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
Correspondence: julia.salzman@stanford.edu

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 *1* cube and 10^3 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. In the second round of SQUICH, up to 1 of each shape is captured and tagged with a ``2" and added to the sampling box. Then, up to 9 of each shape are captured and destroyed. In this round, no cubes are captured, as they have all been removed. In the third round, up to one shape is captured, marked with a ''3'', and up to 88 of each shape are destroyed. The numbers of captured and destroyed molecules in the third round satisfy the property that 10^2 ($1 + 9 + 1 + 89 = 100$) of each shape has been destroyed or captured up to and including this round. In round four, 1 (or 0) of each shape is captured, tagged with a "4", and up to *898* of each shape are destroyed; in round 5, one more shape is captured and up to 8998 are destroyed, and so on. 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 order of magnitude of each shape in the original tube can now be estimated with *5* samples without replacement; while SRS requires on the order of 10^3 samples 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 $10^{y_2-y_1}$ 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

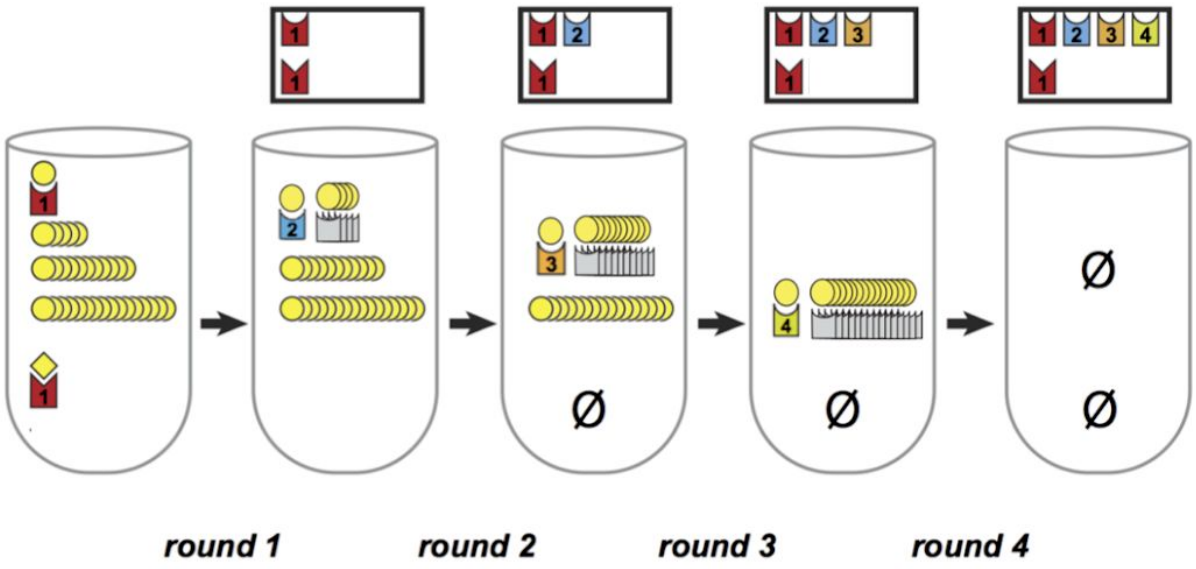

Figure 1: Stylized illustration of SQUICH acting in parallel on spheres and cubes: each interact in parallel with limiting encoders and competitors. When an encoder interacts with a shape, the shape is labeled with that step at which it interacts with the encoder and is brought to a sampling box. The 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.

