An Experimental Evaluation of the Performance of RNA-Seq Mapping Tools

ABSTRACT

Transcriptome sequencing (RNA-Seq) has become a key technology in the field of transcriptomics for quantifying gene expression, detecting novel transcripts, analyzing RNA functions, etc. For organisms with reference genomes, mapping RNA-Seq reads to the genomic sequences is typically the first step to process RNA-Seq data. In the last few years, many algorithms and tools for mapping RNA-Seq reads have been developed. Since the objectives and constraints of these methods are usually different, their performance varies. How to choose the most appropriate mapping tools to analyse a specific RNA-Seq dataset so some particular performance expectations can be met is an important question that bioinformaticians need to address. Here, we provide a systematic experimental evaluation of some state-of-the-artRNA-Seq mapping tools by studying their accuracy in read alignment and junction detection on simulated RNA-Seq data with different sequencing depths,read lengths and rates of substitutions, indels and sequencing error. In addition, the time efficiency and memory usages of the tools will beinvestigated. We are also interested in the impact of paired-end reads on the performance of the tools and the question of to what degree the tools are able to take advantage of parallel computation. Some real data tests will be used to confirm the simulation results. We hope that our study will provide information useful for choosing suitable mapping tools in RNA-Seq data analysis.

1. INTRODUCTION

Transcriptome is an important part of the cell. RNA-seq based on next-generation sequencing(NGS) provides a powerful technique for transcriptome profiling(Wang, Gerstein et al. 2009). Bringing insights to quantifying gene expression, detecting novel transcripts, analyzing RNA functions, and etc., RNA-Seq is proved to be a accurate method to survey the entire transcriptome in a quantitative and high-throughput way than expressed sequence tag(EST) sequencing and microarray technology.

A critical step in analyzing RNA-Seq data is to map millions of reads back to a reference genome. Various mapping tools targeted on mapping DNA, RNA, miRNA or bisulfate reads, have been developed(Fonseca, Rung et al. 2012). With specific consideration to processing junction reads from RNA-Seq, RNA mapping tools plays an important role in analyzing RNA-Seq data.

From early RNA-Seq mapping tools include TopHat(Trapnell, Pachter et al. 2009)
, SpliceMap(Au, Jiang et al. 2010) and MapSplice(Wang, Singh et al. 2010) to newly published tools include OLego(Wu, Anczuków et al. 2013), TrueSight(Li, Li-Byarlay et al. 2012) and STAR(Dobin, Davis et al. 2012), over fifteen RNA-Seq mapping tools have been developed(see table 2). Which tool to choose according to certain time and memory constrain and data features is an issue waits to be addressed.

Some review papers on mapping tools have been published to address this problem. Those papers either present a algorithmic summary (Garber, Grabherr et al. 2011) or summarize mapping tools documentations(Fonseca, Rung et al. 2012), yet provide few recommendation on choosing suitable mapping tools for RNA-Seq data.

This paper aims testing seven state of the art RNA mapping tools, thus provide

information on different features of those tools and better guide choosing suitable tool fitting specific RNA-Seq data.

2. METHODS

Our evaluation consists of three steps: (1)simulate RNA-Seq reads by BEERS(Grant, Farkas et al. 2011)(2)map simulated reads to the reference genome by seven selected RNA mapping tools(3)evaluate the performance of mapping**(see fig 1)**.

