SCIENTIFIC DATA

OPEN

Data Descriptor: Epigenetic and transcriptional profiling of triple negative breast cancer

Andrea A. Perreault[1], Danielle M. Sprunger[2] & Bryan J. Venters[2]

Received: 28 September 2018
Accepted: 22 January 2019
Published: 5 March 2019

The human HCC1806 cell line is frequently used as a preclinical model for triple negative breast cancer (TNBC). Given that dysregulated epigenetic mechanisms are involved in cancer pathogenesis, emerging therapeutic strategies target chromatin regulators, such as histone deacetylases. A comprehensive understanding of the epigenome and transcription profiling in HCC1806 provides the framework for evaluating efficacy and molecular mechanisms of epigenetic therapies. Thus, to study the interplay of transcription and chromatin in the HCC1806 preclinical model, we performed nascent transcription profiling using Precision Run-On coupled to sequencing (PRO-seq). Additionally, we mapped the genome-wide locations for RNA polymerase II (Pol II), the histone variant H2A.Z, seven histone modifications, and CTCF using ChIP-exo. ChIP-exonuclease (ChIP-exo) is a refined version of ChIP-seq with near base pair precision mapping of protein-DNA interactions. In this Data Descriptor, we present detailed information on experimental design, data generation, quality control analysis, and data validation. We discuss how these data lay the foundation for future analysis to understand the relationship between the nascent transcription and chromatin.

Design Type(s)	epigenetic modification identification objective • replicate design
Measurement Type(s)	chromatin immunoprecipitation with exonuclease sequencing assay • precision nuclear run-on sequencing assay
Technology Type(s)	ChIP-exo • PRO-seq
Factor Type(s)	biological replicate • technology type
Sample Characteristic(s)	Homo sapiens • HCC1806 cell

[1]Chemical and Physical Biology Program at Vanderbilt University, Nashville, TN, USA. [2]Department of Molecular Physiology and Biophysics, Vanderbilt Genetics Institute, Vanderbilt Ingram Cancer Center, Vanderbilt University, Nashville, TN, USA. Correspondence and requests for materials should be addressed to B.J.V. (email: bryan.venters@vanderbilt.edu)

Background & Summary

Triple negative breast cancer (TNBC) is a highly aggressive and heterogeneous form of cancer[1]. TNBC is characterized by a lack of expression for estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2). Because TNBC lacks these proteins, targeted drug therapies directed against ER, PR, and HER2 are not possible. Therefore, standard care typically includes surgical resection, radiation, and chemotherapy. Prognosis remains poor for TNBC patients receiving standard care, highlighting the need for new and innovative therapeutic strategies. Emerging efforts have focused on the epigenome as a therapeutic target for TNBC[2]. It is now clear that epigenetic mechanisms play an important role in the pathogenesis, maintenance, and therapeutic resistance of the disease[3]. Consistent with this notion, a recently proposed model for transcriptional addiction in cancer suggests that transcriptional and chromatin regulators are potential targets for therapeutic intervention using innovative approaches[4].

Given that epigenetic and transcriptional regulators are implicated in the pathogenesis of TNBC, there is a need to better understand the mechanisms involved in the epigenetic modulation of genes expressed in TNBC[5]. Functional genomic approaches offer an unbiased, comprehensive glimpse into how transcriptional and chromatin regulators interact with the genome. For example, nascent transcriptional profiling (Precision Run-On sequencing, PRO-seq) and protein-DNA contact profiling (Chromatin Immunoprecipitation Exonuclease and sequencing, ChIP-exo) are state-of-the-art functional genomic tools that enable interrogation of global RNA synthesis and protein-DNA interactions, respectively[6,7]. PRO-seq and ChIP-exo studies generate scientifically valuable datasets due to their unique characteristics and reanalysis potential, which advances the sharing and reuse of scientific data.

PRO-seq and other nascent profiling approaches, such as GRO-seq, are more sensitive at detecting transcriptional responses than traditional assays like RNA-seq because they measure newly synthesized RNA, rather than the steady-state abundance of total RNA synthesis and degradation. This distinction is critical for detecting rapid transcriptional responses to stimuli, such as hormones and drug treatments[8,9]. In addition to using PRO-seq data to compute RNA synthesis rates[10], other biological features may be inferred from the unique data structure, such as transcription factor (TF) binding and enhancer activity[11–13]. Furthermore, computational tools have been developed to mine PRO-seq (or related GRO-seq) data and provide de novo annotation of long noncoding RNAs, microRNAs, and enhancer RNAs[14–17]. Taken together, the reanalysis potential of PRO-seq data is high.

ChIP-exo displays improved resolution and sensitivity over the traditional ChIP-seq method[18]. Rather than sequencing from the distal sonication borders as in ChIP-seq, ChIP-exo enriched DNA fragments are sequenced from the left and right 5' DNA borders of the protein-DNA crosslink site. The precision of the resulting data can be leveraged to provide unique and ultra-high resolution insights into the functional organization of the genome[19]. For example, ChIP-exo was uniquely capable of spatially resolving divergent, initiating, paused, and elongating RNA polymerase II (Pol II) on a genome-wide scale[20–22].

