Predicting the molecular complexity of sequencing libraries

Timothy Daley[1] & Andrew D Smith[2]

Predicting the molecular complexity of a genomic sequencing library is a critical but difficult problem in modern sequencing applications. Methods to determine how deeply to sequence to achieve complete coverage or to predict the benefits of additional sequencing are lacking. We introduce an empirical Bayesian method to accurately characterize the molecular complexity of a DNA sample for almost any sequencing application on the basis of limited preliminary sequencing.

Modern DNA sequencing experiments routinely interrogate hundreds of millions or even billions of reads, often to achieve deep coverage or to observe very rare molecules. Low-complexity DNA sequencing libraries are problematic in such experiments: many sequenced reads will correspond to the same library molecules, and deeper sequencing will either provide redundant data or introduce biases in downstream analyses. When sequencing depth appears insufficient, investigators must decide whether to sequence more deeply from an existing library or to generate another library. If this situation is anticipated during experimental design, investigators can plan to select from several libraries or samples for deep sequencing on the basis of preliminary 'shallow' surveys. The underlying question is how much new information will be gained from additional sequencing? The Lander-Waterman model[1] was essential to understanding traditional Sanger sequencing experiments but does not account for the various biases typical in applications of high-throughput sequencing.

We present an empirical Bayes method for predicting the molecular complexity of sequencing libraries or samples on the basis of data from very shallow sequencing runs. We define complexity as the expected number of distinct molecules that can be observed in a given set of sequenced reads[2]. This function, which we call the complexity curve, efficiently summarizes new information to be obtained from additional sequencing and is generally robust to variation between sequencing runs (**Supplementary Note**). Our method also applies to understanding the complexity of molecular species in a sample (for example, RNA from different isoforms), and as we require no specific assumptions about the sources of biases, our method is applicable in a variety of contexts (**Supplementary Note**).

Consider a sequencing experiment as sampling at random from a DNA library. Distinct molecules in the library have different probabilities of being sequenced, which we assume will change very little upon deep sequencing of the same library. Our goal is to accurately estimate the number of previously unsequenced molecules that would be observed if additional reads were generated.

Capture-recapture statistics has dealt with analogous questions of estimating animal population size or species diversity[3]. We borrow the Good and Toulmin model[4], a classic Poisson nonparametric empirical Bayes model, for the problem of read sampling. On the basis of the initial sequencing experiment, we identify unique molecules by some unique molecular identifier[5] and obtain the frequency of each unique observation (for example, each genomic position, transcript, allele and others). Using these frequencies, we estimate the expected number of molecules that would be observed once, twice and so on, in an experiment of the same size from the same library. The formula takes the form of an alternating power series with the estimated expectations as coefficients (we provide the full derivation in Online Methods).

The power series is extremely accurate for small extrapolations, but major problems are encountered when attempting to extrapolate past twice the size of the initial experiment[4]. At that point, the estimates show extreme variation depending on the number of terms included in the sum. Technically, the series is said to diverge and therefore cannot be used directly to make inferences about properties of experiments more than twice as large as the initial experiment. Methods traditionally applied to help these series converge in practice, including Euler's series transformation ETR[6], are not sufficient when data are on the scale produced in high-throughput sequencing experiments or for long-range predictions.

We investigated a technique called rational function approximation (RFA), which is commonly used in theoretical physics[7]. Rational functions are ratios of polynomials that, when used to approximate a power series, often have a vastly increased radius of convergence. Algorithms to fit an RFA essentially rearrange the information in the coefficients of the original power series, under the constraint that the resulting rational function closely approximates the power series. The convergence properties of RFAs are known to be especially good for a class of functions that includes the Good-Toulmin power series (**Supplementary Note**).

By combining the Good-Toulmin power series with RFAs, we developed an algorithm that can make optimal use of information from the initial sample and accurately predict the properties of sequencing data sets several orders of magnitude larger than the

[1]Department of Mathematics, University of Southern California, Los Angeles, California, USA. [2]Molecular and Computational Biology, University of Southern California, Los Angeles, California, USA. Correspondence should be addressed to A.D.S. (andrewds@usc.edu).