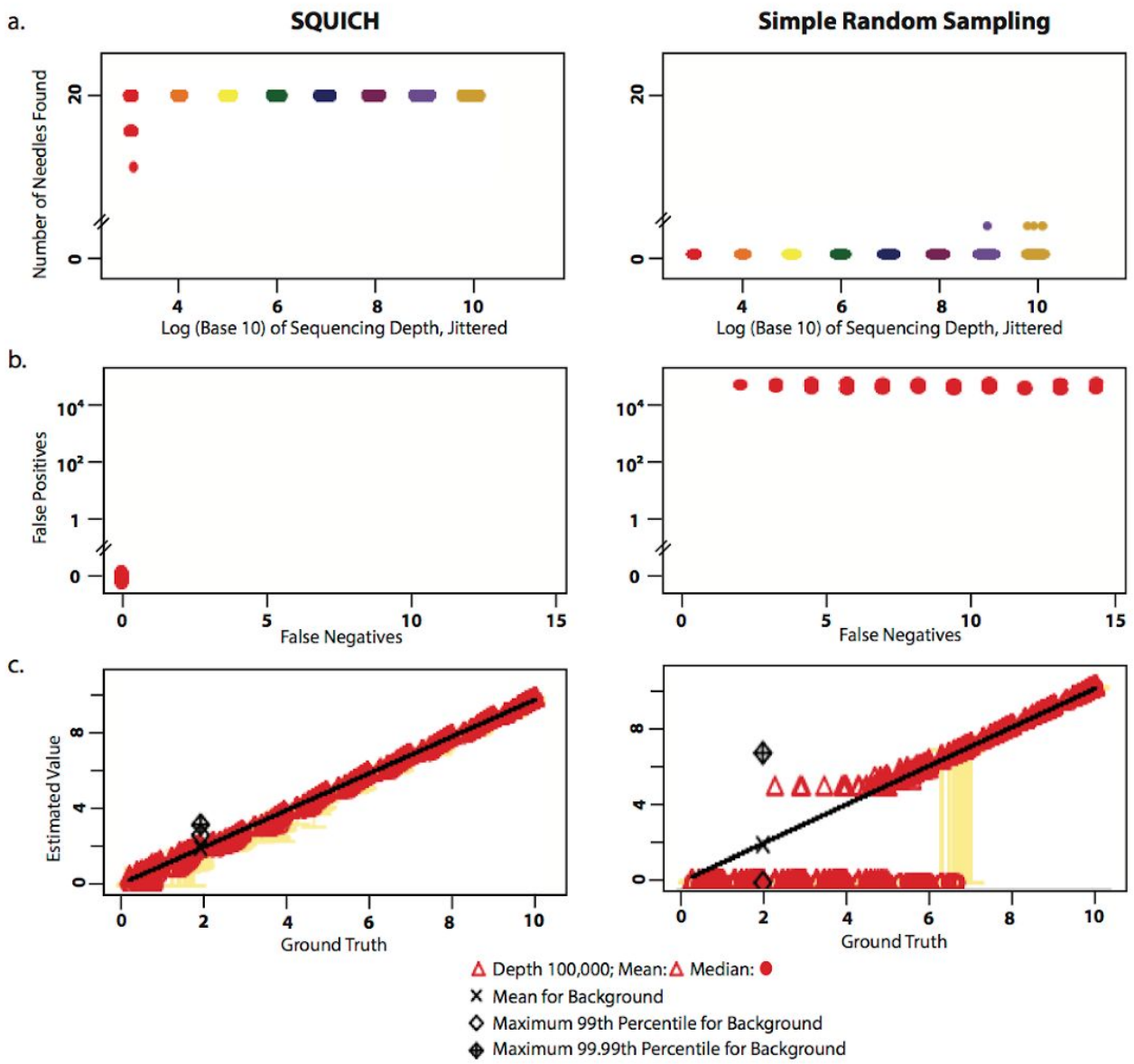

Figure 2a: In a background of two species at abundance 10^{15}, SQUICH (L) detects all 20 species with abundance *100* in all simulations with 10^4 samples; SRS (R) with 10^9-fold more reads fails to achieve the same detection yield (1000 trials, x-axis is $\log_{10}$-scale and jittered). Each sampling 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.

