Chapter 1

Sequencing and Genome Assembly Using Next-Generation Technologies

Niranjan Nagarajan and Mihai Pop

Abstract

Several sequencing technologies have been introduced in recent years that dramatically outperform the traditional Sanger technology in terms of throughput and cost. The data generated by these technologies are characterized by generally shorter read lengths (as low as 35 bp) and different error characteristics than Sanger data. Existing software tools for assembly and analysis of sequencing data are, therefore, ill-suited to handle the new types of data generated. This paper surveys the recent software packages aimed specifically at analyzing new generation sequencing data.

Key words: Next-generation sequencing, Genome assembly, Sequence analysis

1. Introduction

Recent advances in sequencing technologies have resulted in a dramatic reduction of sequencing costs and a corresponding increase in throughput. As data produced by these technologies is rapidly becoming available, it is increasingly clear that software tools developed for the assembly and analysis of Sanger data are ill-suited to handle the specific characteristics of new generation sequencing data. In particular, these technologies generate much shorter read lengths (as low as 35 bp), complicating repeat resolution during both de novo assembly and while mapping the reads to a reference genome. Furthermore, the sheer size of the data produced by the new sequencing machines poses performance problems not previously encountered in Sanger data. This is further exacerbated by the fact that the new technologies make it possible for individual labs (rather than large sequencing centers) to perform high-throughput sequencing experiments, and these labs do not have the computational infrastructure commonly

David Fenyö (ed.), *Computational Biology*, Methods in Molecular Biology, vol. 673,
DOI 10.1007/978-1-60761-842-3_1, © Springer Science+Business Media, LLC 2010

available at large sequencing facilities. In this paper we survey software packages recently developed to specifically handle new generation sequencing data. We briefly overview the main characteristics of the new sequencing technologies and the computational challenges encountered in the assembly of such data; however, a full survey of these topics is beyond the scope of our paper. For more information, we refer the reader to other surveys on sequencing and assembly (1–3).

We hope the information provided here will provide a starting point for any researcher interested in applying the new technologies to either de novo sequencing applications or to resequencing projects. Due to the rapid pace of technological and software developments in this field we try to focus on more general concepts and urge the reader to follow the links provided in order to obtain up-to-date information about the software packages described.

2. Sequencing Technologies

Before discussing the software tools available for analyzing the new generation sequencing data we briefly summarize the specific characteristics of these technologies. For a more in-depth summary, the reader is referred to a recent review by Mardis (1).

2.1. Roche/454 Pyrosequencing

The first, and arguably most mature, of the new generation sequencing technologies is the pyrosequencing approach from Roche/454 Life Sciences. Current sequencing instruments (GS FLX Titanium) can generate in a single run ~500 Mbp of DNA in sequencing reads that are ~400 bp in length (approximately 1.2 million reads per run), while the previous generation instruments (GS FLX) generate ~100 Mbp of DNA in reads that are ~250 bp in length (approximately 400,000 reads per run). Initial versions of mate-pair protocols are also available that generate paired reads spaced by approximately 3 kbp.

The main challenge in analyzing 454 data is the high error-rate in homopolymer regions – sections of DNA comprised of a single repeated base. The 454 sequencing approach is based on a technique called pyrosequencing (4) wherein double-stranded DNA is synthesized from single-strand templates (DNA fragments being sequenced) through the iterative addition of individual nucleotides, and the incorporation of a nucleotide is detected by the emission of light. When encountering a run of multiple identical nucleotides in the template DNA, the amount of light emitted should be proportional to the length of this homopolymer run. This correspondence, however, is nonlinear due to limitations of the optical device used to detect the signal. As a result, the length of homopolymer runs is frequently misestimated by the 454 instrument, in particular for long homopolymer runs.

A 454 sequencing instrument can output copious information, including raw images obtained during the sequencing process. For most purposes, however, it is sufficient to retain the 454 equivalent of sequence traces, information stored in .SFF files. These files contain information about the sequence of nucleotide additions during the sequencing experiment, the corresponding intensities (normalized) for every sequence produced by the instrument and the results of the base-calling algorithm for these sequences. Each called base is also associated with a phred-style quality value (log-probability of error at that base), providing the same information as available from the traditional Sanger sequencing instruments. Note, however, that homopolymer artifacts also affect the accuracy of the quality values – Huse et al. (5) show that the quality values decrease within a homopolymer run irrespective of the actual confidence in the base-calls.

Due to the long reads and availability of mate-pair protocols, the 454 technology can be viewed as a direct competitor to traditional Sanger sequencing and has been successfully applied in similar contexts such as de novo bacterial and eukaryotic sequencing (6, 7) and transcriptome sequencing (8).

2.2. Solexa/Illumina Sequencing

The Solexa/Illumina sequencing technology achieves much higher throughput than 454 sequencing (~1.5 Gbp/run) at the cost, however, of significantly smaller read lengths (currently ~35 bp). Initial mate-pair protocols are available for this technology that generate paired reads separated by ~200 bp and approaches to generate longer libraries are currently being introduced. While the reads are relatively short, the quality of the sequence generated is quite high, with error rates of less than 1%. The sequencing approach used by Solexa relies on reversible terminator chemistry and is, therefore, not affected by homopolymer runs to the same extent as the 454 technology. Note that homopolymers, especially long ones, cause problems in all sequencing technologies, including Sanger sequencing.