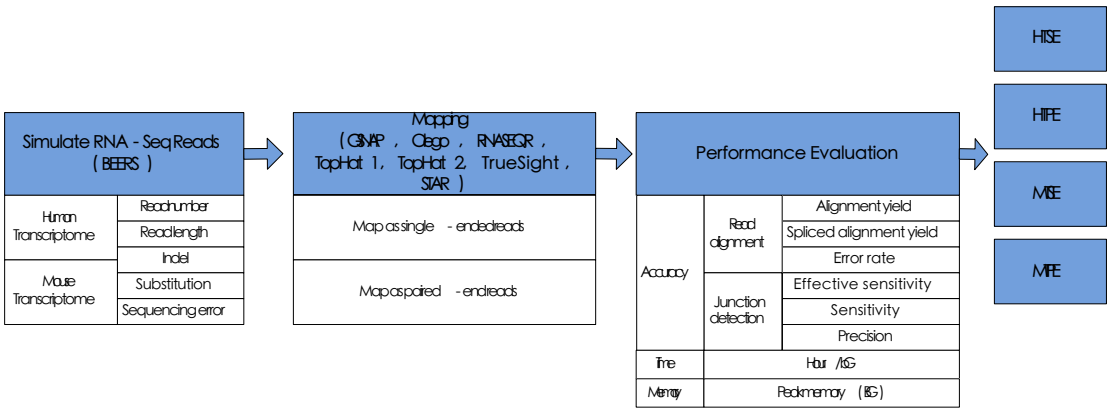

Fig 1. legend. Experimental evaluation process. (1)RNA-Seq dataset of human and mouse transcriptome with different features are generated by BEERS (2) Selected tools map reads to genomic reference dataset by dataset and take reads as single-ended or paired-end separately. (3) Mapping result of each tool on each dataset as single-ended or paired-end reads is evaluated according criteria listed in the figure. (4) Four sets of evaluation results are obtained: HTSE(Evaluation on dataset of Human Transcriptome mapped as Single-ended reads), HTPE(Evaluation on dataset of human transcriptome mapped as paired-end reads), MTSE(Evaluation on dataset of mouse transcriptome mapped as single-ended reads), MTPE(Evaluation on dataset of mouse transcriptome mapped as paired-end reads).

2.1.Experimental design

(1) Simulate RNA-Seq Reads

The "real" genomic locations of RNA-Seq reads generated by sequencing platform, like Illumina, are unknown and different "artifacts" on reads mixed together. So our experimental evaluation depends mainly on simulated reads and separately varying "artifacts". Additionally, real RNA-Seq data is used to confirm the simulated results.

Three kinds of "artifacts", which are substitution, indel and sequencing error, together with read number and read length are five changing factors in simulated RNA-Seq datasets. Datasets with read number of 10, 20, 30, 40, 50, 60, 80, 100 million and read length of 25, 50, 75, 100, 125, 150, 175, 200, 225, 250, 500nt, are generated by BEERS. The default value of read number and read length of the datasets are 30 million and 100nt. Close to "real" statistical values of indel frequency, substitution frequency and sequencing error frequency provided by BEERS, three values of each factor and together with the artifact-free situation (0 value) are adopted for simulating reads. All the "real" statistical values are default setting in BEERS. For indel are 0,

0.0005(default), 0.001, 0.0015. For substitution, 0, 0.005, 0.001(default), 0.01 and 0, 0.005(default), 0.007, 0.009 for sequencing error**(See table)**.

Table 1.

Factor	Values
Read number	10million, 20m, **30m**, 40m, 50m, 60m, 80m, 100m
Read length	25nt, 50nt, 75nt, **100nt**, 125nt, 150nt, 175nt, 200nt, 225nt, 250nt, 500nt
Indel frequency	0, **0.0005**, 0.001, 0.005, 0.01
Substitution frequency	0, **0.001**, 0.005, 0.01
Sequencing error	0, **0.005**, 0.007, 0.009

Note: Values simulating real data are in bold. Simulation datasets are noise(indel or substitution or sequencing error) free, with only one factor varies.

For consideration toward transcriptome variation and computational replication, both human transcriptome and mouse transcriptome are simulated. hg19 of human genome and RefSeq transcriptome annotation is used for HTSE and HTPE. mm9 of mouse genome and 11 merged databases of transcriptome annotation is used for MTSE and MTPE. Additionally, paired-end reads simulated by BEERS are mapped as single-ended reads and paired-end reads to see if paired-end reads could provide extra information for mapping and also act as computational replication.

(2)Real RNA-Seq data

GSE26248: The real data where BEERS empirical gene expression distribution come. http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE26248

GSM958728: ENCODE RNA-Seq data published on July, 2012 of GM12878. http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM958728

(3)Tools to be evaluated(algorithmic classification, reason to be chosen)

Over 30 mapping tools have been published, here we select RNA mapping tools with fairly amount of users and timely update releases. We evaluate TopHat1,TopHat2 GSNAP, RNASEQR, TrueSight, OLego, STAR of the their newest version before Dec. 31st, 2012**(see table).**