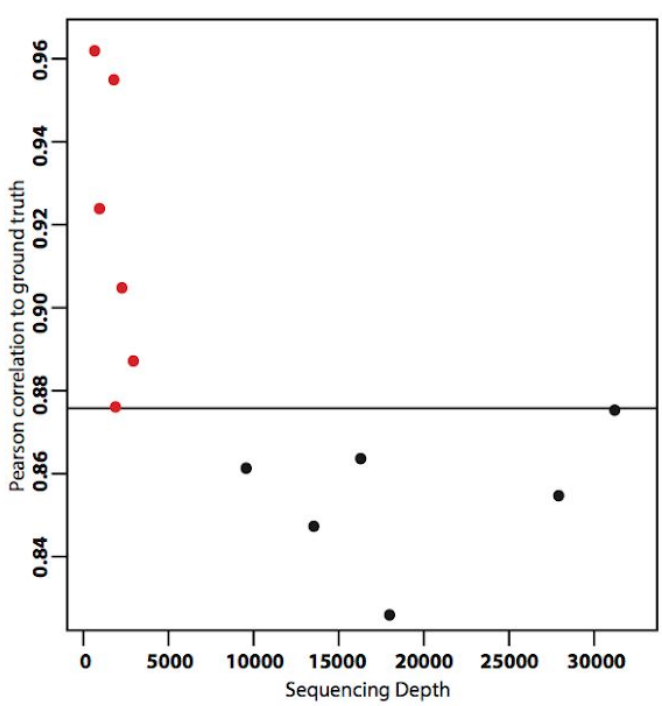

Fig 3: Six SQUICH replicates were sequenced at depths from 1583 to 3305; six conventional sequencing replicates were sequenced to depths of an average of 9 fold greater, from 10345 to 57213 reads. (a) Example of sequencing reads in each code round by SQUICH and SRS (best representative experiment collapsed 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

Methods

Caroline Horn and Julia Salzman

July 31, 2018

1 Probabilty models

1.1 Probabilistic formulation of round-coding, definitions

In non-filtered round coding, the ith SQUICH step is deterministic if $n_i > C_i + D_i$ in which case, C_i molecules are coded and D_i are competed out. Because of this, it is useful to introduce the following unique representation of n_0 where $0 \leq n_0 \leq \sum_{i=1}^{m}(C_i + D_i)$ which uniquely defines l:

$$n_0 = r_l + \sum_{i=1}^{l-1}(C_i + D_i) \qquad \text{with } 0 \leq r_l \leq C_l + D_l \text{ and } 0 \leq l \leq m \tag{1}$$

Randomness is only introduced at step l, the unique step where encoders and competitors outnumber targets in its round-expansion. Because $C_i + D_i > n_i$ in this step, all molecules are coded or competed out.

Definition 1.1 (non-filtered round-coding) *Fix a target sequence and abundance n_0. For $i \geq 1$, and C_i, D_i as the number of encoder i and competitor molecules.*

Let l be defined by Eqn. **??**. *Let n_i denote the number of free molecules of this target at the end of round $i-1$. Let c_i and d_i be the number of molecules that hybridize and extend on the ith encoder or competitor(respectively). In non-filtered round-coding, for $0 \leq i \leq l-1$, $c_i = C_i$ and $d_i = D_i$ and*

$$c_l \sim Hyper(r = C_l, n = C_l + D_l, k = n_l).$$

Definition 1.2 (Filtered round-coding) *Fix constants χ, the discrimination factor and p_t, p_e, p_c, probability of escape of target, encoder and competitor escaping. At step i, if n_i molecules of a target are present in round i,*

$$min(n_i \sum_{j=1}^{k} \binom{n}{k} \chi^{-j}, C_j + D_j)$$

targets are called 'conditionally active'. At this round, with probability p_e and p_c respectively, there are are random number of active encoders and active competitors distributed as $Bin(C_j, p_e)$ and $Bin(D_j, p_c)$ respectively. With probability p_t, each conditionally active target molecule binds an active competitor and active encoder with equal probability. If bound, a target is considered "active". Active targets follow the probability distribution in non-filtered round coding.

