this PDF was downloaded from minerva,
unzipping raw_data.zip using cyberduck failed but in the raw_data unzipped folder was this PDF.

Raw Sequencing Data

The original fluorescence images obtained from high throughput sequencing platforms are transformed to short reads by base calling. These short reads are recorded in FASTQ format, which contains base information (reads) and corresponding sequencing quality information.

1 Summary of result files

1.1 Files in the folder 'RawData'

- ***.fq.gz ('*' denotes sample ids)**

  Compressed FASTQ file created by gzip. Note: 1.fq.gz and 2.fq.gz contain read1 and read2 for paired-end sequencing, respectively or .fq.gz for single-end sequencing.

- **MD5.txt**

  MD5 hash for the compressed FASTQ files. The MD5 hash can be used to verify the integrity of files. If a file has been changed as a result of a faulty file transfer, its MD5 hash would be changed.

2 Explanation on FASTQ Format

Every read is stored in four lines in FASTQ format as follows:

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
GCTCTTTGCCCTTCTCGTCGAAAATTGTCTCCTCATTCGAAACTTCTCTGT
+
@@CFFFDEHHHHFIJJJ@FHGIIIEHIIJBHHHIJJEGIIJJIGHIGHCCF
```

Line 1 begins with a '@' character which is followed by a sequence identifier and an optional description.

Line 2 shows the sequenced bases.

Line 3 begins with a '+' character and is optionally followed by the same sequence identifier (and any description) again.

Line 4 encodes the sequencing quality for each base in line 2, and must contain the same number of characters as bases in line 2. The ASCII value of every character minus 33 equals to the phred-scaled quality score of the sequenced base.

Table Illumina sequence identifier datails

EAS139	The unique instrument name
136	Run ID
FC706VJ	Flowcell ID
2	Flowcell lane
2104	Tile number within the flowcell lane
15343	‘x’-coordinate of the cluster within the tile
197393	‘y’-coordinate of the cluster within the tile
1	Member of a pair, 1 or 2 (paired-end or mate-pair reads only)
Y	Y if the read fails filter (read is bad), N otherwise
18	0 when none of the control bits are on, otherwise it is an even number
ATCACG	Index sequence