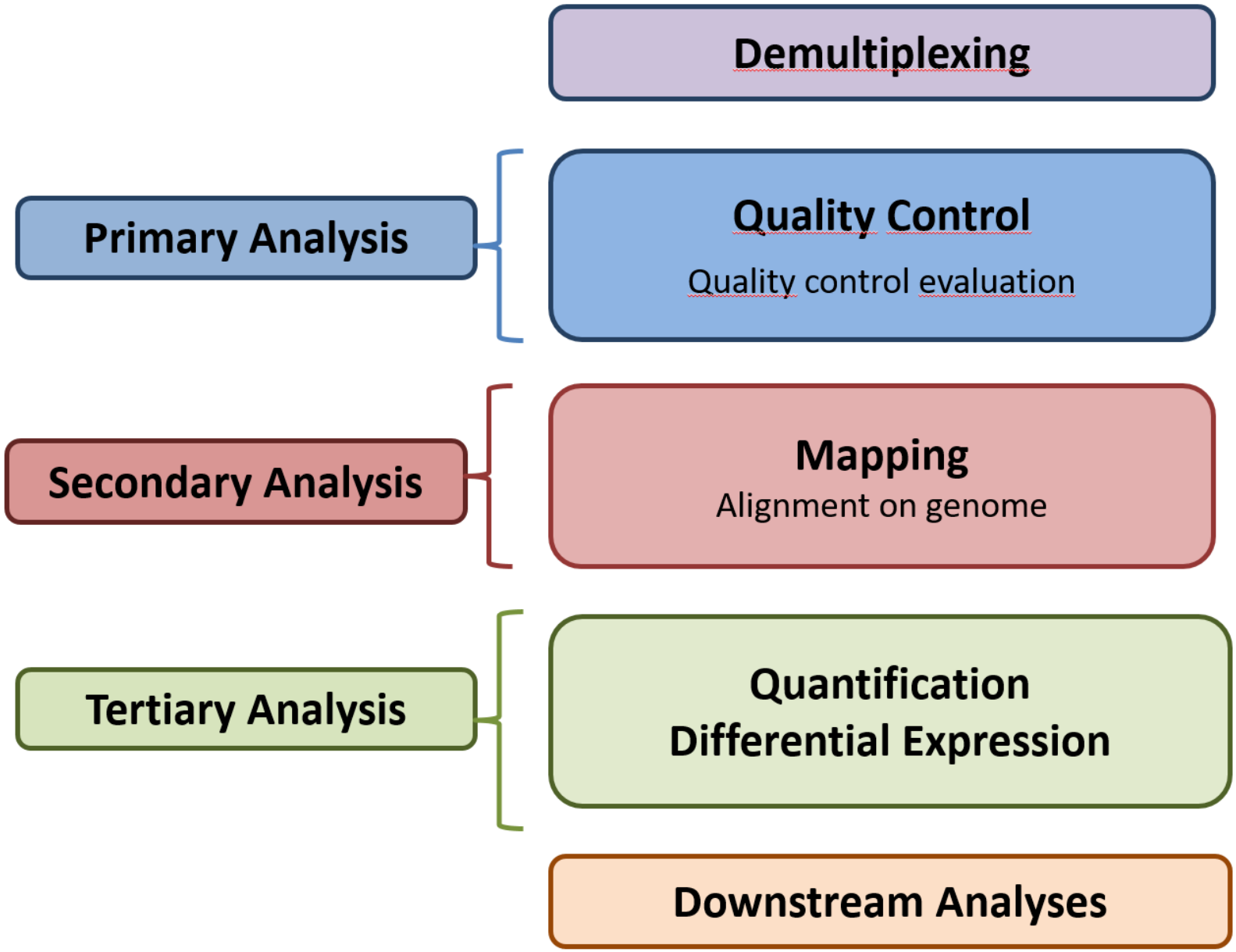
Bioinformatics and Reproducibility

RNA-seq pipeline

1. Assess reads quality (QC)
2. Clean reads from noise
3. Map reads agains the reference genome
4. Visualize/Manipulate aligned reads
5. Post process

