A biodiversity-based approach to development of performance enzymes

Applied metagenomics and directed evolution

Eric J. Mathur, Gerardo Toledo, Brian D. Green, Mircea Podar, Toby H. Richardson, Michael Kulwiec, and Hwai W. Chang*

Diversa Corporation
4955 Directors Place, San Diego, CA 92121
Phone (858) 526-5000
Fax (858) 526-5551
www.diversa.com

** Corresponding author: emathur@diversa.com*

Introduction

Many industrial processes catalyzed by chemical reactions could benefit from the use of enzymes; cost reductions, increased efficiency, improved recovery of products, and reduced use and disposal of toxic chemicals are just some of the advantages that enzymes can deliver. Industrial processes often take place under harsh reaction conditions of temperature and pH that also occur in natural environments. Through the targeted use of applied metagenomics and directed evolution, genes from uncultured microorganisms residing in these extreme environments can be recovered and optimized to produce enzymes with specificities and stabilities tuned to particular industrial reactor conditions.

It has become increasingly clear that we know very little about microorganisms that inhabit the natural environment. During the past hundred years of what might be considered modern microbiology, only about 10,000 strains have been well characterized. Yet, using cultivation-independent approaches to assess microbial diversity, single biotopes containing over 15,000 unique microbial genomes appear to be common in the biosphere. When attempting to isolate microorganisms from extreme environments (extremophiles), the situation is further exacerbated; very few genera of extremophiles are amenable to cultivation under laboratory conditions. As a result, the genes and enzymes produced from extremophilic microorganisms have been difficult to identify, recover, and commercialize.

At Diversa Corporation, we have developed technologies to rapidly obtain enzymes from uncultivatable microorganisms. For the past eleven years, Diversa has worked to advance the field of applied metagenomics by developing and optimizing a series of cultivation-independent recombinant techniques that enable the screening, cloning, and expression of genes and pathways derived from environmental microorganisms at previously unobserved rates. In addition, we have developed complementary directed evolution tools that can be used to improve and optimize the characteristics of the discovered enzymes to match industrial conditions. This Methods paper provides an overview of Diversa's technology platforms and describes how these are used to discover and evolve enzymes from nature. Two examples of commercial products that have been developed will be presented to illustrate this integrative approach for enzyme discovery.

Access to biodiversity

The starting point for all enzyme discoveries is biological diversity, or, simply stated, the total variety of life on earth. This includes genes, species, and ecosystems. The concept of biodiversity conjures up images of tropical rainforests replete with old-growth trees, birds, insects, and other charismatic macrofauna. However, recent molecular taxonomic inventories suggest that the vast majority of phylogenetic and metabolic diversity on earth resides in the genomes of microorganisms, the "blue collar workers" of the environment.

