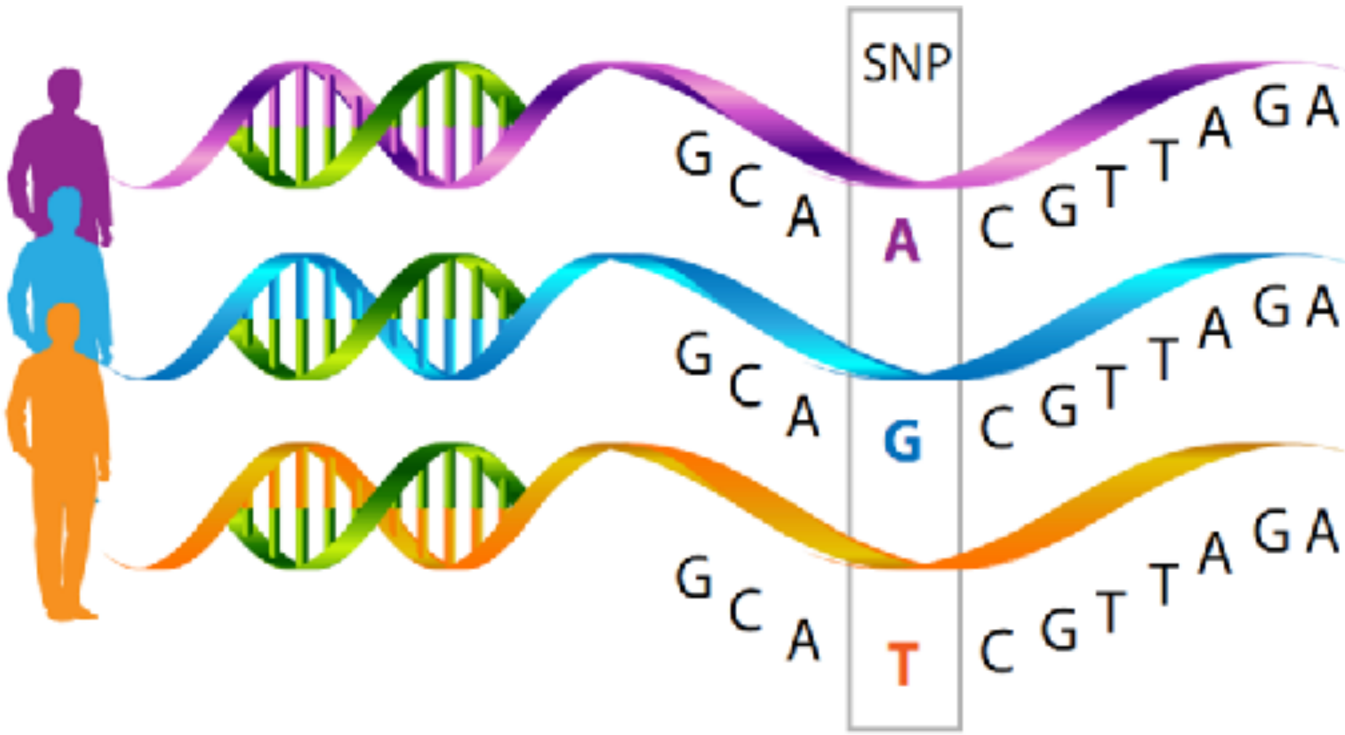
# ComHapDet - A Spatial Community Detection Algorithm for Haplotype Assembly

Abishek Sankararaman, Haris Vikalo, François Baccelli

# Genetic Variations

Different organisms of a species have *similar* genomes.




1 SNP in ~ 1000 nucleotides

Understanding this has effect on human health and medical treatments

Risk to hereditary diseases, effect of drugs on individuals

# Genetic Variations in Humans

Humans are *diploid* - chromosomes come in pairs

A Single Individual's genome

AGGATTCC**A**AGTTA**C**CGAAATTCAGGATTCA**G**GCTTAAATGGCTT

AGGATTCC**G**AGTTA**G**CGAAATTCAGGATTCA**A**GCTTAAATGGCTT

SNP locations are *heterozygous*

# Genetic Variations in Humans

Humans are *diploid* - chromosomes come in pairs

A Single Individual's genome

AGGATTCC**A**AGTTA**C**CGAAATTCAGGATTCA**G**GCTTAAATGGCTT

AGGATTCC**G**AGTTA**G**CGAAATTCAGGATTCA**A**GCTTAAATGGCTT

SNP locations are *heterozygous*

The complete information is provided by Haplotypes

In this ex - (A,C,G) and (G,G,A)

# Genetic Variations in Humans

Humans are *diploid* - chromosomes come in pairs

A Single Individual’s genome

AGGATTCC**A**AGTTA**C**CGAAATTCAGGATTCA**G**GCTTAAATGGCTT

AGGATTCC**G**AGTTA**G**CGAAATTCAGGATTCA**A**GCTTAAATGGCTT

SNP locations are *heterozygous*

The complete information is provided by Haplotypes

In this ex - (A,C,G) and (G,G,A)

Haplotype Assembly - Reconstruct haplotypes from *paired-end reads*

# Haplotype Assembly - Problem

Reconstruct the string from noisy measurements

0 1 1 0 1 1 1 0 1 1 0 0 0 1 1 0 1 $\mathbf{s}$

1 0 0 1 0 0 0 1 0 0 1 1 1 0 0 1 0 $\mathbf{s}^c$

Humans are bi-alletic (binary alphabet)

