Bioinformatics for large-scale tumour sequence analysis

Hongye Wang

Imperial College London

10/12/2019

Supervisor: Fumagalli Matteo, Department of Life Sciences (Silwood Park), Imperial College London

Objectives

To identify and characterise CNV (copy number variation) in tumor cells.

Introduction

Ploidy is the number of complete sets of chromosomes in a cell (1). Individual organisms can be described according to the number of sets of chromosomes present (the "ploidy level"), like monoploid (1 set), diploid (2 sets), and cells which have three or more chromosome set are often described as polyploid. Humans are diploid organisms, carrying two complete sets of chromosomes in their somatic cells (2). However, in some tumor cells, the ploidy usually increased because of the copy number variation. Therefore, we want to estimate the ploidy of tumor cells based on next generation sequencing (NGS) data and compare with the normal somatic cells.

Inferring ploidy levels is one of the most important jobs of our research. However, the current approaches can not meet our need because most of them are based on the frequency and depth of genome, they do not account for genotype uncertainty as well, that make them unreliable on low- and mid-depth sequencing data (2). Instead, we choose HMMploidy as the main tool. Due to the combination of sequencing depth and genotype likelihoods, the effectiveness of HMMploidy is boosted. The great power has been shown at very low coverage in both simulated and real data.

The NGS data which I will use to analyse is collected from the patients affected by sarcoma.

keywords: NGS data, Ploidy, Tumour gene sequence, CNV, HMM

Methods

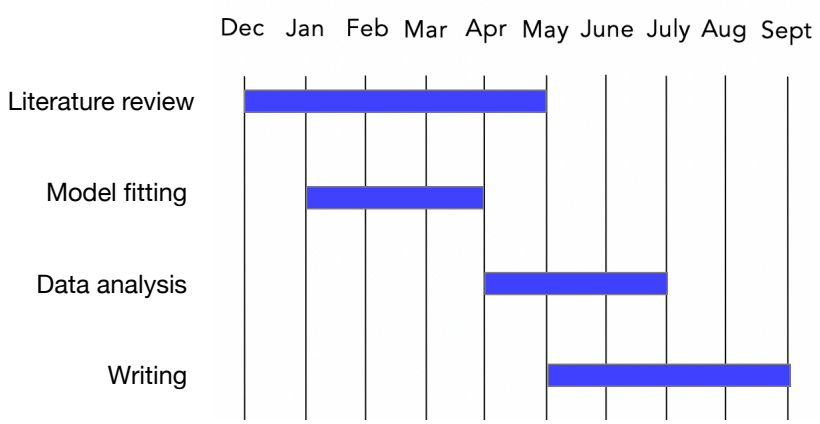
The main method is HMMploidy. The idea is to apply this method and related software to the tumor patients' data. The open source software is available at https://github.com/SamueleSoraggi/HMMploidy. It accepts input files in mpileup format. The framework consists of a python and R scripts (including compiled Rcpp code) (3). Each script can be run with a single-line command from any UNIX shell.

Budget

The budget allocated towards this Master project is 500 pounds. Currently there are no expenses, however future expenses may be towards visiting and learning from interrelated research institution,

attending conference and lab equipment.

Schedule



Supervisor Statement

I have seen and approved the proposal and the budget.

Supervisor: Fumagalli Matteo

Signature: Fumagalli Matteo

Date: 10/12/19

References

[1] Ankit Malhotra, Yong Wang, Jill Waters, Ken Chen, Funda Meric-Bernstam, Ira M Hall, and Nicholas E Navin. Ploidy-seq: inferring mutational chronology by sequencing polyploid tumor subpopulations. *Genome medicine*, 7(1):6, 2015.

[2] Gabriel RA Margarido and David Heckerman. Conpade: genome assembly ploidy estimation from next-generation sequencing data. *PLoS computational biology*, 11(4):e1004229, 2015.

[3] Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.