BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 22 2014, pages 3159–3165
doi:10.1093/bioinformatics/btu540

Genome analysis

Advance Access publication August 8, 2014

Modeling genome coverage in single-cell sequencing

Timothy Daley[1] and Andrew D. Smith[2,*]

[1]Department of Mathematics and [2]Department of Molecular and Computational Biology, University of Southern California, Los Angeles, CA 90089, USA

Associate Editor: John Hancock

ABSTRACT

Motivation: Single-cell DNA sequencing is necessary for examining genetic variation at the cellular level, which remains hidden in bulk sequencing experiments. But because they begin with such small amounts of starting material, the amount of information that is obtained from single-cell sequencing experiment is highly sensitive to the choice of protocol employed and variability in library preparation. In particular, the fraction of the genome represented in single-cell sequencing libraries exhibits extreme variability due to quantitative biases in amplification and loss of genetic material.

Results: We propose a method to predict the genome coverage of a deep sequencing experiment using information from an initial shallow sequencing experiment mapped to a reference genome. The observed coverage statistics are used in a non-parametric empirical Bayes Poisson model to estimate the gain in coverage from deeper sequencing. This approach allows researchers to know statistical features of deep sequencing experiments without actually sequencing deeply, providing a basis for optimizing and comparing single-cell sequencing protocols or screening libraries.

Availability and implementation: The method is available as part of the preseq software package. Source code is available at http://smithlabresearch.org/preseq.

Contact: andrewds@usc.edu

Supplementary information: Supplementary material is available at *Bioinformatics* online.

Received on May 22, 2014; revised on July 10, 2014; accepted on August 4, 2014

1 INTRODUCTION

The capability to sequence the DNA of a single cell is essential to analyzing biological diversity in heterogeneous populations of cells. Single-cell DNA sequencing technology is also necessary in applications like preimplantation genetic diagnosis based on the genotype of an individual cell biopsied from a blastocyst (Sermon *et al*., 2004). Recent efforts have used single-cell sequencing to examine genotypic heterogeneity in tumors (Narayan *et al*., 2012; Navin *et al*., 2011; Xu *et al*., 2012), rates of somatic mutations (Evrony *et al*., 2012), recombination rates in the germ line (Kirkness *et al*., 2013; Lu *et al*., 2012; Wang *et al*., 2012) and probing of the genetic diversity in unculturable bacterial populations such as those naturally occurring in the ocean (Kashtan *et al*., 2014) or the human gut (Pamp *et al*., 2012).

The challenges associated with single-cell genome sequencing are all due to the fact that the relevant DNA only exists in a single copy. For example, the nuclear DNA of a human cell weighs ~80 picograms while most standard library preparations specify a minimum input in the nanogram range (Blainey, 2013). Special protocols are needed to prepare DNA sequencing libraries in single-cell applications. Whole-genome amplification (WGA) is conducted prior to PCR, with the goal of producing more copies of the genome in the form of long amplicons that uniformly cover the original genome.

Biases in WGA can dramatically alter the representations of different parts of the genome in the sequencing library (Sun *et al*., 1995; Hosono *et al*., 2003). Methods have been developed to minimize WGA amplification bias by reducing the limiting volume for multiple displacement amplification (MDA) to avoid exponential preferential amplification (Gole *et al*., 2013; Wang *et al*., 2012) or looping of the amplicons to induce quasi-linear amplification [MALBAC; Zong *et al*. (2012)]. Despite these advances, whole genome amplification remains far from uniform.

A major problem in single-cell and low-input sequencing is the loss of loci in the process of sequencing. There are multiple opportunities for portions of the genome to disappear in the library preparation, making them unavailable for sequencing and subsequent observation. This situation is known as locus dropout and creates significant problems for downstream analysis (Shapiro *et al*., 2013). For diploid cells, locus dropout presents the additional difficulty that the dropout of one allele is easily mistaken for homozygosity. New single molecule sequencing technologies still require some form of whole-genome amplification prior to PCR amplification (Blainey, 2013), suggesting these problems will persist. It is our goal here to investigate for a single cell DNA sequencing library the genome coverage from deep sequencing, which we define as the expected number of bases in the reference genome covered by sequencing using high-throughput short-read technology.

The traditional mathematical model of sequencing assumes that all parts of the genome are represented in the sequencing library in uniform abundance, resulting in a simple Poisson distribution for the number of reads covering each base (Lander and Waterman, 1988). The possibility of unknown dropout implies this model is inadequate for single-cell sequencing. Additionally the uniformity assumption is lost due to a myriad of biases inherent to high-throughput sequencing (Sims *et al*., 2014). These problems still exist for single-cell sequencing experiments but are exacerbated by the low starting material, in addition to biases specific to WGA. One example is the observation that priming efficiency and extension rate of the DNA polymerase $\phi 29$ used in MDA is dependent on nucleotide content, leading to uneven amplification (Pinard *et al*., 2006).

