ELDA: Extreme limiting dilution analysis for comparing depleted and enriched populations in stem cell and other assays

Yifang Hu and Gordon K. Smyth*

Bioinformatics Division,
The Walter and Eliza Hall Institute of Medical Research,
1G Royal Parade, Parkville, Victoria 3052, Australia

15 June 2009

ABSTRACT
ELDA is a software application for limiting dilution analysis (LDA), with particular attention to the needs of stem cell assays. It is the first limiting dilution analysis software to provide meaningful confidence intervals for all LDA data sets, including those with 0% or 100% responses. Other features include a test of the adequacy of the single-hit hypothesis, tests for frequency differences between multiple data sets, and the ability to take advantage of cases where the number of cells in the sample is counted exactly. A webtool at http://bioinf.wehi.edu.au/software/elda/ provides an easy user interface.

Keywords: stem cell assays, single-hit hypothesis, generalized linear models, Clopper-Pearson, likelihood ratio tests, computer software

* Corresponding author: Tel: (+61 3) 9345 2326; Fax: (+61 3) 9347 0852; E-mail address: smyth@wehi.edu.au

1. INTRODUCTION

A limiting dilution assay is an experimental technique for quantifying the proportion of biologically active particles in a larger population (Finney, 1952; Fazekas de St. Groth, 1982; Taswell, 1987). It is a type of dose-response experiment in which each individual culture allows a negative or positive response. Replicates are conducted which vary in the number of active particles tested. The process of dilution of the dose is typically continued to extinction of the response, or close to it. The rate of positive and negative responses at each dose allows the frequency of biologically active particles to be inferred.

Limiting dilution assays have been actively used in a wide variety of biological and scientific contexts for more than a century, most notably for quantifying bacteria (Phelps, 1908), immunocompetent cells (Makinodan and Albridge, 1962) or stem cells (Breivik, 1971). In immunology, limiting dilution assays were popularized by the work of Lefkovits and Waldman (1979) as a systematic technique for the study of B-cells and T-cells and their interactions. In different application areas an individual assay can take on different forms. In stem cell or cancer research, an assay might actually consist of, for example, an in vivo transplantation or injection. In this article, we will use the term “culture” to refer to an individual assay, regardless of the application area.

We use the term limiting dilution analysis (LDA) to refer to the statistical analysis of data from limiting dilution assays. LDA typically assumes the Poisson single-hit model, which assumes that the number of biological active particles in each culture varies according to a Poisson distribution, and a single biologically active cell is sufficient for a positive response from a culture (Greenwood and Yule, 1917; Taswell, 1981). As a statistical technique, LDA applies equally to a range of experimental scenarios which produce dose-response data whether or not these are limiting dilution assays in the strict sense. From this wider point of view, the main requirements are that the cultures are independent and that the frequency of biologically active particles is constant.

The classical aim of LDA is to estimate the active cell frequency (Finney, 1952; Lefkovits and Waldman, 1979; Taswell, 1981). Fisher (1922) showed that the estimator with the best possible precision can be derived by the statistical method of maximum likelihood estimation (MLE). The same estimation strategy was outlined even earlier by McCrady (1915). An efficient computational algorithm for MLE was worked out by Mather (1949) and Finney (1951). The MLE computations for LDA became available in general purpose statistical software after they were shown to fall within the framework of generalized linear models (GLM) by Nelder and Wedderburn (1972) and McCullagh and Nelder (1989). Free open-source GLM software has been available through the R project (www.r-project.org) since the late 1990’s, although this software is designed for mathematicians and statisticians rather than biologists or immunologists. The GLM approach to LDA frequency estimation is also implemented in the Microsoft Windows software application L-Calc (Stem Cell Technologies, www.stemcell.com), and this version has proved highly popular (Omobolaji et al., 2008; Bowie et al., 2007; Chen et al., 2008; Eirew et al., 2008; Huynh et al., 2008; Janzen et al., 2006; Kent et al., 2008; Liang et al., 2007; Maillard

et al., 2008; Oostendorp et al., 2008; Sambandam et al., 2005; Schatton et al., 2008; Walkley et al., 2007).

MLE is not the only efficient estimation strategy for LDA. Taswell (1981) showed that minimum chisquare (MC) estimation has equal or even better accuracy MLE in certain situations, when the number of distinct doses is small but the number of replicates is large. Strijbosch et al (1987) argued that MLE could be further improved by incorporating a jackknife correction for bias. However the difference in performance between these methods is small. MLE remains our method of choice because it provides the most flexible and powerful framework for confidence intervals and hypothesis testing as well as estimation. Unfortunately, Lefkovits and Waldman (1979) recommended a more statistically naïve method for LDA based on least square regression (LS). Taswell (1981) showed LS to be an order-of-magnitude less accurate than either MLE or MC. While LS gives acceptable results when the number of replicate cultures is very large (Lefkovitz and Waldman recommend a minimum of 60 replicate cultures per dose), it proves dangerously unreliable in the common situation that the data is less plentiful (Taswell, 1981).

There are at least two other distinct scientific aims which LDA might have, apart from the classical aim of estimating the active cell frequency. A second common aim is to check the validity of the single-hit hypothesis. A third possible aim, which has so far received little attention, is to compare the active cell frequency between different cell populations. Understanding these aims has a profound influence on the experimental design.

In stem cell research, a very common aim, perhaps the key aim, is to isolate as pure a population of stem cells as possible. In pursuit of this aim, it is common to sort cells according to different markers, and test for stem cell enrichment in the sorted subpopulations. In this process, a precise estimate of stem cell frequency may not be required in populations which are clearly depleted for these cells. Indeed, when an effective stem cell marker is discovered, the sorting process leads naturally to subpopulations which contain no stem cells, and hence give no positive cultures at any dose in an dilution assay (Vaillant et al., 2008). In this situation, it is of interest to establish that the subpopulation is significantly depleted relative to the enriched population or, even better, to place an upper bound on the stem cell frequency which could reasonably be in the depleted population. Pursuing a precise estimate of the active cell frequency would be meaningless. The converse situation also arises. Quintana et al (2008) show that cancer stem cells are more comment than previously appreciated, and present many assays with 100% positive results. In this situation it is of interest to place a lower bound on the stem cell frequency. LDA methods have not so far covered these situations.

