JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Bacterial Sequencing and Assembly for Analysis of Antibiotic Resistance Genes and Mutations

Yunfan Fan

Department of Biomedical Engineering

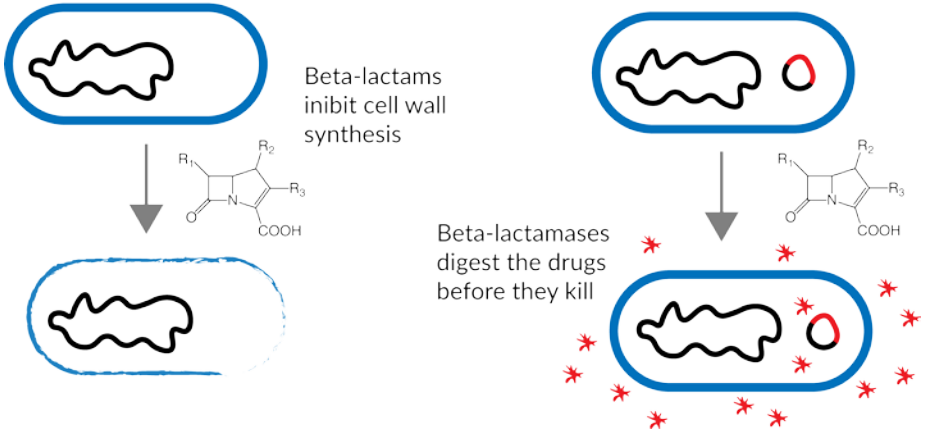
Johns Hopkins University

Timp Lab

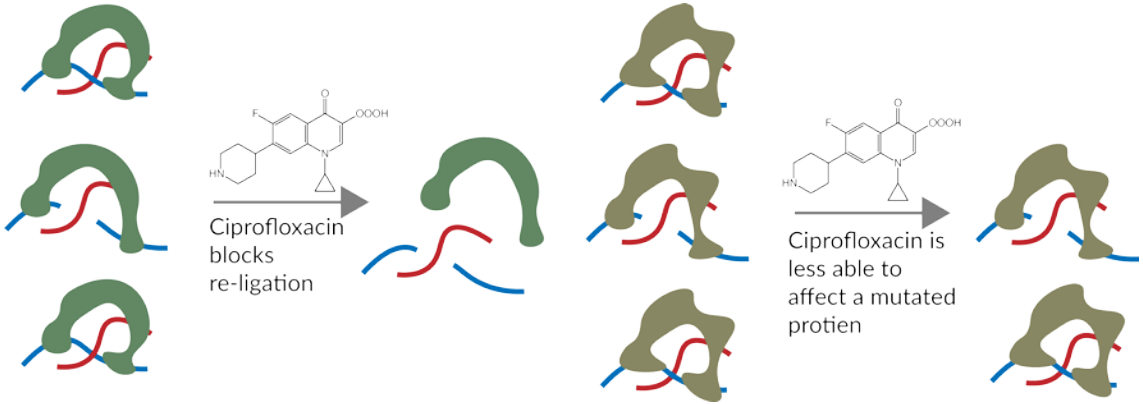
Sequencing, Finishing and Analysis in the Future

May 24, 2018

Antimicrobial Resistance



Resistance genes can be acquired via plasmids.



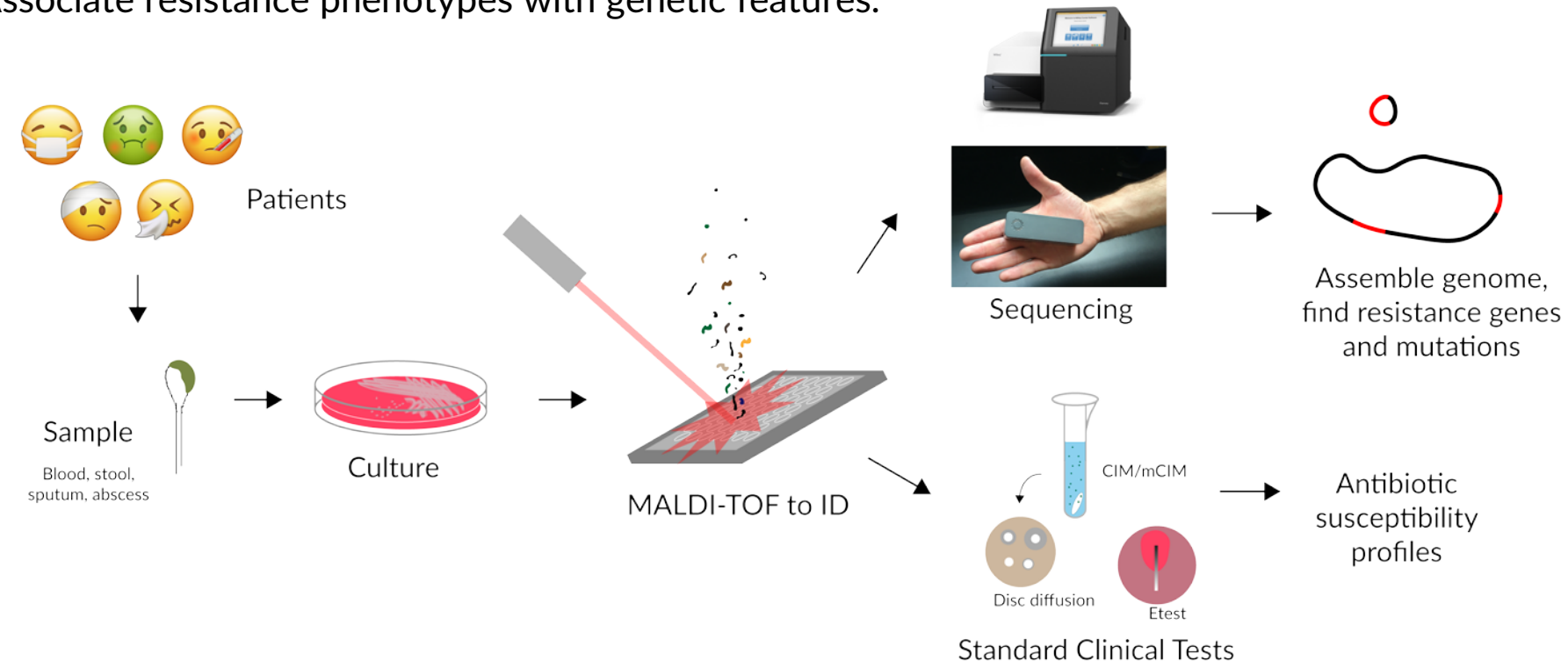
Mutations of otherwise benign genes can confer resistance.

