Supplementary Note

Software

The software preseq is available as Open Source and can be downloaded from **Supplementary Software** or:

`http://smithlab.usc.edu/software/`

The software has been tested on Linux and Mac OSX, but is written in C++ and should be easily ported. The GNU Scientific Library (GSL) is required.

Methods

Contexts in which our model applies

We assume a sequencing experiment randomly samples molecules from the library and that the sampled molecules are observed as sequenced reads. In general we are interested in determining when two reads give different information. In evaluating our method, most of our applications focused on whether sequenced reads came from distinct fragments in the library. In those applications we counted the number of distinct molecules observed as a function of the number of reads sequenced. We required a method to identify when two reads correspond to the same or distinct molecules – such methods have been called "unique molecular identifiers" (UMIs) [14] and popular UMIs include random barcodes or distinct mapping locations. We used distinct mapping locations as the UMI, but this method will be problematic when the sequencing experiment is such that distinct molecules are likely to map to the same location. This happens when sequencing very deeply (as often happens for smaller genomes), or when the nature of the experiment enriches for reads from specific genomic locations, for example in "deep CAGE" when reads come from the 5' ends of transcripts.

Considering the role of the identifier more abstractly is helpful to understand the broad applicability of our method. In the terminology of capture-recapture, the identifier groups captured individuals (the sequenced reads) into classes. This interpretation, along with the generality of the capture-recapture theory, allows our method to be used in a great many contexts. When we examined the ChIP-seq and RNA-seq data to investigate how much additional information would come from additional sequencing (see **Fig. 2**) we also counted distinct genomic windows to which reads mapped. The identifier of a read was the genomic window containing the mapped location of the read. If one were interested only in counting the number of distinct exons from a pre-defined set of exons, then one would associate each mapped read with the exon to which they map. Those reads not mapping to an exon would be ignored, and in a technical sense we would say those are "non-identifiable" reads. Ultimately, all that is required to apply our method in a new sequencing context is some means of determining when a newly sequenced read provides additional information when compared with those already sequenced.

The underlying physical process and potential sources of variation

We assume that in each sequencing experiment the number of reads corresponding to each unique molecule is determined by a probability associated with that molecule. We usually want those probabilities to be proportional to the frequencies of each molecule in the original population. There are numerous potential sources of bias that alter these probabilities. We divide all potential sources of bias into the following three groups to understand how they impact our assumptions.

The type of experiment Prior to sequencing there is often some specific type of experiment conducted on the population of molecules. These experiments alter the proportions of molecules in both intended and unintended ways. One example of intentional alterations include the ChIP step in a ChIP-seq experiment, which results in extreme enrichment for DNA molecules interacting with some protein. Another example is the poly-A purification in RNA-seq, which enriches for mRNA. One example of unintentional alteration to the proportions is the degradation of DNA from bisulfite treatment in BS-seq experiments, which for some protocols appears to have different affects on different parts of the genome.

Library preparation One hopes that the process of library preparation does not alter the proportions of molecules too drastically, but the procedures involved in constructing libraries (e.g. adaptor ligation, amplification, etc.) can have a major impact. These effects are captured by our method, because they are not changed if the library is sequenced twice. We also expect that in the future the variation introduced by the library preparation process will gradually decrease as technology improves.

The sequencing runs The prepared sequencing libraries contain multiple copies of each unique molecule from the original biological sample. We assume that in a given sequencing run, these are sampled uniformly at random, and this is why we refer to "library complexity" as reflected in reads produced by the sequencer. The sequencing process, however, does not sample uniformly from the library. There are two kinds of bias we must consider. First, there is bias that is consistent from one sequencing run to the next, the nature of which depends on the sequencing technology. This kind of bias presents absolutely no problem for our assumptions. Second, there is bias that changes between experiments: the technical variation in sequencing, between runs and between lanes in a given run. In theory these forms of bias are problematic for our model. In practice, we have observed that they have *virtually no impact*, and when they do have an effect it will typically be clear that a sequencing run has encountered problems. Investigations into the technical variance of RNA-seq experiments validate this assumption[6,17]. Note that we are not claiming the variation between sequencing runs will have no impact on the results of the experiment, only that this variation will have very little impact on observed and estimated library complexity between runs when using our method. We obtained library complexity estimates for a single library run on two different machines (in fact, one was an Illumina HiSeq and the other a GAII, **Supplementary Fig. 1**). A sample of reads from the HiSeq run almost perfectly predicts the complexity observed in the full GAII, and vice versa.

Approximating library yield with rational functions

Recall that we derived an unbiased estimator for the marginal yield given by Equation (2) in **Online Methods**:

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

We take rational function approximations of Equation (1) as our estimates for the expected marginal yield.

In the following, we shall let $f(t) = \sum_{j=0}^{\infty} f_j t^j$ be an arbitrary power series. Two common implementations are Padé approximants and truncated continued fractions. A Padé approximant to a power series centered at zero is the unique ratio

$$f(t) \approx p(t)/q(t) = \frac{p_0 + p_1 t + \cdots + p_P t^P}{1 + q_1 t + \cdots + q_Q t^Q}, \tag{2}$$

such that the associated power series agrees with $\sum_{j=0}^{\infty} f_j t^j$ for the first $P + Q + 1$ coefficients. Similarly a truncated continued fraction approximation to a power series is of the form:

$$f(t) \approx \cfrac{a_0}{1 + \cfrac{a_1 t}{1 + \cfrac{\vdots}{1 + a_C t}}}$$

The associated power series agrees with $f(t)$ for the first $C+1$ coefficients. We refer the values $P+Q+1$ and $C+1$ as the the order of the approximations. Since Padé approximants are unique[3], these two forms of optimal rational function approximations are equivalent.

The two representations differ in the ease with which they can be estimated and evaluated. Typical methods to calculate the Padé coefficients involve solving a system $Ax = b$ where A is a Hankel matrix and therefore commonly ill-conditioned[10]. The time complexity for computing the coefficients is $\Theta(Q^3 + P^2)$[3]. This complexity is not desirable if it must be computed frequently, as when large numbers of bootstrap samples are required. Furthermore, direct evaluation of rational functions in the usual representation as a pair of polynomials (*i.e.* the numerator and the denominator of Equation (2)) can be problematic for large t and large degree because the intermediate values can grow independently of the final value.

The coefficients $\{a_0, \ldots, a_C\}$ of the continued fraction representation, on the other hand, can be fit using recursive algorithms like the quotient difference[18] and product-difference algorithms[13]. Each of these takes $\Theta(k^2)$ time to approximate an order k polynomial. By avoiding the inversion of an ill-conditioned matrix the computation of the required coefficients is also more numerically stable. Finally, evaluating the rational function when represented as a continued fraction is more numerically stable as it can be done using Euler's recursion with renormalization[4].

