vcfdist: accurately benchmarking phased variant calls

Tim Dunn

PhD Candidate

University of Michigan

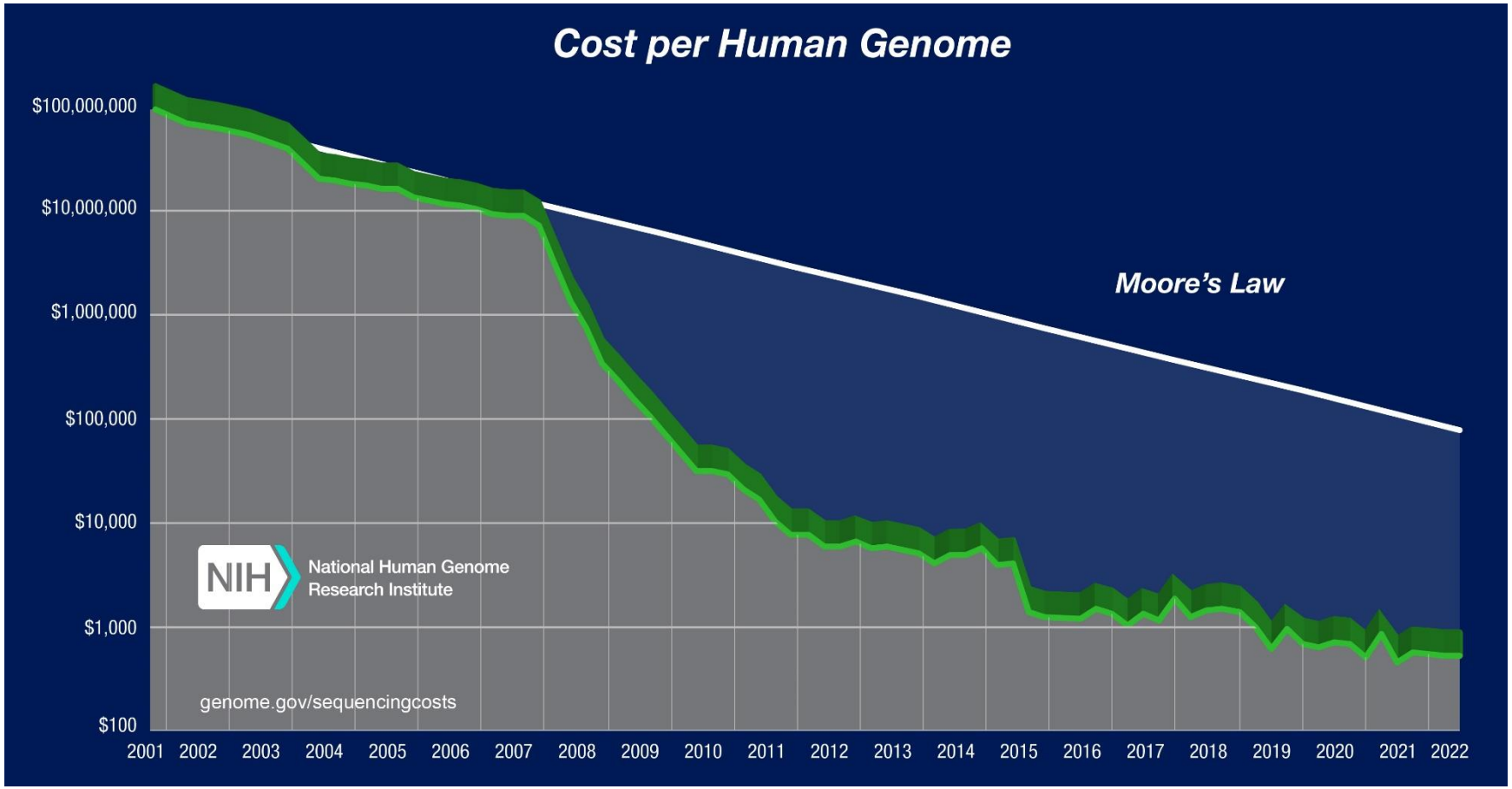
Outline

1. **Background**
2. **Problem #1**
3. **Solution**
4. **Results**
5. **Problem #2**
6. **Solution**
7. **Results**

Outline

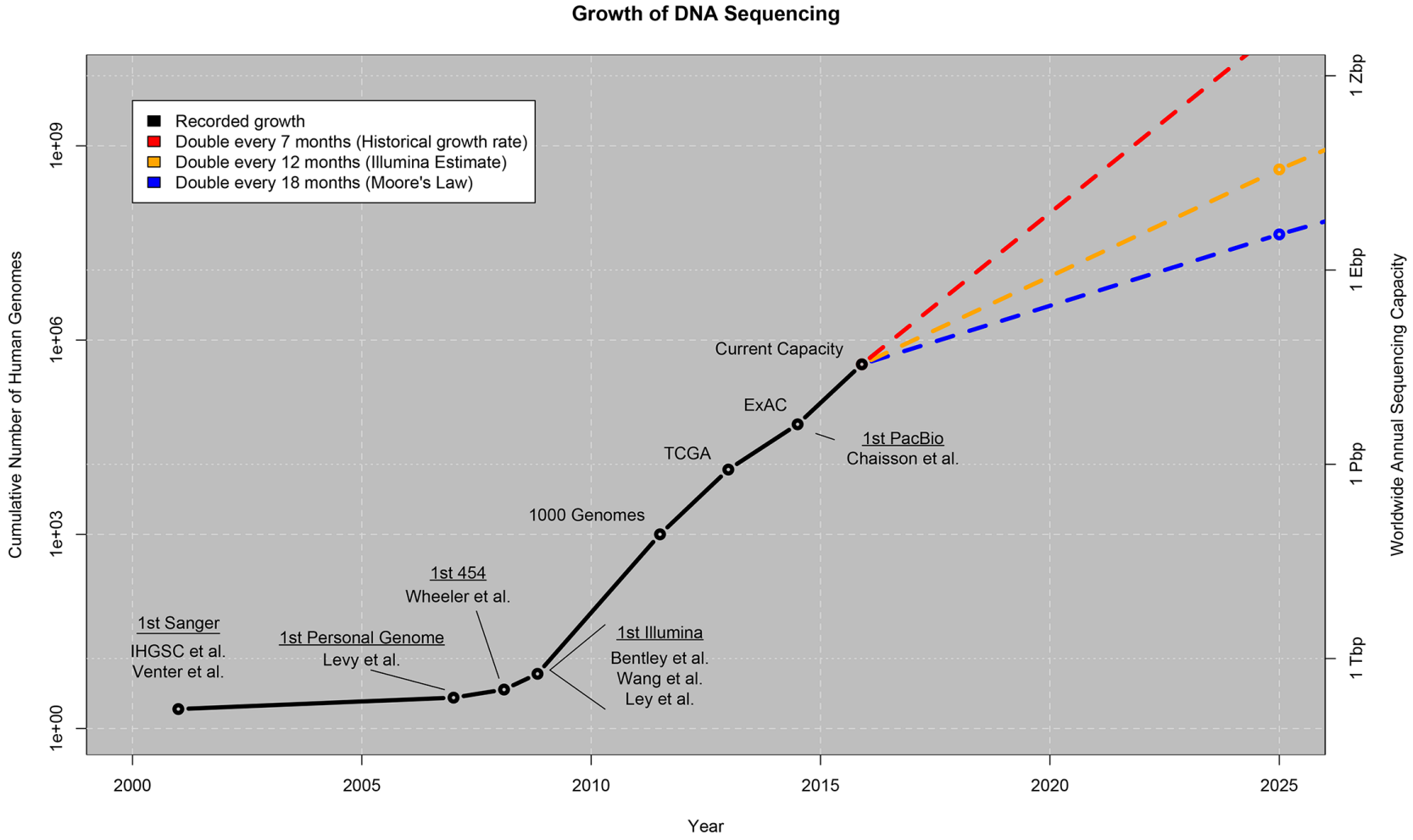
1. **Background:** whole genome sequencing evaluation
2. **Problem #1**
3. **Solution**
4. **Results**
5. **Problem #2**
6. **Solution**
7. **Results**

