Jackman and Birol *Genome Biology* 2010, **11**:202
http://genomebiology.com/2010/11/1/202

Genome Biology

MINIREVIEW

Assembling genomes using short-read sequencing technology

Shaun D Jackman and İnanç Birol*

Abstract

Gigabase-scale genome assemblies are now feasible using short-read sequencing technology, bringing the cost of such projects below the million-dollar mark.

Moore's law is often used as a predictor in the informatics field for the growth of processing power based on the increase in the number of transistors in integrated circuits. It states that, according to the historical trend, this number doubles roughly every 2 years. A similar trend manifests itself in the number of base pairs deposited in the GenBank database, which had a mere 680,338 base pairs (bp) in its December 1982 release. Twenty-seven years later, that number reached 110,118,557,163 bp in its core repository, and 158,317,168,385 bp in the Whole Genome Shotgun sequencing project repository. This increase corresponds to a doubling roughly every 17 months over 3 decades. If this trend is sustained, by the mid-21st century we will have enough sequencing data to cover the genomes of the entire projected human population of 9 billion with more than fivefold redundancy, and have several exabases (10^{18} bp) remaining to sequence other species.

This gap between the rates of growth of informatics and sequencing throughput is exerting a considerable strain on the development of bioinformatics tools to process the sequencing data generated. Hence, we need ever faster and more accurate algorithms to keep up with this increasing gap, much as media-specific compression algorithms such as those used by MP3 and DVD filled the gap between the digital media revolution and its storage requirements. This article focuses on three large and two smaller *de novo* sequencing projects, all published within the last 6 months, with a special emphasis on the recently published giant panda genome [1], which used a so-called next-generation sequencing (next-gen) platform from Illumina.

Of the three major contenders in the next-gen sequencing field, the 454 platform from Roche generates the longest reads, and so its data are suited for *de novo* sequencing studies. However, it is also the most expensive per sequenced base to operate. The SOLiD platform from ABI sequences dinucleotides in color space rather than individual nucleotides. In color space representation, each of the 16 dinucleotides is assigned to one of four dyes. Each nucleotide is interrogated twice, which can improve accuracy, but the fact that each dye is shared by four dinucleotides complicates analysis. Hence, although less expensive to run, the SOLiD platform has mostly been used for resequencing studies. The Illumina platform is on a par with SOLiD in throughput and sequencing cost. However, it generates short-sequence data in nucleotide space and so is suitable for *de novo* sequencing. Although all three platforms were originally marketed for resequencing, with increasing read lengths, improving quality, and the development of protocols for paired-end reads, they are all now being used in *de novo* sequencing studies as well [1,2].

Recent *de novo* assemblies

Three genome projects recently published their results on the assembly and analysis of gigabase-scale genomes. For two of these, the B73 maize genome [3] and the domestic horse genome [4], researchers took the more conventional approach of sequencing clones using capillary technology. In contrast, researchers on the third project - the panda genome [1] - exclusively used Illumina's short-read technology to sequence the complete genome.

The B73 maize genome project followed the approach used by the original human genome project, using a physical map to select a minimum bacterial artificial chromosome (BAC) tiling path, and sequencing and assembling the selected clones to construct the *Zea mays* ssp. *mays* L. genome [3]. The high prevalence of repeat elements, constituting about 85% of the 10-chromosome, 2.3-gigabase genome, necessitated this rather conservative strategy. The project team assembled the 4x to 6x coverage data from capillary (Sanger) sequencing of a

*Correspondence: ibirol@bcgsc.ca
Genome Sciences Centre, British Columbia Cancer Agency, Vancouver, British Columbia V5Z 4E6, Canada

Table 1. Assembly statistics for maize, horse, panda, blue-stain fungus (*G. clavigera*) and *P. syringae* genomes and their cost

	B73 maize	Domestic horse*	Giant panda†	*G. clavigera*†	*P. syringae*†
Genome length	2.3 Gb	2.5-2.7 Gb	2.4-2.5 Gb	32.5 Mb	6.1 Mb
Sequencing technology/ies	Sanger	Sanger	Illumina	Sanger, 454, Illumina	Illumina
Number of contigs	125,325	55,316	198,274	3,361	1,346
Contig N50	40 kb	112 kb	40 kb	32 kb	11 kb
Number of scaffolds	61,161	9,687	81,469	2,322	71
Scaffold N50	76 kb	46 Mb	1.3 Mb	782 kb	317 kb
Estimated sequencing cost	$30 million	$15 million	$0.6 million	$100,000	$4,000

Contiguity statistics are calculated for *contigs and scaffolds 1 kb or longer and †contigs and scaffolds 100 bp or longer.

BAC library of 16,848 clones using Phrap [5], confirmed the assembly by BAC end sequencing, and refined it by sequencing 63 fosmid clones. The resulting assembly contains 125,325 contigs (61,161 scaffolds) with a contig (scaffold) N50 of 40 kb (76 kb), reconstructing 89% of the genome, with N50 denoting the weighted median; for a given assembly, half the genome is assembled in contigs larger than its N50. The estimated cost of the project, excluding the bioinformatics cost, is around US$30 million.

