The Annals of Applied Statistics
2019, Vol. 13, No. 1, 113–131
https://doi.org/10.1214/18-AOAS1197
© Institute of Mathematical Statistics, 2019

CLONALITY: POINT ESTIMATION[1]

BY LU TIAN*, YI LIU†, ANDREW Z. FIRE*, SCOTT D. BOYD* AND RICHARD A. OLSHEN*

Stanford University and Calico Life Sciences LLC†*

Assessments of biological complexity for populations that are of mixed species are central in many biological contexts, including microbiomes, tumor cell population structure, and immune cell populations. Here we address the problem of quantifying the population diversity in experiments where high throughput DNA sequencing is used to distinguish a large number of cell subpopulations. Our model assumes a list of clonal species and their observed frequencies in each of several replicate sequencing libraries. Though the underlying distribution of frequencies cannot be estimated well from data coming from only a small fraction of the total cell population, one can estimate well the population-level *clonality*, defined as the sum of squared underlying fractions of the respective clones, the complement of the Gini–Simpson index. Specifically, we proposed to adaptively combine multiple unbiased estimators of *clonality* derived from pairs of replicates to construct a single estimator without relying on the commonly used but restrictive multinomial assumption. The new estimator performs particularly well for replicates of unequal size. We further illustrate the proposed methods with extensive simulations and a small real data example.

1. Introduction.

A quantitative understanding of the structures of biological populations is central to many questions in science and medicine. Recent advances in analytical chemistry have facilitated detailed descriptions of biological macromolecules, particularly in the area of DNA sequencing. Sampling and classification of molecules from a complex population leads to a natural series of questions about the number of distinct classes of molecules, and their proportions in the population. At one extreme, the underlying population can consist of a single class; at the other extreme, each individual in a population may be the unique representative of its class. A practical challenge in the characterization of populations of biological molecules is that often only a small fraction of the total can be sampled and imperfectly measured. In our case, the molecules are derived from cells in the blood of a patient. Typically, this small fraction of blood from an individual is divided during

Received March 2017; revised June 2018.

[1] Supported in part by Grant 5U19AI090019 and U54AI065359 from the National Institute of Allergy and Infectious Diseases, by Discovery Fund from Stanford University, by Grant R01 HL089778-05 from National Institutes of Heart, Lung, and Blood, by Grant 4R37EB002784 from the National Institute of Biomedical Imaging and Bioengineering, and by Grant NIH UL1TR001085 from the National Institutes of Health.

Key words and phrases. Clonality, V(D)J rearrangements, richness, jackknife.

the experiment into a few (typically 5 to 10) parts, within which the biological population of interest is measured. These parts are termed *replicates*. Specifically, one may measure the cell population composition by amplifying pools of molecules from the blood sample using the celebrated polymerase chain reaction (PCR) and subsequent DNA sequencing. The result of DNA sequencing from each replicate is total numbers of reads of different types, which are the basis for studying the population structure. One complication is that PCR amplification may randomly alter original cell counts. This randomness oftentimes causes violation of the simple multinomial assumption for the observed counts of different cell types. While problems we study bear substantial resemblance to those in ecology, the sample size and underlying model assumption can be very different.

This research is motivated by the need to quantify the diversity of the adaptive human immune system, that of so-called V(D)J rearrangements of particular subsets of T cells and B cells [Schatz and Ji (2011)]. V(D)J rearrangement occurs in the primary lymphoid organs (bone marrow for B cells and thymus for T cells) and in a nearly random fashion rearranges variable (V), joining (J), and in some cases, diversity (D) gene segments. It is a mechanism of genetic recombination that results in the highly diverse repertoire of antibodies/immunoglobulins and T cell receptors. The subgroups of immune cells with particular V(D)J rearrangements are termed clones. If we have two cells drawn independently from the population, then the “collision probability” that two cells belong to the same clone is obviously the sum of squares of the proportions of cells in different clones. Such a “collision probability” is the complement of the Gini–Simpson index [Breiman et al. (1984), Kaplinsky and Aranout (2016)] and a natural measure of the diversity of a population. We term it *clonality* of a population; estimating it is our primary goal. In the rest of the paper, we use *clonality* with respect to the V(D)J rearrangement of immune cells to illustrate our proposal. However, the method has much broader applications.

In addition to *clonality*, Shannon’s entropy is obviously another summary for the distribution of clones. The number of different clones could be of interest as well. The paper by Kaplinsky and Aranout (2016) speaks well to the conundrum that no single quantity suffices. Each of these summaries can be descriptive of a population [Laydon, Bangham and Asquith (2015)]. For estimating the number of distinct clones in ecology, see the celebrated work of Chao (1987, 1989). The question of estimating the proportions of unseen clones can be traced back to the work done in Bletchley Park during World War II [Aldrich (2010), Good (1953), Robbins (1968)]. An analogous problem has been studied in linguistic contexts by Efron and Thisted (1976). In general, it is even more difficult to estimate aforementioned quantities than *clonality* without strict parametric assumptions, and they are not our focus.

Once *clonality* is available for a group of subjects, subsequent analyses can be conducted to examine the association between different phenotypes and immune diversity characterized by *clonality*. For example, one hypothesized cause for inadequate vaccine efficacy among the elderly is their lost immune diversity as a consequence of repeatedly responding to past challenges. This hypothesis can be tested by examining if the *clonality* increases with age. One may also directly associate the vaccine efficacy measured by, for example, antibody titer changes after the vaccination, with *clonality*. Since the true value of *clonality* would never be known considering the huge number of immune cells in our body, we need to use its estimated counterpart in aforementioned analyses. Consequently, the power and efficiency of those analyses depend on the accuracy of *clonality* estimator. As is well known from measurement error literature, using inaccurate estimators for the true *clonality* can greatly dilute the underlying association of interest [Fuller (1987)].

In this paper, we attempt to construct an approximately unbiased estimator with a small variance. In Section 2, we address necessary assumptions and describe the proposed estimation method. Section 3 gives results of extensive simulations for evaluating improvements of the new *clonality* estimator over existing ones. Section 4 is a report on applications to data from a recent study of CD4 T cells. Finally, Section 5 includes further discussion.

2. Method.

2.1. *Model assumptions.* We assume that $\boldsymbol{p} = (p_1, \dots, p_C)'$ is a probability vector whose coordinates represent the relative abundance of $V(D)J$ rearrangements of T cells or B cells for an individual. The dimension C is termed *richness*. Without loss of generality, we assume that $p_1 \geq p_2 \geq \cdots \geq p_C > 0$ for $\boldsymbol{p}$. We wish to estimate *clonality*

$$\theta = \sum_{i=1}^{C} p_i^2.$$

The closer *clonality* is to 1, the more peaked the probability vector $\boldsymbol{p} \in \mathcal{R}^C$ is, and the less diverse the cell population is.

Only the finite set of vectors $\boldsymbol{R}_i$, $1 \leq i \leq n_R$, is observed in practice. The index n_R is the number of *replicates*; each $\boldsymbol{R}_i$ consists of a vector of nonnegative integers $(R_{i1}, \dots, R_{iC})'$, where R_{ij} is the count of sequenced reads from the jth clone in the ith replicate. We first assume that

$$\{\boldsymbol{R}_1, \boldsymbol{R}_2, \dots, \boldsymbol{R}_{n_R}\} \quad \text{are independent.} \tag{2.1}$$

The observed clone frequencies in the ith replicate are represented by the vector $\hat{\boldsymbol{p}}_i = (\hat{p}_{i1}, \dots, \hat{p}_{iC})'$, where $\hat{p}_{ij} = R_{ij} / \sum_{j=1}^{C} R_{ij}$. The denominator of the fraction is the total number of *reads* in replicate i. Since *clonality* $\theta = \boldsymbol{p}'\boldsymbol{p} = \|\boldsymbol{p}\|_2^2$, one obvious estimate of θ is its empirical counterpart

$$n_R^{-1} \sum_{i=1}^{n_R} \|\hat{\boldsymbol{p}}_i\|_2^2.$$

However, the function x^2 is convex. So, Jensen's inequality implies that the cited average is upwardly biased, even if $E(\hat{\boldsymbol{p}}_i|\boldsymbol{p}) = \boldsymbol{p}$.

