MICHIGAN ENGINEERING
UNIVERSITY OF MICHIGAN

vcfdist: accurately benchmarking phased small variant calls

Tim Dunn, Satish Narayanasamy

Introduction: VCF Benchmarking

- With many new emerging sequencing technologies and methods, it's important to accurately assess the relative performance of each option
- This is done by comparing the set of variants called to a known ground truth set of variant calls in VCF format

Technology #1
Reference: ACCGTTGAAG
Query #1: ACAGTAGAAG

	CHROM	POS	REF	ALT
✗	chr14	3	C	A
✓	chr14	6	T	A

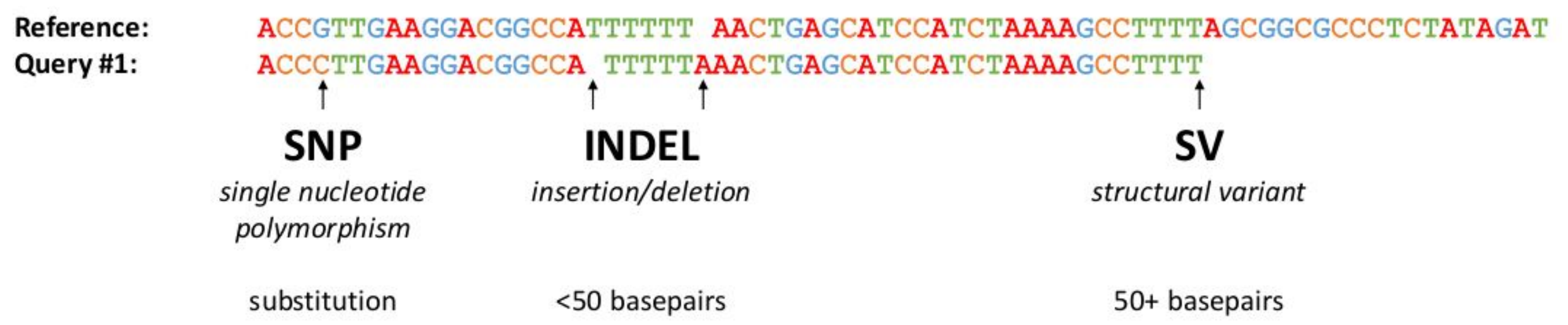
Technology #2
Reference: ACCGTTGAAG
Query #2: ACCGTAGAGG

	CHROM	POS	REF	ALT
✓	chr14	6	T	A
✓	chr14	9	A	G

Ground Truth
Reference: ACCGTTGAAG
Query: ACCGTAGAGG

CHROM	POS	REF	ALT
chr14	6	T	A
chr14	9	A	G

- Performance metrics such as precision and recall are reported independently for each category of variant


Reference: ACCGTTGAAGGACGGCCATTTTTT AACTGAGCATCCATCTAAAAGCCTTTTAGCGGCGCCCTCTATAGAT
Query #1: ACCCTTGAAGGACGGCCA TTTTTAAACTGAGCATCCATCTAAAAGCCTTTT

