Next Generation Sequencing Pipeline Development and Data Analysis

Julian Pierre[1], Jordan Taylor[2], Amit Upadhyay[3], Bhanu Rekepalli[3]

1 Texas Southern University, 2 Austin Peay State University, 3 University of Tennessee

ABSTRACT

DNA sequencing is one of the most important platforms for study in biological systems today. The High-Throughput-next generation sequencing technologies delivers fast, inexpensive, and accurate genome information. Next Gen Sequencing can produce over 100 times more data than methods based on Sanger Sequencing. The next gen sequencing technologies offered from Illumina /Solexa, ABI/SOLiD, 454/Roche, and Helicos has provided unprecedented opportunity for high –throughput functional genomic research. Next gen sequence technology offer novel and rapid ways for genome-wide characterization and profiling of mRNA's, transcription factor regions, and DNA patterns.

Next Generation Sequencing

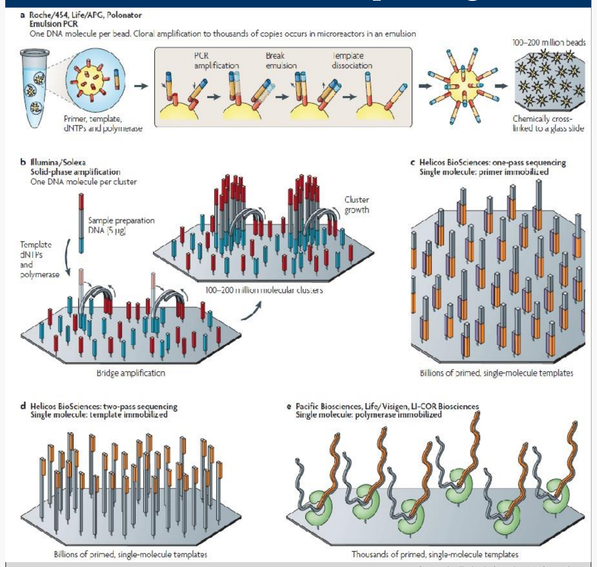

Fig. 1) This is a figure comparing two different methods of Next Generation Sequencing. [2]

Sequencing Technologies – the Next Generation, Micahel L. Metzkerh

- 454/Roche – 454 Life Sciences is a Biotechnology company that is a part of Roche and based in Branford, Connecticut. The center develops ultra-fast high-throughput DNA sequencing methods and tools.
- Illumina/Solexa– Illumina is a company that develops and manufactures integrated systems for the analysis of gene variation. Solexa was founded to develop genome sequencing technology.
- ABI/SOLiD - (Sequencing by Oligonucleotide Ligation and Detection) is a next-generation DNA sequencing technology developed by Life Technologies and has been commercially available since 2006. This next generation technology generates hundreds of millions to billions of small sequence reads at one time.
- Helicos - Helicos's technology images the extension of individual DNA molecules using a defined primer and individual fluorescently labeled nucleotides, which contain a "virtual terminator" preventing incorporation of multiple nucleotides per cycle.

Data Analysis

In Next-Generation Sequencing, data analysis is one of the most expensive processes. While the cost of genome sequencing goes down, the cost of analyzing data is still expensive. In the future, the "$1,000 genome will come with a $20,000 analysis price tag."

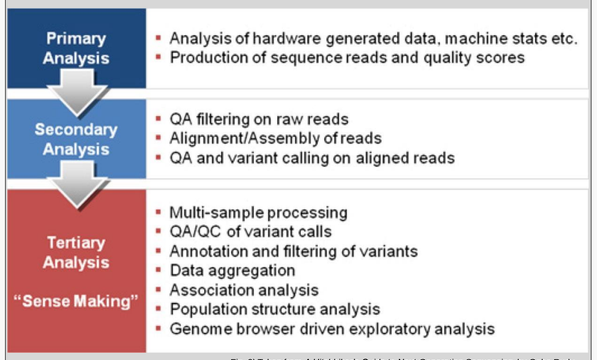

Fig. 2) Taken from *A Hitchhiker's Guide to Next-Generation Sequencing*, by Gabe Rudy

