Genomics Data 2 (2014) 195–198

Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

Data in Brief

Genome-wide epigenetic profiling of breast cancer tumors treated with aromatase inhibitors

Ekaterina Nevedomskaya [a,b], Lodewyk Wessels [b,c], Wilbert Zwart [a,*]

[a] Division of Molecular Pathology, Netherlands Cancer Institute, Amsterdam, The Netherlands
[b] Division of Molecular Carcinogenesis, Netherlands Cancer Institute, Amsterdam, The Netherlands
[c] Delft Bioinformatics Lab, Delft University of Technology, Delft, The Netherlands

ARTICLE INFO

Article history:
Received 22 May 2014
Accepted 24 June 2014
Available online 8 July 2014

Keywords:
Breast cancer
ChIP-seq
Epigenetic modifications
Aromatase inhibitor treatment outcome
Estrogen Receptor

ABSTRACT

Aromatase inhibitors (AI) are extensively used in the treatment of estrogen receptor-positive breast cancers, however resistance to AI treatment is commonly observed. Apart from Estrogen receptor (ERα) expression, no predictive biomarkers for response to AI treatment are clinically applied. Yet, since other therapeutic options exist in the clinic, such as tamoxifen, there is an urgent medical need for the development of treatment-selective biomarkers, enabling personalized endocrine treatment selection in breast cancer. In the described dataset, ERα chromatin binding and histone marks H3K4me3 and H3K27me3 were assessed in a genome-wide manner by Chromatin Immunoprecipitation (ChIP) combined with massive parallel sequencing (ChIP-seq). These datasets were used to develop a classifier to stratify breast cancer patients on outcome after AI treatment in the metastatic setting. Here we describe in detail the data and quality control metrics, as well as the clinical information associated with the study, published by Jansen et al. [1]. The data is publicly available through the GEO database with accession number GSE40867.

Specifications	
Organism/cell line/tissue	Homo sapiens
Sex	Female
Sequencer or array type	Illumina Hiseq 2000 genome analyzer
Data format	Raw: SRA study; processed: BED
Experimental factors	Poor vs. good outcome tumors
Experimental features	Genome-wide binding of Estrogen Receptor α (ERα), as well as histone marks H3K4me3 and H3K27me3, were assessed in tumors from breast cancer patients with good or poor survival outcome after aromatase inhibitors therapy.
Consent	All patients gave their written informed consent before study entry.
Sample source location	Samples were from breast cancer patients, treated at the Erasmus University Medical Center (EMC; Rotterdam, the Netherlands), the Netherlands Cancer Institute/Antoni van Leeuwenhoek hospital (Amsterdam, the Netherlands), and the Translational Cancer Research Unit (Saint Augustinus Hospital, Antwerpen, Belgium).

Direct link to deposited data

Deposited data can be found here: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE40867.

Experimental design, materials and methods

Study population and clinical data

The cohort of 84 metastatic ERα-positive breast cancer patients, who received AI therapy, was selected for evaluation. Tumor material analyzed by genomic profiling was extracted from primary surgery specimens. The patient selection criteria, definitions of follow-up, tumor staging, and response to therapy were previously described by Ramirez-Ardila et al. [2]. Briefly, fresh frozen ERα-positive breast tumor tissue specimens were collected from female patients with primary operable breast cancer and whose metastatic disease was treated with first-line aromatase inhibitors (anastrozole, letrozole, exemestane). Time to progression (TTP) was taken as the end point. Thirteen specimens were selected for chromatin immunoprecipitation (ChIP) and massive parallel sequencing (ChIP-seq) analyses, all on samples with more than 50% ER-positive tumor cells. Poor outcome patients were defined as patients with a TTP < 12 months, whereas good outcome

* Corresponding author.
E-mail address: w.zwart@nki.nl (W. Zwart).

http://dx.doi.org/10.1016/j.gdata.2014.06.023

Table 1
Patient and tumor characteristics for the selected groups.

Patients (n = 13)		
Characteristic	No of patients	
	Good outcome	Poor outcome
	5	8
Age at diagnosis (mean), years	64	60
Age at start therapy (mean), years	68	63
Treatment type		
Anastrozole	2	5
Exemestane	0	1
Exemestane	0	1
Letrozole	3	1
Grade		
1	1	0
2	3	3
3	1	4
ER status		
Negative	0	0
Positive	5	8
PR status		
Negative	0	0
Positive	5	8
HER2 status		
Negative	3	5
Positive	1	1
TTP (median), months	38	6.5

patients were defined as patients with a TTP > 24 months. Clinical characteristics of the selected groups of patients are provided in Table 1 and clinical characteristics per sample are provided in the Supplementary Table 1.

