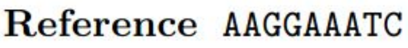
# vcfdist

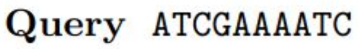
## Accurately benchmarking phased small variant calls in human genomes

Tim Dunn
University of Michigan

# Background: Variant Call Format (VCF) Normalization

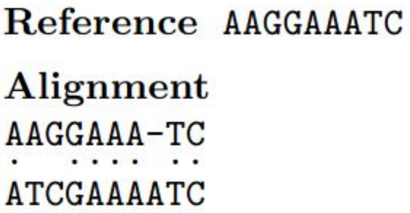


Reference AAGGAAATC



Query ATCGAAAATC

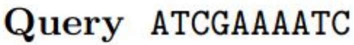
# Background: Variant Call Format (VCF) Normalization



**Reference** AAGGAAATC

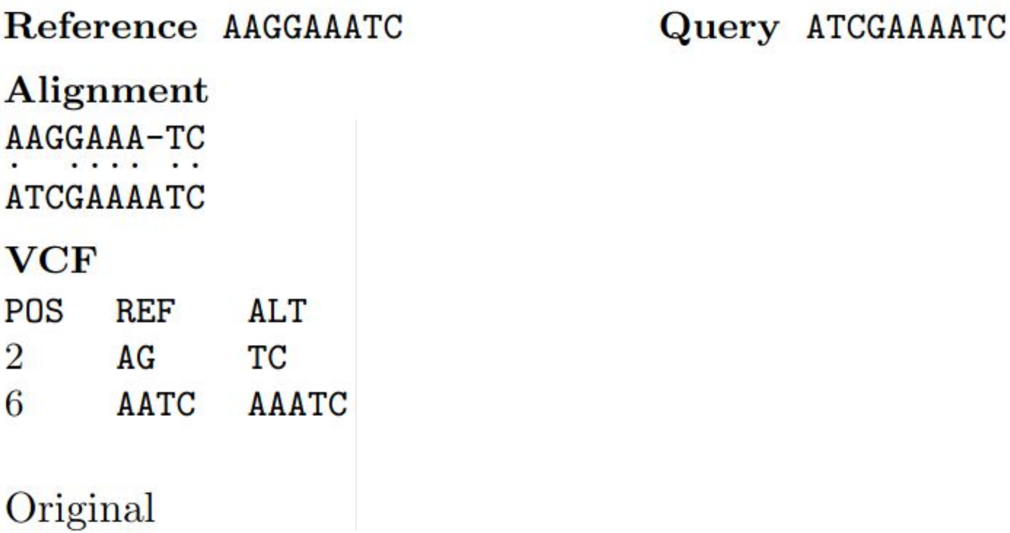
**Alignment**

```
AAGGAAA-TC
.  .... ..
ATCGAAAATC
```



**Query** ATCGAAAATC

# Background: Variant Call Format (VCF) Normalization



**Reference** AAGGAAATC

**Query** ATCGAAAATC

**Alignment**

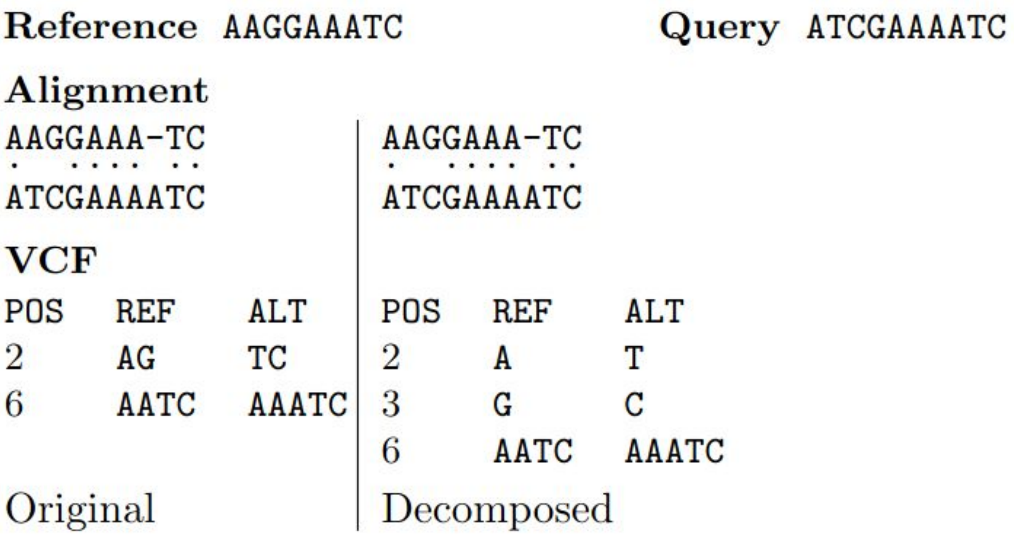
```
AAGGAAA-TC
.  .... ..
ATCGAAAATC
```

**VCF**

| POS | REF | ALT |
|---|---|---|
| 2 | AG | TC |
| 6 | AATC | AAATC |

Original

# Background: Variant Call Format (VCF) Normalization



**Reference** AAGGAAATC **Query** ATCGAAAATC

**Alignment**

```
AAGGAAA-TC     |  AAGGAAA-TC
.  .... ..     |  .  .... ..
ATCGAAAATC     |  ATCGAAAATC
```

**VCF**

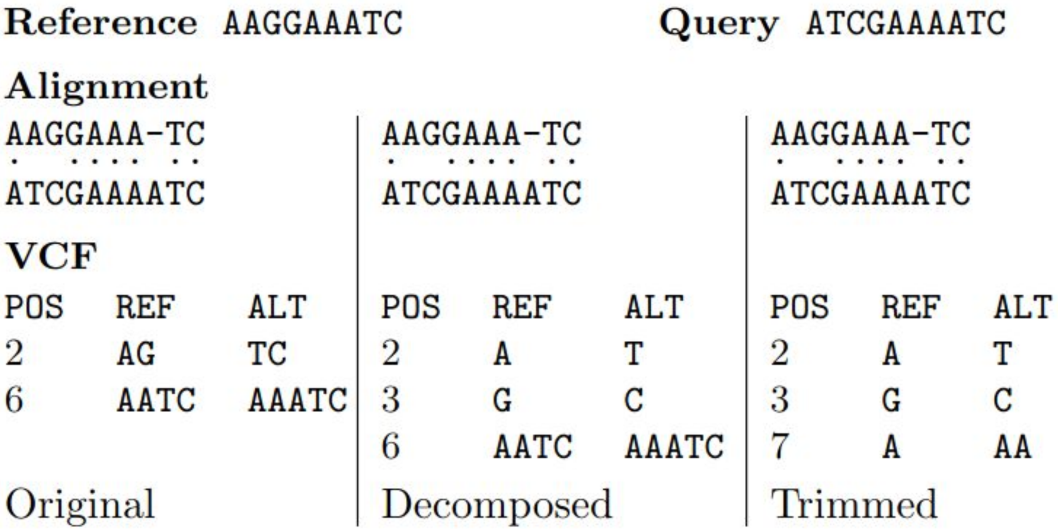
| POS | REF | ALT |
|---|---|---|
| 2 | AG | TC |
| 6 | AATC | AAATC |

Original

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

Decomposed

# Background: Variant Call Format (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 |

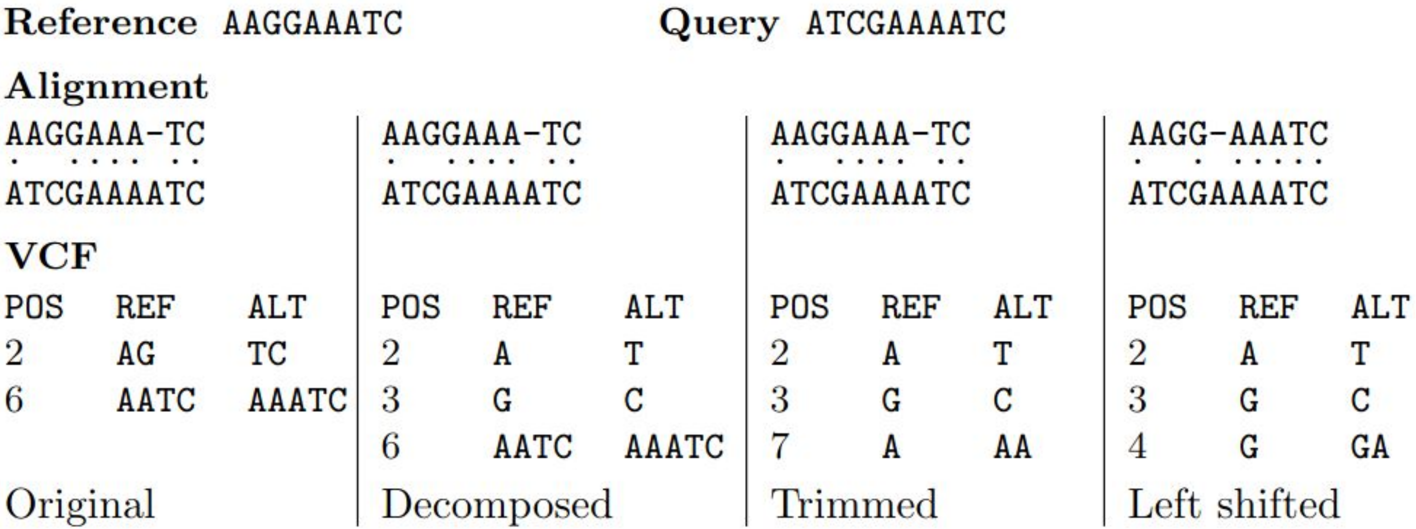
Decomposed

```
AAGGAAA-TC
.  .... ..
ATCGAAAATC
```

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

Trimmed

# Background: Variant Call Format (VCF) Normalization



**Reference** AAGGAAATC **Query** ATCGAAAATC

**Alignment**

```
AAGGAAA-TC | AAGGAAA-TC | AAGGAAA-TC | AAGG-AAATC
.  .... .. | .  .... .. | .  .... .. | .  . .....
ATCGAAAATC | ATCGAAAATC | ATCGAAAATC | ATCGAAAATC
```