Table 2（Supplement）

	Publish	Latest Release	Cite Times	Version	Reason	Link
QPALMA	*Bioinformatics*.2008,	2010.5	87	-	NNR	http://www.raetschlab.org/suppl/qpalma
TopHat	***Bioinformatics*.2009,3**	**2012.11**	**778**	**2.0.6**		**http://tophat.cbcb.umd.edu/**
GSNAP	***Bioinformatics*.2010,2**	**2012.11**	**124**	**-**		**http://research-pub.gene.com/gmap/**
SpliceMap	*Nucleic Acids Research* 2010,5	2010.10	103	3.3.5.2	NNR	http://www.stanford.edu/group/wonglab/SpliceMap/
MapSplice	*Nucleic Acids Research* 2010,8	2011.3	84	2.x	NNR	http://www.netlab.uky.edu/p/bioinfo/MapSplice
SOAPsplice	*Frontiers* 2011,7	2012.2	9	1.9	FU	http://soap.genomics.org.cn/soapsplice.html
RUM	***Bioinformatics*.2011,7**	**2012.10**	**28**	**2.0.x**		**http://www.cbil.upenn.edu/RUM/userguide.php**
RNASEQR	***Nucleic Acids Research* 2011.10**	**-**	**4**	**1.0.2**		**http://hood.systemsbiology.net/rnaseqr.php**

PASSion	*Bioinformatics*.[2012,1]	-	4	1.2.1		https://trac.nbic.nl/passion/
TrueSight	***Nucleic Acids Research*** **[2012]**	**2012.9**	**-**	**0.06**		**http://bioen-compbio.bioen.illinois.edu/TrueSight/**
OLego	***Nucleic Acids Research*** **Accepted**	**2012.11**	**-**	**1.0.8**		**http://zhanglab.c2b2.columbia.edu/index.php/OLego**
GEM	***Nature Method*** **[2012,12]**	**2012.10**	**2**	**-**		**http://algorithms.cnag.cat/wiki/The_GEM_library**
STAR	***Bioinformatics.***	**2013.1**	**1**	**-**		**http://code.google.com/p/rna-star/**
Supersplat	*Bioinformatics*[2010.1]	-	28		NNR	http://mocklerlab.org/tools/1
PALMapper	*Current Protocols in Bioinformatics*[2010.10]	2011.01	9	0.4	NNR, FU	http://cbio.mskcc.org/public/raetschlab/software/palmapper/
MicroRazerS	*Bioinformatics*[2010.10]	2009.11	9	1.0	NNR, FU	http://www.seqan.de/projects/microrazers/

Note: **Bold lines** correspond to evaluated tools. NNR, short for "no new release", means no release during the year of 2012. FU, short for "few users", means less than 10 citations after publishing at least one year. Data collected before Dec. 31st, 2012

2.2.Evaluation criteria:

For the performance of tools on simulated dataset, we mainly check three aspects: accuracy of aligning reads and detecting junctions, time efficiency and memory usage. Several criteria listed in **(see table 3)** are used for these three aspects.

Table 3.

Accuracy	
Read alignment	Alignment yield(Scale: 0-1):
	#(Correct reads' alignment)/ #(Total reads)
	Spliced alignment yield(Scale: 0-1)
	#(Correct junction reads' alignment)/ #(Junction reads in the sample)
	Error yield(Scale: 0-1)
	#(Correct reads' alignment)/ #(Uniquely mapped reads)
Junction detection	Effective sensitivity(Scale: 0-1)
	#(Detected junctions in the sample)/ #(Junction reads in the sample)
	Sensitivity(Scale: 0-1)
	#(Detected junctions in the genome)/ #(Junctions in the genome)
	Precision(Scale: 0-1)
	#(Detected junctions in the sample)/ #(Detected junctions)
Time usage	Time (Unit: h/Gb)
	(Total alignment time)/#(Gb of reads)
Memory usage	Peak memory(GB)

Note: #(X) means the number of X.

With different mapping result reported by tools, reads are classified into three categories: uniquely mapped reads, multiple mapped reads and unmapped reads. Multiple mapped reads are treated as unmapped reads if all hits of one read have equal confidence and as uniquely mapped reads if the tool gives priority of confidence of the hits of one read. The hit with highest confidence will take as this multiple

mapped read's unique hit.

For tools' performance on reads alignment, alignment yield is the percent of reads being correctly mapped and shows tool's overall capacity of finding reads' "real" locations. Spliced alignment yield is the percent of junction reads being correctly mapped and shows tool's capacity of finding junction reads' "real" locations. Error yield is the percent of unique hits being falsely reported and shows tool's likelihood of being wrong. Alignment yield may not always correspond with spliced alignment yield. Error yield's decrease will not guarantee (Spliced) alignment yield's increase.

