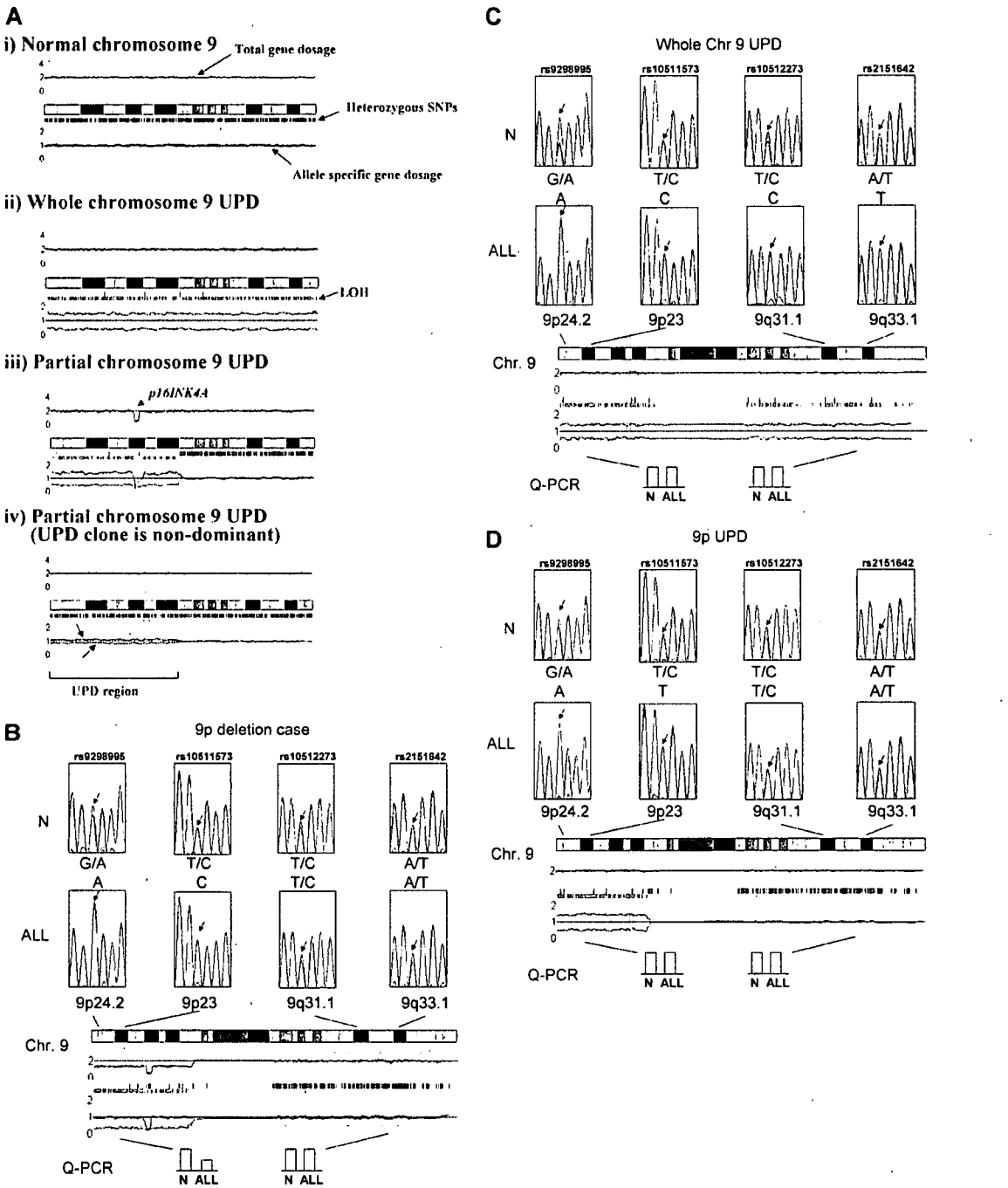

Figure 1. Result of SNP-chip analysis. (A) Results of normal and abnormal chromosomes visualized by CNAG software. Blue lines above each chromosome show total gene dosage; level 2 indicates diploid (2N) amount of DNA, which is normal. Green bars under each chromosome indicate the SNP sites showing heterozygosity in leukemic cells. When heterozygosity is not detected in the leukemic cells but is detected in matched normal controls, the result suggests that the leukemic cells have allelic imbalance (AI) in that region. Pink bars that replaced green ones suggest AI. The bottom lines (green and red lines) in each panel show allele-specific gene dosage levels (one indicates the gene dosage of paternal allele, the other indicates the gene dosage of maternal allele). Level 1 is normal for each gene dosage. (i) Pattern of normal chromosome 9. Blue line is at level 2 (2N DNA). Large number of SNP sites shows normal heterozygosity (green bars under the chromosomes), and no pink bars are detected. Allele-specific gene dosage is at level 1. Panel ii shows the pattern of whole chromosome 9 uniparental disomy (UPD) detected by SNP-chip. Total gene dosage (blue line) is normal (level 2). A number of pink bars (AI) are detected. Allele-specific gene dosage data show that one allele is deleted (level 0) and the other allele is duplicated (level 2). Panel iii shows the pattern of partial UPD. Left half shows the pattern of UPD as described above. Right half shows the pattern of a normal chromosome as described above. This case also has homozygous deletion of *p16INK4A* gene (see that both allele-specific dosage lines [green and red lines] and total gene dosage line [blue line] are at zero). Panel iv shows nondominant UPD. Total gene dosage (blue line) indicates 2N. Allele-specific gene dosage lines (arrows, green and red lines) on left half show that one allele (green line) is lower than normal, and the other allele (red line) is higher than normal. Allele-specific gene dosage on right half show that each allele has same level. (B-D) Validation of SNP-chip data by direct nucleotide sequencing of SNP sites and qPCR. Top panels: direct nucleotide sequencing of SNP sites in ALL samples with matched controls. ALL indicates leukemic samples; N, matched control samples. Heterozygous SNP sites are indicated by arrows. Middle panels: results of SNP-chip data (see Figure 1 legend). Bottom panels: qPCR at each chromosome location. Gene dosage levels were examined using qPCR at indicated chromosomal region. Gene dosage was determined relative to the levels at the 2p21 region. Gene dosage in leukemic cells (ALL) was compared with the matched normal control DNA (N). (B) ALL with 9p hemizygous deletion; homozygous deletion of 9p21 is also detected. (C) ALL with whole chromosome UPD. (D) ALL with 9p UPD.

Table 2. Detection of hyperdiploidy ALL by DNA-index and SNP-chip analysis

	SNP-chip	
	HD	non-HD
DNA index		
HD	44 cases	4 cases
non-HD	30 cases	200 cases

DNA index was measured by FACS as described in "DNA index, immunophenotyping, molecular analysis of chromosomal abnormalities" and DNA index of 278 ALL samples were available. Normal diploid cells have a DNA index of 1.0. When DNA index is the same as or greater than 1.16, the leukemia is defined as hyperdiploid ALL by DNA index. Hyperdiploid ALL detected by SNP-chip analysis had more than 50 chromosomes, which were counted manually.

HD indicates hyperdiploid ALL; and non-HD, nonhyperdiploid ALL.

estimate survival rates. Differences were compared with the 2-sided log-rank test. Event-free survival (EFS) was calculated from diagnosis to the time of the first event (relapse, secondary malignancy, or death from any cause) or to the date of last follow-up.