**VCF**

| POS | REF | ALT |
|---|---|---|
| 2 | AG | TC |
| 6 | AATC | AAATC |

Original

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

Decomposed

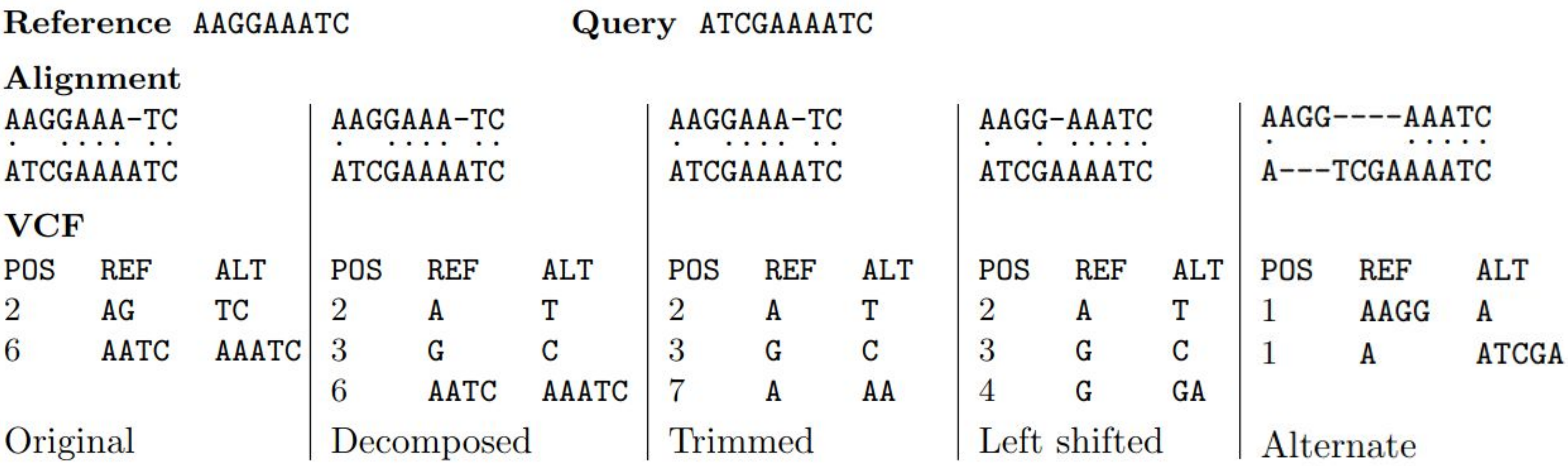
| POS | REF | ALT |
|---|---|---|
| 2 | A | T |
| 3 | G | C |
| 7 | A | AA |

Trimmed

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

Left shifted

# Background: Variant Call Format (VCF) Normalization



**Reference** AAGGAAATC **Query** ATCGAAAATC

**Alignment**

**VCF**

Original

```
AAGGAAA-TC
.  .... ..
ATCGAAAATC
```

| POS | REF | ALT |
|---|---|---|
| 2 | AG | TC |
| 6 | AATC | AAATC |

Decomposed

```
AAGGAAA-TC
.  .... ..
ATCGAAAATC
```

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

Trimmed

```
AAGGAAA-TC
.  .... ..
ATCGAAAATC
```

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

Left shifted

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

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

Alternate

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

| POS | REF | ALT |
|---|---|---|
| 1 | AAGG | A |
| 1 | A | ATCGA |

# Overview

## 1. Key Ideas

A. Standardize complex variant representation

B. Allow partial credit for variant calls

C. Distance-based evaluation metrics

D. Enforce local variant phasing

## 2. Results

Improved stability of variant calling evaluation

## 3. Extension

*"Can we directly evaluate structural variants?"*

# Overview

## 1. Key Ideas

A. Standardize complex variant representation
B. Allow partial credit for variant calls
C. Distance-based evaluation metrics
D. Enforce local variant phasing

## 2. Results

Improved stability of variant calling evaluation

## 3. Extension

*"Can we directly evaluate structural variants?"*

# A. Standardize complex variant representation

"Best alignment normalization" (Bayat, 2016)

*m* = match (0)

*x* = mis-match

*o* = gap opening

*e* = gap extension

# A. Standardize complex variant representation

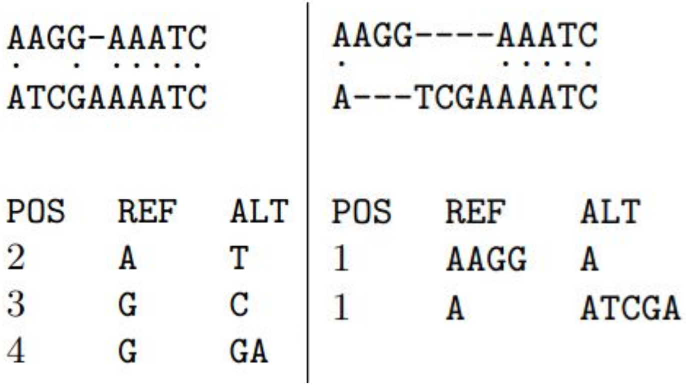
"Best alignment normalization" (Bayat, 2016)

*m* = match (0)

*x* = mis-match

*o* = gap opening

*e* = gap extension



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

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

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

| POS | REF | ALT |
|---|---|---|
| 1 | AAGG | A |
| 1 | A | ATCGA |

x + x + (o+e)

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

# A. Standardize complex variant representation

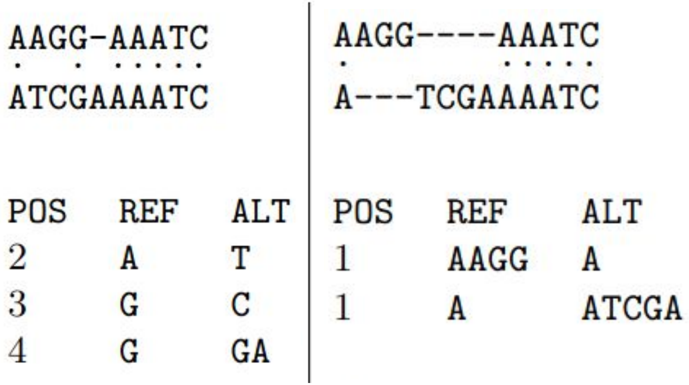
"Best alignment normalization" (Bayat, 2016)

*m* = match (0)

*x* = mis-match (5)

*o* = gap opening (6)

*e* = gap extension (2)



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

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

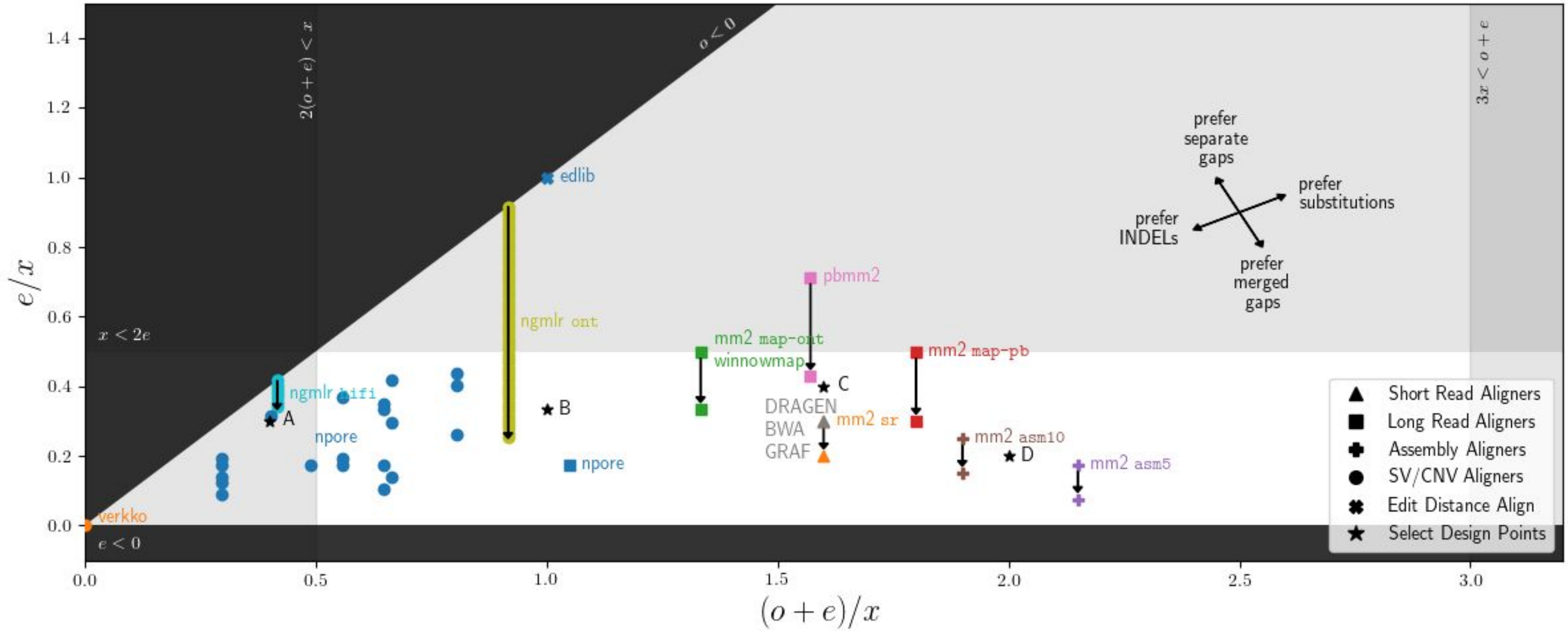
```
AAGG----AAATC
.       .....
A---TCGAAAATC
```

