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 generation sequencing can produce over 100 times more data than methods based on Sanger Sequencing. The next generation sequencing technologies offered from Illumina / Solexa, ABI/SOLiD, 454/Roche, and Helicos has provided unprecedented opportunity for high–throughput functional genomic research. Next generation sequence technologies offer novel and rapid ways for genome-wide characterization and profiling of mRNA's, transcription factor regions, and DNA patterns.

Next Generation Sequencing

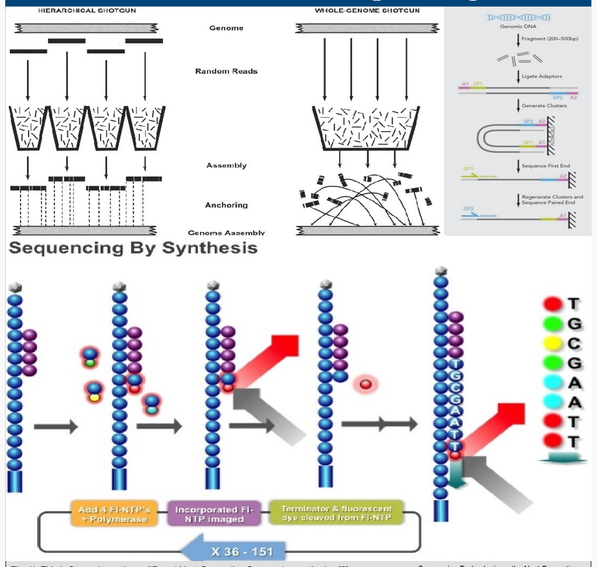


Fig. 1) This is figure shows three different Next Generation Sequencing methods. [2]

Sequencing Technologies – the Next Generation, Michael 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."

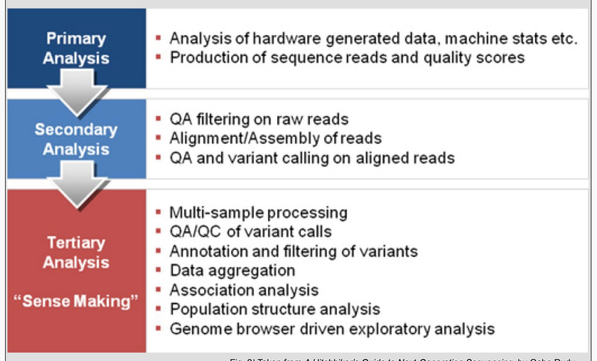


Fig. 2) Taken from *A Hitchhiker's Guide to Next-Generation Sequencing*, by Gabe Rudy

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

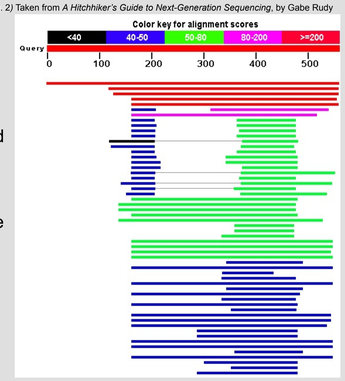


Fig. 3) Taken from bio.davidson.edu/courses. Shows alignment results for yeast.

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.

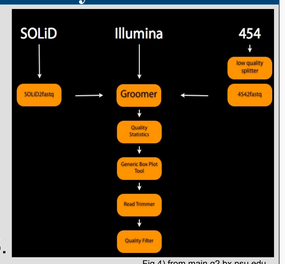


Fig 4) from main.g2.bx.psu.edu

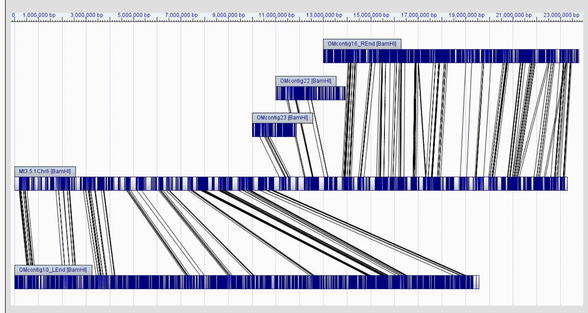


Fig 5) from jcvi.org shows the mapping of chr6 of a Human Genome

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 (Maximum Unique Matches) 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 (Hidden Markov Modeler) 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)

Two FASTA files related to the same nucleotide sequence were input into both BLAST and MUMmer and the results were parsed into tables. Then, the coverage of all hit contigs and nodes from both programs was found.