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

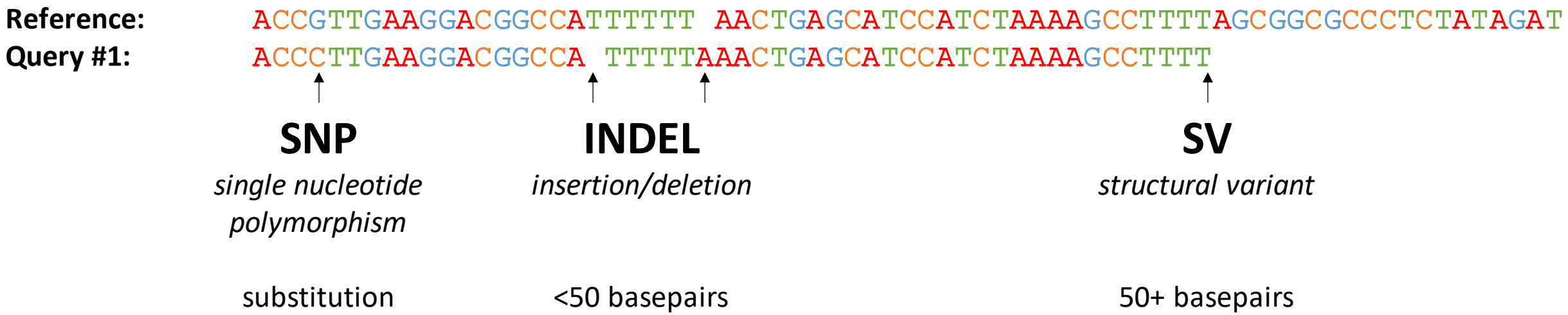
	CHROM	POS	REF	ALT
✓	chr14	6	T	A
✓	chr14	9	A	G

Ground Truth

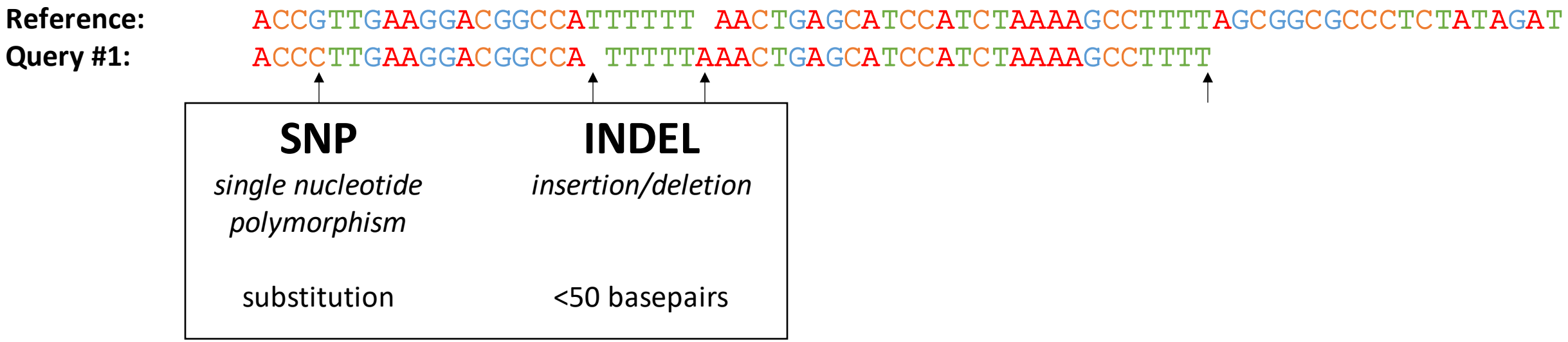
Reference: ACCGTTGAAG
Query: ACCGTAGAGG

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

Benchmarking: *stratification by variant type*



Benchmarking: *small variants only*



Tim Dunn, Satish Narayanasamy. “vcfdist: accurately benchmarking phased small variant calls in human genomes”. Nature Communications, 2023.

Benchmarking: *precision-recall curves*

Reference: ACCGTTGAAGGACGGCCATTTTTT AACTGAGCATCCATCTAAAAGCCTTTTAGCGGCGCCCTCTATAGAT

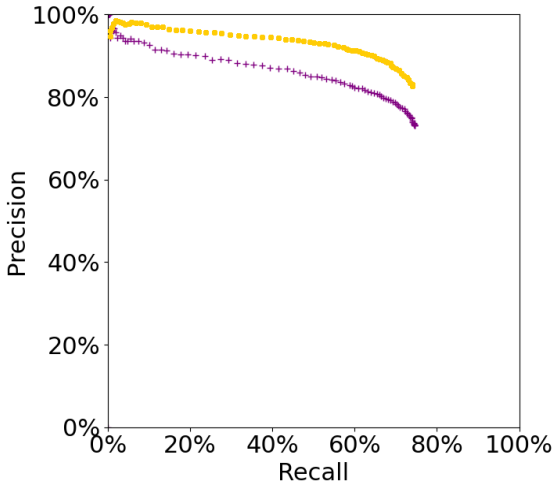
Query #1: ACCCTTGAAGGACGGCCA TTTTTAAACTGAGCATCCATCTAAAAGCCTTTT

SNP	**INDEL**
single nucleotide polymorphism	*insertion/deletion*
substitution	<50 basepairs

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

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

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



Tim Dunn, Satish Narayanasamy. “vcfdist: accurately benchmarking phased small variant calls in human genomes”. Nature Communications, 2023.

Outline

1. **Background:** whole genome sequencing evaluation
2. **Problem #1:** complex variants
3. **Solution**
4. **Results**
5. **Problem #2**
6. **Solution**
7. **Results**

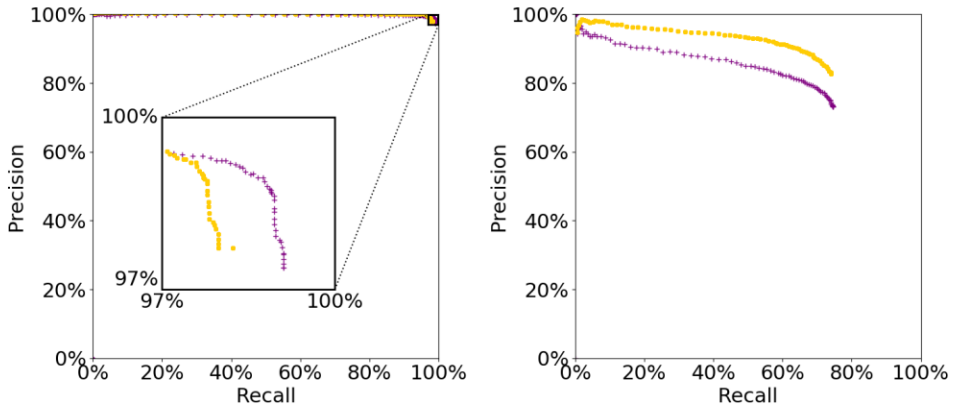
Problem: *variant representation matters*

Reference: ACCGTTGAAGGACGGCCATTTTTT AACTGAGCATCCATCTAAAAGCCTTTTAGCGGCGCCCTCTATAGAT

Query #1: ACCCTTGAAGGACGGCCA TTTTTAAACTGAGCATCCATCTAAAAGCCTTTT

Query #2: ACCCTTGAAGGACGGCCATTTTTA AACTGAGCATCCATCTAAAAGCCTTTT

SNP **INDEL**



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

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

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

Prior work: **vcfeval**

Cleary. et al. "Comparing variant call files for performance benchmarking of next-generation sequencing variant calling pipelines." *BioRxiv,* 2015.

