vcfdist: accurately benchmarking phased variant calls

Tim Dunn

PhD Candidate

University of Michigan

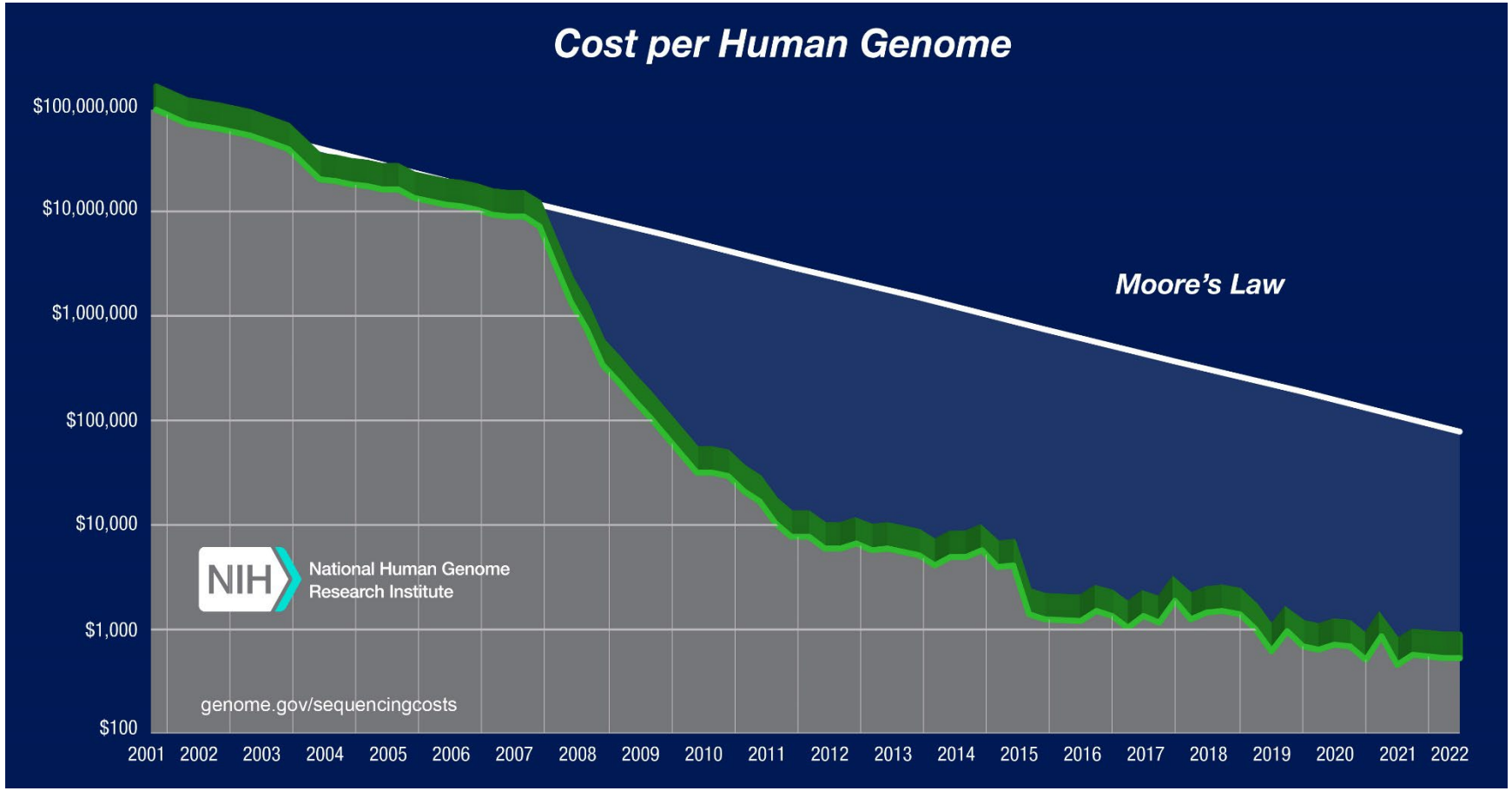
Outline

1. **Context**
2. **Problem**
3. **Discussion**
4. **Solution**
5. **Implementation**
6. **Results**
7. **Next Steps**

Outline

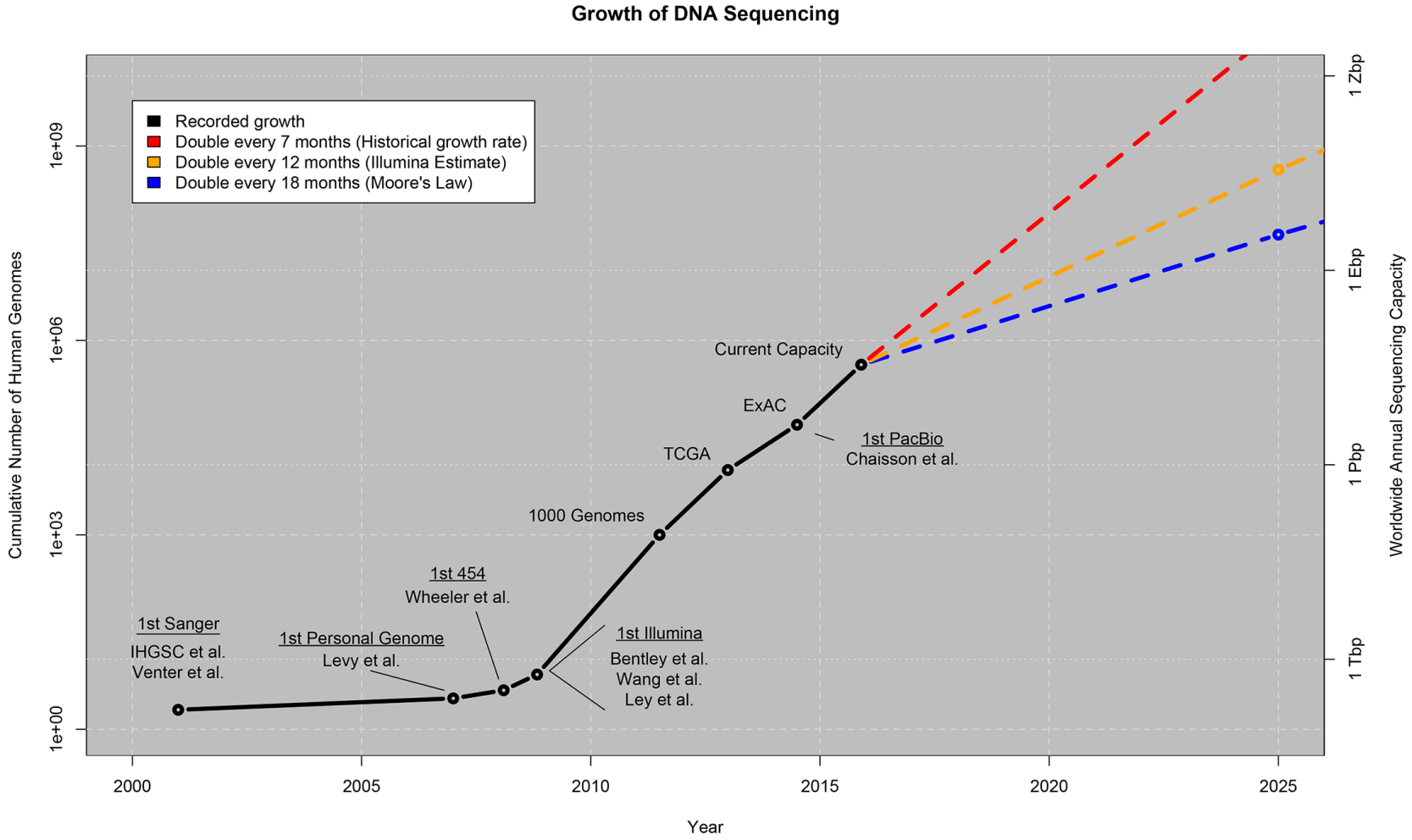
1. **Context:** whole genome sequencing evaluation
2. **Problem**
3. **Discussion**
4. **Solution**
5. **Implementation**
6. **Results**
7. **Next Steps**

Sequencing: *cost is rapidly declining*



NHGRI. "DNA Sequencing Costs: Data". https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data, 2023.

Sequencing: *exponential growth in genomes*



Stephens et al. “Big Data: Astronomical or Genomical?”.PLOS Biology, 2015.

Applications

- Genome wide association studies
- Pharmacogenomics
- Clinical diagnostics
- Benchmarking new methods and tech

Applications: *genome comparison required*

- Genome wide association studies
- Pharmacogenomics
- Clinical diagnostics
- Benchmarking new methods and tech

Comparison: *genomes are mostly identical*

Reference: ACCGTTGAAGGACGGCCATTTTTT AACTGAGCATCCATCTAAAAGCCTTTTAGCGGCGCCCTCTATAGAT

Query #1: ACCCTTGAAGGACGGCCA TTTTTAAACTGAGCATCCATCTAAAAGCCTTTT

Variant Call Format: *difference-based*

POSITION	REFERENCE	ALTERNATE
4	G	C
18	AT	A
25	T	TA
53	TAGCGGCGCCC…	T

Applications: *benchmarking*

- Genome wide association studies
- Pharmacogenomics
- Clinical diagnostics
- **Benchmarking new methods and tech**

Applications: *benchmarking*

- Genome wide association studies
- Pharmacogenomics
- Clinical diagnostics
- **Benchmarking new methods and tech**

Belton et al, “Hi-C: a comprehensive technique to capture the conformation of genomes”. Nature Methods, 2012.
DelveInsight Business Research. “Global DNA Sequencing Market Set to Reach USD 28.85 billion by 2027”. Web, 2022.

Benchmarking: *a simple example*

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

Benchmarking: *a simple example*

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

Benchmarking: *a simple example*

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

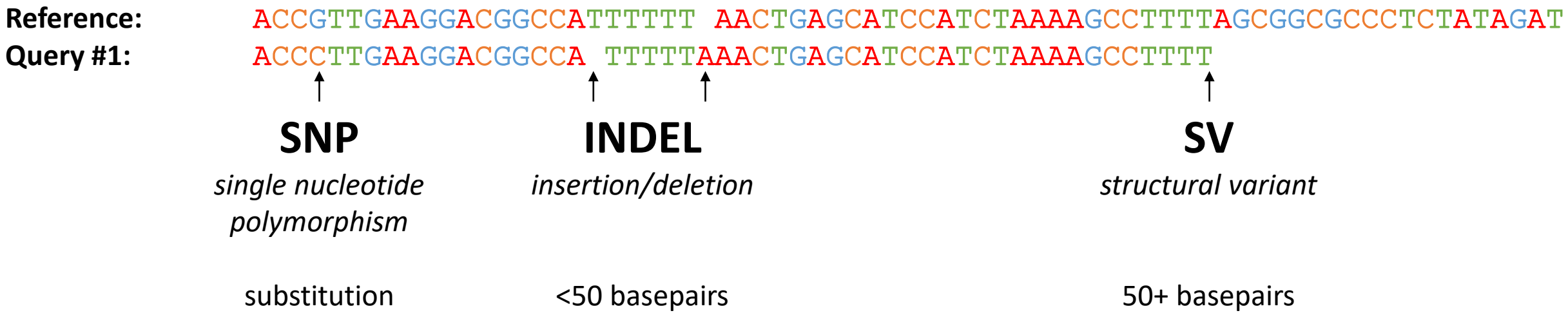
	CHROM	POS	REF	ALT
✓	chr14	6	T	A
✓	chr14	9	A	G

