Updated July 31, 2013

TCGA mRNA-seq Pipeline for UNC data

This document provides a detailed knowledge base of mRNA-seq data processing by UNC as part of the Cancer Genome Atlas Project. Here we provide the references, commands, and known caveats of the bams deposited at CGHub by UNC. These methods are also relevant to the level 3 data for the 'rnaseqv2' platform available at the TCGA Data Coordinating Center (DCC). Further information regarding specific laboratory and sequencing protocols are available in the IDF and SDRF files at the DCC.

Software and References

Mapsplice v12_07	https://webshare.bioinf.unc.edu/public/mRNAseq_TCGA/MapSplice_multithreads_12_07.tar.gz
Mapsplice v2.0.1.9	https://webshare.bioinf.unc.edu/public/mRNAseq_TCGA/MapSplice_multithreads_2.0.1.9.tar.gz
RSEM v1.1.13	http://deweylab.biostat.wisc.edu/rsem/src/rsem-1.1.13.tar.gz
UBU v1.0	https://github.com/mozack/ubu
Reference data	https://webshare.bioinf.unc.edu/public/mRNAseq_TCGA/ Alignments are against hg19 + chrM_rCRS Reformatted versions of TCGA GAF 2.1 files are used for isoform and gene definitions

Workflow Commands

The default workflow is provided below. This workflow used Mapsplice v12_07. Due to sample and protocol differences, some tumor types deviated from this workflow. Mapsplice v2.0.1.9 was used for HNSC, GBM, and OV. Single end options in Mapsplice v2.0.1.9 were used for READ, COAD, and UCEC single end sequenced samples.

1. Format fastq 1 for Mapsplice

```
java -Xmx512M -jar ubu.jar fastq-format --phred33to64 --strip --suffix /1 –in raw_1.fastq --out working/prep_1.fastq >
working/mapsplice_prep1.log
```

2. Format fastq 2 for Mapsplice

```
java -Xmx512M -jar ubu.jar fastq-format --phred33to64 --strip --suffix /2 –in raw_2.fastq --out working/prep_2.fastq >
working/mapsplice_prep2.log
```

3. Mapsplice

```
python mapsplice_multi_thread.py --fusion --all-chromosomes-files hg19_M_rCRS/hg19_M_rCRS.fa --pairend -X 8 -Q fq --chromosome-files-
dir hg19_M_rCRS/chromosomes --Bowtieidx  hg19_M_rCRS/ebwt/humanchridx_M_rCRS -1 working/prep_1.fastq -2 working/prep_2.fastq -o
SAMPLE_BARCODE 2> working/mapsplice.log
```

4. Add read groups

```
java -Xmx2G -jar AddOrReplaceReadGroups.jar INPUT=alignments.bam OUTPUT=working/rg_alignments.bam RGSM=
SAMPLE_BARCODE RGID= SAMPLE_BARCODE RGLB=TruSeq RGPL=illumina RGPU=barcode
VALIDATION_STRINGENCY=SILENT TMP_DIR=working/add_rg_tag_tmp > working/add_rg_tag.log 2> working/add_rg_tag.log
```

5. Convert back to phred33

```
java -Xmx512M -jar ubu.jar sam-convert --phred64to33 --in working/rg_alignments.bam –out working/phred33_alignments.bam >
working/sam_convert.log 2> working/sam_convert.log
```

6. Sort by coordinate

```
samtools sort working/phred33_alignments.bam sorted_genome_alignments
```

7. Flagstat

```
samtools flagstat sorted_genome_alignments.bam > sorted_genome_alignments.bam.flagstat
```

8. Index

```
samtools index sorted_genome_alignments.bam
```

9. Sort by chromosome, then read id

```
perl sort_bam_by_reference_and_name.pl --input sorted_genome_alignments.bam –output working/sorted_by_chr_read.bam --temp-dir . –samtools /datastore/tier1data/nextgenseq/seqware-analysis/software/samtools/samtools > working/sorted_by_chr_read.log 2> working/sorted_by_chr_read.log
```

10. Translate to transcriptome coords

```
java -Xms3G -Xmx3G -jar ubu.jar sam-xlate --bed unc_hg19.bed –in working/sorted_by_chr_read.bam --out working/transcriptome_alignments.bam –order rsem_ref/hg19_M_rCRS_ref.transcripts.fa --xgtags --reverse > working/genome_to_transcriptome.log 2> working/genome_to_transcriptome.log
```

11. Filter indels, large inserts, zero mapping quality from transcriptome bam

```
java -Xmx512M -jar ubu.jar sam-filter --in working/transcriptome_alignments.bam –out working/transcriptome_alignments_filtered.bam --strip-indels --max-insert 10000 --mapq 1 > working/sam_filter.log 2> working/sam_filter.log
```

12. RSEM

```
rsem-calculate-expression --gcr-output-file --paired-end --bam --estimate-rspd -p 8 working/transcriptome_alignments_filtered.bam /datastore/tier1data/nextgenseq/seqware-analysis/mapsplice_rsem/rsem_ref/hg19_M_rCRS_ref rsem > working/rsem.log 2> working/rsem.log
```

13. Strip trailing tabs from rsem.isoforms.results

```
perl strip_trailing_tabs.pl --input rsem.isoforms.results --temp working/orig.isoforms.results > working/trim_isoform_tabs.log 2> working/trim_isoform_tabs.log
```