The first step in Diversa's discovery process involves gaining legal

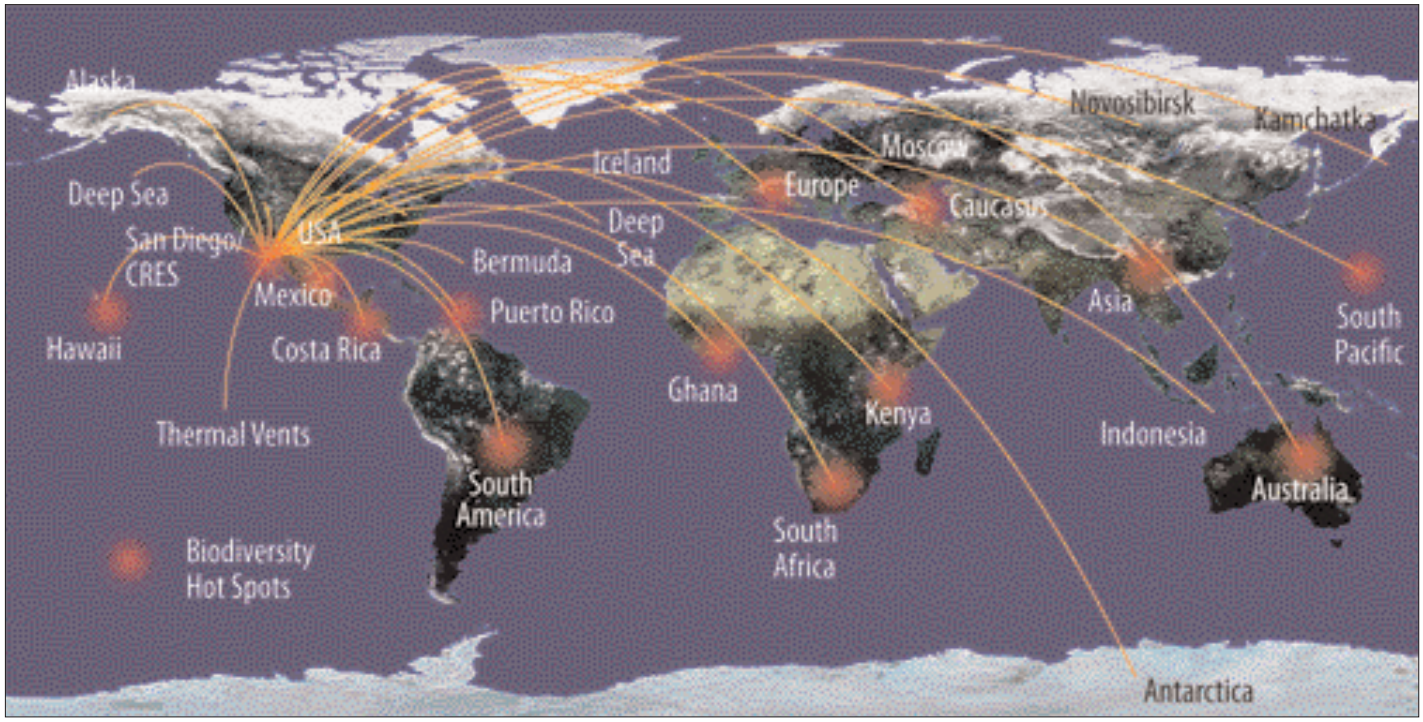

Figure 1. Diversa's access to global hotspots of biodiversity

access to collect samples from biodiverse environments *(see "Bioprospecting ethics and benefits: A model for effective benefit-sharing," p. 255)*. In keeping with the Convention on Biological Diversity (CBD), Diversa has established a bioprospecting program with research institutions and governments around the world, including INBio in Costa Rica; the Institute of Biochemistry in Leon, Ghana; the International Center for Insect Physiology and Ecology (ICIPE) in Kenya; and the Institute of Biochemistry and Physiology of Microorganisms (IBPM) in Pushchino, Russia, among others *(Figure 1)*. Diversa's bioprospecting framework agreements provide legal access to many of the global hotspots of biodiversity. In exchange for rights to patent the genes and gene products recovered from environmental samples and commercialize products, the source institutions and nations receive milestone payments and royalties on the products commercialized. Nonfinancial benefits derived from these collaborations include technology transfer, capacity building, and donations of supplies and equipment. At a higher level, these relationships serve to bolster economic and conservation goals underpinning the medical and agricultural advances needed to combat disease and sustain growing human populations[1].

Discovery platform

The samples collected for our discovery program typically consist of less than 50 grams of soil, sediment, leaf litter, or other environmental materials. The total microbial-community nucleic acids are first extracted from the samples and then converted into highly complex, representative metagenome (environmental) libraries[2]. Depending on the discovery application, the environmental libraries are sized to contain fragments ranging from 3,000 to 150,000 base pairs, and have the complexity to capture the majority of genes from the genomes of more than 15,000 unique microorganisms. Diversa currently possesses more than 4,000 of these metagenome libraries, which represent a vast resource and repository of genetic material for enzyme discovery. Since the libraries can be propagated within Diversa laboratories, there is seldom need to go back to the environment to resample. Thus, Diversa's sample collections minimize impact on the environment.

