INFECTION AND IMMUNITY, Mar. 2006, p. 1500–1504 Vol. 74, No. 3
0019-9567/06/$08.00+0 doi:10.1128/IAI.74.3.1500–1504.2006

Estimating the Relative Contributions of Virulence Factors for Pathogenic Microbes†

Erin E. McClelland,[1]* Paul Bernhardt,[2] and Arturo Casadevall[1]

Department of Medicine, 702 Golding, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, New York 10461,[1] *and Department of Psychology, Westminster College, 1840 South 1300 East, Salt Lake City, Utah 84105*[2]

Received 6 November 2005/Returned for modification 23 November 2005/Accepted 18 December 2005

Many pathogenic microbes have multiple virulence factors that can cause damage to the host and thus contribute to an overall virulence phenotype for that organism. Although current techniques are suitable for demonstrating that a particular microbial characteristic contributes to virulence, no formal approach for defining the relative contributions of multiple virulence factors to overall virulence has been proposed. This paper describes the use of multivariate linear regression to estimate the relative contributions of virulence factors to the overall phenomenon of virulence. The approach is illustrated here with sample calculations of the relative contributions of individual *Cryptococcus neoformans* and *Bacillus anthracis* virulence factors to the overall virulence phenotype. These calculations were derived from a small underpowered experimental data set for the fungus and two larger sets of randomly generated data for both microbes. The major limitation of this method is a requirement for large data sets of microbial strains that differ in virulence and virulence factor expression. Multivariate linear regression can be used to identify the relative levels of importance of virulence factors in virulence studies, and this information can be used to prioritize antigen identification for vaccine development and the design of antimicrobial strategies that target virulence mechanisms.

While the literature contains an abundance of papers describing microbial determinants (virulence factors) that contribute to microbial virulence, no methodology has been proposed for determining the relative contributions of individual virulence factors to the overall virulence phenotype. For example, the pathogenic yeast *Cryptococcus neoformans* has a number of virulence factors that are thought to contribute to virulence, such as capsule (2, 12), melanin synthesis (13, 26), and a variety of secreted enzymes (6, 7, 14, 22). We have investigated the use of multivariate linear regression as a tool to rank the virulence factors in disease importance. Multivariate linear regression is a well-established statistical tool that has been used to analyze the relative contributions of different parameters in other fields, including the shape of coronary arteries on the presence, extent, and severity of disease (8), trace metal levels in tannery effluents in relation to soil and water (24), and the quality of stimulation in the family environment and a child's cognitive development (1). Here we demonstrate how this approach can be applied to a variety of microbes with more than one virulence factor. Indeed, many pathogenic microbes possess adhesion molecules, enzymes, and toxins, each of which can be shown to influence virulence in experimental systems. However, such experiments almost always involve the evaluation of a single virulence factor at a time, despite the fact that virulence is the result of their combined effects. Since new virulence factors are being identified for many microbes (9, 11, 19, 25), there is an urgent need for a systematic approach to determine the relative contributions of virulence factors in pathogenesis. Targets for vaccine and drug design could be selected from this information.

* Corresponding author. Mailing address: Department of Medicine, 702 Golding, Albert Einstein College of Medicine, 1300 Morris Park Avenue, Bronx, NY 10461. Phone: (718) 430-3768. Fax: (718) 430-8701. E-mail: mcclella@aecom.yu.edu.

† Supplemental material for this article may be found at http://iai.asm.org/.

MATERIALS AND METHODS

When considering how each individual virulence factor contributes to an overall virulence phenotype, one must first identify a measure of virulence. Virulence has been defined as the relative capacity of a microbe to cause damage in a host (5). Although virulence is universally considered to be a microbial characteristic, it is expressed only in susceptible hosts and its readout involves the measurement of a host-related parameter (4). Some commonly used measures of virulence are mortality, microbial burden on tissue, or lifetime reproductive success of infected hosts versus uninfected hosts (21). Additional measurements of virulence could include different measures of host damage and the immune response (5).

To obtain an estimate of the relative contributions of various virulence factors by using multivariate linear regression for a given microbe requires (i) a set of related strains that differ in virulence factor expression in a manner that can be quantified and (ii) a system for comparing the relative levels of virulence of these strains. When considering this approach, it is important to have an adequate definition for the word "strain." A strain has been defined as "a group of organisms of the same species, having distinctive [phenotypic] characteristics but not usually considered a separate breed or variety" (http://www.thefreedictionary.com/strain).

