Massively parallel pyrosequencing in HIV research

Frederic D. Bushman[a], Christian Hoffmann[a], Keshet Ronen[a], Nirav Malani[a], Nana Minkah[a], Heather Marshall Rose[a], Pablo Tebas[b] and Gary P. Wang[a]

AIDS 2008, **22**:1411–1415

Keywords: DNA sequence, homogeneous PCR products, pyrosequencing

The new massively parallel sequencing methods are so astonishing that one wonders whether space aliens are secretly behind them. One technician, running a single instrument, can obtain up to approximately 1 billion bases of DNA sequence in a few days. Here we describe the new sequencing methods, briefly present a few applications in HIV research, and then speculate on future directions.

The new massively parallel sequencing methods

Several methods for massively parallel pyrosequencing have recently been commercialized. As an example, consider the use of the 454 Life Sciences pyrosequencing method for metagenomic analysis of woolly mammoth DNA [1,2]. DNA from a mammoth carcass was purified and fragmented, and DNA linkers were ligated to the free ends. DNAs were then denatured and strands annealed to beads conjugated with oligonucleotides complementary to the linker sequences. This step is carried out with very low DNA concentrations so that on average only one strand binds to each bead. Bead-bound DNA is then PCR amplified in an oil–water emulsion, where each water droplet in the emulsion contains on average a single bead. The amplified DNA strands anneal to the beads, yielding beads with many copies of homogeneous PCR products. Pools of up to 400 000 beads are then distributed in a picotiter plate and further manipulations carried out in a custom fluidics station (Fig. 1). A polymerase is used to extend a DNA chain from a bound primer on each strand. The four nucleoside triphosphates are sequentially flowed over the picotiter plate. With each incorporation event, pyrophosphate is liberated into solution (hence 'pyrosequencing'). An enzyme system is present in the aqueous phase that directs incorporation of pyrophosphate into ATP, which in turn activates purified luciferase, also present in the aqueous phase, to produce a flash of light. Each flash from each well is quantified by a charge coupled device camera and the signals detected and stored in a computer. Sequential application of the four nucleotides (nts) allows DNA sequences of approximately 100 nt to be built up several hundred thousand fold at a time. Using this method a detailed comparison of the mammoth and elephant genomes can be carried out (yes, there is also an elephant genome project).

With the improved 454 technology released recently, it is possible to generate reads of approximately 260 nt on approximately 400 000 beads per run, yielding a whopping 100 million bases of DNA sequence in a day or two. An illustrated description of the method can be found at http://www.454.com/enabling-technology/index.asp.

[a]University of Pennsylvania School of Medicine, Department of Microbiology, and [b]University of Pennsylvania School of Medicine, Department of Medicine, Division of Infectious Diseases, Philadelphia, Pennsylvania, USA.

Correspondence to Frederic D. Bushman, University of Pennsylvania School of Medicine, Department of Microbiology, 3610 Hamilton Walk, Philadelphia, PA 19104-6076, USA.

E-mail: Bushman@mail.med.upenn.edu

Received: 8 January 2008; revised: 31 January 2008; accepted: 11 February 2008.

Fig. 1. The 454 GS FLX sequencing station.

A second pyrosequencing technology, commercialized by Solexa/Illumina (San Diego, California, USA), yields shorter sequence reads, only approximately 35 bp, but a single run yields up approximately 1 billion bases of DNA sequence [3,4] (see www.illumina.com/downloads/SS_DNAsequencing.pdf). Table 1 compares the Sanger, 454/Roche (Branford, Connecticut, USA), and Solexa/Illumina methods. A variety of additional technologies are also under development [5].

Pyrosequencing to analyze HIV diversity

The pyrosequencing methods are well suited to addressing questions on the dynamics of HIV quasi-species in response to selective pressures. HIV reverse transcriptase is very error prone, making roughly one base pair substitution mutation per round of replication [6]. The viral populations in infected individuals are also very large, with some 10^{10} virions produced and destroyed per day [7–9]. After infection, this activity quickly results in the formation of a diverse pool, or quasispecies, in which most viral sequences differ from all others.

Table 1. Comparison of sequencing methods.

