SUPPLEMENTARY INFORMATION

Chromatin accessibility pre-determines glucocorticoid receptor binding patterns

Sam John, Peter J. Sabo, Robert E. Thurman, Myong-Hee Sung, Simon C. Biddie, Thomas A. Johnson, Gordon L. Hager, and John A. Stamatoyannopoulos

SUPPLEMENTARY TABLES

SUPPLEMENTARY TABLE 1. **DNase I sensitive regions in the baseline (pre-hormone) and post-hormone (dexamethasone treated) states in the murine mammary adenocarcinoma cell line, 3134**. Columns: Chr,start,stop.

SUPPLEMENTARY TABLE 2. **DNase I hypersensitive sites (DHSs) in the baseline (pre-hormone) state in the murine mammary adenocarcinoma cell line, 3134**. Columns: Chr,start,stop.

SUPPLEMENTARY TABLE 3. **DNase I hypersensitive sites (DHSs) in post dexamethasone-treated 3134 cells**. Columns: Chr,start,stop.

SUPPLEMENTARY TABLE 4. **GR occupancy sites (FDR 0%) in the murine mammary adenocarcinoma cell line, 3134**. Columns: Chr,start,stop.

SUPPLEMENTARY TABLE 5. **Expression analysis (Affymetrix mouse exon array 1.0) of mammary (3134) and pituitary (AtT-20) cells**. Columns: Gene ID, $\log_2$ ratios of expression at 0h, 2h, 4h and 8h (post-dexamethasone treatment), gene symbol.

SUPPLEMENTARY TABLE 6. **GRBE sequence classes with greater than 50 instances in the genome.** Chromatin Context Coefficient (CCC) classes in the murine genome. mCCC = CCC values in computed from mammary (3134) cells. pCCC = CCC values computed from pituitary (AtT-20) cells. Columns: Class logo, -log p of the GRBE, mCCC, pCCC and number of genomic instances.

SUPPLEMENTARY TABLE 7. **DNase I sensitive regions in the baseline (pre-hormone) and post-hormone (dexamethasone treated) states in the murine pituitary cell line, AtT-20**. Columns: Chr,start,stop.

SUPPLEMENTARY TABLE 8. **DNase I hypersensitive sites (DHSs) in the baseline (pre-hormone) state in the murine pituitary cell line, AtT-20**. Columns: Chr,start,stop.

SUPPLEMENTARY TABLE 9. **DNase I hypersensitive sites (DHSs) post-hormone in AtT-20 cells**. Columns: Chr,start,stop.

SUPPLEMENTARY TABLE 10. **GR occupancy sites in the murine pituitary cell line, AtT-20 (FDR 0%)**. Columns: Chr,start,stop.

SUPPLEMENTARY NOTES

Gene expression (exon array) analysis. RNA was extracted from cells either vehicle treated or treated with 100 nM hormone (dexamethasone, Dex) for 0h, 2h, 4h or 8h. 3134 and AtT-20 RNA for microarray analysis was prepared via standard manufacturer protocols (Qiagen, Valencia, Ca) using cells resuspended in Trizol reagent (Invitrogen, Carlsbad, Ca). RNA from untreated and dex treated cells were labeled with biotin-CTP using manufacturer's recommendations (Affymetrix, Santa Clara, Ca). Biotinylated RNA was then used to hybridize mouse exon 1.0 ST arrays. Probe-level CEL files were processed through Affymetrix Expression Console using RMA summarization and median normalization methods. CHP files were generated based on core annotation confidence for exons. For gene-level expression values, expression summary value for each gene was obtained by taking the trimmed mean of all exon-level expression log intensities that were between the upper and lower quartiles (from half the exons around the median log intensity, see Supplementary Table 5).

