Oncogene (2004) 23, 2716–2726
© 2004 Nature Publishing Group All rights reserved 0950-9232/04 $25.00

www.nature.com/onc

High-resolution single-nucleotide polymorphism array and clustering analysis of loss of heterozygosity in human lung cancer cell lines

Pasi A Jänne[1,2], Cheng Li[3,4], Xiaojun Zhao[1], Luc Girard[5,6], Tzu-Hsiu Chen[1], John Minna[5,6], David C Christiani[7,8], Bruce E Johnson[1,2] and Matthew Meyerson*[,1]

[1]Department of Medical Oncology, Dana Farber Cancer Institute, Boston, MA, USA; [2]Department of Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA; [3]Department of Biostatistical Sciences, Dana Farber Cancer Institute, Boston, MA, USA; [4]Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA; [5]Hamon Center for Therapeutic Oncology Research, University of Texas Southwestern Medical Center, Dallas, TX, USA; [6]Departments of Internal Medicine and Pharmacology, University of Texas Southwestern Medical Center, Dallas, TX, USA; [7]Department of Environmental Health, Harvard School of Public Health, Boston, MA, USA; [8]Pulmonary and Critical Care Unit, Department of Medicine, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA

Chromosomal loss of heterozygosity (LOH) is a common mechanism for the inactivation of tumor suppressor genes in human epithelial cancers. Hybridization to single-nucleotide polymorphism (SNP) arrays is an efficient method to detect genome-wide cancer LOH. Here, we survey LOH patterns in a panel of 33 human lung cancer cell lines using SNP array hybridization containing 1500 SNPs. We compared the LOH patterns generated by SNP array hybridization to those previously obtained by 399 microsatellite markers and find a high degree of concordance between the two methods. A novel informatics platform, dChipSNP, was used to perform hierarchical tumor clustering based on genome-wide LOH patterns. We demonstrate that this method can separate non-small-cell and small-cell lung cancer samples based on their shared LOH. Furthermore, we analysed seven human lung cancer cell lines using a novel 10 000 SNP array and demonstrate that this is an efficient and reliable method of high-density allelotyping. Using this array, we identified small regions of LOH that were not detected by lower density SNP arrays or by standard microsatellite marker panels.

Oncogene (2004) **23,** 2716–2726. doi:10.1038/sj.onc.1207329
Published online 29 March 2004

Keywords: lung neoplasm; non-small cell; small-cell carcinoma; loss of heterozygosity; single-nucleotide polymorphism

Introduction

Lung cancer is the leading cause of cancer-related deaths in the United States (Jemal *et al*., 2003). Its genetic basis is complex and involves several genetic alterations including activating mutations in proto-oncogenes (such as *KRAS*) and inactivation of tumor suppressor genes (Osada and Takahashi, 2002; Zabarovsky *et al*., 2002; Sekido *et al*., 2003). Tumor suppressor gene inactivation typically requires two genetic events: a genetic deletion (loss of heterozygosity or LOH) on one of the chromosomal alleles and a smaller genetic event (such as a point mutation, promoter hypermethylation or a small intragenic deletion) on the other allele leading to biallelic inactivation of the gene.

LOH can be detected through allelotyping the DNA from a cancer sample and a corresponding normal control sample with polymorphic markers from each of the chromosomal arms. Traditionally, this involves polymerase chain reaction (PCR)-based amplification of simple sequence length polymorphisms (SSLPs) and the analysis of the amplified products using either gel- or capillary-based electrophoresis. Many studies in lung cancer have focused on LOH analyses in specific regions of the genome and correlated these findings with clinical outcome parameters including survival (Fong *et al*., 1994, 1995; Mitsudomi *et al*., 1996; Tseng *et al*., 1999). However, given the genetic complexity of lung cancers, LOH events from multiple loci likely contribute to the overall phenotype of the tumor. Unbiased genome-wide analyses have also been performed, although most studies have used only a limited number of markers (Virmani *et al*., 1998; Stanton *et al*., 2000). The most comprehensive genome-wide analysis in lung cancer using these methods was performed by Girard *et al*. (2000), who studied a total of 36 lung cancer cell lines (14 small-cell lung cancer (SCLC) cell lines and 22 non-small-cell lung cancer (NSCLC) cell lines) using 399 fluorescent microsatellite markers This study provided a global estimate of LOH in both SCLC and NSCLC cell lines and identified several putative areas that may harbor previously undiscovered tumor suppressor genes.

Single-nucleotide polymorphisms (SNPs) are the most common form of sequence variation in the human genome, occurring approximately every 1200 base pairs (bps) (Sachidanandam *et al*., 2001; Reich *et al*., 2003).

*Correspondence: M Meyerson, Department of Medical Oncology, Dana Farber Cancer Institute, Mayer 446, 44 Binney St., Boston, MA 02115, USA; E-mail: matthew_meyerson@dfci.harvard.edu
Received 4 September 2003; revised 29 October 2003; accepted 30 October 2003

Owing to their frequency in the genome, they provide a high-density method for analyses of polymorphic markers, such as LOH. We and others have previously demonstrated that hybridization to SNP arrays provides an efficient and unbiased method to evaluate genome-wide LOH from human tumors (Lindblad-Toh *et al*., 2000; Mei *et al*., 2000). This method uses simultaneous PCR-based amplification of nearly 1500 genomic loci with hybridization to a high-density oligonucleotide array (Affymetrix HuSNP array). The HuSNP arrays are comparable to SSLP and comparative genomic hybridization in the ability to detect genome-wide LOH, with a reduction in the amount of time and DNA required to perform these analyses (Lindblad-Toh *et al*., 2000). This method has the potential applicability in clinical and diagnostic laboratory setting and is ideally suited for the analysis of large numbers of tumor specimens. The HuSNP method has now been utilized by our group and other investigators to study allelic imbalances in primary human tumors, including samples generated by laser capture microdissection (Janne *et al*., 2002; Primdahl *et al*., 2002; Dumur *et al*., 2003; Hoque *et al*., 2003; Lieberfarb *et al*., 2003; Wang *et al*., 2003). Recently, a high-density SNP array, Mapping 10K array, has been generated that can robustly and reproducibly analyse over 10 000 SNP loci using a genome representation approach (Kennedy *et al*., 2003). This array has not been used previously to examine genome-wide LOH.

Genome-wide approaches such as mRNA expression profiling have led to the development of systematic tools for cancer classification (Eisen *et al*., 1998; Golub *et al*., 1999; Alizadeh *et al*., 2000). We and other groups have performed expression profiling in lung cancer (Bhattacharjee *et al*., 2001; Garber *et al*., 2001; Beer *et al*., 2002). Tumors from specific histologic subtypes (such as SCLC, adenocarcinoma or squamous cell carcinoma) were found to cluster together based on shared expression profiles using hierarchical clustering (Bhattacharjee *et al*., 2001; Garber *et al*., 2001). Furthermore, among adenocarcinomas there are expression-based subgroups that associated with a shorter survival (Bhattacharjee *et al*., 2001; Garber *et al*., 2001; Beer *et al*., 2002).

The dCHIP program (http://www.dchip.org), used in the analysis of gene expression data, has recently been modified (dChipSNP) to be able to analyse LOH data (Li and Wong, 2001; Lin *et al*., 2003). The analysis of SNP arrays and other LOH data using dChipSNP now makes it possible to automate the detection of shared LOH regions and perform hierarchical clustering on tumors based on their shared LOH pattern (Lin *et al*., 2003). Previously, we have used these methods to analyse, perform hierarchical clustering and identify commonly deleted regions in prostate carcinoma specimens, and now apply them to lung carcinoma cell lines (Lieberfarb *et al*., 2003).

