Processing Ion AmpliSeq™ Cancer Panel Data using NextGENe® Software

November 2011

John McGuigan, Megan Manion, Kevin LeVan, CS Jonathan Liu

Introduction

The Ion AmpliSeq™ Cancer Panel uses highly multiplexed PCR in order to generate amplicons from 46 different cancer genes. These amplicons are sequenced on the Ion PGM™, allowing for rapid turnaround and low cost. NextGENe® software allows for a customized analysis of the results, including detection of novel variants or alleles, alleles found at lower frequencies (less than 5%), or alleles in regions with less than 500x coverage. NextGENe includes many useful features, such as quality control reports, functional prediction scoring, and advanced project comparison. When using the Ion Torrent paired-end protocol, the overlapping reads can be merged into high-quality single reads in order to improve sequence accuracy. This can reduce the amount of filtering that is needed and increase the percentage of reads that successfully align. The analysis of this type of data is covered in the "Processing Paired End Data from the Ion PGM™ Sequencer Using the Floton Paired-End Merger in NextGENe® Software" app note.

Procedure

A given dataset can be processed in less than 15 minutes on a desktop computer running a 64-bit Windows operating system. All steps are performed with an easy-to-use point-and-click interface with no scripting required. Three datasets (one 316 chip and two 314 chips) were processed in this analysis.

Format Conversion

- As with all NextGENe projects, the raw data (FASTQ or SFF in this case) is first converted to FASTA format
- Basecall quality scores are used to filter and trim the raw data

Alignment to the human genome

- Alignment is performed against a pre-index human genome reference in order to avoid false positives caused by nonspecific amplification of untargeted regions
- Up to one mismatch is allowed in a read before the reads are broken into seeds.
- Mismatches are called as variants in the mutation report if they pass the mutation filter. Suggested settings require more than 100x coverage, 1% frequency, and 20 total reads with the mutation (Figure 1).
- Variants with F/R read balance < 0.2 and small homopolymer indels with F/R balance < 0.25 are removed from the final report in order to reduce the number of false positives.

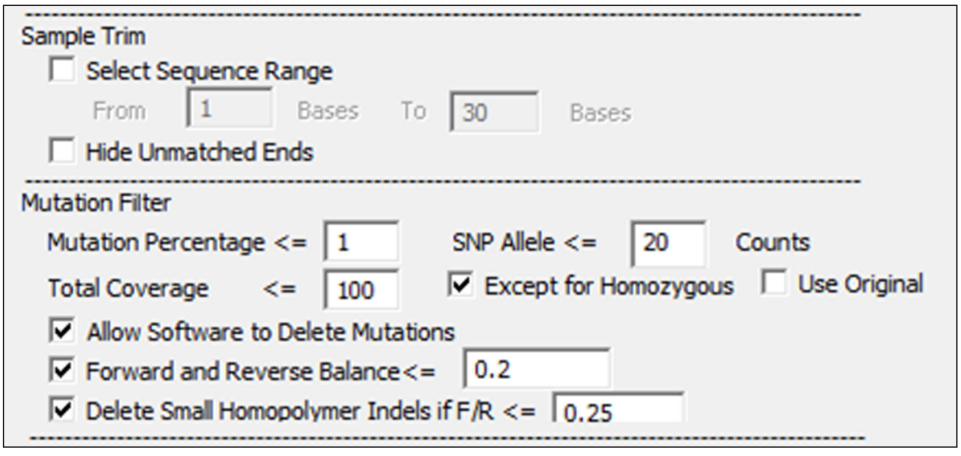


Figure 1 – Suggested mutation filter settings

Project review

- The mutation report, expression report, and coverage curve reports are filtered by loading a BED file specifying amplicon or targeted loci regions. The latter two reports provide information about coverage in the regions of interest. The mutation report has many additional filtering options.