*To whom correspondence should be addressed.

The highly non-uniform molecular abundances and unknown dropout in the sequencing library are not the only problems in specifying a model for the genome coverage in single-cell sequencing experiments. The coverage of local bases will be highly correlated. There is the natural correlation caused by nearby bases being covered by the same read. Additionally we expect broad correlations due correlated molecular abundances of nearby regions. One example is local correlations related to nucleotide content due to the uneven amplification of MDA. These all create problems in mathematically modeling the sequencing process, as misspecification can create significantly biased estimates.

Our aim in this paper is to present a method for estimating the genome coverage of a reference genome in a deep sequencing experiment, based only on information from a shallow initial sequencing run. One key to our method is treating sequenced nucleotides as independent observations despite the fact that the true unit of sampling is the sequenced read. We show that the loss of information caused by this assumption is acceptable. We adapt the non-parametric empirical Bayes approach we developed previously for estimating library complexity (Daley and Smith, 2013), which abstracts the sequencing process as a capture–recapture experiment. By applying our method to publicly available human single-cell sequencing data from a variety of sources and technologies, we demonstrate the method to be accurate and widely applicable. We then investigate practical considerations in applying the method, including methods to reduce the running time and ways to reduce the cost of initial experiments. Finally we apply our method to a broad swath of recent shallow single-cell sequencing experiments to show the variability of genome coverage for differing protocols.

2 THEORY

We assume a sequencing experiment samples molecules from a large pool of DNA fragments in the library. Further, we assume that the number of amplified copies of each DNA fragment is sufficiently large that sampling behaves as with replacement. Each molecule sampled during sequencing corresponds to a sequenced read. Each sequenced read will cover multiple bases in the genome: every sequenced nucleotide covers exactly one base. Our goal is to use information from a shallow sequencing run to predict the number of bases that would be covered after deep sequencing. We call the shallow sequencing run the initial experiment, and we make the essential assumption that the properties of the library do not change between the initial experiment and deeper sequencing.

We define the following symbols:

G = haploid genome length in bp;

L = read length in bp;

N = number of reads sequenced in the initial experiment;

tN = number of reads sequenced in the full experiment;

t = fold extrapolation;

π_i = probability a randomly sequence read covers base i;

$\vec{\pi} = (\pi_1, \ldots, \pi_G)$;

$\vec{\lambda} = N\vec{\pi}$, the expected number of reads covering each position.

Consider the random trial of sequencing an individual read. This can also be considered as L trials, one per nucleotide in the read. The outcomes of these L trials correspond to covering L consecutive bases in the genome. A second read whose origin partially overlaps the first will provide an additional L trials, some of which will cover new bases (Fig. 1). The part of the genome where the two reads overlap, however, will correspond to outcomes observed twice. Although the outcome of each of these LN trials is dependent on $L - 1$ others, this dependence is highly localized. Similarly, for a given base in the genome, the number of trials (sequenced nucleotides) whose outcome corresponds to that base also has a strong local dependence: the number of reads covering any given base is dependent on the number covering $2L - 2$ others.

In adapting the non-parametric empirical Bayes approach of Good and Toulmin (1956) to predict genome coverage, we will follow the compound Poisson formulation elaborated by Efron and Thisted (1976), who applied it to estimating Shakespeare's vocabulary based on his printed works. The way words may be combined in English text follows both grammatical and stylistic constraints. For example, in *Julius Ceasar*, the word '*thou*' precedes '*art*' more than twice as often as it precedes '*the*', despite the word '*the*' appearing 50 times more often than '*art*' in the entire play. Efron and Thisted (1976) argued that this amount of dependence can be ignored in a large enough body of work. Unfortunately, the dependence introduced by sampling nucleotides as contiguous reads is much stronger: by covering the ith base in the genome, one greatly increases the probability that base $i + 1$ will be covered.

For our application, however, the relations $N \gg L$ and $G \gg L$ both hold in practice, meaning that dependence between events (nucleotides in reads) and outcomes (covered bases in the reference genome) are both of extremely limited reach. When considering the outcomes, since we assume that reads are multinomially sampled, as we previously considered (Daley and Smith, 2013), the number of reads covering a position depend only on the $L - 1$ positions downstream and $L - 1$

