PlatinumCNV: a Bayesian Gaussian Mixture Model for Genotyping Copy Number Polymorphisms Using SNP Array Signal Intensity Data

Natsuhiko Kumasaka[1,*], Hironori Fujisawa[2], Naoya Hosono[3], Yukinori Okada[1], Atsushi Takahashi[1], Yusuke Nakamura[4], Michiaki Kubo[3], Naoyuki Kamatani[1]

1 Research Group for Medical Informatics, Center for Genomic Medicine, RIKEN, Tokyo, Tokyo, JAPAN

2 Department of Mathematical Analysis and Statistical Inference, The Institute of Statistical Mathematics, Tachikawa, Tokyo, JAPAN

3 Research Group for Genotyping, Center for Genomic Medicine, RIKEN, Tsurumi, Kanagawa, JAPAN

4 Human Genome Center, Institute of Medical Science, The University of Tokyo, Tokyo, JAPAN

[*]**Correspondence to: Natsuhiko Kumasaka**

Address: RIKEN, 4-6-1, Shirokane-dai, Minato-ku, Tokyo 108-8639, JAPAN

E-mail: kumasaka@src.riken.jp

Phone: +81-3-5449-5708

Fax: +81-3-5449-5564

Abstract

We present a statistical model for allele-specific patterns of copy number polymorphisms (CNPs) in commercial single nucleotide polymorphism (SNP) array data. This model is based on the observation that fluorescent signal intensities tend to cluster into clouds of similar allele-specific copy number (ASCN) genotypes at each SNP locus. To capture the tendency of this clustering to be made vague by instrumental errors, our model allows for cluster memberships to overlap each other, according to a Bayesian Gaussian mixture model (GMM). This approach is flexible, allowing for both absolute scale differences and X/Y scale imbalances of fluorescent signal intensities. The resulting model is also robust toward unobserved ASCN genotypes, which can be problematic for ordinary GMMs. We illustrated the utility of the model by applying it to commercial SNP array intensity data obtained from the Illumina HumanHap 610K platform. We retrieved more than 4,000 allele-specific CNPs, though 99% of them showed rather simple allele-specific CNP patterns with only a single aneuploid haplotype among the normal haplotypes. The genotyping accuracy was assessed by two approaches, quantitative PCR and replicated subjects. The results of both of these approaches demonstrated mean genotyping error rates of 1%. We demonstrated a preliminary genome-wide association study of three hematological traits. The result exhibited that it could form the foundation for new, more effective statistical methods for the mapping of both disease genes and quantitative trait loci with genome-wide CNPs. The methods described in this work are implemented in a software package, PlatinumCNV, available on the Internet.

Key words: Allele-specific copy number; genome-wide association study; oligonucleotide assay; empirical Bayes estimation; quantitative PCR.

Introduction

The success of genome-wide association studies (GWAS) with a number of single-nucleotide polymorphisms (SNPs) prompted us to study other genomic variants such as copy number variations (CNVs), insertions, inversions and translocations [1]. In particular, CNVs have been widely investigated in diverse populations [2, 3, 4, 5, 6], and we expect that they can account for the heritability void that has been left by GWAS with SNPs [7]. Nevertheless, only a few GWAS with CNVs have been conducted thus far [8, 9, 10], and, unlike those with SNPs [11], they have not become popular. A major reason would be that common CNVs, often referred to as copy number polymorphisms (CNPs), are well tagged by nearby SNPs and have been indirectly explored through SNP studies [12, 4], while rare CNVs may be impractical for the statistical power provided by GWAS. Another reason might be the need for statistical models that are both capable of capturing the complex patterns of allele-specific CNPs that slip into common commercial SNP arrays and sufficiently accurate in the genotyping of such CNPs from allele-specific fluorescent signal intensity data. Here, we present such a model and assess its ability to accurately capture the patterns of variations by applying it to detect allele-specific CNPs among genome-wide SNP markers and to infer allele-specific copy number (ASCN) genotypes on an individual scale from the fluorescent signal intensities of commercial SNP arrays.

This model was motivated by the observation that, at each SNP locus, fluorescent signal intensities tend to cluster into clouds of similar ASCN genotypes (Figure 1). This clustering tends to be vague because, as a result of instrumental errors, the intensity clusters that are proportional to the number of copies of a DNA segment that exists in nature will overlap with those that lie away from the center of other clusters. To capture this situation, for each cluster, we assumed that the signal intensities were observed around the (unknown) mean of a two-dimensional normal distribution and allowed for the cluster memberships to uncertainly overlap each other according to a mixture of a finite number of normal components, namely a Gaussian mixture model (GMM). The mixing proportions of the normal components were regarded here as the population frequencies of the ASCN genotypes. These frequencies are thought to follow from the population haplotype frequencies under the Hardy-Weinberg equilibrium (HWE) because most of the CNVs in any individual's genome arise not from de novo mutations but from CNPs shared with other individuals [13]. Our hope is that this assumption will allow the model to efficiently capture complex CNP patterns that are not well captured by an ordinary GMM, while reducing the number of model parameters. This assumption also make it possible to check whether the genotyping was successful in terms of the HWE test statistic (see Material and Methods for details).

We fit this GMM on a marker-by-marker basis; that is, we genotype the CNP for each marker independently across all the samples (in the same manner as SNP genotyping). This "cross-sample" approach seems to be more accurate than "cross-genome" approaches [14, 15, 16, 17, 18] (which often use a hidden Markov model to capture the segmental aneuploid events of each individual throughout the genome) because recent studies using higher-resolution arrays have shown that most CNVs are smaller than was initially reported [13]. These types of methods often struggle to capture common CNPs because they are not well supported by a sufficient number of SNP markers in lower-resolution commercial arrays (see Results). Another existing method that aims to capture CNPs by the cross-sample approach is TriTyper [19]. It employs a maximum likelihood (ML) procedure that minimizes the deviation of each SNP marker from HWE under the triallelic SNP assumption. The current limitation of TriTyper is that it can capture only CNPs with deletions. In this respect, Birdsuite [13] is the first attempt to capture allele-specific CNPs with both deletions and duplications in commercial SNP arrays, although support for chips and platforms other than the Affymetrix SNP 6.0 is currently limited. Indeed, Birdsuite employs a two-dimensional GMM similar to ours, to infer an individual's ASCN genotype and to provide a simultaneous, combined test of the copy number dosage (CN dosage) and the SNP allelic variation. This idea was proposed independently of our work, and the model-fitting strategy is different from ours; see Discussion.

Below, we describe our model, instantiated in software called PlatinumCNV. Our model is presented in a Bayesian framework with a well-conceived prior distribution that essentially performs the fully automated genome-wide genotyping of CNPs. Results using fluorescent signal intensity data from the Illumina HumanHap 610K platform demonstrate that more than 4,000 allele-specific CNPs of various kinds can be retrieved from the commercial SNP array. Because the genotyping accuracy is quite high, the estimated genotyping error rates using quantitative PCR (qPCR) and replicated subjects (genotyped twice in the sample set) are less than or equal to 1% of the mean values. The model also infers the CN dosage and the SNP allelic variation for each individual, which can be used for further analyses, such as linkage disequilibrium (LD) mapping and GWAS. Our hope is that the kinds of models and methods examined here will form the foundation for new, more effective statistical methods for the mapping of both disease genes and quantitative trait loci (QTL) with genome-wide CNPs, using commercial SNP arrays.

Material and Methods

Sample and Signal Intensity Data

Our sample included 30,685 self-identified Japanese individuals who were part of the BioBank Japan Project [20]. The BioBank Japan Project collected human genomic DNA after the donors provided written informed consent to participate in the project. This project was approved by the ethics committees of The Institute of Medical Science, The University of Tokyo and the Center for Genomic Medicine, Institute of Physical and Chemical Research (RIKEN). The subjects in the sample set were genotyped by Illumina HumanHap 610K commercial platforms, and 33 subjects were genotyped twice.

The normalized X/Y signal intensities of 600,470 SNPs on 22 autosomes were obtained from BeadStudio 3.1 software. Before CNP genotyping, we performed quantile normalization [21] across the genome for each subject's X/Y intensities independently. We obtained the manufacturer's annotations for the Illumina platform and calculated the GC contents for the 50 bp oligonucleotide probe sequences provided by the manufacturer. We removed the GC content effect for X/Y intensities independently by using a smoothing spline in R. We further removed the effect of the whole genome amplification for X/Y intensities independently by using LOESS [22].

Bayesian Gaussian Mixture Model

Although we used data from a specific platform, the concepts and approach described here constitute a general strategy that can be applied to any genotyping platform. Oligonucleotide assays, which are available for high-throughput SNP genotyping, usually measure the intensities of two fluorescent labels that are attached to two known alleles. Throughout this manuscript, we refer to those alleles (bases) as A and B. Essentially, at any common SNP marker, three clouds appear in the two-dimensional signal intensity plot (Figure 1a) corresponding to all of the possible pairs of alleles (AA, AB and BB). These clouds are identified as SNP genotypes with extremely high accuracy (see Results). Once the SNP locus is in a CNV region, extra clouds are present on the signal intensity plot (Figure 1b-d). These clouds are essentially seen as arising from a pair of unexpected allele-specific copy number (ASCN) haplotypes that are made up of alleles A and B at the SNP locus. Here, we assume that such haplotypes are composed of up to two alleles, and therefore $J = 6$ ASCN haplotypes

$$\{\text{O, A, B, AA, AB, BB}\} \tag{1}$$

are present at the SNP locus (they are essentially ASCN alleles at the locus, but we call them "haplotypes" to avoid confusion in the notation of the A and B "alleles" at each SNP locus). Here, haplotype O is called "null allele" [13, 19], which indicates that a DNA segment of one to several megabases surrounding the locus was deleted from a chromosome (Figure 2a). The AA, AB and BB haplotypes indicate duplications of segmental aneuploidy, in which a DNA segment surrounding the locus has been repeated in a chromosome (Figure 2a). A and B denote normal haplotypes that are to be genotyped by the SNP arrays as usual. This assumption implies that the number of copies in an individual's genome ranges from zero to four at the locus. This assumption seems necessary and sufficient when we compare it to other methods [15, 14, 23], and it is biologically more plausible than that of {loss, normal, gain} proposed in [18]. We assume that these haplotypes penetrate in a population as ordinary SNP alleles. Therefore, an individual is thought to have drawn a pair of those haplotypes, as an allele-specific CN diplotype, randomly from a population with (expected) haplotype frequencies $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_J)^\top$. Here, each π_j corresponds to the frequency of the jth haplotype in (1). Note that the assumption above is equivalent with the result as we assume the law of HWE. We can formally describe this phenomenon by using the multinomial distribution $\mathcal{M}(n; \boldsymbol{\pi})$ with the number of trials $n = 2$. If we denote $\boldsymbol{z} = (z_1, \ldots, z_J)^\top$ to be the diplotype configuration of the individual (*e.g.*, $\boldsymbol{z} = (0, 1, 0, 0, 1, 0)^\top$ indicating the A/AB diplotype at the locus), the probability of $\boldsymbol{z}$, given the population haplotype frequencies $\boldsymbol{\pi}$, is given by

$$p(\boldsymbol{z}|\boldsymbol{\pi}) = \frac{2}{\prod_{j=1}^{J} z_j!} \prod_{j=1}^{J} \pi_j^{z_j}. \tag{2}$$

Similarly to the haplotype phasing problem [24], such a diplotype $\boldsymbol{z}$ is usually unobserved. In our case, we can only observe the allele-specific signal intensity $\boldsymbol{x} = (x, y)^\top$ instead of $\boldsymbol{z}$, which suggests that we need a probabilistic model linking $\boldsymbol{x}$ and $\boldsymbol{z}$.

By looking at the signal intensities plot, we may confirm the multiple clouds, each of which is made up of the intensities whose source is a similar diplotype of configuration $\boldsymbol{z}$ that is shared among multiple individuals. Without any instrumental error, these clouds should be observed as nodes on a lattice (Figure 2b), and the location of each node is proportional to the numbers of A and B alleles. For example, the cloud nearest to the origin (blue node) indicates a cluster of individuals who have a pair of double O haplotypes; above is a cloud of those who have B and O haplotypes, the right is a cloud of those who have A and O haplotypes, and so on. Here, one may notice that there exist only $K = 15$ nodes as the maximum on the lattice from the ASCN haplotypes in (1), while the total number of diplotypes $\boldsymbol{z}$ of all

kinds is $J \times (J+1)/2 = 21$. Some of the clouds that are derived from different haplotype pairs completely overlap each other (*e.g.*, we may not be able to distinguish the cloud of O/AB from that of A/B). Therefore, we need to introduce another variable, such as ASCN genotype $\mathcal{G}$ ($\mathcal{G} = 1, \ldots, K$), as a medium of $\boldsymbol{x}$ and $\boldsymbol{z}$. As genotype $\mathcal{G}$ is exactly determined by a known diplotype $\boldsymbol{z}$, the probability $\mathcal{G}$, given $\boldsymbol{z}$, becomes

$$p(\mathcal{G}|\boldsymbol{z}) = \begin{cases} 1 & \mathcal{G} \sim \boldsymbol{z}, \\ 0 & \text{otherwise.} \end{cases}$$

Here, the operator "$\sim$" stands for "consistent with" (*e.g.*, the genotype AAB is consistent with both the A/AB and AA/B diplotypes). Therefore, the marginal probability of $\mathcal{G}$ is given by

$$p(\mathcal{G}|\boldsymbol{\pi}) = \sum_{\boldsymbol{z}} p(\mathcal{G}|\boldsymbol{z}) p(\boldsymbol{z}|\boldsymbol{\pi}),$$

which indicates the (expected) genotype frequency of $\mathcal{G}$, given the haplotype frequencies $\boldsymbol{\pi}$.

Supposing that each of the clouds specified by $\mathcal{G}$ follows a two-dimensional normal distribution $\mathcal{N}(\boldsymbol{\mu}_{\mathcal{G}}, \Sigma_{\mathcal{G}})$ with mean vector $\boldsymbol{\mu}_{\mathcal{G}}$ and covariance matrix $\Sigma_{\mathcal{G}}$, then the signal intensity $\boldsymbol{x}$ follows a Gaussian mixture model

$$p(\boldsymbol{x}|\theta) = \sum_{\mathcal{G}=1}^{K} p(\mathcal{G}|\boldsymbol{\pi}) \phi(\boldsymbol{x}|\boldsymbol{\mu}_{\mathcal{G}}, \Sigma_{\mathcal{G}}), \tag{3}$$

where $\phi(\boldsymbol{x}|\boldsymbol{\mu}, \Sigma)$ denotes the normal density and θ denotes a set of all model parameters. To determine the ASCN genotype and haplotype frequencies, we fit the model directly onto clouds in a two-dimensional fluorescent signal plot (Figure 2c). If we observe a set of signal intensities for N individuals, say $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)^\top$, then the posterior probability is given by $p(\theta|X) \propto \prod_{i=1}^{N} p(\boldsymbol{x}_i|\theta) p(\theta)$ with an appropriate prior distribution $p(\theta)$. We employ the Expectation-Maximization (EM) algorithm [25] to obtain the maximum a posteriori (MAP) estimation $\hat{\theta} = \text{argmax}_\theta \; p(\theta|X)$ (provided in the Supplementary Methods). For subsequent analyses, we can infer the MAP ASCN genotype

$$\hat{\mathcal{G}}_i = \underset{\mathcal{G}}{\text{argmax}} p(\mathcal{G}|\boldsymbol{x}_i, \hat{\theta}) \tag{4}$$

and the MAP CN dosage and MAP allelic variation

$$\hat{N}_i = \underset{n\in\{0,\ldots,4\}}{\operatorname{argmax}} \sum_{\mathcal{G};\, \#\mathcal{G}=n} p(\mathcal{G}|\boldsymbol{x}_i, \hat{\theta}), \tag{5}$$
$$\hat{M}_i = \underset{n\in\{-4,\ldots,4\}}{\operatorname{argmax}} \sum_{\mathcal{G};\, \#_{\mathrm{A}\backslash\mathrm{B}}\mathcal{G}=n} p(\mathcal{G}|\boldsymbol{x}_i, \hat{\theta}),$$

for each individual i ($i = 1, \ldots, N$), where $\#\mathcal{G}$ denotes the total number of alleles (A and/or B) that $\mathcal{G}$ contains (*e.g.*, $\#\mathcal{G} = 3$ for genotype AAB and $\#\mathcal{G} = 1$ for B) and $\#_{\mathrm{A}\backslash\mathrm{B}}\mathcal{G}$ denotes the number of allelic differences (*e.g.*, $\#\mathcal{G}_{\mathrm{A}\backslash\mathrm{B}} = 1$ for genotype AAB and $\#_{\mathrm{A}\backslash\mathrm{B}}\mathcal{G} = -2$ for genotype ABBB). Note that the posterior probability of $\mathcal{G}$, given $\boldsymbol{x}_i$, is obtained by Bayes' rule $p(\mathcal{G}|\boldsymbol{x}_i, \hat{\theta}) = p(\mathcal{G}|\hat{\boldsymbol{\pi}})\phi(\boldsymbol{x}_i|\hat{\boldsymbol{\mu}}_{\mathcal{G}}, \hat{\Sigma}_{\mathcal{G}}) / p(\boldsymbol{x}_i|\hat{\theta})$.

