ICG Tips and Tricks
November 2nd, 2021

Processing RNA-Seq data

For those using perun, remember to check for a conda environment for the programs
-conda info –envs |grep -i name_of_program
    -grep -i ignores case so you don't have to worry about Trinity vs trinity
-can ask Balagopal to install/update

For those using compute Canada
    -can check
    https://docs.computecanada.ca/wiki/Available_software
    -can ask for installation

Most programs run on the command line have limited help documents
name_of_program -h (or -help or --h or --help)

Some of the programs mentioned may be available in Galaxy and Geneious

RNA-Seq data

- -Illumina reads
- -50 bps to 300 bps
- -RNA
  - -mRNA
  - -rRNA
  - -small RNAs
- -paired-end
- -single-read

Typically paired-end RNA-seq data comes from the sequencing centre as
-two separate fastq files
    -R1
        -forward reads
    -R2
        -reverse reads

A few centres provide the paired end data as a single interleaved fastq file
    -JGI

Some older data sets have a single file with all R1 reads and then all R2 reads
    -SRA
    -probably makes sense to split them into R1 and R2 reads
        -find number of lines (wc -l filename)
        -divide the number of lines by 2
        head -(number_of_lines/2) > R1_file.fastq
        tail -(number_of_lines/2) > R2_file.fastq

```
@A00516:246:HJV73DSX2:2:1101:2139:1000 1:N:0:AGACCTTG+NCACGTAA
NTGAAGTTCACTTATGTTACCATGCATTATCCACACGCTGGTCCGGGAATATTAACCCGGTTCCCTTTCGATAGGCGGCGCACATGGCGCCCTTGACACGG
+
#FFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00516:246:HJV73DSX2:2:1101:2573:1000 1:N:0:AGACCTTG+NCACGTAA
NTCGTCTGCAAAAGATCTATCACTTCCAAAGTTTGAAATTGAGATTAGCCACAAAACCTTTGACGGTAAGAGCAGCTCGTATGCAACAGGGCCGAGACCTA
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,FFFFFFFFFFFFFFFFFFFFFFFFF:F:F
@A00516:246:HJV73DSX2:2:1101:3803:1000 1:N:0:AGACCTTG+NCACGTAA
CCTGAGAATAGCTGACGTCAGCACCCTCCTTTTCTGCATCAAGCTCTTCAGCTGGATAGTTCTTCTTCCAGAATTGCACCTCGTCTTTCGAATACAACTTC
+
FFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
```

header

sequence

Spacer line

quality scores

Fastq files

-4 lines per read

- -line 1(header) starts with @ followed by a unique identifier
  - -identifier usually followed by a space and additional information like 1 (R1) or 2 (R2), indexes
  - -sometimes the identifier includes the 1 or 2 to designate R1 or R2 (/1 or /2)
  - -other portions indicate instrument name, run id,flowcell id,tile number, x-cord of cluster,y-cord
- -line 2 – the sequence (A,C,G,T,N)
- -line 3 –a spacer line reserved for additional information, sometimes has copy of sequence
- -line 4 –quality scores -note that the sequence line and the quality score line should have the same number of characters

Interleaved fastq files

-R1 read pair followed immediately by R2 read pair

```
@A00516:246:HJV73DSX2:2:1101:1108:1000 1:N:0:AGGATAGC+NACCTTGG
NGAGAGATCAACTTCTTGAGGCGTGCTTGAAATTGTGCCTCTGCGAGAGGTTGGGACCATGATGCGAACCGGAGCTCGCTTGTCGTCAGGTTAATGTTCTG
+
#F:FFF:FF:FFFFFFFFFFFFF:FF:,F:F:FFF,FFFFFFFFF:FFFF,,FFF,FF:FFFFFFFFFFF,FFF,:FFF,,,:FF,F,,FFFF::,,,F,F
@A00516:246:HJV73DSX2:2:1101:1108:1000 2:N:0:AGGATAGC+NACCTTGG
AACAATCAACAGTTGATGGTAACATGCCCTCCTGGTGTACGACCAGGCATGCAGGTGCAATACACTCCGAAAATTAACATGACGAAAAGCGATCTCAGGTT
+
F,FF,:F:FFFF:FFFFF:FFF:FFFFFF::FF:FF:F:,FF:FF:FFFFFF:FF::::,:FF:F:F,FF,,FF:FF:F:FFFF:F:FFFF:,FF:,FF,:
@A00516:246:HJV73DSX2:2:1101:1904:1000 1:N:0:AGGATAGC+NACCTTGG
NTTGTTCCATCGGGAAGCTCGAGGAGCCCTTGCTGAAAGTAGTGAATGATTGCGTTGGCGAACTCCTCCTCCTCGAGCGACCACTTTCCCTTGCGTTCAGC
+
#FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF,FFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFF
@A00516:246:HJV73DSX2:2:1101:1904:1000 2:N:0:AGGATAGC+NACCTTGG
GTCACAAAAGCCTGAGCTGTGATCACGAGGTGTGGCATGGACCGTCCCTCGGCGAGCTATGACAGCCTCGCGTCTTCTGAAATGAGTGAAGACTCGAGCAA
+
FFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFFFFFFFFFFFFF:FFFFFFFFFFFFFFFFF:FFFFFFFF:F,FFFFFFFFFFFFFF:,F:
```

