LETTERS TO THE EDITOR

Copy-Number Variations and Human Disease

To the Editor: The comprehensive mapping of genomic copy-number variations (CNVs) should allow for those variants to be studied for their correlation with disease phenotypes.[1] One line of evidence that was advanced to support potential implications of CNVs for human disease was the overlap of CNVs with chromosomal loci harboring genes known to very often cause well-characterized monogenic illnesses from the OMIM database.[1] But a general concern arises from the list of CNVs that overlap rare disease genes reported in table 5 of the mapping study by Wong et al.[1] For instance, the *BSCL2* gene fell within a CNV region that had a frequency of 3 in 105 subjects. The gene was noted by the authors to be causative for spinal muscular atrophy, distal, type V (MIM 600794),[1] and indeed two missense mutations in *BSCL2* (MIM 606158.0013 and .0014) have been reported for that phenotype. However, 13 other missense mutations in that gene cause Berardinelli-Seip congenital generalized lipodystrophy (BSCL), an extremely rare autosomal recessive disorder affecting ~1 person in 10 million.[2] Thus, the observed frequency of ~3% *BSCL2*-CNV heterozygotes seems high, given the low prevalence of BSCL as ascertained clinically in the general population. Conservative assumption of a *BSCL2*-CNV frequency of 1% would predict a homozygote frequency of a major CNV rearrangement of this region of 1 person in 40,000, a frequency that is much higher than the observed prevalence of BSCL—or of any spinal muscular atrophy subtype, for that matter—in the general population. The same type of disparity between predicted and observed prevalence appears to hold true for several other genes causing very rare homozygous diseases that lie within CNV loci,[1] including *SMA3* (MIM 253400) and *SMA4* (MIM 271150), which cause spinal muscular atrophy subtypes (CNVs seen in 60 of 105 samples), and *GCK* (MIM 138979), which causes neonatal-onset diabetes (CNVs seen in 10 of 105 samples). Obviously, CNVs include both duplications and deletions, and homozygosity for a duplication-type CNV would not necessarily imply the same pathogenic consequence as a deletion-type CNV for an OMIM gene. Thus, calculations of predicted disease frequency should be based on homozygosity for the subset of deletion-type CNVs. But, because deletions would still be expected to represent a sizable subset of CNVs for at least a sizable subset of OMIM genes, the predicted disease rate would still be much higher than the observed frequency of the autosomal recessive disease phenotype in the general population.

There may be some valid biological reasons for the apparent disparities between the observed frequency of CNV heterozygotes and the reported frequencies of the related rare OMIM recessive diseases, including (1) inaccurate disease-frequency estimates in published reports that underestimate the actual frequency of the phenotype in the general population (perhaps subtle or later-onset forms of the phenotypes might be more prevalent in the general population than has been generally recognized), (2) incompatibility of homozygosity for certain completely deleted genes with fetal viability, (3) derivation of CNV-frequency estimates in nonrepresentative "normal control" samples, and (4) rescue of the lost or altered function in homozygotes for a CNV by another gene product. Alternatively, there may be systematic technical reasons for potential disparities leading to overestimation of some CNV frequencies. It would thus be important to validate those CNVs with use of alternative technologies, such as quantitative PCR,[3] and to expand the samples of normal control subjects to determine whether homozygotes for those CNVs exist among "healthy" controls. In addition, studies within families, both healthy and diseased, might help to clarify the potential pathogenicity of some of these CNVs. The findings emphasize the fact that the excitement over the biological reality of CNVs within clinical and research samples should be tempered pending the development of standards and independent wide-scale replications, possibly with use of a variety of detection methods.

Robert A. Hegele

Acknowledgments

R.A.H. is supported by operating grants from the Canadian Institutes of Health Research (MT14030), the Heart and Stroke Foundation of Ontario, Genome Canada through the Ontario Genomics Institute, a Career Investigator award from the Heart and Stroke Foundation of Ontario, the Edith Schulich Vinet Canada Research Chair (Tier I) in Human Genetics, and the Jacob J. Wolfe Distinguished Medical Research Chair.

