NextGENe®
2nd Generation Sequence Analysis Software

NextGENe® Software Analysis of Solid Tumors and Hematological Cancers Using RainDance ThunderBolts™ NGS Panels

April 2015

Authors

John McGuigan[1], Kevin LeVan[1],
Ni Shouyong[1], CS Jonathan Liu[1], Jeff Olson[2],
Omoshile Clement[2], Dan Aiello[2].

[1]SoftGenetics Inc.
100 Oakwood Ave
State College, PA 16803, USA.

[2]RainDance Technologies
749 Middlesex Turnpike
Billerica, MA 01821, USA.

Introduction

A major challenge facing today's cancer researchers is the ability to have highly accurate, rapid, simple, and low-cost tools to identify genetic variants underlying specific phenotypes found in solid tumors and hematological cancers. The combination of the RainDance ThunderBolts™ System, the ThunderBolts NGS gene panels, and SoftGenetics NextGENe® Software enables targeted sequencing for research of important oncogenes that have clinical relevance to cancer diseases. The solution features preloaded analysis templates for each gene panel and pre-defined settings for detecting minor allele frequencies at different sensitivity settings (1% and 5%).

The following application note describes the use of SoftGenetics' NextGENe software for the customized analysis workflow of sequencing data using the ThunderBolts Cancer Panel (6 samples - cell lines, plasma and FFPE) and the new ThunderBolts Myeloid Panel (16 samples - cell lines).

ThunderBolts NGS Enrichment Chemistry & Workflow

The ThunderBolts System features award-winning NGS target enrichment instrumentation, pre-validated NGS gene panels for profiling solid tumors and hematological cancers, as well as open source capabilities for custom assay development. The fast and simple workflow is optimized for Illumina NGS systems and compatible with all types of DNA (FFPE, tissue, plasma, marrow aspirates, or FNA) samples with as little as 10 ng of starting DNA. Researchers gain rapid, accurate mutation profiles for a low overall cost per sample.

The ThunderBolts assay chemistry applies RainDance's proprietary chip-based microfluidic technology to encapsulate PCR reagents, genomic template (amplifiable DNA) and a library of gene-targeted primer pairs in millions of picoliter-volume droplets. For these assays, the ThunderBolts System generates millions of single molecule PCRs reactions with ~8 million 5pL-sized droplets per well (8-well chip) encapsulating either wild type or mutant alleles in each droplet. This allows enrichment of rare allele variants to similar amplification levels as those of high abundance alleles, enabling ultra-sensitive detection of low percentage mutants in the presence of a high wild type background.

Figure 1: A generalized workflow for ThunderBolts NGS target enrichment.

For example, the ThunderBolts Cancer Panel targets mutational hotspots in 50 known oncogenes and tumor suppressors with 230 amplicons from as little as 10ng of amplifiable gDNA. In this example, each sample will yield 16 million droplets (2 wells per sample) containing a total of 10ng of input DNA. There are approximately 3,000 genomic equivalents (GE) per 10 ng of DNA. This results in there being 690,000 droplets with an amplifiable target (3,000 GE x 230 amplicons) or approximately 1 out of 23 droplets with an amplifiable target. The large number of droplets generated by the ThunderBolts System allows for truly digital PCR amplification. NGS-ready libraries are generated after a second PCR step (DirectSeq™ method) that eliminates sequencing library preparation steps by adding Illumina indexing barcodes for sample plexing on a MiSeq, NextSeq, or HiSeq flow cell(s).

NextGENe® Software

NextGENe version 2.4.1 introduces Autorun Templates for quick project setup using various types of library preparations. The ThunderBolts Cancer Panels and ThunderBolts Myeloid Panels are installed by default. They can be loaded into the NextGENe Autorun tool to quickly set up a job to process multiple samples. For each panel, NextGENe includes a template with 5% allele frequency cutoff and a "High Sensitivity" template with 1% allele frequency cutoff. These pre-defined templates cannot be modified, but they may be used to create new custom templates. After processing, researchers can view projects in the NextGENe Viewer and (optionally) exported to Geneticist Assistant (figure 2).

Figure 2: The NextGENe Template Workflow

Panel Templates Streamline Analysis Workflow

Panel templates offer an easier approach for targeted resequencing data in NextGENe. Templates are used to specify all settings for the analysis when creating ngjob files. These ngjob files are detected and processed by the Autorun tool.

