Revista de Bioarqueología **"ARCHAEOBIOS"**
ISSN 1996-5214, Setiembre 2008

ARQUEOBIOS
CENTRO DE INVESTIGACIONES ARQUEOBIOLÓGICAS Y PALEOECOLÓGICAS ANDINAS
http://www.arqueobios.org

ARTÍCULOS DE REVISIÓN

SEQUENCING ANCIENT AND MODERN GENOMES

Gabriel Dorado[1], Víctor Vásquez[2], Isabel Rey[3], Fernando Luque[4], Inmaculada Jiménez[5], Arturo Morales[6], Manuel Gálvez[7], Jesús Sáiz[8], Adela Sánchez[8], Pilar Hernández[9]

[1]Author for correspondence, Dep. Bioquímica y Biología Molecular, Campus Rabanales C6-1-E17, Universidad de Córdoba, 14071 Córdoba (Spain), eMail: <bb1dopeg@uco.es>; [2]Centro de Investigaciones Arqueobiológicas y Paleoecológicas Andinas, ARQUEOBIOS, Apartado Postal 595, Trujillo (Peru); [3]Colección de Tejidos y ADN, Museo Natural de Ciencias Naturales, 28006 Madrid; [4]Servicio Sanidad Exterior, Dependencia de Sanidad, Subdelegación del Gobierno en Huelva, C/. Sanlúcar de Barrameda 7, 21001 Huelva; [5]I.E.S. Abyla, Polígono Virgen de África, s/n, 51001 Ceuta; [6]Dep Biología, Facultad de Ciencias, Universidad Autónoma de Madrid, 28049 Cantoblanco (Madrid); [7]Dep. Radiología y Medicina Física, Unidad de Física Médica, Facultad de Medicina, Avda. Menéndez Pidal s/n, Universidad de Córdoba, 14071 Córdoba; [8]Dep. Farmacología, Toxicología y Medicina Legal y Forense, Facultad de Medicina, Avda. Menéndez Pidal, s/n, Universidad de Córdoba, 14071 Córdoba; [9]Instituto de Agricultura Sostenible (IAS), Consejo Superior de Investigaciones Científicas (CSIC), Alameda del Obispo s/n, 14080 Córdoba

The first generation of DNA sequencing methodologies started in 1975 with the "plus and minus" method of Sanger and Coulson, which required cloning of each read start, for production of single-stranded DNA. In 1977, Maxam and Gilbert publish the "DNA sequencing by chemical degradation" methodology. It was based on the chemical modification and subsequent cleavage of DNA, and became the sequencing method of choice, since it allowed to use purified DNA directly, without cloning. The same year, Sanger published "DNA sequencing by enzymatic synthesis"; a new method which set a new mark for the following 30 years (Sanger et al, 1977, 1992; Wikipedia, 2008a).

The Sanger method allowed reading 25 bases (b) and later on 80 b, using dideoxy terminators. The method was further optimized using fluorescent dideoxynucleotide dyes instead of toxic compounds and radioisotopes, automated detection, increased throughput and accuracy, allowing reads of 1,000 b, to which we have significantly contributed (Lario et al, 1997). These breakthroughs represented a fascinating revolution, allowing to decipher genes initially, and eventually even full genomes (Schuster, 2008), albeit at a very high cost for the latter. Currently, the Sanger method is still used in most laboratories around the world. Yet, such technology is not particularly suitable for archaeological samples, hindered by small and degraded DNA. That is why the first generation of DNA sequencing was not successful enough to sequence ancient DNA (aDNA) genomes. Yet, this scenario has changed with the development of the second generation sequencing platforms.

The second generation (also called "next-generation) platforms of DNA sequencing arrived three years ago, increasing the effectiveness of DNA sequencing by several orders of magnitude (Bonetta, 2006); thus, generating reading of gigabases (Gb) in a single experiment. Four second-generation platforms have been commercialized so far:

1) The 454 Instrument (454 Life Sciences), based on emulsion, sequencing-by-synthesis (SBS) and pyrosequencing. It was released in 2005 (Margulies, 2005), purchased by Roche Diagnostics in 2007 and sold as the "Genome Sequencer 20 System" and the "Genome Sequencer FLX System" (Roche Applied Sciences) <https://www.roche-applied-science.com/sis/sequencing/index.jsp>. The 454

technology started reading 100 b, then 250 b after 16 months, and now more than 400 b (Schuster, 2008).