Ground Truth

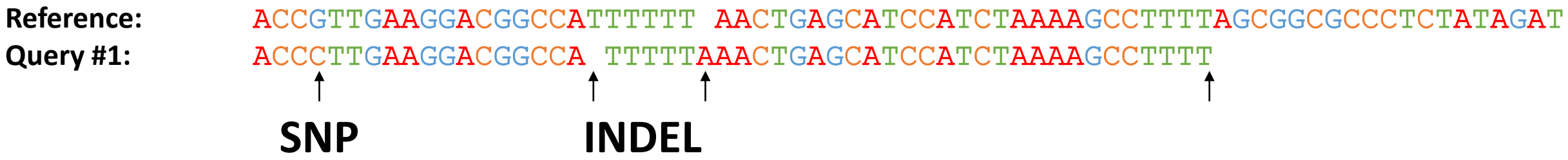
Reference: ACCGTTGAAG
Query: ACCGTAGAGG

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

Benchmarking: *stratification by variant type*



Benchmarking: *precision-recall curves*



$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

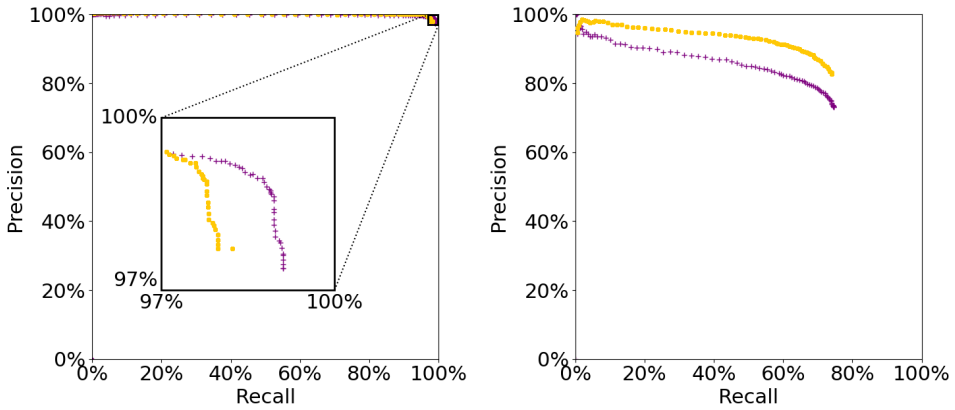
$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Benchmarking: *precision-recall curves*

Reference: ACCGTTGAAGGACGGCCATTTTTT AACTGAGCATCCATCTAAAAGCCTTTTAGCGGCGCCCTCTATAGAT

Query #1: ACCCTTGAAGGACGGCCA TTTTTAAACTGAGCATCCATCTAAAAGCCTTTT

SNP **INDEL**



$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Outline

1. **Context**
2. **Problem:** complex variants
3. **Discussion**
4. **Solution**
5. **Implementation**
6. **Results**
7. **Next Steps**

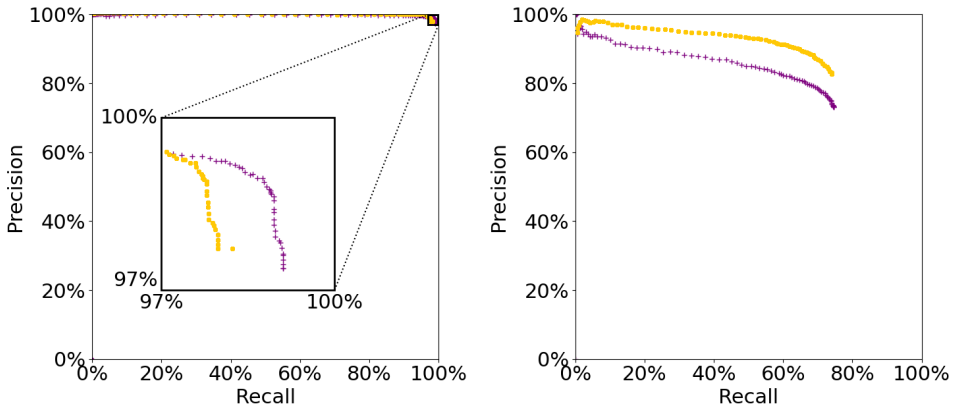
Problem: *evaluation consistency*

Reference: ACCGTTGAAGGACGGCCATTTTTT AACTGAGCATCCATCTAAAAGCCTTTTAGCGGCGCCCTCTATAGAT

Query #1: ACCCTTGAAGGACGGCCA TTTTTAAACTGAGCATCCATCTAAAAGCCTTTT

Query #2: ACCCTTGAAGGACGGCCATTTTTA AACTGAGCATCCATCTAAAAGCCTTTT

SNP **INDEL**



$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2 \cdot \frac{precision \cdot recall}{precision + recall}$$

Challenge: *comparing complex variants*

Ground Truth

Representation #1

Reference: AAGG AAATC
Truth: ATCGAAAATC

CHROM	POS	REF	ALT
chr14	2	A	T
chr14	3	G	C
chr14	4	G	GA

Representation #2

Reference: AAGG AAATC
Truth: A TCGAAAATC

CHROM	POS	REF	ALT
chr14	1	AAGG	A
chr14	1	A	ATCGA

Challenge: *comparing complex variants*

Ground Truth

Representation #1

Reference:	AAGG AAATC
Truth:	ATCGAAAATC

CHROM	POS	REF	ALT
chr14	2	A	T
chr14	3	G	C
chr14	4	G	GA

Representation #2

Reference:	AAGG AAATC
Truth:	A TCGAAAATC

CHROM	POS	REF	ALT
chr14	1	AAGG	A
chr14	1	A	ATCGA

Challenge: *comparing complex variants*

Technology #1

Reference: AAGGAAATC
Query #1: ATCGAAATC

CHROM	POS	REF	ALT
chr14	2	A	T
chr14	3	G	C

Technology #2

Reference: AAGGAAATC
Query #2: A AAATC

CHROM	POS	REF	ALT
chr14	1	AAGG	A

Ground Truth

Representation #1

Reference: AAGG AAATC
Truth: ATCGAAAATC

CHROM	POS	REF	ALT
chr14	2	A	T
chr14	3	G	C
chr14	4	G	GA

Representation #2

Reference: AAGG AAATC
Truth: A TCGAAAATC

CHROM	POS	REF	ALT
chr14	1	AAGG	A
chr14	1	A	ATCGA

Challenge: *comparing complex variants*

Technology #1

Reference: AAGGAAATC
Query #1: ATCGAAATC

	CHROM	POS	REF	ALT
✓	chr14	2	A	T
✓	chr14	3	G	C

SNP Precision: 100%
SNP Recall: 100%
INDEL Precision: NA
INDEL Recall: 0%

Technology #2

Reference: AAGGAAATC
Query #2: A AAATC

	CHROM	POS	REF	ALT
✗	chr14	1	AAGG	A

SNP Precision: NA
SNP Recall: 0%
INDEL Precision: 0%
INDEL Recall: 0%

Ground Truth

Representation #1

Reference: AAGG AAATC
Truth: ATCGAAAATC

CHROM	POS	REF	ALT
chr14	2	A	T
chr14	3	G	C
chr14	4	G	GA

Representation #2

Reference: AAGG AAATC
Truth: A TCGAAAATC

CHROM	POS	REF	ALT
chr14	1	AAGG	A
chr14	1	A	ATCGA

Challenge: *comparing complex variants*

Technology #1

Reference: AAGGAAATC
Query #1: ATCGAAATC

	CHROM	POS	REF	ALT
✗	chr14	2	A	T
✗	chr14	3	G	C

SNP Precision: 100%
SNP Recall: 100%
INDEL Precision: NA
INDEL Recall: 0%

SNP Precision: 0%
SNP Recall: NA
INDEL Precision: NA
INDEL Recall: 0%

Technology #2

Reference: AAGGAAATC
Query #2: A AAATC

	CHROM	POS	REF	ALT
✓	chr14	1	AAGG	A

SNP Precision: NA
SNP Recall: 0%
INDEL Precision: 0%
INDEL Recall: 0%

SNP Precision: NA
SNP Recall: NA
INDEL Precision: 100%
INDEL Recall: 50%

Ground Truth

Representation #1

Reference: AAGG AAATC
Truth: ATCGAAAATC

CHROM	POS	REF	ALT
chr14	2	A	T
chr14	3	G	C
chr14	4	G	GA

Representation #2

Reference: AAGG AAATC
Truth: A TCGAAAATC

CHROM	POS	REF	ALT
chr14	1	AAGG	A
chr14	1	A	ATCGA

Outline

1. **Context**
2. **Problem**
3. **Discussion:** complex variant representation
4. **Solution**
5. **Implementation**
6. **Results**
7. **Next Steps**

Complex variants: *standard VCF normalization*

Reference AAGGAAATC **Query** ATCGAAAATC

Heng Li. “Toward better understanding of artifacts in variant calling from high-coverage samples.” Bioinformatics, 2014.
Tan et al. “Unified representation of genetic variants.” Bioinformatics, 2015.

Complex variants: *standard VCF normalization*

Reference AAGGAAATC **Query** ATCGAAAATC

Alignment

```
AAGGAAA-TC
·  ···· ··
ATCGAAAATC
```

VCF

POS	REF	ALT
2	AG	TC
6	AATC	AAATC

Original

Heng Li. “Toward better understanding of artifacts in variant calling from high-coverage samples.” Bioinformatics, 2014.
Tan et al. “Unified representation of genetic variants.” Bioinformatics, 2015.

Complex variants: *standard VCF normalization*

Reference AAGGAAATC **Query** ATCGAAAATC

Alignment

```
AAGGAAA-TC
·   ···· ··
ATCGAAAATC
```

VCF

POS	REF	ALT
2	AG	TC
6	AATC	AAATC

Original

```
AAGGAAA-TC
·   ···· ··
ATCGAAAATC
```

POS	REF	ALT
2	A	T
3	G	C
6	AATC	AAATC

Decomposed

Heng Li. “Toward better understanding of artifacts in variant calling from high-coverage samples.” Bioinformatics, 2014.
Tan et al. “Unified representation of genetic variants.” Bioinformatics, 2015.

Complex variants: *standard VCF normalization*

Reference AAGGAAATC **Query** ATCGAAAATC

Alignment

```
AAGGAAA-TC
·   ···· ··
ATCGAAAATC
```

VCF

POS	REF	ALT
2	AG	TC
6	AATC	AAATC

Original

```
AAGGAAA-TC
·   ···· ··
ATCGAAAATC
```

POS	REF	ALT
2	A	T
3	G	C
6	AATC	AAATC

Decomposed

```
AAGGAAA-TC
·   ···· ··
ATCGAAAATC
```

POS	REF	ALT
2	A	T
3	G	C
7	A	AA

Trimmed

Heng Li. “Toward better understanding of artifacts in variant calling from high-coverage samples.” Bioinformatics, 2014.
Tan et al. “Unified representation of genetic variants.” Bioinformatics, 2015.

