Efficient and Accurate Clustering for Large-Scale Genetic Mapping

V. Strnadová (Neeley)*, Aydın Buluç*, Jarrod Chapman*, ¶, Joseph Gonzalez §, John Gilbert++, Stefanie Jegelka§, Daniel Rokhsar*,§, ¶, Leonid Oliker*

*Lawrence Berkeley National Labs, ++ UC Santa Barbara, § UC Berkeley, ¶ Joint Genome Institute
Motivation

- High-throughput sequencing methods have produced a flood of inexpensive genetic information

- Genetic maps are important to breeding studies but genetic mapping software is prohibitively slow on large data sets
# The Genetic Mapping Problem

## Data

<table>
<thead>
<tr>
<th></th>
<th>$i_1$</th>
<th>$i_2$</th>
<th>$i_3$</th>
<th>$i_4$</th>
<th>$i_5$</th>
<th>$i_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m_1$</td>
<td>A</td>
<td>B</td>
<td>-</td>
<td>-</td>
<td>A</td>
<td>-</td>
</tr>
<tr>
<td>$m_2$</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>A</td>
</tr>
<tr>
<td>$m_3$</td>
<td>A</td>
<td>A</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>B</td>
</tr>
<tr>
<td>$m_4$</td>
<td>A</td>
<td>-</td>
<td>B</td>
<td>-</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>$m_5$</td>
<td>B</td>
<td>-</td>
<td>B</td>
<td>A</td>
<td>-</td>
<td>A</td>
</tr>
<tr>
<td>$m_6$</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$m_7$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>A</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>$m_8$</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>B</td>
<td>-</td>
<td>A</td>
</tr>
<tr>
<td>$m_9$</td>
<td>A</td>
<td>B</td>
<td>-</td>
<td>B</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$m_{10}$</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>-</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>$m_{11}$</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>$m_{12}$</td>
<td>B</td>
<td>-</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>-</td>
</tr>
<tr>
<td>$m_{13}$</td>
<td>B</td>
<td>B</td>
<td>-</td>
<td>A</td>
<td>A</td>
<td>-</td>
</tr>
<tr>
<td>$m_{14}$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>B</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>$m_{15}$</td>
<td>B</td>
<td>-</td>
<td>-</td>
<td>A</td>
<td>A</td>
<td>B</td>
</tr>
</tbody>
</table>
The Genetic Mapping Problem

<table>
<thead>
<tr>
<th>Data</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$i_1$</td>
<td>$i_2$</td>
<td>$i_3$</td>
<td>$i_4$</td>
<td>$i_5$</td>
<td>$i_6$</td>
</tr>
<tr>
<td>$m_1$</td>
<td>A</td>
<td>B</td>
<td>-</td>
<td>-</td>
<td>A</td>
<td>-</td>
</tr>
<tr>
<td>$m_2$</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>A</td>
</tr>
<tr>
<td>$m_3$</td>
<td>A</td>
<td>A</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>B</td>
</tr>
<tr>
<td>$m_4$</td>
<td>A</td>
<td>-</td>
<td>B</td>
<td>-</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>$m_5$</td>
<td>B</td>
<td>-</td>
<td>B</td>
<td>A</td>
<td>-</td>
<td>A</td>
</tr>
<tr>
<td>$m_6$</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$m_7$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>A</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>$m_8$</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>B</td>
<td>-</td>
<td>A</td>
</tr>
<tr>
<td>$m_9$</td>
<td>A</td>
<td>B</td>
<td>-</td>
<td>B</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$m_{10}$</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>-</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>$m_{11}$</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>$m_{12}$</td>
<td>B</td>
<td>-</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>-</td>
</tr>
<tr>
<td>$m_{13}$</td>
<td>B</td>
<td>B</td>
<td>-</td>
<td>A</td>
<td>A</td>
<td>-</td>
</tr>
<tr>
<td>$m_{14}$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>B</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>$m_{15}$</td>
<td>B</td>
<td>-</td>
<td>-</td>
<td>A</td>
<td>A</td>
<td>B</td>
</tr>
</tbody>
</table>

