nature chemical biology

Design, synthesis and selection of DNA-encoded small-molecule libraries

Matthew A Clark[1,3], Raksha A Acharya[1,3], Christopher C Arico-Muendel[1,3], Svetlana L Belyanskaya[1,3], Dennis R Benjamin[1,3], Neil R Carlson[1,3], Paolo A Centrella[1,3], Cynthia H Chiu[1,3], Steffen P Creaser[1,3], John W Cuozzo[1,3], Christopher P Davie[1,3], Yun Ding[1,3], G Joseph Franklin[1,3], Kurt D Franzen[1,3], Malcolm L Gefter[1,3], Steven P Hale[1,3], Nils J V Hansen[1,3], David I Israel[1,3], Jinwei Jiang[1,3], Malcolm J Kavarana[1,3], Michael S Kelley[1,3], Christopher S Kollmann[1,3], Fan Li[1,3], Kenneth Lind[1,3], Sibongile Mataruse[1,3], Patricia F Medeiros[1,3], Jeffrey A Messer[1,3], Paul Myers[1,3], Heather O'Keefe[1,3], Matthew C Oliff[1,3], Cecil E Rise[1,3], Alexander L Satz[1,3], Steven R Skinner[1,3], Jennifer L Svendsen[1,3], Lujia Tang[1,3], Kurt van Vloten[1,3], Richard W Wagner[1,3], Gang Yao[1,3], Baoguang Zhao[2] & Barry A Morgan[1,3]

Biochemical combinatorial techniques such as phage display, RNA display and oligonucleotide aptamers have proven to be reliable methods for generation of ligands to protein targets. Adapting these techniques to small synthetic molecules has been a long-sought goal. We report the synthesis and interrogation of an 800-million-member DNA-encoded library in which small molecules are covalently attached to an encoding oligonucleotide. The library was assembled by a combination of chemical and enzymatic synthesis, and interrogated by affinity selection. We describe methods for the selection and deconvolution of the chemical display library, and the discovery of inhibitors for two enzymes: Aurora A kinase and p38 MAP kinase.

Combinatorial biochemical methods such as phage display[1,2], RNA display[3,4] and aptamer SELEX[5–7] have emerged as robust approaches to the discovery of biopolymer-derived ligands to biological targets. Taking a cue from biological evolution, these methods employ large populations of species, and then focus the populations by selection under an environmental pressure. The most commonly used selection pressure is binding to a particular biological target. As in natural systems, each species carries an oligonucleotide 'gene' that specifically encodes its structure and serves as a template for its amplification. The selected species are peptide or oligonucleotide sequences, and so their structures can be determined by sequencing the gene. Using such strategies, libraries with 10^8 to 10^{13} members have been successfully interrogated for ligands to protein targets.

The success of biological display techniques has driven efforts to extend the strategy to the discovery of synthetic small-molecule ligands. The ability to generate small-molecule modulators of protein targets is central to the pharmaceutical industry. Currently, pharmaceutical firms can interrogate small-molecule populations with about 10^6 members, through high-throughput screening of corporate compound collections. By analogy with biological display methods, an encoded chemical display technology would increase the screening throughput by several orders of magnitude and require only modest instrumentation. Such a technology could provide a rapid and inexpensive route to the discovery of small-molecule drug leads.

A DNA-based bead-encoding strategy was first proposed in the 1990s[8]. Recent improvements have sought to enable affinity selection of soluble small-molecule mixtures[9]. These new techniques rely on covalent attachment of encoding DNA to the small-molecule library members. In one technique, a modified DNA oligo serves as a template for multistep molecular synthesis[10]. Building blocks are delivered as DNA conjugates, which are appropriately positioned for reaction by complementarity between the template and building block strands. While this strategy offers the potential of accurate, one-pot library synthesis and amplification, the library sizes do not yet approach those of biological methods. The largest reported so far

[1]The former Praecis Pharmaceuticals, Waltham, Massachusetts, USA, presently GlaxoSmithKline, Molecular Discovery Research Boston, Waltham, Massachusetts, USA. [2]GlaxoSmithKline, Computational and Structural Chemistry, King of Prussia, Pennsylvania, USA. [3]Present addresses: GlaxoSmithKline, Molecular Discovery Research Boston, Waltham, Massachusetts, USA (M.A.C., C.C.A.-M., S.L.B., N.R.C., P.A.C., C.H.C., J.W.C., C.P.D., Y.D., G.J.F., D.I.I., M.S.K., C.S.K., K.L., P.F.M., J.A.M., H.O., C.E.R., A.L.S., S.R.S., G.Y. and B.A.M.); Bedford, Massachusetts, USA (R.A.A.); Seattle Genetics, Inc., Bothell, Washington, USA (D.R.B.); Satori Pharmaceuticals, Cambridge, Massachusetts, USA (S.P.C.); ParagonDx, LLC, Morrisville, North Carolina, USA (K.D.F.); Lincoln, Massachusetts, USA (M.L.G.); Ensemble Discovery Corporation, Cambridge, Massachusetts, USA (S.P.H.); Vipergen, Copenhagen, Denmark (N.J.V.H.); AVEO Pharmaceuticals, Inc., Cambridge, Massachusetts, USA (J.J.); Foley & Lardner LLP, Washington, DC, USA (M.J.K.); Medway, Massachusetts, USA (F.L.); Massachusetts College of Pharmacy and Health Sciences, Worcester, Massachusetts, USA (S.M. and J.L.S.); Monsanto, Cambridge, Massachusetts, USA (P.M.); Columbia University, College of Physicians and Surgeons, New York, New York, USA (M.C.O.); Novartis Pharmaceuticals, Cambridge, Massachusetts, USA (L.T.); Biogen IDEC, Cambridge, Massachusetts, USA (K.v.V.); SRU Biosystems, Woburn, Massachusetts, USA (R.W.W.). Correspondence should be addressed to B.A.M. (barry.a.morgan@gsk.com).

