**Supplementary file 3**

Genome-wide analysis of uracil-DNA pattern comparing to ChIP-seq and DNA accessibility data using either GIGGLE search or the Segway genome segmentation tool.

**Collection of HCT116 related ChIP-seq and DNA accessibility data.**

To find colocalizing binding factors and other genomic features, first a HCT116 specific or relevant set of data were collected. On the one hand, from Cistrome database (<http://cistrome.org/db/#/>, (Mei et al., 2017), data reflect the state of 17 July 2019), overall, 542 ChIP-seq data made in HCT116 for transcription factors or histone markers were downloaded as interval (bed) files. Although, these data are still heterogeneous regarding quality and the applied treatments, it is definitely more reasonable than searching in the whole Cistrome database without any restriction for cell types. Only those data were comprised that contained more than 400 intervals (471 files remained). It also has to be noted that no controls (such as input samples in our case) are considered in the evaluation pipeline of Cistrome. To strengthen this dataset, further HCT116-specific ChIP-seq data (36 bed files) were downloaded from ENCODE (<https://www.encodeproject.org/search/?type=Experiment&status=released&replicates.library.biosample.donor.organism.scientific_name=Homo+sapiens&assembly=GRCh38&biosample_ontology.classification=cell+line&biosample_ontology.term_name=HCT116> (ENCODE Project Consortium, 2012)), where evaluation pipeline (<https://www.encodeproject.org/pages/pipelines/>) considers controls and includes many quality measures, and the resulted “replicated peaks” reflect only the consensus peaks of replicates and pseudo-replicates. Further 27 bed files were downloaded from the Ensembl database ((Zerbino et al., 2018), release 97, July 2019, <ftp://ftp.ensembl.org/pub/release-97/regulation/homo_sapiens/Peaks/HCT116/>). Moreover, for colorectal tissues, comprehensive epigenomic data focusing on five core histone marks (H3K4me3, H3K4me1, H3K27me3, H3K9me3, and H3K36me3) were constructed by Roadmap Epigenomics (<https://egg2.wustl.edu/roadmap/web_portal/processed_data.html>, (Kundaje et al., 2015)). From these data, overall 40 bed files (broad- and gappedPeaks) corresponding to E075 Colonic Mucosa (7 experiments); E076: Colon Smooth Muscle (7 experiments); and E106: Sigmoid Colon (6 experiments)) were also integrated to our dataset for comparison with U-DNA-Seq data. These files were originally aligned to the hg19 reference genome, therefore liftOver (<https://github.com/ucscGenomeBrowser/kent>) was applied to convert the coordinates to hg38 as follows.

$ liftOver NAME\_of\_DB\_intervals.bed hg19ToHg38.over.chain NAME\_ofDB\_intervals\_hg38.bed unMapped

*# The file hg19ToHg38.over.chain can be downloaded from the UCSC (*[*http://hgdownload.soe.ucsc.edu/goldenPath/hg19/liftOver/hg19ToHg38.over.chain.gz*](http://hgdownload.soe.ucsc.edu/goldenPath/hg19/liftOver/hg19ToHg38.over.chain.gz)*).*

Finally, to address the published centromeric localization of uracil (Shu et al., 2018), ChIP-seq data on CENPA in HuRef cells were also downloaded (GSM1105684, bw file (Hayden et al., 2013)). Note that CENPA data are not available for HCT116 cells. In this study, reads were aligned to human reference genome hg19, and enrichment was given in bedgraph format containing only the alpha satellite segments. Following the paper (Hayden et al., 2013), those data bins were selected that showed at least two-fold enrichment, and then bed files were generated (using the same procedure that we used for derivation of region interval files from the log2 coverage tracks). Then liftOver was applied to convert hg19 coordinates to hg38 (GSM1105684\_2fold\_enriched\_merged\_hg38.bed). Data from the same experiment appear also in the Cistrome database (40153), where data were simply realigned to hg38 reference genome, although the original paper reported much more careful procedure on mapping reads in the highly repetitive centromeres (Hayden et al., 2013). Other CENPA data in Cistrome database represent results on ectopically expressed CENPA, outside of the centromeres.

