A Bayesian shared component model to analyze copy number data in genetic studies

Juan R González[1,2,4,*], Carlos Abellán[3], Juan J Abellán[3,4]

[1]Center for Research in Environmental Epidemiology (CREAL), Barcelona, Spain.

[2]Institut Municipal d'Investigació Mèdica (IMIM), Barcelona, Spain

[3]Joint Research Unit on Genomics and Health, Centre for Public Health Research (CSISP) and Cavanilles Institute for Biodiversity and Evolutionary Biology, University of Valencia, Valencia, Spain

[4]CIBER Epidemiología y Salud Pública (CIBERESP), Spain

*Corresponding author: Juan R González

Center for Research in Environmental Epidemiology (CREAL)

Doctor Aiguader 88

08003 Barcelona, Spain

E-mail: jrgonzalez@creal.cat

URL: http://www.creal.cat/jrgonzalez/software.htm

Abstract

An important question in genetic studies is to determine those genetic variants, in particular CNVs, that are specific of different groups of individuals. This could help in elucidating differences in disease predisposition and response to pharmaceutical treatments. We propose a Bayesian model designed to analyze thousand of CNVs where only few of them are expected to be associated with a specific phenotype. The model is illustrated by analyzing three major human groups belonging to HapMap data. We also show how the model can also be used for analyzing a case-control study with cases from different diseases. Through a simulation study, we show that the proposed model outperforms other approaches that are typically used to analyze this data. We have developed an `R` package, called `bayesGen` that implements the model.

KEY WORDS: Bayesian hierarchical model; CNVs; HapMap; sub-phenotype

1 Introduction

The aim of genome-wide association studies (GWAS) is to asses the association between single nucleotide polymorphisms (SNPs) and common diseases. Recent GWAS have been successful in discovering SNPs significantly associated with complex diseases [11, 6]. However, published SNP associations account for only a fraction of the genetic component of most common diseases [13]. Lately, several studies have been focused on the association between copy number variants (CNV) and disease. Some reports have suggested a role of rare CNVs (i.e. CNV with low prevalence in the general population)

in susceptibility to neurodevelopmental disorders [22, 21, 23]. Other studies have shown statistically significant associations between common CNVs (i.e. CNV with high prevalence in the general population) and common diseases such as psoriasis [5], Chron's disease [14], HIV-1/AIDS [9], or Alzheimer [19] to name a few. These studies indicate that the identification of DNA copy number is important in understanding the genesis and progression of human diseases.

Several techniques and platforms have been developed for GWAS involving CNVs, such as array-based comparative genomic hybridization (aCGH). For targeted studies, other techniques such as real time PCR, or Multiplex Ligation-dependent Probe Amplification (MLPA) assays have been used to compare the copy number status of particular loci in cases and controls. In both cases, a signal intensity is measured for each CNV as a continuous variable, from which the copy number status is inferred. In many cases, the distribution of the observed CNV probe measurements is continuous and multimodal, representing the unobserved copy number status as a latent variable [10]. Thus, scoring copy number may lead to misclassification and, hence, unreliable results, making it necessary to incorporate uncertainty in the association analysis. So far, two methods have been developed to analyze CNV data that incorporate uncertainty. The first one performs the calling procedure and incorporates the posterior probabilities in a latent class model [10], while the other is based on a likelihood test that combines calling and testing in a single procedure [3]. Despite the existence of these methods CNV association studies often analyze CNVs with very low uncertainty that are not likely genotyping artefacts. For example, in the GWAS

performed in the Myocardial Infarction Genetics Consortium [15] the authors pointed out that: "[for the CNV analysis] as an initial quality control step, [they] removed any variant where more than 10% of the copy calls were uncertain" ([15], Supplementary Material). Such approach allows the use of standard tests such as χ^2, Fisher or Mann-Whitney tests [9, 19, 14, 5] to assess differences between cases and controls.

Herein, we present a Bayesian shared component model for CNV-based association studies. We illustrate the model with a case study to determine those CNVs that are specific of a given population when comparing individuals belonging to different ethnic groups. This model could also be applied to studies where the cases have different subphenotypes. The approach here adapts the model suggested by [1] for genetic association studies based on SNPs, to the context of CNVs.

