Blatantly Commandeered Slides
PacBio RS
Single Molecule Sequencing

Tristan De Buysscher

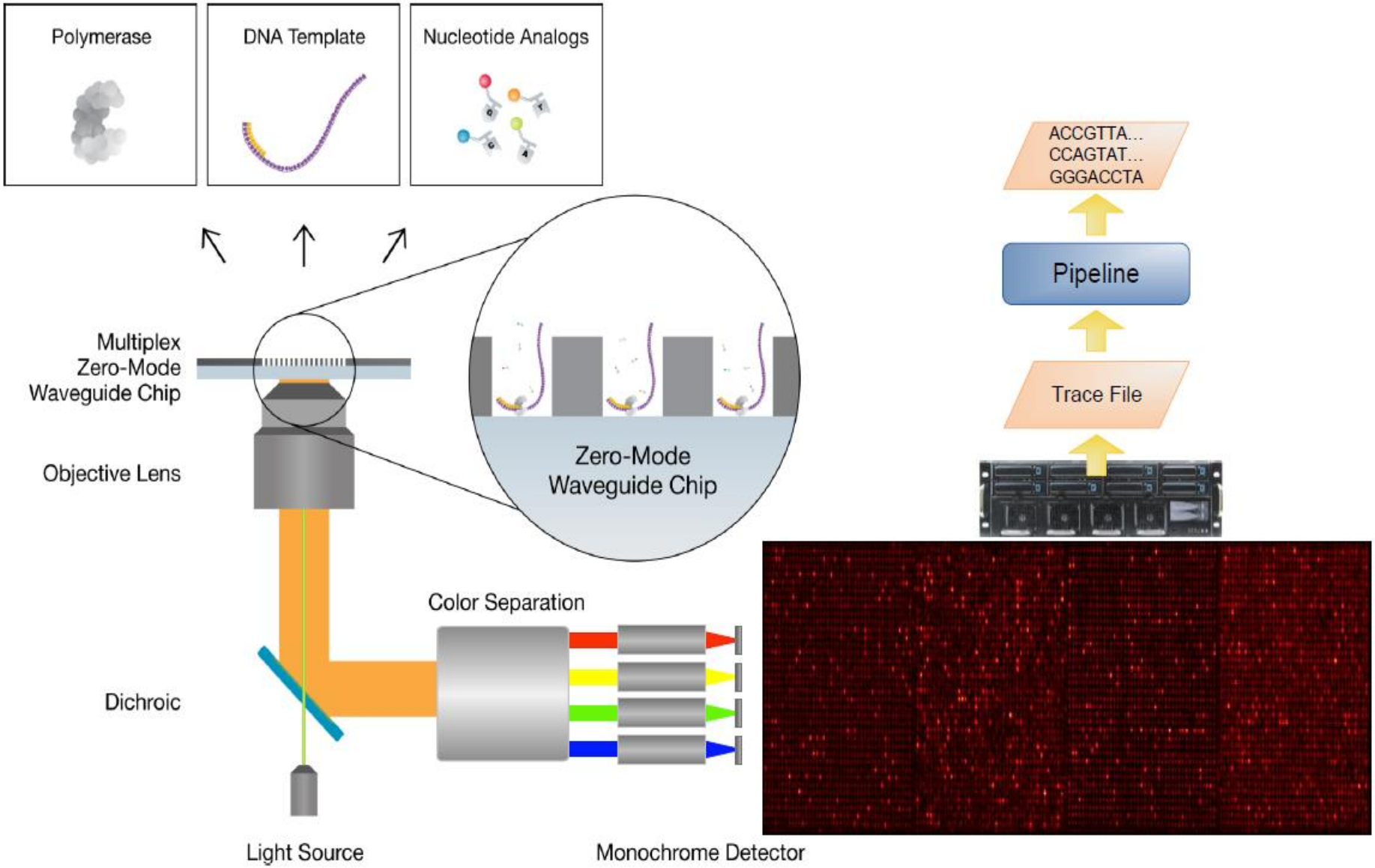
UNC Center for Bioinformatics

NGS 2013

PacBio® *RS*

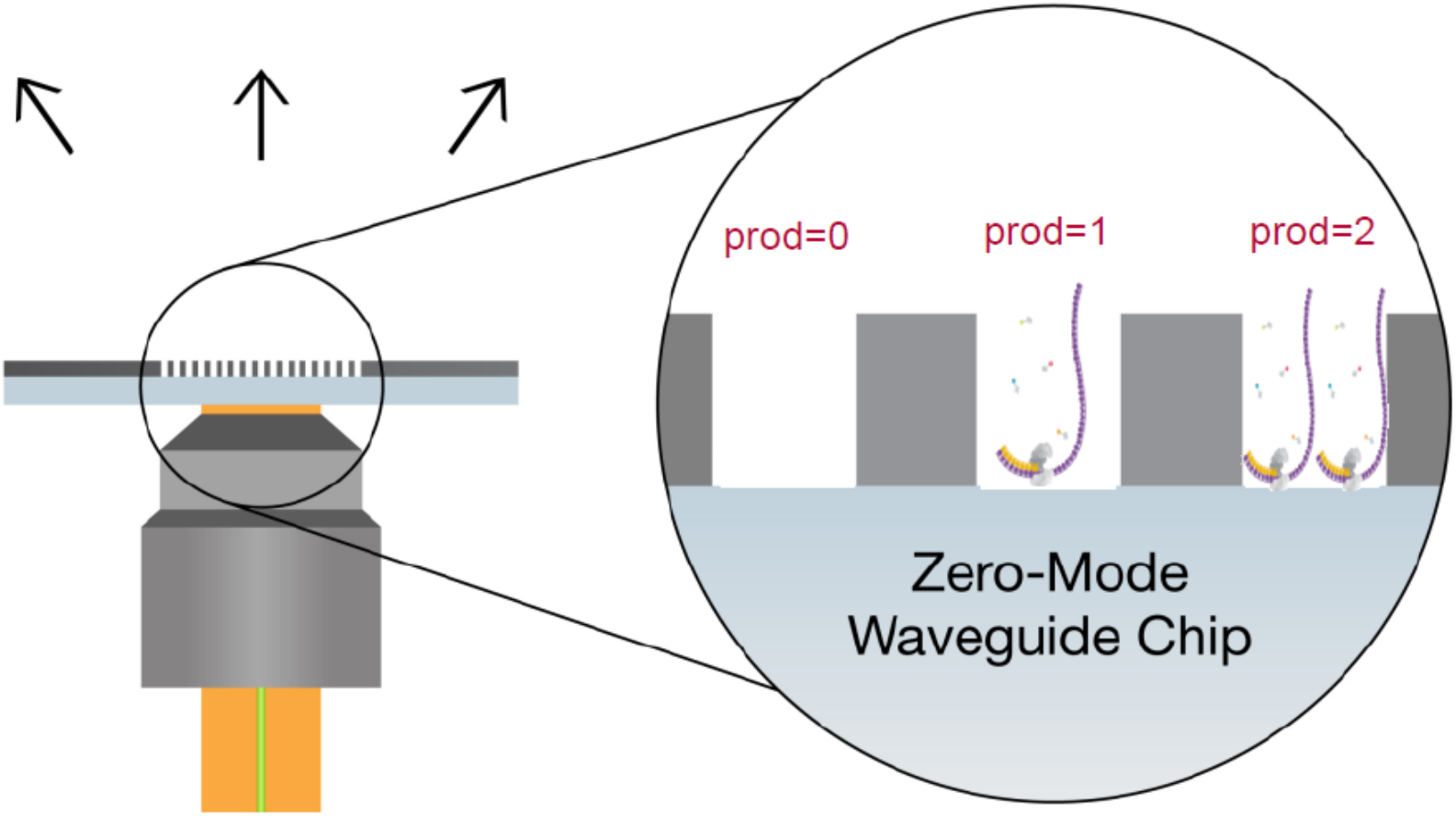
SMRT® Technology

SMRT® Cell


Polymerase
DNA Template
Nucleotide Analogs
Multiplex
Zero-Mode
