Supporting Information for "Quantitative Comparison of Enrichment from DNA-Encoded Chemical Library Selections"

John C. Faver, [†] Kevin Riehle, [†, ‡] David R. Lancia, Jr., [§] Jared B. J. Milbank, [§, ∆] Christopher S. Kollmann, [§] Nicholas Simmons, [†] Zhifeng Yu, [†] and Martin M. Matzuk[†, ¤]

[†]Center for Drug Discovery and Department of Pathology and Immunology, [‡]Bioinformatics Research Laboratory, [¤]Departments of Molecular and Cellular Biology, Molecular and Human Genetics, and Pharmacology and Chemical Biology Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, United States, [§]FORMA Therapeutics Inc, 500 Arsenal Street, Suite 100, Watertown, MA 02472, United States.

Table of Contents

S1. Nomenclature for *n-synthons*

Consider a hypothetical combinatorial library with 3 cycles of split-and-pool chemistry, where each cycle adds 1,000 unique sequence/building block pairs. When analyzing data from a selection, any combination of encoding sequences might be found to be enriched, which would correspond to different combinations of building blocks promoting binding to the target. These combinations of building blocks are often called *n-synthons*, where *n* is the number of cycles in the combination. Each of these *n-synthons* can be evaluated for enrichment in selection data separately by aggregating count data grouped by the different combinations of synthetic cycles in the library. The full set of synthon types in the example 10^3 x 10^3 x 10^3 library are listed in Table S1. When plotting selection data in a standard 3D scatter plot (or "cubic view"), a plane feature in the cubic view represents a single conserved building block from one of the three cycles of this library and hence can be called a *1-synthon* or *mono-synthon*. The example library contains 10^3 different *mono-synthons* in each of three axes (i.e., 10^3 per cycle). Every *mono-synthon* represents a single building block which is a substructure of 10^6 unique molecular structures in the library. *Di-synthons* and *tri-synthons* similarly represent higher dimensional groupings of 2 and 3 cycles of building blocks, respectively. These correspond to lines and points when plotted in the cubic view. Finally, *n-synthons* for the highest dimension *n* within a library (e.g., *tri-synthon* for a 3-cycle library) are sometimes referred to as *singletons*, because they each represent only one molecular superstructure.

Cubic View Representation	Synthon Type	Dimension	Feature Axes[a]	Synthons per axis[b]	Compounds per synthon[c]
Plane	*Mono-synthon*	1	(1, 0, 0) (0, 1, 0) (0, 0, 1)	10^3	10^6
Line	*Di-synthon*	2	(1, 1, 0) (1, 0, 1) (0, 1, 1)	10^6	10^3
Point	*Tri-synthon* *Singleton*	3	(1, 1, 1)	10^9	1

Table S1. Nomenclature for different types of combinatorial features in 3-cycle DNA-encoded libraries. Within a 3-cycle combinatorial library, any combination of chemical building blocks may be found to be important for binding affinity to a target. This leads to seven different feature axes belonging to three different synthon types from which specific features can be evaluated for enrichment. [a]Feature axes are described using a notation indicating the inclusion (1) or exclusion (0) of a cycle in a feature. [b]The total number of synthons per feature axis in a 3-cycle library wherein each cycle contains 10^3 building blocks. [c]The number of unique library molecules represented by a single synthon in the example 3-cycle library.

S2. Comparison of Enrichment Metrics

S2.1. Candidate Enrichment Metrics

We evaluated the set of enrichment metrics in Table S2 for their ability to meet the enumerated criteria for a successful enrichment metric. The *z-score* metric is the often-utilized measure of difference from the mean count in units of standard deviations in counts. The normalized *z-score* metric is similar, but it is normalized by the square root of the number of samples. The *Count_ratio* is the difference from observed counts to the expected count in units of the expected count, *CBV_ratio* is the ratio of the observed count to the Bonferroni-corrected 95% significance critical value from a fitted binomial distribution, l*ogE* is the logarithm of the ratio of observed to expected population fractions, and *Cohen's h* uses the arcsine transformation and is a variance-stabilizing metric for differences in proportions.

Name	Formula	Comments
z-score	$z = \frac{C - E}{\sigma}$	σ from binomial distribution
"normalized" *z-score*	$z_n = \frac{z}{\sqrt{n}}$	Normalizes for sample size
Count_ratio	$C_{rat} = \frac{C - E}{E}$	Trivial interpretation, Blows up when E << 1
CBV_ratio	$CBV_{\text{rat}} = {}^{C}/_{E_{cbv}}$	Signal-to-noise ratio Computationally expensive
logE	$logE = \log\left(\frac{P_o}{P_i}\right)$	Trivial interpretation
Cohen's h	$h = 2\left[\arcsin\left(\sqrt{p_o}\right) - \arcsin\left(\sqrt{p_i}\right)\right]$	Variance stabilizing function for proportions

Table S2. List of evaluated enrichment metrics. For a given synthon using the above metrics, C is the observed count, E is the expected (mean) count, n is the total number of molecules

sampled, σ is the standard deviation in counts from a binomial distribution model, E_{cbv} is the Bonferonni-corrected 95% critical value from a binomial distribution with n and p_i, p_o is the observed population fraction and p_i is the expected population fraction.

Each of the candidate metrics involve comparing observed versus expected populations, and here expected populations were evaluated by using only the diversity of each *n-synthon*. In other words, all synthetic yields were assumed to be equal and therefore features with equal diversity are equally probable to be chosen in a random selection. Where expected counts and standard deviations are required, we modeled the data with binomial distributions. The binomial distribution was utilized because its probability mass function yields the probability of observing *k* counts given a fixed selection probability p_i and the total number of observations *n*. For small expected counts, the binomial distribution closely resembles the Poisson distribution, and for higher expected counts, it closely resembles the normal distribution. Thus, the binomial distribution can simultaneously model the wide range of selection probabilities for *mono-*, *di-*, and *tri-synthon* features in DEL selection data. The enrichment metrics in Table S2 were evaluated under four different scenarios: naïve sequencing, non-enrichment, target-specific enrichment, and variable sampling.

S2.2. Example Scenarios

S2.2.1. Naive Sequencing

Before a library is screened in a selection against a target, it is important to verify that the distribution of members in the unscreened library is reasonably close to the expected distribution. For this reason, it is common to screen each library after synthesis in its unselected, or "naïve", form. Generally, it is expected that counts should follow close to a binomial distribution for n equal to the number of molecules decoded and p_i equal to the probability of random selection. Small errors during synthesis, amplification, and sequencing typically produce deviations from the expected binomial distribution, but in our experience these deviations are much smaller than typical perturbations due to affinity selection. Small random perturbations in count distributions are therefore acceptable in practice. Figure S1 shows the observed enrichment for each *n-synthon* in the sequencing of a naïve library. This 3-cycle library has a size of around 229 million members, and 99 million sequences were read during DNA sequencing. Of these, 81 million sequences represented valid barcodes, and after UMI filtering (see Supporting Information S3), the final data set included just under 15 million sampled library molecules. In this scenario, the expected count for a unique *singleton* in the library is below 1, at 0.065. This implies that every molecule observed is present at a minimum of 15 times the expected count. Since observed molecules in a naïve library sample are presumably observed only due to random selection, successful enrichment metrics should account for this random selection noise, especially for high-diversity features when the expected count is very low.

