

limdil Software for Limiting Dilution Analysis

The Walter and Eliza Hall Institute of Medical Research

Yifang Hu and Gordon K. Smyth

13 Oct 2008

ABSTRACT

limdil is a software application for limiting dilution analysis (LDA). It is the first such software to provide meaningful confidence intervals for all LDA data sets, including those with 0% or 100% responses. Other unique 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 known exactly. limdil is available as a function in the statmod software package for the R computing environment (www.r-project.org). A webtool at <http://bioinf.wehi.edu.au/software/limdil/> provides an easy interface to the R functionality. The supplementary material gives details of the mathematical algorithm implemented in limdil and it is available at http://bioinf.wehi.edu.au/software/limdil/limdil_supp.pdf.

INTRODUCTION

Limiting dilution analysis (LDA) is a technique to determine the frequency of positive cell response in a cell population. Originally developed in immunology, it is now one of the most commonly used statistical methods of stem cell research. LDA uses data from a series of dilution assays to estimate the proportion of cells in a mixed population with a particular ability. In stem cell applications, LDA typically assumes a single-hit model, whereby a successful growth response reveals the presence of at least one stem cell in the cell sample. Dilution assay allows the total number of cells in the sample to be varied so that the probability of a response becomes as informative as possible about the frequency of stem cells.

The single-hit Poisson model underlying LDA can be estimated from suitable data by the statistical technique of generalized linear [1]. A confidence interval can be obtained for the stem cell proportion by computing an asymptotic Wald confidence interval. This approach is implemented in the L-Calc software program for LDA (Stem Cell Technologies, www.stemcell.com).

The generalized linear approach to LDA fails when the proportion of positive responses is 0% or 100% for some assays. In such cases, it is more appropriate to compute exact binomial confidence intervals, similar to the classical intervals proposed by Clopper and Pearson [2]. We have developed a statistical strategy which combines the Clopper-Pearson and generalized linear model approaches, computing exact or asymptotic confidence intervals as appropriate for the data at hand. Our methodology is implemented in the `limdil` function of the `statmod` software package for R and in an easy-to-use webtool. Our software is the first to provide meaningful confidence intervals for all LDA data sets.

The `limdil` software has a number of other unique features. First, a test of the single-hit hypothesis is provided [3]. Secondly, `limdil` is able to test for differences between multiple limiting dilution data sets. In this way, `limdil` helps determine which cell populations are enriched for stem cells. Finally, `limdil` gives an explicit treatment of the special case where the total number of cells in the dilution sample has been counted exactly. Usually, the total number of cells is inferred as an expected value from the degree of dilution. `limdil` is the first software which can take advantage of cases where the number of cells is counted before the assay is completed.

Our methodology was originally developed as part of the search for mammary stem cells [4].

MATERIAL AND METHODS

The command-line version of `limdil` was developed using the R programming language. The webtool version of `limdil` utilizes the R version via a Perl and http interface.

The `limdil` function handles 0% or 100% positive responses as special cases. When there are no positive responses for any assay, an exact one-sided Clopper-Pearson confidence interval is available [2]. When there are 100% positive responses, Clopper-Pearson is not available, but `limdil` computes a conservative exact one-sided confidence interval. Exactly how this is done is explained in Supplementary Information.

In other cases, a binomial generalized linear model with complementary log-log link function is fitted by maximum likelihood, and Wald confidence intervals are computed [1]. The fitted models for different datasets are compared using likelihood ratio tests using the asymptotic chisquare approximation to the log-ratio [1].

The number of positive responses observed out of n_i assays at dose d_i is assumed to be binomial with probability p_i . Here i indexes the dilution and d_i represents the number of cells in the diluted samples. The classic single-hit model, which assumes that only one cell is necessary for a positive response from the sample as a whole, implies the complementary log-log linear model

$$\log(-\log(1 - p_i)) = \beta_0 + \log d_i$$

where β_0 is an unknown intercept parameter to be estimated.

In the classic LDA model, the number of cells is not observed directly. Rather d_i is the expected number of cells in a diluted sample of that volume. In that case, the

proportion of responding cells is estimated as $\lambda = \exp \beta_0$. On the other hand, if the number of cells d_i is counted exactly, then the estimated proportion is

$\lambda = 1 - \exp(-\exp \beta_0)$. If λ is small, then the two formula yield very similar values.

The software outputs confidence intervals for $1/\lambda$, representing the number of cells required on average to obtain one responding cell.