2 proofs

Proof 2.1 (Proof of Claim ??) *Let*

$$z_{i,k} = 10 \ \cdot 1_{y_i > k} + x_i 1_{y_i = k}.$$

Then SQUICH transforms the sampling distribution from ratiometric sampling of $(x_1 10^{y_1}, x_2 10^{y_2})$ *to ratiometric sampling of* $(z_{1,1}, \ldots, z_{1,y_1}, z_{2,2}, \ldots, z_{2,y_2}$ *via the counts from each* s_i*-code pair below:*

sequence	*code 1*	*code 2*	...	*code* y_2
s_1	$z_{1,1}$	$z_{1,2}$	...	z_{1,y_2}
s_2	$z_{2,1}$	$z_{2,2}$	...	z_{2,y_2}

Suppose $x_1 = x_2 = 1$, $y_1 = 1$ *and* y_2 *is large. Then, to achieve detection of* s_1 *with probability* p*, there exists* α_p *such thatSQUICH requires*

$$\alpha_p \frac{y_1 + y_2}{y_1} = \alpha_p (1 + y_2)$$

samples whereas simple random sampling requires $\alpha_p \frac{10^{y_2} + 10^{y_1}}{10^{y_1}} = \alpha_p (10^{y_2 - y_1} + 1)$.

Geometric progression of $\{D + C_i\}_{i=1}^d$ is key to encoding scientific digits, as can be seen by a more general choice of x_i and y_i than in the claim. In this case, let $X_{i,j}$ be the observed counts in the i, j cell in the table above.

Further, the expected number of counts in any cell at depth $\frac{y_1 + y_2}{y_1}$ is α. Standard sampling theory implies that

$$\frac{X_{i,j}}{\alpha} \to z_{i,j}$$

with deviations on the order of $\sqrt{\alpha}$. This implies that with high probability, $\alpha = 100$ suffices for accurate scientific digit estimation.

2.1 Statistical estimators

Definition 2.2 (SQUICH EstiMator 1 (SEM1)) *In a SQUICH procedure with* d *rounds, and satisfying*

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

for $1 < k < m$ *let*

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

and e_k *the* k^{th} *unit vector.*

Given data $X_1, \ldots, X_m$ *counts observed from each round, let* r_k *be the residual sum of squares from of regressing the vector* X *on the two orthogonal predictors* e_k *and* f_k*, and let*

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

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

Then the estimate of $\hat{n_0}$ *given by SEM1 is*

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

Intuitively, the SEM1 uses a discplined and analytically tractable procedure (regression) to identify the round at which the CGA target is limiting compared to competitors and encoders. The term $\bar{X^{(I)}}$ serves to normalize observed values of X_I and can be replaced with any estimator of the sequencing depth, or even omitted in applied settings if it suffices to quantify sepcies up to a universal scalar constant. Conditional on correctly determining that round, the SEM1 estimator is then based on the significant digit by estimating a Poisson random variable sampled with a rate proportional to c_l.

In applied settings, we may use the estimator

$$SEM1(X) := \sum_{j=1}^{I-1} 10^j + 10^I \; min(1, \frac{X_I}{\bar{X^{(I)}}})$$

as intuitively, if no coded molecules are observed at round $I+1$, $SEM1(X) \leq 10^I$.

It is clear how SEM1 can be extended to arbitrary $\{C, D_i\}_{i=1}^d$ for variable C_i, for example, by defining

$$f_i = (\underbrace{\frac{1}{C_1}, \frac{1}{C_2}, \ldots, \frac{1}{C_{i-1}}}_{i-1}, \underbrace{0, \ldots, 0}_{m-i+1})$$

and re-weighting the estimators of $\bar{X}_I$ and $SEM1(X)$ accordingly; for example, the 10^{I-1} term in the estimate of $SEM1(X)$ in the above estimator derives from the base 10 increase in concentration of the encoder and competitor. In addition, in filtered-round coding, if estimates $\tilde{C}_i$ and $\tilde{D}_i$ of the effective C_i and D_i, each down-sampled by molecular escape, can be computed, they could replace C_i and D_i in the SEM1. The following claim is proved in the Supplement.

Claim 2.3 (SEM1 is consistent)

Note that SEM1 could be improved by building strength across samples and using spike-ins such as members of the multiset with ground truth. This allows development of empirical models of the behavior of CGA codes during a SQUICH procedure with increased precision. The full statistical power of spike-ins is a subject for future work. Here, we use spike-ins simply to ensure accurate estimation of tagged molecules in each coding round (see below).

