G&I Genomics & Informatics

pISSN 1598-866X eISSN 2234-0742
Genomics & Informatics 2012;10(2):81-87
http://dx.doi.org/10.5808/GI.2012.10.2.81

REVIEW

Identifying Copy Number Variants under Selection in Geographically Structured Populations Based on *F*-statistics

Hae-Hiang Song[1]*, Hae-Jin Hu[2], In-Hae Seok[3], Yeun-Jun Chung[2]**

[1]Division of Biostatistics, Department of Medical Lifescience, The Catholic University of Korea, College of Medicine, Seoul 137-040, Korea, [2]Department of Microbiology, Integrated Research Center for Genome Polymorphism, The Catholic University of Korea, College of Medicine, Seoul 137-040, Korea, [3]Department of Statistics, Hankuk University of Foreign Studies, Yongin 449-791, Korea

Large-scale copy number variants (CNVs) in the human provide the raw material for delineating population differences, as natural selection may have affected at least some of the CNVs thus far discovered. Although the examination of relatively large numbers of specific ethnic groups has recently started in regard to inter-ethnic group differences in CNVs, identifying and understanding particular instances of natural selection have not been performed. The traditional F_{ST} measure, obtained from differences in allele frequencies between populations, has been used to identify CNVs loci subject to geographically varying selection. Here, we review advances and the application of multinomial-Dirichlet likelihood methods of inference for identifying genome regions that have been subject to natural selection with the F_{ST} estimates. The contents of presentation are not new; however, this review clarifies how the application of the methods to CNV data, which remains largely unexplored, is possible. A hierarchical Bayesian method, which is implemented via Markov Chain Monte Carlo, estimates locus-specific F_{ST} and can identify outlying CNVs loci with large values of F_{ST}. By applying this Bayesian method to the publicly available CNV data, we identified the CNV loci that show signals of natural selection, which may elucidate the genetic basis of human disease and diversity.

Keywords: Bayes theorem, DNA copy number variations, population structure, selection, Wright's F_{ST}

Introduction

It has become known that human genomes differ more as a consequence of structural variation than of single-base-pair differences [1-4]. Structural genomic variants, mainly in the form of copy number variants (CNVs), have recently been rediscovered as contributors to evolution and as pathoetiologic elements for complex human diseases [5, 6]. However, so far, there are not many CNV-based genome-wide association studies or population genetics studies compared with the single-nucleotide polymorphism (SNP)-based counterparts. One notable CNV-based population genetics study was done by Jakobsson *et al*. [7]. They applied SNP, haplotype, and CNV information to reveal the structures of 29 world-wide populations. Even though their CNVs were not as effective as SNPs or haplotypes in discriminating those populations, the authors showed the potential to apply this information to human genetic studies. Other examples of CNV-based disease susceptibility studies are systemic autoimmunity diseases [8, 9], psoriasis [10], and human immunodeficiency virus (HIV) infection and progression to acquired immune deficiency syndrome (AIDS) [11]. Disease could affect the patterns of CNV diversity via natural selection, and higher copy numbers of the immunoregulatory and inflammatory cytokine gene *CCL3L1*, for example, are associated with lower risks of HIV infection and the progression to AIDS [11]; furthermore, in a genome-wide study, the level of population differentiation at this locus was found to be extraordinary compared to that of other CNVs, suggesting that natural selection may have influenced *CCL3L1* copy number in humans [3].

The traditional F_{ST} measure is useful for identifying

Received April 10, 2012; Revised May 11, 2012; Accepted May 17, 2012

***Corresponding author 1:** Tel: +82-2-2258-7225, Fax: +82-2-596-8969, E-mail: hhsong@catholic.ac.kr

****Corresponding author 2:** Tel: +82-2-2258-7343, Fax: +82-2-596-4513, E-mail: yejun@catholic.ac.kr

regions of the genome affected by natural selection. In a recently published review paper, Holsinger and Weir define Wright's *F*-statistics (F_{ST} in particular) [12] and describe methods-of-moment estimates and how F_{ST} estimates should be interpreted [13]. Although the authors also mention the maximum likelihood Bayesian estimates of F_{ST}, their description is very limited in identifying genomic regions under selection and do not present an application to the datasets. They simply compared locus-specific estimates of F_{ST} with its genome-wide distribution, and therefore, probabilities are not attached to those with a higher F_{ST}.

