JOURNAL OF CLINICAL MICROBIOLOGY, Nov. 2001, p. 4190–4192 Vol. 39, No. 11
0095-1137/01/$04.00+0 DOI: 10.1128/JCM.39.11.4190–4192.2001

Determining Confidence Intervals When Measuring Genetic Diversity and the Discriminatory Abilities of Typing Methods for Microorganisms

HAJO GRUNDMANN,[1]* SATOSHI HORI,[1] AND GREGOR TANNER[2]

Division of Microbiology and Infectious Diseases, University Hospital Nottingham,[1] and School of Mathematical Sciences,[2] University of Nottingham, Nottingham NG7 2UH, United Kingdom

Received 21 June 2001/Returned for modification 24 July 2001/Accepted 21 August 2001

We describe here a method for determining confidence intervals for a commonly used index of diversity. This approach facilitates the comparison of the genetic population structure of microorganisms isolated from different environments and improves the objective assessment of the discriminatory power of typing techniques.

The discrimination of organisms on the basis of variable phenotypic or genetic markers is still the mainstay of quantitative microbial ecology and descriptive epidemiology. To determine the diversity of microorganisms in defined environments (ecosystems) or to identify the reproductive success of disease causing organisms, i.e., the spread of particular strains between hosts, genetic typing techniques are deployed which have the ability to distinguish diverse organisms of the same species. Importantly, when one is comparing the diversity of a single species between different ecosystems or comparing the various typing methods used to resolve such differences, a robust statistical approach is required that allows an objective assessment. To this end, indices of diversity have been defined mathematically that are based on the frequency with which organisms of a particular type occur in a population or can be discriminated by a given typing tool (3, 4, 5). Individuals of a population will belong to one of Z types and will occur with frequencies of $\pi_1 \ldots \pi_Z$ such that $\Sigma\pi = 1$. For microorganisms that usually have a very large population size, the genetic diversity (λ) can be described as $\lambda = 1 - \Sigma\pi^2$, which will be the probability that two individuals chosen at random will be of a different type.

Inferences on the diversity of the population involve a sampling process. The index of diversity D, as defined by Simpson (5) and lately utilized for the assessment of the discriminatory power of typing techniques (2, 6), is an unbiased estimate of the true diversity λ of a population based on a sample of n individuals. Simply by chance, different samples will give different results, the difference being due to sample variation and by drawing repeated samples, the precision of the mean estimate for D will improve. If repeated samples of a fixed size n are drawn from the sample population, the values of D will be distributed about λ with the variance σ^2 (5):

$$\sigma^2 = \frac{4}{n}[\sum\pi_j^3 - (\sum\pi_j^2)^2], \quad (1)$$

where π_j is the frequency n_j/n, n_j is the number of strains belonging to the jth type, and n is the total number of strains in the sample population. An estimate of the standard deviation of λ is given by the square root of σ^2, and we propose the following as approximate 95% confidence interval (CI):

$$CI = [D - 2\sqrt{\sigma^2}, D + 2\sqrt{\sigma^2}]. \quad (2)$$

We have applied these equations to determine confidence intervals (i) when assessing the genetic diversity of *Staphylococcus aureus* isolated from healthy carriers in the community as opposed to hospitalised patients and (ii) when comparing the discriminatory power of macrorestriction analysis by using *Sma*I restriction patterns with that of RAPD [random(ly) amplified polymorphic DNA] typing.

By using the same sampling frame, healthy individuals in the community and inpatients of the same age group who had stayed at the University Hospital in Nottingham for more than 3 weeks were sampled by use of swabs taken from the anterior nares. Carriage strains of *S. aureus* obtained from the community were genotyped by *Sma*I macrorestriction analysis, as well as by RAPD typing, by using two different primers. Carriage strains from hospitalized patients were typed by macrorestriction alone. Macrorestriction and RAPD analyses were performed by standard published protocols (7; Harmony [http://www.phls.co.uk/International/Harmony/microtyping.htm]), and genotypes were classified according to conventional criteria, with more than two band differences defining separate genotypes or a cutoff value of 70% for Pearson correlation coefficients discriminating genotypes in the RAPD approach (1, 8).

Among the 117 carriage strains from the community, 57 types were distinguished by macrorestriction analysis. Of 117 carriage strains obtained from the hospital population 55 types were distinguished. The distribution and type frequencies are presented in Tables 1 and 2. The genetic diversity (D) of carriage strains in the community was 97.6%, with a CI of 96.8 to 98.5%, and for carriage strains from hospital patients (D) it equalled 89.5%, with a CI of 84.4 to 94.7% (the CI values did not overlap).

We also compared the discriminatory ability of macrorestriction analysis with RAPD typing for the sample of community carriage isolates. While macrorestriction analysis identified 57

* Corresponding author. Mailing address: Division of Microbiology and Infectious Diseases, University Hospital, Queen's Medical Centre, Clifton Blvd., Nottingham NG7 2UH, United Kingdom. Phone: 44-115-970-9162. Fax: 44-115-970-9233. E-mail: Hajo.Grundmann@Nottingham.ac.uk.

