REPORT

A Bayesian Approach to Copy-Number–Polymorphism Analysis in Nuclear Pedigrees

Konstantina Kosta, Ian Sabroe, Jonathan Göke, Robert J. Nibbs, John Tsanakas, Moira K. Whyte, and M. Dawn Teare

Segmental copy-number polymorphisms (CNPs) represent a significant component of human genetic variation and are likely to contribute to disease susceptibility. These potentially multiallelic and highly polymorphic systems present new challenges to family-based genetic-analysis tools that commonly assume codominant markers and allow for no genotyping error. The copy-number quantitation (CNP phenotype) represents the total number of segmental copies present in an individual and provides a means to infer, rather than to observe, the underlying allele segregation. We present an integrated approach to meet these challenges, in the form of a graphical model in which we infer the underlying CNP phenotype from the (single or replicate) quantitative measure within the analysis while assuming an allele-based system segregating through the pedigree. This approach can be readily applied to the study of any form of genetic measure, and the construction permits extension to a wide variety of hypothesis tests. We have implemented the basic model for use with nuclear families, and we illustrate its application through an analysis of the CNP located in gene *CCL3L1* in 201 families with asthma.

Evidence that segmental copy-number variants (CNVs) represent a significant portion of human genetic variation is accumulating.[1] A CNV is generally defined as a segment of DNA >1 kb and present at variable copy number when compared with a reference genome.[1,2] More than 6,000 such CNVs have been reported (Database of Genomic Variants), and recent genomewide studies estimate that thousands of CNVs exist in the human genome.[1,3] It is likely that these variants lead to phenotypic variation and modification of disease risk through gene-dose or position effects.[2,4,5]

Redon et al.[2] showed that the presence of CNVs is associated with low call rates in SNPs; hence, CNVs tend to occur in regions with low densities of validated SNPs. This is likely to be due in part to the diploid-genome assumption being effectively violated in these CNV regions. The accurate assignment of the copy number (integer count) in an individual will present new challenges to assays,[6] and proposals to use quantitative SNP genotypes to infer CNVs will require more-refined calling algorithms.[2] Accurate quantification of counts of CNV repeats, which can be thought of as allele sizes, is not yet routinely possible; most technologies are able to quantify only the total phenotype or the sum of all alleles detected. CNVs seen in at least 1% of the population are termed copy-number polymorphisms (CNPs) and are good candidates for disease-risk modifiers. Consideration of these polymorphisms in case-control association studies requires no specialized methods, and, when sufficient evidence has accumulated, candidate loci can be further studied and characterized in population-based family studies.[7] In the family-based context, however, the underlying allelic segregation must be inferred from the CNP phenotypes. Although some traditional segregation and linkage analysis tools, such as PAP and LINKAGE, allow genotypes to be inferred from observed phenotype classes, more-recently developed methods aimed at genetic linkage or candidate-gene analysis cannot handle this type of data, since they assume codominant markers. Many CNPs behave as multiallelic systems,[2] and this may lead to nonnegligible error when integers are assigned from quantitative CNP assays. This makes it desirable to develop a method that models the relationship between CNP phenotype and genotype and allows for the CNP phenotype assignment to occur within the statistical segregation analysis itself.

We present a Bayesian graphical model that enables the statistical evaluation of a candidate CNP. Whatever property or aspect of the candidate CNP that is assumed to be associated with disease can then be specifically observed through stochastic elements of the graphical model or the construction of a logical node. A logical node, in this context, means one whose value or state is determined by the states of its parental variables. We have implemented this model in WinBUGS, to analyze nuclear families. The distinctive feature of our implemented method is that the multiple (or repeated) raw assessments of individual CNP phenotype are used directly, rather than use of a single summary measure or an integer value assigned by an al-

From the School of Medicine and Biomedical Sciences, University of Sheffield, Sheffield, United Kingdom (K.K.; I.S.; J.G.; M.K.W.; M.D.T.); Glasgow Biomedical Research Centre, University of Glasgow, Glasgow (R.J.N.); and Pediatric Respiratory Unit, Hippokration General Hospital, Thessaloniki, Greece (J.T.)

