Analyzing Sequence Data from GS GType Targeted Sequencing of Leukemia-Associated Genes using NextGENe® Software

May 2012

John McGuigan, Kevin LeVan, Shouyong Ni, CS Jonathan Liu

Introduction

The GS GType RUNX1 and GS GType TET2/CBL/KRAS Primer Sets can be used to find genetic variations in four cancer-associated genes using the Roche GS FLX or Roche GS Junior sequencing systems. NextGENe is able to rapidly process the sequence data in only a few minutes on a typical desktop computer running a Windows operating system. Processing includes sorting based on MID sequences, alignment, and mutation calling. The quality information can be used to perform additional filtering and trimming on the data when starting from an SFF file. The entire workflow can be automated to run unattended using the NextGENe AutoRun tool.

Two datasets were processed in this analysis- one 12-sample run of the RUNX1 assay and one 3-sample run of the TET2/CBL/KRAS assay. An alignment of one sample is shown in figure 1, where a 12 bp deletion was found in the TET2 gene.

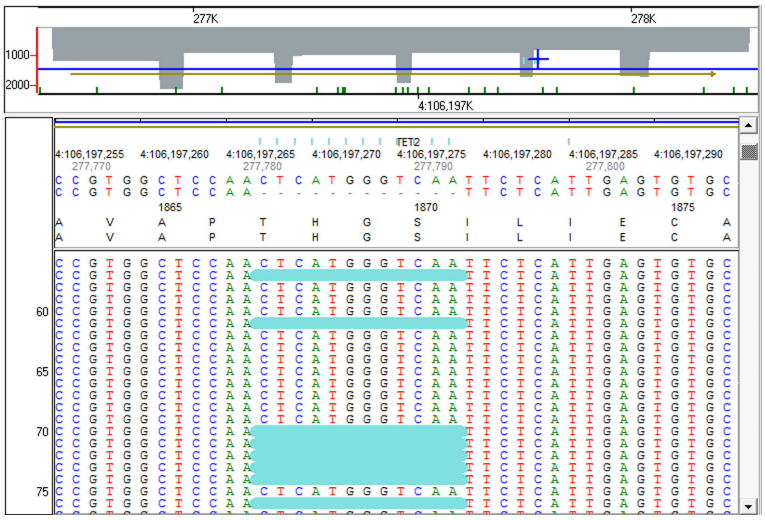

Figure 1: A 12 bp deletion in TET2 detected at approximately 41% frequency

Procedure

1. The sequence reads (SFF or FNA format) are sorted into individual projects bases on MID sequences using the barcode sorting tool (Figure 2). A log file records the results of sorting and MID trimming. Sorting SFF files will generate .qual files containing the basecalling quality scores in addition to the fna files containing the sorted and trimmed reads.

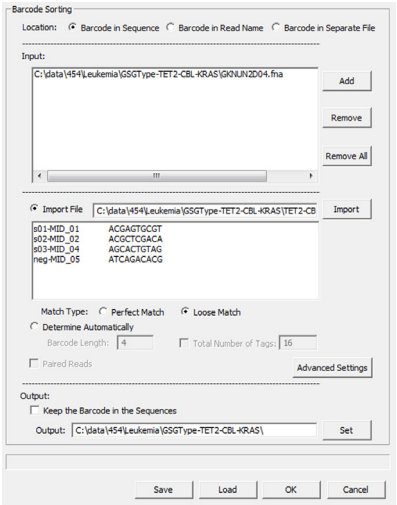

Figure 2: The barcode sorting tool. A tab-delimited text file specifying all four MID sequences was used to sort the FNA file into three samples- the last MID was for a negative control.

2. The sorted sequences are processed in the Sequence Operation Tool (Figure 3). The Sequence Trim function is used to remove primer sequences from the ends of the reads. The list of primer sequences is a tab-delimited text file where each line has a name, the forward primer, and the reverse primer. If an SFF file was used during sorting, the .fna and .qual files can be loaded into the format conversion tool for primer trimming instead and the quality scores can be used for further trimming and filtering of the reads.

