Genome 373: Mapping Short Sequence Reads II

Doug Fowler
The final

Will be in this room on June 6th at 8:30am.

Will be focused on the second half of the course, but will include material from the first half.

You are allowed one 8.5”x11” cheat sheet.

You will be asked to do numerical calculations (e.g., you may want to bring a calculator).
Short Read Mapping Problems

reads

? ?

? ?

? ?

? ?

? ?

genome
Short Read Mapping Problems

Scale

Multiple mapping

Inexact mapping
Hashing Improves Search

• A hash function...

• The hash table...
Hashing Improves Search

• A **hash function** assigns a unique key to each unique data element (DNA sequence in our case)

• The **hash table** is an associative array that describes the relationship between the key and the sequence

<table>
<thead>
<tr>
<th>Key</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Key1</td>
<td>“GCTAGC”</td>
</tr>
<tr>
<td>…</td>
<td>…</td>
</tr>
<tr>
<td>KeyN</td>
<td>“TTTAGC”</td>
</tr>
</tbody>
</table>
Using Hashing to Map Seeds

For seed mapping we:
Using Hashing to Map Seeds

For seed mapping we:

1) Create a hash table for the genome
2) Hash each seed
3) See if the read key exists in the genome hash table (much faster than comparing the sequences directly)
4) We then extend each seed alignment and score them, picking the best one as the correct alignment of the read
Refinement #1: Spaced seeds

- Why use spaced seeds?
- What is the basic concept?
Refinement #1: Spaced seeds

• Split each seed into four “spaced” seeds

• Any read with two (or fewer) mismatches will have at least two seeds that match the reference

• As usual, these hits can be further examined to determine how well they align
Outline

• Using compression to enable alignment

• Taking advantage of paired reads

• Variant calling

• usegalaxy.org registration
What if we could **compress** the genome to make searching it for matching sequence much faster?
Burrows-Wheeler

- The BW algorithm, developed for **data compression**, can do this.
- Hash entire B-W transformed reference genome (only 2 Gb)
Burrows-Wheeler

- The BW algorithm, developed for **data compression**, can do this
  - Hash entire B-W transformed reference genome (only 2 Gb)
  - To align a read, find all single-base matches
The BW algorithm, developed for data compression, can do this:

- Hash entire B-W transformed reference genome (only 2 Gb)
- To align a read, find all single-base matches
- Extend all hits and report perfectly-aligning hits (if any)
- If no match is found, then back up and try a substitution (why?).
Comparison

Spaced seeds
- Requires ~50Gb of memory.
- Runs 30-fold slower.
- Is much simpler to program.

Burrows-Wheeler
- Requires <2Gb of memory.
- Runs 30-fold faster.
- Is much more complicated to program.

Maq

BWA, Bowtie
Short-read analysis software

<table>
<thead>
<tr>
<th>Program</th>
<th>Website</th>
<th>Open source?</th>
<th>Handles ABI color space?</th>
<th>Maximum read length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bowtie</td>
<td><a href="http://bowtie.cbcb.umd.edu">http://bowtie.cbcb.umd.edu</a></td>
<td>Yes</td>
<td>No</td>
<td>None</td>
</tr>
<tr>
<td>BWA</td>
<td><a href="http://maq.sourceforge.net/bwa-man.shtml">http://maq.sourceforge.net/bwa-man.shtml</a></td>
<td>Yes</td>
<td>Yes</td>
<td>None</td>
</tr>
<tr>
<td>Maq</td>
<td><a href="http://maq.sourceforge.net">http://maq.sourceforge.net</a></td>
<td>Yes</td>
<td>Yes</td>
<td>127</td>
</tr>
<tr>
<td>Mosaik</td>
<td><a href="http://bioinformatics.bc.edu/marthlab/Mosaik">http://bioinformatics.bc.edu/marthlab/Mosaik</a></td>
<td>No</td>
<td>Yes</td>
<td>None</td>
</tr>
<tr>
<td>Novoalign</td>
<td><a href="http://www.novocraft.com">http://www.novocraft.com</a></td>
<td>No</td>
<td>No</td>
<td>None</td>
</tr>
<tr>
<td>SOAP2</td>
<td><a href="http://soap.genomics.org.cn">http://soap.genomics.org.cn</a></td>
<td>No</td>
<td>No</td>
<td>60</td>
</tr>
<tr>
<td>ZOOM</td>
<td><a href="http://www.bioinfor.com">http://www.bioinfor.com</a></td>
<td>No</td>
<td>Yes</td>
<td>240</td>
</tr>
</tbody>
</table>
Outline

