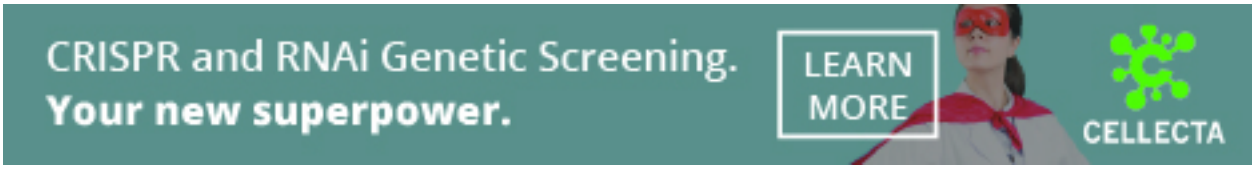
Global Survey of Chromatin Accessibility Using DNA Microarrays

M. Ryan Weil, Piotr Widlak, John D. Minna, et al.

Genome Res. 2004 14: 1374-1381
Access the most recent version at doi:10.1101/gr.1396104



Global Survey of Chromatin Accessibility Using DNA Microarrays

M. Ryan Weil,[1,4,5,6,9] Piotr Widlak,[2,8] John D. Minna,[3,5,7] and Harold R. Garner[1,4,5,6]

[1]*Program in Molecular Biophysics, Division of Cell and Molecular Biology, Southwestern Graduate School of Biomedical Science,* [2]*Department of Molecular Biology,* [3]*Hamon Center for Therapeutic Oncology Research,* [4]*Center for Biomedical Inventions,* [5]*Department of Internal Medicine,* [6]*Eugene McDermott Center for Human Growth and Development, and* [7]*Department of Pharmacology, UT Southwestern Medical Center, Dallas, Texas 75390, USA;* [8]*Department of Experimental and Clinical Radiobiology, Center of Oncology, Gliwice, 44-100, Poland*

An increasing number of studies indicate a central role for chromatin remodeling in the regulation of gene expression. Current methods for high-resolution studies of the relationship between chromatin accessibility and transcription are low throughput, making a genome-wide study impractical. To enable the simultaneous measurement of the global chromatin accessibility state at the resolution of single genes, we developed the Chromatin Array technique, in which chromatin is separated by its condensation state using either the solubility differences of mono- and oligonucleosomes in specific buffers or controlled DNase I digestion and selection of the large refractory (condensed) DNA fragments. By probing with a comparative genomic hybridization style microarray, we can determine the condensation state of thousands of individual loci and correlate this with transcriptional activity. Applying this technique to the breast tumor model cell line, MCF7, we found that when the condensation is homogeneous in the population of cells, expression is inversely proportional to the level of accessibility and the two methods of accessibility-based target selection correlate well. Using functional annotation and comparative genomic hybridization data, we have begun to decipher the possible biological implications of the relationship between chromatin accessibility and expression.

[Supplemental material is available online at www.genome.org and at http://famine.swmed.edu.]

In recent years, the study of transcriptional regulation by epigenetic mechanisms has enjoyed a renaissance because of advances in DNA microarray technology. These developments include the creation of high-throughput CpG methylation resequencing microarrays (Hatada et al. 2002) and advances in using DNA microarrays to probe Chromatin Immuno-Precipitation (ChIP) assays (Ren et al. 2000) on a genomic scale. Even with all these advances, perhaps one of the most important epigenetic regulation systems, chromatin architecture, has been overlooked. By mediating the availability of specific DNA sequences to regulatory proteins, chromatin accessibility in the form of chromatin condensation or relaxation is thought to be a major regulator of transcription (Orphanides and Reinberg 2002). Current methods of studying chromatin architecture either measure the accessibility of the genome as a whole (Banerjee and Hulten 1994) or of a few sub-kilobase regions (Reid et al. 2000), but no technique is currently available to easily and simultaneously measure the chromatin accessibility of the whole genome at kilobase resolution (Urnov 2003; Crawford et al. 2004).

In this paper, we describe a new method for using DNA microarrays to study the global chromatin accessibility state as a measure of nuclease accessibility in relation to expression at the resolution of single genes. The primary method we chose for isolating DNA by its chromatin accessibility state takes advantage of the solubility differences of histone H1-depleted mononucleosomes and histone H1-containing mono- and oligonucleosomes in the presence or absence of $MgCl_2$ and KCl to recover different chromatin fractions based on their activity states. This method's utility was demonstrated by Rose and Garrard (1984) to study the chromatin packing of immunoglobulin light chain genes in relation to their transcription during B-cell development. A second method was optimized to use the preferential sensitivity of transcriptionally active chromatin to DNase I cleavage (Weintraub and Groudine 1976) to recover the relatively resistant regions as the "condensed" fraction using fragment length selection. Both of these methods are currently used in high-resolution, low-throughput chromatin accessibility studies.

To make these techniques both high resolution and high throughput, we optimized microarray-based comparative genomic hybridization (CGH) methods using commercially available probe sets or microarrays to probe the chromatin accessibility state en masse (Pollack et al. 1999; Weil et al. 2002). This "Chromatin Array" allows us to overcome the limited resolution and throughput problems of previous methods (Banerjee and Hulten 1994; Reid et al. 2000) by using the multiplex nature of microarray experiments while retaining the high resolution of low-throughput chromatin accessibility measurement techniques. Because this new type of microarray experiment has a novel output, we developed methods to interpret the chromatin state from the relationship of the condensed fraction's hybridization intensity as compared with the intensity of total genomic DNA. These data can then be related to the absolute RNA expression level measured on an identical microarray.

To demonstrate the utility of the Chromatin Array method, we chose the cell line MCF7 because it is has been extensively studied by other groups (Pollack et al. 1999; Ross et al. 2000). We show that the chromatin solubility assay recovered fractions based on the condensation state of the chromatin, and that the microarray-based measurements could accurately measure the accessibility. The reproducibility of the condensation state mea-

[9]Corresponding author.
E-MAIL rweil@mednet.swmed.edu; FAX (214) 648-1445.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.1396104.

14:1374–1381 ©2004 by Cold Spring Harbor Laboratory Press ISSN 1088-9051/04; www.genome.org

surements was independently verified using two different methods to extract the condensed chromatin for microarray-based measurements.