For association studies, it would be better to perform a statistical test of both the CN dosage and the allelic variation effects, because the observed intensities are allele-specific in nature. Here, we recommend the use of the posterior mean CN dosage $\bar{N}_i$ and allelic variation $\bar{M}_i$, rather than the MAP estimators $\hat{N}_i$ and $\hat{M}_i$, to minimize the differential bias [26]. The mean CN dosage and mean allelic variation can be easily calculated as

$$\bar{N}_i = \sum_{\mathcal{G}=1}^{K} \#\mathcal{G}\ p(\mathcal{G}|\boldsymbol{x}_i, \hat{\theta}), \tag{6}$$
$$\bar{M}_i = \sum_{\mathcal{G}=1}^{K} \#_{\mathrm{A}\backslash\mathrm{B}}\mathcal{G}\ p(\mathcal{G}|\boldsymbol{x}_i, \hat{\theta}).$$

Prior Settings

Here, we explain the need for prior distributions of θ in the GMM: several potential difficulties are present when using real fluorescent signal intensity data. As an example, by using a typical intensity plot as a reference (Supplementary Figure 1a), we often encounter plots with absolute scale differences (Supplementary Figure 1b-c) and signal imbalances in X/Y scales (Supplementary Figure 1d). For those signal plots, it appears necessary to introduce multiple starting points for the EM algorithm. Additionally, unobserved genotype clouds are technically problematic because certain model parameters will be undetermined. For example, if the population haplotype frequency $\pi_1 = 0$, then the parameters of the normal components $\{\boldsymbol{\mu}_{\mathcal{G}}, \Sigma_{\mathcal{G}}\}$ for $\mathcal{G} \in \{\mathrm{O}, \mathrm{A}, \mathrm{B}\}$ are absolutely undetermined. This means that, for ordinary GMMs, superfluous normal components for the unobserved clouds may interrupt other normal components to be fit appropriately on the predefined clouds.

Nevertheless, we fit the GMM of all the K normal components to maintain model flexibility, and we

applied the EM algorithm from the same initial values to reduce the computational burden. To overcome these issues, we introduced well-conceived priors for $\boldsymbol{\pi}$, $\boldsymbol{\mu} = (\boldsymbol{\mu}_1{}^\top, \ldots, \boldsymbol{\mu}_K{}^\top)^\top$ and $\Sigma_\mathcal{G}$ ($\mathcal{G} = 1 \ldots, K$), such that

$$\begin{aligned}
\boldsymbol{\pi} &\sim \mathcal{D}(\boldsymbol{a}), \\
\boldsymbol{\mu} &\sim \mathcal{LN}(\boldsymbol{m}, \delta^{-1} S), \\
\Sigma_\mathcal{G} &\sim \mathcal{W}^{-1}(\lambda V_\mathcal{G}, \lambda - 3);\ \mathcal{G} = 1, \ldots, K.
\end{aligned}$$

We assumed the Dirichlet prior on $\boldsymbol{\pi}$ with parameter values $a_j \ll 1$, $j = 1, \ldots, J$ because the haplotype frequency spectrum among CNPs is highly skewed toward zero and not uniformly distributed (see Results). This prior aided in making the observed haplotype frequency $\hat{\pi}_j = 0$ to reduce redundant parameters, if there is strong evidence from the data that the population haplotype frequency is $\pi_j = 0$, We also assumed the inverse Wishart prior on each covariance matrix $\Sigma_\mathcal{G}$ with an inverse scale matrix $\lambda V_\mathcal{G}$ and $\lambda - 3$ degrees of freedom. This prior essentially avoids the covariance matrix uncertainty for clouds with extremely low (or even zero) frequencies. For the compound vector $\boldsymbol{\mu}$ of all mean vectors $\boldsymbol{\mu}_\mathcal{G}$ ($\mathcal{G} = 1, \ldots, K$), we assumed a log-normal prior with mean vector $\boldsymbol{m}$ and covariance matrix $\delta^{-1} S$. This prior (especially the covariance structure specified by S) played the central role in our model because the values of the compound mean vector $\boldsymbol{\mu}$ were highly correlated in nature so as to keep the lattice structure of the genotype clouds. This prior also greatly aided in solving the absolute scale difference and the X/Y signal imbalance by multiplying δ^{-1} of a large value. As a consequence, the introduction of such prior distributions improved both the efficiency and the accuracy of the model fitting while maintaining the flexibility of the GMM.

To make the prior distributions more efficient, we improved all the hyperparameters, say η, except for δ and λ, of the prior distributions via the empirical Bayes estimation. If we let X_t be the signal intensity at the locus t ($t = 1, \ldots, L$) for all subjects in the sample set, then the empirical Bayes estimation of η among all SNP markers is given by

$$\hat{\eta} = \underset{\eta}{\operatorname{argmax}}\, L(\eta | X_1, \ldots, X_L)$$

where $L(\eta | X_1, \ldots, X_L) \propto \prod_{t=1}^{L} p(X_t | \eta)$ is the marginal likelihood of η. Here we needed to integrate out θ_t from $p(X_t, \theta_t | \eta)$ to obtain $p(X_t | \eta)$. We used the Laplace approximation [27] around $\theta_t = \hat{\theta}_t$, which had already been obtained by the MAP estimation of the model parameters. The use of the approximation could be the best solution to increasing the accuracy of the hyperparameter estimation while reducing the computational complexity of multiple integrations among genome-wide SNP markers (See Supplementary

Methods for more details).

The remaining hyperparameters $\delta > 0$ and $\lambda > 3$ were not estimated via the maximum marginal likelihood framework and were left as open (user-defined) parameters to be calibrated against the condition of the signal intensity data (from the viewpoint of the penalized likelihood method, these values are often referred to as the tuning parameters, which essentially control the impact of the penalties on the likelihood). We predefined the parameters as $\delta = N/3000$ and $\lambda = \max\{N/30, 3\}$, which followed from our informal studies of the CNP discovery and genotyping accuracy among genome-wide markers (data not shown). However, we estimate that such values are appropriate only for data sets that are similar to ours, and it would be better to calibrate these values according to the signal intensities of the data set to be analyzed. For example, if the data are much more heterogeneous than ours in terms of the absolute scale and the X/Y signal imbalance, then δ should be smaller to make the GMM more flexible. If the data are much noisier than ours, that is, if the genotyping clouds often overlap, then λ should be larger to make the GMM less flexible to prevent each normal component from diverging from a region of predefined cloud.

Quality Control Filters

We assessed the quality of the CNP genotyping results analogously to the SNP genotyping results. We introduced quality control filters to exclude CNP loci that did not meet specified thresholds of call rate, haplotype frequencies of deletion and duplications and the HWE test statistic. In addition, we also introduced cut-off values of overlapping rates between normal components.

We first removed the loci in which the EM algorithm did not converge after sufficient iterations (1,000 in our case). The convergence criterion was

$$|\log p(\theta^{(k)}|X) - \log p(\theta^{(k-1)}|X)| < 10^{-8}.$$

Next, we remove the monozygote loci (one of $\hat{\pi}_j$ is larger than 99%) and SNP loci (total frequency of A and B haplotypes is larger than 99%). To obtain the call rate, we treated subjects who meet $p(\boldsymbol{x}_i|\hat{\theta}) < 10^{-6}$ as outliers and assigned to them an "undetermined" genotype (such outliers are specified by "+" in, *e.g.*, Figure 2d). Therefore, the call rate is given by

$$(\text{Call Rate}) = \frac{\sum_{i=1}^{N} I\{p(\boldsymbol{x}_i|\hat{\theta}) > 10^{-6}\}}{N},$$

where $I\{\cdot\}$ is the indicator function. We removed loci with a call rate below 99%. We further assessed the

genotyping quality using the goodness-of-fit test of $p(\mathcal{G}|\hat{\boldsymbol{\pi}})$. Under HWE, the test statistic

$$\chi^2 = \sum_{\mathcal{G}=1}^{K} \frac{[\sum_{i=1}^{N} p(\mathcal{G}|\boldsymbol{x}_i, \hat{\theta}) - Np(\mathcal{G}|\hat{\boldsymbol{\pi}})]^2}{Np(\mathcal{G}|\hat{\boldsymbol{\pi}})}$$

may asymptotically follow a χ^2 distribution with $K - J$ degrees of freedom. We removed the loci with a HWE p-value $< 10^{-6}$, as is usual in a GWAS with SNPs.

We also removed loci that were in excess of a degree to which any pair of two normal densities overlap each other. The degree of overlapping is defined by the p-value that the estimated cluster mean $\hat{\boldsymbol{\mu}}_{\mathcal{G}'}$ is drawn from the other normal component $\mathcal{N}(\hat{\boldsymbol{\mu}}_{\mathcal{G}}, \hat{\Sigma}_{\mathcal{G}})$, such that

$$(\text{Overlapping Rate}) = \int_{d_{\mathcal{G}}(\boldsymbol{x}) > d_{\mathcal{G}}(\hat{\boldsymbol{\mu}}_{\mathcal{G}'})} \phi(\boldsymbol{x}|\hat{\boldsymbol{\mu}}_{\mathcal{G}}, \hat{\Sigma}_{\mathcal{G}}) d\boldsymbol{x}$$

where $d_{\mathcal{G}}(\boldsymbol{x}) = (\boldsymbol{x}^\top \hat{\Sigma}_{\mathcal{G}}^{-1} \boldsymbol{x})^{1/2}$.

Identification of Primer Polymorphisms in Illumina HumanHap 610k

We obtained the top genomic sequences, the oligonucleotide sequences around the SNP markers on the Illumina 610K, from the manufacturer's annotations. To identify from which strand the sequences came, we used BLAST [28] to map the top genomic sequences of 600,470 autosomal SNPs onto the human autosomal reference genome sequence (Build 36). We identified 596,354 autosomal SNPs whose top genome sequences were mapped onto the reference sequence. From this map information, we derived the physical genomic positions at which the 50 bp Illumina probes annealed and determined whether more polymorphisms had been described within these loci in dbSNP (build 130). We identified 79,159 SNPs that had more than one polymorphism within the 50 bp Illumina probe.

Other Statistical Methods

We compared the results of PlatinumCNV with those of two previously mentioned methods, TriTyper and PennCNV. The log R ratio, B allele frequency and SNP genotype for 600,470 loci on 22 autosomes were obtained using BeadStudio 3.1 software, with which we ran PennCNV with default settings. For TriTyper, we used our normalized X/Y intensities after the quantile normalization and GC correction.

qPCR Validation

To assess the accuracy of our method, we used a benchmark CNP data set of discovered 164 loci previously genotyped using qPCR. TaqMan assays were performed following the method described in [29] for 752 subjects drawn from our sample. All primers and TaqMan probes were designed using Primer Express software v2.0 (Applied Biosystems). The RNaseP assay (Applied Biosystems) was used as an internal control. All assays were performed on an ABI 7900 (Applied Biosystems) using TaqMan Gene Expression PCR Master Mix (Applied Biosystems) with 10 ng of genomic DNA in a 10 μl reaction. The cycling conditions were as follows: 95 degrees Celsius for 10 min for the initial denaturation and enzyme activation, followed by 40 cycles each of 95 degrees Celsius for 15 sec and 60 degrees Celsius for 1 min. The copy number was calculated using the comparative *Ct* method [30], where *Ct* indicates the threshold cycle. The validated loci are categorized into three subtypes according to the existing CN genotypes: there are 97 loci with deletions, 25 loci with duplications and 42 loci with both deletions and duplications (which we refer to as complex CNPs). Note that the qPCR assays have been performed before (and independently of) our work. Therefore, the loci were considered to be selected randomly from entire genomes with different allele frequencies (see supplementary online table in Web Resources section for details). All statistical analyses using PlatinumCNV, PennCNV and TriTyper are guaranteed equality in error rate evaluation.

Results

Identified CNP Loci among Genome-wide SNP Markers

We ran PlatinumCNV on our sample, and we obtained 4,256 CNP loci (Table 1) that passed the quality control filters (detailed information on the identified 4,256 CNPs is also provided as a supplementary online table). We observed 3,241 out of 4,256 (76%) CNPs with the ASCN haplotype O whose frequencies were greater than 1% in the population (we refer to these as "deletions"), 981 out of 4,256 (23%) CNPs with at least one of the ASCN haplotypes (AA, AB or BB) whose frequencies were greater than 1% in the population (we refer to these as "duplications") and 34 out of 4,256 (1%) CNPs with both haplotype O and at least one of the AA, AB or BB haplotypes whose frequencies were greater than 1% in the population (we refer to these as "complex CNPs", also known as "multi-allelic CNPs", in, *e. g.*, [31]). For duplications, we further classified CNPs into ones with the ASCN haplotypes AA, AB or BB, whose frequencies in the population were greater than 1% (we refer to these as "AA duplications", "AB duplications" or "BB duplications", respectively), and we found 369 (8.7%) loci with AA duplications, 311 (7.3%) with AB duplications and 326

(7.6%) loci with BB duplications.

The haplotype frequency spectrum clearly shows that the frequencies of aneuploid haplotypes are quite low (Supplementary Figure 2); therefore these distributions were skewed toward the haplotype frequencies being zero (note that this fact is due to the design of the SNP arrays, not to their biological nature). The average frequency for haplotype O was only 8.0%, and for these AA, AB and BB haplotypes it was 3.3%, 1.3% and 1.8%, respectively. In addition, the aneuploid haplotypes of the duplications were rare and almost never seen simultaneously. Therefore, the mean frequency of the total duplication haplotypes was 6.3%.

For a better understanding of the allele-specific features of the identified CNPs, we further classified the identified CNPs into various ASCN genotype patterns. As a combination of ASCN haplotypes whose frequencies are $\geq$1% in the population, we observed 24 different patterns of ASCN genotypes (Supplementary Figure 3a-b). We found that 2,514 out of 4,256 (59%) CNPs were SNPs with deletions, that is, a combination of ASCN haplotypes O, A and B. Additionally, 4,195 out of 4,256 (99%) CNPs arose from the combination of just one extra aneuploid haplotype with the normal A and/or B haplotypes. This fact strongly suggested the impoverishment of complex CNPs in common commercial SNP arrays. The more ASCN genotype clouds a SNP locus exhibited in the signal intensity plot, the more likely that the locus had been removed during the design of the SNP arrays. In fact, the discovery percentage of the complex CNPs was substantially lower than the 5 to 7% reported in recent studies [31, 4]

For these CNPs, we derived genomic regions of known CNVs in the Database of Genomic Variants (DGV; Build 36, Nov. 2010). We found that more than 60% of the total CNPs were observed within the known CNV regions, in contrast to 34% of the total SNP markers. For complex CNPs in particular, 88% were found within the known CNV regions. Although the total number of complex CNPs was quite small, the DGV percentage was significantly higher than expected (Fisher's exact p-value$<$0.001), which suggests that the vast majority of the common and complex CNVs have been explored in the past few years. Interestingly, the DGV percentage for AB duplications was also 88%. An explanation of these results is that these CNPs were significantly enriched within a specific region on the short arm of chromosome 8 (Supplementary Figure 4).

To confirm that CNPs were more likely to be found if they were concentrated within a region, we analyzed the number of identified CNPs that were present next to one another within the scope of SNP marker location. We found that 1,005 out of 4,256 (24%) CNPs spanned multiple adjacent markers in the Illumina 610K, and more than 95% of them were observed in the known CNV regions. This finding implied that those adjacent CNPs might exist within a relatively long CNV region that spans more than 5 kbp on aver-

age (the average distance of the SNP markers in the Illumina 610K is 5 kbp, or approximately 3 billion bp / 610K). In other words, the longer the CNV region, the more likely it is that the region has already been reported. As we expected, 78 out of 311 AB duplications were found next to one another. The number of CNPs was twice as large as the number for AA or BB duplications, and the DGV percentage was almost 100% for those adjacent CNPs.

We also investigated the enrichment and impoverishment of the genomic features associated with the identified CNPs (Table 2). In total, we found 1,646 out of 4,256 (39%) CNPs that were present within 1,196 out of 21,061 RefSeq genes, which was significantly less than for the total of the SNP markers (45% on average). Indeed, this impoverishment, especially in the coding regions (exons), was strongly associated with deletions, which probably reflects a greater purifying selection acting on deletions of genes than on duplications or SNPs [4, 31]. It is also worth noting that AB duplications were more likely to be found in genic regions than in intergenic regions. However, 99 out of the 311 AB duplications were involved with the same gene (*CSMD1*, whose length is more than 2 Mbp). Additionally, this gene contains a number of CNV regions (reported in DGV), in which all 99 AB duplications were involved. Therefore, the results for enrichment in both DGV and genic regions must have been strongly biased by the Illumina 610K array.

By comparing the CN dosages of the CNPs with the genotypes of nearby biallelic SNPs, we found, consistent with earlier studies [12, 4, 13], that deletions exhibit strong LD with SNPs (Table 3 and Supplementary Figure 5). Although the discovery percentage was somewhat lower than the 81% reported in the most recent study [6], 846 out of 1,518 (56%) common deletions were strongly correlated with one or more SNPs (Pearson's $r^2 > 0.8$). On the other hand, only 159 out of 487 (33%) common duplications were tagged by nearby SNPs. One reason would be that the average frequency of total duplications was lower than that of deletions in our CNPs. Another reason might be due to a higher genotyping error rate for such duplications than for deletions, as the genotype clouds of higher CN overlap each other, and misclassifications can then easily occur (see the next section for details).

Finally, we repeated the above analyses for two of the aforementioned methods, TriTyper and PennCNV. PennCNV and TriTyper (with imputation) detected 3,745 and 2,863 CNPs, respectively (Table 1), that met specific thresholds for call rate ($\geq$ 99%), normal haplotype frequency ($<$ 99%) and the HWE test P-value ($\geq 1.0 \times 10^{-6}$). We observed almost similar results by TriTyper as were obtained from the deletions identified by PlatinumCNV (Table 1-3). The average frequency of deletions was 7.4%, and 66% of common deletions were tagged by nearby SNPs after imputation. Interestingly, for such deletions, only 1,509 loci were jointly observed by TriTyper and PlatinumCNV, and 3,086 loci were missed by one or the other (Supplementary

Figure 6). As for PennCNV, we observed an entirely different aspect of CNPs. The average frequencies of deletions and duplications in the population were only 1.1% and 1.3 %, respectively. The number of CNPs tagged with nearby SNPs was only 102 (2.7%) loci in total. There were more than twice as many identified duplications as there were deletions, and almost all loci were described in DGV (Table 1). Indeed, 1,112 out of 2,234 (50%) loci with duplications existed next to one another within a 1.6Mb region overlapping the *CSMD1* gene. Other detailed information on the identified CNPs for these methods is also provided in supplementary online tables (see Web Resources section), along with that for Platinum CNV.