The present study was undertaken to characterize further and validate the efficacy of using the HuSNP array as a method for determining genome-wide LOH in lung cancer cell lines. The cell lines used in this study have all been characterized for LOH using SSLP markers (Girard *et al*., 2000). We analysed the data generated by both the SNP and SSLP methods using the dChipSNP program (Lin *et al*., 2003). We demonstrate that clustering of LOH data can distinguish SCLC from NSCLC with reasonable accuracy. In addition, we provide data using the 10K SNP array (Affymetrix GeneChip® Mapping 10K Array Xba 130) and compare the LOH findings to HuSNP- and SSLP-based methods, noting that the higher density array permits the detection of smaller regions of LOH.

Results

Genome-wide LOH analysis of lung cancer cell lines using HuSNP arrays

We generated LOH profiles from the DNA of 33 lung cancer cell lines and their corresponding lymphoblastoid cell line controls using HuSNP arrays representing 1494 SNPs. The mean call rate was 80.3% and ranged from 71.6 to 86.2%. The average call rate did not vary significantly between the lymphoblastoid (82.2±4.3%) and tumor cell lines (80.3±4.5%). Markers numbers are summarized in Table 1. The number of LOH events, the number of informative loci, the heterozygosity rate and the fractional allelic loss (FAL) are shown for all of the samples in Table 2. The mean LOH% is similar between the SCLC and the NSCLC cell lines.

Virtual markers

In order to compare the data generated using HuSNP- and SSLP-based methods, we generated virtual markers. Owing to the sparse nature of the HuSNP markers, this method allows us to infer a region of LOH based on the call (LOH or retention) of adjacent markers. We generated an LOH profile for each sample using virtual markers. The number of LOH events, the number of informative loci, the heterozygosity rate and the FAL are shown for all of the samples in Table 2. There was a good correlation of sample-wise LOH percentage when we compared actual and virtual markers ($R^2 = 0.94$; Supplemental Figure 1a). We further applied the virtual marker method to the LOH data generated by SSLP and also found a good correlation between LOH data generated by actual and virtual markers ($R^2 = 0.95$; Supplemental Figure 1b).

Table 1 Effective markers

	Mapping 10K array	*HuSNP array*	*SSLP*
No. of mapped markers	9666	1329	390
Mean informative % of actual markers	25%	22%	76%
No. of effective markers	2417	292	296

The total number of markers with known chromosomal positions are shown. The mean informative percentage is used to calculate the number of effective markers

Table 2 LOH analyses of 33 lung cancer cell lines using HuSNP analyses

	Actual markers				*Virtual markers*			
Cell line	*No. of LOH*	*No. of informative loci*	*Heterozygosity (%)*	*FAL (%)*	*No. of LOH*	*No. of informative loci*	*Heterozygosity (%)*	*FAL (%)*
SCLC								
H128	151	284	21.4	53.2	972	1828	60.4	53.2
H209	65	319	24.0	20.4	412	1994	65.9	20.6
H289	159	336	25.3	47.3	894	2045	67.6	43.7
H1184	97	324	24.4	29.9	615	1945	65.3	31.6
H1450	96	277	20.8	34.7	460	1605	53.1	28.6
H1607	143	308	23.2	46.4	830	1859	61.5	44.6
H1672	91	243	18.3	37.4	544	1530	50.6	35.6
H1963	31	311	23.4	10.0	207	1961	64.8	10.6
H2107	66	263	19.8	25.1	517	1788	59.1	28.9
H2141	83	262	19.7	31.7	627	1800	59.5	34.8
H2171	173	333	25.1	52.0	988	1960	64.8	50.4
H2195	108	337	25.4	32.0	582	1955	64.6	29.8
HCC33	86	291	21.9	29.6	434	1859	61.4	23.3
HCC970	123	313	23.6	39.3	828	2010	66.4	41.2
Mean				34.9				34.1
s.d.				12.2				11.8
NSCLC								
H1395	100	263	19.8	38.0	668	1816	60.0	36.8
H1648	119	288	21.7	41.3	688	1751	57.9	39.3
H1819	132	296	22.3	44.6	758	1630	53.9	46.5
H1993	141	321	24.2	43.9	704	1770	58.5	39.8
H2009	166	320	24.1	51.9	870	1796	59.4	48.4
H2087	130	290	21.9	44.8	813	1719	56.8	47.3
H2122	67	304	22.9	22.0	461	1902	62.9	24.2
H2347	61	299	22.5	20.4	350	1956	64.7	17.9
H2887	99	238	17.9	41.6	473	1279	42.3	37.0
HCC44	131	285	21.4	46.0	785	1767	58.4	44.4
HCC78	0	272	20.5	0.0	0	1973	65.2	0.0
HCC193	49	301	22.6	16.3	278	1921	63.5	14.5
HCC515	121	282	21.2	42.9	580	1741	57.6	33.3
HCC827	194	322	24.2	60.2	1122	2226	73.6	50.4
HCC366	98	310	23.3	31.6	530	1820	60.2	29.1
H2052	109	293	22.0	37.2	607	1748	57.8	34.7
H2126	135	272	20.5	49.6	906	1691	56.0	53.6
HCC95	134	289	21.7	46.4	828	1806	59.7	45.8
HCC1171	116	329	24.8	35.3	640	1879	62.1	34.1
Mean				37.6				35.6
s.d.				14.3				13.7

The number of LOH events, the number of informative loci, the percent heterozygosity and the FAL are shown using both actual and virtual markers. The mean and standard deviations are shown separately for NSCLC and SCLC cell lines

Comparison of HuSNP- and SSLP-based methods

We compared the LOH data generated by the HuSNP- and SSLP-based methods. As the SNP and SSLP markers are in different locations, we needed to use the virtual markers to compare LOH calls between the two methods, as the virtual markers generated from each data set are at the same chromosomal locations. After calculating LOH/retention calls for each virtual marker (total of 3025 markers × 33 cell lines = 99825), we examined the distribution of these calls using both the HuSNP and SSLP methods. A total of 51 888 virtual markers were informative in both array types. The LOH/retention call agreement was observed for 50 752 or 97.8% of the markers.

We examined the regional concordance of the HuSNP- and SSLP-based methods by examining each chromosomal arm (Figure 1). As can be seen in Figure 1, there is a high rate of concordance of LOH between the two methods (also see supplemental Table 3) on p- and q-arms on all chromosomes ($R^2 = 0.96$ for p-arms; $R^2 = 0.95$ for q-arms). We also examined the sample-wise LOH percentage correlation between HuSNP and SSLP using both actual and virtual markers (Supplemental Figure 1c and d). In both cases, there was a good correlation between the two methods ($R^2 = 0.88$ for actual markers and 0.89 for virtual markers). Additional comparisons between HuSNP- and SSLP-based methods are shown in supplemental Tables 1 and 2.

