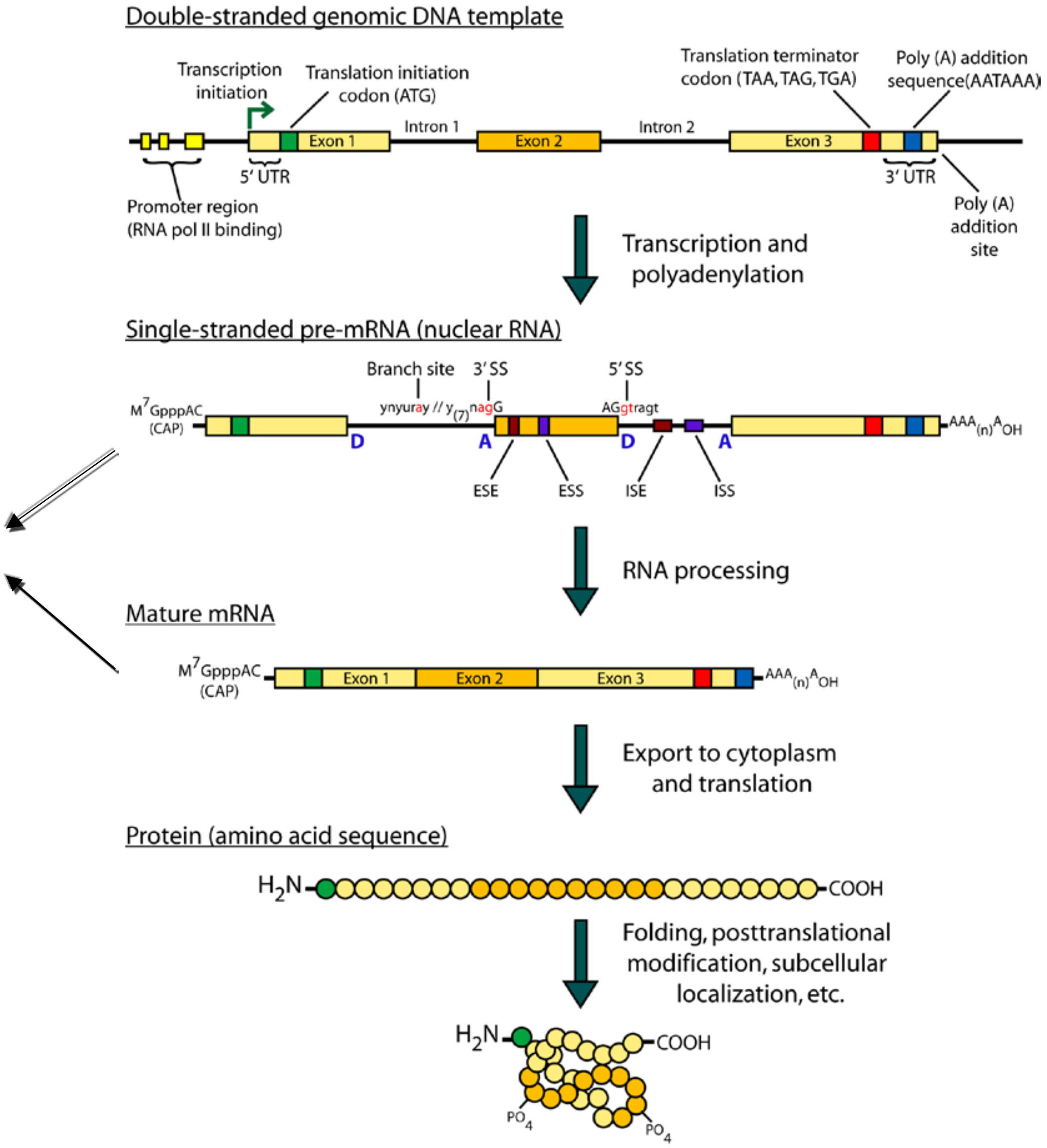
Gene expression

RNA sequencing


Double-stranded genomic DNA template
Transcription initiation
Translation initiation codon (ATG)
Translation terminator codon (TAA, TAG, TGA)
Poly (A) addition sequence(AATAAA)
Exon 1
Intron 1
Exon 2
Intron 2
Exon 3
5′ UTR
3′ UTR
Promoter region (RNA pol II binding)
Poly (A) addition site
Transcription and polyadenylation
Single-stranded pre-mRNA (nuclear RNA)
Branch site
3′ SS
5′ SS
ynyuray // y(7) nagG
AGgtragt
M7GpppAC (CAP)
D
A
D
A
AAA(n)AOH
ESE
ESS
ISE
ISS
RNA processing
Mature mRNA
M7GpppAC (CAP)
Exon 1
Exon 2
Exon 3
AAA(n)AOH
Export to cytoplasm and translation
Protein (amino acid sequence)
H2N
COOH
Folding, posttranslational modification, subcellular localization, etc.
H2N
COOH
PO4
PO4

