[CANCER RESEARCH 64, 3060–3071, May 1, 2004]

An Integrated View of Copy Number and Allelic Alterations in the Cancer Genome Using Single Nucleotide Polymorphism Arrays

Xiaojun Zhao,[1,4] Cheng Li,[2,5] J. Guillermo Paez,[1,3] Koei Chin,[7] Pasi A. Jänne,[1,3] Tzu-Hsiu Chen,[1] Luc Girard,[8,9] John Minna,[8,9] David Christiani,[6] Chris Leo,[1] Joe W. Gray,[7] William R. Sellers,[1,3] and Matthew Meyerson[1,4]

[1]Departments of Medical Oncology and [2]Biostatistical Sciences, Dana-Farber Cancer Institute, Boston, Massachusetts; [3]Departments of Medicine and [4]Pathology, Harvard Medical School, Boston, Massachusetts; [5]Departments of Biostatistics and [6]Environmental Health, Harvard School of Public Health, Boston, Massachusetts; [7]Department of Laboratory Medicine, University of California, San Francisco, California, [8]Hamon Center for Therapeutic Oncology Research, and [9]Departments of Internal Medicine and Pharmacology, University of Texas Southwestern Medical Center, Dallas, Texas

ABSTRACT

Changes in DNA copy number contribute to cancer pathogenesis. We now show that high-density single nucleotide polymorphism (SNP) arrays can detect copy number alterations. By hybridizing genomic representations of breast and lung carcinoma cell line and lung tumor DNA to SNP arrays, and measuring locus-specific hybridization intensity, we detected both known and novel genomic amplifications and homozygous deletions in these cancer samples. Moreover, by combining genotyping with SNP quantitation, we could distinguish loss of heterozygosity events caused by hemizygous deletion from those that occur by copy-neutral events. The simultaneous measurement of DNA copy number changes and loss of heterozygosity events by SNP arrays should strengthen our ability to discover cancer-causing genes and to refine cancer diagnosis.

INTRODUCTION

Chromosomal copy number alterations can lead to activation of oncogenes and inactivation of tumor suppressor genes (TSGs) in human cancers. These genes play key roles in multiple genetic pathways to positively and negatively regulate cell growth, proliferation, apoptosis, and metastasis (1). Many TSGs, including *RB1* (2), *p16* (3), and *PTEN* (4), were originally pinpointed by localizing regions of homozygous deletion. Similarly, regions of chromosome amplification frequently harbor oncogenes, such as *MYC* (5) and *ERBB2* (6). Thus, identification of cancer-specific copy number alterations will not only provide new insight into understanding the molecular basis of tumorigenesis but will also facilitate the discovery of new TSGs and oncogenes.

Comparative genomic hybridization (CGH) is a method to detect chromosomal copy number by comparing hybridization intensity of a tumor and a normal control DNA sample (7). Array-based CGH makes it possible to scan the genome for copy number with high resolution by hybridizing to arrayed genomic DNA or cDNA clones (8–10). To increase sensitivity and specificity, the hybridization of genomic representations to CGH arrays has been developed (11); this is particularly useful for oligonucleotide arrays. Lucito *et al.* (12) have developed recently a new representational oligonucleotide microarray that could achieve an average resolution of 30 kb across the genome. However, currently available array CGH methods cannot simultaneously detect chromosomal loss of heterozygosity (LOH). To combine the detection of cancer copy number with cancer-specific LOH in the same experiments, we have developed an analytical method to detect DNA copy number changes by hybridization of representations of genomic DNA to commercially available single nucleotide polymorphism (SNP) arrays.

SNPs are the most frequent form of DNA variation present in the human genome, where >2 million SNPs have been identified by public efforts.[10] Because of their abundance, even spacing, and stability across the genome, SNPs offer significant diagnostic potential for human diseases including cancers, compared with other polymorphisms such as fragment length polymorphisms and microsatellite markers. Moreover, scoring of SNPs is easily automated, *e.g.,* high-density oligonucleotide arrays have been used for large-scale high-throughput SNP analysis (13).

We and others have demonstrated previously that SNP arrays covering 1,494 SNP loci (HuSNP; Affymetrix) could accurately measure genome-wide LOH (14–19). LOH calls by SNP arrays were consistent with analysis using simple sequence length polymorphisms and CGH (15). Furthermore, our group has demonstrated that hierarchical clustering based on genome-wide LOH patterns can distinguish different types of tumor cells based on their shared LOH (20). A high-density SNP array has been generated recently that can analyze >10,000 SNP loci using a genome representation approach (21). The *Xba*I mapping array is highly robust and reproducible with call rate accuracy well in excess of 99% (21). We have shown that LOH analysis with this high-density array shares high concordance with microsatellite methods, and permits us to detect smaller regions of LOH that are missed by HuSNP array and microsatellite mapping (20).

In the present study, we demonstrate the utility of 10K SNP arrays for characterizing DNA copy number changes including amplification and homozygous deletion from a subset of lung and breast carcinoma cell lines and lung tumors.

MATERIALS AND METHODS

Genomic DNA

We obtained the following genomic DNA HCC1395, HCC1395 BL, HCC1187, HCC1187 BL, HCC1599, HCC1599 BL, HCC1143, HCC1143 BL, HCC38, HCC38 BL, HCC2218, HCC2218 BL, HCC1937, HCC1937 BL, HCC1007, HCC1007 BL, BT-474, UACC-812, and MCF7 from American Type Culture Collection. We obtained the following genomic DNA NA01723, NA04626, NA040695, NA06061, NA03226, and NA01201 from the National Institute of General Medical Sciences Human Genetic Mutant cell repository (Coriell Institute for Medical Research). We prepared genomic DNA from the following cell lines (American Type Culture Collection), NCI-H1648, NCI-BL10, NCI-H2141, NCI-BL2141, NCI-H1395, NCI-BL1395, NCIH128, NCI-BL128, NCI-H289, NCI-BL289, NCI-H2171, NCI-BL2171, NCI-H2107 NCI-BL2107, and the following primary lung carcinomas, 10372 (non-small cell lung carcinoma, unspecified), 18252 (non-small cell lung carcinoma, unspec-

Received 10/21/03; revised 1/27/04; accepted 2/17/04.

Grant support: National Cancer Institute Grants R01CA92824 (D. Christiani), P50CA70907 (J. Minna), 2P30 CA06516–39 (C. Li), 1K12CA87723–01 (P. A. Jänne), and CA58207 (K. Chin and J. W. Gray), the American Cancer Society RSG-03–240-01-MGO (M. Meyerson), the Pew Scholars in the Biomedical Sciences (M. Meyerson), the Flight Attendant Medical Research Institute (M. Meyerson), the Dunkin Donuts Rising Star Award (P.A. Jänne), and the Arthur and Rochelle Belfer Foundation (C. Li and M. Meyerson).

Note: Supplementary data for this article can be found at Cancer Research Online (http://cancerres.aacrjournals.org).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Requests for reprints: M. Meyerson, Department of Medical Oncology, Dana-Farber Cancer Institute, Mayer 430, 44 Binney Street, Boston, MA 02115. Phone: (617) 632-4768; Fax: (617) 632-5998; E-mail: matthew_meyerson@dfci.harvard.edu.

[10] Internet address: http://www.ncbi.nlm.nih.gov/SNP/.

ified), 57588 (squamous cell lung carcinoma), and 83437 (lung adenocarcinoma), using standard methods. The "BL" in cell line names refers to B-lymphoblast; in each cell line pair, the BL cell line serves as a normal control for the cancer-derived cell line.

*Xba*I Mapping Array Hybridization

*Xba*I mapping array 130 (Affymetrix, Inc., Santa Clara, CA) was used in this study. This array covers 10,043 SNP loci distributed on all of the human chromosomes except Y chromosome, resulting in a resolution close to 300 kb. The analyses were performed according to previously described methods (21) and the manufacturer's instructions. In brief, 250 ng of genomic DNA is digested with *Xba*I restriction enzyme, ligated to an adaptor, and amplified by PCR. The resulting amplicons are fragmented, labeled with biotinylated dideoxy ATP using terminal deoxynucleotidyl transferase, and hybridized to the array. Hybridization is detected by incubation with streptavidin-phycoerythrin conjugates, followed by scanning the array for phycoerythrin fluorescence and quantitation using the MAS 5.0 software.

Imaging and Data Analysis

Normalization of Arrays and Model-Based Signal Values. The median perfect match and mismatch probe intensities of the arrays range from 110 to 329, indicating the need for normalization to compare the signals across different arrays. We used the invariant set normalization method (22) to normalize all arrays at the probe intensity level to a baseline array "HCC1937 BL." This method adaptively selects probes that have similar ranks (thus more likely to belong to SNPs that have the same copy numbers) between one array and the baseline array to determine the normalization function.

After normalization, we used a model-based method (23) to obtain the signal values for each SNP in each array. Because the probe response patterns of the three genotypes (AA, BB, and AB) are dissimilar, we defined the new perfect match probe intensity as pmA + pmB and the new mismatch probe intensity as mmA + mmB for each probe quartet of a probe set. This transformation makes the probe intensity pattern and magnitude of a probe set comparable across the genotypes. Then the perfect match/mismatch difference model was applied on the transformed probe-level data to compute model-based signal values. The model-based method weighs probes by their sensitivity and consistency when computing signal values, and image artifacts are also identified and eliminated by the outlier detection algorithm in this step.

