Module 1
Basic Quality Control of Raw Reads

The Fleming Fund
Regional Grants

17 February 2021

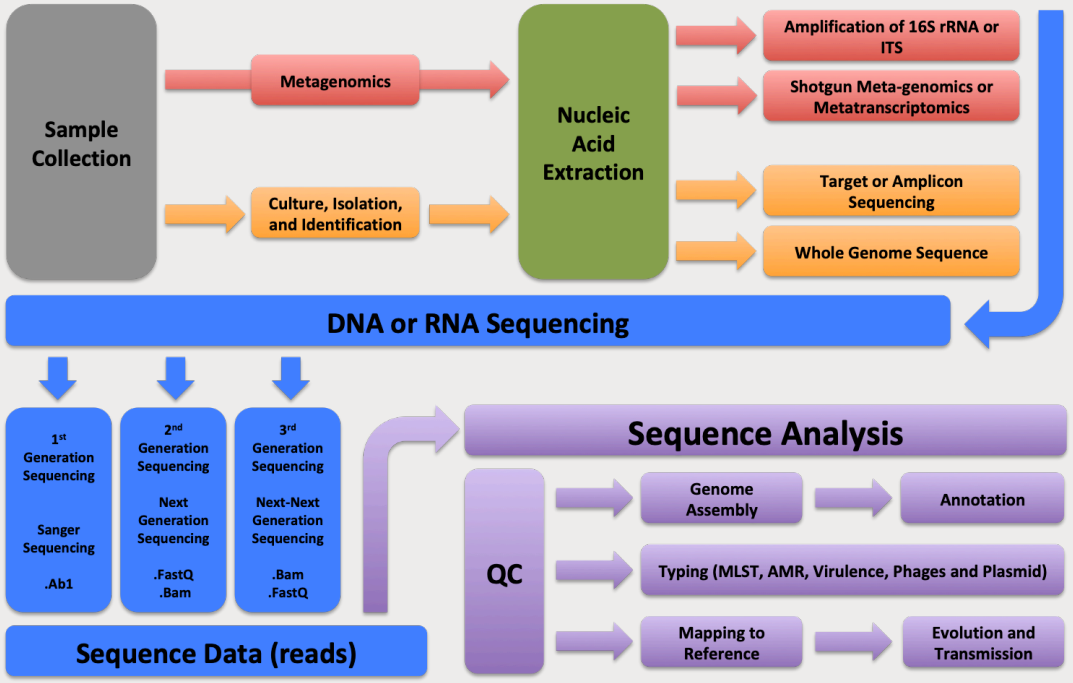
Mushal Allam,

NICD, South Africa

The Fleming Fund
Regional Grants

Introduction

- Modern high throughput sequencers can generate hundreds of millions of sequences in a single run
- Before analysing this sequence to draw biological conclusions you should always perform some simple quality control (QC) checks to ensure that the raw data looks good and there are no problems or biases in your data which may affect how you can usefully use it
- Most sequencers will generate a QC report as part of their analysis pipeline, but this is usually only focused on identifying problems which were generated by the sequencer itself



The Fleming Fund
Regional Grants

FASTQ Format

- FASTQ format has for long been the standard to store short-read sequencing data
- For a **single-read** run, one Read1 (R1) FASTQ file is created for each sample per flow cell lane
- For a **paired-end** run, one R1 and one Read2 (R2) FASTQ file is created for each sample for each lane
- FASTQ files contain the nucleotide sequence and the per-base calling quality for millions of reads
- FASTQ files are compressed and created with the extension *.fastq.gz
- Although file size will depend on the actual number of reads, FASTQ files are typically large (in the order of megabytes and gigabytes)
- Most tools that use FASTQ files as input can handle them in compressed format so, in order to save disk space, it is recommended not to uncompress them

TB17_S12_L001_R1_001.fastq.gz

- **TB17:** the sample name
- **S12:** the sample order in the sample sheet
- **L001** the lane number
- **R1**: represents Read 1
- **001:** represents the last segment
- **.fastq:** represents the FASTQ file
- **.gz:** compressed

