Logarithmic molecular sampling for next-generation sequencing

Caroline Horn[1] and Julia Salzman[1,2]
[1]Department of Biochemistry and [2] Department of Data Science, Stanford University, Stanford, CA 94304

Abstract

Next-generation sequencing enables measurement of chemical and biological signals at high throughput and falling cost. Conventional sequencing requires increasing sampling depth to improve signal to noise discrimination, a costly procedure that is also impossible when biological material is limiting. We introduce a new general sampling theory, Molecular Entropy encodinG (MEG), which uses biophysical principles to functionally encode molecular abundance before sampling. SeQUential DepletIon and enriCHment (SQUICH) is a specific example of MEG that, in theory and simulation, enables sampling at a logarithmic or better rate to achieve the same precision as attained with conventional sequencing. In proof-of-principle experiments, SQUICH reduces sequencing depth by a factor of 10. MEG is a general solution to a fundamental problem in molecular sampling and enables a new generation of efficient, precise molecular measurement at logarithmic or better sampling depth.

Introduction

Measurement of molecular abundance is fundamental to basic science and to medicine; it is a classical statistical problem formalized as the problem of estimating the multiplicities n_i of elements s_i in a so-called multiset, and traditionally solved with simple random sampling (SRS). Next-generation sequencing enables SRS of molecules to be performed at high throughput, but in many applications, current or projected throughput is insufficient for addressing critical problems in medicine (e.g. next-generation biomarker detection), chemistry (e.g. high throughput compound screens) and biology (e.g. single-cell sequencing); it is insufficient because SRS suffers from intrinsic limitations when: (i) the cardinality of the multiset is comparable or large compared to the number of measurements taken; (ii) large discrepancies exist between the n_i; (iii) or when precise detection of small changes between the n_i is required. A variety of molecular technologies have attempted to address inefficiencies, including targeted or semi-unbiased enrichment or depletion of a population of molecules (Boone, De Koker, Callewaert, 2018; Hubank & Schatz, 1994). However, these technologies are only semi-quantitative as they compromise quantification of a set of sequences subject to the depletion and require the depleted or enriched sequences to be prespecified.

To overcome these problems, we sought to design an efficient measurement platform with theoretically tractable principles that could be realized in experiment. This led us to develop Molecular Entropy encodinG (MEG), a new general paradigm for molecular sampling that uses computations performed by molecular ensembles to encode the abundance of each species in a sample before measurement, predicts that one can enable experiment-specific sampling designs that only require log or sub-log scale sampling depth compared to SRS while achieving the same measurement precision.

To illustrate the general theory, we start with a simple stylized example (Example 1) of SeQUential DepletIon and enriCHment (SQUICH), a special case of MEG. Consider a tube containing *10* cubes and 10^4 spheres (Fig. 1); these quantities are unknown to the experimenter who wants to estimate them. Suppose there is a physical process that for any number *n* allows up to, but no more than, *n* shapes of each type to be drawn from the tube. SQUICH uses this procedure as follows: in the first round of SQUICH, one object of each shape (cube and sphere) is captured from the tube, tagged with a ``1'', and placed into a container we call the

sampling box. Then, up to 9 of each shape are captured and destroyed. In the second round of SQUICH, 1 (or 0) of each shape is captured and tagged with a ``2'' and added to the sampling box. In this example, no cubes are captured, as they have all been removed. Then, up to 89 of each shape are captured and destroyed. The numbers of captured and destroyed molecules in the second round satisfy the property that up to and including this round, 10^2 ($1 + 9 + 1 + 89 = 100$) of each shape has been destroyed or captured. In the third round, 1 (or 0) of each shape is captured, tagged with a “3”, and up to *899* of each shape are destroyed; in round 4, one more shape is captured and up to 8999 are destroyed. At the end of 4 rounds, there are 4 cubes and 1 sphere in the sampling box and they are each labeled with the number of the round they were captured in. The abundance of each shape in the original tube can now be estimated with *5* samples without replacement; while SRS requires on the order of 10^2 to make the same inference.

SQUICH is much more general than as presented in Example 1; MEG is yet more general than SQUICH (Methods and Supplement). The numbers in Ex. 1 were arbitrary: the same procedure can operate on, e.g., 10^{15} cubes, allowing a much larger savings in sampling and demonstrating the intuition for why SQUICH enables logarithmic sampling depth compared to SRS.

Informally, three properties enable sampling reductions by SQUICH in Example 1: (1) tagging and removal operating independently on each shape; (2) limiting the number of each shape that is tagged and depleted in each round; (3) sampling only tagged shapes.

The critical properties (1-3) above are fulfilled with nucleic acids replacing the objects of different shapes. Each “shape” in Ex. 1 is replaced by a unique DNA sequence called a target. For each target, sets of DNA oligonucleotides called encoders and competitors that each hybridize with targets are the key to SQUICH. Encoders are libraries of reverse complements of all possible targets which have three critical regions: (1) a region of reverse complementarity to the target; (2) a DNA sequence that is a DNA code representing the round in which the encoder was added to the original tube and; (3) a PCR handle that allows sampling of only targets that extend on encoders. Competitors have the same region of reverse complementarity as encoders, but lack a PCR handle. In each round, targets are hybridized with competitors and encoders, and after hybridization, extend on competitors and encoders (in which case the target is said to be ‘‘coded’’) which serve to tag and pull targets into the sampling box as the physical device did for shapes in Example 1 (Figure S4).

As in Example 1, competitors and encoders are added in limiting amounts (n) at each step so that removal and/or tagging of no more than n of each sequence type can occur in each step. To ensure only coded molecules are sampled, PCR is used as an AND logic gate to selectively sample molecules that are targets AND have extended on encoders (Methods). Competitors can be designed so that targets extending on them can be later retrieved. If targets are in excess of encoders and competitors, the number of targets that extend is limited by the available encoder and competitors. When encoders and competitors are in excess of targets, they compete for binding, which enables estimation of the first significant figures in scientific notation (Results). In addition, the abundance of each competitor and encoder can vary by target as may be desired in certain applications; for example if an experimenter seeks to measure spheres only if they are more abundant than 10^4 copies, 10^4 competitors for spheres could be added in the first round.

Results

SQUICH is simple to embody in experiment and provably enables logarithmic or even sub-logarithmic sampling compared to SRS for precision desired in ubiquitous sequencing applications including estimation of scientific figures formalized in this: **Claim** (Logarithmic sampling with non-filtered round coding): Suppose the abundance of two species are respectively $x_1 10^{y_1}$ and $x_2 10^{y_2}$ with $y_1 < y_2$ and $x_1, x_2 \in \mathbb{Z}^+$ and $0 < p < 1$ fixed.

There is a SQUICH procedure such that $((y_2+1)/(y_1))\log(1/p)$ samples suffice to achieve a probability of detection of at least p; a standard result shows SRS requires at least 10^{y2-y1} samples to detect the second species which implies the sampling depth required by SQUICH is logarithmic compared to SRS. See Supplement for proof.

THe proof of the claim shows how SQUICH can achieve more general sampling reductions such as sub-logarithmic rates with super-geometric increases in the number of competitors per round. Simulation tests of SQUICH performance are given in three common application regimes: (1) detection of rare species in the presence of a large background; (2) small fold changes in a complex population; (3) quantification of each species in a population with high dynamic ranges.