A more sophisticated approach is to assume that $\mathbf{R}_i$ follows a multinomial distribution, a common assumption adopted in ecology [McKane, Alonso and Solé (2004)], and construct the simple unbiased estimators for *clonality* as

$$\sum_{j=1}^{C_i} \frac{R_{ij}(R_{ij}-1)}{R_{i.}(R_{i.}-1)}, \qquad 1 \le i \le n_R,$$

where C_i is the observed *richness* in replicate i. However the multinomial assumption is often too strong and may be violated in intended applications. The underlying reason is that while the number of sampled cells from different clones may follow a multinomial distribution, the sequenced reads from each cell introduce extra randomness, so the distribution of counts of reads from different clones is no longer multinomial. The objective of this paper is to construct a valid and accurate estimator under the weak but plausible assumption that

$$E(\hat{\boldsymbol{p}}_i|R_{i.}) = \boldsymbol{p}. \tag{2.2}$$

To this end, define $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \dots, \varepsilon_{iC})$ by $\boldsymbol{\varepsilon}_i = \hat{\boldsymbol{p}}_i - \boldsymbol{p}$. Thus, $E(\boldsymbol{\varepsilon}_i|R_{i.}) = \mathbf{0}$, the C dimensional vector whose every entry is 0. If $\mathbf{R}_i|R_{i.}$ follows a multinomial distribution with the probability $\boldsymbol{p}$, then clearly (2.2) is satisfied. In fact, it can hold true in cases well beyond the multinomial setting. For example $\mathbf{R}_i|(R_{i.}, \boldsymbol{p}_i)$ may follow a multinomial distribution $\mathrm{MN}(R_{i.}, \boldsymbol{p}_i)$ with a random probability vector $\boldsymbol{p}_i$ following a complicated distribution across replicates, but $E(\boldsymbol{p}_i|\boldsymbol{p}) = \boldsymbol{p}$. This assumption in general is true if there is no systematic "bias" in generating the observed counts from sampled cells such as sample contamination during sequencing.

2.2. *Clonality estimation.* The assumptions (2.1) and (2.2) imply that for $k < l$,

$$E(\hat{p}_{kj}\hat{p}_{lj}|R_{k.}, R_{l.}) = p_j^2$$

for all $1 \le j \le C$. Therefore, we may define

$$\hat{\theta}_{(k,l)} = \hat{\boldsymbol{p}}_k' \hat{\boldsymbol{p}}_l = \sum_{j=1}^{C} \hat{p}_{kj}\hat{p}_{lj},$$

as an unbiased estimator of θ. Note that although we observe only positive $\hat{p}_{lj}$, and C is unknown, $\hat{\theta}_{(k,l)}$ can always be calculated from the observed data. Since $E(\hat{\theta}_{(k,l)}|R_{k.}, R_{l.}) = \theta$, the weighted average

$$\hat{\theta}_w^* = \frac{\sum_{1 \le k < l \le n_R} w(R_{k\cdot}, R_{l\cdot})\hat{\theta}_{(k,l)}}{\sum_{1 \le k < l \le n_R} w(R_{k\cdot}, R_{l\cdot})}$$

is also an unbiased estimator for θ. That is, $E(\hat{\theta}_w^* | R_{1\cdot}, \dots, R_{n_R\cdot}) = \theta$ for any weight function w. A special case for which $w(r_1, r_2) = r_1 r_2$ was employed by Parameswaran et al. (2013).

To construct a better estimator, we need to consider the following fact. Suppose that we have M asymptotically unbiased estimators of θ, $\tilde{\theta}_1, \dots, \tilde{\theta}_M$. We may combine them to obtain a more efficient estimator by computing

$$\tilde{\theta} = \sum_{i=1}^{M} w_i \tilde{\theta}_i,$$

where $(w_1, \dots, w_M)' = C_w \boldsymbol{\Sigma}_M^{-1} \mathbf{1}_M$, C_w is a constant such that $\sum_{i=1}^{M} w_i = 1$, $\mathbf{1}_M$ is a M dimensional vector whose every entry is 1, and $\boldsymbol{\Sigma}_M$ is the variance-covariance matrix of $(\tilde{\theta}_1, \dots, \tilde{\theta}_M)'$. Therefore, to choose a good weighting scheme for $\hat{\theta}_w^*$, we need to study the covariance between $\hat{\theta}_{(k,l)}$ and $\hat{\theta}_{(g,h)}$ for all possible pairs of (k,l) and (g,h). If all four subscripts are distinct, then obviously $\operatorname{Cov}(\hat{\theta}_{(k,l)}, \hat{\theta}_{(g,h)} | R_{k\cdot}, R_{l\cdot}, R_{g\cdot}, R_{h\cdot}) = 0$ due to the assumption of independence (2.1). Our next step is to study the covariance in case that three indices are distinct, but the fourth is common. To that end, without loss of generality, we assume $l = g$ and have

$$\begin{aligned}
&\operatorname{Cov}(\hat{\theta}_{(k,l)}, \hat{\theta}_{(g,h)} \mid R_{k\cdot}, R_{l\cdot}, R_{g\cdot}, R_{h\cdot}) \\
&\quad = E(\hat{\theta}_{(k,l)}\hat{\theta}_{(l,h)} \mid R_{k\cdot}, R_{l\cdot}, R_{h\cdot}) - \theta^2 \\
&\quad = E\{(\boldsymbol{p} + \boldsymbol{\varepsilon}_k)'(\boldsymbol{p} + \boldsymbol{\varepsilon}_l)(\boldsymbol{p} + \boldsymbol{\varepsilon}_l)'(\boldsymbol{p} + \boldsymbol{\varepsilon}_h) \mid R_{k\cdot}, R_{l\cdot}, R_{h\cdot}\} - \theta^2 \\
&\quad = \boldsymbol{p}' E(\boldsymbol{\varepsilon}_l \boldsymbol{\varepsilon}_l' \mid R_{l\cdot}) \boldsymbol{p}.
\end{aligned}$$

For our last consideration, we study the case that $(k,l) = (g,h)$ and have

$$\begin{aligned}
&\operatorname{Cov}(\hat{\theta}_{(k,l)}, \hat{\theta}_{(g,h)} \mid R_{k\cdot}, R_{l\cdot}, R_{g\cdot}, R_{h\cdot}) \\
&\quad = E(\hat{\theta}_{(k,l)}^2 \mid R_{k\cdot}, R_{l\cdot}) - \theta^2 \\
&\quad = \boldsymbol{p}'\{E(\boldsymbol{\varepsilon}_k \boldsymbol{\varepsilon}_k' \mid R_{k\cdot}) + E(\boldsymbol{\varepsilon}_l \boldsymbol{\varepsilon}_l' \mid R_{l\cdot})\}\boldsymbol{p} + \operatorname{Tr}\{E(\boldsymbol{\varepsilon}_k \boldsymbol{\varepsilon}_k' \mid R_{k\cdot}) E(\boldsymbol{\varepsilon}_l \boldsymbol{\varepsilon}_l' \mid R_{l\cdot})\},
\end{aligned}$$

which can be approximated by

$$\boldsymbol{p}'\{E(\boldsymbol{\varepsilon}_k \boldsymbol{\varepsilon}_k' \mid R_{k\cdot}) + E(\boldsymbol{\varepsilon}_l \boldsymbol{\varepsilon}_l' \mid R_{l\cdot})\}\boldsymbol{p}$$

under the assumption that

$$\operatorname{Tr}\{E(\boldsymbol{\varepsilon}_k \boldsymbol{\varepsilon}_k' \mid R_{k\cdot}) E(\boldsymbol{\varepsilon}_l \boldsymbol{\varepsilon}_l' \mid R_{l\cdot})\} \ll \boldsymbol{p}'\{E(\boldsymbol{\varepsilon}_k \boldsymbol{\varepsilon}_k' \mid R_{k\cdot}) + E(\boldsymbol{\varepsilon}_l \boldsymbol{\varepsilon}_l' \mid R_{l\cdot})\}\boldsymbol{p}. \tag{2.3}$$

In our assumption (2.2), the coordinates of $\boldsymbol{\varepsilon}$ are typically small, and one may expect that the first term in (2.3) involving products of four such coordinates tends to be smaller than the second term involving the product of only two. To approximate the aforementioned correlations, we introduce one additional assumption that

$$\psi_i = \frac{\boldsymbol{p}' E(\boldsymbol{\varepsilon}_i \boldsymbol{\varepsilon}_i' \mid R_{i\cdot}) \boldsymbol{p}}{E\{\|\boldsymbol{\varepsilon}_i\|_2^2 \mid R_{i\cdot}\}\theta}, \qquad i = 1, \dots, n_R, \tag{2.4}$$

are independent of the replicate i, that is, equal to a constant $\bar{\psi}$. In the Appendix, we have shown that assumptions (2.3) and (2.4) hold under a simple yet reasonable model.

