cases were also confirmed by allele-specific PCR of SNPs on 9p (table 2). The proportion of 9p UPD–positive components estimated both from allele-specific PCR and from AsCNAR (see the "Material and Methods" section) shows a good concordance (table 2). In some cases, 9p UPD–positive cells account for almost all the *JAK2* mutation–positive population, whereas, in others, they represent only a small subpopulation of the entire *JAK2* mutation–positive population (fig. 5). AsCNAR analysis also disclosed the additional three cases that have 9p gain (9p trisomy) (fig. 4*E*). The 9p trisomy is among the most-frequent cytogenetic abnormalities in MPDs[25] and is implicated in duplication of the mutated *JAK2* allele[6] but could not have been discriminated from UPD or "LOH with CN loss" by use of conventional techniques—for example, allele-specific PCR to measure relative allele dose. Since the proportions of the mutated *JAK2* allele coincide with two-thirds of the observed trisomy components in all three cases, the data suggest that the mutated *JAK2* allele is duplicated in the 9p trisomy cases (table 2). Of particular interest is the unexpected finding of the presence of two discrete populations carrying 9p UPD in three cases, in which the AsCN graph showed a two-phased dissociation along the 9p arm (fig. 4*F*). In the previous observations, homozygous *JAK2* mutations have been reported to be more common in PV cases (~40%) than in ET cases (<~10%). With AsCNAR analysis, the difference in the frequency of 9p UPD becomes more conspicuous; nearly all PV cases (11/11) and IMF cases (9/10) with a *JAK2* mutation had one or more UPD components or other gains of 9p material, whereas only 3 of the 11 *JAK2* mutation–positive ET cases carried a 9p UPD component or gain of 9p ($P = 1.3 \times 10^{-4}$, by Fisher's exact test).

Discussion

The robustness of the AsCNAR method lies in its capacity to measure accurately allele dosage and thereby to detect LOH even in the presence of significant normal cell components, which often occurs in primary tumor samples. In principle, an accurate LOH determination is accomplished only by demonstrating an absolute loss of one parental allele, not simply by detecting AI with conventional allele-measurement techniques. This is especially the case for contaminated samples, where it is essentially impossible to discriminate the origin of the remaining minor-allele component (i.e., differentiating normal cells and tumor cells).[1,3] Nevertheless, and paradoxically, it is these normal cells within the tumor samples that enable determination of AsCNs in AsCNAR. It computes AsCNs on the basis of the strength of heterozygous SNP calls produced from the "contaminated" normal component, which effectively works as "an internal reference," precluding the need for preparing a paired germline reference.

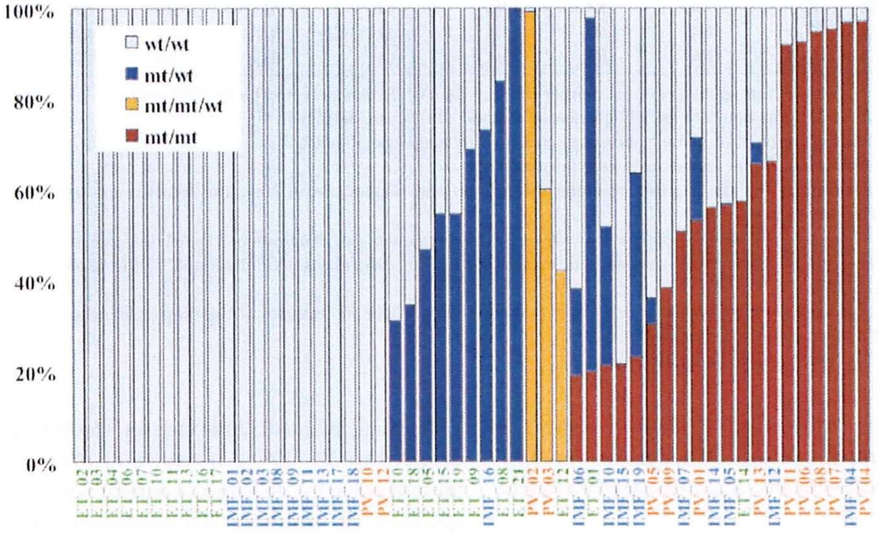

Figure 5. Estimation of tumor populations carrying 9p UPD and the *JAK2* mutation in MPD samples. The populations of 9p UPD–positive components in the 53 MPD cases were estimated by calculation of the mean difference of AsCNs within the UPD regions. Heterozygous (*blue bars*) or homozygous (*red bars*) *JAK2* mutations in MPD samples were also estimated by measurement of *JAK2* mutated alleles and UPD alleles, under the assumption that all the UPD alleles have a *JAK2* mutation. Measurement of *JAK2* mutated alleles was performed by allele-specific PCR. For three cases having trisomy components (*orange bars*), the duplicated allele was assumed to have a *JAK2* mutation, which is the consistent interpretation of the observed fraction of trisomy and mutated *JAK2* alleles for case PV_02 (table 2). mt = *JAK2* mutated allele; wt = wild-type allele.

The figure is available in its entirety in the online edition of *The American Journal of Human Genetics*.

Figure 6. Effects of the use of the different reference sets on signal-to-noise (S/N) ratios in CN analysis. The legend is available in its entirety in the online edition of *The American Journal of Human Genetics*.

It far outperforms the SNP call–based LOH-inference algorithms and other methods and definitively determines the state of LOH by sensing CN loss of one parental allele.

In the previously published algorithms, AsCN analysis was enabled by fitting observed array data to a model constructed from a fixed data set from normal samples.[18,21] However, the model that explicitly assumes integer CNs fails to cope with primary tumor samples that contain varying degrees of normal cell components (PLASQ)[18] (fig. 2). Another algorithm (CARAT) requires a large number of references to construct a model by which AsCNs are predicted, but such a model may not necessarily be properly applied to predict AsCNs for the newly processed samples, if the experimental condition for those samples is significantly different from that for the reference samples, which were used to construct the model (fig. 6 and data not shown).[21] Signal ratios between array data from very different experiments could be strongly biased, to the extent that they can no more be properly compensated by conventional regressions. In contrast, AsCNAR uses just a small number of references simultaneously processed with tumor specimens, to minimize difference in experimental conditions between tumor and references, which act as excellent controls in calculating AsCNs, although references analyzed in short intervals also work satisfactorily (data not shown).

The CN analysis software for the Illumina array provides allele frequencies, as well as CNs, by use of a model-based approach, and, as such, it enables AsCN analysis but seems to be less sensitive for detection of AIs.[26] AsCNAR can be easily adapted to other Affymetrix arrays, including 10K and 500K arrays, and may be potentially applied to Illumina arrays.

The probability of finding at least one concordant SNP between a tumor sample and a set of anonymous references is enough with five references, but use of just one reference provides almost an equivalent AsCN profile to that obtained with its paired reference (fig. 7). The sensitivity and specificity of LOH detection with this algorithm are excellent, even in the presence of significant degrees of normal cell components (~70%–80%), which circumvent the need for purifying the tumor components for analysis—for example, by time-consuming microdissection.

The figure is available in its entirety in the online edition of *The American Journal of Human Genetics*.

Figure 7. CN profile obtained with the use of a varying number of anonymous references. The legend is available in its entirety in the online edition of *The American Journal of Human Genetics*.

Because the AsCNAR algorithm is quite simple, it requires much less computing power and time (several seconds per sample on average laptop computers) than do model-based algorithms. For example, with PLASQ, it takes overnight for model construction and an additional hour for processing each sample.

The high sensitivity of LOH detection by AsCNAR has been validated not only by the analysis of tumor DNA intentionally mixed with normal DNA but also by the analysis of primary leukemia samples. It unveiled otherwise undetected, minor UPD-positive populations within leukemia samples. Especially, the extremely high frequency of 9p UPD or gains of 9p in particular types of *JAK2* mutation–positive MPDs, as well as multiple UPD-positive subclones in some cases, demonstrated how strongly and efficiently a genetic change (point mutation) works to fix the next alteration (mitotic recombination) in the tumor population during clonal evolution in human cancer. Finally, the conspicuous difference in UPD frequency among different MPD subtypes (PV and IMF vs. ET) is noteworthy. This is supported by a recent report that demonstrated the presence of minor subclones carrying exclusively the mutated *JAK2* allele in all PV samples, but in none of the ET samples, by examining a large number of erythroid burst-forming units and Epo-independent erythroid colonies for *JAK2* mutation.[27] Our observation also supports their hypothesis that the biological behavior of these prototypic stem-cell disorders with a continuous disease spectrum could be determined by the components with either homozygous or duplicated *JAK2* mutations.

In conclusion, the AsCNAR with use of high-density oligonucleotide microarrays is a robust method of genomewide analysis of allelic changes in cancer genomes and provides an invaluable clue to the understanding of the genetic basis of human cancers. The AsCNAR algorithm is freely available on our CNAG Web site for academic users.

Acknowledgments

This work was supported by Research on Measures for Intractable Diseases, Health and Labor Sciences Research Grants, Ministry of Health, Labor and Welfare, by Research on Health Sciences focusing on Drug Innovation, by the Japan Health Sciences Foundation, by Core Research for Evolutional Science and Technology, Japan Science and Technology Agency, and by Japan Leukemia Research Fund.

Appendix A

AsCNAR

Quadratic Regression

The $\log_2$ signal-ratio, $\log_2 R^{\mathrm{ref}I}_{AB,i}$ is regressed by the quadratic terms (the length $[L_i]$ and the GC content $[M_i]$ of the PCR fragment of the ith SNP) as

$$\log_2 R^{\mathrm{ref}I}_{AB,i} = \alpha L_i^2 + \beta L_i + \chi M_i^2 + \delta M_i + \gamma + \varepsilon_i \ ,$$

where ε_i is the error term and the coefficients of regressions α, β, χ, δ, and γ are dependent on the reference used and are determined to minimize the residual sum of squares (i.e., $\sum_i \varepsilon_i^2$). Note that the sum is taken for those SNPs that have concordant SNP calls between the tumor and the reference samples.

We suppose that both allele A DNA and allele B DNA follow the same PCR kinetics, and allele-specific ratios $R^{\mathrm{ref}I}_{A,i}$ and $R^{\mathrm{ref}I}_{B,i}$, respectively, can be regressed by the same parameters, as

$$\log_2 \hat{R}^{\mathrm{ref}I}_{A,i} = \log_2 R^{\mathrm{ref}I}_{A,i} - (\alpha L_i^2 + \beta L_i) - (\chi M_i^2 + \delta M_i) - \gamma$$

and

$$\log_2 \hat{R}^{\mathrm{ref}I}_{B,i} = \log_2 R^{\mathrm{ref}I}_{B,i} - (\alpha L_i^2 + \beta L_i) - (\chi M_i^2 + \delta M_i) - \gamma \ ,$$

and the corrected total CN ratio is

$$\hat{R}^{\mathrm{ref}I}_{AB,i} = \begin{cases} \hat{R}^{\mathrm{ref}I}_{A,i} & \text{for } O_i^{\mathrm{tum}} = O_i^{\mathrm{ref}I} = AA \\ \hat{R}^{\mathrm{ref}I}_{B,i} & \text{for } O_i^{\mathrm{tum}} = O_i^{\mathrm{ref}I} = BB \\ \frac{1}{2}\left(\hat{R}^{\mathrm{ref}I}_{A,i} + \hat{R}^{\mathrm{ref}I}_{B,i}\right) & \text{for } O_i^{\mathrm{tum}} = O_i^{\mathrm{ref}I} = AB \end{cases} .$$

Averaging over the References of Concordance SNPs

Concordant reference sets C_i^K and $C_i^{K,\mathrm{hetero}}$ for each SNP S_i for a given set of references, K, are defined as follows:

$$C_i^K = \{\mathrm{ref}I \mid O_i^{\mathrm{tum}} = O_i^{\mathrm{ref}I}, \mathrm{ref}I \in K\}$$

$$C_i^{K,\mathrm{hetero}} = \{\mathrm{ref}I \mid O_i^{\mathrm{tum}} = O_i^{\mathrm{ref}I} = AB, \mathrm{ref}I \in K\} \ ,$$

and the averaged CN ratio, $\tilde{R}^K_{AB,i}$, is provided by

$$\tilde{R}^K_{AB,i} = \frac{1}{\# C_i^K} \sum_{\mathrm{ref}I \in C_i^K} \hat{R}^{\mathrm{ref}I}_{AB,i} \ , \ C_i^K \neq \phi$$

where "#" denotes the number of the elements of the set. Similarly, AsCN ratios are obtained by

$$\tilde{R}^K_{A,i} = \frac{1}{\# C_i^{K,\mathrm{hetero}}} \sum_{\mathrm{ref}I \in C_i^{K,\mathrm{hetero}}} \hat{R}^{\mathrm{ref}I}_{A,i}$$

$$(C_i^{K,\mathrm{hetero}} \neq \phi) \ .$$

$$\tilde{R}^K_{B,i} = \frac{1}{\# C_i^{K,\mathrm{hetero}}} \sum_{\mathrm{ref}I \in C_i^{K,\mathrm{hetero}}} \hat{R}^{\mathrm{ref}I}_{B,i}$$

Exceptional Handling with Regions of Homozygous Deletion, High Amplification, and LOH

To prevent SNPs within the regions that show homozygous deletion or high-grade amplification from being analyzed as "homozygous SNPs," a homozygous SNP S_i in the tumor sample is redefined as a heterozygous SNP with $\hat{O}_i^{\mathrm{tum}} = AB$, if $\max(\log_2 \tilde{R}^K_{A,i}, \log_2 \tilde{R}^K_{B,i}) \leqslant 0.1$ or $\min(\log_2 \tilde{R}^K_{A,i}, \log_2 \tilde{R}^K_{B,i}) \geqslant -0.1$, where $\tilde{R}^K_{A,i}$ and $\tilde{R}^K_{B,i}$ are calculated supposing SNP S_i is heterozygous. These cutoff values (0.1 and −0.1) are determined by receiver operating characteristic (ROC) curve for detection of gain of the larger allele and loss of the smaller allele in a sample containing 20% tumor cells (data not shown). In addition, SNPs within inferred LOH regions are also analyzed as "heterozygous" SNPs.