Waveguide Chip
Objective Lens
Zero-Mode
Waveguide Chip
Color Separation
Dichroic
Light Source
Monochrome Detector
ACCGTTA…
CCAGTAT…
GGGACCTA
Pipeline
Trace File

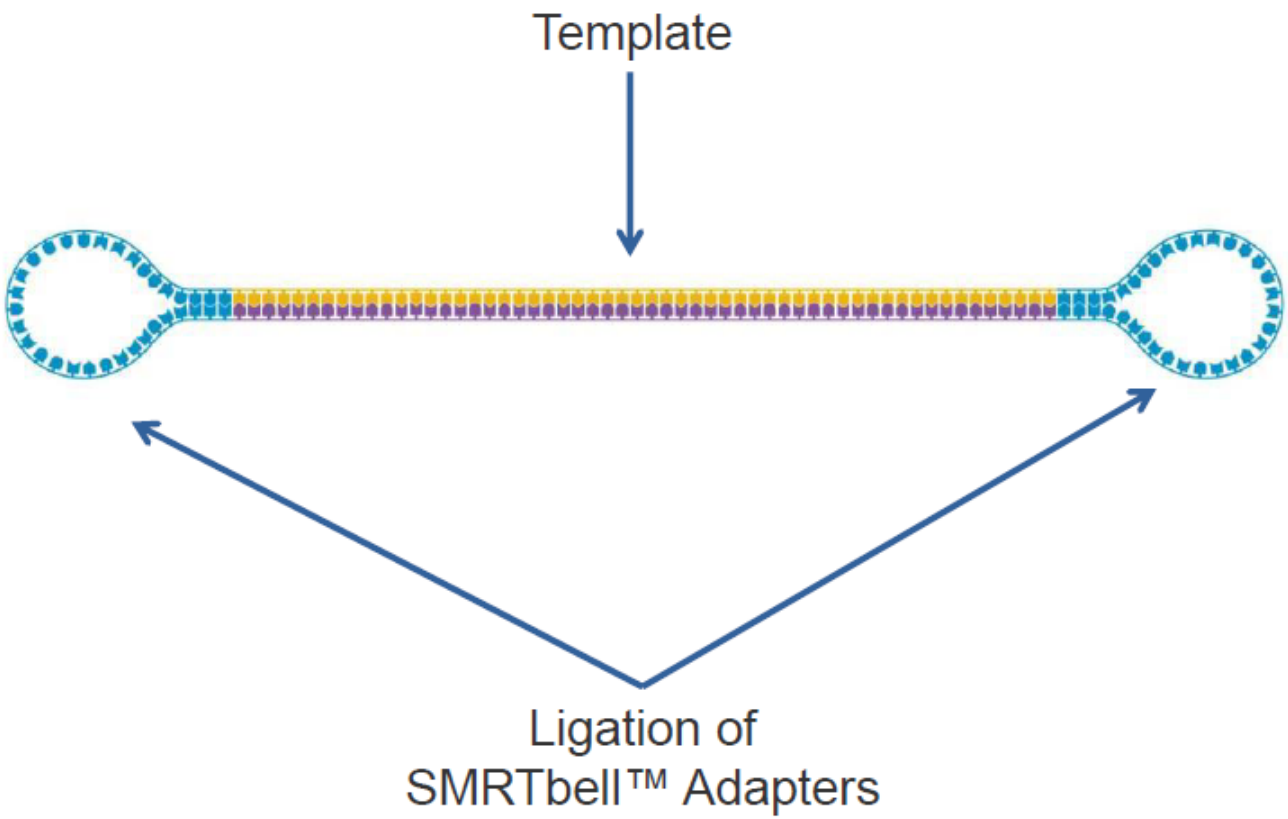
Productivity

- An estimate of the number of active polymerases in a ZMW
- Number varies due to diffusion loading
- Goal: prod=1



