Error correction, assembly and consensus algorithms for MinION data

Jared Simpson

Ontario Institute for Cancer Research
&
Department of Computer Science
University of Toronto

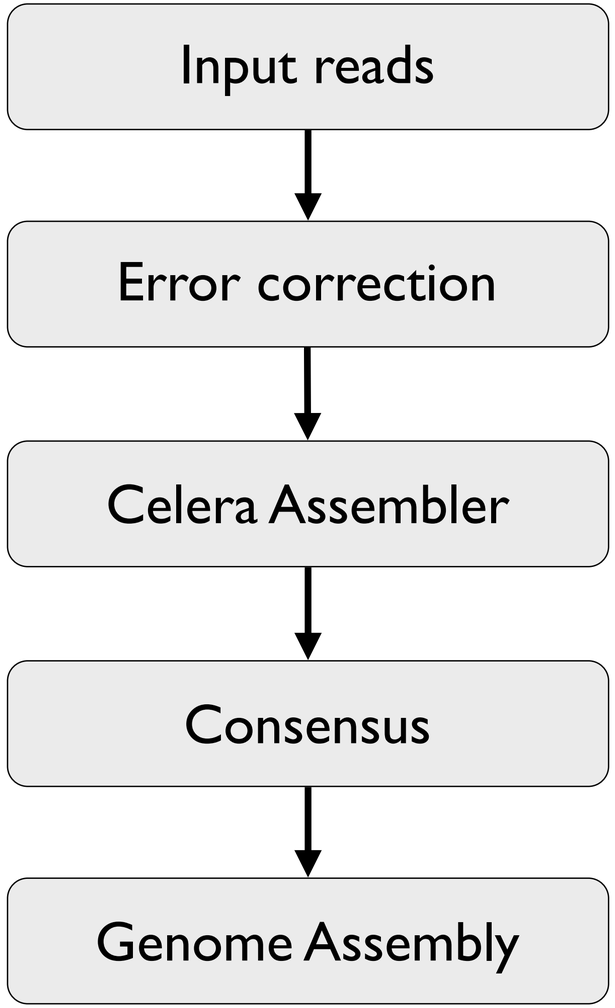
Porecamp - December 2015

An overview of NGS assembly

- Illumina data: short reads, very accurate, very deep
  - nearly all Illumina assembly is based on exact matching algorithms
  - fragmented assemblies

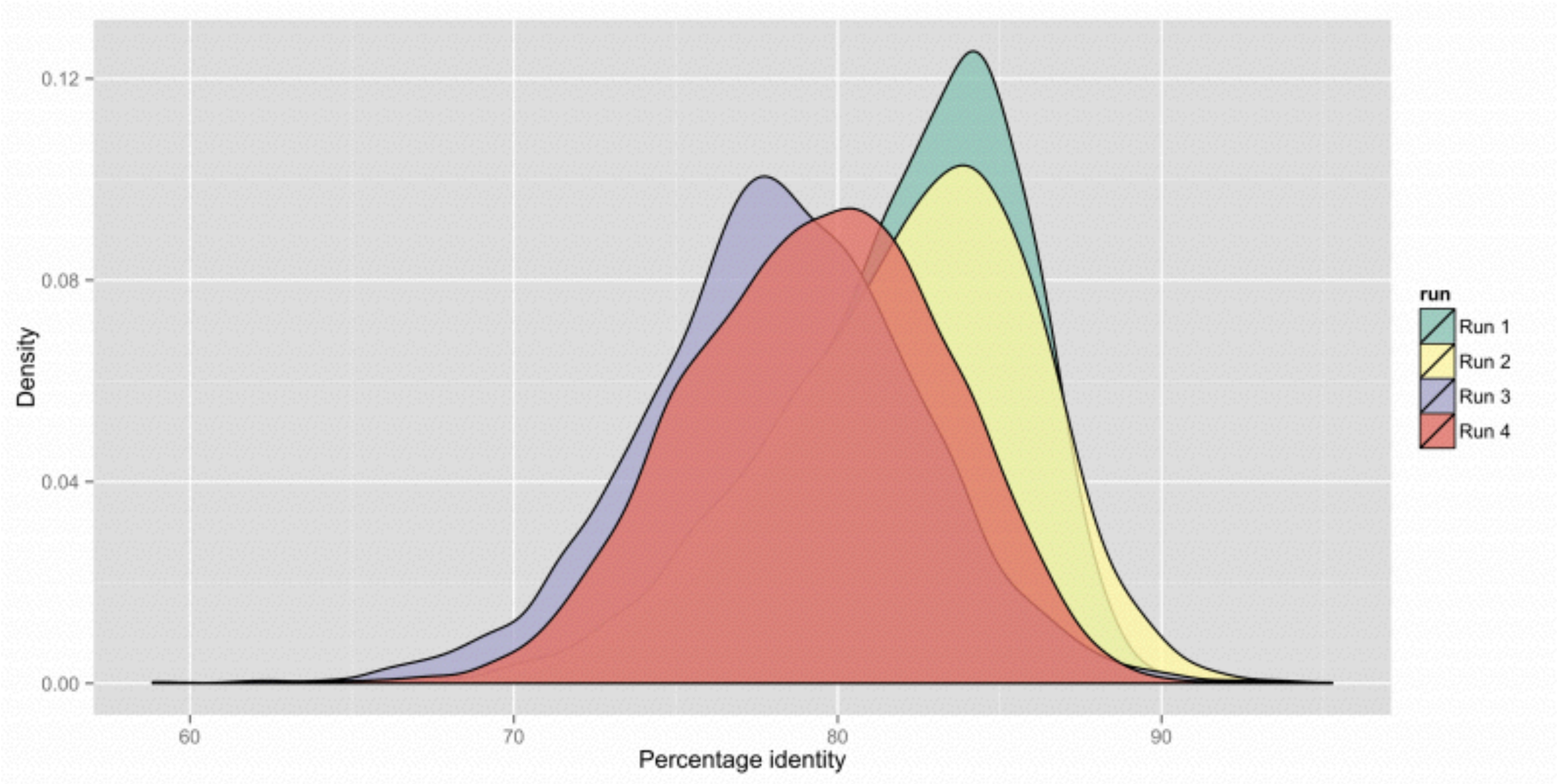
- Algorithms for Illumina data do not work for long, noisy reads
  - PacBio developed a pipeline ("HGAP") to assemble their data
    - We used this recipe as a starting point but with custom components

Long read assembly pipeline



Input Data

- First challenge is finding overlaps for reads with 15-20% errors



Overlap Detection

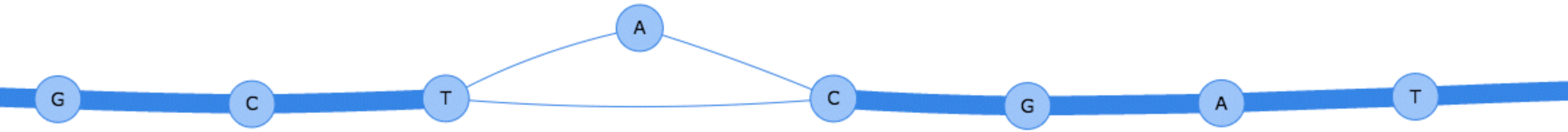
```
2890 -GCCAGAGTCA-AT-GC--TTCCACGCCGGGGTTACCGCCGATAACCGCTAC-CCGTTACGTTAAA-CAGCG--C--TAGTCAGAGGCAATCCGGCAGTA
      ||||| || | || ||  ||||||||||| ||||||||||    ||| ||| |||   |||| |  || ||  |  | | ||| | |||| | |||||
1952 TGCCAG-GT-AGATTGCGCTTCCACGCCGGAGTTACCGCCG----CCGATACGCCGACGCGTTTAGTCATCGAACCGTGGGCAGTGCCAATGC-GCAGTG  32.0%

2979 AAT-CGCCTTTGGTGCGATACTGATCTT-TCTAATAGACTGCTTTCATGTGTT-GCCATTT-GCAAGCGGTCGCCAC--GATAATG---CTGCGTGC-TT
     | | |||||||||||||||||||||||| | || |   |    | ||||||| |  ||||||| || | ||||||||||  |||||||   |||||||| ||
2045 AGTACGCCTTTGGTGCGATACTGATCCTGTC-ACCCG-GAACGTTCATGT-TGCGCCATTTTGC-ACCGGTCGCCACACGATAATGATGCTGCGTGCCTT  25.0%

3069 TCAGTACCGCGC-CAGAAGCTGTT-CACGTG-A-TGCAGACCACTTCAACTGCTGCTGGGATCAGTTTAGCCGAC-C-CTGTCTAA--ATCACCGT-AAC
     |||||  | | | ||||||||||| ||  |  | |||||||| ||||||||||||||| |||||||||| || | | | ||| ||    |||| | | |||
2141 TCAGTGGC-CACTCAGAAGCTGTTTCA-ATCTACTGCAGACC-CTTCAACTGCTGCT-GGATCAGTTT-GCTGGCGCTCTGGCTGTCGATCA-CATCAAC  26.0%

3160 GTCGTATTCATCAACGTGAACTT--CAGTGCCTGCC---TTC--CTTCAG-TCTTCG--GTACAGAAATGTAGTTTTCGATATGA-CGGTA-C---CG-T
     ||||  |||||||||||||||||  |||||||||||   |||  |||||| |||||   |||||||||||||||||||||||| | ||||| |   || |
2235 GTCG-GTTCATCAACGTGAACTTTTCAGTGCCTGCCAGCTTCTGCTTCAGGTCTTCTCAGTACAGAAATGTAGTTTTCGATATCAACGGTATCGGGCGAT  20.0%

3244 -T-ACCA--AA--CGTTCG-CCATCAGAAGGTAATCG-ATGCCTTTACGTGCGCTATAATTAGAT-C--GCTTAAGCATG--GGGC-GGAA--CCGACGA
      | ||||  ||  |||||  ||||||| |  | | || |||||||||||||||  |||| | | | |  |||  | |  |  |||| ||||  |||||||
2334 CTGACCACCAAAACGTTCCACCATCAG-A-CTCAACGGATGCCTTTACGTGCGGAATAAATTGCTGCCCGCTGCA-CCCGCCGGGCCGGAAATCCGACGA  32.0%
```

we use github.com/thegenemyers/daligner to compute overlaps

Partial Order Graphs

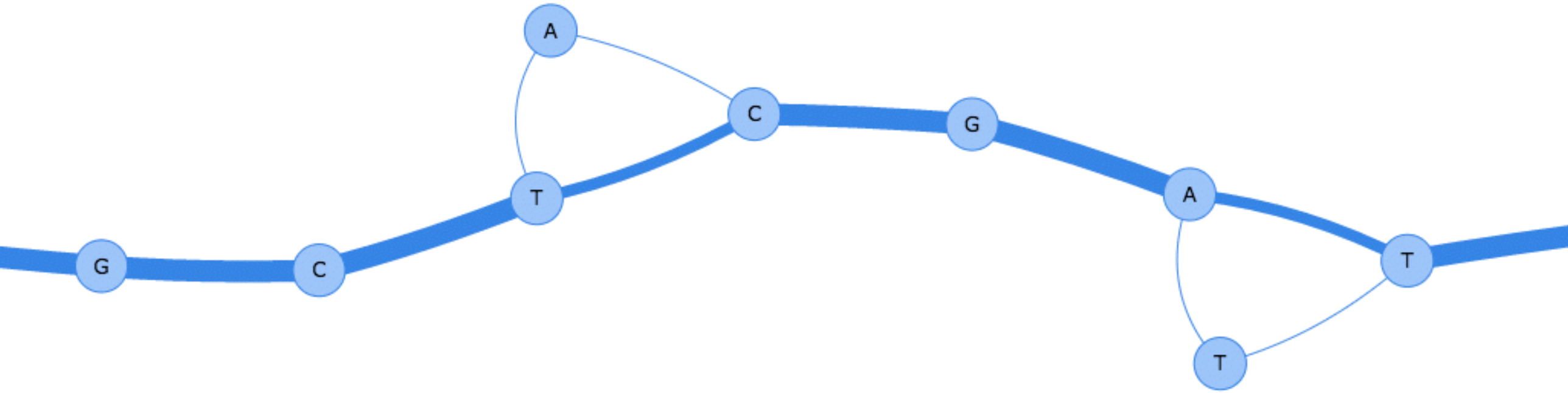
add read GCTACGAT that we want to correct to graph

