ELSEVIER
Journal of Immunological Methods xx (2006) xxx–xxx
JIM Journal of Immunological Methods
www.elsevier.com/locate/jim

Research paper

The automated counting of spots for the ELISpot assay

Natalie Hawkins *, Steve Self, Jon Wakefield

Fred Hutchinson Cancer Research Center, Statistical Center for HIV Aids Research and Prevention, 1100 Fairview Avenue, LE-400, Seattle, WA 98109, United States

Received 15 March 2006; received in revised form 23 June 2006; accepted 10 August 2006

Abstract

An automated method for counting spot-forming units in the ELISpot assay is described that uses a statistical model fit to training data that is based on counts from one or more experts. The method adapts to variable background intensities and provides considerable flexibility with respect to what image features can be used to model expert counts. Point estimates of spot counts are produced together with intervals that reflect the degree of uncertainty in the count. Finally, the approach is completely transparent and "open source" in contrast to methods embedded in current commercial software. An illustrative application to data from a study of the reactivity of T-cells from healthy human subjects to a pool of immunodominant peptides from CMV, EBV and flu is presented.
© 2006 Published by Elsevier B.V.

Keywords: Automated spot counting; ELISpot assay; Image analysis; Generalized linear models

UNCORRECTED PROOF

1. Introduction

T-lymphocyte response to vaccination represents the primary immunogenicity endpoint in Phase I/II trials of current candidate HIV vaccines (Koup et al., 1994; Borrow et al., 1994; Rowland-Jones et al., 1995; Mazzoli et al., 1997; Musey et al., 1997; Ogg et al., 1998; Goh et al., 1999), and the use of a highly standardized, sensitive assay to measure these responses is a critical requirement in the development and evaluation of HIV vaccines. The ELISA-spot or ELISpot assay currently represents the primary method to detect T-cell responses to HIV vaccines in the HIV Vaccine Trials Network. Considerable effort has been made to standardize the reagents and laboratory procedures used in these assays. However methods for the counting of spot-forming units (SFUs), which is used to obtain the final quantitative result of the ELISpot assay, have received somewhat less attention.

Historically, SFUs have been hand-counted by laboratory technicians but such subjective readings introduce significant variability in the assay outcome and are time-consuming. Computer algorithms for the analysis of images of the wells have been employed to automate the process of spot counting (Hudgens et al., 2004). Although automated spot counting algorithms can provide highly standardized assay outcomes, there are challenges to this approach that call into question the ultimate accuracy of these methods. Specifically, there is no "gold standard" for defining an SFU that can explicitly

Abbreviations: CMV, cytomegalovirus; EBV, Epstein–Barr virus; ELISpot, enzyme-linked immunospot; HIV, human immunodeficiency virus; T-cell, T-lymphocyte; SFU, spot-forming unit; TIFF, Tagged Image File Format.

* Corresponding author. Tel.: +1 206 667 7753; fax: +1 206 667 4812.
E-mail address: hawkins@scharp.org (N. Hawkins).

0022-1759/$ - see front matter © 2006 Published by Elsevier B.V.
doi:10.1016/j.jim.2006.08.005

ARTICLE IN PRESS

be used in algorithm design. In addition, such algorithms must integrate an automated method for calibration to background intensity levels that vary from plate to plate and distinguish "true SFUs" from various artifacts that include variable background intensity within wells (e.g., edge effects) and contamination. Examples of images from ELISpot assays that illustrate some aspects of this variability are given in Fig. 1. Numbering from left to right and top to bottom, wells 1, 4 and 5 contain clear artifacts, while there are dark patches close to the edges of a number of wells.

In this work, we propose an automated approach to the analysis of images from ELISpot assays that provides accurate and highly standardized counts of SFUs. In the absence of a gold standard for defining an SFU, we define the conceptual criterion of success for the method as a standardized implementation of the implicit rules for use by a designated expert (or possibly a panel of such experts) in counting SFUs. Specifically, the method uses "training data", composed of SFU counts by an expert, in order to refine the algorithm to produce counts that are accurate reflections of the expert counts but, unlike counts by any human, are uniformly applied from assay to assay. The model-based approach we describe allows the uncertainty in the count to be acknowledged, so that an interval estimate for the

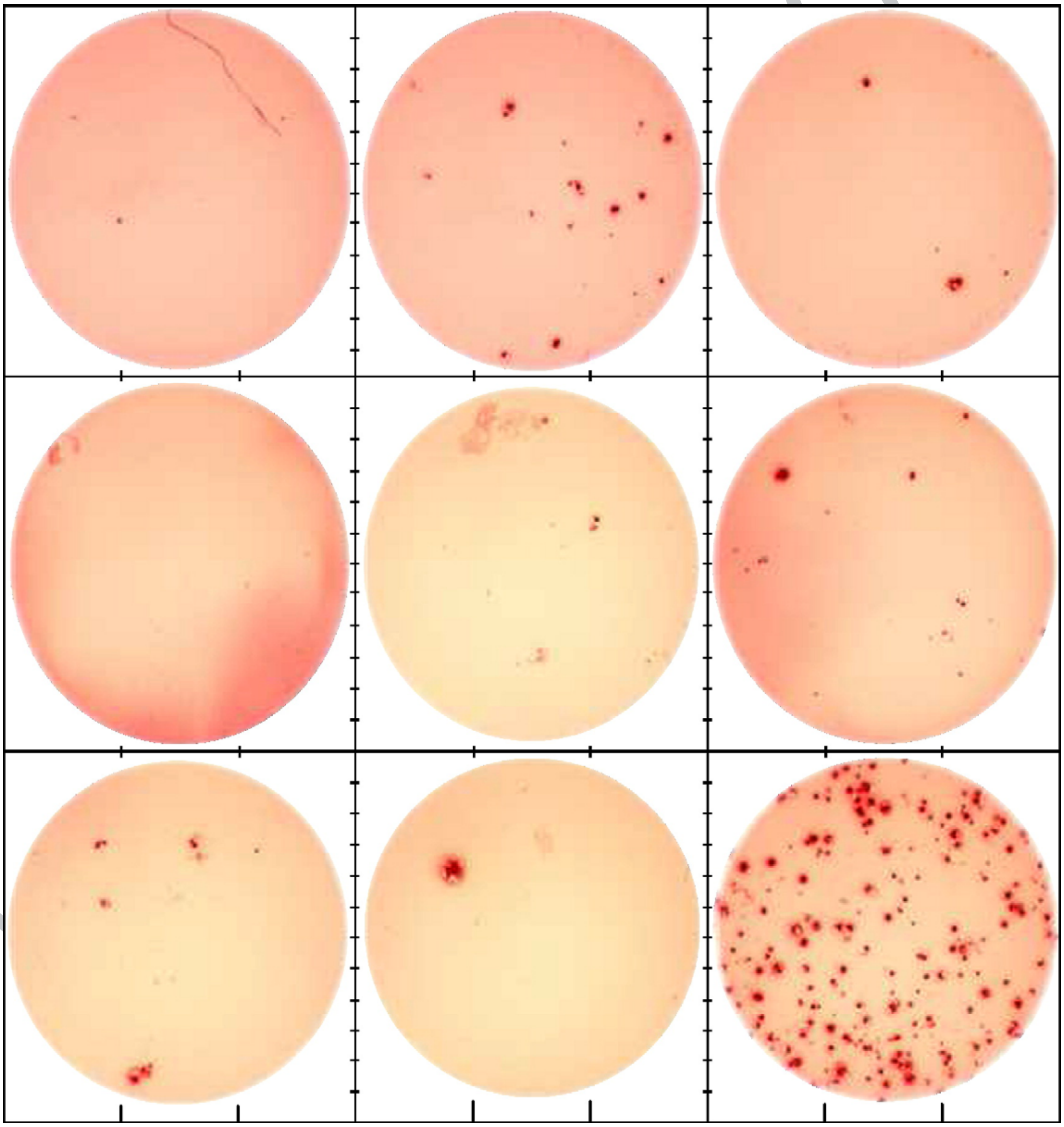

Fig. 1. Nine typical wells, showing spot forming units and various artifacts.