14. Prune isoforms from gene quant file

```
mv rsem.genes.results working/orig.genes.results; sed /^uc0/d working/orig.genes.results > rsem.genes.results
```

15. Normalize gene quant

```
perl quartile_norm.pl -c 2 -q 75 -t 1000 -o rsem.genes.normalized_results rsem.genes.results
```

16. Normalize isoform quant

```
perl quartile_norm.pl -c 2 -q 75 -t 300 -o rsem.isoforms.normalized_results rsem.isoforms.results
```

17. Junction counts

```
java -Xmx512M -jar ubu.jar sam-junc --junctions splice_junctions.txt --in sorted_genome_alignments.bam --out junction_quantification.txt > junction_quantification.log 2> junction_quantification.log
```

18. Exon counts

```
coverageBed -split -abam sorted_genome_alignments.bam -b composite_exons.bed | perl normalizeBedToolsExonQuant.pl composite_exons.bed > bt.exon_quantification.txt 2> bt_exon_quantification.log
```

19. Cleanup large intermediate output

```
rm alignments.bam logs/* working/phred33_alignments.bam working/rg_alignments.bam working/sorted_by_chr_read.bam working/transcriptome_alignments.bam working/transcriptome_alignments_filtered.bam working/prep_1.fastq working/prep_2.fastq > working/cleanup.log
```

Incompatibility of mRNA-seq bams with SAM 1.4

Upon initiation of mRNA-seq data generation by TCGA, publicly available splice aware alignment algorithms did not conform perfectly to the SAM 1.4 specification. MapSplice was chosen based on its performance and features to provide high quality mRNA alignments to TCGA investigators. However, the MapSplice incompatibilities with SAM 1.4, detailed below, prevent accurate reconstruction of the original fastq file using typical tools (i.e. samtools, picard). Thus original fastq files are currently being made available at CGHub and investigators are encouraged to utilize those fastq files when re-analysis from raw sequence is required. Thanks to Olena Morozova and the UCSC team for identifying and generating examples of these incompatibilities.

The incompatibilities below refer specifically to MapSplice versions 12.07 and 2.0.1.9. With the exception of item 1, we have found these incompatibilities apply to a relatively small fraction of reads.

Example records were generated using the command:

```
$ samtools view -X UNCID_1372446.9f678cb6-26c7-4c76-831c-e541319648d1.sorted_genome_alignments.bam | cut -f 1-9
```

1. The /1 or /2 is included in the QNAME. The developers retained these in order to disambiguate read pairs because Mapsplice's interpretation of the paired and proper pair bit flags for fusions, multi-mappers, and unaligned reads precludes typical segregation (see items 2-5). This is the only incompatibility attributable to all records.

```
UNC13-SN749:179:C0V5LACXX:8:1105:15090:63585/2 pPR2 chr1 12078 66 48M = 12140 110
UNC13-SN749:179:C0V5LACXX:8:1105:15090:63585/1 pPr1 chr1 12140 66 48M = 12078 -110
```

2. The properPair (0x0002) bit for secondary, read2, properly paired mappings is not set.

```
UNC13-SN749:179:C0V5LACXX:8:1105:16559:107025/2 pR2s chr1 11630 39 48M = 12058 476
UNC13-SN749:179:C0V5LACXX:8:1105:16559:107025/1 pPr1s chr1 12058 66 48M = 11630 -476
```

3. The paired sequencing (0x0001), read1 (0x0040), or read2 (0x0080) bits are not set for unmapped reads.

```
UNC13-SN749:179:C0V5LACXX:8:1101:2352:2224/1 u * 0 0 * * 0 0
```

4. The properPair (0x0002) bit for candidate fusion alignments is set when the mate is mapped to different chromosome or on the same strand.

```
UNC13-SN749:179:C0V5LACXX:8:2103:3131:103020/2 pP2 chr1 564475 56 48M chrM_rCRS 3953 560569
```

5. Secondary/not-secondary mappings in ProperPair mapping are set incorrectly for candidate fusion alignments.

```
UNC13-SN749:180:D127FACXX:2:1101:1122:191492/1 pPrR1 chr1 565638 63 48M chrM_rCRS 4968 560717
UNC13-SN749:180:D127FACXX:2:1101:1122:191492/2 pPrR2 chrM_rCRS 4968 44 48M = 5088 -73
UNC13-SN749:180:D127FACXX:2:1101:1122:191492/1 pPrR1s chrM_rCRS 5088 63 48M = 4968 167
```

6. The following incompatibilities (6,7) regard candidate fusion alignments, including candidate fusion alignments that are ultimately reported as unaligned. Read1, read2 (0x00c0) bits and TLEN values are erroneous.

```
UNC13-SN749:179:C0V5LACXX:8:2107:20483:40121/1 pPR12s chr1 564732 255 37M11S chrM_rCRS 4187 560545
```

7. The reverse-complement of SEQ is not reported when the reverse-complement (0x0010) bit is set, and the order of QUAL cannot be determined.

8. Mate information is occasionally not recorded

```
UNC13-SN749:179:C0V5LACXX:8:2108:14392:110369/1 pPR1    chr1    564744  255     25M23S  *       0       0
```

Also of note, multi-mappers are reported as separate lines. This is not explicitly forbidden in the SAM 1.4 specification, and this format of reporting multi-mappers was retained due to the requirement (at the time of development) by other downstream RNA-seq algorithms such as Cufflinks and RSEM.