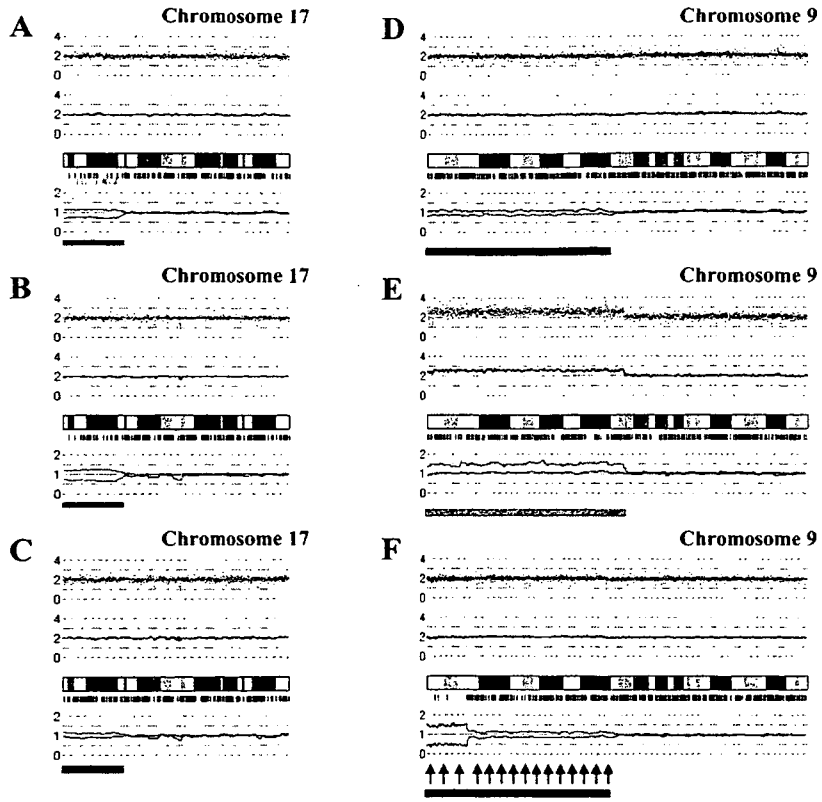

Figure 4. Detection of AI in samples of primary AML and MPD. AsCN analyses disclosed the presence of a small population with 17p UPD in a primary AML specimen (W150673) (93% blasts in microscopic examination) with either a paired sample (*A*) or anonymous reference samples (*B*). The difference of the 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.[23,24] 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. 4*C*), 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[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.

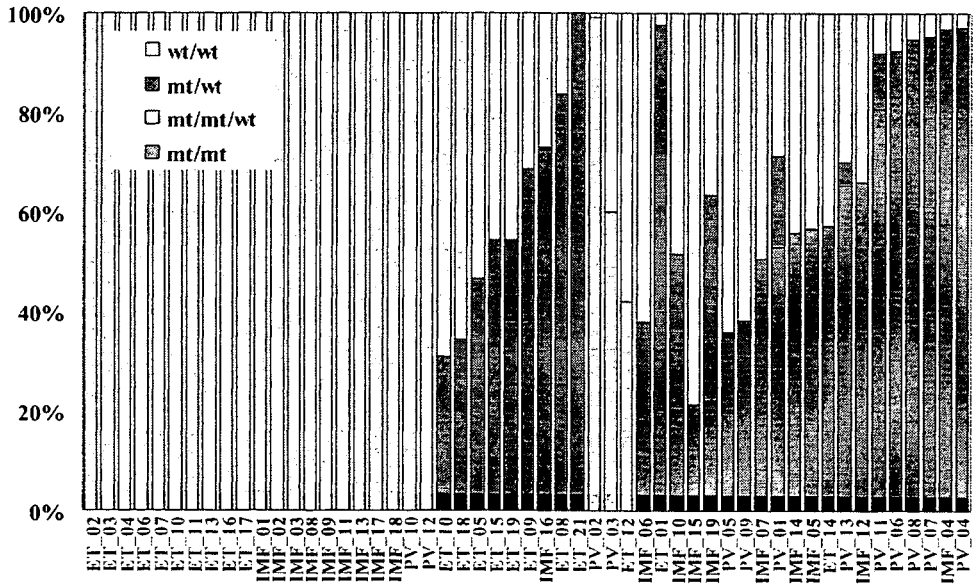

Figure 5. Estimation of tumor populations carrying 9p UPD and the *JAK2* mutation in MPD samples. The populations of 9p UPD–positive components in the 53 MPD cases were estimated by calculation of the mean difference of AsCNs within the UPD regions. Heterozygous (*blue bars*) or homozygous (*red bars*)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

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, $\log_2 R^{refJ}_{AB,i}$ 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

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

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

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

and

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

and the corrected total CN ratio is

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

Averaging over the References of Concordance SNPs

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

$$C_i^K = \{refJ \mid O_i^{tum} = O_i^{refJ}, refJ \in K\}$$

$$C_i^{K,hetero} = \{refJ \mid O_i^{tum} = O_i^{refJ} = AB, refJ \in K\} ,$$

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

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

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

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

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

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

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 $\hat{O}_i^{tum} = AB$, if $\max(\log_2 \tilde{R}^K_{A,i}, \log_2 \tilde{R}^K_{B,i}) \leq 0.1$ or $\min(\log_2 \tilde{R}^K_{A,i}, \log_2 \tilde{R}^K_{B,i}) \geq -0.1$, 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,

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

To do this, instead of testing all possible 2^N combinations of N references, we calculate $SD_K(D)$ for individual references $K = \{ref1\}, \{ref2\}, \{ref3\}, \ldots, \{refN\}$, to order the references such that $SD_1(D) \leq \ldots \leq SD_s(D) \leq SD_{s+1}(D) \leq \ldots \leq SD_N(D)$, 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 $SD_{K(1)}(D) \geq \ldots \geq SD_{K(N_0)}(D) < SD_{K(N_0+1)}(D)$.

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

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

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:

$$Z(\Omega_{i,j}) = \ln \frac{\prod_{i<k<j} \{(1-q)O_k + q(1-O_k)\}}{\prod_{i<k<j} [\{(1-h_k)(1-q) + h_k q\}O_k + \{(1-h_k)q + h_k(1-q)\}(1-O_k)]}$$

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 $Z(\Omega_{I_n,J_n}) > 0$ 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

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

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

and

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

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

(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 $(1/2)\tanh(2\theta)$ 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

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

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

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

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

and

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

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

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

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

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