Here, we outline the steps that constitute testing for outlier loci in population datasets by the maximum likelihood Bayesian method, using BayeScan computer software [14] (Note that the terms of locus and CNV are used interchangeably in this paper). The logic for determining outlier loci is simple. If natural selection favors one allele over others at a particular locus in some populations, the F_{ST} at that locus will be larger than at loci in which among-population differences are purely a result of genetic drift. Therefore, genome scans that compare single-locus estimates of F_{ST} with the genome-wide background might identify regions of the genome that have been subjected to diversifying selection [15].

We will demonstrate F_{ST} estimation by maximum likelihood Bayesian method with the publicly available high-resolution CNV dataset generated by Conrad *et al*. [4]. We will also show that this method identifies several genomic regions showing signals of natural selection. Even though we focus mainly on the CNV data here, the detailed steps of analysis are analogous for other types of molecular marker data, such as microsatellites, SNPs, and amplified fragment-length polymorphisms.

Method-of-moment (or ANOVA) estimates of F_{ST}

F-statistics

Wright [12] introduced *F*-statistics (F_{ST}, F_{IT}, and F_{IS}) as a tool for describing the partitioning of genetic diversity within and among populations that are directly related to the rates of evolutionary processes, such as migration, mutation, and drift. Specifically, *F*-statistics can be defined in many different ways: in terms of variances of allele frequencies, correlations between random gametes, and probabilities that two gametes chosen have different alleles. Depending on the relativity to the subpopulation or to the total population, F_{ST}, F_{IT}, and F_{IS} are defined, where subscript IS refers to 'individuals within subpopulations,' ST to 'subpopulations within the total population,' and IT to 'individuals within the total population.'

Following the work of Cockerham [16], *F*-statistics are defined in terms of the variance components - that is, the total variation in the genetic data is broken down into three components: (a) between subpopulations within the total population (we sometimes say 'between populations'); (b) between individuals within subpopulations; and (c) between gametes within individuals. F_{ST}, F_{IT}, and F_{IS} are defined as the expectations under the model of a/(a + b + c), (a + b)/(a + b + c), and b/(b + c) and estimated by the corresponding sample values [17, 18]. Here, it is perhaps pertinent to mention that when Weir and Cockerham [18] presented these definitions, they assumed a model consisting of an ancestral population from which subpopulations have descended in isolation under the same evolutionary processes. Thus, it is meaningful to have a single measure of population structure; that is, a global F_{ST}, which is an average over subpopulations. However, in identifying candidate loci under natural selection, evidence for locus-specific selection is of interest, and thus the estimators of locus-specific F_{ST} will be described in the next section.

Often, readers may be confused with several terms appearing in population genetics. Wright [12] interprets F_{ST} as a measure of the progress of the subpopulation towards the fixation of one allele of each locus in the absence of mutation and hence called a 'fixation index.' F_{ST} is also interpreted as a measure of shared ancestry with the subpopulations, relative to that in the population, and is thus called the 'coancestry coefficient' [19]. Therefore, if the value of F_{ST} is small, it means that the allele frequencies within each subpopulation are similar; if it is large, it means that the allele frequencies of subpopulations are different. On the other hand, F_{IS} or F_{IT} is defined as the correlation between two gametes that form a zygote relative to the subpopulation or population, and thus, F_{IS} (or F_{IT}) is called the 'inbreeding coefficient' [19].

Estimating F_{ST} by ANOVA methods

The estimators of *F*-statistics proposed by Weir [17] and Weir and Cockerham [18] are based on an analysis of variance (ANOVA) of allele frequencies, equivalently called the method-of-moments estimates. The weighted ANOVA estimates of F_{ST}, $\mathrm{F_{IT}}$, and $\mathrm{F_{IS}}$ may be expressed in terms of the mean sum of squares for gametes (*MSG*), individuals (*MSI*), and populations (we sometimes say 'between subpopulations') (*MSP*), where the mean squares are estimated by an ANOVA model. In estimating F_{ST} specifically for our analysis of CNV data, we need to consider unbalanced samples (i.e., populations of unequal size). However, as the formulas are messy, we present here those for balanced samples. Formulas for unbalanced samples can be found in Rousset (in Appendix A) [20].