The project team for the domestic horse genome reported the second version of the draft *Equus caballus* genome [4], which has 31 pairs of autosomes and one pair of sex chromosomes. Genome length is estimated to be between 2.5 and 2.7 Gb. Sampling the genome of a thoroughbred mare, three clone libraries were generated: 4-kb and 10-kb inserts, and 40-kb fosmids, yielding sequence fold-coverages of 4.96x, 1.42x and 0.40x, respectively, on the capillary sequencing platform to a total of 6.8x coverage. To improve the contiguity of the draft assembly, the team used end sequences of 314,972 BACs derived from a half-brother of the sequenced mare. The horse genome was assembled by Arachne 2.0 [6] to obtain a contig (scaffold) N50 of 112 kb (46 Mb), with about 46% of the assembled genome in repetitive sequences. The use of a whole-genome shotgun approach reduced the cost of this project to half that of the maize project.

The above two projects used capillary sequencing data. In contrast, the giant panda genome project used Illumina sequencing data with an average read length of 52 bp and 73x coverage to assemble the *Ailuropoda melanoleuca* genome [1], which, at an estimated 2.4-2.5 Gb, is of comparable length to the other two genomes. The assembly was performed in two stages using SOAPdenovo [7]. In the first stage, the project team used paired-end sequencing data from 26 fragment libraries with nominal fragment sizes ranging from 110 bp to 570 bp. In the second stage, they used the pairing information from these libraries and from 11 long insert libraries of lengths 2 kb, 5 kb and 10 kb in successive iterations to scaffold the initial contigs. The resulting draft assembly is reported to have a contig (scaffold) N50 of 40 kb (1.3 Mb), reconstructing an estimated 92% of the genome. They also report that 36% of the panda genome is composed of transposable elements. The estimated cost of sequencing for this project is well under $1 million, making it 25 to 50 times more cost-efficient than the B73 maize and horse genome projects.

The extensive use of the Illumina short-read technology, and the longer reads from the 454 machine, for the *de novo* assembly of shorter genomes have been reported at recent conferences, and those studies have started to be published [2,8-10]. Table 1 compares the three genome projects described above and the recent genome assemblies of the filamentous fungus *Grosmannia clavigera* (blue-stain fungus) [2] and of the bacterial pathogen *Pseudomonas syringae* pathovar *tabaci* 11528 [8], both of which used next-gen sequencing.

Arguably, even if state-of-the-art sequencing protocols and bioinformatics tools are used, genomes with high repeat content, such as B73 maize, may still not yield to short-read sequencing. However, if the success and the quality of the paradigm used by the giant panda genome project team is validated and reproduced, new *de novo* sequencing projects for complex genomes will benefit from the reduction in cost as well as the time efficiencies offered by the short-read technologies.

Assembly tools

The enabling paradigm behind the *de novo* assembly of the giant panda genome is based on a de Bruijn graph representation of short sequence overlaps. A de Bruijn graph is a directed graph where vertices are strings of length k and edges represent overlaps of k-1 symbols, or nucleotides in the case of genome sequences. This approach was introduced to the field by Pevzner and co-workers with the Euler software [11], and was made popular by the software Velvet [12]. The first application of the technology for mammalian-sized genomes was demonstrated by Simpson *et al*. using ABySS [13].

Table 2. Effect of the choice of k-mer size on the single-end contig N50 for the giant panda assembly using ABySS 1.1.0

k-mer size	27*	30	32	34	35	36	37	38
Contig N50 (bp)	1,381	1,724	1,863	1,940	1,952	1,942	1,924	1,860

*The k-mer size used for the reported giant panda genome assembly [1].

These tools produce first-pass draft assemblies using a de Bruijn graph, followed by contig merging using paired-end information. For the latter stage, several groups have developed alternative ways of using the information in the read pairs. The ALLPATHS algorithm [14] uses the paired-end information in layers, starting with the large-fragment libraries to build 20 kb regions, called neighborhoods, around unique contigs, called seeds. The short-fragment pairs are then used to assemble the neighborhood, including the repetitive regions between the seeds. The panda assembly [1] also used a similar layered approach to using fragment libraries, but started with the shorter-fragment libraries and proceeded to the longer-fragment libraries.

The authors of Velvet suggest in a subsequent paper [15] that shorter-fragment libraries may be unnecessary. They argue that distance between two nearby contigs can be calculated by comparing their distances, estimated using a large-fragment library, to a third more distant contig. The distance between the two nearby contigs is logically the difference between their distances to the distant contig.

In ABySS [13], multiple libraries of different sized fragments are considered simultaneously. Distances between pairs of contigs are estimated using each fragment library on its own, and the most accurate distance estimates between contig pairs, which typically come from the library with the smallest fragments that span each distance, are retained. After smaller contigs have been merged into larger contigs, cases that could not be resolved in previous iterations are then reconsidered.