(gyrA)

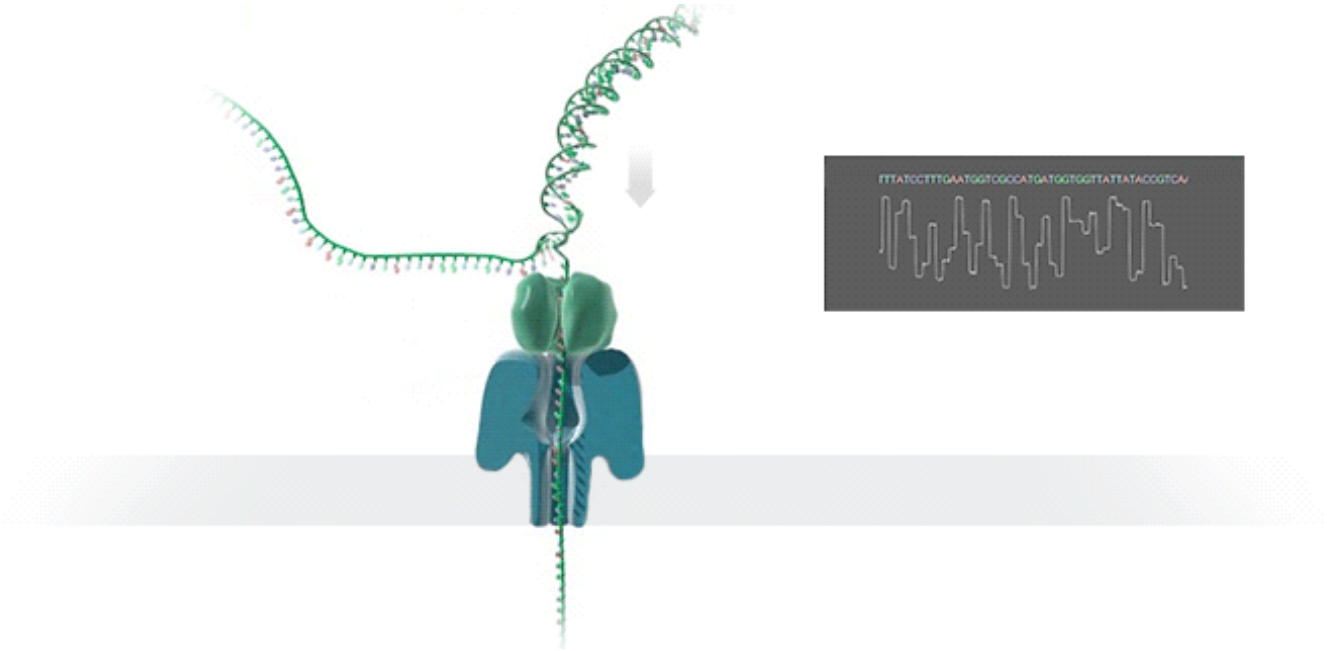
Aldred et al *Biochemistry* (Feb 2014)

Overview

Associate resistance phenotypes with genetic features.



Nanopore Sequencing

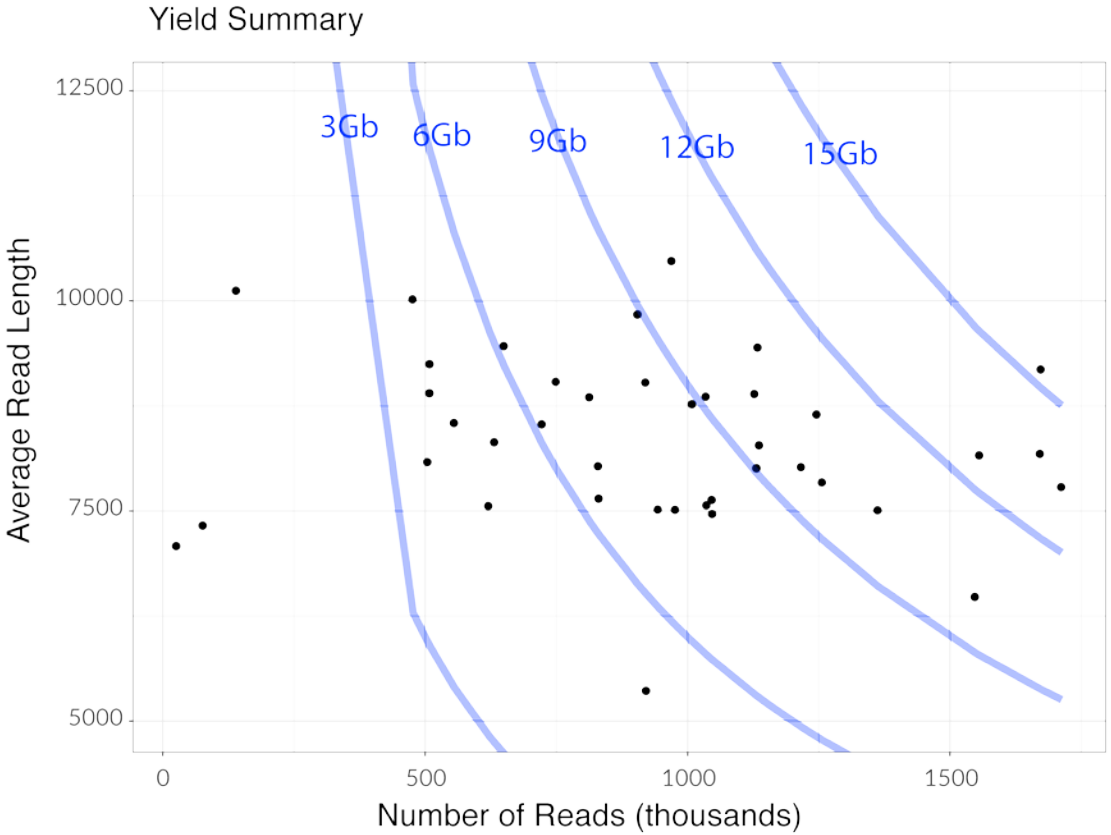


Changes in ionic current are measured as DNA threads through a pore.