**GIGGLE search with U-DNA-Seq regions in the established HCT116 related ChIP-seq dataset.**

After neglecting files that contain less than 400 intervals, overall, 576 bed files remained in the dataset. Because of the limitation of the applied GIGGLE search tool ((Layer et al., 2018) version 1.0, cf. issue #46 [https://github.com/ryanlayer/GIGGLE/issues/46](https://github.com/ryanlayer/giggle/issues/46)), top 100,000 intervals were selected from those interval (bed) files that contained more than 100,000 intervals (overall 50 bed files were cut in this way).

Then GIGGLE search was performed on this set of relevant and good quality data with all the U-DNA-Seq samples corresponding to HCT116 cells. A digestion of the results is shown in Figure 4A, while the whole set of the combo scores is provided in Supplementary file 3-table 1. The GIGGLE search was done as follows:

*# The database interval files as well as the query interval files have to be sorted and gzipped using a script belongs to the GIGGLE package using also bgzip tool that has to be installed in advance (*[*https://github.com/samtools/htslib/releases/*](https://github.com/samtools/htslib/releases/)*, htslib-1.9.tar.bz2).*

$ {PATH}/GIGGLE/scripts/sort\_bed "{PATH}/\*.bed" bed\_sorted 4

$ {PATH}/GIGGLE/scripts/sort\_bed "{PATH}/\*.region.bed" own\_bed\_sorted 4

*# Indexing the database*

$ GIGGLE index -i "bed\_sorted/\*gz" -o bed\_sorted\_b -f -s

# *Running GIGGLE search in this indexed library*

$ GIGGLE search -i bed\_sorted\_b –q own\_bed\_sorted/NAME.filtered\_blacklisted.bin100bp.smooth5k.RPGC.log2.0p2.region.bed.gz -s > NAME.log2.0p2.regions.GIGGLE\_results.csv

Note: In some cases, GIGGLE search might end in an error message “too many open files”. To solve this problem, a soft limit (-Sn) of the possible open files has to be checked and changed on the linux operating system (it is possible to do up to the hard limit (-Hn)).

*# Checking:*

$ ulimit -Sn

$ ulimit -Hn

*# Changing:*

$ ulimit -Sn 4096

ChIP-seq factors corresponding to the 10 best correlating hits for each U-DNA-Seq sample were selected. GIGGLE scores between all seven samples and all experiments corresponding to these factors were plotted excluding those, where data were found not informative (CNOT3, H2B, H3K27me1/2, SKP2, SIRT1, MCM2, and H4K20me1). All data are provided here in Supplementary file 3-table 1.

Supplementary file 3-table 1. Combo scores from GIGGLE search on the full dataset. 



**ChIP-seq for H3K36me3 histone marker in non-treated and RTX treated UGI-expressing HCT116 cells.**

To check, if drug treatment affects the histone marker distribution, a ChIP-seq experiment for H3K36me3 marker was performed on non-treated and RTX treated HCT116 cells that were also expressing UGI (cf. Materials and Methods). The pre-processing of the raw data was performed in the same way as in case of the U-DNA-Seq applying the HCT116 specific blacklist (cf. Supplementary file 1 and Figure 2-figure supplement 2). The statistics on these pre-processing steps are provided here in Supplementary file 3-table 2.

**Supplementary file 3-table 2. Statistics on pre-processing of H3K36me3 ChIP-seq data.**



The peak calling was performed by MACS2 using broad peak option as given below using an input control selected from the ENCODE (ENCFF489VMD).

$ macs2 callpeak -t NAME(IP).filtered\_blacklisted.bam –c ENCFF489VMD.filtered\_blacklisted.bam --broad -g 2830361285 --broad-cutoff 0.05 -n NAME\_ChIPvsCtr\_0p05 --outdir {PATH} --nomodel -f BAMPE --cutoff-analysis

The fold change tracks were also calculated comparing to an input control as given below.

$ bamCoverage -b NAME.filtered\_blacklisted.bam -o NAME.bin100.RPGC.bw --binSize 100 --verbose --normalizeUsing RPGC --effectiveGenomeSize 2792809237 -p 32 --extendReads

$ bigwigCompare -b1 NAME.bin100.RPGC.bw -b2 ENCFF489VMD.filtered\_blacklisted.bin100.RPGC.bw –o NAME\_foldChange.bin100bp.RPGC.ratio.bw -of bigwig --binSize 100 -v -p 32 --skipNAs --operation ratio

These processed files as well as the corresponding raw data were uploaded to the GEO (GSE153407, that is combined with the U-DNA-Seq data under GSE153408). These were compared to each other as well as to the H3K36me3 ChIP-seq foldChange tracks downloaded from the ENCODE using multiBigWigSummary (Figure 4-figure supplement 1).