If the observed coefficients of the original power series arise from moments of a measure defined on the positive real line, then the measure is called a Stieltjes measure and the associated power series is called a series of Stieltjes. The associated rational function approximations can be shown to converge and exhibit additional properties that make their application useful. It can be shown that the approximations where the difference between the degrees of the numerator and denominator is even converge from above, while if this difference is odd the convergence is from below[19] in some neighborhood of zero (**Supplementary Fig. 2**). If the difference is greater than or equal to -1, then the convergence holds for all positive values. We can therefore choose to err on the conservative or liberal side when appropriate.

Instabilities in rational function approximations

The moments of the Stieltjes measure assumed to generate Equation (1) are equal to the expectations of $n_j(1)$, $j = 1, 2, \ldots$ We are using observed estimates of the moments which take the form of random Poisson variables with mean equal to the true moment[9]. By breaking up the observed coefficients into their expected values plus mean zero error terms, we can see that the resulting series is not necessarily Stieltjes. The error terms cause the series to have positive measure on negative values. In practice this is more likely to happen for high frequencies. We commonly observe these as equal to zero but they actually have positive, though small, expectation. Errors in the observed coefficients of the original power series will typically result in poles in the rational function approximation with corresponding zeros in a neighborhood, resulting in approximate cancellation outside a small neighborhood[2]. This phenomenon will be transitory as the order of the approximation changes[11,12] . An additional advantage of the continued fraction representation over the Padé approximation is that we can easily identify the locations of potential defects by using a necessary condition on individual coefficients of the continued fraction. We can then evaluate the continued fraction in neighborhoods of selected points to check for defects, rather than evaluating the rational function through the entire domain. If any of these potential defects is found to actually be a defect, we immediately know the depth at which the continued fraction should be truncated to remove that defect, so we can adjust the continued fraction, erring on the conservative side, without refitting.

Computation and evaluation of continued fraction approximations

Rational function approximations require that the coefficients of the estimated power series be non-zero in order to obtain consistent estimates. We therefore truncate the series (1) to the lowest order zero coefficient before computing the coefficients. In addition, we factor out a power of $t - 1$ so that the power series we approximate has a constant for the lowest order term. When we do any evaluation, we shall evaluate the continued fraction approximation and multiply by $t - 1$.

We use Stieltjes fractions in our approximation, which are of the form

$$(t-1)\cfrac{a_0}{1+\cfrac{a_1(t-1)}{1+\cfrac{\vdots}{1+\cfrac{}{1+a_C(t-1)}}}}. \tag{3}$$

If C is even, then the equivalent Padé approximation will have numerator and denominator of equal degrees. If C is odd, then the denominator will be one degree larger than the numerator. We recall that a continued fraction approximation should have power series coefficients equal to the original power series. Therefore to obtain a continued fraction with an equivalent Padé approximation with numerator d degrees larger than the denominator, we can first take out the first d coefficients, then the remaining terms will be used to form a continued fraction of the form (3), i.e.

$$(t-1)\left(n_1 - n_2\,(t-1) + \ldots + (-1)^{d-1}\,n_d\,(t-1)^{d-1} + (t-1)^d\cfrac{a_0}{1+\cfrac{a_1(t-1)}{1+\cfrac{\vdots}{1+\cfrac{}{1+a_{C-d-1}(t-1)}}}}\right).$$

We refer to the first d coefficients as the offset coefficients and the remaining as the continued fraction coefficients.

For an approximation with equivalent Padé approximant that has numerator with degree d less than the denominator, we need to work with the reciprocal series $g(t) = 1/f(t)$. The reciprocal of the continued fraction approximation of g will be the continued fraction approximation to f.

In our method, we use the reciprocal series with an offset of one to approximate the marginal yield,

$$\sum_{j=1}^{M}(-1)^{j+1}\,(t-1)^j\,n_j \approx (t-1)/\left(\frac{1}{n_1} + (t-1)\cfrac{a_0}{1+\cfrac{a_1\,(t-1)}{1+\cfrac{\vdots}{1+\cfrac{}{1+a_{M-2}\,(t-1)}}}}\right)$$

with M odd. This will ensure that our estimates stay stable for large values of t and are conservative, on average. To ensure stable estimates we require a minimum continued fraction degree of 5, i.e. $M \geq 5$. This implies that our method is only applicable to initial experiments with the first five entries of the read count histogram greater than zero. This will exclude some extremely small initial experiments.

Confidence intervals

We can treat the observed counts, $n_1, n_2, \ldots$ as Poisson random variables with estimated mean equal to the observed value. Since the counts are negatively correlated, the sum of the estimated variances of the observed coefficients is an upper bound for the true variance[9]. This bound is significantly larger than observed in practice and in most cases is unreasonably large. We must therefore resort to bootstrapping to obtain useful confidence intervals.

The time complexity of bootstrapping the histogram is on the order of the number of non-zero entries in the histogram and therefore proportional to the largest observed count. Since this may be very large for experiments such as RNA sequencing, bootstrapping the histogram a sufficient number of times to obtain accurate quantiles ($\approx$ 1000 times[8]) will take too long. There is a natural skew to the confidence intervals upward so we employ the log-normal confidence intervals as suggested by Chao[7].

Euler's series transformation

Euler's series transformation is suggested by Efron & Thisted[9] to improve the convergence of the power series (1) by taking $u = 2(t-1)/(1+(t-1))$. The transformed series then takes the form

$$\sum_{k=1}^{\infty} \xi_k u^k \text{ with } \xi_k = \sum_{j=1}^{k} \binom{k-1}{j-1} 2^{-k} (-1)^{j+1} n_j.$$

This method is particularly suited to series with exponentially decreasing coefficients. They prove that if

$$n_j = L \frac{\Gamma\left(j+\alpha^{-1}\right)}{j!\Gamma\left(\alpha^{-1}\right)} (\alpha\mu)^j (1+\alpha\mu)^{\alpha^{-1}-j}$$

(*i.e.* the read counts are the expected counts from a negative binomial (μ, α) distribution) with $\alpha \geq 1$, then the coefficients $\xi_k, k = 1, 2, \ldots$ are all strictly positive. This implies that the radius of absolute convergence is infinite after applying Euler's transformation. For more complicated distributions this is not necessarily the case and the practical application is confounded by the error introduced by estimating the true count frequencies by the observed count frequencies.

This problem is exacerbated as more terms are used since the noise to signal ratio is significantly larger for higher order terms. In practice this translates smaller radii of convergence as more terms are used. Recall that formula (1) is unbiased only if all observed coefficients are used. Therefore using less terms will result in biased estimates. In the aforementioned situation where the read counts are Negative Binomially distributed, truncating will always result in downward bias.

