

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

Nicholas J Loman, Joshua Quick, Jared T Simpson

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

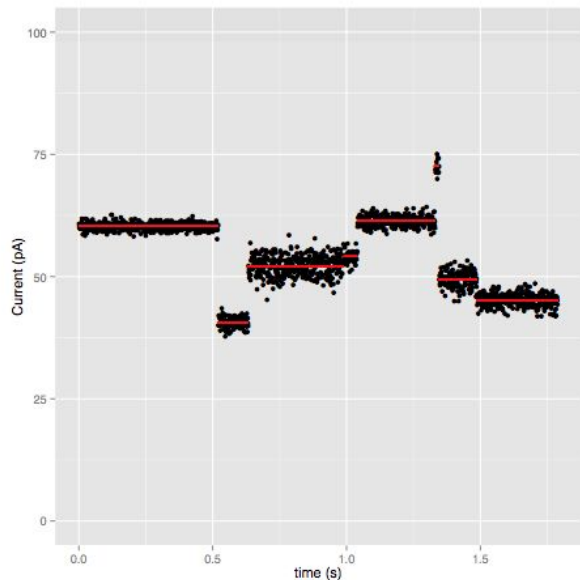


Figure S1: Simulated ideal signal data to illustrate the data that is input into our model. The black points are sampled current levels at a given time. The red lines are the events detected by a feature detection algorithm, which partitions the samples into discrete segments.