Received 20 November 2008; accepted 19 June 2009; published online 2 August 2009; corrected after print 17 September 2009; doi:10.1038/nchembio.211

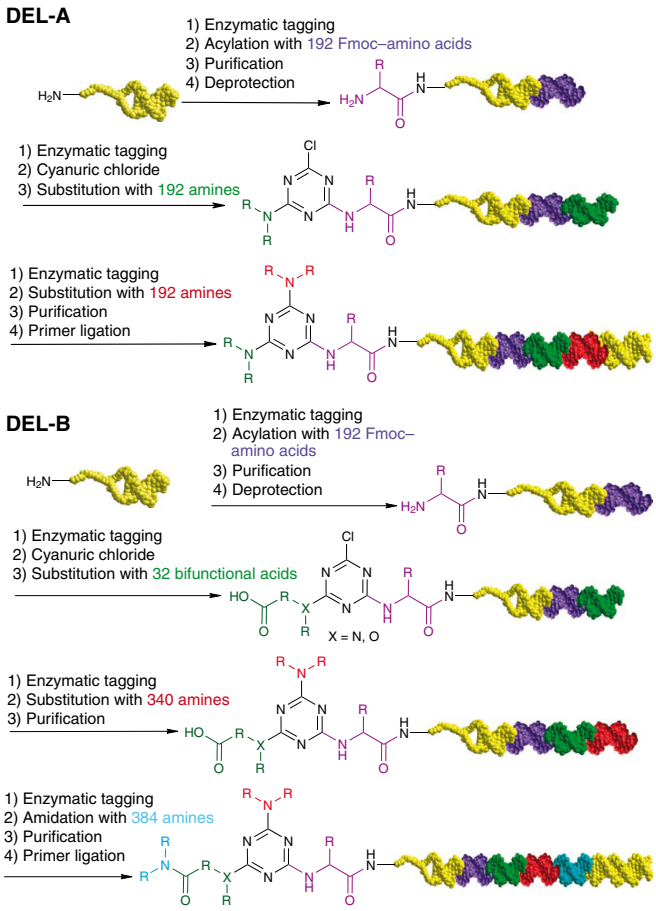


Figure 1 Synthetic schemes for DEL-A and DEL-B.

contained 100,000 compounds (see http://www.ensembleDiscovery.com/publications/pdfs/ensemblePoster_02.pdf).

A similar but distinct method has been reported[11]. Rather than templating the delivery of labeled building blocks, the oligonucleotide controls the physical routing of the library through multistep split-and-pool synthesis[11]. The advantage of this method is that library synthesis requires only 'off-the-shelf' building blocks and reagents. Libraries with 10^8 members have been created using this technique and successfully screened for new ligands[12]. Investigators have reported the interrogation of DNA-encoded libraries using both DNA-based self assembly[13] and encoded multistep synthesis[14]. More recently, workers at Vipergen have reported on a templating strategy employing three-dimensional DNA junctions[15].

We have developed an alternative technique for the synthesis and selection of DNA-encoded libraries (DELs) containing millions to billions of components[16]. In contrast to the methods described above, these DELs were encoded by double-stranded DNA, rather than single-stranded. The libraries were constructed using a combination of enzymatic and chemical synthesis in a split-and-pool format. Here we report the synthesis of two distinct DELs, and selection results against two kinase targets. Through a process of affinity selection, DNA sequencing and chemical synthesis, several families of potent kinase inhibitors were discovered. X-ray crystallographic studies revealed the binding mode of these inhibitors.

RESULTS

Library synthesis

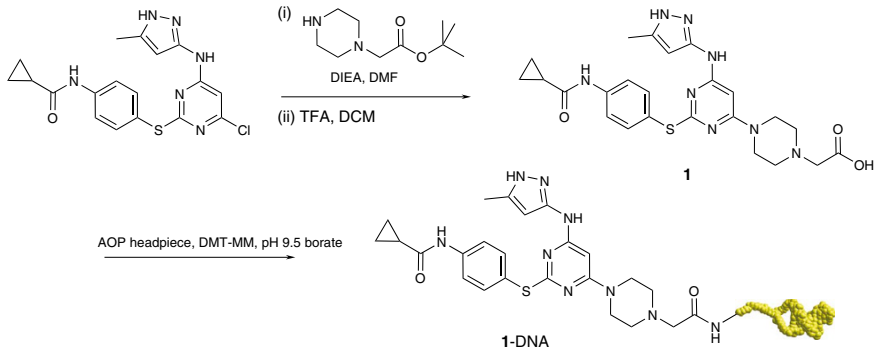
The precursor for library synthesis was a short, covalently linked DNA duplex—the 'headpiece' (**Supplementary Fig. 1**). Our choice of an encoding duplex contrasts with other DNA encoding techniques, which use single-stranded coding oligos. We believed that a duplex structure that buries the heterocyclic bases within the helix would protect the DNA from potential chemical lesions induced by library synthesis. We also speculated that double-stranded DNA was less likely to interfere with, or play a role in, the binding to an enzyme target. The headpiece was appended with Fmoc-15-amino-4,7,10,13-tetraoxapentadecanoic acid (AOP), which served as a spacer between the DNA and the small-molecule portion of the construct. After deprotection, the amine would serve as the starting point for the synthesis of the library. The nonlinked end of the headpiece duplex contained a two-base 3′ overhang, which formed a substrate for subsequent ligation to coding tags. The coding tags were short double-stranded DNA sequences consisting of a 7-base variable region flanked by constant 3′ overhangs. The overhangs, which served as "sticky ends" for ligation[17], were unique to each cycle of synthesis, so that each set of tags could only ligate to the set from the preceding cycle, and not to truncated sequences.

We designed two related libraries based on a triazine scaffold (**Fig. 1**). In both cases, 192 Fmoc amino acids were acylated onto the headpiece. After deprotection, the triazine was installed with cyanuric chloride. For DEL-A, the remaining chlorines on the triazine were substituted stepwise with 192 amines at each position, giving a three cycle library of 7,077,888 components (**Supplementary Dataset 1**). Among the amines at cycle 2, we included 3-amino-4-methyl-*N*-methoxybenzamide (AMMB), a known pharmacophore fragment for p38 mitogen-activated protein kinase (MAPK)[18]. For DEL-B, a small set of 32 amino acids was incorporated in cycle 2, and the number of amines used in cycles 3 and 4 was increased to 340 and 384, respectively, yielding a final library size of 802,160,640 over 4 cycles (see **Supplementary Dataset 2**).

The chemical diversity elements (herein referred as 'synthons') of the library were, with a few exceptions, commercially available reagents. We tested each potential synthon in one or more representative reactions with a DNA-linked substrate, and only those that gave 70–80% or greater conversion, depending on the library scheme, were included in library synthesis (see **Supplementary Methods** for a detailed description). Over 400 Fmoc–amino acids were tested as both electrophiles and nucleophiles to yield the 192 synthons in cycle 1. Over 1,000 amines were also tested for triazine substitution (DEL-A, cycles 2 and 3, and DEL-B, cycle 3) or for carboxylate amidation (DEL-B, cycle 4).