In the actual application of Euler's transformation, the choice of where to truncate is murky. Efron & Thisted choose to truncate the number of terms to ten based upon an examination of the transformed coefficients. Since this must be done on a case by case basis, examining the coefficients is unfeasible for our experiments. We examine the effect of truncation upon the estimates and we observe that if less less terms are used, then the estimates tend to be biased downward, though stable. This means that initial experiments that are small, more uniform, and have fewer terms in the read count histogram tend to perform better with this method. This is not always the case, as we observe both positive and negative bias in our applications (**Fig. S1**). There seems to be no clear pattern of dependency of the bias upon the order of truncation, indicating the unreliability of this method.

Zero-truncated negative binomial distribution

A popular model for the distribution of read counts is the Negative Binomial distribution[1,22], representing a generalization of the Poisson model of read counts to account for the high variance of the counts observed in sequencing experiments. Since we assume that the total number of true molecules contained in the library is unknown, we cannot differentiate between molecules that are not observed due to random chance, but are contained in the library, and molecules that are not contained in the library and will never be observed. We must make inferences only upon the observed read counts and therefore the observed read counts are distributed according to a Zero-truncated Negative Binomial distribution with

$$\Pr(X=j) = \frac{1}{(1+\alpha\mu)^{\alpha^{-1}} - 1} \frac{\Gamma\left(j+\alpha^{-1}\right)}{\Gamma(j+1)\Gamma\left(\alpha^{-1}\right)} \left(\frac{\alpha\mu}{1+\alpha\mu}\right)^j.$$

Parameters are fit by an EM algorithm, with the unobserved zero counts as the missing data. The expected yield can then be calculated as

$$\Delta_{\text{ZTNB}} = D \frac{1-(1+\alpha\mu t)^{-\alpha^{-1}}}{1-(1+\alpha\mu)^{-\alpha^{-1}}},$$

with D equal to the number of distinct reads in the initial experiment.

Bias of parametric inference

Recall that the expected value of the yield of a larger experiment is equal to

$$\Delta(t) = L \int_0^\infty e^{-\lambda}(1 - e^{-\lambda(t-1)})d\mu(\lambda).$$

A parametric generalization of the Poisson model will assume a functional form of the measure $\mu(\lambda)$. For instance the ZTNB model assumes that μ follows a Gamma distribution. For fixed $t > 1$, the function $e^{-\lambda} - e^{-\lambda t}$ is concave in λ. Therefore Jensen's inequality implies that if the assumed form of μ is not sufficiently robust to accurately model the true distribution, then the estimated yield will be lower than the true value. Clearly the accuracy of the estimates will then depend on the specific parametric assumptions, a problem characterized in previous studies[5,23].

Results

We evaluated our method using 16 sequencing data sets from diverse experiment types (**Supplementary Table 1**), divided into two groups of equal size according to prior expectations about their library complexities. For experiments in group 1, which we call uniform input, the underlying population of molecules are expected to be in roughly equal frequency (e.g. full-genome resequencing). For experiments in group 2, variable input, the relative frequencies of molecules are heavily influenced by underlying biology. An example is RNA-seq, where libraries should contain molecules from genes with both high and low expression (see discussion in online methods). We used uniform random samples of reads to simulate the initial experiments. All bisulfite sequencing experiments were mapped with RMAPBS[20]. All RNA-seq experiments were mapped with RMAP and Tophat[21]. The Ion Torrent experiment was mapped with Bowtie2[15] and BWA[16] in Smith-Waterman-like mode. All other experiments were mapped with RMAP and Bowtie2. The results were similar for each library regardless of mapping software (**Supplementary Tables 1**, **2**, and **3**). We discuss the results of our method (RF), Euler's transform applied to equation 1, and estimates based on a Zero-truncated Negative Binomial model of observed read counts (ZTNB). We show the results for all libraries mapped with RMAP and RMAPBS, except for the Ion Torrent library, which we show the Bowtie2 results.

As expected, the ET method is extremely accurate at small extrapolations; otherwise the method diverges (**Supplementary Table 2**, **Figs. S2** and **S3**). The point of divergence and the direction of the divergence is impossible to predict, but convergence is guaranteed up to 2X (fold) extrapolations. We observe that ET tends to stay stable for initial experiments with smaller histograms, i.e. small initial experiments in the uniform input case. We can truncate the histogram of the larger and more variable experiments at a lower depth to obtain convergent estimates. These estimates will be biased, and as discussed above, the direction of the bias is unknown. Furthermore, convergence is not guaranteed so that this strategy is not advisable and can be seen from estimates which remain stable but have massive confidence intervals (i.e. **Fig. S2c**).

In every case examined, the RF method outperforms the ZTNB (**Figs. S2**, **S3**, and **S4**). For estimates at 50X of the initial experiment size, on uniform input data sets the RF method is found to be always within 10% of the true value, with a mean error of 5%. In contrast, the ZTNB method has an average error of over 40%. Interestingly, the ZTNB performs worse with more data (**Fig. S4**, 1M vs 5M initial experiments one-sided t-test, $p = 4.558e - 6$, $n = 18$). This is a consequence of the penalty of misspecifying the degree of bias, indicating that the ZTNB is not sufficiently robust for accurate predictions.

Due to the larger variance natural to the variable input libraries, all methods are less accurate. The RF shows similar performance to ET, when the latter converges, and still significantly outperforms the ZTNB method (14% average error vs. 63% at 50X extrapolation). The ZTNB method tends to predict saturation of the distinct molecules, with little variance in the predictions. The RF do not predict the saturation of the estimates, but tend to be accurate ($< 10\%$ error) for estimates only in the range of 20 to 30X extrapolations. For far ranging extrapolations, the RF method underestimates the distinct molecules with large variation. This is a direct consequence of the larger variation in the variable input experiments. In only one case out of the 48 total simulations does the observed complexity curve lie outside the computed confidence interval, which is expected by pure chance (2.4 expected by chance, 91% probability at least one observed curve lies outside the 95% confidence interval).

Minimum initial experiment size and maximum extrapolation

For all experiments, a minimum sample size of one million mapped reads ensured that the histogram met the criteria for our method. We noted that smaller sample sizes, particularly in the Exome-seq and DNA-seq experiments, resulted in some initial experiments that were not sufficiently diverse for our method. This was not a problem for the variable input libraries, though we did notice that they require a larger initial experiment to accurately estimate the complexity. We believe this is due to an inability to accurately observe the biases in these experiments without a sufficient number of observations. We therefore suggest initial experiment sizes of at least 1M mapped reads for uniform input experiments and 4M reads for variable input experiments. More may be required if the UMI includes random barcodes or both ends of concordantly mapped read for a paired end experiment.

With these suggested minimum initial experiment sizes we have found that our method is accurate to within 10% for extrapolations up to 100 fold of the initial sample size for the uniform input experiments. The natural variance of the variable input experiments imply that our estimates will be much more uncertain. We find that our method is accurate to within 10% for up to 30 fold extrapolation.

