Old Herborn University Seminar Monograph 23:
Intestinal microbiomics: novel indicators of health and disease.
Editors: Peter J. Heidt, Johannes Snel, Tore Midtvedt, and Volker Rusch.
Old Herborn University Foundation, Herborn-Dill, Germany: 7-13 (2010).

EXPLORING MICROBIAL COMMUNITIES USING NEXT-GENERATION DNA SEQUENCING PLATFORMS

LARS ENGSTRAND

Department of Bacteriology, Swedish Institute for Infectious Disease Control, and
Department of Microbiology, Tumour and Cell Biology,
Karolinska Institute, Stockholm, Sweden

SUMMARY

The sequencing of the human genome constituted a starting point in the understanding of human biology at a global scale, yet today there is a growing agreement that human health and disease cannot be understood without considering the microbial communities (microbiome). Now, as the human microbiome project has been launched, a number of international consortiums are starting up with efforts to explore the role of the human microbiota in health and disease. During the last two years molecular microbiology has revolutionized the landscape of microbiology and will continue to do so by providing new solutions for microbe identification and characterization. Next-generation sequencing will open up new areas of research for people in the field of intestinal microbiomics. The high-throughput sequencing platforms we now have access to will hopefully also help to increase our understanding of host-bacteria interactions, immune maturation and mechanisms behind chronic disease development in the gut. The analysis of data resulting from a large-scale sequencing project requires the use of bioinformatics including standard blast analysis, annotation or clustering and assembly competence. In addition, an interdisciplinary approach must be taken on comprising medical, computational, and biotechnology expertise focusing on understanding the human microbial communities and their effect on human health. Application of different "omics"-methods and computational systems biology methods to unique biobanks will be required in order to map out the human microbiome as well as the human cellular machinery interacting it.

TECHNOLOGY PLATFORMS

A human body contains more bacterial than human cells and harbours several microbial ecosystems. The microbes living there and their interactions with the human cells are key components to our health. Malfunctions in these ecosystems clearly lead to everyday health problems currently generating high costs for society and are implicated in health threats such as cancer, type II diabetes, inflammatory bowel disorder, allergy, and obesity. The human microbiome is a complex system with perhaps 1000 microbial species in-

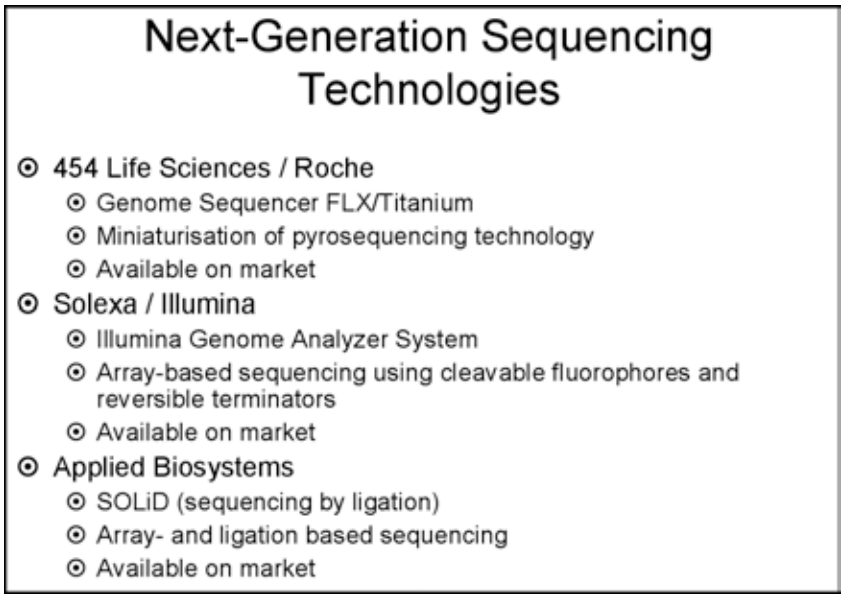

Figure 1: Brief summary of high-troughput technologies available today.

volved, each having around 1000 genes. On top of that, the number of individuals of a species in the microbiome, as well as the pan-genome of a species, is in flux and affected by antibiotic treatment, diet, *et cetera*.

