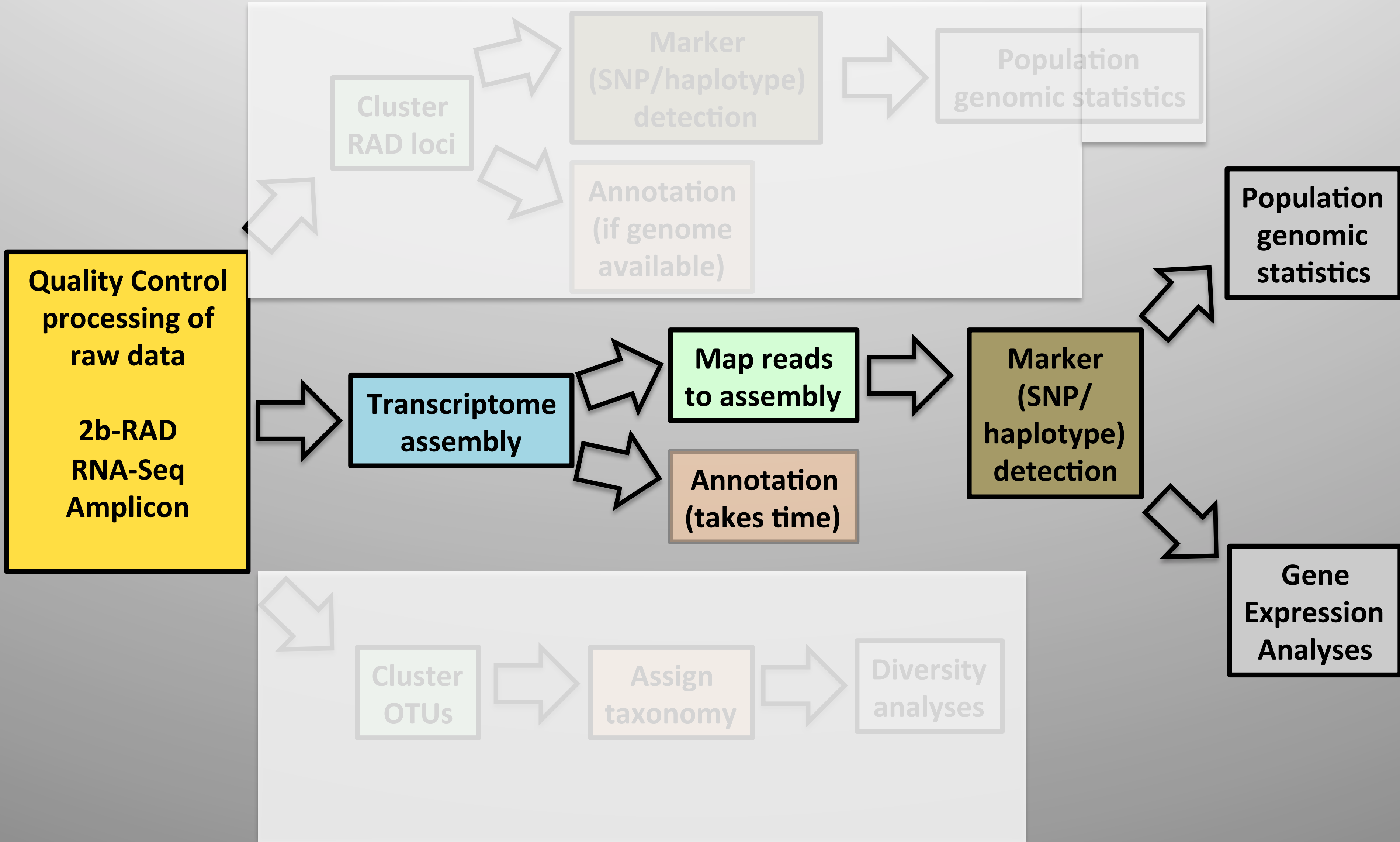
Alignment –
software, algorithms and parameters

Analysis pipeline for RNA-Seq data



Cluster RAD loci

Marker (SNP/haplotype) detection

Population genomic statistics

Annotation (if genome available)

Quality Control processing of raw data

2b-RAD
RNA-Seq
Amplicon

Transcriptome assembly

Map reads to assembly

Annotation (takes time)

