

Next-Gen Sequence Analysis and Utility

The combination of different technologies and bioinformatics tools provide the best results

Next Generation sequencers have now been on the market for quite some time. In the early days the focus was on testing and validating the systems rather than using them in daily routine. A variety of different projects now reveal the pros and cons of each technology. They show that the use of one technology alone does not provide the best results. Rather a combination of two or three technologies provides a more complete, cost-effective analysis. In addition to sequencing, bioinformatic analysis is critically important for gaining an in-depth understanding of the biological significance of the organism.

There are two main parameters which need to be observed to ensure the in-depth understanding and interpretation of the sequence data:

- a) the use of various Next-Gen sequencing systems to take advantage of each technology, and
- b) the combination of different bioinformatic tools and their stepwise application.

The two main differences to traditional Sanger sequencing machines are the generation of giant data sets and a comparatively short read length. These two issues - high data output and short read length - are the main parameters that influence the usefulness for projects and special applications.

Illumina GA II	Roche GS FLX
Reversible terminator-based sequencing method	Pyrosequencing
18, 26 or (2x) 36 bases up to 75bp (available beginning 2009)	Ø 100 bases Ø 250 bases Ø 400 bases Titanium (available autumn 2008)

When choosing the right technology or combination of technologies, it is important to understand the aim of the project, as well as to have access to all available information about the organism. Having performed many projects at GATC, each with a different set of questions and goals, various issues have emerged.

De novo sequencing

For complete *de novo* sequencing, e.g. of prokaryotes and eukaryotes, for which there is no reference sequence from a closely related genome available, we highly recommend a hybrid strategy of GS FLX, Illumina Genome Analyzer II, and Sanger sequencing. The long Sanger reads are required for finishing and gap closure, in particular for long repeat structures, while the Roche/454 technology provides a good data backbone. Homopolymers can be solved using short reads, e.g. generated by the Illumina Genome Analyzer II technology (Fig. 1).

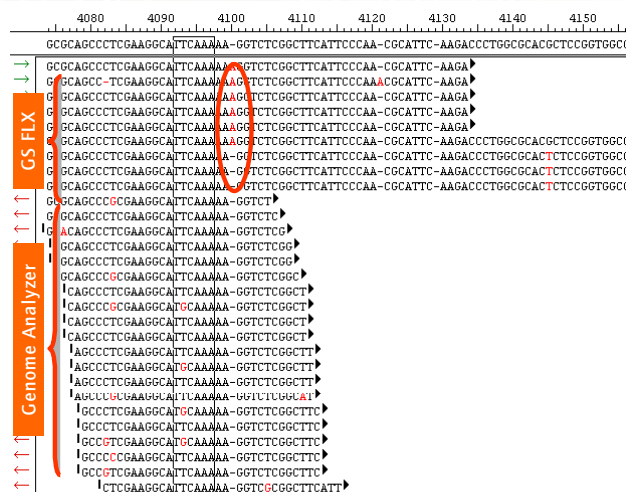


Fig. 1 | Homopolymer Validation
Hybrid Assembly using Illumina GA II and Roche GS FLX

Resequencing

If there is a reference genome available to be used for mapping (e.g. during resequencing projects), the Illumina Genome Analyzer II is the system of choice. It is the best technology for all applications for which short reads are sufficient, e.g. for an identification such as ChIP, small RNA, 3'UTR or SAGE. In addition, the huge number of reads from the Illumina Genome Analyzer II are perfect for quantification projects, e.g. expression level studies.

Mate pair / paired end

The paired end read (compared to the single read mode) provides additional data regarding the orientation of the reads and the distance of one read from its mate (Fig. 2). This information can be used to identify repeats, rearrangements and insertions/deletions.

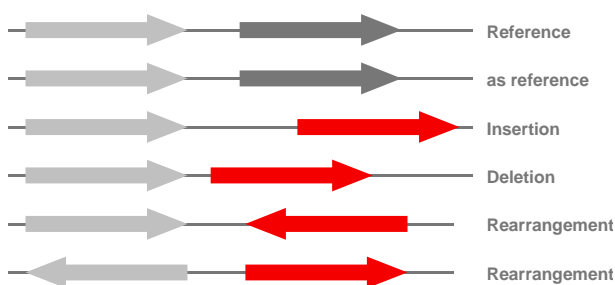


Fig. 2 | mate pairs, paired-end

The detection of such structural variants, e.g. insertions and deletions, translocation, duplications and transposons is a challenge that is being addressed by many research groups and software providers. The insert size of the paired-end libraries can be varied between 200 bp and 3 kb for the technologies.

Project Example | Resequencing on Illumina GA II

A 6.5 MB bacterial genome was sequenced on an Illumina Genome Analyzer II using the single read mode with a coverage between 20-fold and 70-fold. A detailed analysis shows that a 30-fold coverage is optimal. Above 30-fold, no additional information was obtained (Fig. 3).

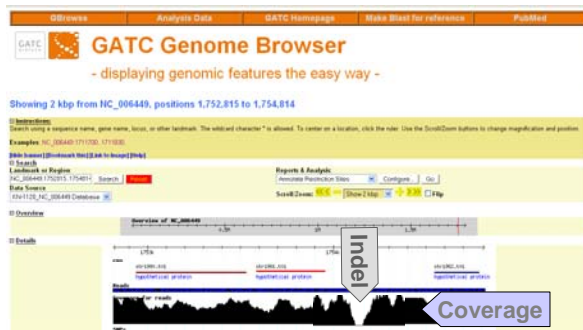


Fig. 3 | GATC Genome Browser desktop application

Using the paired end reads with an insert size of 200 bp simplified the analysis and provided additional information. Structural variants could be detected, and a surplus of reads of 20 % could be mapped to the reference genome.

