

# Barcodes for DNA sequencing with guaranteed error correction capability

A.R. Krishnan, M. Sweeney, J. Vasic, D.W. Galbraith and B. Vasic

Advances in DNA sequencing have resulted in instruments of remarkable performance, including extraordinary base read rates, and enormous sequencing depths. Sample throughput, nevertheless, remains slow, a situation that could be alleviated through sample multiplexing, with the incorporation of DNA barcodes serving to identify the different samples. Given the existence of finite sequencing error rates, reported is the design from BCH codes of DNA barcodes that provide guaranteed correction of errors within these barcodes.

**Introduction:** Next generation sequencing instruments produce hundreds of millions of short reads from a single run, and are poised to increase this output to billions of reads. However, many projects do not need or desire this depth of sequencing for a single sample, so multiplexing many samples together in a run allows this sequencing depth to be divided. In addition to the increase in sequencing capacity, techniques that allow targeted resequencing are being developed and optimised. Targeted resequencing decreases the amount of sequence needed and increases the need to multiplex in order to effectively use the available sequencing capacity. To be able to deconvolve sequencing data for each sample during data analysis, a short oligonucleotide sequence, termed a barcode, is added to each DNA fragment during library preparation. Each sample is labelled with a different barcode, and these DNA barcodes are sequenced along with the DNA or RNA from the sample, either as a paired run or a longer continuous read. Although the sequencing accuracy of next generation technologies is very high, given their extraordinary sequencing depth, sequencing errors become inevitable. As the number of multiplexed samples increases, the likelihood also increases that sequencing errors in the barcodes will prevent the definitive assignment of a sequencing read to a sample, resulting, at best, in loss of data or, at worst, in the transformation of one barcode into another, resulting in sample misclassification. There is, therefore, a need to develop barcodes that can automatically detect and correct errors introduced during sequencing.

**Methodology:** In the recent past, for designing DNA barcodes, researchers have turned to error-correcting codes, hitherto used for digital communication and data storage applications. Briefly, error-correcting codes add redundancy to correct errors in data. In particular, suppose that  $k$  bits need to be transmitted through a noisy channel. An additional  $m$  bits that reflect the parity of combinations of the  $k$  information bits are incorporated to make an error-correcting codeword that has a length of  $n$  bits, where  $n = k + m$ . The parity bits are used to enable the decoder to correctly decode the codeword in the presence of errors. The set of all codewords is referred to as a code. The error-correcting capability of these codes arises by virtue of the enhanced Hamming distance between codewords, as compared to uncoded data, the Hamming distance between two codewords being defined as the number of positions at which the two codewords differ. It can be proved that if, in a code, no two codewords have a Hamming distance less than  $d$ , then the code guarantees the correction of up to  $(d - 1)/2$  errors if  $d$  is odd and  $(d - 2)/2$  errors if  $d$  is even. The quantity  $d$  is referred to as the minimum distance of the code. In general, an  $(n, k, d)$  error-correcting code is of length  $n$  bits and encodes  $k$  information bits, having a minimum distance  $d$ .

In designing DNA barcodes based on error-correcting communication codes, some of the main concerns that need to be addressed are: (i) the error-correcting capability, i.e. the number of sequencing errors that the barcode can correct, (ii) the total number of distinct barcodes available for the design, and (iii) the decoding complexity, i.e. the computational time required for decoding each read of the barcode. Towards this end, Hamady *et al.* [1] used a short-length, binary Hamming code to create DNA barcodes. Though Hamming codes possess low decoding complexity, they have a minimum distance of 3, albeit very low. Indeed, it can be shown that these barcodes cannot guarantee the recovery from even one sequencing error (c.f. [2]). Error-correcting codes based on quaternary alphabets have yielded codes with higher minimum distances [3–6]. Unfortunately, no efficient and systematic decoding algorithms exist for decoding quaternary codes that are of interest in DNA

barcoding. Note that the time required for checking a read barcode against the database of all barcodes grows exponentially with the length of the code, thereby rendering such a naive algorithm unsuitable for all but the smallest-sized barcodes. Hence, we exclude this algorithm from our definition of the systematic decoding algorithm.

In this Letter, we describe DNA barcodes based on a family of binary, linear error-correcting codes termed Bose-Chaudhary-Hocquenghem (BCH) codes [7]. DNA barcodes designed from BCH codes are superior to codes previously designed in that they provide guaranteed error-correcting capability with low decoding complexity.

Fig. 1a shows a schematic diagram of our design methodology for the creation of barcodes from BCH error-correcting codes. For our design, we chose two codes belonging to the class of BCH codes. The first code (henceforth referred to as code I) is a (15, 7, 5) code. The second code (code II) is a (31, 11, 11) code. To make the length even, a bit is randomly appended to codes I and II, thereby making their length 16 bits and 32 bits, respectively. The minimum distance of code I is 5, so it can correct any combination of 2-bit errors. Similarly, code II can correct any combination of 5-bit errors.

The final step in the barcode-design is the conversion of the binary BCH code into a DNA barcode. For this, we used an assignment map that converts the four possible binary tuples, namely, 00, 01, 10, and 11, to one of the four bases. During the conversion, barcodes with homopolymer chains of length four or more were discarded. To increase the number of valid barcodes, we chose a position-dependent assignment map (c.f. [2]).

During decoding of these barcodes, the operations are performed in reverse (Fig. 1b) to obtain an estimate of the decoded barcode. The decoding is considered successful if the decoded barcode is the same as the original barcode.

This design procedure resulted in two barcodes based on code I and code II, which were of length 8 bases and 16 bases, respectively. Barcodes I and II have sizes 126 and 1755, respectively. Analytical evaluation of our barcodes shows that barcodes I and II guarantee correction of up to 1 and 2 sequencing errors, respectively (c.f. [2]).

