BMC Bioinformatics

RESEARCH ARTICLE Open Access

On the association analysis of CNV data: a fast and robust family-based association method

Meiling Liu[1,2†], Sanghoon Moon[3†], Longfei Wang[4†], Sulgi Kim[5], Yeon-Jung Kim[3], Mi Yeong Hwang[3], Young Jin Kim[3], Robert C. Elston[6], Bong-Jo Kim[3*] and Sungho Won[4,7,8*]

Abstract

Background: Copy number variation (CNV) is known to play an important role in the genetics of complex diseases and several methods have been proposed to detect association of CNV with phenotypes of interest. Statistical methods for CNV association analysis can be categorized into two different strategies. First, the copy number is estimated by maximum likelihood and association of the expected copy number with the phenotype is tested. Second, the observed probe intensity measurements can be directly used to detect association of CNV with the phenotypes of interest.

Results: For each strategy we provide a statistic that can be applied to extended families. The computational efficiency of the proposed methods enables genome-wide association analysis and we show with simulation studies that the proposed methods outperform other existing approaches. In particular, we found that the first strategy is always more efficient than the second strategy no matter whether copy numbers for each individual are well identified or not. With the proposed methods, we performed genome-wide CNV association analyses of hematological trait, hematocrit, on 521 Korean family samples.

Conclusions: We found that statistical analysis with the expected copy number is more powerful than the statistic with the probe intensity measurements regardless of the accuracy of the estimation of copy numbers.

Keywords: CNV, Association analysis, Score test, Hematocrit

Background

Copy number variants (CNVs) are widely distributed throughout the human genome [1, 2] and have been considered as important genetic factors for human diseases [3, 4]. Several different methods, such as array comparative genomic hybridization (aCGH) and next generation sequencing, have been suggested to identify CNVs. Thanks to the recent improvement of sequencing technology, sequencing cost decreases very fast and becomes much cheaper. Furthermore, aCGH cannot detect aberrations such as mosaicism that do not result in copy number changes. However, in spite of this advantage of sequencing, aCGH is still cheaper and many array data have already been produced. Thus, it may be a cost effective choice at least for a while. In this report, we focus on CNV analysis with aCGH data–though the proposed method can be readily extended to other types of CNV data.

For aCGH data, gene copy numbers are not directly observed and have to be estimated with their intensity measures for association analyses. True unknown copy numbers will be called as unobserved copy numbers in the remainder of this report. CNV association requires estimation of copy numbers, and several algorithms, such as PennCNV [5], QuantiSNP [6], dChip [7] and GTC [8], have been developed to detect unobserved copy numbers. Then statistical methods such as linear

* Correspondence: kbj6181@cdc.go.kr; won1@snu.ac.kr
†Equal contributors
[3]Division of Structural and Functional Genomics, Center for Genome Science, National Institute of Health, Cheongju-si, Chungcheongbuk-do 363-951, South Korea
[4]Interdisciplinary Program of Bioinformatics, Seoul National University, Seoul 151-742, South Korea
Full list of author information is available at the end of the article

regression and chi-square tests have been utilized to detect CNV association with estimated copy numbers [9]. Barns et al. [10] calculated the posterior probability for each possible copy number, and likelihoods weighted by these posterior probabilities were used to build a likelihood ratio test. As an alternative to CNV analysis using the estimated copy number, the probe intensity measurements can be used to detect the CNV association [11]. The probe intensity is assumed to be proportional to the unobserved copy number, and its distributions can be compared between affected and unaffected individuals. If copy numbers are correctly estimated, the analysis using the expected copy numbers seems to be an efficient choice. However, estimates of copy numbers are often uncertain and this effect has not been carefully considered in the statistical analysis [4]. In this report, we considered both approaches and compared them with simulation studies for a large variety of parameter settings.

For association analysis, phenotypic correlations between individuals have the effect of sample size reduction, and thus independent population-based samples have often been preferred to maximize the statistical efficiency. However, family-based association analyses have been useful for certain scenarios because family members are genetically homogeneous [12, 13]. For instance, FBAT-statistics based on the so-called within-family component [14] are robust in the presence of population substructure and they are often preferred, in particular for candidate gene studies. Within-family component indicates the distribution of non-founders' genotype when their parental genotypes are conditioned. The distribution of founders' genotype is called between-family component and the statistical power of FBAT-statistics has been improved by combining FBAT-statistics with the between-family component in a robust way [11, 15, 16]. This two-stage analysis can achieve efficiency comparable to that of independent samples. However, due to the assumption that between- and within- family components are equally informative, this method can suffer from statistical power of loss if the numbers of founders and nonfounders are different.

In this report we propose two statistics, T_1 and T_2, for CNV association analysis using family-based samples; for T_1, the phenotypes are regressed on the expected copy number, and for T_2 they are regressed directly on the probe intensity measurements. A random effect is included to model the phenotypic covariance between family members, and the variance components for the phenotype are estimated with a restricted likelihood. Our results show that statistical analysis with the expected copy number is usually more efficient than the statistic with probe intensity measurements. We applied the proposed methods to detect CNV association with a hematology-related trait, hematocrit, in Korean family-based samples.

Methods

Notations and the disease model

We assume that K intensity measurements are observed at a particular CNV region for each individual, there are n families, and n_i individuals in family i. For simplicity, we consider only trio families, but the methods can be extended to large extended families. We assume that j = 1, 2 indicates the parents in each family. We let x_{ijk} indicate the observed intensity measurement on probe k for individual j in family i. $\mathbf{X}_{ij}$ indicates the column vector, $(x_{ij1},\ldots,x_{ijK})^T$, for individual j in family i. We let λ_{ij} be the unobserved copy number for individual j in family i, and denote a set of possible realizations of λ_{ij} and their corresponding frequencies respectively as $\boldsymbol{C}$ and Θ. We denote the phenotype for individual j in family i by y_{ij}, and let $\mathbf{Z}_{ij}$ be a vector of measured environmental factors, including an intercept as the first element. The intensity matrix, $\mathbf{X}_i$, and phenotype vector, $\mathbf{Y}_i$, for family i are respectively defined as $\mathbf{X}_i = \left(\mathbf{X}_{i1}^T, \ldots, \mathbf{X}_{in_i}^T\right)^T$ and $\mathbf{Y}_i = (y_{i1}, \ldots, y_{in_i})T$. We include a random effect, $\mathbf{b}_i$, to allow for the phenotypic correlation between family members. $\boldsymbol{\lambda}_i$ and $\boldsymbol{\varepsilon}_i$ indicate respectively an unobserved copy number vector and a measurement error vector for family members in family i. If we let $N = \Sigma_i\ n_i$, an $N \times K$ design matrix $\mathbf{X}$ and an $N \times 1$ vector $\mathbf{Y}$ are respectively obtained by stacking all $\mathbf{X}_i$ and $\mathbf{Y}_i$ vertically. $\boldsymbol{\lambda}$, $\mathbf{b}$ and $\boldsymbol{\varepsilon}$ are $N \times 1$ vectors and are obtained by stacking all $\boldsymbol{\lambda}_i$, $\mathbf{b}_i$ and $\boldsymbol{\varepsilon}_i$ vertically.

Signal model

We assumed that there are some correlations among the probe intensity measurements and the correlation matrix is assumed to be unstructured. We let $\boldsymbol{\gamma}_{\lambda_{ij}}$ and $\boldsymbol{\Sigma}_{\lambda_{ij}}$ be a $K \times 1$ mean vector and a $K \times K$ variance-covariance matrix of the intensity measurements. We assume that $\mathbf{X}_{ij}|\lambda_{ij}$ are identical and independently distributed for i and j, and

$$\mathbf{X}_{ij}|\lambda_{ij}\tilde{}\mathcal{N}\left(\boldsymbol{\gamma}_{ij}, , \boldsymbol{\Sigma}_{ij}\right).$$

If we assume that the correlation matrix is $\mathbf{R}$, the variance-covariance matrix can be expressed as $\boldsymbol{\Sigma}_{\lambda_{ij}} = \mathbf{D}_{\lambda_{ij}}$ $\mathbf{R}\mathbf{D}_{\lambda_{ij}}$, where

$$\mathbf{D}_{\lambda_{ij}} = \begin{bmatrix} \sigma_{1\lambda_{ij}} & 0 & \cdots \\ 0 & \sigma_{2\lambda_{ij}} & \mathbf{O} \\ \vdots & \mathbf{O} & \mathbf{O} \end{bmatrix}.$$

The parameters for $\boldsymbol{\Sigma}_{\lambda_{ij}}$ will be denoted by $\boldsymbol{\Sigma}$, and this proposed model will be called the signal model in the remainder of this manuscript.

