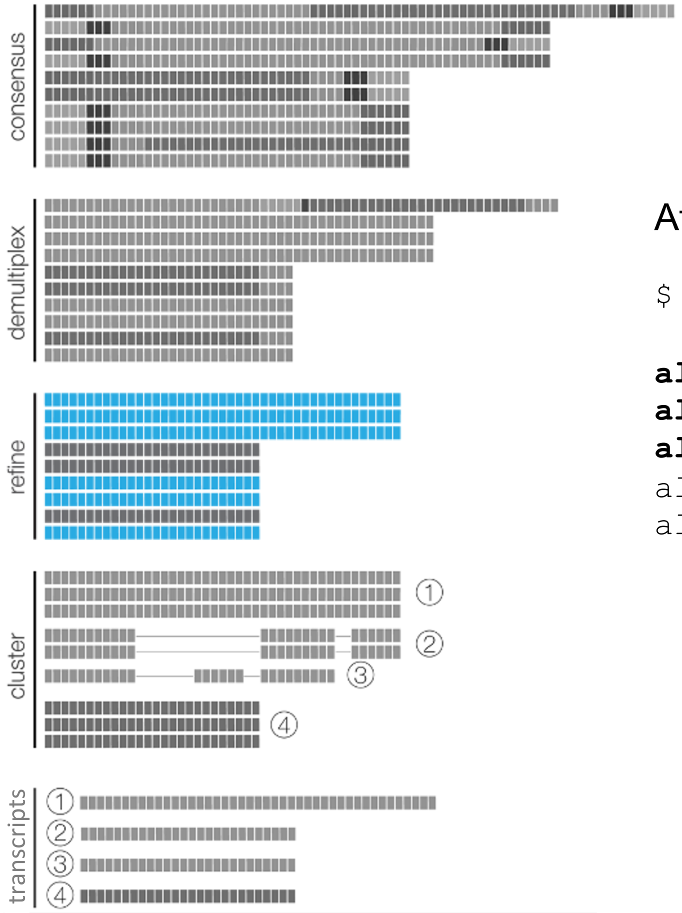
Command line :

```
isoseq3 refine --require-polya\
alz.demult.5p--3p.bam\ isoseq_primers.fasta
alz.flnc.bam
```

```
Input files:
      alz.demult.5p--3p.bam
      isoseq_primers.fasta

Output files:
      alz.flnc.bam

Options:
       --require-polya #if your transcripts have a polyA tail
```



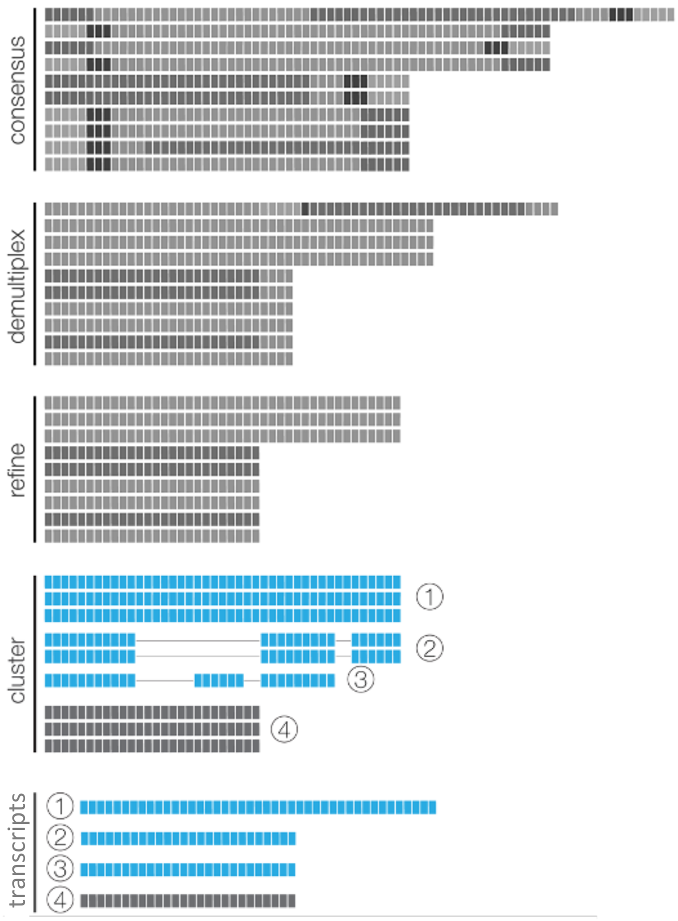
# TRIMMING POLY(A) TAILS AND CONCATEMER REMOVAL