The definition of F-statistics used here is

$$F_{\mathrm{IS}} = \frac{Q_1 - Q_2}{1 - Q_2},\ F_{\mathrm{ST}} = \frac{Q_2 - Q_3}{1 - Q_3},\ \text{and } F_{\mathrm{IT}} = \frac{Q_1 - Q_3}{1 - Q_3},$$

where Q values are probabilities of identity in state: Q_1 among the genes (gametes) within individuals, Q_2 among genes in different individuals within populations, and Q_3 among the populations. The estimates $\hat{Q}$ are expressed in terms of observed frequencies of identical pairs of genes in the sample, with the following relationships:

$$1 - \hat{Q}_1 = MSG,$$

$$\hat{Q}_1 - \hat{Q}_2 = (MSI - MSG)/2,$$

and

$$\hat{Q}_2 - \hat{Q}_3 = (MSP - MSI)/(2n),$$

where n is the sample size of each population. Then, the single locus estimator $\hat{F}_{\mathrm{ST}}$ is given by

$$\hat{F}_{\mathrm{ST}} = \frac{MSP - MSI}{MSP + (n_c - 1) + n_c MSG}, \tag{1}$$

which is found in Weir (1997: 178) [17]. n_c will be defined below. If one needs to obtain the multilocus estimator of $\hat{F}_{\mathrm{ST}}$, it is usual to compute the estimator as a sum of locus-specific numerators over a sum of locus-specific denominators (see Weir [17] and Weir and Cockerham [18]). This is the case that map information for SNPs is obtained for each gene, and a weighted-average F_{ST} from all SNPs is estimated for each gene [18]. For a set of I loci, the multilocus ANOVA estimators are

$$\hat{F}_{\mathrm{IS}} = \frac{\sum_{i=1}^{I} [n_c(MSI - MSG)]_i}{\sum_{i=1}^{I} [n_c(MSI + MSG)]_i},$$

$$\hat{F}_{\mathrm{ST}} = \frac{\sum_{i=1}^{I} [MSP - MSI]_i}{\sum_{i=1}^{I} [MSP + (n_c - 1)MSI + n_c MSG)]_i},$$

$$\hat{F}_{\mathrm{IT}} = \frac{\sum_{i=1}^{I} [MSP + (n_c - 1)MSI - n_c MSG)]_i}{\sum_{i=1}^{I} [MSP + (n_c - 1)MSI + n_c MSG)]_i}, \tag{2}$$

for $n_c = (S_1 - S_2/S_1)/(n - 1)$, where S_1 is the total sample size and S_2 is the sum of squared sample sizes of populations [21]. For convenience, we denote the estimator $\hat{F}_{\mathrm{ST}}$ by F_{ST}.

Rousset [21] explained that the multilocus estimators of Weir [17] and Weir and Cockerham [18] differ slightly, and these two also differ slightly from that proposed by Rousset [21], which assigns more weight to larger samples. In this paper, the GENEPOP software (version 3.4) (http://wbiomed.curtin.edu.au/genepop/) of Rousset was used for the calculation of F_{ST}. In order to distinguish from those of the method-of-moments estimates of Weir [17] and Weir and Cockerham [18], we will call the estimates of GENEPOP ANOVA estimates.

The estimated values of F_{ST} can be negative when levels of differentiation are close to zero and/or sample sizes are small, indicating no population differentiation at these loci [18]. One can assign a value of zero to negative F_{ST} estimates.

Identifying CNVs under Selection Using a Bayesian Method

Identifying selection

Identifying candidate CNVs under natural selection using locus-specific estimates of F_{ST} is of great interest, in view of our increased knowledge of the relationships among human populations from genome-wide patterns of variation. The logic for identifying selection is straightforward. The pattern of genetic differentiation at a neutral locus is completely determined by the demographic history, migration rates among the populations, and the mutation rates at the loci. It is reasonable to assume that all autosomal loci have experienced the same demographic history and migration rates among the populations, and the observed population structure can be largely explained by random drift at neutral loci. However, as the individuals from different populations often vary genetically at a few key sites in their genome, loci showing unusually large amounts of differentiation may indicate regions of the genome that have been subject to positive selection, whereas loci showing unusually small amounts of differentiation may indicate regions of the genome that have been subject to stabilizing (balancing) selection [15, 22-24]. Thus, the outlier method makes it possible to detect divergences in some loci of the genome due to selection.

Bayesian method

Balding and Nichols have proposed the use of population- and locus-specific estimates in the context of a migration-drift equilibrium model [24]. They modeled allele frequencies of biallelic markers using a beta distribution with expectation p and variance $p(1-\mathrm{p})/(1+\theta)$ so that $F_{\mathrm{ST}} = 1/(1+\theta)$. Then, they used a likelihood-based approach to estimate population- and locus-specific F_{ST}.