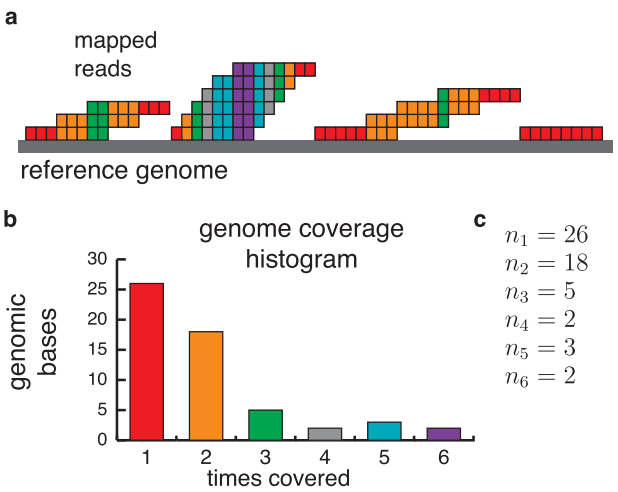

Fig. 1. (a) Schematic with sites colored to indicate how genome coverages are tabulated, (b) the corresponding colored histogram and (c) estimated coefficients for the Good–Toulmin power series

positions upstream. Therefore, the set of dependence is $2L-1$ bases. We can apply a Chen–Stein argument to infer that the limited dependence can be well approximated by assuming independence (Barbour *et al.*, 1992).

Define $n_j(t)$ as the number of bases covered by exactly j reads after sequencing tN reads and let n_j denote the number of bases covered by j reads in the initial experiment (i.e. $n_j = n_j(1)$). We call the full vector $n_1, n_2, \ldots$ the coverage counts histogram. We assume that the number of reads that cover each position follows a Poisson process with rates independently and identically distributed according to μ, an arbitrary distribution with mean NL/G. This is known as the non-parametric compound Poisson model (Wang and Lindsay, 2005) and under this model the expected value of $n_j(t)$ is equal to

$$\mathrm{E}(n_j(t)) = G\int_0^\infty (e^{-\lambda t}(\lambda t)^j/j!)\,d\mu(\lambda). \quad (1)$$

Although we know the number of positions in the genome that are not covered by reads, the value n_0 is not known because it refers only to those positions not covered by the current sequencing but might be covered if the same library were sequenced more deeply. This implies that n_0 is non-identifiable in the non-parametric model Link (2003), so analysis must be done with the identifiable portions of the model, the counts $n_1, n_2, \ldots$. Good and Toulmin (1956) introduced an empirical Bayes approach that, in our context, provides an estimator for the gain in genome coverage from sequencing an additional $(t-1)N$ reads. We refer to the following as the Good–Toulmin estimator:

$$\hat{\Delta}_G(t) = \sum_{j=1}^{\infty} (-1)^{j+1}(t-1)^j n_j. \quad (2)$$

The way these observed counts are obtained is illustrated schematically in Figure 1.

As we previously discussed (Daley and Smith, 2013), the Good–Toulmin estimator is highly accurate when predicting the gain in coverage for small increases in the experiment. Using the Good–Toulmin estimator to predict beyond $t = 2$ is problematic and suffers from extreme instability. In particular, the estimator will diverge to positive or negative infinity depending on whether the largest observed coverage count is odd or even. We introduced rational function approximations to obtain globally stable estimates that still satisfy the nice local properties of the Good–Toulmin estimator (Daley and Smith, 2013). A rational function approximation to a power series is a ratio of polynomials that asymptotically approximates the power series up to a given degree,

$$\sum_{j=1}^{P+Q+1} (-1)^{j+1}(t-1)^j n_j =$$

$$(t-1)\frac{p_0 + p_1(t-1) + \ldots + p_P(t-1)^P}{1 + q_1(t-1) + \ldots + q_Q(t-1)^Q} + O((t-1)^{P+Q+2}).$$

To guide the selection of P and Q, we note that the coverage will asymptote as the number of bases sequenced and the sequencing depth goes to infinity. This indicates that we should choose $P = Q - 1$ so that the rational function behaves the same in the limit as the coverage curve which we approximating. Our observation is that this choice gives superior performance to other rational function approximations (Supplementary Fig. S1), allowing for accurate and stable long-range predictions of the genome coverage.

3 ACCURACY OF PREDICTIONS

We evaluated our method on 19 single-end 100 or 101 nucleotide (nt) single-cell sequencing experiments (Lu *et al.*, 2012; Wang *et al.*, 2012; Zong *et al.*, 2012) that were 'deeply-sequenced', which we defined as at least 100 M reads. We downsampled each library to 5 million (M) reads to simulate the initial sequencing experiment and compared the estimated genome coverage curve to the observed genome coverage curve, calculated by downsampling the library and using BEDTools (Quinlan and Hall, 2010). We measure the accuracy of our estimates by the relative error, which we define as the observed genome coverage minus the estimated divided by the observed.

Even for long-range extrapolations, the estimates remain stable and accurate (Supplementary Figs S2 and S3). When we use 5 M sequenced reads ($0.17\times$ coverage) to predict the genome coverage for 100 M sequenced reads ($3.33\times$ coverage), a 20-fold extrapolation, we observed a mean absolute relative error of <3% (Table S1, Supplementary Figs S2 and S3).

