HOP 1.0:
User manual.

Reference-based homopolymer error correction.

Giorgio Gonnella
Stefan Kurtz
Michael Beckstette

Center for Bioinformatics
University of Hamburg
Bundesstrasse 43
20146 Hamburg (Germany)

gonnella@zbh.uni-hamburg.de
beckstette@zbh.uni-hamburg.de
kurtz@zbh.uni-hamburg.de

August 23, 2012

1 Introduction

HOP [2] is a tool for the correction of homopolymer errors using a reference sequence. Homopolymer errors are particularly frequent in Roche 454 and IonTorrent sequencing data sets.

2 Installation

HOP is provided as 64-bit binary distributions for Linux and Mac and as source code. The software is available from `http://zbh.uni-hamburg.de/hop` and will be included in future releases of *GenomeTools*.

2.1 Binary distributions

In the binary distributions, the `bin` directory contains the *GenomeTools* binary (`bin/gt`). The whole *GenomeTools* is compiled into this single binary, including HOP. On some systems (not MacOS) a statically-linked binary is also available under `bin/static/gt`. See also the `README` file.

If desired, it may be possible to install the *GenomeTools* binary system-wide by following the instructions contained in `INSTALL` file.

2.2 Source code distribution

The source code for HOP is written in `C` and it is based on the *GenomeTools* framework [1]: as such, it is designed to be compilable on any POSIX-compliant operative systems.

The compilation can be done using the `Makefile` provided with *GenomeTools*. From the main directory of the sources, the following will compile *GenomeTools*, including HOP, as a 64bit binary:

```
> make 64bit=yes amalgamation=yes cairo=no curses=no
```

In principle it is possible to compile HOP as a 32-bit binary, by leaving the `64bit=yes` flag out, however this is discouraged, if a 64-bit system is available. Furthermore, on some systems or using some particular compilers, it may be necessary to use the `errorcheck=no` option, if errors deriving from compiler warnings arise during the compilation.

2.3 Checking the installation of HOP

After downloading HOP or compiling it from the source code, you will have a single binary named `gt`, comprising HOP and other tools.

HOP can be executed, entering `gt hop` from the command line. Thus if everything went fine, and the *GenomeTools* binary is in the command line path, the following will output an help text on the available options and command line syntax:

```
> gt hop -help
```

2.4 External programs necessary for the HOP pipeline

Besides HOP you will need a read mapping tool. For our evaluation, we used *bwa* (`http://bio-bwa.sourceforge.net/`), either using the aln/samse/sampe pipeline (for shorter reads, such as the IonTorrent reads) or the bwasw tool (for longer reads such as the 454 reads). More information can be obtained in the BWA manual page (`http://bio-bwa.sourceforge.net/bwa.shtml`).

The output of the mapping tool must be, or must be converted into, in SAM/BAM format. Before feeding it into HOP, the output must be sorted by reference coordinates. For this purpose, one can use the *samtools* (`http://samtools.sourceforge.net/`) `sort` tool. For more information, see also the samtools manual page (`http://samtools.sourceforge.net/samtools.shtml`).

3 Preliminary steps

3.1 1. Preparing the mapping using BWA and Samtools

HOP corrects homopolymer errors based on a mapping of the reads to a reference sequence. This must be in SAM/BAM format and sorted by coordinate on the reference.

It is usually quite easy to prepare the mapping as required. Say the reads to be corrected are `f_reads.fastq` and `r_reads.fastq`. Furthermore, the reference for the correction is in a Fasta file `refseq.fas`. Then, in the following example, the `bwa` tool will be used for the mapping in paired end mode and the results will be sorted using the `samtools` (as a side note, some of the commands could be combined using piping for an increased efficiency).

```
> bwa index refseq.fas
> bwa aln -t 4 -o 3 -q 15 -f f_reads.sai refseq.fas f_reads.fastq
> bwa aln -t 4 -o 3 -q 15 -f r_reads.sai refseq.fas r_reads.fastq
> bwa sampe refseq.fas -f map.sam f_reads.sai r_reads.sai f_reads.fastq r_reads.fastq
> samtools view -b -o map.bam -S map.sam
> samtools sort map.sam sorted
```

The results of this pipeline will be in the `sorted.bam` BAM file, which will be used as an input mapping file for HOP.

