Assessment of Aligner and SNP Caller for Next Generation Sequencing and a Fast and Accurate SNP Detection Method

Weixin Wang[1], Feng Xu[1] ,Panwen Wang[1], Mulin Jun Li[1],Pak Chung Sham[2], JJ Wang[1,*]

[1] Department of Biochemistry, LKS Faculty of Medicine, The University of Hong Kong

[2] Department of Psychiatry, LKS Faculty of Medicine, The University of Hong Kong

*Email: junwen@hkucc.hku.hk Tel: (852) 2831 5075 Office: 1-05 E, Human Research Institute, 5 Sassoon Road, HK

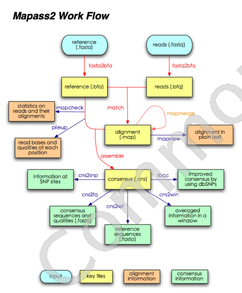
Introduction

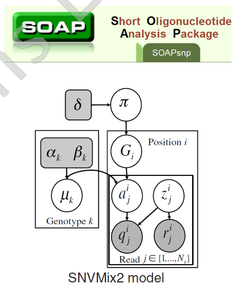
The advent of Next Generation Sequencing (NGS) technology has significantly advanced the sequence-based genomic research and its downstream applications, which include, but not limit to, metagenomics, epigenetics, gene expression, RNA splicing and RNA-seq and ChIP-seq. Alignment and SNPs discovery are two major procedures in NGS data analysis.

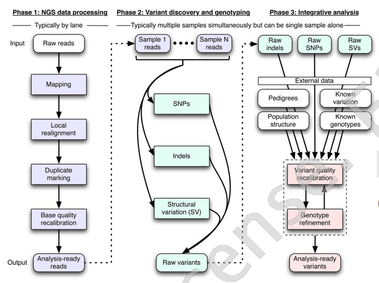
Table 1 | Summary of the representative software tools

Program	Version	Algorithm	Color-space supported	Read length(bp) supported	Gapped	pair-end supported	Can output all(suboptimal) hits	output format
Bowtie	0.12.7	FM-index	Yes	<=1024	no	yes	yes	SAM
BWA	0.5.8c	FM-index	Yes	Arbitrary	yes	yes	yes	SAM
SOAP2	2.2	FM-index	No	<=1024	no	yes	yes	SOAP2
RMAP	2.0.5	hash reads	No	Arbitrary	no	yes	yes	BED
ZOOM	1.5.0	hash reads	Yes	<=240	yes	yes	yes	ZOOM
Maq	0.7.1	hash reads	Yes	<=127	yes	yes	no	Maq
Novoalign	2.07.00	hash ref.	yes[a]	Arbitrary	yes	yes	yes	SAM
SHRiMP	2.1.0	hash ref.	Yes	Arbitrary	yes	yes	yes	SAM

[a]NovoalignCS supports the SOLiD platform







These softwares perform better at locus with higher sequence depth. But when the sequence depth is lower than 10, their performance decreases sharply. To deal with these issues, we developed a novel algorithm, FaSD, to call SNP based on only the bam or pileup file generated from the standard NGS analysis pipeline. We compared our model with existing softwares on both cancer and normal tissues from TCGA project. Assessed by Illumina and Affymetrix SNP arrays, we found that our model has higher accuracy on SNP calling, especially when the depth of sequencing data is low.

Methods

Datasets

Blood derived normal tissue (TCGA-06-0188-10B-01D-0373-08)
Primary tumor tissue (TCGA- 06-0188-01A-01D-0373-08)
Serous Cystadenocarcinoma sample(TCGA-13-0720-01A-01D-0445-10)
Yoruba individual (NA19240)
40 CEU individuals(NA12878, NA12891, NA12892,…)

FaSD model

$$\mathrm{FaSD\_Score} = -\mathrm{alternative\_score} \times \frac{\sum_{i=1}^{\mathrm{Depth}} \log_2\left(P_{(\mathrm{read}_i/\mathrm{ref})}\right)}{\mathrm{Depth}}$$

We used FaSD to call SNPs for each aligned position. The higher the FaSD_Score was, the more probable that the site might be a SNP position.