What you want from RNA-Seq analysis

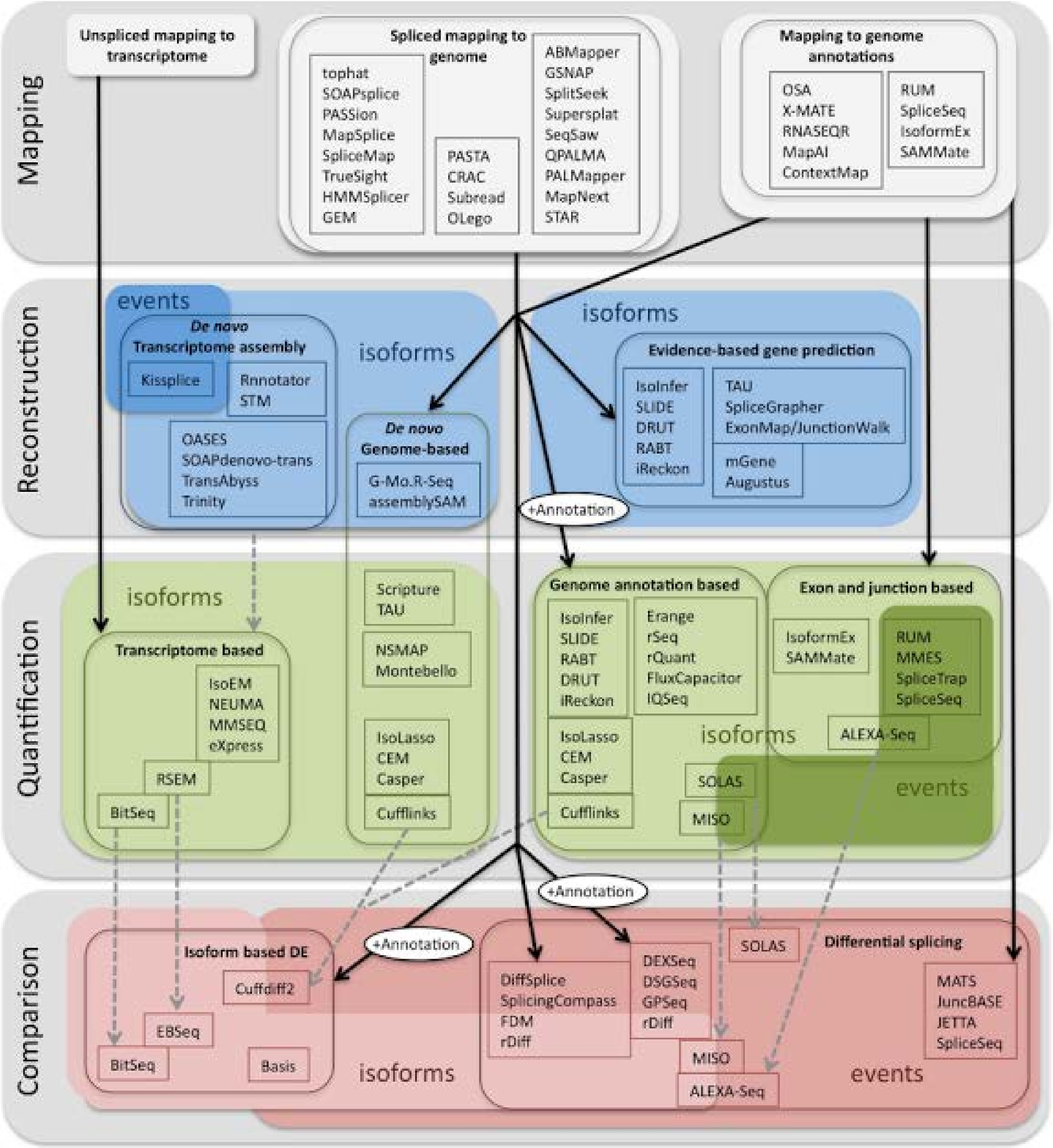
- Gene expression and differential expression
- Alternative expression analysis
- Transcript discovery and annotation
- Mutation discovery
- Fusion detection
- RNA editing
- Allele specific expression

All you have is fastq

- fastq

```
@LAMARCK:3118:C067BACXX:2:1101:2616:22262
GTACACCCCAGAGGCCAGCATTGACTCCACAAAATGATATTGCTACTAGTC
+
<@<DDDADFHH=BBFHIGB?F?GHBCC@3DGGHIG@<DHIGIG<FGGCAFB
@LAMARCK:3118:C067BACXX:2:1101:19818:19913
GTTCCACCTCCAGAAGACCTCATCCAAGCACCTCCTGCAGTACCCAGTACA
+
@CCDDFFFHBHFFCFH>FHIJGIJIIEC1CGAEIIJH@FEGGHIJIDBFGG
@LAMARCK:3118:C067BACXX:2:1101:2616:22262
GTACACCCCAGAGGCCAGCATTGACTCCACAAAATGATATTGCTACTAGTC
+
<@<DDDADFHH=BBFHIGB?F?GHBCC@3DGGHIG@<DHIGIG<FGGCAFB
@LAMARCK:3118:C067BACXX:2:1101:6996:29437
CGTCTTCCAAGTATAGATACATGTCCCCTTCAGTCTTCAGCCTCCTTGGAC
+
CC@FFFEFHHGBFGFHICGJIGCGHIIJIIGGHGCFIGGCHEGIIIIIIII
```

What pipeline to use? What settings?