In this Data Descriptor, we provide a technical validation of twenty-two functional genomic data sets that interrogate the nascent transcription, Pol II binding, and chromatin architecture using the HCC1806 preclinical model for TNBC[24–26]. In this study, we focused on the TNBC HCC1806 cell line and have generated 2 PRO-seq data sets and 20 ChIP-exo data sets (2 biological replicates for each of the following targets: Pol II, H2A.Z, H3K4me3, H3K4me2, H3K4me1, H3K27ac, H3K9ac, H3K27me3, H4K20me1, and CTCF). ChIP-exo mapping of Pol II, a histone variant, and select histone modifications should enable other investigators to use these data sets for their own research to further understand the detailed interplay of Pol II and chromatin in ultra-high resolution in a preclinical model for breast cancer. On average, 36 million uniquely aligned reads were generated for each PRO-seq and ChIP-exo data set (Table 1). To facilitate interpretation of these data, we provide detailed information on experimental design (Fig. 1), sequence quality control analyses (Fig. 2), and biological validation (Fig. 3).

Methods

Tissue culture

The human HCC1806 triple negative breast cancer cell line, basal-like 2 subtype (ATCC) was maintained at 37 °C in 5% CO_2 between 20–80% confluency in RPMI 1640 (Roswell Park Memorial Institute, Gibco 11875-093) containing 10% bovine calf serum (Gibco 16170-078), 1% L-glutamine (Gibco 25030-081), and 1% Penicillin/Streptomycin (Gibco-15146-122).

PRO-seq library preparation

PRO-seq was performed as previously described[7] with isolated nuclei from 25 million cells from two biological replicates. To enable comparisons to drug treatment experiments, cells were treated with vehicle (final 0.03% DMSO (dimethyl sulfoxide)) for 4 h prior to harvest.

ChIP-exo library preparation

ChIP-exo was performed as previously described[6,20] with chromatin extracted from 50 million cells, ProteinG MagSepharose resin (GE Healthcare), and 5 ug of antibody directed against RNA polymerase II (Santa Cruz sc899), H2A.Z (EMD Millipore 07-594), H3K4me3 (Abcam ab8580), H3K4me2 (Abcam ab7766), H3K4me1 (Abcam ab8895), H3K27ac (Abcam ab4729), H4K9Ac (Abcam ab4441), H3K27me3

ChIP target	Antibody	Replicate	SRA identification	Read Length	Total Mapped Reads	Uniquely Mapped Reads	Unique Mapping Rate
PRO-seq		1	SRX4485383	50	46,406,885	35,062,933	76%
		2	SRX4485400	75	87,474,762	43,763,320	50%
		TOTAL			**133,881,647**	**78,826,253**	
Pol2	sc899 (Santa Cruz)	1	SRX4485384	50	82,965,774	63,381,472	76%
		2	SRX4485397	50	32,051,539	24,838,833	77%
		TOTAL			**115,017,313**	**88,220,305**	
H2A.Z	07-594 (EMD Milipore)	1	SRX4485385	50	9,783,691	6,506,998	67%
		2	SRX4485398	50	41,880,546	31,974,125	76%
		TOTAL			**51,664,237**	**38,481,123**	
H3K4me3	ab8580 (Abcam)	1	SRX4485387	50	37,105,094	30,487,398	82%
		2	SRX4485404	75	76,560,088	59,872,974	78%
		TOTAL			**113,665,182**	**90,360,372**	
H3K4me2	ab7766 (Abcam)	1	SRX4485388	50	28,077,305	21,915,585	78%
		2	SRX4485401	75	74,333,953	58,377,458	79%
		TOTAL			**102,411,258**	**80,293,043**	
H3K4me1	ab8895 (Abcam)	1	SRX4485389	50	43,663,792	32,717,084	75%
		2	SRX4485402	50	37,395,250	28,813,014	77%
		TOTAL			**81,059,042**	**61,530,098**	
H3K27Ac	ab4729 (Abcam)	1	SRX4485390	50	32,203,255	17,535,368	54%
		2	SRX4485395	50	43,865,531	31,734,234	72%
		TOTAL			**76,068,786**	**49,269,602**	
H3K9Ac	ab4441 (Abcam)	1	SRX4485392	50	30,684,340	26,874,881	88%
		2	SRX4485394	50	16,339,334	14,037,513	86%
		TOTAL			**47,023,674**	**40,912,394**	
H3K27me3	07-449 (EMD Millipore)	1	SRX4485391	50	36,950,654	29,977,079	81%
		2	SRX4485396	75	37,579,106	32,496,728	86%
		TOTAL			**74,529,760**	**62,473,807**	
H4K20me1	ab9051 (Abcam)	1	SRX4485399	75	79,587,575	62,260,072	78%
		2	SRX4485393	75	44,993,047	36,175,837	80%
		TOTAL			**124,580,622**	**98,435,909**	
CTCF	07-729 (EMD Millipore)	1	SRX4485386	75	65,862,116	48,218,698	73%
		2	SRX4485403	75	61,974,400	52,856,699	85%
		TOTAL			**127,836,516**	**101,075,397**	

Table 1. Sequencing read alignment statistics for PRO-seq and ChIP-exo data sets.

