Nanopore Sequencing

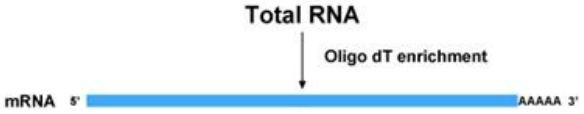
Long reads matter

Sebastian Krautwurst, Kevin Lamkiewicz

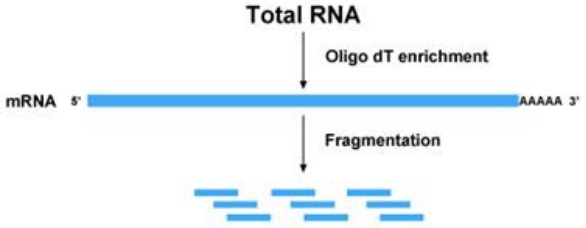
17.01.2019
RNA Bioinformatics and High-Throughput Analysis

Sequenzierung und Assemblierung

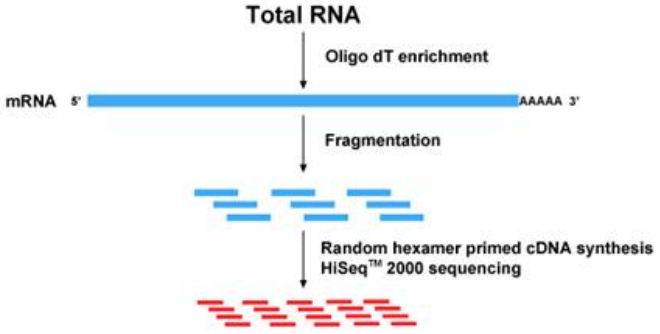
RNA SEQ



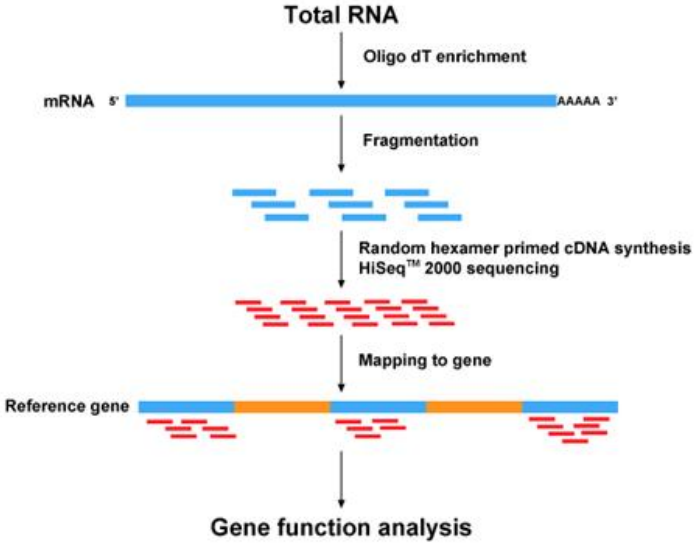
RNA SEQ



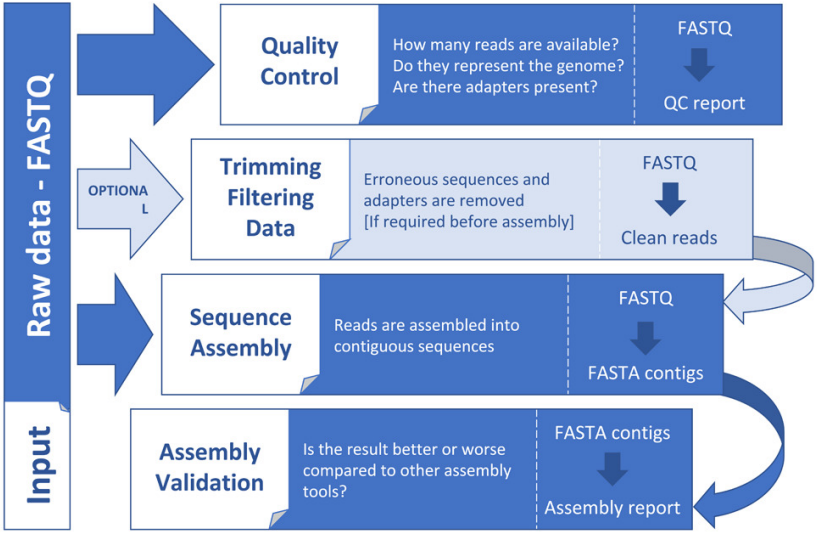
RNA SEQ


Total RNA
Oligo dT enrichment
mRNA 5'
AAAAA 3'
Fragmentation
Random hexamer primed cDNA synthesis
HiSeq™ 2000 sequencing