The data set used for multivariate linear regression should conform to all statistical assumptions, including an adequate sample size (10, 15) (Table 1), normality, screening for univariate and multivariate outliers, and independent variables that are linear and homoscedastic (exhibit similar levels of variance across the range of the outcome variable) (23).

Initially, a data set involving 18 strains (16) was used for the *C. neoformans* analysis. However, due to the small sample size, it was underpowered. Because data sets of sufficient power for *C. neoformans* or *Bacillus anthracis* were not available, we generated demonstration data sets using RandGen (17). RandGen is a computer program that can generate random data with specified distributions and with specified correlations among the variables. The gamma distribution was used to ensure generation of data with means, skewnesses, lower bounds, and standard deviations similar to those seen in actual data sets. RandGen provided random data sets with correlations of predictors to the outcome variable and between the various predictor variables similar to those that might be seen in naturally occurring data sets. For the *C. neoformans* data, actual data from 18 strains (16) were used for the parameters entered into RandGen. RandGen uses these parameters, together with user-entered parameters, to control correlations

TABLE 1. Hierarchical regression analysis results for initial 18-strain data set

Virulence factor(s)	Correlation with time to death (r)	R^2 change	df	F value	P value[a]
Capsule size in vitro	0.530	0.281	1, 14	5.467	0.035
Melanin production	0.415	0.085	1, 13	1.731	0.211 (NS)
Doubling time	−0.173	0.001	1, 12	0.026	0.875 (NS)
GXM release in vitro	−0.202	0.003	1, 11	0.047	0.831 (NS)
Laccase secretion	−0.068	0.002	1, 10	0.040	0.846 (NS)
Phospholipase secretion	−0.170	0.046	1, 9	0.705	0.423 (NS)
Urease secretion	0.190	0.015	1, 8	0.209	0.660 (NS)
Phagocytosis index	0.117	0.181	1, 7	3.269	0.114 (NS)
Macrophage killing	0.155	0.078	1, 6	1.513	0.265 (NS)
All factors but capsule size	0.530	0.538	1, 6	1.021	0.496 (NS)

[a] NS, not significant.

among variables and to create random data sets. For *B. anthracis*, the data set was generated in the same manner as for *C. neoformans*, except the input parameters for RandGen were completely fabricated. These data sets were randomly generated solely to increase the sample size and statistical power in order to illustrate the approach. The subsequent analysis using these data sets should not be interpreted to have any biological meaning for *C. neoformans* or *B. anthracis*.

In the present analysis, the virulence factors used as the predictor variables for *C. neoformans* were capsule size, melanin production, glucuronoxylomannan (GXM) release, urease production, and growth rate (doubling time), because these factors were determined a priori to contribute to virulence. For *B. anthracis*, the virulence factors used as the predictor variables were capsule, toxin, and anthralysin. Predictor variables were entered into the regression equation to determine the unique contribution of each virulence factor to the outcome variable (time to death). Analysis was done via standard, multivariate hierarchical regression using SPSS v.13 for Windows.

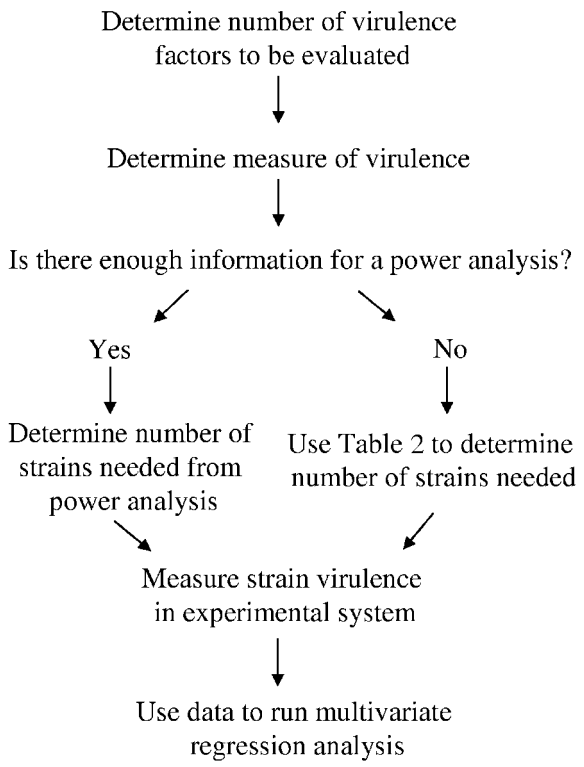

FIG. 1. Approach to determine the relative levels of importance of virulence factors in microbial pathogenesis by using multivariate linear regression.

