BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 1 2013, pages 22–28
doi:10.1093/bioinformatics/bts639

Sequence analysis

Advance Access publication October 26, 2012

A hidden Markov model to identify combinatorial epigenetic regulation patterns for estrogen receptor α target genes

Russell Bonneville and Victor X. Jin*

Department of Biomedical Informatics, The Ohio State University, Columbus, OH 43210, USA

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Many studies have shown that epigenetic changes, such as altered DNA methylation and histone modifications, are linked to estrogen receptor α (ERα)-positive tumors and disease prognoses. Several recent studies have applied high-throughput technologies such as ChIP-seq and MBD-seq to interrogate the altered architectures of ERα regulation in tamoxifen (Tam)-resistant breast cancer cells. However, the details of combinatorial epigenetic regulation of ERα target genes in breast cancers with acquired Tam resistance have not yet been fully examined.

Results: We developed a computational approach to identify and analyze epigenetic patterns associated with Tam resistance in the MCF7-T cell line as opposed to the Tam-sensitive MCF7 cell line, with the goal of understanding the underlying mechanisms of epigenetic regulatory influence on resistance to Tam treatment in breast cancer. In this study, we used ChIP-seq of ERα, RNA polymerase II, three histone modifications and MBD-seq data of DNA methylation in MCF7 and MCF7-T cells to train hidden Markov models (HMMs). We applied the Bayesian information criterion to determine that a 20-state HMM was best, which was reduced to a 14-state HMM with a Bayesian information criterion score of 1.21291×10^7. We further identified four classes of biologically meaningful states in this breast cancer cell model system, and a set of ERα combinatorial epigenetic regulated target genes. The correlated gene expression level and gene ontology analyses showed that different gene ontology terms were enriched with Tam-resistant versus sensitive breast cancer cells. Our study illustrates the applicability of HMM-based analysis of genome-wide high-throughput genomic data to study epigenetic influences on E2/ERα regulation in breast cancer.

Contact: victor.jin@osumc.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on June 14, 2012; revised on October 6, 2012; accepted on October 22, 2012

1 INTRODUCTION

The dysfunction of estrogen receptor α (ERα), an estrogen (E2)-inducible transcription factor, accounts for the growth of ~70% of breast tumors (EBCTCG, 1998). Treatment with tamoxifen (Tam), an estrogen antagonist and a selective ER modulator, has been successful; however, ~25% of the patients that respond to antihormonal therapy relapse within 10 years (EBCTCG, 2005). Many studies have shown that epigenetic changes, such as altered DNA methylation and histone modifications, are linked to ERα-positive tumors and prognoses of the disease (Hsu *et al.*, 2010; Kovalchuk *et al.*, 2007; Kutanzi *et al.*, 2010; Rodriguez *et al.*, 2011). Other recent studies, including ours (Gu *et al.*, 2010; Shen *et al.*, 2011), have applied high-throughput technologies such as chromatin immunoprecipitation sequencing (ChIP-seq) and methyl binding domain sequencing (MBD-seq) to interrogate the altered architectures of ERα regulation in Tam-resistant cells. However, the details of combinatorial epigenetic regulation of ERα target genes in breast cancers with acquired Tam resistance have not yet been fully examined. We sought to develop a computational approach to identify and analyze these epigenetic patterns to gain insight into the epigenetic changes associated with Tam resistance in the MCF7-T cell line compared with the Tam-sensitive MCF7 cell line, and ultimately to understand the underlying mechanisms of epigenetic regulatory influence on Tam resistance in breast cancer.

The hidden Markov model (HMM), originally developed for computerized speech recognition, is a widely used statistical method for computational modeling and analysis of various biological questions (Fischer *et al.*, 2005; Henderson *et al.*, 1997; Qin *et al.*, 2010). The basic principle of a Markov model is that it is a form of a Markov chain (a type of stochastic finite-state machine) in which each state has been extended with a set of outputs and probabilities to emit each output. In an HMM, the state sequence giving rise to the outputs is not visible, but the output sequence is known. Ernst *et al.* used HMMs to find epigenetic states in several human cell types on the basis of epigenetic data (Ernst J. *et al.*, 2011; Ernst and Kellis, 2010). Xu *et al.* (2008) used HMMs to identify differential histone modification sites between two ChIP-seq samples. Given these successful applications of HMMs, we applied the HMM to study epigenetic mechanisms underlying Tam resistance.

In this study, we used ChIP-seq of ERα, RNA polymerase II (PolII), three histone modification marks and MBD-seq data of DNA methylation in both MCF7 and MCF7-T cells to train a first-order HMM. These datasets were used because they were the only datasets available in both MCF7 and MCF7-T at the time this study was begun. We then applied the Bayesian information criterion (BIC) (Schwarz, 1978) to determine a best number of combinatorial epigenetic states. We further identified a set of ERα-regulated combinatorial epigenetic states, including promoter states and several levels of transcription, by their probability of emitting ERα/E2. Finally, gene ontology (GO) analysis was performed through the web-based tool DAVID on the genes in each list to find functional enrichments (Huang *et al.*, 2009a, b) (Fig. 1).

*To whom correspondence should be addressed.

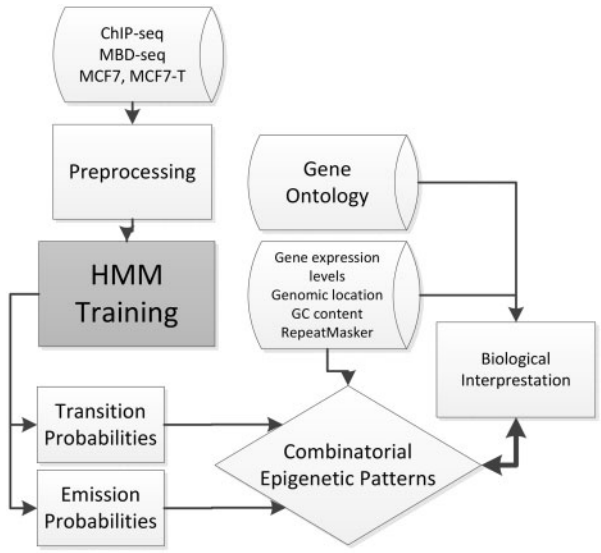

Fig. 1. An overview of the workflow

2 METHODS

2.1 Preprocessing

Each chromosome in the entire genome (hg18) was divided into 1000-bp bins with the exception of chrY (as MCF7 and MCF7-T are female-derived cell lines) and chrM (no histones). The 1000-bp size was decided because the intent of this study is to identify combinatorial regulatory patterns, not necessarily to replicate the annotation scheme developed by Ernst *et al.* Additionally, this study uses fewer datasets than the Ernst studies, potentially increasing the effect of bias or spurious reads in one dataset on the output sequence used for HMM training. A total of 1000 bp was used to mitigate possible spurious reads. For each dataset, the unique matched reads (UMRs) (Supplementary Table S1) were assigned to a bin according to their 5′ end. We determined the mark status for each bin for each of eight datasets as either '0' for non-mark or '1' for mark if the number of reads in the bin is sufficient such that $P<10^{-4}$ under the Poisson distribution, as described by Ernst *et al.* (Ernst and Kellis, 2010). An alphabet of 256 (2^8) observation symbols was constructed by enumerating each possible combination of marks including no marks. This enumeration was performed by mapping each mark to a bit in an integer value (Supplementary Table S1), and for each possible combination of marks, calculating the logical odds ratio of the bits of these marks. For example, observation 137 (0b10001001) corresponds to the presence of DNA methylation (128 = 0b10000000), H3K4me2 (8 = 0b00001000) and ERα/E2 (1 = 0b00000001), and the absence of all other marks. Each bin was assigned an observation according to the combination of marks in the bin as calculated according to the aforementioned cutoff. This was done for both cell lines to produce the HMM training sequences, for a total of 46 sequences and 6 045 312 bins.

2.2 HMM training

For each number of states from 9 to 24, five first-order HMMs were randomly initialized and learned over all observation sequences of both cell lines. The HMMs were learned over the aforementioned alphabet of 256 symbols corresponding to each possible combination of marks. Each HMM was trained for 300 iterations to ensure that they approach convergence (Supplementary Table S2) using the Baum–Welch algorithm (Baum *et al.*, 1970), with a minimum of 10^{-6} enforced for all transition, emission and start probabilities (see Supplementary Material, Minimum probability enforcement). The BIC scores of each HMM were compared, and the lowest was of the third 20-state HMM, 1.21246×10^7. This HMM was chosen as the initial HMM (Supplementary Fig. S1). As States 3, 9, 10, 11, 13 and 14 were called in <350 bins in either cell line (Supplementary Table S3) by the Viterbi decoding algorithm (run over the HMM training observation sequences) (Viterbi, 1967), these states were removed from the model, and the transition probabilities of the remaining states to them were equally divided and added to their transition probabilities to the other remaining states (Supplementary Fig. S2). This was done to simplify the HMM and to maximize the descriptive power of its states. The resulting 14-state HMM (BIC = 1.22497×10^7) was trained for a further 100 iterations to produce the final 14-state HMM, with a BIC score of 1.21291×10^7 (Fig. 2). The log-likelihood of this HMM after each training iteration was calculated to show that our Baum–Welch implementation truly improves the goodness-of-fit of the HMM with each iteration, and that 100 iterations were sufficient for this HMM to approach convergence (Supplementary Fig. S3). To show that this state reduction procedure produced a better fitting 14-state HMM than 14-state HMM training would have alone, 45 additional 14-state HMMs were randomly generated and trained for 300 iterations (Supplementary Table S4), and the significance of the BIC score of the model with state reduction was calculated versus all 50 HMMs initially trained with 14 states ($P<0.001$, Student's *t*-test).

2.3 Combinatorial epigenetic patterns

After the final 14-state HMM was learned, the Viterbi decoding algorithm was used to determine the most likely states of each bin in MCF7 and MCF7-T, using the same observation sequences used for HMM training (Supplementary Fig. S7). The distributions of the states in each bin were then compared with several other datasets by listing the bins occupied by a state for each state in both cell lines. Each list of bins for each state was first subjected to gene annotation, by calculating the distance of each bin to the nearest gene and assigning a label to the bin based on it (see Supplementary Material, Gene annotation) (Lan *et al.*, 2011). Next, the percentage of bins with each state within 2 kb of a transcription start site (TSS) region was computed to detect promoter-associated states. Similarly, the percentage of bins with each state within a gene (5′ end to 3′ end) was computed to detect transcribed-associated states. The percentage of bins with each state within RepeatMasker regions was calculated to find states corresponding to repetitive DNA (Karolchik *et al.*, 2004; Smit *et al.*, 2010). Non-coding RNA regions ($n=2975$) were determined by filtering a set of hg19 genes with a refSeq ID beginning with 'NR' ($n=2979$), then using LiftOver to convert their positions to hg18 coordinates (Hinrichs *et al.*, 2005). The average percent of GC content in bins with each state was calculated by first dividing the gc5base file into 1000-bp bins analogous to those used for HMM training to map genome-wide average GC content, and averaging the GC content in each bin with each state (Karolchik *et al.*, 2004). Bins not covered by the gc5base file were excluded from the calculation of the average.