For tools' performance on junction detection, effective sensitivity is the percent of junctions supported by at least one read in the dataset being correctly reported and shows tool's capacity of detecting junction. Sensitivity is the percent of junctions annotated being correctly reported and also shows tool's capacity of detecting junction. Precision is the percent of reported junctions being correct. Good algorithmic design could find junctions even without read support and results in more detected junctions among annotated junctions.

Time for mapping 1 Gb is used to check tool's time-efficiency. Peak memory of the run of a tool is examined as criteria for time usage.

In most cases, alignment yield and spliced alignment yield, effective sensitivity and sensitivity, spliced alignment yield and effective sensitivity are pairs that will support each other. (Spliced) alignment yield and error yield, (effective) sensitivity and precision are pairs with tradeoffs.

3. RESULTS

3.1.Features of seven RNA mapping tools

Robustness is a nice feature of softwares. RNA mapping tools' robustness upon increased varying factor becomes critical in case of RNA-Seq data of relatively low quality. "Substitution robust", "sequencing error robust" and "indel robust" exhibit tools' robustness upon increased substitution frequency, sequencing error rate and indel frequency. They are features draw from stabilized error yield upon increased varying factors. RNASEQR, TopHat1 and TopHat2 are proved to be substitution robust and sequencing error robust. RNASEQR and TrueSight are proved to be indel robust according to our systematic evaluation**(see table 4)**.

Tools all perform normally for read length in the range of 50nt to 250nt. Short reads with length less than 50nt and long reads with length longer than 250nt present challenge to RNA mapping tools. "Long reads disposition" and "short reads disposition" are features indicating how well they deal with short or long reads. OLego, RNASEQR and STAR show good performance upon long reads. For reads with length 25nt, it puts stress for tools to align junction reads or detect junction correctly, only RNASEQR exhibits capability to deal with short reads **(see table 4).** Default parameter settings are adopted in this case.

OLego, RNASEQR and STAR are "time efficient", which means they could finish mapping 100 million reads with length of 100nt within 10 hours**(see table 4).**

For personal computer with memory of 4GB, OLego, RNASEQR and TopHat1 or 2 work normally. According to our evaluation, these four tools are "memory efficient"**(see table 4).**

Table 4. Features of seven RNA mapping tools

	GSNAP	OLego	RNASEQR	TopHat1	TopHat2	TrueSight	STAR
Substitution robust			√	√	√		
Sequencing error robust			√	√	√	√	
Indel robust			√			√	
Long reads disposition		√	√				√
Short reads disposition			√				
Time efficient		√	√				√
Memory efficient		√	√	√	√		

Substitution robust: stabilizaed error yield upon increased substitution; Sequencing error robust: stabilizaed error yield upon increased sequencing error; Indel robust: stabilizaed error yield upon increased indel; Long reads disposition: capability of dealing with reads ＞ 250nt; Short reads disposition: capability of dealing with reads ＜ 50nt; Time efficient: capability of mapping reads with speed < 1hour/Gb; Memory efficient: capability of mapping reads with peak memory < 4GB. **For statistical data in detail, check HTSE data or MTSE data.**