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. **Background**: whole-genome sequencing evaluation
2. **Problem #1:** complex variants
3. **Solution:** variant normalization
4. **Results**
5. **Problem #2**
6. **Solution**
7. **Results**

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

Original	Decomposed	Trimmed
AAGGAAA-TC	AAGGAAA-TC	AAGGAAA-TC
ATCGAAAATC	ATCGAAAATC	ATCGAAAATC

VCF

Original			Decomposed			Trimmed		
POS	REF	ALT	POS	REF	ALT	POS	REF	ALT
2	AG	TC	2	A	T	2	A	T
6	AATC	AAATC	3	G	C	3	G	C
			6	AATC	AAATC	7	A	AA

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.

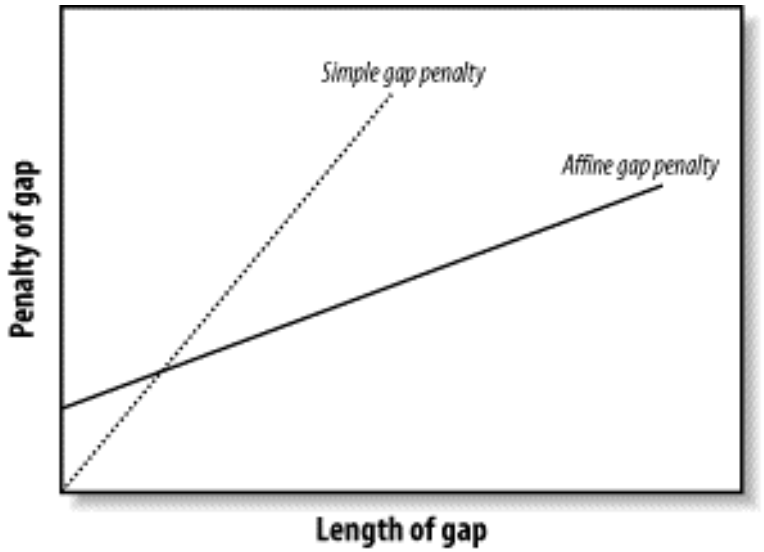
Choosing representations: *best-alignment normalization*