Multiple bases occupy the pore at once – these k-mers produce characteristic current signatures.

Oxford Nanopore Technologies

Nanopore Sequencing



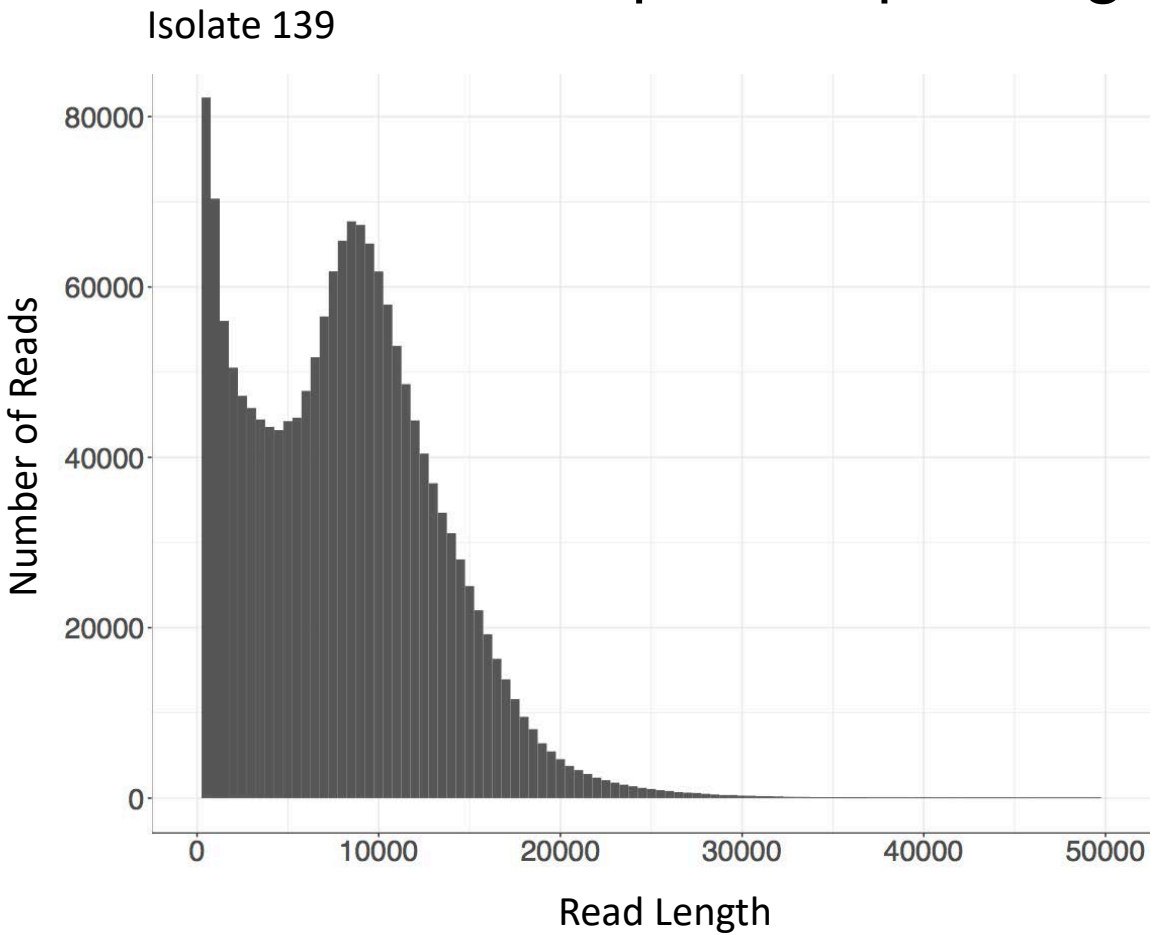
We easily get long reads and high coverage – both critical for good assemblies.

Simner Lab bacterial runs average 8.9Gb in yield. At $500 per flowcell, it costs less than 50 cents for 1X coverage of a large bacterial genome.