Validation and Power Analysis for Association Studies

We assessed the quality of the identified CNPs and inferred genotyping error rates by two approaches: one was to compare the replicated MAP ASCN genotypes for individuals who were genotyped twice in our sample, and the other was to validate the MAP CN dosage using qPCR [29].

We compared the MAP genotypes for each of the 33 replicated subjects. To estimate the error rate, we employed a simple error model and the ML technique (see Supplementary Methods for details). Figure 3a (and Supplementary Table 1) showed that, with the 3,241 deletions, the mean error rate became 0.71% due to the well-separated genotype clouds around the origin of the signal intensity plot. With the 981 duplications, the mean error rate was 2.0%, which was three times larger than that of the deletions. This large value was due to the nature of SNP array intensities, in which the signal is too saturated to distinguish the higher number of copies. The genotyping error rates for complex CNPs were also worse than that for deletions; they also diverged, ranging from 0.0 to 8.9%. As a consequence, the mean error rate for the 33 subjects using 4,256 CNPs was 1.02%. Note that the SNP genotyping error rates were substantially lower than those for the CNPs; the mean error rate was 0.026% among these subjects, which demonstrates how reliable the SNP genotyping technology is.

As a complementary approach, we compared the MAP CN dosage with the qPCR results for a subsample of 752 subjects drawn from our sample (see Material and Methods). For each of the 164 loci genotyped by qPCR assays, we assessed 10 markers (in Illumina 610K) around the assay position and estimated the genotyping error rate by dividing the number of mismatches by the subsample size. Then, we selected a CNP that exhibited the minimum error rate among the 10 markers and that also met specific thresholds of call rate ($\geq$ 99%), normal haplotype frequency ($<$ 99.5%) and the HWE test P-value ($\geq 1.0 \times 10^{-6}$). We observed 95 out of 97 deletions that passed the QC filters, and the mean error rate was 0.25 (Figure 3b and Supplementary Table 1). For duplications and complex CNPs, we observed estimated mean error

rates of 1.23 and 1.27, respectively. These values were much higher than for deletions but lower than those estimated using the replicated subjects. The mean genotyping error rate for all loci was 0.62%.

We repeated the same analyses for PennCNV and TriTyper. For loci with deletions, TriTyper with imputation produced a lower error rate (0.6%) than those of PennCNV and PlatinumCNV with the replicated subjects. This result suggested that TriTyper can take advantage of the accurate genotyping results for biallelic SNPs to successfully boost the concordance rate between the first and replicated samples. However, we found that, with qPCR validation, TriTyper detected not only deletions but also complex CNPs with a sufficient deletion frequency. Because the current TriTyper cannot genotype duplication subjects, the error rate of a complex CNP increases as the proportion of duplication haplotypes increases in the population (Figure 3b). For loci with duplications, PennCNV produced a lower error rate (0.6%) than those of the other methods for the replicated subjects. PennCNV and TriTyper outperformed PlatinumCNV with the replicated subjects, while PlatinumCNV produced substantially lower error rates than the other methods with qPCR validation (Supplementary Table 1). We will discuss this issue later.

We also assessed how the power of an association study decays when genotyping error occurs (see Supplementary Methods for details of the power calculation). As we expected, the power tends to decrease as the genotyping error rate increases, and the reduction rate of the power curve for a rare allele is much larger than for a common allele (Figure 3c-e). In particular, in the case of a risk allele frequency of 3%, the power fell to below 40% after the genotyping error rate reached 5% and was almost 0% when the genotyping error rate reached 15%, although the initial power was higher than 90% (Figure 3e). This result suggested that genotyping error is much more crucial for loci with rare risk alleles than ones with common risk alleles. As the haplotype frequency of allele-specific CNPs in commercial SNP arrays is potentially lower than that of SNPs, we must increase the accuracy of CNP genotyping to retrieve novel variants associated with various phenotypes embedded in genome-wide CNPs.

GWAS of Three Hematological Traits

Using our CNP genotyping result, we conducted preliminary GWAS of three hematological traits, white blood cell (WBC), red blood cell (RBC) and platelet (PLT) counts, which were previously analyzed using genome-wide SNP genotypes in a Japanese population [32]. We enrolled the same subjects (N=14,700) as in [32], which formed the sub-sample of our sample for CNP genotyping. A linear regression model was used to analyze associations of the quantitative traits with CNP genotypes by incorporating age, square of age, gender, the first and second eigenvectors (result of population structure analysis using SNP genotypes)

and affection status of 19 diseases as covariates. To simplify the GWAS, we used only the mean CN dosage in (6) and assumed an additive effect. The inflation factors of P-values, λ_{GC} [33], were as low as 1.003-1.096 (Supplementary Figure 7), which suggested that no structural genotyping error [26] existed in our data. Manhattan plots (Figure 4) showed associations with WBC and PLT in the MHC region that satisfied the genome-wide significance level for the identified CNPs ($P < 1.0 \times 10^{-5} \approx 0.05/4,256$). We identified no novel association for these traits at the genome-wide significance level for total SNP markers ($P < 5.0 \times 10^{-8}$). In addition, all CNPs except rs3130981 could have been tagged by surrounding SNPs with $r^2 > 0.8$ (Table 4).

After careful consideration of LD with adjacent markers, we found that the duplication pattern at rs3130981 was probably because of the 3-base deletion polymorphism (rs9278982) in the vicinity of 12 bases upstream of the CNP, at which the 50-base probe sequence annealed. This phenomenon is often referred to as "primer polymorphism" [19] and occurs when too many SNPs and InDels are concentrated within a short region (*e.g.*, the MHC region) and probe design was failed to correctly capture signal intensities of the target SNP (see Discussion section for details). Because this primer deletion may be linked with the B allele at rs3130981 (thus causing the B allele signal intensity to be lower than usual), we carried out re-clustering of the CNP to highlight the deletion (Supplementary Figure 8) and re-assessed the association with the WBC count. As we expected, the remote deletion was strongly associated with WBC count ($P = 1.8 \times 10^{-10}$) and also strongly correlated with the SNP (rs3094212) that was previously reported (see Table 1 in [32]).

Discussion

The development and validation of novel approaches to accurately mapping ASCN genotypes using the commercial SNP arrays is important for the discovery of novel variants that account for the heritability void left by GWAS with SNPs. The oligonucleotide assay platform that was originally developed for SNP genotyping has been successfully used to capture various kinds of CNPs. Here, we have presented a new statistical model instantiated in software called PlatinumCNV that employs a Bayesian GMM with calibrated prior settings. Our results showed the power of PlatinumCNV in genotyping allele-specific CNPs with sufficient accuracy and also demonstrated that it could form the foundation for new, more effective statistical methods for the mapping of both disease genes and quantitative trait loci with genome-wide CNPs. Although the percentage of CNPs involved in genic regions was lower than for SNPs, and although aneuploid haplotypes were rare in the commercial SNP array, the SNP array-based approach itself may

perform well for the extremely quick identification of genome-wide CNPs for tens of thousands of subjects without any extra cost. This fact makes these platforms a viable and efficient complementary technology to classic qPCR assays, which are essentially impractical for large-scale sample sets.

The Bayesian probabilistic framework of PlatinumCNV may provide considerable flexibility for extending the model to specific applications. In genetic and genomic studies, we often encounter a population in which the HWE assumption does not hold (*e.g.*, a mixture or admixture population or cancer genomes that might strongly deviate from HWE). One extension of our model would be to introduce the "inbreeding model" [34] on the ASCN diplotype $\boldsymbol{z}$. In this case, we would need to introduce only one additional parameter, the inbreeding coefficient f, into the model. Then the model would remain computationally tractable and also would be stable in the convergence of the EM algorithm. Another extension of the current PlatinumCNV would be to allow the number of copies in the genotype to be larger than four. In commercial SNP arrays, the signal intensity is too saturated to distinguish the higher number of copies that are present in nature. However, we found several SNP loci in the Illumina 610K, that exhibited more than four copies of a genotype in practice (Supplementary Figure 9). To capture such higher copy numbers, we can readily introduce extra normal components in the GMM for the ASCN genotype clouds with more than four copies. Here, the issues of both computational burden and algorithmic stability concerning additional model parameters should be addressed properly. Another novel aspect of our work is the extension of our model to deal with multiple adjacent markers for a dense SNP array. This extension can be readily implemented using the hidden Markov model (HMM), where ASCN diplotype state $\boldsymbol{z}_t$ for each locus t ($t = 1, \ldots, T$) is thought of as a hidden state in the Markov chain (as in cnvHap [35]). We expect that a consistent improvement in performance could be obtained by taking into account the correlations among markers (*i.e.* the linkage disequilibrium [LD]). That is, by sharing haplotype information among adjacent markers, the HMM makes it possible to capture the long and rare CNVs crossing multiple markers and also to improve the accuracy of genotyping for highly overlapping clouds for common CNP genotypes.

We believe that our approach is the first application of the Bayesian GMM to the fluorescent signal intensities from commercial SNP array platforms. Here, we mention the potential differences between the model-fitting approaches of PlatinumCNV and Birdsuite. Birdsuite employs possible submodels with different CNP patterns and compares them using several criteria (*e.g.*, BIC and HWE penalty score) to seek a model that better fits the given data. In contrast, we always fit the full model in (3) to the signal intensity data and maximize insight into the complex patterns of CNPs (deletions and duplications simultaneously) by maximizing the posterior probability. The major drawback of our approach is that the model usu-

ally becomes redundant in terms of the model parameters. To overcome this issue, we have introduced well-conceived prior distributions in the Bayesian framework that solve the potential parameter excess. In addition, hyperparameters in the prior distributions were calibrated via empirical Bayes estimation. As a consequence, our approach seems to be more accurate than Birdsuite. Although, we have not compared the results of PlatinumCNV to those of Birdsuite (because of the genotyping platform difference), our qPCR validation revealed that the accuracy of CNP genotyping is slightly better than that reported in [23].

By comparing two other methods (PennCNV and TriTyper) with PlatinumCNV, we still observed minimum error rates for all subtypes of CNP with qPCR validation. It is notable that the estimated error rate with qPCR validation was substantially lower than that found with the replicated subjects. This reduction is partly because we selected the lowest error rate at a locus among 10 adjacent markers that spanned a CNV region, and therefore it might be underestimated when compared to the true underlying error rate. Nevertheless, it is true that one of the CNPs within the CNV region could exhibit such a low error rate. On the other hand, two other methods produced substantially lower error rates with the replicated subjects than with qPCR validation. We suspect that those error rates were underestimated for the following reasons: Although TriTyper could improve the consistency between the first and replicated samples by using nearby SNP information, the imputation method does not essentially correct the genotyping error at the target CNP when a recombination event occurs between a nearby SNP and the CNP. This hypothesis was supported by the qPCR validation; the initial deletion of TriTyper produced lower error rates than for the imputed deletions. In this regard, the error rates for the imputed deletions with the replicated subjects must be underestimated because the result indirectly reflects the genotyping accuracy of nearby SNPs rather than of the target deletions. As for PennCNV, we have already mentioned that the identified CNPs include a massive duplication in which more than 1,000 markers were involved. This region is extraordinary compared with other regions, as the number of loci that support a single CNV region was only 6 as a median value for all regions detected by PennCNV. After removing such extreme CNPs, which involved more than 100 markers, the error rates with the replicated subjects significantly increased (Paired T-test $P = 5.3 \times 10^{-12}$; Supplementary Table 1). This result suggests that the number of loci supporting the CNV region is crucial for the cross-genome approach, to preserve better genotyping accuracy. Therefore, the potential error rate of PennCNV for common and short CNPs would be as much worse than for the qPCR validation. We therefore concluded that both of these methods that rely on information across the genome should be applied for higher-resolution SNP arrays to take advantage of their characteristics.

We also assessed the impact of sample size using three subsamples (N=1,000, 5,000 and 10,000) of our

largest sample (Supplementary Table 2-4). We found that the number of CNPs discovered from the SNP array was strongly affected by the sample size of 1,000. This result might suggest that we need sample size of several thousands for sufficient power to detect genome-wide CNPs. In terms of genotyping accuracy, the larger the sample size we have, the lower the genotyping error rate we get (see Supplementary Table 4 and Supplementary Figure 10 and 11). However, the estimated genotyping error rates did not substantially increase as the sample size decreased; the error rates for total CNPs estimated by the two different approaches were both less than 1% as median values.

Although our model can capture complex patterns of allele-specific CNPs with sufficient accuracy, there is the possibility that a portion of the identified CNPs were not due to segmental aneuploidy but rather to biological artifacts: primer polymorphisms [19] and dispersed segmental duplications [12], which can easily contaminate oligonucleotide assays. In principle, the primer and probe sequences in the oligonucleotide assay should be unique on entire genomes. However, if another variant is present within the immediately adjacent locus that is complementary to a primer (probe) of the target SNP (Supplementary Figure 12a), then it affects the hybridization efficacy of the primer and diminishes the signal intensity. We found that 1,008 out of 4,256 (24%) CNPs involved such primer polymorphisms in the probe region of Illumina 610K (Table 1). We may not be able to distinguish such affected loci by looking at the signal intensity plot alone, because the plots showed a notable resemblance to true deletions and duplications (Supplementary Figure 12b-c). For all markers in the Illumina 610K, the histogram of distances to the primer polymorphisms revealed that the SNP marker is designed so that the 50 bp probe contains fewer polymorphisms within 10 bp of the target (Supplementary Figure 12d), whereas the distribution conditional on the identified deletions or the AA and BB duplications was far from that for the total markers (Supplementary Figure 12e-f). This result may imply that the vast majority of the 1,008 CNPs were not actually CNPs but were instead affected by the primer polymorphisms.

Similar phenomena will occur when a dispersed segmental duplication exists within a CNV region. For instance, the LD mapping of CNPs and biallelic SNPs revealed that the mean CN dosage of rs416858 on chromosome 22 was highly correlated ($r^2 > 0.9$) with SNPs in chromosome 17 (Supplementary Figure 13a-c) around which the top genomic sequence of rs416858 also mapped without any mismatch or gap (Supplementary Figure 13d). In addition, the location at which the sequence was mapped was known to be a common CNV region (DGV; Build 36, Nov. 2010). Therefore, we concluded that the SNP marker (rs416858) is not a CNP (because no duplication exists on chromosome 22), but the segmental aneuploidy on chromosomal 17 was indirectly captured by the SNP marker though the dispersed segmental duplication.

As for the other CNPs, we could not find such concrete evidence of dispersed segmental duplication, but the analysis of the top genomic sequences by BLAST showed that some of the identified CNPs may be affected by dispersed segmental duplications. For 2.4% of total CNPs, the top genomic sequences arose more than twice on different chromosomes, which is approximately 1.7 times more frequently than the total for the SNP markers (Table 1). Here, AB duplications were, in turn, strongly associated with the interchromosomal duplications of the top genomic sequences, whereas AA and BB duplications were not. This fact might suggest that the development of AB duplications is biologically different from that of AA or BB duplications. Similarly, for 1.4% of deletions, the top genomic sequences were unmapped to the reference sequence, which is approximately twice as frequent as that of the total SNP markers. These deletions were seen arising from true aneuploid events. However, we should be careful in deriving the physical genomic positions of such deletions, because targeted loci were missing in the reference sequence.

The results presented in this manuscript are preliminary and were made after an initial analysis of the commercial SNP array data based on a consideration of PlatinumCNV alone. Clearly, further investigation would be required to probe in more detail the biological issues raised here. We would emphasize here that the great value of PlatinumCNV is that it can provide a foundation for GWAS with CNPs using commercial SNP arrays, which will often suggest potential avenues for the subsequent mapping of disease genes and quantitative trait loci. As shown in our preliminary GWAS, we found the deletion polymorphism (rs9278982) in the 3' UTR of the *CDSN* gene, which could have a functional impact and downregulate the amount of WBC, although the association of the *CDSN-PSORS1C1* genes has already been implicated by the tag-SNP in intron (rs3094212). We hope that this kind of analysis will reveal additional insights into various biological questions along with the GWAS with SNPs.

Web Resources

R project (http://www.r-project.org/);

Database of Genomic Variants (http://projects.tcag.ca/variation/);

dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/);

PlatinumCNV software (http://kumasakanatsuhiko.jp/projects/platinumcnv/);

Supplementary Table Online (http://kumasakanatsuhiko.jp/projects/platinumcnv/paper/index.cgi).

Acknowledgments

We thank Prof. Seiya Imoto at the university of Tokyo and Prof. Ryo Yamada at Kyoto university for helpful suggestions and comments. We also thank all technical staff of BioBank Japan Project and Laboratory for Genotyping Development at RIKEN for SNP genotyping. This work was conducted as a part of the BioBank Japan Project.

References

1. Feuk, L., Carson, A. R., et al. (2006) Structural variation in the human genome. Nature Reviews Genetics. 7, 85-97.

2. Sebat, J., Lakshmi, B., et al. (2004) Large-scale copy number polymorphism in the human genome. Science. 305, 525-8.

3. Redon, R., Ishikawa, S., et al. (2006) Global variation in copy number in the human genome. Nature. 444, 444-54.

4. Conrad, D. F., Pinto, D., et al. (2010) Origins and functional impact of copy number variation in the human genome. Nature. 464, 704-12.

5. Park, H., Kim, J. I., et al. (2010) Discovery of common asian copy number variants using integrated high-resolution array CGH and massively parallel dna sequencing. Nat. Genet. 42, 400-5.

6. Mills, R. E., Walter, K., et al. (2011) Mapping copy number variation by population-scale genome sequencing. Nature. 470, 59-65.

7. Manolio, T. A., Collins, F. S., et al. (2009) Finding the missing heritability of complex diseases. Nature. 461, 747-53.