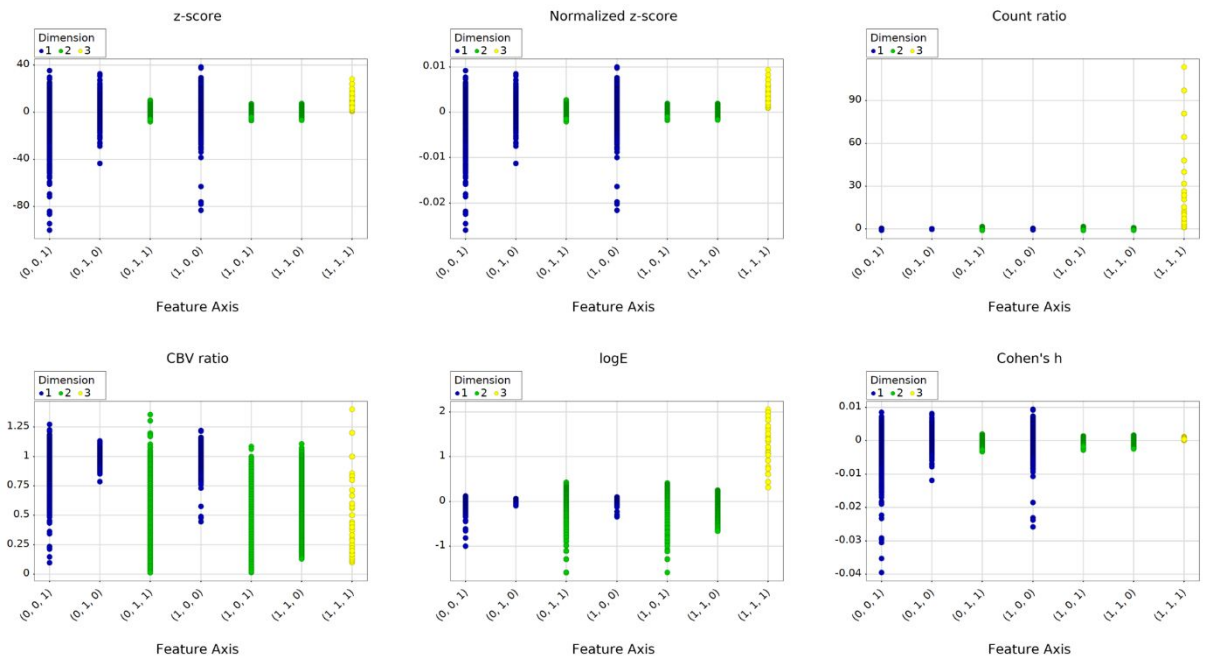

Figure S1. Evaluated enrichment by six different enrichment metrics for each synthon type in a naïve library sample. These data represent the library population distributions before being perturbed by selection with a target. The enrichment metrics are a) *z-score*, b) normalized *z-score*, c) *Count_ratio*, d) *CBV_ratio*, e) *logE*, and f) *Cohen's h.*

In this naïve data set, it was observed that for the *z-score* metrics, enrichment is usually centered close to zero for each synthon type, i.e., the observed populations are, on average, close to the expected populations. The exception is for the (1, 1, 1) feature axis (also known as *singletons* or *tri-synthons*), because both the expected population and sampling ratio are low enough that an observed count of 1 is measured as being significant enrichment (an observed count of 1 is much greater than the expected count of 0.065). This effect is also prominent in the *logE* and *Count_ratio* metrics, where the evaluated enrichment for *singletons* is much larger than that of other *n-synthons*. These two metrics do not evaluate enrichment for different *n-synthon* types

with different values of expected population (i.e., diversity) on the same order of magnitude. On the other hand, the *z-score* metrics evaluate the enrichment of *singletons* to be of the same order of magnitude as other *n-synthons*. The *CBV_ratio* metric shows only a few features with a value greater than 1, which is interpreted as the value above which enrichment would be considered statistically significant at the 95% confidence level. *Cohen's h* tends to tighten the distribution of enrichment values for high diversity features which have low expected probabilities compared to lower-diversity features.

S2.2.2. Non-enrichment.

An ideal enrichment metric should make it easy for the analyst to determine a lack of significant enrichment of features in a library. To investigate how the candidate metrics perform in the non-enrichment scenario, we examined a data set for which we believe that the library contains no binders with significant affinity for the protein target. We evaluated the metrics for the target-selection pool and the NTC pool and plotted the enrichment values of each feature against each other in Figure S2. In the non-enrichment scenario, an ideal metric would be expected to yield 1) low measures of enrichment and 2) similar enrichment for both target and NTC data sets with a small amount of additional random noise. In Figure S2, the *z-score* shows generally higher measured enrichment for the NTC data than the target data, while the normalized *z-score* more closely follows the diagonal line. This is consistent with our observation of 206,220 molecules for the NTC sample compared to 118,446 for the target sample. Thus, the unnormalized *z-score* can be skewed by the number of decoded molecules, while the normalized *z-score* is less sensitive to the amount of sampling. *CBV_ratio* similarly is affected by differences in sampling, while *Cohen's h* evaluates the two data sets to be more equal in enrichment. l*ogE* and

Count_ratio again are very sensitive to expected population and show large differences between synthon types in terms of magnitudes of enrichment.

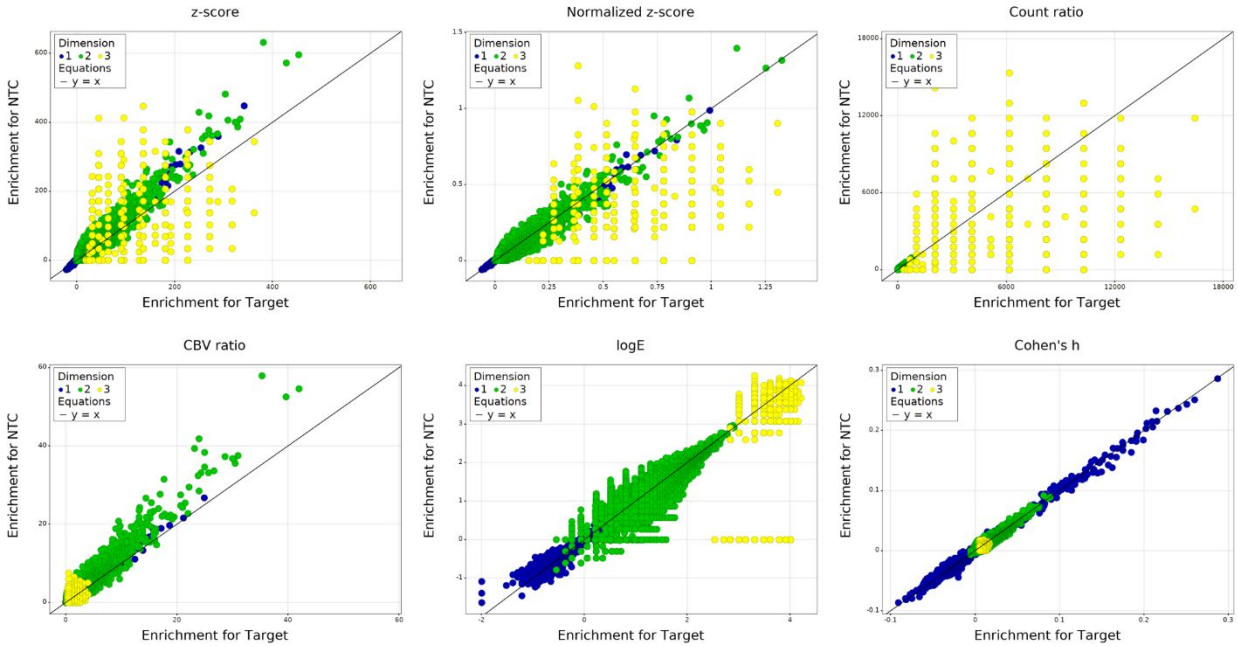

Figure S2. Comparative enrichment plots for a selection with no significant target-specific enrichment. For each enrichment metric, enrichment for the NTC is plotted against the enrichment for the target selection data set. Enrichment is evaluated for each *n-synthon* within the library, and the points are colored by synthon type (dimension). The diagonal y = x line represents equal enrichment in the target and NTC data. The enrichment metrics are a) *z-score*, b) normalized *z-score*, c) *Count_ratio*, d) *CBV_ratio*, e) *logE*, and f) *Cohen's h.*