Please cite this article as: Natalie Hawkins et al., The automated counting of spots for the ELISpot assay, Journal of Immunological Methods (2006), doi:10.1016/j.jim.2006.08.005.

number of spots per well is produced. The method is illustrated using data from a study of the reactivity of T-cells from healthy human subjects to a pool of immunodominant peptides from CMV, EBV and flu.

2. Methods

In this section we describe the method of assigning a spot count to each well, along with an associated interval estimate. The method has two components. First, we pre-process the image using a thresholding and grouping technique to identify interesting areas which we call "globs". Second, based on training data, we formulate a model to predict the number of spots in each glob, based on glob characteristics such as the size of the glob. The resulting model is used to predict the number of spots in a new well, along with an interval estimate.

2.1. Pre-processing

For each well, the raw data originate from a Tagged Image File Format (TIFF) file and consist of pixel-level red, green and blue intensities, displayed in Fig. 1. For processing we use grey scale values by computing a mean of the red, green, and blue values to get an intensity at each pixel. These values range from 0 to 255 and are such that high intensities correspond to background, while low intensities correspond to spots, and to anomalies of the measurement process, such as an errant hair in the well.

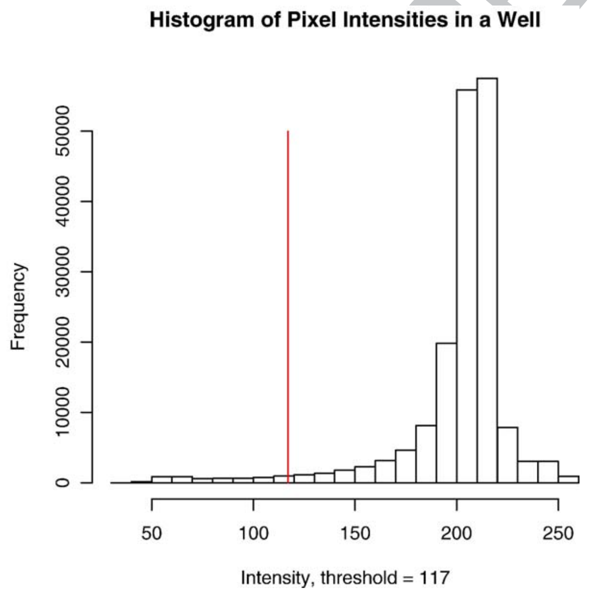

Fig. 2. Histogram of intensities from the ninth well in Fig. 1. The vertical line corresponds to the "threshold".

We use a thresholding technique, followed by a set of grouping rules based on contiguity, to identify interesting areas in the well which we call globs. We start with globs rather than with SFUs, or spots, because the thresholding technique easily identifies globs, but not confluent spots within globs. A glob can contain zero or one or more spots. We use a statistical model, described later, to determine the number of spots within each glob.

We are first required to choose a thresholding value to apply to a well to identify pixels belonging to globs. Through empirical experimentation we chose, for each well, the threshold to be the mean intensity of all pixels in the well minus three standard deviations, the latter calculated over all pixels in the well. Fig. 2 illustrates, with the histogram of intensities for the ninth well in Fig. 1 and the associated threshold.

