# BGI Takes on *De Novo* Assembly and More with the Genome Analyzer

From cucumbers and pandas to the human microbiome and the first Asian diploid genome, the Beijing Genomics Institute is advancing science at billions of bases per day using Illumina's next-generation sequencing technology.

## Introduction

Sequencing 20 gigabases per day, the Beijing Genomics Institute (BGI) at Shenzhen[1] is one of the most productive genome centers in the world. Goals of this powerhouse research facility include improving human health and quality of life, decoding large and small genomes from all five kingdoms, and enhancing agricultural research and industry. How do BGI researchers see themselves achieving their mission? According to Dr. Xiuqing Zhang, Director of Sequencing Platforms, "we are using sequencing-based strategies, relying almost completely on Illumina's sequencing technology. We generate about 19 gigabases per day, or 95% of our sequence data, from the 16 Genome Analyzers we keep in continuous production." Dr. Ruiqiang Li, Director of Bioinformatics, adds, "running these high-throughput systems is simple and cost-effective, allowing us to focus our resources on translating the data into meaningful information."

Using their Genome Analyzers, BGI researchers are pursuing a wide range of discovery applications such as de novo genome assembly, metagenomic analysis, and whole-genome and targeted resequencing. In addition, they collaborate with numerous international partners on large-scale projects and offer leading sequencing and data analysis solutions as a services provider.

## *De Novo* Assembly Using the Genome Analyzer

*De novo* genome assembly using traditional Sanger sequencing is expensive, slow, and labor intensive, limiting the pace at which new genomes are sequenced. Earlier this year, researchers at BGI reduced the cost of sequencing a new genome by 50% using a combination of Genome Analyzer data and traditional Sanger sequencing (Table 1). Zhang adds, "With the Genome Analyzer we can easily obtain and afford much deeper coverage, so we can achieve much higher data quality." Li elaborates, "Detailed analysis of our Genome Analyzer data revealed that on average we are seeing less than one error every 100kb at approximately 50× coverage."

This project was so successful that BGI and collaborators[2] have gone on to sequence and assemble larger genomes. This October, the team finished a draft sequence of the panda genome using data generated exclusively from Genome Analyzers. For this human-sized genome, Li reports that they were able to generate 50–75 bp paired-end reads at 50× coverage in less than one month. He adds, "For our cucumber genome project we were getting about 45–50 bp reads, with panda we are up to 75 bp reads. We see how good the data quality is at 75 bp, and we believe the Genome Analyzer has the ability to continue scaling up to longer reads."

**Figure 1: Beijing Genomics Institute**

The new BGI headquarters in Shenzhen (left) features a high-throughput sequencing facility that continuously runs 16 Genome Analyzers (right).

To match the rapidly increasing scale at which BGI can churn out raw sequence data, Li and his team of 100 bioinformaticians are developing novel analysis tools that can handle massive amounts of short-read sequences. Li says, "Right now we can assemble a large genome like the panda's in about two days using our *de novo* genome assembly software SOAPdenovo, a new part of our Short Oligonucleotide Alignment Program (SOAP)."

Confident in their ability to produce high-quality data and assemble new genomes, BGI has started ten more large plant and animal de novo sequencing projects.

## Path to Personalized Genomics

Besides being leaders in *de novo* sequencing and assembly, BGI researchers are paving the way for personalized genomics. In a recently published Nature paper[3], they describe the first complete sequencing of a diploid Asian genome. Using the Genome Analyzer, they achieved greater than 99.96% coverage of a Han Chinese genome with 22.5× single- and 13.5× paired-end reads.

Analyzing this data using their SOAP software, BGI researchers detected more than 417,000 SNPs not reported in the dbSNP database and identified over 135,000 small indels and 2,600 structural variants. With this contribution, BGI researchers believe they and the greater scientific community are one giant step closer to understanding population and individual genetic variation.

## Studying Disease Through Sequencing

An important aspect of human health and well-being is the complex association between the microbes that live on and in our bodies. As a member of MetaHIT, a 13-member international consortium, BGI is responsible for sequencing the gut microbiome isolated from thousands of individuals. With BGI's proven ability to assemble small and large genomes using Illumina's sequencing system (Tables 1 and 2), the consortium has set out to create a reference set of genes and genomes from intestinal microbes in healthy and diseased individuals. Through this ambitious metagenomics approach, researchers aim to uncover the role played by intestinal microbes in preservation of health and the etiology of a wide range of chronic diseases.

In another international collaboration designed to improve understanding of human health, BGI has set out to discover novel variations that correlate with increased risk of visceral obesity, type 2 diabetes, and hypertension. Teaming up with the Steno Diabetes Center in Denmark and Copenhagen University, BGI researchers are using a targeted re-sequencing strategy to capture and deeply sequence exons, 5' and 3' UTRs, and highly conserved genomic regions from 4,000 individuals. This case-control genome-wide association study is an alternative to array-based genotyping strategies, which, they believe, will offer novel and more comprehensive insight into the common and rare variation underlying these phenotypes of interest.

