Numbers of Copy-Number Variations and False-Negative Rates Will Be Underestimated If We Do Not Account for the Dependence between Repeated Experiments

To the Editor: We read with interest the recent publication of Wong et al.[1] that uses six repeat experiments to provide estimates of copy-number variation (CNV) numbers and false-positive and false-negative rates in the absence of a "gold-standard" set of data. With acceptance of the obvious limitation that such an approach is not making inference about the true CNV population but only that subset that might be detected via this technology, this appears to be an ingenious idea (with echoes of capture-recapture schemes) and is itself worthy of replication.

From the observed values that they report (and reproduced in table 1), Wong et al.[1] estimate that there are 141 true CNVs (i.e., those 141 that were called in more than one experiment). This is based on the observation that, if these data were arising from independent Bernoulli/binomial processes, the probability of calling the same clone twice by chance would be very small. The authors acknowledge that they are underestimating the total number of CNVs, since some of the 340 clones called in only one of the six repeat experiments are likely to be true calls, but they accept 141 as a conservative (for their purpose) estimate of the true number of CNVs.

If one formally fits a statistical model to the vector of observed data, treating it as a mixture of observations from two binomial distributions (one arising from those clones that are truly CNVs and one from those that are not), then one has three parameters to estimate. We need to estimate the proportion of clones that represent true CNVs, from which we can later estimate *n,* the number of CNVs. We denote the probability of correctly calling a CNV within any single experiment as p (one minus the false-negative rate of Wong et al.[1]) and that of correctly ignoring a clone that is not a true CNV within any single experiment as q (one minus the false-positive rate).

The models were fitted using the WinBUGS[2] software package. One would anticipate that both proportions p and q would be near 1, and so beta prior distributions were assigned that reflected this. We presume that the proportion of clones that "are" CNVs is small (probably of the magnitude of 10^{-2}), and we assign a triangular distribution over the region 0–0.4. Convergence was quick, and comparison of prior and posterior distributions gave no cause for concern. Full details of the model and model fit are available as detailed at the authors' Web site.

The values we obtained from this model (given as me-

Table 1. Numbers Observed by Wong et al. and the Abilities of a Binomial and Generalized Binomial Model to Account for Them

Called in No. of Experiments	Observed	95% Credible Intervals	
		Binomial	Generalized Binomial
0	23,911	23,850–23,970	23,850–23,970
1	340	293–396	290–392
2	50	22–53	37–79
3	46	33–65	20–47
4	15	**25–53**	13–35
5	15	8–27	9–28
6	15	**0–7**	5–23

dian, with 95% credible interval in parentheses) suggest that Wong et al.'s estimates[1] were very good. Their estimate for p was 0.547, whereas we found it to be 0.514 (0.472–0.554). Their estimate for q was 0.998, and we found it to be 0.9977 (0.9974–0.9980); their estimate of n was 141, whereas ours was 154 (120–192). When the fact that they had deliberately slightly underestimated n is considered, it seems that their simplified calculation scheme came at little or no cost.

However, one of the advantages of our fitting the full model is that we can estimate the number of calls that should be seen within each of the categories (those called for all six experiments, those called for five of the six experiments, etc.). Credible intervals from the binomial model (table 1) reveal that there are discrepancies between the expected and observed numbers in the tail of the distribution. One explanation for this is that the calls between experiments are not independent; they are, after all, replicates. Thus, a greater proportion of clones called by a few experiments will be called by all experiments than can be accounted for under a binomial model.

One's first instinct when accounting for this dependency might be to place beta distributions on the parameters p and q. We do not take this approach, for three reasons. First, there are computational issues with fitting a model of such complexity to seven observed numbers. Second, such a model suggests a specific form of dependency, and we do not wish to make that restriction. The dependency would be interpreted as being driven by varying effect sizes; a CNV representing several gains would be more likely to be called by each of the experiments than would one representing little gain. However, even if there were both uniform effect sizes for each CNV and uniform levels of evidence, one might wish to account for a dependence arising from the replicate nature of the experiments. Finally, and more trivially, we recognize that it is difficult to marry the concept of a false-negative rate with that of modeling CNVs as coming from some continuum rather than simply being or not being.

Therefore, we chose to use a mixture of generalized binomial models—in particular, the multiplicative generalization presented by Altham[3] that includes one extra parameter θ that both models and provides a diagnostic for the dependence of the experiments. If $\theta < 1$, a positive dependence between experiments is indicated; if $\theta = 1$, then the experiments are modeled as being independent; and if $\theta > 1$, then we are in the unlikely situation in which the responses of different experiments are negatively associated.