Marker (SNP/ haplotype) detection

Population genomic statistics

Gene Expression Analyses

Cluster OTUs

Assign taxonomy

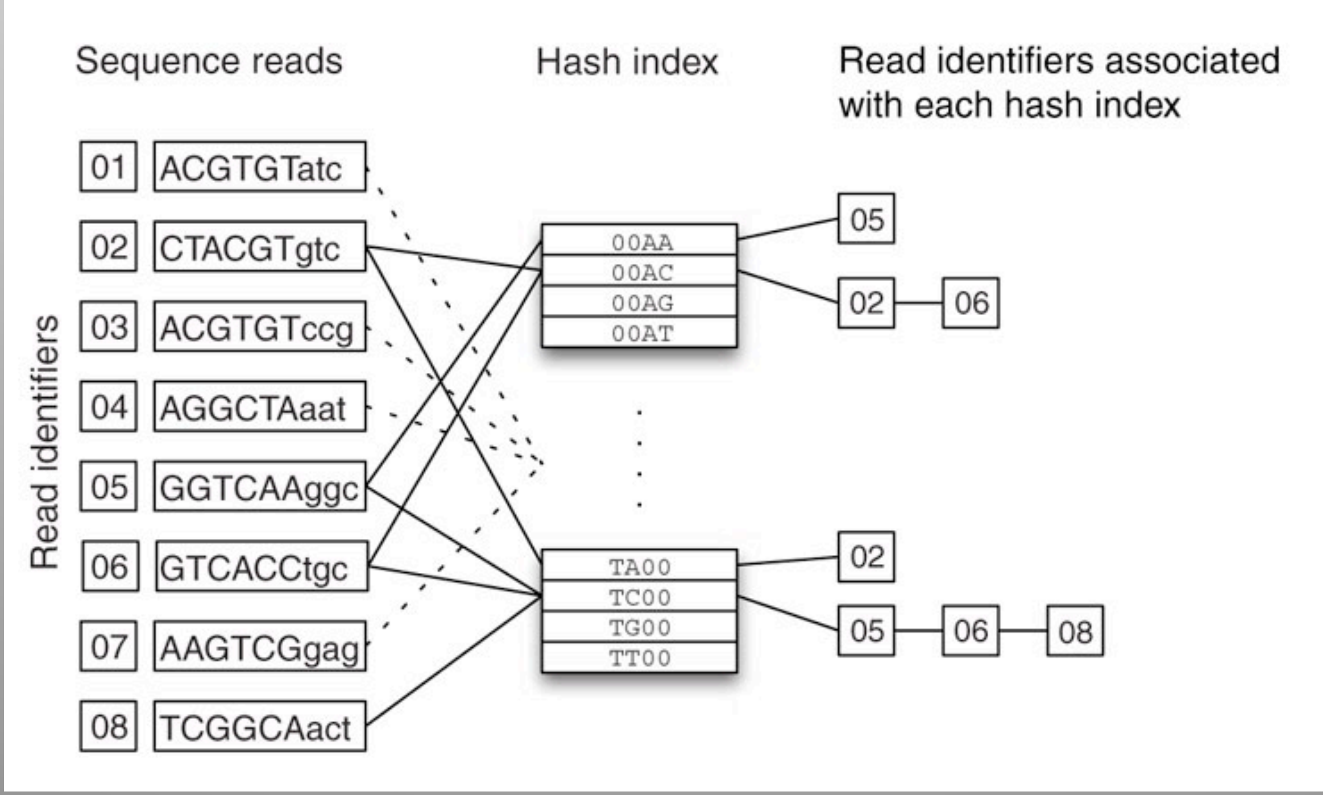
Diversity analyses

Alignment algorithms

- Trying to match full-length sequences everywhere in a genome is unfeasible.
- First step is to weed down possible alignment locations for each read.
- There are 2 different methods of this indexing, hash-based or Burrows-Wheeler transform based.

