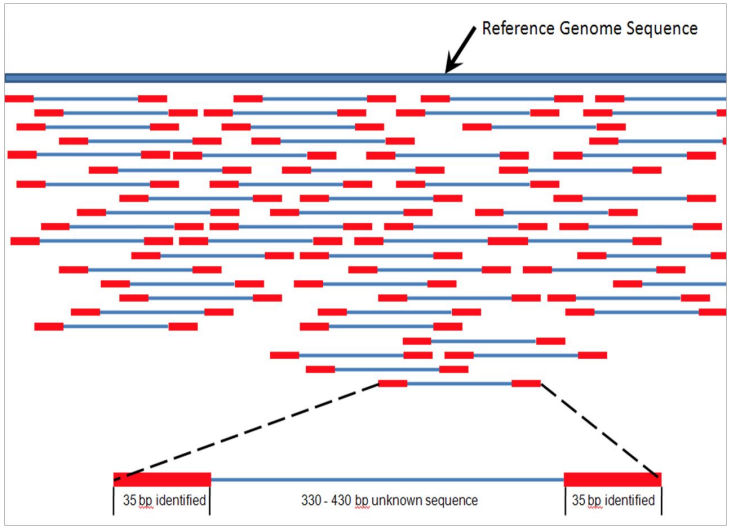
Modeling library complexity in DNA sequencing experiments with Julia

Mukarram Tahir
Larson Hogstrom
Andres Hasfura

MIT 18.337
Fall 2015



SAM (BAM) file format

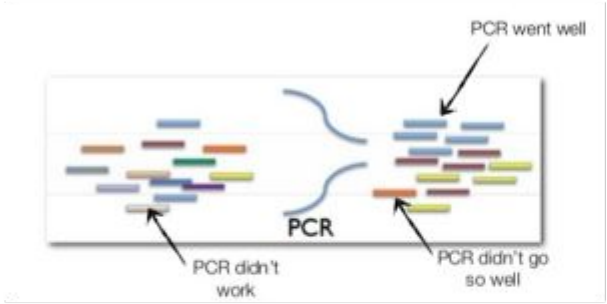
Headers

```
@HD VN:1.5 SO:coordinate
@SQ SN:ref LN:45
r001    99 ref  7 30 8M2I4M1D3M = 37  39 TTAGATAAAGGATACTG *
r002     0 ref  9 30 3S6M1P1I4M *  0   0 AAAAGATAAGGATA    *
r003     0 ref  9 30 5S6M       *  0   0 GCCTAAGCTAA       * SA:Z:ref,29,-,6H5M,17,0;
r004     0 ref 16 30 6M14N5M    *  0   0 ATAGCTTCAGC       *
r003 2064 ref 29 17 6H5M       *  0   0 TAGGC             * SA:Z:ref,9,+,5S6M,30,1;
r001   147 ref 37 30 9M         =  7 -39 CAGCGGCAT         * NM:i:1
```

Alignment data (QNAME, FLAG, RNAME, POS, MAPQ, CIGAR, RNEXT, PNEXT, TLEN, SEQ, Q)

