

Accuracy of Next-Generation DNA Sequencing: illumina Genome Analyzer Re-sequencing of *Staphylococcus aureus* MW2

FARINELLI, Laurent¹; OSTERAS, Magne¹; BAERLOCHER, Loïc¹;
HERNANDEZ, David²; FRANCOIS, Patrice²; SCHRENZEL, Jacques²

¹ FASTERIS SA, 1228 Plan-les-Ouates, Switzerland;

² Genomic Research Laboratory, Geneva University Hospitals, 1211 Geneva 14, Switzerland

Abstract



Next-generation DNA sequencing technologies have revolutionized sequencing and re-sequencing projects by dramatically increasing the speed and reducing the costs of analysis. We were interested by the robustness and particularly the error rates of data obtained using these novel approaches.

We used a shotgun protocol to process the genomic DNA of *Staphylococcus aureus* strain MW2. Using 2 channels on the illumina Genome Analyzer, we obtained 3,857,879 reads of 30-35 bases, representing a 1.3x coverage in reads and a 47x coverage in bases.

Using the MAQ alignment software, we mapped the reads on the published sequence, including the pMW2 plasmid. Only a single base was not covered by any read. Using criteria to select non-ambiguous SNPs, no SNP were detected.

We repeated the sample preparation starting from another genomic DNA sample and obtained 2,684,877 reads from one channel, representing a 1x coverage in reads and a 33x coverage in bases. All the bases of the reference sequences were covered by mapped reads. Again, no non-ambiguous SNPs were detected.

For both samples the Genome Analyzer reads were de novo assembled using the EDENA software (Hernandez et al, submitted). The contigs obtained covered 95% of the reference sequences. The absence of SNPs in the first sample was confirmed by the contigs obtained using EDENA.

These results underline the outstanding reliability of sequence data obtained by this technology, where base call errors can be corrected using a combination of very high genome coverage and adequate analysis tools for the assembly process.

Introduction

Massively parallel DNA sequencing technologies have enabled sequencing or re-sequencing of entire genomes in a matter of weeks at a fraction of the cost of capillary DNA sequencing. The applications of Next-Generation or Ultra-Deep DNA sequencing are not limited to sequencing *per se*, but include almost all fields of genomics, e.g. Digital Gene Expression Profiling (*i.e.* counting the transcripts present in a sample); small RNA discovery and analysis; genotyping; or ChIP-SEQ-like experiments where the binding sites of proteins are monitored on whole genomes.

We used *Staphylococcus aureus* strain MW2 as a model to study the accuracy of data produced by the illumina Genome Analyzer system. Although the error rate in the reads obtained by the Next-Generation systems may be higher than capillary sequence data, the very high coverage enables proof reading of the data as errors are expected to be produced at random while SNPs should be found at the same location in overlapping but not identical reads.

S. aureus is a major pathogen responsible for both healthcare- and community-associated infections.

The illumina Genome Analyzer

The illumina Genome Analyzer system is based on a technology developed in Geneva in 1996, the DNA Colonies (Mayer and Farinelli, 1997), where DNA samples are amplified in situ on a glass slide, like a million-plex PCR.

After primer annealing, the DNA colonies are sequenced in parallel base-by-base using a reversible terminator approach. The illumina Genome Analyzer system (www.morethansequencing.com) uses flow-cells with 8 channels, each channel generating 2-5 mio reads of 35 bases, *i.e.* a 3.5 days run often generates over 1 billion bases.

Up to 10,000,000 DNA colonies / cm²

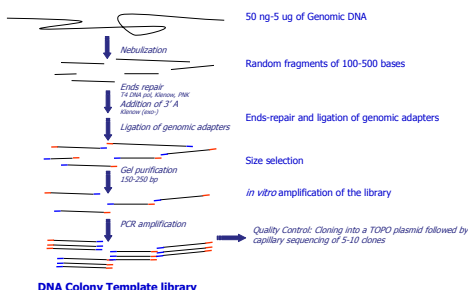
TGCTACGAT...

