

# Biopieces: a bioinformatics toolset and framework

**Martin Asser Hansen**  
m.hansen@imb.uq.edu.au

**Chol Hee Jung**  
c.jung@imb.uq.edu.au

**Selene Fernandez Valverde**  
s.fernandezvalverde@imb.uq.edu.au

**Harald Oey**  
h.oey@imb.uq.edu.au

Institute for Molecular Bioscience, University of Queensland, Australia

## Abstract

The Biopieces are a collection of bioinformatics tools that can be pieced together in a very easy and flexible manner to perform both simple and complex tasks. The Biopieces work on a data stream in such a way that the data stream can be passed through several different Biopieces, each performing one specific task: modifying or adding records to the data stream, creating plots, or uploading data to databases and web services. The Biopieces are executed in a command line environment where the data stream is initialized by specific Biopieces which read data from files, databases, or web services, and output records to the data stream that is passed to downstream Biopieces until the data stream is terminated at the end of the analysis as outlined below:

```
read_data | calculate_something | write_results
```

The following example demonstrates how a Solexa deep sequencing experiment can be analyzed – including removal of adaptor sequence, determining the number of unique sequences, mapping to a specified genome, and uploading the data to the UCSC genome browser for further analysis:

```
read_solexa -i data.solexa | - Initialize data stream from a file.
remove_adaptor -a TCGTATGCC -m 2 | - Remove adaptor sequence allowing for 2 mismatches.
grab -e 'ADAPTOR_POS > -1' | - Get all entries where an adaptor sequence was found.
count_vals -k SEQ | - Determine the occurrences of all sequences.
uniq_vals -k SEQ | - Get all entries with a unique sequence.
merge_vals -k SEQ_NAME,SEQ_COUNT | - Append the sequence count to the sequence name.
vmatch_seq -g hg18 | - Map the sequences to the Human genome using Vmatch.
upload_to_ucsc -d hg18 -t solexa_data -x - Upload the mapping results to the UCSC Genome Browser.
```

The advantage of the Biopieces is that a user can easily solve simple and complex tasks without having any programming experience. Moreover, since the data format used to pass data between Biopieces is text based, different developers can quickly create new Biopieces in their favorite programming language - and all the Biopieces will maintain compatibility.

The Biopieces are open source (GPL) and available online at [www.biopieces.org](http://www.biopieces.org)

**Keywords:** Bioinformatics, Tools, Toolset, Framework

**Acknowledgements:** Danish Agency for Science, Technology and Innovation, Grant number 272-06-0325.