Results

Features of samples

Clinical features of 399 pediatric ALL patients are shown in Table 1. Infant ALL (< 1years of age) were excluded from this study, and 77% (307 cases) of the patients were from 1 to 9 years old. Forty-nine cases of T-cell lineage ALL and 339 cases of B-cell lineage ALL were examined. Ninety-six samples (24%) had *ETV6/RUNX1* fusion, and 6 cases had the *BCR/ABL* fusion gene.

Validation of SNP-chip data

Gene dosage, heterozygous SNPs, allele-specific gene dosage, and allelic composition (loss of heterozygosity [LOH]) was visualized as shown in Figure 1 using our novel analysis software, CNAG for SNP-chip.[11,12] Duplication/amplification, deletion, and UPDs of chromosomes were easily detected (Figure 1A). To validate abnormalities found by SNP-chip, genomic quantitative PCR and direct sequencing of SNP sites at duplicated, amplified, deleted, and UPD regions were performed including chromosome 9. Representative results of validation are shown in Figure 1B-D.

Also, hyperdiploid (HD) ALL defined by DNA index and SNP-chip analysis was compared for selected cases (Table 2). Number of total chromosomes was counted manually in SNP-chip analysis, and ALL with more than 50 chromosomes was defined as HD-ALL by SNP-chip. When DNA index of leukemic cells was same as or greater than 1.16, the sample was defined as HD-ALL by DNA index.[16,17] DNA index of 278 ALL samples were available, and 200 cases were defined as non-HD ALL by both methods. SNP-chip detected more cases of HD-ALL (74 cases) than DNA index. As shown in Figure 1Aiv, SNP-chip can precisely detect gene dosage, and this high sensitivity of SNP-chip analysis permitted more accurate detection of HD-ALL than by the DNA index method. Results of these analyses validated that the abnormalities detected by SNP-chip were reliable.

Three common abnormalities in pediatric ALL

Figure 2A summarizes molecular allelokaryotyping profiles of the 399 ALL cases after clustering with regard to the status of copy number alterations as well as copy number neutral LOH, so-called UPD, showing a number of clusters having common genetic lesions.

Among these clusters, 3 genetic abnormalities were frequently detected: hyperdiploidy (HD, > 50 chromosomes), deletion of the 9p region, and deletion of 12p (Figure 2A,B). The common deleted region (CDR) on 9p involved the *p16INK4A* gene (called p16Del, Figure 2B), and the CDR on 12p involved the *ETV6* gene (called ETV6Del, Figure 2B). Concurrent abnormalities of p16Del and HD were rare ($P < .001$); concurrent abnormalities of ETV6Del and HD also were rare ($P < .001$; Figure 2B). No case had all 3 common abnormalities.

The clinical features of cases with each of these 3 genetic abnormalities were analyzed (Table 3). Individuals with p16Del-ALL frequently were older ($P = .017$), had higher WBC ($P < .001$), and T-cell lineage ALL ($P < .001$). Those with ETV6Del-ALL were more often younger ($P = .009$), non–T-cell lineage ($P = .014$), and *ETV6/RUNX1* fusion gene positive ($P < .001$). Patients having HD-ALL were more frequently younger ($P < .001$), showed lower WBC ($P < .001$), non–T-cell lineage ($P < .001$), and *ETV6/RUNX1* negative ($P < .001$).

Numerical chromosomal abnormalities in pediatric ALL

Numerical chromosome changes were frequently detected in pediatric ALL samples, as summarized in Figure 3A. Numerical change of chromosome 21 (trisomy, tetrasomy, and pentasomy) was the most frequent (134 [34%] cases). We had 8 cases with Down syndrome who had trisomy 21 in their leukemic cells and their matched controls. These 8 cases are excluded in Figure 3A. Most of the numerical abnormalities were detected in HD-ALL cases (Figure S1A) except for those with trisomy 21 (Figure S1B). As for trisomy 21, half (21 cases) occurred in patients with subtypes other than HD (Figure S1B). In HD-ALL, gain of chromosomes was restricted to particular chromosomes, involving chromosomes 4, 6, 8, 10, 14, 17, 18, 21, and X (Figures 2A, 3A).

Nonrandom genetic abnormalities in pediatric ALL detected by SNP-chip

All copy number changes (deletions and duplications/amplifications) detected by SNP-chip analysis are shown in Figure 2A and Figure S2. Small deletions that could not be detected by conventional cytogenetics were sensitively identified, including deletions of 3p14.2 (500 kb), 3q26.32 (700 kb), and Xp21.1 (1 Mb) (Table 4 and Figure S2). Nonrandom chromosomal abnormalities (frequency > 1.5% of all cases) are listed in Table 4. Besides the 3 common genetic abnormalities, duplication of 1q (11%) and deletion of 6q (11.5%) were often detected. In 13 cases with 1q duplication, the duplication began at the *PBX1* gene (Figures 2A, S2). Since gain of the entire or part of either chromosome 21 or X was frequently found in non–HD-ALL, these abnormalities were grouped separately (Table 4).

Recently, other groups of investigators performed SNP-chip analysis on pediatric ALL and found deletions of several transcriptional factors associated with B-cell development including *PAX5* (9p13), *EBF* (5q33), *Ikaros* (7p12.2), *Aiolos* (17q12), *LEF1*(4q25), *RAG1*(11p12), and *RAG2* (11p12).[19,20] We also have found deletion of these genes in our study. However, the frequency of deletion of these genes, except for PAX5, was low (fewer than 2%) and/or the deleted regions contained multiple genes (Table 4; Figure S2 and data not shown).

UPD

One of the major advantages of SNP-chip analysis is capability of sensitive detection of UPD, even in samples suffering from small

