of tyrosine kinase domain are frequently found in infant ALL with *MLL* rearrangements and pediatric ALL with hyperdiploidy. Blood 2004;103:1085–8.

[13] Shimada A, Taki T, Tabuchi K, Tawa A, Horibe K, Tsuchida M, Hanada R, Tsukimoto I, Hayashi Y. *KIT* mutations, and not *FLT3* internal tandem duplication, are strongly associated with a poor prognosis in pediatric acute myeloid leukemia with t(8;21): a study of the Japanese Childhood AML Cooperative Study Group. Blood 2006;107:1806–9.

[14] Schnittger S, Kinkelin U, Schoch C, Heinecke A, Haase D, Haferlach T, Büchner T, Wörmann B, Hiddemann W, Griesinger F. Screening for MLL tandem duplication in 387 unselected patients with AML identify a prognostically unfavorable subset of AML. Leukemia 2000;14:796–804.

[15] Sheng XM, Kawamura M, Ohnishi H, Ida K, Hanada R, Kojima S, Kobayashi M, Bessho F, Yanagisawa M, Hayashi Y. Mutations of the *RAS* genes in childhood acute myeloid leukemia, myelodysplastic syndrome and juvenile chronic myelocytic leukemia. Leuk Res 1997;21:697–701.

[16] Lange BJ, Kobrinsky N, Barnard DR, Arthur DC, Buckley JD, Howells WB, Gold S, Sanders J, Neudorf S, Smith FO, Woods WG. Distinctive demography, biology, and outcome of acute myeloid leukemia and myelodysplastic syndrome in children with Down syndrome: Children's Cancer Group Studies 2861 and 2891. Blood 1998;91:608–15.

[17] Bunin N, Nowell PC, Belasco J, Shah N, Willoughby M, Farber PA, Lange B. Chromosome 7 abnormalities in children with Down syndrome and preleukemia. Cancer Genet Cytogenet 1991;54: 119–26.

[18] Meshinchi S, Stirewalt DL, Alonzo TA, Zhang Q, Sweetser DA, Woods WG, Bernstein ID, Arceci RJ, Radich JP. Activating mutations of RTK/ras signal transduction pathway in pediatric acute myeloid leukemia. Blood 2003;102:1474–9.

[19] Zwaan CM, Meshinchi S, Radich JP, Veerman AJ, Huismans DR, Munske L, Podleschny M, Hahlen K, Pieters R, Zimmermann M, Reinhardt D, Harbott J, Creutzig U, Kaspers GJ, Griesinger F. *FLT3* internal tandem duplication in 234 children with acute myeloid leukemia: prognostic significance and relation to cellular drug resistance. Blood 2003;102:2387–94.

[20] Liang DC, Shih LY, Hung IJ, Yang CP, Chen SH, Jaing TH, Liu HC, Wang LY, Chang WH. *FLT3*–TKD mutation in childhood acute myeloid leukemia. Leukemia 2003;17:883–6.

[21] Shimada A, Taki T, Tabuchi K, Taketani T, Hanada R, Tawa A, Tsuchida M, Horibe K, Tsukimoto I, Hayashi Y. Japanese Childhood AML Cooperative Study Group. Tandem duplications of *MLL* and *FLT3* are correlated with poor prognoses in pediatric acute myeloid leukemia: a study of the Japanese childhood AML Cooperative Study Group (Epub ahead of print: Aug. 30, 2007). Pediatr Blood Cancer 2007.

[22] Yokota S, Kiyoi H, Nakao M, Iwai T, Misawa S, Okuda T, Sonoda Y, Abe T, Kahsima K, Matsuo Y, Naoe T. Internal tandem duplication of the *FLT3* gene is preferentially seen in acute myeloid leukemia and myelodysplastic syndrome among various hematological malignancies: a study on a large series of patients and cell lines. Leukemia 1997;11:1605–9.

[23] Kottaridis PD, Gale RE, Langabeer SE, Frew ME, Bowen DT, Linch DC. Studies of *FLT3* mutations in paired presentation and relapse samples from patients with acute myeloid leukemia: implications for the role of *FLT3* mutations in leukemogenesis, minimal residual disease detection, and possible therapy with FLT3 inhibitors. Blood 2002;100:2393–8.

[24] Abu-Duhier FM, Goodeve AC, Wilson GA, Gari MA, Peake IR, Rees DC, Vandenberghe EA, Winship PR, Reilly JT. *FLT3* internal tandem duplication mutations in adult acute myeloid leukaemia define a high-risk group. Br J Haematol 2000;111:190–5.

[25] Yamamoto Y, Kiyoi H, Nakano Y, Suzuki R, Kodera Y, Miyawaki S, Asou N, Kuriyama K, Yagasaki F, Shimazaki C, Akiyama H, Saito K, Nishimura M, Motoji T, Shinagawa K, Takeshita A, Saito H, Ueda R, Ohno R, Naoe T. Activating mutation of D835 within the activation loop of FLT3 in human hematologic malignancies. Blood 2001;97: 2434–9.

[26] Thiede C, Steudel C, Mohr B, Schaich M, Schakel U, Platzbecker U, Wermke M, Bornhauser M, Ritter M, Neubauer A, Ehninger G, Illmer T. Analysis of FLT3-activating mutations in 979 patients with acute myelogenous leukemia: association with FAB subtypes and identification of subgroups with poor prognosis. Blood 2002;99: 4326–35.

[27] Bacher U, Haferlach T, Schoch C, Kern W, Schnittger S. Implications of *NRAS* mutations in AML: a study of 2502 patients. Blood 2006;107:3847–53.

[28] Kiyoi H, Naoe T, Nakano Y, Yokota S, Minami S, Miyawaki S, Asou N, Kuriyama K, Jinnai I, Shimazaki C, Akiyama H, Saito K, Oh H, Motoji T, Omoto E, Saito H, Ohno R, Ueda R. Prognostic implication of *FLT3* and *N-RAS* gene mutations in acute myeloid leukemia. Blood 1999;93:3074–80.

[29] Ravindranath Y, Abella E, Krischer JP, Wiley J, Inoue S, Harris M, Chauvenet A, Alvarado CS, Dubowy R, Ritchey AK, Land V, Steuber CP, Weinstein H. Acute myeloid leukemia (AML) in Down's syndrome is highly responsive to chemotherapy: experience on Pediatric Oncology Group AML Study 8498. Blood 1992;80:2210–4.

[30] Lie SO, Jonmundsson G, Mellander L, Siimes MA, Yssing M, Gustafsson G. A population-based study of 272 children with acute myeloid leukaemia treated on two consecutive protocols with different intensity: best outcome in girls, infants, and children with Down's syndrome. Nordic Society of Paediatric Haematology and Oncology (NOPHO). Br J Haematol 1996;94:82–8.

[31] Creutzig U, Ritter J, Vormoor J, Ludwig WD, Niemeyer C, Reinisch I, Stollmann-Gibbels B, Zimmermann M, Harbott J. Myelodysplasia and acute myelogenous leukemia in Down's syndrome: a report of 40 children of the AML-BFM Study Group. Leukemia 1996;10:1677–86.