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



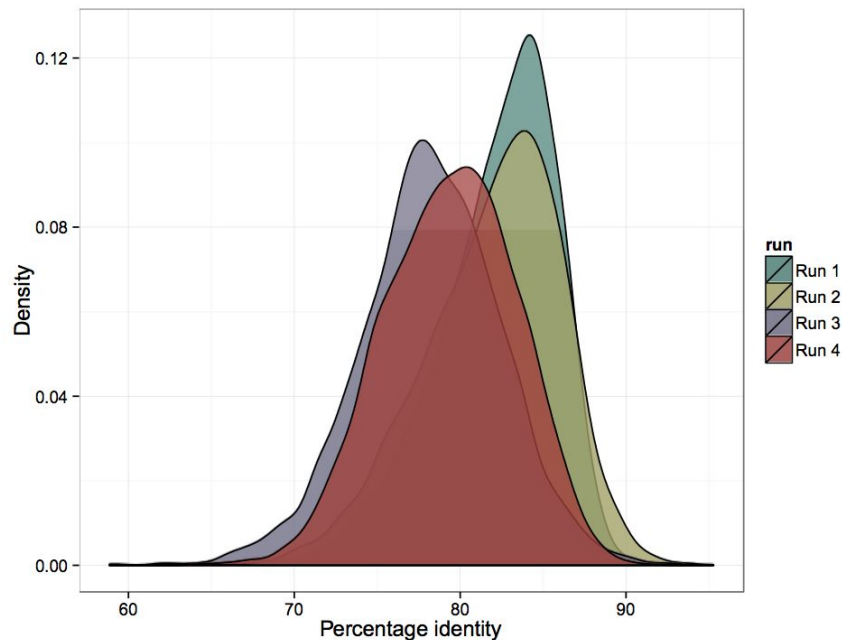
Sequence data

- 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

```
2890 --GCCAGAGTCA-AT-GC--TTCCACGCCGGGTTACCGCGGATAACCGCTAC--CCGTTACGTTAAA-CAGCG--C--TAGTCAGAGGCAATCCGGCAGTA
1952 TGCCAG-GT-AGATTGCGCTTCCACGCCGGAGTTACCGCGG-----CCGATACGCGGACGGCTTTAGTCATGAACCGTGGGAGTGGCCAAATGC-GCAGTG
2970 AAT--CGCCTTTGGTGGGATACTGATCTT--TCTAATAGACTGCTTTTCATGTGTT--GCCATTT--GCAGCGGTGCGCAC--GATAATG--CTGCGTGC-TT
2845 AGTACGCCCTTTGGTGGGATACTGATCTGTC--ACCCG--GAACGTTTCATGT--TGCGCCATTTTGC--ACCGGTGCGCACACAGATAATGATGCTGCGTGCCTT
3869 TCAGTACCGCGC--CAGAAGCTGTT--CACGTG-A-TGCAGACCACTTCAACTGCTGCTGGGATCAGTTAGCGGAC--C-CTGTCTAA--ATCACGCT--AAC
2141 TTAGTGGC--CACTCAGAGGCTGTTTCA--ACTCTCTCGAGAC--CTTCAACTGCTGCT--GGATCAGTT--GCTGGCCTCTGCGTGCATCA--CATCAC
3160 GTGCTATTTCATCAAGTGAACCT--CAGTGCCTGC--TTC--CTTCAG--TCTTCG--GTACAGAAATGATTTTCGATATGA--CGTA-C--CG-T
2235 GTCG-GTTTCATCAAGTGAACCTTTTCAGTGCCTGCCAGCTTCGCTTCAGGTCTTCTCAGTACAGAAATGATTTTCGATATCAAGGATATCGGGCGAT
3244 -T-ACCA-AA--CGTTGC--CAATCAGAAGTAATCG--ATGCTTTACGTGCGCTATAATTAGAT-C--GCTTAAGCATG--GGGC-GGAA--CCGACGA
2334 CTGACCAACAAACGTTCCACCATCAG-A-CTCAACGGATGCTTTACGTGCGGAATTAATGCTGCGCGCTCA--CCGCCGGGCGGAATCCGACGA
```

Consensus and
correction with POA

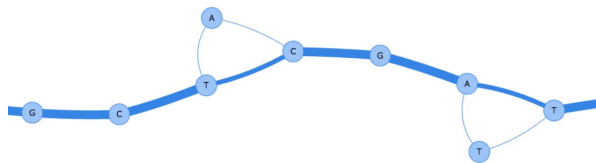
Read GCTACGAT



Add overlapping sequence GCTCGAT

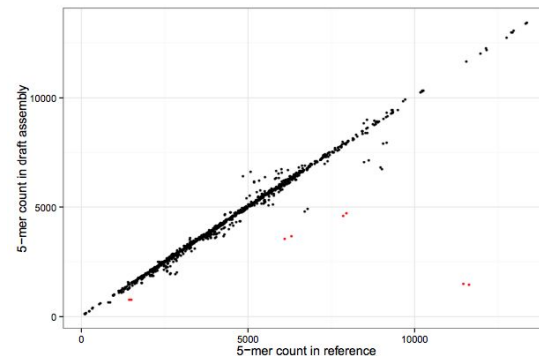


Add overlapping sequence GCTCGATT



Assembly with Celera

Draft: 98.5% accuracy



Polish - Outline

Map reads (bwa-mem with -x ont2d)

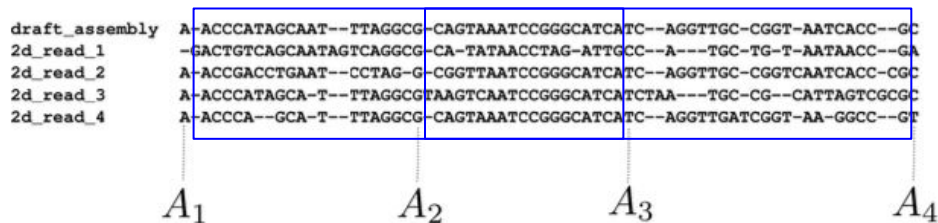
Segment and anchor

Mutate and evaluate

Probabilistic Model:

- ❖ Profile HMM
- ❖ Train Transitions

Segment and Anchor



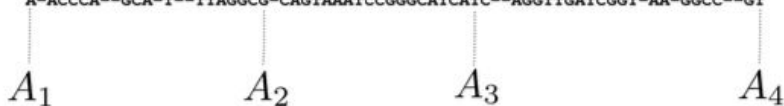
Split alignment into 10kb segments that overlap by 200bp