Web Resource

The URL for data presented herein is as follows:

Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for spinal muscular atrophy, distal, type V; *BSCL2* mutations; *SMA3*; *SMA4*; and *GCK*)

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. Garg A (2004) Acquired and inherited lipodystrophies. N Engl J Med 350:1220–1234
3. Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. Nat Rev Genet 7:85–97

From the Schulich School of Medicine and Dentistry, University of Western Ontario, and Vascular Biology Research Group, Robarts Research Institute, London, Ontario

Address for correspondence and reprints: Dr. Robert A. Hegele, Blackburn Cardiovascular Genetics Laboratory, Robarts Research Institute, 406-100 Perth Drive, London, ON, Canada N6A 5K8. E-mail: hegele@robarts.ca

Am. J. Hum. Genet. 2007;81:414. © 2007 by The American Society of Human Genetics. All rights reserved.
0002-9297/2007/8102-0022$15.00
DOI: 10.1086/521416

Reply to Dr. Robert A. Hegele

To the Editor: Dr. Hegele called attention to the disparities between the observed frequencies of copy-number variations (CNVs) and the reported frequencies of some OMIM diseases.[1(in this issue)] We share Dr. Hegele's concern and emphasize the importance of interpreting CNV data with caution, with respect to correlations with genes or phenotypes.

A number of issues must be considered when reported CNV data are used. One is that the exact boundaries of CNVs are rarely known; thus, direct correlations with genes can be challenging. The *BSCL2* (MIM 606158) gene, noted by Dr. Hegele, overlaps the boundary of a BAC clone that we reported as variable in copy number.[2] In this case, approximately one-third of the *BSCL2* gene overlaps the end of a CNV clone (this can be viewed using our CNV custom track, now publicly available at the UCSC Genome Browser). However, the gene may not be part of the CNV. It is therefore necessary to confirm, using alternative validation technologies, whether a gene is actually affected by a particular CNV. Depending on the detection sensitivities and resolutions of the array platforms, the boundaries of the reported CNVs will be within tens to hundreds of kilobases from the actual boundaries. When a combined data set such as the Database of Genomic Variants[3] is used, it is very important to be mindful of the strengths and weaknesses of the platforms used to derive the data, such as resolution, detection sensitivity, and false-positive and false-negative rates.

In our study, we reported genes as CNV associated if any part of the gene overlapped a BAC clone that we measured to be variable in copy number.[2] There is a need for standardization in the reporting of which genes are potentially associated with CNVs. Furthermore, it is biologically difficult to know when a CNV will influence the expression of a gene, given that position effects are known to influence the expression of genes across hundreds of kilobases.[4]

An additional confounding issue when interpreting CNV data obtained from array comparative genomic hybridization is the comparative nature of the technique. Gains and losses are called in relation to a reference DNA, which will vary by study and are often simplistically interpreted as a single-copy change from diploid. In fact, the exact copy number is often not known for either the sample or the reference. In the case of the CNV associated with the *BSCL2* gene, we detected two samples that showed a gain relative to the reference and one sample that showed a loss in copy number relative to the reference, demonstrating the complexity of changes occurring in the genome. A further consideration is that, in some cases, the baseline copy number could be greater than two. Genes in such regions may be particularly resistant to disease-causing mutations because of functional redundancy.

In summary, CNV data provide valuable information for studies involving human genetics, and the abundance of CNVs means that they are likely to include or influence many genes; however, the data need to be used with caution.

Kendy K. Wong, Ronald J. deLeeuw,
Carolyn J. Brown, and Wan L. Lam

Web Resources

The URLs for data presented herein are as follows:

Database of Genomic Variants, http://projects.tcag.ca/variation/
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for *BSCL2*)
UCSC Genome Browser, http://genome.ucsc.edu/goldenPath/customTracks/custTracks.html

References

1. Hegele RA (2007) Copy-number variations and human disease. Am J Hum Genet 81:414–415 (in this issue)
2. 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
3. Zhang J, Feuk L, Duggan GE, Khaja R, Scherer SW (2006) Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. Cytogenet Genome Res 115:205–214
4. Kleinjan DJ, van Heyningen V (1998) Position effect in human genetic disease. Hum Mol Genet 7:1611–1618

From the Department of Cancer Genetics and Developmental Biology, British Columbia Cancer Research Centre (K.K.W.; R.J.d.; W.L.L.), and the Department of Medical Genetics, University of British Columbia (C.J.B.), Vancouver
Address for correspondence and reprints: Dr. Kendy K. Wong, 675 W. 10th Avenue, Vancouver, BC, V5Z 1L3 Canada. E-mail: kwong@bccrc.ca

Am. J. Hum. Genet. 2007;81:414. © 2007 by The American Society of Human Genetics. All rights reserved.
0002-9297/2007/8102-0023$15.00
DOI: 10.1086/521416

