PAM-read analysis demonstration

08 October 2010
Mark Okada

This document is a walkthrough of the PAM-read analysis using the demo dataset.
Refer to Demo_datasets.doc and README.txt for detailed information on the steps and output formats.

1 Poly-A Mate (PAM) Read analysis

Demo data set

```
[mokada@xhost09 data_pam]$ ls -l
total 28
-rw-r--r-- 1 mokada users 1292 Oct  5 12:48 cmd.txt
-rw-r--r-- 1 mokada users 2261 Oct  5 01:04 contigs.fa
-rw-r--r-- 1 mokada users 7827 Oct  5 01:40 genes.fa
-rw-r--r-- 1 mokada users 3996 Oct  5 01:07 pam_forward.txt
-rw-r--r-- 1 mokada users 3960 Oct  5 01:07 pam_reverse.txt
-rw-r--r-- 1 mokada users  227 Oct  5 02:07 polyafinder.conf
```

Extract the PAM reads from the raw input reads

```
[mokada@xhost09 data_pam]$ polyareads.pl -p -f pam_forward.txt -r pam_reverse.txt -config
polyafinder.conf
# Extracting forward reads pam_forward.txt
/home/mokada/wtss/bin/extractCutoffPolyA.pl -a -c 40 -f pam_forward.txt -r pam_reverse.txt -of
polya_out.pam.1.forwardT.fastq -or polya_out.pam.2.reverseT.fastq
```