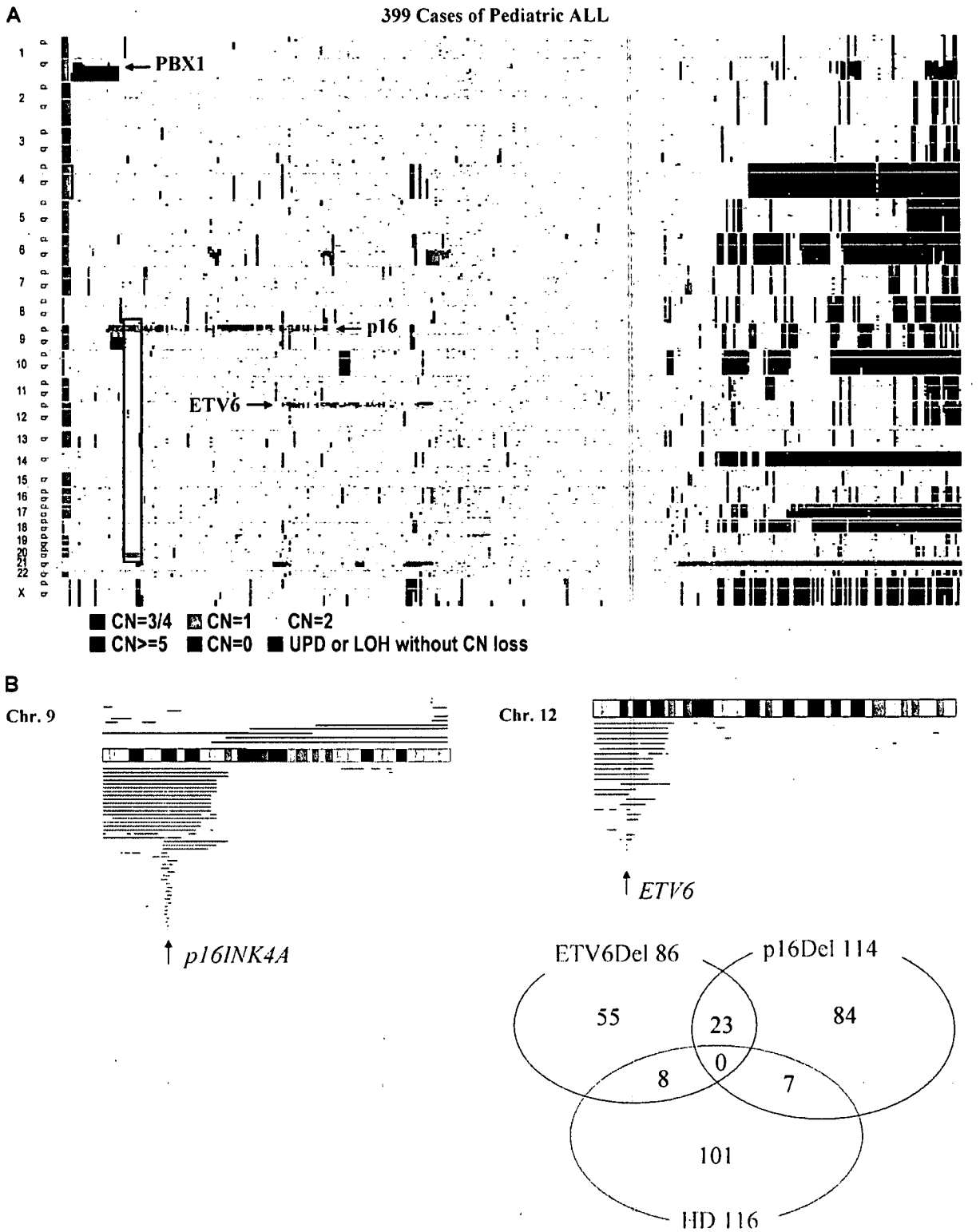

Figure 2. Allelokaryotyping of pediatric ALL. (A) Genetic clustering of 393 ALL samples. Genetic status of each chromosomal region is visualized. Vertical axis: chromosomes, p: short arms, q: long arms. Horizontal axis: individual cases. CN: copy number of alleles. UPD: uniparental disomy. Locations of *PBX1*, *INK4A/ARF*(p16), and *ETV6* genes are indicated. A rectangle indicates cases having 9p and 20q deletions. (B) Three common genomic abnormalities detected in ALL by SNP-chip analysis. Left panel: Deletion of 9p is frequently detected; the arrow indicates the commonly deleted region (CDR) where the p16INK4A gene is located. Right panel: Deletion of 12p often occurs. The arrow indicates that the CDR is where the ETV6 gene is localized. Green lines under the chromosome indicate the deleted regions in individual cases. Brown lines above the chromosome indicate the duplicated regions. Only representative cases are shown. Green and red bands on idiograms indicate centromeres and noncoding regions, respectively. Bottom panel: Venue diagram of 3 common abnormalities detected in this study. Numbers of respective cases in each category are indicated. HD: ALL with hyperdiploid (chromosomes > 50). ETV6Del: ALL with deletion of ETV6 gene. p16Del: ALL with deletion of p16INK4A gene.

Table 3. Clinical features of ALL cases associated with one of three common genetic abnormalities

	Genetic abnormality, no. (%)	Others, no. (%)	P
p16Del-ALL			
Age			
1 to 9 y	76 (68)	231 (80)	—
Older than 9 y	35 (32)	57 (20)	.017
WBC			
Below $10^2 \times 10^9$/L	85 (77)	277 (96)	—
Over $10^2 \times 10^9$/L	26 (23)	11 (4)	.001
Non-T lineage	71 (62)	263 (96)	—
T-lineage	83 (38)	11 (4)	.001
ETV6Del-ALL			
Age			
1 to 9 y	75 (87)	232 (74)	—
Older than 9 y	11 (13)	81 (26)	.009
Non-T lineage	89 (95)	261 (85)	—
T-lineage	4 (5)	45 (15)	.015
ETV6/RUNX1			
Positive	53 (66)	43 (15)	—
Negative	27 (34)	243 (85)	.001
HD-ALL			
Age			
1 to 9 y	101 (89)	206 (72)	—
Older than 9	13 (11)	79 (28)	.001
WBC			
Below $10^2 \times 10^9$/L	112 (98)	250 (88)	—
Over $10^2 \times 10^9$/L	2 (2)	35 (12)	.001
Non-T lineage	110 (100)	229 (82)	—
T-lineage	0 (0)	49 (18)	.001
ETV6/RUNX1	HD-ALL	Others	
Positive	8 (8)	88 (34)	—
Negative	97 (92)	173 (66)	.001

p16Del-ALL indicates ALL with deletion of *p16INK4A* gene; ETV6Del-ALL, ALL with deletion of *ETV6* gene; HD-ALL, ALL with hyperdiploidy (chromosomes >50); —, not applicable; WBC, white blood cell count in peripheral blood ($\times 10^9$/L) at diagnosis; *and ETV6/RUNX1, ETV6/RUNX1* fusion was examined by RT-PCR and/or FISH analysis.

tumor content; UPD in samples with as low as 20% of tumor contents are clearly identified (Figure 1Aiv). Whole and partial chromosome UPDs were observed in 95 cases; 6 cases showed both whole and partial chromosome UPD (Figure 3B). Most of the whole chromosome UPD was detected in HD-ALL (Figure 3C). UPD of whole chromosome 9 was the most common whole chromosome UPD (Figure 3B). In contrast, UPD involving part of chromosomes was most often found in non–HD-ALL cases (Figure 3C). Recurrent partial chromosome UPD was detected in many chromosomal regions (Figure 3B). We frequently found whole chromosome 9 UPD (18 cases) and 9p UPD (30 cases) (Figure 3B). *INK4A* deletion was often found in 9p UPD (23 of 30, 77%), while it was rare in whole chromosome 9 UPD (1 of 18, 6%) (Figure 3D).

Relationship between genetic abnormalities

Recurrent abnormalities described above were compared with each other (3 common abnormalities and 26 nonrandom alterations) to detect relationships between these abnormalities (Table S1). Strong correlations between abnormalities of 12p and 21q were detected, duplications of 12p and 21q often occurred simultaneously, and duplications of 21q were accompanied by deletion of the *ETV6* gene that was localized on 12p. ETV6Del ALL frequently had additional changes, including duplications (21q and 1q) and deletions (3p21, 1q, *FHIT*, 15q, and 4q). Abnormalities involving chromosome X, including *DMD* (Xp21.2) deletion, were frequently accompanied by deletions of 8p, 4q, and 6q. Deletion of 20q often occurred with either *p16INK4A* deletion or duplication of 21q (Figure 2A and Table S1).