***De novo* motif discovery and motif matching**. We used the well-established MEME algorithm[33] to search for motifs in the top 500 GR ChIP peaks (150 bp width). GR peaks are defined as either pre- or re-programmed peaks. DNase I sensitive sites in Dex- and Dex+ samples which overlapped peaks in Dex+ GR ChIP samples were designated pre-programmed GR peaks while DNase I sensitive regions in Dex+ but not Dex- samples which overlapped peaks in Dex+ GR ChIP samples were defined as re-programmed GR peaks. The sequence was repeat masked prior to searching. Settings of a minimum and maximum motif size of 8 bp and 40 bp respectively with a maximum of 100 motifs were used for the search. The resulting motifs were searched against the Transfac database using Tomtom to identify known motifs. Unknown motifs with MEME e-values $< 10^{-3}$ were further screened by comparing average p-values for sites within DNase I sensitive regions that overlapped GR ChIP peaks with DNase I sensitive regions that did not overlap GR ChIP peaks. Stronger motifs, those with lower p-values, $< 10^{-4}$, in the GR ChIP peaks vs DNase I sensitive regions alone were considered candidates for further analysis. These motifs were used to scan pre-programmed DNase I sensitive regions sites using MAST with a p-value cutoff of 10^{-3}. The logo for the motifs identified in 3134 and AtT-20 are in

Figure 4 and Supplementary Figure 7d.

Classification of glucocorticoid receptor binding element (GRBE) classes. MEME was applied to 3134 GR ChIP-Seq data in order to discover enriched motifs within GR bound regions in an unbiased fashion. The input for MEME was genomic sequences that correspond to the top 500 GR ChIP peaks of width 150 bp each. The most highly enriched sequence pattern conformed to a previously known palindromic Glucocorticoid Receptor Binding Element (GRBE motif). A 15-mer position weight matrix for a GR binding motif was constructed from this MEME output. We scanned the repeat-masked mouse genome for sequence matches to the GRBE motif by applying the algorithm MAST[33] with the GRBE position weight matrix as input. All matches with the MAST position p-value less than 0.001 were identified, based on the default random sequence model. Their genomic coordinates and DNA sequences were retrieved for subsequent classification. We then grouped the individual 15-mer instances (~2.2 million) into motif sequence classes based on their nucleotide usage at non-degenerate positions of the GRBE motif. For this purpose, motif positions 3, 7, 8, 9, and 13 had little information content and were considered degenerate. There were 2,866 GRBE sequence classes that had at least 50 GRBE occurrences genome-wide (**Supplementary Table 6).**

Computation of Chromatin Context Coefficient (CCC). To ensure statistical validity in assessing the effect of chromatin context, we identified 2,866 GRBE classes with 50 or more genomic instances (range: 50 -7,385), of which 1,100 were statistically well-defined. For each GRBE class, CCC is defined as the ratio of [the proportion of GR binding to a specific GRBE sequence in open chromatin] relative to [the proportion of GR binding to the same GRBE sequence in closed chromatin] (**Figure 2c**). In this schema, high values of CCC represent high chromatin context-dependence of GR binding, while low values indicate GRBEs that are relatively insensitive to local chromatin context. The latter category represents sites that have the potential to escape the dominant effect of the chromatin structural landscape, and initiate local remodeling. Notably, no CCC values <1 were observed, demonstrating that GR binding was universally enhanced by residence of a specific GR recognition element within accessible chromatin prior to hormone.

Filtering for sequence artifacts related to altered genomic copy numbers. Sequencing artifacts derived from altered genomic copy numbers of specific elements or small regions (e.g., satellite sequences) are frequently observed in both DNase I and ChIP data, and typically manifest as a high concentration of tags in a small area. We attempt to remove these artifacts in two ways. First, we observe that satellite repeats are a significant source of artifacts, and we therefore simply mask satellite repeat regions from our final hotspot and peak sets. Next, we apply a scanning procedure across the genome that identifies 50bp windows (each containing at least 5 mapped tags) that contain at least 80% of the tags in a 250bp surrounding window. A priori, 50bp is significantly smaller than the expected size of a *bona fide* GR binding event or a DNase I hypersensitive site. We mask all such flagged artifact regions from our final hotspot and peak sets.