We conducted library synthesis using a split-and-pool strategy in 96-well plates (see Methods and **Supplementary Methods**). The starting duplex was arrayed into wells and ligated to cycle 1 DNA tags using T4 ligase. Following gel analysis of all wells to confirm quantitative ligation, the DNA was precipitated with ethanol. The pellets were then redissolved in buffer and subjected to acylation with Fmoc–amino acids. Progress of the chemistry was checked in a subset of wells by LCMS. After completion, the wells were pooled. In most cases, the pooled product was purified using reverse-phase HPLC. The product was again split into plates for entry into the next cycle of synthesis. After the last cycle of synthesis, a 30-base-pair primer sequence was ligated onto the library. This primer included a short randomized region that served as a control for PCR artifacts during subsequent sequencing.

It should be noted that because of the large diversity and high molecular weights of the library, it was difficult to accurately determine the reaction yield of any individual chemical transformation



Scheme 1 Synthesis of DNA-linked Aurora A kinase inhibitor VX-680. DIEA, *N*,*N*-diisopropylethylamine; DMF, *N*,*N*-dimethyformamide; TFA, trifluoroacetic acid; DCM, dichloromethane; DMT-MM, 4-(4,6-dimethoxy-1,3,5-triazin-2-yl)-4-methylmorpholinium chloride.

subsequent to cycle 1. Gross shifts in the average mass of the ensemble provided evidence that conversion was occurring (see **Supplementary Methods**), but a yield could not be calculated, nor could byproducts be identified or removed. Thus we relied on the validation procedure to ensure that all building blocks were competent toward the desired reaction. Of course, efficient reactivity in a model system was no assurance that an equally efficient transformation occurred in the hugely complex milieu of the library. But as we shall see, because affinity selection and deep sampling generated families of related structures, the impact of individual reaction failures during library synthesis was minimized.

Proof of concept

Before selection of the DELs against enzyme targets, we undertook a series of experiments aimed at demonstrating the feasibility of affinity selection. Since the DELs could not be amplified, we presumed there would be a limit to how many rounds of selection they could undergo. Due to physical losses incurred during the selection process, we found that more than three or four rounds of selection would reduce the population to undetectable levels. Based on the data available from SELEX-type techniques, we anticipated that the population obtained after three rounds of selection would be very heterogeneous. Population convergence usually required up to 10 or more selection rounds in other techniques, and so we predicted that high-affinity binders would comprise only a small fraction of the population after three rounds. We realized that only a very high sampling depth would allow characterization of such a poorly enriched population. Put simply, the more sequences we could obtain, the less convergence of the population would be needed to resolve a signal. In most biochemical display techniques, only 100–1,000 sequences are obtained, using traditional cloning methods. Newer high-throughput sequencing technologies can yield over 100,000 individual sequences in a single experiment[19]. We hypothesized that the use of such high-capacity sequencing methods would allow us to characterize an enriched population obtained from only a few rounds of selection[20].

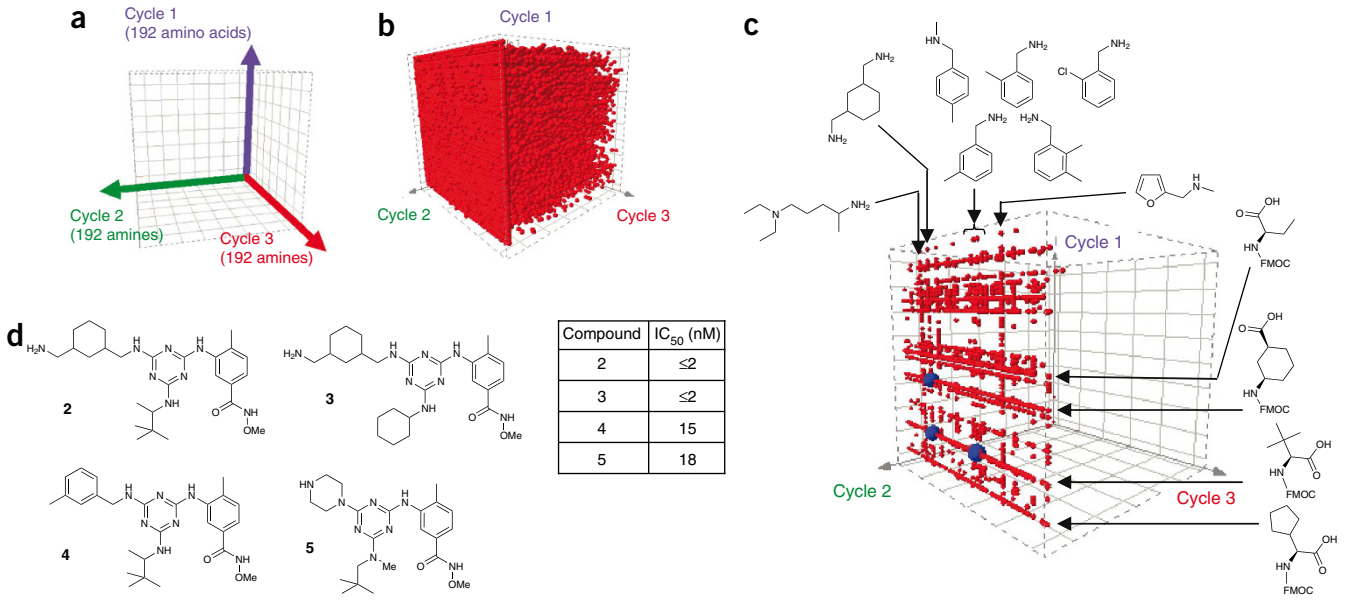
To test this hypothesis, we synthesized an analog of the Aurora A kinase inhibitor VX-680 (ref. 21) and attached it to the AOP headpiece (**Scheme 1**; also see **Supplementary Methods**). The resulting inhibitor-DNA conjugate **1** was tested for inhibition of Aurora A kinase and was found to retain a significant portion of the activity of the parent (K_i = 20 nM for **1**-DNA compared to 1 nM for VX-680). We then turned to the process of affinity selection against Aurora A. In order to efficiently separate the binding library members from the nonbinders, we used a protein construct that incorporated the affinity tag 6His. The affinity tag would allow the target to be isolated by immobilization on a solid matrix such as IMAC (immobilized metal affinity chromatography) resin. Once the target was immobilized, the removal of nonbinding library members would be achieved by simply washing the resin. Binders could then be eluted by heat denaturation of the target followed by additional rinsing.

In practice, two selection methods were typically used (see **Supplementary Methods**). In method A, we incubated an aliquot of library with protein. Blocking agents such as sheared salmon sperm DNA and bovine serum albumin were included in the buffer to reduce nonspecific binding. We immobilized the protein on resin and washed away the nonbinders. In method B, we first immobilized the protein on the resin at a concentration sufficient to saturate the matrix. The library aliquot was then added to the immobilized protein target. After incubation, the nonbinders were rinsed away. In both methods, we then denatured the protein and recovered the binding population. This eluant was then incubated with fresh protein to start the next round of selection. Typically three rounds of selection were performed. In order to rule out molecules whose enrichment was due to binding to the resin matrix, we conducted parallel selections in an identical fashion, but without the addition of protein target ('no target control'). Library members that were observed to be enriched in the no target control were omitted from the analysis of target-based selection experiments.