Globs are identified in the well by first comparing each well pixel to the threshold. If the pixel intensity is below the threshold, the pixel is called a glob pixel, and globs are formed from glob pixels based on contiguity of those pixels. For one pixel globs, none of the possible eight pixels surrounding the one glob pixel is a glob pixel. For multiple-pixel globs, each pixel in the glob must be touching another glob pixel in, at least, one of the possible eight positions surrounding the pixel. Once the globs have been identified, we drop small, light globs since, in discussion with the lab technicians, these do not correspond to real spots. "Small" means less than 10 pixels and "light" corresponds to average intensity greater than 95% of the threshold value used to make the glob/not-glob pixel assignment (recall that high intensity values mean that the spot is light, not dark). As an example, the left-hand panel of Fig. 3 reproduces the ninth well in Fig. 1, with the right-hand panel showing the globs that have been identified using the thresholding technique.

Next we formulate a statistical model, based on training data, which can be used to predict the number of spots within each glob and, as a result, the number of spots in a new well, along with a confidence interval.

2.2. Training data

We use a set of training data to build a predictive statistical model, based on glob characteristics, which can be used to predict the number of SFUs, or spots, in a well, along with an interval estimate. The statistical model requires, as input, data from globs identified in the well.

The training data consist of glob data from 50 wells, selected from three plates. For each glob we obtained an "expert" count of the number spots within the glob. The

Please cite this article as: Natalie Hawkins et al., The automated counting of spots for the ELISpot assay, Journal of Immunological Methods (2006), doi:10.1016/j.jim.2006.08.005.

ARTICLE IN PRESS

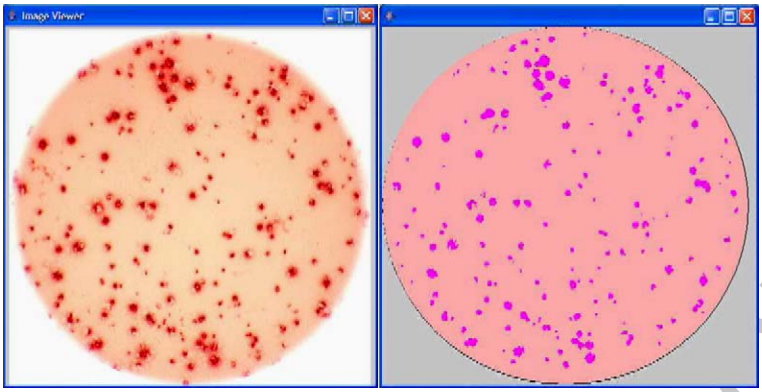

Fig. 3. The image on the right shows the globs identified in the well on the left using the thresholding technique. This well is the ninth well in Fig. 1.

"expert" count of the number of spots within each glob was provided by a senior immunologist. We provided the expert with an Excel spreadsheet which contained one page per well. On each page we displayed the original TIFF image of the well, along with numbered, computer-generated arrows super-imposed on the image pointing to globs, which we had identified using the thresholding and grouping technique described above. In areas of high congestion, outlines were drawn to separate globs. To the right of the image, a data entry area was provided with a column displaying the glob numbers and an empty column for the number of spots judged to be within each glob. The expert examined each image, and entered the number of spots for each glob.

Discussions with the expert revealed a set of rules that were used when counting spots. True spots are dark in the center and slightly fuzzy on the edges. False spots are either: (1) very faint and/or very small, (2) clustered at the edges of the well, (3) aligned in a hair-like pattern (indicates a cracked well), or (4) look like debris (very dark and often not circular). The characteristics of the globs that we chose to investigate were based on these rules, and on our empirical observations of what glob characteristics were important predictors of the number of spots in each glob.

The nine glob characteristics were: (1) glob size, (2) median intensity within glob, (3) ratio of maximum glob intensity to minimum glob intensity, (4) variance of glob intensity, (5) ratio of variance of glob intensity to mean glob intensity, (6) median distance of the glob from the center of the well, (7) whether or not the glob is located near the edge of the well, which is defined as whether or not the median distance of the glob from the center of the well is greater than 75% of the longest radius in the well (the well is almost, but not quite a perfect circle), (8) the percent of the pixels in the box which bounds the glob which are glob pixels, (9) the square of the log of the ratio of the dimensions (height and width) of the box which bounds the glob.