(EMD Millipore 07-449), H4K20me1 (Abcam ab9051), and CTCF (EMD Millipore 07-729). Two biological replicates were prepared for each ChIP target. To enable future comparisons to drug treatment experiments, cells were treated with vehicle (final 0.03% DMSO (dimethyl sulfoxide)) for 4 h prior to harvest. Libraries were sequenced using an Illumina NextSeq500 sequencer as single-end reads 50 or 75 nucleotides in length (Table 1).

Sequence read alignment and quality control

The base call quality for each sequenced read was assessed using the FastQC program (bioinformatics.babraham.ac.uk/projects/fastqc/) (Fig. 2a and Supp. Figs. 1–3). Sequence reads (fastq files) were aligned to the human hg19 reference genome build using BWA-MEM algorithm with default parameters[27]. The resulting bam files were first sorted using the Samtools Sort function, and then bam index files were generated using the Samtools Index function[28]. The purpose of bam index files is to enable viewing of raw sequencing data in a genome browser. Next, genome-wide read coverage and enrichment were assesses using deepTOOLS fingerprint plots[29] (Fig. 2b and Supp. Figs 4–6).

Biological validation

To estimate variance across biological replicates, the Pearson correlation coefficient for pairwise gene Reads Per Kilobase of genome per Million reads (RPKM) was computed (Fig. 2c, Supp. Fig. 7) using the

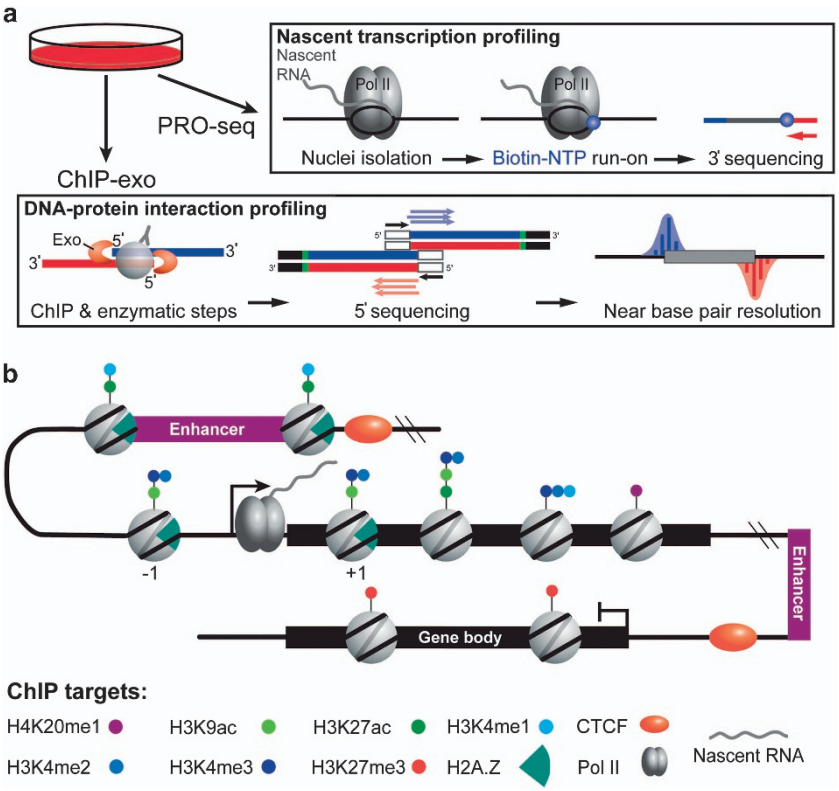

Figure 1. Experimental design and overview of ChIP targets. (**a**) HCC1806 cells were cultured and harvested for PRO-seq and ChIP-exo. PRO-seq measures nascent transcription. ChIP-exo identifies the exonuclease left and right borders that flank protein-DNA interactions. (**b**) Illustration of the genomic context for ChIP targets: Pol II, H2A.Z, H3K4me3, H3K4me2, H3K4me1, H3K27ac, H3K9ac, H3K27me3, H4K20me1, and CTCF.

HOMER (Hypergeometric Optimization of Motif EnRichment) suite[30]. Briefly, bam files were converted to tag directories using the makeTagDirectory function with the –genome, –checkGC, and –format options. To quantify and normalize tags within gene body regions to RPKM, the analyzeRepeats function was used with the –rpkm and –d options (2019SciDataVenters_RPKM.xlsx, Data Citation 1).

bedTOOLS was used to convert files from bam to Bigwig, and then the ChAsE (Chromatin Analysis and Exploration) suite was used to display the read distribution relative to the TSS from Bigwig files (Fig. 3a,b, Supp. Fig. 8)[33]. Raw sequencing tags were binned, smoothed, and RPKM computed using the deepTOOLS genomeCoverage tool (20 bp bin, 100 bp sliding window)[29]. Smoothed RPKM signal was visualized with Integrative Genomics Viewer (IGV) (Fig. 3c)[32].

Code availability

Below is a list of software used in this study.

FastQC v0.11.2 (www.bioinformatics.babraham.ac.uk/projects/fastqc/)
[27]BWA-MEM v0.7.13
[28]Samtools v1.3.1
[30]HOMER v4.6
[31]ChAsE v1.0.11
[29]deepTOOLS v2.2.4
[33]bedTOOLS v2.24.0
[32]IGV v2.3.77.