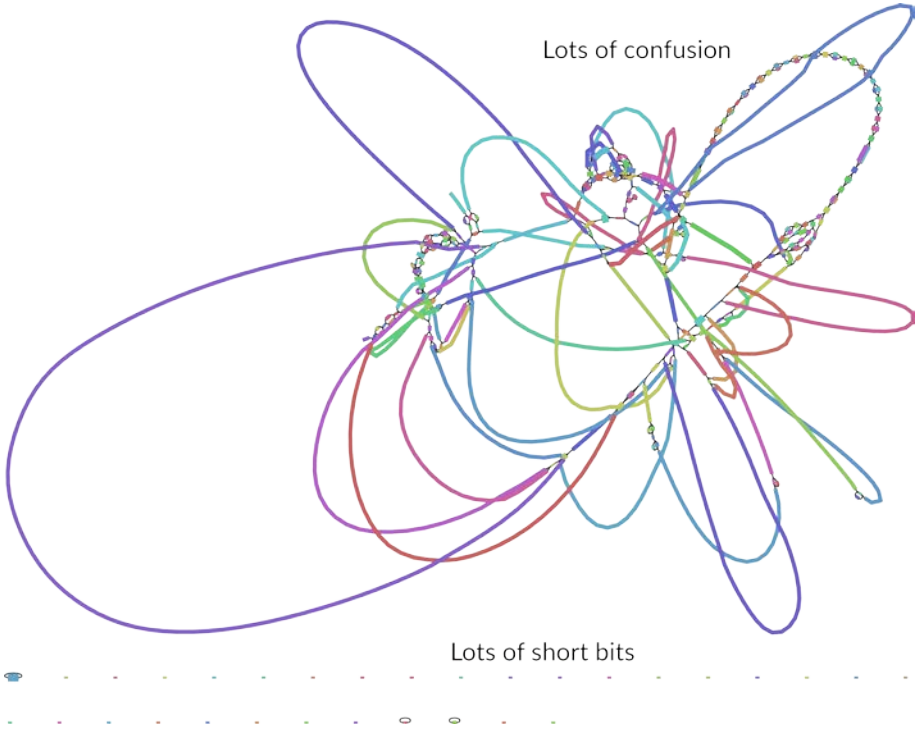
Nanopore Sequencing

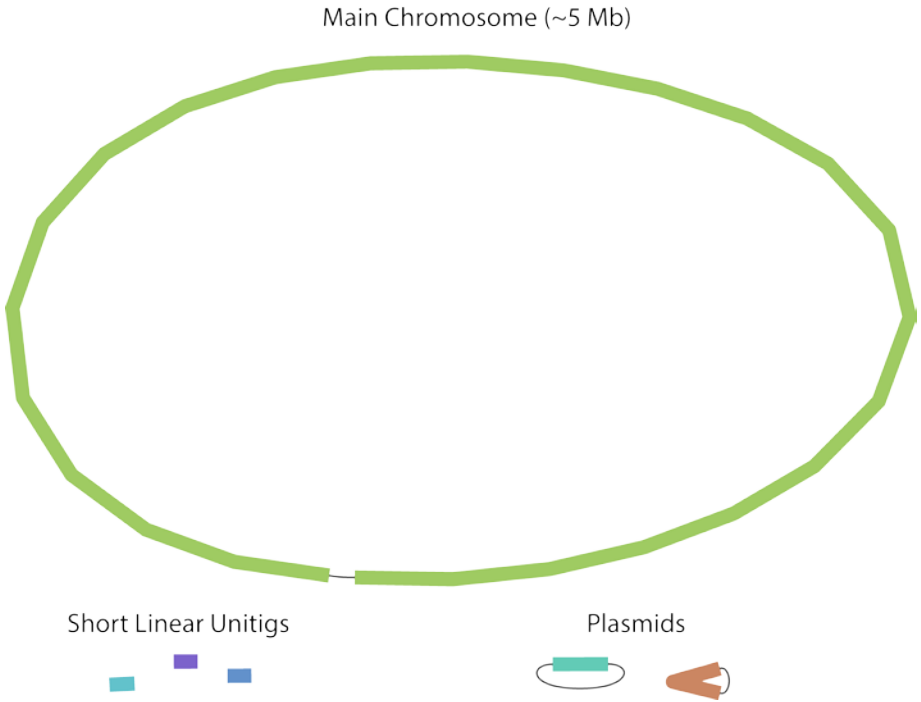
Library prep essentially consists of two ligation steps, so small circular DNA are easily missed.

Shearing (to about 10kb) is necessary to linearize plasmids.



Very Contiguous Assemblies





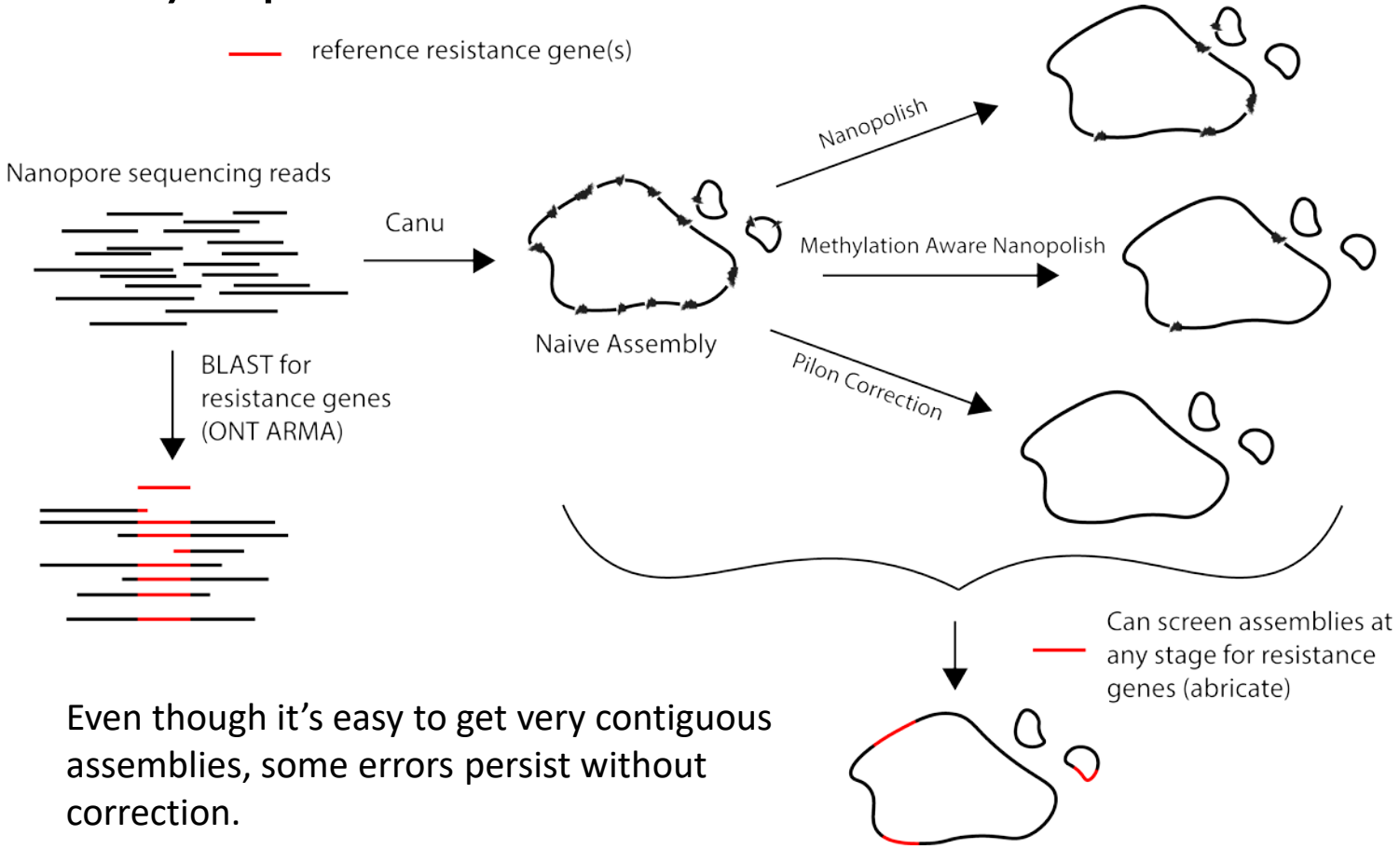
Only Illumina reads

(~200X coverage, v2 150bp PE)