Effect of UMI choice on accuracy of our predictions

To demonstrate that our method can accurately predict yield from larger sequencing experiments when a different form of UMI is used we analyzed the data from Kivioja et al.[14], specifically the RNA-seq data from *Drosophila* S2 cells (ENA accession ERR048992). We mapped the data to the entire dm3 assembly using Bowtie2 with default settings, resulting in 59,953,425 uniquely mappable reads.

We conducted two separate experiments. In the first experiment, we sampled 2M reads uniformly at random (the initial experiment), and counted the frequencies of distinct molecules based on random barcodes. We then extrapolated the complexity curve and also produced the true complexity curve based on the full set of 60M mapped reads (**Fig. S5a**).

In the second experiment, we similarly sampled 2M reads uniformly at random, but this time we ignored the random barcodes and distinguished unique reads based only on mapping position. We used the frequencies to extrapolate a complexity curve, and also produced the true complexity curve that would have resulted from using the same UMI method (i.e. mapping locations) on the entire set of 60M reads (**Fig. S5b**).

Use of random barcodes is of course expected to increase the number of distinct molecules that can be observed. We can see that the random barcodes permit roughly 39M distinct molecules to be identified, and our prediction is virtually perfect (**Fig. S5a**). This is an easy case for our method, because the complexity is very high, and the curve almost linear.

When only distinct mapping locations are used, there is an inherent ceiling of approximately 168.7M molecules that can be distinguished, since that is the size of the dm3 *Drosophila* genome assembly (including the chrU and chrUextra chromosomes with the download available from the UCSC Genome Bioinformatics FTP site). Moreover, since this is RNA-seq data, and most of the reads will reflect the most abundant RNA species, the actual ceiling based on mapping location might be far lower. The true complexity curve using this UMI method is somewhat linear through the 2M range of the initial experiment, but then shows saturation. Considering the range of the 60M mapped reads, only 1.26M distinct locations are mapped. Our predicted complexity curve again follows the true curve quite closely, despite the saturating behavior.

There are two conclusions to be drawn from this study. First, if there is reason to expect the reads in a given experiment will saturate the reference to which they are being mapped, then random barcodes are really necessary to get the most from the data (as was shown by Kivioja et al.[14]). In general we expect the optimal UMI methods will be application-specific, but in many cases when saturation is not an issue the mapping locations will be sufficient.

The second conclusion to be drawn is that our method is able to accurately predict the yield in either case, and so can be used regardless of the chosen UMI method.

References

1. S. Anders and W. Huber. Differential expression analysis for sequence count data. *Genome Biol*, 11(10):R106, 2010.

2. G.A. Baker. Defects and the convergence of padé approximants. *Acta Applicandae Mathematicae*, 61(1):37–52, 2000.

3. G.A. Baker and P. Graves-Morris. *Padé approximants*, volume 59. Cambridge University Press, 1996.

4. G. Blanch. Numerical evaluation of continued fractions. *Siam Review*, 6(4):383–421, 1964.

5. D. Böhning and D. Schön. Nonparametric maximum likelihood estimation of population size based on the counting distribution. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(4):721–737, 2005.

6. J.H. Bullard, E. Purdom, K.D. Hansen, and S. Dudoit. Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments. *BMC bioinformatics*, 11(1):94, 2010.

7. A. Chao. Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, 43:783–791, 1987.

8. B. Efron. *The jackknife, the bootstrap, and other resampling plans*, volume 38. SIAM, 1982.

9. Bradley Efron and Ronald Thisted. Estimating the number of unseen species: How many words did Shakespeare know? *Biometrika*, 63(3):pp. 435–447, 1976.

10. P. Feldmann and R.W. Freund. Efficient linear circuit analysis by pade approximation via the lanczos process. *Computer-Aided Design of Integrated Circuits and Systems, IEEE Transactions on*, 14(5):639 –649, may 1995.

11. J. Gilewicz and M. Pindor. Padé approximants and noise: rational functions. *Journal of computational and applied mathematics*, 105(1):285–297, 1999.

12. Jacek Gilewicz and Yuri Kryakin. Froissart doublets in Padé approximation in the case of polynomial noise. In *Proceedings of the Sixth International Symposium on Orthogonal Polynomials, Special Functions and their Applications (Rome, 2001)*, volume 153, pages 235–242, 2003.

13. R. G. Gordon. Error Bounds in Equilibrium Statistical Mechanics. *Journal of Mathematical Physics*, 9:655–663, May 1968.

14. T. Kivioja, A. Vähärautio, K. Karlsson, M. Bonke, M. Enge, S. Linnarsson, and J. Taipale. Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods*, 9:72–74, 2012.

15. B. Langmead and S.L. Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357–359, 2012.

16. H. Li and R. Durbin. Fast and accurate short read alignment with burrows–wheeler transform. *Bioinformatics*, 25(14):1754–1760, 2009.

17. J.C. Marioni, C.E. Mason, S.M. Mane, M. Stephens, and Y. Gilad. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome research*, 18(9):1509–1517, 2008.

18. J. H. McCabe. The quotient-difference algorithm and the Padé table: an alternative form and a general continued fraction. *Math. Comp.*, 41(163):183–197, 1983.