The anonymized clinical data were deposited in the Gene Expression Omnibus database (GEO; [3]) under accession number GSE40867.

Chromatin immunoprecipitations and sequencing

Chromatin immunoprecipitation (ChIP) was performed as described before [1]. To obtain input material, tumor samples were cryosectioned (30 × 30 mm sections) prior to further processing for ChIP-seq as described before [7]. For each ChIP, 10 mg of antibody and 100 mL of Protein A magnetic beads (Invitrogen) were used. Antibodies against ERα (SC-543; Santa Cruz), H3K4me3 (ab8580; Abcam), and H3K27me3 (07–449; Millipore) were used.

ChIP DNA was amplified as described [1,4]. Sequences were generated by the Illumina Hiseq 2000 genome analyzer (using 50 bp reads), and aligned to the Human Reference Genome (assembly hg19, February 2009). Non-ChIP input DNA from a randomly selected tumor was sequenced as an input control. Enriched regions of the genome were identified by comparing the ChIP samples to input using the MACS peak caller [5] version 1.3.7.1 with default parameters, except for the p-value cutoff that was set at 10^{-7}. Details on the number of reads obtained, the percentage of reads aligned, and the number of peaks called can be found in Table 2. ChIP-seq data and sample annotations were deposited in GEO under accession number GSE40867.

Quality control

Prior to analysis, visual inspection of the regions known to typically bind ERα or contain histone modifications was performed using the Integrative Genome Viewer IGV 2.1 (www.broadinstitute.org/igv/). Examples of such regions are provided in Fig. 1A. As expected, ERα

Table 2
Read count, number of peaks and quality parameters.

GEO accession	ChIP	Total reads	Mapped reads (%)	No of peaks	Fraction of reads in peaks, %	NSC	RSC
GSM1003708	ERα	23,760,885	21,964,709 (92.4)	524	0.15	1.02	0.48
GSM1003709	H3K4me3	22,772,852	20,520,046 (90.1)	16,384	26.35	1.64	1.72
GSM1003710	H3K27me3	26,990,559	26,122,735 (96.8)	14,890	3.56	1.01	0.46
GSM1003711	ERα	23,802,294	22,002,226 (92.4)	2255	0.6	1.02	0.47
GSM1003712	H3K4me3	22,591,289	20,813,064 (92.1)	16,857	31.03	1.61	1.49
GSM1003713	H3K27me3	22,096,326	21,343,362 (96.6)	10,078	2.91	1.01	0.32
GSM1003714	ERα	20,789,758	17,832,808 (85.8)	15,381	4.38	1.09	0.89
GSM1003715	H3K4me3	23,075,271	20,411,990 (88.5)	25,111	11.97	1.07	0.53
GSM1003716	H3K27me3	19,103,286	17,130,759 (89.7)	4008	1.48	1.02	0.32
GSM1003717	ERα	22,555,195	21,115,239 (93.6)	2726	0.84	1.04	0.76
GSM1003718	H3K4me3	19,872,399	18,226,845 (91.7)	16,320	9.05	1.05	0.44
GSM1003719	H3K27me3	23,961,464	22,493,285 (93.9)	3085	0.64	1.02	0.39
GSM1003720	ERα	16,604,876	15,605,068 (94.0)	13,575	3.61	1.09	1.03
GSM1003721	H3K4me3	10,238,004	9,467,187 (92.5)	19,012	6.69	1.09	0.33
GSM1003722	H3K27me3	22,530,535	21,625,249 (96.0)	33,661	9.13	1.04	0.83
GSM1003723	ERα	19,902,396	18,778,288 (94.4)	6387	1.46	1.04	0.7
GSM1003724	H3K4me3	20,235,985	18,151,245 (89.7)	18,351	41.23	1.83	1.56
GSM1003725	H3K27me3	24,169,596	23,067,266 (95.4)	30,588	7.44	1.02	0.42
GSM1003726	ERα	16,011,312	13,905,708 (86.8)	2287	0.58	1.02	0.41
GSM1003727	H3K27me3	16,423,400	15,482,959 (94.2)	28,514	7.34	1.03	0.57
GSM1003728	ERα	21,552,073	17,908,925 (83.1)	709	0.72	1.02	0.31
GSM1003729	H3K4me3	27,693,755	25,171,058 (90.9)	27,023	15.47	1.16	1.03
GSM1003730	H3K27me3	27,372,177	24,765,816 (90.5)	11,395	1.28	1.02	0.67
GSM1003731	ERα	15,620,215	14,134,239 (90.5)	5170	2.48	1.05	0.72
GSM1003732	H3K4me3	20,741,336	18,816,604 (90.7)	26,821	14.94	1.1	0.57
GSM1003733	H3K27me3	21,310,477	20,553,892 (96.4)	27,122	3.28	1.02	0.49
GSM1003734	ERα	18,169,785	16,090,891 (88.6)	19,716	5.43	1.14	1.19
GSM1003735	H3K4me3	26,621,106	24,586,405 (92.4)	23,785	1.58	1.11	0.87
GSM1003736	H3K27me3	26,069,531	25,135,569 (96.4)	59,910	22.03	1.06	1.23
GSM1003737	ERα	20,867,111	18,925,868 (90.7)	1110	0.29	1.01	0.37
GSM1003738	H3K4me3	20,012,887	17,988,530 (89.9)	16,427	22.24	1.38	1.41
GSM1003739	H3K27me3	23,750,330	22,949,910 (96.6)	4256	0.79	1.01	0.33
GSM1003740	ERα	13,499,179	12,530,097 (92.8)	924	1.76	1.02	0.33
GSM1003741	H3K4me3	26,027,543	24,076,775 (92.5)	26,396	11.24	1.13	0.96
GSM1003742	H3K27me3	31,996,441	30,602,420 (95.6)	7067	1.13	1.03	0.86
GSM1003743	Input	27,097,497	25,588,905 (94.4)				