1. Open the Autorun tool from the NextGENe "Tools" menu.
2. Open the "Job File Editor" from the Autorun "Tool" menu (Figure 3).

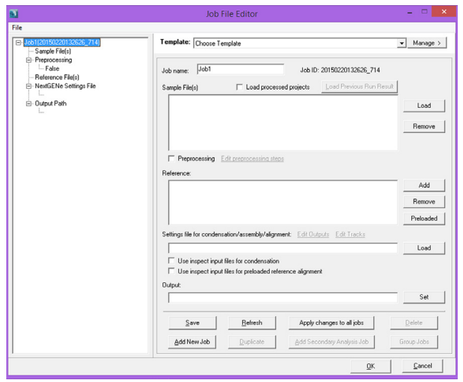


Figure 3 - The Autorun Job File Editor

3. Select a template from the drop-down. Templates for the ThunderBolts Cancer Panel and Myeloid Panel are included with the software. The "Manage" button allows the templates to be modified and saved as new templates, or for completely custom templates to be created.
4. Load the raw data (fastq files).

5. Select a preloaded reference to use for alignment.
6. Set an output location.
7. Click the “Group Jobs” button to open the grouping tool (Figure 4). This makes it easy to split the list of raw data files into separate sample projects.

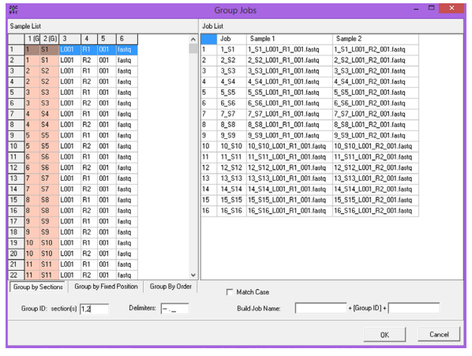


Figure 4 - The Grouping Tool

8. Click OK to save the ngjob file.
9. Set the autorun tool to check the directory containing the saved ngjob file. Start the autorun tool to begin processing.

The provided templates for ThunderBolts panels include multiple preprocessing steps (quality trimming, adapter trimming, and primer trimming) and settings for alignment and variant calling.

Rapid Results

Speed

16 ThunderBolts Myeloid Panel samples and six ThunderBolts Cancer Panel samples were processed. With enough RAM, it is possible to process multiple samples simultaneously by simply changing a setting in the autorun tool. This can greatly increase the processing speed, especially when using computers with multiple cores. The samples were completed in approximately three hours when run with up to four simultaneous jobs on a server computer with eight cores and 64 GB of RAM.

When run one-at-a-time on a laptop computer with a dual-core 2 GHz processor and eight GB of RAM the total processing time was approximately seven hours for the first set of samples and two hours for the second.

Alignment and Pre-processing

The total read counts (counting a pair of reads as two) in each sample at different stages of the analysis (raw data, after pre-processing, and after alignment) are shown in Figure 5. The pre-processing included an adapter trimming and quality trimming step in addition to primer trimming and filtering out reads shorter than 100 bp after primers were trimmed

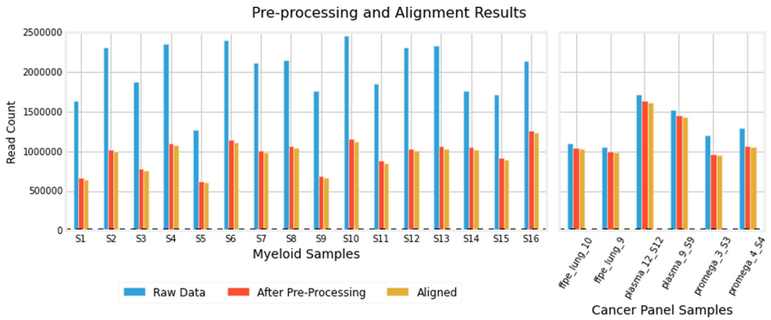
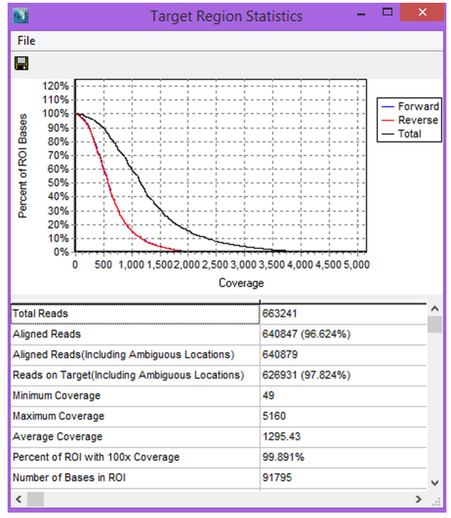


