Structural Variation Detection using NextGENe® software

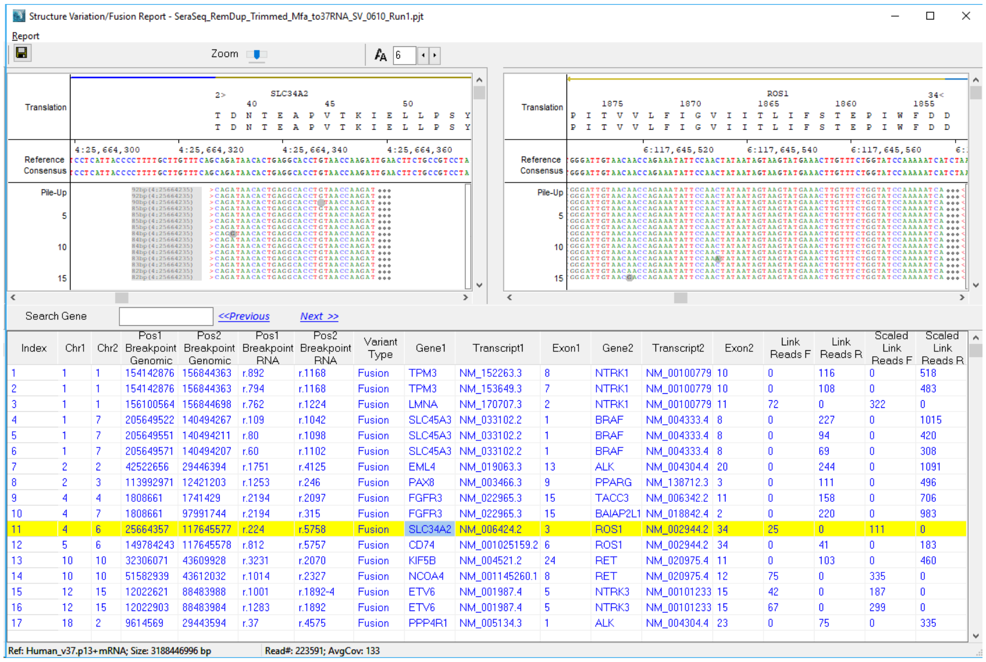
February 2021

Kevin LeVan, Jonathan Liu, Megan McCluskey, Ni Shouyong

Introduction

Over 68,000 structural variants (SVs) have been reported in the human genome, including deletions, duplications/insertions, and translocations such as gene fusions (1). Alone, there are more than 2,200 catalogued gene fusions (2) making genomic disorders an important area of research (3).

There are several techniques available for detecting SVs. Microarrays, for one, can be useful for detecting copy number variation but are limited to detecting events relative to specific design probes (4). Next Generation Sequencing can produce higher resolution results and enable the identification of novel events as well as the simultaneous detection of SNPs and small indels.