Under assumptions (2.3) and (2.4),

$$\mathrm{Cov}(\hat{\theta}_{(k,l)}, \hat{\theta}_{(g,h)} \mid R_{k.}, R_{l.}, R_{g.}, R_{h.}) = \theta\bar{\psi} \sum_{i\in\{k,l\}\cap\{g,h\}} E\{\|\boldsymbol{\varepsilon}_i\|_2^2 \mid R_{i.}\}.$$

Therefore, the optimal weighting scheme depends on only a good approximation to $E\{\|\boldsymbol{\varepsilon}_k\|_2^2 \mid R_{k.}\}$. To this end, we approximate $E\{\|\boldsymbol{\varepsilon}_k\|_2^2 \mid R_{k.}\}$ by the reciprocal of the kth diagonal element of the $n_R \times n_R$ matrix $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{p}}^{-1}$, where

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{p}} = \begin{pmatrix} \max(\|\hat{p}_1\|_2^2, \hat{\theta}) & \hat{\theta} & \cdots & \hat{\theta} \\ \hat{\theta} & \max(\|\hat{p}_2\|_2^2, \hat{\theta}) & \cdots & \hat{\theta} \\ \cdots & \cdots & \cdots & \cdots \\ \hat{\theta} & \hat{\theta} & \cdots & \max(\|\hat{p}_{n_R}\|_2^2, \hat{\theta}) \end{pmatrix}$$

is a regularized version of the covariance matrix of the random vectors $(\hat{p}_{1i}, \dots, \hat{p}_{n_R i})'$, and $\hat{\theta}$ is an initial estimator for θ such as that proposed by Boyd et al. (2009) and Parameswaran et al. (2013). Specifically, this amounts to estimating $E\{\|\boldsymbol{\varepsilon}_k\|_2^2 \mid R_{k.}\}$ by

$$\hat{e}_k = (\|\hat{p}_k\|_2^2 - \hat{\theta})_+ + \left(\sum_{j\neq k} \frac{1}{(\|\hat{p}_j\|_2^2 - \hat{\theta})_+} + \frac{1}{\hat{\theta}}\right)^{-1}.$$

In other words, while $\|\hat{p}_k\|_2^2 - \hat{\theta}$ should approximate $E\{\|\boldsymbol{\varepsilon}_k\|_2^2 \mid R_{k.}\}$, this new estimator adjusts the naive estimator upwardly to avoid negative values. Therefore, we may construct an estimator for $\boldsymbol{\Sigma}_0$, the $n_R(n_R-1)/2 \times n_R(n_R-1)/2$ dimensional variance-covariance matrix for the estimators $\{\hat{\theta}_{(k,l)}, 1 \le k < l \le n_R\}$ up to a constant multiplier. Denote the estimator by $\hat{\boldsymbol{\Sigma}}_0$, whose entry indexed by (k,l) and (g,h) is 0 if $\{k,l\}$ and $\{g,h\}$ are entirely distinct and $\sum_{i\in\{k,l\}\cap\{g,h\}} \hat{e}_i$, otherwise.

Next, we may construct a set of estimators for *clonality* based on $\{\hat{\theta}_{(k,l)}, 1 \le k < l \le n_R\}$ with different weighting schemes. For the first estimator $\hat{\theta}_1$, the weight for $\hat{\theta}_{(k,l)}$ is proportional to $R_{k.}R_{l.}$, which is proposed by Parameswaran et al. (2013). For the second estimator $\hat{\theta}_2$, the weights are proportional to $\hat{\boldsymbol{\Sigma}}_0^{-1}\mathbf{1}_{n_R(n_R-1)/2}$. In practice, we have found that the estimated covariance matrix is likely to be (nearly) singular, which results in large variability in weights used in the linear combination. Therefore, we consider three new estimators $\hat{\theta}_i$, $i = 3, 4, 5$, whose weights are proportional to $\widehat{\boldsymbol{\Sigma}}_j^{-1}\mathbf{1}_{n_R(n_R-1)/2}$, where $\widehat{\boldsymbol{\Sigma}}_j$, $j = 3, 4, 5$, are regularized counterparts of $\widehat{\boldsymbol{\Sigma}}_0$. Specifically,

$$\widehat{\boldsymbol{\Sigma}}_3 = \frac{1}{2}(\hat{\boldsymbol{\Sigma}}_0 + \bar{e}\mathbf{I}),$$

$$\widehat{\boldsymbol{\Sigma}}_4 = \frac{1}{2}\{\hat{\boldsymbol{\Sigma}}_0 + \mathrm{diag}(\hat{\boldsymbol{\Sigma}}_0)\}$$

and

$$\widehat{\mathbf{\Sigma}}_5 = \frac{1}{2}\{\hat{\mathbf{\Sigma}}_0 + \operatorname{diag}(\hat{\mathbf{\Sigma}}_0 - e_{\min}) + e_{\min}\mathbf{1}_{n_R(n_R-1)/2}\mathbf{1}'_{n_R(n_R-1)/2}\},$$

where $\mathbf{I}$ is the $n_R(n_R-1)/2$ by $n_R(n_R-1)/2$ identity matrix, $\bar{e} = n_R^{-1}\sum_{k=1}^{n_R}\hat{e}_k$, and $e_{\min} = \min\{\hat{e}_1, \dots, \hat{e}_{n_R}\}$. The choice of these three regularization schemes is out of convenience and simplicity: the first two are analogous to the ridge regularization to the covariance and correlation matrices, respectively, and the last one partially offsets the regularization of the second by adding a small positive constant to all off-diagonal elements of the matrix. There are other types of regularization in estimating the covariance matrix [see Fan, Liao and Liu (2016) and its references]. One unique aspect of our problem is that we focus on approximating the vector $\mathbf{\Sigma}_0^{-1}\mathbf{1}_{n_R(n_R-1)/2}$ rather than the covariance matrix itself.

Without prior knowledge, it is not clear which of the $\hat{\theta}_j$ performs better than others for the observed data. Instead of choosing among them, our proposed estimator is a new linear combination of the aforementioned five estimators $\{\hat{\theta}_j, j = 1, \dots, 5\}$. The weight again is proportional to $\widehat{\mathbf{\Sigma}}_\theta^{-1}\mathbf{1}_5$, where $\widehat{\mathbf{\Sigma}}_\theta$ is the jackknife estimator for the variance-covariance matrix of the random vector $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \dots, \hat{\theta}_5)'$ [Miller (1974)]. Specifically, for $j = 1, \dots, n_R$, we calculate $\hat{\boldsymbol{\theta}}$ without using the jth replicate. Denoting the resulting vector by $\hat{\boldsymbol{\theta}}^{(-j)}$, we have

$$\widehat{\mathbf{\Sigma}}_\theta = \frac{n_R - 1}{n_R}\sum_{j=1}^{n_R}(\hat{\boldsymbol{\theta}}^{(-j)} - \hat{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}^{(-j)} - \hat{\boldsymbol{\theta}})'.$$

In summary, the final estimator $\hat{\theta}_F$ can be constructed via the following steps:

1. Obtain $\{\hat{\theta}_{(k,l)}\}$ and $\{\hat{e}_k\}$.
2. Obtain $\widehat{\mathbf{\Sigma}}_j$, $j = 0, 3, 4, 5$ and $\hat{\theta}_j$, $1 \le j \le 5$ as linear combinations of $\hat{\theta}_{(k,l)}$.
3. Obtain $\widehat{\mathbf{\Sigma}}_\theta$ and $\hat{\theta}_F$ as the corresponding linear combination of $\{\hat{\theta}_j, 1 \le j \le 5\}$.

REMARK 1. Even with the proposed regularization, $\hat{\theta}_j$ (and $\hat{\theta}_F$) may sometimes be negative! One remedy is to force all the weights to be nonnegative by introducing additional regularization [Tian et al. (2005)]. Specifically, let

$$\mathbf{w}(\lambda) = \{\widehat{\mathbf{\Sigma}}_j + \lambda \operatorname{diag}(\widehat{\mathbf{\Sigma}}_j)\}^{-1}\mathbf{1}_{n_R(n_R-1)/2}$$

for $\lambda \ge 0$. We may choose a small λ so that all the components of $\mathbf{w}(\lambda)$ are nonnegative. The weights used in combining estimators are then obtained by rescaling the corresponding $\mathbf{w}(\lambda)$ so that the rescaled weights sum to 1. One formal way to identify the nonnegative weights is to solve the quadratic programming problem

$$\min_{\boldsymbol{w}} \boldsymbol{w}'\widehat{\mathbf{\Sigma}}_j\boldsymbol{w} \qquad \text{subject to } \boldsymbol{w} \ge \mathbf{0},\ \boldsymbol{w}'\mathbf{1}_{n_R(n_R-1)/2} = 1,$$

where the inequality $\boldsymbol{w} \ge \mathbf{0}$ is component-wise.