Complex variants: *standard VCF normalization*

Reference AAGGAAATC **Query** ATCGAAAATC

Alignment

Original	Decomposed	Trimmed	Left shifted
AAGGAAA-TC	AAGGAAA-TC	AAGGAAA-TC	AAGG-AAATC
· · ···· ··	· · ···· ··	· · ···· ··	· · · ·····
ATCGAAAATC	ATCGAAAATC	ATCGAAAATC	ATCGAAAATC

VCF

POS	REF	ALT
2	AG	TC
6	AATC	AAATC

POS	REF	ALT
2	A	T
3	G	C
6	AATC	AAATC

POS	REF	ALT
2	A	T
3	G	C
7	A	AA

POS	REF	ALT
2	A	T
3	G	C
4	G	GA

Heng Li. “Toward better understanding of artifacts in variant calling from high-coverage samples.” Bioinformatics, 2014.
Tan et al. “Unified representation of genetic variants.” Bioinformatics, 2015.

Complex variants: *standard VCF normalization*

Reference AAGGAAATC **Query** ATCGAAAATC

Alignment

Original	Decomposed	Trimmed	Left shifted	Alternate
AAGGAAA-TC	AAGGAAA-TC	AAGGAAA-TC	AAGG-AAATC	AAGG----AAATC
ATCGAAAATC	ATCGAAAATC	ATCGAAAATC	ATCGAAAATC	A---TCGAAAATC

VCF

Original			Decomposed			Trimmed			Left shifted			Alternate		
POS	REF	ALT	POS	REF	ALT	POS	REF	ALT	POS	REF	ALT	POS	REF	ALT
2	AG	TC	2	A	T	2	A	T	2	A	T	1	AAGG	A
6	AATC	AAATC	3	G	C	3	G	C	3	G	C	1	A	ATCGA
			6	AATC	AAATC	7	A	AA	4	G	GA			

Heng Li. “Toward better understanding of artifacts in variant calling from high-coverage samples.” Bioinformatics, 2014.
Tan et al. “Unified representation of genetic variants.” Bioinformatics, 2015.

Complex variants: *standard VCF normalization*

Reference AAGGAAATC **Query** ATCGAAAATC

Alignment

```
AAGGAAA-TC
·  ···· ··
ATCGAAAATC
```

VCF

POS	REF	ALT
2	AG	TC
6	AATC	AAATC

Original

```
AAGGAAA-TC
·  ···· ··
ATCGAAAATC
```

POS	REF	ALT
2	A	T
3	G	C
6	AATC	AAATC

Decomposed

```
AAGGAAA-TC
·  ···· ··
ATCGAAAATC
```

POS	REF	ALT
2	A	T
3	G	C
7	A	AA

Trimmed

```
AAGG-AAATC
·  · ·····
ATCGAAAATC
```

POS	REF	ALT
2	A	T
3	G	C
4	G	GA

Left shifted

```
AAGG----AAATC
·       ·····
A---TCGAAAATC
```

POS	REF	ALT
1	AAGG	A
1	A	ATCGA

Alternate

Heng Li. “Toward better understanding of artifacts in variant calling from high-coverage samples.” Bioinformatics, 2014.
Tan et al. “Unified representation of genetic variants.” Bioinformatics, 2015.

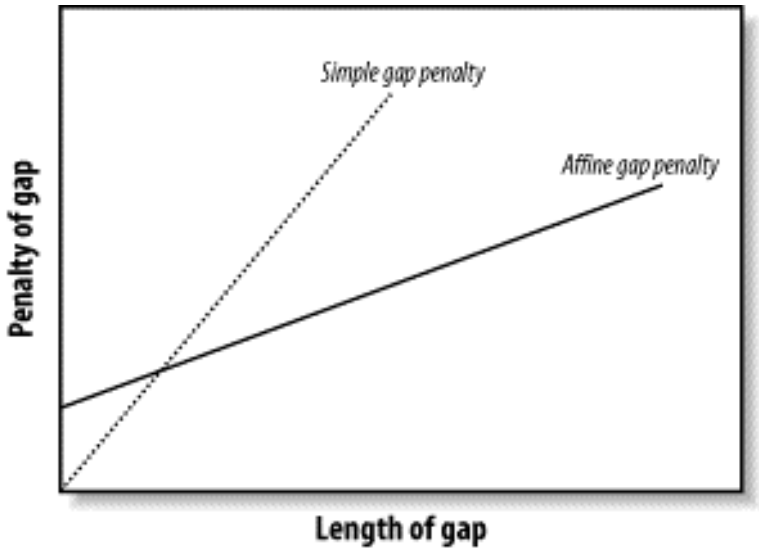
Choosing representations: *best-alignment normalization*

m = match

x = mis-match

o = gap opening

e = gap extension



Bayat et al. “Improved VCF normalization for accurate VCF comparison”. Oxford Bioinformatics, 2017.

Choosing representations: *best-alignment normalization*

m = match

x = mis-match

o = gap opening

e = gap extension

Option #1

```
AAGG-AAATC
.  .  .....
ATCGAAAATC
```

POS	REF	ALT
2	A	T
3	G	C
4	G	GA

x + x + (o+e)

Option #2

```
AAGG----AAATC
.       .....
A---TCGAAAATC
```

POS	REF	ALT
1	AAGG	A
1	A	ATCGA

(o+3e) + (o+4e)

Bayat et al. “Improved VCF normalization for accurate VCF comparison”. Oxford Bioinformatics, 2017.

Choosing representations: *best-alignment normalization*

$m = 0$ = match

$x = 5$ = mis-match

$o = 6$ = gap opening

$e = 2$ = gap extension

Option #1

```
AAGG-AAATC
.  .  .....
ATCGAAAATC
```

POS	REF	ALT
2	A	T
3	G	C
4	G	GA

x + x + (o+e)

18

Option #2

```
AAGG----AAATC
.       .....
A---TCGAAAATC
```

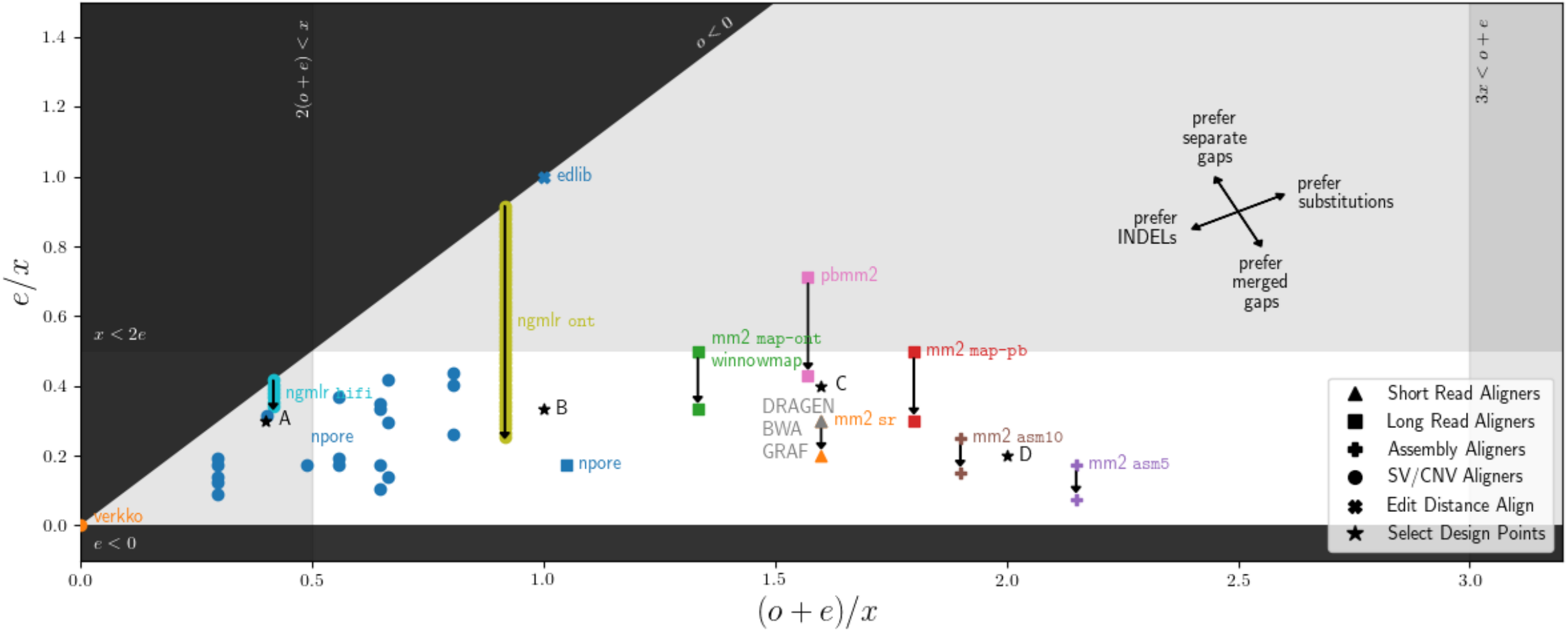
POS	REF	ALT
1	AAGG	A
1	A	ATCGA

(o+3e) + (o+4e)

26

Bayat et al. “Improved VCF normalization for accurate VCF comparison”. Oxford Bioinformatics, 2017.

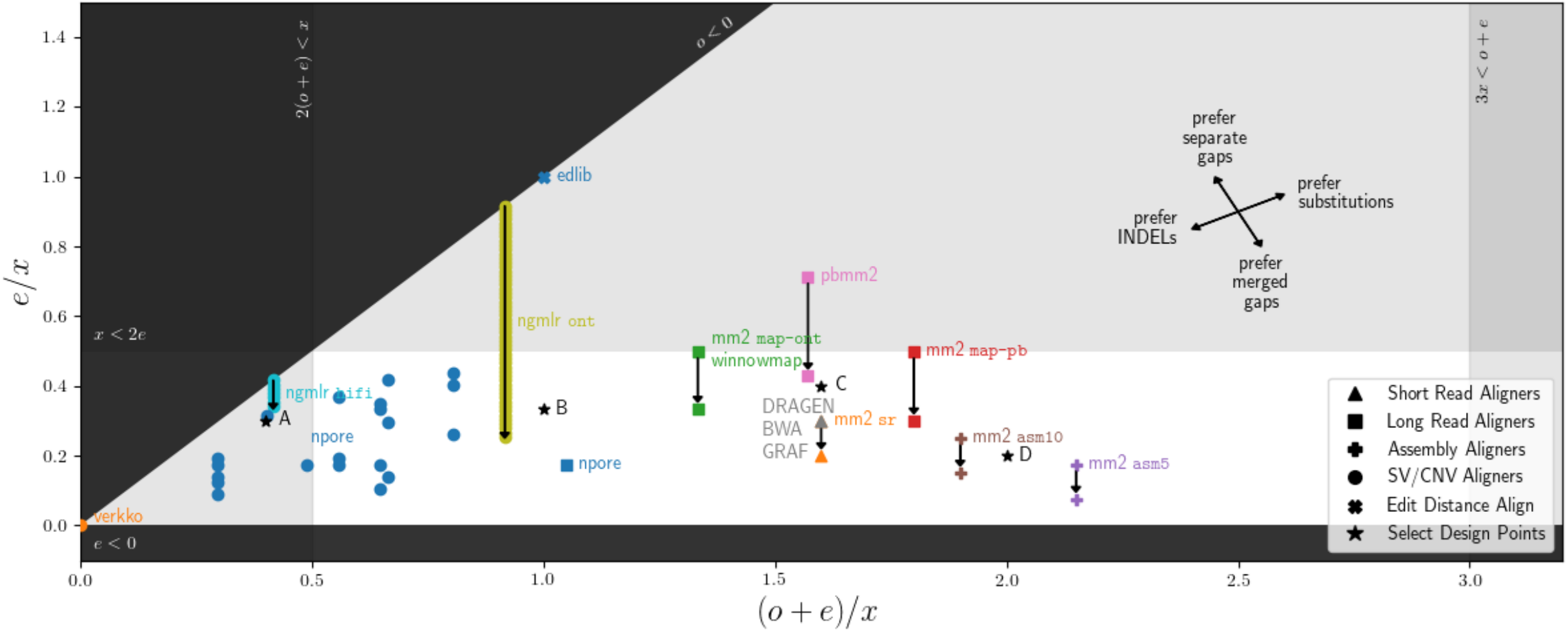
Alignment-based normalization design space



To what extent do parameters matter?