Growing demand in this area of research has fuelled the development of more efficient genomic sequencing methods (*Mardis*, 2008; *Shendure* and *Ji*, 2008). Such methods are already several orders of magnitude more efficient than the Sanger capillary-array electrophoresis machines (for example Applied Biosystems 3730xl) that were used in the human genome project (Figures 2 and 3). Massively parallel DNA sequencing platforms have not only reduced the cost of DNA sequencing but also moved the technology from major genome centres to individual investigators. The new platforms will dramatically accelerate biological and biomedical research, by enabling the comprehensive analysis of genomes to become inexpensive, routine and widespread. Three commercial systems are briefly described in Figure 1.

Roche/454 FLX pyrosequencer

Multiple whole prokaryote genomes can easily be sequenced with the 454 FLX and Titanium systems (*Margulies* et al., 2005; *Ronaghi* et al., 1996). This high-throughput sequencing in real time technology provides long reads (up to 400 base pairs) which facilitates the completion of near-finished draft sequences on a single instrument run. Large-size genomic DNA samples are randomly fragmented into small 300- to 800-base-pair fragments for shotgun sequencing. Addition of adapters to the fragments creates a library of DNA fragments, which is immobilized on DNA capture beads whereafter a PCR amplification takes place in water-in-oil microreactors, resulting in millions of copies of the template. Finally the microreactor is broken and beads carrying single-stranded DNA templates

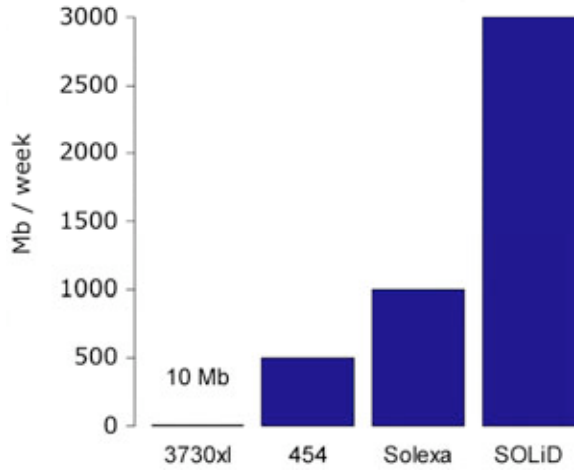

Figure 2: Comparison of yesterdays and todays technologies (Mb output per week)

are individually sequenced on a pico titre plate device. The generated sequences are then assembled into a number of unordered contigs using specific assembler software. Finally a consensus sequence is generated.

The sequencing depth achieved with 454 FLX Titanium sequencing systems ensures accurate characterization of microbial or bacterial diversity, sensitive detection of even rare mutations, and rapid discovery of disease causing agents (*Rothberg* and *Leamon*, 2008). Furthermore, this system for ultra-high-throughput DNA sequencing is used for de novo sequencing and resequencing of genomes, metagenomics and targeted sequencing of DNA regions of interest. The newest version generates up to 500 million bases per 10-hour instrument run. The key-advantage of this technology is read-length (up to 400 base pairs which is necessary in *de novo* assembly and metagenomics). However, a major limitation is that no prevention of multiple

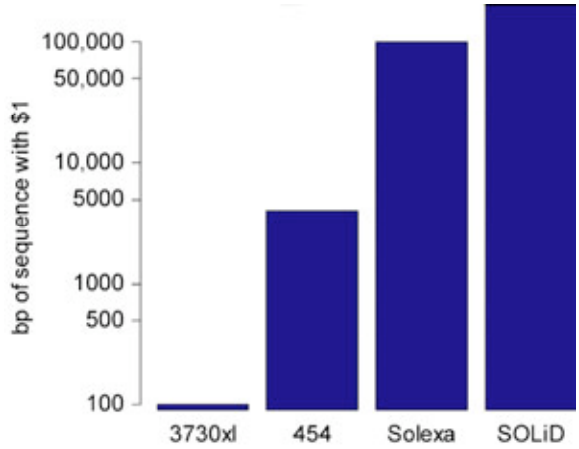

Figure 3: Comparison of different technologies and cost (number of base pairs per US $).

incorporations at a given cycle is provided which leads to homopolymer errors.