| POS | REF | ALT |
|---|---|---|
| 1 | AAGG | A |
| 1 | A | ATCGA |

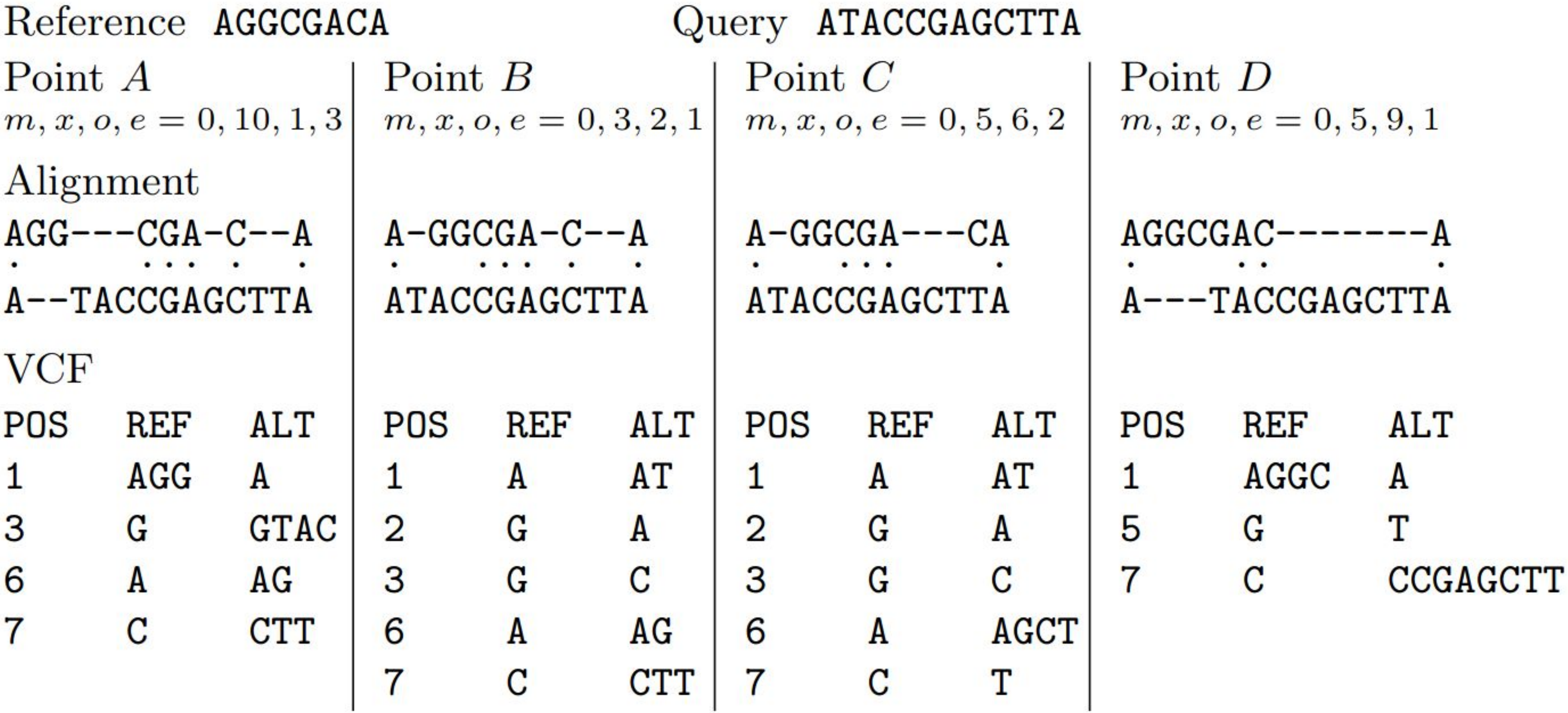
18

26

# A. Standardize complex variant representation



# A. Standardize complex variant representation



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 |

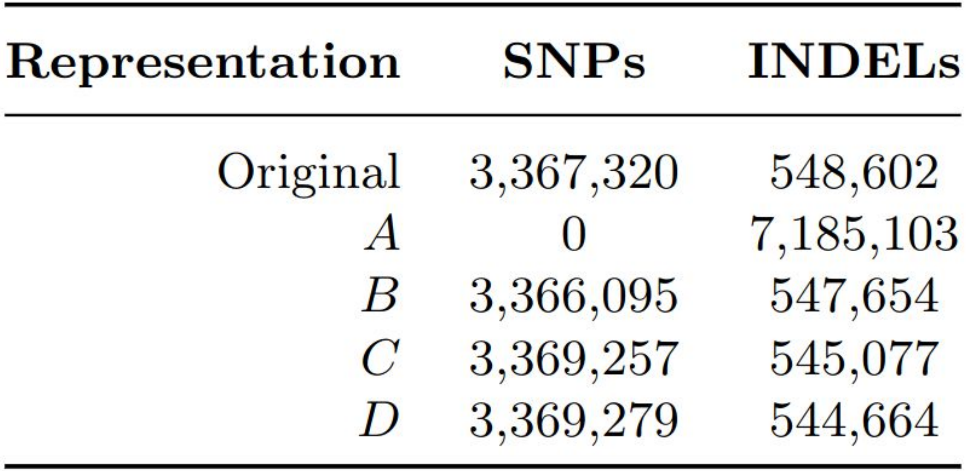
Point *D*

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

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

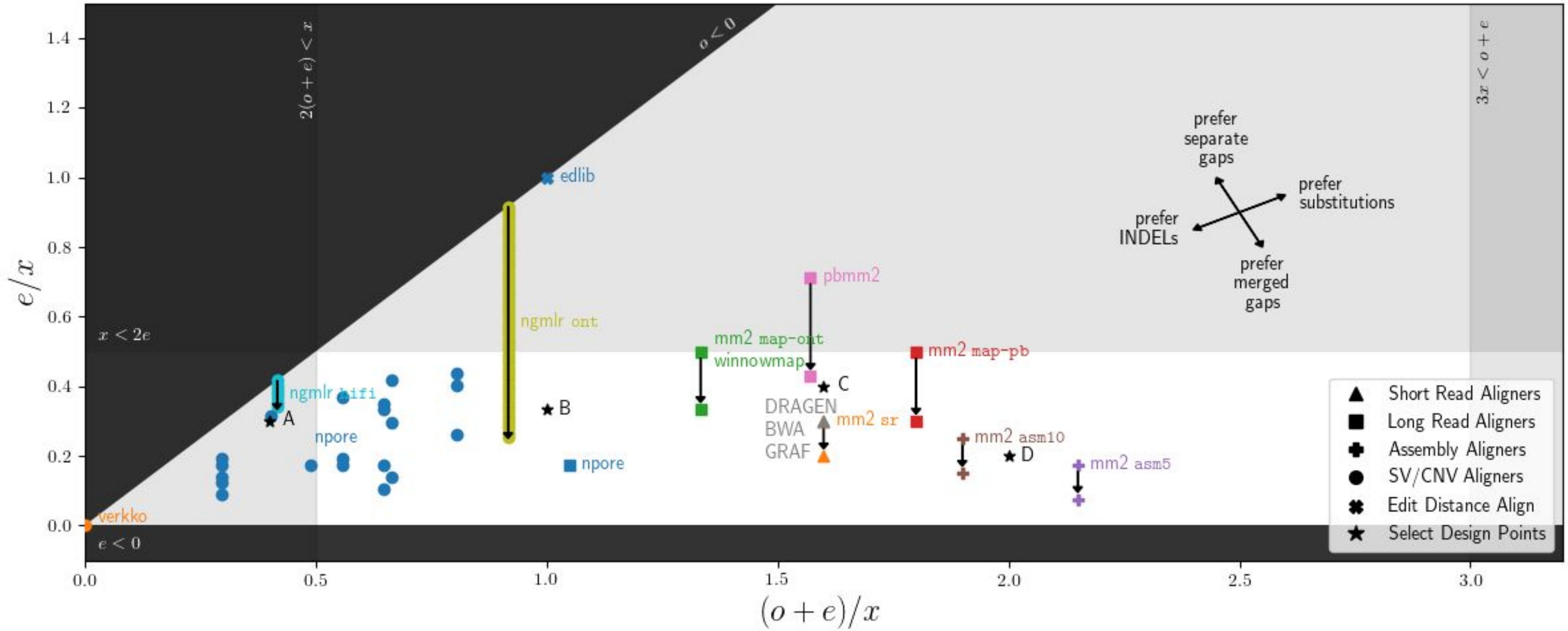
| POS | REF | ALT |
|---|---|---|
| 1 | AGGC | A |
| 5 | G | T |
| 7 | C | CCGAGCTT |

# A. Standardize complex variant representation



| 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 |

# A. Standardize complex variant representation



# Overview

## 1. Key Ideas

A. Standardize complex variant representation

B. Allow partial credit for variant calls

C. Distance-based evaluation metrics

D. Enforce local variant phasing

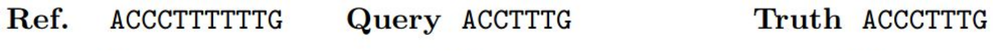
## 2. Results

Improved stability of variant calling evaluation

## 3. Extension

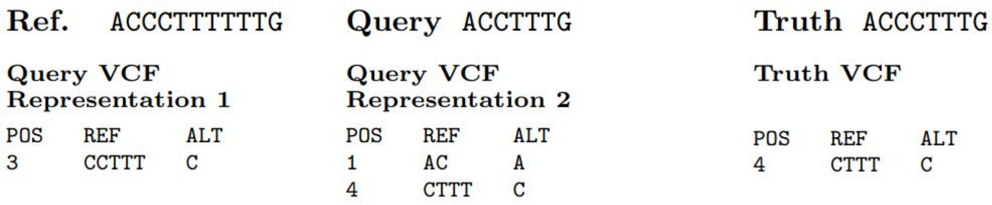
*"Can we directly evaluate structural variants?"*

# B. Allow partial credit for variant calls



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

# B. Allow partial credit for variant calls



**Ref.** ACCCTTTTTTG

**Query VCF**
**Representation 1**

| POS | REF | ALT |
|---|---|---|
| 3 | CCTTT | C |

**Query** ACCTTTG

**Query VCF**
**Representation 2**

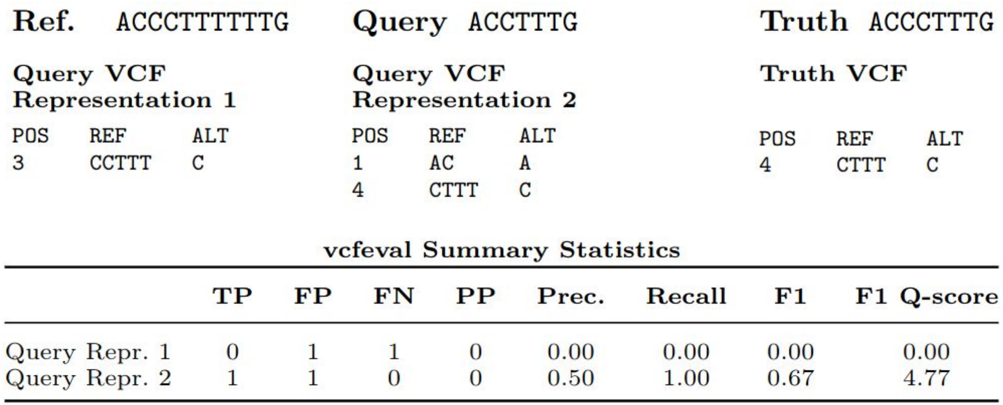
| POS | REF | ALT |
|---|---|---|
| 1 | AC | A |
| 4 | CTTT | C |