S2.2.3. Enrichment

The main goal of any DEL analysis strategy is to enable straightforward detection of target-specific enrichment of any *n-synthons* of a library. We evaluated the performance of the

enrichment metrics using a data set for which there was specific enrichment of a family of molecular structures with affinity for a protein target. In Figure S3, *logE* shows little separation between highly-enriched and lowly-enriched features, causing interpretation to be nontrivial. *Count_ratio* treats high-diversity features very differently than low-diversity features, measuring their enrichment with very different magnitudes and thereby precluding simple and simultaneous analysis of all *n-synthons*. On the other hand, the remaining four metrics very clearly distinguish the family of highly-enriched *n-synthons* from the rest of the library. *Cohen's h* tended to give higher enrichment values for low-diversity features like *mono-synthons*, while *z-score*, normalized *z-score*, and *CBV_ratio* show higher enrichment for higher diversity *di-* and *tri-synthons*.

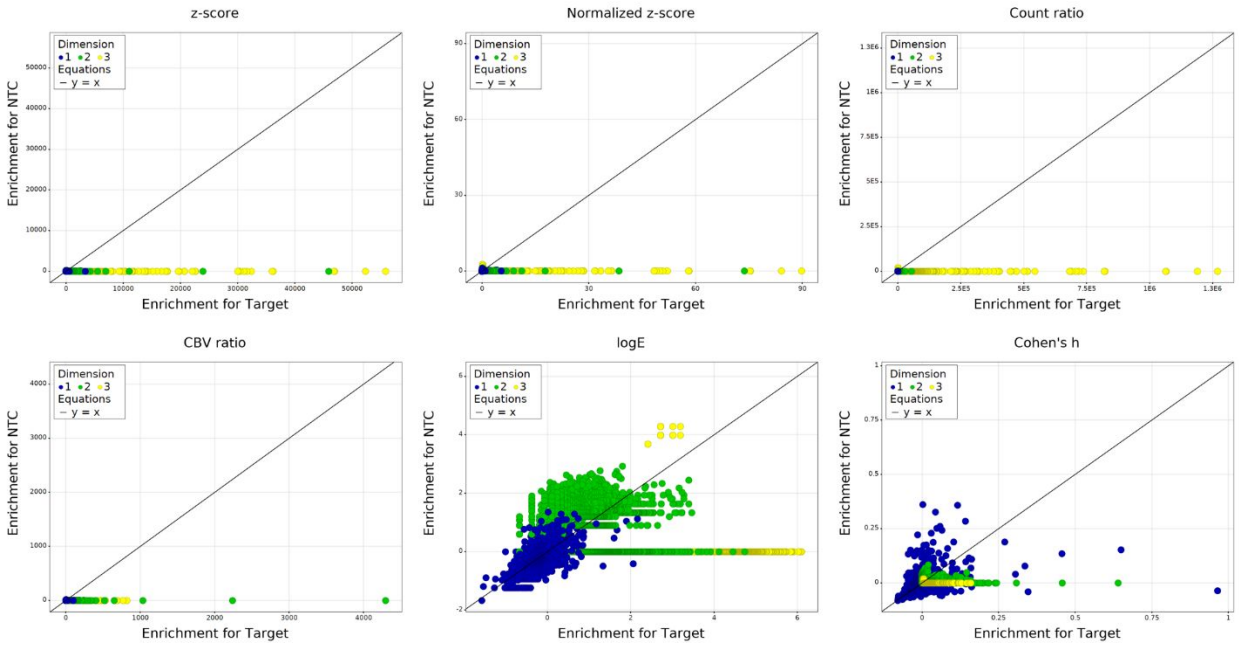

Figure S3. Comparative enrichment plots for a selection with significant target-specific enrichment. For each enrichment metric, enrichment for the Non-Target Control (NTC) is plotted against the enrichment for the target selection data set. Enrichment is evaluated for each

n-synthon within the library, and the points are colored by synthon type (dimension). The diagonal y = x line represents equal enrichment in the target and NTC data. The enrichment metrics are a) *z-score*, b) normalized *z-score*, c) *Count_ratio*, d) *CBV_ratio*, e) *logE*, and f) *Cohen's h*.

S2.2.4. Variable Sampling

One of the most important requirements for a useful enrichment metric is an insensitivity to sampling. Sampling insensitivity is required to compare enrichment in multiple selection experiments against each other without being affected by sampling bias. If a sampling bias is present in the enrichment metric, then it would be difficult to determine if a feature is enriched due to target-specific binding rather than the sampling bias. To examine this property, we evaluated the metrics on two data sets: one experimental data set with target-specific enrichment, and the same dataset with 90% of the decoded ligands randomly removed from the data. Thus, the two samples represent the same selection experiment but with a ten-fold difference in sampling. The two data sets are compared in Figure S4. The *z-score* and *CBV_ratio* metrics clearly show bias for the higher-sampled data set, while the normalized *z-score*, *Count_ratio*, and *Cohen's h* appear to be insensitive to the ten-fold difference in sampling. In comparison, l*ogE* shows much larger deviations between the two data sets, especially for lower-enriched features.