Due to the inherent complexity of microbial diversity, these metagenome libraries typically must contain 10^8 primary clones. (Depending on the insert size, this number of clones represents approximately 1X coverage of 10,000 3Mb genomes). In order to identify and recover enzyme candidates that fit a particular industrial performance profile, one must rapidly sift through these complex metagenome libraries to find the genes and gene products of interest. The term "applied metagenomics" refers to the combination of high throughput functional and sequence-based approaches designed to rapidly clone and express genes from metagenome libraries (uncultured microorganisms). Shotgun sequencing of metagenome libraries as a method of gene discovery has become more practical, as the cost of sequencing at genomic centers (such as the US Department of Energy's Joint Genome Institute) is now less that $0.001 per base. The art in sequence-based metagenome discovery lies in the high throughput recovery and expression of full-length gene products from shotgun sequence reads which often do not contain full-length genes. (See *Conclusion and Future Outlook* for more details on strategies for sequence-based gene discovery). In terms of functional or activity-based cloning using traditional agar plate-based methods, screening of even one of Diversa's

metagenome library would require 10,000 Petri plates, each containing 10,000 clones. This is simply not practical in an industrial setting.

To solve the throughput challenge, Diversa developed an ultrahigh throughput screening platform, Gigamatrix™ *(Figure 2)*. This system utilizes proprietary 400,000-well microtiter plates with the same footprint as traditional 96-well plates; screening one metagenome library with Gigamatrix technology can now be accomplished in less than three hours[3]. Another ultrahigh throughput proprietary screening platform, SingleCell™, is based on fluorescence-activated cell sorting (FACS). While FACS has been used as a tool for sorting eukaryotic cells, Diversa has developed and optimized the technology for sorting and screening recombinant bacterial cells, using either functional or sequence-based screens, at rates exceeding 10^7 cells per second[4].

An example of how the power of accessing biodiversity coupled with high throughput screening techniques can result in a commercial product is Luminase™, an enzyme resulting from an extensive search for xylanase enzymes with specific functionalities for use in the bio-bleaching of pulp for the production of paper. Incorporation of an enzymatic pretreatment step prior to the chemical pulp bleaching step can dramatically reduce the use of harsh chemical oxidizers such as chlorine dioxide. Activity at high pH and high temperatures and, most important, functional activity in laboratory-scale bio-bleaching assays were prerequisites for identifying candidates for this enzyme product. Ultimately, the Luminase enzyme was derived from a sample collected from an alkaline hot spring in the volcanic Uzon Caldera, in Kamchatka, Russia. Microbial nucleic acids were extracted from hot spring sediments and converted into metagenome libraries that were subsequently used in a functional screen for xylanase activity. A unique xylanase was discovered with properties and activity profiles consistent with the pH and temperature conditions required for pulp pretreatment. Moreover, it demonstrated functional activity in a bag bio-bleaching assay that is used in pulp mills to measure the effectiveness of bleaching systems.

The Luminase enzyme hydrolyzes the lignocellulose associated with the brown color of raw pulp and permits a reduction of up to 22% of chlorine dioxide usage during the chemical bleaching step. Not only is the Luminase product beneficial for the environment by decreasing the use and disposal of harsh chemicals, it represents a significant cost savings to pulp processors. In contrast to the long product development cycles of pharmaceutical products, the Luminase enzyme was developed in 30 months from sample collection to an EPA approval. It was introduced in late 2004 and is currently under evaluation through multiple large-scale pulp mill trials across the United States.

DirectEvolution® technologies

Diversa has developed two complementary gene-evolution technologies, Gene Site Saturation Mutagenesis (GSSM) and Tunable GeneReassembly™, designed to optimize enzyme candidates. Diversa's GSSM technology is a method of creating a family of related genes that all differ from a parent gene by at least a single amino acid change at any defined position[5,6]. This GSSM technology can produce all possible amino acid substitutions at every position within a polypeptide chain, removing the need for prior knowledge about the protein structure and allowing all possibilities to be tested in an unbiased manner. The family of variant genes created using GSSM technology is then available to be screened for enzymes (or antibodies) with improved qualities, such as increased ability to work at high temperature, increased reaction rate, resistance to denaturing chemicals, or other properties important in specific chemical processes. Beneficial mutations can then be combined in a combinatorial fashion using the Tunable GeneReassembly process to create a single highly improved version of the protein.