Partial Order Graphs



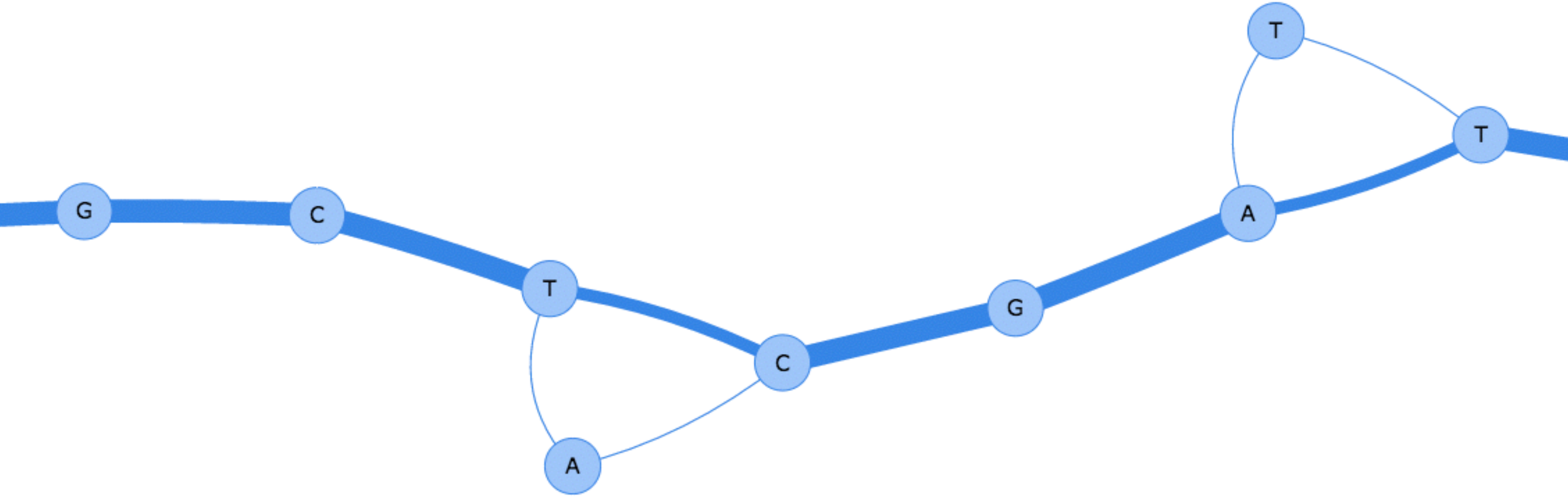
add sequence GCTCGAT to graph

Partial Order Graphs



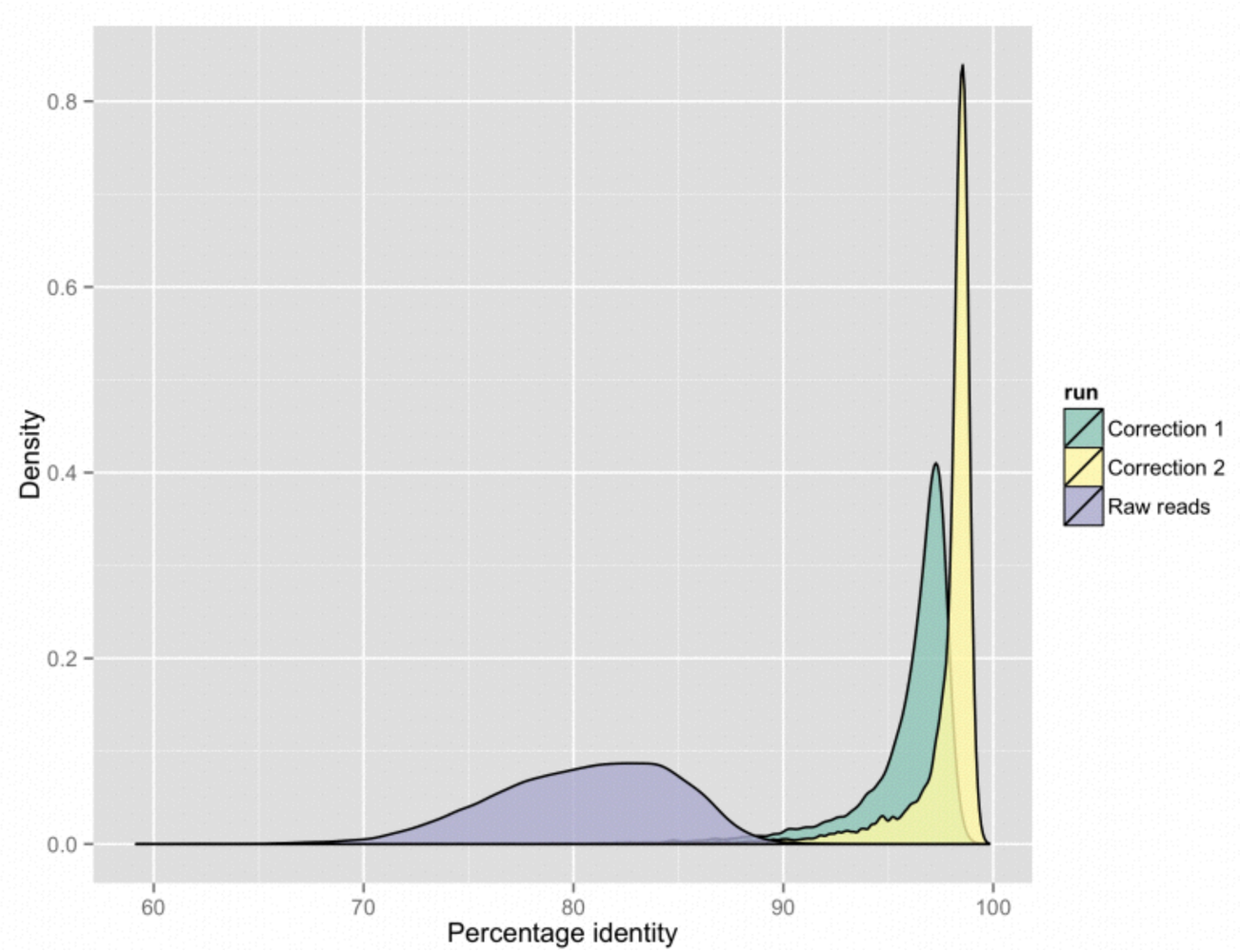
add sequence GCTCGATT to graph

Partial Order Graphs



maximum weight path GCTCGAT is the corrected read

Error Correction



Contig Assembly

Celera Assembler produces one contig at 98.5% identity

```
Query: 61     gacaaccagatttatctgtcgatttcgctccccttcggtaacggtggtccgcttggctat 120
              ||||||||||||||||||||||||||||||||||||||||||||||||||| | ||||||||
Sbjct: 750537 gacaaccagatttatctgtcgatttcgctccccttcggtaacggtggtcgggttggctat 750478

Query: 121    gacatgcaaa-cagtagccacagcaccatacaccgcatgtcgtggaacgatacgctggat 179
              |||||||||| |||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 750477 gacatgcaaaacagtagccacagcaccatacaccgcatgtcgtggaacgatacgctggat 750418

Query: 180    gaacgtaatagctggggcatgtctgccggactgcaatccgatcgttagaaccggacaatg 239
              |||||||||||||||||||||||||||||||||||||||||||     ||||||||||
Sbjct: 750417 gaacgtaatagctggggcatgtctgccggactgcaatccgatcgt-----ccggacaatg 750363

Query: 240    gagccca--tgagcggtaactatcagcacctgagttcagcgggtgagtgggatatttctg 297
              |||||||  |||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 750362 gagcccaggtgagcggtaactatcagcacctgagttcagcgggtgagtgggatatttctg 750303
```

Assembly Polishing

- Consensus problem is viewed as choosing a sequence *C'* that maximizes the probability of the event data