Received March 6, 2007; accepted for publication May 9, 2007; electronically published August 8, 2007.

Address for correspondence and reprints: Dr. M. Dawn Teare, Mathematical Modelling and Genetic Epidemiology, School of Medicine, University of Sheffield, Beech Hill Road, Sheffield, S10 2RX, United Kingdom. E-mail: m.d.teare@sheffield.ac.uk

Am. J. Hum. Genet. 2007;81:808–812. 0002-9297/2007/8104-0023$15.00
DOI: 10.1086/520096

gorithm. We assume that the CNP phenotype results from the sum of two independently inherited alleles and that each allele consists of a discrete number of repeats (which may include a null allele). This, of course, may not be true for all CNPs, since the repeated homologous sequence may be located at multiple sites throughout the genome.

The presented implementation has been constructed to evaluate the candidate CNP *CCL3L1* isoform (MIM 601395), which contains a segment that is present in multiple copies. The variation in copy number in this gene segment has been well studied. Within populations, lower copy number is associated with both risk of HIV-1 infection and more rapid progression to AIDS[8] and with reduced risk of Kawasaki disease.[9] Variation in copy number within this gene may affect susceptibility to or progression of other diseases of the autoimmune and inflammatory response systems, such as asthma. This polymorphism has therefore been investigated as a candidate CNP in a series of 201 nuclear families ascertained because of the presence of at least two affected offspring with asthma.

The graphical model is presented in figure 1. In founders, copy-number alleles are assumed to be sampled from a Poisson distribution, where the rate, λ, may be different in each population (as observed by Gonzalez et al.[8] for this same polymorphism). Therefore, the distribution of copy-number phenotype (the sum of the two allele lengths) in the population will follow a Poisson distribution with rate 2λ. In this specific implementation, of interest is whether there is a Mendelian-transmission distortion, in which the lighter allele is transmitted to affected offspring.

The various nodes in figure 1 represent the following.

λ_i: Poisson rate, used in prior population-allele frequency distribution in subpopulation *i*.

$\alpha_{ijl}^{(m)}$ (and $\alpha_{ijl}^{(f)}$): Allele *l* for the mother (and father) in nuclear family *j* in population *i*.

$\gamma_{ij1}^{(m)}$ (and $\gamma_{ij1}^{(f)}$): The "lighter" of the two alleles in the mother (and father).

$\gamma_{ij2}^{(m)}$ (and $\gamma_{ij2}^{(f)}$): The "heavier" of the same two alleles.

τ_a (and τ_u): The probability that a parent transmits the heavier allele to affected (and unaffected) offspring.

$\alpha_{ij1}^{(a_k)}$ (and $\alpha_{ij2}^{(a_k)}$): The allele inherited by affected offspring a_k from the mother (and father) of family *j* in population *i*.

$\alpha_{ij1}^{(u_k)}$ (and $\alpha_{ij2}^{(u_k)}$): The allele inherited by unaffected offspring u_k from the mother (and father) of family *j* in population *i*.

$\phi_{ijr}^{(m)}$ (and $\phi_{ijr}^{(f)}$): CNP-phenotype assay replicate *r* for the mother (and father) of family *j* in population *i*.

$\phi_{ijr}^{(a_k)}$ (and $\phi_{ijr}^{(u_k)}$): CNP-phenotype assay replicate *r* for affected (and unaffected) offspring *k* in family *j* in population *i*.

σ^2: Copy-number–phenotype assay variance.

$\text{rep}_{ij}^{(m)}$ (and $\text{rep}_{ij}^{(f)}$): Number of replicate copy-number–phenotype assays for mother (and father) in family *j* in population *i*.

$\text{rep}_{ij}^{(a_k)}$ (and $\text{rep}_{ij}^{(u_k)}$): Number of replicate copy-number–phenotype assays for affected (and unaffected) offspring.

na_{ij} (and nu_{ij}): Number of affected (and unaffected) offspring in family *j* in population *i*.