Duplication events arise in DNA sequencing

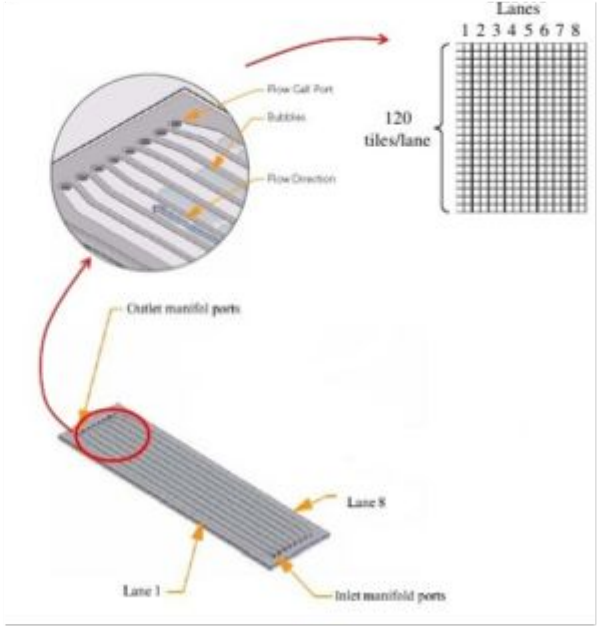
PCR duplicates



Optical duplicates

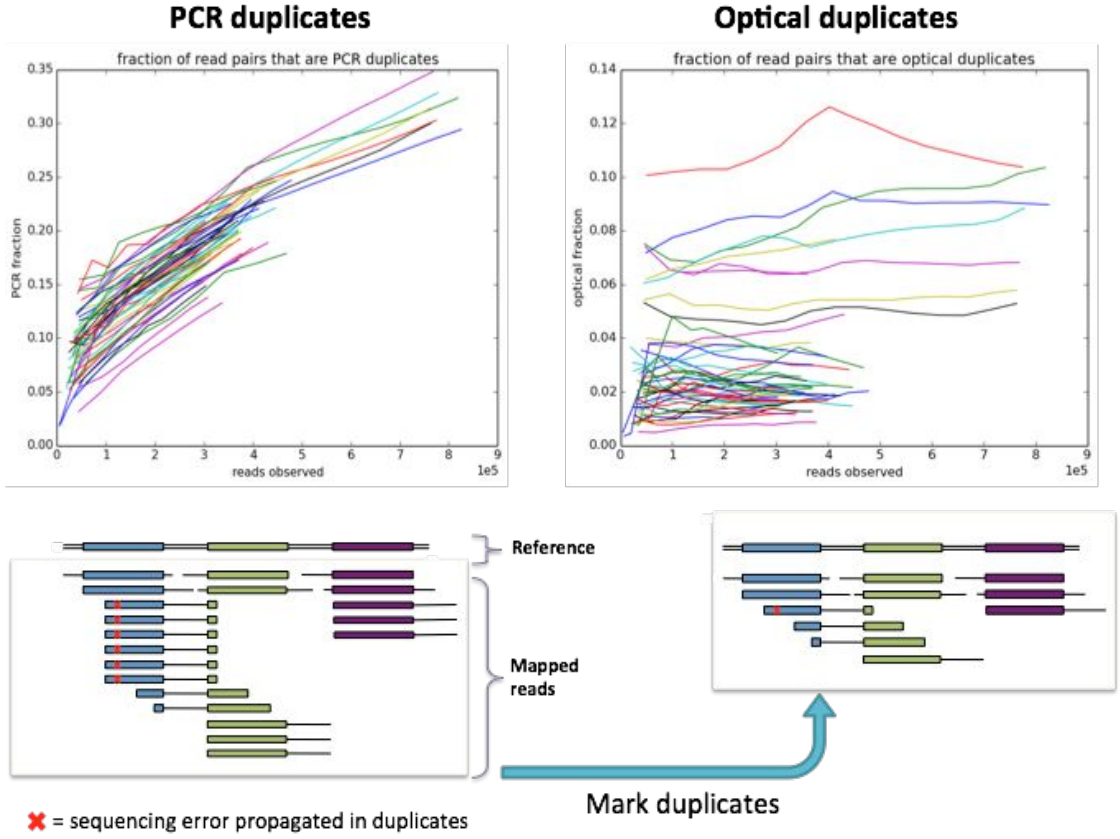
Read names have the following form:

```
@identifier:lane:tile:x:y
```