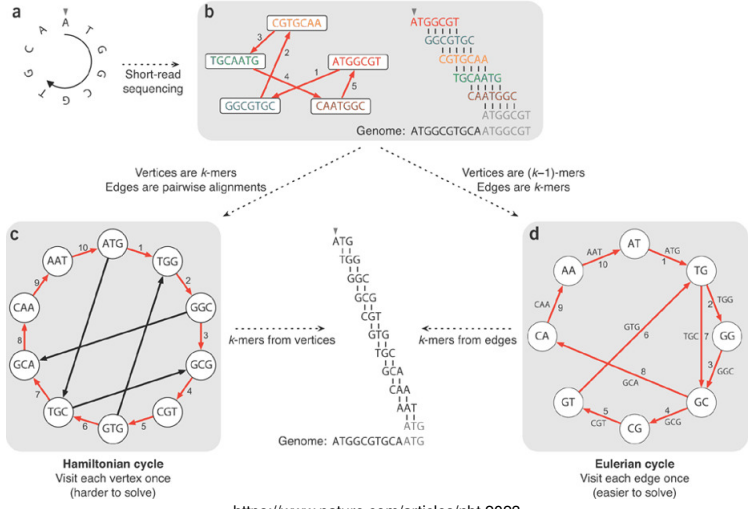
RNA SEQ


Total RNA
Oligo dT enrichment
mRNA 5'
AAAAA 3'
Fragmentation
Random hexamer primed cDNA synthesis
HiSeq™ 2000 sequencing
Mapping to gene
Reference gene
Gene function analysis

ASSEMBLIERUNGS-PIPELINE



K-MER UND DE BRUIJN GRAPHEN



https://www.nature.com/articles/nbt.2023

Unser heutiger Use-Case: Mykoplasmen

CRASHKURS: MYKOPLASMEN

Was sagt uns Wikipedia?

Mycoplasma, deutsch auch als Mykoplasmen bezeichnet sind eine Gattung sehr kleiner Bakterien aus der Klasse der Mollicutes. Im Gegensatz zu allen anderen Bakterien fehlt ihnen eine Zellwand. [...]
Mykoplasmen sind meist parasitär, intra- und extrazellulär lebende Bakterien, die beim Menschen und Wirbeltieren die Ursache für zahlreiche Krankheiten sind. [...]
Mit einer Größe von 580 –– 1.380 kbp haben die Gattungen Mycoplasma und Ureaplasma das kleinste Genom der zur Auto-Replikation befähigten Prokaryonten [...]
Ihr Genom weist meist einen relativen niedrigen Guanin-Cytosin (GC) Gehalt auf und ihre Zellmembran enthält Cholesterin, das sonst nur bei Eukaryoten gefunden wird.

CRASHKURS: MYKOPLASMEN

Was sagt uns Wikipedia?

Mycoplasma, deutsch auch als Mykoplasmen bezeichnet sind eine Gattung sehr kleiner Bakterien aus der Klasse der Mollicutes. Im Gegensatz zu allen anderen Bakterien fehlt ihnen eine Zellwand. [...]
Mykoplasmen sind meist parasitär, intra- und extrazellulär lebende Bakterien, die beim Menschen und Wirbeltieren die Ursache für zahlreiche Krankheiten sind. [...]
Mit einer Größe von 580 –– 1.380 kbp haben die Gattungen Mycoplasma und Ureaplasma das kleinste Genom der zur Auto-Replikation befähigten Prokaryonten [...]
Ihr Genom weist meist einen relativen niedrigen Guanin-Cytosin (GC) Gehalt auf und ihre Zellmembran enthält Cholesterin, das sonst nur bei Eukaryoten gefunden wird.

