A complete bacterial genome assembled de novo using only nanopore sequencing data

Nicholas J Loman, Joshua Quick, Jared T Simpson

Yunfan Fan
Johns Hopkins Genomics Journal Club
July 20, 2017

Long Read Sequencing

Oxford Nanopore (this paper), Pacific Biosciences

- Uses:
  - Span large variations
  - Resolve tricky regions (repeats)
  - Speed, portability (ONT)
  - Direct base modification, RNA detection (ONT)

- Weaknesses
  - Error rate (ONT R7.3 chemistry: 15 - 20%)
  - Cost/Yield

Nanopore Data

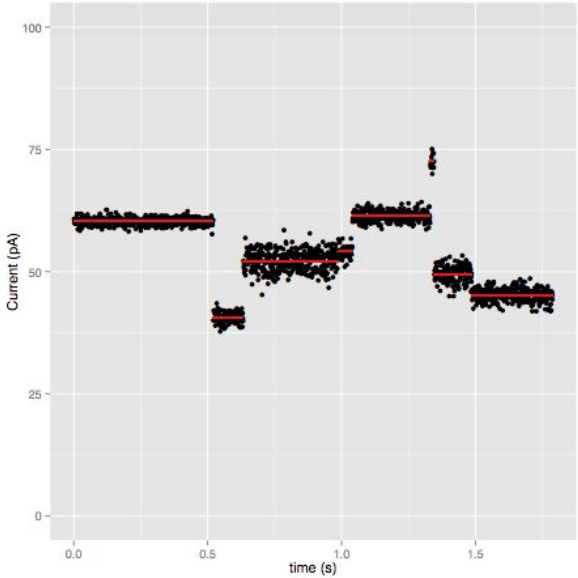


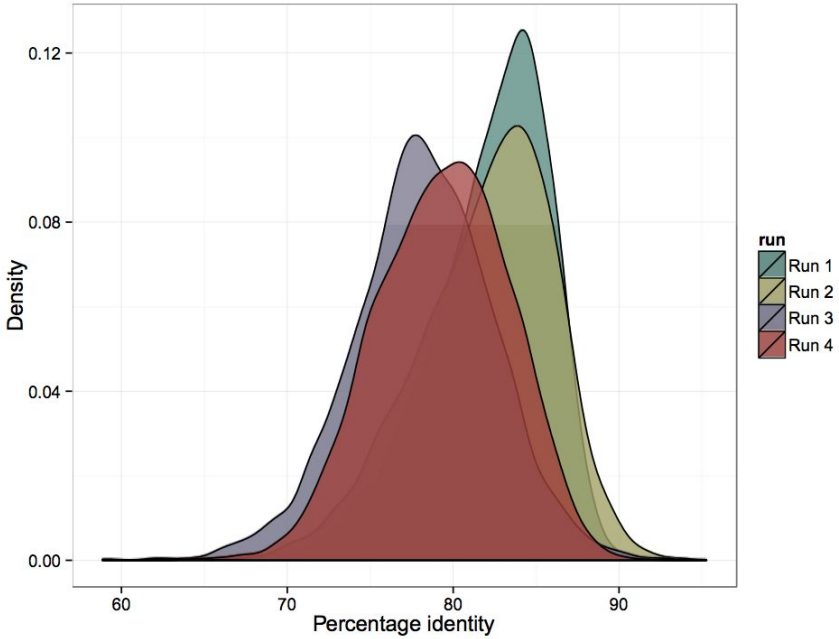
Figure S1: Simulated ideal signal data to illustrate the data that is input into our model. The black points are sampled current levels at a given time. The red lines are the events detected by a feature detection algorithm, which partitions the samples into discrete segments.