$$C' = \arg\max_{S \in \mathcal{C}} P(\mathcal{D}|S)$$

where

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

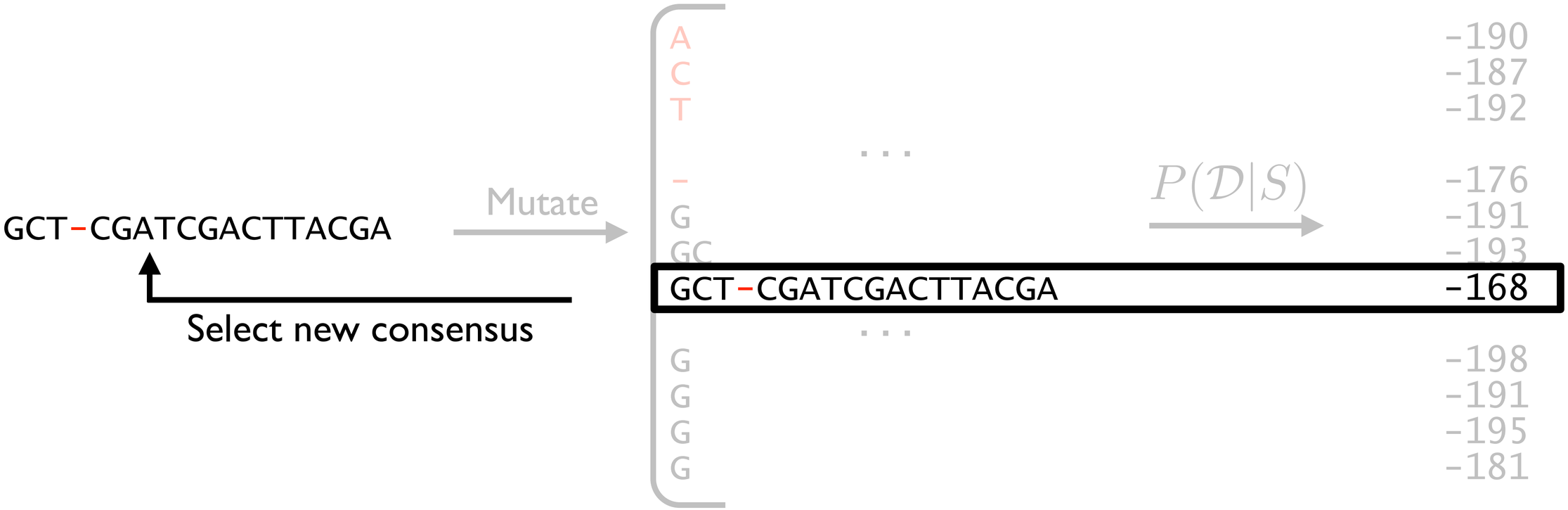
Selecting a Consensus

GCTACGATCGACTTACGA $\xrightarrow{\text{Mutate}}$

ACTACGATCGACTTACGA	-190
CCTACGATCGACTTACGA	-187
TCTACGATCGACTTACGA	-192
...	
-CTACGATCGACTTACGA	-176
G-TACGATCGACTTACGA	-191
GC-ACGATCGACTTACGA	-193
GCT-CGATCGACTTACGA	-168
...	
GACTACGATCGACTTACGA	-198
GCCTACGATCGACTTACGA	-191
GGCTACGATCGACTTACGA	-195
GTCTACGATCGACTTACGA	-181

$\xrightarrow{P(\mathcal{D}|S)}$

Selecting a Consensus

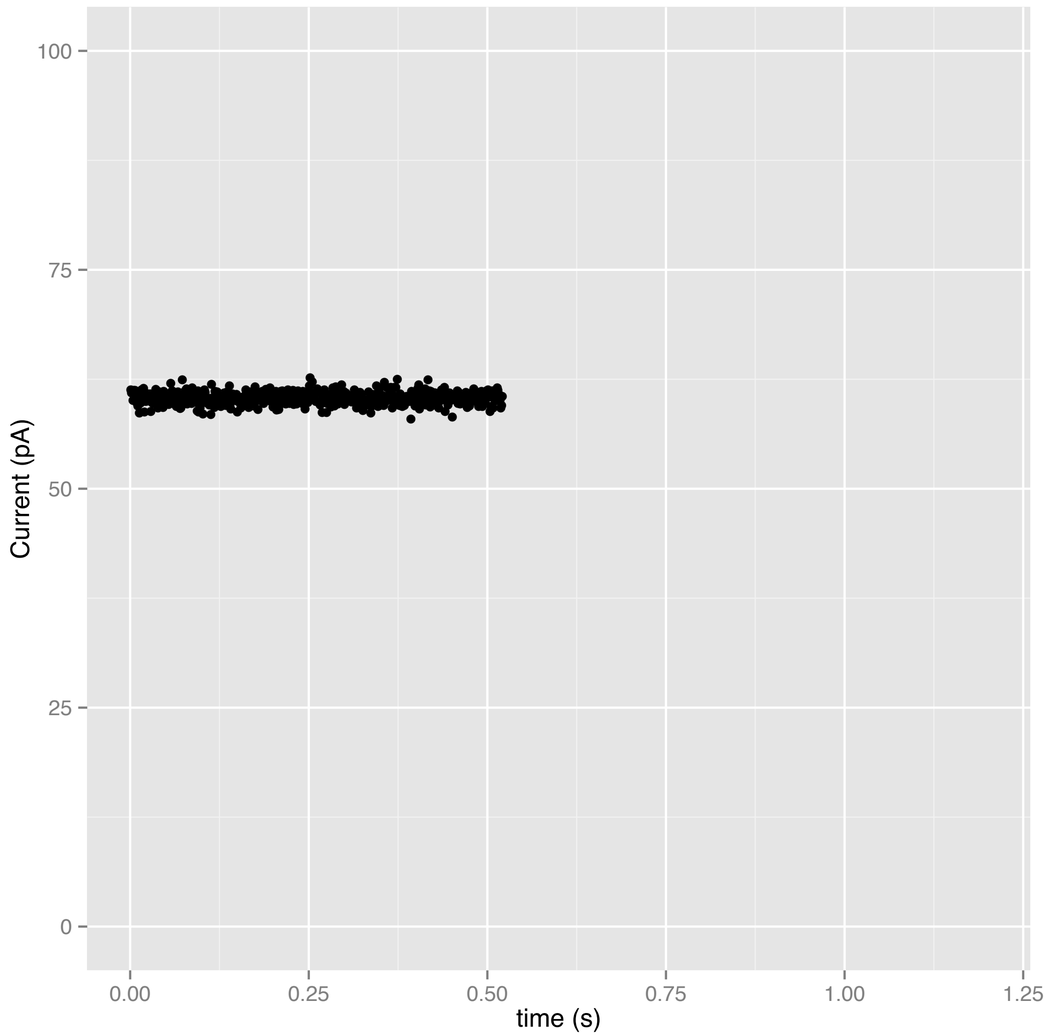


Pore Models

5-mer	μ_k	σ_k
AAAAA	53.5	1.3
AAAAC	54.2	0.9
...	...	...
TTTTG	65.3	1.8
TTTTT	67.1	1.4

Generating Events

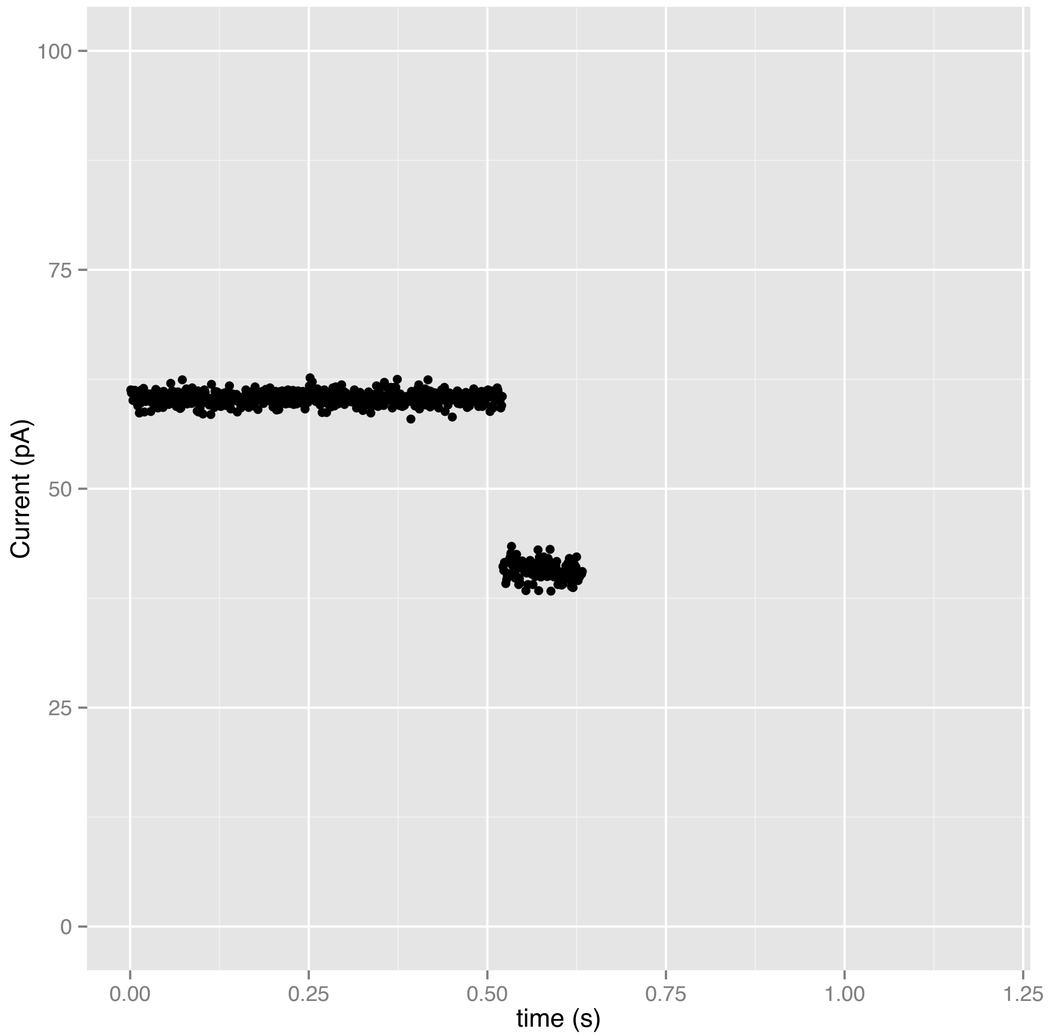
- What do we expect events from a given sequence to look like?



Generating Events

- What do we expect events from a given sequence to look like?