Having good statistical power to check the single-hit model requires that wide range of different dilutions are used, with a moderate to large number of replicate cultures and with a worthwhile number of both positive and negative results. Many lack of fit tests have been proposed (Stein, 1922; Moran, 1954ab; Armitage, 1959; Cox, 1962; Shortley and Wilkins, 1965; Gart and Weiss, 1967; Thomas, 1972; Lefkovitz and Waldman, 1979; Taswell, 1984; Bonnefoix and Sotto, 1994; Bonnefoix et al, 1996, Bonnefoix et al, 2001). Some of the tests are graphically motivated (Shortley and Wilkins, 1965; Gart and Weiss, 1967; Bonnefoix et al, 2001). Lefkovitz and Waldman

(1979) also emphasize the need to plot the data to check the assumptions. Two major types of deviation from the model can detected. Firstly, there is the multi-hit possibility, whereby the single-hit hypothesis might be false, and some sort of mechanism involving multiple cells might in fact contribute to a positive culture response. In this case, the proportion of positive assays is likely to increase more rapidly than expected as the cell dose is increased. Secondly, the single-hit model might be correct but the assays may not be homogeneous in terms of the active cell frequency. In this case, the proportion of positive cultures is likely to increase more slowly as the dose increases than the classic model would predict, although rapid increase is also possible if the heterogeneity is correlated with dose. These two possibilities correspond respectively to curves bending down and curves bending up in the plots of Lefkovitz and Waldman (1979). However these two possibilities have not always been clearly been distinguished in the literature. Cox (1962) and Thomas (1972) test a particular multi-hit model, although this test is relatively difficult to implement and interpret. Shortley and Wilkins (1965) and Gart and Weiss (1967) concentrate on heterogeneity whereas Bonnefoix et al (1996) concentrate on the single-hit hypothesis. However these are all regression based tests which are straightforward to implement and interpret, and have good properties in small samples. The Pearson goodness of fit tests proposed by Lefkovitz and Waldman (1979) and Taswell (1981) do not distinguish the two types of deviation. Pearson tests also have poor power (Bonnefoix and Sotto, 1994), and are unreliable when the number of replicate cultures is small (McCullagh, 1985).

In many immunological contexts, the only practical way to assess the single-hit hypothesis is by way of the statistical tests described above. However there are experimental situations for which it is worthwhile and practical to validate the assumption experimentally. Shackleton et al (2006), Quintana et al (2008), Leong et al (2008) and Vermeulen et al (2008) validate the single-hit hypothesis experimentally, by confirming a single input cell in each culture by microscope visualization, before the assay is conducted. The fact that any of the single-cell assays lead to a positive response is then proof that a single cell is sufficient. Where the single-hit hypothesis can be confirmed experimentally, as in these cases, the need to validate the hypothesis statistically in each and every assay is no longer compelling, although the need to check heterogeneity remains. If the single-hit model can be assumed, then the active cell frequency may be accurately estimated from a limited number of distinct dilutions, provided that a worthwhile number of positive and negative cultures are available from at least one dilution.

Counting the number of cells also has the consequence that the number of cells no longer follows a Poisson distribution, but rather is a fixed quantity. This means that the classical Poisson model of LDA does not apply.

This article describes a coherent approach to LDA which includes extreme data situations, multiple populations and non-Poisson situations. The approach is implemented in the ELDA (Extreme LDA) webtool for LDA. ELDA provides a convenient interface for users without any need to download or install software. ELDA implements the GLM approach to LDA, with a number of extensions to cover situations commonly seen in current stem cell and other medical research, but not covered by classical analysis. Hypothesis tests are provided, using standard GLM theory, to compare active cell frequencies between two or more cell populations.

Although these tests use standard GLM theory, they have not been fully available previously in specialist LDA software. In a novel extension, one-sided confidence intervals provided for the active cell frequency when 0% or 100% positive responses are observed at all doses. The tradition assumption that the number of cells follows a Poisson distribution is also varied to allow for the possibility that the number of cells in the culture is observed exactly. We show that the GLM framework still applies, with a minor modification, even when the total number of cells is not Poisson but is fixed. The graphical displays recommended by Lefkovitz and Waldman (1979) are included but with efficient estimation of the active cell frequency.

We give tests of heterogeneity and the single-hit hypothesis which are adapted from Gart and Weiss (1967) and Bonnefoix et al (1996) and which take advantage of the GLM framework. The GLM test has the best performance of the goodness of fit tests in small samples, and it also has to ability to distinguish heterogeneity of samples from multi-hit alternatives.

ELDA has already proved valuable for LDA in a wide variety of high-profile research areas (Diaz-Guerra et al., 2007, Hosen et al., 2007, Leong et al., 2008, Quintana et al., 2008, Shackleton et al 2006; Siwko et al., 2008, Vaillant et al., 2008, Vermeulen et al., 2008).

The ELDA webtool is described in Section 2, and Section 3 gives examples of usage. These two sections are written for readers wishing to use the webtool. Section 4 gives details of the statistical methodology for readers wanting the mathematical background. Section 5 finishes with discussion and conclusions.

2. THE ELDA WEBTOOL

ELDA is an online tool for limiting dilution analysis. Users simply cut and paste a table of data into the web page. There is no need to download software or to undertake any programming.
ELDA accepts an input data table of three or four columns, separated by any combination of commas, spaces or tabs (Table 1). Users can type the data directly into the webpage text field, or can simply cut and paste the whole table from any spreadsheet application. Each row of data gives results for a particular cell dose. The columns are:

1. Dose: number of cells in each culture
2. Tested: number of cultures tested
3. Response: number of positive cultures
4. Group (Optional): label for the population group to which cells belong

By default, ELDA computes a 95% confidence for the active cell frequency in each population group. The user can vary the confidence level by entering whatever level is desired. Additional output can be requested by clicking any of the following checkboxes:

1. Plot: is a graphical display desired?
2. Compare groups: should frequencies be compared between multiple groups?
3. Model adequacy: should the adequacy of the single-hit model be tested?
4. Observed or expected: is the actual number of cells observed directly?

If all checkboxes are ticked, ELDA outputs the following results:

1. Estimated confidence intervals for the frequency of active cells in each group (Table 2).
2. A plot of the log proportion of negative cultures vs the number of cells, with a trend line representing the estimated active cell frequency (Figure 1).
3. Tests for pair-wise differences in active cell frequency between groups.
4. An overall test for differences between the population groups. (This is analogous to a one-way anova test.)
5. Goodness of fit tests. These is conducted using data from all the groups at once.

