



Next Generation Bioinformatics Services

Data analysis for Next Generation sequencing data

by Dr. Kerstin A. Stangier | k.stangier@gatc-biotech.com

Some years ago, when the new sequencing technologies entered the market, they were called "Next Generation Sequencers". The technical capability of the systems, particularly embodied by the Roche / 454 GS FLX and the Illumina / Solexa Genome Analyzer, generate huge amounts of sequence information.

The most important difference to traditional Sanger sequencing is the tremendous increase in data output. However, the read length of all Next-Gen machines which are commercially available, is shorter than the read length generated using the Sanger method.

These two points: high data output and shorter read length, are the parameters which influence the bioinformatic analysis.

To interpretate and understand the biological relevance behind the data, Next Generation Bioinformatic services are required.

Illumina GA II	Roche GS FLX
Reversible terminator-based sequencing method	Pyrosequencing
(2x) 18, 26 or 36 bases up to 100 bases (available beginning 2009)	Ø 100 bases Ø 250 bases Ø 400 bases Titanium (available autumn 2008)
28 / 40 million reads	400,000 reads (short and standard); approx. one million reads (Titanium)

8 discrete channels per flow cell / one flow cell per run

16 samples per pico titer plate

The final goals of the sequencing project and the information available about the organism under study determine the choice of the most adequate and most effective sequencing technology or combination of technologies.

In the same way that the project's overall goals define the application of different techniques, the bioinformatic tools must be carefully selected from the available pipelines.

Project example 1 | De novo Sequencing - Hybrid strategies

The recommended approach for a *de novo* sequencing project is a combination of technologies.

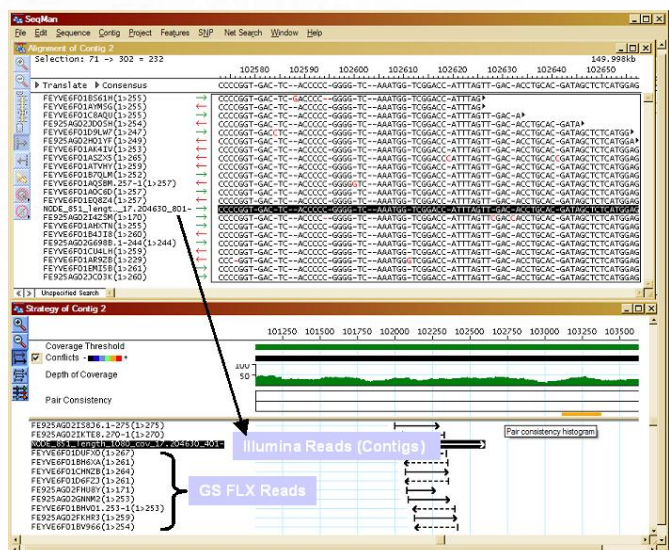
With a 15fold to 20fold coverage on a Roche GS FLX (standard or using the new Titanium kit) a good data backbone can be created. The starting material can either be whole genomic DNA or BAC libraries.

In order to achieve a deeper coverage and to resolve regions having homopolymer repeats, an additional 20fold to 30fold coverage on an Illumina Genome Analyzer II is necessary. Using the paired-end mode of the system additional sequence data is obtained with regard to the read orientation and the distance of one read to its mate.

Using libraries with different insert sizes provides a good basis for a deep and thorough quality check of the initial assembly. Therefore, the insert size of the paired-end libraries should be varied between 200 bp and 3 kb.

In order to sequence the complete DNA including repetitive regions, sequence data obtained from classic Sanger sequencing systems are highly useful. The cloned inserts used for Sanger sequencing are very helpful for finishing and gap closure using primer walking.

A variety of different bioinformatic tools is available for hybrid assemblies. SeqMan NGen™ of DNASTAR Inc. is one of the recommended tools for Next Generation sequence assembly of Illumina, Roche and Sanger data.