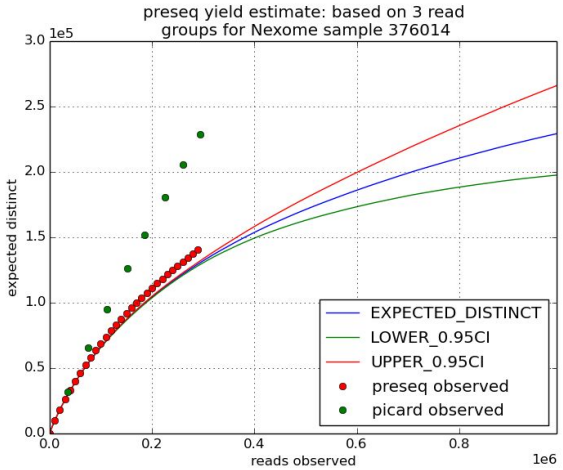


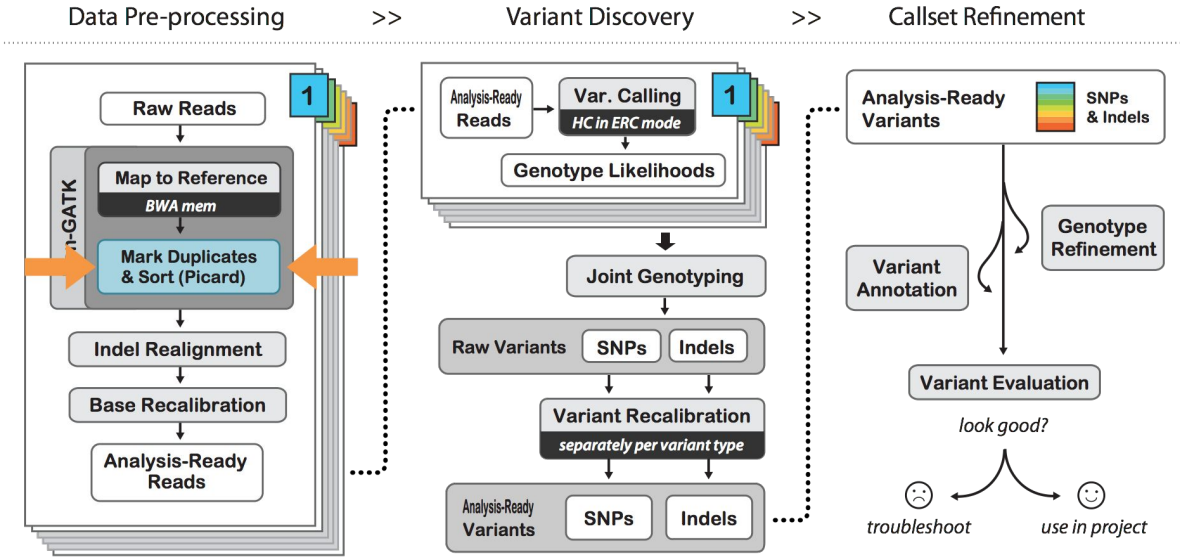
http://www.slideshare.net/randyt/next-generation-sequencing-course-oct-2-seq-error-mapping
http://www.slideshare.net/invitrue/illumina-guide-to-high-throughput-sequencing

Optical and PCR duplication events arise at different rates as a sequencing experiment proceeds