This formulation was extended later to consider multi-allelic loci [19]. Falush *et al.* [25] considered the non-equilibrium fission model that subpopulations evolve in isolation after splitting from an ancestral population. We will not distinguish the migration-drift equilibrium model and nonequilibrium fission model, since both lead to the same multinomial-Dirichlet distribution (or beta-binomial for biallelic loci) for the subpopulation allele frequencies [26]. The main difference between the two models resides in the interpretation given to F_{ST}: in the case of the migration-drift equilibrium model, F_{ST} measures how divergent each subpopulation is from the total population, while in the case of the fission model, it measures the degree of genetic differentiation between each descendant population and the ancestral population.

We consider a collection of *J* subpopulations and a set of *I* loci. Let K_i be the number of alleles at the *i* th locus. The extent of differentiation between subpopulation *j* and the ancestral population at locus *i* is measured by F_{ST}^{ij} . Let $\mathbf{p}_i = \{p_{ik}\}$ denote the allele frequencies of the ancestral population at locus *i*, where p_{ik} is the frequency of the allele *k* at locus $i(\sum_k P_{ik} = 1)$. We use $\mathbf{p} = \{\mathbf{p}_i\}$ to denote the entire set of allele frequencies of the ancestral population and $\widetilde{P}_{ij} = \{\widetilde{P}_{ijk}\}$ to denote the current allele frequencies at locus *i* for subpopulation *j*. Under the model and the definitions above, the allele frequencies at locus *i* in subpopulation *j* follow a Dirichlet distribution with parameters $\theta_j\mathbf{p}_i$, a Bayesian prior distribution,

$$\widetilde{P}_{ij} \sim \mathrm{Dir}(\theta_{ij}p_{i1}, \cdots, \theta_{ij}p_{iK_i}), \qquad (3)$$

where $\theta_{ij} = 1/F_{\mathrm{ST}}^{ij} - 1$. The extent of differentiation at locus *i* between subpopulation *j* and the ancestral population is measured by F_{ST}^{ij} and is the result of its demographic history. The full prior distribution across loci and populations is given by

$$\pi(\widetilde{\mathrm{p}} \mid \mathrm{p}, \theta) = \prod_{i=1}^{I}\prod_{j=1}^{J}\pi(\widetilde{\mathrm{p}}_{ij} \mid \mathrm{p}_i, \theta_{ij}). \qquad (4)$$

We need a hierarchical model for locus- and population-specific effects to identify candidate loci under natural selection. It is no doubt that population- and locus-specific estimates of F_{ST}^{ij} by the moments (or ANOVA) methods are likely to be inaccurate, especially for loci with a small number of different alleles. As an alternative, Beaumont and Balding [22] proposed a familiar logistic regression model to decompose population- and locus-specific F_{ST}^{ij} coefficients (bounded between 0 and 1) into a population-specific component, β_j, shared by all loci and a locus-specific component, α_i, shared by all populations [22]:

$$\log\left[\frac{F_{\mathrm{ST}}^{ij}}{1-F_{\mathrm{ST}}^{ij}}\right] = \log\left[\frac{1}{\theta_{ij}}\right] = \alpha_i + \beta_j \qquad (5)$$

The locus-specific effects express mutation and some forms of selection, and population-specific effects express migration rates, population sizes, and population-specific mating patterns [22]. The advantage of this formulation is that instead of estimating locus- and population-specific $I \cdot J$ F_{ST}^{ij} coefficients (as in the method of moments or ANOVA), we only have to estimate *I* the parameters α_i and the *J* parameters β_j. With the estimates of α_i and β_j, F_{ST}^{ij} (and equivalently $\theta_{ij} = \exp(-(\alpha_i + \beta_j))$ can be estimated.

Above all, the estimate of α_i is our objective to detect outlier loci as selection candidates. Departure from neutrality at a given locus is assumed when α_i is significantly different from 0 at that locus. A positive value of α_i suggests diversifying selection, whereas negative values suggest balancing selection. The posterior probability that a locus is subject to selection, $P(\alpha_i \neq 0)$, is estimated directly from the Markov Chain Monte Carlo method (MCMC) (see below).