Impact of nonrandom genetic abnormalities on prognosis

We analyzed prognosis of cases showing nonrandom abnormalities listed in Table 4 and found that the recurrent abnormalities had no impact on the event-free survival (EFS; data not shown) except for amplification/duplication of chromosome 9q. Our initial, early analysis found that EFS of pediatric ALL patients was not impacted by *ETV6* deletion either with or without *ETV6/RUNX1*. Nine cases with 9q amplification/duplication showed a poor prognosis (6 patients relapsed within 3 years; Figure S3A), although the number of cases having this abnormality is too small to reach a significant conclusion. Of these cases, 3 also had duplication of part of chromosome 22 (Figure S3B), suggesting that duplication of 9q is part of an extra copy of the Philadelphia chromosome. These 3 cases showed *BCR/ABL* positively by FISH/RT-PCR analysis (data not shown). Two other cases showed high copy number amplification that encompassed *ABL* and *NUP214* genes (Figure S3C), suggesting these cases had a *NUP214/ABL* fusion.[21] The ALL cells of these 2 cases were steroid-resistant and T-cell lineage phenotype.

Children with HD-ALL without gain of either chromosome 17 or 18 had a worse prognosis (Figure 4). Furthermore, children with HD-ALL and no extra copies of chromosomes 17 and 18 had a significantly worse prognosis ($P < .001$), with 53% EFS at 5 years compared with a 90% 5-year EFS in the other HD-ALL cohort (Figure 4).

Discussion

SNP-chip analysis is a reliable method to detect gene dosage, which was validated by direct sequencing of SNP sites and quantitative PCR in this study. To detect HD-ALL, DNA index is not a sensitive method since contaminated normal cells (DNA index 1.0) decrease the levels of DNA index of hyperdiploid leukemic cells. Although karyotyping is a good method to detect HD-ALL, sufficient number of high-quality chromosomal metaphases is not always obtained from the leukemic cells. SNP-chip analysis may be a more useful and reliable method to detect this subtype of ALL.

Molecular allelokaryotyping of a large series of pediatric ALL samples showed 3 major abnormalities: deletion of *p16INK4A*, deletion of *ETV6*, and hyperdiploidy. Besides these 3 common abnormalities, a number of novel, nonrandom changes were found in ALL. Some of them showed a very narrow commonly deleted region, which was limited to one target gene, including *FHIT* (3p14.2), *TBL1XR1* (3q26.3), and *DMD* (Xp21.2). *DMD* is the causative gene for Duchenne-type muscular dystrophy.[22] While germ-line inactivating mutations of this gene cause the disease, no association has been made between this disease and cancers, including ALL. Since *DMD* is an extremely large gene (2.4 Mb), deletion of it may occur as a result of instability of genomic DNA in ALL cells.

Other investigators and we have found in pediatric ALL a number of deleted genes, including transcriptional factors involved in B-cell differentiation.[19,20] However, since no point mutations of those genes, except PAX5, were detected,[19] it is unclear that these transcriptional factors associated with B-cell development are target genes of these deletions. Mullighan et al[19] and we have found that PAX5 gene is frequently involved in deletions and translocations (N.K., S.O., M.Z., et

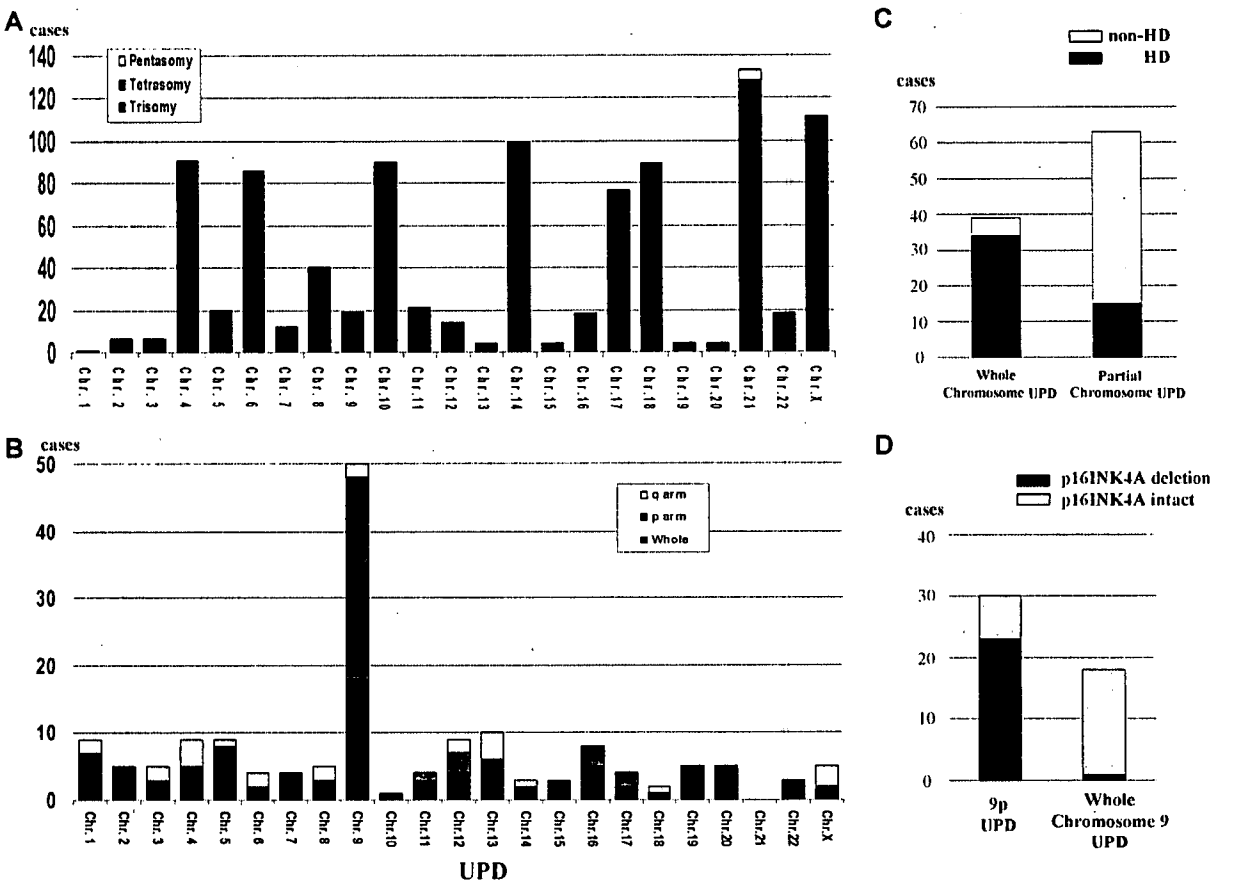

Figure 3. Numerical chromosomal changes and uniparental disomy in pediatric ALL. (A) Frequency of pentasomy/tetrasomy/trisomy affecting each chromosome. For X chromosome, trisomy (105 cases) contains trisosomy X in male pateints (67 cases) and disomy X in female patients (38 cases). All tetrasomy X were female patients. (B) Frequency of uniparental disomy (UPD). Whole: cases with whole chromosome UPD; p arm: cases with UPD of short arm; and q arm: cases with UPD of long arm. Chr: chromosome. UPD involving X chromosome was detected only in female cases. (C) Distribution of whole and partial chromosome UPD in HD and non–HD-ALL. Whole chromosome UPD is frequently detected in HD-ALL. Thirty-four cases with whole chromosome UPD were HD-ALL. Partial UPD is frequently detected in non–HD-ALL. Fifteen of 63 cases with partial UPD were HD-ALL. (D) Frequency of deletion of p16*INK4A* gene in whole chromosome 9 UPD and 9p UPD. Twenty-three cases showed deletion of p*16INK4A*, out of a total of 30 cases with 9p UPD. One case had p16*INK4A* deletion from a total 18 ALL samples with whole chromosome 9 UPD.

al, manuscript submitted), suggesting that impairment of PAX5 is associated with leukomogenesis.