Phenotype model

We assume that phenotypes are quantitative. We consider a standard linear mixed model for phenotypes that consists of CNV effects, additive polygenic effects, and measurement error. If we denote the $w \times w$ identity matrix by $\mathbf{I}_w$, the measurement error $\boldsymbol{\varepsilon}$ is assumed to follow the multivariate normal distribution with mean $\mathbf{0}$ and variance $\sigma_\varepsilon^2 \mathbf{I}_N$. The phenotypic correlations between family members are usually explained by a polygenic effect, $\mathbf{b}$, and we assume $\mathbf{b}$ follows a multivariate normal distribution. We let $\pi_{ijj'}$ be the kinship coefficient between individuals j and j' in family i, we let d_{ij} be the inbreeding coefficient for individual j in family i, we denote $\boldsymbol{\Phi}_i$ by the matrix

$$\begin{bmatrix} 1+d_{i1} & 2\pi_{i12} & \cdots & 2\pi_{i1n_i} \\ 2\pi_{i21} & 1+d_{i2} & \cdots & 2\pi_{i2n_i} \\ \vdots & \vdots & \ddots & \vdots \\ 2\pi_{in_i1} & 2\pi_{in_i2} & \cdots & 1+d_{in_i} \end{bmatrix},$$

and we let

$$\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\Phi}_1 & 0 & 0 & \cdots \\ 0 & \boldsymbol{\Phi}_2 & 0 & \cdots \\ 0 & 0 & \boldsymbol{\Phi}_3 & \cdots \\ \vdots & \vdots & \vdots & \mathrm{O} \end{bmatrix}.$$

The kinship coefficient between two subjects indicates the probability that two alleles randomly selected from each subject are identical by decent, and the inbreeding coefficient of a subject means the probability that his or her two alleles are identical by descent. Then, if we let the variance of the polygenic effect be σ_g^2, $\mathbf{b}$ follows the multivariate normal distribution with mean $\mathbf{0}$ and variance covariance matrix, $\sigma_g^2 \boldsymbol{\Phi}$. In the presence of population substructure, the empirical correlation matrix estimated with large-scale SNP data can replace $\boldsymbol{\Phi}$ to provide robustness to the proposed method [17, 18]. If we condition on the true copy number vector $\boldsymbol{\lambda}$, the linear model for the phenotype is

$$\mathbf{Y} = \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\lambda}\beta + \mathbf{b} + \boldsymbol{\varepsilon}, \text{ where } \mathbf{b} \sim \mathcal{N}\left(0, \sigma_g^2 \boldsymbol{\Phi}\right), \boldsymbol{\varepsilon} \sim \mathcal{N}\left(0, \sigma_\varepsilon^2 \mathbf{I}_N\right). \tag{1}$$

Copy number model

For disease copy number region we assume that there are M different unobserved copy numbers in the population. We further assume that the frequency of subjects with c_m copy numbers is θ_m in the population. We let $\boldsymbol{C} = \{c_1, \ldots, c_M\}$ and $\Theta = \{\theta_1, \ldots, \theta_M\}$, where $\theta_1 + \ldots + \theta_M = 1$. We denote maternal and paternal copy numbers of individual j in family i by λ_{ij}^1 and λ_{ij}^2 respectively, and we assume that λ_{ij} (= $\lambda_{ij}^1 + \lambda_{ij}^2$) for founders follows the multinomial distribution under Hardy-Weinberg equilibrium. It should be noted that λ_{ij} can be any element in $\boldsymbol{C}$. We assume no *de novo* CNVs and we assume that parental CNVs are transmitted to their offspring in a Mendelian fashion. For simplification, we consider nuclear families but the proposed method can be easily extended to the extended families. The probability of the ordered copy numbers for subjects in nuclear family i becomes

$$P\left(\left(\lambda_{i1}^1, \lambda_{i1}^2\right), \left(\lambda_{i2}^1, \lambda_{i2}^2\right), \ldots, \left(\lambda_{in_i}^1, \lambda_{in_i}^2\right)\right) = P\left(\left(\lambda_{i1}^1, \lambda_{i1}^2\right)\right) P\left(\left(\lambda_{i2}^1, \lambda_{i2}^2\right)\right) \times \prod_{j=1}^{n_i} P\left(\left(\lambda_{ij}^1, \lambda_{ij}^2\right) | \left(\lambda_{i1}^1, \lambda_{i1}^2\right), \left(\lambda_{i2}^1, \lambda_{i2}^2\right)\right).$$

Here, for $j = 1$ or 2,

$$P\left(\left(\lambda_{ij}^1, \lambda_{ij}^2\right)\right) = \begin{cases} \theta_m^2, & \mathit{if}\ \lambda_{ij}^1 = \lambda_{ij}^2 = c_m \\ 2\theta_m \theta_{m'}, & \mathit{if}\ \lambda_{ij}^1 = c_m, \lambda_{ij}^2 = c_{m'}, \\ & c_m \neq c_{m'} \end{cases}$$

and for $j = 3, \ldots, n_i$,

$$P\left(\left(\lambda_{ij}^1, \lambda_{ij}^2\right) | \left(\lambda_{i1}^1, \lambda_{i1}^2\right), \left(\lambda_{i2}^1, \lambda_{i2}^2\right)\right) = \begin{cases} 1/4, & \mathit{if}\ \lambda_{ij}^1 = \lambda_{i1}^l, \lambda_{ij}^2 = \lambda_{i2}^{l'}, \\ & l = 1, 2, l' = 1, 2 \\ 0, & \mathit{otherwise} \end{cases}.$$

We let Λ_{ij} be the set of possible maternal and paternal copy number pairs for individual j in family i, for which the sum is equal to λ_{ij}, as follows:

$$\Lambda_{ij} = \left\{ \left(\lambda_{ij}^{*1}, \lambda_{ij}^{*2}\right) | \lambda_{ij}^{*1} + \lambda_{ij}^{*2} = \lambda_{\mathrm{ij}} \right\}.$$

Then the joint probability of $\lambda_{i1}, \ldots, \lambda_{in_i}$ for individuals in family i is

$$P(\lambda_{i1}, \ldots, , \lambda_{in_i}) = \sum_{\left(\lambda_{i1}^{*1}, \lambda_{i1}^{*2}\right) \in \Lambda_{i1}} \cdots \sum_{\left(\lambda_{in_i}^{*1}, \lambda_{in_i}^{*2}\right) \in \Lambda_{in_i}} P\left(\left(\lambda_{i1}^{*1}, \lambda_{i1}^{*2}\right), \ldots, \left(\lambda_{in_i}^{*1}, \lambda_{in_i}^{*2}\right)\right).$$

If we assume that $\boldsymbol{\lambda}$ and $\mathbf{b}$ are missing values for the EM algorithm, our full likelihood is

$$f(\mathbf{X}, \mathbf{Y}, \boldsymbol{\lambda}, \mathbf{b} | \mathbf{Z}, \boldsymbol{\Phi}, \boldsymbol{\alpha}, \beta, \boldsymbol{\gamma}, \boldsymbol{\Sigma}, \boldsymbol{\Theta}) = f(\mathbf{X} | \boldsymbol{\lambda}; \boldsymbol{\gamma}, \boldsymbol{\Sigma}) \cdot f\left(\mathbf{Y}, \mathbf{b} | \mathbf{Z}, \boldsymbol{\lambda}; \boldsymbol{\alpha}, \beta, \sigma_g^2, \sigma_\varepsilon^2\right) \cdot f(\boldsymbol{\lambda} | \boldsymbol{\Theta}) \cdot \tag{2}$$

Parameter estimation with the EM algorithm

To derive a score test for CNV association analysis, β in the phenotype model was assumed to be 0, and the variance component parameters were estimated with the restricted maximum likelihood (REML) method. The copy number vector $\boldsymbol{\lambda}$ and the random effect vector $\mathbf{b}$ are considered as missing variables for the EM algorithm, and the conditional expectation of a complete data log-likelihood was maximized to estimate all the parameters. Individuals were separated with K-means clustering [19], and the empirical mean and co-variance matrix were used as the initial values for the signal model.

