



## New sequencers and opportunities in genome and transcriptome research

Dr. Kerstin A. Stangier | Heike Hegele

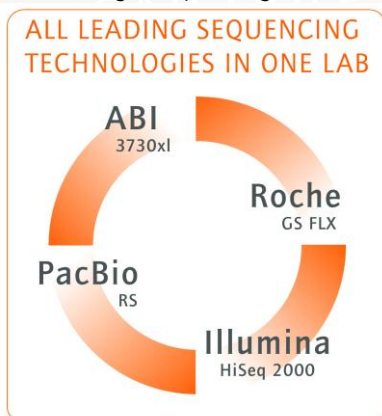
In recent years the development of new sequencing methods has opened up a wide variety of opportunities for using the DNA sequence to provide answers to scientific problems. This trend is continuing in 2011.

### Fast-moving technological developments in sequencing

In the 10 years since the first draft sequence of the human genome was published, the world of sequencing has undergone fundamental changes and further developments (Table 1). This development continues to this day, opening up more and more options for DNA analysis and thus allowing completely new approaches in genome and transcriptome research.

In addition to the automatic Sanger sequencing, which has been in existence since the middle of the 1990s, the established technologies now also include those methods which are used in instruments made by Roche, Illumina and Life Technologies. The GS FLX sequencer (Roche), the successor to the original GS 20 which was launched commercially in 2005, achieves an average read length of 400 bases and an output of more than 400 megabases per run. Its read length makes this instrument particularly suitable for the *de novo* sequencing of larger organisms, amplicon and cDNA analysis. Illumina launched its Genome Analyzer a short time later in 2006 and Life Technologies followed in 2007 with its SOLiD. Both systems have high outputs of up to 300 gigabases per run (Illumina HiSeq 2000, a more advanced version of the Genome Analyzer) and short reads (up to 150 bp for Illumina, up to 75 bp for Life Technologies) and so are primarily used for resequencing and quantitative analysis.

Last year Ion Torrent (now part of Life Technologies) and Pacific Biosciences have announced the launch of a new instrument.



### Pacific Biosciences RS

The Pacific Biosciences PacBio RS sequencer introduces several new features to sequencing: It is the first instrument to detect DNA in real time. Moreover, no amplification step is required during sample preparation, because every sequence read derives from a single DNA molecule (single molecule sequencing). A further benefit is the long read length of more than 1,000 base pairs, which can currently only be achieved with conventional Sanger-based sequencing, but not by other high-throughput instruments. These long reads lead to a new quality in the sequence data for many applications. The *de novo* sequencing of larger organisms serves as an example. The longer the reads, the more restrictive parameters for assembling individual reads into a large piece (contig) can be chosen. This results in a significantly higher quality of the assembled sequence. Larger transcripts can be covered with one single read. This means that individual reads no longer have to be assembled to complete transcripts, and the errors arising when this analysis is necessary are avoided.

The technological development of the PacBio RS can directly detect epigenetic changes such as chemically modified bases by detecting the changed kinetics when the base is incorporated [1]. Investigations of these epigenetic changes have so far concentrated on the methylation of DNA cytosine bases. These investigations can be undertaken using methods which are based on subsequent sequencing, such as bisulfite treatment or the enrichment of methylated DNA (e.g. methylated DNA immunoprecipitation (MeDIP)). None of the methods mentioned detects the modified base directly, however. Moreover, most approaches detect methylation, but not other modifications such as hydroxymethylation or similar, for example [2].

Furthermore, the analysis has so far been limited to modifications to the cytosine. However, it is known that in bacteria, in particular, adenine methylation also plays an important role. These shortcomings in the investigation of modified bases are overcome with direct sequencing, which allows new insights into regulatory elements of DNA. A further new possibility is offered by using a reverse transcriptase instead of the DNA polymerase: The direct sequencing of RNA. By using an RNA polymerase the existing RNA can be sequenced directly and in real time without rewriting it into cDNA beforehand or amplifying it.

### Possible application: Microbial ecosystems

Conventional methods for the phylogenetic characterization of microbial ecosystems have so far usually been based on DNA analysis. GATC Biotech is currently developing an approach which uses complete RNA as the starting material. It involves combining different sample preparations and sequencing methods, such as the normalization of cDNA followed by sequencing on the GS FLX, with subsequent bioinformatics analysis. This analysis allows phylogenetic studies based on total RNA. It can also serve as the basis for quantitative expression analysis with the Illumina HiSeq 2000.

