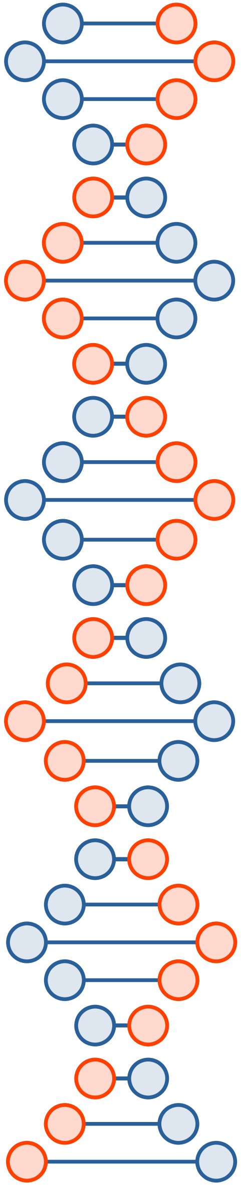
# Using readmers and hapmers in assessing phase switching after read error correction of Oxford Nanopore Sequences

**Jean P. Elbers**
jean.elbers@gmail.com
Institute of Medical Genetics, Center for Pathobiochemistry and Genetics, Medical University of Vienna

**Tamara Löwenstern**
tamara.loewenstern@meduniwien.ac.at
Institute of Medical Genetics, Center for Pathobiochemistry and Genetics, Medical University of Vienna

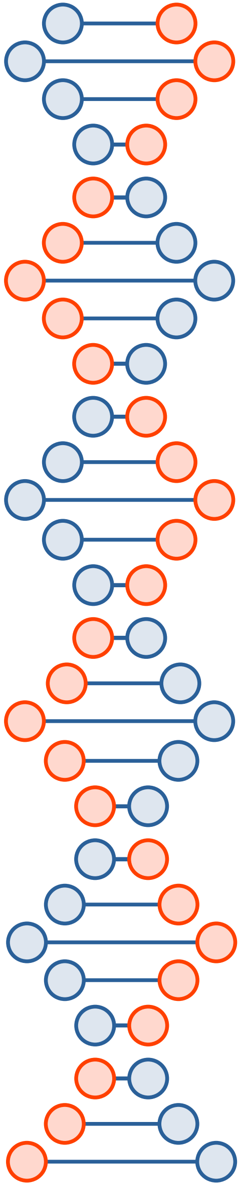
**David Horner**
david.horner@meduniwien.ac.at
Institute of Medical Genetics, Center for Pathobiochemistry and Genetics, Medical University of Vienna

**Franco Laccone**
franco.laccone@meduniwien.ac.at
Institute of Medical Genetics, Center for Pathobiochemistry and Genetics, Medical University of Vienna



# Outline

- **Introduction**
- Methods
- Results
- Wrap up



# Outline

- **Introduction**

k-mers

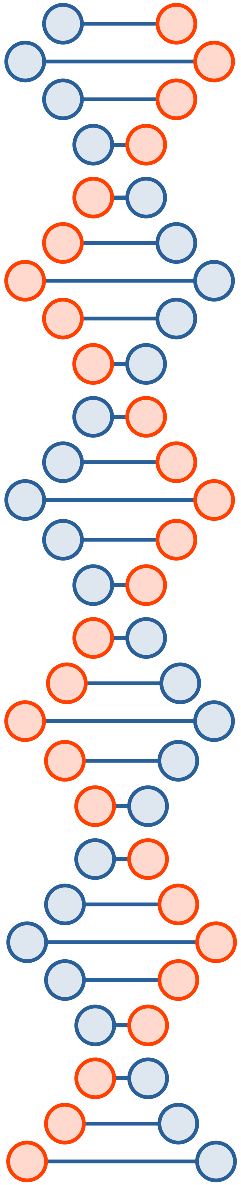
readmers

hapmers

phase switching

Error correction

Why do we care? – Research Question

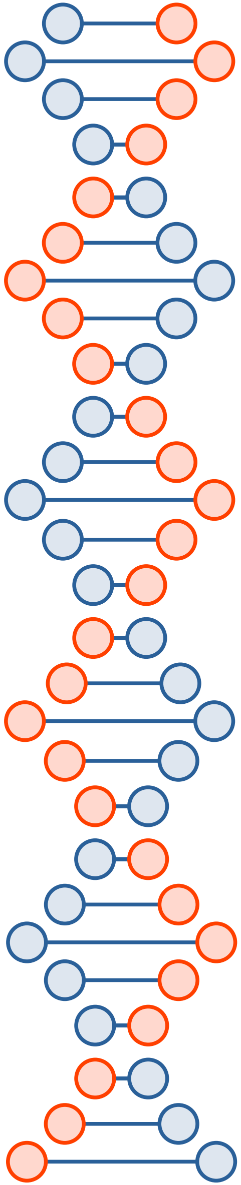


# K-mers, Readmers, & Hapmers, oh my….

- DNA has no “words”
  - Can break DNA into pieces of “k-length” called k-mers

```
k-mer length = 4
```

ACTGCA

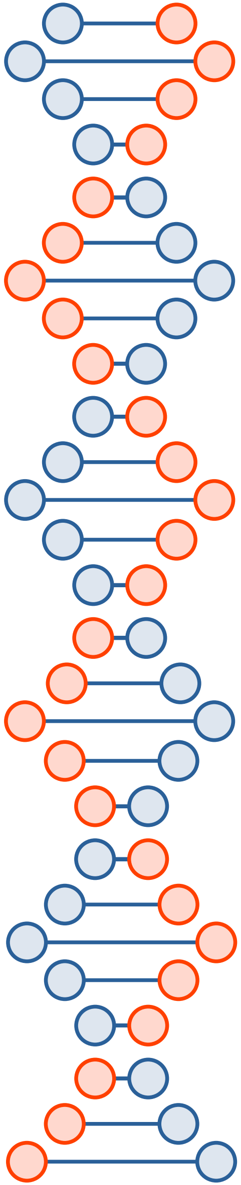


# K-mers, Readmers, & Hapmers, oh my….

- DNA has no “words”
  - ▹ Can break DNA into pieces of “k-length” called k-mers

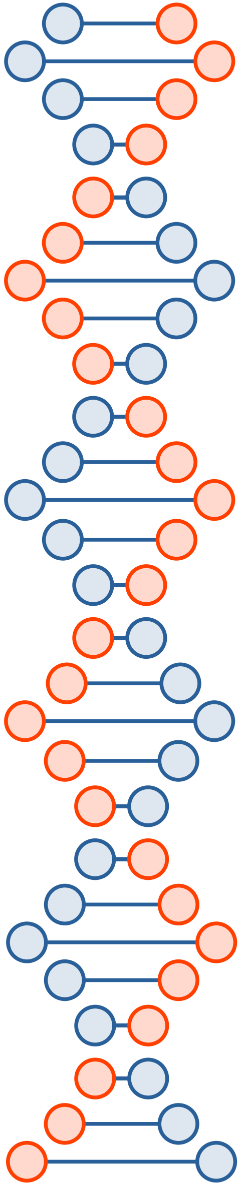
```
k-mer length = 4
```

```
ACTGCA
actg      4-mer #1
 ctgc     4-mer #2
  tgca    4-mer #3
```



# K-mers, Readmers, & Hapmers, oh my….

- K-mers from sequencing reads ---
    - “Readmers”



# K-mers, Readmers, & Hapmers, oh my….

- K-mers from sequencing reads ---
  - “Readmers”

    *A FASTQ read:*

```
@seq1
ACTGCATAGC
+
6655EEACDC
```

# K-mers, Readmers, & Hapmers, oh my….

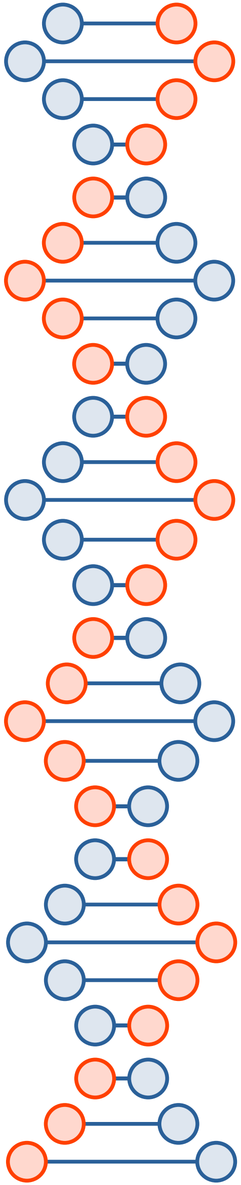
- K-mers unique to a haplotype ---
  - “Hapmers”

using 8-mers on “+” strand

# K-mers, Readmers, & Hapmers, oh my….

- K-mers unique to a haplotype ---
  - “Hapmers”