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
```

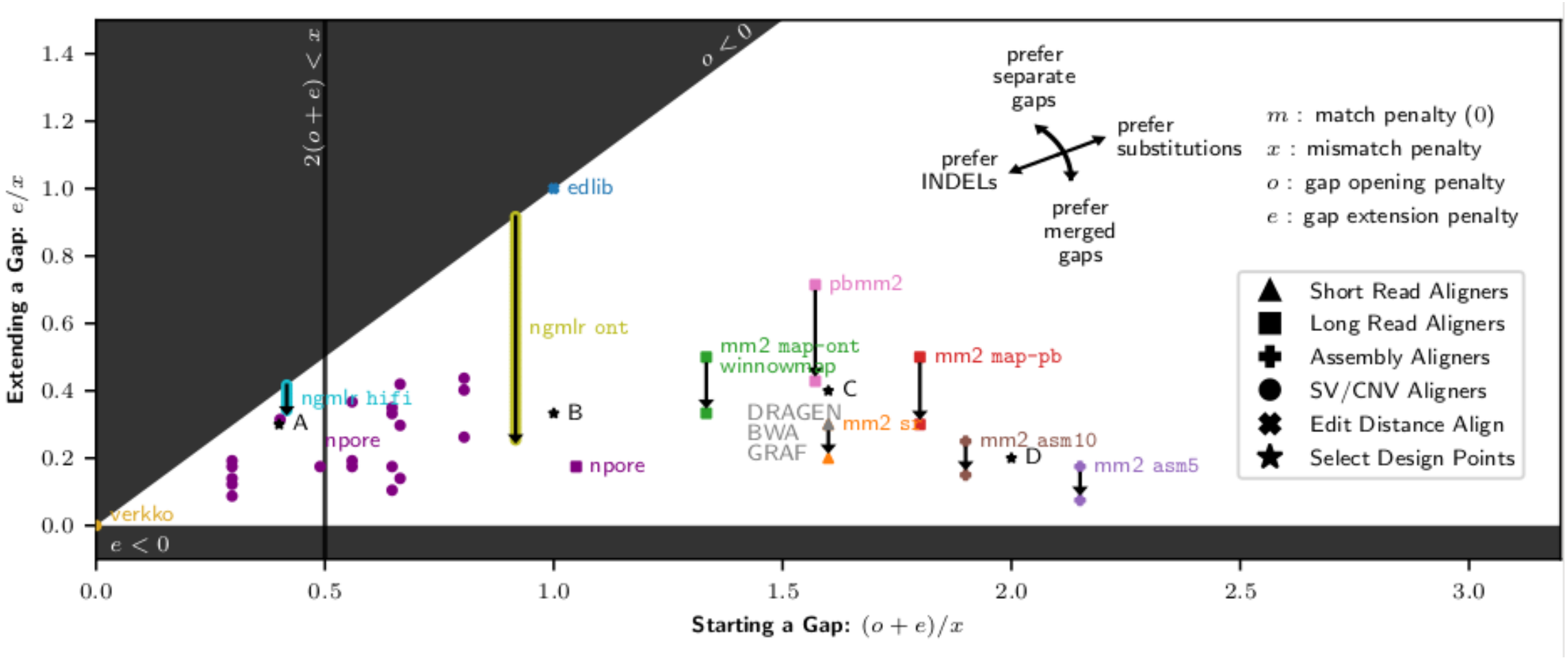
POS	REF	ALT
1	AAGG	A
1	A	ATCGA

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

26

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

Alignment-based normalization design space



Tim Dunn, Satish Narayanasamy. “vcfdist: accurately benchmarking phased small variant calls in human genomes”. Nature Communications, 2023.

Outline

1. **Background**: whole-genome sequencing evaluation
2. **Problem #1:** complex variants
3. **Solution:** variant normalization
4. **Results:** stable evaluations
5. **Problem #2**
6. **Solution**
7. **Results**

Example: *tandem repeat benchmark representation*

Dataset	SNPs	INDELS
Original Representation	917,255	431,545
Normalized At Point C	502,076	461,258

English et al. “Benchmarking of small and large variants across tandem repeats”. bioRxiv, 2023.

Example: *tandem repeat benchmark representation*

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)

Other Contributions

- Efficient variant clustering
- Allow inexact variant matches
- Use phasing information for evaluation
- Distance based evaluation metrics