RNA-seq pipeline



QC

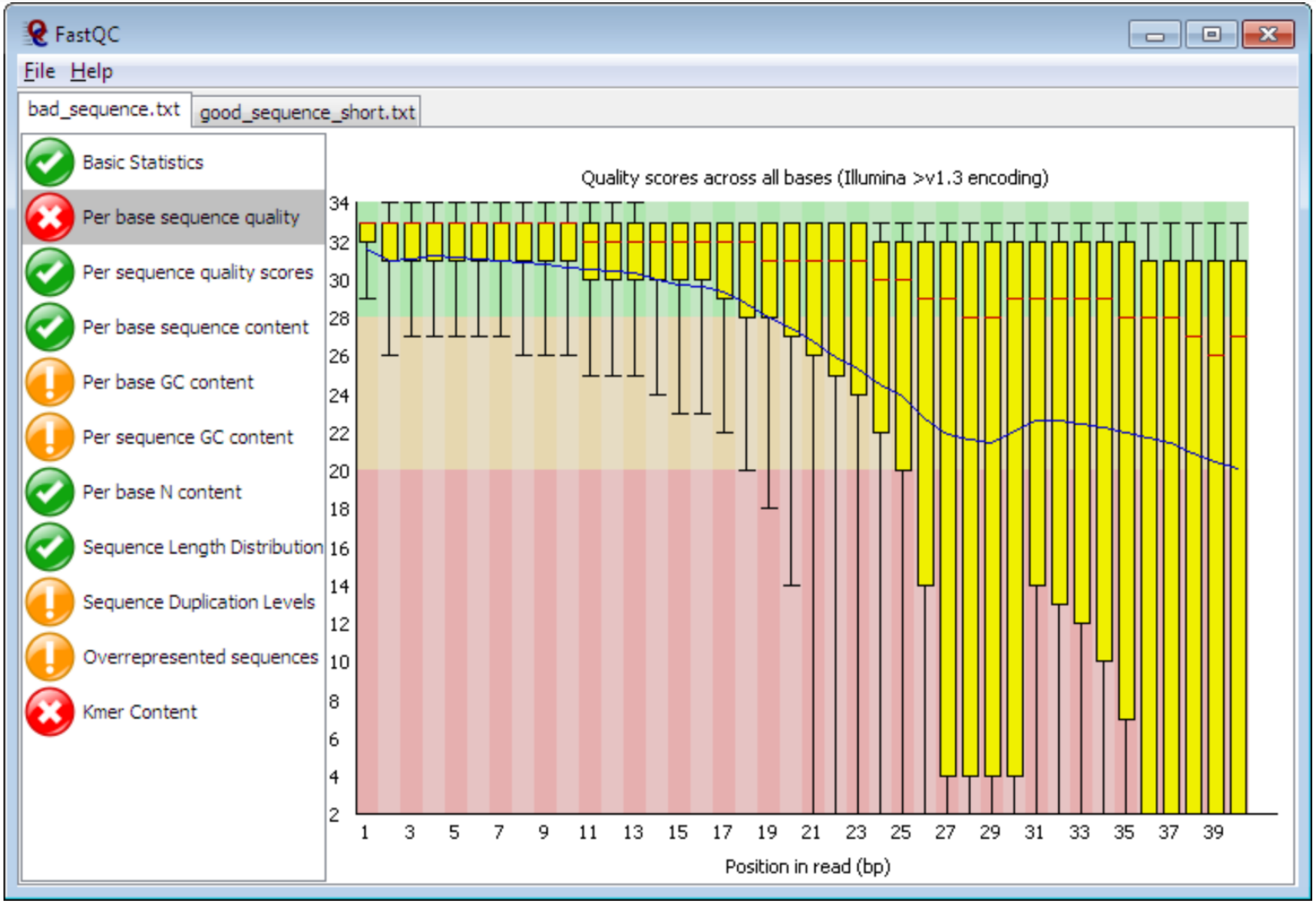


Figure 2: FastQC

QC

- 04-10-18: Version 0.11.8 released
  - Fixed a performance bug in highly duplicated sequences
  - Changed the behaviour of the sequence length module when run with --nogroup
  - Other minor bug fixes
- 10-01-18: Version 0.11.7 released
  - Fixed a crash if the first sequence in a file was shorter than 12bp
- 21-12-17: Version 0.11.6 released
  - Disabled the Kmer plot by default
  - Fixed a bug when long custom adapters were being used
  - Changed the tile number cutoff to accommodate the novaseq
  - Fixed various format changes in nanopore data from ONT
  - Added new Clontech sequences to the contaminant list
  - Added a --min-length option to remove short sequences
  - Added an option to specify the output name of data streamed into the program
- 08-03-16: Version 0.11.5 released
  - Fixed the smallRNA adapter sequence so that abundance isn't under-represented in the adapter content plot
  - Fixed a bug in the warn / error code for the per-base sequence content plot
  - Fixed a typo in the documentation for the duplication plot
- 09-10-15: Version 0.11.4 released
  - Changed the OSX launcher to not rely on the internal JVM framework, but use any command line java which is found
  - Fixed a typo in one of the adapter sequences
  - Fixed a bug which meant that some file extensions weren't removed from report names in non-interactive mode
  - Made the per-tile module not collect any stats if it's disabled in limits.txt
  - Fixed a bug in the calculation of duplication for highly duplicated, ordered files with very small numbers of sequences
  - Fixed an incorrect error flag in the per-base quality module where there were less than 100 observations in a read group
- 25-3-15: Version 0.11.3 released
  - Fixed a bug when disabling the per-tile plot from limits.txt
  - Fixed a bug which caused the program to continue when processing of multiple files was actually complete
  - Fixed a bug which meant format selection in the interactive application didn't work
  - Added checks for mis-itentifying tile numbers in confusing sample ids