The remainder of the paper unfolds as follows. Section 2 describes HapMap data used to determine those CNVs that are specific of any of three major ethnic groups. The Bayesian shared component model is introduced in Section 3, where the likelihood, prior, hyperpriors and the inferential process are specified. The estimating method is given in Section 3.1. The model is illustrated in Section 4, where a genetic analysis of CNVs for three major ethnics groups belonging to HapMap data is presented. A simulation study is presented in Section 5. Our conclusions are given in Section 6.

2 Motivating data

The motivating data were collected from a genetic study conducted at the Center for Genomic Regulation (CRG) in Barcelona, Spain. The study was aimed to determine those CNVs that are specific for major human ethnic groups included in the HapMap project (e.g., African, Asian or European) [2]. This type of data can help in the understanding of some Mendelian diseases such as cystic fibrosis [4] or deafness [7] that present different prevalence in the different populations. In addition, the genomic variants that are population-specific can guide to drug discovery. For example, the existing population variability in the acetylating activity of the N-acetyltransferase 2 (NAT2) gene makes possible to determine those ethnic groups that are more susceptible to develop some diseases [24].

As previously mentioned, a very simple approach to determine the CNVs that are specific to each population is to compare the observed CNV frequency in individuals from different ethnic groups [2]. One of the main limitations of this approach is that the number of copies may vary between 0 up to 6 and therefore χ^2, Fisher or Mann-Whitney tests can be underpowered. In addition most of the CNVs analyzed have similar frequencies accross ethnic groups, and only a few, if any, show differences between them. Therefore, the use of a shared component model can be very useful in the context of CNVs.

3 The Bayesian Model

Let $\{X_{ijp} \in D\}$ be the number of copies of the jth CNV, for the ith individual of population p, where D denotes the set of indices for the observed data, $i = 1, \ldots, n$ (number of individuals), $j = 1, \ldots, c$ (number of CNVs) and $p = 1, \ldots, P$ (number of populations). We assume that all individuals in the same population group have the same chance of having a number of copies in a given CNV, then we observe $X_{ijp} \in \{0, 1, 2, 3, 4, \ldots\}$.

Now, let $Y_{jp} = \sum_{i=1}^{n_{jp}} \frac{X_{ijp}}{n_{jp}}$ be the average number of copies found in the jth CNV of the pth population, where n_{jd} denotes the number of individuals in population p with non-missing information for the jth CNV. Then, by the central limit theorem [18], and assuming independence among individuals we have

$$Y_{jp} \dot{\sim} N(\mu_{jp}, \nu_p^2), \tag{1}$$

where μ_{jp} is the mean frequency for CNV j in population p and ν_p^2 is the variation of the average of CNV frequencies in population p.

We introduce the next shared component formulation with Gaussian likelihood to decompose the variability of μ_{jp}

$$\mu_{jp} = \alpha_p + \beta_p \cdot \theta_j + \lambda_{jp}, \tag{2}$$

where α_p is a population-specific intercept, θ_j is the component shared by all populations, β_p denotes the loading of the common component into population p and λ_{jp} encodes the population-specific components. In order to make the model as flexible as possible we have considered that ν_p^2 depends

on the population group p. However, a simpler model can also be fitted by considering that Y_{jp} has the same variance for each population group, ν^2.

Figure (1) depicts a schematic representation of our model. Notice that this formulation considers that no reference group is available (i.e control group). The formulation can be changed to accomodate the possibility of having a control group. For example, in the context of a case-control study where different diseases and only one group of control individuals is available. This is the case of the Wellcome Trust Case Control Consortium (WTCCC) study where 7 common diseases are compared with a unique group of controls [25] and thousands of CNVs were analyzed.