The *TAF1/DYT3* Multiple Transcript System in X-Linked Dystonia-Parkinsonism

To the Editor: The X-linked dystonia-parkinsonism syndrome (XDP or DYT3 [MIM #314250]) is a severe adult-onset movement disorder that originated by founder effect in the Philippine island of Panay.[1] The disease gene was identified in 2003 and was described as a "multiple transcript system." It is composed of several of the 38 known *TAF1* (TATA-box binding protein–associated factor 1) exons and an additional 5 of the then-unknown exons that lie 3′ of *TAF1* exon 38.[2] The latter five exons can either be spliced to known *TAF1* exons (variants 1 and 2) or be transcribed separately (variants 3 and 4). Five disease-specific single-nucleotide changes (DSCs) and a small deletion were detected within this transcript system. One of the DSCs (DSC3) is located in a transcribed exon. These findings have now been confirmed by Makino and coworkers in the March issue of the *Journal*.[3] Whereas, in the original study, the *DYT3* critical region was sequenced by PCR in a patient, Makino et al.[3] resequenced this region in BAC clones constructed from a patient's DNA. The resequencing confirmed the DSCs described elsewhere[2] and detected a previously unrecognized retrotransposon (SVA [SINE, VNTR, and *Alu*] element) in intron 32 of *TAF1* in close proximity to DSC10. Makino et al.[3] also confirmed the various splice variants of *TAF1* that were found earlier.[2] Furthermore, the study by Makino et al.[3] validates our use of several DSCs in the routine molecular genetic diagnosis of XDP.[4]

It is currently not known whether or to what extent the DSCs in *TAF1/DYT3* are involved in the disease process. Makino et al.[3] implicate the SVA retrotransposon in intron 32 of *TAF1* in the pathology of XDP. They present several findings supporting an important role for SVA in XDP. In particular, they provide evidence that the SVA affects expression of a transcript splice variant described elsewhere[2] that includes exon 34′ of *TAF1*. However, the data are not entirely convincing.

1. The article by Makino et al.[3] implies that there is only one splice variant of *TAF1* that includes exon 34′. This is not accurate. There are other splice variants of *TAF1* that also include exon 34′—for example, splice variants including exons 34′ and 32′ and splice variants including exons 34′, 3, and 4.[2]
2. Figure 5*a* in the work of Makino et al.[3] shows dramatic reduction of expression of an exon 34′–containing transcript but also demonstrates reduced expression of the common form of *TAF1* in a patient's caudate nucleus. Although antibodies were directed against TAF1 polypeptides (and not specifically against the exon 34′ isoform), their figure 5*f* implies complete absence of TAF1 in the patient's caudate. This cannot be explained by gliosis alone, since calcineurin antibodies definitely identified neurons in the patient's caudate (see their figure 5*f*).
3. Although hypermethylation was shown at CpG sites of SVA, a correlation between this epigenetic modification and the postulated specifically reduced expression of the exon 34′ transcript was not shown.
4. When postmortem brain is used for the quantitative ascertainment of gene expression, a high degree of variation has to be kept in mind. Apart from biological reasons, different postmortem times, storage conditions, etc. account for this variation. This might explain why transcript variant 34′ of *TAF1* is specifically reduced in some experiments but the common form of *TAF1* is also affected in others (their fig. 5*a* and 5*f*).
5. Makino et al.[3(p402)] indicate that the decrease in expression of the exon 34′ transcript is "the cause rather than the result of neuronal loss in the caudate nucleus…." This argument is not entirely convincing, since this transcript is also reduced in cortex and nucleus accumbens that do not show major neuronal loss. Makino et al.[3] do not provide evidence of neuronal subtype-specific expression and function of the exon 34′ transcript that might explain the discrepancy.