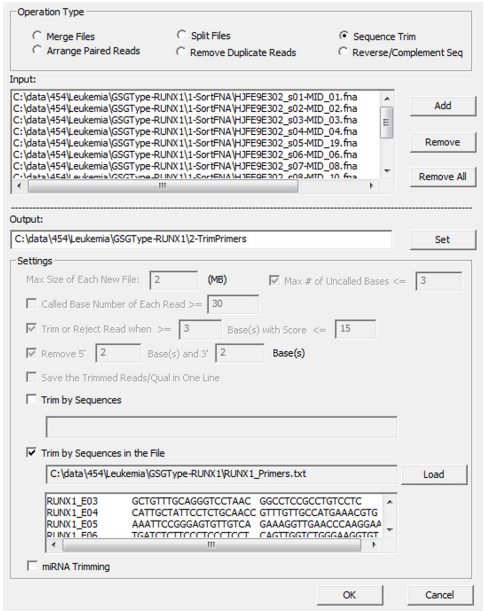

Figure 3: The Sequence Operation Tool

3. The reads are aligned to Genbank reference files and mutations are called. Alignment settings were adjusted from the default settings in order to improve sensitivity and specificity (Figure 4). At least 15 occurrences of the SNP must be found at a position with at least 250x coverage and at least 5% of the reads at that position must show the mutation. Homopolymer indel calls were removed if the forward/reverse balance was less than 0.5- the forward reads with the indel must be at least half the number of reverse reads, or the number of reverse reads must be at least half the number of forward reads. The Forward and Reverse Balance filter was also applied- the ratio of forward reads for the mutant allele must be at least 30% of the ratio of forward reads for the normal allele.

4. Mutation filtering settings in the NextGENe Viewer can be adjusted in order to further improve specificity. In this case only mutations within 5 bp of a coding sequence were kept after filtering in the viewer. The mutations were annotated with dbSNP135 (built into NextGENe references) and dbNSFP v1.1. The database of Non-synonymous Functional Predictions (dbNSFP) includes several different prediction scores [1]. NextGENe is able to import PhyloP, PolyPhen-2, SIFT, MutationTaster, LRT, and 1000 Genomes frequency information from this database. The next major release of NextGENe will include support for the Catalog of Somatic Mutation in Cancer (COSMIC) database [2].

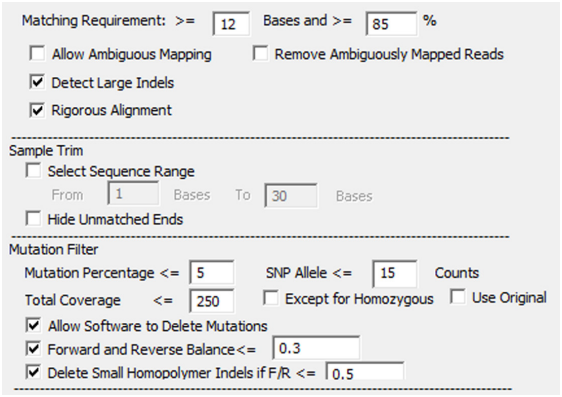

Figure 4: Alignment settings used in this analysis.

Results

Barcode sorting and alignment results are summarized in tables 1 and 2. The reads were distributed evenly among all of the samples and nearly all of the reads were aligned.

	01 MID-01	02 MID-02	03 MID-04
Sorted Reads	26,842	25,797	25,792
Percent of Total	34.22%	32.89%	32.88%
Aligned Reads	26,806	25,768	25,758
Aligned %	99.87%	99.89%	99.87%

Table 1: Analysis results for TET2/CBL/KRAS samples

	01 MID-01	02 MID-02	03 MID-03	04 MID-04	05 MID-19	06 MID-06	07 MID-08	08 MID-10	09 MID-13	10 MID-18	11 MID-15	12 MID-16
Sorted Reads	6,100	5,705	5,157	5,870	6,049	5,236	5,291	4,858	4,694	5,646	4,990	5,182
Percent of Total	9.41%	8.80%	7.96%	9.06%	9.34%	8.08%	8.17%	7.50%	7.24%	8.71%	7.70%	8.00%
Aligned Reads	6,082	5,682	5,139	5,858	6,032	5,226	5,263	4,834	4,668	5,615	4,967	5,131
Aligned %	99.7%	99.6%	99.7%	99.8%	99.7%	99.8%	99.5%	99.5%	99.4%	99.5%	99.5%	99.0%

Table 2: Analysis results for RUNX1 samples

11 mutations were called in the 3 TET2/CBL/KRAS samples, and 5 different mutations were called in the RUNX1 samples. The mutant allele percentages are shown in tables 3 and 4.

			s01- MID_01	s02- MID_02	s03- MID_04	
CBL	c.1139T>CT	380 L>PL		58.06%		
	c.1253T>CT	418 S>SF			52.60%	
KRAS	c.34G>GT	12 G>CG	12.62%			rs12193530
	c.57G>GT	19 L>FL	18.45%			rs12193538
	c.58A>AT	20 T>TS	18.45%			
TET2	c.3782G>AG	1261 R>HR			43.66%	
	c.3894_3895 insT	FS (1299)		44.15%		
	c.5284A>AG	1762 I>IV		50.63%		rs2454206
	c.5600_c.5611 delCTCATGGGTCAA	delTHGS (1867-1870)			41.50%	
	c.5736T>CT	1912 H>HH	48.45%			

Table 3: Mutation calling results for CBL/KRAS/TET2 samples.
Mutations highlighted in purple were found in the dbSNP v135 database.

