HTSeq - a Python framework to work with high-throughput sequencing data

Dragoș Alin Rotaru

University of Bucharest

24 April, 2015

Overview for section 1

Motivation

- Low cost DNA sequencing

Motivation

- Low cost DNA sequencing
- Technologies that can take a snapshot of RNA at a moment (RNA-Seq)

Motivation

- Low cost DNA sequencing
- Technologies that can take a snapshot of RNA at a moment (RNA-Seq)
- Size of human genome is about 3,234.83 Mb (Mega-basepairs)

Motivation

- Low cost DNA sequencing
- Technologies that can take a snapshot of RNA at a moment (RNA-Seq)
- Size of human genome is about 3,234.83 Mb (Mega-basepairs)
- Dealing with HTS(High Throughput Sequencing) data is inevitably

Overview for section 2

Current problems

- Most tools are built only for specific purposes

Current problems

- Most tools are built only for specific purposes
- Part of them: not open-source

Current problems

- Most tools are built only for specific purposes
- Part of them: not open-source
- Part of the others: not suitable for high performance tasks

What is HTSeq?

- Framework for Python
- Written in Cython
- Great and easy for writing customizable scripts
- High level scripts
- Well documented

What is HTSeq?

HTSeq 0.6.1p2 documentation » previous | next | index

Table Of Contents

Previous topic

HTSeq: Analysing high-throughput sequencing data with Python

Next topic

Prequisites and installation

This Page

Show Source

Quick search

Go

Enter search terms or a module, class or function name.

HTSeq: Analysing high-throughput sequencing data with Python

HTSeq is a Python package that provides infrastructure to process data from high-throughput sequencing assays.

- Please see the chapter *A tour through HTSeq* first for an overview on the kind of analysis you can do with HTSeq and the design of the package, and then look at the reference documentation.
- While the main purpose of HTSeq is to allow you to write your own analysis scripts, customized to your needs, there are also a couple of stand-alone scripts for common tasks that can be used without any Python knowledge. See the *Scripts* section in the overview below for what is available.
- For downloads and installation instructions, see *Prequisites and installation*.

Paper

HTSeq is described in the following publication:

Simon Anders, Paul Theodor Pyl, Wolfgang Huber
HTSeq — A Python framework to work with high-throughput sequencing data
Bioinformatics (2014), in print, online at doi:10.1093/bioinformatics/btu638

If you use HTSeq in research, please cite this paper in your publication.

Documentation overview

- *Prequisites and installation*

  Download links and installation instructions can be found here

- *A tour through HTSeq*

  The Tour shows you how to get started. It explains how to install HTSeq, and then demonstrates typical analysis steps with explicit examples. Read this first, and then see the Reference for details.

- *A detailed use case: TSS plots*

  This chapter explains typical usage patterns for HTSeq by explaining in detail three different solutions to the same

HTSeq - Documentation Website

Overview for section 3

Back to biology... - Phred Quality Scores

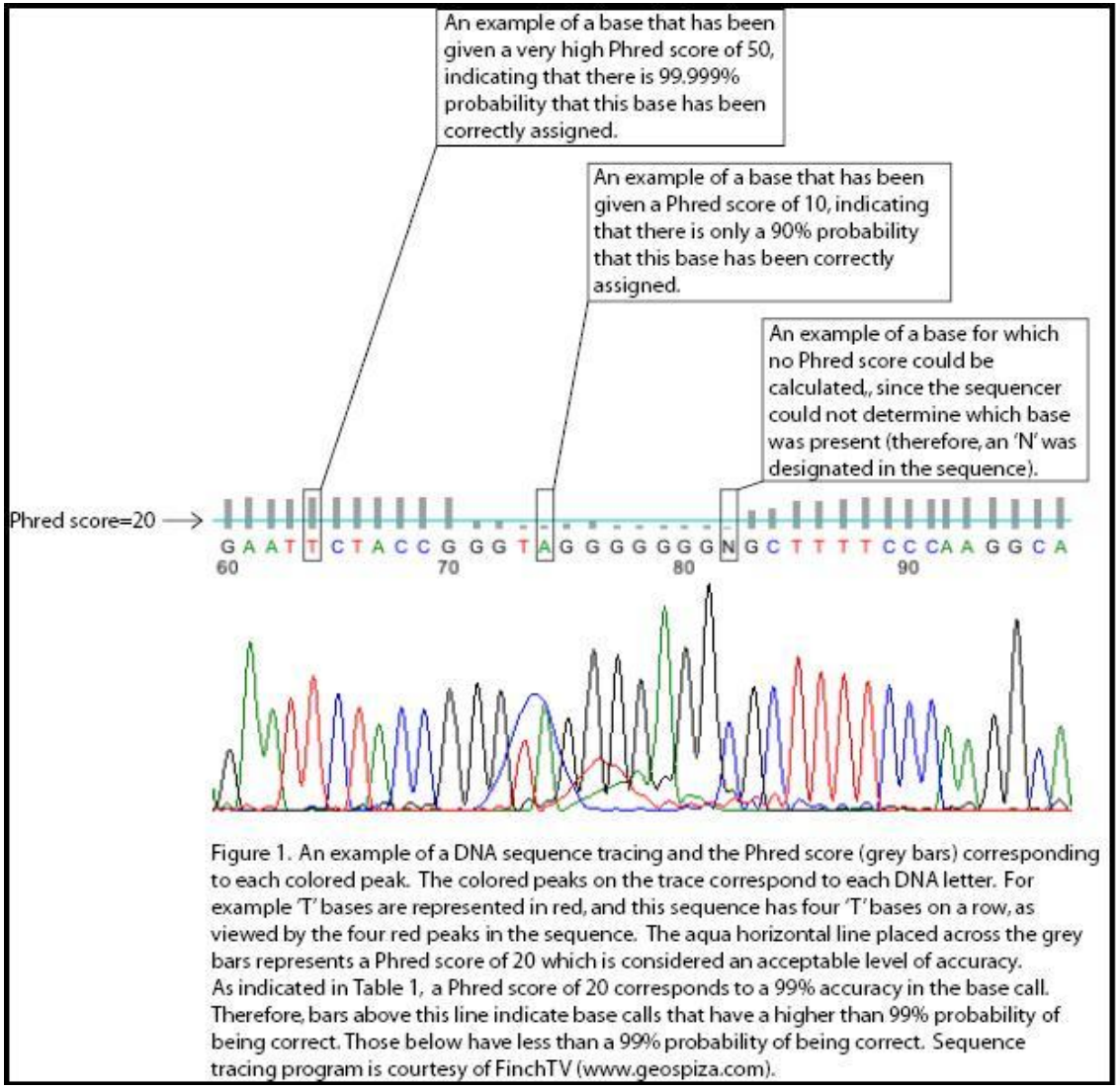


Figure 1. An example of a DNA sequence tracing and the Phred score (grey bars) corresponding to each colored peak. The colored peaks on the trace correspond to each DNA letter. For example 'T' bases are represented in red, and this sequence has four 'T' bases on a row, as viewed by the four red peaks in the sequence. The aqua horizontal line placed across the grey bars represents a Phred score of 20 which is considered an acceptable level of accuracy. As indicated in Table 1, a Phred score of 20 corresponds to a 99% accuracy in the base call. Therefore, bars above this line indicate base calls that have a higher than 99% probability of being correct. Those below have less than a 99% probability of being correct. Sequence tracing program is courtesy of FinchTV (www.geospiza.com).

Phred Quality Scores - Source: Wiki

File formats?

- FASTA files - representing nucleotide (A-C-G-T) or peptide (aminoacid) sequences
- FASTQ files - usually nucleotide sequences, this time with their scores

File formats?

```
@HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
TTAATTGGTAAATAAATCTCCTAATAGCTTAGATNTTACCTTNNNNNNNNNNTAGTTTCTTC
+HWI-EAS209_0006_FC706VJ:5:58:5894:21141#ATCACG/1
efcfffffcfeefffcffffffddf‘feed]‘]_Ba_^__[YBBBBBBBBBBRTT\]][]dc
```

More file formats

- SAM - Sequence Alignments Map

More file formats

- SAM - Sequence Alignments Map
- SAM can be derived from FASTQ files by DNA aligning reads with some reference genome

More file formats

- BAM

More file formats

- BAM
- Binary version of SAM

More file formats

- BAM
- Binary version of SAM
- There are different tools which does the convertion between SAM and BAM or others (samtools)

Now HTSeq; SAM - Aligned; FASTQ - Phred score

- Fortunately, HTSeq have built-in parsers for those files!

Now HTSeq; SAM - Aligned; FASTQ - Phred score

- Fortunately, HTSeq have built-in parsers for those files!
- Can perform quality assessments on SAM and FASTQ directly with htseq-qa command

Now HTSeq - counting

- Given a list of features, genes, exons etc, how many reads map to them?