Wir schauen heute genauer auf *Mycoplasma bovis*.

MYCOPLASMA BOVIS

Chen et al., Complete Genome Sequence of Mycoplasma bovis Strain 08M, Genome Announcements, 2017,

doi: 10.1128/genomeA.00324-17

Aus einer Publikation, 2017...

The complete M. bovis 08M genome contains a single circular chromosome of 1,016,753 bp with an overall G+C content of 29.27%.[...]
The coding density of the genome is 89.38%. The genome contained 62 copies of interspersed nuclear elements and 87 copies of tandem repeats.

MYCOPLASMA BOVIS

Chen et al., Complete Genome Sequence of Mycoplasma bovis Strain 08M, Genome Announcements, 2017,

doi: 10.1128/genomeA.00324-17

Aus einer Publikation, 2017...

The complete M. bovis 08M genome contains a single circular chromosome of 1,016,753 bp with an overall G+C content of 29.27%.[...]
The coding density of the genome is 89.38%. The genome contained 62 copies of interspersed nuclear elements and 87 copies of tandem repeats.

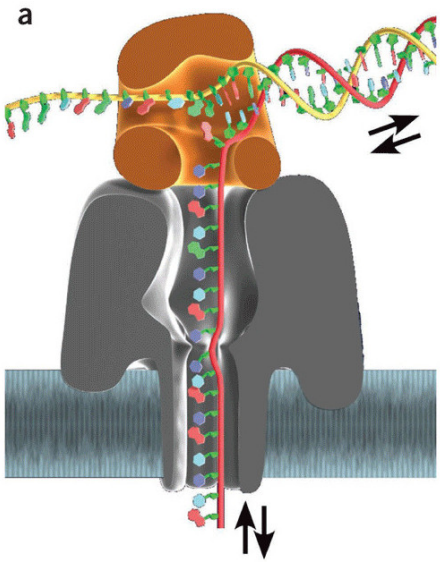
Stellen wir uns vor, wir haben Illumina short-reads für *M. bovis 08M*. Was für Probleme könnten bei der bioinformatischen Erstellung eines kompletten Genoms auftreten?

3rd Generation Sequencing – Oxford Nanopore Technologies

PROMO-VIDEO!

https://www.youtube.com/watch?v=E9-Rm5AoZGw

Zusammenfassung: Nanopore Schema



MinION von ONT

Vor- und Nachteile

Vorteile:

- Plattform limitiert nicht die read Länge
- Echtzeit-Sequenzierung
- in development: Selektive Sequenzierung
- anpassbar (Protokolle, Adaptoren, etc.)
- DNA und RNA Sequenzierung
- Modifikationen von Nukleotiden messbar

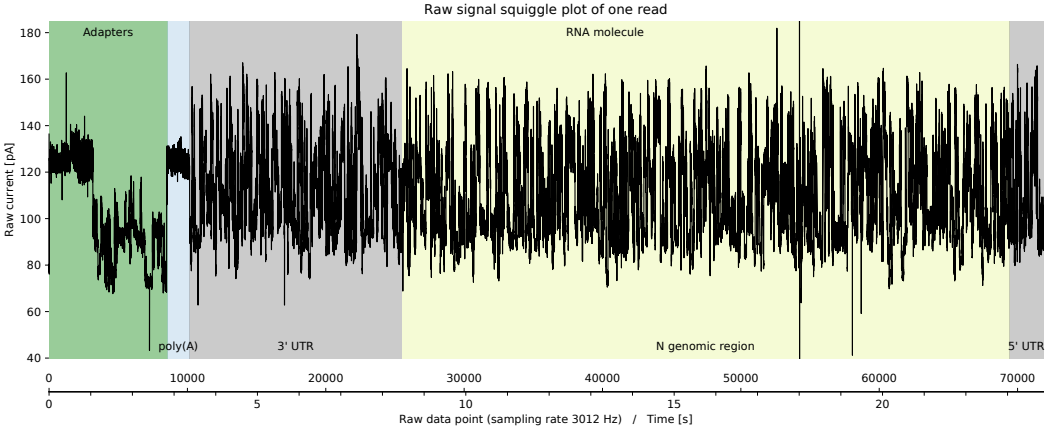
Nachteile:

