0273-2289/05/120/0029/$30.00

Efficiency Control in Large-Scale Genotyping Using Analysis of Variance

Geert T. Spijker,[1] Marcel Bruinenberg,[2] and Gerard J. te Meerman*,[1]

[1]Department of Medical Genetics, [2]Medical Biology, Department of Pathology and Laboratory Medicine, University of Groningen, A. Deusinglaan 4, 9713 AW Groningen, the Netherlands, E-mail: G.J.te.Meerman@med.rug.nl

Received September 29, 2003; Revised August 9, 2004; Accepted August 12, 2004

Abstract

The efficiency of the genotyping process is determined by many simultaneous factors. In actual genotyping, a production run is often preceded by small-scale experiments to find optimal conditions. We propose to use statistical analysis of production run data as well, to gain insight into factors important for the outcome of genotyping. As an example, we show that analysis of variance (ANOVA) applied to the first-pass results of a genetic study reveals important determinants of genotyping success. The largest factor limiting genotyping appeared to be interindividual variation among DNA samples, explaining 20% of the variance, and a smaller reaction volume, sizing failure, and differences among markers all explained ~10%. Other potentially important factors, such as sample position within the plate and reusing electrophoresis matrix, appeared to be of minor influence. About 55% of the total variance could be explained by systematic factors. These results show that ANOVA can provide valuable feedback to improve genotyping efficiency. We propose to adjust genotype production runs using principles of experimental design in order to maximize genotyping efficiency at little additional cost.

Index Entries: Laboratory procedure; polymerase chain reaction; genotyping; efficiency; analysis of variance.

Introduction

Huge amounts of genotypes are produced in the course of mapping genetic factors for complex diseases. Hundreds to thousands of people are

*Author to whom all correspondence and reprint requests should be addressed.

genotyped for tens to hundreds of different markers. Presently, this is mostly done in a semiautomated way using 384-well thermal cyclers, pipetting robots, and capillary sequencers capable of generating data for 96 samples and up to 20 genetic markers at the same time. With increased use of association-based techniques using repeat markers and/or single nucleotide polymorphisms, the amount of data to be generated is likely to increase even more *(1–3)*.

In this industrial type of process, it is important to keep first-pass genotyping success rates as high as possible, to limit workload and costs. Therefore, it is crucial to identify factors limiting genotyping success. An obvious approach is to design and carry out specific experiments to improve the efficiency of the genotyping process. Although this has proven to be useful and valuable research (e.g., see refs. *4–6*), it is expensive. In fact, it can be considered rather inefficient to generate data only to improve genotyping efficiency. An alternative and potentially more efficient approach would be to use the information inherently available in the results of the standard genotyping process, by statistical analysis regarding underlying quality-related issues. Generally, precise information is available about handling and routing of specimens, positions in plates, polymerase chain reaction (PCR) conditions, concentrations, and so on. These conditions are often quite systematic and *orthogonal* (i.e., perpendicular), like the factorial experiments that are routinely analyzed using analysis of variance (ANOVA) *(7,8)*. This raises the possibility that ANOVA could be applied to identify factors limiting genotyping efficiency.

The basis of ANOVA is estimation of the amount of variance in an outcome or response variable that is explained by explanatory variables. The data are split into relevant subgroups defined by the explanatory variable, and the amount of variance *within* these subgroups is compared with the amount of variance *between* these subgroups. Subgroups that have small variance within, and large variance between them, mark variables that have great impact on the outcome variable. The probability that this amount of variance is explained by chance, rather than by the grouping parameter, can be assessed by an *F*-test. The influence of explanatory variables can be assessed *quantitatively* by comparing the amount of variance that is explained. This comparison is mostly made by comparing the so-called sum of squares. The *quality* of the influence of the different levels of the variable can be assessed by inspection of the coefficient that is estimated for the model *(8)*.

The purpose of the present study was to investigate whether ANOVA can be applied to routine genotyping data, in order to identify factors limiting genotyping efficiency. The approach is applied to a data set to illustrate the value of the approach, although the exact results will probably be specific to our setting.

Materials and Methods

Genotyping

As material, we used first-pass genotyping results for 28 microsatellite markers in a study on cardiovascular disease using 1308 subjects possibly related, but without pedigree information. Markers were in 11 different genomic regions on nine different chromosomes. Primer sequences were obtained from The Genome Database (www.gdb.org) and NCBI (http://www.ncbi.nlm.nih.gov/genome/sts). Oligonucleotides were ordered with one primer 5' labeled with a fluorochrome FAM, HEX (Biolegio, Malden, the Netherlands), or NED (Applied Biosystems, Foster City, CA).