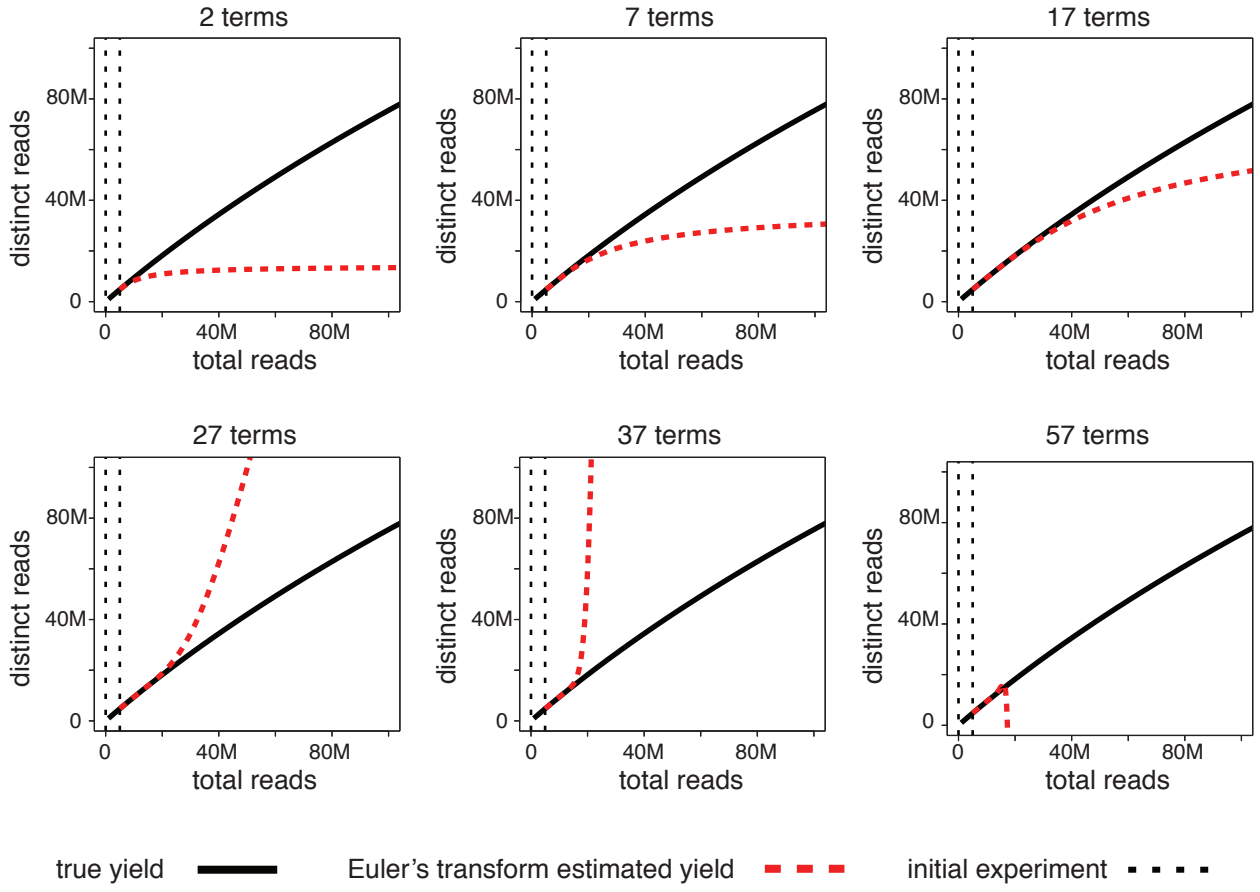
19. Barry Simon. The classical moment problem as a self-adjoint finite difference operator. *Adv. Math.*, 137(1):82–203, 1998.

20. A.D. Smith, W.Y. Chung, E. Hodges, J. Kendall, G. Hannon, J. Hicks, Z. Xuan, and M.Q. Zhang. Updates to the RMAP short-read mapping software. *Bioinformatics*, 25(21):2841–2842, 2009.

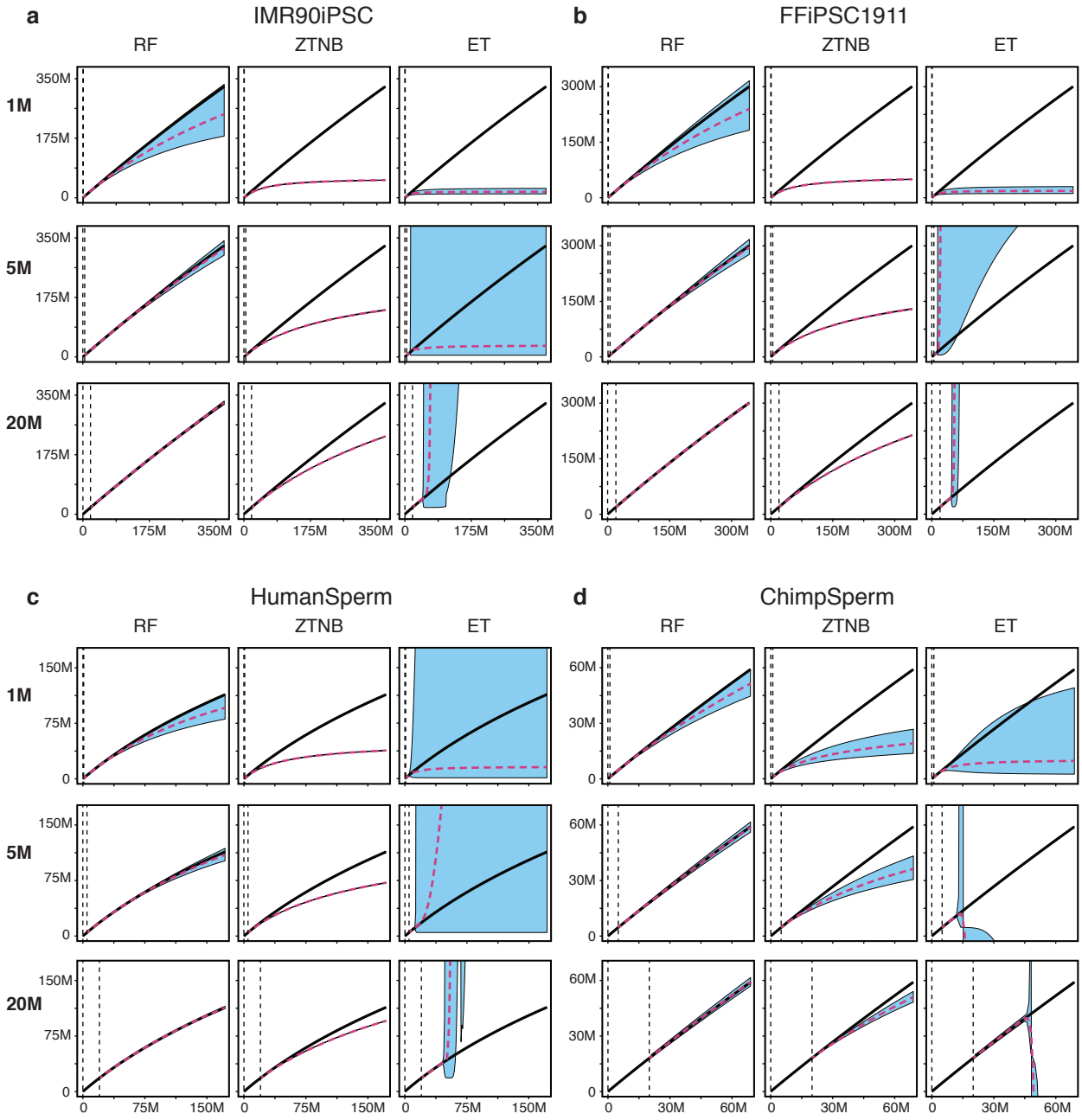
21. C. Trapnell, L. Pachter, and S.L. Salzberg. Tophat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, 2009.

22. P. Uren, E. Bahrami-Samani, S. Burns, M. Qiao, F. Karginov, E. Hodges, G. Hannon, J. Sanford, L. Penalva, and A. Smith. Site identification in high-throughput RNA-protein interaction data. *Bioinformatics*, In Press, 2012.