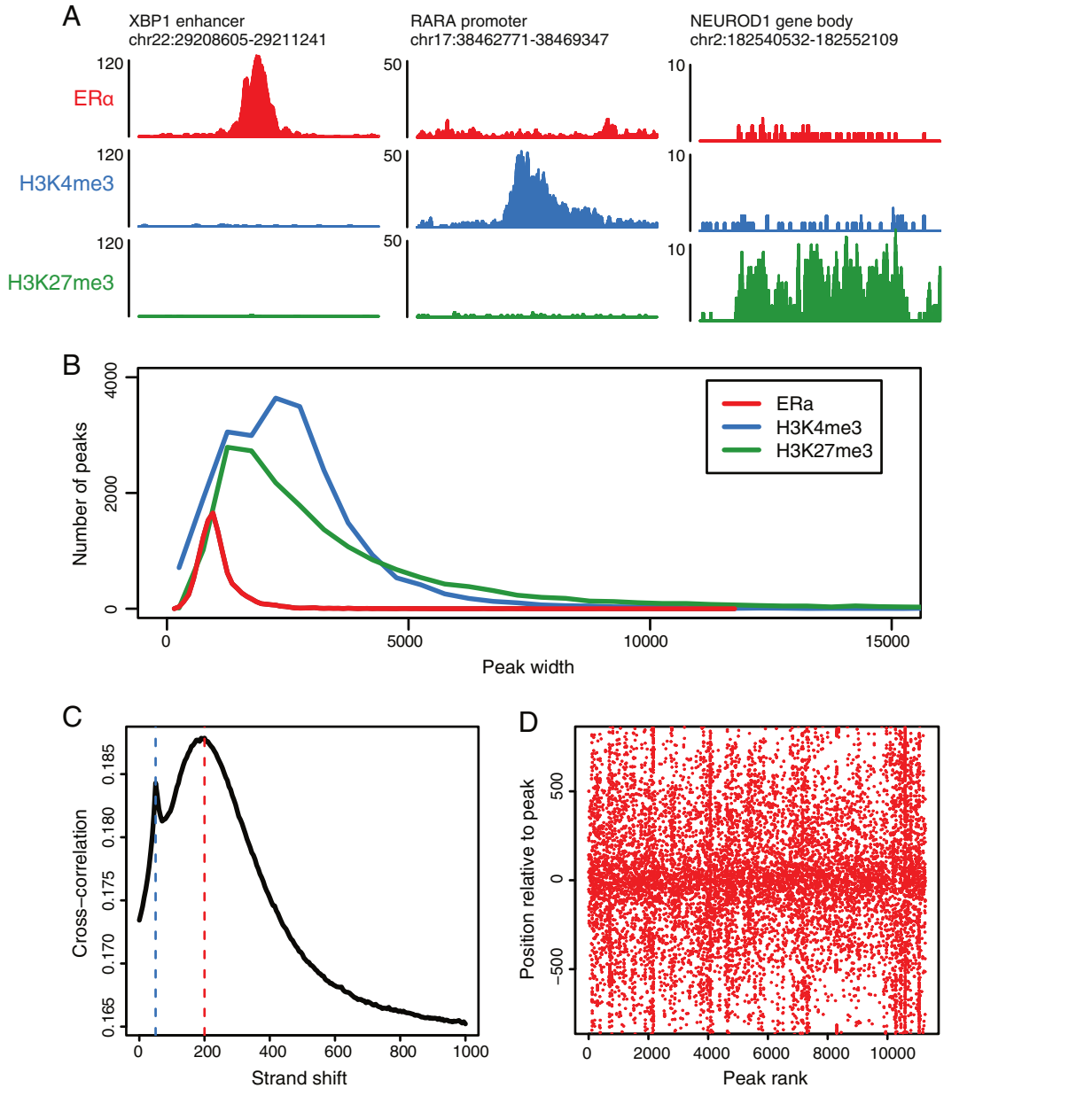

Fig. 1. Quality control and data metrics of ChIP-seq data. (A) Example genomic regions with distinct and unique signal of ERα (red), H3K4me3 (blue), and H3K27me3 (green) binding events. Genomic coordinates are indicated. Tag count is shown for each position. (B) Distribution of peak widths in different ChIP-seq datasets. (C) Example of a cross-correlation plot. Blue dashed line indicates the 'phantom' peak corresponding to the read length, red dashed line marks the peak of the fragment length. (D) Distribution of ERα motifs relative to the peak position of ERα binding events.

peaks were found at the enhancers of known estrogen-responsive genes (e.g. XBP1 (Fig. 1A), RARA, GREB1), H3K4me3 signal was observed at promoters of estrogen-responsive genes and H3K27me3 marked genes not expressed in breast tissue, such as NEUROD1. (Fig. 1A). The peaks of H3K4me3 histone modification are often wider than the peaks of ERα binding [6], while the transcription repressive histone mark H3K27me3 can cover large areas, including full gene bodies [7], which also results in the identification of broad peaks for this histone mark. Peak widths for all three datasets are illustrated by the density distributions as depicted in Fig. 1B.

There is no current consensus on the quality control metrics for ChIP- and enrichment-based technologies, such as ChIP-seq, GRO-seq and others. Commonly, the number of reads and peaks detected is reported. The total number of reads, number of aligned reads and number of peaks for each ChIP-seq sample are shown in Table 2. A few quality control procedures have been suggested in the literature [8,9], however their use is not established practice and some of them may not be applicable to a large variety of ChIP-seq data.