Sequence Analysis refers to the process of subjecting a DNA, RNA or peptide sequence to a wide range of analytical methods to:

- Compare sequences to find similarities and infer if they are Homologous
- To identify the features of the sequence such as gene structure, distribution, introns and exons, and regulation of gene expression
- Identify Sequence differences and variations such as mutations

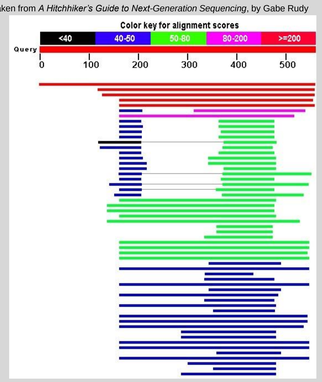

Fig. 3 Taken from bio.davidson.edu/courses. Shows alignment results for yeast.

Genome Assembly

Galaxy is an open, web-based platform for data intensive biomedical research. It can be used on its own free public server where you can perform, reproduce, and share complete analyses.

An example of how Galaxy reflects its data is shown in Fig 5.

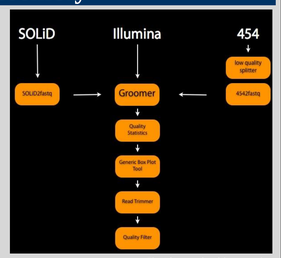

Fig 4 from main.g2.bx.psu.edu

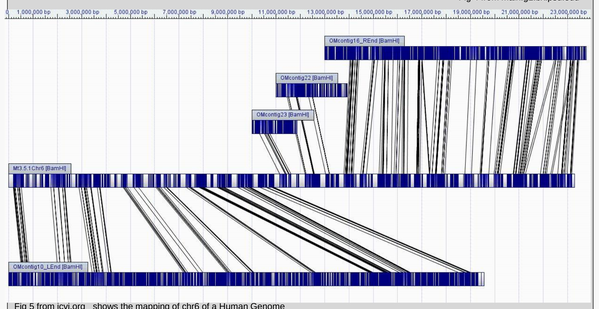

Fig 5 from jcvi.org shows the mapping of chr6 of a Human Genome

Downstream Analysis

Next Gen Sequencing uses a wide array of tools to obtain results based on the genome sequence. The most widely used Tools are BLAST, HMMER, and MUMmer.

- BLAST (Basic Local Alignment Search Tool) is a multi-sequence alignment tool developed by NIH (National Institute of Health). It is used find similar regions in different sequences and then compare their similarities.
- MUMmer is a rapid alignment system used for rapidly aligning entire genomes. It can also align incomplete genomes and can easily handle 1000's of contigs from a shotgun sequencing project.
- HMMER is used for searching sequence databases for homologs of protein sequences, and for making protein sequence alignments. It implements methods using probabilistic models called profile hidden Markov models (HMMs)

This is a small sample from our contig database in a FASTA format (.fna). This file contains all of the contig Id's, the length of the contig sequence, the number of reads, and the sequence itself.

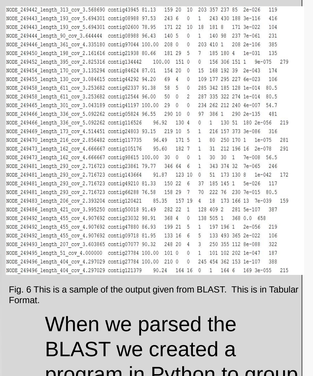

Fig. 6 This is a sample of the output given from BLAST. This is in Tabular Format.

Parsed Results (BLAST)

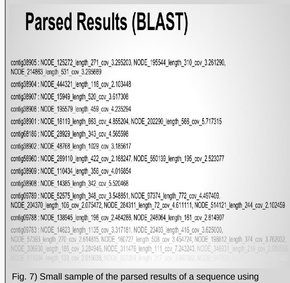

Fig. 7) Small sample of the parsed results of a sequence using BLAST. This shows how many NODE's are matched with each contig.