• Using compression to enable alignment

• Taking advantage of paired reads

• Variant calling

• usegalaxy.org registration
When sequencing, we start with a template.
HTS Facilitates Paired-End Sequencing

We’ve seen how HTS allows us to acquire a short read from the template
HTS Facilitates Paired-End Sequencing

We can also acquire a second paired read from every template!
HTS Facilitates Paired-End Sequencing

How would we do this in an Illumina (bridge PCR) sequencer?
HTS Facilitates Paired-End Sequencing

After the first read, remove the synthesized (non-surface bound) strand and anneal a second primer!
HTS Facilitates Paired-End Sequencing

Paired-read information is extremely useful!

<table>
<thead>
<tr>
<th>Read 1</th>
<th>Read 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATGCTAG...</td>
<td>GCCTGCG...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>TGCGAAAA...</td>
<td>CCCTTTA...</td>
</tr>
</tbody>
</table>
Paired Reads Can Reveal Structural Variation

What do we mean by structural variation?
Paired Reads Can Reveal Structural Variation
Paired Reads Can Reveal Structural Variation
Paired Reads Can Reveal Structural Variation

If we had just single end, short reads, what would we actually observe in each of these situations (focus on the simple deletion)?
Paired Reads Can Reveal Structural Variation

The only informative reads would be ones that spanned the new junction between deleted sequences.
Paired Reads Can Reveal Structural Variation

First, shear genomic DNA to a specific size
Paired Reads Can Reveal Structural Variation

Then generate HTS-compatible fragments and sequence
Paired Reads Can Reveal Structural Variation

Type II restriction enzymes cut 20-30 bases away from their recognition sequence!
Paired Reads Can Reveal Structural Variation

__________ Expected mapping distance

Next, map paired reads to reference genome and determine (mapped) distance between them
Paired Reads Can Reveal Structural Variation

Next, map paired reads to reference genome and determine (mapped) distance between them
Paired Reads Can Reveal Structural Variation

Structural variation can be inferred by deviation from the expected mapping distance.

- Reads that map closer than expected
- Reads that map further than expected
- Expected mapping distance
Paired Reads Can Reveal Structural Variation

**Observation:** Read pairs have a larger than expected distance compared to reference

**Conclusion:**
Paired Reads Can Reveal Structural Variation

**Observation:** Read pairs have a larger than expected distance compared to reference

**Conclusion:** The donor genome must contain an insertion relative to the reference
Paired Reads Can Reveal Structural Variation

**Observation:** Read pairs have a shorter than expected distance compared to reference

**Conclusion:**
Paired Reads Can Reveal Structural Variation

**Observation:** Read pairs have a shorter than expected distance compared to reference

**Conclusion:** The donor genome must contain a deletion relative to the reference
Paired Reads Can Reveal Structural Variation

Observation: The orientation of one of the read pairs is flipped, relative to the reference genome

Conclusion:
Paired Reads Can Reveal Structural Variation

**Observation:** The orientation of one of the read pairs is flipped, relative to the reference genome

**Conclusion:** The donor genome must contain an inversion relative to the reference genome
Paired Reads Can Reveal Structural Variation

Observation: Read pairs originate from different chromosomes
Conclusion:
Paired Reads Can Reveal Structural Variation

Observation: Read pairs originate from different chromosomes
Conclusion: The donor genome has undergone a translocation
Paired Reads Can Reveal Structural Variation

This strategy, DNA-Paired End Tagging (DNA-PET) can resolve many different types of structural variation.
Outline

• Using compression to enable alignment

• Taking advantage of paired reads

• Variant calling

• usegalaxy.org registration
Variation in the human genome

The 1000 Genomes Project found ~15 million single nucleotide polymorphisms

Variation in the human genome

And many insertions/deletions of varying sizes…

Detecting SNPs With HTS Data

A SNP is a single nucleotide change. **Genotyping** individuals involves determining some or all of the SNPs they carry (e.g. for association studies, etc.)
We Can Create a Read Pileup on the Reference