- Raw data
  - Ionic current readings as DNA passes through a protein pore
    - 5 bases occupy pore at once (now 6 with R9)
  - 4 kHz sampling rate
- Sequence data
  - Current readings segmented into events (edge detection)
    - Mean, standard deviation, duration
  - Translate into 5-mers
    - Metrichor (RNN, previously HMM)

Assembly of E. *coli* - Data

Four MinION runs:

- E. coli K-12, a genome of 4.6 Mb
- 22,270 reads
- 133.6 Mb ~ 29x coverage



Assembly of E. *coli* - Outline

Primary Assembly

- Compute Overlaps (Daligner)
- Multiple Sequence Alignment (POA)
- Celera Assembler

Polish

- Map reads to primary assembly (bwa-mem)
- Segment and anchor
- Compute consensus

Primary Assembly

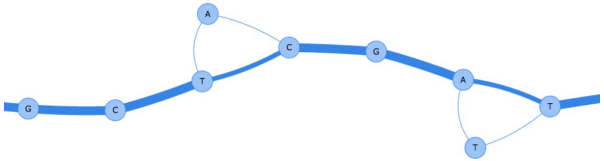
Compute read overlaps with Daligner

Consensus and correction with POA

Read GCTACGAT

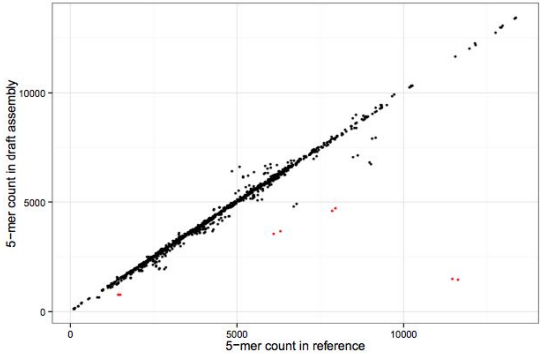
Add overlapping sequence GCTCGAT

Add overlapping sequence GCTCGATT



Assembly with Celera

Draft: 98.5% accuracy



Figures from Jared Simpson (LC 2015)

Polish - Outline

Map reads (bwa-mem with -x ont2d)

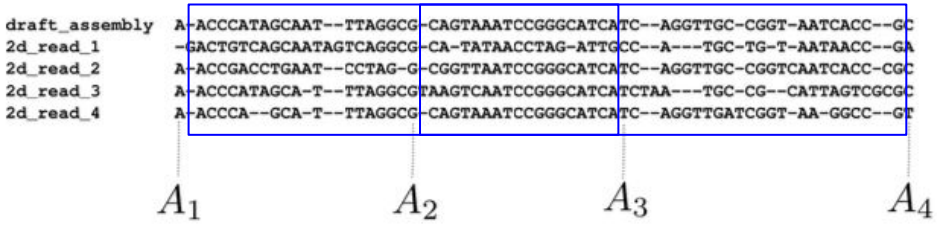
Segment and anchor

Mutate and evaluate

Probabilistic Model:

- Profile HMM
- Train Transitions

Segment and Anchor



Split alignment into 10kb segments that overlap by 200bp

Assign anchor points every 50bp - define a mapping between draft assembly and events in reads

Compute consensus for each segment between two anchors

Segment and Anchor

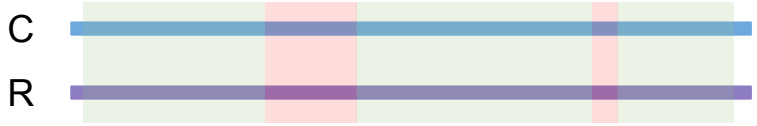
Split alignment into 10kb segments that overlap by 200bp

Assign anchor points every 50bp - define a mapping between draft assembly and events in reads

Compute consensus for each segment between two anchors

Mutate and Evaluate

Block replacement method



Add to set of candidates *C*:

Choose best candidate from *C*, repeat until convergence

Mutation method

Candidates are all sequences with edit distance 1 from current consensus

Choose best candidate to be new C, repeat until convergence

Update 10kb segment

Probabilistic Model

Choose best candidate:

$$C' = \underset{S \in \mathcal{C}}{\operatorname{argmax}}\, P(\mathcal{D}|S)$$

$$P(\mathcal{D}|S) = \prod_{k=1}^{r} P(e_{i,k}, e_{i+1,k}, \ldots, e_{j,k}|S, \Theta)$$

