Analysis of Illumina Microbial Assemblies

Alicia Clum[1] (aclum@lbl.gov), Brian Foster[1], Jeff Froula[1], Kurt LaButti[1], Alex Sczyrba[1], Alla Lapidus[1] and Tanja Woyke [1]
[1]Lawrence Berkeley National Laboratories
DOE Joint Genome Institute, Walnut Creek

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

Introduction

Since the emerging of second generation sequencing technologies, the evaluation of different sequencing approaches and their assembly strategies for different types of genomes has become an important undertaken. Next generation sequencing technologies dramatically increase sequence throughput while decreasing cost, making them an attractive tool for whole genome shotgun sequencing.

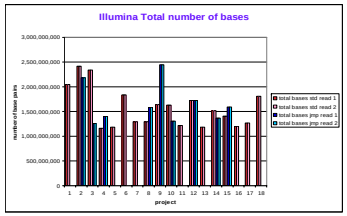
To compare different approaches for de-novo whole genome assembly, appropriate tools and a solid understanding of both quantity and quality of the underlying sequence data are crucial. Here, we performed an in-depth analysis of short-read Illumina sequence assembly strategies for bacterial and archaeal genomes. Different types of Illumina libraries as well as different trim parameters and assemblers were evaluated. Results of the comparative analysis and sequencing platforms will be presented. The goal of this analysis is to develop a cost-effective approach for the increased throughput of the generation of high quality microbial genomes.

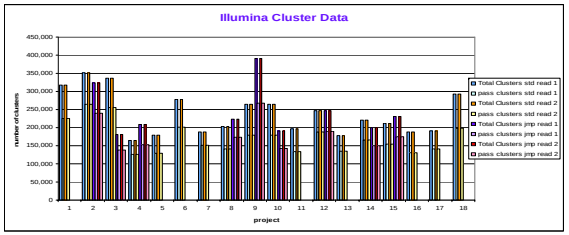
Methods

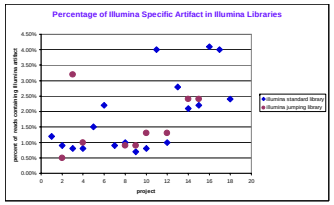
Eighteen bacterial and archaeal genomes of diverse GC and repeat content were chosen for the analysis. Illumina standard and Illumina jumping libraries were generated and one lane of each was sequenced. Quality was assessed and trimming was done if necessary. Datasets were assembled with Velvet, or Newbler in combination with a 454 paired end library (where available) and compared to the quality draft (QD) assembly. All genomes analyzed have a finished reference so we were able to evaluate how much of the reference was covered. For all charts, genomes are ordered by ascending GC content.

