How-To: SNP and INDEL detection

April 23, 2014
Lumenogix NGS SNP and INDEL detection

Mutation Analysis
Identifying known, and discovering novel genomic mutations, has been one of the most popular applications of next generation sequencing (NGS) technologies. In this Case Study, we describe several approaches for performing identification of mutations, including single nucleotide polymorphisms (SNPs), short insertions and deletions (INDELs) as well as other genomic variations such as intermediate and long INDELs, tandem duplications, substitutions and inversions. We also compare different mutation discovery algorithms, discuss the issues of mutation analysis specific to haploid organisms and describe the process of mutation detection using RNASeq data.

Mutation Detection Process
In most cases, there are 4 basic steps for detecting SNPs and short INDELs using NGS data:

1. The raw samples should be examined using a quality assessment tool. For this purpose, we tend to rely on FASTQC (unpublished, cited by [8], which provides easy-to-use numeric and graphical assessment of each sample’s quality.

2. The samples should be mapped (aligned) to a reference genome. Several popular tools are available. Unfortunately, these tools do not always produce equivalent results. For more information, please see our previous Case Study: Impact of Mapping Algorithms on SNP Discovery that compared SNP detection results using 3 different mapping algorithms bowtie [2], bowtie2 [1] and BWA [3].

3. After mapping, we recommend performing assessment of coverage across either the entire genome (if doing whole genome mutation analysis) or across the specific enrichment subset (in most cases, this subset is either the exome or a gene enrichment panel).

4. Once the samples are mapped, detection of SNPs and INDELs can be performed using several available algorithms. In this Case Study, we will take a look at the difference between two of the most popular tools GATK [5] and Samtools [4].

Several additional complexities will be discussed:
1. The impact of removing optical duplicates and performing realignment.

2. Working with haploid species.

3. Performing mutation analysis on RNASeq data.

4. Detecting genomic variations other than SNPs and short INDELs.

**Quality Assessment**

While most contract research organizations (CROs) and core facilities will provide a quality report with each data set, it is important to perform an independent, unbiased review of the data. Integrated within Lumenogix NGS, is FASTQC, a fast and convenient quality assessment tool. Please also note that performing an in-depth quality check will not only help to double-check the vendors quality assessment, but also can help to guide the next step mapping of the reads to reference genome. For example, the data in Figure 1 were generated by a sequencing CRO based in Asia.

Note that the quality of the bases is very high until approximately base 60. If we want to obtain only the highest-quality data (and for SNP detection quality is extremely important), we may want to map only the first 60 bases of each read and discard the last 40. While this may significantly reduce our coverage (in the current example, we would lose up to 40% of our data), it also would significantly increase reliability of the mutations that we do detect.

Within Lumenogix NGS, trimming of the reads prior to mapping is extremely easy. On the request submission screen (Figure 2), simply indicate how many bases are to be trimmed, and from which side of the reads. In addition to checking data quality using FASTQC, it also is helpful to map a subset of reads of each sample to NCBI’s nucleotide collection (nr/nt) [7].

This can help to detect if the samples have been inadvertently mixed up by the vendor or core facility. While these mix-ups are rare, they can waste a lot of research time and money, if not detected early. Lumenogix NGS automatically aligns a small number of randomly selected reads to NCBI’s databases during the initial QC request (Figure 3).

Performing mapping using Lumenogix NGS is straightforward; the user needs to select the reference genome, set parameters and, if appropriate, arrange sample pairing and provide pair-related information (Figure 4).
Figure 1: Quality score across all bases

Figure 2: Submitting mapping request with quality trimming
Figure 3: Lumenogix NGS QC results  FASTQC and BLAST information

Figure 4: Arranging paired samples for mapping
Mapping Reads to the Genome

In recent years, mapping of NGS reads to the genome has become a standard service offered at no additional cost by most CROs and core facilities that provide a mapped BAM [4] file as a deliverable. (The costs of the mapping are built into the sequencing service). However, we believe that mapping to the genome is an important step and should be performed with care. Clearly, which reference genome is used to map the reads is important. It is also important to consider which version of the genome is used for mapping. NCBI releases updates (revisions) of the human genome assembly every few months. The latest release notes of the human genome assembly can be found here. And, as was shown earlier, information generated in the QC step can help select appropriate mapping parameters. We also should remember that algorithm used to map reads to genome also will impact the final results (Table 1).

Assessing Coverage

Once the reads have been mapped, we can now assess actual coverage—for a given base in the reference, how many reads are present at that position. Please note that actual coverage can be substantially different from the estimated coverage based on the quantity of data generated. This discrepancy is due to multiple factors:

1. The overall quality of the data was increased by removing some reads from the data set.

<table>
<thead>
<tr>
<th>Mapping tool</th>
<th>Reads on Target</th>
<th>SNPs called</th>
<th>Novel SNPs</th>
<th>Novel high-quality SNPs</th>
<th>INDELs called</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bowtie</td>
<td>55%-77%</td>
<td>370</td>
<td>207</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>Bowtie2</td>
<td>98%-99%</td>
<td>5,821</td>
<td>5,337</td>
<td>353</td>
<td>239</td>
</tr>
<tr>
<td>BWA</td>
<td>76%-90%</td>
<td>690</td>
<td>557</td>
<td>33</td>
<td>268</td>
</tr>
<tr>
<td>Ion Torrent</td>
<td>94%-98%</td>
<td>97</td>
<td>NA</td>
<td>NA</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1: Mapping results summary
2. Sequencing does not produce reads that are evenly distributed across the genome (or the target enrichment). Some areas of the genome may be more difficult to sequence so they may be under-represented (for example GC-rich regions). Other areas, especially repeated regions, may be over-represented.