Replicate-concordant data sets. We define replicate concordant sets for DNase I and GR ChIP as follows. We use the generally more conservative definitions imposed by replicate concordance for the DNase I sets used in the CCC and aggregate plot (**Figure 2 and Supplementary Figure 4**) analyses. For both GR ChIP and DNase I experiments, we generated two replicates for each condition (with and without Dex), for a total of four individual sets per experiment and tissue. For DNase I, we define regions of replicate concordance within a fixed condition as the intersection of merged, minimally thresholded (z>=2) hotspots from each replicate. We then combine the tags from both replicates for that condition and call and score FDR-thresholded hotspots and peaks in the combined tag set. An FDR-thresholded replicate concordant DNase I set is defined as the FDR-thresholded set in the combined dataset, intersected with the replicate concordant regions for that condition. We do not restrict replicate concordant Dex- hotspots to those that overlap Dex+ hotspots. For GR ChIP, we take a more conservative approach of thresholding both before and after taking intersections. For each condition we call FDR-thresholded peaks and hotspots in each replicate separately, and then define regions of replicate concordance as the intersection of merged hotspots at the FDR 0% level from both replicates. We then enforce a specified degree of replicate concordance by further thresholding the results by density (sliding window tag counts), considering only FDR 0% peaks from either replicate whose values are over a given absolute value, as follows. When each replicate is considered separately, we define the degree of replicate concordance as the percentage of each replicate's density thresholded peaks that fall in the replicate concordant regions. This percentage generally increases as the density threshold increases. We take as the 99% replicate concordant set the peaks from the larger replicate thresholded by density at a level to achieve 99% replicate concordance by this measure.

Calculation of enrichment p-values. Throughout the text we provide p-values for the overlap of one set of genomic features with another. In most instances we use the binomial distribution (R function pbinom) for these calculations. For the enrichment p-values in the GRBE/AP-1/HNF3 motif analysis section, we use the one-sided tests for the relationship between two proportions (R function prop.test).

Analysis of deep DNase-seq data vs. GR occupancy. In 3134 cells, 71% of GR occupancy sites (5,865 sites) were localized within the 2.1% of the genome defined by pre-existing (i.e., pre-hormone or baseline) DNaseI sensitive regions (DNaseI hotspots; $P<10^{-300}$). However, we noticed that an additional 13% of GR sites were localized within 2kb around these regions. Because chromatin accessibility varies as a continuous function of genome position, and observed DNaseI sensitive regions exhibit >200-fold dynamic range in total tag counts between the weakest and strongest sites, we surmised that a sequencing depth of ~25 million uniquely mapping reads had significantly under-sampled the true accessible chromatin compartment. To delineate accessible chromatin more completely, we sequenced both hormone-naïve and dexamethasone-treated DNaseI samples to a total depth of ~101 million uniquely mapping reads per condition, and recomputed sites of significantly elevated DNaseI sensitivity.and DHSs Deeper sequencing identified an additional 188,560 DHSs (276,050 vs 87,490, both at FDR 1%), and expanded annotation of significantly DNaseI sensitive regions (11.9% vs 2.1% of genome; **Supplementary Table 1-3**). Comparing GR binding patterns with this more completely delineated accessible chromatin compartment revealed that 88.3% of GR binding sites localized within pre-hormone DNaseI sensitive regions ($P<10^{-300}$).

Identification of genomic clusters of GR occupancy sites.. The genomic regions exhibiting several GR binding sites in close proximity were retrieved by the following procedure. We filtered out ChIP peaks with maximum tag density lower than 5 percentile of tag density values from all the peaks. The thresholded peaks were scanned from the beginning to the end of each chromosome and consecutive peaks within 25kb were considered to belong in a same cluster. The final set of GR binding clusters were defined to be those that have at least 3 peaks and are at least 2kb wide.

