TECHNICAL REPORT

Ovation® Fusion Detection BaseSpace® Application

I. Introduction

The NuGEN Ovation Fusion Detection Application is designed to detect gene fusions in human sequencing data produced from libraries prepared with NuGEN's Ovation Fusion Panel Target Enrichment System or Ovation Custom Fusion Target Enrichment panels.

The Ovation Target Enrichment Systems utilize Single Primer Enrichment Technology (SPET), a novel approach for targeted resequencing of genomic DNA or cDNA for targeted RNA analysis, and are suitable for a wide range of target sizes from a few kilobases to over 10 megabases. This method uses a single targeting probe that hybridizes to the target region and then extends through the region of interest. This approach eliminates the difficulty of designing specific PCR primer pairs while maintaining a high specificity of recovered target sequences in the final library. This technique is highly flexible and appropriate for use in the targeted analysis of a wide range of genomic markers including mutations, SNP's, indels, gene fusions, alternately spliced transcripts and copy number variants.

This application is designed for use with paired-end sequencing data from any Illumina sequencing platform. Data input requirements for this application is human sequencing data in FASTQ format (including both the paired end reads and the 14 bp index read) as well as the NuGEN-provided manifest file containing probe and target information.

II. Description

The data is first demultiplexed using all 96 available barcodes provided in the Ovation Target Enrichment System. Subsequently, the linker sequence is trimmed followed by assignment of probes to their respective derived reads. After an optional deduplication step, the reads undergo quality trimming and alignment to the UCSC human reference genome hg19 using STAR. Finally, the alignment is filtered and fusions are identified using STAR-Fusion. A summary table is generated, providing general data analysis metrics detailing the total number of reads, probe assignment rates, duplicate rates, alignment rates and the number of gene fusions detected for each sample. Each sample-specific summary provides a graphical representation of the fusion sequence, coding frame and translation. An additional summary table provides genomic coordinates of the fusion breakpoint, number of supporting reads and a p-value for each fusion identified in the sample.

NuGEN
imagine more from less®

III. Sequencing and Data Upload

Setting Up a Sequencing Run to Capture the N6 Duplicate Information

A 14 bp index read is recommended when sequencing libraries generated with the Ovation Target Enrichment System. This allows sequencing of the 8-base barcode, as well as 6 random bases that immediately follow the barcode sequence. The additional 6 bases are used for PCR duplicate read determination during data analysis. To capture this information, ensure that the sequencer is configured properly to sequence a 14-base index read.

Since this application initially performs a demultiplexing step, we recommend using a multiplexed fastq file as input into the Application to analyze multiple samples simultaneously. For example, on a MiSeq set up the sequencing sample sheet with a generic sample name (i.e.: "sample") and NNNNNNNNNNNNNN (14 Ns) as the index, so that library demultiplexing will be skipped and only one multiplexed fastq file will be generated. Alternatively, the sample sheet can include the 8-base barcodes followed by NNNNNN, although each sample will have to be uploaded and processed individually in the NuGEN Ovation Fusion Detection BaseSpace Application (not recommended).

Illumina sequencing instruments do not provide a simple way to obtain the sequence information contained in the 14-base pair index read including the 6 random bases that are necessary for duplicate read determination. Several recommended methods to generate the necessary index fastq file are provided below.

MiSeq Instruments

Parsing multiplex runs using the MiSeq built-in Illumina software replaces the barcode sequence from each library with a numerical substitute. This removes the duplicate information provided by the N6 sequence present after the barcode. To retrieve this information using the MiSeq instrument, we recommend modification of the MiSeq config file to allow generation of an index fastq file during data analysis. This will generate a 14 base index file that is compatible with the NuGEN Ovation Fusion Detection BaseSpace Application.

If you are unfamiliar with editing the config file, it is recommended to request assistance from Illumina Technical Support to make this modification. The steps are as follows:

1. Stop the MiSeq Reporter process.
2. Locate the "MiSeq Reporter.exe.config" file located in C:/Illumina/MiSeq Reporter
3. Open config file and search for a line that reads:

   "<add key="CreateFastqForIndexReads" value="0"/>".

   - If this line is present, change the value from "0" to "1".
   - If this line is not present, add the line to the config file using the add keys function under the appSettings tab.
4. Restart the MiSeq reporter process.
5. Requeue the run for data analysis if required.

Other Illumina Sequencers

Use the method described below to generate the read and index fastq files for use with the NuGEN Ovation Fusion Detection BaseSpace Application using bcl2fastq2 version 2.17.

1. Navigate to the location of the run folder (referred to as RunFolder in this document) and rename the sample sheet (i.e. SampleSheet.csv.bak).
2. To generate the run and index fastq files use the following command:

   "/location/bcl2fastq --runfolder-dir . --output-dir ./Data/Intensities/BaseCalls/ --use-bases-mask y*,y*,y* --minimum-trimmed-read-length 0 --mask-short-adapter-reads 0".

   The generated fastq files can be uploaded to BaseSpace for input into the NuGEN Ovation Fusion Detection BaseSpace Application.