The advantages of this model are that it is suitable for use when only the summary data are available (such as in this case). Moreover, it is particularly easy to deal with situations such as this, where every clone features in the same number of experiments. Finally, it reduces to the binomial model when only one experiment is performed, meaning that $1 - q$ and $1 - p$ still represent the false-positive and false-negative rates, respectively, for a single experiment and are related to those rates as the number of experiments increases. Also, as noted, it reduces to the binomial model when $\theta = 1$, providing a simple test for the hypothesis of independence.

We have chosen to fit a mixture of two generalized binomial models with a common θ parameter, but arguments could also be made for separate θ parameters or indeed for a mixture of a generalized binomial model for the CNV clones and a standard binomial model for the non-CNV clones. These alternatives lead to no essential differences in the results, except that the estimate of q tends to be a little greater. The prior distribution given to θ was log-normal and reasonably symmetric about 1, so that we might interpret departure from the value of 1 as a test of the independence of the experiments. Fitting our mixture of two generalized binomial distributions, we find that the 95% credible interval for θ is 0.61–0.78, thus showing strong evidence of dependence between responses to the repeated experiments and further suggesting that the binomial model is not adequate.

By accounting for the dependence between experiments, the model provides a better fit to the observations, in terms of the values in each contingency cell (table 1), with regard to both the credible intervals and the χ^2 statistic for the goodness of fit (8.9 as opposed to 64.7). The Deviance Information Criterion, which compensates for the extra complexity of the generalized model, is reduced to 47 from a value of 87 for the binomial model.

However, our estimate of p for a single experiment, which takes into account the dependence, is now merely 0.394 (0.340–0.449). This is to be expected if we believe the responses to be dependent. The estimate for q is less dramatically altered. One consequence of having a lower value of p than previously thought (or, in the language of Wong et al.,[1] a higher false-negative rate) is that we are likely to be missing more CNVs, so our estimate of the number of CNVs increases to 399 (212–1,139). This is 2.5 times the estimate that arises from the model that assumed independence.

CNVs are, of course, heterogeneous, and, as we have stressed, there undoubtedly exist classes of CNVs that we could not detect with this technology. Therefore, we must assume that the true number of CNVs (for as much as the concept is sensible) is greater still. It is also the case that any sizable heterogeneity between the repeated experiments (in terms of levels of noise, etc.) would impinge on the interpretation of our results; however, we doubt that heterogeneity great enough to change our overall conclusions would have been tolerated in any laboratory.

In conclusion, whereas experimental validation of course remains the ideal when practicable, we applaud the concept of replicated experiments in attempting to estimate such values. However, we caution that failing to take the dependence into account can lead to underestimation of

the false-positive and false-negative rates and, perhaps more crucially, the of true number of CNVs to be identified.

ANDY G. LYNCH, JOHN C. MARIONI, AND SIMON TAVARÉ

Acknowledgments

We are supported by grants from Cancer Research UK. We thank two anonymous referees and our colleagues for useful discussions and comments.

Web Resource

The URL for data presented herein is as follows:

Authors' Web site, http://www.damtp.cam.ac.uk/user/jcm68/AJHG.html (for download of the WinBUGS code for the models and for further information about the models)

References

1. Wong KK, deLeeuw RJ, Dosanjh NS, Kimm LR, Cheng Z, Horsman DE, MacAulay C, Ng RT, Brown CJ, Eichler EE, et al (2007) A comprehensive analysis of common copy-number variations in the human genome. Am J Hum Genet 80:91–104
2. Spiegelhalter DJ, Thomas A, Best NG, Lunn D (2003) WinBUGS user manual, version 1.4.1. Medical Research Council Biostatistics Unit, Cambridge, United Kingdom
3. Altham PME (1978) Two generalizations of the binomial distribution. Appl Stat 27:162–167

From the Computational Biology Group, Department of Oncology (A.G.L.; J.C.M.; S.T.), and Department of Applied Mathematics and Theoretical Physics (J.C.M.; S.T.), University of Cambridge, Cambridge, United Kingdom

Address for correspondence and reprints: Dr. Andy Lynch, Department of Oncology, University of Cambridge, Cancer Research UK Cambridge Research Institute, Robinson Way, Cambridge CB2 0RE, United Kingdom. E-mail: andy.lynch@cancer.org.uk

Am. J. Hum. Genet. 2007;81:414.

0002-9297/2007/8102-0026$15.00
DOI: 10.1086/521416