Data Records

PRO-seq and ChIP-exo bigwig data files from merged replicates were deposited in the NCBI Gene Expression Omnibus (GEO) (Data Citation 2). GEO linked PRO-seq and ChIP-exo bam data files for each replicate were deposited in the Sequence Read Archive (SRA) (Data Citation 3). Table 1 contains sequencing statistics for each data set and linked to its SRA identification number.

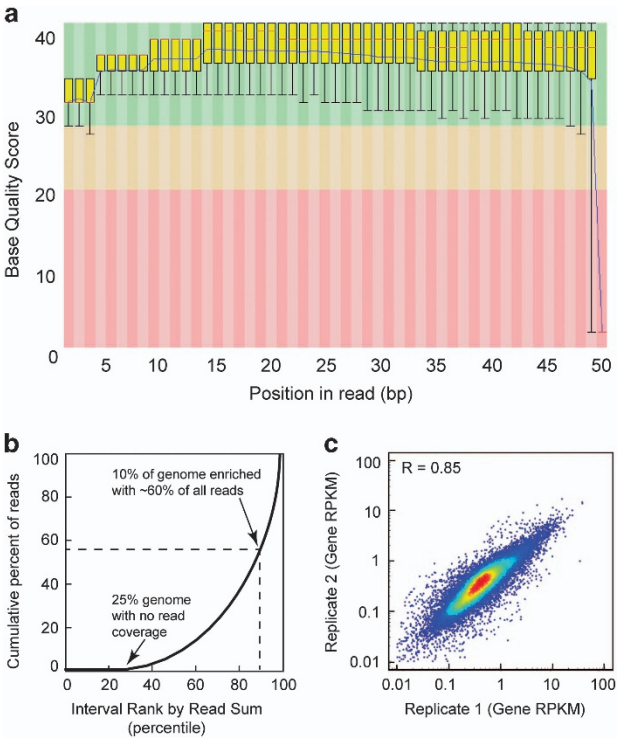

Figure 2. **Quality control, enrichment analysis, and reproducibility for PRO-seq and ChIP-exo data.** (**a**) Box-plot distribution of base quality scores are shown for H2A.Z ChIP-exo replicate 2. A score greater than 30 (green region) indicates a high confidence base call. (**b**) ChIP enrichment analysis plot that displays the cumulative percent of total reads found in a given percent of the mappable human genome. No ChIP enrichment would result in a diagonal trace. (**c**) Scatter plot correlation analysis for H2A.Z ChIP-exo biological replicates as measured by the Spearman correlation coefficient R-values (upper left corner).

Technical Validation

Overview of experimental design

In this study, functional genomic experiments using HCC1806 cells were designed with two primary goals in mind. First, PRO-seq data sets were generated to specifically measure nascent transcription. Second, the ChIP targets (Pol II, H2A.Z, H3K4me3, H3K4me2, H3K4me1, H3K27ac, H3K9ac, H3K27me3, H4K20me1, and CTCF) were selected so that Pol II binding and chromatin architecture may be examined on a genome-scale at high precision (Fig. 1). Histone modifications follow patterns of enrichment and delineate specific regions in the genome. For example, H3K4me1 and H3K27Ac are known to be found around distal enhancer regions, along with the histone variant H2A.Z. H3K4me1/2/3 and H3K9Ac are associated with promoter proximal regions, where H3K27ac and H2A.Z are also present[23,34–43]. H4K20me1 is critical for proper cell cycle progression and is typically depleted at promoters, but enriched in the body of genes[44]. Repressive marks and structural transcription factors, such as H3K27me3 and CTCF respectively, insulate regions of the genome that are not actively transcribed[35,45,46]. Taken together, reanalysis of this collection of data should enable new biological insights into chromatin dynamics in a pre-clinical breast cancer model. Below, we briefly describe the rationale and considerations for sequencing data analysis with respect to general read quality, genome alignment, ChIP enrichment, replicate correlation, and biological validation.

Raw sequence quality control analyses

To assess the quality of the raw sequencing data sets, base call scores were analyzed using the FastQC program and displayed as a box plot distribution at each base position (Fig. 2a and Supp. Figs. 1–3). The average base quality score for a majority of the data sets in the present study fell within the high confidence range (base quality score of 30–40, green region).

Raw sequence reads were aligned to the hg19 build of the human genome. On average, 46 million total aligned reads were generated for each PRO-seq and ChIP-exo data sets (Table 1). Because of the ambiguity of reads that align to multiple locations throughout the genome, we only retain uniquely aligned reads for subsequent analyses. On average, 36 million uniquely aligned reads were obtained per data set, representing an average unique alignment rate of 76%.