Only Nanopore reads

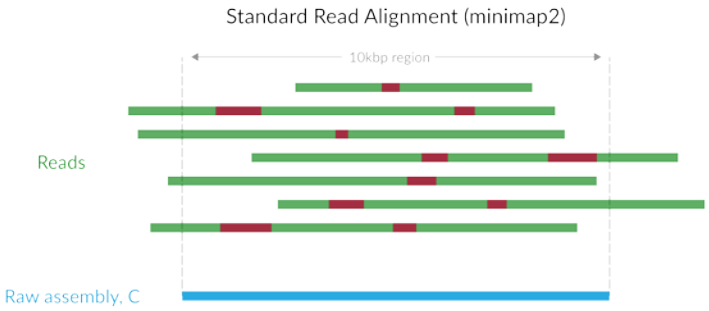
(~2000X coverage, r9.4 flowcell)

Assembly Pipeline

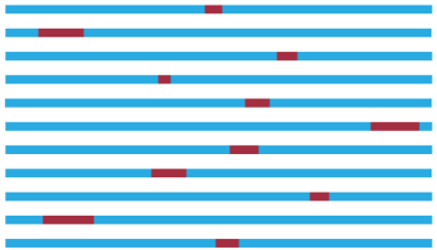


Even though it's easy to get very contiguous assemblies, some errors persist without correction.

Nanopolish Consensus

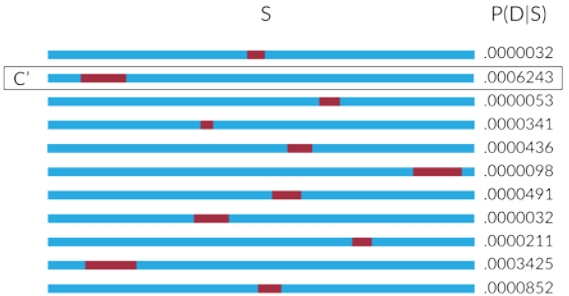


Based on non-matching stretches of sequence between the assembly and the reads and all possible 1bp changes, generate a list of candidate improvements to C, called S

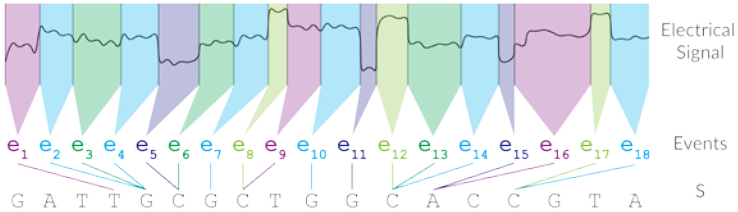


Sum over all possible alignments and compute the probability of observing the event data given S, P(D|S)

Set the best S as the new C, and repeat



Choose the S that maximizes the probability of observing the event data

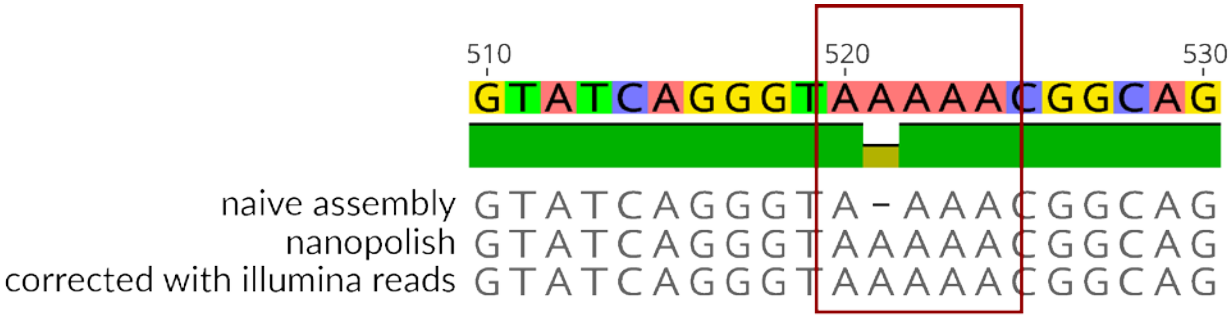


Use raw electrical data to correct assemblies.

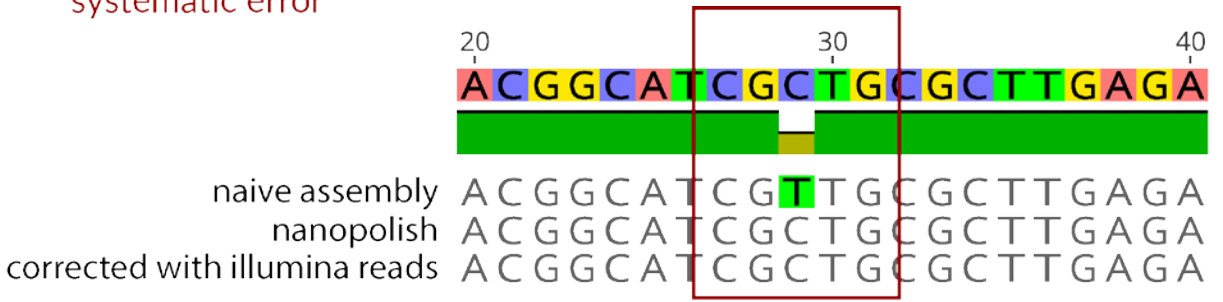
Jared Simpson: https://github.com/jts/nanopolish

Nanopolish Consensus

Homopolymers are systematically mis-called on this platform, causing indels and difficulty in determining potential phenotypic consequences.



nanopolish fixes most homopolymers indels, the most prominant type of systematic error