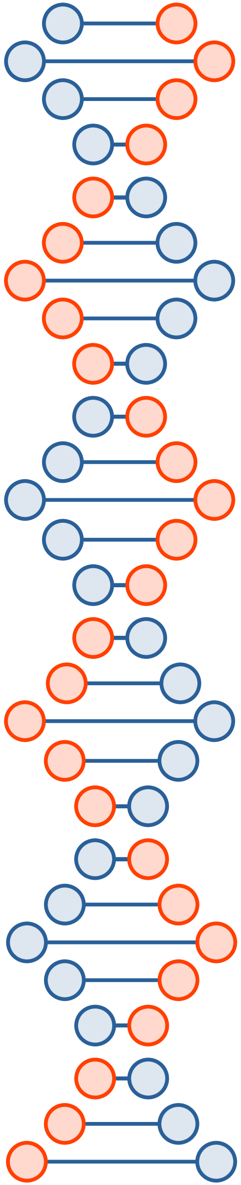
```
using 8-mers on “+” strand

ATTGCATA  + strand maternal haplotype


ACTGAATA  + strand paternal haplotype
```



# K-mers, Readmers, & Hapmers, oh my….

- K-mers unique to a haplotype ---
  - “Hapmers”

```
using 8-mers on “+” strand

ATTGCATA  “+” strand maternal haplotype
attgcata

ACTGAATA  “+” strand paternal haplotype
actgaata
```

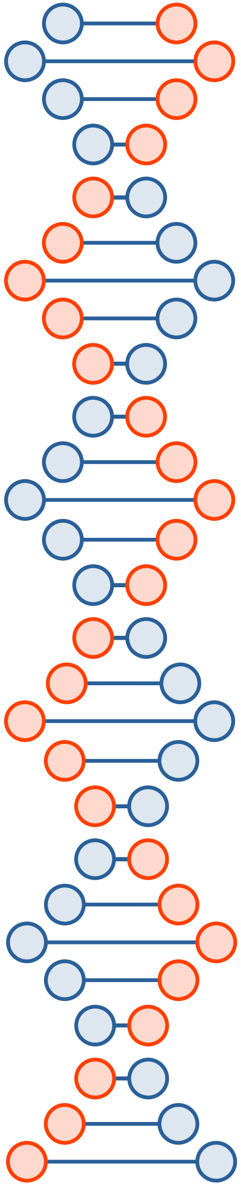


# K-mers, Readmers, & Hapmers, oh my….

- K-mers unique to a haplotype ---
  - “Hapmers”

```
using 8-mers on “+” strand

ATTGCATA  “+” strand maternal haplotype
attgcata

ACTGAATA  “+” strand paternal haplotype
actgaata
```

Hapmers

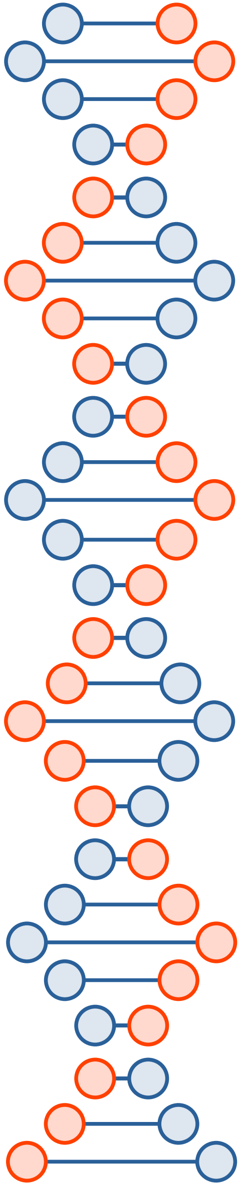


# Phase Switching

- At the read level, read contains info of 1> haplotype

ATTGCATA “+” strand maternal haplotype

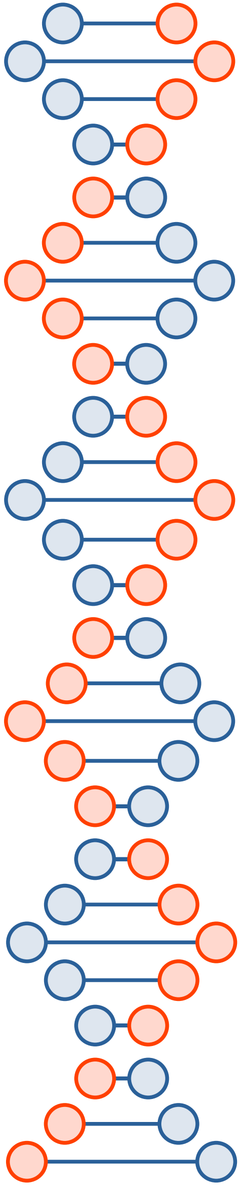
ACTGAATA “+” strand paternal haplotype



# Phase Switching

- At the read level, read contains info of 1> haplotype

```
ATTGCATA  “+” strand maternal haplotype

ACTGAATA  “+” strand paternal haplotype


@seq1
ACTGCATA
+
CCEEDDAD
```



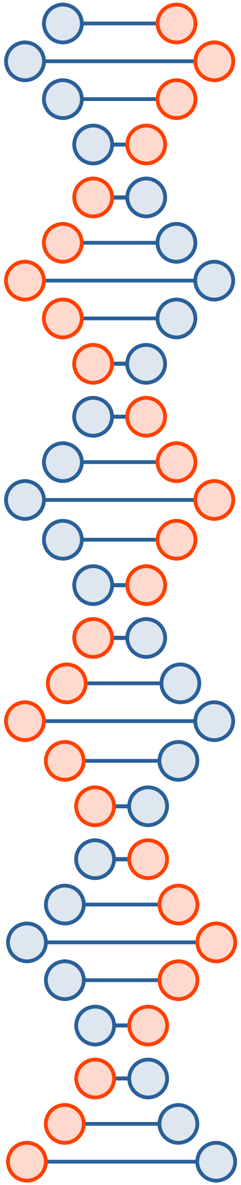
# Phase Switching

- At the read level, read contains info of 1> haplotype

```
ATTGCATA  “+” strand maternal haplotype

ACTGAATA  “+” strand paternal haplotype
```

```
@seq1
ACTGCATA
+
CCEEDDAD
```

Phase Switching!



# Phase Switching

- At the read level, read contains info of 1> haplotype

```
ATTGCATA  “+” strand maternal haplotype

ACTGAATA  “+” strand paternal haplotype

@seq1
ACTGCATA
+
CCEEDDAD
```

In this example:

If counting 8-mers, then there will be

**No matching readmers to hapmers**

# Error correction

- Improve accuracy by “consensus”

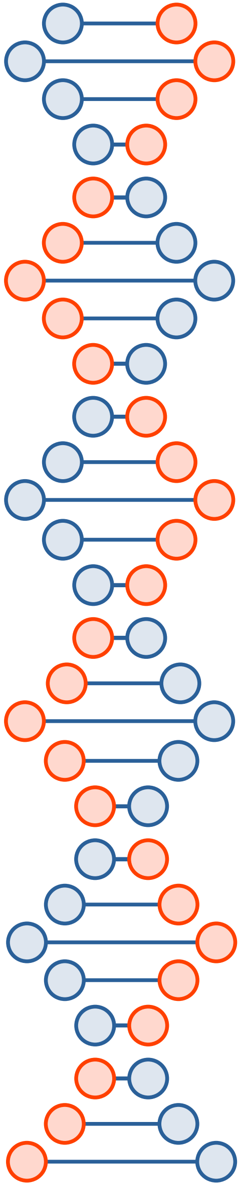
    1. Herro

# Error correction

- Improve accuracy by “consensus”

1. Herro

2. Brutal Rewrite



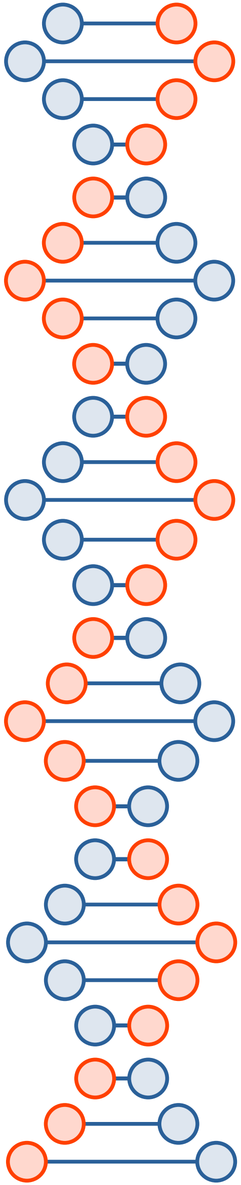
# Error correction

- Improve accuracy by “consensus”

