Detecting Gene Mutations in Cancer

androwis abumoussa | janice spence phd | john spence phd | richard burack md/phd
Background

sanger sequencing | next generation sequencing
DNA Sequencing

**THE SANGER METHOD:** Single-stranded DNA is mixed with a primer and split into four aliquots, each containing DNA polymerase, four deoxyribonucleotide triphosphates and a replication terminator. Each reaction proceeds until a replication-terminating nucleotide is added. The mixtures are loaded into separate lanes of a gel and electrophoresis is used to separate the DNA fragments. The sequence of the original strand is inferred from the results. (See p. 40 for an illustration of a high-speed DNA sequencer.)
DNA Sequencing

biology
DNA Sequencing

alignment
DNA Sequencing

map reduce

Coverage across reference
SRR022913_bwa_backtrack_sorted.bam

Coverage (X)

GC (%)

Position (bp)

Coverage — GC content(%) — mean GC content
DNA Sequencing | Alignment

- Maximize Similarity
- Bound Search Space
- Make assumptions to do this
Targeted Resequencing

1. Read data
2. Ref Sequences
3. Map reads
4. Align Reads
5. Site frequency data
6. Filter out noise
7. SNV data

8. Read data
9. Ref Sequences
10. Map reads
11. Partition reads using ROAs
12. Form dictionaries & count words
13. Threshold - low frequency words
14. ROA Dictionaries
15. Partial Assembly Verification
16. Verified Dictionaries
17. Filter out noise
18. SNV data

Custom Reference Set
Local word assembly

Iterative refinement
Loop until converged
Cell Pathways
Cell Pathways

cancer

KS

CD83

myc-1

Coverage across reference

Coverage

Reference Position

Coverage

Reference Position

Coverage

Reference Position

Coverage

Reference Position
Next Gen Sequencing

Whole genome
- Predominant applications:
  - Structural variants
  - Point mutations
  - Copy number variation

Whole-exome (1%)
- Predominant applications:
  - Point mutations
  - Copy number variation

PCR amplicon
- Predominant applications:
  - Point mutations
  - Deletions

Transcriptome RNA
- Predominant applications:
  - Gene expression
  - Gene fusions
  - Splice variants

Exon capture transcriptome
- Predominant applications:
  - Gene expression
  - Gene fusions
  - Splice variants
Big Data in Genetics

Parallel Data Structures | Map Reduce | Distributed Computing
Parallel Data Structures

trie | burrow-wheeler’s algorithm
Trie Formation

read: accgcgtaag
read: actgcgttag
read: actgcgtatg
read: accgcgtaag
read: actgcgtatg
read: accgcgtaag
read: accgcgtaag
read: actgcgtatg
read: accgcgtaag
read: ac...
Trie Formation

read : accgcgttgaag

Step 1 : create suffixes
read : accgcgttgaag$
ccgcgttgaag$
cgcgttgaag$
gcgttgaag$
cgttgaag$
tttgaag$
aag$
ag$
g$
$

Step 2 : alphabetize suffixes
read : aag$
accgcgttgaag$
ag$
cgcgttgaag$
ccgcgttgaag$
cgttgaag$
gcgttgaag$
tttgaag$
aag$
ag$
g$
$

Trie Formation

read : accgcgtaag
hash : 1346579028

Step 2 : alphabetize suffixes

read : 0 aag$
     1 accgcgtaag$
     2 ag$
     3 ccgcgtaag$
     4 cgcgtaag$
     5 cgtaag$
     6 gcgtaag$
     7 gtaag$
     8 g$
     9 taag$

Step 3 : build trie
Burrows-Wheeler

Reference: cccgcgttaagaccgcgttaagaccgcgttaaga...
Burrows-Wheeler (actual)

Reference: cccgtaagaccgcgtaagaccgcgtaaga...
Map-Reduce in Genetics

Regions of Analysis
ROA Partitioning

Ref: ...taagaaccgcgtaagaaccgcgtaagaaccgcgtaagaaccgcgtaagaaccgcgtaagaaccgcgtaagaaccgc...