  - Added the SOLID smallRNA adapter to the standard search set
  - Fixed a bug when extracting casava names from uncompressed fastq files
  - Added support for processing files of Oxford Nanopore reads
- 6-6-14: Version 0.11.2 released
  - Fixed incorrect warn/fail defaults for per-seq quality plot
  - Fixed memory leaks in Kmer and per-seq quality modules
  - Added an option to use a custom limits file
  - Fixed a bug in the naming of the folder inside the zip output file
  - Fixed a bug in the --extract option
- 2-6-14: Version 0.11.1 released
  - Added configurable warn/fail thresholds for all modules
  - Allow modules to be selectively turned off
  - Added a per-tile quality plot for Illumina libraries
  - Added an adapter content plot
  - Improved the duplication plot
  - Improved the Kmer module
  - Used embedded graphics in the HTML output so you can distribute a single file
  - Added the ability to read data from stdin
  - Changed how base grouping works to better accommodate long reads
  - Dropped support for Solexa64 format (NB **not** Phred 64 which is still supported)
- 3-5-12: Version 0.10.1 released
  - Added a workround to allow the analysis of concatenated gzipped files
  - Fixed a bug when FastQC was installed in a path containing characters needing to be escaped in a URL
  - Added an option to specify the location of the java interpreter on the command line
- 9-9-11: Version 0.10.0 released
  - Added a Casava mode to sanely process the multiple fastq files produced by the latest illumina pipeline
  - Fixed a bug in Kmer analysis which missed of the last possible Kmer in each sequence
  - Fixed a classpath bug if using the wrapper script under windows
- 31-8-11: Version 0.9.6 released
  - Fixed a crash in libraries where every sequence ended in poly-N
  - Fixed the launch wrapper to set the classpath correctly on OSX
- 16-8-11: Version 0.9.5 released
  - Fixed a bug in text output for the per-base sequence content module
  - Made progress reporting absolute, and not approximate
  - Added a print CSS style so reports are printable again