To conservatively model SQUICH performance in simulation, we introduce a set of engineered DNA target sequences (which we term "CGA libraries" of length n) consisting of any molecule matching the format [(C/G)A)^n]; CGA libraries are targets behave like the shapes did in Example 1. Competitors and encoders for CGA libraries consist of all reverse compliments of the CGA library with auxiliary sequences that identify them as competitors or encoders (described above, Methods, Tables S4,S5). Equilibrium thermodynamics of CGA are modeled in simulations to include inefficiencies and mismatches in oligonucleotide hybridization when the minimum edit distance between targets is one (Wang & Zhang, 2015; Zhang, Chen, Yin, 2012)(Methods). SQUICH can perform more favorably than in our simulation when targets have minimum edit distance of four or more, a design achieved with sphere packing theory (Conway & Sloane, 2011); that is, CGA codes are a convenient way to explain, model and experimentally embody SQUICH, but SQUICH performance is optimized by different designs of targets. For example, experiments in this paper were performed with oligonucleotides containing degenerate bases (Methods).

Simulation 1 models the “needle in haystack” problem with two species at abundance 10^x where $x=15$, with *20* “needle” species at abundance *100*. As predicted by theory, SQUICH robustly identifies all needles across all trials with less than *2000* samples (Fig. 2a). SRS detects at most one needle with 10^{11} samples in 1000 simulations and requires more than 10^{15} samples for the same recall as achieved by SQUICH with 2000 samples (Supplement and basic statistical theory). For $x = 15$, this implies SQUICH reduces sampling depth by a factor of 10^{11}; Table S1, Supplement for other values of x.

Simulation 2 tests SQUICH performance where a subset of species (here 20) are 2-fold more abundant than a background of complexity >*260,000*, modeling complexity needed to detect duplication events of > ~10kb in the human genome, or a 2-fold enrichment in a chemical or high throughput pooled CRISPR screen. We designed a statistical estimator for SQUICH to identify species enriched above background (Methods). With this estimator, 10^5 samples suffice for recovering a median of *18* of the *20* enriched species with 0 false positives (FP) across 1000 trials. To achieve a zero FP rate, SRS requires 10^8 samples, requiring at least 10^3-fold higher sampling depth than SQUICH (Fig 2b). This simulation demonstrates that SQUICH provides flexibility by designing encoder and competitor abundance increases over rounds. Varying competitor and encoder abundance tunes false positive and false negative rates separately, overcoming an intrinsic limitation of SRS where false positives are functionally related to false negatives as a function of sampling depth.

To simulate measurement of native molecules such as RNA or microbial DNA with high dynamic ranges, we modeled each RNA species as a specific CGA code. The molecular biological procedure for converting RNA, DNA or protein to CGA codes is straightforward (Fig. S3; Supplement). In Simulation 3, we test the performance of SQUICH when the distribution of sampled species fills a high dynamic range ($x10^y$ for x=1,..,9 and y=0,...,9), as arises in measurement of protein and environmental microbial DNA (Fig. S3). SQUICH fails to detect only 428 of more than 5000 species at a sampling depth by 10^5 (Fig. 2c); SRS has a drop-out of 3706

species. The log MSE log (Table S3) for SQUICH is lower than SRS at depths up to and including 10^{10}. SQUICH performance with 10^6 samples also exceeds SRS at depth 10^{11}; at this depth for SQUICH, a mean of only 4 molecules dropout of sampling.

We also simulated sequencing of a single cells with a dynamic range of transcripts with ~4000 transcripts expressed, and ~1500 transcripts significantly above basal expression which we set to 100 (n=2236), including expression of 10 transcripts at each value $x\ 10^y$ for x=1,..,9 and y=0,...,4, and 100 additional transcripts at each level 1:10. With 10^5 samples, SQUICH has comparable performance to SRS with 10^7 samples as measured by drop-out (~*2%)* and log MSE (Fig S6; Supplement) a significant improvement over high dropout rates in single-cell sequencing (Vallejos, Risso, Scialdone, Dudoit, Marioni, 2017); SRS at 10^5 samples has a drop-out rate of roughly 50%, evidence that SQUICH could significantly improve transcript detection in massive throughput single-cell sequencing. In summary, simulation shows that SQUICH exceeds performance of SRS by 100-1000 or more fold in diverse problems including detection of expression of rare species, small fold changes and quantifying species at high dynamic ranges.

SQUICH, as modeled in simulation, can be directly applied to primary biological samples whenever an orthogonal barcode is introduced into the sample, e.g. pooled chemical or genetic screens, with gains in sampling precision illustrated above. To test SQUICH in real next-generation sequencing experiments, we designed a synthetic target library of complexity 2^{18} = 262,144, similar to the CGA code set and manually added a set of individual species ranging from $81x$ to $80{,}000x$ fold over background (Methods, Table S4-5). SQUICH was carried out with 10-fold increases in total molecules in each round, low encoder amounts in the first round and constant encoder amounts in rounds 2-6 (Methods). Six SQUICH libraries prepared with two levels of encoder in round one were sequenced to a mean depth of 2187 reads. Six conventional libraries that model SRS with experimental error introduced during library preparation were sequenced to a mean depth of 19759. In all SQUICH replicates, Pearson and rank correlation between ground truth and estimated abundance exceed all replicates of conventional libraries, despite SQUICH libraries being sequenced at >9 -fold lower depth (Methods, Fig. 3, Tables S4-6).

To control for the high leverage of species with high abundance on correlation values, we used a conservative measure of performance of SQUICH vs. conventional sequencing using a rank based method (Supplement, Methods). 3 of the 6 SQUICH runs sequenced exceeded performance of all 6 conventional sequencing experiments. In addition, 5 out of 6 replicates were statistically significantly more sensitive than SRS with no noise introduced during sequencing (p<.05 in 5 out of 6 replicates; p=0.138 in one replicate, labeled CH52.03, Table S3). No p-values were significant for conventional experiments. We developed a statistical method to control for variable sampling depths in SQUICH and conventional libraries, and used it to estimate SQUICH efficiency compared to conventional sequencing; it also estimated that proof-of-principle SQUICH experiments achieves a 10x reduction in sequencing depth compared to conventional sequencing.

Discussion

MEG, and a specific example, SQUICH, is a new framework for quantifying each of a large number (millions or more) species of molecules in a pool, one of the most ubiquitous and important molecular measurement problems today. MEG theory can be applied to any molecular sampling problem, though here we focus on DNA. Small molecules, proteins and RNA can be tagged with DNA sequences, so common assays and screens all reduce to SQUICH, and more generally MEG, measurement. In applications where the sample is limiting, such as biomedical testing, increasing sampling depth is impossible, as sample amplification introduces extra sources of measurement error. In these areas, MEG may be especially important. The flexibility of the sampling distribution provided by MEG expands the scope of statistical algorithms that can be

used for estimation. Further, MEG provides key advantages when integrated with modern statistical approaches that use assumptions of sparsity to both improve precision in signal detection and reduce resource cost.

For example, SQUICH could be an ideal platform to measure massive single-cell RNA profiles. To illustrate the design of SQUICH for single-cell RNA-seq, we provide a molecular mapping strategy to combine cell barcodes and gene identity into a single target code as a concise input into SQUICH (Fig. S3). Because this strategy involves hybridization, it has a further unique advantage to improve performance in single-cell applications: multiple target codes can be mapped to the same molecule (e.g. RNA) through hybridization in (Fig S1) with the potential to reduce drop-out, resolve isoforms and overcome 3' bias or the requirement of a poly-A tail.