3.2.Accuracy

(1)Read number

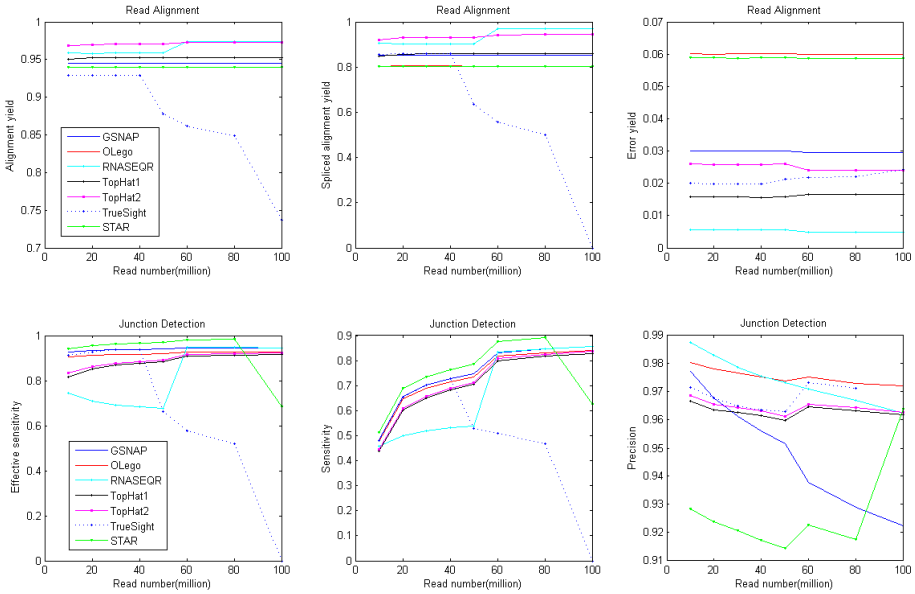

Fig 2. htse.rn.acu.tif

For read alignment, with reads' coverage increase, all tools except for TrueSight have alignment yield stabilized around 0.95 and spliced alignment yield stabilized around 0.9. Compared with OLego and STAR which have error yield stabilized around 0.06, GSNAP, TopHat2, TrueSight, TopHat1, RNASEQR yields lower error, which are all less than 0.05.

For junction detection, with reads' coverage increase, which also means more junctions being included, effective sensitivity and sensitivity increase, which means more junctions being detected. The precision of tools decrease which indicates that there are new junction being included but not correctly detected.

(2)Read length

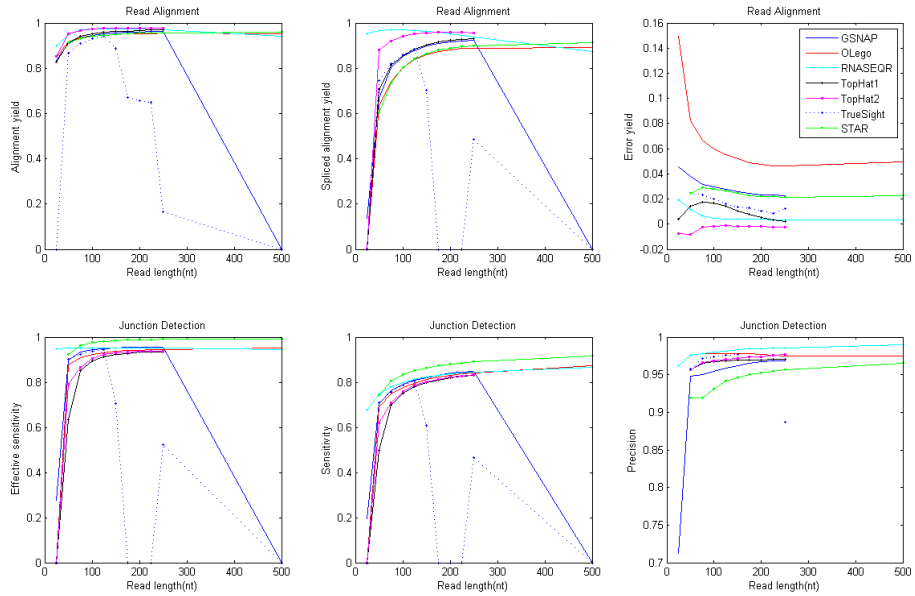

Fig 3. (read length)

