#The first script summarizes the summary.txt file for all samples

It is located at:
`/bonn_data/Bonn_0_fastq/fastqcRawdata/summarizing_fastq_stats.sh`

#It can be run by simply writing `$ ./summarizing_fastq_stats.sh` and it outputs to screen and to a file

summarizing_fastq_stats.sh

```
#!/bin/bash

# summarizing_fastq_stats.sh is a bash program made by vifehe to
summarize the statistics outputs from fastaq/summary.txt
#
#The summary output is:
#[vifehe@detritus 1_paraparesia_fastaq]$ cat
paraparesia_fastq_QC/SN7570192_15190_P4H11_L5150_1_sequence.fq_fastqc/s
ummary.txt
#PASS   Basic Statistics
SN7570192_15190_P4H11_L5150_1_sequence.fq.gz      1
#PASS   Per base sequence quality
SN7570192_15190_P4H11_L5150_1_sequence.fq.gz  2
#PASS   Per sequence quality scores
SN7570192_15190_P4H11_L5150_1_sequence.fq.gz  3
#PASS   Per base sequence content
SN7570192_15190_P4H11_L5150_1_sequence.fq.gz  4
#PASS   Per base GC content
SN7570192_15190_P4H11_L5150_1_sequence.fq.gz      5
#WARN   Per sequence GC content
SN7570192_15190_P4H11_L5150_1_sequence.fq.gz      6
#PASS   Per base N content
SN7570192_15190_P4H11_L5150_1_sequence.fq.gz      7
#PASS   Sequence Length Distribution
SN7570192_15190_P4H11_L5150_1_sequence.fq.gz  8
#WARN   Sequence Duplication Levels
SN7570192_15190_P4H11_L5150_1_sequence.fq.gz  9
#PASS   Overrepresented sequences
SN7570192_15190_P4H11_L5150_1_sequence.fq.gz  10
#PASS   Kmer Content    SN7570192_15190_P4H11_L5150_1_sequence.fq.gz
11

idir=fastqcRawdata_P1
ofile=${idir}.sumstats

touch $ofile


printf "#BS = Basic statistics\n#PBSQ = Per base sequence
quality\n#PSQS = Per sequence quality scores\n#PBSQ = Per base sequence
content\n#bCG = Per base GC content\n#sGC = Per sequence GC
```

```
content\n#bN = Per base N content\n#SLD = Sequence Length
Distribution\n#SDL = Sequence Duplication Levels\n#OS = Overrepresented
sequences\n#KC = Kmer
Content\nSample\tBS\tPBSQ\tPSQS\tPBSQ\tbGC\tsGC\tbN\tSLD\tSDL\tOS\tKC\n
" >> $ofile

for x in $idir/*.fq_fastqc/summary.txt
do
    echo $x
    sample=(`echo $x | awk -F "/" {'print $2'} | awk -F"_" {'print $4"-
"$5'}`) #this should output L5150-1
    echo $sample
    basic_stats=(`cat $x | sed -n '1p' | awk -F"\t" {'print $1'}`) #
this should output the filter status of basic statistics
    echo $basic_stats
    per_base_seq_qual=(`cat $x | sed -n '2p' | awk -F"\t" {'print
$1'}`) # this should output the filter status of basic statistics
    echo $per_base_seq_qual
    per_seq_qual_scores=(`cat $x | sed -n '3p' | awk -F"\t" {'print
$1'}`) # this should output the filter status of basic statistics
        echo $per_seq_qual_scores
    per_base_seq_content=(`cat $x | sed -n '4p' | awk -F"\t" {'print
$1'}`) # this should output the filter status of basic statistics
        echo $per_base_seq_content
    per_base_GC_content=(`cat $x | sed -n '5p' | awk -F"\t" {'print
$1'}`) # this should output the filter status of basic statistics
        echo $per_base_GC_content
    per_seq_GC_content=(`cat $x | sed -n '6p' | awk -F"\t" {'print
$1'}`) # this should output the filter status of basic statistics
        echo $per_seq_GC_content
    per_base_N_content=(`cat $x | sed -n '7p' | awk -F"\t" {'print
$1'}`) # this should output the filter status of basic statistics
        echo $per_base_N_content
    seq_length_distr=(`cat $x | sed -n '8p' | awk -F"\t" {'print $1'}`)
# this should output the filter status of basic statistics
        echo $seq_length_distr
    seq_dupl_level=(`cat $x | sed -n '9p' | awk -F"\t" {'print $1'}`) #
this should output the filter status of basic statistics
        echo $seq_dupl_level
    overrepresented=(`cat $x | sed -n '10p' | awk -F"\t" {'print $1'}`)
# this should output the filter status of basic statistics
        echo $overrepresented
    kmer_content=(`cat $x | sed -n '11p' | awk -F"\t" {'print $1'}`) #
this should output the filter status of basic statistics
        echo $kmer_content

    printf
"$sample\t$basic_stats\t$per_base_seq_qual\t$per_seq_qual_scores\t$per_
base_seq_content\t$per_base_GC_content\t$per_seq_GC_content\t$per_base_
N_content\t$seq_length_distr\t$seq_dupl_level\t$overrepresented\t$kmer_
```

```
content\n" >> $ofile
done
```

a bit of the output is:

```
[vifehe@detritus fastqcRawdata]$ head -n20  fastqcRawdata_P1.sumstats
#BS = Basic statistics
#PBSQ = Per base sequence quality
#PSQS = Per sequence quality scores
#PBSQ = Per base sequence content
#bCG = Per base GC content
#sGC = Per sequence GC content
#bN = Per base N content
#SLD = Sequence Length Distribution
#SDL = Sequence Duplication Levels
#OS = Overrepresented sequences
#KC = Kmer Content
Sample  BS    PBSQ    PSQS    PBSQ    bGC    sGC    bN    SLD    SDL    OS
KC
MND1014-2   PASS    PASS    PASS    PASS    PASS    WARN    PASS    PASS
WARN    PASS    PASS
MND116-1    PASS    PASS    PASS    PASS    PASS    WARN    PASS    PASS
WARN    PASS    PASS
MND116-2    PASS    PASS    PASS    PASS    PASS    WARN    PASS    PASS
WARN    PASS    PASS
MND126-1    PASS    PASS    PASS    PASS    PASS    WARN    PASS    PASS
WARN    PASS    PASS
MND126-2    PASS    PASS    PASS    PASS    PASS    WARN    PASS    PASS
WARN    PASS    PASS
MND1405-1   PASS    PASS    PASS    PASS    PASS    WARN    PASS    PASS
WARN    PASS    PASS
MND1405-2   PASS    PASS    PASS    PASS    PASS    WARN    PASS    PASS
WARN    PASS    PASS
MND1493-1   PASS    PASS    PASS    PASS    PASS    WARN    PASS    PASS
WARN    PASS    PASS
```