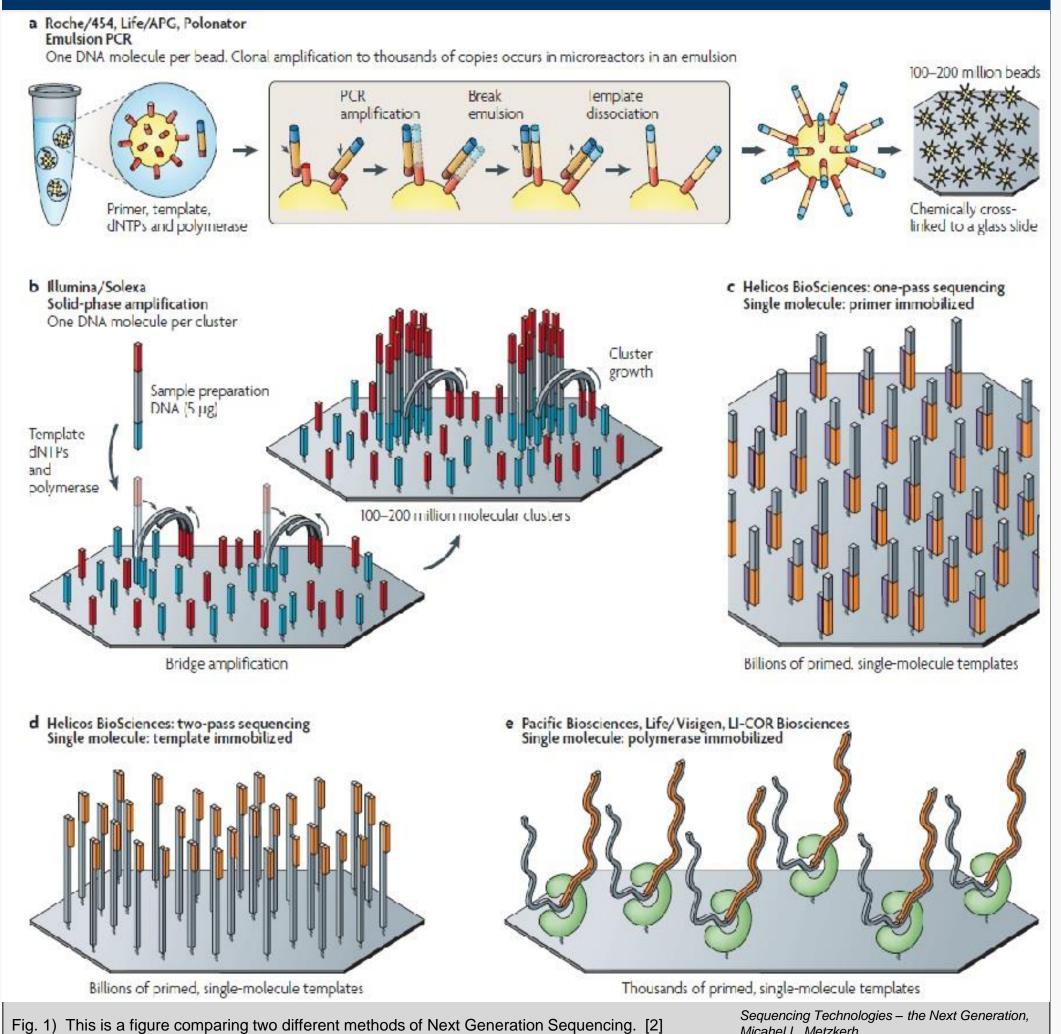


Next Generation Sequencing Pipeline Development and Data Analysis

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.