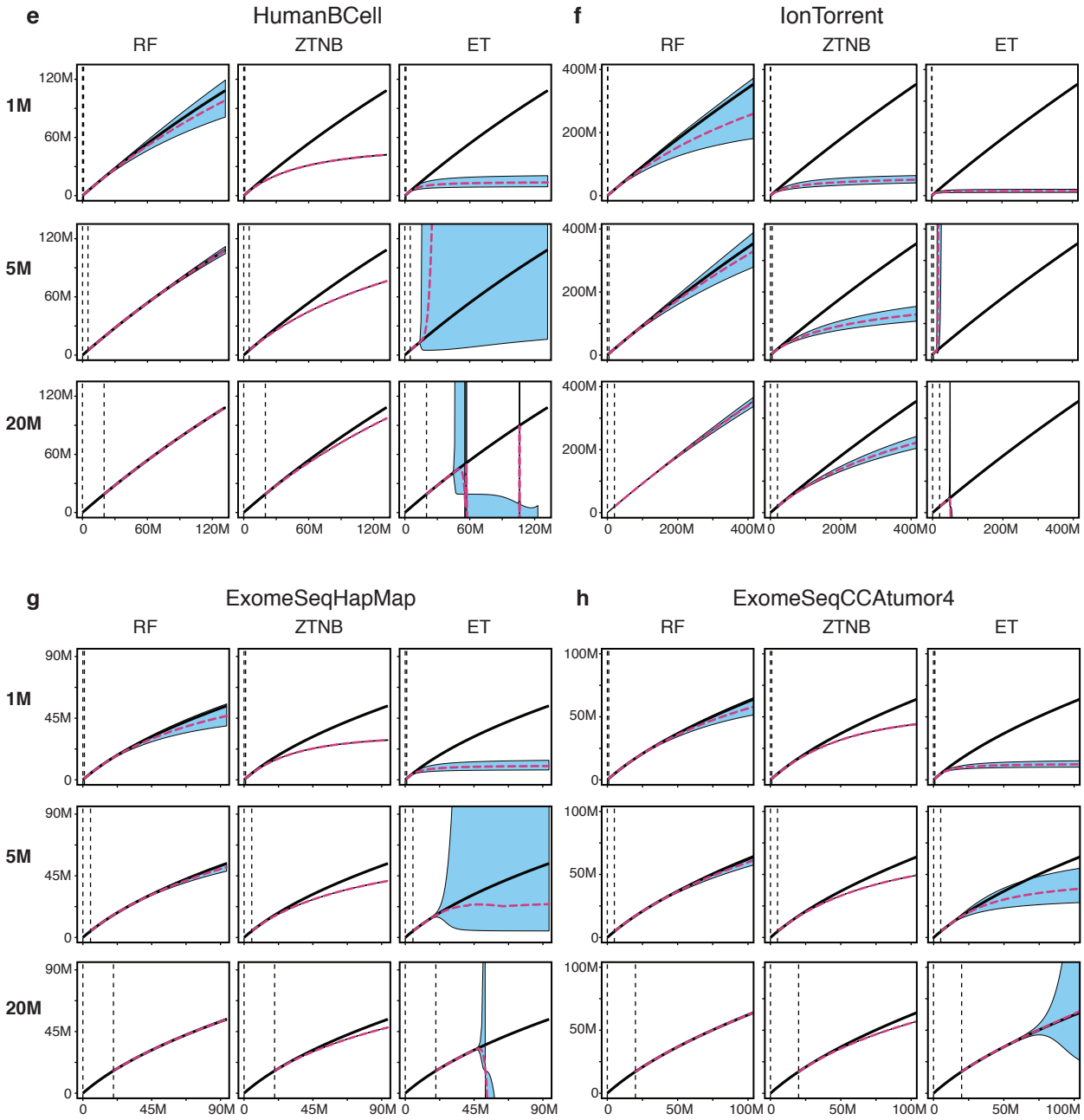
23. Ji-Ping Wang. Estimating species richness by a poisson-compound gamma model. *Biometrika*, 97(3):727–740, 2010.

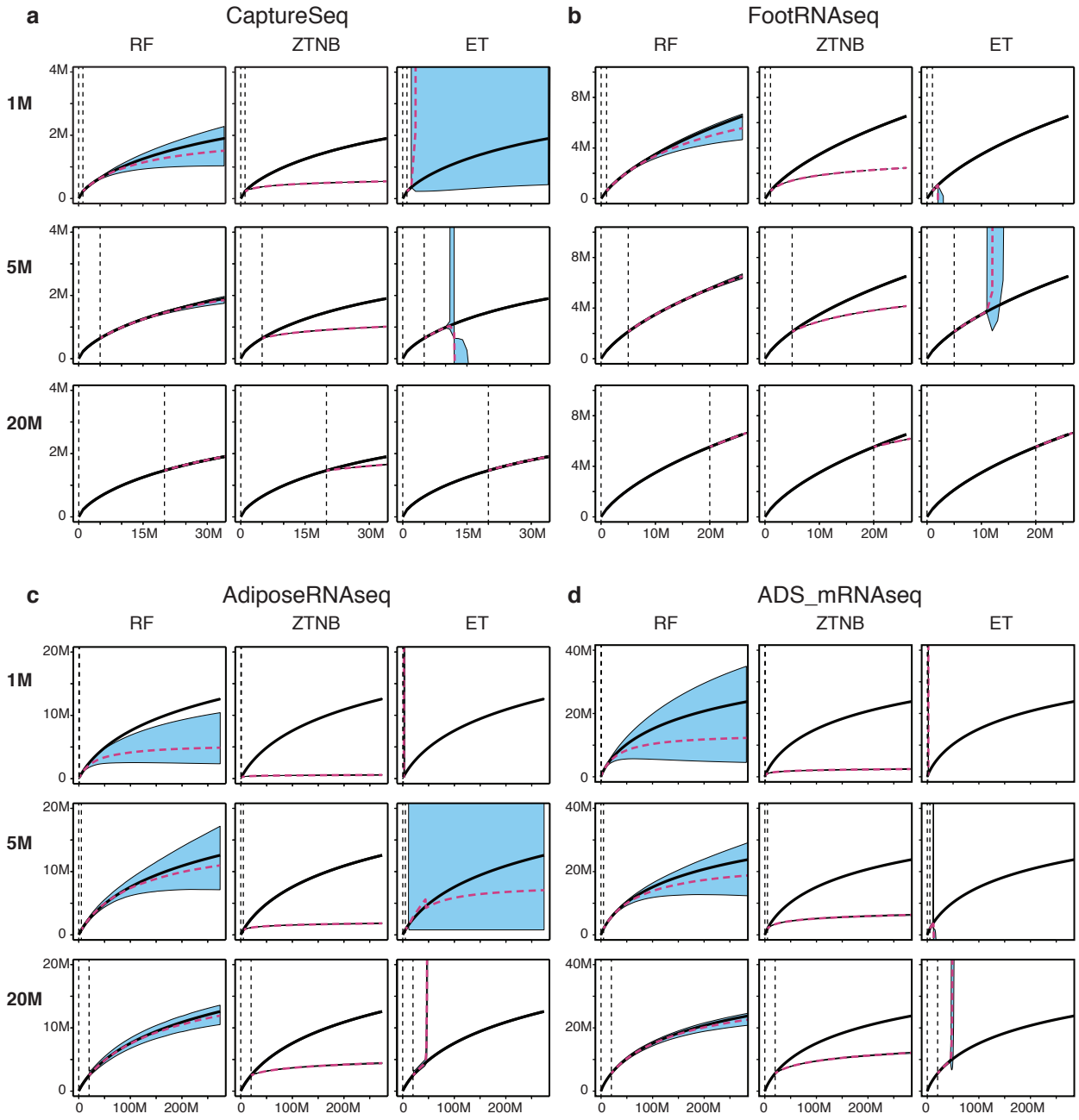


Supplementary Fig. S 1: Divergence when using Euler's transform. Estimated yield for the human sperm BS-seq library using Euler's transform applied to Equation (2) in **Online Methods** truncated at 2, 7, 17, 27, 37 and 57 terms.