Representation	SNPs	INDELs
Original	3,367,320	548,602
A	0	7,185,103
B	3,366,095	547,654
C	3,369,257	545,077
D	3,369,279	544,664

Design point A: *structural and copy number variants*



SV / CNV Analysis: *only recently enabled by long reads*

- **2014**: NIST/GIAB initial small variant benchmark (77% of GRCh38)
- **2019**: NIST/GIAB small variant benchmark expansion (84% of GRCh38)
- **2020**: NIST/GIAB structural variant benchmarks
- **2022**: NIST/GIAB challenging small variants (92% of GRCh38)
- **2023**: NIST/GIAB tandem repeat benchmarks

- **2022**: T2T Consortium “The first complete human genome”

Motivation: *GIAB tandem repeat benchmark*

Dataset	SNPs	INDELS
Original GIAB TR	917,255	431,545
Normalized GIAB TR	502,076	461,258

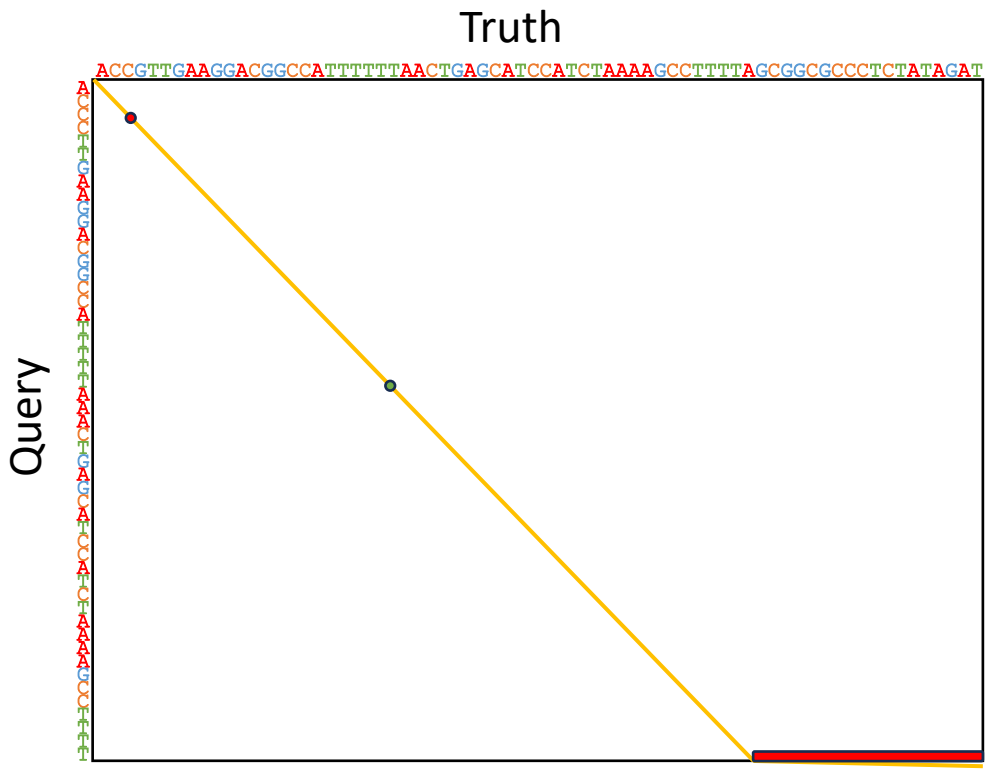
Outline

1. **Context**
2. **Problem**
3. **Discussion**
4. **Solution:** new comparison methods
5. **Implementation**
6. **Results**
7. **Next Steps**

Idea #1: *sequence-based evaluation metrics*

Reference: ACCGTTGAAGGACGGCCATTTTTT AACTGAGCATCCATCTAAAAGCCTTTTAGCGGCGCCCTCTATAGAT
Query #1: ACCCTTGAAGGACGGCCA TTTTTAAACTGAGCATCCATCTAAAAGCCTTTT
Truth: ACCGTTGAAGGACGGCCA TTTTTAAACTGAGCATCCATCTAAAAGCCTTTTAGCGGCGCCCTCTATAGAT

- Edit Distance
- Distinct Edits
- Alignment Distance



Idea #2: *standardize complex variant representation*

Reference: ACCGTTGAAGGACGGCCATTTTTT AACTGAGCATCCATCTAAAAGCCTTTTAGCGGCGCCCTCTATAGAT

Query #1: ACCCTTGAAGGACGGCCA TTTTTAAACTGAGCATCCATCTAAAAGCCTTTT

Query #2: ACCCTTGAAGGACGGCCATTTTTA AACTGAGCATCCATCTAAAAGCCTTTT

Query

POS	REFERENCE	ALTERNATE
4	G	C
18	AT	A
25	T	TA
53	TAGCGGCG…	T

POS	REFERENCE	ALTERNATE
4	G	C
24	T	A
53	TAGCGGCG…	T

Idea #3: *allow partial credit for variant calls*

Reference: ACCGTTGAAGGACGGCCATTTTTTAACTGAGCATCCATCTAAAAGCCTTTTAGCGGCGCCCTCTATAGAT
Query #1: ACCCTTGAAGGACGGCCATTTTTAAACTGAGCATCCATCTAAAAGCCTTTT 20 bases
Truth: ACCCTTGAAGGACGGCCATTTTTAAACTGAGCATCCATCTAAAAGCCTTTTAG 18 bases

Query

POS	REFERENCE	ALTERNATE
4	G	C
24	T	A
53	TAGCGGCG…	T

Truth

POS	REFERENCE	ALTERNATE
4	G	C
24	T	A
55	GCGGCG…	G

Idea #3: *allow partial credit for variant calls*

Reference: ACCGTTGAAGGACGGCCATTTTTTAACTGAGCATCCATCTAAAAGCCTTTTAGCGGCGCCCTCTATAGAT
Query #1: ACCCTTGAAGGACGGCCATTTTTAAACTGAGCATCCATCTAAAAGCCTTTT 20 bases
Truth: ACCCTTGAAGGACGGCCATTTTTAAACTGAGCATCCATCTAAAAGCCTTTTAG 18 bases

Query

	POS	REFERENCE	ALTERNATE
✓	4	G	C
✓	24	T	A
✗	53	TAGCGGCG…	T

Truth

	POS	REFERENCE	ALTERNATE
✓	4	G	C
✓	24	T	A
✗	55	GCGGCG…	G

Idea #3: *allow partial credit for variant calls*

Reference: ACCGTTGAAGGACGGCCATTTTTTAACTGAGCATCCATCTAAAAGCCTTTTAGCGGCGCCCTCTATAGAT
Query #1: ACCCTTGAAGGACGGCCATTTTTAAACTGAGCATCCATCTAAAAGCCTTTT (20 bases)
Truth: ACCCTTGAAGGACGGCCATTTTTAAACTGAGCATCCATCTAAAAGCCTTTTAG (18 bases)

Query

	POS	REFERENCE	ALTERNATE
✓	4	G	C
✓	24	T	A
.9 ✓ + .1 ✗	53	TAGCGGCG…	T

Truth

	POS	REFERENCE	ALTERNATE
✓	4	G	C
✓	24	T	A
.9 ✓ + .1 ✗	55	GCGGCG…	G

Idea #4: *enforce local variant phasing*

Reference: GAGCC

Query #1:
1 GACCC
2 GTGAC

Phased Query

POS	REF	ALT	GENOTYPE
2	A	T	0|1
3	G	C	1|0
4	C	A	0|1

"Correct" Query haps:
1 GACCC
2 GTGAC

Idea #4: *enforce local variant phasing*

Reference: GAGCC

Query #1:

1 GACCC

2 GTGAC

Phased Query

POS	REF	ALT	GENOTYPE
2	A	T	0|1
3	G	C	1|0
4	C	A	0|1

“Correct” Query haps:

1 GACCC

2 GTGAC

Idea #4: *enforce local variant phasing*

Reference: GAGCC

Query #1:

1 GACCC

2 GTGAC

Phased Query

POS	REF	ALT	GENOTYPE
2	A	T	0|1
3	G	C	1|0
4	C	A	0|1

"Correct" Query haps:

1 GACCC

2 GTGAC

Unphased Query

POS	REF	ALT	GENOTYPE
2	A	T	0/1
3	G	C	0/1
4	C	A	0/1

Idea #4: *enforce local variant phasing*

Reference: GAGCC

Query #1:

1 GACCC
2 GTGAC

Phased Query

POS	REF	ALT	GENOTYPE
2	A	T	0|1
3	G	C	1|0
4	C	A	0|1

"Correct" Query haps:

1 GACCC
2 GTGAC

Unphased Query

POS	REF	ALT	GENOTYPE
2	A	T	0/1
3	G	C	0/1
4	C	A	0/1

"Correct" Query haps:

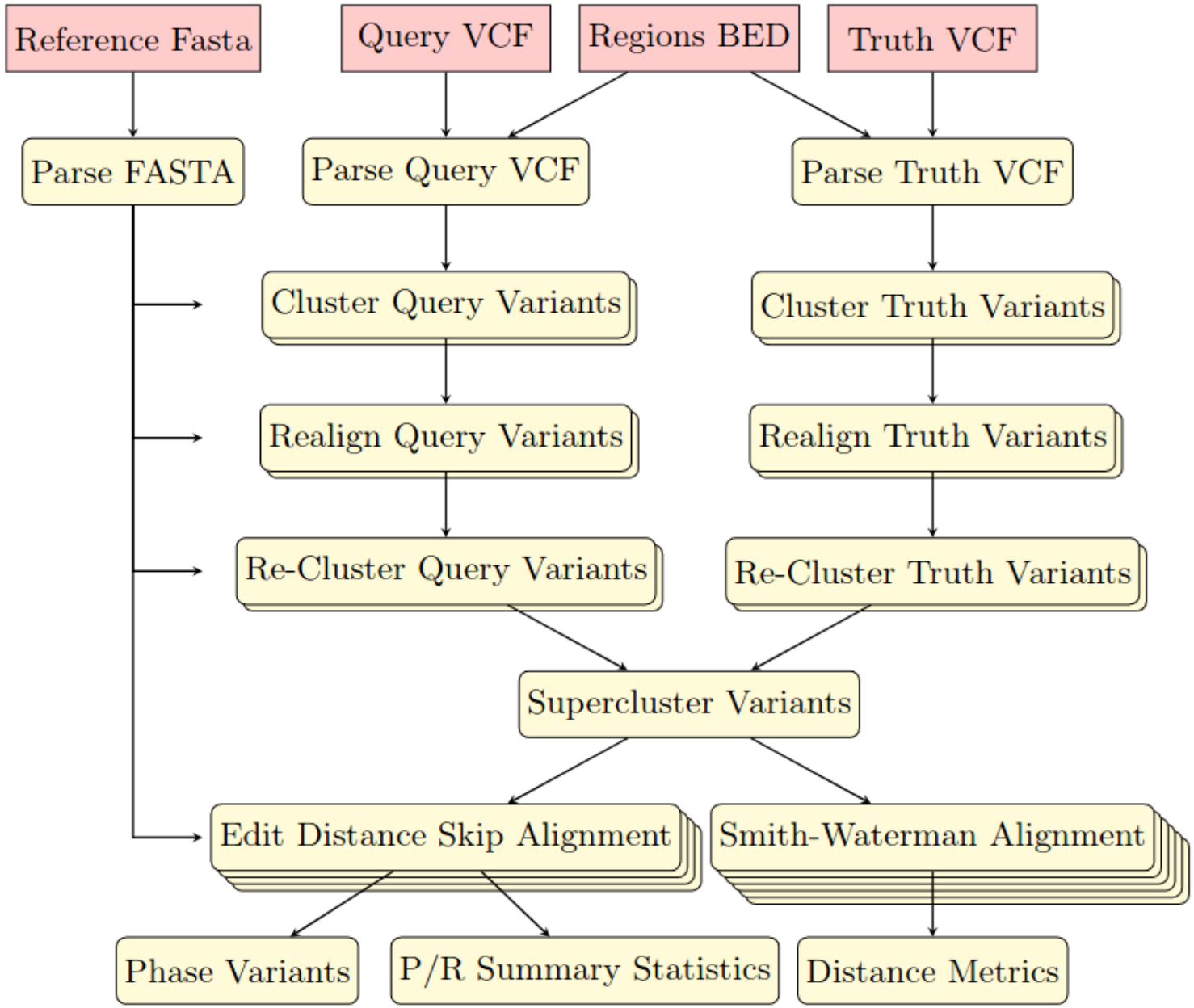
1 GAGCC GAGAC GACCC GTGCC
2 GTCAC GTCCC GTGAC GACAC

