Bioorganic & Medicinal Chemistry Letters 20 (2010) 4188–4192

Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry Letters

journal homepage: www.elsevier.com/locate/bmcl

High-throughput sequencing for the identification of binding molecules from DNA-encoded chemical libraries

Fabian Buller [a], Martina Steiner [a], Jörg Scheuermann [a], Luca Mannocci [b], Ina Nissen [c], Manuel Kohler [d], Christian Beisel [c], Dario Neri [a,*]

[a] Institute of Pharmaceutical Sciences, Department of Chemistry and Applied Biosciences, ETH Zurich, Wolfgang-Pauli-Strasse 10, CH-8093 Zurich, Switzerland
[b] Philochem AG, c/o ETH Zurich, Wolfgang-Pauli-Strasse 10, CH-8093 Zurich, Switzerland
[c] Department of Biosystems Science and Engineering, ETH Zurich, Mattenstrasse 26, CH-4058 Basel, Switzerland
[d] Center for Information Sciences/Databases, ETH Zurich, Mattenstrasse 26, CH-4058 Basel, Switzerland

ARTICLE INFO

Article history:
Received 6 April 2010
Revised 12 May 2010
Accepted 13 May 2010
Available online 20 May 2010

Keywords:
DNA-encoded chemical library
High-throughput sequencing
Illumina
454
Drug discovery

ABSTRACT

DNA-encoded chemical libraries are large collections of small organic molecules, individually coupled to DNA fragments that serve as amplifiable identification bar codes. The isolation of specific binders requires a quantitative analysis of the distribution of DNA fragments in the library before and after capture on an immobilized target protein of interest. Here, we show how Illumina sequencing can be applied to the analysis of DNA-encoded chemical libraries, yielding over 10 million DNA sequence tags per flow-lane. The technology can be used in a multiplex format, allowing the encoding and subsequent sequencing of multiple selections in the same experiment. The sequence distributions in DNA-encoded chemical library selections were found to be similar to the ones obtained using 454 technology, thus reinforcing the concept that DNA sequencing is an appropriate avenue for the decoding of library selections. The large number of sequences obtained with the Illumina method now enables the study of very large DNA-encoded chemical libraries (>500,000 compounds) and reduces decoding costs.

The identification of small molecules, capable of specific binding to target proteins of interest, remains one of the most significant challenges for the discovery of new pharmaceuticals. DNA-encoded chemical libraries are large collections of small organic molecules, individually tagged with unique DNA fragments serving as amplifiable identification bar codes.[1–10] These libraries can be easily panned on target proteins immobilized on solid support.[2,5–7] Since the DNA-code distribution can be analyzed by PCR amplification and subsequent high-throughput sequencing before and after protein capture, DNA-encoded chemistry represents a new tool for the facile discovery of small organic binding molecules,[5–7,9–11] which can be used as such or can be further optimized by medicinal chemistry techniques.[7,12,13]

DNA-encoded compound libraries can be constructed by the stepwise combinatorial assembly of small molecule building blocks and the parallel extension of the DNA-code for each building block.[5,6,8] Both single pharmacophore[5,6,8] and dual pharmacophore library designs[4,14] can be considered, and a variety of different methods can be applied for linking chemical structures to their cognate DNA sequence tag.[1,3,5,6,8,11,15,16]

The first DNA-encoded chemical libraries of a few hundred compounds were decoded using DNA-microarray technology or by subcloning, bacterial amplification and conventional Sanger sequencing.[4,8,16,17] With the availability of novel methods for high-throughput DNA sequencing, we demonstrated the application of Roche's 454-technology for the deconvolution of amplicon mixtures corresponding to DNA-encoded compound libraries before and after selection.[5,7,18] 454 technology, which relies on emulsion PCR and subsequent pyrosequencing of DNA-coated beads in high-density picoliter reactors (picotiter plate),[18] is increasingly used for the decoding of single pharmacophore libraries.[5–7,11] However, while DNA-encoded libraries of thousands of compounds could be oversampled with 20–50,000 sequence tags corresponding to 1/8 of a 454 picotiter plate or >1 million sequence tags corresponding to a full 454 plate,[7,18,19] these numbers and the associated costs may become limiting factors when larger libraries are considered and when oversampling in decoding procedures is desirable.[6,11]

Illumina sequencing technology relies on the attachment of single-stranded DNA fragments to a solid surface (flow cell with eight flow lanes), bridge amplification of the single-molecule DNA templates and the subsequent sequencing by synthesis using reversible terminators.[19–23] In this study, we compared Illumina and 454 sequencing technologies for the analysis of DNA-encoded chemical library selections. In addition, we developed a novel sample preparation scheme, which allowed the mixing of different samples in the same sequencing experiment. Our results demonstrate that similar enrichment profiles could be observed with both

* Corresponding author. Tel.: +41 44 6337401.
E-mail address: neri@pharma.ethz.ch (D. Neri).

0960-894X/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmcl.2010.05.053

sequencing technologies, thus reinforcing the concept that high-throughput sequencing is an appropriate method for the analysis of DNA-encoded chemical library selections. Furthermore, the large number of sequence tags obtained from Illumina sequencing allowed the decoding of a 1 million DNA-encoded-compound library, indicating the general applicability of this sequencing technology for the analysis of DNA-encoded chemical libraries.

Illumina technology for the decoding of DNA-encoded chemical libraries was initially tested with a 4000 compound library obtained via Diels–Alder cycloaddition reactions, consisting of two sets of building blocks A (code 1) and B (code 2) (Fig. 1).[7,8] The library can be panned on immobilized antigen and the code distribution is subsequently analyzed by high-throughput sequencing before and after selection (Fig. 1).[7] When the library is decoded using 454 technology, specific DNA sequences are introduced by PCR (Fig. 1).[5,7,18] In order to enable Illumina sequencing, we replaced the 454 adapter sites with sequences recommended for the genomic sequencing using Illumina technology (Fig. 1; Supplementary data). The Illumina adapter sites are necessary for immobilization and bridge PCR amplification on solid support,[19,21,23] as well as for hybridization of the standard Illumina genomic DNA sequencing primer (Supplementary data). In order to reduce sequencing time and costs, it may be convenient to tag individual selections with a short DNA sequence ('tag'), thus allowing to mix PCR products corresponding to different selections, while retrieving all the necessary information by the computerized analysis of the experimental sequences (Fig. 1).

High-throughput sequencing results of the DNA-encoded chemical library can be visualized in three-dimensional plots, with 20×200 code combinations in the *xy* plane and sequence counts for individual library codes on the *z*-axis (Fig. 2). The sequence data obtained from 454 technology (50,000 sequences for each experiment, 1/8 of a picotiter plate)[7] are compared with the corresponding plots generated from the Illumina sequencing experiment (normalized to 1.3 million sequences for each experiment). The sequence count profiles for library selections on streptavidin and trypsin, chosen as model target proteins for these experiments, revealed strikingly similar distributions for the two sequencing technologies (Fig. 2).[7] For the streptavidin selection (Fig. 2A and B), the enrichment of compounds with building block code 2 = 200 indicated the preferential capture of library compounds containing a D-desthiobiotin moiety, a nanomolar binder to streptavidin.[24] In Fig. 2C and D, the enrichment profiles on trypsin revealed a selective capture of molecules containing a benzamidine moiety (code 2 = 6).[7]

To evaluate the Illumina sequence distribution of the library before selection, we used as a model a negative binomial distribution, in analogy to our previous analysis of 454 sequencing data.[7] As expected, an excellent agreement was observed between the experimental sequence count distribution of the library before selection