- File contains millions of records
  - Each record has four lines, represents ONE sequence
    - Line 1 – the **name,** starts with **@**
    - Line 2 – the **sequence, starts at new line**
    - Line 3 – some **other** stuff, optional, starts with **+**
    - Line 4 – the **quality scores, starts at new line**

base = T
score = F = 37

```
@SN1083:379:H8VA1ADXX:2:1101:1248:2144 1:N:0:12
CCTAAATGGTGCCATGCTAGGAGGCCGTGCCCTTCTTGAAAAGTTGTATGTGAA
+
BBBFFFFFFBFFFIIIIFI<FFIIIIIFIIIIFBFIIIIIIIIFFFIIIIFIII
```

Table 1 ASCII Characters Encoding Q-scores 0–40

Symbol	ASCII Code	Q-Score	Symbol	ASCII Code	Q-Score	Symbol	ASCII Code	Q-Score
!	33	0	/	47	14	=	61	28
"	34	1	0	48	15	>	62	29
#	35	2	1	49	16	?	63	30
$	36	3	2	50	17	@	64	31
%	37	4	3	51	18	A	65	32
&	38	5	4	52	19	B	66	33
'	39	6	5	53	20	C	67	34
(	40	7	6	54	21	D	68	35
)	41	8	7	55	22	E	69	36
*	42	9	8	56	23	F	70	37
+	43	10	9	57	24	G	71	38
,	44	11	:	58	25	H	72	39
-	45	12	;	59	26	I	73	40
.	46	13	<	60	27			

The Fleming Fund
Regional Grants

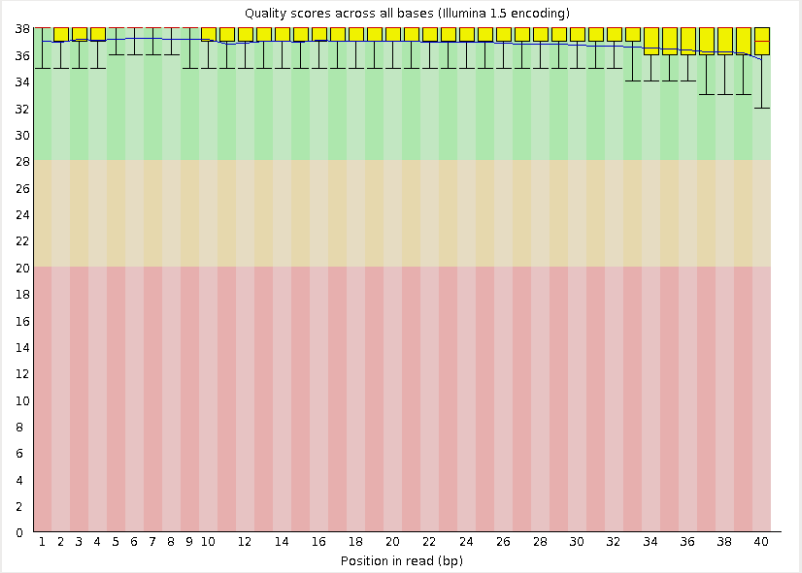
Phred Quality

- Phred quality score (Q) is a measure of the quality of the identification of the nucleobases generated by automated DNA sequencing

- If Phred score assigns a quality score of 30 (Q30) to a base, the chances that this base is called incorrectly are 1 in 1000

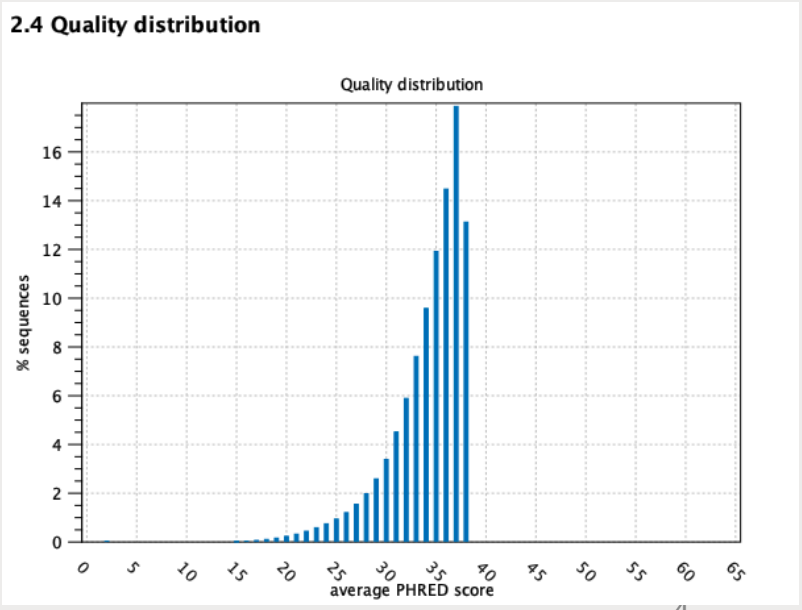
Phreq Quality score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1 000	99.9%
40	1 in 10 000	99.99%
50	1 in 100 000	99.999%
60	1 in 1 000 000	99.9999%