Index	Chr1	Chr2	Pos1 Breakpoint Genomic	Pos2 Breakpoint Genomic	Pos1 Breakpoint RNA	Pos2 Breakpoint RNA	Variant Type	Gene1	Transcript1	Exon1	Gene2	Transcript2	Exon2	Link Reads F	Link Reads R	Scaled Link Reads F	Scaled Link Reads R
1	1	1	154142876	156844363	r.892	r.1168	Fusion	TPM3	NM_152263.3	8	NTRK1	NM_00100779	10	0	116	0	518
2	1	1	154142876	156844363	r.794	r.1168	Fusion	TPM3	NM_153649.3	7	NTRK1	NM_00100779	10	0	108	0	483
3	1	1	156100564	156844698	r.762	r.1224	Fusion	LMNA	NM_170707.3	2	NTRK1	NM_00100779	11	72	0	322	0
4	1	7	205649522	140494267	r.109	r.1042	Fusion	SLC45A3	NM_033102.2	1	BRAF	NM_004333.4	8	0	227	0	1015
5	1	7	205649551	140494211	r.80	r.1098	Fusion	SLC45A3	NM_033102.2	1	BRAF	NM_004333.4	8	0	94	0	420
6	1	7	205649571	140494207	r.60	r.1102	Fusion	SLC45A3	NM_033102.2	1	BRAF	NM_004333.4	8	0	69	0	308
7	2	2	42522656	29446394	r.1751	r.4125	Fusion	EML4	NM_019063.3	13	ALK	NM_004304.4	20	0	244	0	1091
8	2	3	113992971	12421203	r.1253	r.246	Fusion	PAX8	NM_003466.3	9	PPARG	NM_138712.3	3	0	111	0	496
9	4	4	1808661	1741429	r.2194	r.2097	Fusion	FGFR3	NM_022965.3	15	TACC3	NM_006342.2	11	0	158	0	706
10	4	7	1808661	97991744	r.2194	r.315	Fusion	FGFR3	NM_022965.3	15	BAIAP2L1	NM_018842.4	2	0	220	0	983
11	4	6	25664357	117645577	r.224	r.5758	Fusion	SLC34A2	NM_006424.2	3	ROS1	NM_002944.2	34	25	0	111	0
12	5	6	149784243	117645578	r.812	r.5757	Fusion	CD74	NM_001025159.2	6	ROS1	NM_002944.2	34	0	41	0	183
13	10	10	32306071	43609928	r.3231	r.2070	Fusion	KIF5B	NM_004521.2	24	RET	NM_020975.4	11	0	103	0	460
14	10	10	51582939	43612032	r.1014	r.2327	Fusion	NCOA4	NM_001145260.1	8	RET	NM_020975.4	12	75	0	335	0
15	12	15	12022621	88483988	r.1001	r.1892-4	Fusion	ETV6	NM_001987.4	5	NTRK3	NM_00101233	15	42	0	187	0
16	12	15	12022903	88483984	r.1283	r.1892	Fusion	ETV6	NM_001987.4	5	NTRK3	NM_00101233	15	67	0	299	0
17	18	2	9614569	29443594	r.37	r.4575	Fusion	PPP4R1	NM_005134.3	1	ALK	NM_004304.4	23	0	75	0	335

Ref: Human_v37.p13+mRNA; Size: 3188446996 bp   Read#: 223591; AvgCov: 133

Figure 1: NextGENe Viewer's Structural Variation Report is an interactive tool for displaying both breakpoints of each detected SVs.

NextGENe software can align reads simultaneously to genomic and transcriptomic references. Reads from RNA samples can align across exon junctions more easily, enabling higher sensitivity for the detection of breakpoints. Next generation sequencing reads from instrument manufactures such as Illumina® can be analyzed by NextGENe software for the detection of SVs. Reads with a high level of mismatch at an end are split into two reads and remapped to the DNA/RNA reference. These Link Reads identify the breakpoints of structural variations including gene fusions.

NextGENe software includes templates to easily analyze gene fusion samples, streamlining the analysis of many SVs.

Procedure

NextGENe software has many integrated features, including tools to import references, to merge paired reads into a single long read, to batch process samples, and to view the results.