3. EXAMPLES OF USAGE

3.1 Confidence intervals and tests

A key facility of ELDA is the ability to handle extreme data situations. Table 1 shows a small data example which illustrates some of the capabilities of the software. This gives data on the frequency of repopulating mammary cells from a tumorigenic mouse model (Vaillant et al., 2008). Here a positive assay is one which results in a visible mammary epithelial outgrowth. In this experiment, the wild-type cells did not produce any outgrowths, although this might be due to insufficient cell numbers. MMTV-Wnt-1 is the enriched population. The interest is to estimate the repopulating cell frequency in the MMTV-Wnt-1, and to establish rigorously that it is enriched relative to wild-type. Also of interest is to place an upper bound on the proportion of repopulating cells which could be in the wild-type.

Despite the absence of responses, ELDA is able to compute an upper bound of 1/701 for the frequency of repopulating wild-type cells, representing an exact one-sided 95% confidence interval. Meanwhile MMTV-Wnt-1 cells are estimated to have a $1/464$ repopulating cell frequency. This frequency difference is statistically significant ($\chi_1^2 = 7$, $P = 0.0083$), confirming that the tumorigenic mouse cells are enriched for repopulating cells relative to wild-type.

With older software, researchers were forced to add a notional, in reality false, positive result to the wild-type cells in Table 1, so as to be able to run a conventional GLM analysis. This is naturally not to be recommended, as it misrepresents the true results. In this case, adding one imaginary positive response to the wild-type data would increase the upper bound on the repopulating frequency to 1/284, a several-fold overestimate compared to the correct value of 1/701. It may conceivably be that the wild-type cells contain no repopulating cells, in which case even one positive response would be a qualitative misrepresentation of reality.

3.2 Least squares vs maximum likelihood

The popular least squares regression method proposed by Lefkovitz and Waldman (1979) is not adequate when the number of cultures is small. Suppose, for example, that five cultures were assayed at each of the cell doses 100, 200, 300, 400, 500 and 600, and suppose that the number of positive cultures was 2, 3, 2, 5, 5 and 5 respectively. The MLE estimate of the active cell frequency is 1/205 with 95% confidence interval 1/332 to 1/126. The LS estimate of the active cell frequency is

1/361, not much more than half the MLE estimate and well outside the confidence interval. Part of the reason for this discrepancy is because LS is unable to constructively use much of the data. LS ignores the three doses giving 100% successes. However these are valid observations which contribute good information to the active cell frequency. Removing them biases the estimator, underestimating the proportion of active cells. Another part of the reason is that LS fails to weight the remaining observations correctly. LS gives equal weight to all the remaining data values whereas they are actually highly heteroscedastic on the log-proportion scale. Taswell (1981) has noted that examples in the literature are common where the scientific results would have changed had a more efficient LDA method been used.

3.4 Experimentally verified cell numbers

In traditional applications of LDA, the dosage of cells is controlled by serial dilution, a process which affects the average number of cells rather than being deterministic in terms of controlling individual cells. A scientifically much stronger strategy, where it is available, is to confirm the single-hit hypothesis experimentally by visualizing or otherwise delivering individual cells to the culture assays. In this way, Shackleton et al (2006, Supplementary Table 6) visualized individual cells microscopically from a cell population enriched for mammary stem cells. The assays consisted of growing mammary outgrowths from the individual cells. Six outgrowths were produced from 102 transplants in three independent experiments (Table 3). Here there is an experimentally verified single cell in each culture, as opposed to a variable Poisson number of cells. It turns out that the effect of treating the cell dose as fixed rather than random has only a modest affect on the estimated stem cell frequency, which is 1/17 compared to 1/16.5 if the cell dose is treated as Poisson.

Quintana et al (2008) describes another example. Flow-cytometrically isolated human melanoma cells derived from xenografts from four patients were diluted into Terasaki microwells such that wells containing single cells could be identified by phase contrast microscopy (Quintana et al., 2008). Wells visually containing a single cell were mixed with Matrigel and injected into NOD/SCID $Il2rg^{-/-}$. For many patients, all cultures gave positive results, for example the results in Table 4 for patient 214. In this case, the lower bound for the frequency is active cells decreases to 1/218 instead of 1/211 when the fact that the number of cells is observed exactly is taken into account.

3.5 Testing the single-hit hypothesis

If the cell characteristic being examined in a limiting dilution analysis does not satisfy the single-hit hypothesis, then any results from the analysis are likely to be misleading. Therefore a test of this hypothesis is a valuable routine check on the validity of the data. The GLM approach to LDA provides an excellent framework for statistical goodness of fit tests of the single-hit model, as previously observed by Bonnefoix et al (1996). Bonnefoix et al (1996) proposed a goodness of fit test in the form of a t-test derived from the GLM model. We use a likelihood ratio test in place of a t-statistic approach, because of greater power and much improved performance in small samples (Fears et al, 1996). We applied our test to the simulated and real data sets in Bonnefoix & Sotto (1994). These datasets included simulations of a two-hit Poisson model, a helper two-target Poisson model, a suppressor two-target Poisson model, and

a real dataset for which the Pearson chi-square goodness of fit test is clearly rejected. Our test correctly rejects the single-hit hypothesis for all four data sets, a performance not achieved by any of the five statistical tests considered by Bonnefoix and Sotto (1994).

Table 5 gives another data example, on tumor-forming frequencies of cells extracted from mice, in a study of a p53 mouse model of breast cancer. The data and estimated active cell frequency is shown in Figure 1. For this data, the single-hit model is unambiguously rejected ($\chi_1^2 = 8.226$, $P = 0.004$). Closer examination of this data revealed that the problem was not with the single-hit hypothesis itself, but with the fact that the data was obtained by combining several independent experiments conducted at different times with markedly different tumor-forming frequencies between experiments. In this case, the test of the model alerted the researchers to the need to control and adjust for variation between different runs of the experiment. In other situations, rejection of the single-hit hypothesis could indicate the need for a more complex two-hit or multi-hit biological model. This example shows that testing the single-hit hypothesis is useful also as a check on the consistency of the experiment from a technical point of view.