In situ sequencing TTTTTTTGT...

The identity of each base of a DNA colony is read off from sequential images using a reversible terminator approach: after each cycle the 3' is de-blocked and the fluorophore is removed.

Genomic DNA Shotgun Sample Preparation

The whole sample preparation and bioinformatics data analysis was performed twice from genomic DNA isolated from *Staphylococcus aureus* strain MW2, for which the complete genomic sequence is available.

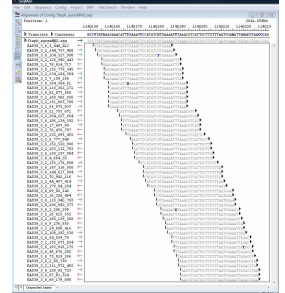


Mapping of the Reads on the Reference Genome

The images from the Genome Analyzer were converted into 35 bases reads using the Solexa Data Analysis pipeline.

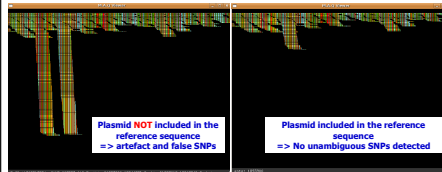
The reads were mapped on the MW2 reference sequence NC_003923, using DNASTAR or MAQ. We further processed the data to estimate the coverage and identify SNPs.

□ Mapping using the DNASTAR SeqMan Genome Assembler software



□ Mapping using the MAQ software

MAQ	Experiment 1	Experiment 2
Genomic DNA prep	Jan 2007	Jan 2005
Ref. / size	NC_003923 / 2.82 Mb AP004832 pMW2 / 20.6 kb	
Reads	3'857'879 (2 channels)	2'684'877 (1 channel)
Reads mapped (%)	3'606'310 (93.5%)	2'530'239 (94.2%)
Coverage gDNA	44.8	30.7
Coverage pMW2	245	92.6
Bases not covered	1	0
SNPs	0	0

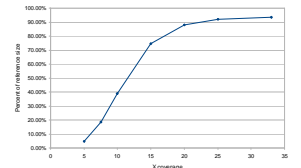


- ✓ Over 30x coverage
- ✓ The Whole Genome is covered by reads
- ✓ Detection of higher coverage regions: the pMW2 plasmid
- ✓ No SNPs were detected

de novo Assembly

de novo assembly of short reads is a real challenge as millions of reads must be compared one with each other. We used the beta 2 version of EDENA software (Hernandez et al, submitted) using standard parameters.

EDENA	Experiment 1	Experiment 2
Genomic DNA prep	Jan 2007	Jan 2005
Ref. / size	NC_003923 / 2.82 Mb AP004832 pMW2 / 20.6 kb	
Reads	3'857'879 (2 channels)	2'684'877 (1 channel)
Contigs	1726	2468
Total length	2'702'725	2'633'369
Ctigs mapped	1726	2462
Coverage	95%	92.3%



Conclusions

- **Accuracy:** The assembled sequences obtained using the illumina Genome Analyzer system are highly accurate as no errors were detected in 2 independent experiments.
- **Coverage:** A very high coverage of over 30x is achieved on a 2.8 Mb genome using the reads of a single channel. The uniform coverage demonstrates that no bias was introduced during sample preparation or Genome Analyzer sequencing.
- **SNPs:** The fact that no SNP was detected is surprising, as we could have expected differences with the reference sequence obtained by capillary sequencing or the appearance of spontaneous mutations.
- **de novo assembly:** The EDENA software produces contigs that cover over 95% of a *S. aureus* genome. This software was also used successfully to assemble the transcriptome from a plant for which no reference sequence was available.

References

- 1) Mayer P. and Farinelli L., 1997, Patent application WO 98/44151.
- 2) Hernandez D., Francois P., Farinelli L., Osteras M., and Schrenzel J., *De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer*, Submitted