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

```
draft_assembly  A-ACCCATAGCAAT--TTAGGCG-CAGTAAATCCGGGCATCATC--AGGTTGC-CGGT-AATCACC--GC
2d_read_1       -GACTGTCAGCAATAGTCAGGCG-CA-TATAACCTAG-ATTGCC--A---TGC-TG-T-AATAACC--GA
2d_read_2       A-ACCGACCTGAAT--CCTAG-G-CGGTTAATCCGGGCATCATC--AGGTTGC-CGGTCAATCACC-CGC
2d_read_3       A-ACCCATAGCA-T--TTAGGCGTAAGTCAATCCGGGCATCATCTAA---TGC-CG--CATTAGTCGCGC
2d_read_4       A-ACCCA--GCA-T--TTAGGCG-CAGTAAATCCGGGCATCATC--AGGTTGATCGGT-AA-GGCC--GT
```



A_1 A_2 A_3 A_4

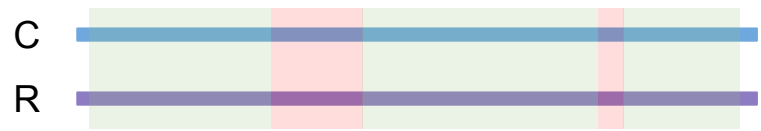
Split alignment into 10kb segments
that overlap by 200bp

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:

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

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

Forward Algorithm to compute over all paths in
Profile HMM

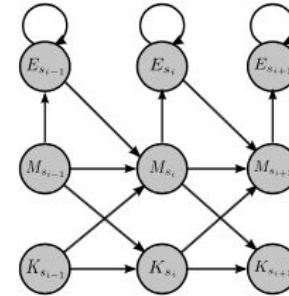


Figure S4: The state structure of the Profile Hidden Markov Model we use to calculate the probability of a sequence of events given a known sequence.

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

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

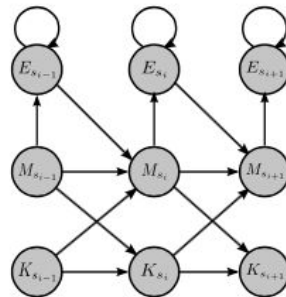


Figure S4: The state structure of the Profile Hidden Markov Model we use to calculate the probability of a sequence of events given a known sequence.

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

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

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

$$t(M_{s_{i-1}} \rightarrow M_{s_i}) = 1 - t(M_{s_{i-1}} \rightarrow K_{s_i}) - t(M_{s_{i-1}} \rightarrow E_{s_{i-1}})$$

$$t(E_{s_{i-1}} \rightarrow E_{s_{i-1}}) = p_{ee}$$

$$t(E_{s_{i-1}} \rightarrow M_{s_i}) = 1 - t(E_{s_{i-1}} \rightarrow E_{s_{i-1}})$$

$$t(K_{s_{i-1}} \rightarrow K_{s_i}) = f(|\mu_{s_{i-1}} - \mu_{s_i}|)$$

$$t(K_{s_{i-1}} \rightarrow M_{s_i}) = 1 - t(K_{s_{i-1}} \rightarrow K_{s_i})$$

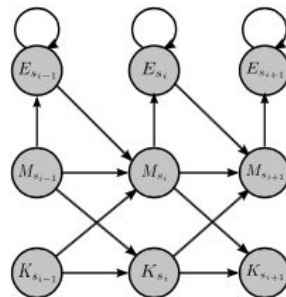


Figure S4: The state structure of the Profile Hidden Markov Model we use to calculate the probability of a sequence of events given a known sequence.

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

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

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

$$t(M_{s_{i-1}} \rightarrow M_{s_i}) = 1 - t(M_{s_{i-1}} \rightarrow K_{s_i}) - t(M_{s_{i-1}} \rightarrow E_{s_{i-1}})$$

$$t(E_{s_{i-1}} \rightarrow E_{s_{i-1}}) = p_{ee}$$

$$t(E_{s_{i-1}} \rightarrow M_{s_i}) = 1 - t(E_{s_{i-1}} \rightarrow E_{s_{i-1}})$$

$$t(K_{s_{i-1}} \rightarrow K_{s_i}) = f(|\mu_{s_{i-1}} - \mu_{s_i}|)$$

$$t(K_{s_{i-1}} \rightarrow M_{s_i}) = 1 - t(K_{s_{i-1}} \rightarrow K_{s_i})$$

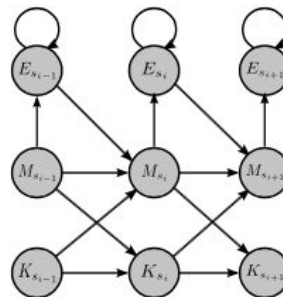


Figure S4: The state structure of the Profile Hidden Markov Model we use to calculate the probability of a sequence of events given a known sequence.