Perform the indexing, alignment and analysis. Outputs tab delimited files (.tsv).

```
[mokada@xhost09 data_pam]$ polyafinder.pl -f polya_out.pam.1.forwardT.fastq -r
polya_out.pam.2.reverseT.fastq -t genes.fa -a -cont contigs.fa -capp -config polyafinder.conf
# Appending poly-A file polya_out.transPolyA.fa from transcript file(s) genes.fa
cat genes.fa | /home/mokada/wtss/bin/appendpolyAs.pl -a -length 50 -getLength -dup > polya_out.transPolyA.fa

# Indexing transcript file polya_out.transPolyA.fa
/home/pubseq/BioSw/bwa/bwa-0.5.4/bwa index -a is polya_out.transPolyA.fa
[bwa_index] Pack FASTA... 0.00 sec
[bwa_index] Reverse the packed sequence... 0.00 sec
[bwa_index] Construct BWT for the packed sequence...
[bwa_index] 0.00 seconds elapse.
[bwa_index] Construct BWT for the reverse packed sequence...
[bwa_index] 0.00 seconds elapse.
[bwa_index] Update BWT... 0.00 sec
[bwa_index] Update reverse BWT... 0.00 sec
[bwa_index] Construct SA from BWT and Occ... 0.00 sec
[bwa_index] Construct SA from reverse BWT and Occ... 0.00 sec

# Aligning forward reads to transcripts
/home/pubseq/BioSw/bwa/bwa-0.5.4/bwa aln polya_out.transPolyA.fa polya_out.pam.1.forwardT.fastq >
polya_out.pam.aln.forward.sai
[bwa_aln] 17bp reads: max_diff = 2
[bwa_aln] 38bp reads: max_diff = 3
[bwa_aln] 64bp reads: max_diff = 4
[bwa_aln] 93bp reads: max_diff = 5
[bwa_aln] 124bp reads: max_diff = 6
[bwa_aln] 157bp reads: max_diff = 7
[bwa_aln] 190bp reads: max_diff = 8
[bwa_aln] 225bp reads: max_diff = 9
[bwa_aln_core] calculate SA coordinate... 0.00 sec
```

```
[bwa_aln_core] write to the disk... 0.00 sec
[bwa_aln_core] 12 sequences have been processed.

# Aligning reverse reads to transcripts
/home/pubseq/BioSw/bwa/bwa-0.5.4/bwa aln polya_out.transPolyA.fa polya_out.pam.2.reverseT.fastq >
polya_out.pam.aln.reverse.sai
[bwa_aln] 17bp reads: max_diff = 2
[bwa_aln] 38bp reads: max_diff = 3
[bwa_aln] 64bp reads: max_diff = 4
[bwa_aln] 93bp reads: max_diff = 5
[bwa_aln] 124bp reads: max_diff = 6
[bwa_aln] 157bp reads: max_diff = 7
[bwa_aln] 190bp reads: max_diff = 8
[bwa_aln] 225bp reads: max_diff = 9
[bwa_aln_core] calculate SA coordinate... 0.00 sec
[bwa_aln_core] write to the disk... 0.00 sec
[bwa_aln_core] 13 sequences have been processed.

# Saving forward read alignments to TSV format
/home/pubseq/BioSw/bwa/bwa-0.5.4/bwa samse -n 20 polya_out.transPolyA.fa polya_out.pam.aln.forward.sai
polya_out.pam.1.forwardT.fastq | /home/mokada/wtss/bin/samse2csv.pl -fromend -u
polya_out.pam.aln.forward.unmapped > polya_out.pam.aln.forward.tsv

# Saving reverse read alignments to TSV format
/home/pubseq/BioSw/bwa/bwa-0.5.4/bwa samse -n 20 polya_out.transPolyA.fa polya_out.pam.aln.reverse.sai
polya_out.pam.2.reverseT.fastq | /home/mokada/wtss/bin/samse2csv.pl -fromend -u
polya_out.pam.aln.reverse.unmapped > polya_out.pam.aln.reverse.tsv

################################
# Created PAM alignment table output:
# Forward PAM alignments: polya_out.pam.aln.forward.tsv
# Reverse PAM alignments: polya_out.pam.aln.reverse.tsv
#

# Appending polyA and polyT to contigs
cat contigs.fa | /home/mokada/wtss/bin/appendpolyAs.pl -t -a -length 50 -getLength >
polya_out.contigPolyAT.fa
# Indexing contig file polya_out.contigPolyAT.fa
/home/pubseq/BioSw/bwa/bwa-0.5.4/bwa index -a is polya_out.contigPolyAT.fa
[bwa_index] Pack FASTA... 0.00 sec
[bwa_index] Reverse the packed sequence... 0.00 sec
[bwa_index] Construct BWT for the packed sequence...
[bwa_index] 0.00 seconds elapse.
[bwa_index] Construct BWT for the reverse packed sequence...
[bwa_index] 0.00 seconds elapse.
[bwa_index] Update BWT... 0.00 sec
[bwa_index] Update reverse BWT... 0.00 sec
[bwa_index] Construct SA from BWT and Occ... 0.00 sec
[bwa_index] Construct SA from reverse BWT and Occ... 0.00 sec

# Aligning forward reads to contigs
/home/pubseq/BioSw/bwa/bwa-0.5.4/bwa aln polya_out.contigPolyAT.fa polya_out.pam.1.forwardT.fastq >
polya_out.pam.contig.aln.forward.sai
[bwa_aln] 17bp reads: max_diff = 2
[bwa_aln] 38bp reads: max_diff = 3
[bwa_aln] 64bp reads: max_diff = 4
[bwa_aln] 93bp reads: max_diff = 5
[bwa_aln] 124bp reads: max_diff = 6
[bwa_aln] 157bp reads: max_diff = 7
[bwa_aln] 190bp reads: max_diff = 8
[bwa_aln] 225bp reads: max_diff = 9
[bwa_aln_core] calculate SA coordinate... 0.00 sec
[bwa_aln_core] write to the disk... 0.00 sec
[bwa_aln_core] 12 sequences have been processed.

# Aligning reverse reads to contigs
/home/pubseq/BioSw/bwa/bwa-0.5.4/bwa aln polya_out.contigPolyAT.fa polya_out.pam.2.reverseT.fastq >
polya_out.pam.contig.aln.reverse.sai
[bwa_aln] 17bp reads: max_diff = 2
[bwa_aln] 38bp reads: max_diff = 3