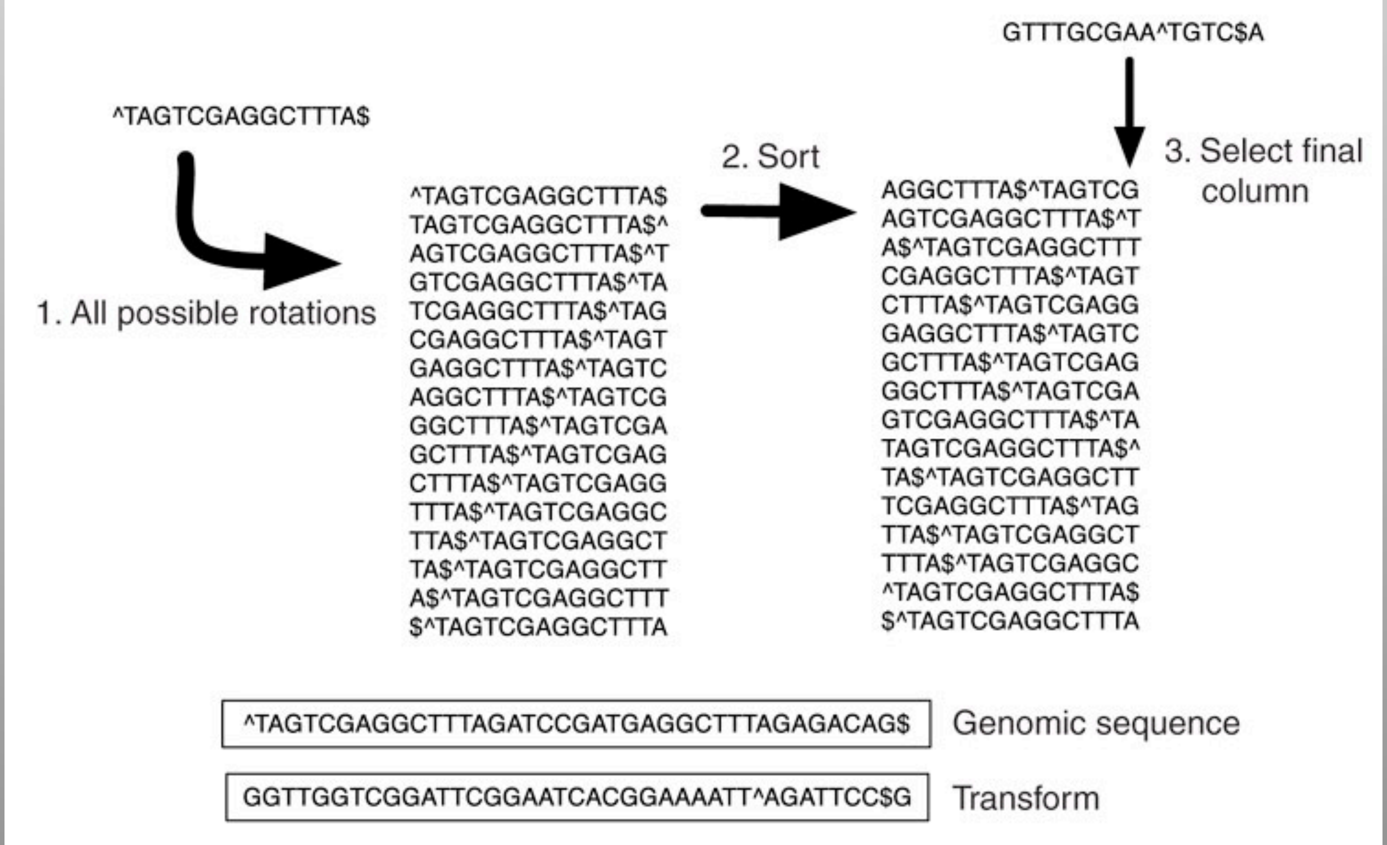
Alignment software

- Hash-based
  - MAQ, SOAP, ELAND, SHRiMP, ZOOM, BFAST, MOSAIK



Alignment software