Complexity curve calculations through marking read duplicates

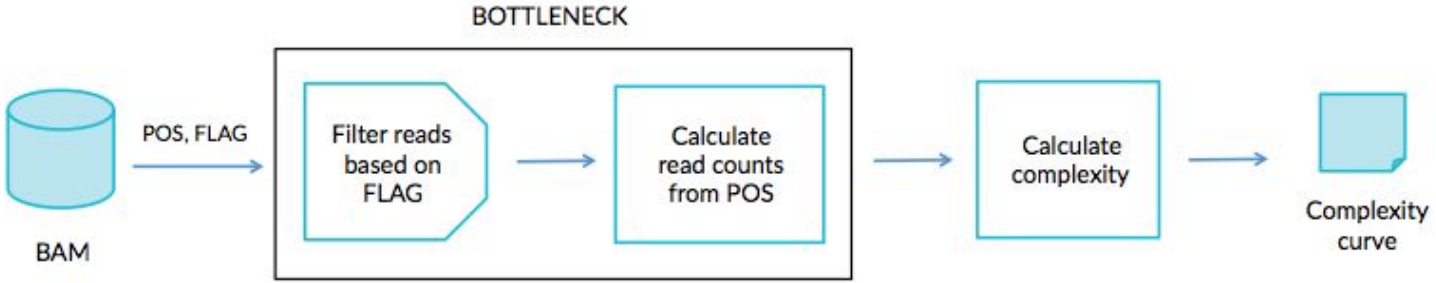




preseq.cpp

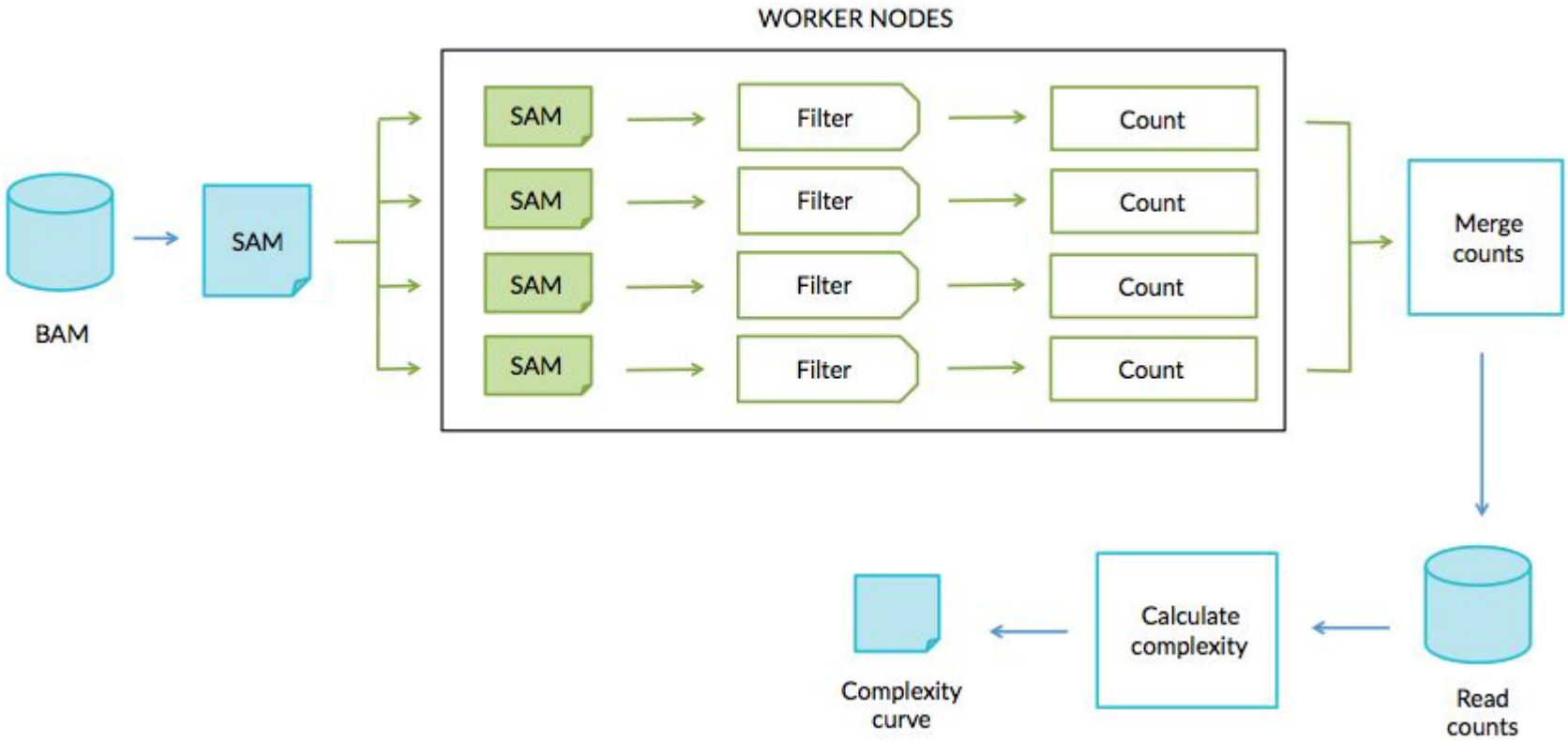
Complexity curve is constructed from BAM file using `preseq`

Code originally written in C++ by Smith lab at USC - directly interfaces with `samtools` for reading BAM files