To support the data analysis and interpretation, we used the Stanford Microarray Database (SMD) to validate our expression findings (Sherlock et al. 2001). Although the condensation state and expression measurement of a single gene may be of great value in transcriptional discovery, the biological relevance of the data on a global scale is possibly even more valuable. By relating function as defined by the Gene Ontology (GO) database (Ashburner et al. 2000) to the condensation state of large groups of genes, specific accessibility signatures of functionally related genes can be identified. These signatures are based on the different functional gene groupings of a particular accessibility state, and the differences in functional group assignments observed across the different accessibility states (Jimenez-Sanchez et al. 2001). These signatures can then be used to uniquely define a cell line. By comparing the signatures of multiple cell lines, it should be possible to identify the disease- and tissue-specific components of the signatures. Analysis of the accessibility data in light of both the condensation state of single genes as well as its global relationship to gene function makes the development of the Chromatin Array method a novel and important addition to study chromatin structure–function relationships.

RESULTS AND DISCUSSION

The Chromatin Array Accurately Measures the Accessibility State of the DNA Recovered by the Chromatin Solubility Assay

The chromatin solubility assay first uses micrococcal nuclease to generate mono- and oligonucleosomes that are separated into three fractions designated S1, S2, and P. The transcriptionally active DNA is found in the S1 and P fractions, which in MCF7 comprise ~68% of the total DNA. The S1 fraction is depleted in histone H1 and enriched in the high mobility group (HMG) proteins and heterogeneous ribonucleoproteins particles (HnRNPs), both of which are known to be associated with actively transcribed chromatin (Huang et al. 1986). Likewise, the P fraction is highly enriched in nonhistone proteins, and with further digestion, it can be partially converted to the S1 fraction (Rose and Garrard 1984; Huang et al. 1986). The S2 fraction represents ~32% of the total DNA and contains nucleosomes stoichiometrically associated with histone H1 and highly deficient in nonhistone proteins (Rose and Garrard 1984). This S2 fraction operationally represents the most condensed chromatin fraction as indicated by previous studies that have demonstrated that histone H1 is responsible for the formation of higher-order chromatin structures and is a general repressor of transcriptional activity (Allan et al. 1981; Croston et al. 1991). Figure 1 illustrates the differences between the S1, S2, and P fractions in terms of the protein compositions (Fig. 1A), DNA fractionation patterns (Fig. 1B), and as a molecular model of the digestion process (Fig. 1C). For the chromatin solubility assay study, the S2 fraction was chosen because it is the least sensitive to digestion variations during fractionation and therefore is the most reproducible.

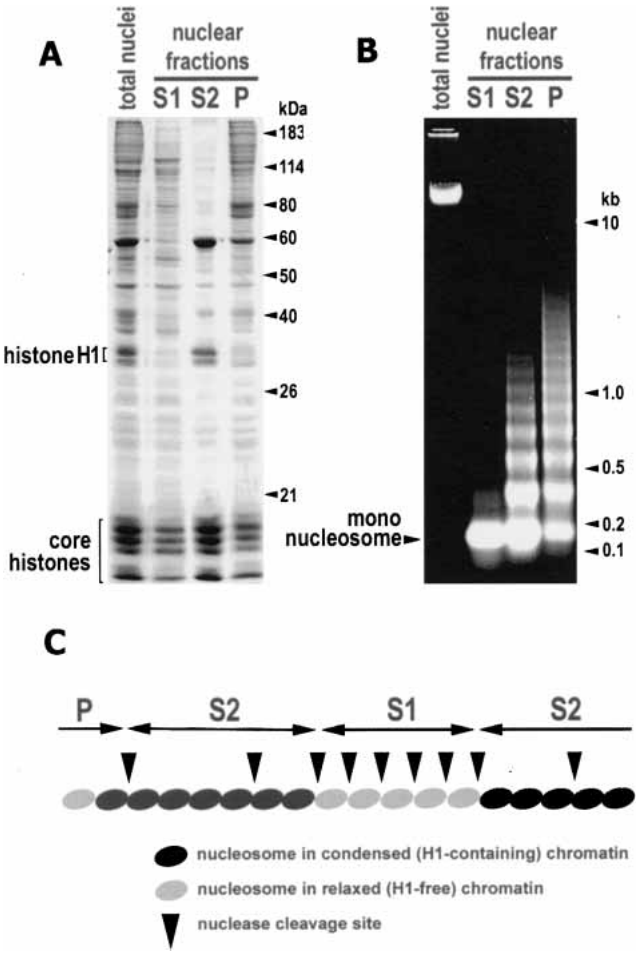


Figure 1 The chromatin solubility assay is used to extract the S1, S2, and P fractions, and the result of that process is depicted as the patterns seen in the gels of the resulting protein and DNA. The protein gel (*A*) shows the depletion of the H1 histone and the enrichment of nonhistone proteins in the S1 and P fractions. The DNA gel (*B*) shows the difference in the extent of the digestion between the fractions. The model (*C*) illustrates the basic concept of the chromatin fractionation assay.

Table 1. Measurements of the Chromatin Array Method's Reproducibility

	# Genes	Reproducibility	Concordance
Group 1: C/E >2	2620	86%	84%
Group 2: C/E <0.5	1493	76%	53%
Group 3: C/E <2 to >0.5	3842	69%	64%
Genes passing all filters	7955	—	—

The number of genes (19,437 possible) in each group that pass all data possessing filters is shown. Reproducibility refers to the percentage of genes that yield similar results in an independent replicate experiment of chromatin solubility fractionation on a different array platform. Concordance refers to the percentage of genes between the merged fragment length selection data and the chromatin solubility data that are consistently classifiable to each group.

The isolated S2 fraction, total genomic DNA, and RNA were fluorescently labeled and hybridized to the microarrays, and the data were extracted as discussed in the Methods section. From a total of 19,437 unique genes on the microarrays, 8860 genes passed the quality filters as discussed in the Methods section (see Table 1). Overall, the average of the S2 fraction in the experimental series, after thresholding the raw intensities, was a 2.25-fold depletion from the average of the total genomic DNA reference, likely because the microarrays used were designed for gene expression studies and are therefore biased toward genes known to be commonly expressed.