- Burrows-Wheeler Transform based
  - BOWTIE, BWA, SOAP2



Burrows-Wheeler aligner

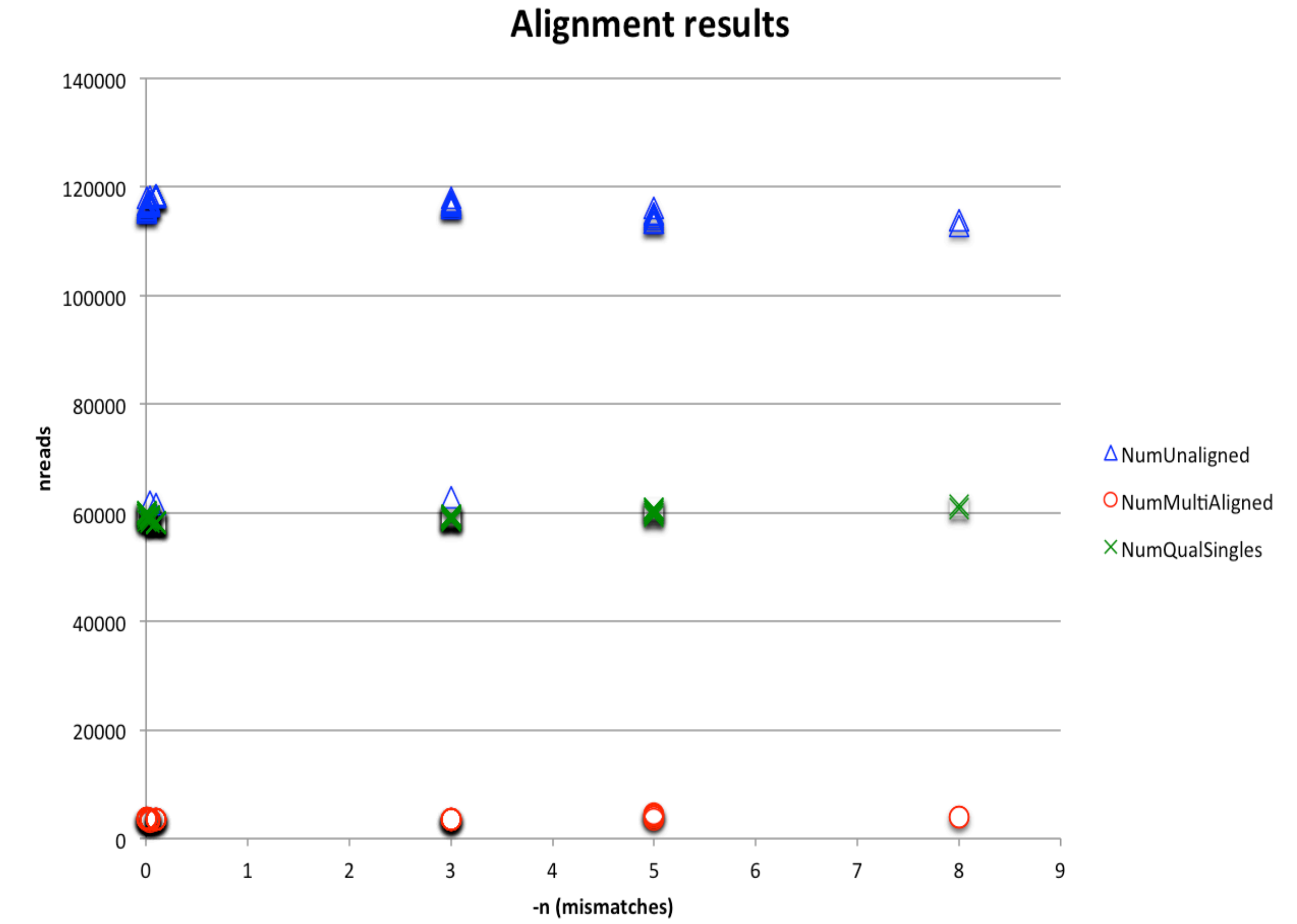
http://bio-bwa.sourceforge.net/bwa.shtml

