Louis Gil Acevedo

louis.gil@upr.edu

Prof. Humberto Ortiz

humberto.ortiz@upr.edu

Computer Science Department - Rio Piedras

Universidad de Puerto Rico, Recinto de Rio Piedras

May, 2019

Cancer Detection from Kmers

INTRODUCTION

Differential expression techniques have been used to identify genes related to abnormalities in the expression of genes when compared to a normal control sample. Using these techniques many diseases such as cancer have been genotypically described in order to understand how the disease works in a network of genes. In this study, we aim to observe if differential expression at a raw sequence level in order to see if we can identify variants related to cancer at a gene level. Effectively detecting if a tumor is cancerous without having to assemble the reads, making the process faster. First, we observe if there are any kmers present in a cancerous tumor compared with their germline tissue as a control. Our data is derived from the Personal Genome Project, which aims to create an open source accessible genomic and medical records of patient data available. Here we used FastQ sample from Jay Lake a science fiction author that made his whole genome publicly available (https://my.pgp-hms.org/profile/huDCD45D) he was diagnosed and passed away from cancer. As his last stand against the disease, he decided his data should be publicly available in order to help students and researchers around the world become smarter about human health and life in general. (http://www.jlake.com/2013/12/31/cancerscience-my-entire-genome-has-gone-open-source/)

METHOD

We acquired or data from the Personal genome project. We then cleaned the data using Trimmomatic and used Sanquishe's version of diffhash to calculate the number of genes present in each sample. Given the large size of the files and having diffhash at its early stages it was not able to analyze these files with only one core. Given this challenge, we subsampled the data by grabbing the first four million lines and running diffhash on these subsamples. Because we only have one sample of each condition the R script that worked with the statistics was not going to yield any interesting results without more samples. But we received a file containing the kmer and the amount of times found in a file called hashcounts. From this, we extracted wich kmers where only present in the tumor saple and then summed one to every sample in order to calculate the fold change and filter for over positive one or under negative one.

PRELIMINARY RESULTS

We were able to identify 8,977 kmers only found in the tumor subsample and the following genes annotated as upregulated for having a fold change greater or equal to positive one, and downregulated genes with fold change lower or equal to negative one.

Upregulated Genes	Downregulated Genes	Total Significant Fold Changes
16,816	13,132	29,948

CONCLUSION

We were able to find kmers only present in the tumor subsample, this may give a good indication that there could be kmers that identify and describe the disease. Throughout the semester I also had the opportunity to speak with Titus Brown and he confirmed our suspicion that variants could be identified from kmers. His lab published a paper where they identify kmers related to autism, they did

this by subtracting any kmers that appeared in the human genome reference and the patient's parents in order to only have novel kmers. This study confirmed our intuition that it is possible to identify variants through kmers yet cancer is a very difficult disease where many different mutations may cause the same outcome, also there are other mutations other that insertion and deletion that may make identify these kmers harder in cancer.

FUTURE WORK

- Use diff-hash with multiple samples of tumors and their corresponding germline normal tissue.
- Test in a different disease that is easier to identify, with small well-known mutations.
- Make diffhash multi-threaded, in order to analyze big files such as the ones we attempted to use.

REFERENCES

Cmero, M., Davidson, N., & Oshlack, A. (2018). Fast and accurate differential transcript usage by testing equivalence class counts: Supplementary figures and tables. doi:10.1101/501106

Standage, D. S., Brown, C. T., & Hormozdiari, F. (2019). Kevlar: A mapping-free framework for accurate discovery of de novo variants. doi:10.1101/549154

Gonick, L., & Wheelis, M. (2005). *The cartoon guide to genetics*. New York, NY: Collins Reference, an imprint of HarperCollins.