2.3. Statistical modeling

Based on training data, we aim to form a model, which entails selecting glob characteristics on the basis of their ability to predict the number of spots in each glob. We build all possible models having from just one to all nine of the glob characteristics as covariates ($2^9-1=511$ models), as well as all possible combinations involving interaction terms with the discrete glob characteristic edge (an additional 6305 models). A cross-validating procedure, described later, is used to select the best model from the complete set of 6816 possible models. The best model can then be used to predict the number of spots in any future wells, based on the glob characteristics of those wells.

We select a set of n training wells, pre-processed as described in Section 2.1, containing a set of globs with glob characteristics X_{ij} for glob j within training well i; accompanying each well and glob is a number of spots, Y_{ij}, $i=1, \ldots, n$, $j=1, \ldots, g_i$, as counted by the lab technician.

Since the outcome is discrete, a natural starting point for analysis is a Poisson model with mean number of counts $E[Y_{ij}|X_{ij}]$. Unfortunately such a model is deficient in the sense that the Poisson assumption constrains the variance to equal the mean. As described in McCullagh and Nelder (1989), a more flexible working model assumes that $\mathrm{var}(Y_{ij}|X_{ij})=\kappa\times E[Y_{ij}|X_{ij}]$, so that κ allows the variance to deviate from that under a Poisson model. We also assume that the mean takes the log-linear form

$$\log E[Y_{ij}|X_{ij}] = X_{ij}\beta,$$

UNCORRECTED PROOF

Please cite this article as: Natalie Hawkins et al., The automated counting of spots for the ELISpot assay, Journal of Immunological Methods (2006), doi:10.1016/j.jim.2006.08.005.

Table 1
Summary of parameter estimates from best-fitting model

Characteristic	Estimate	Standard error	*p*-value
Located near edge	1.20	0.859	0.164
Height–width ratio	−0.0901	0.3186	0.7775
Median intensity in glob	−0.0325	0.00362	2.0×10^{-16}
Variance of intensities in glob	0.00447	0.000427	2.0×10^{-16}
Ratio of variance to mean intensities in glob	−0.606	0.0608	2.0×10^{-16}
Glob size	0.000105	0.000313	0.737
Median distance of glob from the center of the well	−0.000308	0.000955	0.747
Ratio of max to min intensity in glob	0.279	0.0867	0.00135
Edge × height–width ratio	−1.74	0.634	0.00620
Edge × size	0.000418	0.000532	0.433
Edge × median distance from center of well	−0.00560	0.00420	0.183

though our method could use any form. For example, the method we describe could be applied to any parametric or semi-parametric model including logic regression, generalized additive models, or splines, see Hastie, Tibshirani, and Friedman (2000) for more detail on these methods. A quasi-likelihood method of inference, as described in McCullagh and Nelder (1989), is used to estimate the parameters of the model; this method has the advantage of requiring the specification of the first two moments of the data, without making a distributional assumption. The method we describe can also be used with specific distributional assumptions, if these appear reasonable in any particular application. We also use sandwich estimation (Royall, 1986) to provide empirical estimates of the standard errors. This approach provides a consistent estimator of the standard errors, given independent glob counts.

The over-dispersion parameter, along with sandwich estimation, is designed to account for components of variation that are attributed to well and/or plate. Although there are methods for improving prediction error of counts for one well using data from other wells on the same plate, in our experience working with laboratory scientists, they prefer to make prediction for each well independently. We wish to have a general method and not one which needs retuning in each different scenario.

Once we have selected the best predictive model of the type described above, based on the training data, the model can be used to predict the number of spots in a new well. Let X_j denote the glob characteristics of a new well containing $j=1, \ldots, n_{\text{new}}$, globs, for which we require an estimate of the number of spots, call this $\hat{\theta}$. Once estimates $\hat{\beta}$ and $\hat{\kappa}$ are obtained, a prediction is available via $\hat{\theta} = \sum_{j=1}^{n_{\text{new}}} \exp(X_j\hat{\beta})$, which is an unbiased estimate.

