DOI: 10.1002/anie.201003880

The Development of DNA Sequencing: From the Genome of a Bacteriophage to That of a Neanderthal**

*Uschi Sundermann, Susanna Kushnir, and Frank Schulz**

fluorescence · gene sequencing · genetic code · nucleotides

In 1977 Sanger et al. reported the first genome sequence ever determined: the roughly 5000 base pairs of a bacteriophage genome.[1] In the same year Sanger published his dideoxy method for DNA sequencing,[2a] an experimental technique which in the following decades would revolutionize modern biochemistry and bring Sanger his second Nobel Prize in Chemistry.[2b] The aim of deciphering the human genome spurred a tremendous jump in the development of the Sanger sequencing technology (Table 1). The Human Genome Project (HGP), initiated in 1990, led to a factorylike upscaling of sequencing capacities in the participating institutes. Through optimization and automatization of each step of the sequencing process, the elucidation of complex genomes slowly came within reach.

In the early 1990s, the improvements in Sanger technology enabled the sequencing of small bacterial genomes[3] and already in 1996, the genome of *Saccharomyces cerevisiae*, baker's yeast, was described.[4] In 2001, one decade after its project's commencement, the first draft of the human genome was published independently and in parallel by the Human Genome Consortium and Celera Genomics.[5,6] This initial draft was brought close to completion in 2004.[7] The HGP certainly was a milestone in biochemistry, but the methods applied were time-consuming and expensive. A broader application, whether in personalized medicine or for the routine sequencing of microorganisms, still seemed too ambitious at that time. Despite the significant drop in sequencing costs during the HGP from approximately 10 US$ to 0.09 US$ per nucleobase (see Table 1), the total costs of the human genome sequencing summed up to roughly three billion US$.[8–10]

Table 1: Comparison of cost and expenditure of time for different sequencing techniques.[10] Only commercially available techniques of the first and second generation are considered. Mbp: 1×10^6 base pairs; Gbp: 1×10^9 base pairs.

Year	$/Mbp	Days/Gbp	Comment
1977	n.d.	n.d.	dideoxy method
1990	10×10^6	n.d.	HGP is launched
1995	1×10^6	n.d.	introduction of capillary electrophoresis
1998	5×10^5	n.d.	
2002	9×10^4	260[a]	end of HGP
2005	60	3.1[b]	Roche 454 GS FLX
2006	2	2.3[b]	Illumina Solexa 1G
2007	2	1.6[b]	AB SOLiD System

[a] Time requirement is estimated for the whole capacity of the HGP. [b] Time requirement is calculated for a single instrument. n.d. = not determined.

It was in early 2010, only few years after its commencement, that the Neanderthal Genome Project, led by Svante Pääbo in Leipzig, was brought to completion.[11] The approximately 3.2 billion base pairs of the Neanderthal genome were deciphered from 40000-year-old small fragments of ancient DNA. Clearly, the starting position for this project was significantly less favorable than for the HGP, owing to the poor condition of the old genetic material and its comparably limited availability. But an impressive jump in DNA-sequencing technology pushed forward a significantly faster and more economical genome analysis of our prehistoric relative.

This jump in development is heralding a new era in biochemical research. The very first steps towards a truly broad application of genome sequencing were taken independently through "sequencing by synthesis" developed by 454 Life Sciences[12] led by Jonathan Rothberg, and through "multiplex polony sequencing" developed by Shendure et al.[13] Both groups used fluorescence detection, which enabled simultaneous sequencing of several hundred thousand DNA fragments from tiny amounts of template—a major improvement over the 96-well format used in the dideoxy method. This impressive parallelization was one reason that the first version of the genome sequencer of 454 Life Sciences was already operating at a sixth of the cost of the Sanger method. However, early in its development, sequencing by synthesis experienced initial difficulties. Of major