Figure 5

The coverage curve report makes it easy to examine the depth of coverage across all bases in the targeted amplicons. The graphical summary view of the coverage curve report for Myeloid sample “S1” is shown in Figure 6. It is easy to see that approximately 90% of the bases have at least 500x coverage, and over 99% of the bases have at least 100x coverage. Figure 7 shows a summary of the minimum, average, and maximum coverage across targeted bases in all of the samples.



Total Reads	663241
Aligned Reads	640847 (96.624%)
Aligned Reads(Including Ambiguous Locations)	640879
Reads on Target(Including Ambiguous Locations)	626931 (97.824%)
Minimum Coverage	49
Maximum Coverage	5160
Average Coverage	1295.43
Percent of ROI with 100x Coverage	99.891%
Number of Bases in ROI	91795

Figure 6 - Coverage Curve Summary Report

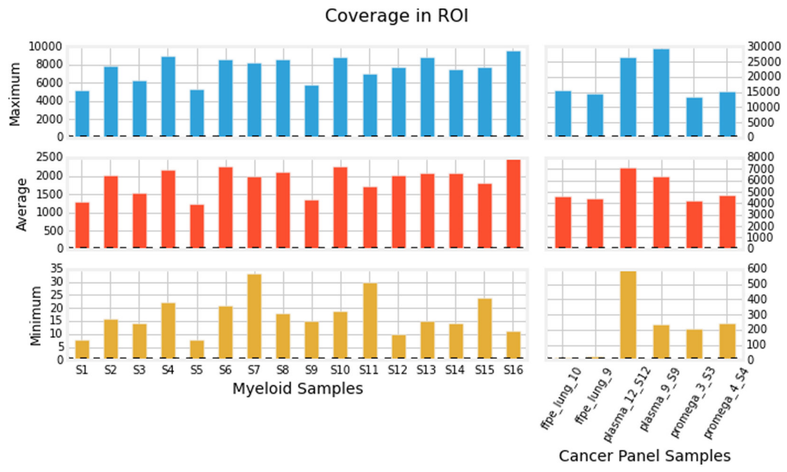


Figure 7 - Summary of Coverage in the ROI

Accurate Variant Calling

Figure 8 shows the number of variants called in each sample grouped by the type of variant (substitution, insertion or deletion). Samples within a given panel had similar numbers of variants- the Myeloid Panel is approximately three times larger than the Cancer Panel, likely explaining the difference between panels. The High Sensitivity panels will call additional low-level variants but additional false positives are likely.

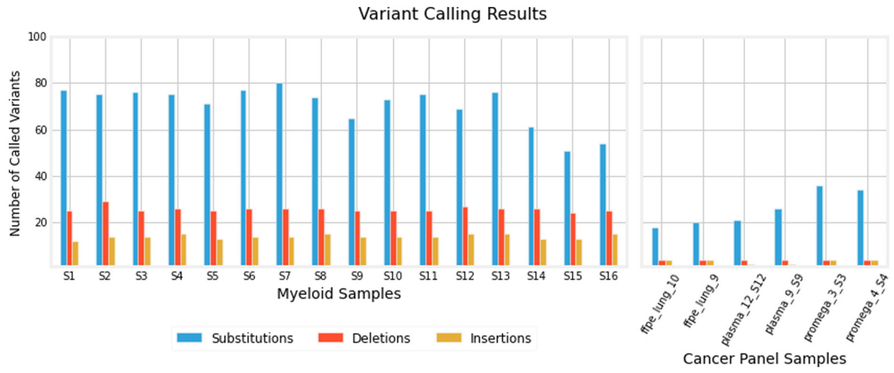


Figure 8

Discussion

There are several preprocessing steps performed on the data in each panel.