Mapping

Unspliced mapping to transcriptome

Spliced mapping to genome

tophat
SOAPsplice
PASSion
MapSplice
SpliceMap
TrueSight
HMMSplicer
GEM

PASTA
CRAC
Subread
OLego

ABMapper
GSNAP
SplitSeek
Supersplat
SeqSaw
QPALMA
PALMapper
MapNext
STAR

Mapping to genome annotations

OSA
X-MATE
RNASEQR
MapAl
ContextMap

RUM
SpliceSeq
IsoformEx
SAMMate

Reconstruction

events

De novo Transcriptome assembly

Kissplice

Rnnotator
STM

OASES
SOAPdenovo-trans
TransAbyss
Trinity

isoforms

De novo Genome-based

G-Mo.R-Seq
assemblySAM

isoforms

Evidence-based gene prediction

IsoInfer
SLIDE
DRUT
RABT
iReckon

TAU
SpliceGrapher
ExonMap/JunctionWalk

mGene
Augustus

+Annotation

Quantification

isoforms

Transcriptome based

IsoEM
NEUMA
MMSEQ
eXpress

RSEM

BitSeq

Scripture
TAU

NSMAP
Montebello

IsoLasso
CEM
Casper

Cufflinks

Genome annotation based

IsoInfer
SLIDE
RABT
DRUT
iReckon

Erange
rSeq
rQuant
FluxCapacitor
IQSeq

IsoLasso
CEM
Casper

Cufflinks

isoforms

SOLAS

MISO

Exon and junction based

IsoformEx
SAMMate

RUM
MMES
SpliceTrap
SpliceSeq

ALEXA-Seq

events

Comparison

+Annotation

+Annotation

Isoform based DE

Cuffdiff2

EBSeq

BitSeq

Basis

isoforms

DiffSplice