We notice that our choice of order of the rational function approximation tends to give conservative estimates of the genome coverage (Supplementary Figs S2 and S3). Theoretically, the distribution of the curves should be symmetric about the mean, but we observe a downward skew to curves (D'Agostino skewness test (D'Agostino, 1970) on the relative error for the 20-fold extrapolations $P < 3\,\mathrm{E}\text{-}16$). We investigated bootstrapping to reduce the skew and lower the variance of our estimates (Breiman, 1996). The bootstrapped median shows significant improvement over the simple extrapolations and even the bootstrapped mean (Supplementary Fig. S4), indicating that the use of bootstrapping and aggregating the median curve leads to more accurate estimates.

4 PRACTICAL CONSIDERATIONS

4.1 Binning to approximate coverage

The major computational demands in the extrapolation algorithm come from bootstrapping the coverage counts histogram to reduce variance of the estimator (see Supplementary Information). The time required for resampling is a function of the number of possible outcomes for each event, in this case individual genome positions. An approach to reduce the running time is therefore to partition the genome into non-overlapping bins and use the bins as the outcomes of each random trial. If we had an estimate for the number of bins that would be covered after deep sequencing from the library, the genome coverage could be estimated by multiplying that number of bins by their size (or their average size, if the bins had varying sizes). We refer to this approach as a binning based estimator. It is not difficult to see that counts for bins can be supplied to the Good–Toulmin

estimator in the same way as counts for individual nucleotides as illustrated in Figure 1.

We assume the genome has been partitioned into equal sized bins. There are several possible schemes one may follow to construct a binning based estimator. If we say that a read covers a bin only if it completely covers the bin then we will consistently underestimate the coverage, and this excludes use of bins larger than the read length. Conversely, if we say a read covers a bin if there is any overlap then we will consistently overestimate genome coverage. A reasonable approach is to say a read covers a bin with probability proportional to the number of nucleotides in the read that cover the bin (Supplementary Fig. S5). This ensures that, for a single read, the probability any position is covered by that read is unbiased. Unfortunately one can demonstrate that even for binning based estimators that are unbiased for a single read, they will be biased for multiple reads (Supplementary Information). This also works in the other direction, if a strategy is biased for a fixed number of multiple reads then it will be biased for any other number of reads.

The binning-based estimator does indeed slightly underestimate coverage (Supplementary Information, Supplementary Fig. S6). Despite this, we observe that the binned coverage counts histogram is approximately proportional to single base resolution coverage counts histogram (Supplementary Fig. S7). This holds even when the bin size is larger than the read length, so that reads are randomly thrown out or extended to cover whole bin while on average covering the same number of bases (Fig. 2a, Supplementary Fig. S7). The gain from binning is a reduction in running time proportional to the bin size (Fig. 2b, Supplementary Information). As can be seen from Supplementary Figure S6, this comes at an acceptable cost in accuracy.

The binning procedure naturally introduces variance in the estimated coverage count histogram and therefore also the bin size due to the random nature of the binning. This variance increases with the bin size, as portions of reads are more likely to be thrown out or to be extended with increasing bin size. On the other hand, the binning will group close neighboring bases together and reduce the variance introduced when treating close bases as independent. Accordingly the variance of the estimated coverage should achieve a minimum lying somewhere between single base pair and the read length (Fig. 2c, Supplementary Fig. S6 and S9).

4.2 Effects of read length on coverage estimates

At present the vast majority of whole-genome sequencing experiments, including single-cell experiments, use Illumina sequencing technology. Reads from this technology correspond to the ends of DNA fragments. Suppose one sequences L nucleotide reads that map unambiguously back to the reference genome, and that the DNA fragments in the library all have length greater than L. Using information about the genome coverage from that experiment, one could compute what the genome coverage would have been had the reads been of length $L + 1$. This can be done by 'extending' the reads in silico, which amounts to pretending the reads were longer.

The cost of a sequencing experiment is approximately a linear function of (i) the number of reads sequenced, and (ii) the length of reads. For a variety of reasons, including optmization of total

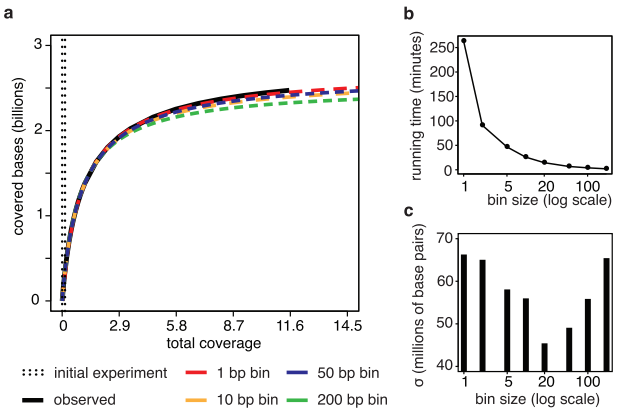

Fig. 2. (a) Observed coverage curve versus estimated using bin sizes of 1, 10, 50 and 200 using the same 5 M 100 base pair single end read initial experiment. (b) Running times to extrapolate the curve for different bin sizes. (c) Observed variance of the full extrapolation estimates (to 398 M mapped reads) for different bin sizes from 250 independent 5 M read downsampled initial experiments

throughput at sequencing centers, investigators are often presented with a relatively limited set of options. Often there are options for a few standard read lengths that include both 'short' and 'long' options (at present, e.g. 36 and 100 nt). We asked whether an approach of extending reads in silico could help minimize the cost of the initial experiment while still accurately predicting genome coverage in deeper sequencing with true longer reads.