GTGCTGCTCTTTTAGCTCCGAAGACTGCTTCTC
TCACTTCTCTTTTAGCTCCGAAGACTGCTTCTC
CACTTCTCCTTTAGCTCCGAAGACTGCTTCTC
TTCTTCTTTTAGCTCCGAAGACTGCTTCTC
TGCTCCTTTTAGCTCCGAAGACTGCTTCTC
TCTCCTTTTAGCTCCGAAGACTGCTTCTC
TCTCCTTTAGCTCCGAAGACTGCTTCTC
GCTCCTTTAGCTCCGAAGACTTCTC
TCTCCTTTAGCTCCGAAGACTTCTC
TTAGCTCCGAAGACTTCTC
AGCTCCGAAGACTTCTC
GCTCCGAAGACTTCTC
CGCCGAAGCTTCTCACTTTAGTTCCA
TCCGAAGACTTCTCACTTTAGTTCCA
ACCGTGAAGCTTCTCACTTTAGTTCCA
CCGAAGACTTCTCACTTTAGTTCCA
AAACTGCTTCTTCAAGTTCCA
AAGCTTCTTCTCAAGTTCCA
TGCTTCTTCTCAAGTTCCAGTCACTG
CTTCTTCTTCTCAAGTTCCAGTCACTG
TCTCTTAGGCTCACTTTCTCAAGTTCCAGTCACTG
TCTCTTAGGCTCACTTTCTCAAGTTCCAGTCACTG
AGAGCTATATCGAGATCCAGGAGAGCCGAGCTTCTTAGGATCAGTCTCACTGACACATCTCGCTGCAGAGATCCTATAGAGTCCAGATGAGCC
Where is the Variant in this Read Pileup?

Given this pileup, where would you call a SNP?

next-gen sequencing reads

gtctgctcttttagctccgagaactgcttctca
tcacttctcttttagctccgagaactgcttctcag
caacctctcttttagctccgagaactgcttctctagg
tttttcttttagctccgagaactgcttctctcaggtcct
tgctctctttagctccgagaactgcttctcttaggtcct
tctctcttttagctccgagaactgcttctcttaggtcca
tctccttttagctccgagaactgcttctcttagtcca
gctccttttagctccgagaactgcttctctatgtcca
ttccttttagctccgagaactgcttctcttagtcca

ttagctccgagaactgcttctctaggtccaacatct
agctccgagaactgcttctctcaggtccaacatctca
gctccgagaactgcttctcttaggtccaacatctcaac
cgccgagagctgctctcttagtccaacactcaactctct
tccgagaactgcttctctcttaggtccaacatctcaactg
accgtgaactgcttctctcaggtccaacatctcaactg
cgagaaactgcttctctcttagtccaacatctcactgac

dagagcttttcacaggtccaaacatctcaactgacaacaca
dagagcttttcacaggtccaaacatctcaactgacaacaca
tgctttctcaggtccaaacatctcactgacaactctg
ctctcttaggtccaaacatctcactgacaactctgcttc
tctctcaggtccaaacatctcactgacaactctgctgc
tcagagctatagctgatccaggagagccgagctctctaggtccagcatctcaactgacaactctgactgcagagatcctatagagtcagatgagcc
Where is the Variant in this Read Pileup?

Given this pileup, where would you call a SNP?

next-gen sequencing reads

reference genome

AGAGCTATATCGAGATCCCAAGGAGAGGCCGAGCTCTCCTACCTCAGCATCTCACTGACACACCTCGCTGCAGAGATCCTATAGAGTCCAGATGAGCC
Where is the Variant in this Read Pileup?

This one is pretty obvious... but it’s not always that clear

next-gen sequencing reads

reference genome
Where is the Variant in this Read Pileup?

What about this position?

next-gen sequencing reads

reference genome
Where is the Variant in this Read Pileup?

Here, 11/26 reads are C and the rest are T... about half
Where is the Variant in this Read Pileup?

What’s going on here?

next-gen sequencing reads

reference genome
We need to discriminate between heterozygous (expect 50% of reads to be allele 1, 50% allele 2) and homozygous (100% either allele 1 or allele 2)
The Human Genome is Diploid

We need to discriminate between heterozygous (expect 50% of reads to be allele 1, 50% allele 2) and homozygous (100% either allele 1 or allele 2).