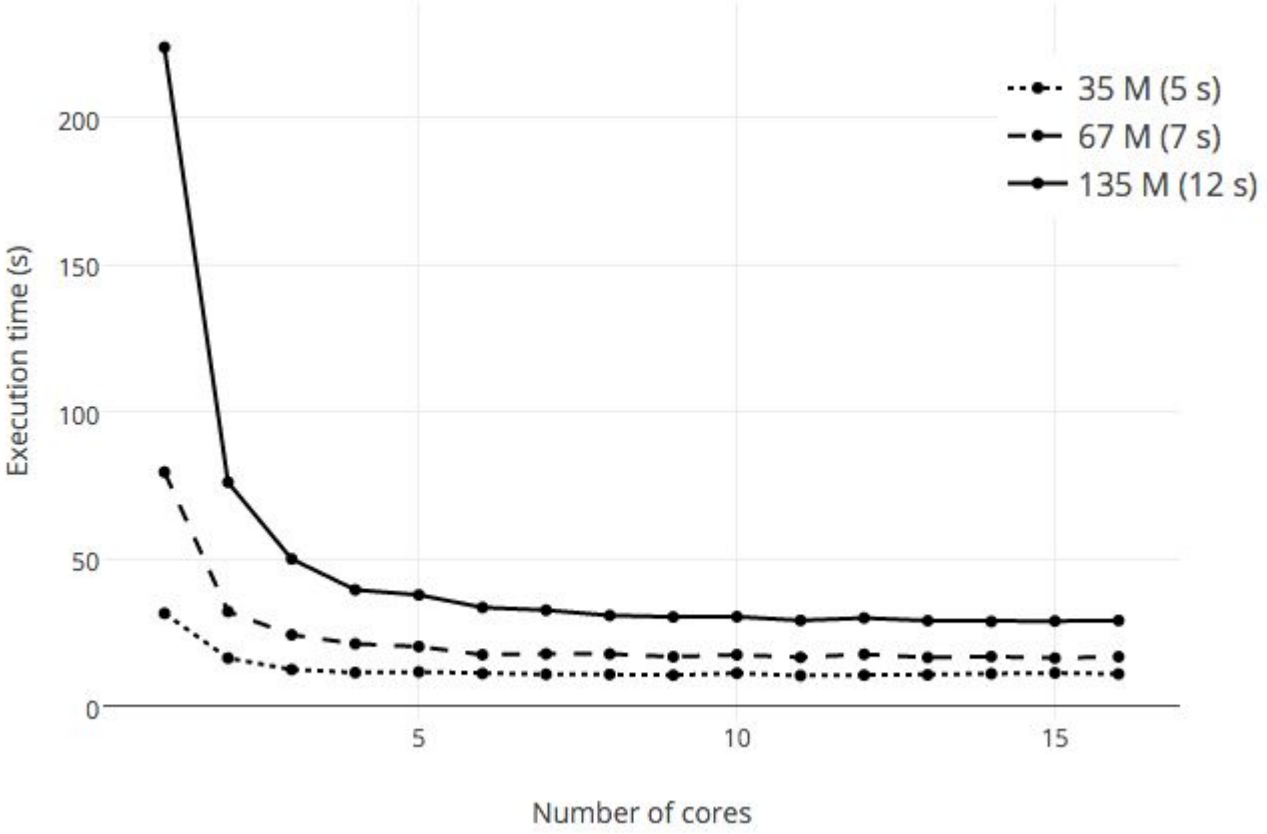


Daley, Timothy, and Andrew D. Smith. "Predicting the molecular complexity of sequencing libraries." Nature methods 10.4 (2013): 325-327.