PACIFIC
BIOSCIENC

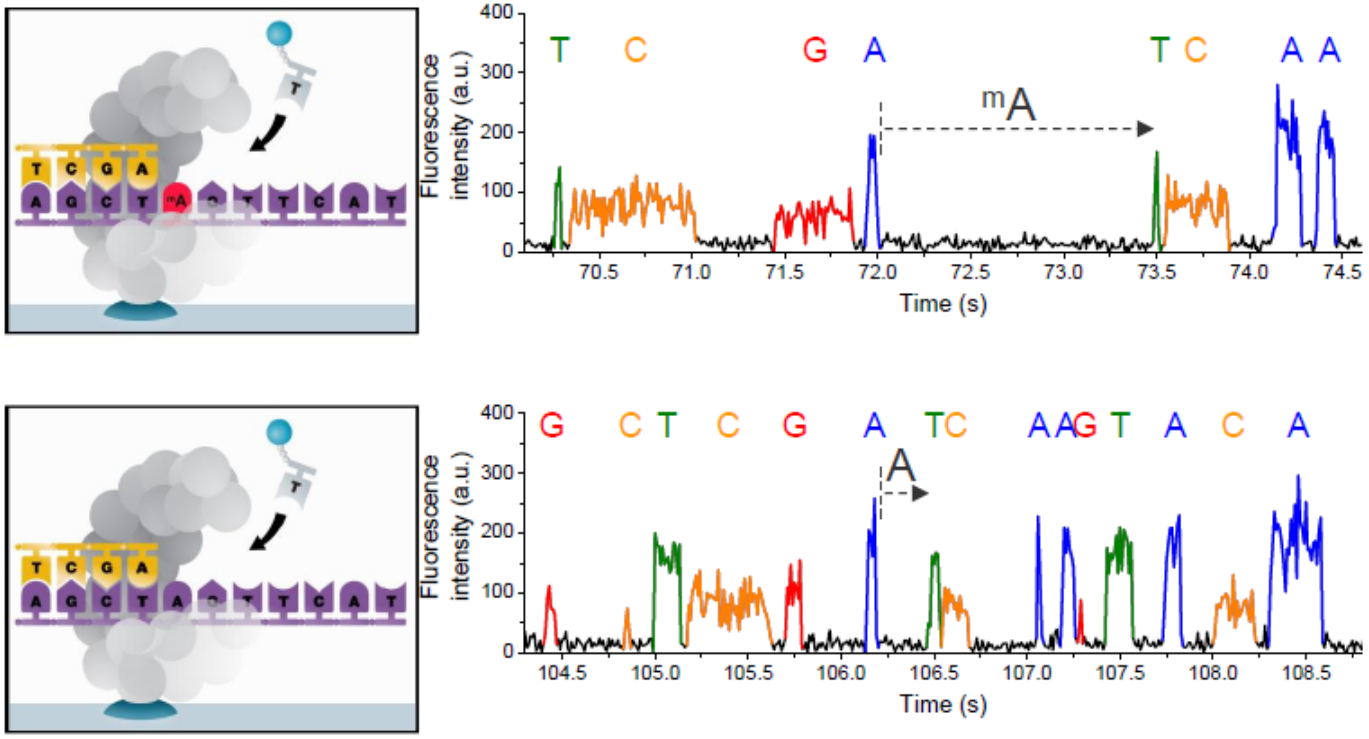
SMRTbell™ Sample Preparation



SMRTbell™ Sample Preparation

PACIFIC
BIOSCIENC

Kinetic Information

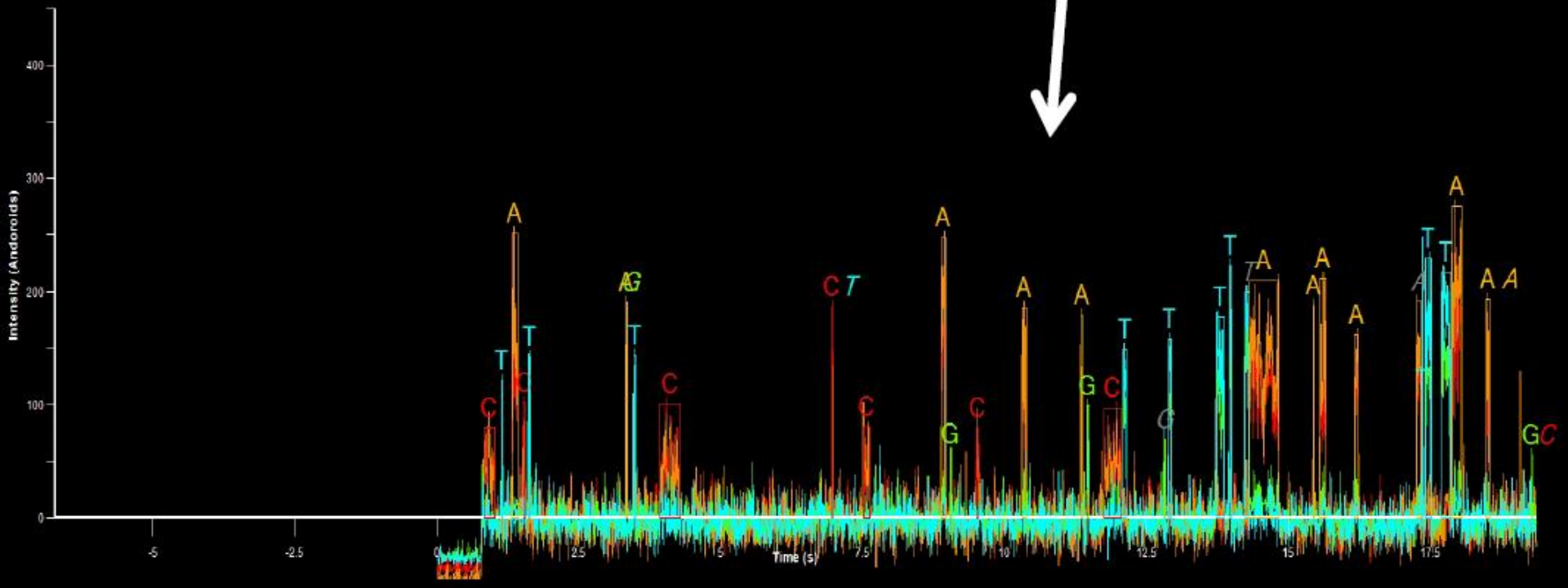


- Differentiation between modified and non-modified bases
  - Epigenetics, DNA damage, new, novel modifications
- Direct observation (*e.g.* no bisulfite)

PACIFIC BIOSCIENC

Signal Processing and Base Calling

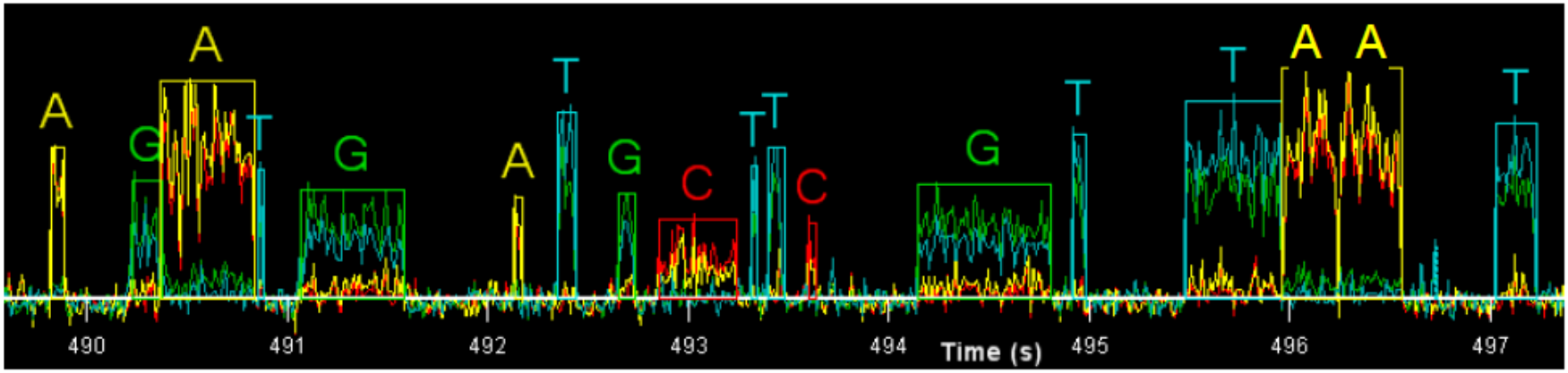
Converting pulses of light into DNA bases and kinetic measures



PACIFIC BIOSCIENCE

PulseToBase Summary

- Single-molecule pulse events are sequential: No phasing problem, no Sanger limit!
- Main kinetic information retained in the **bas.h5** output files are Inter-pulse duration (IPD) and Pulse Width (PW)
- Quality Values
  - Substitution
  - Insertion
  - Deletion
  - Merge
  - Sum of all error probabilities