[*] U. Sundermann, Dr. S. Kushnir, Prof. Dr. F. Schulz
Technische Universität Dortmund, Fakultät für Chemie
Otto-Hahn-Strasse 6, 44221 Dortmund (Germany)
Fax: (+49) 231-133-2498
E-mail: frank3.schulz@tu-dortmund.de
Homepage: http://www.chemie.tu-dortmund.de/schulz

U. Sundermann, Prof. Dr. F. Schulz
Max-Planck-Institut für Molekulare Physiologie
Abteilung für Chemische Biologie
Otto-Hahn-Strasse 11, 44227 Dortmund (Germany)

[**] We acknowledge generous financial support from the Beilstein Institut zur Förderung der chemischen Wissenschaften and the Fonds der Chemischen Industrie (Liebig stipend to F.S. and a predoctoral stipend to U.S.). U.S. is a fellow of the International Max-Planck-Research School of Chemical Biology.

concern was the comparably short read length and the relatively low accuracy of the sequencing reactions; in both, the didesoxy technique was superior. In contrast in 2005 "sequencing by synthesis" method was only in its infancy. Before long, the read length was increased from 100 to 250, to 400–500 bases using today's version.[14] Introduced only shortly after the 454 Genome sequencer, two further competing instruments entered the market, the Illumina Solexa Plattform and Applied Biosciences SOLiD Sequencing. Both techniques promise to reduce the sequencing costs and increase the sequencing throughput even further. However, these advances come at the cost of a reduced sequencing read length, which hampers the application in repetitive sequence areas and complicates the de novo sequencing of genomes in certain cases.

Together, these techniques make up the so-called second generation of sequencing technology—with partially complementary strengths and weaknesses (the current distribution of the systems in the literature is shown in Figure 1). All second-generation sequencing techniques share an initial step in which the genomic DNA is fragmented, analogous to Sanger sequencing. Subsequently, the resulting fragments are not cloned and amplified in vivo, as in the old technique, but made available for sequencing through a polymerase chain reaction (PCR) step (see Figure 2). The ensuing sequencing step follows different procedures, but in each case high-resolution fluorescence cameras are used as detectors. This enables tremendous parallelization and a concomitant high throughput at a rather low cost per sequenced base pair (see Table 1). The Neanderthal Genome Project relied on a combination of the 454 GS FLXD and Illumina Solexa GAII systems.[11] From ancient bones, recovered from four different sites, several hundred milligrams of bone substance were extracted and used for DNA isolation.

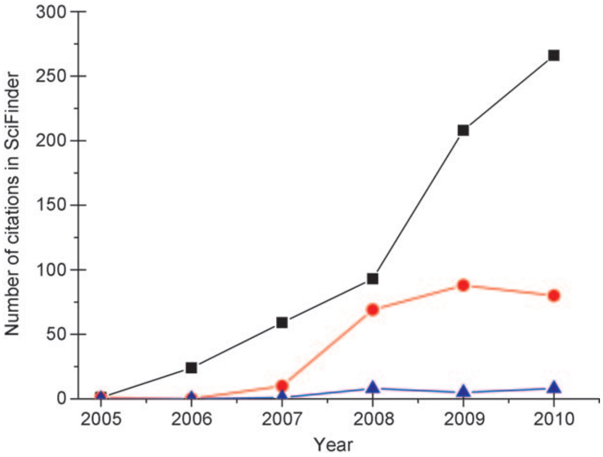

Figure 1. Result of a SciFinder literature search for commercialized sequencing technologies of the second generation as the number of listed citations per year (as of June 2010, results for 2010 extrapolated). Black squares: Roche 454 GS FLX; red circles: Illumina Solexa Sequencer; blue triangles: Applied Biosystems SOLiD System.