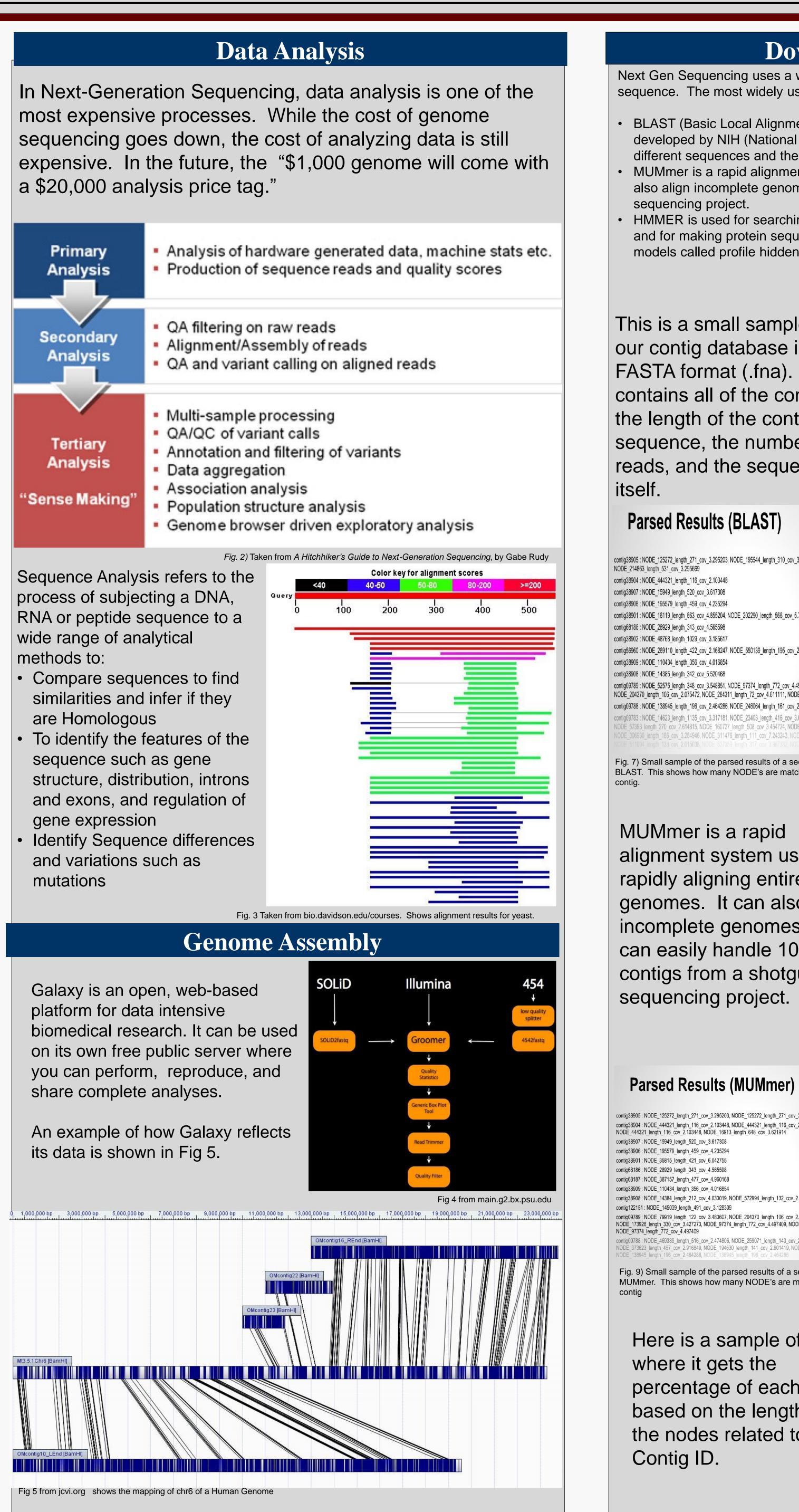
Next Generation Sequencing



- 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.