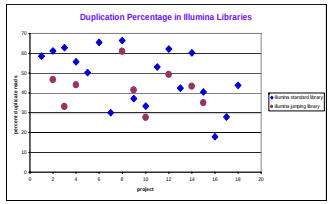
project number	species name	genome size (bp)	GC%	Illumina 300 bp (std) library	insert size	sd	Illumina Jumping library	median insert of correctly oriented reads	median insert from velvet	sd from velvet	454 PE library insert size	Technologies in QD	QD assembler	Num of Repeats
1	Brachyspira murdochii DSM 12563	3,241,805	28	GWZO	196	206					n/a	Sanger (8kb), 454-FLX	pga	193
2	Arcobacter nitrofigilis DSM 07299	3,192,235	29	GUYB	278	79	GWHN	4647	4506	2262	26907 +/- 6726	454-Ti, 454-Ti-PE	newbler	131
3	Ilyobacter polytropus CuHBu1, DSM 2926	3132299	35	GUYC	278	79	GWHP	3797	414	1898	13990 +/- 3497	454-Ti, 454-Ti-PE	newbler	439
4	Ignisphaera aggregans AQ1.S1, DSM 17230	1875953	36	GONU	232	58	GOYZ	2191	2135	853	15001 +/- 3750	454-Ti, 454-Ti-PE	newbler	51
5	Acetohalobium arabaticum Z-7288, DSM 5501	2,469,596	37	GWZP	204	73					10395 +/- 2598	454-Ti, 454-Ti-PE	newbler	268
6	Archaeoglobus profundus Av18, DSM 5631	1563423	42	GUYA	262	72					15524 +/- 3881	454-Ti, 454-Ti-PE	newbler	47
7	Pedobacter heparinus HIM 762-3, DSM 2366	5,167,387	42	GWZS	267	75					n/a	Sanger (8kb & 40kb)	pga	126
8	Thermosphaera aggregans M11TL, DSM 11486	1,316,595	47	GUYI	280	82	GWSO	3345	396	1535	8298 +/- 2074	454-Ti, 454-Ti-PE, Illumina	newbler	11
9	Fibrobacter succinogenes S85 ATCC 19169	3842636	48	GONW	231	53	GOZA	2287	2240	820	20581 +/- 5145	454-FLX, 454-Ti-PE	newbler	156
10	Spirochaeta smaragkinae DSM 11293	4653970	48	GONY	241	60	GUPN	3828	3674	2117	12036 +/- 3009	454-Ti, 454-Ti-PE	newbler	123
11	Arcanobacterium haemolyticum DSM 20595	1986158	53	GWZU	195	80					8317 +/- 2079	454-Ti, 454-Ti-PE, Illumina	newbler	72
12	Palaeococcus ferophilus DSM 13482	2217824	54	GUYG	280	80	GWSI	3642	200	1487	12867 +/- 3216	454-Ti, 454-Ti-PE	newbler	n/a
13	Alicyclobacillus acidocaldarius acidocaldarius 104-IA, DSM 446	2,669,688	62	GWZX	262	72					n/a	Sanger (8kb), 454-FLX	pga	172
14	Olsenella uli VPI, DSM 7084	2051896	65	GUYF	291	83	GWSH	3592	218	1422	9273 +/- 2318	454-Ti, 454-Ti-PE, Illumina	newbler	37
15	Segniliparus rotundus DSM 44985	3157527	67	GUYH	269	73	GWSN	3699	284	1717	3949 +/- 987	454-Ti, 454-Ti-PE, Illumina	newbler	84
16	Streptosporangium roseum NI 9100, DSM 43021	10369532	71	GWZY	214	32					n/a	Sanger (8kb & 40kb)	pga	493
17	Conexibacter woesei ID131577, DSM 14684	6,359,369	73	GWZZ	278	84					n/a	Sanger (8kb & 40kb), 454-FLX	pga	65
18	Cellulomonas flavigena 134, DSM 20109	4123179	74	GXAA	253	65					n/a	Sanger (8kb & 40kb)	pga	83

Read QC and Analysis

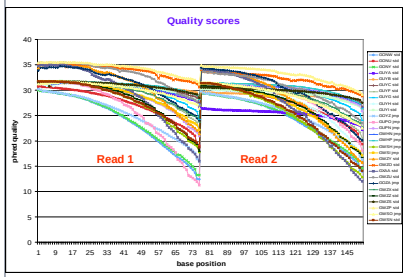


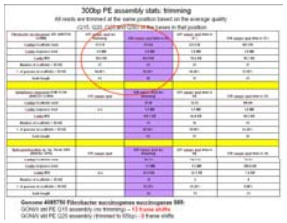






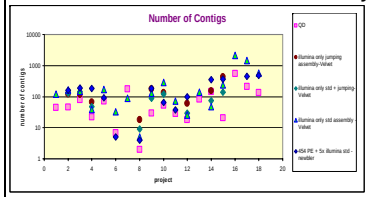
A random sample of 5% of the data from the lane was used to generate the percentages of duplication and Illumina artifact.

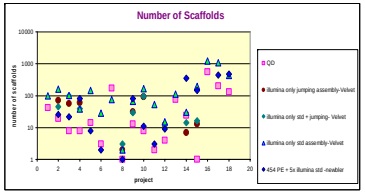


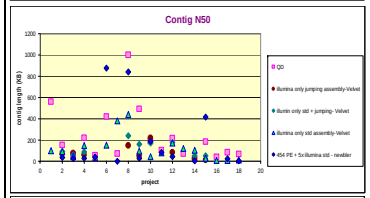


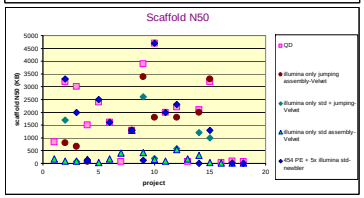
Based on tests of trimming by quality, we decided to trim reads to a Sanger Q20 phred score for assembly if the quality values were below that threshold. Striking a balance between quality and read length is important.