To correlate HMM states with actual epigenetic marks, the proper mark combination was first selected, as described earlier [by mapping each mark to a bit in an integer value, see Supplementary Tables S5–S7, column Output(s)]. These mark combinations were enforced to exclude spurious bins, by filtering for those bins that actually have the marks that would be expected from that states' emission probabilities and interpretations. The bins containing the desired state were intersected with the bins containing the desired mark with BEDTools' intersectBed (Quinlan and Hall, 2010). The bins were then filtered according to their region (for example, only the bins in State 8 within a 5_TSS region were selected). The reported expression level of each gene is that of the splice variant with the highest expression level (see Supplementary Material, Microarray data).

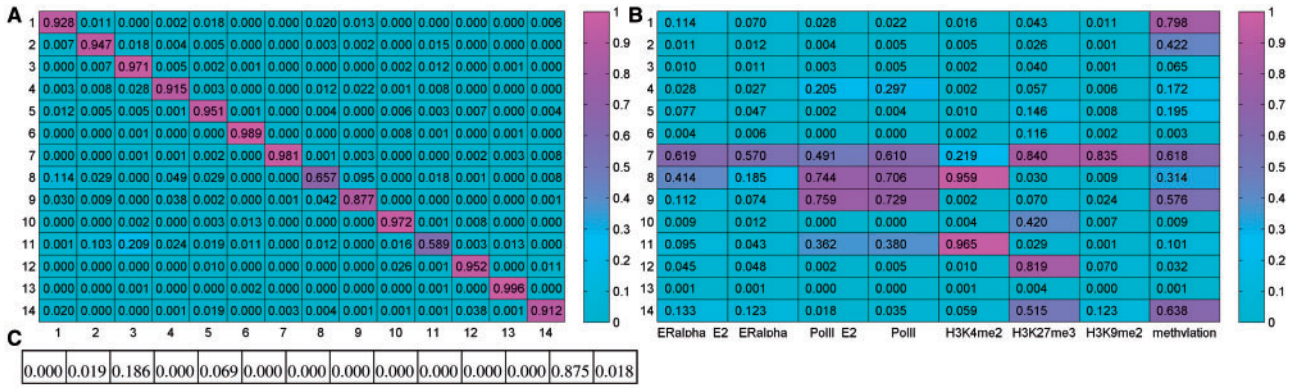

Fig. 2. (**A**) The transition probabilities of the final 14-state HMM. The transitions are from the states on the *y*-axis to the *x*-axis. (**B**) The emission probabilities of each mark independent of others of the final 14-state HMM (see Supplementary Material, Independent mark emission probabilities). (**C**) The start probabilities of the final 14-state HMM

2.4 Gene expression and GO analysis

The average microarray expression in bins with each state was computed similarly to the average GC content, using the expression of the highest-expressed splice variant of each gene (Fan *et al*., 2006). GO analysis of the genes in each list in each cell line was performed using DAVID (Huang *et al*., 2009a, b). The annotations from SP_PIR_KEYWORDS, GOTERM_BP_3 (biological process) and GOTERM_MF_3 (molecular function) were used.

3 RESULTS

3.1 Determination of combinatorial epigenetic patterns

We applied an HMM to *de novo* identification of combinatorial epigenetic patterns for ERα regulation in both MCF7 and MCF7-T cells. In addition to the 'omics data publicly available, ChIP-seq of H3K27me3, H3K4me2, PolII and ERα, the latter two performed separately in both estrogen-treated (ERα/E2 and PolII/E2) and untreated cells (ERα/control and PolII/control), and MBD-seq for DNA methylation (Gu *et al*., 2010; Shen *et al*., 2011; Welboren *et al*., 2009), we have performed ChIP-seq on histone H3 lysine 9 dimethylation (H3K9me2) in both MCF7 and MCF7-T (see Supplementary Material, ChIP-seq for H3K9me2). This resulted in a total of ~108 million UMRs in MCF7 and ~111 million UMRs in MCF7-T for all eight datasets (Supplementary Table S1).

A 14-state HMM (BIC = 1.21291×10^7) was trained over all chromosomes except chrY in both cell lines, with each combination of epigenetic marks as outputs (Fig. 2). The optimal state sequence was then determined for both cell lines with this HMM (Supplementary Fig. S8). Particularly by its emission probabilities (Fig. 2B), our model revealed spatial relationships between epigenetic marks with ERα/E2 binding, and we were able to infer specific types of functional elements associated with a given combinatorial epigenetic pattern, such as promoter, enhancer, transcribed, heterochromatin and mapping bias/copy number variation. The comparison between the model's output probabilities and the actual output frequencies for each output under each state in each bin yielded $R^2 = 0.9898$ in MCF7 and $R^2 = 0.9889$ in MCF7-T (Pearson correlation, Supplementary Figs S5B and S6).

3.2 Interpretation of combinatorial epigenetic patterns

Next, we integrated other genomic information with our HMM training results to propose potential biological interpretations of each of the detected combinatorial epigenetic patterns (Fig. 3) (see Supplementary Material, Additional data sources). We were able to identify four classes of biologically meaningful states in this breast cancer cell model system.

3.2.1 Promoter states States 8 and 11 were determined to correspond to promoter regions, primarily because 58.06 and 53.76%, respectively, of the bins with States 8 and 11 were within ±2 kb of a TSS region. This is further supported by their high emission probabilities for PolII and H3K4me2. Additionally, both States 8 and 11 had relatively high GC content (53.80 and 50.50%, respectively), as expected of the majority of promoters (Saxonov *et al*., 2006). State 8 was classified as active promoter because of the high average microarray expression value of genes associated with it (9.32). State 11 was similarly classified as weak promoter because of its lower (compared with State 8) but still relatively high average expression value (7.99). In addition, State 11 has a probability of 0.209 to transition to State 3 (low transcribed and hypomethylated), and State 8 has a probability of 0.095 to transition to State 9 (active transcribed) and a probability of 0.114 to transition to State 1 (low transcribed and hypermethylated). These promoter to transcribed transition probabilities further support the assignment of States 8 and 11 as promoters, especially given that these probabilities would be lowered by the presence of genes on the − strand (where the promoter region is encountered last when reading the + strand from the 5′ to 3′ direction). It is also possible that State 8 is at least partially influenced by ERα, as it has a relatively high emission probability for ERα/E2 (0.414); however, the average microarray expression values differ little with E2 treatment, and they differ by an average of 0.311 between MCF7 and MCF7-T.