Clean

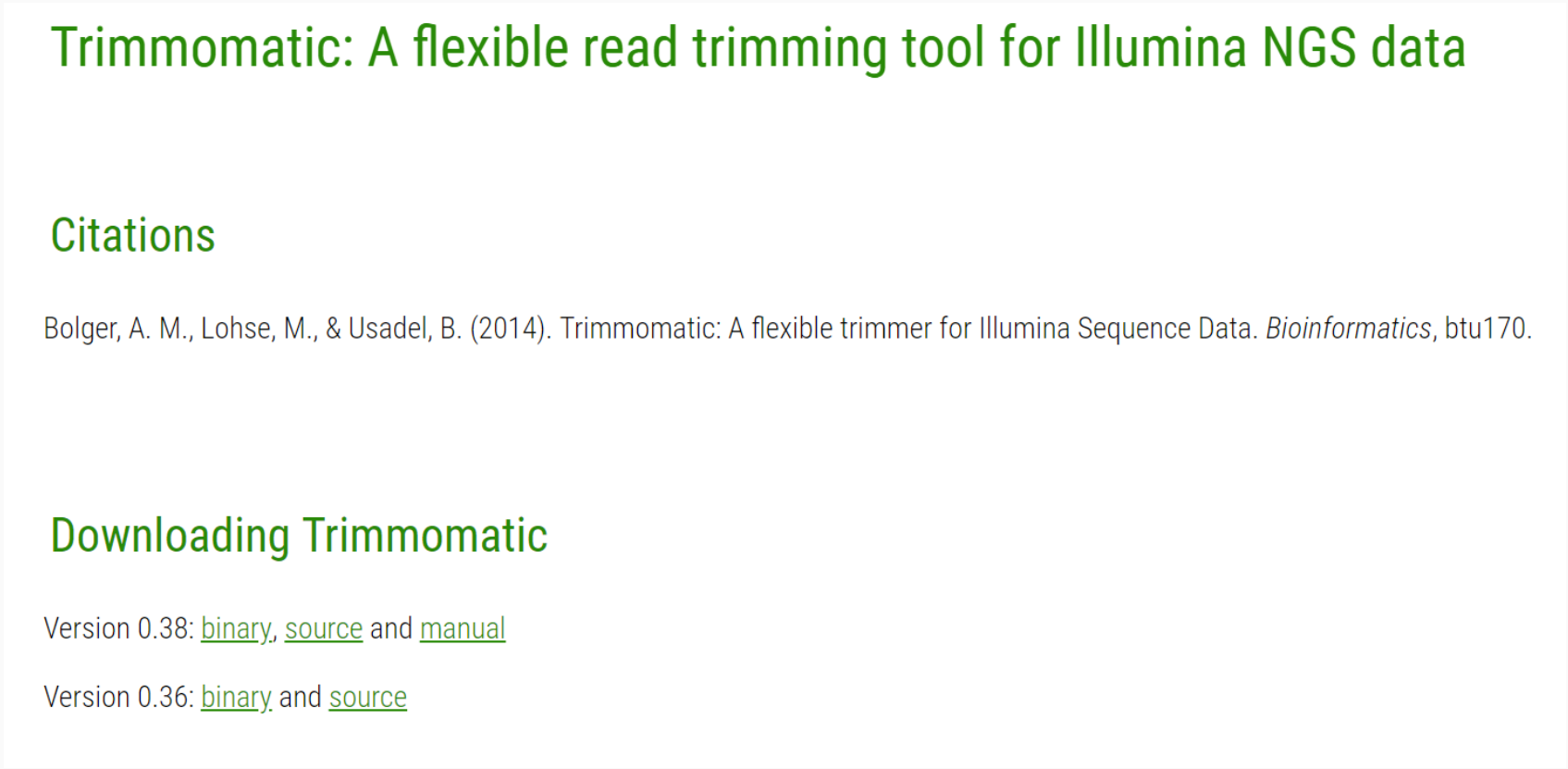


Figure 4: Trimmomatic

Mapping

Bioinformatics. 2013 Jan 1;29(1):15-21. doi: 10.1093/bioinformatics/bts635. Epub 2012 Oct 25.

STAR: ultrafast universal RNA-seq aligner.

Dobin A[1], Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR.

Author information

Abstract

MOTIVATION: Accurate alignment of high-throughput RNA-seq data is a challenging and yet unsolved problem because of the non-contiguous transcript structure, relatively short read lengths and constantly increasing throughput of the sequencing technologies. Currently available RNA-seq aligners suffer from high mapping error rates, low mapping speed, read length limitation and mapping biases.