SplicingCompass
FDM
rDiff

DEXSeq
DSGSeq
GPSeq
rDiff

SOLAS

Differential splicing

MATS
JuncBASE
JETTA
SpliceSeq

MISO

ALEXA-Seq

events

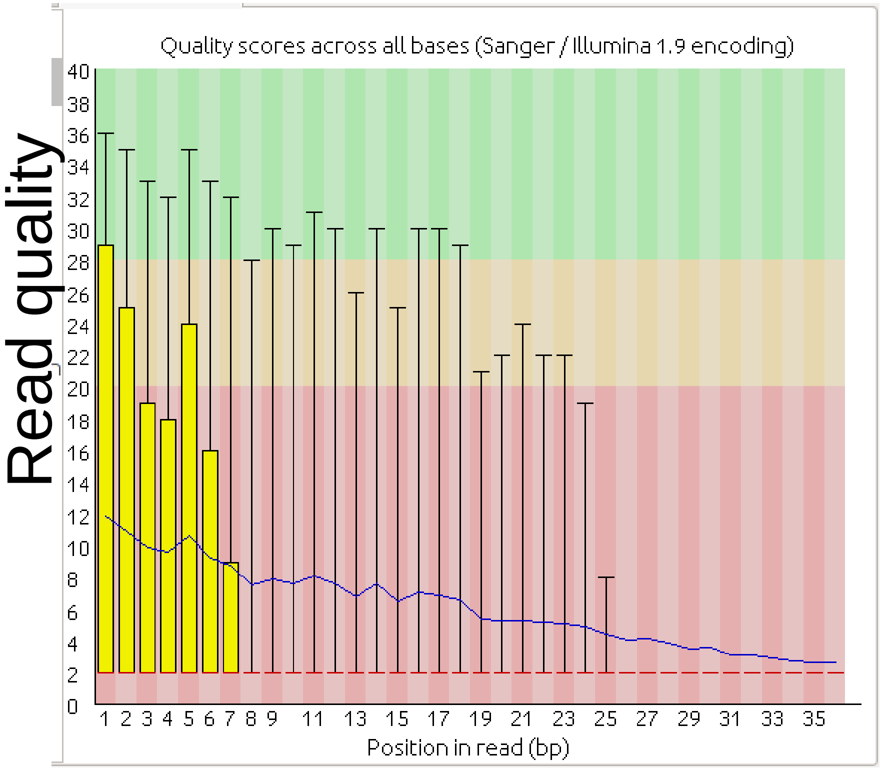
RNA Sequence Analysis

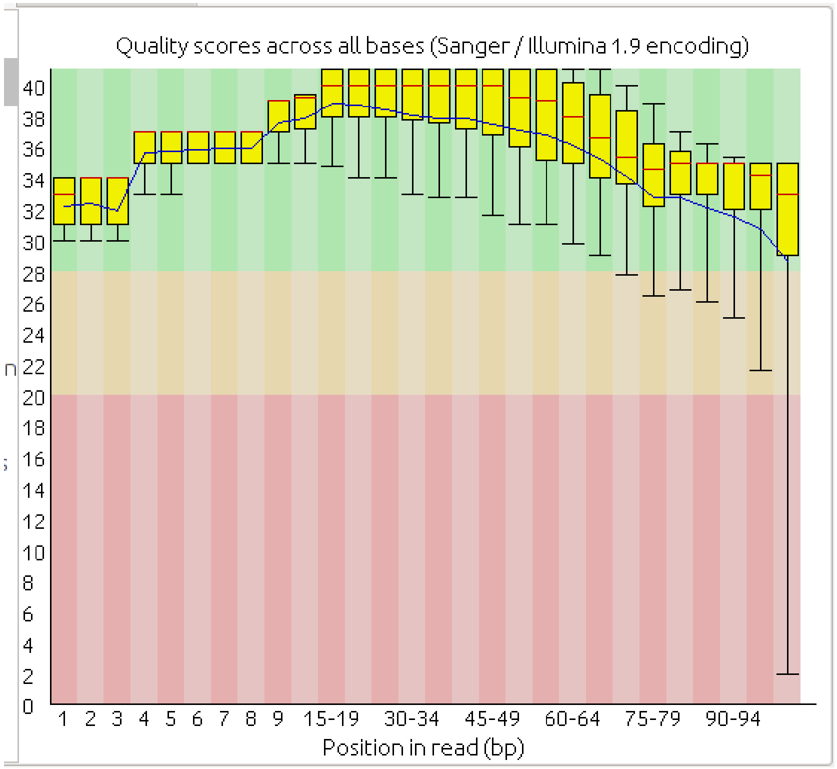
FASTQ

- fastq

```
@LAMARCK:3118:C067BACXX:2:1101:2616:22262
GTACACCCCAGAGGCCAGCATTGACTCCACAAAATGATATTGCTACTAGTC
+
<@<DDDADFHH=BBFHIGB?F?GHBCC@3DGGHIG@<DHIGIG<FGGCAFB
@LAMARCK:3118:C067BACXX:2:1101:19818:19913
GTTCCACCTCCAGAAGACCTCATCCAAGCACCTCCTGCAGTACCCAGTACA
+
@CCDDFFFHBHFFCFH>FHIJGIJIIEC1CGAEIIJH@FEGGHIJIDBFGG
@LAMARCK:3118:C067BACXX:2:1101:2616:22262
GTACACCCCAGAGGCCAGCATTGACTCCACAAAATGATATTGCTACTAGTC
+
<@<DDDADFHH=BBFHIGB?F?GHBCC@3DGGHIG@<DHIGIG<FGGCAFB
@LAMARCK:3118:C067BACXX:2:1101:6996:29437
CGTCTTCCAAGTATAGATACATGTCCCCTTCAGTCTTCAGCCTCCTTGGAC
+
CC@FFFEFHHGBFGFHICGJIGCGHIIJIIGGHGCFIGGCHEGIIIIIIII
```

Is the output any good?





Their data sucks!

Our data

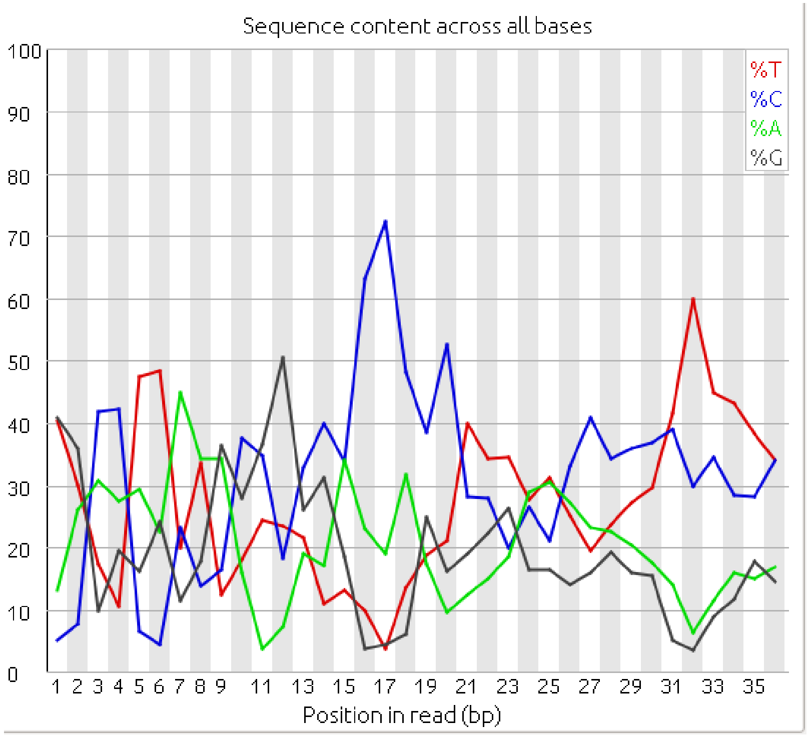
Quality score of 10 means 90% of bases are correct

20 means 99% of bases are correct

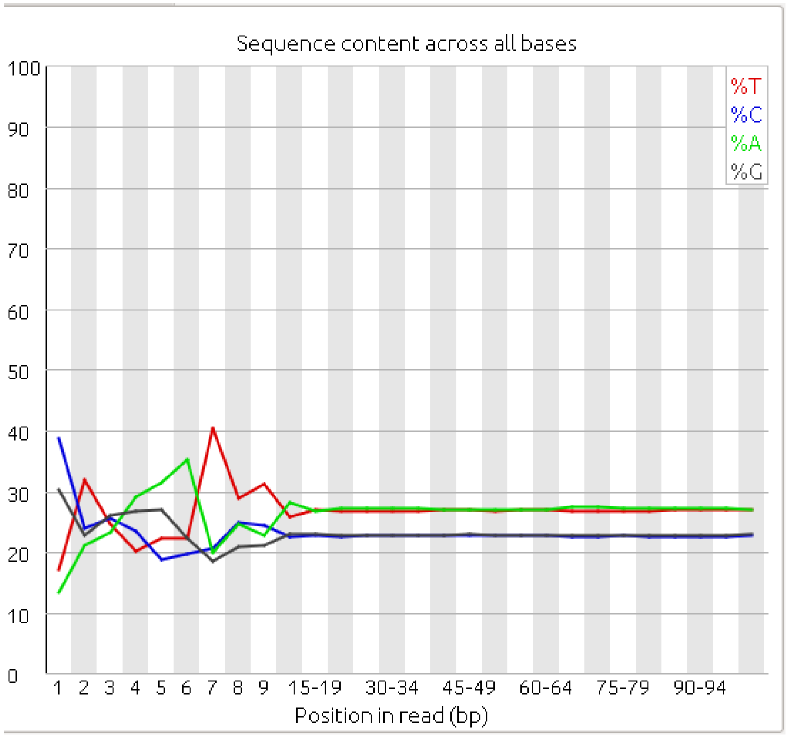
30 means 99.9% of bases are correct, etc.

Genome Res 21: 410-421

Is the output any good?