Cluster

Linkage group 1

$\{m_3, m_4, m_7, m_10, m_5, m_{12}, m_{11}\}$

Linkage group 2

$\{m_8, m_1, m_{15}, m_2, m_9\}$

(missing data)
## The Genetic Mapping Problem

### Data

<table>
<thead>
<tr>
<th>$i_1$</th>
<th>$i_2$</th>
<th>$i_3$</th>
<th>$i_4$</th>
<th>$i_5$</th>
<th>$i_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m_1$</td>
<td>A</td>
<td>B</td>
<td>-</td>
<td>-</td>
<td>A</td>
</tr>
<tr>
<td>$m_2$</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>$m_3$</td>
<td>A</td>
<td>A</td>
<td>-</td>
<td>-</td>
<td>B</td>
</tr>
<tr>
<td>$m_4$</td>
<td>A</td>
<td>-</td>
<td>B</td>
<td>-</td>
<td>B</td>
</tr>
<tr>
<td>$m_5$</td>
<td>B</td>
<td>-</td>
<td>B</td>
<td>A</td>
<td>-</td>
</tr>
<tr>
<td>$m_6$</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>-</td>
</tr>
<tr>
<td>$m_7$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>$m_8$</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>B</td>
<td>-</td>
</tr>
<tr>
<td>$m_9$</td>
<td>A</td>
<td>B</td>
<td>-</td>
<td>B</td>
<td>-</td>
</tr>
<tr>
<td>$m_{10}$</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>-</td>
<td>A</td>
</tr>
<tr>
<td>$m_{11}$</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>$m_{12}$</td>
<td>B</td>
<td>-</td>
<td>A</td>
<td>B</td>
<td>A</td>
</tr>
<tr>
<td>$m_{13}$</td>
<td>B</td>
<td>B</td>
<td>-</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>$m_{14}$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>B</td>
<td>A</td>
</tr>
<tr>
<td>$m_{15}$</td>
<td>B</td>
<td>-</td>
<td>-</td>
<td>A</td>
<td>A</td>
</tr>
</tbody>
</table>

### Linkage Groups

- **Linkage group 1**
  - Cluster: $m_{13}, m_{14}, m_4, m_7, m_3, m_{10}, m_5, m_{12}, m_{11}$
  - Order: $m_3, m_6, m_{10}, m_4, m_7, m_13, m_{12}, m_{11}, m_{14}$

- **Linkage group 2**
  - Cluster: $m_2, m_8, m_1, m_{15}$
  - Order: $m_2, m_1, m_{15}, m_8$
The Genetic Mapping Problem

<table>
<thead>
<tr>
<th></th>
<th>$i_1$</th>
<th>$i_2$</th>
<th>$i_3$</th>
<th>$i_4$</th>
<th>$i_5$</th>
<th>$i_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m_1$</td>
<td>A</td>
<td>B</td>
<td>-</td>
<td>-</td>
<td>A</td>
<td>-</td>
</tr>
<tr>
<td>$m_2$</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>A</td>
</tr>
<tr>
<td>$m_3$</td>
<td>A</td>
<td>A</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>B</td>
</tr>
<tr>
<td>$m_4$</td>
<td>A</td>
<td>-</td>
<td>B</td>
<td>-</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>$m_5$</td>
<td>B</td>
<td>-</td>
<td>B</td>
<td>A</td>
<td>-</td>
<td>B</td>
</tr>
<tr>
<td>$m_6$</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>-</td>
<td>A</td>
</tr>
<tr>
<td>$m_7$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>A</td>
<td>B</td>
<td>-</td>
</tr>
<tr>
<td>$m_8$</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>B</td>
<td>-</td>
<td>A</td>
</tr>
<tr>
<td>$m_9$</td>
<td>A</td>
<td>B</td>
<td>-</td>
<td>B</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$m_{10}$</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>-</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>$m_{11}$</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>$m_{12}$</td>
<td>B</td>
<td>-</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>-</td>
</tr>
<tr>
<td>$m_{13}$</td>
<td>B</td>
<td>B</td>
<td>-</td>
<td>A</td>
<td>A</td>
<td>-</td>
</tr>
<tr>
<td>$m_{14}$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>B</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>$m_{15}$</td>
<td>B</td>
<td>-</td>
<td>-</td>
<td>A</td>
<td>A</td>
<td>B</td>
</tr>
</tbody>
</table>