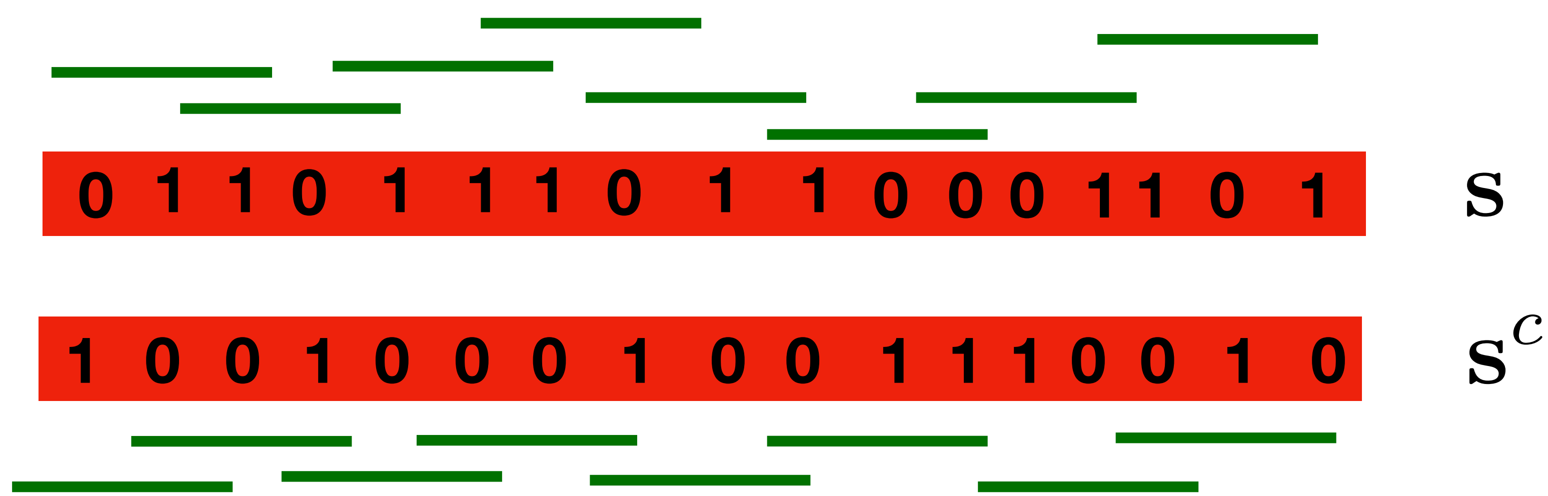
Each *paired-read* consists of

- The underlying string $\mathbf{s}$ or $\mathbf{s}^c$ that is unknown
- A set of locations that is known
- Noisy measurement of the unknown chosen string at the known chosen locations

**_Read 1_ - Positions - 2,10 Values:000,011**

# Haplotype Assembly - Problem

Reconstruct the string from noisy measurements




In our work, we consider *paired-end read measurements*

*Read 1* - Positions - 2,10 Values:000,011
*Read 2* - Positions - 4,11 Values:00,01
*Read 3* - Positions - 1,9 Values:0010,01101
*Read 4* - Positions - 2,11 Values:00011,01

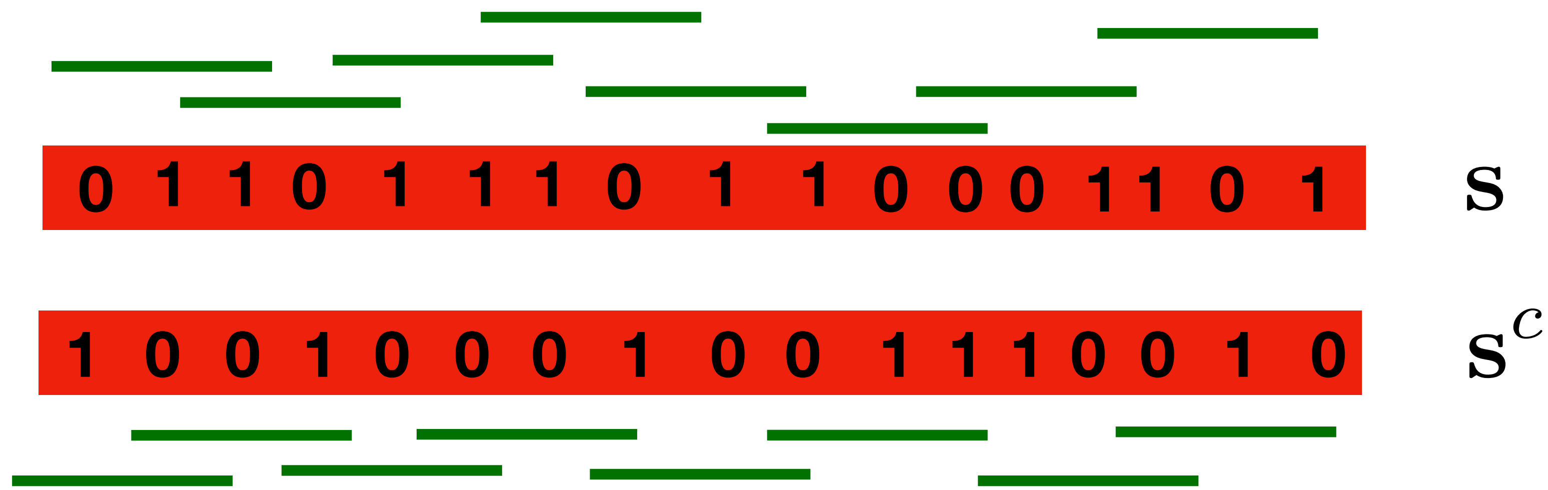
.
.
.