REMARK 2. Our final estimator $\hat{\theta}_F$ is also a linear combination of $\{\hat{\theta}_{(k,l)}, 1 \leq k < l \leq n_R\}$. Ideally we want to estimate *clonality* as a direct linear combination of these $n_R(n_R - 1)/2$ unbiased estimators $\{\hat{\theta}_{(k,l)}, 1 \leq k < l \leq n_R\}$. However, the optimal weights of such a linear combination depend on $\mathbf{\Sigma}_0$, the variance-covariance matrix of $\{\hat{\theta}_{(k,l)}, 1 \leq k < l \leq n_R\}$. Without additional and often unverifiable assumptions, it is not estimable with a limited number of replicates. For example, it is clear that the jackknife procedure does not apply here. Consequently, the performance of these five proposed initial estimators as direct combinations of $\hat{\theta}_{(k,l)}$ can be poor in practice. The biggest advantage of constructing the final estimator from these five rather than the original $n_R(n_R - 1)/2$ simple estimators is that while the proposed estimator $\widehat{\mathbf{\Sigma}}_0$ may be inaccurate due to violating assumption (2.3) or (2.4), the variance-covariance matrix of the five estimators can be reliably estimated via the nonparametric jackknife procedure. Thus, the weights of $\hat{\theta}_{(k,l)}$ used in constructing $\hat{\theta}_F$ are still adaptive to the data without relying entirely on the quality of the estimator of $\mathbf{\Sigma}_0$. Because of this adaptivity, we have found that the empirical performance of $\hat{\theta}_F$ is fairly robust. When n_R is small, it is prudent to consider fewer initial estimators since the jackknife procedure may fail to estimate accurately the covariances of initial estimators.

3. Simulation study. In this section, we show the results from a series of simulations to study the finite sample performance of the proposed estimator by comparing it with some existing alternatives. We simulate the observed data mimicking the underlying mechanism of the sequencing process. Specifically, the counts $\mathbf{R}_i$ are generated via the following steps:

1. Simulate the number of cells of different clones from a multinomial distribution $\mathrm{MN}(n_i, \tilde{\boldsymbol{p}}_i)$ for the ith replicate, where the total number of cells in the replicate, n_i, is given a priori, $\tilde{\boldsymbol{p}}_i \sim \mathrm{Dirichlet}(100C, \boldsymbol{p})$, reflects the inhomogeneity of clone distributions in serum samples and $\boldsymbol{p}$ is the true clone probability.

2. Denote the number of cells of the kth clone by $n_{ik}, k = 1, \ldots, C$. Simulate the number of reads originating from each cell independently from a distribution $\mathcal{L} \times 2^{\mathcal{S}}$, where $\mathcal{L} \sim \mathrm{Bin}(1, 0.1)$ and $\mathcal{S} \sim \mathrm{Bin}(12, 0.75)$ represent if the DNA fragment has successfully ligated and the number of successful PCR cycles, respectively.[2] Let the generated reads be $\{\zeta_1, \ldots, \zeta_{n_i}\}$, where n_i is the total number of sampled cells. Without loss of generality, we assume that all cells are sorted so that the first n_{i1} cells are from the first clone, the next n_{i2} cells are from the second clone, and so forth. Then there are $r_{i1} = \sum_{j=1}^{n_{i1}} \zeta_j$ reads from the first clone

[2] The choice of these parameters quantitatively but not qualitatively affects the simulation results presented in this section.

and

$$r_{ij} = \sum_{j=n_{i1}+\cdots+n_{i(j-1)}+1}^{n_{i1}+\cdots+n_{ij}} \zeta_j$$

reads from the jth clone for $j = 2, \ldots, C$.

3. Simulate the number of sequenced (observed) reads of the jth clone from independent Poisson distributions, that is,

$$R_{ij} \sim \text{Poisson}\left(\frac{r_{ij}\tilde{R}_i}{\sum_{j=1}^{C} r_{ij}} \right), \qquad j = 1, \ldots, C,$$

where $\tilde{R}_i$ is the expected sequencing depth for the ith replicate.

The observed data consist of all nonzero R_{ij}s. Thus,

$$E(\hat{p}_{ij}) = E\left(\frac{\sum_{j=n_{i1}+\cdots+n_{i(j-1)}+1}^{n_{i1}+\cdots+n_{ij}} \zeta_j}{\sum_{j=1}^{n_i} \zeta_j} \right) = E\left(\frac{n_{ij}}{n_i} \right) = E(\tilde{p}_{ij}) = p_j;$$

and therefore it satisfies our key assumption (2.2). It is also clear that $\mathbf{R}_i = (R_{i1}, \ldots, R_{iC})'$ does not follow a multinomial distribution.

For each simulated data set, we constructed five estimators of *clonality*: (1) the proposed estimator $\hat{\theta}_{\text{regF}}$ with regularization to ensure nonnegative weights; (2) the weighted estimator $\hat{\theta}_{w_1} = \hat{\theta}_w^*$ with $w(r_1, r_2) = r_1 r_2$; (3) the weighted estimator $\hat{\theta}_{w_2} = \hat{\theta}_w^*$ with $w(r_1, r_2) = r_1 r_2/(r_1 + r_2)$; (4) the naïve estimator

$$\hat{\theta}_{\text{Naive}} = \sum_{j=1}^{C} \frac{R_{.j}^2}{R_{..}^2},$$

where $R_{..} = \sum_{j=1}^{C} R_{.j}$; and (5) the unbiased estimator under the multinomial assumption

$$\hat{\theta}_{\text{MN}} = \sum_{j=1}^{C} \frac{R_{.j}(R_{.j} - 1)}{R_{..}(R_{..} - 1)}.$$

The performance of an estimator of *clonality* $\hat{\hat{\theta}}$ is measured by the base 2 log-transformed empirical mean squared error (MSE)

$$\log_2\{E(\hat{\hat{\theta}} - \theta_0)^2\}$$

using estimators from 500 independently simulated datasets. We have also estimated the relative bias for each estimator defined as

$$E(\hat{\hat{\theta}} - \theta)/\theta.$$

In the first set of simulations, we let $n_R = 8$, $C = 10^6$ and $p_j \propto \{\log(1+j)\}^{-1}$. For the total numbers of cells sampled across replicates, we let $(n_1, \ldots, n_8)'$ be $\mathbf{n}_1 = (8, \ldots, 8)' \times 10^5$ or $\mathbf{n}_2 = (16, 16, 8, 6, 4, 2, 1, 1)' \times 10^5$. For sequencing depths across replicates, we let $(\tilde{R}_1, \ldots, \tilde{R}_8)$ be $\tilde{\mathbf{R}}_1 = (4, \ldots, 4)' \times 10^5$ or $\tilde{\mathbf{R}}_2 = (2, 2, 4, 4, 4, 4, 8, 8)' \times 10^5$. In the simulation settings 1–4, we consider all four different combinations of numbers of sampled cells and sequencing depths in the order of $(\mathbf{n}_1, \tilde{\mathbf{R}}_1)$, $(\mathbf{n}_1, \tilde{\mathbf{R}}_2)$, $(\mathbf{n}_2, \tilde{\mathbf{R}}_1)$, and $(\mathbf{n}_2, \tilde{\mathbf{R}}_2)$. We also considered $p_j \propto j^{-0.25}$, $j^{-0.5,}$ and j^{-1} in simulation settings 5–8, 9–12, and 13–16, respectively. The true *clonality* was 1.01×10^{-6}, 1.12×10^{-6}, 3.60×10^{-6}, and 7.9×10^{-3} for aforementioned four choices of $\boldsymbol{p}$. Simulation results on MSE and bias are summarized in Tables 1 and 2, respectively. The results show clearly that the performance of the proposed estimator is superior to other competitors across

