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)

method	# reads aligned	error rate
Best unique hit	1,686,129	0.439%
2-diff	1,476,373	0.002%
SE MapQ>=10	1,665,959	0.079%
SE MapQ>=40	1,461,179	0.002%
PE MapQ>=10	1,756,368	0.016%
PE MapQ>=40	1,671,328	0.002%

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