1. Herro

2. Brutal Rewrite

3. Peregrine_2021



# Error correction

- Improve accuracy by “consensus”

1. Herro

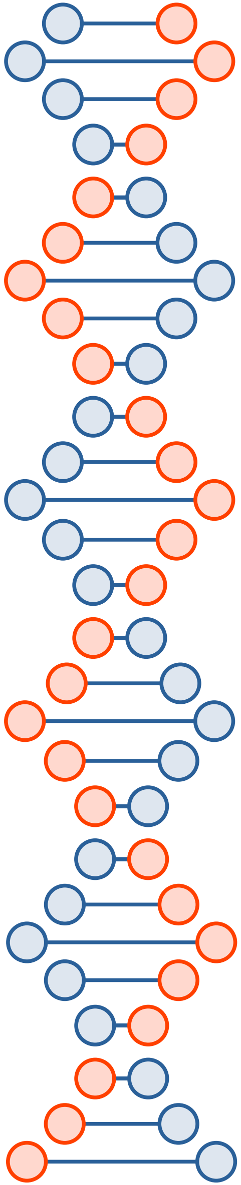
2. Brutal Rewrite

3. Peregrine_2021

4. DeChat

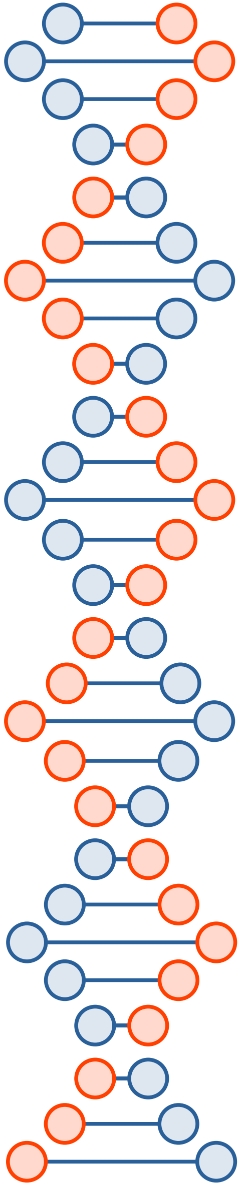
# Question

- Since we do not know what sequences belong to which haplotype *a priori* , can error correcting reads cause phase switching?



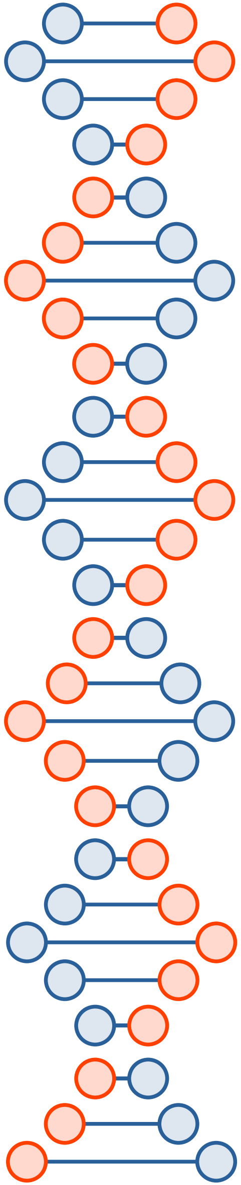
# Question

- Since we do not know what sequences belong to which haplotype *a priori* , can error correcting reads cause phase switching?
- How can we estimate phase switching at the read level?



# Outline

- Introduction
- **Methods**
- Results
- Wrap up



# Outline

- **Methods**

  error correction

  read alignment

  analyze alignments

  k-mers

# Methods

- Error correct reads

# Methods

- Error correct reads

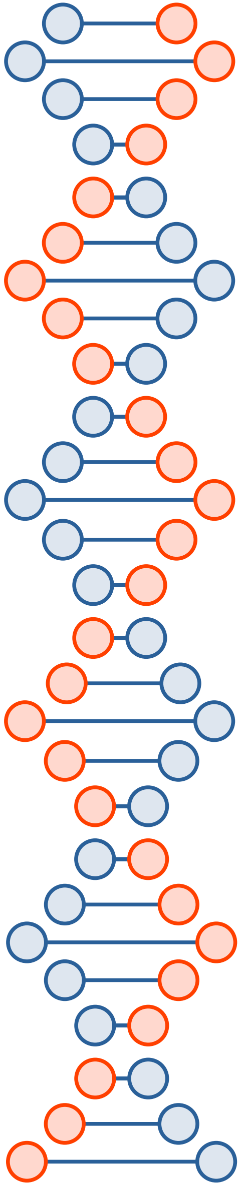
  1. Run Herro

# Methods

- Error correct reads

    1. Run Herro

    2. Run Brutal Rewrite on Herro-corrected reads



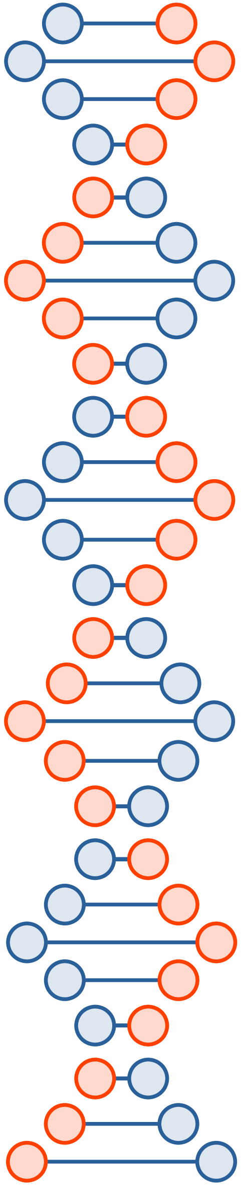
# Methods

- Error correct reads

1. Run Herro

2. Run Brutal Rewrite on Herro-corrected reads

3. Run Peregrine_2021 on Brutal Rewrite-corrected reads



# Methods

- Error correct reads

  1. Run Herro

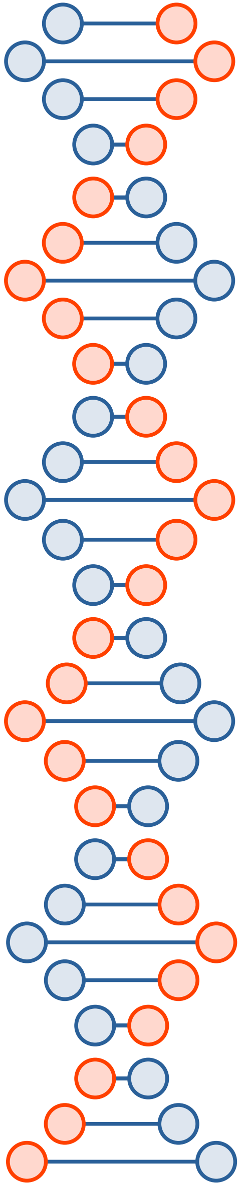
  2. Run Brutal Rewrite on Herro-corrected reads

  3. Run Peregrine_2021 on Brutal Rewrite-corrected reads

  4. Run DeChat on Peregrine_2021-corrected reads

# Methods

- Error correct reads
- Align raw & corrected reads to each HG002 haplotype separately

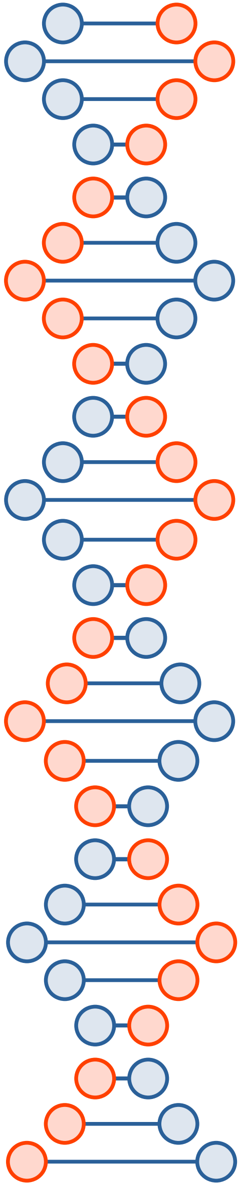


# Methods

- Error correct reads
- Align raw & corrected reads to each HG002 haplotype separately
- Collect alignment info

