Data Processing Strategies for Bacterial Genome Re-Sequencing Projects Using Illumina Solexa Technology

Ponnala L, Bukowski R, VanEe J, Schweitzer P, Simpson K, Sun Q

Life Sciences Core Laboratories Center, Cornell University, Ithaca, NY

ABSTRACT

Due to intensive genome rearrangements and high selection pressure, bacterial genomes are very dynamic, which makes it difficult to align short-read sequencing data to the reference genomes. In this study, we explore using the Illumina Solexa sequencing technology to do comparative genomics studies of multiple invasive *E. coli* isolates. Illumina Solexa reads are aligned to multiple finished *E. coli* genomes. Bbrowse is used to visualize the coverage and SNP data. Within each ortholog group, the protein sequence of the known genomes with the highest number of Illumina Solexa coverage is used as the template to assemble the new strain's sequence. Codon based analysis is carried out in each ortholog group.

INTRODUCTION

Cornell University Life Sciences Core Laboratories Center (CLC): The CLC provides an array of life sciences shared research resources and services to all Cornell University and to outside institutions. The CLC includes fee-for-service research, technology testing and development, and educational components. The CLC is part of a Center for Advanced Technology in Life Science Enterprise designated by New York State. The mission of the CLC is to promote research in the life sciences with advanced technologies in a shared resource environment. The resources of the CLC facilities are open to all investigators at Cornell University and to investigators at other academic institutions and at commercial enterprises. The CLC is composed of eight core facilities covering DNA sequencing and genotyping, microarrays, proteomics and mass spectrometry, protein production and characterization, microscopy and imaging, transgenics, bioinformatics, bio-IT, and advanced technology assessment (http://cores.lifesciences.cornell.edu).

Computational Biology Service Unit (CBSU) The CBSU is the bioinformatics core of the Life Sciences Core Laboratories Center (CLC). The CBSU provides research, software and hardware support for life sciences research, including research collaboration, software development and maintenance of database resources. The facility was founded in 2001 as a computational resource for the Tri-institutional Collaboration among Cornell University, Rockefeller University, and Memorial Sloan-Kettering Cancer Center. The CBSU is hosted by the Cornell University Center for Advanced Computing (CAC). In 2006, the CBSU became part of the Life Sciences Core Laboratories Center. Also in 2006, the CBSU was chosen to become one of ten Microsoft High-Performance Computing Institutes worldwide. The CBSU has a 677 node cluster to support analysis applications and has expertise in large scale sequencing data management.

New sequencing technologies services: The CLC DNA Sequencing and Genotyping Facility has two ABI 3730xl 96-capillary array sequencing instruments. The CLC currently offers the Illumina Solexa Genome Analyzer as a standard core facility service and recently offered the Roche 454 GS-FLX as a core service for a five month demonstration period. The Illumina Solexa instrument is being purchased and funding for long term placement of a 454 GS-FLX is being pursued. For these new sequencing technologies, the CLC DNA Sequencing and Genotyping Facility, in collaboration with the DNA Microarrays, Computational Biology, and Information Technology core facilities, provides project consultation, sample preparation, data generation, and analysis support. The DNA Sequencing and Genotyping Facility provides expertise in high throughput sequencing projects. The DNA Microarrays Facility provides expertise in RNA based applications. The Computational Biology Services Unit has developed and offers informatics pipelines for both the 454 and Illumina Solexa platforms. These pipelines are in a state of continuous development.

Comparison of sequencing technologies:

	ABI 3730xl	Roche 454 GS-FLX	Illumina Solexa
Average read length	750 bases	250 bases	35 bases
# of parallel reads	96	400,000	60 million
# bases per full run	< 100,000	> 100 million	> 1 million (1GB)
Run time per full run	2 hours	7.5 hours	3 days
Raw data output scale (bytes)	MB	GB	TB
Cost per full run (with 1 sample)	$50-$500	$7500	$3500

RESOURCES

Life Sciences Core Laboratories Center resources support for next generation sequencing:

DNA Sequencing and Genotyping : Illumina Solexa Genome Analyzer (currently available), Roche 454 Genome Sequencer FLX (recently available), two ABI 3730xl DNA Analyzers, three Tecan Genesis liquid handling robots (two model RSP150 and one model 200), Beckman Biomek FX liquid handling robot, MD spectraPlus plate reader, Nanodrop 1000 spectrophotometer, Sigma 4-15C centrifuge, Fisher Marathon 3200R centrifuge, six PE 9700 thermocyclers, ABI 7900 HT Sequence Detection System, Illumina BeadStation 500GX.

DNA Microarrays: Affymetrix GeneChip System with SNP chip and Targeted Genotyping (TG) upgrades, including model 3000 Scanner, model 640 hybridization oven, three model 450 Fluidics Stations, Agilent 2100 Bioanalyzer, access to Illumina BeadStation 500GX.

Information Technology Services: 15 Linux and Windows-based file, database and application servers totaling >4 TB RAID storage, 12 TB of backup/archive capacity.