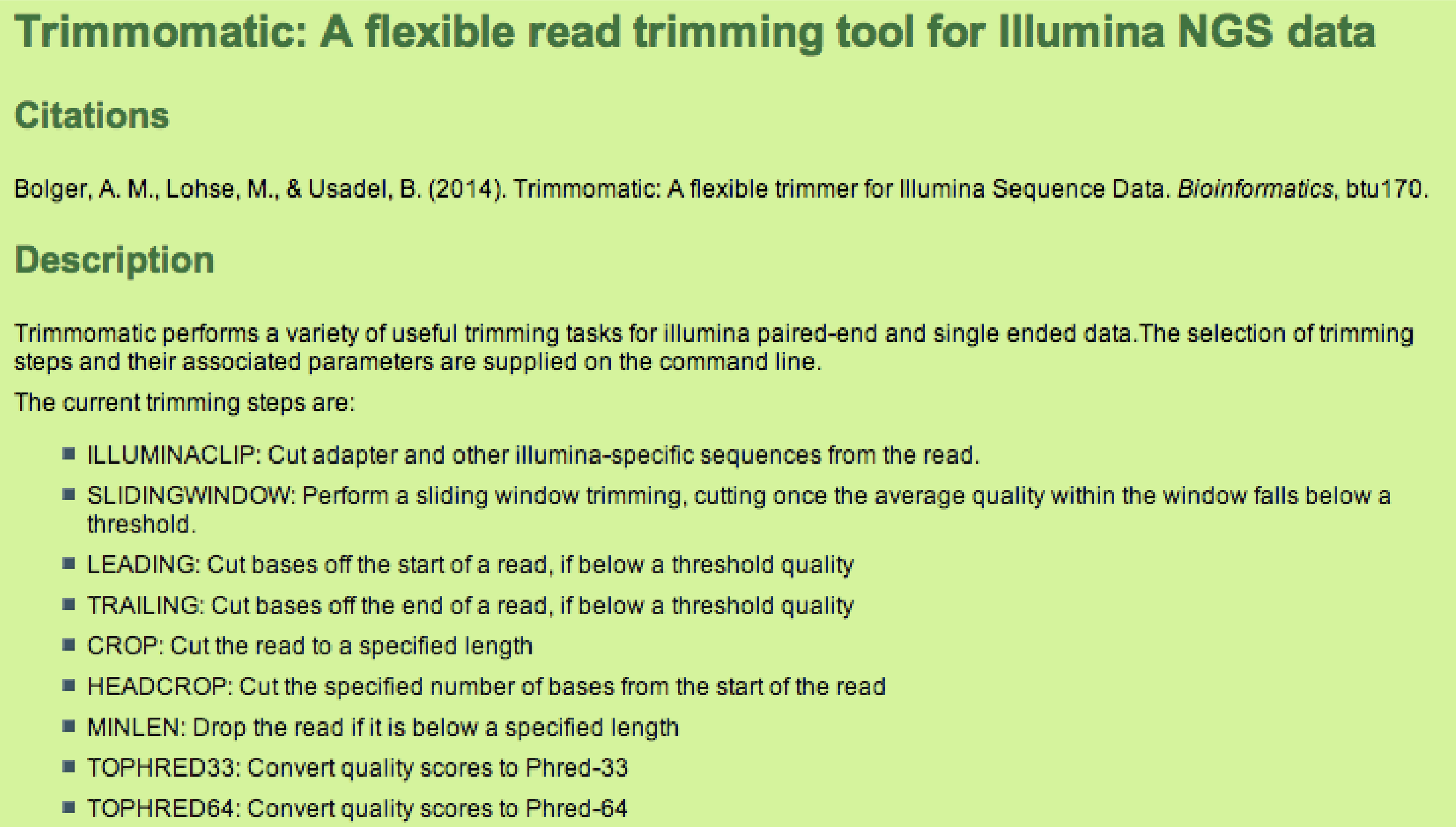
Their data



Our data

Genome Res 21: 410-421

Trimming (optional)



Trimmomatic: A flexible read trimming tool for Illumina NGS data

Citations

Bolger, A. M., Lohse, M., & Usadel, B. (2014). Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, btu170.

Description

Trimmomatic performs a variety of useful trimming tasks for illumina paired-end and single ended data.The selection of trimming steps and their associated parameters are supplied on the command line.

The current trimming steps are:

- ILLUMINACLIP: Cut adapter and other illumina-specific sequences from the read.
- SLIDINGWINDOW: Perform a sliding window trimming, cutting once the average quality within the window falls below a threshold.
- LEADING: Cut bases off the start of a read, if below a threshold quality
- TRAILING: Cut bases off the end of a read, if below a threshold quality
- CROP: Cut the read to a specified length
- HEADCROP: Cut the specified number of bases from the start of the read
- MINLEN: Drop the read if it is below a specified length
- TOPHRED33: Convert quality scores to Phred-33
- TOPHRED64: Convert quality scores to Phred-64

General themes of RNA-seq workflows

- Each type of RNA-seq analysis has distinct requirements and challenges but also a common theme:
  1. Obtain raw data (convert format)
  2. Identify/assemble reference and align reads
  3. Process alignment with a tool specific to the goal
     - e.g. cufflinks, rsem for expression analysis
  4. Post process
     - Import into downstream software (R, Matlab, etc.)
  5. Summarize and visualize
     - Create gene lists, prioritize candidates for validation, etc.

Numerous possible analysis strategies

- Two major branches
  - Direct alignment of reads (spliced or unspliced) to genome or transcriptome
  - Assembly of reads followed by alignment
- Recently Kmer comparison (Sailfish)