Obviously, the molecular pathological mechanism in XDP remains unknown. The involvement of one or several of the described DSCs, either alone or in concert with the SVA retrotransposon, certainly cannot be ruled out. Here, a function of DSC3 is intriguing, since it is located in an exon that can be part of all major splice variants of the *TAF1/DYT3* transcript system. However, intronic SNPs can also affect gene expression, as was recently shown with the *SORL1* susceptibility gene for late-onset Alzheimer disease.[5]

Several other aspects of the article need further clarification. In figure 3, patients are shown carrying the "disease-specific" 6.1-kb SVA fragment, but other patients (right panel of fig. 3) show the "wild-type" fragment. Can the SVA fragment occur in healthy persons as well, or has a sample mix-up occurred? Makino et al.[3] claim that exon 2, described elsewhere[2] (3′ of *TAF1* exon 38), is derived from *ING2*. This is not the case, since the *ING2* pseudogene overlaps with exon 2 on the opposite strand. Exon 38 of *TAF1* is skipped when further 3′ exons are used in a transcript. This was shown in cDNAs isolated from a brain cDNA bank and in RT-PCR experiments.[2] Makino et al.,[3] however, report the presence of this exon in these alternative transcripts. Provided that this is no RT-PCR artifact, this does not disprove previous findings of the absence of exon 38 in some splice variants that include additional 3′ exons.

In conclusion, many issues remain unresolved as to both the normal function of *TAF1* variants in various tissues and the role of disease-specific changes in the *TAF1/DYT3* multiple transcript system in patients with XDP.

Ulrich Müller, Thilo Herzfeld, and Dagmar Nolte

Acknowledgments

Our work was supported by Deutsche Forschungsgemeinschaft grant Mu668/12.

Web Resource

The URL for data presented herein is as follows:

Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for XDP or DYT3)

References

1. Kupke KG, Lee LV, Viterbo GH, Arancillo J, Donlon T, Müller U (1990) X-linked recessive torsion dystonia in the Philippines. Am J Med Genet 36:237–242
2. Nolte D, Niemann S, Müller U (2003) Specific sequence changes in multiple transcript system DYT3 are associated with X-linked dystonia parkinsonism. Proc Natl Acad Sci USA 100: 10347–10352
3. Makino S, Kaji R, Ando S, Tomizawa M, Yasuno K, Goto S, Matsumoto S, Tabuena D, Maranon E, Dantes M, et al (2007) Reduced neuron-specific expression of the *TAF1* gene is associated with X-linked dystonia-parkinsonism. Am J Hum Genet 80:393–406
4. Evidente VGH (2005) X-linked dystonia-parkinsonism syndrome. Gene Reviews (http://www.genetests.org/servlet/access?qry=evidente&submit=Search&id=8888891&prg=j&db=genestar&site=gc&fcn=author&key=2ufeYApfIefX3) (accessed June 8, 2007)
5. Rogaeva E, Meng Y, Lee JH, Gu Y, Kawarai T, Zou F, Katayama T, Baldwin CT, Cheng R, Hasegawa H, et al (2007) The neuronal sortilin-related receptor SORL1 is genetically associated with Alzheimer disease. Nat Genet 39:168–177

From the Institut für Humangenetik, Justus-Liebig Universität, Giessen, Germany

Address for correspondence and reprints: Dr. Ulrich Müller, Institut für Humangenetik, Justus-Liebig Universität, Schlangenzahl 14, 35392 Giessen, Germany. E-mail: ulrich.mueller@humangenetik.med.uni-giessen.de

Am. J. Hum. Genet. 2007;81:414. © 2007 by The American Society of Human Genetics. All rights reserved.
0002-9297/2007/8102-0024$15.00
DOI: 10.1086/521416

TAF1 as the Most Plausible Disease Gene for XDP/DYT3

To the Editor: We address the concerns of Dr. Müller and his colleagues[1(in this issue)] regarding our recent article in the *Journal*.[2] Previously, Müller et al. reported that five disease-specific single-nucleotide changes (DSCs) in a “multiple transcript system” (*MTS*) were associated with XDP (MIM #314250).[3] Our article was the first to report *TAF1* (MIM *313650) as the most plausible disease gene,[2] and it is regrettable that no one has succeeded in confirming *MTS* transcripts by any standard technologies, such as northern blot, to provide the information on length and abundance of the transcripts. Therefore, the term “*TAF1/DYT3* multiple transcript system” is ambiguous and misleading. *TAF1* and *MTS* are different genes that have distinct functions, although some alternative splicing exons of *TAF1* are shared with some *MTS* transcripts. In particular, we claim that the neuron-specific isoform of *TAF1* was discovered by us.[2]

The following are the responses to the points raised by Müller et al.[1]

1. As listed in our table 5, the splicing variant of *MTS* that includes exons 32′ and 34′ was never detected by our TaqMan assay.[2] Our results from long RT-PCR analysis consistently show that exons 3 and 4 may be just an additional part of the 3′ UTR of *TAF1,* rather than part of *MTS* (fig. 4*a*).[2]

2. As described in the figure legend, our figure 5*f* shows weak immunoreactivity of *TAF1* in neurons in the patient’s caudate and glial cells in both tissues but never implies complete absence of *TAF1* in the patient’s caudate.[2]

3. The possible mechanism between the epigenetic modification and the reduced neuron-specific expression of the *TAF1,* including the neuron-specific isoform that contains exon 34′, which we discovered, were described in the “Discussion” section of our article.[2]

4. The postmortem brain was immediately frozen after the patient’s death, so the point raised about our postmortem sample seems to be excessively speculative.[2] However, we are ready for a confirmatory examination of other frozen specimens, to give credence to our findings.