Results

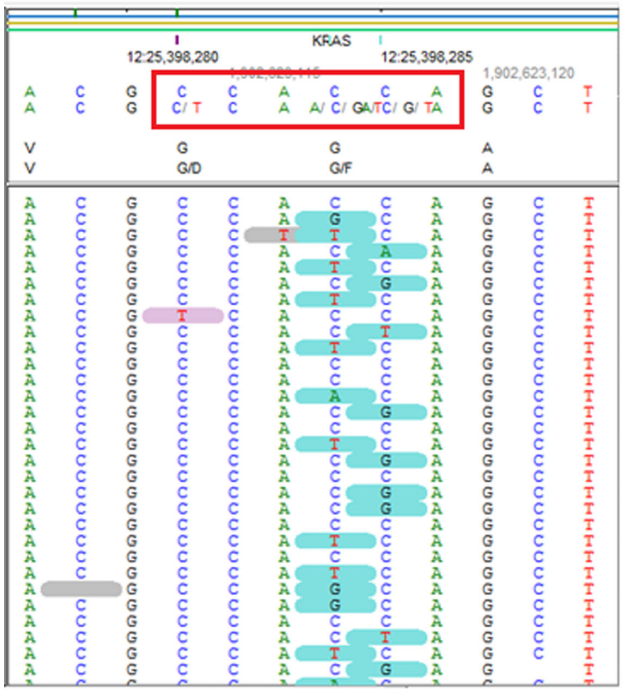
Table 1 lists the results of format conversion. Roughly 90% of the original reads were kept for each sample. Table 2 lists alignment results- around 95% of reads were successfully aligned, and about 95% of those reads were aligned in the amplicon regions. Figure 2 shows several mutations detected in a KRAS amplicon. One mutation is known (listed in the dbSNP database) and two others are complex, with multiple mutant alleles. Figure 3 lists the number of mutations found in each project. All 37 mutations called by the Ion AmpliSeq™ Cancer Variant Caller plugin were also found by NextGENe.

	FOZ-214 (316)	SUR-173 (314)	KER-780 (314)
Total Reads	1,167,000	578,080	569,813
Kept Reads	1,037,550	537,355	539,215
% Kept	88.91%	92.96%	94.63%

Table 1 – Format Conversion Results

	FOZ-214	SUR-173	KER-780
Aligned Reads	976,766	510,793	520,010
% Aligned	94.14%	95.06%	96.44%
Reads In Amplicon Regions	943,691	484,538	494,918
% in Regions	96.61%	94.86%	95.17%
Average Coverage in Amplicons	2,667x	1,621x	1,685x

Table 2 – Alignment Results



Chr 12 Position	Alleles			
	A	C	G	T
25,398,281	0.05	**95.34**	0.00	4.61
25,398,282	0.00	**99.84**	0.04	0.07
25,398,283	**99.4**	0.00	0.38	0.22
25,398,284	2.23	**79.22**	4.45	14.11
25,398,285	3.22	**44.67**	33.10	18.82
25,398,286	**99.43**	0.04	0.35	0.15

Figure 2 - A series of mutations found in KRAS. The first (purple) is a known mutation found in dbSNP. The other two have multiple mutant alleles at various frequencies. Reference allele frequencies are listed in bold.