[32] Kurosawa H, Tsuboi T, Shimaoka H, Okuya M, Nakajima D, Matsunaga T, Hagisawa S, Sato Y, Sugita K, Eguchi M. Long-term remission in an acute monoblastic leukemia patient with Down syndrome after cord blood transplantation [In Japanese]. Rinsho Ketsueki 2005;46:274–7.

[33] Yamaguchi Y, Fujii H, Kazama H, Iinuma K, Shinomiya N, Aoki T. Acute myeloblastic leukemia associated with trisomy 8 and translocation 8;21 in a child with Down syndrome. Cancer Genet Cytogenet 1997;97:32–4.

ARTICLE

Highly Sensitive Method for Genomewide Detection of Allelic Composition in Nonpaired, Primary Tumor Specimens by Use of Affymetrix Single-Nucleotide–Polymorphism Genotyping Microarrays

Go Yamamoto,* Yasuhito Nannya,* Motohiro Kato, Masashi Sanada, Ross L. Levine, Norihiko Kawamata, Akira Hangaishi, Mineo Kurokawa, Shigeru Chiba, D. Gary Gilliland, H. Phillip Koeffler, and Seishi Ogawa

Loss of heterozygosity (LOH), either with or without accompanying copy-number loss, is a cardinal feature of cancer genomes that is tightly linked to cancer development. However, detection of LOH is frequently hampered by the presence of normal cell components within tumor specimens and the limitation in availability of constitutive DNA. Here, we describe a simple but highly sensitive method for genomewide detection of allelic composition, based on the Affymetrix single-nucleotide–polymorphism genotyping microarray platform, without dependence on the availability of constitutive DNA. By sensing subtle distortions in allele-specific signals caused by allelic imbalance with the use of anonymous controls, sensitive detection of LOH is enabled with accurate determination of allele-specific copy numbers, even in the presence of up to 70%–80% normal cell contamination. The performance of the new algorithm, called "AsCNAR" (allele-specific copy-number analysis using anonymous references), was demonstrated by detecting the copy-number neutral LOH, or uniparental disomy (UPD), in a large number of acute leukemia samples. We next applied this technique to detection of UPD involving the 9p arm in myeloproliferative disorders (MPDs), which is tightly associated with a homozygous *JAK2* mutation. It revealed an unexpectedly high frequency of 9p UPD that otherwise would have been undetected and also disclosed the existence of multiple subpopulations having distinct 9p UPD within the same MPD specimen. In conclusion, AsCNAR should substantially improve our ability to dissect the complexity of cancer genomes and should contribute to our understanding of the genetic basis of human cancers.

Genomewide detection of loss of heterozygosity (LOH), as well as copy-number (CN) alterations in cancer genomes, has drawn recent attention in the field of cancer genetics,[1-3] because LOH has been closely related to the pathogenesis of cancers, in that it is a common mechanism for inactivation of tumor suppressor genes in Knudson's paradigm.[4] Moreover, the recent discovery of the activating Janus kinase 2 gene (*JAK2* [MIM *147796]) mutation that is tightly associated with the common 9p LOH with neutral CNs, or uniparental disomy (UPD), in myeloproliferative disorders (MPDs)[5-8] uncovered a new paradigm—that a dominant oncogenic mutation may be further potentiated by duplication of the mutant allele and/or exclusion of the wild-type allele—underscoring the importance of simultaneous CN detection with LOH analysis. On this point, Affymetrix GeneChip SNP-detection arrays, originally developed for large-scale SNP typing,[9] provide a powerful platform for both genomewide LOH analysis and CN detection.[10-12] On this platform, the use of large numbers of SNP-specific probes showing linear hybridization kinetics allows not only for high-resolution LOH analysis at ~2,500–150,000 heterozygous SNP loci but also for accurate determination of the CN state at each LOH region.[12-14] Unfortunately, however, the sensitivity of the currently available algorithm for LOH detection by use of SNP arrays may be greatly reduced when they are applied to primary tumor specimens that are frequently heterogeneous and contain significant normal cell components.

In this article, we describe a simple but highly sensitive method to detect allelic dosage (CNs) in primary tumor specimens on a GeneChip platform, with its validations, and some interesting applications to the analyses of primary hematological tumor samples. It does not require paired constitutive DNA of tumor specimens or a large set of normal reference samples but uses only a small number of anonymous controls for accurate determination of allele-specific CN (AsCN) even in the presence of significant

From the Departments of Hematology/Oncology (G.Y.; Y.N.; M.S.; A.H.; M. Kurokawa; S.O.), Regeneration Medicine for Hematopoiesis (S.O.), Pediatrics (M. Kato), and Cell Therapy and Transplantation Medicine (S.C.; S.O.), and The 21st Century Center of Excellence Program (Y.N.; M.S.; S.O.), Graduate School of Medicine, University of Tokyo, and Core Research for Evolutional Science and Technology, Japan Science and Technology Agency (S.O.), Tokyo; Division of Hematology, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston (R.L.L.; D.G.G.); and Hematology/Oncology, Cedars-Sinai Medical Center/University of California–Los Angeles School of Medicine, Los Angeles (N.K.; H.P.K.)

Received February 2, 2007; accepted for publication April 12, 2007; electronically published June 5, 2007.

Address for correspondence and reprints: Seishi Ogawa, The 21st Century COE Program, Department of Regeneration Medicine for Hematopoiesis, Department of Cell Therapy and Transplantation Medicine, Graduate School of Medicine, University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo 113-8655, Japan. E-mail: sogawa-tky@umin.ac.jp

* These two authors contributed equally to this work.

Am. J. Hum. Genet. 2007;81:114–126. © 2007 by The American Society of Human Genetics. All rights reserved. 0002-9297/2007/8101-0011$15.00
DOI: 10.1086/518809

proportions of normal cell components, thus enabling reliable genomewide detection of LOH in a wide variety of primary cancer specimens.

Material and Methods

Samples and Microarray Analysis

Genomic DNA extracted from a lung cancer cell line (NCI-H2171) was intentionally mixed with DNA from its paired lymphoblastoid cell line (LCL) (NCI-BL2171) to generate a dilution series, in which tumor contents started at 10% and increased by 10% up to 90%. The ratios of admixture were validated using measurements of a microsatellite (*D3S1279*) within a UPD region on chromosome 3 (data not shown). The nine mixed samples, together with nonmixed original DNAs (0% and 100% tumor contents), were analyzed with GeneChip 50K Xba SNP arrays (Affymetrix). Microarray data corresponding to 5%, 15%, 25%,..., and 95% tumor content were interpolated by linearly superposing two adjacent microarray data sets after adjusting the mean array signals of the two sets. Both cell lines were obtained from the American Type Culture Collection (ATCC). Genomic DNA was also extracted from 85 primary leukemia samples, including 39 acute myeloid leukemia (AML [MIM #601626]) samples and 46 acute lymphoblastic leukemia (ALL) samples, and was subjected to analysis with 50K Xba SNP arrays. Of the 85 samples, 34 were analyzed with their matched complete-remission bone marrow samples. DNA from 53 MPD samples—13 polycythemia vera (PV [MIM #263300]), 21 essential thrombocythemia (ET [MIM #187950]), and 19 idiopathic myelofibrosis (IMF [MIM #254450])—43 of which had been studied for *JAK2* mutations,[8] were also analyzed with 50K Xba SNP arrays. Microarray analyses were performed according to the manufacturer's protocol,[15] except with the use of LA *Taq* (Takara) for adaptor-mediated PCR. Also, DNA from 96 normal volunteers was used for the analysis. All clinical specimens were made anonymous and were incorporated into this study in accordance with the approval of the institutional review boards of the University of Tokyo and Harvard Medical School.