Owing to sequencing and sampling errors, etc we rarely get exactly 50%.
In practice, we can consider loci homozygous when >80% of reads are identical.

The key point is we use the frequency of base calls to decide what is true for each position in the genome.
Sequencing error introduces an additional complication... how much can we trust a particular base call in a pileup? Should we treat every base call the same, as we did in the last example?
Fortunately, we are given an error estimate with each base call. This error is determined by the sequencer itself based on a variety of parameters. We use the Phred scale for sequencing error.
Sequencing Error and Variant Calling

This is a trace from a single capillary Sanger sequencer – each peak represents a different sized DNA fragment terminated by a fluorescent nucleotide

In Sanger sequencing, the Phred score is based on the chromatogram
In Sanger sequencing, the Phred score is based on the chromatogram.

Sequencing Error and Variant Calling

Which would you trust more?
Why?

In Sanger sequencing, the Phred score is based on the chromatogram.
Different HTS platforms use different parameters to calculate Phred scores.

Sequencing Error and Variant Calling

High signal, low noise = high score

High signal, high noise = low score

Different HTS platforms use different parameters to calculate Phred scores.
Phred Quality Scores

Let $P$ be the probability of an incorrect base call

$$Q = -10 \times \log_{10}(P)$$

A Phred score ($Q$) is a transformation of the probability of an incorrect base call
A Phred score ($Q$) is a transformation of the probability of an incorrect base call

\[ Q = -10 \times \log_{10}(P) \]

### Phred quality scores are logarithmically linked to error probabilities

<table>
<thead>
<tr>
<th>Phred Quality Score</th>
<th>Probability of incorrect base call</th>
<th>Base call accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1 in 10</td>
<td>90%</td>
</tr>
<tr>
<td>20</td>
<td>1 in 100</td>
<td>99%</td>
</tr>
<tr>
<td>30</td>
<td>1 in 1000</td>
<td>99.9%</td>
</tr>
<tr>
<td>40</td>
<td>1 in 10,000</td>
<td>99.99%</td>
</tr>
<tr>
<td>50</td>
<td>1 in 100,000</td>
<td>99.999%</td>
</tr>
<tr>
<td>60</td>
<td>1 in 1,000,000</td>
<td>99.9999%</td>
</tr>
</tbody>
</table>

A Phred score ($Q$) is a transformation of the probability of an incorrect base call
Phred Quality Scores

Let $P$ be the probability of an incorrect base call

$$Q = -10 \times \log_{10}(P)$$

In practice, high-throughput sequencer error is somewhere between 1:100 and 1:1000 (e.g. many of the short reads contain at least one error)
Let’s say that green average $Q = 40$ and red average $Q = 40$… what call would you make?
Sequencing Error and Variant Calling

Most likely to be heterozygous A/G at that site
Let's say that green average Q = 40 and red average Q = 5… what call would you make?
Sequencing Error and Variant Calling

Most likely to be $G$ at that site
Sequencing Error and Variant Calling

Formally, we can **weight** the information provided by each read at each position by the quality score.
To see how this works, let’s consider a simple example where we have three reads and a haploid genome.
**Sequencing Error and Variant Calling**

Let’s first convert the Q scores into error probabilities

$$Q = -10 \times \log_{10}(P)$$

<table>
<thead>
<tr>
<th>Base call</th>
<th>Q Phred</th>
<th>P(error)</th>
<th>P(not error)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>10</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>A</td>
<td>10</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>G</td>
<td>30</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>
### Sequencing Error and Variant Calling

Let’s also calculate the probability that each base call is not an error.