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

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

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

$$t(M_{s_{i-1}} \rightarrow M_{s_i}) = 1 - t(M_{s_{i-1}} \rightarrow K_{s_i}) - t(M_{s_{i-1}} \rightarrow E_{s_{i-1}})$$

$$t(E_{s_{i-1}} \rightarrow E_{s_{i-1}}) = p_{ee}$$

$$t(E_{s_{i-1}} \rightarrow M_{s_i}) = 1 - t(E_{s_{i-1}} \rightarrow E_{s_{i-1}})$$

$$t(K_{s_{i-1}} \rightarrow K_{s_i}) = f(|\mu_{s_{i-1}} - \mu_{s_i}|)$$

$$t(K_{s_{i-1}} \rightarrow M_{s_i}) = 1 - t(K_{s_{i-1}} \rightarrow K_{s_i})$$

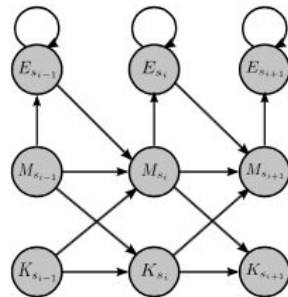


Figure S4: The state structure of the Profile Hidden Markov Model we use to calculate the probability of a sequence of events given a known sequence.

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

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

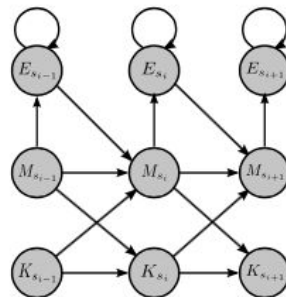


Figure S4: The state structure of the Profile Hidden Markov Model we use to calculate the probability of a sequence of events given a known sequence.

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

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

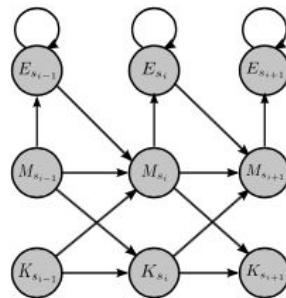


Figure S4: The state structure of the Profile Hidden Markov Model we use to calculate the probability of a sequence of events given a known sequence.

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

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

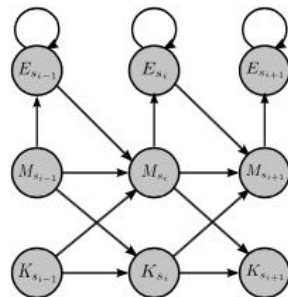


Figure S4: The state structure of the Profile Hidden Markov Model we use to calculate the probability of a sequence of events given a known sequence.

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

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

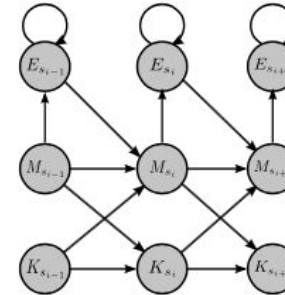


Figure S4: The state structure of the Profile Hidden Markov Model we use to calculate the probability of a sequence of events given a known sequence.

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

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

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

Emission distributions from M state:
known

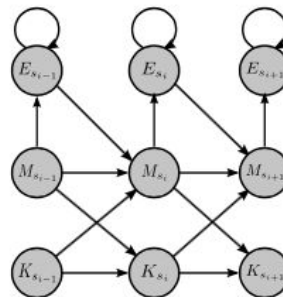


Figure S4: The state structure of the Profile Hidden Markov Model we use to calculate the probability of a sequence of events given a known sequence.