*Read m* - Positions - 21,40. Values:0,01100

Handle inaccuracies in practice (not necessarily 2 strands) in the sequel

# Haplotype Assembly - Problem

Reconstruct the string from noisy paired-end read measurements



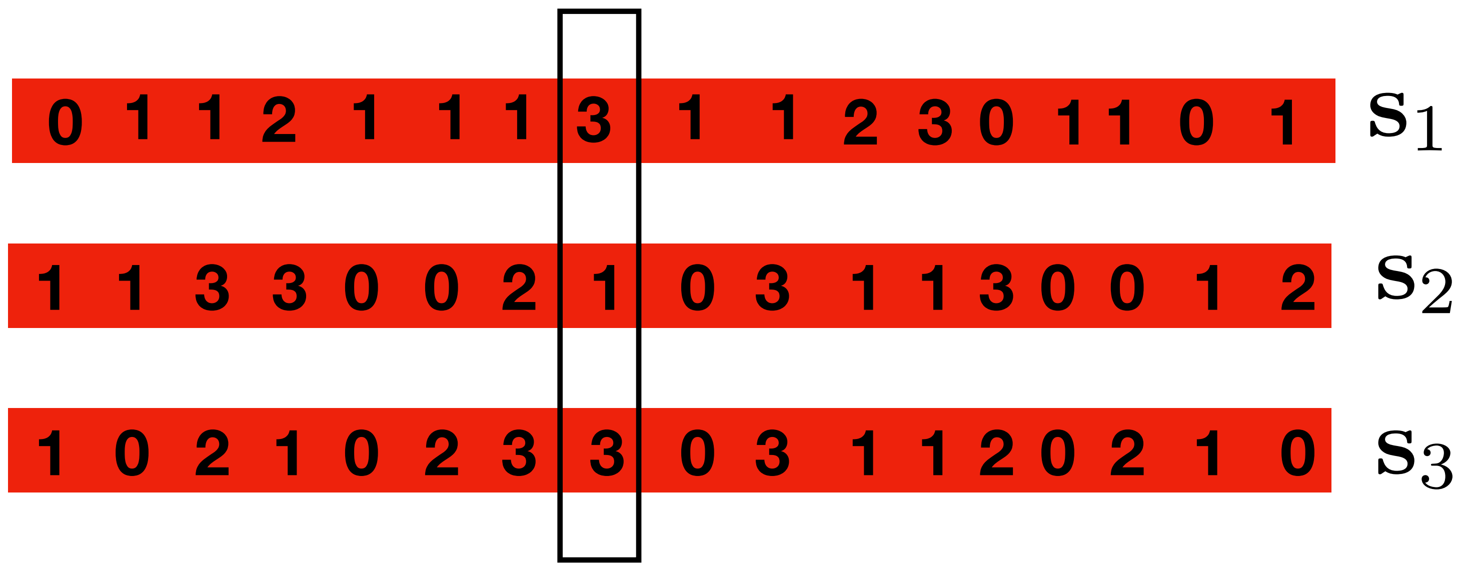

Fundamental and *challenging* problem in computational genomics

( NP Hard [Bonnizzoni e.al. ’16])

Approximations and heuristics for binary alphabet - long history

# Haplotype Assembly - Problem

Reconstruct the string from noisy paired-end read measurements




*At all positions, not all strings are identical*

Fundamental and *challenging* problem in computational genomics
( NP Hard [Bonnizzoni e.al. '16])

Approximations and heuristics for binary alphabet - long history

We consider the general case of multiple strings and multiple alphabets
*(Plant Species)*

# Prior Work

Majority of prior work focussed on binary alphabet case.

Hapcut - [Bansal et.al. ’08]

HapCompass - [Aguiar. et.al. ’12]

HapTree - [Berger et.al. ’14]

SDHaP - [Das et.al. ’15]

HPoP - [Xie et.al.’16]

BP - [Puljiz et.al.’16]

AltHap - [Hashemi et.al.’18]

*Only prior method to work for polyploid polyallelic case*

# Proposed Algorithm

1) Create a weighted spatial graph $G$

2) Cluster nodes(reads) into those originating from same haplotype

3) Reconstruct position by position as the majority alphabet among the reads estimated to come from this haplotype and covering the position

# Proposed Algorithm

1) Create a weighted spatial graph $G$

Reads -> Nodes with spatial embedding

2) Cluster nodes(reads) into those originating from same haplotype

3) Reconstruct position by position as the majority alphabet among the reads estimated to come from this haplotype and covering the position

# Proposed Algorithm

1) Create a weighted spatial graph $G$

Reads -> Nodes with spatial embedding

2) Cluster nodes(reads) into those originating from same haplotype

Euclidean Community Detection

3) Reconstruct position by position as the majority alphabet among the reads estimated to come from this haplotype and covering the position

# Proposed Algorithm

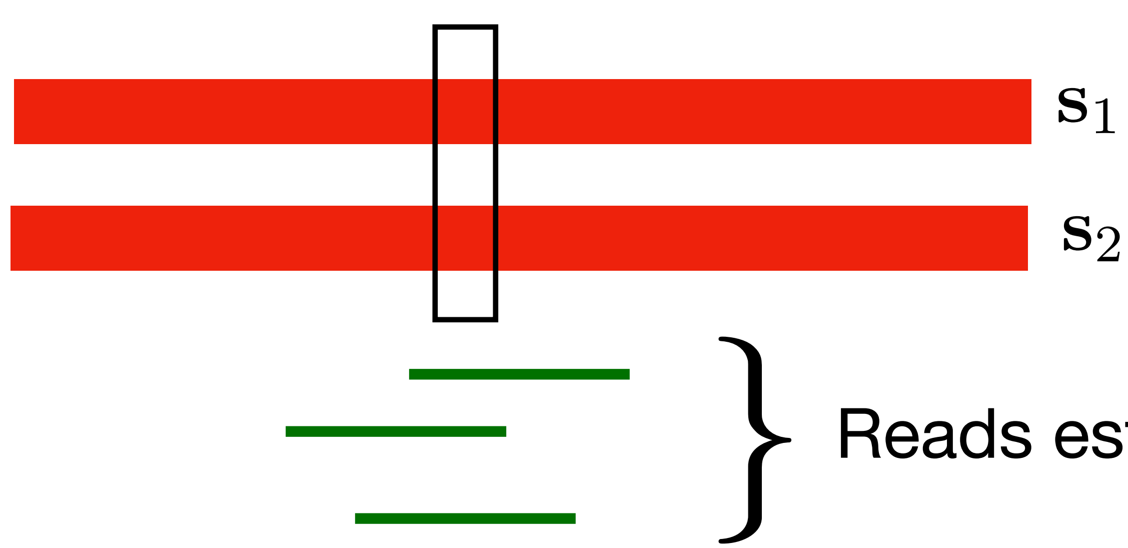
1) Create a weighted spatial graph $G$

Reads -> Nodes with spatial embedding

2) Cluster nodes(reads) into those originating from same haplotype

Euclidean Community Detection

3) Reconstruct position by position as the majority alphabet among the reads estimated to come from this haplotype and covering the position



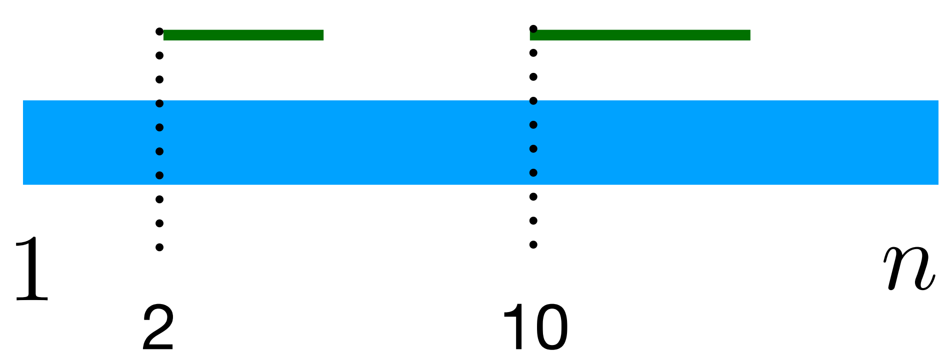
# Constructing the Weighted Spatial Graph

