Next-generation sequencing of highly multiplexed samples: the NuGEN Encore™ 384 Multiplex System

Chris Raymond, Jill Magnus, Bonnie Kwong, Leah Turner and Joe Don Heath
NuGEN Technologies, Inc. 201 Industrial Road, Suite 310, San Carlos, CA 94070

NGS — The Challenge of Leveraging Increases in Sequencing Output

Advances in DNA sequencing throughput afforded by highly parallel, reversible-terminator sequencing platforms have completely redefined our view of biological inquiry. Crisp, high-resolution, digital views of whole-genome biology provide both profound clarity and significance. As with any disruptive technology, implementation creates turbulence. In the case of NGS, accelerated sequencing throughput has created challenges in upstream sample preparation and downstream data analysis. Initially, sample preparation was particularly problematic because it required individuals with highly specialized molecular biology skills for the hands-on execution of complex protocols. Fortunately, automated solutions for NGS library preparation have arrived, and it is reasonable to expect broad dissemination of turnkey solutions for routine library generation in the very near future. What remains problematic is that some sequencing projects involve large numbers of samples that require relatively shallow sequencing coverage. The sequencing costs of such projects are often prohibitive, calling for methods to multiplex high sample numbers in the same sequencing run. The obvious solution is to combine many samples into a single sequencing lane, and there are now many reports in the literature describing ad hoc methods and strategies for multiplexing samples prior to sequencing. Here we introduce a commercial solution for deep multiplexing of NGS samples that builds on our Encore™ NGS Multiplex System. The Encore 384 Multiplex System consists of a refined set of 384 molecularly "bar-coded" library adaptors that enable deep multiplexing of sequencing samples within a library preparation system that is compatible with high-throughput automation. The version of the product discussed in this White Paper contains adaptor sequences compatible with Illumina sequencing platforms.

Barcode Design

As with previous versions of NuGEN's Encore NGS Multiplex adaptors, three key considerations were addressed in the adaptor design (Figure 1). First, the codes must be "unambiguous," meaning that each code in our 384-plex collection is separated from all other codes in the barcode family by at least three discrete base pair differences. This ensures that sequencing reads with one or two base calling errors in the code region are not mis-assigned to the wrong sample. Perhaps more importantly, the most frequent errors in NGS libraries are introduced by micro-deletions of one or two nucleotides from the 3′ end of library adaptor molecules. Our codes are resistant to mis–assignments caused by these micro-deletion events. It was also important to make our unambiguous codes as short as possible so that the majority of sequencing resource is devoted to the sequencing of the sample inserts. We determined that a code length of eight nucleotides was necessary and sufficient to create the family of 384 unambiguous barcodes. The

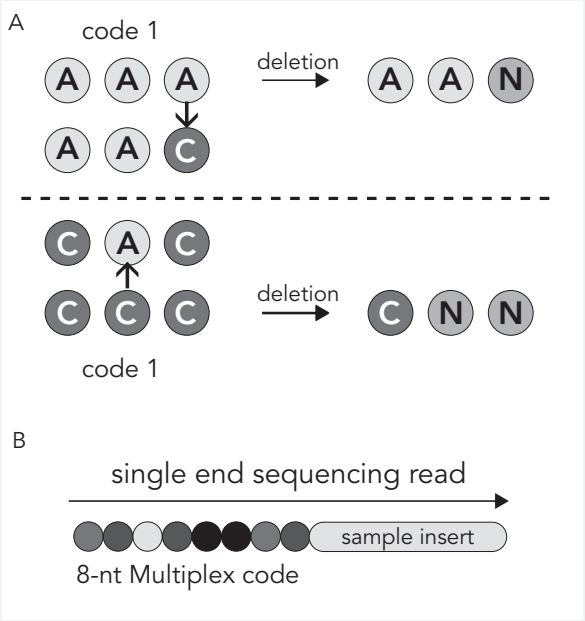
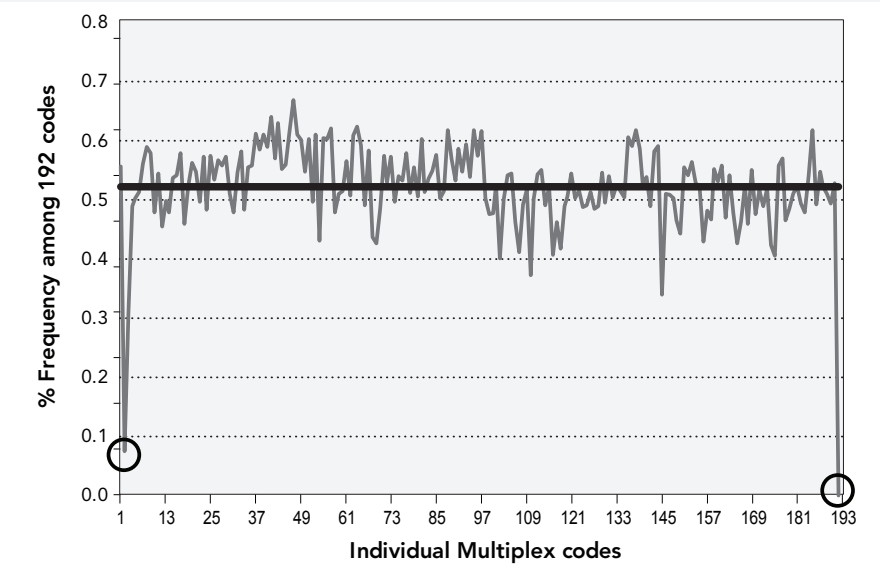