Sample clustering based on genome-wide LOH events

The dChipSNP program was used to perform sample clustering based on genome-wide LOH events. The analysis was repeated for LOH data generated using SSLP. We examined three representative LOH score thresholds: 0.18, 0.22 and 0.46. Using the lowest threshold (0.18), the tumor cells cluster predominantly into SCLC and NSCLC based on their shared LOH (data not shown). However, the branch lengths are short and not well distinguished from the rest of the branches. As the threshold is increased to 0.22 and then to 0.46, we observe better separation of the sample branches (Figure 2a). However, with the higher threshold, three of the SCLC cell lines (HCC33, H1963 and H1672) now fall out of the main SCLC cluster and are clustered together with the NSCLCs. The large-cell carcinoma cell line H2126 even now clusters among the SCLC cell lines. Three main clusters become evident. The left branch contains mainly adenocarcinomas, the middle branch a combination of adenocarcinomas, other NSCLC and two SCLC cell lines (H1672 and HCC33) and the right branch predominately SCLC cell lines (Figure 2a). Hierarchical clustering of LOH data generated from the 33 cell lines using the SSLP-based methods was also performed using dChipSNP (Figure 2b). A cluster analysis with a threshold of 0.46 was also used and demonstrates similar separation of sample branches into three main clusters. However, there are some differences among the cells lines within each cluster. Similar to the HuSNP-based clustering, two of the SCLC cell lines (HCC33 and H1963) cluster among the NSCLC cell lines (Figure 2b). However, two adenocarcinoma cell lines (H1819 and H2087) cluster within the SCLC cell lines. We performed the analysis to determine whether these observed clustering results based on shared LOH are related to the histologic subtypes of the cell lines (SCLC *vs* NSCLC). We find that the

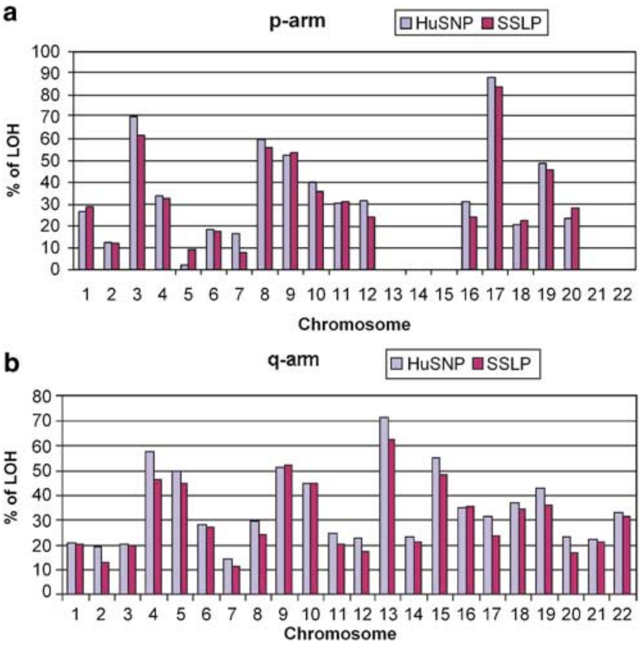

Figure 1 Comparison of LOH using HuSNP- and SSLP-based methods. Shown are percent LOH as determined by each of the methods for each of the chromosomal arms (**a**) p-arm, (**b**) q-arm

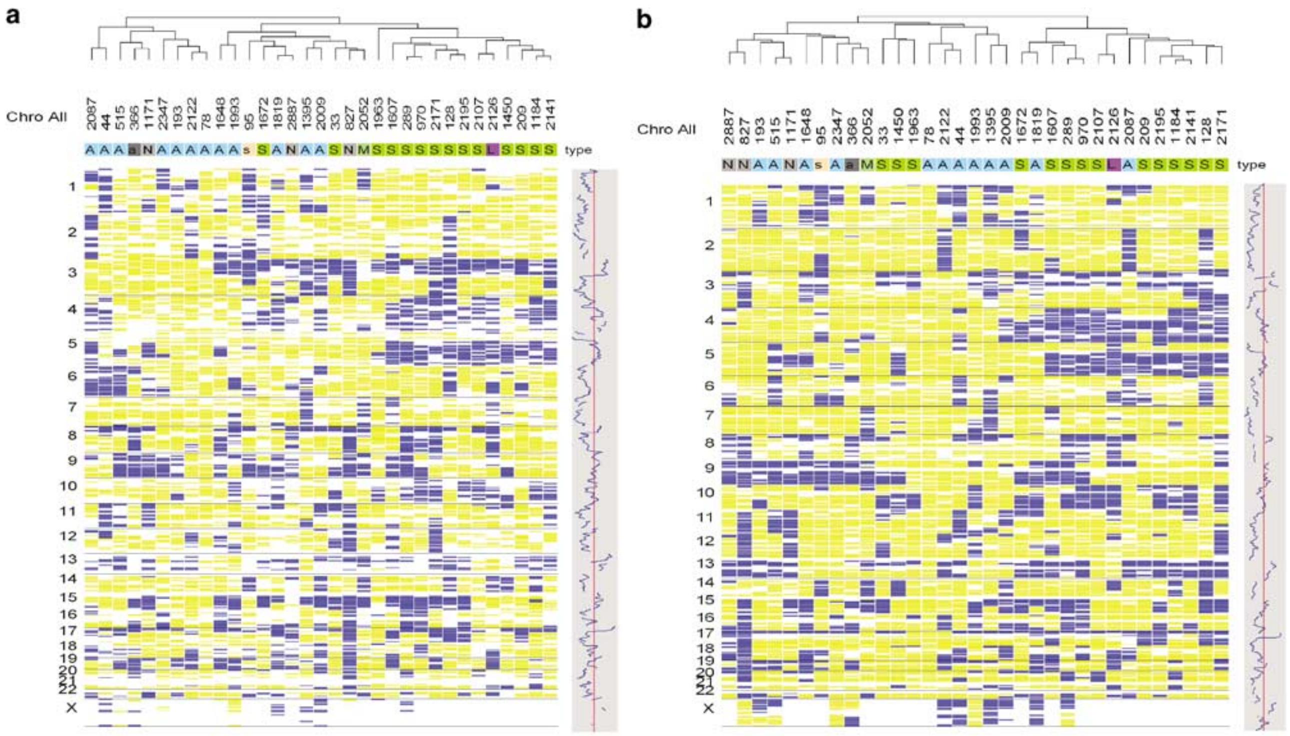

Figure 2 (**a**) Hierarchical clustering of tumor cell lines based on genome-wide LOH using HuSNP analysis. Threshold = 0.46. A: adenocarcinoma; S: small-cell lung carcinoma; L: large-cell carcinoma; N: non-small-cell lung carcinoma not otherwise specified; s (orange): squamous cell cancer; a: adenosquamous cell carcinoma; M: mesothelioma; blue, LOH; yellow, retention (**b**). Hierarchical clustering of tumor cell lines based on genome-wide LOH using SSLP analysis. Threshold = 0.46

clusters (SCLC *vs* NSCLC) based on LOH using either HuSNP or SSLP are significantly related to the histologic subtype of the cell line ($P<0.001$ by χ^2 for HuSNP and SSLP; $P<0.0001$ for HuSNP; and $P=0.0005$ for SSLP by Fisher's exact test).

LOH analysis using Mapping 10K SNP arrays

We analysed two control DNA specimens in duplicate and seven lung cancer cell lines with their corresponding lymphoblastoid cell lines using the GeneChip Mapping 10K array. The call rates for these samples are shown in Table 3. The concordance rate between duplicate specimens was extremely high. The call rates were slightly higher for the normal cell lines compared to the tumor cell lines (94.2 *vs* 89.6%, respectively), but both were higher than those observed with the HuSNP array.