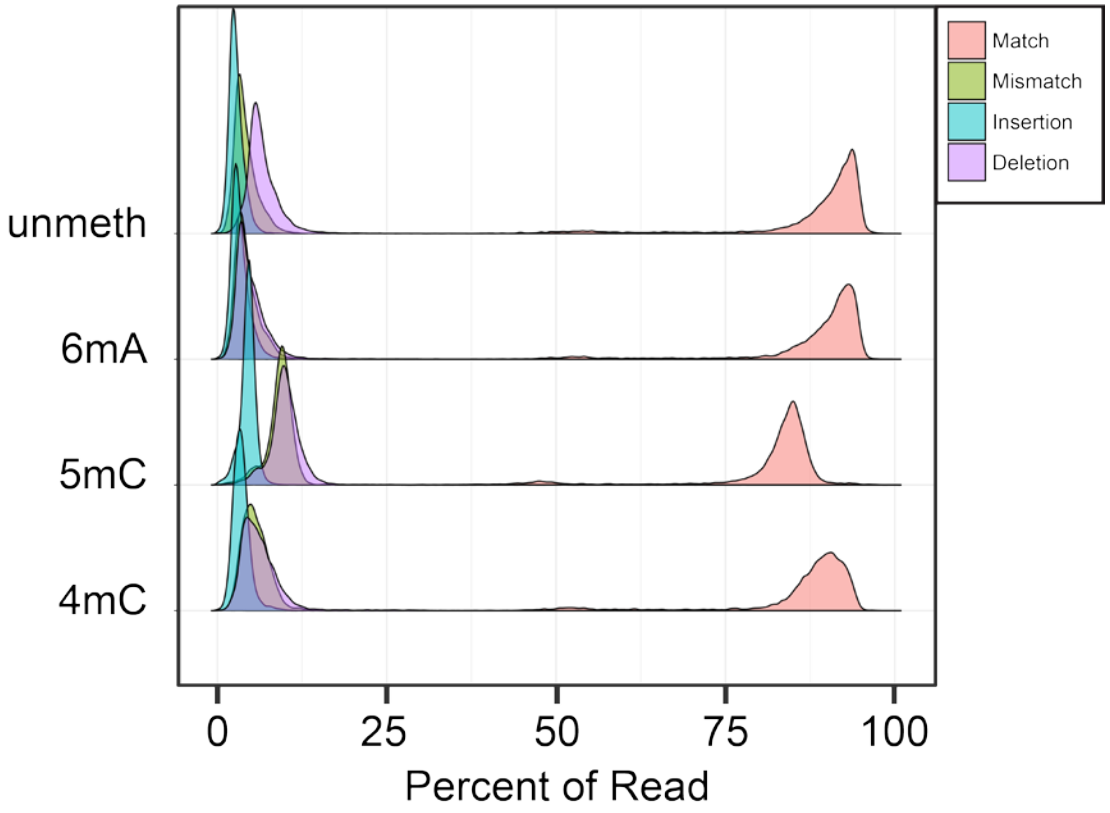
nanopolish fixes most random errors, not associated with homopolymers or methylation

Methylation Associated Error

Because the sequencer physically interrogates DNA molecules, chemical modifications like methyl marks also cause systematic error.

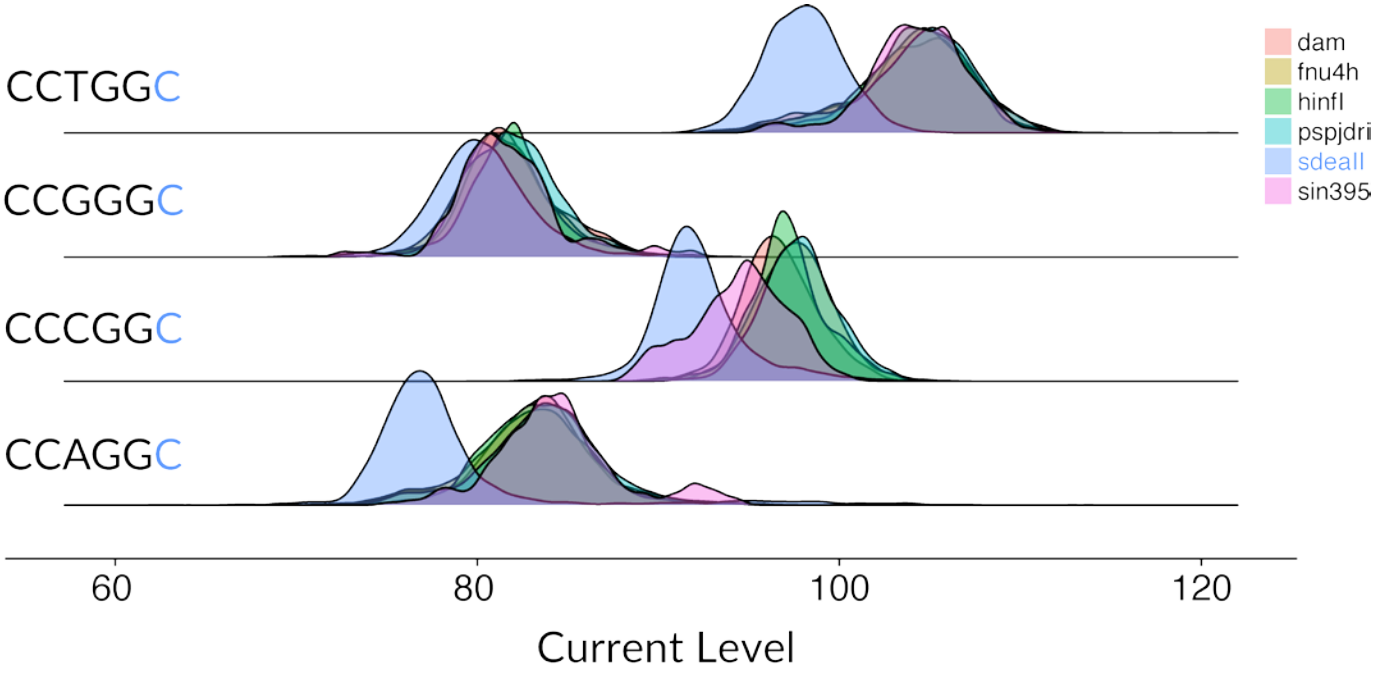
E. coli DNA with controlled methylation types cause an increase in non-matches when reads are aligned to a reference.

