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

Andrea A. Perreault¹, Danielle M. Sprunger² & Bryan J. Venters²

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                                       |

¹Chemical and Physical Biology Program at Vanderbilt University, Nashville, TN, USA. ²Department 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. 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.

It is now clear that epigenetic mechanisms play an important role in the pathogenesis, maintenance, and therapeutic resistance of the disease. 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.

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. 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. 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. In addition to using PRO-seq data to compute RNA synthesis rates, other biological features may be inferred from the unique data structure, such as transcription factor (TF) binding and enhancer activity.

ChIP-exo displays improved resolution and sensitivity over the traditional ChIP-seq method. Rather than sequencing from the distal sonication borders as in ChIP-seq, ChIP-exo enriched DNA fragments are sequenced from the left and right 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. 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.

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. 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% CO2 between 20–80% confluence 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 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 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
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. The resulting bam files were first sorted using the Samtools Sort function, and then bam index files were generated using the Samtools Index function. 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 assessed using deepTOOLS fingerprint plots (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

---

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

| ChIP target | Antibody       | Replicate | SRA identification | Read Length | TotalMappedReads | UniquelyMappedReads | UniqueMappingRate |
|-------------|----------------|-----------|--------------------|-------------|------------------|---------------------|-------------------|
| 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 Millipore) | 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          |                   |
| H3K1me3     | 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          |                   |
| H3K1me2     | ab7766 (Abcam) | 1         | SRX4485388         | 50          | 28,077,305       | 21,915,385          | 78%               |
|             |                | 2         | SRX4485401         | 75          | 74,333,953       | 58,377,458          | 79%               |
|             |                | TOTAL     |                    |             | 102,411,258      | 80,293,043          |                   |
| H3K1me1     | 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       | 17,353,168          | 54%               |
|             |                | 2         | SRX4485393         | 50          | 43,865,531       | 31,734,234          | 72%               |
|             |                | TOTAL     |                    |             | 76,068,786       | 49,269,602          |                   |
| H3K9Ac      | ab4441 (Abcam) | 1         | SRX4485392         | 50          | 30,844,340       | 26,874,481          | 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          | 37,379,106       | 32,496,728          | 86%               |
|             |                | TOTAL     |                    |             | 74,529,760       | 62,473,807          |                   |
| H4K20me1    | ab9051 (Abcam) | 1         | SRX4485399         | 75          | 79,387,575       | 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         |                   |
HOMER (Hypergeometric Optimization of Motif EnRichment) suite. 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). Raw sequencing tags were binned, smoothed, and RPKM computed using the deepTOOLS genomeCoverage tool (20 bp bin, 100 bp sliding window). Smoothed RPKM signal was visualized with Integrative Genomics Viewer (IGV) (Fig. 3c).

Code availability
Below is a list of software used in this study.

- FastQC v0.11.2 (www.bioinformatics.babraham.ac.uk/projects/fastqc/)
- BWA-MEM v0.7.13
- Samtools v1.3.1
- HOMER v4.6
- ChAsE v1.0.11
- deepTOOLS v2.2.4
- bedTOOLS v2.24.0
- 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.
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. H4K20me1 is critical for proper cell cycle progression and is typically depleted at promoters, but enriched in the body of genes. Repressive marks and structural transcription factors, such as H3K27me3 and CTCF respectively, insulate regions of the genome that are not actively transcribed. 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).

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).
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

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.
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. 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 position. 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. Garnpis, 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/engp.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 Triple-negative Breast Cancer Genomics Proteomics. *Breast* **29**, 75–85 https://doi.org/10.1016/j.canlet.2016.03.001 (2016).
5. Jubes, 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).
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.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

The Creative Commons Public Domain Dedication waiver http://creativecommons.org/publicdomain/zero/1.0/ applies to the metadata files made available in this article.

© The Author(s) 2019