For small sampling depth, SEM1 may have poor performance because it fails to select I properly due to high variablility in each coding round. SEM1 could be modified to allow for the values in f_k to take a functional form, such as linear which would result in SEM1 estimating I by regression on two predictors. In larger samples, it will be robust to escape of targets as occurs in filtered round coding. Before turning to sampling theory, we introduce a second estimator family (SEM2).

Definition 2.4 (SQUICH EstiMator (SEM2)) *In a SQUICH procedure with d rounds, and satisfying*

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

for $1 < k < m$ *let*

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

and e_k *the ith unit vector.*

Given data $X_1, \ldots, X_m$ *counts observed from each round, the the estimate of* $\hat{n_0}$ *given by SEM1 is*

$$SEM2(X) := \gamma \; argmax_{1<i\leq m} \left(X_i 10^i + X_{i-1} 10^{(i-1)} \right) 1_{X_j > 0 \; \forall j \leq i}$$

and γ *is a scalar normalizing constant which is a function of sequencing depth.*

As with SEM1, SEM2 has straightforward generalizations to variable encoder and competitor amounts. In addition, in filtered-round coding, if estimates $\tilde{C}_i$ and $\tilde{D}_i$ of the effective C_i and D_i, each down-sampled by molecular escape, can be computed, they could replace C_i and D_i in the SEM2. Intuitively, SEM2 will over-estimate n_0 any time a single molecule is sampled at a round that exceeds $log(n_0)$. However, SEM2 may have better performance than SEM1 in small sample sizes where SEM1 fails to estimate I properly. Other estimators in the SEM2 family replace $1_{X_j>0\ \forall j\leq i}$, a measure that X_j is not molecular noise but rather represents a true signal that $n_0 > 10^j$, with an arbitrary function of X_j where $j \leq i$.

Claim 2.5 (Statistical properties of SEM2) *SEM2 is a consistent estimator of $\hat{n_0}$ and for a geometric progression of $\{b_i\}_{i=1}^m$ in base b where $\sum_{i=1}^k (C_i + D_i) = b^k$, with $\frac{1}{\gamma} > ||X_i||_\infty$, in non-filtered round coding,*

$$P(SEM2(X) > b\, n_0) = 0.$$

The proof of the claim is straightforward and left to the reader.

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 empirical behavior of spike-ins to quantify unknown molecular abundance.

Definition 2.6 (SQUICH EstiMator (SEM3): interpolation through spikeins) *Let A be the set of abundances used for spike-ins, S_a denote the set of spike-ins at abundance a and n_a be the number at abundance a.*

Let X_i be the vector of observed sequencing counts of the ith spike-in. Let X be the vector of observed counts of the target of unknown abundance. Define the following estimator of $SEM3(X)$:

$$SEM3(X) = argmin_{a\in A} \frac{1}{n_a} \sum_{X_i \in S_a} ||X_i - X||_2.$$

Then $SEM3(X)$ is consistent if there exists $a \in A$ such that the target is at abundance a.

It is clear that SEM3 is consistent for values of the parameters that cioncide with with values of the spike-ins and that the bias for n_0 where $n_0(i) \leq n_0 \leq n_0(i+1)$ is bounded by $|n_0(i) - n_0(i+1)|$.

3 Simulation conditions

Parameter regimes were as described in the text and in the parameters in the files multi.squish.data.simulation.r and do.multi.squish.R.

For simulations, SQUICH and simple random sampling were both performed using sampling from the multinomial distribution with fixed sample size as described in the text. Spikeins and species with unknown abundance were both pooled in silico, subjected to SQUICH or left unchanged in the case of simple random sampling trials, and this pool was then sampled.

Simulations that involved SQUICH proceeded as follows. We use the simulation conditions that model the conditions of the "CGA code" as described in the main text. One unit of target, encoder and competitor are each considered to be a mixture of one molecule of each CGA target sequence or its reverse compliment respectively. For each round, the molecules of code used in each round and the cumulative sum of total molecules of competitor and code used up until and including this round were used as parameters to calculate the number of molecules of competitor and code per round (an arbitrary parameterization). In each round, the number of active targets, encoders and competitors were calculated using the following methodology to model noise.