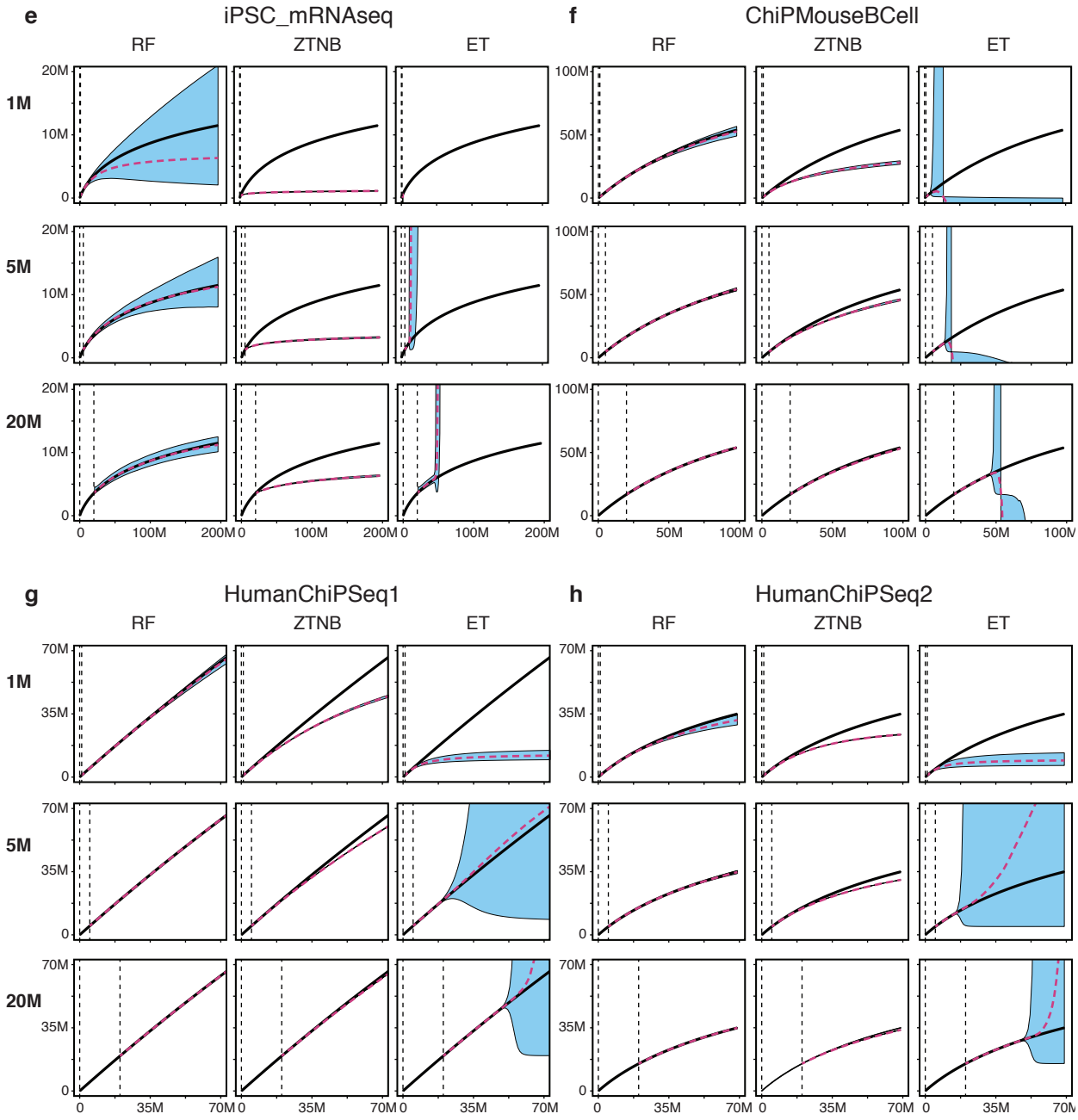


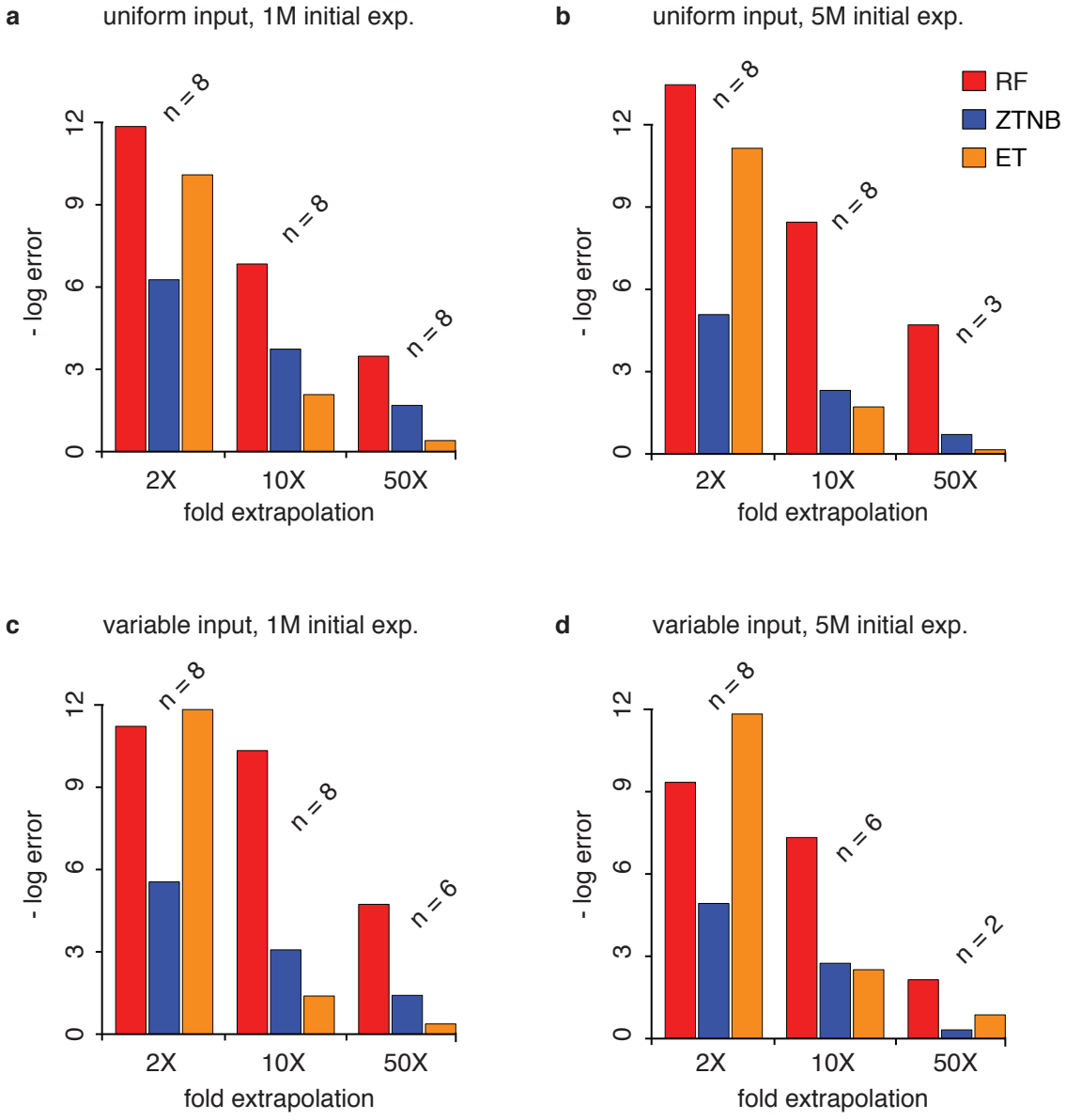
Supplementary Fig. S 2: Detailed comparison of methods predicting complexity of uniform input libraries. In each graph the observed complexity curve is plotted (solid black) along with the estimated complexity curve (dashed red) and the corresponding confidence interval. For each data set, initial experiment sizes of 1M, 5M and 20M sampled reads are presented for each of the three methods: rational function (RF), zero-truncated negative binomial (ZTNB) and Euler's transform (ET). **(a)** IMRiPSC90 BS-seq, **(b)** FFiPSC1911 BS-seq, **(c)** HumanSperm BS-seq, **(d)** ChimpSperm BS-seq, **(e)** HumanBCell BS-seq, **(f)** IonTorrent DNA-seq, **(g)** ExomeSeqHapMap, and **(h)** ExomeSeqCCAtumor4.


e
HumanBCell
f
IonTorrent
RF
ZTNB
ET
1M
5M
20M
120M
60M
0
400M
200M
0
60M
120M
200M
400M
g
ExomeSeqHapMap
h
ExomeSeqCCAtumor4
90M
45M
100M
50M

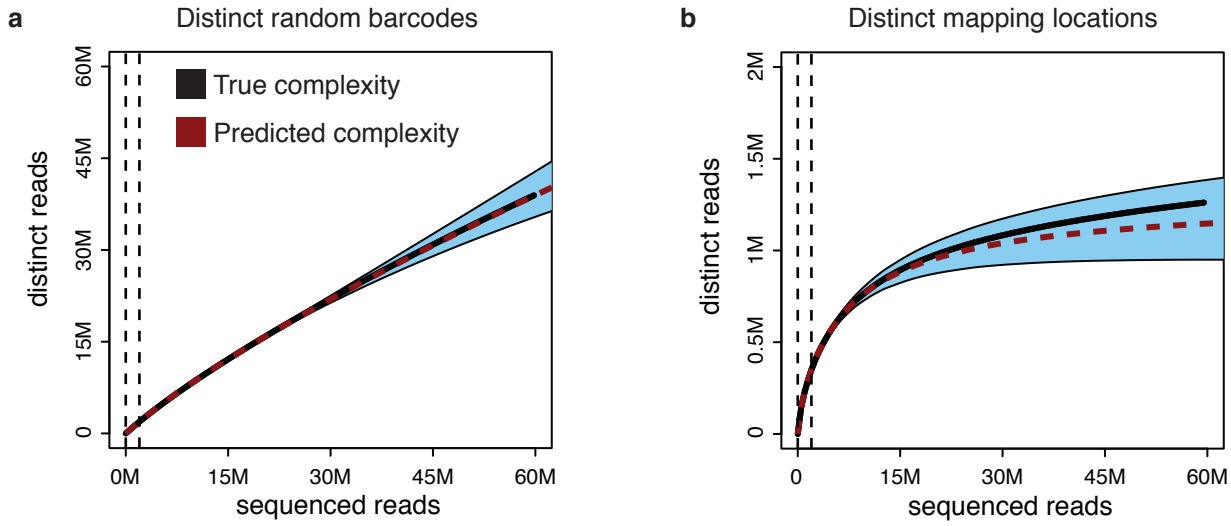


Supplementary Fig. S 3: Detailed comparison of methods predicting complexity of variable input libraries. In each graph the observed complexity curve is plotted (solid black) along with the estimated complexity curve (dashed red) and the corresponding confidence interval. For each data set, initial experiment sizes of 1M, 5M and 20M sampled reads are presented for