We describe the work flow in some detail, because this is helpful in understanding the statistical analysis presented. DNA was isolated elsewhere using a QIAamp® 96 DNA spin blood kit (Qiagen, Hilden, Germany). DNA and reaction mix were pipetted (in two phases) into 384-well plates using a Biomek® 2000 pipetting robot (Beckman Coulter, Allendale, NJ). PCR was done at volumes of either 5 or 10 µL. Each well contained 30 ng of DNA, 0.25 U of Taq DNA polymerase (Amersham Pharmacia Biotech, Uppsala, Sweden), 0.2 m*M* dNTP (Roche, Basel, Switzerland), 2.5 m*M* $MgCl_2$, 10 m*M* Tris-HCl (pH 9.0), 50 m*M* KCl (Amersham Pharmacia Biotech), and 0.25 µ*M* of each primer. Thermal cycling was done on a PTC-225® thermal cycler (MJ Research, Waltham, MA). Cycling started with denaturation at 96°C for 10 min, followed by 35 cycles each consisting of denaturation at 95°C for 30 s; annealing at 55, 57, 60, or 63°C for 30 s; and an extension at 72°C for 1 min. The last cycle ended with an extension at 72°C for 30 min. Reaction products were pooled into four predefined panels according to expected allele range and fluorescent label. A total of 2.3 µL of the pooled reaction products was mixed with 2.5 µL of milliQ water and 0.2 µL of ET-400R size standard (Amersham Pharmacia Biotech). The mixture was loaded onto 96-well capillary sequencers MegaBACE™ 1000 (Amersham Pharmacia Biotech). Run time was 65 min at 10 kV. Three different runs were done before new acrylamide matrix was injected into the capillaries. Scoring of the traces was done using Genetic Profiler 1.1 (Amersham Pharmacia Biotech), with visual control of all scored genotypes.

Statistical Analyses

ANOVA was conducted using S-PLUS® 2000 (release 3; Insightful, Seattle, WA). The data set was prepared using Microsoft® Excel 2000. The dependent variable was genotype success (1 = yes, 0 = no). Explanatory variables are presented in Table 1. The 1308 samples were at fixed positions within four different 384-well plates. Each marker was determined for all samples in a 384-well plate, making marker and plate orthogonal parameters. The 384-well plates were filled using eight-channel pipets, so the same channel of the pipet filled two adjacent pairs of rows in

Table 1
Description of Explanatory Variables

Variable	Description	No. of levels
Sample	1308 different subjects	1308
Row	8 rows in 96-well plate	8
Column	12 columns in 96-well plate	12
Well	Well position in 96-well plate	96
Quadrant	4 quadrants in a 384-well plate	4
96-Well plate	Plate number of 96-well plate	14
384-Well plate	Plate number of 384-well plate	4
Marker	The 28 different markers	28
Label	Type of fluorescent marker label (FAM, HEX, or NED)	3
Reaction volume	5 or 10 µL of reaction mix added	2
No. of injections	Electrophoresis matrix used three times	3
Sizing possible	Whether the program is able to localize sizer peaks	2

Table 2
Definition of Quadrants[a]

384-Well plate			
Row	Column	Example wells	Quadrant
Odd	Odd	A1, A3, C1, E1, etc.	1
Odd	Even	A2, A4, C2, E2, etc.	2
Even	Odd	B1, B3, D1, F1, etc.	3
Even	Even	B2, B4, D2, F2, etc.	4

[a]The layout that was used to distribute four different 96-well plates over a 384-well plate is described (384-well plates have 16 rows and 24 columns, and 96-well plates have 8 rows and 12 columns).

the 384-well plate. This made it useful to assess whether different rows had different chances of success. Row and column were mutually orthogonal, as well as orthogonal to 384-well plate number (because each plate had the same number of rows and columns), and orthogonal to marker (because each marker was determined for each plate). Row and column determined well position; thus, the amount of variation explained by well position could be statistically modeled as an interaction effect between "Row" and "Column." Four different 96-well plates were filled from one 384-well plate in fixed pattern, which we called "quadrants" (see Table 2 for a definition). Thus, 384-well plate number and quadrant were orthogonal, with 96-well plate number as the interaction between these two. Finally, markers could be grouped according to type of fluorescent label. This could be analyzed by nesting the parameter "Marker" in "Label."