Sequencing method[a]	Read length (bp)	Reads/run[b]	Cents/base[c]
Sanger	850	384	0.4
454/Roche	260	400 000	0.015
Solexa/Illumina	35	30 000 000	0.0006
ABI/SOLiD	35 (or 2X25)	20 000 000	Unknown

[a]'Sanger' indicates the standard dideoxy chain terminator method. '454/Roche' and 'Solexa/Illumina' indicate pyrosequencing methods implemented by the indicated companies. 'ABI/SOLiD' indicates a method based on hybridization of short labeled oligonucleotides implemented by Applied Biosystems (Foster City, California, USA).
[b]Each 'run' will take about a day for the Sanger and 454/Roche methods and several days for Solexa/Illumina method. Run times for the ABI/SOLiD method are uncertain because the technology is the newest of the four.
[c]All cost estimates are very approximate because they are influenced by discounts associated with throughput and scale, and whether factors such as equipment depreciation are included in the calculation – thus, costs will vary widely among sequencing centers.

Pyrosequencing offers an improved means of characterizing sequence variation present in such large populations. One application is quantification of rare drug-resistant mutations in treated individuals failing antiretroviral therapy. Current guidelines recommend that whenever an HIV-positive individual is initiating or changing therapy, possible drug-resistant alleles should be assayed and treatment choices adjusted accordingly. However, the most commonly used genotyping methods only provide information on the most abundant sequence variants. Evidence suggests that rare drug-resistant variants, when subjected to the selective pressures of drug treatment, can quickly grow out and become the predominant form, leading to treatment failure [10,11].

Two reports have applied pyrosequencing to the detection and characterization of rare drug-resistant variants [12,13]. Shafer and colleagues [13] purified RNA from patient virions, reverse transcribed to generate cDNA, sheared the cDNA product, and carried out pyrosequencing as described for mammoth DNA. Hoffmann *et al.* [12] took a different approach, PCR amplifying short regions of interest, then pyrosequencing the amplicons directly. Both groups also analyzed control cloned viral stocks, allowing empirical determination of the rates of false-positive calls and thereby permitting convincing demonstrations of detection of drug-resistant mutations present as less than 1% of the population.

Another study [14] applied pyrosequencing to analyzing viral tropism in quasispecies inferred from V3 loop sequences. As inhibitors that block coreceptor binding come into widespread use, there will be an increased need for deep profiling of coreceptor usage prior to initiating therapy, and pyrosequencing would be an ideal tool for this application.

The Hoffmann approach [12] applied a twist that is likely to be useful in many applications. They indexed PCR products by adding short DNA bar codes to their PCR primers. Sequence reads extended across the bar codes, allowing different samples to be distinguished after sequencing in a mixed pool. Running a single 454 plate is fairly expensive, but bar coding allows many samples to be run on the same plate, thereby reducing the cost per sample [12,15].

Pyrosequencing to analyze HIV integration targeting

Detailed studies of the placement of HIV integration sites in the human genome [16–19] have yielded a wealth of new insights into HIV biology. Related studies [20,21] on integration site distributions are also important in the gene therapy field, in which integration of therapeutic retroviral vectors has resulted in insertional activation of proto-oncogenes and clinical adverse events. The first genome-wide study of HIV integration targeting [22] used Sanger sequencing and reported approximately 500 integration sites [16]. A recent study using pyrosequencing yielded 40 000 sites. Each boost in the number of integration sites has led to new insights into mechanism. The pyrosequencing study of HIV integration, for example, yielded data indicating that integration in chromosomes *in vivo* usually takes place on nucleosome-bound DNA (Fig. 2)[23–25].

Pyrosequencing data sets are just beginning to be generated for samples from gene therapy trials, including trials to treat HIV [26], which will allow much deeper investigation of vector integration site distributions. One question for future studies will be to what degree patterns in deep integration site data help forecast impending clinical adverse events.