1 GTCAC GTCCC GTGAC GACAC
2 GAGCC GAGAC GACCC GTGCC

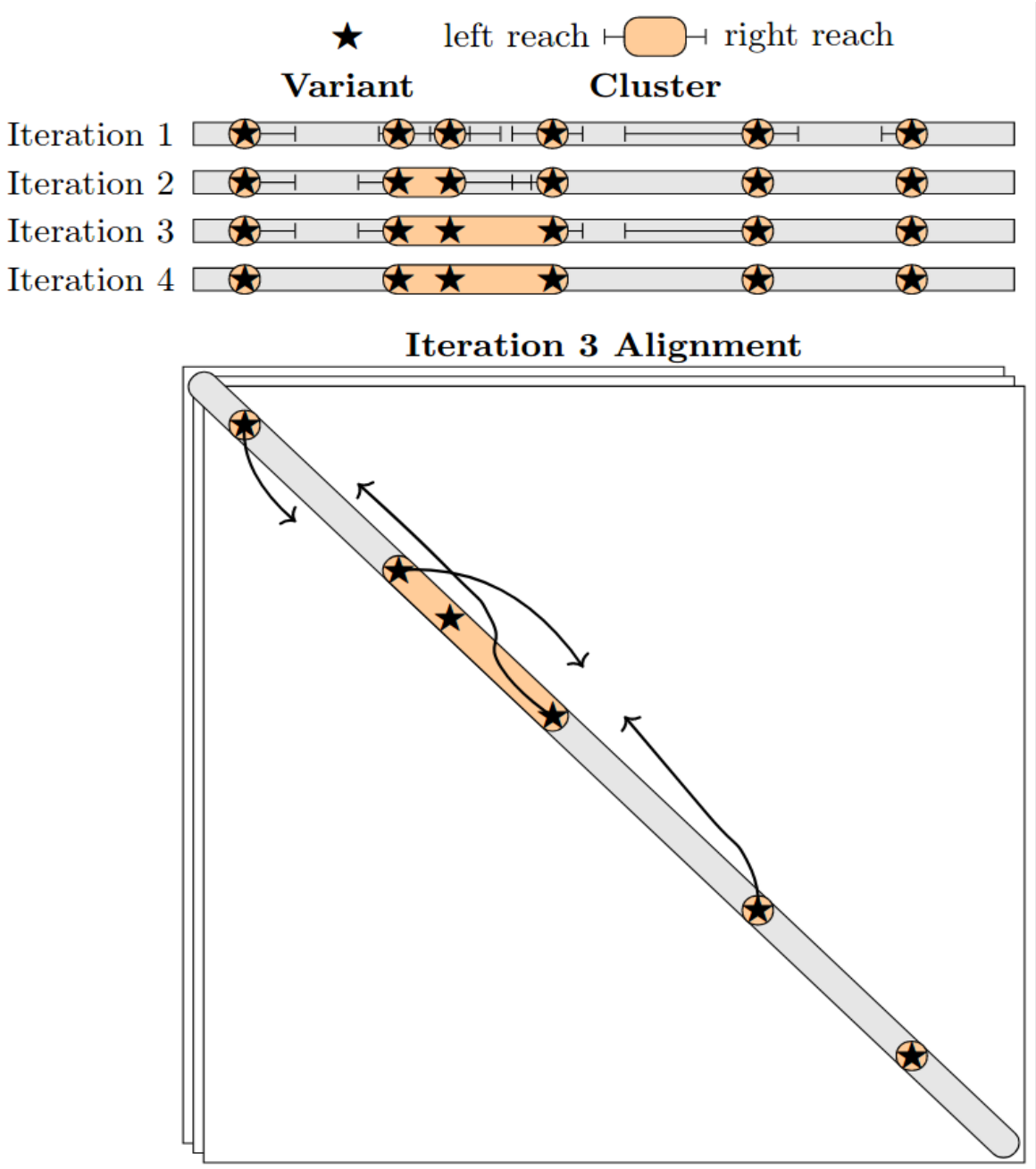
Outline

1. **Context**
2. **Problem**
3. **Discussion**
4. **Solution**
5. **Implementation:** dynamic programming / alignment
6. **Results**
7. **Next Steps**

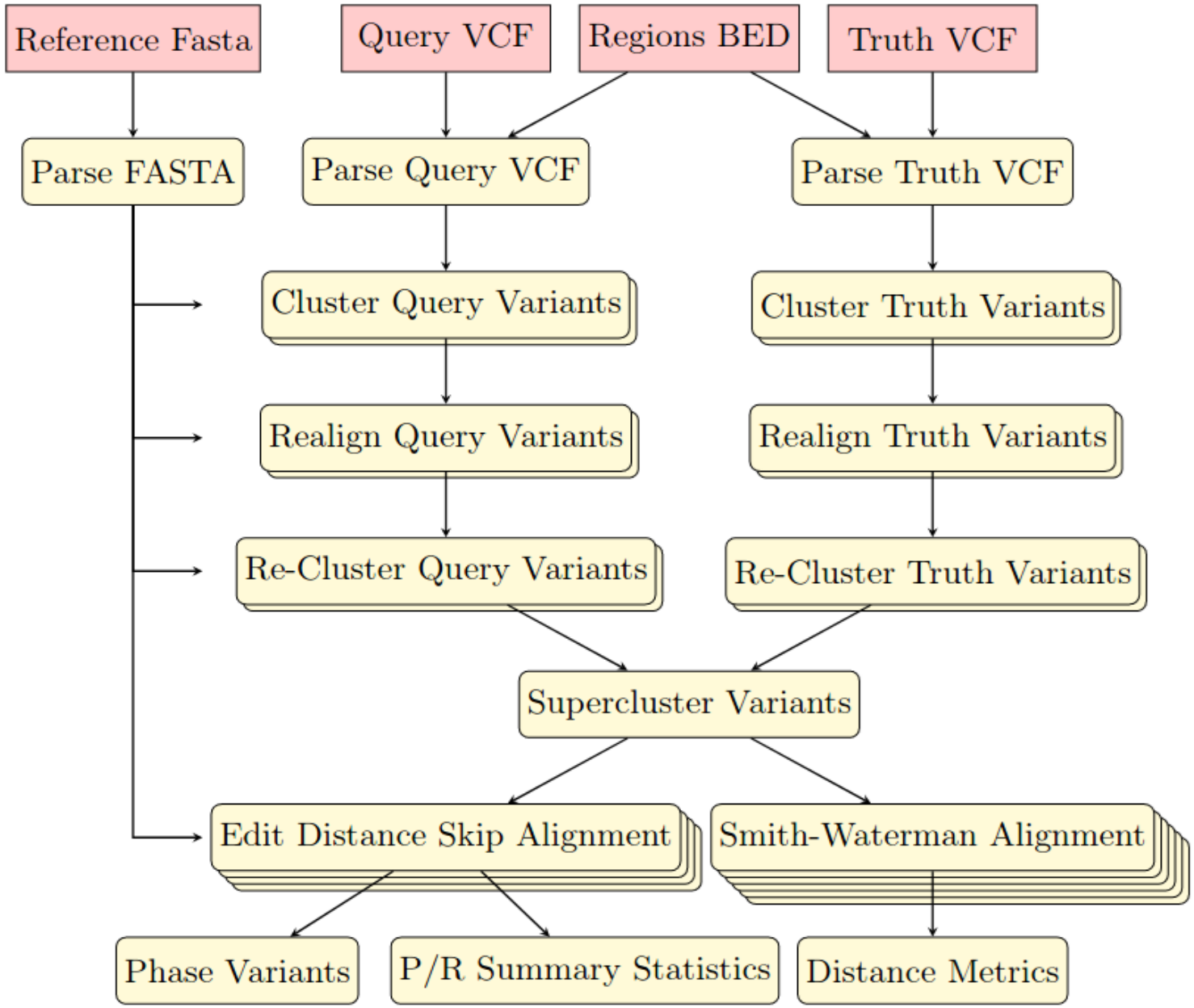
Overview


Reference Fasta
Query VCF
Regions BED
Truth VCF
Parse FASTA
Parse Query VCF
Parse Truth VCF
Cluster Query Variants
Cluster Truth Variants
Realign Query Variants
Realign Truth Variants
Re-Cluster Query Variants
Re-Cluster Truth Variants
Supercluster Variants
Edit Distance Skip Alignment
Smith-Waterman Alignment
Phase Variants
P/R Summary Statistics
Distance Metrics

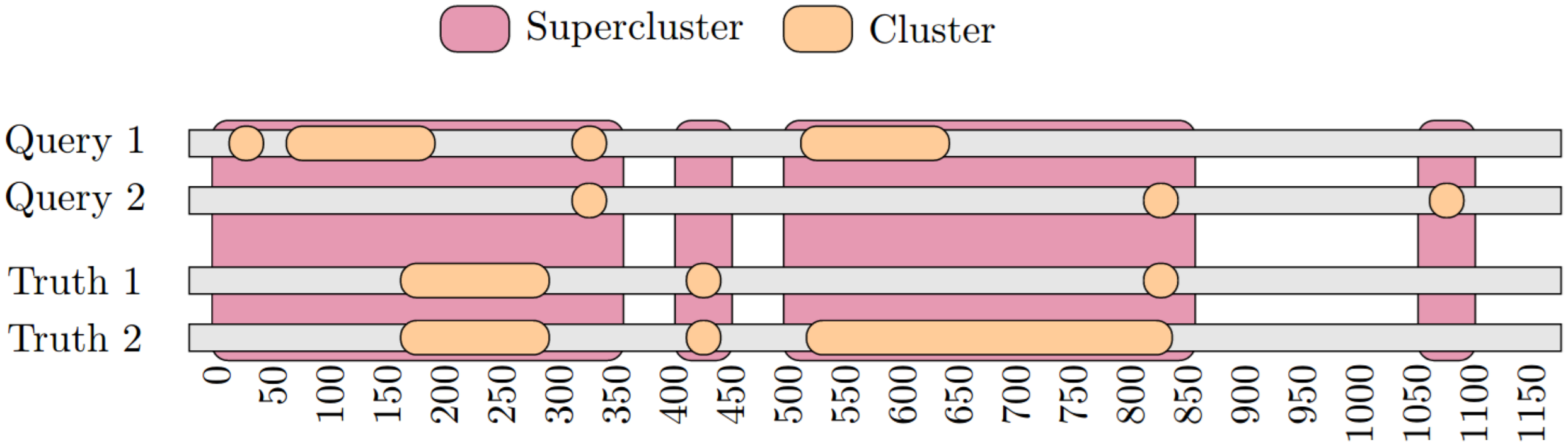
Clustering:
iterative bi-directional wavefront alignment