To complete the Bayesian formulation, the prior and hyperprior distribution for the model parameters are needed. Our basic principle in specifying these distributions is to let the likelihood dominate over the prior information. Therefore, non-informative prior distributions are assumed. We also refer to previous similar studies that specify prior distributions in this way. We assumed the following priors

$$\begin{aligned} \alpha_p &\sim \text{Normal}(0, 1000) \\ \theta_j &\sim \text{Normal}(0, \sigma_\theta^2) \\ \lambda_{jp} &\sim t_4(0, \sigma_p^2) \\ \beta_p &\sim \text{Normal}(0, 100) \end{aligned}$$

and non-informative hyperpriors for the standard deviations of the random

effects

$$\sigma_\theta, \sigma_p \sim \text{Normal}(0, 100) \cdot \mathbf{1}_{(0,+\infty)}$$

For the sake of the identifiability we fixed $\sigma_\theta^2 = 1$. These priors and hyperpriors are commonly used for full Bayesian statistical inference when information about the model parameters are not available. However, the specific components, λ_{jp}, were considered as a zero-mean t-distribution with 4 degrees of freedom and unknown variances to account for large values. The priors and hyperpriors for the asymmetric formulation (e.g. having a control group and different diseases) are mainly the same, except that we consider $\beta_1 = 1$, where β_1 correspond to the reference population.

3.1 Estimation of model parameters

The JAGS software (available at `http://www-fis.iarc.fr/martyn/software/jags/`) was used to carry out MCMC posterior sampling using the `R` package `rjags` [16]. We ran the sampler 40,000 iterations and considered estimates based on the last 30,000 runs, allowing a burn-in of 10,000 iterations. Two chains were run for each of the models. Convergence was assessed from time series and $Q - Q$ plots. We also used the "potential scale reduction factor" diagnostic proposed by Gelman and Rubin [8].

MCMC is computationally intensive, even more in the case of analyzing genetic data where normally thousands of genes are analyzed. To overcome this difficulty, we also used the Integrated Nested Laplace Approximation (INLA) approach to make statistical inference of our model. INLA provides a fast (it gives answers in minutes and seconds where MCMC

requires hours and days), deterministic alternative to MCMC [20]. The only difference between both approaches is that the model based on INLA replaces the t distributions for Normals. This in principle could shrink CNV-disease risk associations more than the original model, but it runs much faster and it can be applied to GWAS. In any case, the t distribution can easily be incorporated when available (`http://www.r-inla.org/`). We have developed an `R` package called `bayesGen` that incorporates both estimating processess as well as some tools for displaying model parameters and evaluating model convergence. The package is available at `http://www.creal.cat/jrgonzalez/software.htm`

4 Example: Genomic differences between human populations

In this section some results are shown for the application of the proposed model to the HapMap data (`http://hapmap.ncbi.nlm.nih.gov/`). Armengol et al. [2] showed some CNVs that are present in different frequencies accross individuals belonging to three human populations (YRI-Yoruba in Ibadan, Nigeria, CEU-Utah residents with ancestry from Northern and Western Europe; and CHB/JPT-Han Chinese in Beijing, China and Japanese in Tokyo, Japan), representatives of sub-Saharan Africa, Europe and East Asia, respectively. The authors, in a preliminary step, used aCGH and BAC-based platforms to identify CNVs with different frequencies in the three populations using pools of individuals. This yielded a total of 111 loci whose copy number state frequencies differed among populations. In order

to confirm the changes detected with the aCGH platforms, they performed validation experiments using MLPA on individual DNAs from the HapMap samples. In total they analyzed 152 CNVs (genes). Overall, they found 33 CNVs that were specific to any of the three populations after applying standard statistical tests (χ^2 or Fisher tests).

The final data set we use for illustration purposes, consists of 120 CNVs (we removed 32 CNVS that were not variable among populations) and 261 individuals (56 CEU, 58 YRI and 147 CHB/JPT) belonging to the MLPA experiment. Therefore, our data consists in a 261×120-dimensional matrix with values corresponding to the observed copy number status $X_{ijp} \in \{0, 1, 2, 3, 4\}$. After aggregating the counts of each number of copies over the individuals in each population for each CNV we fit the model 3 to the aggregated data Y_{jp} where $j = 1, \ldots, 120$ and $p \in \{\text{CEU, YRI, CHB/JPT}\}$. Using the `bayesCNVassoc` function in the `bayesGen R` package we ran two chains of 200,000 iterations. We discarded the first 20,000 and kept every 50th to reduce autocorrelation in the chains. Inference is therefore based on (thinned) samples of size 4,000. We assessed convergence using graphical techniques and the Gelman-Rubin method and no symptoms of non-convergence were detected. To keep the false discovery rate under control when evaluating whether a specific component was statistically significant or not, we computed credible intervals at 99.98% level (in the frequentist framework this would be equivalent to a Bonferroni correction $0.05/120 \sim 0.0002$) for λ_{jp}'s.

