Mapping single molecule sequencing reads using basic local alignment with successive refinement(BLASR): application and study.

2013-11-18 Lab meeting

Yifan Yang

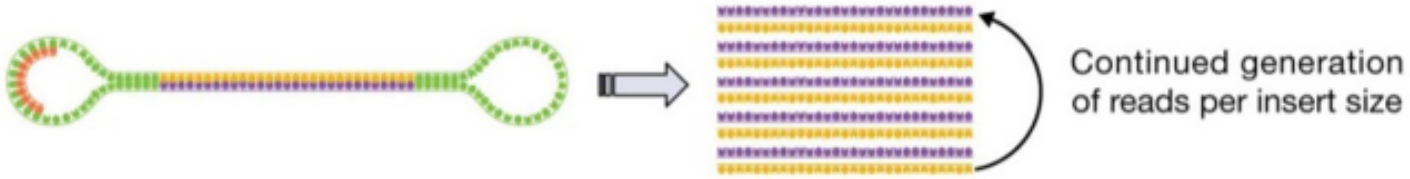
- Background
- Scientific question
- Algorithm the paper introduced
- Discussion

PacBio® Data Characteristics

Depending on the insert size of the library, the PacBio® *RS* can be optimized to generate longer sequences or shorter but higher quality sequences.

Continuous Long Reads (CLR). A large insert library (e.g. 6-10kb) results in long CLR reads up to 10 kb:

Circular Consensus Sequence (CCS). A short insert library (e.g. 500-1000 bp) favors multiple passes around each circular SMRTbell™ construct. The sequence generated by multiple observations of a single DNA molecule can be summarized as a higher quality (>99% accuracy) consensus sequence.



Background: Advantage v.s disadvantage of PacBio SMS

Disadvantage:

PacBio Long reads: ~11% error (15%)

Error: uniformly distributed insertions and deletions, very few substitution

Advantage:

mean= 2246bp

max= 23,000bp

Single molecule sequencing

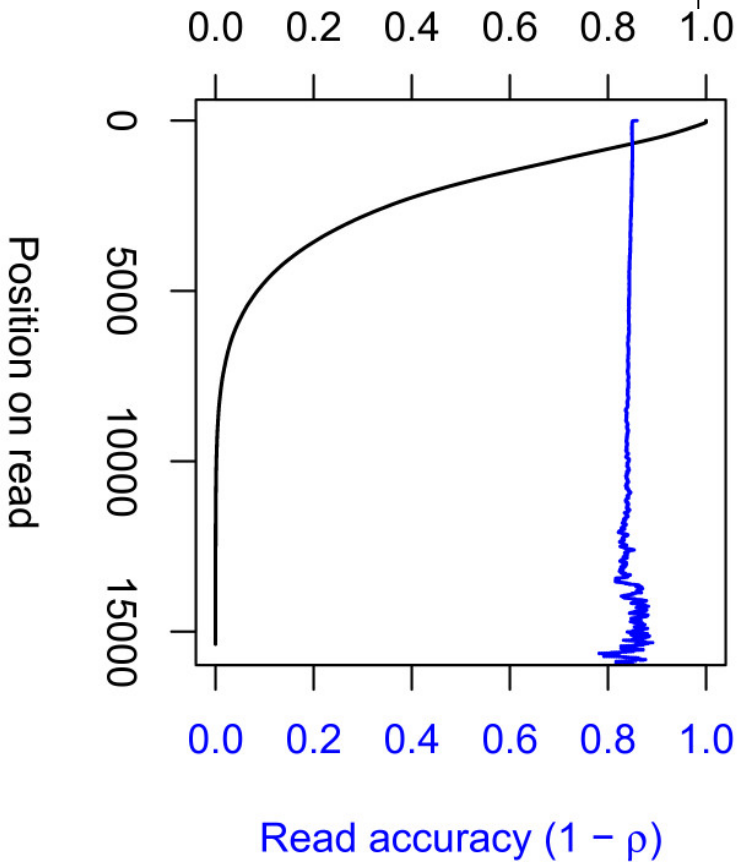
Available for RNA-seq-long reads RNA-seq