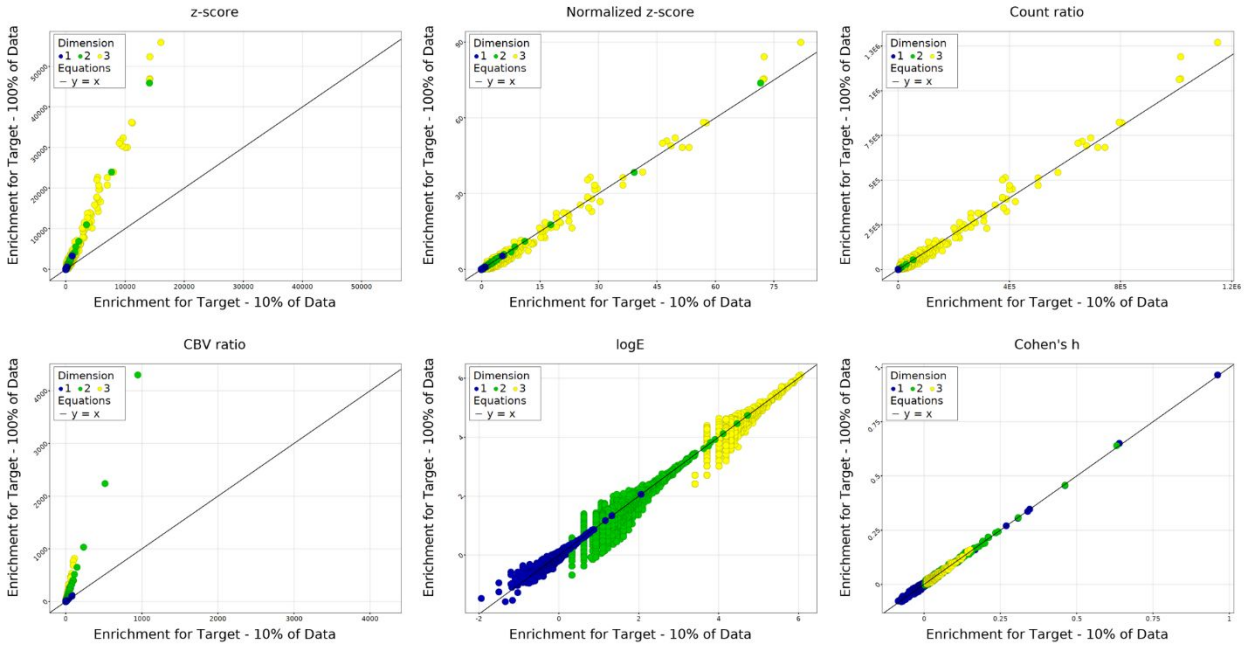

Figure S4. Comparative enrichment plots for a selection with significant target-specific enrichment and a subsampled version of the same data set. For each enrichment metric, enrichment for the full target data set is plotted against the enrichment for the same data set with 90% of decoded molecules randomly removed. Enrichment is evaluated for each *n-synthon* within the library, and the points are colored by synthon type (dimension). The diagonal y = x line represents equal enrichment at the two levels of sampling. The enrichment metrics are a) *z-score*, b) normalized *z-score*, c) *Count_ratio*, d) *CBV_ratio*, e) *logE*, and f) *Cohen's h*.

S3. Counting Unique Molecules

S3.1. Introduction

Accurately counting molecules from a sampled DEL pool is a critical precursor to data analysis. Errors in the form of insertions, deletions, and substitutions from PCR-amplification or DNA sequencing introduce uncertainties in the counting of each library member in a sequenced sample. Addressing this issue for the codon regions of DNA barcodes is straightforward: before DEL synthesis, one can preselect sets of codon sequences which have a specified minimum edit distance (Levenshtein distance)[1] between any pair. This method of codon selection decreases the probability of miscalling one building block for another in the presence of sequencing errors. It additionally allows for more accurate copy counting in the presence of read errors. For example, by using codon sets with a minimum edit distance of 3 between any pair, observed codon sequences which differ by an edit distance of 1 can be safely considered as the same encoding sequence.

A different strategy is needed to address errors in sequencing and decoding degenerate UMI (Unique Molecular Identifier) regions of DNA barcodes. The purpose of the UMI is to uniquely label individual molecules of a DEL before PCR-amplification of the DNA barcodes.[2] By including randomized UMI sequences, identical DNA barcodes read after PCR-amplification can be inferred as having come from the same original library molecule, while barcodes which differ only in the UMI region can be identified as distinct library molecules which represent the same small molecular structure. Thus, when counting observations for analysis, multiple DNA sequences which are observed to be identical including the UMI should be counted only once (as they are PCR copies of the same library molecule), while multiple sequences which are identical

except for the UMI should be counted as distinct library molecules. Without error corrections during analysis, sequence perturbations and read errors in UMIs have the effect of artificially increasing library member counts, as UMI sequences which were once identical become distinct UMIs with mismatches.[3] Given an estimated 0.1% substitution rate for the Illumina HiSeq as an example, one should expect an average of 1 per 100 incorrect UMI sequences in a barcode with UMI length of 10 (assuming no additional sources of error).[4] In the best-case scenario, these errors would affect all library members equally and thus lower the final signal-to-noise ratio. In the worst-case, these errors might alter the observed count distributions if there is any bias in the sequences which are altered or misread. The presence of errors in UMI sequences can be observed in uncorrected data (where a simple unique UMI counter is used), as the PCR copy count per library molecule distribution often has a significant spike at 1. This effect is demonstrated in Figure S5, left, where the copy count per molecule distribution from using the simple unique counter on a data set from a non-target control (NTC; i.e., a selection in which no target protein is included) is shown in red.

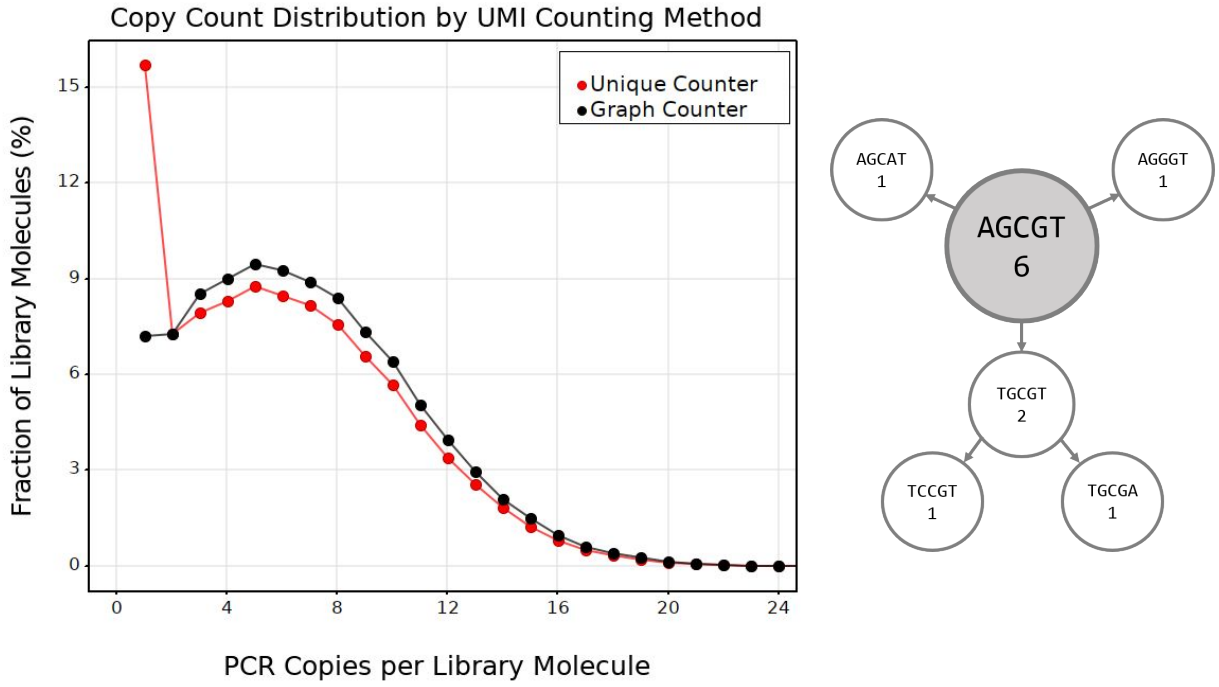

Figure S5. Directed graph-based strategy to count unique molecules while addressing sequencing errors in DNA barcodes. **Left)** Copy count distribution (number of degenerate copies per library molecule) from an NTC data set using the unique UMI counting method (red) and the directed graph-based method (black). Due to the presence of errors in the observed sequences, the unique UMI counting method yields a significant number of single UMI sequences per ligand, which produces a discontinuity in the distribution at a copy count of one. This in turn has the effect of artificially inflating count data from DEL selections. The graph-based counter shows a much smoother distribution of degenerate UMI counts per molecule. **Right)** Example of the directed graph-based counting scheme. A set of unique molecular identifier sequences is arranged into a directed graph structure, in which each node represents a unique sequence weighted by the number of observations, and the edges represent the condition of the edit distance and count difference between nodes meeting set thresholds (e.g., edit distance <=2, and

count ratio >=2). The directions of the edges represent likely perturbations in the sequences. Rather than counting each unique UMI sequence as unique molecules from a DEL sample pool, each completed graph is considered to be one unique molecule for which the DNA barcode has been subject to sequencing errors.