Forward Algorithm to compute over all paths in Profile HMM

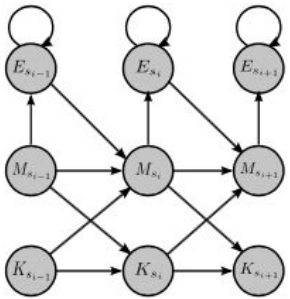


Figure S4: The state structure of the Profile Hidden Markov Model we use to calculate the probability of a sequence of events given a known sequence.

E: 5-mer has emitted an extra event
K: 5-mer has not emitted an event
M: 5-mer has emitted an event

Probabilistic Model - Transition Probabilities

$$t(M_{s_{i-1}} \to K_{s_i}) = f(|\mu_{s_{i-1}} - \mu_{s_i}|)$$
$$t(M_{s_{i-1}} \to E_{s_{i-1}}) = p_{me}(1 - t(M_{s_{i-1}} \to K_{s_i}))$$
$$t(M_{s_{i-1}} \to M_{s_i}) = 1 - t(M_{s_{i-1}} \to K_{s_i}) - t(M_{s_{i-1}} \to E_{s_{i-1}})$$
$$t(E_{s_{i-1}} \to E_{s_{i-1}}) = p_{ee}$$
$$t(E_{s_{i-1}} \to M_{s_i}) = 1 - t(E_{s_{i-1}} \to E_{s_{i-1}})$$
$$t(K_{s_{i-1}} \to K_{s_i}) = f(|\mu_{s_{i-1}} - \mu_{s_i}|)$$
$$t(K_{s_{i-1}} \to M_{s_i}) = 1 - t(K_{s_{i-1}} \to K_{s_i})$$

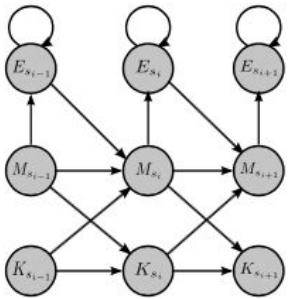


Figure S4: The state structure of the Profile Hidden Markov Model we use to calculate the probability of a sequence of events given a known sequence.

E: 5-mer has emitted an extra event
K: 5-mer has not emitted an event
M: 5-mer has emitted an event

Probabilistic Model - Transition Probabilities

$$t(M_{s_{i-1}} \rightarrow K_{s_i}) = f(|\mu_{s_{i-1}} - \mu_{s_i}|)$$
$$t(M_{s_{i-1}} \rightarrow E_{s_{i-1}}) = p_{me}(1 - t(M_{s_{i-1}} \rightarrow K_{s_i}))$$
$$t(M_{s_{i-1}} \rightarrow M_{s_i}) = 1 - t(M_{s_{i-1}} \rightarrow K_{s_i}) - t(M_{s_{i-1}} \rightarrow E_{s_{i-1}})$$
$$t(E_{s_{i-1}} \rightarrow E_{s_{i-1}}) = p_{ee}$$
$$t(E_{s_{i-1}} \rightarrow M_{s_i}) = 1 - t(E_{s_{i-1}} \rightarrow E_{s_{i-1}})$$
$$t(K_{s_{i-1}} \rightarrow K_{s_i}) = f(|\mu_{s_{i-1}} - \mu_{s_i}|)$$
$$t(K_{s_{i-1}} \rightarrow M_{s_i}) = 1 - t(K_{s_{i-1}} \rightarrow K_{s_i})$$

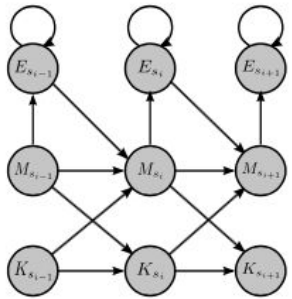


Figure S4: The state structure of the Profile Hidden Markov Model we use to calculate the probability of a sequence of events given a known sequence.