- geringerer Durchsatz als Illumina
- Entwicklungsphase
- hohe Fehlerrate
  - DNA: $\sim$ 6%
  - RNA: $\sim$ 10%
- Protokolle müssen i.d.R. optimiert werden

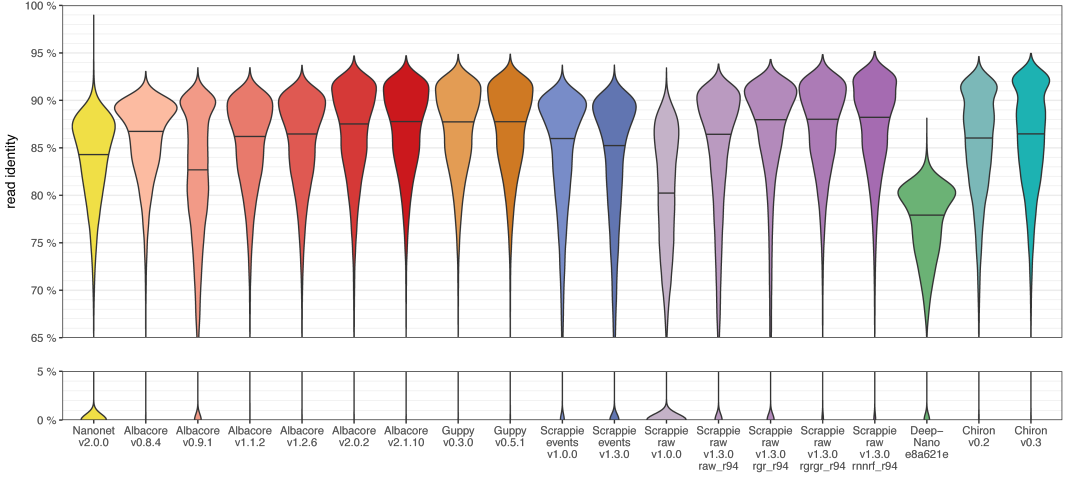
High-End Labore

Bioinformatische Anwendungen und Herausforderungen mit ONT

BASECALLING – ROHSIGNAL



Basecalling – Vom Signal zum Nukleotid



https://github.com/rrwick/Basecalling-comparison

PREPROCESSING

Wir haben jetzt 2 Möglichkeiten mit unseren Daten umzugehen:

Sequence-based

Signal-based

PREPROCESSING

Wir haben jetzt 2 Möglichkeiten mit unseren Daten umzugehen:

Sequence-based

https://github.com/rrwick/Porechop

Signal-based

PREPROCESSING

Wir haben jetzt 2 Möglichkeiten mit unseren Daten umzugehen:

Sequence-based

https://github.com/rrwick/Porechop

Signal-based

https://github.com/rrwick/Deepbinner

ASSEMBLIERUNG MIT LONG-READS

`https://github.com/rrwick/Unicycler`

- basiert auf SPAdes
- wurde exklusiv für bakterielle Isolate entwickelt
- ermöglicht auch hybrid-assemblies

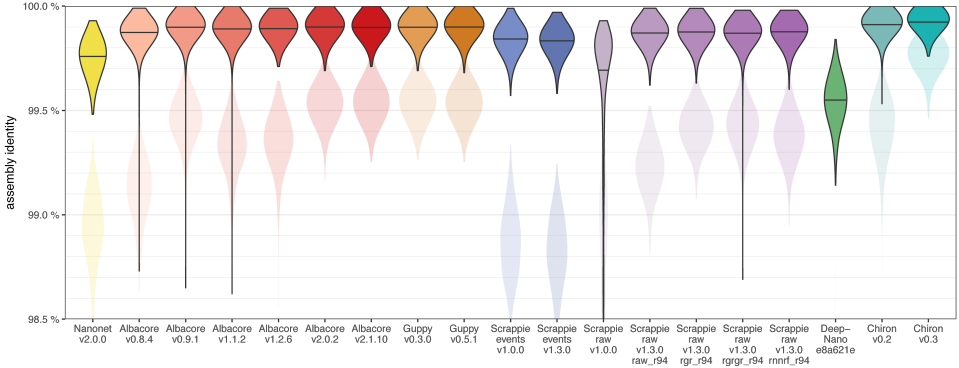
- Alternativen: `Canu`, `metaSPAdes`, `Flye`

SCHNELLES MAPPING

- `minimap2` ist der Mapping Standard
- sehr schnell, aber dafür ungenau

`https://github.com/lh3/minimap2`

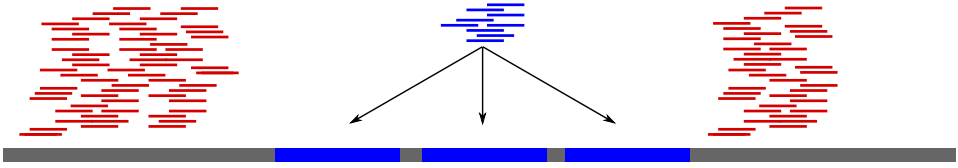
QUALITÄTSVERBESSERUNG MIT NANOPOLISH



https://github.com/rrwick/Basecalling-comparison

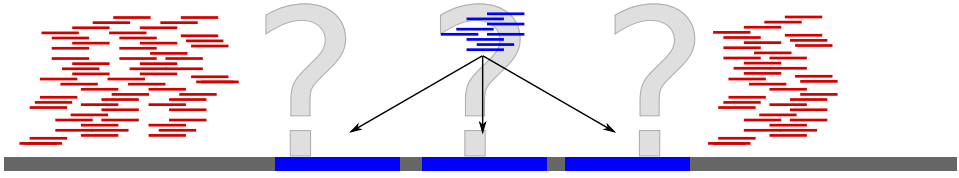
Illumina und Nanopore kombinieren

Idee: Mit short-read Daten die long-read Daten korrigieren:



Illumina und Nanopore kombinieren

Idee: Mit short-read Daten die long-read Daten korrigieren:



Illumina und Nanopore kombinieren

Idee: Mit short-read Daten die long-read Daten korrigieren:

Modifikationen messen und charakterisieren

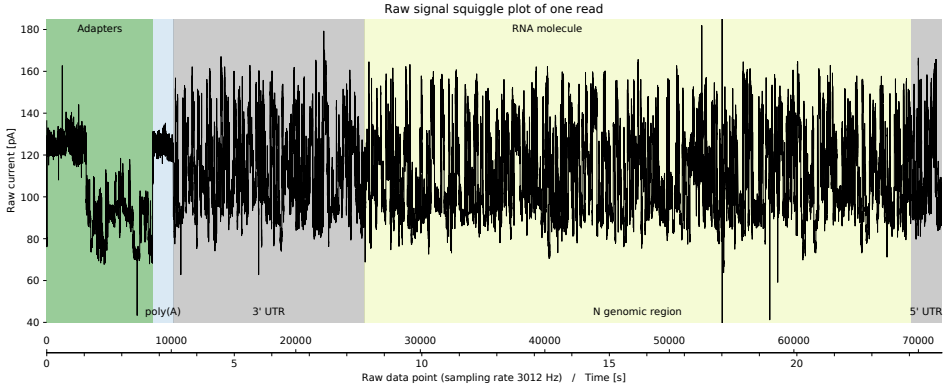
TOMBO

- Tombo is a suite of tools primarily for the identification of modified nucleotides from nanopore sequencing data.