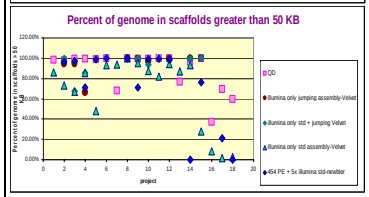
Assembly Analysis











Please note that the charts for Number of Contigs and Number of Scaffolds are log scale. Also projects 3,14, and 18 were assembled with 15x of illumina data (not 5x). Projects are in order of ascending GC content. The correct sequence orientation for jumping data is ← →. This data was reverse complemented before assembly because velvet expects pairs that face each other.

All Illumina only assemblies are with Velvet, the 454 paired end + 5x Illumina data is assembled with Newbler, and the QD assembly (see chart listing QD assembler). Note: There aren't scaffold fasta files for PGA QD assemblies so for those assemblies the statistics for scaffolds and contigs are the same.

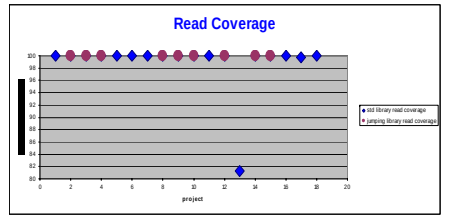
Conclusions

- Quality remains an important aspect to sequencing. Increased read quality leads to better quality assemblies and less frameshifts due to consensus errors.
- Because basic biology remains the same, longer insert libraries are needed to scaffold more complex genomes for finishing.
- Most assemblies with 5x Illumina data and 454 paired end data look comparable to the QD assembly.
- Illumina reads cover nearly 100% of genomes but Illumina contigs cover between 68.7% to 99.97% of the genome.
- High GC genomes continue to be problematic .

Future Developments

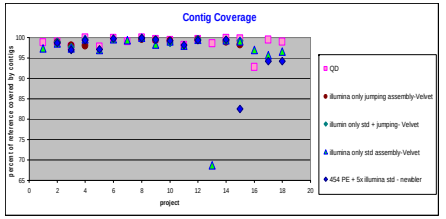
- Test merging the output of different assemblies (the same dataset with different assemblers or different datasets with the same assembler). Preliminary work has been done to test this using minimus from the AMOS package.
- Test and analyze larger insert jumping libraries on a read and assembly level.
- Improve jumping libraries to increase the amount of correctly oriented pairs by making the ligation more efficient.
- Test optimal assembly coverage for Illumina data for the purpose of barcoding and pooling microbes.

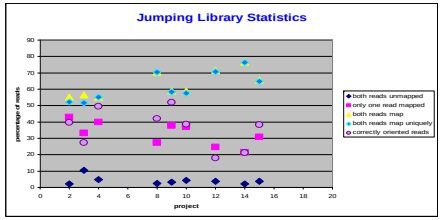
Alignment Analysis



In most cases Illumina reads cover ~100% of the genome. Assembled contigs, however, cover between 68.7% to 99.97% of the genome.

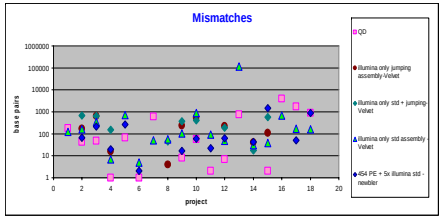
Reads and contigs were aligned to the reference using Arachne's QueryLookupTable.

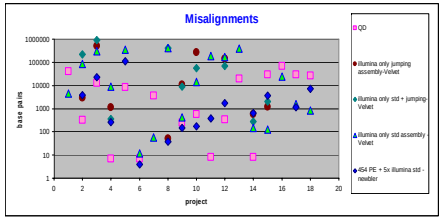




For jumping data correctly oriented data faces ← →. There is a relatively high percent of reads where only one of the pairs maps to the reference. A threshold of 90% ID was used so it is possible the pair was lower quality and had errors. The percentage of correctly oriented pairs varies from ~18-52%. Ideally we would like to increase this amount.

To validate and compare assemblies we counted mismatches and misalignments for each assembly as compared to the reference. These charts are in log scale.





This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.