Ig-HTS-Cleaner README

The program cleans the 454 reads from MID (molecular identification) tags and primers, and discards too-short or too-long sequences according to the input file as detailed in our paper:
Michaeli, M., Noga, H., Tabibian-Keissar, H., Barshack, I. and Mehr, R., (2012). Automated cleaning and pre-processing of immunoglobulin gene sequences from high-throughput sequencing. Frontiers in Immunology 3(December), 1-16.

Preparations:

- Please make sure your computer has a Java Runtime Environment installed. You can freely download it at: http://www.java.com/en/download/.
- Please save the attached JAR file in the same directory where all input files are saved.
- Running the program in Windows operating system is done by double clicking on the JAR file, after it is saved in the same directory with all input files.
- Although compatible with all operating systems, it is preferable to run this program on UNIX for large file sizes. After the JAR file is saved in the same directory with all input files, open a terminal window, enter this directory and use the following command: java -jar Ig-HTS-Cleaner.jar
- Please remember to change the name of the example input file into "input.txt".

Input:

Ig-HTS-Cleaner works on *.fna and *.qual files that are generated during the 454 run.
In addition, the user should prepare an input file exactly named "input.txt". This file includes all information for the run: MID tags used (if any), primers, requested length range and a table of samples and their used tags. For more clarifications regarding the input file please see the paper (Figure 2). Please find below an example of an input.txt file and follow the instructions using this example.

----------EXAMPLE START----------

```
organism
human
chain
h
quality threshold
20
maximum mismatches allowed
2
fraction of primer to search
0.75
range to search mids in
10
range to search primers in
25
minimal mid length
5
mids
ACGAGTGCGT
ACGCTCGACA
AGACGCACTC
AGCACTGTAG
ATCAGACACG
ATATCGCGAG
#
forward
TGCGMCAGGCCCCYGGACAAR
ARGRAAGGCCCTGGAGTGG
CCGCCAGGCTCCAGGSAAG
MGGGAAGGGRCTGGAGTGG
GAAAGGCCTGGAGTGGATGGG
TTGAGTGGCTGGGRAGGAC
#
reverse
TGACCRKGGTHCCYTGGCCC
#
minimum length
200
```

Fig. 2. An example of the input.txt file content. In bold-words/characters that should always appear. Organism - represents the organism to which the sequences belong, 'human' in this example. Chain - represents the chain of the Ig (h- heavy, l- lambda, k- kappa). Quality threshold – the minimal average score for a sequence allowed. Maximum mismatches allowed – the number of mismatches the user allows when primers are being searched. '2' means that when primers are being searched, the sequence can contain 2 insertions/ deletions or substitutions in the primer's sequence. Fraction of primer to search - in case the full primer was not found, the program searches only the given fraction of the primer from the side closer to the gene. Range to search mids/primers in - the search is executed on a limited range of bases at the ends of the read. Minimal MID length - in case the full MID was not found, the program searches for a perfect match of the minimal length of the MID from the side closer to the gene. Mids - a list of the MID tags that have been used in the current sequencing. The program automatically numbers the MID tags according to the insertion order. At the end of each list, a "#" should appear, see example. In case no MIDs were used, leave only the title and the "#". Forward - a list of the forward primers that have been used in the current sequencing, used for identification of the primers. At the end of each list, a "#" should appear, see example. Reverse - a list of the reverse primers that have been used in the current sequencing, used for identification of the primers. At the end of each list, a "#" should appear, see example. Minimum length - two values, the first is the minimal length for the first filtering of the data (minL1). The second value represents the minimal length that is legitimate for the genes in between the primers (minL2).

```
150
maximum length
400
360


table
1      1      sample1
3      3      sample2
2      2      sample3
3      1      sample4
5      3      sample5
6      2      sample6
#
```

Maximum length - two values, the first is the maximal length for the first filtering of the data (maxL1). The second value represents the maximal length that is allowed for the genes in between the primers (maxL2). Table - contains the MID tag combination per each sample that was sequenced. MID tag numbers should coordinate with their serial number in the above list. This enables the program to attribute each sequence to its corresponding sample. Each line should contain: number of the forward MID tag/tab/ number of the reverse MID tag /tab/ sample id (see example). At the end of each list, a "#" should appear, see example. In case no MIDs were used, put 0 as the number of forward and reverse MIDs.

----------EXAMPLE END----------

Output:

The program generates a number of files according to the number of samples listed in the table in the input file, which contain sequences belonging to each sample. The files contain the tag combinations in their names, for example: 'sample4_3_5.txt' where 'sample4' is the name of the sample, '3' is the forward tag and '5' is the reverse tag.

In addition, Ig-HTS-Cleaner generates three files containing sequences, as follows.

- failedInFindMIDs.txt – contains all sequences that did not have identifiable MID tags at both ends.
- failedInFindPrimers.txt – contains all sequences that did not have identifiable primers at both ends.
- failedInCheckLength.txt – contains all sequences that were too short or too long, based on the limits defined in the input file.

In addition, the program generates a log file (named log.txt), into which all statistics regarding the sequences are written. The first part of the log file gives details of how many sequences were identified, how many did not contain primers or were not in the right length, and how many were left for each sample. The second part contains some statistics of the run; for example, how many reads did not contain tags. Below is an example of the two parts.

----------EXAMPLE - FIRST PART ----------

Sample	Total	Failed in primers	% out of total	Failed in length	% out of total	% out of remaining	Failed in quality	% out of total	% out of remaining	Total remaining	Avg score
sample1	113	16	14.16	0	0	0	1	0.88	1.03	97	32
sample2	437	174	39.81	3	0.68	1.14	0	0	0	260	25
sample3	664	240	36.14	1	0.15	0.23	0	0	0	423	24
sample4	469	10	2.13	0	0	0	3	0.69	0.65	459	24
sample5	770	4	0.52	0	0	0	2	0.26	0.26	766	24

----------EXAMPLE SECOND PART----------

```
Total number of reads: 3591
How many reads were not within the first length boundaries and were discarded from
further analysis: 368
How many did not have MIDs at both ends of the sequence: 288
How many did not have primers at both ends of the sequence (either forward or reverse
primer, or both): 444
How many were not within the second length range: 4
How many were shorter than the range: 4
How many were longer than the range: 0
How many sequences had average quality scores below the threshold: 6
How many sequences are in a sense orientation out of the ok sequences: 1404
How many sequences are in a antisense orientation out of the ok sequences: 601
How many succeeded in partial match of forward primer: 157
How many succeeded in partial match of reverse primer: 84
How many failed in finding a forward primer: 444
How many failed in finding a reverse primer: 0
```

How many had primers, but both were sense or antisense, which probably means chimeric sequences: 0
How many have different MID combinations than those given in the input, out of those that have MIDs: 0
How many out of the total sequences could not be identified by their MIDs (percent): 8

MIDs were searched 5 nts from the side closer to the primers in order to deal with deletion of the edges.
The partial match was based on the following criteria:
taking 0.75 of primers' length from the side closer to the gene,
searching it in the 25 nts edges of the gene- depending on the orientation of the primer,
allowing maximum 2 mismatches (insertions/ deletions/ substitutions).
The threshold for the average quality score per sequence was: 20

----------EXAMPLE END----------