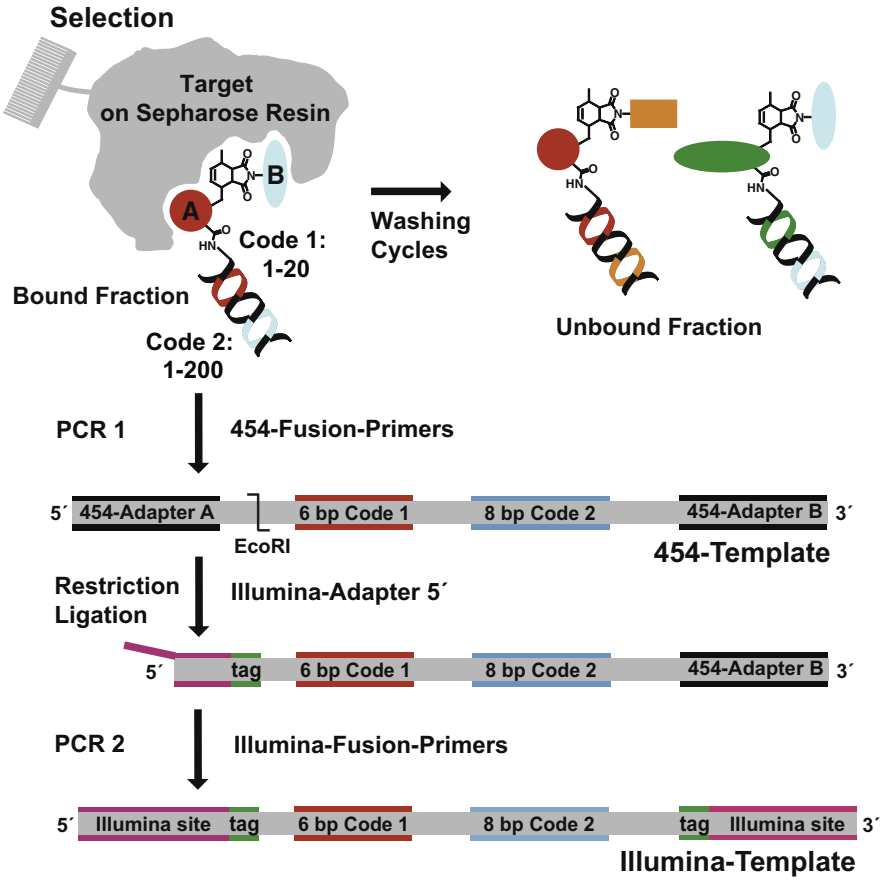

Figure 1. Scheme of a DNA-encoded chemical library selection and subsequent sample preparation for decoding using 454 or Illumina high-throughput sequencing. A 4000-member library containing two chemical building blocks linked by a Diels–Alder cycloaddition reaction and coding regions for each building block is panned against a target protein immobilized on sepharose resin.[8] Non-binding conjugates are washed away before PCR amplification with 454-fusion primers.[8] The resulting amplicon mixture is decoded using 454 high-throughput sequencing.[7] To compare the decoding results with Illumina high-throughput sequencing, the adapter sites are changed to the Illumina flanking sites using a two step procedure comprised of (i) an *Eco*RI restriction and subsequent ligation of a 5′ phosphorylated, partially double stranded Illumina-adpater oligonucleotide and (ii) a second PCR amplification using Illumina-Fusion primers. The introduction of barcodes ('tags') allows the mixing of different PCR-products in the same Illumina flow lane for parallel sequencing. A 5′ tag is introduced adjacent to the restriction site by ligation and a 3′ tag is added by PCR.