## Access to BGI's Expertise

With a roomful of Genome Analyzers, renowned researchers, and a team of bioinformaticians that is expected to grow to 200 members by next summer, BGI has positioned itself as a leading genomics solutions services provider. Zhang says one of the strongest advantages BGI offers is "that we are a genome center and a research center. We can rapidly generate data and we have a very strong bioinformatics team that can help customers with data analysis for all kinds of projects."

**Table 1: De Novo Sequencing and Assembly iof the Cucumber Genome with the Genome Analyzer**

| **Genome Analyzer** | | **Sanger Sequencing** | |
|---|---|---|---|
| **Sequencing** | | | |
| Coverage | 50× | Coverage | 4× |
| Read Length | ~50 bp | Fosmid Clone Physical Coverage | 10× |
| Paired-End Insert Size | 150 bp–2kb | Fosmid Clone Insert Size | 40kb |
| **Assembly** | | | |
| Contig N50 | ~5kb | Contig N50 | ~9kb |
| Scaffold N50 | ~60kb (Paired-end reads only) | Scaffold N50 | ~80kb (2–7kb plasmid ends) ~500kb (40kb fosmid ends) |
| **TOTAL** | | | |
| Length of Genome: | 290 Mb | | |
| Genomic Coverage: | 90% based on 90,307 UniGene fragments, 95% based on seven finished fosmids/BACs | | |

Data courtesy of Beijing Genomics Institute

**Table 2: Small Genome De Novo Assembly With the genome Analyzer**

| | Bacteria | Bacteria | Bacteria | Fungi |
|---|---|---|---|---|
| Genome Size | 6.6 Mb | 4.5 Mb | 4.5 Mb | 34 Mb |
| Sequencing Depth | 50× | 60× | 50× | 50× |
| Read Length | 35 bp | 35 bp | 45 bp | 45 bp |
| PE Insert | 130 bp | 130 bp | 130 bp/2kb | 200/500 bp |
| Scaffold N50 | 126kb | 31kb | 314kb | 90kb |

Data courtesy of Beijing Genomics Institute

## Reference


1. www.genomics.org.cn
2. In addition to researchers at BGI-Shenzhen, the current participants in this project consist of scientists from all around the globe, including researchers from the Kunming Institute of Zoology at the Chinese Academy of Sciences; the Institute of Zoology at the Chinese Academy of Sciences (Beijing); Chengdu Research Base of Giant Panda Breeding; the China Research and Conservation Center for the Giant Panda (Wolong); the Beijing Institute of Genomics, the Chinese Academy of Sciences; Beijing Genomics Institute (BGI); BGI-Hangzhou; the University of Alberta (Canada); Cardiff University (UK); Fudan University (Shanghai); Sichuan University; Southeast University (Nanjing); Sun Yat-Sen University (Guangzhou); the University of California at Berkeley; the University of Copenhagen; the University of Hong Kong; the University of Washington (Seattle); the World Wide Fund for Nature, China; and the Zoological Society of San Diego.
3. Wang J, Wang W, Li R, Li Y, Geng T, et al. (2008) The diploid genome sequence of an Asian individual. Nature 456: 60–65.


## Additional Information

Visit www.illumina.com or contact us at the address below to learn more about Illumina sequencing applications and products.

**Figure 2: Supercomputing Center**

Using a supercomputer that draws upon more than 100 CPUs and 500 GB of RAM, BGI researchers can assemble a human-sized genome in two days.

**Illumina, Inc.** • 9885 Towne Centre Drive, San Diego, CA 92121 USA • 1.800.809.4566 toll-free • 1.858.202.4566 tel • techsupport@illumina.com • illumina.com

FOR RESEARCH USE ONLY

© 2010 Illumina, Inc. All rights reserved.
Illumina, illumina*Dx*, Solexa, Making Sense Out of Life, Oligator, Sentrix, GoldenGate, GoldenGate Indexing, DASL, BeadArray, Array of Arrays, Infinium, BeadXpress, VeraCode, IntelliHyb, iSelect, CSPro, GenomeStudio, Genetic Energy, HiSeq, and HiScan are registered trademarks or trademarks of Illumina, Inc. All other brands and names contained herein are the property of their respective owners. Pub. No. 770-2008-020 Current as of 15 December 2008

illumina®

AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
TCAACGTACCGTAACGAACGTATCATTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTT
CGACGAAAAGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACC
AACGTACCATTAAGAGCTACCGTGCAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGACGAAAAGAATGATAA
AGAATGATAACAGTAACACACTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAAC
GATTACTTGATCCACTGATTCAACGTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGCTTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAGCAACGACG
CGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGTCTGTTAACCTTAAGATTACTTGATCCACTGATTCAACGTACCGTAACGAACGTATCAATTGAGACTAAATATTAACGTACCATTAAGAGCTACCGTGCAACGAAAAGAATGATAACAGTA