From Figure 2 it can be seen that, as expected (Foe 1978; Janicki et al. 2004), the condensation state of the genes (ex-

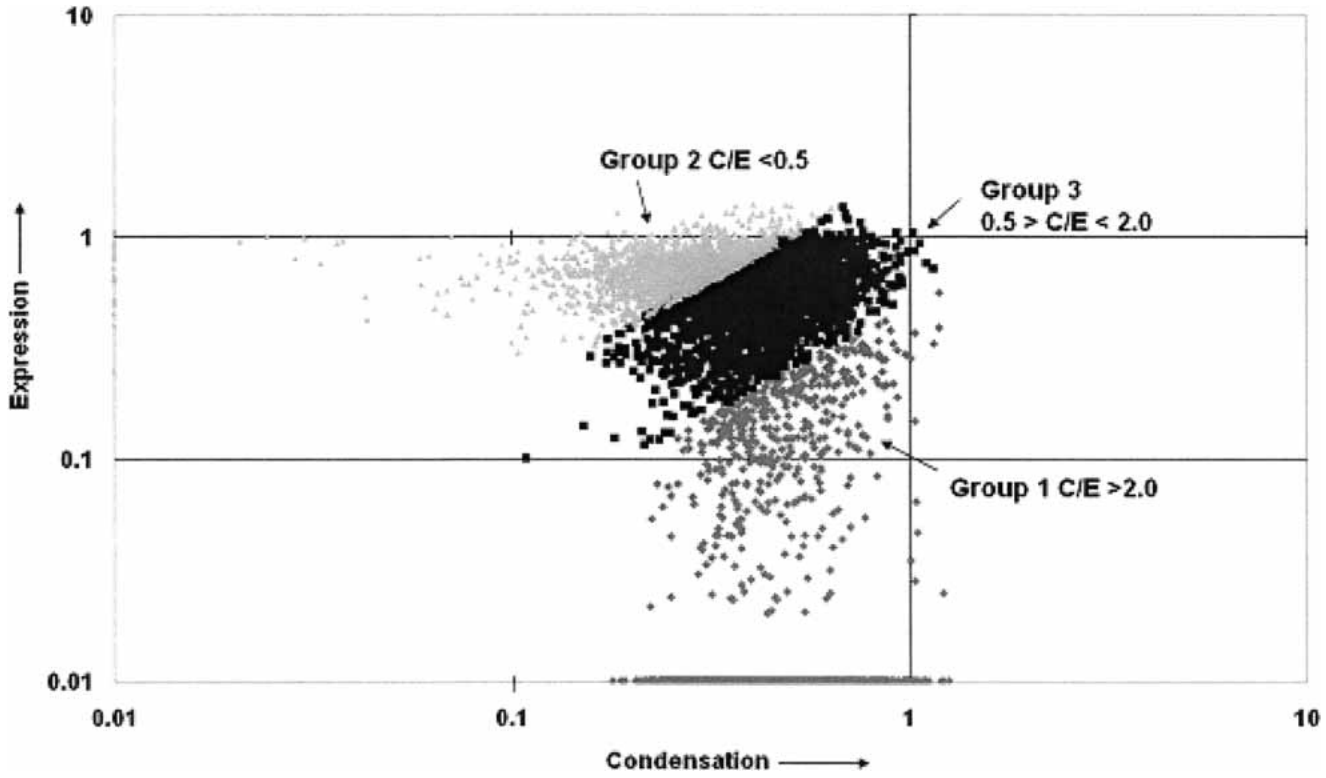


Figure 2 The log scale scatter plot of the co-normalized intensities (sequence abundances) illustrates the relationship between condensation (the ratio of the S2 fraction intensity to total genomic DNA intensity) and the absolute RNA expression. The light gray triangles are Group 2, which has a condensation to expression (C/E) ratio of <0.5. The black squares represent Group 3, which has indeterminate accessibility and a C/E ratio of <2.0 to >0.5. The gray diamonds represent Group 1, which has a C/E ratio of >2.0.

pressed as a ratio between the S2 fraction signal and total genomic DNA signal) shows a weak inverse correlation with the relative expression level, with a linear regression line slope of −0.3 and a Pearson correlation of −0.16. However, it also can be seen that the Chromatin Array data indicate a continuum of accessibility states. Because the accessibility data alone do not provide a means to delimit distinct condensation states with any degree of certainty, for the analysis it was necessary to use the nature of the known relationship between condensation and expression to understand and interpret the data. By using the ratio of the condensation state (as measured by relative intensity) to the RNA expression data set and delimiting the groups by the standard twofold difference in intensity as minimum statistically significant change, the data can separated into three groups with a high degree of confidence in the assignment. These three groups shown in Figure 2 are Group 1, which has a condensation/expression (C/E) ratio of >2.0 (most condensed and expressed at low levels); Group 2, which has a C/E ratio <0.5 (least condensed and most highly expressed); and Group 3, with a C/E ratio between <2.0 and >0.5 (indeterminate accessibility).

It is clear in Figure 2 that genes of indeterminate accessibility (Group 3), which show a direct correlation between accessibility and expression, represent a significant portion of the genes. The genes in this group at first glance are simply the set that cannot be assigned to a condensation group with high confidence. By filtering the data to remove the probes that cross-hybridize to nonfunctional elements, like pseudogenes, from the S2 fraction DNA and probes that are possibly not specific in a CGH application, the remaining genes likely to be heterogeneously expressed in a population of cells (Cho et al. 1998; Su et al. 2002). This analysis combined with directed experimentation has shown that this group has the potential to be a very rich source for new discoveries about chromatin's role in transcriptional control.

Fragment Length Selection Experiments Confirm the Results of the Chromatin Solubility Assay

Because a high-throughput validation of the chromatin solubility assay results is not possible by traditional means, a second technique to separate the condensed fraction was used for independent verification of the accessibility state while controlling for experimental artifacts. Fragment-length-based selection of condensed chromatin was the method of choice because it works on fundamentally different principles than the chromatin solubility assay. The fragment length selection process is based on the fact that active chromatin is preferentially sensitive to DNase I digestion (Weintraub and Groudine 1976). To achieve the experimental equivalent of a S2 fraction, we optimized a method for chromatin cleavage and subsequent electrophoretic fragment length selection to enrich for DNA that is condensed, and therefore is relatively resistant to digestion (see Fig. 3.) The 40% fraction, possessing fragment lengths from 6 to 23 kb, was selected as a tradeoff between enrichment for highly condensed DNA and generating sufficient material for the assay.

