JOHNS HOPKINS
UNIVERSITY

Zanvyl Krieger School of Arts and Sciences
Advanced Academic Programs

410.734.81 and 410.734.82
Practical Introduction to Metagenomics

Topic: Sample collection and sequencing

Instructor: Joshua Orvis

Overview

This figure shows a possible flow for metagenome experiments, and this week's focus is on the first few steps – sampling through sequencing.

We'll consider the wide array of different environments targeted for metagenomic sampling and some of the considerations that need to be made in protocols for many of them.

Next, we'll dive in to the rapidly-changing world of "next-generation" DNA sequencing and learn how they are applied to different types of metagenome experiments.

Note: Several publications will be referenced in this lecture apart from the assigned literature readings, such as the one referenced under the figure on the right. They are optional, and you can find the full reference and/or links in the last slide, "Further reading."

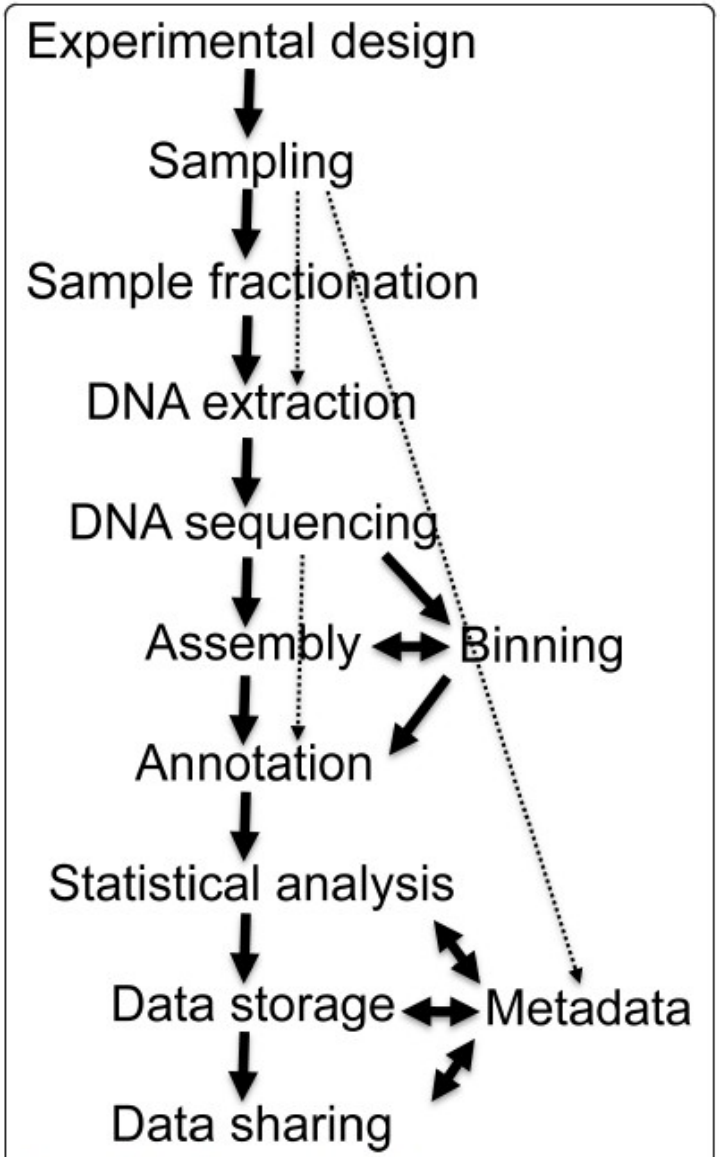


Figure 1 Flow diagram of a typical metagenome projects. Dashed arrows indicate steps that can be omitted.

[1] Thomas, T et al. 2012

DNA sampling / extraction

There are some important general considerations applicable to most microbiome experiments that we'll consider first:

Biofilm from the acid mine drainage site study

- **DNA extracted should be representative of all cells present in the sample.**

  This should be intuitive, but if your goal is to describe the full taxonomic or functional makeup of a sample you need to take steps to ensure you've taken steps to ensure each member of the microbial community is represented.