After completion, you will see the following files:

```
$ ls -ltrh

alz.flnc.bam
alz.flnc.bam.pbi
alz.flnc.consensusreadset.xml
alz.flnc.filter_summary.json    #isoseq3 refine reports
alz.flnc.report.csv
```

# ISOFORMS



Command line :

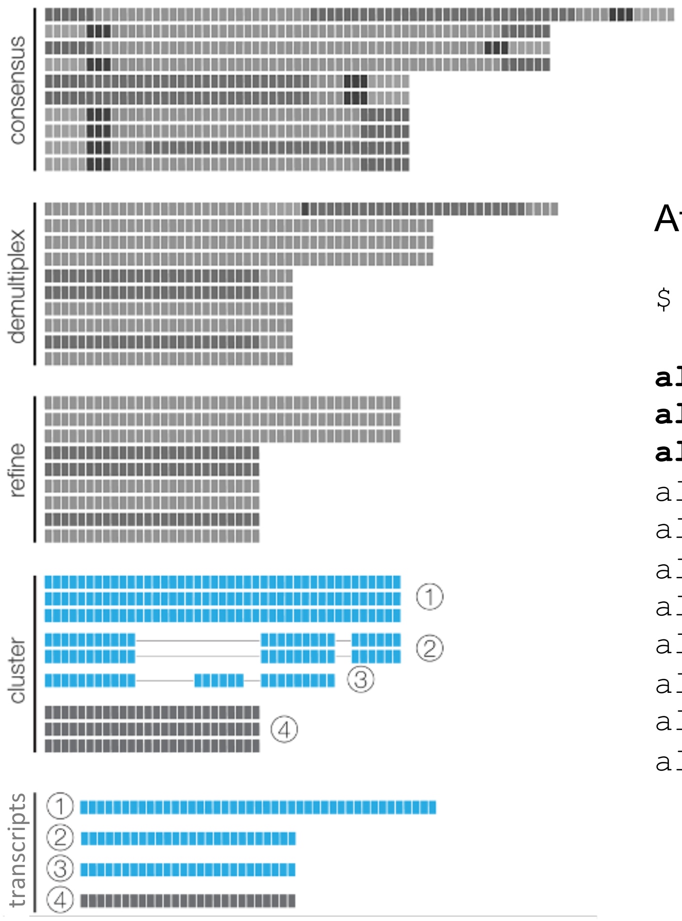
```
isoseq3 cluster alz.flnc.bam  alz.polished.bam \
--verbose --use-qvs
```

```
Input files:
      alz.flnc.bam

Output files:
     alz.polished.bam

Options:
     --verbose #if your transcripts have a polyA tail
     --use-qvs #Use CCS QVs, sets --poa-cov 100
```

# ISOFORMS



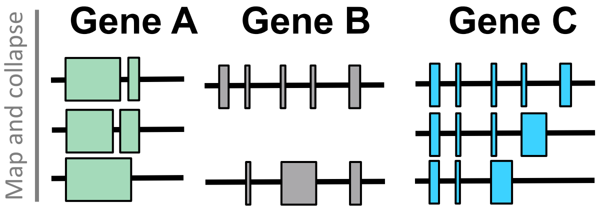
After completion, you will see the following files:

Because the ccs input is Polished, the isoseq3 **cluster output is already polished!**

```
$ ls -ltrh

alz.polished.bam
alz.polished.bam.pbi
alz.polished.transcriptset.xml
alz.polished.cluster                  #isoseq3 cluster reports
alz.polished.cluster_report.csv
alz.polished.hq.bam
alz.polished.hq.bam.pbi          #high quality isoforms(≥0.99)
alz.polished.hq.fasta.gz
alz.polished.lq.bam
alz.polished.lq.bam.pbi          #low quality isoforms(<0.99)
alz.polished.lq.fasta.gz
```