This study showed that numerical chromosomal changes and UPD were common genomic abnormalities in pediatric ALL. Interestingly, whereas trisomy 21 was the most common numerical chromosomal abnormality, UPD of chromosome 21 was not detected in our study. Even in 8 cases with Down syndrome who had trisomy 21 in their constitutional DNA, UPD of chromosome 21 was not detected. Although UPD of chromosome 21 in leukemic cells of patients with Down syndrome has been reported,[8] it may be a rare event in pediatric ALL.

Chromosomal mis-segregation occurs when duplicated chromosomes separate improperly during cell division,[23] leading to numerical chromosomal changes in leukemic cells, including HD-ALL. Most of the whole chromosome UPD was detected in HD-ALL, suggesting that whole chromosome UPD is due also to chromosomal mis-segregation. In contrast, UPD involving part of chromosomes was most often found in non–HD-ALL cases. This may suggest that these partial UPDs are not caused by mis-segregation, but by mitotic recombination, which can usually cause chromosomal exchange.[24]

UPD involving chromosome 9 or 9p is a very common abnormality in pediatric ALL. *INK4A* gene may be a target gene of 9p UPD since most of the cases with 9p UPD had deletion of *INK4A* (23 of 30 cases with 9p UPD). For UPD involving whole chromosome 9, INK4A is not a target gene, since only one case with whole chromosome 9 UPD had deletion of this gene (1 of 18, 6%). Which gene is the target of whole chromosome 9 UPD is unclear. Furthermore, 7 cases with 9p UPD had intact *INK4A*, and the target gene of 9p UPD in these cases is also unknown. This is the first report showing that whole chromosome 9 UPD and 9p UPD are common abnormalities in pediatric ALL. Although Mullighan et al also analyzed a large number of pediatric ALL patients by SNP-chip,[19] they did not report this abnormality.

UPD on 9p is associated with activating point mutations of *JAK2* in myeloproliferative disorders (MPDs).[26-28] However, point mutation of *JAK2* in ALL is very rare.[29,30] We examined for *JAK2* mutations at "hot-spots" (exons 12 and 14)[18,26-28] in these 7 cases of ALL with 9p UPD in which deletion of *INK4A* was not detected. We found no mutations of *JAK2* in these cases. Another unidentified gene(s) in the region may be mutated in these cases.

In this study, we found that absence of gain of chromosomes 17 and 18 had an adverse impact on the prognosis of children with HD-ALL. Another large-scale study found that gain of chromosome 17 was associated with a better prognosis in HD-ALL.[31] Although this previous study reported that gain of chromosomes 4

Table 4. Recurrent genetic abnormalities detected by SNP-chip

Chromosomal sites	Type of abnormality	No. of cases (%)	Candidate genes
1q	Duplication	44 (11)	—
1q	Deletion	11 (3)	—
3p21	Deletion	6 (2)	—
3p14.2	Deletion	6 (2)	*FHIT*
3q26.3	Deletion	10 (3)	*TBL1XR1*
4q31	Deletion	7 (2)	—
6q	Deletion	46 (11)	—
7p	Deletion	10 (3)	—
7q34	Deletion	7 (2)	—
8p	Deletion	13 (3)	—
8q	Duplication	9 (2)	—
9q	Dup/amp	9 (2)	*ABL*
10p	Duplication	7 (2)	—
10q24	Deletion	12 (3)	—
11q	Deletion	24 (6)	—
12p	Duplication	13 (3)	—
13q14.2	Deletion	14 (4)	*RB1*
15q	Deletion	7 (2)	—
17p	Deletion	8 (2)	*TP53*
17q	Duplication	10 (3)	—
17q11.2	Deletion	7 (2)	*NF1*
20p12.2	Deletion	6 (2)	—
20q	Deletion	13 (3)	—
Xp21.2	Deletion	8 (2)	*DMD*
Gain of Chr. 21 or 21q in non-HD ALL cases	—	37/283 (13)	—
Gain of Chr. X or Xq in non-HD ALL cases	—	23/283 (8)	—

Nonrandom chromosomal abnormalities (frequency >1.5% of all cases) are listed. 9p deletion and 12p deletions are separately shown in Figure 2B.
HD indicates hyperdiploid (>50 chromosomes); dup/amp, duplication and amplification of the region; and —, not applicable.

and 10 also was associated with a better prognosis, our study showed that change in number of these chromosomes did not have an impact on prognosis. Even though the size of our study is relatively large, it might not be able to detect some important factors associated with survival because the number of cases of this ALL subtype enrolled in our study was too small or advances in treatment of pediatric ALL may have eliminated several factors that previously influenced prognosis.

One of the limitations of SNP-chip analysis is that it cannot detect balanced translocations, which are common abnormalities in ALL, since this technique can detect only allelic dosage. In our correlation study of genomic abnormalities, a strong correlation was found between abnormalities involving 12p and 21q as described above. This, in part, reflected translocations of chromosome 12p and 21q (*ETV6/RUNX1* fusion). Another correlation that we found between p16Del (on 9p) and deletion of 20q, may reflect dic(p13;q11).[9,20,32,33] These strong correlations of gains or loss of genetic materials may, therefore, sometimes suggest unbalanced translocations in ALL. Combination of SNP-chip and karyotyping will be a very strong technique to examine all genomic abnormalities in malignant cells.

Molecular allelokaryotyping of a series of 399 pediatric ALL samples has defined the range of genetic changes that occur in childhood ALL, including those associated with a poor prognosis. SNP-chip analysis is a novel technique that allows a thorough

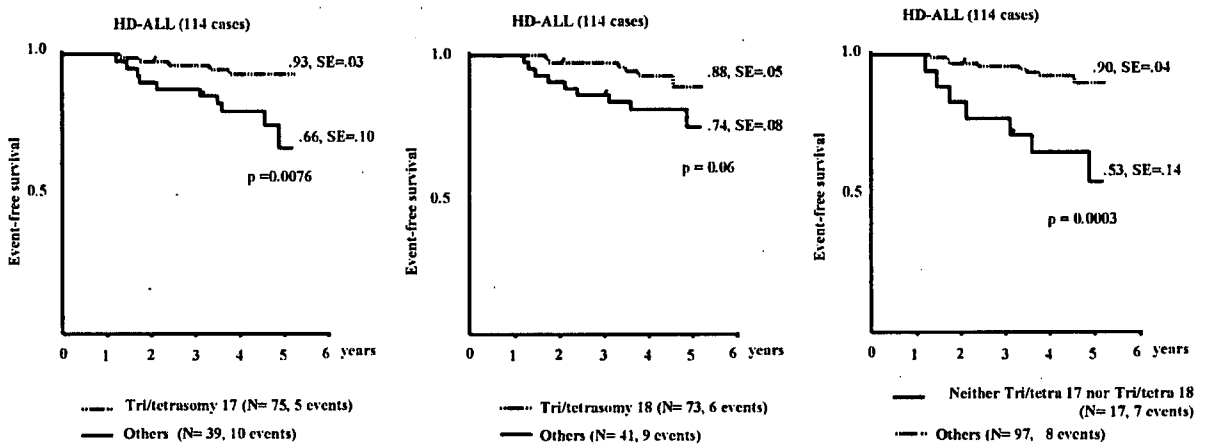

Figure 4. Prognostic impact of chromosomes 17 and 18. Absence of a gain of either chromosome 17 (left panel) or chromosome 18 (middle panel) is associated with poor prognosis in patients with HD-ALL; concurrent absence of a gain of both chromosomes 17 and 18 (right panel) is associated with very poor prognosis. Tri/tetra 18 and 17: HD-ALL with trisomy or tetrasomy 18 or 17. Others: HD-ALL without gain of chromosomes 17 and/or 18.

interrogation of the genome in ALL and identification of clinically significant subgroups of patients.