Reference Selection

The optimized set of references is selected that minimizes the SD of total CN at the diploid region D,

$$\mathrm{SD}_K(D) = \sqrt{\frac{\sum_{i \in D, C_i^K \neq \phi} (\log_2 \tilde{R}^K_{AB,i})^2}{\#\{i \mid i \in D, C_i^K \neq \phi\} - 1}} \ .$$

To do this, instead of testing all possible 2^N combinations of N references, we calculate $\mathrm{SD}_K(D)$ for individual references $K = \{\mathrm{ref}1\}, \{\mathrm{ref}2\}, \{\mathrm{ref}3\}, \ldots, \{\mathrm{ref}N\}$, to order the references such that $\mathrm{SD}_1(D) \leqslant \ldots \leqslant \mathrm{SD}_s(D) \leqslant \mathrm{SD}_{s+1}(D) \leqslant \ldots \leqslant \mathrm{SD}_N(D)$, where $1, 2, 3, \ldots, s, s+1, \ldots, N$ denotes the ordered references. The optimal set $K(N_0) = \{1, 2, 3, \ldots, N_0\}$ is determined by choosing N_0 that satisfies $\mathrm{SD}_{K(1)}(D) \geqslant \ldots \geqslant \mathrm{SD}_{K(N_0)}(D) < \mathrm{SD}_{K(N_0+1)}(D)$.

Note that, in principle, a diploid region cannot be unequivocally determined without doing single-cell–based analysis—for example, FISH or cytogenetics. Otherwise, a diploid region is empirically determined by setting the CN-minimal regions with no AI as diploid, which provides correct estimation of the ploidy in most cases (data not shown).

The figure is available in its entirety in the online edition of *The American Journal of Human Genetics.*

Figure C1. Inference of LOH on the basis of heterozygous SNP calls. The legend is available in its entirety in the online edition of *The American Journal of Human Genetics.*

Appendix C
Inference of LOH Based on Heterozygous SNP Calls

For a given contiguous region $\Omega_{i,j}$ between the ith and jth SNPs ($i \leq j$) and for the complete set of observed SNP calls therein, $O(\Omega_{i,j})$, consider the log likelihood ratio

$$Z(\Omega_{i,j}) \equiv \ln \frac{P(O(\Omega_{i,j}) \mid \Omega_{i,j} \in \mathrm{LOH})}{P(O(\Omega_{i,j}) \mid \Omega_{i,j} \notin \mathrm{LOH})} ,$$

where the ratio is taken between the conditional probabilities that the current observation, $O(\Omega_{i,j})$, is obtained under the assumption that $O(\Omega_{i,j})$ belongs to LOH or not. We assume a constant miscall rate ($q = 0.001$) for all SNP and use the conditional probability that the kth SNP is heterozygous (h_k), depending on the observed $k-1$th SNP call, for partially taking the effect of linkage disequilibrium into account:

$$Z(\Omega_{i,j}) =$$

$$\ln \frac{\prod_{i \leq k \leq j} \{(1-q)O_k + q(1-O_k)\}}{\prod_{i \leq k \leq j} [\{(1-h_k)(1-q) + h_k q\}O_k + \{(1-h_k)q + h_k(1-q)\}(1-O_k)]}$$

where h_k is calculated using the data from the 96 normal Japanese individuals, whereas O_k takes either 1 or 0, depending on the kth SNP call, with 1 for a homozygous call and 0 for a heterozygous call. For each chromosome, a set of regions, $\Omega_{I_n,J_n}(J_{n-1} < I_n \leq J_n, J_0 = 0)$ ($n = 1,2,3,\ldots$), can be uniquely determined as follows.

Beginning with the SNP at the short arm end (S_0), find the SNP S_{I_n} that satisfies $Z(\Omega_{I_n,I_n}) > 0$ and $Z(\Omega_{i,i}) \leq 0$ for $J_{n-1} < \forall i < I_n$ (fig. C1). Identify the SNP S_{J^+}, such that $Z(\Omega_{I_n,j}) > 0$ for $I_n \leq \forall j \leq J^+$ and $Z(\Omega_{I_n,J^+ + 1}) \leq 0$, or that S_{J^+} is the end of the chromosome (fig. C1). Then, put J_n as $\arg\max_j Z(\Omega_{I_n,j})(I_n \leq j \leq J^+)$ (fig. C1). This procedure is iteratively performed, beginning the next iteration with the SNP S_{J_n+1}, until it reaches to the end of the long arm, generating a set of nonoverlapping regions, $\Omega_{I_1,J_1}, \Omega_{I_2,J_2}, \Omega_{I_3,J_3}, \ldots \Omega_{I_n,J_n}, \ldots$. LOH inference is now enabled by testing each $Z(\Omega_{I_n,J_n})$ against a threshold (25), which is arbitrarily determined from the ROC curve for LOH determination on a DNA sample from a lung cancer cell line, NCI-H2171 (fig. C1). This algorithm is implemented in our CNAG program, which is available at our Web site.

Appendix E
Algorithm for Detection of AI With or Without LOH

The regions with AI are inferred from the AsCN data by use of an HMM, where the real state of AI (a hidden state) is inferred from the observed states of difference in AsCNs of the two parental alleles, which are expressed as dichotomous values ("preset" or "absent") according to a threshold (μ). The emission probabilities at the ith SNP locus (Si) are

$$P(|\log_2 \tilde{R}^K_{A,i} - \log_2 \tilde{R}^K_{B,i}| \leq \mu | Si \in \mathrm{AI}) = \beta$$

$$P(|\log_2 \tilde{R}^K_{A,i} - \log_2 \tilde{R}^K_{B,i}| > \mu | Si \in \mathrm{AI}) = 1 - \beta$$

and

$$P(|\log_2 \tilde{R}^K_{A,i} - \log_2 \tilde{R}^K_{B,i}| > \mu | Si \in \overline{\mathrm{AI}}) = \alpha$$

$$P(|\log_2 \tilde{R}^K_{A,i} - \log_2 \tilde{R}^K_{B,i}| \leq \mu | Si \in \overline{\mathrm{AI}}) = 1 - \alpha$$

(see also the "Material and Methods" section and appendix A for calculation of $\tilde{R}^K_{A,i}$ and $\tilde{R}^K_{B,i}$).

The parameters (μ, α, and β) are determined by the results of 10%, 20%, and 30% tumor samples. Sensitivity and specificity are calculated with varying threshold (μ), where sensitivity is defined as the ratio of detected SNPs of UPD region detected in the 100% tumor sample, specificity is defined as the ratio of nondetected SNPs in normal samples, and α and β parameters are determined from mixed tumor-sample data for each threshold value. Sensitivity and specificity are relatively stable and are within the acceptable range when the threshold is between 0.05 and 0.15 in 20% and 30% tumor samples (fig. E1). We used 0.12, 0.17, and 0.06 for μ, α, and β, respectively, on the basis of 20% tumor-sample data.

Considering that UPD is caused by a process similar to recombination, the Kosambi's map function (1/2)tanh(2θ) is used for transition probability, where θ is the distance between two SNPs, expressed in cM units; for simplicity, 1 cM should be 1 Mbp. Thus, the most likely underlying, hidden, real states of AI are calculated for each SNP according to Vitervi's method, by which AI-positive regions are defined by contiguous SNPs with "present" AI calls flanked by either chromosomal end or an "absent" AI call. Next, to determine the LOH status for each AI-positive region (Γ), AsCN states at each SNP locus within Γ are

The figure is available in its entirety in the online edition of *The American Journal of Human Genetics.*

Figure E1. Sensitivity and specificity for determination of AI, LOH, and UPD. The legend is available in its entirety in the online edition of *The American Journal of Human Genetics.*

inferred as "reduced (R)" and "not reduced ($\bar{R}$)" for the smaller AsCNs, and "increased (I)" and "not increased ($\bar{I}$)" for the larger AsCNs, using similar HMMs from the "observed CN states" of the smaller and the larger AsCNs, which are expressed as dichotomous values according to thresholds μ_S and μ_L, respectively. The emission probabilities of these models are

$$P[\min(\log_2 \tilde{R}^K_{A,i}, \log_2 \tilde{R}^K_{B,i}) < \mu_S | Si \in \mathrm{R}] = 1 - \beta_S$$

$$P[\min(\log_2 \tilde{R}^K_{A,i}, \log_2 \tilde{R}^K_{B,i}) \geqslant \mu_S | Si \in \mathrm{R}] = \beta_S$$

$$P[\min(\log_2 \tilde{R}^K_{A,i}, \log_2 \tilde{R}^K_{B,i}) < \mu_S | Si \in \bar{\mathrm{R}}] = \alpha_S$$

$$P[\min(\log_2 \tilde{R}^K_{A,i}, \log_2 \tilde{R}^K_{B,i}) \geqslant \mu_S | Si \in \bar{\mathrm{R}}] = 1 - \alpha_S$$

and

$$P[\max(\log_2 \tilde{R}^K_{A,i}, \log_2 \tilde{R}^K_{B,i}) > \mu_L | Si \in \mathrm{I}] = 1 - \beta_L$$

$$P[\max(\log_2 \tilde{R}^K_{A,i}, \log_2 \tilde{R}^K_{B,i}) \leqslant \mu_L | Si \in \mathrm{I}] = \beta_L$$

$$P[\max(\log_2 \tilde{R}^K_{A,i}, \log_2 \tilde{R}^K_{B,i}) > \mu_L | Si \in \bar{\mathrm{I}}] = \alpha_L$$

$$P[\max(\log_2 \tilde{R}^K_{A,i}, \log_2 \tilde{R}^K_{B,i}) \leqslant \mu_L | Si \in \bar{\mathrm{I}}] = 1 - \alpha_L \ .$$

These parameters (μ_S, α_S, β_S, μ_L, α_L, and β_L) are determined by evaluating sensitivities and specificities of the results for 10%, 20%, and 30% tumor samples, where sensitivities and specificities are calculated the same way as was AI. Sensitivity and specificity are relatively stable for μ_S between -0.03 and -0.13 and are relatively stable for μ_L between 0.04 and 0.09 in 20% and 30% tumor samples (fig. E1). We employed $\mu_S = -0.1$, $\alpha_S = 0.3$, $\beta_S = 0.26$, $\mu_L = 0.08$, $\alpha_L = 0.27$, and $\beta_L = 0.31$ on the basis of the data for 20% tumor content.

Web Resources

The URLs for data presented herein are as follows:

ATCC, http://www.atcc.org/common/cultures/NavByApp.cfm
BACPAC Resources Center, http://bacpac.chori.org/
CNAG, http://www.genome.umin.jp/
dChip, http://www.dchip.org/
Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/Omim/ (for *JAK2,* AML, PV, ET, and IMF)
PLASQ, http://genome.dfci.harvard.edu/˜tlaframb/PLASQ/

References

1. Mei R, Galipeau PC, Prass C, Berno A, Ghandour G, Patil N, Wolff RK, Chee MS, Reid BJ, Lockhart DJ (2000) Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays. Genome Res 10:1126–1137
2. Horvath A, Boikos S, Giatzakis C, Robinson-White A, Groussin L, Griffin KJ, Stein E, Levine E, Delimpasi G, Hsiao HP, et al (2006) A genome-wide scan identifies mutations in the gene encoding phosphodiesterase 11A4 (*PDE11A*) in individuals with adrenocortical hyperplasia. Nat Genet 38:794–800
3. Lindblad-Toh K, Tanenbaum DM, Daly MJ, Winchester E, Lui WO, Villapakkam A, Stanton SE, Larsson C, Hudson TJ, Johnson BE, et al (2000) Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. Nat Biotechnol 18:1001–1005
4. Knudson AG (2001) Two genetic hits (more or less) to cancer. Nat Rev Cancer 1:157–162
5. Baxter EJ, Scott LM, Campbell PJ, East C, Fourouclas N, Swanton S, Vassiliou GS, Bench AJ, Boyd EM, Curtin N, et al (2005) Acquired mutation of the tyrosine kinase JAK2 in human myeloproliferative disorders. Lancet 365:1054–1061
6. James C, Ugo V, Le Couedic JP, Staerk J, Delhommeau F, Lacout C, Garcon L, Raslova H, Berger R, Bennaceur-Griscelli A, et al (2005) A unique clonal JAK2 mutation leading to constitutive signalling causes polycythaemia vera. Nature 434:1144–1148
7. Kralovics R, Passamonti F, Buser AS, Teo SS, Tiedt R, Passweg JR, Tichelli A, Cazzola M, Skoda RC (2005) A gain-of-function mutation of JAK2 in myeloproliferative disorders. N Engl J Med 352:1779–1790
8. Levine RL, Wadleigh M, Cools J, Ebert BL, Wernig G, Huntly BJ, Boggon TJ, Wlodarska I, Clark JJ, Moore S, et al (2005) Activating mutation in the tyrosine kinase JAK2 in polycythemia vera, essential thrombocythemia, and myeloid metaplasia with myelofibrosis. Cancer Cell 7:387–397
9. Kennedy GC, Matsuzaki H, Dong S, Liu WM, Huang J, Liu G, Su X, Cao M, Chen W, Zhang J, et al (2003) Large-scale genotyping of complex DNA. Nat Biotechnol 21:1233–1237
10. Zhao X, Li C, Paez JG, Chin K, Janne PA, Chen TH, Girard L, Minna J, Christiani D, Leo C, et al (2004) An integrated view of copy number and allelic alterations in the cancer genome using single nucleotide polymorphism arrays. Cancer Res 64:3060–3071
11. Huang J, Wei W, Zhang J, Liu G, Bignell GR, Stratton MR, Futreal PA, Wooster R, Jones KW, Shapero MH (2004) Whole genome DNA copy number changes identified by high density oligonucleotide arrays. Hum Genomics 1:287–299
12. Bignell GR, Huang J, Greshock J, Watt S, Butler A, West S, Grigorova M, Jones KW, Wei W, Stratton MR, et al (2004) High-resolution analysis of DNA copy number using oligonucleotide microarrays. Genome Res 14:287–295
13. Wang ZC, Buraimoh A, Iglehart JD,Richardson AL (2006) Genome-wide analysis for loss of heterozygosity in primary and recurrent phyllodes tumor and fibroadenoma of breast using single nucleotide polymorphism arrays. Breast Cancer Res Treat 97:301–309
14. Zhou X, Mok SC, Chen Z, Li Y, Wong DT (2004) Concurrent analysis of loss of heterozygosity (LOH) and copy number abnormality (CNA) for oral premalignancy progression using the Affymetrix 10K SNP mapping array. Hum Genet 115:327–330
15. Matsuzaki H, Dong S, Loi H, Di X, Liu G, Hubbell E, Law J, Berntsen T, Chadha M, Hui H, et al (2004) Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. Nat Methods 1:109–111
16. Nannya Y, Sanada M, Nakazaki K, Hosoya N, Wang L, Hangaishi A, Kurokawa M, Chiba S, Bailey DK, Kennedy GC, et al (2005) A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. Cancer Res 65:6071–6079
17. Beroukhim R, Lin M, Park Y, Hao K, Zhao X, Garraway LA, Fox EA, Hochberg EP, Mellinghoff IK, Hofer MD, et al (2006) Inferring loss-of-heterozygosity from unpaired tumors using