3. As we can see from Table 1, different mapping tools will result in a different number of reads being mapped to target, which will result in a different coverage.

4. Number of reads aligned to target also can be impacted by mapping parameters. Setting looser criteria will result in more mapped reads; setting more stringent criteria will result in fewer reads on target (but higher quality results).

A typical coverage report provides information about the percentage of the genome (or a selected subset) with 100x, 20x and lower coverage (Figure 5). Additional usefulness of the coverage assessment is the availability of the No Coverage report. This may help to identify large and very large deletions.
Mutation Detection

Once all mapping and assessment has been completed, the final step is SNP/INDEL detection. At this time, the two most popular SNP detection tools are Samtools and GATK; both have been fully integrated into Lumenogix NGS. We will look at the differences that are due to different mutation detection tools. The data for this analysis came from human exome, captured by Agilent SureSelect v4+UTR, mapped with BWA. In both cases, duplicates were removed using Picard and realignment was done using GATK. We further filtered the SNPs selecting only those that were novel, meaning not in dbSNP [9], and had coverage, SNP quality, genotype quality and mapping quality $\geq 20$.

The selection was made using the Lumenogix NGS filter function. Within Lumenogix NGS, the initial SNP analysis detects all possible variations and then the built-in filter function can be applied to select the appropriate data set (Figure 6). This approach allows researchers to keep direct control over the collection of SNPs with which they are working.

In this case (Figure 7, Table 2), GATK detected about three times as many SNPs as Samtools; however, there does exist a small number of high quality SNPs that are detected by Samtools but not GATK.
Figure 7: Comparing high quality SNPs detected using Samtools and GATK

<table>
<thead>
<tr>
<th>Detection tool</th>
<th>SNPs detected</th>
<th>Novel high-quality SNPs</th>
<th>Unique SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Samtools</td>
<td>562,620</td>
<td>1,794</td>
<td>136</td>
</tr>
<tr>
<td>GATK</td>
<td>550,870</td>
<td>4,561</td>
<td>2,903</td>
</tr>
</tbody>
</table>

Table 2: Comparing SNP numbers between Samtools and GATK
SNP and INDEL How-To  

A Lumenogix White Paper

<table>
<thead>
<tr>
<th>Detection tool</th>
<th>INDELS detected</th>
<th>High-quality INDELS</th>
<th>Unique INDELS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Samtools</td>
<td>30,896</td>
<td>6,135</td>
<td>1,603</td>
</tr>
<tr>
<td>GATK</td>
<td>9,923</td>
<td>6,790</td>
<td>2,258</td>
</tr>
</tbody>
</table>

Table 3: Comparing INDEL numbers between Samtools and GATK

Unlike SNPs, Samtools appears to detect approximately the same number of high quality short INDELs as GATK (Figure 8, Table 3), at least in this particular data set. High quality INDELs are defined as INDELs with coverage ≥ 20 and mapping quality ≥ 20.

Unlike SNPs, both Samtools and GATK detect a roughly equivalent number of high quality INDELs with both tools having about 30% (26% for Samtools and 33% for GATK) of the INDELs unique to that algorithm.

The Impact of optical duplicate removal and realignment on SNP and INDEL discovery

We also looked at the impact made by applying the duplicate removal and realignment procedure to the mapped reads prior to calling SNPs. In this case, we also used BWA mapped data and defined high quality SNPs as
### Table 4: Impact on duplicate removal and realignment on SNP discovery

<table>
<thead>
<tr>
<th>Processing</th>
<th>SNPs</th>
<th>High-quality novel SNPs</th>
<th>INDELs</th>
<th>High quality INDELs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Duplicate removal only</td>
<td>551,876</td>
<td>4,759</td>
<td>9,008</td>
<td>6,324</td>
</tr>
<tr>
<td>Realignment only</td>
<td>577,083</td>
<td>5,054</td>
<td>10,926</td>
<td>7,426</td>
</tr>
<tr>
<td>Neither</td>
<td>578,542</td>
<td>5,175</td>
<td>9,851</td>
<td>6,907</td>
</tr>
<tr>
<td>Both</td>
<td>550,870</td>
<td>4,561</td>
<td>9,923</td>
<td>6,790</td>
</tr>
</tbody>
</table>

Looking at the results without preprocessing (meaning without duplicate removal and realignment), we see a substantial difference between the high quality novel SNPs identified with and without preprocessing (Figure 9). This is an indication of the number of false-positives removed by processing the data prior to SNP/INDEL calling. In the case of INDEL discovery, the results are similar, although with a smaller difference (Figure 10).