Observed and Inferred DNA Copy Number. For each SNP, the signal values of all of the normal cell lines were averaged to obtain the mean signal of 2 copy (male X chromosomes are multiplied by 2 before averaging), and the observed copy number is defined as (observed signal/mean signal of two copy) * 2, and visualized either log 2 ratio displayed in blue to white then to red color scale (Fig. 6*B*) or white (0 copy) to red color scales. In general, we assume a diploid genome in the absence of specific average DNA content data, but experimental values for mean copy number, derived from flow cytometry, can be substituted for the 2-copy assumption and will give more reliable results. To infer the DNA copy number from the raw signal data, we used the Hidden Markov Model (HMM; Ref. 24). First, we specify that for each SNP the observed signal values are random values drawn from a t distribution with parameters determined by the underlying real copy number (Fold*2) and the estimated mean signals and their SDs in the normal samples: (Signal − Mean * Fold / Std *Fold)~ $t(40)$. These distributions give the "emission probabilities" of the HMM. Secondly, we assume that the copy number changes are caused by genetic recombination events: for a particular sample, the larger the genetic distance between the two markers, the more likely it is that recombination (thus a copy number change) will happen within the interval. The Haldane's map function. $\theta = \frac{1}{2}(1 - e^{-2d})$; Ref. 25) is used to convert the genetic distance d between two SNP markers to the probability (2θ) that the copy number of the second marker will return to the background distribution of copy numbers in this sample and thus independent from the copy number of the first marker. These probabilities are used as the "transition probabilities" of HMM that determine how *d,* the real copy number of one marker, provides information of the real copy number of the adjacent marker. Thirdly, we estimate the background distribution of copy numbers in each sample in two rounds. The proportion of chromosome regions that have a particular copy number is set to fixed values in the first round [0.9 for 2 copy, $0.1/(N-1)$ for copy 0 to N except 2, where N is the maximal allowed copy number in inference]. The HMM is run as described below and then the inferred copy numbers are used to re-estimate the sample-specific background distribution of the copy numbers. After this, the HMM model is rerun to obtain the final results. These background distributions are used as the "initial probabilities" of HMM specifying the likelihood of observing a particular copy number at the beginning of the p-arm and also used together with the "transition probabilities" to determine the dependency of the copy number values of two adjacent markers as described above.

A HMM model with these probabilities specifies the joint probability of the unobserved copy number and the observed signal values, and the Viterbi algorithm (24) was then used to obtain the most likely underlying copy number path of SNPs in a chromosome (in the p-arm to q-arm ordering), given the observed signal values. The algorithm works by analyzing one chromosome at a time. The HMM was applied to all of the chromosomes and all of the samples separately, and the best paths were defined as the inferred DNA copy number values. The inferred copy number is visualized in the same way as the observed copy number.

The above described analysis methods are implemented in the dChip software Version 1.3 (26), which is freely available to academic users.

Array CGH Analysis

Bacterial artificial chromosome (BAC) DNA microarrays were obtained from the core facility at University of California San Francisco Comprehensive Cancer Center[11] and performed as described (8). The images were analyzed as described elsewhere (8). Data were normalized to the median raw CY3:CY5 ratio and converted to log base 2 to weight gains and losses equally.

Quantitative Real-Time PCR

Quantitative real-time PCR was performed on a PRISM 7700 sequence detector (Applied Biosystem, Foster City, CA) by using a QuantiTect SYBR Green kit (Qiagen, Inc., Valencia, CA). We have quantified each tumor DNA by comparing the target locus to the reference Line-1, a repetitive element for which copy numbers per haploid genome are similar among all of the human normal and neoplastic cells (27). Quantification is based on standard curves from a serial dilution of human normal genomic DNA. The relative target copy number level was also normalized to normal human genomic DNA as calibrator. Copy number change of target gene relative to the Line-1 and the calibrator were determined by using the formula $(T_{target}/T_{Line\text{-}1})/(C_{target}/C_{Line\text{-}1})$, where T_{target} and $T_{Line\text{-}1}$ are quantity from tumor DNA by using target and Line-1, and C_{target} and $C_{Line\text{-}1}$ are quantity from calibrator by using target and Line-1. PCRs for each primer set were performed in at least triplicate, and means were reported. Conditions for quantitative PCR reaction were as follows, one cycle of 50°C for 15 min, one cycle of 94°C for 2 min, 40 cycles of 94°C for 20 s, 56°C for 20 s, and 70°C for 20 s. At the end of the PCR reaction, samples were subjected to a melting analysis to confirm specificity of the amplicon. Primers were designed by using Primer 3[12] to span a 100–150-bp nonrepetitive region and were synthesized by Invitrogen (Carlsbad, CA). Each primer set was subsequently compared with the human genome using the basic local alignment search tool algorithm to determine its uniqueness. All of the primer sets were additionally confirmed to generate a single desired size amplicon evaluated by gel electrophoresis. For homozygous deletion, the presence or absence of PCR products was also evaluated by agarose gel electrophoresis. Primer sequences for each target used in this study are published as the supporting information (Supplementary Table 1).

RESULTS

Analysis of DNA Copy Number Changes. To evaluate the ability of SNP arrays to detect DNA copy number changes, we began by analyzing cell lines with defined DNA chromosomal copy number. Genomic DNA, isolated from cells containing one to five copies of the X chromosome, was digested, amplified, labeled, and hybridized to SNP arrays. After normalizing the total signal from each sample, we computed the probe signal representing each SNP. The ratio of each

[11] Internet address: http://ml.ucsf.edu/cores/arrays_bac.asp.
[12] Internet address: http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi.

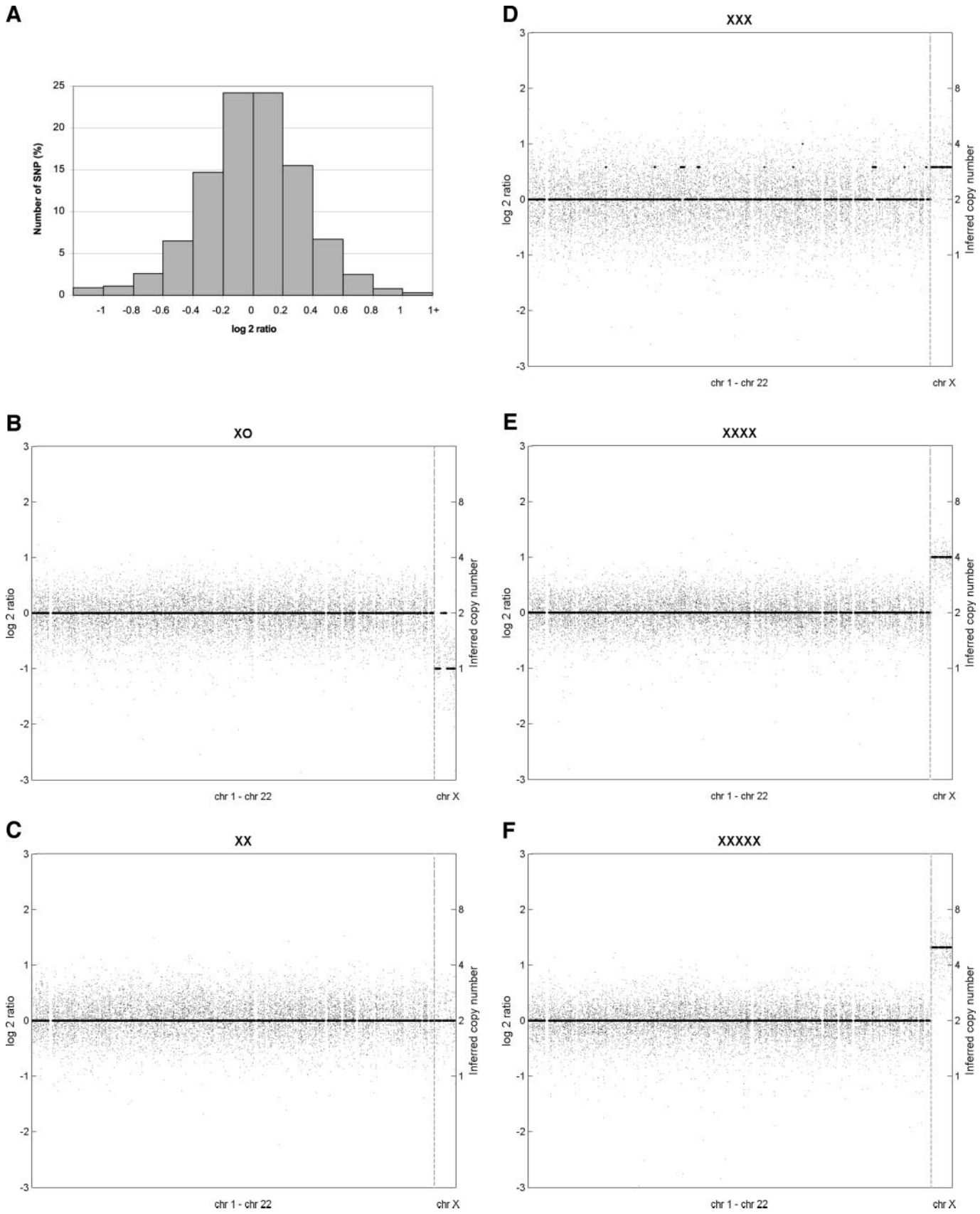

Fig. 1. Measurement of X chromosome copy number by single nucleotide polymorphism (*SNP*) array hybridization. *A,* the distribution of log 2 copy number signal ratios for all autosomal SNPs in DNA from XO, XX, 3X, 4X, and 5X cell lines. *B–F,* scatter plots of the log 2 copy number ratios (*black dots, left axis*) and inferred copy numbers from the dChip.SNP program (*black bars, right axis*) for single SNP array hybridizations of DNA from XO (*B*), XX (*C*), 3X (*D*), 4X (*E*), and 5X (*F*) cell lines. Comparisons are with respect to a reference of pooled normal DNA. Both *Y* axes are plotted against the physical position of the marker (*X* axis) in the genome, starting from chromosome 1 to chromosome X. Borders between the autosomal and X chromosomal SNPs are indicated by *dashed vertical bars*. Note that rare SNPs have extremely low hybridization signal (log2 < -3) in each cell line DNA, possibly due to *Xba*I polymorphism or probe hybridization failure; these SNPs cannot be visualized in this graph.