We predict that MEG's design enables even further sampling reductions by providing a platform to convert measurement of nucleic acids into target codes that can be measured by approaches such as compressed sensing, which is not possible achieved with traditional sequencing (Candès, 2006; Cleary, Cong, Lander, Regev, 2017). SQUICH and MEG enable experiment-specific sampling paradigms that lead to future sampling reductions, for example to measure molecules only when their abundance is above a prespecified value. In proof-of-principle SQUICH experiments achieves 10x reduction, we foresee much greater fold reduction by increasing hamming distance between sequences in the pool of targets, competitors and encoders, and increasing purity of oligosynthesis, and by experimental designs that enable specific sampling of only species exceeding or depleted by a prespecified fold. This can be achieved by SQUICH by varying the abundance of each competitor (or encoder) target-by-target, so that for example, either encoders in early rounds are omitted, resulting in only sampling species exceeding a fixed threshold, or increasing encoders in early rounds and decreasing competitors to sample species at low abundance more deeply (unpublished work). In summary, MEG is a new approach for overcoming fundamental limitations in molecular sampling and could enable a new generation of efficient, precise biochemical measurement, from screens to detection of rare species in the blood and single-cell sequencing at an unprecedented resolution, with large numbers of potential variations and platforms.

Figures

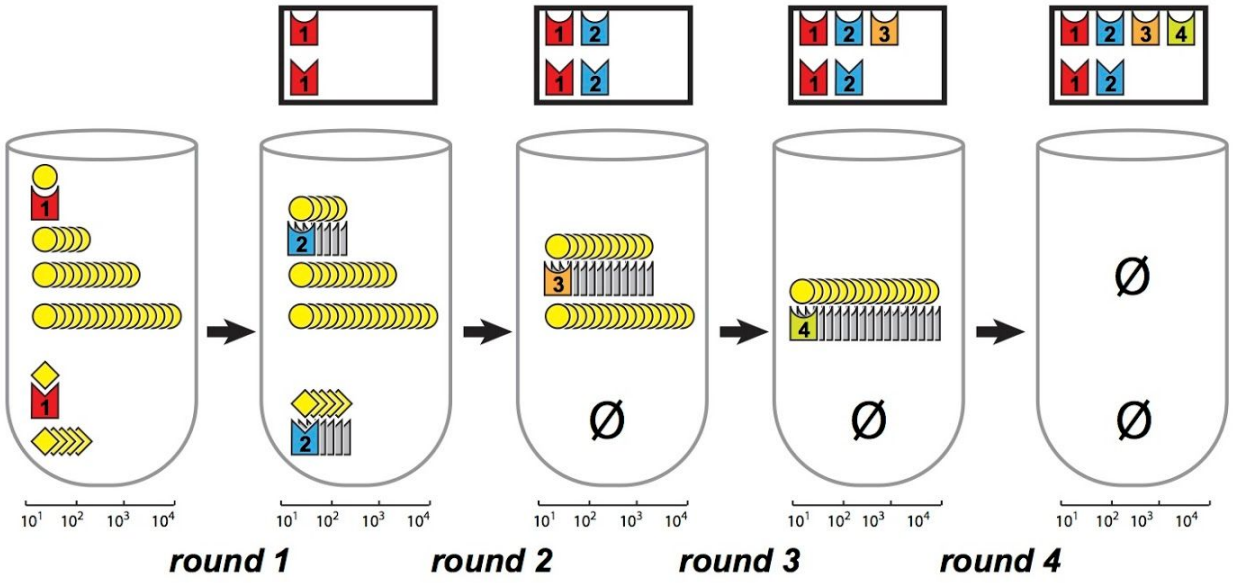

Figure 1: Stylized illustration of SQUICH acting in parallel on spheres and cubes: each

interact in parallel with limiting encoders and competitors. When an encoder interacts with a shape, the shape is labeled with that step at which it interacts with the encoder and is brought to a sampling box. The process continues sequentially: encoders are constant amounts while competitor abundance increases (here geometrically). If this geometric increase is in base 10, observation of a tag with the number "4" implies that the original number of molecules in the tube exceeded 10^4. The total number of shapes in the sampling box is low, requiring very few samples to fully sample it.