Likelihood for allele counts

The data of codominant markers consist of allele counts obtained from samples of size n_{ij}. We use α_{ijk} to denote the number of alleles *k* observed at locus *i* in the sample from subpopulation *j*. Thus, $n_{ij} = \sum_k \alpha_{ijk}$. The full dataset can be presented as a matrix $\mathbf{N} = \{\mathbf{a}_{\mathbf{ij}}\}$, where $\mathbf{a}_{\mathbf{ij}} = \{\alpha_{ij1}, \alpha_{ij2}, \cdots, \alpha_{ijK_i}\}$ is the allele count at locus *i* for subpopulation *j*. The observed allele frequencies, $\mathbf{a}_{\mathbf{ij}}$, can be considered as sampled from the true alleles frequencies $\widetilde{\mathrm{P}}_{\mathrm{ij}}$ and therefore can be described by the multinomial distribution:

$$\alpha_{ij} \sim Multinomia\ l\ \{n_{ij}; \widetilde{P}_{ij1}, \widetilde{P}_{ij2}, \cdots \widetilde{P}_{ijK_i}\} \qquad (6)$$

We adopt the multinomial-Dirichlet distribution as a marginal distribution of $\mathbf{a}_{\mathbf{ij}}$, the allele frequency counts at a locus within a subpopulation, and the reasons of its adoption are explained in Foll and Gaggiotti [26]:

$$P(\mathbf{a}_{\mathbf{ij}} \mid \mathbf{p}_{\mathrm{i}}, \alpha_i, \beta_j) = \frac{n_{ij}!\Gamma(\theta_{ij})}{\Gamma(n_{ij}+\theta_{ij})}\prod_{k=1}^{k_i}\frac{\Gamma(a_{ijk}+\theta_{ij}p_{ik})}{a_{ijk}!\Gamma(\theta_{ij}p_{ik})}.$$

The joint likelihood for allele counts is obtained by multiplying across all loci and populations:

$$L(\mathbf{p}, \alpha, \beta) = \prod_{i=1}^{I}\prod_{j=1}^{J}P(\mathbf{a}_{\mathbf{ij}} \mid \mathbf{p}_{\mathrm{i}}, \alpha_i, \beta_j). \qquad (7)$$

Since the allele frequencies in the ancestral population $\mathbf{p}_i = \{p_{ik}\}$ are unknown, they are estimated by introducing a noninformative Dirichlet prior, $\mathbf{p}_{\mathrm{i}} \sim \mathrm{Dir}\ (1, \cdots, 1)$, into the Bayesian model [27].

The full Bayesian model is given by

$$\pi(\mathbf{p}, \alpha, \beta \mid \mathrm{N}) \propto L(\mathbf{p}, \alpha, \beta)\pi(\mathbf{p})\pi(\alpha)\pi(\beta). \qquad (8)$$

Following Beaumont and Balding [22], the Gaussian prior distributions are used for the population effects β_j and for the locus effects α_i. Their means and variances are chosen to improve convergence and for each F_{ST}^{ij} to have non-negligible density over almost the whole interval from 0 to 1.

Identifying CNVs subject to selection

As described above, the effect of selection is parameterized by α_i in the logistic regression. Thus, two models in the logistic, one that includes both effects of α_i (i.e., $\alpha_i \neq 0$) and β_j (selection model, M2) and another one that does not include the effect of selection (i.e., $\alpha_i = 0$) and only includes the effect of β_j (neutral model, M1), are considered.

We can decide in a Bayesian framework whether or not there is selection at all. This is done by estimating the posterior probabilities for two models for each locus in the logistic regression: neutral (M1) and selection (model 2, M2), on the basis of a dataset $\mathbf{N} = \{ \mathbf{a}_{ij} \}$. We use a reversible jump MCMC algorithm [28] to estimate the posterior probability of each one of these models, and this is done separately for each locus *i*. We can then have posterior probability that a locus is subject to selection - that is, $P(\alpha_i \neq 0)$. This probability is estimated directly from the output of the MCMC by simply counting the number of times α_i is included in the model (see Foll and Gaggiotti [27] for a more detailed explanation).