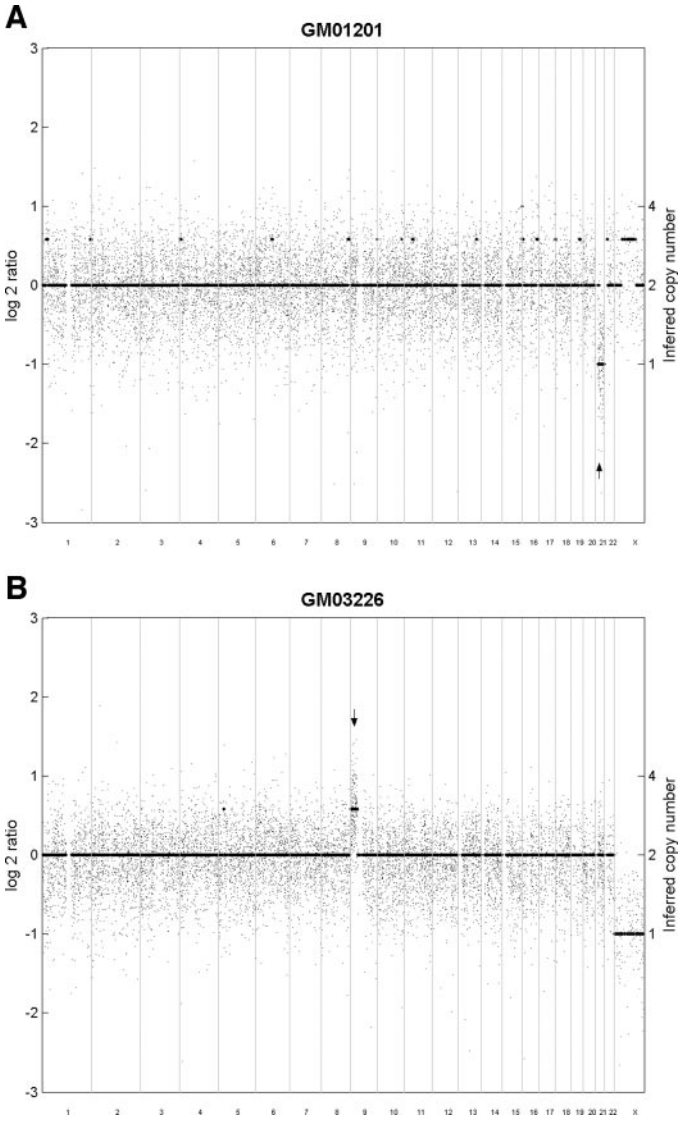

Fig. 2. Measurement of single-copy autosomal changes by single nucleotide polymorphism (SNP) array hybridization. Scatter plots of the log 2 copy number ratios (*black dots, left axis*) and inferred copy numbers from the dChip.SNP program (*black bars, right axis*) for single SNP array hybridizations of DNA from GM02101 and GM03236 cell lines. SNPs from each chromosome are separated by a *vertical line*. *A,* single-copy loss of chromosome 21 in cell line GM01201 (*arrow*). *B,* single-copy gain of 9pter to 9p13 in cell line GM03236 (*arrow*).

SNP hybridization signal was then computed for the XO, XX, XXX (3X), XXXX (4X), and XXXXX (5X) cell line genomic DNAs with respect to an average of normal XX genomic DNA samples (Fig. 1). The log 2 ratios of the raw signal for all 9684 of the SNPs located on autosomes were distributed normally with 50% of the absolute magnitude of the log 2 ratios <0.21 and 75% of the absolute ratios <0.37 (Fig. 1*A*).

The observed copy numbers of all 178 X-chromosome SNPs on the array were positively correlated with the known X chromosome copy number of the cell lines. Fifty percent of X-chromosome SNPs had correlation coefficients with known copy number >0.95, and 75% of X-chromosome SNPs had correlations >0.89. Furthermore, the relationship between observed and predicted copy number for X-chromosomal SNP loci from these different experiments fits a linear regression with an R^2 correlation of 0.9961, arguing that the SNP array-based signal fits the actual copy number within this range (Supplementary Fig. 1).

This correlation suggests that a quantitative model could be developed to predict chromosomal copy number based on SNP array hybridization intensity. To do so, we have implemented a novel analytical method to infer the copy number of each SNP based on a hidden Markov model within the dChipSNP computational platform (see details in "Materials and Methods").

The observed copy number ratios for X-chromosome SNPs (scatter plots) and the inferred copy numbers (black bars) correlated with the known copy number for each of the known karyotypes (Fig. 1, *B–F*). The XO genomic DNA showed a decreased X-chromosome hybridization signal compared with XX controls but little or no change in the raw and inferred autosomal signals (Fig. 1*B*). The experimental XX cell line DNA showed a constant inferred copy number of 2 for both the autosomes and the X chromosome (Fig. 1*C*), whereas the X-chromosomal signal was increased in samples with 3, 4, or 5 copies of the X chromosome, as reflected by inferred copy numbers of 3 for 3X, 4 for 4X, and 5 for 5X (Fig. 1, *D–F*). Overall, the inferred DNA copy number for the autosomes was accurately predicted as diploid for 99.2% of SNPs. This result suggests that the inferred copy number calculations based on SNP array hybridization intensity approximate closely to the actual copy number.

To additionally validate our ability to measure DNA copy changes from autosomes, we measured two otherwise diploid cell lines containing cytogenetically mapped partial or whole-chromosome copy number gains or losses. SNP array hybridization analysis shows both decreased raw SNP hybridization ratios (black dots) and an inferred copy number (black bar) of 1 for the entire chromosome 21 (Fig. 2*A*, red arrow), which is lost in the GM01201 cell line. This analysis also shows increased hybridization intensities and a copy number gain to 3 copies within chromosome 9p (Fig. 2*B,* red arrow) for which the GM03236 cell line is triploid. The inferred copy number analysis showed that 96.7% and 99.9% of the SNP loci from the two cell lines were predicted as two copies, in these otherwise diploid cell lines.

Detection of Chromosome Amplifications in Cancer Cell Line DNA. Because the above analyses from samples of known copy number demonstrate that SNP array hybridization and the dChip model can detect DNA copy number changes to a reasonable degree of accuracy, we applied quantitative analysis of SNP array hybridization to human cancer-derived samples. In total, we examined 18 lung and breast cancer cell line DNA samples together with 15 normal blood control cell line DNA samples; 3 cell line DNA samples were unmatched. When analyzed in arbitrary units (perfect match-mismatch), the median intensity by array ranged from 110 to 329. For a typical array "HCC1187 BL," the median probe intensity was 147, with the 10^{th} percentile at 68 and the 90^{th} percentile at 567.

Quantitative analysis of SNP array data from the cancer cell line samples revealed a variety of candidate copy number alterations, including both low-level and high-level amplifications, as well as hemizygous and homozygous deletions. Copy number analyses were similar regardless of whether the reference sample was paired normal DNA or pooled normal DNA (data not shown). Raw data are available at our website.[13]

An example of cancer-specific amplification is shown for the male small cell lung cancer cell line, H2171 (Fig. 3). A genome-wide view reveals a variety of large regions with triploid or haploid DNA content, including a single-copy X chromosome as expected and several regions of high-copy amplification (Fig. 3*A*). These include inferred copy numbers of ≥ 7 in a 1.7–2.6 megabase region of chromosome 8q12 and in a 1.1–2.1 megabase region of chromosome 8q24

[13] Internet address: http://research.dfci.harvard.edu/meyersonlab/snp/snp.htm.