Some explanatory variables, however, are not orthogonal to all other parameters. First, the software cannot always locate the sizer peaks within

Table 3
Multiway ANOVA[a]

Explanatory variable	df	Range of sum of squares	Variance explained (%)
Sizing possible	1	505.1–505.1	7.2
Reaction volume	1	694.9–702.3	10
No. of injections	2	1.752–11.10	<0.1 ($p = 0.002$)
384-Well plate	3	85.48–88.43	1.2
Quadrant	3	21.14–26.95	0.3
96-Well plate	7*	140.2–150.0	2.1
Row	7	97.42–97.48	1.4
Column	11	40.74–40.75	0.58
Well	77*	301.4–301.4	4.3
Label	2	140.7–140.9	2.0
Marker nested in label	25*	378.0–379.1	5.4
Sample	1199*	1418–1418	20.3
Residual	34,855	3132–3132	44.9

[a]ANOVA using multiple explanatory variables was repeated several times; the order of nonorthogonal variables was varied. Reported sum of squares are minimum and maximum observed values. All p values are $<1 \times 10^{-7}$ unless otherwise specified. Asterisks mark factors with a reduced number of df compared to the number of levels, owing to correlation between explanatory variables; *see* Table 4.

Table 4
Nested Parameters

Variable	Already estimated df	Remaining df
28 Markers	2 for Label	25
96 Wells	7 for Row 11 for Column	77
14, 96-Well plates	3 for 384-Well plate 3 for Quadrant	7
1308 Samples	3 for 384-Well plate 3 for Quadrant 7 for 96-Well plate 77 for Well	1199

a trace. When this is the case, no single marker genotype can be determined for this sample-panel combination, even when clear marker peaks are visible. This problem is mostly seen at some capillaries that have supposedly worn out, or when, by chance, insufficient sizing mix is added to that well. For this reason, the data were analyzed conditional on sizing success. Second, the acrylamide electrophoresis matrix was reused three times before new matrix was injected into the capillaries. This causes correlation between markers. Third, the first part of the reactions was done using a reaction volume of 5 instead of 10 µL. For factors that are not orthogonal, the amount of variance cannot always be separated from related factors. S-PLUS

handles the parameters in hierarchical order. Therefore, ANOVA was repeated, using different formulas, shuffling the nonorthogonal parameters relative to the remaining parameters. Reported for each parameter were minimal and maximal sum of squares.

Regarding the number of degrees of freedom (df), as can be seen from Table 3, the df of a parameter is not always the number of levels minus 1. This is owing to the fact that these parameters are nested. For example, the markers fall in either of three categories, according to "Label." Therefore, after fitting "Label" (2 df), only 26 coefficients remain to be estimated, at the cost of 25 df. This is illustrated in Table 4.

Results

The results of the ANOVA are shown in Table 3. All results are highly significant, with p values below 1×10^{-7} unless otherwise specified. The emphasis is therefore on quantitative evaluation of the results.

Traces in which the scoring software cannot locate enough peaks of the sizer standard cannot yield any scored genotypes. This explained 7.2% of the variance in scored genotypes. The reaction volume explained 10% of the variation in scored genotypes. The smaller volume did not work as well as the reaction volume of 10 µL. Reusing the electrophoresis matrix explained at most 0.2% of the total variance, which shows that reuse of the matrix had a negligible effect.

Of the remaining parameters, "Sample" explained the most variance: >20%. The 96-well plate number explained much more variance (2.1%) than the combination of 384-well plate number (1.2%) and quadrant (0.3%). However, the four 96-well plates within a 384-well plate were subjected to PCR at the same time. This suggests that differences associated or confounded with thermal cyclers have less influence on genotyping efficiency than the other steps in the procedure.

A similar reasoning can be used when comparing the effects of "Row" and "Column." "Row" explained more variance than "Column" (1.4 vs 0.58%). This suggests some systematic effect of the eight-channel pipet, which could deliver different amounts of fluid to different rows. The influence on whether the genotype can be scored is, however, small; it explained only 1.4% of the variance. "Well" explained 4.3% of the variance but could not be analyzed independently from "Sample"; as the same samples were always in the same position on the plates, the amount of variance explained by "Well" is an overestimation, because it is partly owing to differences among samples.

