Elizabeth Helton

Professor Aleshunas

MATH 3220

Comparative Genomics Find My Friends Analysis

Problem Description:

Comparative genomics is a highly used field of research that allows for a comparison analysis of genome sequences from different species. Being able to compare like species to different types of organisms allows researchers to identify similarities and differences that may be used for discovering new ways to battle human diseases. The problem with working genomes and an infinite amount of organism is that it quickly becomes a lot of material to work with. This is where Bioconductor and packages such as 'Find My Friends' are used as a possible answer to help speed up the comparison. Since 'Find My Friends' is a relatively new package it will have some faults. To properly look at how the package may respond to different variations of data, I used Bacteriophages and ran them as one gene. I compared this result to the example work space that the author Thomas Pedersen created. In this research project, the goal to this research is to identify if 'Find My Friends' is an effective enough package that will compare genomes from various species to look for similarity.

Background:

Bioconductor is an open source and developmental software program that provide tools and comprehensive genomic data (Bioconductor 2003). With it being open sourced it allows researchers to create reproducible research that can be used with analyzing biological data. It is also primarily used in R programming language. Bioconductor uses packages to solve various biological issues. The main function of Bioconductor packages is to provide an analysis of DNA microarray, sequence flow, SNP and other data (Bioconductor 2003). This use of analysis, comprehension and visual aid of genomic data is beneficial for working with comparative genomics.

Comparative genomics is a form of biological research that compares complete genome sequences of various species of organisms (NIH 2015). This can be done using various tools as well as computer based analysis. Through comparing the characteristics of the organism's researchers can highlight regions of similarity and differences in the genome. This comparison or similarity can reveal what the function and structure of human genes are like and be able to develop new strategies to fight human diseases (NIH 2015). There are various benefits that can come out of comparing genomics, mainly centered around evolution. Being able to identify DNA sequences that have been preserved in a variety of organisms over millennia can help researchers identify genes that a necessary for life, as well as focuses on genomic signals that can control gene function throughout various species. It also helps identify which genes relate to certain biological systems which could then ultimately help form a new way to treat human diseases and improve health (NIH 2015). This conservation of genes can also help evolution researchers by identifying evolutionary relationships between various species and identifying their differences in DNA. By doing this it can possibly be used to formulate hypothesis on appearance, behavior

or biology of organisms and how they differ over time. Over time DNA sequencing has become more well-known and less expensive allowing for comparative genomics to branch out into new areas such as agriculture, biotechnology and zoology. Using comparative genomics has allowed evolutionists to create new branches on the evolutionary tree, as well as helping the veterinarian field on finding ways to improve the health of domesticated animals. Comparative genomics also highlights new strategies that may be beneficial for conserving rare or endanger species of animals.

When using Rstudio there are various ways to view comparative genomics, but one new way that doesn't use an algorithm is the package Find My Friends. Most comparative genomic packages use pangenomes as a way to look for similarity among various species. So, what is a pangenome? A pangenome is a collection of all the genes of related organisms that are grouped together in some form of homology. Pangenomes can be used as organizational tools, as well as a way to collect genomic information within a group of organisms (Pedersen 2017). Find My Friends uses a unique strategy of using alignment free sequence comparisons based on the decomposition of sequences into K-mer vectors (Pedersen 2017). K-mer vectors are words that are equal in length. This comes from creating a 'window' of size K over the sequences and the K-mer vector value is the count of each unique word in the window. Using this decomposition strategy, the sequence similarity problem then turns into a vector similarity problem. The one that is used by this package is cosine similarity, which uses the cosine of the angle between two vectors. If the vector is identified as stringently positive, then that value is set into either being a 0 or 1 with 1 having 100% similarity. Find My Friends also uses another way of grouping that is different from K-mers. In most of the other comparative genome packages the final grouping of similarities is derived straight from the main similarity group. But in Find my Friends approach this main similarity group is kept broad to contain large gene groups. These groups are then looked at for finer details in regarding to sequence similarity, neighborhood similarity and sequence length (Pedersen 2017). By doing this the package can quickly observe gene pairs for similarity, which finally results in a linear scaling.