<table>
<thead>
<tr>
<th>Base call</th>
<th>Q Phred</th>
<th>P(error)</th>
<th>P(not error)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>10</td>
<td>0.1</td>
<td>0.9</td>
</tr>
<tr>
<td>A</td>
<td>10</td>
<td>0.1</td>
<td>0.9</td>
</tr>
<tr>
<td>G</td>
<td>30</td>
<td>0.001</td>
<td>0.999</td>
</tr>
</tbody>
</table>

**Sequencing Error and Variant Calling**

TCCGAGAACTGTGTTCTCTAGGTCCA\textcolor{red}{A}CATCTCACTG
ACCGTGAGACTGTGTTCTC\textcolor{red}{AC}GTCCA\textcolor{red}{A}ATCTCACTG
CGAGACTGTTCTCTAGGTCCA\textcolor{red}{G}CATCTCACTGAC
AGAGCTATATGAGATCCAGGAGGCGGAGCTCTCTAGGTCACTCATCCTACCTGAGATCACTCGCTGAGATCAGATGAGTCAGATGAGCC
Next, let’s compute the probability that the true genotype is an $A$ and the $G$ is an error – what should this calculation look like?
Next, let’s compute the probability that the true genotype is an A and the G is an error – what should this calculation look like?

\[ p(A) = 0.9 \times 0.9 \times 0.001 = 0.00081 \]
Sequencing Error and Variant Calling

TCCGAGAACTGCTTCTCTAGGTTCCAATCTCACTG
ACCGTGAACTGCTTCTCCAGGTCCAATCTCACTG
CGAGAACTGCTTCTCTAGGTTCACATCTCACTGAC
AGAGCTATATCGAGATCCAGGAGACCGAGCTCTCTAGGTTCCAAGCATCTCACTGACACACTCGCTGCAGAGATCCTATAGAGTCAGATGAGTCAGATGAGCC

<table>
<thead>
<tr>
<th>Base call</th>
<th>Q Phred</th>
<th>P(error)</th>
<th>P(not error)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>10</td>
<td>0.1</td>
<td>0.9</td>
</tr>
<tr>
<td>A</td>
<td>10</td>
<td>0.1</td>
<td>0.9</td>
</tr>
<tr>
<td>G</td>
<td>30</td>
<td>0.001</td>
<td>0.999</td>
</tr>
</tbody>
</table>

\[ p(A) = 0.9 \times 0.9 \times 0.001 = 0.00081 \]

How about the probability that the true genotype is an G and the A’s are errors?
Sequencing Error and Variant Calling

<table>
<thead>
<tr>
<th>Base call</th>
<th>Q Phred</th>
<th>P(error)</th>
<th>P(not error)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>10</td>
<td>0.1</td>
<td>0.9</td>
</tr>
<tr>
<td>A</td>
<td>10</td>
<td>0.1</td>
<td>0.9</td>
</tr>
<tr>
<td>G</td>
<td>30</td>
<td>0.001</td>
<td>0.999</td>
</tr>
</tbody>
</table>

\[ p(A) = 0.9 \times 0.9 \times 0.001 = 0.00081 \]
\[ p(G) = 0.1 \times 0.1 \times 0.999 = 0.00999 \]

So, what call should we make?
Sequencing Error and Variant Calling

<table>
<thead>
<tr>
<th>Base call</th>
<th>Q Phred</th>
<th>P(error)</th>
<th>P(not error)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>10</td>
<td>0.1</td>
<td>0.9</td>
</tr>
<tr>
<td>A</td>
<td>10</td>
<td>0.1</td>
<td>0.9</td>
</tr>
<tr>
<td>G</td>
<td>30</td>
<td>0.001</td>
<td>0.999</td>
</tr>
</tbody>
</table>

\[ p(A) = 0.9 \times 0.9 \times 0.001 = 0.00081 \]
\[ p(G) = 0.1 \times 0.1 \times 0.999 = 0.00999 \]

So, what call should we make? G
Outline

• Using compression to enable alignment

• Taking advantage of paired reads

• Variant calling

• usegalaxy.org registration
What is Galaxy?

Galaxy is a free, open source web platform for running all sorts of computational analyses including pretty much all of the sequencing-related stuff we’ve discussed!

You can use it at usegalaxy.org
Go to User -> Register and register, if you haven’t already… if you have, login
References

• High throughput sequencing

• Short read mapping
  Li, Ruan and Durbin, “Mapping short DNA sequencing reads and calling variants using mapping quality scores.” Genome Research, 2008


• Variant calling
  Li, Ruan and Durbin, “Mapping short DNA sequencing reads and calling variants using mapping quality scores.” Genome Research, 2008

  Li “A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data.” Bioinformatics 2011