FIGURE 1. The Encore 384 Multiplex System. A) Unambiguous barcodes differ by at least three nucleotides. In this simplified example of code 1 (AAA) and code 2 (CCC), samples can, in principle, be grouped despite a single base calling error (↓). The assignment of a read to a code is also resistant to 3′ deletions in the code. In practice, only perfectly called codes are generally used in data sets, and the unambiguous feature prevents cross-talk of sequencing reads between distinct samples. B) Encore Multiplex codes are inserted upstream of sample inserts such that the code and the sample sequence occur on the same single read.

NuGEN®
imagine more from less™

A



B

% reads with error-free codes	% correctly coded reads mapping to reference genomes	% coded and mapped reads mapping to the correct genome
94.1%	**97.4%**	**99.99%**

FIGURE 2. Encore 384 Multiplex System proof-of-principle DNA sequencing. A) One hundred and ninety-two coded libraries were pooled and sequenced. The frequency of each barcode is shown by the gray line. The black, horizontal line represents the ideal frequency. Two barcodes (circles) were not well represented; one was a control sample while the source of the other drop-out is likely due to a reagent addition error. B) Percent mapping frequencies for the Multiplex data set.

second key design criterion was that our codes are situated in-line with the sequencing reads of sample inserts. This means that our multiplexing strategy is compatible with simple and economical single-read sequencing across leading NGS platforms.

Finally, current NGS base calling software relies on an equivalent balance of all four DNA bases in the first few bases sequenced in order to accurately calibrate base calls. Our barcodes are precisely balanced with all four bases at all eight positions within the codes. Importantly, subsets of our adaptors used in the Encore 384 Multiplex System (e.g., 48 or 96 codes) are also base balanced.

Proof of Principle

To validate the performance of our Encore 384 Multiplex System, we created 96 coded libraries with *Escherichia coli* (MG1655) genomic DNA and 96 coded libraries with human genomic DNA. Following PCR amplification of each library individually, these 192 libraries were pooled and sequenced as one sample on a single lane of an Illumina GAIIx sequencing run. We extracted two key metrics from the data (Figure 2). First, we evaluated the balanced representation of barcodes. Ninety-four percent of all reads possessed a complete code, and all but two codes were represented near the expected frequency of 0.5%; one code was intentionally left out of the library while the other underrepresented code was due to a reagent addition error. Second, we determined that 97.7% of all reads mapped to one of the two reference genomes, and "cross-talk" between libraries, i.e., reads from barcodes assigned to human libraries that inadvertently mapped to *E. coli* and vice versa, was less than one per ten thousand reads (0.01%). These performance metrics are consistent with our earlier releases of unambiguous barcodes in the Encore NGS Multiplex Systems.

What Sample Types?

We envision at least four immediate applications for high multiplex sequencing enabled by the Encore 384 Multiplex System. These include 1) sequencing of a limited number of PCR amplicons derived from a large number of samples,

2) (re-)sequencing of plasmid isolates from individual shotgun libraries, 3) simultaneous evaluation of experimental parameters that may impact highly specific NGS library applications and 4) quality control assessment of sequencing libraries prior to more in-depth sequencing. A subset of these barcodes (e.g., 48 or 96 codes) may also be useful for sequencing small viral or prokaryote genomes where extreme sequencing depth is not required in the experimental design.

The Encore 384 Multiplex System is available from NuGEN as four separate 96-reaction configurations. Each of these solutions has sufficient reagents to construct 96 independent NGS libraries, each with a unique barcoded adaptor for the Illumina platform. These products may be used in any combination to enable multiplex sequencing with 96, 192, 288 or 384 unique barcoded libraries. The Encore 384 Multiplex System has been fully automated on the Sciclone NGS Workstation (Caliper Life Sciences).

For more information on the Encore 384 Multiplex System or any of NuGEN's products, please email marketing@nugeninc.com

NuGEN Technologies, Inc.

Headquarters USA

201 Industrial Road, Suite 310
San Carlos, CA 94070 USA
Toll Free Tel: 888.654.6544
Toll Free Fax: 888.296.6544
custserv@nugeninc.com
techserv@nugeninc.com

Europe

P.O. Box 149
6680 AC Bemmel
The Netherlands
Tel: +31-13-5780215
Fax: +31-13-5780216
europe@nugeninc.com

For our international distributors contact information, visit our website

www.nugeninc.com