**Supplemental Experimental Procedures for *diffloop***

**Preprocessing**

Assays such as ChIA-PET enrich for chromatin interactions where interacting loci, termed anchors, are bound by a protein of interest. Anchors linked by paired-end tags/reads (PETs) represent distal regions of DNA that co-localize in three-dimensional space. The preprocessed data input to *diffloop* consists of interaction counts between putatively interacting anchors, as represented in the graphical abstract. Raw data preprocessing pipelines that produce the standard bed paired-end (.bedpe) input data format. A number of software solutions, including the Mango and ChIA-PET2 pipelines, exist to process raw sequencing reads from a ChIA-PET assay into putative loops in the bedpe format that is imported into *diffloop* (Li et al., 2016; Li et al., 2010; Paulsen et al., 2014; Phanstiel et al., 2015)*.* These tools typically consider interactions for a single sample at a time when identifying statistically significant loops. For the purposes of differential analysis, it is important to make the determination of valid 3D interactions using all samples collectively on order to define a common set of loops. We therefore recommend the use of inclusive preprocessing parameters that do not filter out interactions, retaining all data for import and subsequent filtering in *diffloop*. In Mango, retaining is achieved by setting the reportallpairs flag to TRUE. *diffloop* collates these read counts and assembles a list of anchors and a counts matrix to be utilized in differential association as shown inthe graphical abstract.

When loops data is imported into *diffloop*, the precise anchor loci may very slightly from sample to sample. As defining a common feature space is necessary for differential testing, users have the option to pad loop anchors when using the **loopsMakes()** function. This function first loops over each file to determine the union of loop anchors, merging and padding anchors as specified by the user to define a common feature or loop space. One this set of common loops is established, the PETs that support each loop in each sample are imported and the counts matrix is populated. In effect, the counts matrix provides the level of evidence supporting each putative loop (row) for each sample (column). Under this construct, familiar techniques such as differential testing, normalization, and principal component analyses can be applied to the counts matrix.

**diffloop Data Structure**

The core functionality of *diffloop* is designed around **loops** objects, a novel S4 class implemented in the package. Each **loops** object contains five slots that provide an efficient storage of three-dimensional data in the R environment. Specifically, the anchors slot contains a GRanges (Lawrence et al., 2013) object specifying the genomic coordinates of the DNA anchors identified through peak calling; the interactions slot represents each loop (row) by a pair of indices (two columns) specifying its two anchors; the counts matrix summarizes the number of supporting PETs for each loop (row) per sample (column); the colData slot contains per-sample information, such as group labels and normalizing constants; and the rowData slot, which provides per-loop annotation, such as loop width, loop type, and statistical significance measures.

**ENCODE ChIA-PET Data**

All ChIA-PET data in this study was generated as part of the ENCODE Project and downloaded from the Sequence Read Archive (SRA)(Consortium, 2012). The format of raw ChIA-PET data is .fastq files that correspond to paired-end reads from a sequencing experiment. For our preprocessing, we used the default parameters in Mango (Phanstiel et al., 2015) except for specifying that all interactions be preserved (reportallpairs = TRUE) and ChIP peaks be extended by 1,000 base pairs (peakslop = 1000) rather than the default 500 bp. Additionally, we specified linker sequences previously described in the ENCODE ChIA-PET protocol, (Consortium, 2012) which also correspond to the default parameters in Mango. **Table S1 (A)** provides an overview of the ChIA-PET samples used in this study, including the raw read counts, the location of the data on GEO, as well as the number of PETs used in *diffloop* after data processing with Mango.

