De novo assembly of the Pseudomonas syringae pv. syringae B728a genome using Illumina/Solexa short sequence reads
Rhys A. Farrer, Eric Kemen, Jonathan D.G. Jones & David J. Studholme
The Sainsbury Laboratory, Colney Lane, Norwich, UK

Abstract
Illumina's Genome Analyzer generates ultra-short sequence reads, typically 36 nucleotides in length, and is primarily intended for resequencing. We tested the potential of this technology for de novo sequence assembly on the 6 Mbp genome of Pseudomonas syringae pv. syringae B728a with several freely available assembly software packages. Using an unpaired data set, VELVET assembled >96% of the genome into contigs with an N50 length of 8289 nucleotides and an error rate of 0.33%. EDENA generated smaller contigs (N50 was 4192 nucleotides) and comparable error rates. SSAKE and VCAKE yielded shorter contigs with very high error rates. Assembly of paired-end sequence data carrying 400 bp inserts produced longer contigs (N50 up to 15 628 nucleotides), but with increased error rates (0.5%). Contig length and error rate were very sensitive to the choice of parameter values. Noncoding RNA genes were poorly resolved in de novo assemblies, while >90% of the protein-coding genes were assembled with 100% accuracy over their full length. This study demonstrates that, in practice, de novo assembly of 36-nucleotide reads can generate reasonably accurate assemblies from about 40× deep sequence data sets. These draft assemblies are useful for exploring an organism's proteomic potential, at a very economic low cost.

Introduction
‘Next Generation’ sequencing technologies such as Illumina’s Genome Analyzer (GA), ABI’s SOLiD and Roche’s 454 GS FLX enable a new approach to genome sequencing, generating gigabases of sequence data in a few days for a few thousand US dollars (Holt & Jones, 2008). The relatively long read lengths generated by 454 GS FLX technology make it suitable for de novo assembly of bacterial genomes. In contrast, the Illumina GA generates shorter sequence reads, typically 36 nucleotides in length, and was originally intended for resequencing genomes where a previously sequenced, very closely related reference genome is available. However, the GS FLX platform is on the order of 30 times more expensive than the Solexa platform for the same quantity of sequence (Holt & Jones, 2008). Illumina’s recent introduction of a paired-end reads option improves the prospects for de novo assembly using the Solexa platform. Therefore, we wanted to evaluate the potential of Illumina’s Solexa sequencing technology for determining the complete sequence of a 6 Mbp bacterial genome de novo, that is, without using a reference sequence to aid assembly.

Recently, several algorithms for de novo sequence assembly from short reads have been implemented as freely available software packages including VELVET, EDENA, SSAKE, VCAKE, EULER-SR and SHARCGS (see Materials and methods for references). A subset of these (VELVET and SSAKE) was able to exploit the additional information found in paired data sets. We wanted to know which of these assembly methods would give the best performance with a 42× deep Solexa data set of 36-nucleotide reads from a 6 Mb bacterial genome (Pseudomonas syringae pv. syringae B728a). Each of the assembly methods takes two or more input parameters, and so we also wanted to know how robust the assembly is with respect to the choice of parameter values. Finally, we aimed to assess whether the use of paired reads yielded a significant improvement in sequence assembly quality.

Materials and methods
The previously published sequence of P. syringae pv. syringae B728a (Feil et al., 2005) (Refseq: NC_007005) was downloaded from the NCBI FTP site (ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Pseudomonas_syringae_pv_B728a).
Solexa data were assembled using VELVET 0.6 (Zerbino & Birney, 2008), EDENA 2.1.1 (Hernandez et al., 2008), SSAKE 3.2 (Warren et al., 2007) and VCAKE 1.0 (Jecck et al., 2007).

We also attempted assemblies using EULER-SR 3.0 (Chaisson & Pevzner, 2007) and SHARCOS v1.2.8 (Dohm et al., 2007), but these consistently ran out of random access memory (a maximum of 32 Gb was available on our hardware) before completion. Butler et al. (2008) also propose an algorithm for assembly for paired-read data, but this is designed for multiple DNA libraries of differing insert lengths, and a production version of the software has not yet been released.

Solexa reads were aligned to the reference sequence using MAQ 0.6 (Li et al., 2008), allowing up to two mismatches. Duplicate reads were removed using MAQ’s rmdup function.