Acknowledgments

We thank the Parker Hughes Fund and National Institues of Health grants for supporting this study. N.K. is supported by the fellowship from The Tower Cancer Research Foundation. H.P.K. holds the Mark Goodson Chair in Oncology Research at Cedars Sinai and is a member of the Jonsson Cancer Center and the Molecular Biology Institute of UCLA. This work was also supported by grant-in-aid from Department of Health, Welfare. and Labor and from MEXT of the Japanese government, by EU grant FOOD-CT-2005-016320, and a grant from the Deutsche Krebshilfe to C.R.B. The ALL-BFM 2000 trial is supported by 50-2698-Schr 1 of the Deutsche Krebshilfe.

Authorship

Contribution: N.K. and S.O. designed this study, performed experiments, analyzed the data, and wrote the paper. M.Z., K.H., R.K., and W.-D.L. analyzed the data and wrote the paper. M.K., T.F., C.W.M., J.H., and M.S. analyzed the data. M.S., G.Y. and Y.N. performed experiments and analyzed the data. M.S. designed this study and wrote the paper. C.R.B. and H.P.K. designed this study, analyzed the data, and wrote the paper.

Conflict-of-interest disclosure: The authors declare no competing financial interests.

Correspondence: Seishi Ogawa, Department of Regeneration Medicine for Hematopoiesis, Graduate School of Medicine, University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8655, Japan; e-mail: sogawa-tky@umin.ac.jp; or Norihiko Kawamata, Hematology/Oncology, Cedars-Sinai Medical Institute, 8700 Beverly Blvd, Los Angeles, CA 90048; e-mail: kawamatan@cshs.org.

References

1. Armstrong SA, Look AT. Molecular genetics of acute lymphoblastic leukemia. J Clin Oncol. 2005;23:6306-6315.
2. Pui CH, Evans WE. Treatment of acute lymphoblastic leukemia. N Engl J Med. 2006;354:166-178.
3. Pui CH, Relling MV, Downing JR. Acute lymphoblastic leukemia. N Engl J Med. 2004;350:1535-1548.
4. Gunderson KL, Steemers FJ, Lee G, Mendoza LG, Chee MS. A genome-wide scalable SNP genotyping assay using microarray technology. Nat Genet. 2005;37:549-554.
5. Garraway LA, Widlund HR, Rubin MA, et al. Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. Nature. 2005;436:117-122.
6. Raghavan M, Lillington DM, Skoulakis S, et al. Genome-wide single nucleotide polymorphism analysis reveals frequent partial uniparental disomy due to somatic recombination in acute myeloid leukemias. Cancer Res. 2005;65:375-378.
7. Teh MT, Blaydon D, Chaplin T, et al. Genomewide single nucleotide polymorphism microarray mapping in basal cell carcinomas unveils uniparental disomy as a key somatic event. Cancer Res. 2005;65:8597-8603.
8. Rogan PK, Close P, Blouin JL, et al. Duplication and loss of chromosome 21 in two children with Down syndrome and acute leukemia. Am J Med Genet. 1995;59:174-181.
9. International HapMap Consortium. A haplotype map of the human genome. Nature. 2005;437:1299-1320.
10. Matsuzaki H, Dong S, Loi H, et al. Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. Nat Methods. 2004;1:109-111.
11. Nannya Y, Sanada M, Nakazaki K, et al. A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. Cancer Res. 2005;65:6071-6079.
12. Yamamoto G, Nannya Y, Kato M, et al. Highly sensitive method for genomewide detection of allelic composition in nonpaired, primary tumor specimens by use of affymetrix single-nucleotide-polymorphism genotyping microarrays. Am J Hum Genet. 2007;81:114-126.
13. Bene MC, Castoldi G, Knapp W, et al. Proposals for the immunological classification of acute leukemias: European Group for the Immunological Characterization of Leukemias (EGIL). Leukemia. 1995;9:1783-1786.
14. Ludwig WD, Rieder H, Bartram CR, et al. Immunophenotypic and genotypic features, clinical characteristics, and treatment outcome of adult pro-B acute lymphoblastic leukemia: results of the German multicenter trials GMALL 03/87 and 04/89. Blood. 1998;92:898-909.
15. Schrappe M, Reiter A, Ludwig WD, et al. Improved outcome in childhood acute lymphoblastic leukemia despite reduced use of anthracyclines and cranial radiotherapy: results of trial ALL-BFM 90. German-Austrian-Swiss ALL-BFM Study Group. Blood. 2000;95:3310-3322.
16. Look AT, Melvin SL, Williams DL, et al. Aneuploidy and percentage of S-phase cells determined by flow cytometry correlate with cell phenotype in childhood acute leukemia. Blood. 1982;60:959-967.
17. Whitehead VM, Vuchich MJ, Lauer SJ, et al. Accumulation of high levels of methotrexate polyglutamates in lymphoblasts from children with hyperdiploid (greater than 50 chromosomes) B-lineage acute lymphoblastic leukemia: a Pediatric Oncology Group study. Blood. 1992;80:1316-1323.
18. Scott LM, Tong W, Levine RL, et al. JAK2 exon 12 mutations in polycythemia vera and idiopathic erythrocytosis. N Engl J Med. 2007;356:459-468.
19. Mullighan CG, Goorha S, Radtke I, et al. Genome-wide analysis of genetic alterations in acute lymphoblastic leukaemia. Nature. 2007;446:758-764.
20. Kuiper RP, Schoenmakers EF, van Reijmersdal SV, et al. High-resolution genomic profiling of childhood ALL reveals novel recurrent genetic lesions affecting pathways involved in lymphocyte differentiation and cell cycle progression. Leukemia. 2007;21:1258–1266.
21. Graux C, Cools J, Melotte C, et al. Fusion of NUP214 to ABL1 on amplified episomes in T-cell acute lymphoblastic leukemia. Nat Genet. 2004;36:1084-1089.
22. Koenig M, Hoffman EP, Bertelson CJ, Monaco AP, Feener C, Kunkel LM. Complete cloning of the Duchenne muscular dystrophy (DMD) cDNA and preliminary genomic organization of the DMD gene in normal and affected individuals. Cell. 1987;50:509-517.
23. Kops GJ, Weaver BA, Cleveland DW. On the road to cancer: aneuploidy and the mitotic checkpoint. Nat Rev Cancer. 2005;5:773-785.
24. Kotzot D. Complex and segmental uniparental disomy (UPD): review and lessons from rare chromosomal complements. J Med Genet. 2001;38:497-507.
25. Baxter EJ, Scott LM, Campbell PJ, et al. Acquired mutation of the tyrosine kinase JAK2 in human myeloproliferative disorders. Lancet. 2005;365:1054-1061.
26. James C, Ugo V, Le Couedic JP, et al. A unique clonal JAK2 mutation leading to constitutive signalling causes polycythaemia vera. Nature. 2005;434:1144-1148.
27. Kralovics R, Passamonti F, Buser AS, et al. A gain-of-function mutation of JAK2 in myeloproliferative disorders. N Engl J Med. 2005;352:1779-1790.
28. Levine RL, Wadleigh M, Cools J, et al. Activating mutation in the tyrosine kinase JAK2 in polycythemia vera, essential thrombocythemia, and myeloid metaplasia with myelofibrosis. Cancer Cell. 2005;7:387-397.
29. Levine RL, Loriaux M, Huntly BJ, et al. The JAK2V617F activating mutation occurs in chronic myelomonocytic leukemia and acute myeloid leukemia, but not in acute lymphoblastic leukemia or chronic lymphocytic leukemia. Blood. 2005;106:3377-3379.
30. Sulong S, Case M, Minto L, Wilkins B, Hall A, Irving J. The V617F mutation in Jak2 is not found in childhood acute lymphoblastic leukaemia. Br J Haematol. 2005;130:964-965.
31. Sutcliffe MJ, Shuster JJ, Sather HN, et al. High concordance from independent studies by the Children's Cancer Group (CCG) and Pediatric Oncology Group (POG) associating favorable prognosis with combined trisomies 4, 10, and 17 in children with NCI Standard-Risk B-precursor Acute Lymphoblastic Leukemia: a Children's Oncology Group (COG) initiative. Leukemia. 2005;19:734-740.
32. Rieder H, Schnittger S, Bodenstein H, et al. dic(9;20): a new recurrent chromosome abnormality in adult acute lymphoblastic leukemia. Genes Chromosomes Cancer. 1995;13:54-61.
33. Heerema NA, Maben KD, Bernstein J, Breitfeld PP, Neiman RS, Vance GH. Dicentric (9;20)(p11;q11) identified by fluorescence in situ hybridization in four pediatric acute lymphoblastic leukemia patients. Cancer Genet Cytogenet 1996;92:111-115.