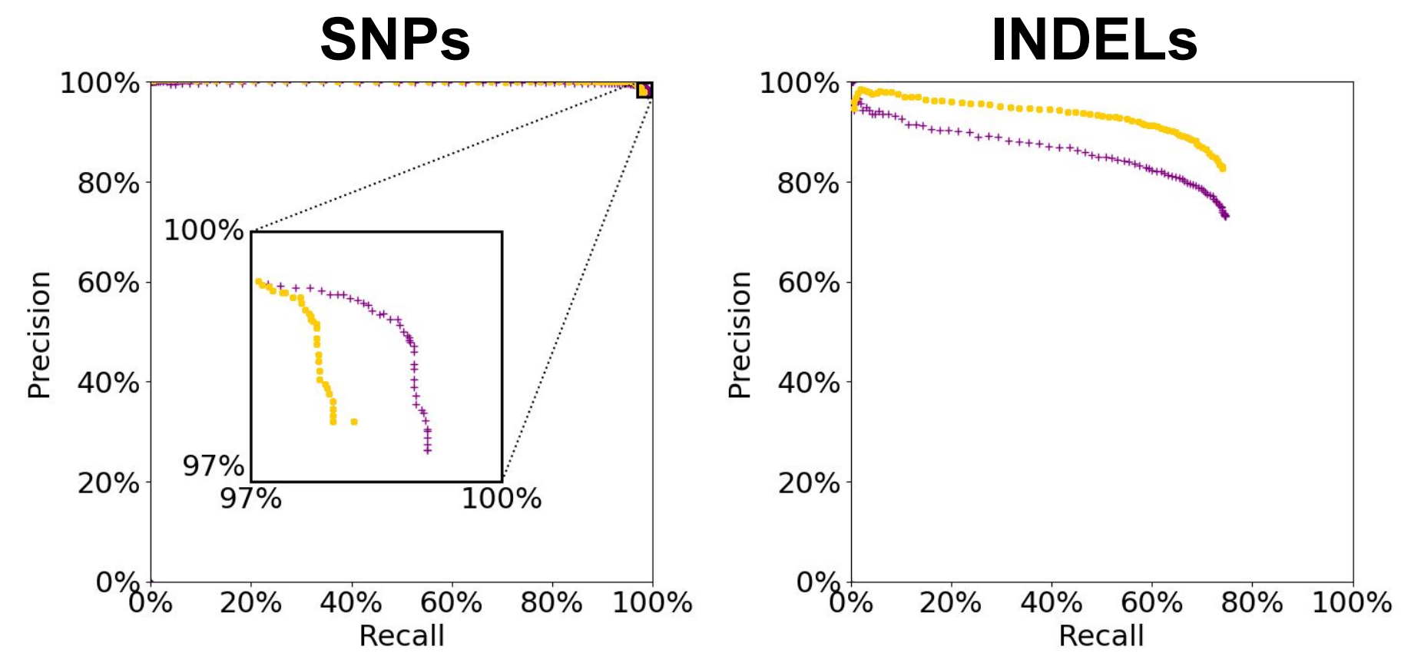
SNP
single nucleotide polymorphism
substitution

INDEL
insertion/deletion
<50 basepairs

SV
structural variant
50+ basepairs

Motivation

- Generating SNP/INDEL precision-recall curves for phased VCFs using vcfeval[1] does not result in stable evaluations
- The following figure shows benchmarking results for two VCFs which both encode the exact same underlying query sequence, but prefer different variant representations