The Tunable GeneReassembly technology allows blending of gene sequences independent of sequence homology[7]. Multiple crossovers can be introduced at precise positions within the genes. The complexity of the variant library can be fine tuned by the number of parental genes used and the average number of crossovers allowed in the reaction. Moreover, the crossover frequencies can be modulated to reflect the resilience of the targeted gene family to mutations. For example, for a gene family which is mutation-sensitive, crossover frequencies can be optimized to reflect PCR-based DNA shuffling strategies; however, if the gene family is more resilient to mutations,

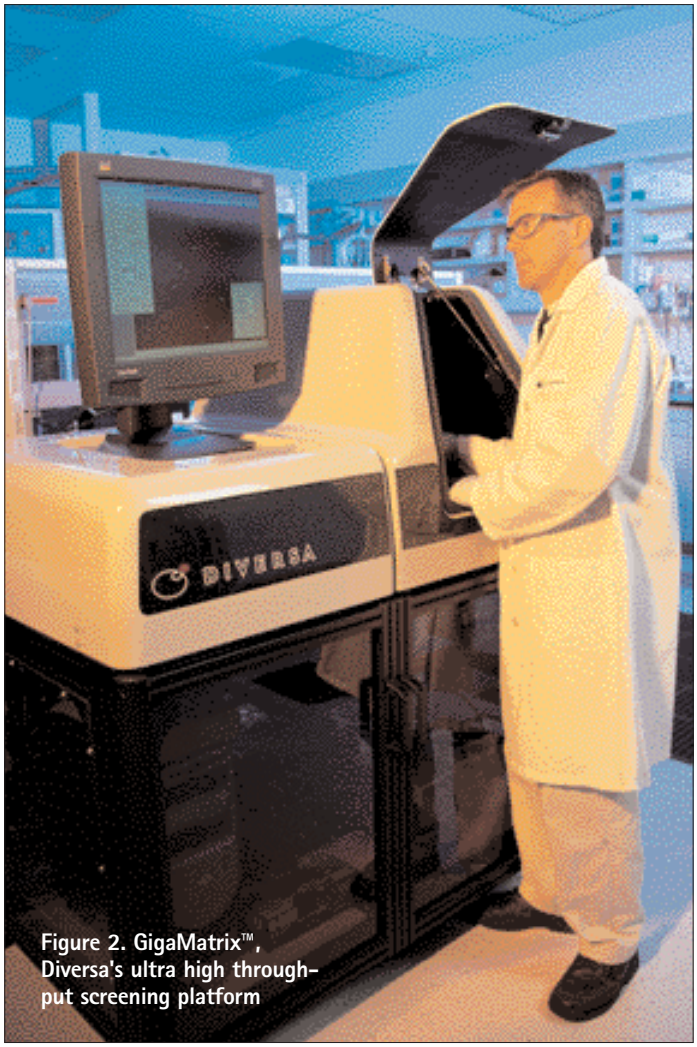