Length: 34 BP
## Dictionary Formation

Ref: \( \ldots \text{aagaaccgcgt} \text{aagaaccgcgt} \text{aagaaccgcgt} \text{aagaaccgcgt} \text{aagaaccgcgt} \text{aagaaccgcgt} \text{aagaaccgcgt} \text{aagaaccgcgt} \ldots \)

<table>
<thead>
<tr>
<th>ROA</th>
</tr>
</thead>
<tbody>
<tr>
<td>aagaaccgcgtaagaaccgcgtaagaaccgcgtaagaaccgcgtaagaaccgcgtaagaaccgcgtaagaaccgcgtaagaaccgcgtaagaaccgcgtaagaaccgcgtaagaaccgcgtaagaaccgcgtaagaaccgcgtaagaaccgcgtaagaaccgcgtaagaaccgcgt</td>
</tr>
</tbody>
</table>
## Dictionary Formation

<table>
<thead>
<tr>
<th>ROA</th>
<th>REF:</th>
<th>Word 1</th>
<th>Word 2</th>
<th>Word 3</th>
<th>Word 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>acgtggttacctgtacgttttgggaccaatgca</td>
<td>8657</td>
<td>1033</td>
<td>427</td>
<td>98</td>
<td>2</td>
</tr>
<tr>
<td>acgtggttacctgtacgttttgggaccaatgca</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>acgtggttacctgtacgtaatgttttgggaccaatgca</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>acgtggttacctgtacgtaatgttttgggaccaatgca</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>acgtggttacctgtacgtaatgttttgggaccaatgca</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>acgtggttacctgtacgtaatgttttgggaccaatgca</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note: The sequence ROA is shown, along with a reference (REF) and counts for different words.
## Dictionary Reconstruction

### Initial dictionary

<table>
<thead>
<tr>
<th></th>
<th>Reference Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>$t = 1$</td>
<td></td>
</tr>
<tr>
<td>Word (68)</td>
<td>Reference (50)</td>
</tr>
<tr>
<td></td>
<td>Word (68)</td>
</tr>
<tr>
<td></td>
<td>Reference (50)</td>
</tr>
<tr>
<td></td>
<td>Word (68)</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>Word (68)</td>
<td>Reference (50)</td>
</tr>
<tr>
<td>Word (68)</td>
<td>Reference (50)</td>
</tr>
<tr>
<td>Reference (50)</td>
<td>Word (68)</td>
</tr>
<tr>
<td>Reference (50)</td>
<td>Reference (50)</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

### $t = n$

<table>
<thead>
<tr>
<th></th>
<th>Reference Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word (68)</td>
<td>Reference (50)</td>
</tr>
<tr>
<td>Word (68)</td>
<td>Reference (50)</td>
</tr>
<tr>
<td>Word (68)</td>
<td>Reference (50)</td>
</tr>
<tr>
<td>Word (68)</td>
<td>Reference (50)</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
<tr>
<td>Word (68)</td>
<td>Reference (50)</td>
</tr>
<tr>
<td>Word (68)</td>
<td>Reference (50)</td>
</tr>
<tr>
<td>Word (68)</td>
<td>Reference (50)</td>
</tr>
<tr>
<td>Word (68)</td>
<td>Reference (50)</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>
Repeat

1. Read data
2. Ref Sequences
3. Map reads
4. Partition reads using ROAs
5. Form dictionaries & count words
6. Threshold - low frequency words
7. ROA Dictionaries
8. Partial Assembly Verification
9. Verified Dictionaries
10. Filter out noise
11. SNV data
12. Custom Reference Set
13. Local word assembly
14. Iterative refinement
15. Loop until converged
Distributed Systems in Genetics
Distributed Systems