Tim Dunn, Satish Narayanasamy. “vcfdist: accurately benchmarking phased small variant calls in human genomes”. Nature Communications, 2023.

Example: *inexact complex variant matches*

Query:

CHROM	POS	REF	ALT	CALL	CREDIT
chr1	976722	C	CAGGAACCGCCTCCCACTCCCCCCACAACCCCGGGAACCGCCTCCCACTC CCCCCGCAACCCCGGGAACCGCCTCCCACTCCCCCCGCAACCCC	INS TP	0.979
chr1	976745	G	A	SNP TP	0.979

Truth:

CHROM	POS	REF	ALT	CALL	CREDIT
chr1	976715	A	ACAACCCCAGGAACCGCCTCCCACTCCCCCCA	INS TP	0.979
chr1	976747	C	CAACCCCGGGAACCGCCTCCCACTCCCCCCG	INS TP	0.979
chr1	976777	G	A	SNP TP	0.979
chr1	976811	C	CAACCCCGGGAACCGCCTCCCACTCCCCCCG	INS TP	0.979
chr1	976840	C	G	SNP TP	0.979
chr1	976841	G	A	SNP TP	0.979

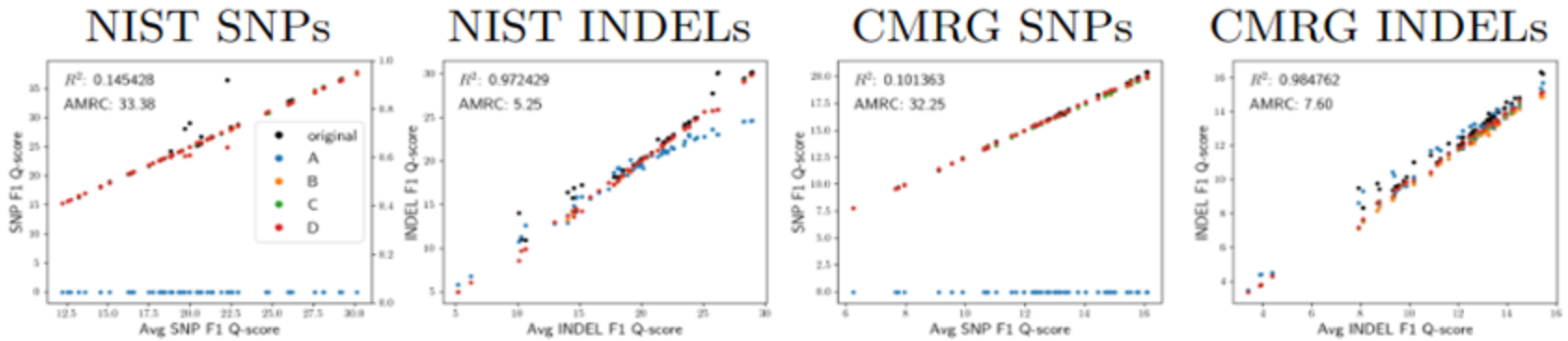
Dataset: *PrecisionFDA Truth Challenge V2*

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

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

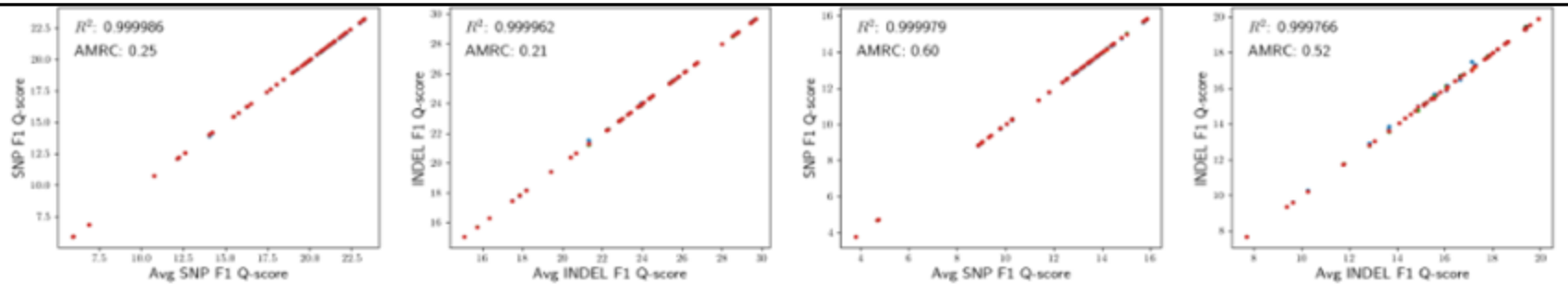
Results: *stable performance across representations*

vcfeval
precision/recall
metrics



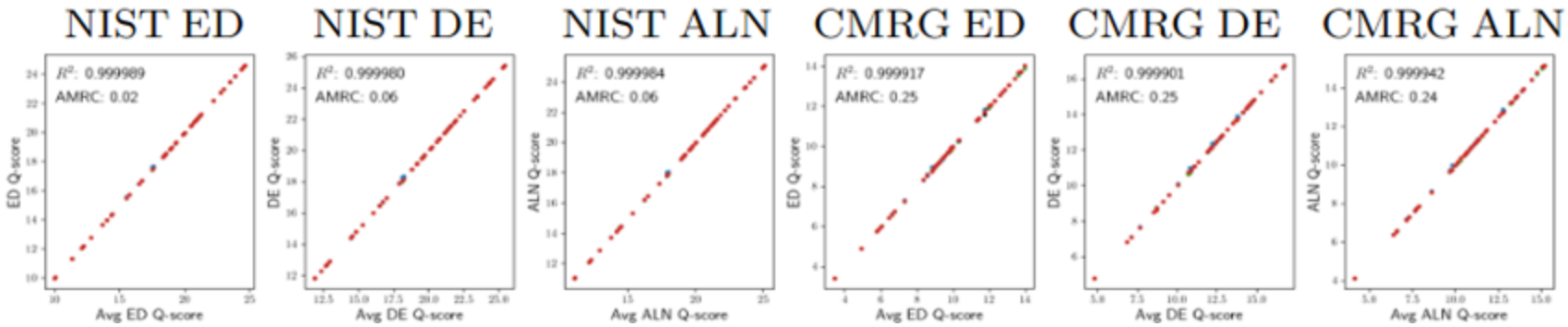
Prior Work

vcfdist
precision/recall
metrics



This Work

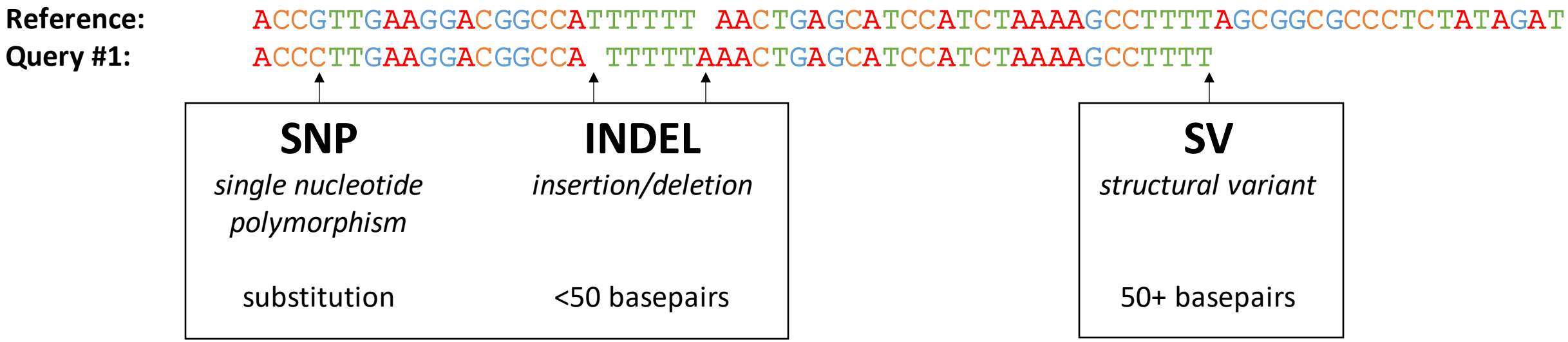
vcfdist