Comparison of LOH analyses using SSLP, HuSNP and Mapping 10K array methods

We compared the LOH calls generated by the HuSNP, SSLP and Mapping 10K arrays using seven cells lines and their control lymphoblastoid DNA. The number of LOH events, the number of informative loci, the heterozygosity rate and the FAL are shown for all of the samples in Table 4. We also generated virtual markers (see Materials and methods) in order to be able to compare the three methods and performed similar analyses of LOH (Table 4). For the seven cell lines, the total number of virtual markers analysed was 21175 (3025 virtual markers per cell line). The total number of virtual markers in common between HuSNP and Mapping 10K array was 12289. The rate of agreement in these markers between the two arrays was 98.5% (12100/12289 markers). For SSLP and 10K arrays, the number of virtual markers in common was 15 162, with

Table 3 Performance of the Mapping 10K array

Cell line	*Call rate (%)*	*Concordance*	*Cell line pair*	*Normal*	*Tumor*
Control 1	93.0		BL2141/H2141	94.5	89.5
Control 1	84.9	99.99	BL289/H289	94.2	90.1
Control 2	93.9		BL128/H128	94.3	90.2
Control 2	93.8	99.99	BL10/H1648	94.2	89.8
			BL1395/H1395	95.2	88
Mean	91.4		BL2171/H2171	94.5	92.4
s.d.	4.4		BL2107/H2107	92.8	87.2
			Mean	94.2	89.6
			s.d.	0.7	1.7

The call rates and concordance between replicates for control DNA specimens are shown. The call rates, the mean and standard deviation are shown for seven tumor and corresponding normal lymphoblastoid cell lines

Table 4 LOH analyses of seven lung cancer cell lines using both HuSNP, SSLP and the Mapping 10K array analyses

		Actual markers				*Virtual markers*			
Cell line	*Array*	*No. of LOH*	*No. of informative loci*	*Heterozygosity (%)*	*FAL (%)*	*No. of LOH*	*No. of informative loci*	*Heterozygosity (%)*	*FAL (%)*
SCLC									
H128	HuSNP	151	284	21.4	53.2	972	1828	60.4	53.2
	SSLP	143	316	81.0	45.2	1005	2255	74.5	44.6
	10K	1386	2382	24.6	58.2	1273	2586	85.5	49.2
H2141	HuSNP	83	262	19.7	31.7	627	1800	59.5	34.8
	SSLP	81	285	73.1	28.4	697	2315	76.5	30.1
	10K	1032	2256	23.3	45.7	941	2504	82.8	37.6
H289	HuSNP	159	336	25.3	47.3	894	2045	67.6	43.7
	SSLP	131	300	76.9	43.7	658	2127	70.3	31.0
	10K	1250	2348	24.3	53.2	1236	2670	88.3	46.3
H2107	HuSNP	66	263	19.8	25.1	517	1788	59.1	29.0
	SSLP	78	397	78.7	25.4	537	2302	76.1	23.3
	10K	795	2197	22.7	36.2	701	2513	83.1	27.9
H2171	HuSNP	173	333	25.1	52.0	988	1960	64.8	50.4
	SSLP	141	289	74.1	48.8	1070	2249	74.3	47.6
	10K	1396	2732	28.2	51.1	1310	2610	86.3	50.2
NSCLC									
H1395	HuSNP	100	263	19.8	38.0	668	1816	60.0	36.8
	SSLP	110	305	78.2	36.1	816	2370	78.3	34.4
	10K	1128	2361	24.4	47.8	1145	2601	86.0	44.0
H1648	HuSNP	119	288	21.7	41.3	688	1751	57.9	39.3
	SSLP	121	309	79.2	39.2	800	2203	72.8	36.3
	10K	1206	2533	26.2	47.6	1175	2548	84.2	46.1

The number of LOH events, the number of informative loci, the percent heterozygosity and the FAL are shown using both actual and virtual markers a 95.5% rate of agreement (14474/15162). The 10K array also provides calls (either LOH or retention) for 71.7 and 22.3% of the loci that were identified as noninformative by the HuSNP and SSLP methods, respectively. Additional details on the analyses between the three methods are provided in supplemental Tables 4–7.

We examined specific regions of the genome to determine if there are areas of LOH that were ambiguous or not identified by HuSNP and/or SSLP and could be identified using the Mapping 10K array due to the greater number of markers. Examples of these are shown in Figures 3 and 4. LOH events were not identified on chromosome 10 using HuSNP analysis of sample pairs BL1648/H1648 and BL128/H128 (Figure 3). By SSLP, each sample pair has one marker that demonstrates an LOH event. However, the adjacent markers are not informative and are separated by a significant genetic distance. Using the Mapping 10K array, small regions of LOH with a high density of informative loci were identified on chromosome 10 in both the cell lines (see circled section in Figure 3). The proximal and distal ends of the deletions are clearly identified and the single LOH events identified using SSLP fall within these regions. On chromosome 20, there is a single area of LOH identified by HuSNP in sample pair BL1648/H1648 (Figure 4). The corresponding region using SSLP has four tandem LOH events, suggesting a deletion of 20q. This is confirmed using the 10K array, which demonstrates multiple contiguous deleted markers on 20q. The LOH events identified using HuSNP and SSLP fall within the region defined by the Mapping 10K array (circled regions in Figure 4). We also examined the Mapping 10K SNP array data for regions of LOH that were not identified by either SSLP or HuSNP. Meaningful areas of deletion were defined as stretches of at least three contiguously deleted SNPs (Table 5). Multiple new small regions of LOH were identified using the Mapping 10K arrays in four out of the seven cell line pairs examined. We further compared the differences in the size of LOH regions detected by SSLP, HuSNP and the Mapping 10K array (Table 6) using these seven cell lines. The Mapping 10K array identified more than twice the number of LOH regions compared to SSLP or HuSNP. The minimum, mean and median sizes of these regions were substantially smaller