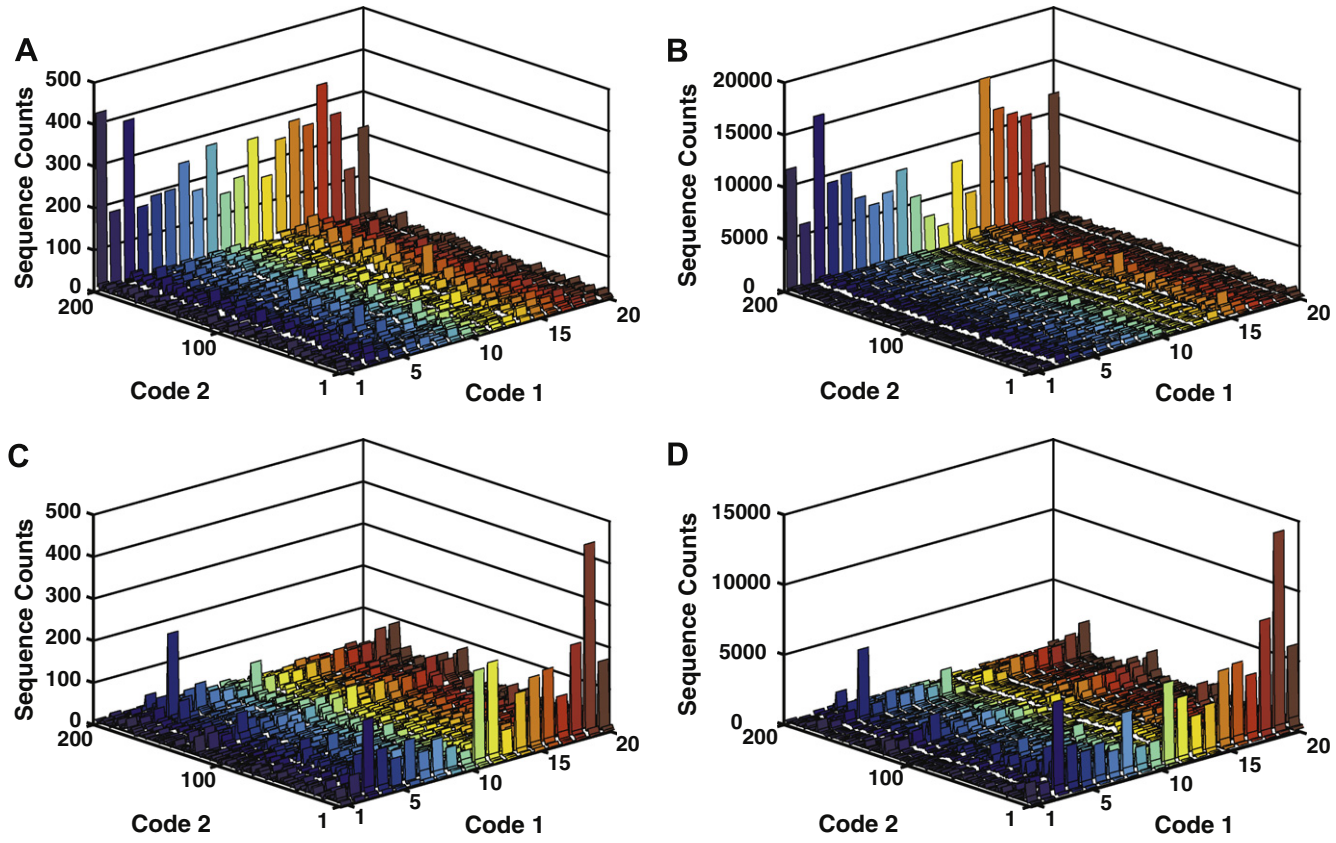

Figure 2. Sequence count profiles after high-throughput sequencing of a DNA-encoded chemical library using 454 or Illumina technology. The sequence profiles display 20×200 code combinations in the *xy* plane with the sequence counts for individual library codes on the *z*-axis. Selections were performed against the target proteins streptavidin (A and B) and trypsin (C and D). The previously described 454 decoding results (A and C, 50,000 sequences evaluated)[8] can be compared to the decoding by Illumina technology (B and D, 1.3 million sequences evaluated). Similar sequence count profiles between the two technologies indicate that both technologies can be applied for the quantification of DNA-codes after selection.

and the fitted negative binomial density function (Fig. 3A).[7] A quantile-quantile plot of the experimental and theoretical distributions represented a particularly useful tool for visualizing the model fit (i.e., equal distribution of all library members in the library before selection, Fig. 3B), whereas deviations from the diagonal indicate enrichment of specific binders after strepatavidin (Fig. 3C) and trypsin selection (Fig. 3D).[7] This model can also be used for the calculation of a one-sided *p*-value for each of the 4000 library codes, using as a null hypothesis the negative binomial probability density function based on the library PCR before selection, thus providing a statistical basis for the selection of enriched compounds which may deserve further analysis (e.g., resynthesis in the absence of the DNA tag).[7] For example, all twenty desthiobiotin conjugates in the streptavidin selection and the highest-enriched benzamidine conjugates in the trypsin selection exhibited *p*-values smaller than 0.001.

Sequencing power becomes particularly useful when library size grows (e.g., >100,000 compounds). Figure 4 depicts a pseudo-four-dimensional plot of the code distribution after selection for a library containing 1,000,000 compounds, obtained via Diels–Alder reactions (Buller et al., manuscript in preparation).[8] More than one million tags were analyzed in an inexpensive manner using Illumina technology, while the same task would have required the use of multiple picotiter plates using 454 technology.