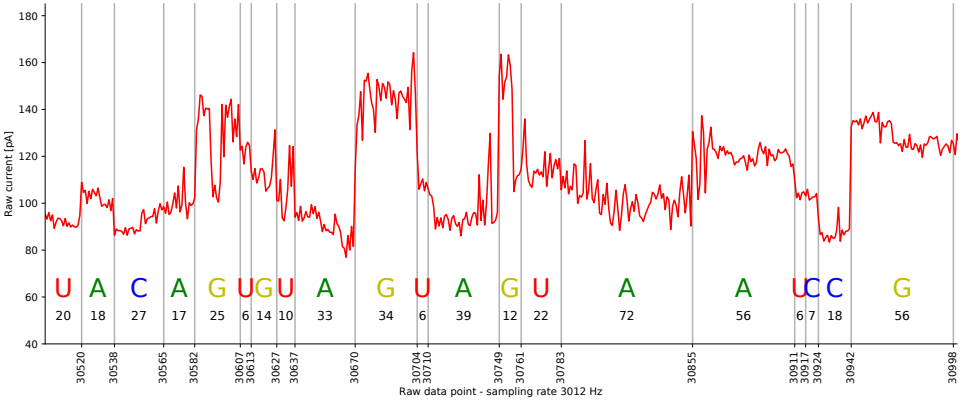
TOMBO

- Tombo is a suite of tools primarily for the identification of modified nucleotides from nanopore sequencing data.
- Tombo also provides tools for the analysis and visualization of raw nanopore signal.

RAW SIGNAL


Raw signal squiggle plot of one read
Adapters
poly(A)
3' UTR
RNA molecule
N genomic region
5' UTR
Raw current [pA]
180
160
140
120
100
80
60
40
0
10000
20000
30000
40000
50000
60000
70000
0
5
10
15
20
Raw data point (sampling rate 3012 Hz) / Time [s]

Events


Raw current [pA]
180
160
140
120
100
80
60
40
U A C A G U G U A G U A G U A A U C C G
20 18 27 17 25 6 14 10 33 34 6 39 12 22 72 56 6 7 18 56
30520 30538 30565 30582 30607 30613 30627 30637 30670 30704 30710 30749 30761 30783 30855 30911 30917 30924 30942 30998
Raw data point - sampling rate 3012 Hz

TOMBO COMMANDS

- Re-squiggle (Raw Signal Genomic Alignment)

TOMBO COMMANDS

- Re-squiggle (Raw Signal Genomic Alignment)
- Modified Base Detection

TOMBO COMMANDS

- Re-squiggle (Raw Signal Genomic Alignment)
- Modified Base Detection
- Text Output, Plotting, Read filtering

RE-SQUIGGLE (RAW SIGNAL GENOMIC ALIGNMENT)

- Aligns raw signal to genomic/transcriptomic sequence

RE-SQUIGGLE (RAW SIGNAL GENOMIC ALIGNMENT)

- Aligns raw signal to genomic/transcriptomic sequence
- Assumes provided reference sequence is correct

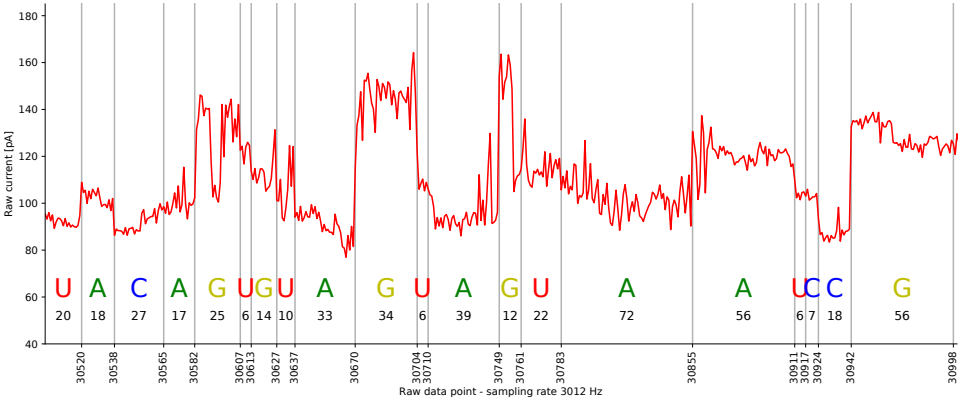
RE-SQUIGGLE (RAW SIGNAL GENOMIC ALIGNMENT)

- Aligns raw signal to genomic/transcriptomic sequence
- Assumes provided reference sequence is correct
- Adds mapped genomic location and the raw signal to sequence assignment to FAST5 files

Re-squiggle (Raw Signal Genomic Alignment)