[bwa_aln] 64bp reads: max_diff = 4
[bwa_aln] 93bp reads: max_diff = 5
[bwa_aln] 124bp reads: max_diff = 6
[bwa_aln] 157bp reads: max_diff = 7
[bwa_aln] 190bp reads: max_diff = 8
[bwa_aln] 225bp reads: max_diff = 9
```

```
[bwa_aln_core] calculate SA coordinate... 0.00 sec
[bwa_aln_core] write to the disk... 0.00 sec
[bwa_aln_core] 13 sequences have been processed.

# Saving forward read alignments to TSV format
/home/pubseq/BioSw/bwa/bwa-0.5.4/bwa samse -n 20 polya_out.contigPolyAT.fa
polya_out.pam.contig.aln.forward.sai polya_out.pam.1.forwardT.fastq | /home/mokada/wtss/bin/samse2csv.pl -
fromend -u polya_out.pam.contig.aln.forward.unmapped > polya_out.pam.contig.aln.forward.tsv

# Saving reverset read alignments to TSV format
/home/pubseq/BioSw/bwa/bwa-0.5.4/bwa samse -n 20 polya_out.contigPolyAT.fa
polya_out.pam.contig.aln.reverse.sai polya_out.pam.2.reverseT.fastq | /home/mokada/wtss/bin/samse2csv.pl -
fromend -u polya_out.pam.contig.aln.reverse.unmapped > polya_out.pam.contig.aln.reverse.tsv

# Find forward reads that map only to contigs
/home/mokada/wtss/bin/getallunmappedfromTSV.pl polya_out.pam.aln.forward.unmapped
polya_out.pam.contig.aln.forward.tsv > polya_out.pam.contig.remapped.forward.tsv

# Find reverset reads that map only to contigs
/home/mokada/wtss/bin/getallunmappedfromTSV.pl polya_out.pam.aln.reverse.unmapped
polya_out.pam.contig.aln.reverse.tsv > polya_out.pam.contig.remapped.reverse.tsv

################################
# Created PAM contig alignment table output:
# Forward PAM contig alignments: polya_out.pam.contig.remapped.forward.tsv
# Reverse PAM contig alignments: polya_out.pam.contig.remapped.reverse.tsv
#


[mokada@xhost09 data_pam]$ ls -l
total 160
-rw-r--r-- 1 mokada users 1292 Oct  5 12:48 cmd.txt
-rw-r--r-- 1 mokada users 2261 Oct  5 01:04 contigs.fa
-rw-r--r-- 1 mokada users 7827 Oct  5 01:40 genes.fa
-rw-r--r-- 1 mokada users 3996 Oct  5 01:07 pam_forward.txt
-rw-r--r-- 1 mokada users 3960 Oct  5 01:07 pam_reverse.txt
-rw-r--r-- 1 mokada users  227 Oct  5 02:07 polyafinder.conf
-rw-r--r-- 1 mokada users 2515 Oct  5 12:20 polya_out.contigPolyAT.fa
-rw-r--r-- 1 mokada users    9 Oct  5 12:20 polya_out.contigPolyAT.fa.amb
-rw-r--r-- 1 mokada users   99 Oct  5 12:20 polya_out.contigPolyAT.fa.ann
-rw-r--r-- 1 mokada users  944 Oct  5 12:20 polya_out.contigPolyAT.fa.bwt
-rw-r--r-- 1 mokada users  606 Oct  5 12:20 polya_out.contigPolyAT.fa.pac
-rw-r--r-- 1 mokada users  944 Oct  5 12:20 polya_out.contigPolyAT.fa.rbwt
-rw-r--r-- 1 mokada users  606 Oct  5 12:20 polya_out.contigPolyAT.fa.rpac
-rw-r--r-- 1 mokada users  328 Oct  5 12:20 polya_out.contigPolyAT.fa.rsa
-rw-r--r-- 1 mokada users  328 Oct  5 12:20 polya_out.contigPolyAT.fa.sa
-rw-r--r-- 1 mokada users 1619 Oct  5 12:20 polya_out.pam.1.forwardT.fastq
-rw-r--r-- 1 mokada users 1755 Oct  5 12:20 polya_out.pam.2.reverseT.fastq
-rw-r--r-- 1 mokada users  208 Oct  5 12:20 polya_out.pam.aln.forward.sai
-rw-r--r-- 1 mokada users  403 Oct  5 12:20 polya_out.pam.aln.forward.tsv
-rw-r--r-- 1 mokada users  200 Oct  5 12:20 polya_out.pam.aln.forward.unmapped
-rw-r--r-- 1 mokada users  292 Oct  5 12:20 polya_out.pam.aln.reverse.sai
-rw-r--r-- 1 mokada users  799 Oct  5 12:20 polya_out.pam.aln.reverse.tsv
-rw-r--r-- 1 mokada users   67 Oct  5 12:20 polya_out.pam.aln.reverse.unmapped
-rw-r--r-- 1 mokada users  208 Oct  5 12:20 polya_out.pam.contig.aln.forward.sai
-rw-r--r-- 1 mokada users  789 Oct  5 12:20 polya_out.pam.contig.aln.forward.tsv
-rw-r--r-- 1 mokada users  195 Oct  5 12:20 polya_out.pam.contig.aln.forward.unmapped
-rw-r--r-- 1 mokada users  148 Oct  5 12:20 polya_out.pam.contig.aln.reverse.sai
-rw-r--r-- 1 mokada users  264 Oct  5 12:20 polya_out.pam.contig.aln.reverse.tsv
-rw-r--r-- 1 mokada users  362 Oct  5 12:20 polya_out.pam.contig.aln.reverse.unmapped
-rw-r--r-- 1 mokada users  789 Oct  5 12:20 polya_out.pam.contig.remapped.forward.tsv
-rw-r--r-- 1 mokada users  264 Oct  5 12:20 polya_out.pam.contig.remapped.reverse.tsv
-rw-r--r-- 1 mokada users 8025 Oct  5 12:20 polya_out.transPolyA.fa
-rw-r--r-- 1 mokada users    9 Oct  5 12:20 polya_out.transPolyA.fa.amb
-rw-r--r-- 1 mokada users  189 Oct  5 12:20 polya_out.transPolyA.fa.ann
-rw-r--r-- 1 mokada users 2964 Oct  5 12:20 polya_out.transPolyA.fa.bwt
-rw-r--r-- 1 mokada users 1953 Oct  5 12:20 polya_out.transPolyA.fa.pac
-rw-r--r-- 1 mokada users 2964 Oct  5 12:20 polya_out.transPolyA.fa.rbwt
-rw-r--r-- 1 mokada users 1953 Oct  5 12:20 polya_out.transPolyA.fa.rpac
-rw-r--r-- 1 mokada users 1000 Oct  5 12:20 polya_out.transPolyA.fa.rsa
-rw-r--r-- 1 mokada users 1000 Oct  5 12:20 polya_out.transPolyA.fa.sa
```