- **Sufficient amounts of high-quality nucleic acids must be obtained for subsequent library production and sequencing.**

  Different sequencing protocols have very different source DNA requirements. We'll see specifics later in this lecture.

[1] Thomas, T et al. 2012

Fractionation

- If a host is involved fractionation or selective lysis might be enough to ensure minimal contamination.

- Physical fractionation also applies in studies where only part of the community is targetted, (ie, only viruses.)

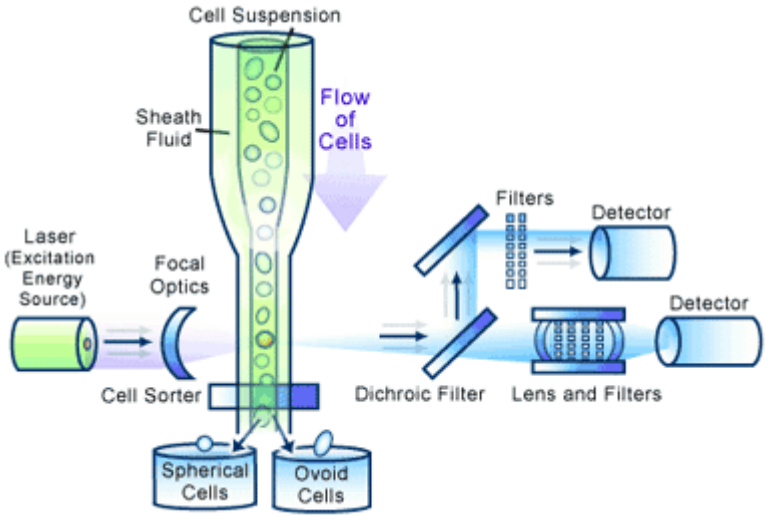
- Physical isolation & indirect lysis has been shown in soil as necessary to improve biases in microbial diversity, DNA yield and fragment lengths (Delmont, et al. 2011)

- There are several fractionation techniques, such as:

Filtration

Centrifugation

Flow cytometry



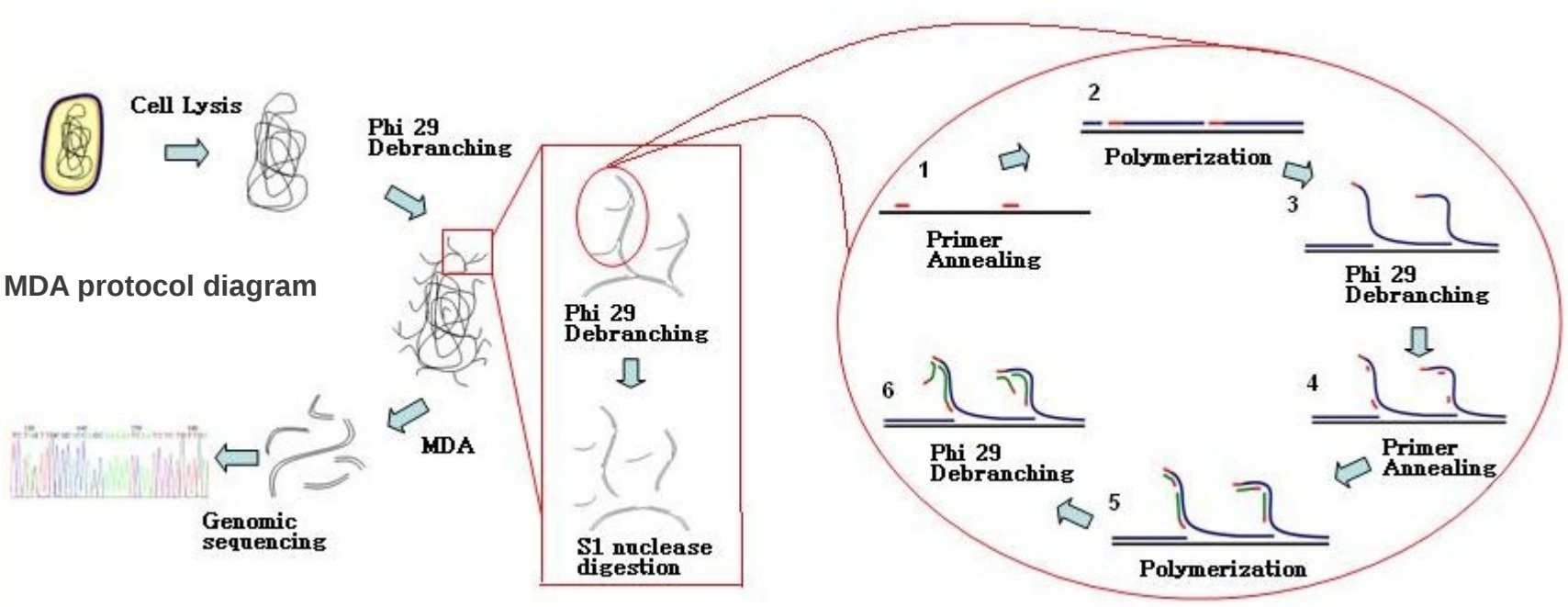
[1] Thomas, T et al. 2012

Amplification / MDA

Sometimes amplification is necessary to achieve the high ng or mg amounts of DNA needed by some sequencing protocols. This is especially true if sample size is limited or in low-biomass habitats.

The multiple displacement amplification (MDA) protocol can amplify a single cell's worth of DNA to generate 1-2 µg in as little as three hours.

Problems with MDA include reagent contamination, chimera formation and sequence bias. These are reviewed in [3], [4].



Time for a pretty break.

Bioluminescent phytoplankton in the Maldives

Delmont study

Metagenomic comparison of direct and indirect soil DNA extraction approaches.
Delmont TO, Robe P, Clark I, Simonet P, Vogel TM.
J Microbiol Methods. 2011 Sep;86(3):397-400.
PMID: 21723887

This paper provides a glimpse at the process of DNA sampling/extraction and illustrates the difference in results based on the protocol used.

Specifically, it contrasts the “direct” and “indirect” methods of extracting DNA from soil samples.

What are the differences between these two methods?

Contents lists available at ScienceDirect
Journal of Microbiological Methods
journal homepage: www.elsevier.com/locate/jmicmeth

Note

Metagenomic comparison of direct and indirect soil DNA extraction approaches

Tom O. Delmont [a], Patrick Robe [b], Ian Clark [c], Pascal Simonet [a], Timothy M. Vogel [a,*]

[a] Environmental Microbial Genomics, Ecole Centrale de Lyon, Université de Lyon, 36 avenue Guy de Collongue, 69134 ECULLY, France
[b] LibraGen, 3 rue des Satellites, 31400 Toulouse, France
[c] Rothamsted Research, Harpenden, Hertfordshire AL5 2JQ, UK

ARTICLE INFO

Article history:
Received 21 March 2011
Received in revised form 15 June 2011
Accepted 18 June 2011
Available online 25 June 2011

Keywords:
Soil metagenomics
DNA extraction
Nycodenz

ABSTRACT

Full pyrosequencing runs of both direct-extracted (high yield, low DNA size) and indirect-extracted DNA (low yield, high DNA size) from the same prairie soil show that the sequence distribution of the majority of the metabolic functions and species detected were statistically similar. Although some microbial functions differed at the 95% confidence interval in bootstrap analyses, the overall functional diversity was the same.