   Note: In order to parse the data during the fastq file generation, modify the sample sheet to remove the six N's located at the end of the barcodes. Run bcl2fastq using the command:

   "/location/bcl2fastq --runfolder-dir . --output-dir ./Data/Intensities/BaseCalls/ --sample-sheet SampleSheet.csv --use-bases-mask y*,i8y*,y* --minimum-trimmed-read-length 0 --mask-short-adapter-reads 0".

   This command will produce an R1 fastq file with the forward read, an R2 fastq files containing just the N6 information and, if present, an R3 fastq file containing the reverse read.
3. The fastq files will be located in /RunFolder/Data/Intensities/BaseCalls/ unless specified otherwise.

Importing Sequencing Data into BaseSpace: Forward and Reverse Sequencing Reads

If you are able to access your files in BaseSpace, skip this step. If you are unable to access your files directly in BaseSpace, manually upload the required files using the instructions below. This will first require that the files follow the Illumina naming convention, described below:

> <sample name>_S<sample number>_L<lane (0-padded to 3 digits)>_R<read number>_<set number (0-padded to 3 digits>.fastq.gz
>
> i.e.: Sample_S1_ L001_ R1_001.fastq.gz, Sample_S1_ L001_ R2_001.fastq.gz

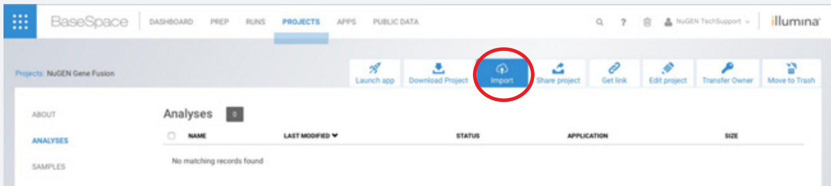
BaseSpace only allows the import of one sample at a time. Each unique sample fastq file will need to be uploaded individually, so the use of multiplexed fastq files is recommended. Use of paired-end data is required and it is recommended that both the forward and reverse reads are imported together, as they will then be associated with the same sample within BaseSpace. In order to do this, first click on the 'Import' icon within a BaseSpace Project (Figure 1).

This will take you to the page shown in Figure 2.

Either click on 'select files' and navigate to the sequencing files on your computer, or 'Drag and drop' the files onto the window. The file name of the forward and reverse read should be identical, with exception of the read number. For example, "mysample_S0_L001_R1_001.fastq.gz" and "mysample_S0_L001_R2_001.fastq.gz" (Figure 3).

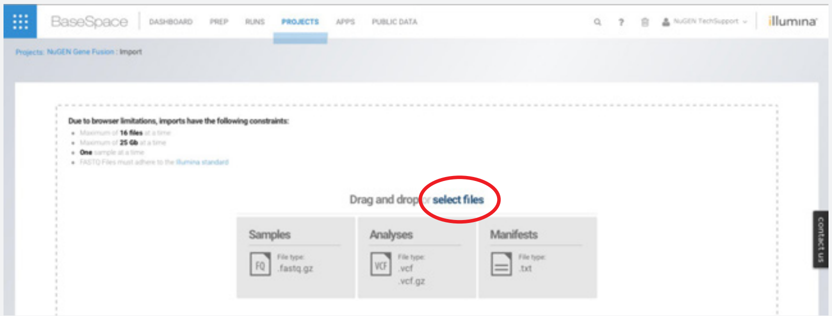
Both reads will then be associated with the sample name assigned to it, which can be entered in the grey field at the left of the screen; in the example above, this is "Sample". The import can take a considerable amount of upload time depending on your run and files

FIGURE 1. Import icon in BaseSpace



No matching records found

FIGURE 2. Import window in BaseSpace

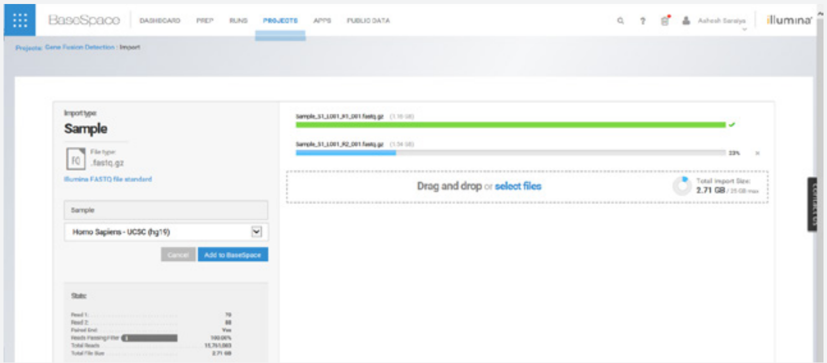


Due to browser limitations, imports have the following constraints:

- Maximum of 16 files at a time
- Maximum of 25 Gb at a time
- One sample at a time
- FASTQ Files must adhere to the Illumina standard