This formulation declares the basic model. It is straightforward to extend it to include slight variations or to examine other hypotheses, such as copy-number–threshold

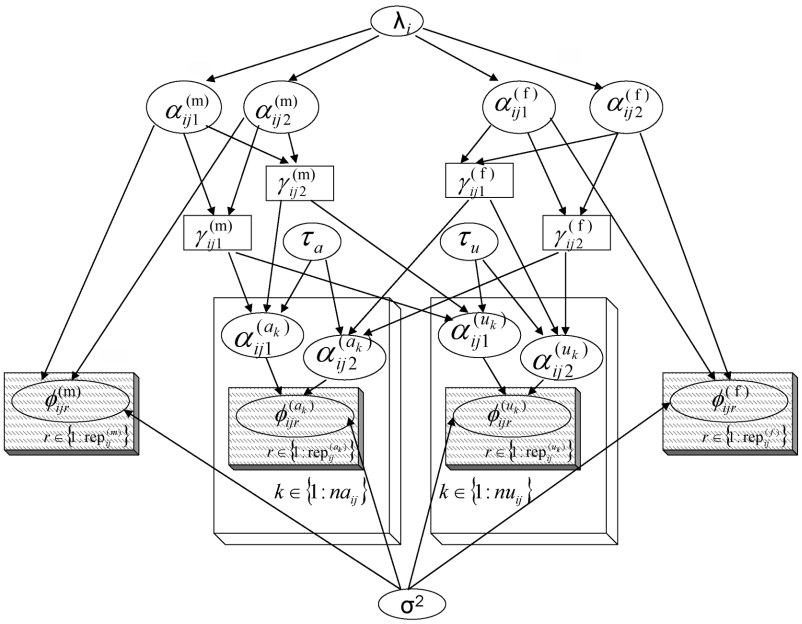

Figure 1. Model represented as a directed acyclic graph. Ovals represent stochastic nodes, rectangles represent logical nodes, and shaded rectangles summarize repeated structures.

effects. Adding further logical nodes enables the estimation of identity-by-descent (IBD) sharing probabilities.

The true copy-number phenotype for a single individual is determined by the sum of their two inherited discrete allele lengths. Therefore, the observed CNP phenotype ϕ_{ijr} is assumed to be a normally distributed variable with mean $\alpha_{ij1} + \alpha_{ij2}$ and variance σ^2. In practice, the phenotype must be treated as a censored variable, since this assay is truncated at zero. When null alleles exist, some individuals will be effectively homozygous for the null allele, and the assay may yield small positive values. Any copy-number assay replicate returning a zero is treated as a censored observation with true value <0.

Founder alleles are sampled from the Poisson distribution, and the genotype is then configured so the first allele is the lighter (if heterozygous). An inheritance vector then identifies which of the two configured parental alleles has been transmitted to offspring. If the gene polymorphism is not associated with disease, you would expect to see equally likely transmission of light or heavy alleles, consistent with Mendelian segregation. This configuration enables the transmission parameter to be estimated and, hence, the alternative hypothesis to be tested. The graphical model distinguishes between affected and unaffected offspring, so unaffected offspring are used only to assist in the inference of parental genotypes.

Nuclear families from two European populations, representing a subset of a larger multicenter study,[10] were collected from centers in Sheffield, United Kingdom, and in Thessaloniki, Greece. The U.K. families include a small number of unaffected offspring (table 1). The estimation of copy-number phenotype was determined by quantitative real-time PCR, by use of a previously validated technique.[11] The copy number was quantified at least twice; if a clear, discrete copy-number count was not achieved, further replicates were generated until the mean value approached an integer value. A similar protocol was followed by Gonzalez et al.[8] This procedure was employed for every member of the nuclear family. Figure 2 shows a selected U.K. pedigree illustrating the variable number of copy-number assays performed within one family. This selected pedigree also provides an example of how assigning the integer closest to the mean, as outlined above, can lead to Mendelian inconsistencies. In this case, the inconsistency was able to be resolved if, for example, the father's phenotype was 3 or the unaffected offspring's phenotype

Table 1. Details of Nuclear Family Size and Distribution of Affected Siblings in the Two Samples