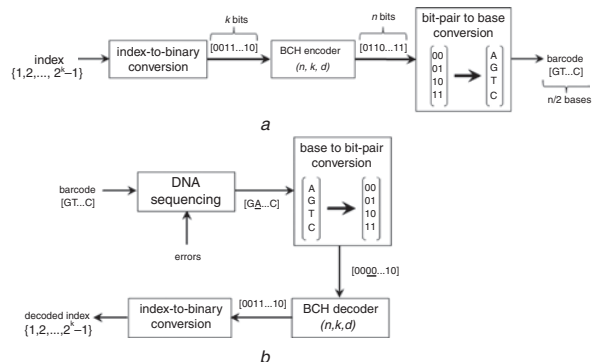


**Fig. 1** Schematic representation of design model and of decoding DNA barcodes

a First step is conversion of index to a binary representation. Subsequently, binary representation encoded using BCH code. Finally, conversion from bit-pair to bases yields final barcode  
b Barcode read is converted to binary BCH code, which is then decoded to yield final decoding index

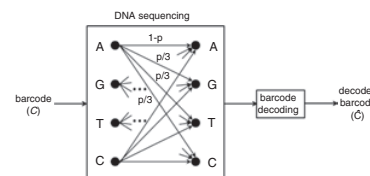


**Fig. 2** Schematic diagram of simulated study

With probability  $p/3$ , channel changes transmitted symbol to one of three remaining symbols, and transmits symbol without errors with probability  $(1 - p)$

To compare our barcodes with the barcode of [1], we conducted experiments that simulated DNA sequencing (Fig. 2). We considered the probability of sequencing errors,  $p$ , ranging from 0.03 to 0.001 (sequencing accuracy of 97–99.999%). This range includes the error

rates for the major NexGen sequencing technologies. From the results (Fig. 3), we observe that codes I and II have more than one/two orders of magnitude lower probability of decoding failure, when compared to barcodes of Hamady *et al.* For better quality sequencing (smaller values of  $p$ ) the improvement is many orders of magnitude. Furthermore, it can be seen that code II outperforms code I. This can be attributed to the fact that code II has a better error-correcting capability than code I.

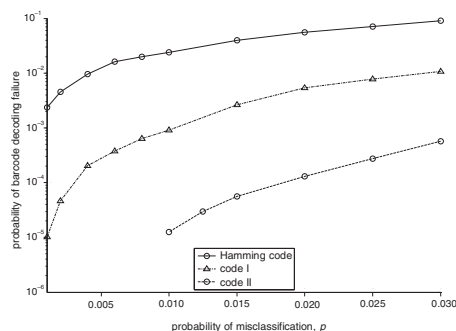


**Fig. 3** Performance comparison of barcodes in simulated study

**Conclusion:** We have introduced a barcode-design based on BCH codes. We have also presented two exemplar barcodes created using this design. Our codes provide improved resilience towards sequencing errors while also providing low-complexity decoding. Analytical evaluation of codes I and II proves that they can correct 1 and 2 sequencing errors, respectively, which is an improvement on the barcodes of Hamady *et al.* Furthermore, simulated studies indicate that our codes provide better error correction. This will result in fewer barcoded NexGen sequencing reads having to be discarded because sequencing errors prevented them from being definitely classified.

**Acknowledgment:** This work was funded by NSF under grants CCF-0830245 and CCF-0963726.

© The Institution of Engineering and Technology 2011

3 January 2011

doi: 10.1049/el.2010.3546

A.R. Krishnan and B. Vasic (Department of Electrical and Computer Engineering, University of Arizona, 1230 E. Speedway Blvd., Tucson, AZ 85721, USA)

E-mail: ananthak@ece.arizona.edu

M. Sweeney, J. Vasic and D.W. Galbraith (Department of Plant Sciences, 1140 E. South Campus Drive, Tucson, AZ 85721, USA)

## References

- 1 Hamady, M., Walker, J.J., Harris, J.K., Gold, N.J., and Knight, R.: 'Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex', *Nature Methods*, 2008, **5**, pp. 235–237
- 2 Krishnan, A.R., Vasic, B., Vasic, J., Galbraith, D.W., and Vasic, B.: 'Barcodes for DNA sequencing with guaranteed error correction capability: supplementary material', (Available from: [http://www.ece.arizona.edu/~ananthak/ElectronicsLetter/DNABarcode\\_SupplementaryMaterial.doc](http://www.ece.arizona.edu/~ananthak/ElectronicsLetter/DNABarcode_SupplementaryMaterial.doc))
- 3 Gaborit, P., and King, O.D.: 'Linear constructions for DNA codes', *Theor. Comput. Sci.*, 2005, **334**, pp. 99–113
- 4 Bogdanova, G.T., Brouwer, A.E., Kapralov, S.N., and Östergård, P.R.J.: 'Error-correcting codes over an alphabet of four elements', *Codes Cryptogr.*, 2001, **23**, pp. 333–342
- 5 Abualrub, T., and Ghayeb, A.: 'On the construction of cyclic codes for DNA computing', *J. Franklin Inst.*, 2006, **343**, pp. 448–457
- 6 Schmid, W.C., and Schurer, R.: 'MinT – online database of optimal parameters of  $(t, m, s)$ -nets', (Available from: <http://mint.sbg.ac.at/index.php>)
- 7 Lin, S., and Costello, D.: 'Error control coding: fundamentals and applications' (Prentice-Hall, Englewood Cliffs, NJ, USA, 2004, 2nd edn)