By the way, our results show that it is not possible to test the Poisson assumption of the so-called single-hit Poisson model, because the same GLM model applies even when the cell dosage is not Poisson.

4. RATIONALE FOR THE APPROACH

4.1 Generalized linear models

In this section, we outline the statistical methodology behind the ELDA software. We begin by outlining the GLM approach to LDA. Alternative introductions to GLMs can be found in Collett (1991) and Bonnefoix et al (1996).

The fundamental property of limiting dilution assays is that each culture results in positive or negative result. Write p_i for the probability of a positive result given that the expected number of cells in the culture is d_i . If n_i independent cultures are conducted as dose d_i , then the number of positive results follows a binomial distribution.

Write λ for the proportion of active cells in the cell population, so that the expected number of active cells in the culture is λd_i . If the cells behave independently, i.e., if there are no community effects amongst the cells, and if the cell dose is controlled simply by dilution, then the actual number of cells in each culture will vary according to a Poisson distribution. A culture will give a negative result only if there are no active cells in the assay. The Poisson probability formula tells us that this occurs with probability

$$1 - p_i = \exp(-\lambda d_i)$$

This formula can be linearized by taking taking logarithms of both sides, as

$$\log(1 - p_i) = -\lambda d_i$$

or, taking logarithms again,

$$\log(-\log(1-p_i)) = \log\lambda + \log d_i$$

This last formula is the famous complementary log-log transformation from Mather (1949). The first log formula is used by Lefkovitz and Waldman (1979), but the second log-log formula is superior for statistical purposes, because the intercept term $\log\lambda$ turns out to follow a normal distribution much more closely than does λ itself.

Although the binomial distribution and the complementary log-log transformation seem far removed from ordinary linear regression, GLM theory tells us that we can treat the above model philosophically as if it was a simple linear regression, using much of our intuition from linear regression (McCullagh and Nelder, 1989). Using the GLM approach, we fit the simple regression model

$$g(p_i) = \alpha + \beta x_i \qquad (1)$$

to the observed proportion of positive responses. Here g is the complementary-log-log transformation, and $g(p_i)$ is the quantity that the regression is predicting. The predictor x_i is the log-dose, and the intercept α is the log active cell fraction, the quantity that we want to estimate. The regression slope β is equal to one. An important point is that, when the regression model is estimated, the slope is kept fixed at this value rather than being estimated de novo, so α is the only unknown quantity. (In GLM terminology, x_i is known as an offset).

Fitting the GLM yields the MLE estimate $\hat{\alpha}$ of α. A confidence interval is given by

$$\hat{\alpha} \pm z\ \mathrm{se}(\hat{\alpha})$$

where $\mathrm{se}(\hat{\alpha})$ is the standard error of the estimate and z is the critical value of the normal distribution, e.g., $z = 1.96$ for a 95% confidence interval. To get back to the active cell frequency we simply exponentiate and invert the estimate and the confidence interval, i.e., $1/\hat{\lambda} = \exp(-\hat{\alpha})$. The software outputs confidence intervals for $1/\lambda$, representing the number of cells required on average to obtain one responding cell.

4.2 When the number of cells is observed

In the previous subsection, we outlined the classic Poisson single-hit model, common to almost all treatments of LDA since 1917. We show now however that the Poisson assumption is not required. Suppose that the number of cells d_i is each culture is counted exactly, by some experimental means. In that case, d_i is a fixed quantity, and the use of Poisson probabilities is inappropriate. The probability that any particular cell is not an active cell is $1-\lambda$. The probability of a negative response for the culture is the probability that none of the d_i cells are active, i.e.,

$$1 - p_i = (1-\lambda)^{d_i}$$

This gives

$$\log(1-p_i) = d_i \log\lambda$$

and

$$\log(-\log(1-p_i)) = \log(-\log\lambda) + \log d_i$$

In other words

$$g(p_i) = \alpha + x_i$$

where α is now log-log the active cell proportion. This shows that the GLM approach to LDA remains valid even when the number of cells is a fixed quantity. The only change from the previous subsection is that the intercept is log-log the active cell proportion instead of the log-proportion. The active cell frequency is now related to the intercept by $\hat{\lambda} = 1 - \exp(-\exp\hat{\alpha})$. If λ is small, then this and the previous formula yield very similar values.

The question of observing or not observing the cell counts exactly is important from a theoretical point of view, because it shows that the basic statistics of LDA do not flow from the Poisson distribution, as it generally claimed. In practical terms, the issue has a substantial impact on the estimated active cell frequency when the active cell is very high. When the active cell proportion is a very small proportion of the total number of cells, as is usually the case, the impact on the estimated frequency is relatively small.

4.3 Multiple groups

The fitted models for different datasets are compared using likelihood ratio tests using the asymptotic chisquare approximation to the log-ratio (Collett, 1991). If there are multiple groups, we fit the model

$$g(p_{gi}) = \alpha_g + x_i$$

where α_g is the intercept for group g. We test pairwise differences between the groups, i.e., we test equality of each pair of α_g. When there are more than two groups, we also test the overall null hypothesis that all the α_g are equal against that alternative that there are at least some differences between the groups.

4.4 Testing goodness of fit

A number of authors have proposed that the goodness of fit of the single-hit Poisson model can be tested using Pearson goodness of fit tests (Lefkovitz and Waldman, 1979; Taswell, 1981). However these are general purpose tests which are not focused on any particular type of departure from the model. They therefore have relatively poor power to distinguish the specific departures of interest in LDA, for example to distinguish multi-hit from single-hit models (Bonnefoix and Sotto, 1994). Even more importantly, the Pearson goodness of fit test is a statistically valid test only when the number of replicate cultures is large (McCullagh, 1985). This makes the test untrustworthy in many typical situations in which LDA is applied.

More focused and effective tests can be constructed by examining the pattern with which the proportion of positive cultures increases as the dose number of cells increases. A simple and natural way to do this is to extend the simple regression model (1) to have a slope parameter

$$g(p_i) = \alpha + \beta x_i \qquad (2)$$

and then to test whether the slope is equal to one. A regression slope less than one implies that the proportion of positive cultures is less responsive to dose than it should be, and is suggestive of heterogeneity (Armitage, 1959; Shortley and Wilkins, 1965). Heterogeneity can take different forms. For example, different cultures might contain different proportions of active cells, especially if the experiment combines biological replicates or if it has been conducted of a period time using different batches of material. Alternatively, heterogeneity could indicate variation in host sensitivity to active particles. In the stem cell context, heterogeneity might mean that the stem cell frequency varies between cultures, or that the stem cells vary in potency between

cultures, i.e., in the probability with which an individual stem cell produces an observable result. A regression slope greater than one in (2) implies that the proportion of positive cultures is hyper responsive to dose, and is suggestive of multi-hit alternatives (Taswell, 1984).