8. McCarroll, S. A., Huett, A., et al. (2008) Deletion polymorphism upstream of irgm associated with altered irgm expression and crohn's disease. Nat. Genet. 40, 1107-12.

9. Glessner, J. T., Wang, K., et al. (2009) Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. Nature. 459, 569-73.

10. Myocardial Infarction Genetics Consortium,, Kathiresan, S., et al. (2009) Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. Nat. Genet. 41, 334-41.

11. Hindorff, L. A., Sethupathy, P., et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. PNAS. 106, 9362-7.

12. Wellcome Trust Case Control Consortium, (2010) Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature. 464, 713-20.

13. McCarroll, S. A., Kuruvilla, F. G., et al. (2008) Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat. Genet. 40, 1166-74.

14. Wang, K., Li, M., et al. (2007) PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res. 17, 1665-74.

15. Colella, S., Yau, C., et al. (2007) QuantiSNP: an objective Bayes hidden-Markov model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res. 35, 2013-25.

16. Fridlyand, J., Snijders, A. M., et al. (2004) Hidden Markov models approach to the analysis of array cgh data. J. Multivariate Anal. 90, 132-153.

17. Marioni, J. C., Thorne, N. P., et al. (2006) Biohmm: a heterogeneous hidden Markov model for segmenting array cgh data. Bioinformatics. 22, 1144-1146.

18. Pique-Regi, R., Monso-Varona, J., et al. (2008) Sparse representation and bayesian detection of genome copy number alterations from microarray data. Bioinformatics. 24, 309-18.

19. Franke, F., de Kovel, C. G., et al. (2008) Detection, imputation, and association analysis of small deletions and null alleles on oligonucleotide arrays. Am. J. Hum. Genet. 82, 1316-33.

20. Nakamura, Y. (2007) The biobank japan project. Clin. Adv. Hematol. Oncol. 5, 696-7.

21. Bolstad, B. M., Irizarry, R. A., et al. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. Bioinformatics. 19, 185-93.

22. Marioni, J. C., Thorne, N. P., et al. (2007) Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. Genome Biology. 8, R228.

23. Korn, J. M., Kuruvilla, F. G., et al. (2008) Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat. Genet. 40, 1253-60.

24. Excoffier, L., Slatkin, M. (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. Mol. Biol. Evol. 12, 921-927.

25. McLachlan, G. J., Krishman, T. (1997) The EM Algorithm and Extensions. New York: John Wiley & Sons.

26. Plagnol, V., Cooper, J. D., et al. (2007) A method to address differential bias in genotyping in large-scale association studies. PLoS Genet. 3, e74.

27. Pawitan, Y. (2001) In all likelihood: statistical modelling and inference using likelihood. New York: Oxford University Press.

28. Altschul, S. F., Gish, W., et al. (1990) Basic local alignment search tool. J. Mol. Biol. 215, 403-10.

29. Hosono, N., Kato, M., et al. (2009) CYP2D6 genotyping for functional-gene dosage analysis by allele copy number detection. Clinical Chemistry. 55, 1546-54.

30. Bodin, L., Beaune, P. H., et al. (2005) Determination of cytochrome P450 2D6 (CYP2D6) gene copy number by real-time quantitative pcr. J. Biomed. Biotechnol. 3, 248-253.

31. The International HapMap Consortium, (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature. 449, 851-61.

32. Kamatani, Y., Matsuda, K., et al. (2010) Genome-wide association study of hematological and biochemical traits in a japanese population. Nat Genet. 42, 210-215.

33. Devlin, B., Roeder, K. (1999) Genomic control for association studies. Biometrics 55, 997-1004.

34. Wakefield, J. (2010) Bayesian methods for examining hardy-weinberg equilibrium. Biometrics. 66, 257-65.

35. Coin, L. J. M., Asher, J. E., et al. (2010) cnvHap: an integrative population and haplotype-based multiplatform model of SNPs and CNVs. Nature Methods. 7, 541-6.

36. Magnus, J. R., Neudeker, H. (1988) Matrix Differential Calculus with Applications in Statistics and Econometrics. New York: John Wiley.

37. Louis, T. A. (1982) Finding the observed information matrix when using the em algorithm. Journal of the Royal Statistical Society. Series B. 44, 226-33.

38. Chapman, J. M., Cooper, J. D., et al. (2003) Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. Hum. Hered. 56, 18-31.

Figure Titles and Legends

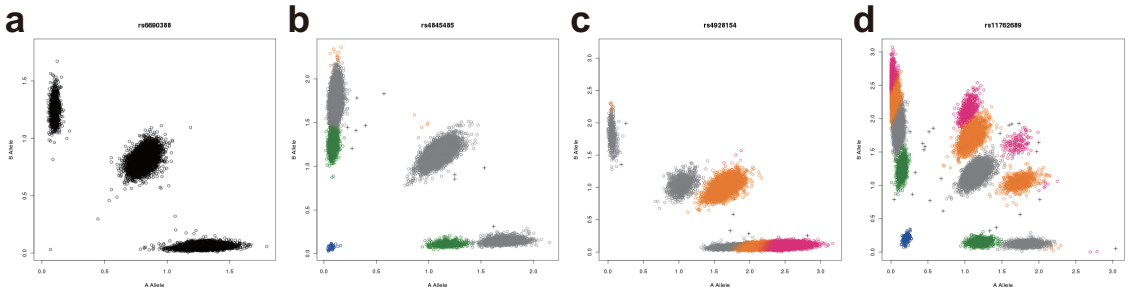

Figure 1. Example of fluorescent signal intensity plots at four SNP loci in Illumina 610K. (a) Fluorescent singal intensity plot at rs6690388 shows a typical SNP pattern with three clouds. Here the x-axis corresponds to the allele A and the y-axis corresponds to the allele B, respectively. (b) Fluorescent singal intensity plot at rs4845485 within a CNV region reported in, *e.g.*, [13], where subjects with deletions were observed. (c) Fluorescent singal intensity plot at rs4928154 within a CNV region reported in, *e.g.*, [4], where subjects with duplications were observed. (d) Fluorescent singal intensity plot at rs11762689 within both CNV region reported in, *e.g.*, [5], where subjects with both deletions and duplications were observed.

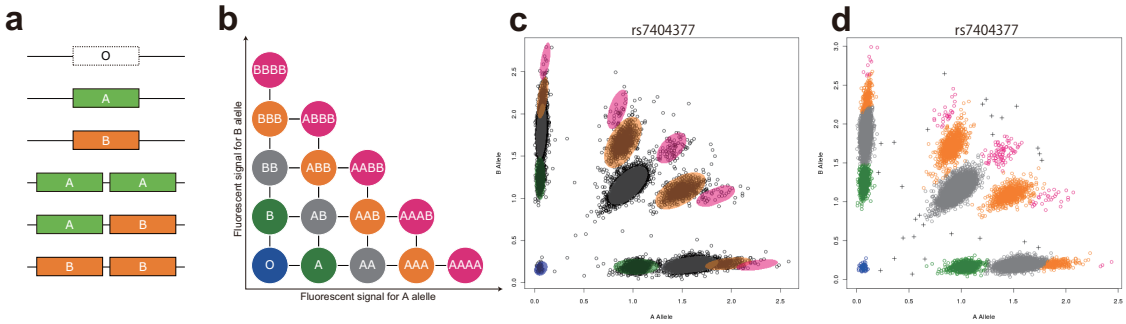

Figure 2. Concepts and approach of CNP genotyping using a Gaussian mixture model. (a) Stylized picture of the allele-specific copy number (ASCN) haplotypes. A rectangle with a letter A or B indicates a homologous DNA segment ranging from 1 to several megabases surrounding the targeted SNP locus on a chromosome. The rectangle with the letter O indicates the homologus segment is deleted from a chromosome. (b) Stylized picture of the possible 15 ASCN genotype clouds $\mathcal{G}$ on a two-dimensional fluorescent signal intensity plot. The x-axis corresponds to the fluorescent signal of the allele A and the y-axis corresponds to the fluorescent signal of the alelle B. A pair of two ASCN haplotypes specifies a ASCN genotype whose location is proportional to the number of alleles A and B involved in the genotype. (c) The Gausian mixture model (GMM) based on the fifteen ASCN genotypes fitted on the real fluorescent signal intensity plot at rs7404377. A colored ellipse shows 99% prediction interterval of each normal component. (d) The signal intensity plot at rs7404377 colored by the MAP genotype infered by the GMM. The undetermined genotype is specified by the black cross "+".

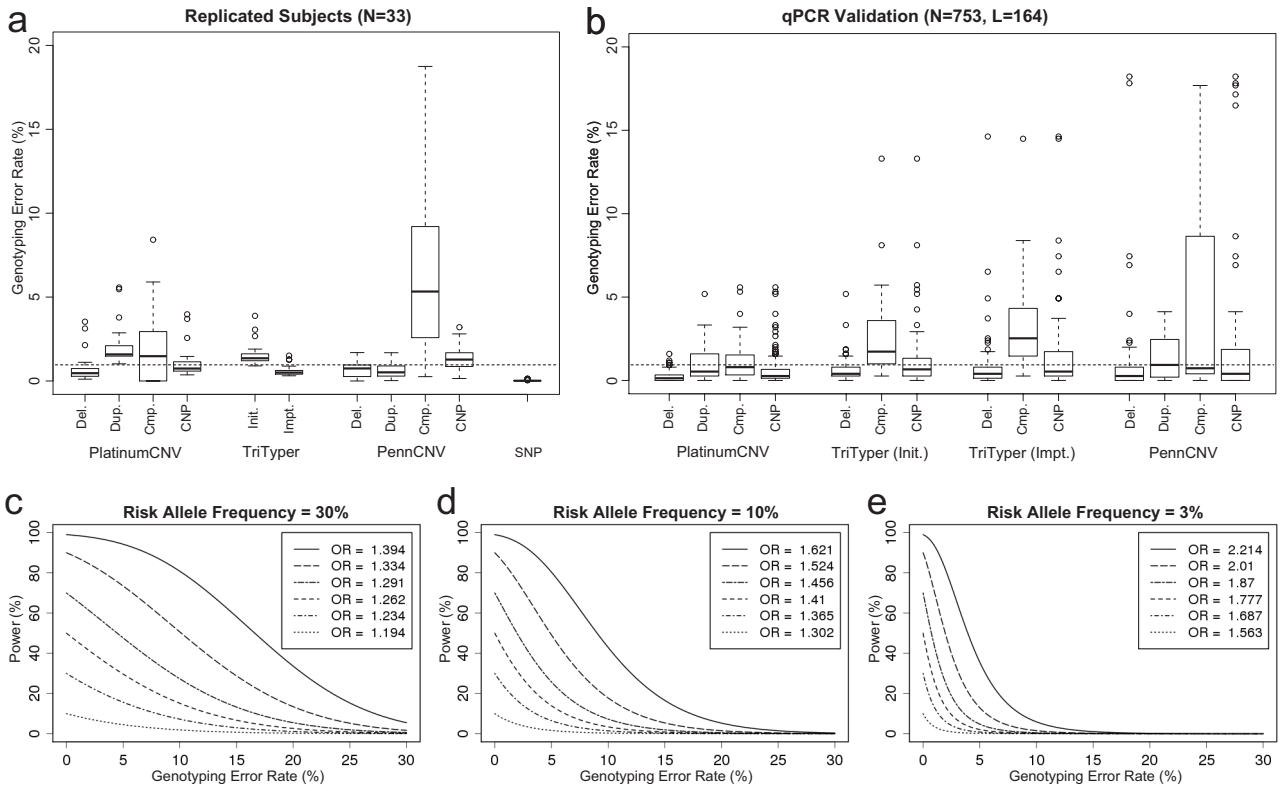

Figure 3. Estimated genotyping error rates and expected power curves. (a) Genotyping error rates estimated by using replicated subjects. These error rates were estimated for deletions (Del.), duplications (Dup.), complex CNPs (Cmp.) and total CNPs (CNP) genotyped by PlatinumCNV in conjunction with the initial genotype (Init.) and imputed genotype (Impt.) genotyped by TriTyper and deletions (Del.), duplications (Dup.), complex CNPs (Cmp.) and total CNPs (CNP) genotyped by PennCNV along with SNP genotypes (SNP). Note that the error rates more than 20% were not shown. (b) Genotyping error rates estimated by comparing with qPCR result. These error rates were estimated for deletions (Del.), duplications (Dup.), complex CNPs (Cmp.) and total CNPs (CNP) genotyped by PlatinumCNV in conjunction with the initial genotype (Init.) and imputed genotype (Impt.) genotyped by TriTyper and deletions (Del.), duplications (Dup.), complex CNPs (Cmp.) and total CNPs (CNP) genotyped by PennCNV. Note that the error rates more than 20% were not shown. (c) Expected power curves against the genotyping error rate under MAF=30%. (d) Expected power curves against the genotyping error rate under MAF=10%. The initial powers at the error rate of 0% are the same as the previous figure. (e) Expected power curves against the genotyping error rate under MAF=3%. The initial powers at the error rate of 0% are the same as the previous figures.

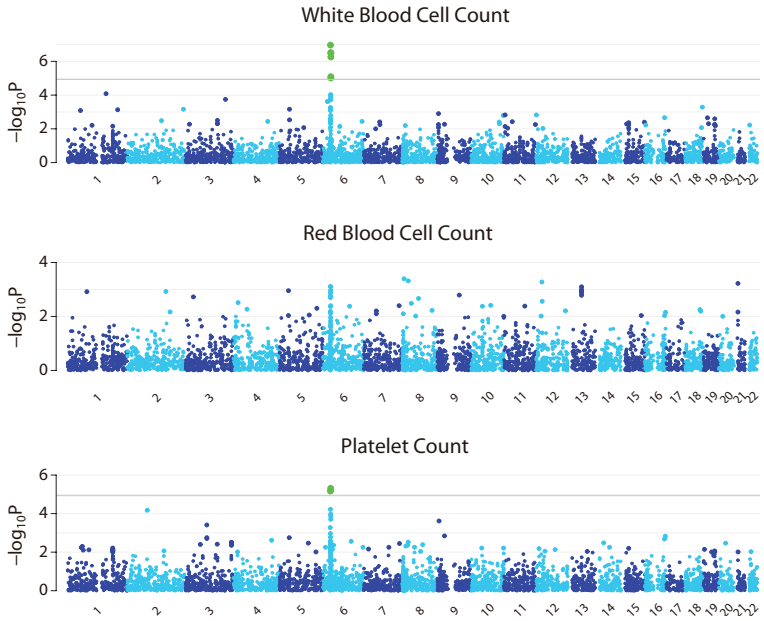

Figure 4. Manhattan plots of three hematological traits. Manhattan plots showing the $-\log_{10}(P\text{-value})$ of the 4,256 identified CNPs in the GWAS for white blood cell, red blood cell and platelet counds. The genetic loci that satisfied the significance level of $P < 1.0 \times 10^{-5}$ for the CNPs are colored in green.

Tables

Table 1. Summary of the identified CNPs

	Total (DGV%*)	≥ 2 adjacent markers (DGV%)	≥ 1 primer Polymorphisms (DGV%)	BLAST (RefSeq; Build 36)	
				Interchromosomal (DGV%)	Unmapped (DGV%)
PlatinumCNV					
Total CNPs	4,256 (62.08)	1,005 (95.52)	1,008 (54.76)	104 (67.31)	55 (92.73)
Deletions	3,241 (61.28)	858 (96.27)	678 (54.42)	50 (68.00)	46 (91.30)
Total Duplications†	981 (63.81)	140 (90.71)	319 (54.23)	53 (66.04)	8 (100.0)
AA Duplications	369 (55.01)	33 (93.94)	148 (50.68)	10 (80.00)	6 (100.0)
AB Duplications	311 (87.78)	78 (97.44)	52 (90.38)	39 (61.54)	1 (100.0)
BB Duplications	326 (51.84)	35 (74.29)	124 (44.35)	7 (57.14)	1 (100.0)
Complex CNPs	34 (88.24)	7 (100.0)	11 (90.91)	1 (100.0)	1 (100.0)
TriTyper					
Deletions (initial)	4047 (56.76)	1974 (72.95)	900 (53.33)	50 (62)	31 (83.87)
Deletions (imputed)	2863 (58.4)	1424 (76.33)	613 (52.53)	41 (63.41)	24 (83.33)
PennCNV					
Total CNPs	3745 (99.52)	3733 (99.52)	721 (99.86)	195 (100)	95 (100)
Deletions	946 (98.52)	942 (98.51)	153 (100)	57 (100)	32 (100)
Duplications	2234 (99.87)	2230 (99.87)	460 (99.78)	56 (100)	23 (100)
Complex CNPs	388 (100)	387 (100)	84 (100)	62 (100)	28 (100)
Total SNPs	600,470 (33.9)	- (-)	79,159 (35.4)	8,683 (59.92)	4,111 (56.09)

*The percentage of CNPs detected within the CNV regions reported in Database of Genomic Variants (Build 36; Nov. 2010).

† The subclassifications for the total duplications overlapped each other but were almost entirely mutually exclusive.