Drag and drop or select files

FIGURE 3. Importing files into BaseSpace



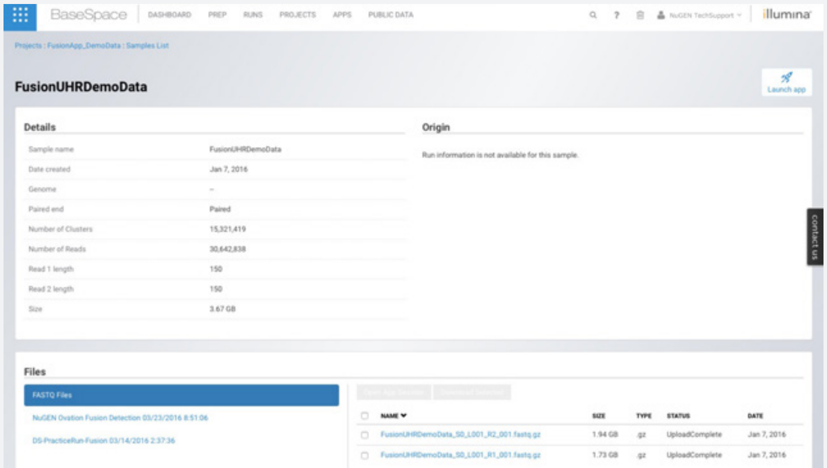
Sample_S1_L001_R1_001.fastq.gz

Sample_S1_L001_R2_001.fastq.gz 23%

Sample

Homo Sapiens - UCSC (hg19)

FIGURE 4. Sample information displayed in BaseSpace


BaseSpace
DASHBOARD
PREP
RUNS
PROJECTS
APPS
PUBLIC DATA
NuGEN TechSupport
illumina
Projects : FusionApp_DemoData : Samples List
Launch app
contact us
Files
FASTQ Files
NuGEN Ovation Fusion Detection 03/23/2016 8:51:06
DS-PracticeRun-Fusion 03/14/2016 2:37:36

FusionUHRDemoData

Details

Sample name	FusionUHRDemoData
Date created	Jan 7, 2016
Genome	–
Paired end	Paired
Number of Clusters	15,321,419
Number of Reads	30,642,838
Read 1 length	150
Read 2 length	150
Size	3.67 GB

Origin

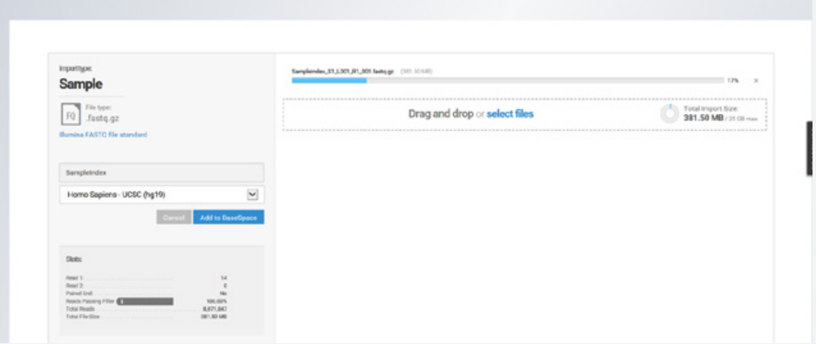
Run information is not available for this sample.

NAME	SIZE	TYPE	STATUS	DATE
FusionUHRDemoData_S0_L001_R2_001.fastq.gz	1.94 GB	.gz	UploadComplete	Jan 7, 2016
FusionUHRDemoData_S0_L001_R1_001.fastq.gz	1.73 GB	.gz	UploadComplete	Jan 7, 2016

must conform to BaseSpace's upload requirements. At the time of writing this manual, the file size upper limit is 25 GB.

After the upload is complete, click on the "Add to BaseSpace" button to add the sample and the sample information will be displayed as shown in Figure 4.

FIGURE 5. Importing additional required files to BaseSpace



Import type:

Sample

File type: .fastq.gz

Illumina FASTQ file standard

SampleIndex

Homo Sapiens - UCSC (hg19)

Cancel | Add to BaseSpace

Stats:

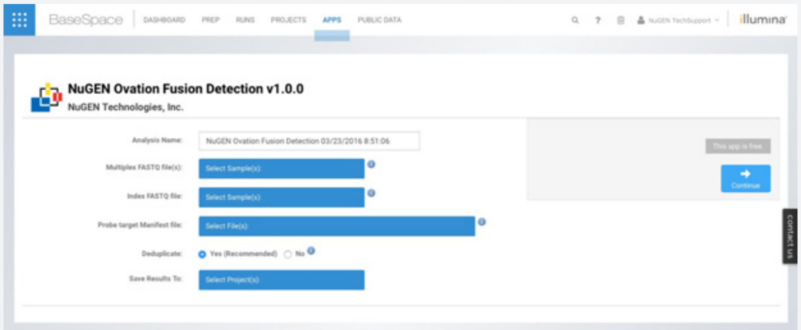
Read 1:	14
Read 2:	0
Paired End:	No
Reads Passing Filter	100.00%
Total Reads	8,671,847
Total File Size	381.50 MB

SampleIndex_S1_L001_R1_001.fastq.gz (381.50 MB) 17%

Drag and drop or select files

Total Import Size: 381.50 MB / 25 GB max

FIGURE 6. Setting up data analysis in BaseSpace



NuGEN Ovation Fusion Detection v1.0.0

NuGEN Technologies, Inc.