In line with the assumption above, we expect that in silico read extension will work as long as the mapped reads find their correct mapping positions at shorter read lengths. While this places a lower bound on the lengths of reads that can be used, there are only small differences in mappability for reads >36 nt (Derrien *et al*., 2012).

To evaluate this approach, we examined four libraries given by Sequencing Read Archive accession numbers SRX151616, SRX202787, SRX204160 and SRX205367, which correspond respectively to an MDA haploid, MALBAC diploid, MDA diploid and MALBAC haploid libraries. We downsampled 5 M reads originally of length 100 bp, and then truncated the original reads at 50 nt prior to mapping. We tested the results for extending the reads by 50 bp *in silico* versus extending the fold-extrapolation (Supplementary Fig. S10). We see that the artificially extended reads consistently overestimate the genome coverage (mean relative error across all libraries = 0.018). On the other hands, extending the fold-extrapolation on the mapped truncated reads shows an increase in relative error compared to the full length reads (mean relative error of –0.028 across all libraries), not unexpected, as we are estimating further away from the initial experiment. This indicates that artificially extending the reads *in silico* is the incorrect strategy and we should simply extend the fold extrapolation.

We next examined the effect of read length on the estimates while holding constant the total number of nucleotides sequenced. For each of the four libraries, we downsampled 5 M reads at 100 nt, and 10 M reads at 50 nt. The 50 nt read samples consistently yield more accurate predictions of genome

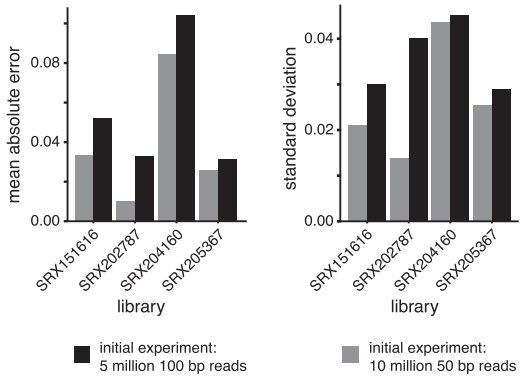

Fig. 3. Mean absolute error and standard deviation when extrapolating to the full library, which is 25.4, 79.6, 93.8 and 23.9 fold for SRX151616, SRX202787, SRX204160 and SRX205367, respectively

coverage and have lower variance (Fig. 3, Supplementary Fig. S11). This result suggests that any bias associated with mappability and fragment length variance, at least for 50 nt reads and the human reference genome, may be acceptable to reduce costs for library test runs.

4.3 Does including duplicate reads impact predictions?

It is standard in many sequencing applications to remove all but one read among a set believed to be amplified copies of the same original DNA fragment (Sims *et al*., 2014). Multiple options are available for identifying when reads are duplicates (Kivioja *et al*., 2012). In the context of single-cell genotyping, one often seeks to use duplicates to correct sequencing errors, a concept that has been called 'single-molecule consensus' reads (Hiatt *et al*., 2013).

For most current methods of identifying amplification duplicates, all duplicate reads for a given molecule will map to the same position in the genome. As a consequence, the genome coverage for a given sequencing experiment will be identical whether or not duplicates are removed. Moreover, in theory the number of distinct reads that cover a particular position should also approximately follow a Poisson process. This observation suggests two possible avenues for using the Good–Toulmin genome coverage estimator to predict coverage from deeper sequencing. The direct approach is to ignore duplicate reads when making predictions based on the test run. A different approach would be to make estimates using only unique reads, and then to make predictions based on (i) the number of unique reads in a deeper experiment, using the method of Daley and Smith (2013), and (ii) to then predict genome coverage conditional on the number of unique reads sequenced in the full experiment (Supplementary Fig. S12). This second approach includes two predictive steps, but neither of these steps would be extrapolating as far as the direct approach. And we know that the non-parametric empirical Bayes framework is more accurate when extrapolations are less extreme.

Although genome coverage estimates with duplicates and without duplicates should be equal, the variances may differ. Therefore we compare the results from the same library under both cases extrapolated to the same genome coverage achieved at 100 M reads with duplicates included, so that the fold-extrapolation is less for duplicates removed (Supplementary Fig. S13). For some cases (MALBAC libraries), the performance with duplicates removed comparable to including duplicates. But for the MDA haploid libraries, the performance is worse (paired *t*-test, $P < 3\text{E}-16$). We therefore suggest that the genome coverage is predicted with duplicated included, but note that it can also be done with duplicates removed.