The Aurora A inhibitor-DNA conjugate **1** was elaborated with 4 cycles of tags and a closing primer, and introduced into DEL-A at the concentration of a single library member (that is, 1 in 7 million-fold dilution). After 3 rounds of affinity selection using method A, the positive control occurred 971 times in 66,201 sequence reads, for a ratio of 1:70. This result corresponded to a 100,000-fold enrichment during selection, which gave us confidence that high-affinity binders could be enriched and identified from complex mixtures. This result also illustrated the importance of the high-throughput sequencing. If we had obtained only 100 sequences, the known inhibitor may not have been distinguishable from the background signal.

We next turned to selection of DEL-A against p38 MAPK. As discussed above, we had designed DEL-A to contain a known family of p38 MAPK inhibitors, by virtue of the inclusion of AMMB as a synthon in cycle 2. Selection of DEL-A against 6His-tagged p38 MAPK was performed using method A, followed by PCR and DNA sequencing. We found it convenient to visualize the selection output as a cubic scatter plot (**Fig. 2**). Each axis of the plot represented one cycle of synthesis, and contained a number of locations equal to the number of synthons used in that cycle. Thus any member of the three-cycle DEL-A could be represented at a unique location in the cubic space. This visualization allows families of related structures to appear as lines (2 synthons in common) or planes (1 synthon in common). Some molecules were found to be more highly represented than others in the selected population; this was reflected by their DNA sequence appearing more often during sequencing. In order to focus on the most highly selected families, we typically removed the low-copy-number molecules from the analysis.

In this example, we obtained 190,363 sequence reads, which encompassed 89,314 unique sequences. The visualizations in **Figure 2b**,**c** show the population before and after removing all sequences that



Compound	IC_{50} (nM)
2	≤2
3	≤2
4	15
5	18

Figure 2 Selection of DEL-A against p38 MAPK. (**a**) Library populations can be viewed in three dimensions using Spotfire Decision Site 8.1.1 (Spotfire US). For DEL-A each axis of the cube consists of 192 locations, representing the synthons used in a given cycle of synthesis; consequently, each of the 7,077,088 components can be uniquely identified by a single point in the cubic space. (**b**) Unfiltered population obtained from DEL-A selection with p38 MAPK. There are 89,314 individual structures represented. (**c**) Same population after removing the low occurring molecules (less than 10 copies); 1,970 structures remain. Points corresponding to compounds **2**, **3** and **4** are colored blue. The structures and locations of some selected synthons are shown. (**d**) Structures and assay results of selected compounds **2**, **3** and **4**. The structures are very similar to reported p38 MAPK inhibitors such as **5**.

occurred less than 10 times, leaving 61,929 reads of 1,970 uniques. While filtering at 10 copies was arbitrary, it did remove the background and bring into focus the features of the selected population. After filtering, 33% of the total sequence reads remained; this is a very high level of signal in our experience, and it reflects the fact that DEL-A was designed to contain inhibitors of p38 MAPK (for a list of the compounds that remained after filtering, see **Supplementary Dataset 3**).

As expected, the population selected by p38 MAPK was dominated by AMMB at cycle 2. This is illustrated in **Figure 2c** as a highly populated plane, whose location corresponds to that of AMMB along the cycle 2 axis. Additionally, it was found that a short, branched alkyl amine was preferred as the cycle 1 substituent on the triazine, and a diamine or benzylic amine was preferred at cycle 3. Synthesis of several members of this family without the DNA tag yielded compounds **2**, **3** and **4**, which were potent inhibitors of p38 MAP kinase (half-maximal inhibitory concentrations (IC_{50}s) = 4–10 nM, **Fig. 2d**). The structure-activity relationship (SAR) defined by the library selection was broadly in agreement with results obtained by conventional medicinal chemistry approaches, as exemplified by compound **5** (IC_{50} = 18 nM)[18].

Selection of Aurora A kinase inhibitors

We next turned our attention to Aurora A kinase. Selections with Aurora A were highly reproducible; in two independent selections using method A, 63% of the selected molecules were found to occur in both populations (data not shown). Examination of the structures of selected families revealed that the compounds were unrelated to known Aurora A kinase inhibitors, although they did contain features broadly associated with kinase inhibition[22]. One family strongly selected by both methods was defined by the presence of 6-aminoquinoline as the cycle 2 synthon, with several amines preferred in cycle 3, and no observable preference in cycle 1. Notably, different cycle 3 amines were preferred using method A compared to method B, as evidenced by the different positions of the vertical lines within the 6-aminoquinoline plane (**Fig. 3a**).

A second family was defined by the presence of 7-azatryptophan (7-AT) in cycle 1, with a variety of amines at cycles 2 and 3. Since 7-AT did not strongly correlate with any particular synthon in cycles 2 or 3, it did not form obvious lines in the cube views. Nevertheless, 7-AT was the most prevalent cycle 1 synthon observed in both selections, as demonstrated by the histograms in **Figure 3b** (for the output of both selections, see **Supplementary Datasets 4** and **5**).

In contrast to the p38 MAPK results discussed above, the signal in the Aurora selections was much lower. For method A, we obtained 65,230 reads of 57,378 unique sequences. Filtering sequences with <4 copies gave well-defined families in the cube plot, but left only 821 reads of 148 uniques. This corresponds to 1.2% of the total data, compared to 33% for p38 MAPK after filtering. Method B was similar, with 1.0% of the total reads generating the filtered cube plots in **Figure 3a**. Again, without deep sequencing, identifying these features would have been impossible.

In order to elucidate the binding site of the selected molecules, a competition selection was performed. The ATP-competitive inhibitor VX-680 (not attached to DNA) was added to the selection buffer, and the Aurora A selection was repeated using method A. After sequencing, the previously selected families were no longer apparent. This result indicated that the selected molecules were likely interacting with the ATP-binding site.

A representative set of the selected molecules was synthesized (**Fig. 3c**). For the purposes of small-molecule synthesis, the linker to the encoding DNA had to be "cleaved" at some position. To achieve this, cycle 1 amino acids were sometimes replaced with their decarboxy analogs: thus, 7-AT was replaced with 7-azatryptamine. Assay data indicated that deletion of the linking carboxamide did not negatively affect potency (see compounds **6** and **7**). After purification, the molecules were subjected to biochemical assay. The inhibition constants ranged from 0.27 to 7.6 μM. Notably, the most potent compound in the series shared elements of both the selected families (compound **11**).