The discovery of small organic molecules capable of specific binding to target proteins of interest is of fundamental importance not only for the Pharmaceutical Industry, but also for the growing field of Chemical Biology, where selective inhibitors can be used to dissect molecular pathways and interrogate protein function.[25] With advances in genome technologies, new targets become available, while certain previously known proteins remain 'undruggable' (particularly in the field of protein–protein interactions) using existing technologies.[26,27] The high-throughput screening of compound libraries is limited by costs and logistics. Major pharmaceutical companies can individually screen hundreds of thousands organic molecules in specialized assays.[28] However, these procedures are not always successful and hardly accessible for academic laboratories.

DNA-encoded chemical libraries offer a direct avenue to compound collections of unprecedented size.[4–7] As organic molecules are individually tagged by a distinctive DNA sequence, they can be used as a mixture, with panning procedures analogous to the ones used for the isolation of monoclonal antibodies from phage display libraries.[29] The use of DNA-encoded chemical libraries, however, crucially relies on the quantification of the relative abundance of individual compounds before and after selection on the target protein of interest. In this Article, we have shown that both 454 and Illumina high-throughput sequencing technologies are suitable for library decoding. These findings indicate that (i) the use of DNA sequence counts is an appropriate method for the quantification of the individual DNA-compound conjugates, and (ii) the procedure appears to be independent of the sequencing technology chosen (Figs. 2 and 3). While for small libraries both 454 and Illumina technologies provide the sequence information necessary for the identification of binding compounds, factors such as cost-per-base, sequence length and error rates influence the choice of a given sequencing method once library size grows. In our on-going decoding experiments, DNA-encoded chemical libraries (i.e., short