5 COMPARING LIBRARY PREP PROTOCOLS

There is little available data on the performance of specific single-cell library preparation protocols based on deeply sequencing the libraries. Existing information exists only for a few protocols (Geigl *et al*., 2009; Zong *et al*., 2012) or for bacterial genomes (Pinard *et al*., 2006; Zhang *et al*., 2006). The reason is likely the cost associated with deeply sequencing libraries purely to determine the degree to which they differ. This has almost certainly also been a barrier to advances in single-cell library protocols: evaluating each protocol variation is a major cost.

Because of its accuracy, our estimator completely eliminates the cost barrier to developing and refining single cell library protocols. Libraries can be constructed and their deep properties inferred with high accuracy without actually deeply sequencing. To illustrate this concept, we examined a multitude of low-coverage single-cell sequencing experiments from a variety of recent human studies (Evrony *et al*., 2012; Gole *et al*., 2013; Hou *et al*., 2013; Kirkness *et al*., 2013; McConnell *et al*., 2013; Navin *et al*., 2011; Ni *et al*., 2013). We limited our analysis to 209 libraries with at least 500 M total nucleotides mapped to autosomes (equal to $0.17\times$ coverage). Because our error in estimating genome coverage increases with fold-extrapolation, we downsampled all libraries to exactly 500 M nucleotides mapped to the autosomes, avoiding any effects of variable initial sample size. We estimated genome coverage using a bin size of 5. We then compare the estimates at 10 gigabases ($\approx 3.5\times$ coverage) and 30 gigabases ($\approx 10.4\times$ coverage) mapped to the autosomes. A summary of all libraries can be found in Table S2.

Karyotype and ploidy can have a major impact on the apparent genome coverage in a single-cell experiment. Conceptually, higher copy-number for chromosomes, or large parts thereof, means greater opportunity to observe reads mapping back to the corresponding part of a reference genome. It can be difficult to resolve such issues for an individual cell without prior knowledge of genotype. In terms of our evaluations, these difference mean that cells of differing karyotype or ploidy may not be directly comparable. For example Gole *et al*. (2013) had to restrict their comparison to the MALBAC method in diploid data to only a few chromosomes of the original data (Zong *et al*., 2012) and to pooled haploid libraries (Lu *et al*., 2012) due to unusual karyotype of the cell line used to benchmark the method. To test the effect of ploidy we take matching first and second polar bodies from Hou *et al*. (2013), resulting in 42 paired samples. Every pair was excised, prepared, and sequenced at the same time, which we hope will result in minimal batch effects. We find that the diploid libraries have significantly higher coverage than the haploid libraries for all levels of coverage considered (Supplementary Fig. S14), confirming that ploidy has a significant effect on dropout.

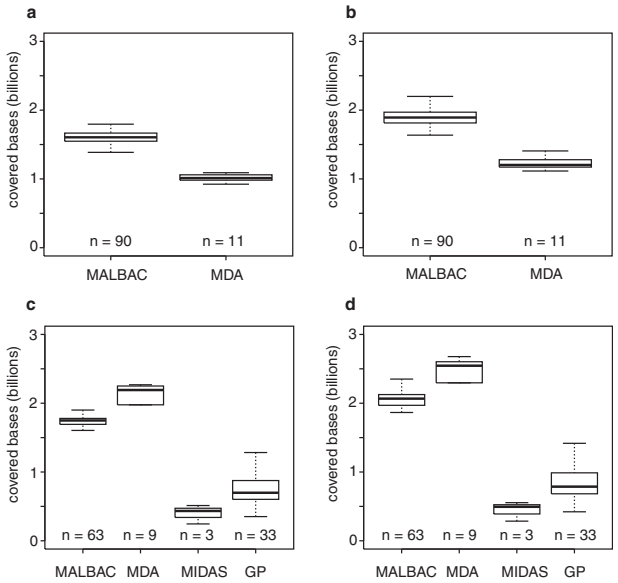

Fig. 4. Expected genome coverage for evaluated haploid single-cell libraries at sequencing depth of (a) 3.5× and (b) 10.4×, and for diploid single-cell libraries at sequencing depth of (c) 3.5× and (d) 10.4×

We next examine the differences in library preparation protocols, keeping libraries from haploid and diploid cells separate. For the haploid case, two studies have produced data from human sperm cells amplified using MDA (Kirkness *et al*., 2013) and second polar bodies and female pronuclei amplified with MALBAC (Hou *et al*., 2013). We observed a significant difference in estimated coverage for both 3.5× extrapolation and 10.4× extrapolation (Fig. 4a and b, two-sided *t*-test $P < 5E-11$ for both the 3.5× and 10.4× coverage). With only one study for each group we can not identify if the difference is due to the differing protocols (MDA or MALBAC) or other factors.