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. 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 non-mixed 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*),

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

and

$$S^{\mathrm{ref}l}_{A,i} = \sum \mathrm{PM}^{\mathrm{ref}l}_{A,i} \ , \quad S^{\mathrm{ref}l}_{B,i} = \sum \mathrm{PM}^{\mathrm{ref}l}_{B,i} \ , \quad (I = 1,2,3,\ldots,N) \ ,$$

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

$$R^{\mathrm{ref}l}_{A,i} = \frac{S^{\mathrm{tum}}_{A,i}}{S^{\mathrm{ref}l}_{A,i}} \qquad (\text{for } O^{\mathrm{tum}}_i = O^{\mathrm{ref}l}_i),$$

$$R^{\mathrm{ref}l}_{B,i} = \frac{S^{\mathrm{tum}}_{B,i}}{S^{\mathrm{ref}l}_{B,i}}$$

and the total CN ratio is calculated as follows:

$$R^{\mathrm{ref}l}_{AB,i} = \begin{cases} R^{\mathrm{ref}l}_{A,i} & \text{for } O^{\mathrm{tum}}_i = O^{\mathrm{ref}l}_i = AA \\ R^{\mathrm{ref}l}_{B,i} & \text{for } O^{\mathrm{tum}}_i = O^{\mathrm{ref}l}_i = BB \\ \frac{1}{2}\left(R^{\mathrm{ref}l}_{A,i} + R^{\mathrm{ref}l}_{B,i}\right) & \text{for } O^{\mathrm{tum}}_i = O^{\mathrm{ref}l}_i = AB \end{cases} \quad (I = 1,2,3,\ldots,N) \ .$$

For CN estimations, however, $R^{\mathrm{ref}l}_{AB,i}$, $R^{\mathrm{ref}l}_{A,i}$, and $R^{\mathrm{ref}l}_{B,i}$ 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 $\hat{R}^{\mathrm{ref}l}_{AB,i}$, $\hat{R}^{\mathrm{ref}l}_{A,i}$, and $\hat{R}^{\mathrm{ref}l}_{B,i}$, 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 $\bar{R}^K_{AB,i}$, $\bar{R}^K_{A,i}$, and $\bar{R}^K_{B,i}$. Note that $\bar{R}^K_{A,i}$ and $\bar{R}^K_{B,i}$ 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

$$\Lambda_K = \{\bar{R}^K_{AB,i}\} \ ,$$

and provisional AsCN profiles are obtained by

$$\Lambda^{\mathrm{large}}_K = \{\max(\bar{R}^K_{A,i}, \bar{R}^K_{B,i})\}$$

$$\Lambda^{\mathrm{small}}_K = \{\min(\bar{R}^K_{A,i}, \bar{R}^K_{B,i})\} \ .$$

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.

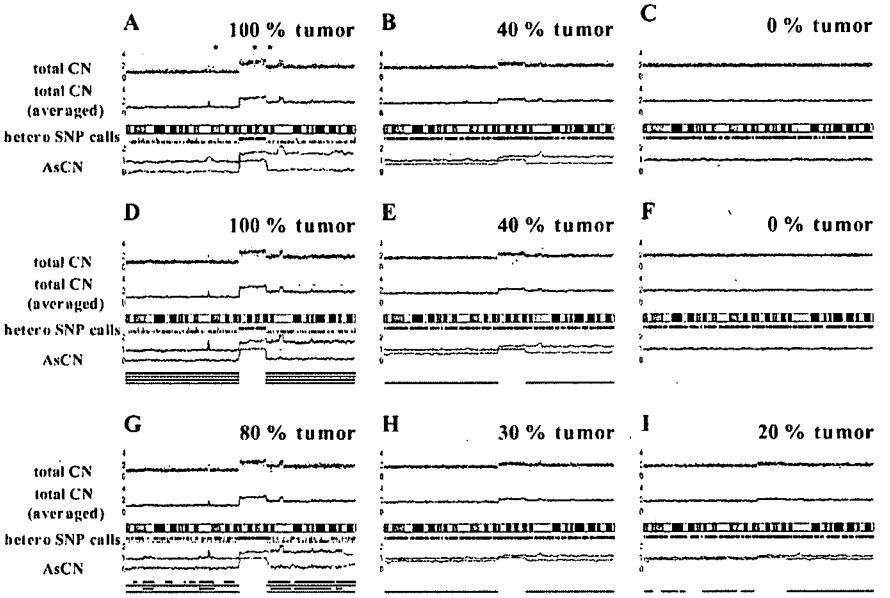

Figure 1. AsCN analysis with or without paired DNA. DNA from a lung cancer cell line (NCI-H2171) was mixed with DNA from an LCL (NCI-BL2171) established from the same patient at the indicated percentages and was analyzed with GeneChip 50K Xba SNP arrays. AsCNs, as well as total CNs, were analyzed using 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

$$\Lambda^{\text{large}} = \left|\max\left(\hat{R}_{A,i}^{\text{refSelf}}, \hat{R}_{B,i}^{\text{refSelf}}\right)\right|$$

and

$$\Lambda^{\text{small}} = \left|\min\left(\hat{R}_{A,i}^{\text{refSelf}}, \hat{R}_{B,i}^{\text{refSelf}}\right)\right| .$$

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, $|\log_2 \hat{R}_{A,i}^{K} - \log_2 \hat{R}_{B,i}^{K}|$, 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