Comparing SNP and INDEL discovery results based on partial preprocessing (duplicate removal only and realignment only), shows that the artifacts are removed by both procedures (Figures 11 and 12), although realignment makes a more significant contribution especially for INDEL discovery.

### Working with Non-diploid Organisms

While extremely popular tools, it should be noted that both Samtools and GATK were originally developed to handle human (diploid) data. To handle non-diploid data, Lumenogix NGS also offers FreeBayes, an algorithm developed to handle haploid and polyploid organisms. As in the case of Samtools and GATK, removal of duplicates and realignment preprocessing options are available for FreeBayes. As expected, the three SNP calling algorithms do select a somewhat different collection of mutations from the same data (Table 5). In this case, we used a Staphylococcus aureus data set. Please note that high quality SNPs are those with coverage, mapping quality, SNP quality and genotype quality $\geq 20$ and novel SNPs are those that are not in dbSNP. High quality INDELs are defined as those INDELs with coverage and mapping quality $\geq 20$. 

those having coverage, mapping quality, genotype quality and SNP quality $\geq 20$. Novel SNPs are defined as those SNPs that are not in dbSNP. High quality INDELs are defined as those INDELs with coverage and mapping quality $\geq 20$. 

Looking at the results without preprocessing (meaning without duplicate removal and realignment), we see a substantial difference between the high quality novel SNPs identified with and without preprocessing (Figure 9). This is an indication of the number of false-positives removed by processing the data prior to SNP/INDEL calling. In the case of INDEL discovery, the results are similar, although with a smaller difference (Figure 10).

Comparing SNP and INDEL discovery results based on partial preprocessing (duplicate removal only and realignment only), shows that the artifacts are removed by both procedures (Figures 11 and 12), although realignment makes a more significant contribution especially for INDEL discovery.

### Working with Non-diploid Organisms

While extremely popular tools, it should be noted that both Samtools and GATK were originally developed to handle human (diploid) data. To handle non-diploid data, Lumenogix NGS also offers FreeBayes, an algorithm developed to handle haploid and polyploid organisms. As in the case of Samtools and GATK, removal of duplicates and realignment preprocessing options are available for FreeBayes. As expected, the three SNP calling algorithms do select a somewhat different collection of mutations from the same data (Table 5). In this case, we used a Staphylococcus aureus data set. Please note that high quality SNPs are those with coverage, mapping quality, SNP quality and genotype quality $\geq 20$ and novel SNPs are those that are not in dbSNP.
Figure 9: Comparing SNP discovery results with and without preprocessing

Figure 10: Comparing INDEL discovery results with and without preprocessing
Figure 11: Comparing the SNPs discovered using partial preprocessing

Figure 12: Comparing INDELs discovered with partial preprocessing
SNP and INDEL How-To  
A Lumenogix White Paper

<table>
<thead>
<tr>
<th>Detection tool</th>
<th>SNPs detected</th>
<th>Quality novel SNPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Samtools</td>
<td>196</td>
<td>161</td>
</tr>
<tr>
<td>FreeBayes</td>
<td>334</td>
<td>277</td>
</tr>
<tr>
<td>GATK</td>
<td>299</td>
<td>122</td>
</tr>
</tbody>
</table>

Table 5: Comparing number of bacterial SNPs discovered by Samtools, GATK and FreeBayes

the dbSNP.

**Mutation detection using RNASeq data**

While mutation detection is normally done using DNASeq data, it is also possible to perform mutation detection on RNASeq data. (Meaning data that were generated by sequencing the samples RNA.) The RNASeq reads are first mapped to the genome, in this case utilizing a splice junction mapping tool such as TopHat or TopHat 2.0 [10]. Once the reads are mapped, Samtools, GATK and FreeBayes are available for mutation analysis in Lumenogix NGS. Launching the analysis is a simple operation that can be done using a point-and-click interface on any Internet-enabled device, including smartphones and tablet computers (Figure 13).

**Detecting other genomic variations**

It should be stressed that the popular mutation detection algorithms, such as Samtools and GATK, have all been optimized for SNP discovery. While they are also capable of detecting insertions and deletions, those will be fairly short  typically less than 10 nucleotides. To detect longer INDELs, as well as other genomic variations (GVs), alternative algorithms must be utilized. Most of these algorithms take advantage of the paired-reads configuration (see appendix) and are based on analysis of the difference between expected distance between paired reads and actual (mapped) distance. A good review of using paired reads for discovery of structural variants is available here: [6]. For handling identification of intermediate and long insertions and deletions, as well as substitutions, tandem duplications and a few other genomic break points, Lumenogix NGS integrates Pindel [11]. Running Pindel requires mapping of paired DNASeq data using BWA (unfortunately, this analysis is not available for RNASeq data). Launching the request is another simple point-and-click command (Figure 14).
Figure 13: Launching mutation analysis (GATK) from mapped RNASeq data

Figure 14: Launching analysis to detect genomic variations using paired reads