RESULTS

The approach used to apply the method of multivariate linear regression to a pathogenic microbe is illustrated in Fig. 1. For *C. neoformans*, multivariate hierarchical linear regression was initially performed on a set of 18 strains for which experimental data were available (16) (Table 1). Even with this relatively small set, the capsule size in vitro emerges as the dominant virulence factor but the other parameters do not reach statistical significance. Because this was a small sample size and was severely underpowered, RandGen was used to generate a larger demonstration data set that had sufficient power for multivariate hierarchical linear regression. Although this data set in the subsequent analysis was hypothetical, the values used were generated within constraints imposed by the experimental values listed in Table 1 and, consequently, are not completely random. By varying the specification of the correlation between predictors, it was possible to demonstrate the impact of these correlations on power (Table 2). In naturally occurring data sets, the various virulence factors would certainly have some correlation with each other due to the fact that they are often linked by underlying mechanisms, such as coordinate gene regulation. To the extent that the researcher seeks relatively uncorrelated predictors, sample sizes may be markedly reduced. As Table 2 indicates, a lower correlation

TABLE 2. Estimation of sample size by using different degrees of correlation between predictors[a]

No. of virulence factors	Mean effect size of predictor on outcome variable	No. of strains needed[b]			
		r_{xx} = 0.50	r_{xx} = 0.40	r_{xx} = 0.30	r_{xx} = 0.20
2	Large (0.45)	64	49	37	29
	Medium (0.30)	231	179	141	113
	Small (0.15)	559	433	343	276
3	Large (0.45)	120	79	52	33
	Medium (0.30)	455	314	218	150
	Small (0.15)	1,109	772	542	379
5	Large (0.45)	279	160	85	39
	Medium (0.30)	1,116	692	419	240
	Small (0.15)	2,752	1,731	1,070	632

[a] Derived from Maxwell (15).

[b] Type I error rate, 0.05; type II error rate, 0.20; power, 0.8. r_{xx} is the average size of correlation among the predictors.

TABLE 3. Hierarchical regression analysis results for *C. neoformans*, with high correlation between predictors

Virulence factor	Correlation with time to death (*r*)	R^2 change	df	*F* value	*P* value[a]
Capsule size	0.435	0.190	1, 277	64.82	<0.001
Melanin	0.511	0.105	1, 276	40.98	<0.001
GXM release	−0.330	0.001	1, 275	0.25	0.62 (NS)
Urease	0.450	0.007	1, 274	2.68	0.10 (NS)
Doubling time	−0.223	0.014	1, 273	5.70	0.02

[a] NS, not significant.

between the predictors (r_{xx}) yields a reduction in required sample size (number of strains) to obtain acceptable power. Thus, two demonstration data sets were created for *C. neoformans*. One was based on high intercorrelations between the predictors, necessitating a large sample size. The second was based on low correlations between the predictors, allowing a great reduction in the sample size, as suggested in Table 2. Therefore, to find the large effects of five predictors of time to death for *C. neoformans* (capsule size, melanin production, GXM release, urease production, and growth rate) when the correlation between predictors is high, the randomly generated sample size is 279 (see Table S1 in the supplemental material). When the correlation between predictors is small in comparison to the effect size, a smaller sample size of 39 (see Table S2 in the supplemental material) can suffice to find the large effects of the predictors.

The regression summary of five predictors of time to death for *C. neoformans* when the correlation between predictors is high is given in Table 3. Results of the regression indicated that capsule size and melanin production significantly accounted for 19.0% and 10.5% of the variance of time to death, respectively. In vitro growth rate (doubling time) accounted for 1.4% of the variance of time to death, after accounting for the previously entered variables of interest. Urease production and GXM release did not account for a significant portion of the variance for this demonstration data set.

The regression summary of five predictors of time to death for *C. neoformans* when the correlation between predictors is small is given in Table 4. Results of the regression indicated that capsule size and melanin production significantly accounted for 41.4% and 16.2% of the variance of time to death, respectively. GXM release, in vitro growth rate (doubling time), and urease production did not account for a significant portion of the variance of time to death for this demonstration data set.

TABLE 4. Hierarchical regression analysis results for *C. neoformans*, with low correlation between predictors

Virulence factor	Correlation with time to death (*r*)	R^2 change	df	*F* value	*P* value[a]
Capsule size	0.643	0.414	1, 37	26.10	<0.001
Melanin	0.571	0.162	1, 36	13.76	0.001
GXM release	−0.323	0.002	1, 35	0.15	0.70 (NS)
Urease	0.344	0.005	1, 34	0.38	0.54 (NS)
Doubling time	0.075	0.038	1, 33	3.33	0.08 (NS)

[a] NS, not significant.