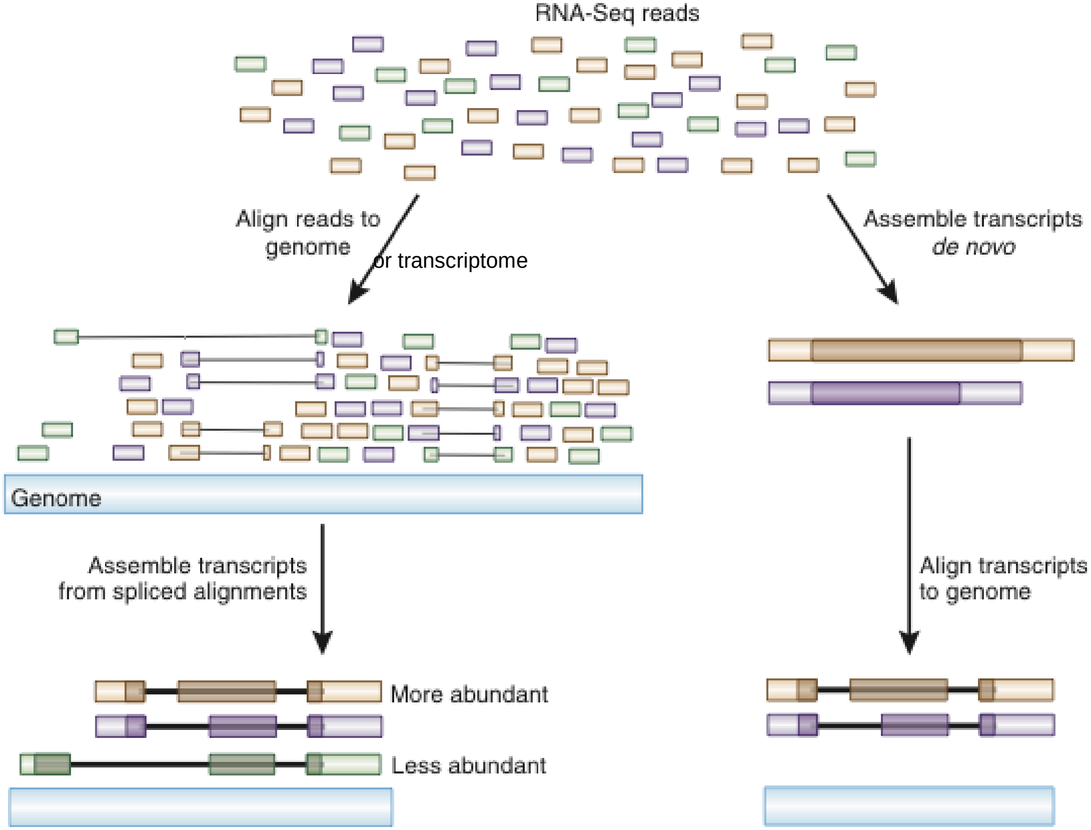

RNA-Seq reads
Align reads to genome
or transcriptome
Assemble transcripts de novo
Genome
Assemble transcripts from spliced alignments
Align transcripts to genome
More abundant
Less abundant

Image from Haas & Zody, 2010

What aligner to use?

- Bowtie 1 or 2
- BWA (aln, mem)
- STAR
- Subread
- Tophat
- Novoalign
- GSNAP
- GEM
- BLAT
- more all the time

They all do well and poorly in different ways, mostly due to tricks employed to make them go faster.

All are free/open source except Novoalign.

*aligners designed to align RNAseq to genomes

Alignment

- Find genomic/transcriptomic source of reads
- Requires a reference
  - Typically compiled from a fasta file
  - Can be genome/transcriptome.
  - Can contain custom sequences
  - How do you get this fasta file?
    - It exists in a database (Xenbase)
    - You assemble it (Trinity, Oases)
- Many next-gen aligners available, two main approaches
  - Store reference in memory and map each read successively (Bowtie2)
    - memory footprint proportional to genome size
  - Store reads in memory and scan across reference (Star)
    - memory footprint proportional to read number

Reference choice

- Alignment to genome
  - Allow reads from unannotated loci, introns *et cetera* to align to their correct locations… potential for new biological insights
- Alignment to transcriptome
  - Computationally cheap (fast!)
  - Spliced (exon junction) reads map correctly
  - Pairing distance and junction reads may help distinguish individual isoforms (informative/unique regions of transcripts)
- Assembly
  - Can provide a more long-range view of transcripts
  - Allows detection of chimeric transcripts and resolution of ‘breakpoints’
  - May be done with or without existing genome

Drawbacks for each strategy

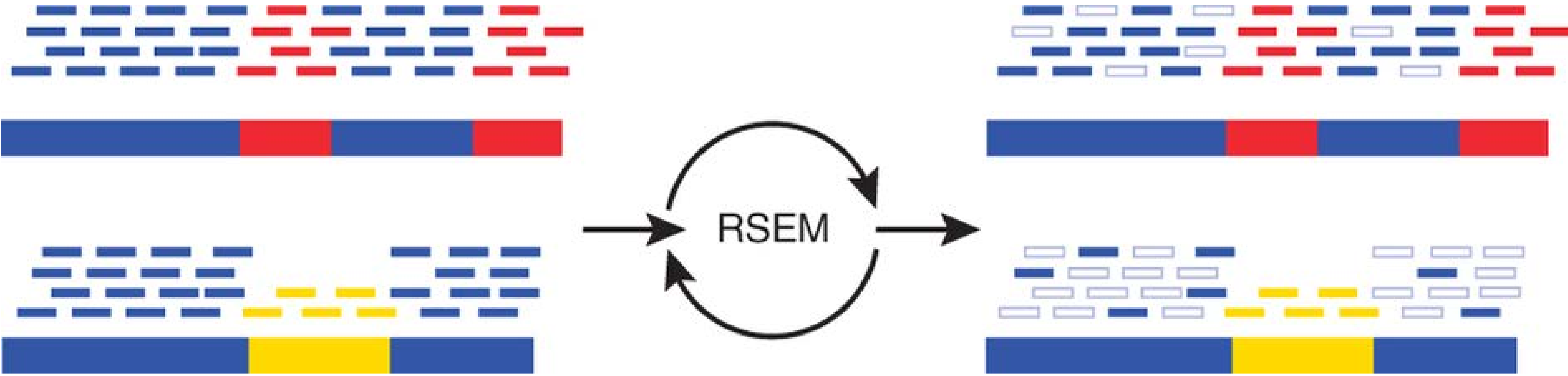
- Alignment to genome
  - Computationally expensive
  - reads spanning splice junctions are lost
- Alignment to transcriptome
  - Reads deriving from non-genic structures may be ‘forcibly’ (and erroneously) aligned to genes
- Assembly
  - Low expression = difficult/impossible to assemble
  - Misassemblies/fragmented contigs due to repeats
  - Requires vast amounts of memory