The number of active targets at round i was computed as follows:

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

where $p.escape$ was a value representing failure of the target to hybridize to either of the compeitor or coder, modeling noise such as binding of oligos to the tube in a molecular biology experiment. Due to numerical conventions in R, if $n[i]$ exceeded 10^8, we assigned active targets a value equal to its expectation. The number that "escaped" at this round will always carry through to $n[i+1]$, thus a necessary condition that $n[i+1] = 0$ is that $escaped = 0$ above.

We computed active competitors and active encoders molecules in a similar way, but included a term for bias due to off-target binding using a variable thermo.bias. Thermo.bias is an inflation factor modeling imperfect matches between targets and encoder or competitor: with up k mismatches with a CGA code of length n is modeled as

$$thermo.bias = \sum_{i=1}^{k} \binom{n}{k} \chi^{-i}.$$

In our simulations, we chose $k = 3$ and χ as described in the text and below. Given this variable, active competitors and encoders are a random variable with binomial distribution $p = 1 - p.failure$ where $p.failure$ is the probability of competitor or encoder escape, and n equal to the number of competitor or encoder molecules multiplied by the inflation factor; again, due to numerical conventions in R, we assigned the active molecules to be their expectation if they exceeded 10^8. A code example is below:

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

if (binom.nd <=10^8){
    active.deg = rbinom(1,binom.nd,1-p.degfailure)
          } else {
    active.deg = binom.nd*(1 - p.degfailure)
          }
```

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

Given these calculations, if at this round $n[i]$ exceeds the sum of active encoders or competitors, $n[i]$ is reduced by the sum of these two numbers, and the number of encoded $n[i]$ at this round is set to the number of active encoders as calculated above. If $n[i]$ is less than this number, the number of coded molecules is set to a hypergeometric random variable

$$coded[i] = rhyper(nn = 1, m = active.code, n = active.deg, k = n.active)$$

and $n[i+1]$ is set to the number that escaped.

3.1 Parameters used

3.2 Experimental Methods

In the experiment without the competitor, a 600 pM solution of the target (disregarding the contribution from the spike ins) was extended using an equimolar pool of the code oligonucleotides (aggregate concentration of 600 pM). Native taq polymerase (Invitrogen) was added in the appropriate buffer and dNTPs, and the solution underwent 5 cycles of denaturing, annealing, and extension. This product was the template for sequencing library 1.

In the SQUICH experiment, the target pool at 600 pM was extended with the first code oligonucleotide (150 pM final) for 5 cycles. After this first extension, a second code (at 150 pM) and a competitor oligonucleotide (at 600

pM) were added and the solution subjected to thermocycling. The addition of code and competitor was performed 5 times with thermocycling between each step. At each step, competitor concentration increased 10-fold to a final concentration of 155 nM. Libraries were prepared for sequencing using standard protocols: sequencing barcodes were added using Phusion polymerase with 15 cycles, and the products were purified via AmpureXP magnetic beads.

The products were then analyzed using the Agilent Bioanalyzer (DNA High Sensitivity) and sequenced on a MiSeq.

Oligo identities are listed in Supplemental Table 1.

Calculations for theoretical performance:

```
rmultd=c(80, 800,800,8000,80000,80000,rep(1,2^18-6))
samples=c( 4742)
REPS=1000
n.geq80=rep(0,REPS)
for (i in 1:REPS){
y=rmultinom(1,samples,rmultd)
n.geq80[i]=sum(y[7:2^18]>=y[1])
# result: 6%
}
```

Publication Acknowledgement As with all Stanford Service Centers, credit must be given to the Stanford Functional Genomics Facility for data that results in a publication. If the work done at SFGF produces data resulting in a figure in a publication, you are required to acknowledge the Stanford Functional Genomics Facility in the publication. Further, if SFGF staff members provided significant experimental design, data interpretation, or other intellectual contribution (as evaluated by the PI), then it is expected that these individuals will be coauthors on the publication.

4 Supplement

4.1 RNAseq fold change

Here, we systematically profile two performance metrics for SQUICH: (1) detection of low abundance species in a background of high abundance species, ie drop-out; (2) L_2 loss of estimated fold change between estimated abundance and ground truth.