Model (2) goes most naturally with a plot of log-log proportion versus log(dose) (Shortley and Wilkins, 1965), but it can be interpreted easily also in terms of the log-proportion plots suggested by Lefkovitz and Waldman (1979). Slope β greater than one corresponds to concave downward curves in Lefkovitz and Waldman (1979) whereas slope less than one corresponds to convex upward curves.

Bonnefoix et al (1996) pointed out the hypothesis that $\beta = 1$ in model (2) can be tested very easily and naturally in the framework of generalized linear models. Bonnefoix et al (1996) recommended a t-statistic to test $\beta = 1$ against the two-sided alternative that $\beta \neq 1$. In the context of MLE and asymptotic tests, t-statistics which divide a parameter estimate by its standard error are known as Wald tests. Because of the well known shortcomings of Wald tests in binomial GLMs (Fears et al, 1996), we use a likelihood ratio test instead (Cox and Hinkley, 1974).

A further refinement is possible. Gart and Weiss (1967) pointed out that model (2) is not fully efficient for testing heterogeneity. An improved test is obtained by regressing on dose instead of log(dose), i.e.,

$$\log(-\log(1-p_i)) = \alpha + x_i + \beta d_i \qquad (3)$$

and then testing $\beta = 0$ versus $\beta < 0$.

We adapt the tests of Gart and Weiss (1967) and Bonnefoix et al (1996) as follows. We conduct a one-sided test of $\beta = 1$ vs $\beta > 1$ in model (2) to test multi-hit alternatives, and a one-sided tests of $\beta = 0$ vs $\beta < 0$ in model (3) to test heterogeneity. The two tests are conducted using likelihood score statistics, which are locally most powerful for testing one-sided alternatives (Cox and Hinkley, 1974). Because we are conducting two tests instead of one, the p-values are adjusted for multiple testing using the method of Holm (1979).

4.5 One-sided confidence intervals

The ELDA function handles 0% or 100% positive responses as special cases. ELDA use a strategy for one-sided confidence intervals adapted from (Clopper and Pearson, 1934). The probability of observing entirely negative responses in every culture at all doses is

$$p_0 = \prod_{i=1}^{m} (1-p_i)^{n_i d_i}$$

where n_i is the number of cultures at dose d_i and m is the number of separate doses in the experiment. When no positive responses for any assay, an exact one-sided confidence interval is obtained by solving for λ such that

$$p_0 = \alpha$$

where α is the type I error rate required, e.g., $\alpha = 0.05$ for a 95% confidence interval. The confidence interval for the active cell proportion is $(0, \lambda_0)$ where λ_0 is the solution for λ. This interval communicates our confidence that the active less

proportion is less than λ_0 and could be as low as zero. The confidence interval for the active cell frequency is the inverse of this, from $1/\lambda_0$ to infinity. The infinite bound has the meaning that an infinite number of cells might need to be observed before an active cell is found.

The probability of observing entirely positive response is

$$p_1 = \prod_{i=1}^{m} \left\{1-(1-p_i)^{d_i}\right\}^{n_i}$$

When there are 100% positive responses, an exact one-sided confidence interval is obtained by solving for λ such that

$$p_1 = \alpha$$

The equation is solved using a globally convergent Newton iteration. The confidence interval for the active cell proportion is $(\lambda_0, 1)$ where λ_0 is the solution for λ. This interval communicates our confidence that the active less proportion is at least λ_0 and could be as high as 100%. The confidence interval for the active cell frequency is the inverse of this, from 1 to $1/\lambda_0$. The lower bound has the meaning that only one cell needs to be observed to find an active cell, i.e., all cells are active.

The same strategy applies whether the number of cells is Poisson or is observed exactly, the only difference being the mathematical form of the dependence of the p_i on λ.

4.6 Command-line software

As well as providing the freely available webtool, we also provide the underlying algorithms and computer code in the statmod (Statistical Modelling) package for the R programming environment (www.r-project.org). R is the world's most popular open-source statistical software (Vance, 2009). The command-line version is not intended for most readers of this journal, but may be invaluable for programmers or biostatisticians wishing to build on the capabilities of ELDA. Full documentation is provided online using the conventions and facilities of the R programming environment (http://cran.r-project.org/web/packages/statmod/). The webtool version of ELDA utilizes the R version via a Perl and http interface.

5. DISCUSSION AND CONCLUSION

Despite more than a century of methodological development for LDA, the best methods have not generally be available to immunologists because of lack of easily accessible software.

The ELDA webtool gives researchers access to optimal LDA statistical techniques without the need to install software or to undertake any programming. The aims are (i) to give confidence intervals for the active cell frequency, (ii) to compare the active cell frequency across multiple cell subpopulations and (iii) to check the single-hit model. ELDA will handle any valid limiting dilution assay data set, without the need to edit or remove data cases, such as those which arise when the proportion of positive cultures approaches 0 or 100%. All data is incorporated into the analysis using appropriate statistical methods.

A particular motivation has been to provide rigorous methods to compare cell subpopulations which are depleted or enriched for stem cells. This gives rise to several issues which have not been addressed in traditional LDA. In this context, the focus in on comparing populations and placing bounds on stem cell frequencies, rather than the traditional focus of LDA on estimating frequencies. To this end, one-sided confidence intervals are developed for stem cell frequency for subpopulations which produce 0% or 100% positive cultures in a limited number of trials. This allows researchers to place an upper bound the stem cell frequency which could feasibly remain in a depleted population, and to place an lower bound on the stem cell frequency in a highly enriched population.

Another need which arises in stem cell research is the need accommodate small numbers of replicates in a statistically consistent and defensible manner. To this end, ELDA gives emphasis to statistical methods which behave well when the number of replicates is small. Hence emphasis is given to exact methods, and likelihood ratio tests are chosen over t-test methods which rely on standard errors and which are known to behave poorly in small samples. The literature shows that LDA is frequently used when the number of replicate cultures is moderate to small, in fact this might be closer to the norm than an exception.