	No. of Families with					
	No. of Affected Siblings			No. of Unaffected Siblings		
Family Origin	2	3	4	0	1	2
United Kingdom	84	14	1	81	15	3
Greece	96	6	0	98	4	0

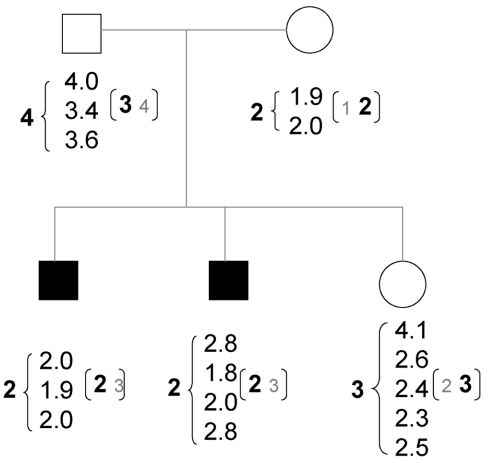

Figure 2. One selected U.K. family. Blackened symbols represent individuals affected with asthma. The column of numbers below each individual lists the replicate CNP assay measure for each. The number to the left of the column represents the integer (closest to the mean) assignment. The numbers to the right in parentheses reflect the most likely (in larger, bold type) and the next most likely underlying phenotype taken from the marginal posterior probability for each individual.

was 0, 2, 4, or 6. Assigning phenotype by the integer closest to the mean resulted in Mendelian inconsistencies detected in 20 of the 201 families. Obviously, some inconsistencies might be expected to be due to nonpaternity, but that cause can be excluded in this series, since all families were screened for nonpaternity before their use in genomewide linkage studies.[10]

The graphical model was implemented as a WinBUGS procedure (CNPrep) allowing for up to six replicate assays per individual. The censoring function was applied only to the first two assay results, since, for this assay at least, a null-copy phenotype was easy to classify. Because of this declared experimental design, missing replicate assays were assumed to be "missing at random."

The combined use of inheritance vectors, which point to the alleles transmitted, and the quantitative estimate of the true copy-number phenotype means that the Markov-chain iterations do not suffer from the problem of reducibility often experienced in Markov-chain pedigree analysis.[12] The assumption that the phenotype is measured with error allows the founder alleles to be sampled from the full prior distribution. Although some founder genotypes are extremely unlikely, they are not impossible.

The implemented procedure needed a long burn-in but did not appear to be sensitive to starting values, and convergence was assessed using standard diagnostics.[13] Multiple runs with three simultaneous sampling chains showed no evidence of reducibility or convergence to local maxima. There is strong evidence that the allele distributions in the Greek and U.K. populations are different (table 2), but there is only suggestive evidence of a trans-

Table 2. Parameter Estimates Generated by WinBUGS Analysis

Parameter[a]	Mean	2.5% CI	97.5% CI
Greek Poisson rate	1.08	.98	1.18
U.K. Poisson rate	1.32	1.21	1.44
Transmission probability[b]	.47	.43	.53
z_0	.22	.17	.27
z_1	.50	.44	.57
z_2	.27	.22	.32

[a] The terms z_0, z_1, and z_2 represent the probabilities that a pair of affected siblings share 0, 1, or 2 alleles IBD.

[b] Probability of transmission of the heavy allele to affected offspring by heterozygote.

mission distortion favoring transmission of the lighter allele to affected offspring ($\hat{\tau}_a = 0.47$ [95% CI 0.43–0.53]). A weak tendency for affected sibs to share two alleles IBD was also observed.

Figure 2 shows the most likely distribution of copy-number phenotype for each person in the selected pedigree, based on the results of the WinBUGS procedure (this is shown by the bold number in parentheses on the right of the column of repeated values). These values represent the most (and next most) likely underlying integer phenotype for each individual taken from the posterior distribution for each person, not the joint distribution for the whole family.