1. Format conversion (FASTQ to FASTA) with adapter sequence trimming, quality trimming (3 or more bases with a score <= 10), and quality filtering (median score < 20, more than 3 uncalled bases)
2. Trimming the 5' primer
3. Trimming the 3' primer and removing reads that are less than 100 bp long

After preprocessing the data is aligned to a whole genome reference. An attempt is made to match reads perfectly, then reads are aligned using a seed-based approach. As seen in figure 6 for sample "S1", this results in most reads aligning to the reference (over 96% in this sample) with most of these being aligned to targeted regions (over 97% in this sample). Reads are soft-clipped when multiple mismatches occur near the end of the read.

Default settings specify that all positions meeting these criteria are called as variants:

- At least 100x total coverage at the position
- At least 5 occurrences of the variant
- At least 5% variant frequency (1% in the "High Sensitivity" templates)
- Forward/Reverse balance is greater than 0.5 (0.8 for homopolymer indels) unless the variant is present at a frequency above 80%. The balance is calculated as the smaller value of forward/reverse or reverse/forward.
- The position is in the targeted regions.

All of the necessary files (primer sequences, adapter sequence, regions of interest) are included as part of the panel. Any setting(s) can be adjusted to create a new panel.

After processing, the projects can be visualized in the NextGENe Viewer (Figure 9). New reports can be created and saved in addition to those already created during processing, which include the mutation report, several coverage curve reports (with different coverage cutoffs), a distribution report, and an expression report (showing read counts per target). The track manager tool allows for optional tracks to be imported into the NextGENe reference folder. These tracks (including COSMIC[1], ClinVar [2], and dbNSFP [3]) can then be queried automatically for every new project or after a project has been run.

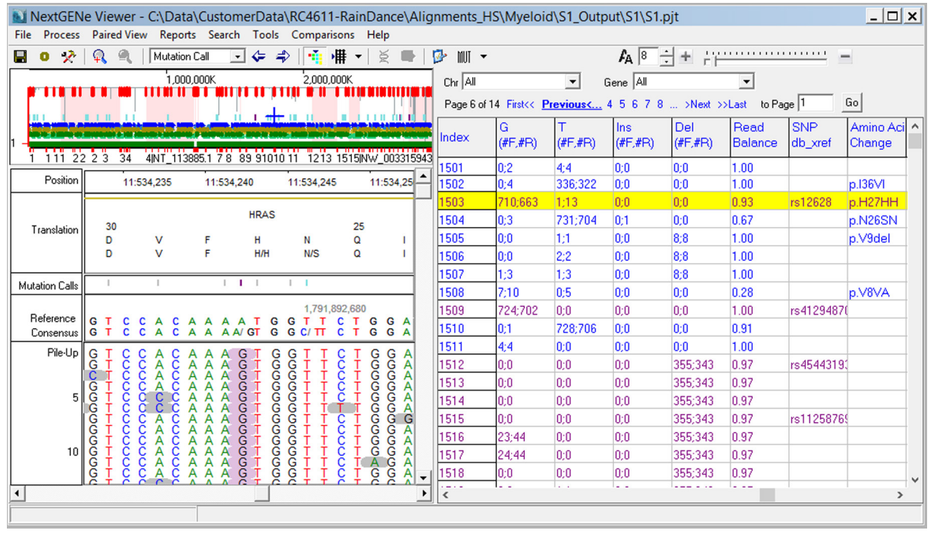


Figure 9 – A ThunderBolts Myeloid Panel project as seen in the Viewer

Projects may also be compared side-by-side in the Variant Comparison Tool. Figure 10 shows a comparison of all 16 Myeloid Panel samples with the same variant present in all 3 samples that are selected for visualization. This tool allows for filtering based on the presence or absence of variants in specific projects in addition to the many filters available in the single-project mutation report.