All treatments of LDA over past century have relied on the assumption of Poisson variation in the number of cells in a culture. However in a number of recent stem cell applications, the number of cells is observed explicitly. This means that the cells do not follow a Poisson distribution, and that appealing to Poisson probabilities to derive the single-hit statistical model are invalid. ELDA handles this situation explicitly. We show that the generalized linear model approach to ELDA can still be used in this situation, although the relationship of the model coefficients to the active cell frequency is changed.

ELDA implements a test of the single-hit hypothesis similar to that of Bonnefoix et al. (1996). However a likelihood ratio test is used in place of a t-statistic approach, because of greater power and much improved performance in small samples.

In summary, ELDA is applicable in all common LDA situations. It provides a valuable resource for stem cell, cancer and immunological research.

ACKNOWLEDGMENT

Thanks to Mark Shackleton, Francois Vaillant, Jane Visvader and Geoff Lindeman for valuable discussions and feedback and for the use of unpublished data. Keith Satterley created the original web interface for ELDA.

REFERENCES

Armitage, P., 1959, An examination of some experimental cancer data in light of the one-hit theory of infectivity titrations. J. NATIONAL CANCER INSTITUTE 23, 1313-1330.

Bonnefoix, T. and Sotto, J.J., 1994, The standard chi2 test used in limiting dilution assays is insufficient for estimating the goodness-of-fit to the single-hit Poisson model. J. IMMUNOL. METHODS 167, 21-33.

Bonnefoix, T., Bonnefoix, P., Verdiela, P. and Sotto, J.J., 1996, Fitting limiting dilution experiments with generalized linear models results in a test of the single-hit Poisson assumption. J. IMMUNOL. METHODS 194, 113–119.

Bonnefoix, T., Bonnefoix, P., Callanan, M., Verdiel, P. and Sotto, J.J., 2001, Graphical representation of a generalized linear model-based statistical test estimating the fit of the single-hit Poisson model to limiting dilution assays. J. IMMUNOLOGY 167, 5725-5730.

Bowie, M.B., Kent, D.G., Dykstra, B., McKnight, K.D., McCaffrey, L., Hoodless, P.A. and Eaves, C.J., 2007, Identification of a new intrinsically timed developmental checkpoint that reprograms key hematopoietic stem cell properties. PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES 104, 5878.

Breivik, H., 1971, Haematopoietic stem cell content of murine bone marrow, spleen, and blood. Limiting dilution analysis of diffusion chamber cultures. J CELL PHYSIOL 78, 73-78.

Chen, W., Kumar, A.R., Hudson, W.A., Li, Q., Wu, B., Staggs, R.A., Lund, E.A., Sam, T.N. and Kersey, J.H., 2008, Malignant Transformation Initiated by Mll-AF9: Gene Dosage and Critical Target Cells. CANCER CELL 13, 432-440.

Cho, R.W., Wang, X., Diehn, M., Shedden, K., Chen, G.Y., Sherlock, G., Gurney, A., Lewicki, J. and Clarke, M.F., 2007, Isolation and molecular characterization of cancer stem cells in MMTV-Wnt-1 murine breast tumors. STEM CELLS 26, 364–371.

Clopper, C.J. and Pearson, E.S., 1934, The use of confidence or fiducial limits illustrated in the case of the binomial. BIOMETRIKA 26, 404–413.

Collett, D., 1991, MODELLING BINARY DATA. Chapman & Hall, London.

Cox, D.R., 1962, Further tests of separate families of hypotheses. J. ROY. STATIST. SOC. B24, 406-424.

Cox, D.R. and Hinkley, D.V., 1974, THEORETICAL STATISTICS. Chapman & Hall, London.

Diaz-Guerra, E., Vernal, R., del Prete, M.J., Silva, A. and Garcia-Sanz, J.A., 2007, CCL2 inhibits the apoptosis program induced by growth factor deprivation, rescuing functional T cells. J. IMMUNOL. 179, 7352-7357.

Eirew, P., Stingl, J., Raouf, A., Turashvili, G., Aparicio, S., Emerman, J.T. and Eaves, C.J., 2008, A method for quantifying normal human mammary epithelial stem cells with in vivo regenerative ability. NATURE MEDICINE 14, 1384-1389.

Fazekas de St. Groth, S., 1982, The evaluation of limiting dilution assays. J. IMMUNOL. METHODS 49, R11-R23.

Fears, TR, Benichou, J, Gail, MH (1996) A reminder of the fallibility of the Wald statistic. THE AMERICAN STATISTICIAN 50, 226-7.

Finney, D.J., 1951, The estimation of bacterial densities from dilution series. J. HYG. 49, 26-35.

Finney, D.J., 1952, STATISTICAL METHOD IN BIOLOGICAL ASSAY 1ST,2ND AND 3RD EDS., Charles Griffin, London.

Fisher, R.A., 1922, On the mathematical foundations of theoretical statistics. PHILOS. TRANS. R. SOC. LONDON SER. A. 222, 309-368.

Gart, J.J. and Weiss, G.H., 1967, Graphically oriented tests for host variability in dilution experiments. BIOMETRICS 23, 269-284.

Greenwood, M., and Yule, G. U., 1917, On the statistical interpretation of some bacteriological methods employed in water analysis. J. HYG. 16, 36-54.

Holm, S., 1979, A simple sequentially rejective multiple test procedure. SCANDINAVIAN JOURNAL OF STATISTICS 6, 65-70.

Hosen, N., Yamane, T., Muijtjens, M., Pham, K., Clarke, M.F. and Weissman, I.L., 2007, Bmi-1-green fluorescent protein-knock-in mice reveal the dynamic regulation of bmi-1 expression in normal and leukemic hematopoietic cells. STEM CELLS 25, 1635-1644.

Huynh, H., Iizuka, S., Kaba, M., Kirak, O., Zheng, J., Lodish, H.F. and Zhang, C.C., 2008, Insulin-Like Growth Factor-Binding Protein 2 Secreted by a Tumorigenic Cell Line Supports Ex Vivo Expansion of Mouse Hematopoietic Stem Cells. STEM CELLS 26, 1628.