Soil metagenomic approaches require access to high quality DNA in order to construct clone libraries and DNA sequences in sufficient quantity (or representativity) to begin to understand soil microbial ecology (Vogel et al., 2009). Soil DNA extraction is a key step for these metagenomic approaches (Bertrand et al., 2005; Frostegård et al., 1999; Lakay et al., 2007; Delmont et al., 2011) and can be separated in two general strategies. The first strategy, which is more commonly used, is direct DNA extraction and consists of cell lysis directly within a soil sample (e.g. in 1 g of soil) (Ogram et al., 1987; Van Elsas et al., 1997). With the second strategy, indirect DNA extraction, cells are first removed from a soil (e.g. 60 g of soil) and subsequently lysed (Berry et al., 2003; Jacobsen and Rasmussen, 1992). This method separates bacterial and archaeal cells from eukarya cells to some extent by using a density gradient (e.g., Nycodenz density gradient: (Bertrand et al., 2005; Courtois et al., 2001; Lefevre et al., 2008)). Of course, this approach is not the best strategy when eukaryotic sequences are of interest or for studying interactions between eukarya and bacteria or archaea, but can be helpful when eukarya are to be excluded or when high DNA fragments are required (e.g. to construct fosmids and cosmids clones (around 40 kb inserts)). A critical methodological aspect is the DNA yield especially when large quantities are needed for high throughput sequencing or cloning.

Previous studies concluded that in spite of a lower purity, the DNA yield in terms of mass of DNA per mass of soil is greater with direct than indirect extraction (Leff et al., 1995; Steffan et al., 1988) – up to 100-fold greater (Courtois et al., 2001; Roh et al., 2006). A critical question is whether the reduced DNA yield of the indirect extraction strategy results in a significant loss of functional diversity. Due to our current inability to sequence an entire soil metagenome (roughly 10^{15} bp), only the relative genome proportions in the extracted DNA pool can be compared in order to assess the accessibility of the soil microbial genetic richness. The two approaches were compared by analyzing pyrosequencing data from each method, including different sampling strategies due to the important differences in terms of the soil quantity required. Both functional and taxonomical distributions were compared to examine differences in apparent community diversity based on pyrosequencing of DNA resulting from either a direct DNA extraction approach with less soil or an indirect DNA extraction approach with more soil.

Samples were collected from the untreated control plot (3 d) of Park Grass, Rothamsted (England) in March 2009. The Park Grass soil is an internationally-recognized resource and is targeted as a reference for soil metagenomics (Vogel et al., 2009). It is classified as Chromic Luvisol and is a silty clay loam (pH 5.2 measured in H_2O). Soil samples (the top 21 centimeters) were collected during the day with soil cores, fractioned vertically and then homogenized manually by thorough mixing. For both direct and indirect DNA extraction, we used the FastPrep® lysing matrix (MP biomedical). While this approach might not detect all genera present in a soil (Delmont et al., 2011), this protocol is relatively stringent and is thought to lyse the majority of the cells (Howeler et al., 2003; Lakay et al., 2007).

The direct DNA extraction from the 0 to 21 cm core was done as follows: the soil was cored with a 2.5 cm diameter core from 0 to 21 cm depth. The core was subsampled every 3 cm (7 subsamples). Each subsample was mixed manually before DNA extraction. These subsamples were used for direct DNA extraction. Direct DNA

* Corresponding author at: Environmental Microbial Genomics, Laboratoire AMPERE, Ecole Centrale de Lyon, Université de Lyon, 36 avenue Guy de Collongue, 69134 ECULLY, France. Tel.: +33 4 72 18 65 14; fax: +33 4 78 43 37 17.
E-mail address: tvogel@ec-lyon.fr (T.M. Vogel).
URL: http://www.GenomEnviron.org (T.M. Vogel).

0167-7012/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.mimet.2011.06.013

Delmont study

Soil environments tend to have complex ecologies, and extraction techniques used there are important to ensure enough high-quality DNA for sufficient *representativity*.

Direct extraction

This is the most common and involves lysis of cells directly within the soil sample.

Uses around 1g of soil.

The DNA from all cells lysed will be present, regardless of source cell types.

This results in lower purity than indirect extraction, but the yield is up to 100x greater.

Indirect extraction

Cells are removed and separated using one of several possible methods and then lysed.

Around 60g of soil is used

Usually separates sequences by cell size, which has the affect of selecting for/against eukaryotes, prok/arch and viruses.