high-density oligonucleotide SNP arrays. PLoS Comput Biol 2:e41

18. Laframboise T, Harrington D, Weir BA (2007) PLASQ: a generalized linear model-based procedure to determine allelic dosage in cancer cells from SNP array data. Biostatistics 8: 323–336
19. Kralovics R, Teo SS, Li S, Theocharides A, Buser AS, Tichelli A, Skoda RC (2006) Acquisition of the V617F mutation of JAK2 is a late genetic event in a subset of patients with myeloproliferative disorders. Blood 108:1377–1380
20. Wang L, Ogawa S, Hangaishi A, Qiao Y, Hosoya N, Nanya Y, Ohyashiki K, Mizoguchi H, Hirai H (2003) Molecular characterization of the recurrent unbalanced translocation der(1;7)(q10;p10). Blood 102:2597–2604
21. Huang J, Wei W, Chen J, Zhang J, Liu G, Di X, Mei R, Ishikawa S, Aburatani H, Jones KW, et al (2006) CARAT: a novel method for allelic detection of DNA copy number changes using high density oligonucleotide arrays. BMC Bioinformatics 7:83
22. Dugad R, Desai U (1996) A tutorial on hidden Markov models. Technical report SPANN-96.1. Signal Processing and Artificial Neural Networks Laboratory, Bombay, India
23. Raghavan M, Lillington DM, Skoulakis S, Debernardi S, Chaplin T, Foot NJ, Lister TA, Young BD (2005) Genome-wide single nucleotide polymorphism analysis reveals frequent partial uniparental disomy due to somatic recombination in acute myeloid leukemias. Cancer Res 65:375–378
24. Fitzgibbon J, Smith LL, Raghavan M, Smith ML, Debernardi S, Skoulakis S, Lillington D, Lister TA, Young BD (2005) Association between acquired uniparental disomy and homozygous gene mutation in acute myeloid leukemias. Cancer Res 65:9152–9154
25. Najfeld V, Montella L, Scalise A, Fruchtman S (2002) Exploring polycythaemia vera with fluorescence in situ hybridization: additional cryptic 9p is the most frequent abnormality detected. Br J Haematol 119:558–566
26. Peiffer DA, Le JM, Steemers FJ, Chang W, Jenniges T, Garcia F, Haden K, Li J, Shaw CA, Belmont J, et al (2006) High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. Genome Res 16: 1136–1148
27. Scott LM, Scott MA, Campbell PJ, Green AR (2006) Progenitors homozygous for the V617F mutation occur in most patients with polycythemia vera, but not essential thrombocythemia. Blood 108:2435–2437

Human Molecular Genetics, 2007, Vol. 16, No. 20 2494–2505
doi:10.1093/hmg/ddm205
Advance Access published on July 31, 2007

Evaluation of genome-wide power of genetic association studies based on empirical data from the HapMap project

Yasuhito Nannya[1,2,4], Kenjiro Taura[3], Mineo Kurokawa[1], Shigeru Chiba[2] and Seishi Ogawa[2,4,*]

[1]Department of Hematology/Oncology, [2]Department of Cell Therapy and Transplantation Medicine, Graduate School of Medicine and [3]Department of Information and Communication Engineering, Graduate School of Information Science, University of Tokyo, Tokyo 113-8655, Japan and [4]Core Research for Evolutional Science and Technology, Japan Science and Technology Agency, Saitama 332-0012, Japan

Received April 7, 2007; Revised and Accepted July 22, 2007

With recent advances in high-throughput single nucleotide polymorphism (SNP) typing technologies, genome-wide association studies have become a realistic approach to identify the causative genes that are responsible for common diseases of complex genetic traits. In this strategy, a trade-off between the increased genome coverage and a chance of finding SNPs incidentally showing a large statistics becomes serious due to extreme multiple-hypothesis testing. We investigated the extent to which this trade-off limits the genome-wide power with this approach by simulating a large number of case-control panels based on the empirical data from the HapMap Project. In our simulations, statistical costs of multiple hypothesis testing were evaluated by empirically calculating distributions of the maximum value of the χ^2 statistics for a series of marker sets having increasing numbers of SNPs, which were used to determine a genome-wide threshold in the following power simulations. With a practical study size, the cost of multiple testing largely offsets the potential benefits from increased genome coverage given modest genetic effects and/or low frequencies of causal alleles. In most realistic scenarios, increasing genome coverage becomes less influential on the power, while sample size is the predominant determinant of the feasibility of genome-wide association tests. Increasing genome coverage without corresponding increase in sample size will only consume resources without little gain in power. For common causal alleles with relatively large effect sizes [genotype relative risk ≥1.7], we can expect satisfactory power with currently available large-scale genotyping platforms using realistic sample size (∼1000 per arm).

INTRODUCTION

Genome-wide association studies have been proposed as a strategy to identify genetic factors with small to moderate genetic effects in the development of human diseases, as typically assumed for a common disease common variant (CDCV) model (1). In this strategy, a disease-associated locus is identified through single nucleotide polymorphisms (SNPs) that show 'significantly' different allele frequencies between affected (cases) and unaffected (controls) individuals, and a large number of SNPs are tested for association in an attempt to realistically identify such SNPs (2,3). Although only a theoretical perspective a decade ago (1), with the unprecedented advance in large-scale genotyping technologies (4–6), it has now become a realistic approach to exploring the genetic basis of human disease (7,8). In addition, recent efforts in the International HapMap Project to understand the genetic diversity among human populations (9,10) have greatly contributed to clarifying the extent to which the number of marker SNPs could be reduced to achieve given genome coverage, or how much genome coverage can be obtained with a given marker SNP set by optimally 'tagging' untyped SNPs based on the linkage disequilibrium (LD) observed in the human genome (11–16).

*To whom correspondence should be addressed to: Department of Cell Therapy and Transplantation Medicine, The 21st Century COE Program, Graduate School of Medicine, University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo 113-8655, Japan. Tel: +81 358008741; Fax: +81 358046261; Email: sogawa-tky@umin.ac.jp

Meanwhile, the major interest of the most researchers, who plan genetic association studies, would be the practical success rates in such attempts and their efficient study designs, rather than mere genome coverage (17,18), because increase in genome coverage might not be linearly translated into gain in power (19,20). In addition, the more SNPs are genotyped to achieve better genome coverage, the higher hurdle is imposed for a target allele to be detected.

This dilemma, known as the trade-off between increased genome coverage and the consequent inflation of null statistics due to extreme multiple testing, is a unique feature of genetic association studies, and is best described by considering the distributions of test statistics for markers truly associated with a causative allele ('causal distribution') and for all other markers ('null distribution') (21). Regardless of the properties of the causative SNP and whether one or more tagging strategies are used, the null distribution for a given marker set depends on its genome coverage in the study population. In particular, the null distribution with complete genome coverage is related to the overall diversity of the human genome and should substantially shift to the right (7,8,22). On the other hand, for a given disease model, the size of the test statistic expected for the causative SNPs is limited by the number of samples to be analyzed, once they are directly captured by one or more marker SNPs. After all, the feasibility of genome-wide association studies, or the required sample size to obtain realistic power, is determined by the overall diversity of the human genome, or given restricted study resources, the diversity of the human genome determines the property of disease-associated SNPs that can be detected with this approach.

Our questions are, therefore, how diverse is the human genome in view of conducting genome-wide association studies, how much power could be obtained to identify causative SNPs given that diversity and how the typical study parameters affects the power in that situation? To answer these questions, we need to evaluate both null and causal distributions in a quantitative manner. Because both distributions intrinsically depend on the LD structure within N (typically $>\sim 10^{5-6}$) interrelated marker SNPs and the particular location of causative SNPs within the genome, they cannot be calculated in an algebraic manner, but need to be estimated based on the observed data of human genome variations (10,21). So we approach these issues by extensively simulating a large number of case-control panels under both null and alternative scenarios based on the data from the International HapMap Consortiums (9,10), and assess the feasibility and efficient designs of whole genome association studies by estimating the genome-wide power that would be obtained using this genetic approach under varying study conditions.

RESULTS

Estimation of null distributions of the maximum χ^2 statistics

In considering the issue of multiple testing in genetic association studies, it is convenient to evaluate the maximum value of the χ^2 statistic [$\max(\chi^2)$] in all the marker SNPs that are truly unrelated to the causative SNP (21). Different statistics can be used (23–26), but the power calculated for this statistic, i.e. the probability of $\max(\chi^2)$ indicating a true association, will provide a reasonable bottom line to discuss the feasibility of typical genetic association studies (21). When all N marker SNPs are independent, the null distribution for $\max(\chi^2)$ is given as

$$\varphi_{\mathrm{N}}(\chi^2) = \frac{\mathrm{d}}{\mathrm{d}\chi^2}\{\phi(\chi^2)^{\mathrm{N}}\},$$

where $\phi(\chi^2)$ is the cumulative density function of the χ^2 distribution (d.f. = 1). However, since SNPs in real marker sets are variably degenerated due to the presence of LD between adjacent SNPs, we empirically estimated the distribution of $\max(\chi^2)$ for a series of marker sets by simulating 10 000 null case-control panels, where each panel was generated by randomly resampling phased chromosomes from the HapMap data sets, and $\max(\chi^2)$ was calculated for each case-control panel. Although the number of resampled chromosomes for each case-control panel (i.e. the sample size) does not significantly affect the distributions (data not shown), there arises some concern about the possibility of underestimating the null distributions due to resampling from very limited numbers of chromosomes, because the latter procedure could restrict the freedom of allelic segregation within the same chromosome. To address this issue, we progressively divided the whole genome into larger numbers of sub-blocks consisting of 10 000 to 10 SNPs in the HapMap Phase II set, and resampled these sub-blocks to simulate distributions of $\max(\chi^2)$. Reducing the mean block size down to 7.1 kb, these divisions allow for greater freedom of allelic segregation, but does not significantly affect the $\max(\chi^2)$ distributions until the resampled block size becomes smaller than the mean LD length (27), indicating that our simulations are not likely to substantially underestimate the null distributions (Supplementary Material, Figure S1).

Figure 1 A shows the simulated null distributions in the CEU panel for varying numbers of randomly selected SNPs ('correlated' SNP sets). The number of segregating or polymorphic markers contained in each random set is designated as Ns. The theoretical distribution for the same numbers (Ns) of 'independent' SNPs, $\varphi_{\mathrm{Ns}}(\chi^2)$, is also provided (Fig. 1B). The null distribution increases as the number of randomly selected SNP markers increases, and in a random 1000K set containing 681K segregating SNPs, the threshold χ^2 value that provides a genome-wide *P*-value of 0.05 or 0.01 becomes as large as 27.6 or 30.5, respectively. On the other hand, reflecting the growing inter-marker LD intensity, the empirical distributions gradually deviate from the theoretical ones, $\varphi_{\mathrm{Ns}}(\chi^2)$'s, for increasing Ns within the corresponding marker sets, underscoring the importance of considering inter-marker LD to avoid overestimation of the statistical threshold for multiple testing, especially for higher marker density.

