**AR sequencing analysis pipeline – command line download**

**Purpose:** This document describes the bioinformatics analysis of antigen receptor sequencing data starting from the fastq file download from Basespace and including the Interrogate analysis.

The analysis can be done on a **local computer** or on a **Compute Canada cluster**. Transfer of fastq files from Basespace to a local computer using the ‘Download’ option on Basespace will create a different folder structure than the download using command line (as is the case when using ComputeCanada for the analysis). The following refers to the download via command line and analysis on a Compute Canada cluster.

#========= Transfer fastq files from Basespace to Compute Canada ==============

See separate document:

mkdir -p "Data/Basespace"

mv ds\* Data/Basespace/

#========= the resulting folder structure looks like this ==============

ls ds.0baa9e37cb924a74aa950860a791ea2f/

18-069171-2D1P3C1L1P1\_L001\_ds.0baa9e37cb924a74aa950860a791ea2f.json

Parry-Lola\_S23\_L001\_R1\_001.fastq.gz

Parry-Lola\_S23\_L001\_R2\_001.fastq.gz

#========= Clone folder with analysis scripts to correct location ==============

git clone https://github.com/theKellerLab/IntrgAnalysis

Unzip and copy to the following folder (e.g. 'Sequencing/)

‘Sequencing/Runs/FASTQ\_Generation\_2018-11-21\_19\_28\_48Z-138638908/’

The file tree should now look like this

‘Sequencing/Runs/FASTQ\_Generation\_2018-11-21\_19\_28\_48Z-138638908/’

Data

BasespaceData

IntrgAnalysis-master

0\_Move&QcFastqFiles

...

#====== Combine fastq in one folder =============

bash 0\_Move&QcFastqFiles/0\_moveOriginal.sh

Creates folder ‘Original’ with the following folder & file name:

16-091599-1\_D1P1C1P1C1/

16-091599-1\_D1P1C1P1C1\_L001\_R1\_001.fastq.gz

16-091599-1\_D1P1C1P1C1\_L001\_R2\_001.fastq.gz

No modification of script required

#====== Option 1: using fastqc & multiqc & trimmomatic (see fastp approach below) ===========

Advantage:

* fastqc output is amenable to multiQC analysis

Disadvantage:

* QC & trimming are two separate steps
* trimmomatic is apparently slow compared to other programs (<https://academic.oup.com/bioinformatics/article/34/17/i884/5093234>)

#====== run QC & multiQC =============

bash 1\_qcOriginal.sh

#====== trim reads =====

bash 2\_trim30.sh

#====== run QC & multiQC on trimmed reads =============

bash 3\_qcTrimmed30

#====== Option 2: using fastp ===========

Advantage:

* QC and trimming are done in one step
* fastp is faster than trimmomatic

Disadvantages:

* fastp output seems not amenable to multiQC analysis

1\_qcAndTrim\_fastp.sh

#========= upload to Interrogate ============

Upload