The graphical model outlined in figure 1 constrains the genotype distribution within the pedigree to be consistent with Mendelian segregation. It does not allow for mutation at the CNP, nor does it allow a CNP phenotype to be the "sum" of allele lengths at several unlinked loci. Modifications to incorporate these facets are straightforward, but more prior knowledge or additional data sets allowing characterization through hierarchical modeling would be needed. Currently, available data suggest that CNPs are generally stably inherited (i.e., consistent with a high heritability),[14] but more work is needed to distinguish between technical artifacts, high error rates, and high mutability.

Further refinement of the model would include exploration of appropriate prior distributions for allele frequencies and models allowing assay variance to be a function of the underlying true copy-number phenotype. Our use of the Poisson as the prior for allele frequencies has the benefit that a single parameter defines the distribution; however, for the investigated CNP *CCL3L1,* this was not entirely justified, since the range of allele lengths observed was wider than a Poisson would predict. One effect of the Poisson assumption is illustrated in figure 2. Since the Poisson rate for the U.K. set finds shorter alleles more likely, the father is therefore more likely to have phenotype 3 than phenotype 4, and this resolves the potential Mendelian inconsistency highlighted before between the father and unaffected offspring.

When studying related individuals for segregation analysis of candidate loci, it is not only possible but also beneficial to include raw, possibly replicate, genotype or phenotype measurements. It is beneficial for two reasons: (1) it is more efficient to incorporate the error model into the analysis and to use the raw data directly than to take summary assignments and compose an error function that relates to the summary assignment and the underlying true value, and, (2) although the true Mendelian system is assumed to be discrete, if the quantitative measures can be assumed to be drawn from continuous distributions, then this facilitates the use of a Markov chain–Monte Carlo (MCMC) approach,[12] which allows a general graphical model to form the basis for any genetic hypothesis to be tested. Our approach is illustrated through a simple example implemented in WinBUGS. WinBUGS was chosen because of its favorable environment for flexible model development, but many alternative Bayesian techniques could be used equally well. It should be possible to incorporate this integrated genetic-marker model into existing genetic-analysis tools that use MCMC or hidden Markov models, such as MORGAN or MERLIN, thereby taking advantage of their many other algorithmic and computational features.

This approach and the realization that presence of CNVs can be detected through SNP-driven technology suggest that the model can be extended to SNP analysis. In genetic association studies, SNPs have become the marker of choice[15] because of their high genomic density and technical advantages over the more polymorphic but less frequent and technically more difficult microsatellites. In spite of their technical advantages, not all discovered SNPs meet required quality-control (QC) standards, and some genomic regions are relatively sparsely covered by SNPs.[16] It is now clear that some of the QC failures may be due to the presence of copy-number–variable regions.[2,17] This makes it desirable to work toward models able to simultaneously analyze SNP and CNP observations together, to more precisely identify the segregation of genetic material through pedigrees, and hence to formally assess the disease risk. A recent genetic linkage analysis that mapped autism risk loci[18] took advantage of the information on individual CNV status contained in the SNP arrays. Because of the current lack of both integrated statistical methods and firm knowledge of suitable and reliable inheritance models for CNVs, the patterns of CNV clustering were used to subclassify families rather than to inform the genetic linkage process. The analysis of further detailed experimental data such as those reported in the linkage analysis[18] is required to ensure appropriate models integrating the effect of CNP on neighboring SNPs.

Genomewide case-control association studies are an efficient way to screen large numbers of genetic loci for association with multifactorial disease. Candidate regions suggested by the genomewide studies may be followed up in population-based family resources, since they provide a powerful framework to quantify gene-gene and gene-environment interactions. CNPs present a promising class of candidate loci, but the error associated with measuring

the population-based variation may be large when assigning phenotype to individuals, requiring consideration of the error or the assignment within the statistical analysis. We have outlined an efficient approach to do this, using a Bayesian graphical model for candidate-gene segregation analysis that jointly models the relationship between (single or replicate) quantification assays and true underlying genotype and/or phenotype. As we have demonstrated, this basic model is easily adapted to suit the nature of the data studied and the hypothesis in question.