In the expectation step, we calculate posterior probabilities for each possible value of the unobserved copy number using the estimates from the previous iteration. We use the superscript (ω) to indicate the estimate at the ω-th iteration. The posterior probability of $\boldsymbol{\lambda}$ is obtained by

$$\begin{aligned}&P\left(\boldsymbol{\lambda}|\mathbf{X},\mathbf{Y},\mathbf{Z},\boldsymbol{\Phi},\boldsymbol{\alpha}^{(\omega)},\hat{\beta}^{(\omega)},\hat{\boldsymbol{\gamma}}^{(\omega)},\hat{\boldsymbol{\Sigma}}^{(\omega)},\hat{\boldsymbol{\Theta}}^{(\omega)}\right)\\&=\frac{f\left(\mathbf{X},\mathbf{Y},\boldsymbol{\lambda}|\mathbf{Z},\boldsymbol{\Phi},\hat{\boldsymbol{\alpha}}^{(\omega)},\hat{\beta}^{(\omega)},\hat{\boldsymbol{\gamma}}^{(\omega)},\hat{\boldsymbol{\Sigma}}^{(\omega)},\hat{\boldsymbol{\Theta}}^{(\omega)}\right)}{\sum_{\lambda'}f\left(\mathbf{X},\mathbf{Y},\boldsymbol{\lambda}'|\mathbf{Z},\boldsymbol{\Phi},\hat{\boldsymbol{\alpha}}^{(\omega)},\hat{\beta}^{(\omega)},\hat{\boldsymbol{\gamma}}^{(\omega)},\hat{\boldsymbol{\Sigma}}^{(\omega)},\hat{\boldsymbol{\Theta}}^{(\omega)}\right)}.\end{aligned}$$

Under the null hypothesis, this posterior probability becomes

$$\begin{aligned}&P\left(\boldsymbol{\lambda}|\mathbf{X},\mathbf{Y},\mathbf{Z},\boldsymbol{\Phi},\hat{\boldsymbol{\alpha}}^{(\omega)},\beta=0,\hat{\boldsymbol{\gamma}}^{(\omega)},\hat{\boldsymbol{\Sigma}}^{(\omega)},\hat{\boldsymbol{\Theta}}^{(\omega)}\right)\\&=\frac{f\left(\mathbf{X}|\boldsymbol{\lambda};\hat{\boldsymbol{\gamma}}^{(\omega)},\hat{\boldsymbol{\Sigma}}^{(\omega)}\right)f\left(\boldsymbol{\lambda}|\hat{\boldsymbol{\Theta}}^{(\omega)}\right)}{\sum_{\lambda'}f\left(\mathbf{X}|\boldsymbol{\lambda}';\hat{\boldsymbol{\gamma}}^{(\omega)},\hat{\boldsymbol{\Sigma}}^{(\omega)}\right)f\left(\boldsymbol{\lambda}'|\hat{\boldsymbol{\Theta}}^{(\omega)}\right)}.\end{aligned}$$

The copy number with the largest posterior density was assumed to be the true copy number $\hat{\boldsymbol{\lambda}}^{(\omega+1)}$ for each individual in the $(\omega+1)$-th iteration. For the missing variable $\mathbf{b}$, if we let $\mathbf{V}=\sigma_g^2\boldsymbol{\Phi}+\sigma_\varepsilon^2\mathbf{I}_N$ and $\mathbf{e}=\mathbf{Y}-\mathbf{Z}\boldsymbol{\alpha}-\boldsymbol{\lambda}\beta$, the posterior mean of $\mathbf{b}$ in the $(\omega+1)$-th iteration is estimated as

$$\hat{\mathbf{b}}^{(\omega+1)}=\hat{\sigma}_g^{(\omega)2}\boldsymbol{\Phi}\hat{\mathbf{V}}^{(\omega)-1}\hat{\mathbf{e}}^{(\omega)}.$$

In the maximization step, all parameters are estimated by maximizing the expected log-likelihood of

$$f\left(\mathbf{X},\mathbf{Y},\boldsymbol{\lambda}^{(\omega)},\mathbf{b}^{(\omega)}|\mathbf{Z},\boldsymbol{\Phi},\hat{\boldsymbol{\alpha}}^{(\omega)},\hat{\beta}^{(\omega)},\hat{\boldsymbol{\gamma}}^{(\omega)},\hat{\boldsymbol{\Sigma}}^{(\omega)},\hat{\boldsymbol{\Theta}}^{(\omega)}\right).$$

$\boldsymbol{\gamma}$ and $\boldsymbol{\Sigma}$ are updated by the sample mean and sample variance-covariance matrix. $\boldsymbol{\alpha}$ and β in the phenotype model are estimated by

$$\begin{aligned}&\hat{\boldsymbol{\alpha}}^{(\omega+1)},\hat{\beta}^{(\omega+1)})\\&=\left[\begin{bmatrix}Z\\\hat{\boldsymbol{\lambda}}^{(\omega)}\end{bmatrix}\hat{\mathbf{V}}^{(\omega)-1}\begin{pmatrix}\mathbf{Z}&\hat{\boldsymbol{\lambda}}^{(\omega)}\end{pmatrix}\right]^{-1}\begin{bmatrix}Z\\\hat{\boldsymbol{\lambda}}^{(\omega)}\end{bmatrix}\hat{\mathbf{V}}^{(\omega)-1}\mathbf{Y}.\end{aligned}$$

The variance parameters, σ_g^2 and σ_ε^2, are updated as

$$\begin{aligned}\hat{\sigma}_g^{(\omega+1)2}&=\hat{\sigma}_g^{(\omega)2}\\&\quad+\frac{1}{N}\mathrm{tr}\left(\hat{\mathbf{b}}^{(\omega)}\hat{\mathbf{b}}^{(\omega)T}-\hat{\sigma}_{g^{(\omega)4}}\boldsymbol{\Phi}\hat{\mathbf{P}}^{(\omega)}\boldsymbol{\Phi}\right)\hat{\sigma}_\varepsilon^{(\omega+1)2}\\&=\hat{\sigma}_\varepsilon^{(\omega)2}+\frac{1}{N}\mathrm{tr}\left(\hat{\mathbf{e}}^{(\omega)}\hat{\mathbf{e}}^{(\omega)T}-\hat{\sigma}_\varepsilon^{(\omega)4}\hat{\mathbf{P}}^{(\omega)}\right),\end{aligned}$$

where $\hat{\mathbf{P}}^{(\omega)}=\hat{\mathbf{V}}^{(\omega-1)-1}-\hat{\mathbf{V}}^{(\omega-1)-1}\mathbf{X}\left(\mathbf{X}^T\hat{\mathbf{V}}^{(\omega-1)-1}\mathbf{X}\right)^{-1}\mathbf{X}^T\hat{\mathbf{V}}^{(\omega-1)-1}$. Last, θ_k is updated with the following best linear unbiased estimator [20]:

$$\begin{aligned}\hat{\theta}_k^{(\omega+1)}&=\frac{1}{2}\left(1_N^T\boldsymbol{\Phi}^{-1}1_N\right)^{-1}1_N^T\boldsymbol{\Phi}^{-1}\\&\quad\times\begin{bmatrix}P\left(\lambda_{11}^1=c_k|\mathbf{X},\mathbf{Y},\mathbf{Z},\boldsymbol{\Phi},\hat{\boldsymbol{\alpha}}^{(\omega)},\hat{\beta}^{(\omega)},\hat{\boldsymbol{\gamma}}^{(\omega)},\hat{\boldsymbol{\Sigma}}^{(\omega)},\hat{\boldsymbol{\Theta}}^{(\omega)}\right)\\\vdots\\P\left(\lambda_{nn_n}^1=c_k|\mathbf{X},\mathbf{Y},\mathbf{Z},\boldsymbol{\Phi},\hat{\boldsymbol{\alpha}}^{(\omega)},\hat{\beta}^{(\omega)},\hat{\boldsymbol{\gamma}}^{(\omega)},\hat{\boldsymbol{\Sigma}}^{(\omega)},\hat{\boldsymbol{\Theta}}^{(\omega)}\right)\end{bmatrix}.\end{aligned}$$