In BayeScan, selection decision (model choice decision), with a specific goal of detecting outlier loci in mind, is actually performed by using posterior odds (PO). For this, we first need to describe 'Bayes factors.' For the two models of neutral (M1) and selection (M2), the Bayes factor (BF) for model M2 is given by BF = $P(\mathbf{N} \mid \mathrm{M2})/P(\mathbf{N} \mid \mathrm{M1})$, where $\mathbf{N} = \{ \mathbf{a}_{ij} \}$ is a dataset and *P* indicates probability. This BF provides a degree of evidence in favor of one model versus another. In the context of multiple testing - that is, testing a large number of loci simultaneously - we also need to incorporate our skepticism about the chance that each locus is under selection. This is done by setting the prior odds for the neutral model $P(\mathrm{M1})/P(\mathrm{M2})$. We make selection decisions by using PO, which is defined as PO = $P(\mathrm{M2}|\mathbf{N})/P(\mathrm{M1}|\mathbf{N})$ = BF × $P(\mathrm{M2})/\mathrm{P}(\mathrm{M1})$. PO are simply the ratio of posterior probabilities and indicate how much more likely the model with selection (M2) is compared to the neutral model (M1) (GENEPOP software). PO of 0.5-1.0, 1.0-1.5, 1.5-2.0, and 2-∞ are, respectively, considered as substantial, strong, very strong, and decisive evidence for selection [14].

A big advantage of posterior probabilities is that they directly allow the control of the false discovery rate (FDR), where FDR is defined as the expected proportion of false positives among outlier loci. Controlling for FDR has a much greater power than controlling for family-wise error rates using Bonferroni correction, for example [29]. In BayeScan, by first setting a target FDR, the PO threshold achieving this FDR is determined, and outliers equal to or above this PO threshold are listed in the output of R that is provided along with BayeScan.

Empirical Distribution of F_{ST} Based on CNV

Study populations

We obtained the high-resolution CNV data by Conrad *et al.* [4] from the Database of Genomic Variants (http://projects.tcag.ca/variation/). The dataset consists of 90 Han Chinese (CHB) and Japanese (JPT), 180 Caucasians of European descent (CEU), and 180 Yoruba (YRI) samples. Among the 450 individuals, we removed 19 individuals with more than 400 missing CNV genotypes. Each CNV was analyzed as biallelic dominant data, classified as neutral versus non-neutral (losses or gains).

Empirical distribution of F_{ST}

To assess population differentiation with an allele frequency spectrum of CNVs, the locus-specific estimates by ANOVA method and by hierarchical Bayesian method were obtained from all autosomes. The estimated values by the ANOVA method (moments estimates) can be negative when levels of differentiation are close to 0 and/or sample sizes are small, indicating no population differentiation at these loci [18]. In the hierarchical Bayesian method, the locus-specific estimate values are all positive, since they are generated by posterior distributions, which is an important benefit of a Bayesian method.

With the CNV data of this study, the summary statistics of F_{ST} was 0.0853 ± 0.1195 (median, 0.0360; range, −0.0071 to 0.8994) for the ANOVA method and 0.2459 ± 0.0115 (median, 0.2462; range, 0.2193 to 0.6166) for the Bayesian method. The empirical distributions of F_{ST} are found in Fig. 1. As can be seen from the Figure, low F_{ST} values are prevalent, and a large variability persisted in the ANOVA method, while F_{ST} distribution was sharply focused in the Bayesian method. The comparison of empirical distributions of the ANOVA and Bayesian methods demonstrates that Bayesian estimates perform better than ANOVA in identifying CNVs affected by natural selection.

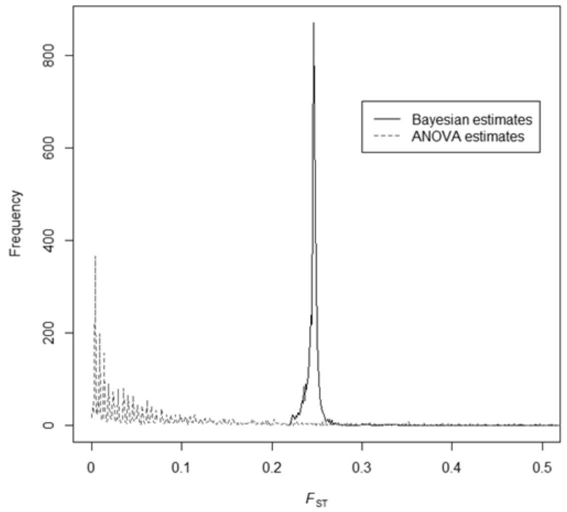

Fig. 1. Empirical distributions of copy number variants (CNV) data with the three population groups. CNV data are from Conrad *et al.* [4].