RECEIVED 23 OCTOBER 2012; ACCEPTED 18 JANUARY 2013; PUBLISHED ONLINE 24 FEBRUARY 2013; DOI:10.1038/NMETH.2375

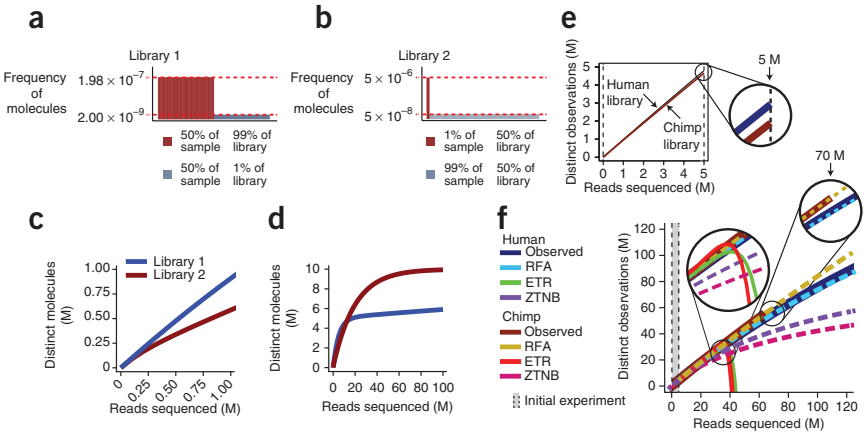

Figure 1 | Difficulties in predicting library complexity from initial shallow sequencing. (**a**,**b**) Frequency of molecules in two hypothetical libraries containing 10 million distinct molecules. Half of the molecules (5 million) make up 99% of library 1 (**a**), whereas only 10,000 molecules make up half of library 2 (**b**). (**c**,**d**) Observed complexity of both libraries after a shallow (1 million (M) reads) sequencing run (**a**) and after deeper sequencing (**b**). (**e**) Initial observed complexity from 5 M sequenced reads for human and chimp DNA bisulfite sequencing libraries. (**f**) Observed complexity after additional sequencing and predicted complexity based on the initial experiments in **e**. Observed complexity curves cross, with the chimp sperm sample yielding more distinct observations after sequencing 22 M reads. Estimates using RFA and ETR predict crossing (though ETR becomes unstable), whereas ZTNB does not.

initial 'shallow' sequencing run. We implemented our methods as a command-line software package licensed under the General Public License and available as **Supplementary Software** or at http://smithlab.usc.edu/software/librarycomplexity/.

We use an example to illustrate how naive analysis can lead to incorrect complexity predictions (**Fig. 1a–d**). In the example, two hypothetical libraries have complexity curves that initially appear linear (**Fig. 1c**) but eventually cross (**Fig. 1d**). Such extreme behavior can actually arise in practice. As the initial sample, we used a small sample of reads from human and chimp sperm bisulfite sequencing experiments[8] (**Supplementary Table 1**) and produced complexity curves for the libraries (**Fig. 1e**). Both complexity curves appear linear in the initial 5-million-read experiment, with the chimp library curve exhibiting a lower trajectory; on the basis of this information, a naive analysis might predict that the chimp library would saturate first. However, the complexity curves cross after deeper sequencing (at 22 million reads; **Fig. 1f**), with the chimp library yielding more distinct observations.

On the basis of the initial sample of 5 million reads, we estimated the complexity of these two libraries using the RFA as well as ETR applied to the Good-Toulmin power series and a zero-truncated negative binomial (ZTNB). The ZTNB is commonly used to model count data that are not Poisson-distributed, and ETR is the traditional method used to improve convergence of the Good-Toulmin series. ETR initially gives accurate estimates, but these diverge and are useless after 40 million reads. The ZTNB estimates exhibit a substantial downward bias (more than 35% error for both libraries) and do not predict that the complexity curves cross, indicating that this distribution does not account for library biases. The RFA estimates the complexity of both libraries almost perfectly. For the human library, this amounts to extrapolating to over 30 times the size of the initial sample while only incurring 4% error (**Supplementary Table 2**).