Function (lookup table) of expected event means. Calculated from alignment of reads to previous C using Viterbi, counting the types of transitions.

E: 5-mer has emitted an extra event
K: 5-mer has not emitted an event
M: 5-mer has emitted an event

Probabilistic Model - Transition Probabilities

$$t(M_{s_{i-1}} \to K_{s_i}) = f(|\mu_{s_{i-1}} - \mu_{s_i}|)$$
$$\boxed{t(M_{s_{i-1}} \to E_{s_{i-1}}) = p_{me}(1 - t(M_{s_{i-1}} \to K_{s_i}))}$$
$$t(M_{s_{i-1}} \to M_{s_i}) = 1 - t(M_{s_{i-1}} \to K_{s_i}) - t(M_{s_{i-1}} \to E_{s_{i-1}})$$
$$t(E_{s_{i-1}} \to E_{s_{i-1}}) = p_{ee}$$
$$t(E_{s_{i-1}} \to M_{s_i}) = 1 - t(E_{s_{i-1}} \to E_{s_{i-1}})$$
$$t(K_{s_{i-1}} \to K_{s_i}) = f(|\mu_{s_{i-1}} - \mu_{s_i}|)$$
$$t(K_{s_{i-1}} \to M_{s_i}) = 1 - t(K_{s_{i-1}} \to K_{s_i})$$

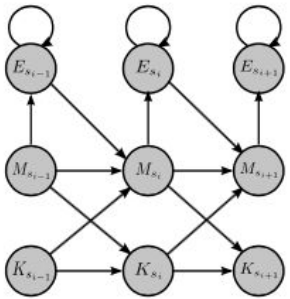


Figure S4: The state structure of the Profile Hidden Markov Model we use to calculate the probability of a sequence of events given a known sequence.

Leverage knowledge that you haven't gone to a K state, and scale.

E: 5-mer has emitted an extra event
K: 5-mer has not emitted an event
M: 5-mer has emitted an event

Probabilistic Model - Transition Probabilities

$$t(M_{s_{i-1}} \to K_{s_i}) = f(|\mu_{s_{i-1}} - \mu_{s_i}|)$$
$$t(M_{s_{i-1}} \to E_{s_{i-1}}) = p_{me}(1 - t(M_{s_{i-1}} \to K_{s_i}))$$
$$\boxed{t(M_{s_{i-1}} \to M_{s_i}) = 1 - t(M_{s_{i-1}} \to K_{s_i}) - t(M_{s_{i-1}} \to E_{s_{i-1}})}$$
$$t(E_{s_{i-1}} \to E_{s_{i-1}}) = p_{ee}$$
$$t(E_{s_{i-1}} \to M_{s_i}) = 1 - t(E_{s_{i-1}} \to E_{s_{i-1}})$$
$$t(K_{s_{i-1}} \to K_{s_i}) = f(|\mu_{s_{i-1}} - \mu_{s_i}|)$$
$$t(K_{s_{i-1}} \to M_{s_i}) = 1 - t(K_{s_{i-1}} \to K_{s_i})$$

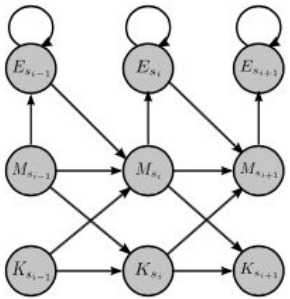


Figure S4: The state structure of the Profile Hidden Markov Model we use to calculate the probability of a sequence of events given a known sequence.