The analysis of Solexa/Illumina data poses several challenges. First of all, a single run of the machine produces hundreds of gigabytes of image files that must be transferred to a separate computer for processing. In addition to the sheer size of the data generated, a single Solexa run results in ~50 million reads leading to difficulties in analyzing the data, even after the images have been processed. Finally, the short length of the reads generated complicates de novo assembly of the data due to the inability to span repeats. The short reads also complicate alignment to a reference genome in resequencing applications, both in terms of efficiency and due to the increased number of spurious matches caused by short repeats.

Analogous to 454 sequencing, the output from an Illumina sequencing instrument contains a wealth of information, including raw image data that could be reprocessed to take advantage of new base-calling algorithms. In practice, however, these data are rarely retained due to the large memory requirements. For most applications it is sufficient to use the sequence trace information encoded in an SRF file – a newly developed format for encoding new generation sequencing data. When just the sequence and quality information are needed, these data are usually stored in a FASTQ file (an extension of the FASTA format that combines sequence and quality data) and represents quality values in a compressed (one character per base) format.

2.3. ABI/SOLiD Sequencing

The ABI/SOLiD technology generates data with characteristics similar to that generated by Solexa/Illumina instruments, albeit at higher throughput (~3 Gbp/run). Challenges in image storage and processing that are present with Solexa technology are therefore also there for the ABI/SOLiD instrument. The latter, however, integrates computer hardware with the sequencing machine, eliminating the need to transfer large image files for analysis purposes.

A major challenge in analyzing SOLiD data stems from the sequencing-by-ligation approach used in this technology. Specifically, the sequencing of a DNA template is performed by iteratively interrogating pairs of positions in the template with semi-degenerate oligomers of the form NNNACNNN, where N indicates a degenerate base. Each oligomer is tagged with one of four colors, allowing the instrument to "read" the sequence of the template. Note, however, that each color is associated with four distinct pairs of nucleotides, complicating the determination of the actual DNA sequence. In fact, the sequence of colors observed by the instrument during the sequencing process is not sufficient to decode the DNA sequence – rather it is necessary to also know the first base in the sequence (the last base within the sequencing adapter). The lack of a direct correspondence between the sequencing signal and the DNA sequence complicates the analysis of SOLiD data in the presence of errors. A single sequencing error (missing or incorrect color) can result in a "frame-shift" that affects the remainder of the DNA sequence decoded by the instrument. Note that this phenomenon is similar to that encountered during gene translation from three-letter codons. Due to this property of SOLiD data, most software tools attempt to operate in "color space" in order to avoid having to consider all possible frame-shift events during data analysis. This also makes it difficult to apply SOLiD data in de novo assembly applications.

File formats for representing SOLiD data are still being developed and a SOLiD-specific extension to the SRF format is expected in the near future.

2.4. Others

We presented in more detail the three technologies outlined above because they are the only technologies currently deployed on a large scale within the community. It is important to note, however, that new sequencing technologies are being actively developed and several will become available in the near future. For example, Helicos Biosciences have recently reported the sale of the first instruments of a high-throughput, single-molecule (requiring no amplification) sequencing technology (9). Also, recently, Pacific Biosciences have described a new technology characterized by substantially longer read lengths and higher throughputs than the technologies currently available (10). These advances underscore the dynamic nature of research on DNA sequencing technologies, and highlight the fact that the information we provide in this article is necessarily limited to the present and might become partly obsolete in the near future.

2.5. NCBI Short Read Archive

The large volumes of data generated by the new technologies as well as the rapidly evolving technological landscape are posing significant challenges to disseminating and storing this data. To address these challenges and provide a central repository for new generation data, the NCBI has established the Short Read Archive, an effort paralleling the successful Trace Archive – a repository of raw Sanger sequence information. The Short Read Archive (http://ncbi.nlm.nih.gov/Traces/sra) already contains a wealth of data generated through the 454 and Illumina technologies, including data from the 1,000 Genomes project – an effort to sequence the genomes of 1,000 human individuals. In addition to being a data repository, the Short Read Archive is actively involved in efforts to standardize data formats used to represent new generation data, efforts that resulted in the creation of the . SFF format (454) and the .SRF format meant to become a universal format for representing sequence information.