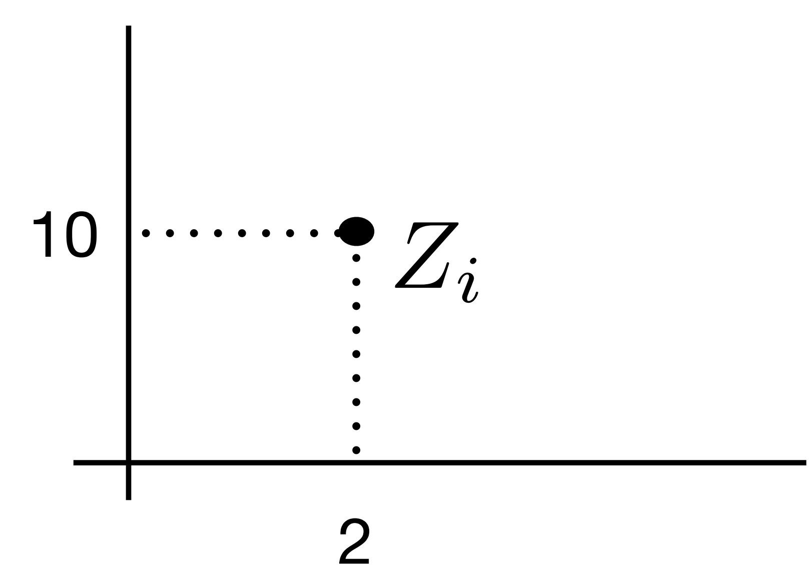
## Nodes -> Reads

**_Read i_ - Positions - 2,10  Values:000,011**

Two node features - *unknown* haplotype and *known* positions

$Z_i \in \{1, \cdots, k\}$ $\qquad$ $(2, 10) \in \{1, \cdots, n\}^2$



# Constructing the Weighted Spatial Graph

## Nodes -> Reads

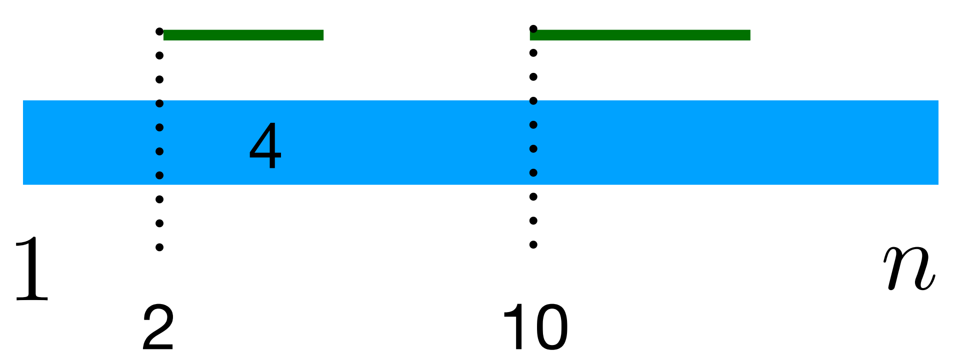
**Read *i* - Positions - 2,10 Values:000,011**

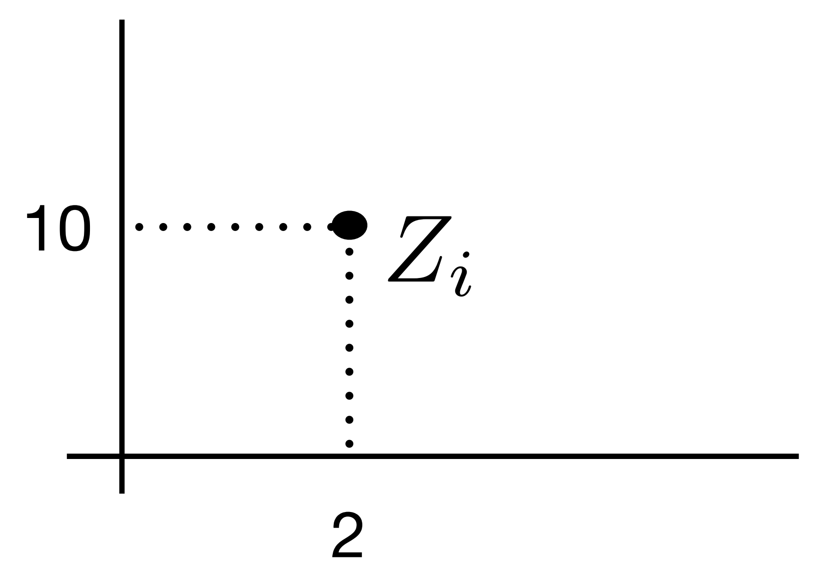
Two node features - *unknown* haplotype and *known* positions

$Z_i \in \{1, \cdots, k\}$ $(2,10) \in \{1, \cdots, n\}^2$

## Edge weight

$$w_{ij} = \frac{\#\text{Sites the reads agrees on} - \#\text{Sites the reads differs}}{\#\text{Total number of overlapping sites}}$$



# Constructing the Weighted Spatial Graph

Nodes -> Reads

***Read i*** **- Positions - 2,10 Values:000,011**

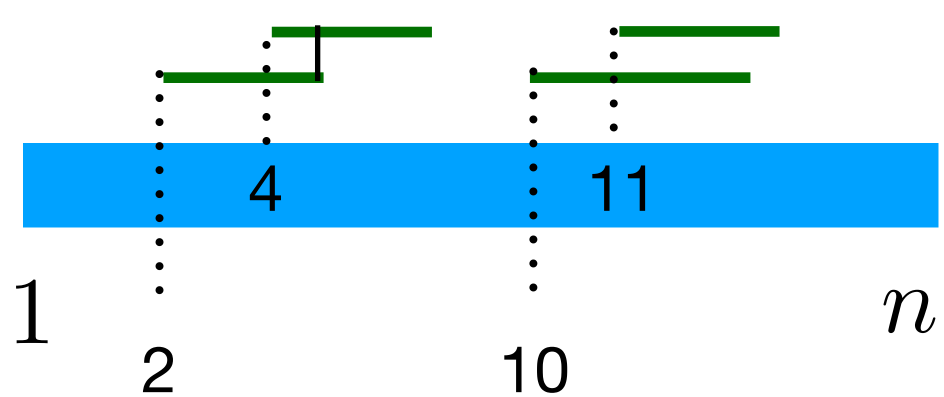
Two node features - *unknown* haplotype and *known* positions