(3)Sequencing error

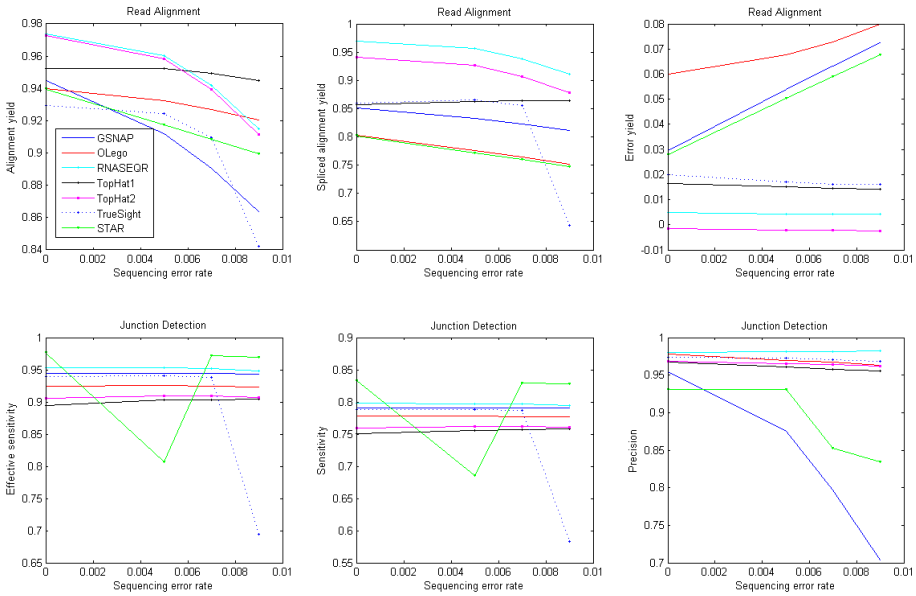

Fig 4.(Sequencing error)
For read alignment, all tools are affected by more sequencing error. By yielding more error, OLego, GSNAP and STAR are proved to be more sequencing error sensitive and TopHat1, TopHat2, TrueSight and RNASEQR are sequencing error robust with stabilized error yield.
For junction detection, STAR is more aggressive by demonstrating the highest effective sensitivity and sensitivity and lowest precision which means STAR finds the most junctions but makes the largest mistakes. GSNAP is secondly aggressive. Other tools are not much affected by more sequencing error for maintaining stabilized effective sensitivity, sensitivity and precision.

(4)Indel

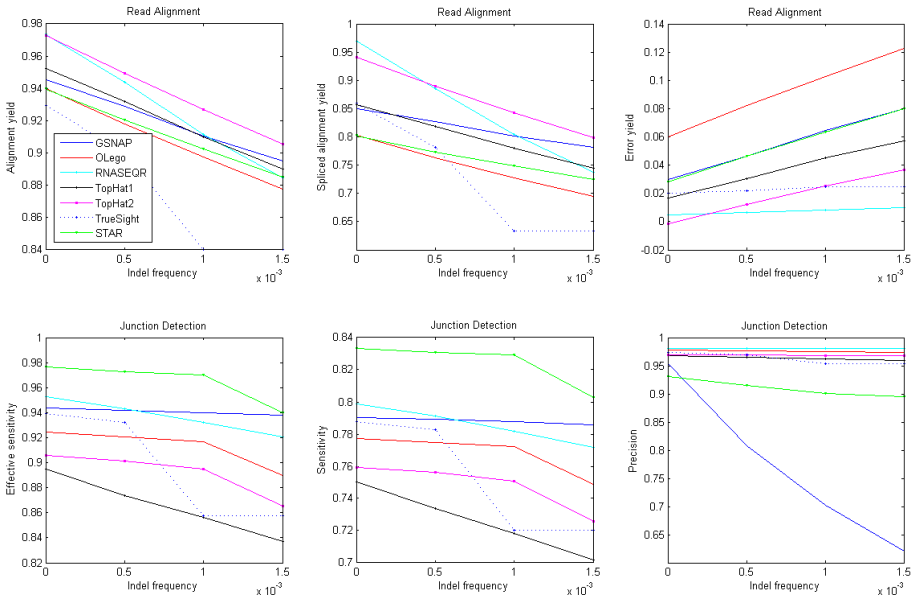

Fig 5.(Indel)
For read alignment, all tools' performance are affected by more indels. Only two tools which are TrueSight and RNASEQR show robustness upon more indels and other tools include OLego, GSNAP, STAR, TopHat1 and TopHat2 yield increased error with more indels in the sample.
For junction detection, STAR and GSNAP are more aggressive by demonstrating the highest effective sensitivity and sensitivity and lowest precision which means they find the most junctions but makes the largest mistakes. Other tools are not much affected by indel for maintaining stabilized precision.