2. End Junction reads

Extract the EJ reads from the raw reads

```
[mokada@xhost09 data_ej]$ polyareads.pl -e -f ej_forward.txt -r ej_reverse.txt

# Extracting forward reads ej_forward.txt
cat ej_forward.txt | /home/mokada/wtss/bin/getrawreadendtag.pl -n >> polya_out.ej.1.forwardT.fastq

# Extracting reverse EJ reads ej_reverse.txt
cat ej_reverse.txt | /home/mokada/wtss/bin/getrawreadendtag.pl -n >> polya_out.ej.2.reverseT.fastq

# Getting mate read names forward-T
cat polya_out.ej.1.forwardT.fastq polya_out.ej.2.reverseT.fastq | /home/mokada/wtss/bin/getejreadnames.pl >
polya_out.ej.readnames.tsv

# Extracting mate of EJ read pairs forward-T
cat ej_forward.txt | /home/mokada/wtss/bin/extractmatereads.pl polya_out.ej.readnames.tsv >>
polya_out.ej.1.mate.fastq

# Extracting mate of EJ reads reverse-T
cat ej_reverse.txt | /home/mokada/wtss/bin/extractmatereads.pl polya_out.ej.readnames.tsv >>
polya_out.ej.2.mate.fastq
```

Perform the alignment to the genes and analyse. Output is the tab delimited file .tsv. The mate pair files are those EJ reads and alignments which have a mapped mate pair upstream from the polyadenylation site, within range and on the opposite strand.

```
[mokada@xhost09 data_ej]$ polyafinder.pl -e -a -f polya_out.ej.1.forwardT.fastq -r
polya_out.ej.2.reverseT.fastq -t genes.fa -mf polya_out.ej.1.mate.fastq -mr
polya_out.ej.2.mate.fastq -conf polyafinder.conf

# Appending poly-A file polya_out.transPolyA.fa from transcript file(s) genes.fa
cat genes.fa | /home/mokada/wtss/bin/appendpolyAs.pl -a -length 50 -getLength -dup > polya_out.transPolyA.fa

# Indexing transcript file polya_out.transPolyA.fa
/home/pubseq/BioSw/bwa/bwa-0.5.4/bwa index -a is polya_out.transPolyA.fa
[bwa_index] Pack FASTA... 0.00 sec
[bwa_index] Reverse the packed sequence... 0.00 sec
[bwa_index] Construct BWT for the packed sequence...
[bwa_index] 0.00 seconds elapse.
[bwa_index] Construct BWT for the reverse packed sequence...
[bwa_index] 0.00 seconds elapse.
[bwa_index] Update BWT... 0.00 sec
[bwa_index] Update reverse BWT... 0.00 sec
[bwa_index] Construct SA from BWT and Occ... 0.00 sec
[bwa_index] Construct SA from reverse BWT and Occ... 0.00 sec

# Calculating number of poly-A's in each transcript
/home/mokada/wtss/bin/calcPolyAinTails.pl -i polya_out.transPolyA.fa -o -50 >
polya_out.transPolyA.fa.taillengths.txt

# Aligning forward EJ read to transcript
/home/pubseq/BioSw/bwa/bwa-0.5.4/bwa aln polya_out.transPolyA.fa polya_out.ej.1.forwardT.fastq >
polya_out.ej.aln.forward.sai
[bwa_aln] 17bp reads: max_diff = 2
[bwa_aln] 38bp reads: max_diff = 3
[bwa_aln] 64bp reads: max_diff = 4
[bwa_aln] 93bp reads: max_diff = 5
[bwa_aln] 124bp reads: max_diff = 6
[bwa_aln] 157bp reads: max_diff = 7
[bwa_aln] 190bp reads: max_diff = 8
[bwa_aln] 225bp reads: max_diff = 9
[bwa_aln_core] calculate SA coordinate... 0.00 sec
[bwa_aln_core] write to the disk... 0.00 sec
[bwa_aln_core] 6 sequences have been processed.

# Aligning reverse EJ read to transcript
/home/pubseq/BioSw/bwa/bwa-0.5.4/bwa aln polya_out.transPolyA.fa polya_out.ej.2.reverseT.fastq >
polya_out.ej.aln.reverse.sai
[bwa_aln] 17bp reads: max_diff = 2
[bwa_aln] 38bp reads: max_diff = 3
```

```
[bwa_aln] 64bp reads: max_diff = 4
[bwa_aln] 93bp reads: max_diff = 5
[bwa_aln] 124bp reads: max_diff = 6
[bwa_aln] 157bp reads: max_diff = 7
[bwa_aln] 190bp reads: max_diff = 8
[bwa_aln] 225bp reads: max_diff = 9
[bwa_aln_core] calculate SA coordinate... 0.00 sec
[bwa_aln_core] write to the disk... 0.00 sec
[bwa_aln_core] 18 sequences have been processed.

# Extracting forward reads that have only trimmed reads align
/home/pubseq/BioSw/bwa/bwa-0.5.4/bwa samse -n 20 polya_out.transPolyA.fa polya_out.ej.aln.forward.sai
polya_out.ej.1.forwardT.fastq | /home/mokada/wtss/bin/samse2csv.pl -fromend -offset
polya_out.transPolyA.fa.taillengths.txt -endread -negalign | /home/mokada/wtss/bin/procendtagtsv.pl -s
polya_out.ej.aln.stats.forward.txt -a > polya_out.ej.aln.forward.tsv

# Extracting reverse reads that have only trimmed reads align
/home/pubseq/BioSw/bwa/bwa-0.5.4/bwa samse -n 20 polya_out.transPolyA.fa polya_out.ej.aln.reverse.sai
polya_out.ej.2.reverseT.fastq | /home/mokada/wtss/bin/samse2csv.pl -fromend -offset
polya_out.transPolyA.fa.taillengths.txt -endread -negalign | /home/mokada/wtss/bin/procendtagtsv.pl -s
polya_out.ej.aln.stats.reverse.txt -a > polya_out.ej.aln.reverse.tsv

################################
# Created EJ alignment table output:
# Forward EJ alignments: polya_out.ej.aln.forward.tsv
# Reverse EJ alignments: polya_out.ej.aln.reverse.tsv
#

# Aligning forward reads which are mates of reverse EJ reads
/home/pubseq/BioSw/bwa/bwa-0.5.4/bwa aln polya_out.transPolyA.fa polya_out.ej.1.mate.fastq >
polya_out.ej.aln.mate.forward.sai
