NextGENe Software Analysis Using the NEBNext Direct® Cancer Hotspot Panel

September 2016

Kevin LeVan[1], Ni Shouyong[1], CS Jonathan Liu[1], Cynthia Hendrickson[2], Andrew Barry[3], Bjoern Textor[3]
1. SoftGenetics, State College, PA
2. Directed Genomics, Ipswich, MA
3. New England Biolabs, Ipswich, MA

Introduction

Cancer researchers are in need of easy-to-use and highly accurate methods for the detection of variants in cancer samples. The NEBNext Direct technology, offered by New England BioLabs® Inc. in conjunction with SoftGenetics NextGENe software enable the study of several disease causing genes. NEBNext utilizes Unique Molecular Identifiers (UMIs), allowing for the detection of PCR duplicates, and NextGENe software is capable of removing these duplicates, increasing the allele frequency accuracy.

Procedure

NEBNext Chemistry

NEBNext Direct libraries were created by shearing 100ng of genomic DNA to a median size of 200bp using Covaris ultrafocused acoustic shearing. DNA was denatured and hybridization was carried out using biotinylated oligonucleotide baits specific to the NEBNext Direct Cancer HotSpot Panel. Captured DNA was bound to streptavidin beads and separated from unbound material. Enzymatic removal of off-target sequence was performed, followed by a series of enzymatic manipulations to convert the captured material into Illumina sequencer compatible libraries. The resulting libraries contain two index barcodes incorporated into the universal adaptors. The 3' adaptor contains an 8 bp, sample-specific index used to disambiguate samples pooled prior to sequencing. The 5' adaptor contains a 12 bp randomized sequence that serves as a Unique Molecule Index (UMI), tagging individual molecules allowing marking of duplicates created during PCR amplification (Figure 1). For full details, please refer to the Product Manual for the NEBNext Direct Cancer HotSpot Panel available for download from NEB.com. Prepared libraries were sequenced on an Illumina MiSeq using 2x75bp paired-end v2 chemistry.