**Truth** ACCCTTTG

**Truth VCF**

| POS | REF | ALT |
|---|---|---|
| 4 | CTTT | C |

# B. Allow partial credit for variant calls



**Ref.** `ACCCTTTTTG`

**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 |

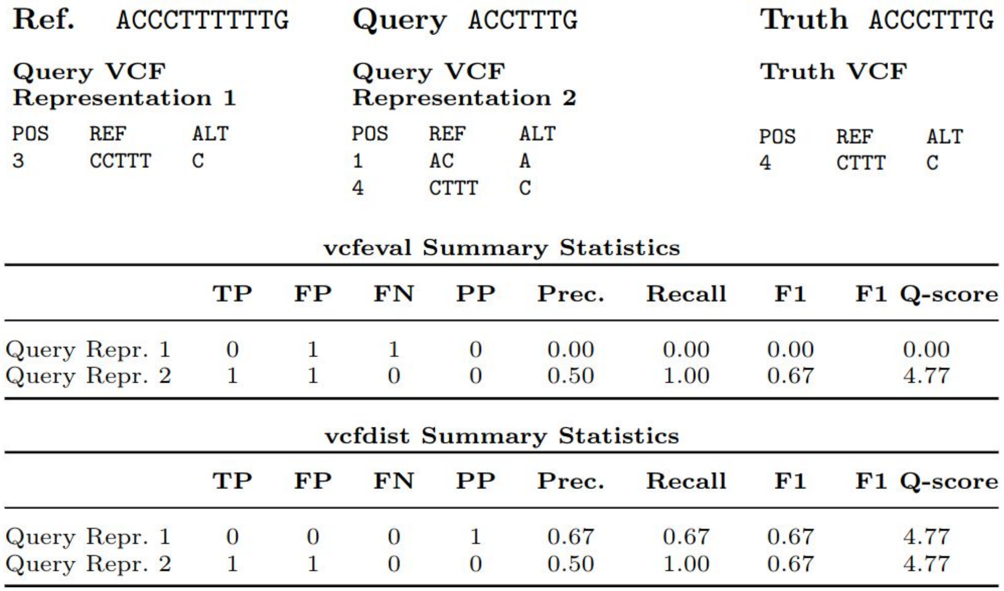
**Truth VCF**

| POS | REF | ALT |
|---|---|---|
| 4 | CTTT | C |

**vcfeval Summary Statistics**

| | TP | FP | FN | PP | Prec. | Recall | F1 | F1 Q-score |
|---|---|---|---|---|---|---|---|---|
| Query Repr. 1 | 0 | 1 | 1 | 0 | 0.00 | 0.00 | 0.00 | 0.00 |
| Query Repr. 2 | 1 | 1 | 0 | 0 | 0.50 | 1.00 | 0.67 | 4.77 |

# B. Allow partial credit for variant calls



**Ref.** ACCCTTTTTG

**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 | F1 Q-score |
|---|---|---|---|---|---|---|---|---|
| Query Repr. 1 | 0 | 1 | 1 | 0 | 0.00 | 0.00 | 0.00 | 0.00 |
| Query Repr. 2 | 1 | 1 | 0 | 0 | 0.50 | 1.00 | 0.67 | 4.77 |

**vcfdist Summary Statistics**

| | TP | FP | FN | PP | Prec. | Recall | F1 | F1 Q-score |
|---|---|---|---|---|---|---|---|---|
| Query Repr. 1 | 0 | 0 | 0 | 1 | 0.67 | 0.67 | 0.67 | 4.77 |
| Query Repr. 2 | 1 | 1 | 0 | 0 | 0.50 | 1.00 | 0.67 | 4.77 |

# Overview

## 1. Key Ideas

A. Standardize complex variant representation
B. Allow partial credit for variant calls
C. Distance-based evaluation metrics
D. Enforce local variant phasing

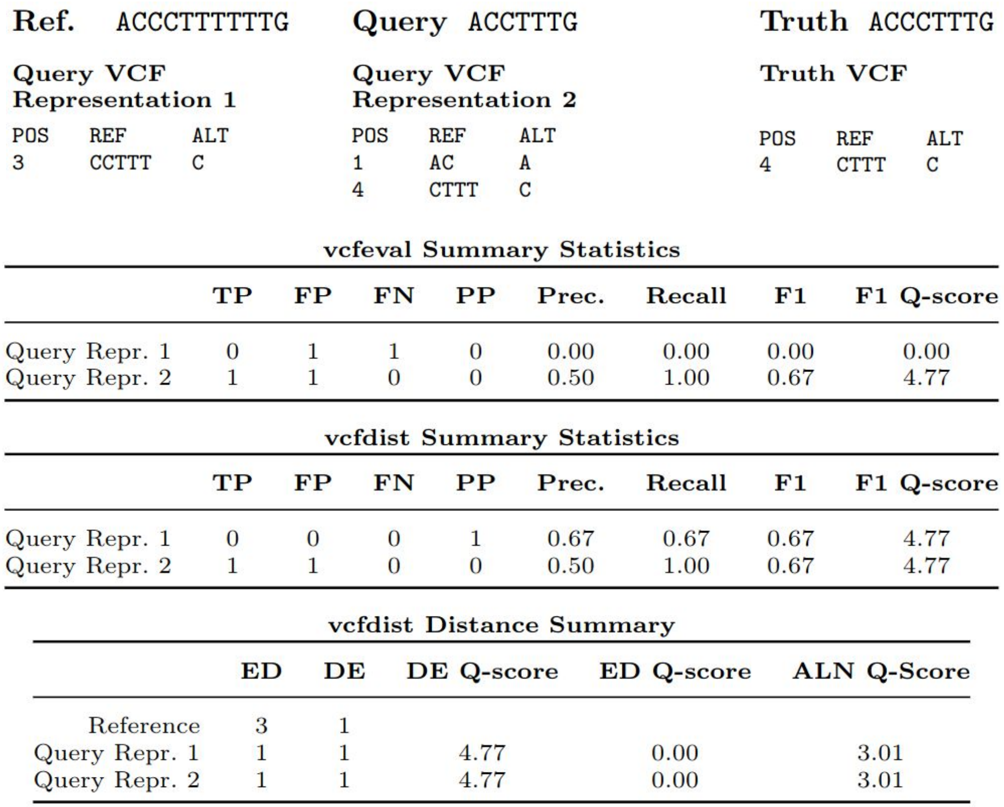
## 2. Results

Improved stability of variant calling evaluation

## 3. Extension

*"Can we directly evaluate structural variants?"*

# C. Distance based evaluation metrics



**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 | F1 Q-score |
|---|---|---|---|---|---|---|---|---|
| Query Repr. 1 | 0 | 1 | 1 | 0 | 0.00 | 0.00 | 0.00 | 0.00 |
| Query Repr. 2 | 1 | 1 | 0 | 0 | 0.50 | 1.00 | 0.67 | 4.77 |

**vcfdist Summary Statistics**

| | TP | FP | FN | PP | Prec. | Recall | F1 | F1 Q-score |
|---|---|---|---|---|---|---|---|---|
| Query Repr. 1 | 0 | 0 | 0 | 1 | 0.67 | 0.67 | 0.67 | 4.77 |
| Query Repr. 2 | 1 | 1 | 0 | 0 | 0.50 | 1.00 | 0.67 | 4.77 |

**vcfdist Distance Summary**

| | ED | DE | DE Q-score | ED Q-score | ALN Q-Score |
|---|---|---|---|---|---|
| Reference | 3 | 1 | | | |
| Query Repr. 1 | 1 | 1 | 4.77 | 0.00 | 3.01 |
| Query Repr. 2 | 1 | 1 | 4.77 | 0.00 | 3.01 |

# Overview

## 1. Key Ideas

A. Standardize complex variant representation
B. Allow partial credit for variant calls
C. Distance-based evaluation metrics
D. Enforce local variant phasing

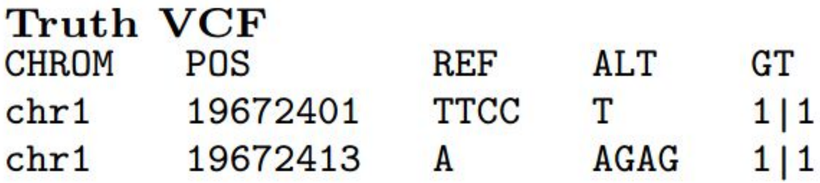
## 2. Results

Improved stability of variant calling evaluation

## 3. Extension

*"Can we directly evaluate structural variants?"*

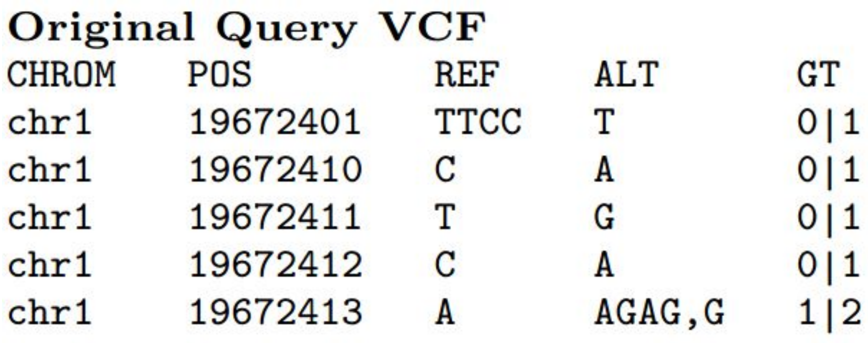
# D. Enforce local variant phasing



**Truth VCF**

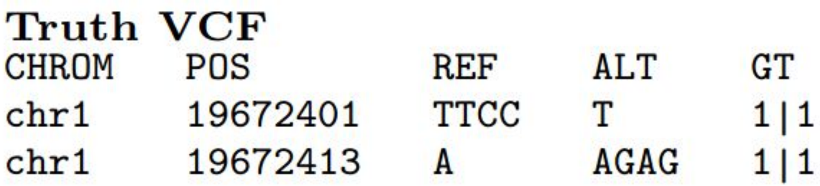
| CHROM | POS | REF | ALT | GT |
|---|---|---|---|---|
| chr1 | 19672401 | TTCC | T | 1\|1 |
| chr1 | 19672413 | A | AGAG | 1\|1 |

