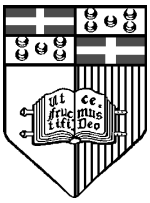
The Malta Human Genome Project

Progress Report

Sara Ann Abdilla (188396M)

Supervisor(s): Jean Paul Ebejer



Faculty of ICT

University of Malta

December 2016

Submitted in partial fulfillment of the requirements for the degree of B.Sc. Computing Science (Hons.)

Contents

Abstract: The abstract should act as a stand-alone (very) brief description of the whole story: The context, the solution, how effective it was found to be. There is no better way to learn how to write an abstract than by carefully reading the abstracts of good papers. This is usually the last part of the report to be written. circa 50 words

1 Introduction & Motivation

Over the years, more and more human reference genomes are being globally assembled. A reference genome is a digital database consisting of DNA sequences; each individual's DNA corresponding to an arrangement of approximately 3 billion bases. The possible nucleotide bases are adenine, cytosine, guanine and thymine - oftenly referred to by the letters A, C, G and T respectively [1, 2].

Genome assembly/sequencing technologies are rapidly advancing and their costs are decreasing; both factors being due to the fact that their importance is becoming more widely known. DNA variations and mutations may correlate to diseases so discovering any approximately matching alignments between the reference genome and the DNA sequencing reads being analysed could very well help future medical diagnosis and treatments [3, 4, 5].

The University of Malta is developing a National Maltese Human Reference Genome for this reason whereby whole genome sequencing on certain Maltese DNA samples (provided by the Malta BioBank) will be performed. The American human genome sequencing facility Complete Genomics, which was founded in 2006, will be a partner in this project.

Thus, computation wise, a data visualisation tool along with a genome browser need to be constructed in order to aid in this endeavour.

2 Reasoning for Nontriviality of Problem

Globally, large genome projects are being sequenced rapidly. Examples of such projects include the 1000 Genomes Project and the International Cancer Genome Project [6, 7, 8]. Malta should aim to be a part of this endevaour so that further studies can be conducted involving a larger variety of genes (i.e. the Maltese genes). After all, while the reading of an individual's DNA shows the likeliness of that person developing a disease, the reading of a nations DNA shows why that population is more likely to develop a disease [3]. Such analyses can therefore prove to not only medically aid global research but also nation-wide studies.

3 Background Research and Literature Review

The development of a genome assembly technology consists of multiple stages; the main ones being genome compression, sequence alignment, alignment visualisation and genome browser construction; all the steps following one another. The following points detail some research which has already been conducted in these areas.

3.1 Genome Compression Tools

3.2 Sequence Aligners

3.3 Read Mapping Algorithms

4 Aims and Objectives

The aim of this project is to build tools for the analysis of sequenced genomes from the Malta Human Genome Project. The research areas studied are bioinformatics, big data, data storage, data visualisation and data analysis among others.

The main objectives of the system are as follows:

1. The alignment of a reference genome against a number of sequencing reads such as that done by Lee et al [9] and Li et al [6] among others;
2. The data visualiation of human genomes;
3. The construction of a genome browser using novel and established components in order to reference DNA for comparitive genomics and to analyse DNA mutations;
4. A comparitive review of existing methods against all the above points.

5 Methods and Techniques Used or Planned

This section details the components which have been implemented along with a description of future plans.

5.1 Genome Compression

5.2 Sequence Alignment

5.3 Alignment Visualisation

5.4 Planned Methods

6 Proposed Evaluation Stategy

The main evaluation techniques proposed are as follows:

1. Comparitive review of the developed human genome visualisation tool with existing methods;
2. Reference DNA and analyse DNA mutations using the constructed genome browser.

7 Expected Deliverables

1. The Final Year Project (FYP) which will include: all the relevant background information required to understand said project, a detailed explanation of the system and the evaluation results;
2. The implementation of the designed and developed system;
3. The documentation which will explain to the users how the system should be employed.

   The milestone schedule is represented in the following Gantt chart:

References

[1] L. Hunter, *Artificial Intelligence and Molecular Biology*. AAAI Press, 445 Burgess Drive, Menlo Park, California 94025, USA: MIT Press, 1993.

[2] A. M. Lesk, *Introduction to Genomics*. Great Claredon Street, Oxford, OX2 6DP, UK: Oxford University Press, 2 ed., 2012.

[3] "The hidden history of the maltese genome," *Think Magazine*, vol. 16, pp. 19–25, 4 2016.

[4] "Blood, genes, you," *Think Magazine*, vol. 16, pp. 26–31, 4 2016.

[5] "Heartbreakers," *Think Magazine*, vol. 16, pp. 32–37, 4 2016.

[6] H. Li and R. Durbin, "Fast and accurate short read alignment with burrowswheeler transform," tech. rep., Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK, 5 2009.

[7] M. H.-Y. Fritz, R. Leinonen, G. Cochrane, and E. Birney, "Efficient storage of high throughput dna sequencing data using reference-based compression," tech. rep., United Kingdom, 1 2011.

[8] L. Huang, V. Popic, and S. Batzoglou, "Short read alignment with populations of genomes," tech. rep., 2013.

[9] D. Lee, F. Hormozdiari, H. Xin, F. Hach, O. Mutlu, and C. Alkan, "Fast and accurate mapping of complete genomics reads," tech. rep., 10 2014.

[10] B. Berger, J. Peng, and M. Singh, "Computational solutions for omics data," tech. rep., USA, 3 2014.

[11] X. Chen, S. Kwong, and M. Li, "A compression algorithm for dna sequences and its applications in genome comparison," tech. rep., New York, NY, USA, 4 2000.

[12] M. Ruffalo, T. LaFramboise, and M. Koyutrk, "Comparative analysis of algorithms for next-generation sequencing read alignment," tech. rep., 7 2011.

[13] J. Shang, F. Zhu, W. Vongsangnak, Y. Tang, W. Zhang, , and B. Shen, "Evaluation and comparison of multiple aligners for next-generation sequencing data analysis," tech. rep., 3 20114.

[14] A. M. Lesk, *Introduction to Bioinformatics*. Great Claredon Street, Oxford, OX2 6DP, UK: Oxford University Press, 4 ed., 2014.

[15] D. W. Mount, *Bioinformatics: Sequence and Genome Analysis*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, USA: Cold Spring Harbor Laboratory Press, 2001.