We used CGVIEW (Stothard & Wishart, 2005) to generate views of the genome.

**Library preparation for Illumina sequencing**

DNA was prepared from bacteria grown in L-medium using the Puregene Genomic DNA Purification Kit (Genta Systems Inc., Minneapolis) according to the manufacturer’s instructions. A library for Illumina Paired-End sequencing was prepared from 5 μg DNA using a Paired-End DNA Sample Prep Kit (Pe-102-1001, Illumina Inc., Cambridge, UK). DNA was fragmented by nebulization for 6 min at a pressure of 32 psi. For end-repair and phosphorylation, sheared DNA was purified using the QIAquick Nucleotide Removal Kit (Qiagen, Crawley, UK). The end-repaired DNA was A-tailed and adaptors were ligated according to the manufacturer’s instructions.

Size fractionation and purification of ligation products were performed using a 5% polyacrylamide gel run in Tris borate EDTA at 180 V for 120 min. Gel slices containing DNA in the 500 ± 10 bp range were cut. DNA was then extracted using 0.3 M sodium acetate and 2 mM EDTA (pH 8.0), followed by ethanol precipitation.

Using 18 PCR cycles with primers PE1.0 and PE2.0 supplied by Illumina, 5′ adaptor extension and enrichment of the library were performed. The library was finally purified using a QIAquick PCR Purification Kit and adjusted to a concentration of 10 nM in 0.1% Tween. The stock was kept at −20 °C until used.

**Sequencing**

The flow cell was prepared according to the manufacturer’s instructions using a Paired-End Cluster Generation Kit (Pe-103-1001) and a Cluster Station. Sequencing reactions were performed on a 1G GA equipped with a Paired-End Module (Illumina Inc.). Five picomolar of the library was used to achieve c. 20 000–25 000 clusters per tile.

**Metrics of assembly quality**

We used several metrics to assess the quality of each de novo genome assembly. In addition to the mean and median contig lengths, we included metrics for the breadth of genome coverage and the accuracy of the contig sequences, based on alignment of contigs against the published PsB genome. Contigs were aligned against this reference genome with BLAT (Kent, 2002) using a minimum identity threshold of 90%. For each contig, we considered only the ‘best’ hit against the genome. The best hit was defined as the hit with the highest number of matching nucleotides. Where there was a tie between hits with an equal number of matches, we considered the hit with the fewest gap nucleotides as the best.

To calculate the breadth of genome coverage, we counted how many BLAT hits overlapped each of the 6 093 698 nucleotide positions in the reference genome and counted how many of those positions were covered by at least one contig match.

There were several types of discrepancies between our de novo assemblies and the published PsB genome sequence, presumably due to errors in the assembly and possibly the presence of contaminating DNA. The discrepancies included unalignable sequences, single-base substitutions (mismatches in the alignments), insertions and deletions (gaps in the alignments). For each assembly, we counted the total number of inserted and deleted nucleotides, mismatches and unaligned nucleotides and then expressed this total number of errors as a percentage of the total assembly length (i.e. the sum of lengths of all the contigs). We also expressed the total number of single-base substitutions, insertions, deletions and ambiguous nucleotides as a fraction of the genome coverage.

The Solexa sequence data have been submitted to the Short Read Section of the European Nucleotide Archive at the European Bioinformatics Institute (http://www.ebi.ac.uk/emb/) as Project ID 32555. The raw Solexa data and the assemblies can also be downloaded from the authors’ website at http://tinyurl.com/68aeq3.

**Results and discussion**

**Solexa sequencing**

We generated 3 551 133 pairs of 36-nucleotide reads inserted with a mean length of 400 nucleotides separating the two members of a pair. This is equivalent to a total of 255 681 576 nucleotides of sequence representing c. 42 times the size of the 6 093 698 nucleotide genome.