Using the delta method to obtain the variance of $\hat{\theta}$, we obtain an approximate 95% interval for the total number of spots that is given by:

$$\sum_{j=1}^{n_{\text{new}}} \exp(X_j\hat{\beta}) \pm 1.96 \times \left[\left\{\sum_{j=1}^{n_{\text{new}}} \exp(X_j\hat{\beta})X_j\right\}\hat{V}\left\{\sum_{j=1}^{n_{\text{new}}} X_j^T \exp(X_j\hat{\beta})\right\}\right]^{1/2}$$

where $\hat{V}$ is the sandwich estimate of the variance of $\hat{\beta}$.

3. Results

We wish to use the training data to decide on which of the 9 glob characteristics are important predictors of the number of spots that each glob contains, in order to find the model which would best serve as a predictive model. Specifically we have a total of $K=6816$ models, this set consisting of all possible models containing or not-containing each of the 9 glob characteristics, as well as all possible interaction models containing an interaction with the discrete glob characteristic, edge. We use a cross-validation technique, in which we use 49 of the training wells to estimate the parameters of model, M_k, $k=1, \ldots, K$, and then predict the number of spots in

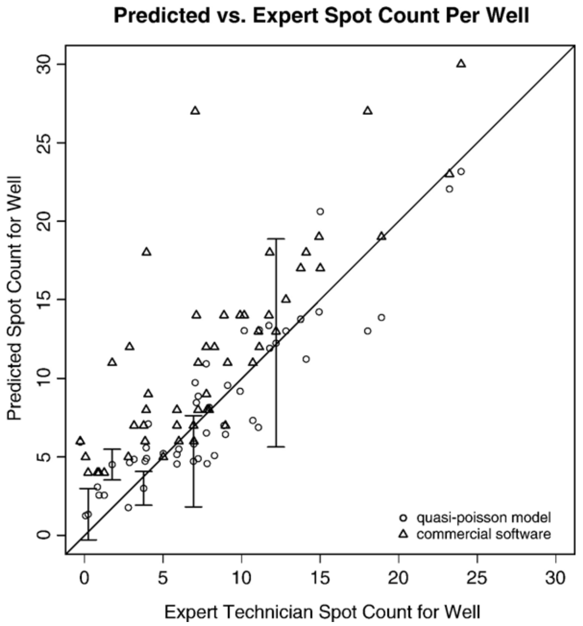

Fig. 4. Number of spots as predicted by the model-based approach and the current automated lab method, for 50 wells.

UNCORRECTED PROOF

Please cite this article as: Natalie Hawkins et al., The automated counting of spots for the ELISpot assay, Journal of Immunological Methods (2006), doi:10.1016/j.jim.2006.08.005.

ARTICLE IN PRESS

the 50th well; repeating this procedure and leaving out a different well each time, gives a set of predictions $\hat{Y}_{ij}^{k}$ under model k, so that we can calculate the model assessment sum of squares criteria

$$SS_k = \sum_{i=1}^{n} \sum_{j=1}^{g_i} (Y_{ij} - \hat{Y}_{ij}^{k})^2,$$

$k=1, \ldots, K$. After training the model with data from globs from 50 wells, we found the best model, based on the minimum SS_k.

The best model was found to contain eight glob characteristics and three interaction terms with the glob characteristic edge: (1) edge, (2) height–width ratio, defined as the square of the log of the ratio of the dimensions (height and width) of the box which bounds the glob, (3) median intensity, (4) variance of the intensity, (5) variance of the intensity divided by the mean intensity, (6) size, (7) median distance from the center of the well, (8) the ratio of the maximum intensity to the minimum intensity; and interactions of edge with: (1) height–width ratio, (2) size, and (3) median distance from the center of the well. Once we have decided upon this model we re-estimate the coefficients based on all 50 wells. Table 1 contains the resulting estimates, along with their standard errors.

