Cutadapt removes adapter sequences from high-throughput sequencing reads

Marcel Martin
Department of Computer Science, TU Dortmund, Germany

Abstract
When small RNA is sequenced on current sequencing machines, the resulting reads are usually longer than the RNA, and therefore contain parts of the 3' adapter. That adapter must be found, and removed, error-tolerantly from each read before read mapping. Previous solutions are either hard to use or do not offer required features, in particular, support for colour-space data.
As an easy-to-use alternative, we developed the command-line tool cutadapt, which supports 454, Illumina and SOLiD (colour space) data, offers two adapter trimming algorithms, and has other useful features.
Cutadapt, including its MIT-licensed source code, is available for download at http://code.google.com/p/cutadapt/.

Introduction

The lengths of individual nucleotide sequences (*reads*) output by second-generation sequencing machines have reached 35, 50, 100 bps and more. When DNA or RNA molecules are sequenced that are shorter than this length, especially in small RNA sequencing experiments, the machine sequences into the adapter ligated to the 3' end of each molecule during library preparation. Consequently, the reads that are output contain the sequence of the molecule of interest and also the adapter sequence. Essential first tasks during analysis of such data are therefore to find the reads containing adapters and to remove the adapters where they occur. Only the relevant part of the read is passed on to further analysis. In some cases, finding adapters is a sign of contamination, and the reads containing them must be discarded entirely. For both tasks, we suggest the use of cutadapt, a user-friendly command-line program that supports a variety of file formats produced by second-generation sequencers. It especially supports colour-space data as produced by Applied Biosystems' SOLiD sequencer.

Other solutions for adapter trimming exist. Some software libraries, such as HTSeq by Simon Anders (http://www-huber.embl.de/users/anders/HTSeq/) and Biostrings (Pages et al., http://bioconductor.org/packages/release/bioc/html/Biostrings.html) offer the essential error-tolerant trimming routines, but HTSeq does not consider insertions and deletions, and both require the user to be able to write their own programs that use those routines. Some read-mapping tools, such as SOAP (version 1) by Li *et al.* [1], MAQ by Li *et al.* [2] and Novoalign (http://novocraft.com/) can also trim adapters, but this is only useful if the reads are to be mapped with the respective program. If one wants to use a different mapping tool without trimming capability, such as BWA [3], the solution is to use a stand-alone adapter trimmer, such as cutadapt. We are aware of two other tools of this type. Vectorstrip is part of the EMBOSS package [4]. It was originally developed to recognise and remove vector sequence contamination, which makes it slightly cumbersome to use in high-throughput sequencing experiments. Vectorstrip does not find partial adapter matches and does not support colour space. The program fastx_clipper, part of the FASTX-Toolkit by Assaf Gordon (http://hannonlab.cshl.edu/fastx_toolkit/) is another command-line tool in the same spirit as cutadapt, but it also does not support colour space data.

Cutadapt is the only stand-alone tool that can correctly trim colour-space reads. It supports FASTQ, FASTA and also SOLiD .csfasta/.qual input files. It outputs results in FASTA or FASTQ format; gzip-compression of input or output files is automatically detected.

Implementation

Cutadapt is mainly written in Python, but for speed, the alignment algorithm is implemented in C as a Python extension module. The program was developed on Ubuntu Linux, but tested also on Windows and Mac OS X. It should also work on other platforms for which Python is available.

Features

The program was initially developed to trim 454 sequencing data collected by Zeschnigk *et al.* [5]. As insertions and deletions within homopolymer runs are common in 454 data, cutadapt

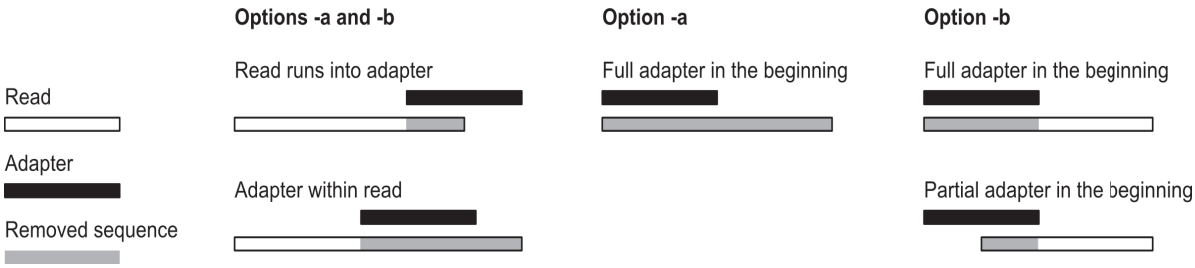

Figure 1. This illustration shows all possible alignment configurations between the read and adapter sequence. There are two different trimming behaviours, triggered by whether option "-a" or "-b" is used to provide the adapter sequence. Note that the case "Partial adapter in the beginning" is not possible with option "-a", as the alignment algorithm prevents it.

supports gapped alignment. For small RNA data analysis by Schulte *et al*. [6], the program was modified to support trimming of colour-space reads. It has also been tested and works well on Illumina data.