Evaluation of the inter-marker LD

The intensity of the inter-marker LD in a given marker set is more simply evaluated by fitting the simulated distribution to a theoretical one for independent Nc makers, $\varphi_{\mathrm{Nc}}(\chi^2)$ (see Methods). Irrespective of marker sets, fitting is finely

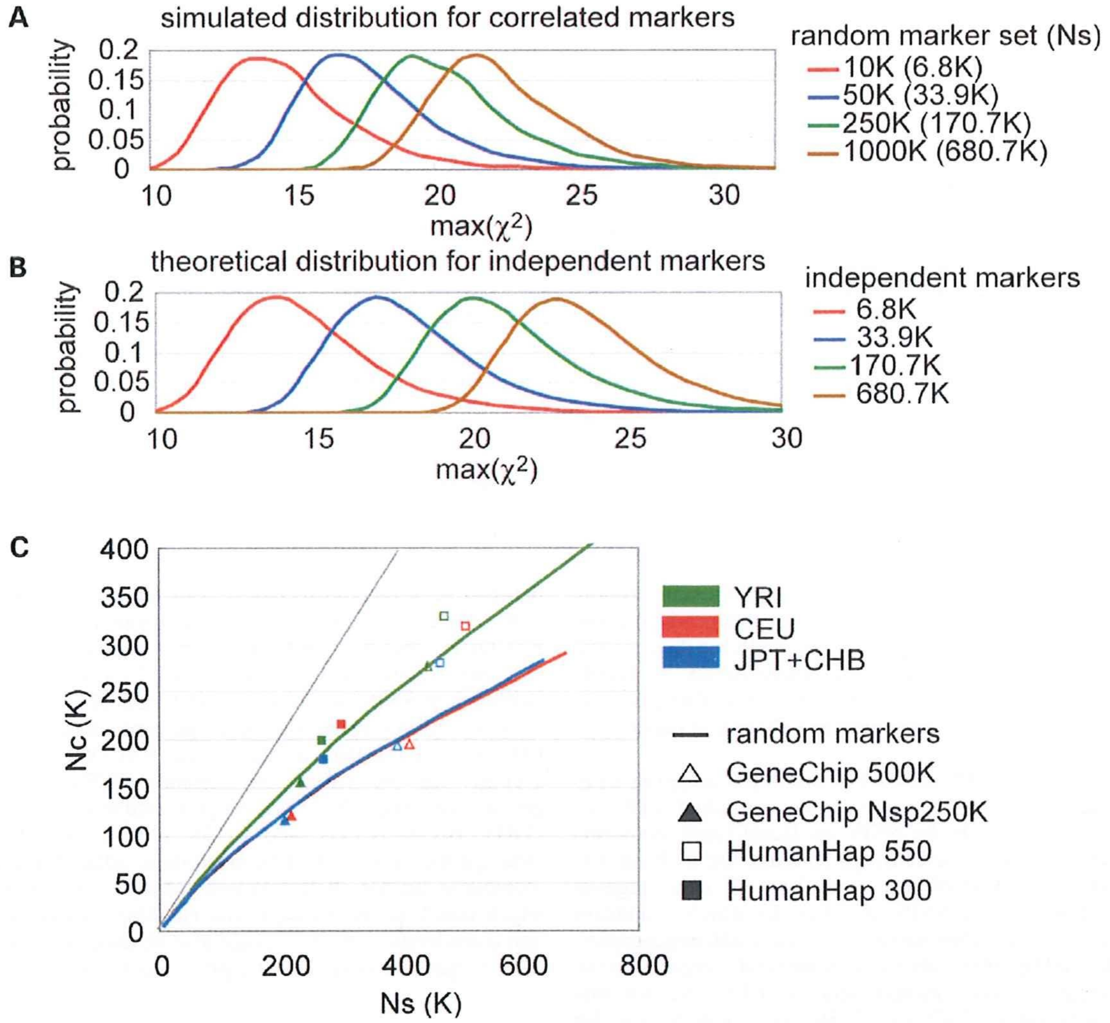

Figure 1. Null distributions of max(χ^2) and the effective number of independent SNPs (Nc) for various marker sets. Distributions of max(χ^2) for all null SNPs (null distributions) were simulated for increasing numbers of randomly selected SNP markers in the CEU panel. Ten thousand null panels, each consisting of 1000 cases and 1000 controls, were generated for the indicated marker sets by randomly resampling phased autosomal chromosomes from the HapMap Phase II data in CEU (**A**). Theoretical null distributions corresponding to each SNP set, $\varphi_{Ns}(\chi^2)$, were calculated assuming all Ns segregating SNPs therein are independent (**B**). The effective numbers of hypothetical independent SNPs (Nc) were estimated by fitting simulated null distributions to theoretical ones for Nc independent SNPs, $\varphi_{Nc}(\chi^2)$, for the indicated SNP sets, and are plotted against the number of segregating SNPs of the corresponding marker set (Ns) for different HapMap panels (**C**).

performed except in the vicinity of the maximal points (Supplementary Material, Figure S2). In particular, the distribution in extreme χ^2 values is satisfactorily approximated to provide a rough estimate of the nominal *P*-value for given genome-wide thresholds as confirmed by the concordance of the upper *p* point in the simulated distribution with the upper p/Nc point in the χ^2 distribution (d.f. = 1) (Bonferroni) (Table 1). In this formulation, it is reasonable to regard Nc as the number of hypothetical independent SNPs equivalent to the corresponding marker set, where the null distribution for a large number of mutually degenerated SNPs is described by an integer and the mean intensity of the inter-marker LD is measured through the Nc/Ns ratio.

Nc values were calculated for a variety of randomly selected SNP marker sets and plotted against the number of segregating SNP markers therein (Fig. 1C). As the Phase II data contain most of the SNPs in commercially available platforms, including Affymetrix® GeneChip® and Illumina® HumanHap® arrays (28–30), Nc values were also evaluated for these platforms (Supplemental Material, Table S1). Note that the numbers of segregating SNP markers varies among different HapMap panels, even though the same numbers of SNPs are randomly selected for each panel (Supplementary Material, Figure S3). Figure 1C illustrates how the degree of degeneration within marker SNPs increases in different HapMap panels as more marker SNPs are selected.

Table 1. Size of null distributions of max(χ^2) in various marker sets in the CEU panel

Platform	Ns	Nc	Fold degeneration	$P = 0.05$			$P = 0.01$		
				Nominal P^a	Actual^b	Bonferroni^c	Nominal P^a	Actual^b	Bonferroni^c
Random 10K	6.8K	6K	1.1	7.99×10^{-6}	19.94	19.86	1.57×10^{-6}	23.06	22.95
Random 30K	20.6K	17K	1.2	2.86×10^{-6}	21.91	21.85	5.73×10^{-7}	25.00	24.95
Random 50K	33.9K	27K	1.3	1.76×10^{-6}	22.84	22.74	4.01×10^{-7}	25.69	25.84
Random 125K	85.1K	60K	1.4	7.39×10^{-6}	24.51	24.28	1.56×10^{-7}	27.51	27.39
Random 250K	170.7K	105K	1.6	4.52×10^{-7}	25.46	25.36	9.04×10^{-8}	28.57	28.47
Random 500K	340.4K	179K	1.9	2.45×10^{-7}	26.64	26.39	5.39×10^{-8}	29.57	29.50
Random 1000K	680.7K	290K	2.3	1.48×10^{-7}	27.62	27.32	3.41×10^{-8}	30.46	30.44
GeneChip 500K	417.8K	196K	2.1	2.05×10^{-7}	26.99	26.56	4.94×10^{-8}	29.74	29.68
GeneChip Nsp250K	219.4K	120K	1.8	3.69×10^{-7}	25.85	25.62	7.94×10^{-8}	28.82	28.73
GeneChip 100K	101.3K	62K	1.6	7.75×10^{-7}	24.42	24.34	1.38×10^{-7}	27.75	27.45
HumanHap 300	305.1K	215K	1.4	2.18×10^{-7}	26.87	26.74	4.06×10^{-8}	30.12	29.86
HumanHap 550	513.8K	318K	1.6	1.41×10^{-7}	27.71	27.50	2.90×10^{-8}	30.77	30.62
HapMap Phase II	2557.4K	603K	4.2	7.09×10^{-8}	29.04	28.74	1.48×10^{-8}	32.08	31.86
ENCODE 7 regions	7.7K	1.3K	5.8						

[a] Nominal *P*-value to reach given experiment-wide significance obtained from actual distribution.
[b] The upper 1–*P* point of the actual null distribution.
[c] The argument of χ^2 distribution (d.f.=1) for cumulative density 1 − *P*/Nc.

For example, 681K segregating SNPs within a random 1000K set in the CEU panel are equivalent to independent 290K SNPs, indicating that in this panel, these SNPs are degenerated 2.3-fold. On the other hand, the degeneration in 1000K random markers is reduced to 1.8-fold for the YRI panel, as expected from the lower inter-marker LD for this panel compared to that of CEU.

The SNPs on the Affymetrix® GeneChip® mapping array sets are degenerated to the same degree as random SNP sets, reflecting the fact that the SNPs on GeneChip® platforms are virtually randomly selected. In contrast, the SNPs on the Illumina® HumanHap300 are selected by efficiently tagging the HapMap Phase I SNPs in CEU, in which redundant SNPs are effectively eliminated (28). As a result, degeneration in the HumanHap300 is substantially reduced compared to the corresponding random marker sets. In CEU, Nc for this 305.1K segregating SNP set (215K Nc) exceeds that for 417.8K segregating SNPs on GeneChip® 500K set (196K), as predicted by the higher genome coverage of the former set (see Table 1 and Supplementary Material, Figure S4). The tagging for CEU also increases the Nc in JPT+CHB, suggesting that tagging in one panel is also effective to a certain degree for another (31,32). The tagging seems to be less efficient in YRI, because the Nc value of HumanHap300® in YRI is less deviated from that of the random marker set with a corresponding Ns. In HumanHap550®, more tag SNPs are selected from YRI, which contributes to the relative increase in Nc for this marker set compared to that for the corresponding random marker SNP set.

Estimation of Nc for common SNPs in complete genome coverage

It is particularly interesting to calculate the Nc values for the ENCODE regions, in which human variations have been most densely explored. Currently 10 regions have been extensively genotyped in the ENCODE Project (http://www.hapmap.org/downloads/encode1.html.en), of which we used 7 regions that had been randomly chosen from the genome. A total of 7741, 9832 and 7396 SNPs are segregated in these seven ENCODE regions, and they are equivalent to 1340 (5.8-fold), 2580 (3.8-fold), and 1460 (5.1-fold) hypothetical independent SNPs, in the CEU, YRI, and JPT+CHB panels, respectively. Assuming the entire genome shows the similar LD intensity to that in the seven ENCODE regions on average, the Nc values for common SNPs in complete genome coverage (Nc^G) are roughly estimated to be 1971K (YRI), 1023K (CEU), and 1115K (JPT+CHB) (Table 2), although the values would be much more inflated if rare polymorphisms [minor allele frequency (MAF) <0.01], many of which could not be found in the HapMap panels, are taken into consideration. Nc/Nc^G could also be used as another indicator of genome coverage of a given marker set.

Causal distribution of max(χ^2)

In view of power estimation, our next interest was the expected size of causal distributions relative to that of the inflated null distributions under varying disease/study parameters that affect the former distributions. To illustrate this, we simulated causal distributions of max(χ^2) for representative CEU alleles assumed to be causative (Fig. 2). Two thousand case-control panels were generated for each simulation, in which phased HapMap SNPs within 500 Kb around the causative locus were randomly resampled assuming a multiplicative model with varying genotype relative risks (GRRs) and the max(χ^2) was calculated for the resampled marker SNPs on GeneChip® 500K. Prevalence of the trait was set to 0.05. While the χ^2 threshold for genome-wide p of 0.05 could inflate from 19.9 for the random 10K set (6K Nc; semi-solid line) to as high as 29.8 for complete genome coverage (1023K Nc^G; dotted lines), these costs of multiple testing are acceptable when LD capture of the causative SNP by one or more markers with high correlation coefficient (r^2) can create large causal distributions with practical sample sizes (Fig. 2D–F), i.e. when the causal allele is common

Table 2. The number of corresponding independent markers

	ENCODE[a]	Whole genome[b]	All Phase II[c]
YRI	2580	1971K	1049K
CEU	1340	1023K	603K
JPT + CHB	1460	1115K	632K

[a]Nc values calculated for combined SNPs from seven regions.
[b]Nc of ENCODE regions are extrapolated to the entire genome.
[c]Nc of all SNPs in the HapMap Phase II.

(MAF > 0.2) and has a large GRR (>1.7) (Fig. 2A, D and G). In contrast, in the case where the causal allele with smaller MAF value (<0.2) or with a modest to weak GRR (<1.5) is to be detected, the trade-off between increased chance to capture the allele with higher r^2 using more markers and the accompanying cost of multiple testing can offset the power to varying degrees (Fig. 2A–C, G–I). The effect of 'collaborative' capture, i.e. the probability of detecting an association by one of the multiple surrounding marker SNPs other than the SNPs showing max(r^2), creates measurable gain in causal distributions and overall power, but does not essentially influence the above observations (Supplementary Material, Figure S5).

Estimation of genome-wide power

Based on the above consideration, we estimated the genome-wide power in genetic association studies for common (MAF ≧ 0.05) causal alleles with weak to moderate genetic effects. To do this, after assuming all the common SNPs in the human genome being equally causative, we used two sets of SNPs, the Ref$^{\mathrm{ENCODE}}$ and the Ref$^{\mathrm{Phase\ II\ 5Kb}}$ sets (see Methods), as references that are considered as random sampling from the entire SNPs. For each putative causative SNP, we simulated case-control panels as described in the previous section, and calculated the single point power as the proportion of simulated panels whose max(χ^2) exceeded a predetermined χ^2 threshold corresponding to a genome-wide $P = 0.01$ or 0.05 for each marker set. For genome-wide power, each single point power was averaged for all common SNPs within the reference set. For the Ref$^{\mathrm{Phase\ II\ 5Kb}}$ set, over-representation of the direct association was adjusted based on the estimated genome coverage of the Phase II data set (see Methods). Figure 3 shows the genome-wide power in the CEU panel that was calculated for the Ref$^{\mathrm{Phase\ II\ 5Kb}}$ for moderate to small effect sizes (i.e. GRR ≦ 1.7) assuming various parameter values. The calculation on the Ref$^{\mathrm{ENCODE}}$ set provides a largely equivalent estimation of the power (Supplementary Material, Figure S6), although the power is expected to be less reliable for smaller marker sets, reflecting their poor representation of the genome.

Under strong genetic effects (GRR ≧ 2.0) and large sample sizes (≧ 1500/arm), the power tends to saturate as the number of randomly selected SNPs increases (≧ 250K), because most of the common SNPs would be already captured by one or more marker SNPs with enough r^2 (Supplementary Material, Figure S4), and the capture causes large shifts of causal distributions to the extent that the cost of multiple testing is trivial (Fig. 2). On the other hand, when causative SNPs with weak to moderate genetic effects are detected with insufficient sample numbers, causal distributions cannot exceed large thresholds resulting from extreme multiple testing, even though more and more SNPs are captured by strong LD. With increasing effect size and sample number, the genome coverage is less influential except for smaller numbers of marker SNPs (<250K). The power gain obtained with increased genome-coverage tends to be offset by the increased cost of multiple testing. After all, in most scenarios, genome coverage is less influential on power when ≧ 250K random markers or equivalent tag SNPs are used. In contrast, the effect of sample numbers is predominant. To detect weak genetic effects (GRR ≦ 1.3), the number of samples becomes critical. More than 4000 samples per arm will be required, but the requirement of genome coverage is not substantially increased when more than 250K randomly selected SNPs or their equivalents are used (Fig. 3A). Given a higher genetic effect, this dependence on sample size is dramatically ameliorated, but the genome coverage remains less influential.