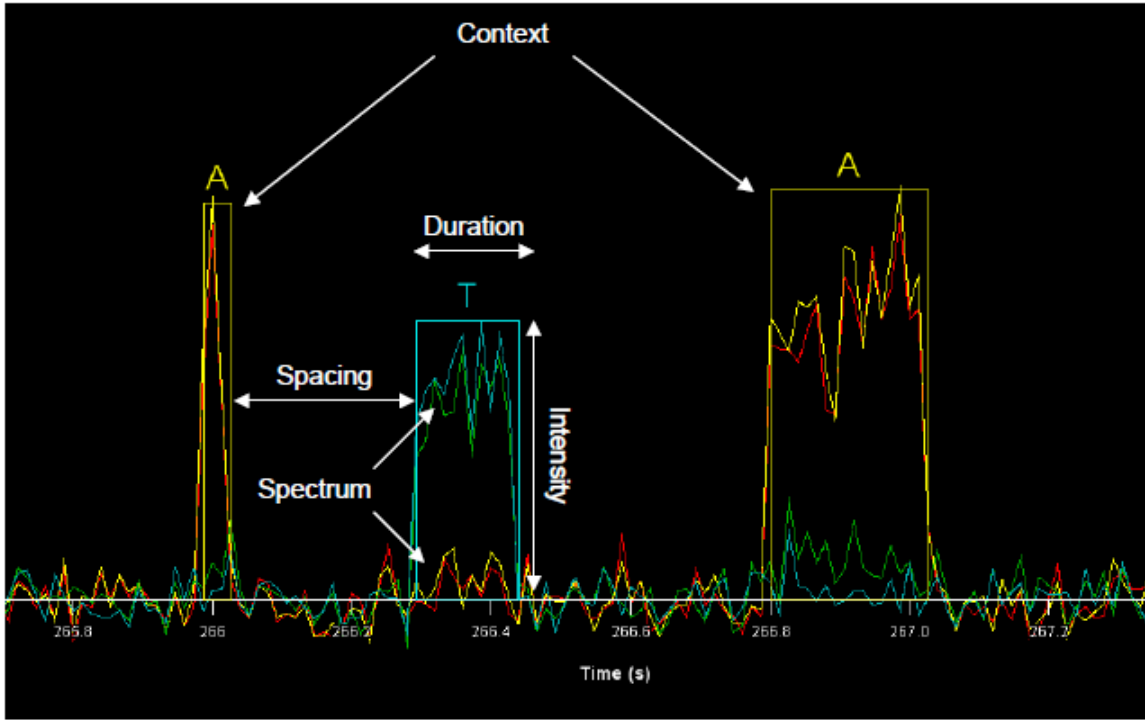


PACIFIC BIOSCIENC

PulseToBase Inputs

- Receive observation list from TraceToPulse

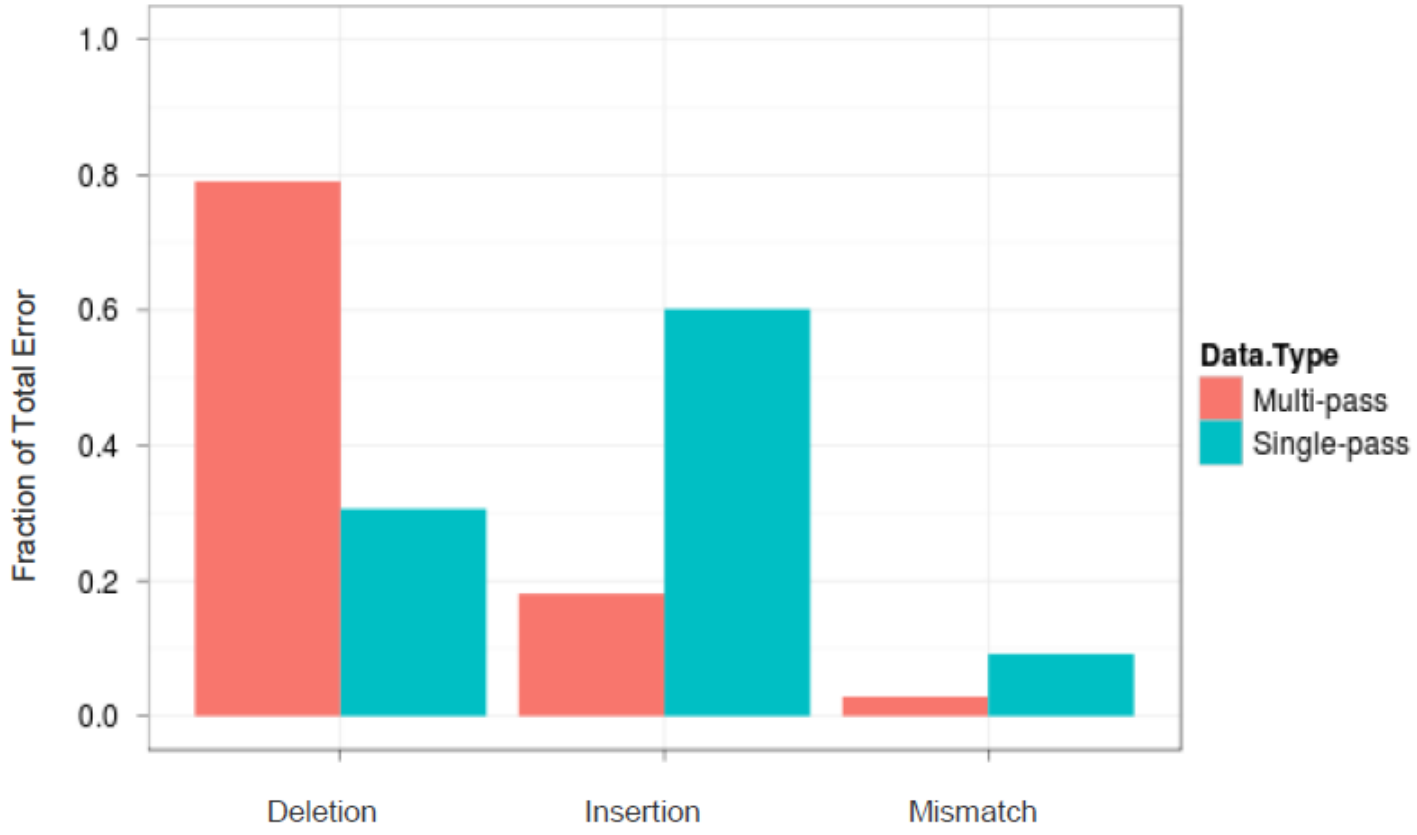
- Each pulse has vector of associated measurements
  - Duration
  - Spectrum
  - Intensity
  - Spacing to neighbors
  - Local context
  - etc…



PACIFIC BIOSCIENCES

Typical Error Profiles

Errors are random and dominated by indels



Universal SMRTbell™ Template

Standard Sequencing for Continuous Long Reads (CLR)

Large Insert Sizes

- Recommended Insert Size: > 2 kb
- Recommended Movie Collection Time: 1 x 90 min