distance
metrics



Outline

1. **Background**: whole-genome sequencing evaluation
2. **Problem #1:** complex variants
3. **Solution:** variant normalization
4. **Results:** stable evaluations
5. **Problem #2:** separate evaluation of small and structural variants
6. **Solution**
7. **Results**

Problem: *separate evaluations for small and SVs*



Example: *variant matches across size categories*

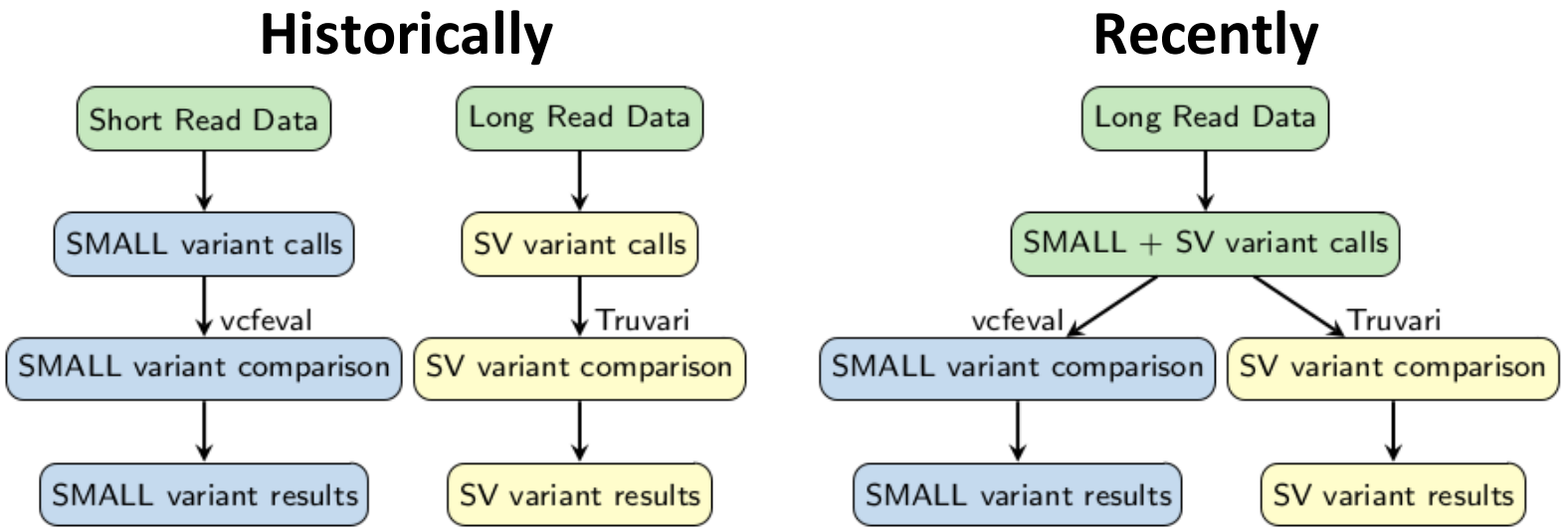
Query:

CHROM	POS	REF	ALT	CALL	CREDIT
chr1	976722	C	CAGGAACCGCCTCCCACTCCCCCCACAACCCCGGGAACCGCCTCCCACTC CCCCCGCAACCCCGGGAACCGCCTCCCACTCCCCCCGCAACCCC	INS TP	0.979
chr1	976745	G	A	SNP TP	0.979

Truth:

CHROM	POS	REF	ALT	CALL	CREDIT
chr1	976715	A	ACAACCCCAGGAACCGCCTCCCACTCCCCCCA	INS TP	0.979
chr1	976747	C	CAACCCCGGGAACCGCCTCCCACTCCCCCCG	INS TP	0.979
chr1	976777	G	A	SNP TP	0.979
chr1	976811	C	CAACCCCGGGAACCGCCTCCCACTCCCCCCG	INS TP	0.979
chr1	976840	C	G	SNP TP	0.979
chr1	976841	G	A	SNP TP	0.979

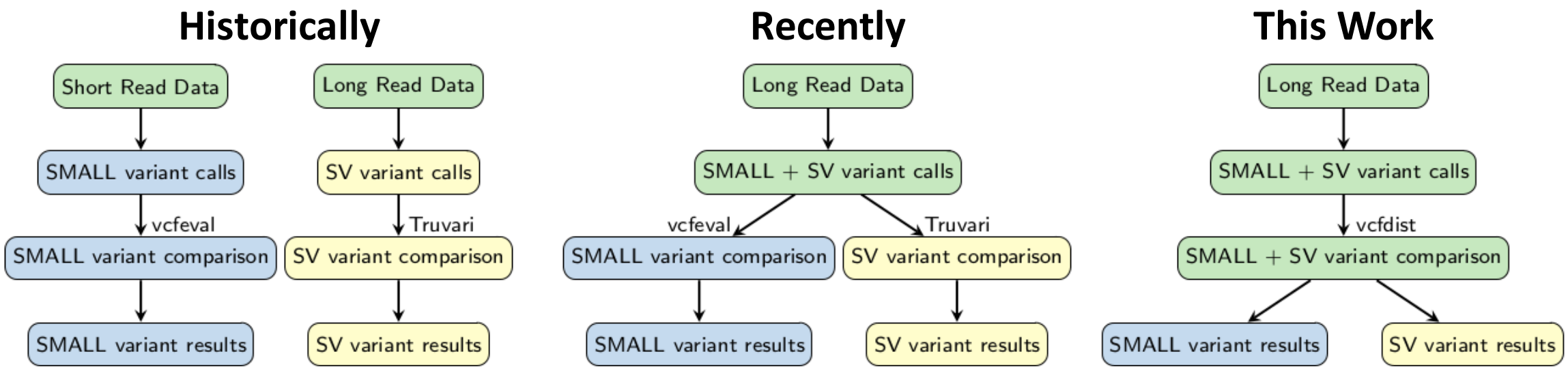
Why: *short-read mapping, inaccurate long reads*



Outline

1. **Background**: whole-genome sequencing evaluation
2. **Problem #1:** complex variants
3. **Solution:** variant normalization