Analysis Name: NuGEN Ovation Fusion Detection 03/23/2016 8:51:06

Multiplex FASTQ file(s): Select Sample(s):

Index FASTQ file: Select Sample(s):

Probe target Manifest file: Select File(s):

Deduplicate: ● Yes (Recommended) ○ No

Save Results To: Select Project(s):

This app is free

Continue

Importing Sequencing Data Into BaseSpace: Index Read and Manifest File

In order to identify PCR duplicates, the NuGEN Ovation Fusion Detection BaseSpace Application requires a 14 bp index read. Due to the way that BaseSpace manages FASTQ files, the index file name must be changed to be unique from the forward and reverse read fastq files while conforming to the Illumina standard naming convention. To maintain this convention, it is recommended that "Index" be incorporated in the sample name to reduce confusion and designate the file as "R1". An example index read FASTQ name is provided below:

SampleIndex_S0_L001_R1_001.fastq.gz

This file can then be imported using the import function, as described previously. In this case only a single file will be imported, and it should be assigned a different sample name from the forward and reverse reads (Figure 5).

Use of the NuGEN Ovation Fusion Detection BaseSpace Application requires an additional file, called a manifest file that identifies the target regions and probe sequences used for target enrichment. This file is specific to the standard or custom Ovation Target Enrichment System Panel that was used to generate the sequencing data. A single target file will be used for all samples in a sequencing run. If multiple targeted resequencing experiments are multiplexed on the same sequencing run, the application will need to be run multiple times using different manifest files. The manifest file can be added to a BaseSpace Project using the Import function, as described previously.

This manifest file will have the format probe_target_manifest_ETXXXX_1_XXX.txt, and can be obtained from NuGEN Technical Support (Email: techserv@nugen.com). For custom panels, please include the DesignID (format ET####) specific to the custom design; for standard panels, such as the Ovation Fusion Panel Target Enrichment System, please provide the product name or number.

IV. Running the Application

Once files have been uploaded to BaseSpace, you can go to the NuGEN Ovation Fusion Detection Application to begin your analysis (Figure 6).

The following fields require input:

Analysis Name: A unique name for the analysis to be performed. Note that multiple analyses can be run on the same dataset by selecting different options, so it is best to be as descriptive as possible in this field.

Multiplex FASTQ file(s): Clicking on "Select Sample(s)" will navigate to a window that displays the available Sample Names for use. Select the desired dataset; note that paired end datasets will have both the forward and reverse read associated with one sample name. Only samples with more than 10,000 reads will be analyzed.

Index FASTQ file: Click on "Select Sample(s)" and choose the index read from the corresponding sequencing run.

Probe Target Manifest file: Click on "Select File(s)" and navigate to the NuGEN-supplied manifest file for the target region used with the system. This file is supplied by NuGEN during the design process. For more information on obtaining this file, contact techserv@nugen.com.

Deduplicate: Select whether you would like to deduplicate the reads before analysis. Deduplication is recommended for standard gene fusion detection and is set as the default.

Save Results To: Click on "Select Project(s)" and choose a Project folder to save the data.

After the files are in place and the desired functions are selected, click on the blue 'Continue' button to start the analysis.