RESULTS: To align our large (>80 billon reads) ENCODE Transcriptome RNA-seq dataset, we developed the Spliced Transcripts Alignment to a Reference (STAR) software based on a previously undescribed RNA-seq alignment algorithm that uses sequential maximum mappable seed search in uncompressed suffix arrays followed by seed clustering and stitching procedure. STAR outperforms other aligners by a factor of >50 in mapping speed, aligning to the human genome 550 million 2 × 76 bp paired-end reads per hour on a modest 12-core server, while at the same time improving alignment sensitivity and precision. In addition to unbiased de novo detection of canonical junctions, STAR can discover non-canonical splices and chimeric (fusion) transcripts, and is also capable of mapping full-length RNA sequences. Using Roche 454 sequencing of reverse transcription polymerase chain reaction amplicons, we experimentally validated 1960 novel intergenic splice junctions with an 80-90% success rate, corroborating the high precision of the STAR mapping strategy.

AVAILABILITY AND IMPLEMENTATION: STAR is implemented as a standalone C++ code. STAR is free

Figure 5: STAR

Mapping


Tags
2.7.0c
6 days ago e205c13 zip tar.gz Notes
2.7.0b
9 days ago 66f06aa zip tar.gz
2.7.0a
21 days ago 4247708 zip tar.gz
2.6.1d
on 16 Nov 2018 909a3b9 zip tar.gz Notes
2.6.1c ...
on 17 Oct 2018 ffd8416 zip tar.gz Notes
2.6.1b
on 6 Sep 2018 456c589 zip tar.gz Notes
2.6.1a
on 15 Aug 2018 45f7bd7 zip tar.gz Notes
2.6.0c
on 11 May 2018 391d99f zip tar.gz Notes
2.6.0b
on 3 May 2018 305d810 zip tar.gz
2.6.0a
on 24 Apr 2018 2b95e47 zip tar.gz

Visualize/Manipulate

Create index from BAM w/

Samtools

Samtools is a suite of programs for interacting with high-throughput sequencing data. It consists of three separate repositories:

Samtools	Reading/writing/editing/indexing/viewing SAM/BAM/CRAM format
BCFtools	Reading/writing BCF2/VCF/gVCF files and calling/filtering/summarising SNP and short indel sequence variants
HTSlib	A C library for reading/writing high-throughput sequencing data

Samtools and BCFtools both use HTSlib internally, but these source packages contain their own copies of htslib so they can be built independently.

Visualize/Manipulate

Tags

1.9 ...
on 18 Jul 2018 b24d812 zip tar.gz Notes Downloads

1.8 ...
on 3 Apr 2018 61e9762 zip tar.gz Notes Downloads

1.7 ...
on 26 Jan 2018 6d79411 zip tar.gz Notes Downloads

1.6 ...
on 28 Sep 2017 f4dc22a zip tar.gz Notes Downloads

1.5 ...
on 20 Jun 2017 f510fb1 zip tar.gz Notes Downloads

1.4.1 ...
on 8 May 2017 8f8600a zip tar.gz Notes Downloads

1.4 ...
on 13 Mar 2017 0a859b1 zip tar.gz Notes Downloads

1.3.1 ...
on 22 Apr 2016 897c002 zip tar.gz Notes Downloads

1.3 ...
on 15 Dec 2015 ed3191b zip tar.gz Notes Downloads

1.2 ...
on 2 Feb 2015 255f97d zip tar.gz Notes Downloads

Post process

look at data w/bash

Setup

In order to perform the analysis you must have:

1. An environemnt w/all the required software
2. The reference genome
3. The raw reads to analyize
4. Hopefully automate the analysis

Setup: software

You can:

- create install eveything from scratch
- reinstall software with conda
- re-use the `unimiPhD` conda environment

Setup: conda environment

```
conda activate unimiPhD
```

or

```
source activate unimiPhD
```

The ref. fasta

As a toy example we can download only the Human chromosome 19

```
wget ftp://ftp.ensembl.org/pub/release-95/fasta/homo_sapiens/c
```