The technology has enabled a number of peer-reviewed studies in diverse research fields such as cancer and infectious disease research, drug discovery, marine biology, anthropology, palaeontology and many more. The value of the 454 system for bacterial sequencing applications is underscored by a number of important studies including a *M. tuberculosis* study that resulted in the identification of the first tuberculosis-specific drug candidate in 40 years (*Andries* et al., 2005). The 454 pyrosequencing has so far been the method of choice for exploring the human microbiota (*Andersson* et al., 2008).

Massive parallel 454-tagsequencing targeting the 16S rRNA gene

The most commonly utilized target for microbiome analysis is the ubiquitous gene coding for small sub-unit ribosomal RNA. The 16S rRNA gene (16S rDNA) has historically proved to be the most accurate gene for studies of bacterial diversity, evolution, as well as phylogenetic analysis. The 16S rDNA consists of consensus sequences, universal for all procaryotes, and variable sequences that are specific for particular groups or species of bacteria. Hypervariable sequences, that may be unique for certain strains within a species, are contained within the variable areas. We have developed a novel approach to study the human microbiota using the 454-pyrosequencing platform targeting 16S rDNA (*Andersson* et al., 2008 and Figure 4).

Illumina/Solexa Genome Analyzer

Illumina sequencing technology, or the Solexa platform, allows for selection of any SNP or probe, enabling dense, uniform coverage across the genome and the ability to target any genomic region (*Turcatti* et al., 2008; *Adessi* et al., 2000). This platform is based on massively parallel sequencing of millions of fragments using a reversible terminator-based sequencing chemistry. The technology, together with a software application, allows a scalable system that many consider cost-effective and accurate. It relies on the attachment of randomly fragmented genomic DNA to an optically transparent surface. Attached DNA fragments are extended and bridge amplified to create an ultra-high density sequencing flow cell with ≥50 million clusters, each containing ~1,000 copies of the same template. These templates are sequenced using a four-color DNA sequencing-by-synthesis technology that employs reversible terminators with removable fluorescent dyes. This approach ensures high accuracy and true base-by-base sequencing, eliminating sequence-context specific errors and enabling sequencing through repetitive sequences. After completion of the first read, the templates can be regenerated *in situ* to enable a second >36 bp read from the opposite end of the fragments. A paired-end module directs the regeneration and amplification operations to prepare the templates for the second round of sequencing. Once the original templates are cleaved and removed, the reverse strands undergo sequencing-by-synthesis. The second round of sequencing occurs at the opposite end of the templates, generating >36 bp reads for a total of >3 Gb of data which is an obvious advantage when sequencing large genomes. The short read-length (35 bp) is a limitation and will not provide enough information for 16S identification, but compared to 454 pyrosequencing, homopolymer errors are less of an issue with this technology.