Figure 2. GigaMatrix™, Diversa's ultra high throughput screening platform

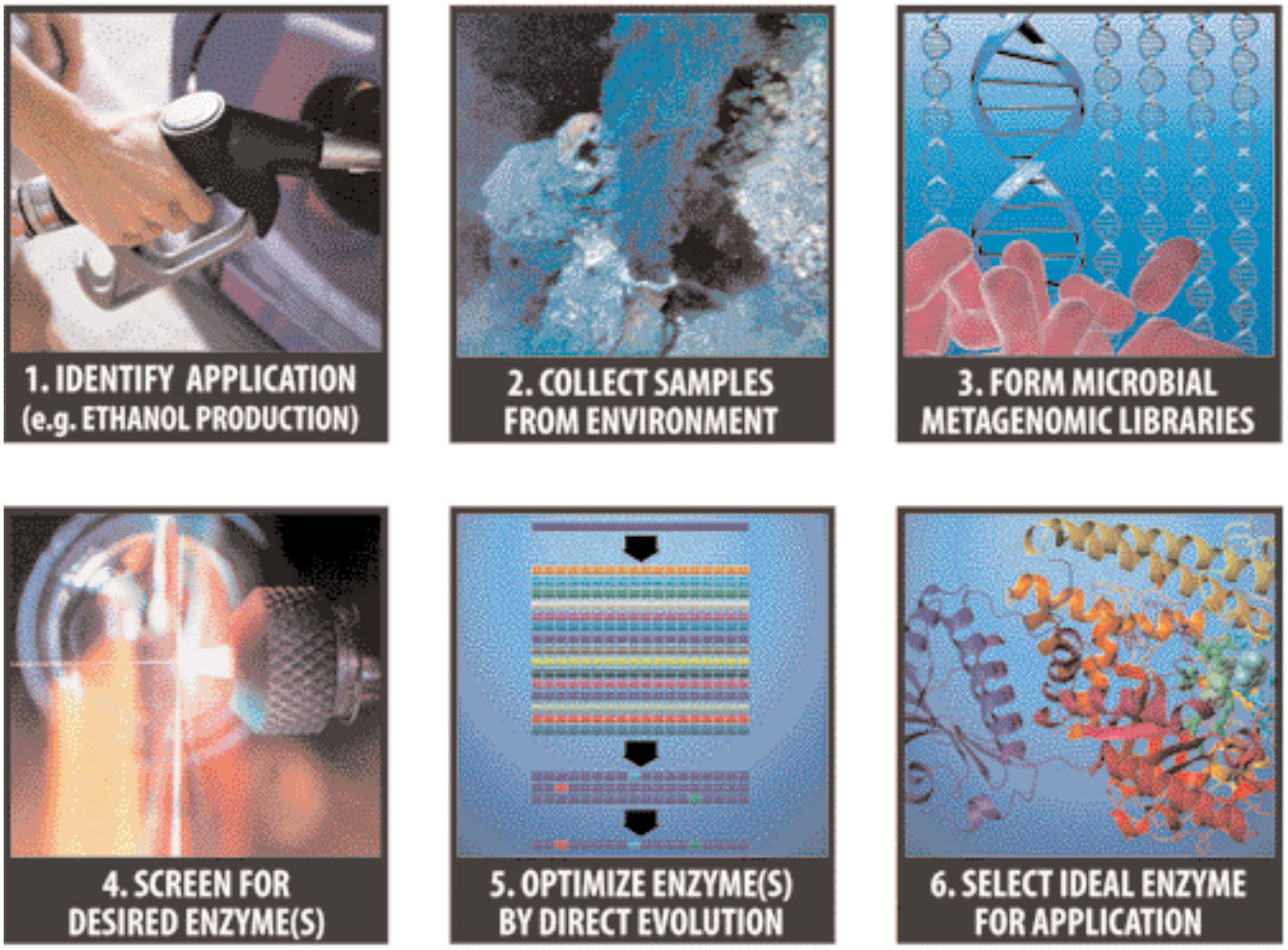

Figure 3. Diversa's discovery and evolution technology platforms

crossover frequencies can be increased accordingly. In addition, any structural information available can be incorporated into crossover-point decisions, and codon usage can be optimized during the reassembly process to maximize expression in the selected production host. As such, Diversa's Tunable GeneReassembly method represents a next-generation gene-blending evolution method where the location and extent of crossovers is in the hands of the investigator and not dependent on DNA polymerases. The method can be used to improve promoters, domains, proteins, and even entire pathways.