Identifying the number of clusters

The optimal M was chosen with the silhouette score which quantifies whether objects in the same cluster stay together and objects in different clusters are well separated [21]. We denote the Euclidean distance between $\mathbf{X}_{ij}$ and $\mathbf{X}_{i^{'}j^{'}}$ by $d_{ij,i'j'}$, and denote the number of individuals whose copy numbers are c_m by $n(c_m)$. If the estimated copy number $\hat{\lambda}_{ij}$ for individual j in family i is assumed to be c_m, we let the average distance to the rest of the cluster be

$$a_{ij}=\frac{1}{n(c_m)}\sum_{\left\{(i^{'},j^{'})|\hat{\lambda}_{i^{'}j^{'}}=c_m\right\}}d_{ij,i^{'}j^{'}},$$

and the minimum average distance to other clusters be

$$b_{ij}=\min\left\{\frac{\sum_{\{(i',j')|\hat{\lambda}_{i'j'}=c_m\}}d_{ij,i'j'}}{n(c_{m^{'}})}\middle|m'\neq m,\ m'=1,\ldots,M\right\}.$$

Then the silhouette score for individual j in family i is defined as

$$sil_{ij}=\frac{b_{ij}-a_{ij}}{\max\left\{a_{ij},,b_{ij}\right\}}.$$

If sil_{ij} is close to one, it indicates that the corresponding individual is well-clustered, whereas if sil_{ij} is close to −1, it means that the individual is badly clustered. If sil_{ij} is close to zero, there may exist a better cluster for the corresponding individual. Therefore, we first estimated the copy numbers for each individual for different choices of M. Then we calculated silhouette scores for all the individuals, and the value of M that maximized the mean silhouette score was considered as the optimal choice.

Statistical inference

The Wald and likelihood ratio tests for the proposed likelihood are computationally intensive, and CNV association analysis with large families may not be feasible on a genome-wide scale. Therefore, we provide two score statistics based on Eq. (2); one is based on the estimated copy number and the other is based on the probe intensity measurement itself. First, the copy numbers and parameters for variance components are estimated from the likelihood under the null hypothesis. The expected copy numbers are assumed to be the unknown true copy numbers. Then Rao's score test statistic is

$$T_1 = \left((\mathbf{Y}-\mathbf{Z}\boldsymbol{\alpha})^T \mathbf{V}^{-1} \boldsymbol{\lambda} \right)^T \times \left(\boldsymbol{\lambda}^T \mathbf{V}^{-1} \boldsymbol{\lambda} - \left(\mathbf{Z}^T \mathbf{V}^{-1} \boldsymbol{\lambda} \right)^T \left(\mathbf{Z}^T \mathbf{V}^{-1} \mathbf{Z} \right)^{-1} \left(\mathbf{Z}^T \mathbf{V}^{-1} \boldsymbol{\lambda} \right) \right)^{-1} \times \left((\mathbf{Y}-\mathbf{Z}\boldsymbol{\alpha})^T \mathbf{V}^{-1} \boldsymbol{\lambda} \right),$$

and T_1 follows the chi-square distribution with a single degree of freedom under H_0 (See Additional file 1: Text 1 for details). If there exists no inverse matrix of $\mathbf{V}$, the generalized inverse matrix [22] can be utilized.

However, T_1 is based on the estimates of the expected copy numbers and its performance may depend on the accuracy of $\hat{\boldsymbol{\lambda}}$. We therefore also provide the statistic T_2, based directly on the probe intensity measurements. It should be noted that, contrary to T_1, T_2 does not need one to estimate the unknown copy number and the computation is less intensive. We let $\boldsymbol{\Psi}$ be the empirical variance-covariance matrix between individuals and $\mathbf{I}_N$ be the $N \times N$ dimensional identical matrix,

$$\mathbf{v} = \frac{\mathrm{tr}\left((\mathbf{I}_N - \mathbf{S}_2) \boldsymbol{\Psi} (\mathbf{I}_N - \mathbf{S}_2)^T \mathbf{V}^{-1} (\mathbf{I}_N - \mathbf{S}_1) \right)}{\mathrm{tr}\left((\mathbf{I}_N - \mathbf{S}_2) \boldsymbol{\Psi} \boldsymbol{\Phi}^{-1} \right)} \times \mathbf{X}^T (\mathbf{I}_N - \mathbf{S}_2)^T \boldsymbol{\Phi}^{-1} (\mathbf{I}_N - \mathbf{S}_2) \mathbf{X},$$

and

$$\mathbf{u}^T = \mathbf{Y}^T (\mathbf{I}_N - \mathbf{S}_1)^T \mathbf{V}^{-1} (\mathbf{I}_N - \mathbf{S}_2) \mathbf{X},$$

where

$$\mathbf{S}_1 = \mathbf{Z}\left(\mathbf{Z}^T \mathbf{V}^{-1} \mathbf{Z} \right)^{-1} \mathbf{Z}^T \mathbf{V}^{-1}, \text{ and } \mathbf{S}_2 = 1_N \left(1_N^T \boldsymbol{\Phi}^{-1} 1_N \right)^{-1} 1_N^T \boldsymbol{\Phi}^{-1}.$$

If we denote the rank of $\mathbf{v}$ by r, T_2 is defined by

$$T_2 = \mathbf{u}^T \mathbf{v}^{-1} \mathbf{u} \sim \chi^2 (df = r) \text{ under } H_0.$$

The detailed derivation of T_2 is shown in Additional file 1: Text 2. In particular, we can utilize a transformed value for $\mathbf{X}$ in T_2. For instance, the mean intensity measurement over all probes or the first principal component (PC) score can be utilized, and then T_2 follows the chi-square distributions with a single degree of freedom. Implementation of the methods is assembled in an R package PedCNV, which is available from CRAN.

Simulation studies

Data generation

We conducted simulation studies to evaluate the performance of the proposed methods and, for computational simplicity, we simulated just 300 parent-offspring trios. We considered two scenarios; (1) $M = 3$, $\boldsymbol{C} = \{0, 1, 2\}$ and $\Theta = \{(1-\theta)^2, 2\theta(1-\theta), \theta^2\}$, and (2) $M = 6$, $\boldsymbol{C} = \{0, 1, 2, 3, 4, 5\}$ and $\Theta = \{(1-\theta)^5, 5\theta(1-\theta)^4, 10\theta^2(1-\theta)^3, 10\theta^3(1-\theta)^2, 5\theta^4(1-\theta), \theta^5\}$. Copy numbers for offspring were generated with simulated Mendelian transmission. We assumed $K = 7$ probe intensities were measured for a CNV region, and each intensity, x_{ijk}, was generated from a normal distribution with

$$E(x_{ijk}) = \begin{cases} s_k \lambda_{ij} + p_{\lambda_{ij}}, & k = 1,2,3 \\ s_k + p_{\lambda_{ij}}, & k = 4,5,6,7 \end{cases}, \mathrm{var}(x_{ijk}) = \left(z \cdot \lambda_{ij} + q_v \right)^2.$$

Here the dissimilarity between probe intensity measurements in different clusters for probe k is proportional to the value of s_k and we considered three scenarios by using three different choices of s_k: badly separated clusters (BSC), moderately separated clusters (MSC) and well separated clusters (WSC). The different means for the probe intensities were provided by $p_{\lambda_{ij}}$ generated from $N(0, 0.9(\lambda_{ij} + 1.5)^2)$. The variance of each probe intensity measurement was provided by q_v generated from $\Gamma(0.025, 0.0016^2)$. The parameter settings in the signal model are described in Additional file 1: Table S1.