Table 2. Functional impact of identified CNPs

	Total gene overlap (%*)	Within gene			Intergenic (%)
		Exon (%)	UTR (%)	Intron (%)	
PlatinumCNV					
Total CNPs	1,646 (38.67)	51 (1.2)	61 (1.43)	1,534 (36.04)	2,610 (61.33)
Deletions	1,158 (35.73)	24 (0.74)	38 (1.17)	1,096 (33.82)	2,083 (64.27)
Total Duplications	476 (48.52)	27 (2.75)	22 (2.24)	427 (43.53)	505 (51.48)
AA Duplications	167 (45.26)	10 (2.71)	6 (1.63)	151 (40.92)	202 (54.74)
AB Duplications	177 (56.91)	4 (1.29)	7 (2.25)	166 (53.38)	134 (43.09)
BB Duplications	143 (43.87)	13 (3.99)	10 (3.07)	120 (36.81)	183 (56.13)
Complex CNPs	12 (35.29)	0 (0.00)	1 (2.94)	11 (32.35)	22 (64.71)
TriTyper					
Deletions (initial)	1482 (36.62)	40 (0.99)	47 (1.16)	1395 (34.47)	2565 (63.38)
Deletions (imputed)	1069 (37.34)	30 (1.05)	25 (0.87)	1014 (35.42)	1794 (62.66)
PennCNV					
Total CNPs	1721 (45.95)	51 (1.36)	44 (1.17)	1626 (43.42)	2024 (54.05)
Deletions	270 (28.54)	21 (2.22)	21 (2.22)	228 (24.1)	676 (71.46)
Duplications	1290 (57.74)	24 (1.07)	15 (0.67)	1251 (56)	944 (42.26)
Complex CNPs	124 (31.96)	4 (1.03)	7 (1.8)	113 (29.12)	264 (68.04)
Total SNPs	268,645 (44.74)	11,537 (1.92)	9,160 (1.53)	247,948 (41.29)	331,825 (55.26)

* The percentage within each (row) category.

Table 3. Number of CNPs tagged by nearby SNPs ($r^2 \geq 0.8$)*

	PlatinumCNV(%**)	TriTyper(%)		PennCNV(%)
		initial	imputed	
Total CNPs	1,196 (28.1)	–	–	102 (2.7)
Deletions	1,020 (31.5)	1,087 (23.3)	1,227 (38.3)	83 (8.8)
Common†	845 (55.9)	891 (42.4)	924 (66.8)	49 (39.8)
Low-frequency ††	175 (10.1)	187 (7.4)	303 (16.6)	34 (4.1)
Duplications	174 (17.7)	–	–	0 (0.0)
Common	159 (32.6)	–	–	0 (0.0)
Low-frequency	15 (3.0)	–	–	0 (0.0)
Complex CNPs	2 (5.9)	–	–	19 (4.9)
Common	2 (33.3)	–	–	0 (–)
Low-frequency	0 (0)	–	–	0 (0.0)

* Maximum of (Pearson's) correlation coefficient between each CN dosage and a nearby SNP were calculated.

** Percentage of loci detected within the CNP subtypes.

† Corresponding haplotype frequency is $\geq$ 5.0%.

††Corresponding haplotype frequency is 1.0-5.0%.

Table 4. Suggestive associations with white blood cell (WBC) count and platelet (PLT) counts

SNP	Chr	Posi	Gene	CN Frequency			Beta	*P*	Proxy	$R^{2\dagger}$	Primer
		(Build 36)		0	1	2	(s.e.)		SNP		polymorphisms
WBC											
rs3130981	6	31,191,792	*CDSN-PSORS1C1*	0.0	0.80	0.20	0.088 (0.016)	1.1×10^{-7}	rs9263607	0.77	rs9278982 (3base indel)
rs3130981R*	–	–	–	0.34	0.66	0.0	0.086 (0.014)	1.8×10^{-10}	rs3094212	0.87	–
rs9269287	6	32,647,375	*HLA-DRB1*	0.75	0.25	0.0	0.073 (0.014)	3.2×10^{-7}	rs9271100	0.99	
PLT											
rs130065	6	31,230,479	*CCHCR1*	0.12	0.88	0.0	-0.084 (0.019)	6.8×10^{-6}	rs35718543	0.98	rs130075
rs204994	6	32,262,976	*PBX2*	0.11	0.89	0.0	-0.091 (0.020)	4.9×10^{-6}	rs34012154	0.92	

* ASCN genotypes were reclustered so that the primer deletion at rs9278982 is captured (see main text for details).

† Pearson's correlation coefficient between the mean CN dosage of CNP and the genotype of proxy SNP.

Supplementary Figures and Tables

PlatinumCNV: a Bayesian Gaussian Mixture Model for Genotyping Copy Number Polymorphisms Using SNP Array Signal Intensity Data

Natsuhiko Kumasaka[1,*], Hironori Fujisawa[2], Naoya Hosono[3], Yukinori Okada[1], Atsushi Takahashi[1], Yusuke Nakamura[4], Michiaki Kubo[3], Naoyuki Kamatani[1]

1 Research Group for Medical Informatics, Center for Genomic Medicine, RIKEN, Tokyo, Tokyo, JAPAN

2 Department of Mathematical Analysis and Statistical Inference, The Institute of Statistical Mathematics, Tachikawa, Tokyo, JAPAN

3 Research Group for Genotyping, Center for Genomic Medicine, RIKEN, Tsurumi, Kanagawa, JAPAN

4 Human Genome Center, Institute of Medical Science, The University of Tokyo, Tokyo, JAPAN

[*]**Correspondence to: Natsuhiko Kumasaka**

Address: RIKEN, 4-6-1, Shirokane-dai, Minato-ku, Tokyo 108-8639, JAPAN

E-mail: kumasaka@src.riken.jp

Phone: +81-3-5449-5708

Fax: +81-3-5449-5564

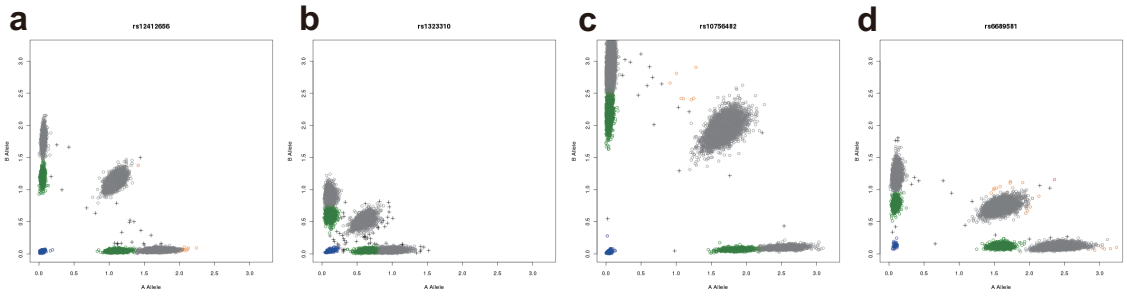

Figure 1. Fluorescent signal intensity plots in different scales. (a) Typical signal intensity plot of deletions where ASCN genotypes AA and BB are located around X and Y coordinates of 2.0, respectively. (b) Signal intensities are 2 times weaker than those in the typical plot. (c) Signal intensities are 1.5 times stronger than those in the typical plot. (d) Signal intensities for the X coordinate is two times stronger than those for the Y coordinate.

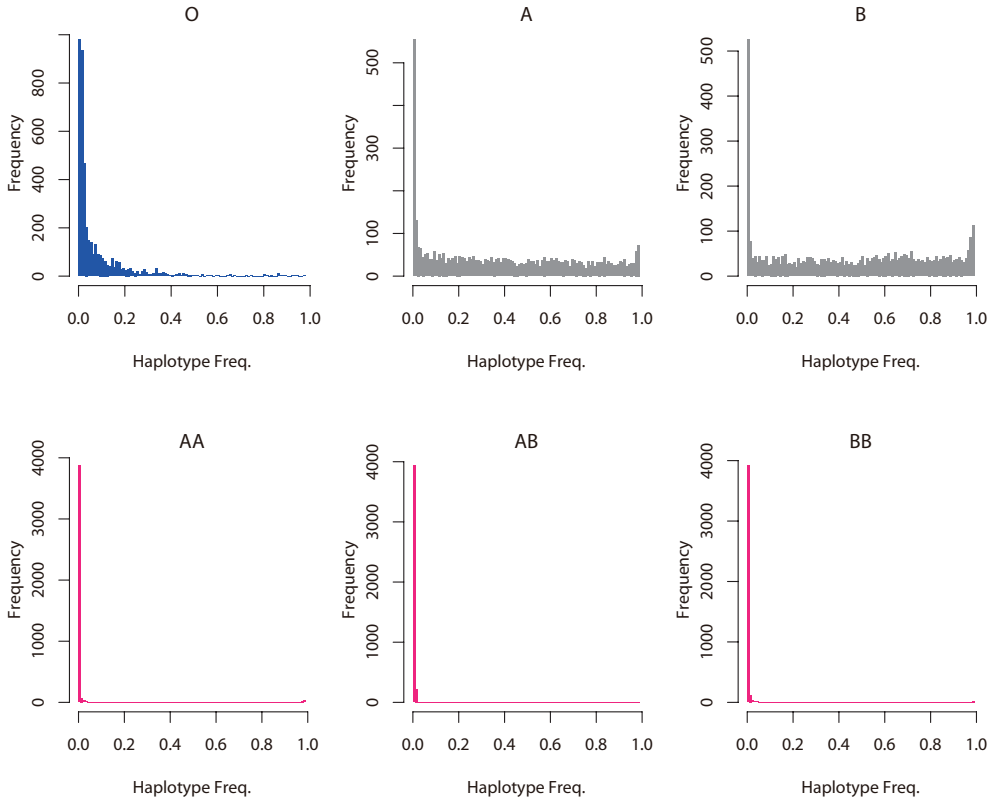

Figure 2. Identified allele-specific CNPs. Haplotype frequency spectrum for each ASCN haplotype among the identified CNPs.

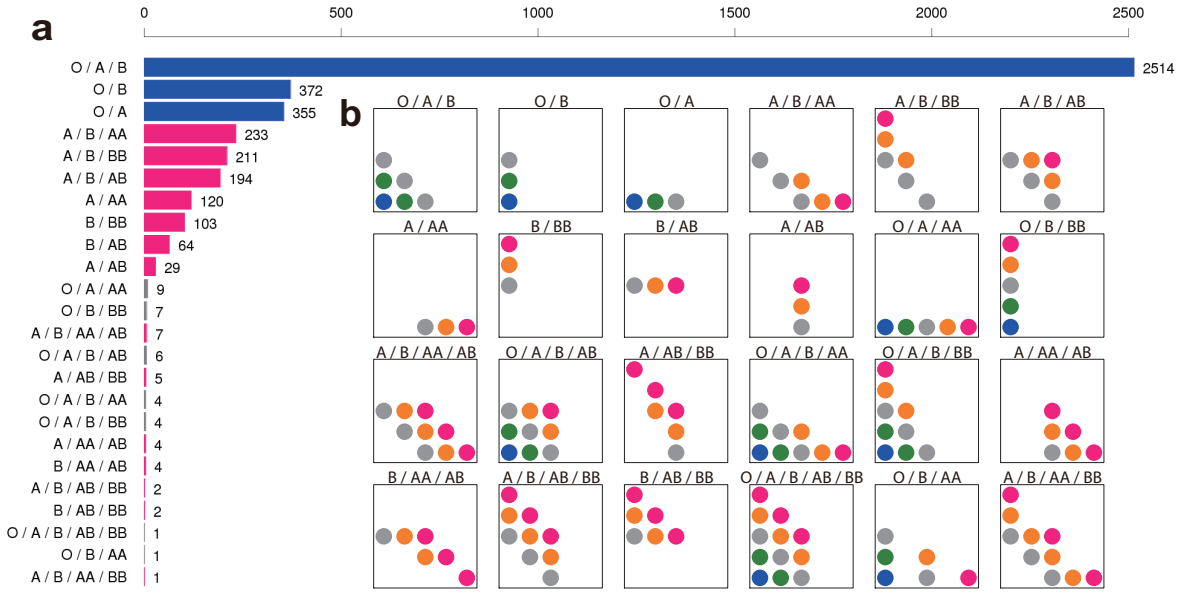

Figure 3. Identified allele-specific CNPs. (a) The number of different CNP genotype patterns with ASCN haplotypes whose frequencies were larger than 1% in the population. Bar color summarizes the major classification of CNPs, such as deletions (blue), duplications (pink) and complex CNPs (gray). (b) Stylized picture of observed ASCN genotype patterns. The color and location of each circle is compatible with Figure 2b shown in the main text.

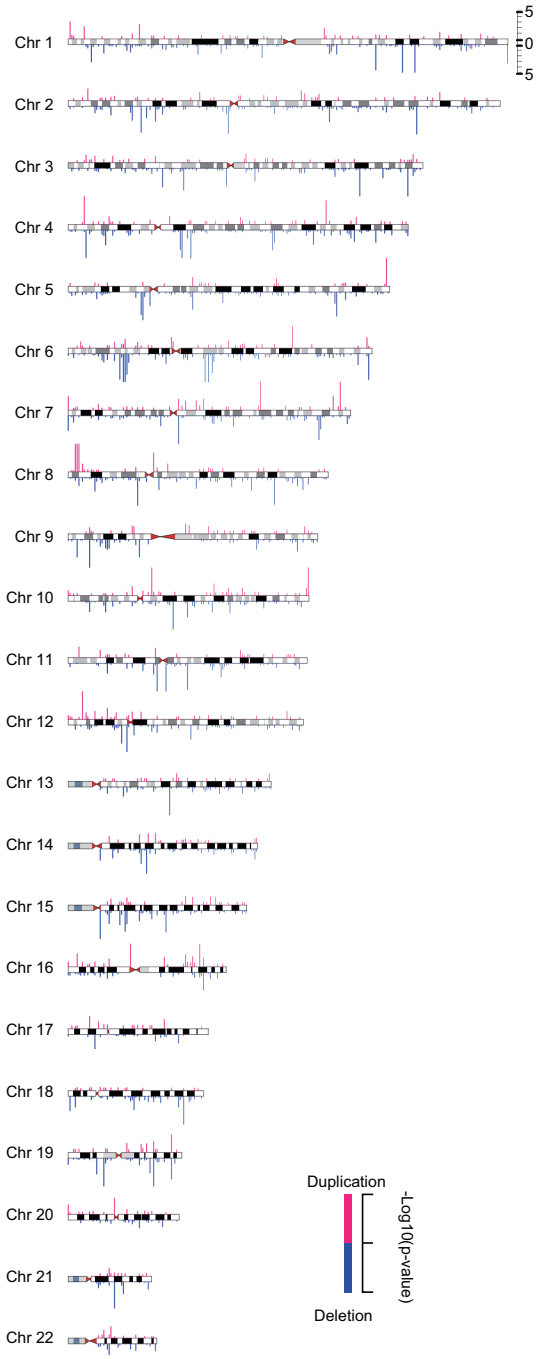

Figure 4. Enrichment of identified CNPs on 22 autosomes against SNP markers. The number of the identified CNPs in each 1 Mbp bin is compared with that of SNPs loaded in the Illumina 610K, and the CNP/SNP ratio for each bin is calculated throughout the 22 autosomes. The Fisher's exact test of the CNP/SNP ratio against the total number of identified CNPs out of the total SNPs is performed for each bin, and $\log_{10}$ p-values (with odds-ratio > 1.0) are shown. Here we divided CNPs into two subtypes: the deletions (blue bars) and the duplications (pink bars), and the complex CNPs were included in both two subtypes.

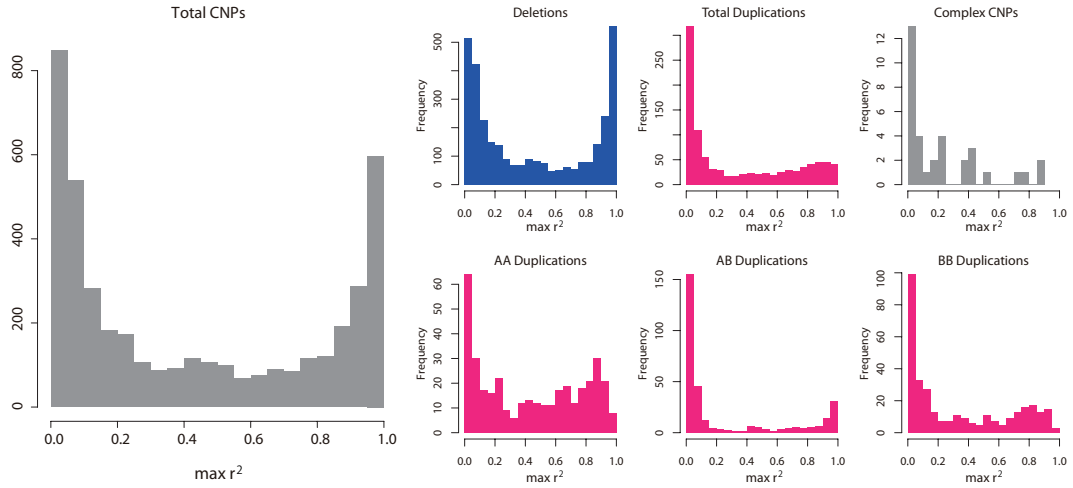

Figure 5. Histogram of maximum (Pearson's) correlation r^2 between each CNV and a nearby SNP.

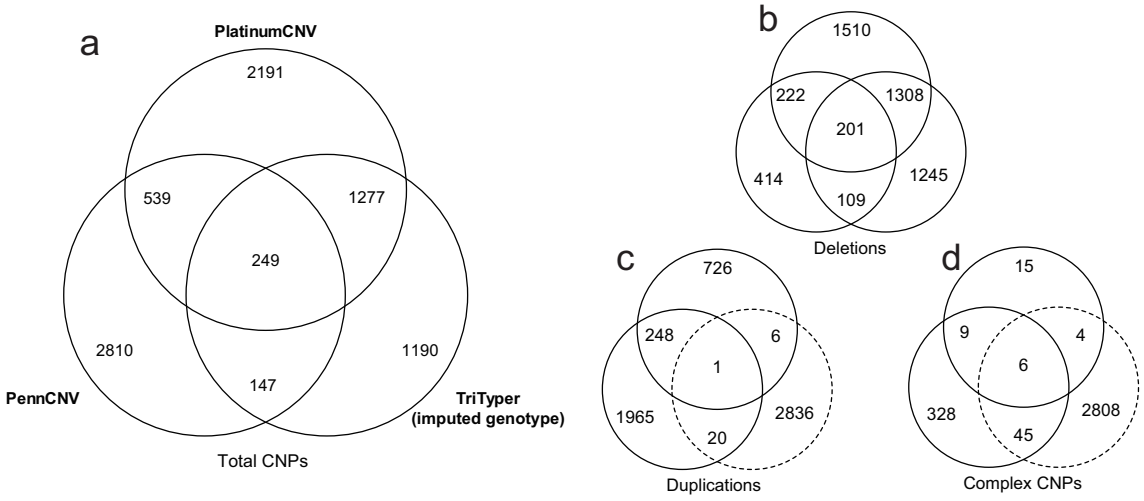

Figure 6. Venn diagrams showing the numbers of CNPs identified by PlatinumCNV, PennCNV and TriTyper (imputed genotype). (a) Total CNPs discovered by the three methods. (b) Deletions discovered by the three methods. (c) Total duplications discovered by PennCNV and PlatinumCNV. Deletions discovered by TriTyper (imputed genotypes) were also shown by the circle with dashed line suggesting that those 27 overlapping loci may be false positive loci for either TriTyper or PlatinumCNV and PennCNV. (d) Complex CNPs discovered by PennCNV and PlatinumCNV. Deletions discovered by TriTyper (imputed genotypes) were also shown by the circle with dashed line.