Generates one pass on each molecule sequenced

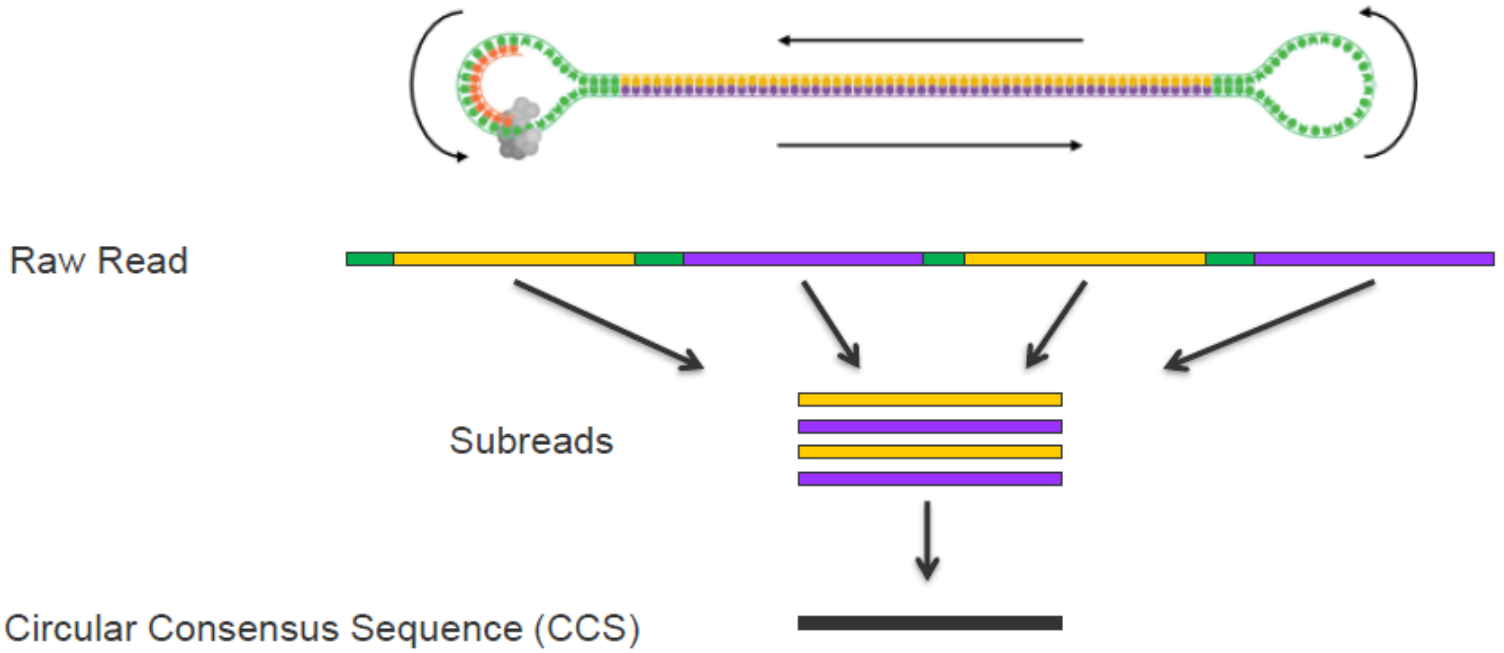
Circular Consensus Sequencing (CCS)

Small Insert Sizes

- Recommended Insert Size: 250 bp-2 kb
- Recommended Movie Collection Time: 2 x 45 min

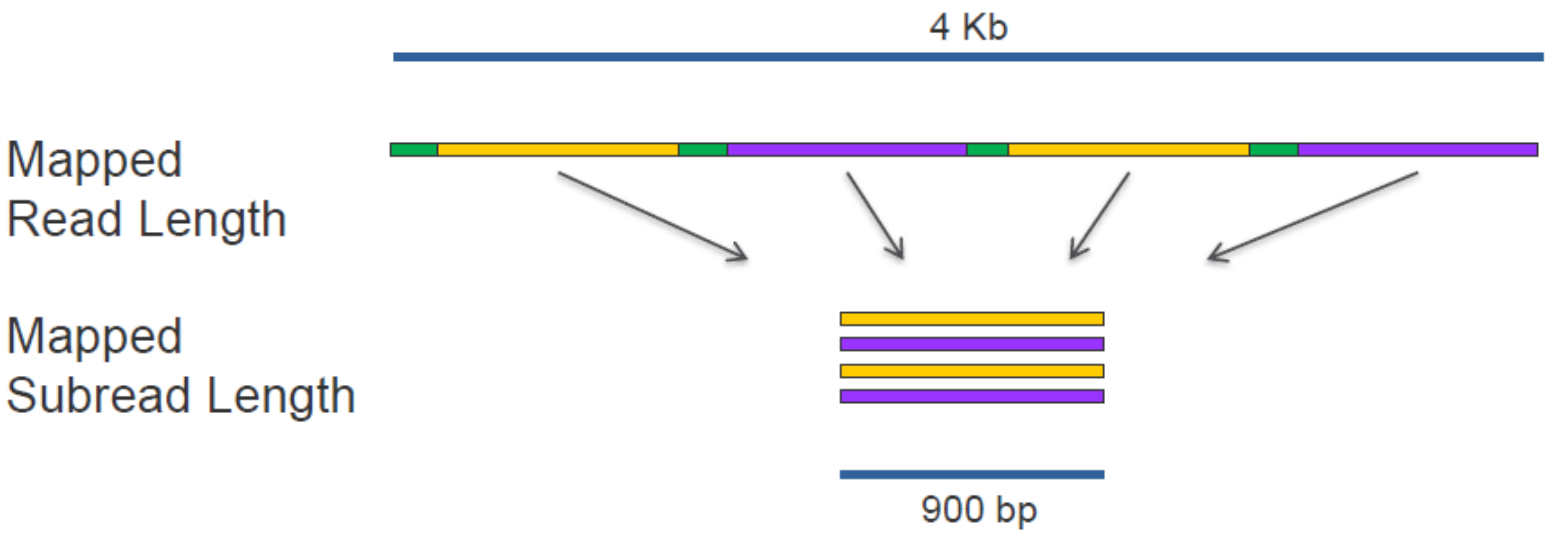
Generates multiple passes on each molecule sequenced