Quality Check
FastQC



CLC bio

QIAGEN



The Fleming Fund
Regional Grants

FASTQ Trimming

- Quality trimming based on quality scores
  - ≥Q20
- Ambiguity trimming to trim off e.g. stretches of Ns
  - ~~NNNNNNNNNNNN~~TACTACATAGATACAGATACTTA~~NNNNNN~~
- Adapter trimming
  - ~~CTGTCTCTTATACACATCT~~TACTACATAGATACAGATACTTA~~CTGTCTCTTATACACATCT~~
- Base trim to remove a specified number of bases at either 3' or 5' end
  - ~~ATACACATCT~~TACTACATAGATACAGATACTTA~~CTGTCTCTTA~~
- Length trimming to remove reads shorter or longer than a specified threshold
  - ~~TACTACATAGA~~ ≤15
  - ~~TACTACATAGA~~ ≥1000

Trimm-omatic

PRINSEQ

FASTX toolkit

FASTX-Toolkit
FASTQ/A short-reads pre-processing tools

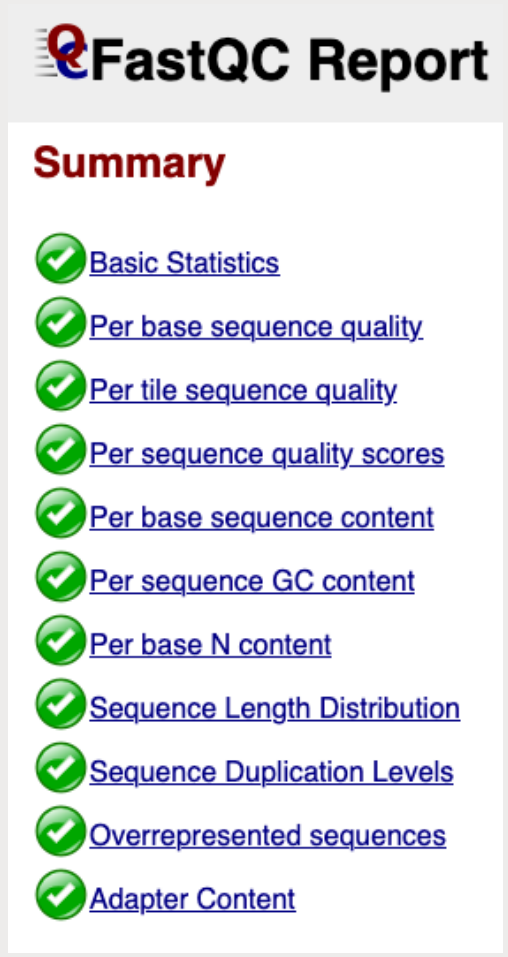
Babraham Institute Babraham Bioinformatics

Trim Galore

The Fleming Fund
Regional Grants

FastQC

- FastQC provide a QC report which can spot problems which originate either in the sequencer or in the starting library material
- FastQC can be run in one of two modes:
  - It can either run as a stand alone interactive application for the immediate analysis of small numbers of Fastq files
  - or it can be run in a non-interactive mode where it would be suitable for integrating into a larger analysis pipeline for the systematic processing of large numbers of files
- FastQC output the results in html format