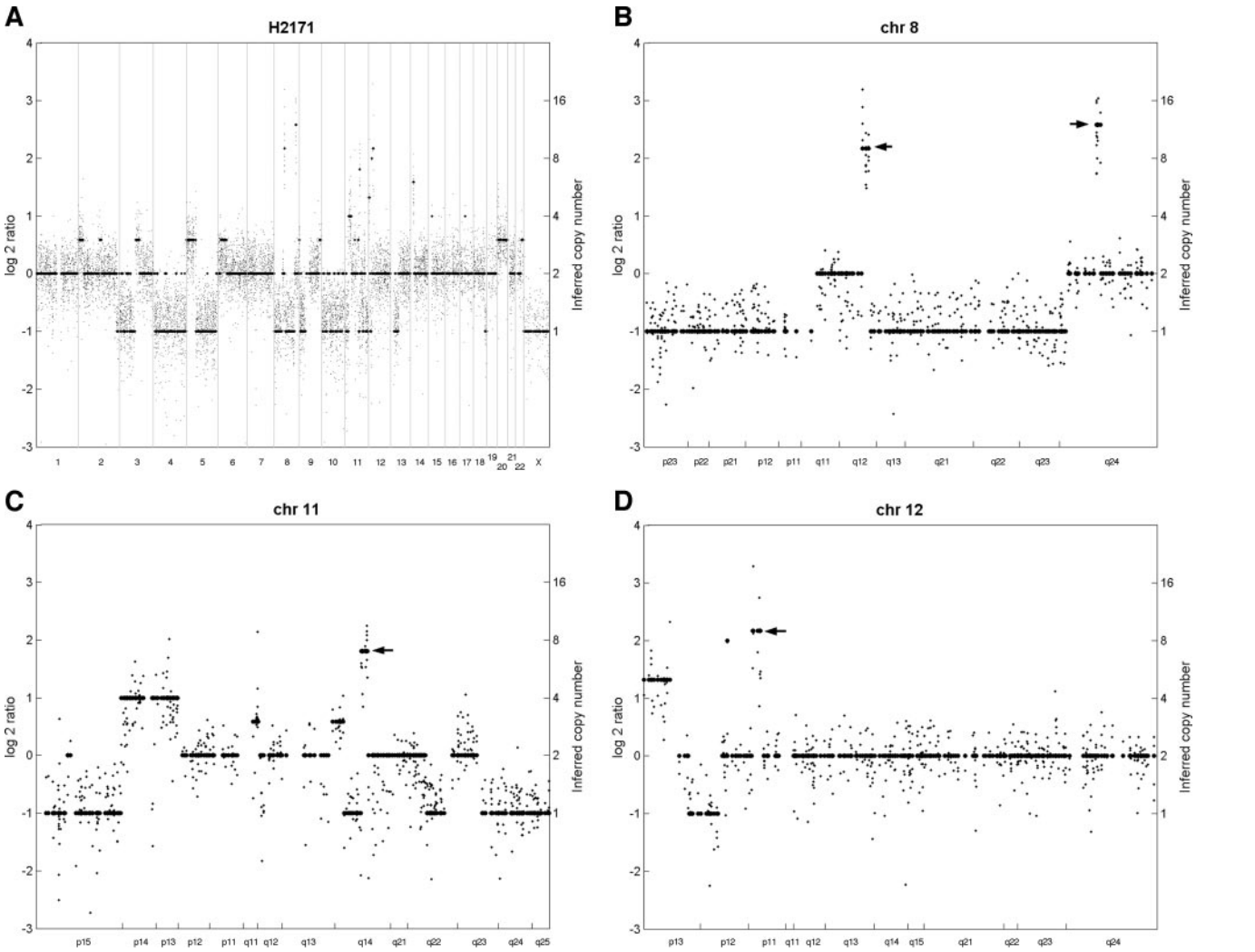

Fig. 3. Single nucleotide polymorphism array analysis of chromosome amplification in the NCI-H2171 small cell lung carcinoma cell line. *A,* copy number analysis across the autosomes and X chromosome. Regions of copy number amplification ≥7, as inferred by dChip.SNP are shown by *black arrows* for (*B*) chromosome 8, (*C*) chromosome 11, and (*D*) chromosome 12; each of these regions was validated by quantitative real-time PCR (Table 1). Cytobands are indicated in the *X* axis for each chromosome.

Table 1 *Predicted regions of amplification*[a]

Cell line or tumor	Size range[b] (Mb)	Cytoband[c]	Candidate gene[d]	Inferred copy number[e]	Measured copy number[f]
H1395	0.06–0.47	8q24.12	*NOV*	9	43.23
H1395	0–0.73	8q24.21	*MYC*	7	37.45
H1395	0–0.62	20q11.23-q13.11		9	60.63
H2171	1.72–2.63	8q12.1-q12.3		9	33.80
H2171	1.08–2.08	8q24.13-q24.21	*MYC*	12	43.57
H2171	1.67–2.11	11q14.1-q14.2		7	14.52
H2171	1.86–3.21	12p11.23-p11.22		9	23.14
HCC1143	0.46–6.3	11q13.1-11q13.4	*CCND1*	7	25.44
HCC1143	1.98–3	12q14.3-q15	*DYRK2*	9	10.05
HCC1599	4.5–5.76	19q12-q13.12	*CCNE1*	7	13.48
HCC2218	0.13–0.99	17q11.2		11	32.90
HCC2218	1.69–2.82	17q25.1		9	22.97
BT-474	2.07–4.64	17q12-q21.2	*ERBB2*	9	35.27
BT-474	2.74–5.12	20q13.2-q13.31	*BCAS1*	14	36.91
UACC-812	1.88–4.7	13q14.2-q14.3		7	5.15
UACC-812	7.29–8.59	13q21.31-q21.33		7	14.86
UACC-812	3.33–4.1	13q22.2-q31.1		7	8.64
UACC-812	2.73–3.24	13q31.3		10	21.97
MCF7	2.84–3.74	3p14.2-p14.1		9	23.90
MCF7	0.78–2.46	20q13.2-20q13.31	*BCAS1*	13	35.89
10372 (tumor)	5.91–6.21	12q12-q13.11		7	9.85

[a] Predicted regions containing at least two single nucleotide polymorphisms (SNPs) with inferred copy number ≥7.
[b] Based on hg12 human genome assembly. Minimal and maximal size range based on the SNP positions and locations described in Supplementary Table 2.
[c] Based on hg12 position (Supplementary Table 2).
[d] Known or previous candidate oncogene within maximal region.
[e] Inferred by dChip analysis of SNP arrays.
[f] Measured by quantitative real-time PCR with candidate or randomly selected region with reference to LINE-1 control. The locus tested and primer sequences are described in Supplementary Table 1.

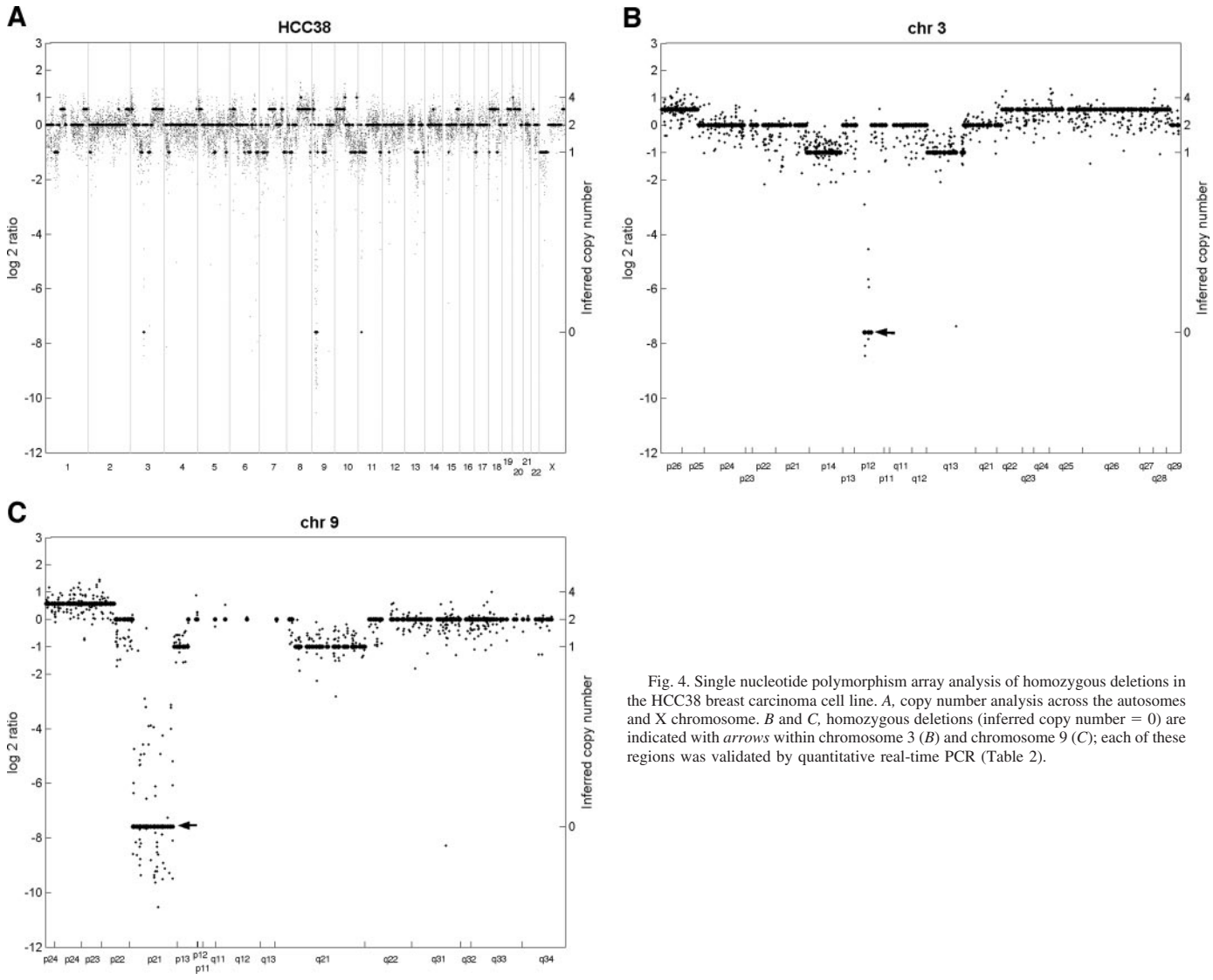

Fig. 4. Single nucleotide polymorphism array analysis of homozygous deletions in the HCC38 breast carcinoma cell line. *A,* copy number analysis across the autosomes and X chromosome. *B* and *C,* homozygous deletions (inferred copy number = 0) are indicated with *arrows* within chromosome 3 (*B*) and chromosome 9 (*C*); each of these regions was validated by quantitative real-time PCR (Table 2).