# Methods

- Error correct reads
- Align raw & corrected reads to each HG002 haplotype separately
- Collect alignment info
- Filter alignments (primary alignments,
  mapping quality Q60,
  longest alignment block per read)



# Methods

- Error correct reads
- Align raw & corrected reads to each HG002 haplotype separately
- Collect alignment info
- Filter alignments (primary alignments,

  mapping quality Q60,

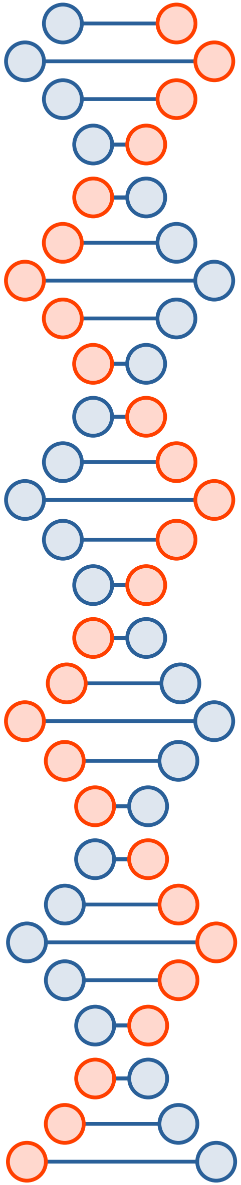
  longest alignment block per read)
- Analyze k-mers for same reads (~ 1 million)

# Methods

- Error correct reads
- Align raw & corrected reads to each HG002 haplotype separately
- Collect alignment info
- Filter alignments (primary alignments,  
  mapping quality Q60,  
  longest alignment block per read)
- Analyze k-mers for same reads (~ 1 million)
- Histograms (percent matching readers to hapmers)

# K-mers and hapmers

- Keep alignments with same alignment block length on mat. and pat. chrom. (query_start/stop same)



# K-mers and hapmers

- Keep alignments with same alignment block length on mat. and pat. chrom. (query_start/stop same)

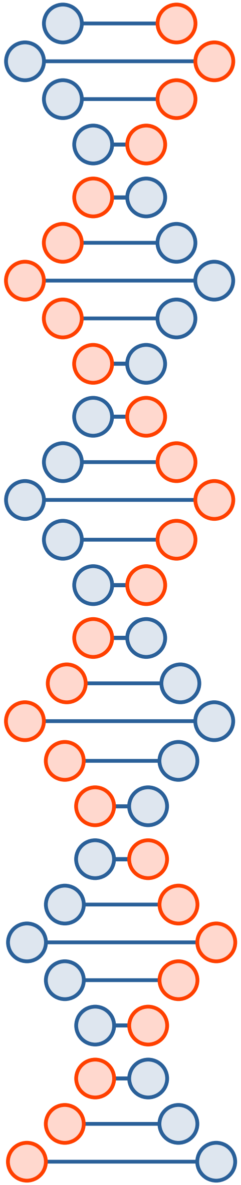
```
Alignments to mat. chrom.
query_start query_stop
1                 1000

Alignments to pat. chrom.
query_start query_stop
1                 1000
```

# K-mers and hapmers

- Count readmers from query_start to query_stop

```
Alignments to mat. chrom.
query_start query_stop
1                 1000
```

```
Alignments to pat. chrom.
query_start query_stop
1                 1000
```



# K-mers and hapmers

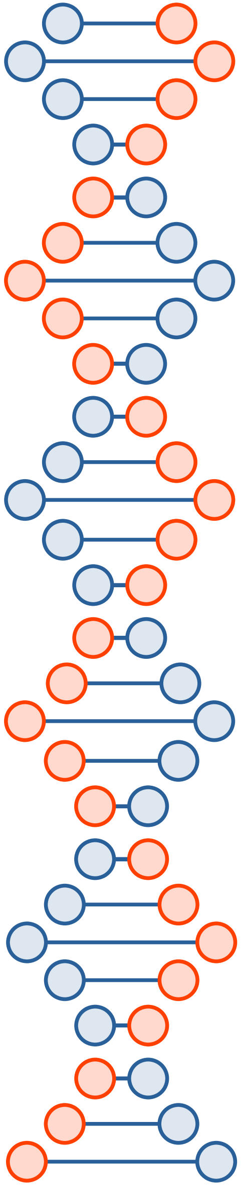
- Count k-mers from each haplotype from each target_start to target_stop

Alignments to mat. chrom.

| target_start | target_stop |
|---|---|
| 10000 | 11000 |

Alignments to pat. chrom.

| target_start | target_stop |
|---|---|
| 20000 | 21001 |

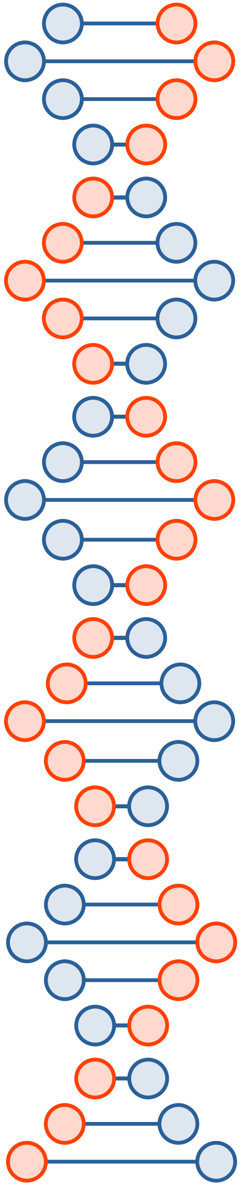


# K-mers and hapmers

- Count readmers from query_start to query_stop
- Count k-mers from each haplotype from each target_start to target_stop
- Determine hapmers

```
ATTGCATA  “+” strand maternal haplotype
attgcata

ACTGAATA  “+” strand paternal haplotype
actgaata
```

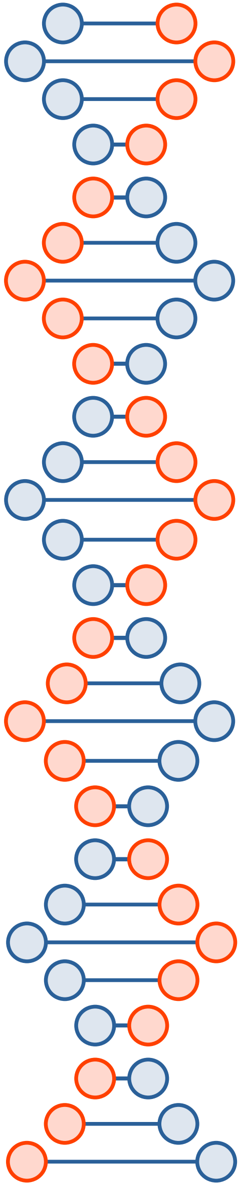
Hapmers



# K-mers and hapmers

- Intersect readmers and hapmers

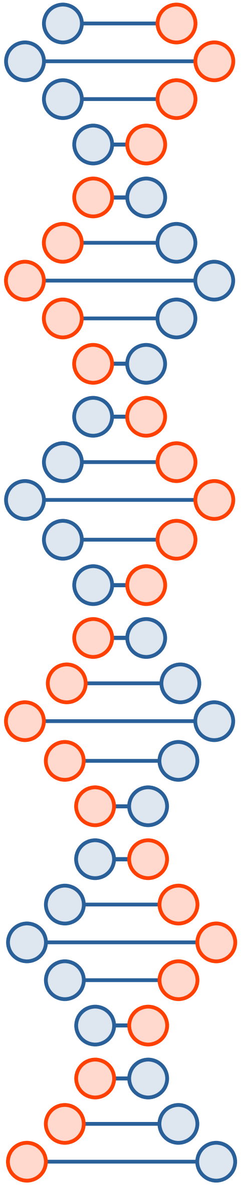
| read_id | %match_mat_hapmers | %match_pat_hapmers |
|---|---|---|
| seq1 | 75 | 25 |
| seq2 | 25 | 75 |
| seq3 | 0 | 0 |
| seq4 | 100 | 0 |