Example of K-mers:

GATTCGATTAG -> ATT: 2

CGA: 1

GAT: 2

TAG: 1

TCG: 1

TTA: 1

TTC: 1

When viewing the results of Find My Friends, Bioconductor is very helpful because it easily can set the pangenome matrix as an Expression Set Object or just as a regular matrix to use. But Find my Friends also incorporates a few additional tools to look at the results even more. It uses 2 calculations that can calculate statistics on the two main gene groupings in the dataset; generally, gene groups and organisms (Pedersen 2017). These two functions are called

groupStat and orgStat(Pedersen 2017). The final results can also be viewed in various types of graphs for visual interpretation of organization.

Methodology

In this experiment I used a package that is found in Bioconductor. This package is called Find My Friends. Find My Friends creates pangenomes of genomes from various organisms, this then looks at the similarity and differences between the genes. The first step that was taken was to figure out what organisms would run well through the package, and based off the sample done by the author and a suggestion from one of my professors, I chose to go with Bacteriophages. Bacteriophages are a small type of virus that lives on other bacteria as host cells, in the end it destroys the host cell. I gathered my bacteriophage genomes from a dataset called 'Actinobacteriophage Database'. The package requires that the genomes be in a fasta file, so I selected the genome that had this listed. Looking through the database I chose bacteriophages that infect the host genus called Mycobacterium. The sample that is done by the author Thomas Pedersen used mycoplasma bacteria, so I thought this data might be similar. Samples that were selected were discovered in different years and places, incase location had anything to do with the gene variation. Next, I formulated a dataset using 10 bacteriophage genomes and grouped them together in a folder on my computer. I then ran this folder into Rstudio and created a pangenome using the pangenome function in FindMyFriends. Once I created my genome parameters were added in to help with the calculation of the pangenome. The calculation part is where this package differs from the rest. It combines gene groups based on lower similarity thresholds, and the largest member of each gene group becomes the representative of the group. Next paralogs were looked at to ensure that genes from the same genome were not grouped together. Once this was done, our results were calculated either by matrix or graphs. The main graphs that were looked at for the bacteriophage dataset was a plotStat, plot Evolution, Kmer similarity graph, and a dendrogram graph.

Assumptions:

Find My Friends will have more errors in identifying similarity of expressivity due to it being a relatively new package. There will also be problems with the evolution plot because it will be biased towards organisms. It was also assumed that the bacteriophage datasets that were given would already have genes identified.

Experimental Design:

The first step that was taken in this research project was to find a dataset that I could run through the package. Since the package works best with microbial organisms and the sample that the Thomas Pedersen did was a bacteria called *Mycoplasma pneumonia* and *hyopneumonia*(Pedersen 2017*).* The organisms that I chose to go with were bacteriophages, this is because bacteriophages are a smaller version of bacteria. Bacteriophages rely on bacteria as host cells to survive. My dataset consisted of 10 bacteriophages from the Mycobacterium host genus. The bacteriophages all varied in discovery time and location. The 10 bacteriophages that I used where; Bobby, Cjw1, Dori, Giles, Kalah2, Lilbit, Petra 64142, ShereKhan, Spongebob, and Webster2. Bobby and Khala2 were both discovered at Webster U. After the organisms were selected, a dataset had to be created. This took a while to figure out how to do based off of the notes that were left by the creator. In order to create the dataset, you had to create a file somewhere in your working directory in Rstudio and put all of the fasta files of the genomes in.