5. We hypothesize that neuronal death depends on difference of local conditions in various brain tissues—for example, free radical production, calcium flux, local temperature, and dominant neurotransmitters, as illustrated for dopamine receptor D2 (*DRD2* [MIM *126450]) in figure 5*c*.[2]

For the family shown in the upper right of figure 3,[2] the picture scanned from the x-ray film was accidentally misaligned when the margin was trimmed. Our original x-ray film exactly shows the concordant signal with all other patients, as described in our article.[2] For the disease-specific SVA retrotransposon insertion, there is a striking discrepancy between the results from of our work[2] and the work reported by Nolte et al.[3] Nolte and colleagues stated that they sequenced 260 kb of the critical interval in a patient with XDP[3]; however, they did not find the insertion. We studied[2] the patients with XDP who had the same STR/DSCs haplotype as that of the patients examined by Nolte and colleagues.[3] It might be argued that the element was inserted only in the families we studied who carried no etiological significance in XDP. Nolte and colleagues claimed that they determined the genomic sequence beyond 260 kb by “cycle sequencing of overlapping PCR products,”[3(p10347)] without presenting any detailed information about the experimental conditions and without submitting their genomic sequence to any public database. Such a mutation search analysis requires the *completeness* of the sequence to justify its conclusion in the published work,[3] so readers such as us might interpret the sequence determined by Nolte and colleagues to be continuous and complete. It is, however, commonly believed that it is extremely difficult or almost impossible to determine such a large genomic sequence with use of the PCR-based sequence method, because of the complexity of the human genome and the well-known technical limitations, especially for long-range PCR. We ask Dr. Nolte and colleagues to submit their completed sequence to a public database, to strengthen the quality of their findings and to enable any researcher interested in this field to compare with our complete sequence (DNA Database of Japan accession number AB191243).

We hope that Dr. Müller and colleagues will succeed in determining the complete and accurate structure and abundance of *MTS* transcripts by means of various standard experiments, including northern blot, probe-hybridization screening of unbiased cDNA libraries, and TaqMan assay, and then present a hypothesis about what leads to the loss of striatal neurons by DSC3 on the *MTS* gene.

Gen Tamiya, Satoshi Makino, and Ryuji Kaji

Web Resources

The accession number and URLs for data presented herein are as follows:

DNA Database of Japan, http://www.ddbj.nig.ac.jp/Welcome-e.html (for the complete genomic sequence of the *DYT3* region [accession number AB191243])

Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for XDP, *TAF1,* and *DRD2*)

References

1. Müller U, Herzfeld T, Nolte D (2007) The *TAF1/DYT3* multiple transcript system in X-linked dystonia-parkinsonism. Am J Hum Genet 81:415–417 (in this issue)
2. Makino S, Kaji R, Ando S, Tomizawa M, Yasuno K, Goto S, Matsumoto S, Tabuena D, Maranon E, Dantes M, et al (2007) Reduced neuron-specific expression of the *TAF1* gene is associated with X-linked dystonia-parkinsonism. Am J Hum Genet 80:393–406
3. Nolte D, Niemann S, Müller U (2003) Specific sequence changes in multiple transcript system *DYT3* are associated with X-linked dystonia parkinsonism. Proc Natl Acad Sci USA 100: 10347–10352

From the Department of Neurology and Neuroscience, University of Tokushima Graduate School of Medicine, Tokushima, Japan

Address for correspondence and reprints: Dr. Gen Tamiya, Department of Neurology and Neuroscience, University of Tokushima Graduate School of Medicine, Tokushima 770-8503, Japan. E-mail: gtamiya@genetix-h.com

Am. J. Hum. Genet. 2007;81:414. © 2007 by The American Society of Human Genetics. All rights reserved.
0002-9297/2007/8102-0025$15.00
DOI: 10.1086/521416

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	95% Credible Intervals: 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

Note.—Discrepancies are shown in bold.