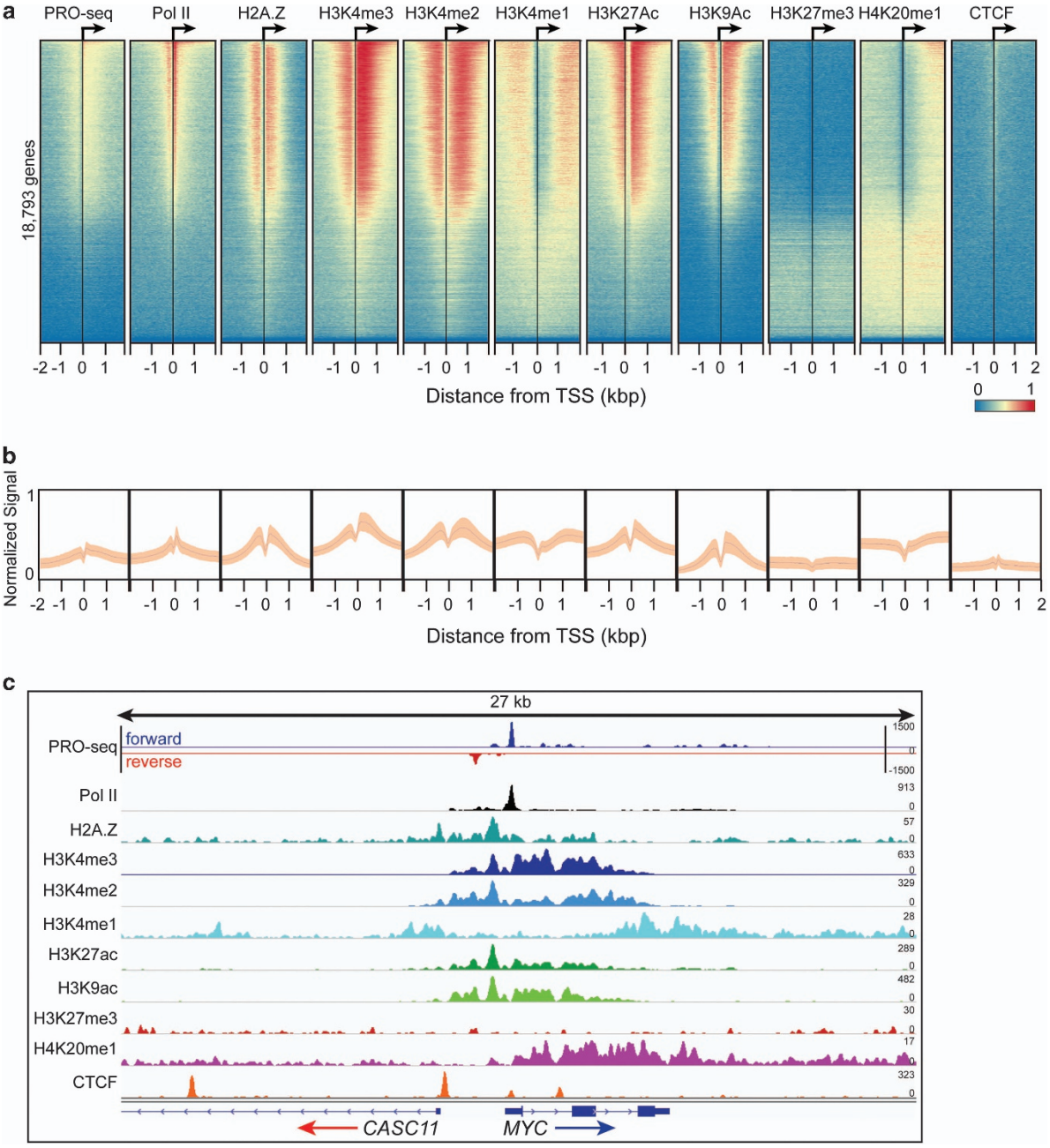

Figure 3. **Genomic distribution of RPKM normalized signal for PRO-seq and ChIP-exo targets (Pol II, H2A.Z, H3K4me3, H3K4me2, H3K4me1, H3K27ac, H3K9ac, H3K27me3, H4K20me1, and CTCF).** (**a**) Row-linked heatmaps show RPKM normalized number of reads across a 4 kb genomic interval in 40 bp bins relative to the TSS. Heatmaps were generated from merged biological replicate pairs for each data set. Regions are sorted in descending order based on average row tag density for Pol II. Each row represents a gene, with 18,793 genes displayed. Red and blue reflect high and low read densities, respectively. (**b**) Composite plots below each heatmap quantify the normalized tag density. The central trace denotes the average tag density for each 40 bp bin and the orange fill reflects the standard deviation. (**c**) Genome browser view of PRO-seq and ChIP-exo signal for the indicated targets in HCC1806 cells shown at the *MYC* gene. Tag distributions were smoothed and RPKM normalized using deepTOOLS. Traces were generated from merged biological replicate pairs.

Two critical questions for assessing ChIP sequencing data quality are: 1) how much of the genome is represented by a given experiment? and 2) to what extent did the ChIP assay enrich for specific regions of the genome? Typically, high genome coverage and strong ChIP enrichment are desirable in ChIP experiments. To determine genome coverage and ChIP enrichment simultaneously, we used the

deepTOOLS suite to perform a fingerprint analysis (Fig. 2b). In the case of H2A.Z ChIP-exo (Fig. 2b), the fingerprint plot trace intersects the x-axis at 25, indicating 75% genome coverage. In fingerprint plots, a rightward deflection of the trace indicates the extent of ChIP enrichment. Given a point along the trace that is the point of intersection from the axes, the corresponding values on the x- and y-axes denote the percent of genome and the percent of all uniquely aligned reads, respectively. Together, these values reflect ChIP enrichment.