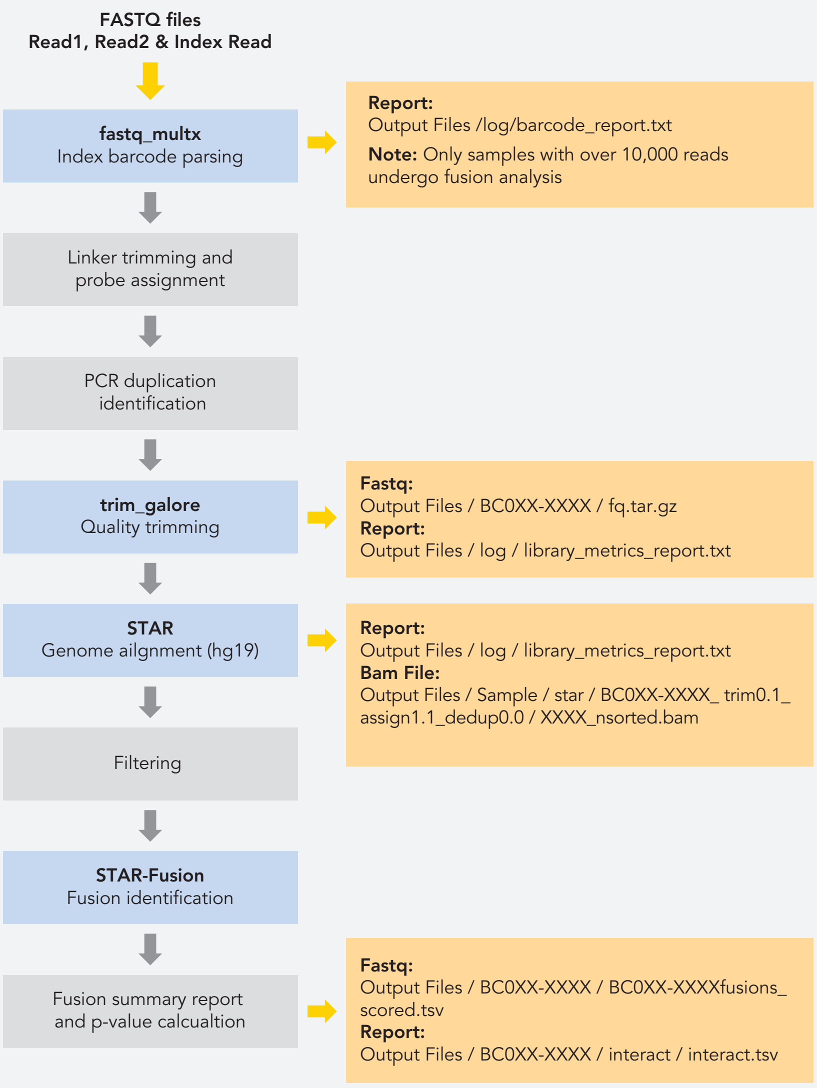
V. Analysis Procedure

The general pre-processing workflow and output files can be seen in Figure 7.

When the analysis of all samples is complete you will receive an e-mail notification and the status of the analysis on the Analysis Info page will change to "Complete". Data processing times depend primarily on the size of the dataset and are typically complete within 6 hours.

If the analysis failed, the failure detail(s) will be stored in the log file available on the analysis info page (see below for details). For troubleshooting help with NuGEN Ovation Fusion Detection Application analysis failures, please

FIGURE 7. **Pre-processing workflow and output files**


FASTQ files
Read1, Read2 & Index Read
fastq_multx
Index barcode parsing
Report:
Output Files /log/barcode_report.txt
Note: Only samples with over 10,000 reads undergo fusion analysis
Linker trimming and probe assignment
PCR duplication identification
trim_galore
Quality trimming
Fastq:
Output Files / BC0XX-XXXX / fq.tar.gz
Report:
Output Files / log / library_metrics_report.txt
STAR
Genome ailgnment (hg19)
Report:
Output Files / log / library_metrics_report.txt
Bam File:
Output Files / Sample / star / BC0XX-XXXX_ trim0.1_ assign1.1_dedup0.0 / XXXX_nsorted.bam
Filtering
STAR-Fusion
Fusion identification
Fusion summary report and p-value calcualtion
Fastq:
Output Files / BC0XX-XXXX / BC0XX-XXXXfusions_ scored.tsv
Report:
Output Files / BC0XX-XXXX / interact / interact.tsv

contact NuGEN's Technical Support team (techserv@nugen.com) with the log file copy pasted into an email to help us better assist you.

VI. Reviewing and Downloading Analysis Results

After completing the analysis, a number of different files and reports are available, and are described below.

Analysis Info: Clicking on this tab shows the status of the analysis (Figure 8).

Clicking on 'log files' will navigate to a page that contains downloadable log files of the analysis pipeline (Figure 9). The 'output-xxxx.log' file contains the commands used for data analysis and can be replicated or modified to analyze the data in your own analysis environment. The remaining log files are specific for the BaseSpace environment.

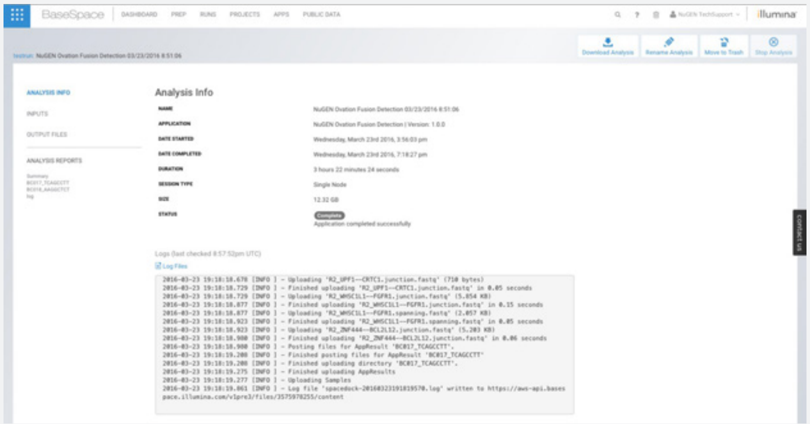
Inputs: Clicking on inputs navigates to a page that contains a summary of the input files and settings for the analysis run (Figure 10).

Output Files: Clicking on this tab navigates to a page with folders for each sample and a log folder (Figure 11). The log folder contains the demultiplexing (barcode_report.txt) and general alignment (library_metrics_report.txt) log files.

Each sample folder contains additional folders and a number of files available for download (Figure 12).

- **interact folder:** Tab-delimited report of the identified fusions and their sequence information. This information is provided in the graphical representation for each fusion (see Sample tab).
- **log folder:** Provides sample specific log reports.
- **star folder:** Contains the genomic alignment used for gene fusion detection.