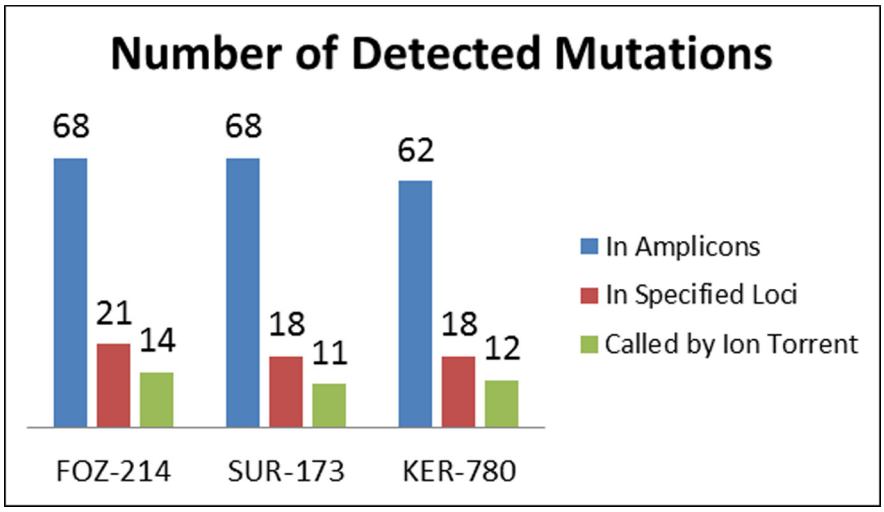
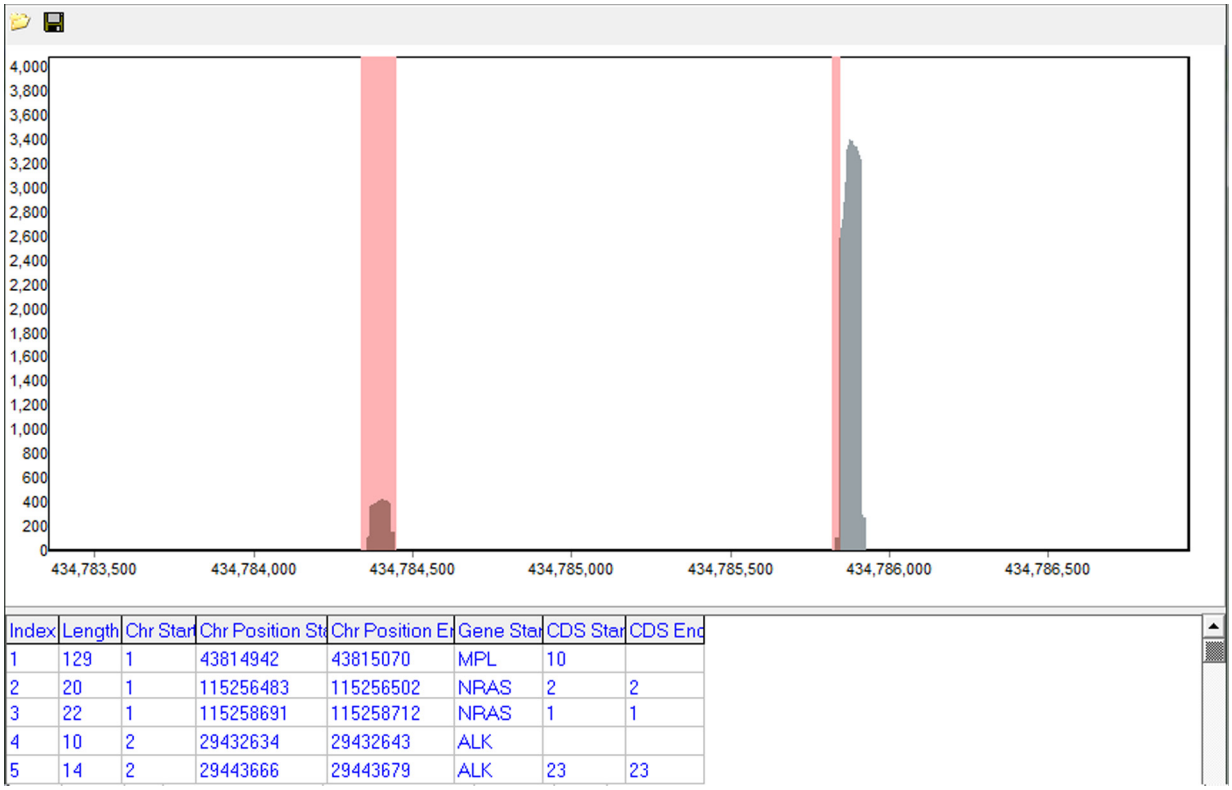


Figure 3 – Number of detected mutations in each project. NextGENe was able to detect more mutations than the built-in plugin because it can examine a larger portion of the amplicons and detect lower frequency alleles.

Discussion

When choosing alignment settings, it is important to consider the expected results. Increasing the minimum depth of coverage will reduce the number of false positives (even at lower mutation frequencies), but it may also decrease sensitivity. Setting a minimum number of mutant allele reads will allow for detection of low frequency variants in high coverage regions without allowing low frequency false positives to be called in low coverage regions. The coverage curve report is very useful for measuring potential loss of sensitivity (figure 4). The expression report can also be used the measure coverage (figure 5). This can be done for entire amplicons or for specific loci.