For example, the H2A.Z ChIP-exo fingerprint trace reveals that 10% of the genome is enriched with 60% of all uniquely aligned reads, suggesting strong enrichment in the H2A.Z ChIP-exo data set (Fig. 2b). Fingerprint plots for other replicates showed similar patterns of genome coverage and ChIP enrichment (Supp. Figs 4–6). Theoretically, complete genome coverage with no enrichment would be result in a trace with a slope equal to one that intersects the origin (eg: whole genome sequencing wherein 50% of the genome is contains 50% of all aligned reads).

Biological validation

After verifying the quality of the raw sequencing data, we next sought to provide evidence of biological validity for the data. First, we determined the extent to which biological replicates were reproducible using correlation scatter plots (Fig. 2c). For each gene, the RPKM was computed using the HOMER suite (Data Citation 1). Pearson correlation coefficients (R values) were computed for pairwise correlation plots of gene RPKM across biological replicates. For example, biological replicates for H2A.Z ChIP-exo analysis displayed an R value of 0.85, indicating high reproducibility (Fig. 2c). Overall, correlation analysis resulted in an average R-values of 0.86 (Supp. Fig. 7). Because CTCF is typically enriched in intergenic regions rather than within gene bodies, correlation analysis compared peak RPKM.

Given that certain histone modifications are consistently found at distinct regions of the stereotypical gene, analyzing global patterns of ChIP signal relative to TSSs is a useful method to assess biological validation[47]. For example, it is well established that once Pol II initiates transcription of genes in metazoans, Pol II moves into a stable paused state 30–50 bp downstream of the TSS. Nascent transcription profiling with PRO-seq enables quantification of RNA synthesis and largely coincides with Pol II ChIP-exo density. H3K4me1/2/3 and H3K9Ac are associated with promoter proximal regions, surrounding the TSS of active genes in combination with H3K27Ac. The histone variant H2A.Z is consistently incorporated into the + 1 nucleosome of actively transcribed genes. Distal to gene promoters, H3K4me1 and H3K27Ac have been used as predictive marks of enhancers, which regulate the transcription of their target genes in a distance and orientation independent manner.

Thus, to examine global patterns of ChIP enrichment, the Chromatin Analysis and Exploration (ChAsE) heatmap tool was used to align ChIP signal merged from both biological replicates to TSSs (Fig. 3a, sorted by max peak; and Supp. Fig. 8, sorted by max peak position). Quantification of signal density relative to TSSs is displayed as a composite plot below each heatmap (Fig. 3b). As expected, Pol II was strongly enriched at the pause site just downstream of the TSS. H2A.Z enrichment at the -1 and + 1 nucleosomes immediately flanked Pol II density. H3K4me2, H3K4me3, and H3K9ac were enriched at the + 1 nucleosome as well, but also spread into the body of gene, overlapping the + 1, + 2, and + 3 nucleosome positions. Interestingly, H3K27ac density was similar to H3K9ac but avoided the + 1 nucleosome position. H3K4me1 and H4K20me1 density excluded promoter regions, but were enriched further downstream into the gene body. In contrast to the other histone modifications and consistent with its association to gene repression, H3K27me3 was enriched at genes with the least Pol II and PRO-seq density. Lastly, as expected, gene bodies largely lacked CTCF signal. To examine individual examples of global patterns, RPKM normalized tracks for PRO-seq and ChIP-exo signal were displayed using the Integrative Genome Viewer (IGV), and displayed at the *MYC* gene locus (Fig. 3c). Finally, a comparison of Pol II ChIP-exo and ChIP-seq signal at promoters (Supp. Fig. 9) shows that Pol II ChIP-exo more clearly resolves adjacent peaks of Pol II enrichment corresponding to Pol II pausing just after the TSS and divergent transcription just upstream of the TSS. This is evident both as a genome-wide pattern (Supp. Fig. 9a,b) and in several anecdotal examples (Supp. Fig. 9c).

Taken together, the data presented in this Data Descriptor represents high quality next generation sequencing data that are biologically valid, and should be useful to future studies that seek to understand the interplay of Pol II and chromatin in high resolution on a global scale.

References

1. Lehmann, B. D. & Pietenpol, J. A. Clinical implications of molecular heterogeneity in triple negative breast cancer. *Breast* **24** (Suppl 2): S36–S40 https://doi.org/10.1016/j.breast.2015.07.009 (2015).
2. Garmpis, N. *et al.* Histone Deacetylases as New Therapeutic Targets in Triple-negative Breast Cancer: Progress and Promises. *Cancer Genomics Proteomics* **14,** 299–313 https://doi.org/10.21873/cgp.20041 (2017).
3. Flavahan, W. A., Gaskell, E. & Bernstein, B. E. Epigenetic plasticity and the hallmarks of cancer. *Science* **357** https://doi.org/10.1126/science.aal2380 (2017).
4. Bradner, J. E., Hnisz, D. & Young, R. A. Transcriptional Addiction in. *Cancer. Cell* **168,** 629–643 https://doi.org/10.1016/j.cell.2016.12.013 (2017).
5. Judes, G. *et al.* High-throughput <<Omics>>technologies: New tools for the study of triple-negative breast cancer. *Cancer Lett* **382,** 77–85 https://doi.org/10.1016/j.canlet.2016.03.001 (2016).
6. Perreault, A. A. & Venters, B. J. The ChIP-exo Method: Identifying Protein-DNA Interactions with Near Base Pair Precision. *J Vis Exp* https://doi.org/10.3791/55016 (2016).