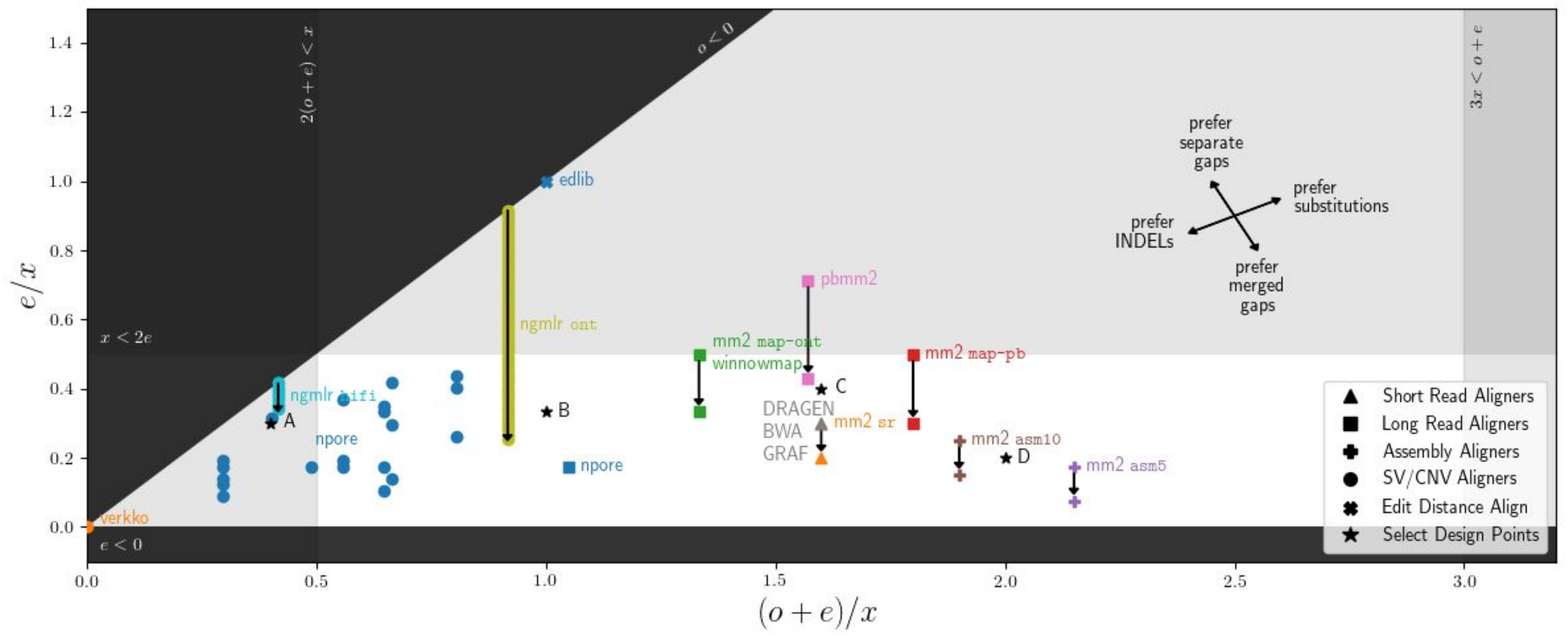


1. Variant Clustering

- The human genome is prohibitively large for exact alignment of entire chromosomes, and must be partitioned into independent subproblems for efficient analysis
- We present an alignment-based algorithm for identifying potential variant dependencies and grouping variants into independent clusters

2. Variant Normalization

- We present a thorough exploration of "Best Alignment Normalization" [2] affine-gap parameters



Reference AGGCGACA   Query ATACCGAGCTTA

Point A
$m, x, o, e = 0, 10, 1, 3$

Alignment

```
AGG---CGA-C--A
.     ... .  .
A--TACCGAGCTTA
```

VCF

POS	REF	ALT
1	AGG	A
3	G	GTAC
6	A	AG
7	C	CTT

Point B
$m, x, o, e = 0, 3, 2, 1$

```
A-GGCGA-C--A
.   ... .  .
ATACCGAGCTTA
```

POS	REF	ALT
1	A	AT
2	G	A
3	G	C
6	A	AG
7	C	CTT

Point C
$m, x, o, e = 0, 5, 6, 2$

```
A-GGCGA---CA
.   ...    .
ATACCGAGCTTA
```

POS	REF	ALT
1	A	AT
2	G	A
3	G	C
6	A	AGCT
7	C	T

Point D
$m, x, o, e = 0, 5, 9, 1$

```
AGGCGAC-------A
.    ..       .
A---TACCGAGCTTA
```

POS	REF	ALT
1	AGGC	A
5	G	T
7	C	CCGAGCTT

3. Enforce Local Variant Phasing

- Previous work vcfeval[1] was designed for short read variant calls and assumes all variant calls are unphased
- For vcfdist, local phasing is enforced within each cluster of variants and arbitrary phase swaps are allowed to occur between clusters

4. Partial Credit for Variant Calls

- Using a novel alignment algorithm, we can assign partial credit to mostly-correct calls
- This allows inexact matches for long or complex variants

Ref. ACCCTTTTTTG   **Query** ACCTTTG   **Truth** ACCCTTTG

Query VCF Representation 1

POS	REF	ALT
3	CCTTT	C

Query VCF Representation 2

POS	REF	ALT
1	AC	A
4	CTTT	C

Truth VCF

POS	REF	ALT
4	CTTT	C

vcfeval Summary Statistics

	TP	FP	FN	PP	Prec.	Recall	F1
Query Repr. 1	0	1	1	0	0.00	0.00	0.00
Query Repr. 2	1	1	0	0	0.50	1.00	0.67