Power in different HapMap panels and in commercially available platforms

Power is significantly reduced in YRI compared to CEU and JPT+CHB for any marker set (Fig. 4A–C). The lower power in YRI is mainly due to the lower 'relative' genome coverage of the marker set (Nc/Nc^{G}), rather than the higher cost of type I errors in this population.

The Illumina® HumanHap® series are commercially available platforms that incorporate the tagging theory, in which marker SNPs were selected to efficiently tag the CEU SNPs in the Phase I data set. Tagging seems to be effective, since HumanHap300® in the Ref$^{\mathrm{Phase\ II\ 5Kb}}$ set shows slightly higher power than the GeneChip® 500K in CEU, although the power is slightly biased by the higher representation of the Phase I SNPs in the Ref$^{\mathrm{Phase\ II\ 5Kb}}$ set (Fig. 4D). HumanHap300® shows comparable power to that of GeneChip® 500K, but the power of HumanHap300® is significantly reduced in YRI. In HumanHap550®, more tag SNPs from YRI and JPT+CHB were added to HumanHap300®, the power is more improved in YRI and in JPT+CHB, but the power is also increased to a lesser degree in CEU reflecting a transferability of tag SNPs between CEU and JPT+CHB. The power of various commercially available platforms with various sample sizes are shown in Figure 4E (adaptive threshold) and in Supplementary Material, Figure S7 (fixed threshold). Genome coverage and power of HumanHap550® in the CEU are comparable to those of the random 1000K set (Supplementary Material, Figure S4), an equivalent to Human SNP Array 6.0® that is planned by Affymetrix® (Fig. 4E). Nevertheless, and in spite of the significant difference in cost, the gain of power in HumanHap550® is not so prominent. Also note that the power calculation for HumanHap550® could be slightly biased by using the subset of the Phase II SNPs as a reference.

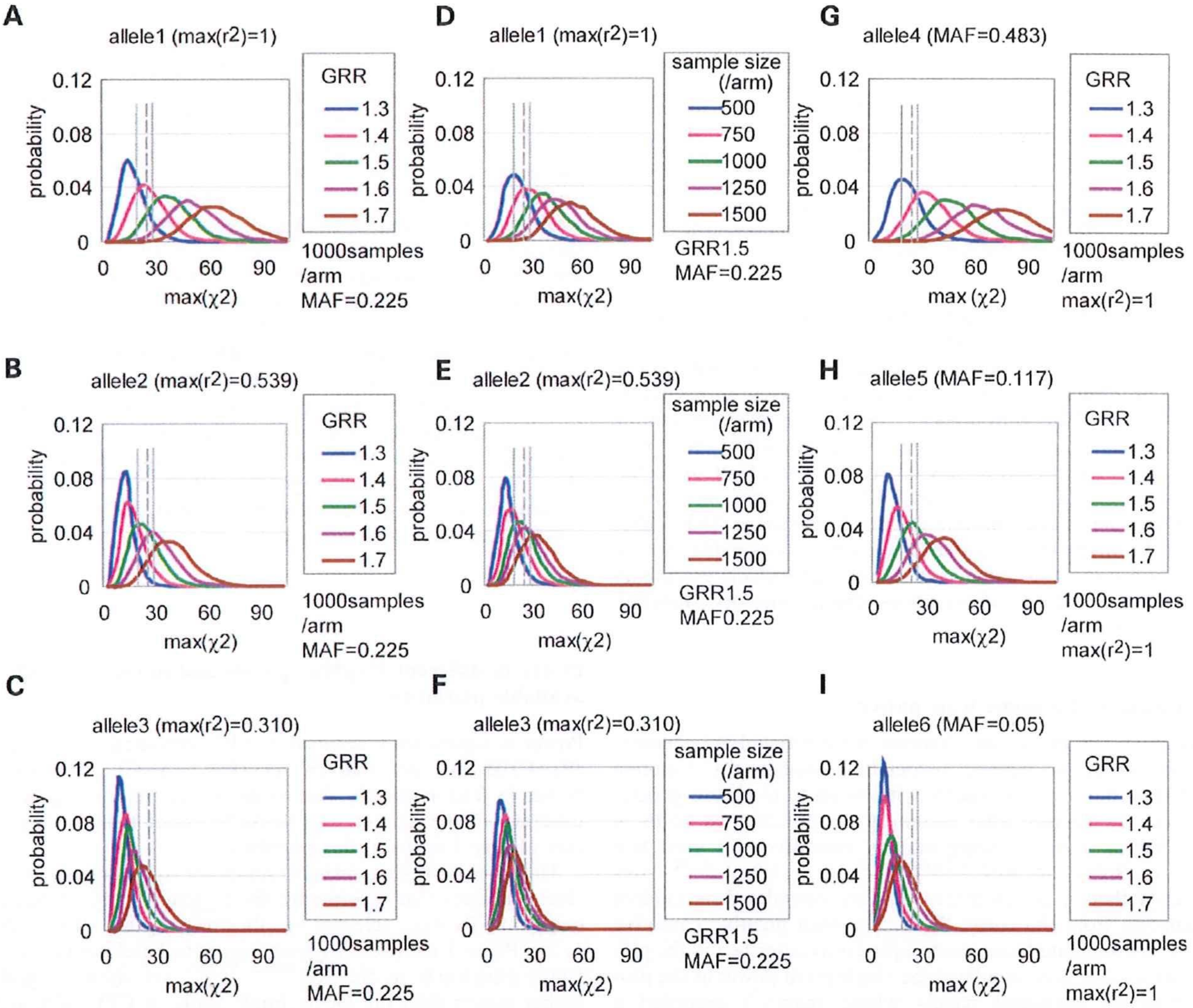

Figure 2. Enhancement of causal distributions by various parameters. Combined effects of LD [in max(r^2)] and effect size (in GRR) on causal distributions under constant sample size (1000/arm) and MAF value (0.225) (**A**–**C**), LD and sample size under constant effect size (GRR = 1.5) and MAF value (0.225) (**D**–**F**), and MAF and effect size under constant sample size (1000/arm) and LD [max(r^2) = 1.0] (**G**–**I**), are illustrated based on the simulations for six representative CEU alleles analyzed on GeneChip® 500K [rs9782915 in (A and D); rs7543006 in (B and E); rs731030 in (C and F); rs6603803 in (G); rs3052 in (H); rs1307490 in (I)]. Thresholds for genome-wide *P*-value of 0.05 are indicated for random 10K (solid lines), GeneChip 500K (dashed lines), and complete genome coverage (dotted lines), corresponding to Nc values of 6K, 196K, and 1023K (Nc^G), respectively. Effects of collaborative capture by nearby markers are incorporated, but they are generally small (Supplementary Material, Figure S5).

Power depends on allele frequencies of causative alleles

Power strongly depends on MAF of causative alleles, and detecting rare causative alleles is very difficult (Fig. 2) (8,20) for two reasons. First, rare variants are difficult to capture in high r^2 values. With currently available platforms (GeneChip® 500K or HumanHap550®), most SNPs with more than 0.10 MAF values are captured in high r^2, which could be effectively detected in high power given moderate GRRs (≧ 1.5) and sample size (≧ 1000/arm) (Fig. 5). In contrast, capturing rare causal SNPs (MAF < 0.10) requires many more marker SNPs or their combinations than capturing common SNPs at the more cost of multiple hypothesis testing. Second, even when captured in high r^2 with one or more marker SNPs, associations with these rare SNPs are more difficult to detect than those with common SNPs (Fig.5). In common diseases, the existence of multiple phenocopy variants would further compromise detection (multiple rare variants) (33,34). Thus, regardless of genome coverage, power is consistently lower for less common SNPs (Fig. 6A and C). To detect rare causative SNPs, we need not only to invest in genotyping large numbers of marker SNPs with

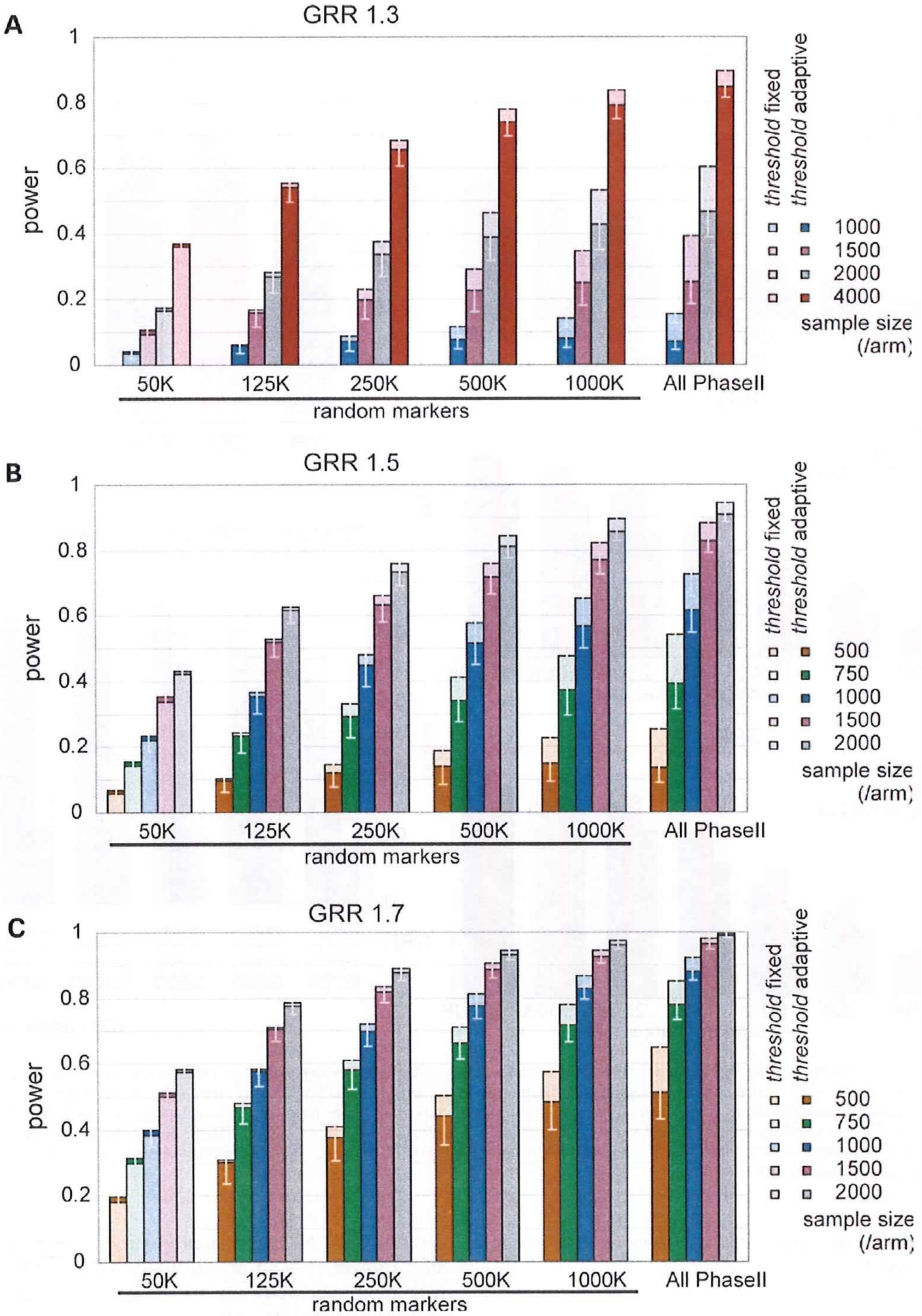

Figure 3. Genome-wide power of association studies for common causal alleles with weak to moderate genetic effects. Genome-wide power was calculated in CEU by averaging single point power for each putative causal allele over all common (MAF ≧ 0.05) SNPs in the $\text{Ref}^{\text{Phase II 5Kb}}$ reference set, with increasing marker and sample sizes for small to moderate GRRs (1.3–1.7) in multiplicative disease models. Power was computed using adaptive thresholds for max(χ^2) that provides a genome-wide *P*-value of 0.05 (dark columns) or using a fixed threshold ($P = 1 \times 10^{-6}$; light columns) for each marker set. The power with an adaptive threshold for a genome-wide *P*-value of 0.01 was also indicated by a lower bar within each column.