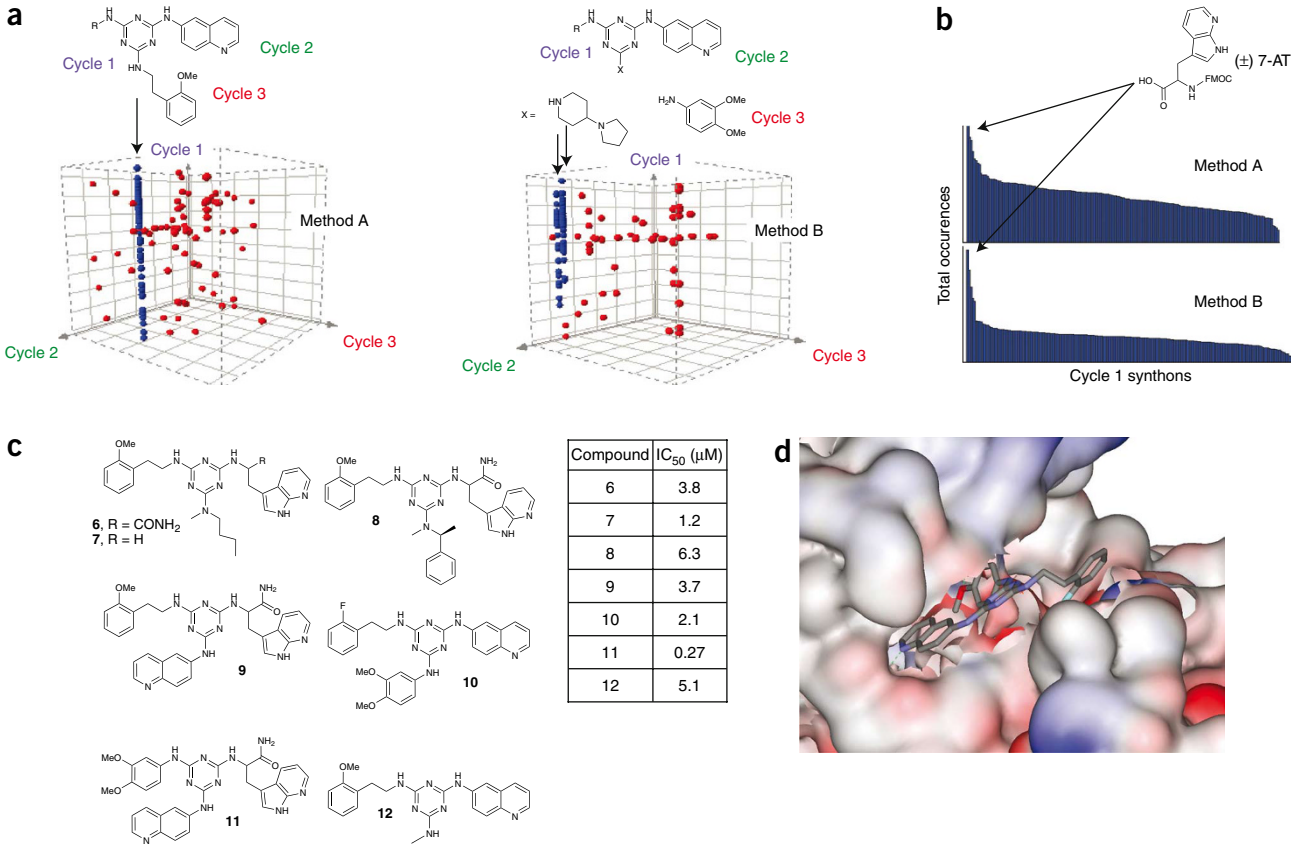


Figure 3 Selection of DEL-A against Aurora A kinase. (**a**) Selection output from DEL-A selections against Aurora A kinase. Low occurrence molecules (<4 copies) have been removed. Both selections show lines in the cycle 2 plane corresponding to 6-aminoquinoline. In method A, the line comprises 2-methoxyphenethylamine at cycle 3. In method B, there are two lines corresponding to 3,4-dimethoxyaniline and 4-pyrrolidinopiperidine at cycle 3. (**b**) Histograms showing the total occurrences of cycle 1 synthons after Aurora A selections. Using both methods, 7-AT is the highest occurring cycle 1 synthon. (**c**) Structures of synthesized compounds and their activity data. The compounds consist of members of both the 7-AT and the 6-aminoquinoline families. (**d**) Crystal structure of compound **10** bound to Aurora A kinase, showing the 2-fluorophenethylamine moiety exposed to solvent. Image created in Accelrys DS Viewer Pro 5.0.

A crystal structure of compound **10** bound to the ATP site of Aurora A kinase was obtained (**Fig. 3d**). The structure showed that the quinoline pointed toward the hinge region, with the ring nitrogen atom involved in a hydrogen bond with the amide N-H of Ala213. The ether oxygens of the dimethoxyaniline formed a bifurcated hydrogen bond with the side chain amine of the catalytic Lys162. Notably, poor electron density was observed for the 2-fluorophenethylamine moiety, which appeared to be disordered and exposed to solvent. The selection data showed little preference for the cycle 1 synthon, so it is not surprising that the crystal structure shows no strong interaction between this region of the inhibitor and the protein. In addition, cycle 1 was the site of the DNA attachment point in the library, so this position would need to provide a path to solvent for the oligonucleotide.

Selection of p38 MAP kinase inhibitors

Confident that the mechanics of synthesis, selection and deconvolution of DELs were operating correctly, we turned our attention to the 800-million-member DEL-B. When selected against p38 MAPK, DEL-B yielded several families, broadly defined by the identity of the amino acid of cycle 2. Since DEL-B contained four dimensions of variability, it could not be viewed in a single cube as was DEL-A. Instead, an array of cubes was needed. We chose to represent DEL-B as an array of 32 cubes, in which each cube corresponded to a different cycle 2 synthon, and the cube axes corresponded to cycles 1, 3 and 4.

We chose to focus on the family containing benzimidazole-5-carboxylate at cycle 2. Heterocycle-substituted benzimidazoles of this type had been shown previously to be key substructures of kinase inhibitors[23]. The selected structures corresponded to a conserved trisynthon, comprising a bicyclic tyrosine derivative at cycle 1, the benzimidazole at cycle 2, and tetrahydroquinoline at cycle 4 (see **Supplementary Dataset 6**). A representative compound from this family, **13**, was synthesized and shown to be a potent inhibitor of p38 MAPK (**Fig. 4**). The synthesis of compounds of this class was complicated by the formation of two regioisomers during the substitution of the triazine with the benzimidazole. In the case of **13**, these isomers could not be separated, and a mixture of isomers was submitted for bioassay. In other cases, the isomers proved separable (*vide infra*). Upon separation, we demonstrated that it was the 6-isomer that was the active component.