7. Kwak, H., Fuda, N. J., Core, L. J. & Lis, J. T. Precise maps of RNA polymerase reveal how promoters direct initiation and pausing. *Science* **339,** 950–953 https://doi.org/10.1126/science.1229386 (2013).
8. Hah, N. *et al.* A rapid, extensive, and transient transcriptional response to estrogen signaling in breast cancer cells. *Cell* **145,** 622–634 https://doi.org/10.1016/j.cell.2011.03.042 (2011).
9. Zhao, Y. *et al.* High-Resolution Mapping of RNA Polymerases Identifies Mechanisms of Sensitivity and Resistance to BET Inhibitors in t(8;21) AML. *Cell Rep* **16,** 2003–2016 https://doi.org/10.1016/j.celrep.2016.07.032 (2016).
10. Danko, C. G. *et al.* Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells. *Mol Cell* **50,** 212–222 https://doi.org/10.1016/j.molcel.2013.02.015 (2013).
11. Hait, T. A., Amar, D., Shamir, R. & Elkon, R. FOCS: a novel method for analyzing enhancer and gene activity patterns infers an extensive enhancer-promoter map. *Genome Biol* **19,** 56 https://doi.org/10.1186/s13059-018-1432-2 (2018).
12. Wang, Z., Martins, A. L. & Danko, C. G. RTFBSDB: an integrated framework for transcription factor binding site analysis. *Bioinformatics* **32,** 3024–3026 https://doi.org/10.1093/bioinformatics/btw338 (2016).
13. Danko, C. G. *et al.* Identification of active transcriptional regulatory elements from GRO-seq data. *Nat Methods* **12,** 433–438 https://doi.org/10.1038/nmeth.3329 (2015).
14. Azofeifa, J. G., Allen, M. A., Lladser, M. E. & Dowell, R. D. An Annotation Agnostic Algorithm for Detecting Nascent RNA Transcripts in GRO-Seq. *IEEE/ACM Trans Comput Biol Bioinform* **14,** 1070–1081 https://doi.org/10.1109/TCBB.2016.2520919 (2017).
15. Chae, M., Danko, C. G. & Kraus, W. L. groHMM: a computational tool for identifying unannotated and cell type-specific transcription units from global run-on sequencing data. *BMC Bioinformatics* **16,** 222 https://doi.org/10.1186/s12859-015-0656-3 (2015).
16. Liu, Q. *et al.* Identification of active miRNA promoters from nuclear run-on RNA sequencing. *Nucleic Acids Res* **45,** e121 https://doi.org/10.1093/nar/gkx318 (2017).
17. Sun, M., Gadad, S. S., Kim, D. S. & Kraus, W. L. Discovery, Annotation, and Functional Analysis of Long Noncoding RNAs Controlling Cell-Cycle Gene Expression and Proliferation in Breast Cancer Cells. *Mol Cell* **59,** 698–711 https://doi.org/10.1016/j.molcel.2015.06.023 (2015).
18. Rhee, H. S. & Pugh, B. F. Comprehensive genome-wide protein-DNA interactions detected at single-nucleotide resolution. *Cell* **147,** 1408–1419 https://doi.org/10.1016/j.cell.2011.11.013 (2011).
19. Venters, B. J. Insights from resolving protein-DNA interactions at near base-pair resolution. *Brief Funct Genomics* **17,** 80–88 https://doi.org/10.1093/bfgp/elx043 (2018).
20. Pugh, B. F. & Venters, B. J. Genomic Organization of Human Transcription Initiation Complexes. *PloS one* **11,** e0149339 https://doi.org/10.1371/journal.pone.0149339 (2016).
21. Chen, Y. *et al.* Principles for RNA metabolism and alternative transcription initiation within closely spaced promoters. *Nat Genet* **48,** 984–994 https://doi.org/10.1038/ng.3616 (2016).
22. Shao, W. & Zeitlinger, J. Paused RNA polymerase II inhibits new transcriptional initiation. *Nat Genet* **49,** 1045–1051 https://doi.org/10.1038/ng.3867 (2017).
23. Rhee, H. S., Bataille, A. R., Zhang, L. & Pugh, B. F. Subnucleosomal structures and nucleosome asymmetry across a genome. *Cell* **159,** 1377–1388 https://doi.org/10.1016/j.cell.2014.10.054 (2014).
24. Gazdar, A. F. *et al.* Characterization of paired tumor and non-tumor cell lines established from patients with breast cancer. *Int J Cancer* **78,** 766–774 (1998).
25. Lehmann, B. D. *et al.* Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J Clin Invest* **121,** 2750–2767 https://doi.org/10.1172/JCI45014 (2011).
26. Neve, R. M. *et al.* A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. *Cancer Cell* **10,** 515–527 https://doi.org/10.1016/j.ccr.2006.10.008 (2006).
27. Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26,** 589–595 https://doi.org/10.1093/bioinformatics/btp698 (2010).
28. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 https://doi.org/10.1093/bioinformatics/btp352 (2009).
29. Ramirez, F., Dundar, F., Diehl, S., Gruning, B. A. & Manke, T. deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res* **42,** W187–W191 https://doi.org/10.1093/nar/gku365 (2014).
30. Heinz, S. *et al.* Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* **38,** 576–589 https://doi.org/10.1016/j.molcel.2010.05.004 (2010).
31. Younesy, H. *et al.* An interactive analysis and exploration tool for epigenomic data. *Computer Graphics Forum* **32,** 91–100 https://doi.org/10.1111/cgf.12096 (2013).
32. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat Biotechnol* **29,** 24–26 https://doi.org/10.1038/nbt.1754 (2011).
33. Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr Protoc Bioinformatics* **47,** 11 12 11-11 12 34 https://doi.org/10.1002/0471250953.bi1112s47 (2014).
34. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A* **107,** 21931–21936 https://doi.org/10.1073/pnas.1016071107 (2010).
35. Heintzman, N. D. *et al.* Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature* **459,** 108–112 https://doi.org/10.1038/nature07829 (2009).
36. Heintzman, N. D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat Genet* **39,** 311–318 https://doi.org/10.1038/ng1966 (2007).
37. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat Biotechnol* **28,** 817–825 https://doi.org/10.1038/nbt.1662 (2010).
38. Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473,** 43–49 https://doi.org/10.1038/nature09906 (2011).
39. Brunelle, M. *et al.* The histone variant H2A.Z is an important regulator of enhancer activity. *Nucleic Acids Res* **43,** 9742–9756 https://doi.org/10.1093/nar/gkv825 (2015).
40. Cauchy, P., Koch, F. & Andrau, J. C. Two possible modes of pioneering associated with combinations of H2A.Z and p300/CBP at nucleosome-occupied enhancers. *Transcription* **8,** 179–184 https://doi.org/10.1080/21541264.2017.1291395 (2017).
41. Chen, P., Wang, Y. & Li, G. Dynamics of histone variant H3.3 and its coregulation with H2A.Z at enhancers and promoters. *Nucleus* **5,** 21–27 https://doi.org/10.4161/nucl.28067 (2014).
42. Mavrich, T. N. *et al.* Nucleosome organization in the Drosophila genome. *Nature* **453,** 358–362 https://doi.org/10.1038/nature06929 (2008).
43. Segala, G., Bennesch, M. A., Pandey, D. P., Hulo, N. & Picard, D. Monoubiquitination of Histone H2B Blocks Eviction of Histone Variant H2A.Z from Inducible Enhancers. *Mol Cell* **64,** 334–346 https://doi.org/10.1016/j.molcel.2016.08.034 (2016).
44. Beck, D. B., Oda, H., Shen, S. S. & Reinberg, D. PR-Set7 and H4K20me1: at the crossroads of genome integrity, cell cycle, chromosome condensation, and transcription. *Genes Dev* **26,** 325–337 https://doi.org/10.1101/gad.177444.111 (2012).

