MapReduce for accurate error correction of next-generation sequencing data

Assoc. Prof. Liang Zhao

School of Computing and Electronic Information
Guangxi University & Taihe Hospital

Oct 5th, 2016
What is next-generation sequencing (NGS)?
- High throughput, e.g., millions of sequences per run
- Low cost, e.g., $1000 per human genome

What does NGS data look like?
Applications of NGS data analysis

- **De Novo** genome assembly
- Genome re-sequencing
- Genetic variations:
  - Single Nucleotide Polymorphisms (SNPs)
  - Small insertions/deletions (indels)
  - Structural variations
- Linking genetic variations to diseases
  - Genome-wide association studies (GWAS)
  - Functional categorization of SNPs
- RNA-Seq:
  - Gene expression
  - Exon-intron structure

SNP associated trait categories on Human Chromosome 6 by 2014. The figure is obtained from EBI: [http://www.ebi.ac.uk/fgpt/gwas/images/timeseries/gwas-2014-05.png](http://www.ebi.ac.uk/fgpt/gwas/images/timeseries/gwas-2014-05.png)
Errors in NGS data

- **Types of errors:**
  - **Substitution**
    - Error rate of Illumina sequences: 0.5% ~ 2.5%
    - Other platforms: negligible
  - **Insertion/deletion (indel)**
    - Error rate of Roche 454 sequences: 1.5% ~ 5%
    - Error rate of PacBio sequences: 15% ~ 20%
    - Error rate of Oxford Nanopore sequences: 25% ~ 40%
    - Error rate of Illumina sequences: negligible

```
TCTGACTGCAACGGGCAATAT--GTCTCTGT
GGGTCTCTGTTGACTGCAGC
ACGGGCACTA
GAGTGCAACG
TAT--GTCTCTGACT-CAACGGG
GGCAATAT--GTCGTCTG ... 2%3%
2%
3%2%2%
2% 2% 2%
TGCA%
TCTG% CTGA% TGAC% GACT% ACTG% CTGC%
(a)%
(b)%
(c)%
```

(error correction)

(GXU, THH)
**Extra complexity introduced by errors**

- **De novo assembly**: De Bruijn graph-based
  - Branches
  - Bubbles
  - Tips

- **Mapping**:
  - Incorrect place
  - Multiple places
  - Unable to map

- **Variants calling**:
  - False positive
  - SNP occurs in 1/300
Existing error correction approaches

- K-spectrum-based approaches
- Suffix tree-/array-based approaches
- Multiple sequence alignment-based-based approaches
- Cluster-based approaches
- Probabilistic model-based approaches

(a) (b) (c)
K-spectrum-based approach

- **Algorithm briefing**
  - Decompose reads into k-mers;
  - Count the frequencies of k-mers;
  - Substitute the k-mer having low frequency to the nearest high one.

- **Bloom Filter**
- **graph model**

- **Pros & cons:**
  - Pretty fast
  - Good scalability
  - Very sensitive to k
Suffix tree-/array-based approach

- **Algorithm briefing**
  - Construct suffix tree/array for all the reads;
  - Count the frequencies of all the nodes;
  - Substitute the branch having low frequency to its neighbor with high frequency.

- **Pros & cons:**
  - k is flexible
  - Memory consuming
  - Slow

![Suffix tree/array diagram](image-url)
Multiple sequence alignment-based approach

- Algorithm briefing
  - Group reads by k-mers;
  - Perform multiple sequence alignment;
  - Edit reads by using the consensus of the alignment.

- Pros & cons:
  - More accurate
  - Not very sensitive to k
  - Time and space complexity are very high
Existing approaches cannot guarantee the completeness of coverage.

Suppose read length is $l$, per base error rate is $e$, position coverage is $d$, and $k$-mer size is $k$, then the expected number of $k$-mers (the same $k$-mer) cover each position is:

$$d' = d \times \frac{l - k + 1}{l} \times (1 - e)^k$$
Two-layered MapReduce framework

alignment_1

alignment_2

alignment_3

graph_1

graph_2

graph_3
The first layer of MapReduce

reads

(groups)

(k, j, i, l)

k-mer

error correction

(GXU, THH)
The first layer of MapReduce

- **Input:**
  - A set of Paired-end Read \( R \).

- **Goal:**
  - Fishing out prospective erroneous bases from \( R \).