Proprietary barcoding systems for DNA smaller than whole genomes

GATC has developed propriety barcoding systems that allow an additional level of parallel processing with a virtually unlimited increase in the number of samples processed. These barcoding methods are especially well suited to cDNA libraries, *de novo* sequencing of BACs, fosmids, viruses, or organelles such as chloroplasts or mtDNA, whose sequencing would otherwise be impractical and expensive.

Unlike other methods, GATC's system is suitable for use with the Roche GS FLX and Illumina Genome Analyzer II. The technique is highly efficient, resulting in 99.9% of sequences successfully tagged, which lead to high cost savings. The length of the tags is optimized so that they are sufficient for high-quality sequence sorting but also make the data loss acceptable when using Next-Gen sequencers which produce micro reads of 50 bases or less.

The tags are nucleotides which are read during the Next-Gen sequencing process. Following sequencing, samples are sorted according to their tag.

The barcoding system is exclusively available to GATC Biotech's next-generation sequencing customers.

Next-Gen data analysis and data assembly

At the moment, there is no reliable algorithm available for *de novo* assembly of short reads below 50 bp. Some software solutions have recently been developed that address these challenges. However, few have been thoroughly tested or validated and approved. Moreover, since the sequencing systems continue to double or triple their data output every few months, software applications are continuously pushed to their limits. For example, the new Illumina Genome Analyzer II provides a tremendous increase of reads per run compared to the former version. The Roche GS FLX Titanium (release announced for autumn this year) will also considerably enhance read length and number of reads per run.

As the bioinformatic analysis is crucial for the project's success, different tools must be used for different applications. Sequence information from small organelles or from BACs with many repeats need a different bioinformatic tool than data sets for hybrid assemblies.

GATC offers a wide range of bioinformatic solutions for genome assembly, transcriptome analysis or small RNA analysis. Proprietary tools are used for the analysis and handling of Next-Gen sequencing data, while third party tools are available for *de novo* assembly of sequence data from the Illumina Genome Analyzer II or improved hybrid assembly.

Data Assembly and Visualisation

SeqMan NGen™ by DNASTAR uses a unique algorithm to assemble fragment data sequenced using Illumina, Roche 454 and Sanger technologies on a desktop computer. The software can be used in a wide range of genomic assembly and re-sequencing types of Next-Gen sequencing projects.

SeqMan NGen is designed to work closely with the SeqMan Pro™ assembly module in Lasergene™. For example, reads of paired-end libraries can easily be displayed and visualise the differences between mates using a color code.



Fig. 5 | Paired-end analysis: Red shows that there is a deviation of expected insert sizes

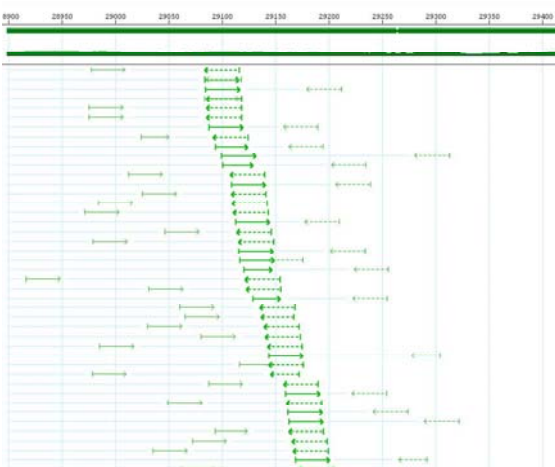


Fig. 4 | Paired-end analysis: Green shows that reads have expected insert sizes

The advantage is that it minimizes the time required to learn how to efficiently use the software and it permits users to work with features in Lasergene they have learned how to use.

Contact GATC:

customerservice@gatc-biotech.com

Germany +49 (0) 7531 81 60 29
 France +33 (0) 4 91 82 84 88
 Sweden +46 (0) 8 655 3609
 United Kingdom +44 (0) 1223 42 10 11

Data Visualisation in GATC Genome Browser

GATC Genome Browser desktop applications for data visualization allow a convenient overview of e.g. whole genome *de novo* or re-sequencing, ChIP or SAGE experiments.

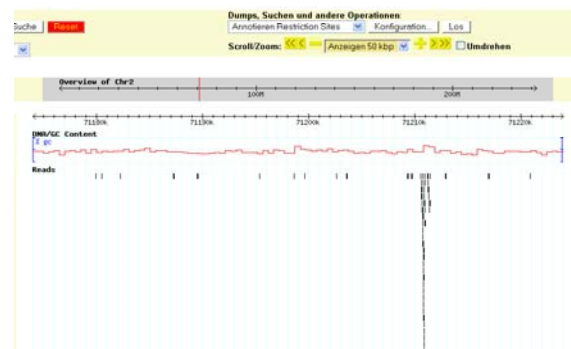


Fig. 6 | Transcriptome/ChIP-Seq Analysis

The visualization of the coverage helps to identify InDels and rearrangements within the genome. SNPs, coding regions and other annotations can also be displayed.