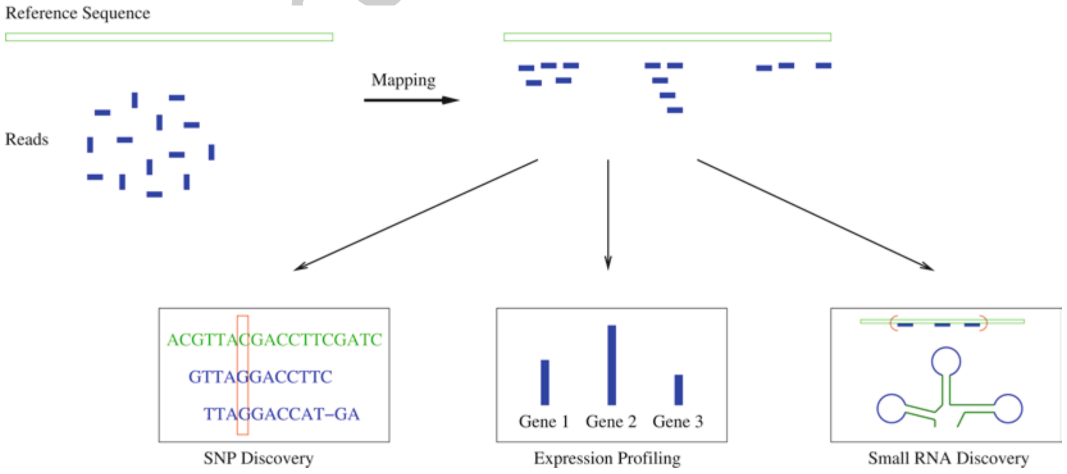
3. Assembly Programs

The assembly of sequences from a shotgun-sequencing project is typically a challenging computational task akin to solving a very large one-dimensional puzzle. Several assembly programs have been described in the literature (such as **Celera Assembler** (11), **ARACHNE** (12, 13), and **PHRAP** (14)) and have been successfully used to assemble the genomes of a variety of organisms – from viruses to humans. These programs were designed when Sanger sequencing was the only technology available and were therefore tailored to the characteristics of the data. With the advent of new technologies there has been a flurry of efforts to cope with the characteristics of the new datasets. An important consideration is the reduced read length and the limited form of mated read libraries. These make the assembly problem even more difficult as we discuss in Subheading 3.2. What the new technologies do offer is the ability to sequence genomes to high redundancy (every base in the genome is represented in many reads) and in a relatively unbiased manner. Managing the corresponding flood of information effectively is an important challenge facing new computational tools.

3.1. Mapping and Comparative Assembly

In many sequencing projects, an assembled genome of a related organism is available and this can dramatically simplify the assembly task. The task of assembly is then often translated to one of matching sequences to the reference genome and de novo assembly of just the polymorphic regions from the unmatched reads. This strategy has been widely used for resequencing projects (15). It has also been used to assemble closely related bacterial strains (16). The strategy of sequencing and mapping to a reference genome has also been used in a variety of other applications – from discovering novel noncoding RNA elements (17) to profiling methylation patterns (18) (see Subheading 4 for more examples). The general pipeline for these applications is outlined in Fig. 1 with mapping of reads to a reference being an important common component.

In recent years, several programs have been developed to handle the challenge of mapping a large collection of reads onto a reference genome while accounting for sequencing errors and polymorphisms. These programs often trade-off flexibility in matching policy – how many mismatches and indels they can handle – in



this figure will be printed in B/W

[AU1] Fig. 1. Read mapping and its applications. Mapping programs are widely used to align reads to a reference while allowing some flexibility in terms of mismatches and indels and a policy for handling ambiguous matches. The matches are then processed in different ways depending on the application of interest.

order to improve computational efficiency and the size of their memory footprints. For the longer reads from Sanger and 454 sequencing, programs such as **MUMmer** (19) and **BLAT** (20) provide the right trade-off between efficiency and flexibility in matching policy; they allow many mismatches and indels and are correspondingly slower.

The large volume of reads from Illumina and SOLiD sequencing have spurred the development of a new set of tools. In order to efficiently handle large amounts of short-read data, these programs attempt to find the right balance between alignment sensitivity and performance. Performance is generally achieved by constructing efficient indexes of either the reference genome or the set of reads, allowing the rapid identification of putative matches which are then refined through more time-intensive algorithms. Further improvements in performance arise from the handling of reads that map within repeat regions – most programs only report a few (or even just one) of the possible mappings. Finally, these programs allow only a few differences between a read and the reference genome and frequently do not allow indels. The choice of alignment program and corresponding parameters ultimately depends on the specific application: for example, in SNP discovery it is important to allow for differences between the reads and the reference beyond those expected due to sequencing errors, while in CHIP-seq experiments (21), exact or almost-exact alignments are probably sufficient. We review some of the popular mapping programs below.

MAQ (22) (it stands for *Mapping and Assembly with Quality*) is designed to map millions of very short reads accurately to a reference genome by taking into account the quality values associated with bases. In addition, MAQ also assigns to every mapped read, an assessment of the quality of the mapping itself. This information allows MAQ to perform well in SNP-calling applications. MAQ constructs an index of the reads, therefore its memory footprint is proportional to the size of the input and the authors recommend performing the alignment in chunks of two million sequences. MAQ only allows for mismatches in the alignment (no indels) and randomly assigns a read to one of several equally good locations when multiple alignments are possible (though this behavior can be modified through command-line parameters). Furthermore, MAQ can utilize mate-pair information in order to disambiguate repetitive matches. MAQ was originally developed for Illumina data, though it can also handle SOLiD sequencing using a transformation of the reference sequence into color space.

The inputs to MAQ are provided in FASTA (reference) and FASTQ (reads) formats and the output consists of a list of matches with associated qualities. MAQ also includes modules for SNP calling, as well as a viewer **Maqview** that provides a graphical representation of the alignments.

The source code is available for download at http://maq.sourceforge.net under the GNU Public License.

SOAP (23) which stands for *Short Oligonucleotide Alignment Program* indexes the reference instead of the reads and therefore its memory footprint should be constant irrespective of the number of reads processed. Its alignment strategy allows alignments with one short indel (1–3 bp) in addition to mismatches between the read and the reference. Its treatment of reads with multiple alignments can be tuned through command-line parameters. Like MAQ, SOAP also provides support for mate-pairs, and includes a module for SNP calling. In addition, SOAP provides an iterative trimming procedure aimed at removing low quality regions at the ends of reads, as well as specialized modules for small RNA discovery and for profiling of mRNA tags.

