**AR sequencing analysis pipeline**

**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 the latter data from Basespace to ComputeCanada is covered in a separate document by Skyler. The following refers to a local analysis.

#========= Download sequence data from Basespace ==============

Download fastq file to local directory (for SK: ‘Sequencing/Runs/’)

Data will be dumped in a folder analogous to this:

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

Make additional dirs and move sequencing data:

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

#========= A comment on file structure ==============

Default file format after download:

Folder/

Read1file.fastq

Read2file.fastq

Depending on how the sample sheet was setup, there are currently various versions of folder & file names. Examples:

Run11 – MRD-LN:

axel-0-3D1\_L001-ds.20b13682e59946d987419a36b177b704/

axel-0-3D1\_S1\_L001\_R1\_001.fastq

axel-0-3D1\_S1\_L001\_R1\_001.fastq

Run22 – MRD-blood:

daisy-00-1D1P1C1P1C1\_L001-ds.fd23f8d8edd44ca6b50b75e6d7d74825/

Elliott-Daisy\_S1\_L001\_R1\_001.fastq.gz

Elliott-Daisy\_S1\_L001\_R2\_001.fastq.gz

Run21 - CNS

17-076750-2D1P4C1P1C1\_L001-ds.0ef4dbe5260f462a8e853019b8c3a81e/

Hegde-Tess\_S28\_L001\_R1\_001.fastq

Hegde-Tess\_S28\_L001\_R2\_001.fastq

**Problem1:** The name that will prevail through the Interrogate analysis is the file name. With the current setup, this would be ‘OwnerLast$Patient’, which is an equivocal designation if there are more than one sample per patient (e.g. MRD study). We want the sequencing ID (format xxxxxD\_P\_C\_L\_P\_C\_) to be included.

**Solution:** Combine a portion of the folder name (daisy-00-1D1P1C1P1C1) PLUS a portion from the file name (L001\_R2\_001.fastq.gz)

**Problem2:** The format of the folder/file names differs between run depending on what the database was capable off at the time of sequencing.

Solution: Set up new sample sheet with current version of database and re-run the analysis on BaseSpace. Note, the database does currently not allow pooling of different primer sets after PCR1 but that is being developed. Hence, start analyzing sequencing runs that only used one primer set (i.e. MRD runs).

Default file SequencingIDs:

16-091599-1\_D1P1C1P1C1

Pending: standardize control names

Patterns of sample combinations:

1. **2-2 pattern**: 2 samples per submission, 2 replicates per sample. Note that submission, samples and replicates are coded through distinct parts of the string. E.g. CNS or MRD-blood samples:
   * 16-091599-1\_D1P1C1P1C1
   * 16-091599-1\_D1P2C1P1C1
   * 16-091599-2\_D1P1C1P1C1
   * 16-091599-2\_D1P2C1P1C1
2. **1-2** **pattern**: 1 sample per submission, 2 replicates per sample. E.g. MRD-LN or diagnostic samples

The pattern is important for the structure of downstream R analysis scripts

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

Go to:

https://github.com/theKellerLab/IntrgAnalysis

Clone or Download > Download ZIP

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

#====== 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 =============

1\_qcOriginal.sh

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

2\_trim30.sh

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

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 ============