Table 1 shows the estimates for the population-specific intercepts α_p for the shared component model assuming a symmetric formulation. The

specific intercept for all three populations, α_p, is around 2 as expected. Regarding the specific component for each population we found that only 31 CNVs were population-specific (Figure 2). By looking at the estimates of ν_p we observe that $\nu_{\text{CEU}} = 0.0756$, while $\nu_{\text{CHB/JPT}} = 0.0306$ and $\nu_{\text{CEU}} = 0.0362$. This indicates that there is more variability among european individuals, which decreases the power of finding any specific CNV for european population.

Armengol et al. [2] found 33 population-specific CNVs after using χ^2 or Fisher tests. In order to compare the performance of both approaches we tested the existence of population stratification (i.e. genetic differences among individuals) using a principal component analysis (PCA) as suggested in [17]. Armengol et al., estimated that 30% of total variance is explained by the two first principal components (PC1 16.6%, and PC2 13.4%) using 33 CNVs. In our case, with only 31 CNVs, the two first principal components explain a 38.3% of the total variability (PC1 22.1%, and PC2 16.2%) indicating that our subset of variants better discriminates the individuals.

5 Simulation Studies

As in real data set we can only illustrate the methods and the truth about which CNVs are really associated with each population is unknown, we carried out a small-scale simulation study to evaluate our proposed model mimicking the real data analysis of major human populations in some aspects. We considered three different populations with 500 and 2,000 CNVs. Only two of the CNVs were in a different proportion for one population (i.e. these two CNVs were specific for such group of individuals). We simulated 3 different scenarios for the trully associated CNVs. The first one considers that the two CNVs are hihgly associated with one of the population (OR=2.0), the second one consider a moderate increase on risk (OR=1.5), while the third one is designed to study the performance of our proposed method in a low risk scenario (OR=1.2). The simulation emulates a likely association between thousands of genes and disease. In genetic studies only a few of the analyzed genes are trully associated with the phenotype of interest. For instance, the WTCCC analyzed 3,432 CNVs among different diseases and only found 3 loci where CNVs were associated with disease [25].

The copy number status for the CNVs were simulated considering two types of CNV data. The first one assumes that CNVs were common, meaning that they can be tagged by a SNPs, and hence the copy number status can only be $\{0, 1, 2\}$. The second scenario considers polymorphic CNVs taking values $\{0, 1, 2, 3, 4, 5, 6\}$. In both cases we simulated CNVs assuming Hardy-Weinberg equilibrium and the allelic frequencies were randomly selected form $U(0.05, 0.5)$. We compared the results obtained from our pro-

posed Bayesian shared component model with those obtained with a χ^2 test, non-parametric Kruskall-Wallis test and multinomial logistic regression comparing the null model versus the model including the CNV using the likelihood ratio test. Bonferroni correction was used in order to deal with multiple comparisons. We also computed corrected credible intervals for the specific components. Given that the Bonferroni-like correction requires estimation of extreme percentiles for the posterior distribution, which are difficult to be obtained from MCMC samples, we computed a credible interval based on the normal approximation. Finally, we considered the posterior probability as an alternative criterion to detect significant CNVs. We compared the different approaches by computing the true and false positive rate (TPR and FPR, respectively) in 500 simulations.