Table 2 *Predicted regions of homozygous deletion*[a]

Cell line or tumor	Size range[b] (Mb)	Cytoband[c]	Candidate gene[d]	Inferred copy number[e]	Measured copy number[f]
NCI-H1648	0.21–0.5	3p14.2	*FHIT*	0	0.00012
NCI-H1648	0.39–0.68	9p21.3	*P16*	0	0.00021
NCI-H1648	1.88–4.5	Xq21.31-Xq21.33		0	0.00090
NCI-H2141	0.26–0.83	10p12.1		0	0.015
HCC1187	0.05–1.01	14q23.2		0	0.0069
HCC1599	4.06–4.16	4q35.1-q35.2		0	0.00028
HCC1937	0.76–2.04	10q21.3		0	0.0012
HCC38	2.48–5.86	3p12.3-p12.2		0	0.00019
HCC38	10.05–10.72	9p21.3-p21.1	*P16*	0	0.00063
HCC1395	0.43–2.03	6q16.1		0	0.00033
HCC1395	2.88–4.43	6q16.3-q21		0	0.000081
HCC1395	0.12–1.6	11p13-p12		0	0.000083
HCC1395	7.99–10.37	13q14.3-q21.2		0	0.0013
HCC1395	0.74–2.53	Xq21.1-q21.2		0	0.000083
MCF7	0.22–0.4	3q13.31		0	0.12
10372	0.84–1.01	1p13.1-p12		0	2.31
10372	1.83–6.04	19p13.3		0	1.21

[a] Predicted regions of at least 1 kb in size containing at least two single nucleotide polymorphisms (SNPs) with inferred copy number = 0.
[b] Based on hg12 human genome assembly. Minimal and maximal size range based on the SNP positions and locations described in Supplementary Table 3.
[c] Based on hg12 position (Supplementary Table 3).
[d] Known or previous candidate tumor suppressor gene within maximal region.
[e] Inferred by dChip analysis of SNP arrays.
[f] Measured by quantitative real-time PCR with candidate or randomly selected region with reference to LINE-1 control. The locus tested and primer sequences are described in Supplementary Table 1.

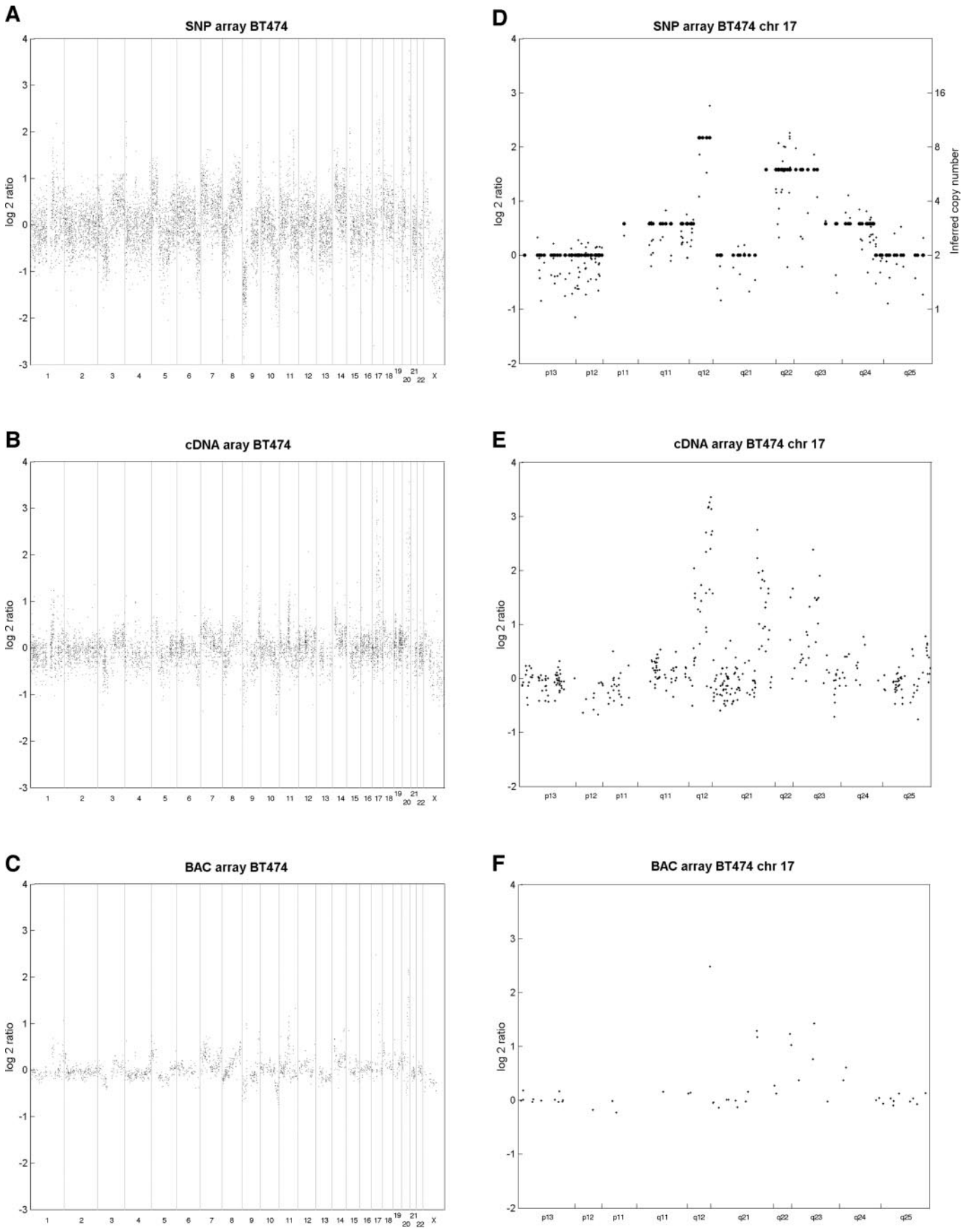

Fig. 5. Comparison of single nucleotide polymorphism (*SNP*) array, cDNA array, and bacterial artificial chromosome (*BAC*) array log 2 copy number ratios (*scatter plots*) for BT474 breast carcinoma cell line DNA. All clones are mapped based on the hg12 assembly. *A–C,* autosomes and X chromosome; *D--F,* chromosome 17; and *G–I,* chromosome 20. *A*, *D*, and *G,* SNP array; *B*, *E*, and *H,* cDNA array; and *C, F*, and *I,* BAC array. *G–I,* hemizygous deletion within 20q12 deletion is indicated by the *arrow*.