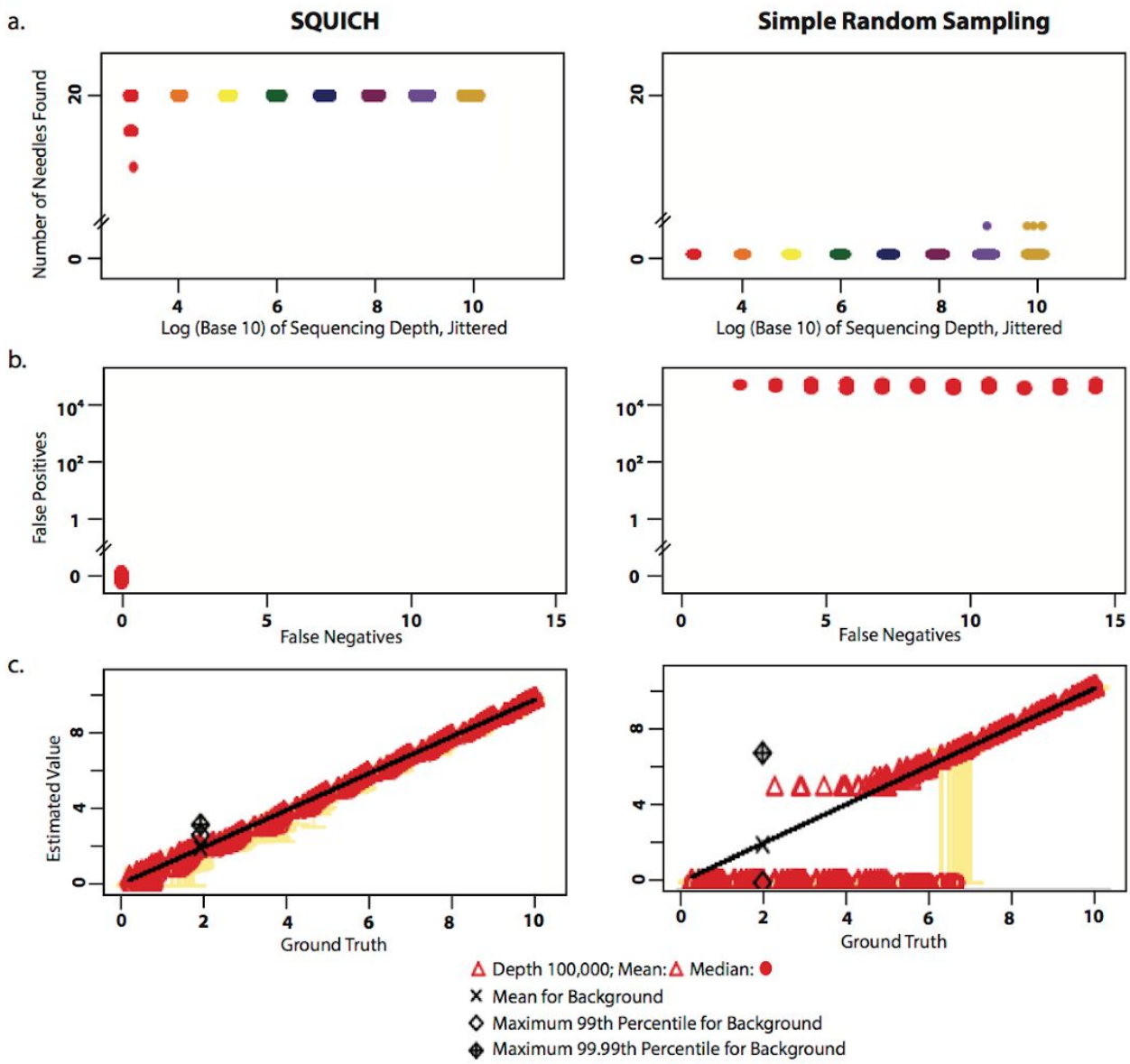

Figure 2a: In a background of two species at abundance 10^{15}, SQUICH (L) detects all 20 species with abundance *100* in all simulations with 10^4 samples; SRS (R) with 10^9-fold more reads fails to achieve the same detection yield (1000 trials, x-axis is $\log_{10}$-scale and jittered). Each sampling depth is depicted in a unique color. **b:** SQUICH enables detection of small fold changes, here 2 fold in 20 species, in a background of *>260,000* species. With 10^5 samples, all true positives are recovered with 0 false positives. To achieve the same performance with SRS, 10^3-fold more samples are required (Supplement) (x- and y-values jittered). **2c:** Detection performance of SQUICH and SRS (100 replicates each) in quantifying species across 10 orders of magnitude. 10 species were assigned an abundance of each value (1:10) $10^{0:9}$; and the remaining ~3000 species were set to background level of 100. Yellow bars denote the 25-75th quantiles of measurement for SQUICH and SRS; SQUICH enables detection of small and large molecular abundances across a dynamic

range of 10 orders of magnitude with 100,000 samples (top L). At this depth, more than half of species drop-out of SRS sampling and log MSE error is More than 10^10 samples are required by SRS to achieve the same rates of drop-out and log MSE is roughly 2x SQUICH at this depth (Supplement; Table S2).

a.

code1	code2	code3	code4	code5	code6	Ground Truth	SEM
165	344	310	148	10	0	79954	277211
39	82	95	43	8	2	79954	160782
49	88	54	37	2	0	7996	65349
7	13	8	9	1	0	800	19974
0	9	13	2	0	0	800	5507
0	4	2	1	0	0	1	1765
0	1	1	1	0	0	1	1545
1	1	3	0	0	0	80	701
0	2	2	0	0	0	1	512

b.

Code Sum	Ground Truth
5128	79955
1399	79955
1000	7996
337	800.5
115	1
109	1
95	800.5
59	1
54	1
47	1
28	80.95
28	1
26	1

c.

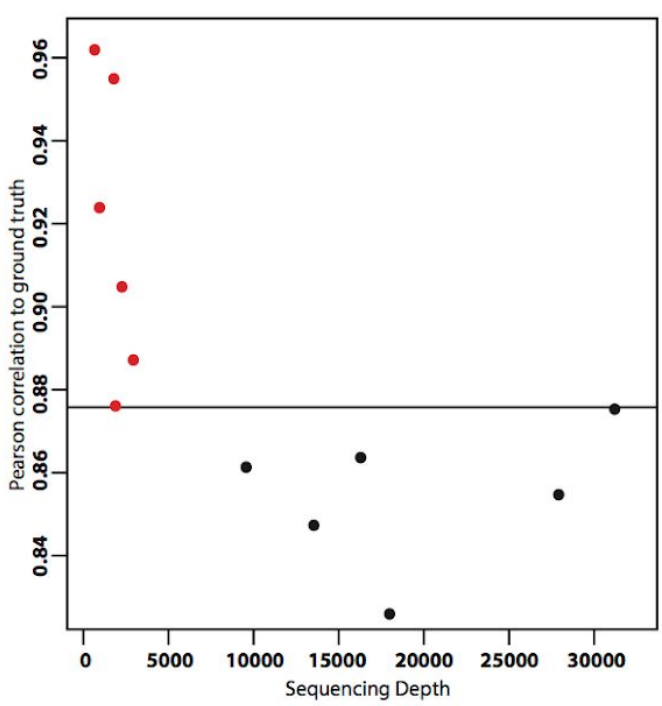

Fig 3: Six SQUICH replicates were sequenced at depths from 1583 to 3305; six conventional sequencing replicates were sequenced to depths of an average of 9 fold greater, from 10345 to 57213 reads. (a) Example of sequencing reads in each code round by SQUICH and SRS (best representative experiment collapsed over two technical replicates for SQUICH and SRS shown); (b) Pearson correlation between estimated counts and ground truth of each SQUICH replicate exceeded each replicate of conventional sequencing although the correlation in one replicate of SQUICH (depth 2407) exceeded the best conventional library (depth 30653) by only ~ .00015.

Acknowledgements

The authors thank Peter Kim for many useful discussions throughout the project and for an early suggestion of thermodynamic modeling in simulations; Andrew Fire for many useful discussions and critical reading of the manuscript; Peter Wang for useful discussions and generating graphics; Robert Bierman and Julia Olivieri for help with data parsing and critical reading of the manuscript; Eric Freeman for useful discussions, running scripts and generating plots; the Salzman Lab and Rhiju Das for reading of the manuscript. DNA sequencing was performed by the Stanford Functional Genomics Facility, in particular, we thank Xuhuai Ji for running all samples.

Bibliography

Cleary, C., Cong, L., Lander, E., & Regev, A. (2017). Composite measurements and molecular compressed sensing for highly efficient transcriptomics. bioRxiv, doi: https://doi.org/10.1101/091926

Vallejos, C. A., Risso, D., Scialdone, A., Dudoit, S., & Marioni, J. C. (2017). Normalizing single-cell RNA sequencing data: Challenges and opportunities. Nature Methods, 14(6), 565-571. doi:10.1038/nmeth.4292

Zadeh, J. N., Steenberg, C. D., Bois, J. S., Wolfe, B. R., Pierce, M. B., Khan, A. R., . . . , Pierce, N. A. (2010). NUPACK: Analysis and design of nucleic acid systems. Journal of Computational Chemistry, 32(1), 170-173. doi:10.1002/jcc.21596

Candès, E. (n.d.). Compressive sampling. Proceedings of the International Congress of Mathematicians Madrid, August 22–30, 2006, 1433-1452. doi:10.4171/022-3/69

Gu, W., Crawford, E. D., O' Donovan, B. D., Wilson, M. R., Chow, E. D., Retallack, H., & Derisi, J. L. (2015). Depletion of Abundant Sequences by Hybridization (DASH): using Cas9 to remove unwanted high-abundance species in sequencing libraries and molecular counting applications. doi:10.1101/031708

Conway, J. H., & Sloane, N. J. (2011). Sphere packings, lattices and groups. New York: Springer.