Index	Length	Chr Start	Chr Position St	Chr Position En	Gene Start	CDS Start	CDS End
1	129	1	43814942	43815070	MPL	10	
2	20	1	115256483	115256502	NRAS	2	2
3	22	1	115258691	115258712	NRAS	1	1
4	10	2	29432634	29432643	ALK		
5	14	2	29443666	29443679	ALK	23	23

Figure 4 - The coverage curve report can be used to find positions (pink) lacking the desired level of coverage (gray)

Software	NextGENe V 2.												
Project Name	FOZ-214.pjt												
Date/Time	11/8/2011 10:4												
Total Reads	1037550												
Matched Rea	976766												
Instrument													
Application	SNP/Indel Disc												
Index	Chr	Chr Position S	Chr Position E	Gene	CDS	Length	Min Counts	Max Counts	Average C	Read Cou	Forward	RPKM	RPK
1	1	43814942	43815070	MPL; +	10	129	1	105	67.49	121	67	993.9530	0.1282
2	1	115256483	115256582	NRAS; -	2	100	1	6491	3484.60	6707	2880	71071.99	7.1072
3	1	115258691	115258791	NRAS; -	1	101	2	7615	4301.60	7815	4862	81993.19	8.2813
4	10	43609046	43609184	RET; +	10	139	10	245	154.40	298	96	2271.808	0.3158
5	10	43609916	43610030	RET; +	11	115	1	1653	1124.30	1814	934	16715.12	1.9222
6	10	43613776	43613879	RET; +	13	104	1	229	129.88	249	124	2537.091	0.2639
7	10	43615541	43615646	RET; +	15	106	1	871	610.18	1015	410	10146.82	1.0756
8	10	43617380	43617479	RET; +	16	100	1	8888	4753.10	9253	4252	98051.16	9.8051
9	10	89624220	89624321	PTEN; +	1	102	1	14578	8478.60	15390	9355	159885.3	16.3083
10	10	89685236	89685387	PTEN; +	3	152	1	2090	1156.20	3010	1968	20984.23	3.1896
11	10	89711859	89711958	PTEN; +	6	100	1	5649	3132.70	5804	2863	61503.18	6.1503
12	10	89717556	89717667	PTEN; +	7	112	1	797	485.66	894	497	8458.428	0.9473
13	10	89717672	89717803	PTEN; +	7	132	1	86	42.32	99	61	794.7517	0.1049
14	10	89720663	89720774	PTEN; +	8	112	5	8707	4526.90	9370	5011	88652.65	9.9291
15	10	89720770	89720911	PTEN; +	8	142	1	211	126.21	258	95	1925.313	0.2734
16	10	123257972	123258071	FGFR2; -	11	100	1	883	529.28	914	555	9685.373	0.9685
17	10	123274713	123274841	FGFR2; -	8	129	2	7645	3965.50	8730	6285	71712.47	9.2509
18	10	123279456	123279565	FGFR2; -	6	110	1	6969	4294.00	7400	3974	71286.81	7.8415
19	10	123279612	123279745	FGFR2; -	6	134	1	7296	4387.50	8508	3147	67281.06	9.0157
20	11	533798	533933	HRAS; -	2	136	3	5263	3702.90	6803	3476	53006.81	7.2089

Figure 5- The expression report can be used to examine the coverage in each amplicon

NextGENe's whole genome alignment algorithm has three steps- Match perfect reads, match reads with some number of mismatches, and finally a seeded alignment. Shorter amplicons will benefit from the second step because there may not be enough bases on either side of a mutation to align perfectly matching seeds. Longer seed sizes will improve alignment specificity, and fewer seeds will improve speed. After alignment several projects can be compared and filtered against one another. Compound heterozygous, shared/different, low coverage, and Mendelian inheritance filtering are all possible (figure 6).