A crucial factor in the Neanderthal Genome Project was the very high risk of contaminating DNA from human or other sources; specially developed procedures to eliminate this potential problem were needed. One procedure was to use project-specific sequences as tags for the immobilization of the genomic DNA fragments. Furthermore, initial investigations showed that approximately 99% of the recovered DNA was of microbial origin, a consequence of the old age of the specimens. The Pääbo group used restriction enzymes that preferentially cleave GC-rich bacterial DNA sequences to reduce the level of microbial contamination. This was successful in enriching the Neanderthal DNA four- to sixfold but led to an unavoidable loss of genome coverage; this was accepted as it would lead to a significant improvement of the quality of the sequence at onefold coverage. Another, probably more important, aspect was the potential contamination by human DNA, which led to strong criticism of the project.[15] Pääbo and co-workers used two different statistical analyses to quantify the contamination level. Both techniques revealed a contamination level of less than 1%.

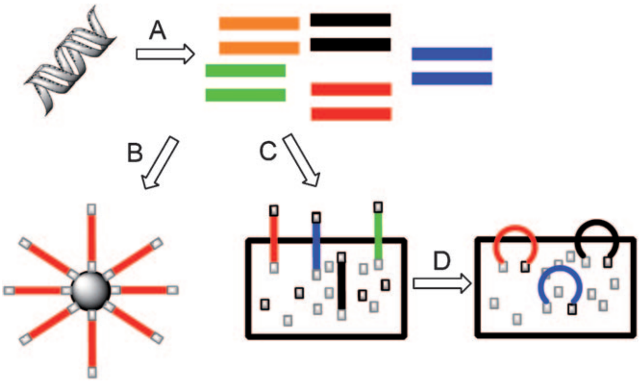

Figure 2. Overview of individual steps in second-generation sequencing technology. Initially, genomic DNA is fragmented by shear forces (A). Subsequently, the obtained fragments are ligated in vitro to small DNA fragments and, in the case of 454 and polony sequencing, immobilized on solid beads (B). The DNA is amplified by an emulsion PCR, thus avoiding the "cloning bias" intrinsic in Sanger technology. The solid beads are scattered and the sequencing reaction is carried out without chromatographic steps employing fluorescence cameras. In the Solexa technology, the oligonucleotide-tagged genomic DNA fragments are hybridized to a solid surface (C) and then amplified by bridge *PCR* (D). This leads to clusters of identical DNA sequences on the surface, which can afterwards be sequenced and detected by fluorescence microscopy.

To commence the investigation of the obtained Neanderthal sequence, it was compared to the genomes of five modern humans (one San from South Africa, one Yoruba from West Africa, one Papua New Guinean, one Han Chinese, and one French European), the chimpanzee genome, and the human reference genome. Interestingly, a significant relationship between the Neanderthal genome and the European and the Asian genomes was detected, but not to the African genomes. The flow of genetic information was directed towards the modern human and made up to several percent of today's genomes of Asians and Europeans. Overall, Pääbo et al. find the Neanderthal genome to be very similar to the modern human genome; the Neanderthals were close relatives to the anatomically modern humans and substantial interbreeding most likely occurred after humans left sub-Saharan Africa.

The decryption of the Neanderthal genome shows unequivocally the impressive capacity of the second generation of sequencing technology which, within a short time, has found widespread application. Current applications range from the analysis of metagenomes and transcriptomes, to *"deep sequencing"* for the identification of single nucleotide polymorphisms, to de novo sequencing as in the case of the Neandertal genome. However, there is still a long way to go to reach the frequently discussed 1000$ genome.[9] For this reason, a third generation, also called the "next next generation", of sequencing technologies is under intense investigation.[10] None of these techniques has reached the market yet, not to mention applications beyond model experiments. But if technology development keeps up the current pace, in a few years from now genome sequencing may well have become a routine technique in biochemical laboratories. The impact of this development on science as well as on society is not yet clear. But certainly it will be wide-reaching.

Received: June 25, 2010
Published online: September 22, 2010

[1] F. Sanger, G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, J. C. Fiddes, C. A. Hutchison, P. M. Slocombe, M. Smith, *Nature* **1977**, *265*, 687 – 695.

[2] a) F. Sanger, S. Nicklen, A. R. Coulson, *Proc. Natl. Acad. Sci. USA* **1977**, *74*, 5463 – 5467; b) F. Sanger, *Angew. Chem.* **1981**, *93*, 937 – 944.