3.2.2 Transcribed states States 1, 2, 3, 4 and 9 were determined to correspond to transcribed regions, primarily because the percentages of their bins within transcribed regions were the highest among the states with the exception of the promoter States 8 and 11 (57.30, 52.73, 47.96, 74.79 and 67.10% for States 1, 2, 3, 4 and 9, respectively). In addition, high proportions

A

State	Interpretation
1	low transcribed, hypermethylated
2	low transcribed, medium methylated
3	low transcribed, hypomethylated
4	medium transcribed
5	weakly repressed
6	heterochromatin
7	mapping bias/CNV
8	active promoter
9	highly transcribed
10	repetitive
11	weak promoter
12	polycomb repressed
13	low signal/ unmappable
14	double repressed

B

State:	1	2	3	4	5	6	7	8	9	10	11	12	13	14
5_TSS (MCF7)	2879	1680	2589	608	2737	1615	130	11442	758	2206	7597	1061	1030	952
3_Core	4251	2558	5083	4190	3045	1906	134	893	7609	2393	155	978	1119	434
5_Proximal	17917	14915	28630	7560	22191	8825	527	3665	6417	12952	1952	6059	4674	2268
3_Proximal	12767	8943	18493	9792	11797	7087	422	1731	12345	9099	604	3360	3823	1177
Intragenic	82751	67924	228145	91495	90425	117561	4933	7876	37663	120469	4690	42577	20536	10050
5_Distal	25404	28540	98088	7638	72976	62760	1639	1893	2807	76838	2182	25070	25201	5232
3_Distal	16472	14086	57389	4242	41475	47402	1251	1071	1812	53277	1145	15744	20147	3816
Gene_Desert	8208	8263	100585	2634	52126	440290	2985	414	741	162425	1174	38276	244854	4965
5_TSS (MCF7-T)	997	3146	3703	350	723	1991	178	5808	415	2447	14635	772	1352	767
3_Core	1732	5703	8728	3522	884	2471	157	603	4579	2570	589	789	1857	564
5_Proximal	8036	23845	43886	4525	6059	11143	713	2170	3593	14271	5200	4485	7278	3348
3_Proximal	5652	17410	28448	7223	3417	9174	527	1081	6673	9275	1729	2873	6240	1718
Intragenic	41624	128830	280263	36301	32723	133460	5644	4283	18925	121509	9804	38458	57671	17600
5_Distal	12770	45854	115361	3461	24803	76496	2313	1096	1379	80689	4894	20923	36674	9555
3_Distal	7857	26990	62416	1974	13205	55780	1679	601	859	54536	2614	14689	29107	7022
Gene_Desert	3584	16831	69669	1331	25476	438780	4609	328	384	171823	2351	48196	275445	9133
% +-2kb TSS (MCF7)	4.2760	3.1700	1.3952	1.7197	2.2185	0.4960	2.1296	59.5998	3.5380	1.0620	55.7003	1.6856	0.6631	6.2366
% +-2kb TSS (MCF7-T)	3.3385	3.1045	1.8344	2.2237	1.6162	0.5775	2.2187	56.5247	3.8199	1.1450	51.8127	1.2547	0.7071	3.1203
% +-2kb TSS (average)	3.8073	3.1373	1.6148	1.9717	1.9174	0.5368	2.1742	58.0623	3.6790	1.1035	53.7565	1.4702	0.6851	4.6784
% RepeatMasker (MCF7)	19.3409	27.8560	32.7650	25.2975	29.8300	35.1245	29.0991	9.0806	18.3886	29.8140	10.7595	25.0772	14.8281	19.5092
% RepeatMasker (MCF7-T)	20.7958	24.0993	31.5780	24.5676	30.7009	34.5296	29.0961	10.1753	17.6515	29.4198	11.5793	26.1798	19.1880	22.6447
% RepeatMasker (average)	20.0684	25.9777	32.1715	24.9326	30.2654	34.8271	29.0976	9.6279	18.0201	29.6169	11.1694	25.6285	17.0080	21.0770
% GC content (MCF7)	51.6954	46.3535	39.9826	39.8676	43.8178	36.8236	41.6485	54.0898	45.5364	38.8173	50.2088	40.3071	39.7186	47.5090
% GC content (MCF7-T)	50.3390	48.1407	41.2974	40.8270	42.1260	37.3148	41.5063	53.5086	46.2270	38.9243	50.7854	39.2698	39.2538	45.1812
% GC content (average)	51.0172	47.2471	40.6400	40.3473	42.9719	37.0692	41.5774	53.7992	45.8817	38.8708	50.4971	39.7884	39.4862	46.3451
% transcript (MCF7)	56.4404	51.3107	45.8186	78.2286	33.4594	18.7097	46.8347	57.1123	68.0522	29.8822	49.2179	34.6652	7.8255	38.7935
% transcript (MCF7-T)	58.1664	54.1423	50.0932	71.3531	33.1746	19.9936	40.9102	55.5291	66.1423	28.6857	46.9964	32.1043	15.6165	38.9502
% transcript (average)	57.3034	52.7265	47.9559	74.7909	33.3170	19.3516	43.8725	56.3207	67.0973	29.2840	48.1071	33.3847	11.7210	38.8719
avg. microarray expression (control, MCF7)	7.0275	6.9002	7.1251	8.4654	6.1105	4.6864	7.4004	9.1470	9.2658	4.8607	7.9069	5.1013	6.1210	5.7784
avg. microarray expression (E2, MCF7)	7.1402	6.8933	7.0830	8.4833	6.1668	4.7319	7.5596	9.1842	9.3378	4.9125	7.8686	5.1573	6.1777	5.8388
avg. microarray expression (control, MCF7-T)	7.5524	7.0687	7.3500	8.8595	6.6317	5.0657	7.7169	9.4815	9.4850	5.0998	8.1137	5.3198	6.7376	6.6340
avg. microarray expression (E2, MCF7-T)	7.5570	7.0262	7.2894	8.8142	6.5260	4.9780	8.1537	9.4710	9.4760	5.0174	8.0800	5.2630	6.6878	6.5438
avg. microarray expresion range	0.5295	0.1754	0.2670	0.3941	0.5212	0.3793	0.7533	0.3344	0.2192	0.2391	0.2451	0.2185	0.6166	0.8556
avg. microarray expression (total)	7.3192	6.9721	7.2119	8.6556	6.3587	4.8655	7.7076	9.3209	9.3912	4.9726	7.9923	5.2104	6.4310	6.1987
# bins (MCF7)	170649	146909	539002	128159	296772	687446	12021	28985	70152	439659	19499	133125	321384	28894
% of bins (MCF7)	5.6457	4.8603	17.8321	4.2399	9.8183	22.7431	0.3977	0.9589	2.3209	14.5455	0.6451	4.4042	10.6325	0.9559
# bins (MCF7-T)	82252	268609	612474	58687	107290	729295	15820	15970	36807	457120	41816	131185	415624	49707
% of bins (MCF7-T)	2.7212	8.8865	20.2628	1.9416	3.5495	24.1276	0.5234	0.5283	1.2177	15.1231	1.3834	4.3401	13.7503	1.6445
most common annotation (MCF7)	Intragenic	Intragenic	Intragenic	Intragenic	Intragenic	Gene_Desert	Intragenic	5_TSS	Intragenic	Gene_Desert	5_TSS	Intragenic	Gene_Desert	Intragenic
most common annotation (MCF7-T)	Intragenic	Intragenic	Intragenic	Intragenic	Intragenic	Gene_Desert	Intragenic	5_TSS	Intragenic	Gene_Desert	5_TSS	Gene_Desert	Gene_Desert	Intragenic