S3.2. Graph-based UMI counting

To address sequence errors in the UMI region, we have adopted the strategy of Smith *et al.* which involves a directed graph-based counting scheme (Figure S5, Right).[5] The method assumes that all sampled DNA barcodes have been decoded and that each unique decoded library member is associated with a set of UMI sequences. For each decoded library member, the set of associated UMIs and their populations is read, and the most populated UMI is set as the root node of a directed graph. UMI sequences which are within a set edit distance D and meet a set count ratio threshold R are added as child nodes. Remaining UMIs are likewise added to the graph as children of the root node or its descendants. This process is repeated until no more UMI sequences can be added to the graph. The next highest populated UMI which has not yet been inserted into a graph is then assigned to be the root node of a new graph, and then children are similarly added to this new graph. The process continues until each UMI sequence is assigned a position in one of the directed graphs. Each completed graph then represents a single unique library molecule, wherein the root node is interpreted as having been the original UMI sequence, and its child nodes are interpreted as products of errors during PCR-amplification or sequencing. We have observed that this directed graph-based counting strategy generally improves agreement between theoretical and observed count distributions in naïve (unscreened) library data sets and

removes significant discontinuities in copy count per molecule distributions (Figure S5, Left, black).

S3.3. Triazine DEL Naïve data set

The naive sequencing of the 174,145,836-member triazine DEL provides an excellent example of the effect of errors in UMI sequences. In Table S3, various molecule counting schemes are compared to the theoretical count distribution based on a binomial distribution model. Since each counting method results in a different total number of decoded molecules, n, the counting schemes must be compared to different evaluations of the binomial probability mass function with different values of n. We first compared the unique counter (U) to the binomial distribution model (U*). We found that 512 library members had observed counts above the expected noise level of 4. We additionally examined the performance of the graph-based methods, labeled as G(D, R) where G represents a graph built with an edit distance parameter D and a count ratio parameter R. Thus G(2, 2) corresponds to a graph built with child nodes added with edit distance less than or equal to 2 and count ratio of parent count to child count greater than or equal to 2. The corresponding binomial count distributions are provided in the table and labeled as G(D, R)*. We observed that for this data set, the G(2, 1) model was able to generate count data which matched expected noise levels from a binomial distribution model.

k	U	U*	G(2, 2)	G(2, 2)*	G(2, 1)	G(2, 1)*	G(3, 1)	G(3, 1)*
1	5,080,555	6,450,921	5,554,544	5,953,857	5,818,097	5,684,023	5,820,547	5,681,712
2	687,794	124,171	269,890	105,448	30,192	95,947	27,765	95,867
3	73,296	1,593	22,281	1,245	233	1,080	210	1,078
4	6,368	15	1,696	11	3	9	3	9
5	473	0	104	0	0	0	0	0
6	36	0	9	0	0	0	0	0
7	3	0	1	0	0	0	0	0
n	6,704,105	6,704,102	6,168,532	6,168,532	5,879,192	5,879,193	5,876,719	5,876,716
R^2		0.98688		0.99909		0.99987		0.99986

Table S3. Comparison of unique molecule counting schemes. Counting unique molecules in a sequenced DEL sample is affected by errors in the degenerate unique molecular identifier (UMI) sequence. The strict unique counter method (U) has a count distribution which overcounted some library members and thus library members were observed with higher counts (k) than the expected random noise level from a binomial distribution model (U*). Using the graph-based G(2, 2) counter alleviated these high counts somewhat, but observed counts did not meet expected noise levels unless the G(2, 1) or G(3, 1) counters were used. By using such directed-graph based approaches, errors in UMI sequences can be addressed and unique molecules can more accurately be counted.

S3.4. References

(1) Navarro, G. A Guided Tour to Approximate String Matching. *ACM Computing Surveys* **2001**, *33* (1), 31–88.

(2) Fu, G. K.; Hu, J.; Wang, P.-H.; Fodor, S. P. A. Counting Individual DNA Molecules by the Stochastic Attachment of Diverse Labels. *Proc. Natl. Acad. Sci.* **2011**, *108* (22), 9026–9031.

(3) Islam, S.; Zeisel, A.; Joost, S.; Manno, G. L.; Zajac, P.; Kasper, M.; Lönnerberg, P.; Linnarsson, S. Quantitative Single-Cell RNA-Seq with Unique Molecular Identifiers. *Nat. Methods* **2013**, *11* (2), 163–166.

(4) Goodwin, S.; McPherson, J. D.; McCombie, W. R. Coming of Age: Ten Years of next-Generation Sequencing Technologies. *Nat. Rev. Genet.* **2016**, *17* (6), 333–351.

(5) Smith, T.; Heger, A.; Sudbery, I. UMI-Tools: Modeling Sequencing Errors in Unique Molecular Identifiers to Improve Quantification Accuracy. *Genome Res.* **2017**, *27* (3), 491–499.

S4. Synthesis of DELs

S4.1. General Information

S4.1.1. Materials for the synthesis of DNA-encoded libraries.

DTSU ("DEC-Tec Starting Unit") **1** (Figure S6) and 5'-phosphorylated oligonucleotides were obtained from LGC Biosearch Technologies and assessed for purity through the general analytical procedure for DNA oligonucleotides; oligonucleotide sequences were designed on principles designed to maximize sequence-reads (discussed within the main text of the manuscript). A special "spike-in" 10-mer DNA oligomer functionalized with a primary amine and cholesterol tag was obtained from Sigma to monitor chemical steps during post-pooling manipulations (as the greasy oligo has a very different chromatographic retention time). T4 DNA ligase was obtained from Enzymatics (Qiagen) and the activity was experimentally determined through test DNA oligomer ligations. Chemical building blocks and reagents were sourced from a variety of suppliers and aliquots of building blocks were stored in acetonitrile or mixed aqueous acetonitrile solutions in Tracetraq barcoded tubes (Biosero) with either screw- or septa-caps. Barcoded tubes were read using a SampleScan 96 scanner (BiomicroLab) and decoded using Vortex software (Dotmatics). All buffers, including HEPES 10X ligation buffer (300 mM 2-[4-(2-hydroxyethyl)piperazin-1-yl]ethanesulfonic acid, 100 mM $MgCl_2$, 100 mM dithiothreitol, 10 mM adenosine triphosphate, pH 7.8) and basic borate buffer (250 mM sodium borate/boric acid, pH 9.5), were prepared in-house. Library working solutions were prepared using DNAse free ultra-pure water (Invitrogen), HPLC-grade acetonitrile (Fisher) or high-purity absolute ethanol (Koptec). LC/MS running solvents were made from Optima LC/MS grade water (Fisher), Optima LC/MS grade methanol (Fisher), 99+% purity hexafluoroisopropanol (Sigma)

and HPLC-grade triethylamine (Fisher). Solutions were generally transferred or pooled utilizing Biotix brand pipette tips and reservoirs (various sizes), reactions were generally performed in polypropylene, 96-well, deep-well plates (USA Scientific, various sizes), plates were sealed for incubation with AlumaSeal II foil seals (Excel Scientific) and large volume DNA precipitations were performed in polypropylene 250 mL screw-cap bottles (from various vendors). Heated reactions were either performed in ep384 Mastercyclers (Eppendorf) or in laboratory ovens (Fisher). Solutions were centrifuged in either Avanti J-30I or Allegra X-15R centrifuges (Beckman-Coulter). Optical density measurements were made using a Biophotometer (Eppendorf).