AsCN Analyses Using Anonymous Control Samples (AsCNAR)

SNP typing on the GeneChip platform uses two discrete sets of SNP-specific probes, which are arbitrarily but consistently named "type *A*" and "type *B*" SNPs, at every SNP locus, each consisting of an equal number of perfectly matched probes (PM_As or PM_Bs) and mismatched probes (MM_As or MM_Bs). For AsCN analysis, the sums of perfectly matched probes (PM_As or PM_Bs) for the *i*th SNP locus in the tumor (tum) sample and reference samples (ref1, ref2,..., ref*N*),

$$S_{A,i}^{\mathrm{tum}} = \sum \mathrm{PM}_{A,i}^{\mathrm{tum}}\ ,\ S_{B,i}^{\mathrm{tum}} = \sum \mathrm{PM}_{B,i}^{\mathrm{tum}}$$

and

$$S_{A,i}^{\mathrm{ref}l} = \sum \mathrm{PM}_{A,i}^{\mathrm{ref}l}\ ,\ S_{B,i}^{\mathrm{ref}l} = \sum \mathrm{PM}_{B,i}^{\mathrm{ref}l}\ ,\ (l = 1,2,3,\dots,N)\ ,$$

are compared separately at each SNP locus, according to the concordance of the SNP calls in the tumor sample (O_i^{tum}) and the SNP calls in a given reference sample ($O_i^{\mathrm{ref}l}$),

$$R_{A,i}^{\mathrm{ref}l} = \frac{S_{A,i}^{\mathrm{tum}}}{S_{A,i}^{\mathrm{ref}l}} \qquad \left(\text{for } O_i^{\mathrm{tum}} = O_i^{\mathrm{ref}l}\right), \qquad R_{B,i}^{\mathrm{ref}l} = \frac{S_{B,i}^{\mathrm{tum}}}{S_{B,i}^{\mathrm{ref}l}}$$

and the total CN ratio is calculated as follows:

$$R_{AB,i}^{\mathrm{ref}l} = \begin{cases} R_{A,i}^{\mathrm{ref}l} & \text{for } O_i^{\mathrm{tum}} = O_i^{\mathrm{ref}l} = AA \\ R_{B,i}^{\mathrm{ref}l} & \text{for } O_i^{\mathrm{tum}} = O_i^{\mathrm{ref}l} = BB \\ \frac{1}{2}\left(R_{A,i}^{\mathrm{ref}l} + R_{B,i}^{\mathrm{ref}l}\right) & \text{for } O_i^{\mathrm{tum}} = O_i^{\mathrm{ref}l} = AB \end{cases} \quad (l = 1,2,3,\dots,N)\ .$$

For CN estimations, however, $R_{AB,i}^{\mathrm{ref}l}$, $R_{A,i}^{\mathrm{ref}l}$, and $R_{B,i}^{\mathrm{ref}l}$ are biased by differences in mean array signals and different PCR conditions between the tumor sample and each reference sample and need to be compensated for these effects to obtain their adjusted values $\tilde{R}_{AB,i}^{\mathrm{ref}l}$, $\tilde{R}_{A,i}^{\mathrm{ref}l}$, and $\tilde{R}_{B,i}^{\mathrm{ref}l}$, respectively (appendix A).[16]

These values are next averaged over the references that have a concordant genotype for each SNP in a given set of references (*K*), and we obtain $\tilde{R}_{AB,i}^{K}$, $\tilde{R}_{A,i}^{K}$, and $\tilde{R}_{B,i}^{K}$. Note that $\tilde{R}_{A,i}^{K}$ and $\tilde{R}_{B,i}^{K}$ are calculated only for heterozygous SNPs in the tumor sample (see appendix A for more details).

A provisional total CN profile Λ_K is provided by

$$\Lambda_K = \{\tilde{R}_{AB,i}^{K}\}\ ,$$

and provisional AsCN profiles are obtained by

$$\Lambda_K^{\mathrm{large}} = \{\max(\tilde{R}_{A,i}^{K},\tilde{R}_{B,i}^{K})\}$$

$$\Lambda_K^{\mathrm{small}} = \{\min(\tilde{R}_{A,i}^{K},\tilde{R}_{B,i}^{K})\}\ .$$

These provisional analyses, however, assume that the tumor genome is diploid and has no gross CN alterations, when the coefficients are calculated in regressions. In the next step, the regressions are iteratively performed using a diploid region that is truly or is expected to be diploid, to determine the coefficients on the basis of the provisional total CN, and then the CNs are recalculated.

Finally, the optimized set of references is selected that minimizes the SD of total CN at the diploid region by stepwise reference selection, as described in appendix A.

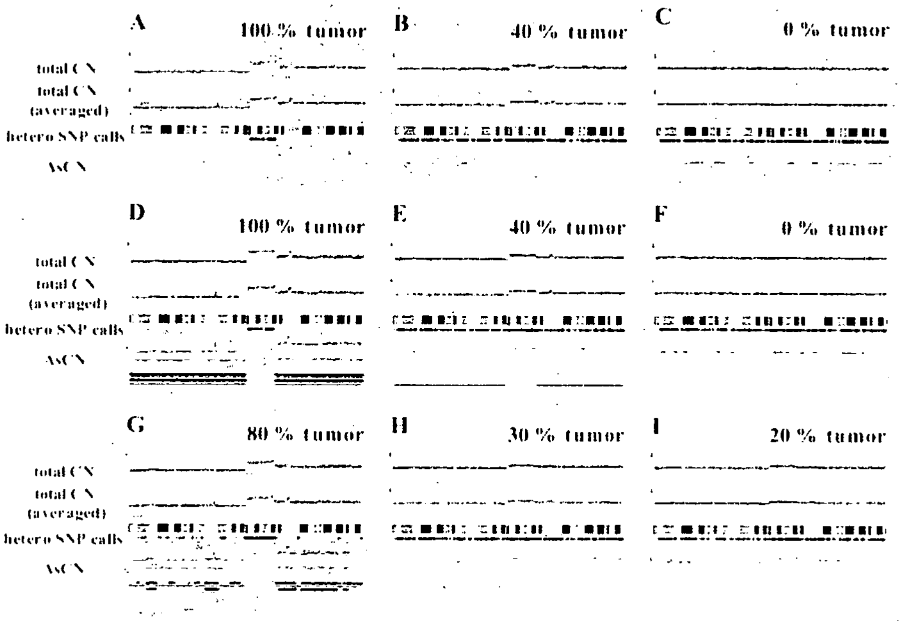

Figure 1. AsCN analysis with or without paired DNA. DNA from a lung cancer cell line (NCI-H2171) was mixed with DNA from an LCL (NCI-BL2171) established from the same patient at the indicated percentages and was analyzed with GeneChip 50K Xba SNP arrays. AsCNs, as well as total CNs, were analyzed using either the paired reference sample (NCI-BL2171) (*upper panels, A–C*) or samples from unrelated individuals simultaneously processed with the tumor samples (*middle and lower panels, D–I*). On each panel, the upper two graphs represent total CNs and their moving averages for the adjacent 10 SNPs, whereas moving averages of AsCNs for the adjacent 10 SNPs are shown below (*red and green lines*). Green and pink bars in the middle are heterozygous (hetero) calls and discordant SNP calls between the tumor and its paired reference, respectively. At the bottom of each panel, LOH regions inferred from AsCNAR (*orange*), SNP call–based LOH inference of CNAG (*blue*), dChip (*purple*), and PLASQ (*light green*) are depicted. Asterisks (*) indicate the loci at which total CNs were confirmed by FISH analysis (data not shown). The calibrations of CN graphs are linearly adjusted so that the mean CNs of null and single alleles should be 0 and 1, respectively.