DNA samples courtesy of NEB



Blow et al *PLoS Genetics* (Feb 2016)

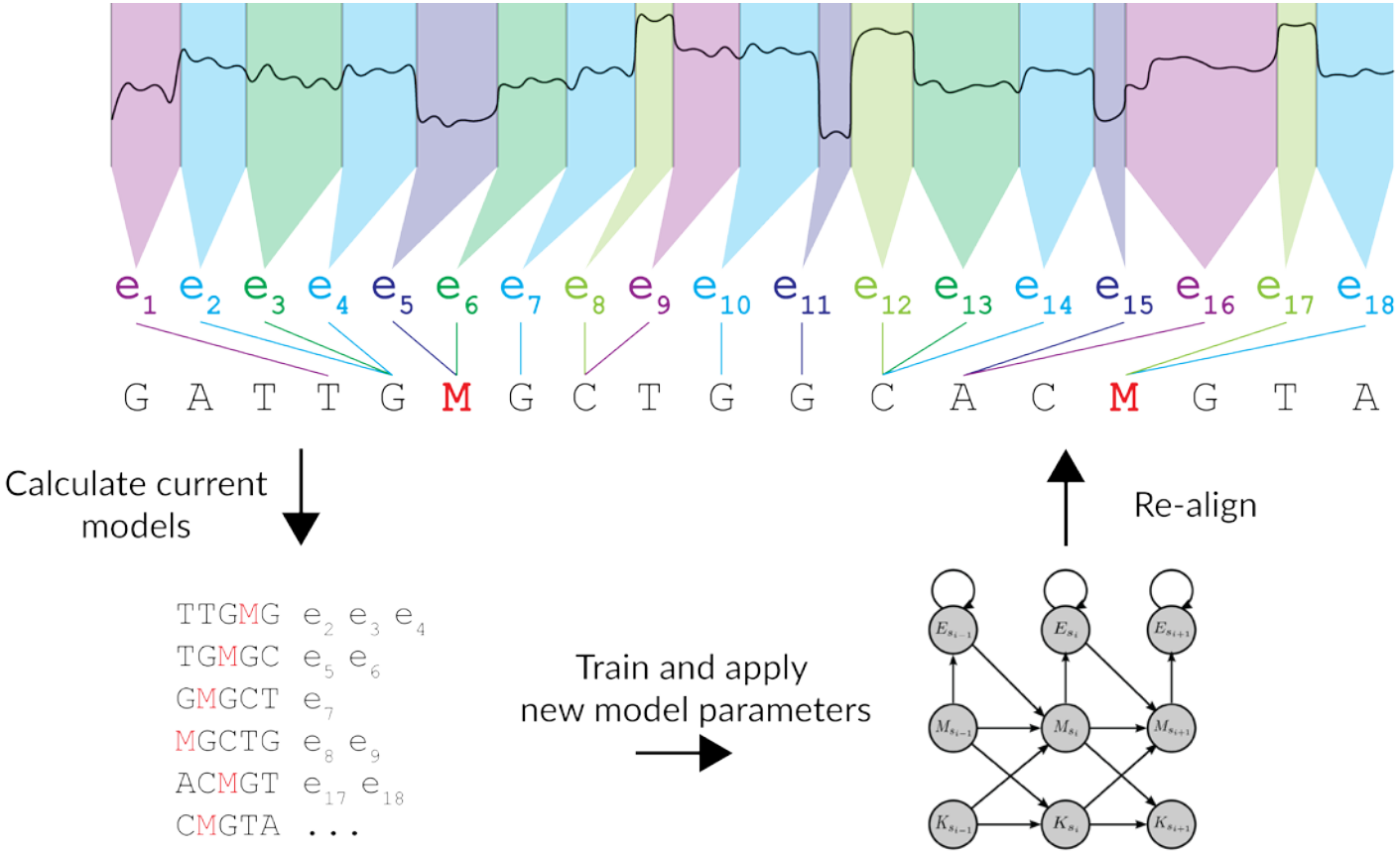
Methylation Associated Error



Distinct changes in current distribution at methylation motifs can cause basecalling errors in all the reads, which persist through assembly.

Methylation Associated Error

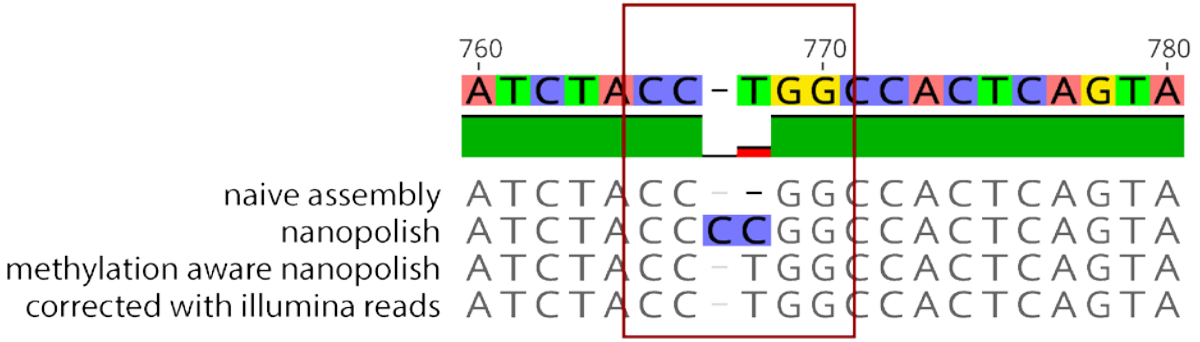
We can address this problem by training models specifically for methylation motifs, using a similar HMM scheme to align electrical data to a reference.



http://simpsonlab.github.io/2017/10/31/methylation-aware/

Methylation Aware Nanopolish

With methylation aware correction, nanopore-only assemblies can achieve in the range of 99.8% identity with assemblies corrected using Illumina reads.