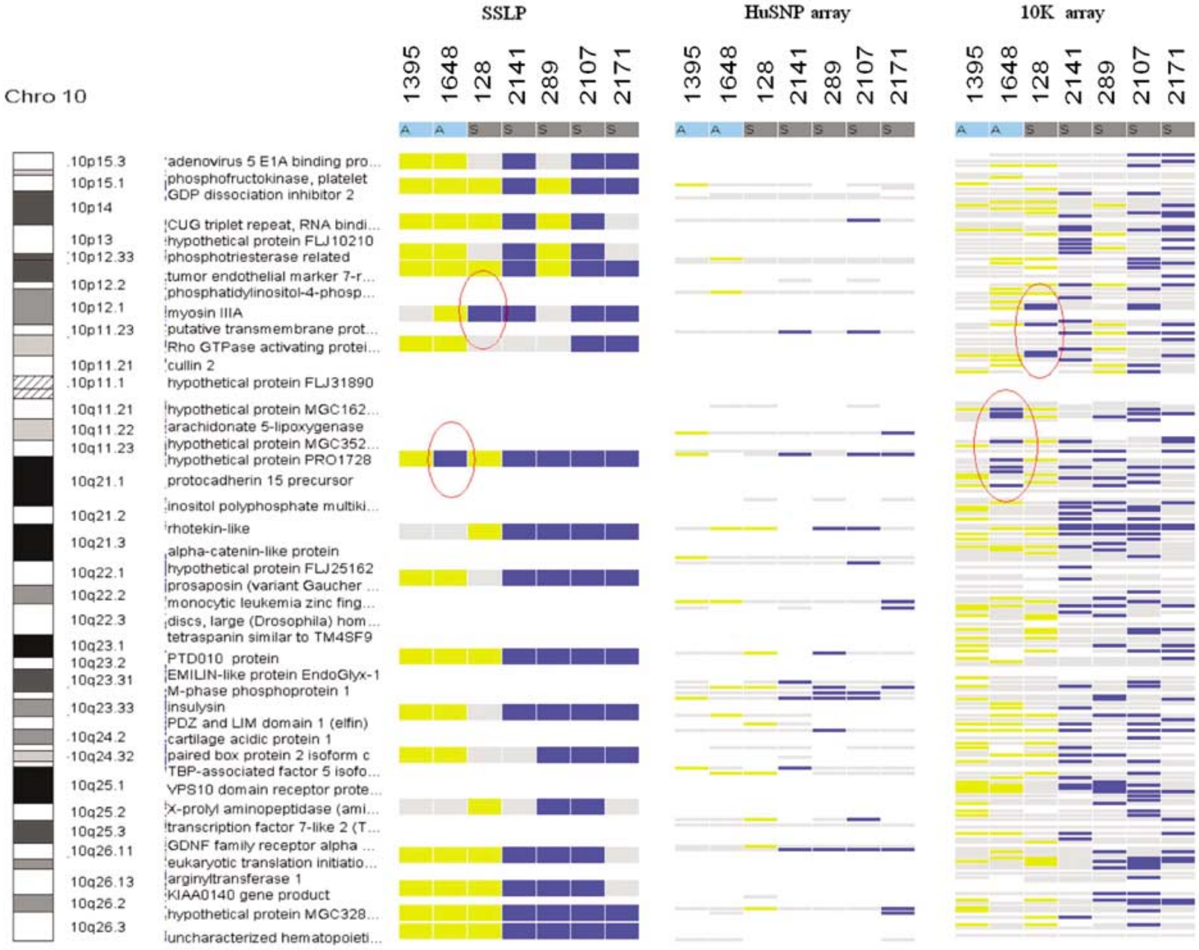

Figure 3 The Mapping 10K array identifies regions of LOH that are missed by HuSNP. Blue, LOH; yellow, retention; gray, uninformative

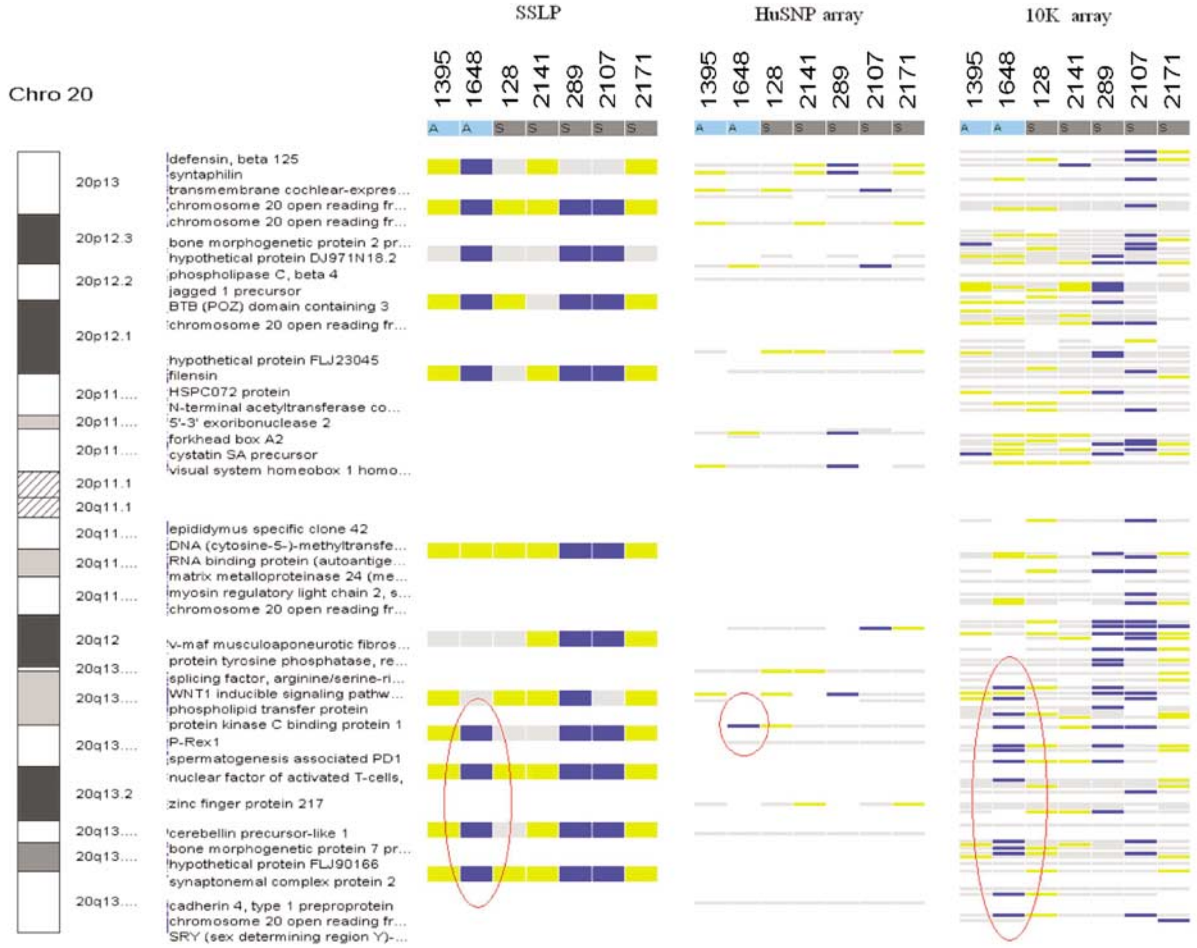

Figure 4 The Mapping 10K array helps define a region of LOH identified by either SSLP or HuSNP. Blue, LOH; yellow, retention; gray, uninformative

Table 5 Novel regions of LOH identified by the Mapping 10K array

Cell line	H1395	H1648	H2171	H2141
Histology	Adeno	Adeno	SCLC	SCLC
LOH regions	2p24.1 2p22.2 2p15 2p12 4q28.2 15q26.3	6q22.31–6q23.1 7q31.31	11p14.1 11p12	7q31.2 7q36.2 8p12 16q11.1–16q11.2

Shown are the chromosomal locations of novel regions of LOH, not identified by SSLP or HuSNP, in four lung cancer cell lines

Table 6 Comparison of the sizes of LOH regions detected using SSLP, HuSNP or the Mapping 10K array in seven lung cancer cell lines

Method	*No. of regions*	*Minimum*	*Maximum*	*Mean*	*Median*
SSLP	94	17.5	189.8	81.6	71.8
HuSNP	82	12.7	241.6	95.3	94.4
Mapping 10K	190	1.6	199.0	44.7	27.5

All sizes are listed in megabases (Mb)

using the Mapping 10K array compared with the other two methods (Table 6). The maximum size of the LOH regions was similar with all three methods.

Discussion

The present study describes genome-wide LOH analyses of human lung cancer cell lines using SNP arrays. These same cell lines have been examined for LOH using 399 SSLP markers distributed throughout the entire genome (Girard *et al*., 2000). We validated our SNP assay data by comparing it to the SSLP-based method with the aid of a novel informatics platform dChipSNP (Lin *et al*., 2003). Using these methods, we find a very high degree of concordance between the two types of LOH analyses as analysed by sample-wise LOH percentage or by chromosomal arm. By examining individual virtual markers, we find a 97.8% agreement between the two methods. The excellent agreement of the virtual marker calls using data from different platforms (SSLP and HuSNP) supports the validity of our method. Our results to date also represent the highest resolution of allelotyping and most detailed analyses of individual loci in NSCLC and SCLC lung carcinoma cell lines using two independent methods.