Two main problems of short reads RNA-seq:

Ambiguous reads mapping

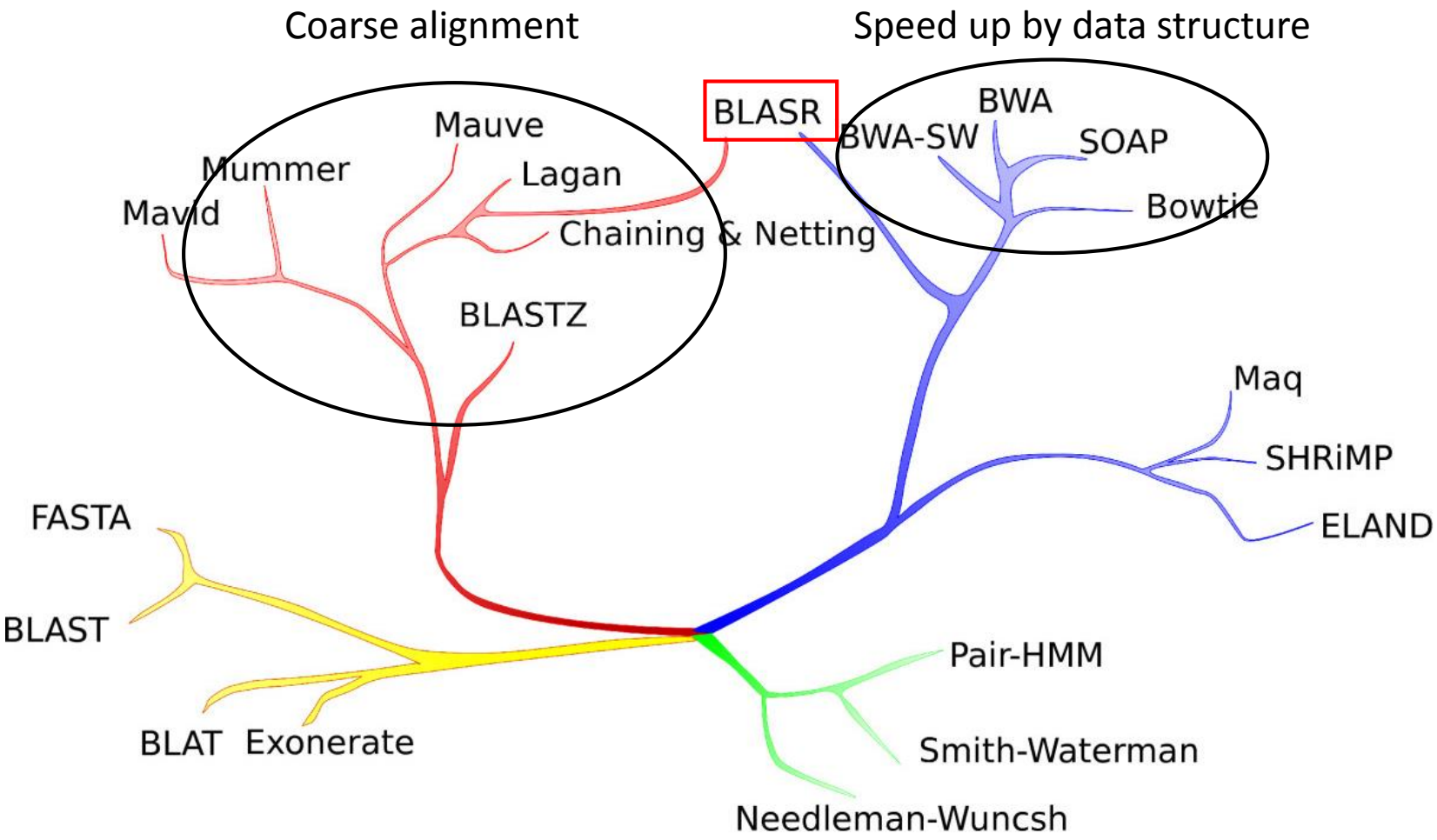
Assembly for transcripts

Will PacBio SMS be a future?