The selection showed no preference for the identity of the cycle 3 amine. We reasoned that this variability indicated that the region of cycle 3 was not interacting strongly with the target. We synthesized additional molecules in which the cycle 3 amine was removed from the structure completely. Both the ethoxy- (**14a** and **14b**) and the hydroxy- (**15a** and **15b**) substituted triazines showed very good

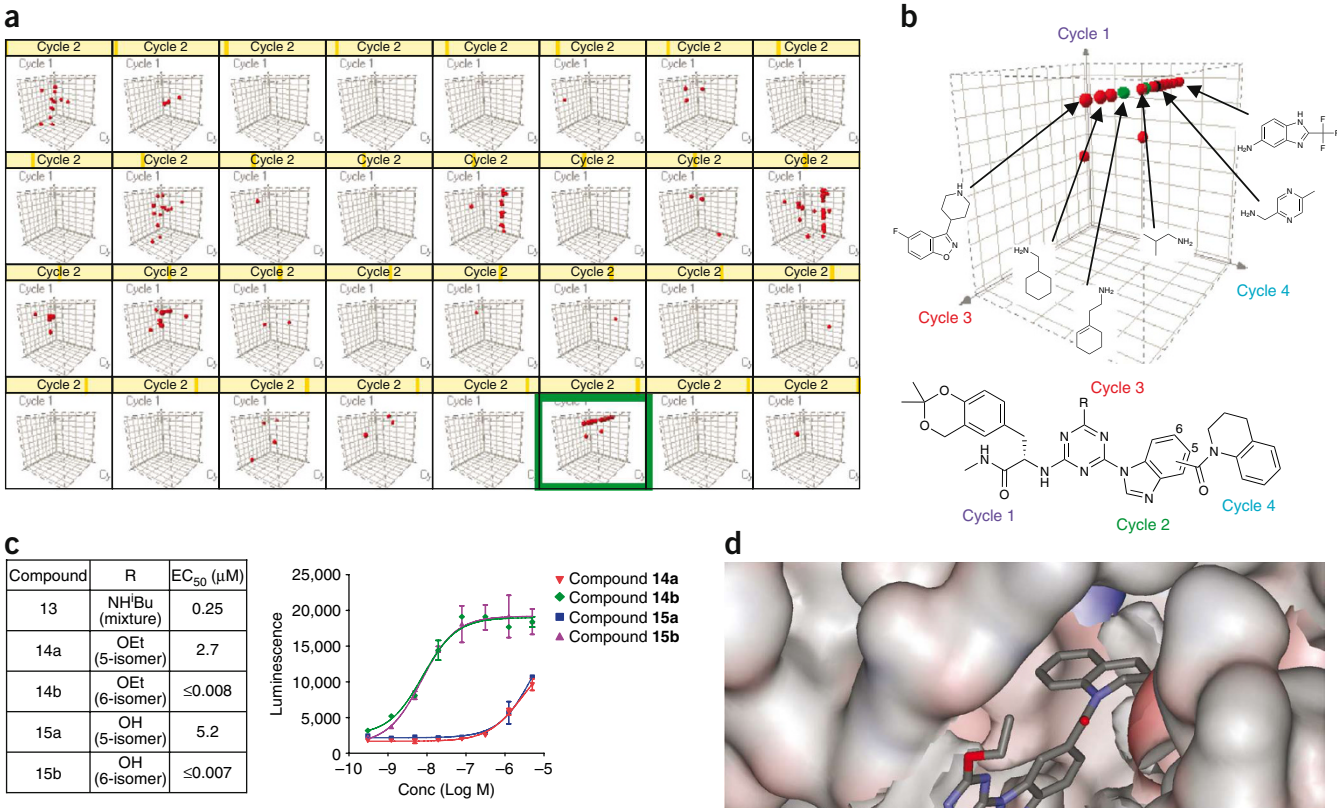


Figure 4 Selection of DEL-B against p38 MAPK. (**a**) The array of cubes shows the selection output from DEL-B against p38 MAPK. Molecules with <2 copies have been omitted. Each cube corresponds to a cycle 2 synthon. The cube axes represent cycles 1, 3 and 4. (**b**) The cube corresponding to benzimidazole-5-carboxylic acid is shown in detail. The coloring of the points indicates the copy number of the selected species (red = 2 copies, black = 3 copies, green = 4 copies). The line of points indicates a family of compounds all sharing the structure shown at bottom; several examples of selected cycle 3 synthons are shown. (**c**) Compounds corresponding to this family were synthesized and assayed against p38 MAPK. Error bars represent the s.e.m. of the data. (**d**) Crystal structure of compound **14b** bound to p38 MAP kinase. The DNA-linking carboxamide is exposed to solvent. Image created in Accelrys DS Viewer Pro 5.0.

potency. These structures are potential side products of the library synthesis, arising from hydrolysis during the substitution reaction or ethanolysis during DNA precipitation. We found that both these reactions are promoted by attachment of the poorly electron-donating benzimidazole to the triazine nucleus. We suggest that it is likely that such products are the actual species selected during the selection process. Thus the selection data demonstrated the key residues and allowed a highly active "tri-synthon" substructure to be identified and synthesized.

For this selection we obtained 67,959 reads encompassing 67,757 unique sequences. After filtering the sequences with <2 copies, 355 reads remained, encompassing 153 unique sequences. At 0.52%, this selection shows weaker signal than those conducted with DEL-A. However, since DEL-B is roughly 100 times larger than DEL-A, we hypothesized that such low signal was not necessarily indicative of a weaker binding event. Indeed, the potencies of the DEL-B hits are superior to those of the DEL-A Aurora inhibitors and comparable to those of the literature-inspired DEL-A p38 MAPK inhibitors.

In an effort to elucidate its binding mode, a crystal structure of **14b** complexed with p38 MAPK was obtained. X-ray diffraction revealed that **14b** bound to the ATP site of p38 MAPK, in a mode similar to other inhibitors. The tetrahydroquinoline moiety occupied the hydrophobic pocket adjacent to the ATP site. The 3-position nitrogen of the benzimidazole formed a hydrogen bond with the main-chain amide of Met109 in the hinge region. Notably, the bicyclic side chain of the cycle 1 amino acid was bent back into the pocket where the ribose moiety binds. As expected, the carboxamide group in the cycle 1 amino acid pointed toward solvent. This observation is consistent with the need for this position to accommodate a passage to solvent for the linker and the DNA.