Cutadapt can search for multiple adapters in a single run of the program and will remove the best matching one. It can optionally search and remove an adapter multiple times, which is useful when (perhaps accidentally) library preparation has led to an adapter being appended multiple times. It can either trim or discard reads in which an adapter occurs. Reads that are outside a specified length range after trimming can also be discarded.

To decrease the number of random hits, a minimum overlap between the read and adapter can be specified. In addition to adapter trimming, low-quality ends of reads can be trimmed using the same algorithm as BWA. Cutadapt is thoroughly unit-tested. The program is actively maintained, and many features have been added in response to requests by users.

Performance

In theory, adapter trimming with cutadapt is dominated by the time needed to compute alignments, which is O(*nk*), where *n* is the total number of the characters in all reads, and *k* is the sum of the length of the adapters. In practice, other operations, such as reading and parsing the input files, take up more than half of the time. With 35 bp colour-space reads and an adapter of length 18, cutadapt trims approximately 1 million reads per minute (0.06 ms per read) on a single core of a 2.66 GHz Intel Core 2 processor.

Colour-space reads

Cutadapt correctly deals with reads given in SOLiD colour space. When an adapter is found, the adapter and the colour preceding it must be removed, as that colour encodes the transition from the small RNA into the adapter sequence and could otherwise lead to a spurious mismatch during read mapping. Cutadapt can produce output compatible with MAQ [2] and BWA [3]. These tools need FASTQ files in which the colours are not encoded by the digits 0-3, but by the letters ACGT (so-called double encoding), and in which the last primer base (given in nucleotide space) and the first colour are removed.

Usage example

In this simple example, the adapter AACCGG is trimmed from reads in the compressed file in.fastq.gz. The result is written to out.fastq:

```
cutadapt -a AACCGG in.fastq.gz > out.
fastq
```

Other use cases are documented in the README file in the cutadapt distribution, and on the website. Full documentation is available by typing "`cutadapt --help`" on the command line.

Algorithm

In the following, a character is a nucleotide or a colour (encoded by 0-3). The first step in processing a single read is to compute optimal alignments between the read and all given adapters.

Cutadapt computes either 'regular' or slightly modified semi-global alignments. Regular semi-global alignments, also called end-space free alignments [7, Chapter 11.6.4], do not penalise initial or trailing gaps. This allows the two sequences to shift freely relative to each other. When the "-a" parameter is used to provide the sequence of an adapter, the adapter is assumed to be ligated to the 3' end of the molecule, and the behaviour of

cutadapt is therefore to remove the adapter and all characters after it. With regular semi-global alignments, a short, usually random, match that overlaps the beginning of a read would lead to the removal of the entire read. We therefore require that an adapter starts at the beginning or within the read. This is achieved by penalising initial gaps in the read sequence, which is the only modification to regular overlap alignment. See Figure 1, right column.

Regular semi-global alignment is used when the location of the adapter is unknown (assumed when the "-b" parameter is used). Then, if the adapter is found to overlap the beginning of the read, all characters *before* the first non-adapter character are removed. See Figure 1 for an illustration of all possible cases.

After aligning all adapters to the read, the alignment with the greatest number of characters that match between read and adapter is considered to be the best one. Next, the error rate e/l is computed, where e is the number of errors, and l is the length of the matching segment between read and adapter. Finally, if the error rate is below the allowed maximum, the read is trimmed as shown in Figure 1.

Conclusion

Cutadapt solves a small, but important task within sequencing pipelines, especially those for small RNA. It offers an easy-to-use command-line interface. If colour space is to be processed, then cutadapt is the only standalone tool that supports this.

Acknowledgements

Part of this work was funded by Mercator Research Center Ruhr, Germany, grant Pr-2010-0016.

Competing interest statement

None declared

Bibliography

1. Li R, Li Y, Kristiansen K, Wang J (2008) SOAP: short oligonucleotide alignment program. Bioinformatics, 24(5), 713–714.
2. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res., 18(11), 1851–1858.
3. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25(14), 1754–1760.
4. Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet., 16(6), 276–277.
5. Zeschnigk M, Martin M, Betzl G, Kalbe A, Sirsch C, Buiting K, Gross S, Fritzilas E, Frey B, Rahmann S, Horsthemke B (2009). Massive parallel bisulfite sequencing of CG-rich DNA fragments reveals that methylation of many X-chromosomal CpG islands in female blood DNA is incomplete. Hum. Mol. Genet., 18(8), 1439–1448.
6. Schulte JH, Marschall T, Martin M, Rosenstiel P, Mestdagh P, Schlierf S, Thor T, Vandesompele J, Eggert A, Schreiber S, Rahmann S, Schramm A (2010) Deep sequencing reveals differential expression of microRNAs in favorable versus unfavorable neuroblastoma. Nucleic Acids Res., 38(17), 5919–5928.
7. Gusfield D (1997) Algorithms on Strings, Trees and Sequences. Cambridge University Press.