Julian Pierre¹, Jordan Taylor², Amit Upadhyay³, Bhanu Rekepalli³

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



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

Parsed Results (BLAST)

: NODE_125272_length_271_cov_3.295203, NODE_195544_length_310_cov_3.261290, 63_length_531_cov_3.295669
: NODE_444321_length_116_cov_2.103448
: NODE_15949_length_520_cov_3.617308
: NODE_195579_length_459_cov_4.235294
: NODE_16119_length_663_cov_4.855204, NODE_202290_length_566_cov_5.717315
: NODE_28929_length_343_cov_4.565598
: NODE_48768_length_1029_cov_3.185617
: NODE_269110_length_422_cov_2.168247, NODE_550139_length_195_cov_2.523077
: NODE_110434_length_356_cov_4.016854
: NODE_14385_length_342_cov_5.520468
: NODE_52575_length_348_cov_3.548851, NODE_97374_length_772_cov_4.497409, 70_length_106_cov_2.075472, NODE_284311_length_72_cov_4.611111, NODE_514121_length_244_cov_2
: NODE_138945_length_196_cov_2.464286, NODE_246064_length_161_cov_2.614907
: NODE_14623_length_1135_cov_3.317181, NODE_23403_length_416_cov_3.625000,
3 length 270 cov 2.614815, NODE 160727 length 508 cov 3.454724, NODE 198812 length 374 cov 3

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

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.

Parsed Results (MUMmer)