DISCUSSION

To our knowledge, this work represents the first use of DNA encoding for the discovery of new small-molecule enzyme inhibitors. Although these DELs cannot be amplified and subjected to iterative affinity maturation, we have found that the degree of convergence achieved after three rounds of selection is sufficient to identify families of ligands. High-throughput sequencing of the enriched populations provides a broad picture of the chemical space of the binding families. Instead of obtaining a handful of unique structures, one sees families of related compounds. Observing logical chemical relationships among the binding population provides evidence that a structure-based binding event is occurring. These relationships provide a basis for SAR hypotheses that can be tested through conventional synthesis and assay.

In contrast to functional screening of discrete compounds, we have found that affinity selection of encoded mixtures is surprisingly tolerant of heterogeneity arising from alternative reaction outcomes during library synthesis. In the case of benzimidazole-containing inhibitors of p38 MAP kinase, it was possible that an alternative reaction product at cycle 3 gave rise to the actual selected species. It was unnecessary to deduce whether such an alternative pathway resulted in unchanged starting material, hydrolysis, ethanolysis or some other byproduct. Alternatively, all encoded building blocks at this position might have been tolerated in the binding interaction, perhaps because they were exposed to solvent. In either case, since the remainder of the molecule was specifically selected, we inferred that the building blocks at those cycles must have been crucial to target binding, and therefore incorporated during library synthesis. These inferences were used to design and assemble discrete molecules that demonstrated inhibitory activity and constituted the minimal pharmacophore for potent inhibition.

This reliance on familial trends to form structural hypotheses renders the enrichment of any particular compound less relevant than the total enrichment of the family. From our observations, copy number of an individual family member is not predictive of the potency of the synthesized compound. Again turning to the case of DEL-B against p38, if one accepts that cycle 3 failure gave rise to the selected species, it follows that all the points along the line in **Figure 4** define the same compound; yet the copy numbers of the points along that line vary from 1 to 4 copies.

We believe that this technique could revolutionize the discovery of small-molecule modulators of biological targets. In terms of cost, throughput and reliability, we believe that DNA-encoded chemical display libraries offer a real alternative to, and could surpass, current techniques such as high-throughput screening or molecular microarrays. A primary factor for the success of any molecular discovery technique is the chemical diversity of the accessible structures. Chemical diversity, in turn, depends on the repertoire of synthetic transformations that can be employed in library synthesis. Fortunately, an examination of the chemical literature reveals that the aqueous conditions required for synthesis of the DELs are amenable to many chemical reactions, including carbon-carbon bond-forming reactions[24,25]. A variety of useful synthetic transformations have been developed, optimized and employed in library synthesis. The results of these efforts will be reported at a future date.

We also believe that this technique may help answer questions around the nature of chemical and numerical diversity. Results with biopolymers seem to point to library size being the main determinant in the success of the selection process. As pointed out previously[12], whether this relationship is unique to biopolymers or whether it extends to all classes of chemical structure has been impossible to assess. Now, with the methods described here and elsewhere, it should be possible to study the roles of numerical diversity and "privilege" in ligand discovery.

In summary, we have devised and developed a chemical display technology for the synthesis and selection of encoded combinatorial libraries. The integrity of the encoded library has been validated through the synthesis and inhibitory activity of cognate molecules. DNA encoding allows the selection of hundreds of millions of discrete molecules in a single container with a minimal amount of protein. By using high-throughput sequencing techniques, the selection output can be deconvoluted to yield families of ligands and emerging trends of SAR and selectivity before biochemical assay. We suggest that the technology described in this report represents a significant advance in the ability to create and screen large numbers of small molecules.

METHODS

Preparation of library starting materials. See the **Supplementary Methods**. The procedures for the synthesis of DEL-A are described below.

DEL-A cycle 1. A 1 mM solution of the primer-elongated AOP headpiece (19.2 μmol, 19.2 ml) was split into 384 wells (50 nmol per well). To each well, we then added 20 μl of 10× ligation buffer, 2.0 μl T4 DNA ligase and 25 μl water. An aliquot of 1 of 384 tag solutions (100 μl of 1 mM stock solutions in water) was added to each well, and ligation was allowed to proceed at 16 °C for 16 h. After ligation, 5 M NaCl (10% by volume) and 2.5 volumes of cold ethanol were added to each vessel to precipitate the DNA. The DNA pellets recovered after centrifugation were each dissolved in 50 μl 150 mM borate pH 9.5. The plates were cooled to 4 °C, and to each well we added 40 equivalents of 1 of 192 Fmoc-protected amino acids (12.6 μl of a 150 mM DMF solution), followed by 40 equivalents of DMT-MM (7.6 μl of a 250 mM water solution). The acylation reactions were allowed to proceed for 18 h at 4 °C. After completion, the reactions were pooled, precipitated with ethanol and purified by reverse-phase HPLC, yield = 7.4 μmol. The lyophilized product was then deprotected by exposure to 20 ml of 10% (v/v) piperidine in water for 4 h. The deprotected product was precipitated. Since 384 tags but only 192 chemical building blocks were used, each building block was encoded by two different tags.

DEL-A cycle 2. The cycle 1 product (7.4 μmol) was dissolved in 7.4 ml water and split into 384 wells (19.2 μl per well). To each well we then added 7.8 μl of 10× ligation buffer, 0.77 μl T4 DNA ligase and 10.9 μl water. DNA tags were then added (38.4 μl of 1 mM stocks in water), and the ligations were allowed to proceed at 16 °C for 16 h. The DNA was precipitated as above, and pellets were dissolved in 19.2 μl per well of 150 mM borate pH 9.5. The library was cooled to 4 °C. To each well we added 10 equivalents of cyanuric chloride (1 μl of a 200 mM stock in acetonitrile). After 1 h, each well received 50 equivalents of an amine (4.8 μl of a 200 mM stock in acetonitrile or dimethylacetamide). The substitution was allowed to proceed for 16 h at 4 °C. The library was then pooled and precipitated, and taken into the next cycle without purification. A total of 192 amines and 384 tags were used, so that each synthon corresponded to 2 different DNA tags.

DEL-A cycle 3. The cycle 2 product (7.4 μmol) was dissolved in 7.4 ml water and split into 192 wells (38.4 μl per well). To each well we then added 15 μl of 10× ligation buffer, 1.5 μl T4 DNA ligase and 21 μl water. DNA tags were then added (95 μl of 1 mM stocks in water, 2.5 eq of tags), and the ligations were allowed to proceed at 16 °C for 16 h. The DNA was precipitated as above and dissolved in 38.4 μl per well of 150 mM borate pH 9.5. To each well we added 45 equivalents of amine (9 μl of a 200 mM stock in acetonitrile or dimethylacetamide). A total of 192 amines and 192 tags were used. The substitution was allowed to proceed for 6 h at 80 °C. The library was then pooled, precipitated and purified to give 3.9 μmol of product (19% final yield). This quantity is sufficient for 780 selection experiments at 5 nmol per selection.