Import Reference

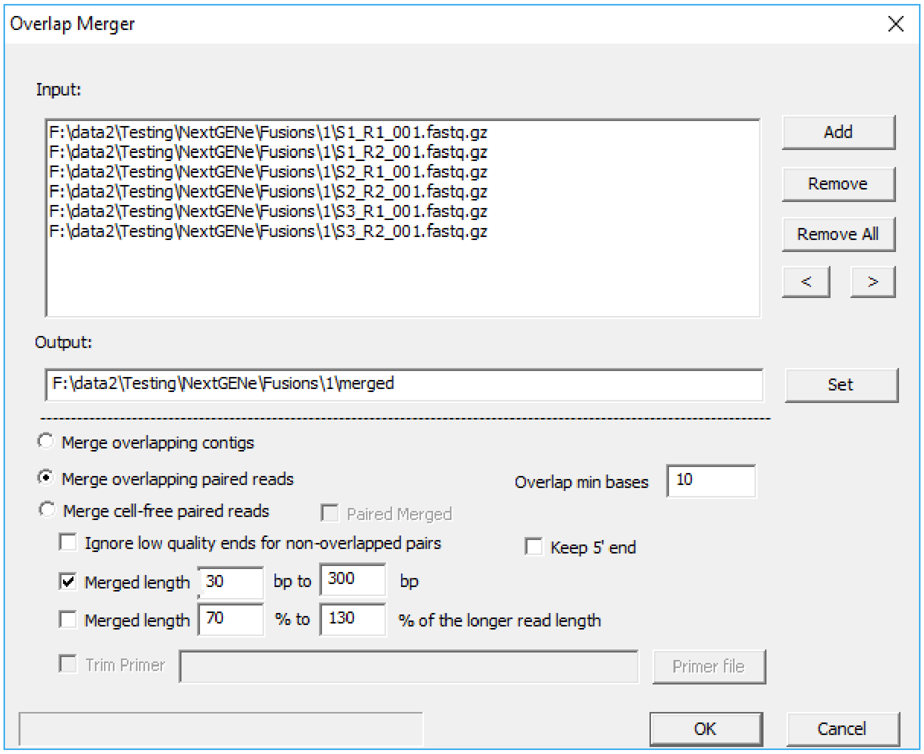
For detection of structural variations with samples derived from RNA, NextGENe can align to a reference that contains both DNA and RNA. Open the NextGENe Tools menu and select the Reference and Track Manager to import references such as Human_v37.p13+mRNA and Human_v38.p12+mRNA.

Overlap Merger Tool

The Overlap Merger tool can be used for Illumina Paired End files to generate a single read from the two pairs (see Figure 2).

- Open the Overlap Merger tool in the NextGENe "Tools" menu
- Add the R1.fastq and R2.fastq files into the Overlap Merger's Input field
- Select the Merge overlapping paired reads option (Overlap min bases 10)
- Select Keep 5' end
- Select Merge length of 30-300bp
- Click OK

The Overlap Merger opens a message when the process is complete. The resultant files with a PairMerge.fasta suffix are ready to be aligned.



Overlap Merger

Input:

F:\data2\Testing\NextGENe\Fusions\1\S1_R1_001.fastq.gz
F:\data2\Testing\NextGENe\Fusions\1\S1_R2_001.fastq.gz
F:\data2\Testing\NextGENe\Fusions\1\S2_R1_001.fastq.gz
F:\data2\Testing\NextGENe\Fusions\1\S2_R2_001.fastq.gz
F:\data2\Testing\NextGENe\Fusions\1\S3_R1_001.fastq.gz
F:\data2\Testing\NextGENe\Fusions\1\S3_R2_001.fastq.gz

Output:

F:\data2\Testing\NextGENe\Fusions\1\merged

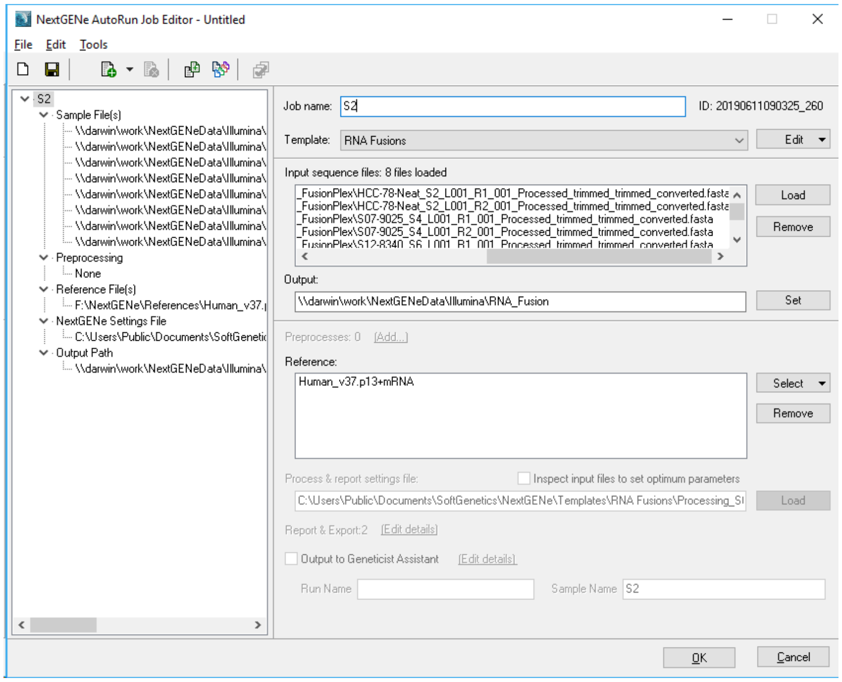
○ Merge overlapping contigs
● Merge overlapping paired reads Overlap min bases 10
○ Merge cell-free paired reads □ Paired Merged
□ Ignore low quality ends for non-overlapped pairs □ Keep 5' end
☑ Merged length 30 bp to 300 bp
□ Merged length 70 % to 130 % of the longer read length
□ Trim Primer