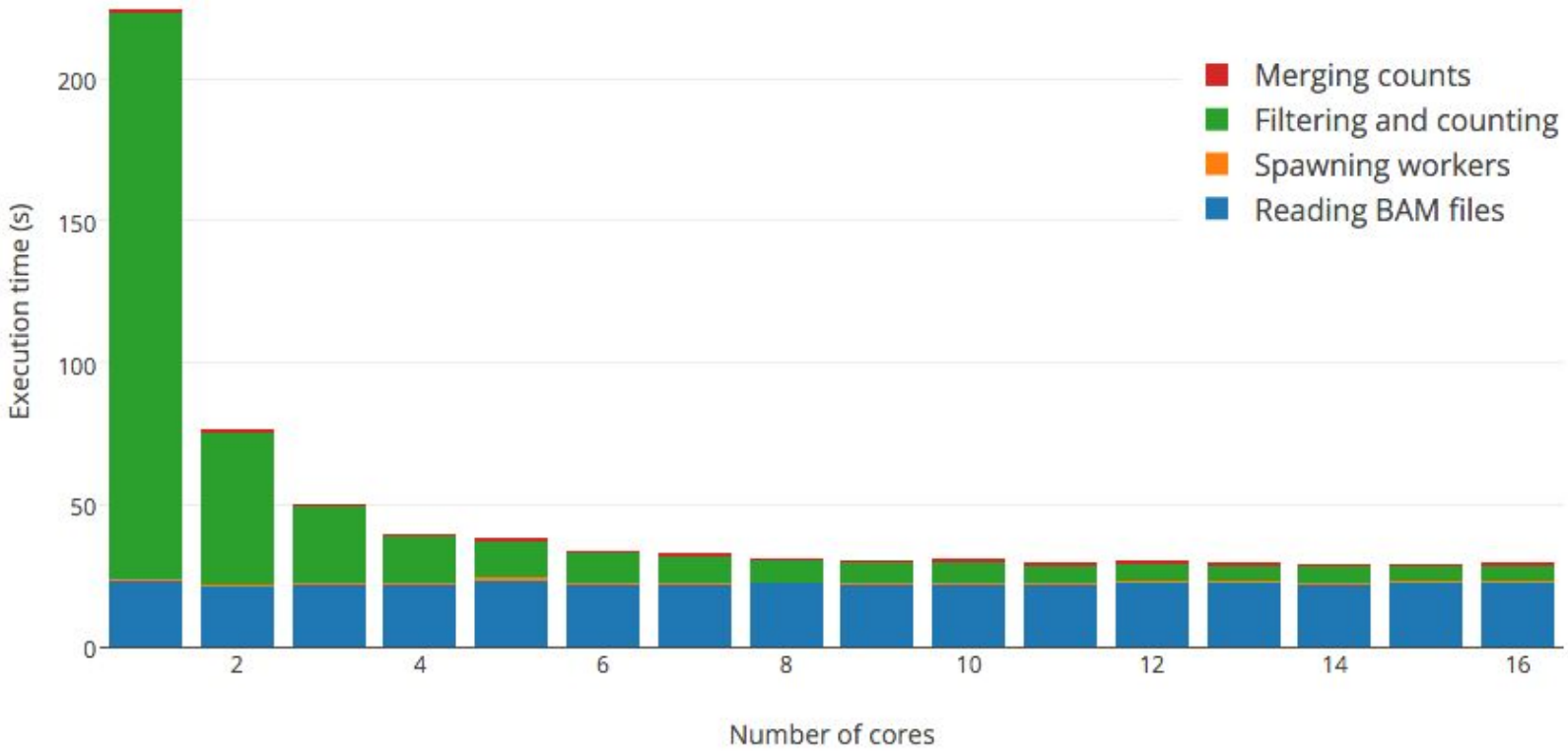
preseq.jl


WORKER NODES
BAM
SAM
SAM
Filter
Count
SAM
Filter
Count
SAM
Filter
Count
SAM
Filter
Count
Merge counts
Read counts
Calculate complexity
Complexity curve

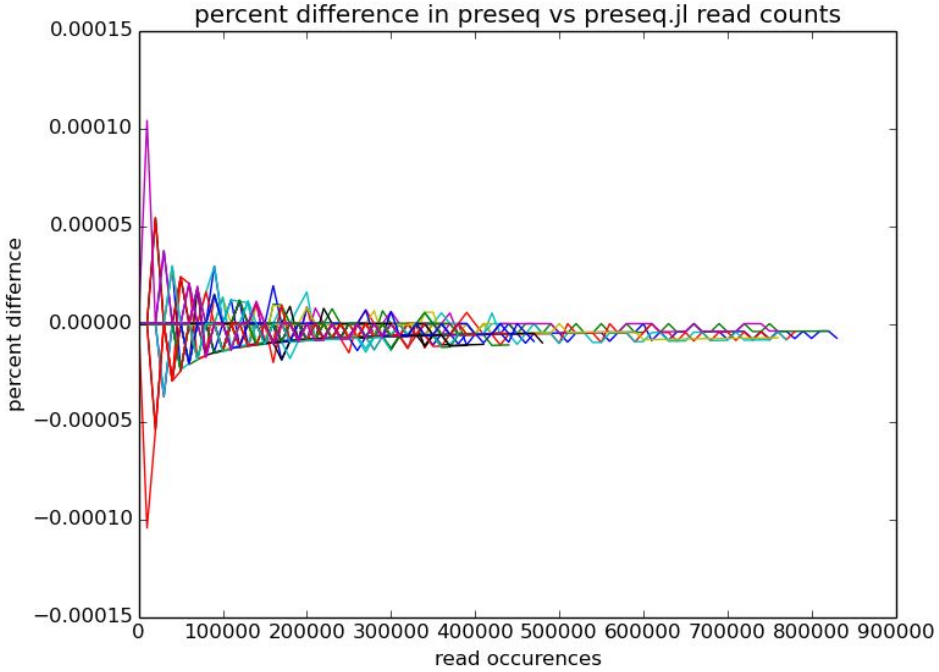
Speedup curves for various BAM file sizes