2) The multiplex polony sequencing protocol (Shendure, 2005) is similar to the previous one, yet much cheaper, using off-the-shelf instrumentation and reagents. Shotgun genomic libraries are amplified on microbeads by emulsion PCR. Then they are used as templates for sequencing by fluorescent nonamer ligation reactions on a microscope slide, generating millions of 26-bp reads (Porreca et al, 2008), so any laboratory can deploy it.

3) The Genome Analyzer System (Solexa) combined SBS chemistry with terminators and cluster technology. The company was acquired by Illumina in 2007, delivering the "Genome Analyzer Sequencing System" <http://www.illumina.com/pages.ilmn?ID=204>. Such technology generates tenfold more reads than the 454 one, but with only 35 b or less in length.

4) The SOLiD System (Applied Biosystems) uses a ligation-based chemistry (Chi, 2008) and was released in 2007 <http://solid.appliedbiosystems.com>.

The second generation DNA sequencing platforms differ from traditional sequencing methods in two ways. First, rather than sequencing a few individual DNA clones (eg., 96 sequencing templates on a contemporary Sanger capillary sequencer), hundreds of thousands (454 system), thousands to millions (polony protocol), or even tens of millions (Solexa and SOLiD) of DNA molecules are sequenced in parallel, using much smaller reaction volumes (Schuster, 2008). Second, the sequences obtained are much shorter (25-50 nucleotides for the polony, Solexa and SOLiD technologies, and 200-400 nucleotides for the 454 system) than those generated by traditional sequencing (Graveley, 2008), although the cost of such instruments is much higher (about $500,000) than those using the Sanger chemistry ($10,000 to $100,000), which can be also carried out on much cheaper manual instrumentation with radioisotopes or fluorescent dyes.

As a practical example of these advancements, the first sequencing of the first human (Homo sapiens) genome required hundreds of machines operating 24 hours a day for 13 years, at a cost of over $300 million. The project began in 1990, releasing a working draft of the genome in 2000, and a more complete one in 2003, with further analysis still being published (HGP, 2008; HUGO, 2008; Wikipedia, 2008b). The task was in fact compared in time and cost to the Apollo project, that placed a man on the Moon. Later on, the diploid genome from both chromosomes of a single person (J. Craig Venter) was read by whole-genome shotgun sequencing, requiring 10 years and $70 million, using the optimized Sanger technology (Levy et al, 2007). Instead, the Watson genome was sequenced in just two months for $1 million using the 454 Life Sciences machine (Chi, 2008; Wheeler et al, 2008). Another inspiring example is the sequencing of the platypus (*Ornithorhynchus anatinus*) genome, revealing unique signatures of evolution, with genes that appear in reptiles, or birds and other mammals. This fascinating mixture of features in the genome of the ornithorhynchus provides many clues about the role and evolution of the genomes of mammals (Warren et al, 2008).

Since the Sanger approach is prohibitively expensive for nuclear DNA (nuDNA) projects, many laboratories now rely solely on the second generation sequencing data, combining the advantages of relatively long reads of the 454 system with the low operating costs of Solexa or SOLiD systems, or implementing the polony protocol.

The third generation (also called "next-next-generation") of DNA sequencing has just arrived this year, with revolutionary single-molecule chemistries:

1) The HeliScope Single Molecule Sequencer from Helicos BioSciences <http://www.helicosbio.com> has been released this year. It allows accurate reads of 25 to 45 bases for thousands of millions (billions) of strands on a single run now (producing over 2 Gb

of sequence data per day), and up to one billion bases per hour in the future <http://www.helicosbio.com/Portals/0/Videos/tSMS-How_It_Works.flv>. That is because it uses "true Single Molecule Sequencing" (tSMS), reading single DNA strands (Blow, 2008; Harris, 2008).

2) VisiGen Biotechnologies <http://visigenbio.com> has not been released yet, promising massively parallel arrays (microarrays) of nanomachines, with a sequencing rate of one Mb/sec/machine (over 86 Gb of sequence data per day) <http://visigenbio.com/flash/stream/visigen_movie_6mb.swf>, reading also single molecules.