FIGURE 8. Analysis information tab



testrun: NuGEN Ovation Fusion Detection 03/23/2016 8:51:06

Analysis Info

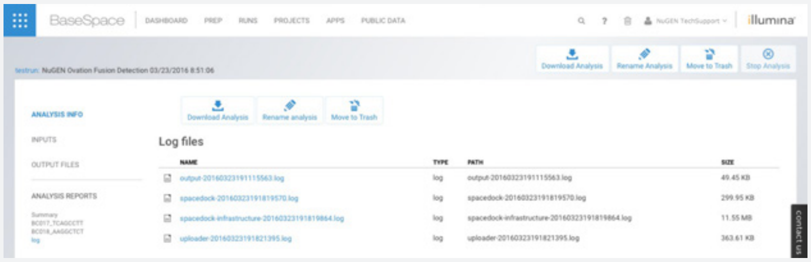
NAME	NuGEN Ovation Fusion Detection 03/23/2016 8:51:06
APPLICATION	NuGEN Ovation Fusion Detection | Version: 1.0.0
DATE STARTED	Wednesday, March 23rd 2016, 3:56:03 pm
DATE COMPLETED	Wednesday, March 23rd 2016, 7:18:27 pm
DURATION	3 hours 22 minutes 24 seconds
SESSION TYPE	Single Node
SIZE	12.32 GB
STATUS	Complete Application completed successfully

Logs (last checked 8:57:52pm UTC)

Log Files

```
2016-03-23 19:18:18.678 [INFO ] - Uploading 'R2_UPF1--CRTC1.junction.fastq' (710 bytes)
2016-03-23 19:18:18.729 [INFO ] - Finished uploading 'R2_UPF1--CRTC1.junction.fastq' in 0.05 seconds
2016-03-23 19:18:18.729 [INFO ] - Uploading 'R2_WHSC1L1--FGFR1.junction.fastq' (5.054 KB)
2016-03-23 19:18:18.877 [INFO ] - Finished uploading 'R2_WHSC1L1--FGFR1.junction.fastq' in 0.15 seconds
2016-03-23 19:18:18.877 [INFO ] - Uploading 'R2_WHSC1L1--FGFR1.spanning.fastq' (2.057 KB)
2016-03-23 19:18:18.923 [INFO ] - Finished uploading 'R2_WHSC1L1--FGFR1.spanning.fastq' in 0.05 seconds
2016-03-23 19:18:18.923 [INFO ] - Uploading 'R2_ZNF444--BCL2L12.junction.fastq' (5.203 KB)
2016-03-23 19:18:18.980 [INFO ] - Finished uploading 'R2_ZNF444--BCL2L12.junction.fastq' in 0.06 seconds
2016-03-23 19:18:18.980 [INFO ] - Posting files for AppResult 'BC017_TCAGCCTT'.
2016-03-23 19:18:19.200 [INFO ] - Finished posting files for AppResult 'BC017_TCAGCCTT'
2016-03-23 19:18:19.200 [INFO ] - Finished uploading directory 'BC017_TCAGCCTT'.
2016-03-23 19:18:19.275 [INFO ] - Finished uploading AppResults
2016-03-23 19:18:19.277 [INFO ] - Uploading Samples
2016-03-23 19:18:19.861 [INFO ] - Log file 'spacedock-20160323191819570.log' written to https://aws-api.bases
pace.illumina.com/v1pre3/files/3575978255/content
```

FIGURE 9. Downloadable log files

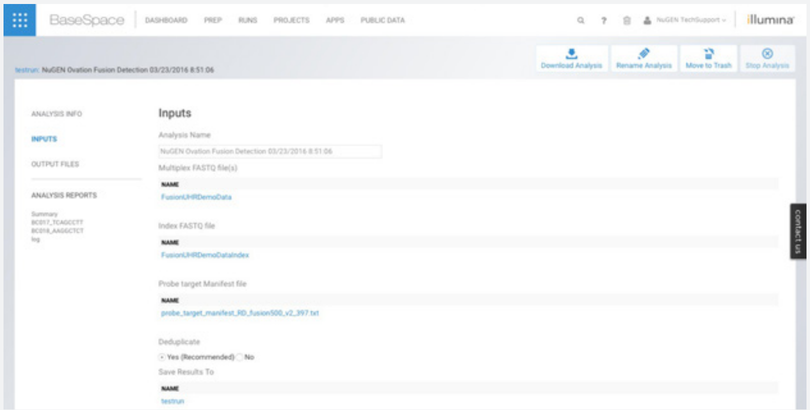


testrun: NuGEN Ovation Fusion Detection 03/23/2016 8:51:06

Log files

NAME	TYPE	PATH	SIZE
output-20160323191115563.log	log	output-20160323191115563.log	49.45 KB
spacedock-20160323191819570.log	log	spacedock-20160323191819570.log	299.95 KB
spacedock-infrastructure-20160323191819864.log	log	spacedock-infrastructure-20160323191819864.log	11.55 MB
uploader-20160323191821395.log	log	uploader-20160323191821395.log	363.61 KB