Not K state and not E state

E: 5-mer has emitted an extra event
K: 5-mer has not emitted an event
M: 5-mer has emitted an event

Probabilistic Model - Transition Probabilities

$$t(M_{s_{i-1}} \to K_{s_i}) = f(|\mu_{s_{i-1}} - \mu_{s_i}|)$$
$$t(M_{s_{i-1}} \to E_{s_{i-1}}) = p_{me}(1 - t(M_{s_{i-1}} \to K_{s_i}))$$
$$t(M_{s_{i-1}} \to M_{s_i}) = 1 - t(M_{s_{i-1}} \to K_{s_i}) - t(M_{s_{i-1}} \to E_{s_{i-1}})$$
$$\boxed{t(E_{s_{i-1}} \to E_{s_{i-1}}) = p_{ee}}$$
$$t(E_{s_{i-1}} \to M_{s_i}) = 1 - t(E_{s_{i-1}} \to E_{s_{i-1}})$$
$$t(K_{s_{i-1}} \to K_{s_i}) = f(|\mu_{s_{i-1}} - \mu_{s_i}|)$$
$$t(K_{s_{i-1}} \to M_{s_i}) = 1 - t(K_{s_{i-1}} \to K_{s_i})$$

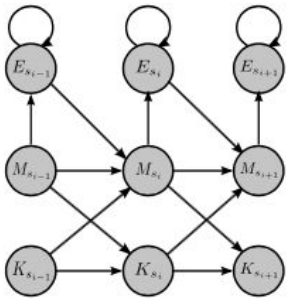


Figure S4: The state structure of the Profile Hidden Markov Model we use to calculate the probability of a sequence of events given a known sequence.

Trained from sample of reads (Viterbi to previous C)

E: 5-mer has emitted an extra event
K: 5-mer has not emitted an event
M: 5-mer has emitted an event

Probabilistic Model - Transition Probabilities

$$t(M_{s_{i-1}} \to K_{s_i}) = f(|\mu_{s_{i-1}} - \mu_{s_i}|)$$
$$t(M_{s_{i-1}} \to E_{s_{i-1}}) = p_{me}(1 - t(M_{s_{i-1}} \to K_{s_i}))$$
$$t(M_{s_{i-1}} \to M_{s_i}) = 1 - t(M_{s_{i-1}} \to K_{s_i}) - t(M_{s_{i-1}} \to E_{s_{i-1}})$$
$$t(E_{s_{i-1}} \to E_{s_{i-1}}) = p_{ee}$$
$$\boxed{t(E_{s_{i-1}} \to M_{s_i}) = 1 - t(E_{s_{i-1}} \to E_{s_{i-1}})}$$
$$t(K_{s_{i-1}} \to K_{s_i}) = f(|\mu_{s_{i-1}} - \mu_{s_i}|)$$
$$t(K_{s_{i-1}} \to M_{s_i}) = 1 - t(K_{s_{i-1}} \to K_{s_i})$$

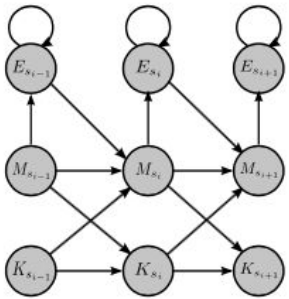


Figure S4: The state structure of the Profile Hidden Markov Model we use to calculate the probability of a sequence of events given a known sequence.

Did not stay in E state

E: 5-mer has emitted an extra event
K: 5-mer has not emitted an event
M: 5-mer has emitted an event

Probabilistic Model - Transition Probabilities

$$t(M_{s_{i-1}} \to K_{s_i}) = f(|\mu_{s_{i-1}} - \mu_{s_i}|)$$
$$t(M_{s_{i-1}} \to E_{s_{i-1}}) = p_{me}(1 - t(M_{s_{i-1}} \to K_{s_i}))$$
$$t(M_{s_{i-1}} \to M_{s_i}) = 1 - t(M_{s_{i-1}} \to K_{s_i}) - t(M_{s_{i-1}} \to E_{s_{i-1}})$$
$$t(E_{s_{i-1}} \to E_{s_{i-1}}) = p_{ee}$$
$$t(E_{s_{i-1}} \to M_{s_i}) = 1 - t(E_{s_{i-1}} \to E_{s_{i-1}})$$
$$\boxed{t(K_{s_{i-1}} \to K_{s_i}) = f(|\mu_{s_{i-1}} - \mu_{s_i}|)}$$
$$t(K_{s_{i-1}} \to M_{s_i}) = 1 - t(K_{s_{i-1}} \to K_{s_i})$$

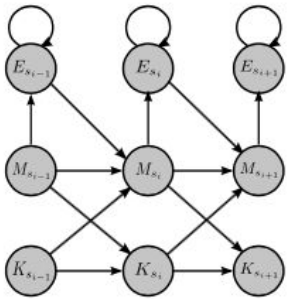


Figure S4: The state structure of the Profile Hidden Markov Model we use to calculate the probability of a sequence of events given a known sequence.