$ multiBigwigSummary bins -b ENCFF238GBP.bigWig ENCFF334KFI.bigWig ENCFF514ZYW.bigWig NT\_UGI\_H3K36me3\_ChIP\_foldChange.bin100bp.RPGC.ratio.bw RTX\_UGI\_H3K36me3\_ChIP\_foldChange.bin100bp.RPGC.ratio.bw -o mbws\_Fig4FigSup1\_H3K36me3ChIPseq.npz -v -p 32

$ plotCorrelation --corData mbws\_Fig4FigSup1\_H3K36me3ChIPseq.npz --corMethod pearson --whatToPlot heatmap -o mbws\_Fig4FigSup1\_H3K36me3ChIPseq.png -T mbws\_Fig4FigSup1\_H3K36me3ChIPseq --skipZeros --removeOutliers --plotNumbers --colorMap RdPu

GIGGLE scores between U-DNA-seq results and these H3K36me3 ChIP-seq peaks were calculated as it was described for other ChIP-seq data above (Figure 4A, Figure 4-source data 1).

The fold change track files were also included in the Segway analysis (Figure 4B, Figure 4-source data 2).

**Genome segmentation analysis of U-DNA-Seq results and ChIP-seq data from the ENCODE using Segway genome segmentation algorithm.**

22 independent ChIP-seq experiments available for HCT116 cells in the ENCODE database were selected for this type of analysis (see Supplementary file 3-table 3). The corresponding ‘fold change over control’ tracks were combined with our ChIP-seq data on H3K36me3 as well as with our U-DNA-Seq data to a single genomedata file using Genomedata software (Hoffman, Buske, & Noble, 2010). In this analysis, merged ratio tracks of U-DNA enrichment were used that were calculated by bigwigCompare of the deepTools package (Ramírez et al., 2016) in the format of bw files and converted to BedGraph format. File conversions were done using BigWigToWig tool (Kuhn, Haussler, & Kent, 2013). The genome segmentation analysis was confined to the core chromosomes (1 to 22, X, and Y) of GRCh38 reference genome, and the regions of our HCT116 specific blacklist were also excluded. The applied command lines are provided below.

File conversion and filtration in case of bigWig files from the ENCODE database:

$ bigWigToWig NAME.bigwig NAME.wig

$ awk '{ if ($1 == "chr1" || $1 == "chr2" || $1 == "chr3" || $1 == "chr4" || $1 == "chr5" || $1 == "chr6" || $1 == "chr7" || $1 == "chr8" || $1 == "chr9" || $1 == "chr10" || $1 == "chr11" || $1 == "chr12" || $1 == "chr13" || $1 == "chr14" || $1 == "chr15" || $1 == "chr16" || $1 == "chr17" || $1 == "chr18" || $1 == "chr19" || $1 == "chr20" || $1 == "chr21" || $1 == "chr22" || $1 == "chrX" || $1 == "chrY") {print $0} }' NAME.wig > NAME.filtered.wig

File conversion and filtration in case of bw files of U-DNA-Seq ratio tracks:

$ bigWigToWig NAME.bw NAME.bdg

$ awk '{ if ($1 == "chr1" || $1 == "chr2" || $1 == "chr3" || $1 == "chr4" || $1 == "chr5" || $1 == "chr6" || $1 == "chr7" || $1 == "chr8" || $1 == "chr9" || $1 == "chr10" || $1 == "chr11" || $1 == "chr12" || $1 == "chr13" || $1 == "chr14" || $1 == "chr15" || $1 == "chr16" || $1 == "chr17" || $1 == "chr18" || $1 == "chr19" || $1 == "chr20" || $1 == "chr21" || $1 == "chr22" || $1 == "chrX" || $1 == "chrY") {print $0} }' NAME.bdg > NAME.filtered.bdg

Preparing genomedata file using genomedata-load (Hoffman et al., 2010):

$ genomedata-load -t {SAMPLE1}={FILE1} -t {SAMPLE2}={FILE2} … -s core\_regions.fna mergedload

The genome data was trained by the segway train (Chan et al., 2017; Hoffman et al., 2012) where the number of labels (types of characteristic patterns) was defined as 25, also providing the same blacklist that was used for the U-DNA-Seq data.