Table 2 shows the TPR and FPR for the different methods in the case of analyzing common CNVs. Across all scenarios, as expected, the TPR decreases when the ORs for the significant CNVs decrease. The TPR are almost 100% in all cases since only two of the CNVs (500 or 2,000) were simulated with a signal different from 0. We observed that the Bayesian shared component model outperforms the other methods in the case of having low and moderate risk effects. This finding is important since common CNVs can be tagged by SNPs and their risks are expected to be about 1.15-1.45. For example in the context of CNVs that can be tagged by a SNPs, de Cid et al. [5] found that the risk of having one copy of the LCE gene increased by 41% the chance of having psoriasis. We finally noticed that non-parametric tests are not able to detect the two significant CNVs in any situation, suggesting that such method is not a good choice for the analysis of CNV data

with a very small number of significant signals. On the other hand, Table 3 shows the TPR and FPR in the case of analyzing complex/polymorphic CNVs. Overall, the results are the same as those obtained for the case of analyzing common SNPs, showing even more differences between Bayesian model and the other methods. This can be explained by the fact that by simulating CNV with number of copies between 0 and 6, the number of individuals in each category is reduced. In this situation, the power of using methods based on the observed number of individuals in each category decreases.

6 Conclusion

Here we considered the problem of determining genetic variants that are specific to different subgroups of individuals or different subphenotypes when thousand of variants are analyzed and only a few of them are truly associated to a given group. In particular, we focused on describing how to find specific CNVs to the three major ethnic groups. We have implemented a Bayesian shared component model to decompose the observed variability in the number of copies of each CNV into two components: shared and specific. Simulation results showed a better performance than other existing methods.

We established the CNVs that are specific to each population by computing credible intervals of the posterior mean of the specific components and their posterior probabilities. In order to avoid false positive results, we adopted a Bonferroni-like correction. Therefore credible intervals require estimation of extreme percentiles. This may lead to some difficulties when using MCMC samples. Thus, we also calculated credible intervals based on normal approximation. Simulation studies showed that this method slightly outperforms the method based on percentiles.

The model has been formulated using a hierarchical structure. Therefore, it is straightforward to add further levels of hierarchy if needed. For instance, CNVs can be in the same pathway or may have the same function. Thus, this information can be incorporated in the model in order to better estimate the effect of each CNV as described in [12].

We conclude that our proposed model is useful to discover specific ge-

netic variants for different subgroups of individuals. This could help in determining differences in disease predisposition or response to pharmaceutical treatments. Estimating model parameters can be very time consuming, however we have developed an R package (`bayesGen`) that not only includes MCMC methods but also a fast estimation of the posterior distribution based on INLA that provides estimates for a whole chromosome in a few minutes.

Acknowledgments

We thank Prof. Havard Rue for help with the implementation of the shared component model using INLA. We also thank Xavier Estivill and Lluś Armengol for providing access to the HapMap data. This work has been supported by the Spanish Ministry of Science and Innovation (MTM2008-02457 to JRG) and Grants GVPRE/2008/010 and AP-055/09 from Generalitat Valenciana (JA).

References

[1] J. J. Abellan, C Abellan, and J. R. Gonzalez. A Bayesian shared component model for genome association studies. Technical Report 1120, COBRA, 2010.

[2] L Armengol, S Villatoro, JR González, L Pantano, M García-Aragonés, R Rabionet, M Cáceres, and X Estivill. Identification of copy number

variants defining genomic differences among major human groups. *PLoS ONE*, 4(9):e7230+, September 2009.

[3] C. Barnes, V. Plagnol, T. Fitzgerald, R. Redon, J. Marchini, D. Clayton, and M. E. Hurles. A robust statistical method for case-control association testing with copy number variation. *Nature Genetics*, 40(10):1245–52, 2008.

[4] J. L. Bobadilla, M. Macek, J. P. Fine, and P. M. Farrell. Cystic fibrosis: A worldwide analysis of CFTR mutationscorrelation with incidence data and application to screening. *Human Mutation*, 19:575–606, 2002.