Like before, not emitting an event is a function of expected current levels

E: 5-mer has emitted an extra event
K: 5-mer has not emitted an event
M: 5-mer has emitted an event

Probabilistic Model - Transition Probabilities

$$t(M_{s_{i-1}} \to K_{s_i}) = f(|\mu_{s_{i-1}} - \mu_{s_i}|)$$
$$t(M_{s_{i-1}} \to E_{s_{i-1}}) = p_{me}(1 - t(M_{s_{i-1}} \to K_{s_i}))$$
$$t(M_{s_{i-1}} \to M_{s_i}) = 1 - t(M_{s_{i-1}} \to K_{s_i}) - t(M_{s_{i-1}} \to E_{s_{i-1}})$$
$$t(E_{s_{i-1}} \to E_{s_{i-1}}) = p_{ee}$$
$$t(E_{s_{i-1}} \to M_{s_i}) = 1 - t(E_{s_{i-1}} \to E_{s_{i-1}})$$
$$t(K_{s_{i-1}} \to K_{s_i}) = f(|\mu_{s_{i-1}} - \mu_{s_i}|)$$
$$\boxed{t(K_{s_{i-1}} \to M_{s_i}) = 1 - t(K_{s_{i-1}} \to K_{s_i})}$$

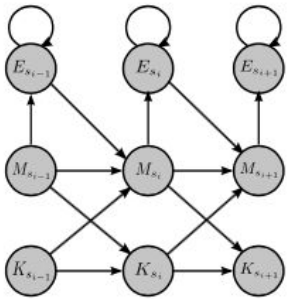


Figure S4: The state structure of the Profile Hidden Markov Model we use to calculate the probability of a sequence of events given a known sequence.

Has not gone to another K state

E: 5-mer has emitted an extra event
K: 5-mer has not emitted an event
M: 5-mer has emitted an event

Probabilistic Model - Emission Probabilities

$$P(e_i|\pi_k = (i, M_{s_j})) = \mathcal{N}(\mu_{s_j}, \sigma^2_{s_j})$$
$$P(e_i|\pi_k = (i, E_{s_j})) = \mathcal{N}(\mu_{s_j}, (v\sigma_{s_j})^2)$$

Emission distributions from M state: known

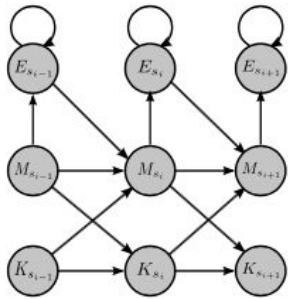


Figure S4: The state structure of the Profile Hidden Markov Model we use to calculate the probability of a sequence of events given a known sequence.

E: 5-mer has emitted an extra event
K: 5-mer has not emitted an event
M: 5-mer has emitted an event

Probabilistic Model - Emission Probabilities

$$P(e_i|\pi_k = (i, M_{s_j})) = \mathcal{N}(\mu_{s_j}, \sigma^2_{s_j})$$
$$P(e_i|\pi_k = (i, E_{s_j})) = \mathcal{N}(\mu_{s_j}, (v\sigma_{s_j})^2)$$

Emission distributions from E state: same as distribution from M, but variance is scaled by v=1.75

