What affects alignment?

- Repeat structure of the reference (sequenceability)
- Read base quality
  - Low quality ⇒ fewer reads aligned
  - Low quality ⇒ creating false hits (mainly in repeats)
- Sensitivity of alignment algorithm
  - Missing true hits (also related to repeats)
- Mate pair information
Achieving alignment reliability

- Best unique hit (or 1-diff)
  ✓ A simple improvement: discard an alignment if there are many 1-mismatch-away hits.

- 2-diff (the second best hit is at least 2-mismatch away)
  ✓ Maybe requiring to see 3- or 4-mismatch hits

- Predefine regions where alignments are reliable

- Regarding alignment as a stochastic procedure
  ✓ Mapping reads to the most probable position
  ✓ Phred-scaled prob. of the alignment being wrong
Using mate-pair information

- How mate pairs help?
  - Increase the mappability of the reference
  - Increase the reliability of alignment
  - Find short indels

- Complication to alignment
  - Mapping a pair simultaneously; Otherwise, we lose the ability to recover short repeats
  - If algorithm indexes the genome: joint mapping is relatively easy
  - If algorithm indexes reads: sliding window
  - Complicated to work with predefined unique regions
Alignment accuracy (simulation)

<table>
<thead>
<tr>
<th>method</th>
<th># reads aligned</th>
<th>error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Best unique hit</td>
<td>1,686,129</td>
<td>0.439%</td>
</tr>
<tr>
<td>2-diff</td>
<td>1,476,373</td>
<td>0.002%</td>
</tr>
<tr>
<td>SE MapQ&gt;=10</td>
<td>1,665,959</td>
<td>0.079%</td>
</tr>
<tr>
<td>SE MapQ&gt;=40</td>
<td>1,461,179</td>
<td>0.002%</td>
</tr>
<tr>
<td>PE MapQ&gt;=10</td>
<td>1,756,368</td>
<td>0.016%</td>
</tr>
<tr>
<td>PE MapQ&gt;=40</td>
<td>1,671,328</td>
<td>0.002%</td>
</tr>
</tbody>
</table>
Miscellaneous issues

- What accuracy do we need?
  - SNP calling: possible to combine mapping accuracy to the model
  - Structural variation: no 2-mismatch-way hits for reliable calls

- Implementation:
  - Memory: less than 1GB memory per process is ideal for parallelization (multi-threading helps)
  - File size (seq+qual+read_name+pos) and indexing

- Discussion topics:
  - Reference bias
  - non-independent of wrong read alignments