[5] R de Cid, E Riveira-Munoz, PL Zeeuwen, J Robarge, W Liao, EN Dannhauser, E Giardina, PE Stuart, R Nair, C Helms, G Escaramis, E Ballana, G Martn-Ezquerra, M den Heijer, M Kamsteeg, I Joosten, EE Eichler, C Lazaro, RM Pujol, L Armengol, G Abecasis, JT Elder, G Novelli, JA Armour, PY Kwok, A Bowcock, J Schalkwijk, and X Estivill. Deletion of the late cornified envelope LCE3B and LCE3C genes as a susceptibility factor for psoriasis. *Nature genetics*, 41(2):211–215, 2009.

[6] P Donnely. Progress and challenges in genome-wide association studies in humans. *Nature*, 456:728–731, 2008.

[7] P Gasparini, R Rabionet, G Barbujani, S Melhionda, M Petersen, K Brondum-Nielsen, A Metspalu, E Oitmaa, Pisano M, P Fortina, L Zelante, and X. Estivill. High carrier frequency of the 35delG deaf-

ness mutation in European populations. Genetic Analysis Consortium of GJB2 35delG. *European Journal Human Genetics*, 8:19–23, 2000.

[8] A. Gelman and D.B. Rubin. Inference from iterative simulation using multiple sequences (with discussion). *Statistical Science*, (7):457–472, 1992.

[9] E. Gonzalez, H. Kulkarni, H. Bolivar, A. Mangano, R. Sanchez, G. Catano, and et al. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science*, 307(5714):1434–40, 2005.

[10] J. R. Gonzalez, I. Subirana, G. Escaramis, S. Peraza, A. Caceres, X. Estivill, and L. Armengol. Accounting for uncertainty when assessing association between copy number and disease: a latent class model. *BMC Bioinformatics*, 10:172, 2009.

[11] L.A. Hindorff, H.A. Junkins, J.P. Mehta, and T.A. Manolio. A catalog of published genome-wide association studies. *[accessed, 14 September 2010]. Available at http://www.genome.gov/26525384*, 2010.

[12] R. J. Hung, P Brennan, C Malaveille, S Porru, F Donato, P Boffetta, and J. S. Witte. Using hierarchical modeling in genetic association studies with multiple markers: Application to a case-control study of bladder cancer. *Cancer Epidemiology, Biomarkers and Prevention*, (13):1013, 2004.

[13] T.A. Manolio, FS Collins, NJ Cox, DB Goldstein, LA Hindorff, DJ Hunter, MI McCarthy, EM Ramos, LR Cardon, A Chakravarti,

JH Cho, AE Guttmacher, A Kong, L Kruglyak, E Mardis, CN Rotimi, M Slatkin, D Valle, AS Whittemore, M Boehnke, AG Clark, EE Eichler, G Gibson, JL Haines, TF Mackay, SA McCarroll, and PM Visscher. Finding the missing heritability of complex diseases. *Nature*, 461:747–753, 2009.

[14] SA McCarroll, A Huett, P Kuballa, SD Chilewski, A Landry, P Goyette, MC Zody, JL Hall, SR Brant, JH Cho, RH Duerr, MS Silverberg, KD Taylor, JD Rioux, D Altshuler, MJ Daly, and RJ Xavier. Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nature Genetics*, 40(9):1107–1112, 2008.

[15] Myocardial Infarction Genetics Consortium. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nature genetics*, 41(3):334–341, 2009.

[16] Martyn Plummer. *rjags: Bayesian graphical models using MCMC*, 2009. R package version 1.0.3-13.

[17] Alkes L. Price, Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8):904–909, July 2006.

[18] John A. Rice. *Mathematical Statistics and Data Analysis*. Duxbury Press, 2nd edition, 1995.

[19] A. Rovelet-Lecrux, D. Hannequin, G. Raux, N. Le Meur, A. Laquerriere, A. Vital, C. Dumanchin, S. Feuillette, A. Brice, M. Vercelletto, F. Dubas, T. Frebourg, and D. Campion. App locus duplication causes autosomal dominant early-onset alzheimer disease with cerebral amyloid angiopathy. *Nat Genet*, 38(1):24–6, 2006.

[20] H. Rue, S. Martino, and N Chopin. Approximate bayesian inference for latent gaussian models using integrated nested laplace approximations (with discussion). *Journal of the Royal Statistical Sociery, Series B*, (71):319–392, 2009.