Allele-specific analysis using a constitutive reference, refSelf, is provided by

$$A^{\mathrm{large}} = \{\max(\tilde{R}^{\mathrm{refSelf}}_{A,i},\tilde{R}^{\mathrm{refSelf}}_{B,i})\}$$

and

$$A^{\mathrm{small}} = \{\min(\tilde{R}^{\mathrm{refSelf}}_{A,i},\tilde{R}^{\mathrm{refSelf}}_{B,i})\} \;.$$

Computational details of AsCNAR are provided in appendix A.

Comparison with Other Algorithms

dChip[17] and PLASQ[18] were downloaded from their sites, and the identical microarray data were analyzed using these programs. Since PLASQ requires both Xba and Hind array data, microarray data of mixed tumor contents for Hind arrays were simulated by linearly superimposing the tumor cell line (NCI-H2171) and LCL (NCI-BL2171) data at indicated proportions.

Statistical Analysis

Significance of the presence of allelic imbalance (AI) in a given region, Γ, called as having AI by the hidden Markov model (HMM), was statistically tested by calculating *t* statistics for the difference in AsCNs, $|\log_2\tilde{R}^{k}_{A,i} - \log_2\tilde{R}^{k}_{B,i}|$, between Γ and a normal diploid region, where the tests were unilateral. Significance between the numbers of UPDs detected by the SNP call–based method and by AsCNAR was tested by one-tailed binominal tests. *P* values for AI detection by allele-specific PCR were calculated by one-tailed *t* tests, comparing triplicates of the target sample and triplicates of five normal samples that have heterozygous alleles in the SNP.

Detection of the JAK2 *Mutation and Measurements of Relative Allele Doses*

The *JAK2* V617F mutation was examined by a restriction enzyme–based analysis, in which PCR-amplified *JAK2* exon 12 fragments were digested with *Bsa*XI, and the presence of the undigested fragment was examined by gel electrophoresis.[5] Relative allele dose between wild-type and mutated *JAK2* was determined by measuring allele-specific PCR products for wild-type and mutated *JAK2* alleles by

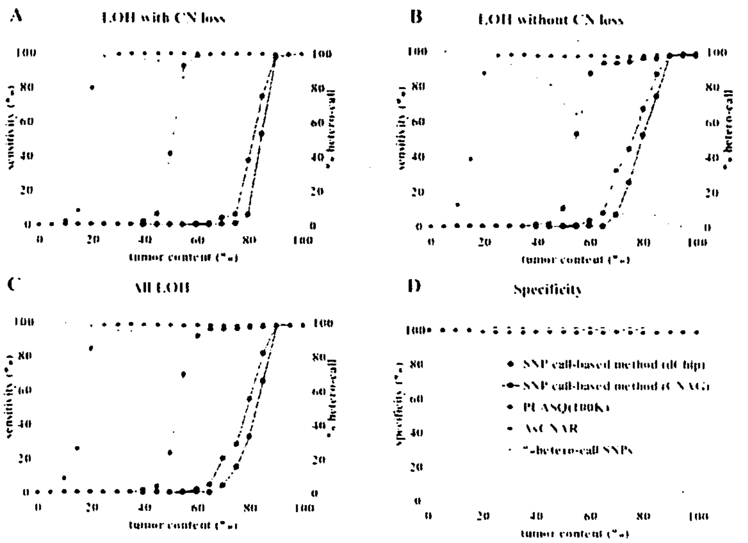

Figure 2. Sensitivity and specificity of LOH detection for intentionally mixed tumor samples. Sensitivity of detection of LOH with or without CN loss (*A* and *B*) in different algorithms were compared using a mixture of the tumor sample (NCI-H2171) and the paired LCL sample (NCI-BL2171). The results for all LOH regions are shown in panel C, and the specificities of LOH detection are depicted in panel D. For precise estimation of sensitivity and specificity, we defined the SNPs truly positive and negative for LOH as follows. The tumor sample and the paired LCL sample were genotyped on the array three times independently, and we considered only SNPs that showed the identical genotype in the three experiments. SNPs that were heterozygous in the paired LCL sample and were homozygous in the tumor sample were considered to be truly positive for LOH, and SNPs that were heterozygous both in the paired LCL sample and in the tumor sample were considered to be truly negative. Proportions of heterozygous SNP calls (%hetero-call) that remained in LOH regions of each sample are also shown in panels A–C.

capillary electrophoresis by use of the 3100 Genetic Analyzer (Applied Biosystems), as described in the literature.[19] Likewise, the fraction of tumor components having 9p and other UPDs was measured by either allele-specific PCR or STR PCR,[7,19] by use of the primers provided in appendix B [online only]. The percentage of UPD-positive cells (%UPD(+)) was also estimated as the mean difference of AsCNs for heterozygous SNPs within the UPD region divided by that for homozygous SNPs within an arbitrary selected normal region:

$$\%\mathrm{UPD}(+) = \frac{E(|\bar{R}^{K}_{A,i} - \bar{R}^{K}_{B,i}|_{i\text{: hetero SNPs in UPD region}})}{E(|\bar{R}^{K}_{A,j} - \bar{R}^{K}_{B,j}|_{j\text{: homo SNPs with normal CN}})} ,$$

where AsCNs for the denominator were calculated as if the homozygous SNPs were heterozygous. However, in those samples with a high percentage of UPD-positive components, the heterozygous SNP rate in the UPD region decreased. For such regions, we calculated the percentage of UPD-positive cells by randomly selecting 30% (the mean heterozygous SNP call rate for this array) of all the SNPs therein and by assuming that they were heterozygous SNPs. Cellular composition of *JAK2* wild-type (wt) and mutant (mt) homozygotes (wt/wt and mt/mt) and heterozygotes (wt/mt) in each MPD specimen was estimated assuming that all UPD components are homozygous for the *JAK2* mutation. The fractions of the wt/mt heterozygotes in cases with a 9p gain were estimated assuming that the duplicated 9p alleles had the *JAK2* mutation. Throughout the calculations, small negative values for wt/mt were disregarded.

FISH

FISH analysis was performed according to the previously published method, to confirm the absolute total CNs in NCI-H2171.[20] The genomic probes were generated by whole-genome amplification of FISH-confirmed RP11 BAC clones 169N13 (3q13; CN = 2), 227F7 (8q24; CN = 2), 196H14 (12q14; CN = 2), 25E13 (13q33; CN = 2), 84E24 (17q24; CN = 2), 12C9 (19q13; CN = 2), 153K19 (3q13; CN = 3), 94D19 (3p14; CN = 1), 80P10 (8q22; CN = 1), and 64C21 (13q12-13; CN = 1), which were obtained from the BACPAC Resources Center at the Children's Hospital Oakland Research Institute in Oakland, California.