GWAS with Three Hematological Traits

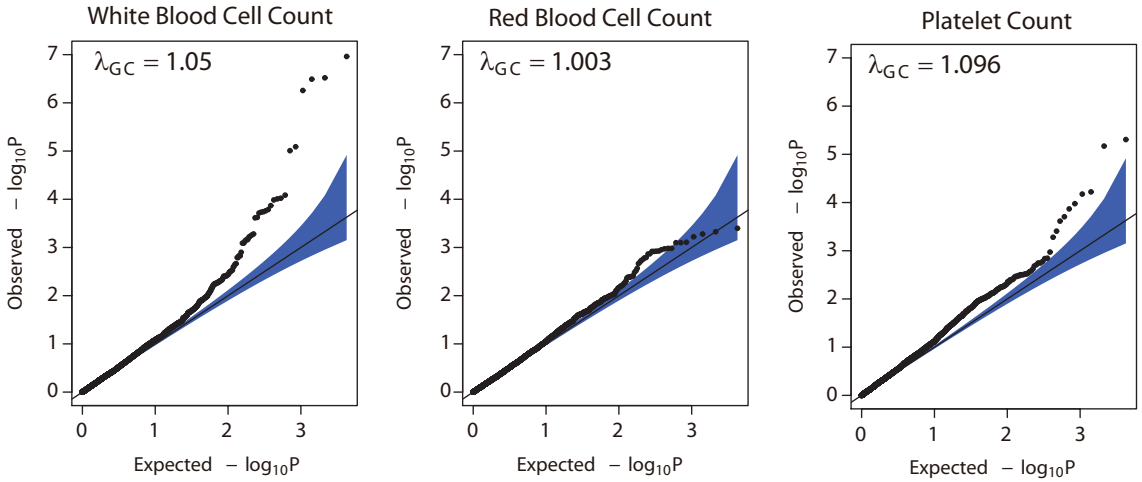

Figure 7. Quantile-Quantile plots of P-values for three hematological traits.

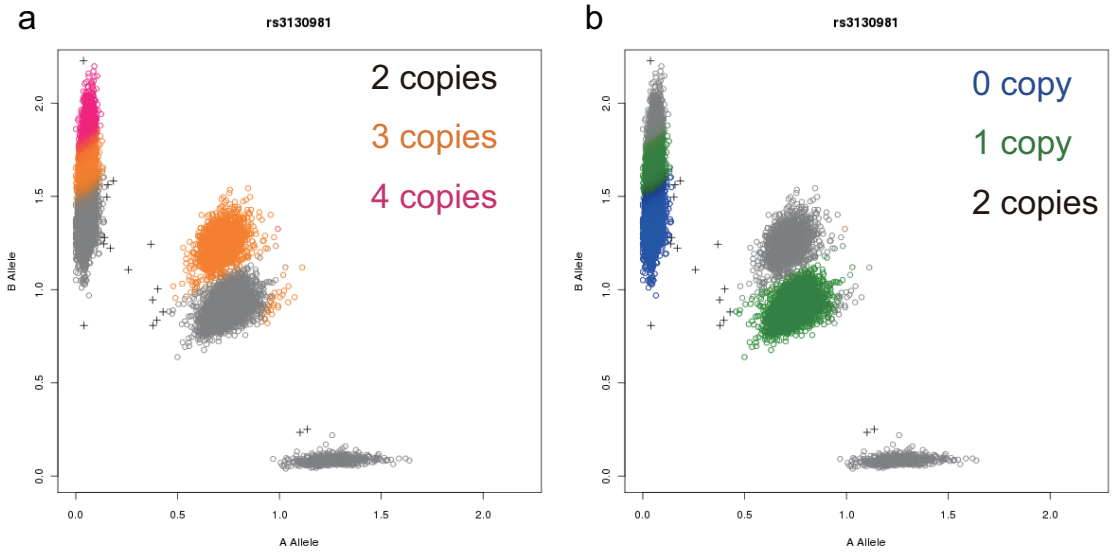

Figure 8. Reclustering to capture primer polymorphism. (a) Signal intensity plot before reclustering. The plot shows a clear duplication pattern. (a) Signal intensity plot after reclustering. The cloud of ASCN genotype BB is regarded as that of genotype O so as to capture a deletion polymorphism at rs9278982 (see main text for details).

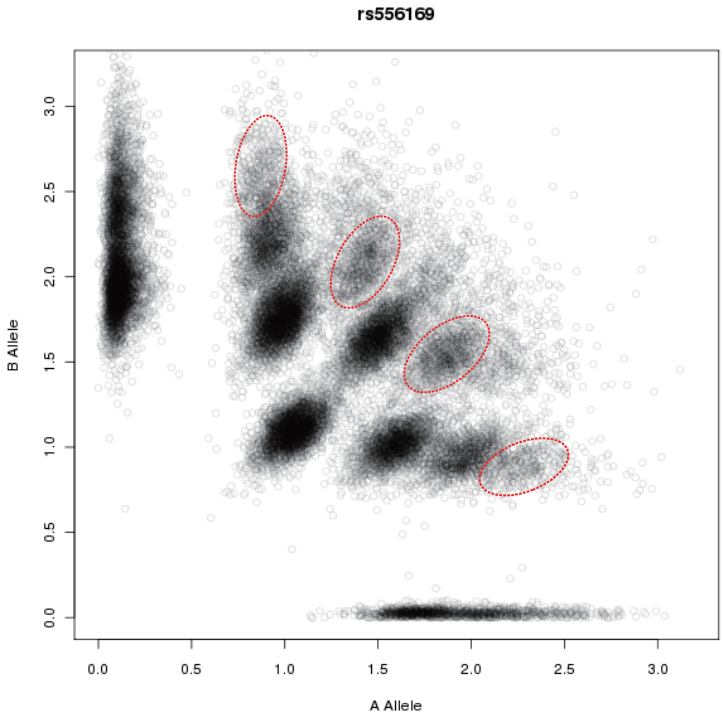

Figure 9. Higher order duplications at rs556169. Genotype clouds of 5 copies are shown by red circles. Those clouds are not capable of capturing by current PlatinumCNV.

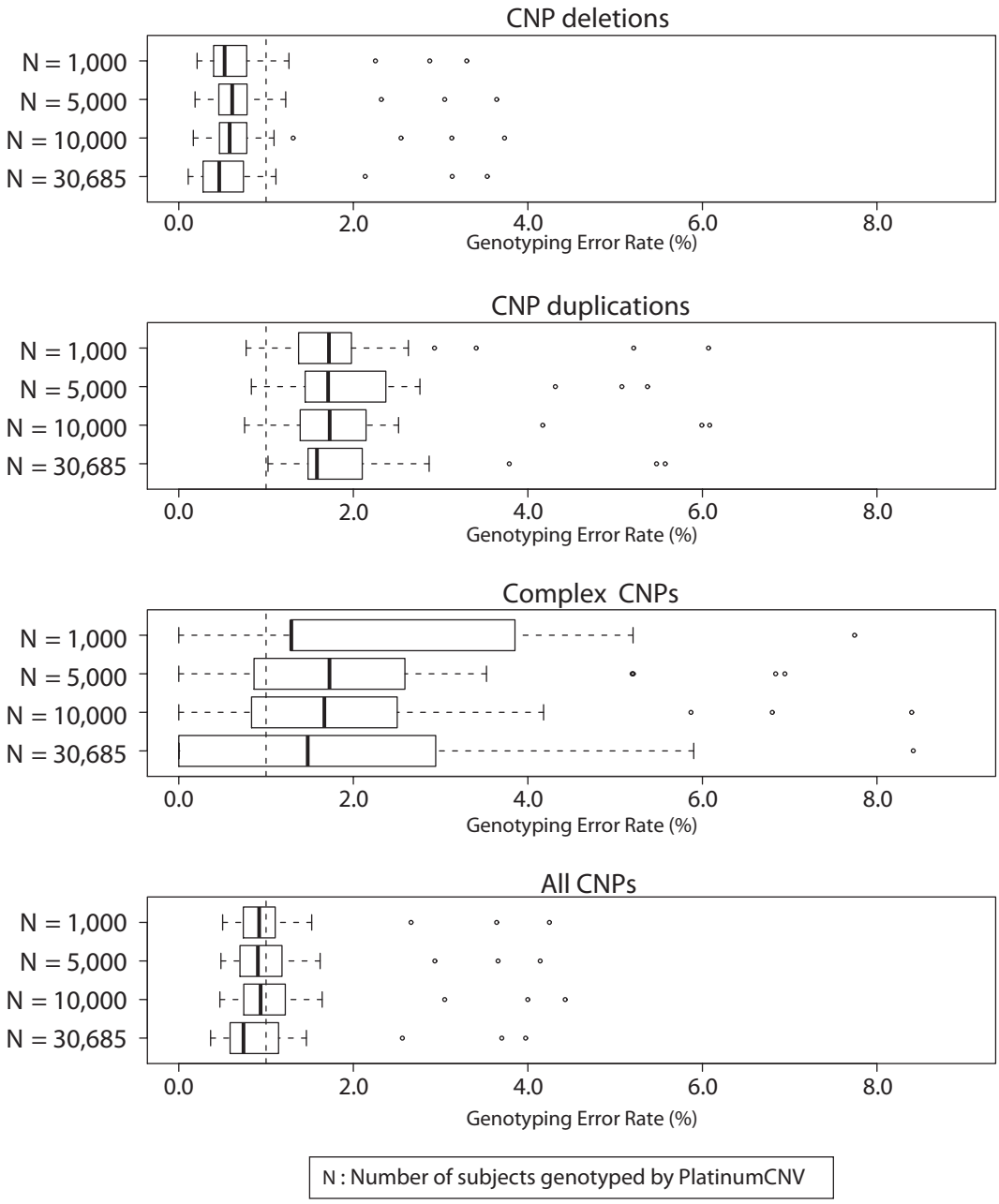

Figure 10. Estimated error rates for replicated samples with different sample size (N =1,000, 5,000 and 10,000).

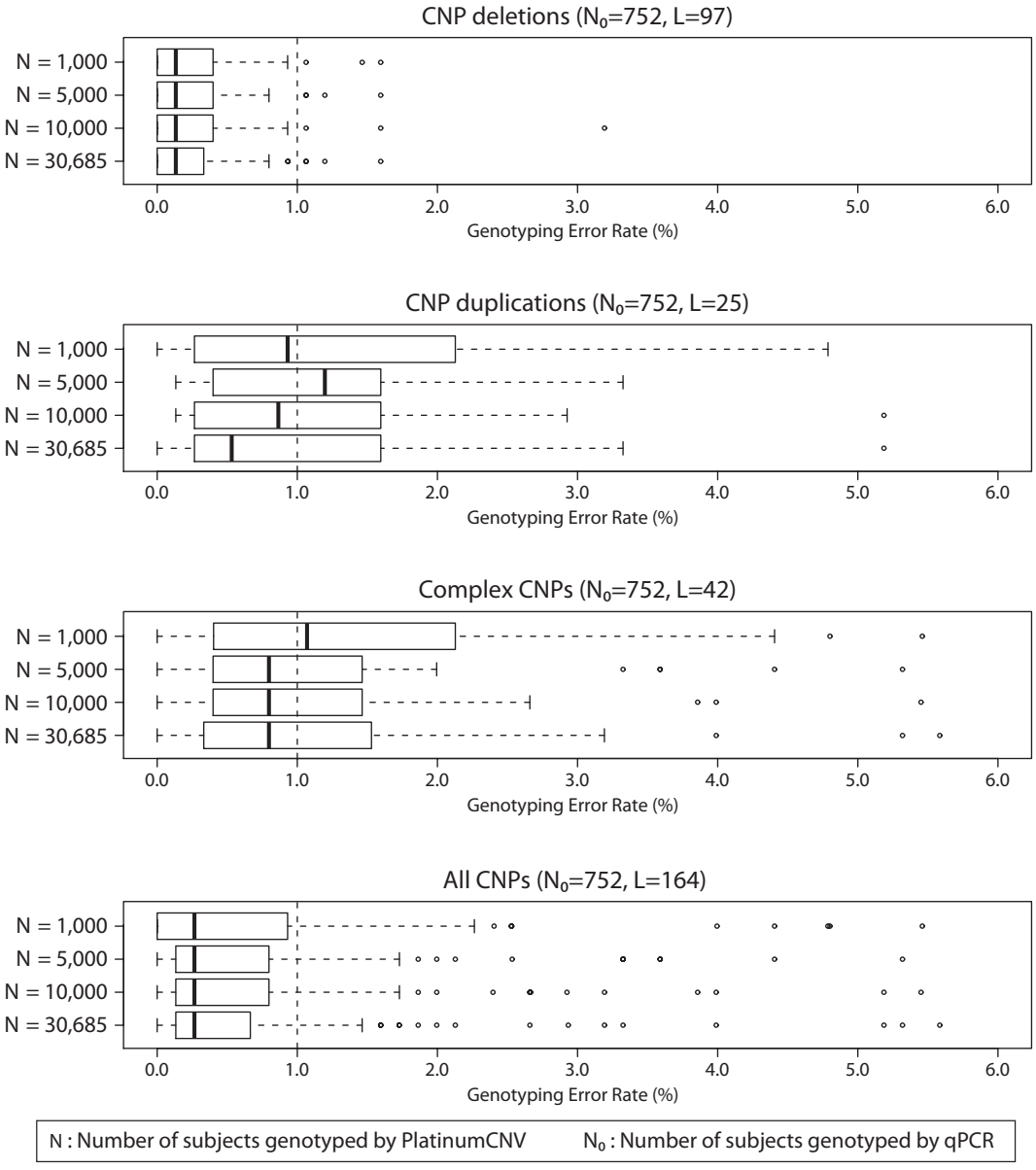

Figure 11. Estimated error rates using qPCR validation data with different sample size (N =1,000, 5,000 and 10,000).

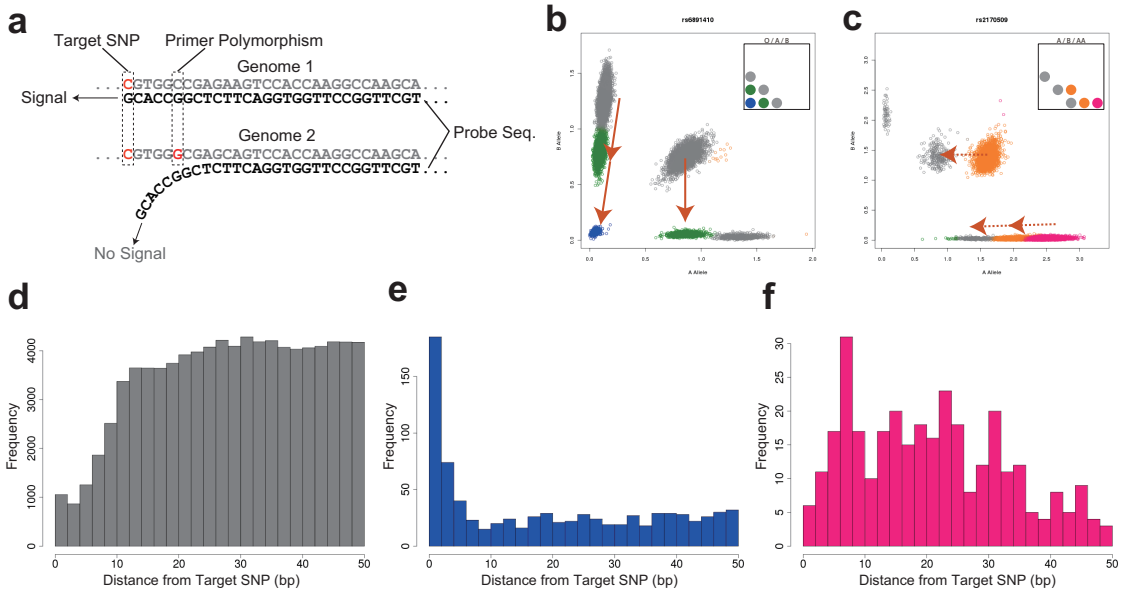

Figure 12. Impact of primer polymorphisms. (a) Stylized picture of a primer polymorphism existing in a genome where the oligonucleotide probe is annealed. In this example, both genome 1 and 2 have a allele (base) C at the target SNP locus. However genome 2 results in no fluorescent signal because of the unexpected variation at 5 bp away from the target. (b) Signal intensity plot of rs6891410 shows the typical deletion pattern but the signal intensity is, in fact, diminished by the known primer polymorphism (rs11953167) 2 bp away. The minor allele frequency of rs11953167 (17%) observed from the HapMap data (JPT) was almost the same as the estimated frequency (17.04%) of the ASCN haplotype O. (c) Signal intensity plot of rs2170509 shows the typical duplication patter with AA haplotype, but the signal intensity is, in fact, half diminished by the known primer polymorphism (rs3823071) 7 bp away. The minor allele frequency (11%) of rs3823071 observed from the HapMap data (JPT) was similar to the estimated frequency (9.81%) of the ASCN haplotype A. (d) Distribution of distances between target SNPs and primer polymorphisms for total SNPs in Illumina 610K that have more than one primer polymorphisms within the 50bp probe region. (e) Distribution of these distances conditional on CNP loci with deletions. The primer polymorphisms were concentrated from 0 to 10 bp from the target. (f) Distribution of these distances conditional on CNP loci with AA or BB duplications. The primer polymorphisms were predominantly from 5 to 30 bp from the target.