From Raw Reads to CCS



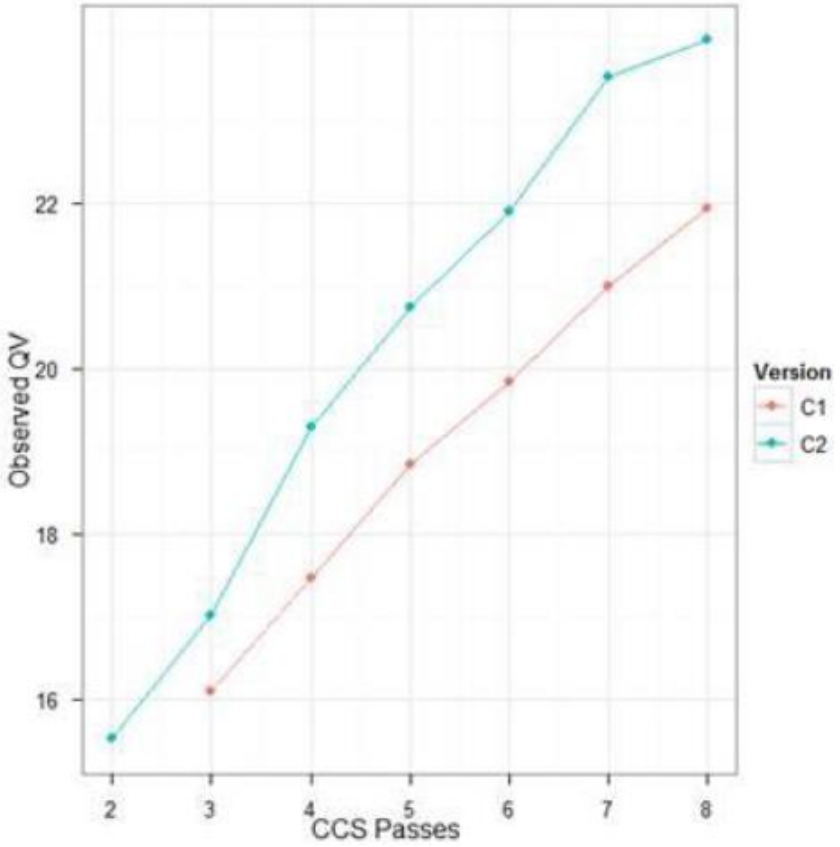
- Read Length: Length of the raw read
- Mapped Read Length: A composite of all mapped subreads and adaptors
- Sub-reads (purple and gold) are separated by adapter sequences (green)
- ≥2 full passes required for CCS
- CCS or individual subreads can be used for subsequent analysis

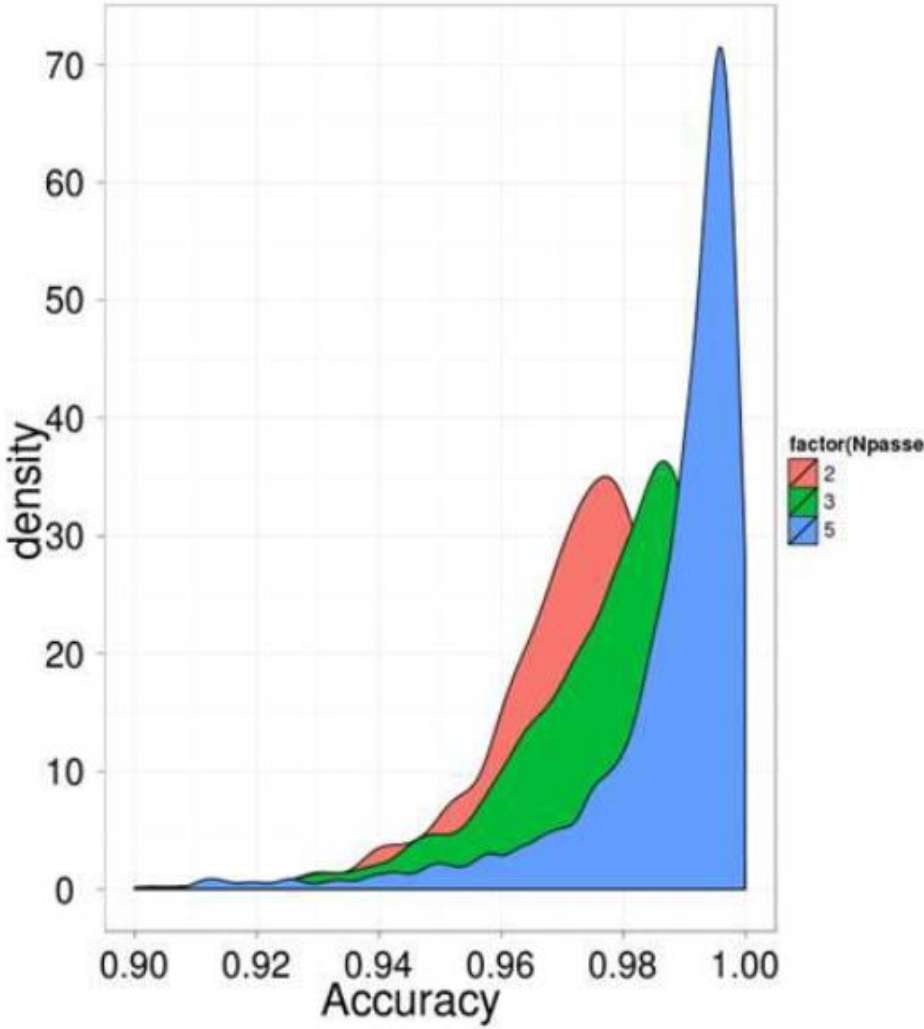
Mapped Subread vs. Mapped Read Length



Mapped Read Length	Mapped Sub-read Length
Measure of ZMW sequencing productivity	Measure of scientifically applicable sequence
Upper bound by speed and fidelity of the polymerase and movie time	Upper bound by insert size and loading effects

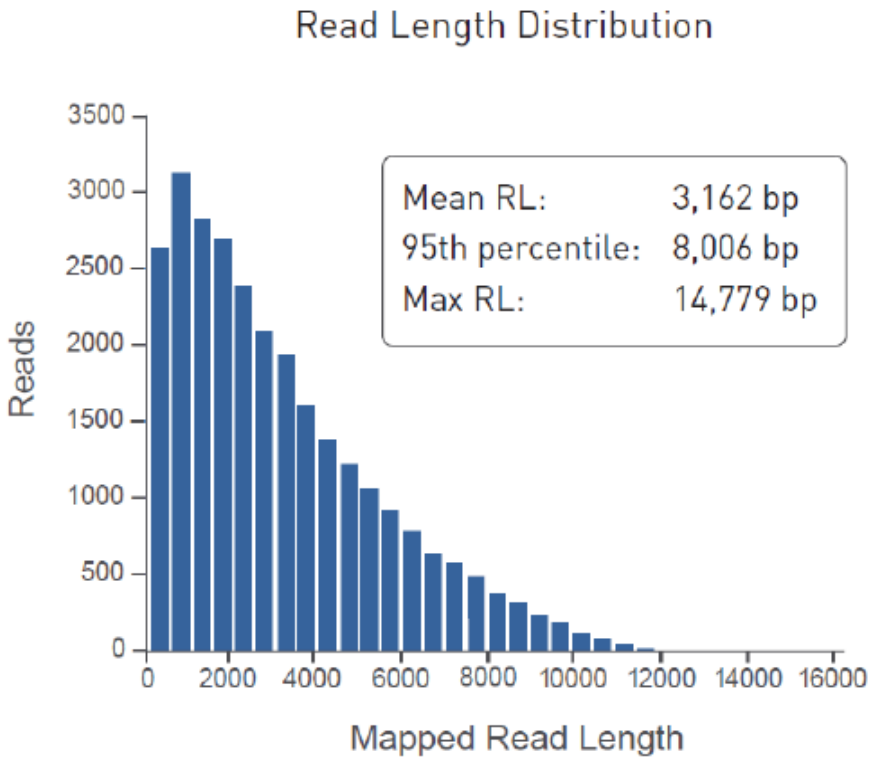
CCS Consensus Quality improvements with C2 and v1.3 upgrad