TABLE 1
The $\log_2(MSE)$ *of* $\hat{\theta}_{\text{regF}}$ *and alternatives including* $\hat{\theta}_{\text{Naive}}$, $\hat{\theta}_{\text{MN}}$, $\hat{\theta}_{w_1}$, *and* $\hat{\theta}_{w_2}$ *in estimating clonality based on* 500 *simulations when the sequencing reads do not follow a multinomial distribution*

Setting	$(\mathbf{n}_1, \mathbf{R}_1)$	$(\mathbf{n}_1, \mathbf{R}_2)$	$(\mathbf{n}_2, \mathbf{R}_1)$	$(\mathbf{n}_2, \mathbf{R}_2)$
		$p_j \propto \{\log(1+j)\}^{-1}$		
$\hat{\theta}_{\text{regF}}$	−52.82	−52.96	−52.66	−52.49
$\hat{\theta}_{w_1}$	−52.82	−52.84	−51.59	−49.89
$\hat{\theta}_{w_2}$	−52.82	−52.94	−51.59	−50.68
$\hat{\theta}_{\text{Naive}}$	−36.24	−35.71	−33.10	−30.80
$\hat{\theta}_{\text{MN}}$	−36.51	−35.90	−33.19	−30.83
		$p_j \propto j^{-0.25}$		
$\hat{\theta}_{\text{regF}}$	−52.72	−52.68	−52.39	−52.36
$\hat{\theta}_{w_1}$	−52.72	−52.59	−51.27	−49.62
$\hat{\theta}_{w_2}$	−52.72	−52.67	−51.27	−50.44
$\hat{\theta}_{\text{Naive}}$	−36.24	−35.71	−33.10	−30.80
$\hat{\theta}_{\text{MN}}$	−36.51	−35.90	−33.19	−30.83
		$p_j \propto j^{-0.5}$		
$\hat{\theta}_{\text{regF}}$	−47.60	−47.53	−47.19	−47.01
$\hat{\theta}_{w_1}$	−47.59	−47.37	−46.08	−44.96
$\hat{\theta}_{w_2}$	−47.59	−47.53	−46.08	−45.49
$\hat{\theta}_{\text{Naive}}$	−36.24	−35.71	−33.10	−30.80
$\hat{\theta}_{\text{MN}}$	−36.51	−35.90	−33.19	−30.83
		$p_j \propto j^{-1.0}$		
$\hat{\theta}_{\text{regF}}$	−27.59	−27.47	−25.25	−25.02
$\hat{\theta}_{w_1}$	−27.72	−27.58	−26.18	−25.15
$\hat{\theta}_{w_2}$	−27.72	−27.76	−26.17	−25.66
$\hat{\theta}_{\text{Naive}}$	−27.71	−27.47	−26.17	−24.93
$\hat{\theta}_{\text{MN}}$	−27.71	−27.47	−26.17	−24.93

TABLE 2

The relative bias of $\hat{\theta}_{\text{regF}}$ and alternatives including $\hat{\theta}_{\text{Naive}}$, $\hat{\theta}_{\text{MN}}$, $\hat{\theta}_{w_1}$, and $\hat{\theta}_{w_2}$ in estimating clonality based on 500 *simulations when the sequencing reads do not follow a multinomial distribution*

Setting	$(\mathbf{n}_1, \mathbf{R}_1)$	$(\mathbf{n}_1, \mathbf{R}_2)$	$(\mathbf{n}_2, \mathbf{R}_1)$	$(\mathbf{n}_2, \mathbf{R}_2)$
		$p_j \propto \{\log(1+j)\}^{-1}$		
$\hat{\theta}_{\text{regF}}$	0.010	0.009	0.009	0.010
$\hat{\theta}_{w_1}$	0.010	0.009	0.009	0.010
$\hat{\theta}_{w_2}$	0.010	0.009	0.009	0.010
$\hat{\theta}_{\text{Naive}}$	3.473	4.178	10.30	22.89
$\hat{\theta}_{\text{MN}}$	3.164	3.903	9.994	22.62
		$p_j \propto j^{-0.25}$		
$\hat{\theta}_{\text{regF}}$	0.009	0.009	0.009	0.009
$\hat{\theta}_{w_1}$	0.009	0.009	0.009	0.009
$\hat{\theta}_{w_2}$	0.009	0.009	0.009	0.009
$\hat{\theta}_{\text{Naive}}$	3.12	3.757	9.265	20.59
$\hat{\theta}_{\text{MN}}$	2.84	3.510	9.987	20.34
		$p_j \propto j^{-0.5}$		
$\hat{\theta}_{\text{regF}}$	0.003	0.002	0.004	0.002
$\hat{\theta}_{w_1}$	0.003	0.003	0.003	0.004
$\hat{\theta}_{w_2}$	0.003	0.002	0.003	0.004
$\hat{\theta}_{\text{Naive}}$	0.974	1.172	2.891	6.421
$\hat{\theta}_{\text{MN}}$	0.887	1.095	2.804	6.343
		$p_j \propto j^{-1.0}$		
$\hat{\theta}_{\text{regF}}$	0.001	0.001	−0.002	0.000
$\hat{\theta}_{w_1}$	0.001	0.001	−0.001	0.000
$\hat{\theta}_{w_2}$	0.000	0.000	−0.001	0.000
$\hat{\theta}_{\text{Naive}}$	0.001	0.001	0.001	0.003
$\hat{\theta}_{\text{MN}}$	0.001	0.001	0.001	0.003

all cases. For example, the MSE of the proposed estimator is 80% less than that of $\hat{\theta}_{w_1}$ or $\hat{\theta}_{w_2}$ and only a tiny fraction of that of $\hat{\theta}_{\text{MN}}$, when $p_j \propto \{\log(1+j)\}^{-1}$ with $(\mathbf{n}_2, \tilde{\mathbf{R}}_2)$. The performance of $\hat{\theta}_{\text{MN}}$ assuming a multinomial distribution for $\mathbf{R}_i$ is approximately the same as that of the naïve estimator and substantially worse than those based on $\hat{\theta}_{(k,l)}$s. In the combination of $\mathbf{n}_1$ and $\tilde{\mathbf{R}}_1$, all replicates are actually identically distributed. One may expect that $\hat{\theta}_{w_1} = \hat{\theta}_{w_2}$ with equal weights for all $\hat{\theta}_{(k,l)}$ is optimal. It is confirmed by the simulation. On the other hand, the proposed estimator performs similarly well. In general, the proposed estimator has almost no bias, while both $\hat{\theta}_{\text{Naive}}$ and $\hat{\theta}_{\text{MN}}$ can be severely biased, which contributes to their inflated MSEs. When $\rho = -1$, all estimators including the naïve estimator perform quite well. It may be due to the observation that in such a setting, the

TABLE 3

The $\log_2(MSE)$ *of* $\hat{\theta}_{\text{regF}}$ *and alternatives including* $\hat{\theta}_{\text{Naive}}$, $\hat{\theta}_{\text{MN}}$, $\hat{\theta}_{w_1}$, *and* $\hat{\theta}_{w_2}$ *in estimating clonality based on* 500 *simulations when the sequencing reads follow a multinomial distribution*

	$p_j \propto j^{-0.25}$		$p_j \propto j^{-0.5}$		$p_j \propto j^{-1}$	
Setting	$\mathbf{n}_1$	$\mathbf{n}_2$	$\mathbf{n}_1$	$\mathbf{n}_2$	$\mathbf{n}_1$	$\mathbf{n}_2$
$\hat{\theta}_{\text{regF}}$	−56.63	−61.01	−48.99	−51.86	−28.74	−31.75
$\hat{\theta}_{w_1}$	−56.63	−61.07	−48.98	−51.95	−28.81	−31.90
$\hat{\theta}_{w_2}$	−56.63	−60.84	−48.98	−51.66	−28.81	−31.57
$\hat{\theta}_{\text{Naive}}$	−38.57	−44.69	−38.57	−44.69	−28.81	−31.94
$\hat{\theta}_{\text{MH}}$	−56.42	−57.39	−48.97	−51.97	−28.81	−31.94

clonality is determined by the cell probabilities of few large clones, which can be estimated accurately.

In the second set of simulations, we let $\mathbf{R}_i$ follow a simple multinomial distribution $\text{MN}(\tilde{R}_i, \boldsymbol{p})$. In this case, one may expect that $\hat{\theta}_{\text{MN}}$ is unbiased and performs relatively well. We also let $(n_1, \ldots, n_8)'$ be $\mathbf{n}_1$ or $\mathbf{n}_2$ and $p_j \propto \{\log(j+1)\}^{-1}$ and $j^{-\rho}$ for $\rho = 0.25, 0.5$, and 1. The results are summarized in Table 3. The results show that under the multinomial assumption, $\hat{\theta}_{\text{MN}}$ indeed performs better than others in most settings; however, the loss in the finite sample performance for the proposed estimators is at most moderate.