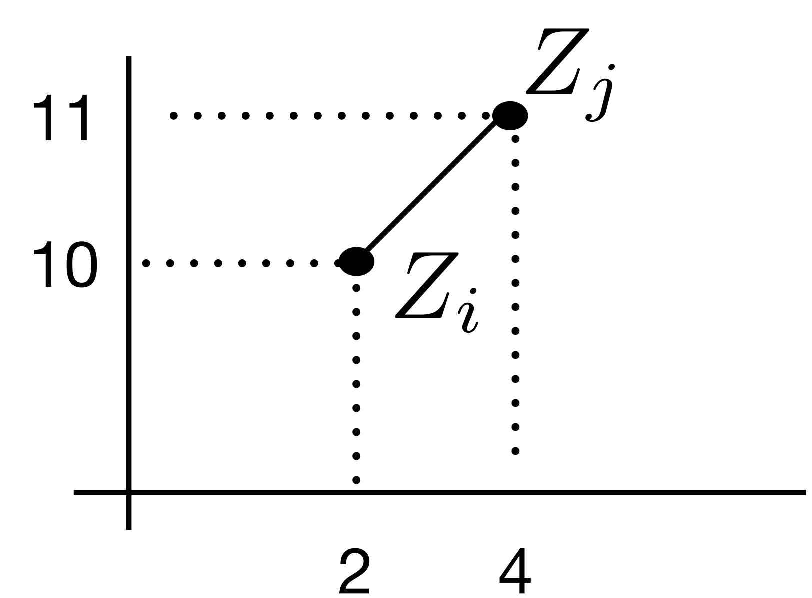
$Z_i \in \{1, \cdots, k\}$ $\quad (2, 10) \in \{1, \cdots, n\}^2$

Edge weight $\quad w_{ij} = \dfrac{\#\text{Sites the reads agrees on} - \#\text{Sites the reads differs}}{\#\text{Total number of overlapping sites}}$

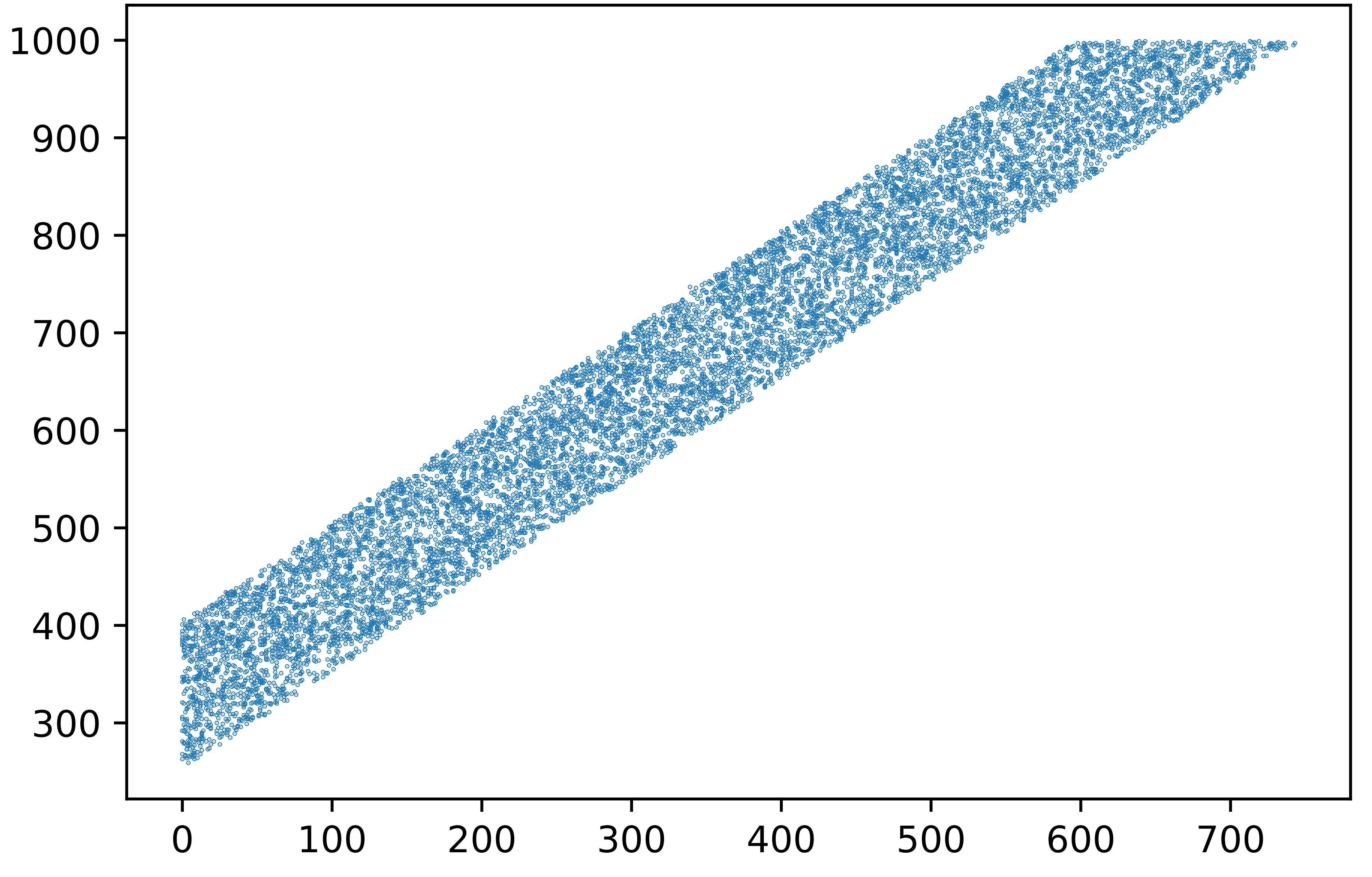
***Read j*** **- Positions - 4,11 Values:01,101**

$w_{ij} = \dfrac{2-1}{2+1}$ $\quad$ Overlapping Sites $= \{4, 10, 11\}$



# Example Node Embeddings of Reads




Benchmark simulation data with 4 strings and string length 700.