Development of Diversa's Ultra-Thin™ enzyme product serves as a demonstration of the power of combining natural discovery with Diversa's DirectEvolution technologies *(Figure 3)*. Ultra-Thin is an alpha-amylase used in corn wet and dry mill applications for the production of syrups and ethanol. The optimal genes that were used for development of the Ultra-Thin product were recovered from a black smoker sample collected at a deep-sea hydrothermal vent by the submersible Alvin. Hydrothermal vent chimneys are known to be colonized by hyperthermophilic Archaea that can grow at temperatures exceeding 110°C. The process of starch liquefaction occurs at temperatures of 105°C and pH of 4.5 and thus is similar to conditions in and around hydrothermal vents. Through a combination of sequence-based and functional applied metagenomics, three amylase candidates were selected, each of which exhibited one of the optimal characteristics for the performance specification: optimal activity at pH 4.5, optimal thermal stability at 105°C, and optimal expression in the selected production host. The three genes were blended, and a variant that possessed the optimal characteristics of all three parental genes was selected for product development[8]. The Ultra-Thin product is a specialty enzyme whose performance characteristics dramatically outperform competitor enzymes according to our tests and is currently being marketed by Valley Research (South Bend, Indiana) and used in starch liquefaction for the production of ethanol.

Conclusion and future outlook

This Methods paper has provided an overview of Diversa's discovery platform including both applied metagenomics and evolution

technologies, highlighting the value of these approaches in the context of product development. While metagenomics has become popular since the turn of the century, particularly with microbial ecologists, applied metagenomics was developed over ten years ago at Diversa. Back in the mid-nineties when Karl Woese and Norman Pace were refining the art of molecular phylogeny (PCR amplifying ribosomal genes from the environment)[9], a handful of Diversa scientists asked the question, Why can't we clone protein encoding genes from uncultured microorganisms?. Diversa has pioneered the field of applied metagenomics and has demonstrated the incredible potential for gene discovery from uncultured microbial diversity for specific use in the discovery of novel recombinant enzymes[10-14].

Until recently, shotgun sequencing of metagenome libraries was not a cost-effective or productive method for cloning and expressing genes from uncultured microorganisms. Bacterial genes are typically about 1 kb in size, while sequence reads are currently 750 bp. For this reason, shotgun sequencing rarely finds complete genes in one read. In addition, almost all microbial communities are inherently complex, and thus the paradigm of closing genomes or even building large scaffolds no longer holds true for metagenomics. In order to maximize the utility of genomic data generated by shotgun sequencing of highly complex metagenome libraries, a gene-centric (as opposed to genome-centric) approach must be utilized. Such a strategy looks at the global metabolism features of a microbial community and identifies physiological traits and enzymatic specificities without necessarily connecting them to a particular species or genome. These genes and gene fragments, termed "environmental gene tags," or EGTs[15], are not only important from the gene-discovery perspective and metabolic profiling of the community, but can also be used as markers in the form of gene arrays for rapidly screening environmental libraries for the presence of a desired trait, enzymatic specificity, or for monitoring fluctuations in the microbial community composition.

Diversa has an ongoing applied metagenomics collaboration with the Department of Energy's Joint Genome Institute and the California Institute of Technology focused on understanding how nature has developed efficient strategies for conversion of cellulosic materials to fermentable sugars. The model system to initiate these studies involves a systems approach to understanding the genomics and biochemistry of how the microbial symbionts residing in the hindguts of Costa Rica and Keyna higher termites convert wood material into fermentable sugars, hydrogen, and methane; certain species can efficiently convert wood particles into fermentable sugars in less than 24 hours. To accomplish the aims of this applied metagenomics program, a multipronged approach has been initiated that utilizes comparative metagenomics, taxonomic microbial inventories, and activity-based screening to identify, clone, and express the microbial community genes and pathways that produce the enzymes, small molecule mediators, and cofactors required for hydrolysis of lignocellulosic substrates. These techniques, combined with the novel cultivation methods[16] and complete genome sequencing of previously uncultured species, along with proteomic and metabolomic studies of the strains and termite hindgut lumen fluids, will enable a glimpse into how nature has solved the complex problem of biomass conversion. As we gain a better understanding of these natural processes and begin to elucidate the underlying mechanisms that then will become more amenable to biotechnological improvements, the possibility arises for industrial-scale bioreactors based on termite symbiont applied metagenomics that efficiently convert a wide range of agricultural wastes into fermentable sugars and alternative fuels.