TABLE 1. Frequency of PFGE types among community carriage strains of *S. aureus*

PFGE type	No. of types	Frequency (%)
CP1	9	7.7
CP2	8	6.8
CP3	7	6.0
CP4	6	5.1
CP5	5	4.3
CP6	5	4.3
CP7	4	3.4
CP8	4	3.4
CP9	4	3.4
CP10	3	2.6
CP11	3	2.6
CP12	3	2.6
CP13	3	2.6
CP14	2	1.7
CP15	2	1.7
CP16	2	1.7
CP17	2	1.7
CP18	2	1.7
CP19	2	1.7
CP20	2	1.7
CP21	2	1.7
CP22	2	1.7
CP23 to CP57	1[a]	0.9[b]
Total	117	100.0

[a] One isolate each of these 35 types.
[b] The frequency for each of these 35 types.

TABLE 2. Frequency of PFGE types among hospital carriage strains of *S. aureus*

PFGE type	No. of types	Frequency (%)
HP1	37	31.6
HP2	5	4.3
HP3	4	3.4
HP4	4	3.4
HP5	3	2.6
HP6	3	2.6
HP7	3	2.6
HP8	3	2.6
HP9	2	1.7
HP10	2	1.7
HP11	2	1.7
HP12	2	1.7
HP13	2	1.7
HP14	2	1.7
HP15	2	1.7
HP16	2	1.7
HP17 to HP55	1[a]	0.9[b]
Total	117	100.0

[a] One isolate each of these 39 types.
[b] The frequency for each of these 39 types.

*Sma*I restriction patterns, 26 types could be discriminated when RAPD results generated by different primers were combined (Table 3). The index of diversity based on the combined RAPD grouping was 89.9%, and the CI was 86.5 to 93.3%.

For nominal scale data, such as phenotypic or genetic markers, there is no mean or median that could serve as a reference for a central tendency or a measure of variability. Instead, we can invoke the concept of diversity to determine the distribution of observations among categories. We suggest use of the standard deviation for Simpson's index of diversity as a measure of dispersion around the true diversity, λ. Two times the standard deviation on either side of the measured value should roughly include 95% of all of the expected distribution of the sample mean and is thus an approximate measure for the confidence with which various diversity indices can be estimated. For highly diverse populations or typing techniques with extreme abilities to discriminate genotypes, the number of classes (genotypes) will increase with increasing sample size. Thus, the chance that two randomly sampled isolates differ will increase, i.e., the index of diversity becomes dependent on the sample size. Therefore, it is important to compare samples of roughly the same size as long as the number of classes depends on the sample size.

Using this algorithm, we concluded that the observed difference of the genetic diversity with nonoverlapping CIs among community and hospital carriage strains of *S. aureus* reflect two truly distinct population structures shaped by differential ecological constraints. Likewise, we have measured the discriminatory capacity of RAPD and macrorestriction analysis with sufficient precision and can be confident that the inability of RAPD typing to achieve the same degree of discrimination as macrorestriction analysis is an inherent property of these methods. We believe that an estimation of confidence intervals when calculating the index of diversity greatly aids the comparison of genetic diversity in different environments, as well as the ability to objectively address the discriminatory potential of diverse typing systems.

TABLE 3. Frequency of RAPD types among community carriage strains of *S. aureus*

RAPD type	No. of types	Frequency (%)
CR1	30	25.6
CR2	13	11.1
CR3	9	7.7
CR4	8	6.8
CR5	7	6.0
CR6	7	6.0
CR7	7	6.0
CR8	6	5.1
CR9	6	5.1
CR10	5	4.3
CR11	2	1.7
CR12	2	1.7
CR13	2	1.7
CR14 to CR26	1[a]	0.9[b]
Total	117	100

[a] One isolate each of these 13 types.
[b] The frequency for each of these 13 types.

REFERENCES

1. **Grundmann, H. J., K. J. Towner, L. Dijkshoorn, P. Gerner-Smidt, M. Maher, H. Seifert, and M. Vaneechoutte.** Multicenter study using standardized protocols and reagents for evaluation of reproducibility of PCR-based fingerprinting of *Acinetobacter* spp. J. Clin. Microbiol. **35:**3071–3307.
2. **Hunter, P. R., and M. A. Gaston.** 1988. Numerical index of the discriminatory ability of typing systems: an application of Simpson's index of diversity. J. Clin. Microbiol. **26:**2465–2466.
3. **Li, W. H., and D. Graur.** 1991. Genetic polymorphisms, p. 35–40. *In* W. H. Li and D. Graur (ed.), Fundamentals of molecular evolution. Sinauer Associates, Sunderland, Mass.
4. **Shannon, C. E.** 1948. A mathematical theory of communication. Bell Syst. Tech. J. **27:**376–423; 623–656.
5. **Simpson, E. H.** 1949. Measurement of diversity. Nature **163:**688.
6. **Struelens, M. J., and the European Study Group on Epidemiological Markers of the European Society for Clinical Microbiology and Infectious Diseases.** 1996. Consensus guidelines for appropriate use and evaluation of microbial epidemiologic typing systems. Clin. Microbiol. Infect. **2:**2–11.
7. **Tambic, A., E. G. Power, R. M. Anthony, and G. L. French.** 1997. Analysis of an outbreak of non-phage-typeable methicillin-resistant *Staphylococcus aureus* by using a randomly amplified polymorphic DNA assay. J. Clin. Microbiol. **35:**3092–3097.
8. **Tenover, F. C., R. D. Arbeit, R. V. Goehring, P. A. Mickelsen, B. E. Murray, D. H. Persing, and B. Swaminathan.** 1995. Interpreting chromosomal DNA restriction patterns produced by pulsed-field gel electrophoresis: criteria for bacterial strain typing. J. Clin. Microbiol. **33:**2233–2239.