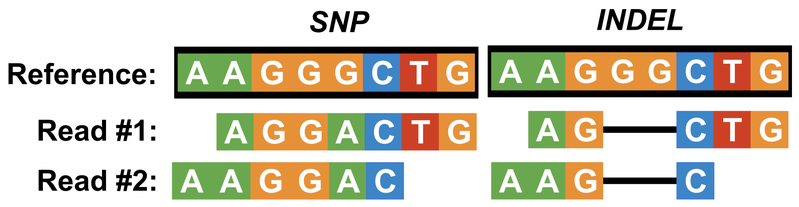
*n*PoRe: *n*-Polymer Realigner for improved pileup-based variant calling

Tim Dunn, David Blaauw, Reetu Das, Satish Narayanasamy — University of Michigan, CSE

Background:

Variant calling requires the identification of mutations present in sequenced reads relative to an expected reference genome. Small variants are classified into Single Nucleotide Polymorphisms, or SNPs, and Insertions/Deletions, or INDELs.



Although nanopore sequencing can identify SNPs with high accuracy, a current weakness is recalling small INDELs.

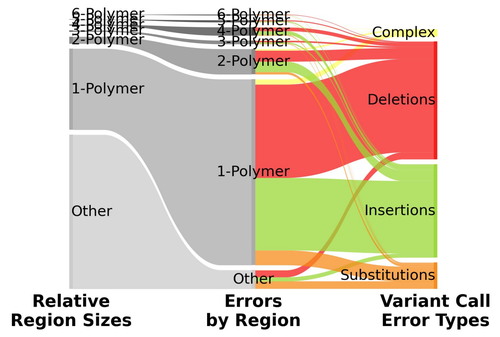
Nanopore R9.4.1 Variant Calling Accuracy [1]

SNPs: 99.7% precision, 99.7% recall

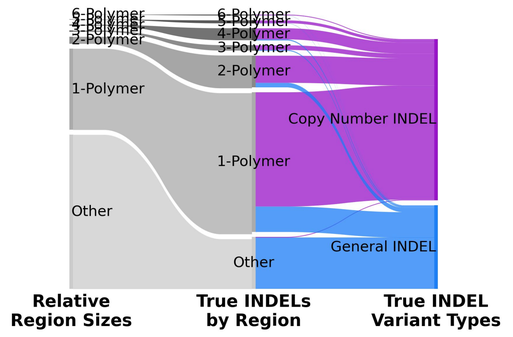
INDELs: 92.8% precision, **76.0% recall**

Motivation:

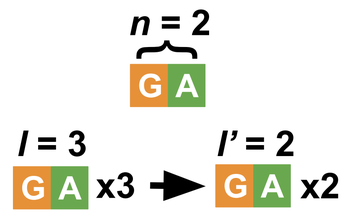
Nanopore sequencers frequently mis-call the length of repeated low-complexity DNA sequences. Existing aligners do not take this failure mode into account, leading to **inconsistent alignments** which complicate variant calling.



Most variant calling errors can be attributed to INDEL errors within n-polymer regions.



Most true INDEL mutations are variations in the number of copies of an n-polymer.



We define:

- ***n***, the length of the repeated unit
- ***I***, the expected number of repeated copies
- ***I'***, the measured number of repeated copies
- **"n-polymer"**: repeated sequence where $n \leq 6$ and $I \geq 3$

Algorithm: Baseline: Affine Gap Smith-Waterman

$G_{open} = 5$

$G_{extend} = 2$

S

Reference \ Basecall	A	C	G	T
A	0.01	6.16	4.82	6.08
C	6.06	0.01	6.25	4.78
G	4.79	6.25	0.01	6.05
T	6.07	4.81	6.13	0.01

D ■

i \ j-1	j-1	j
i	$\blacksquare G_{extend}$	$\blacksquare \star_0$

I ♦

i-1 \ j	j
i-1	$\blacklozenge G_{extend}$
i	$\blacklozenge \star_0$

M ★

i-1 \ j-1	j-1	j
i-1	$\star S[R[j], r[i]]$	$\blacklozenge G_{open}$
i	$\blacksquare G_{open}$	★

Extension: n-Polymer Realignment

1. Generate n-polymer reference annotations

Reference:	A	T	A	T	A	T	A	T	T	T	T	A	A	A	G	C	G	C	G	C	
n=1	***l:***	0	0	0	0	0	0	0	5	5	5	5	5	3	3	3	0	0	0	0	0	0
	idx:	0	0	0	0	0	0	0	1	2	3	4	0	1	2	0	0	0	0	0	0	0
n=2	***l:***	4	3	4	3	4	3	4	0	0	0	0	0	0	0	0	3	0	3	0	3	0
	idx:	0	0	1	1	2	2	3	0	0	0	0	0	0	0	0	0	0	1	0	2	0