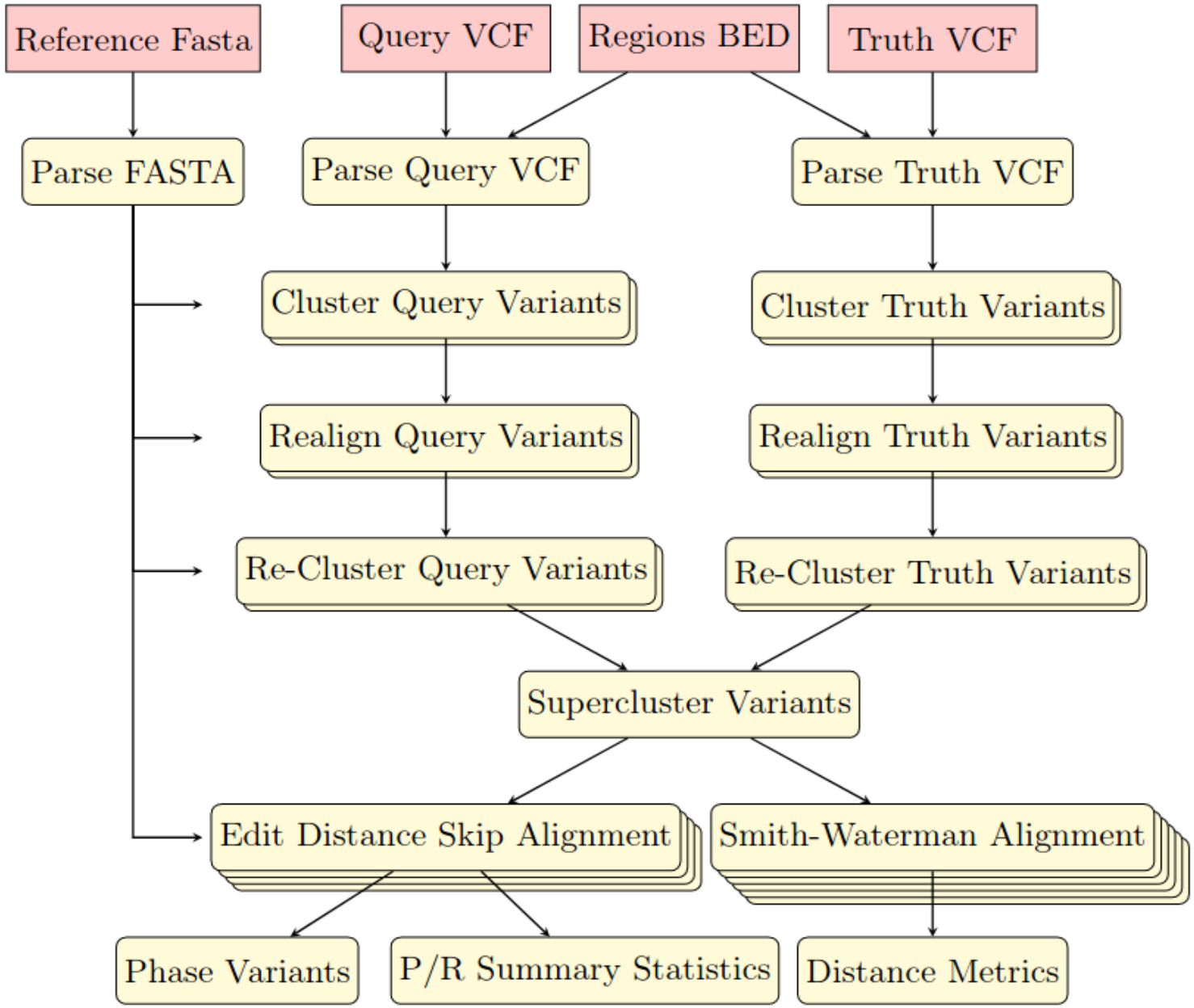
Overview


Reference Fasta
Query VCF
Regions BED
Truth VCF
Parse FASTA
Parse Query VCF
Parse Truth VCF
Cluster Query Variants
Cluster Truth Variants
Realign Query Variants
Realign Truth Variants
Re-Cluster Query Variants
Re-Cluster Truth Variants
Supercluster Variants
Edit Distance Skip Alignment
Smith-Waterman Alignment
Phase Variants
P/R Summary Statistics
Distance Metrics