- **Procedures:**
  - Map all the reads of \( R \) into groups:
    - The keys are the \( k \)-mers;
    - The values are the tuples, \((\kappa, j, i)\), representing the \( k \)-mer \( \kappa \) is in read \( r_j \) starting at position \( i \);
    - The tuples having the same \( \kappa \) are assigned to the same group;
  - Perform multiple reads alignment, taking the \( k \)-mer \( \kappa \) as seed.
  - Identify positions from the alignments having inconsistent bases composition.
  - Recombine reads covering the same position that has been identified as erroneous.
The first layer of MapReduce

- Completes the coverage by collecting reads from multiple groups.
- Improves the accuracy markedly.
The second layer of MapReduce

- **Input:**
  - The prospective erroneous positions with covering reads provided.

- **Goal:**
  - Correct erroneous bases of all the reads.

- **Procedure:**
  - Map all the positions to computing units:
    - The key is the position;
    - The value is the reads covering the position.
  - Correct the prospective erroneous bases through the following statistics:
    \[
    L_{x/x_0} = \log \frac{\prod_j I(\hat{j} = x) * p_j + I(\hat{j} \neq x) * (1 - p_j)/3}{\prod_j I(\hat{j} = x_0) * p_j + I(\hat{j} \neq x_0) * (1 - p_j)/3}
    \]
    where \(I(\cdot)\) is a indicator function, \(p_j\) is the probability that the base \(j\) called correctly, \(x_0\) is the base having the largest support, and \(x\) be the prospective erroneous base to be corrected.
## Data sets

- **Data sets used for performance evaluation**

<table>
<thead>
<tr>
<th>Data set</th>
<th>Genome name</th>
<th>Genome size (mbp)</th>
<th>Read length (bp)</th>
<th>Coverage</th>
<th>Number of paired-end reads</th>
<th>Per base error rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1</td>
<td>S. aueus</td>
<td>2.8</td>
<td>101</td>
<td>46.3×</td>
<td>1,294,104</td>
<td>1.17</td>
</tr>
<tr>
<td>R2</td>
<td>R.sphaeroides</td>
<td>4.6</td>
<td>101</td>
<td>33.6×</td>
<td>766,646</td>
<td>1.28</td>
</tr>
<tr>
<td>R3</td>
<td>H.sapiens 14</td>
<td>88.3</td>
<td>101</td>
<td>38.3×</td>
<td>16,757,120</td>
<td>0.86</td>
</tr>
<tr>
<td>R4</td>
<td>B. impatiens</td>
<td>249.2</td>
<td>124</td>
<td>150.8×</td>
<td>303,118,594</td>
<td>0.96</td>
</tr>
<tr>
<td>D1</td>
<td>E.Coli</td>
<td>4.6</td>
<td>101</td>
<td>30.0×</td>
<td>689,927</td>
<td>1.35</td>
</tr>
<tr>
<td>D2</td>
<td>S. cerevisiae</td>
<td>12.4</td>
<td>101</td>
<td>60.2×</td>
<td>3,599,533</td>
<td>1.53</td>
</tr>
<tr>
<td>D3</td>
<td>H.sapiens 22</td>
<td>41.8</td>
<td>101</td>
<td>30.0×</td>
<td>6,209,209</td>
<td>1.47</td>
</tr>
<tr>
<td>D4</td>
<td></td>
<td></td>
<td>150</td>
<td>60.0×</td>
<td>8,361,240</td>
<td>1.59</td>
</tr>
</tbody>
</table>