Hierarchical clustering approaches can be used to classify cancer samples using genome-wide approaches such as mRNA expression profiling. We now show that tumor classification can be performed by hierarchical clustering of LOH regions, using the dChipSNP software. Using this method, we are able to demonstrate that tumor cells lines cluster into three main groupings regardless of the method of LOH assessment. However, there are some minor differences among the individual cell lines within the clusters. The distribution of markers and the method of LOH assessment between SSLP and HuSNP likely account for these differences. We would expect even better cluster separation in studies using the Mapping 10K array. One of the groups contains predominately SCLC, another predominately adenocarcinoma cell lines and third a mixture of tumor cell lines. The large-cell carcinoma cell line H2126 clusters among the SCLC samples. Using mRNA expression profiling, large-cell tumors have also been found to cluster closest with SCLCs (Garber *et al*., 2001).

Tumors bearing different LOH patterns may be more advanced or associated with different clinical behaviors (Hoque *et al*., 2003). We did not analyse sufficient numbers of adenocarcinoma cell lines to be able to determine if there are several adenocarcinoma clusters based on shared LOH as has been demonstrated by shared mRNA expression (Bhattacharjee *et al*., 2001; Garber *et al*., 2001; Beer *et al*., 2002). However, using HuSNP, we do find two clusters among the NSCLC cell lines and perhaps with greater numbers of samples additional groupings would be found. Interestingly, the cluster containing predominately adenocarcinomas (Figure 2a) contains tumors cell lines with no LOH on chromosome 3, while the second NSCLC cluster contains tumors with LOH on chromosome 3. These findings also need to be validated in primary tumors and the significance of LOH patterns correlated with clinical outcome parameters. We have previously demonstrated that using laser capture microdissection we are able to isolate relative pure populations of tumor cells from paraffin-embedded specimens and perform HuSNP analyses (Janne *et al*., 2002; Lieberfarb *et al*., 2003). Our pilot studies demonstrated that there was a significant difference in the mean LOH percentage between adenocarcinomas and bronchioloalveolar cell carcinomas (Janne *et al*., 2002). Additional studies using paraffin-embedded tumor specimens and correlating the LOH patterns with clinical outcome(s) are underway. These techniques may ultimately also be more adaptable to clinical specimens compared with gene expression profiling as paraffin-embedded tumor specimens are more commonly available than fresh frozen tumor specimens.

Although the HuSNP method is higher throughput, faster and automated compared with SSLP-based analyses, neither method is perfect. The main disadvantage of the HuSNP method is the much lower rate of informative loci compared with SSLP-based methods (mean 22.3 *vs* 76.4%). Significant regions of the genome have few or no informative markers, thus limiting the utility of these approaches. Using both methods together, the combined informative rate is 84% and both methods provide calls for loci that were not informative by the other method (supplemental Tables 1 and 2). However, a combined analysis using both methods is unlikely to be practical for future studies.

Higher-density arrays would help overcome some of these limitations. The present study represents the first examination of the 10 000 SNP array. We have compared the results using the Mapping 10K array in seven cell lines both of which have also been examined using HuSNP and SSLP markers. We find approximately 10 times as many informative loci throughout the genome compared with HuSNP (Table 4). The Mapping 10K array also identifies 71.7 and 22.3% of the loci that were noninformative by HuSNP or SSLP, respectively (supplemental Tables 4–7). With the increased resolution, smaller areas of deletions that are missed by HuSNP (see Figures 3 and 4) and areas of single LOH can now be further delineated. We also find several small areas of LOH and ones that were previously not identified using SSLP or HuSNP (Tables 5 and 6). Since the Mapping 10K array will aid significantly in the detection of small areas of LOH and provide the tools to begin isolating genes in such deletions, it should be the LOH detection method of choice for future studies.

Materials and methods

Tumor cell lines

In all, 33 lung cancer cell lines (14 SCLCs and 19 NSCLCs) and their normal matched lymphoblastoid cell lines were used in this study and have been characterized previously (Girard *et al*., 2000). They were grown in cell culture in RPMI 1640 supplemented with 5 or 10% fetal bovine serum and DNA was prepared using standard methods.

HuSNP analyses

HuSNP analyses of normal and tumor cell DNA was performed according to previously published methods (Lindblad-Toh *et al*., 2000). The primary PCR, using 24 pools of primer pairs, was performed using a total of 300 ng of DNA according to the manufacturer's recommendations. A 1 : 1000 dilution of each pool was reamplified using biotinylated T7 and T3 primers as described previously (Lindblad-Toh *et al*., 2000). PCR products from each of the secondary pools were verified by using agarose-gel electrophoresis. Secondary PCR products were then concentrated (Microcon-10 spin column, Amicon Bioseparations, Bedford, MA, USA), denatured and hybridized to HuSNP arrays for 16 h at 44°C and 40 r.p.m. The following day, the arrays were washed on the Affymetrix fluidics stations, stained with strepavidin–phycoerythrin and biotinylated-anti-strepavidin antibody, and scanned using the HP GeneArray Scanner (Hewlett-Packard, Palo Alto, CA, USA) all according to the manufacturer's recommended conditions (HuSNP Mapping Assay Manual, Affymetrix P/N 700308) and as described previously (Lindblad-Toh *et al*., 2000).

GeneChip Mapping 10K array analyses

Preparation of DNA target DNA from tumor and control lymphoblastoid cell lines were prepared according to the Early Access Mapping Assay protocol as follows: DNA (250 ng) was digested with *Xba*I in 20 μl (New England Biolabs, Boston, MA, USA) and then ligated with a linker (supplied by Affymetrix, Inc.) by T4 DNA ligase (New England Biolabs, Boston, MA, USA) in a final volume of 25 μl. Four individual 100 μl mixtures were subsequently set up, containing 10 μl of 1 : 4 dilution of ligated DNA, 2.5 mM $MgCl_2$, 250 μM dNTPs, 10 U Amplitaq Gold (PE Biosystems, Foster City, CA, USA) and 0.75 μM PCR primer (supplied by Affymetrix, Inc.). Mixtures were denatured for 3 min at 95°C, followed by 35 cycles of 95°C for 20 s, 59°C for 15 s and 72°C for 15 s, and a final extension of 72°C for 7 min. The four PCR products were combined and concentrated using QIAquick columns (Qiagen, Inc.). Volumes were adjusted with distilled water to 47.5 μl containing 9–20 μg DNA.

Target labeling, hybridization, washing and staining The PCR products were fragmented by incubating them for 30 min at 37°C, in a final volume of 55 μl containing 47.5 μl of the PCR products and DNAse (supplied by Affymetrix, Inc.). After fragmentation, 50 μl of this mixture was used to make a 65 μl labeling mixture, containing 0.0154 mM biotinylated-ddATP (Perkin-Elmer Life Sciences), 0.23–0.46 U (stock 15–30 U/ml) TdT (Promega) and incubated for 16 h at 37°C, while the remaining 5 μl was examined for successful fragmentation using 4% NuSieve 3 : 1 plus agarose (Cambrex)-gel electrophoresis. The labeled DNA was diluted to a final volume of 240 μl containing 2.92 M tetramethylammonium chloride, 0.0125% Tween-20, 12.5 μg/ml Human Cot-1 DNA, 37.5 pM Oligo B2, 0.125 mg/ml herring sperm DNA, 6.25 mM EDTA, 2.71 × Denhardt's solution, 5.4% DMSO, 0.061 M MES and denatured for 10 min at 95°C. After 10 s on ice and 47.5°C for 10 min, 200 μl of each sample was injected into the chips and hybridized for 16–18 h in a rotating oven (Affymetrix, Inc.) at 47.5°C and 60 r.p.m. The following day, each array was washed and stained on the Affymetrix fluidics station. Arrays were first washed for six cycles with nonstringent Wash Buffer A (6 × SSPE (BioWhittaker Molecular Applications/ Cambrex) + 0.01% Tween-20) at 25°C and then for six cycles with stringent Wash Buffer B (0.6 × SSPE, 0.01% Tween-20) at 45°C. Arrays were subsequently incubated for 10 min at 25°C using 10 μg/ml streptavidin (Pierce) in the Stain Buffer (6 × SSPE, 0.01% Tween-20, 1 × Denhardt's solution) in 500 μl, followed by washing for six cycles with nonstringent Wash Buffer A. Subsequently, the chips were stained for 10 min at 25°C with 5 μg/ml biotinylated anti-streptavidin antibody (Vector Laboratories) in 500 μl of stain buffer and for 10 min at 25°C with 10 μg/ml streptavidin, R-phycoerythrin conjugate (Molecular Probes) in 500 μl of stain buffer.