These advancements will reduce the current sequencing price from one to two orders of magnitude, thus allowing the development of "personal genomics": to sequence the entire human genome of any person in less than one day, for $1,000 or less (Mardis, 2006; Milos, 2008; VonBubnoff, 2008; Schuster, 2008). The reduction of the sequencing error rate, as well a the improvement of current computer microprocessors and software bioinformatics tools can be exploited to prevent the analytical bottleneck that such huge amount of data could create (Díaz et al, 2008a,b; Schuster, 2008). Thus, it is expected that by the end of next year, there will be complete genome sequences of at least draft quality for more than 1,000 bacteria and archaea and 100 eukaryotes (Liolios et al, 2008), and for even larger numbers of organelles, plasmids viruses, viroids and virusoids.

Not surprisingly, all these technologies are having a significant impact on archaeology and related sciences, allowing to sequence more than 25,000 base pairs (bp) of nuDNA of a Pleistocene cave bear (*Ursus spelaeus*) bone dated 40,000 years old (Noonan et al, 2005), 13 million bp (Mbp) of nuDNA from a 28,000 year-old mammoth (*Mammuthus primigenius*) bone dated 28,000 years old (Poinar et al, 2006) and 1 Mbp of nuDNA from Neandertal (*Homo neanderthalensis*) bones dated 38,000 years old (Green et al, 2006; Noonan et al, 2006). Nevertheless, these feats are not easily accomplished. First, the amount of DNA typically obtained from such old samples like the Neanderthal is less than 5%, when compared to modern samples. In other words, the aDNA genome projects require 20 times more sequencing than a modern DNA genome project. Second, the aDNA is usually broken and chemically altered (damaged). Last but not least, the aDNA is often contaminated with modern DNA, which is particularly relevant for human DNA. Overcoming these serious obstacles on aDNA genomic projects may require multiple-fold coverage of a given region and resequencing, as well as using the new emulsion PCR (emPCR) and sequencing platforms (Blow et al, 2008). Yet, the reward is exciting, allowing sequencing the genomes and thus analyzing at the molecular level the flora and fauna of at least the last 100,000 years (Schuster, 2008).

Additionally, besides the de novo assemblies of genomes, it is now economically sound to resequence genomes from different samples, organisms or individuals, using a previously available reference sequence to guide the assembly. This approach requires much less coverage (8 to 12 fold) than assembling genomes de novo (25 to 70 fold), and has been used to sequence ancient mitochondrial genomes from human hairs (Gilbert, 2007, 2008), thus enabling population studies (Schuster, 2008). Mitochondrial DNA (mtDNA) has several advantages when compared to nuDNA for aDNA studies: there are thousands of mitochondrial genomes per cell, it shows maternal inheritance, and has an accelerated mutation rate.

On the other hand, the third generation sequencing methods are so powerful that they allow to carry out studies not only on structural genomics, but also on functional genomics and sequence census (Wold and Myers, 2008), including: i) ChIP-Seq, which is based on chromatin immunoprecipitation (ChIP), to map the in vivo DNA sequences occupied by a DNA-binding protein; ii) mRNA-Seq to study gene expression (Graveley, 2008); and iii) Methyl-Seq to analyze methylation patterns. These procedures can be also applied to ancient DNA, as long as suitable DNA, DNA-protein or mRNA can be isolated from such samples.

In summary, all this demonstrates that we are not really in the post-genomics era as some may have thought, but rather starting the genomics era. Not only because of these brave new technologies, but also because only a tiny number of genomes have been sequenced so far, when compared to the enormous amount of biological entities that harbors the planet Earth: from virusoids, viroids and viruses to prokaryotes and eukaryotes. So, expect great news on this genomics era of nucleic acid sequencing, because the race has started, again, and is here to stay for a long, long time (Dorado, 2008).

Acknowledgements:

Supported by grants AGL2006-12550-C02-01 & AGL2006-12550-C02-02 of "Ministerio de Educación y Ciencia", Project 041/C/2007 of "Consejería de Agricultura y Pesca, Junta de Andalucía", and Grupo PAI AGR-248 of "Junta de Andalucía" (Spain).

References

Blow N (2008): DNA sequencing: generation next-next. Nature Methods 5: 267-272.

Bonetta L (2006): Genome sequencing in the fast lane. Nature Methods 3: 141-146.