# D. Enforce local variant phasing



Original Query VCF

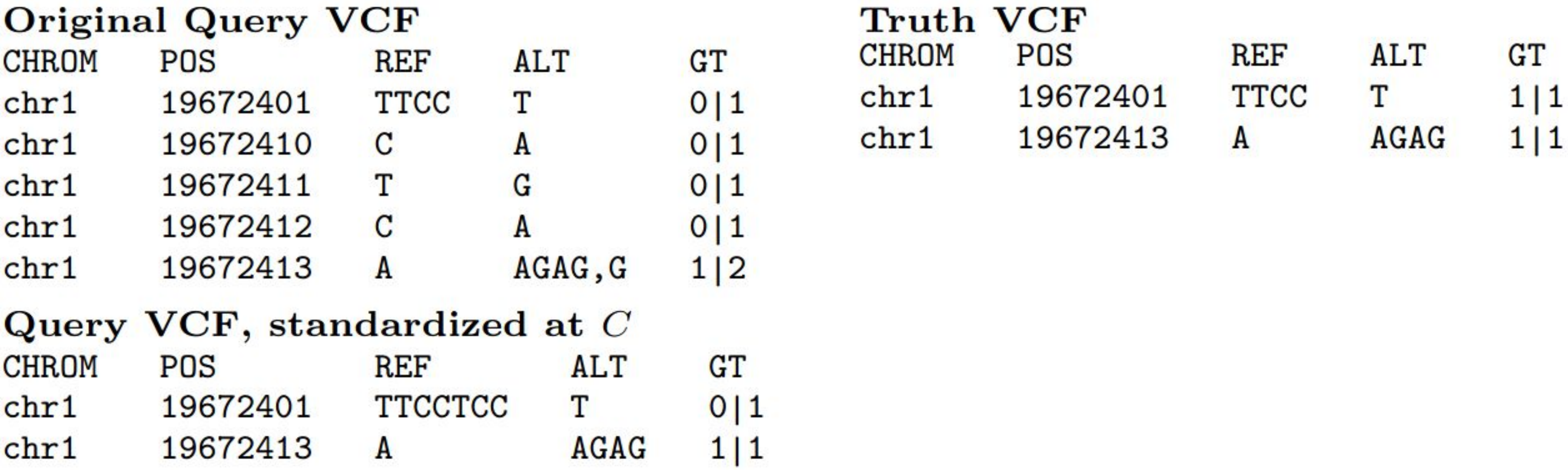
| CHROM | POS | REF | ALT | GT |
|---|---|---|---|---|
| chr1 | 19672401 | TTCC | T | 0\|1 |
| chr1 | 19672410 | C | A | 0\|1 |
| chr1 | 19672411 | T | G | 0\|1 |
| chr1 | 19672412 | C | A | 0\|1 |
| chr1 | 19672413 | A | AGAG,G | 1\|2 |



Truth VCF

| CHROM | POS | REF | ALT | GT |
|---|---|---|---|---|
| chr1 | 19672401 | TTCC | T | 1\|1 |
| chr1 | 19672413 | A | AGAG | 1\|1 |

# D. Enforce local variant phasing



**Original Query VCF**

| CHROM | POS | REF | ALT | GT |
|---|---|---|---|---|
| chr1 | 19672401 | TTCC | T | 0\|1 |
| chr1 | 19672410 | C | A | 0\|1 |
| chr1 | 19672411 | T | G | 0\|1 |
| chr1 | 19672412 | C | A | 0\|1 |
| chr1 | 19672413 | A | AGAG,G | 1\|2 |

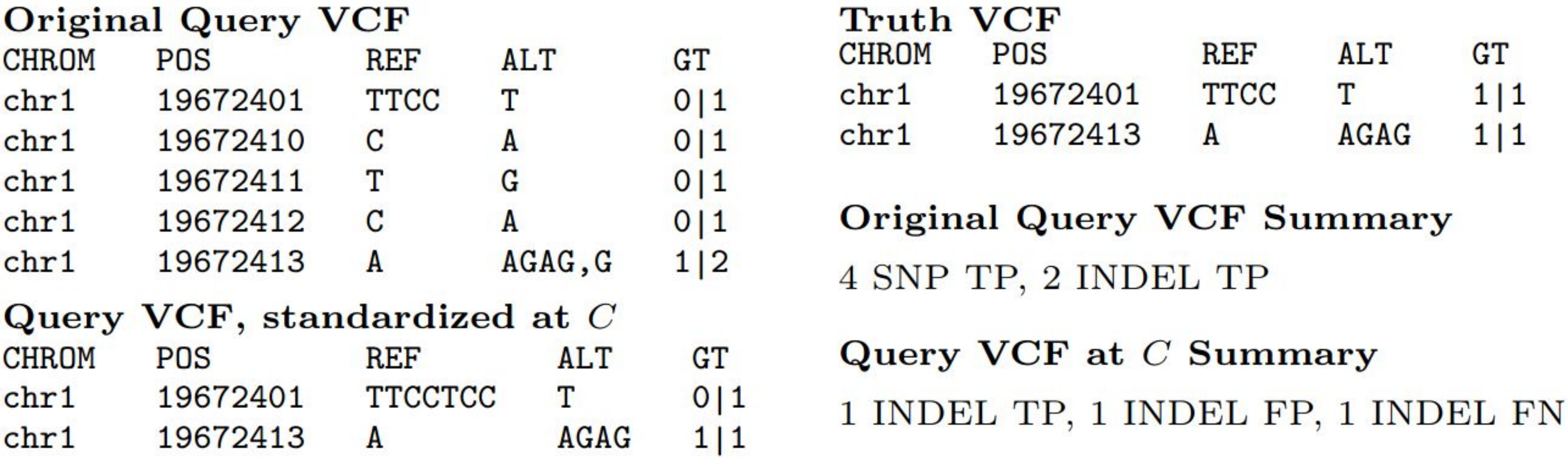
**Query VCF, standardized at $C$**

| CHROM | POS | REF | ALT | GT |
|---|---|---|---|---|
| chr1 | 19672401 | TTCCTCC | T | 0\|1 |
| chr1 | 19672413 | A | AGAG | 1\|1 |

**Truth VCF**

| CHROM | POS | REF | ALT | GT |
|---|---|---|---|---|
| chr1 | 19672401 | TTCC | T | 1\|1 |
| chr1 | 19672413 | A | AGAG | 1\|1 |

# D. Enforce local variant phasing



**Original Query VCF**

| CHROM | POS | REF | ALT | GT |
|---|---|---|---|---|
| chr1 | 19672401 | TTCC | T | 0\|1 |
| chr1 | 19672410 | C | A | 0\|1 |
| chr1 | 19672411 | T | G | 0\|1 |
| chr1 | 19672412 | C | A | 0\|1 |
| chr1 | 19672413 | A | AGAG,G | 1\|2 |

**Query VCF, standardized at *C***

| CHROM | POS | REF | ALT | GT |
|---|---|---|---|---|
| chr1 | 19672401 | TTCCTCC | T | 0\|1 |
| chr1 | 19672413 | A | AGAG | 1\|1 |

**Truth VCF**

| CHROM | POS | REF | ALT | GT |
|---|---|---|---|---|
| chr1 | 19672401 | TTCC | T | 1\|1 |
| chr1 | 19672413 | A | AGAG | 1\|1 |