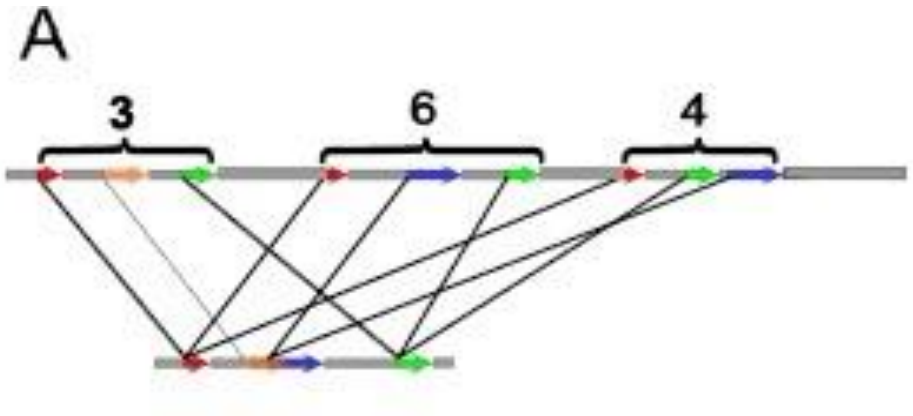


Question:
How to map the reads to genome with high error rate?