374_length_772_cov_4.497409 18 : NODE_469389_length_516_cov_2.474806, NODE_259071_length_143_cov_2.237762,	
 4321_length_116_cov_2.103448, NODE_16913_length_648_cov_3.621914 4321_length_1520_cov_3.617308 46 : NODE_195579_length_459_cov_4.235294 41 : NODE_35815_length_421_cov_6.042755 46 : NODE_28929_length_343_cov_4.565598 47 : NODE_110434_length_356_cov_4.960168 49 : NODE_110434_length_212_cov_4.033019, NODE_572994_length_132_cov_2.022727 51 : NODE_14384_length_491_cov_3.128309 49 : NODE_145039_length_491_cov_3.128309 49 : NODE_79919_length_122_cov_3.483607, NODE_204370_length_106_cov_2.075472, 3926_length_330_cov_3.427273, NODE_97374_length_772_cov_4.497409, NODE_97374_length_772_cov_374_length_772_cov_4.497409, NODE_97374_length_772_cov_374_length_772_cov_4.497409, NODE_138945_length_196_362_alength_457_cov_2.916849, NODE_194630_length_141_cov_2.801419, NODE_138945_length_196_ 	5 : NODE_125272_length_271_cov_3.295203, NODE_125272_length_271_cov_3.295203
 16 : NODE_195579_length_459_cov_4.235294 11 : NODE_35815_length_421_cov_6.042755 16 : NODE_28929_length_343_cov_4.565598 17 : NODE_387157_length_477_cov_4.960168 19 : NODE_110434_length_356_cov_4.016854 18 : NODE_14384_length_212_cov_4.033019, NODE_572994_length_132_cov_2.022727 151 : NODE_145039_length_491_cov_3.128309 19 : NODE_79919_length_122_cov_3.483607, NODE_204370_length_106_cov_2.075472, 3926_length_330_cov_3.427273, NODE_97374_length_772_cov_4.497409, NODE_97374_length_772_cov 374_length_772_cov_4.497409 18 : NODE_469389_length_516_cov_2.474806, NODE_259071_length_143_cov_2.237762, 3623_length_457_cov_2.916849, NODE_194630_length_141_cov_2.801419, NODE_138945_length_196_ 	
M1: NODE_35815_length_421_cov_6.042755 i6: NODE_28929_length_343_cov_4.565598 i7: NODE_387157_length_477_cov_4.960168 i9: NODE_110434_length_356_cov_4.016854 i8: NODE_14384_length_212_cov_4.033019, NODE_572994_length_132_cov_2.022727 51: NODE_14384_length_491_cov_3.128309 i9: NODE_79919_length_122_cov_3.483607, NODE_204370_length_106_cov_2.075472, 3926_length_330_cov_3.427273, NODE_97374_length_772_cov_4.497409, NODE_97374_length_772_cov 374_length_772_cov_4.497409 i8: NODE_469389_length_516_cov_2.474806, NODE_259071_length_143_cov_2.237762, 3623_length_457_cov_2.916849, NODE_194630_length_141_cov_2.801419, NODE_138945_length_196_	7:NODE_15949_length_520_cov_3.617308
16 : NODE_28929_length_343_cov_4.565598 17 : NODE_387157_length_477_cov_4.960168 19 : NODE_110434_length_356_cov_4.016854 18 : NODE_14384_length_212_cov_4.033019, NODE_572994_length_132_cov_2.022727 51 : NODE_145039_length_491_cov_3.128309 19 : NODE_79919_length_122_cov_3.483607, NODE_204370_length_106_cov_2.075472, 3926_length_330_cov_3.427273, NODE_97374_length_772_cov_4.497409, NODE_97374_length_772_cov 374_length_772_cov_4.497409 18 : NODE_469389_length_516_cov_2.474806, NODE_259071_length_143_cov_2.237762, 3623_length_457_cov_2.916849, NODE_194630_length_141_cov_2.801419, NODE_138945_length_196_	6 : NODE_195579_length_459_cov_4.235294