Results

SNP Call–Based Genomewide LOH Detection by Use of SNP Arrays

When a pure tumor sample is analyzed with a paired constitutive reference on a GeneChip Xba 50K array, LOH is easily detected as homozygous SNP loci in the tumor spec-

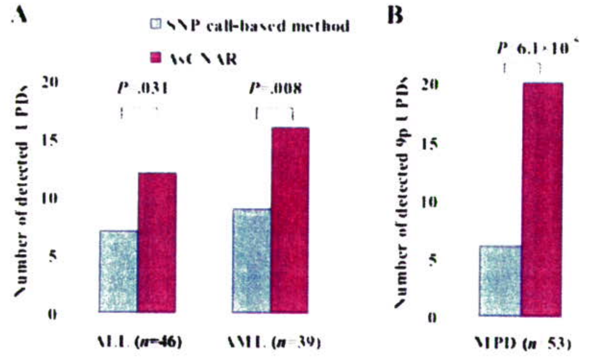

Figure 3. The number of UPD regions for acute leukemia and MPD samples detected by either the SNP call–based method or AsCNAR. The number of UPD regions for ALL and AML samples detected by the SNP call–based method or by AsCNAR is shown in panel A, and the number of 9p UPDs for MPD samples detected by the two methods is shown in panel B. Some samples have more than one UPD region. Details of UPD regions are given in table 1. Significance between the numbers of UPDs detected by the SNP call–based method and by the AsCNAR method was tested by one-tailed binomial tests.

imen that are heterozygous in the constitutive DNA (fig. 1*A*, pink bars). In addition, given a large number of SNPs to be genotyped, the presence of LOH is also inferred from the grossly decreased heterozygous SNP calls, even in the absence of a paired reference (fig. 1*D*). The accuracy of the LOH inference would depend partly on the algorithm used but more strongly on the tumor content of the specimens. Thus, our SNP call–based LOH inference algorithm in CNAG (appendix C), as well as that of dChip,[17] show almost 100% sensitivity and specificity for pure tumor specimens. But, as the tumor content decreases, the LOH detection rate steeply declines (fig. 1*G*), and, with <50% tumor cells, no LOH can be detected, even when complete genotype information for both tumor and paired constitutive DNA is obtained (fig. 1*B*, 1*E*, 1*H*, and 1*I*).

LOH Detection Based on AsCN Analysis

On the other hand, the capability of allele-specific measurements of CN alterations in cancer genomes is an excellent feature of the SNP array–based CN-detection system that uses a large number of SNP-specific probe sets.[16,18,21] When constitutive DNA is used as a reference, AsCN analysis is accomplished by separately comparing the SNP-specific array signals from the two parental alleles at the heterozygous SNP loci in the constitutive genomic DNA.[16] It determines not only the total CN changes but also the alterations of allelic compositions in cancer genomes, which are captured as the split lines in the two AsCN graphs (fig. 1*A* and 1*B*). In this mode of analysis, the presence of LOH can be detected as loss of one parental allele, even in specimens showing almost no discordant calls (fig. 1*B*).

AsCNAR

The previous method for AsCN analysis, however, essentially depends on the availability of constitutive DNA, since AsCNs are calculated only at the heterozygous SNP loci in constitutive DNA.[16] Alternatively, allele-specific signals can be compared with those in anonymous references on the basis of the heterozygous SNP calls in the tumor specimen. In the latter case, the concordance of heterozygous SNP calls between the tumor and the unrelated sample is expected to be only 37% with a single reference. However, the use of multiple references overcomes the low concordance rate with a single reference, and the expected overall concordance rate for heterozygous SNPs and for all SNPs increases to 86% and 92%, respectively, with five unrelated references (appendix D [online only]). Thus, for AsCNAR, allele-specific signal ratios are calculated at all the concordant heterozygous SNP loci for individual references, and then the signal ratios for the identical SNPs are averaged across different references over the entire genome. For the analysis of total CNs, all the concordant SNPs, both homozygous and heterozygous, are included in the calculations, and the two allele-specific signal ratios for heterozygous SNP loci are summed together. Since AsCNAR computes AsCNs only for heterozygous SNP loci in tumors, difficulty may arise on analysis of an LOH region in highly pure tumor samples, in which little or no heterozygous SNP calls are expected. However, as shown above, such LOH regions can be easily detected by the SNP call–based algorithm, where AsCNAR is formally calculated assuming all the SNPs therein are heterozygous. Thus, the AsCNAR provides an essentially equivalent result to that from AsCN analysis using constitutional DNA, with similar sensitivity in detecting AI and LOH (compare fig. 1*A* with 1*D* and 1*B* with 1*E*).

As expected from its principle, AsCNAR is more robust in the presence of normal cell contaminations than are SNP call–based algorithms. To evaluate this quantitatively, we analyzed tumor DNA that was intentionally mixed with its paired normal DNA at varying ratios in 50K Xba SNP arrays, and the array data were analyzed with AsCNAR. To preclude subjectivity, LOH regions were detected by an HMM-based algorithm, which evaluates difference in AsCNs in both parental alleles (appendix E).[22] As the tumor content decreases, the SNP call–based LOH inference fails to detect LOH because of the appearance of heterozygous SNP calls from the contaminated normal cell component (fig. 1*E* and 1*G*–1*I*), but these heterozygous SNP calls, in turn, make AsCNAR operate effectively.