FIGURE 10. Inputs tab



testrun: NuGEN Ovation Fusion Detection 03/23/2016 8:51:06

Inputs

Analysis Name

NuGEN Ovation Fusion Detection 03/23/2016 8:51:06

Multiplex FASTQ file(s)

NAME
FusionUHRDemoData

Index FASTQ file

NAME
FusionUHRDemoDataIndex

Probe target Manifest file

NAME
probe_target_manifest_RD_fusion500_v2_397.txt

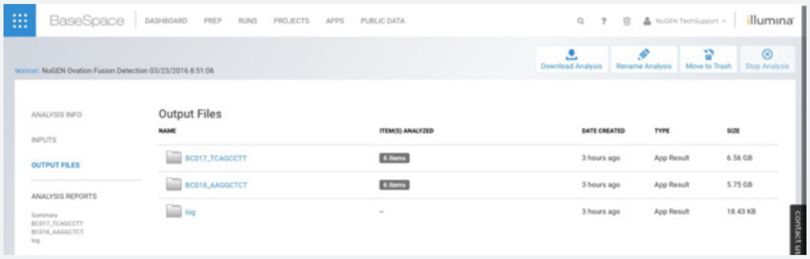
Deduplicate

● Yes (Recommended) ○ No

Save Results To

NAME
testrun

FIGURE 11. Output files



testrun: NuGEN Ovation Fusion Detection 03/23/2016 8:51:06

Output Files

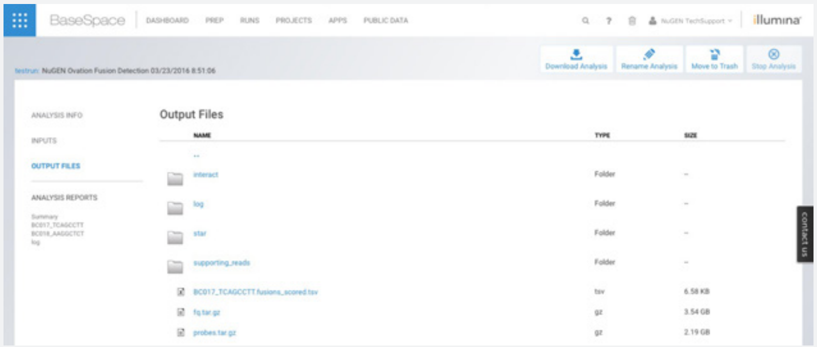
NAME	ITEM(S) ANALYZED	DATE CREATED	TYPE	SIZE
BC017_TCAGCCTT	6 items	3 hours ago	App Result	6.56 GB
BC018_AAGGCTCT	6 items	3 hours ago	App Result	5.75 GB
log	–	3 hours ago	App Result	18.43 KB

- **supporting_reads folder:** Contains supporting junction and spanning reads for each gene fusion identified in the sample.
- **sample.fusions_scored.tsv file:** Tab-delimited summary of the gene fusions identified by the App. This information is provided in the gene fusion table for each sample.
- **fq.tar.gz file:** Fastq file containing the probe assigned, deduplicated, quality trimmed reads for this sample.
- **probes.tar.gz file:** Tab-delimitated sample reads after probe assignment and deduplication.

Summary: Clicking on 'Summary' will show a table that summarizes the probe assignment rate, duplication rate, alignment rate and number of fusions identified for the samples that were processed. Figure 13 shows an example of this table; note below the table is an explanation of each of the metrics that are reported.

Sample: Clicking on each sample shows a graphical representation of a selected fusion and a table of fusions identified in the sample as shown in Figure 14. The graphical representation provides information about the identified fusion sequence, amino acid translation and maintenance of the fusion coding frame. Relevant supporting fusion reads can be quickly downloaded using the provided links under either 'Junction File' or 'Spanning File'. Graphical representations of other identified fusions can be visualized by clicking on the appropriate fusion. Furthermore, clicking on the individual genes involved in the fusion provides a link to the corresponding gene page in the COSMIC database.

FIGURE 12. Sample-specific output files

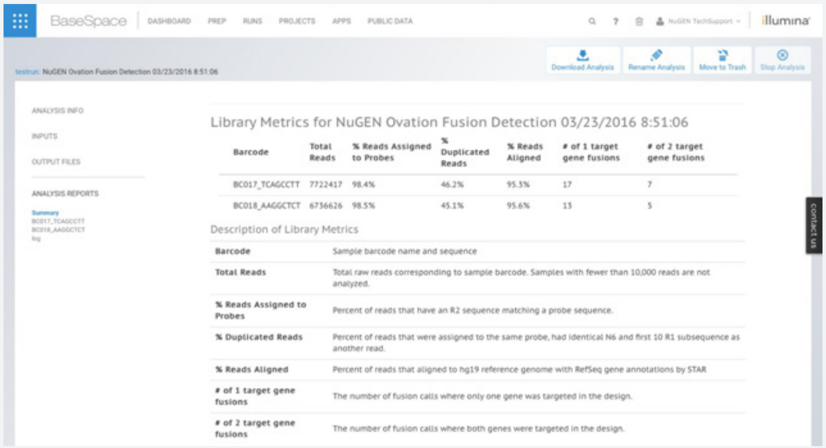


testrun: NuGEN Ovation Fusion Detection 03/23/2016 8:51:06

Output Files

NAME	TYPE	SIZE
..		
interact	Folder	–
log	Folder	–
star	Folder	–
supporting_reads	Folder	–
BC017_TCAGCCTT.fusions_scored.tsv	tsv	6.58 KB
fq.tar.gz	gz	3.54 GB
probes.tar.gz	gz	2.19 GB