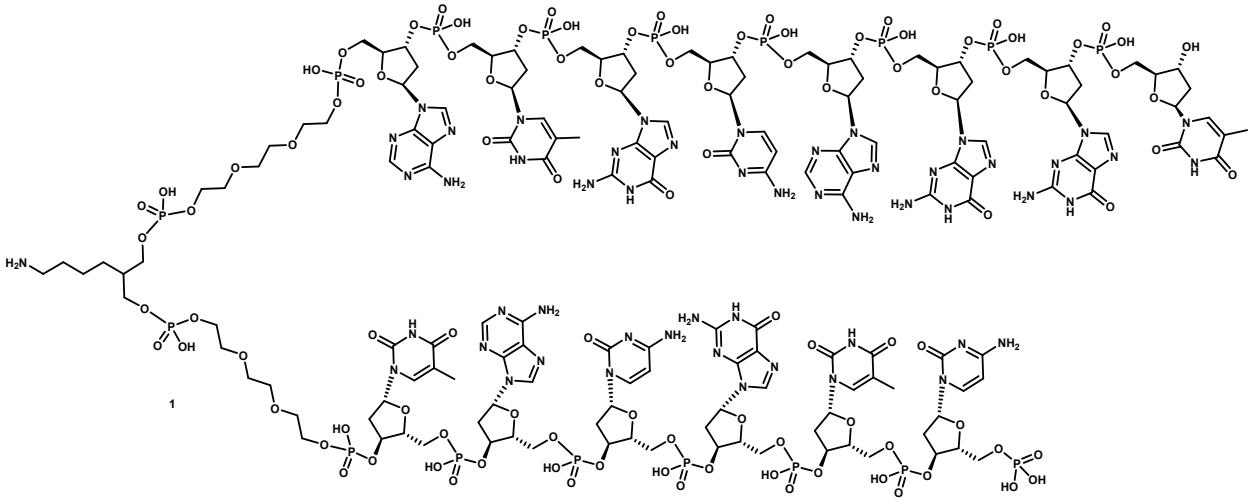

Figure S6. Structure of "DTSU" **1** (5'-Phos-CTGCAT-Spacer 9-Amino C7-Spacer 9 ATGCAGGT 3').

S4.1.2. General analytical procedure for the analysis of DNA oligonucleotide compositions.

A Vanquish UHPLC system was integrated with LTQ XL ion trap mass spectrometer (ThermoFisher Scientific) for LC/MS analysis of oligonucleotides. Injection amounts were typically 5–10 µL containing 50–200 pmol DNA analyte.

LC/MS Parameters for Thermo Vanquish UHPLC with LTQ Ion Trap MS Instrument

(i) LC settings

Column: Thermo DNAPac RP (2.1 x 50 mm, 4µm)
Solvent A: 15mM triethylamine (TEA)/100mM hexafluoroisopropanol (HFIP) in water
Solvent B: 15mM TEA/100mM HFIP in 50% methanol
Solvent C: Methanol
Flow rate: 0.65 mL/min
Run time: 2 mins (gradient)
Column temperature: 100 °C (post column cooler at 40 °C)

(ii) MS settings

Source: ESI in negative mode
Spray voltage: 4100 V
Source heater temperature: 390 °C
Sheath Gas: 28 (instrument units)
Auxiliary Gas: 8 (instrument units)
Sweep Gas: 2 (instrument units)
Capillary temperature: 350 °C
Capillary voltage: -33.0 V
Tube lens: -92.0 V
MS Scan: 500 – 2000 *m/z*

Samples were analyzed on a Thermo Vanquish UHPLC system coupled to an electrospray LTQ ion trap mass spectrometer. An ion-pairing mobile phase comprising of 15mM TEA/100mM HFIP in a water/methanol solvent system was used in conjunction with an oligonucleotide column Thermo DNAPac RP (2.1 x 50 mm, 4µm) for all the separations. All mass spectra were acquired in the full scan negative-ion mode over the mass range 500–2000m/z. The data analysis was performed by exporting the raw instrument data (.RAW) to an automated biomolecule deconvolution and reporting software (ProMass) which uses a novel algorithm known as ZNova

to produce artifact-free mass spectra. The following deconvolution parameters were applied: peak width 3.0, merge width 0.2, minimum and normalize scores of 2.0 and 1.0 respectively. The noise threshold was set at S/N 2.0. The processed data was directly exported to Microsoft Excel worksheets for further data comparisons. A sample MS analysis using ProMass software is presented in Figure S7.

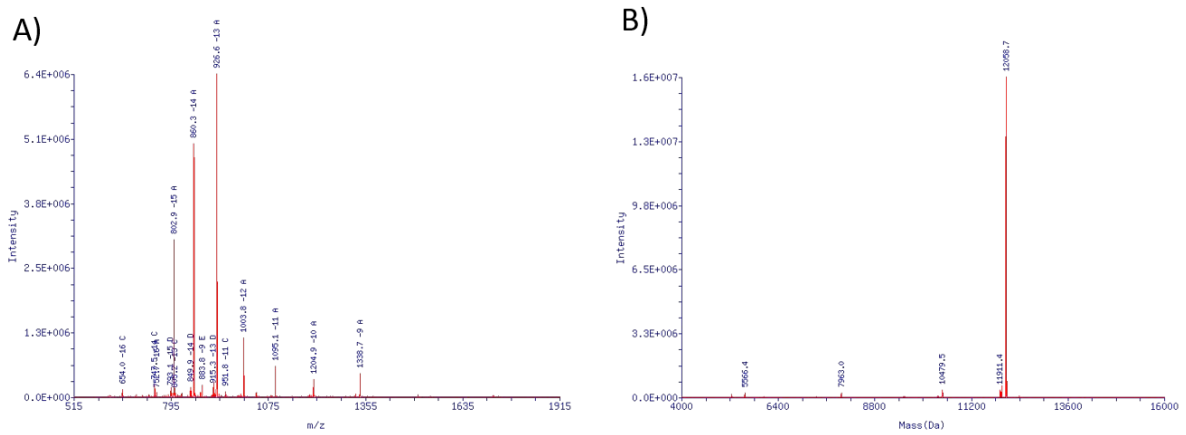

Figure S7. Representative analysis of oligonucleotide MS data on the HP ("Headpiece") **2**. A) The crude MS data showing the various m/z ions observed in the 500–2000 mass region; B) The deconvoluted spectrum showing the parent ion mass (12059, the expected molecular weight of **2**).