The microarray assays and data analysis of the fragment length selected samples were carried out as described in the Methods section. Genes that showed a threefold intensity change with respect to total genomic DNA reference were operationally defined as the condensed (for genes with increased intensity) or relaxed (for genes with decreased spot intensity) sets, respectively. The gene lists from the fragment length selection process

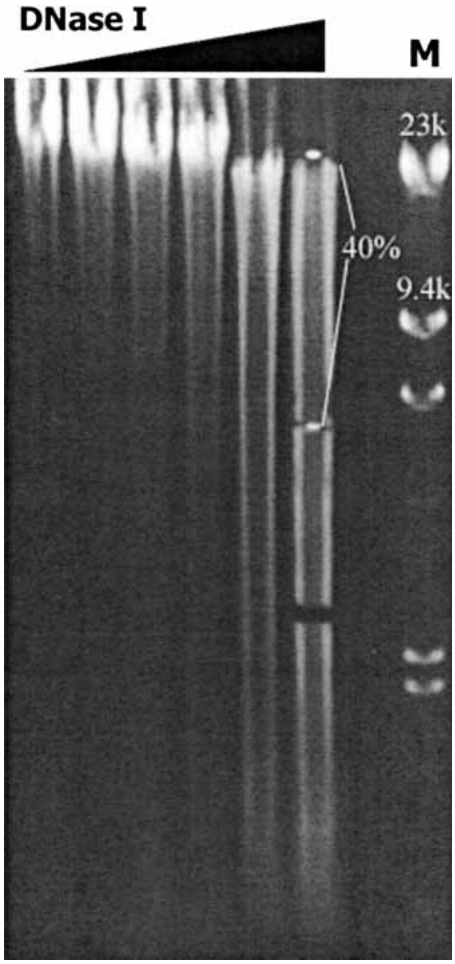


Figure 3 A fragment length selection gel showing the distribution of DNase I-cleaved fragments is used to identify the upper 40% of the total lane intensity to direct the excision of the packing enriched region. The lane digested with the highest concentration of DNase I (64 U) was chosen for its uniform distribution of fragments around the 5-kb median. This section of the gel was excised, and the DNA was recovered for labeling. (M) λ HindIII DNA mass standard.

were merged with the chromatin solubility assay results for comparison (see Table 1), and 84% of the condensed genes correlated between the two assays. The genes predicted to be relaxed showed a 53% concordance based on raw intensity comparisons. It was expected that the fragment length selection method would be more error-prone for relaxed genes because the recovery used for the fragment selection process was not optimized for the separation of the indeterminate accessibility set from the relaxed set. The strong correlation (84%) between these two techniques demonstrates that the chromatin array concept based on chromatin solubility is a robust method for accessibility analysis with an acceptable error rate in terms of the data volume.

Chromatin Accessibility Data Combined With Expression Levels Increases the Significance of Functional Relationships

The accessibility data were divided into three groups for analysis based on the relationship between the co-normalized sequence abundances in the condensation data and RNA expression data (Table 1). The groups were filtered using GO annotation (Ashburner et al. 2000) to identify the functional terms that were statistically overrepresented. The same analysis was repeated using the expression data alone. Table 2 shows that the addition of the chromatin array data substantially improved the significance scores for each identified term. The first group of 2620 genes (in Fig. 2, gray diamond) was the condensed group and showed a condensation to expression ratio of >2.0 with 86% reproducibility between replicate fractionations on different microarray types. Within this group, the genes had a statistically significant (p-values < 1.4E-2) overlap with seven major GO terms (Table 2). Upon inspection of the list of genes susceptible to pseudogene cross-hybridization, we determined that genes that would have been assigned to groups 1 and 2 obey the inverse relationship between condensation and expression, making false assignment to a GO term unlikely. Rerunning the analysis including the 289 genes with known pseudogenes strengthened the p-values of the relationships and identified a new statistically significant overlap with the GO term "DNA repair," a term that is known to be associated with oncogenesis.

The second group of 1493 genes (Fig. 2, light gray triangles) was the relaxed group and has a condensation to expression ratio of <0.5. The reproducibility between replicate fractionations on different style arrays for this group was 76%. With the genes that have known pseudogenes excluded, no GO term could be statistically associated with this group. We again reran the analysis with the 192 previously excluded genes with pseudogenes. This time the analysis found that several statistically significant GO terms were represented by this group (with p-values < 1.8E-2), most notably the ribosomal genes and RNA-binding genes (Table 2), both of which are constitutively expressed housekeeping genes and are known to be excluded from the S2 fraction (Davis et al. 1983) but also have large numbers of pseudogenes (Zhang et al. 2002). From the divergence seen in the GO terms that dominate the condensed and relaxed sets, clear differences in the gene functional types in each fraction can be identified. The division goes deeper, because whereas several signal transduction genes are in the relaxed fraction and are, indeed, transcribed, the subcategories of signal transduction genes are different between the groups (Table 2), and it is these differences that define the unique signature of a cell line.

The last group of 3842 genes (Fig. 2, black squares) displays indeterminate accessibility with a condensation to expression ratio of <2.0 but >0.5. This group showed 69% reproducibility between replicate fractionations on different style microarrays. The genes in this group cannot be identified from expression data alone, thus in Table 2 the "significance E" column is empty. However, the sample heterogeneity highlighted by the condensation data is also measured by expression analysis as well. Because of this, the genes represented by Group 3 are likely a source of a large amount of known and unknown error in expression analysis. Based on closer inspection of the relationship between condensation and RNA expression, this group was determined to consist of three subgroups. The first of these subgroups included 1166 genes that showed poor signal reliability because the signal was marginally above threshold in one sample and below threshold in the other. The second subgroup of 1412 genes showed a significant change in raw signal, but because normalization converted the change to less than twofold, they were excluded from further analysis. The last subgroup was the truly indeterminate set and had 1264 genes that showed both condensation and expression. Of these genes, 458 were involved in "cellular maintenance" with a statistically significant (p-values < 1.0E-3) association to several GO terms (Table 2). Of the remaining 806 genes, 428 genes had no annotated function and thus they could not be considered further. The remaining 378 showed a statistically significant (p-values < 8.5E-4) overlap with the GO term "nucleic acid binding." From these results, we conclude that this subgroup