Acknowledgments

We thank all the families in Thessaloniki, Greece, and Sheffield, United Kingdom, for their participation in this study. Thanks also to James E. Pease at Imperial College for helpful discussions. The recruitment and phenotyping of the families was funded by GlaxoSmithKline as part of the Genetics of Asthma International Network (GAIN) Study and via an unrestricted educational grant. This project was funded by European Respiratory Society long-term research fellowship award 2003-023 (to K.K.).

Web Resources

The URLs for data presented herein are as follows:

CNPrep, http://www.dawn-teare.staff.shef.ac.uk/
Database of Genomic Variants, http://projects.tcag.ca/variation/
LINKAGE, http://linkage.rockefeller.edu/soft/linkage/
MERLIN, http://www.sph.umich.edu/csg/abecasis/Merlin/
MORGAN, http://www.stat.washington.edu/thompson/Genepi/MORGAN/
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nih.gov/Omim/ (for *CCL3L1* isoform)
PAP Pedigree Analysis Package, http://hasstedt.genetics.utah.edu/
WinBUGS, http://www.mrc-bsu.cam.ac.uk/bugs/

References

1. Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. Nat Rev Genet 7:85–97
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. 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
4. Inoue K, Lupski JR (2002) Molecular mechanisms for genomic disorders. Annu Rev Genomics Hum Genet 3:199–242
5. Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C, et al (2007) Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315:848–853
6. Armour JA, Palla R, Zeeuwen PL, den Heijer M, Schalkwijk J, Hollox EJ (2007) Accurate high-throughput typing of copy number variation using paralogue ratios from dispersed repeats. Nucleic Acids Res 35:e19
7. Barrett JH, Sheehan NA, Cox A, Worthington J, Cannings C, Teare MD (2007) Family based studies and genetic epidemiology: theory and practice. Hum Hered 64:146–148
8. Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, Nibbs RJ, Freedman BI, Quinones MP, Bamshad MJ, et al (2005) The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 307:1434–1440
9. Burns JC, Shimizu C, Gonzalez E, Kulkarni H, Patel S, Shike H, Sundel RS, Newburger JW, Ahuja SK (2005) Genetic variations in the receptor-ligand pair *CCR5* and *CCL3L1* are important determinants of susceptibility to Kawasaki disease. J Infect Dis 192:344–349
10. Pillai SG, Chiano MN, White NJ, Speer M, Barnes KC, Carlsen K, Gerritsen J, Helms P, Lenney W, Silverman M, et al (2006) A genome-wide search for linkage to asthma phenotypes in the Genetics of Asthma International Network families: evidence for a major susceptibility locus on chromosome 2p. Eur J Hum Genet 14:307–316
11. Townson JR, Barcellos LF, Nibbs RJ (2002) Gene copy number regulates the production of the human chemokine CCL3-L1. Eur J Immunol 32:3016–3026
12. Sheehan NA (2000) On the application of Markov chain Monte Carlo methods to genetic analysis on complex pedigrees. Int Stat Rev 68:83–110
13. Gammerman D (1997) Markov chain Monte Carlo: stochastic simulation for Bayesian inference. Chapman and Hall, London
14. Locke DP, Sharp AJ, McCarroll SA, McGrath SD, Newman TL, Cheng Z, Schwartz S, Albertson DG, Pinkel D, Altshuler DM, et al (2006) Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. Am J Hum Genet 79:275–290
15. Cordell HJ, Clayton DG (2005) Genetic association studies. Lancet 366:1121–1131
16. Nicolae DL, Wen XQ, Voight BF, Cox NJ (2006) Coverage and characteristics of the Affymetrix GeneChip Human Mapping 100K SNP set. PLoS Genet 2:e67
17. Carlson CS, Smith JD, Stanaway IB, Rieder MJ, Nickerson DA (2006) Direct detection of null alleles in SNP genotyping data. Hum Mol Genet 15:1931–1937
18. Szatmari P, Paterson AD, Zwaigenbaum L, Roberts W, Brian J, Liu X-Q, Vincent JB, Skaug JL, Thompson AP, Senman L, et al (2007) Mapping autism risk loci using genetic linkage and chromosomal rearrangements. Nat Genet 39:319–328