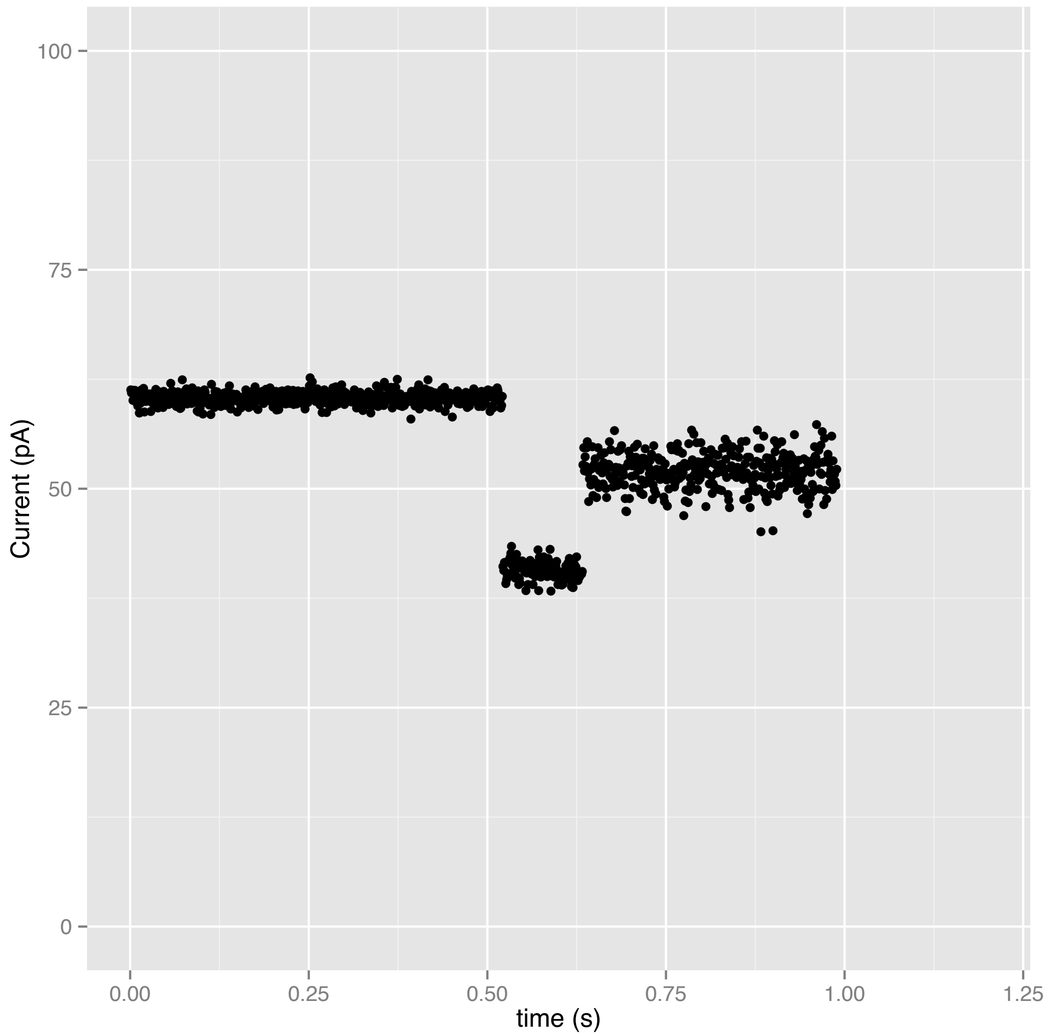
Sample Current



Generating Events

- What do we expect events from a given sequence to look like?

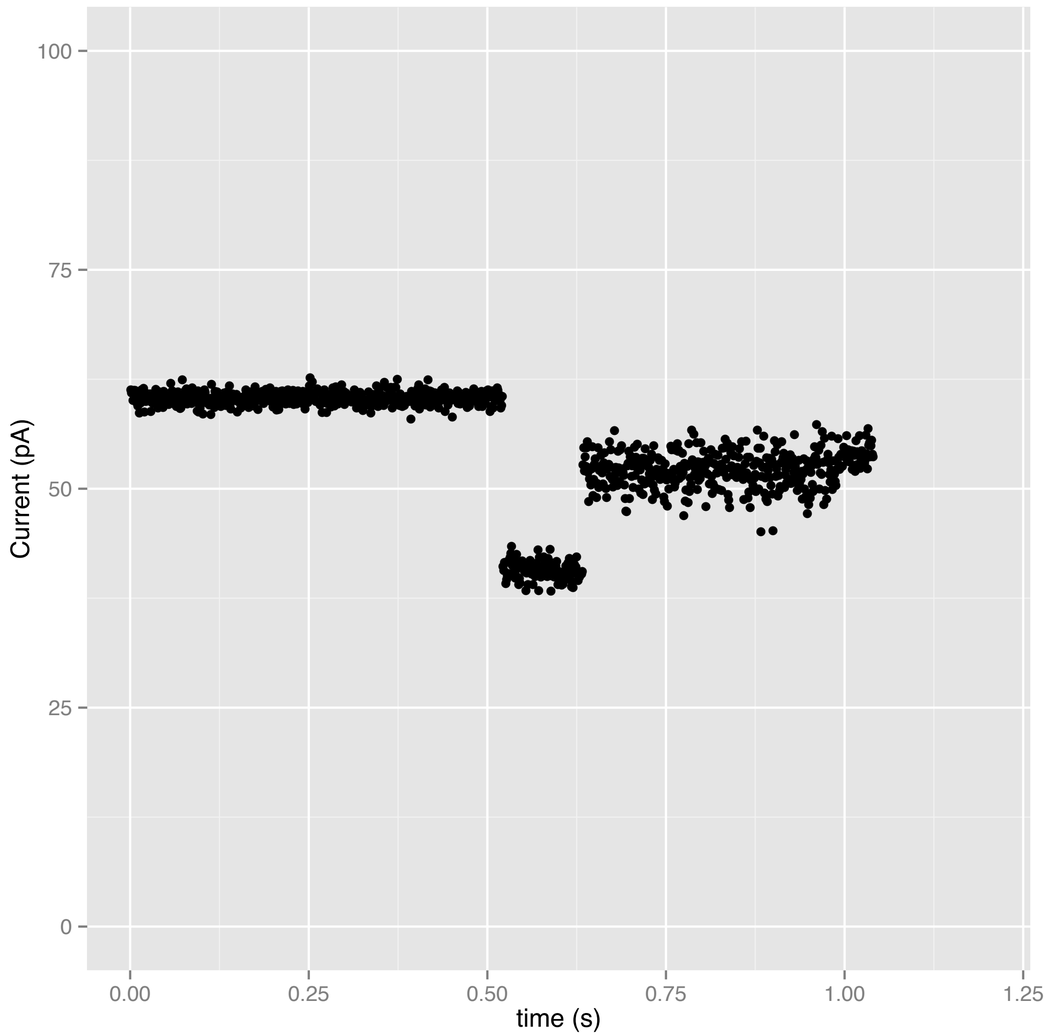
Sample Current



Generating Events

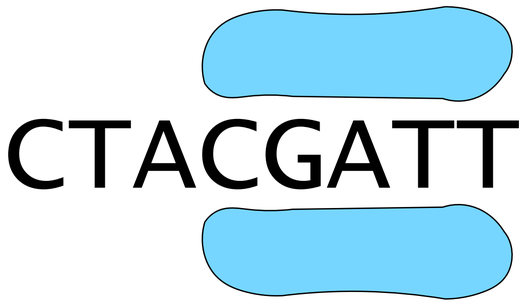
- What do we expect events from a given sequence to look like?

Sample Current

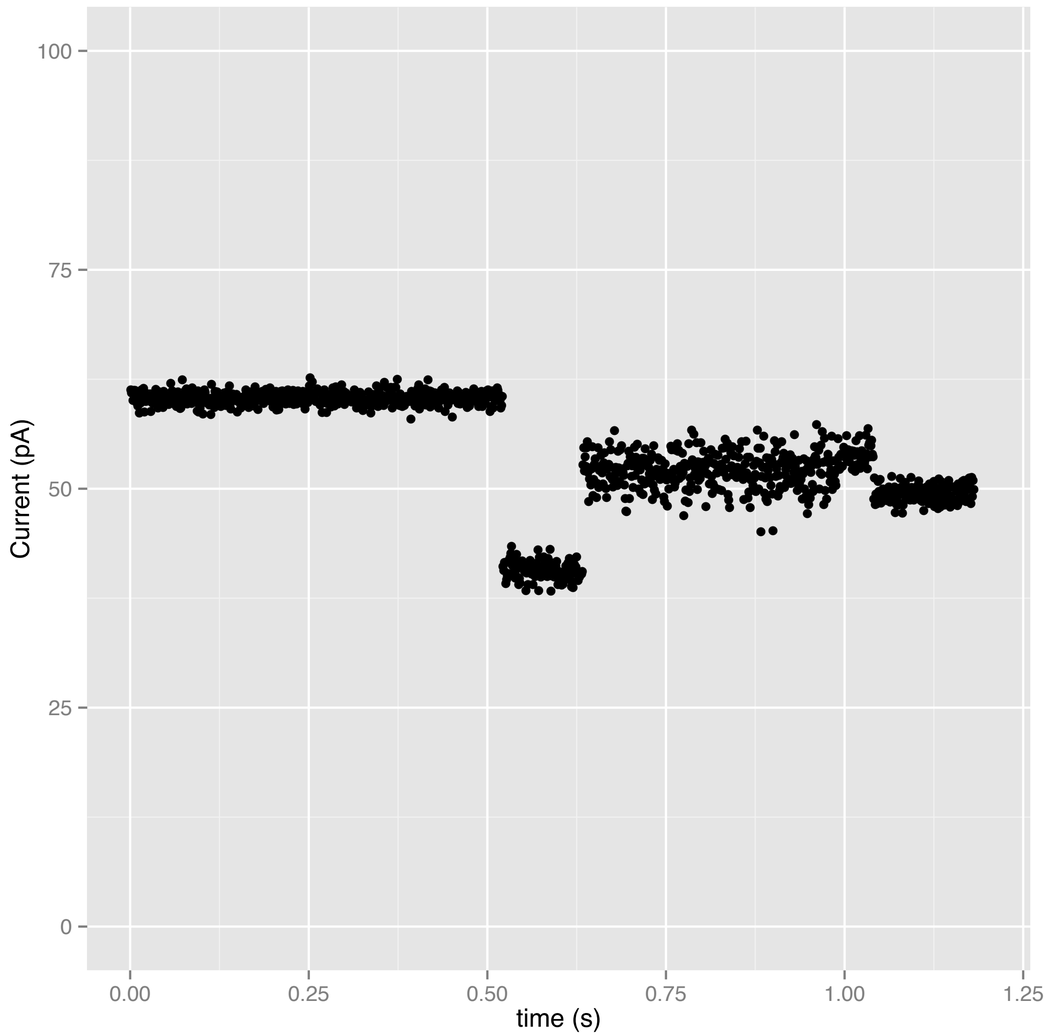


Generating Events

- What do we expect events from a given sequence to look like?