RNA-seq processing

- -QC (quality control)
  - -length
  - -quality
  - -level of duplication/redundancy

- -Trimming
  - -removing adapter sequences
  - -removing poor quality
  - -removing bad sequences
  - -removing short sequences

QC (quality control)

Most widely used QC program for Illumina reads

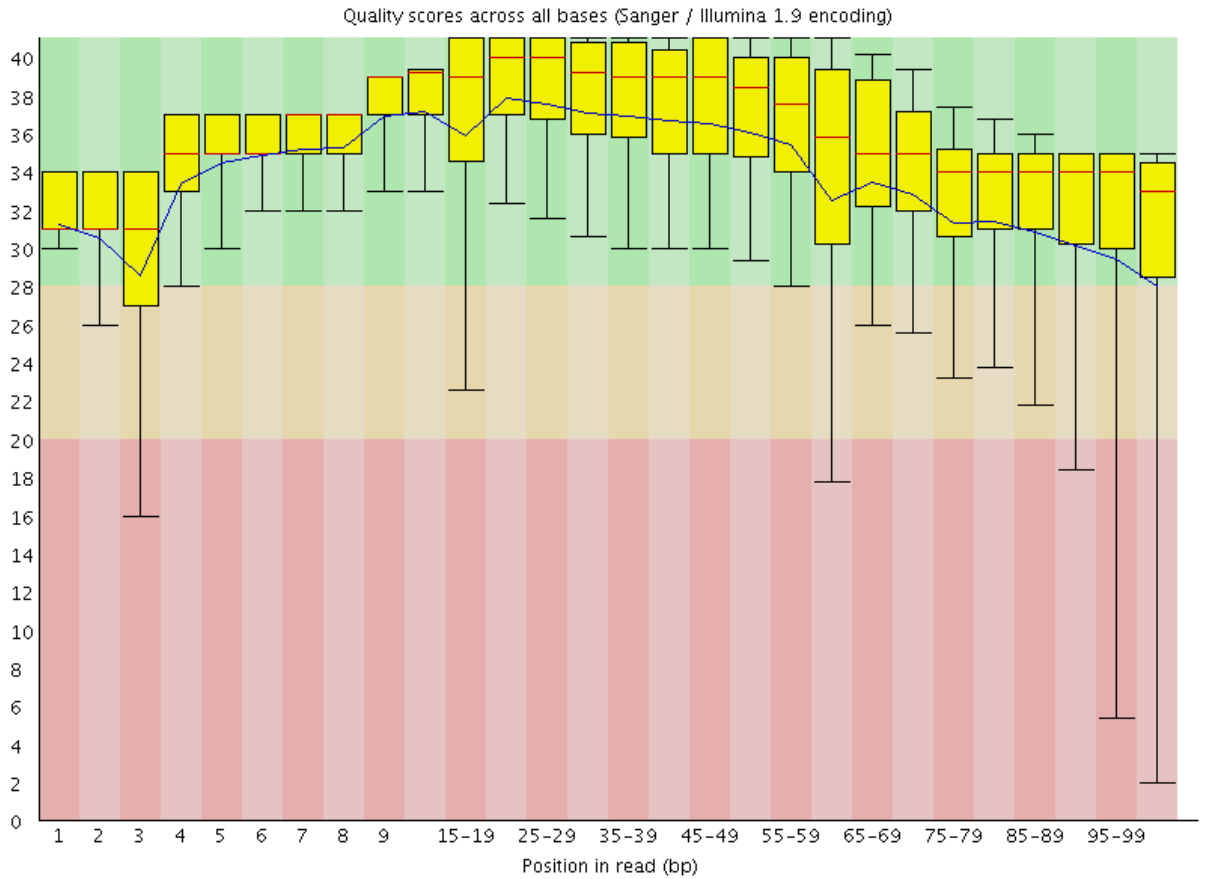
- -FastQC
  https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
- -graphical interface
- -command line
  - -generates html files for visual inspection