[3] F. R. Blattner, G. Plunkett, III, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, Y. Shao, *Science* **1997**, *277*, 1453 – 1462.

[4] R. A. Clayton, O. White, K. A. Ketchum, J. C. Venter, *Nature* **1997**, *387*, 459 – 462.

[5] International Human Genome Consortium, *Nature* **2001**, *409*, 860 – 921.

[6] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. D. Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R.-R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Y. Wang, A. Wang, X. Wang, J. Wang, M.-H. Wei, R. Wides, C. Xiao, C. Yan et al., *Science* **2001**, *291*, 1304 – 1351.

[7] International Human Genome Consortium, *Nature* **2004**, *431*, 931 – 945.

[8] F. S. Collins, M. Morgan, A. Patrinos, *Science* **2003**, *300*, 286 – 290.

[9] R. F. Service, *Science* **2006**, *311*, 1544 – 1546.

[10] P. K. Gupta, *Trends Biotechnol.* **2008**, *26*, 602 – 611.

[11] R. E. Green, J. Krause, A. W. Briggs, T. Maricic, U. Stenzel, M. Kircher, N. Patterson, H. Li, W. Zhai, M. H.-Y. Fritz, N. F. Hansen, E. Y. Durand, A.-S. Malaspinas, J. D. Jensen, T. Marques-Bonet, C. Alkan, K. Prufer, M. Meyer, H. A. Burbano, J. M. Good, R. Schultz, A. Aximu-Petri, A. Butthof, B. Hober, B. Hoffner, M. Siegemund, A. Weihmann, C. Nusbaum, E. S. Lander, C. Russ, N. Novod, J. Affourtit, M. Egholm, C. Verna, P. Rudan, D. Brajkovic, Z. Kucan, I. Gusic, V. B. Doronichev, L. V. Golovanova, C. Lalueza-Fox, M. de La Rasilla, J. Fortea, A. Rosas, R. W. Schmitz, P. L. F. Johnson, E. E. Eichler, D. Falush, E. Birney, J. C. Mullikin, M. Slatkin, R. Nielsen, J. Kelso, M. Lachmann, D. Reich, S. Pääbo, *Science* **2010**, *328*, 710 – 722.

[12] M. Margulies, M. Egholm, W. E. Altman, S. Attiya, J. S. Bader, L. A. Bemben, J. Berka, M. S. Braverman, Y.-J. Chen, Z. Chen, S. B. Dewell, L. Du, J. M. Fierro, X. V. Gomes, B. C. Godwin, W. He, S. Helgesen, C. H. Ho, G. P. Irzyk, S. C. Jando, M. L. I. Alenquer, T. P. Jarvie, K. B. Jirage, J.-B. Kim, J. R. Knight, J. R. Lanza, J. H. Leamon, S. M. Lefkowitz, M. Lei, J. Li, K. L. Lohman, H. Lu, V. B. Makhijani, K. E. McDade, M. P. McKenna, E. W. Myers, E. Nickerson, J. R. Nobile, R. Plant, B. P. Puc, M. T. Ronan, G. T. Roth, G. J. Sarkis, J. F. Simons, J. W. Simpson, M. Srinivasan, K. R. Tartaro, A. Tomasz, K. A. Vogt, G. A. Volkmer, S. H. Wang, Y. Wang, M. P. Weiner, P. Yu, R. F. Begley, J. M. Rothberg, *Nature* **2005**, *437*, 376 – 380.

[13] J. Shendure, G. J. Porreca, N. B. Reppas, X. Lin, J. P. McCutcheon, A. M. Rosenbaum, M. D. Wang, K. Zhang, R. D. Mitra, G. M. Church, *Science* **2005**, *309*, 1728 – 1732.

[14] J. M. Rothberg, J. H. Leamon, *Nat. Biotechnol.* **2008**, *26*, 1117 – 1124.

[15] J. D. Wall, S. K. Kim, *PLoS Genet.* **2007**, *3*, e175.