Here we employed quality control measures suggested by the ENCODE consortium for assessing the quality of the data [8]. It is, however, important to mention that ENCODE guidelines are used in the analysis of the data from cell line experiments. Data from tumor samples, used in the current study, are more difficult to process due to intrinsic intra-tumor heterogeneity and biological variation. Therefore, we cannot expect our tumor sample-based ChIP-seq data to fully meet the criteria used for the cell line data. The minimal fraction of reads in peaks as prescribed by ENCODE (1%), which is an indicator of ChIP efficiency, was met in almost 80% of the samples (Table 2). Cross-correlations of positive and negative strands were calculated using publicly available scripts (http://code.google.com/p/phantompeakqualtools) [10,11]. An example of a cross-correlation plot can be seen in Fig. 1C. Dominant fragment and read lengths were calculated from the cross-

correlations, and the related measures, namely Normalized Strand Coefficient (NSC) and Relative Strand Correlation (RSC), were assessed. As can be seen from Table 2, not all the samples meet the ENCODE criteria of NSC > 1.05 and RSC > 0.8. The best results for these parameters are achieved in the H3K4me3 data with over 90% meeting the NSC criterion and over 60% meeting the RSC criterion. Overall, the quality metrics for ERα ChIP-seq have lower values than those for the histone marks. However, it is not surprising for a number of reasons. First, immunoprecipitation of chromatin with histone marks is more efficient as histones are the intrinsic part of the chromatin, whereas ERα is a transcription factor not integrated in the structure of chromatin. Second, being a hormone-dependent transcription factor, ERα chromatin interactions are dependent on the physiological levels of E2, which may be at non-saturated levels within the tumor and could vary from patient to patient. Third, as shown before, high quality ChIP-seq datasets with limited number of genuine binding sites may produce low NSC and RSC values [8].