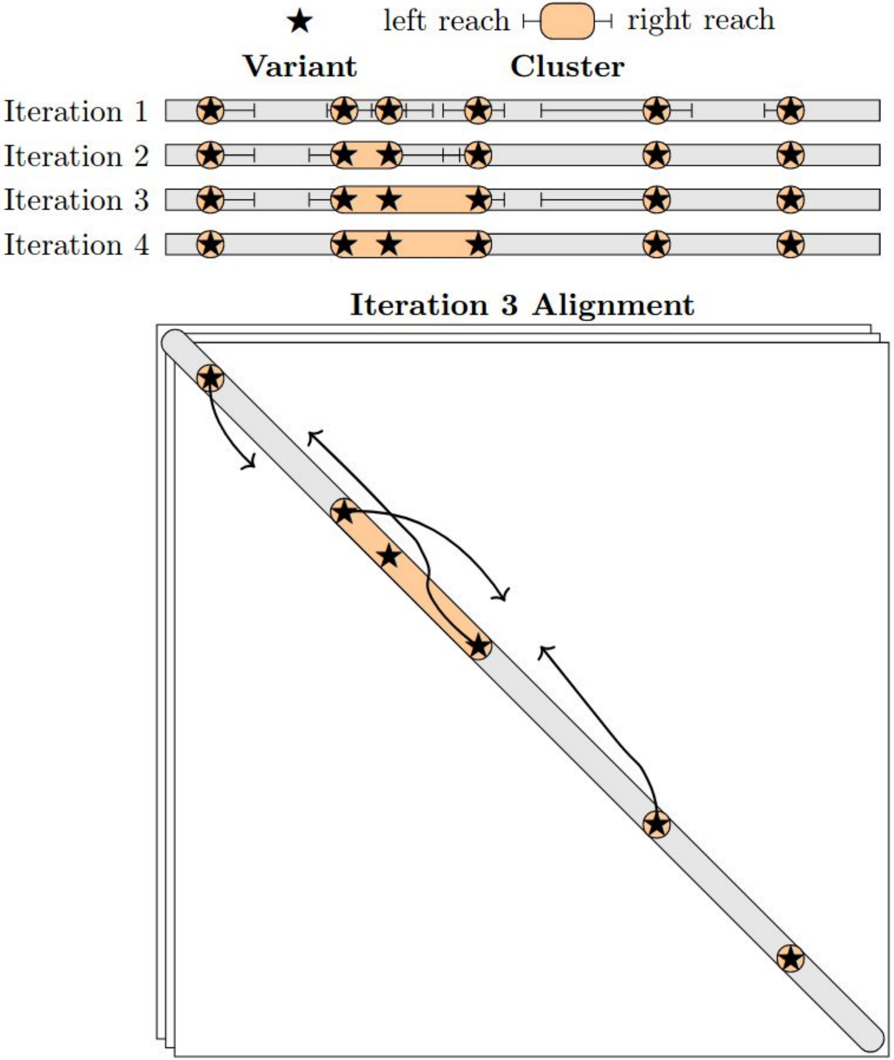
**Original Query VCF Summary**

4 SNP TP, 2 INDEL TP

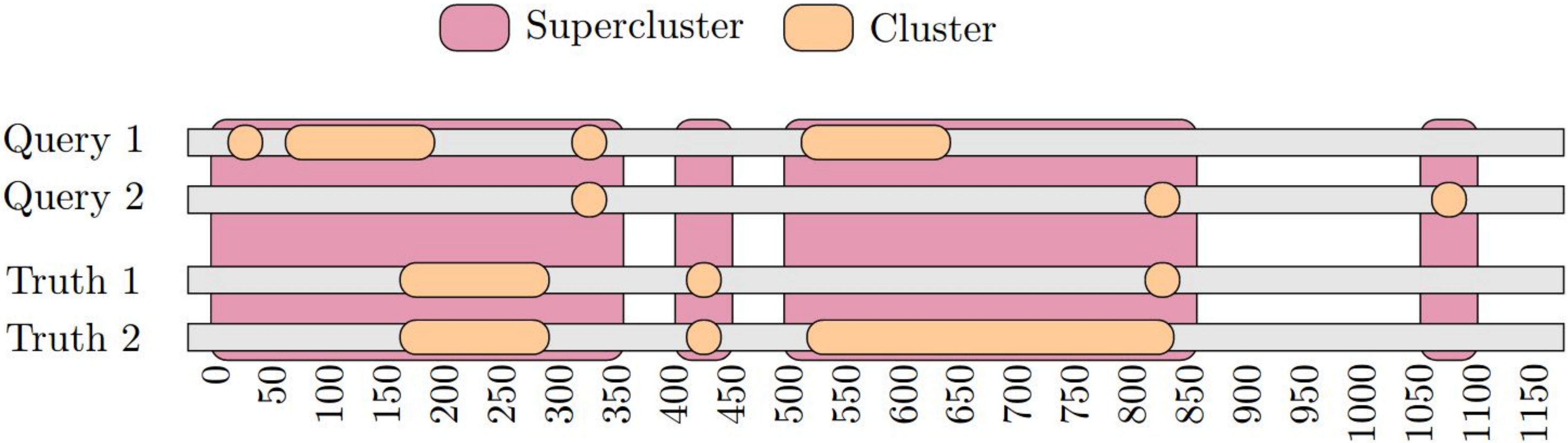
**Query VCF at *C* Summary**

1 INDEL TP, 1 INDEL FP, 1 INDEL FN

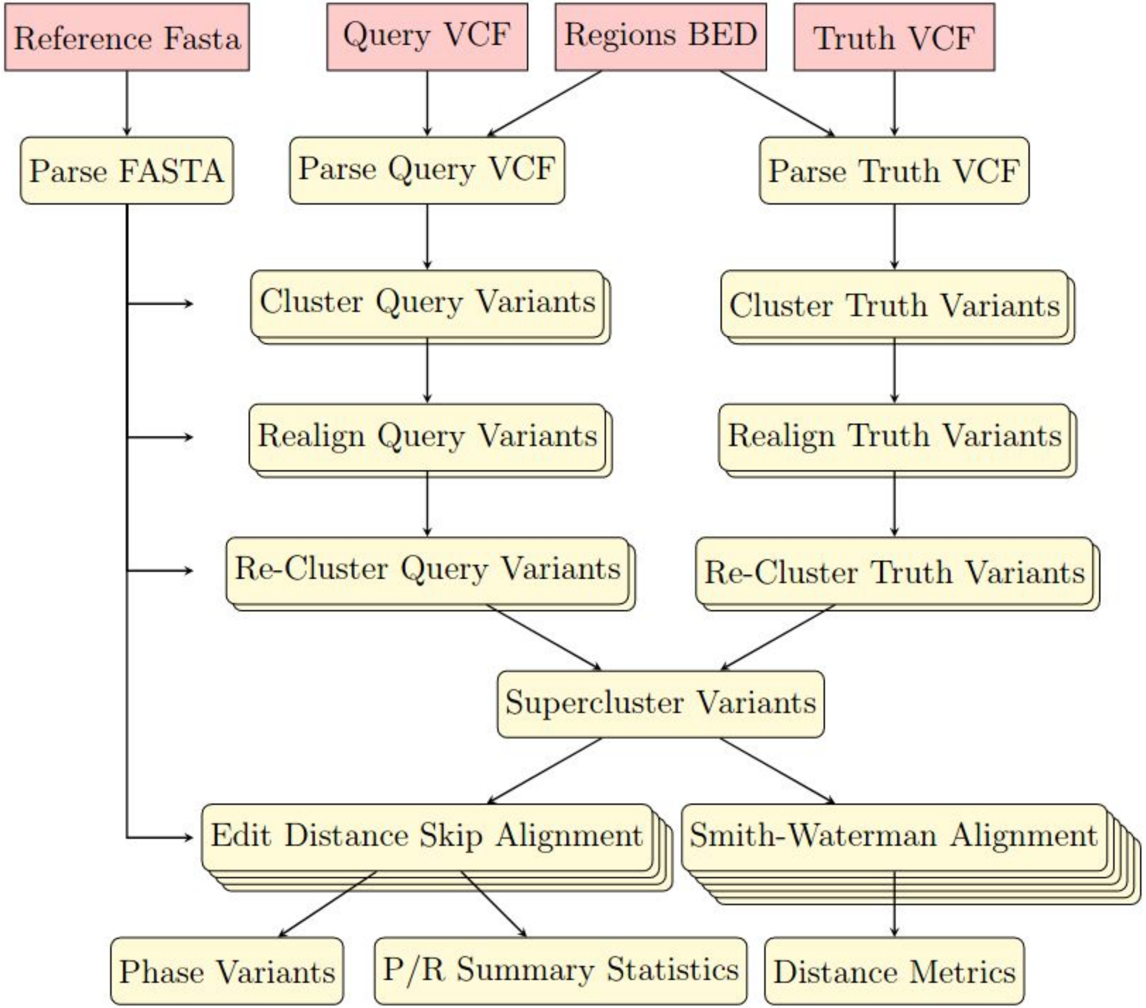
# Clustering


★ Variant
left reach
Cluster
right reach
Iteration 1
Iteration 2
Iteration 3
Iteration 4
Iteration 3 Alignment

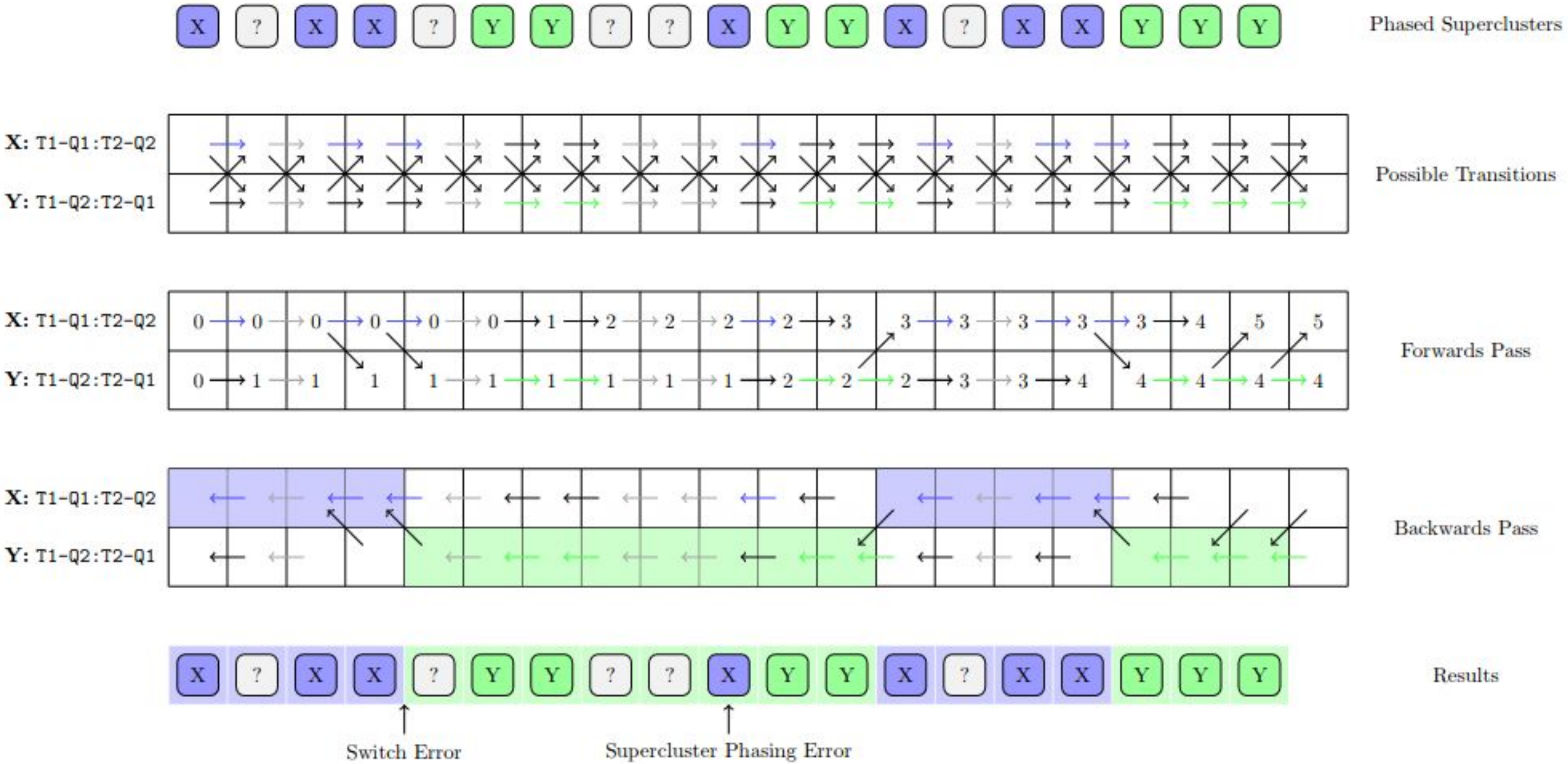
# Superclustering



# 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

# Phasing


X ? X X ? Y Y ? ? X Y Y X ? X X Y Y Y
Phased Superclusters
X: T1-Q1:T2-Q2
Y: T1-Q2:T2-Q1
Possible Transitions
0 0 0 0 0 0 1 2 2 2 2 3 3 3 3 3 3 4 5 5
0 1 1 1 1 1 1 1 1 1 2 2 2 3 3 4 4 4 4 4
Forwards Pass
Backwards Pass
X ? X X ? Y Y ? ? X Y Y X ? X X Y Y Y
Results
Switch Error
Supercluster Phasing Error