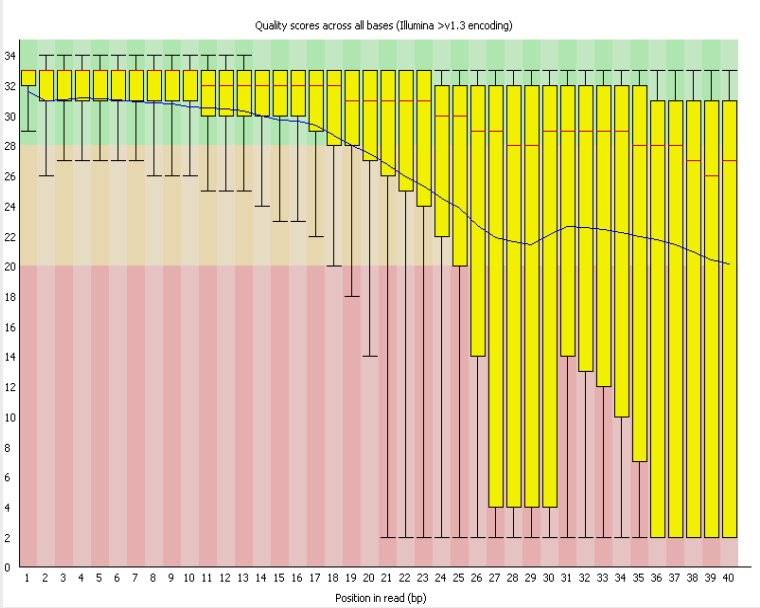


https://www.bioinformatics.babraham.ac.uk/projects/fastqc

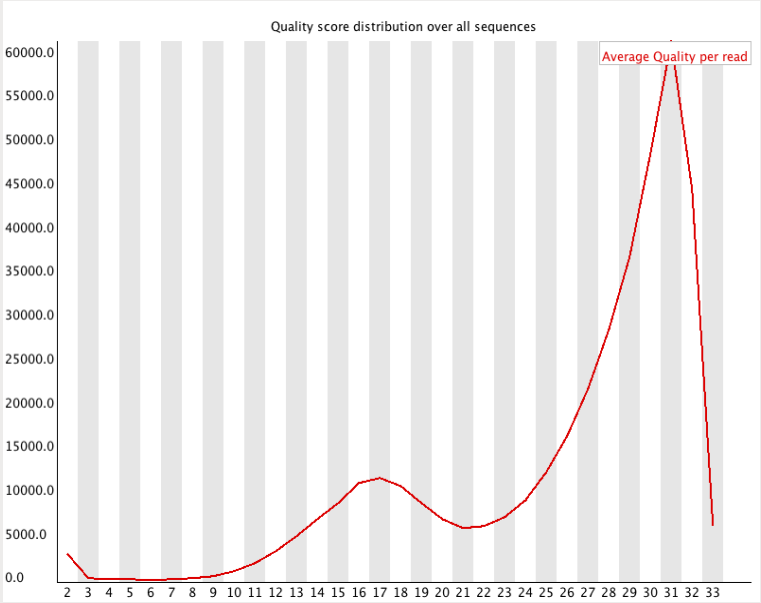
FastQC: Analysis Modules

Measure	Value
Filename	good_sequence_short.txt
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	250000
Sequences flagged as poor quality	0
Sequence length	40
%GC	45