[21] J Sebat, B Lakshmi, D Malhotra, J Troge, C Lese-Martin, T Walsh, B Yamrom, S Yoon, A Krasnitz, J Kendall, A Leotta, D Pai, R Zhang, Y Lee, J Hicks, SJ Spence, AT Lee, K Puura, T Lehtimaki, D Ledbetter, PK Gregersen, J Bregman, JS Sutcliffe, V Jobanputra, W Chung, D Warburton, MC King, D Skuse, DH Geschwind, TC Gilliam, K Ye, and M Wigler. Strong association of de novo copy number mutations with autism. *Science*, 316:445–449, 2007.

[22] P. Stankiewicz and A.L. Beaudet. Use of array cgh in the evaluation of dysmorphology, malformations, developmental delay, and idiopathic mental retardation. *Curr. Opin. Genet. Dev.*, 17:182–192, 2007.

[23] The International Schizopherenia Consortium. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature*, 455:237–241, 2008.

[24] K. P. Vatsis, K. J. Martell, and W. W. Weber. Diverse point mutations

in the human gene for polymorphic N-acetyltransferase. *Proc Natl Acad Sci U S A*, 88:6333–6337, 1991.

[25] Wellcome Trust Case Control Consortium. Genome-wide association study of cnvs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, 464(7289):713–720, 2010.

Group	Parameter	median (95%CI)
CEU	α_1	1.953 (1.892, 2.016)
YRI	α_2	1.989 (1.943, 2.036)
CHB/JPT	α_3	1.974 (1.926, 2.029)

Table 1: Posterior median and 95% credibility intervals for population-specic intercepts.

					Bayesian Shared Model		
	# SNPs	χ^2	K-W	Multinomial regression	Posterior Distribution	Normal Approximation	Posterior Probability
high risk scenario (OR=2.0)							
TPR	2000	100.00	0	100.00	100.00	100.00	100.00
TNR	2000	100.00	100	100.00	99.97	99.99	99.94
TPR	500	100.00	0	100.00	100.00	100.00	100.00
TNR	500	99.99	100	99.99	99.88	99.94	99.76
moderate risk scenario (OR=1.5)							
TPR	2000	55.25	0	55.50	71.50	71.50	77.00
TNR	2000	100.00	100	100.00	99.98	99.99	99.95
TPR	500	68.50	0	66.25	78.25	86.75	84.50
TNR	500	99.98	100	99.99	99.91	99.94	99.82
low risk scenario (OR=1.2)							
TPR	2000	14.25	0	14.50	22.50	23.50	28.75
TNR	2000	100.00	100	100.00	99.99	100.00	99.98
TPR	500	25.25	0	25.75	32.50	33.25	36.50
TNR	500	99.99	100	99.99	99.99	99.99	99.98

Table 2: Results for the simulation described in Section 5 for the case of having common CNVs. The scenarios are described in that section. We compare four different approaches: χ^2 test, Kruskall-Wallis (K-W), Multinomial regression using likelihood ratio test, and our proposed Bayesian model. The comparison was based on computing the True Positive and Negative Rates, TPR and TNR respectively.

	# SNPs	χ^2	K-W	Multinomial regression	Bayesian Shared Model		
					Posterior Distribution	Normal Approximation	Posterior Probability
moderate risk scenario (OR=1.5)							
TPR	2000	24.32	0	32.1	57.34	58.22	57.90
TNR	2000	100.00	100	100.00	99.99	99.98	99.97
TPR	500	40.10	0	50.42	69.85	70.2	70.61
TNR	500	99.99	100	99.99	99.95	99.96	99.90
low risk scenario (OR=1.2)							
TPR	2000	7.86	0	9.75	19.23	20.64	21.57
TNR	2000	100.00	100	100.00	99.99	99.99	99.98
TPR	500	9.62	0	11.71	22.62	23.88	24.20
TNR	500	99.99	100	99.99	99.99	99.99	99.98

Table 3: Results for the simulation described in Section 5 for the case of having polymorphic CNVs. The scenarios are described in that section. We compare four different approaches: χ^2 test, Kruskall-Wallis (K-W), Multinomial regression using likelihood ratio test, and our proposed Bayesian model. The comparison was based on computing the True Positive and Negative Rates, TPR and TNR respectively.