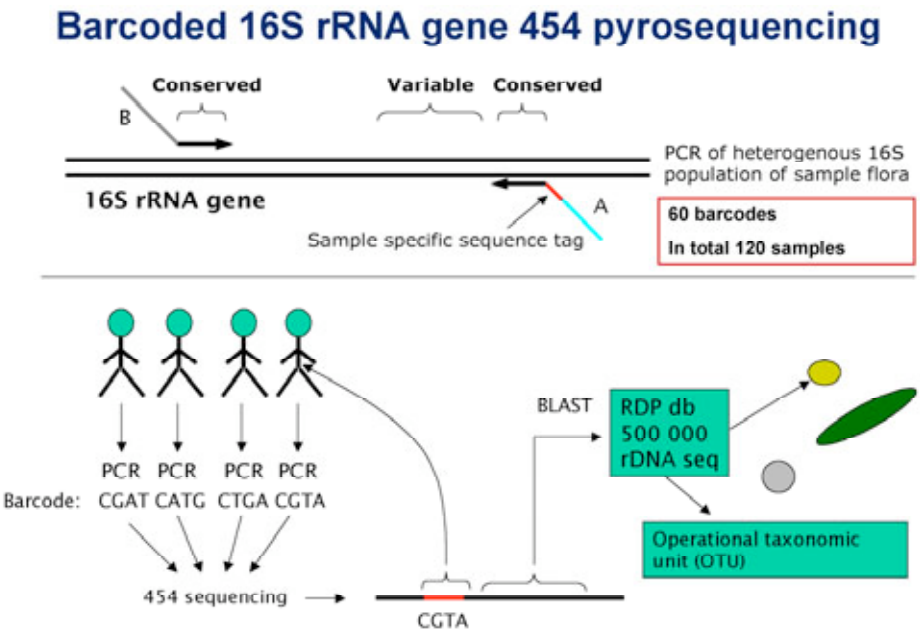

Figure 4: DNA is extracted from each sample and the 16 S rRNA gene contents are amplified by PCR using conserved 16S primers. By using short sample-specific sequence tags, incorporated during the initial 16S PCR reactions, each sequence obtained on the pico titer plate is traced back to its original sample. Thousands of PCR products from each sample are sequenced providing a representative picture of the composition of the microbiota in each individual. The Ribosomal Database Project including more than 500 000 rDNA sequences provides information for species identification (OTU). The capacity of the method is today 120 samples per run at an approximate cost of 80 USD per sample. The 454 pyrosequencing platform allowed us to explore the gut microbiota with a sequencing depth that ensures an accurate characterization. Similar approaches for analyses of the gut microbiota have been reported over the past 6 months (*Huse* et al., 2008; *Dethlefsen* et al., 2008; *Keijser* et al., 2008).

Applied Biosystems SOLiD™ System

Sequencing by ligation generates DNA by measuring the serial ligation of an oligonucleotide. This technology is used in the SOLiD system (*Dressman* et al., 2003; *Shendure* et al., 2005). All fluorescently labelled oligonucleotide probes are present simultaneously and compete for incorporation. After each ligation, the fluorescence signal is measured and then cleaved before another round of ligation takes place. The SOLiD system is a massively parallel genomic analysis platform that supports a wide range of applications. The flexibility of two independent flow cells allows multiple experiments in a single run. The SOLiD system can cost effectively complete large-scale sequencing and with a reference sequence for a microorganism, it is possible to perform comparative sequencing or re-sequencing to characterize the genetic diversity within the organism's species or between closely related species. The throughput is greater than 30 Gb per run but the read-length is limited to 35 bp and consequently not useful for 16S identification.

COMPARISON OF NEXT-GENERATION SEQUENCING TECHNOLOGIES

	454	Solexa	SOLiD
Read length	250-400 bp	25-35 bp	25-35 bp
Reads	1 M	30 M	90 M
Data	500 Mb	3 Gb	30 Gb
Scale-up of # reads	+	+++	+++
Future increase of read length	++	+	+
Access to instruments	+++	+++	+++
Drawbacks	High error rate for homopolymers	Short read length	Short read length
Advantages	Long read length	Easy to scale up	Easy to scale up

CONSIDERATIONS AND CHALLENGES

The cost of sequencing is steadily decreasing, facilitating sequencing of increasing number of microbes and even metagenomics. Many envision that the near future improvements of cost efficiency for DNA sequencing will effectively eliminate sequencing as a bottleneck in biomedical research. The next-generation DNA sequencing platforms will without doubts be applied for a variety of goals within the human microbiome research field. A list of applications that will raise new challenges for experimental design and interpretation of the results have recently been described (*Shendure* and *Ji*, 2008; *Mardis*, 2008; *Rothberg* and *Leamon*, 2008). However, the large amount of data achieved by these instruments must lead to biologically meaningful insights and hopefully also to clinical strategies in the future.