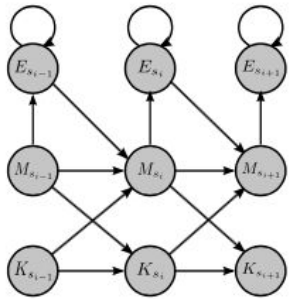
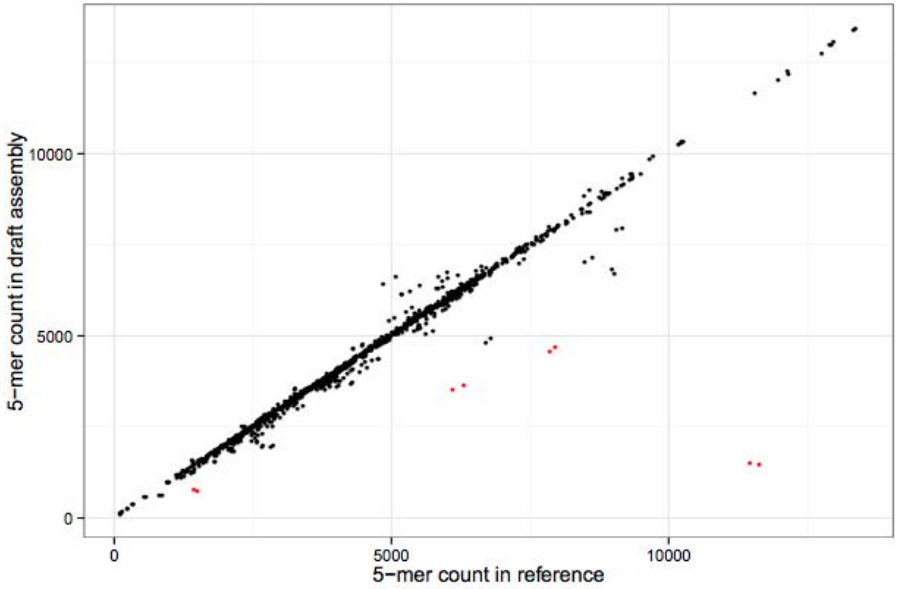


Figure S4: The state structure of the Profile Hidden Markov Model we use to calculate the probability of a sequence of events given a known sequence.

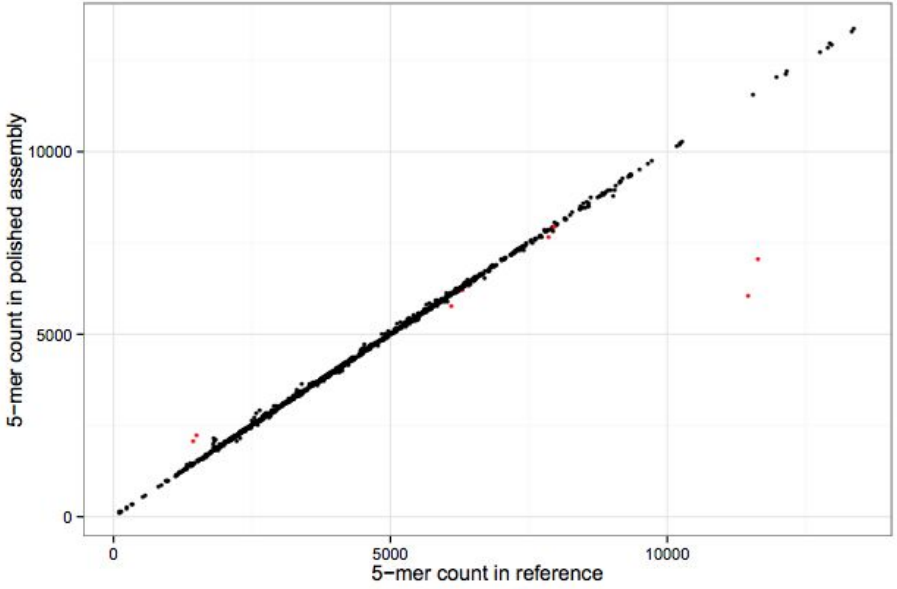
E: 5-mer has emitted an extra event
K: 5-mer has not emitted an event
M: 5-mer has emitted an event