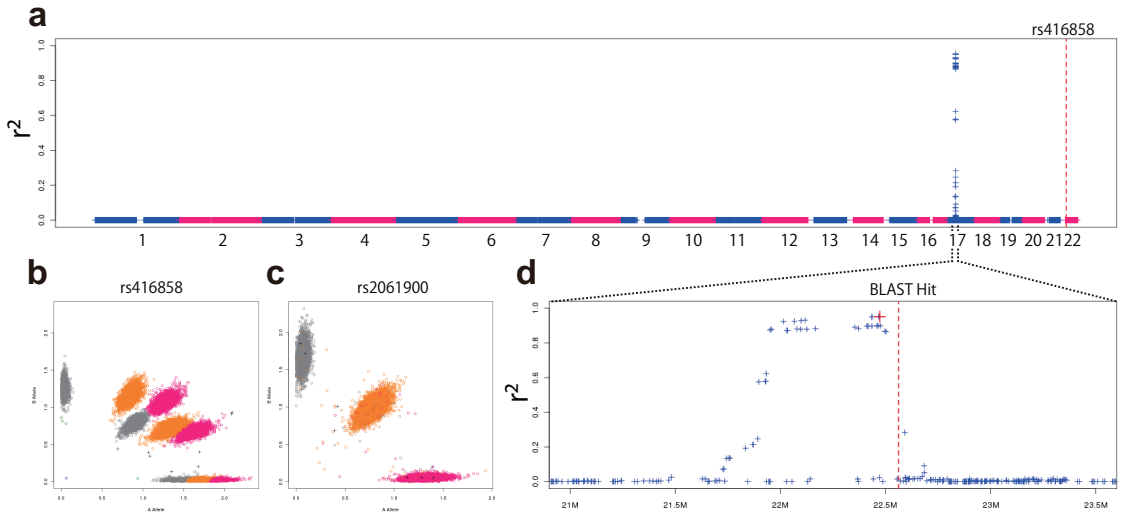

Figure 13. Instance of dispersed segmental duplication. (a) Plot of r^2 values between the mean CN dosage at rs416858 and genome-wide SNPs in Illumina 610K. (b) Fluorescent signal intensity plot at rs416858 shows a typical duplication pattern with ASCN haplotype AA and AB. (c) Fluorescent signal intensity plot at rs2061900 colored according to the number of copies observed at rs416858. (d) Plot of r^2 values around the peak. The red cross shows the value at rs2061900. The dashed red line shows the location (22,561,133 bp) at which the top genomic sequence of rs416858 (chromosome 22; 16,872,749) mapped. Note that a lot of CNVs are reported around the region (see DGV; Build 36, Nov. 2010).

Table 1. Summary of genotyping accuracy.

Variation Category†	Replication subjects (N=33)			qPCR validation (N=752)		
	No. Loci assessed*	Mean error (%)	Median error (%)	No. Loci assessed**	Mean error (%)	Median error (%)
PlatinumCNV						
Deletion	3,241	0.71	0.46	95	0.25	0.13
Duplication	981	2.01	1.58	21	1.23	0.53
Complex	34	2.00	1.48	35	1.27	0.80
Total CNP	4,256	1.02	0.74	151	0.62	0.27
TriTyper						
Initial deletions						
Deletion	–	–	–	74	0.66	0.40
Complex	–	–	–	20	2.96	1.73
Total CNP	4,047	1.55	1.35	94	1.15	0.66
Imputed deletions						
Deletion	–	–	–	61	1.00	0.4
Complex	–	–	–	15	3.64	2.53
Total CNP	2,863	0.57	0.50	76	1.52	0.53
PennCNV						
All loci						
Deletion	946	0.68	0.75	68	2.74	0.27
Duplication	2,234	0.62	0.51	23	8.60	0.93
Complex	388	6.56	5.33	34	7.4	0.73
Total CNP	3,745	1.35	1.28	125	5.08	0.40
Conditional on short CNVs (<100 markers)						
Deletion	779	0.82	0.91	–	–	–
Duplication	704	2.08	1.92	–	–	–
Complex	373	6.83	5.55	–	–	–
Total CNP	2,033	2.51	2.37	–	–	–
SNP	479,654	0.0258	0.0127	–	–	–

* We selected loci that met specific threshold of Call Rate > 0.99, normal haplotype frequency < 0.99 and $P_{\mathrm{HWE}} > 1.0 \times 10^{-6}$.

**We selected loci that met specific threshold of Call Rate > 0.99, normal haplotype frequency < 0.995 and $P_{\mathrm{HWE}} > 1.0 \times 10^{-6}$ out of 164 loci genotyped by qPCR assays.

† Categories are determined by observed frequencies of deletions and duplications for the error rate estimation using the replicated subjects; and observed frequencies of deletions and duplications of qPCR validation data for the qPCR validation.

Table 2. Summary of the identified CNPs (N=1,000, 5,000 and 10,000)

	Total (DGV%*)	$\geq$ 2 adjacent markers (DGV%)	$\geq$ 1 primer Polymorphisms (DGV%)	BLAST (RefSeq; Build 36)	
				Interchromosomal (DGV%)	Unmapped (DGV%)
N**=10,000**					
Total CNPs	4274 (62.35)	1080 (95.65)	1023 (53.86)	91 (60.44)	52 (86.54)
Deletion	3019 (60.55)	814 (95.82)	631 (54.52)	38 (60.53)	42 (88.1)
Total Duplications	1195 (65.69)	245 (94.69)	375 (51.47)	51 (60.78)	9 (77.78)
AA Duplications	375 (51.47)	30 (83.33)	147 (38.78)	7 (85.71)	7 (85.71)
AB Duplications	451 (91.8)	189 (98.41)	87 (95.4)	37 (54.05)	0 (–)
BB Duplications	383 (49.09)	28 (82.14)	143 (37.76)	9 (66.67)	2 (50)
Complex	60 (86.67)	21 (100)	17 (82.35)	2 (50)	1 (100)
N**=5,000**					
Total CNPs	4146 (62.42)	969 (95.46)	997 (53.06)	81 (62.96)	52 (84.62)
Deletion	2945 (60.92)	755 (95.89)	622 (54.02)	33 (57.58)	41 (85.37)
Total Duplications	1143 (65)	198 (93.43)	361 (50.42)	46 (65.22)	10 (80)
AA Duplications	375 (53.87)	34 (85.29)	147 (40.82)	11 (72.73)	10 (80)
AB Duplications	391 (91.05)	132 (98.48)	79 (91.14)	30 (56.67)	0 (–)
BB Duplications	395 (50.13)	36 (83.33)	138 (37.68)	9 (77.78)	0 (–)
Complex	58 (87.93)	16 (100)	14 (78.57)	2 (100)	1 (100)
N**=1,000**					
Total CNPs	3588 (61.98)	758 (97.63)	919 (53.32)	71 (66.2)	46 (84.78)
Deletion	2382 (60.75)	537 (97.58)	537 (53.45)	31 (54.84)	35 (88.57)
Total Duplications	1167 (63.67)	208 (97.6)	370 (52.43)	39 (74.36)	10 (70)
AA Duplications	362 (53.04)	22 (90.91)	135 (42.22)	11 (45.45)	8 (87.5)
AB Duplications	436 (92.43)	167 (100)	100 (90)	26 (76.92)	0 (–)
BB Duplications	392 (42.09)	23 (86.96)	138 (35.51)	9 (88.89)	2 (0)
Complex	39 (87.18)	13 (100)	12 (75)	1 (100)	1 (100)
Total SNPs	600,470 (33.9)	– (–)	79,159 (35.4)	8,683 (59.92)	4,111 (56.09)

*The percentage of CNPs detected within the CNV regions reported in Database of Genomic Variants (Build 36; Nov. 2010).

Table 3. Functional impact of identified CNPs (*N*=1,000, 5,000 and 10,000)

	Total gene overlap (%*)	Within gene			Intergenic (%)
		Exon (%)	UTR (%)	Intron (%)	
***N*=10,000**					
Total CNPs	1731 (40.5)	43 (1.01)	59 (1.38)	1629 (38.11)	2543 (59.5)
Deletion	1125 (37.26)	26 (0.86)	38 (1.26)	1061 (35.14)	1894 (62.74)
Total Duplications	583 (48.79)	16 (1.34)	20 (1.67)	547 (45.77)	612 (51.21)
AA Duplication	155 (41.33)	5 (1.33)	5 (1.33)	145 (38.67)	220 (58.67)
AB Duplication	258 (57.21)	2 (0.44)	6 (1.33)	250 (55.43)	193 (42.79)
BB Duplication	174 (45.43)	9 (2.35)	9 (2.35)	156 (40.73)	209 (54.57)
Complex	23 (38.33)	1 (1.67)	1 (1.67)	21 (35.0)	37 (61.67)
***N*=5,000**					
Total CNPs	1661 (40.06)	49 (1.18)	56 (1.35)	1556 (37.53)	2485 (59.94)
Deletion	1109 (37.66)	29 (0.98)	37 (1.26)	1043 (35.42)	1836 (62.34)
Total Duplications	539 (47.16)	20 (1.75)	19 (1.66)	500 (43.74)	604 (52.84)
AA Duplication	157 (41.87)	6 (1.6)	6 (1.6)	145 (38.67)	218 (58.13)
AB Duplication	225 (57.54)	5 (1.28)	7 (1.79)	213 (54.48)	166 (42.46)
BB Duplication	166 (42.03)	9 (2.28)	8 (2.03)	149 (37.72)	229 (57.97)
Complex	13 (22.41)	0 (0.0)	0 (0.0)	13 (22.41)	45 (77.59)
***N*=1,000**					
Total CNPs	1448 (40.36)	36 (1.0)	61 (1.7)	1351 (37.65)	2140 (59.64)
Deletion	886 (37.2)	23 (0.97)	39 (1.64)	824 (34.59)	1496 (62.8)
Total Duplications	546 (46.79)	13 (1.11)	21 (1.8)	512 (43.87)	621 (53.21)
AA Duplication	155 (42.82)	3 (0.83)	8 (2.21)	144 (39.78)	207 (57.18)
AB Duplication	245 (56.19)	2 (0.46)	9 (2.06)	234 (53.67)	191 (43.81)
BB Duplication	156 (39.8)	8 (2.04)	6 (1.53)	142 (36.22)	236 (60.2)
Complex	16 (41.03)	0 (0.0)	1 (2.56)	15 (38.46)	23 (58.97)
Total SNPs	268,645 (44.74)	11,537 (1.92)	9,160 (1.53)	247,948 (41.29)	331,825 (55.26)

* The percentage within each (row) category.

Table 4. Summary of genotyping accuracy for different sample size (N=1,000, 5,000 and 10,000).

SNP	Replication subjects (N=33)			qPCR validation (N=752)		
	No. Loci assessed*	Mean error (%)	Median error (%)	No. Loci assessed**	Mean error (%)	Median error (%)
N=10,000						
Deletion	3,019	0.832	0.583	84	0.266	0.133
Duplication	1,195	2.057	1.729	17	1.228	0.865
Complex	60	2.081	1.668	37	1.178	0.798
Total CNPs	4,274	1.191	0.938	138	0.824	0.266
N=5,000						
Deletion	2,945	0.836	0.612	95	0.244	0.133
Duplication	1,143	2.031	1.711	21	1.241	1.197
Complex	58	2.07	1.727	37	1.233	0.798
Total CNPs	4,146	1.183	0.907	153	0.824	0.266
N=1,000						
Deletion	2,382	0.743	0.525	96	0.242	0.133
Duplication	1,167	1.961	1.721	22	1.284	0.931
Complex	39	2.224	1.289	36	1.933	1.07
Total CNPs	3,588	1.154	0.921	154	0.824	0.266

* We selected loci that met specific threshold of Call Rate > 0.99, normal haplotype frequency < 0.99 and $P_{\mathrm{HWE}} > 1.0 \times 10^{-6}$.

**We selected loci that met specific threshold of Call Rate > 0.99, normal haplotype frequency < 0.995 and $P_{\mathrm{HWE}} > 1.0 \times 10^{-6}$ out of 164 loci genotyped by qPCR assays.

Supplementary Methods

Maximum a Posteriori (MAP) Estimation of Model Parameters

Let $\boldsymbol{x}_i = (x_i, y_i)^\top$ be the two-dimensional signal intensity vector and $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{iJ})^\top$ be the diplotype configuration for individual i ($i = 1, \ldots, N$). The complete-data likelihood under the Gaussian mixture model in (3) is given by

$$L^c(\theta|X, Z) = \prod_{i=1}^{N} \sum_{\mathcal{G}=1}^{K} p(\boldsymbol{z}_i|\boldsymbol{\pi}) p(\mathcal{G}|\boldsymbol{z}_i) \phi(\boldsymbol{x}_i|\boldsymbol{\mu}_\mathcal{G}, \Sigma_\mathcal{G}),$$

where $X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)^\top$, $Z = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N)^\top$ and $\theta^\top = (\boldsymbol{\pi}^\top, \boldsymbol{\mu}_1{}^\top, \ldots, \boldsymbol{\mu}_K{}^\top, \mathrm{vech}(\Sigma_1)^\top, \ldots, \mathrm{vech}(\Sigma_K)^\top)$. Here, we use the half-vec operator $\mathrm{vech}(\cdot)$ to avoid parameter redundancy [36]. The posterior probability is then given by

$$\begin{aligned} p(\theta|X, Z) &\propto L^c(\theta|X, Z) p(\theta) \\ &= p(\boldsymbol{\pi}|\boldsymbol{a}) p(\boldsymbol{\mu}|\boldsymbol{m}, \delta^{-1}S) \prod_{\mathcal{G}=1}^{K} p(\Sigma_\mathcal{G}|\lambda, V_\mathcal{G}) \prod_{i=1}^{N} \sum_{\mathcal{G}=1}^{K} p(\boldsymbol{z}_i|\boldsymbol{\pi}) p(\mathcal{G}|\boldsymbol{z}_i) \phi(\boldsymbol{x}_i|\boldsymbol{\mu}_\mathcal{G}, \Sigma_\mathcal{G}), \end{aligned}$$

where $p(\boldsymbol{\pi}|\boldsymbol{a})$, $p(\boldsymbol{\mu}|\boldsymbol{m}, \delta^{-1}S)$ and $p(\Sigma_\mathcal{G}|\lambda, V_\mathcal{G})$ denote the probability (density/mass) functions of the Dirichlet, log-normal and inverse Wishart distributions, respectively. We used the EM algorithm to obtain the (MAP) estimation of θ. The Q-function of the posterior probability is given by

$$\begin{aligned} \tilde{Q}(\theta|\hat{\theta}) &= \mathbb{E}_{Z|X,\hat{\theta}}[\log p(\theta|X, Z)] \\ &= Q(\theta|\hat{\theta}) + \log p(\boldsymbol{\pi}|\boldsymbol{a}) + \log p(\boldsymbol{\mu}|\boldsymbol{m}, \delta^{-1}S) + \sum_{\mathcal{G}=1}^{K} \log p(\Sigma_\mathcal{G}|\lambda, V_\mathcal{G}), \end{aligned}$$

where $Q(\theta|\hat{\theta}) = \mathbb{E}_{Z|X,\hat{\theta}}[\log L^c(\theta|X, Z)]$ is the Q-function of the complete-data log likelihood. Therefore, the E-step is given by

$$\begin{aligned} \bar{\boldsymbol{z}}_i^{(k)} &= \sum_{\boldsymbol{z}_i} \boldsymbol{z}_i p(\boldsymbol{z}_i|\boldsymbol{x}_i, \theta^{(k)}), \\ u_{i\mathcal{G}}^{(k)} &= \sum_{\boldsymbol{z}_i \sim \mathcal{G}} p(\boldsymbol{z}_i|\boldsymbol{x}_i, \theta^{(k)}), \\ p(\boldsymbol{z}_i|\boldsymbol{x}_i, \theta^{(k)}) &= \frac{p(\boldsymbol{z}_i|\boldsymbol{\pi}^{(k)}) \phi(\boldsymbol{x}_i|\boldsymbol{\mu}_\mathcal{G}^{(k)}, \Sigma_\mathcal{G}^{(k)})\big|_{\mathcal{G}\sim\boldsymbol{z}_i}}{\sum_{\boldsymbol{z}_i} p(\boldsymbol{z}_i|\boldsymbol{\pi}^{(k)}) \phi(\boldsymbol{x}_i|\boldsymbol{\mu}_\mathcal{G}^{(k)}, \Sigma_\mathcal{G}^{(k)})\big|_{\mathcal{G}\sim\boldsymbol{z}_i}}, \end{aligned}$$

and the corresponding M-step is given by

$$\boldsymbol{\pi}^{(k+1)} = \frac{1}{2N + a_0 - J}\left(\sum_{i=1}^{N} \tilde{\boldsymbol{z}}_i^{(k)} + \boldsymbol{a} - \mathbf{1}\right),$$
$$\Sigma_{\mathcal{G}}^{(k+1)} = \frac{\sum_{i=1}^{N} u_{i\mathcal{G}}^{(k)} (\boldsymbol{x}_i - \boldsymbol{\mu}_{\mathcal{G}}^{(k+1)})(\boldsymbol{x}_i - \boldsymbol{\mu}_{\mathcal{G}}^{(k+1)})^\top + \lambda V_{\mathcal{G}}}{\sum_{i=1}^{N} u_{i\mathcal{G}}^{(k)} + \lambda}; \;\; \mathcal{G} = 1, \ldots, K.$$