- R1 and R4 are real data sets.
- D1 to D4 are simulated data sets.
<table>
<thead>
<tr>
<th>data</th>
<th>corrector</th>
<th>reca</th>
<th>prec</th>
<th>gain</th>
<th>pber*</th>
<th>data</th>
<th>corrector</th>
<th>reca</th>
<th>prec</th>
<th>gain</th>
<th>pber*</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MEC</td>
<td>0.893</td>
<td>0.924</td>
<td>0.874</td>
<td>0.103</td>
<td></td>
<td>MEC</td>
<td>0.944</td>
<td>0.963</td>
<td>0.894</td>
<td>0.120</td>
</tr>
<tr>
<td></td>
<td>Coral</td>
<td>0.803</td>
<td>0.858</td>
<td>0.728</td>
<td>0.210</td>
<td></td>
<td>Coral</td>
<td>0.663</td>
<td>0.970</td>
<td>0.642</td>
<td>0.460</td>
</tr>
<tr>
<td></td>
<td>Racer</td>
<td>0.822</td>
<td>0.929</td>
<td>0.760</td>
<td>0.190</td>
<td></td>
<td>Racer</td>
<td>0.921</td>
<td>0.949</td>
<td>0.872</td>
<td>0.150</td>
</tr>
<tr>
<td></td>
<td>BFE,S</td>
<td>0.409</td>
<td>0.650</td>
<td>0.189</td>
<td>0.879</td>
<td></td>
<td>BFE,S</td>
<td>0.722</td>
<td>0.989</td>
<td>0.714</td>
<td>0.340</td>
</tr>
<tr>
<td></td>
<td>SGA</td>
<td>0.817</td>
<td>0.927</td>
<td>0.753</td>
<td>0.196</td>
<td></td>
<td>SGA</td>
<td>0.726</td>
<td>0.995</td>
<td>0.716</td>
<td>0.323</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R2</td>
<td>MEC</td>
<td>0.874</td>
<td>0.936</td>
<td>0.814</td>
<td>0.260</td>
<td></td>
<td>MEC</td>
<td>0.836</td>
<td>0.884</td>
<td>0.746</td>
<td>0.271</td>
</tr>
<tr>
<td></td>
<td>Coral</td>
<td>0.690</td>
<td>0.779</td>
<td>0.495</td>
<td>0.430</td>
<td></td>
<td>Coral</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>Racer</td>
<td>0.797</td>
<td>0.890</td>
<td>0.699</td>
<td>0.230</td>
<td></td>
<td>Racer</td>
<td>0.541</td>
<td>0.703</td>
<td>0.313</td>
<td>0.484</td>
</tr>
<tr>
<td></td>
<td>BFE,S</td>
<td>0.558</td>
<td>0.965</td>
<td>0.538</td>
<td>0.390</td>
<td></td>
<td>BFE,S</td>
<td>0.018</td>
<td>0.003</td>
<td>-0.517</td>
<td>0.862</td>
</tr>
<tr>
<td></td>
<td>SGA</td>
<td>0.641</td>
<td>0.966</td>
<td>0.613</td>
<td>0.319</td>
<td></td>
<td>SGA</td>
<td>0.457</td>
<td>0.636</td>
<td>0.195</td>
<td>0.607</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R3</td>
<td>MEC</td>
<td>0.999</td>
<td>0.996</td>
<td>0.995</td>
<td>0.007</td>
<td></td>
<td>MEC</td>
<td>0.997</td>
<td>0.984</td>
<td>0.983</td>
<td>0.031</td>
</tr>
<tr>
<td></td>
<td>Coral</td>
<td>0.998</td>
<td>0.986</td>
<td>0.984</td>
<td>0.024</td>
<td></td>
<td>Coral</td>
<td>0.995</td>
<td>0.954</td>
<td>0.947</td>
<td>0.081</td>
</tr>
<tr>
<td></td>
<td>Racer</td>
<td>0.998</td>
<td>0.981</td>
<td>0.980</td>
<td>0.032</td>
<td></td>
<td>Racer</td>
<td>0.994</td>
<td>0.957</td>
<td>0.949</td>
<td>0.077</td>
</tr>
<tr>
<td></td>
<td>BFE,S</td>
<td>0.980</td>
<td>0.998</td>
<td>0.979</td>
<td>0.033</td>
<td></td>
<td>BFE,S</td>
<td>0.984</td>
<td>0.997</td>
<td>0.981</td>
<td>0.029</td>
</tr>
<tr>
<td></td>
<td>SGA</td>
<td>0.991</td>
<td>0.999</td>
<td>0.990</td>
<td>0.016</td>
<td></td>
<td>SGA</td>
<td>0.970</td>
<td>0.997</td>
<td>0.968</td>
<td>0.043</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D1</td>
<td>MEC</td>
<td>0.987</td>
<td>0.915</td>
<td>0.896</td>
<td>0.340</td>
<td></td>
<td>MEC</td>
<td>0.996</td>
<td>0.939</td>
<td>0.928</td>
<td>0.281</td>
</tr>
<tr>
<td></td>
<td>Coral</td>
<td>0.973</td>
<td>0.762</td>
<td>0.670</td>
<td>0.510</td>
<td></td>
<td>Coral</td>
<td>0.971</td>
<td>0.783</td>
<td>0.702</td>
<td>0.467</td>
</tr>
<tr>
<td></td>
<td>Racer</td>
<td>0.880</td>
<td>0.466</td>
<td>0.670</td>
<td>1.800</td>
<td></td>
<td>Racer</td>
<td>0.883</td>
<td>0.487</td>
<td>-0.017</td>
<td>1.610</td>
</tr>
<tr>
<td></td>
<td>BFE,S</td>
<td>0.881</td>
<td>0.892</td>
<td>0.774</td>
<td>0.350</td>
<td></td>
<td>BFE,S</td>
<td>0.909</td>
<td>0.897</td>
<td>0.795</td>
<td>0.328</td>
</tr>
<tr>
<td></td>
<td>SGA</td>
<td>0.883</td>
<td>0.907</td>
<td>0.792</td>
<td>0.340</td>
<td></td>
<td>SGA</td>
<td>0.891</td>
<td>0.889</td>
<td>0.819</td>
<td>0.316</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>D2</td>
<td>MEC</td>
<td>0.996</td>
<td>0.939</td>
<td>0.996</td>
<td>0.340</td>
<td></td>
<td>MEC</td>
<td>0.996</td>
<td>0.939</td>
<td>0.928</td>
<td>0.281</td>
</tr>
<tr>
<td></td>
<td>Coral</td>
<td>0.973</td>
<td>0.762</td>
<td>0.670</td>
<td>0.510</td>
<td></td>
<td>Coral</td>
<td>0.971</td>
<td>0.783</td>
<td>0.702</td>
<td>0.467</td>
</tr>
<tr>
<td></td>
<td>Racer</td>
<td>0.880</td>
<td>0.466</td>
<td>0.670</td>
<td>1.800</td>
<td></td>
<td>Racer</td>
<td>0.883</td>
<td>0.487</td>
<td>-0.017</td>
<td>1.610</td>
</tr>
<tr>
<td></td>
<td>BFE,S</td>
<td>0.881</td>
<td>0.892</td>
<td>0.774</td>
<td>0.350</td>
<td></td>
<td>BFE,S</td>
<td>0.909</td>
<td>0.897</td>
<td>0.795</td>
<td>0.328</td>
</tr>
<tr>
<td></td>
<td>SGA</td>
<td>0.883</td>
<td>0.907</td>
<td>0.792</td>
<td>0.340</td>
<td></td>
<td>SGA</td>
<td>0.891</td>
<td>0.889</td>
<td>0.819</td>
<td>0.316</td>
</tr>
</tbody>
</table>