Table 1. CN-Neutral LOH in Primary Acute Leukemia

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

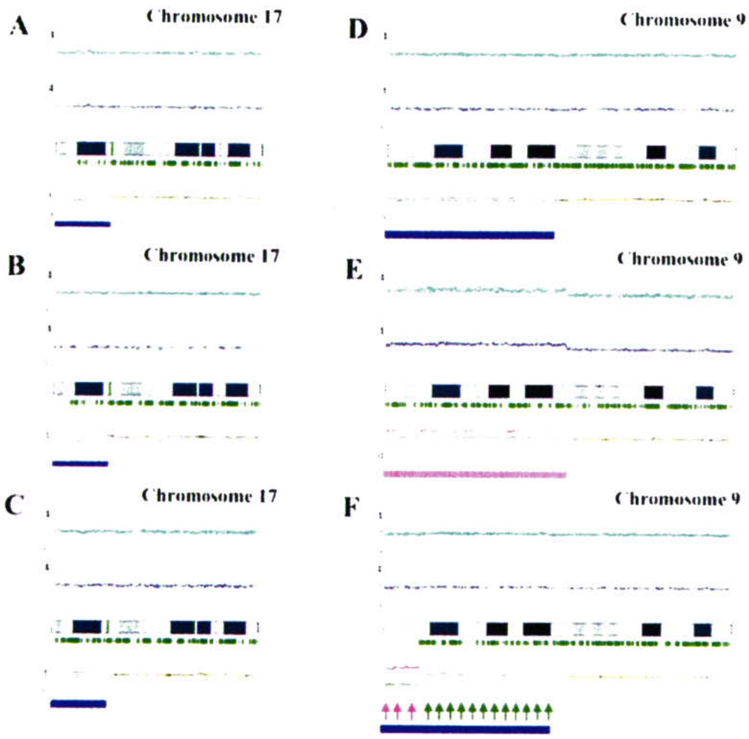

Figure 4. Detection of AI in samples of primary AML and MPD. AsCN analyses disclosed the presence of a small population with 17p UPD in a primary AML specimen (W150673) (93% blasts in microscopic examination) with either a paired sample (*A*) or anonymous reference samples (*B*). The difference of the mean CNs of the two parental alleles is statistically different between panels A (0.38) and B (0.55) ($P < .0001$, by *t* test), which is explained by the residual tumor component within the bone marrow sample in complete remission (1% blast) used as a paired reference (W150673CR) (*C*). AI in the 9p arm was also sensitively detected in *JAK2* mutation–positive MPD cases. UPD may be carried only by a very small population (~20% estimated from the mean deviation of AsCNs in 9p) (IMF_10) (*D*), or by two discrete populations within the same case (PV_06), as indicated by two-phased dissociation of AsCN graphs (*pink and green arrows*) (*F*). AI in 9p is mainly caused by UPD but may be caused by gains of one parental allele without loss of the other allele (*E*), both of which are not discriminated by conventional allele measurements. Blue and pink bars are UPD and AI calls, respectively, from the HMM-based LOH detection algorithm. Other features are identical to those indicated in figure 1.

In fact, this algorithm precisely identifies known LOH regions, as well as regions with AI, in intentionally mixed tumor samples containing as little as 20% (for LOH without CN loss) to 25% (LOH with CN loss) tumor contents (fig. 2*A*–2*C*). Note that this large gain in sensitivity is obtained without the expense of specificity, which is very close to 100%, as observed with other algorithms (fig. 2*D*). In AsCNAR, small regions of AI (<1 million bases in length) are difficult to detect in samples contaminated with normal cells. However, such regions are also difficult to detect using other algorithms (data not shown).

Identification of UPD in Primary Tumor Samples

To examine further the strength of the newly developed algorithms for AsCN and LOH detection, we explored UPD regions in 85 primary acute leukemia samples, including 39 AML and 46 ALL samples, on GeneChip 50K Xba SNP arrays, since recent reports identified frequent (~20%) occurrence of this abnormality in AML.[14,24] In the SNP call–based LOH inference algorithm, 16 UPD regions were identified in 14 cases, 8 (20.5%) AML and 6 (13.0%) ALL. However, the frequencies were almost doubled with the AsCNAR algorithm; a total of 28 UPD loci were identified in 25 cases, including 14 (35.9%) AML and 11 (23.9%) ALL (fig. 3*A* and table 1). In 5 of the 25 UPD-positive cases, a matched remission sample was available for AsCN analysis, which provided essentially the same results as AsCNAR, except for one relapsed AML case (W150673). In the latter case, a discrepancy in AsCN shifts in 17p UPD occurred between AsCN analysis with and without a constitutive reference, with more CN shift detected with anonymous references (fig. 4*A* and 4*B*). The discrepancy was, however, explained by the unexpected detection of a subtle UPD change in 17p in the reference sample by

Table 2. AI of 9p in *JAK2* Mutation-Positive MPDs

Case	9p Status by AsCNAR			Detection by SNP Call-Based Method[a]	% *JAK2* Mutation[b]	Allele-Specific PCR[c]		
	Type	Break Point[d]	%UPD[e]			SNP	%UPD[f]	*P*[g]
PV_02	Gain	42.9	99	NA	63	*rs2009991*	84	.004
PV_03	Gain	Whole	60	NA	39	*rs10511431*	63	.008
PV_04	UPD	37.0	93	D	95	5Homo	5Homo	5Homo
PV_08	UPD	34.2	91	D	93	5Homo	5Homo	5Homo
PV_07	UPD	23.8	88	D	90	5Homo	5Homo	5Homo
PV_06	UPD[h]	7.1/35.3	83	D	93	5Homo	5Homo	5Homo
PV_11	UPD	31.2	68	D	76	5Homo	5Homo	5Homo
PV_13	UPD	28.1	66	ND	48	*rs1416582*	64	.001
PV_01	UPD	20.9	56	ND	62	*rs10511431*	49	.007
PV_09	UPD	30.8	38	ND	30	*rs10491558*	32	.020
PV_05	UPD	23.5	32	ND	33	*rs1374172*	31	.010
IMF_04	UPD	33.8	79	D	90	5Homo	5Homo	5Homo
IMF_05	UPD	37.0	58	ND	57	*rs1416582*	49	.004
IMF_07	UPD	20.3	52	ND	50	*rs1416582*	57	.005
IMF_12	UPD[h]	26.8/42.9	52	ND	66	5Homo	5Homo	5Homo
IMF_14	UPD[h]	22.8/33.8	45	ND	56	*rs1374172*	35	.015
IMF_19	UPD	34.4	26	ND	43	*rs10511431*	33	.017
IMF_10	UPD	34.6	21	ND	36	*rs1374172*	21	.049
IMF_15	UPD	33.8	21	ND	17	*rs10511431*	20	.084
IMF_06	UPD	35.3	17	ND	28	*rs1374172*	20	.048
IMF_16	(−)	NA	NA	NA	37	NA	NA	NA
ET_12	Gain	Whole	42	NA	27	*rs2009991*	36	.046
ET_14	UPD	42.9	63	ND	45	*rs1374172*	54	.006
ET_01	UPD	35.4	19	ND	59	*rs10511431*	33	.017
ET_05	(−)	NA	NA	NA	23	NA	NA	NA
ET_08	(−)	NA	NA	NA	42	NA	NA	NA
ET_09	(−)	NA	NA	NA	34	NA	NA	NA
ET_10	(−)	NA	NA	NA	16	NA	NA	NA
ET_15	(−)	NA	NA	NA	27	NA	NA	NA
ET_18	(−)	NA	NA	NA	17	NA	NA	NA
ET_19	(−)	NA	NA	NA	27	NA	NA	NA
ET_21	(−)	NA	NA	NA	55	NA	NA	NA

NOTE.—NA = not applied; (−) = neither UPD nor gain of 9p was detected by AsCNAR analysis.

[a] D = UPD was detected by SNP call-based method; ND = not detected.

[b] Percentage of *JAK2* mutant alleles, as measured by allele-specific PCR.

[c] 5Homo = all five tested SNPs were homozygous.

[d] Position of the break point from the p-telomeric end (values are in Mb). The location of *JAK2* corresponds to 5 Mb.

[e] Percentage of tumor cell populations with either UPD or gain of 9p, as determined by AsCNAR analysis.

[f] Percentage of tumor cell populations with either UPD or gain of 9p, as determined by the allele-specific PCR.

[g] *P* values were derived from one-tailed *t* tests comparing triplicate analyses of the target sample and triplicate analyses of five normal samples.

[h] Two UPD-positive populations exist.

AsCNAR ($P<.0001$, by *t* test) (fig. 4C), which offset the CN shift in the relapsed sample, although it was morphologically and cytogenetically diagnosed as in complete remission.