Janzen, V., Forkert, R., Fleming, H.E., Saito, Y., Waring, M.T., Dombkowski, D.M., Cheng, T., DePinho, R.A., Sharpless, N.E. and Scadden, D.T., 2006, Stem-cell ageing modified by the cyclin-dependent kinase inhibitor p16 INK4a. NATURE 443, 421-426.

Kent, D.G., Dykstra, B.J., Cheyne, J., Ma, E. and Eaves, C.J., 2008, Steel factor coordinately regulates the molecular signature and biologic function of hematopoietic stem cells. BLOOD 112, 560.

Lefkovits, I. and Waldmann, H., 1979, LIMITING DILUTION ANALYSIS OF CELLS IN THE IMMUNE SYSTEM. Cambridge University Press, Cambridge.

Leong, K.G., Wang, B.E., Johnson, L. and Gao, W.Q., 2008, Generation of a prostate from a single adult stem cell. NATURE 456, 804-808.

Liang, Y., Jansen, M., Aronow, B., Geiger, H. and Van, Z.G., 2007, The quantitative trait gene latexin influences the size of the hematopoietic stem cell population in mice. NATURE GENETICS 39, 178-188.

Maillard, I., Koch, U., Dumortier, A., Shestova, O., Xu, L., Sai, H., Pross, S.E., Aster, J.C., Bhandoola, A., Radtke, F. and Pear, W.S., 2008, Canonical Notch Signaling Is Dispensable for the Maintenance of Adult Hematopoietic Stem Cells. CELL STEM CELL 2, 356-366.

Makinodan, T., and Albright, J.F., 1962, Cellular variation during the immune response: One possible model of cellular differentiation. J. CELL. COMP. PHYSIOL 60, 129-144.

Mather, K., 1949, The analysis of extinction time data in bioassay. BIOMETRICS 5, 127-143.

McCrady, M.H., 1915, The numerical interpretation of fermentation-tube results. J. INFECT. DIS 17, 183-212.

McCullagh, P., 1985, On the Asymptotic Distribution of Pearson's Statistic in Linear Exponential-Family Models. INTERNATIONAL STATISTICAL REVIEW 53, 61-67.

McCullagh, P. and Nelder, J.A., 1989, GENERALIZED LINEAR MODELS 2ND EDN, Chapman and Hall, London.

Moran, P.A.P., 1954a, J. HYG. 52, 189.

Moran, P.A.P., 1954b, J. HYG. 52, 444.

Nelder, J.A., and Wedderburn, R.W., 1972, Generalized linear models. JOURNAL OF THE ROYAL STATISTICAL SOCIETY A135, 370-384.

Omobolaji, O.A., In-Kyung P., Dalong Q., Michael, P., Michael, W. B. and Michael, F.C., 2008, Long-term haematopoietic reconstitution by Trp53-/-p16 Ink4a-/-p19 Arf-/-multipotent progenitors. NATURE 453, 228-232.

Oostendorp, R.A., Gilfillan, S., Parmar, A., Schiemann, M., Marz, S., Niemeyer, M., Schill, S., Hammerschmid, E., Jacobs, V.R., Peschel, C. and Götze, K.S., 2008, Oncostatin M-mediated Regulation Of KIT-Ligand-Induced ERK Signaling Maintains Hematopoietic Repopulating Activity Of Lin-CD34+ CD133+ Cord Blood Cells. STEM CELLS 26, 2164-2172.

Phelps, E. B., 1908, A method for calculating the numbers of B. coli from the results of dilution tests. PUBL. HEALTH PAP. REP. 33, 9-13.

Quintana, E., Shackleton, M., Sabel, M.S., Fullen, D.R., Johnson, T.M. and Morrison, S.J., 2008, Efficient tumour formation by single human melanoma cells. NATURE 456, 593-598.

Sambandam, A., Maillard, I., Zediak, V.P., Xu, L., Gerstein, R.M., Aster, J.C., Pear, W.S. and Bhandoola, A., 2005, Notch signaling controls the generation and differentiation of early T lineage progenitors. NAT IMMUNOL. 6, 663-670.

Schatton, T., Murphy, G.F., Frank, N.Y., Yamaura, K., Waaga-Gasser, A.M., Gasser, M., Zhan, Q., Jordan, S., Duncan, L.M., Weishaupt, C., Fuhlbrigge, R.C., Kupper, T.S., Sayegh, M.H. and Frank, M.H., 2008, Identification of cells initiating human melanomas. NATURE 451, 345-349.

Shortley, G. and Wilkins, J.R., 1965, Independent-action and birth-death models in experimental microbiology. BACTERIOLOGICAL REVIEWS 29, 102-141.

Shackleton, M., Vaillant, F., Simpson, K.J., Stingl, J., Smyth, G.K., Asselin-Labat, M.-L., Wu, L., Lindeman, G.J., and Visvader, J.E., 2006, Generation of a functional mammary gland from a single stem cell. NATURE 439, 84–88.

Siwko, S.K., Dong, J., Lewis, M.T., Liu, H., Hilsenbeck, S.G. and Li, Y., 2008, Evidence that an early pregnancy causes a persistent decrease in the number of functional mammary epithelial stem cells--implications for pregnancy-induced protection against breast cancer. STEM CELLS 26, 3205-3209.

Stein, M.F., 1922, ENG. CONT. 57, 445.

Strijbosch, L.W., Buurman, W.A., Does, R.J., Zinken, P.H. and Groenewegen, G., 1987, Limiting dilution assays. Experimental design and statistical analysis. J. IMMUNOLO. METHODS 97, 133-140.

Taswell, C., 1981, Limiting dilution assays for the determination of immunocompetent cell frequencies. I. Data analysis. J. IMMUNOL. 126, 1614-1619.

Taswell, C., 1984, Limiting dilution assays for the determination of immunocompetent cell frequencies. III. Validity tests for the single-hit Poisson model. J. IMMUNOL. METHODS 72, 29.

Taswell, C., 1987, Limiting dilution assays for the separation, characterization, and quantitation of biologically active particles and their clonal progeny. CELL SEPARATION: METHODS AND SELECTED APPLICATIONS, VOL.4, 109-145.

Thomas, D.G., 1972, Tests of fit for a one-hit vs. two-hit curve. APPL. STAT. 21, 103.

T.S., Sayegh, M.H. and Frank, M.H., 2008, Identification of cells initiating human melanomas. NATURE 451, 345.