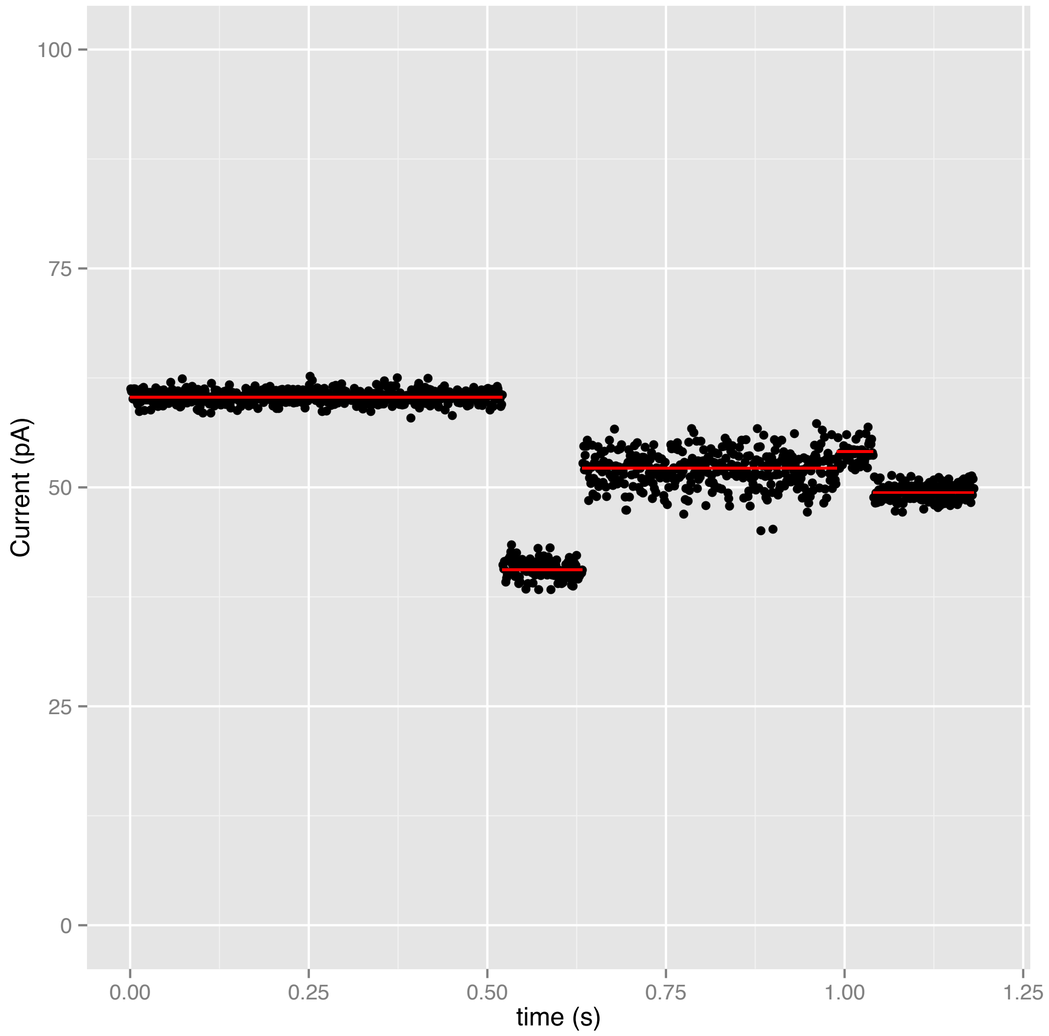


Sample Current



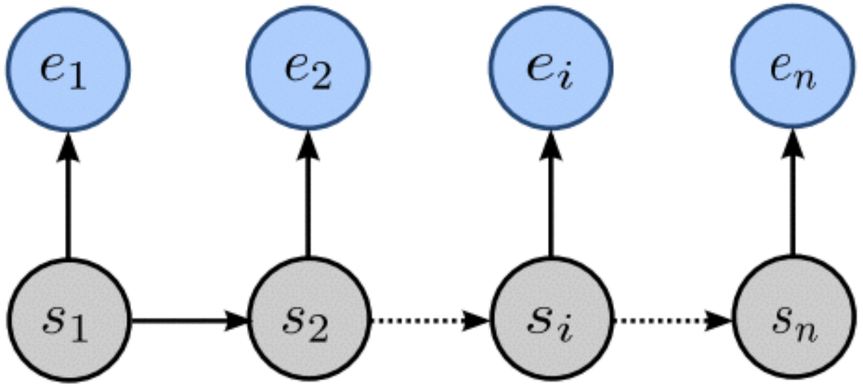
Event Detection

Event	mean current (pA)	current stdv	duration (s)
1	60.3	0.7	0.521
2	40.6	1.0	0.112
3	52.2	2.0	0.356
4	54.1	1.2	0.291
5	49.5	1.5	0.141



A simple model

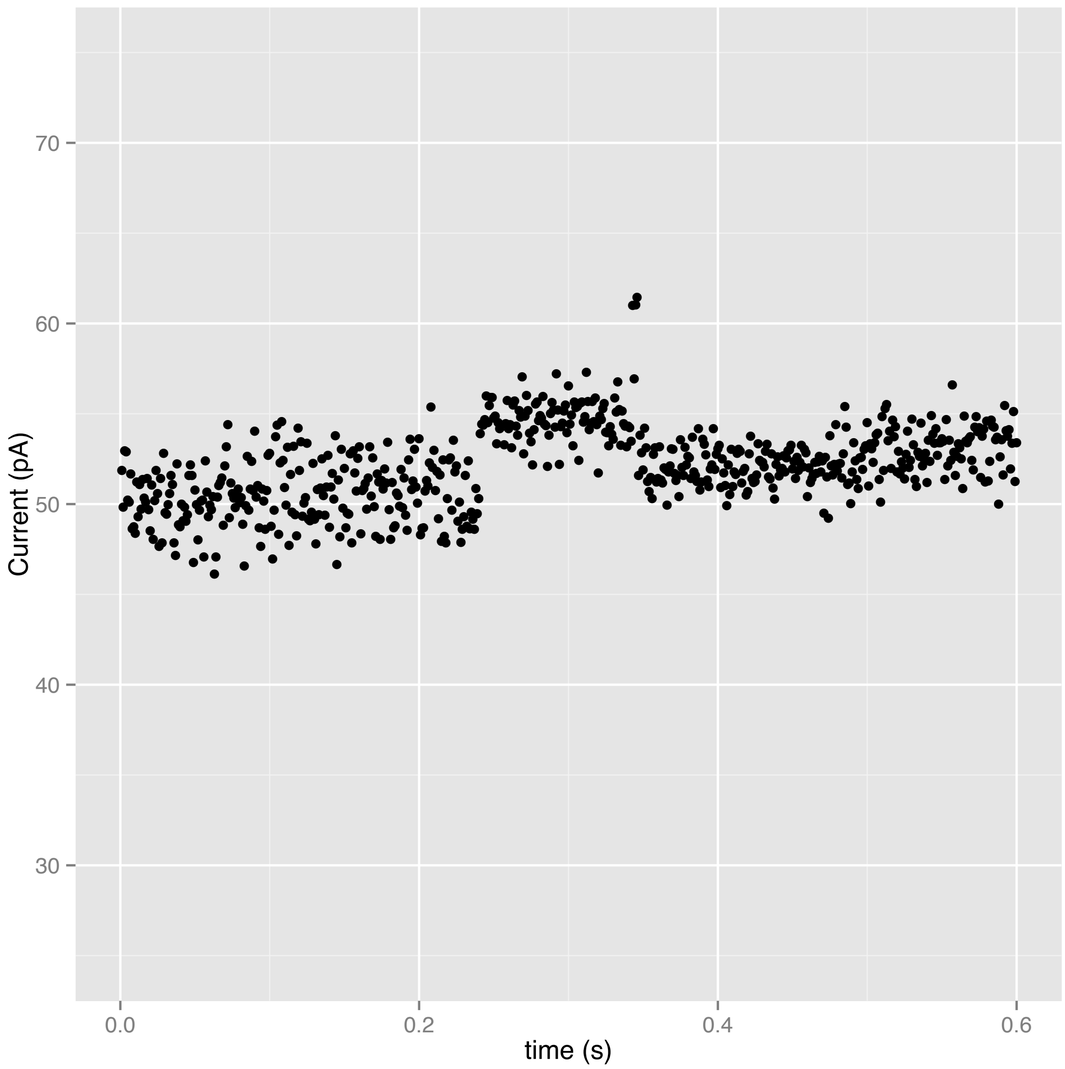
- What is the probability of observing events E given a sequence S?
- Assuming for the moment there are no missing or extra events:



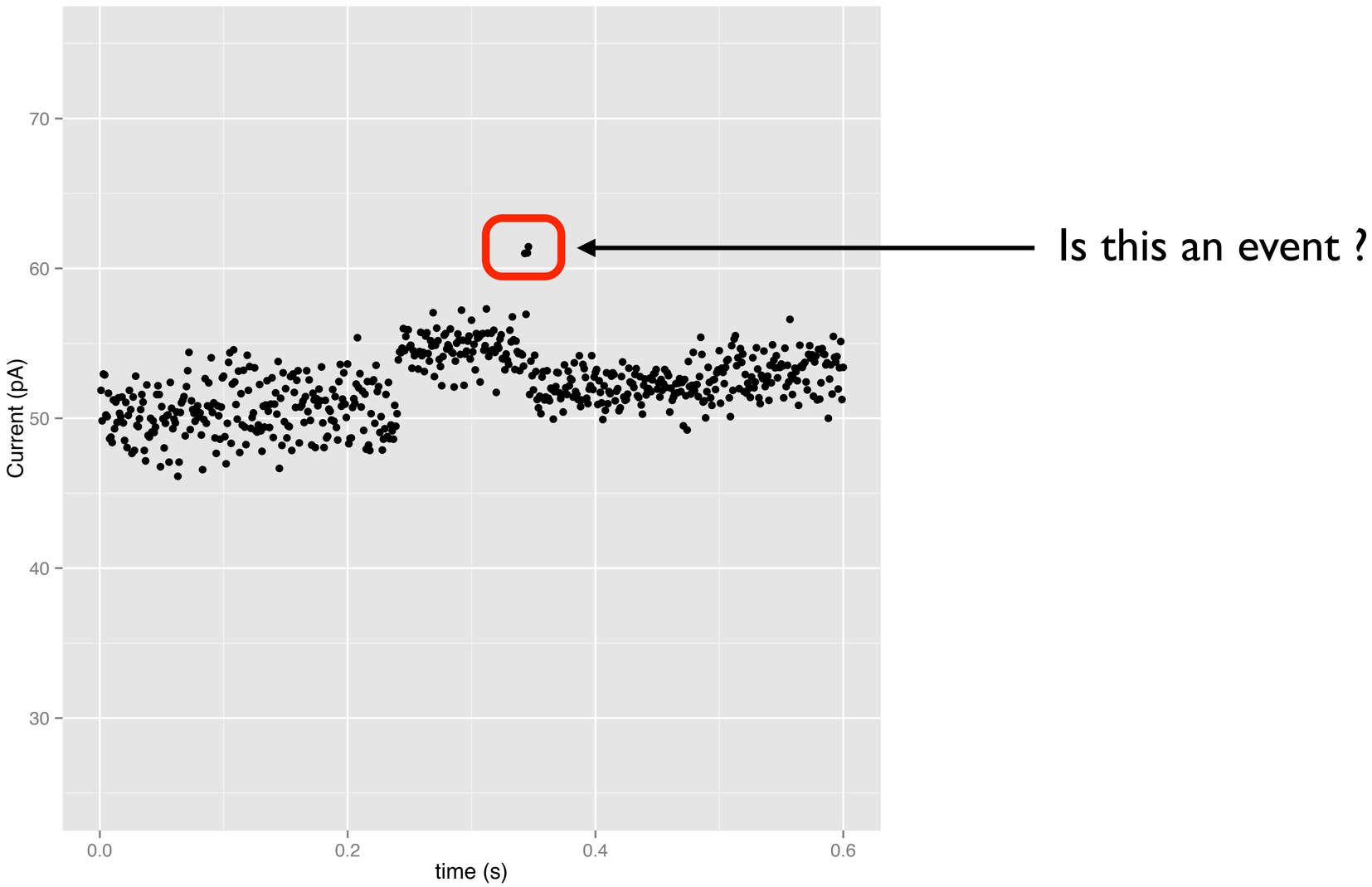
$$P(e_1, e_2, ..., e_n | s_1, s_2, ..., s_n, \mathbf{\Theta}) = \prod_{i=1}^{n} P(e_i | s_i, \mu_{s_i}, \sigma_{s_i})$$