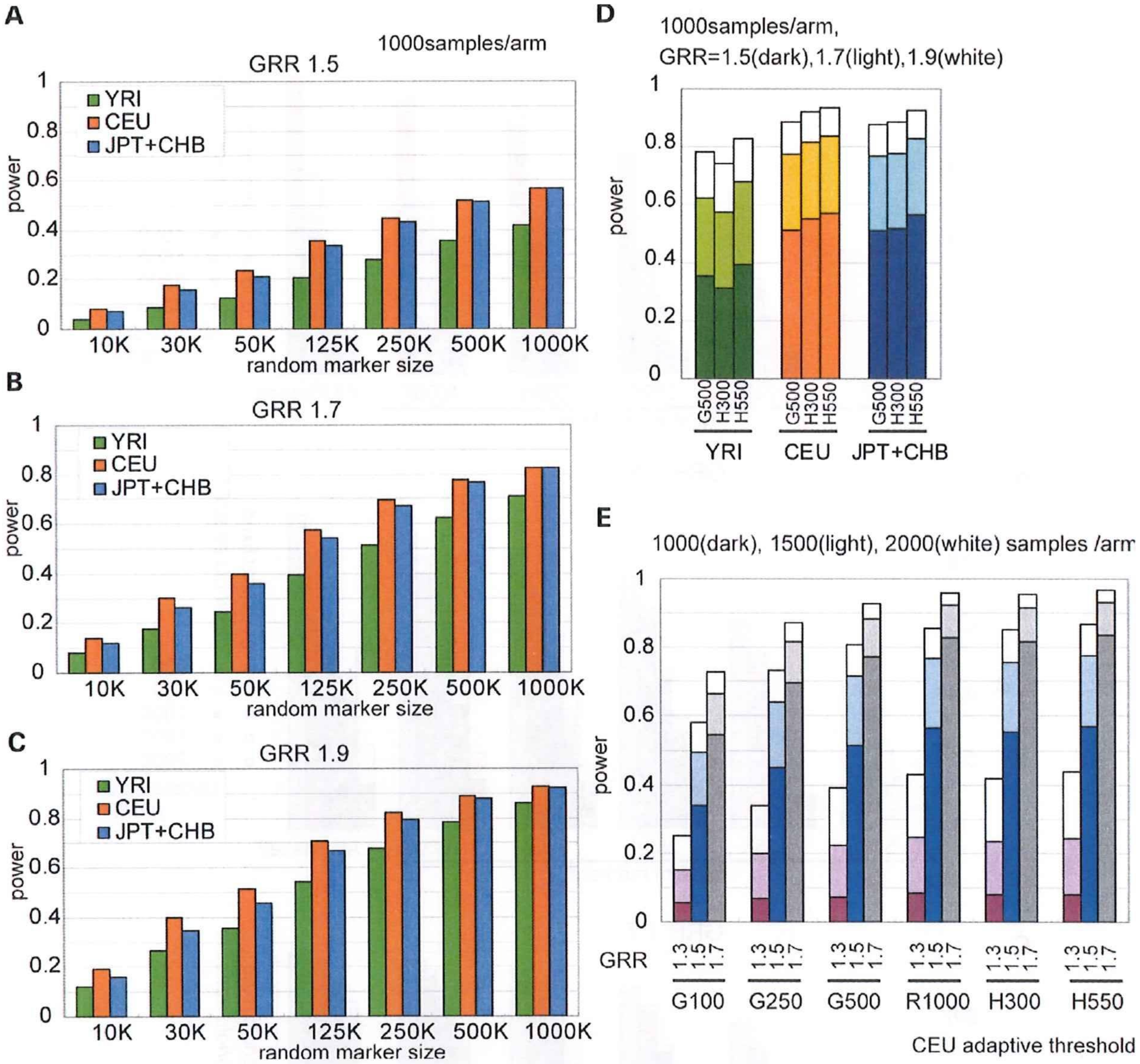

Figure 4. Comparison of power in different HapMap panels and in commercially available genotyping platforms. Genome-wide power was calculated for different HapMap panels in a variety of marker sets, including indicated numbers of randomly selected SNP markers for GRR=1.5 (**A**), GRR=1.7 (**B**), and GRR=1.9 (**C**). Statistical thresholds were adjusted to provide genome-wide *P*-values of 0.05. Genome-wide power was also calculated for commercially available genotyping platforms in different HapMap panels (**D**) and varying sample numbers and effect sizes for CEU (**E**). The examined platforms are GeneChip® 100K (G100), GeneChip® Nsp250K (G250), GeneChip® 500K (G500), HumanHap300® (H300) and HumanHap550® (H550). Power in a random 1000K set (R1000) is shown for comparison in E.

low MAF values by any means, but also to increase the sample size (Fig. 6B and C).

Discussion

Through the current analysis, we empirically determined the size of test statistics for causal as well as null markers under varying degrees of genome coverage and realistic study parameters, and thereby demonstrated how genome-wide power is affected by the interplay between genome-coverage and other determinants. Here it is appropriate to compare the performance [power ($1 - \beta$) or sensitivity] of the different SNP sets with their specificity (or $1 - \alpha$) being constant by applying adaptive thresholds, where α denotes genome-wide type I error probability. In addition, the power calculated in this way is directly related to false positive report probability (FPRP), which is simply expressed as $1/[1+(1 - \beta)/\alpha]$, which is approximately extended to $1/[1+m(1 - \beta)/\alpha]$ assuming a total of m independent causative loci having the same effect size. Note that α is a constant for all SNP sets,

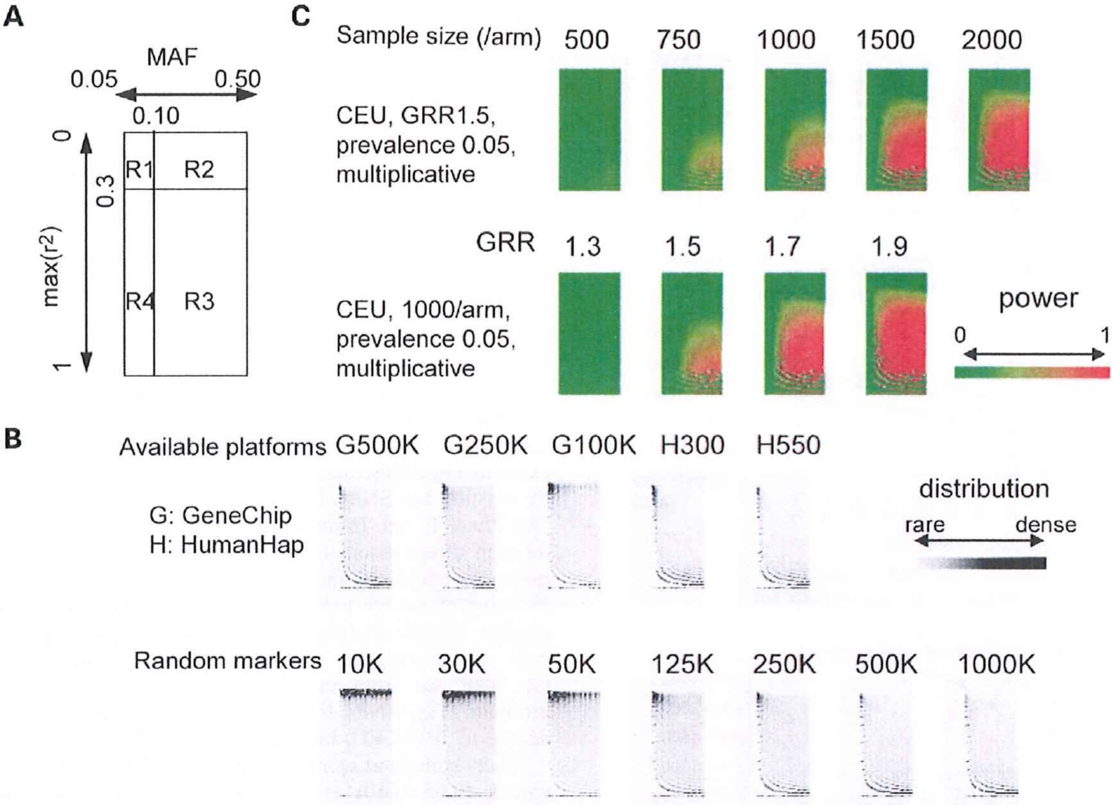

Figure 5. Impact of allele frequencies and genome coverage on genome-wide power. Reference SNPs randomly selected from the Phase II CEU set (Ref$^{\text{Phase II 5Kb}}$) are plotted onto a panel according to their MAF and the max(r^2) within the indicated marker set, and assigned into four categories; sub-common and weakly proxied SNPs [MAF < 0.10 and max(r^2) < 0.3] SNPs (R1), common and weakly proxied SNPs (MAF ≧ 0.10 and max(r^2) < 0.3) SNPs (R2), common and strongly proxied SNPs [MAF ≧ 0.10 and max(r^2) ≧ 0.3] (R3), or sub-common and strongly proxied SNPs [MAF < 0.10 and max(r^2) ≧ 0.3] (R4). (**A**). Distributions of these SNPs are shown by gray-scaled density for different marker set, where the SNP distribution shifts downward as the genome coverage improves (**B**). GeneChip® 500K, 250K (NspI), 100K, HumanHap300®, and HumanHap550® are designated as G500K, G250K, G100K, H300K, and H550K, respectively. On the other hand, neglecting the collaborative capture effect, the power for SNPs with a given MAF and max(r^2) value is largely determined by GRR and sample size. Distributions of the power are color-coded for different parameter sets as indicated (**C**). Genome-wide power is roughly estimated by taking the product sum of corresponding cells in both panels.

i.e. 0.05 or 0.01. So from our simulations, readers will easily evaluate the power and FPRP expected form given SNP set, sample size and predicted effect size. As long as practical power (for example, $1 - \beta > \alpha$) is obtained, FPRP is expected to less than 0.5, which will be satisfactory for initial discovery studies.

We estimated genome-wide thresholds based on the simulations using small numbers of HapMap chromosomes. In real studies, the threshold should be determined using their own applicable data sets, where diploid, rather than phased, chromosomes could be used when enough samples are analyzed. A larger number of chromosomes should contain more numbers of rare segregating SNPs, but these rare SNPs would not increase χ^2 thresholds substantially (22).

In terms of the effective number of independent SNPs (Nc) in various marker sets, the diversity of the human genome is likely to be on the order of 1000K in CEU and the corresponding nominal P-value giving a genome-wide α error of 0.05 is 5×10^{-8}. For moderate GRRs (≦ 1.5), this threshold could be overcome with ≦ 1500 samples per arm for very common SNPs (MAF > 0.20), but for less common SNPs or those with a small genetic effect (GRR=1.1–1.2), extremely large numbers of samples will be required (Supplementary Material, Figure S8), which urges moves toward sharing typing data across multiple groups as exemplified in recent reports that identified predisposing factors with very modest genetic effects for type 2 diabetes (35–37). The diversity of our genome may not allow for detecting very rare causative alleles (<0.01) with even smaller genetic effects (i.e. GRR < 1.1) using this approach (Fig. 6D).

Under these limitations, several issues should be considered to efficiently exploit study resources and to increase the chance of finding a true association. First, for the increased genome coverage to be effectively translated into power, it needs to be accompanied by a corresponding increase in sample size. When sample numbers are small relative to the effect size, the cost of multiple testing largely offsets the expected increase in the test statistics for causal alleles with

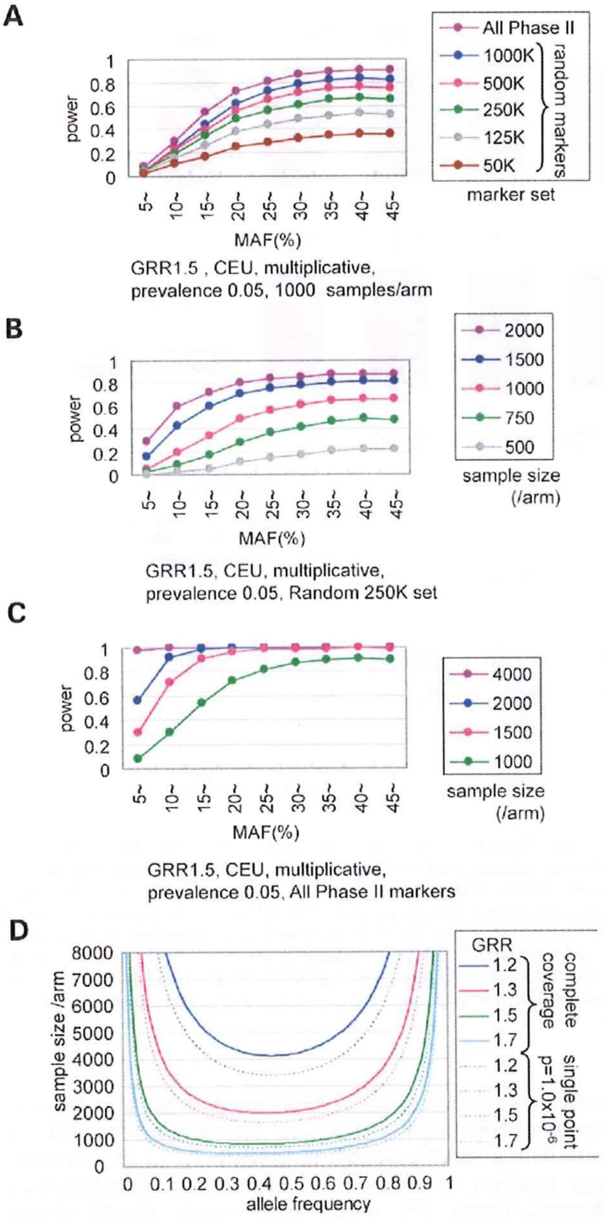

Figure 6. Effects of allele frequency on simulated power. Distribution of power on MAF in association studies are shown for varying marker sets under a constant sample size (1000 /arm) (**A**), and for varying sample sizes under a fixed marker set; GeneChip® 250K (**B**) or a hypothetical complete marker set (**C**). CEU was used for simulations with fixed GRR (1.5) and disease prevalence (0.05). The sample size that is required for detecting a causative allele with 80% power was calculated for GRRs of 1.2, 1.3, 1.5 and 1.7, assuming complete genome coverage in a multiplicative model (**D**). The significance threshold for genome-wide *P*-values of 0.05 is set assuming complete genome coverage (Nc^G=1023K, solid lines) or independent 50K markers (single point *P*-value =1 × 10^{-6}, Nc=50K, broken lines).

no measurable gain in power, and can even exceed the gain in causal distributions (Fig. 4). Increasing genome coverage with insufficient sample sizes would only consume resources with no substantial benefit in power. In addition, power tends to saturate in higher genome coverage and the effect of increasing the number of marker SNPs is less prominent compared to that of increasing sample sizes. In most simulated situations, more power is expected by doubling the sample size than by doubling the number of maker SNPs. For example, our simulations predict that doubling the sample size using GeneChip® Nsp 250K is almost certainly more efficient than analyzing half of the samples with both Nsp 250K+Sty 250K (Supplementary Material, Figure S9).

The tagging strategy or statistical imputation is effective for increasing genome coverage with limited numbers of marker SNPs (21,38,39), although it does not save the cost of multiple-hypothesis testing. The efficiency of generating a tag SNP set with higher genome coverage, however, is increasingly compromised. The additional gain in power becomes smaller with increasing genome coverage, while more and more effort will be required to find additional independent tag SNPs, because many SNPs are already captured by existing tag SNPs. In addition, we simulated power using 'All Phase II' set. In the sense that all references are captured through direct association, this marker set provides the ultimate coverage of the genome. Considering that modest increase of power using 'All Phase II' set compared with random 1000K set (Fig. 3), multimarker tagging presumably may not push up the power profoundly. Transferability of a tag SNP set from one population to another is also a problem. Tag SNPs for CEU are transferable to a certain degree to JPT+CEU, but they are less effective for YRI.