Analysis of 9p UPD in MPDs

Another interesting application of the AsCNAR is the analysis of allelic status in the 9p arm among patients with MPD, which includes PV, ET, and IMF. According to past reports, ~10% (in ET) to ~40% (in PV) of MPD cases with the activating *JAK2* mutation (V617F) show evidence of clonal evolution of dominant progeny that carry the homozygous *JAK2* mutation caused by 9p UPD.[5,7,8] In our series that included 53 MPD cases, the *JAK2* mutation was detected in 32 (60%), of which 13 (41%) showed >50% mutant allele by allele measurement with the use of allele-specific PCR, and thus were judged to have one or more populations carrying homozygous *JAK2* mutations (table 2). This frequency is comparable to that reported elsewhere.[8] However, when the same specimens were analyzed with 50K Xba SNP arrays by use of the AsCNAR algorithm, 20 of the 32 *JAK2* mutation–positive cases were demonstrated to have minor UPD subpopulations (table 2 and fig. 3*B*), in which as little as 17% of UPD-positive populations were sensitively detected (fig. 4*D*). In fact, these minor (<50%) UPD-positive populations in these 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[20] 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.

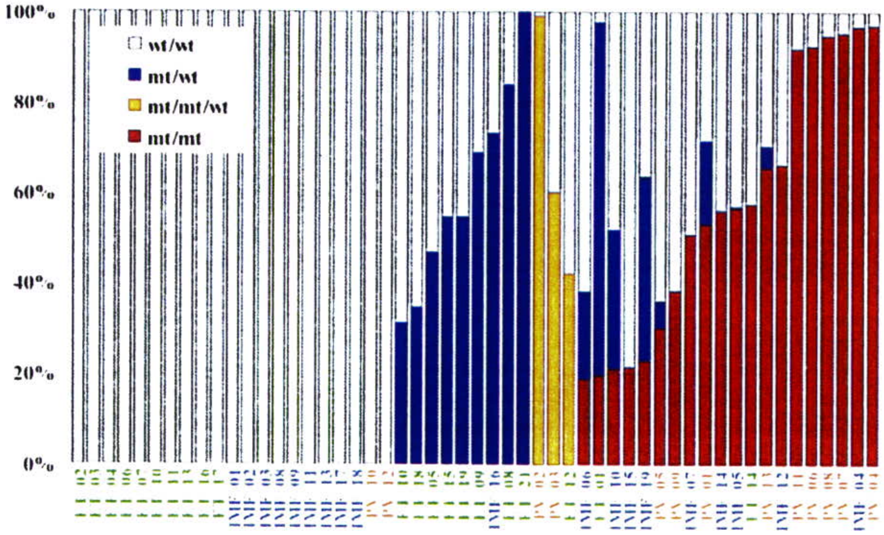

Figure 5. Estimation of tumor populations carrying 9p UPD and the *JAK2* mutation in MPD samples. The populations of 9p UPD–positive components in the 53 MPD cases were estimated by calculation of the mean difference of AsCNs within the UPD regions. Heterozygous (*blue bars*) or homozygous (*red bars*) *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.

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.

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.*

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, $\log_2 R^{\mathrm{ref}l}_{AB,i}$ 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

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

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 $R^{\mathrm{ref}l}_{A,i}$ and $R^{\mathrm{ref}l}_{B,i}$, respectively, can be regressed by the same parameters, as

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

and

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

and the corrected total CN ratio is

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

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:

$$C_i^K = \{\mathrm{ref}l \,|\, O_i^{\mathrm{tum}} = O_i^{\mathrm{ref}l}, \mathrm{ref}l \in K\}$$

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

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

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

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

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

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

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

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,

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

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,…,$s$, $s+1$,…, 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 *i*th and *j*th SNPs ($i \le j$) and for the complete set of observed SNP calls therein, $O(\Omega_{i,j})$, consider the log likelihood ratio

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

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 *k*th SNP is heterozygous (h_k), depending on the observed $k-1$th SNP call, for partially taking the effect of linkage disequilibrium into account:

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

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

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 *k*th SNP call, with 1 for a homozygous call and 0 for a heterozygous call. For each chromosome, a set of regions, $\Omega_{I_n,J_n}(J_{n-1} < I_n \le J_n, J_0 = 0)$ ($n = 1,2,3,\ldots$), 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 $Z(\Omega_{i,i}) \le 0$ for $J_{n-1} < \forall i < I_n$ (fig. C1). Identify the SNP S_{J^+}, such that $Z(\Omega_{I_n,j}) > 0$ for $I_n \le \forall j \le J^+$ and $Z(\Omega_{I_n,J^{+}+1}) \le 0$, or that S_{J^+} is the end of the chromosome (fig. C1). Then, put J_n as $\arg\max_j Z(\Omega_{I_n,j})(I_n \le j \le J^+)$ (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 *i*th SNP locus (*Si*) are

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

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

and

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

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

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

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

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

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

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

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

and

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

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

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

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

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 $\geq$1.7], we can expect satisfactory power with currently available large-scale genotyping platforms using realistic sample size ($\sim$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(χ^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(χ^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(χ^2) is given as

$$\varphi_N(\chi^2) = \frac{d}{d\chi^2}\{\phi(\chi^2)^N\}.$$

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(χ^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(χ^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(χ^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(χ^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, $\varphi_{Ns}(\chi^2)$, 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, $\varphi_{Ns}(\chi^2)$'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 Ne makers, $\varphi_{Ne}(\chi^2)$ (see Methods). Irrespective of marker sets, fitting is finely