We further validated that the peaks detected in ERα ChIP-seq data are genuine signal and correspond to the binding sites of ERα. Called peaks that were found in at least two tumor samples were considered for analysis, resulting in 11,262 peaks for ERα dataset. This high number of consensus peaks illustrates the quality of the data available for the analysis. We subsequently defined the locations of ERα motifs in these peaks by using the ScreenMotif tool from the Cistrome (cistrome.org). As seen from the Fig. 1D, the motifs are clearly concentrated around the center of identified peaks. This illustrates that despite the NSC and RSC metrics having marginal values, the ERα peaks detected present a genuine signal. R scripts for analysis are available upon request.

Discussion

Here we describe a unique dataset, in which we profiled the chromatin binding landscapes of ERα, H3K4me3 and H3K27me3 in primary human ERα-positive luminal breast tumor specimens. Patients were treated in the metastatic setting with AIs, and survival data are available and provided in the public data repositories. With this, our datasets consist of two parts: clinical and ChIP-seq data. Clinical data includes outcome upon treatment with aromatase inhibitors and other important clinic-pathological characteristics. ChIP-seq data comprises genome-wide profiling of estrogen receptor (ERα) binding to chromatin, promoter-specific histone modification H3K4me3 and transcription repressive histone mark H3K27me3. This dataset has been recently used in a publication for finding epigenetic signatures related to the outcome upon aromatase inhibitors treatment for metastatic breast cancer [1].

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gdata.2014.06.023.

Acknowledgments

The authors would like to thank all contributors to the original paper [4]. Wilbert Zwart is supported by an Alpe d'HuZes foundation/KWF Dutch Cancer Society Bas Mulder Award (NKI2014-6711) and a Veni grant from The Netherlands Organisation for Scientific Research (NWO) (916.12.009). This work was supported by A Sisters Hope.

References

[1] M.P.H.M. Jansen, et al., Hallmarks of aromatase inhibitor drug resistance revealed by epigenetic profiling in breast cancer. Cancer Res. 73 (2013) 6632–6641.
[2] D.E. Ramirez-Ardila, et al., Hotspot mutations in PIK3CA associate with first-line treatment outcome for aromatase inhibitors but not for tamoxifen. Breast Cancer Res. Treat. 139 (2013) 39–49.
[3] T. Barrett, et al., NCBI GEO: mining millions of expression profiles—database and tools. Nucleic Acids Res. 33 (2005) D562–D566.
[4] D. Schmidt, et al., ChIP-seq: using high-throughput sequencing to discover protein–DNA interactions. Methods 48 (2009) 240–248.
[5] Y. Zhang, et al., Model-based analysis of ChIP-Seq (MACS). Genome Biol. 9 (2008) R137.
[6] M.G. Guenther, S.S. Levine, L.A. Boyer, R. Jaenisch, R.A. Young, A chromatin landmark and transcription initiation at most promoters in human cells. Cell 130 (2007) 77–88.
[7] M.D. Young, et al., ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity. Nucleic Acids Res. 39 (2011) 7415–7427.
[8] S.G. Landt, et al., ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res. 22 (2012) 1813–1831.
[9] M.-A. Mendoza-Parra, W. Van Gool, M.A. Mohamed Saleem, D.G. Ceschin, H. Gronemeyer, A quality control system for profiles obtained by ChIP sequencing. Nucleic Acids Res. 41 (2013) e196.
[10] P.V. Kharchenko, M.Y. Tolstorukov, P.J. Park, Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat. Biotechnol. 26 (2008) 1351–1359.
[11] G.K. Marinov, A. Kundaje, P.J. Park, B.J. Wold, Large-scale quality analysis of published ChIP-seq data. G3 (Bethesda) 4 (2014) 209–223.