Vaillant, F., Asselin-Labat, M.L., Shackleton, M., Forrest, N.C., Lindeman, G.J. and Visvader, J.E., 2008, The mammary progenitor marker CD61/B3 integrin identifies cancer stem cells in mouse models of mammary tumorigenesis. CANCER RES 68, 7711-7717.

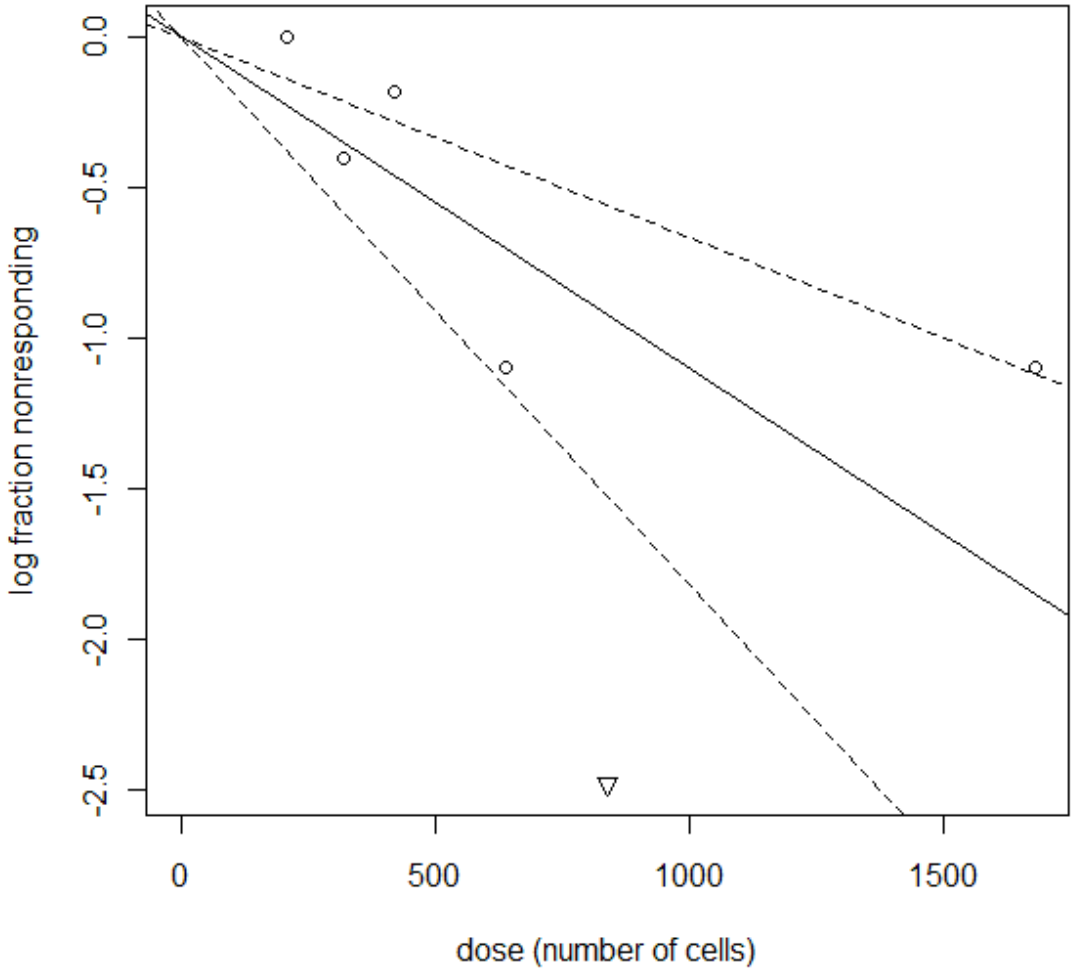
Vance, A (2009). Data Analysts Captivated by R's Power. New York Times, 7 January 2009. http://www.nytimes.com/2009/01/07/technology/business-computing/07program.html

Vermeulen, L., Todaro, M., de sousa Mello, F., Sprick, M.R., Kemper, K., Perez Alea, M., Richel, D.J., Stassi, G. and Medema, J.P., 2008, Single-cell cloning of colon cancer stem cells reveals a multi-lineage differentiation capacity. PROC NATL ACAD SCI U S A 105, 13427-13432.

Walkley, C.R., Shea, J.M., Sims, N.A., Purton, L.E. and Orkin, S.H., 2007, Rb Regulates Interactions between Hematopoietic Stem Cells and Their Bone Marrow Microenvironment. CELL 129, 1081-1095.

FIGURE

Figure 1
A log-fraction plot of the limiting dilution model fitted to the data in Table 5. The slope of the line is the log-active cell fraction. The dotted lines give the 95% confidence interval. The data value with zero negative response at dose 840 is represented by a down-pointing triangle.



TABLES

Table 1
Limiting dilution data showing the frequency of repopulating mammary cells from a tumorigenic mouse model (Vaillant et al, 2008). CD29loCD24+CD61+ cells from pre-neoplastic tissue of wild-type or MMTV-Wnt-1 mouse glands were transplanted into the cleared mammary fat pads of BALB/c recipients. Shown is the number of assays giving positive outgrowths.

Dose	Tested	Response	Group
100	9	0	Wild-type
200	6	0	Wild-type

50	13	1	MMTV-Wnt-1
100	19	3	MMTV-Wnt-1
200	6	3	MMTV-Wnt-1

Table 2
95% confidence intervals for repopulating mammary cell frequency for the data in Table 1 "∞" denotes infinity.

Group	Lower	Estimate	Upper
Wild-type	∞	∞	701
MMTV-Wnt-1	970	464	222

Table 3
Mammary epithelial outgrowths derived from single $Lin^{-}CD29^{hi}CD24^{+}$ cells.

Supporting cells	Number of outgrowths	Number of transplants
-	2	32
-	2	32
+	2	38

Table 4
Transplantation of melanoma cells mixed with Matrigel for patient 214.

Cells	Number of tumors	Number of injections
1000	6	6
200	6	6

Table 5
Limiting dilution data for tumor-forming frequency of cells from a p53 mouse model of breast cancer. Dose is the number of cells, Tested in the number of assays, Response is the number of positive assays yielding tumor growth.

Dose	Tested	Response
210	6	0
320	5	2
420	6	1
640	5	4
840	6	6
1680	4	4