2. Define conditions for valid n-polymer INDELs

$c_1 = l > 0$ — start of repeat unit

$c_2 = l > 0 \text{ and } idx == 0$ — start of n-polymer

$c_3 = R[j+1 : j+1+n] == r[i+1 : i+1+n]$ — next n bases of r match R

3. Calculate n-polymer INDEL cost matrix

$$N[n,i,j] \approx -\log \mathbb{P}(n,i,j) \approx -\log \frac{C_N[n,i,j] + \epsilon}{\text{sum}(C_N[n,i,:]) + \epsilon}$$

4. Extend alignment algorithm to allow n-polymer INDELs

D ■

i \ j-1	j-1	j
i	$\blacksquare G_{extend}$	$\blacksquare \star_0$

I ♦

i-1 \ j	j
i-1	$\blacklozenge G_{extend}$
i	$\blacklozenge \star_0$

M ★

i-1 \ j-1	j-1	j
i-1	$\star S[R[j], r[i]]$	$\blacklozenge G_{open}$
i	$\blacksquare G_{open}$	$\star$ $\bullet N[n,l,l-1]$ if c_2 $\blacktriangle N[n,l,l+1]$ if c_2, c_3

L ▲

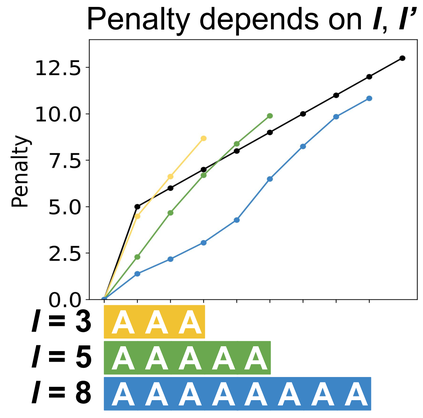
i-run \ j	j
i-run	$\blacktriangle N[n,l,1+l+run/n]$ if c_1, c_3
i-1	
i	$\star_0$
i+n	▲

S ●

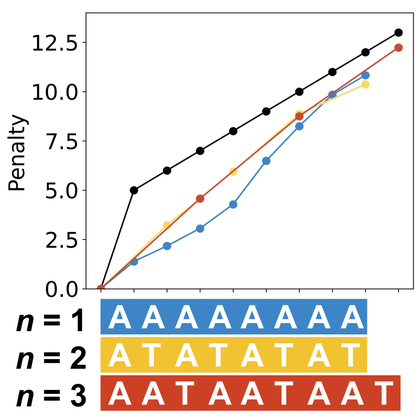
i \ j-run	j-run	···	j-1	j	···	j+n
i	$\bullet N[n,l,l-1-run/n]$ if c_1	···		$\star_0$	···	●

Gap (Deletion) Penalties:

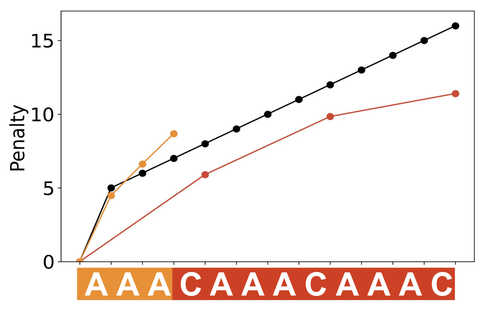
Penalty depends on *I*, *I'*



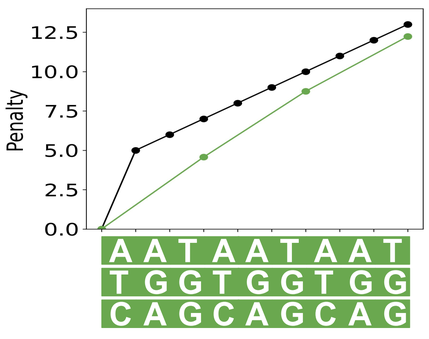
Penalty depends on *n*, only fully copy variations allowed