Other methods. Library synthesis and analysis, the structure of the headpiece, positive control synthesis, selection protocols and crystallographic data are provided in the **Supplementary Methods**.

Accession codes. Protein Data Bank: Coordinates and structure factors have been deposited for Aurora A with **10** (accession code 3HA6) and p38 with **14b** (accession code 3HA8).

Note: Supplementary information and chemical compound information is available on the Nature Chemical Biology website.

Published online at http://www.nature.com/naturechemicalbiology/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Smith, G.P. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science* **228**, 1315–1317 (1985).
2. Barbas, C.F. III. Recent advances in phage display. *Curr. Opin. Biotechnol.* **4**, 526–530 (1993).

3. Wilson, D.S., Keefe, A.D. & Szostak, J.W. The use of mRNA display to select high-affinity protein-binding peptides. *Proc. Natl. Acad. Sci. USA* **98**, 3750–3755 (2001).
4. Frankel, A., Li, S., Starck, S.R. & Roberts, R.W. Unnatural RNA display libraries. *Curr. Opin. Struct. Biol.* **13**, 506–512 (2003).
5. Ellington, A.D. & Szostak, J.W. *In vitro* selection of RNA molecules that bind specific ligands. *Nature* **346**, 818–822 (1990).
6. Joyce, G.F. Amplification, mutation and selection of catalytic RNA. *Gene* **82**, 83–87 (1989).
7. Tuerk, C. & Gold, L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science* **249**, 505–510 (1990).
8. Brenner, S. & Lerner, R.A. Encoded combinatorial chemistry. *Proc. Natl. Acad. Sci. USA* **89**, 5381–5383 (1992).
9. Melkko, S., Dumelin, C.E., Scheuermann, J. & Neri, D. Lead discovery by DNA-encoded chemical libraries. *Drug Discov. Today* **12**, 465–471 (2007).
10. Li, X. & Liu, D.R. DNA-templated organic synthesis: nature's strategy for controlling chemical reactivity applied to synthetic molecules. *Angew. Chem. Int. Ed.* **43**, 4848–4870 (2004).
11. Halpin, D.R., Lee, J.A., Wrenn, S.J. & Harbury, P.B. DNA display. III. Solid-phase organic synthesis on unprotected DNA. *PLoS Biol.* **2**, 1031–1038 (2004).
12. Wrenn, S.J., Weisinger, R.W., Halpin, D.R. & Harbury, P.B. Synthetic ligands discovered by in vitro selection. *J. Am. Chem. Soc.* **129**, 13137–13143 (2007).
13. Melkko, S., Scheurmann, J., Dumelin, C.E. & Neri, D. Encoded self-assembling chemical libraries. *Nat. Biotechnol.* **22**, 568–574 (2004).
14. Buller, F. *et al.* Design and synthesis of a novel DNA-encoded chemical library using Diels-Alder cycloadditions. *Bioorg. Med. Chem. Lett.* **18**, 5926–5931 (2008).
15. Hansen, M.H. *et al.* A Yoctoliter-scale DNA reactor for small-molecule evolution. *J. Am. Chem. Soc.* **131**, 1322–1327 (2009).
16. Morgan, B. *et al.* Synthesis of combinatorial libraries containing encoding oligonucleotide tags. WO patent application 2007/053358 (2007).
17. Kinoshita, Y. & Nishigaki, K. Enzymic synthesis of code regions for encoded combinational chemistry (ECC). in *Nucleic Acids Symposium Series No. 34*, 201–202 (Oxford University Press, Oxford, 1995).
18. Leftheris, K. *et al.* The discovery of orally active triaminotriazine aniline amides as inhibitors of p38 MAP kinase. *J. Med. Chem.* **47**, 6283–6291 (2004).
19. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).
20. Mannoci, L. *et al.* High-throughput sequencing allows the identification of binding molecules isolated from DNA-encoded chemical libraries. *Proc. Natl. Acad. Sci. USA* **105**, 17670–17675 (2008).
21. Harrington, E. *et al.* VX-680, a potent and selective small-molecule inhibitor of the Aurora kinases, suppresses tumor growth in vivo. *Nat. Med.* **10**, 262–267 (2004).
22. Mortlock, A.A. *et al.* Progress in the development of selective inhibitors of Aurora Kinases. *Curr. Top. Med. Chem.* **5**, 807–821 (2005).
23. Burns, C.J., Wilks, A.F. & Bu, X. Preparation of pyrazine derivatives as kinase inhibitors. WO patent 2005/054230 (2005).
24. Western, E.C., Daft, J.R., Johnson, E.M., II, Gannett, P.M. & Shaughnessy, K.H. Efficient one-step Suzuki arylation of unprotected halonucleosides, using water-soluble palladium catalysts. *J. Org. Chem.* **68**, 6767–6774 (2003).
25. Gartner, Z.J., Kanan, M.W. & Liu, D.R. Expanding the reaction scope of DNA-templated synthesis. *Angew. Chem. Int. Ed.* **41**, 1796–1800 (2002).

Erratum: Design, synthesis and selection of DNA-encoded small-molecule libraries

Matthew A Clark, Raksha A Acharya, Christopher C Arico-Muendel, Svetlana L Belyanskaya, Dennis R Benjamin, Neil R Carlson, Paolo A Centrella, Cynthia H Chiu, Steffen P Creaser, John W Cuozzo, Christopher P Davie, Yun Ding, G Joseph Franklin, Kurt D Franzen, Malcolm L Gefter, Steven P Hale, Nils J V Hansen, David I Israel, Jinwei Jiang, Malcolm J Kavarana, Michael S Kelley, Christopher S Kollmann, Fan Li, Kenneth Lind, Sibongile Mataruse, Patricia F Medeiros, Jeffrey A Messer, Paul Myers, Heather O'Keefe, Matthew C Oliff, Cecil E Rise, Alexander L Satz, Steven R Skinner, Jennifer L Svendsen, Lujia Tang, Kurt van Vloten, Richard W Wagner, Gang Yao, Baoguang Zhao & Barry A Morgan

Nat. Chem. Biol. **5, 647–654 (2009); published online 2 August 2009; corrected after print 17 September 2009**

In the version of this article initially published, the IC_{50} values in the table in Figure 3c were listed as nM instead of μM. The error has been corrected in the HTML and PDF versions of the article.