A statistical test of the single-hit model can be obtained by fitting the two-parameter model

$$\log(-\log(1 - p_i)) = \beta_0 + \beta_1 \log d_i$$

and testing whether $\beta_1 = 1$ [3]. This tests whether the data supports the assumptions of the single-hit hypothesis. limdil tests this hypothesis using a likelihood ratio test.

RESULTS

Table 1 shows data from an experiment on breast cancer stem cells [5]. The aim is to identify cell markers which isolate cells populations enriched for tumorigenic cells.

The data compares cells from Thy and non-Thy isolations. Thy cells are selected by a series of markers including Thy1+, CD24+, CD49f+ and CD45-. The non-Thy cells are similar but are Thy1-.

Table 2 shows confidence intervals output by limdil for the this data. Thy cells are estimated to have a 1/203 tumor-forming frequency whereas non-Thy cells have a much lower 1/54905 tumor-forming frequency. This frequency difference is statistically highly significant using a likelihood ratio test ($\chi_1^2 = 32.9$, $P = 9.7e - 9$).

Meanwhile, there is no reason to doubt the adequacy of the single-hit model

($\chi_1^2 = 0.49$, $P = 0.49$).

Table 3 gives data on the frequency of repopulating mammary cells from a tumorigenic mouse model [4]. 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. Despite the absence of responses, limdil 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.

In general, the limdil web-tool accepts an input data table in the same format as Tables 1 and 3. The user sets the confidence level required (95% default) and indicates whether the actual number of cells observed directly. The limdil output gives (1) estimated confidence intervals for the frequency of positive cells in each group, (2) pairwise comparisons between groups and an overall test of group differences and (3) a test of the single-hit model.

DISCUSSION AND CONCLUSION

While earlier LDA software programs, such as L-Calc, are able to compute confidence intervals for the breast cancer stem cell data example, limdil is the only software to the authors knowledge which can compare the two groups or assess the single-hit hypothesis. In the mammary stem cell example, limdil is the only software which can provide a confidence interval for the wild-type group with zero responses

or compare the wild-type and tumorigenic models. While the examples here have two groups, limdil can handle an arbitrary number of groups with the same ease.

limdil offers more flexible and powerful statistical analysis of limiting dilution assays than previously available. limdil is available as a command-line function in R for use by bioinformaticians and programmers, and as an easily accessible webtool for biologists. It provides a valuable resource for stem cell and immunological research.

ACKNOWLEDGMENT

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

REFERENCES

1. Collett D. Modelling Binary Data. Boundary Row, London: Chapman & Hall, 1991: 112-113.
2. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *BIOMETRIKA* 1934;26: 404–413.
3. Bonnefoix T, Bonnefoix P, Verdiela P et al. Fitting limiting dilution experiments with generalized linear models results in a test of the single-hit Poisson assumption. *JOURNAL OF IMMUNOLOGICAL METHODS* 1996;194: 113–119.

4. Francois V, Marie-Liesse A, Mark S et al. The mammary progenitor marker CD61/B3 integrin identifies cancer stem cells in mouse models of mammary tumorigenesis. *CANCER RESEARCH* 2008;68: 7711-7717.
5. Cho RW, Wang X, Diehn M et al. Isolation and molecular characterization of cancer stem cells in MMTV-Wnt-1 murine breast tumors. *STEM CELLS* 2007;26: 364–371.

Table 1. Limiting dilution data comparing two isolations of tumorigenic cells. Dose is the number of cells, Tested in the number of assays, Response is the number of positive assays yielding tumor growth. Cells with Thy markers are enriched for tumorigenic cells.

Dose	Tested	Response	Group
500	2	2	Thy
100	1	0	Thy
50	9	2	Thy
10000	12	2	non-Thy
5000	15	2	non-Thy
2000	20	0	non-Thy

Table 2. 95% confidence intervals for tumorigenic cell frequency.

Group	Lower	Estimate	Upper
Thy	581	203	71
non-Thy	146062	54905	20639

Table 3. Limiting dilution data showing the frequency of repopulating mammary cells from a tumorigenic mouse model. CD29^{lo}CD24⁺CD61⁺ cells from preneoplastic 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 4. 95% confidence intervals for repopulating mammary cell frequency. “Inf” denotes an infinite or unbounded quantity.

Group	Lower	Estimate	Upper
Wild-type	Inf	Inf	701
MMTV-wnt-1	970	464	222