# Overview

## 1. Key Ideas

A. Standardize complex variant representation
B. Allow partial credit for variant calls
C. Distance-based evaluation metrics
D. Enforce local variant phasing

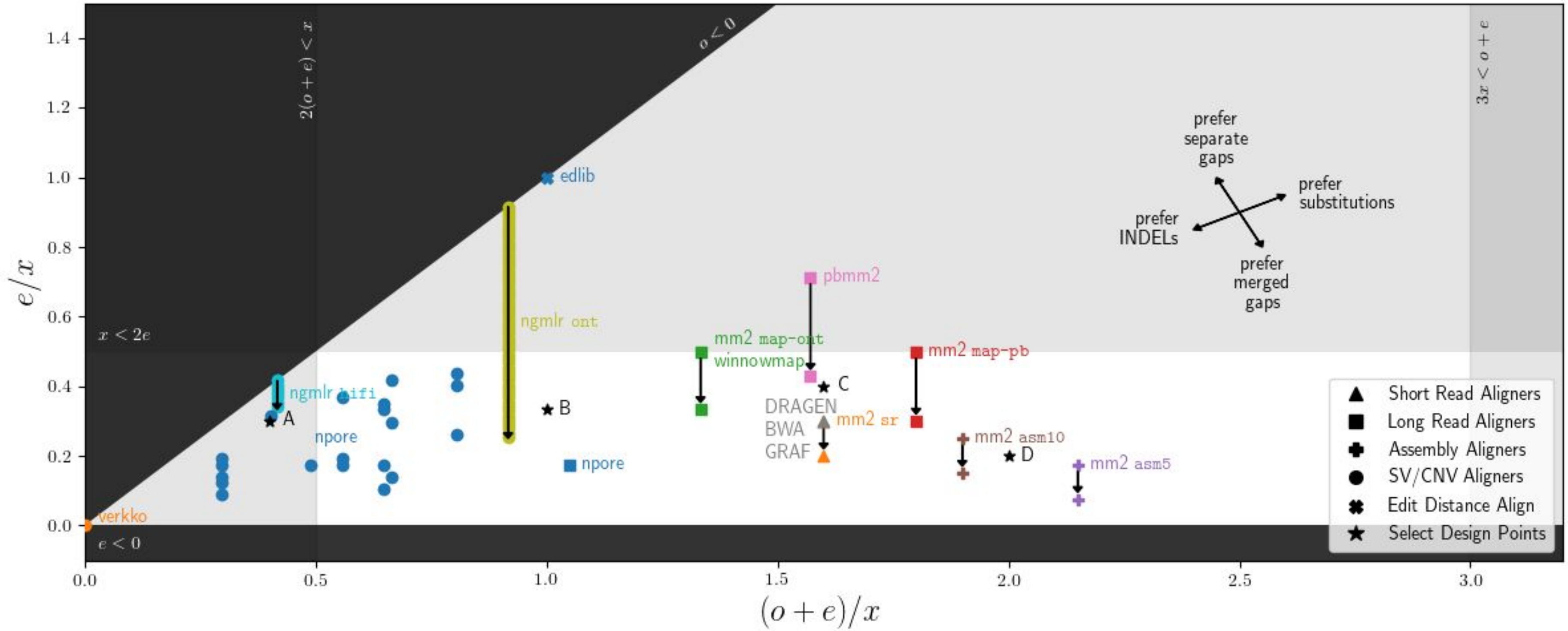
## 2. Results

Improved stability of variant calling evaluation

## 3. Extension

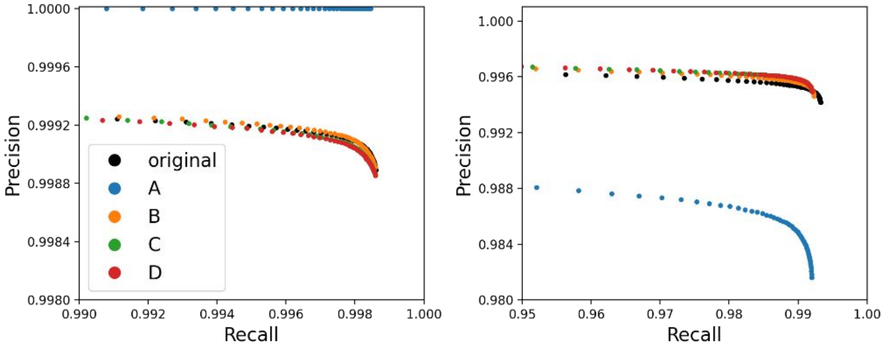
*"Can we directly evaluate structural variants?"*

# Stable performance across representations

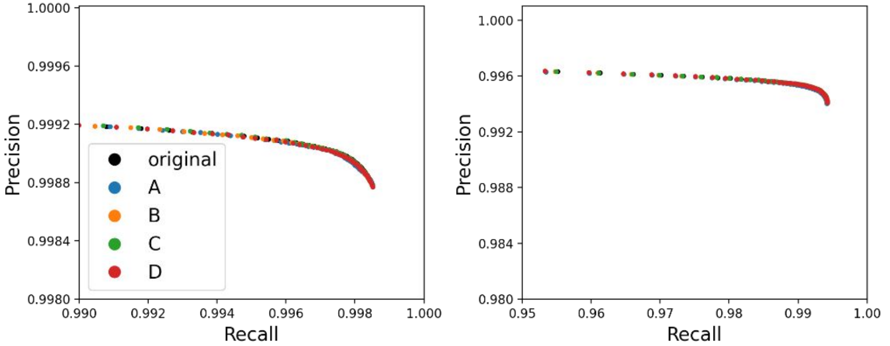


# Stable performance across representations

NIST v4.2.1 SNPs NIST v4.2.1 INDELs



vcfdist, **without** normalization or partial credit



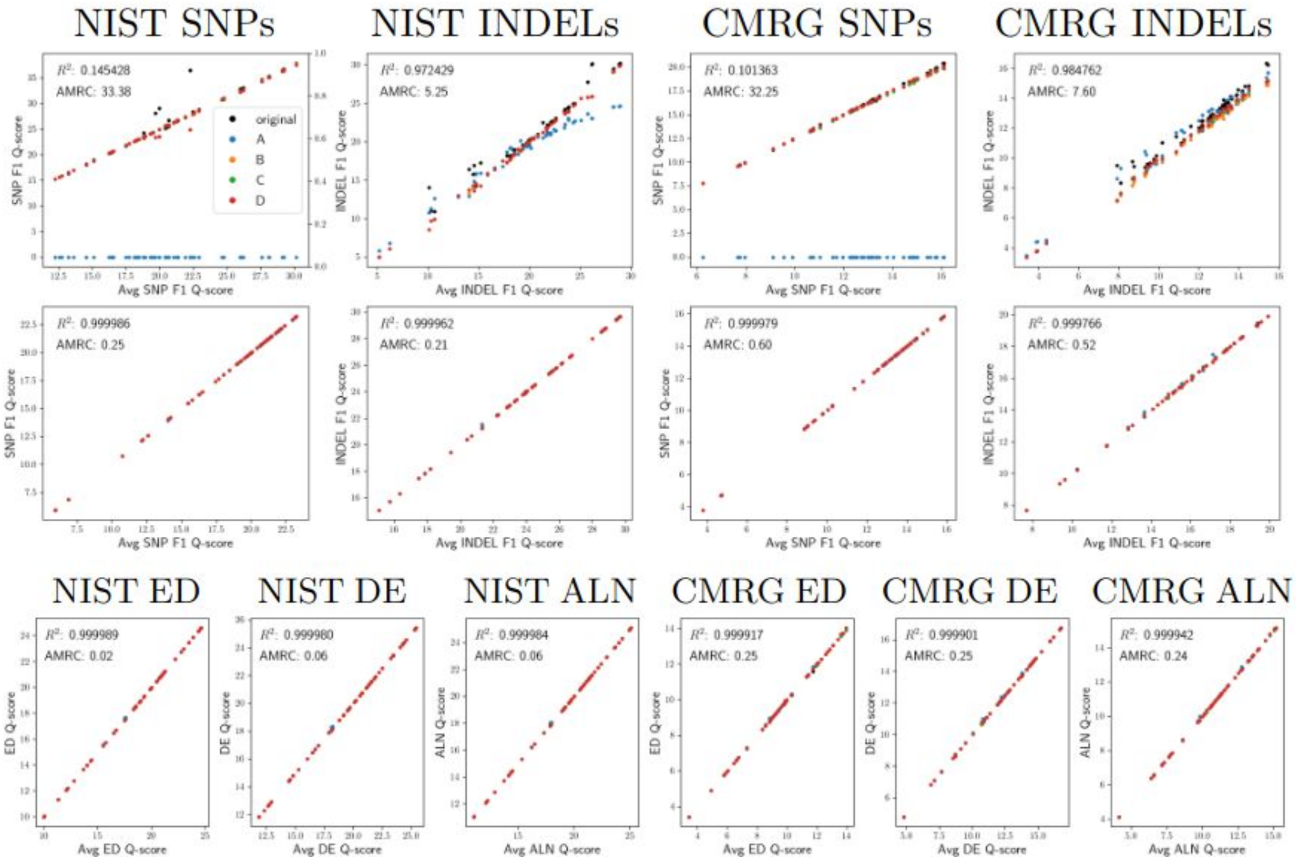
vcfdist, **with** normalization and partial credit