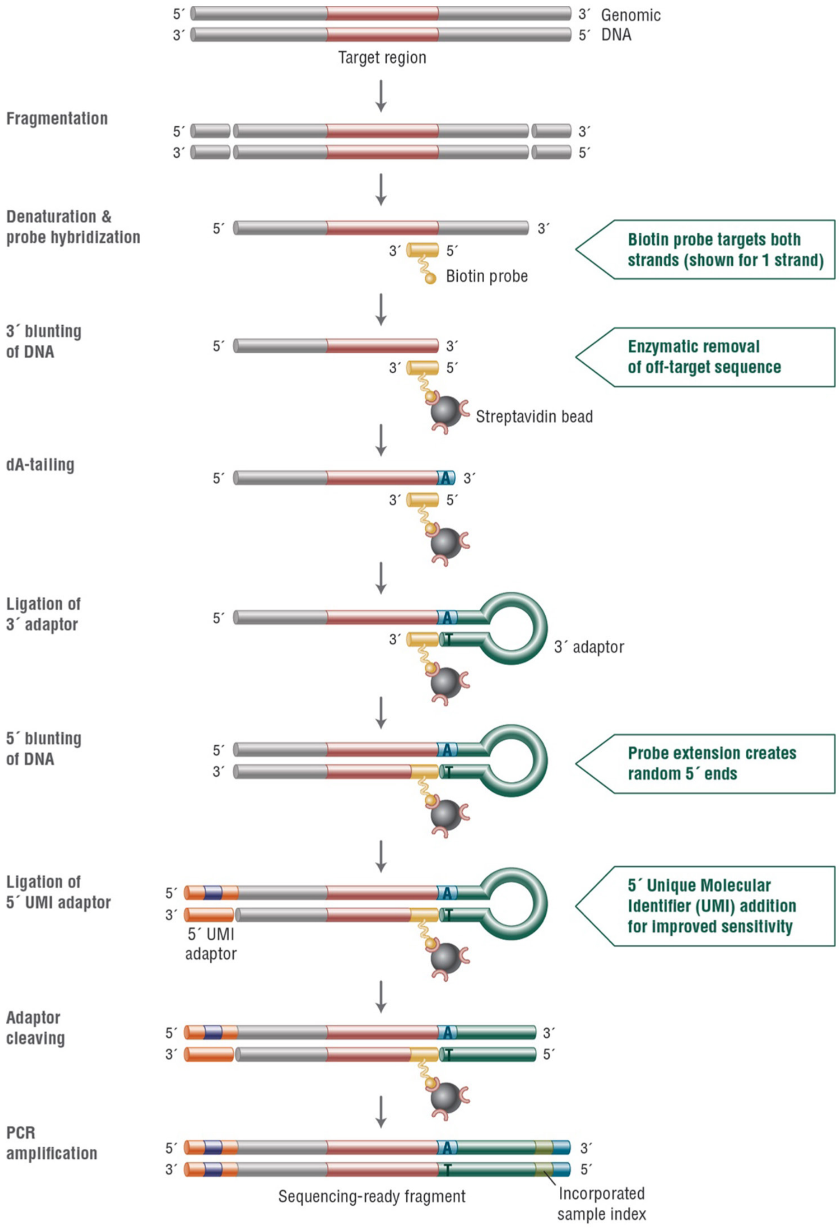

5´
3´
3´ Genomic
5´ DNA
Target region
Fragmentation
Denaturation & probe hybridization
Biotin probe
Biotin probe targets both strands (shown for 1 strand)
3´ blunting of DNA
Enzymatic removal of off-target sequence
Streptavidin bead
dA-tailing
A
Ligation of 3´ adaptor
T
3´ adaptor
5´ blunting of DNA
Probe extension creates random 5´ ends
Ligation of 5´ UMI adaptor
5´ UMI adaptor
5´ Unique Molecular Identifier (UMI) addition for improved sensitivity
Adaptor cleaving
PCR amplification
Sequencing-ready fragment
Incorporated sample index

Figure 1: NEBNext Direct Cancer HotSpot Panel chemistry

NextGENe Software

Preinstalled panel templates are included in NextGENe software for NEBNext chemistries, including the Cancer HotSpot Panel. 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. Any setting(s) can be adjusted to create a new template. 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.
3. Select a template from the drop-down list. Templates for the NEBNext Direct Cancer HotSpot Panel are included with the software. The "Save As" button allows the templates to be modified into 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. This makes it easy to split the list of raw data files into separate sample projects.
8. Click OK to save the ngjob file.
9. Start the AutoRun tool to begin processing.

Discussion

The NextGENe NEBNext Direct Template automatically guides the sample files through a series of steps, including Removal of PCR duplicates, Format Conversion, trimming of adapters, alignment to the human genome and variant calling. Variants will be called when all of the

following criteria are satisfied:

- Percentage of reads with variant is greater than 1.5%
- Variant is found in more than 3 reads
- Total coverage is more than 100 reads
- Variant forward/reverse balance ratio more than 0.2 (0.8 for homopolymer indels)
- Variant is within target region

A pair of samples, S12 and S13, takes under 12 minutes to analyze on a desktop computer with an i7 processor and 16GB RAM. Each sample started with over 1.2 million pairs of reads (Figure 2).