In any simulated scenarios, detecting SNPs with lower MAF values (0.05–0.10) is very difficult using whole genome approaches, which is especially true for SNPs with less than 0.05 MAF values. In this situation, genome coverage to capture these rare SNPs becomes definitely important, but the required increase in the sample size is greater for rare SNPs than for common ones.. Effort to devising SNP sets for these rare alleles, or exhaustive multimarker tests (21,38), is not likely to be rewarding unless their genetic effects are substantially large.

MATERIALS AND METHODS

HapMap data sets

The phased genotyping data of the HapMap Phase II (release 21) were obtained from the International HapMap Project web site (http://www.hapmap.org/downloads/phasing/2006-07_Phase II/) (10). It includes the data from 60 CEU parents (120 chromosomes), 60 YRI parents (120 chromosomes) and the combined set of 45 JPT and 45 CHB unrelated individuals (180 chromosomes), and is provided in three discrete sets ('all', 'consensus', and 'phased'), of which we used the former two sets for analysis. The 'all' set contains the comprehensive data of all SNPs genotyped in each population including non-segregating sites, and the 'consensus' set consists of the intersection of 'all' sets from the three population panels. The 'all' sets contain 3755 469, 368 5205 and 3776 850 SNPs for CEU, YRI and JPT+CHB, respectively, and the 'consensus' set includes 3535 396 SNPs.

Marker sets and the references for power calculation

We generated a series of marker sets consisting of 10K, 30K, 50K, 125K, 250K, 500K and 1000K SNPs, by randomly selecting SNPs from the Phase II 'all' sets for each HapMap panel. The number of segregating SNPs in each set is denoted as Ns and shown in Table 1for CEU panel. Because the Phase II 'all' set contains most of the SNPs on commercially available platforms, including Affymetrix® GeneChip® 500K (Nsp+Sty), 250K (Nsp), 100K (Hind+Xba), Illumina® HumanHap300®, and HumanHap550® (Supplementary Material, Table S1), the intersectional SNPs of these platforms with the Phase II 'all' set were incorporated into the analysis as representative SNPs of each commercial set. Annotation files for SNPs on GeneChip® series are available from the Affymetrix® web site (http://www.affymetrix.com/products/application/whole_genome.affx). The SNP information of HumanHap® series was kindly provided by Illumina® Inc. A subset of the Phase II SNPs, referred to as 'Ref$^{\mathrm{Phase\ II\ 5Kb}}$', was constructed and used as a reference in the calculation of genome-wide powers by randomly selecting SNPs from the 'consensus' set so that each SNP is, on average, 5 Kb apart from the adjacent SNPs. Combined SNPs from the 10 ENCODE regions, denoted as Ref$^{\mathrm{ENCODE}}$, were used as an alternative reference set. Only common SNPs (MAF $\geq$ 0.05) were included in the power calculations as putative causal alleles.

Simulation of case-control panels under the null hypothesis and fitting simulated distributions

Null distributions in genetic association studies are considered for only vaguely defined ensembles having limited population sizes, e.g. all adult Japanese eligible for a study. To obtain asymptotic distributions, we generated 10 000 null case-control panels by randomly resampling phased autosomal chromosomes from the 'all' set of CEU, YRI and JPT+CHB. Simulations were performed with different sample numbers, i.e. 500, 750, 1000, 1500, 2000 and 4000 per single arm. For each case-control panel, the maximum χ^2 value (max(χ^2); d.f.=1) in the standard allele test was calculated for different marker sets to obtain empirical null distributions of max(χ^2).

The simulated distributions, $\Phi(\chi^2)$, were fitted to the null distribution for hypothetical Nc independent SNPs, $\varphi_{\mathrm{Nc}}(\chi^2)$, by the least squares method as follows:

$$\mathrm{Nc} = \arg\min_{\mathrm{N}} \int (\varphi_{\mathrm{N}}(\chi^2) - \Phi(\chi^2))^2 d\chi^2$$

The Gnu Scientific Library was used to handle these functions.

Simulation of case-control studies and calculation of power

We consider multiplicative disease models showing a prevalence *e*, and assume a single causative allele whose MAF and GRR are P ($\geq$0.05) and γ, respectively. Given the penetrance for *AA*, *Aa* and *aa* genotypes as f_{AA}, f_{Aa}, and f_{aa}, respectively, expected genotype frequencies in the case and control panels are given as,

$$P(AA|\mathrm{case}) = \frac{p^2 f_{AA}}{e}$$

$$P(Aa|\mathrm{case}) = \frac{2p(1-p)f_{Aa}}{e}$$

$$P(aa|\mathrm{case}) = \frac{(1-p)^2 f_{aa}}{e}$$

$$P(AA|\mathrm{control}) = \frac{p^2(1-f_{AA})}{1-e}$$

$$P(Aa|\mathrm{control}) = \frac{2p(1-p)(1-f_{Aa})}{1-e}$$

$$P(aa|\mathrm{control}) = \frac{(1-p)^2(1-f_{aa})}{1-e}$$

where

$$e = p^2 f_{AA} + 2p(1-p)f_{Aa} + (1-p)^2 f_{aa}$$

$$f_{AA} = \gamma^2 f_{aa}, \quad f_{Aa} = \gamma f_{aa}$$

According to these allele frequencies, we generated 2000 case-control panels under the alternative hypothesis by resampling a predetermined number of phased chromosomes, and calculated max(χ^2) of the marker SNPs for each panel, where the calculations were performed only for those marker SNPs that are within 500 Kb from the putative causal SNP. The proportion of simulated case-control panels whose max(χ^2) exceeded the upper 95 or 99% point of the corresponding null distribution for that marker set was defined as the power. The genome-wide power was computed by averaging each power for all SNPs within the reference set. As the number of marker SNPs increases, up to as high as 1000K, there is a considerable chance of detecting direct associations, i.e. the causative SNP is included in the marker set. Assuming 7500K common SNPs within the human genome (17), the Phase II data set includes one-fourth (2167K common SNPs in CEU) of all the common SNPs. Based on this estimation, we excluded three-fourths of the direct associations from the calculation of genome-wide power to avoid overestimating its chance. The adjustment of direct association, however, has little influence on the results. This correction was not applied to the power calculation on the Ref$^{\mathrm{ENCODE}}$ set, because it represents the nearly complete data set for those regions.

Computational resources

All simulations were run on the GXP clustering computer system in the Department of Information and Communication Engineering, Graduate School of Information Science, University of Tokyo.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at HMG Online.

ACKNOWLEDGEMENTS

This work is totally indebted to the achievement of the International HapMap Consortium and we thank all the people who participated in the project. We also thank Jun Ohashi for helpful discussions.

Conflict of Interest statement. None declared.

FUNDING

This work was supported by Research on Measures for Intractable Diseases, Health and Labor, Sciences Research Grants, Ministry of Health, Labor and Welfare, Research on Health Sciences focusing on Drug Innovation, the Japan Health Sciences Foundation, and Core Research for Evolutional Science and Technology (CREST), Japan Science and Technology Agency.

REFERENCES

1. Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science*, **273**, 1516–1517.
2. Kruglyak, L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.*, **22**, 139–144.
3. Risch, N.J. (2000) Searching for genetic determinants in the new millennium. *Nature*, **405**, 847–856.
4. Syvanen, A.C. (2001) Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nat. Rev. Genet.*, **2**, 930–942.
5. Kennedy, G.C., Matsuzaki, H., Dong, S., Liu, W.M., Huang, J., Liu, G., Su, X., Cao, M., Chen, W., Zhang, J. *et al.* (2003) Large-scale genotyping of complex DNA. *Nat. Biotechnol.*, **21**, 1233–1237.
6. Fan, J.B., Chee, M.S. and Gunderson, K.L. (2006) Highly parallel genomic assays. *Nat. Rev. Genet.*, **7**, 632–644.
7. Hirschhorn, J.N. and Daly, M.J. (2005) Genome-wide association studies for common diseases and complex traits. *Nat. Rev. Genet.*, **6**, 95–108.
8. Wang, W.Y., Barratt, B.J., Clayton, D.G. and Todd, J.A. (2005) Genome-wide association studies: theoretical and practical concerns. *Nat. Rev. Genet.*, **6**, 109–118.
9. The International HapMap Consortium (2003) The International HapMap Project. *Nature*, **426**, 789–796.
10. The International HapMap Consortium (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.
11. Johnson, G.C., Esposito, L., Barratt, B.J., Smith, A.N., Heward, J., Di Genova, G., Ueda, H., Cordell, H.J., Eaves, I.A., Dudbridge, F. *et al.* (2001) Haplotype tagging for the identification of common disease genes. *Nat. Genet.*, **29**, 233–237.
12. Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M. *et al.* (2002) The structure of haplotype blocks in the human genome. *Science*, **296**, 2225–2229.
13. Carlson, C.S., Eberle, M.A., Rieder, M.J., Yi, Q., Kruglyak, L. and Nickerson, D.A. (2004) Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.*, **74**, 106–120.
14. Halldorsson, B.V., Istrail, S. and De La Vega, F.M. (2004) Optimal selection of SNP markers for disease association studies. *Hum. Hered.*, **58**, 190–202.
15. Zhang, K., Qin, Z., Chen, T., Liu, J.S., Waterman, M.S. and Sun, F. (2005) HapBlock: haplotype block partitioning and tag SNP selection software using a set of dynamic programming algorithms. *Bioinformatics*, **21**, 131–134.
16. Ao, S.I., Yip, K., Ng, M., Cheung, D., Fong, P.Y., Melhado, I. and Sham, P.C. (2005) CLUSTAG: hierarchical clustering and graph methods for selecting tag SNPs. *Bioinformatics*, **21**, 1735–1736.
17. Barrett, J.C. and Cardon, L.R. (2006) Evaluating coverage of genome-wide association studies. *Nat. Genet.*, **38**, 659–662.
18. Pe'er, I., de Bakker, P.I., Maller, J., Yelensky, R., Altshuler, D. and Daly, M.J. (2006) Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nat. Genet.*, **38**, 663–667.
19. Ohashi, J. and Tokunaga, K. (2001) The power of genome-wide association studies of complex disease genes: statistical limitations of indirect approaches using SNP markers. *J. Hum. Genet.*, **46**, 478–482.
20. Zondervan, K.T. and Cardon, L.R. (2004) The complex interplay among factors that influence allelic association. *Nat. Rev. Genet.*, **5**, 89–100.
21. de Bakker, P.I., Yelensky, R., Pe'er, I., Gabriel, S.B., Daly, M.J. and Altshuler, D. (2005) Efficiency and power in genetic association studies. *Nat. Genet.*, **37**, 1217–1223.
22. Neale, B.M. and Sham, P.C. (2004) The future of association studies: gene-based analysis and replication. *Am. J. Hum. Genet.*, **75**, 353–362.
23. Dudbridge, F. and Koeleman, B.P. (2003) Rank truncated product of *P*-values, with application to genomewide association scans. *Genet. Epidemiol.*, **25**, 360–366.
24. Hoh, J. and Ott, J. (2003) Mathematical multi-locus approaches to localizing complex human trait genes. *Nat. Rev. Genet.*, **4**, 701–709.
25. Hoh, J., Wille, A. and Ott, J. (2001) Trimming, weighting, and grouping SNPs in human case-control association studies. *Genome Res.*, **11**, 2115–2119.
26. Zaykin, D.V., Zhivotovsky, L.A., Westfall, P.H. and Weir, B.S. (2002) Truncated product method for combining *P*-values. *Genet. Epidemiol.*, **22**, 170–185.
27. De La Vega, F.M., Isaac, H., Collins, A., Scafe, C.R., Halldorsson, B.V., Su, X., Lippert, R.A., Wang, Y., Laig-Webster, M., Koehler, R.T. *et al.* (2005) The linkage disequilibrium maps of three human chromosomes across four populations reflect their demographic history and a common underlying recombination pattern. *Genome Res.*, **15**, 454–462.
28. Gunderson, K.L., Steemers, F.J., Lee, G., Mendoza, L.G. and Chee, M.S. (2005) A genome-wide scalable SNP genotyping assay using microarray technology. *Nat. Genet.*, **37**, 549–554.
29. Matsuzaki, H., Dong, S., Loi, H., Di, X., Liu, G., Hubbell, E., Law, J., Berntsen, T., Chadha, M., Hui, H. *et al.* (2004) Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat. Methods*, **1**, 109–111.
30. Steemers, F.J., Chang, W., Lee, G., Barker, D.L., Shen, R. and Gunderson, K.L. (2006) Whole-genome genotyping with the single-base extension assay. *Nat. Methods*, **3**, 31–33.
31. Tenesa, A. and Dunlop, M.G. (2006) Validity of tagging SNPs across populations for association studies. *Eur. J. Hum. Genet.*, **14**, 357–363.
32. de Bakker, P.I., Burtt, N.P., Graham, R.R., Guiducci, C., Yelensky, R., Drake, J.A., Bersaglieri, T., Penney, K.L., Butler, J., Young, S. *et al.* (2006) Transferability of tag SNPs in genetic association studies in multiple populations. *Nat. Genet.*, **38**, 1298–1303.
33. Pritchard, J.K. (2001) Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.*, **69**, 124–137.
34. Slager, S.L., Huang, J. and Vieland, V.J. (2000) Effect of allelic heterogeneity on the power of the transmission disequilibrium test. *Genet. Epidemiol.*, **18**, 143–156.
35. Scott, L.J., Mohlke, K.L., Bonnycastle, L.L., Willer, C.J., Li, Y., Duren, W.L., Erdos, M.R., Stringham, H.M., Chines, P.S., Jackson, A.U. *et al.* (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science*, **316**, 1341–1345.
36. Saxena, R., Voight, B.F., Lyssenko, V., Burtt, N.P., de Bakker, P.I., Chen, H., Roix, J.J., Kathiresan, S., Hirschhorn, J.N., Daly, M.J. *et al.* (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*, **316**, 1331–1336.
37. Zeggini, E., Weedon, M.N., Lindgren, C.M., Frayling, T.M., Elliott, K.S., Lango, H., Timpson, N.J., Perry, J.R., Rayner, N.W., Freathy, R.M. *et al.* (2007) Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science*, **316**, 1336–1341.
38. Lin, S., Chakravarti, A. and Cutler, D.J. (2004) Exhaustive allelic transmission disequilibrium tests as a new approach to genome-wide association studies. *Nat. Genet.*, **36**, 1181–1188.
39. Weale, M.E., Depondt, C., Macdonald, S.J., Smith, A., Lai, P.S., Shorvon, S.D., Wood, N.W. and Goldstein, D.B. (2003) Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene SCN1A: implications for linkage-disequilibrium gene mapping. *Am. J. Hum. Genet.*, **73**, 551–565.