ACKNOWLEDGMENTS

The authors thank the entire Diversa staff, as the work described here is the result of teamwork from all the departments. Special acknowledgment to Paul Zorner, Connie Hansen, and Jay Short for suggestions for this paper. In addition, the authors acknowledge the work of INBio in Costa Rica and ICIPE in Kenya, Jared Leadbetter from the California Institute of Technology, and Eddy Rubin from the US DoE Joint Genome Institute, for the termite collections and collaborative work.

REFERENCES

1. Mathur EJ, Costanza C, Christoffersen L, Erickson C, Sullivan M, Bene M, and Short JM. An overview of bioprospecting and the Diversa model. *IP Strategy* 11, 1-20 (2002).

2. Robertson D, Chaplin J, DeSantis G, Podar M, Madden M, Chi E, Richardson T, Milan A, Miller M, Weiner D, Wong K, McQuaid J, Farwell B, Preston L, Tan X, Snead M, Keller M, Mathur E, Kretz P, Burk M, and Short JM. Exploring nitrilase sequence space for enantioselective catalysis. *Appl Environ Microbiol* 70, 2429-2436 (2004).

3. Lafferty M and Dycaico MJ. GigaMatrix: A novel ultrahigh throughput protein optimization and discovery platform. *Meth Enzymol* 388, 119-134 (2004).

4. Short, JM and Keller M. High throughput screening for novel bioactivities. US Patent No. 6,872,526 (2005).

5. Short JM. Saturation mutagenesis in directed evolution. US Patent No. 6,764,835 (2004).

6. DeSantis G, Wong K, Farwell B, Chatman K, Zoulin Z, Tomlinson G, Huang H, Tan X, Bibbs L, Chen P, Kretz K, and Burk MJ. Creation of a productive, highly enantioselective nitrilase through gene site saturation mutagenesis (GSSM). *J Am Chem Soc* 125, 11476 (2003).

7. Short JM. Synthetic ligation reassembly in directed evolution. US Patent No. 6,537,776 (2003).

8. Pace NR. A molecular view of microbial diversity and the biosphere. *Science* 276, 734-740 (1997).

9. Robertson DE, Mathur EJ, Swanson RV, Marrs BL, and Short JM. The discovery of new biocatalysis from microbial diversity. *SIM News* 46, 3-8 (1996).

10. Robertson DE, Mathur EJ, Swanson RV, Marrs BL, and Short JM. The discovery of new biocatalysis from microbial diversity. *SIM News* 46, 3-8 (1996).

11. Short JM. Recombinant approaches for accessing biodiversity. *Nat Biotechnol* 15, 1322-1323 (1997).

12. Short JM. Protein activity screening of clones having DNA from uncultured microorganisms. US Patent No. 5,958,672 (1999).

13. Short JM. Gene expression library produced from DNA from uncultivated microorganisms and methods for making the same. US Patent No. 6,280,926 (2001).

14. Brennan Y, Callen WN, Christoffersen L, Dupree P, Goubet F, Healey S, Hernandez M, Keller M, Li K, Palackai N, Sittenfeld A, Tamayo G, Wells S, Hazlewood G. P, Mathur EJ, Short JM, Robertson DE, and Steer, BA. Unusual microbial xylanases from insect guts. *Appl Environ Microbiol* 70, 3609-3617 (2004).

15. Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, Bork P, Hugenholtz P, and Rubin EM. Comparative metagenomics of microbial communities. *Science* 308, 554-557 (2005).

16. Zengler K, Toledo G, Rappe M, Elkins J, Mathur EJ, Short JM, and Keller M. Cultivating the uncultured. *Proc Natl Acad Sci USA* 99, 15681-15686 (2002).