FIGURE 13. Library metrics report



testrun: NuGEN Ovation Fusion Detection 03/23/2016 8:51:06

Library Metrics for NuGEN Ovation Fusion Detection 03/23/2016 8:51:06

Barcode	Total Reads	% Reads Assigned to Probes	% Duplicated Reads	% Reads Aligned	# of 1 target gene fusions	# of 2 target gene fusions
BC017_TCAGCCTT	7722417	98.4%	46.2%	95.3%	17	7
BC018_AAGGCTCT	6736626	98.5%	45.1%	95.6%	13	5

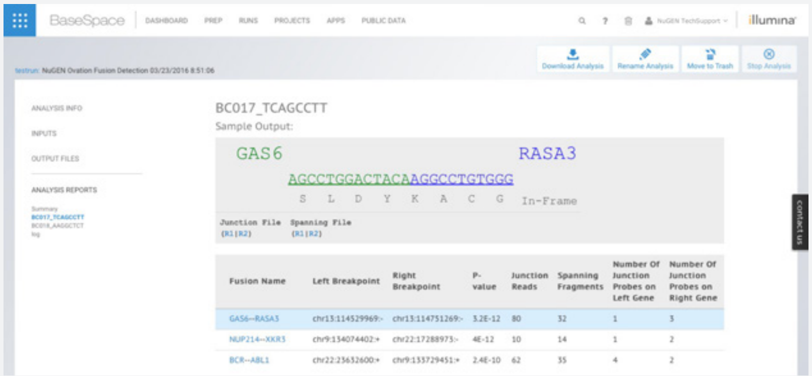
Description of Library Metrics

Barcode	Sample barcode name and sequence
Total Reads	Total raw reads corresponding to sample barcode. Samples with fewer than 10,000 reads are not analyzed.
% Reads Assigned to Probes	Percent of reads that have an R2 sequence matching a probe sequence.
% Duplicated Reads	Percent of reads that were assigned to the same probe, had identical N6 and first 10 R1 subsequence as another read.
% Reads Aligned	Percent of reads that aligned to hg19 reference genome with RefSeq gene annotations by STAR
# of 1 target gene fusions	The number of fusion calls where only one gene was targeted in the design.
# of 2 target gene fusions	The number of fusion calls where both genes were targeted in the design.

The table (shown in Figure 14) provides additional information:

- **Fusion Name:** Genes involved in the identified fusion.
- **Left Breakpoint:** Genomic location of the left gene fusion breakpoint.
- **Right Breakpoint:** Genomic location of the right gene fusion breakpoint.
- **p-value:** Probability of observing a fusion between two genes with equal or better number of supporting probes.
- **Junction reads:** Number of reads that overlap the fusion junction.
- **Spanning reads:** Number of reads that span the two genes involved in the fusion.
- **Number of junction probes on left gene:** Number of probes on the left gene that provide gene fusion supporting junction reads.
- **Number of junction probes on right gene:** Number of probes on the right gene that provide gene fusion supporting junction reads.

FIGURE 14. Analysis report



testrun: NuGEN Ovation Fusion Detection 03/23/2016 8:51:06

BC017_TCAGCCTT

Sample Output:

```
GAS6                                   RASA3
      AGCCTGGACTACAAGGCCTGTGGG
       S  L  D  Y  K  A  C  G  In-Frame
Junction File   Spanning File
(R1|R2)         (R1|R2)
```

Fusion Name	Left Breakpoint	Right Breakpoint	P-value	Junction Reads	Spanning Fragments	Number Of Junction Probes on Left Gene	Number Of Junction Probes on Right Gene
GAS6--RASA3	chr13:114529969:-	chr13:114751269:-	3.2E-12	80	32	1	3
NUP214--XKR3	chr9:134074402:+	chr22:17288973:-	4E-12	10	14	1	2
BCR--ABL1	chr22:23632600:+	chr9:133729451:+	2.4E-10	62	35	4	2

NuGEN Technologies, Inc.

Headquarters USA

201 Industrial Road, Suite 310
San Carlos, CA 94070 USA
Toll Free Tel: 888.654.6544
Toll Free Fax: 888.296.6544
custserv@nugen.com
techserv@nugen.com

Europe

P.O. Box 109
9350 AC Leek
The Netherlands
Tel: +31-13-5780215
Fax: +31-13-5780216
europe@nugen.com

For our international distributors contact information, visit our website

www.nugen.com

For research use only.