S4.1.3. General procedure for the isolation of DNA library material from aqueous solutions (ethanol precipitation)

Based on the theoretical solution volume *n* (ignoring any loss from heating, etc.), *n*/20 volume of a 5 M NaCl stock solution was added and the solution was gently mixed. Then absolute ethanol (3*n* volume, 75% v/v final ethanol concentration) was added, the solution was thoroughly mixed, and then stored at -20 °C overnight to precipitate the DNA. The resulting slurry was centrifuged (10,000 × G for 1 h), the supernatant decanted, an addition 2*n* volume of chilled 75% ethanol (v/v) was added, and the pellet was centrifuged again (10,000 × G for 30 min). After decantation

of the supernatant, the pellet was dried (in open air or under gentle vacuum) and reconstituted in neutral water or buffer (to a concentration of ~1 mM; assessed by optical density measurements). The solution was then centrifuged (10,000 × G for 10 min) to pellet any left-over solids (unremoved chemical building blocks or byproducts, denatured ligase, etc.), and the solution was transferred to leave these solids behind. The DNA may undergo a second round of precipitation if the purity is insufficient (as assessed by the general analytical procedure). In addition, if the initial solution contains high amounts of organic co-solvent or chaotropic reagents (e.g., piperidine), the solution may be diluted with neutral water (by n or $2n$) to enhance the overall precipitation yield. Typically, precipitations were conducted in polypropylene 96-well plates or polypropylene bottles which can withstand high centrifugal speeds. However, polypropylene is incompatible with piperidine—reactions with this reagent were run in fluorinated ethylene propylene (FEP) bottles and spun with a maximum speed of 4,000 × G.

S4.1.4. General procedure for the ligation of DNA oligonucleotides

To a ~1 mM solution of the HP-containing library intermediate (1 equiv), a premixed solution of the duplex oligonucleotide ("codon") with the appropriate overhang was added (1 mM stock soln in neutral water, 1.05-1.1 equiv). Separately, a master mix consisting of additional water, HEPES 10X ligation buffer, and T4 DNA ligase was prepared and added to the wells or container with mixing and incubated at room temperature overnight. The concentration of the HP-contain library intermediate in the final solution was 0.24 mM (thus the amount of HEPES 10X ligation buffer was 1/10th of this final volume). The amount of T4 DNA ligase stock added depended on the assayed activity of the ligase batch—however we routinely observe activities greater than 200X (i.e., full ligation observed with the addition of ligase stock 1/200th overall volume). After

the overnight incubation, the ligation progress was assessed by LC/MS with the general analytical procedure (due to the large MW increase, the ligation is obvious even on pooled post cycle 1 samples) as well by gel electrophoresis. If incomplete, additional buffer, ligase or codon may be added. Typically, ligation samples were run on a denaturing 6% TBE-Urea gel (Invitrogen), in TBE buffer at 180–200 V for 30–40 min. Gels were stained with ethidium bromide, visualized with a Gel Doc (Bio rad) or equivalent imager, and assessed for transformation into a new, higher-MW band (similar to the analysis applied to thin-layer chromatography). A typical gel result is shown in Figure S8.

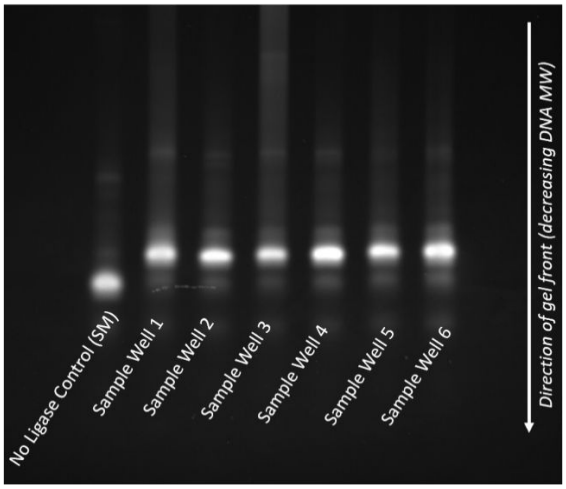

Figure S8. Representative 6% TBE-Urea gel for the analysis of ligation of DNA codons. As shown here, the disappearance of the no ligase control well's band (the starting material) to a higher-MW band signifies a finished codon ligation.

S4.1.5. General architecture of the Main Build of the triazine DEL

The DNA encoded library described in this research article is a three-cycle library (three encoded chemistry steps). DTSU **1** is elaborated on one end with a linker on which the small molecule portion is iteratively built up, and duplexed pairs of DNA oligonucleotides ("codons") are ligated

on the other end. The DTSU is extended to a longer region that will act as a primer in post-selection amplification (DTSU + first overhang + forward primer unit), followed by three 11-bp regions used to encode each of the three chemical transformations. In between these regions are complimentary 2-bp overhangs used to ensure efficient and selective annealing/codon ligation. These three cycles represent the Main Build of the library—the DNA material may be further ligated/elaborated for use in selection experiments.

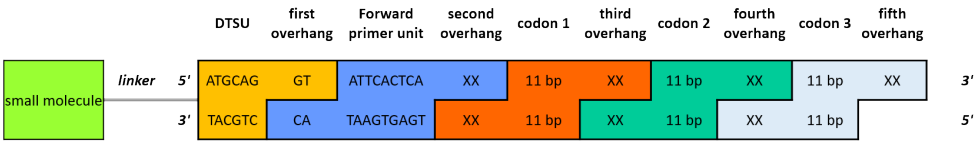

Figure S9. Sequence architecture of the Main Build of the triazine DEL.

S4.2. Synthesis of the triazine DEL

S4.2.1. Overall synthetic sequence of the triazine DEL

The triazine DEL was constructed in three cycles around a triazine scaffold. First, DTSU **1** was modified with a PEG-based linker, followed by the large scale ligation of an 11-bp DNA duplex needed for a primer region to form the library starting material **2**. Cycle 1 consisted of codon 1 ligation/nucleophilic substitution into cyanuric chloride/nucleophilic substitution of amines and amino acids/pooling, cycle 2 consisted of nucleophilic substitution of amines/codon 2 ligation/pooling, and cycle 3 consisted of codon 3 ligation/"reverse" acylation of amines/pooling. Although generally ligation before a chemical step is the preferred order within a cycle (to avoid inhibition of ligase activity from residual chemical building blocks), the order was reversed in

cycle 2 to avoid quenching the electrophilic, substituted triazine intermediate. This synthetic sequence is depicted in Figure S10.