FastQC

-series of modules analyzing aspects of read data
-can used to get a sense of how well the library sequenced
-can be used to spot significant problems with the library
-CANNOT be used to micromanage the library
    -just trends
-some of the analyses are kind of useless/misleading
-results need to be interpreted in the context of what you want from the library
    -some libraries may be intentionally biased
-don’t spend too much time poring over the results

-for a detailed and user friendly explanation of what each module does and shows see https://dnacore.missouri.edu/PDF/FastQC_Manual.pdf

The **“Per base sequence quality”** plot provides the distribution of quality scores across all bases at each position in the reads.



Things to note

-colours correspond to very good (green), okay/probably keep (brown), poor quality (pink)

-phred scores are logarithmically linked to error probabilities

-20 is 99% correct, 30 is 99.9 %, 40 is 99.99% correct

-quality scores usually start to dip in the last 10 positions, and sometimes at the beginning

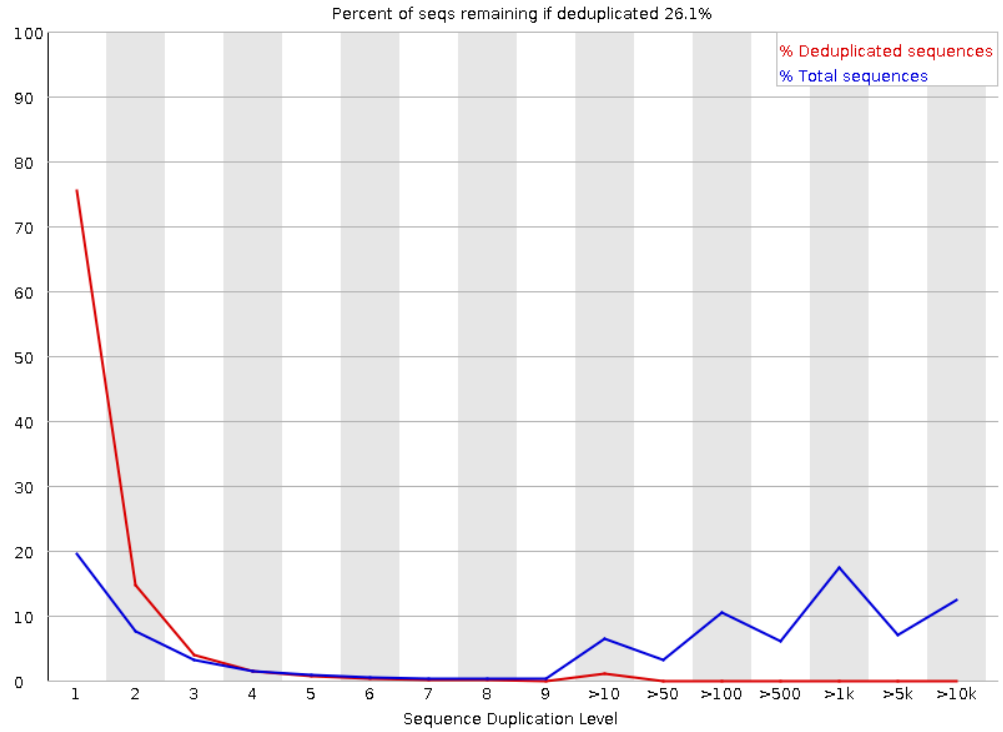
-makes a guess for the phred quality scoring code

-you should check both R1 and R2

Overrepresented sequences
-can indicate contamination
-can indicate biased library
-what type of library is it and how was it constructed
-only list those sequences representing more than 0.1% of total reads
-only test first 100,000 reads
-reads truncated to 50 bases