In sequencing applications that identify genomic intervals such as protein-binding sites in chromatin immunoprecipitation and sequencing (ChIP-seq) or expressed exons in RNA sequencing (RNA-seq), the number of distinct molecules in the library may be of secondary interest to the number of distinct genomic intervals identified after processing mapped reads. To demonstrate the broad applicability of our method (discussed in the **Supplementary Note**), we investigated how well our method could predict the number of non-overlapping genomic windows identified in a ChIP-seq experiment (1 kilobase) and an RNA-seq experiment (300 base pairs) using an initial 5 million reads. We used non-overlapping windows for simplicity, but more sophisticated methods of identifying binding sites or exons are equally applicable. For the ChIP-seq experiment (CCCTC-binding factor; mouse B cells[9]), the number of distinct reads did not reach saturation even after sequencing 90 million reads (**Fig. 2a**), whereas the number of identified windows saturated after approximately 25 million reads (**Fig. 2b**). RFA predicted this saturation correctly (**Fig. 2b**) and estimated library complexity with very high accuracy (**Fig. 2a**). The ZTNB overestimated the saturation of identified windows at 4 million reads, more than possible in the mouse genome and severely underestimated the yield of distinct reads. The RNA-seq experiment (Human adipose-derived mesenchymal stem cells[10]) did not saturate for

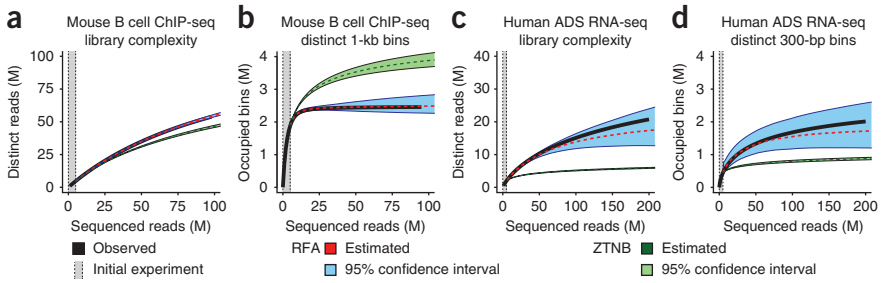

Figure 2 | Library complexity can be estimated in terms of distinct molecules sequenced or distinct loci identified. (**a–d**) Comparison of estimated complexity curves for RFA and ZTNB using 5 million (M) sequenced reads from an initial ChIP-seq experiment (CCCTC-binding factor; mouse B cells; **a**,**b**) and an RNA-seq experiment (human adipose-derived mesenchymal stem cells (ADS); **c**,**d**). Complexity is defined in terms of distinct observed reads in **a** and **c**, and in terms of distinct occupied nonoverlapping bins in **b** and **d**.

either distinct reads (**Fig. 2c**) or identified windows (**Fig. 2d**), suggesting additional sequencing from this library would yield more information. Only the RFA accurately predicted absence of saturation for both windows and reads, with considerably lower relative error than the ZTNB at 200 million sequenced reads (**Supplementary Table 3**).

Sequencing data will always be subject to some amount of technical variation between sequencing instruments or even between runs on the same machine. We applied our method to data from a single library sequenced on different instruments (using slightly differing sequencing technologies) and found that complexity estimates were within the range expected owing to stochastic noise (**Supplementary Fig. 1**). To have an impact on library complexity estimates, run-to-run variation must be dramatic and would likely be caused by detectable sequencing error at levels sufficient to warrant discarding the run.

As the cost, throughput and read lengths of sequencing technologies improve, the usefulness of methods for understanding molecular complexity in a DNA sample will increase. The approach we described, which is based on RFA to the power series of Good and Toulmin, can be applied to an immense diversity of sequencing applications (**Supplementary Note**). As the age of clinical sequencing approaches, substantial resources will be dedicated to refining quality control, protocol optimization and automation; methods for evaluating libraries will be essential to controlling costs and interpreting the results of sequencing that potentially could inform clinical decisions.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Supplementary information is available in the online version of the paper.

ACKNOWLEDGMENTS

We thank S. Tavaré, M. Waterman, P. Calabrese, G. Hannon, and members of the Hannon lab and the Smith lab for their help, advice and input. This work was supported by US National Institutes of Health National Human Genome Research Institute grants (R01 HG005238 and P50 HG002790).