Once this file was created, it had to be loaded into Rstudio so that the pangenome could be created. To create the pangenome you had to use the pangenome function that is found in the package. One thing that was quickly noticed about this and my dataset is that my organisms were accounted for only having one gene per organism. This ended up being because that fasta files that I used kept the genome in base pairs and didn't identify the genes. I left the pangenome how it was with each organism reading as only having one gene to see how the package would be able to analyze the data. After the pangenome is created the author adds additional information about the organisms on the pangenome using a different package. This package comes from CRAN and links with NCBI databases to gain other outside information. In the case of the bacteriophage dataset, this was not able to run. The next step that is used is the addition of parameters, such as; groupPrefix, coreThreshold, kmerSize, lowerLimit, and flank size (Pedersen 2017). These parameters help with the calculation of pangenomes. Calculating the pangenome is used by a function called cdhitGrouping(Pedersen 2017). This function works by repeatedly coming genes from lower similarity thresholds. During each step the longest gene is isolated as the model for the repeated combining. It is ideal to set the threshold to the lowest number possible so that genes coming from the same group can be clustered together (Pedersen 2017). The grouping function is not the only function that is needed for calculating, the other function is a neighborhoodSplit or a kmerSplit. These functions go into each gene group and compares the members based on their sequence similarity, chromosomal neighborhood similarity, sequence length and genome membership (Pedersen 2017). Using the dataset bacteriophages did not allow for a neighborhood split in genes, so I had to use a kmerSplit. This caused a kmerSplit on the pangenome to create 5gene groups. Post-processing part of this package consist of paralogue linking. Paralogue linking causes a link to be created based on groups that have similarity. It can be created by calculating a Kmer similarity between those representatives of each gene group. Thomas Pedersen then goes to show that it is ideal to remove genes that are no longer functioning correctly due to a frameshift. Removing inactive genes can help improve a general analysis. I chose not to remove any genes since I was working with a small dataset of genes. Once the post-processing is done the investigation of results begin, there are various forms of investigation depending on what is looked at. I chose to look at the pangenome matrix as a ExpressionSet object, a groupStat with plotStat graph, Evolution plot, and a Kmer similarity plot. Comparing these to Thomas Pedersen's work show that the package was still able to work with the organisms being labeled as only having one gene.

Results and Discussions:

After having worked with this package I was able to conclude that this relatively new package is a pretty useful package in terms of comparative genomics. Comparing the results created from my pangenome with that of Thomas Pedersen's work, I was able to see some similar results as well as drastic differences. Figure 1 and 2 shows the pangenome that was created. **Figure 1** 'mypang' is the pangenome that I created using the 10 organisms. Five gene groups were defined after being ran through cdhitGrouping. All of the genes were forced into accessory and singleton genes with no core. **Figure 2** is Thomas Pedersen's work showing that he created a pangenome using 9 organisms and from the genes there was 3141 gene groups created. But also looking at this figure we see that all of the genes groups are either accessory or singleton, which is interesting. **Figure 3** shows a plotstat graph of the bacteriophages for visual representation of the ratio of accessory to singleton gene groups, as well as the number of genes found in each organism. **Figure 4** shows the plotstat graph of Thomas Pedersen's work with

mycoplasma bacterium. **Figure 5** is the evolution plot for my pangenome that looks at the number organisms compared to number of gene groups in terms of singleton, accessory, core, and total number of gene groups. This shows that the core gene groups disappeared rather quickly after calculation. Whereas accessory had a steady incline and singleton had a bit of variation around 1-2 gene groups. **Figure 6** shows that Thomas Pedersen had a slower decline in core gene groups as apposed to mine. The singleton declined in number as more organisms were added. One main similarity between **Figure 5** and **Figure 6** is the steady increase in accessory. And finally, I looked at the Kmer similarity graphs. **Figure 7** is a Kmer heat plot based off my pangenome I created. The darker the image the closer in similarity it has, looking at **Figure 7** there are 2 main sets of similarities. These 2 sets are Giles and Webster2, as well as Bobby and Kalah2. **Figure 8** shows the similarity as well with looking at the various organisms that were used. These results show that further research would be needed on this topic to identify why in both cases the core values disappear so fast. Even though this package is relatively new, it provides a different way of creating comparative genomics and has works at a quick rate. But there are always downsides to working with new items or packages such as 'Find My Friends'.