SOAP is available for download at http://soap.genomics.org.cn as a Linux executable.

SHRiMP (unpubl.) is one of the first alignment programs specifically targeted at SOLiD data, though Illumina data can also be processed. This program uses a spaced-seed index followed by Smith–Waterman alignment to provide full alignment accuracy and flexibility. Since SHRiMP uses a full dynamic programming approach for alignment instead of heuristics, it is considerably slower than MAQ or SOAP, even though the implementation of the Smith–Waterman algorithm is parallelized through vectored operations supported by Intel and AMD processors. In addition to SOLiD data, SHRiMP now also supports data generated by the Helicos technology.

SHRiMP is available from http://compbio.cs.toronto.edu/shrimp as both source code and precompiled binaries.

Bowtie (24) is the first of a new-breed of fast and memory-efficient short-read aligners based on the compact Burrows–Wheeler index (both MAQ and SOAP now offer BWT-based indices), used to index the reference sequence. While following the same alignment policies as MAQ and SOAP, Bowtie is typically more than an order of magnitude faster, aligning more than 20 million reads per hour to the human genome on a typical workstation. Unlike other aligners, Bowtie allows the index for a genome to be precomputed, reducing the overall alignment time and making it easier to parallelize the alignment process. Furthermore, the indexing structure used is space-efficient, requiring just over 1 GB for the entire human genome.

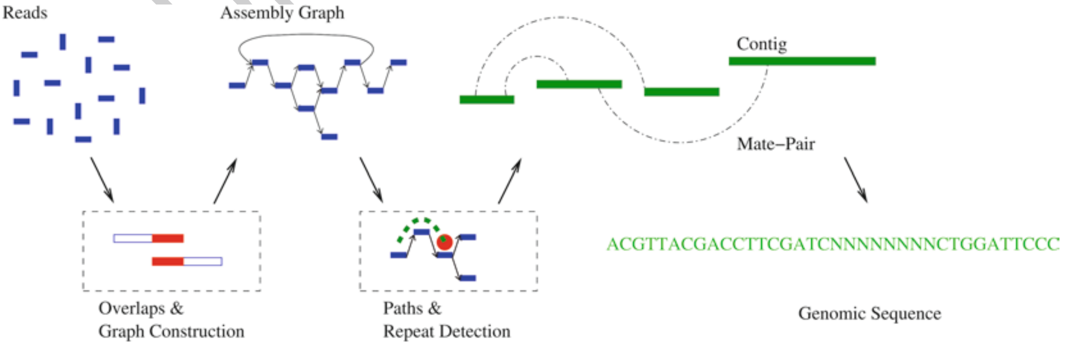
Bowtie is available at http://bowtie-bio.sourceforge.net as an open-source package together with an associated program called **TopHat** to map splice junctions from RNA-seq experiments.

Other programs. Several other programs are available for the alignment of short reads and more will likely become available in the near future. Among the most widely used is **Eland**, the aligner provided by Illumina with their sequencing instruments. This program is proprietary and unpublished and we cannot provide any additional information on its performance. Another commercial offering is SX-OligoSearch from Synamatix, a program that is provided together with the specialized hardware necessary to run it. Finally, **SeqMap** (25), **RMAP** (http://rulai.cshl.edu/rmap), and **ZOOM** (26) are other aligners that have been recently reported in the literature. The latter is based on a spaced-seed index and appears to be very efficient; however, the code can currently only be obtained by direct request from the authors.

Postalignment analysis. Several of the alignment programs described above provide additional modules for postprocessing the set of alignments in order to identify SNPs, discover small RNAs or analyze transcriptome profiling data or splicing patterns. The resulting alignments can also be provided as input to a comparative assembler such as **AMOScmp** (27) to construct local assemblies of the set of reads in a "template-guided" fashion. In a recent work, Salzberg et al. (16) demonstrated the use of this tool with Solexa data in bacterial sequencing, and have also proposed an approach to leverage similarity at the amino acid level to construct gene-centric assemblies of the data.

3.2. De Novo Assembly

In the absence of a reference genome, researchers typically rely on de novo assembly programs to reconstruct the sequences represented in the shotgun sequencing reads. An overview of the assembly process is presented in Fig. 2. The de novo assembly of a genome relies on the assumption that two reads that overlap significantly in their sequence are likely to represent neighboring segments of a genome. This assumption is, however, violated when the overlapping sequence is part of a repetitive region in the genome and recognizing such regions is an important part of



this figure will be printed in b/w

[AU1] Fig. 2. Overview of de novo assembly. De novo assembly programs typically use the overlap between reads to construct a graph structure. After some simplifications of the graph, unambiguous paths in the graph are used to reconstruct contiguous sequences (contigs). Information such as the presence of mate-pair links between contigs may also allow the construction of gapped sequences (scaffolds).

genome assembly. The short read lengths of the new sequencing technologies entail that even short genomic repeats (that tend to be more frequent as well) can introduce ambiguities into the assembly process. As a result, the output from the assemblers is often a highly fragmented picture of the genome. Despite these limitations, several sequencing projects have successfully used short-read technologies.