# Euclidean Community Detection

Task - Cluster nodes of $G$ according to haplotypes the read originates from

Not standard graph clustering due to presence of spatial labels

Key Structure in the graph $G$

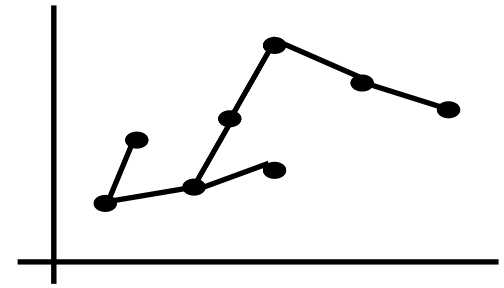
1. Edges are localized in space

2. On average, larger weight between nodes of the same cluster(haplotype)

3. The density of reads belonging to different clusters are identical in space

# Key Structure in G

1. Edges are localized in space



Paired-end reads are typically short

2. On avg, larger weight between nodes of the same cluster(haplotype)

$$w_{ij} = \frac{\#\text{Sites the reads agrees on} - \#\text{Sites the reads differs}}{\#\text{Total number of overlapping sites}}$$

0 1 1 0 1 1 1 0 1 1 0 0 0 1 1 0 1 $\mathbf{s}$

1 0 0 1 0 0 0 1 0 0 1 1 1 0 0 1 0 $\mathbf{s}^c$

3. The density of reads belonging to different clusters are identical in space

*In each location (x,y) of G, a read(node) is equally likely to be from any haplotype*

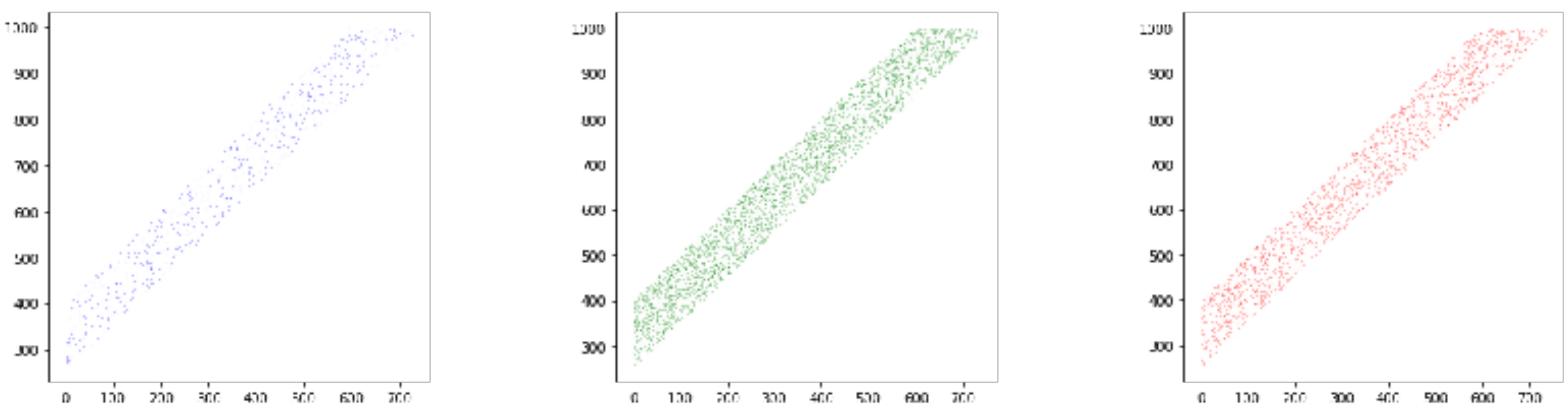
# Key Structure in G

3. The density of reads belonging to different clusters are identical in space

*In each location (x,y) - a read is equally likely to be from any haplotype*

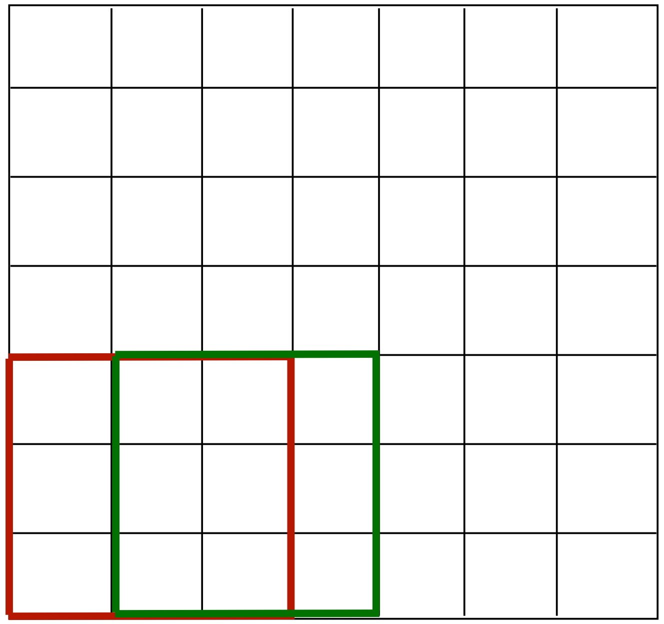
The impact of this assumption

Standard Spectral Clustering ignoring spatial data fails



Spatially unbalanced clusters are recovered.

# Euclidean Community Detection



Algorithm

1) Spectral Clustering of sub-graph in each block

2) Sequentially synchronize the clusters in different blocks

Statistical benefits

- Increased Precision as a node is in multiple boxes and hence, multiple estimates
- Regularization for equal density of haplotypes in space