Does the reduced yield of indirect extraction result in a significant loss of functional diversity? What about the lower purity of direct extraction? Comments in the abstract don't precisely agree with findings in the body text. Discuss this in the forums.

When I review a Perl script:

DNA sequencing

The Hopkins course 410.666 – Genomic Sequencing and Analysis provides a thorough coverage of different sequencing technologies.

Here I'm going to summarize several of the most common ones used in metagenomics. These include:

- Sanger sequencing
- 454 / Roche
- Illumina / Solexa
- ABI SOLiD
- Ion Torrent / Ion Proton
- PacBio

After reviewing these slides check back on this lesson in Blackboard to find videos illustrating several of these technologies.

DNA crystal desk lamp.
Build your own on Instructables

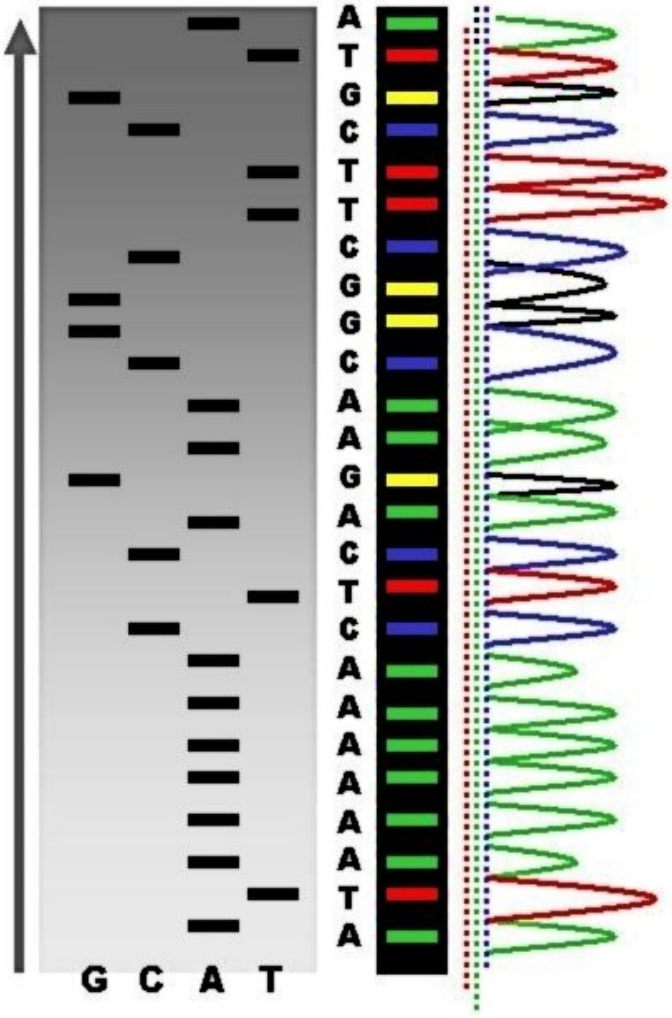
Sequencing - Sanger

Though there were previous methods of sequencing available, Sanger's chain-terminator method was one used in the first-generation sequencing machines. It involves replication of single-stranded templates using a pool of regular and modified (ddNTP) bases. Each time a ddNTP gets incorporated the template extension is halted.

This happens in so many copies of the template that random probability ensures chain termination at each base position. The sequences then separate by size when run through an agarose gel (or capillary) and the ddNTPs can be detected in order, revealing the sequence.

<u>The Good</u>: Still the gold standard for sequencing because of its low error rate, 700bp read length and large insert sizes. Great for generating near-complete genomes in low diversity metagenomic experiments.

<u>The Bad</u>: Labor-intensive cloning process, biased against genes toxic for the cloning host, and high cost (400k/Gbp)



[1] Thomas, T et al. 2012

Sequencing – 454 / Roche

Uses emulsion-PCR to clonally amplify random DNA fragments attached to microscopic beads. The beads are put into picotitre plate wells and pyrosequenced.