Fig. 3. The biological interpretations (**A**) of the HMM states and the correlations performed to identify them (**B**) (see Supplementary Material, Gene annotation). For all rows except gene annotation (the first 16 rows), cells are shaded according to relative value within the row (from light to dark). For gene annotation, cells are shaded according to relative value within each state

of these bins were in intragenic regions (49.18, 47.35, 44.15, 68.40 and 52.91%, respectively). State 9 was classified as highly transcribed because of its high average expression value (9.39) and its high probabilities for emitting PolII/E2 and PolII/control (0.759 and 0.729, respectively). State 3 was classified as low transcribed and hypomethylated because of its lower average expression value (7.21), very low emission probability of DNA methylation (0.065) and very low probabilities for emitting PolII/E2 and PolII/control (0.003 and 0.005, respectively). State 2 was classified as low transcribed and medium methylated because of its lower average expression value (6.97), intermediate emission probability of DNA methylation (0.422) and very low emission probabilities of PolII with and without E2 treatment (0.004 and 0.005, respectively). State 1 was classified as low transcribed and

hypermethylated because of its lower average expression value (7.31), very high emission probability of DNA methylation (0.798) and very low emission probabilities of PolII/E2 and PolII/control (0.028 and 0.022, respectively). State 4 was classified as medium transcribed because of its intermediate average expression value (8.66) and emission probabilities of PolII/E2 and PolII/control (0.205 and 0.297, respectively). The average expression level of State 1 increased somewhat with E2 treatment, and State 1 has a 0.114 probability to emit ERα/E2 and a 0.070 probability to emit ERα/control. E2 may at least partially activate some of these regions of the genome.

3.2.3 Repressed states States 5 and 14 appear to correspond to repressed intragenic regions, owing to their medium–high probabilities of being found in intragenic regions (30.48 and 35.18%, respectively) and their low average expression values (6.36 and 6.20, respectively). State 14 has higher probabilities for H3K27me3 (0.515) and methylation (0.638). State 5 has lower probabilities for H3K27me3 (0.146) and methylation (0.195). These probabilities support the assignment of States 5 and 14 as repressed. The probabilities of H3K27me3 primarily distinguish these repressed states from the low transcribed States 1 and 2. However, State 14 has a probability of 0.1038 to emit no mark, 0.1174 for H3K27me3 alone, 0.2070 for methylation alone and 0.2092 for H3K27me3 and methylation together (Supplementary Fig. S4, note the bands at State 5 and 14, output 160). The corresponding probabilities for State 5 are 0.5975, 0.1072, 0.1431 and 0.0234. In sum, whenever State 14 emits H3K27me3 or methylation, it emits both 39.21% of the time, versus 8.55% for State 5. Therefore, State 14 was classified as double repressed and State 5 as merely weakly repressed. State 12 was classified as polycomb repressed because of its very low average expression value (5.21) and its high probability for H3K27me3 (0.819), which is commonly associated with polycomb repression (Zhou *et al*., 2011).