Superclustering: *simple reference distance heuristic*

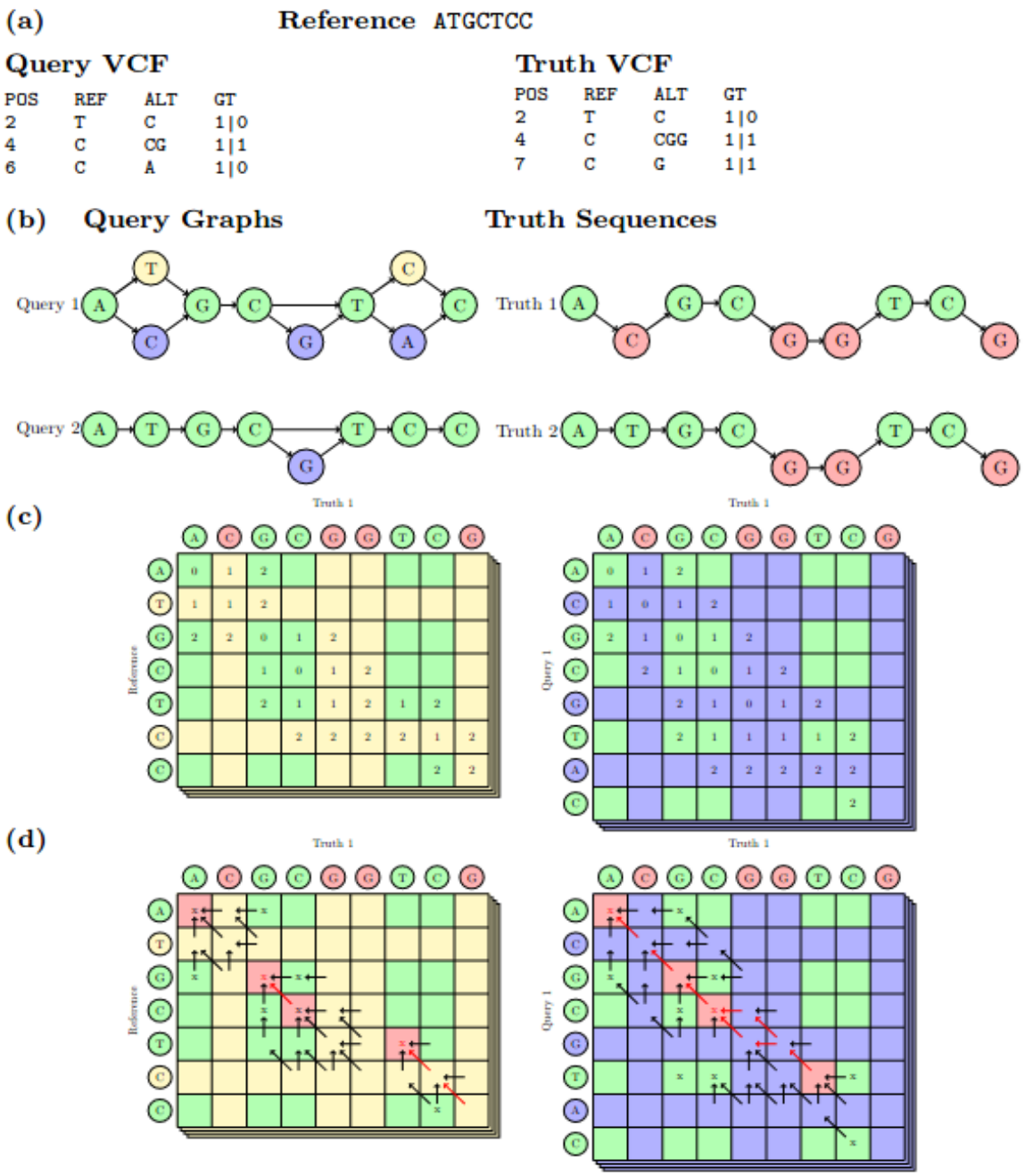


Overview

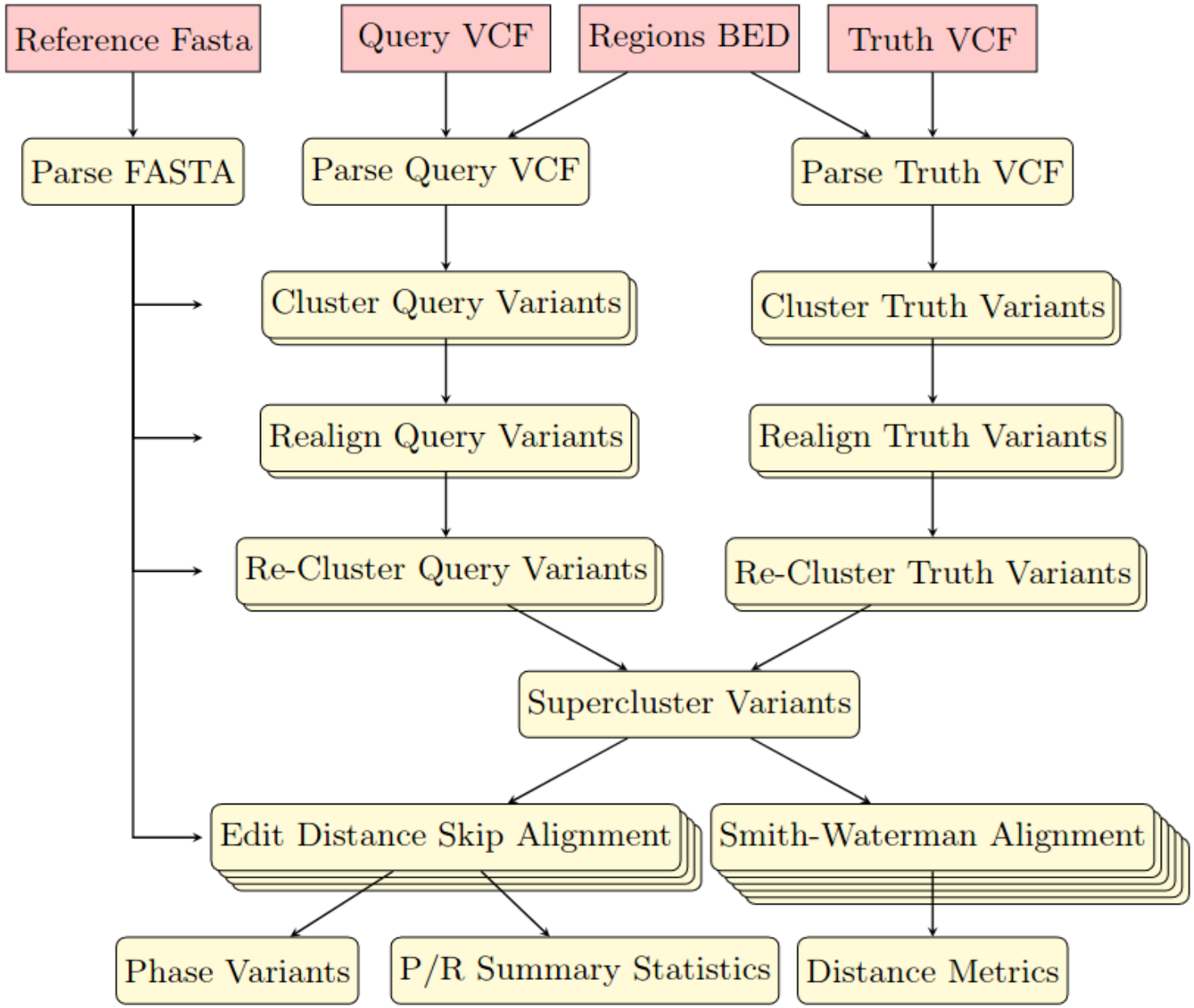

Reference Fasta
Query VCF
Regions BED
Truth VCF
Parse FASTA
Parse Query VCF
Parse Truth VCF
Cluster Query Variants
Cluster Truth Variants
Realign Query Variants
Realign Truth Variants
Re-Cluster Query Variants
Re-Cluster Truth Variants
Supercluster Variants
Edit Distance Skip Alignment
Smith-Waterman Alignment
Phase Variants
P/R Summary Statistics
Distance Metrics

Precision/Recall:

edit distance, allows skipping FP query variants, backtracking



Overview


Reference Fasta
Query VCF
Regions BED
Truth VCF
Parse FASTA
Parse Query VCF
Parse Truth VCF
Cluster Query Variants
Cluster Truth Variants
Realign Query Variants
Realign Truth Variants
Re-Cluster Query Variants
Re-Cluster Truth Variants
Supercluster Variants
Edit Distance Skip Alignment
Smith-Waterman Alignment
Phase Variants
P/R Summary Statistics
Distance Metrics

Outline

1. **Context**
2. **Problem**
3. **Discussion**
4. **Solution**
5. **Implementation**
6. **Results:** improved evaluation stability
7. **Next Steps**

Example #1: *complex variant normalization*

Original VCF: *GIAB Tandem Repeats*

chr20 278985 A C
chr20 278986 C G
chr20 278990 G C
chr20 278993 C A
chr20 278994 G GGGAGGGAGGGCGGGACGGAGGGA
GGGAGGGAGGGACGGAGGGCGGGACGGCGGGAGGGCGGGAC
GGAGGGACGGAGGGAGGGCGGGACGGAGGGCGGGAGGGCGG
GACGGAGGGAGGGAGGGAGGGAGGGCGGGACGGAGGGAGGG
AGGGCGGGACGGAGGGAGGGAGGGAGGGACGGAGGGCGGGA
CGGCGGGAGGGCGGGACGGAGGGACGGAGGGAGGGCGGGAC
GGAGGGCGGGAGGGCGGGACGGAGGGAGGGAGGGCGGGACG
GAGGGACGGAGGGAGGGAGGGCGGGACGGAGGGAGGGAGGG
CGGGACGGAGGGACGGAGGGAGGGAGGGCGGGACGGAGGGA
GGGAGGGAGGGACGGAGGGCGGGACGGCGGGAGGGCGGGAC
GGAGGGACGGAGGGAGGGCGGGACGGAGGGCGGGAGGGAGG
GAGGGCGGGACGGAGGGAGGGAGGGAGGGAGGGCGGGACGG
AGGGAGGGAGGGAGGGAGGGACGGAGGGACGGAGGGAGGGA
GGGAGGGAGGGACGGAGGGCGGGACGGAGGGAGGGAGGGCG
GAGGGAGGGAGGGCGGGACGGAGGGAGGGAGGGAGGGACGG
AGGGCGGGACGGAGGGAGGGAGGGC
chr20 278998 C G
chr20 279001 C A
chr20 279022 C G
chr20 279029 A C
chr20 279033 C A
chr20 279038 C T
chr20 279045 C A
chr20 279069 A C

12 SNPs
1 INS (622bp)

Normalized VCF: *vcfdist design point C*

chr20 278984 G GCGGGACGGAGGGAGGGAGGGCG
GGACGGAGGGAGGGAGGGAGGGACGGAGGGCGGGA
CGGCGGGAGGGCGGGACGGAGGGACGGAGGGAGGG
CGGGACGGAGGGCGGGAGGGCGGGACGGAGGGAGG
GAGGGAGGGAGGGCGGGACGGAGGGAGGGAGGGCG
GGACGGAGGGAGGGAGGGAGGG
chr20 279069 A AGGGCGGGACGGAGGGACGGAGG
GAGGGAGGGCGGGACGGAGGGAGGGAGGGCGGGAC
GGAGGGACGGAGGGAGGGAGGGCGGGACGGAGGGA
GGGAGGGAGGGACGGAGGGCGGGACGGCGGGAGGG
CGGGACGGAGGGACGGAGGGAGGGCGGGACGGAGG
GCGGGAGGGAGGGAGGGCGGGACGGAGGGAGGGAG
GGAGGGAGGGCGGGACGGAGGGAGGGAGGGAGGGA
GGGACGGAGGGACGGAGGGAGGGAGGGAGGGAGGG
ACGGAGGGCGGGACGGAGGGAGGGAGGGCGGAGGG
AGGGAGGGCGGGACGGAGGGAGGGAGGGACG
GAGGGCGGGACGGAGGGAGGGAGGGCGGAGGGAGG
GAGGGCGGGACGGAGGGAGGGAGGGCGGGAGGGAT
GGAGGGAGGGAGGGCGGGACGGAGGGAGGGC

2 INS (438bp, 184bp)

Example #2: *complex variant near-equivalence*

Query:

CHROM	POS	REF	ALT	CALL	CREDIT
chr1	976722	C	CAGGAACCGCCTCCCACTCCCCCCACAACCCCGGGAACCGCCTCCCACTC		
CCCCCGCAACCCCGGGAACCGCCTCCCACTCCCCCCGCAACCCC				INS PP	0.979167
chr1	976745	G	A	SNP PP	0.979167

Truth:

CHROM	POS	REF	ALT	CALL	CREDIT
chr1	976715	A	ACAACCCCAGGAACCGCCTCCCACTCCCCCCA	INS PP	0.979167
chr1	976747	C	CAACCCCGGGAACCGCCTCCCACTCCCCCCG	INS PP	0.979167
chr1	976777	G	A	SNP PP	0.979167
chr1	976811	C	CAACCCCGGGAACCGCCTCCCACTCCCCCCG	INS PP	0.979167
chr1	976840	C	G	SNP PP	0.979167
chr1	976841	G	A	SNP PP	0.979167

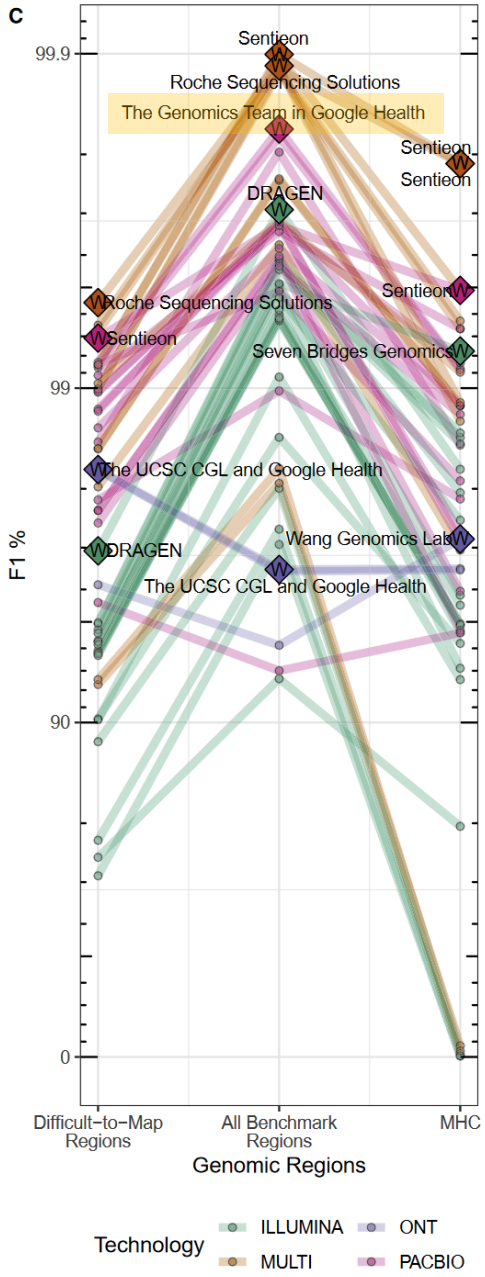
Dataset: *PrecisionFDA Truth Challenge V2*

- 64 whole genome sequencing submissions
- Illumina, PacBio, ONT, and Multi

Olson et al. “PrecisionFDA Truth Challenge V2: Calling variants from short and long reads in difficult-to-map regions.” Cell, 2022.

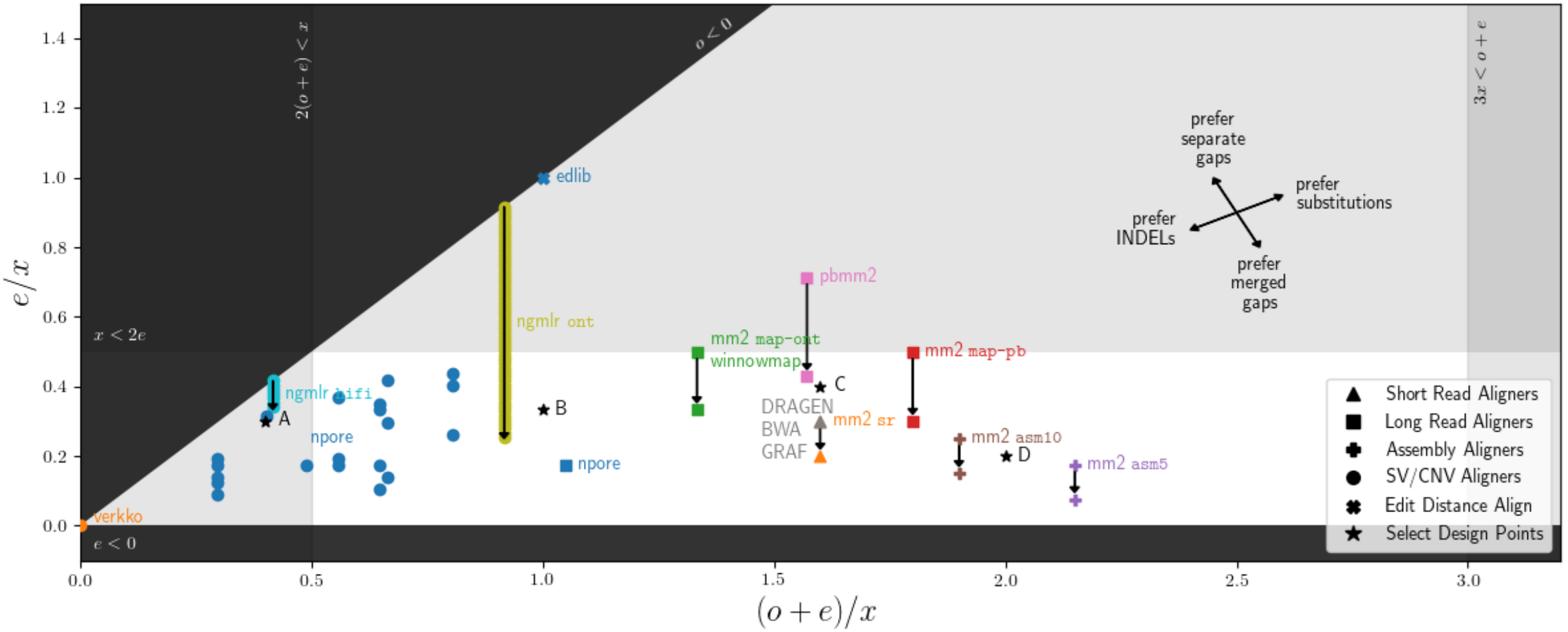
Dataset: *PrecisionFDA Truth Challenge V2*

- 64 whole genome sequencing submissions
- Illumina, PacBio, ONT, and Multi



Olson et al. “PrecisionFDA Truth Challenge V2: Calling variants from short and long reads in difficult-to-map regions.” Cell, 2022.