LITERATURE

Adessi, C., Matton, G., Ayala, G., Turcatti, G., Mermod, J.J., Mayer, P., and Kawashima, E.: Solid phase DNA amplification: Characterisation of primer attachment and amplification mechanisms. Nucleic Acids Res. 28, E87 (2000).

Andersson, A.F., Lindberg, M., Jakobsson, H., Bäckhed, F., Nyrén, P., and Engstrand, L.: Comparative analysis of human gut microbiota by barcoded pyrosequencing. PLoS ONE 3, e2836 (2008).

Andries, K., Verhasselt, P., Guillemont, J., Göhlmann, H.W., Neefs, J.M., Winkler, H., Van Gestel, J., Timmerman, P., Zhu, M., Lee, E., Williams, P., de Chaffoy, D., Huitric, E., Hoffner, S., Cambau, E., Truffot-Pernot, C., Lounis, N., and Jarlier, V.: A diarylquinoline drug active on the ATP synthase of *Mycobacterium tuberculosis*. Science 307, 223-227 (2005).

Dethlefsen, L., Huse, S., Sogin, M.L., and Relman, D.A.: The pervasive effects of an antibiotic on the human gut microbiota, as revealed by deep 16S rRNA sequencing. PLoS Biol. 6, e280 (2008).

Dressman, D., Yan, H., Traverso, G., Kinzler, K.W., and Vogelstein, B.: Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. Proc. Natl. Acad. Sci. USA 100, 8817-8822 (2003).

Huse, S.M., Dethlefsen, L., Huber, J.A., Mark Welch, D., Relman, D.A., and Sogin, M.L.: Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. PLoS Genet. 4, e1000255 (2008).

Keijser, B.J., Zaura, E., Huse, S.M., van der Vossen, J.M., Schure,n F.H., Montijn, R.C., ten Cate, J.M., and Crielaard, W.: Pyrosequencing analysis of the oral microflora of healthy adults. J. Dent. Res. 87, 1016-1020 (2008).

Mardis, E.R.: Next-generation DNA sequencing methods. Annu. Rev. Genomics Hum. Genet. 9, 387-402 (2008).

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z., Dewell, S.B., Du, L., Fierro, J.M., Gomes, X.V., Godwin, B.C., He, W., Helgesen, S., Ho, C.H., Irzyk, G.P., Jando, S.C., Alenquer, M.L., Jarvie, T.P., Jirage, K.B., Kim, J.B., Knight, J.R., Lanza. J.R., Leamon, J.H., Lefkowitz, S.M., Lei, M., Li, J., Lohman, K.L., Lu, H., Makhijani, V.B., McDade, K.E., McKenna, M.P., Myers, E.W., Nickerson, E., Nobile, J.R., Plant, R., Puc, B.P., Rona,n M.T., Roth, G.T., Sarkis, G.J., Simons, J.F., Simpson, J.W., Srinivasan, M., Tartaro, K.R., Tomasz, A., Vogt, K.A., Volkmer, G.A., Wang, S.H., Wang, Y., Weiner, M.P., Yu, P., Begley, R.F., and Rothberg, J.M.: Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376-380 (2005).

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M., and Nyrén, P.: Real-time DNA sequencing using detection of pyrophosphate release. Anal. Biochem. 242, 84-89 (1996).

Rothberg, J.M. and Leamon, J.H.: The development and impact of 454 sequencing. Nat. Biotechnol. 26, 1117-1124 (2008).

Shendure, J. and Ji, H.: Next-generation DNA sequencing. Nat. Biotechnol. 26, 1135-1145 (2008).

Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., Wang, M.D., Zhang, K., Mitra, R.D., and Church, G.M.: Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309, 1728-1732 (2005).

Turcatti, G., Romieu, A., Fedurco, M., and Tairi, A.P.: A new class of cleavable fluorescent nucleotides: Synthesis and optimization as reversible terminators for DNA sequencing by synthesis. Nucleic Acids Res. 36, e25 (2008).