45. Phillips, J. E. & Corces, V. G. CTCF: master weaver of the genome. *Cell* **137,** 1194–1211 https://doi.org/10.1016/j.cell.2009.06.001 (2009).
46. Tang, Z. *et al.* CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell* **163,** 1611–1627 https://doi.org/10.1016/j.cell.2015.11.024 (2015).
47. Venkatesh, S. & Workman, J. L. Histone exchange, chromatin structure and the regulation of transcription. *Nat Rev Mol Cell Biol* **16,** 178–189 https://doi.org/10.1038/nrm3941 (2015).

Data Citations

1. Perreault, A. A., Sprunger, D. M. & Venters, B. J. *figshare* https://doi.org/10.6084/m9.figshare.7473374 (2018).
2. Perreault, A. A., Sprunger, D. M. & Venters, B. J. *NCBI Gene Expression Omnibus* GSE118033 (2018).
3. Perreault, A. A., Sprunger, D. M. & Venters, B. J. *NCBI Sequence Read Archive* SRP155750 (2018).

Acknowledgements

The corresponding author of this study had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis. Special thanks to Jennifer Pietenpol for kindly sharing the HCC1806 cell line. We would also like to thank Scott Hiebert for helpful advice and assistance with the PRO-seq assay. Thanks to Tyler Hansen and Zenab Mchaourab for assistance with bioinformatic analyses.

Author Contributions

A.P. conducted bioinformatics data analysis, prepared figures, and wrote the manuscript. D.W. performed experiments. B.V. designed experiments, conceived bioinformatic analyses, prepared figures, and wrote the manuscript.

Additional Information

Supplementary information accompanies this paper at http://www.nature.com/sdata

Competing interests: The authors declare no competing interests.

How to cite this article: Perreault, A. A. *et al.* Epigenetic and transcriptional profiling of triple negative breast cancer. *Sci. Data.* 6:190033 https://doi.org/10.1038/sdata.2019.33 (2019).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.