For the diploid case data from four protocols is available: (i) first polar bodies (Hou *et al*., 2013) and circulating tumor cells (Ni *et al*., 2013) amplified with MALBAC; (ii) neurons (Evrony *et al*., 2012; Gole *et al*., 2013) and a single lymphocyte (Gole *et al*., 2013) amplified with MDA; (iii) neurons (Gole *et al*., 2013) amplified with MIDAS; and (iv) breast tumor cells (Navin *et al*., 2011) and neurons (McConnell *et al*., 2013) amplified by Sigma-Aldritch GenomePlex universal oligonucleotide primers (GP).

MALBAC and MDA libraries exhibit significantly higher expected genome coverage than those based on MIDAS or GP, at both depths considered (Fig 4c and d). MALBAC and MDA also exhibit substantial variability in expected genome coverage; these are the only protocols for which data was used from multiple studies. For MALBAC, the three libraries showing lowest expected genome coverage are from circulating tumor cells (the three lowest MALBAC points in Fig. 4c and d). These have the possibility of being polyploid or aneuploid, so by the logic above we would expect these libraries to have predicted genome coverage at least that of the diploid first polar bodies. This indicates that cell type and lab (which would account for slight differences in library preparation protocol) may account for a large portion of the variability and there are opportunities for optimization of protocols to improve whole genome amplification.

6 DISCUSSION

We described a method for predicting the genome coverage gained from deeper sequencing of a single-cell genome sequencing library based on a compound Poisson model of sequencing. By ignoring local dependence, we can approximate the number of bases covered by additional sequencing with a non-parametric empirical Bayes estimator. This estimator is extremely accurate for predicting additional coverage from relatively small amounts of additional sequencing but suffers from large instabilities for large amounts of additional sequencing. Applying rational function approximations removes the instability and allows us to make accurate long-range predictions.

The running time of the algorithm may be unreasonably long for single base resolution estimates. To facilitate researchers in obtaining quick and accurate estimates, we introduced a strategy to reduce the running time of the algorithm significantly, with a small cost in accuracy, by randomly binning reads. By choosing the bin size, the researcher has the option to control how quickly estimates can be obtained, keeping in mind the trade-off of accuracy and variance.

There is appreciable variability in genome coverage both for the deeply sequenced libraries and for the extrapolated low-coverage libraries; this variability exists even for libraries originating in the same lab using the same protocol. In such cases our method can help for selection of the best libraries to sequence deep. This can help researchers in knowing the trade-off between sequencing depth and observed loci prior to committing resources for deep sequencing. This is particularly important for studies involving single nucleotide variation that require deep sequencing rather than broader variation such as copy number variation. One field where this is becoming increasingly important is full genome pre-implantation genetic diagnosis, which until recently was considered impossible due to technological constraints (Geraedts and De Wert, 2009).

Finally, a major barrier to the development of new technologies or optimization of current protocols is the resources required to compare genome coverage across libraries. Naive use of shallow test sequencing runs to compare libraries is often misleading, as samples that initially appear to be high complexity may suffer from large locus dropout (Supplementary Fig. S13). The method we have presented can provide the information required for deep evaluation of libraries without deep sequencing. Though we presented our analysis to the problem of sequencing single human cells, the method is equally applicable to sequencing projects with a reference genome from low (Parkinson *et al*., 2012) or highly degraded (Prüfer *et al*., 2014) input, bacterial samples with a reference genome, or when mapping to a reference genome of a closely related species with unknown overlap (Enk *et al*., 2014).

ACKNOWLEDGEMENT

We would like to thank Manuel Lladser, Michael Waterman and Norm Arnheim for their helpful discussions and insight. We would also like to thank the Smith lab group for their comments, help and patience.

Funding: This work was supported by US National Institute of Health National Heath Genome Research Institute grants (no. R01 HG005238-01 and P50 HG002790-06).

Conflict of interest: The authors are co-inventors of a patent application related to the topic of this manuscript. Non-commercial use of the method and associated software is free of charge and Open Source software has been provided by the authors.

REFERENCES

Barbour,A.D. *et al.* (1992) Compound poisson approximation for nonnegative random variables via Stein's method. *Ann. Probab.*, **20**, 1843–1866.

Blainey,P.C. (2013) The future is now: single-cell genomics of bacteria and archaea. *FEMS Microbiol. Rev.*, **37**, 407–427.

Breiman,L. (1996) Bagging predictors. *Mach. Learn.*, **24**, 123–140.

D'Agostino,R.B. (1970) Transformation to normality of the null distribution of G1. *Biometrika*, **57**, 679–681.

Daley,T. and Smith,A.D. (2013) Predicting the molecular complexity of sequencing libraries. *Nat. Methods*, **10**, 325–327.

Derrien,T. *et al.* (2012) Fast computation and applications of genome mappability. *PLoS One*, **7**, e30377.

Efron,B. and Thisted,R (1976) Estimating the number of unseen species: how many words did Shakespeare know? *Biometrika*, **63**, 435–447.