- Aligns raw signal to genomic/transcriptomic sequence
- Assumes provided reference sequence is correct
- Adds mapped genomic location and the raw signal to sequence assignment to FAST5 files
- Builds index for fast access

EVENTS


Raw current [pA]
180
160
140
120
100
80
60
40
U A C A G U G U A G U A G U A U C C G
20 18 27 17 25 6 14 10 33 34 6 39 12 22 72 56 6 7 18 56
30520 30538 30565 30582 30607 30613 30627 30637 30670 30704 30710 30749 30761 30783 30855 30911 30917 30924 30942 30998
Raw data point - sampling rate 3012 Hz

Modified Base Detection

- Three methods to find modified bases:

Modified Base Detection

- Three methods to find modified bases:
  - Alternative model
  - *De novo*
  - Sample compare

Modified Base Detection

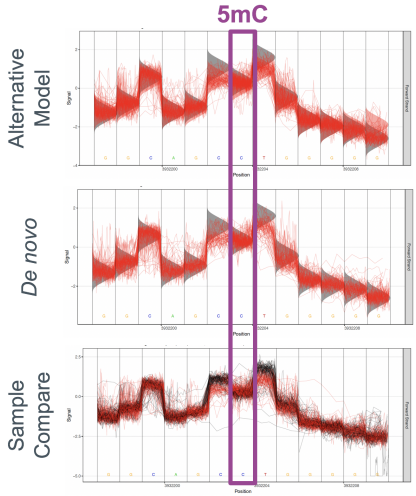
- Three methods to find modified bases:
  - Alternative model
  - *De novo*
  - Sample compare
- DNA and RNA

METHODS

MODIFIED BASE DETECTION

Tombo provides three methods for the identification of non-standard bases

	Advantages	Disadvantages
Alternative Model	• Known alt. base • Exact alt. loc. • Good accuracy	• Requires alt. model estimation
De novo	• Apply to any sample	• High error rate • Inexact location • Alt. base unknown
Sample Compare	• Best AUC • Most robust	• Inexact location • Alternative base unknown



https://nanoporetech.github.io/tombo/_images/testing_method_comparison.png

Specific Alternative Base Method

Recommended:

- Specific Alternative Base Method

Specific Alternative Base Method

Recommended:

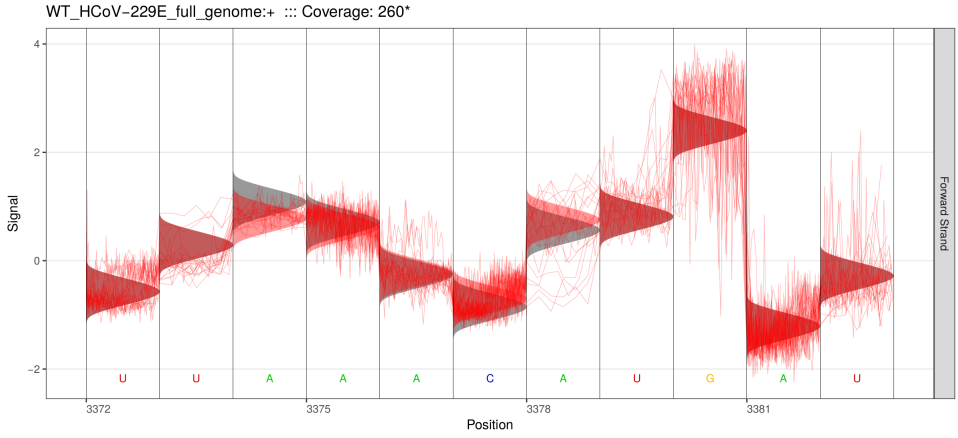
- Specific Alternative Base Method
  - Canonical model compared to:
  - DNA: 5-methylcytosine (5mC) and N6-methyladenosine (6mA)
  - RNA: only 5mC

Specific Alternative Base Method

Recommended:

- Specific Alternative Base Method
  - Canonical model compared to:
  - DNA: 5-methylcytosine (5mC) and N6-methyladenosine (6mA)
  - RNA: only 5mC
- Produces a statistic similar to a log likelihood ratio