**Quality Control in diffloop**

We filtered amplified or deleted copy number variation (CNV) regions in either of K562 or MCF-7 using the **removeRegion()** function to reduce the chance that genome alterations would bias differential loop calls. We next retained only those loops whose interaction counts were significantly higher than that expected based on the background chromatin interaction frequency using a threshold of 0.01 on q-values generated by the **mangoCorrection()** function. *diffloop* aggregates count data across all samples, providing more power to call valid loops than analyzing each sample individually. In order to further reduce the multiple testing burden, we further restricted loops to those with a minimum of 2 samples with at least 2 PETs per loop (similar as to what was used previously (Ji et al., 2016)) using the **filterLoops()** function. The low counts associated with the discarded loops would preclude meaningful inference about between group differences.

One peculiar feature of the data was the presence of “discordant” loops that were highly variable between the replicates. Setting a threshold of five or more counts in one replicate but zero in the other identified 337 such loops. Some of these discordant loops appear significantly differential as a result of the variance shrinkage performed in the association model. For example, while a loop with counts of 45 and 0 for one group and 0 and 0 for another is classified as differential, this finding is unlikely to be reliable. Many (166 of 337 or 49%) of these identified discordant loops (such as the example noted above) were removed using the **filterLoops()** function since they don’t meet the criterion of being present in 2+ samples.

**Loop Annotation**

The **annotateLoops()** function classifies each loop as enhancer-promoter, promoter-promoter, enhancer-enhancer, or no special annotation for POL2 loops. Additional annotation for loops mediated by CTCF or cohesin subunits is also currently supported. Moreover, loops that connect distal regulatory elements and gene promoters can be selected using the **keepEPloops()** function. For the analyses described in this manuscript, enhancer regions were defined by a 1kb radius around an H3K27ac peak for either cell type. Other epigenetic markings may also be suitable for nuanced logic in defining promoter and regulatory regions. For our analyses, peaks were downloaded from GEO using the accession numbers provided in **Table S1 (B)**. Promoter regions were defined as being within a 1kb radius of a RefSeq transcription start site.

**Differential Loop Calling**

To determine differential topological features, *diffloop* uses the edgeR package to model PET counts from the loops object counts matrix using a negative binomal distribution (Robinson et al., 2010). To account for variation in sequencing depth, a per-sample size factor is included as an offset in the model, as similarly implemented in DESeq2. Loop-wise dispersion estimates are stabilized using an empirical Bayes shrinkage procedure. In effect, the implementation of differential loop calling in *diffloop* applies the RNA-Seq gene read count test from edgeR to loop PET counts. We’ve additionally implemented a modified version of the limma-voom framework for identifying differential loops using precision weights though the negative binomial regression is presently the default in *diffloop*. (Law et al., 2014)

**Differential Expression Analyses**

Paired 75 base pair reads from PolyA RNA-Seq for each of the K562 and MCF-7 cell lines from the ENCODE Project were processed (GEO series GSE33480). These data included two samples for K562 (GSM958729) and three samples for MCF-7 (GSM958745). An additional replicate was processed for K562 (GSM646524) for a balanced differential expression analysis. Each samples’ reads were individually aligned using Tophat v1.0.14 (Trapnell et al., 2009) and hg19 RefSeq reference transcriptome counts were generated using HTSeq 0.6.1.(Anders et al., 2015) Differential expression was performed using DESeq2 v1.11.45 (Love et al., 2014) to determine variable gene expression between the two cancer cell lines. Enhancer-promoter loops that uniquely linked to a single transcription start site were annotated with the summary statistics from DESeq2 using the **annotateLoops.dge()** function. While this function has additional parameters to handle loops that do not clearly link to a single transcription start site, all analyses including transcription annotation retained only enhancer-promoter loops where the “promoter” anchor mapped to within 1kb of a single transcription start site in the hg19 Refseq build.