Alignment options

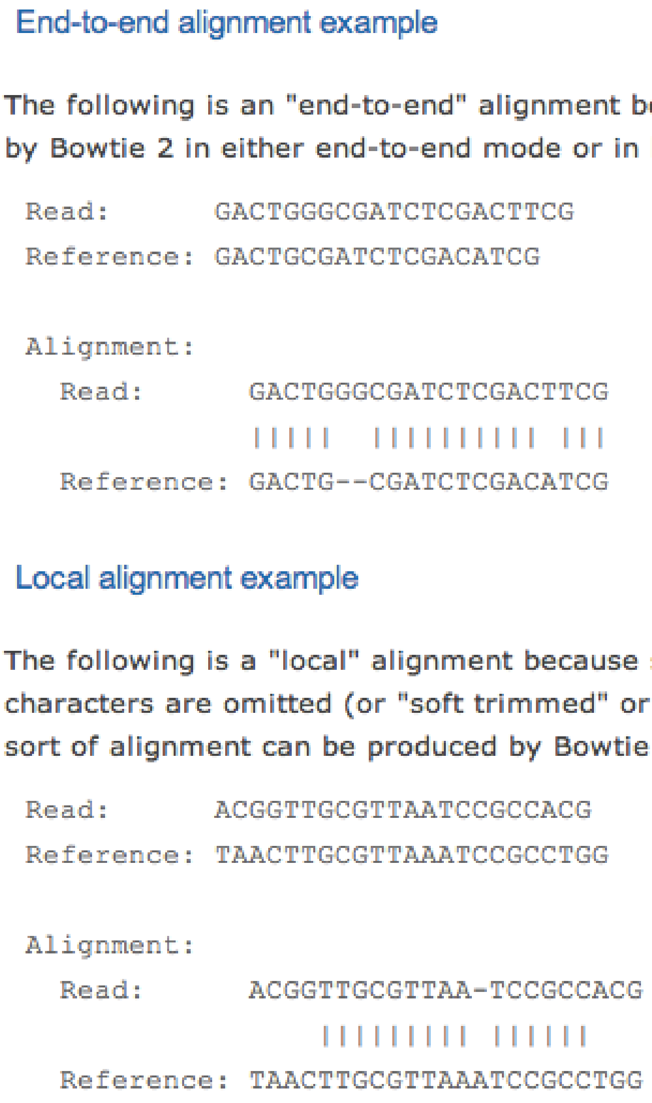
- Multi-map reads (i.e. reads that can be placed in multiple locations with equal or good enough score)
  - have only one “best” location reported
  - Report all alignments (or up to some maximal number)
- Single versus paired-end reads
  - paired-end reads are positioned with more precision, but if the reference is imprecise you will lose reads
- Local versus global alignment
  - One seed versus multiple seeds
- How many mismatches are allowed
  - within a seed or total
- Format/details of alignment output vary
  - standardization towards SAM/BAM format (header)

How to count how many times reads go to a specific transcript or position?

What if a read aligns equally to more than one place?



Nature Protocols 8:149



End-to-end alignment example

The following is an "end-to-end" alignment because it involves all the characters in the read. Such an alignment can be produced by Bowtie 2 in either end-to-end mode or in local mode.

```
Read:      GACTGGGCGATCTCGACTTCG
Reference: GACTGCGATCTCGACATCG

Alignment:
  Read:      GACTGGGCGATCTCGACTTCG
             |||||  |||||||||| |||
  Reference: GACTG--CGATCTCGACATCG
```

Local alignment example

The following is a "local" alignment because some of the characters at the ends of the read do not participate. In this case, 4 characters are omitted (or "soft trimmed" or "soft clipped") from the beginning and 3 characters are omitted from the end. This sort of alignment can be produced by Bowtie 2 only in local mode.

```
Read:      ACGGTTGCGTTAATCCGCCACG
Reference: TAACTTGCGTTAAATCCGCCTGG

Alignment:
  Read:      ACGGTTGCGTTAA-TCCGCCACG
                 ||||||||| ||||||
  Reference: TAACTTGCGTTAAATCCGCCTGG
```

Installation is easy

<skip this, we have done it>

```
>wget http://sourceforge.net/projects/bowtie-bio/files/bowtie2/2.2.2/bowtie2-2.2.2-linux-x86_64.zip
>unzip bowtie2-2.2.2-linux-x86_64.zip
```

<skip this, we have done it>

Go to your directory and type:

```
>alias bowtie2='/opt/xenopus/bowtie2-2.2.2/bowtie2'
>bowtie2
```

bowtie2 [options]* -x <bt2-idx> -q myfastqfile.fq

- The chief differences between Bowtie 1 and Bowtie 2 are:

1. For reads >50 bp Bowtie 2 is generally faster, more sensitive, and uses less memory than Bowtie 1.

2. Bowtie 2 supports gapped alignment with affine gap penalties. Number of gaps and gap lengths are not restricted, except by way of the configurable scoring scheme. Bowtie 1 finds just ungapped alignments.

3. Bowtie 2 supports local alignment, which doesn't require reads to align end-to-end. Local alignments might be "trimmed" ("soft clipped") at one or both extremes in a way that optimizes alignment score. Bowtie 2 also supports end-to-end alignment which, like Bowtie 1, requires that the read align entirely.