It is designed to work closely with the SeqMan Pro™ assembly module in Lasergene. One of the advantages is that it minimizes the time required to learn how to efficiently use the software as it permits users to work with features in Lasergene they are already familiar with. In addition, the files from SeqMan NGen can be imported into SeqMan Pro. SeqMan Pro displays dual-end sequence data when available. The alignment window enables the user to locate and to edit any nucleotide, SNP or sequence fragment and evaluate edited consensus sequences.

Project example 2 | Resequencing project using the Illumina Genome Analyzer

For a resequencing project, where a reference sequence is available, the recommended approach is a 25fold to 30fold coverage obtained from an Illumina Genome Analyzer. The short reads can be aligned and mapped to the reference sequence using several bioinformatic tools. Many mapping applications allow the dynamic variation of alignment match length to optimize the amount of informative data. Changing these parameters should be tested carefully as they have a considerable influence on the quality of the alignment. GATC uses proprietary software tools for the subsequent analysis of insertions and deletions. Just as for *de novo* sequencing projects, paired-end reads also provide additional information for a more accurate assembly for resequencing projects.

A combination of single reads and/or different library insert sizes for paired-end sequencing helps significantly to identify insertions and deletions as well as rearrangements and other differences to the reference sequence.

Beside for resequencing projects, the Illumina Genome Analyzer is the most suitable technology for all applications for which short reads are sufficient, e.g. for methods such as ChIP, small RNA, 3' UTR or SAGE. In addition, the huge number of reads from the Genome Analyzer II are well-suited for quantification projects, e.g. cDNA expression level studies.

Project example 3 | Transcriptome Analysis using the Roche GS FLX

For sequencing whole transcripts, the GS FLX is the right choice. The longer reads of the Titanium chemistry will improve the sequence assembly as it will be easier to differentiate between similar transcripts which are highly homologous, but which originate in different genes. For rare transcript searching and analysis, normalized cDNA libraries are the preferred starting material. The normalization leads to a sharp decrease in the representation of the most abundant transcripts (for

example housekeeping genes) in cDNA samples, consequently enhancing the gene discovery rate of GS FLX sequencing. The required coverage also depends on the cDNA library. If a standard library is used, a higher coverage needs to be generated. cDNA or EST clustering and alignment is a complex process that requires effort to optimize the analysis conditions. Detected SNPs can be listed either with or without annotated reference sequence:

Alignment reference format

Position	Coverage	RefBase	ConsBase	QualitySums	Bases	AvgQuality
1	1	A	A	28->A	A	1.00
2	2	G	G	62->G	GG	1.00
...						
100000	55	T	T	1597->T	TTTTTTTT...	1.00
100001	55	C	C	1558->C, 7->N	CCCCCCCC...	1.00

SNP table

a) without associated annotation

Position	Coverage	RefBase	ConsBase	QualitySums	Bases	AvgQuality
19182	32	T	G	881->G, 23->N	GGGGGGGG...	0.98
23226	5	T	C	143->C, 3->N	CCCCc	1.00
24975	36	A	T	1256->T, 19->N	TTTTTTTT	1.00

b) with associated annotation

Position	Coverage	RefBase	ConsBase	QualitySums	Bases	AvgQual	GeneID
1556955	49	G	A	1494->A	AAAA... (ID 946002)	1.00	ddpD
1557073	37	G	T	1118->T	TTTT... (ID 946028)	1.00	ddpC

Conclusion

With the Next Generation sequencing technologies, a wide range of applications are now affordable and within reach.

The systems can be used for applications ranging from metagenome and genome sequencing (*de novo* or resequencing) to transcriptome analysis (e.g. cDNA, SAGE), to regulome studies (e.g. ChIP, microRNA).

Having performed many projects, each with a different set of questions and goals, various issues have emerged:

1. Depending on the goals of the project, different sequencing technologies or combination of technologies should be applied
2. The project goals and technology used will dictate the analysis routines involving a wide range of rapidly changing bioinformatic tools.

In-depth knowledge of the strength and limitations of the Next Generation technologies and of the available analysis tools is the most crucial issue in project design and analysis.