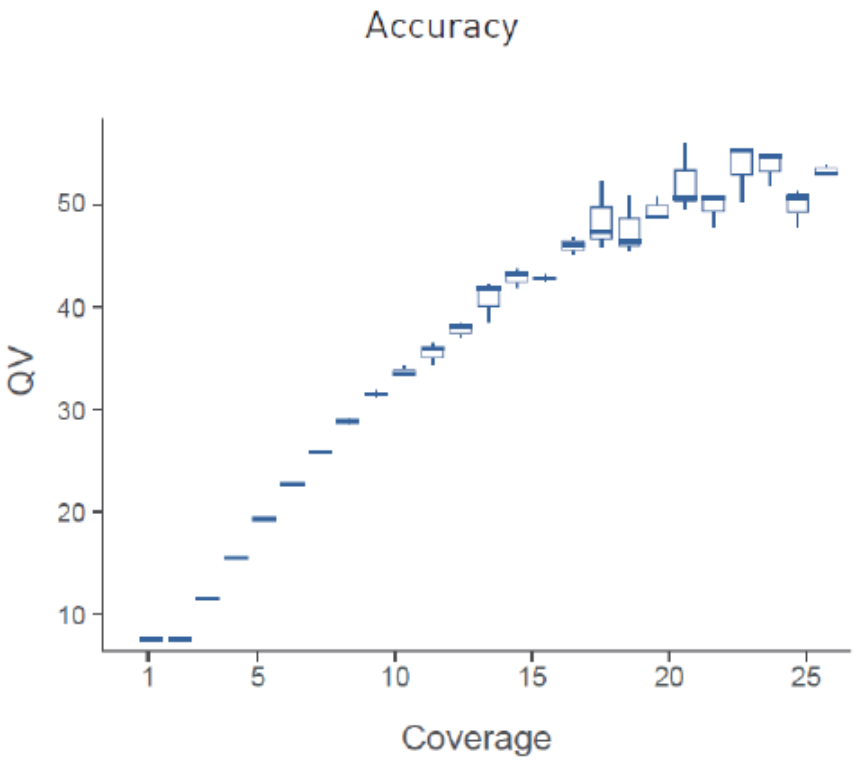




Consensus Accuracy improves with higher coverage due to random error profile

Exponential ReadLength Distribution and Consensus Accuracy

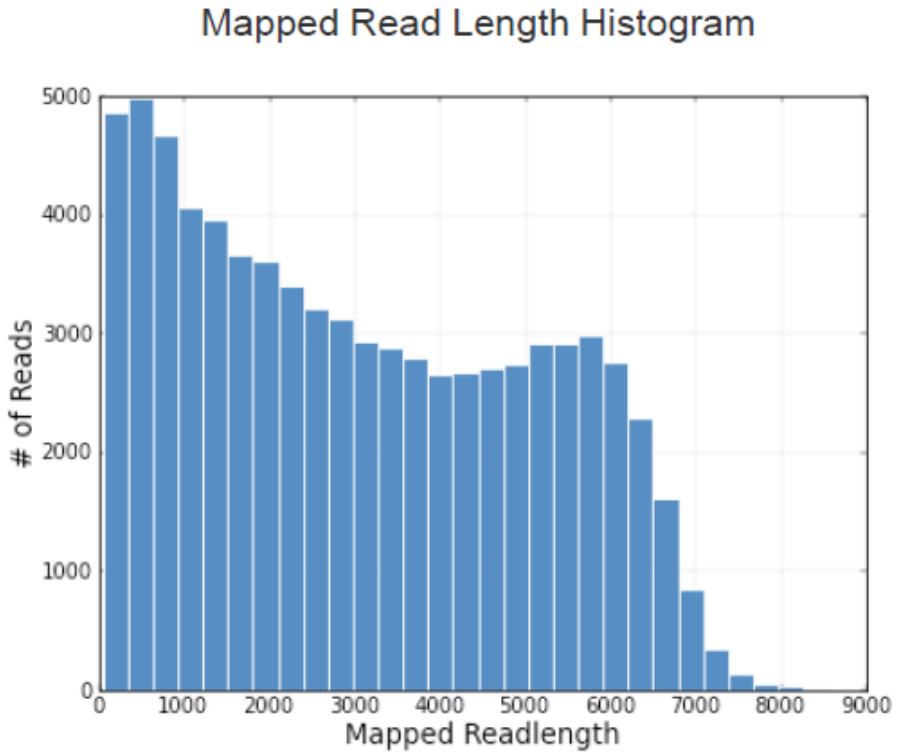


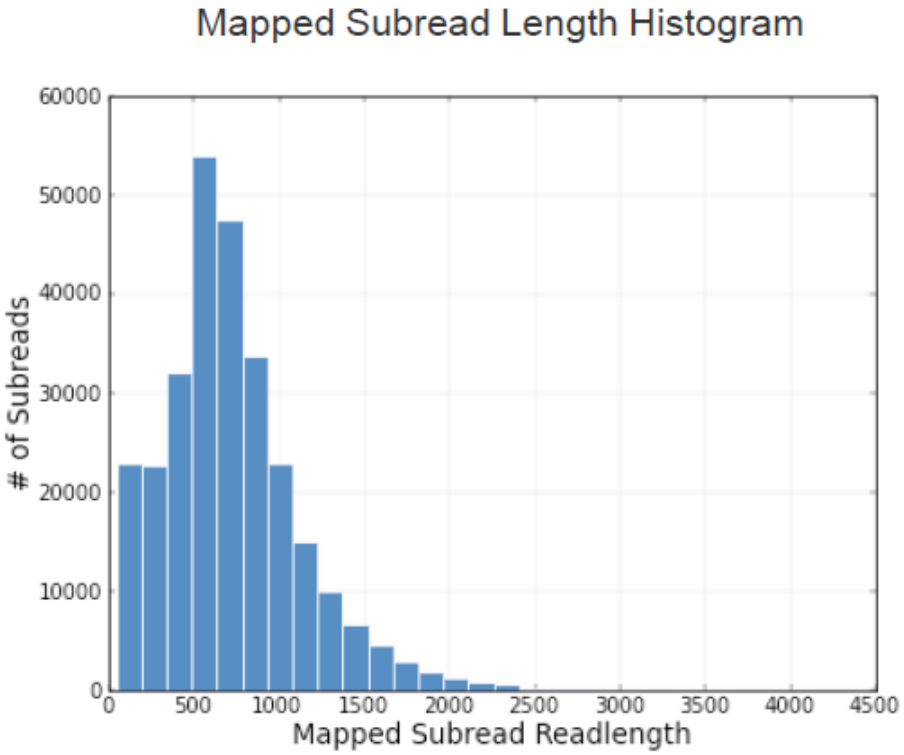


Based on data from *E. coli* with 10 kb libraries using a 90 minute movie.