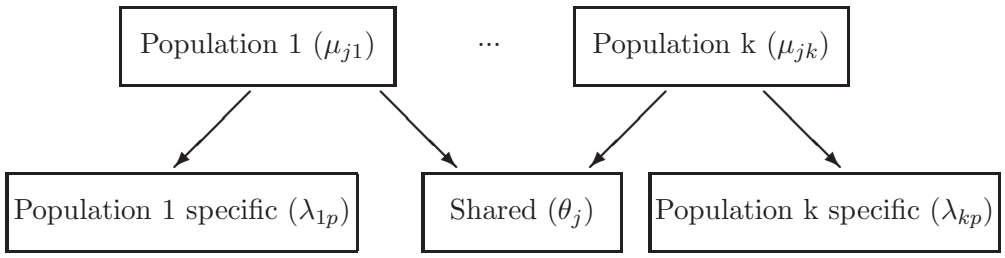

Figure 1: Schematic representations of the shared component model using a symmetric formulation (i.e., no reference group)

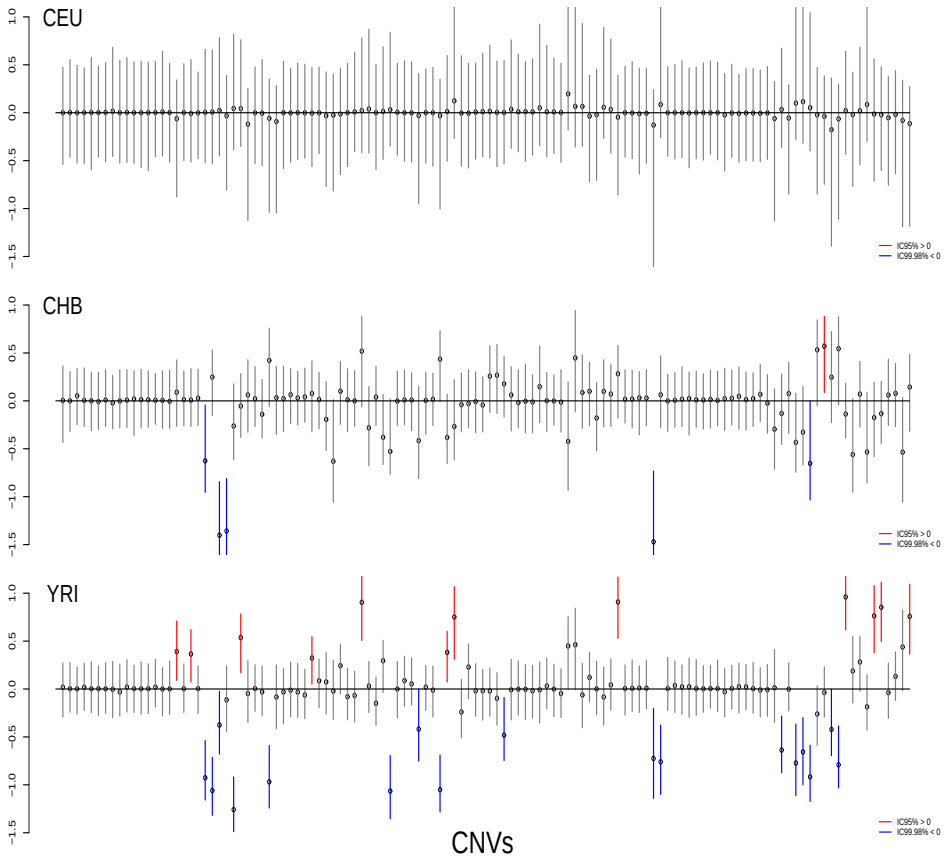

Figure 2: Estimates of specific components, λ_{jp}, for each CNV and different human populations. Each point represents the posterior medians, while segments show its 99.98% credibility intervals. CNVs that are statistically significant specific of each population are coloured in red (gains) and blue (losses).