$$P(e_i | k, \mu_k, \sigma_k) = \mathcal{N}(\mu_k, \sigma_k^2)$$

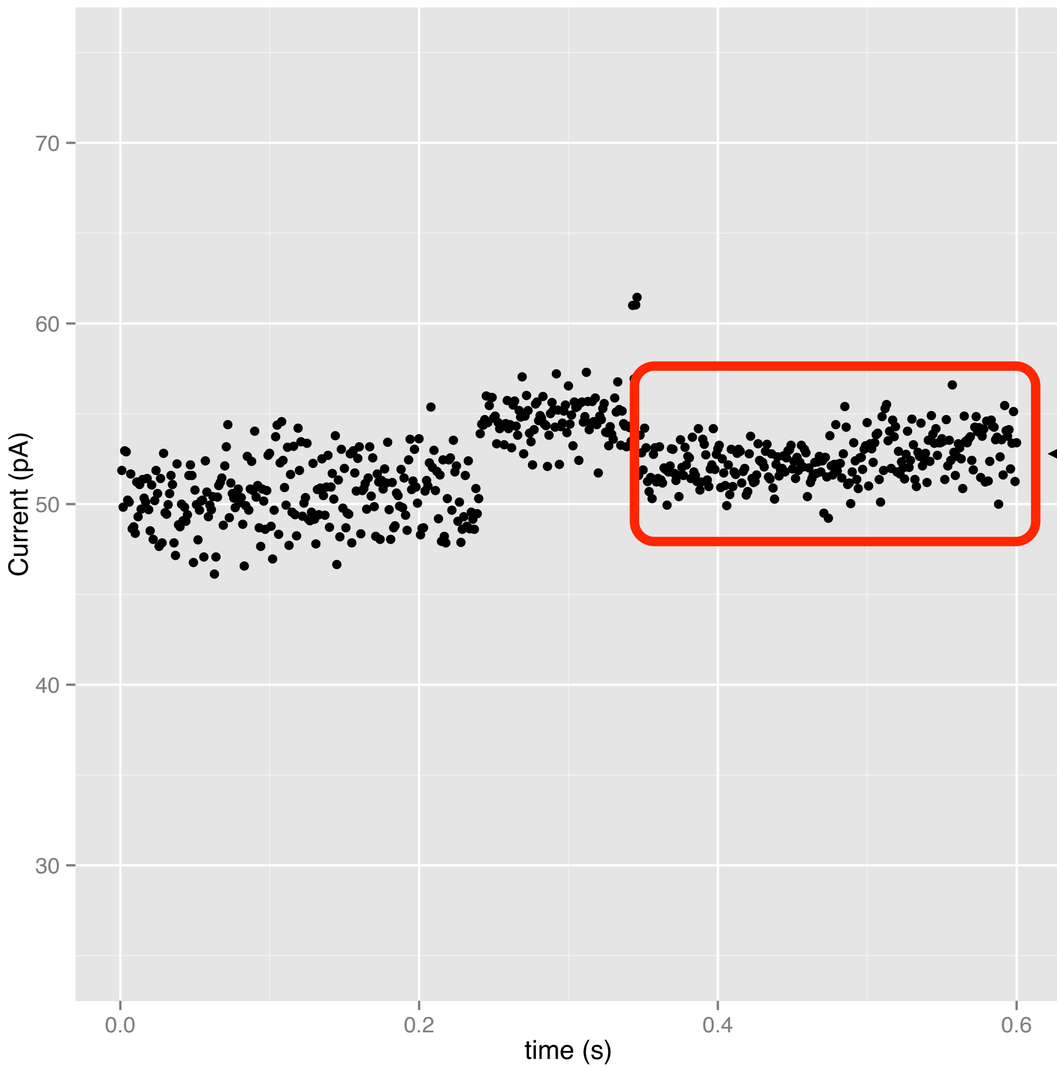
Complications



Complications



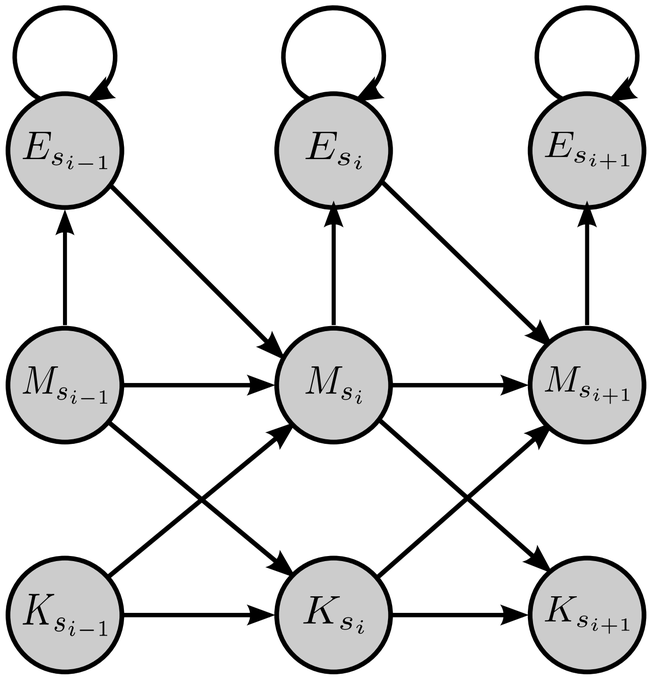
Complications



One event or two ?

Nanopore HMM

- $P(\mathcal{D}|S)$ must consider:
  - over segmentation
  - under segmentation
  - missed short events
- HMM:
  - M states: match event to 5-mers
  - E states: extra obs. of an event
  - K states: no event obs. for 5-mer

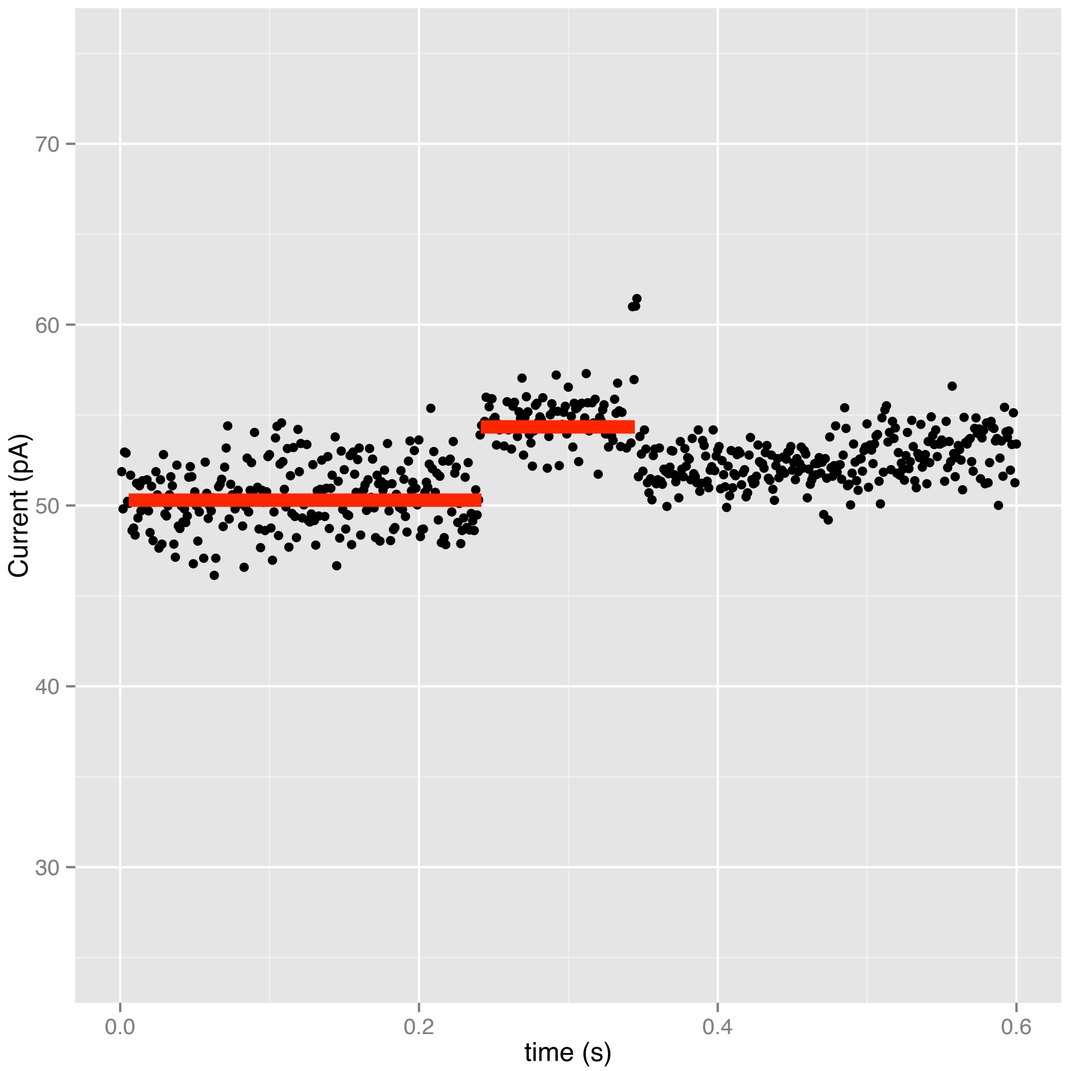


$$P(\pi, e_1, e_2, ..., e_n|S, \mathbf{\Theta}) = \prod_{i=1}^{n} P(e_i|\pi_i, \mu_{s_i}, \sigma_{s_i}) P(\pi_i|\pi_{i-1}, S)$$

$$P(e_1, e_2, ..., e_n|S, \mathbf{\Theta}) = \sum_{\pi} P(\pi, e_1, e_2, ..., e_n|S, \mathbf{\Theta})$$