The parameter "Marker" explained 7.4%. When comparing the efficiency of markers grouped according to the fluorescent label used (FAM, HEX, or NED), this explained 2.0% of the total variance. The fraction of scored genotypes was higher for FAM- and HEX-labeled markers than for NED-labeled markers. Forty-five percent of the variance remains unexplained by these factors.

Discussion

We demonstrated how routine genotyping data could be used to identify some basic factors limiting genotyping efficiency. Our study shows that owing to the regularities of the semiautomated genotyping process, it can be viewed as an experiment designed to test the influence of potentially important factors. ANOVA enabled us to assess these factors quantitatively, some showing considerable influence, whereas others were negligible. In principle, every variable that is present in a sufficiently systematic way can be analyzed for its influence on the genotyping success.

We have shown for the present data set that differences among DNA samples is the largest identifiable factor limiting genotyping efficiency, explaining about 20% of the variation. This means that theoretically most efficiency can be gained by optimizing this factor. In future experiments, slight variations in the amount of DNA per reaction could be introduced to determine whether this variability is owing to differences in DNA concentration, rather than DNA quality. Other important factors were sizing problems (probably attributable to insufficient sizer DNA or worn-out capillaries), differences among markers, and using a reaction volume of 5 instead of 10 μL, all explaining <10%. An important result is that several potentially important factors have only a small influence: the type of marker label used, reusing the electrophoresis matrix, and position of the sample within the plate. There are several processes that could cause the genotyping efficiency to depend on position of the sample within the plate, such as, when not all channels of the eight-channel pipet deliver the same amount of fluid, systematic evaporation of fluid during the thermal cycling, or irregular temperature distribution within the thermal cycler. However, because only 4.3% of the variance is explained by position within the PCR plates, none of these factors was of major influence in the present study.

The importance of our data is, however, not in the actual results, because it is likely that these results are specific for our setting. The novelty is, rather, in the realization that techniques such as ANOVA, applied to data that are already present, can help to assess genotyping efficiency and to build confidence that specific factors have little influence, thus avoiding time-consuming optimization procedures. In principle, every variable that is present in a sufficiently systematic way can be analyzed for its influence on the genotyping success. Owing to the abundant amount of genotyping data, the sensitivity for small effects is high, even using a crude measure of genotyping success or failure; a factor explaining only ~0.1% of the variance still had a highly significant p value (Table 3).

Now that the potential of ANOVA is shown, the information content could be increased even further by introducing small controlled and designed experimental variations into the genotyping process. Examples are varying the reaction concentration of one of the reagents, or varying the position of DNA samples within the PCR plates (although the latter may be difficult to attain in a standard work flow). Another application would

be in quality control: ANOVA could be implemented in an automated way, to track changes in variance. When the amount of variance explained by a factor increases much compared to earlier results, something could be wrong. For example, a large amount of failure in one row could indicate problems with a multichannel pipet, or an increase in the amount of sizing problems could indicate that the capillaries should be replaced.

In conclusion, we have shown that ANOVA can easily be applied to routine genotyping data. This provides valuable feedback on factors hindering genotyping and offers starting points for refined designs, in order to maximize genotyping efficiency at a minimal load of additional genotyping.

Acknowledgments

We thank the anonymous reviewer for detailed comments that assisted in improving this article. This research was supported by a grant from the Netherlands Organization of Scientific Research (NWO).

References

1 Risch, N. and Merikangas, K. (1996), *Science* **273,** 1516, 1517.
2 Risch, N. J. (2000), *Nature* **405,** 847–856.
3 Carlson, C. S., Eberle, M. A., Rieder, M. J., Smith, J. D., Kruglyak, L., and Nickerson, D. A. (2003), *Nat. Genet.* **33,** 518–521.
4 Crivellente, F. and McCord, B. R. (2002), *J. Capillary Electrophor.* **7,** 73–80.
5 Lahiri, D. K. and Schnabel, B. (1993), *Biochem. Genet.* **31,** 321–328.
6 Saunders, G. C., Dukes, J., Parkes, H. C., and Cornett, J. H. (2001), *Clin. Chem.* **47,** 47–55.
7. Dobson, A. J. (2000), *An Introduction to Generalized Linear Models*, 2nd ed., Chapman & Hall, London.
8. Altman, D. G. (1991), *Practical Statistics for Medical Research*, Chapman & Hall, London.