**Linkage group 1:**

- $m_3$
- $m_6$
- $m_{10}$
- $m_4$
- $m_7$
- $m_2$
- $m_1$
- $m_5$
- $m_{12}$
- $m_{11}$
- $m_{13}$
- $m_{14}$

**Linkage group 2:**

- $m_{15}$
- $m_8$
- $m_9$
The Need for Large-Scale Clustering in Genetic Mapping

• Hundreds of thousands of genetic markers available, but current software can only handle up to \(~10,000\) markers

• A major bottleneck is the linkage-group-finding phase

• Popular mapping tools all handle this phase the same way, with an \(O(M^2)\) clustering algorithm for \(M\) markers
The Need for Large-Scale Clustering in Genetic Mapping

- Hundreds of thousands of genetic markers available, but current software can only handle up to \(~10,000\) markers
- A major bottleneck is the linkage-group-finding phase
- Popular mapping tools all handle this phase the same way, with an \(O(M^2)\) clustering algorithm for \(M\) markers

Our solution: A \textit{fast, scalable} clustering algorithm tailored to genetic marker data
### Data

<table>
<thead>
<tr>
<th></th>
<th>$i_1$</th>
<th>$i_2$</th>
<th>$i_3$</th>
<th>$i_4$</th>
<th>$i_5$</th>
<th>$i_6$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$m_1$</td>
<td>A</td>
<td>B</td>
<td>-</td>
<td>-</td>
<td>A</td>
<td>-</td>
</tr>
<tr>
<td>$m_2$</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>A</td>
</tr>
<tr>
<td>$m_3$</td>
<td>A</td>
<td>A</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>B</td>
</tr>
<tr>
<td>$m_4$</td>
<td>A</td>
<td>-</td>
<td>B</td>
<td>-</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>$m_5$</td>
<td>B</td>
<td>-</td>
<td>B</td>
<td>A</td>
<td>-</td>
<td>A</td>
</tr>
<tr>
<td>$m_6$</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$m_7$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>A</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>$m_8$</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>B</td>
<td>-</td>
<td>A</td>
</tr>
<tr>
<td>$m_9$</td>
<td>A</td>
<td>B</td>
<td>-</td>
<td>B</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>$m_{10}$</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>-</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>$m_{11}$</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>$m_{12}$</td>
<td>B</td>
<td>-</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>-</td>
</tr>
<tr>
<td>$m_{13}$</td>
<td>B</td>
<td>B</td>
<td>-</td>
<td>A</td>
<td>A</td>
<td>-</td>
</tr>
<tr>
<td>$m_{14}$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>B</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>$m_{15}$</td>
<td>B</td>
<td>-</td>
<td>-</td>
<td>A</td>
<td>A</td>
<td>B</td>
</tr>
</tbody>
</table>

### Clustering

Cluster:

- $m_{13}m_{14}$
- $m_4m_7$
- $m_3m_6m_{11}$
- $m_{12}$

Linkage group 1:

- $m_{13}$
- $m_{14}$
- $m_4$
- $m_7$
- $m_3$
- $m_6$
- $m_{11}$
- $m_{12}$

Linkage group 2:

- $m_{15}$
- $m_2$
- $m_1$
- $m_8$
- $m_9$
Standard Approach to Genetic Marker Clustering