Phenotypes were generated based on Eq. (1). For the phenotype model, we assumed that there was a single covariate for $\mathbf{Z}$ which was independently generated for each individual from the standard normal distribution. σ_ε^2 and σ_g^2 were assumed to be 1. For our simulations, we considered trios and $\boldsymbol{\Phi}_i$ becomes

$$\begin{bmatrix} 1 & 0 & .5 \\ 0 & 1 & .5 \\ .5 & .5 & 1 \end{bmatrix}.$$

Analysis of a hematological trait

Subjects

Hematocrit indicates the volume percentage of red blood cells in blood and red blood cells transfer oxygen from the lungs to body tissues. Some diseases such as anemia are related to hematocrit and we conducted association analyses of hematocrit to identify CNVs related to anemia. We used the same DNA samples as were used in Lee et al. [23]. Five hundred fifty-one individuals from 59 families including 216 Granular corneal dystrophy type 2 patients and 324 unaffected controls were genotyped with Illumina HumanCNV 370 K-Duo Beadchip. Clinical information for 30 individuals was missing. Therefore, 521 individuals were used for the association analysis. All subjects enrolled in this study were of Korean ethnicity. Basic characteristics of our samples are summarized in Table 1.

Table 1 Basic characteristics of study participants and hematological trait

Variables	Discovery (family)	Replication (cohort)
Sample size (n)	521	4694
Age (years)	38.2 ± 18.3	54.0 ± 9.0
Male (%)	45.7%	47.1%
Hematocrit (%)	41.3 ± 4.3	41.1 ± 4.5

CNV discovery

All samples were genotyped with NimbleGen HD2 3 × 720 K aCGH which contains more than 720,000 probes. Around 360,000 probes were designed based on previously reported CNVs, and the other probes were spaced uniformly throughout the whole genome as a backbone. Sample NA10851obtained from the HapMap lymphoblastoid cell line (LCL) DNA was used as a reference, and NimbleScan version 2.5 was used to process the array image files (.tif) according to the manufacturer's protocol. Extracted signal intensity was transformed to log2 ratio with hg18/NCBI build 36. Subsequently, we set the log2 ratio thresholds less than −0.25 for a deletion and greater than 0.25 for a duplication, with more than 10 consecutive probes required to assign a CNV.

CNV selection

We used a reciprocal overlap threshold > 50% to find CNVs with similar boundaries for association analysis. According to this threshold, clusters of overlapping CNVs at the sample level are merged into one CNV. Overlapping CNVs with very different sizes and sequentially connected CNVs were excluded from further study. Moreover, we selected CNV clusters which are well-separated and have multi-class CNVs in order to assign individuals to copy-number classes with high confidence [24]. In total 500 CNVs were utilized for association analyses.

CNV association

PedCNV was applied to an association study with a hematological trait: hematocrit (Hct). The association of CNVs with Hct was analyzed using T_1 and T_2, with age, age^2 and sex included as covariates. The resulting statistics were adjusted by using genomic control to allow for population substructure.

CNV validation by PCR experiment

To confirm CNV genotypes, a PCR using the AccuPrime Taq DNA Polymerase High Fidelity (invitrogen, CA, USA) was performed on 10–16 individuals selected from each cluster (Additional file 1: Figure S1 (A)). The primers were designed to give rise to amplicons with different lengths to detect both the deleted (690 bp) and normal (1519 bp) alleles (Additional file 1: Table S2). Genomic locations for designed primers based on human genome assembly hg18 were converted to those based on hg19 by liftOver of the UCSC genome browser. PCR was carried out on a GeneAmp PCR system 9700 (Applied Biosystems, Calif., USA) with the following PCR conditions: 5 min at 95 °C, followed by 33 cycles of 30s at 95 °C, 30s at 60 °C, 2 min at 68 °C, and final extension at 68 °C for 7 min. The resulting PCR products were visualized by electrophoresis separation on a 1.5% agarose gel with Safe-Pinky DNA gel staining solution (Genedepot, TX, USA). Moreover, to confirm exact break-points of the CNVs, PCR products were sequenced using an ABI 3730 DNA analyzer (Applied Biosystems, CA, USA).

Replication study

We have previously implemented KGVDB, which includes 3601 multi-class CNVs and their tagging SNPs, from 4694 community-based cohort samples, as a part of the Korean Genome Epidemiology Study (KoGES) [25]. We used these unrelated individual samples to pursue replication of the identified CNV from the discovery association study. Table 1 shows a summary of the participants' characteristics. In short, all the 4694 samples were also genotyped with NimbleGen HD2 3 × 720 K aCGH. The NA10851 sample was again used as a reference. NimbleScan version 2.5 was used to extract signal intensity. Subsequently, quality control, such as normalization and waviness correction, was conducted using the R package (http://cran.r-project.org) and WaveNorm [26]. For CNV detection, the Genome Alteration Detection Analysis algorithm (GADA) was used with T = 10, alpha = 0.2 and MinSegLen = 10. Moreover, an average log2 ratio of ±0.25 was set as a cut-off value [25]. Among the detected CNVs, we selected those CNVs having a similar boundary with any CNV significant in the discovery association study. Additional file 1: Figure S2 shows the overall process of the replication study.

CNV validation of replication study samples

To verify whether an estimated CNV genotype using cohort samples is true or not, we carried out quantitative PCR (qPCR) using the TaqMan Copy number Assay (Life Technologies, Foster City, CA, USA) according to the manufacturer's guidelines. A pre-designed TaqMan probe (Assay ID: Hs04965547_cn) was used to validate the existence of the CNV. All experiments were replicated three times to enhance the validation accuracy. The samples used for validation were randomly selected from each genotype (Additional file 1: Figure S1 (B)). Copy number genotype for each sample was calculated by Copy caller v2.0 (Applied Biosystems, Calif., USA) using the manufacturer's guideline.

Results

Evaluation with simulated data

Clustering

With the simulated data we evaluated the accuracy of estimating M and the estimated copy numbers for each individual when the true M was assumed to be known. The results from the proposed method were compared with CNVtools [10]. The probe intensity measurements were generated under the three different scenarios: BSC, MSC and WSC. For each individual, we calculated from the probe intensity measurements the mean, the first PC score and the fewest PC scores that explain more than 90% of the variation; they are denoted by mean, PC1 and PC.9 respectively. In addition to the original probe intensity measurements (RAW), we used the mean, PC1 and PC.9, for the proposed method and the results were compared with CNVtools. For CNVtools, the mean, PC1 and the one-dimensional canonical correlation transformed vector of the probe intensity measurements were used.

Additional file 1: Tables S3 and S4 show the accuracy of the estimated value of M from 1000 replicates using PedCNV and CNVtools. The proposed method using PC1 was always the most accurate, followed by the proposed method using PC.9. The results from the proposed method performed better than CNVtools. CNVtools had a tendency to choose a larger number of clusters, and the results were rarely consistent, even when the clusters were well separated. CNVtools selects the number of clusters using a Bayesian information criterion [10], while the proposed method selects it with a silhouette score, which appears to be a better choice. In Additional file 1: Table S5, M was set to be the true value 3 for all methods, and the relative proportions of individuals for whom the estimated copy number was consistent with the true copy number were calculated from 1000 replicates under the null and alternative hypotheses. Additional file 1: Table S5 shows that the proposed method based on PC1 was the most accurate, followed by the proposed method based on PC.9. Therefore we conclude that basing our method on PC1 may be a reasonable choice.

Association analysis

In order to evaluate the proposed statistics T_1 and T_2, we simulated the probe intensity measures for BSC, MSC and WSC, and phenotypes were generated under the null and alternative hypotheses. Seven probe intensity measurements were generated, so that T_2 followed the chi-square distribution with seven degrees of freedom under the null hypothesis. For the statistical validity of the proposed methods, empirical type-1 error estimates at the various significance levels were calculated from 5000 replicates; Table 2 shows that for our methods the nominal significance levels were always preserved under BSC, MSC and WSC. The quantile quantile (QQ) plots in Fig. 1 also indicate the validity of T_1 and T_2.