TABLE 5. Hierarchical regression analysis results for *B. anthracis*

Virulence factor	Correlation with time to death (*r*)	R^2 change	df	*F* value	*P* value
Capsule	−0.456	0.208	1, 148	38.77	<0.001
Anthrax toxin	−0.565	0.199	1, 147	49.39	<0.001
Anthralysin	−0.365	0.020	1, 146	5.14	0.025

To illustrate the method for a microbe with fewer identified virulence determinants, we fabricated a randomly generated sample set for *B. anthracis* by using three predictors of virulence (capsule, toxin, and anthralysin), with a sample size of 150 (see Table S3 in the supplemental material). This is the sample size suggested in Table 2 to find larger effects of predictors that have low correlation with each other. The regression summary of three predictors of time to death for *B. anthracis* is given in Table 5. For this hypothetical data set, capsule, anthrax toxin, and anthralysin significantly account for 20.8%, 19.9%, and 2.0% of the variance of time to death, respectively. Notably, anthralysin was moderately correlated with time to death, yet the small correlation with other predictors makes it possible to expose its contribution. If the correlation with other predictors were higher, a larger number of strains would be required. Therefore, researchers are advised to be sensitive to the degree of correlation between virulence factors when determining the sample size (number of strains) for a study.

DISCUSSION

Assuming that a measurement for virulence that produces quantitative data for all measures of strain virulence exists and that a strain set that expresses various virulence factors in different measurable proportions is available, then multivariate linear regression can be used to estimate the relative contribution of each virulence factor to the overall virulence phenomenon. In addition, multivariate linear regression can also be used to determine if there are undiscovered virulence factors that contribute to the overall phenomenon of virulence.

There are three different types of multivariate linear regression: standard multiple regression, sequential (hierarchical) regression, and statistical regression. The choice of which regression method to employ is dependent on the nature of the research question (23). In standard multiple regression, all variables are entered into the same block so that they are each assessed at the same time. Thus, standard regression illustrates how much of the outcome variable is explained by all of the variables at once. Standard regression is an assessment of a complete model when there is no intention of examining the contributions of individual predictors. In hierarchical regression, the variables are entered in different blocks in a specific order, with the order of entry resulting from theoretical or logical importance. Thus, hierarchical regression illustrates the unique contribution of each predictor variable to the variance in the outcome variable, while taking into account the contribution of previously identified significant predictors. In statistical regression, the variables are entered or removed in different blocks in an order that is specified by statistical criteria. Therefore, statistical regression is an exploratory method use-

ful for selecting which variables best predict the outcome variable when there is no theoretical rationale for a priori prioritization (23). To answer the question of how multiple virulence factors contribute to overall virulence, hierarchical regression is the best type of linear regression to use when the researcher has an a priori expectation about the relative importance of each virulence factor, as in the case of *C. neoformans*. Statistical regression is appropriate when the researcher has no a priori expectations or when there are numerous factors that may or may not contribute to virulence. In that case, statistical regression can be utilized to remove factors that do not contribute to virulence.

When considering multivariate linear regression analysis, it is essential to use an adequate sample size. The sample size (number of strains) required is a function of the number of virulence factors measured and the questions asked. To determine the dominant virulence factor (a large effect), a relatively small number of strains is needed. Conversely, to determine the virulence factor that contributes the least to pathogenesis (a small effect), a relatively large number of strains is needed. The number of strains required can be estimated from a power analysis or the use of a general rule of thumb. A good explanation of the parameters needed for a power analysis can be found at the website http://www.statsoft.com/textbook/stpowan.html. If a power analysis cannot be done, as a general rule of thumb, 10 datum points are required for each predictor variable in the regression, although Maxwell (15) suggests that this is overly optimistic. Maxwell recognized that the correlation among the predictors has a powerful effect on the power to find even large effect sizes. Reduction of the magnitude of the correlation between predictors results in smaller sample sizes to obtain equivalent power (Table 2).

Using two microbial examples and a demonstration data set of strains that differ in virulence factor expression, we show the usefulness of multivariate linear regression analysis to determine the relative levels of importance of virulence factors in microbial pathogenesis. The method relies on established statistical principles and uses readily available commercial software. Comparing the results from the small experimental data set and the larger randomly generated data set based on experimentally derived input parameters shows that, even with a small sample size, *C. neoformans* capsule contributes most to virulence (Table 1).