Computational benefits

- Only perform clustering on small sub-graphs

# Proposed Algorithm

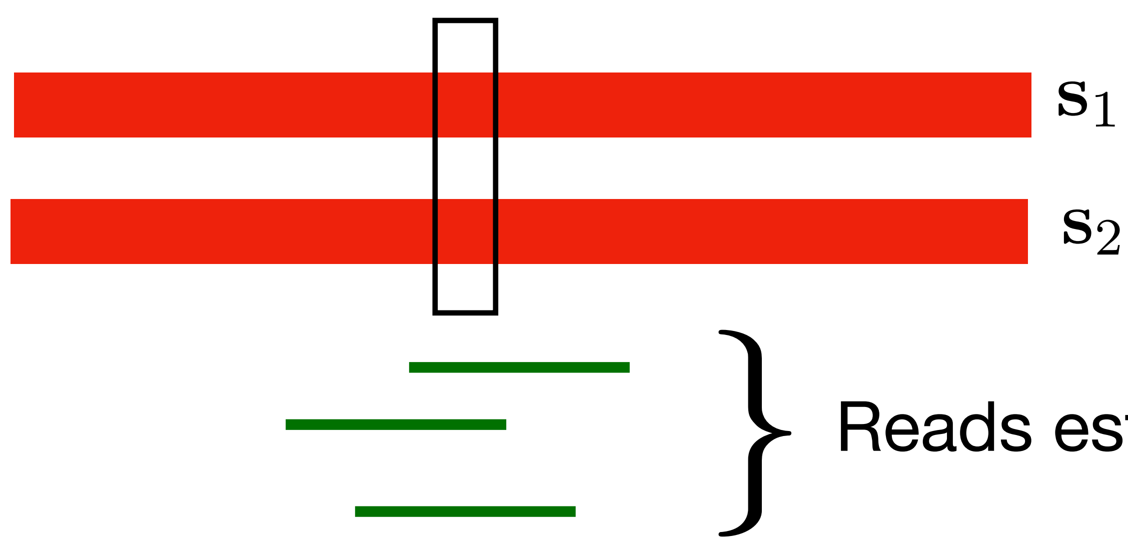
1) Create a weighted spatial graph $G$

Reads -> Nodes with spatial embedding

2) Cluster nodes(reads) into those originating from same haplotype

Euclidean Community Detection

3) Reconstruct position by position as the majority alphabet among the reads estimated to come from this haplotype and covering the position




Reads estimated in step 2 to come from $\mathbf{s}_2$

# Performance Metrics

1. CPR (Correct Phasing Rate) - Fraction of sites correctly recovered

(Needs ground truth to compute) $$\max_{\pi \in S_k} \frac{1}{m} \sum_{i=1}^{m} \prod_{l=1}^{k} \mathbf{1}_{\widehat{s}_l[i] = s_{\pi(l)}[i]}$$

0 1 1 0
1 0 0 1

Ground truth

or

0 1 1 1
1 0 0 0

CPR = 0.75

The two possible permutations of estimates

# Performance Metrics

1. CPR (Correct Phasing Rate) - Fraction of sites correctly recovered

(Needs ground truth to compute)
$$\max_{\pi\in S_k} \frac{1}{m}\sum_{i=1}^{m}\prod_{l=1}^{k}\mathbf{1}_{\widehat{s}_l[i]=s_{\pi(l)}[i]}$$

| | | | |
|---|---|---|---|
| 0 | 1 | 1 | 0 |
| 1 | 0 | 0 | 1 |

Ground truth

or

| | | | |
|---|---|---|---|
| 0 | 1 | 1 | 1 |
| 1 | 0 | 0 | 0 |

CPR = 0.75

The two possible permutations of estimates

2. MEC (Minimum Error Correction)

How many values in each read fails to align with estimates

(No Ground truth knowledge)
$$\sum_{u=1}^{n}\min_{l\in\{1,\cdots,k\}}\sum_{i=1}^{m}\mathbf{1}_{\text{Read } u \text{ covers site } i}\mathbf{1}_{\hat{s}^{(u)}[i]\neq s_l[i]}$$

```
Read 1 - Pos 1  Values:111
Read 2 - Pos 3  Values:00
```

| | | | |
|---|---|---|---|
| 1 | 0 | 0 | 0 |
| 0 | 1 | 1 | 1 |

MEC = 1

Estimated String

# Experimental Evaluation

Problem Parameters

1. Coverage - Average number of reads covering any site
2. Error Probability - The error made by reads in reporting sites

# Experimental Evaluation

## Problem Parameters

1. Coverage - Average number of reads covering any site
2. Error Probability - The error made by reads in reporting sites

Competing baselines - AltHap [Hashemi et.al' 18] and HPoP [Xie.et.al.'16]

| Cov | Err | ComHapDet | | | | AltHap | | | | HPoP | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | CPR | MEC | t(sec) | $\sigma$(CPR) | CPR | MEC | t(sec) | $\sigma$(CPR) | CPR | MEC | t(sec) | $\sigma$(CPR) |
| 7 | 0.05 | 99.24 | **662.7** | 18.34 | 0.28 | **99.99** | 960.7 | 13.46 | 0.01 | 99.8 | 961.5 | 3.1 | 0.12 |
| | 0.1 | 98.18 | **1289.13** | 18.88 | 0.45 | **99.86** | 1871.25 | 13.84 | 0.14 | 99.4 | 1868.5 | 3.42 | 0.3 |
| | 0.2 | 80.49 | **2640** | 18.2 | 1.6 | **85.9** | 4844.1 | 13.69 | 1.3 | 84.8 | 3862.7 | 3.53 | 8.64 |
| 10 | 0.05 | 99.86 | **923.4** | 29.23 | 0.11 | **99.99** | 1352.9 | 15.43 | 0.01 | **99.99** | 1354.92 | 1.75 | 0.03 |
| | 0.1 | 99.47 | **1831.13** | 27.05 | 5.29 | 98.09 | 3132.3 | 15.5 | 0.8 | **99.84** | 2667.46 | 3.14 | 0.38 |
| | 0.2 | 91.85 | **3575.86** | 27.88 | 1.35 | **92.82** | 5231.85 | 24.24 | 1.3 | 88.29 | 5488.2 | 3.33 | 11.52 |
| 15 | 0.05 | 99.98 | **1382.73** | 52.13 | 0.03 | 99.97 | 2034.5 | 29.98 | 0.05 | **100** | 2022.47 | 8.013 | 0 |
| | 0.1 | **99.91** | **2772.93** | 56.37 | 0.13 | 99.9 | 3989.65 | 39.1 | 0.03 | 99.9 | 3986.5 | 7.3 | 0.04 |
| | 0.2 | **97.91** | **5283.6** | 50.02 | 0.38 | 96.80 | 7646.25 | 39.2 | 0.56 | 96.72 | 7788.95 | 6.94 | 1.8 |

Synthetic Diploid-Biallelic data. Haplotype length 1000 with paired end Average read length 2