Hubank, M., & Schatz, D. (1994). Identifying differences in mRNA expression by representational difference analysis of cDNA. Nucleic Acids Research, 22(25), 5640-5648. doi:10.1093/nar/22.25.5640

Boone, M., De Koker, A., & Callewaert, N. (2018). Capturing the 'ome': The expanding molecular toolbox for RNA and DNA library construction. Nucleic Acids Research, 46(6), 2701-2721. doi:10.1093/nar/gky167

Wang, J. S., & Zhang, D. Y. (2015). Simulation-guided DNA probe design for consistently ultraspecific hybridization. Nature Chemistry, 7(7), 545-553. doi:10.1038/nchem.2266

Zhang, D. Y., Chen, S. X., & Yin, P. (2012). Optimizing the specificity of nucleic acid hybridization. Nature Chemistry, 4(3), 208-214. doi:10.1038/nchem.124

Owczarzy, R., Tataurov, A. V., Wu, Y., Manthey, J. A., Mcquisten, K. A., Almabrazi, H. G., . . . , Peek, A. S. (2008). IDT SciTools: A suite for analysis and design of nucleic acid oligomers. Nucleic Acids Research, 36(Web Server). doi:10.1093/nar/gkn198

Sano, T., Smith, C., & Cantor, C. (1992). Immuno-PCR: Very sensitive antigen detection by means of specific antibody-DNA conjugates. Science, 258(5079), 120-122. doi:10.1126/science.1439758

Robinson, M. D., & Smyth, G. K. (2007). Moderated statistical tests for assessing differences in tag abundance. Bioinformatics, 23(21), 2881-2887. doi:10.1093/bioinformatics/btm453

Mcintyre, L. M., Lopiano, K. K., Morse, A. M., Amin, V., Oberg, A. L., Young, L. J., & Nuzhdin, S. V. (2011). RNA-seq: Technical variability and sampling. BMC Genomics, 12(1). doi:10.1186/1471-2164-12-293

1 Methods

1.1 Formal definition of SQUICH procedure

A sequential SQUICH procedure with d rounds is defined as follows: let C_i and D_i be the number of encoders and competitors, respectively, added in the ith round. The partial sums $b_j := \sum_{i=1}^{j}(C_i + D_i)$ together comprise what we call the base system and are assumed to satisfy $0 < b_1 < b_2 < \ldots b_d$. An intuitive example of a base system $\{b_i\}_{i=1}^{d}$ is a geometric progression with $b_i = 10^i$. In each round, targets are sequentially hybridized with a pool of encoders and competitors, designed such that through step i (inclusive), a total of b_i encoders and competitors have been added to the reaction. Formally, we give the following definition.

Definition 1.1 (SQUICH procedure of duration d for the base system $\{b_i\}_{i=1}^{d}$) *Consider a multiset T_0 of targets. The abundances of each of the targets are the parameters of interest. Fix an integer $d > 0$, called the duration, and the base system of d natural numbers $0 < b_1 < b_2 < \ldots b_d$, and fix*

$$\begin{aligned} C_i &= \textit{units of the encoder in round } i\textit{, and} \\ D_i &= \textit{units of the competitor, defined such that } \sum_{j=1}^{i}(C_j + D_j) = b_i. \end{aligned}$$

The active pool P_0 at time 0 is defined to be $P_0 = T_0$. For each SQUICH step $1, \ldots, d$, define the active pool P_i recursively as follows The SQUICH procedure of duration d in base $\{b_i\}_{i=1}^{d}$ is defined as follows:

for $i = 1, \ldots, d$ ***do***

Incubate the oligonucleotide pool P_{i-1} with C_i and D_i, anneal and extend to encode the identity of targets that extend on the encoders for round i.

Define the resulting active oligo pool as P_i.

end for

The reaction is terminated after step d and molecules that have extended on codes are sequenced.

When a target extends on an encoder, we say it has been "coded".

To measure high dynamic ranges and also for estimation of scientific digit notation, an example of a useful choice of $\{b_i\}$ is when b_i is geometric progression in base 10, namely $b_i = 10^i$ for all i, and C_i is constant. Because the number of encoders is constant, the maximum number of molecules of each species coded in each round is at most that constant, regardless of the abundance of the species.

1.2 Statistical Estimators for SQUICH

SQUICH enables flexible statistical algorithms for estimation; estimator families with unique performance characteristic are given below.

To state the estimators, we assume a SQUICH procedure with d rounds satisfies

$$\sum_{i=1}^{j}(C_i + D_i) = 10^j$$

for $1 \leq j \leq d$, and $C_i = 10$ for all i. We introduce a convenient entity to define the first estimator: for $1 < j < d$ let

$$f_j := (\underbrace{1, \ldots, 1}_{j-1}, \underbrace{0, \ldots, 0}_{m-j+1})$$

and let e_j denote the jth unit vector.

After SQUICH, data is generated by sampling with replacement from the pool of coded molecules. The counts observed from each round are denoted by $\mathbf{X} = X_1, \ldots, X_d$.

Definition 1.2 (SQUICH EstiMator 1 (SEM1) family) *Let r_i be the residual sum of squares from regressing the vector $\mathbf{X}$ on the two orthogonal predictors e_i and f_i, and let*

$$I := \frac{1}{N} argmin_{1 \leq i \leq m} r_i.$$

$$\overline{X^{(I)}} := \left(\frac{1}{I-1} \sum_{i=1}^{I-1} X_i \right)^{-1}.$$

Then the estimate of $\hat{n}_0$ given by SEM1 is

$$SEM1(\mathbf{X}) := \sum_{j=1}^{I-1} 10^j + 10^I X_I \overline{X^{(I)}}.$$

Note that in the discussion before the definition we have assumed that the C_i are constant. If they are not, then we assume that they satisfy $C_i > 0$ and define $\mathbf{Y} = \left(\frac{X_1}{C_1}, \frac{X_2}{C_2}, \ldots, \frac{X_d}{C_d} \right)$; then we proceed as above to determine I by using the residual sums of squares from instead regressing $\mathbf{Y}$ on e_i and f_i.

Intuitively, the SEM1 family uses a disciplined and analytically tractable procedure (regression) to identify the round in which the CGA target is limiting compared to competitors and encoders. The term $\overline{X^{(I)}}$ serves to normalize observed values of X_I and can be replaced with any estimator of the sequencing depth. Conditional on correctly determining the key round of interest I, the SEM1 estimator is then based on estimating the first significant figure by estimating a random variable sampled with a rate proportional to the number of molecules coded in round i. SEM1 family estimators can control for experimental bias and can be modified to produce unbiased estimates when the quantities C_i vary with i.

Other choices of the normalizing factor $\overline{X^{(I)}}$ are also possible. One straightforward method is to use spike-ins. Given that the ith spike-in has abundance y_i, let $y_{i,j}$ be the expected number of molecules of the ith spike-in coded in round j. Let $Y_{i,j}$ be the counts observed when sampling Y_i in round j. Then one can use

$$\overline{X^{(I)}} := \frac{\sum_j y_{I,j}}{\sum_j Y_{I,j}}$$

For small sampling depth, SEM1 may have poor performance because it fails to select I properly due to high variability in each coding round. In this case other choices of I can be used, for example, as the last round where all codes up to and including this round are observed:

$$I := \max_j 1_{X_j > 0 \ \forall j \leq i}.$$

Spike-ins are an integral part of each SQUICH experiment. They are designed to bracket the abundance of the range of values to be measured in a SQUICH experiment and will be used to define an estimator that uses the empirical behavior of spike-ins to quantify unknown molecular abundance.