i7: NODE_387157_length_477_cov_4.960168 i9: NODE_110434_length_356_cov_4.016854 i8: NODE_14384_length_212_cov_4.033019, NODE_572994_length_132_cov_2.022727 51: NODE_145039_length_491_cov_3.128309 i9: NODE_79919_length_122_cov_3.483607, NODE_204370_length_106_cov_2.075472, 3926_length_330_cov_3.427273, NODE_97374_length_772_cov_4.497409, NODE_97374_length_772_cov 374_length_772_cov_4.497409 i8: NODE_469389_length_516_cov_2.474806, NODE_259071_length_143_cov_2.237762, 3623_length_457_cov_2.916849, NODE_194630_length_141_cov_2.801419, NODE_138945_length_196_	1:NODE_35815_length_421_cov_6.042755
9 : NODE_110434_length_356_cov_4.016854 18 : NODE_14384_length_212_cov_4.033019, NODE_572994_length_132_cov_2.022727 51 : NODE_145039_length_491_cov_3.128309 19 : NODE_79919_length_122_cov_3.483607, NODE_204370_length_106_cov_2.075472, 3926_length_330_cov_3.427273, NODE_97374_length_772_cov_4.497409, NODE_97374_length_772_cov 374_length_772_cov_4.497409 18 : NODE_469389_length_516_cov_2.474806, NODE_259071_length_143_cov_2.237762, 3623_length_457_cov_2.916849, NODE_194630_length_141_cov_2.801419, NODE_138945_length_196_	6 : NODE_28929_length_343_cov_4.565598
8 : NODE_14384_length_212_cov_4.033019, NODE_572994_length_132_cov_2.022727 51 : NODE_145039_length_491_cov_3.128309 9 : NODE_79919_length_122_cov_3.483607, NODE_204370_length_106_cov_2.075472, 3926_length_330_cov_3.427273, NODE_97374_length_772_cov_4.497409, NODE_97374_length_772_co 374_length_772_cov_4.497409 18 : NODE_469389_length_516_cov_2.474806, NODE_259071_length_143_cov_2.237762, 3623_length_457_cov_2.916849, NODE_194630_length_141_cov_2.801419, NODE_138945_length_196_	7 : NODE_387157_length_477_cov_4.960168
51: NODE_145039_length_491_cov_3.128309 99: NODE_79919_length_122_cov_3.483607, NODE_204370_length_106_cov_2.075472, 3926_length_330_cov_3.427273, NODE_97374_length_772_cov_4.497409, NODE_97374_length_772_cov 374_length_772_cov_4.497409 18: NODE_469389_length_516_cov_2.474806, NODE_259071_length_143_cov_2.237762, 3623_length_457_cov_2.916849, NODE_194630_length_141_cov_2.801419, NODE_138945_length_196_	9 : NODE_110434_length_356_cov_4.016854
9: NODE_79919_length_122_cov_3.483607, NODE_204370_length_106_cov_2.075472, 3926_length_330_cov_3.427273, NODE_97374_length_772_cov_4.497409, NODE_97374_length_772_cov 374_length_772_cov_4.497409 18: NODE_469389_length_516_cov_2.474806, NODE_259071_length_143_cov_2.237762, 3623_length_457_cov_2.916849, NODE_194630_length_141_cov_2.801419, NODE_138945_length_196_	8 : NODE_14384_length_212_cov_4.033019, NODE_572994_length_132_cov_2.022727
3926_length_330_cov_3.427273, NODE_97374_length_772_cov_4.497409, NODE_97374_length_772_co 374_length_772_cov_4.497409 i8 : NODE_469389_length_516_cov_2.474806, NODE_259071_length_143_cov_2.237762, 3623_length_457_cov_2.916849, NODE_194630_length_141_cov_2.801419, NODE_138945_length_196_	51 : NODE_145039_length_491_cov_3.128309
3623_length_457_cov_2.916849, NODE_194630_length_141_cov_2.801419, NODE_138945_length_196_	3926_length_330_cov_3.427273, NODE_97374_length_772_cov_4.497409, NODE_97374_length_772_cov
	3623_length_457_cov_2.916849, NODE_194630_length_141_cov_2.801419, NODE_138945_length_196_c