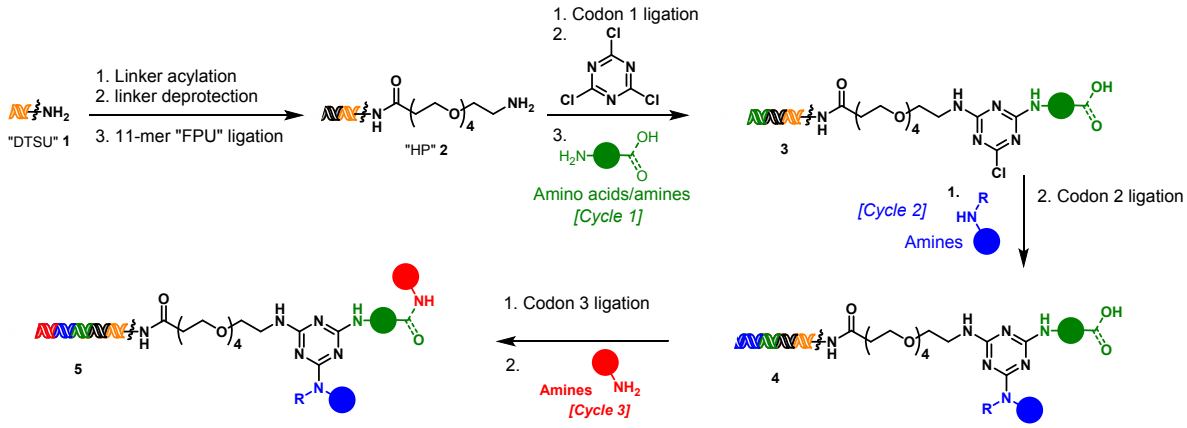

Figure S10. Synthetic sequence of the triazine DEL's Main Build.

S4.2.2. Representative preparation of "HP" 2.

To four 250-mL fluorinated ethylene propylene centrifuge bottles each charged with DTSU **1** (25 µmol in 6.43 mL water, 1 equiv, 3.89 mM), aqueous sodium borate buffer (6250 µmol, 25 mL of a 250 mM aq. soln, 250 equiv, pH 9.5), CH_3CN (6 mL) and a solution of Fmoc-15-amino-4,7,10,13-tetraoxapentadecanoic acid (1250 µmol, 2.5 mL of a 500 mM soln in CH_3CN, 50 equiv) was added. After brief mixing, 4-4,6-dimethoxy-1,3,5-triazin-2-yl)-4-methylmorpholinium chloride ("DMTMM", 1250 µmol, 2.5 mL of a 500 mM soln in H2O, 50 equiv) was added and the soln was incubated at 25 °C for 2 h. With verification of reaction completion by LC/MS (by formation of expected product, MW = 5461), the DNA was isolated by the general procedure. To the reconstituted DNA solutions (~1 mM), an aq. soln of piperidine (5 mL, 10% piperidine v/v) was added. After incubation at 25 °C for 3 h, the full deprotection of *N*-Fmoc was verified by LC/MS (formation of expected product, MW = 5240). After DNA

isolation by the general procedure, the pellets were reconstituted in H_2O (10 mL) and the solutions combined (total volume of 49.8 mL after extra transfer washes). Optical density measurements of this soln indicated an approximate concentration of 1.82 mM (90.6 μmol, 90.6% yield). This material was then ligated with two 5'Phos 11-mer DNA oligomers ("forward primer unit") and isolated by the listed general procedures to provide the fully elongated HP **2** in near quantitative yield from the linker intermediate. This material was used without further purification.

S4.2.3. Procedure for Cycle 1

After portioning 1,040 wells on 11 plates with 100 nmol of HP **2**, each well was ligated with codon 1 by the general ligation procedure, followed by ethanol precipitation by the general procedure. The pellets (~100 nmol, 1 equiv) were reconstituted in water (100 μL) and 250 mM pH 9.5 Borate buffer (100 μL, 25,000 nmol, 250 equiv). After cooling to 4 °C, cyanuric chloride (25 μL, 40 mM in CH_3CN, 1000 nmol, 10 equiv) was added, and the wells were monitored for triazine addition by LC/MS. After complete addition, a collection of 1,040 amines and amino acids were added to individual wells (25 μL, 200 mM in CH_3CN/water, 5000 nmol, 50 equiv) and the reactions were left overnight at 4 °C. After analysis of all wells and controls by LC/MS, the DNA was precipitated by the general procedure. After reconstitution, the wells were quickly pooled and precipitated by the general procedure. In addition, a separate control of a small amount of the library pool with a triazine-functionalized "spike-in" oligo was monitored by LC/MS to ensure that residual contaminants were not reacting with the on-DNA diamino-chloro-triazine intermediates (no significant reaction was detected). After reconstitution, an estimated

yield of cycle 1 was determined by OD measurement of the pooled solution (*est*. 74.5 µmol, 1.57 mM, 47.5 mL, 74.5% yield).

S4.2.4. Procedure for Cycle 2

The cycle 1 pool was distributed to 1010 wells in 11 plates (47 µL, 73.8 nmol, 1 equiv) and 250 mM pH 9.5 Borate buffer (73.8 µL, 18450 nmol, 250 equiv) and CH_3CN (23.3 µL) were added. Then 1008 amines (18.45 µL, 200 mM in CH_3CN/water, 3690 nmol, 50 equiv) were added to individual wells and the plates were heated to 80 °C for 6 h (two wells received no amine as an encoded control). In addition, a series of non-library, parallel controls that mimicked library reaction substrates and conditions were included on several empty plate wells, and some library wells were augmented with a triazine-functionalized "spike in" control oligo (8 nmol) to monitor the post-pool transformations by LC/MS. After analysis and positive confirmation of all control data, each well underwent ethanol precipitation by the general procedure. After reconstitution and ligation of codon 2 by the general procedure, all cycle 2 wells were pooled and precipitated by the general procedure. An OD measurement of the reconstituted cycle 2 pool indicated near quantitative recovery (*est.* 74.500 µmol, 124 mL, 0.56 mM, quant).

S4.2.5. Procedure for Cycle 3

The cycle 2 library pool was distributed to 1040 wells (71 nmol, 127.7 µL, 1 equiv) and the cycle 3 codon was ligated by the general procedure. After reconstitution in water (75 µL), 250 mM pH 9.5 borate buffer (71 µL, 17750 nmol, 250 equiv) and CH_3CN (40 µL) were added. Then 1040 amines (28.4 µL, 200 mM in water/CH_3CN, 5680 nmol, 80 equiv) were added to individual wells, followed by 4-4,6-dimethoxy-1,3,5-triazin-2-yl)-4-methylmorpholinium

chloride (“DMTMM”, 56.8 µL, 20448 nmol, 360 mM soln in water, 288 equiv) and the plates were incubated at 30 °C overnight. As in cycle 2, a series of non-library parallel control wells were set up and some wells were augmented with an appropriately functionalized triazine “spike in” control oligo (8 nmol). After analysis of all control well data, the library wells were precipitated by the general procedure, reconstituted, pooled, and again precipitated by the general procedure. The final yield of the cycle 3 pool was estimated by OD (62.99 µmol, 81.8 mL, 0.77 mM, 84.5 % yield), although a secondary measurement by comparing the intensity of the 3-cycle library band to a known standard (Low Molecular Weight DNA Ladder from New England Biolabs) on a native TBE gel suggested a recovery of ~52 µmol.

S4.2.6. Preparation of amplifiable triazine DEL samples (“shots”) for selection experiments

On small scale (1–20 nmol of completed library), the triazine DEL material was ligated with two DNA oligonucleotides containing a DNA segment encoding the library design, a segment encoding the experimental usage, a degenerate segment serving as the UMI region, a segment increasing sequencing diversity and a terminal primer segment to allow PCR amplification. After ethanol precipitation and reconstitution, the amount of amplifiable library material within prepared shots was subsequently quantified by qPCR and shots were used without further purification. Alternatively, portions of the library were ligated on large scale (1–5 µmol) with a duplexed 12-bp DNA codon to encode the library design followed by small scale ligation (1–20 nmol) of the remaining regions needed for the amplifiable shot.