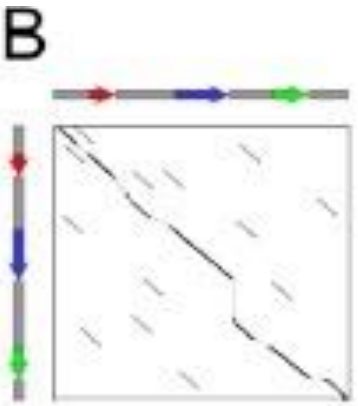
Algorithms relationship



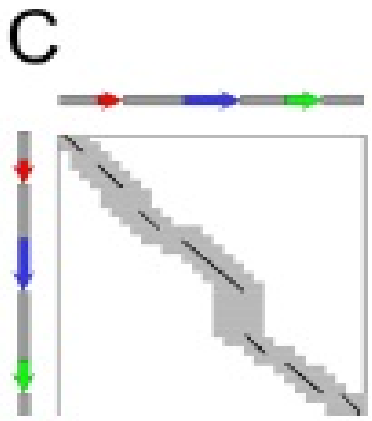
3 steps Algorithm



Find anchor



rough alignment



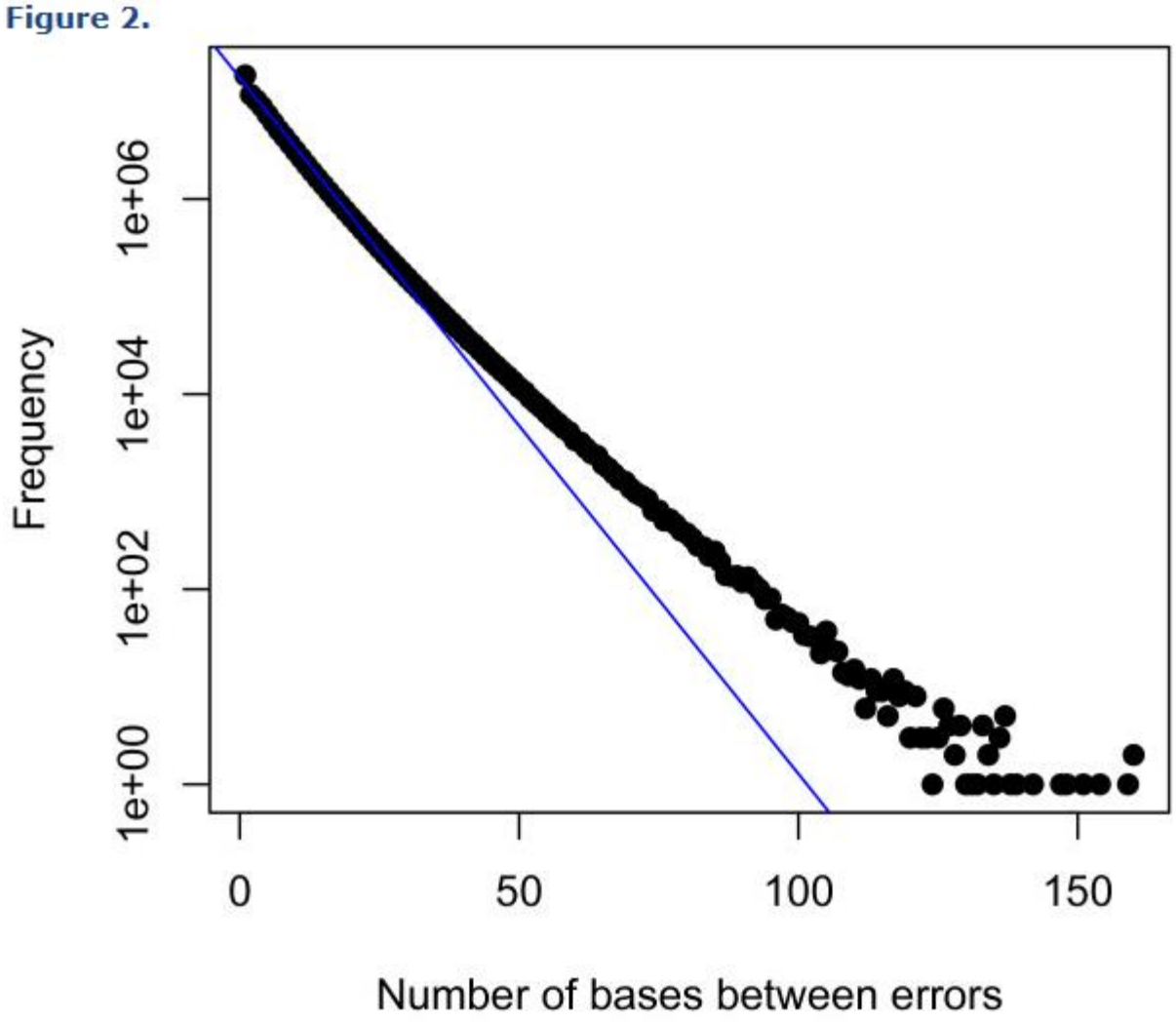
refinement alignment

$$s_{i,j} = \min \begin{cases} s_{i-1,j-1} + \begin{cases} 0 & \text{if } r_i = g_j \\ \mathcal{S}_i & \text{if } r_i \neq g_j, \hat{\mathcal{S}}_i = g_j \\ \text{MISMATCHPRIOR} & \text{otherwise} \end{cases} \\ s_{i-1,j} + \mathcal{I}_i \\ s_{i,j-1} + \begin{cases} \mathcal{D}_i & \text{if } \hat{\mathcal{D}}_i = g_{j-1} \\ \text{DELETIONPRIOR} & \text{otherwise.} \end{cases} \end{cases}$$

Two critical questions:

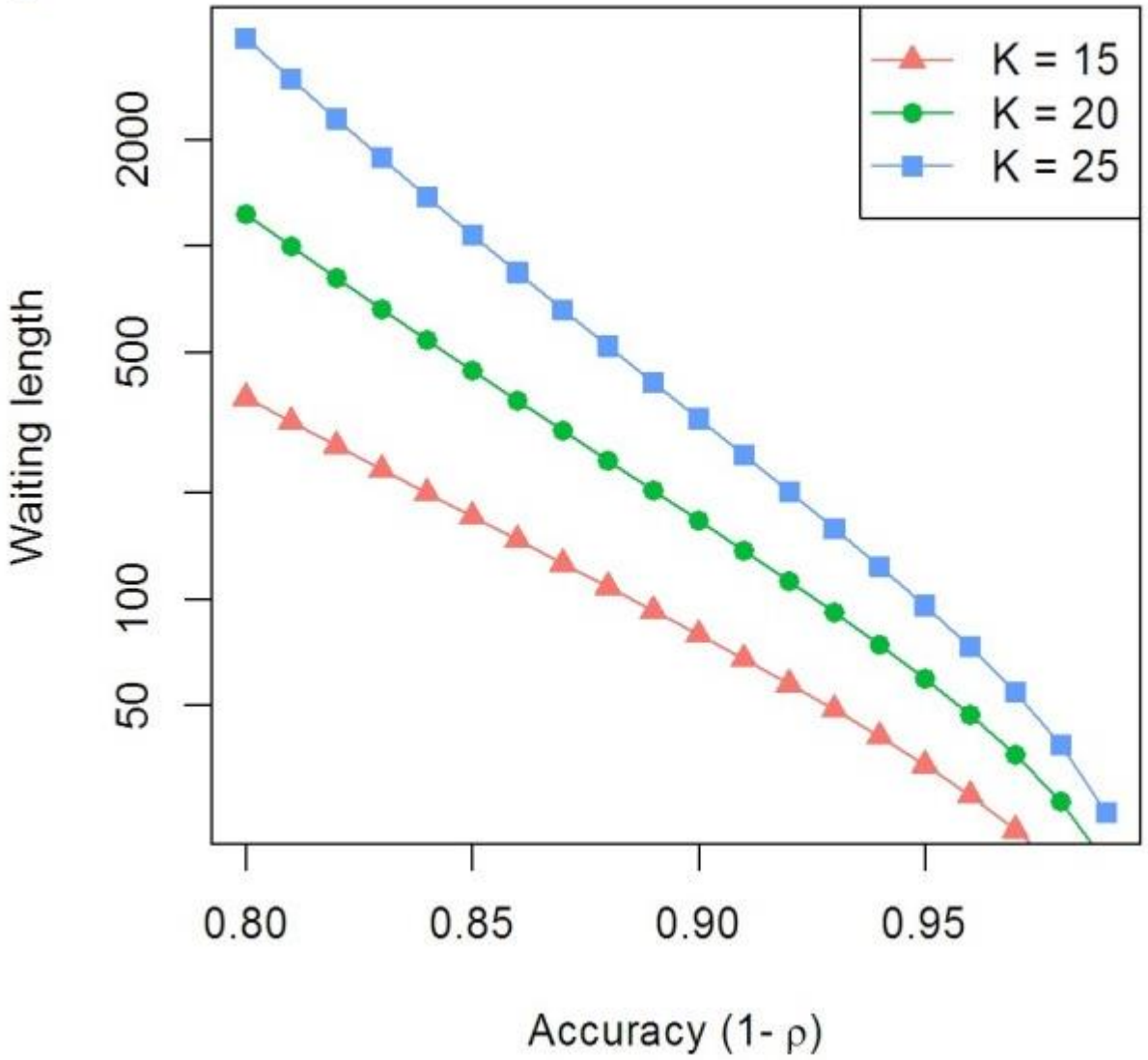
1. How to define the length of anchor K?
2. Whether anchor number N will influence mapping?

How to define the length of anchor K?