Table 2. The Functional Signatures of the Primary Groups

GO Term Name	# Overlapping	Genes in Term	Significance C/E	Significance E
Group 1 C/E >2 (2620)				
Structural proteins	316	1576	1.21E-11	3.94E-10
Plasma membrane	247	1192	3.15E-10	1.56E-08
Transporter	268	1462	2.04E-05	1.13E-05
Metabolism	168	856	2.32E-05	4.84E-05
Signal transduction	525	1142	2.06E-04	7.00E-03
Transcription factor	154	856	1.44E-02	1.56E-02
DNA repair	13	106	1.90E-02	—
Wnt receptor signaling*	5	12	6.72E-03	—
Group 2 C/E <0.5 (1493)				
Ribosome	59	139	6.11E-24	4.88E-19
RNA binding	72	374	5.53E-08	1.02E-02
Cytosol	51	226	9.44E-08	1.02E-03
Mitochondrial	65	459	1.78E-02	—
Cell motility*	10	108	9.56E-03	—
Group 3 C/E <2 to >0.5 (3842)				
Signal transduction	252	1142	2.54E-82	—
Serine threonine kinase	29	245	4.93E-10	—
Proliferation	32	426	4.17E-06	—
Cell adhesion	31	445	4.03E-05	—
Cell cycle regulators	22	365	4.20E-03	—
Nucleic acid binding	67	1925	8.50E-04	—

The Gene Ontology term names of the primary functional groups represented by the three major groups are shown. The number of genes that overlap with the term and the total number of genes in the term are shown for comparison. The significance value of the calculated relationship for each term from the combined data set and from the expression data alone is given next. Terms with — in the last column could not be found by expression alone. Significance was calculated from the overlap of the master list to the test list, and tested using a Bonferroni-corrected hypergeometric distribution to determine a confidence value to reject the null hypothesis. Terms with asterisks were not identified as dominant terms in the initial search, but were found to be dominate subterms in that set.

of genes for which accessibility is indeterminate are not errors in the method but represent the genes that are heterogeneously expressed in the population because of cell cycle timing or environmental or genetic factors.

By Removing a Source of Heterogeneity in the Cell Population, Genes of Indeterminate Accessibility Can Be Properly Assigned

To test the assertion that heterogeneity in the sample is largely responsible for the genes with indeterminate accessibility (Group 3), we controlled the cell cycle state using serum starvation, which permits the capture of a homogeneous population of cells for chromatin fractionation. Flow cytometry is used to determine the fraction of cells in each phase of the cell cycle to estimate the expected heterogeneity of the population (Darzynkiewicz and Juan 1997). Of the cells from the unsynchronized culture, 65% were in G_1, and the subsequent chromatin fractionation experiment yielded ~32% of the DNA in the S2 fraction. By serum starving a culture for 24 h then allowing 6 h of recovery with 10% FBS before fractionation, >80% of the population was in the G_1 phase and the percent of the DNA in the S2 fraction increased to 42%. This change is due to the fact that in the early G_1 phase the chromatin is in its most condensed form (Warters and Lyons 1992). The increase in the S2 fraction DNA causes a 16% increase in overall intensity after normalization of the synchronized chromatin array from the intensity of the chromatin array done on the unsynchronized sample. Of the genes that differ by at least twofold between the synchronized and unsynchronized samples, the GO term "cell cycle regulators" is significantly (p-value < 6.0E-3) overrepresented. The majority of these 61 cell cycle genes are measured as having more accessible chromatin in synchronized cells, which correlates with the high expression levels measured in the unsynchronized culture. However, a small subset of genes go against this trend and show an increase in condensation such as *CDKN2B* (*NM_078487*), a tumor suppressor involved in preventing cells from entering G_1. Also of note is a strong increase in condensation of the mitotic motor *KNSL4* (*NM_07317*), which is only expressed in proliferating cells indicating that the synchronized cells are not ready to cycle. The results of the synchronization experiment indicate that by reducing a source of heterogeneity, the condensation state of the affected genes can be better correlated to the expression state. Although the synchronization experiment showed that the genes we expected to change did, and did so in a manner that we could predict, 69% of genes that changed were not reported to be involved in cell cycle regulation.

Accessibility Data Provides Significant Insight in the Relationship Between Gene Amplification and Deletions and Transcription

Using the publicly available CGH data for MCF7 from the Breast aCGH project (Pollack et al. 2002), we built a map of gene amplifications and deletions in MCF7. We chose this data set because it is widely available and has been rigorously reviewed. We found that genes that are amplified fell into two groups. The first group contained genes that were highly condensed with low expression, which was characterized by the GO terms in Table 3 with p-values < 2.6E-4. Conversely, the second group had genes that were relaxed and showed high expression. This group was dominated by the GO terms "ribosomal" and "molecular chaperone," which are broad groups of "housekeeping" genes (see Table 3).

Genes having CGH measured copy number loss, high condensation, and weak expression map to the GO terms "cell growth and maintenance," "cell communication," and "tumor suppressors" with p-values < 2.9E-3 (Table 3). These terms in-

Table 3. The Functional Signatures of Chromosomal Alterations Measured by Accessibility State

GO term name	No. overlapping	Genes in term	Significance
Amplified with increased condensation			
Cell communication	147	3844	4.68E-09
Cell growth and maintenance	131	3385	5.11E-08
Transcription factors	46	926	2.57E-04
Amplified with decreased condensation			
Ribosome	6	93	4.39E-03
Molecular chaperone	5	96	1.02E-02
Copy numer lose with increased condensation			
Cell growth and maintenance	49	3385	6.64E-05
Cell communication	52	3844	2.04E-04
Tumor suppressor	5	177	2.87E-03

The Gene Ontology term names of the functional signatures represented in the CGH data as measured by accessibility state. Also shown is the number overlapping with the term and the total genes in the term with the significance values. Significance was calculated from the overlap of the master list to the test list, and tested using a Bonferroni-corrected hypergeometric distribution test to determine a confidence value to reject the null hypothesis.

clude the gene functions traditionally associated with transformation, immortalization, and general oncogenesis, and likely represent a substantial portion of the tumor-specific signature of MCF7. The genes that both showed a copy number loss as seen by CGH and had reduced condensation (but strong expression) did not map to any GO terms with meaningful significance.