From the coefficients we see that globs classified as near the edge are more likely to contain more spots. The more rectangular the glob is, as measured by the height–width ratio, the less likely it is to contain more spots. Darker globs (as measured by lower median intensity) are more likely to contain more spots, while more constant intensity within a glob implies fewer spots. As the ratio of the variance of the intensity to the mean intensity increases the number of spots decreases. Globs containing more pixels are more likely to contain more spots. Globs that are located further from the center of the well are more likely to contain fewer spots (reflecting the anomalies that occur towards the outside of the well, see Fig. 1, wells 4 and 6 in particular). Finally, greater maximum to minimum intensities suggest more spots also. Looking at the interaction terms we see that globs near the edge and more rectangular (as measured by the height–width ratio) are likely to contain fewer spots. Larger globs near the edge are more likely to contain more spots, and globs classified as near the edge but which are closer to the edge are likely to contain fewer spots. The non-significance of four of the variables and two of the interaction terms, is perhaps surprising but it is the combination of variables that is important from a prediction point of view.

Fig. 4 shows the estimated number of spots in each of the 50 wells from our method, versus those from the laboratory expert. Also shown are the estimates from the automated method currently used by the lab. For clarity, for a small collection of wells we include our confidence interval, based on the sandwich estimator of the variance. For plotting, we have jittered the values on the x-axis slightly to uncover points which might be overlapping so that all 100 points are visible on the plot. We see that the model predictions are more accurate relative to the expert technician, than is the commercial software being used by the lab. As confirmation of this we can evaluate the average bias, given by $1/n \sum_{i=1}^{n} (Y_i - \hat{Y}_i)$, and the mean squared error (MSE), given by $1/n \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$, where Y_i and Y_i are the observed and predicted number of spots in well i, for each of the model-based and current automated lab methods. For the model-based approach we obtain an average bias and MSE of 0.0336 and 5.68, while for the current automated lab method we obtained average bias and MSE of 3.49 and 26.4. Hence we see the model-based approach provides more accurate predicted numbers of spots, as measured by both bias and precision; in particular the commercial software provides an overcount of the number of spots.

4. Discussion

There is no "gold standard" method of spot counting to which automated methods can be compared. In the absence of such a standard, expert opinion with all of its associated vagaries, represents the standard by which automated methods must be judged. However expert opinion must first be operationally defined. We have operationally defined expert opinion in this work as the counts made on our training data set by a senior immunologist with whom we have collaborated. This has served our purpose of providing a realistic and pertinent illustration of a specific application of our proposed method. A broader definition based on a panel of immunologists might also have been used. We leave to future work the development of a more extensive set of training data together with an associated consensus expert opinion of spot counts that might provide a more definitive and broadly applicable counting algorithm based on our methods.

The accuracy of an automated counting method refers to how faithfully the method replicates the counts from expert opinion on average (over globs). Our proposed method is trained directly from expert opinion using statistical methods that guarantee (in large samples) such accuracy. We expect that this will provide a more

UNCORRECTED PROOF

Please cite this article as: Natalie Hawkins et al., The automated counting of spots for the ELISpot assay, Journal of Immunological Methods (2006), doi:10.1016/j.jim.2006.08.005.

accurate reproduction of counts based on expert opinion than other methods that are indirectly “calibrated”.

Assessing the precision of automated methods is challenging because there is innate non-systematic variability in expert opinion. This variability is reflected in the fact that expert recounts do not always result in exactly the same number of spots per well. This component of random variation will be inherited by any automated method. The proposed counting method is based on measurable characteristics of globs and, to the extent that these characteristics capture all factors considered systematically by experts in their counts, the automated methods will faithfully replicate the expert opinion up to the aforementioned random variability. We expect that a certain amount of systematic variation in expert counts will not be captured by readily measurable glob characteristics so that automated methods will inevitably be somewhat more variable than the theoretical minimum variation defined by recount variability. However, the proposed method is completely flexible with respect to the set of measurable glob characteristics that can be considered as possible predictors with practical limits on this set imposed only by the size of the training data set. Thus, with an extensive training data set and careful elicitation of the glob characteristics and other factors considered by experts in performing their counts, it is reasonable to expect that the proposed method will reproduce the systematic variation in expert counts.