To evaluate statistical efficiency, empirical power estimates for T_1 and T_2 were calculated from 2000 replicates under the alternative hypothesis, and compared with the FBAT statistic which directly utilizes the intensity measurement [11]. We considered various choices of β, and the probe intensity measurements for BSC, MSC and WSC were generated. Table 3 shows that the proposed statistics T_1 and T_2 performed better than FBAT, and T_1 was always more powerful than T_2 under all scenarios. The power loss of T_2 compared to T_1 is the largest for BSC and we can conclude that the statistical power of T_2 is more affected by proportions of noise in the probe intensity measurement. We calculated empirical type 1 error and power estimates when $M = 6$ in Additional file 1: Tables S6 and S7, and the same patterns as $M = 3$ are observed. Moreover, we compared T_1 with the statistic with the most probable copy number, and found that T_1 is more powerful to estimate the parameters, especially under the BSC scenario (Additional file 1: Tables S8 and S9). Therefore, we conclude that T_1 should be selected for a CNV association analysis.

Table 2 Empirical type 1 error estimates ($M = 3$)

		Significance Level			
		.005	.05	.1	.2
BSC	T1	0.0060 ± 0.0021	0.0504 ± 0.0061	0.1018 ± 0.0084	0.2082 ± 0.0113
	T2	0.0056 ± 0.0021	0.0550 ± 0.0063	0.1008 ± 0.0086	0.2072 ± 0.0112
MSC	T1	0.0048 ± 0.0019	0.0486 ± 0.0060	0.1006 ± 0.0083	0.2104 ± 0.0113
	T2	0.0048 ± 0.0019	0.0472 ± 0.0059	0.0956 ± 0.0082	0.1884 ± 0.0108
WSC	T1	0.0056 ± 0.0021	0.0516 ± 0.0061	0.0962 ± 0.0082	0.2006 ± 0.0111
	T2	0.0048 ± 0.0019	0.0498 ± 0.0060	0.0968 ± 0.0082	0.1922 ± 0.0109

The 95% confidence intervals of empirical type I error estimates for the proposed methods were calculated from 5000 replicates at four significance levels under BSC, MSC and WSC, when there are three copy number clusters

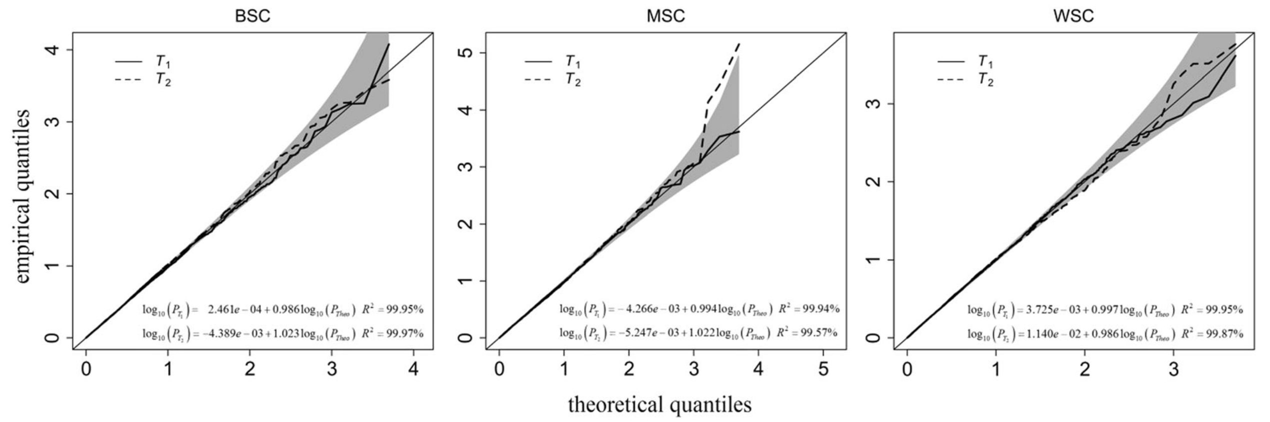

Fig. 1 The QQ plots without genome control for T_1 and T_2 from simulated data. The empirical *p*-values adjusted by genomic control for the proposed methods were calculated under the null hypothesis with 5000 replicates under BSC, MSC and WSC, and their QQ plots are shown

Table 3 Empirical power estimates ($M = 3$)

Significance Level			β					
			.1	.2	.3	.4	.5	.6
.001	BSC	T1	0.0135	0.1390	0.4830	0.8410	0.9830	0.9985
		T2	0.0065	0.0510	0.2370	0.5745	0.8860	0.9795
		FBAT	5e-4	0.0060	0.0340	0.1300	0.3570	0.6005
	MSC	T1	0.0160	0.1570	0.5530	0.8740	0.9885	1.0000
		T2	0.0085	0.0685	0.3200	0.6900	0.9290	0.9945
		FBAT	0.0000	0.0100	0.0695	0.2505	0.5575	0.8385
	WSC	T1	0.0195	0.1615	0.5375	0.8935	0.9910	0.9980
		T2	0.0075	0.0815	0.3300	0.7240	0.9545	0.9970
		FBAT	0.0010	0.0115	0.0915	0.3240	0.6460	0.8955
.01	BSC	T1	0.0710	0.3585	0.7510	0.9605	0.9990	1.0000
		T2	0.0265	0.1780	0.4630	0.7900	0.9685	0.9950
		FBAT	0.0155	0.0455	0.1540	0.3775	0.6620	0.8430
	MSC	T1	0.0725	0.3805	0.8070	0.9690	0.9990	1.0000
		T2	0.0340	0.2100	0.5640	0.8660	0.9790	0.9980
		FBAT	0.0150	0.0550	0.2400	0.5385	0.8155	0.9645
	WSC	T1	0.0800	0.3795	0.7925	0.9740	0.9985	1.0000
		T2	0.0370	0.2115	0.5730	0.8855	0.9920	0.9995
		FBAT	0.0175	0.0710	0.2740	0.6350	0.8775	0.9740
.05	BSC	T1	0.1905	0.5930	0.9075	0.9920	1.0000	1.0000
		T2	0.0975	0.3505	0.6880	0.9080	0.9910	0.9990
		FBAT	0.0575	0.9610	0.3660	0.6310	0.8555	0.9525
	MSC	T1	0.2050	0.6190	0.9295	0.9900	1.0000	1.0000
		T2	0.1110	0.3915	0.7650	0.9495	0.9970	1.0000
		FBAT	0.0725	0.2080	0.4990	0.7585	0.9305	0.993
	WSC	T1	0.2095	0.6145	0.9260	0.9950	0.9995	1.0000
		T2	0.1255	0.3995	0.7650	0.9530	0.9990	1.0000
		FBAT	0.0790	0.2240	0.5460	0.8415	0.9705	0.9960

The empirical power for the proposed methods have been estimated at various significance levels based on 2000 replicates for different values of β under BSC, MSC and WSC, when there are three copy number clusters. The score test using the inferred CNVs is denoted by T_1. The score test using the intensity measurements is denoted by T_2. For comparison, we also calculated the power using FBAT

Results of real data analysis

CNV association

500 well-separated multi-class CNVs were chosen for an association study. The 0.05 genome-wide significance level by Bonferroni correction for 500 CNVs is 10^{-4} and association analyses of Hct were conducted with the proposed methods. Figure 2 shows QQ and Manhattan plots for the statistics T_1 and T_2. We listed the most significant results of T_1 and T_2 respectively in Table 4. There is no genome-wide significant CNV and this is partially attributable to the insufficient sample size. In our analyses, 521 subjects are utilized, and if the effect size is 0.206 and sigma is 0.957, 1563 subjects are required to achieve 0.8 power at the 10^{-4} significant level. The difference between T_1 and T_2 may be attributable to the low accuracy of the clustering, because the performance of T_1 depends on the accuracy of the clustering. However, T_2 models the relationship between intensity and phenotypes without estimating copy numbers; but there is also the possibility of poor fit, including nonnormality.