When we parsed the BLAST we created a program in Python to group all the nodes related to a specific Contig ID. This grouping is helpful as some Contigs have multiple nodes. Making it easier to search for the desired results

MUMmer is a rapid alignment system used for rapidly aligning entire genomes. It can also align incomplete genomes and can easily handle 1000's of contigs from a shotgun sequencing project.

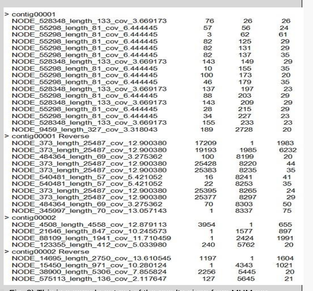

Fig. 8) This is a sample out put of the results given from MUMmer. These are in stdout format.

Parsed Results (MUMmer)

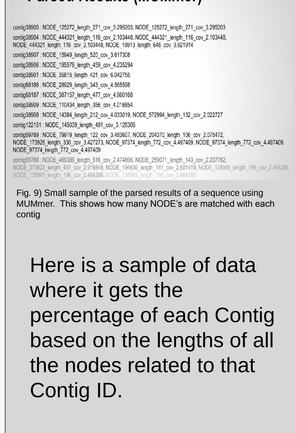

Fig. 9) Small sample of the parsed results of a sequence using MUMmer. This shows how many NODE's are matched with each contig

We used a similar test on MUMmer as we did on BLAST and because of the way that MUMmer aligns its data sequences, it came out with a smaller results than the BLAST results by about 2,000 fewer hits.

Here is a sample of data where it gets the percentage of each Contig based on the lengths of all the nodes related to that Contig ID.

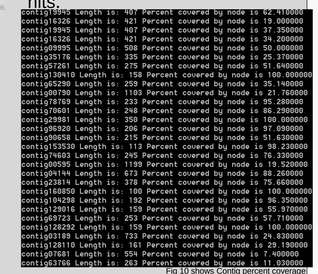

Fig 10 shows Contig percent coverage

Comparisons

Once the results were found using both the BLAST and MUMmer search tools, we created a program to see which sequencing tool had the most hits per contig. The total number of contigs in the database file is 160,749 and the total number of nodes in the query file is 552,305. BLAST returned a total of 123,070 hits and MUMmer returned a total of 121,829 hits. We also found the percent amount that each Node covers each contig. This shows how much each Node is matching with each contig it hits to.

Conclusion and Future Work

The Future of Next Gen Sequencing can be broken down into a variety of categories such as personalized medicine, drug development, climate change, and other life science fields

- Personalized Medicine is a medical model that proposes the customization of medical decision to tailor an individual
- Drug Development is the process of bringing a new drug to the market once a lead compound has been identified through drug discovery.
- Climate change is the active study of past and future theoretical models which uses the past climate data to make future projections.

REFERENCES

- http://www.roche.com/research_and_development/r_d_overview/r_d_sites.htm?id=18
- http://www.illumina.com/technology/solexa_technology.ilmn
- http://blast.ncbi.nlm.nih.gov/Blast.cgi
- https://main.g2.bx.psu.edu/u/dan/p/fastq
- http://ori.dhhs.gov/education/products/n_illinois_u/datamanagement/datopic.htmll
- http://www.jcvi.org/medicago/include/images/chr6.BamHI.maps.jpg
- Gabe Rudy, (2010) *A Hitchhiker's Guide to Next-Generation Sequencing,* :1-9, Golden Helix
- [1] John D. McPherson, (2009) *Next-Generation Gap,* 6:1-4, Nature Methods Supplement
- [2]Michael L. Metzker, *(2010) Sequencing Technologies, - the next generation,* 11:1-5, Nature Reviews
- Md. Fakruddin,Khanjada Shahnewaj Bin mannan, (2012) *Next Generation sequencing technologies – Principles and prospects,* 6:1-9, Research and Reviews in Biosciences

CONTACT INFO

Julian Pierre – julz_pierre@yahoo.com
Jordan Taylor – jtaylor74@my.apsu.edu
Amit Upadhyay – aupadhy1@utk.edu
Bhanu Rekepalli – brekapal@utk.edu