The CGH data also yielded one additional relationship that was not expected. Genes found in the indeterminate accessibility group (Group 3) showed strong ties to CGH copy number changes including both copy number loss (p-value = 3.0E-8) and gain (p-value = 9.5E-8). The 1166 genes from Group 3 that were marginally above threshold strongly overlapped (p-value = 4.5E-25) with amplified regions. In summary, the addition of the CGH data to the accessibility and expression data helps refine the interpretation of the full data set.

Accessibility Differences Show Patterning Based on Chromosomal Location

It is known that in cancer the alteration of the gene copy number can modify transcriptional activity to provide a selective advantage. However, because the regions that are altered are often large and contain genes that could be detrimental if deregulated, additional regulatory mechanisms are required to provide fine control of these regions (Pollack et al. 2002; Albertson et al. 2003). During the CGH and accessibility analysis, we identified numerous regions of copy number change that shared a similar condensation state. We found that amplified and condensed regions, sometimes only a few 100 kb wide, exhibiting low expression were interspersed at chromosomal positions 11q13, 7q21.11 to 7q22, 5q31 to 5q32, and 1q31.1 to 1q32. Amplified regions that primarily exhibit low condensation and increased expression were also seen. The 5q11 to 5q12 region is the most tantalizing example of this relationship, with the gene of greatest interest in that region being *DHFR* (*NM_000791*; Sullivan and Bickmore 2000), which by amplification can confer resistance to the chemotherapeutic compound methotrexate (Banerjee et al. 2002). Instances of low condensation and high expression from regions with copy number loss were also seen and include 22q13.1, which is the location of *H1F0* (*NM_005318*), a histone 1 variant known to be involved in maintaining basal gene repression and differentiation. The regions with copy number loss, high condensation, and low expression were also numerous and include the 3p21 region, which contains numerous tumor-suppressor genes, and in the 22q13.2 region, the location of the tumor suppressor *RBX1* (*NM_014248*). In conclusion, the use of accessibility information to link the chromosomal state to expression is fundamental to the understanding of diseases like cancer, where large, enigmatic chromosomal alterations occur long before the selection that would make such changes oncogenic (Albertson et al. 2003). It is this ability to make transcriptional discoveries by mapping accessibility changes to transcriptional control of genes or regions that makes the Chromatin Array method powerful.

METHODS

Cell Culture

The cell line used for this study was MCF7 (American Type Culture Collection number HTB-22), which is a well-studied female breast cancer model line. The cells were grown in RPMI with 10% fetal bovine serum (FBS) to 90% confluence. Harvesting was done using 0.05% trypsin with two washes in sterile 4°C phosphate-buffered saline. The harvested cells were counted and aliquoted for nuclease digestion. The remaining cells were pelleted and flash-frozen for DNA or RNA extraction.

Cell Cycle Synchronization by Serum Starvation and Population Analysis

For the synchronized cultures, the cells were allowed to grow to 80% confluence under normal conditions. Then the media was removed, and the cells were washed twice in RPMI without FBS. The washed cells were grown in RPMI alone for 24 h to serum-starve and synchronize. The media was then replaced with RPMI supplemented with 10% FBS, and the cells were allowed to recover for 6 h before harvesting for nuclease digestion. An aliquot was ethanol-fixed according to the protocol developed by Darzynkiewicz, and stained with propidium iodide. The stained cells were analyzed by flow cytometry to determine the proportion of cells in each phase of the cell cycle.

Micrococcal Nuclease Digestion and the Chromatin Fractionation Assay

The chromatin fractionation assay was performed as described by Rose and Garrard (1984). Cells were lysed by incubation with a hypotonic buffer supplemented with nonionic detergent NP-40; the nuclei were washed, resuspended in digestion buffer, and incubated with micrococcal nuclease to ~10% acid solubility. The nuclei were then centrifuged to obtain the S1 fraction, which represents the chromatin that was soluble in a buffer containing 5 mM $MgCl_2$ and 100 mM KCl. The nuclei were then resuspended in 2 mM EDTA for lysis, and the S2 fraction was recov-

ered as the supernatant following centrifugation. The pellet that remains after removal of the S2 fraction was the P fraction. The DNA was purified from the samples by digestion with Proteinase K, extraction with phenol/chloroform, and ethanol precipitation. DNA was resuspended in Tris/EDTA (pH 8.0), treated with RNase, and repurified as above.

DNase I Digestion and Fragment Length Selection

The DNase I digestion protocol was based on Roque et al. (1996). Cells were permeabilized by 0.11% lysolecithin, and incubated for 10 min in the presence of DNase I at concentrations up to 64 U/2.0 $\times$ 10^7 cells. The DNA was purified as described above, and 15 µg of each sample was separated electrophoretically on 0.8% Seakem GTG agarose (BMA) gels in 1$\times$ TAE for 10 h at 60 V and then stained with ethidium bromide and visualized under UV light. The band that represents the sample with the uniform distribution of fragments from 23 kb to 1 kb with a median fragment length ~5 kb was selected from the DNase I concentration series (Fig. 3), and the upper portion of the lane containing 40% of the total DNA by intensity was excised. The sample was then cut into pieces before being frozen at −20°C, snap-thawed at 37°C, and placed in tubes with 0.22-µm filters (Gelman). The samples were eluted by centrifugation at 13,000*g* for 10 min. The elutant was concentrated using a 30,000-Da centrifuge filter (Millipore) and resuspended in 20 µL of nuclease-free water.

DNA Labeling

Fluorescent labeling was achieved by random priming according to the protocol devised by Pollock et al. (1999) for their microarray-based Comparative Genomic Hybridizations (CGH) and optimized for use as a genomic standard (Weil et al. 2002). Direct labeling of the sample was accomplished by the incorporation of Cy5 fluorophore (Amersham Biosciences) into ~2 µg of sample. An undigested control sample of total genomic DNA from the same cell line and at the same concentration was labeled with the Cy3 fluorophore (Amersham Biosciences) for use as the genomic reference. Sample purification and blocking was performed as specified in our protocol.