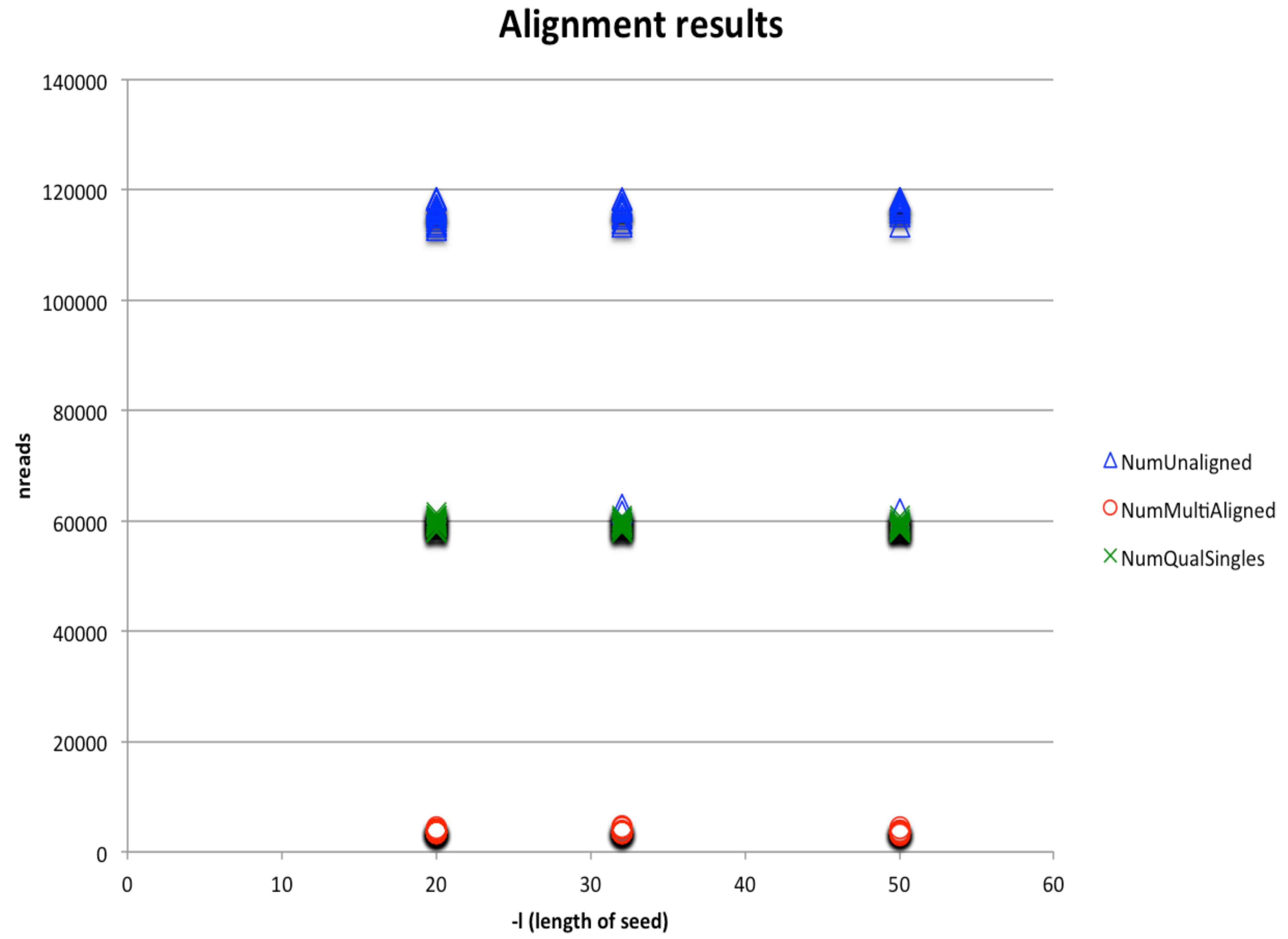
- Aligns to a reference using BWT
- (Some of the) important parameters that WILL affect the result:
  - -n : maximum edit distance if INT,
    maximum fraction of missing alignment if FLOAT [0.04]
  - -l : take the first INT subsequence as seed [32]
  - -k : maximum edit distance in seed [2]
  - -o : maximum number of gap opens [1]
- The combination of these (and more) will affect the number of reads aligning uniquely to one location in the reference, but also the number aligning to multiple locations, as well as the alignment **speed**.