Computational Biology: 425-node Sun Linux cluster with 280 GB local HD space; 252-node Windows cluster with 344 GB local HD space; 6 web and general purpose servers with 1.1 TB total HD space; 4 file and database servers with 15.5 TB total HD space. We are in the process of obtaining a large shared-memory machine with 128 GB RAM.

RESULTS

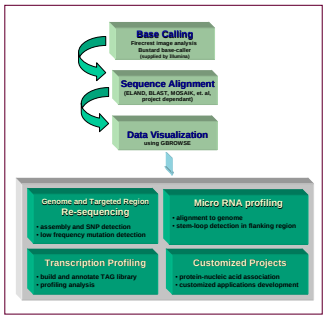
Data Analysis

High Performance Computing Cluster

Sequencing data processed on a 677 node cluster.

Informatics Pipeline

Illumina Solexa



Alignment to Reference Genome

using ELAND and BLAST

The quality scores of the Illumina Solexa reads decline after 25 cycles. Running ELAND using 32 base pairs leads to lower coverage.

ELAND	% unique matches with 0 to 2 errors	% multiple matches with 0 to 2 errors	total base pairs that match to the genome	average coverage on genome
32mer	35.14%	0.60%	44910880	8.6
32mer->25mer progressive run	54.38%	0.40%	66620557	12.7
25mer	54.27%	0.99%	54189400	10.4

Running ELAND progressively using 32 to 25 base pairs can increase coverage, while keeping the multiple match numbers low.

ELAND	U0	U1	U2	R1	R2	NM	QC	R0
32	374027	526970	502468	8738	7267	2560953	5190	8681
31	150	312	130973	9228	7141	2428711	4223	1410
30	156	288	127396	9613	7129	2299870	3426	1425
29	94	124	113075	9789	7164	2185090	3330	1372
28	54	151	107114	10006	7182	2076264	3312	1290
27	130	198	101036	10123	7197	1973761	2994	1325
26	373	405	91407	10256	7087	1881358	1907	1282
25	194	231	94612	10395	6988	1785271	1506	1411

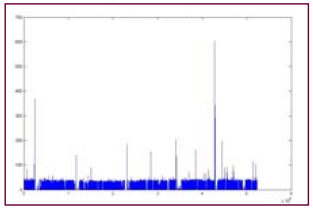
Running BLAST with small word size allows more flexible parameter setting, but would not increase the coverage for this project.

Genome Coverage

From Illumina Solexa Run

Genome coverage (bp)	3,842,953	73.5%
# of genes with > 50% coverage	3922	72.9%
# of genes with > 75% coverage	3851	71.6%
# of genes with > 90% coverage	3584	66.6%

Bacterial genome size: 5,231,428 bp
Number of Genes: 5,379

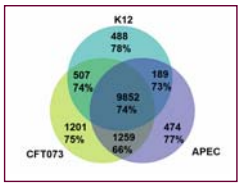


Coverage depth of Illumina Solexa reads on the bacterial genome

Comparison of Genome Coverage

on Three *E. coli* Reference Genomes

Genome	Length (bp)	Coverage (bp)		#Genes	#Genes >90% covered	
K12	4,639,675	4,074,616	87.8%	4,133	3,579	86.6%
APEC	5,082,025	3,854,383	75.8%	4,458	3,125	70.1%
CFT073	5,231,428	3,842,953	73.5%	5,379	3,584	66.6%



Ortholog groups of the three genomes and gene coverage in each category

SNP Identification

Length of genome	5231428	
Number of bases at coverage level 0	1388475	26.5%
Number of bases at coverage level (1-3)	98572	1.9%
Number of bases at coverage level (4-20)	2588970	49.5%
Number of bases at coverage level >20	1155411	22.1%

SNPs are identified by counting alleles at each polymorphic site with 3 or higher coverage.

Genome Browser Tools

Genome browser view of *E. coli* assembly with a tract that we added showing depth of coverage from Solexa reads.

CONCLUSIONS

1. Due to the dynamic genome structures of bacteria, multiple reference genomes are necessary to increase the coverage of the short reads obtained from the Illumina Solexa platform.
2. Running ELAND progressively from 32 bp to 25 bp read lengths yields greater genome coverage than using only 32 bp or 25 bp read lengths. Choices of alignment algorithms and parameter settings are largely dependant on the goals of the project. Different data processing pipelines are needed for different projects.
3. Short sequencing reads can be used as a cost effective method for codon based analysis for most of the genes on the genome.

Contact Information

http://cores.lifesciences.cornell.edu

Life Sciences Core Laboratories Center
George Grills
Director of Operations
Director Advanced Tech. Assessment
LSCLC_director@cornell.edu

DNA Sequencing and Genotyping
Peter Schweitzer, Director
LSCLC_dna@cornell.edu

DNA Microarrays
Wei Wang, Director
LSCLC_microarray@cornell.edu

Information Technology Services
James VanEe
LSCLC_its@cornell.edu

Computational Biology Service Unit
Qi Sun, Co-Director
LSCLC_cbsu@cornell.edu