[bwa_aln] 17bp reads: max_diff = 2
[bwa_aln] 38bp reads: max_diff = 3
[bwa_aln] 64bp reads: max_diff = 4
[bwa_aln] 93bp reads: max_diff = 5
[bwa_aln] 124bp reads: max_diff = 6
[bwa_aln] 157bp reads: max_diff = 7
[bwa_aln] 190bp reads: max_diff = 8
[bwa_aln] 225bp reads: max_diff = 9
[bwa_aln_core] calculate SA coordinate... 0.00 sec
[bwa_aln_core] write to the disk... 0.00 sec
[bwa_aln_core] 9 sequences have been processed.

# Aligning reverse reads which are mates of forward EJ reads
/home/pubseq/BioSw/bwa/bwa-0.5.4/bwa aln polya_out.transPolyA.fa polya_out.ej.2.mate.fastq >
polya_out.ej.aln.mate.reverse.sai
[bwa_aln] 17bp reads: max_diff = 2
[bwa_aln] 38bp reads: max_diff = 3
[bwa_aln] 64bp reads: max_diff = 4
[bwa_aln] 93bp reads: max_diff = 5
[bwa_aln] 124bp reads: max_diff = 6
[bwa_aln] 157bp reads: max_diff = 7
[bwa_aln] 190bp reads: max_diff = 8
[bwa_aln] 225bp reads: max_diff = 9
[bwa_aln_core] calculate SA coordinate... 0.00 sec
[bwa_aln_core] write to the disk... 0.00 sec
[bwa_aln_core] 3 sequences have been processed.

# Saving forward mate reads to TSV file
/home/pubseq/BioSw/bwa/bwa-0.5.4/bwa samse -n 20 polya_out.transPolyA.fa polya_out.ej.aln.mate.forward.sai
polya_out.ej.1.mate.fastq | /home/mokada/wtss/bin/samse2csv.pl -fromend -offset
polya_out.transPolyA.fa.taillengths.txt > polya_out.ej.aln.mate.forward.tsv

# Saving reverse mate reads to TSV file
/home/pubseq/BioSw/bwa/bwa-0.5.4/bwa samse -n 20 polya_out.transPolyA.fa polya_out.ej.aln.mate.reverse.sai
polya_out.ej.2.mate.fastq | /home/mokada/wtss/bin/samse2csv.pl -fromend -offset
polya_out.transPolyA.fa.taillengths.txt > polya_out.ej.aln.mate.reverse.tsv

# Find forward matching mate reads and save to TSV
/home/mokada/wtss/bin/findMatePairedMappings.pl -r polya_out.ej.aln.reverse.tsv -m
polya_out.ej.aln.mate.forward.tsv -a -dmin 100 -dmax 250 > polya_out.ej.aln.matematched.forward.tsv

# Find reverse matching mate reads and save to TSV
/home/mokada/wtss/bin/findMatePairedMappings.pl -r polya_out.ej.aln.forward.tsv -m
polya_out.ej.aln.mate.reverse.tsv -a -dmin 100 -dmax 250 > polya_out.ej.aln.matematched.reverse.tsv
```

```
################################
# Filtered EJ alignments which have corresponding mate reads:
# Forward EJ alignments: polya_out.ej.aln.matematched.forward.tsv
# Reverse EJ alignments: polya_out.ej.aln.matematched.reverse.tsv
#
```

```
[mokada@xhost09 data_ej]$ ls -l
total 108
-rw-r--r--  1 mokada users  236 Oct  5 16:38 cmd.txt
-rw-r--r--  1 mokada users 2144 Oct  5 15:48 ej_forward.txt
-rw-r--r--  1 mokada users 1970 Oct  5 15:49 ej_reverse.txt
-rw-r--r--  1 mokada users 1909 Oct  5 15:36 genes.fa
-rw-r--r--  1 mokada users  227 Oct  5 02:07 polyafinder.conf
-rw-r--r--  1 mokada users  774 Oct  5 16:29 polya_out.ej.1.forwardT.fastq
-rw-r--r--  1 mokada users 1219 Oct  5 16:29 polya_out.ej.1.mate.fastq
-rw-r--r--  1 mokada users  358 Oct  5 16:29 polya_out.ej.2.mate.fastq
-rw-r--r--  1 mokada users 2324 Oct  5 16:29 polya_out.ej.2.reverseT.fastq
-rw-r--r--  1 mokada users  136 Oct  5 16:29 polya_out.ej.aln.forward.sai
-rw-r--r--  1 mokada users  126 Oct  5 16:29 polya_out.ej.aln.forward.tsv
-rw-r--r--  1 mokada users  244 Oct  5 16:29 polya_out.ej.aln.mate.forward.sai
-rw-r--r--  1 mokada users  592 Oct  5 16:29 polya_out.ej.aln.mate.forward.tsv
-rw-r--r--  1 mokada users  502 Oct  5 16:29 polya_out.ej.aln.matematched.forward.tsv
-rw-r--r--  1 mokada users  126 Oct  5 16:29 polya_out.ej.aln.matematched.reverse.tsv
-rw-r--r--  1 mokada users  124 Oct  5 16:29 polya_out.ej.aln.mate.reverse.sai
-rw-r--r--  1 mokada users  197 Oct  5 16:29 polya_out.ej.aln.mate.reverse.tsv
-rw-r--r--  1 mokada users  280 Oct  5 16:29 polya_out.ej.aln.reverse.sai
-rw-r--r--  1 mokada users  502 Oct  5 16:29 polya_out.ej.aln.reverse.tsv
-rw-r--r--  1 mokada users   27 Oct  5 16:29 polya_out.ej.aln.stats.forward.txt
-rw-r--r--  1 mokada users   27 Oct  5 16:29 polya_out.ej.aln.stats.reverse.txt
-rw-r--r--  1 mokada users  365 Oct  5 16:29 polya_out.ej.readnames.tsv
-rw-r--r--  1 mokada users 1958 Oct  5 16:29 polya_out.transPolyA.fa
-rw-r--r--  1 mokada users    9 Oct  5 16:29 polya_out.transPolyA.fa.amb
-rw-r--r--  1 mokada users   45 Oct  5 16:29 polya_out.transPolyA.fa.ann
-rw-r--r--  1 mokada users  756 Oct  5 16:29 polya_out.transPolyA.fa.bwt
-rw-r--r--  1 mokada users  479 Oct  5 16:29 polya_out.transPolyA.fa.pac
-rw-r--r--  1 mokada users  756 Oct  5 16:29 polya_out.transPolyA.fa.rbwt
-rw-r--r--  1 mokada users  479 Oct  5 16:29 polya_out.transPolyA.fa.rpac
-rw-r--r--  1 mokada users  264 Oct  5 16:29 polya_out.transPolyA.fa.rsa
-rw-r--r--  1 mokada users  264 Oct  5 16:29 polya_out.transPolyA.fa.sa
-rw-r--r--  1 mokada users   28 Oct  5 16:29 polya_out.transPolyA.fa.taillengths.txt
```