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

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

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

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

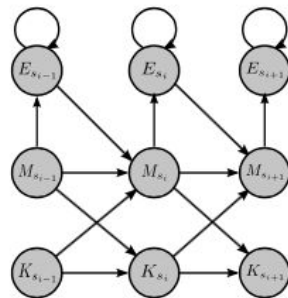
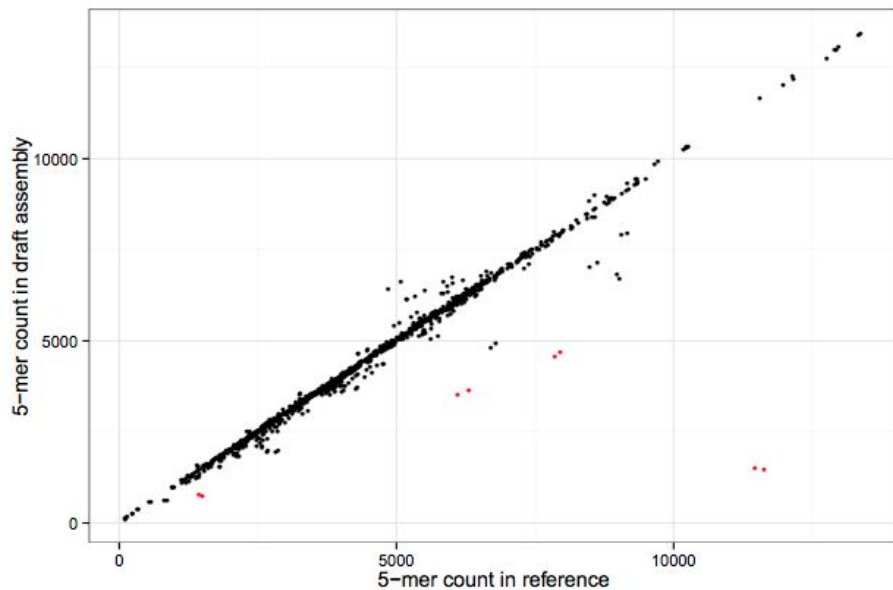


Figure S4: The state structure of the Profile Hidden Markov Model we use to calculate the probability of a sequence of events given a known sequence.

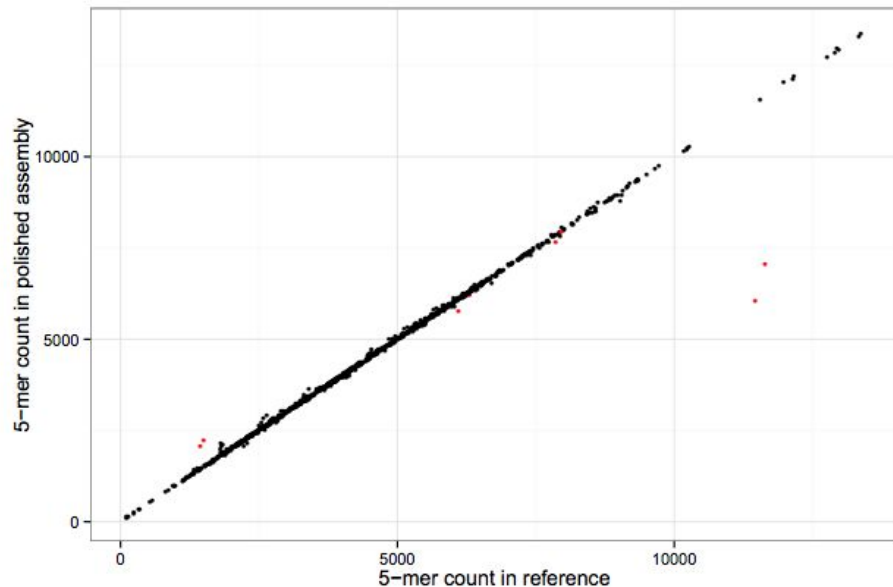
E: 5-mer has emitted an extra event
K: 5-mer has not emitted an event
M: 5-mer has emitted an event

Results

Draft: 98.5% accuracy



Polished: 99.5% accuracy



Results