# Experimental Evaluation

## Synthetic Triploid-Tetraallelic data

| Coverage | Err Rate | ComHapDet | | | | | AltHap | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | CPR | MEC | t(sec) | $\sigma$(CPR) | M-CPR | CPR | MEC | t(sec) | $\sigma$(CPR) | M-CPR |
| 7 | 0.002 | **98.6** | **97** | 76.7 | 0.88 | 99.5 | 88.95 | 687 | 295.22 | 13.97 | 92.97 |
| | 0.01 | **93.78** | **662.1** | 81.25 | 10.794 | 96.95 | 88.69 | 966.2 | 289.75 | 17.5 | 92.44 |
| | 0.05 | **97.11** | **1504.7** | 75.52 | 1.571 | 98.9 | 80.13 | 2887.4 | 332.1 | 20.27 | 86.31 |
| 10 | 0.002 | **99.75** | **93.7** | 137.5 | 0.168 | 99.91 | 83.67 | 1215.4 | 593.19 | 20.65 | 88.42 |
| | 0.01 | **99.67** | **413.1** | 135.9 | 0.21 | 99.89 | 92.72 | 1029.1 | 592.74 | 14.59 | 95.36 |
| | 0.05 | **99.44** | **2021.9** | 139.78 | 0.27 | 99.77 | 92.73 | 3632.0 | 592.44 | 14.59 | 95.36 |
| 15 | 0.002 | **99.91** | **124.6** | 300.35 | 0.11 | 99.97 | 89.89 | 1725 | 708.5 | 16.07 | 94 |
| | 0.01 | **99.88** | **611.1** | 307.88 | 0.07 | 99.95 | 95.96 | 1628.6 | 781 | 9.82 | 97.58 |
| | 0.05 | **99.86** | **2981.5** | 297.19 | 0.15 | 99.95 | 87.43 | 6721.3 | 713.3 | 20.36 | 92.09 |

Haplotype length 1000 with paired end average read length 2

# Experimental Evaluation

## Synthetic Tetraploid-Tetraallelic data

| Coverage | Err Rate | ComHapDet | | | | | AltHap | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | CPR | MEC | t(sec) | $\sigma$(CPR) | M-CPR | CPR | MEC | t(sec) | $\sigma$(CPR) | M-CPR |
| 7 | 0.002 | **79.97** | **1316.25** | 143.48 | 20.27 | 91.8 | 76.08 | 1388.6 | 521.36 | 20.81 | 87.49 |
| | 0.01 | 79.09 | **1640.0** | 118.52 | 17.84 | 91.8 | **79.86** | 1812.8 | 515.78 | 20.45 | 88.05 |
| | 0.05 | 68.34 | 3722.8 | 129.66 | 13.98 | 87.29 | **83.59** | **3481.9** | 503.13 | 20.23 | 91.97 |
| 10 | 0.002 | **98.86** | **193.1** | 253.32 | 1.42 | 99.64 | 71.92 | 1979.7 | 594.3 | 15.5 | 85.58 |
| | 0.01 | **99.17** | **585.9** | 261.81 | 0.41 | 99.76 | 85.44 | 1779.4 | 585 | 18.53 | 92.10 |
| | 0.05 | **98.2** | **2727.7** | 238.56 | 0.64 | 99.51 | 78.55 | 5331.4 | 667.49 | 15.55 | 89.65 |
| 15 | 0.002 | **99.75** | **182.7** | 487.02 | 0.22 | 99.93 | 85.21 | 2614.6 | 684.45 | 18.39 | 92.01 |
| | 0.01 | **99.75** | **806.5** | 482.74 | 0.18 | 99.94 | 83.53 | 3973.7 | 684.13 | 17.41 | 92.61 |
| | 0.05 | **99.0** | **4101.4** | 523.78 | 298.88 | 99.65 | 95.13 | 6397.6 | 682.51 | 14.47 | 97.38 |

Haplotype length 1000 with paired end average read length 2

# Experimental Evaluation

## Synthetic Hexaploid-Tetraallelic data

| Coverage | Err Rate | ComHapDet | | | | | AltHap | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | CPR | MEC | t(sec) | $\sigma$(CPR) | M-CPR | CPR | MEC | t(sec) | $\sigma$(CPR) | M-CPR |
| 10 | 0.002 | **78.89** | 2256.6 | 551.11 | 15.62 | 94.05 | 75.97 | **2022.9** | 977.93 | 20.01 | 90.59 |
| | 0.01 | **84.09** | **2250.4** | 563.20 | 13.96 | 95.83 | 70.39 | 3533.7 | 919.85 | 19.88 | 86.82 |
| | 0.05 | 48.77 | 9578.4 | 526.31 | 25.55 | 81.86 | **75.76** | **7440.7** | 1222.07 | 17.85 | 90 |
| 15 | 0.002 | **99.3** | **308.2** | 1295.63 | 0.3 | 99.87 | 70.36 | 4960.6 | 1780.37 | 25.19 | 87.32 |
| | 0.01 | **97.44** | **1528.5** | 1359.14 | 5.42 | 99.37 | 77.68 | 5493.4 | 1624.56 | 23.17 | 89.94 |
| | 0.05 | **94.74** | **6554.2** | 1207.46 | 11.654 | 98.65 | 65.89 | 13751.6 | 2406.31 | 19.04 | 87.205 |
| 20 | 0.002 | **99.52** | **382.8** | 2097.09 | 0.23 | 99.91 | 77.13 | 7095.1 | 7561.21 | 19.35 | 91.85 |
| | 0.01 | **99.51** | **1654.3** | 2116.48 | 0.22 | 99.9 | 87.32 | 5905.4 | 6862.06 | 17.98 | 96.05 |
| | 0.05 | **99.58** | **7912.8** | 2298.87 | 0.17 | 99.92 | 65.06 | 23381.8 | 8563.43 | 24.5 | 86.85 |

Haplotype length 1000 with paired end average read length 2

# Experimental Evaluation

Real Tetraploid-Biallelic data from Chromosome 5 of Potato

| Method | MEC Score | t(secs) |
|---|---|---|
| ComHapDet | 17738 | 207 |
| AltHap | 14580 | 105 |
| HPoP | 10596 | 102 |
| HapCompass | 12497 | 375 |
| HapTree | 46617 | 215 |

All reads are not exactly paired end

- Single ended reads are placed on the diagonal
- If a read has 3 or more strands, then they are split into multiple paired and/or single ended reads

Our method has a poorer MEC compared to others

- Low coverage (~4) in the dataset
- Tetraploid balletic is challenging for our model (because edge weights become biased)

MEC only a *proxy* of true performance absent ground truth.

# Conclusions

A novel methodology to assemble both diploid and polyploid haplotypes

Key observation - spatial graph representation of paired end reads is useful

New clustering algorithm to cluster graphs with spatial labels

\

Thank You