Chi KR (2008): The year of sequencing. Nature Methods 5: 11-14.

Díaz D, Claros MG, Falgueras J, Hernández P, Caballero JA, Dorado G, Gálvez S (2008a): DemAlign/Omega-Jalview: an algorithm/viewer for fast discovering differences in similar sequences. Accelrys Science & Technology Forum (Paris, France; <http://accelrys.com/events/seminars/science-and-technology-forums>)

Díaz D, Falgueras J, Claros MG, Guerrero DD, Hernández P, Caballero JA, Dorado G, Gálvez S (2008b): Integrating bioinformatics workflow tools: Omega-Brigid and Scitegic Pipeline Pilot. Accelrys Science & Technology Forum (Paris, France; <http://accelrys.com/events/seminars/science-and-technology-forums>)

Dorado G (ed) (2008): "Molecular Markers, PCR, Bioinformatics and Ancient DNA - Technology and Applications". Science Publishers (New York, NY, USA). In press.

Gilbert MT, Kivisild T, Grønnow B, Andersen PK, Metspalu E, Reidla M, Tamm E, Axelsson E, Götherström A, Campos PF, Rasmussen M, Metspalu M, Higham TF, Schwenninger JL, Nathan R, De Hoog CJ, Koch A, Møller LN, Andreasen C, Meldgaard M, Villems R, Bendixen C, Willerslev E (2008): Paleo-Eskimo mtDNA genome reveals matrilineal discontinuity in Greenland. Science 320: 1787-1789.

Gilbert MT, Tomsho LP, Rendulic S, Packard M, Drautz DI, Sher A, Tikhonov A, Dalén L, Kuznetsova T, Kosintsev P, Campos PF, Higham T, Collins MJ, Wilson AS, Shidlovskiy F, Buigues B, Ericson PG, Germonpré M, Götherström A, Iacumin P, Nikolaev V, Nowak-Kemp M, Willerslev E, Knight JR, Irzyk GP, Perbost CS, Fredrikson KM, Harkins TT, Sheridan S, Miller W, Schuster SC (2007): Whole-genome shotgun sequencing of mitochondria from ancient hair shafts. Science 317: 1927-1930.

Graveley BR (2008): Power sequencing. Nature 453: 1197-1198.

Green RE, Krause J, Ptak SE, Briggs AW, Ronan MT, Simons JF, Du L, Egholm M, Rothberg JM, Paunovic M, Pääbo S (2006): Analysis of one million base pairs of Neanderthal DNA. Nature 444: 275-276.

Harris TD, Buzby PR, Babcock H, Beer E, Bowers J, Braslavsky I, Causey M, Colonell J, Dimeo J, Efcavitch JW, Giladi E, Gill J, Healy J, Jarosz M, Lapen D, Moulton K, Quake SR, Steinmann K, Thayer E, Tyurina A, Ward R, Weiss H, Xie Z (2008): Single-molecule DNA sequencing of a viral genome. Science 320: 106-109.

HGP (2008): Human Genome Project. Web <http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml> and <http://www.ornl.gov/sci/techresources/Human_Genome/project/progress.shtml>.

HUGO (2008): Human Genome Organisation. Web <http://www.hugo-international.org>.

Lario A, González A, Dorado G (1997): Automated laser-induced fluorescence DNA sequencing: equalizing signal-to-noise ratios significantly enhances overall performance.

Analytical Biochemistry 247: 30-33.

Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF, Denisov G, Lin Y, MacDonald JR, Pang AW, Shago M, Stockwell TB, Tsiamouri A, Bafna V, Bansal V, Kravitz SA, Busam DA, Beeson KY, McIntosh TC, Remington KA, Abril JF, Gill J, Borman J, Rogers YH, Frazier ME, Scherer SW, Strausberg RL, Venter JC (2007): The diploid genome sequence of an individual human. PLoS Biol 5: e254.

Liolios K, Mavromatis K, Tavernarakis N, Kyrpides NC (2008): The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res 36(Database issue): D475-D479.

Mardis ER (2006): Anticipating the 1,000 dollar genome. Genome Biol 7: 112.1-112.5.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM (2005): Genome sequencing in microfabricated high-density picolitre reactors. Nature 437: 376-380.

Milos P (2008): Helicos BioSciences. Pharmacogenomics 9: 477-480.