The observed gain in estimation accuracy can greatly increase the power of the subsequent statistical analysis concerning *clonality*. For example, one may be interested in comparing *clonality* between two groups of patients. Assuming that $p_j \propto j^{-0.25}$ in the first group and $\propto \{\log(1+j)\}^{-1}$ in the second group, that is, $\theta = 1.12 \times 10^{-6}$ versus 1.01×10^{-6}, the power depends on the effect size of the test, which is the signal to noise ratio and can be easily estimated via simulations. Treating the test based on the proposed estimator as the reference, the asymptotic relative efficiency (ARE) of tests based on $\hat{\theta}_{w1}$, $\hat{\theta}_{w2}$, $\hat{\theta}_{\text{Naive,}}$ and $\hat{\theta}_{\text{MH}}$ is estimated as 0.25, 0.25, 0.01, and 0.01, respectively, for the setting with $\mathbf{n}_2$ and $\tilde{\mathbf{R}}_1$. The ARE between two tests is the ratio of squared effective sizes and can be interpreted as the inverse of the ratio of the sample sizes needed to achieve the same power. Therefore, the test based on the proposed estimator is substantially more powerful than alternatives, which require at least 300% more patients to reach a comparable power. To better quantify the gain, we conducted a more realistic simulation. Under the same setup, we generate clone frequencies for two groups of patients. For the ith patient from group k, the probability $p_j \propto j^{-\rho_{ik}}$, $j = 1, \ldots, 10^6$, where ρ_{ik} is drawn from beta distribution $\text{Beta}(100, 300)$ and $\text{Beta}(150, 400)$ for $k = 1$ and 2, respectively. We estimated the power of the unequal variance t-test (Behrens–Fisher test) comparing *clonality* between two groups of 15 patients each under different combinations of sequencing depth and cell number based on 500 simulated datasets. Although strictly speaking, the normality assumption of the t-test

TABLE 4
The empirical power of two sample t-test based on θ, $\hat{\theta}_{\mathrm{regF}}$, $\hat{\theta}_{\mathrm{Naive}}$, $\hat{\theta}_{\mathrm{MN}}$, $\hat{\theta}_{w_1}$, *and* $\hat{\theta}_{w_2}$ *comparing the clonality of* $p_j \propto j^{-\rho}$, $\rho \sim \mathrm{Beta}(100, 300)$ *with that of* $p_j \propto j^{-\rho}$, $\rho \sim \mathrm{Beta}(150, 400)$ *with* 15 *patients per group*

Setting	$(\mathbf{n}_1, \mathbf{R}_1)$	$(\mathbf{n}_1, \mathbf{R}_2)$	$(\mathbf{n}_2, \mathbf{R}_1)$	$(\mathbf{n}_2, \mathbf{R}_2)$
θ	74%	74%	74%	74%
$\hat{\theta}_{\mathrm{regF}}$	72%	73%	72%	73%
$\hat{\theta}_{w_1}$	71%	70%	24%	7%
$\hat{\theta}_{w_2}$	71%	70%	24%	7%
$\hat{\theta}_{\mathrm{Naive}}$	72%	73%	65%	52%
$\hat{\theta}_{\mathrm{MH}}$	72%	73%	65%	61%

is violated here, the t-test is still approximately valid and commonly used in practice.[3] The results are summarized in Table 4. It is clear that the power of the test using the proposed estimator is almost as good as that using the true *clonality*, while the tests based on other estimators may suffer substantial power loss under selected settings.

The power loss is not the only consequence of using an inaccurate estimator and we conducted an additional set of simulations to examine the dilution of the correlation coefficient caused by not using the true *clonality.* To this end, we simulate *clonality* for 1000 patients, where $p_j \propto j^{-\rho}$, $j = 1, \dots, 10^6$, and $\rho \sim \mathrm{Beta}(60, 300)$. The simulated ρs are in the range of $[0.1, 0.25]$. We then simulate an independent variable $Z = (\rho + \epsilon_\rho)/\sqrt{2}$, where the independent error ϵ_ρ follows the same beta distribution as ρ. We calculate the correlation coefficient between Z and the true *clonality* as well as the estimators thereof. The results are summarized in Table 5. The proposed estimator has the smallest dilution for the correlation coefficient between Z and the true *clonality* across all settings. Dilution from other estimators can be big, which would have a substantial negative impact on the estimation and hypothesis testing for the correlation coefficient of interest.

4. Example. In this section we illustrate the application of our approach in a recent study conducted by Qi et al. (2014). The T-cell receptors (TCRs) diversity plays a crucial role in determining the ability of the immune system to efficiently respond to various pathogenic challenges [Qi et al. (2014)]. It is desirable to investigate human TCR diversity by sequencing a large number of sequences and extrapolating the information based on an appropriate statistical method [Boyd et al. (2009)]. Among the several metrics for TCR diversity are *richness* and *clarity*;

[3] Based on additional simulations (not shown here; available from the first author), the empirical Type I errors based on different estimators of clonality varied from 0.03 to 0.06 with 15 patients per group, further confirming the validity the t-test.

TABLE 5
The estimated correlation coefficient between Z and θ, $\hat{\theta}_{\text{regF}}$, $\hat{\theta}_{\text{Naive}}$, $\hat{\theta}_{\text{MN}}$, $\hat{\theta}_{w_1}$, and $\hat{\theta}_{w_2}$ based on 1000 *patients*

Setting	$(\mathbf{n}_1, \mathbf{R}_1)$	$(\mathbf{n}_1, \mathbf{R}_2)$	$(\mathbf{n}_2, \mathbf{R}_1)$	$(\mathbf{n}_2, \mathbf{R}_2)$
θ	0.72	0.72	0.72	0.72
$\hat{\theta}_{\text{regF}}$	0.65	0.66	0.62	0.60
$\hat{\theta}_{w_1}$	0.60	0.53	0.12	0.03
$\hat{\theta}_{w_2}$	0.60	0.53	0.12	0.03
$\hat{\theta}_{\text{Naive}}$	0.65	0.65	0.46	0.27
$\hat{\theta}_{\text{MH}}$	0.65	0.66	0.46	0.36

we focus on the latter. In the study by Qi et al. (2014), five replicate TCR libraries of CD4 naïve T cells, CD4 memory T cells, CD8 naïve T cells, and CD8 memory T cells are sequenced from seven participants. The sequencing depth $R_{i.}$ varied from 8.9×10^4 to 7.4×10^5 with a median level of 3.2×10^5. Figure 1 shows the observed cumulative proportions of CD4 naïve and memory T cells from the

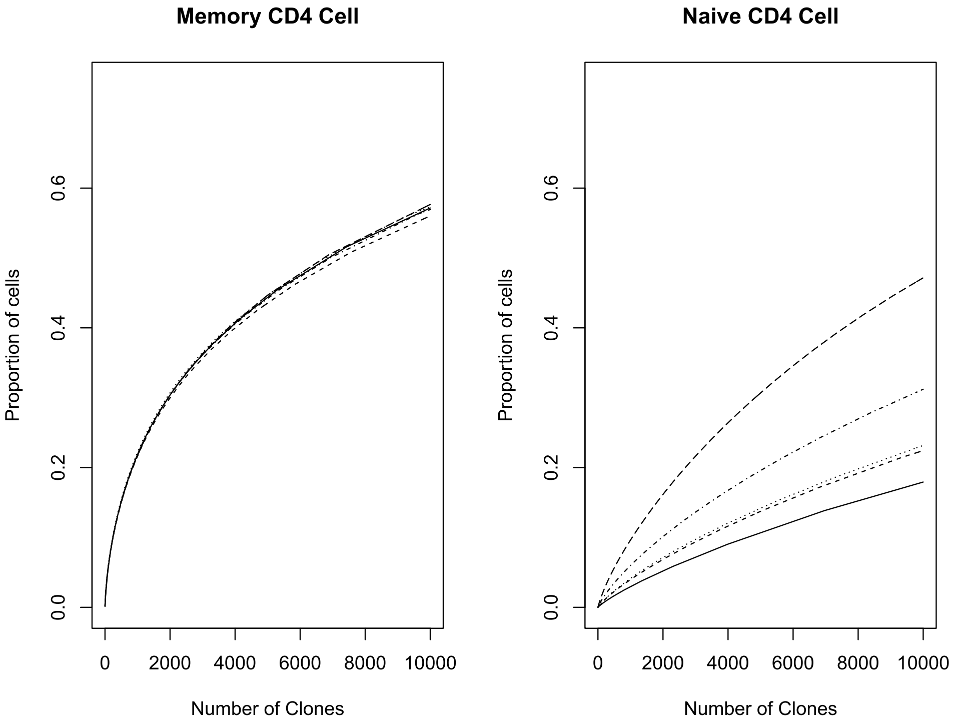

FIG. 1. *The cumulative proportion of CD*4 *naïve and memory T cells from the top* 10,000 *clones for a* 33 *year-old female. Five curves represent results from independent replicates.*