Note
-small list of overrepresented reads is not a true measure of the duplication level
-take the top overrepresented sequences and do a blastn at NCBI (more than likely from rRNA)
-the numbers quoted for deduplication are rough estimates and can be influenced by various factors (sequencing errors, only using the first 50 bps)

Sequence Duplication Levels



Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GTCGGTTGATACCATTTTGCCAACATCTTTAGTGCACGAATGTAAAACCA	135214	0.15860539298687196	No Hit
CGCGCCAGTTCTGAAGCGGCTGTTCAGCGCCCAGGGAGAGACCGCGAGGT	90784	0.10648920967444335	No Hit
CGGCATAGTTTATGGTTAAGACTAGGACGGTATCTGATCGTCTTTGATCC	85482	0.10026998833925324	No Hit

Adapter removal

-what is the adapter
    -literature can be a little confusing
        -some of the terms thrown around
            -adapter
            -primer
            -index
            -barcode
            -tag
            -insert

-some genome centres will remove adapters for you prior to releasing the data
    -beware of old data
    -to be on the safe side- analyze the library yourself
    -know what kit was used to make your library

Trimming for adapter sequence and quality

Two schools of thought

-brute force
- -remove a defined number of bases 5’ and 3’
- -fast
- -no thinking required
- -usually more than enough data

-refined
- -trim 3’ based on quality scores for each individual read
  - -need to determine values
    - -what phred values are acceptable?
    - -how many poor quality bases in a row?
- -check for adapter sequences for each individual read
  - -dependent on good quality sequences
    - -can miss if poor quality or read not long enough for match

Alternative View of Trimming

-some programs don't require trimming at all

-read mapping programs

-can do soft clipping of poor quality bases

-still need to trim off leftover adapter sequence

-still need to remove short reads

Mixture model

-you may end up doing some brute force trimming and more refined trimming

-common trimmomatic parameters

HEADCROP:15 LEADING:20 TRAILING:20 MINLEN:40

-remove first 15 bases from every read

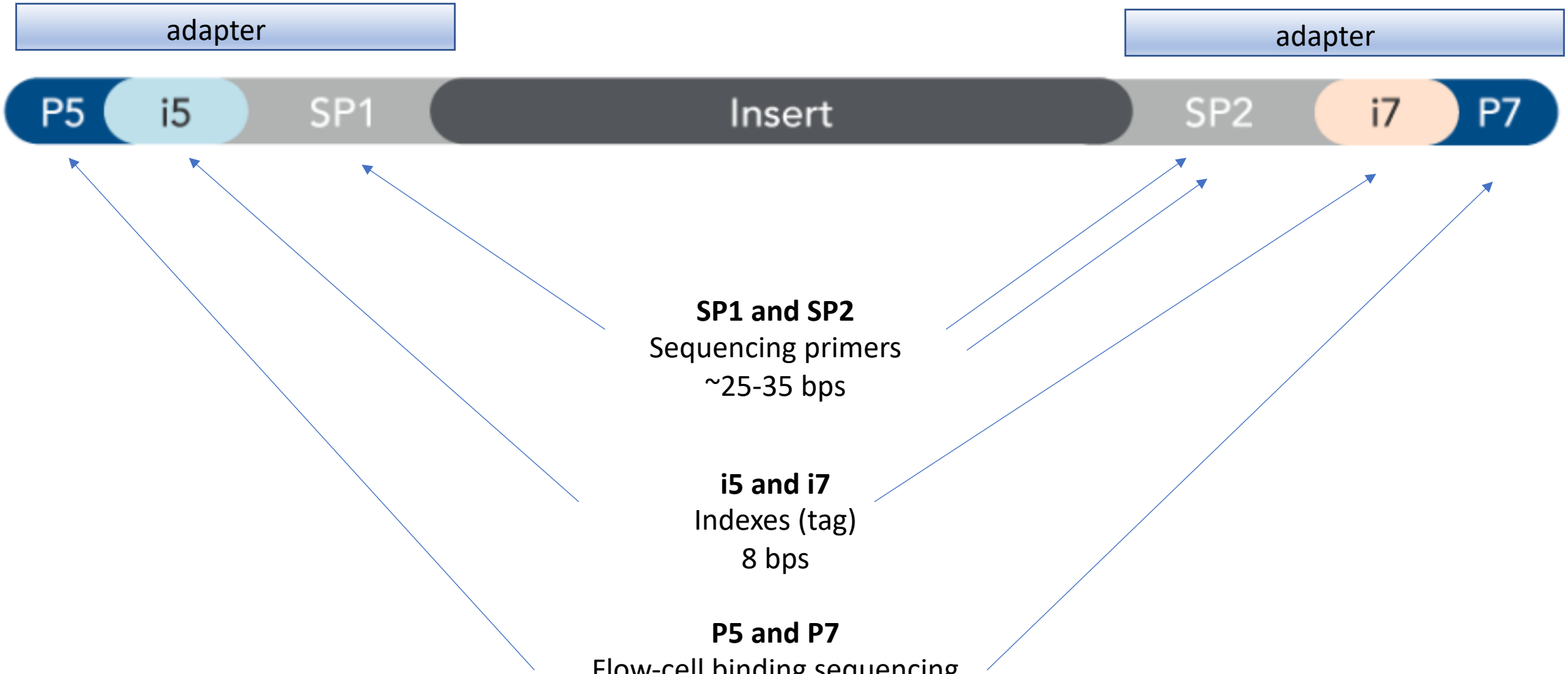
-remove bases from 5' until first base with quality score of 20