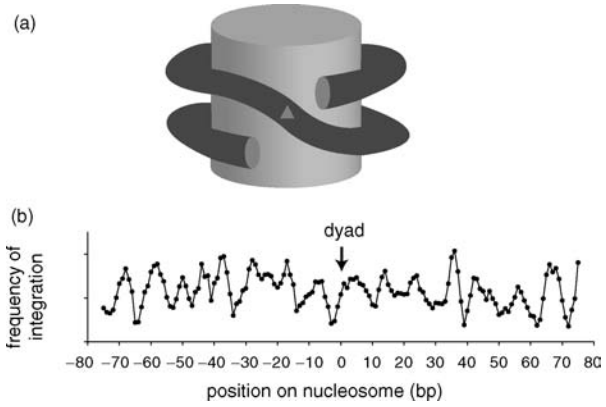

Fig. 2. Use of pyrosequencing data to show that HIV integration takes place on nucleosomal DNA *in vivo*. (a) Schematic diagram of DNA wrapped on a nucleosome. The dyad axis (center of two-fold symmetry) is shown by the diamond. (b) Data from pyrosequencing analysis of 40 000 sites of HIV DNA integration, showing frequency as a function of position on the nucleosome. For each of the integration sites, the position of the underlying nucleosome was predicted from the primary sequence using the method of Segal *et al.* [23]. Nucleosomes were aligned at their centers of symmetry and the integration frequency quantified across the full data set moving outward from the dyad axis. A periodic pattern of high and low frequency integration was found with a period of about 10.5 bases. This pattern is exactly as expected for DNA bound on the nucleosome surface, where integration is favored in sequential out-ward facing major grooves [24,25].

Pyrosequencing to characterize unknown opportunistic infections

Clinicians are occasionally confronted with apparent infections that cannot be easily attributed to a known pathogen. Given concerns about bioterrorism, there is an urgent need to identify the infectious agents quickly in such cases. For AIDS patients, opportunistic infections that would be cleared in immunocompetent individuals can cause morbidity in the immunocompromised, potentially resulting in challenges in identifying the agent. A wealth of new molecular methods are becoming available for identifying new pathogens, with pyrosequencing a major addition.

The discovery of a virus potentially responsible for colony collapse disorder (CCD) of honey bees provides a striking example [27]. North American bee colonies have been failing at alarming rates, and an infectious agent implicated. Pyrosequencing of a large number of samples from affected and unaffected colonies revealed that the presence of Israeli acute paralysis virus was strongly associated with CCD, and follow-up quantitative PCR studies strengthened the connection. Experimental infection studies are now needed to test causality.

Pyrosequencing also provides potent new methods for analyzing bacterial populations. Many bacterial diseases likely involve not only single pathogens but also the full microbial community in which that pathogen resides – for example, obesity and Crohn's disease are proposed to involve community-wide alterations in gastrointestinal microbiota [28–30]. Pyrosequencing of DNA samples from uncultured bacterial communities can identify the members of a community and their relative abundance in a rapid and cost-effective fashion [31,32]. For example, 454 pyrosequencing of segments of the 16S RNA genes from uncultured communities of marine bacteria has revealed numerous previously undiscovered bacterial taxa [31,33].

Of particular interest to HIV research, a pyrosequencing study [34] of the gastrointestinal microbiota present in simian immunodeficiency virus (SIV)-infected macaques has shown consistent changes associated with chronic colitis accompanying disease progression. These findings set the stage for an investigation of pathogenic mechanisms, particularly in connection with recent proposals for involvement of gastrointestinal flora in inflammation and lentiviral disease progression [35].

Pyrosequencing in hypothesis testing

Pyrosequencing can be used not just to sequence genomes, but also as an end-point assay in a mechanistic experiment, as one might once have used a p24 assay or

Southern blot. For example, in an application in the integration targeting field, HIV integration in cells containing or lacking the targeting factor PC4 and SFRS1 interacting protein 1/lens epithelium-derived growth factor/p75 (PSIP1/LEDGF/p75) was compared using pyrosequencing. The analysis revealed that cells lacking the cofactor showed reduced integration in transcription units ([18]; see also [17,19]).

In another example, Shuman and colleagues [36] identified mitoxantrone as an antipoxviral agent and sought to identify the viral target of the inhibitor. They picked mutant vaccinia viruses insensitive to the drug and then they sequenced the entire 195 kb mutant genomes from mutants and compared them to the wild type. A mutation in the ligase gene was found to be selectively present in the genomes of the drug-resistant variants. Thus, pyrosequencing allowed efficient identification of the molecular target of a new antiviral agent in a large viral genome.

Looking ahead

In just approximately 2 years since the 454 sequencing system has been commercially available, the technology has improved considerably and other companies are introducing competing platforms. It is virtually certain that the already amazing new sequencing technologies will be further improved in the years to come and that these new methods will transform HIV research. Some present and possible future applications of pyrosequencing to HIV research are summarized as follows:

(1) Quantifying rare drug-resistant mutations in HIV quasispecies
(2) Quantifying immune escape mutations in quasispecies under immune pressure
(3) Quantifying coreceptor usage in quasispecies
(4) Genome-wide monitoring of integration site selection
(5) Identification of novel pathogens
(6) Analysis of effects of lentiviral infection on gastrointestinal microbiota
(7) Affordable genotyping for resource-limited settings.