/UMmer. This shows how many NODE's are matched with each

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.

NODE_249442_length_313_cov_3.568690	contig43945	81.13	159	20	10	203	357	237	85	2e-026	119	
NODE_249443_length_193_cov_5.694301	contig08988	97.53	243	6	0	1	243	430	188	3e-116	416	
NODE_249443_length_193_cov_5.694301	contig02600	78.95	171	22	10	18	181	8	171	3e-022	104	
NODE_249444_length_90_cov_3.644444	contig08988	96.43	140	5	0	1	14 0	98	237	7e-061	231	
NODE_249446_length_361_cov_4.335180	contig97044	100.00	208	0	0	203	410	1	208	2e-106	385	
NODE_249450_length_198_cov_2.161616	contig21938	80.66	181	29	5	7	185	180	4	1e-031	135	
NODE_249452_length_395_cov_2.825316	contig134442	100	.00	151	0	0	156	306	151	1 9e-	075	279
NODE_249454_length_170_cov_3.135294	contig84624	87.01	154	20	0	15	168	192	39	2e-043	174	
NODE_249455_length_130_cov_3.084615	contig44292	94.20	69	4	0	109	177	295	227	6e-023	106	
NODE_249458_length_611_cov_3.253682	contig62337	91.38	58	5	0	285	342	185	128	1e-014	80.5	
NODE_249458_length_611_cov_3.253682	contig12544	96.00	50	0	2	287	335	322	274	1e-014	80.5	
NODE_249465_length_301_cov_3.043189	contig41197	100.00	29	0	0	234	262	212	240	4e-007	54.7	
NODE_249466_length_336_cov_5.092262	contig05824	96.55	290	10	0	97	386	1	290	2e-135	481	
NODE_249466_length_336_cov_5.092262	contig116526	96.	92	130	4	0	1	130	51	180 2e-	056	219
NODE_249469_length_173_cov_4.514451	contig24803	93.15	219	10	5	1	216	157	373	3e-086	316	
NODE_249470_length_216_cov_2.856482	contig117735	96.	49	171	5	1	80	250	170	1 1e-	075	281
NODE_249473_length_162_cov_4.666667	contig105176	95.	60	182	7	1	31	212	196	16 2e-	078	291
NODE_249473_length_162_cov_4.666667	contig98615	100.00	30	0	0	1	30	30	1	7e-008	56.5	
NODE_249481_length_293_cov_2.716723	contig23861	79.77	346	64	6	1	343	374	32	7e-065	246	
NODE_249481_length_293_cov_2.716723	contig143664	91.	87	123	10	0	51	173	130	8 1e-	042	172
NODE_249481_length_293_cov_2.716723	contig49210	81.33	150	22	6	37	185	145	1	5e-026	117	
NODE_249481_length_293_cov_2.716723	contig66288	76.58	158	29	7	70	222	76	230	7e-015	80.5	
NODE_249483_length_206_cov_2.393204	contig120421	85.	35	157	19	4	18	173	166	13 7e-	039	159
NODE_249486_length_421_cov_3.995250	contig50018	91.49	282	22	1	128	409	2	281	5e-107	387	
NODE_249492_length_455_cov_4.907692	contig23032	98.91	368	4	0	138	505	1	368	0.0 65	В	
NODE_249492_length_455_cov_4.907692	contig47880	86.93	199	21	5	1	197	196	1	2e-056	219	
NODE_249492_length_455_cov_4.907692	contig09718	81.95	133	16	6	5	133	493	365	2e-022	106	
NODE_249493_length_207_cov_3.603865	contig07077	90.32	248	20	4	3	250	355	112	8e-088	322	
NODE_249495_length_51_cov_4.000000	contig27784	100.00	101	0	0	1	101	102	202	1e-047	187	
NODE_249496_length_404_cov_4.297029	contig27784	100.00	210	0	0	245	454	362	153	1e-107	388	
NODE_249496_length_404_cov_4.297029	contig121379	90.	24	164	16	0	1	164	6	169 3e-	055	215