Figure 2: Overlap Merger can be used to combine paired end reads with these settings.

NextGENe AutoRun Tool

A preinstalled template is included in NextGENe software for structural variation detection. Panel templates are used to specify all settings for the analysis when creating ngjob files (see Figure 3). Any setting(s) can be adjusted to create a new template. These ngjob files are detected and processed by the AutoRun tool.

- Open the AutoRun tool from the NextGENe "Tools" menu.
- Open the "Job File Editor" from the AutoRun "Tool" menu.
- Select a template from the drop-down list. A template for the structural variation detection is 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.
- Load the PairMerge.fasta output from the Overlap Merger Tool.
- Select one of the Human+mRNA preloaded references to use for alignment.
- Set an output location.
- 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.
- Click OK to save the ngjob file.
- Start the AutoRun tool to begin processing.



S2
- Sample File(s)
  - \\darwin\work\NextGENeData\Illumina\
  - \\darwin\work\NextGENeData\Illumina\
  - \\darwin\work\NextGENeData\Illumina\
  - \\darwin\work\NextGENeData\Illumina\
  - \\darwin\work\NextGENeData\Illumina\
  - \\darwin\work\NextGENeData\Illumina\
  - \\darwin\work\NextGENeData\Illumina\
  - \\darwin\work\NextGENeData\Illumina\
- Preprocessing
  - None
- Reference File(s)
  - F:\NextGENe\References\Human_v37.|
- NextGENe Settings File
  - C:\Users\Public\Documents\SoftGenetic
- Output Path
  - \\darwin\work\NextGENeData\Illumina\

Job name: S2 ID: 20190611090325_260

Template: RNA Fusions Edit

Input sequence files: 8 files loaded

```
_FusionPlex\HCC-78-Neat_S2_L001_R1_001_Processed_trimmed_trimmed_converted.fasta
_FusionPlex\HCC-78-Neat_S2_L001_R2_001_Processed_trimmed_trimmed_converted.fasta
_FusionPlex\S07-9025_S4_L001_R1_001_Processed_trimmed_trimmed_converted.fasta
_FusionPlex\S07-9025_S4_L001_R2_001_Processed_trimmed_trimmed_converted.fasta
_FusionPlex\S12-8340_S6_L001_R1_001_Processed_trimmed_trimmed_converted.fasta
```

Output: \\darwin\work\NextGENeData\Illumina\RNA_Fusion

Preprocesses: 0 (Add...)

Reference: Human_v37.p13+mRNA

Process & report settings file: □ Inspect input files to set optimum parameters

C:\Users\Public\Documents\SoftGenetics\NextGENe\Templates\RNA Fusions\Processing_SI

Report & Export:2 (Edit details)

□ Output to Geneticist Assistant (Edit details)

Run Name: Sample Name: S2

Figure 3: AutoRun can be used to save panel templates and batch process samples.

Results

The Seraseq Fusion RNA Mix v2 from SeraCare Life Sciences, Inc. includes many gene fusions, like TPM3-NTRK1, LMNA-NTRK1, SLC45A3-BRAF, EML4-ALK, PAX8-PPARG, FGFR3-TACC3, FGFR3-BAIAP2L1, SLC34A2-ROS1, CD74-ROS1, KIF5B-RET, NCOA4-RET, and ETV6-NTRK3.