Definition 1.3 (SQUICH EstiMator (SEM2): Interpolation through spike-ins) *Let A be the set of abundances used for spike-ins, and let S_a denote the set of spike-ins at abundance a.*

Let X_i be the vector of observed sequencing counts of coded molecules for the ith spike-in. For any target of unknown abundance, let X be the vector of observed counts of coded molecules for the target. Then let

$$SEM2(X) = argmin_{a \in A} \frac{1}{|S_a|} \sum_{X_i \in S_a} ||X_i - X||_2.$$

In fact, SEM2(X) is consistent if there exists $a \in A$ such that the target has abundance a.

1.3 Simulation Methods

The abundances of the species in the target pool are fixed numbers. SQUICH was simulated following Definition 1.2 with additional variance introduced by modeling of thermodynamic effects below. After SQUICH coding, each coded species was sampled with replacement and compared to simple random sampling (SRS) directly from the original pool.

1.3.1 Modeling of thermodynamic effects

Molecular noise in SQUICH can be decomposed into noise from hybridization inefficiency and noise from off target hybridization. In our model, we use a worst-case for SQUICH: we assume that targets, codes and competitors are all oligonucleotides such that for each code (respectively competitor, encoder), at any variable nucleotide in the code (respectively competitor, encoder), there exists another code (respectively competitor, encoder) one mismatch away from it. This is less favorable than an optimal design of targets that one could take from sphere packing theory, in which each pair of targets would have edit distance of larger than 1.

Off target hybridization effects are modeled as follows: we assume that the number of coded and competed targets will be a function of a discrimination factor χ that measures the relative probability of mismatching between perfectly and imperfectly complementary oligos. Simulations use values of χ of 1/50 and 1/100, conservative estimates of reports in the literature: 1/100 and 1/500 (Wang et al, 2015; Zhang et al, 2012).

For an encoder or competitor e with complement $t^{(0)}$ and having i mismatches to a target $t^{(i)}$, (Wang et al, 2015; Zhang et al, 2012).show that there is an oligonucleotide design such that

$$\frac{P(\text{e is bound to } t^{(i)})}{P(\text{e is bound to } t^{(0)})} = \frac{1}{\chi^i}.$$

If there are N molecules of each specific encoder or competitor sequence, and there are n regions of the sequence taking the value G or C, then the above implies that the expected number of targets hybridizing to any encoder or competitor having up to k mismatches is

$$N \sum_{i=1}^{k} \binom{n}{i} \chi^{-i}. \tag{1}$$

$k = 5$ is used in simulations; the energetic penalty of 4 or more mismatches is very high (NUPACK; Zadeh et al, 2010), and 5 allows for further caution.

This approximation is conservative as other targets will also be competing with encoders and competitors that mismatch to targets. Further, perfect complementarity is not required to fulfill the conditions for SQUICH. Many of the sub-optimal energetic (i.e. imperfect complement) configurations in SQUICH are indistinguishable from optimal energetic hybridization where all duplexes are perfect. An example of this is as follows. Let X and X' denote the vector of perfectly duplexed targets and encoders during a SQUICH round, and X'_π denote a permutation of the entries in X' such that X can prime on X'_π in each entry. It is clear that for any permutation π, the sequence of X after it primes on X'_π will not depend on π. Finally CGA codes are suboptimal because each is hamming distance 1 from n other target codes. Our simulation results could have been improved by consider a set of targets where each pair is 4 or more mismatches away. In this case, thermodynamics predicts almost no imperfect duplex binding.

Probability of failure of encoder, competitor and target to hybridize have each been set to 0.1, corresponding to reports by Integrated DNA Technologies (IDT) of typical empirical efficiencies of annealing and extension in thermocycling conditions, the molecular procedures performed in SQUICH.

1.3.2 Simulation implementation

Spike-ins and species with unknown abundance were both pooled in silico.

Simulations model the SQUICH procedure on each species independently. One unit of encoder (competitor) is considered to be a mixture of one of each reverse complement of every CGA target sequence, with an auxiliary encoder-specific (competitor-specific) tag. A SQUICH numeral system and number of encoder molecules were used as inputs to the simulation. In each round, the number of active targets were calculated as (and encoders and competitors similarly).

```
active targets = available targets - escaped targets
```

where the number of escaped targets is determined by the following R code:

```
if (n[i]<=10^8){
    escaped=rbinom(1,n[i],p.escape)
} else {
    escaped= n[i]*p.escape + ceil(rnorm(0,1) * sqrt( n[i]*p.escape*(1-p.escape)))
}
```

where *p.escape* is a value representing the probability of the failure of the target to hybridize to either of the encoders or the competitors, modeling noise such as binding of oligos to the tube in an experiment. Due to conventions of variable types in R, if $n[i]$ (the number of active targets in round i) exceeded 10^8, the normal approximation to the binomial was used rather than the function rbinom; this is precise for large values. The number that failed to hybridize and thus "escaped" in round i carry through to the next round; thus a necessary condition that $n[i+1] = 0$ is that $escaped = 0$ in the previous round.

We computed active competitors and active encoders molecules analogously, including an inflation factor due to off-target binding using a variable thermo.bias which is defined to be the value in Eq. (1).

Given this variable, active competitors and encoders are random variables that follow the binomial distribution $B(n, p)$ with n equal to the number of competitor or encoder molecules multiplied by the inflation factor (1 + thermo.bias), and with p given by 1 minus some probability of failure, which is the probability of either competitor or encoder escape. Again, due to conventions of variable types in R, if the number of active molecules exceeded 10^8, we assigned the number of active molecules using a normal approximation to the binomial . A code example for the case of competitors follows; here comp[i] denotes the number of competitors added in round i and active.comp is the number of active competitors in the round.

```
binom.comp=floor(comp[i]*(1+thermo.bias))

if (binom.comp <=10^8){
    active.comp = rbinom(1,binom.comp,1-p.comp.failure)
} else {
    active.comp = binom.comp*(1 - p.comp.failure)
                 + ceil(rnorm(0,1) * sqrt(comp[i]*p.comp.failure*(1-p.comp.failure)))
}
```

This method treats all molecules as independent, and also inflates the variance (a 'worse case') for SQUICH performance by modeling all noise independently.

Then, if in the ith round, the number of active targets $n[i]$ exceeds the sum of the numbers of active encoders and active competitors, the number of potential targets in the next round is the number of active targets reduced by the sum of these two numbers and increased by the number of escaped targets; also, the number of encoded targets at this round is determined to be the number of active encoders as calculated above (as in this case, all active encoders are assumed to bind with targets). If $n[i]$ is less than the sum of the numbers of active encoders and active competitors, then the encoders need not all bind (as they, informally, compete with the competitors), and so the number of coded molecules is set to a hypergeometric random variable

```
coded[i] = rhyper(N=1, m=active.encoders, n=active.comp, k=n.active)
```

and $n[i+1]$ is set equal to the number that escaped. There is an approximation used similar to that above if the number of active targets is large.

1.3.3 Simulation conditions

In all simulations, p.escape, p.degfailure and p.codefailure were set to 0.1 in all simulations, and discrimination was set to 1/50.

For simulation 1, 20 needles at abundance 100 were simulated in a background of two species at values 10^x as reported in the text and supplement. Formally, two spike-ins were used: one set to 100 and the abundance of the value 10^x for programming consistency, though this value was treated as a second species of 10^x and not used for estimation purposes as it would be if it were a spike-in. Encoder abundance was 10 and the vector of total molecules per round with the ith component being total molecules in round i were code molecules in addition to $(0, 10^2, 10^3, 10^4, 10^5, 10^6, 10^7, 10^8)$.