Overlaps are allowed

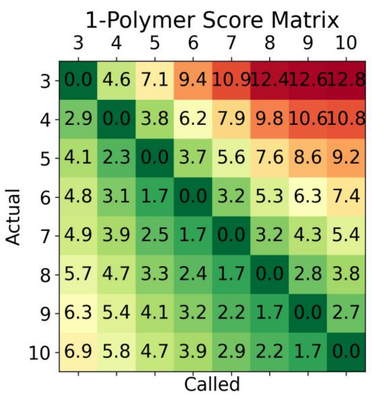


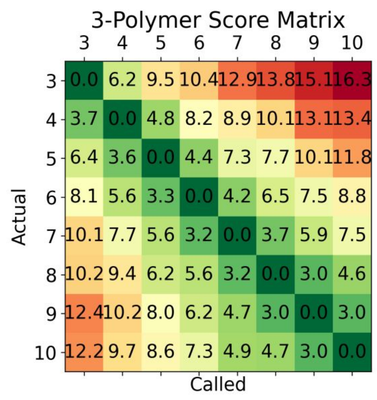
Sequence is not considered

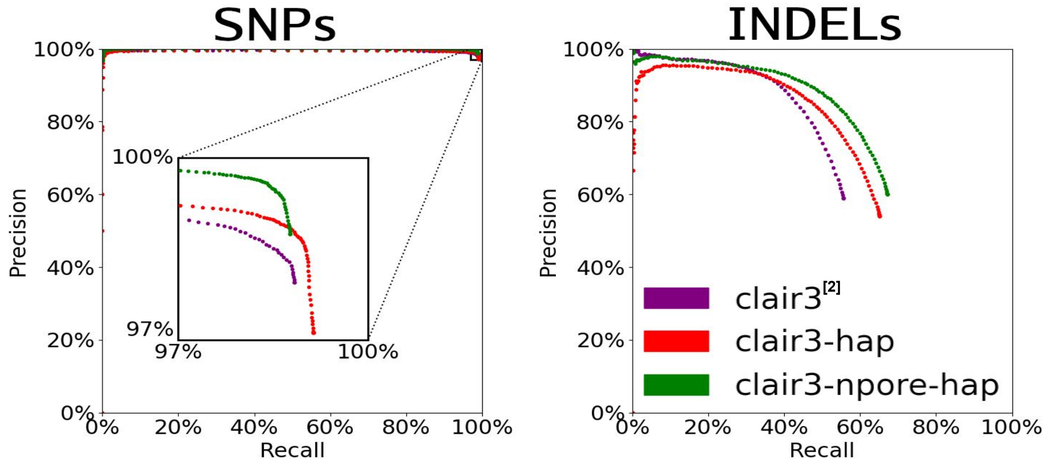


Results:

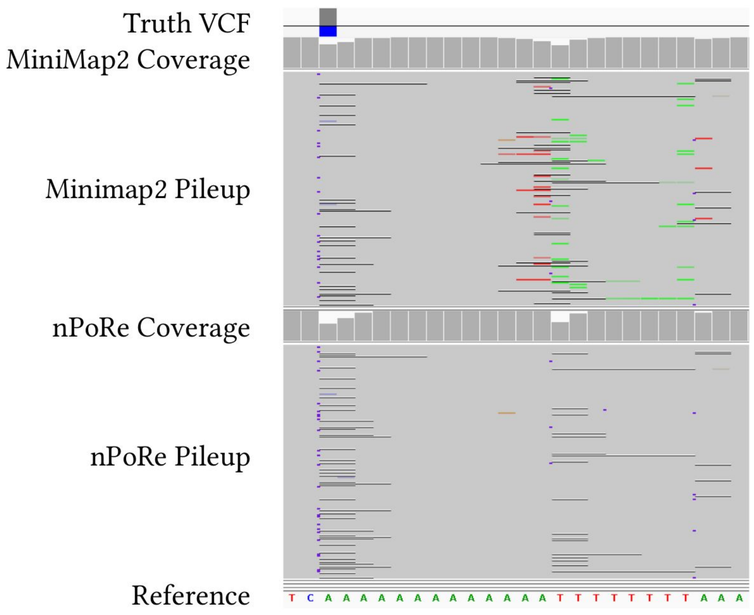
Negative log probability cost matrices (***N***) for n-polymer INDELs. In general, INDELs were more common for lower ***n***.







Realignment improved both INDEL and SNP pileup-based variant calling accuracy. Unfortunately, after using Clair3's full alignment model, similar accuracy improvements were not observed.



nPoRe realignment results in greater pileup concordance in comparison to MiniMap2[3].

In this example, nPoRe always aligns errors as two separate homopolymer length mis-calls. MiniMap2 sometimes calls substitutions instead and merges INDELs where possible, resulting in an inconsistent pileup.

More Information:

This project was supported by the Kahn Foundation, NSF Grant 2030454, and NSF Graduate Research Fellowship 1841052.

GitHub Code
https://github.com/TimD1/nPoRe

bioRxiv Pre-Print
https://www.biorxiv.org/content/10.1101/2022.02.15.480561v1

References

[1] Shafin, et. al. "Haplotype-aware variant calling with PEPPER Margin DeepVariant enables high accuracy in nanopore long-reads." Nature Methods, 2021. *https://github.com/kishwarshafin/pepper*

[2] Zheng, et. al. "Symphonizing pileup and full-alignment for deep learning-based long-read variant calling." bioRxiv, 2022.

[3] Li, Heng. "Minimap2: pairwise alignment for nucleotide sequences."