3. Optional steps – ranking transcripts and visualization

3.1 PAM read visualization

Combine the .tsv files for the next steps (note that there is no distinction between forward and reverse results)

```
[mokada@xhost09 data_pam]$ cat polya_out.pam.aln.forward.tsv polya_out.pam.aln.reverse.tsv >
polya_out.pam.aln.all.tsv
[mokada@xhost09 data_pam]$ cat polya_out.pam.contig.remapped.forward.tsv
polya_out.pam.contig.remapped.reverse.tsv > polya_out.pam.contig.remapped.all.tsv
```

Extract read mapping information from the raw reads to .bed and .wig formats (compatible with UCSC genome browser).

```
[mokada@xhost09 data_pam]$ cat pam_forward.txt pam_reverse.txt | getremappedBED.pl
polya_out.unmapped.reads polya_out.pam.1.forwardT.fastq polya_out.pam.2.reverseT.fastq >
polya_out.map.bed
```

** Optionally, if the raw Illumina reads are not available use BWA to align the PAM FASTQ reads (combine polya_out.pam.1.forwardT.fastq and polya_out.pam.2.reverseT.fastq) to the genome, and use the script

```
bwa aln <genome.fa> <input.fastq> > <pam.sai>
```

bwa samse -n 20 <genome.fa> <pam.sai> <input.sai> | samse2bed.pl > <pam.bed>

Convert the .bed files to .wig for easier viewing.

```
[mokada@xhost09 data_pam]$ cat polya_out.map.bed | bed2Wig.pl > polya_out.map.wig
```