CNV validation

Among 500 multi-class CNVs, the CNV region (chr7:81279592–81280418) was randomly selected for evaluation of CNV genotype estimation. In total, 41 subjects were selected from each CNV cluster and a PCR experiment was conducted for them. Among these samples, 38 subjects (92.7%) had the same copy numbers as the estimates from the proposed methods (Additional file 1: Figures S3 and S4).

Discussion

Even though CNV has been expected to be an important genetic factor for many diseases, CNV association analysis has often been limited because of uncertainty of the copy number, and several statistical methods [8, 27] have been proposed to handle this uncertainty. However, even though some of the existing methods are relatively accurate, the estimated copy numbers are not accurate in some situations, which might cause a power loss for CNV association analysis. In this report, we propose new statistical methods for CNV association analysis with family-based samples. With extensive simulations, we showed that the proposed methods perform much better than the existing approaches. The proposed method was implemented in the R package, pedCNV and the main function in our R package was implemented with C++. We found that association analyses of 300 trios were completed within one minute using an Intel (R) Xeon (R) E5-2620 0 CPU at 2.00GHz, with a single node and 80 gigabyte memory.

Furthermore, the proposed method is flexible and can be extended to various scenarios. First, the proposed methods consist of T_1 and T_2. The former is based on the estimated copy number and the latter is on the probe intensity measurements. Our simulation studies show that the most efficient statistic is always the statistic with the expected copy numbers. However, if the accuracy of the estimated copy numbers is not clear and there is a systematic bias, the statistical power of T_1 can be substantially affected, and some modification can be made to the proposed methods. For instance, the minimum of the *p*-values for T_1 and T_2 could be considered as a test statistic and permutation-based *p*-values could be calculated. Alternatively, the posterior probabilities for each copy number estimated from the E step in T_1 can be utilized as classified copy numbers. These modifications are computationally feasible and may provide

Table 4 The most significant results of T_1 and T_2 from analyzing the family data

Chr	Position	0/1/2	T_1	T_2
8	94141469–94142527	42/226/253	1.38e-03	4.67e-02
5	147534018–147534337	119/265/137	1.19e-02	3.92e-03

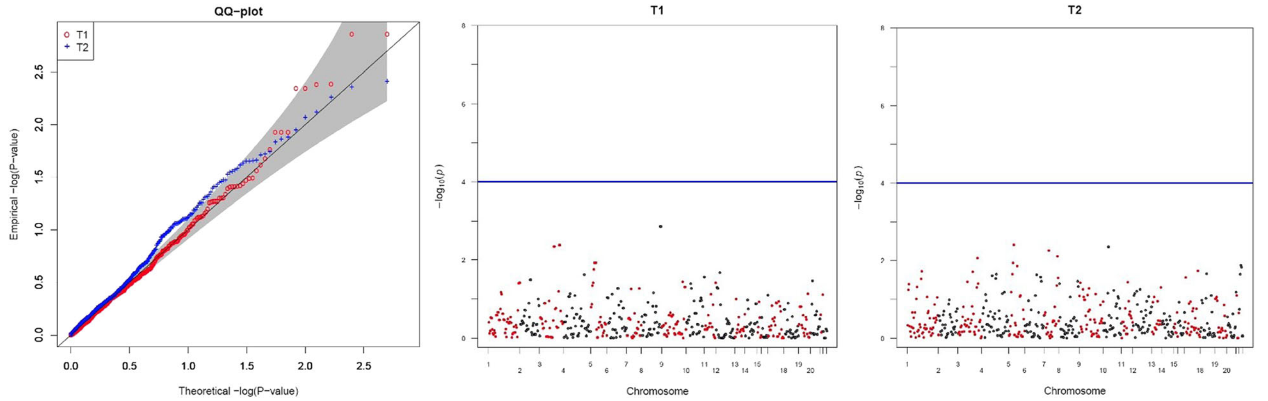

Fig. 2 The QQ plot and Manhattan plots for T_1 and T_2 from analysis of the family data

less sensitive results compared to T_1 and T_2. Second, the presence of population substructure has been known to be a factor that leads to violation of the assumptions underlying statistical association analysis. In our real data analysis, the genomic control approach [28] was adopted, but the linear mixed model is known to be the most efficient if the polygenic effects are substantial [29]. The correlations between individuals can be estimated with large-scale genetic data such as genome-wide SNPs, and this can be incorporated into the phenotype model in the proposed method. Third, the proposed methods can be simply extended to the sequencing data with a minor modification even though it only applied to aCGH data in this report. This will be investigated in our future work.

However, in spite of the practical advantage of the proposed methods, there exist some limitations, and further investigation is necessary. First, the incorporation of Mendelian transmission into the signal model induces a substantial computational burden for large families. In our PedCNV package, Mendelian transmission for a signal model is considered, but only for nuclear families. We found with simulation studies that the drop of accuracy is not substantial when Mendelian transmission is not considered, but its effect can be substantial if only a few large families are available. A peeling algorithm [30] has been developed that minimizes the computation of likelihoods for large families and it will be implemented in the PedCNV package. Second, the proposed method assumes that there is no de novo mutation and recombination. In such cases, the statistic T_2 may be a better choice. Third, it has been observed that the bias in CNV calls can be different between parents and offspring, and our first statistic, T_1, can suffer from this differential bias. Our simulation studies do not examine any such violation of statistical assumptions, but its effect on T_1 could be substantial in CNV association analysis with large families. Third, copy numbers for each individual were identified by calculating the expectation of copy numbers using the posterior probability and the expected copy numbers were utilized as λ in T_1. Although this maximum likelihood approach for classification can yield inconsistent estimators of parameters [31, 32], the simulation studies show that the accuracy of this method is higher. Thus we continued adopting this method in spite of its deficiencies.

In recent decades various types of genetic data have been used to detect the genetic factors underlying many diseases and many disease susceptibility loci have been found. Even though CNVs have been expected to be an important genetic factor, the findings of CNV association analysis have been limited and the proposed methods may bridge this gap by alleviating the issue of copy number uncertainty.

Conclusion

PedCNV presents a computationally efficient R package that provides two statistics for family-based CNV association analysis: first, the copy number is estimated by maximum likelihood and association of the estimated copy number with the phenotype is tested; second, the observed probe intensity measurements is directly used to detect association of CNV with the phenotypes of. The simulation studies showed that the proposed methods outperform other existing approaches. In particular, we found that statistical analysis with the expected copy number is more powerful than the statistic with the probe intensity measurements regardless of the accuracy of the estimation of copy numbers.

Additional files

Additional file 1: A complete report for all experiments performed for this work. Text 1. Rao's score test statistic with the expected copy number. Text 2. Rao's score test statistic with the probe intensity measurements. **Figure S1.** Results of the clustering analysis with family samples (A) and cohort samples (B). **Figure S2.** Schematic representation of the strategy for the CNV analysis. **Figure S3.** Validation results. **Figure S4.** Validation results of replication samples by TaqMan qPCR experiment. **Table S1.** Specification of parameters for the signal model used in the simulation studies. **Table S2.** Primer information for CNV validation. **Table S3.** Accuracy of copy number clusters identified with PedCNV. **Table S4.** Accuracy of copy number clusters identified with CNVtools. **Table S5.** Accuracy of copy number estimated with silhouette score. **Table S6.** Empirical type 1 error estimates when $M = 6$. **Table S7.** Empirical power estimates when $M = 6$. **Table S8.** Empirical power estimates of T_1 and T_1^* which use the expected copy number and the most probable copy number respectively. **Table S9.** Estimated parameters of T_1 and T_1^* which use the expected copy number and the most probable copy number respectively. (PDF 814 kb)

Additional file 2: Simulated data with 2000 replicates for different values of β under WSC, when there are three copy number clusters. (ZIP 138685 kb)

Abbreviations

aCGH: array comparative genomic hybridization; BSC: Badly separated clusters; CNV: Copy number variation; GADA: Genome Alteration Detection Analysis; Hct: Hematocrit; KoGES: Korean Genome Epidemiology Study; LCL: Lymphoblastoid cell line; MSC: Moderately separated clusters; PC: Principle component; qPCR: quantitative PCR; QQ plot: Quantile quantile plot; REML: Restricted maximum likelihood; WSC: Well separated clusters

Acknowledgements

We do appreciate that Korean Genome Analysis Project (4845-301) provided the genome data, and that all three anonymous reviewers and editor for valuable suggestions and constructive comments, which greatly help us to improve our manuscript.