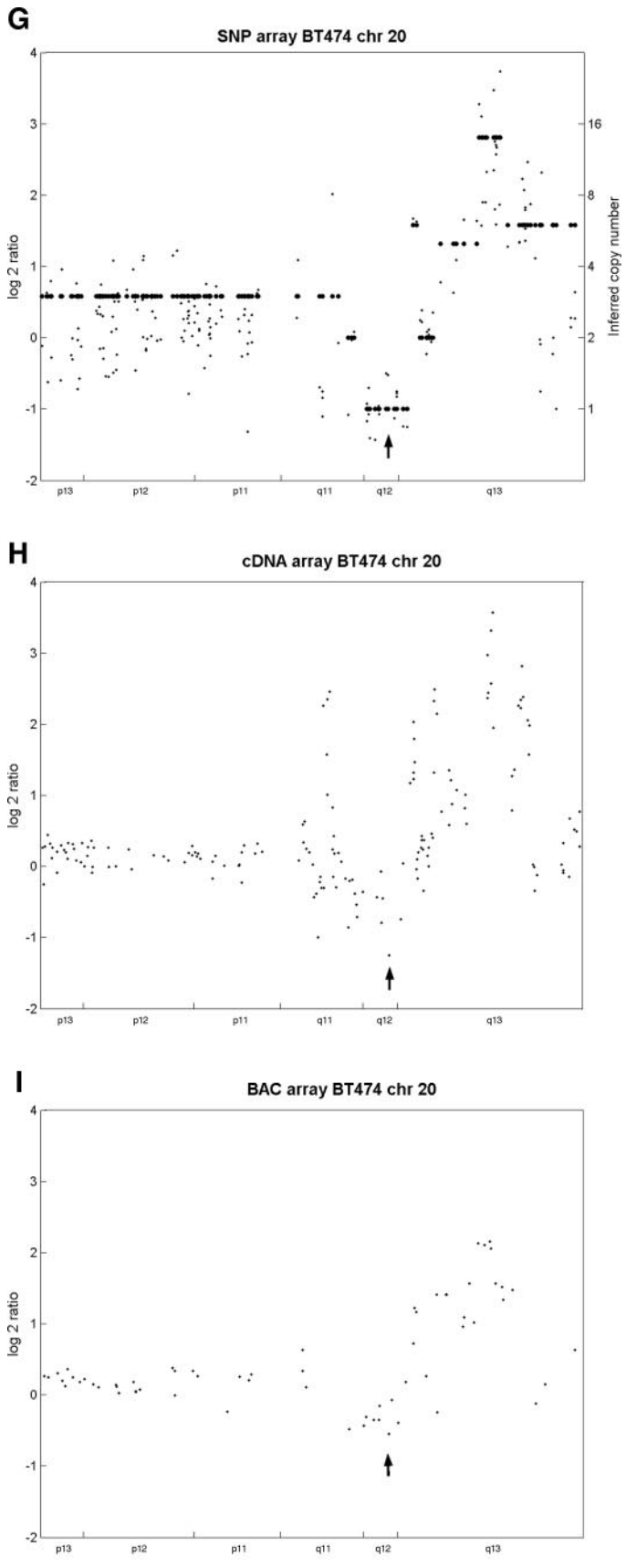

Fig. 5. *Continued.*

encompassing the *MYC* locus (Fig. 3*B*; Table 1), a 1.7–2.1 megabase region of chromosome 11q14 (Fig. 3*C*; Table 1), and a 1.9–3.2 megabase region of chromosome 12p11 (Fig. *3D*; Table 1).

In total, the SNP array hybridization identified 21 candidate regions of high-level DNA amplification, arbitrarily defined as an inferred copy number of ≥7 (Table 1). Two regions, one encompassing *MYC* (NCI-H1395 and NCI-H2171) and another encompassing the *BCAS1* gene (BT-474 and MCF7) were subject to predicted high-level amplification in more than one sample. Cancer cell-specific chromosome amplifications also included some other known regions of amplification, including regions of 11q13.3 (*CCND1*), 19q12-q13.12 (*CCNE1*), and 12q14.1-q15 (*DYRK2*; Table 1). Moreover, SNP array quantitation of DNA from the BT-474 and UACC-812 breast cancer cell lines could detect recurrent amplifications in chromosomes 17 and 20. For example, the *ERRB2* amplicon is seen to be highly amplified in UACC-812 (data not shown).

Additionally, we were able to detect several novel amplicons, including amplification of the *NOV* gene in NCI-H1395 cell DNA and a large amplicon in UACC- 812 cells from 13q14.2 to 13q31.3 with copy number as high as 11 (Table 1). The resolution of amplification detection will depend on the density of the SNP array used, but we have identified high-copy amplifications of <500 kb in maximum size and confirmed these amplifications by quantitative real-time PCR (Table 1; Supplementary Table 1).

The copy numbers of the predicted amplified regions were validated by quantitative real-time PCR (Table 1). The magnitude of the amplification was generally underestimated by the SNP array hybridization intensity. The SNP array inferred copy number of the tested regions ranged from 7 to 12, whereas the quantitative PCR-derived copy number ranged from 5.15 to 60.63. This general underestimation most likely reflects the saturation of the SNP arrays at high copy number, but it is conceivable that local copy number variations between the SNP locus and the quantitative PCR locus may also contribute to the discrepancy. Additionally, we have additionally confirmed that 1.5–3-fold (3–6 copy) changes in copy number could be predicted with reasonable accuracy (Supplementary Fig. 2). For example, we evaluated amplification of the *MYC* locus in several samples with lower predicted copy numbers for the region. Samples with predicted copy numbers of 3 had a mean of 3.04 with a SD of 0.59 by quantitative PCR, whereas samples with SNP array predicted copy numbers of 4 had a mean of 4.80 with a SD of 1.38 by quantitative PCR.

SNP Array Identification of Homozygous Deletions. SNP array analysis is able to detect homozygous deletions in cancer cell line DNA in genome-wide scans. Two examples of homozygous deletions from a breast cancer cell line, HCC38, are shown in Fig. 4. Our criteria for homozygous deletion require the presence of at least 2 SNPs that cover an area of >1 kb in addition to an inferred copy number of 0; these eliminate candidate regions that may be caused by *Xba*I polymorphism together with LOH. The HCC38 cell line contains three regions with an inferred copy number of 0 (Fig. 4*A*), including larger regions on chromosome 3p12 (Fig. 4*B*; Table 1) and chromosome 9p21 (Fig. 4*C*; Table 1), as well as one small region, that do not meet the criteria. The chromosome 3p12 deletion has been described previously (28), whereas the 9p21.3-p21.1 deletion, encompassing the *CDKN2A* locus, has not been reported previously. The median observed copy number for the two regions of homozygous deletions predicted in HCC38 (Fig. 4, *B* and *C*) was 0.0082 (log 2 ratio of −7.6 compared with diploid) and the 95th percentile of the observed copy number was 0.46 (log 2 ratio of −2.2 compared with diploid).

In total, 15 candidate regions of homozygous deletions were identified by the dChipSNP algorithm (Table 2); one region, the CDKN2A locus on chromosome 9p21 was identified in two samples. Thirteeen

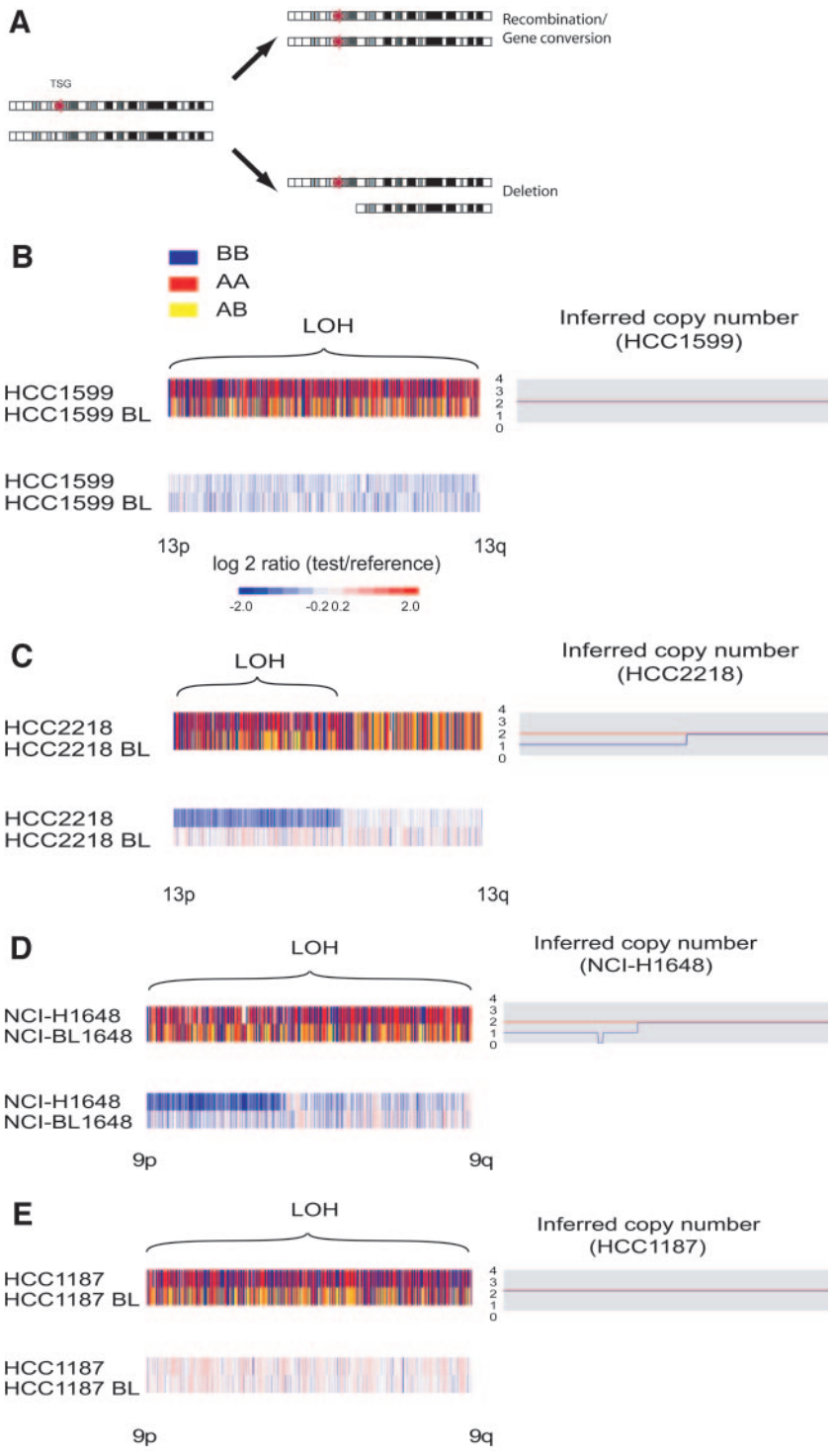

Fig. 6. Single nucleotide polymorphism array analysis can distinguish different genetic mechanisms that lead to loss of heterozygosity (*LOH*). *A,* different chromosomal mechanisms can cause LOH. In this cartoon, a mutated tumor suppressor gene indicated by a *red star* become homozygous because the remaining wild-type allele is lost by a copy-neutral event such as recombination or gene conversion (*top right*) or by hemizygous deletion (*bottom right*). *B* and *C,* LOH analysis (*top left*), copy number analysis (*bottom left*), and copy number quantitation of chromosome 13 in cell lines HCC1599 and HCC2218. In the LOH panel, *yellow* denotes heterozygosity (*AB*), whereas *red* (*AA*) and *blue* (*BB*) denote homozygosity. Note that LOH regions show heterozygous SNP markers in the normal that are reduced to homozygosity in the cell line. *D* and *E,* LOH within chromosome 9 associated with copy number loss and an interstitial homozygous deletion in NCI-H1648, contrasted with LOH and copy number maintenance in HCC1187.

