The advent of Next Generation Sequencing (NGS) technology has significantly advanced the sequence-based genomic research and its downstream applications; which include, but not limited 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.

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 (NA19250)

40 CEL individual(NA12878, NA12891, NA12892, ...)

**FaSD model**

\[ \text{FaSD.Score} = \sum_{\text{Depth}} \log \left( \frac{1}{P_{\text{ref}}/P_{\text{alt}}} \right) \]

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.

\[
\begin{align*}
0 & : \frac{\text{pseudo.score, when }}{\text{N} \text{(0.999)}} \text{ (0.001)}^{N} \text{ is max} \\
1 & : \frac{\text{pseudo.score, when }}{\text{N} \text{(0.500)}} \text{ (0.500)}^{N} \text{ is max} \\
2 & : \frac{\text{pseudo.score, when }}{\text{N} \text{(0.999)}} \text{ (0.001)}^{N} \text{ is max}
\end{align*}
\]

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**

<table>
<thead>
<tr>
<th>Program</th>
<th>Category</th>
<th>Version</th>
<th>Interface</th>
<th>Indexrate (kb/m)</th>
<th>Peak memory (gigabyte)</th>
<th>Alignment (kb/m)</th>
<th>Peak memory (gigabyte)</th>
<th>Result (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bowtie 2</td>
<td>BWT</td>
<td>0.12.7</td>
<td>3.4-3.6</td>
<td>1.2-2.2.2</td>
<td>2.2-3.5</td>
<td>2.2-3.5</td>
<td>67.7</td>
<td>0.01</td>
</tr>
<tr>
<td>DRA</td>
<td></td>
<td>0.5.8</td>
<td>1.46-4.2</td>
<td>1.5-2.4-12</td>
<td>5.0-7.9</td>
<td>7.9-12-7</td>
<td>40.9</td>
<td>0.01</td>
</tr>
<tr>
<td>SOAP2</td>
<td>MAQ</td>
<td>2.16.8</td>
<td>3.26-3.6</td>
<td>1.5-6.3</td>
<td>4.8-10</td>
<td>9.8-11</td>
<td>72.9</td>
<td>0.01</td>
</tr>
<tr>
<td>Bcftools</td>
<td></td>
<td>2.07.0</td>
<td>2.13.145</td>
<td>1.5-13.5-2.8</td>
<td>10-12-2.1</td>
<td>12-12-2.1</td>
<td>60.9</td>
<td>0.01</td>
</tr>
<tr>
<td>GenomeBam</td>
<td>NGX</td>
<td>2.0.5</td>
<td>N/A</td>
<td>5.1-10.18-1.6</td>
<td>10.0-32</td>
<td>11.0-42</td>
<td>55.9</td>
<td>0.01</td>
</tr>
<tr>
<td>SOAP2</td>
<td></td>
<td>2.0.5</td>
<td>1.94-2.0</td>
<td>1.5-16.3-1.6</td>
<td>10.0-32</td>
<td>11.0-42</td>
<td>55.9</td>
<td>0.01</td>
</tr>
<tr>
<td>SOAP2</td>
<td></td>
<td>2.0.5</td>
<td>1.94-2.0</td>
<td>1.5-16.3-1.6</td>
<td>10.0-32</td>
<td>11.0-42</td>
<td>55.9</td>
<td>0.01</td>
</tr>
<tr>
<td>SOAP2</td>
<td></td>
<td>2.0.5</td>
<td>1.94-2.0</td>
<td>1.5-16.3-1.6</td>
<td>10.0-32</td>
<td>11.0-42</td>
<td>55.9</td>
<td>0.01</td>
</tr>
<tr>
<td>SOAP2</td>
<td></td>
<td>2.0.5</td>
<td>1.94-2.0</td>
<td>1.5-16.3-1.6</td>
<td>10.0-32</td>
<td>11.0-42</td>
<td>55.9</td>
<td>0.01</td>
</tr>
</tbody>
</table>

**Lower Sequence Coverage in the Regulatory Regions**

- GC-rich (green)
- GC-poor (red)
- repetitive elements-rich (green) - poor (red)

**References**

6. R. Li, Y. Li, X. Fang et al., *Genome Res* 19 (6), 1124 (2009).