Table 1. Identified outlier CNVs potentially subject to positive selection by BayeScan

CNP_id	Chr	Start	Stop	Proportion of non-neutral			F_{ST}
				CEU	CHB + JPT	YRI	
CNVR2664.1	5	159282379	159283692	0.90	1.00	0.00	0.623
CNVR371.1	1	153926378	153931106	0.00	0.00	0.85	0.544

CNV, copy number variants; CEU, Caucasians of European descent; CHB, Han Chinese; JPT, Japanese; YRI, Yoruba.

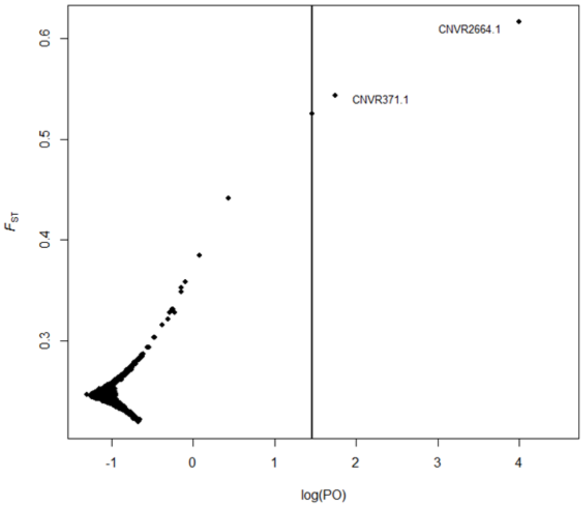

Fig. 2. Spacial scans for identification of F_{ST} outlier copy number variants (CNVs) potentially subject to positive selection by BayeScan. CNV data are from Conrad *et al*. [4]. PO, posterior odds.

Bayesian outlier detections

We searched the genome regions that show signals of natural selection by the Bayesian likelihood method implemented via reversible jump MCMC in BayeScan software [27]. For each CNV, BayeScan directly estimates the probability that a CNV is under positive selection, from which the PO are computed [27]. The posterior odds indicate how much more likely the model with selection is compared to the neutral model, and posterior odds of 5.0 was chosen as a threshold in the detection of CNVs under positive selection. This directly allows us to control the FDR, the expected proportion of false positives among outlier CNVs [29]. In our analysis, we chose FDR to be 0.05. In the BayeScan application, first following 10 pilot runs of 5,000 iterations and an additional burn-in of 50,000 iterations, we used 50,000 iterations (sample size of 5,000 and a thinning interval of 10).

Outliers detected by BayeScan

BayeScan identified the three outlier CNVs with Conrad *et al*.'s [4] CNV data (Table 1, Fig. 2). The most decisive outlier is CNVR2664.1 (chromosome 5), in which most CEU (90%) and all CHB + JPT populations are outlying gain or loss status. The next decisive outlier is CNVR371.1 (chr. 1), in which only the YRI population has non-neutral CNVs.

Conclusion

In general, the observed population structure observed by hierarchical clustering, multidimensional scaling (MDS) plots, and tree plots can be largely explained by random drift at neutral CNVs. However, as individuals from different populations often vary genetically at a few key sites in their genome, only the outlier method makes it possible to detect accelerated divergence in some CNV regions of the genome due to local selection. This local selection is detected relative to the majority of CNVs, in which among-population differences are purely a result of genetic drift.

Successful outlier detection depends on reliable and obtainable estimates of F_{ST} and also on sampling variances of F_{ST}. Very large variances that are associated with single locus moment estimates of F_{ST} preclude the use of these estimates to detect selection in spite of the fact that sampling variances will decrease with the number of alleles at a locus and with the numbers of populations sampled. In this respect, the availability of locus- and population-specific Bayesian estimates of F_{ST} provides a set of tools for identifying genomic regions or populations with unusual evolutionary histories. The most important benefits of Bayesian estimates and its selection method are that the Bayesian methods allow probability statements to be made about F_{ST} and can be extended to explore the relationship with demographic or environmental covariates in the model [26]. Furthermore, likelihood-based Bayesian methods have the flexibility to accommodate missing data. However, implementations of Bayesian methods may be computationally demanding.

Selection provides information about the adaptation to a wide range of habitats and climates [30], and thus, interpreting the story of human adaptation is an interesting research area for the studies of evolution and disease processes in the future. Thus, more studies of outlier detection need to be replicated with large population-based data.