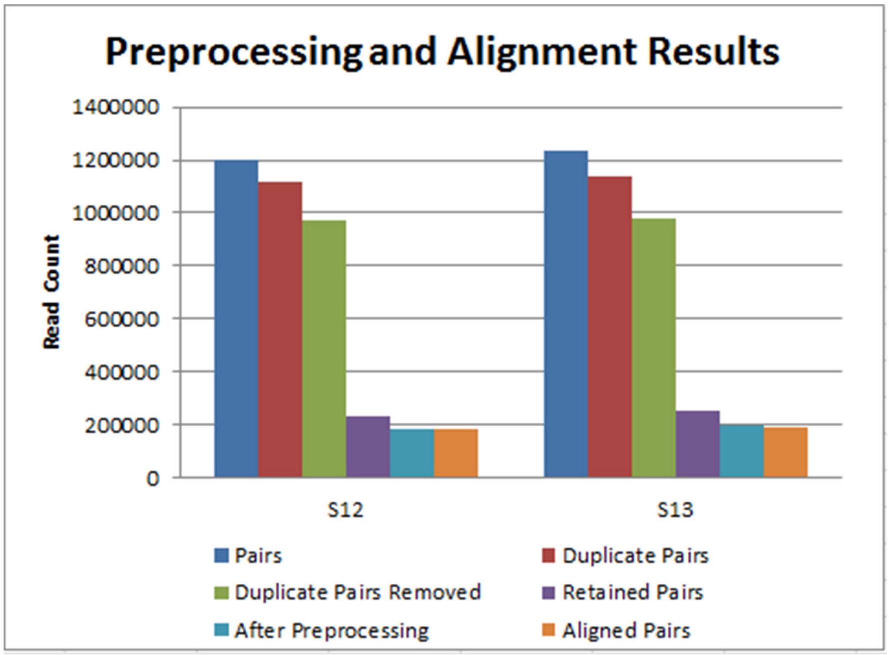
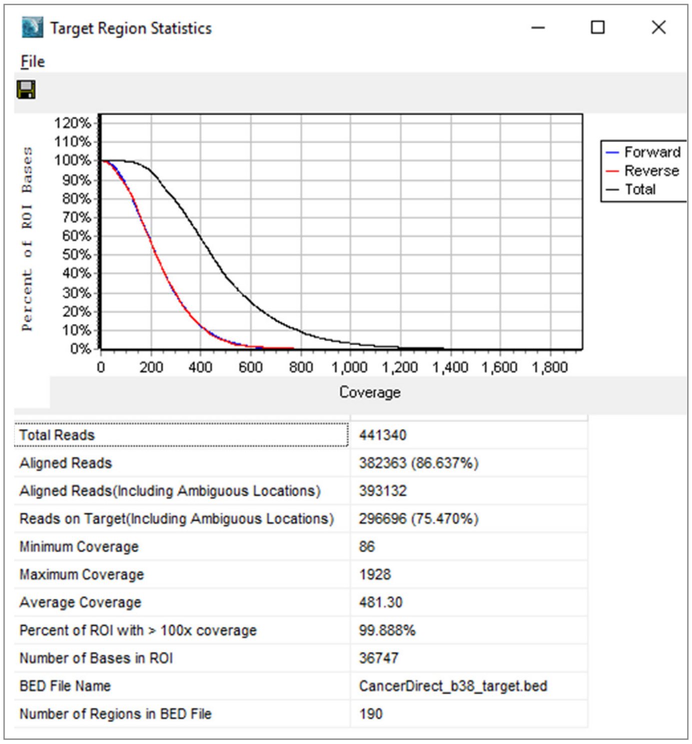


Figure 2: Read count statistics

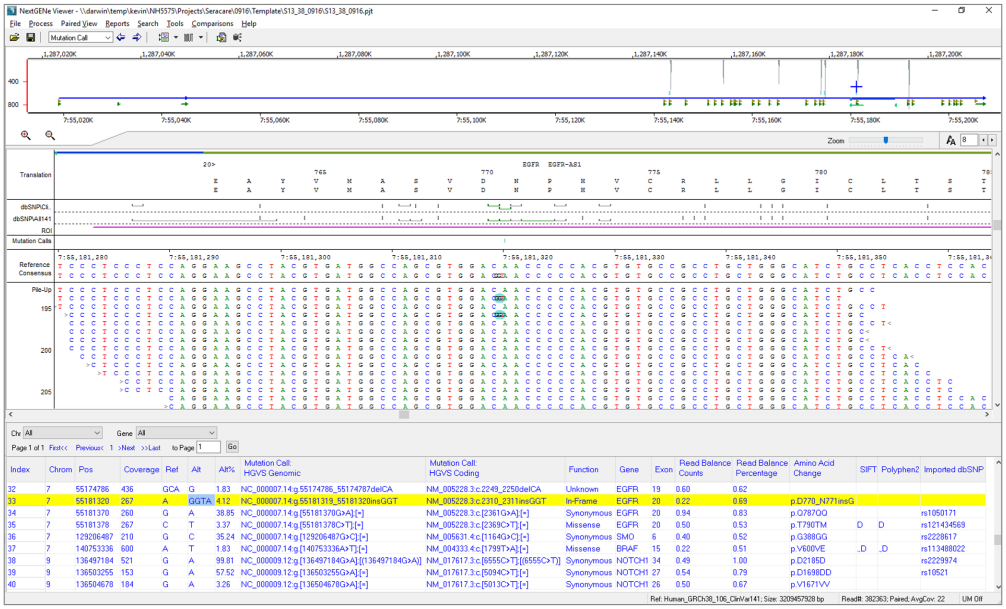
NextGENe software uses UMIs within the Illumina I2 files to identify PCR duplicates. The pair of duplicates with the highest total score is maintained to be processed along with unique paired reads, while duplicate reads are removed from further processing. This process left about 200,000 unique pairs for alignment of each of these samples to the human genome.

Figure 3: Sample S13 coverage statistics 99.9% of the target regions had over 100x coverage, as shown in Figure 3. Counting unique reads, the target had a minimum coverage of 86 reads, average coverage of 481 reads and a maximum coverage of 1928 reads.