Results

Draft: 98.5% accuracy

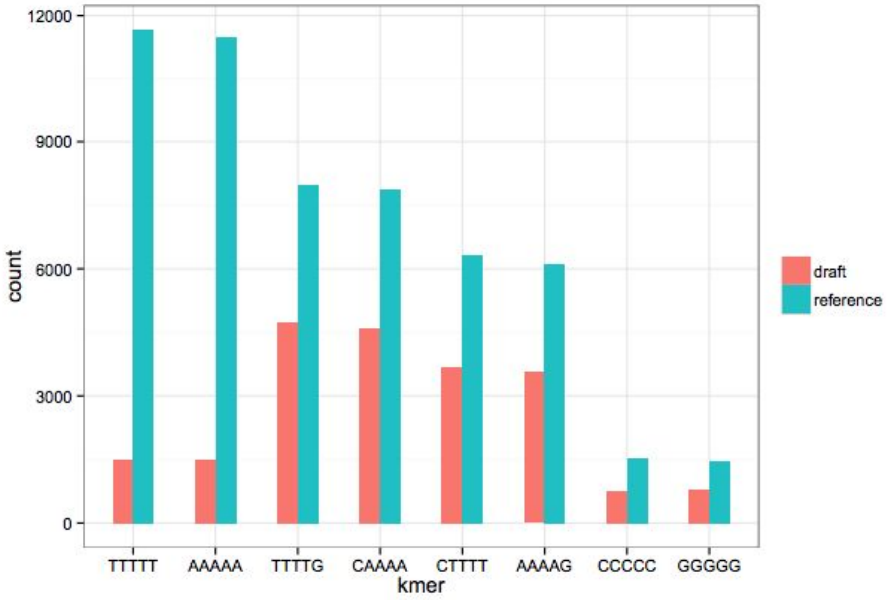


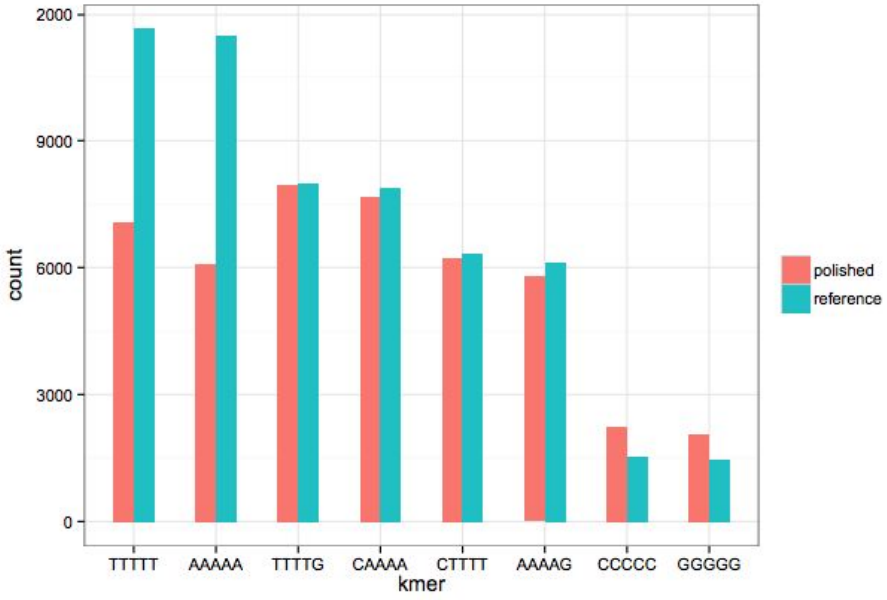
Polished: 99.5% accuracy



Figures from Jared Simpson (LC 2015)

Results


12000
9000
6000
3000
0
count
TTTTT
AAAAA
TTTTG
CAAAA
CTTTT
AAAAG
CCCCC
GGGGG
kmer
draft
reference


2000
9000
6000
3000
0
count
TTTTT
AAAAA
TTTTG
CAAAA
CTTTT
AAAAG
CCCCC
GGGGG
kmer
polished
reference