Figure 1

```
> mypang
An object of class pgFull

The pangenome consists of 10 genes from 10 organisms
5 gene groups defined

    Core|
Accessory|=======================================
Singleton|==========

Genes are translated
```

Figure 2:

```
mycoPan
## An object of class pgFullLoc
##
## The pangenome consists of 12247 genes from 9 organisms
## 3141 gene groups defined
##     Core|
##Accessory|=========================================:
## Singleton|======
## Genes are translated
```

Figure 3:

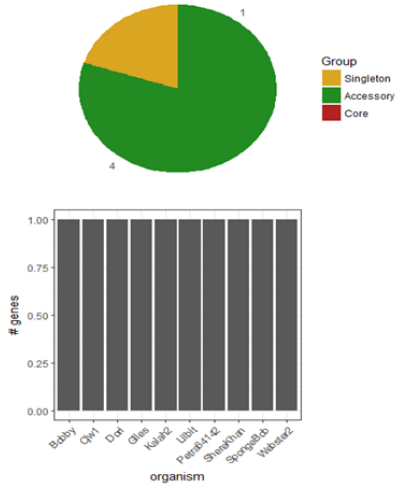


Figure 4:

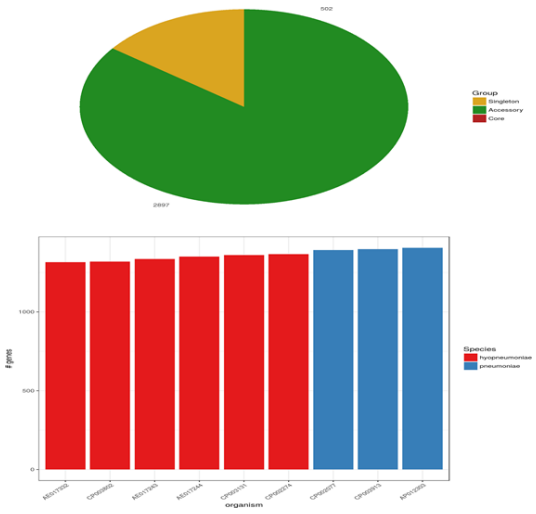


Figure 5:

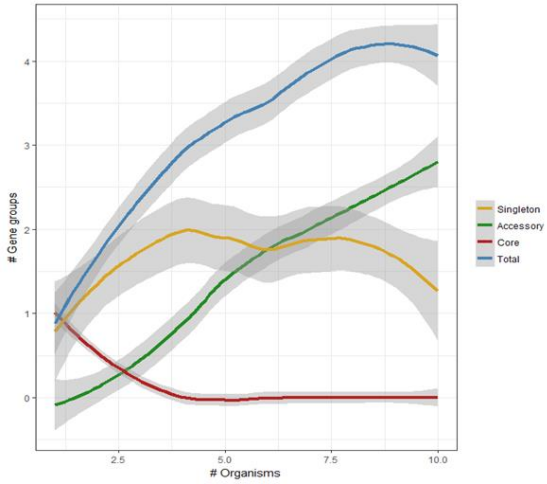


Figure 6:

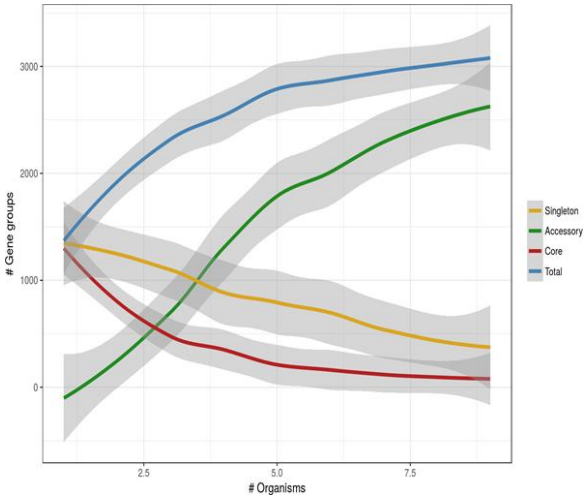


Figure 7:

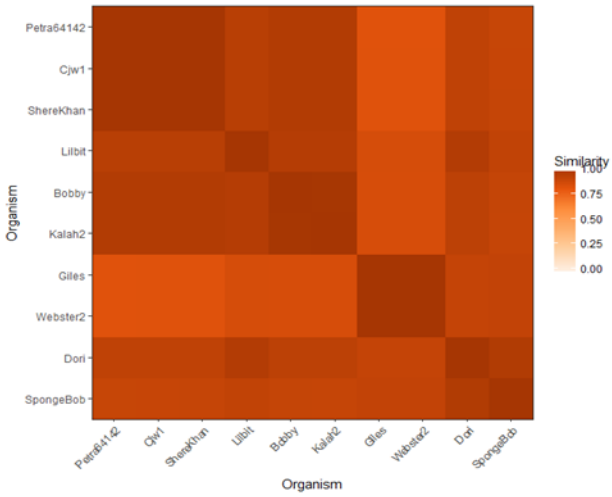
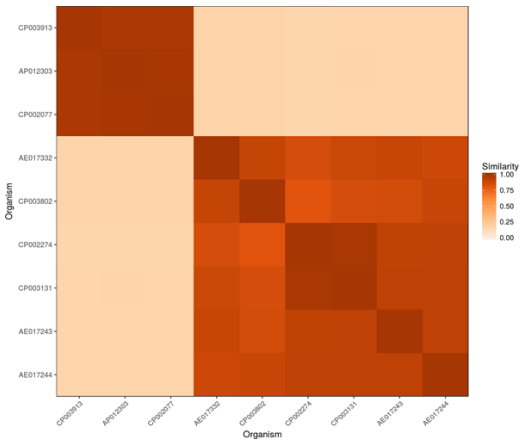


Figure 8:



Issues:

Many issues were ran into in this project, the first one being identifying how to create a dataset and load it into the package. Once that was figured out several errors kept coming up when I was trying to run the code. It turned out that the genomes I loaded from my dataset didn't have the genes already labeled it just counted the entire genome with base pairs as one gene. This caused a different path to be created to see how well the package would continue working with these 10 organisms if they only have one gene. Extra data that was applied to the example work space that the author created was not applicable to the dataset I used because of the difference in organisms; bacteria and bacteriophages. The last issue that I ran into is that I was not able to completely run through the sample code using my dataset because of those errors I previously encountered. This caused for some vital features to not be looked at such as neighborhoods.

Conclusions and Further Work:

Based on the research that was conducted this package is a pretty good resource for creating comparative genomics. With it being roughly two months old, it still will need more research and a better outline for being able to reproduce the authors work. Calculating the genome had a pretty fast response rate of 3minutes. This fast response will be beneficial when working with many different species and an infinite number of genes. Even though I only worked with 10 organisms that were classified as having only 1 gene per organism, the package was able to still run and get drastic results compared to the authors work.

Some further work that could be done would be to repeat this project but instead of keeping the bacteriophages as only having one genome, find a way to cluster the individual genes together found in the database. Once all of the genes have been clustered for each organism, then repeat this process to see how well the package is able to run through the data. Something else that could be used as further research is to run a given set of data through Micro Array package limma and Affymetrix to look at expressivity in the genes and genome. After running the data through microarrays, to then turn this data and run it through Find my Friends and compare the similarity and emphasis on the genes. Lastly, you could also do a comparison project of running data through Find My Friends and compare it to results you get back from running data through other comparative genomic packages such as Roary that use strict algorithms.

References:

Bioconductor. (2003). Bioconductor. Retrieved November 11, 2017, from http://bioconductor.org/

NIH. (2015, November 3). Comparative Genomics Fact Sheet. Retrieved November 11, 2017, from https://www.genome.gov/11509542/comparative-genomics-fact-sheet/comparative-genomics-fact-sheet/

Pedersen, T. L. (2017, October 30). Creating pangenomes using FindMyFriends. Retrieved November 2, 2017, from https://www.bioconductor.org/packages/devel/bioc/vignettes/FindMyFriends/inst/doc/FindMyFriends_intro.html