# Stable performance across representations

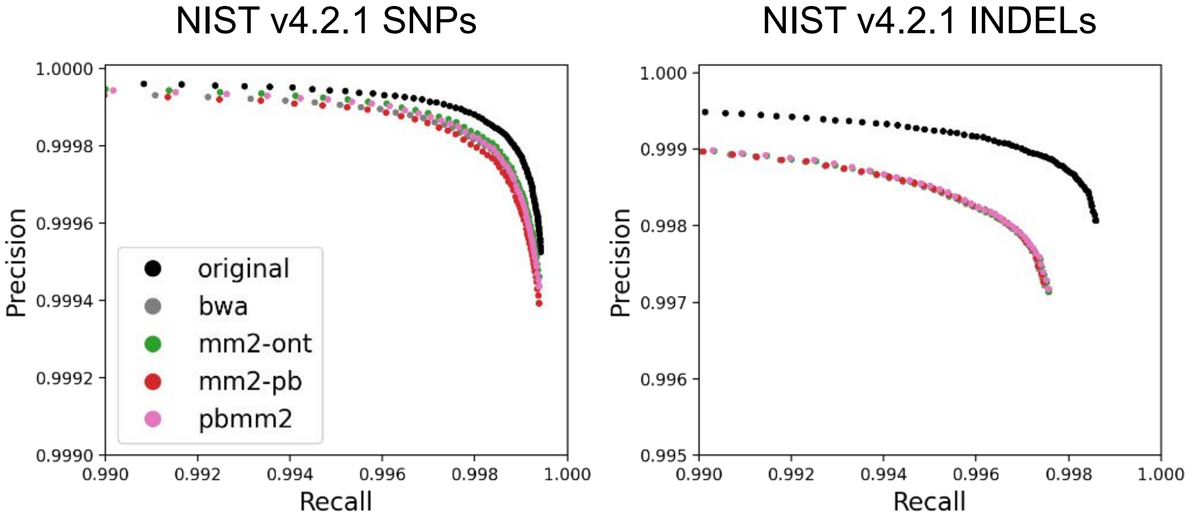
**vcfeval** precision/recall metrics

**vcfdist** precision/recall metrics

**vcfdist** distance metrics



# Bias towards fragmented variants



NIST v4.2.1 SNPs

NIST v4.2.1 INDELs

**vcfeval**

# Overview

## 1. Key Ideas

A. Standardize complex variant representation
B. Allow partial credit for variant calls
C. Distance-based evaluation metrics
D. Enforce local variant phasing

## 2. Results

Improved stability of variant calling evaluation

## 3. Extension

*"Can we directly evaluate structural variants?"*

# Motivation

A. **Single tool** to handle all genomic variation: SNPs, INDELs, SVs, TRs…

B. **Alignment-based**

- Variant representation has little/no impact
- Results don't depend on threshold heuristics

C. **Partial credit**

- Can treat SVs with same methods as SNPs and small INDELs
- Most SV calls aren't exactly correct

# Tools for variant calling evaluation

| About | | | | Variant Types | | | Phasing | | | Credit | Aln Invariance | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Tool** | **Lang** | **Release** | **Stars** | **SNPs / INDELs** | **small SVs** | **large SVs** | **none** | **local** | **global** | **near match** | **exact seq** | **near seq** |
| **vcfdist** | C++ | 2023 | 18 | ✅ | ? | ❌ | ❌ | ✅ | ✅ | ✅ | ✅ | ✅ |
| **rtg vcfeval** | Java | 2015 | 232 | ✅ | ❌ | ❌ | ✅ | ❌ | ✅ | ❌ | ✅ | ✅ |
| **xcmp hap.py** | C++ | 2019 | 343 | ✅ | ❌ | ❌ | ✅ | ❌ | ✅ | ❌ | ✅ | ✅ |
| **VarMatch** | C++ | 2016 | 9 | ✅ | ❌ | ❌ | ✅ | ❌ | ❌ | ❌ | ✅ | ✅ |
| **TruVari** | Python | 2018 | 222 | ❌ | ✅ | ✅ | ❌ | ? | ✅ | ✅ | ✅ | ❌ |
| **hap-eval** | Python | 2022 | 11 | ❌ | ✅ | ✅ | ✅ | ✅ | ✅ | ✅ | ✅ | ❌ |
| **TT-Mars** | Python | 2021 | 16 | ❌ | ✅ | ✅ | ❌ | ? | ✅ | ✅ | ❌ | ❌ |
| **SVanalyzer** | Perl | 2017 | 65 | ❌ | ✅ | ✅ | ? | ? | ✅ | ✅ | ❌ | ❌ |

# A simple example

**Query:** Verkko Assembly (Zook)

| CHROM | POS | REF | ALT | CALL | CREDIT |
|---|---|---|---|---|---|
| chr1 | 893791 | AAAAAAAAAAAATATATATATATATATATATATATATAT | A | DEL PP | 0.972222 |

**Truth:** GIAB TR Benchmark (English)

| CHROM | POS | REF | ALT | CALL | CREDIT |
|---|---|---|---|---|---|
| chr1 | 893789 | AAAAAAAAAAAAAATATATATATATATATATATATATAT | A | DEL PP | 0.972222 |

# A more complex example

**Query:** 94 base insertion

| CHROM | POS | REF | ALT | CALL | CREDIT |
|---|---|---|---|---|---|
| chr1 | 976722 | C | CAGGAACCGCCTCCCACTCCCCCCACAACCCCGGGAACCGCCTCCCACTCCCCCCGCAACCCCGGGAACCGCCTCCCACTCCCCCCGCAACCCC | INS PP | 0.979167 |
| chr1 | 976745 | G | A | SNP PP | 0.979167 |

**Truth:** Three ~31 base insertions

| 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 |

# GIAB TR equivalent representations

## Original

```
chr20    278985    A    C
chr20    278986    C    G
chr20    278990    G    C
chr20    278993    C    A
chr20    278994    G    GGGAGGGAGGGCGGGACGGAGGGA
GGGAGGGAGGGACGGAGGGCGGGACGGCGGGAGGGCGGGACG
GAGGGACGGAGGGAGGGCGGGACGGAGGGCGGGAGGGCGGGA
CGGAGGGAGGGAGGGAGGGAGGGCGGGACGGAGGGAGGGAGG
GCGGGACGGAGGGAGGGAGGGAGGGACGGAGGGCGGGACGGC
GGGAGGGCGGGACGGAGGGACGGAGGGAGGGCGGGACGGAGG
GCGGGAGGGCGGGACGGAGGGAGGGAGGGCGGGACGGAGGGA
CGGAGGGAGGGAGGGCGGGACGGAGGGAGGGAGGGCGGGACG
GAGGGACGGAGGGAGGGAGGGCGGGACGGAGGGAGGGAGGGA
GGGACGGAGGGCGGGACGGCGGGAGGGCGGGACGGAGGGACG
GAGGGAGGGCGGGACGGAGGGCGGGAGGGAGGGAGGGCGGGA
CGGAGGGAGGGAGGGAGGGAGGGCGGGACGGAGGGAGGGAGG
GAGGGAGGGACGGAGGGACGGAGGGAGGGAGGGAGGGAGGGA
CGGAGGGCGGGACGGAGGGAGGGAGGGCGGAGGGAGGGAGGG
CGGGACGGAGGGAGGGAGGGAGGGACGGAGGGCGGGACGGAG
GGAGGGAGGGC
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 (C)

```
chr20    278984    G    GCGGGACGGAGGGAGGGAGGGCG
GGACGGAGGGAGGGAGGGAGGGACGGAGGGCGGGACGGCG
GGAGGGCGGGACGGAGGGACGGAGGGAGGGCGGGACGGAG
GGCGGGAGGGCGGGACGGAGGGAGGGAGGGAGGGAGGGCG
GGACGGAGGGAGGGAGGGCGGGACGGAGGGAGGGAGGGAG
GG
chr20    279069    A    AGGGCGGGACGGAGGGACGGAGG
GAGGGAGGGCGGGACGGAGGGAGGGAGGGCGGGACGGAGG
GACGGAGGGAGGGAGGGCGGGACGGAGGGAGGGAGGGAGG
GACGGAGGGCGGGACGGCGGGAGGGCGGGACGGAGGGACG
GAGGGAGGGCGGGACGGAGGGCGGGAGGGAGGGAGGGCGG
GACGGAGGGAGGGAGGGAGGGAGGGCGGGACGGAGGGAGG
GAGGGAGGGAGGGACGGAGGGACGGAGGGAGGGAGGGAGG
GAGGGACGGAGGGCGGGACGGAGGGAGGGAGGGCGGAGGG
AGGGAGGGCGGGACGGAGGGAGGGAGGGAGGGACGGAGGG
CGGGACGGAGGGAGGGAGGGCGGAGGGAGGGAGGGCGGGA
CGGAGGGAGGGAGGGCGGGAGGGATGGAGGGAGGGAGGGC
GGGACGGAGGGAGGGC
```

2 INS (438bp, 184bp)

# Total true positive tandem repeat variants

| whole genome | Original | Normalized | Difference |
|---|---|---|---|
| TR Bench SNP TP | 980,432 | 610,522 | -37.7% |
| TR Bench INDEL TP | 519,114 | 564,916 | +8.8% |
| Verkko SNP TP | 552,776 | 564,982 | +2.2% |
| Verkko INDEL TP | 412,113 | 418,818 | +1.6% |

# Distance-based metrics vs precision and recall

| chr20:1-3,000,000 | Original | Normalized |
|---:|:---:|:---:|
| **SNP Precision** | 97.42% | 96.23% |
| **SNP Recall** | 93.88% | 98.38% |
| **F1 SNP Qscore** | 13.58 | 15.68 |
| **INDEL Precision** | 79.09% | 80.11% |
| **INDEL Recall** | 98.03% | 97.43% |
| **F1 INDEL Qscore** | 9.05 | 9.18 |
| **Edit Distance** | 750 | 750 |
| **Distinct Edits** | 34 | 34 |
| **Alignment Qscore** | 12.83 | 12.83 |

# Comparison with TruVari (<1000bp)

| whole genome | TruVari v3 | vcfdist | Difference |
|---|---|---|---|
| **TR Bench TP** | 1,187,250 | 1,499,546 | +26.3% |
| **Verkko TP** | 778,520 | 964,889 | +23.9% |

# Summary

- **There are still challenges regarding complex variant representations**
- vcfdist makes progress on these challenges, and works with SVs
- vcfdist is currently too inefficient for large SV evaluations
- Lots of room for improving vcfdist's evaluation speed
- **Need for discussion on metrics and best practices**

# Questions?

github.com/timd1/vcfdist