# MAP



Command line：

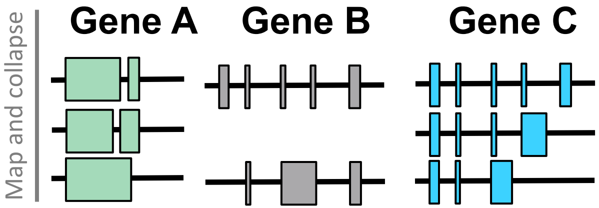
```
pbmm2 align hg38.fa alz.polished.hq.bam
alz.aligned.bam
-j 24 --preset ISOSEQ -sort --log-level INFO
```

```
Input files:
     alz.polished.hq.bam
     hg38.fa

Output files:
      alz.aligned.bam

Options:
     -j 24 #Number of threads to use
     --preset ISOSEQ #select the alignment mode
     --sort #Generate sorted BAM file
     --log-level INFO #show progress
```

# MAP

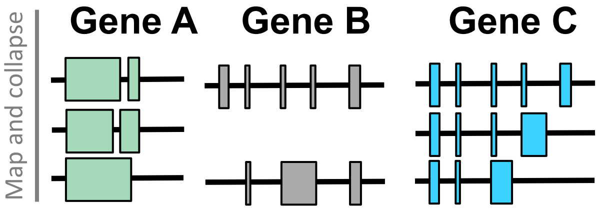


After completion, you will see the following files:

```
$ ls -ltrh

alz.aligned.bam
alz.aligned.bam.bai
```

# COLLAPSE



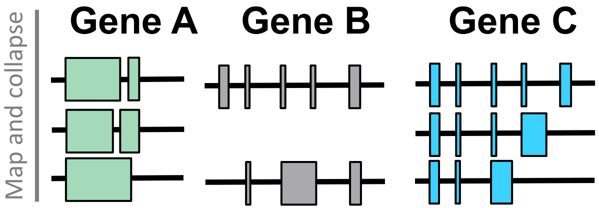
Command line :

```
isoseq3 collapse alz.aligned.bam alz.collapsed.gff
```

```
Input files:
      alz.aligned.bam

Output files:
      alz.collapsed.gff
```

# COLLAPSE



After completion, you will see the following files:

```
$ ls -ltrh

alz.collapsed.report.json
alz.collapsed.abundance.txt      #report, stats and list
alz.collapsed.read_stat.txt
alz.collapsed.group.txt
alz.collapsed.gff
alz.collapsed.fasta
```

# ISO-SEQ ANALYSIS TERMINOLOGY

| NAME | ABBR | EXPLANATION |
|---|---|---|
| **Full-Length Reads** | FL Reads | CCS reads with 5’ and 3’ cDNA primers removed |
| **Full-Length, Non-Concatemer Reads** | FLNC Reads | CCS reads with 5’ and 3’ cDNA primers, polyA tail, and concatemers removed |
| **High-Quality Isoforms** | HQ Isoforms | Polished transcript sequences with predicted accuracy ≥99% & ≥2 FLNC |
| **Low-Quality Isoforms** | LQ Isoforms | Polished transcript sequences with predicted accuracy <99% & ≥2 FLNC |

PACBIO®

www.pacb.com

For Research Use Only. Not for use in diagnostic procedures. © Copyright 2020 by Pacific Biosciences of California, Inc. All rights reserved. Pacific Biosciences, the Pacific Biosciences logo, PacBio, SMRT, SMRTbell, Iso-Seq, and Sequel are trademarks of Pacific Biosciences. Pacific Biosciences does not sell a kit for carrying out the overall No-Amp Targeted Sequencing method. Use of these No-Amp methods may require rights to third-party owned intellectual property. BluePippin and SageELF are trademarks of Sage Science. NGS-go and NGSengine are trademarks of GenDx. FEMTO Pulse and Fragment Analyzer are trademarks of Agilent Technologies Inc.

All other trademarks are the sole property of their respective owners.