**Software paper for submission to MEDSCI 736 University of Auckland**

**(1) Overview**

Title

“ePygenetics: An extraction tool for epigenetic data”.

Paper Authors

1. *Chalmers, Callum*

2. *O’Sullivan, Justin*

*3. Pireva, Kreshnik*

Paper Author Roles and Affiliations

1. Development Team, Faculty of Medical and Health Sciences UoA

2. Product Owner, Liggins Institute UoA

3. Scrum Master, UoA

Abstract

ePygenetics is a python module that retrieves epigenetic data from .wig files. It allows the user to input the data they require from the files, finds the data within the file and outputs the data to a database. This allows geneticists to easily extract the epigenetic status of a single nucleotide polymorphism (SNP) in a particular cell line so this can be considered in any analyses. It is licensed under the MIT license and the source code is freely available for use and modification at: <https://github.com/UOA-MEDSCI-736/CallumChalmers29-crispy-disco>.

Keywords

Software, Python, Genetics, Epigenetics, SNPs

Introduction

ePygenetics is a python module that allows the user to extract epigenetic data from .wig files. The module was created to allow geneticists to quickly and easily access the epigenetic status of single nucleotide polymorphisms (SNPs) from DNAse hypersensitivity data.

Currently, epigenetic data is not considered in most genome-wide association studies, potentially leading to a reduction in the power of this method [1]. While .wig files can be viewed in programmes such as the UCSC Genome Browser, this software is not effective at looking at multiple SNPs at once or comparing multiple cell lines. ePygenetics can perform both these functions by exporting the data as a filterable database which can be easily viewed or integrated into another programme. This allows epigenetic status to be considered when doing genetic analyses, potentially providing the missing piece of the puzzle.

**Implementation and architecture**

ePygenetics was written over a period of 3 months in the Python programming language. The source code can be downloaded from its online GitHub repository. The software reads .wig files, extracts data based on user input and adds the data to a csv database. The software has two main sets of functions: Add a cell line and add a SNP. Add a cell line is how a data file is loaded into the programme. It asks the user to enter to cell line they would like to add and creates a new column in the database for that cell line. The programme then automatically searches the loaded data file for all the SNPs (rows) in the database and populates the database. Add a SNP is how the user tells the programme what data to extract from the files. It asks the user which SNP they would like to add, adds a row to the database for that SNP and then searches that SNP in any files loaded into the programme.

How the software was implemented, with details of the architecture where relevant. Use of relevant diagrams is appropriate. Please also describe any variants and associated implementation differences.

**Quality control**

The software has been functionally tested throughout the development process in the Linux environment. It has also been unit tested for functions which have a clear input and output using the Pytest module. Details of recreating this testing environment can be found in the documentation.

Detail the level of testing that has been carried out on the code (e.g. unit, functional, load etc.), and in which environments. If not already included in the software documentation, provide details of how a user could quickly understand if the software is working (e.g. providing examples of running the software with sample input and output data).

**(2) Availability**

***Operating system***

Ubuntu 16.04 LTS

***Programming language***

Python 3.5.2

***Additional system requirements***

None

***Dependencies***

None

***List of contributors***

None

***Software location:***

***Archive*** To be decided

***Name:*** The name of the archive

***Persistent identifier:*** e.g. DOI, handle, PURL, etc.

***Licence:*** Open license under which the software is licensed

***Publisher:*** Name of the person who deposited the software

***Version published:***The version number of the software archived

***Date published:*** dd/mm/yy

**Code repository**

***Name:*** UOA-MEDSCI-736/CallumChalmers29-crispy-disco

***Identifier:*** <https://github.com/UOA-MEDSCI-736/CallumChalmers29-crispy-disco>

***Licence:*** Attribution-ShareAlike 3.0 New Zealand

***Date published:*** 29/09/2016

**Emulation environment** (if appropriate)

***Name:*** N/A

***Identifier:*** N/A

***Licence:*** N/A

***Date published:*** N/A

***Language***

New Zealand English

**(3) Reuse potential**

The software can be used to read fixed step .wig files with a track type of wiggle\_0, a step of 20 and a span of 20. It cannot be used to read files with a different track type and cannot be used to read variable step .wig files. Altering This should include the use cases for the software, and also details of how the software might be modified or extended (including how contributors should contact you) if appropriate. Also you must include details of what support mechanisms are in place for this software (even if there is no support).

**Acknowledgements**

We would like to acknowledge the whole Medsci 736 teaching staff for helping CC learn the skills necessary to complete this project

**Funding statement**

Vice Chancellors Strategic Development Fund 2015, University of Auckland

**Competing interests**

CC has no competing interests, check other authors

**References**

[1] Hardison, R.C., 2012. Genome-wide epigenetic data facilitate understanding of disease susceptibility association studies. *Journal of Biological Chemistry*, *287*(37), pp.30932-30940. doi: [10.1074/jbc.R112.352427](https://dx.doi.org/10.1074%2Fjbc.R112.352427)

**Copyright Notice**

Authors who publish with this journal agree to the following terms:

Authors retain copyright with the work simultaneously licensed under a [Creative Commons Attribution License](http://creativecommons.org/licenses/by/3.0/) that allows others to share the work with an acknowledgement of the work's authorship and initial publication in this journal.

Authors are able to enter into separate, additional contractual arrangements for the non-exclusive distribution of the journal's published version of the work (e.g., post it to an institutional repository or publish it in a book), with an acknowledgement of its initial publication in this journal.

By submitting this paper you agree to the terms of this Copyright Notice, which will apply to this submission if and when it is published by this journal.