**De novo assemblies using unpaired reads**

Although these sequence data consisted of paired reads, we first wanted to explore the potential to assemble unpaired...
data; hence, we treated the 3,551,133 pairs of reads as 7,102,266 unpaired reads for the first set of assemblies. For each of the assembly methods, we performed a ‘parameter scan,’ in which we ran the programs using a wide range of combinations of parameter values. We assessed the quality of each assembly using the metrics described in Materials and methods. Full results are listed in Supporting Information. The results are summarized in Fig. 1. The ‘best’ assembly should lie near the top left corner of the graph in Fig. 1, having a low error rate and a high N50 contig length. Each run of VELVET and EDENA took several minutes, whereas each SSAKE and VCAKE run took several days. Therefore, for SSAKE and VCAKE, we performed only a limited number of runs using parameter values close to those recommended in the software’s documentation. Both SSAKE and VCAKE generated assemblies with a very large number of errors. The vast majority of these errors were nucleotides deleted in the assemblies with respect to the reference genome sequence, that is, SSAKE and VCAKE frequently generated mis-assemblies where noncontiguous regions of the genome are placed together in contigs. Clearly, VELVET gave the best balance of contig length and accuracy for the assembly of unpaired reads, with EDENA a close second. VELVET gave relatively long contigs (N50 = 6963 nucleotides), with errors covering 0.2% of the length of the assembly (using hash-length = 21 and cutoff = 8). The majority of these errors were regions of the assembly that did not align to the reference. Of the 5089 protein-coding genes in the annotation of the Pss B728a genome, 4205 (82.6%) had full-length matches identical to a contig from this assembly.

It is possible that artefacts of Solexa sequencing technology (rather than the assembly process per se) might contribute to errors in the de novo assembly. To test this, we generated a simulated data set consisting of 6.7 million error-free 36-nucleotide sequences sampled at random from the Pss B728a genome (data not shown). This simulated data set yielded longer contigs (N50 = 13,771 nucleotides). When compared against the reference genome sequence, these contigs had no insertions and deletions and contained a total of 86 single-nucleotide mismatches. The 1369 contigs covered 97.41% of the reference genome sequence. Real data inevitably contain some errors and sampling biases (Dohm et al., 2008), leading to poorer assemblies than can be generated from idealized simulated data. However, even using unbiased, error-free data sets, at least 2.5% of the genome is unresolvable from 36-nucleotide reads.

Assembly of paired-end reads

It has been widely assumed that paired-end reads, whereby a relatively long (with respect to read-length) template is sequenced at both ends, can lead to a significant improvement in de novo assembly from short reads. The VELVET and

Fig. 1. Contig lengths and error rates for de novo assemblies of the Pss B728a genome from 7.1 million 36-nucleotide Solexa reads. Assemblies were generated using VCAKE, VELVET, SSAKE and EDENA, ignoring pairings between reads. Assemblies were also generated from the same data set using the paired reads (for VELVET and SSAKE only). Each assembly was aligned against the reference sequence (Refseq: NC_007005) using BLAT (Kent, 2002) with a minimum identity threshold of 90%. For each contig, only the best BLAT hit was considered. The total number of errors in an assembly comprised the number of unaligned nucleotides plus mismatches and nucleotides inserted into or deleted from the assembly compared with the reference. The number of errors is expressed as a percentage of the size of the assembly (i.e. the sum of all contig lengths).
SSAKE sequence assembly programs are able to exploit paired-end information. Therefore, we repeated our parameter scans using these programs, this time in the paired mode. Full results are listed in Supporting Information and summarized in Fig. 1. Assembly of paired reads using SSAKE generated relatively short contigs (N50 of > 1.5 kb) with very high error rates ( > 47%), which was no better than the performance with unpaired reads. On the other hand, using VELVET, the paired-read assembly yielded longer contigs (N50 up to 15.6 kb), but with higher error rates. The paired-read VELVET assembly giving the best balance between contig length and accuracy (hash-length = 23, cutoff = 9) had an

---

Fig. 2. Effects of parameter values and quantity of raw sequence data on VELVET assemblies of paired reads. VELVET takes two parameters: ‘hash length’ and ‘cutoff’. The 3.55 million pairs of Solexa reads were assembled using VELVET in the paired-end mode and a range of values for ‘hash length’ and ‘cutoff’. For each assembly, we calculated the N50 contig length (a and b) and the error rate (c and d) (see the legend for Fig. 1 for a description of error rate). We explored the effects of data quantity on the breadth of genome coverage and contig length (e and f, respectively) by assembling random subsets of the 3.55 million read-pairs over a range of parameter values.
N50 contig length of 12.5 kb, with errors covering 0.5% of the length of the assembly. Compared with the best unpaired VELVET assemblies, this represented a significant increase in contig length, but with increased error rates. Of the 5089 protein-coding genes annotated in Pss B728a, 4639 (91.1%) had identical full-length matches to a contig from this assembly. Supporting Information Table S7 lists the 451 genes that were not fully and accurately assembled. Many of these genes belong to families with multiple members encoded in the Pss B728a genome such as 16 encoding ABC transporters, five encoding response regulators and 11 encoding YD repeat proteins.