-remove bases from 3' until last base with quality score 20 or higher

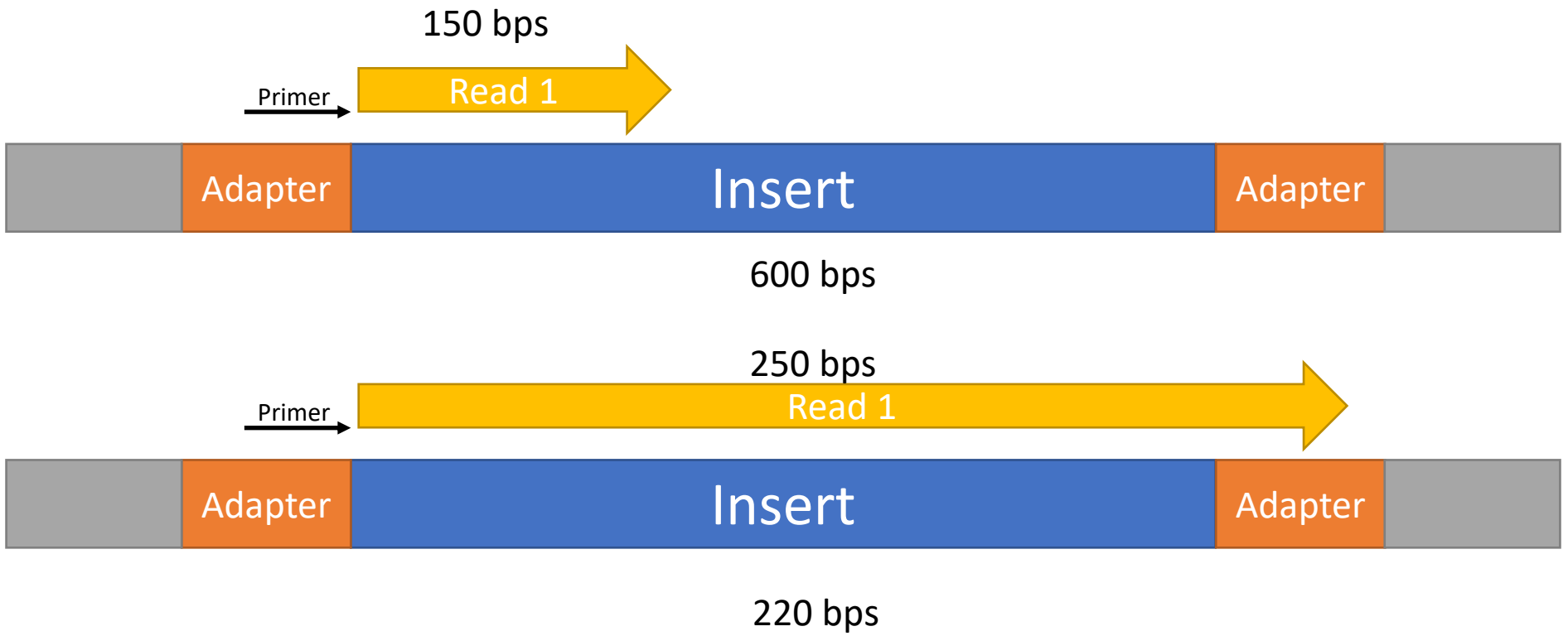
-remove any reads that are shorter than 40 bases after all the trimming