3.2.4 Other states State 13 was classified as low signal/unmappable because of the high percentage of its bins in gene desert regions (70.60%) and its very low emission probabilities for all marks. State 7 was classified as mapping bias/copy number variation (CNV) because of its high emission probabilities for all marks including those not expected to co-occur (such as PolII and H3K27me3), implying that this state corresponds to ChIP artifacts rather than actual mark combinations. State 10 was classified as repetitive because of the high percentage of its bins in RepeatMasker regions (29.62%), its low emission probabilities for all marks except H3K27me3 (0.420) and the high proportion of its bins in gene desert regions (37.27%). State 6 was classified as heterochromatin because of the high percentage of its bins in gene desert regions (62.05%), its low emission probability for all marks except H3K27me3 (0.116), its very low average expression value (4.87), its prevalence in the genome (23.44% of all bins) and the high percentage of its bins in RepeatMasker regions (34.83%).

3.3 Identification of ERα combinatorial epigenetic regulated genes

Our goal is to uncover combinatorial epigenetic influence on ERα regulated genes in MCF7 versus MCF7-T cells. A total of 1 219 650 bins (40.35% of all bins) were in a different state between MCF7 and MCF7-T (Supplementary Fig. S12). Because of the number of different bins, we could not directly compare the state sequences between the cell lines. We believe that this is due in part to noise in the original ChIP data. Thus, we selected States 1, 7, 8, 9 and 14 for further analysis, as they have a probability >0.1 of emitting ERα with E2 treatment in at least one cell line (we select these E2-associated states, as Tam is an E2 antagonist). State 1 was correlated with DNA methylation (see Section 2, Supplementary Table S5) and filtered for intragenic bins (as it was assigned as low transcribed and hypermethylated) (v Fig. S13). All bins in State 7 were used for the analysis. State 8 was correlated with DNA methylation, H3K4me2, ERα/E2, ERα/control, PolII/E2 and PolII/control, and filtered for TSS bins (as it was assigned as active promoter). State 9 was correlated with DNA methylation, ERα/E2, PolII/E2 and PolII/control, and filtered for intragenic bins (as it was assigned as highly transcribed). State 14 was correlated with DNA methylation, H3K27me3, H3K9me2, ERα/E2 and ERα/control.

These results support the assignment of State 7 as mapping bias/CNV. A total of 132 genes were shared by both the MCF7 and MCF7-T lists. This state was distributed consistently (Pearson $R^2 = 0.898$) between the cell lines as expected (as the reference sequence used was the same for all ChIP alignments) (Supplementary Table S8), and the high degree of overlap (91.0% of the MCF7 genes and 67.7% of the MCF7-T genes) observed here is consistent with this conclusion.

A total of 2154 genes were found to have methylated intragenic bins in State 1 with ERα/E2 binding and no ERα/control binding in MCF7, while 339 genes were similarly found in MCF7-T (Fig. 4). Additionally, the average expression level with E2 treatment of the MCF7-T genes was higher (8.0940 in MCF7-T versus 7.2170 in MCF7, $P < 6 \times 10^{-9}$, Benjamini $< 7 \times 10^{-8}$ and two-tailed Welch's *t*-test) (Benjamini and Hochberg, 1995). A total of 213 of these genes were found (and included) in both lists. The average expression level of the genes in the MCF7 list in E2-treated MCF7-T is 7.1893 (SD = 2.4311), and the average expression level of the genes in the MCF7-T list in E2-treated MCF7 is 7.9295 (SD = 2.4545) ($P < 5 \times 10^{-7}$ and Benjamini $< 5 \times 10^{-6}$ of significant expression difference between them). ERα/E2 binding without ERα/control binding is clearly more common in MCF7 than MCF7-T, as evident with the much higher number of genes found in MCF7, as well as the

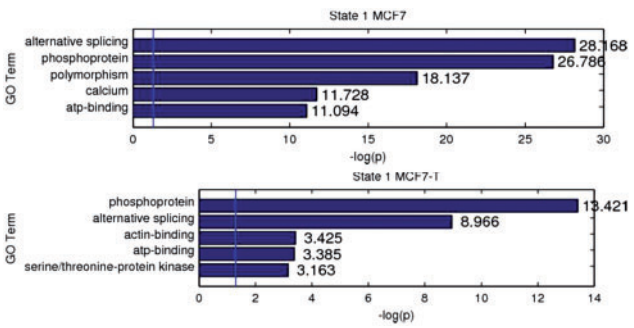

Fig. 4. The top five GO terms in the State 1 gene lists, in MCF7 and MCF7-T. The vertical blue line is at $P = 0.05$. See Supplementary Figure S10 for Benjamini corrected *P*-values

actual emission frequencies of ERα/E2 in MCF7-T (Supplementary Fig. S5B). MCF7-T is apparently able to maintain similar expression levels of these genes even without ERα/E2 binding. The genes in the MCF7-T list appear to simply be more active than the other genes discussed here (Supplementary Fig. S11). This may be related to the ERα/E2 binding observed in both cell lines. Also, the MCF7 gene list was strongly enriched for the GO term 'regulation of cell communication' ($P<10^{-10}$ and Benjamini $<10^{-7}$).

The number of genes with bins in State 8 and 9 dropped significantly in MCF7-T compared with MCF7, as was observed with State 1. The expression levels of the genes in the State 8 lists changed little between MCF7 and MCF7-T, or with and without E2 treatment. The State 9 MCF7-T gene lists had higher expression levels than MCF7 ($P<0.02$ and Benjamini <0.04 in all cases). All four of the States 8 and 9 MCF7-T gene lists were enriched with the GO term 'purine nucleotide binding', but the *P*-values for many GO terms were relatively high (Supplementary Fig. S9), in some cases failing to reach statistical significance with Benjamini correction (Supplementary Fig. S10). This is most likely because of the relatively small size of these gene lists, as well as the fact that they are not necessarily functionally related to each other, as they were selected on the basis of the combinatory epigenetic patterns identified by the HMM. The relatively large size of the MCF7 gene lists also diminished the significance of the GO terms found in them, as well as their lack of guaranteed functional relation.