**Integration of DNA Methylation and ChIP-Seq Data**

*diffloop* provides means for integration of processed epigenetic data as shown in **Figure 2.** To demonstrate this functionality, raw probe intensities from the Illumina 450k array were processed using minfi v1.3.1(Aryee et al., 2014) and exported as .bedgraph format files for both the K562 and MCF-7 cell lines. Per-anchor methylation was computed by averaging the Beta methylation estimates across all CpGs contained in the specific anchor using the **annotateAnchors.bed()** function. Bigwigs H3K27ac were downloaded from GEO accessions as specified in **Table 1 (B)**. Similar to the methylation values, per-anchor intensities for the ChIP-Seq and chromatin accessibility were computed by averaging over all values contained in the specific anchor using the **annotateAnchors.bigwig()** function.

**Visualization**

Within the *diffloop* package infrastructure, several functions related to the **loopPlot()** function call enable the visualization of differential loops for selected samples. For the specific plots shown in **Figure 1** that also show epigenetic tracks (i.e. H3K27ac), we visualize these samples using a shiny app available at dnalandscaper.aryeelab.org. On this site, a detailed explanation of how one can visualize their own looping samples in the browser is available in the “Guide” page.

**Gene set enrichment**

Using the full set of loops that were examined in differential testing, we assessed MsigDB hallmark gene sets (Subramanian et al., 2005) using a Wilcoxon rank sum test, which yieldednine pathways enriched (FDR q < 0.1) for genes with differential enhancer-promoter loops have existing evidence of relevance to the two cell types and cancers. For loops stronger in MCF7, four pathways were significantly enriched with the following FDR q-values: ESTROGEN\_RESPONSE\_EARLY (q = 7.05E-09), TNFA\_SIGNALING\_VIA\_NFKB (q = 4.97E-06), P53\_PATHWAY (0.019), ESTROGEN\_RESPONSE\_LATE (q = 0.022). For loops stronger in K562, the following pathways were significantly enriched: MYC\_TARGETS\_V1 (q = 0.00017), E2F\_TARGETS (q = 0.00017), G2M\_CHECKPOINT (q = 0.00017), REACTIVE\_OXIGEN\_SPECIES\_PATHWAY (q = 0.0089), HEME\_METABOLISM (q = 0.087). These results suggest that variable looping affects pathways genome-wide that are in part responsible for variation in cellular phenotypes.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **(A) Summary Statistics for ChIA-PET from GSE33664** | | | | |
| **Local Name** | **SRR(s)** | **# unique loops** | **# useful PETs** | **# raw PETs** |
| K562\_r1 | SRR372747 | 31,721 | 95,210 | 25,778,343 |
| K562\_r2 | SRR372748 | 50,072 | 161,053 | 37,685,853 |
| MCF7\_r1 | SRR372741 | 23,329 | 50,216 | 29,098,917 |
| MCF7\_r2 | SRR372742 | 36,199 | 85,640 | 40,474,778 |

|  |  |  |
| --- | --- | --- |
| **(B) GEO Sources of ENCODE Epigenetic and RNA-Seq Data used in Integrative Analyses** | | |
| **Data Type** | **K562** | **MCF-7** |
| 450k Methylation Arrays | GSM999341 | GSM999373 |
| CNV Regions | GSM999287 | GSM999333 |
| H3K27ac Peaks | GSM733656 | GSM946850 |
| RNA-Seq Data | GSM646524,GSM958729 | GSM958745 |