Scanning and allelotype generation The chips were scanned using the HP (Hewlett-Packard, Palo Alto, CA, USA) scanner according to the manufacturer's recommended conditions (HuSNP Mapping Assay Manual, Affymetrix P/N 700308). Hybridization signal was detected by Affymetrix Microarray Suite 5.0 software (Affymetrix, Inc.). Genotype calls were generated using the Genotyping Tools software. The data were analysed using the dChipSNP software (Lin *et al*., 2003).

Markers

In order to compare data generated by SSLP- and SNP-based methods, all markers were mapped into the UCSC hg13 genome assembly (http://genome.ucsc.edu/). When a marker was mapped to two or more genome positions, one position was randomly selected to use in the analyses. For the SNP-based methods, a minority of the SNPs in both the HuSNP and Mapping 10K arrays have not been mapped to any chromosomal location and are not included in the analyses. Using these criteria, the number of mapped markers and the number of effective markers (defined as the number of mapped markers times the mean heterozygosity) used in the analyses are shown in Table 1.

LOH calls

For HuSNP and Mapping 10K array data, we used the Affymetrix genotyping software (Affymetrix GeneChip 5.0) to examine the SNP hybridization patterns and make SNP calls for all loci in each of the tumor cell lines and their corresponding lymphoblastoid cell lines. One of the three LOH calls are then assigned by dChipSNP for a loci in a pair of normal and tumor cell lines according to the following rules: L, loss (AB in normal, A or B in tumor); R, retention (AB in both normal and tumor); and N, noninformative or no call (AB_A or AB_B (implying either AB or A or B, respectively), or no call in normal or tumor). The overall call rate by the software was calculated as the number of SNPs assigned AA, BB or AB divided by the total number of SNPs in the array. For the 10K array, we calculated the concordance rate between duplicates specimens by only examining loci where an SNP assignment could be made. We calculated the heterozygosity rate by dividing the number of informative loci over the number of mapped markers for each of the specimens and by using either the SSLP- or SNP-based methods. FAL is defined as the number of LOH events divided by the total number of informative loci. We also obtained the LOH data generated using the SSLP-based methods and compared it to the SNP-based methods (Girard *et al*., 2000). Since microsatellite alterations (MAs) cannot be evaluated using the SNP-based methods, we assigned a locus defined as LOH/MA as an L call, and those assigned as HET/MA as R calls. The dChipSNP software can then used to make LOH calls for the SSLP-based data.

Virtual markers

To account for the uneven marker distribution in the SNP-based analyses, we used 3025 virtual markers to cover the genome at 1 megabase (Mb) intervals. A 1 Mb distance was selected to balance the average between marker distance of 2 Mb in the HuSNP array and 300 Kb in the 10K array. The LOH calls of virtual markers are inferred by using 'Region with the same boundary, 10 Mb' method in the dChipSNP program (Lin *et al*., 2003). Using this method, we first declare a chromosomal region to have the 'same boundary' in one sample if there are two informative markers with the same L or R call on the two boundaries of the region, and there are no other informative markers in the region. Then, a virtual marker falling in such a region is inferred to have the same call as the boundaries, if the shorter distance between the virtual marker and the two boundaries is less than 10 Mb. Virtual markers are called as N if they do not fall in any regions with the same boundary. Thus, virtual markers between consecutive L or R calls are inferred as long as the distance between consecutive calls are smaller than 20 Mb. We also performed a more complicated statistical virtual marker inference method based on a hidden Markov model and obtained similar results (data not shown). Complete details of these methods are discussed and presented in full in Lin *et al*. (2003).

We also inferred virtual markers for the data set generated by the SSLP-based method. The size of the PCR fragments ranged from 108 to 448 bps; thus we only considered their start positions when they were used to infer the LOH status for virtual markers at 1 Mb intervals.

Data generated with the virtual markers were then used to calculate the number of informative loci, the heterozygosity and the FAL as was carried out for the actual marker set. This was repeated for the data sets generated by the SNP- and SSLP-based methods. Furthermore, we examined the correlation between the rate of LOH obtained using the mapped actual markers and the virtual markers for each of the tumor cell lines using the data generated by both the SNP- and SSLP-based methods.

Sample clustering based on genome-wide LOH events

The dChipSNP program provides the ability to cluster tumor specimens based on their shared LOH regions. We computed an LOH score for virtual markers, if the weighted number of informative actual markers within 10 Mb on each side is $\geqslant 5$. The score is the weighted LOH% across samples of the nearby actual markers and is between 0 and 1. A high LOH score indicates that many samples have LOH events in the nearby region.

This score is plotted on the right side of the LOH data picture in blue. By adjusting the score threshold line (in red), one can highlight (denoted by blue) the markers or genes in the chromosome regions with LOH scores exceeding the threshold. We define such regions as the selected chromosomal LOH regions.

The virtual markers in the above selected chromosomal LOH regions are used for sample clustering. The distance between two samples is defined as the ratio between the number of selected virtual markers having the same loss or retention call and the number of selected virtual markers informative in both samples. Average linkage is used for hierarchical clustering. Thus, if two samples share loss or retention of many chromosomal regions, they will cluster closely together.

By changing the LOH score threshold from 0.01 to 1.00, we can examine progressively 100 sample clustering trees. Using a larger threshold, the clustering is driven more by the regions with LOH events. This thresholding is similar to gene filtering in expression data; some regions with background noise are not used in the clustering and as a result this leads to better separation of cluster branches.

Comparison of LOH analysis using SSLP- and SNP-based methods

LOH percentages determined by SSLP- and SNP-based methods were compared for each of the cell lines. We examined the correlation between the LOH percentage obtained by SSLP to that obtained by SNP and calculated an overall correlation coefficient. We repeated the analyses using both actual and virtual markers and calculated a correlation between the two methods over the whole genome. We also examined the LOH correlation for each of the chromosomal arms. We used the LOH calls generated for virtual markers to be able to compare the SNP- and SSLP-based methods. The percentage of LOH calls among informative virtual markers across all of the 33 samples are computed for the p- and q-arms separately and for the SNP- and SSLP-based methods separately. The p-arms were defined as a region from the beginning of a chromosome up to a defined chromosome position near the centromere and the q-arms were defined as a region from that defined position to the end of the chromosome (data not shown). These position points are based on the UCSC hg13 genome assembly file (cytoBand.txt).