For any virus, when understanding the structure of a complex swarm is important, pyrosequencing is an attractive tool. Interactions of HIV with other viruses could be explored; for example, interactions with hepatitis C virus (HCV). HCV quasispecies are probably more complex than HIV quasispecies and understanding of HCV swarms is less advanced than for HIV. Pyrosequencing of HCV populations from well chosen patient populations should be highly informative and studies of the effects of coinfection with HIV are of considerable interest.

Studies of human genetic determinants of HIV disease course will likely be accelerated by the new methods. A couple of examples of complete sequencing of individual human genomes have been reported and more are on the way. Such studies on AIDS patients with different disease responses should help identify new human determinants of disease transmission and progression.

Lastly, pyrosequencing may make some molecular diagnostic methods affordable in resource-limited settings. Methods for characterizing HIV drug resistance are unavailable in many areas of the developing world. Single pyrosequencing runs are expensive (for the 454 method, approximately US$ 15 000 per plate), but using DNA bar coding, several hundred samples may readily be analyzed per plate and larger numbers should be accessible with further refinement. There would be many challenges to implementation, but ultimately pyrosequencing technology could make a variety of molecular diagnostics affordable for those who presently lack access.

References

1. Poinar HN, Schwarz C, Qi J, Shapiro B, Macphee RD, Buigues B, *et al*. **Metagenomics to paleogenomics: large-scale sequencing of mammoth DNA.** *Science* 2006; **311**:392–394.
2. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, *et al*. **Genome sequencing in microfabricated high-density picolitre reactors.** *Nature* 2005; **437**:376–380.
3. Bentley DR. **Whole-genome re-sequencing.** *Curr Opin Genet Dev* 2006; **16**:545–552.
4. Bennett S. **Solexa Ltd.** *Pharmacogenomics* 2004; **5**:433–438.
5. Hall N. **Advanced sequencing technologies and their wider impact in microbiology.** *J Exp Biol* 2007; **210**:1518–1525.
6. Preston BD. **Reverse transcriptase fidelity and HIV-1 variation.** *Science* 1997; **275**:228–229.
7. Coffin JM. **HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy.** *Science* 1995; **267**:483–486.
8. Ho DD, Neumann AU, Perelson AS, Chen W, Leonard JM, Markowitz M. **Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection.** *Nature* 1995; **373**:123–126.
9. Perelson AS, Neumann AU, Markowitz M, Leonard JM, Ho DD. **HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time.** *Science* 1996; **271**:1582–1586.
10. Palmer S, Kearney M, Maldarelli F, Halvas EK, Bixby CJ, Bazmi H, *et al*. **Multiple, linked human immunodeficiency virus type 1 drug resistance mutations in treatment-experienced patients are missed by standard genotype analysis.** *J Clin Microbiol* 2005; **43**:406–413.
11. Jourdain G, Ngo-Giang-Huong N, Le Coeur S, Bowonwatanuwong C, Kantipong P, Leechanachai P, *et al*. **Intrapartum exposure to nevirapine and subsequent maternal responses to nevirapine-based antiretroviral therapy.** *N Engl J Med* 2004; **351**:229–240.
12. Hoffmann C, Minkah N, Leipzig J, Wang G, Arens MQ, Tebas P, *et al*. **DNA bar coding and pyrosequencing to identify rare HIV drug resistance mutations.** *Nucleic Acids Res* 2007; **35**:e91.
13. Wang C, Mitsuya Y, Gharizadeh B, Ronaghi M, Shafer RW. **Characterization of mutation spectra with ultra-deep pyrosequencing: application to HIV-1 drug resistance.** *Genome Res* 2007; **17**:1195–1201.
14. Tsibris A, Russ C, Paredes R, Arnaut R, Honan T, Cahill P, *et al.* **Detection and quantification of minority HIV-1 env V3 loop sequences by ultra-deep sequencing: preliminary results**. International Medical Press, 15th International HIV Drug Resistance Workshop; 2006.