of the 15 candidates could be confirmed by quantitative and non-quantitative PCR analysis. These included previously unreported candidate regions in chromosomes 4q35.1–4q35.2, 10q21.3, 14q23.2, and Xq21.32-Xq21.33, as well as previously described regions in 3p12, 3p14, and the *CDKN2A* locus on 9p21 (Table 2). Candidate homozygous deletion regions from primary tumors could not be validated in this study (Table 2), consistent with the need for purified samples for homozygous deletion detection (see Fig. 7, below). The smallest region of homozygous deletion that was confirmed by quantitative real-time PCR was a maximum of 502 kb in size (Table 2).

Comparison between Copy Number Analyses Using SNP, cDNA, and BAC Arrays. To evaluate the performance of SNP arrays, we compared raw copy number ratios across the genome for DNA from the BT474 breast cancer cell line using the SNP array (our data), cDNA array (9), and BAC array platforms. This cell line was chosen because it has well-characterized amplicons within chromosomes 17 and 20 (8, 9). The correlation coefficients (*R*) between each pair of the three platforms are 0.70 (SNP array *versus* BAC array), 0.62 (SNP array *versus* cDNA array), and 0.76 (BAC array *versus* cDNA array). Each platform provided a generally similar view of the

Table 3 *Confirmation of copy number changes in LOH[a] region by quantitative real time PCR*

Marker	Cell line	Location[b] Mb (Chromosome)	Inferred copy number[c]	Measured copy number[d]
RB1	HCC2218	47 (13)	1	0.95
RB1	HCC1599		2	1.55
TBC1D4	HCC2218	74 (13)	2	2.35
TBC1D4	HCC1599		2	1.95
TPP2	HCC2218	102 (13)	2	1.66
TPP2	HCC1599		2	1.51
SLC1A1	NCI-H1648	5 (9)	1	1.00
SLC1A1	HCC1187		2	2.31
CDKN2A	NCI-H1648	22 (9)	0	0.00021
CDKN2A	HCC1187		2	1.58
GSN	NCI-H1648	115 (9)	2	1.63
GSN	HCC1187		2	2.32

[a] LOH, loss of heterozygosity.
[b] Location based on hg12 human genome assembly. The loci tested are described in Supplementary Table 1.
[c] Inferred by dChip analysis of single nucleotide polymorphism arrays.
[d] Measured by quantitative real-time PCR of candidate regions with reference to LINE-1 control. The primer sequences are described in Supplementary Table 1.

genome (Fig. 5, *A–C*). All types of arrays readily detected two broad amplicons within 17q12 and 17q22–23 (Fig. 5, *D–F*). Similarly, an amplicon located in 20q13 and a region of single-copy loss located in 20q12, confirmed previously by BAC array CGH and cDNA array CGH (8, 9), were also observable in all three of the arrays (Fig. 5, *G–I*). In addition, we compared copy number data between SNP array and BAC array for a breast cancer cell line, HCC1937, with multiple copy number changes (29). Both SNP and BAC arrays detected a remarkably similar pattern of copy number changes (data not shown). In conclusion, we find that SNP arrays, cDNA arrays, and BAC arrays detect generally the same types of copy number variation in the same locations.

Analysis of Distinct Mechanisms Involved in LOH. Many TSGs are inactivated by a recessive mutation in one allele followed by the loss of the other wild-type allele (30). Thus, a tumor cell can arise after LOH at a relevant TSG locus. An analysis of the *RB1* locus in retinoblastoma led to the proposal that a variety of different genetic events underlie LOH, including point mutation, hemizygous deletion, mitotic nondisjunction, and mitotic recombination and gene conversion (31). Whereas hemizygous deletion leads to copy number reduction (Fig. 6*A,* right lower diagram), the other LOH mechanisms do not lead to DNA copy number changes (Fig. 6*A*, right upper diagram).

Because high-density SNP arrays can efficiently detect both copy number changes and LOH, we reasoned that we could discriminate between underlying LOH mechanisms by analyzing copy number changes. We observed that some LOH regions do not exhibit copy number changes. For example, in HCC1599 and HCC1187 cell lines, the entire chromosome 13 and 9, respectively, undergo LOH as detected by genotyping analysis, but there is no change in copy number (Fig. 6*B* and 6*E*), suggesting that these LOH events could be caused by copy-neutral events such as mitotic nondisjunction followed by duplication of one parental chromosome. In contrast, chromosome 13pter-q22 undergoes LOH in HCC2218, and copy number analysis indicates loss of one copy (Fig. 6*C*), whereas the remainder of chromosome 13 shows retention of heterozygosity and a diploid copy number, suggesting that this LOH event might be caused by hemizygous deletion.

Similarly, all of chromosome 9 undergoes LOH in NCI-H1648, but copy number analysis suggests that the LOH on 9p is due to hemizygous deletion, whereas the LOH on 9q is due to a copy-neutral mechanism (Fig. 6*D*). Each of the copy number values of the regions of LOH or retention, described above, was confirmed by quantitative real-time PCR of selected loci (Table 3).

Analysis of Copy Number Changes in Mixed Samples and Primary Tumors. The detection of copy number changes in tumor DNA could be confounded by the presence of DNA from surrounding

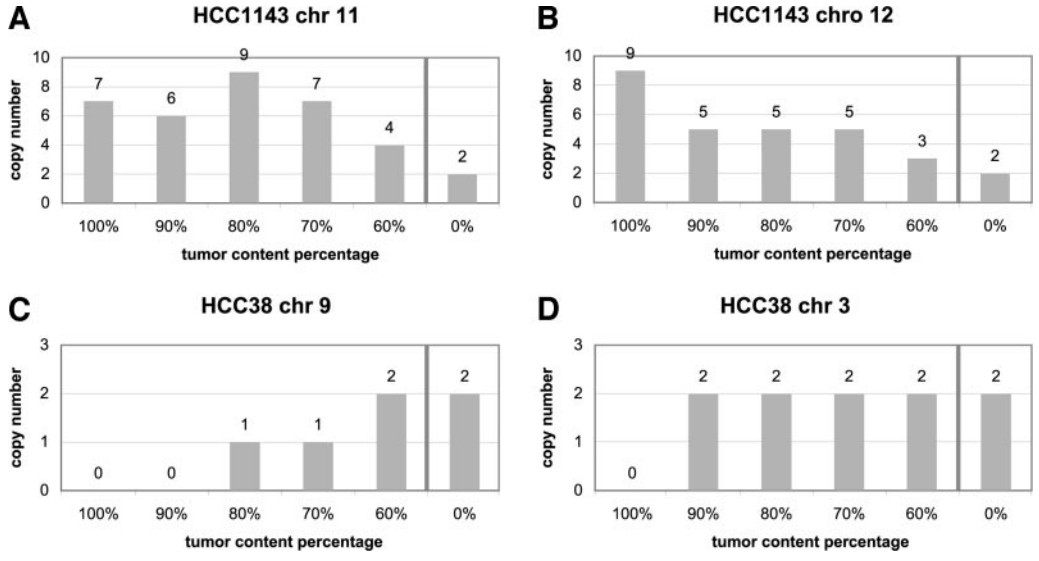

Fig. 7. Mixing experiment shows the effect of tumor DNA content on single nucleotide polymorphism array detection of amplifications and homozygous deletions. *A* and *B,* detection of amplifications on chromosomes 11 (*A*) and 12 (*B*) in a mixture of HCC1143 tumor cell line DNA (100% to 60% to 0%) with HCC1143 BL control cell line DNA. *C* and *D,* detection of homozygous deletions in chromosome 9 (*C*) and chromosome 3 (*D*) in a mixture of HCC38 tumor cell line DNA (100% to 60% to 0%) with HCC38 BL control cell line DNA.

non-neoplastic tissue. We have previously performed tumor-normal mixing experiments to assess the effects of contaminating non-neoplastic cells on LOH calls using SNP arrays and found that the best performance was achieved with 90% tumor purity and above (15). To determine the utility of SNP array for DNA copy number analysis in tumor samples, we performed mixing experiments using two cancer cell lines, HCC1143 and HCC38, each with two confirmed amplifications and homozygous deletions (Tables 1 and 2), and the corresponding B-lymphoblast "normal control" cell lines, HCC1143 BL and HCC38 BL. DNA from each tumor cell line was mixed with matched normal DNA at ratios of 0, 60, 70, 80, 90, and 100% tumor cell line DNA. Copy number amplifications could still be detected in mixtures containing 60% tumor DNA (Fig. 7, *A* and *B*). However, the accuracy of scoring homozygous deletions with dChipSNP drops off steeply with decreasing purity of tumor where the deletion of 9p21 in HCC38 cell line is detectable with tumor DNA purity of ≥90% (Fig. 7*C*), whereas the deletion of chromosome 3p12 is detectable only at 100% tumor purity (Fig. 7*D*). These results suggest that, unlike amplification, detection of homozygous deletion is considerably dependent on tumor purity.