Specific Alternative Base Method



TEXT OUTPUT

- Genome Browser File Output

TEXT OUTPUT

- Genome Browser File Output
  - wiggle
  - bedgraph

Text Output

- Genome Browser File Output
  - wiggle
  - bedgraph
- Genome Sequence Output

Text Output

- Genome Browser File Output
  - wiggle
  - bedgraph
- Genome Sequence Output
  - Sequence around locations with many modified reads

Plotting Output

- Genome Anchored Plotting

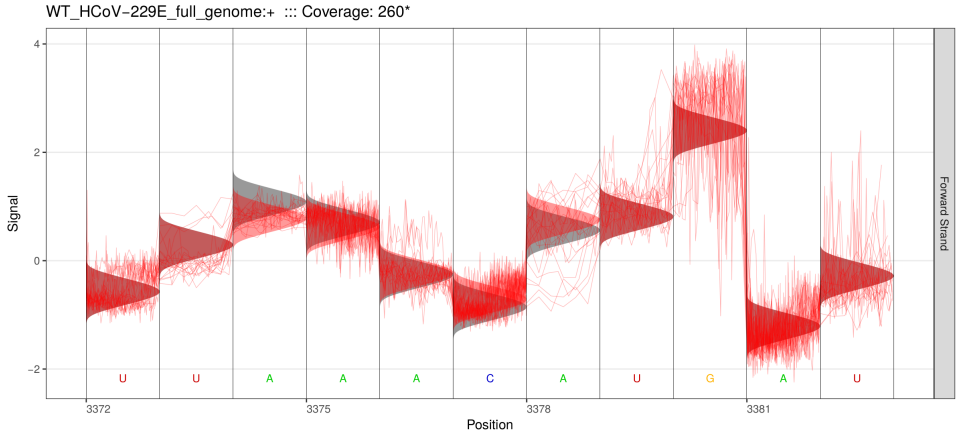
Plotting Output

- Genome Anchored Plotting
  - Specific position
  - Most significant

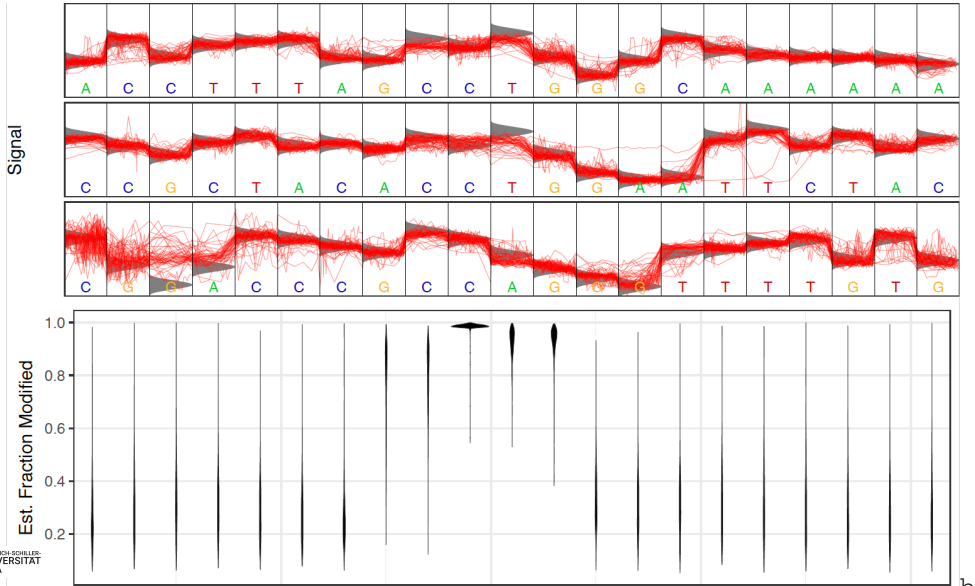
Plotting Output

- Genome Anchored Plotting
  - Specific position
  - Most significant
- Motif-centered Statistic Plotting

Genome Anchored Plotting



Motif-centered Statistic Plotting



DOCUMENTATION

https://nanoporetech.github.io/tombo/examples.html