**Table S1**: Published data used in this manuscript for sample analysis with *diffloop*. (A) Accession of GEO sequencing reads for each of the ChIA-PET samples with summary statistics. (B) Accessions of the epigenetic, genetic, and transcriptomic data used for downstream analyses in *diffloop*.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Region** | **k562\_r1** | **k562\_r2** | **mcf7\_r1** | **mcf7\_r2** | **Width** | **logFC** | **PValue** | **FDR** | **Promoter Gene(s)** |
| chr6:26193587-26208554 | 169 | 242 | 0 | 0 | 8724 | -10.23 | 2.03E-50 | 1.30E-47 | HIST1H3D, HIST1H2AD, HIST1H2BF, HIST1H4E |
| chr17:56403949-56417755 | 155 | 213 | 0 | 0 | 8477 | -10.07 | 1.58E-46 | 7.20E-44 | BZRAP1 |
| chr6:74221945-74234813 | 127 | 190 | 0 | 0 | 6750 | -9.85 | 1.88E-41 | 6.91E-39 | EEF1A1 |
| chr7:100024885-100035440 | 91 | 133 | 0 | 0 | 6621 | -9.35 | 1.12E-31 | 2.28E-29 | ZCWPW1, MEPCE, PPP1R35 |
| chr2:85763824-85776025 | 75 | 125 | 0 | 0 | 7324 | -9.18 | 7.70E-29 | 1.31E-26 | MAT2A |
| chr12:52582290-52599042 | 0 | 0 | 55 | 103 | 11460 | 9.97 | 2.14E-64 | 2.05E-61 | KRT80 |
| chr14:38051267-38066780 | 0 | 0 | 62 | 87 | 11045 | 9.90 | 8.59E-62 | 7.47E-59 | TTC6 |
| chr3:161087371-161126836 | 0 | 0 | 49 | 68 | 35464 | 9.55 | 4.61E-51 | 3.39E-48 | SPTSSB |
| chr17:39676448-39706653 | 0 | 0 | 40 | 74 | 22387 | 9.50 | 7.60E-50 | 4.55E-47 | KRT15, KRT19 |
| chr20:45984705-46003273 | 0 | 0 | 52 | 56 | 11916 | 9.44 | 5.29E-48 | 2.66E-45 | ZMYND8 |

**Table S2**: Top differential loops between MCF-7 and K562. The 5 most significant differential enhancer-promoter loops between these two cell types both more prevalent in MCF-7 and more prevalent in K562 are displayed. The region specified spans the anchors per loop. Annotation is listed alongside the summary statistics of each loop, which includes the number of reads that support each loop per sample. The final column lists all the genes of all promoter regions within 1kb of the loop anchor. The strongest shown in this table represent binary chromatin states between the two cell lines.

**REFERENCES**

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq--a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166-169.

Aryee, M.J., Jaffe, A.E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A.P., Hansen, K.D., and Irizarry, R.A. (2014). Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics *30*, 1363-1369.

Consortium, E.P. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.

Ji, X., Dadon, D.B., Powell, B.E., Fan, Z.P., Borges-Rivera, D., Shachar, S., Weintraub, A.S., Hnisz, D., Pegoraro, G., Lee, T.I.*, et al.* (2016). 3D Chromosome Regulatory Landscape of Human Pluripotent Cells. Cell Stem Cell *18*, 262-275.

Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014). voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol *15*, R29.

Lawrence, M., Huber, W., Pages, H., Aboyoun, P., Carlson, M., Gentleman, R., Morgan, M.T., and Carey, V.J. (2013). Software for computing and annotating genomic ranges. PLoS Comput Biol *9*, e1003118.

Li, G., Chen, Y., Snyder, M.P., and Zhang, M.Q. (2016). ChIA-PET2: a versatile and flexible pipeline for ChIA-PET data analysis. Nucleic Acids Res.

Li, G., Fullwood, M.J., Xu, H., Mulawadi, F.H., Velkov, S., Vega, V., Ariyaratne, P.N., Mohamed, Y.B., Ooi, H.S., Tennakoon, C.*, et al.* (2010). ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. Genome Biol *11*, R22.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol *15*, 550.

Paulsen, J., Rodland, E.A., Holden, L., Holden, M., and Hovig, E. (2014). A statistical model of ChIA-PET data for accurate detection of chromatin 3D interactions. Nucleic Acids Res *42*, e143.

Phanstiel, D.H., Boyle, A.P., Heidari, N., and Snyder, M.P. (2015). Mango: a bias-correcting ChIA-PET analysis pipeline. Bioinformatics *31*, 3092-3098.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139-140.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S.*, et al.* (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A *102*, 15545-15550.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics *25*, 1105-1111.