Because the sample size requirement can be quite large and the number of strains available often limits researchers, we recognize the practical limitations involved in using this method. Thus, one alternative to a larger sample size is to select for comparison virulence factors that are relatively uncorrelated (independent) with each other. If predictor variables that are relatively uncorrelated with each other and with the most virulent predictor are used, fewer strains are needed because each variable is allowed to uniquely contribute to the variance in the dependent variable and fewer variables are used in the model. Another alternative is to decrease the number of virulence factors tested, either by deleting one variable or by combining two or more variables into one variable. Unfortunately, this can sometimes lead to specification error and biased parameter estimates (15). Finally, another alternative is to conduct the regression analysis with the strains available and describe any significant results as preliminary findings (due to the small sample size used). However, note that a significant result may be harder to find with a small sample size, due to problems with power, which makes it more difficult to detect significant results.

While this method will identify the virulence factor most significantly related to the outcome variable (time of death, in our simulations), the greatest limitation is the requirement for increasingly large numbers of different strains as the virulence factors multiply. In fact, the investigator would have to not only generate the strains but also test them in a suitable system to measure virulence, which could prove prohibitive with regard to animal costs. However, this second limitation may be mostly overcome in those situations with the use of signature-tagged mutagenesis (18, 20) or the use of alternative hosts, such as amoebae, slime mold, and *Caenorhabditis elegans*, to measure relative levels of virulence (3). However, while these alternate hosts would be extremely useful in measuring virulence, their use adds another limitation to the method in that even though many microbial virulence factors may be evolutionarily conserved and can cause damage in multiple hosts, there may be a limited correlation to real human disease.

Another way to use multivariate regression analysis that would be perhaps more related to real human disease would be to collect clinical microbial strains from patients and try to correlate virulence factor expression and clinical disease. Results from these kinds of analyses would be very important in vaccine and antimicrobial drug design. However, due to differences in virulence factor expression among microbes, the method may have greater relevance for vaccine design, especially if an effective vaccine requires a cocktail of multiple virulence factor antigens.

In summary, we demonstrate how one can estimate the relative contributions of virulence factors to the overall virulence phenotype by applying multivariate regression analysis. The method requires only an appropriate sample size and a system that yields quantitative measures of virulence. Since the greatest limitation of the method is that the amount of independent measurements rises rapidly as the virulence factors multiply, knowledge of this relationship may promote the development of high-throughput techniques for measurement of virulence.

ACKNOWLEDGMENTS

We thank Megan McClelland, Oscar Zaragoza, and John Kircher for statistical advice and critical discussion.

This study was supported in part by NIH grants GM-071421, AI033142, AI033774, AI052733, AI057158-03, and HL059842 to A.C.

REFERENCES

1. **Andrade, S. A., D. N. Santos, A. C. Bastos, M. R. Pedromonico, N. de Almeida-Filho, and M. L. Barreto.** 2005. Family environment and child's cognitive development: an epidemiological approach. Rev. Saude Publica **39:**606–611. (In Portuguese.)
2. **Bulmer, G. S., M. D. Sans, and C. M. Gunn.** 1967. *Cryptococcus neoformans*. I. Nonencapsulated mutants. J. Bacteriol. **94:**1475–1479.
3. **Casadevall, A.** 2005. Host as the variable: model hosts approach the immunological asymptote. Infect. Immun. **73:**3829–3832.
4. **Casadevall, A., and L. Pirofski.** 2001. Host-pathogen interactions: the attributes of virulence. J. Infect. Dis. **184:**337–344.
5. **Casadevall, A., and L. A. Pirofski.** 1999. Host-pathogen interactions: redefining the basic concepts of virulence and pathogenicity. Infect. Immun. **67:**3703–3713.
6. **Cox, G. M., H. C. McDade, S. C. Chen, S. C. Tucker, M. Gottfredsson, L. C. Wright, T. C. Sorrell, S. D. Leidich, A. Casadevall, M. A. Ghannoum, and**