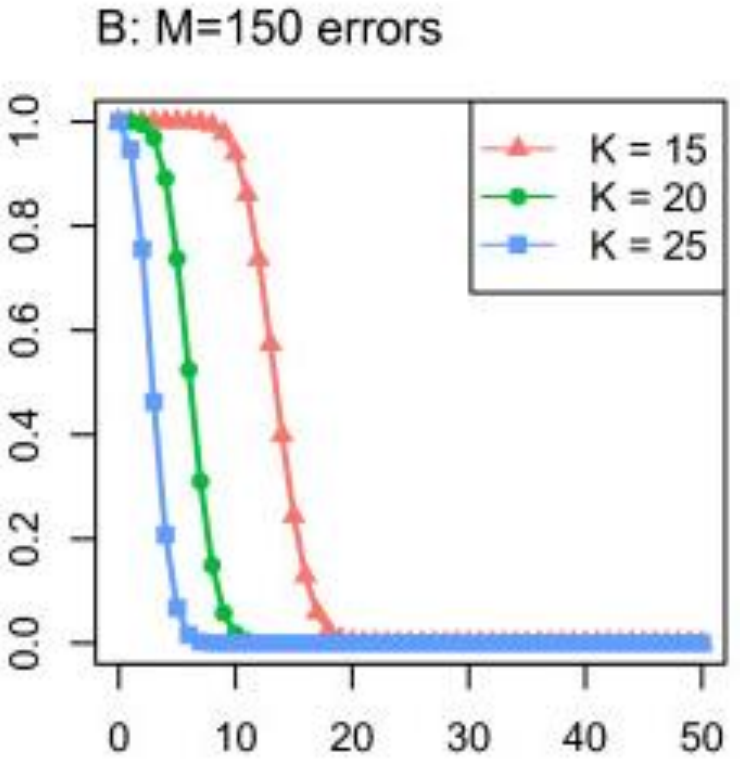


How many bases have to be at least sequenced to get a K error-free length?

Figure 3.