One advantage of the proposed method is that interval estimates of spot counts are naturally produced that reflect the degree of uncertainty in the count. This interval estimate can be used as a component of the assay quality control process to reflect reliability of counts delivered for each well. The estimated variability in spot count at the well level can also form the basis for a similar estimate of variability for summary measures of response that combine spot counts over multiple wells (e.g. total response across peptide-treated wells net of response in negative control wells).

Finally, the proposed method provides a completely transparent “open-source” approach for spot counting that is in contrast to proprietary methods embedded in commercial software that often function as a black-box. In the current atmosphere that places considerable value on standardization of reagents and operating procedures for immunologic assays used in the development and evaluation of HIV vaccines (Klausner et al., 2003), the proposed method represents a natural approach to extending this standardization to the final critical step of the assay process.

References

Borrow, P., Lewicki, H., Hahn, B.H., Shaw, G.M., Oldstone, M.B., 1994. Virus-specific CD8-cytotoxic T-lymphocyte activity associated with control of viremia in primary human immunodeficiency virus type 1 infection. Journal of Virology 68, 6103.

Goh, W.C., Markee, J., Akridge, R.E., et al., 1999. Protection against human immunodeficiency virus type 1 infection in persons with repeated exposure: evidence for T cell immunity in the absence of inherited CCR5 coreceptor defects. Journal of Infectious Diseases 179, 548.

Hastie, T., Tibshirani, R., Friedman, J., 2000. The Elements of Statistical Learning, Data Mining, Inference and Prediction. Springer, Verlag.

Hudgens, M., Self, S., Chiu, Ya-Lin, et al., 2004. Statistical considerations for the design and analysis of the ELISpot assay in HIV-1 vaccine trials. Journal of Immunological Methods 288, 19.

Klausner, R.D., Fauci, A.S., Corey, L., 2003. The need for a global HIV vaccine enterprise. Science 300, 2036.

Koup, R.A., Safrit, J.T., Cao, Y., Andrews, C.A., McLeod, G., Borkowsky, W., Farthing, C., Ho, D.D., 1994. Temporal association of cellular immune responses with the initial control of viremia in primary human immunodeficiency virus type 1 syndrome. Journal of Virology 68, 4650.

Mazzoli, S., Trabattoni, D., Lo Caputo, S., et al., 1997. HIV-specific mucosal and cellular immunity in HIV-seronegative partners of HIV-seropositive individuals. Nature Medicine 3, 1250.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, Second edition. Chapman and Hall.

Musey, L., Hughes, J., Schacker, T., Shea, T., Corey, L., McElrath, M.J., 1997. Cytotoxic-T-cell responses, viral load, and disease progression in early human immunodeficiency virus type 1 infection. New England Journal of Medicine 337, 1267.

Ogg, G.S., Jin, X., Bonhoeffer, S., Dunbar, P.R., Nowak, M.A., Monard, S., Segal, J.P., Cao, Y., Rowland-Jones, S.L., Cerundolo, V., Hurley, A., Markowitz, M., Ho, D.D., Nixon, D.F., McMichael, A.J., 1998. Quantitation of HIV-1-specific cytotoxic T lymphocytes and plasma load of viral RNA. Science 279, 2103.

Rowland-Jones, S., Suttone, J., Ariyoshi, K., et al., 1995. HIV-specific cytotoxic T-cells in HIV-exposed but uninfected Gambian women. Nature Medicine 1, 59.

Royall, R., 1986. Model robust confidence intervals using maximum likelihood estimators. International Statistical Review 54, 221.

UNCORRECTED PROOF