For simulation 2, a complexity of 2^{18} was used; 20 of these species were 2000 and the remainder were 1000. Spikes used for SQUICH were 52 copies of 2000 and 2 copies of each of species with abundance $(1000, 1500, 2000)$. The vector of encoder abundance with the ith component representing the code in round i was 100 in each round; the vector of total molecules per round with the ith component being total molecules in round i were code molecules in addition to $(1000, 1200, 1400, 1600, 1800, 2000)$. Spikes for multinomial were $(1:10) \times 10^{(2:5)}$. The encoded molecules in the first three rounds of SQUICH were omitted, as modeled by replacing encoders $1:3$ with a competitor. For programming consistency, we modeled this by removing encoded molecules from rounds 1-3 from sampling.

For simulation 3, a complexity of 2^{12}. For high dynamic range, 10 species each were set to each value in $(1:10)x10^(0:9)$. For modeling single cell data, 10 values each were set to each range in $(1:10) \times 10^{(0:4)}$ and 100 at each copy number from 1:10. Spike-ins were set to 3 copies each of $(1, 5, 10) \times 10^{(5:9)}$ for high abundance dynamic range and spike-ins were set to 3 copies each of $(1, 5, 10) \times 10^5$ for single cell trials. The remaining species were set to a background level of 100. The vector of encoder abundance with the ith component representing the code in round i was $(5, 10, 100, 100, 100, 100, 100, 100, 1000, 1000, 1000)$; the vector of total molecules per round with the ith component being total molecules in round i were code molecules in addition to $(0, 10^2, 10^3, 10^4, 10^5, 10^6, 10^7, 10^8, 10^9, 10^{10}, 10^{11})$. Each simulation was performed with 100 replications.

1.3.4 Estimators used

1. For Simulation 1 ("Needle in haystack"), the same criterion for detection was applied to SQUICH and SRS: if a sequence s_i was sampled, it was considered to be detected. For reasons of programming consistency, for simulations, we described results for having two species of abundant variety at 10^x and 20 at abundance 10. Formally, although completely equivalent from the sampling and results perspective, one of the abundant species was assigned the name 'spike-in' in the code. In addition, we added one spike-in at abundance 100 in all simulations.

2. For Simulation 2 ("Small fold change"), which models a background species of abundance 1000 and an above background species of 2000, SEM2 was used to call the abundance of the SQUICH sample. The same criterion for detection was applied for both SQUICH and SRS: if the estimated abundance exceeded 1700, the sequence was said to be above background and was said to be 0 otherwise.

3. For Simulation 3 ("Dynamic range"), a modification of SEM1 (described below) was used to call the abundance of the SQUICH sample.

In all three simulation types, standard quantification for simple random sampling based on spike-in normalization (see below) was used.

1.3.5 Normalization of SRS models

To normalize counts from SRS, i.e. to infer an estimate of the number of molecules present from the number of samples, we use the fact that the actual quantities of the spike-ins are known. So we use the conventional estimator based on spike-ins, which is given by

$$r := \frac{\sum_{s \in S} \text{actual}_s}{\sum_{s \in S} \text{obs}_i}$$

where S indexes the spike-ins used, obs_s is the observed count for spike s and actual_s is the (known) abundance of the sth spike. The number of observed counts for each species is multiplied by r to provide the estimate of abundance.

1.3.6 Implementation of SQUICH estimator for Simulation 3

We now describe how SQUICH estimates the abundance of a single species in the modification of SEM1 used for Simulation 3 (dynamic range); we denote the resulting estimate by $\hat{n}$.

Note that for SQUICH estimators, normalization is only required for SEM1 as other estimators use the known values of the spike-ins to provide estimates.

In this method, the key round I is identified using the approach described for the definition of SEM1 with regression.

Recall that for a particular target, we define X to be the vector of observed counts of coded molecules for that target. We estimate $\hat{n}$ in two parts:

$$\hat{n} = \sum_{i=1}^{I-1} \hat{n}_i + \hat{n}_I \tag{2}$$

by analogy with the definition of SEM1, where $\hat{n}_i$ is the estimated number of molecules depleted in round i. We use an estimate of each $\hat{n}_i$ that corrects for: the probability of code escape (p_e); for the probability of encoder/competitor escape (p_d); for the fact that the number of encoders C_i varies; and for mismatching (χ).

First, spike-ins are used to compute a normalization factor like the quantity $\bar{X}^{(I)}$ described in $SEM1$. For round i and spike-in s, let $\exp_{is}$ be the number of expected coded molecules of i in the simulation of SQUICH; this is determined by going through the simulation procedure and using expected values when randomization takes place. (Recall that the true original numbers of spike-ins are known.) Then we denote the number of samples observed for spike-in s in round i (i.e. the number of observed samples after sampling the underlying molecular pool after the SQUICH procedure) by obs_{is}.

Simulations use the term

$$\overline{X^{(I)}} = \frac{\sum_s \exp_{Is}}{\sum_s \text{obs}_{Is}}.$$

Intuitively, multiplication of observed samples for any species in round I by $\overline{X^{(I)}}$ adjusts the observed sampled molecules in each round to arrive at an estimate of the underlying number of coded targets.

The estimate $\hat{n}_I$ used for the number of targets remaining in the key round i is:

$$\hat{n}_I = X_I \overline{X^{(I)}} \frac{C_I + D_I}{C_I} \frac{1}{1 - p_e}$$

The estimates for the the number of target molecules depleted in earlier rounds is given, for $1 \leq i \leq I - 1$, by:

$$\hat{n}_i := \left(1 + \sum_{i=1}^{k} \binom{n}{i} \chi^{-i}\right) (C_i(1 - p_d) + D_i(1 - p_d)),$$

where k is a parameter; as noted above, $k = 5$ is used in simulations. Then these last two expressions can be used in equation (2) to arrive at a final estimate of abundance.

1.4 Experimental Methods

SQUICH Encoding steps Oligonucleotides are labeled JSON# (Table S4). A target pool containing the degenerate target library JSON342 and the individual oligonucleotides added above background JSON343-JSON348, was prepared. In the encoding experiments, the final concentrations were 600 pM for JSON342, 180 pM for JSON348 and JSON347, 18 pM of JSON346, 1.8 pM of JSON345 and JSON344 and 180 fM of JSON343. The encoding/competition extension reactions were performed in the presence of 20 mM Tris-HCl (pH 8.0), 50 mM KCl, 50 nM of each dNTP (NEB), 12.5 nM dideoxy-CTP (GE Healthcare), 1.5 mM MgCl2 (NEB) and 1.25 units of Native taq polymerase (Invitrogen). The initial volume was 50 L.

Experiments 1-6 are SQUICH; experiments 1-3 (CH52.01-CH52.03) were performed with identical conditions andencoder oligonucleotides were added in each round. Experiments 4-6 (CH52.03-CH52.06) were performed with identical conditions and received competitor instead of the encoder in the first round. For experiments 1-3, 1 L of the encoding oligonucleotide, JSON350, was added to 49 L of the target/enzyme solution to a final concentration of 37.5 pM. For experiments 4-6, 1 L of the competitor oligonucleotide, JSON454, was added to a final concentration of 37.5 pM. The reaction mixtures were then subjected to the following thermocycler program: 95C for 5 min followed by five cycles 65C for 30 sec and 72C for 30 sec; followed by a 5 min hold at 72C and then 4C. This is designated Round 1 of the SQUICH procedure.