(1) Compute the similarity between all $O(M^2)$ pairs of markers, producing a complete graph with $M$ vertices.
(1) Compute the similarity between all $O(M^2)$ pairs of markers, producing a complete graph with $M$ vertices

- Similarity function is the “LOD score”, a logarithm of odds that two markers are genetically linked
LOD Score

Compares the likelihood of obtaining test data if the two markers are indeed linked, to the likelihood of observing the same data purely by chance:

\[
LOD(m_i, m_j) = \log_{10} \frac{P(\text{linkage}_{ij})}{P(\text{no linkage}_{ij})}
\]

Formally,

\[
LOD(m_i, m_j) = \log_{10} \left( \frac{1 - \theta_{ij}}{NR_{ij}} \right)
\]

Where:

\[
R_{ij} = \text{number of recombinant offspring}
\]

\[
NR_{ij} = \text{number of nonrecombinant offspring}
\]

\[
\theta_{ij} = \text{recombination fraction, i.e.} \frac{R_{ij}}{R_{ij} + NR_{ij}}
\]
LOD Score

Compares the likelihood of obtaining test data if the two markers are indeed linked, to the likelihood of observing the same data purely by chance:

\[
\text{LOD}(m_i, m_j) = \log_{10} \frac{P(\text{linkage}_{ij})}{P(\text{no linkage}_{ij})}
\]

Formally,

\[
\text{LOD} = \log_{10} \frac{(1 - \theta_{ij}) \tilde{R}_{ij} \theta_{ij} R_{ij}}{0.5^{R_{ij} + \tilde{R}_{ij}}}
\]

Where:
- \(R_{ij}\) = number of recombinant offspring
- \(\tilde{R}_{ij}\) = number of nonrecombinant offspring
- \(\theta_{ij}\) = recombination fraction, i.e. \(\frac{R_{ij}}{R_{ij} + \tilde{R}_{ij}}\)
LOD Score

Compares the likelihood of obtaining test data if the two markers are indeed linked, to the likelihood of observing the same data purely by chance:

\[
LOD(m_i, m_j) = \log_{10}\left(\frac{(1 - \frac{1}{3})^2 \left(\frac{1}{3}\right)^1}{0.5^3}\right) = 0.074
\]

Formally,

\[
LOD = \log_{10}\left(\frac{(1 - \theta_{ij})\bar{R}_{ij}\theta_{ij}R_{ij}}{0.5^{R_{ij}+\bar{R}_{ij}}}\right)
\]

Where:

- \(R_{ij}\) = number of recombinant offspring
- \(\bar{R}_{ij}\) = number of nonrecombinant offspring
- \(\theta_{ij}\) = recombination fraction, i.e. \(\frac{R_{ij}}{R_{ij} + \bar{R}_{ij}}\)
Standard Approach to Genetic Marker Clustering

(1) Compute the similarity between all $O(M^2)$ pairs of markers, producing a complete graph with $M$ vertices

- Similarity function is the “LOD score”, a logarithm of odds that two markers are genetically linked

(2) Cut all edges below a LOD threshold
Standard Approach to Genetic Marker Clustering

(1) Compute the similarity between all $O(M^2)$ pairs of markers, producing a complete graph with $M$ vertices

- Similarity function is the “LOD score”, a logarithm of odds that two markers are genetically linked

(2) Cut all edges below a LOD threshold

(3) The resulting connected components = linkage groups
Our Approach: The BubbleCluster Algorithm

Primary assumption: Clusters have a “linear structure”
Our Approach: The BubbleCluster Algorithm

Primary assumption: Clusters have a “linear structure”

Key idea: Maintain a set of representative or “sketch” points which reveal the cluster structure
The BubbleCluster Algorithm: Overview

**Input:** set of markers, LOD threshold $\tau$, non-missing threshold $\eta$, low-quality threshold $c$, cluster size threshold $\sigma$