4. There is no upper limit on read length in Bowtie 2. Bowtie 1 had an upper limit of around 1000 bp.

5. Bowtie 2 allows alignments to overlap ambiguous characters (e.g. Ns) in the reference. Bowtie 1 does not..

7. Bowtie 2's paired-end alignment is more flexible. E.g. for pairs that do not align in a paired fashion, Bowtie 2 attempts to find unpaired alignments for each mate.

8. Bowtie 2 reports a spectrum of mapping qualities, in contrast fo Bowtie 1 which reports either 0 or high.

Input and output for alignment: fasta/fastq and sam

- fasta

```
@LAMARCK:3118:C067BACXX:2:1101:2616:22262
GTACACCCCAGAGGCCAGCATTGACTCCACAAAATGATATTGCTACTAGTC
```

- fastq

```
@LAMARCK:3118:C067BACXX:2:1101:2616:22262
GTACACCCCAGAGGCCAGCATTGACTCCACAAAATGATATTGCTACTAGTC
+
<@<DDDADFHH=BBFHIGB?F?GHBCC@3DGGHIG@<DHIGIG<FGGCAFB
```

- sam

```
LAMARCK:3118:C067BACXX:2:1101:2616:22262    0
            E2F4|ENSG00000205250|c.XGI_TC419474|JGIv7b.000015416_1879975-1902638-      685         255
            51M         *
0    0      GTACACCCCAGAGGCCAGCATTGACTCCACAAAATGATATTGCTACTAGTC
            <@<DDDADFHH=BBFHIGB?F?GHBCC@3DGGHIG@<DHIGIG<FGGCAFB
NH:i:1      HI:i:1      AS:i:50     nM:i:0
```

Input and output for alignment: fasta/fastq and sam

- Examine samfile
- Convert a samfile back to fastq

```
>cd /opt/xenopus/reads
> awk 'BEGIN{FS="\t"; OFS=FS;}FNR<10{print $1,$10,$11}' overexpression_expt.sam
```

  - and now this

```
> awk 'BEGIN{FS="\t"; OFS=FS;}FNR<10{print "@"$1; print $10; print "+";print $11}'
    overexpression_expt.sam > ~/overexpression_expt.fastq
```

- Run bowtie2
  - find the index path
  - find the fastq file you made

```
bowtie2 -x /opt/xenopus/indexes/LAEVIS_7.1 ~/overexpression_expt.fq -S
    overexpression_expt.my.sam
```

Alignment quality

- Scores: higher = more similar

An alignment score quantifies how similar the read sequence is to the reference sequence aligned to. The higher the score, the more similar they are. A score is calculated by subtracting penalties for each difference (mismatch, gap, etc) and, in local alignment mode, adding bonuses for each match.

The scores can be configured with the --ma (match bonus), --mp (mismatch penalty), --np (penalty for having an N in either the read or the reference), --rdg (affine read gap penalty) and --rfg (affine reference gap penalty) options.

- End-to-end alignment score example

A mismatched base at a high-quality position in the read receives a penalty of -6 by default. A length-2 read gap receives a penalty of -11 by default (-5 for the gap open, -3 for the first extension, -3 for the second extension). Thus, in end-to-end alignment mode, if the read is 50 bp long and it matches the reference exactly except for one mismatch at a high-quality position and one length-2 read gap, then the overall score is -(6 + 11) = -17.

The best possible alignment score in end-to-end mode is 0, which happens when there are no differences between the read and the reference.

Alignment quality

```
> wget http://sourceforge.net/projects/samstat/files/samstat.tgz
> tar –xzvf samstat.tgz
> cd samstat/src/
> make
> alias samstat='/home/virginia/samstat/src/samstat'
> samstat ~/overexpression_expt.sam
```

Downstream processing of aligned output

- samtools
  - compress samfile (.sam) into binary file (.bam
    - much smaller and required for many steps
    - samtools view –bhS mysamfile.sam > mybamfile.bam
  - sort bamfile
    - samtools sort mybamfile.bam mybamfile.sorted
  - index bamfile
    - samtools index mybamfile.sorted.bam
  - find reads that maps to regions
    - samtools view –q 20 mybamfile.sorted.bam <your region>

Common questions: Should I remove duplicates for RNA-seq?

- Maybe… more complicated question than for DNA
- Concern.
  - Duplicates may correspond to biased PCR amplification of particular fragments
  - For highly expressed, short genes, duplicates are expected even if there is no amplification bias
  - Removing them may reduce the dynamic range of expression estimates
- Assess library complexity and decide…
- If you do remove them, assess duplicates at the level of paired-end reads (fragments) not single end reads

```
> samtools rmdup [-sS] <input.srt.bam> <out.bam>
```

Remove potential PCR duplicates: if multiple read pairs have identical external coordinates, only retain the pair with highest mapping quality.

Getting expression data from alignments

- Expression values can be tabulated for individual gene loci, transcripts, exons and splice junctions
- Gene expression values typically reported in RPK
  - Number of reads per kb of exonic bases per million reads in the library

RPKM

(1e+9)*GeneReads/(ExonGeneLen*TotalMappableReads);

  - Compensates for variable library size and over-representation of reads from longer transcripts
- Various software available
  - Cufflinks (*Trapnell et al, 2010. PMID: 20436464*), probably supplants
  - RSEM
  - DE-Seq

Getting expression data from alignments

```
> wget http://cufflinks.cbcb.umd.edu/downloads/cufflinks-2.2.1.Linux_x86_64.tar.gz
> tar -xzvf cufflinks-2.2.1.Linux_x86_64.tar.gz
> alias cufflinks='cufflinks-2.2.1.Linux_x86_64/cufflinks'
> cufflinks mybamfile.bam
```