pber* is pber × 10⁻⁴; gain = (TP-FP)/(TP+FN); reca = TP/(TP+FN); prec = TP/(TP+FP).
The experiments are conducted on R3.

MEC is less sensitive to the change of coverage.
High impact of $k$ on $k$-spectrum-based approach

Experiments are carried out on R3.

The size of $k$ has high impact on existing $k$-spectrum-based approaches.
• Experiments are carried out on R3.
• The size of $k$ has low impact on MEC.
Running time comparison

Environment of experiments:
- CPU: 2 six-core Intel Xeon X5690 3.47GHz
- RAM: 96G
Environment of experiments:

- CPU: 2 six-core Intel Xeon X5690 3.47GHz
- RAM: 96G

![RAM usage comparison chart](chart.png)
Concluding remarks

- **MEC** is an accurate approach for correcting NGS substitution errors. It has the following advantages:
  - Markedly better accuracy
    - completeness of coverage
  - Tolerant to various size of \( k \)
    - \( k \)-mers are only used to group reads but not for correcting errors
  - Easy to deploy on cloud computing platform
    - Both identifying prospective erroneous bases and correcting errors can be carried out in parallel.

- **Future directions:**
  - Improve the time and space complexity.


Acknowledgment

- Prof. Limsoon Wong, NUS
- Prof. Jinyan Li, UTS

- National Science Foundation of China (No. 31501070)
- Scientific Research Foundation of GXU (No. XGZ150316)
Good scalability of MEC

![Graph showing time vs. number of nodes with error correction.]
Implementation of coverage completion