**Output:** set of clusters $C$ and set of representative points $R$

**Phase I:** Build initial set of clusters and set of representative points using high-quality markers (those with at least $\eta$ non-missing entries)

**Phase II:** Add low quality markers (less than $\eta$ non-missing entries) to initial set of clusters

**Phase III:** Attempt to merge small clusters with large
Iteration $i$:

find $r_{MAX} := r_j$ for which $LOD(m, r_j)$ is maximal;

set $C_{MAX} := C_K \in C$ containing $r_{MAX}$
**The BubbleCluster Algorithm**

If \( \text{LOD}(m, r_{MAX}) < \text{LOD\_threshold} \)

\[ \text{LOD}(m, r_{MAX}) \]

\[ m \]

\[ r_{MAX} \]

\[ C_1 \]

\[ C_2 \]
The BubbleCluster Algorithm

If \( LOD(m, r_{MAX}) < LOD\_threshold \)

\[ C = C \cup \{m\} \]
The BubbleCluster Algorithm

Else If ( IS_INTERIOR(r_{MAX}) )

\[ LOD(m, r_{MAX}) \]
The BubbleCluster Algorithm

\[ C_1 = C_{MAX} \]

Else If ( \text{IS\_INTERIOR}(r_{MAX}) )
\[ C_{MAX} = C_{MAX} \cup \{m\} \]
The BubbleCluster Algorithm

Else If ( IS_EXTERIOR(m, r_{MAX}) )

\[ m \]

\[ C_1 = C_{MAX} \]

\[ r_{MAX} \]

\[ C_2 \]
The BubbleCluster Algorithm

Else If ( IS_EXTERIOR(m, r_{MAX}) )
Add m to representative points of C_{MAX}
Add m to C_{MAX}
The BubbleCluster Algorithm

$c_1 = c_{MAX}$

Else  // $m$ is interior to the outer point $r_{MAX}$
The BubbleCluster Algorithm

Else  // $m$ is interior to the outer point $r_{MAX}$
Add $m$ to $C_{MAX}$
The BubbleCluster Algorithm

If $m$ has a LOD score above the threshold to two clusters,
The BubbleCluster Algorithm

If \( m \) has a LOD score above the threshold to two clusters, then merge the clusters and add \( m \) to the merged cluster.
End of Phase I

Stop when all markers with at least $\eta$ non-missing entries have been processed

**Running time:** $O(|H| \log_2 |H| + |H||R|)$

where: $|H|$ = size of high-quality marker set,

    $|R|$ = size of representative point set

Phase II: add low-quality markers

Phase III: merge small clusters
BubbleCluster Parameters

LOD threshold

\[ \text{LOD} = \log_{10} \left( \frac{1 - \theta}{\frac{1}{2} R + N_R} \right) \]

Highest achievable LOD:
\[ \lim_{R \to 0} \log_{10} \left( \frac{1 - \theta}{\frac{1}{2} R + N_R} \right) = \log_{10} 2 \]
Recall the LOD score: 

\[
LOD = \log_{10} \frac{(1 - \theta_{ij}) R_{ij} \theta_{ij}^{R_{ij}}}{0.5 R_{ij} + \bar{R}_{ij}}
\]
BubbleCluster Parameters

LOD threshold

Non-missing threshold: determines how many markers are high-quality

Recall the LOD score: \( LOD = \log_{10} \frac{(1 - \theta_{ij}) R_{ij} \theta_{ij}}{0.5^{R_{ij} + R_{ij}}} \)

Example LOD: \( \log_{10} \frac{(1 - \frac{1}{3})^2 \left(\frac{1}{3}\right)^1}{0.5^3} = 0.074 \)
Recall the LOD score: \[ LOD = \log_{10} \frac{(1 - \theta_{ij})^{\bar{R}_{ij}} \theta_{ij}^{R_{ij}}}{0.5^{R_{ij} + \bar{R}_{ij}}} \]