They expect to find at least 10 anchors for one read

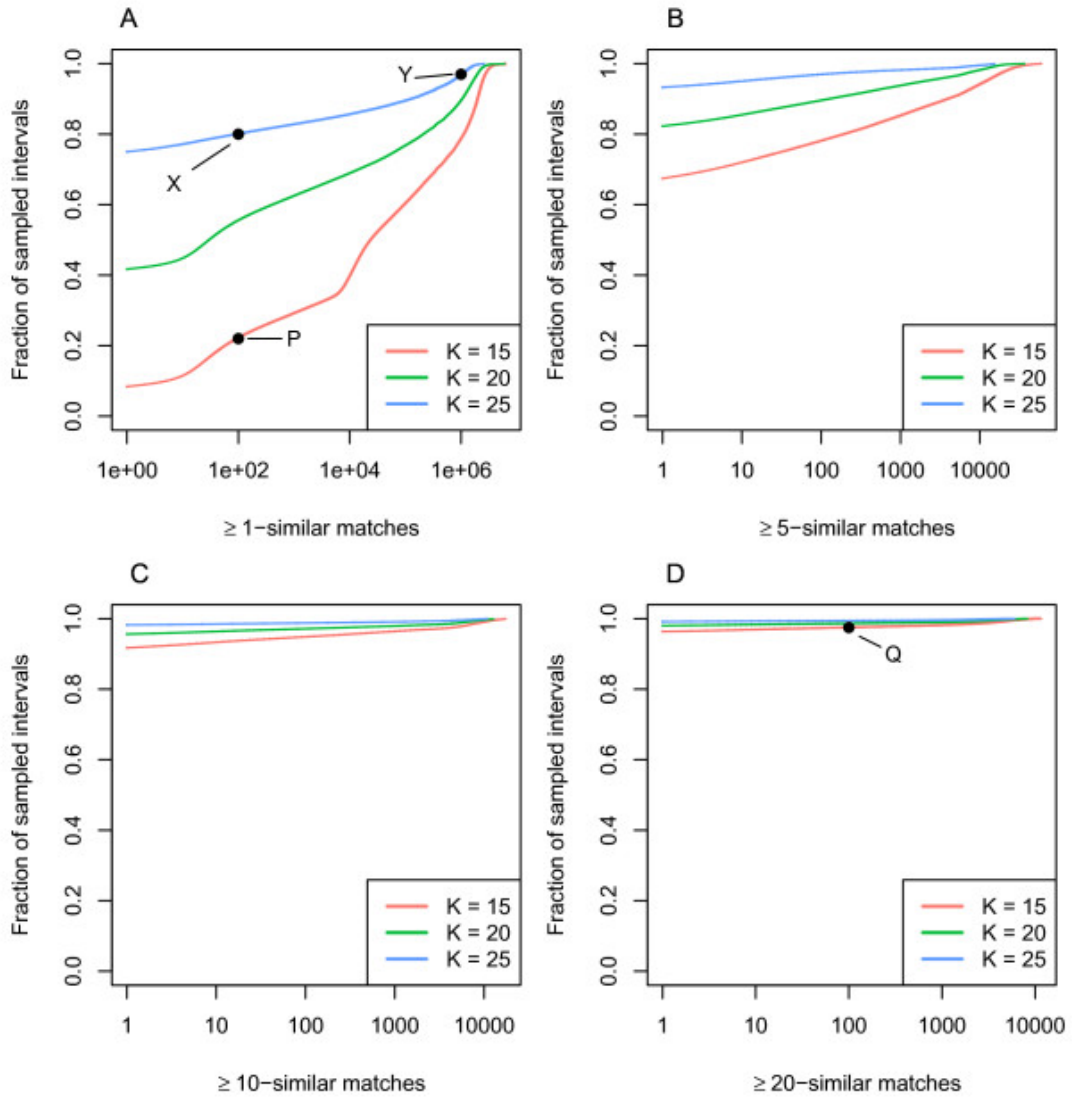


X: number of anchors
Y: probability of sequencing at least X anchors.

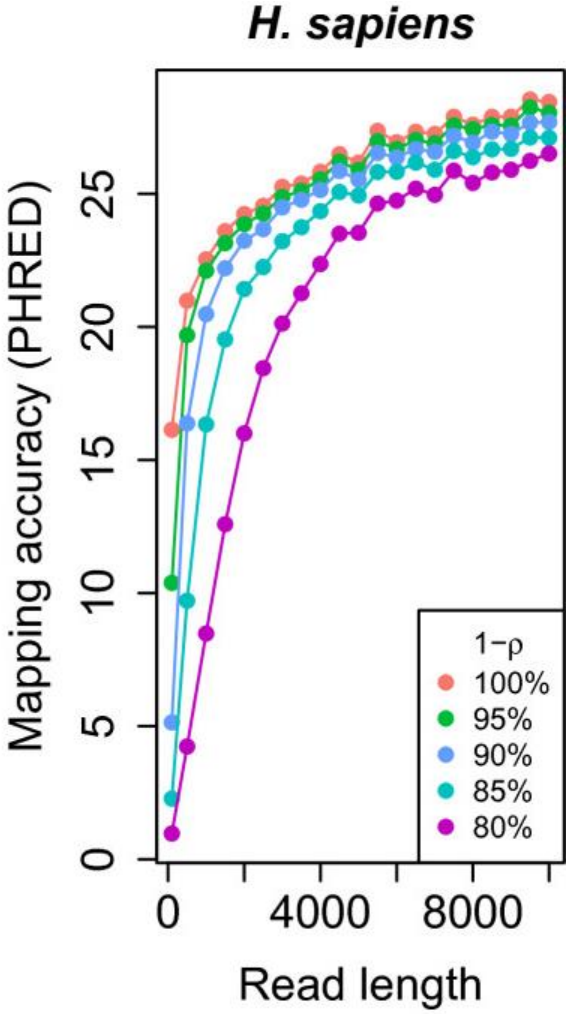
L = 1000

Similar regions need more anchors.

Figure 5.