# K-mers and hapmers

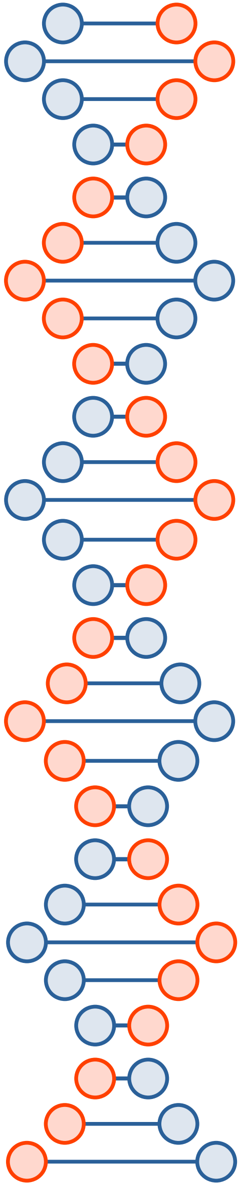
- Intersect readmers and hapmers

<table>
<tr><th>read_id</th><th>%mat</th><th>%pat</th><th>calc Abs. Val of diff</th></tr>
<tr><td>seq1</td><td>75</td><td>25</td><td>|75-25 | = 50%</td></tr>
<tr><td>seq2</td><td>25</td><td>75</td><td>|25-75 | = 50%</td></tr>
<tr><td>seq3</td><td>0</td><td>0</td><td>Discard</td></tr>
<tr><td>seq4</td><td>100</td><td>0</td><td>|100-0| = 100%</td></tr>
</table>