top clones of a 33 year-old female. From the figure, it is clear that the top 10,000 clones account for a bigger proportion of CD4 memory T cells than naïve CD4 T cells, reflecting the relative evenness of the distribution of clone sizes of naïve T cells. Due to the limited number of sampled T cells ($\sim 10^{-4}\%$ out of the total T-cell repertoire) as well as the sequencing depth, the estimation for the entire distribution of clone sizes is very difficult. However, because there were multiple independent measurements per individual, we may apply the proposed method to estimate *clonality* for a given subset of T cells for that individual. Specifically, we calculate the proposed estimator $\hat{\theta}_{\text{regF}}$ as well as the naïve estimator $\hat{\theta}_{\text{Naive}}$ for *clonality* of CD4 naïve and memory T cells for each of the seven participants. First, the *clonality* of CD4 memory T cells is several orders of magnitudes greater than that of the CD4 naïve T cells, confirming our previous observation on the evenness of the distribution of the naïve T cell clones. Second, for the same reason, while the proposed and naïve estimators yield similar estimates of *clonality* for memory T cells, they become very different for naïve T cells. For example, for the individual in Figure 1, the *clonality* estimates $(\hat{\theta}_{\text{Naive}}, \hat{\theta}_{\text{regF}}) = (8.35 \times 10^{-5}, 8.06 \times 10^{-5})$ for CD4 memory T cells and $(3.31 \times 10^{-6}, 3.34 \times 10^{-8})$ for CD4 naïve T cells. This is similar to the pattern observed in the simulation study, and we believe that the proposed estimator is much more reliable than its naïve counterpart for estimating *clonality* of more uniformly distributed CD4 naïve T cell clones.

We may estimate the variance of $\log(\hat{\theta}_{\text{regF}})$ by its jackknife variance estimator denoted by $\hat{\sigma}_\theta$, that is, performing a second round of jackknife in addition to the jackknife step used for estimating the variance-covariance matrix of $(\hat{\theta}_1, \dots, \hat{\theta}_5)'$. We then may construct a 95% confidence interval for *clonality* by approximating the distribution of $\{\log(\hat{\theta}_{\text{regF}}) - \log(\theta)\}/\hat{\sigma}_\theta$ with a t-distribution of $n_R - 1$ degree of freedom. For the same individual above, the 95% confidence interval of *clonality* is $(7.80 \times 10^{-5}, 8.33 \times 10^{-5})$ for the memory CD4 T cells and $(2.16 \times 10^{-8}, 5.14 \times 10^{-8})$ for the naïve CD4 T cells. Due to the small number of replicates, these confidence intervals need to be interpreted cautiously.

Last, we compare *clonality* of naïve CD4 T cells between three participants younger than 40 and four participants older than 70 using a two sample t-test. *Clonalities* are log transformed before the test. The p value based on $\hat{\theta}_{\text{regF}}$ is 0.08 with the estimated average log-transformed *clonality* being -11.92 for old participants and -15.16 for young participants, suggesting that the immune diversity of older participants is substantially lower than that of younger participants with a marginal statistical significance. The test based on $\hat{\theta}_{\text{Naive}}$ yields similar result. However, the estimated average log-transformed *clonality* becomes -11.60 for old participants and -12.59 for young participants, representing a much smaller group difference. We have also compared *clonality* of memory CD4 T cells between three young and four old participants. The results based on $\hat{\theta}_{\text{regF}}$ and $\hat{\theta}_{\text{Naive}}$ are almost identical. It is not a surprise, since two methods yield similar estimates of *clonality* for memory T cells as mentioned above. Based on $\hat{\theta}_{\text{regF}}$, the estimated average log-transformed *clonality* is -7.43 for old participants and -9.29 for young participants and the two-sided p-value is 0.049, also confirming the hypothesis that immune diversity in old participants is lower than that in young participants. The sample size of the analysis is limited, partly due to the sequencing cost. We may expect that with continuously dropping cost, similar data may be collected in a much larger cohort of patients, allowing more sophisticated analysis with clearer advantages of using the proposed estimator.

5. Discussion. Estimating various functionals of $\boldsymbol{p}$ is particularly important in evaluating the health of the immune system of a patient. This may be a patient with a hematopoietic malignancy. It is known that such patients may have compromised immune systems of severely depleted cell diversity [Laydon, Bangham and Asquith (2015), Section 2 and its references]. In such a case, *clonality* can be a sensitive health indicator of the immune system. The goal of this paper is to demonstrate that from data on frequencies of V(D)J rearrangements, our new approach can accurately estimate the underlying *clonality*. We note that this methodology is applicable to data sets in many other biological contexts, where data are obtained by repeated sampling from a large population, and the proportions of different species within the total population are of interest. Population, whose diversity is of potential interest, ranges from molecules to cells, to microorganisms, and to large ecosystems.

If the multinomial assumption is problematic, for example, when the sampling probability varies from capture to recapture, the proposed method is especially useful. It is possible to consider more direct relaxation of the multinomial assumption such as employing the Dirichlet-multinomial model. Such parametric assumptions may help to estimate the variance-covariance matrix $\boldsymbol{\Sigma}_0$ and, more importantly, facilitate the statistical inference of the *clonality* estimator, which warrants further research.

The method can also be employed to estimate other functionals of $\boldsymbol{p}$ given adequate number of replicates. For example, $\eta = \sum_{i=1}^{C} p_i^3$, which is useful for further differentiating immune systems with similar *clonalities*, can be estimated by suitably combining $\{\sum_{i=1}^{C} \hat{p}_{ji}\hat{p}_{ki}\hat{p}_{li}, 1 \le j < k < l \le n_R\}$. The optimal choice of the weighting scheme used in such a combination is more complicated and a topic of future research.

The R-package *lymphclon* implementing our approach can be found in the CRAN repository.

APPENDIX: MODEL ASSUMPTIONS ON ESTIMATING $\boldsymbol{\Sigma}_0$

The proposed estimation procedure for $\boldsymbol{\Sigma}_0$ relies on assumptions (2.3) and (2.4). To assess their plausibility in practice, we consider the following data generating process for the ith replicate: first, n_i cells are sampled and numbers of cells from different clones follow a multinomial distribution $\mathrm{MN}(n_i, \boldsymbol{p})$; second, the observed count vector $\mathbf{R}_i$ is generated from another multinomial distribution $\mathrm{MN}(R_{i.}, \hat{\boldsymbol{p}}_{Ri})$, where $\hat{\boldsymbol{p}}_{Ri}$ is a random probability vector with $E(\hat{\boldsymbol{p}}_{Ri}|\hat{\boldsymbol{p}}_{Ni}) = \hat{\boldsymbol{p}}_{Ni}$ and $\mathrm{var}(\hat{\boldsymbol{p}}_{Ri}|\hat{\boldsymbol{p}}_{Ni}) = \tilde{\boldsymbol{\Sigma}}/n_i$, $\hat{\boldsymbol{p}}_{Ni} = (n_{i1}, \dots, n_{iC})'/n_i$, and n_{ij} is the number of sampled cells from clone j. Under these simple assumptions, $\mathbf{R}_i | R_{i.}$ does not follow the multinomial distribution typically assumed in the literature. One may calculate that

$$\boldsymbol{p}' E(\boldsymbol{\varepsilon}_k \boldsymbol{\varepsilon}_k' \mid R_{k.})\boldsymbol{p} = \left(\frac{1}{n_k} + \frac{1}{R_{k.}} - \frac{1}{n_k R_{k.}}\right)(\eta - \theta^2) + \left(\frac{1}{n_k} - \frac{1}{n_k R_{k.}}\right)\boldsymbol{p}'\tilde{\boldsymbol{\Sigma}}\boldsymbol{p},$$