To rank the transcripts and to create UCSC linked URLs:

```
getpolyTmapcoord_extracols.pl -t <IN_TSV> [-b <OUT_BED>] [-z <ZOOM>] [-c CUTOFF1,C2,C3,...,Cn] [-u 'http://genome.ucsc.edu/cgi-bin/hgTracks?org=<ORGANISM>&db=<DB>&position='] [-o] [-d 500] > <RANK_OUT>
```

```
[mokada@xhost09 data_pam]$ ls -l
total 184
-rw-r--r-- 1 mokada users 1292 Oct  5 12:48 cmd.txt
-rw-r--r-- 1 mokada users 2261 Oct  5 01:04 contigs.fa
-rw-r--r-- 1 mokada users 7827 Oct  5 01:40 genes.fa
-rw-r--r-- 1 mokada users 3996 Oct  5 01:07 pam_forward.txt
-rw-r--r-- 1 mokada users 3960 Oct  5 01:07 pam_reverse.txt
-rw-r--r-- 1 mokada users  227 Oct  5 02:07 polyafinder.conf
-rw-r--r-- 1 mokada users 2515 Oct  5 12:20 polya_out.contigPolyAT.fa
-rw-r--r-- 1 mokada users    9 Oct  5 12:20 polya_out.contigPolyAT.fa.amb
-rw-r--r-- 1 mokada users   99 Oct  5 12:20 polya_out.contigPolyAT.fa.ann
-rw-r--r-- 1 mokada users  944 Oct  5 12:20 polya_out.contigPolyAT.fa.bwt
-rw-r--r-- 1 mokada users  606 Oct  5 12:20 polya_out.contigPolyAT.fa.pac
-rw-r--r-- 1 mokada users  944 Oct  5 12:20 polya_out.contigPolyAT.fa.rbwt
-rw-r--r-- 1 mokada users  606 Oct  5 12:20 polya_out.contigPolyAT.fa.rpac
-rw-r--r-- 1 mokada users  328 Oct  5 12:20 polya_out.contigPolyAT.fa.rsa
-rw-r--r-- 1 mokada users  328 Oct  5 12:20 polya_out.contigPolyAT.fa.sa
-rw-r--r-- 1 mokada users 1023 Oct  5 12:21 polya_out.map.bed
-rw-r--r-- 1 mokada users 1473 Oct  5 12:21 polya_out.map.wig
-rw-r--r-- 1 mokada users 1619 Oct  5 12:20 polya_out.pam.1.forwardT.fastq
-rw-r--r-- 1 mokada users 1755 Oct  5 12:20 polya_out.pam.2.reverseT.fastq
-rw-r--r-- 1 mokada users 1202 Oct  5 12:21 polya_out.pam.aln.all.tsv
-rw-r--r-- 1 mokada users  208 Oct  5 12:20 polya_out.pam.aln.forward.sai
-rw-r--r-- 1 mokada users  403 Oct  5 12:20 polya_out.pam.aln.forward.tsv
-rw-r--r-- 1 mokada users  200 Oct  5 12:20 polya_out.pam.aln.forward.unmapped
-rw-r--r-- 1 mokada users  292 Oct  5 12:20 polya_out.pam.aln.reverse.sai
-rw-r--r-- 1 mokada users  799 Oct  5 12:20 polya_out.pam.aln.reverse.tsv
-rw-r--r-- 1 mokada users   67 Oct  5 12:20 polya_out.pam.aln.reverse.unmapped
-rw-r--r-- 1 mokada users  208 Oct  5 12:20 polya_out.pam.contig.aln.forward.sai
-rw-r--r-- 1 mokada users  789 Oct  5 12:20 polya_out.pam.contig.aln.forward.tsv
-rw-r--r-- 1 mokada users  195 Oct  5 12:20 polya_out.pam.contig.aln.forward.unmapped
-rw-r--r-- 1 mokada users  148 Oct  5 12:20 polya_out.pam.contig.aln.reverse.sai
-rw-r--r-- 1 mokada users  264 Oct  5 12:20 polya_out.pam.contig.aln.reverse.tsv
-rw-r--r-- 1 mokada users  362 Oct  5 12:20 polya_out.pam.contig.aln.reverse.unmapped
-rw-r--r-- 1 mokada users 1053 Oct  5 12:21 polya_out.pam.contig.remapped.all.tsv
-rw-r--r-- 1 mokada users  789 Oct  5 12:20 polya_out.pam.contig.remapped.forward.tsv
-rw-r--r-- 1 mokada users  264 Oct  5 12:20 polya_out.pam.contig.remapped.reverse.tsv
-rw-r--r-- 1 mokada users 8025 Oct  5 12:20 polya_out.transPolyA.fa
-rw-r--r-- 1 mokada users    9 Oct  5 12:20 polya_out.transPolyA.fa.amb
-rw-r--r-- 1 mokada users  189 Oct  5 12:20 polya_out.transPolyA.fa.ann
-rw-r--r-- 1 mokada users 2964 Oct  5 12:20 polya_out.transPolyA.fa.bwt
-rw-r--r-- 1 mokada users 1953 Oct  5 12:20 polya_out.transPolyA.fa.pac
-rw-r--r-- 1 mokada users 2964 Oct  5 12:20 polya_out.transPolyA.fa.rbwt
-rw-r--r-- 1 mokada users 1953 Oct  5 12:20 polya_out.transPolyA.fa.rpac
-rw-r--r-- 1 mokada users 1000 Oct  5 12:20 polya_out.transPolyA.fa.rsa
-rw-r--r-- 1 mokada users 1000 Oct  5 12:20 polya_out.transPolyA.fa.sa
-rw-r--r-- 1 mokada users 4630 Oct  5 12:21 polya_out.unmapped.reads
```

