diagenode

A Hologic Company

Diagenode trimming tools for CATS Library Preparation Kits

(DNA, RNA)

Introduction

This document describes the usage of the Diagenode trimming tools for CATS Library Preparation Kits (DNA, RNA), which consists of two executable file, CATS_trimming_r1.sh and CATS_trimming_r2.sh.

They are designed to be easy to use. They are based on the trimming tool cutadapt (*), so be sure it is installed on your computer before running the Diagenode trimming tools. We recommend version 1.9 or higher. If cutadapt is not installed and you try to run the trimming tools, an error will result cutadapt is not found. If you are adept at (bio)informatics and you would like to customize the tools to your needs, feel free to open the files with a text editor and modify or add the necessary parameters. For the usage of cutadapt, please refer to the cutadapt manual.

(*) Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal, 17(1):10-12, May 2011. DOI: http://dx.doi.org/10.14806/ej.17.1.200

General usage

The scripts use a shell language, designed to be run in a Linux/Unix, Macintosh or a similar shell. They are executable by default, so you can download them and directly run them by opening a command line terminal and entering the file names and the path where you have downloaded them. If this does not work, please check the permissions of your downloaded file. Otherwise, you can also use the sh command with the file names (plus the path to them) to run a shellscript.

The trimming tools only require the name of the input file. Example usage (assuming the current directory is where the read files are):

```
sh /path/to/download/CATS_trimming_r1.sh MySample_R1.fastq
```

The output will be a file containing the trimmed reads, and the file name will be the same as the name of the input file plus the trimmed_ prefix. Thus in the above example the output file will be named trimmed_MySample_R1.fastq, and it will be located in the same folder as the input file. Note that trimming also works directly on compressed fastq files bearing the appropriate suffix, like MySample_R1.fastq.gz. In this case, the output file will also be a gzip compressed archive, with a name trimmed_MySample_R1.fastq.gz.

You should run the trimming tools from the directory of the read files. The trimming tool files can be anywhere, and you can access it by specifying the path to it, but when you designate the read file name, only the file name itself should be given. Other formats are currently not supported. If you also write the path to it (e.g. /path/to/reads/MySample_R1.fastq) the program will exit with an I/O error.

During its operation, cutadapt in the trimming tools will output a report on the screen, which details parameters, runtime and trimming statistics, in both per adapter and summary reports. If you want to save this report, redirect both the output and error messages of the trimming tools to a file, for example like this:

```
/path/to/download/CATS_trimming_r1.sh MySample_R1.fastq &> report.txt
```

The CATS library is designed to be sequenced on Illumina sequencers, and thus these trimming tools are also optimized for Illumina sequencing data. However, they may work on reads from other sequencers as well. Single reads are meant to be trimmed with CATS_trimming_r1.sh. If you have paired-end reads, use CATS_trimming_r1.sh for read 1 and CATS_trimming_r2.sh for read 2.

Troubleshooting

Issue: I cannot run the trimming tools. Permission denied.

Check the file permissions. The downloaded files should be executable for you, or your user group. If you have at least reading permissions, you should be able to run them with the sh command. Note that you might need administrator rights to change file permission.

Issue: I cannot run the trimming tools. Cutadapt: command not found.

Check if cutadapt is installed on your computer. We recommend version 1.9 or higher. If cutadapt is installed, check that it is in the PATH variable, so it can be run outside of its installation directory. You can use the export command to add the cutadapt installation directory to the PATH variable. As an alternative, one of the easiest way to make cutadapt available systemwide is to make a symbolic link (similar to a Windows shortcut) to a directory which is already in the PATH variable, like /usr/bin. For that you can use this command:

```
ln -s /path/to/cutadapt/installdir/cutadapt /usr/bin
```

To check which folders are in the PATH variable, use:

echo $PATH

Note that you might need administrator rights to export directories to PATH or create a symbolic link to /usr/bin.

Issue: I cannot run the trimming tools.

Check your error messages, and if you cannot solve the problem, please contact our customer support. Note that our tools (and cutadapt) will not run properly under Windows; it is designed for Linux and other Unix-like operating systems. It will also run on a Mac OS.

Issue: The trimming tools seem to work, but I cannot find the output files and/or I get an I/O error message.

The output file(s) should be in the same dierctory as the input file(s), no matter where the trimming tools are located. However, if you run the tools from another directory than where the input files are, and you specify the path to the input file, it will cause an error as cutadapt will not be able to generate an output file. Please use only the file names of the input files as an argument, and always run the trimming tools from the same directory where the input files are. To navigate to that directory, you can use the cd command before you run the trimming tools, either on the command line, or e.g. in your shell script.

```
cd /path/to/reads/
```

Issue: The trimming tools cannot recognize my read files..

Please check your file format. Our trimming tools can work with raw reads in fastq or fasta format, but not a mix of those, neither other formats, like sam or bam. Also note that you cannot use our tools in a pipeline (e.g. cat MySample_R1.fastq | CATS_trimming_r1.sh), you always have to specify the input file name as an argument to our trimming tools.

Issue: My read files are compressed, but the trimming tools doesn't recognize them.

Currently the following compressions are supported: gzip, bzip2 and xz. These are recognized by their suffixes, namely .gz, .bz2 and .xz, respectively. If you have a different suffix, e.g. MyFile.fq.gz_reads, then it will not be recognized, even if the file is a properly gzip compressed fastq file.

Issue: The trimming tools seem to be running, but no reads are removed from the input files.

The trimming tools are designed to remove artifacts characteristic of reads generated with our CATS kit. If it is run on other datasets, it might not find anything to remove. Also, check if your reads are trimmed already. In any case, the trimming tools output a report on the trimming process and statistics on the screen while they are running, so you can monitor exactly what is searched for and what has been removed.

Issue: I have paired-end reads. I ran the two trimming tools on the two read files, but after trimming I have different read numbers in the R1 and R2 files.

It is possible that there will be orphan reads after the trimming, because one read of the pair has become too short and is discarded, while the other read passes the filters and remains in the readset. In any way it should not cause any issues downstream, as most aligners have no problem with handling paired-end datasets with unequal read numbers, and depending on the settings they either omit the orphan reads or align them as single reads. You can also use various tools to remove/extract orphan reads either from the readsets or the alignments.

Issue: I have another problem that is not listed here.

Please contact our customer support, describing the problem with as much detail as possible (what files you are working with, what commands you used, what are the error messages, what is your operating environment etc.). You can do so via e-mail (Customer.Support@diagenode.com) or through our web interface: Technical Support

www.diagenode.com