each of the three methods: rational function (RF), zero-truncated negative binomial (ZTNB) and Euler's transform (ET). All experiments were mapped with RMAP, except for IonTorrent which is mapped with Bowtie2. **(a)** CaptureSeq, **(b)** FootRNAseq, **(c)** AdiposeRNAseq, **(d)** ADS_mRNAseq, **(e)** iPSC_mRNAseq, **(f)** ChIPMouseBCell, **(g)** HumanChIPseq1, and **(h)** HumanChIPseq2.


e
iPSC_mRNAseq
f
ChiPMouseBCell
RF
ZTNB
ET
1M
5M
20M
20M
10M
0
100M
50M
200M
g
HumanChiPSeq1
h
HumanChiPSeq2
70M
35M



Supplementary Fig. S 4: Comparison of the average relative error for the RF, ZTNB, and ET methods on negative log scale. If the estimates diverged, the contribution was set to zero. **(a)** 1M read initial experiments taken from the uniform input libraries. **(b)** 5M read initial experiments taken from the uniform input libraries. **(c)** 1M read initial experiments taken from the variable input libraries. **(d)** 5M read initial experiments taken from the variable input libraries.



Supplementary Fig. S 5: Effect of different UMI methods. Predicted and actual complexity curves for the fly RNA-seq data from Kivioja et al. (2012). **(a)** Using random barcodes and unique mapping locations as UMIs. **(b)** Using only unique mapping locations as UMIs. Initial experiment size (between vertical dashed lines) was 2M reads.