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

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

> contig00001 NODE_528348_length_133_cov_3.669173 NODE_55298_length_81_cov_6.444445 NODE_55298_length_81_cov_6.444445 NODE_55298_length_81_cov_6.444445 NODE_55298_length_81_cov_6.444445 NODE_55298_length_81_cov_6.444445 NODE_55298_length_81_cov_6.444445 NODE_55298_length_81_cov_6.444445 NODE_55298_length_81_cov_6.444445 NODE_55298_length_81_cov_6.444445 NODE_55298_length_81_cov_6.444445 NODE_528348_length_133_cov_3.669173 NODE_5298_length_81_cov_6.444445 NODE_5298_length_81_cov_6.444445 NODE_5298_length_81_cov_6.444445 NODE_5298_length_81_cov_6.444445 NODE_5298_length_81_cov_6.444445 NODE_5298_length_81_cov_6.444445 NODE_5298_length_81_cov_6.444445 NODE_5298_length_81_cov_6.444445 NODE_528348_length_133_cov_3.669173 NODE_528348_length_133_cov_3.669173 NODE_528348_length_81_cov_6.444445 NODE_528348_length_81_cov_6.444445 NODE_528348_length_81_cov_6.444445 NODE_528348_length_81_cov_6.444445 NODE_528348_length_81_cov_6.444445 NODE_528348_length_81_cov_6.444445 NODE_528348_length_81_cov_6.444445 NODE_528348_length_81_cov_6.444445 NODE_528348_length_81_cov_6.444445 NODE_528348_length_81_cov_6.444445 NODE_528348_length_81_cov_6.444445 NODE_528348_length_81_cov_6.444445 NODE_528348_length_81_cov_6.444445 NODE_528348_length_81_cov_6.444445 NODE_528348_length_81_cov_6.444445 NODE_528348_length_81_cov_6.444445 NODE_528348_length_81_cov_6.444445 NODE_528348_length_81_cov_6.444445 NODE_528348_length_927_cov_3.318043	76 57 3 82 82 82 143 10 100 46 137 88 143 28 34 155 189	26 56 62 125 131 137 149 155 173 179 197 203 209 215 227 233 2728	26 24 61 29 29 35 29 35 20 35 20 35 29 29 29 29 29 29 23 23 20					
NODE_373_length_25487_cov_12.900380 NODE_373_length_25487_cov_12.900380 NODE_484364_length_69_cov_3.275362 NODE_373_length_25487_cov_12.900380 NODE_373_length_25487_cov_12.900380 NODE_540481_length_57_cov_5.421052 NODE_540481_length_57_cov_5.421052 NODE_373_length_25487_cov_12.900380 NODE_373_length_25487_cov_12.900380 NODE_373_length_25487_cov_12.900380 NODE_373_length_25487_cov_12.900380 NODE_345997_length_70_cov_13.057143 > contig00002	17209 19193 100 25428 25383 16 22 25395 25395 25377 70 1	1 1985 8199 8220 8235 8241 8253 8265 8297 8303 8337	1983 6232 20 44 35 41 35 24 29 50 75					
NODE_4508_length_4558_cov_12.879113 NODE_21646_length_847_cov_10.245573 NODE_88109_length_1941_cov_11.710459 NODE_123355_length_412_cov_5.033980	3954 1 1 240	1 1577 2424 5762	655 897 1991 20					
contig00002 Reverse NODE_14695_length_2750_cov_13.610545 NODE_15450_length_971_cov_10.280124 NODE_38900_length_5306_cov_7.855824 NODE_575113_length_136_cov_2.117647	1197 1 2256 127	1 4343 5445 5645	1604 1021 20 21					
Fig. 8) This is a sample out put of the results These are in stdout format.	given from	n MUMmei						
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. contig19945 Length is: 407 Percent cou contig16326 Length is: 421 Percent cou								
contig19945 Length is: 407 Percent concontig09995 Length is: 421 Percent concontig09995 Length is: 508 Percent concontig57261 Length is: 335 Percent concontig130410 Length is: 158 Percent concontig65290 Length is: 158 Percent concontig00790 Length is: 1103 Percent concontig78769 Length is: 233 Percent concontig70601 Length is: 248 Percent concontig96920 Length is: 350 Percent concontig96920 Length is: 215 Percent concontig96920 Length is: 113 Percent concontig96920 Length is: 113 Percent concontig90658 Length is: 113 Percent concontig153530 Length is: 113 Percent concontig00595 Length is: 1199 Percent concontig00595 Length is: 1199 Percent concontig00595 Length is: 1199 Percent concontig160850 Length is: 100 Percent concontig160850 Length is: 192 Percent concontig160850 Length is: 159 Percent concontig129016 Length is: 159 Percent concontig129016 Length is: 159 Percent concontig128292 Length is: 159 Percent concontig128110 Length is: 161 Percent concontig07681 Length is: 554 Percent concontig07681 Length is: 263 Percent concontig07681 Length is: 263 Percent concontig07666 Length	pered by n pered by n	ode is 37 ode is 34 ode is 50 ode is 50 ode is 51 node is 1 ode is 35 node is 95 ode is 95 ode is 97 ode is 97 ode is 97 ode is 97 ode is 51 node is 1 ode is 75 node is 1 node is 1 node is 57 node is 57 node is 57 node is 57 node is 24 node is 24 node is 7. ode is 7. ode is 7.	.350000 .200000 .370000 .640000 00.00000 .140000 1.760000 .280000 .290000 .290000 .630000 8.230000 8.230000 .330000 9.520000 .660000 00.00000 6.350000 5.970000 .710000 00.00000 .830000 9.190000 400000 .030000					

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.



Comparisons

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_over view/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/dataman agement/datopic.htmll

 http://www.jcvi.org/medicago/include/images/chr6.BamHI.m aps.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