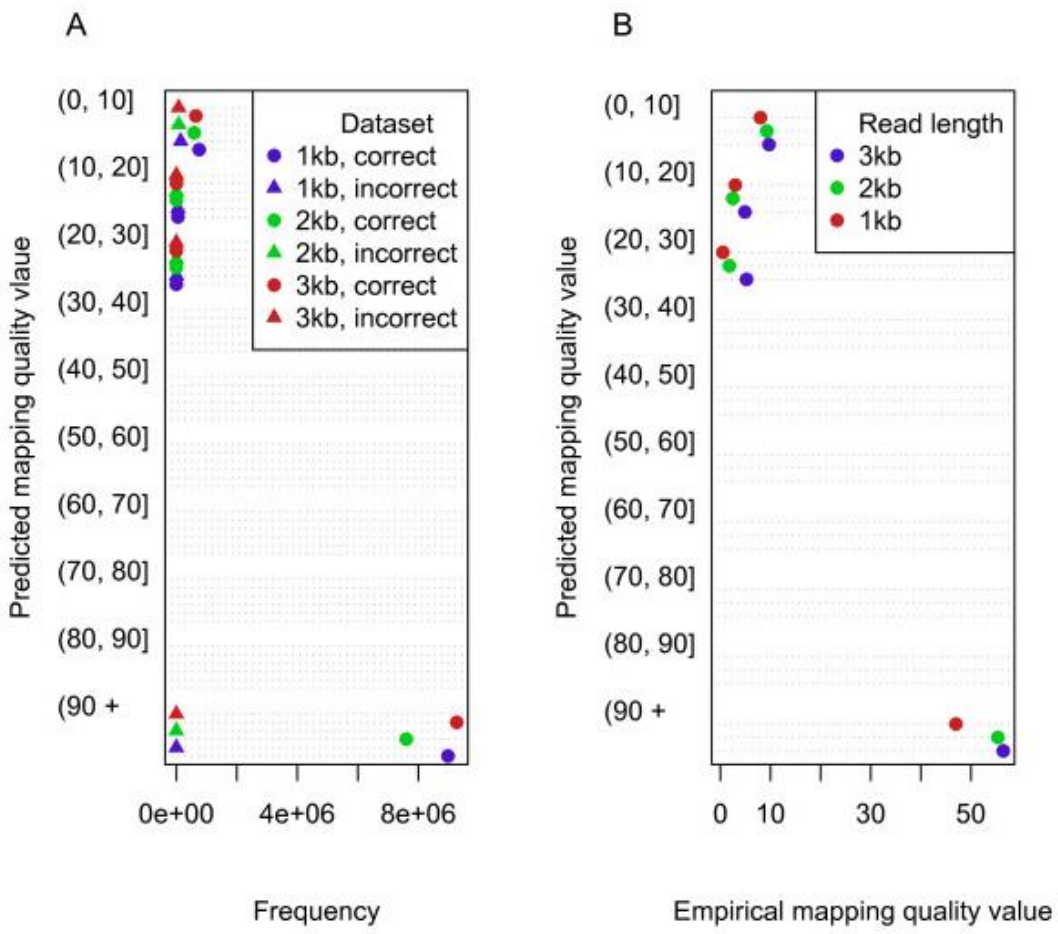


As the read length increases, the mapping quality increases!



Simulated datasets

Figure 8.



“Real” datasets btw different methods

Table 2

A comparison of the BLASR, BWA-SW, and BLAT methods on *E. coli* reads

Method	Number of aligned reads	Number of aligned bases	Run time
BLASR-SA	94057	230.8 M	20m 54s
BLASR-BWT	94527	230.1 M	33m 57s
BWA-SW	97729	132.4 M	434m 5s
BLAT	99530	181.7 M	4724m 40s

Table 3

A comparison of the BLASR, and BWA-SW methods on simulated reads

Method	Correctly mapped		Incorrectly mapped		Skipped	Runtime	Memory
	reads	bases	reads	bases	reads		footprint
E. coli							
BLASR-SA	108789	266.5M	229	0.38M	3766	48m 18s	202 MB
BLASR-BWT	108795	265.3M	259	0.45M	3604	59m 39s	46 MB
BWA-SW	111192	261.9M	1835	0.91M	3005	223m 57s	190 MB
H. sapiens							
BLASR-SA	41726	102.3M	1074	1.89M	413	92m 26s	14.7 GB
BLASR-BWT	41582	101.7M	1159	1.75M	472	53m 26s	8.1 GB
BWA-SW	40381	96.3M	292	1.16M	1554	105m 24s	4.2 GB

Discussion

- How to evaluate method?
- Long reads rna-seq is promising?

Thanks!