Acknowledgments

This study was supported by a grant from the Korea Healthcare Technology R&D Project (A092258), Ministry of Health and Welfare, Republic of Korea.

References

1. Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, *et al*. Large-scale copy number polymorphism in the human genome. *Science* 2004;305:525-528.
2. Tuzun E, Sharp AJ, Bailey JA, Kaul R, Morrison VA, Pertz LM, *et al*. Fine-scale structural variation of the human genome. *Nat Genet* 2005;37:727-732.
3. Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, *et al*. Global variation in copy number in the human genome. *Nature* 2006;444:444-454.
4. Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, *et al*. Origins and functional impact of copy number variation in the human genome. *Nature* 2010;464:704-712.
5. Lupski JR. Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits. *Trends Genet* 1998;14:417-422.
6. Stankiewicz P, Lupski JR. Genome architecture, rearrangements and genomic disorders. *Trends Genet* 2002;18:74-82.
7. Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, *et al*. Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 2008;451: 998-1003.
8. Fanciulli M, Norsworthy PJ, Petretto E, Dong R, Harper L, Kamesh L, *et al*. *FCGR3B* copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nat Genet* 2007;39:721-723.
9. Yang Y, Chung EK, Wu YL, Savelli SL, Nagaraja HN, Zhou B, *et al*. Gene copy-number variation and associated polymorphisms of complement component C4 in human systemic lupus erythematosus (SLE): low copy number is a risk factor for and high copy number is a protective factor against SLE susceptibility in European Americans. *Am J Hum Genet* 2007;80:1037-1054.
10. Hollox EJ, Huffmeier U, Zeeuwen PL, Palla R, Lascorz J, Rodijk-Olthuis D, *et al*. Psoriasis is associated with increased beta-defensin genomic copy number. *Nat Genet* 2008;40: 23-25.
11. Gonzalez E, Kulkarni H, Bolivar H, Mangano A, Sanchez R, Catano G, *et al*. The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science* 2005;307:1434-1440.
12. Wright S. The genetical structure of populations. *Ann Hum Genet* 1949;15:323-354.
13. Holsinger KE, Weir BS. Genetics in geographically structured populations: defining, estimating and interpreting F(ST). *Nat Rev Genet* 2009;10:639-650.
14. Foll M. *BayeScan v2.0 User Manual*. BayeScan, 2010.
15. Akey JM, Zhang G, Zhang K, Jin L, Shriver MD. Interrogating a high-density SNP map for signatures of natural selection. *Genome Res* 2002;12:1805-1814.
16. Cockerham CC. Analyses of gene frequencies. *Genetics* 1973;74:679-700.
17. Weir BS. *Genetic Data Analysis II: Methods for Discrete Population Genetic Data*. Sunderland: Sinauer Associates, 1996.
18. Weir BS, Cockerham CC. Estimating F-statistics for the analysis of population structure. *Evolution* 1984;38:1358-1370.
19. Balding DJ. Likelihood-based inference for genetic correlation coefficients. *Theor Popul Biol* 2003;63:221-230.
20. Rousset F. Inferences from spatial population genetics. In: *Handbook of Statistical Genetics* (Balding DJ, Bishop MJ, Cannings C, eds.). Chichester: Wiley, 2001. pp. 239-269.
21. Rousset F. genepop'007: a complete re-implementation of the genepop software for Windows and Linux. *Mol Ecol Resour* 2008;8:103-106.
22. Beaumont MA, Balding DJ. Identifying adaptive genetic divergence among populations from genome scans. *Mol Ecol* 2004;13:969-980.
23. Vitalis R, Dawson K, Boursot P. Interpretation of variation across marker loci as evidence of selection. *Genetics* 2001;158: 1811-1823.
24. Beaumont MA, Nichols RA. Evaluating loci for use in the genetic analysis of population structure. *Proc R Soc Lond Series B Biol Sci* 1996;263;1619-1626.
25. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* 2003;164:1567-1587.
26. Foll M, Gaggiotti O. Identifying the environmental factors that determine the genetic structure of populations. *Genetics* 2006;174:875-891.
27. Foll M, Gaggiotti O. A genome-scan method to identify selected loci appropriate for both dominant and codominant markers: a Bayesian perspective. *Genetics* 2008;180:977-993.
28. Green PJ. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* 1995;82: 711-732.
29. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Methodol* 1995;57:289-300.
30. Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. *PLoS Biol* 2006;4:e72.