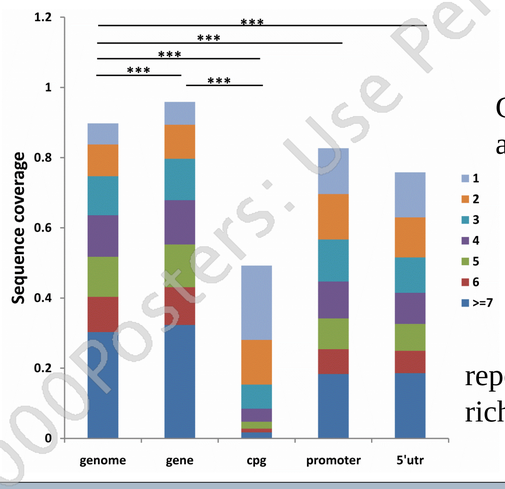
$$\mathrm{Alternative\_Score} = \begin{cases} 0 + \mathrm{pseudo\_score}, \text{when} \binom{N}{m} (0.999)^m (0.001)^{N-m} \text{is max} \\ 1 + \mathrm{pseudo\_score}, \text{when} \binom{N}{m} (0.500)^m (0.500)^{N-m} \text{is max} \\ 2 + \mathrm{pseudo\_score}, \text{when} \binom{N}{m} (0.001)^m (0.999)^{N-m} \text{is max} \end{cases}$$

N was the depth of the reads, and n was the occurrence of reference allele at the position. We added a pseudo_score to avoid Alternative_Score = 0. By default, we used pseudo_score = 0.01.

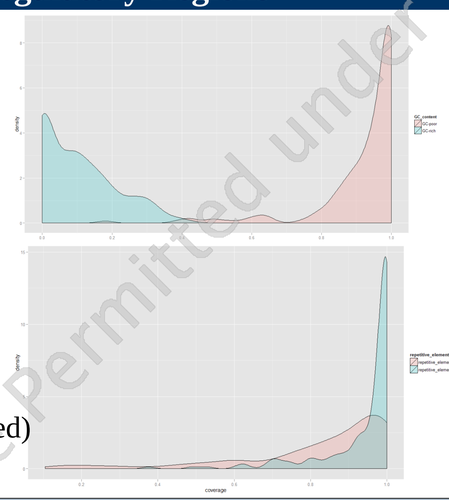
Performance Assessment of Aligners

Program	Category	Version	Index time (h:m:s)	Peak Memory footprint (gigabyte)	Alignment time (h:m:s)	Peak memory footprint (gigabyte)	Reads aligned (%)
Bowtie[a]	BWT	0.12.7	3:43:36	5.5	2:22:36	2.9	67.55
BWA[b]		0.5.8c	1:46:42	1.5	8:24:12	5.0	72.99
SOAP2[c]		2.20	1:45:54	2.3	10:22:26	6.8	60.93
RMAP[d]	Hash reads	2.0.5	N/A[e]	N/A	10:15:18	10.0	55.98
ZOOM[f]		1.5.0	N/A[e]	N/A	7:01:53	10.2	62.86
Maq[g]		0.7.1	0:01:56	0.34	39:10:43	8.1	71.94
Novoalign[h]	S-W	2.07.06	0:06:28	13.5	144:25:35	13.1	77.65
SHRiMP[i]		2.1.0	4:08:13	12.0	1065:10:05	12.0	81.23

Lower Sequence Coverage in the Regulatory Regions



GC–rich (green) and –poor (red)



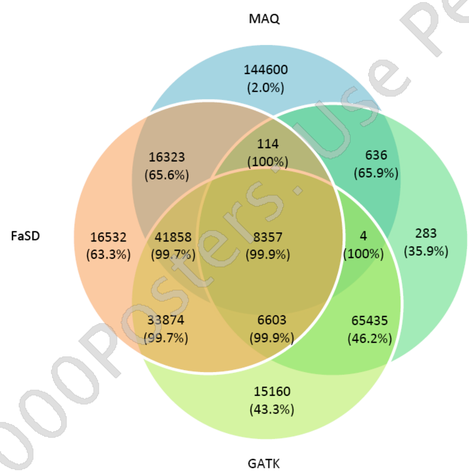
repetitive elements–rich (green) –poor (red)

Performance Assessment of SNP Callers

Performance evaluation on SNPs covered by arrays

	Illumina	Affymetrix	FaSD	MAQ	SOAPsnp	SNVmix2	GATK
Affymetrix	0.997(0.996)						
FaSD	0.882(0.927)	0.891(0.926)					
MAQ	0.397(0.401)	0.436(0.435)	0.449(0.430)				
SOAPsnp	0.417(0.409)	0.437(0.434)	0.449(0.430)	0.997(0.996)			
SNVmix2	0.157(0.182)	0.251(0.277)	0.274(0.290)	0.733(0.778)	0.733(0.779)		
GATK	0.804(0.842)	0.839(0.875)	0.848(0.857)	0.476(0.465)	0.486(0.475)	0.312(0.315)	
Bcftools	0.865(0.898)	0.905(0.928)	0.958(0.960)	0.508(0.465)	0.503(0.453)	0.352(0.336)	0.979(0.975)

The first number in each cell is the concordance between corresponding SNP callers in the normal datasets, the number in the parentheses is the concordance in the tumor datasets. The average depth of both normal and tumor datasets was 10X.