Sample Hybridization

Sample hybridization was performed using the protocol we optimized (Weil et al. 2002) with modifications described here. The microarrays initially used were produced by the UTSW microarray core and were spotted 70-nt oligonucleotides from Operon (QIAGEN). The arrays were printed on poly-L-lysine-coated slides and consist of ~22,000 UniGene clusters (Schuler et al. 1996) and ~2000 blanks, in addition to the control spots. The hybridization was done at 62°C using a recirculating waterbath, in single slide hybridization chambers (Telechem), under 24 $\times$ 60-mm coverslips. The chromatin solubility assay experiments were repeated on the MWG Human 30K oligonucleotide microarrays (MWG Biotech AG) using only the A and B arrays for a total unique gene count of 19,473. The hybridization temperature was adjusted according to the MWG protocols to 42°C. Both types of microarrays were hybridized for 16–20 h prior to washing. The three-buffer washing process is standard, with the first buffer (0.2% SDS and 2$\times$ SSC) warmed to 30°C. The second buffer was 0.2$\times$ SSC, and the third buffer was 0.1$\times$ SSC. The microarray was then dried by centrifugation. The dry microarray was scanned in an Axon 4000B scanner at a resolution of 10 µm/pixel, with adjustments to the photomultiplier tube voltage used to balance the signal in the two channels.

Expression Array

The expression array analysis was performed following the protocols published by Eisen and Brown (1999). RNA was extracted from the reserved snap-frozen cells using the RNeasy kit (QIAGEN). The resulting RNA was quantified and quality-checked by the Bioanalyzer (Agilent). The RNA (20 µg each) was labeled with the Cyscribe kit (Amersham) using Cy3 or Cy5 fluorophores. By replicating the same RNA in both channels, the dye-dependent variables were removed, and each array produced two data points for each gene. The RNA was then degraded by the addition of a sodium hydroxide solution at 70°C for 10 min, and the samples were subsequently neutralized with an equal amount of hydrochloric acid. The samples were washed as before, and the appropriate blocking agents and buffers were added. Hybridization conditions were similar to the conditions used for the packing microarrays (Weil et al. 2002).

Data Analysis and Interpretation

Using the GenePix 4.0 analysis package (Axon), the resulting images were overlaid with a grid to define the spots, and the data were manually flagged to remove poor quality spots (Fielden et al. 2002). The data were extracted from the seven chromatin microarrays and the six expression repeats and imported into GeneSpring 6.0 (Silicon Genetics) for normalization and analysis. The normalization of the chromatin microarrays consisted of dividing the S2 fraction signal by the genomic DNA control signal. The expression data were normalized to the median of the genes across the experimental series. The cross-gene error model was used to remove any genes that do not have sufficient signal to exceed the calculated noise threshold of 125 intensity units in the experimental series. To minimize the inaccuracies introduced by pseudogene cross-hybridization, all genes that are known to have a pseudogene (Zhang et al. 2003) were removed from the primary analysis. The additional GO annotation was done using the "find similar lists" function in GeneSpring. The Simplified Gene Ontology lists used represent only the subset of genes in each term that is actually probed for on the microarray because the lists are generated by parsing the gene annotation file. The similarity was determined using a Bonferroni-corrected hypergeometric distribution test to find the significance of the overlap of the user-generated or master list to all the possible Simplified Gene Ontology lists or test lists. Only statistically significant lists with *p*-values of ~0.01 were selected for manual review of the content before being included as an overlapping term.

The expression results were verified for a subset of genes in each group using the SOURCE database (Diehn et al. 2003) to annotate the experimental data with data from published array experiments (Sherlock et al. 2001), serial analysis of gene expression (SAGE) data (Lash et al. 2000), and sequence data from Ensembl (Hubbard et al. 2002). The accessibility and expression data for this experimental series are stored in an MIAME-compliant database (Brazma et al. 2001) that can be accessed from our Web site at http://famine.swmed.edu.

Additionally, we found that a comparison of total digested DNA to total undigested naked DNA as a secondary normalization was not necessary. The observed differences between total digested to undigested DNA related solely to a small number of the most sensitive DNA sequences, which were also seen as the most relaxed genes in the Chromatin Array. This group was excluded from the analysis.

ACKNOWLEDGMENTS

This work was supported by NIH/NCI grant R33-CA81656, NIH/NCI SPORE grant 50CA70907, and NIH grant T32-HL07360. We thank William T. Garrard for his assistance in the development of these methods and for his critical reading of the manuscript.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

Albertson, D.G., Collins, C., McCormick, F., and Gray, J.W. 2003. Chromosome aberrations in solid tumors. *Nat. Genet.* **34:** 369–376.

Allan, J., Cowling, G.J., Harborne, N., Cattini, P., Craigie, R., and Gould, H. 1981. Regulation of the higher-order structure of chromatin by histones H1 and H5. *J. Cell Biol.* **90:** 279–288.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al. 2000. Gene ontology: Tool for the unification of biology. The Gene