Due to its longer read lengths 454 sequencing is a popular approach for de novo sequencing of bacterial genomes and increasingly for larger eukaryotic genomes as well. The **Newbler** assembler (http://www.roche-applied-science.com) that is distributed with 454 instruments has been used to assemble 454 data in several sequencing projects. The Newbler assembler supports mate-pairs and can do comparative as well as hybrid assembly (see Subheading 3.2.2). With sufficient read coverage (>20×) it generally produces accurate and conservative assemblies containing few misassemblies due to repeats. The consensus sequence of the resulting contigs is of high quality despite the relatively common sequencing errors in homopolymer regions within 454 sequence data.

The Celera Assembler (http://wgs-assembler.sourceforge.net) (11), originally developed for the assembly of large mammalian genomes from Sanger data, has recently been extended to allow assembly of 454 data as well as of mixtures of 454 and Sanger data. Both Celera Assembler and Newbler directly accept 454 data as input in .SFF format and produce outputs in both FASTA format and in several more detailed assembly formats, including the popular ACE format used by the phred-phrap-consed suite of programs.

For assembling the even shorter reads from Illumina and SOLiD, several assembly programs have recently been developed. In order to deal with the large volume of reads, early programs such as SSAKE (28), VCAKE (29), and SHARCGS (30) relied on a simple greedy approach to assembly. However, two new programs (Edena and Velvet) that are based on a graph-theoretic approach to assembly were shown to produce more accurate and larger assemblies and we describe them in more detail here. Note that even the best assemblers generate highly fragmented assemblies from short-read data (~35 bp), leading to contigs in the range of just a few to tens of thousands of basepairs instead of hundreds to millions of bases common in 454 and Sanger assemblies. These programs are, thus, better suited for the assembly of targeted regions, such as individual genes, or data generated in CHIP-seq experiments.

Edena (31), which stands for Exact DE Novo Assembler was designed for assembling Illumina sequences based on a classic overlap-layout-consensus assembly framework. To avoid spurious overlaps, Edena restricts itself to exact matches and this also allows it to compute overlaps efficiently. In addition, Edena incorporates some heuristic approaches to simplify the overlap graph and only linear sections of the graph are assembled into sequences. In addition to this conservative approach, Edena also allows for a non-strict mode which can create longer sequences but with an increased chance of incorrect assembly. Experiments with ~35 bp Illumina reads for a few bacterial genomes have shown that Edena can very accurately assemble them into sequences that are on average a few kilobases long. These assemblies were performed on a desktop computer with 4 GB of memory and in less than 20 min. The Edena program is available for download at http://www.genomic.ch/edena as a linux executable (an experimental windows executable is also available). The program takes a FASTA or FASTQ file of reads as input and produces a FASTA file of assembled sequences as output. It also allows the user flexibility in choosing an overlap size, trimming of reads and filtering of short assemblies. The current implementation of Edena does not handle mated reads.

Velvet (32) is an open-source program that uses a de Bruijn graph-based approach to assembly (33). Correspondingly, while the graph construction step is simplified, the program relies on several error-correction heuristics to improve the structure of the graph. The program also has a module to use mated reads to disambiguate some repeat structures and join contigs. Using simulated mated reads, this approach was shown to produce much longer contigs in prokaryotic genomes. Velvet is available for download at http://www.ebi.ac.uk/~zerbino/velvetand has been tested on Linux, MacOS, and Windows systems with Cygwin. It accepts reads in FASTA as well as FASTQ format and its output is a set of assembled sequences in a FASTA file as well as an AMOS compatible assembly file. Velvet also allows the user to choose the overlap size and can filter sequences that have a low read coverage.

ABySS (34) is a new parallelized sequence assembly program based on the de Bruijn graph approach that can efficiently do de novo assembly of relatively large datasets (billions of reads). It also allows for the use of paired-end information to produce longer contigs. ABySS can take in reads in FASTA format and produce contigs in FASTA format and is available as an open-source package at http://www.bcgsc.ca/platform/bioinfo/software/abyss.

Other software. The **Minimus** assembler (35) which is part of the AMOS package of open-source assembly tools (http://amos.sourceforge.net), like Edena is based on an overlap-layout-consensus framework for assembly. Due to its modular structure, Minimus can easily be adapted for various sequencing technologies and a version for Illumina sequences (http://amos.sourceforge.net/docs/pipeline/minimus.html) is also available. The **ALLPATHS** program (36) is a new short-read assembler, based on the Eulerian assembly strategy, that aims to explicitly present assembly ambiguity to the user in the form of a graph. The authors plan to release a production version of the program soon.

3.2.1. Scaffolding

While programs such as Edena and Minimus do not directly handle information about mated reads, a scaffolding program such as **Bambus** (37) can use this information to stitch together contigs into larger sections of the genome (aka scaffolds). Bambus is available at http://amos.sourceforge.net/docs/bambus. Note that Newbler, Celera Assembler, Velvet, and ALLPATHS can use mate-pair information directly to guide the assembly process and generate larger contigs and scaffolds (where some of the intervening regions can be ambiguous).

Another promising new approach to scaffold short-read sequences is based on Optical Mapping technology (38) (http://www.opgen.com). Optical Maps are a form of restriction maps where genomic DNA is fragmented using a restriction enzyme and the fragment sizes are measured. In optical mapping, both fragment sizes and the order in which they occur within the genome can be determined. This genome-wide map provides an ideal reference to determine the order of the sequences assembled from a shotgun sequencing project. For a typical prokarytoic sequencing project more than 90% of the genome can be scaffolded using these maps into a single genome-wide scaffold (39). The open-source **SOMA** package is specifically designed to map short-read assemblies onto one or more optical maps and scaffold them and is available for download and as a webservice at http://www.cbcb.umd.edu/soma.

3.2.2. Hybrid Assembly

As discussed in Subheading 2 the various sequencing technologies have different advantages and disadvantages, some of which are complementary. Correspondingly there is an interest in constructing hybrid assemblies that, for example, can combine mated reads from one technology with high coverage reads from another. In recent work, Goldberg et al. (40) showed that high-quality assemblies of microbial genomes can be obtained in a cost-effective manner using a Sanger-454 hybrid approach. In order to assemble the data, they relied on an ad-hoc approach where sequences assembled from 454 reads using Newbler were shredded in-silico before assembly with other Sanger reads using the Celera Assembler. Recently, more carefully tuned versions of the Celera Assembler (41) and Newbler have been released that can perform true Sanger-454 hybrid assemblies. Assemblers that are fine-tuned to incorporate various other mixtures of sequence data are still an active area of research.

4. Applications

The dramatic reduction in the cost of sequencing using next-generation technologies has led to widespread adoption of sequencing as a routine research technique. On the one hand, the traditional use of sequencing, i.e., to reconstruct the genomes of a range of model organisms and pathogenic microbes has received a boost. Researchers are now looking to sequence several individuals and strains of the same species to understand within species variation. While in some cases these related genomes can be assembled based on the reference, in others, de novo assembly programs are required. The other popular use for sequencing has been as a substitute for common array based techniques for studying mRNA expression, transcription-factor-binding sites and methylation patterns, among others. These applications rely on read mapping programs followed by application-specific analysis as shown in Fig. 1. Here we highlight a few of the diverse collection of problems that are being impacted by the availability of new sequencing platforms and the computational tools to analyze the data.

4.1. Variant Discovery

High-throughput sequencing has enabled researchers to study the extent of variability in our genomes both in terms of single base mutations as well as larger structural changes that are much more common than we once believed. The current approach for these studies is to map reads to a reference genome to detect changes from the reference and is aided by the array of mapping programs available as detailed in Subheading 3.1. In addition to general-mapping programs such as MAQ that are well-suited for SNP calling and have a built-in procedure to do so, there are other programs that are specifically designed for SNP calling. The **PyroBayes** program is one such tool that was developed to specifically take into account the characteristics of 454 reads, given a set of read mappings (available at http://bioinformatics.bc.edu/marthlab/PyroBayes). The **ssahaSNP** program is another (http://www.sanger.ac.uk/Software/analysis/ssahaSNP), that performs both the mapping and SNP calling for Illumina sequences and also includes a module for indel discovery. Tools to detect larger structural variations based on mated reads are an active area of current research (42).

4.2. Metagenomics

Metagenomics studies where a collection of organisms are sequenced together are, in principle, the prime application for new sequencing technologies that enable cheap and relatively bias-free sequencing. The crucial impediment however is the ability to assemble and annotate the short reads from these technologies.

In recent years, several programs have been designed for classification and gene-finding in 454 reads. Programs such as **MEGAN** (43) and **CARMA** (44), in particular, have had some success using translated BLAST searches to classify and annotate 454 reads. Ideally, annotation and gene-finding of metagenomic sequences would be preceded or done in tandem with assembly of the short reads. Assembly algorithms tuned for metagenomics datasets, especially those based on short reads are, however, still being developed (45), and there is much work to be done in this direction. As read lengths for Illumina and SOLiD sequencing increase, metagenomics studies are likely to more widely use these technologies in the future.

4.3. Small RNA Discovery

Sequencing in combination with computational filters provides an ideal approach to discover various noncoding small RNAs whose regulatory importance is increasingly being apparent. The dramatic decrease in the cost of sequencing has enabled researchers to detect even rarely transcribed elements and fortunately the length of these elements is small enough for them to be profiled with short reads. Analyzing reads from such a project involves mapping them to a reference genome and using annotations on the genome and RNA structure prediction programs to filter out uninteresting loci. The analysis pipeline typically needs to be tailored to the sequencing platform used and the kinds of small RNA that the researchers are interested in. The **miRDeep** package (17), for example, was specifically designed to analyze sequences for microRNAs and more such packages are likely to be made available in the near future. In another recent work, Moxon et al. (46) describe a set of webservices to analyze large datasets of plant small RNA sequences to find various plant-specific RNA elements.

5. Conclusion

In this chapter, we provided an overview of the tools available for assembling and analyzing the new generation sequencing technologies that have emerged in recent years. As these technologies have only recently become available and research on new technologies is ongoing, the associated software tools are also continuously being adapted. Therefore, the information provided here is just a starting point, rather than a complete survey of the field. We hope this information provides the necessary background and we urge the reader to follow the links provided within the text in order to obtain up-to-date information about the software tools described here.

References

1. Mardis, E. R. (2008) The impact of next generation sequencing technology on genetics, *Trends Genet* **24**, 133–141.
2. Pop, M. (2004) Shotgun sequence assembly, *Adv Comput* **60**, 193–248.
3. Pop, M., and Salzberg, S. L. (2008) Bioinformatics challenges of new sequencing technology, *Trends Genet* **24**, 142–149.
4. Ronaghi, M., Uhlen, M., and Nyren, P. (1998) A sequencing method based on real-time pyrophosphate, *Science* **281**, 363–365.
5. Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L., and Welch, D. M. (2007) Accuracy and quality of massively parallel DNA pyrosequencing, *Genome Biol* **8**, R143.
6. Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L., Jarvie, T. P., Jirage, K. B., Kim, J. B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P., Begley, R. F., and Rothberg, J. M. (2005) Genome sequencing in microfabricated high-density picolitre reactors, *Nature* **437**, 376–380.
7. Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y. J., Makhijani, V., Roth, G. T., Gomes, X., Tartaro, K., Niazi, F., Turcotte, C. L., Irzyk, G. P., Lupski, J. R., Chinault, C., Song, X. Z., Liu, Y., Yuan, Y., Nazareth, L., Qin, X., Muzny, D. M., Margulies, M., Weinstock, G. M., Gibbs, R. A., and Rothberg, J. M. (2008) The complete genome of an individual by massively parallel DNA sequencing, *Nature* **452**, 872–876.
8. Emrich, S. J., Barbazuk, W. B., Li, L., and Schnable, P. S. (2007) Gene discovery and annotation using LCM-454 transcriptome sequencing, *Genome Res* **17**, 69–73.
9. Harris, T. D., Buzby, P. R., Babcock, H., Beer, E., Bowers, J., Braslavsky, I., Causey, M., Colonell, J., Dimeo, J., Efcavitch, J. W., Giladi, E., Gill, J., Healy, J., Jarosz, M., Lapen, D., Moulton, K., Quake, S. R., Steinmann, K., Thayer, E., Tyurina, A., Ward, R., Weiss, H., and Xie, Z. (2008) Single-molecule DNA sequencing of a viral genome, *Science* **320**, 106–109.
10. Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., Peluso, P., Rank, D., Baybayan, P., Bettman, B., Bibillo, A., Bjornson, K., Chaudhuri, B., Christians, F., Cicero, R., Clark, S., Dalal, R., Dewinter, A., Dixon, J., Foquet, M., Gaertner, A., Hardenbol, P., Heiner, C., Hester, K., Holden, D., Kearns, G., Kong, X., Kuse, R., Lacroix, Y., Lin, S., Lundquist, P., Ma, C., Marks, P., Maxham, M., Murphy, D., Park, I., Pham, T., Phillips, M., Roy, J., Sebra, R., Shen, G., Sorenson, J., Tomaney, A., Travers, K., Trulson, M., Vieceli, J., Wegener, J., Wu, D., Yang, A., Zaccarin, D., Zhao, P., Zhong, F., Korlach, J., and Turner, S. (2009) Real-time DNA sequencing from single polymerase molecules, *Science* **323**, 133–138.
11. Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., Kravitz, S. A., Mobarry, C. M., Reinert, K. H., Remington, K. A., Anson, E. L., Bolanos, R. A., Chou, H. H., Jordan, C. M., Halpern, A. L., Lonardi, S., Beasley, E. M., Brandon, R. C., Chen, L., Dunn, P. J., Lai, Z., Liang, Y., Nusskern, D. R., Zhan, M., Zhang, Q., Zheng, X., Rubin, G. M., Adams, M. D., and Venter, J. C. (2000) A whole-genome assembly of Drosophila, *Science* **287**, 2196–2204.
12. Batzoglou, S., Jaffe, D. B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J. P., and Lander, E. S. (2002) ARACHNE: a whole-genome shotgun assembler, *Genome Res* **12**, 177–189.
13. Jaffe, D. B., Butler, J., Gnerre, S., Mauceli, E., Lindblad-Toh, K., Mesirov, J. P., Zody, M. C., and Lander, E. S. (2003) Whole-genome sequence assembly for mammalian genomes: Arachne 2, *Genome Res* **13**, 91–96.
14. Green, P. (1994) Statistical aspects of imaging, *Stat Methods Med Res* **3**, 1–3.
15. Hillier, L. W., Marth, G. T., Quinlan, A. R., Dooling, D., Fewell, G., Barnett, D., Fox, P., Glasscock, J. I., Hickenbotham, M., Huang, W., Magrini, V. J., Richt, R. J., Sander, S. N., Stewart, D. A., Stromberg, M., Tsung, E. F., Wylie, T., Schedl, T., Wilson, R. K., and Mardis, E. R. (2008) Whole-genome sequencing and variant discovery in C. elegans, *Nat Methods* **5**, 183–188.
16. Salzberg, S. L., Sommer, D. D., Puiu, D., and Lee, V. T. (2008) Gene-boosted assembly of a novel bacterial genome from very short reads, *PLoS Comput Biol* **4**, e1000186.

17. Friedlander, M. R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., and Rajewsky, N. (2008) Discovering microRNAs from deep sequencing data using miRDeep, *Nat Biotechnol* **26**, 407–415.

18. Down, T. A., Rakyan, V. K., Turner, D. J., Flicek, P., Li, H., Kulesha, E., Graf, S., Johnson, N., Herrero, J., Tomazou, E. M., Thorne, N. P., Backdahl, L., Herberth, M., Howe, K. L., Jackson, D. K., Miretti, M. M., Marioni, J. C., Birney, E., Hubbard, T. J., Durbin, R., Tavare, S., and Beck, S. (2008) A Bayesian deconvolution strategy for immunoprecipitation-based DNA methylome analysis, *Nat Biotechnol* **26**, 779–785.

19. Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S. L. (2004) Versatile and open software for comparing large genomes, *Genome Biol* **5**, R12.

20. Kent, W. J. (2002) BLAT–the BLAST-like alignment tool, *Genome Res* **12**, 656–664.

21. Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. (2007) Genome-wide mapping of in vivo protein-DNA interactions, *Science* **316**, 1497–1502.

22. Li, H., Ruan, J., and Durbin, R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores, *Genome Res* **18**, 1851–1858.

23. Li, R., Li, Y., Kristiansen, K., and Wang, J. (2008) SOAP: short oligonucleotide alignment program, *Bioinformatics* **24**, 713–714.

24. Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, *Genome Biol* **10**, R25.

25. Jiang, H., and Wong, W. H. (2008) SeqMap: mapping massive amount of oligonucleotides to the genome, *Bioinformatics* **24**, 2395–2396.

26. Lin, H., Zhang, Z., Zhang, M. Q., Ma, B., and Li, M. (2008) ZOOM! Zillions of oligos mapped, *Bioinformatics* **24**, 2431–2437.

27. Pop, M., Phillippy, A., Delcher, A. L., and Salzberg, S. L. (2004) Comparative genome assembly, *Brief Bioinform* **5**, 237–248.

28. Warren, R. L., Sutton, G. G., Jones, S. J., and Holt, R. A. (2007) Assembling millions of short DNA sequences using SSAKE, *Bioinformatics* **23**, 500–501.

29. Jeck, W. R., Reinhardt, J. A., Baltrus, D. A., Hickenbotham, M. T., Magrini, V., Mardis, E. R., Dangl, J. L., and Jones, C. D. (2007) Extending assembly of short DNA sequences to handle error, *Bioinformatics* **23**, 2942–2944.

30. Dohm, J. C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2007) SHARCGS, a fast and highly accurate short-read assembly algorithm for de novo genomic sequencing, *Genome Res* **17**, 1697–1706.

31. Hernandez, D., Francois, P., Farinelli, L., Osteras, M., and Schrenzel, J. (2008) De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer, *Genome Res* **18**, 802–809.

32. Zerbino, D. R., and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs, *Genome Res* **18**, 821–829.

33. Pevzner, P. A., Tang, H., and Waterman, M. S. (2001) An Eulerian path approach to DNA fragment assembly, *Proc Natl Acad Sci U S A* **98**, 9748–9753.

34. Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., and Birol, I. (2009) ABySS: a parallel assembler for short read sequence data, *Genome Res* **19**, 1117–1123.

35. Sommer, D. D., Delcher, A. L., Salzberg, S. L., and Pop, M. (2007) Minimus: a fast, lightweight genome assembler, *BMC Bioinformatics* **8**, 64.

36. Butler, J., MacCallum, I., Kleber, M., Shlyakhter, I. A., Belmonte, M. K., Lander, E. S., Nusbaum, C., and Jaffe, D. B. (2008) ALLPATHS: de novo assembly of whole-genome shotgun microreads, *Genome Res* **18**, 810–820.

37. Pop, M., Kosack, D. S., and Salzberg, S. L. (2004) Hierarchical scaffolding with Bambus, *Genome Res* **14**, 149–159.

38. Samad, A., Huff, E. F., Cai, W., and Schwartz, D. C. (1995) Optical mapping: a novel, single-molecule approach to genomic analysis, *Genome Res* **5**, 1–4.

39. Nagarajan, N., Read, T. D., and Pop, M. (2008) Scaffolding and validation of bacterial genome assemblies using optical restriction maps, *Bioinformatics* **24**, 1229–1235.

40. Goldberg, S. M., Johnson, J., Busam, D., Feldblyum, T., Ferriera, S., Friedman, R., Halpern, A., Khouri, H., Kravitz, S. A., Lauro, F. M., Li, K., Rogers, Y. H., Strausberg, R., Sutton, G., Tallon, L., Thomas, T., Venter, E., Frazier, M., and Venter, J. C. (2006) A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes, *Proc Natl Acad Sci U S A* **103**, 11240–11245.

41. Miller, J. R., Delcher, A. L., Koren, S., Venter, E., Walenz, B. P., Brownley, A., Johnson, J., Li, K., Mobarry, C., and Sutton, G. (2008)

Aggressive assembly of pyrosequencing reads with mates, *Bioinformatics* **24**, 2818–2824.

42. Lee, S., Cheran, E., and Brudno, M. (2008) A robust framework for detecting structural variations in a genome, *Bioinformatics* **24**, i59–i67.
43. Huson, D. H., Auch, A. F., Qi, J., and Schuster, S. C. (2007) MEGAN analysis of metagenomic data, *Genome Res* **17**, 377–386.
44. Krause, L., Diaz, N. N., Goesmann, A., Kelley, S., Nattkemper, T. W., Rohwer, F., Edwards, R. A., and Stoye, J. (2008) Phylogenetic classification of short environmental DNA fragments, *Nucleic Acids Res* **36**, 2230–2239.
45. Ye, Y., and Tang, X. (2008) in "Proceedings of the Seventh Annual International Conference on Computational Systems Bioinformatics", Stanford, CA.
46. Moxon, S., Schwach, F., Dalmay, T., Maclean, D., Studholme, D. J., and Moulton, V. (2008) A toolkit for analysing large-scale plant small RNA datasets, *Bioinformatics* **24**, 2252–2253.