The number in each cell is the number of SNPs in the corresponding category. The percentage under the number is the proportion of SNPs that were confirmed by the Affymetrix SNP array. The FaSD, GATK, Bcftools, and MAQ called 123661, 171291, 81432, and 211892 SNPs in total, respectively. The average depth of this dataset was 10X.

Performance evaluation on SNPs not covered by arrays

	High_MAQ	FaSD	MAQ	SOAPsnp	GATK
FaSD	0.419±0.002				
MAQ	0.271±0.001	0.267±0.001			
SOAPsnp	0.266±0.001	0.264±0.001	0.981±0.001		
GATK	0.415±0.001	0.626±0.002	0.315±0.001	0.308±0.001	
Bcftools	0.383±0.001	0.613±0.002	0.295±0.001	0.293±0.001	0.681±0.002

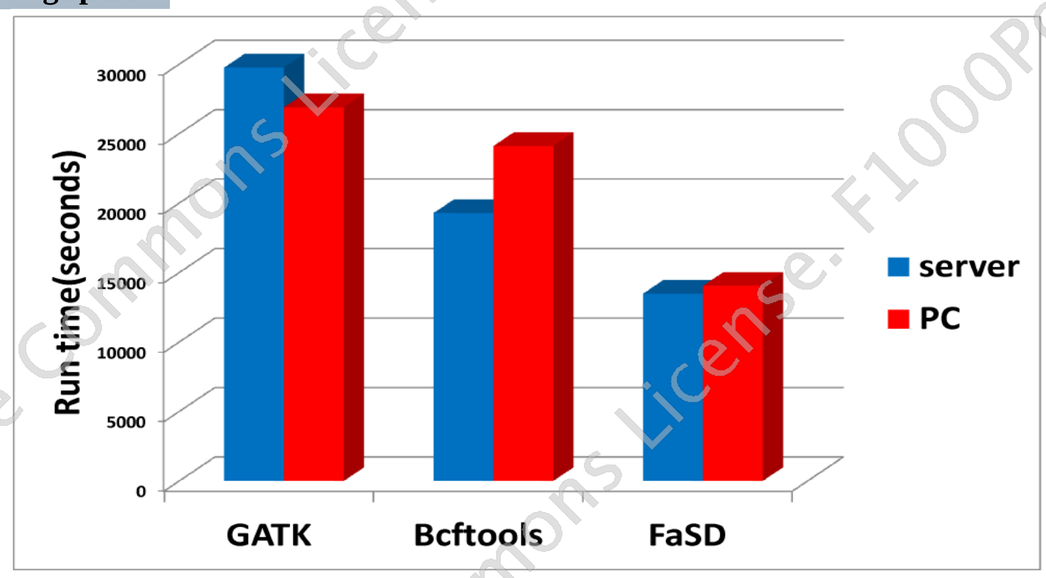
The number in each cell is the mean of non-reference concordance and standard deviation. The average depth of this dataset was 4X. High_MAQ represents the high-depth data called by MAQ, and is the benchmark.

Performance evaluation on pooled data

	High_MAQ	FaSD	GATK
FaSD	0.557(0.556)		
GATK	0.489(0.379)	0.637(0.641)	
Bcftools	0.535(0.353)	0.573(0.673)	0.520(0.603)

The first number in each cell is the non-reference concordance on the basis of pooled data, the number in the parentheses is the non-reference concordance based on the corresponding individual low coverage dataset. High_MAQ was used as the benchmark.

Processing speed



The average depth of this tumor dataset was 10X (30 gigabases).

References

1. P. C. Ng and S. Henikofghf, *Annu Rev Genomics Hum Genet* **7**, 61 (2006).
2. B. C. Kim, W. Y. Kim, D. Park et al., *BMC Bioinformatics* **9 Suppl 1**, S2 (2008).
3. J. O. Yang, W. Y. Kim, and J. Bhak, *Hum Mutat* **30** (12), E1010 (2009).
4. M. Hariharan, V. Scaria, and S. K. Brahmachari, *BMC Bioinformatics* **10**, 108 (2009).
5. H. Li, J. Ruan, and R. Durbin, *Genome Res* **18** (11), 1851 (2008).
6. R. Li, Y. Li, X. Fang et al., *Genome Res* **19** (6), 1124 (2009).
7. R. Li, Y. Li, K. Kristiansen et al., *Bioinformatics* **24** (5), 713 (2008).
8. R. Goya, M. G. Sun, R. D. Morin et al., *Bioinformatics* **26** (6), 730 (2010).
9. R. Sachidanandam, D. Weissman, S. C. Schmidt et al., *Nature* **409** (6822), 928 (2001).
10. Z. Zhao and E. Boerwinkle, *Genome Res* **12** (11), 1679 (2002).