(5)Substitution

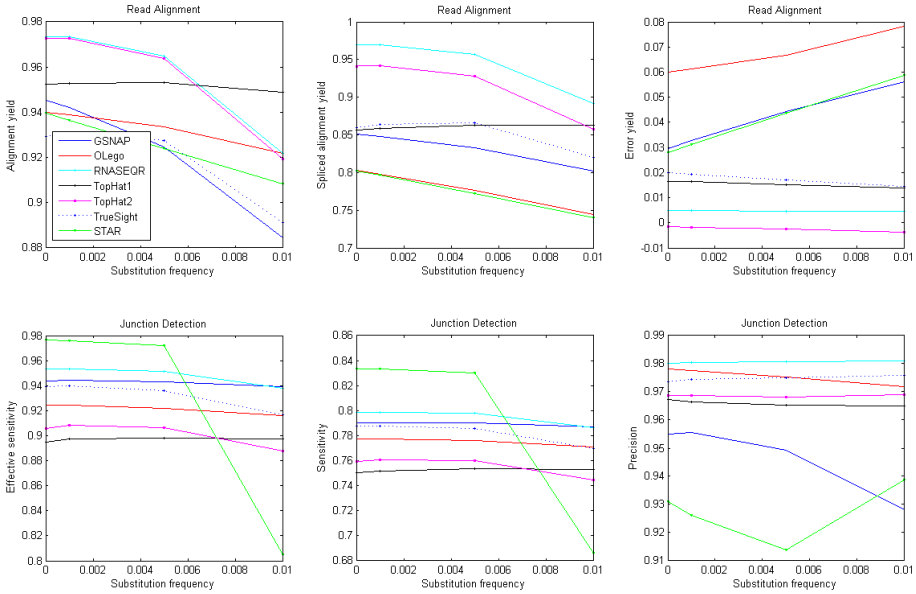

Fig 6. (substitution)
For read alignment, all tools are affected by more substitution. By yielding more error, OLego, GSNAP and STAR are proved to be more substitution sensitive and TopHat1, TopHat2, TrueSight and RNASEQR are substitution robust with stabilized error yield.
For junction detection, STAR is more aggressive by demonstrating the highest effective sensitivity and sensitivity and lowest precision which means STAR finds the most junctions but makes the largest mistakes. GSNAP is secondly aggressive. Other tools are not much affected by more substitutions for maintaining stabilized effective sensitivity, sensitivity and precision.