For the second round, 1 L of the second encoding oligonucleotide, JSON351, was added to all 6 reactions to a final concentration of 75 pM. One L of the competitor, JSON454, was added all reactions to a final concentration of 600 pM. The reactions were then subjected to thermocycling with the same program as Round 1. This completes Round 2 of the procedure.

In Rounds 3-6, 1 L of an additional encoder (JSON352-355) and 1 L of competitor were added at each step, followed by the thermocycling program described above. Each encoder was added to a final concentration of 75 pM each. The concentration of added competitor increased by 10-fold during each step from Round 2 to Round 6 (600pM, 6 nM, 60 nM, 600 nM, 6 M final). The cumulative concentration of competitor was over 6.7 M. Adjustments were made to compensate for the increased volume of the reactions during the procedure.

Experiments CH53.01-CH53.06 model conventional sequencing. An equimolar mixture of the six encoding oligonucleotides, JSON350 to JSON355, was added simultaneously to the target pool reaction mixture to a final collective concentration of 600 pM (100 pM each). This mixture was subjected to the following PCR program: 95C for 5 min followed by five cycles 65C for 30 sec and 72C for 30 sec; followed by a 5 min hold at 72C and then 4C. We also performed another model of conventional sequencing in CH52.07-CH52.09 by performing SQUICH but omitting competitor. These experiments had worse performance than conventional sequencing in CH53.01-CH53.06 and were omitted from analysis.

Some experiments, PCR was done in duplicate (denoted A and B in Table S6) and sequenced reads from each experiment were pooled. The sequencing barcodes were added using standard protocols and sequenced using custom primers (gift from the Jin Billy Li lab (Table S5). The 50 L reaction mixtures were composed of 25 L of the SQUICH reaction with 1X Phusion HF buffer (NEB), 200 M dNTPs, 0.5 M of each primer, and 1 unit of Phusion DNA Polymerase (NEB). We used the following PCR program: 98C for 2 min; four cycles of 98C for 10 s, 55C for 10 s, and 72C for 30 s; nine cycles of 98C for 10 s and 72C for 30 s; and 72C for 5 min. The PCR products were then purified via AmpureXP magnetic beads (Beckman Coulter) at a ratio of 1:1.8. The purified products were quantified on the QUBIT fluorometer and pooled. The final pool was quantified by the Stanford Functional Genomics Facility using the Agilent Bioanalyzer (DNA High Sensitivity).

DNA sequencing DNA sequencing was performed on an Illumina MiSeq instrument (San Diego, CA) by the Stanford Functional Genomics Facility using the MiSeq Reagent Nano Kit v2 (300-cycles), Paired End setting, 60 bases per cycle.Data Parsing was performed using custom scripts to identify reads containing targets, assign codes, and perform statistical analysis described in detail in Analysis of Experimental Data.

1.5 Analysis of Experimental data

Experimental data was generated blindly, that is, the estimation procedure was chosen before the experiment was performed, and all tests in this sequencing run are reported. Parsing was performed by scanning each sequencing read for a target code matching our designed target code libraries (Table S4), and identifying the sequence encoding the round the respective encoder was added appearing adjacent to the target code (Table S4, S5). We required that both R1 and R2 of the sequencing reads match the same target code and also the encoder sequences matched in R1 and R2. Parsed data is given in Table S6. Note that we refer to spike-ins in this experiment as molecules where ground truth is known, as they are a way to estimate the efficiency of SQUICH. These spike-ins are not used for normalization, as they are for simulation. Two technical replicates (SQUICH) were pooled and analyzed with a modified SEM1 estimator. We defined $b_i = f^i$ where

$$f = 10^{\log_{10}(8000)/5}$$

based on previous experimental results that suggest p_e and p_d reduce the value of the SQUICH numeral system from a theoretical value of $b_i = 10^i$ to $b_i = 6^i$ in our proof of principle experiment.

We defined the SQUICH SEM1 estimator as

```
x[,sem:=code1*f+(code2*f^2)*(dig2 &!(dig3))+(f^2*code2+code3*f^3)*(dig3
&!(dig4))+(f^3*code3+code4*f^4)*(dig4
&!(dig5))+(f^4*code4+code5*f^5)*(dig5
&!(dig6))+(f^5*code5+code6*f^6)*dig6]
```

where

```
dig i
```

is the indicator that codes from a given sequence were observed up until round i but not observed in round $i+1$. If SEM above is 0, for example if 1 count for code 2 and code 5 are observed, we set SEM to the sum of all codes.

For control libraries, the species abundance was estimated as the total counts observed between each sequence.

We then computed two metrics: correlation (Spearman and Pearson) and M_1, defined as

$$M_1 := \sum_{i=1}^{6} \pi_i \sigma_i$$

where σ_i is the rank of the ith spike-in based on the empirical rank of SEM1 (for SQUICH) or on total counts (for conventional sequencing) and π_i is the rank of the ith spike-in.

To normalize for variation in sampling depth and to estimate the fold reduction enabled by SQUICH compared to conventional libraries, we use units of effective sequencing depth (ESD). Given experimental performance measured in a metric such as correlation of M_1, ESD is an estimate of the depth required by SRS without noise introduced during sequencing to achieve, on average, comparable performance as the experiment.

To calculate ESD, given an experimental depth of d_0 reads, for $\alpha = .1, .2, .3, .4, .5, .6, .7, .8, .9$ and $\alpha = 1, 2, 3, 4, 5, 6, 7$ we simulated 1000 replications of SRS at depth αd_0 and reported the fraction of replicates of SRS at depth αd_0 that exceeded the observed value of M_1. This fraction is an empirical p-value for M_1 under the Null Hypothesis that the experiment performs as well as SRS at depth αd_0.

For experiment i, the empirical p-value for a fixed α is denoted as $\hat{p}_i^{(\alpha)}$. A basic property of p-values implies that if SQUICH (conventional sequencing) were performing equivalently to SRS with αd_0 samples, $\hat{p}_i^{(\alpha)}$ would be distributed as uniform (0,1). For each α, we computed the Kolmogorov Smirnov goodness of fit test statistic between $\{\hat{p}_i^{(\alpha)}\}$

and the uniform distribution (0,1). Call the value of this test statistic $X(\alpha)$. Then ESD is the value of α giving the best fit, or, formally,

$$ESD := \text{argmin}_{\alpha} X(\alpha).$$

Pooled, the six SQUICH replicates had an ESD of about 5, meaning to achieve the same precision, 5 fold fewer samples are required compared to the number expected with SRS without experimental noise. ESD of conventional sequencing is .5, meaning 2x more samples than predicted to be needed by SRS are required to overcome experimental noise. Taking the ratio of SQUICH ESD to conventional ESD gives the estimate that SQUICH achieves a 10x depth reduction compared to conventional sequencing.

Our experiments are conservative in comparison with conventional sequencing. Our experimental model of conventional sequencing is in a regime where tags were limiting, which attenuates dynamic range usually seen in SRS and increases performance. Second, our experiments are with unpurified oligos and using degenerate bases (Fig S1), meaning any target sequence has other target sequences 1 mismatch away.