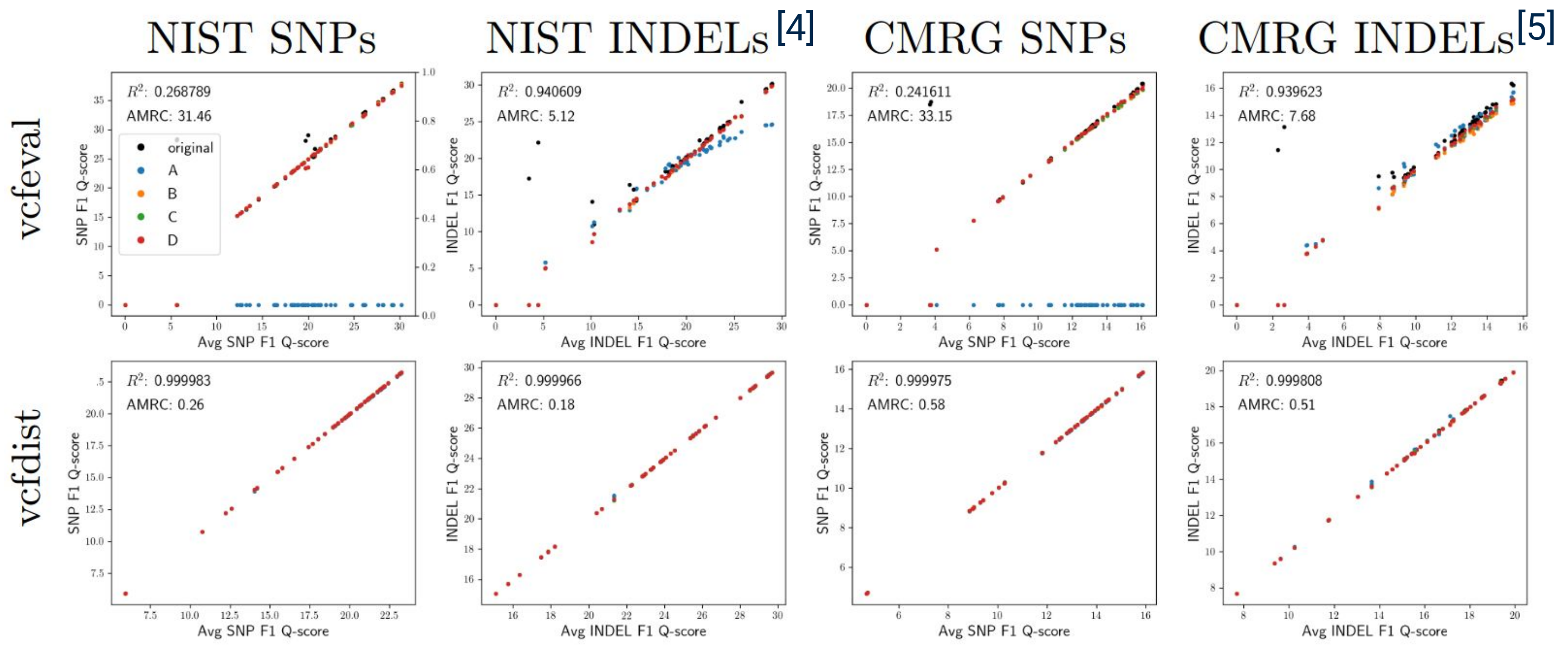
vcfdist Summary Statistics

	TP	FP	FN	PP	Prec.	Recall	F1
Query Repr. 1	0	0	0	1	0.67	0.67	0.67
Query Repr. 2	1	1	0	0	0.50	1.00	0.67

5. Alignment Distance Metrics

- Since SNP and INDEL variants are ill-defined in the case of complex variants, we can avoid classifying variants and simply align the query and truth sequences
- High-level summary information regarding the alignment can then be used to evaluate performance
  - **Edit Distance:** the total number of bases which differ
  - **Distinct Edits:** the total number of variants
  - **Alignment Distance:** the affine-gap alignment score

Results[3]



References

[1] Cleary et al. *"Comparing variant call files for performance benchmarking of next-generation sequencing variant calling pipelines"*. bioRxiv, 2015.
[2] Bayat et al. *"Improved VCF normalization for accurate VCF comparison"*. Bioinformatics, 2017.
[3] Olson et al. *"PrecisionFDA Truth Challenge V2: Calling variants from short and long reads in difficult-to-map regions"*. Cell Genomics, 2022.
[4] Wagner et al. *"Benchmarking challenging small variants with linked and long reads"*. Cell Genomics, 2022.
[5] Wagner et al. *"Curated variation benchmarks for challenging medically relevant autosomal genes"*. Nature Biotechnology, 2022.

Additional Information

Funding: This project was supported by the National Science Foundation Graduate Research Fellowship under Grant 1841052. Any opinion, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Paper: doi.org/10.1101/2023.03.10.532078

Code: github.com/TimD1/vcfdist