4. **Results:** stable evaluations
5. **Problem #2:** separate evaluation of small and structural variants
6. **Solution:** joint evaluation of small and structural variants
7. **Results**

Solution: *joint small and SV evaluation*

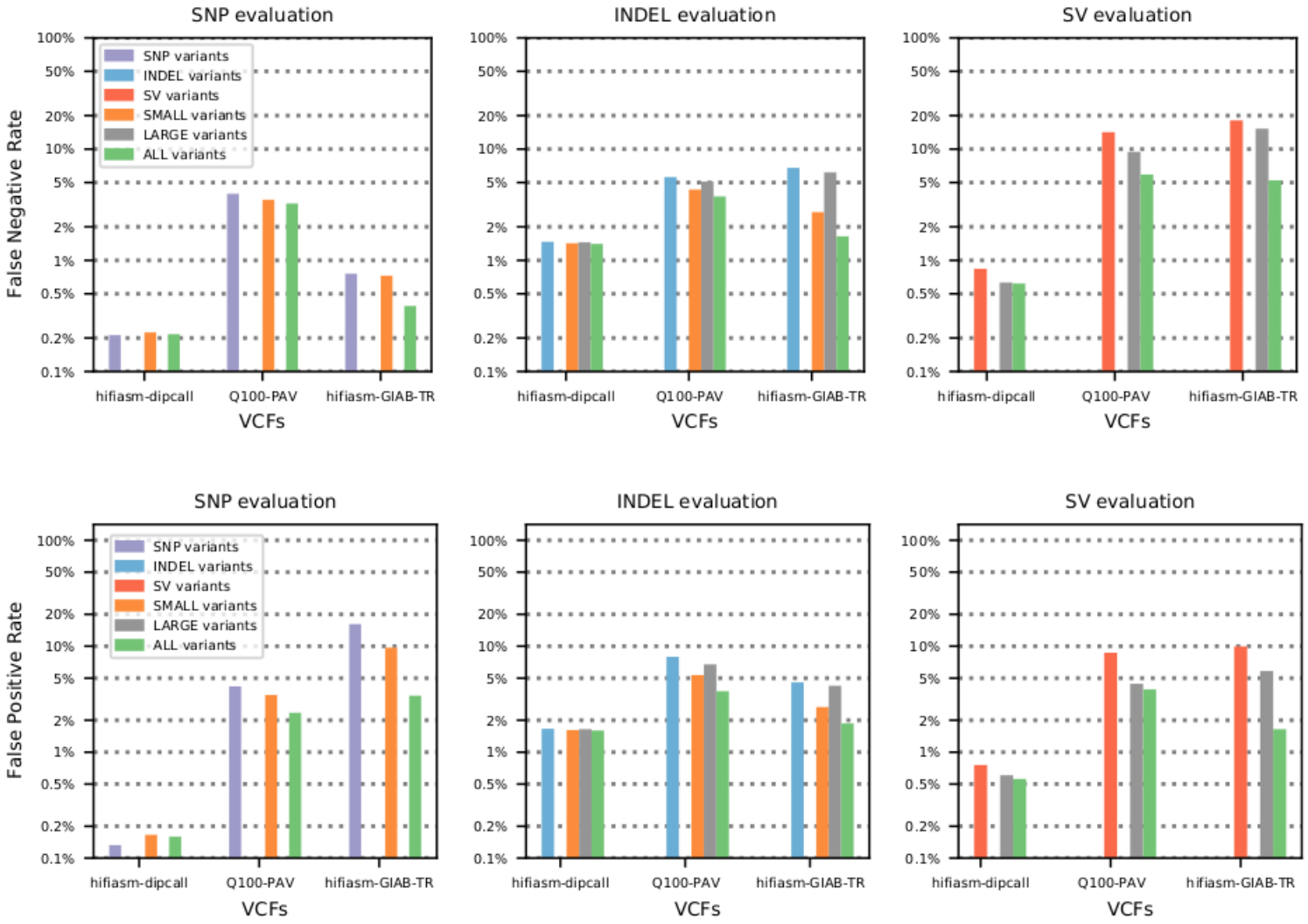


Dunn et al. “Jointly benchmarking small and structural variant calls with vcfdist”. bioRxiv, 2024.

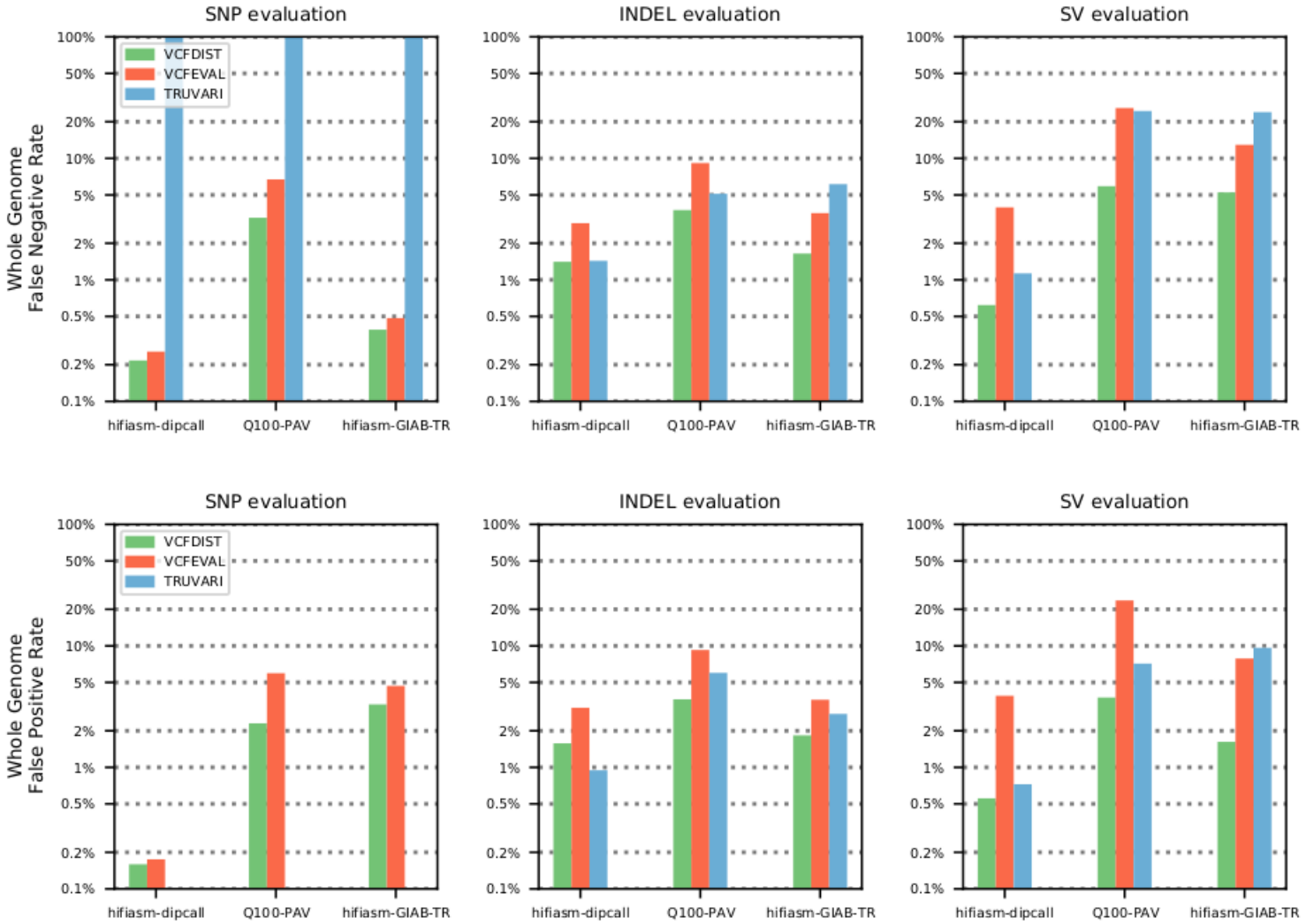
Outline

1. **Background**: whole-genome sequencing evaluation
2. **Problem #1:** complex variants
3. **Solution:** variant normalization
4. **Results:** stable evaluations
5. **Problem #2:** separate evaluation of small and structural variants
6. **Solution:** joint evaluation of small and structural variants
7. **Results**: accurate evaluations

Results: *joint small and SV evaluation*



Results: *comparison to prior work*



Results: *better phasing evaluations*

Dataset	Tool	Switches	Flips
hifiasm-dipcall	WhatsHap	610	396
	vcfdist	494	390
Q100-PAV	WhatsHap	324	433
	vcfdist	6	52
hifiasm-GIAB-TR	WhatsHap	1074	1004
	vcfdist	494	396

Results: *better phasing evaluations*

CONTIG	POS	REF	ALT	FORMAT	TRUTH	QUERY