Funding

This study was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by National Research Foundation of Korea Grant by the Korean Government (NRF-2014S1A2A2028559) and the Ministry of Education (NRF-2016R1D1A1A09919610), and an intramural grant from the Korea National Institute of Health (2012-73004-00). The funding body was not involved in the design or conclusion of our study.

Availability of data and materials

Our R package, PedCNV, can be downloaded from https://cran.r-project.org/web/packages/PedCNV/index.html. Simulated data with 2000 replicates for different values of β under WSC when three copy number clusters are considered, can be downloaded as Additional file 2. In the real data analysis, we used the same DNA samples as were used in Lee et al. [23].

Authors' contributions

Conceived and designed the experiments: ML and SW. Performed the experiments: ML and LW. Analyzed and interpreted the data: ML, SM, LW and SW. Drafted the manuscript: ML, SM, LW and SW. Edited the manuscript: SK, YJK, MYH, YJK, RCE and BJK. All authors read and approved the final version of the manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

All subjects from this study provided written informed consent and the institutional ethics committees of participating institutions approved the experimental protocols (approved IRB number: 2011-08CON-10-P).

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details

[1]Department of Applied Statistics, Chung-Ang University, Seoul 156-756, South Korea. [2]Department of Bioinformatics and Computational Biology, Iowa State University, Ames, IA 50011, USA. [3]Division of Structural and Functional Genomics, Center for Genome Science, National Institute of Health, Cheongju-si, Chungcheongbuk-do 363-951, South Korea. [4]Interdisciplinary Program of Bioinformatics, Seoul National University, Seoul 151-742, South Korea. [5]Naver Labs, 235 Pangyoyeok-ro, Bundang-gu, Seongnam-si, Gyeonggi-do 13494, South Korea. [6]Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH 44106, USA. [7]Department of Public Health Science, Seoul National University, Seoul 151-742, South Korea. [8]Institute of Health and Environment, Seoul National University, Seoul 151-742, South Korea.

Received: 15 June 2016 Accepted: 31 March 2017
Published online: 18 April 2017

References

1. Sharp AJ, Cheng Z, Eichler EE. Structural variation of the human genome. Annu Rev Genomics Hum Genet. 2006;7:407–42.
2. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nat Rev Genet. 2006;7(2):85–97.
3. Lupski J. Genomic Disorders: The Genomic Basis of Disease. J Med Genet. 2008;45:S32.
4. McCarroll SA, Altshuler DM. Copy-number variation and association studies of human disease. Nat Genet. 2007;39:S37–42.
5. Wang K, Li MY, Hadley D, Liu R, Glessner J, Grant SFA, Hakonarson H, Bucan M. PennCNV: An integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. 2007;17(11):1665–74.
6. Colella S, Yau C, Taylor JM, Mirza G, Butler H, Clouston P, Bassett AS, Seller A, Holmes CC, Ragoussis J. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res. 2007;35(6):2013–25.
7. Li C. Automating dChip: toward reproducible sharing of microarray data analysis. BMC Bioinformatics. 2008;9:231.
8. Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K, et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet. 2008;40(10):1253–60.
9. Kim JH, Hu HJ, Yim SH, Bae JS, Kim SY, Chung YJ. CNVRuler: a copy number variation-based case-control association analysis tool. Bioinformatics. 2012; 28(13):1790–2.
10. Barnes C, Plagnol V, Fitzgerald T, Redon R, Marchini J, Clayton D, Hurles ME. A robust statistical method for case-control association testing with copy number variation. Nat Genet. 2008;40(10):1245–52.
11. Ionita-Laza I, Perry GH, Raby BA, Klanderman B, Lee C, Laird NM, Weiss ST, Lange C. On the analysis of copy-number variations in genome-wide association studies: a translation of the family-based association test. Genet Epidemiol. 2008;32(3):273–84.
12. Bansal V, Libiger O, Torkamani A, Schork NJ. Statistical analysis strategies for association studies involving rare variants. Nat Rev Genet. 2010;11(11):773–85.
13. Shi G, Rao DC. Optimum Designs for Next-Generation Sequencing to Discover Rare Variants for Common Complex Disease. Genet Epidemiol. 2011;35(6):572–9.
14. Laird NM, Horvath S, Xu X. Implementing a unified approach to family-based tests of association. Genet Epidemiol. 2000;19 Suppl 1:S36–42.
15. Murphy A, Won S, Rogers A, Chu JH, Raby BA, Lange C. On the genome-wide analysis of copy number variants in family-based designs: methods for combining family-based and population-based information for testing dichotomous or quantitative traits, or completely ascertained samples. Genet Epidemiol. 2010;34(6):582–90.
16. Won S, Wilk JB, Mathias RA, O'Donnell CJ, Silverman EK, Barnes K, O'Connor GT, Weiss ST, Lange C. On the analysis of genome-wide association studies in family-based designs: a universal, robust analysis approach and an application to four genome-wide association studies. PLoS Genet. 2009; 5(11):e1000741.
17. Thornton T, McPeek MS. ROADTRIPS: Case-control Association Testing with Partially or Completely Unknown Population and Pedigree Structure. Am J Hum Genet. 2010;86(2):172–84.
18. Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 2006;38(8):904–9.
19. Hartigan JA. Clustering algorithms. New York: Wiley; 1975.
20. McPeek MS, Wu XD, Ober C. Best linear unbiased allele-frequency estimation in complex pedigrees. Biometrics. 2004;60(2):359–67.
21. Rousseeuw PJ. Silhouettes–a Graphical Aid to the Interpretation and Validation of Cluster-Analysis. J Comput Appl Math. 1987;20:53–65.
22. Rao CR, Mitra SK. Generalized inverse of matrices and its applications. New York: Wiley; 1971.
23. Lee EJ, Kim KJ, Kim HN, Bok J, Jung SC, Kim EK, Lee JY, Kim HL. Genome-wide scan of granular corneal dystrophy, type II: confirmation of chromosome 5q31 and identification of new co-segregated loci on chromosome 3q26.3. Exp Mol Med. 2011;43(7):393–400.
24. Craddock N, Hurles ME, Cardin N, Pearson RD, Plagnol V, Robson S, Vukcevic D, Barnes C, Conrad DF, Giannoulatou E, et al. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature. 2010;464(7289):713–20.
25. Moon S, Jung KS, Kim YJ, Hwang MY, Han K, Lee JY, Park K, Kim BJ. KGVDB: a population-based genomic map of CNVs tagged by SNPs in Koreans. Bioinformatics. 2013;29(11):1481–3.
26. Marioni JC, Thorne NP, Valsesia A, Fitzgerald T, Redon R, Fiegler H, Andrews TD, Stranger BE, Lynch AG, Dermitzakis ET, et al. Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. Genome Biol. 2007;8(10):R228.
27. Fiegler H, Redon R, Andrews D, Scott C, Andrews R, Carder C, Clark R, Dovey O, Ellis P, Feuk L, et al. Accurate and reliable high-throughput detection of copy number variation in the human genome. Genome Res. 2006;16(12): 1566–74.
28. Devlin B, Roeder K. Genomic control for association studies. Biometrics. 1999;55(4):997–1004.
29. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E. Efficient control of population structure in model organism association mapping. Genetics. 2008;178(3):1709–23.
30. Elston RC, Stewart J. A general model for the genetic analysis of pedigree data. Hum Hered. 1971;21(6):523–42.
31. Bryant BG. Large-sample results for optimization-based clustering methods. J Classif. 1991;8(1):31–44.
32. Bryant PG, Williamson JA. Asymptotic Behaviour of Classification Maximum Likelihood Estimates. Biometrika. 1978;65:273–81.