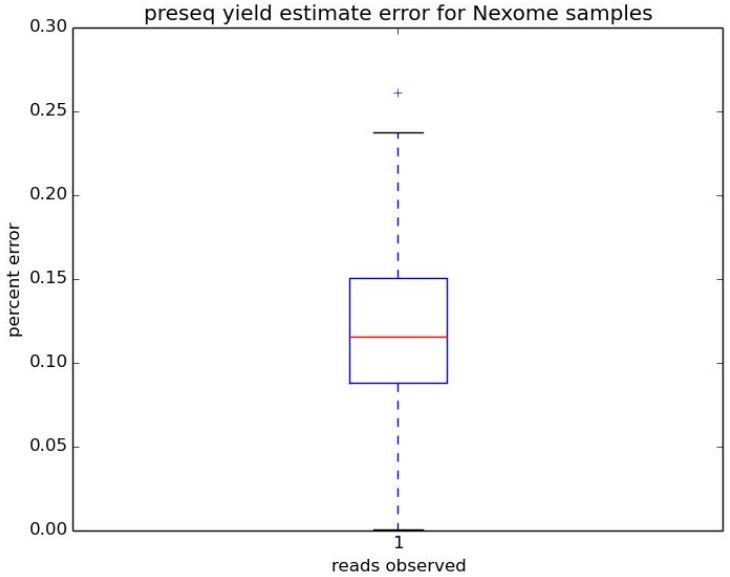


Execution time breakdown (135 M BAM file)

