illumina®

# One Run, Eight Bacterial Genomes: The Potential of the MiSeq® Personal Sequencer

With 10-fold higher output than other small-scale next-generation sequencing platforms, the MiSeq system will enable Tim Stinear, Ph.D. to spend less time generating data and more time analyzing it.

Tim Stinear, Ph.D. is a Principal Research Fellow in the Department of Microbiology and Immunology at the University of Melbourne. His research group, consisting of 10 students, post docs, and research assistants, uses a variety of genomic and molecular biology techniques to study how certain groups of bacteria cause human disease. In particular, they focus on two topics: how mycobacteria cause diseases such as tuberculosis and leprosy, and how hospital pathogens develop antibiotic resistance. To increase their understanding of what these bacteria are doing and how they are evolving, the lab is performing comparative genomics studies using next-generation sequencing.

We recently caught up with Dr. Stinear between sessions at the 16th International Symposium on the Biology of Actinomycetes (ISBA) and spoke with him about his experience with different next-generation sequencing platforms.

*Q: Will you be presenting at the ISBA conference?*

**Tim Stinear (TS):** Yes. I'll be presenting data on what we've found from extensive whole-genome sequencing of many different strains of mycobacteria. We've uncovered how *Mycobacterium ulcerans*, an unusual group of mycobacteria that cause skin ulcers in people, specifically Buruli ulcer, has evolved and the special tricks it's developed to cause disease in humans.

### Comparison of *Enterococcus faecium de novo* Sequencing Data*

| | MiSeq (2 Gb spec) | PGM (316–100 bp) |
|---|---|---|
| **Raw read length** | 2 × 151 (overlapping) | 100 (nominal) |
| **Processed read length** | 220* | 120 |
| **No. of reads (millions)** | 17.4 | 1.47 |
| **Yield (Gb)** | 2.633 (84.2% > Q30) | 0.197 (avg. Q20) |
| **Genome coverage** | 877× | 66× |
| **Assembler** | Velvet 1.1.06 (k=95) | Newbler 2.6 |
| **Genome assembly (bp)** | 3,053,394 | 2,980,036 |
| **N50** | 45,964 | 22,750 |
| **Raw SNP count** | 9,709 | 5,636 |

MiSeq generates about 10-fold more data with 10-fold higher quality than the Ion Torrent PGM, enabling more complete *de novo* assembly and greater genome coverage.

* Contributed by Dr. Tim Stinear and Dr. Torsten Seemann (Victorian Bioinformatics Consortium, Monash University, Australia)

† Processed read length trimmed to library insert size

Tim Stinear, Ph.D. and his group at the University of Melbourne use a variety of genomics and molecular biology techniques, including next-generation sequencing to understand how certain groups of bacteria evolve and cause disease in humans.

*Q: Can you tell us about your experience with the MiSeq system?*

**TS:** We sequenced a strain of *Enterococcus faecium*, a hospital pathogen commonly found in Australia and many other parts of the world, including the U.S. We have also sequenced this bacterium on the Ion Torrent Personal Genome Machine, the PGM, so we were able to make a direct comparison between the two platforms.

*Q: How did the output compare between the MiSeq system and the PGM?*

**TS:** MiSeq output is significantly better than that of the Ion Torrent in terms of the quantity and quality of sequence. Those two basic parameters are about 10-fold higher in MiSeq compared to the PGM. For example, the total number of sequence reads we obtained from MiSeq for this strain of *Enterococcus* was 17.4M. For the same strain run on the Ion Torrent, using the 316 chip, we obtained 1.47M reads. A bit more than 10-fold less. I'm not saying the PGM is under performing. It's actually performing to specification and I hasten to add that the Ion Torrent has been perfectly fine for all we've needed. But in a side-by-side comparison, MiSeq gives us ten times as much data.