Highest achievable LOD: \[ \lim_{R_{ij} \to 0} \log_{10} \frac{(1 - \theta_{ij})^{\bar{R}_{ij}} \theta_{ij}^{R_{ij}}}{0.5^{R_{ij} + \bar{R}_{ij}}} = \bar{R}_{ij} \log_{10} 2 \]
Evaluation Metric: $F$-score

Given a golden standard clustering, the $F$-score measures the quality of another clustering by comparing it to the golden standard

Range: $0 \rightarrow 1$

The $F$-score is a harmonic mean of precision and recall

- An $F$-score of 1 indicates perfect precision and perfect recall for every golden standard cluster
### Results: BubbleCluster on Real Data Sets

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Size</th>
<th>Time</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Barley</td>
<td>64K</td>
<td>15 sec</td>
<td>0.9993</td>
</tr>
<tr>
<td>Switchgrass</td>
<td>113K</td>
<td>8.9 min</td>
<td>0.9745</td>
</tr>
<tr>
<td>Switchgrass</td>
<td>548K</td>
<td>1.9 hrs</td>
<td>0.9894</td>
</tr>
<tr>
<td>Wheat</td>
<td>1.58M</td>
<td>1.22 hrs</td>
<td>N/A*</td>
</tr>
</tbody>
</table>

*Results under review at Genome Biology*
## Comparison of Clustering Algorithms for Simulated Data

<table>
<thead>
<tr>
<th>Clustering Method</th>
<th>12.5K Markers</th>
<th>25K Markers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F-score</td>
<td>Time</td>
</tr>
<tr>
<td>JoinMap</td>
<td>0.99964</td>
<td>14 min</td>
</tr>
<tr>
<td>MSTMap</td>
<td>0.99964</td>
<td>4.5 min</td>
</tr>
<tr>
<td>PIC</td>
<td>0.47024</td>
<td>11 sec (+ 4min)</td>
</tr>
<tr>
<td>BubbleCluster</td>
<td><strong>0.99964</strong></td>
<td><strong>6 sec</strong></td>
</tr>
</tbody>
</table>

Simulated data created with Nicholas Tinker’s *Spaghetti* Software.
Effect of the LOD threshold

200K markers, 300 individuals, 35% missing rate

Fixed missing entry threshold, increasing LOD threshold

<table>
<thead>
<tr>
<th>LOD threshold</th>
<th>5</th>
<th>10</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-Score</td>
<td>0.6225</td>
<td>0.9999</td>
<td>0.9999</td>
<td>0.9999</td>
<td>0.9999</td>
<td>0.9999</td>
</tr>
<tr>
<td>Time (s)</td>
<td>48.6</td>
<td>67.0</td>
<td>70.9</td>
<td>82.0</td>
<td>106</td>
<td>170</td>
</tr>
</tbody>
</table>

\[ \tau_1 \text{ vs. } \tau_2 \]
## Effect of the missing data threshold

### 200K markers, 300 individuals, 35% missing rate

Fixed LOD threshold, increasing non-missing entry threshold

<table>
<thead>
<tr>
<th>Non-missing threshold</th>
<th>132</th>
<th>166</th>
<th>172</th>
<th>179</th>
<th>186</th>
<th>192</th>
</tr>
</thead>
<tbody>
<tr>
<td>F-Score</td>
<td>0.9999</td>
<td>0.9999</td>
<td>0.9992</td>
<td>0.9930</td>
<td>0.9610</td>
<td>0.8948</td>
</tr>
<tr>
<td>Time (s)</td>
<td>82.0</td>
<td>84.6</td>
<td>82.7</td>
<td>83.0</td>
<td>81.7</td>
<td>82.0</td>
</tr>
</tbody>
</table>

VS.

VS.

VS.

VS.

VS.
Conclusion

By exploiting the structure underlying genetic marker clusters, we were able to design a fast clustering algorithm tailored to genetic marker data.
Conclusion

By exploiting the structure underlying genetic marker clusters, we were able to design a fast clustering algorithm tailored to genetic marker data.

