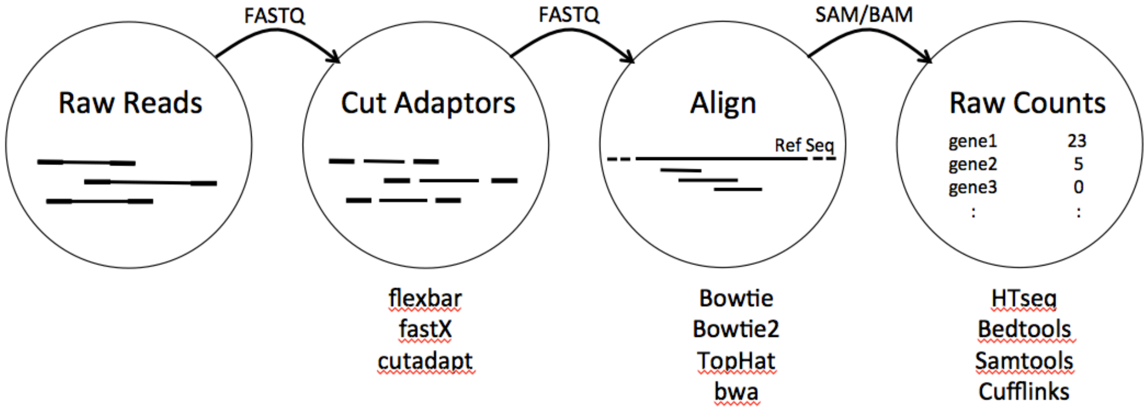
RNA-seq cheat sheet

RNA-seq basic pipeline steps and software:



Calculating the number of reads in a FASTQ file:

```
grep ^@ <file.fastq> | wc —l
```

`grep ^@` will get the lines that start with '@' and `wc —l` will count the number of those lines. These commands together will give you the exact number of reads in `<file.fastq>.`

```
wc -l <file.fastq>
```

You can also run `wc —l` by itself to get the total number of lines in the file `<file.fastq>`. You will need to divide that number by 4 to get the number of reads (Each read in a FASTQ file occupies 4 lines).

Cutting off the adaptor sequences:

```
flexbar —n 1 —t <new_file_name> —r <reads1.fastq>
—p <reads2.fastq> —a adaptors.fasta —f <fastq>
```

`—t <new_file_name>` - this will be used to make output files. For each trimmed reads file the software will add 1 or 2 (depending on the reads) to the output file name, like this `<new_file_name1.fastq>`.
`—r <reads1.fastq> —p <reads2.fastq>` - your raw reads files. If you only have reads 1, input `—r <reads1.fastq>` only.
`—a adaptors.fasta` - the adaptors file. This can also be in `.fna` format.
`—f <fastq>` - format of your reads input files. In this case its FASTQ, but it could also be FASTA.

More info on how to install and run flexbar on TACC or on your own computer:
http://barricklab.org/twiki/bin/view/Lab/ProtocolsFlexbarCommands

Mapping reads:

If you are using TACC, make sure to load the following modules before running the software:

```
module load bowtie
module load tophat
```

Tophat:

```
tophat —p 4 —G <gtf_ref_seq_file> -o <output.sam> --no-novel-juncs <reference_index>
<reads1.fastq> <reads2.fastq>
```

`—p 4` - thread number
`—G <gtf_ref_seq_file>` - reference sequence. Tophat will take either GFF or GTF file formats.
`-o <output.sam>` - name of the output file
`--no-novel-juncs` - this will suppress the search for novel splice junctions (or potential introns).
`<reference_index>` - path to the index files made with bowtie2:

```
bowtie2-build <ref_seq.fa> <reference_index>
```

`<ref_seq.fa>` - reference sequence in FASTA format. This step only accepts the FASTA format.
`<reference_index>` - this will be the name of the output directory (and input for tophat)

Bowtie2:

```
bowtie2 —x <reference_index> -1 <reads1.fastq> -2 <reads2.fastq> -S <output.sam>
```

if you only have reads 1, use `-U <reads1.fastq>` instead of `-1 <reads1.fastq> -2 <reads2.fastq>`. Input variable names are the same as in the example above.

Converting between BAM and SAM files:

If you are using TACC, don't forget to load the module:

```
module load samtools
```

```
samtools view -h <output.bam> > <output.sam>
samtools view -b -S <output.sam> > <output.bam>
```

Sorting BAM files:

```
samtools sort <output.bam> <output_name>
```

Quantifying gene counts:

If you are using TACC, don't forget to load the module:

```
module load htseq
```

```
htseq-count —m <mode> -t <feature_type> -i <id_attribute> <output.sam> <gtf_ref_seq_file>
```

`—m <mode>` - specifies how to count reads. The options for `<mode>` are `union, intersection_strict, intersection_nonempty.`
`-t <feature_type>` - generally this should be `exon` or the 3rd column in your GTF file.
`-i <id_attribute>` - gene name/gene id/etc. to use in the output file.

Normalizing gene count:

DESeq package available here:
http://bioconductor.org/packages/release/bioc/html/DESeq2.html

DESeq2 tutorial:
http://www.bioconductor.org/packages/release/bioc/vignettes/DESeq2/inst/doc/DESeq2.pdf

Other RNA-seq tutorials:

Introduction to Next Gen Sequencing:
https://wikis.utexas.edu/display/bioiteam/SSC+Intro+to+NGS+Bioinformatics+Course

Additional reading:

1. Wagner, Günter P., Koryu Kin, and Vincent J. Lynch. "Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples." *Theory in Biosciences* 131.4 (2012): 281-285.
2. Trapnell, Cole, Lior Pachter, and Steven L. Salzberg. "TopHat: discovering splice junctions with RNA-Seq." *Bioinformatics* 25.9 (2009): 1105-1111.
3. Langmead, Ben, and Steven L. Salzberg. "Fast gapped-read alignment with Bowtie 2." *Nature methods* 9.4 (2012): 357-359.
4. Wall, Jeffrey D., et al. "Estimating genotype error rates from high-coverage next-generation sequence data." *Genome research* (2014): gr-168393.