Cell type-specific and shared clusters were obtained by the following. The above algorithm was applied to the GR ChIP dataset from 3134 or AtT-20, and GR binding clusters were identified independently in each cell line. For 3134-specific clusters, we chose GR binding clusters identified in 3134 that are at least 1kb away from the nearest GR binding clusters found in AtT20. AtT20-specific clusters were obtained in a similar manner. Shared GR binding clusters were defined as those in 3134 that overlapped an AtT-20 GR binding cluster by more than 80% of their bp width.

Comparison of gene expression profiles before and after hormone induction. Comparison of the gene expression profiles of naïve and hormone-treated 3134 cells revealed 500 differentially regulated genes (235 up-regulated at least 2-fold, and 265 down-regulated; **Supplementary Table 5**). However, the average expression of genes near GR binding clusters (see methods for cluster definition) was not significantly altered (**Supplementary Figure 7a-c**).

Statistical analyses of Chromatin Context Coefficients. For each motif sequence class, we enumerated GRBEs within open or accessible chromatin (n_o) and those within closed or inaccessible chromatin (n_c), using the pre-hormone DNase I dataset (o=open and c=closed). Within each category, GR bound GRBEs were counted (n_o^{GR} and n_c^{GR}, respectively). Specifically, GRBEs within accessible chromatin were required to overlap hotspots thresholded to 0% FDR from the pre-hormone DNase I dataset. GRBEs were considered within inaccessible chromatin if they did not overlap with any unthresholded, merged hotspot from the pre-hormone DNase I dataset. GR-bound GRBEs were identified as those overlapping GR ChIP peaks. Then we defined CCC = $(n_o^{GR} / n_o) / (n_c^{GR} / n_c)$. Note that CCC is formally defined for classes with nonzero n_o. CCC is infinity (Inf) if $n_c^{GR} = 0$ and the other three counts are not.

We evaluated the statistical significance of the difference in GR bound proportions for closed and open chromatin by applying Fisher exact test to each submotif class. For this, the following contingency table was considered for each submotif class:

Chromatin	GR bound	GR unbound
closed	n_c^{GR}	$n_c - n_c^{GR}$
open	n_o^{GR}	$n_o - n_o^{GR}$

Fisher exact test was applied for each sequence class with the 2-sided null hypothesis of no chromatin effect on GR bound proportions. The procedure identified all classes with CCC = 0 (ie. no GR bound GRBEs within open chromatin) to have arisen by chance due to the small sample size of the set of GRBEs within open chromatin. The implementation of these criteria restricted our analysis to 1100 statistically well-defined GRBE classes in 3134 (668 classes in AtT-20). The other extreme CCC value of Inf (ie. no GR bound GRBEs within inaccessible chromatin), however, occurred for sub-motif classes whose contingency tables were highly significant by the Fisher exact test above. Almost every nonzero CCC value was virtually identical to the inverse of the estimated odds ratio for the corresponding contingency table according to the Fisher exact test. Therefore, we used the nonzero CCC values themselves in our comparison of the two cell lines.

Comparison of Chromatin Context Coefficients from two cell types. mCCC (mammary CCC ie. from 3134) was calculated from ChIP/DNase I-seq dataset for the mammary cell line 3134, and pCCC (pituitary CCC ie. from AtT-20) was obtained using the dataset for the pituitary cell line, AtT-20. Both versions of CCC could be formally defined for 579 classes (nonzero n_0 in both 3134 and AtT-20). Additional 7 classes for which mCCC = pCCC = 0 (i.e. statistically unreliable by Fisher exact test above) were filtered out from comparison. A random permutation test was performed to assess the significance of the observed correlation between mCCC and pCCC. Hundred random permutations of pCCC data were obtained by shuffling the class IDs of the pCCCs. For each permutation, classes with nonzero finite CCC values in both 3134 and randomized AtT-20 datasets were chosen for correlation calculation at logscale. The random occurrences of cases, where the correlation was greater than or equal to the observed, were counted for a p-value calculation (see **Supplementary Figure 8e**).