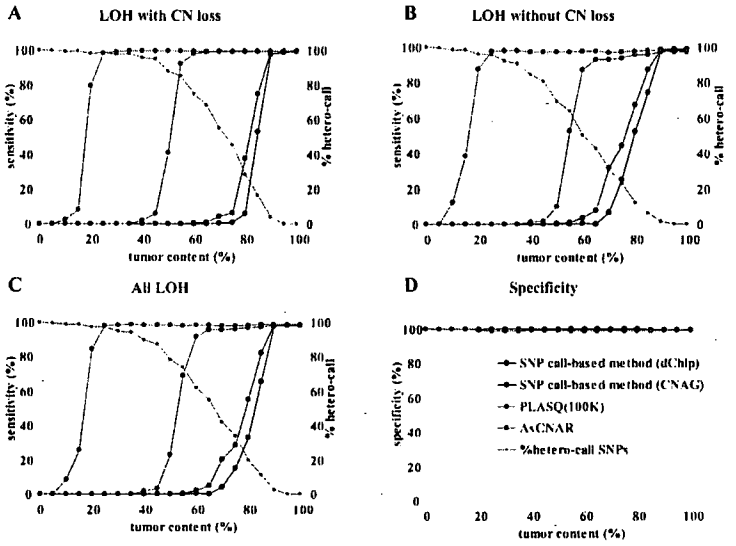

Figure 2. Sensitivity and specificity of LOH detection for intentionally mixed tumor samples. Sensitivity of detection of LOH with or without CN loss (*A* and *B*) in different algorithms were compared using a mixture of the tumor sample (NCI-H2171) and the paired LCL sample (NCI-BL2171). The results 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:

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

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-

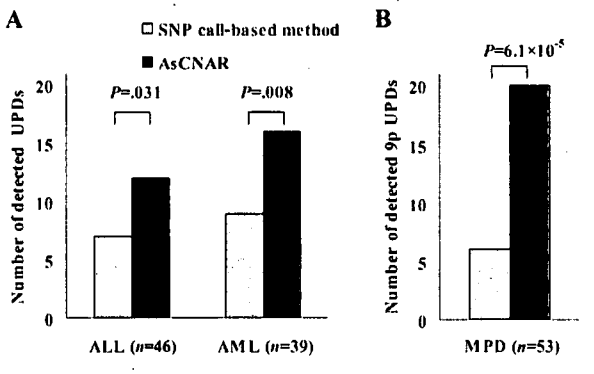

Figure 3. The number of UPD regions for acute leukemia and MPD samples detected by either the SNP call–based method or AsCNAR. The number of UPD regions for ALL and AML samples detected by the SNP call–based method or by AsCNAR is shown in panel A, and the number of 9p UPDs for MPD samples detected by 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*.

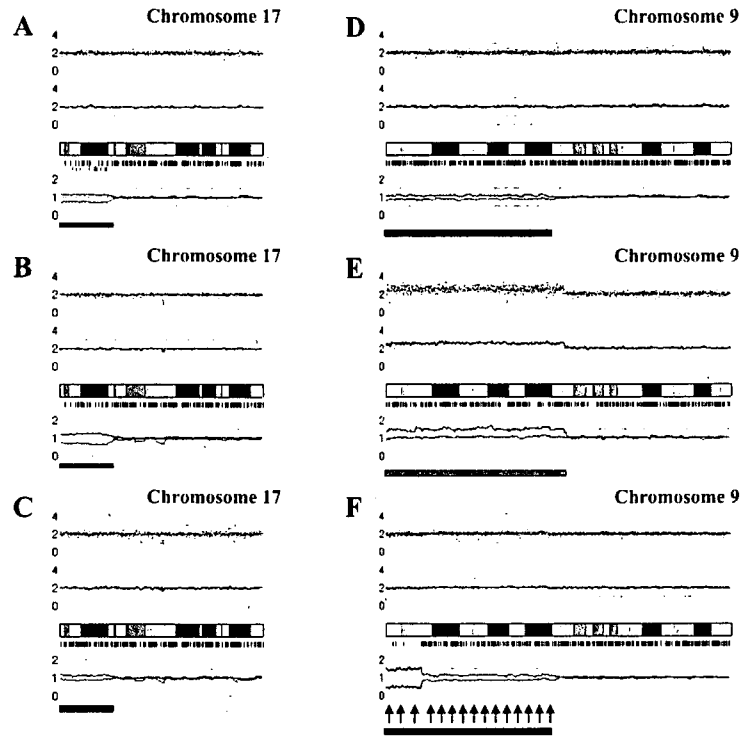

Figure 4. Detection of AI in samples of primary AML and MPD. AsCN analyses disclosed the presence of a small population with 17p UPD in a primary AML specimen (W150673) (93% blasts in microscopic examination) with either a paired sample (*A*) or anonymous reference samples (*B*). The difference of the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

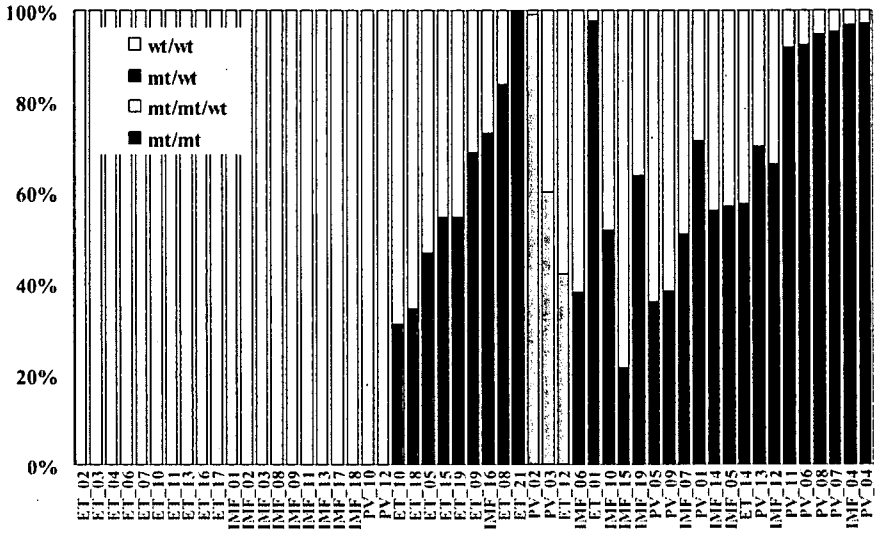

Figure 5. Estimation of tumor populations carrying 9p UPD and the *JAK2* mutation in MPD samples. The populations of 9p UPD–positive components in the 53 MPD cases were estimated by calculation of the mean difference of AsCNs within the UPD regions. Heterozygous (*blue bars*) or homozygous (*red bars*) *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.[20] 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, $\log_2 R^{refI}_{AB,i}$ 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

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

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

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

and

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

and the corrected total CN ratio is

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

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:

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

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

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

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

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

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

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

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

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 $\max(\log_2 \bar{R}^K_{A,i}, \log_2 \bar{R}^K_{B,i}) \leq 0.1$ or $\min(\log_2 \bar{R}^K_{A,i}, \log_2 \bar{R}^K_{B,i}) \geq -0.1$, where $\bar{R}^K_{A,i}$ and $\bar{R}^K_{B,i}$ 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,

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

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:

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

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

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

Appendix E
Algorithm for Detection of AI With or Without LOH

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

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

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

and

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

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

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

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

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

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

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

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

and

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

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

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

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

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