Fig. 3. Global view of de novo assemblies of the Pss B728a genome using EDENA and VELVET (in both paired and unpaired modes) from 7.1 million 36-nucleotide Solexa reads. EDENA takes two parameters: ‘overlap’ and ‘truncate’. We used values of 20 and 2, respectively. For the unpaired VELVET assembly, the chosen values were 21 and 8 for the ‘hash length’ and ‘cutoff’ parameters. For the paired VELVET assembly, the chosen values were 23 and 9 for the ‘hash length’ and ‘cutoff’ parameters. Each assembly was aligned against the reference sequence (Refseq: NC_007005) using BLAT (Kent, 2002) with a minimum identity threshold of 90%. For each contig, only the best BLAT hit was considered. Solexa paired reads were also aligned against the genome using MAQ (allowing two mismatches), and the relative depth of coverage was plotted.
The quantity of sequence data required for assembly

We assembled random subsets of the 3,551,133 pairs of 36-nucleotide reads using VELVET in the paired mode. Figure 2e and f shows that the breadth of genome coverage and the length of the contigs depend on the quantity of sequence data, i.e. the depth of genome coverage. Between five- and 30-fold depth of coverage, the increase in N50 contig length is approximately exponential, but the rate of increase appears to decline drastically above about 35-fold deep coverage. However, even at a relatively deep coverage, the choice of parameter values is critical and a poor choice will lead to an assembly with very short contigs.

Fig. 4. Examples of genomic regions absent from EDENA and VELVET assemblies. (a) Region of the Pss B728a genome encoding an integrase and a transposase, which was absent from all three assemblies. (b) Region of the Pss B728a genome encoding several noncoding RNA genes, which was absent from all three assemblies. RNA71, RNA72, RNA73 and RNA74 encode 23S rRNA, tRNA-ala, tRNA-Ile and 16s rRNA genes, respectively.
Unassembled regions of the genome

Figure 3 illustrates the coverage of the reference genome by examples of the ‘best’ EDENA assemblies, VELVET unpaired assemblies and VELVET paired-read assemblies. In all of the VELVET and EDENA assemblies, at least 3% of the reference genome was absent from the assembly. For the ‘best’ unpaired VELVET assembly, there were 842 unassembled regions between one and 5781 nucleotides in length; 42 of these were > 1 kb in length. When we examined the unassembled genomic regions, we found that they frequently encoded noncoding RNAs or mobile genetic elements. Table 1 demonstrates that although about 98% of the protein-coding regions of the genome are assembled, only about 52–55% of the RNA-encoding sequence and 91–92% of intergenic sequence are assembled. It is perhaps not surprising that noncoding RNAs are difficult to assemble from short reads, because they tend to contain inverted repeat sequences, which are necessary for the formation of RNA secondary structures (Tinoco & Bustamante, 1999). Furthermore, it is difficult to resolve the many tRNA genes and mobile elements that form families of closely related sequences (see Fig. 4). In practice, the different assembly methods show some complementarity in their ability to resolve ‘difficult-to-assemble’ genomic features (see examples in Fig. 5). This complementarity suggests that it might be possible to improve assembly quality by integrating contigs from several different assemblies.

Discrepancies with the published sequence of Pss B728a

As a byproduct of this study, we found eleven discrepancies between our Solexa deep sequence data and the published genome (Supporting Information). These discrepancies are supported both by alignments of unassembled Solexa reads and by the de novo assemblies (see Figs S1 and S2). Specifically, we found a single-nucleotide G → A substitution at position 5408961 of the reference genome. This would result in transformation of a threonine to isoleucine in RpoB’, the beta-prime subunit of RNA-dependent RNA polymerase.