Pyrosequencing – Sequential addition of 4 dNTPs, the polymerization of which releases pyrophosphate, which is enzymatically converted to light. 1.2 million of these reactions are detected in parallel via a CCD camera.

Uses emulsion-PCR to clonally amplify random DNA fragments attached to microscopic beads. The beads are put into picotitre plate wells and pyrosequenced.

Pyrosequencing – Sequential addition of 4 dNTPs, the polymerization of which releases pyrophosphate, which is enzymatically converted to light. 1.2 million of these reactions are detected in parallel via a CCD camera.

[1] Thomas, T et al. 2012

Sequencing – 454 / Roche (cont'd)

The Bad:

• ePCR produces artificial replicate sequences, affecting abundance estimations. These can be handled with bioinformatics applications.
• Often gets homopolymer counts wrong, meaning repetitive strings of any given base are problematic. Assemblers have been parameterized to account for this.

The Good:

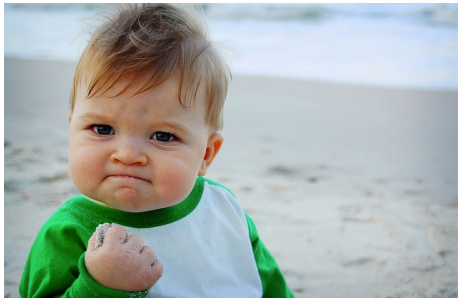
• $20k/Gbp with 600-800 bp read length
• 10s of nanograms of DNA needed for single-ended libraries, a microgram for paired-ends
• 12-sample multiplexing, so a single run = ~500 Mbp
• 1-day runtime

[1] Thomas, T et al. 2012

Sequencing – Illumina / Solexa

Random DNA fragments are immobilized on a surface and undergo solid-surface PCR, generating dense clusters of identical fragments. These are then sequenced with reversible terminators in a sequence-by-synthesis process.

<u>The Bad</u>:
• Some datasets have shown high error rates at read ends, which is improved by 'clipping' reads.
• Assembly might introduce bias due to suppression of low abundant species.
• 10-day runtime (!)



<u>The Good</u>:
• Cost: ~ $50 / Gbp with successful applications in metagenomics including draft genome generation in a complex dataset.
• Protocol needs 20ng for single-read, 500-1000ng for paired-end
• HiSeq2000 has 16 channels with 100s of millions of reads each, with a total of ~60 Gbp per channel.
• Read length is ~150bp and can be paired-end and strand-specific.

[1] Thomas, T et al. 2012

Sequencing: Other platforms

ABI SOLiD: Ultra-low error rate but only 50bp reads. Used mostly for mapping to reference genomes.

Ion Torrent / Ion Proton: Uses protons released during DNA polymerization to detect nucleotide incorporation. Promises 454 throughput with >100bp reads.

PacBio: Technology based on single-molecule, real-time detection in zero-mode waveguide wells. "Strobing" mimics paired-end reads but accuracy is currently 85%.

Complete Genomics: DNA nanoballs with combinatorial probe ligation. 35bp read length is extremely limiting in application.

We are not going to use data from any of these in this course, but I wanted them listed because many groups do use some of these technologies.

[1] Thomas, T et al. 2012

Paper: Field guide to ...

Field guide to next-generation DNA sequencers

TRAVIS C. GLENN
Department of Environmental Health Science and Georgia Genomics Facility, Environmental Health Science Building, University of Georgia, Athens, GA 30602, USA

Abstract

The diversity of available 2nd and 3rd generation DNA sequencing platforms is increasing rapidly. Costs for these systems range from <$100 000 to more than $1 000 000, with instrument run times ranging from minutes to weeks. Extensive trade-offs exist among these platforms. I summarize the major characteristics of each commercially available platform to enable direct comparisons. In terms of cost per megabase (Mb) of sequence, the Illumina and SOLiD platforms are clearly superior (≤$0.10/Mb vs. >$10/Mb for 454 and some Ion Torrent chips). In terms of cost per nonmultiplexed sample and instrument run time, the Pacific Biosciences and Ion Torrent platforms excel, with the 454 GS Junior and Illumina MiSeq also notable in this regard. All platforms allow multiplexing of samples, but details of library preparation, experimental design and data analysis can constrain the options. The wide range of characteristics among available platforms provides opportunities both to conduct groundbreaking studies and to waste money on scales that were previously infeasible. Thus, careful thought about the desired characteristics of these systems is warranted before purchasing or using any of them. Updated information from this guide will be maintained at: http://dna.uga.edu/ and http://tomato.biol.trinity.edu/blog/.