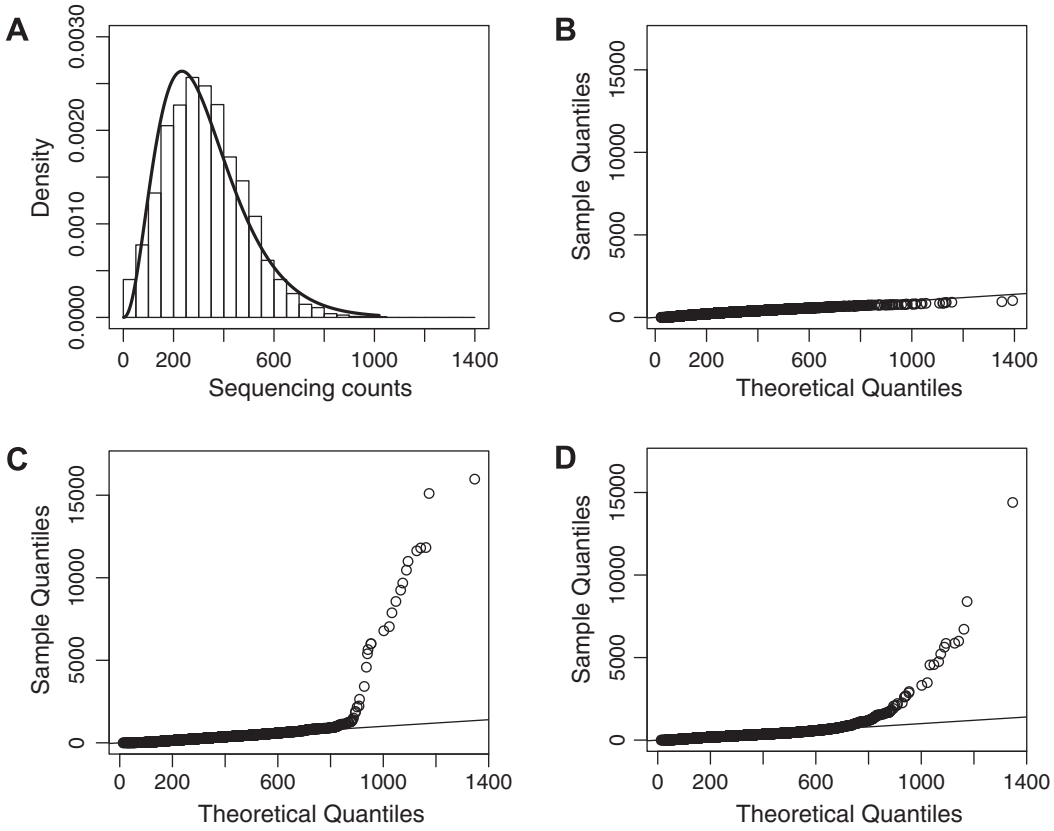

Figure 3. Statistical analysis of the Illumina high-throughput sequencing. Histogram of the sequence counts (A) obtained from the Illumina high-throughput sequencing of the DNA-encoded library before selection. The solid curve is the fitted negative-binomial probability density function (NB). The associated quantile–quantile plot (B) of sample versus theoretical NB quantiles indicates a good fit. Quantile–quantile plots of sample versus theoretical NB quantiles for the streptavidin selection (C) and the trypsin selection (D) show deviations from the diagonal revealing enriched DNA-codes that do not belong to the background distribution.

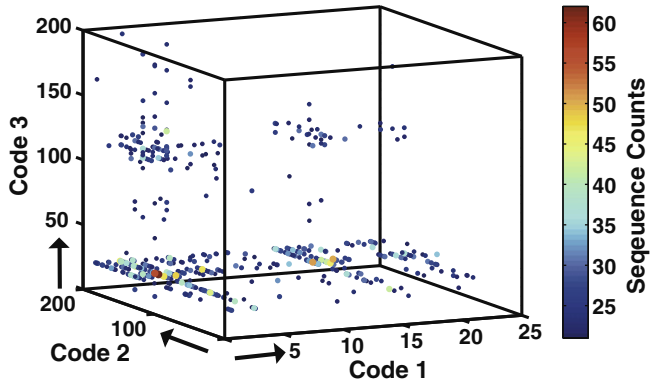

Figure 4. Illumina high-throughput sequencing is used for the decoding of a 1 million DNA-encoded compound library. The DNA-conjugates were prepared in a three-step split and pool synthesis using Diels–Alder chemistry and three encoding steps[7,8] (manuscript in preparation). The library was selected against immobilized antigen (matrixmetalloprotease 1a).[33] The pseudo four-dimensional plot displays the three library codes on the axis and the dots indicate sequence counts (higher than 20 sequences), while color and dot-size indicate the absolute sequence counts. The cut-off at 20 sequence counts was chosen arbitrarily to enhance the visual identification of enriched DNA-codes after selection. For this experiment, 1.4 million of sequences were evaluated and 0.55 million different code-combinations were found.

sequences of <100 bp) sequenced with the Illumina instrument have typically yielded ~10 millions of sequences of 76 bp length for each flow-lane (reagent costs per flow-lane: ca. 1100 EUR), while the 454 instrument yielded ~1 million of sequences of 100 bp length per picotiter plate (reagent costs per plate: ca. 3500 EUR).[30] In our hands, Illumina technology appears to be suitable for the decoding of libraries containing one million compounds (Fig. 4). While larger DNA-encoded chemical libraries can be synthesized (e.g., 800 million compounds),[6] this can only happen at the expense of growing molecular weights for the individual compounds (i.e., >600 Da). Molecules for pharmaceutical development properties typically need to be <500 Da and to fulfill other structural and physical criteria.[31,32] Thus, the need for the development of smaller focused DNA-encoded chemical libraries of drug-like pharmaceutical compounds remains,[31,32] while larger DNA-encoded chemical libraries with multiple building blocks can be explored in the context of targeting extracellular targets (e.g., in the development of injectable protein–protein interaction inhibitors).[26]

In conclusion, we believe that Illumina sequencing methods are appropriate for the decoding of DNA-encoded chemical libraries containing thousands or even millions of compounds. With growing library sizes and with the need to perform panning selections in multiple experimental conditions, we anticipate that developments in this field will be greatly influenced by advances in sequencing technologies. It will be interesting to see whether sequencing costs of existing technologies decrease, or whether alternative DNA-sequencing methodologies (e.g., parallel sequencing by ligation or single molecule sequencing)[19] can be used for the decoding of chemical libraries.