Optional – extract genomic poly-A or poly-T windows to find possible false positives. Output is .wig (compatible with UCSC genome browser)

```
[mokada@xhost09 data_pam]$ cat genome.fa | findgenomicpolya.pl -t > polya_out.genomicAT.wig
```

Notes

-genome sequence is not included
-this is very time consuming, an option is to split the fasta file into separate chromosomes and run in parallel

3.2 EJ read visualization

There are two ways of generating images for the EJ reads, 1) aligning the EJ reads to the genome directly, or 2) extract the EJ read position from the Illumina raw reads (if available). The former is more reliable and accurate but will take more time, especially to index the genome.

1) Method 1: Align EJ reads to the genome
In this case, assume all_chr_mouse.fasta is the genome of the organism (mm9 mouse in this case, not included)

Index genome (all_chr_mouse.fasta) and align

```
[mokada@xhost09 data_ej]$ /home/pubseq/BioSw/bwa/bwa-0.5.4/bwa index -a bwtsw all_chr_mouse.fasta
> log_index.txt 2> log_index2.txt
[mokada@xhost09 data_ej]$ cat polya_out.ej.aln.matematched.both.tsv | awk '{print $1}' >
polya_out.ej.aln.matematched.readnames
[mokada@xhost09 data_ej]$ cat polya_out.ej.1.forwardT.fastq polya_out.ej.2.reverseT.fastq |
fastqsubset.pl -i polya_out.ej.aln.matematched.readnames -t > polya_out.ej.both.fastq
[mokada@xhost09 data_ej]$ /home/pubseq/BioSw/bwa/bwa-0.5.4/bwa aln all_chr_mouse.fasta
polya_out.ej.both.fastq > polya_out.ej.both.sai
```

Extract the alignments that have mate read support by supplying .tsv file. A .bed file is created

```
[mokada@xhost09 data_ej]$ cat polya_out.ej.aln.matematched.forward.tsv
polya_out.ej.aln.matematched.reverse.tsv > polya_out.ej.aln.matematched.both.tsv
[mokada@xhost09 data_ej]$ /home/pubseq/BioSw/bwa/bwa-0.5.4/bwa samse -n 20 all_chr_mouse.fasta
polya_out.ej.both.sai polya_out.ej.both.fastq | samse2bed.pl -t
polya_out.ej.aln.matematched.both.tsv -r > polya_out.ej.map.bed
```

Convert .bed to .wig format (optional)

```
[mokada@xhost09 data_ej]$ cat polya_out.ej.map.bed | bed2Wig.pl > polya_out.ej.map.wig
```

The .wig and .bed files can be uploaded as customisable tracks in the UCSC browser

Rank transcripts by number and quality of aligned EJ reads, provide a genome browser URL

```
[mokada@xhost09 data_ej]$ getpolyTmapcoord_extracols.pl -t polya_out.ej.aln.matematched.both.tsv
-b polya_out.ej.map.bed -z 1500 -c -20,-15,-10,-5,-1 -e -5 -o -d 50 -u
'http://genome.ucsc.edu/cgi-bin/hgTracks?org=mouse&db=mm9&position=' >
polya_out.ej.map.transcripts.txt
-1      -5      -10     -15     -20     transcript      url
0       3       8       10      10      BC054387        http://genome.ucsc.edu/cgi-
bin/hgTracks?org=mouse&db=mm9&position=chr8:127079948-127081447
```

2) Method 2: Extract EJ coordinates from the raw Illumina reads

Extract the alignments that have mate read support by supplying .tsv file. A .bed file is created

```
[mokada@xhost09 data_ej]$ cat ej_forward.txt ej_reverse.txt | getremappedBED.pl
polya_out.unmapped.reads polya_out.ej.1.forwardT.fastq polya_out.ej.2.reverseT.fastq >
polya_out.map2.bed
```

Convert .bed to .wig format (optional)

```
[mokada@xhost09 data_ej]$ cat polya_out.map2.bed | bed2Wig.pl > polya_out.map2.wig
```

```
[mokada@xhost09 data_ej]$ cat polya_out.ej.aln.matematched.forward.tsv
polya_out.ej.aln.matematched.reverse.tsv > polya_out.ej.aln.matematched.both.tsv
```

Rank transcripts by number and quality of aligned EJ reads, provide a genome browser URL

```
[mokada@xhost09 data_ej]$ getpolyTmapcoord_extracols.pl -t polya_out.ej.aln.matematched.both.tsv
-b polya_out.map2.bed -z 1500 -c -20,-15,-10,-5,-1 -e -5 -o -d 50 -u 'http://genome.ucsc.edu/cgi-
bin/hgTracks?org=mouse&db=mm9&position=' > polya_out.ej.map2.transcripts.txt
```

-1	-5	-10	-15	-20	transcript	url
10	10	4	0	0	BC054387	http://genome.ucsc.edu/cgi-bin/hgTracks?org=mouse&db=mm9&position=chr8:127080117-127081616