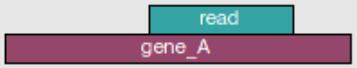
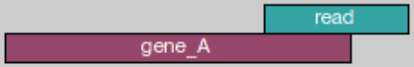
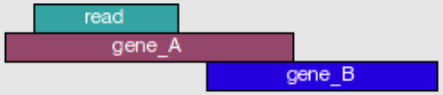
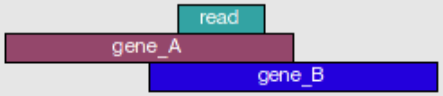
Now HTSeq - counting

- Given a list of features, genes, exons etc, how many reads map to them?
- Can use prebuilt htseq-count command or custom scripts

Now HTSeq - counting

- Given a list of features, genes, exons etc, how many reads map to them?
- Can use prebuilt htseq-count command or custom scripts
- Advanced data structures already implemented such as GenomicInterval or GenomicArrayofSets

Now HTSeq - counting

	union	intersection _strict	intersection _nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

HTSeq count modes - Source: HTSeq WebSite

Now HTSeq - transcription start sites

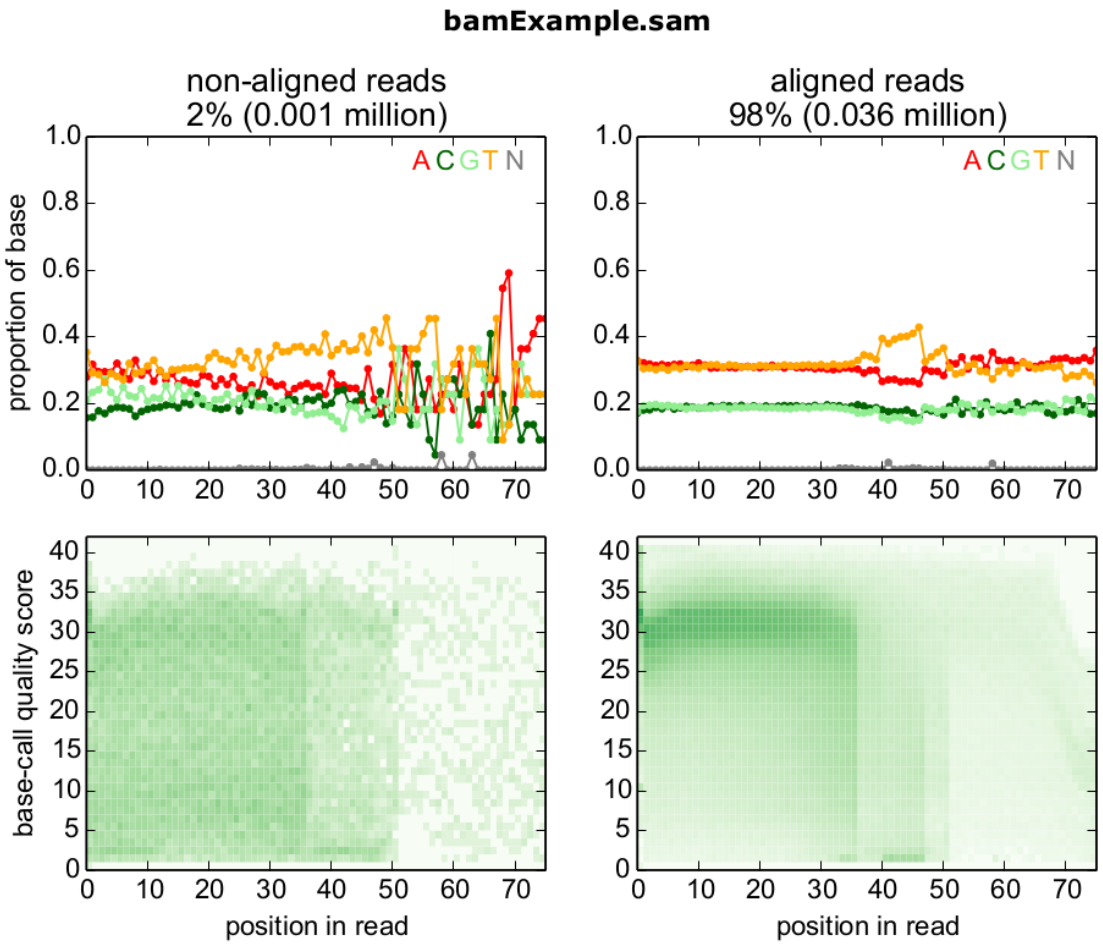
- Compute a 'window' profile from an interval with respect to some features

Overview for section 4

Playing with HTSeq and different files

- Realize that in practice you can't find real SAM files
- BAM files have at least 2GB size
- Samtools doesn't work very well when it comes to converting BAM to SAM
- Managed to get a QA plot

Playing with HTSeq and different files



Toy Example

HTSeq

Live demonstration

Conclusions

- HTSeq is a strong tool for manipulating High-Throughput data

Conclusions

- HTSeq is a strong tool for manipulating High-Throughput data
- Often, theory differs from practice

Thank you!