GNUMap: Unbiased Probabilistic Mapping of Next-Generation Sequencing Reads

Nathan Clement
Computational Sciences Laboratory
Brigham Young University
Provo, Utah, USA
Next-Generation Sequencing (Solexa/Illumina)
Problem Statement

- Map next-generation sequence reads with variable nucleotide confidence to a model reference genome that may be different from the subject genome.
  - **Speed**
    - Tens of millions of reads to a 3Gbp genome
  - **Accuracy**
    - Mismatches included?
    - Repetitive regions
  - **Visualization**
Workflow
Indexing the genome

- Fast lookup of possible hit locations for the reads
  - Hashing groups locations in the genome that have similar sequence content
    - k-mer hash of exact matches in genome can be used to narrow down possible match locations for reads
  - Sorting genome locations provides for content addressing of genome
- GNUMAP uses indexing of all 10-mers in the genome as seed points for read mapping
Building the Hash Table

Sliding window indexes all locations in the genome

ACTGAACCATACGGGTACTGAACCATGAATGGCACCTATACGAGATACGCCATAC
Alignment

- Given a possible genome match location, determine the quality of the match
- If you call bases in the read
  - Every base gets the same weight in the alignment, no matter what the quality
  - Later bases in the read that have lower quality have equal weight in the alignment with high quality bases at the start of the read
- GNUMap uses a Probabilistic Needleman-Wunsch to align reads found with seed points from the genome hash
Probabilistic Needleman Wunsch

- Uses PWM in calculation of alignment score
- Allows for probabilistic mismatches and gaps
- Greater ability to map reads of variable confidence

<table>
<thead>
<tr>
<th>j</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>PWM A</td>
<td>0.059</td>
<td>0.000</td>
<td>0.172</td>
<td>0.271</td>
<td>0.300</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>0.108</td>
<td>0.320</td>
<td>0.136</td>
<td>0.209</td>
<td>0.330</td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>0.305</td>
<td>0.317</td>
<td>0.317</td>
<td>0.164</td>
<td>0.045</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>0.526</td>
<td>0.578</td>
<td>0.375</td>
<td>0.356</td>
<td>0.325</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>NW</th>
<th>T</th>
<th>T</th>
<th>T</th>
<th>T</th>
<th>T</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>-2</td>
<td>-4</td>
<td>-6</td>
<td>-8</td>
<td>-10</td>
<td></td>
</tr>
<tr>
<td>T</td>
<td>-2</td>
<td>0.052</td>
<td>-1.948</td>
<td>-3.948</td>
<td>-5.948</td>
<td>-7.948</td>
</tr>
<tr>
<td>T</td>
<td>-4</td>
<td>-1.844</td>
<td>0.208</td>
<td>-1.792</td>
<td>-3.792</td>
<td>-5.792</td>
</tr>
<tr>
<td>C</td>
<td>-6</td>
<td>-3.844</td>
<td>-1.792</td>
<td>-0.520</td>
<td>-2.448</td>
<td>-4.448</td>
</tr>
<tr>
<td>A</td>
<td>-8</td>
<td>-5.844</td>
<td>-3.792</td>
<td>-2.374</td>
<td>-0.978</td>
<td>-2.978</td>
</tr>
<tr>
<td>C</td>
<td>-10</td>
<td>-7.844</td>
<td>-5.792</td>
<td>-4.131</td>
<td>-2.774</td>
<td>-1.318</td>
</tr>
</tbody>
</table>
Assignment

• Given a read that has matches to possibly multiple locations in the genome, assign the read to locations where it matches
  ▫ Repeat Masking – Discard reads that match to repeat regions.
    • Half of the human genome contains repeat regions, so you are not able to map to those regions
    • Many regulatory regions are repeated in the genome
  ▫ Map to all locations – Repeat regions will be over-represented since one read will generate multiple hits
  ▫ Pick a random location – Biased if there are small numbers of reads

• GNUMap uses probabilistic mapping to allocate a share of the read to matching locations in the genome according to the quality of the match
Equation for probabilistic mapping

\[ G_{M_j} = \frac{Q_{M_j} n_{M_j} Q_{M_j}}{\sum_{k \neq j} n_{M_k} Q_{M_k}} \]

- Allows for multiple sequences of different matching quality.
- Includes probability of each read coming from any genomic position.
Alignment

Read from sequencer: GGGTACAACCATTAC

Read is added to both repeat regions proportionally to their match quality.
Which Program to Use?

- Many different programs. How do they relate?
  - ELAND (included with Solexa 1G machine)
  - RMAP (Smith et al., BMC Bioinformatics 2008)
  - SOAP (Li et al., Bioinformatics 2008)
  - SeqMap (Jiang et al., Bioinformatics 2008)
  - Slider (Malhis et al., Bioinformatics 2008)
  - Zoom (Lin et al., Bioinformatics 2008)
  - Bowtie (Langmead et al., Genome Biology 2009)
  - ...
Simulation Studies

• Ambiguous reads cause:
  1. Missed (unmapped) regions
  2. Too many mapped regions (noise)
Simulation Studies

![Diagram showing ROC curves for different algorithms and change rates. The x-axis represents Estimated spikes, and the y-axis represents Real spikes. The lines show the performance of gnumap, rmap, seqmap, soap, and novocraft. The legend indicates the change rates: 60% with 10 changes, 60% with 15 changes, and 15 changes maxed.]
Actual Data

- ETS1 binding domain
- Repetitive region
Future Plans

- Removal of adaptor sequences
- Methylation analysis
- Paired-end reads
- SOLiD color space
Acknowledgements

Evan Johnson
Quinn Snell
Mark Clement
Huntsman Cancer Institute