3.2 2. Preparing the reference sequence

The reference sequence is provided to HOP in GtEncseq format. Converting a sequence in standard sequence formats such as Fasta into GtEncseq is very easy. The following will convert the sequence in `refseq.fas` into a GtEncseq (see also `gt encseq encode -help` for more information):

```
> gt encseq encode refseq.fas
```

If the reference comprises multiple sequences (e.g. contigs, or multiple chromosomes), it is important that the order of the sequences used for the encode step must be the same which is specified in the

header of the SAM/BAM input file. This is usually not a problem, expecially if all reference sequences are contained in a single MultiFasta file, which is both used for the mapping tool and the generation of the GtEncseq.

4 Running HOP

HOP requires the following input:

- the mapping of the reads to the reference (sorted SAM/BAM file)
- the reference sequence (GtEncseq format)
- the uncorrected reads

The user shall choose one of the following correction modes:
`-aggressive`, `-moderate`, `-conservative` or `-expert`. The aggressive, moderate and conservative modes are three presets of different correction thresholds parameters. The aggressive mode will tend to correct more, but will be less precise. The conservative mode is the most precise, although more errors will remain. The expert mode allows to setup the parameters manually. For more information run `gt hop -help`.

The following is an example of correction the reads in `f_reads.fastq` and `r_reads.fastq` using HOP and a reference sequence contained in `refseq.fas`. The `-moderate` option will be used.

```
> gt hop -moderate -ref refseq.fas \
  -map sorted.bam -reads f_reads.fastq r_reads.fastq
```

The output will be in this case in the two files `hop_f_reads.fastq` and `hop_r_reads.fastq`. The prefix added by HOP (in this case `hop`, which is the default) can be changed using the `-outprefix` option.

5 Restricting corrections to coding sequences

If desired, it is possible to restrict corrections to coding sequences only. The idea is that coding sequences tend to be more conserved, thus homopolymer length differences of the reads to the reference are more likely to be errors in corrections. In other words, this will help increasing the precision of the correction (reduce false positives), but of course the overall sensitivity will be reduced, as only errors in coding sequences will be corrected.

Currently, restricting to coding sequences is only available if a single reference sequence is used (that is no collection of sequences, such as contigs or multiple chromosomes). We plan to eliminate this limitation in future versions of HOP.

The annotation must be provided in `GFF` format, version 3. This is a standard annotation format. Furthermore, the annotation must be sorted by coordinate, which can be obtained using the `gt gff3` tool, using the `-sort` option. In the following example, the `refseq.gff` annotation will be used to restrict correction to coding sequences:

```
> gt gff3 -sort -o sorted.gff refseq.gff
> gt hop -moderate -ref refseq.fas -ann sorted.gff \
  -map sorted.bam -reads f_reads.fastq r_reads.fastq
```

The default behaviour when the `-ann` option is used, is to limit correction to homopolymers inside the `CDS` features. However, another feature type can be specified, e.g. `exon`, using the `-ft` option of HOP.

5.1 Troubleshooting

In some cases, the GFF file may be slightly not conforming to the standard or suffer under some problems which disrupts parsing. In these cases, standard POSIX utilities can be sometimes be useful, to obtain a subset of the GFF file, which can be used for HOP. In the following example, using `awk` and `sort`, a sorted subset of the `refseq.gff` annotation is prepared, and saved as `cdsonly.gff` which only contains the `CDS` features:

```
> echo '##gff-version 3' > cdsonly.gff
> awk '$3 == "CDS"' refseq.gff | sort -k4 -n >> cdsonly.gff
```

6 Help from the command line

Using `gt hop -help` will output more information regarding each command line option. Some command line options are used in `-expert` mode only. To output help also regarding these options use `gt hop -help+`.

References

[1] Gremme, G. (2011). The GENOMETOOLS genome analysis system. `http://genometools.org`.

[2] Gonnella G and Kurtz S and Beckstette M (2012). HOP: a reference-based homopolymer error corrector. *Submitted*