Next, we tested our copy number analysis approach on five primary lung tumor samples. We detected one chromosomal amplification in a primary lung tumor, which was subsequently confirmed by real-time quantitative PCR (Table 1), whereas two homozygous deletions detected in these tumors appear to be false positive (Table 2). These results suggest that whole genome amplification of tumor sample dissected from laser capture microdissection will be the best approach to isolate DNA from primary tumor samples.[14]

DISCUSSION

DNA copy number changes, such as amplifications and deletions, frequently cause oncogene activation and TSG inactivation in cancer. We have shown above that hybridization to arrays of >10,000 SNPs can effectively detect homozygous deletions, hemizygous deletions, and amplifications simultaneously with LOH detection.

This study represents the first application of SNP arrays in genome-wide screening for DNA copy number changes in human cancers. Comparison with BAC and cDNA array analysis shows that the three platforms give generally comparable results. The noise of individual measurements is generally lower using BAC arrays, but the possible density of markers is greater with SNP or other representational oligonucleotide arrays (arrays representing 120,000 SNPs have now been generated). Furthermore, the SNP array approach offers the unique possibility to analyze copy number and LOH simultaneously using the same platform. Thus, this makes it possible to distinguish copy- reducing from copy-neutral genetic mechanisms underlying LOH events.

As part of this work, we have developed the signal analysis module in the dChipSNP platform, which is highly automated and freely available to the scientific community, for copy number analyses and for correlating copy changes readily with cytoband and gene information.[15] These analytic methods could also be adapted to other copy number platforms. Upon further refinement, the SNP array methods should also permit analysis of allele-specific amplification.

Many tumor suppressor and oncogene loci have been identified by pinpointing recurrently deleted or amplified chromosomal regions. CGH, fluorescence *in situ* hybridization, and other techniques have revealed many recurrent copy number changes in a variety of tumors. In this study, we have identified many known regions, such as homozygous deletion of chromosome 9p21 and amplification of chromosomes 8q24 and 17q21. These regions harbor well-characterized TSGs such as *CDKN2A* as well as oncogenes such as *MYC* and *ERBB2*, which are implicated in lung and breast tumorigenesis. In addition, we discovered several novel homozygous deletions and high-level amplifications (Tables 1 and 2). Although the interpretation of these regions must be cautious given the presence of genome instability in cancer cell lines, the surveying of additional cancer specimens will help to address their significance.

The high density of SNP arrays may also make possible the characterization of haplotype structures to analyze cancer predisposition. Furthermore, the detection of single-copy changes with SNP arrays suggest that these arrays could be used to study other genetic diseases in addition to cancers, such as Down, Prader Willi, Angelman, and cri du chat syndromes. The SNP arrays may find application as diagnostic as well as research reagents in this area.

In conclusion, we have demonstrated that SNP array hybridization is a highly efficient method for evaluating genome-wide copy number changes. Whereas the novel deleted and amplified regions discovered in this study may already be significant, application of the SNP array approach to large cancer data sets should prove highly fruitful in discovering cancer-specific genomic alterations.

ACKNOWLEDGMENTS

We thank Pamela Mole, Maura Berkeley, and Dr. Ed Fox at Dana-Farber Cancer Institute microarray facility for valuable assistance with *Xba*I mapping arrays. We thank Dr. Adi Gazdar for discussions and tumor cell lines and Jiajun Gu for preparing figures.

REFERENCES

1. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell 2000;100:57–70.
2. Friend SH, Bernards R, Rogelj S, et al. A human DNA segment with properties of the gene that predisposes to retinoblastoma and osteosarcoma. Nature 1986;323:643–6.
3. Kamb A, Gruis NA, Weaver-Feldhaus J, et al. A cell cycle regulator potentially involved in genesis of many tumor types. Science 1994;264:436–40.
4. Li J, Yen C, Liaw D, et al. PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer. Science 1997;275:1943–7.
5. Little CD, Nau MM, Carney DN, Gazdar AF, Minna JD. Amplification and expression of the c-myc oncogene in human lung cancer cell lines. Nature 1983;306:194–6.
6. Di Fiore PP, Pierce JH, Kraus MH, Segatto O, King CR, Aaronson SA. erbB-2 is a potent oncogene when overexpressed in NIH/3T3 cells. Science 1987;237:178–82.
7. Kallioniemi A, Kallioniemi OP, Sudar D, et al. Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors. Science 1992;258:818–21.
8. Pinkel D, Segraves R, Sudar D, et al. High resolution analysis of DNA copy number variation using comparative genomic hybridization to microarrays. Nat Genet 1998; 20:207–11.
9. Pollack JR, Perou CM, Alizadeh AA, et al. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. Nat Genet 1999;23:41–6.
10. Solinas-Toldo S, Lampel S, Stilgenbauer S, et al. Matrix-based comparative genomic hybridization: biochips to screen for genomic imbalances. Genes Chromosomes Cancer 1997;20:399–407.
11. Lucito R, West J, Reiner A, et al. Detecting gene copy number fluctuations in tumor cells by microarray analysis of genomic representations. Genome Res 2000;10:1726–36.
12. Lucito R, Healy J, Alexander J, et al. Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. Genome Res 2003;13:2291–305.
13. Wang DG, Fan JB, Siao CJ, et al. Large-scale identification, mapping, and genotyping of single- nucleotide polymorphisms in the human genome Science 1998;280:1077–82.
14. Mei R, Galipeau PC, Prass C, et al. Genome-wide detection of allelic imbalance using human SNPs and high-density DNA arrays. Genome Res 2000;10:1126–37.
15. Lindblad-Toh K, Tanenbaum DM, Daly MJ, et al. Loss-of-heterozygosity analysis of small-cell lung carcinomas using single-nucleotide polymorphism arrays. Nat Biotechnol 2000;18:1001–5.
16. Hoque MO, Lee CC, Cairns P, Schoenberg M, Sidransky D. Genome-wide genetic characterization of bladder cancer: a comparison of high- density single-nucleotide polymorphism arrays and PCR-based microsatellite analysis. Cancer Res 2003;63: 2216–22.
17. Lieberfarb ME, Lin M, Lechpammer M, et al. Genome-wide loss of heterozygosity analysis from laser capture microdissected prostate cancer using single nucleotide

[14] G. J. Paez, R. Beroukhim, J. Lee, et al. Genome coverage and sequence fidelity of f29 polymerase based multiple strand displacement whole genome amplification, submitted for publication.

[15] Internet address: http://www.dchip.org.

polymorphic allele (SNP) arrays and a novel bioinformatics platform dChipSNP. Cancer Res 2003;63:4781–5.
18. Primdahl H, Wikman FP, von der Maase H, Zhou XG, Wolf H, Orntoft TF. Allelic imbalances in human bladder cancer: genome-wide detection with high-density single-nucleotide polymorphism arrays. J Natl Cancer Inst 2002;94:216–23.
19. Wang ZC, Lin M, Wei L-J, et al. Loss of heterozygosity and its correlation with expression profiles in subclasses of invasive breast cancers Cancer Res 2004;64:64–71.
20. Jänne PA, Li C, Zhao X, et al. High-resolution single nucleotide polymorphism array and clustering analysis of loss of heterozygosity in human lung cancer cell lines. Oncogene 2004; 23:2716–26.
21. Kennedy GC, Matsuzaki H, Dong S, et al. Large- scale genotyping of complex DNA. Nat Biotechnol 2003;21:1233–7.
22. Li C, Hung Wong W. Model-based analysis of oligonucleotide arrays: model validation, design issues and standard error application. Genome Biol 2001;2:RESEARCH0032.
23. Li C, Wong WH. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. Proc Natl Acad Sci USA 2001;98:31–6.
24. Dugad R, Desai UB. A tutorial on Hidden Markov Models. Technical report, No. SPANN-96.1, 1996.
25. Lange K. Mathematical and Statistical Methods for Genetic Analysis. New York: Springer-Verlag, 2002.
26. Li C, Wang WH. DNA-Chip Analyzer (dChip). In: The Analysis of Gene Expression Data: Methods and Software. Parmigiani G, Garrett ES, Irizarry R, Zeger SL, editors. New York: Springer, 2003. p. 120–41.
27. Wang TL, Maierhofer C, Speicher MR, et al. Digital karyotyping. Proc Natl Acad Sci USA 2002;99:16156–61.
28. Sundaresan V, Chung G, Heppell-Parton A, et al. Homozygous deletions at 3p12 in breast and lung cancer. Oncogene 1998;17:1723–9.
29. Albertson DG, Pinkel D. Genomic microarrays in human genetic disease and cancer. Hum Mol Genet 2003;12 Spec No 2:R145–52.
30. Knudson AG, Jr. Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci USA 1971;68:820–3.
31. Cavenee WK, Dryja TP, Phillips RA, et al. Expression of recessive alleles by chromosomal mechanisms in retinoblastoma. Nature 1983;305:779–84.