The phylogenetic results are verified using standard methods, such as the derivation of primers from conserved regions of 16S RNA, PCR amplification of fragments around 400 bp long and sequencing on the GS FLX. For sequencing on the PacBio RS longer regions of 1,000 or more bases of the 16S RNA can be amplified in order to then read them with one single read. This enhances the specificity of the results.

The planned further developments of the PacBio sequencer, such as the direct sequencing of RNA and the detection of modified bases also allow a new, more detailed look into regulatory mechanisms of microbial ecosystems to which only limited access is available at present.

### Sequencing of human samples

Different combinations of sample preparation and sequencing technology are used for the sequencing of human samples (Fig. 1), the precise combination depending on the project aim. Starting from genomic DNA, libraries with different insert size can be created. Inserts of around 300 bp are preferably used to analyze SNPs and small insertions and deletions. If larger rearrangements or translocations are also of interest, libraries with inserts of 3 kb upwards are used.

If specific regions are investigated, enrichment methods or conventional PCR are used. What all preparations have in common is that the Illumina HiSeq 2000 is used for the subsequent sequencing. This sequencer currently provides the greatest output and highest throughput as well as the highest quality. GATC Biotech's proprietary barcoding system makes it possible to perform a simultaneous analysis of different samples in pools.

It is also possible to pool UTR libraries used for expression profiling, or ChIP, MeDip and smallRNA samples. Thus regulatory studies can be carried out very efficiently.

The long reads of the Pacific Biosciences RS technology also facilitate the identification of larger insertions and deletions.

The so-called "strobe reads" (shorter reads, which are distributed at specific distances on large fragments) facilitate the identification of further genetic variations, such as large rearrangements and translocations. "Strobe read" sequencing obviates the need for the often difficult preparation of large inserts libraries. Here the direct sequencing of modified bases, especially methylated cytosine, will provide a more detailed insight into the mechanisms of cancer development, for example.



### Outlook

Top quality results are only achieved when different sample preparations, the latest sequencing technologies and adapted bioinformatics tools are combined. The next generation of sequencing instruments will be fully commercialized in 2011. This is by no means the end of the road for innovations, however, because companies such as Oxford Nanopores, Roche and IBM have announced further developments on the way to the "\$1,000 human genome".

### Literature

- [1] Flusberg B A, Webster, Dale R, Lee J H et al. (2010) Direct detection of DNA methylation during single-molecule, real-time sequencing. *Nat Methods* 7(6): 461-465
- [2] Harris R A, Wang T, Coarfa C et al. (2010) Comparison of sequencing-based methods to profile DNA methylation and identification of monoallelic epigenetic modifications. *Nat Biotechnol.* 28: 1097-1105

### Customer Service

(D) +49 (0) 7531 81 60 68  
 (F) +33 (0) 4 91 82 84 88  
 (GB) +44 (0) 207 691 4921  
 (S) +46 (0) 8 655 3609

customerservice@gatc-biotech.com  
 www.gatc-biotech.com



### Author



Dr. Kerstin A. Stangier  
 Director Business  
 Development



Heike Hegele  
 Product Manager Next  
 Generation Sequencing

**Table 1: Technical details of the leading sequencing systems**

Company	Illumina	Life Technologies	Pacific Biosciences	Roche Diagnostics
machine	HiSeq 2000	ABI 3730 xl	PacBio RS	GS FLX
since	2006 (GA by Solexa)	1987 (ABI 3700 by Applied Biosystems)	2010	2005 (GS20 by 454 Life Science)
device	flowcell w/ channels	capillaries	SMRT Cell w/ zero-mode waveguides	PicoTiterPlate w/ wells
nucleotides	four-colour fluorescence labeled, base linked	four-colour fluorescence labeled, base linked	four-colour labeled, phospholinked	natural nucleotides
amplification	bridging PCR	PCR	none	emulsion PCR
sequencing	cyclic reversible termination	non-reversible chain termination	single molecule real time sequencing	pyrosequencing
signal	laser-induced fluorescence	laser-induced fluorescence	laser-induced fluorescence pulses	light emission
detection	CCD camera	CCD camera	single-photon sensitive CCD array	CCD camera
raw data format	intensity files	electropherograms	fluorescence trace pulses	flowgrams