Comparison of SNP LOH analysis using HuSNP and 10K array

We also analysed seven tumor cell lines, chosen at random, and their corresponding normal lymphoblastoid cell lines using the Gene Chip Mapping 10K array. These included two adenocarcinomas (H1395 and H1638) and five SCLC (H128, H2141, H289, H2107 and H2171) cell lines. The analysis of the number of LOH events and the rates of heterozygosity for both actual and virtual markers were analysed with the methods described for data generated with the HuSNP array. Using dChipSNP, we examined regions of LOH using the Mapping 10K array that were not identified using either HuSNP or SSLP. We defined a meaningful region of deletion as one that contained at least three contiguously deleted SNPs.

Acknowledgements

This work is supported by grants from the National Cancer Institute R01CA92824 (DC), the National Institutes of Health 2P30 CA06516-39 (CL) and 1K12CA87723-01 (PAJ), the American Cancer Society RSG-03-240-01-MGO (MM), the Pew Scholars in the Biomedical Sciences (MM), the Flight Attendant Medical Research Institute (MM), the Dunkin Donuts Rising Star Award (PAJ), The Arthur and Rochelle Belfer Foundation (CL and MM), Lung Cancer SPORE P50CA70907 (JM) and The Longenbaugh Foundation (JM).

References

Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, Boldrick JC, Sabet H, Tran T, Yu X, Powell JI, Yang L, Marti GE, Moore T, Hudson Jr J, Lu L, Lewis DB, Tibshirani R, Sherlock G, Chan WC, Greiner TC, Weisenburger DD, Armitage JO, Warnke R, Levy R, Wilson W, Grever MR, Byrd JC, Botstein D, Brown PO and Staudt LM. (2000). *Nature*, **403,** 503–511.

Beer DG, Kardia SL, Huang CC, Giordano TJ, Levin AM, Misek DE, Lin L, Chen G, Gharib TG, Thomas DG, Lizyness ML, Kuick R, Hayasaka S, Taylor JM, Iannettoni MD, Orringer MB and Hanash S. (2002). *Nat. Med.*, **8,** 816–824.

Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, Ladd C, Beheshti J, Bueno R, Gillette M, Loda M, Weber G, Mark EJ, Lander ES, Wong W, Johnson BE, Golub TR, Sugarbaker DJ and Meyerson M. (2001). *Proc. Natl. Acad. Sci. USA*, **98,** 13790–13795.

Dumur CI, Dechsukhum C, Ware JL, Cofield SS, Best AM, Wilkinson DS, Garrett CT and Ferreira-Gonzalez A. (2003). *Genomics*, **81,** 260–269.

Eisen MB, Spellman PT, Brown PO and Botstein D. (1998). *Proc. Natl. Acad. Sci. USA*, **95,** 14863–14868.

Fong KM, Zimmerman PV and Smith PJ. (1994). *Genes Chromosomes Cancer*, **10,** 183–189.

Fong KM, Zimmerman PV and Smith PJ. (1995). *Cancer Res.*, **55,** 220–223.

Garber ME, Troyanskaya OG, Schluens K, Petersen S, Thaesler Z, Pacyna-Gengelbach M, van de Rijn M, Rosen GD, Perou CM, Whyte RI, Altman RB, Brown PO, Botstein D and Petersen I. (2001). *Proc. Natl. Acad. Sci. USA*, **98,** 13784–13789.

Girard L, Zochbauer-Muller S, Virmani AK, Gazdar AF and Minna JD. (2000). *Cancer Res.*, **60,** 4894–4906.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD and Lander ES. (1999). *Science*, **286,** 531–537.

Hoque MO, Lee CC, Cairns P, Schoenberg M and Sidransky D. (2003). *Cancer Res.*, **63,** 2216–2222.

Janne PA, Tanenbaum DM, Beheshti J, Mark EJ, Johnson BE and Meyerson ML. (2002). *Proc. Am. Assoc. Can. Res.*, **43,** 3147a.

Jemal A, Murray T, Samuels A, Ghafoor A, Ward E and Thun MJ. (2003). *CA Cancer J. Clin.*, **53,** 5–26.

Kennedy GC, Matsuzaki H, Dong S, Liu WM, Huang J, Liu G, Su X, Cao M, Chen W, Zhang J, Liu W, Yang G, Di X, Ryder T, He Z, Surti U, Phillips MS, Boyce-Jacino MT, Fodor SP and Jones KW. (2003). *Nat. Biotechnol.*, **21,** 1233–1237.

Li C and Wong WH. (2001). *Proc. Natl. Acad. Sci. USA*, **98,** 31–36.

Lieberfarb ME, Lin M, Lechpammer M, Li C, Tanenbaum DM, Febbo PG, Wright RL, Shim J, Kantoff PW, Loda M, Meyerson M and Sellers WR. (2003). *Cancer Res.*, **63,** 4781–4785.

Lin M, Wei L-J, Sellers WR, Lieberfarb M, Wong WH and Li C. (2003). *Bioinformatics*, (in press).

Lindblad-Toh K, Tanenbaum DM, Daly MJ, Winchester E, Lui WO, Villapakkam A, Stanton SE, Larsson C, Hudson TJ, Johnson BE, Lander ES and Meyerson M. (2000). *Nat. Biotechnol.*, **18,** 1001–1005.

Mei R, Galipeau PC, Prass C, Berno A, Ghandour G, Patil N, Wolff RK, Chee MS, Reid BJ and Lockhart DJ. (2000). *Genome Res.*, **10,** 1126–1137.

Mitsudomi T, Oyama T, Nishida K, Ogami A, Osaki T, Sugio K, Yasumoto K, Sugimachi K and Gazdar AF. (1996). *Clin. Cancer Res.*, **2,** 1185–1189.

Osada H and Takahashi T. (2002). *Oncogene*, **21,** 7421–7434.

Primdahl H, Wikman FP, von der Maase H, Zhou XG, Wolf H and Orntoft TF. (2002). *J. Natl. Cancer Inst.*, **94,** 216–223.

Reich DE, Gabriel SB and Altshuler D. (2003). *Nat. Genet.*, **33,** 457–458.

Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL, Hunt SE, Cole CG, Coggill PC, Rice CM, Ning Z, Rogers J, Bentley DR, Kwok PY, Mardis ER, Yeh RT, Schultz B, Cook L, Davenport R, Dante M, Fulton L, Hillier L, Waterston RH, McPherson JD, Gilman B, Schaffner S, Van Etten WJ, Reich D, Higgins J, Daly MJ, Blumenstiel B, Baldwin J, Stange-Thomann N, Zody MC, Linton L, Lander ES and Altshuler D. (2001). *Nature*, **409,** 928–933.

Sekido Y, Fong KM and Minna JD. (2003). *Annu. Rev. Med.*, **54,** 73–87.

Stanton SE, Shin SW, Johnson BE and Meyerson M. (2000). *Genes Chromosomes Cancer*, **27,** 323–331.

Tseng JE, Kemp BL, Khuri FR, Kurie JM, Lee JS, Zhou X, Liu D, Hong WK and Mao L. (1999). *Cancer Res.*, **59,** 4798–4803.

Virmani AK, Fong KM, Kodagoda D, McIntire D, Hung J, Tonk V, Minna JD and Gazdar AF. (1998). *Genes Chromosomes Cancer*, **21,** 308–319.

Wang ZC, Lin M, Wei L-J, Li C, Miron A, Lodeiro G, Harris L, Ramaswamy S, Tanenbaum DM, Meyerson M, Iglehart JD and Richardson A. (2003). *Cancer Res.*, (in press).

Zabarovsky ER, Lerman MI and Minna JD. (2002). *Oncogene*, **21,** 6915–6935.

Supplementary Information accompanies the paper on Oncogene website (http://www.nature.com/onc) and also at research.dfci.harvard.edu/meyersonlab/snp/janne-oncogene2004