Keywords: 2nd and 3rd generation sequencing, 454, Helicos, Illumina, Ion Torrent, Life Technologies, massively parallel sequencing, Pacific Biosystems, Roche, SOLiD

Received 17 March 2011; revision accepted 22 March 2011

Background

DNA sequencing technologies and platforms are being updated at a blistering pace, so much so that reviews of sequencing platforms resemble the work of Sisyphus. It is important, however, for molecular ecologists to keep pace with these technologies, because they are transforming what we can do, how we should do it, and how much it will cost. Institutions and researchers are committing up to a million dollars to purchase massively parallel sequencing instruments. Such purchases lock laboratories and institutions into specific paths for large annual expenditures in both consumable supplies and service contracts. Differences in instrument engineering, platform chemistry and economics related to design constrain what can be done with those instruments once they are purchased.

Several recent major announcements and acquisitions make this an opportune time to evaluate available platforms and what is likely to be available in the immediate future. In this brief guide, I summarize instruments currently available and those that have been announced by major companies. Although several of these platforms have very different strengths touted by the vendors, the weaknesses are often much less clear. I have therefore summarized available information in tables with categories of primary interest to purchasers and to users so that direct comparisons can be made. I will use the convention of 2nd generation to indicate a platform that requires amplification of the template molecules prior to sequencing, 3rd generation to indicate platforms that sequence directly individual DNA molecules, and next-generation sequencing (NGS) platforms to generically indicate 2nd or 3rd generation instruments.

This guide is intended to provide information for readers with little or advanced understanding of NGS platforms. I assume, however, that readers who are not familiar with these systems are learning details by: reading relevant publications (e.g. Mardis 2008; Shendure & Ji 2008; Ansorge 2009; Richardson 2010; Tautz *et al.* 2010), reading information at company and independent websites and talking with staff of the companies making NGS instruments.

My purpose is not to explain how these systems work in detail (that information is readily available from the sources noted above), but instead to focus on generally important traits of these systems and to provide relevant details for prospective buyers and users. In particular, my goal is to present information useful to researchers

Correspondence: Travis C. Glenn, Fax: 706 542 7472;
E-mail: travisg@uga.edu

The previous slides gave brief overviews of each of several popular sequencing technologies as of mid-2011.

This excellent review by Glenn expands on these in great detail describing different read types, costs per run and sample, instrument and service contract costs, advantages and disadvantages of each, etc.

“No single platform can do everything that users will want and do it well or economically.”

Updates to these tables with 2012 values are being maintained by the author here:

http://www.molecularecologist.com/next-gen-fieldguide/

Further reading

[1] “Metagenomics – a guide from sampling to data analysis.”
Torsten Thomas, Jack Gilbert and Folker Meyer
http://www.microbialinformaticsj.com/content/2/1/3

[2] “Flow cytometry: A technology to count and sort cells.”
Megan Simmer – Science Creative Quarterly
http://www.scq.ubc.ca/flow-cytometry-a-technology-to-count-and-sort-cells/

[3] “Genomic DNA amplification by the multiple displacement amplification (MDA) method.”
Lasken RS.
http://www.ncbi.nlm.nih.gov/pubmed/19290880

[4] “Pyrosequence analysis of unamplified and whole genome amplified DNA from hydrocarbon-contaminated groundwater.”
Abbai NS, Govender A, Shaik R, Pillay B.
http://www.ncbi.nlm.nih.gov/pubmed/21656086