chr1	32,653,646	T	G	GT:BD:BC	0|1:TP:1.0	0|1:TP:1.0
chr1	32,653,657	TTTG	T	GT:BD:BC	0|1:TP:1.0	.:.:.
chr1	32,653,658	TTG	T	GT:BD:BC	.:.:.	0|1:TP:1.0
chr1	32,653,658	TTG	T	GT:BD:BC	1|0:TP:1.0	.:.:.
chr1	32,653,659	TG	T	GT:BD:BC	.:.:.	1|0:TP:1.0
chr1	32,653,665	TG	T	GT:BD:BC	.:.:.	1|1:TP:1.0
chr1	32,653,666	G	T	GT:BD:BC	1|1:TP:1.0	.:.:.

	Position	37	46	47	58	59	60	61	66	67
	Reference	GTTTTTTTT	T	TTTTTTTTTT	T	T	G	TTTTTT	G	TTTT
Haplotype 1	Truth	GTTTTTTTT	T	TTTTTTTTTT	T			TTTTTT	T	TTTT
	Query	GTTTTTTTT	T	TTTTTTTTTT	T	T		TTTTTT		TTTT
Haplotype 2	Truth	GTTTTTTTT	G	TTTTTTTTTT				TTTTTT	T	TTTT
	Query	GTTTTTTTT	G	TTTTTTTTTT	T			TTTTTT		TTTT

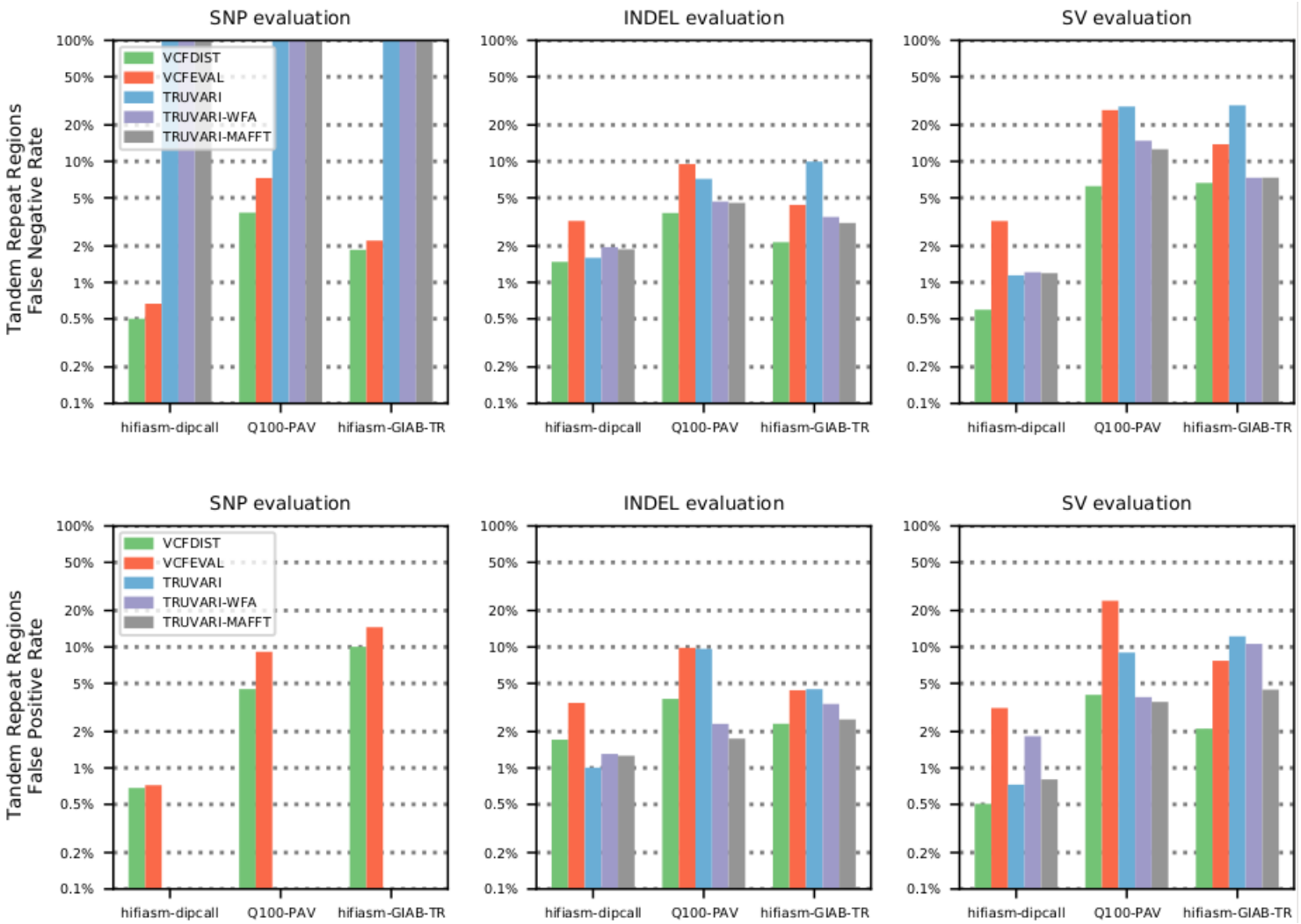
Conclusion

`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

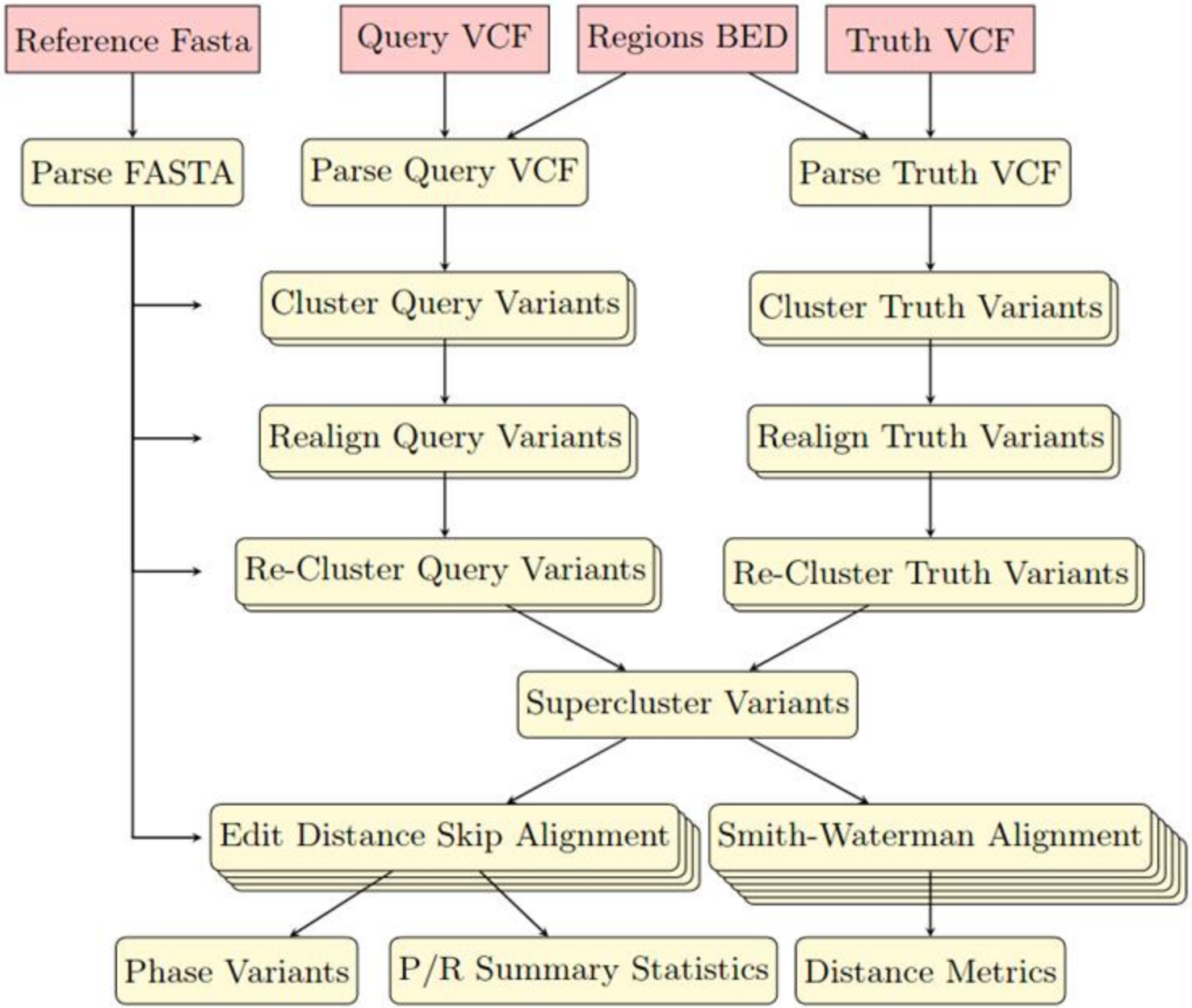
Supplementary Slides

Results: *comparison to prior work*

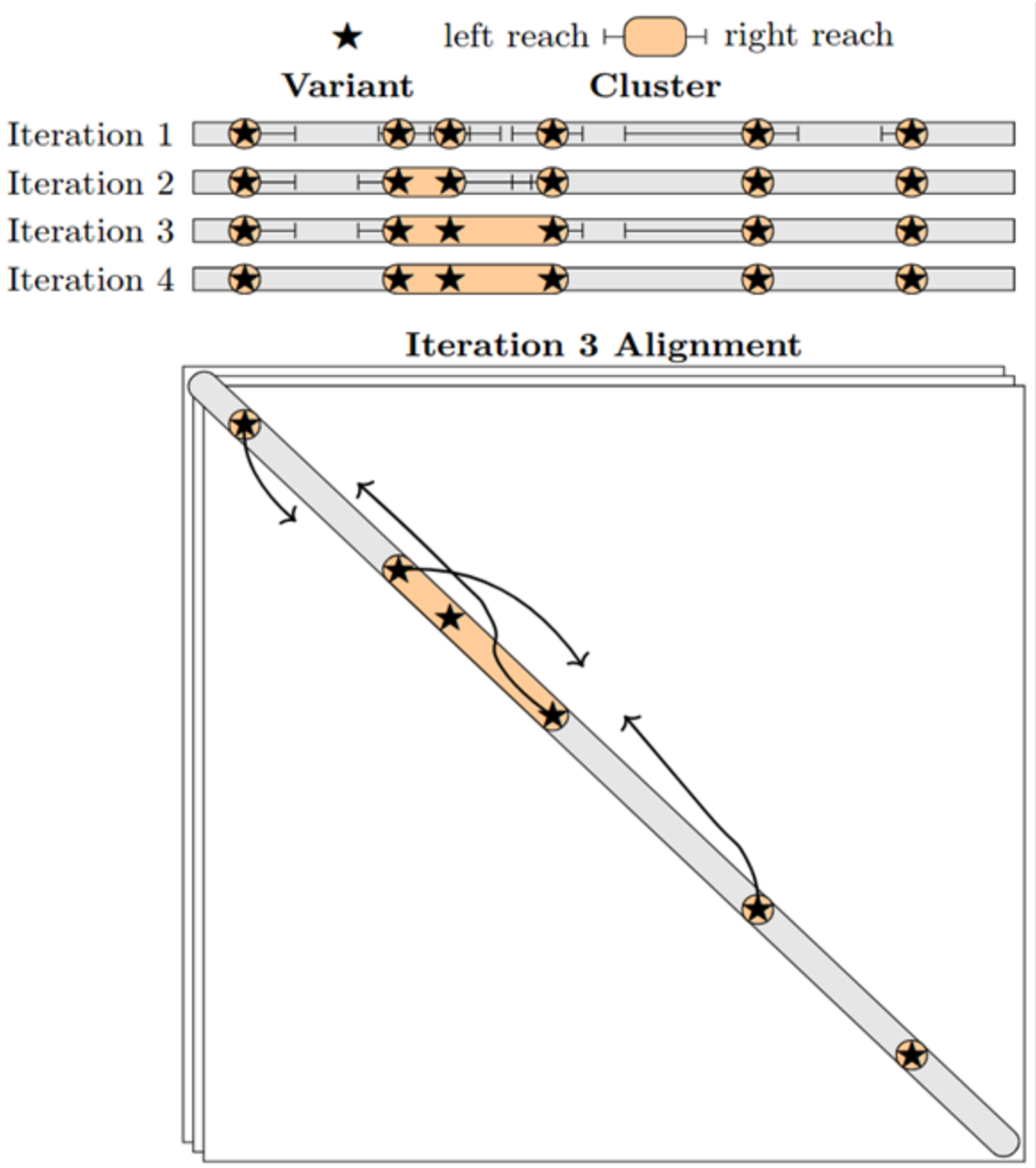


Supplementary Slides: *Implementation*

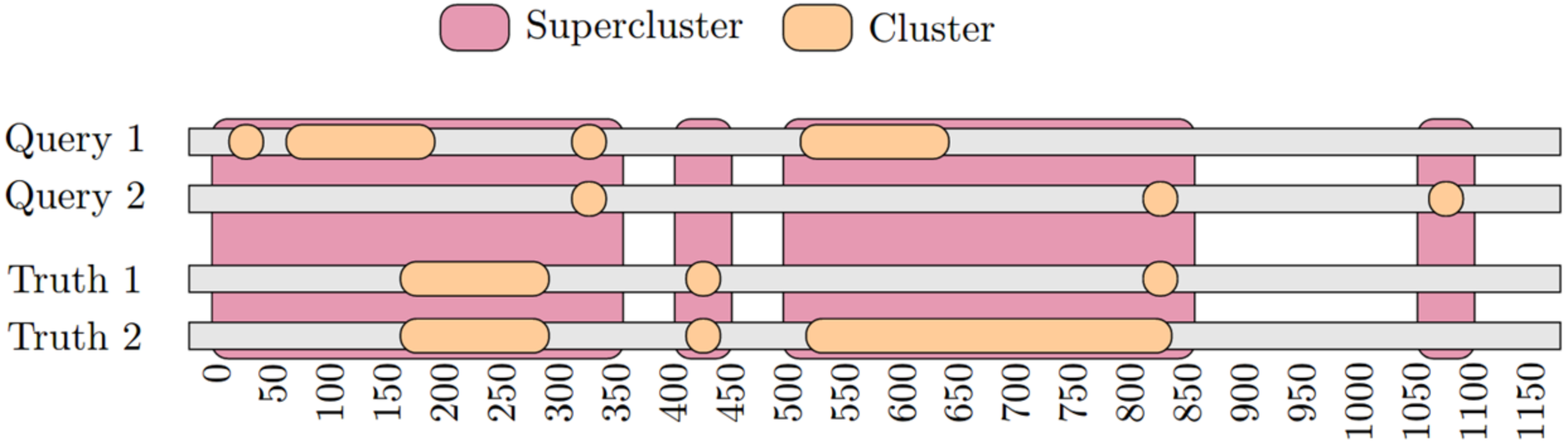
Overview


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



Superclustering: *simple reference distance heuristic*



Precision/Recall:

edit distance, allows skipping FP query variants, backtracking