Adapters

Typical arrangement for paired-end reads



Depending on the insert size and the length of the read the read may run into the adapter at the other end
This means you have to remove the adapter sequence at the end of the read
The insert sizes will vary somewhat so the length of the adapter sequence at the 3’ end will vary from 0-the entire adapter



-almost all Ilumina reads do not have adapter sequences at the 5’
-Nextera mate pair reads have 5’ adapter sequences
-sequencing starts at the first base of the DNA insert
-Illumina reads may have adapter sequences on the 3’ end

How to know what the adapter sequences are?

Check with the sequencing centre

-may be part of the information they provide with your results

-may be part of their FAQs

Library Kit information

-https://support.illumina.com/bulletins/2016/12/what-sequences-do-i-use-for-adapter-trimming.html

-https://support.illumina.com/downloads/illumina-adapter-sequences-document-1000000002694.html

Trimming programs

Trimmomatic

-integrated into some RNA-seq analysis packages

Cutadapt

https://github.com/marcelm/cutadapt/

Fastx_toolkit

Illumina FASTQ Toolkit BaseSpace app

HTStream

https://bioinformatics.ucdavis.edu/software

Why doesn’t the program work?!?!?

What is wrong with my files?

Interleaved fastq file

-program wants separate r1 and r2 file

Solution

Seqtk

seqkit split2 –p 2 myinterleavedfile.fq

-you will have to rename the output files to something more informative

Reformat.sh from BBMap suite

reformat.sh in=myinterleavedfile.fq out1=myR1.fq out2=myR2.fq

The program wants fasta files but you have fastq files

Solution

-convert to fasta

Seqtk

seqtk seq –a nameofile.fq > thenewname.fa

-will also work with .fq.gz files

fastx_tookit

fastq_to_fasta -n -i myfile.fq -o mynewfile.fa

-n keeps sequences with unknown (N) nucleotides)

Wrong quality information

-some older data uses phred 64 quality scoring

How to tell which scoring is used?

Solution

-most programs have an option to indicate which quality scoring to use

-convert the scores

-BBMap

-**reformat.sh in=reads.fq out=reformatted.fq qin=64 qout=33**

-get a friendly bioinformatician to write a conversion script

Fastq Quality scoring


```
S - Sanger        Phred+33,  raw reads typically (0, 40)
X - Solexa        Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64,  raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64,  raw reads typically (3, 41)
    with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
    (Note: See discussion above).
L - Illumina 1.8+ Phred+33,  raw reads typically (0, 41)
P - PacBio        Phred+33,  HiFi reads typically (0, 93)
```

Most modern fastq files used Phred+33

-scores generally between 0(very low) and 41 (very high)

-most common types

-Phred+64 -any lower case letters, any upper case letters K and after

-Phred+33 –any numbers, any upper case letters A-J

Rnaseq files are named incorrectly

-during the processing you may have altered the name of the files considerably
-some programs require that the name of the files has a certain extension
    -fa, fasta, fastq, fq

Solution

-check the documentation for the program to see if they require a particular extension
-change the name
    mv oldname newname.proper_extension

After processing the files are no longer properly paired

-most programs using paired end data require that the pairs in the two files, R1 and R2 be in the same order

-your program will most likely quit or produce garbage

How to tell

-fast method

wc -l r1file

wc –l r2file

-the files should have the same number of lines

Solution

-use a trimming program that will preserve the order

-check documentation and what options are available

-if one of the reads is removed its pair in the other file will be removed

-run a program on the file to make the paired-end files paired again

-fastq_pair

fastq_pair file1.fastq file2.fastq

-bbmap

repair.sh in=aa in2=bb out=aanew out2=bbnew outs=singles repair=t -Xmx300m

I have contamination. How do I get rid of the reads?

Maybe a lot of your reads are derived from thing you don't care about (now)

-spike in rna

-rRNA

-take up computational power and time

Solution

-identify the sequence(s) you don't care about

-rRNA database

-use a mapping program (bowtie2) to map all your reads against the library of sequences you don't care about

-make the output those reads that don't map successfully to your library of sequences you don't care about

-make sure the reduced output files (R1 and R2) are still in order and have the same number of reads