Enk,J.M. *et al.* (2014) Ancient whole genome enrichment using baits built from modern DNA. *Mol. Biol. Evol.*, **31**, 1292–1294.

Evrony,G.D. *et al.* (2012) Single-neuron sequencing analysis of L1 retrotransposition and somatic mutation in the human brain. *Cell*, **151**, 483–496.

Geigl,J.B. *et al.* (2009) Identification of small gains and losses in single cells after whole genome amplification on tiling oligo arrays. *Nucleic Acids Res.*, **37**, e105.

Geraedts,J. and De Wert,G. (2009) Preimplantation genetic diagnosis. *Clin. Genet.*, **76**, 315–325.

Gole,J. *et al.* (2013) Massively parallel polymerase cloning and genome sequencing of single cells using nanoliter microwells. *Nat. Biotechnol.*, **31**, 1126–1132.

Good,I. and Toulmin,G. (1956) The number of new species, and the increase in population coverage, when a sample is increased. *Biometrika*, **43**, 45–63.

Hiatt,J.B. *et al.* (2013) Single molecule molecular inversion probes for targeted, high-accuracy detection of low-frequency variation. *Genome Res.*, **23**, 843–854.

Hosono,S. *et al.* (2003) Unbiased whole-genome amplification directly from clinical samples. *Genome Res.*, **13**, 954–964.

Hou,Y. *et al.* (2013) Genome analyses of single human oocytes. *Cell*, **155**, 1492–1506.

Kashtan,N. *et al.* (2014) Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. *Science*, **344**, 416–420.

Kirkness,E.F. *et al.* (2013) Sequencing of isolated sperm cells for direct haplotyping of a human genome. *Genome Res.*, **23**, 826–832.

Kivioja,T. *et al.* (2012) Counting absolute numbers of molecules using unique molecular identifiers. *Nature methods*, **9**, 72–74.

Lander,E.S. and Waterman,M.S. (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, **2**, 231–239.

Link,W.A. (2003) Nonidentifiability of population size from capture-recapture data with heterogeneous detection probabilities. *Biometrics*, **59**, 1123–1130.

Lu,S. *et al.* (2012) Probing meiotic recombination and aneuploidy of single sperm cells by whole-genome sequencing. *Science*, **338**, 1627–1630.

McConnell,M.J. *et al.* (2013) Mosaic copy number variation in human neurons. *Science*, **342**, 632–637.

Narayan,A. *et al.* (2012) Ultrasensitive measurement of hotspot mutations in tumor DNA in blood using error-suppressed multiplexed deep sequencing. *Cancer Res.*, **72**, 3492–3498.

Navin,N. *et al.* (2011) Tumour evolution inferred by single-cell sequencing. *Nature*, **472**, 90–94.

Ni,X. *et al.* (2013) Reproducible copy number variation patterns among single circulating tumor cells of lung cancer patients. *Proc. Natl Acad. Sci. USA*, **110**, 21083–21088.

Pamp,S.J. *et al.* (2012) Single-cell sequencing provides clues about the host interactions of segmented filamentous bacteria (SFB). *Genome Res.*, **22**, 1107–1119.

Parkinson,N.J. *et al.* (2012) Preparation of high-quality next-generation sequencing libraries from picogram quantities of target DNA. *Genome Res.*, **22**, 125–133.

Pinard,R. *et al.* (2006) Assessment of whole genome amplification-induced bias through high-throughput, massively parallel whole genome sequencing. *BMC Genomics*, **7**, 216.

Prüfer,K. *et al.* (2014) The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature*, **505**, 43–49.

Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Sermon,K. *et al.* (2004) Preimplantation genetic diagnosis. *Lancet*, **363**, 1633–1641.

Shapiro,E. *et al.* (2013) Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nat. Rev. Genet.*, **14**, 618–630.

Sims,D. *et al.* (2014) Sequencing depth and coverage: key considerations in genomic analyses. *Nat. Rev. Genet.*, **15**, 121–132.

Sun,F., Arnheim,N. and Waterman,M.S. (1995) Whole genome amplification of single cells: mathematical analysis of PEP and tagged PCR. *Nucleic acids research*, **23**, 3034–3040.

Wang,J. *et al.* (2012) Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. *Cell*, **150**, 402–412.

Wang,J.P.Z. and Lindsay,B.G. (2005) A penalized nonparametric maximum likelihood approach to species richness estimation. *J. Am. Stat. Assoc.*, **100**, 942–959.

Xu,X. *et al.* (2012) Single-cell exome sequencing reveals single-nucleotide mutation characteristics of a kidney tumor. *Cell*, **148**, 886–895.

Zhang,K. *et al.* (2006) Sequencing genomes from single cells by polymerase cloning. *Nat. Biotechnol.*, **24**, 680–686.

Zong,C. *et al.* (2012) Genome-wide detection of single-nucleotide and copy-number variations of a single human cell. *Science*, **338**, 1622–1626.