After calling mutations, some functional prediction information is available from the dbNSFP database (1). This includes PolyPhen-2, SIFT, MutationTaster, LRT, and PhyloP in addition to 1000 genomes frequencies. dbNSFP information for these three projects is summarized in figure 7. Soon additional databases will be supported including the Sanger COSMIC (Catalog of Somatic Mutations in Cancer) database (2).

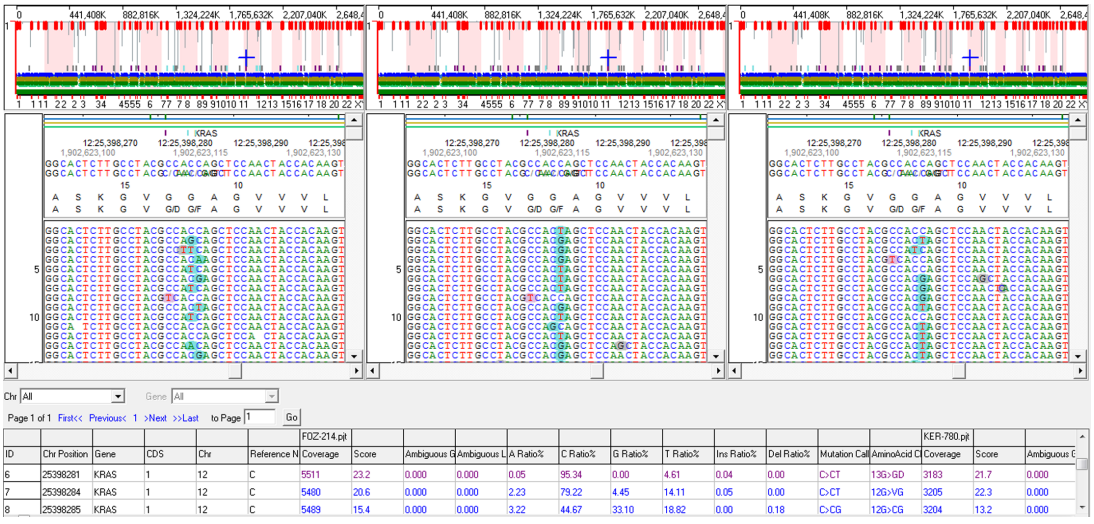


Figure 6 - The Variant Comparison Tool showing a mutation found in all 3 samples in a KRAS amplicon. From the left: FOZ-214, KER-780, SUR-173

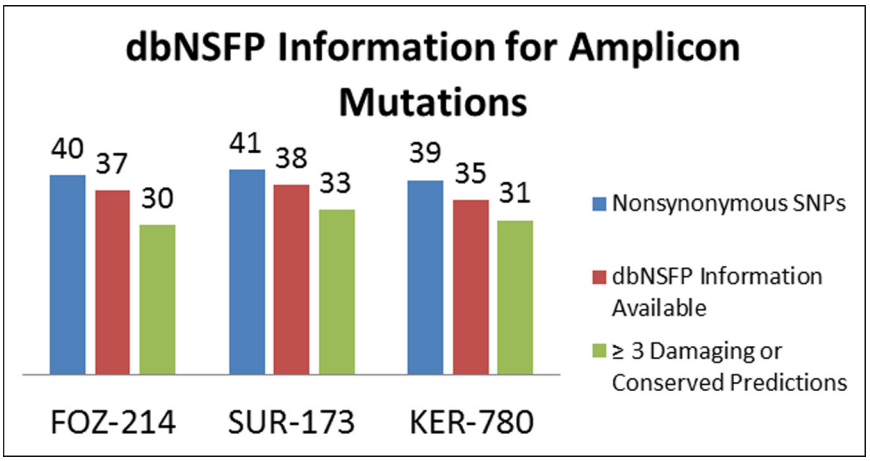


Figure 7 – Summary of dbNSFP Information. Most non-synonymous SNPs had dbNSFP information available, and most of these positions had several scores predicting damage/conservation.

Acknowledgements

We would like to thank Life Technologies for supplying the AmpliSeq data 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, 894-899 (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-D950 (2010).