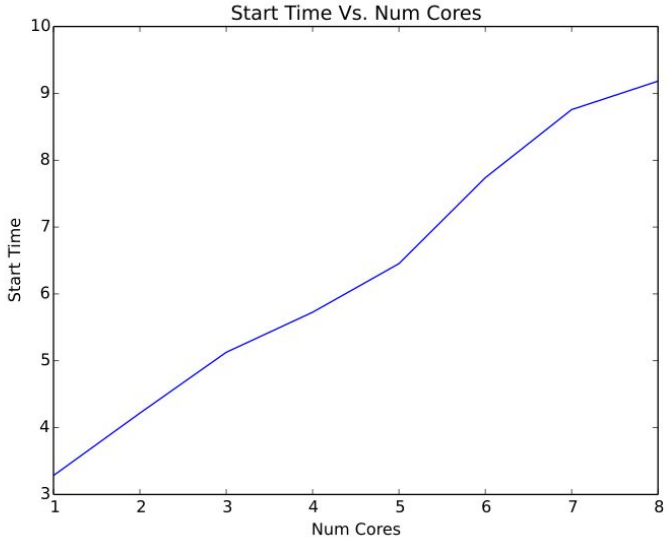


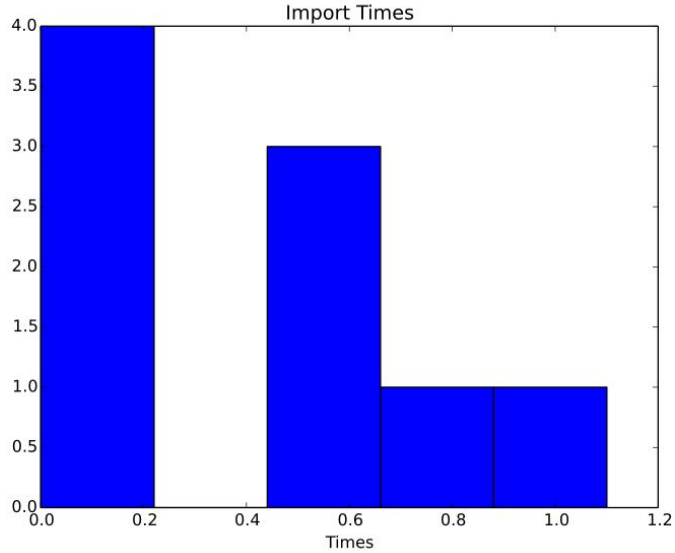
error analysis



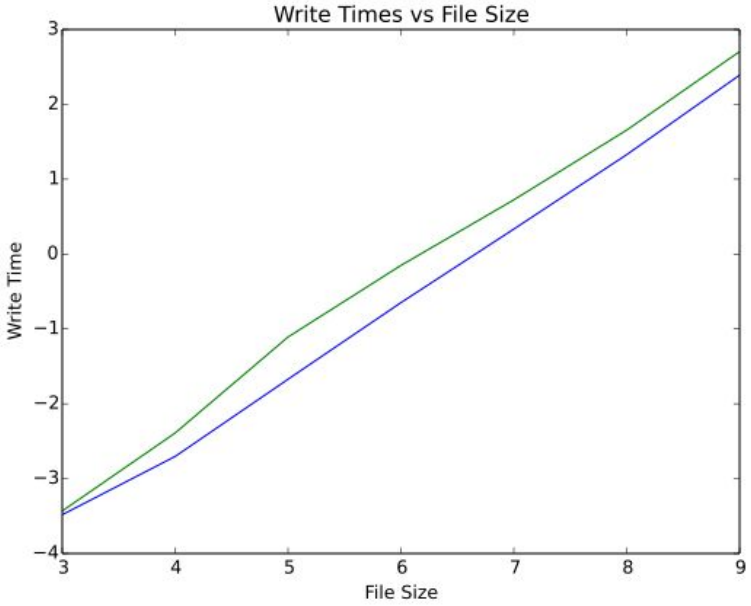


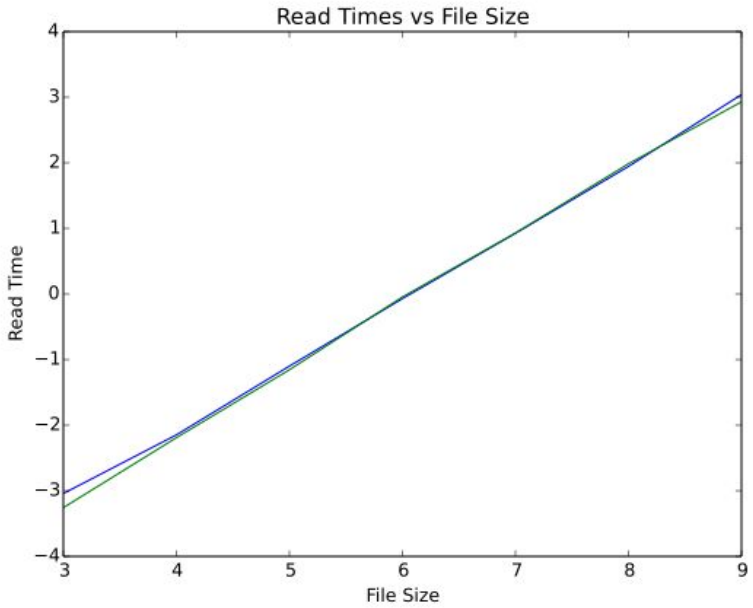
Costs of Using Julia





I/O Cost


Write Times vs File Size
Write Time
File Size
3
2
1
0
−1
−2
−3
−4
3
4
5
6
7
8
9


Read Times vs File Size
Read Time
File Size
4
3
2
1
0
−1
−2
−3
−4
3
4
5
6
7
8
9

Contributions

- Parallelized read count calculation
- Implemented Julia wrapper for molecular complexity prediction package `preseq.cpp`
- Benchmarked molecular complexity calculation on 16 core machine, currently testing with larger data sets on production servers
- Analyzed portions of Julia language which hindered speedup
- **Preseq reference**: Daley, Timothy, and Andrew D. Smith. "Predicting the molecular complexity of sequencing libraries." Nature methods 10.4 (2013): 325-327.