We chose to look at genes with bins in State 14 containing H3K27me3, DNA methylation, H3K9me2 and ERα binding (both with and without E2). Most notably, the average expression levels of the MCF7-T genes were significantly elevated with and without E2 treatment ($P<0.02$ and Benjamini <0.04 in both cases) (Supplementary Tables S6–S7). Although few genes were found in either cell line (and only 11 genes are shared between them), these genes are clearly little affected by E2 treatment in both cell lines, independent of the Tam resistance of MCF7-T ($P>0.93$ in both cases, as would be expected given this). The significance of the GO terms was low, probably because of the small size of the gene lists (Supplementary Fig. S9 and S10).

4 DISCUSSION

Using a HMM, we have identified states reflecting combinatorial epigenetic patterns for ERα regulated target genes in MCF7 and MCF7-T. We have interpreted each of the states to determine their biological significance. With our results, we have determined lists of genes affected by ERα binding with E2 in both MCF7 and MCF7-T, considering a variety of other epigenetic effects. Although these results do not conclusively elucidate the epigenetic mechanisms underlying Tam resistance, they nonetheless illustrate the applicability of HMM-based analysis of genome-wide high-throughput genomic data to study epigenetic influence on E2/ERα regulation in breast cancer.

Although the application of GO analysis did not yield statistically significant terms for every gene list, the results of this study nonetheless suggest several possible future directions. For instance, this study was confined to HMM analysis and gene listing. Perhaps, application of gene network analysis could reveal key genes responsible for the regulation of the genes in the lists formulated by this study. The lists of genes may provide a starting point for further analyses, such as microarray assays of tumor samples focusing on these genes.

The methods used here should be applicable in cases other than Tam resistance in MCF7-T, using other genome-wide epigenetic datasets than those used in this study. The HMM training and identification of combinatorial regulated genes is extensible to more than two cell lines [O(n) in HMM training time and space, where n = number of cell lines, but this is parallelizable with a sufficient number of compute cores]. For example, this can be used to interrogate combinatorial epigenetic regulation patterns in multiple types of cancer, or multiple patient samples with the same type of cancer.

In the HMM implementation used in this study, each possible combination of marks is enumerated as a possible output of the HMM. This permits more straightforward analysis of combinatorial marks (for example, whether marks X and Y are found together more often or not), which is critical to this study. Ernst *et al.* used a multivariate HMM approach, in which the output probabilities of each mark are considered independently of each other (Ernst and Kellis, 2012). Although multiple marks are allowed in each bin with this model, it does not directly model the frequency with which marks co-occur (which would be calculated after Viterbi decoding). For example, consider a toy HMM with two marks and one state, learned over a sequence of two bins. If Bin 1 has Mark 1 and Bin 2 has Mark 2, the state in both an Ernst HMM and one of our HMMs would have an emission probability of 50% for each mark. However, if Bin 1 has both Marks 1 and 2, and Bin 2 has no marks, the Ernst HMM would be the same, and our HMM would have an emission probability of 100% for both marks and 0% for each mark alone.

Future studies could greatly benefit from the incorporation of additional data in the HMM. Although many MCF7 datasets are available, datasets were available for only the eight epigenetic marks used in this study in both MCF7 and MCF7-T at the time this study was begun (a dataset must be available in all cell lines studied to be usable in the HMM). Histone acetylation data could broaden the context of the HMM analysis by providing histone epigenetic data separate from the methylations analyzed in this study. For example, H3K27ac data could support the H3K4me2 data and possibly allow the detection of enhancers enriched in both marks (Zhou *et al.*, 2011), as well as better differentiation between active promoter and weak promoter states. H3K9me3 data may permit differentiation between constitutive and facultative heterochromatin. H3K36me3 data could allow differentiation between exons and introns. ChIP-seq of ERα and PolII with Tam and E2 treatment would allow more direct investigation of Tam resistance.

The potential utility of further next-generation sequencing (NGS) data is most powerfully demonstrated by the scope of the work of Ernst *et al.* (Ernst and Kellis, 2010), in which the authors were able to define 51 HMM states by using datasets of 18 acetylations, 20 methylations and 3 other marks. Though it would undoubtedly be impractical with current technology to perform 41 NGS experiments each in two cell lines, it is clear that the addition of other carefully selected datasets can improve the discriminatory power of HMM-based methods such as those used by Ernst *et al.* and this study. Additionally, the usage of additional datasets can reduce the impact of biases in individual

datasets produced by differing experimental protocols, as the contribution of each dataset to the observation sequences used for HMM training is reduced. This may allow the state sequences in each cell line to be directly compared instead of comparing state-associated genes, as well as reduction of the 1000-bp bin size for higher resolution.

The addition of further NGS datasets may be complicated by the increased computational and memory requirements potentially necessitated by additional data. Additional datasets may require additional states to be included in HMMs. As the computational and memory costs of our implementation of the Baum–Welch algorithm (in particular, the gamma values) grow according to $O(nb^2)$, where n = number of bins and b = number of states, this would become prohibitive with increasing numbers of states. For example, the gamma values for 20 states, 1000-bp bins and two cell lines with long double precision (16 bytes with GCC on the x64 architecture) would require about 36 GB of memory ($3022656 \times 2 \times 16 \times 20 \times 20$ bytes). With 30 states, this increases to ~81 GB. Although not all of this must be allocated at once (as forward–backward is performed separately on each chromosome to allow it to be parallelized), this would nonetheless require a computer with a great deal of available memory. However, given the historical decrease of the cost of memory, this may become less of an issue in the future.