Genome-Wide, High-Resolution Detection of Copy Number, Loss of Heterozygosity, and Genotypes from Formalin-Fixed, Paraffin-Embedded Tumor Tissue Using Microarrays

Sharoni Jacobs,[1] Ella R. Thompson,[2,3] Yasuhito Nannya,[4] Go Yamamoto,[4] Raji Pillai,[1] Seishi Ogawa,[5] Dione K. Bailey,[1] and Ian G. Campbell[2,3]

[1]Affymetrix, Inc., Santa Clara, California; [2]Victorian Breast Cancer Research Consortium Cancer Genetics Laboratory, Peter MacCallum Cancer Centre, East Melbourne, Victoria, Australia; [3]Department of Pathology, University of Melbourne, Parkville, Victoria, Australia; and Departments of [4]Hematology/Oncology and [5]Regeneration Medicine for Hematopoiesis, University of Tokyo, Tokyo, Japan

Abstract

Formalin-fixed, paraffin-embedded (FFPE) material tends to yield degraded DNA and is thus suboptimal for use in many downstream applications. We describe an integrated analysis of genotype, loss of heterozygosity (LOH), and copy number for DNA derived from FFPE tissues using oligonucleotide microarrays containing over 500K single nucleotide polymorphisms. A prequalifying PCR test predicted the performance of FFPE DNA on the microarrays better than age of FFPE sample. Although genotyping efficiency and reliability were reduced for FFPE DNA when compared with fresh samples, closer examination revealed methods to improve performance at the expense of variable reduction in resolution. Important steps were also identified that enable equivalent copy number and LOH profiles from paired FFPE and fresh frozen tumor samples. In conclusion, we have shown that the Mapping 500K arrays can be used with FFPE-derived samples to produce genotype, copy number, and LOH predictions, and we provide guidelines and suggestions for application of these samples to this integrated system. [Cancer Res 2007;67(6):2544–51]

Introduction

The challenges associated with DNA derived from formalin-fixed, paraffin-embedded (FFPE) samples have prevented widespread application of FFPE DNA to many of the technologies available for high-quality DNA, although some options with lower genomic coverage are available (1–3). In this study, we show the feasibility and limitations of a genome-wide assessment of genotype, loss of heterozygosity (LOH), and copy number using FFPE DNA on the Affymetrix Mapping 500K array set, which includes the Mapping 250K Nsp Array and the Mapping 250K Sty Array (Santa Clara, CA). These arrays use a process termed whole-genome sampling analysis (WGSA; ref. 4), in which genomic DNA is digested and ligated to adaptors. A subset of digested fragments are then PCR amplified in a complexity reduction step before hybridization to the arrays. PCR proved to be the critical step when processing FFPE samples.

We compared several extraction methods to determine which protocol provides FFPE DNA most suitable for array analysis and found that a PCR-based assessment of DNA quality predicted the downstream performance of FFPE DNA samples better than age of FFPE sample. We identified a necessity for (*a*) *in silico* compensation against fragment size bias and (*b*) a fragment size filter during analysis of FFPE samples. We tested our new guidelines for FFPE DNA qualification and analysis on archival samples of various tissue types, storage times, and location sources. Quality of FFPE DNA varied but the methods outlined by this study enabled prediction of performance. These results show that FFPE DNA can be suitable for a combined study of genotype, LOH, and copy number on a whole-genome scale.

Materials and Methods

Sample selection and DNA extraction. Five primary endometrioid ovarian cancers were selected without screening for the initial portion of this study. For each sample set, normal lymphocytic DNA, fresh tumor tissue, and FFPE tissue were analyzed. Samples were collected between 1993 and 1999 as part of a larger study of ovarian cancer in women living in and around Southampton, United Kingdom (5). At the time of collection, DNA was extracted from blood samples and fresh tumor biopsies were snap frozen in liquid nitrogen. A portion of each frozen tumor biopsy was sectioned to assess the proportion of tumor. For samples 526T and 594T, microdissection was done (6) to obtain DNA with a >80% tumor component. DNA was extracted from the fresh frozen tissue using a salt chloroform method (7).

In 2002, a portion of each frozen tumor biopsy was formalin fixed and paraffin embedded as described previously (8), with all tumors fixed in 10% neutral buffered formalin for <24 h at room temperature. At the time of DNA extraction, the FFPE tumors had been embedded in paraffin blocks for 3 years. Five sections (10 μm) were deparaffinized twice in xylene (5 min) and rehydrated in 100%, 90%, and 70% ethanol (1 min each). The sections were stained with hematoxylin (4 min) and washed with water (1 min), acid alcohol (10 s), water (1 min), Scott's tap water (1 min), and water (1 min). The sections were then stained with eosin (3 min), rinsed with water (10 s), and dehydrated in 70%, 90%, and 100% ethanol (30 s each). Tumor cells were manually microdissected under a dissecting microscope as described previously (6) to obtain high-purity (>80%) tumor DNA. The tumor component for sample 594 was high enough that it was not stained or microdissected. DNA was extracted from the five endometrioid FFPE tissues using a modified Qiagen protocol (Valencia, CA; described below). Following DNA extraction from FFPE tissue, a salt precipitation DNA cleanup was done as described in the Affymetrix GeneChip Mapping Assay Manuals.

For the study of independent sample sets, DNA was extracted from FFPE tissue from 17 breast tumors and 8 colorectal tumors. FFPE blocks were collected from 11 pathology laboratories and ranged in age from 1 to 17 years. The formalin fixation and paraffin embedding protocols used for these tissues are not known but are likely to be quite varied. For breast

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

S. Jacobs and E.R. Thompson contributed equally to this work.

Conflict of Interest Statement: S. Jacobs, R. Pillai, and D.K. Bailey are employees of Affymetrix, Inc.

Requests for reprints: Sharoni Jacobs, Affymetrix, Inc., 3420 Central Expressway, Santa Clara, CA 95051. Phone: 408-731-5880; Fax: 408-481-0435; E-mail: sharoni_jacobs@affymetrix.com.

doi:10.1158/0008-5472.CAN-06-3597

tumors, 10 μm sections were deparaffinized, stained with H&E, and manually microdissected (described above). The colorectal tumors were not stained or microdissected due to their high tumor component. DNA was extracted from breast and colorectal tissues (described below), and as before, a salt precipitation DNA cleanup was done.

The collection and use of tissues for this study were approved by the appropriate institutional ethics committees.

Trial of DNA extraction methods for FFPE tissue. Five DNA extraction methods were trialed using whole 20 μm sections from three FFPE blocks. The methods that were compared were the MagneSil Genomic Fixed Tissue System (Promega,[6] Madison, WI), ChargeSwitch gDNA Micro Tissue kit (Invitrogen,[7] Carlsbad, CA), PureGene (Gentra Systems,[8] Minneapolis, MN), DNeasy Tissue kit (Qiagen[9]), and a phenol/chloroform extraction. With the exception of the DNeasy Tissue kit and phenol/chloroform, the extractions were done according to the manufacturer's instructions. The extractions done with the DNeasy Tissue kit and with phenol/chloroform both were modified to include an initial incubation at 95°C for 15 min followed by 5 min at room temperature as described previously (9), before being digested with proteinase K for 3 days at 56°C in a rotating oven with periodic mixing and fresh enzyme added each 24 h. A salt precipitation was done on DNA from all five extraction methods.

DNA quality assessment and preparation. The extracted DNA was quantified using UV spectroscopy at 260 nm. Random amplified polymorphic DNA-PCR (RAPD-PCR; ref. 10) was done to assess the quality of DNA and maximum fragment lengths as described previously using 50, 5, or 0.5 ng DNA (11). Qiagen HotStarTaq was used, with 0.4 units per reactions (Qiagen[9]). Products were visualized with ethidium bromide on a 3% gel.

Preparation and application of DNA to the mapping arrays. Matched fresh and FFPE samples were analyzed on the Affymetrix GeneChip Human Mapping 10K v2 Xba Array and 50K Xba Array and prepared using the Mapping 10K v2 Assay kit and the Mapping 100K Assay kit (Affymetrix)[10] The only exception to the manufacturer's protocol was that 10 cycles were added to the PCR cycling conditions for each FFPE sample.

Matched fresh tumor, FFPE tumor, and normal samples were assayed using the Mapping 250K Nsp Assay kit and the Mapping 250K Sty Assay kit[10] and hybridized to the 250K arrays. The 500K assay was done according to the manufacturer's protocol, beginning with 250 ng DNA. Ninety micrograms of PCR product were fragmented and labeled, using additional PCRs when necessary for FFPE breast and colorectal samples.

Data analysis. Genotype calls were produced using the dynamic model algorithm (12) by the Affymetrix GeneChip Genotyping Analysis Software version 4.0. A stringent *P* value cutoff threshold of 0.26 was used. Concordance was determined by calculating the number of single nucleotide polymorphisms (SNP) that gave the same call in both fresh frozen and FFPE DNA from the same tumor and dividing this number by the total number of SNPs that were called in both samples.

LOH predictions were produced using dChipSNP software (dChip2005_f4 version[11]; ref. 13). LOH values were inferred using the Hidden Markov Model and restricting to SNPs on fragment sizes ≤700 bp.

Copy number estimates for ovarian tumor samples using 500K data were determined by pairing tumor and matching normal samples in CNAG_v2.0.[12] Nonpaired, nonmatching references were used for copy number prediction of 10K and 50K data. Log 2 ratios were imported into Spotfire DecisionSite (Spotfire,[13] Somerville, MA) and the Affymetrix Integrated Genome Browser for visualization and comparison. Copy number estimates for breast and colon FFPE tumors were done using data from 48 HapMap samples (available online[10]) as a reference.

Estimated inter-SNP mean and median distances after exclusion of fragment sizes >700 bp were determined by first calculating the distance between all SNPs on each chromosome. Distances were then sorted per chromosome in descending order and the largest distances (representing centromeres) were removed for each chromosome, except for the acrocentric chromosomes 13 to 15 and 21 to 22.

Pearson (linear) correlations were calculated in Partek Genomics Suite (Partek,[14] St. Louis, MO).

Microsatellite analysis. Nine microsatellite markers were used to assess LOH at three loci: chromosome 1q (D1S2816, D1S413, and D1S1726), chromosome 7p (D7S691, D7S670, and D7S2506), and chromosome 14q (D14S1011, D14S258, and D14S1002). Regions were selected where array-based LOH analysis showed discordant LOH results for fresh and FFPE-derived DNA. The forward primer was labeled with a 5′-fluorescent dye (FAM or HEX). The samples were analyzed using a 3130 Genetic Analyzer (Applied Biosystems,[15] Foster City, CA) with POP7 polymer. An assessment of LOH was done using GeneMapper version 3.7. LOH was scored by calculation of the ratio of tumor DNA peaks (T1/T2) compared with that in the normal DNA to give a relative ratio (T1/T2)/(N1/N2). A ratio of 0 indicates complete allele loss and a ratio of 1 indicates no LOH. A ratio of <0.5 was scored as indicative of LOH.

Results

DNA extraction from FFPE tissue. Five DNA extraction methods (phenol/chloroform, Qiagen DNeasy Tissue kit, Invitrogen ChargeSwitch, Promega MagneSil, and Gentra PureGene) were tested on consecutive sections from different FFPE ovarian tumor biopsies. Phenol/chloroform and modified Qiagen protocols (see Materials and Methods) provided the highest DNA yield as determined by UV spectroscopy; these yields were 2.2 times more than the average yield from any of the other three extraction protocols (Fig. 1*A*). RAPD-PCR, which uses primers of 10 bps to produce a ladder of amplicons, was also done to assess both amplification efficiency and maximum product size for each extraction protocol (11). Compared with DNA extracted from fresh lymphocytes, the FFPE-derived DNA from all extraction methods yielded consistently smaller PCR fragments, with a maximum reliable size of ~ 800 bp (Fig. 1*A*). Phenol/chloroform and modified Qiagen extractions produced more intense and consistent PCR fragments across dilutions, suggesting that products were relatively free of contaminant inhibitors (Fig. 1*A*). DNA extracted with these two methods was processed through the PCR step of the Mapping 50K Xba Assay to further assess amplification efficiency. In this test, the modified Qiagen extraction provided a slightly higher PCR yield on average than the phenol/chloroform method (21.4 μg compared with 19.2 μg) and was therefore chosen for DNA extraction from FFPE tissues in this study.

Mapping 500K array performance. Five matched sets; each containing (*a*) nontumor, non-FFPE lymphocytic DNA, (*b*) fresh frozen ovarian tumor DNA, and (*c*) FFPE ovarian tumor DNA; were assessed for performance on the Mapping 500K arrays. All five FFPE samples had been stored for 3 years and provided average RAPD-PCR maximum amplicon sizes from 526 to 800 bp. During the PCR step of the Mapping assay, amplification products from all five FFPE tumors were concentrated <700 bp, a fragment size range that was reduced compared with non-FFPE samples (Fig. 1*B*). Decreased yield from the Mapping PCRs (Table 1) accompanied the decrease in amplicon size distributions. FFPE samples produced

[6] http://www.promega.com
[7] http://www.invitrogen.com
[8] http://www.gentra.com
[9] http://www.qiagen.com
[10] http://www.affymetrix.com
[11] http://www.dchip.org
[12] http://www.genome.umin.jp/
[13] http://www.spotfire.com

[14] http://www.partek.com
[15] http://www.appliedbiosystems.com