Using the data gathered from both BLAST and MUMmer, the frequency of the amount covered for each contig was plotted. From Fig 6), it can be inferred that MUMmer hit more accurately for contigs.

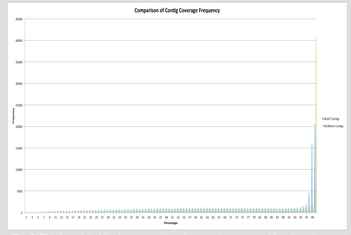
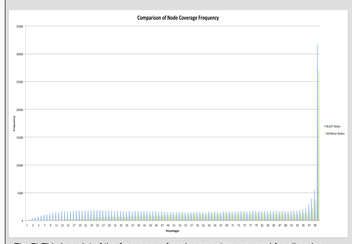


Fig. 6) This is a plot of the frequency of each percentage covered for all contigs. BLAST is in blue, MUMmer is in green.



The same process was done with the Nodes. From Fig 7), it can be inferred that BLAST hit more accurately with nodes. However, there are more BLAST results with lower coverage.

Fig. 7) This is a plot of the frequency of each percentage covered for all nodes. BLAST is in blue, MUMmer is in green.

Using the coverage of each individual contig ID, the results for both BLAST and MUMmer were plotted. While BLAST hit more contigs, there are more contigs with a higher coverage that were hit by MUMmer.

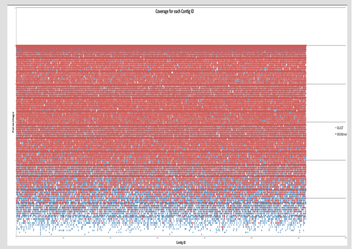
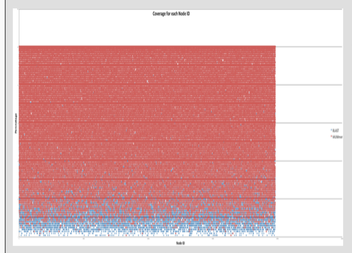


Fig. 8) This is a plot of the coverage of each Contig. BLAST points are blue, MUMmer points are red.



The same process was done with the Nodes. While BLAST hit more Nodes, there are more Nodes that hit with a lower coverage using BLAST.

Fig. 9) This is a plot of the coverage of each Node. BLAST points are blue, MUMmer points are red.

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. From the results, MUMmer hit more accurately than BLAST while BLAST hit more contigs than MUMmer.

Conclusion and Future Work

The future of next generation sequencing can be broken down into a variety of categories such as personalized medicine, bio fuels, climate change, and other life science fields.

- Personalized Medicine is a medical model that proposes the customization of medical decision to tailor an individual
- Bio Fuels present a source of alternative energy. Microalgal biofuels use algae to synthesize the fuel. In order to optimize the process, an understanding of the gene-function relationship of algae would prove helpful.
- Climate change is the active study of past and future theoretical models which uses the past climate data to make future projections.

In conclusion, we hope to contribute the knowledge we have gained to contribute to fields such as these.

REFERENCES

- http://www.roche.com/research_and_development/r_d_overview/r_d_sites.htm?id=18
- http://www.pnas.org/content/99/6/3712/F1.expansion.html
- http://www.yerkes.emory.edu/nhp_genomics_core/Services/Sequencing.html
- 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
- Misra N., Panda P. K., Parida B. K., Mishra B. K., (2012) Phylogenomic Study of Lipid Genes Involved in Mocroalgal Biofuel Production – Candidate Gene Mining and Metabolic Pathway Analyses, *Evolutionary Bioinformatics* 8:545-564, doi: 10.4137/EBO.S10159

CONTACT INFO

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