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. © 2007 by The American Society of Human Genetics. All rights reserved.
0002-9297/2007/8102-0026$15.00
DOI: 10.1086/521416

Reply to Lynch et al.

To the Editor: The letter by Lynch et al.[1(in this issue)] described the application of more-robust statistical modeling for the determination of false-negative and false-positive rates in our copy-number variation (CNV) study. Their conclusion is that the inclusion of dependence in the model increases the false-negative rate while leaving the false-positive rate unaltered.

These findings raised key questions as to what methodology should be employed to quantify false-positive and false-negative rates in CNV data. To determine false-detection rates, are single experiments repeated multiple times preferable to single replication of many experiments or, alternatively, use of self- versus self-hybridization experiments? Within currently published CNV studies,[2–4] which were derived from different array platforms, these methods have been employed individually or in combinations in some studies, whereas others employed completely different methods of quality assessment.[5] Clearly, there is a need for standardization of methods for determining these rates.

We acknowledge that our analysis of false-positive and false-negative rates did not account for the dependence between repeated experiments, although Lynch et. al.[1] determined that the false-positive rate (denoted as "q") was not "dramatically altered."[1(p419)] In fact, on the basis of their criteria, we have gained confidence in a greater number of CNV calls than the 800 reported as "high-frequency CNVs" in our original publication[4(p99)]—that is, an additional 736 CNVs seen in only 2 of the 95 individuals (see data set 2 in the online version of our article[4]). The increase in the false-negative rate (i.e., decrease in p) would have broad implications. If the false-negative rate is as high as Lynch et al. proposed (~60%), the benefit of repeating every experiment with the fluorochromes reversed and eliminating the CNVs not seen in both experiments (also know as "flip-fluor experiments") would be offset by the erroneous elimination of a major portion of real data. Specifically, by achieving a relatively small false-positive rate, flip-fluor repeat experiments (with a false-negative rate of 60%) will capture only 16% of the true CNVs in a given experiment. This raises the question of whether such a practice would be unacceptable if we wish to identify all CNVs in the human population.

Currently, there are >6,000 CNVs noted in the Database of Genomic Variants that affect >3,500 loci.[6] The meta-analysis of the various CNV studies is a major challenge. With the diverse array of platforms employed, it is important to consider the advantages and limitations of each study, since array resolution, DNA reference, genome coverage, and cohort composition vary greatly. Given the limited overlap between individual studies and the indication by Lynch et al.[1] that we are vastly underestimating their prevalence, there are likely tens of thousands of CNVs to be discovered.

Ronald J. deLeeuw, Kendy K. Wong,
Raymond T. Ng, and Wan L. Lam

Web Resource

The URL for data presented herein is as follows:

Database of Genomic Variants,http://projects.tcag.ca/variation/

References

1. Lynch AG, Marioni JC, Tavaré S (2007) Numbers of copy-number variations and false-negative rates will be underestimated if we do not account for the dependence between repeated experiments. Am J Hum Genet 81:418–420 (in this issue)
2. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, et al (2006) Global variation in copy number in the human genome. Nature 444: 444–454
3. Sharp AJ, Locke DP, McGrath SD, Cheng Z, Bailey JA, Vallente

RU, Pertz LM, Clark RA, Schwartz S, Segraves R, et al (2005) Segmental duplications and copy-number variation in the human genome. Am J Hum Genet 77:78–88

4. 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
5. Hinds DA, Kloek AP, Jen M, Chen X, Frazer KA (2006) Common deletions and SNPs are in linkage disequilibrium in the human genome. Nat Genet 38:82–85
6. Zhang J, Feuk L, Duggan GE, Khaja R, Scherer SW (2006) Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. Cytogenet Genome Res 115:205–214

From the Department of Cancer Genetics and Developmental Biology, British Columbia Cancer Research Centre (R.J.d.; K.K.W.; W.L.L.), and the Department of Computer Science, University of British Columbia (R.T.N.), Vancouver

Address for correspondence and reprints: Dr. Ronald J. deLeeuw, 675 W. 10th Avenue, Vancouver, BC V5Z 1L3, Canada. E-mail: rdeleeuw@bccrc.ca

Am. J. Hum. Genet. 2007;81:414. © 2007 by The American Society of Human Genetics. All rights reserved.

0002-9297/2007/8102-0027$15.00

DOI: 10.1086/521416