Acknowledgments

This work was supported by the ETH Zurich, the Swiss National Science Foundation, the Scholarship Fund of the Swiss Chemical Society (SSCI), Philochem AG, KTI [Grant number 8868.1 PFDL-LS]

and the Gebert-Rüf Foundation. We thank Dr. Yixin Zhang and Hannes Röst for help with the computational analysis of the data and Dr. Jean-Paul Gapian Bianké for reviewing the Letter.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bmcl.2010.05.053.

References and notes

1. Brenner, S.; Lerner, R. A. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89*, 5381.
2. Scheuermann, J.; Dumelin, C. E.; Melkko, S.; Neri, D. *J. Biotechnol.* **2006**, *126*, 568.
3. Tse, B. N.; Snyder, T. M.; Shen, Y.; Liu, D. R. *J. Am. Chem. Soc.* **2008**, *130*, 15611.
4. Melkko, S.; Scheuermann, J.; Dumelin, C. E.; Neri, D. *Nat. Biotechnol.* **2004**, *22*, 568.
5. Mannocci, L.; Zhang, Y.; Scheuermann, J.; Leimbacher, M.; De Bellis, G.; Rizzi, E.; Dumelin, C.; Melkko, S.; Neri, D. *Proc. Natl. Acad. Sci. U.S.A.* **2008**, *105*, 17670.
6. Clark, M. A.; Acharya, R. A.; Arico-Muendel, C. C.; Belyanskaya, S. L.; Benjamin, D. R.; Carlson, N. R.; Centrella, P. A.; Chiu, C. H.; Creaser, S. P.; Cuozzo, J. W.; Davie, C. P.; Ding, Y.; Franklin, G. J.; Franzen, K. D.; Gefter, M. L.; Hale, S. P.; Hansen, N. J.; Israel, D. I.; Jiang, J.; Kavarana, M. J.; Kelley, M. S.; Kollmann, C. S.; Li, F.; Lind, K.; Mataruse, S.; Medeiros, P. F.; Messer, J. A.; Myers, P.; O'Keefe, H.; Oliff, M. C.; Rise, C. E.; Satz, A. L.; Skinner, S. R.; Svendsen, J. L.; Tang, L.; van Vloten, K.; Wagner, R. W.; Yao, G.; Zhao, B.; Morgan, B. A. *Nat. Chem. Biol.* **2009**, *5*, 647.
7. Buller, F.; Zhang, Y.; Scheuermann, J.; Schafer, J.; Buhlmann, P.; Neri, D. *Chem. Biol.* **2009**, *16*, 1075.
8. Buller, F.; Mannocci, L.; Zhang, Y.; Dumelin, C. E.; Scheuermann, J.; Neri, D. *Bioorg. Med. Chem. Lett.* **2008**, *18*, 5926.
9. Halpin, D. R.; Harbury, P. B. *PLoS Biol.* **2004**, *2*, 1022.
10. Wrenn, S. J.; Weisinger, R. M.; Halpin, D. R.; Harbury, P. B. *J. Am. Chem. Soc.* **2007**, *129*, 13137.
11. Hansen, M. H.; Blakskjaer, P.; Petersen, L. K.; Hansen, T. H.; Hojfeldt, J. W.; Gothelf, K. V.; Hansen, N. J. *J. Am. Chem. Soc.* **2009**, *131*, 1322.
12. Pellecchia, M.; Bertini, I.; Cowburn, D.; Dalvit, C.; Giralt, E.; Jahnke, W.; James, T. L.; Homans, S. W.; Kessler, H.; Luchinat, C.; Meyer, B.; Oschkinat, H.; Peng, J.; Schwalbe, H.; Siegal, G. *Nat. Rev. Drug Disc.* **2008**, *7*, 738.
13. Keseru, G. M.; Makara, G. M. *Nat. Rev. Drug Disc.* **2009**, *8*, 203.
14. Scheuermann, J.; Dumelin, C. E.; Melkko, S.; Zhang, Y.; Mannocci, L.; Jaggi, M.; Sobek, J.; Neri, D. *Bioconjug. Chem.* **2008**, *19*, 778.