A sample was run on an Illumina MiSeq instrument and produced about 255K 2x150bps reads. Trimming of adapters yielded reads averaging 112bp. After utilizing the Overlap Merger tool, 90.5% of the pairs were merged into single reads with an average length of 140bp, extending the average length by about 30bp, (see Figure 4).

NextGENe software was run on a standard Windows Desktop computer. 223K (96.4%) of these reads are aligned to the DNA/RNA reference genome in 3 minutes and Structural Variation detection is completed in 8 additional minutes by NextGENe software. Reports in TXT, VCF and PDF formats are automatically generated.

The final project can be opened in the NextGENe Viewer software, which offers an interactive tool for viewing the SVs and producing additional reports (see Figure 1). Reads that are split and mapped across breakpoints are called Link Reads, and about 2K of these reads are mapping to all 12 expected fusion events in the SeraCare fusion controls.

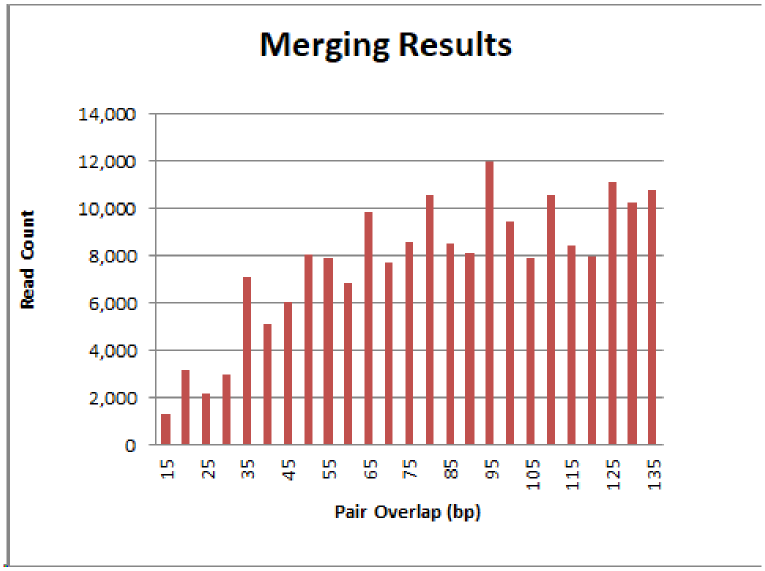


Figure 4: Overlap Merger converted over 200K paired end reads into single reads averaging 140bp. The original paired reads averaged 112bp, and about 2000 pairs had only 15 bases of overlap with each other, roughly doubling the read length.:

Discussion

One feature of the NextGENe Overlap Merger tool can create a single read from paired end reads. Depending on the template size, this could increase 2x150bp reads to a single 290bp read. This step utilizes the quality information within the FASTQ files to assemble the pair with the highest base quality. Longer reads can span more exons, improving the accuracy structural variation breakpoint detection. NextGENe software contains a powerful tool for the analysis of structural variations from high throughput sequencers. Reads are aligned to a reference including both DNA and RNA to enhance sensitivity, reads spanning SVs are split into pseudo paired reads to better identify breakpoints.

NextGENe software is widely used for the analysis of samples from Illumina and Ion Torrent instruments and can also be used for applications such as SNP/Indel discovery, including somatic detection and rare disease research, Copy Number Variation detection, RNA-Seq, de novo assembly and more.

References

1. An integrated map of structural variation in 2,504 human genomes. **Sudmant, Peter H., Eichler, Evan E. and Korbel, Jan O**. s.l. : Nature, 2015, Nature, Vol. 526, pp. 75-81. 75-81.

2. A comprehensive transcriptional portrait of human cancer cell lines. **Klijn, Christiaan and Zhang, Zemin**. s.l. : Nature Biotechnology, 2014, Vol. 33. 306-312.

3. Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. **Shaw, Christine J. and Lupski, James R.** s.l. : Human Molecular Genetics, 2004, Vol. 13. 57–64.

4. Genome structural variation discovery and genotyping. **Alkan, Can, Coe, Bradley P. and Eichler, Evan E.** s.l. : Nature Reviews Genetics, 2011, Vol. 12. 363-376.