J. R. Perfect. 2001. Extracellular phospholipase activity is a virulence factor for Cryptococcus neoformans. Mol. Microbiol. **39:**166–175.
7. **Cox, G. M., J. Mukherjee, G. T. Cole, A. Casadevall, and J. R. Perfect.** 2000. Urease as a virulence factor in experimental cryptococcosis. Infect. Immun. **68:**443–448.
8. **Demirbag, R., and R. Yilmaz.** 2005. Effects of the shape of coronary arteries on the presence, extent, and severity of their disease. Heart Vessels **20:**224–229.
9. **Dziva, F., P. M. van Diemen, M. P. Stevens, A. J. Smith, and T. S. Wallis.** 2004. Identification of Escherichia coli O157:H7 genes influencing colonization of the bovine gastrointestinal tract using signature-tagged mutagenesis. Microbiology **150:**3631–3645.
10. **Green, S. B.** 1991. How many subjects does it take to do a regression analysis? Multivar. Behav. Res. **26:**499–510.
11. **Hubálek, M., L. Hernychová, M. Brychta, J. Lenèo, J. Zechovská, and J. Stulík.** 2004. Comparative proteome analysis of cellular proteins extracted from highly virulent *Francisella tularensis* ssp. *tularensis* and less virulent *F. tularensis* ssp. *holarctica* and *F. tularensis* ssp. *mediaasiatica*. Proteomics **4:**3048–3060.
12. **Kozel, T. R., G. S. Pfrommer, A. S. Guerlain, B. A. Highison, and G. J. Highison.** 1988. Role of the capsule in phagocytosis of Cryptococcus neoformans. Rev. Infect. Dis. **10**(Suppl. 2)**:**S436–S439.
13. **Kwon-Chung, K. J., and J. C. Rhodes.** 1986. Encapsulation and melanin formation as indicators of virulence in *Cryptococcus neoformans*. Infect. Immun. **51:**218–223.
14. **Liu, L., R. P. Tewari, and P. R. Williamson.** 1999. Laccase protects *Cryptococcus neoformans* from antifungal activity of alveolar macrophages. Infect. Immun. **67:**6034–6039.
15. **Maxwell, S. E.** 2000. Sample size and multiple regression analysis. Psychol. Methods **5:**434–458.
16. **McClelland, E. E., W. T. Perrine, W. K. Potts, and A. Casadevall.** 2005. The relationship of virulence factor expression to evolved virulence in mouse-passaged *Cryptococcus neoformans* lines. Infect. Immun. **73:**7047–7050.
17. **Miller, J.** 2002. RandGen. A program for generating random numbers (version 1.2). http://psy.otago.ac.nz/miller/. University of Otago, Dunedin, New Zealand.
18. **Nelson, R. T., J. Hua, B. Pryor, and J. K. Lodge.** 2001. Identification of virulence mutants of the fungal pathogen Cryptococcus neoformans using signature-tagged mutagenesis. Genetics **157:**935–947.
19. **Nguyen, M. H., S. Cheng, and C. J. Clancy.** 2004. Assessment of Candida albicans genes expressed during infections as a tool to understand pathogenesis. Med. Mycol. **42:**293–304.
20. **Potvin, E., D. E. Lehoux, I. Kukavica-Ibrulj, K. L. Richard, F. Sanschagrin, G. W. Lau, and R. C. Levesque.** 2003. In vivo functional genomics of Pseudomonas aeruginosa for high-throughput screening of new virulence factors and antibacterial targets. Environ. Microbiol. **5:**1294–1308.
21. **Poulin, R., and C. Combes.** 1999. The concept of virulence: interpretations and implications. Parasitol. Today **15:**474–475.
22. **Salas, S. D., J. E. Bennett, K. J. Kwon-Chung, J. R. Perfect, and P. R. Williamson.** 1996. Effect of the laccase gene, *CNLAC1*, on virulence of *Cryptococcus neoformans*. J. Exp. Med. **184:**377–386.
23. **Tabachnick, B. G., and L. S. Fidell.** 2001. Using multivariate statistics, 4th ed. Allyn & Bacon, Boston, Mass.
24. **Tariq, S. R., M. H. Shah, N. Shaheen, A. Khalique, S. Manzoor, and M. Jaffar.** 2005. Multivariate analysis of trace metal levels in tannery effluents in relation to soil and water: a case study from Peshawar, Pakistan. J. Environ. Manag. doi:10.1016/j.jenvman.2005.05.009. [Epub ahead of print.]
25. **Wang, P., G. M. Cox, and J. Heitman.** 2004. A Sch9 protein kinase homologue controlling virulence independently of the cAMP pathway in Cryptococcus neoformans. Curr. Genet. **46:**247–255.
26. **Wang, Y., P. Aisen, and A. Casadevall.** 1995. *Cryptococcus neoformans* melanin and virulence: mechanism of action. Infect. Immun. **63:**3131–3136.

Editor: W. A. Petri, Jr.