15. Gartner, Z. J.; Tse, B. N.; Grubina, R.; Doyon, J. B.; Snyder, T. M.; Liu, D. R. *Science* **2004**, *305*, 1601.
16. Dumelin, C. E.; Trussel, S.; Buller, F.; Trachsel, E.; Bootz, F.; Zhang, Y.; Mannocci, L.; Beck, S. C.; Drumea-Mirancea, M.; Seeliger, M. W.; Baltes, C.; Muggler, T.; Kranz, F.; Rudin, M.; Melkko, S.; Scheuermann, J.; Neri, D. *Angew. Chem., Int. Ed.* **2008**, *47*, 3196.
17. Melkko, S.; Zhang, Y.; Dumelin, C. E.; Scheuermann, J.; Neri, D. *Angew. Chem., Int. Ed.* **2007**, *46*, 4671.
18. Margulies, M.; Egholm, M.; Altman, W. E.; Attiya, S.; Bader, J. S.; Bemben, L. A.; Berka, J.; Braverman, M. S.; Chen, Y. J.; Chen, Z.; Dewell, S. B.; Du, L.; Fierro, J. M.; Gomes, X. V.; Godwin, B. C.; He, W.; Helgesen, S.; Ho, C. H.; Irzyk, G. P.; Jando, S. C.; Alenquer, M. L.; Jarvie, T. P.; Jirage, K. B.; Kim, J. B.; Knight, J. R.; Lanza, J. R.; Leamon, J. H.; Lefkowitz, S. M.; Lei, M.; Li, J.; Lohman, K. L.; Lu, H.; Makhijani, V. B.; McDade, K. E.; McKenna, M. P.; Myers, E. W.; Nickerson, E.; Nobile, J. R.; Plant, R.; Puc, B. P.; Ronan, M. T.; Roth, G. T.; Sarkis, G. J.; Simons, J. F.; Simpson, J. W.; Srinivasan, M.; Tartaro, K. R.; Tomasz, A.; Vogt, K. A.; Volkmer, G. A.; Wang, S. H.; Wang, Y.; Weiner, M. P.; Yu, P.; Begley, R. F.; Rothberg, J. M. *Nature* **2005**, *437*, 376.
19. Pettersson, E.; Lundeberg, J.; Ahmadian, A. *Genomics* **2009**, *93*, 105.
20. Bennett, S. T.; Barnes, C.; Cox, A.; Davies, L.; Brown, C. *Pharmacogenomics* **2005**, *6*, 373.
21. Bennett, S. *Pharmacogenomics* **2004**, *5*, 433.
22. Morozova, O.; Marra, M. A. *Genomics* **2008**, *92*, 255.
23. Adessi, C.; Matton, G.; Ayala, G.; Turcatti, G.; Mermod, J. J.; Mayer, P.; Kawashima, E. *Nucleic Acids Res.* **2000**, *28*, E87.
24. Dumelin, C. E.; Scheuermann, J.; Melkko, S.; Neri, D. *Bioconjug. Chem.* **2006**, *17*, 366.
25. Lehar, J.; Stockwell, B. R.; Giaever, G.; Nislow, C. *Nat. Chem. Biol.* **2008**, *4*, 674.
26. Wells, J. A.; McClendon, C. L. *Nature* **2007**, *450*, 1001.
27. Kramer, R.; Cohen, D. *Nat. Rev. Drug Disc.* **2004**, *3*, 965.
28. Mayr, L. M.; Bojanic, D. *Curr. Opin. Pharmacol.* **2009**, *9*, 580.
29. Winter, G.; Griffiths, A. D.; Hawkins, R. E.; Hoogenboom, H. R. *Annu. Rev. Immunol.* **1994**, *12*, 433.
30. Fox, S.; Filichkin, S.; Mockler, T. C. *Methods Mol. Biol.* **2009**, *553*, 79.
31. Lipinski, C. A. *J. Pharmacol. Toxicol. Methods* **2000**, *44*, 235.
32. Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. *Adv. Drug Deliv. Rev.* **2001**, *46*, 3.
33. Pfaffen, S.; Hemmerle, T.; Weber, M.; Neri, D. *Exp. Cell Res.* **2009**, *316*, 836.