Total Reads	441340
Aligned Reads	382363 (86.637%)
Aligned Reads(Including Ambiguous Locations)	393132
Reads on Target(Including Ambiguous Locations)	296696 (75.470%)
Minimum Coverage	86
Maximum Coverage	1928
Average Coverage	481.30
Percent of ROI with > 100x coverage	99.888%
Number of Bases in ROI	36747
BED File Name	CancerDirect_b38_target.bed
Number of Regions in BED File	190

The NextGENe NEBNext templates are set to detect variants above 1.5%. The target contains 190 regions encompassing over 36Kbps. For sample S13, 57 variants were identified (51 substitutions, 2 insertions and 4 deletions). This includes 100% of the expected variants, with an allele percentage as low as 1.81%, total coverage as low as 153 reads and balance as low as 0.34.


NextGENe Viewer - \\darwin\temp\kevin\NH5575\Projects\Seracare\0916\Template\S13_38_0916\S13_38_0916.pjt

File Process Paired View Reports Search Tools Comparisons Help
Mutation Call
Translation
dbSNPCli
dbSNPAll141
ROI
Mutation Calls
Reference
Consensus
Pile-Up
EGFR EGFR-AS1
Zoom
Chr All
Gene All
Page 1 of 1 First<< Previous< 1 >Next >>Last to Page 1 Go

Index	Chrom	Pos	Coverage	Ref	Alt	Alt%	Mutation Call: HGVS Genomic	Mutation Call: HGVS Coding	Function	Gene	Exon	Read Balance Counts	Read Balance Percentage	Amino Acid Change	SIFT	Polyphen2	Imported dbSNP
32	7	55174786	436	GCA	G	1.83	NC_000007.14:g.55174786_55174787delCA	NM_005228.3:c.2249_2250delCA	Unknown	EGFR	19	0.60	0.62				
33	7	55181320	267	A	GGTA	4.12	NC_000007.14:g.55181319_55181320insGGT	NM_005228.3:c.2310_2311insGGT	In-Frame	EGFR	20	0.22	0.69	p.D770_N771insG			
34	7	55181370	260	G	A	38.85	NC_000007.14:g.[55181370G>A];[=]	NM_005228.3:c.[2361G>A];[=]	Synonymous	EGFR	20	0.94	0.83	p.Q787QQ			rs1050171
35	7	55181378	267	C	T	3.37	NC_000007.14:g.[55181378C>T];[=]	NM_005228.3:c.[2369C>T];[=]	Missense	EGFR	20	0.50	0.53	p.T790TM	D	D	rs121434569
36	7	129206487	210	G	C	35.24	NC_000007.14:g.[129206487G>C];[=]	NM_005631.4:c.[1164G>C];[=]	Synonymous	SMO	6	0.40	0.52	p.G388GG			rs2228617
37	7	140753336	600	A	T	1.83	NC_000007.14:g.[140753336A>T];[=]	NM_004333.4:c.[1799T>A];[=]	Missense	BRAF	15	0.22	0.51	p.V600VE	.,D	.,D	rs113488022
38	9	136497184	521	G	A	99.81	NC_000009.12:g.[136497184G>A];[(136497184G>A)]	NM_017617.3:c.[6555C>T];[(6555C>T)]	Synonymous	NOTCH1	34	0.49	1.00	p.D2185D			rs2229974
39	9	136503255	153	G	A	57.52	NC_000009.12:g.[136503255G>A];[=]	NM_017617.3:c.[5094C>T];[=]	Synonymous	NOTCH1	27	0.54	0.79	p.D1698DD			rs10521
40	9	136504678	184	G	A	3.26	NC_000009.12:g.[136504678G>A];[=]	NM_017617.3:c.[5013C>T];[=]	Synonymous	NOTCH1	26	0.50	0.67	p.V1671VV			

Ref: Human_GRCh38_106_ClinVar141; Size: 3209457928 bp   Read#: 382363; Paired; AvgCov: 22   UM Off

Figure 4: NextGENe Viewer

After processing, the projects can be visualized in the NextGENe Viewer (Figure 4). 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), and an expression report (showing read counts per target). The Reference and Track Manager tool allows for optional tracks to be imported into the NextGENe software. These tracks (including COSMIC[1], ClinVar [2], and dbNSFP [3]) can then be queried automatically for every new project, giving additional information for each variant.

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.