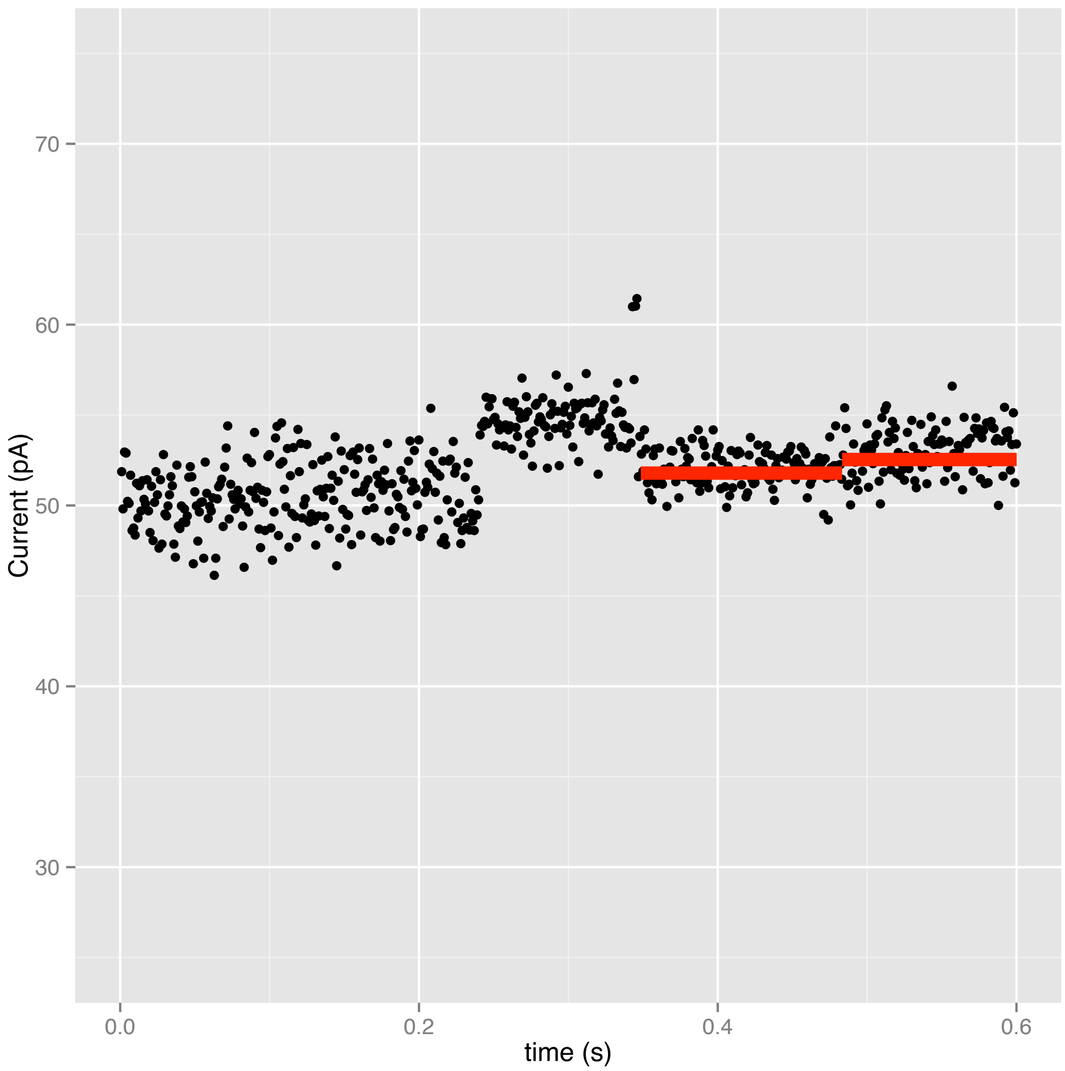
Transition Probabilities

- Probability of not observing an event is a function of absolute difference between (expected) current

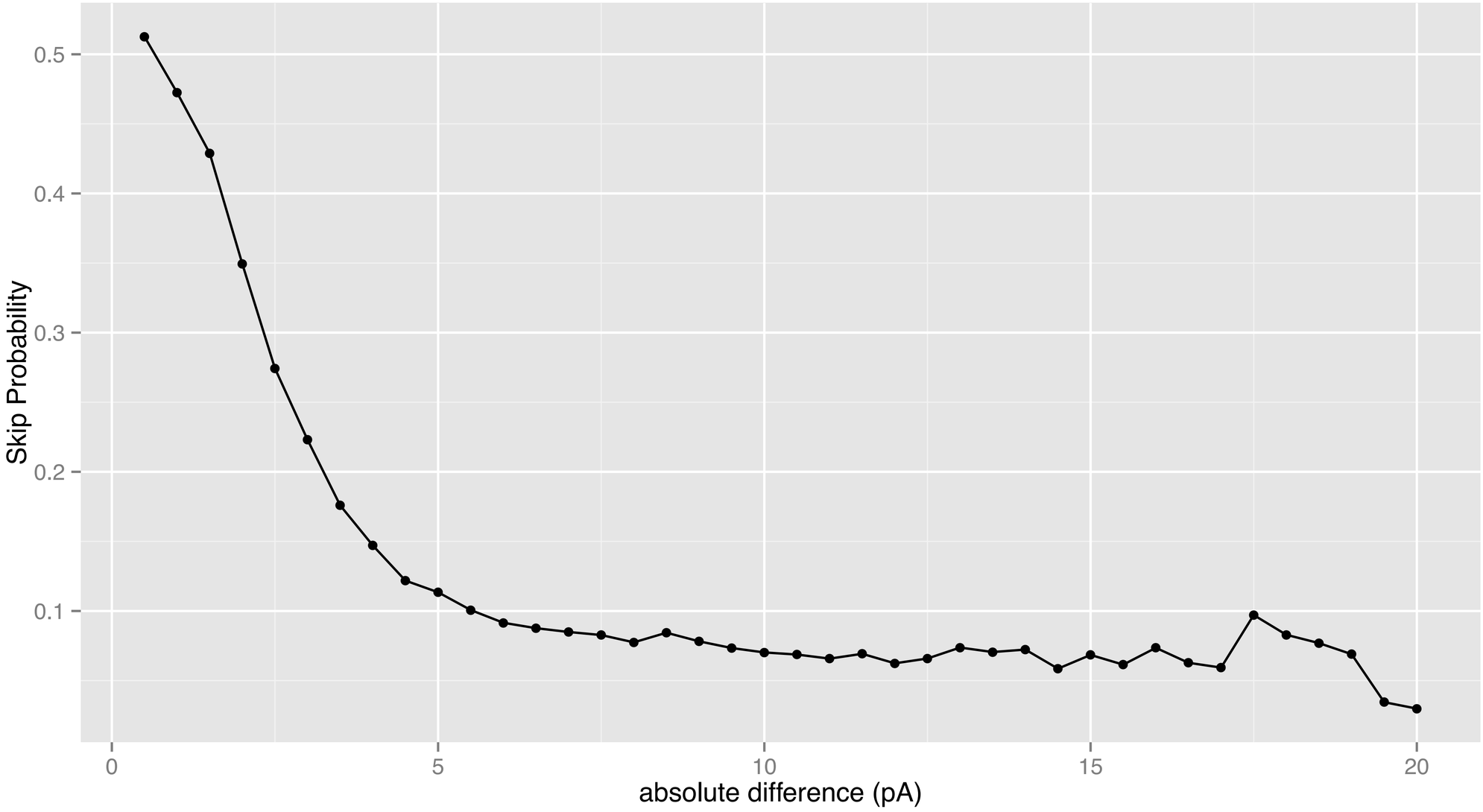


Transition Probabilities

- Probability of not observing an event is a function of absolute difference between (expected) current