There is no closed form of $\boldsymbol{\mu}^{(k+1)}$, so we can use the Newton-Raphson method with the logarithmic transformation $\boldsymbol{\nu} = \log \boldsymbol{\mu}$, such that

$$\boldsymbol{\nu}^{(m+1)} = \boldsymbol{\nu}^{(m)} + \left(- \frac{\partial^2 \tilde{Q}}{\partial \boldsymbol{\nu} \partial \boldsymbol{\nu}^\top}\bigg|_{\boldsymbol{\nu}=\boldsymbol{\nu}^{(m)}} \right)^{-1} \frac{\partial \tilde{Q}}{\partial \boldsymbol{\nu}^\top}\bigg|_{\boldsymbol{\nu}=\boldsymbol{\nu}^{(m)}}$$

with

$$\frac{\partial \tilde{Q}}{\partial \boldsymbol{\nu}^\top} = \frac{\partial Q}{\partial \boldsymbol{\nu}^\top} - \delta S^{-1}(\boldsymbol{\nu} - \boldsymbol{m}) - \mathbf{1},$$
$$\frac{\partial^2 \tilde{Q}}{\partial \boldsymbol{\nu} \partial \boldsymbol{\nu}^\top} = \frac{\partial^2 Q}{\partial \boldsymbol{\nu} \partial \boldsymbol{\nu}^\top} - \delta S^{-1},$$

where

$$\frac{\partial Q}{\partial \boldsymbol{\nu}_{\mathcal{G}}^\top} = n_{\mathcal{G}} \mathrm{diag}(\boldsymbol{\mu}_{\mathcal{G}}) \Sigma_{\mathcal{G}}^{-1} (\bar{\boldsymbol{x}}_{\mathcal{G}} - \boldsymbol{\mu}_{\mathcal{G}}),$$
$$\frac{\partial^2 Q}{\partial \boldsymbol{\nu}_{\mathcal{G}} \partial \boldsymbol{\nu}_{\mathcal{G}}^\top} = -n_{\mathcal{G}} \mathrm{diag}(\boldsymbol{\mu}_{\mathcal{G}}) \Sigma_{\mathcal{G}}^{-1} \mathrm{diag}(\boldsymbol{\mu}_{\mathcal{G}}) + \mathrm{diag}\left[\frac{\partial Q}{\partial \boldsymbol{\nu}_{\mathcal{G}}}\right],$$

for $\mathcal{G}, \mathcal{G}' = 1, \ldots, K$. Here, $n_{\mathcal{G}} = \sum_{i=1}^{N} u_{i\mathcal{G}}$, $\bar{\boldsymbol{x}}_{\mathcal{G}} = \sum_{i=1}^{N} u_{i\mathcal{G}} \boldsymbol{x}_i / n_{\mathcal{G}}$, and $\mathbf{1}$ is the vector of all ones. Note that the Hessian $\partial^2 Q / \partial \boldsymbol{\nu} \partial \boldsymbol{\nu}^\top$ is a block diagonal matrix whose elements are $\partial^2 Q / \partial \boldsymbol{\nu}_{\mathcal{G}} \partial \boldsymbol{\nu}_{\mathcal{G}}^\top$, $\mathcal{G} = 1, \ldots, K$. Here, we set $\boldsymbol{\nu}^{(0)} = \log \boldsymbol{\mu}^{(k)}$ and obtained $\boldsymbol{\mu}^{(k+1)} = \exp(\boldsymbol{\nu}^{(M)})$ after the sufficient number of iterations M.

We iteratively applied the E and M steps until the posterior probability converged to the maximum. The initial parameters of the EM algorithm are given by $\boldsymbol{\mu}^{(0)} = \exp(\boldsymbol{m})$, $\Sigma_{\mathcal{G}}^{(0)} = V_{\mathcal{G}}$ and $\boldsymbol{\pi}^{(0)} = \boldsymbol{a}/a_0$, where $a_0 = \sum_{j=1}^{J} a_j$.

Empirical Bayes Estimation of Hyper-parameters with Laplace Approximation

Let $\eta = (\boldsymbol{a}^\top, \boldsymbol{m}^\top, \mathrm{vech}(S)^\top, \mathrm{vech}(V_1)^\top, \ldots, \mathrm{vech}(V_K)^\top)^\top$ be a hyperparameter vector and $\theta_t = (\boldsymbol{\pi}_t^\top, \boldsymbol{\mu}_t^\top, \mathrm{vech}(\Sigma_{t1})^\top, \ldots, \mathrm{vech}(\Sigma_{tK})^\top)^\top$ be a model parameter vector at locus t ($t = 1, \ldots, T$). The marginal log likelihood with

respect to η is then given by

$$l(\eta) = \log \prod_{t=1}^{T} \int p(X_t, \theta_t|\eta) d\theta_t.$$

Then, we apply the Laplace approximation on the marginal likelihood

$$\begin{aligned} l(\eta) &= \log \prod_{t=1}^{T} \int \exp\{\log p(X_t, \theta_t|\eta)\} d\theta_t \\ &\approx \log \prod_{t=1}^{T} |H_t|^{-1/2} p(X_t, \hat{\theta}_t|\eta) \\ &= \sum_{t=1}^{T} \log p(X_t, \hat{\theta}_t|\eta) - \frac{1}{2} \log |H_t|, \end{aligned}$$

where

$$H_t = -\frac{\partial^2}{\partial \theta_t \partial \theta_t^\top} \log p(X_t, \theta_t|\eta)\big|_{\theta_t = \hat{\theta}_t}$$

and

$$\hat{\theta}_t = \underset{\theta_t}{\operatorname{argmax}}\ p(\theta_t|X_t, \eta).$$

It is obvious that $\hat{\theta}_t$ is the posterior mode at locus t, which has already been given by the MAP estimation at the locus t. Note that Louis [37] showed that H_t can be expressed as

$$H_t = \mathbb{E}_{Z_t|X_t,\hat{\theta}_t}[I_c(\theta_t)] - \mathbb{E}_{Z_t|X_t,\hat{\theta}_t}[\boldsymbol{S}_c(\theta_t)\boldsymbol{S}_c(\theta_t)^\top],$$

where $\boldsymbol{S}_c(\theta_t) = \partial \log p(\theta_t|X_t, Z_t)/\partial\theta_t$ is the complete-data score vector and $I_c(\theta_t) = -\partial^2 \log p(\theta_t|X_t, Z_t)/(\partial\theta_t \partial\theta_t^\top)$ is the complete-data information matrix. This leads to

$$\begin{aligned} \frac{\partial \log |H_t|}{\partial \eta} &= \operatorname{vec}(H_t^{-1})^\top \frac{\partial}{\partial \eta} \operatorname{vec}(I(\hat{\theta}_t|\eta)) \\ \frac{\partial^2}{\partial \eta \partial \eta^\top} \log |H_t| &= -\frac{\partial \operatorname{vec}(I(\hat{\theta}_t|\eta))^\top}{\partial \eta^\top} (H_t^{-1} \otimes H_t^{-1}) \frac{\partial \operatorname{vec}(I(\hat{\theta}_t|\eta))}{\partial \eta}, \end{aligned}$$

where $I(\theta_t|\eta) = -\partial^2 \log p(\theta_t|\eta)/(\partial\theta_t \partial\theta_t^\top)$. Then, we have

$$\begin{aligned}
\hat{\boldsymbol{m}} &= \frac{1}{L}\sum_{t=1}^{T} \hat{\boldsymbol{\nu}}_t, \\
\hat{S} &= \frac{1}{L}\sum_{t=1}^{T} [(\hat{\boldsymbol{\nu}}_t - \hat{\boldsymbol{m}})(\hat{\boldsymbol{\nu}}_t - \hat{\boldsymbol{m}})^\top + \hat{\mathrm{Var}}(\hat{\boldsymbol{\nu}}_t)], \\
\hat{V}_{\mathcal{G}} &= \left(\frac{\lambda}{(\lambda - 3)L} \sum_{t=1}^{T} \left(\Sigma_{t\mathcal{G}}\right)^{-1} \right)^{-1},
\end{aligned}$$

as the empirical Bayes estimators, where $\hat{\boldsymbol{\nu}}_t = \log \hat{\boldsymbol{\mu}}_t$. Note that, there is no closed form of the Dirichlet parameters $\boldsymbol{a}$, but we can obtain $\hat{\boldsymbol{a}}$ using the Newton-Raphson method with the logarithmic transformation $\boldsymbol{b}^{(k)} = \log \boldsymbol{a}^{(k)}$, such that

$$\boldsymbol{b}^{(m+1)} = \boldsymbol{b}^{(m)} + \left(- \frac{\partial^2 l}{\partial \boldsymbol{b} \partial \boldsymbol{b}^\top}\bigg|_{\boldsymbol{b}=\boldsymbol{b}^{(m)}} \right)^{-1} \frac{\partial l}{\partial \boldsymbol{b}^\top}\bigg|_{\boldsymbol{b}=\boldsymbol{b}^{(m)}}$$

with

$$\begin{aligned}
\frac{\partial l}{\partial b_j} &= e^{b_j} \sum_{t=1}^{T} \left[\psi(a_0) - \psi(a_j) + \log \pi_{tj} - \frac{\hat{\mathrm{Var}}(\hat{\pi}_{tj})}{2(\hat{\pi}_{tj})^2} \right], \\
\frac{\partial^2 l}{\partial b_j \partial b_k} &= \frac{\partial l}{\partial b_j}\delta_{jk} + e^{b_j + b_k} \sum_{t=1}^{T} \left[\psi_1(a_0) - \psi_1(a_j)\delta_{jk} + \frac{\hat{\mathrm{Cov}}(\hat{\pi}_{tj}, \hat{\pi}_{tk})^2}{2(\hat{\pi}_{tj})^2(\hat{\pi}_{tk})^2} \right],
\end{aligned}$$

where $\psi(\cdot)$ and $\psi_1(\cdot)$ indicate the digamma function and the trigamma function respectively, and δ_{jk} denotes the Kronecker delta. Here, we set $\boldsymbol{b}^{(0)} = \log \boldsymbol{a}^{(0)}$ and obtain $\hat{\boldsymbol{a}} = \exp(\boldsymbol{b}^{(M)})$ after the sufficient number of iterations M.

Note that, in our study, these hyperparameters and the model parameters in the previous subsection were alternatively updated to maximize the marginal likelihood and the posterior probability for each CNP locus. However, as for the software package, it may not be realistic to apply such an alternative optimization scheme within a limited time. Therefore we provide our $\hat{\eta}$ as a default, the most likely hyperparameters for the Bayesian GMM.

Error Rate Estimation by Using Common Subjects

Let us consider the ASCN genotypes $\mathcal{G}_t^{(1)}$ and $\mathcal{G}_t^{(2)}$ in (4), independently observed twice for an individual at locus t ($t = 1, \ldots, L$). For the true underlying diplotype configuration $\boldsymbol{z}_t$ at locus t, the simplest error

model would be

$$p(\mathcal{G}_t|\boldsymbol{z}_t,\varepsilon) = \begin{cases} 1-\varepsilon & \text{if } \mathcal{G}_t \sim \boldsymbol{z}_t, \\ \varepsilon/(K-1) & \text{otherwise,} \end{cases}$$

with a constant genotyping error rate ε among all CNP loci. Then, the complete-data log likelihood for the multiple observation $G = \{(\mathcal{G}_t^{(1)}, \mathcal{G}_t^{(2)});\ t = 1\ldots, L\}$ and $Z = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_L\}$ can be written as

$$\begin{aligned} \log p(G, Z|\varepsilon) &= \log \prod_{t=1}^{L} p(\mathcal{G}_t^{(1)}, \mathcal{G}_t^{(2)}, \boldsymbol{z}_t|\varepsilon, \hat{\boldsymbol{\pi}}_t) \\ &= \sum_{t=1}^{L} \left[\log p(\mathcal{G}_t^{(1)}|\boldsymbol{z}_t, \varepsilon) + \log p(\mathcal{G}_t^{(2)}|\boldsymbol{z}_t, \varepsilon) + \log p(\boldsymbol{z}_t|\hat{\boldsymbol{\pi}}_t)\right], \end{aligned}$$

where $\hat{\boldsymbol{\pi}}_t$ denotes the MAP estimation of ASCN haplotype frequencies at locus t, and $p(\boldsymbol{z}_t|\boldsymbol{\pi}_t)$ is given in (2). Here, we simply assume that all CNP loci are in linkage equilibrium; that is, $\boldsymbol{z}_t$ and $\boldsymbol{z}_{t'}$ are independent of each other for $t \neq t'$. We use the EM algorithm to obtain the ML estimation $\hat{\varepsilon}$. Note that the Q-function is given by

$$\begin{aligned} Q(\varepsilon|\hat{\varepsilon}) &= \mathbb{E}_{Z|G,\hat{\varepsilon}}[\log p(G, Z|\varepsilon)] \\ &= \sum_{t=1}^{L} p(\boldsymbol{z}_t|\mathcal{G}_t^{(1)}, \mathcal{G}_t^{(2)}, \hat{\varepsilon}) \left[\log p(\mathcal{G}_t^{(1)}|\boldsymbol{z}_t, \varepsilon) + \log p(\mathcal{G}_t^{(2)}|\boldsymbol{z}_t, \varepsilon)\right] + \text{const}, \end{aligned}$$

with

$$p(\boldsymbol{z}_t|\mathcal{G}_t^{(1)}, \mathcal{G}_t^{(2)}, \hat{\varepsilon}) = \frac{p(\mathcal{G}_t^{(1)}|\boldsymbol{z}_t, \hat{\varepsilon}) p(\mathcal{G}_t^{(2)}|\boldsymbol{z}_t, \hat{\varepsilon}) p(\boldsymbol{z}_t|\hat{\pi}_t)}{\sum_{\boldsymbol{z}_t} p(\mathcal{G}_t^{(1)}|\boldsymbol{z}_t, \hat{\varepsilon}) p(\mathcal{G}_t^{(2)}|\boldsymbol{z}_t, \hat{\varepsilon}) p(\boldsymbol{z}_t|\hat{\pi}_t)}.$$

Statistical Power Calculation Under Genotyping Error

Here, we simply assume that a given CNP is diallelic, with either a deletion or a duplication, and association studies are performed with the MAP CN dosage in (5). Note that, under the assumption of the same genotyping error rate in cases and controls, the test statistic is not inflated in the null hypothesis even if we use the MAP CN dosage rather than the mean CN dosage. Let $X \in \{0, 1, 2\}$ be the (estimated) MAP CN dosage and $Z \in \{0, 1, 2\}$ be the (unobservable) true CN dosage. Here, we assume that dosage indicates the number of risk alleles. According to the simplest error model of X given Z, such that

$$p(X = k|Z = j) = \begin{cases} 1-\varepsilon & \text{if } k = j, \\ \varepsilon/2 & \text{if } k \neq j, \end{cases}$$

the distribution of X follows from

$$p(X) = \sum_{j=0}^{2} p(X|Z=j)p(Z=j).$$

When we assume the binomial distribution on Z with the (expected) allele frequency π, then X is beta-binomially distributed with probability parameter π^* and the variance inflation factor $1+f$ so that

$$X \sim \text{BetaBin}\left(2; \frac{\pi^*(1-f)}{f}, \frac{(1-\pi^*)(1-f)}{f}\right), \tag{7}$$

where $\pi^* = \pi + 3(1-2\pi)\varepsilon/4$ and

$$f = \frac{8(1-3\pi+3\pi^2)\varepsilon - 9(1-2\pi)^2\varepsilon^2}{[4\pi+3(1-2\pi)\varepsilon][4-4\pi-3(1-2\pi)\varepsilon]}.$$

Note that f can be considered to be the inbreeding coefficient in population genetics.

For such a diallelic CNP, the trend test statistic with no genotyping error ($\varepsilon = 0$) is already known [38]. It approximately follows a χ^2 distribution with the non-central parameter

$$\lambda \approx \frac{2N(\pi_1-\pi_0)^2\phi(1-\phi)}{\bar{\pi}(1-\bar{\pi})},$$

where N is the total sample size, ϕ is the proportion of the subjects that are cases, and π_1, π_0 and $\bar{\pi}$ are the (expected) frequencies of the risk allele in the cases, controls and whole sample, respectively. The odds ratio, $1+\beta$, enters this expression via the allele frequencies, by relating π_1 to π_0,

$$\pi_1 = \frac{(1+\beta)\pi_0}{1+\beta\pi_0}.$$

Analogously to the case of no genotyping error, the non-central parameter for the χ^2 test statistic with a genotyping error $\varepsilon > 0$ becomes

$$\lambda^* \approx \frac{2N(\pi_1^*-\pi_0^*)^2\phi(1-\phi)}{\bar{\pi}^*(1-\bar{\pi}^*)(1+f_1\phi+f_0(1-\phi))},$$

where $\{\pi_1^*, f_1\}$ and $\{\pi_0^*, f_0\}$ are the parameters of beta-binomial distributions for cases and controls derived from the definition in (7), and $\bar{\pi}^* = \pi_1^*\phi + \pi_0^*(1-\phi)$.

According to the above model, we calculated the power for different allele frequencies using the follow-

ing conditions:

- 2,000 cases and 3,000 controls (*i.e.*, $N = 5,000$ and $\phi = 0.4$).
- The set of risk allele frequencies was {30%, 10% and 3%}.
- The set of initial powers was {99%, 90%, 70%, 50%, 30% and 10%}.
- The significance level was set to $\alpha = 10^{-7}$.

For different risk allele frequencies, we assigned the same initial power at the error rate of zero (*i.e.*, $\varepsilon = 0$). Different odds ratios were assumed for each of the allele frequencies to retain the same initial power.