In addition, there are many subjective parameters in this study. For example, 1000-bp bins were used, primarily to mitigate ChIP-seq noise. Reducing the bin size can potentially increase the resolution of the state assignments; however, it would increase the sensitivity of state assignment to ChIP-seq noise. The cutoff of $P < 10^{-4}$ for the number of reads in a bin required to call that mark present was selected solely because of its successful use in previous work (Ernst and Kellis, 2010). This cutoff may potentially be overly strict or overly permissive in this study. We chose to select states with at least a 0.1 probability of emitting ERα/E2 for further analysis. Although we used this threshold to choose states with some relation to ERα binding, it is admittedly arbitrary. Perhaps a statistical method could be devised to select states for gene identification in future work.

Additional data files are available at http://motif.bmi.ohio-state.edu/ERHMM.

ACKNOWLEDGEMENTS

The authors appreciate the laboratory of Dr Tim Huang at the University of Texas Health Science Center, San Antonio, for providing the unpublished ChIP-seq data of H3K9me2 in MCF7 and MCF7-T cells. They also thank Ms Sandya Liyanarachchi for statistical advice and Ms Kim Leonard for technical proofreading. They thank other laboratory members for their discussions.

Funding: The PhRMA foundation and the Department of Biomedical Informatics, Wexner Medical Center at the Ohio State University, Columbus, OH, USA.

Conflict of Interest: none declared.

REFERENCES

Baum,L.E. *et al.* (1970) A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.*, **41**, 164–171.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.

Early Breast Cancer Trialists' Collaborative Group (EBCTCG). (1998) Tamoxifen for early breast cancer: an overview of the randomised trials. *Lancet*, **351**, 1451–1467.

Early Breast Cancer Trialists' Collaborative Group (EBCTCG). (2005) Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. *Lancet*, **365**, 1687–1717.

Ernst,J. and Kellis,M. (2010) Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.*, **28**, 817–825.

Ernst,J. *et al.* (2011) Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, **473**, 43–49.

Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–206.

Fan,M. *et al.* (2006) Diverse gene expression and DNA methylation profiles correlate with differential adaptation of breast cancer cells to the antiestrogens tamoxifen and fulvestrant. *Cancer Res.*, **66**, 11954–11966.

Fischer,B. *et al.* (2005) NovoHMM: a hidden Markov model for de novo peptide sequencing. *Anal. Chem.*, **77**, 7265–7273.

Gu,F. *et al.* (2010) Inference of hierarchical regulatory network of estrogen-dependent breast cancer through ChIP-based data. *BMC Syst. Biol.*, **4**, 170.

Henderson,J. *et al.* (1997) Finding genes in DNA with a hidden Markov model. *J. Comput. Biol.*, **4**, 127–141.

Hinrichs,A.S. *et al.* (2005) The UCSC genome browser database: update 2006. *Nucleic Acids Res.*, **34** (**Suppl. 1**), D590–D598.

Hsu,P. *et al.* (2010) Estrogen-mediated epigenetic repression of large chromosomal regions through DNA looping. *Genome Res.*, **20**, 733–744.

Huang,D.W. *et al.* (2009a) Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.*, **37**, 1–13.

Huang,D.W. *et al.* (2009b) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.*, **4**, 44–57.

Karolchik,D. *et al.* (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493–D496.

Kovalchuk,O. *et al.* (2007) Estrogen-induced rat breast carcinogenesis is characterized by alterations in DNA methylation, hsitone modifications and aberrant microRNA expression. *Cell Cycle*, **6**, 2010–2018.

Kutanzi,K. *et al.* (2010) Reversibility of pre-malignant estrogen-induced epigenetic changes. *Cell Cycle*, **9**, 3078–3084.

Lan,X. *et al.* (2011) W-ChIPeaks: a comprehensive web application tool for processing ChIP-chip and ChIP-seq data. *Bioinformatics*, **27**, 428–430.

Qin,Z.S. *et al.* (2010) HPeak: an HMM-based algorithm for defining read-enriched regions in ChIP-Seq data. *BMC Bioinformatics*, **11**, 369.

Quinlan,A.R. and Hall,I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Rodriguez,B. *et al.* (2011) Estrogen-mediated epigenetic repression of the imprinted gene cyclin dependent kinase inhibitor 1C in breast cancer cells. *Carcinogenesis*, **32**, 812–821.

Saxonov,S. *et al.* (2006) A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. *Proc. Natl Acad. Sci. USA*, **103**, 1412–1417.

Schwarz,G.E. (1978) Estimating the dimension of a model. *Ann. Stat.*, **6**, 461–464.

Shen,C. *et al.* (2011) A modulated empirical Bayes model for identifying topological and temporal estrogen receptor α regulatory networks in breast cancer. *BMC Syst. Biol.*, **5**, 67.

Smit,A. *et al.* (2010) *RepeatMasker Open-3.0*. http://www.repeatmasker.org (16 October 2011, date last accessed).

Viterbi,A. (1967) Error bounds for convolutional codes and an asymptotically optimum decoding algorithm. *IEEE Trans. Inf. Theory*, **13**, 260–269.

Welboren,W.J. *et al.* (2009) ChIP-Seq of ERα and RNA polymerase II defines genes differentially responding to ligands. *EMBO J.*, **28**, 1418–1428.

Xu,H. *et al.* (2008) An HMM approach to genome-wide identification of differential histone modification sites from ChIP-seq data. *Bioinformatics*, **24**, 2344–2349.

Zhou,V.W. *et al.* (2011) Charting histone modifications and the functional organization of mammalian genomes. *Nat. Rev. Genet.*, **12**, 7–18.