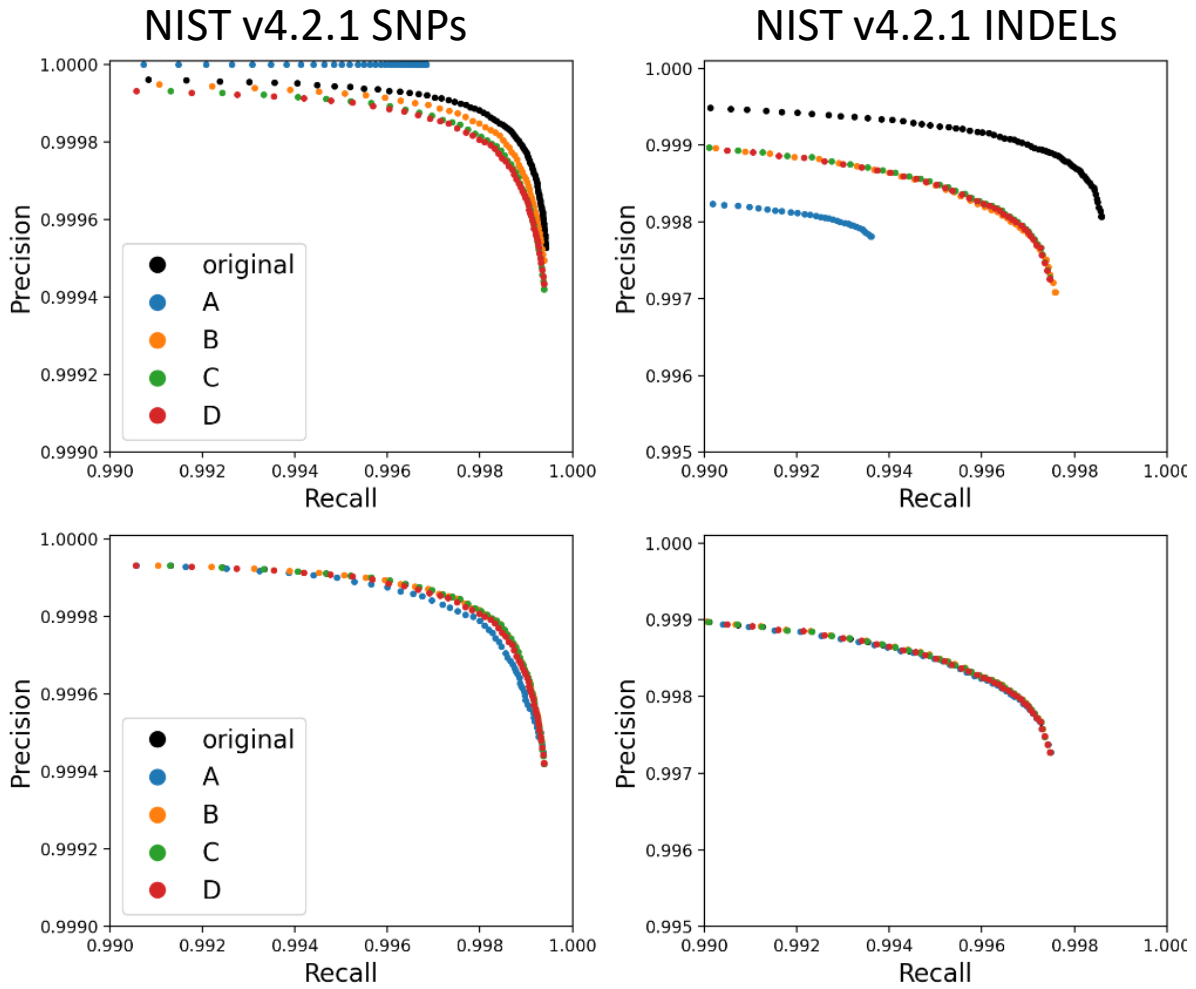
Analysis: *select design points*



Results: *normalization fixes representation bias*

Before

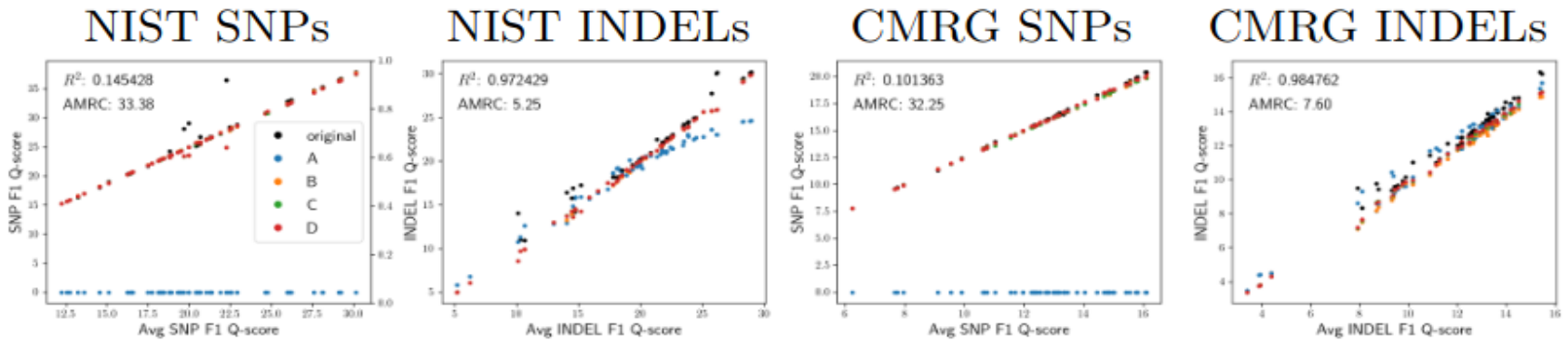
After



Prior Work
vcfeval

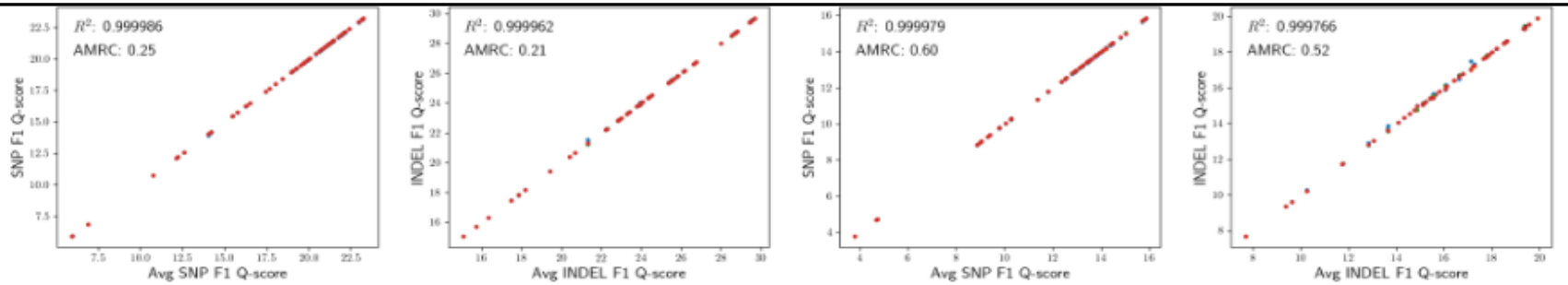
Results: *stable performance across representations*

vcfeval precision/recall metrics



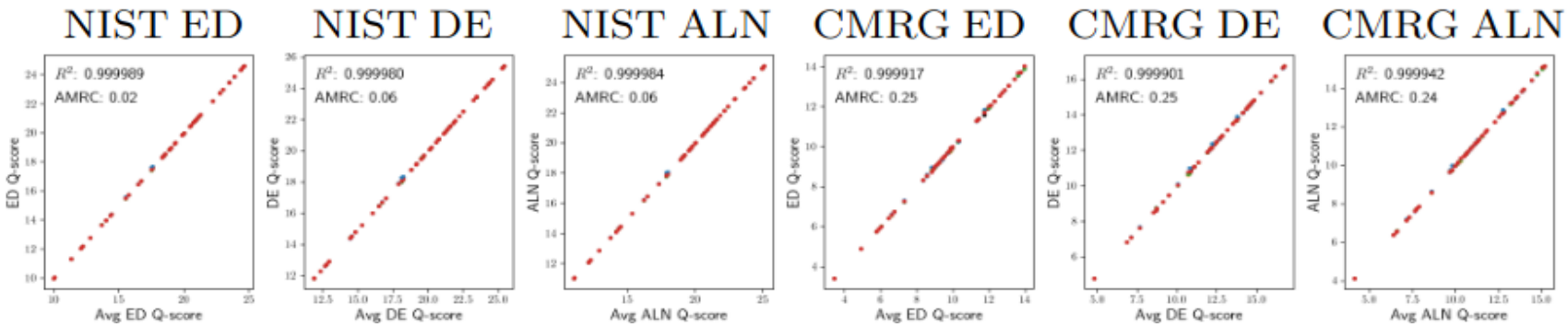
Prior Work

vcfdist precision/recall metrics



This Work

vcfdist distance metrics



Outline

1. **Context**
2. **Problem**
3. **Discussion**
4. **Solution**
5. **Implementation**
6. **Results**
7. **Next Steps:** structural and unphased variants

Next Steps: *structural and unphased variants*

NIST Collaboration

- vcfdist now works with structural variants up to 10,000bp
- Comprehensive evaluation of recent Verkko assemblies
- Simultaneous benchmarking of SNPs/INDELs/TRs/SVs

Next Steps: *structural and unphased variants*

NIST Collaboration

- vcfdist now works with structural variants up to 10,000bp
- Comprehensive evaluation of recent Verkko assemblies
- Simultaneous benchmarking of SNPs/INDELs/TRs/SVs

Planned Research

- Extend vcfdist's alignment algorithm to more general graphs
- This allows vcfdist to evaluate unphased query variant call sets

Conclusion

`github.com/`**`TimD1/vcfdist`**

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

Expected Graduation: **Summer 2024**