The ref. fasta

```
gunzip Homo_sapiens.GRCh38.dna.chromosome.19.fa.gz
```

The ref. annotation

And then the annotation in GTF format

```
wget ftp://ftp.ensembl.org/pub/release-95/gtf/homo_sapiens/Hom
```

The ref. annotation

Uncompress the file

```
gunzip Homo_sapiens.GRCh38.95.gtf.gz
```

Build the index

```
STAR --runThreadN 4 --runMode genomeGenerate --genomeDir ./Ger
```

Error

```
Feb 15 00:55:42 ..... started STAR run
Feb 15 00:55:42 ... starting to generate Genome files

Genome_genomeGenerate.cpp:208:genomeGenerate: exiting because
Solution: check that the path exists and you have write permis

Feb 15 00:55:45 ...... FATAL ERROR, exiting
Command exited with non-zero status 109
```

Build the index

```
mkdir GenomeDir
```

```
STAR --runThreadN 4 --runMode genomeGenerate --genomeDir ./GenomeDir --sjd
```

Index build

```
Feb 15 00:56:27 ..... started STAR run
Feb 15 00:56:27 ... starting to generate Genome files
Feb 15 00:56:29 ... starting to sort Suffix Array. This may ta
Feb 15 00:56:30 ... sorting Suffix Array chunks and saving the
Feb 15 00:57:49 ... loading chunks from disk, packing SA...
Feb 15 00:57:53 ... finished generating suffix array
Feb 15 00:57:53 ... generating Suffix Array index
Feb 15 00:58:49 ... completed Suffix Array index
Feb 15 00:58:49 ..... processing annotations GTF
Feb 15 00:59:07 ..... inserting junctions into the genome indi
Feb 15 00:59:59 ... writing Genome to disk ...
Feb 15 00:59:59 ... writing Suffix Array to disk ...
Feb 15 01:00:04 ... writing SAindex to disk
Feb 15 01:00:18 ..... finished successfully
```

Get the reads

E-MTAB-2319

```
curl -o ERR431583_1.fastq.gz ftp://ftp.sra.ebi.ac.uk/vol1/fast
curl -o ERR431583_2.fastq.gz ftp://ftp.sra.ebi.ac.uk/vol1/fast
```

Get the reads

E-MTAB-2319

From the local server

```
scp user@192.168.200.213:ERR431583_1.fastq.gz ERR431583_1.fast
scp user@192.168.200.213:ERR431583_2.fastq.gz ERR431583_2.fast
```

QC w/fastqc

```
fastqc -t 2 ERR431583_1.fastq.gz ERR431583_2.fastq.gz
```

Chop the size

```
zcat ERR431583_1.fastq.gz | wc -l
$ 58937804
zcat ERR431583_2.fastq.gz | wc -l
$ 58937804
```

that is the number of lines of each file.

Chop the size

```
58,937,804/4 = 14,734,451
```

Chop the size

Grap the first 1mln reads

```
zcat ERR431583_1.fastq.gz | head -n 4000000 | gzip > ERR431583
zcat ERR431583_2.fastq.gz | head -n 4000000 | gzip > ERR431583
```

Trimming

```
trimmomatic PE -threads 4 -phred33 ERR431583_downsize_1.fastq.
```

Trimming

```
Input Read Pairs: 15494812 Both Surviving: 14734451 (95.09%) F
TrimmomaticPE: Completed successfully
```

Alignment

```
STAR --genomeDir ./GenomeDir \
     --runThreadN 4 \
     --readFilesIn R1_P.fastq.gz R2_P.fastq.gz \
     --readFilesCommand zcat \
     --genomeLoad LoadAndRemove \
     --outFileNamePrefix MySample_ \
     --outReadsUnmapped Fastx \
     --outSAMstrandField intronMotif \
     --outFilterIntronMotifs RemoveNoncanonicalUnannotated \
     --quantMode GeneCounts \
     --outSAMtype BAM SortedByCoordinate \
     --limitBAMsortRAM 5000000000
```

Sorting reads

```
samtools sort -o MySample_SortedByName.bam -O bam -n -@ 4 BAMF
samtools index MySample_SortedByName.bam
```