Noonan JP, Coop G, Kudaravalli S, Smith D, Krause J, Alessi J, Chen F, Platt D, Pääbo S, Pritchard JK, Rubin EM (2006): Sequencing and analysis of Neanderthal genomic DNA. Science 314: 1113-1118.

Noonan JP, Hofreiter M, Smith D, Priest JR, Rohland N, Rabeder G, Krause J, Detter JC, Pääbo S, Rubin EM (2005): Genomic sequencing of Pleistocene cave bears. Science 309: 597-599.

Poinar HN, Schwarz C, Qi J, Shapiro B, Macphee RD, Buigues B, Tikhonov A, Huson DH, Tomsho LP, Auch A, Rampp M, Miller W, Schuster SC (2006): Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA. Science 311: 392-394.

Porreca GJ, Shendure J, Church GM (2006): Polony DNA sequencing. In: Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA, Struhl K (eds): "Current Protocols in Molecular Biology". Vols 1-4, Chapter 7: Unit 7.8. Greene & John Wiley (New York).

Sanger F, Nicklen S, Coulson AR (1977): DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci USA 74: 5463-5467.

Sanger F, Nicklen S, Coulson AR (1992): DNA sequencing with chain-terminating inhibitors. Biotechnology 24: 104-108.

Schuster SC (2008): Next-generation sequencing transforms today's biology. Nature Methods 5: 16-18.

Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, Rosenbaum AM, Wang MD, Zhang K, Mitra RD, Church GM (2005): Accurate multiplex polony sequencing of an evolved bacterial genome. Science 309: 1728-1732.

VonBubnoff A (2008): Next-generation sequencing: the race is on. Cell 132 :721-723.

Warren WC, Hillier LW, Marshall Graves JA, Birney E, Ponting CP, Grützner F, Belov K, Miller W, Clarke L, Chinwalla AT, Yang SP, Heger A, Locke DP, Miethke P, Waters PD, Veyrunes F, Fulton L, Fulton B, Graves T, Wallis J, Puente XS, López-Otín C, Ordóñez GR, Eichler EE, Chen L, Cheng Z, Deakin JE, Alsop A, Thompson K, Kirby P, Papenfuss AT, Wakefield MJ, Olender T, Lancet D, Huttley GA, Smit AF, Pask A, Temple-Smith P, Batzer MA, Walker JA, Konkel MK, Harris RS, Whittington CM, Wong ES, Gemmell NJ, Buschiazzo E, Vargas Jentzsch IM, Merkel A, Schmitz J, Zemann A, Churakov G, Kriegs JO, Brosius J, Murchison EP, Sachidanandam R, Smith C, Hannon GJ, Tsend-Ayush E, McMillan D, Attenborough R, Rens W, Ferguson-Smith M, Lefèvre CM, Sharp JA, Nicholas KR, Ray DA, Kube M, Reinhardt R, Pringle TH, Taylor J, Jones RC, Nixon B, Dacheux JL, Niwa H, Sekita Y, Huang X, Stark A, Kheradpour P, Kellis M, Flicek P, Chen Y, Webber C, Hardison

R, Nelson J, Hallsworth-Pepin K, Delehaunty K, Markovic C, Minx P, Feng Y, Kremitzki C, Mitreva M, Glasscock J, Wylie T, Wohldmann P, Thiru P, Nhan MN, Pohl CS, Smith SM, Hou S, Renfree MB, Mardis ER, Wilson RK (2008): Genome analysis of the platypus reveals unique signatures of evolution. Nature 453: 175-183.

Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT, Gomes X, Tartaro K, Niazi F, Turcotte CL, Irzyk GP, Lupski JR, Chinault C, Song XZ, Liu Y, Yuan Y, Nazareth L, Qin X, Muzny DM, Margulies M, Weinstock GM, Gibbs RA, Rothberg JM (2008): The complete genome of an individual by massively parallel DNA sequencing. Nature 452: 872-876.

Wikipedia (2008a): DNA sequencing. Web <http://en.wikipedia.org/wiki/DNA_sequencing>.

Wikipedia (2008b): Human Genome Project. Web <http://en.wikipedia.org/wiki/Human_Genome_Project>.

Wold B, Myers RM (2008): Sequence census methods for functional genomics. Nature Methods 5: 19-21.