Basic Statistics



Per Base Sequence Quality

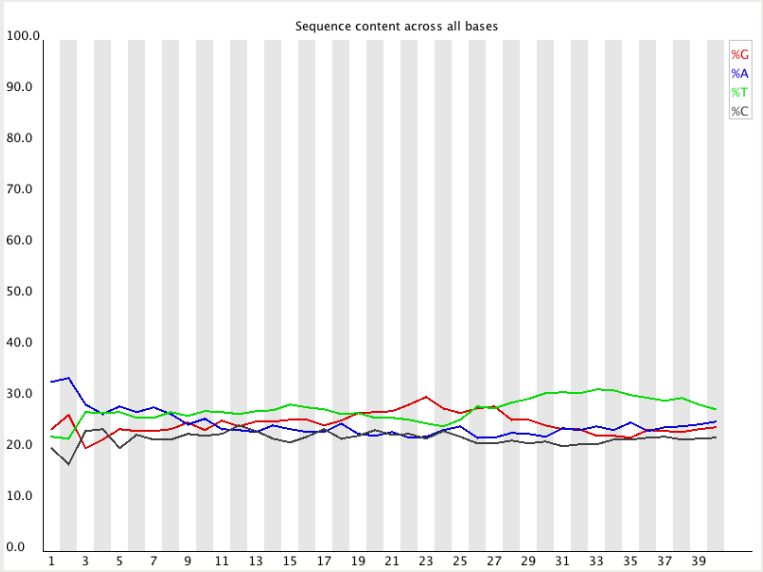


Per Sequence Quality Scores

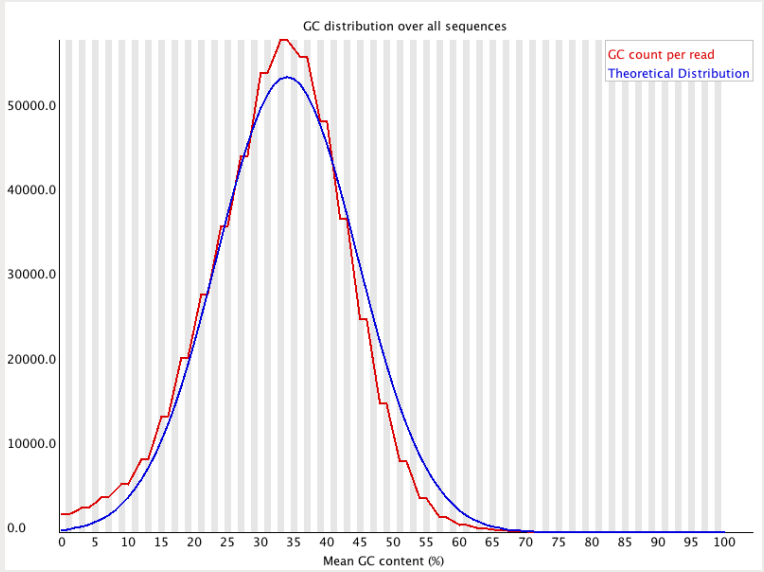
https://www.bioinformatics.babraham.ac.uk/projects/fastqc