$$E(\|\boldsymbol{\varepsilon}_k\|_2^2 \mid R_{k.}) = \left(\frac{1}{n_k} + \frac{1}{R_{k.}} - \frac{1}{n_k R_{k.}}\right)(1 - \theta) + \left(\frac{1}{n_k} - \frac{1}{n_k R_{k.}}\right)\mathrm{Tr}(\tilde{\boldsymbol{\Sigma}})$$

and $\mathrm{Tr}\{E(\boldsymbol{\varepsilon}_k \boldsymbol{\varepsilon}_k' \mid R_{k.})E(\boldsymbol{\varepsilon}_l \boldsymbol{\varepsilon}_l' \mid R_{l.})\}$

$$\begin{aligned} &= \mathrm{Tr}\left[\left\{\left(\frac{1}{n_k} + \frac{1}{R_{k.}} - \frac{1}{n_k R_{k.}}\right)\boldsymbol{\Sigma}_p + \left(\frac{1}{n_k} - \frac{1}{n_k R_{k.}}\right)\tilde{\boldsymbol{\Sigma}}\right\}\right. \\ &\qquad \times \left.\left\{\left(\frac{1}{n_l} + \frac{1}{R_{l.}} - \frac{1}{n_l R_{l.}}\right)\boldsymbol{\Sigma}_p + \left(\frac{1}{n_l} - \frac{1}{n_l R_{l.}}\right)\tilde{\boldsymbol{\Sigma}}\right\}\right], \end{aligned}$$

where $\eta = \sum_{j=1}^C p_j^3$, and $\boldsymbol{\Sigma}_p = \mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}'$. Therefore, if $\tilde{\boldsymbol{\Sigma}} \approx 0$, that is, $\hat{\boldsymbol{p}}_{Ri} \approx \hat{\boldsymbol{p}}_{Ni}$, then

$$\boldsymbol{p}' E(\boldsymbol{\varepsilon}_k \boldsymbol{\varepsilon}_k' \mid R_{k.})\boldsymbol{p} \approx \left(\frac{1}{n_k} + \frac{1}{R_{k.}} - \frac{1}{n_k R_{k.}}\right)(\eta - \theta^2),$$

$$E(\|\boldsymbol{\varepsilon}_k\|_2^2 \mid R_{k.}) \approx \left(\frac{1}{n_k} + \frac{1}{R_{k.}} - \frac{1}{n_k R_{k.}}\right)(1 - \theta)$$

and $\mathrm{Tr}\{E(\boldsymbol{\varepsilon}_k \boldsymbol{\varepsilon}_k' \mid R_{k.})E(\boldsymbol{\varepsilon}_l \boldsymbol{\varepsilon}_l' \mid R_{l.})\}$

$$\approx \left(\frac{1}{n_k} + \frac{1}{R_{k.}} - \frac{1}{n_k R_{k.}}\right)\left(\frac{1}{n_l} + \frac{1}{R_{l.}} - \frac{1}{n_l R_{l.}}\right)(\theta + \theta^2 - 2\eta),$$

which implies that

$$\psi_i \approx \frac{(\eta - \theta^2)}{\theta(1 - \theta)} \geq 0,$$

independent of i, and the term $\mathrm{Tr}\{E(\boldsymbol{\varepsilon}_k \boldsymbol{\varepsilon}_k' \mid R_{k.})E(\boldsymbol{\varepsilon}_l \boldsymbol{\varepsilon}_l' \mid R_{l.})\}$ is negligible when $\min(n_k, R_{k.}, n_l, R_{l.}) \gg \theta/\eta$. The latter is true when the number of sampled cells and sequencing depth are adequately large.

Acknowledgements. Drs. Qian Qi and Jorg Goronzy provided the data used in the example section. Bonnie Chung supplied technical help. We are grateful to two anonymous referees, Associated Editor, and Editor for their insightful comments and constructive suggestions, which greatly improved the manuscript.

REFERENCES

ALDRICH, R. J. (2010). *GCHQ*: *The Uncensored Story of Britain's Most Secret Intelligence Agency*. Harper Collins, London.

BOYD, S. D., MARSHALL, E. L., MERKER, J. D., MANIAR, J. M., ZHANG, L. N., SAHAF, B., JONES, C. D., SIMEN, B. B., HANCZARUK, B., NGUYEN, K. D., NADEAU, K. C., EGHOLM, M., MIKLOS, D. B., ZEHNDER, J. L. and FIRE, A. Z. (2009). Measurement and clinical monitoring of human lymphocyte clonality by massively parallel V-D-J pyrosequencing. *Sci*. *Transl*. *Med*. **1** 12a23.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. Wadsworth Advanced Books and Software, Belmont, CA. MR0726392

CHAO, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* **43** 783–791. MR0920467

CHAO, A. (1989). Estimating population size for sparse data in capture-recapture experiments. *Biometrics* **45** 427–438. MR1010510

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

FAN, J., LIAO, Y. and LIU, H. (2016). An overview of the estimation of large covariance and precision matrices. *Econom*. *J*. **19** C1–C32. MR3501529

FULLER, W. A. (1987). *Measurement Error Models*. Wiley, New York. MR0898653

GOOD, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* **40** 237–264. MR0061330

KAPLINSKY, J. and ARANOUT, R. (2016). Robust estimates of overall immune-repertoire diversity from high throughput measurements on samples. *Nat*. *Commun*. **7** 11881. DOI:10.1038/ncomms11881.

LAYDON, D. J., BANGHAM, C. R. M. and ASQUITH, B. (2015). Estimating T-cell repertoire diversity: Limitations of classical estimators and a new approach. *Philos*. *Trans*. *R*. *Soc*. *B* **370** 20140291. DOI:10.1098/rstb.2014.0291.

MCKANE, A. G., ALONSO, D. and SOLÉ, R. V. (2004). Analytic solution of Hubbell's model of local community dynamics. *Theor*. *Popul*. *Biol*. **65** 67–73.

MILLER, R. G. (1974). The jackknife—a review. *Biometrika* **61** 1–15. MR0391366

PARAMESWARAN, P., LIU, Y., ROSKIN, K. M., JACKSON, K. K., DIXIT, V. F., LEE, J. Y., ARTILES, K. S., ZOMPI, S., VARGAS, M. J., et al. (2013). Convergent antibody signatures in human dengue. *Cell Host Microbe* **13** 691–700.

QI, Q., LIU, Y., CHENG, Y., GLANVILLE, J., ZHANG, D., LEE, J.-Y., OLSHEN, R. A., WEYAND, C. M., BOYD, S. and GORONZY, J. J. (2014). Diversity and clonal selection in human T cell repertoire. *Proc*. *Natl*. *Acad*. *Sci*. *USA* **111** 13139–13144.

ROBBINS, H. E. (1968). Estimating the total probability of the unobserved outcomes of an experiment. *Ann*. *Math*. *Stat*. **39** 256–257. MR0221695

SCHATZ, D. G. and JI, Y. (2011). Recombination centres and the orchestration of V (D) J recombination. *Nat*. *Rev*., *Immunol*. **11** 251–263.

TIAN, L., GREENBERG, S. A., KONG, S. S., ALTSCHULER, J., KOHANE, I. S. and PARK, P. J. (2005). Discovering statistically significant pathways in expression profiling studies. *Proc*. *Natl*. *Acad*. *Sci*. **102** 13544–9.

L. TIAN
R. A. OLSHEN
DEPARTMENT OF BIOMEDICAL DATA SCIENCE
STANFORD UNIVERSITY SCHOOL OF MEDICINE
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305
USA
E-MAIL: lutian@stanford.edu
olshen@stanford.edu

Y. LIU
CALICO LIFE SCIENCES LLC
SOUTH SAN FRANCISCO, CALIFORNIA 94080
USA
E-MAIL: liu.yi.pei@gmail.com

A. Z. FIRE
DEPARTMENT OF GENETICS
DEPARTMENT OF PATHOLOGY
STANFORD UNIVERSITY SCHOOL OF MEDICINE
STANFORD UNIVERISTY
STANFORD, CALIFORNIA 94305
USA
E-MAIL: afire@stanford.edu

S. D. BOYD
DEPARTMENT OF PATHOLOGY
STANFORD UNIVERSITY SCHOOL OF MEDICINE
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305
USA
E-MAIL: sboyd1@stanford.edu