A second single-nucleotide substitution (C → G) was found at position 951882, resulting in a substitution from typtophan to serine in pentapeptide-repeat protein Psyr_0840. A further nine substitutions were found, all of which fell within intergenic regions. The rest of the genome sequence was exactly concordant with the published sequence. These 11 substitutions may represent errors in the published sequence or may reflect real genetic differences between two samples of the strain.

Concluding remarks

The Illumina Solexa sequencing platform, aside from its primary application of resequencing, might have potential for cost-effective and rapid de novo sequencing, especially when used with paired-end libraries. We have critically assessed the qualities of de novo assemblies based on experimentally generated short sequence reads (36 nucleotides) using a variety of available assembly methods over a wide range of parameter values. For assembly of unpaired reads representing a 42 × deep coverage of the genome, we found that, in practice, VELVET gave the best balance of speed, contig length and accuracy while EDENA was nearly as good. VELVET was able to exploit paired-read information leading to a significant increase in contig length (N50 up to 15.5 vs. 8.3 kb), but with increased rates of mis-assembly errors. The cause of these errors is not clear, but the paired-read assembly algorithms are still under active development and it is likely that there will be further improvements in accuracy. In particular, ALLPATHS (Butler et al., 2008), an implementation of which is expected to be released soon, shows excellent potential. Also, since this analysis was performed, a new version of VELVET (version 0.7) has been released, which includes a much-improved algorithm for exploiting paired reads (D. Zerbino, pers. commun.).

Given sufficient depth of coverage (at least about 30 × deep), it is possible to assemble de novo contigs spanning at least 96% of the genome. The unassembled portion of the genome is enriched for noncoding RNA sequences and mobile genetic elements rather than protein-coding genes. Overall, this study demonstrates that the
Solexa technology combined with freely available sequence assembly software can be used to generate reasonably accurate assemblies from about 40 times deep sequence data sets. These draft assemblies typically consist of hundreds of contigs rather than a single closed chromosome and contain non-negligible numbers of errors. However, these draft assemblies are useful at the very least for exploring an organism’s proteomic potential, with > 90% of the protein-coding genes being assembled with 100% accuracy over their full length. This is ideal for applications such as ‘effector profiling’ (Almeida et al., 2009), at least for prokaryotic pathogens. It remains to be seen how well these technologies will perform with larger, more complex genomes.

Fig. 5. Examples of genomic regions showing differences in performance between assembly methods. (a) Region of the Pss B728a genome encoding a hypothetical gene, which was absent from the VELVET assemblies, but present in the EDENA assembly. (b) Region of the Pss B728a genome encoding an ABC transporter, which was incomplete in the unpaired assemblies (EDENA and VELVET), but intact in the paired VELVET assembly.
Acknowledgements

D.J.S. and J.D.G.J. are supported by the Gatsby Charitable Foundation. R.F. was supported by a scholarship funded by the UK Biotechnology & Biological Sciences Research Council’s competitive strategic grant to the John Innes Centre. E.K. is supported by a DFG research fellowship (KE 1509/1-1). The authors are grateful to Jodie Pike and Michael Burrell for excellent technical support, and to Sophien Kamoun, Daniel Zerbino, Daniel MacLean and Naveed Ishaque for useful discussions and comments on the manuscript.

Note added in proof

Using VELVET 0.7.18, which became available after this manuscript was accepted for publication, we assembled the Pss B728a Solexa dataset into contigs with N50 length of up to 165 kilobases covering 97.44% of the genome and the following error rates: 0.027% mismatches, 0.5% insertions and 2.5% deletions. This represents a very significant improvement in assembly quality.

References


Supporting Information

Additional Supporting Information may be found in the online version of this article:

Fig. S1. Discrepancy between published Pss B728a genome sequence and Solexa sequence in the rpoB' gene (Psyr_0840).

Fig. S2. Discrepancy between published Pss B728a genome sequence and Solexa sequence in the pentapeptide-repeat protein gene Psyr_0840.

Table S1. Summary statistics for (unpaired) assemblies using VELVET.

Table S2. Summary statistics for (unpaired) assemblies using VELVET.

Table S3. Summary statistics for (unpaired) assemblies using SSAKE.

Table S4. Summary statistics for (unpaired) assemblies using SSAKE.

Table S5. Summary statistics for (unpaired) assemblies using SSAKE.

Table S6. Summary statistics for (unpaired) assemblies using SSAKE.

Table S7. Genomic regions that failed to be assembled.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.