(6)Paired-end reads

3.3.Time efficiency and memory usage

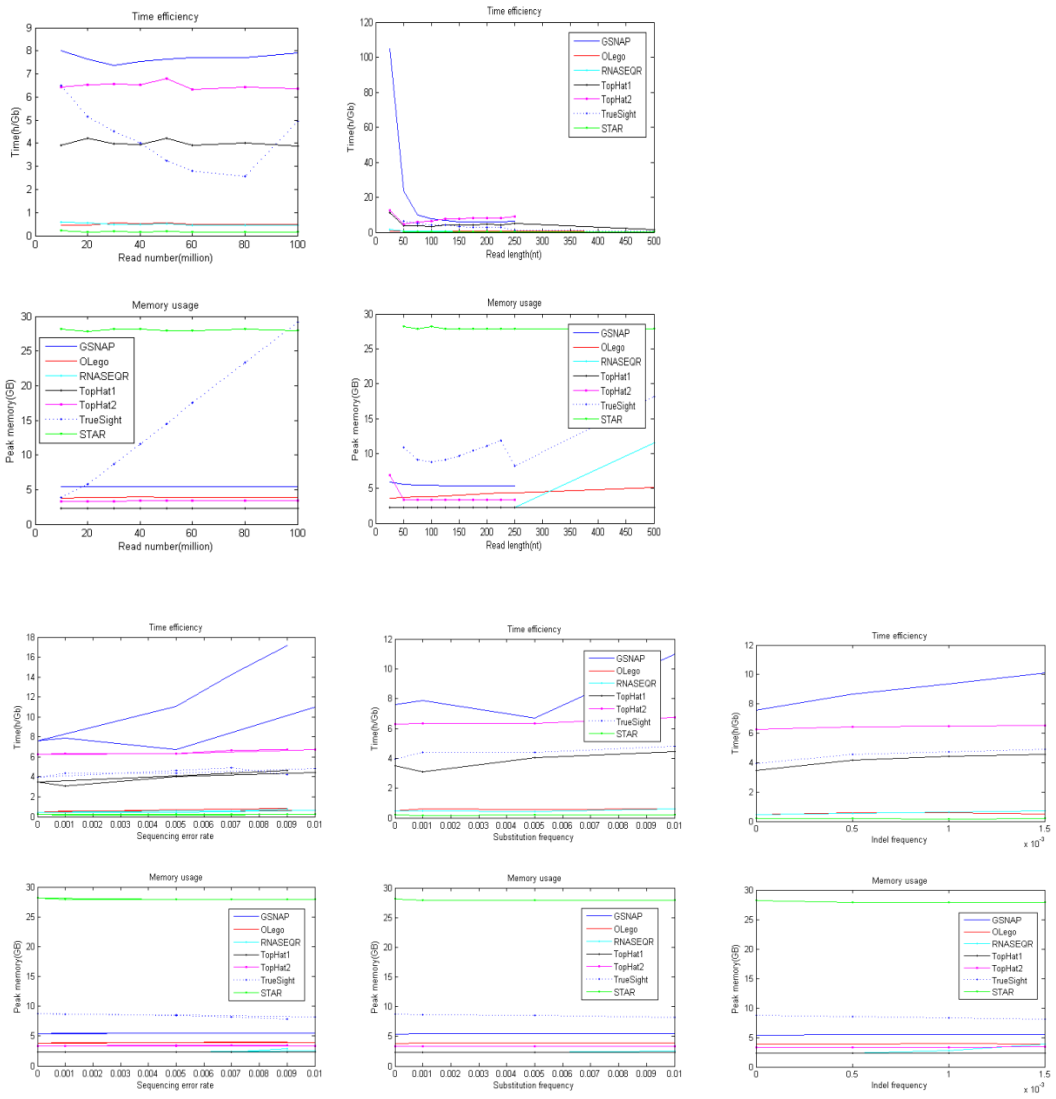

Fig 7.

Table 5. Average time and peak memory use of tools

	GSNAP	OLego	RNASEQR	TopHat1	TopHat2	TrueSight	STAR
AVR(Time), hour/Gb	7.649	0.514	0.526	4.046	6.570	4.680	0.186
VAR(Time)	0.145	0.036	0.041	0.120	0.142	1.262	0.023
AVR(Peak memory), GB	5.422	3.839	2.307	2.305	3.354	8.910	28.041
VAR(Peak memory)	0.025	0.040	0.000	0.000	0.012	8.356	0.157

For time efficiency, we take hours used to map 1 Gb as criteria. Fast tools, include STAR with speed 0.177 hour/Gb, OLego with 0.515 hour/Gb and RNASEQR with 0.526 hour/Gb, could finish mapping 1 Gigabase reads in 1 hour. Other tools, include GSNAP, TopHat2, TopHat1 are

7.649 hour/Gb, 6.570 hour/Gb and 4.046 hour/Gb fast(Table). TrueSight has **a good feature** to increase speed from 6.497 hour/Gb to 2.560 hour/Gb with read coverage increase(Figure).
For memory usage, we take peak memory of one run as criteria. OLego, RNASEQR, TopHat1, TopHat2 use less than 4GB memory maximally. Because of time and memory trade-off, STAR needs 28GB memory.

3.4.Real data(accuracy, time usage, memory usage)

Percent of uniquely mapped reads

4. DISCUSSION

Varying factors like substitution, sequencing error, indel have been explored separately. In real situation, all varying factor will mix together. It's hard to evaluate which tool solve the problem best. A gold standard for choosing most suitable tool is impossible to find. But memory efficiency will become critical if users are using laptop computer conducting RNA-Seq mapping. Moreover, with RNA-Seq technique develops, throughput and read length tend to increase, then time efficiency and long reads disposition will be non-trivial.
Interesting problems not explored in this paper, include mapping result's influence on isoform's abundance estimation using tools like cufflinks and also parallel mapping's improvement to each RNA mapping tools.

5. Reference

Au, K. F., H. Jiang, et al. (2010). "Detection of splice junctions from paired-end RNA-seq data by SpliceMap." Nucleic Acids Research.
Dobin, A., C. A. Davis, et al. (2012). "STAR: ultrafast universal RNA-seq aligner." Bioinformatics.
Fonseca, N. A., J. Rung, et al. (2012). "Tools for mapping high-throughput sequencing data." Bioinformatics.
Garber, M., M. G. Grabherr, et al. (2011). "Computational methods for transcriptome annotation and quantification using RNA-seq." Nat Meth 8(6): 469-477.
Grant, G. R., M. H. Farkas, et al. (2011). "Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM)." Bioinformatics 27(18): 2518-2528.
Li, Y., H. Li-Byarlay, et al. (2012). "TrueSight: a new algorithm for splice junction detection using RNA-seq." Nucleic Acids Research.
Trapnell, C., L. Pachter, et al. (2009). "TopHat: discovering splice junctions with RNA-Seq." Bioinformatics 25(9): 1105-1111.
Wang, K., D. Singh, et al. (2010). "MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery." Nucleic Acids Research 38(18): e178.
Wang, Z., M. Gerstein, et al. (2009). "RNA-Seq: a revolutionary tool for transcriptomics." Nat Rev Genet 10(1): 57-63.
Wu, J., O. Anczuków, et al. (2013). "OLego: fast and sensitive mapping of spliced mRNA-Seq reads using small seeds." Nucleic Acids Research.