15. Binladen J, Gilbert MT, Bollback JP, Panitz F, Bendixen C, Nielsen R, *et al*. **The use of coded PCR primers enables high-throughput sequencing of multiple homolog amplification products by 454 parallel sequencing.** *PLoS ONE* 2007; **2**:e197.
16. Schroder AR, Shinn P, Chen H, Berry C, Ecker JR, Bushman F. **HIV-1 integration in the human genome favors active genes and local hotspots.** *Cell* 2002; **110**:521–529.
17. Ciuffi A, Llano M, Poeschla E, Hoffmann C, Leipzig J, Shinn P, *et al*. **A role for LEDGF/p75 in targeting HIV DNA integration.** *Nat Med* 2005; **11**:1287–1289.
18. Marshall H, Ronen K, Berry C, Llano M, Sutherland H, Saenz D, *et al*. **Role of PSIP1/LEDGF/p75 in lentiviral infectivity and integration targeting**. *PLoS One*; **2**:e1340.
19. Shun MC, Raghavendra NK, Vandegraaff N, Daigle JE, Hughes S, Kellam P, *et al*. **LEDGF/p75 functions downstream from preintegration complex formation to effect gene-specific HIV-1 integration.** *Genes Dev* 2007; **21**:1767–1778.
20. Hacein-Bey-Abina S, Von Kalle C, Schmidt M, McCormack MP, Wulffraat N, Leboulch P, *et al*. **LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1.** *Science* 2003; **302**:415–419.
21. Hacein-Bey-Abina S, von Kalle C, Schmidt M, Le Deist F, Wulffraat N, McIntyre E, *et al*. **A serious adverse event after successful gene therapy for X-linked severe combined immunodeficiency.** *N Engl J Med* 2003; **348**:255–256.
22. Wang GP, Ciuffi A, Leipzig J, Berry CC, Bushman FD. **HIV integration site selection: analysis by massively parallel pyrosequencing reveals association with epigenetic modifications.** *Genome Res* 2007; **17**:1186–1194.
23. Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, Moore IK, *et al*. **A genomic code for nucleosome positioning.** *Nature* 2006; **442**:772–778.
24. Pruss D, Bushman FD, Wolffe AP. **Human immunodeficiency virus integrase directs integration to sites of severe DNA distortion within the nucleosome core.** *Proc Natl Acad Sci U S A* 1994; **91**:5913–5917.
25. Pryciak PM, Varmus HE. **Nucleosomes, DNA-binding proteins, and DNA sequence modulate retroviral integration target site selection.** *Cell* 1992; **69**:769–780.
26. Levine BL, Humeau LM, Boyer J, MacGregor RR, Rebello T, Lu X, *et al*. **Gene transfer in humans using a conditionally replicating lentiviral vector.** *Proc Natl Acad Sci U S A* 2006; **103**:17372–17377.
27. Cox-Foster DL, Conlan S, Holmes EC, Palacios G, Evans JD, Moran NA, *et al*. **A metagenomic survey of microbes in honey bee colony collapse disorder.** *Science* 2007; **318**:283–287.
28. Manichanh C, Rigottier-Gois L, Bonnaud E, Gloux K, Pelletier E, Frangeul L, *et al*. **Reduced diversity of faecal microbiota in Crohn's disease revealed by a metagenomic approach.** *Gut* 2006; **55**:205–211.
29. Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI. **An obesity-associated gut microbiome with increased capacity for energy harvest.** *Nature* 2006; **444**:1027–1031.
30. Ley RE, Backhed F, Turnbaugh P, Lozupone CA, Knight RD, Gordon JI. **Obesity alters gut microbial ecology.** *Proc Natl Acad Sci U S A* 2005; **102**:11070–11075.
31. Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, *et al*. **Microbial diversity in the deep sea and the underexplored 'rare biosphere'.** *Proc Natl Acad Sci U S A* 2006; **103**:12115–12120.
32. Liu Z, Lozupone C, Hamady M, Bushman FD, Knight R. **Short pyrosequencing reads suffice for accurate microbial community analysis.** *Nucleic Acids Res* 2007; **35**:e120.
33. Huber JA, Welch DB, Morrison HG, Huse SM, Neal PR, Butterfield DA, *et al*. **Microbial population structures in the deep marine biosphere.** *Science* 2007; **318**:97–100.
34. McKenna P, Hoffmann C, Leipzig J, Aye PP, Lackner A, Liu Z, *et al*. **The macaque gut microbiome in health, lentiviral infection and inflammatory bowel disease**. *PloS Pathog* 4, e20.
35. Brenchley JM, Price DA, Schacker TW, Asher TE, Silvestri G, Rao S, *et al*. **Microbial translocation is a cause of systemic immune activation in chronic HIV infection.** *Nat Med* 2006; **12**:1365–1371.
36. Deng L, Dai P, Ciro A, Smee DF, Djaballah H, Shuman S. **Identification of novel antipoxviral agents: mitoxantrone inhibits vaccinia virus replication by blocking virion assembly.** *J Virol* 2007; **81**:13392–13402.