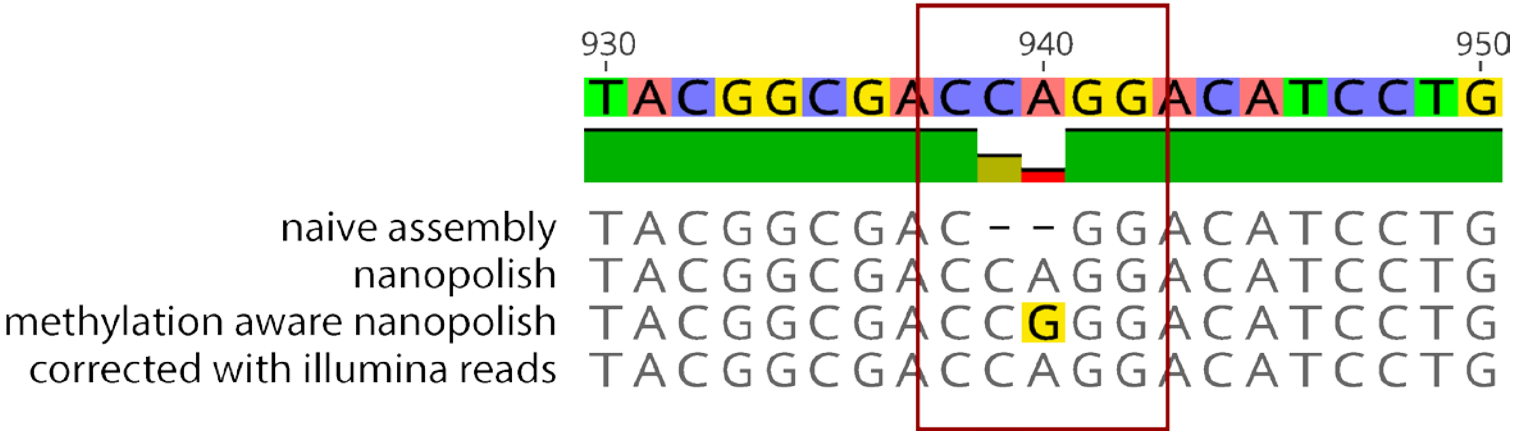


models need to be trained on methylated k-mers in order to correct MTase motifs (dcm MTase= CCWGG)

Raw Assembly	Nanopolish Corrected	Methylation Aware Nanopolish Corrected
98.89%	99.57%	99.76%

http://rebase.neb.com/rebase/rebase.html

Methylation Aware Nanopolish



Not a perfect correction scheme yet. We need a more universal training set in order to attempt to address all possible methylation associated errors.

NB: methylation-aware mode is still experimental

Consequential Mutations

```
>KLPN_133_tig00000001:3594157-3595237- coverage:100.00 score=2120 edit_distance=6
D   115   39(M) A:    115 GAAAATGGTC   AAAAATGGT  homopolymer  premature stop at AA63
I   123   41(G) :G    123 GTCGGgCGAGC  TCGGCGAGC  [unknown]    premature stop at AA89
S   294   98(L) G:A   295 CGTCTaGCGTT  GTCTGGCGT  [unknown]    synonymous
S   688  230(W) T:C   689 AAGCCcGGGCG  AGCCTGGGC  motif CCTGG  AA230|W:R
D   946  316(W) T:    946 GTACCGGTAC   TACCTGGTA  motif CCTGG  premature stop at AA323
I  1049  350(Q) :G   1049 CGACCgAGGCG  GACCAGGCG  motif CCAGG  frame shift QAAV+:RGGR+
```

1. event type
2. WT base loc
3. Amino acid
4. base changes
5. ctg base loc
6. ctg context (5+m+5)
7. WT context (4+1+4)
8. context status
9. mutation impact

Geo Pertea: https://github.com/gpertea/pwasm (under development)

Consequential Mutations

```
>KLPN_154_tig00000001_pilon:2586451-2589085- coverage:100.00 score=5196 edit_distance=12
S       247     83(S)   TC:AT   249     CGACatCGCGG     CGACTCCGC       [unknown]       AA83|S:I
S       552     184(P)  A:G     553     CCGCCgCATAA     CGCCACATA       [unknown]       synonymous
S       699     233(Y)  T:C     700     GCCTAcCGCAC     CCTATCGCA       [unknown]       synonymous
S       726     242(I)  C:T     727     TACATtCGCGC     ACATCCGCG       [unknown]       synonymous
S       1167    389(N)  C:T     1168    GCCAAtATCGA     CCAACATCG       [unknown]       synonymous
S       1251    417(L)  T:C     1252    GATCTcGGTAA     ATCTTGGTA       [unknown]       synonymous
S       1260    420(V)  C:T     1261    AACGTtGCGGC     ACGTCGCGG       [unknown]       synonymous
S       1329    443(V)  A:G     1330    GGCGTgCGCGA     GCGTACGCG       motif GTA       synonymous
S       1347    449(Y)  T:C     1348    TACTAcCTGAC     ACTATCTGA       [unknown]       synonymous
S       1458    486(A)  G:A     1459    ATCGCaGAGCT     TCGCGGAGC       [unknown]       synonymous
S       2502    834(V)  A:G     2503    GTGGTgGGTCT     TGGTAGGTC       [unknown]       synonymous
>KLPN_162_tig00000001_pilon:2487031-2489663- coverage:100.00 score=5196 edit_distance=12
S       247     83(S)   TC:AT   249     CGACatCGCGG     CGACTCCGC       [unknown]       AA83|S:I
S       259     87(D)   G:T     260     TATACtACACC     ATACGACAC       [unknown]       AA87|D:Y
S       552     184(P)  A:G     553     CCGCCgCATAA     CGCCACATA       [unknown]       synonymous
S       1050    350(I)  T:C     1051    GACATcATCGC     ACATTATCG       [unknown]       synonymous
S       1251    417(L)  T:C     1252    GATCTcGGTAA     ATCTTGGTA       [unknown]       synonymous
```

```
stance=12
n]          AA83|S:I
n]          synonymous
n]          synonymous
n]          synonymous
n]          synonymous
```

```
stance=12
n]          AA83|S:I
n]          AA87|D:Y
n]          synonymous
n]          synonymous
n]          synonymous
```

In gyrA, mutations in amino acid 83 are associated with fluoroquinolone resistance, and we can detect these in our isolates.

Future Directions

Better training for methylation polishing

Local assemblies or alignment consensus
to speed up pipeline for point of care use.

Acknowledgements

Timp Lab
Dr. Winston Timp
Stephanie Hao
Rachael Workman
Norah Sadowski
Isac Lee
Norah Sadowski
Tim Gilpatrick
Roham Razaghi
Evi Mercken

Salzberg Lab
Geo Pertea

Simner Lab
Dr. Trish Simner
Yehudit Bergman

Simpson Lab
Dr. Jared Simpson
Dr. P.C. Zuzarte
Dr. Matei David
Dr. L. J. Dursi

National Human Genome Research Institute
1R01HG009190-01A1

National Institute of Allergy and Infectious Diseases
1R21AI130608-01 (Simner)

CIHR IRSC
Canadian Institutes of Health Research
Instituts de recherche en santé du Canada

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Thank you!

Me: Yunfan Fan
yfan7@jhu.edu

Timp Lab
@timp0
Slides will be available at www.timplab.org