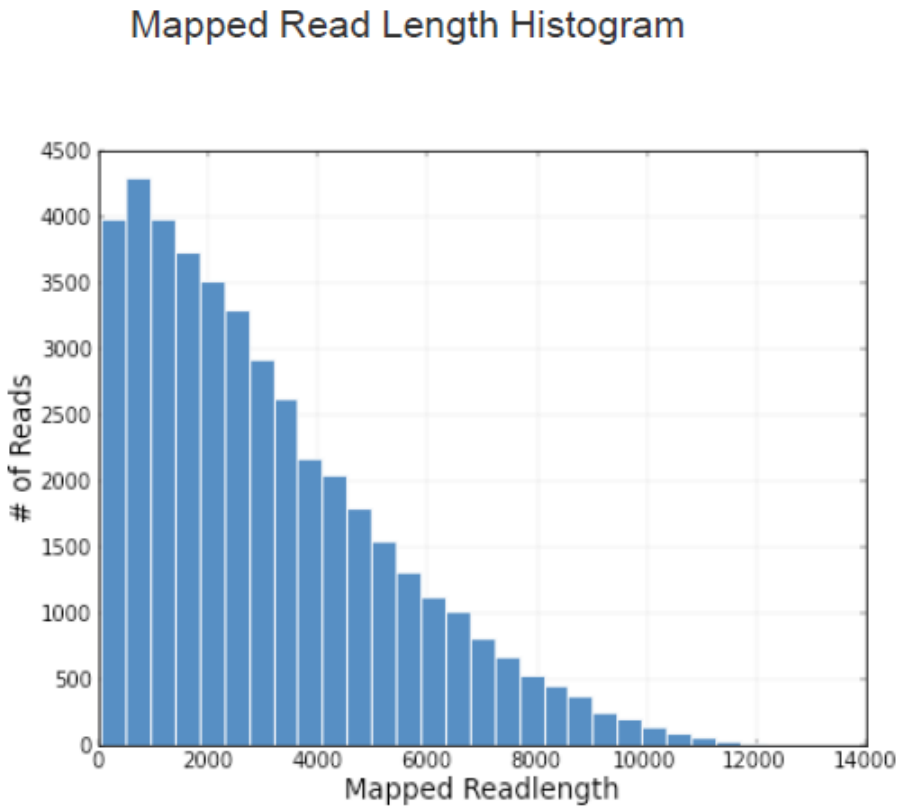
PACIFIC BIOSCIENCE

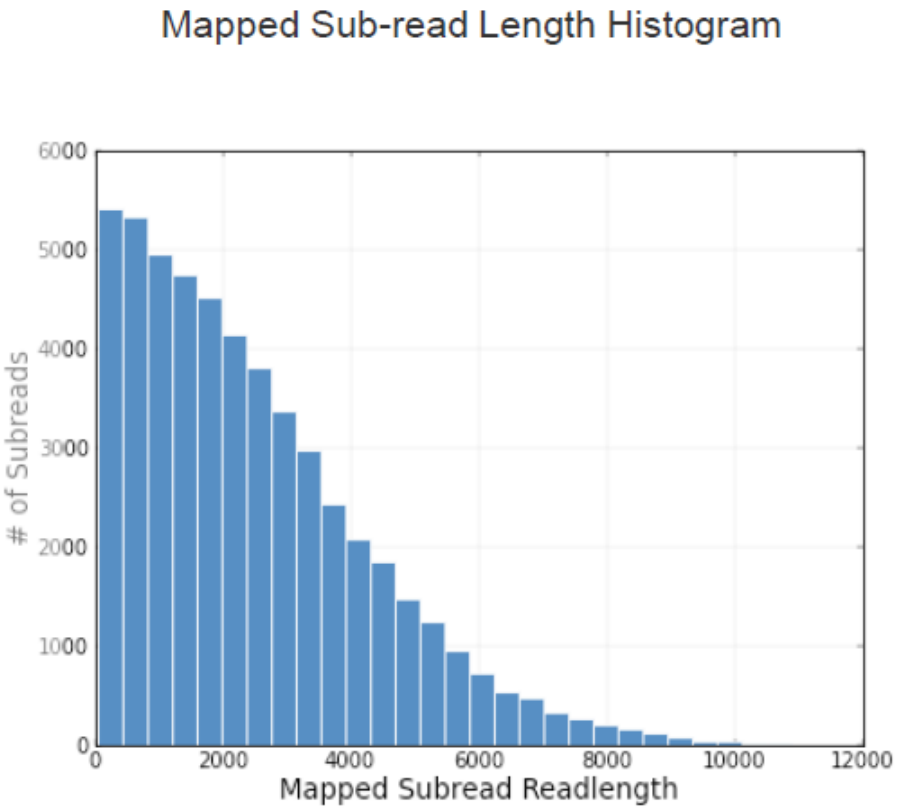
Comparison of Mapped Read Length & Subread Length Distributions for a 2 kb Library





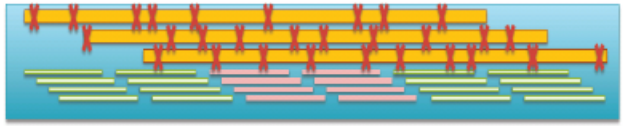
Comparison of Mapped Read Length & Sub-read Length Distributions for a 10 kb Library



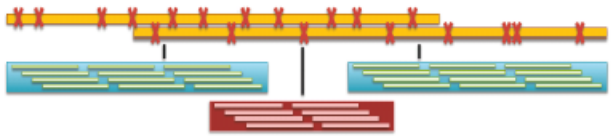


SMRT® Assembly Tools

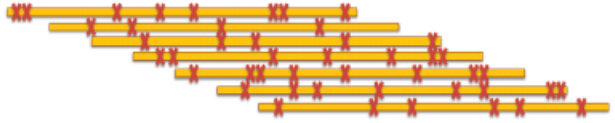
- SMRT Hybrid: The hybrid assembly of error-corrected reads
  - Celera® Assembler
  - P_ErrorCorrection/Allora
  - ALLPATHS-LG
  - MIRA



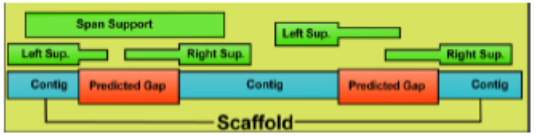
- SMRT Scaffolding: Using PacBio CLR to scaffold existing contigs
  - AHA



- SMRT *de novo*: The assembly of PacBio CLR data only
  - Allora



- SMRT Gap Filling: Using PacBio CLR to fill gaps in existing scaffolds
  - PBJelly



PACIFIC BIOSCIENC