In terms of actual DNA sequence, we got 2.6 Gb of sequence from MiSeq compared to 197 Mb from the PGM. Again, an order of magnitude more on the MiSeq system. The average quality score was Q30 on the MiSeq compared to Q20 on the Ion Torrent. We did some

basic analysis by mapping the reads we obtained from both platforms to our reference genome. We got 877 times coverage of our reference genome with MiSeq, compared to 66 with the Ion Torrent sequence. You are getting significantly more data and higher quality data from a MiSeq run.

*Q: How do the platforms compare on* de novo *assembly?*

**TS:** *De novo* assembly of MiSeq data using the Velvet assembler gave us a total genome size of 3.05 Mb. We performed the same process with our PGM data, using a requisite different assembler, and got a smaller genome of 2.98 Mb. One other *de novo* assembly metric, N50, which measures how well data was assembled, was 45,964 bp for MiSeq data, compared to 22,750 bp for the Ion Torrent data. The higher N50 of MiSeq data suggests that it performed better than the PGM with *de novo* assembly.

> "You are getting significantly more data and higher quality data from a MiSeq run."

*Q: Are the results of the comparison between MiSeq and the PGM what you expected?*

**TS:** It was what we'd been told to expect, and it certainly delivered. We got significantly more data with MiSeq. For us, we don't need 700 times coverage of a bacterial genome, or 877 times coverage, which is what we got. But now we can multiplex up to eight or more samples in a run. We can obtain good economies of scale by multiplexing, which we just can't do on the PGM using the 316 chip. It just doesn't have the coverage to allow us to multiplex to that extent. I can see running several bacterial genomes in one MiSeq run and still getting good, sufficient coverage for very accurate SNP calling or even *de novo* assembly.

*Q: How will the amount of data generated by the MiSeq system change your study design?*

**TS:** We'll be able to reduce the cost of each run. This means that we can now perform experiments that use genomics that we couldn't do before. For example, we just finished a project to sequence

> "MiSeq is very attractive to labs, like mine, that are taking control of genome sequencing themselves."

100 bacterial genomes that cost us about $50,000. That's a substantial amount of money when you're relying on grants. If we can reduce that cost 10-fold or even more, which is what we should be able to do with high-yield runs like the one we've just done on MiSeq, then we can contemplate more of these experiments. It's this broad-scale genomic approach that is unlocking important knowledge about how strains are developing antibiotic resistance. We can conceive much bolder experiments that really get to the heart of what the bacteria are doing when we throw antibiotics at them. We'll be doing more of the same, but generating a lot more information in the process, really answering some important questions.

*Q: In your opinion, what is the biggest benefit offered by the MiSeq system?*

**TS:** For us, it's the increased yield and high quality of the sequence that MiSeq delivers. That's pretty clear cut and quite impressive. MiSeq is very attractive to labs, like mine, that are taking control of genome sequencing themselves. We're no longer dependent on the queue, waiting for service providers to batch up their samples until they've got enough to do a run. Or, if they have problems with sample tracking, dealing with logistics errors, or runs that don't work. Having total control over these parameters is very important to me. It means that I can be much more efficient with generating data and get on to the real business of analyzing it.

Learn more about the MiSeq system at www.illumina.com/miseq

**Illumina** • 1.800.809.4566 toll-free (U.S.) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

FOR RESEARCH USE ONLY

© 2012 Illumina, Inc. All rights reserved.
Illumina, illumina*Dx*, BaseSpace, BeadArray, BeadXpress, cBot, CSPro, DASL, DesignStudio, Eco, GAIIx, Genetic Energy, Genome Analyzer, GenomeStudio, GoldenGate, HiScan, HiSeq, Infinium, iSelect, MiSeq, Nextera, Sentrix, SeqMonitor, Solexa, TruSeq, VeraCode, the pumpkin orange color, and the Genetic Energy streaming bases design are trademarks or registered trademarks of Illumina, Inc. All other brands and names contained herein are the property of their respective owners.
Pub. No. 770-2012-002 Current as of 03 May 2012

illumina®

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGAGC
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA