ERNEST ORLANDO LAWRENCE
BERKELEY NATIONAL LABORATORY

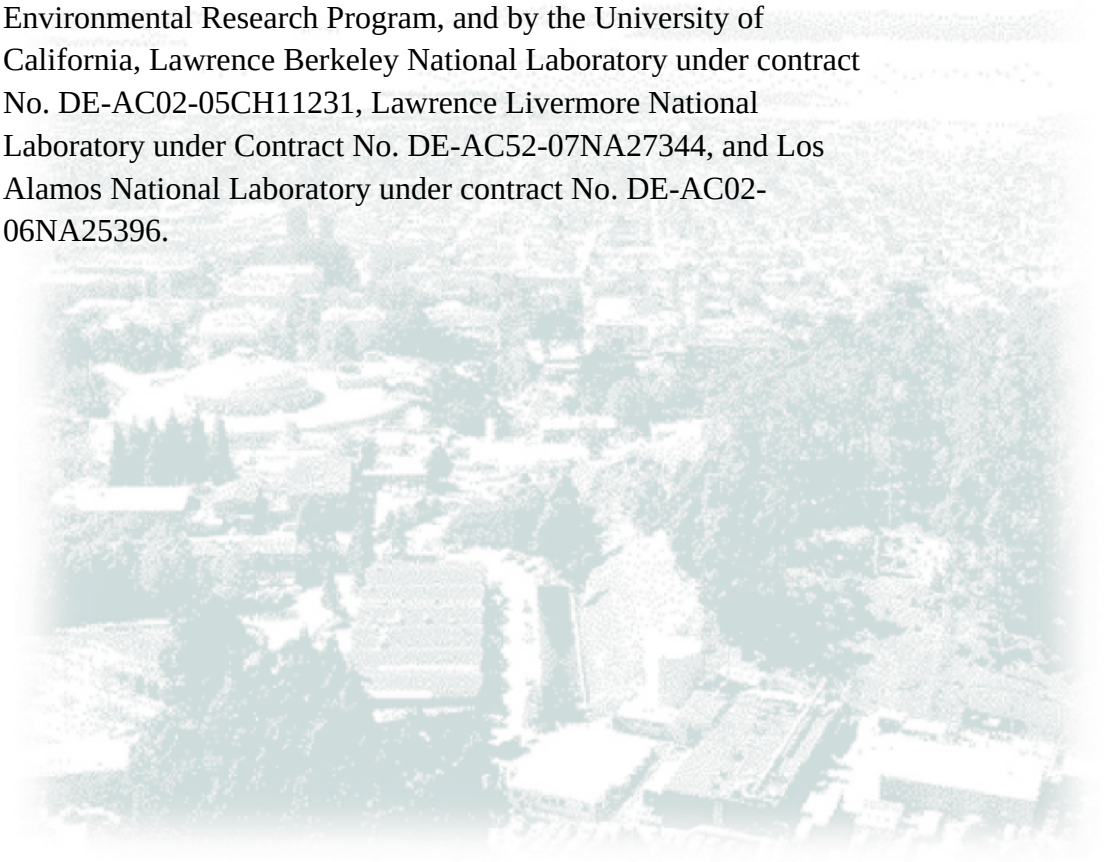
Title: Genome sequence databases (overview): Sequencing and assembly

Author(s): Alla Lapidus[1],

Author Affiliations:[1]Department of Energy, Joint Genome Institute, Walnut Creek, CA 945981

Date: 06/10/09

Funding: This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.



GENOME SEQUENCE DATABASES (OVERVIEW): : SEQUENCING AND ASSEMBLY

Alla L. Lapidus

Affiliations and contact information:

Alla L. Lapidus
Joint Genome Institute
2800 Mitchell Dr.
Walnut Creek, CA 94598, US
e-mail: alapidus@lbl.gov
phone: 1 (925) 296-5778
fax: 1 (925) 296-5850

Summary

From the date its role in heredity was discovered, DNA has been generating interest among scientists from different fields of knowledge: physicists have studied the three dimensional structure of the DNA molecule, biologists tried to decode the secrets of life hidden within these long molecules, and technologists invent and improve methods of DNA analysis. The analysis of the nucleotide sequence of DNA occupies a special place among the methods developed. Thanks to the variety of sequencing technologies available, the process of decoding the sequence of genomic DNA (or whole genome sequencing) has become robust and inexpensive. Meanwhile the assembly of whole genome sequences remains a challenging task. In addition to the need to assemble millions of DNA fragments of different length (from 35 bp (Solexa) to 800 bp (Sanger)), great interest in analysis of microbial communities (metagenomes) of different complexities raises new problems and pushes some new requirements for sequence assembly tools to the forefront. The genome assembly process can be divided into two steps: draft assembly and assembly improvement (finishing). Despite the fact that automatically performed assembly (or draft assembly) is capable of covering up to 98% of the genome, in most cases, it still contains incorrectly assembled reads. The error rate of the consensus sequence produced at this stage is about 1/2000 bp. A finished genome represents the genome assembly of much higher accuracy (with no gaps or incorrectly assembled areas) and quality (~1 error/10,000 bp), validated through a number of computer and laboratory experiments.

Keywords: DNA sequencing, whole-genome shotgun assembly, contig, scaffold, read, misassembly, genome finishing

Defining statement

Putting together data produced during the sequencing stage of a genomic project is comparable to solving a complicated puzzle made of several million pieces. Thanks to the combined efforts of enthusiastic software developers and biologists studying genome structure and functionality, a number of successful algorithms and approaches for genome assembly have been created. High quality genomic assembly lays a solid foundation for functional genome analysis.

Glossary

Assembler –computer program that pieces together overlapping reads to reconstruct the original sequence.

Captured gap (a sequence gap) –unsequenced area between two contigs spanned by at least one subclone.

Contig – " a set of gel readings that are related to one another by overlap of their sequences. All gel readings belong to one and only one contig, and each contig contains at least one gel reading. The gel readings in a contig can be summed to form a contiguous consensus sequence and the length of this sequence is the length of the contig." (Staden, 1980).

Finishing – the process of improving a draft assembly composed of shotgun sequencing reads, resolving misassembled regions, closing sequence gaps and validating low quality regions to produce a highly accurate finished DNA sequence (less than 1 error in 10,000 bp).

Gap - unsequenced area between two contigs.

Paired (sister) reads – sequences generated from both ends of a DNA fragment. Such reads are oriented towards each other and the distance between them is equal to the template length.

Quality Score - the probability of a wrong base-call. A phrap quality score of X corresponds to an error probability of approximately 10-X/10 (a score of 30 means that the error probability is 1/1000 or 99.9% accuracy for a base in the assembled sequence).

Read - gel reading generated during the DNA sequence process.

Repeats – sequences of varying lengths found in multiple copies in the genome.

Scaffold – a group of ordered and oriented contigs.

Sequencing by Synthesis (SBS) – a sequencing approach, which involves multiple parallel micro-sequencing addition events occurring on a surface, where data from each round is detected by imaging.

SNP- single-nucleotide polymorphism.

Uncaptured (physical gap) – unsequenced area between two contigs with no subclones spanning it.

See Also the Following Articles

DNA Sequencing and Genomics
Genome Sequence Databases (overview): : Genomic, Construction of Libraries
Metagenomics
Databases and Online Resources

I. DNA sequencing

I. DNA sequencing

The knowledge of the particular order of nucleotide bases of genomic DNA is widely used both in fundamental biological research and in many applications. This information plays a significant role in medical and forensic studies and is of great value for the pharmaceutical industry.

The era of DNA sequencing began about 30 years ago, when two methods of primary DNA structure determination appeared almost simultaneously. One of them was developed by A. Maxam and W. Gilbert at Harvard University and was based on the use of chemical modifications of DNA. The other method was published by F. Sanger and A. Coulson from Cambridge and is called the chain-termination method. Since then significant changes have been introduced both in the Sanger method (which has been completely automated and therefore has found wider application than Gilbert's method) and in a variety of other sequencing technologies.

Modern approaches are currently aimed at significantly reducing the costs associated with sequence determination. In 2004, the National Human Genome Research Institute (NHGRI) set a task for the global scientific community to lower the cost of sequencing of an individual human genome to $1000, with an intermediate goal being to reduce the cost of sequencing a mammalian-sized genome to $100,000.

This initiative resulted in rapid proliferation of new sequencing technologies including (i) pyrosequencing (Margulies, 2005), (ii) base-by-base sequencing by synthesis (Balasubramanian S, 2001), (iii) sequencing by ligation (Albrecht, 2000), (iv) nanopore sequencing (Park, 2007; Rhee, 2007; Wang, 2001) and (v) single-molecule sequencing by synthesis in real time (Levene, 2003). These newest technologies are currently at different stages of development and implementation: while pyrosequencing and sequencing by synthesis (SBS) approaches are mature enough to be used in second-generation sequencing machines already available on the market, other technologies are still at the early stages of testing (Strezoska, 1991).

As compared to the traditional Sanger sequencing, the greatest advantage of the first three approaches is very high level of parallelization, as they are capable of performing up to 10^7 reactions in one experiment. In addition, the tiny volume of the individual wells (picoliters) in which sequencing reactions are running, and high data density (ten thousand times higher, than in case of the latest microelectrophoresis-based capillary instruments) significantly decreases the volume of reagents necessary to

perform the reactions. Another positive aspect of the new methods is the ability to circumvent the step of DNA cloning and propagation in E. coli cells, thus avoiding the problems of biased genome coverage due to the presence of genomic areas, which are hard or even impossible to clone in E. coli (so called "unclonable" areas). The downside of the newest sequencing technologies, as compared to the Sanger approach, is the length of the reads produced. The 454 Genome Sequencer 20 System instrument produces reads of about 110 bp, although the read length is expected to increase to 200 bp (Genome Sequencer FLX System) and become even longer when further upgrades are made. Meanwhile, technologies using SBS (Braslavsky, 2003) and sequencing by ligation (Applied Biosystem SOLiD; polony based approaches (Shendure, 2005)) produce reads of 25-35 bp long. Such short reads impose certain restrictions on the use of new technologies: while 454 sequencing has been successfully applied to de-novo genome sequencing resulting in a gapped assembly with an error rate of about 1/3000 – 1/5000 bp, the ultra short reads produced by Solexa technology (www.illumina.com) on the other hand appear to be useful for re-sequencing purposes only, when a high quality reference sequence is available for alignment (Bentley, 2006). The use of 25-35 bp reads for de-novo genome sequencing appears to be very problematic.

Among the numerous methods still on the drawing board, special attention should be given to real-time sequencing of DNA molecules (Middendorf, 1992). Such methods use very small amounts of genomic DNA and require supersensitive detection methods. The success of these methods critically depends on the quality of genomic DNA used in the experiment, because any damage to DNA in the process of its isolation would result in corrupted sequence. Theoretically, the read length produced by this type of technology is equal to that of a full genome. A number of development teams have used nanopores (Chen, 2004) or an electron microscope (Glover, 2004) for signal detection. When nanopore-based detection is used, DNA moves through 1.5 nm-wide pores in an electric field and a detector measures the change in conductance within the pore (Winters-Hilt, 2003). ZS Genetics, Inc. is working on a single molecule sequencing method, which uses a single-stranded DNA molecule to form a complementary strand via the incorporation of labeled nucleotides thus making the DNA molecule visible under a Transmission Electron Microscope (TEM). The company claims that their new system will be capable of sequencing DNA strands of 20,000 bp or longer. If this level of performance is achieved, read lengths of this size will greatly simplify subsequent steps of processing raw genome sequencing data and genome assembly.

The list of companies and university-based groups of researchers working on improving sequencing technology goes on. Some of them have already become commercially successful; others are still only approaching this goal, while some will probably fail (Table1). It is not clear today if any of the new methods will be able to completely replace Sanger sequencing. It is possible that a combination of different approaches will allow the production of high quality sequence in good time and at very low cost.

<Table 1 near here>

A. Microbial Genome Project

Decoding the genome of a microbe (or Microbial Genome Project) is a complex scientific task composed of complete genome sequence determination and functional analysis of the genes of the organism being sequenced. Thus each genomic project consists of sequencing, assembly and annotation stages.

When starting a project, it is useful to find out the size of the genome, its G+C content, number of chromosomes (some microbial cells contain more than one chromosome) and potential presence of plasmids (circular or linear extrachromosomal elements). Choosing the most appropriate sequencing and assembly strategy is as important as deciding which tool to use for the subsequent functional analysis of the genome.

B. Sequencing Strategies

Three sequencing strategies have been used in genomic projects. The first one is based on the use of ordered collections of overlapping BAC (bacterial artificial chromosome) or YAC (yeast artificial chromosome) clones (Krzywinski, 2004; Kunst, 1997). To follow this approach one should make a large-insert library (BAC, YAC), map clones to the genome using fingerprinting (Cole, 1998; Krzywinski, 2004) or hybridization (Azevedo, 1993), pick the minimal set of clones covering the entire genome, sequence and assemble each clone from this set separately. This strategy appeared to be very labor intensive and was not widely used.

In 1995 the Whole-Genome Shotgun (WGS) sequencing approach was introduced (Fleischmann, 1995) and it quickly became the most popular strategy for microbial genome sequencing. According to this strategy, each sequencing project requires genomic DNA (gDNA) isolation, library construction, sequencing of the subclones, assembly of sequenced data and genome finishing (Fig.1).

For the project to succeed, it is vital to begin with high molecular weight gDNA which makes possible the creation of three genomic libraries with different insert length. Typically, the insert sizes are set at 3-Kb, 8-Kb and 40 Kb thus creating 3kb, 8kb and fosmid DNA libraries. Library inserts are then sequenced from both ends, resulting in approximately 8-9X genome coverage with paired reads. Reads are aligned against each other by using various genome assemblers to produce a draft assembly, which consists of contigs linked into larger scaffolds based on the information about read pairs. The genome closure stage (finishing) is used to solve all possible misassemblies within contigs and to close gaps in (sequencing or captured gaps) and between (physical or uncaptured gaps) scaffolds to build a single high quality contig spanning the entire genome (note that final assembly might contain more than one contig if the genome consists of a larger number of replicons).

<Figure 1 near here>

Before moving on to a more detailed description of the WGS assembly and finishing stages, it is necessary to mention that new sequencing technologies gave birth to a third sequencing strategy. This combined approach uses sequencing data produced by

different sequencing platforms. For instance, a combination of pyrosequencing data and traditional Sanger reads was successfully applied to a number of microbial projects (Copeland, 2007; Goldberg, 2006). The combined approach has its own advantages and drawbacks (which will be discussed later on), but seems to be promising for faster and more cost effective microbial genome project completion.

C. Libraries

WGS sequencing begins with random shearing of gDNA and further cloning of small fragments of DNA into different cloning vectors. Depending on vector capacity to accept foreign DNA fragments of different length, small- or large-insert genomic libraries can be created. Libraries of different insert size play different roles in the genome assembly: whereas small-insert clones (3 kb on average) are needed to obtain the necessary genome coverage, larger insert libraries (8 kb, cosmids, fosmids or BACs) serve to verify the accuracy of contig assembly and to order and orient them (Myers, 2000).

When dealing with small-insert libraries, it is important to be certain that each vector contains only one insertion. Simultaneous cloning of several random DNA fragments will lead to chimeric clones and thus will cause problems for the later stages of genome assembly. A number of approaches, such as vector/insertion ratio optimization in the cloning ligation reaction, efficient size selection, and the use of specially designed vectors allow minimization of the probability of multiple fragment co-cloning.

Large-insert libraries are usually created by cloning DNA fragments of 25-200 kb into cosmid (Collins, 1978), fosmid (Kim, 1992) or BAC (Shizuya, 1992) vectors. In addition to the fact that such vectors can carry long fragments of foreign DNA, they are present in one or two copies in E.coli cells (low copy number vs. high copy number vectors (Nakano, 1995; Norrander, 1983) used for 3 and 8kb libraries). Low copy number vectors are very useful when so-called “poisonous sequences” (i.e. sequences whose presence or expression interferes with the biology of the host organism in some way resulting in host death) are cloned. Mechanisms of toxicity differ; they include cases of open reading frames which code for toxic proteins, operators that reduce effective concentration of an essential DNA-binding regulator, certain A+T-rich DNA fragments, which can serve as strong promoters in E.coli and initiate transcription of genes encoded by the vector itself at a very high level. It was shown, for example, that some membrane and ribosomal proteins (Luo, 1997) are toxic for E.coli cells when highly expressed. This has been confirmed by numerous experimental observations that some areas of bacterial genomes are represented only by large-insert clones.

The next step is to sequence the inserts from the clone libraries. The WGS strategy does not mean that a complete sequence of the entire insert cloned into the vector will be produced. Typically, inserts are sequenced from both ends resulting in two sequencing reads of about 700 bp each. These pairs of reads, otherwise known as sister reads, form a mate pair. The sister reads are oriented towards each other. The distance between them is within the mean library insert size plus or minus three times the standard deviation of the insert size. When assembled (see below) sister reads may appear in the same or a different contig. Information about mate-paired reads is a very important input for assembly tools (assemblers).

If we assume that the library clones (subclones) cover the chromosome evenly (*i.e.* there is no cloning bias) and that the length of each read is approximately 700 bp, sequencing of 715 clones will be needed to cover (1x read coverage) a 1 MB bacterial genome once. To obtain 1x clone coverage of the same 1 MB genome one needs to have 333 clones from the 3 kb library or 125 clones from the 8 kb library. In practice, clones are not randomly distributed and some areas remain unsequenced (gaps) even if 8X read coverage is provided (Fig.2).

<Figure 2 near here>

II. Assembly

A. Pre-processing of raw data

The next step of the WGS project is to assemble the DNA reads produced during the sequencing stage in order to reconstruct the entire genome. The assembly of tens of thousands of reads into a few contigs (preferably into one – the ultimate goal of any sequencing project) is a very complicated computational task. So, what makes the whole genome assembly so challenging?

Sequenced reads are never 100% error free. Physical limitations of the gel electrophoresis technique, chemical artifacts and human mistakes contribute significantly to assembly complications. The second equally or even more important factor further complicating the assembly is the presence of areas of identical or nearly identical sequence (repeats) in DNA molecules.

A number of software tools were developed to address these issues. Computer programs like ABI-basecaller (http://www.appliedbiosystems.com), pregap4 (http://staden.sourceforge.net) and Phred (Ewing, 1998) analyze raw data produced by the automated DNA sequencing machines by converting them into a sequence of bases (basecalling) and assigning quality scores to each nucleotide. Because of its high accuracy, the Phred base caller became the most widely used tool. Phred quality scores range from 4 to about 60 and define the probability that the base call is correct. A base with a quality score of 20 or higher is usually considered a high quality base. "Phred20" score means that the probability of the base being called incorrectly is 1 in 100. Quality at the beginning of the read and closer to its end is lower than in the middle of the read. This means that the error rate is higher (sometimes significantly) towards the end of the read, which makes assembly more difficult.

Another group of computer programs (available both as independent tools and as part of larger assembly packages) was developed to mark (pregap4 – part of the Staden package [http://staden.sourceforge.net]) or trim out (Lucy (Chou, 2001), Arachne trimming module (Batzoglou, 2002), MAGIC-SPP (Liang, 2006)) low quality tails of the reads and to mask or remove cloning vector sequences (Cross-Match provided by Phrap (http://www.phrap.org); vector_clip (Staden package-gap4), Lucy, etc), mostly found in the beginning of the reads but also occasionally detected at the end of the read if the insert is very short. Such pre-assembly data treatment significantly improves assemblers' performance and increases assembly quality.

B. Assemblers

Generally speaking, most assemblers follow a three-phase approach: overlap, layout, and consensus. During the first phase, the assembler determines which pairs of reads overlap; these reads may have been drawn from the same DNA region. During the layout phase, the assembler places the reads, thereby attempting to reconstruct contiguous sequences of DNA, or contigs, and to order and orient contigs to build scaffolds. During the final phase, the assembler uses the reads that have been placed in order to generate the consensus sequence - the assembler's best reconstruction of the original DNA sequence. Advanced algorithms employed by modern assemblers are also able to address such important problems as repeat locations and at least their partial resolution and identification of chimerical reads.

The first assembler was described by Roger Staden (Staden, 1979) and has been significantly improved over the following few years (Bonfield, 1995; Staden, 1982). The current version of this assembler – GAP4 (Genome Assembly Program) – "contains all the tools that would be expected from an assembly program plus many unique features and a very easily used interface" (http://staden.sourceforge.net/manual).

One of the most commonly used assembly programs is Phrap (Green, 1994). In order to detect overlaps, Phrap scans the set of reads to find pairs with perfectly matching subsequences and passes these sequences to a banded Smith-Waterman algorithm to find the optimal local alignment. Read couples that are near-duplicates are retained, while combined read couples that do not have good matching segments are considered to be not a match and are discarded. The LLR score, a combined measure of match length and quality, is computed for each retained match. In the layout phase, Phrap constructs contig layouts using matches in decreasing order of LLR score. Version 3 of this assembler does not use paired read constraints and does not generate scaffold information. A separate program, such as Bambus (Pop, 2004a), may be used to order and orient contigs. Phrap 4 (also known as Southwest Parallel Software (SPS) Phrap) is an updated version of Phrap which, unlike Phrap 3, uses paired reads information and produces fewer misassemblies. Phrap 4 can create scaffold information using mate pair links, with an option to link contigs only if two mate pairs from the same library link the two contigs. Phrap 4 can also be instructed to ignore singleton mate pairs in the assembly. For consensus sequence generation Phrap analyses all individual reads and picks bases with the highest Phred quality score. Phrap is one of the best tools for working with low-quality reads and for assembling low-coverage nonrepetitive regions (Yang, 2002).

Another widely used assembler is Arachne (Batzoglou, 2002; Jaffe, 2003). To perform the overlapping stage Arachne finds all reads that share a k-mer, an identical sequence of k bases (word), but excludes (masks) the reads that occur with extremely high frequency. In practice $k = 24$ is used. This exclusion improves the efficiency of overlap detection, and is thought to exclude high fidelity repeat sequences. For layout, Arachne finds pairs of paired reads with sequence overlaps at both ends. Arachne iteratively identifies and extends "paired pairs", forming mini-contigs, and then assembles true contigs by avoiding assembling across repeat boundaries. Some of these contigs will represent repeat regions (they are identified by high depth of coverage, and by conflicting mate pair links to other contigs). The non-repeat contigs are organized into scaffolds if they are linked by at least two mate pair links; priority is given to contigs

that are separated by short distances and spanned by many mate pair links. Arachne then attempts to fill in the gaps between contigs by finding a path of repeat contigs across each gap, guided by mate pair links. The improved versions of Arachne first construct a set of contigs that are unlikely to include a misassembly, and then iteratively improving the initial assembly through read placement and contig breaking at suspicious points. Arachne is quite efficient at repeat resolution, with the exception of tandem repeats. To build the consensus sequence Arachne converts pair-wise alignments of reads into multiple alignments and merges overlapping adjacent contigs in scaffolds. Quality scores are used to create and evaluate read alignments. This very powerful assembler was also successfully used for mammalian genome assembly (Jaffe, 2003).

One more state-of-the-art program, PCAP (Huang, 2003), uses a parallelized local alignment technique and length of the word (k) of 12. For layout, pairs of reads are placed in order of decreasing local alignment score (Huang, 1996). The quality of the current layout is then determined by checking mate pair constraints. Corrections are made to the region with the largest number of constraint violations. The process of constraint assessment and correction is repeated until no areas with constraint problems are found. Finally, assembler attempts to close gaps between linked contigs using the reads overlapping the ends of the contigs together with reads considered to lie in repetitive regions. PCAP computes a consensus sequence for contigs based on an alignment of reads in contigs and uses both quality values and coverage information for each base (Huang, 1999). The latest version of this assembler PCAP.REP (Huang, 2006) does not use the constant word length to overlap reads. The developers implemented the idea of using superword for more effective overlapping of repetitive reads. Superwords consist of more than one k-mer and can be of different length. It was demonstrated that the new method of overlap detection is very efficient in both unique and repetitive regions and that PCAP.REP produces more accurate and contiguous assemblies of whole-genome datasets.

The AMOS Comparative assembler (AMOS-Cmp) skips the overlap step (Pop, 2004b). It aligns reads of the target genome to the reference genome of a closely related organism by using a modified version of the MUMmer algorithm (Kurtz, 2004). Authors call this approach alignment-layout-consensus, since the assembler produces consensus from the alignment. However, even closely related organisms can significantly differ from each other due to insertions, deletions, rearrangements, repetitive regions and lateral gene transfer thus making comparative assembly challenging. The AMOS-Cmp is aimed to overcome these problems. Taking into consideration the increasing number of high quality completed reference genomes in public depositories, the comparative (assisted) assembly approach has a bright future.

C. Algorithms for second-generation sequencing data assembly

As was already mentioned above, second generation sequencing technologies are becoming a reality. Most of them produce short or ultra short reads (20-200bp) and the number of reads for a given genomic project is orders of magnitude higher than what is obtained in the case of the Sanger method. This makes usage of currently available assembly tools nearly impossible. Another problem with using traditional assemblers in connection with next generation sequencing technologies is the fact that each of these

platforms generates data in different formats (flowgrams, images, text format (fasta), etc) and that these formats differ from the one generated by Sanger sequencers. As a result, there is an urgent need for new computational methods for analyzing massive amounts of very short reads and for managing this overwhelming volume of data. New algorithms are needed to analyze short reads in a number of applications, including de novo genome assembly and visualization, polymorphism detection, gene expression, and methagenomics.

The only known case of currently available assemblers being used for short-read sequence assembly was described in a publication by EULER assembler developers (Pevzner, 2001). A modified version of EULER was used to assemble simulated reads of length 80-200 bp created to obtain 30x coverage for several test cases (BACs and viral and bacterial genomes) (Chaisson, 2004). The same main problems – repeats and sequencing errors – complicate short-read assembly, but shorter reads result in a larger number of repeats, which creates additional difficulties for assembly tools (repeats that are shorter than the read length). The EULER assembly approach, which performs fragment assembly by finding an Eulerian path through a de Bruijn graph representation of a genome, works well on error-free reads (Pevzner, 2001). To avoid or minimize the influence of the error rate, the modified version of EULER addresses the problem of errors by correcting them prior to assembly, thus eliminating alignment and base calling steps. However, in order to adapt to variations of error rates and types in second generation sequencing methods, the corrective stage needs to be optimized (Chaisson, 2004). Experiments with the EULER assembler allowed assembly of short reads into relatively long contigs and thus demonstrated the feasibility of such assembly for de novo draft sequencing. It was also predicted that significant finishing efforts would be required to resolve repeats and misassemblies to produce the finished genome.

454 Life Science company has developed its own de novo flow-space assembler optimized for an increased number of 80-120 bp flow-based reads (Margulies, 2005). The assembler – Newbler™ – consists of three modules performing the overlap-layout-consensus operations and uses only high quality flowgrams for better performance (it is stated that quality scores used by Newbler™ are in good correlation with phred scores). Errors most typical for the 454 platform are the ones related to homopolymers. They are caused by over- or under- estimation of the intensity of the signal or by their combination.

To overlap reads the Overlapper module compares all flowgrams by using a hashing indexing method. The next module – Unitigger – groups the overlapped reads into “unitigs”. As defined in (Margulies, 2005) a “unitig is a collection of reads whose overlaps between each other are consistent and uncontested by reads external to the unitig. Unitigs are constructed from consistent chains of maximal depth overlaps.” Created unitigs then go through the all-against-all comparison to join the overlapping ones and to break the contigs at the repeat boundaries (Multialigner). For high quality consensus generation the Newbler™ assembler averages the signals for each individual assembly position. This assembler was first tested on several microbial genome assemblies (Margulies, 2005) and is currently employed by all laboratories where 454 instruments are being used.

Another problem that was raised by the novel DNA sequencing technologies is the need to manipulate even shorter reads of 15-30bp long. The main goal of this group

of technologies is to provide cheap and rapid methods for human genome re-sequencing. Tremendous amounts of very short reads produced by high throughput DNA sequencing instruments for a target genome can be aligned against a reference genome to detect the differences between two genomes. Is it possible to create an assembler capable of putting together millions of primer-sized reads without using a reference? Analytical studies performed by Steven Skiena's group at Stony Brook, NY (Skiena, 2007) gave a positive response to this question. They state that microbial genomes can be reliably assembled "with a coverage of 500 under realistic error rate" (http://www.algorithm.cs.sunysb.edu/shorty/files/paper.pdf). Based on their estimate, even with this huge coverage the cost of de novo genome sequencing will not exceed $100 per megabase.

SSAKE (the Short Sequence Assembly by progressive K-mer search and 3' read Extension) program was developed in an attempt to achieve the goal of de novo genome assembly using 25-bp reads (Warren, 2007). Simulated error-free data sets created for bacteriophage (PhiX174), SARS virus (SARS TOR2), bacterial (Haemophilus influenzae) and metagenomic (Sargasso sea) projects were aggressively assembled by using a prefix tree. Data sets are stored in hash tables and the assembly process goes through a number of iterations searching for progressively shorter 3'-most k-mers for every new read added into the assembly. This approach allowed a complete assembly of the PhiX174 genome and generation of reasonably long contigs covering the non-repetitive regions of viral, microbial and even metagenomic samples.

These are the first attempts to satisfy the growing need for advanced assembly tools and it is expected that a number of sophisticated assemblers will appear in the near future. The recent release by Synamatix (Malaysia-based company) of a whole package of tools for short-read genome assembly, visualization and annotation (www.synamatix.com) supports this projection.

D. Assembly of metagenome shotgun sequences

Interest in genomic analysis of microbial communities (Riesenfeld, 2004b) led to the appearance of a relatively new group of sequencing projects – metagenomic projects. The whole-genome shotgun sequencing approach was successfully used for a number of uncultivated microbial community projects (Chen, 2005). Despite this fact, assembly of shotgun-sequenced metagenomic DNA poses a serious challenge to traditional assembly methods that were developed to handle relatively uniform sequences derived from isolated microbes. In addition to the typically very large size of metagenomic sequencing projects, potential problems observed in metagenomic assemblies include chimeric contigs produced by co-assembly of sequencing reads originating from different species, and non-uniform sequence coverage resulting in significant under- and over-representation of certain community members. Communities with one or two highly abundant members represented by several strains can add to the complexity of metagenome assembly due to the presence of extensive strain-level heterogeneity. As a result, often times a non-uniform assembly of the genomes of these abundant members is produced. Depending on the assembler used and sequencing read depth, some fragments are resolved into strain-specific contigs corresponding to different haplotypes, while others are co-assembled into composite contigs with strain-specific variations appearing as single-nucleotide polymorphisms (Markowitz, 2006). Large-scale genome

rearrangements and the presence of mobile genetic elements (phages, transposons) in the abundant community members result in assembly break points in the areas of synteny breakdown. Many parameters of metagenomic assemblies remain unknown, including the nature and extent of chimeric contigs , influence of the level of polymorphism on co-assembly of reads into composite contigs, overall quality of assemblies and binning (process of separation of scaffolds into species-specific groups), etc. Assembly accuracy is essential for metagenomic projects since it has much more influence on subsequent analysis and interpretation of metagenomic data than in cases of individual microbes due to coverage-related increases in consensus error rate. Currently there is no single best pipeline to follow since each assembly program, each gene prediction algorithm, and each method of binning possesses its own set of benefits and problems. Moreover, each environment has its own particular complexities that are best dealt with using a combination of several tools for analysis. The high complexity and heterogeneity of metagenomic data call for an evaluation of the performance of different assembly tools on metagenomic sequences and require modifications to the assemblers and additional quality control throughout the entire process (Mavromatis, 2007). To improve the quality of metagenomic assemblies, reads should be stringently quality- and vector-trimmed prior to assembly (tools like Lucy (Chou, 2001) or MAGIC-SPP (Liang, 2006) can be used for that) and all possible attempts should be made to control quality of metagenomic assemblies produced. A number of assemblers containing modules to correct sequencing errors and to automatically detect and correct assembly problems (Arachne (Jaffe, 2003), Atlas (Havlak, 2004), Phusion (Mullikin, 2003), AutoEditor (Gajer, 2004), EULER-AIR (Zhi D, 2007)), can be used in metagenomic project assembly.

Depending on the complexity of the microbial community and the representativity of its members, assembly can contain a significant number of unassembled reads (up to 100%). This can complicate gene prediction and functional annotation of metagenomic sequences. Different approaches, as for example the comparative assembly approach (using publicly available completely sequenced individual microbial genomes as references (Pop, 2004b)), different HMM-based microbial gene finders (it was shown (Besemer, 1999) that GeneMark family sofware (Besemer, 2005) can be applied for gene finding in segments with length starting from 400 bp; FGENESB (Solovyev, 1997), GLIMMER (Salzberg, 1998) and BLAST search can all be used for data analysis.

Directed sequencing is another approach used for metagenomic projects. This is based on the selective sequence of large-insert library clones of interest (Riesenfeld, 2004a; Uchiyama, 2005). The set of clones are screened for a desired function or the presence of phylogenetic markers. Recently, pyrosequencing technology, which does not require cloning of environmental samples, has been used to generate environmental genome sequences from two sites in the Soudan Mine, Minnesota, USA (Edwards, 2006). Shotgun and directed sequence approaches can be combined to help each other: random sequencing of large-insert libraries guide the selection of clones for complete sequencing using either traditional Sanger sequencing or one of the emerging second generation DNA sequencing methods (Balasubramanian, 2001; Margulies, 2005). This combination brings together the advantages of broader coverage provided by shotgun sequencing with the ability to sample specific genome areas in low abundance organisms without over-sequencing more abundant members of the microbiome. Despite the difference in sequencing strategies, processing of data generated by both approaches faces similar

challenges due to inherent incompleteness and lower quality of the sequence data, and in many cases due to unknown origin of each sequence fragment. Development of new sequencing strategies brings new demands to data processing methods, such as assembly and annotation of pyrosequencing reads that are very short and characterized by high error rate at homopolymeric runs.

Tools specific for metagenomic sequence assembly (especially ones capable of dealing with data produced by different sequencing platforms), gene prediction and functional annotation are at the early stage of development. The same could be said of the process of recording metagenome sequence information in traditional depositories like GenBank and EMBL.

III. Genome assembly improvement and finishing

None of the sophisticated computer algorithms described above is able to automatically reconstruct the entire genome from sequencing reads. They all produce so-called "draft" assemblies, which are never perfect. Typically, there are problems such as misassembled areas usually caused by repetitive regions (repeats); sequence and physical gaps; areas of low coverage and/or poor quality. The better the assembler can handle such problems, the higher quality draft will be created and the faster and easier the assembly improvement step (finishing) will be. Finishing usually starts after the draft assembly is ready. Finishing is the process of transforming a draft assembly into a finished one. During this step all repeats are identified and assembled correctly, all misassemblies are resolved, all gaps are closed, and all bases are identified with high accuracy. Hus finishing is the process of incrementally improving an assembly by using computational tools, techniques and experimental protocols. Despite the fact that a solid suite of software tools and web applications have already been created, finishing genomes remains a labor-intensive process (Schatz, 2007). It requires experienced personnel and laboratory experiments to support finishing strategies.

It is commonly believed that finishing begins by the ordering and orientation of contigs (scaffolding) for subsequent gap closing. However, the experience collected during the course of the Microbial Finishing Program at the DOE Joint Genome Institute (www.jgi.doe.gov) demonstrates that it is more efficient to begin the process of draft assembly improvement with misassembly resolution. It is rather logical, since incorrectly assembled contigs lead to the construction of erroneous scaffolds, created on the basis of false connections. In addition to real gaps, such assemblies will contain pseudo gaps. All this in turn will lead to further complications and the lengthening of the finishing process.

The areas of potential misassemblies can be recognized at different stages of genome assembly via the application of different methodologies. Some assemblers can identify and even correct repetitive areas during the assembly process (Fig. 3). They use paired reads information (detecting areas of clone mates inconsistencies including reads placed too far apart or too close or incorrectly oriented paired reads), statistical information (identify areas of significantly higher coverage than the average) or mask out repeats (in this case additional efforts for repeats assembly are required). Up to 80% of the repetitive areas can be addressed automatically, while the remaining ones (the most ambiguous) need additional laboratory experiments and manual efforts for their resolution (Mulyukov, 2002). Additional tools were developed for more effective

detection of repeats. They can be used after the initial draft assembly has been created. For instance, RepeatMasker (Green, 1994) locates repetitive areas and excludes them (masks) from further searches for similarity regions. In addition to the exact repeats, a tool named REPuter recognizes degenerate repeats thus allowing for a certain rate of sequence errors (Kurtz, 2001). It is able to detect not just direct repeats, but also palindromic repeats and other closely related sequence features. Tools like equicktandem (Rice, 2000), mrep (Kolpakov, 2003), and Tandem Repeat Finder (Benson, 1999) find tandem repeats of certain size by using statistical (equicktandem) or heuristic (mrep, Tandem Repeat Finder) algorithms for their detection. Another software package – Vmatch – represents a collection of programs, for solving large-scale sequence matching tasks (http://www.vmatch.de).

<Figure 3 near here>

Repetitive elements can be of different size and may differ from each other by one or more bases. Most repeats of 2-3 kb are due to insertion elements. Multiple copies of ribosomal DNA sequences represent longer repeats of 5-8 kb. In some microbial genomes, repetitive elements of 70-100kb were observed. They represent duplications of large areas of the genome, which could have occurred during the course of evolution of the particular microbe in question. Besides being differentiated by their length, repeats can be split into three categories by type: direct, inverted or tandem repeats (Fig. 4). While direct repeats and different copies of tandem repeats can collapse in one location or become rearranged during the assembly, copies of inverted repeats may be assembled in the wrong orientation thus creating pseudo physical gaps (Fig. 4). By using the Miropeats program (Parsons, 1995) one can draw a graph that will help to distinguish tandem repeats, inverted repeats and palindromes; however, biological duplication events and assembly mistakes cannot be distinguished by this program.

<Figure 4 near here>

A commonly used approach to repeat resolution consists of the following steps: repeat localization, identification of the reads belonging to each repetitive element, pooling of (grouping) the defined reads, their assembly into separate subassemblies for each individual copy of a repeat and exporting the resulting consensus back to the main assembly as one "long read" (fake read) for each subassembly. Tandem repeats represent the greatest problem in successful repeat resolution. They are especially troublesome if the length of one copy is longer than the insert sizes of the libraries available for the sequencing project. The reads belonging to different copies of a repeat and entirely lying within the repetitive region collapse into one pile during draft assembly. This means that assembler produces fewer copies of the repeat than the finished sequence contains. Because the library clones are not uniformly distributed over the genome, it is not sufficient just to estimate the extent of increased coverage in the area to discover the number of copies of a repeat gathered in such piles. An in vitro transposon insertion strategy involving a random insertion of a yeast transposable element into a repeat-covering, circular plasmid is one of the most powerful experimental approaches for tandem repeat resolution (Liu, 1987). Transposon "bombing" allows random

identification of new sequencing start points within the repeat and to generate new sequence directed away from the insertion points.

After all the misassemblies have been resolved, it is time to fill in the gaps between the contigs, in order to produce contiguous sequence for each DNA. In practice, the process of closing gaps between the correctly assembled contigs can be started in parallel with repeat resolution.

Paired reads information and the knowledge of the library insert size distribution allow ordering and orientation of correctly assembled contigs and organizing them into scaffolds (Fig. 2). Software tools named scaffolders were created to assist finishers in this complex task. Consed (Gordon, 1998) and Bambus (Pop, 2004) are among the most popular tools of this type. These and other similar programs (for example, scaffolders within the Celera assembler (Myers, 2000) and Arachne) provide a global overview of the interdependency between the contigs and help with selection of particular clones that will help bridging gaps and planning of finishing experiments.

Small sequence (captured) gaps can be closed by custom primer walking on the existing clones that span gaps. Such clones can be distinguished by the presence of the sequence from each end of the insert in two separate contigs. During the primer walking procedure a custom primer designed for the end of the contig is used to perform sequencing reactions to extend sequence information into the gap. This process is repeated until the entire gap-spanning plasmid is sequenced. It is faster and more cost efficient to close captured gaps of 5 kb or longer by transposon "bombing" of gap-spanning clones (Epicenter Biotechnologies) or by producing a shatter library (McMurray, 1998) of appropriate plasmids. In the shatter method, the chosen DNA template is sonicated into fragments of 100-300 bp, which are then sub-cloned into a vector and sequenced by using standard vector primers.

Many captured gaps are difficult to sequence by primer walking. Problems usually arise due to the presence of strong secondary structures very common in genomes with GC content higher than 65% and hairpin structures and/or long homopolymer stretches in DNA molecules. Sequencing very short inserts produced by shatter libraries usually helps with most DNA structures because only a small part of the secondary structure or hairpin is cloned. Currently available commercial kits and alternative approaches developed to help struggle with difficult templates were carefully described by Jan Kieleczawa (Kieleczawa, 2006).

Linking information, usually provided by sister reads, is not available for physical gaps, existing between contigs or scaffolds. Sometimes the comparative analysis of closely related reference and target genomes can provide a clue for contig mapping. If the end sequences from the two contigs encode two different parts of the same protein, it may be assumed that these contigs should be linked. It was also shown (http://img.jgi.doe.gov/cgi-bin/pub/main.cgi) that the gene composition of some operons is very well conserved between genomes. This information can also be used to link the contigs containing parts of such operons at their ends (Lapidus, 2002). Several software tools were developed to assist with this analysis by comparing the entire genomes or contigs from different assemblies against each other and against reference genome(s) (GMPTB - http://www.pasteur.fr, MUMmer (Kurtz, 2004), Projector2 (van Hijum, 2005) etc). Such assisted assembly strategies go beyond just helping with gap closing. They offer an opportunity to sequence new organisms to less than the typical sequence

depth. For those cases in which none of the approaches described above is helpful, direct laboratory experiments need to be performed. For example, a technique called optical mapping compares the assembled contigs against the collection of restriction maps of a DNA molecule. Each map produced by a rare cutting restriction enzyme serves to order and orient contigs and scaffolds separated by uncaptured gaps (Reslewic, 2005). Selected restriction fragments can be used as templates to produce the sequence for gaps.

When no other templates are available, PCR products generated across the gaps can be used to map contigs or scaffolds and to produce sequence for the gaps. In order to build a PCR map, unique primers are designed for the end of each contig (it is necessary to make sure that primers correspond to regions outside of repeats) and used in a PCR experiment to test whether a particular pair of primers links the contigs. This approach is not feasible in the case of large numbers of contigs to be mapped, since it requires too many PCR reactions to be performed and analyzed.

An improved version of the combinatorial PCR method – Multiplex PCR – was developed to optimize this process (Sorokin, 1996). This approach is based on the simultaneous use of multiple primers (up to 32) in mapping experiments with further analysis of which two primers made the PCR fragment (Fig.5 schematically represents the use of Multiplex PCR approach for mapping of four contigs).

The advantage of the Multiplex PCR method versus the combinatorial PCR approach, where each end primer is verified against all others except the one that was designed for the second end of the same contig/scaffold, is that it requires less PCR reactions to map the contigs. Thus, to order and orient 4 contigs (Fig. 5) one should perform 9 multiplex reactions instead of 24 combinatorial ones. The difference becomes more prominent when larger numbers of contigs are involved in the experiment (only 17 multiplex PCR reactions vs. 120 combinatorial PCR reactions are needed to map 8 contigs). The PCR products thus obtained are sequenced to close the gaps between mapped contigs or scaffolds (Lapidus, 1997).

Sequencing and mapping complications associated with strong secondary structures and unclonable areas present significantly less challenges for the second generation of sequencing technologies, as the majority of them rely on assembly of very short overlapping fragments and do not require fragment cloning in E.coli. As a result, 454 contigs, assembled by the Newbler assembler effectively cover physical gaps, produced by WGS. The highest success rate was observed for genomes with low GC% (< 40%), for which a combination of Sanger and 454 sequencing was used. Such a combination allowed closing of up to 100% of uncaptured gaps (Fig. 6) (unpublished data). The 454/Sanger combined approach (Goldberg, 2006) also allowed the completion of the uncultured Gram-positive bacterium, *Candidatus Desulforudis audaxviator*, from a low biodiversity water fraction collected at 2.8 km depth in South Africa (Chivian, 2006).

<Figure 5 near here>

In addition to the above-mentioned problems, a typical whole genome draft assembly contains low quality regions and regions poorly covered by clones (1x). In order for the genome to be considered finished such areas should be improved by re-sequencing of the existing clones or of the appropriate PCR products (so called polishing step). Solexa and other new technologies, developed specifically for genome re-

sequencing seem to be very promising for faster and cheaper polishing of the microbial genomes. Results recently reported by the Broad Institute at the Cold Spring Harbor Conference (May, 2007) support this notion: a Solexa run aligned to a 454-assembled bacterial genome identified (and allowed correction of) 98% of frameshifting indels. More importantly, the remaining frameshifts were in fact true indels in the genome.

To completely wrap up the microbial project it is necessary to confirm the correctness of the final assembly and the accuracy of the finished sequence. The sequence quality standard used for Human Genome sequence data and named "Bermuda" standard (1 error per 10,000 bp, Marshall, 2001) has also been accepted in microbial sequencing. Final human inspection step supported by visualization tools (DNPTrapper (Arner, 2006), Hawkey (Schatz, 2007), Orchid – www.shgc.stanford.edu/informatics/orchid.html) aims to verify that all bases of the produced consensus are supported with enough coverage, that consensus quality corresponds to the agreed standards, that all produced reads are correctly assembled (consistent) and that all repetitive areas are resolved (www.jgi.doe.gov).

<Figure 6 near here>

Progress in DNA sequencing technology, improvements in finishing strategies and tools, as well as the availability of a number of assemblers and advanced methods for genome annotation has significantly reduced the time required for genome closure. Despite this fact, the effort and time required for genome closure depends on the quality of the whole genome shotgun libraries created for the project, GC content of the genome, the size and frequency of identical or nearly identical repetitive structures, and the number of regions that can not be cloned or are hard to clone in E.coli. A finished genome represents the genome assembly of high accuracy and quality (with no gaps), verified and confirmed through a number of computer and lab experiments. The value of complete microbial genome sequences has been long established and appreciated by the scientific community (Fraser, 2002). In addition to the genes that differ by a very few bases from each other and thus co-assemble in draft assembly, the genes coding for proteins toxic for E.coli remain undetected by the analysis of incomplete genomes. Furthermore, new sequencing technologies, developed for genome re-sequencing (Solexa, SOLiD, Helicos) dictate a bigger need for high quality references. A significant number of microbial genomes still needs to be sequenced to full completion for successful use of new sequencing platforms in many areas of general microbiology, ecology, evolutionary studies etc.

The DOE Joint Genome Institute is currently launching a project named GEBA (Genomic Encyclopedia for the Bacteria and Archaea), which is focused on filling gaps in the Tree of Life and has the long-term goal "to generate reference genomes for every major and minor group of bacteria and archaea. This could represent on the order of 5,000 genomes" (http://www.jgi.doe.gov/News/primer/primer043007.pdf).

IV. NCBI Trace and Assembly Archive

The Trace Archive was established in 2001 to collect raw data produced at sequencing centers around the world. This depository is a collaborative effort between NCBI (www.ncbi.nlm.nih.gov/Genbank) and European Molecular Biology Laboratory (EMBL/ENSEMBL) (www.ebi.ac.uk/embl) and it currently (at the moment of paper preparation) contains more than 22 trillion bytes of data. The amount of data in the archive doubles every 10 months and new sequencing technologies will result in an even higher rate of data increase in the future.

NCBI's Trace and Assembly Archives provide direct access to the raw traces and assemblies and give researchers a unique opportunity not only to reconstruct the assembly from raw shotgun and finishing reads obtained from the Trace Archive, but also to verify the quality of the assembled genome and its accuracy. Any potential or real frame shifts detected in the course of genome annotation can be thus double checked and corrected based on the details of assembly.

Because different sequencing centers and individual researchers use different tools to assemble sequences, the generated assemblies may come in a variety of formats. Regardless of the specific format, the data can still be submitted to The Assembly Archive, since the database accepts files in popular .ACE format. Outputs of the assemblers like TIGR and Celera assemblers can be converted to .ACE or to Assembly Archive format using open source conversion tools available at the AMOS website (www.amos.sourceforge.net)

<Figure 7 near here>

Fig.7 illustrates one small region of multiple sequence alignment of reads to the *Salinospora tropica* CNB440 genome sequenced at JGI (http://genome.jgi-psf.org/mic_cur1.html). Another function made available by The Assembly Archive is the ability to see the individual bases in the DNA sequence by examining the aligned traces (Fig.8).

<Figure 8 near here>

This is the feature that opened the door for analysis of single nucleotide polymorphisms (SNPs) in a broad range of eukaryotes (Salzberg, 2004) and will be useful in the future for the haplotype analysis of metagenomic assemblies. Currently the haplotype analysis of metagenomic data can be performed in IMG/M (http://img.jgi.doe.gov/cgi-bin/m/main.cgi), an experimental metagenome data management and analysis system developed by the DOE Joint Genome Institute (Markowitz, 2006) (Fig.9)

<Figure 9 near here>

References:

Albrecht, G., Brenner, S., DuBridge, R. B. et al. (2000). Massively parallel signature sequencing by ligation of encoded adaptors, Patent, United States, Application Number 6013445.

Arner, E., Tammi, M.T., Tran, A.N. et al. (2006). DNPTrapper: an assembly editing tool for finishing and analysis of complex repeat regions. BMC Bioinformatics 7, 155-165.

Azevedo, V., Alvarez, E., Zumstein, E. et al. (1993). An ordered collection of Bacillus subtilis DNA segments cloned in yeast artificial chromosomes. Proc. Natl. Acad. Sci. USA 90, 6047-6051.

Balasubramanian, S., and Bentley, D.R. (2001). Polynucleotide arrays and their use in sequencing, Vol. Patent WO 01/157248.

Batzoglou, S., Jaffe, D.B., Stanley, K. et al. (2002). ARACHNE: A whole genome shotgun assembler. Genome Research 12, 177–189.

Beigel, R., Alon, N., Apaydin, M.S. et al. (2001). An Optimal procedure for gap closing in whole genome shotgun sequencing. Proceedings of the Fifth Annual International Conference on Computational biology. RECOMB, 22 - 30.

Benson, G. (1999). Tandem Repeats Finder: a program to analyze DNA sequences. Nucleic Acids Research 27, 573–580.

Bentley, D. R. (2006). Whole-genome re-sequencing. Curr Opin Genet Dev 16, 545-552.

Besemer, J., and Borodovsky, M. (1999). Heuristic approach to deriving models for gene finding. Nucleic Acids Research 27, 3911-3920.

Besemer, J., and Borodovsky, M. (2005). GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Research 33, 451-454.

Bolotin, A., Mauger, S., Malarme, K. et al. (1999). Low-redundancy sequencing of the entire Lactococcus lactis IL1403 genome. Antonie Van Leeuwenhoek 76, 27-76.

Bonfield, J. K., Smith, K.F., and Staden, R. (1995). A new DNA sequence assembly program. Nucleic Acids Research 23, 4992-4999.

Braslavsky, I., Hebert, B., Kartalov, E. et al. (2003). Sequence information can be obtained from single DNA molecules. Proc Natl Acad Sci USA 100, 3960–3964.

Chaisson, M., Pevzner, P., and Tang, H. (2004). Fragment assembly with short reads. Bioinformatics 20, 2067-2074.

Chen, K., and Pachter,L. (2005). Bioinformatics for Whole-Genome Shotgun sequencing of microbial communities. PLoS Computational Biology 1, 106-112.

Chen, P., Gu, J., Brandin, E. et al. (2004). Probing single DNA molecule transport using fabricated nanopores. Nano Letters 4, 2293-2298.

Chivian, D., Alm, E.J., Brodie, E.L. et al. (2006). 106th General Meeting of the American Society for Microbiology, Orlando, Florida.

Chou, H. H., and Holmes, M.H. (2001). DNA sequence quality trimming and vector removal. Bioinformatics 17, 1093-1104.
Cole, S. T., Brosch, R., Parkhill, J. et al. (1998). Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature 393, 537-544.

Collins, J., and Hohn, B. (1978). Cosmids: a type of plasmid gene-cloning vector that is packageable in vitro in bacteriophage lambda heads. Proc Natl Acad Sci USA 75, 4242-4246

Copeland, A., Lucas, S., Lapidus, A. et al. (2007). Petrotoga mobilis SJ95 whole genome shotgun sequencing project. GenBank Submission, Accession Number: AAZB00000000.

Edwards, R. A., Rodriguez-Brito, B., Wegley, L. et al. (2006). Using pyrosequencing to shed light on deep mine microbial ecology under extreme hydrogeologic conditions. BMC Genomics 7, 57-70.

Ewing, B., Hillier, L., Wendl, M.C. et al. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Research 8, 175-185

Fleischmann, R. D., Adams, M.D., White, O. et al. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269, 496-512.

Fraser, C. M., Eisen, J.A., Nelson, K.E. et al. (2002). The value of complete microbial genome sequencing (you get what you pay for). J Bacteriol 184, 6403-6405.

Gajer, P., Schatz, M. and Salzberg, S.L. (2004). Automated correction of genome sequence errors. Nucleic Acids Research 32, 562-569.

Glover, W. (2004). Systems and methods of analyzing nucleic acid polymers and related components. USPTO 20060029957.

Goldberg, S. M., Johnson, J., Busam, D. et al. (2006). A Sanger/pyrosequencing hybrid approach for the generation of high-quality draft assemblies of marine microbial genomes. Proc. Natl. Acad. Sci. USA 103, 11240-11245.

Gordon, D., Abajian, C., Green, P. (1998). Consed: a graphical tool for sequence finishing. Genome Research 8, 195-202.

Green, P. (1994). PHRAP documentation. http://www.phrap.org.

Hardenbol, P., Baner, J, Jain, M. et al. (2003). Multiplexed genotyping with sequence-tagged molecular inversion probes. Nature Biotechnology 21, 673-678.

Havlak, P., Chen, R., Durbin, K.J. et al. (2004). The Atlas genome assembly system. Genome Research 14, 721-732.

Huang, X. (1996). An improved sequence assembly program. Genomics 33, 21-31.

Huang, X., and Madan, A. (1999). CAP3: A DNA sequence assembly program. Genome Research 9, 868-877.

Huang, X., Wang, J., Aluru, S. et al. (2003). PCAP: A Whole-Genome Assembly Program. Genome Research 13, 2164 – 2170.

Huang, X., Yang, S.-P., Chinwalla, A.T. et al. (2006). Application of a superword array in genome assembly. Nucleic Acids Research 34, 201-205.

Jaffe, D. B., Butler, J., Gnerre, S. et al. (2003). Whole-Genome Sequence assembly for mammalian genomes: Arachne 2. Genome Research 13, 91 – 96.

Kieleczawa, J. (2006). Fundamentals of sequencing of difficult templates - an overview. Journal of Biomolecular Techniques 17, 207–217.

Kim, U. J., Shizuya, H., de Jong PJ, et al. (1992). Stable propagation of cosmid-sized human DNA inserts in an F-factor based vector. Nucleic Acids Research 20, 1083-1085.

Kolpakov, R., G. Bana, G. and Kucherov, G. (2003). mreps: efficient and flexible detection of tandem repeats in DNA. Nucleic Acid Research 31, 3672-3678.

Krzywinski, M., Wallis, J., Gösele, C. et al. (2004). Integrated and sequence-ordered BAC- and YAC-based physical maps for the rat genome. Genome Research 14, 766-779.

Kunst, F., Ogasawara, N., Moszer, I. et al. (1997). The complete genome sequence of the Gram-positive bacterium Bacillus subtilis. Nature 390, 249-256.

Kurtz, S., Choudhuri, J.V., Ohlebusch, E. et al. (2001). REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Research 29, 4633–4642.

Kurtz, S., Phillippy, A., Delcher, A. L. et al. (2004). Versatile and open software for comparing large genomes. Genome Biol 5, R12.

Lapidus, A., Galleron, N., Sorokin, A. et al. (1997). Sequencing and functional annotation of the Bacillus subtilis genes in the 200 kb rrnB-dnaB region. Microbiology 143, 3431-3441.

Lapidus, A., Galleron, N., Andersen, J.T. et al. (2002). Co-linear scaffold of the Bacillus licheniformis and Bacillus subtilis genomes and its use to compare their competence genes. FEMS Microbiol Lett 209, 23-30.

Levene, M. J., Korlach, J., Turner, S.W. et al. (2003). Zero-mode waveguides for single-molecule analysis at high concentrations. Science 299, 682–686.

Liang, C., Sun, F., Wang, H. et al. (2006). MAGIC-SPP: a database-driven DNA sequence processing package with associated management tools. BMC Bioinformatics 7, 115-130.

Liolios, K., Tavernarakis,N., Hugenholtz,P., et al. (2006). The Genomes On Line Database (GOLD) v.2: A monitor of genome projects worldwide. Nucleic Acid Research 34, 332–334.

Liu, L., Whalen, W., Das, A. et al. (1987). Rapid sequencing of cloned DNA using a transposon for bidirectional priming: sequence of the Escherichia coli K-12 avtA gene. Nucleic Acids Research 15, 9461–9469.

Luo, Y., Glisson, J.R., Jackwood, M.W. et al. (1997). Cloning and characterization of the major outer membrane protein gene (ompH) of Pasteurella multocida X-73. J Bacteriol 179, 7856-7864.

Margulies, M., Egholm, M., Altman, W.E. et al. (2005). Genome sequencing in open microfabricated high density picoliter reactors. Nature 437, 376-380.
Markowitz, V. M., Ivanova, N., Palaniappan, K. et al. (2006). An experimental metagenome data management and analysis system. Bioinformatics 22, e359.

Marshall, E. (2001). Bermuda rules: community spirit, with teeth. Science 291, 1192.

Mavrommatis, K., Ivanova, N., Shapiro, H. et al. (2007). Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. Nature Methods doi:10.1038/nmeth1043.

McMurray, A. A., Sulston, J.E. and Quail, M.A. (1998). Short-insert libraries as a method of problem solving in genome sequencing. Genome Research 8, 562-566.

Middendorf, L. R., Bruce J.C., Bruce RC, et al. (1992). Continuous, on-line DNA sequencing using a versatile infrared laser scanner and electrophoresis apparatus. Electrophoresis 13, 487-494.

Mullikin, J. C., and Ning, Z. (2003). The phusion assembler. Genome Research 13, 81-90.

Mulyukov, Z., and Pevzner, P.A. (2002). EULER-PCR: finishing experiments for repeat resolution. Pacific Symposium Biocomputing, 199-210.

Myers, E. W., Sutton, G.G., Delcher, A.L. et al. (2000). A whole-genome assembly of Drosophila. Science 287, 2196 - 2204.

Nakano, Y., Yoshida, Y. and Yamashita, Y. (1995). Construction of a series of pACYC-derived plasmid vectors. Gene 162, 157-158.

Norrander, J., Kempe, T. and Messing J. (1983). Construction of improved M13 vectors using oligodeoxynucleotide-directed mutagenesis. Gene 26, 101-106.

Park, S. R., Peng H. and Ling XS. (2007). Fabrication of nanopores in silicon chips using feedback chemical etching. . Small 3, 116-119.

Parsons, J. D. (1995). Miropeats: graphical DNA sequence comparisons. Comput. Applic. Biosci. 11, 615-619.

Pevzner, P., and Tang, H. (2001). Fragment assembly with double-barreled data. Bioinformatics, 225-233.

Pop, M., Phillippy, A., Delcher, A.L. et al. (2004). Comparative genome assembly. Brief Bioinform 5, 237-248.

Pop, M., Kosack, D. and Salzberg, S. L. (2004). Hierarchical scaffolding with Bambus. Genome Research 14, 149-159.

Reslewic, S., Zhou, S., Place, M. et al. (2005). Whole-Genome Shotgun Optical Mapping of Rhodospirillum rubrum. Applied and enviromental microbiology 71, 5511–5522.

Rhee, M., and Burns, M.A. (2007). Nanopore sequencing technology: nanopore preparations. Trends in Biotechnology 25, 174-181.

Rice, P., Longden,I. and Bleasby,A. (2000). EMBOSS: The european molecular biology open software suite. Trends Genet 16, 276-277.

Riesenfeld, C. S., Schloss,P.D. and Handelsman,J. (2004). Metagenomics: genomic analysis of microbial communities. Annual Review of Genetics 38, 525–552.

Riesenfeld, C. S., Goodman, R.M. and Handelsman, J. (2004). Uncultured soil bacteria are a reservoir of new antibiotic resistance genes. Environ Microbiol 6, 981–989.

Salzberg, S. L., Delcher, A.L., Kasif, S. et al. (1998). Microbial gene identification using interpolated Markov models. Nucleic Acids Research 26, 544-548.

Salzberg, S. L., Church, D., DiCuccio, M., et al. (2004). The Genome Assembly Archive: A new public resource. PLoS Biology 2, e311.

Schatz, M. C., Phillippy, A.M., Shneiderman, B. et.al. (2007). Hawkeye: an interactive visual analytics tool for genome assemblies. Genome Biol 8, R:34.

Shendure, J., Porreca, G.J. Reppas, N.B. et al. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309, 1728–1732.

Shizuya, H., Birren, B., Kim, U.J., et al. (1992). Cloning and stable maintenance of 300-kilobase-pair fragments of human DNA in Escherichia coli using an F-factor-based vector. Proc. Natl. Acad. Sci. USA 89, 8794-8797.

Skiena, S. (2007). Assembly for double-ended short-read sequencing technology. In "RECOMB 2007".

Solovyev, V., and Salamov, A. (1997). The Gene-Finder computer tools for analysis of human and model organisms genome sequences. Proc Int Conf Intell Syst Mol Biol 5, 294-302.

Sorokin, A., Lapidus, A., Capuano, V. et al. (1996). A new approach using multiplex long accurate PCR and yeast artificial chromosomes for bacterial chromosome mapping and sequencing. Genome Research 6, 448-453.

Staden, R. (1979). A strategy of DNA sequencing employing computer programs. Nucleic Acids Research 6, 2601-2610.

Staden, R. (1980). A new computer method for the storage and manipulation of DNA gel reading data. Nucleic Acids Research 8, 3673-3693.

Staden, R. (1982). Automation of the computer handling of gel reading data produced by the shotgun method of DNA sequencing. Nucleic Acids Research 10, 4731-4751.

Strezoska, Z., Paunesku, T., Radosavljević, D. et al. (1991). DNA sequencing by hybridization: 100 bases read by a non-gel-based method. Proc Natl Acad Sci USA 88, 10089–10093.

Tyson, G. W., Chapman,J., Hugenholtz,P. et al. (2004). Community Structure and Metabolism through Reconstruction of Microbial Genomes from the Environment. Nature 428, 37–43.

Uchiyama, T., Abe, T., Ikemura, T. et al. (2005). Substrate-induced gene-expression screening of environmental metagenomic libraries for isolation of catabolic genes. Nat Biotechnol 23, 88–93.

van Hijum, S. A. F. T., Zomer, A.L., Kuipers, O.P. et al. (2005). Projector 2: contig mapping for efficient gap-closure of prokaryotic genome sequence assemblies. Nucleic Acids Research 33, W560–W566.

Venter, J. C., Remington,K., Heidelberg et al. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. Science 304, 66–74.

Wang, H., and Branton, D. (2001). Nanopores with a spark for single molecule detection. Nature Biotechnology 19, 622–623.

Warren, R. L., Sutton, G.G., Jones, S.J. et al. (2007). Assembling millions of short DNA sequences using SSAKE. Bioinformatics 23, 500-501.

Winters-Hilt, S., Vercoutere, W., DeGuzman, V.S. et al. (2003). Highly accurate classification of Watson-Crick basepairs on termini of single DNA molecules. Biophys J 84, 967–976.

Yang, S. (2002). Comparison of genomic assemblers. Biology and Technology, February 2002, Marco Island, Florida. GCorp Inc., Waltham, MA.

Zhi, D., Keich, U., Pevzner, P., et al. (2007). Correcting base-assignment errors in repeat regions of shotgun assembly. Transactions on Computational Biology and Bioinformatics 4, 54-64.

Table 1.

Table 1. Companies involved in DNA sequencing technology development

Company names	Approach	Throughput	Read length (bp)	Application	Web site addresses	Reference
454 Life Sciences Corp.	Pyrosequencing	40 MB/4.5 hours	100-200	de-novo sequence resequencing	www.454.com	Margulies, 2005
Solexa/Illumina	Sequencing by synthesis (SBS)	1GB/run	25-35	resequencing expression profiling	www.illumina.com	Balasubramanian, 2001
Agencourt Biosciences Corp.	SBS (polony based technique)	140 bp/sec	26	resequencing	www.agencourt.com	Shendure, 2005
Helicos Bioscience Corp.	True Single Molecule Sequencing (tSMS™) SBS	1 GB/day	10-20	resequencing gene expression	www.helicosbio.com	Braslavsky, 2003
Genovoxx	AnyGene Technology™ (SBS)	n/a	15-20	resequencing	www.genovoxx.de	www.genovoxx.de
Applied Biosystems, Inc.	Sequencing by oligonucleotide ligation and detection (SOLiD™)	100-500 MB/day	25	resequencing gene expression	www.appliedbiosystems.com	Albrecht,2000
Perlegene	Sequencing by hybridrization (SBH)	1GB/run	up to 100	genotyping	www.perlegen.com	Strezoska, 1991
NABsys, Inc.	Hybridization-Assisted Nanopore Sequencing (HANS)	n/a	genome size	de-novo sequence resequencing	www.nabsys.com	Wang, 2001; Rhee, 2007; Park, 2007
ZS Genetics, Inc.	Single molecule sequencing (Tranasmission Electron Microscopy)	10^7 bp/hour	20000	gene expression	www.zsgenetics.com	Glover, 2004
Li-Cor, Inc.	Single molecule sequencing (Infrared Fluorescence)	n/a	genome size	de-novo sequencing resequencing	www.licor.com	Middendorf, 1992
Visigen Biotechnologies, Inc.	Single molecule sequencing in real time	1MB/sec	genome size	de-novo sequencing resequencing	www.visigenbio.com	Levene, 2003

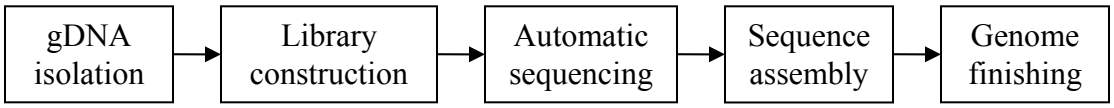


Fig.1. Whole-Genome Shotgun sequencing steps.

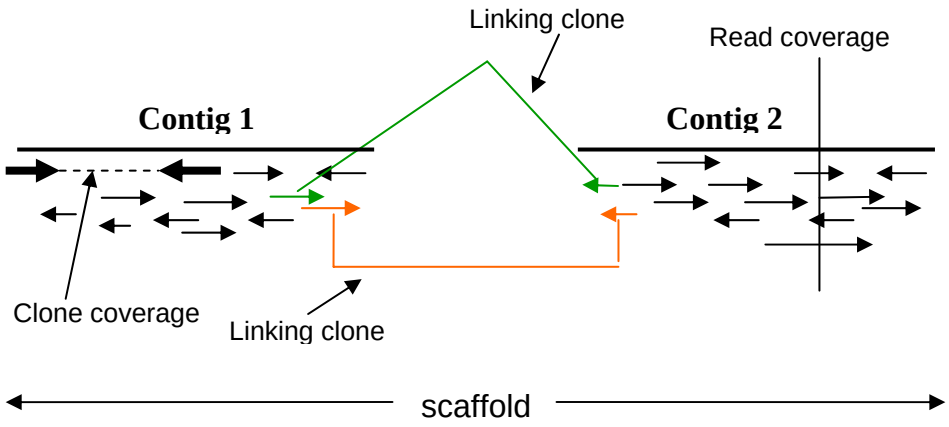
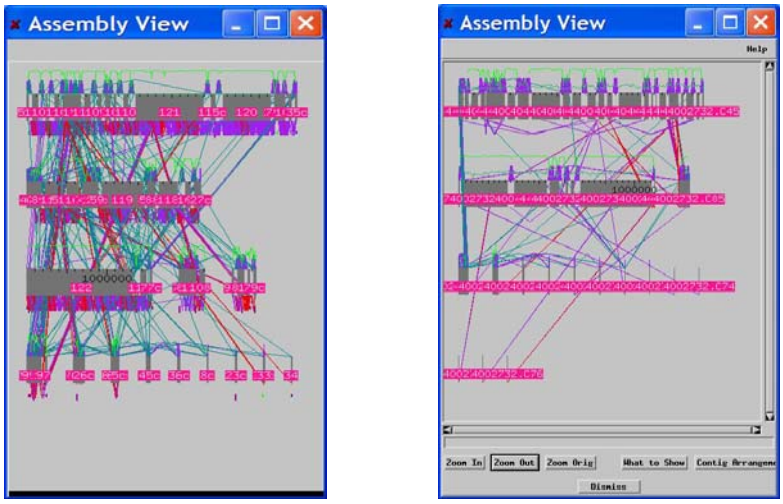


Fig.2. Read and clone coverage.



A. B.

Fig. 3. Draft assembly views of Burkholderia phytofirmans PsJN (GenBank AAUH00000000.)
A. assembly produced by phrap; B. assembly produced by PGA.
Grey boxes represent contigs. Purple lines above the contigs represent clones that span gaps between contigs and join them. Lines of different colors below or between contigs indicate misassembled paired reads.

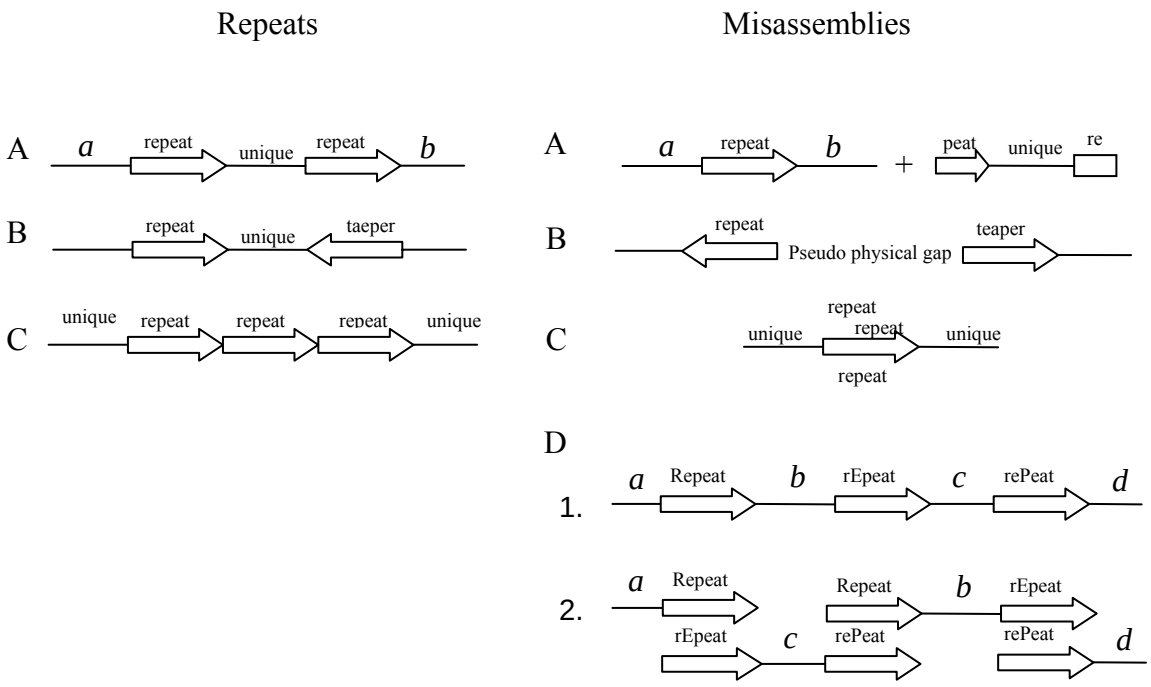


Fig.4. Types of repeats and potential misassemblies
Repeats: A – direct repeat; B – inverted repeat; C – tandem repeat
Misassemblies: A – incorrectly linked non repetitive areas *a* and *b*, located up- and downstream of the direct repeat; B – pseudo physical gap caused by incorrect placement of copies of inverted repeat; C – collapsed copies of tandem repeat; D – rearrangement of identical or nearly identical copies of repeats (1 – correct order; 2 – misassembled area).

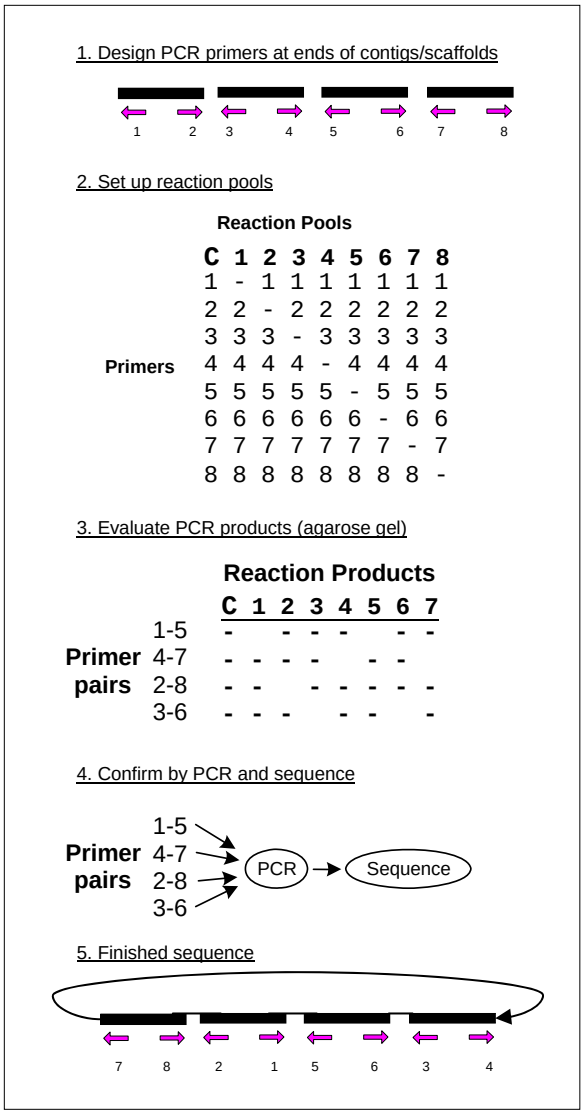
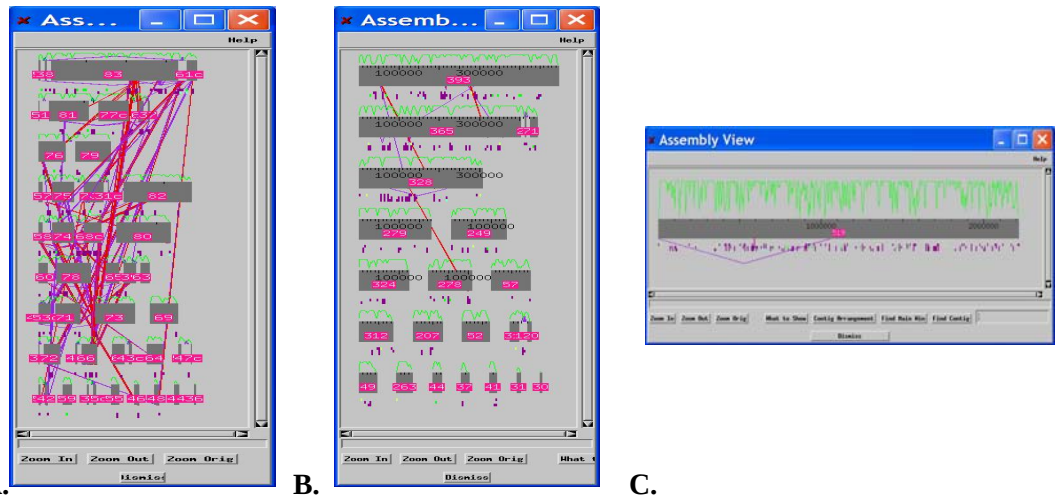


Fig.5. Multiplex PCR approach: schematic representation.
(1) Contigs are represented by black boxes; arrows indicate primers. (2) Primers mixed in each experimental reaction. C – control mix, containing all designed primers. (3) Scheme of electrophoretic analysis. (4) Confirmation by using identified primers pairs. (5) Final contig composition.



A. **B.** **C.**

Fig.6. Assembly Views of Desulforudis audaxviator.

A – draft assembly; **B –** assembly after repeat resolution and two rounds of primer walk; **C** - assembly after repeat resolution and two rounds of primer walk combined with 454 data (for details see legend for Fig.3).

Fig.7. Assembly Viewer display of Salinospora tropica CNB440 draft assembly (www.ncbi.nlm.nih.gov/Traces/assembly/viewer/assmviewer.cgi?id=gnl|TRACE_ASSM|285369). The overlapping traces comprising the assembly and detailed alignments are shown.

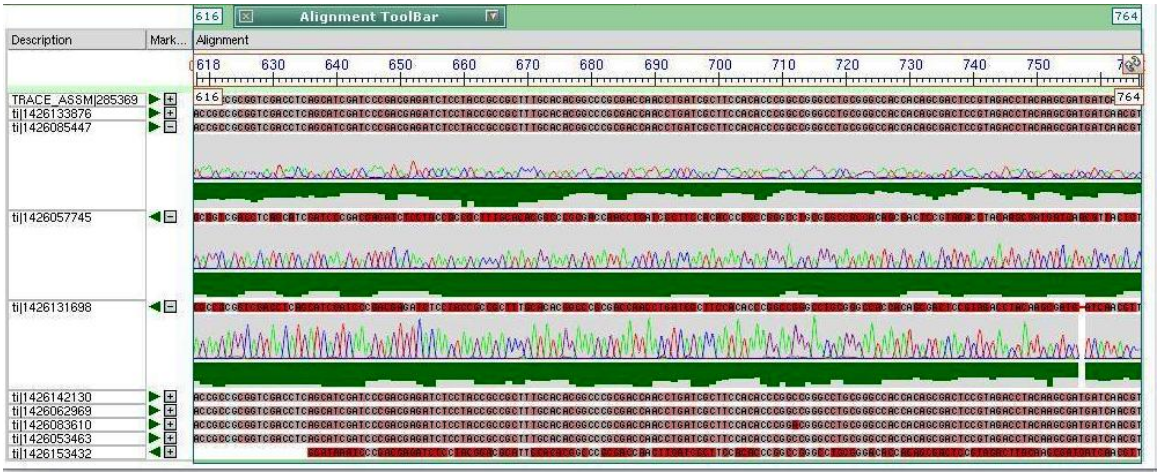


Fig.8. The underlying sequences and Traces from the draft assembly of Salinospora tropica CNB440.

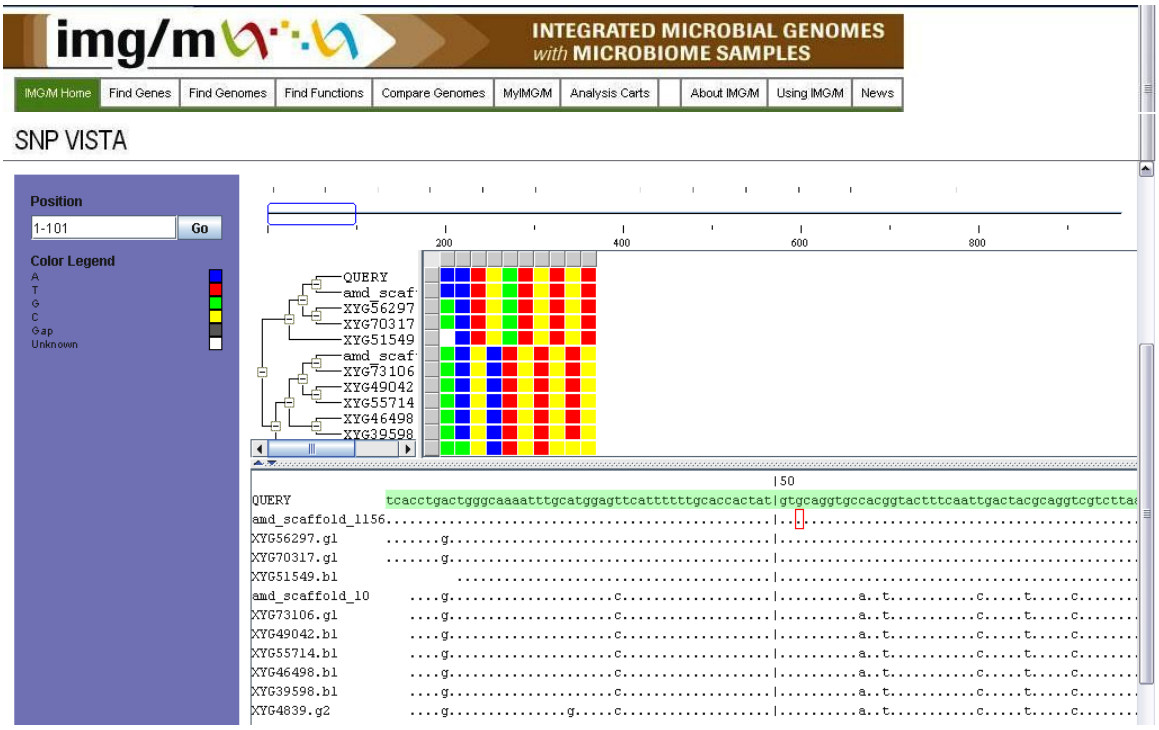


Fig.9. Snapshot of the single nucleotide polymorphism (SNP) VISTA viewer for the metagenomic project of Acid Mine Drainage.