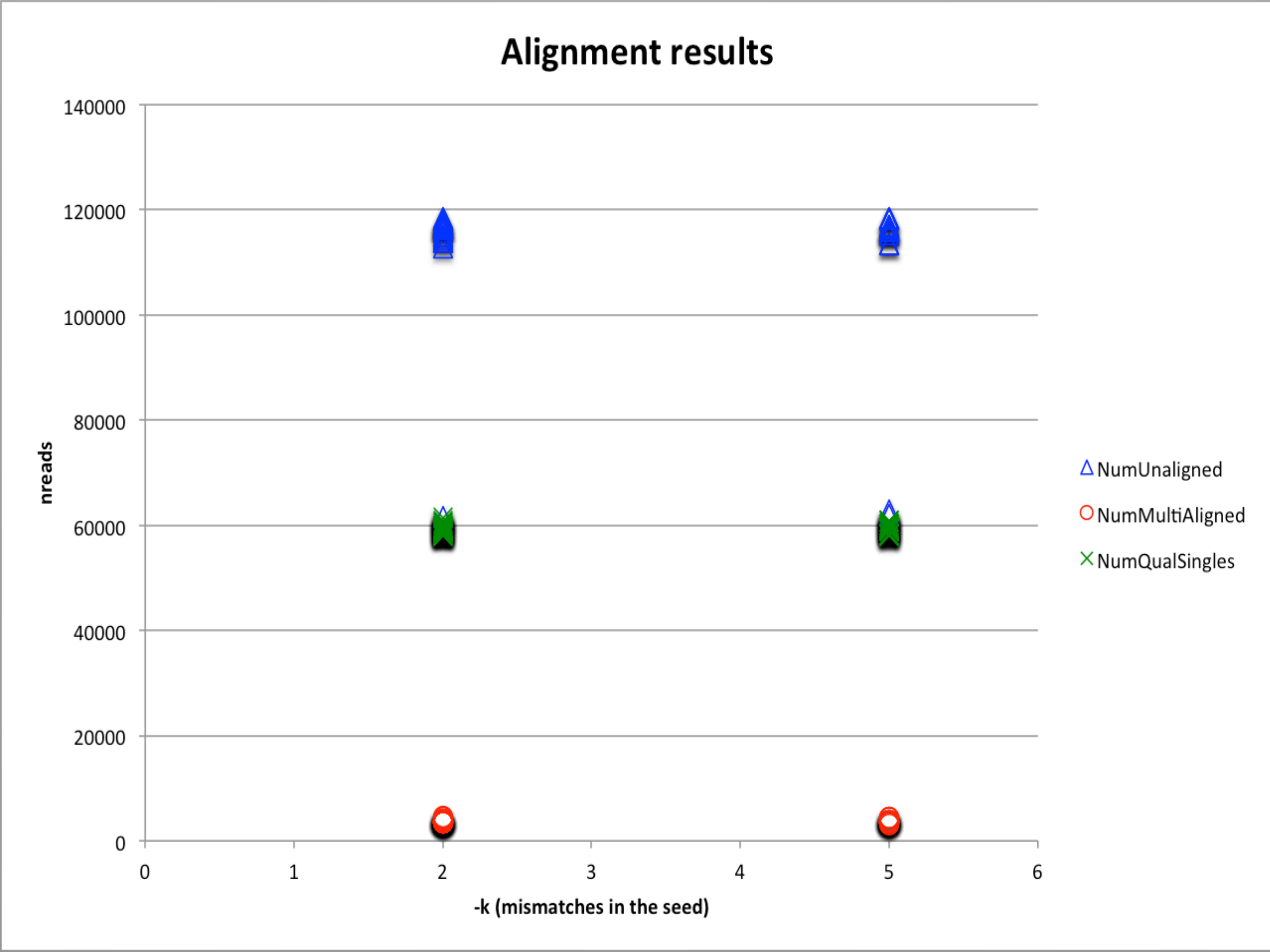
Experiment

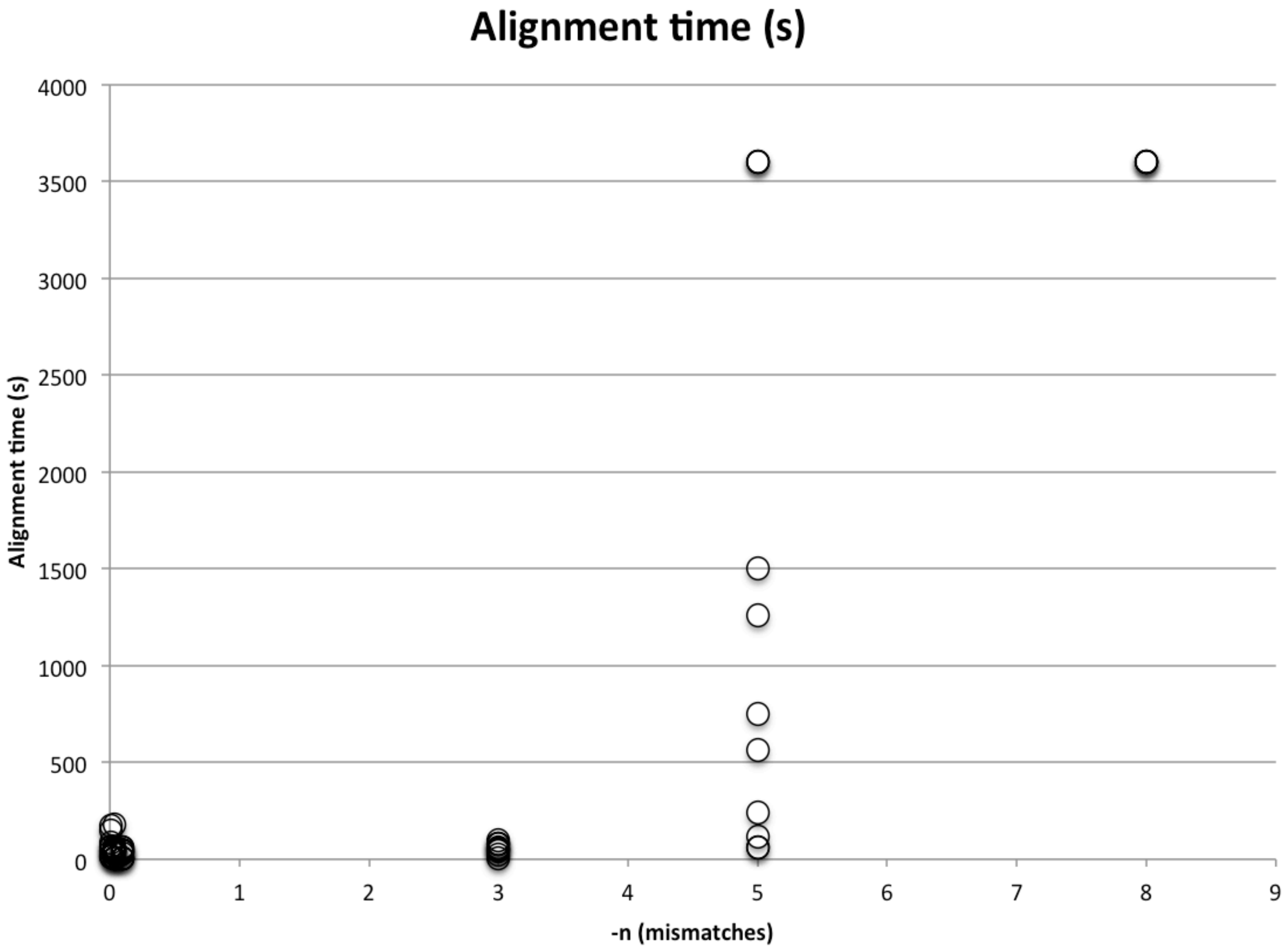
At a course at Cornell 2013, 85 students were each given identical datasets, and were instructed to align with different parameters.

The results provide a little glimpse into which parameters affect alignment the most.