AUTHOR CONTRIBUTIONS

T.D. and A.D.S. designed the method, implemented the software, performed the analysis and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Lander, E. & Waterman, M. *Genomics* **2**, 231–239 (1988).
2. Chen, Y. *et al. Nat. Methods* **9**, 609–614 (2012).
3. Fisher, R.A., Corbet, S. & Williams, C.B. *J. Anim. Ecol.* **12**, 42–58 (1943).
4. Good, I.J. & Toulmin, G.H. *Biometrika* **43**, 45–63 (1956).
5. Kivioja, T. *et al. Nat. Methods* **9**, 72–74 (2012).
6. Efron, B. & Thisted, R. *Biometrika* **63**, 435–447 (1976).
7. Baker, G. & Graves-Morris, P. *Pade Approximants* (Cambrige University Press, Cambridge, UK, 1996).
8. Molaro, A. *et al. Cell* **146**, 1029–1041 (2011).
9. Ribeiro de Almeida, C. *et al. Immunity* **35**, 501–513 (2011).
10. Lister, R. *et al. Nature* **471**, 68–73 (2011).

ONLINE METHODS

Modeling the sequencing process. Consider a sequencing experiment in which a total of $M = tN$ reads are sequenced, for some $1 < t < \infty$ and $N < M$. We fix a size N subset of the reads and refer to this subset as the initial experiment. The remaining $(t - 1)N$ reads are sequenced from the same library in the extended experiment and we refer to the union of the initial and extended experiment as the complete experiment. Although our terminology might suggest that the initial and extended experiments are conducted separately and at different times, this is not necessary, and we only require that the properties of the library are unchanged between the initial and extended experiments. Concretely, we assume the total number of distinct genomic molecules contained in the library remains constant as well as the underlying properties of the library, such as the relative frequencies of the fragments. We seek to use the information obtained during the initial experiment to determine properties of the complete experiment, in particular the number of molecules sequenced in the complete experiment that were unsequenced in the initial experiment.

Estimating the yield of a library. The following derivation closely follows that in ref. 6. Let $n_j(t)$ denote the number of molecules sequenced j times in the complete experiment, and let $n_j = n_j(1)$ be the number of molecules sequenced j times in the initial experiment. Assume that L is the unobserved true total number of distinct molecules in the library and $\pi = \{\pi_i;\ i = 1,\ldots, L\}$ are the probabilities, for each molecule in the library, that a sequenced read corresponds to that molecule. Furthermore assume that $\pi_i > 0$ for all i, so that L includes only those molecules that are in the library and can be observed and identified. Define $\lambda_i = N\pi_i$ and assume each λ_i is independently and identically distributed according to distribution $\mu(\lambda)$ with finite second moment. We assume the number of reads observed from molecule i follows a Poisson process with rate λ_i. Therefore if tN reads are sequenced, the expected number of molecules observed j times is

$$E(n_j(t)) = L\int_0^\infty e^{-\lambda t}(\lambda t)^j / j!\, d\mu(\lambda) \tag{1}$$

Let $\Delta(t)$ denote the marginal yield between the initial and complete experiments, equal to the increase in the number of distinct observed reads resulting from the extended experiment. This is equal to the expected number of unobserved reads after the initial experiment minus the expected number of unobserved reads following the complete experiment:

$$\begin{aligned}\Delta(t) &= E(n_0(1)) - E(n_0(t)) \\ &= L\int_0^\infty e^{-\lambda}(1 - e^{-\lambda(t-1)})\, d\mu(\lambda) \\ &= \sum_{j=1}^{\infty} (-1)^{j+1}(t-1)^j E(n_j(1))\end{aligned} \tag{2}$$

The last equality is obtained by expanding $1 - \exp(-\lambda(t - 1))$ as a power series centered at $t = 1$, reordering the integration and summation, and invoking identity (1). As the observed frequency of count j is always an unbiased estimator for $E(n_j(1))$ regardless of the distribution μ

$$\Delta(t) = \sum_{j=1}^{\infty} (-1)^{j+1}(t-1)^j n_j \tag{3}$$

is an unbiased estimator for the marginal library yield. Therefore, an unbiased estimator of the total library yield can be calculated by adding the marginal yield and the observed number of distinct reads in the initial experiment. The case of extrapolating to $t = \infty$ is equivalent to estimating the library size L, which is a desirable quantity, but is unidentifiable and therefore has no unbiased estimator without additional assumptions[11,12].