Ontology Consortium. *Nat. Genet.* **25:** 25–29.
Banerjee, S. and Hulten, M.A. 1994. Sperm nuclear chromatin transformations in somatic cell-free extracts. *Mol. Reprod. Dev.* **37:** 305–317.
Banerjee, D., Mayer-Kuckuk, P., Capiaux, G., Budak-Alpdogan, T., Gorlick, R., and Bertino, J.R. 2002. Novel aspects of resistance to drugs targeted to dihydrofolate reductase and thymidylate synthase. *Biochim. Biophys. Acta* **1587:** 164–173.
Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C.A., Causton, H.C., et al. 2001. Minimum information about a microarray experiment (MIAME)—Toward standards for microarray data. *Nat. Genet.* **29:** 365–371.
Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J., et al. 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* **2:** 65–73.
Crawford, G.E., Holt, I.E., Mullikin, J.C., Tai, D., Blakesley, R., Bouffard, G., Young, A., Masiello, C., Green, E.D., National Institutes Of Health Intramural Sequencing Center, et al. 2004. Identifying gene regulatory elements by genome-wide recovery of DNase hypersensitive sites. *Proc. Natl. Acad. Sci.* **101:** 992–997.
Croston, G.E., Kerrigan, L.A., Lira, L.M., Marshak, D.R., and Kadonaga, J.T. 1991. Sequence-specific antirepression of histone H1-mediated inhibition of basal RNA polymerase II transcription. *Science* **251:** 643–649.
Darzynkiewicz, Z. and Juan, G. 1997. *Current protocols in cytometry* (ed. J.P. Robinson), pp. 7.5.1–7.5.24. J. Wiley & Sons, New York.
Davis, A.H., Reudelhuber, T.L., and Garrard, W.T. 1983. Variegated chromatin structures of mouse ribosomal RNA genes. *J. Mol. Biol.* **167:** 133–155.
Diehn, M., Sherlock, G., Binkley, G., Jin, H., Matese, J.C., Hernandez-Boussard, T., Rees, C.A., Cherry, J.M., Botstein, D., Brown, P.O., et al. 2003. SOURCE: A unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Res.* **31:** 219–223.
Eisen, M.B. and Brown, P.O. 1999. DNA arrays for analysis of gene expression. *Methods Enzymol.* **303:** 179–205.
Fielden, M.R., Halgren, R.G., Dere, E., and Zacharewski, T.R. 2002. GP3: GenePix post-processing program for automated analysis of raw microarray data. *Bioinformatics* **18:** 771–773.
Foe, V.E. 1978. Modulation of ribosomal RNA synthesis in *Oncopeltus fasciatus*: An electron microscopic study of the relationship between changes in chromatin structure and transcriptional activity. *Cold Spring Harb. Symp. Quant. Biol.* **42 Pt 2:** 723–740.
Hatada, I., Kato, A., Morita, S., Obata, Y., Nagaoka, K., Sakurada, A., Sato, M., Horii, A., Tsujimoto, A., and Matsubara, K. 2002. A microarray-based method for detecting methylated loci. *J. Hum. Genet.* **47:** 448–451.
Huang, S.Y., Barnard, M.B., Xu, M., Matsui, S., Rose, S.M., and Garrard, W.T. 1986. The active immunoglobulin κ chain gene is packaged by non-ubiquitin-conjugated nucleosomes. *Proc. Natl. Acad. Sci.* **83:** 3738–3742.
Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., et al. 2002. The Ensembl genome database project. *Nucleic Acids Res.* **30:** 38–41.
Janicki, S.M., Tsukamoto, T., Salghetti, S.E., Tansey, W.P., Sachidanandam, R., Prasanth, K.V., Ried, T., Shav-Tal, Y., Bertrand, E., Singer, R.H., et al. 2004. From silencing to gene expression: Real-time analysis in single cells. *Cell* **116:** 683–698.
Jimenez-Sanchez, G., Childs, B., and Valle, D. 2001. Human disease genes. *Nature* **409:** 853–855.
Lash, A.E., Tolstoshev, C.M., Wagner, L., Schuler, G.D., Strausberg, R.L., Riggins, G.J., and Altschul, S.F. 2000. SAGEmap: A public gene expression resource. *Genome Res.* **10:** 1051–1060.
Orphanides, G. and Reinberg, D. 2002. A unified theory of gene expression. *Cell* **108:** 439–451.
Pollack, J.R., Perou, C.M., Alizadeh, A.A., Eisen, M.B., Pergamenschikov, A., Williams, C.F., Jeffrey, S.S., Botstein, D., and Brown, P.O. 1999. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat. Genet.* **23:** 41–46.
Pollack, J.R., Sorlie, T., Perou, C.M., Rees, C.A., Jeffrey, S.S., Lonning, P.E., Tibshirani, R., Botstein, D., Borresen-Dale, A.L., and Brown, P.O. 2002. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc. Natl. Acad. Sci.* **99:** 12963–12968.
Reid, J.L., Iyer, V.R., Brown, P.O., and Struhl, K. 2000. Coordinate regulation of yeast ribosomal protein genes is associated with targeted recruitment of Esa1 histone acetylase. *Mol. Cell* **6:** 1297–1307.
Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., et al. 2000. Genome-wide location and function of DNA binding proteins. *Science* **290:** 2306–2309.
Roque, M.C., Smith, P.A., and Blasquez, V.C. 1996. A developmentally modulated chromatin structure at the mouse immunoglobulin κ 3′ enhancer. *Mol. Cell. Biol.* **16:** 3138–3155.
Rose, S.M. and Garrard, W.T. 1984. Differentiation-dependent chromatin alterations precede and accompany transcription of immunoglobulin light chain genes. *J. Biol. Chem.* **259:** 8534–8544.
Ross, D.T., Scherf, U., Eisen, M.B., Perou, C.M., Rees, C., Spellman, P., Iyer, V., Jeffrey, S.S., Van de Rijn, M., Waltham, M., et al. 2000. Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.* **24:** 227–235.
Schuler, G.D., Boguski, M.S., Stewart, E.A., Stein, L.D., Gyapay, G., Rice, K., White, R.E., Rodriguez-Tome, P., Aggarwal, A., Bajorek, E., et al. 1996. A gene map of the human genome. *Science* **274:** 540–546.
Sherlock, G., Hernandez-Boussard, T., Kasarskis, A., Binkley, G., Matese, J.C., Dwight, S.S., Kaloper, M., Weng, S., Jin, H., Ball, C.A., et al. 2001. The Stanford Microarray Database. *Nucleic Acids Res.* **29:** 152–155.
Su, A.I., Cooke, M.P., Ching, K.A., Hakak, Y., Walker, J.R., Wiltshire, T., Orth, A.P., Vega, A.R., Sapinoso, L.M., Moqrich, A., et al. 2002. Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl. Acad. Sci.* **99:** 4465–4470.
Sullivan, B.A. and Bickmore, W.A. 2000. Unusual chromosome architecture and behaviour at an HSR. *Chromosoma* **109:** 181–189.
Urnov, F.D. 2003. Chromatin as a tool for the study of genome function in cancer. *Ann. NY Acad. Sci.* **983:** 5–21.
Warters, R.L. and Lyons, B.W. 1992. Variation in radiation-induced formation of DNA double-strand breaks as a function of chromatin structure. *Radiat. Res.* **130:** 309–318.
Weil, M.R., Macatee, T., and Garner, H.R. 2002. Toward a universal standard: comparing two methods for standardizing spotted microarray data. *Biotechniques* **32:** 1310–1314.
Weintraub, H. and Groudine, M. 1976. Chromosomal subunits in active genes have an altered conformation. *Science* **193:** 848–856.
Zhang, Z., Harrison, P., and Gerstein, M. 2002. Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. *Genome Res.* **12:** 1466–1482.
Zhang, Z., Harrison, P.M., Liu, Y., and Gerstein, M. 2003. Millions of years of evolution preserved: A comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res.* **13:** 2541–2558.

WEB SITE REFERENCES

http://famine.swmed.edu; the Garner lab GeneTraffic Server.

Received April 11, 2003; accepted in revised form April 30, 2004.