The Fleming Fund
Regional Grants

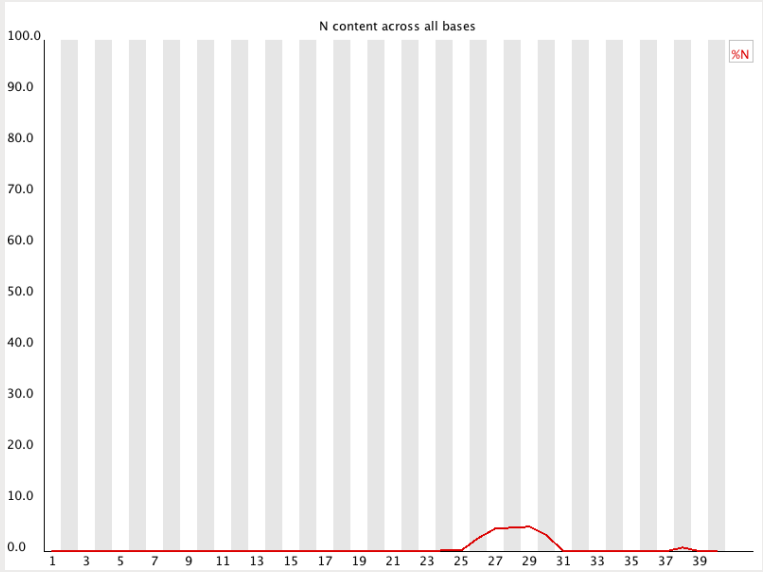
FastQC: Analysis Modules



Per Base Sequence Content



Per Sequence GC Content

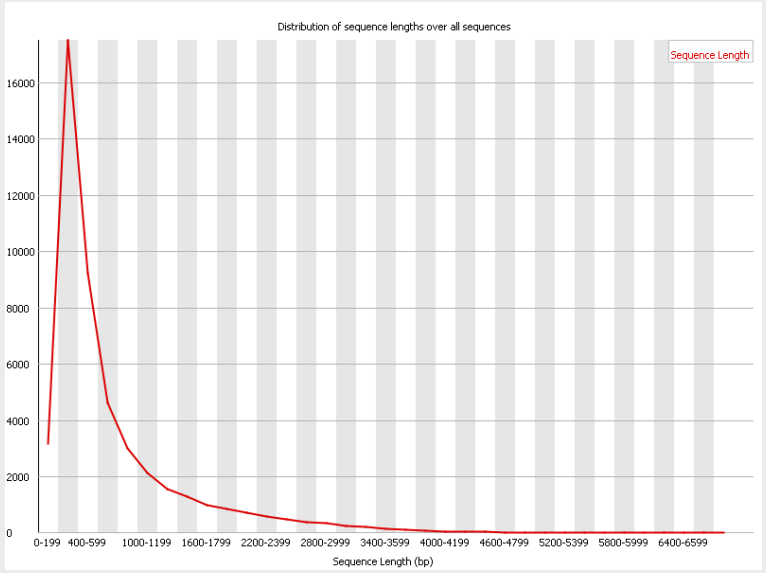


Per Base N Content

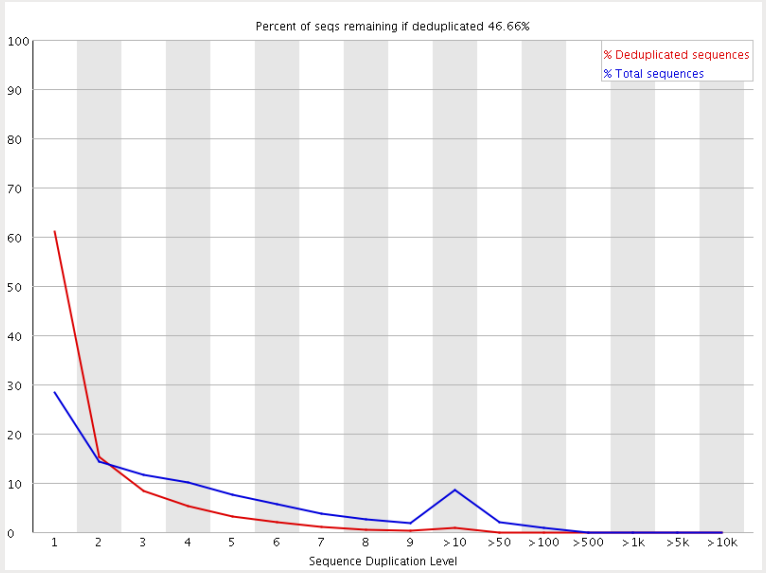
https://www.bioinformatics.babraham.ac.uk/projects/fastqc

The Fleming Fund
Regional Grants

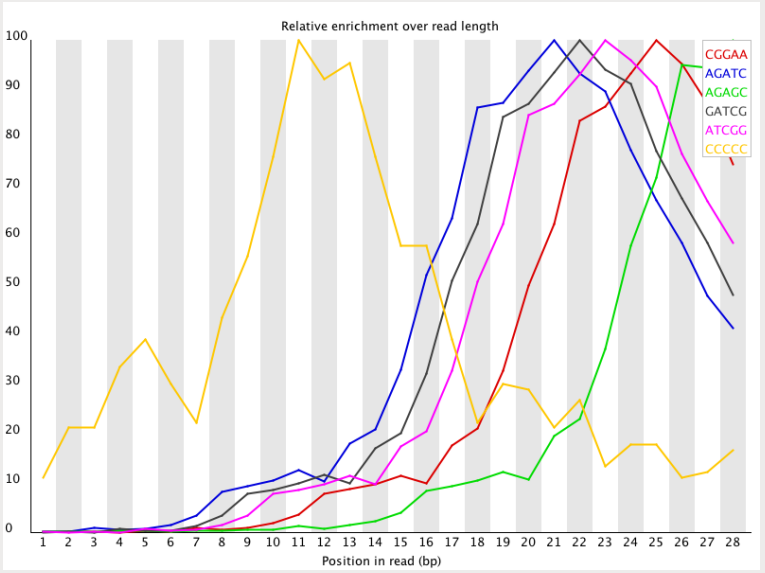
FastQC: Analysis Modules



Sequence Length Distribution



Duplicate Sequences



Kmer Content

https://www.bioinformatics.babraham.ac.uk/projects/fastqc

The Fleming Fund
Regional Grants

Thank you

NOGUCHI MEMORIAL INSTITUTE FOR MEDICAL RESEARCH
UNIVERSITY OF GHANA , LEGON

UNIVERSITY OF IBADAN
RECTE SAPERE FONS

NATIONAL INSTITUTE FOR COMMUNICABLE DISEASES
Division of the National Health Laboratory Service

KCRI Kilimanjaro Clinical Research Institute
Academic Centre for Evidence Based Health Interventions
An Institution of the Good Samaritan Foundation

UKaid
from the British people

MOTT MACDONALD

SEQAFRICA

DTU

This programme is being funded by the UK Department of Health and Social Care.
The views expressed do not necessarily reflect the UK Government's official policies.