This elegant result, originally derived in the 1950s (ref. 4), presents substantial difficulties in direct practical application. Equation (3) is only guaranteed to converge for $t \leq 2$, which corresponds to extrapolating to only twice the number of observations in the initial experiment. Accordingly, most applications of this formula are restricted to this range. Applications associated with deep sequencing, however, require accurate estimates far outside this range, implying we need to increase the radius of convergence of the power series. Previously suggested methods, such as ETR, do not perform well for large values of t, even for experiments significantly smaller than a typical sequencing experiment[13].

Rational function approximation. ETR is a common tool in increasing the radius of convergence for divergent series or to speed up the rate of convergence for difficult to sum series, particularly alternating series[14]. The transformed series is a power series in the variable $u = 2(t - 1)/t$ (**Supplementary Note**), so that the transformed series form is a rational function in t. We hypothesized that more accurate results would be obtained by considering approximations in this larger class. For a given function and its power series, the coefficients of the optimal RFA of fixed degrees P and Q (**Supplementary Note**, equation 2), are determined by requiring the first P + Q + 1 coefficients of the associated power series to be equal to the first P + Q + 1 coefficients of the original power series (equation 3)[7]. The fundamental idea is that the initial experiment gives us reliable information on how the function behaves in a neighborhood around $t = 1$; our approximation should use this information while remaining globally well-behaved.

RFAs can be shown to converge for a wide variety of functions[7] but are more useful for approximating certain classes of functions. One such class is the set of alternating power series with coefficients arising from moments of a positive measure on the real line, including the familiar probability measures. Such power series are often called series of Stieltjes[15]. The power series of equation (2) falls within this class when the true expected frequencies, $E(n_j(1))$, are known exactly. Clearly we can only estimate the $E(n_j(1))$, which we do by counting reads. However, the absolute amount of information contributing to our estimates of the $E(n_j(1))$ is extremely large, especially when we compare the number of reads sequenced in a small sequencing run (that is, millions) with the numbers arising from traditional capture-recapture experiments (for example, hundreds[13] or thousands[6] of captures).

When applied to Stieltjes series, RFAs have the fascinating property that the direction of their convergence depends on the

doi:10.1038/nmeth.2375

relationship between the order of the polynomials in the numerator and denominator[15]. If the order of the approximation (P + Q + 1) is odd then convergence is from below. For a fixed *t*, successive approximations obtained from increasing both P and Q by one increase toward the true value. This implies that any odd order approximation will be conservative when the $E(n_j(1))$ are known exactly. If the order is even then the convergence is from above and the resulting approximations will be liberal (**Supplementary Fig. 2**). We can therefore choose to only consider odd order approximations and our estimates will tend to be conservative.

Two common and equivalent implementations of RFAs are Padé approximants and truncated continued fractions. We chose to implement the approximations using continued fractions for several reasons. First, computing the coefficients for a truncated continued fraction expansion is both asymptotically faster and numerically more stable than directly computing the Padé approximant coefficients. Using the quotient-difference algorithm, the coefficients of the continued fraction can be computed in time that is a quadratic function of the degree of the truncated power series being approximated[16]. The straightforward methods typically associated with Padé approximants require inverting a matrix (often ill-conditioned), which requires cubic time and may be numerically unstable (**Supplementary Note**).

Second, representing rational functions as continued fractions easily circumvents direct evaluation of the high-order polynomials in the numerator and the denominator. Using Euler's recursion with renormalization evaluates continued fractions exactly while ensuring the magnitudes of intermediate values remain manageable[17].

Finally, and most importantly for our application, the continued fraction representation provides a natural means of exploring several very similar RFAs to the same original power series. When instabilities arise in the approximations (**Supplementary Note**), they can usually be avoided by adjusting the order of the numerator and denominator of the rational function. Using the continued fraction representation, such adjustments can be made without recomputing coefficients: lower-order approximants are obtained by successive truncation of a high-order continued fraction.

11. Link, W. *Biometrics* **59**, 1123–1130 (2003).
12. Mao, C. & Lindsay, B. *Ann. Stat.* **35**, 917–930 (2007).
13. Keating, K., Quinn, J., Ivie, M. & Ivie, L. *Ecol. Appl.* **8**, 1239–1249 (1998).
14. Hardy, G. *Divergent series* (Oxford University Press, London, 1949).
15. Simon, B. *Adv. Math.* **137**, 82–203 (1998).
16. McCabe, J.H. *Math. Comput.* **41**, 183–197 (1983).
17. Blanch, G. *SIAM Rev.* **6**, 383–421 (1964).