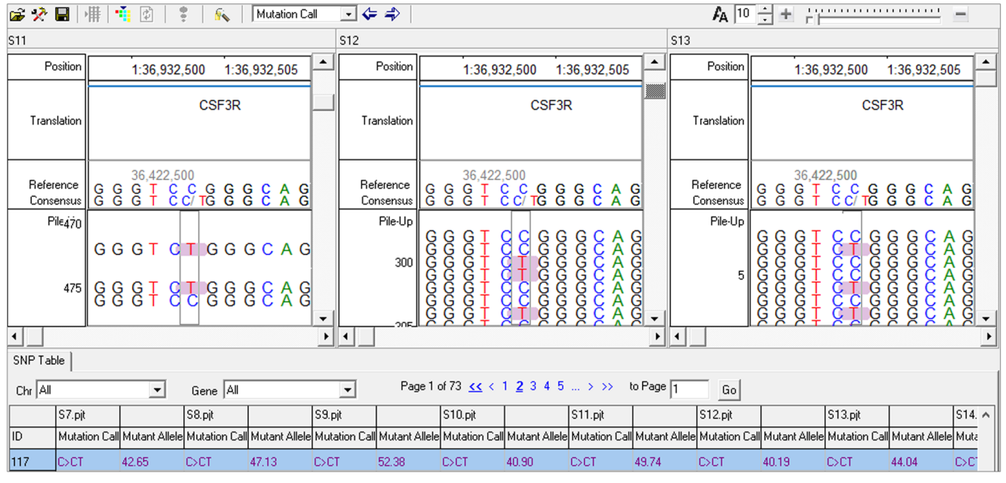


Figure 10 - Variant Comparison Tool

Most of the Myeloid samples in this analysis were created from a mixture of three different cell line samples. This was used to determine expected frequencies of known variants in the mixtures in order to compare to the detected frequencies. The left side of figure 11 shows this comparison for known SNPs that were detected using the high sensitivity template. Variants with expected frequencies between 1% and 5% were detected with good sensitivity, but using the high sensitivity template also introduces a number of false positives at low frequencies. The right side of figure 11 shows the number of false positive SNP calls in each sample that have a frequency above certain values. All tested samples had less than 10 false positive calls at the 5% level, and a much higher number of false positives with lower frequencies.

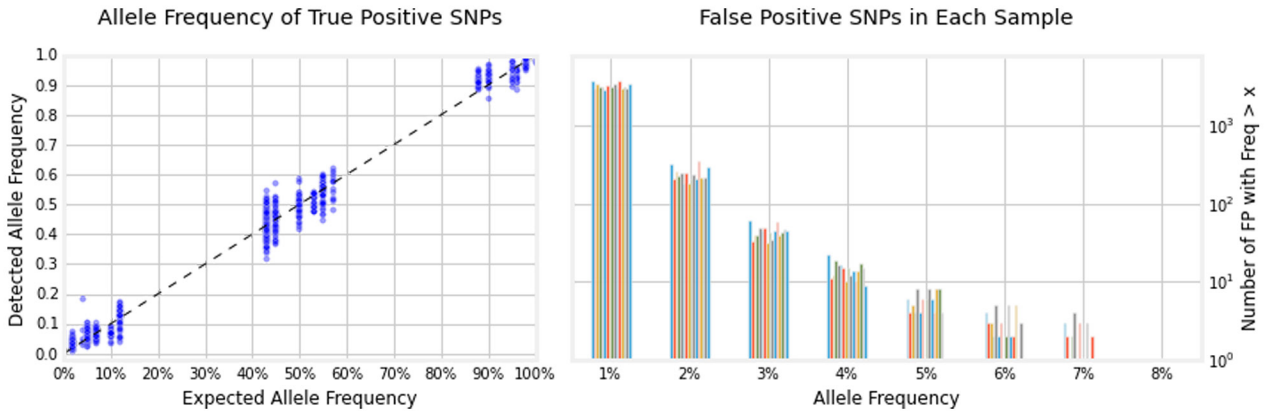


Figure 11 – Allele Frequencies for True Positive SNPs (left) and False Positive SNP Calls (right).

References

1. Forbes, Simon A., et al. "COSMIC: exploring the world's knowledge of somatic mutations in human cancer." Nucleic acids research 43.D1 (2015): D805-D811.
2. Landrum, Melissa J., et al. "ClinVar: public archive of relationships among sequence variation and human phenotype." Nucleic acids research (2013): gkt1113.
3. Liu, Xiaoming, Xueqiu Jian, and Eric Boerwinkle. "dbNSFP v2. 0: A Database of Human Non-synonymous SNVs and Their Functional Predictions and Annotations." Human mutation 34.9 (2013): E2393-E2402.