# Outline

- Introduction
- Methods
- **Results**
- Wrap up



# Outline

- **Results**

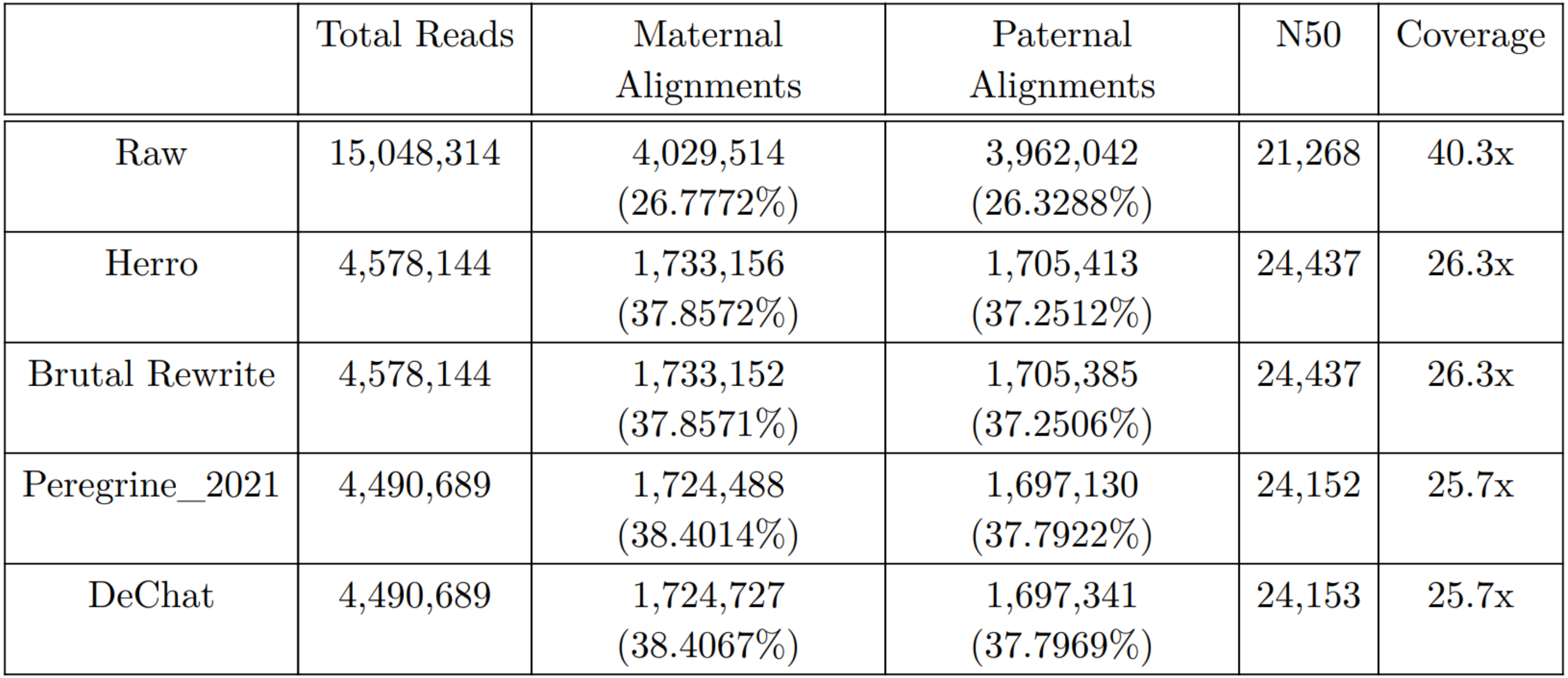
  read alignments

  matching hapmers

  histograms

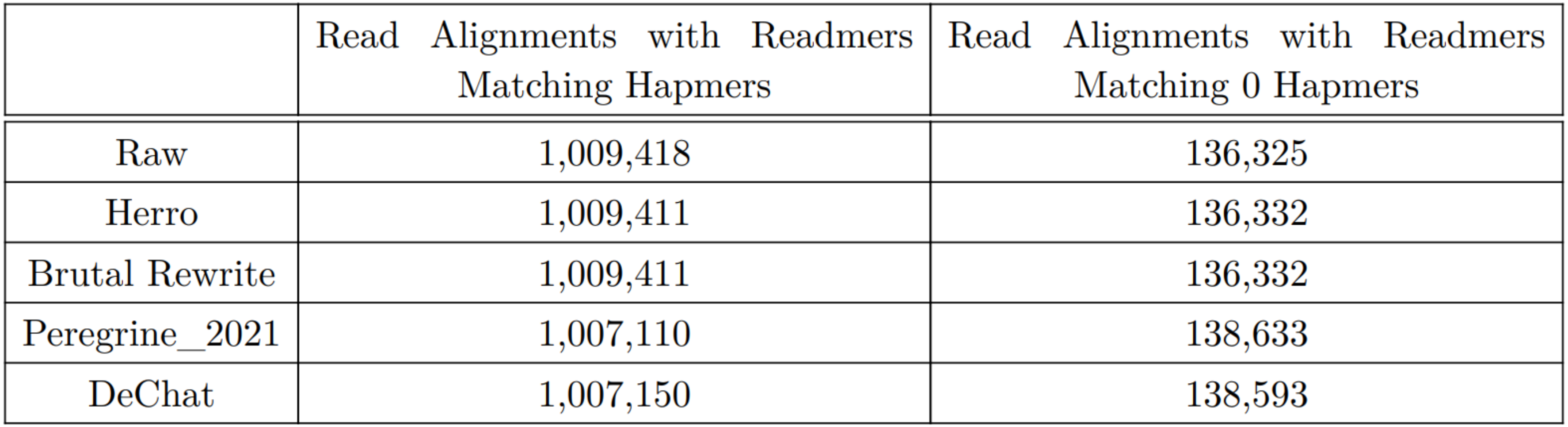
  read error rates

# Results

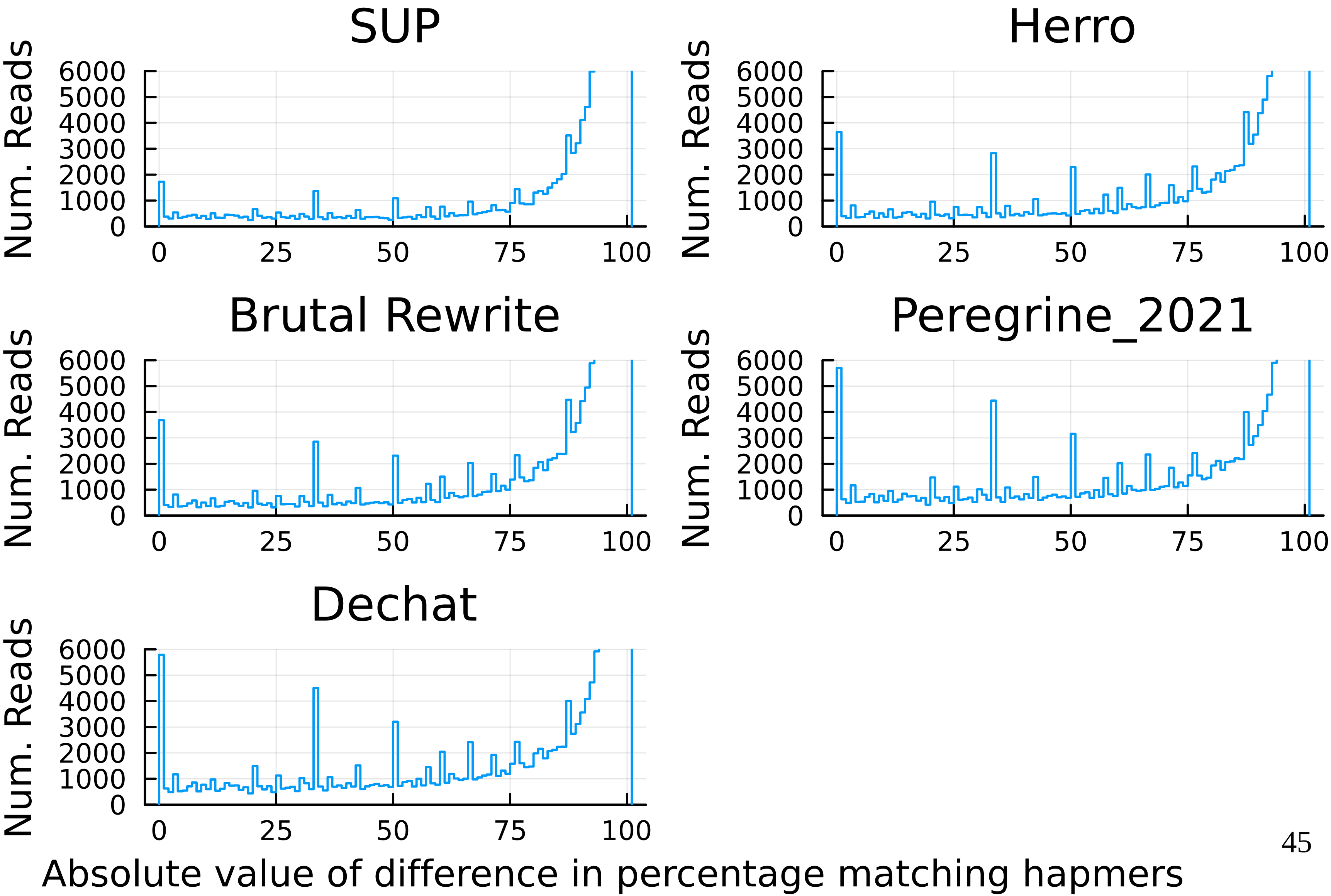


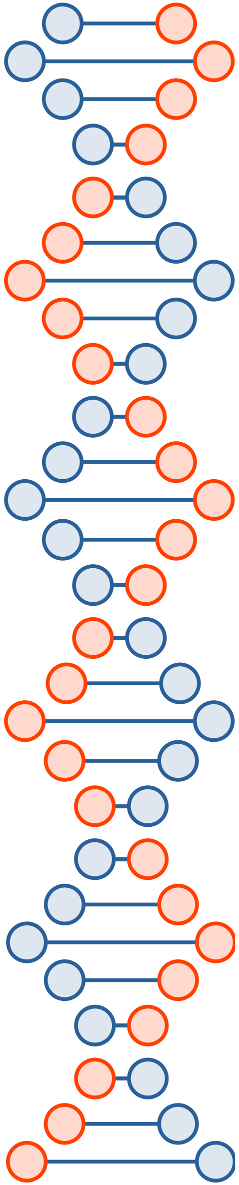
| | Total Reads | Maternal Alignments | Paternal Alignments | N50 | Coverage |
|---|---|---|---|---|---|
| Raw | 15,048,314 | 4,029,514 (26.7772%) | 3,962,042 (26.3288%) | 21,268 | 40.3x |
| Herro | 4,578,144 | 1,733,156 (37.8572%) | 1,705,413 (37.2512%) | 24,437 | 26.3x |
| Brutal Rewrite | 4,578,144 | 1,733,152 (37.8571%) | 1,705,385 (37.2506%) | 24,437 | 26.3x |
| Peregrine_2021 | 4,490,689 | 1,724,488 (38.4014%) | 1,697,130 (37.7922%) | 24,152 | 25.7x |
| DeChat | 4,490,689 | 1,724,727 (38.4067%) | 1,697,341 (37.7969%) | 24,153 | 25.7x |

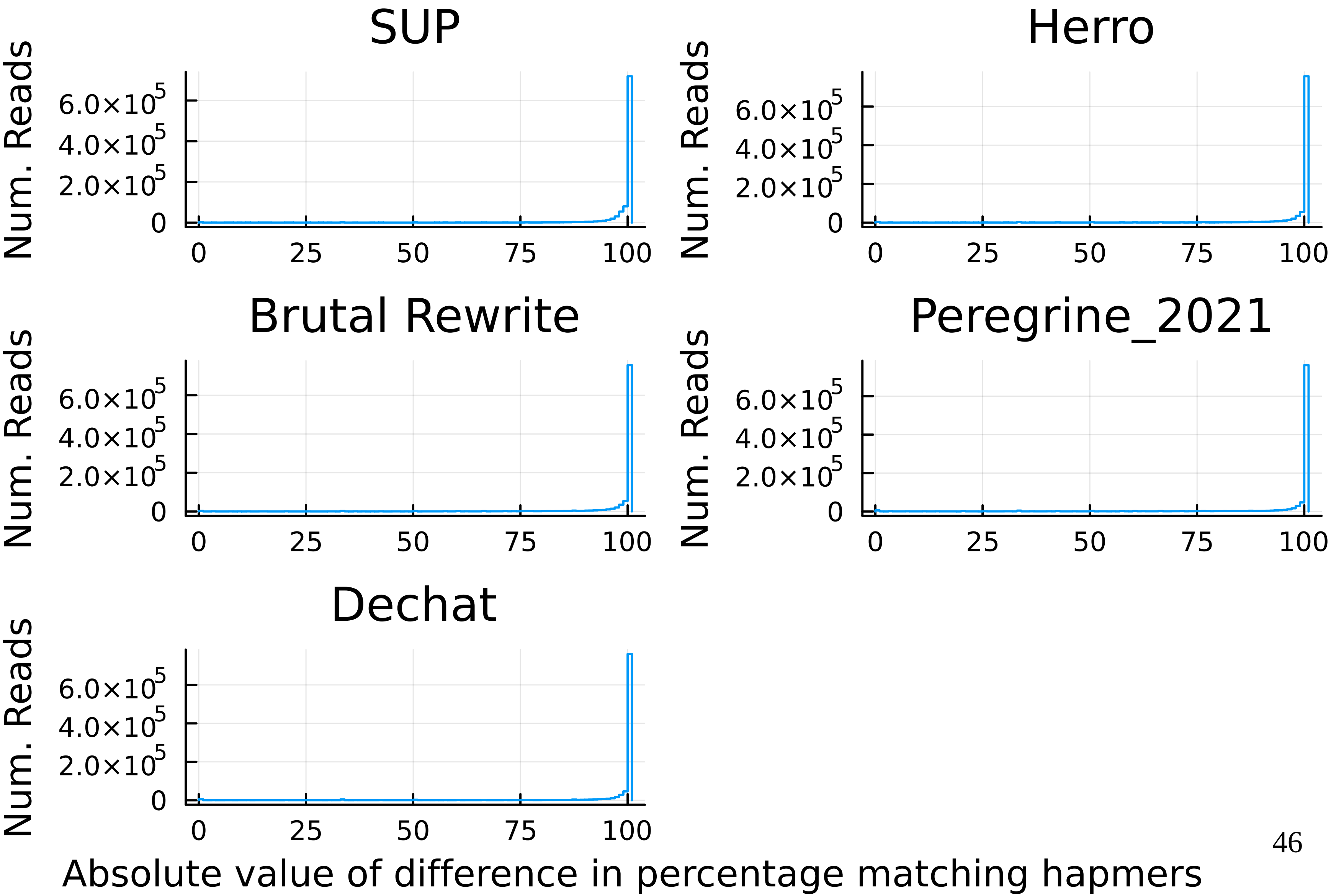
# Results



| | Read Alignments with Readmers Matching Hapmers | Read Alignments with Readmers Matching 0 Hapmers |
|---|---|---|
| Raw | 1,009,418 | 136,325 |
| Herro | 1,009,411 | 136,332 |
| Brutal Rewrite | 1,009,411 | 136,332 |
| Peregrine_2021 | 1,007,110 | 138,633 |
| DeChat | 1,007,150 | 138,593 |


SUP
Herro
Brutal Rewrite
Peregrine_2021
Dechat
Num. Reads
0
1000
2000
3000
4000
5000
6000
0
25
50
75
100
Absolute value of difference in percentage matching hapmers