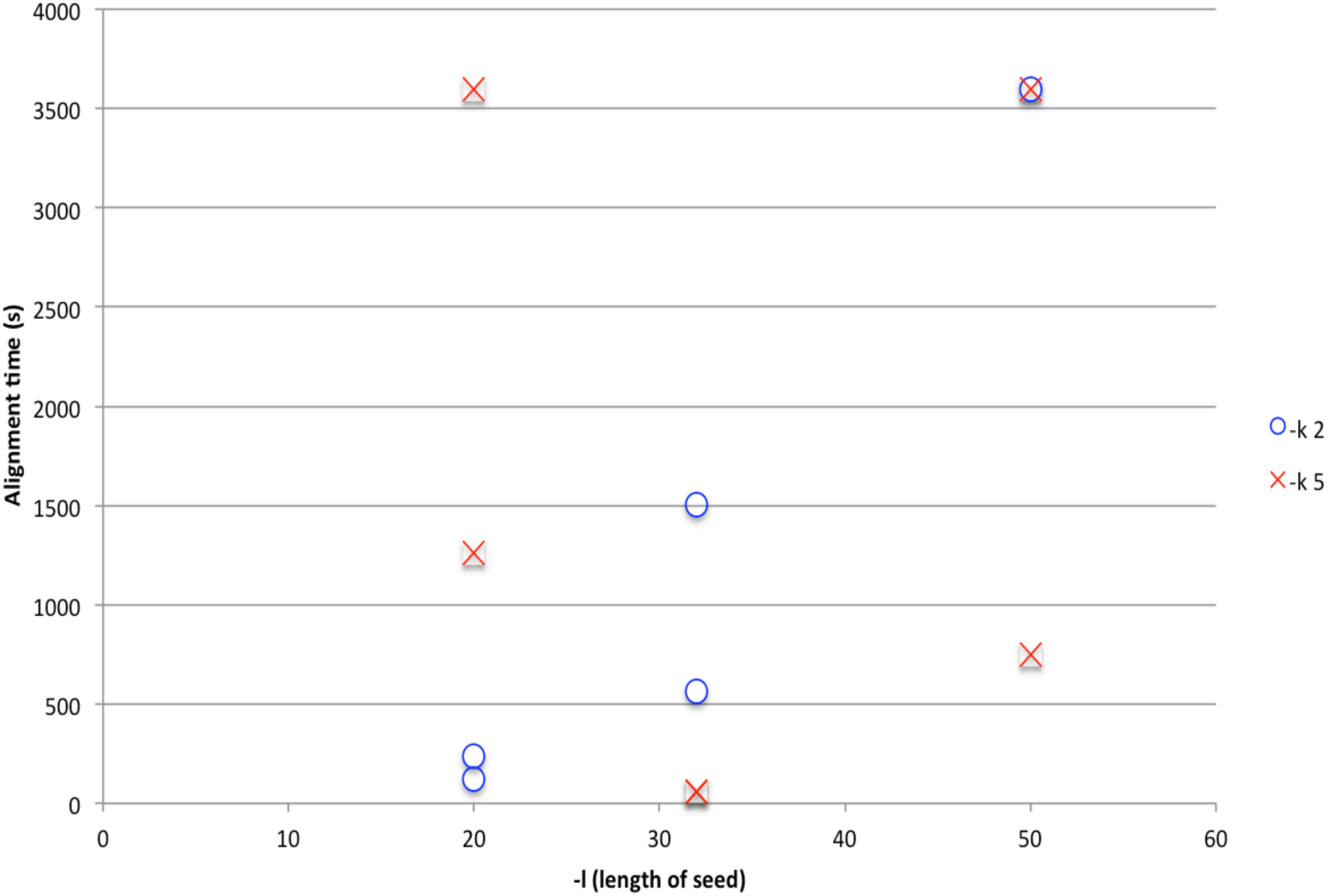

Alignment results
nreads
140000
120000
100000
80000
60000
40000
20000
0
0
1
2
3
4
5
6
7
8
9
-n (mismatches)
△ NumUnaligned
○ NumMultiAligned
× NumQualSingles


Alignment results
nreads
140000
120000
100000
80000
60000
40000
20000
0
0
10
20
30
40
50
60
-l (length of seed)
NumUnaligned
NumMultiAligned
NumQualSingles


Alignment results
nreads
140000
120000
100000
80000
60000
40000
20000
0
0
1
2
3
4
5
6
-k (mismatches in the seed)
△ NumUnaligned
○ NumMultiAligned
× NumQualSingles


Alignment time (s)
4000
3500
3000
2500
2000
1500
1000
500
0
Alignment time (s)
0
1
2
3
4
5
6
7
8
9
-n (mismatches)

Alignment times at -n 5



The sam/bam file format

- There are many alignment programs out there, most output the same format: SAM (Sequence Alignment/Map).
- Tab-delimited text file, contains one line of information per read
- Read name, alignment position, mapping quality (phred scale), read sequence and quality scores
- The last few columns are reserved for user-defined information, such as sample name, treatment or location.
- More info on the format here: http://**sam**tools.sourceforge.net/**SAM**1.pdf

Example:

```
HWI-ST141_0365:2:1101:2983:2114#TTAGGC        16       contig16104    1423    37       42M     *
GAGAATGTGAAGGCCAAGATCCAAGACAAGGAAGGGATTCCCCCAGACC       :?@DD;;B===CAA3CGCA;A?AFB3C3F+:C)::C)?D>:9B?@@?BDD
RG:Z:FR51       XT:A:U NM:i:0 X0:i:1   X1:i:0   XM:i:0  XO:i:0  XG:i:0   MD:Z:42
```

The sam/bam file format

- sam files are LARGE, since each read from the original data is represented as a a line, along with additional information.

- By converting these to binary file format, (zeroes and ones) we can compress these file by a factor of 4.

- This also allows the computer to search through the files faster.

- However, it does make them impossible to read by humans, unless converted back to text format first.

- These binary versions of sam files are called **bam** files.