Producing the best possible de Bruijn graph assembly requires optimizing the fundamental parameter of k-mer size, which determines the length of significant overlaps for contig growth. Li *et al*. [1] report obtaining a single-end contig N50 of 1,483 bp using k = 27 with SOAPdenovo [7]. Reassembling their cleaned sequence data using ABySS 1.1.0 [13] without paired-end information, we obtained a contig N50 of 1,381 bp using k = 27, and an improved N50 of 1,952 using k = 35 (see Table 2). This shows that although the contiguity of the final panda assembly is already adequate for a genome of this size, it might be improved further by using a larger k-mer size.

The five genomes noted in this article have different levels of completeness, and the cost estimates we report are based on a number of assumptions and on the summary numbers reported in the respective studies. Furthermore, they exclude any costs related to the bioinformatics activities. As such, the sequencing costs are not directly comparable. Nevertheless, at face value, a pattern emerges that favors the short-read technology. This is not news, certainly, as it is the underlying premise of the next-gen platforms, yet the short-read assembly studies cited show that bioinformatics is catching up with the pace of data generation by these platforms. Thus, with software tools maturing and experimental protocols being refined, the number of genomes assembled with short reads will increase, and their size will expand.

Competing interests

The authors declare that they have no competing interests.

Published: 28 January 2010

References

1. Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, Zhang Z, Zhang Y, Wang W, Li J, Wei F, Li H, Jian M, Li J, Zhang Z, Nielsen R, Li D, Gu W, Yang Z, Xuan Z, Ryder OA, Leung FC, Zhou Y, Cao J, Sun X, Fu Y, *et al.*: **The sequence and *de novo* assembly of the giant panda genome.** *Nature* 2009, **463:**311-317.
2. Diguistini S, Liao NY, Platt D, Robertson G, Seidel M, Chan SK, Docking TR, Birol I, Holt RA, Hirst M, Mardis E, Marra MA, Hamelin RC, Bohlmann J, Breuil C, Jones SJ: **De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data.** *Genome Biol* 2009, **10:**R94.
3. Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, *et al.*: **The B73 maize genome: complexity, diversity, and dynamics.** *Science* 2009, **326:**1112-1115.
4. Wade CM, Giulotto E, Sigurdsson S, Zoli M, Gnerre S, Imsland F, Lear TL, Adelson DL, Bailey E, Bellone RR, Blöcker H, Distl O, Edgar RC, Garber M, Leeb T, Mauceli E, MacLeod JN, Penedo MC, Raison JM, Sharpe T, Vogel J, Andersson L, Antczak DF, Biagi T, Binns MM, Chowdhary BP, Coleman SJ, Della Valle G, Fryc S, Guérin G, *et al.*: **Genome sequence, comparative analysis, and population genetics of the domestic horse.** *Science* 2009, **326:**865-867.
5. **Phrap** [http://www.phrap.org/phredphrap/phrap.html]
6. Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP, Lander ES: **ARACHNE: a whole-genome shotgun assembler.** *Genome Res* 2002, **12:**177-189.
7. **SOAP: short oligonucleotide analysis package** [http://soap.genomics.org.cn/soapdenovo.html]
8. Studholme DJ, Ibanez SG, MacLean D, Dangl JL, Chang JH, Rathjen JP: **A draft genome sequence and functional screen reveals the repertoire of type III secreted proteins of *Pseudomonas syringae* pathovar *tabaci* 11528.** *BMC Genomics* 2009, **10:**395.
9. Steuernagel B, Taudien S, Gundlach H, Seidel M, Ariyadasa R, Schulte D, Petzold A, Felder M, Graner A, Scholz U, Mayer KF, Platzer M, Stein N: **De novo 454 sequencing of barcoded BAC pools for comprehensive gene survey and genome analysis in the complex genome of barley.** *BMC Genomics* 2009, **10:**547.
10. Farrer RA, Kemen E, Jones JD, Studholme DJ: **De novo assembly of the *Pseudomonas syringae* pv. *syringae* B728a genome using Illumina/Solexa short sequence reads.** *FEMS Microbiol Lett* 2009, **291:**103-111.
11. Pevzner PA, Tang H, Waterman MS: **An Eulerian path approach to DNA fragment assembly.** *Proc Natl Acad Sci USA* 2001, **98:**9748-9753.
12. Zerbino DR, Birney E: **Velvet: algorithms for de novo short read assembly**

using de Bruijn graphs. *Genome Res* 2008, **18:**821-829.
13. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I: **ABySS: a parallel assembler for short read sequence data.** *Genome Res* 2009, **19:**1117-1123.
14. Butler J, MacCallum I, Kleber M, Shlyakhter IA, Belmonte MK, Lander ES, Nusbaum C, Jaffe DB: **ALLPATHS: de novo assembly of whole-genome shotgun microreads.** *Genome Res* 2008, **18:**810-820.
15. Zerbino DR, McEwen GK, Margulies EH, Birney E: **Pebble and rock band: heuristic resolution of repeats and scaffolding in the Velvet short-read *de novo* assembler.** *PLoS ONE* 2009, **4:**e8407.

doi:10.1186/gb-2010-11-1-202
Cite this article as: Jackman SD, Birol I: **Assembling genomes using short-read sequencing technology.** *Genome Biology* 2010, **11:**202.