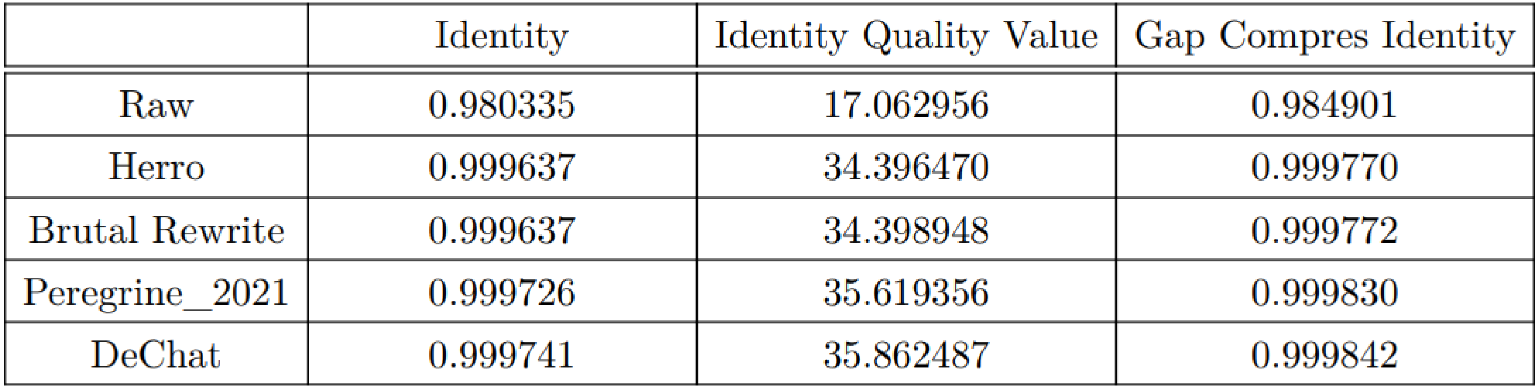
SUP

Herro

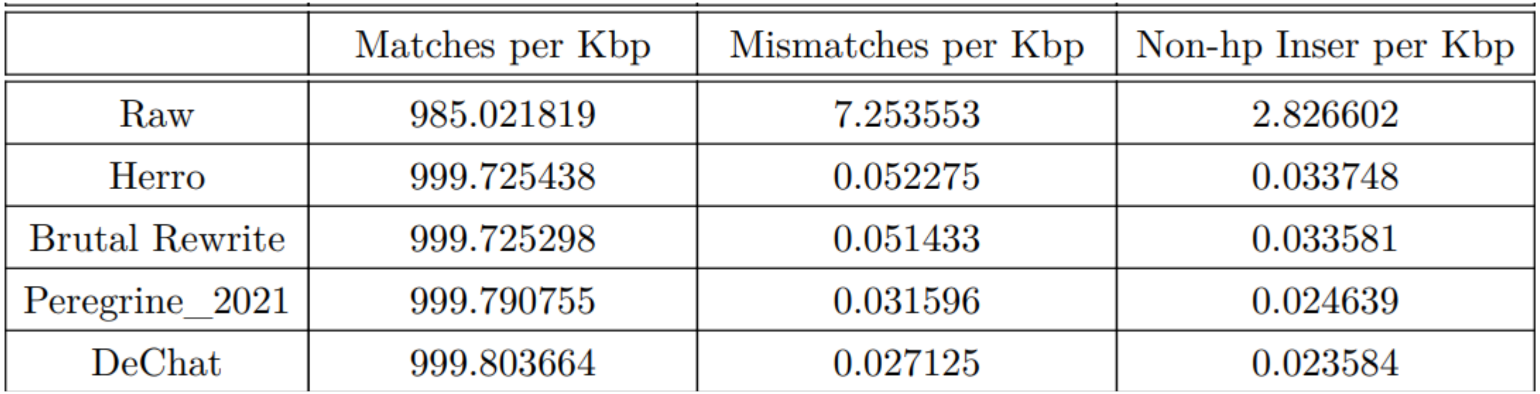
Brutal Rewrite

Peregrine_2021

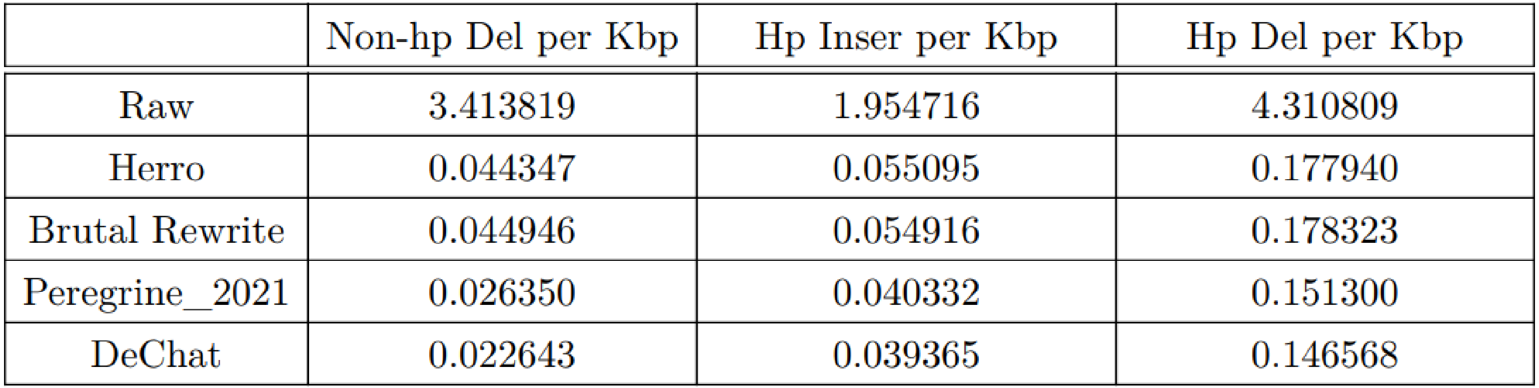
Dechat



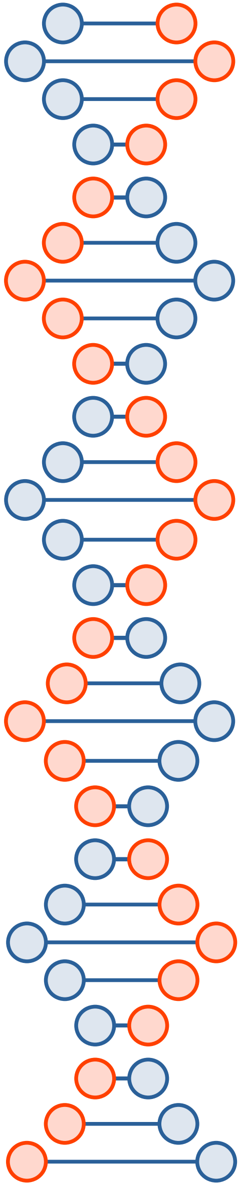
| | Identity | Identity Quality Value | Gap Compres Identity |
|---|---|---|---|
| Raw | 0.980335 | 17.062956 | 0.984901 |
| Herro | 0.999637 | 34.396470 | 0.999770 |
| Brutal Rewrite | 0.999637 | 34.398948 | 0.999772 |
| Peregrine_2021 | 0.999726 | 35.619356 | 0.999830 |
| DeChat | 0.999741 | 35.862487 | 0.999842 |
| Illumina | 0.994539 | 22.627025 | 0.994579 |



| | Matches per Kbp | Mismatches per Kbp | Non-hp Inser per Kbp |
|---|---|---|---|
| Raw | 985.021819 | 7.253553 | 2.826602 |
| Herro | 999.725438 | 0.052275 | 0.033748 |
| Brutal Rewrite | 999.725298 | 0.051433 | 0.033581 |
| Peregrine_2021 | 999.790755 | 0.031596 | 0.024639 |
| DeChat | 999.803664 | 0.027125 | 0.023584 |
| Illumina | 994.562660 | 5.376677 | 0.01667 |



|  | Non-hp Del per Kbp | Hp Inser per Kbp | Hp Del per Kbp |
|---|---|---|---|
| Raw | 3.413819 | 1.954716 | 4.310809 |
| Herro | 0.044347 | 0.055095 | 0.177940 |
| Brutal Rewrite | 0.044946 | 0.054916 | 0.178323 |
| Peregrine_2021 | 0.026350 | 0.040332 | 0.151300 |
| DeChat | 0.022643 | 0.039365 | 0.146568 |
| Illumina | 0.029777 | 0.007443 | 0.030885 |



# Outline

- Introduction
- Methods
- Results
- **Wrap up**

# Wrap up

- Can assess phase switching at read level

# Wrap up

- Can assess phase switching at read level

- Error-correction methods here introduce very little phase switching overall

# Wrap up

- Can assess phase switching at read level

- Error-correction methods here introduce very little phase switching overall

  Peregrine-2021 may introduce more phase switching than others tested

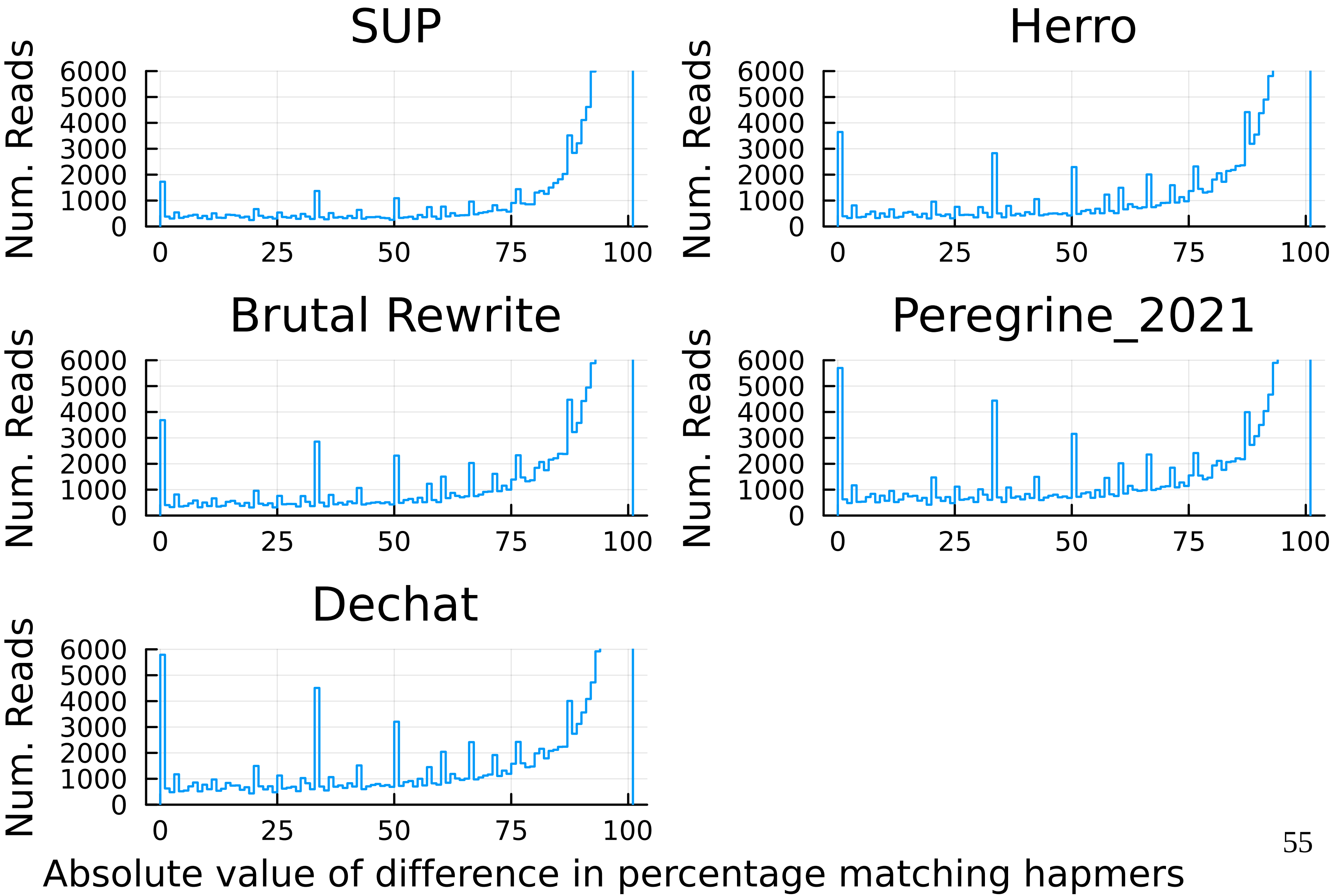
# Wrap up

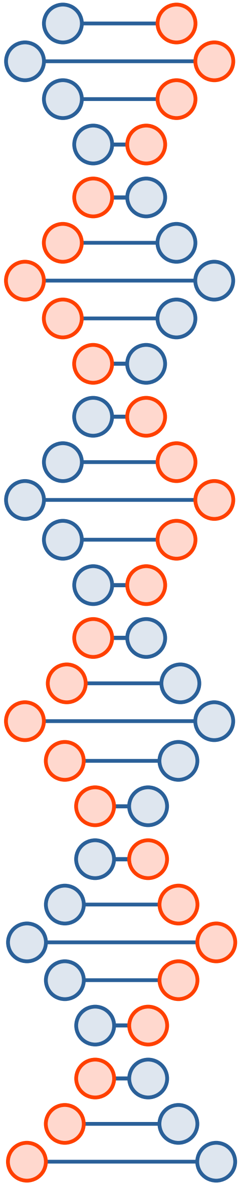
- Can assess phase switching at read level

- Error-correction methods here introduce very little phase switching overall

    Peregrine-2021 may introduce more phase switching than others tested

    - change program settings?


SUP
Herro
Brutal Rewrite
Peregrine_2021
Dechat
Num. Reads
0
1000
2000
3000
4000
5000
6000
0
25
50
75
100
Absolute value of difference in percentage matching hapmers



# Wrap up

- Can assess phase switching at read level
- Error-correction methods here introduce very little phase switching overall
- Thoughts to explore:

# Wrap up

- Can assess phase switching at read level

- Error-correction methods here introduce very little phase switching overall

- Thoughts to explore:

  What are reads with 0 matching hapmers?

# Wrap up

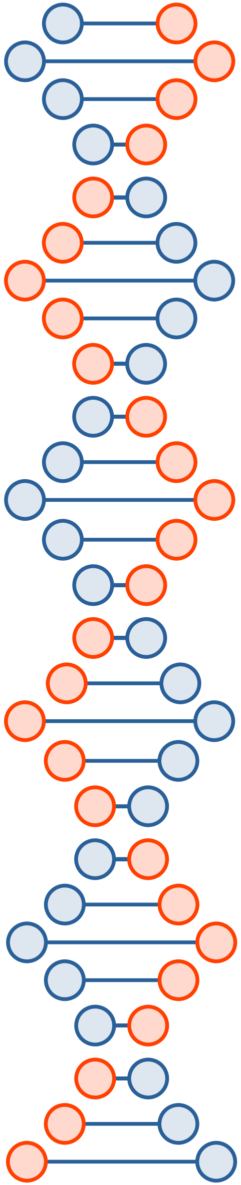
- Can assess phase switching at read level

- Error-correction methods here introduce very little phase switching overall

- Thoughts to explore:

  What are reads with 0 matching hapmers?

  Regions without hapmers



# Wrap up

- Can assess phase switching at read level

- Error-correction methods here introduce very little phase switching overall

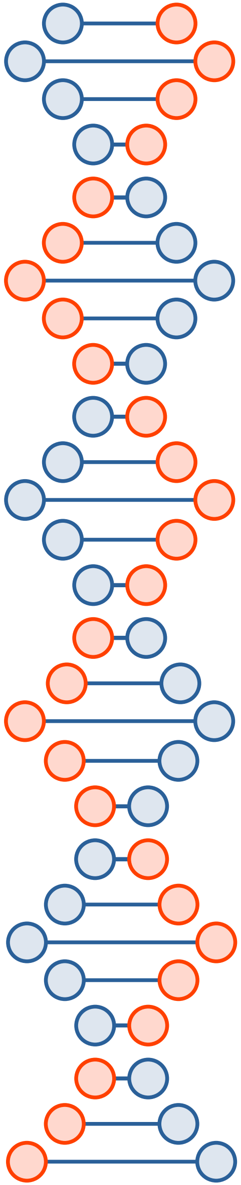
- Thoughts to explore:

  What are reads with 0 matching hapmers?

  Regions without hapmers

  Regions with few hapmers & phase switching

  so that readmers do not match hapmers



# Acknowledgements

- MedUni Wien High Performance Computing Cluster
- Heng Li (BWA, minimap2, SAMtools developer)

# Thanks for your time!

NANO LAB