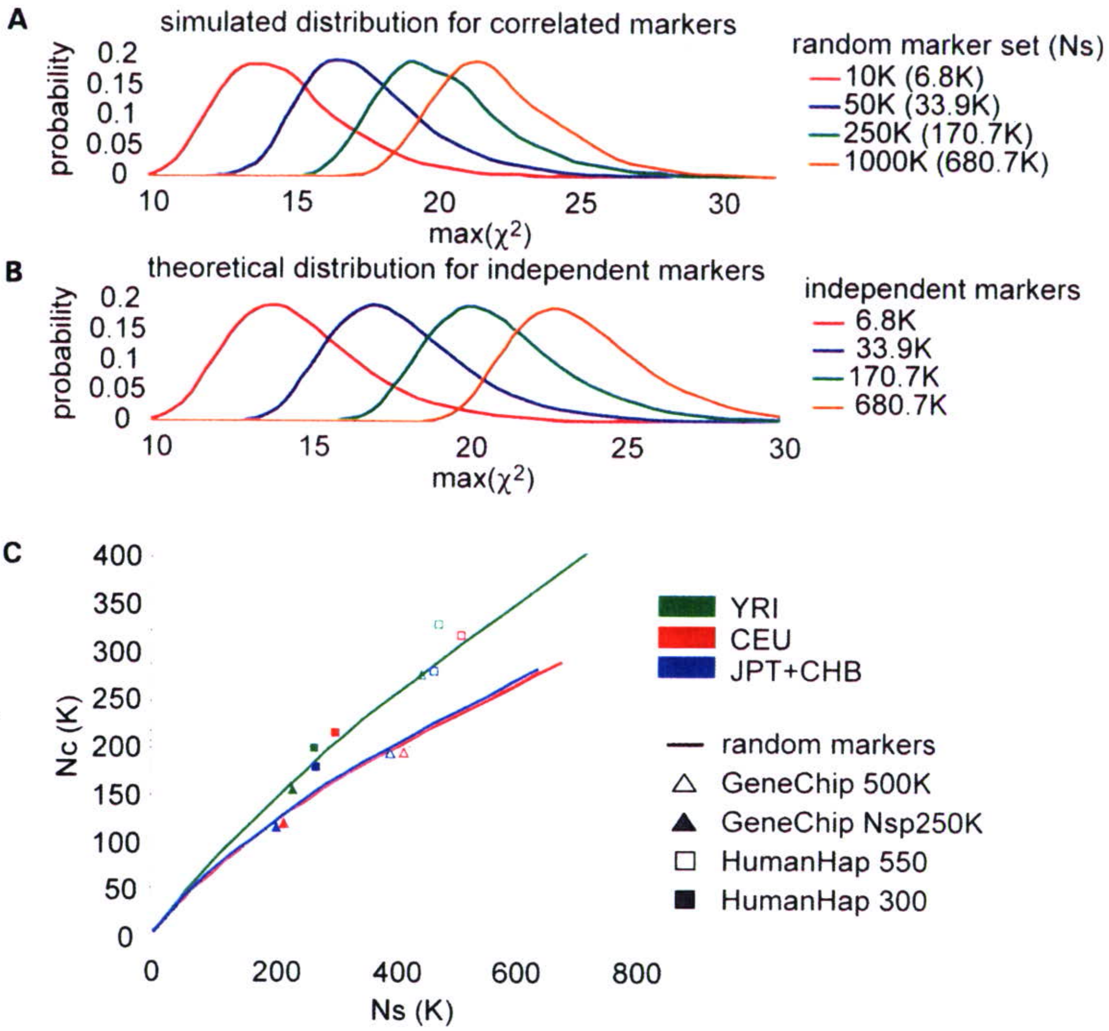

Figure 1. Null distributions of max(χ^2) and the effective number of independent SNPs (Ne) for various marker sets. Distributions of max(χ^2) for all null SNPs (null distributions) were simulated for increasing numbers of randomly selected SNP markers in the CEU panel. Ten thousand null panels, each 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, $\varphi_{N_s}(\chi^2)$, were calculated assuming all Ns segregating SNPs therein are independent (**B**). The effective numbers of hypothetical independent SNPs (Ne) were estimated by fitting simulated null distributions to theoretical ones for Ne independent SNPs, $\varphi_{N_e}(\chi^2)$, 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/Ne point in the χ^2 distribution (d.f. = 1) (Bonferroni) (Table 1). In this formulation, it is reasonable to regard Ne 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 Ne/Ns ratio.

Ne 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), Ne 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] Ne values calculated for combined SNPs from seven regions.
[b] Ne of ENCODE regions are extrapolated to the entire genome.
[c] Ne 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 $\geqq 0.05$) 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 $\leqq 1.7$) 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 $\geqq 2.0$) and large sample sizes ($\geqq 1500$/arm), the power tends to saturate as the number of randomly selected SNPs increases ($\geqq 250$K), 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 $\geqq 250$K random markers or equivalent tag SNPs are used. In contrast, the effect of sample numbers is predominant. To detect weak genetic effects (GRR $\leqq 1.3$), 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 (Ne/Ne$^{(c)}$), 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.

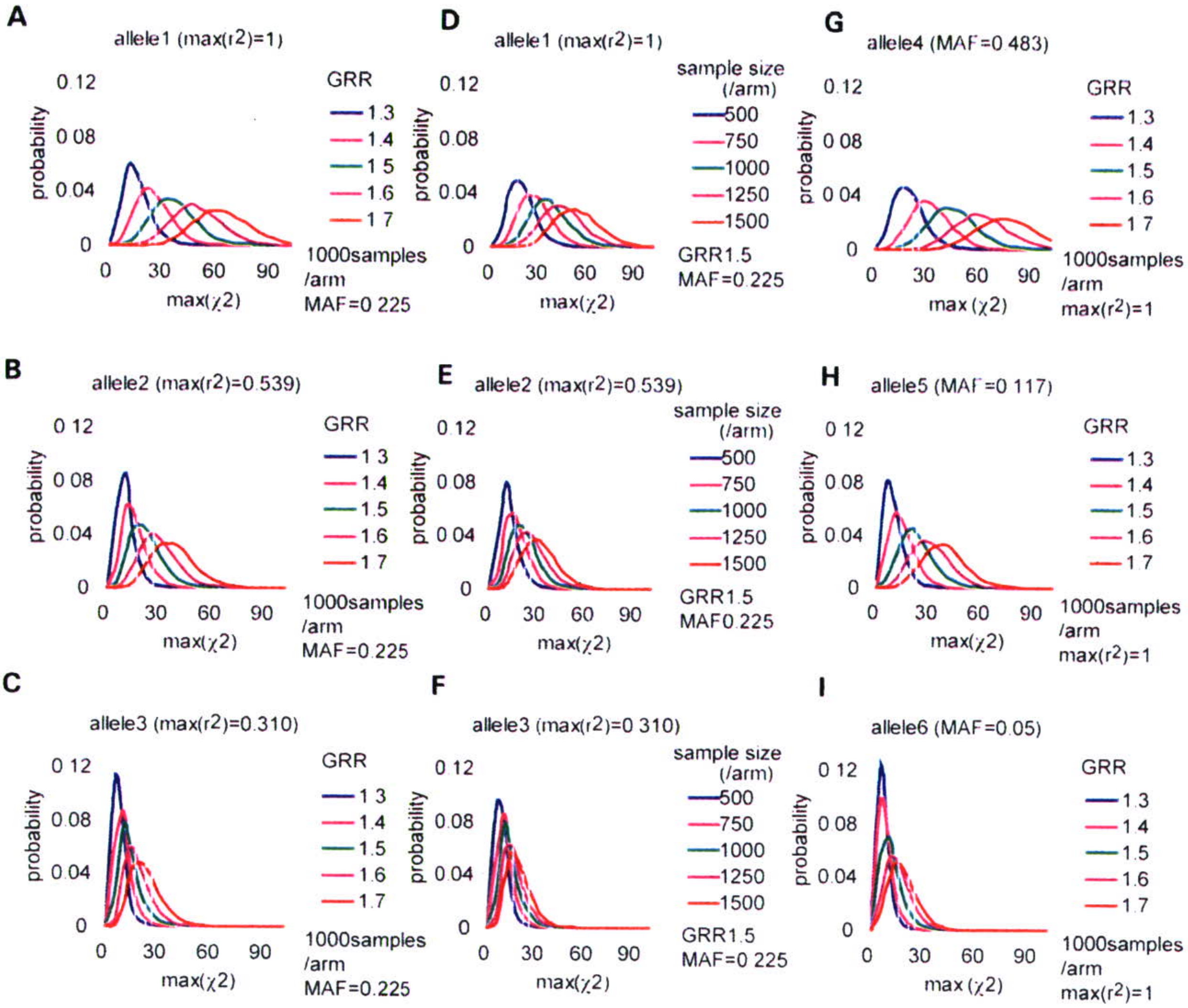

Figure 2. Enhancement of causal distributions by various parameters. Combined effects of LD [in max(r^2)] and effect size (in GRR) on causal distributions under constant sample size (1000/arm) and MAF value (0.225) (**A**–**C**), LD and sample size under constant effect size (GRR = 1.5) and MAF value (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 Ne values of 6K, 196K, and 1023K (Ne^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