		1	2	3	4	5	7	12	
c.261_262 insGA	FS (89)					6.73%	6.71%		
c.331_336 delACCCTG	del TL (111-112)		14.45%		16.20%				
c.482T>TC	161 L>LP	20.38%		18.71%					
c.1025T>TC	342 I>IT			8.11%				6.52%	
c.1389C>GC	463 P>PP	48.04%			49.72%		51.83%		rs61750222

Table 4: Mutation calling results for RUNX1 samples.
Samples 6, 8, 9, 10, and 11 did not have any called mutations after filtering.

Discussion

NextGENe is able to very rapidly process samples from the GS GType RUNX1 and GS GType TET2/CBL/KRAS primer sets in order to find potentially important mutations. This includes MID sorting and trimming, alignment, and mutation calling. NextGENe is also able to annotate the mutations that were found using dbSNP, the dbNSFP database, and (soon) the COSMIC database. Each alignment took less than a minute to run, and all of the pre-alignment processing was done in less than 5 minutes. All of these steps can be fully automated in order to make processing samples even faster and easier.

In this experiment two long (>= 6 bp) in-frame deletions were detected: a 12 bp deletion in TET2 (figure 1) and a 6bp deletion in RUNX1 (figure 5). Multiple SNPs were detected, including a T>TC mutation found near 20% frequency in two RUNX1 samples (figure 6). That mutation is predicted by PhyloP to be at a conserved position, and is predicted to be damaging by four different functional prediction scores (PolyPhen-2, SIFT, Mutation Taster, and LRT).

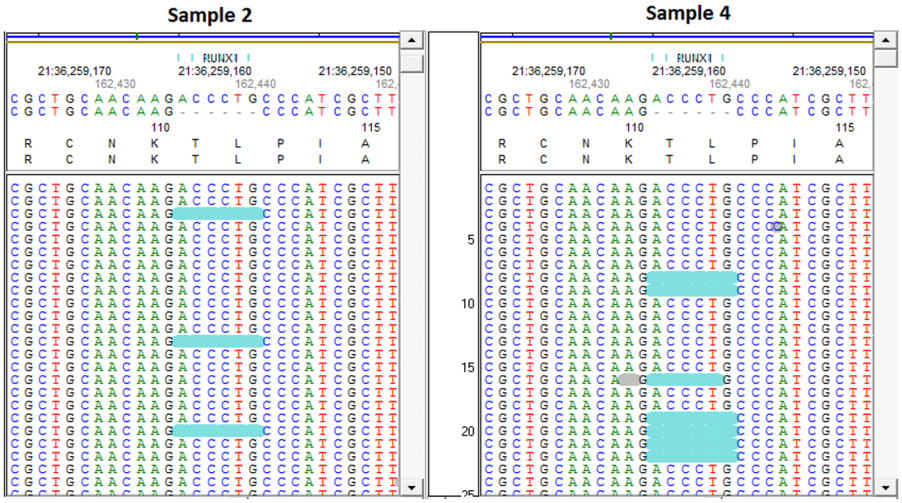

Figure 5: A 6bp deletion found in two RUNX1 samples

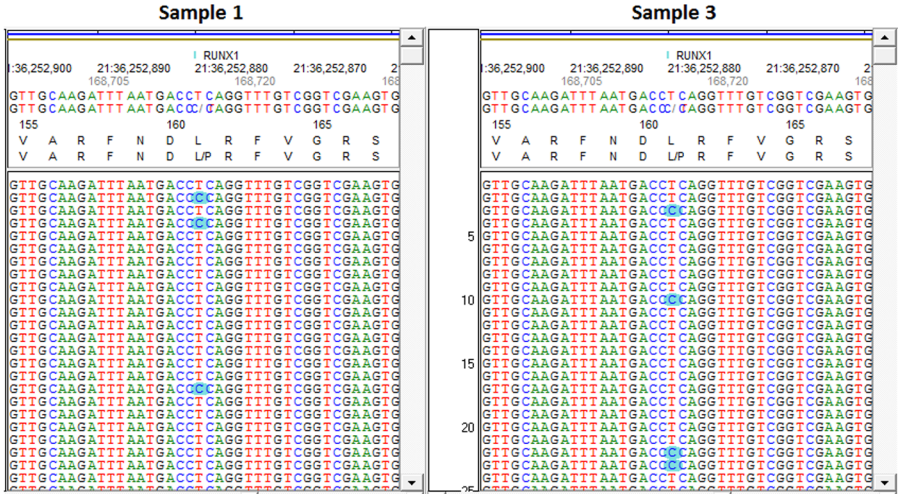

Figure 6: A SNP found in two RUNX1 samples that is predicted to be damaging

Acknowledgements

We would like to thank Roche for supplying the datasets used in this analysis.

References

1. Liu, X., Jian, X. & Boerwinkle, E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Human mutation 32, 894-9 (2011).
2. Forbes, S.A. et al. COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. Nucleic acids research 39, D945-50 (2011).