While remaining highly accurate, we outperform popular existing tools in both runtime and scalability.

<table>
<thead>
<tr>
<th>Clustering Method</th>
<th>12.5K Markers</th>
<th>25K Markers</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F-score</td>
<td>Time</td>
</tr>
<tr>
<td>JoinMap</td>
<td>0.99964</td>
<td>14 min</td>
</tr>
<tr>
<td>MSTMap</td>
<td>0.99964</td>
<td>4.5 min</td>
</tr>
<tr>
<td>PIC</td>
<td>0.47024</td>
<td>11 sec (+ 4 min)</td>
</tr>
<tr>
<td>BubbleCluster</td>
<td><strong>0.99964</strong></td>
<td><strong>6 sec</strong></td>
</tr>
</tbody>
</table>
Future Work

• Use representative points as starting point for ordering phase
Future Work

• Use representative points as starting point for ordering phase

• Provide a more thorough theoretical analysis of achievable clustering as well as order accuracy given assumptions about error and missing data rates

• Develop efficient and accurate, large-scale genetic mapping software
Thank You

Code for BubbleCluster soon available at: www.ucsb.edu/~veronika
Backup Slides
Choosing LOD and non-missing threshold

**Goal: minimize** $P$(mistake) and maximize $F$-score

Let $p = P(LOD(m_i, m_j) > LOD_{threshold} \mid x_i \in C_i, x_j \in C_j, i \neq j)$

Let $\tau$ = LOD threshold

➢ By definition of the LOD score, $p = \frac{1}{10^\tau}$

Let $n_{comp}$ = number of LOD comparisons we make

Then, if we want to ensure that $(1 - p)^{n_{comp}} < 1 - \varepsilon$ then we need:

$$\tau > \log_{10}\left(\frac{1}{1 - (1 - \varepsilon)^{1/n_{comp}}}\right)$$

At the same time, we want to include a marker in the high-quality set only if we expect that it will achieve a LOD of $\tau$ or greater with another marker, requiring:

$$n_{nm} > \frac{\tau}{(1 - \mu) \log_{10} 2}$$

Where $\mu$ is the missing rate and $n_{nm}$ is the number of non-missing entries in the marker
Evaluation Metric for Cluster Quality

F-score (range: 0 to 1)

- Given a “golden standard clustering”, the F-score measures the quality of a clustering as follows:
- The F-score between a golden standard cluster $g$ and a test cluster $c$ is a harmonic mean of precision $P$ and recall $R$:
  \[
  F_{score}(g, c) = \frac{2PR}{P + R}
  \]
- The overall F-score between the golden standard clustering $G$ and a test clustering $C$ is a weighted average of the F-scores for each golden standard cluster $g$:
  \[
  overall_{Fscore}(G, C) = \frac{1}{m} \sum_{g \in G} |g| * \max_{c \in C} F_{score}(g, c)
  \]
Local Linearity Assumption

Although the LOD score does not obey the triangle inequality, we assume that it does at close ranges and with enough data.

\[
\text{LOD}(m, r_2) < \text{LOD}(m, r_1) \\
\text{LOD}(m, r_2) < \text{LOD}(r_1, r_2) \\
\text{LOD}(m, r_1) > \text{LOD}(r_1, r_2)
\]

\( m \) is a new boundary point.
The Effect of the LOD threshold

The standard approach to clustering can be viewed as single linkage clustering.

Time complexity: $O(M^2)$
The Effect of the LOD threshold

A high LOD threshold ensures that the only edges that remain in the completely connected graph are between markers that are extremely likely to be genetically linked.
The Effect of the LOD threshold

Example: LOD threshold = 10
The Effect of the LOD threshold

Example: LOD threshold = 8
The Effect of the LOD threshold

Example: LOD threshold = 7

Linkage group 1

Linkage group 2