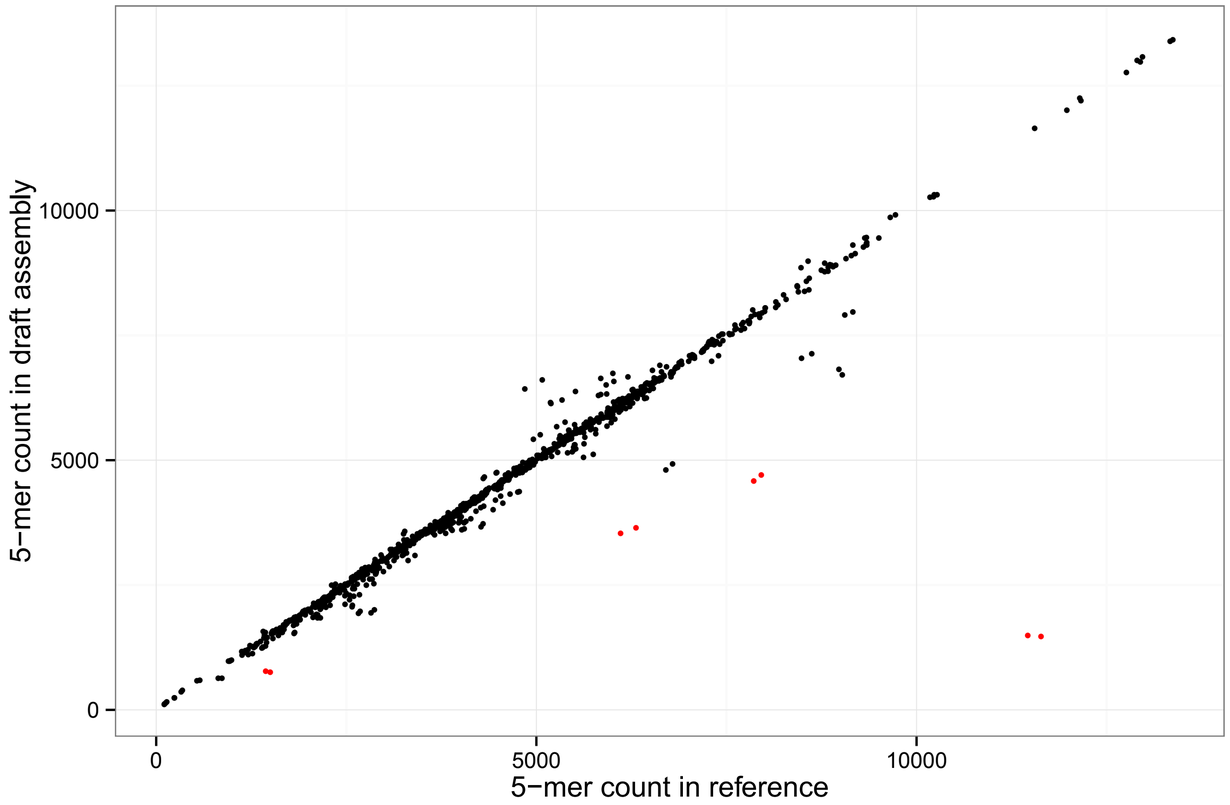


Transition Probabilities

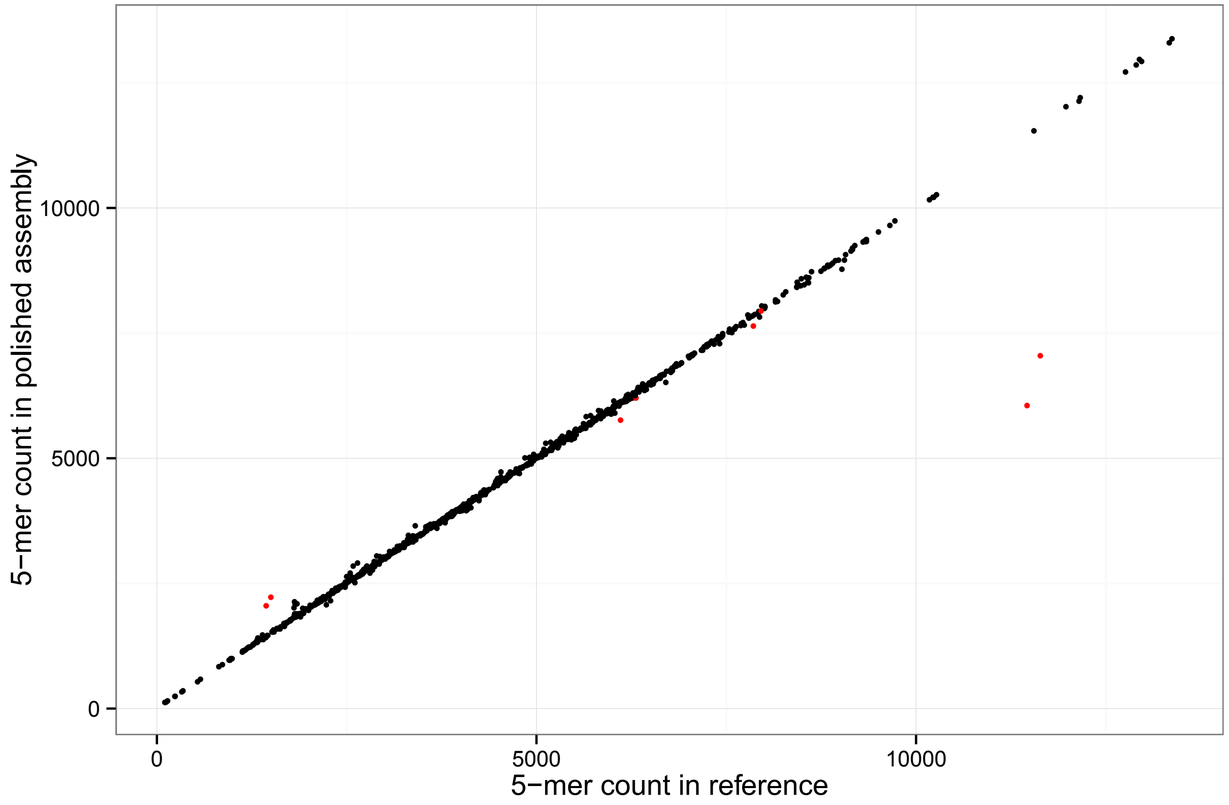


Assembly Accuracy

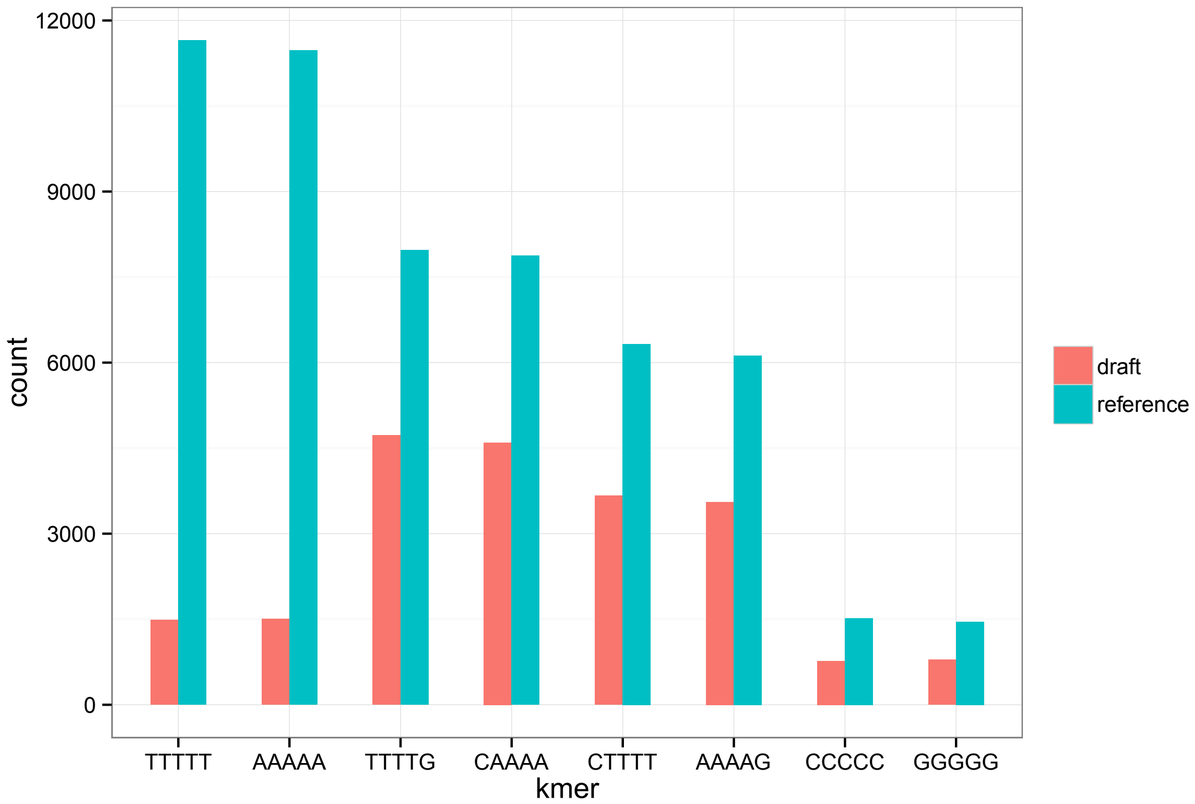
Draft: 98.5% accuracy

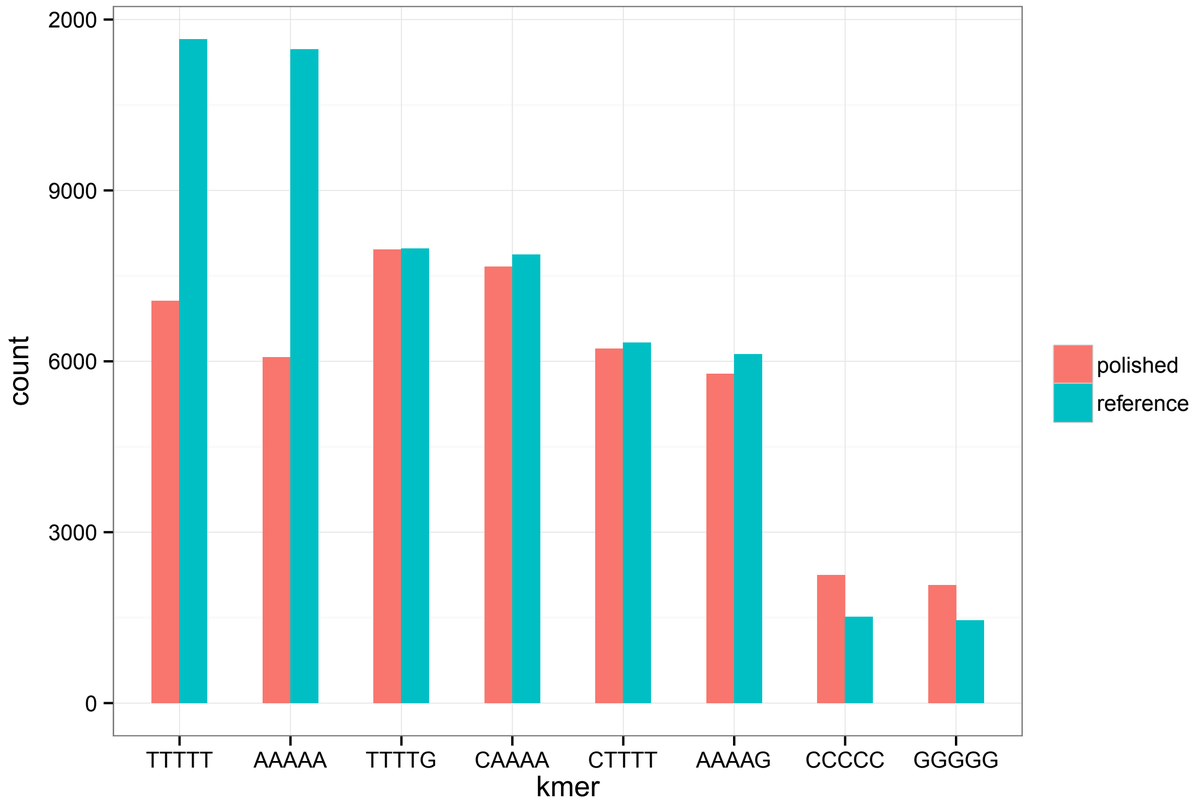


Polished: 99.5% accuracy



Assembly Accuracy





Aligning Events to a Reference

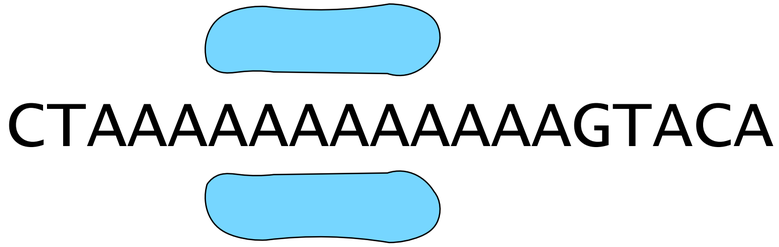
- HMM can also align events to a reference genome

contig	position	reference_kmer	read_index	strand	event_index	event_level_mean	event_length	model_kmer	model_mean	model_stdv
gi|556503834|ref|NC_000913.3|	10000	ATTGC	1	c	27470	50.57	0.022	ATTGC	50.58	1.02
gi|556503834|ref|NC_000913.3|	10001	TTGCG	1	c	27471	52.31	0.023	TTGCG	51.68	0.73
gi|556503834|ref|NC_000913.3|	10001	TTGCG	1	c	27472	53.05	0.056	TTGCG	51.68	0.73
gi|556503834|ref|NC_000913.3|	10001	TTGCG	1	c	27473	54.56	0.011	TTGCG	51.68	0.73
gi|556503834|ref|NC_000913.3|	10002	TGCGC	1	c	27474	65.56	0.012	TGCGC	66.96	2.91
gi|556503834|ref|NC_000913.3|	10002	TGCGC	1	c	27475	69.97	0.071	TGCGC	66.96	2.91
gi|556503834|ref|NC_000913.3|	10003	GCGCT	1	c	27476	67.11	0.017	GCGCT	68.08	2.20
gi|556503834|ref|NC_000913.3|	10004	CGCTG	1	c	27477	69.47	0.052	CGCTG	69.84	1.89

- Read about it here:
  - http://simpsonlab.github.io/2015/04/08/eventalign/

Planned Improvements

- Improve detection of homopolymers (dwell times?)



- SNP calling/genotyping
- Improve scalability to handle larger genomes

Code

- Code:
  - github.com/jts/nanocorrect (error correction)
  - github.com/jts/nanopolish (signal-level algorithms)
  - github.com/jts/nanopore-paper-analysis (reproduce our paper)

Methylation