$ SEGWAY\_CLUSTER=local segway train mergedload --resolution=100 --num-labels=25 --minibatch-fraction=0.01 --exclude-coords=blacklist\_merged\_d500.bed train25\_merged\_blacklisted

Then genomic segments with these labels were identified with segway annotate (Chan et al., 2017; Hoffman et al., 2012):

$ SEGWAY\_CLUSTER=local segway annotate mergedload train25\_merged\_blacklisted Segway\_annot\_out

The signal distribution was calculated using Segtools (Buske, Hoffman, Ponts, Le Roch, & Noble, 2011) and was plotted by Seaborn (MatPlotLib modul in python, (Hunter, 2007)) (cf. Figure 4B).

$ segtools-signal-distribution segway.bed mergedload

import pandas as pd

import seaborn as sb

df = pd.read\_csv("Segway norm\_data.txt", delimiter="\t")

df = df.set\_index(df.columns[0])

s=sb.heatmap(df,xticklabels=True,yticklabels=True,linewidths=0.5,linecolor="black",cmap="RdYlBu\_r",annot=False,annot\_kws={'size':4})

figure = s.get\_figure()

figure.savefig('plot\_v.png', dpi=400)

exit()

**Supplementary file 3-table 3: The composition of the genomedata file**.

|  |  |
| --- | --- |
| {LABEL} | {FILE} |
| WT | WT\_HCT116\_merged\_IP\_vs\_son.bin100bp.smooth5000.RPGC.ratio.filtered.bdg |
| NT\_UGI | NT\_UGI\_HCT116\_merged\_IP\_vs\_son.bin100bp.smooth5000.RPGC.ratio.filtered.bdg |
| NT\_UGI\_MMR | NT\_UGI\_HCT116MMR\_merged\_IP\_vs\_son.bin100bp.smooth5000.RPGC.ratio.filtered.bdg |
| 5FdUR\_UGI | 5FdUR\_UGI\_HCT116\_merged\_IP\_vs\_son.bin100bp.smooth5000.RPGC.ratio.filtered.bdg |
| 5FdUR\_UGI\_MMR | 5FdUR\_UGI\_HCT116MMR\_merged\_IP\_vs\_son.bin100bp.smooth5000.RPGC.ratio.filtered.bdg |
| RTX\_UGI | RTX\_UGI\_HCT116\_merged\_IP\_vs\_son.bin100bp.smooth5000.RPGC.ratio.filtered.bdg |
| RTX\_UGI\_MMR | RTX\_UGI\_HCT116MMR\_merged\_IP\_vs\_son.bin100bp.smooth5000.RPGC.ratio.filtered.bdg |
| EZH2\_pT487 | ENCFF011JWE.filtered.wig |
| H3K79me2 | ENCFF127XQD.filtered.wig |
| H3K4me3 | ENCFF144ZRX.filtered.wig |
| ZFX | ENCFF168KEG.filtered.wig |
| H3K9ac | ENCFF187IFT.filtered.wig |
| H2AFZ | ENCFF236ZOJ.filtered.wig |
| TCF7L2 | ENCFF241JHM.filtered.wig |
| YY1 | ENCFF293WBY.filtered.wig |
| JUND | ENCFF415KJH.filtered.wig |
| H4K20me1 | ENCFF431PNC.filtered.wig |
| H3K36me3 | ENCFF514ZYW.filtered.wig |
| H3K9me3 | ENCFF542HPZ.filtered.wig |
| EZH2 | ENCFF604ZUW.filtered.wig |
| ZBTB33 | ENCFF616IIZ.filtered.wig |
| CTFC | ENCFF620LDT.filtered.wig |
| H3K4me1 | ENCFF774BWO.filtered.wig |
| H3K4me2 | ENCFF783QRO.filtered.wig |
| RNApol2A\_pS5 | ENCFF794LVU.filtered.wig |
| H3K9me2 | ENCFF807PSQ.filtered.wig |
| RNApol2A | ENCFF856HUR.filtered.wig |
| H3K27me3 | ENCFF984BVG.filtered.wig |
| H3K27ac | ENCFF984WLE.filtered.wig |
| NT\_H3K36me3 | NT\_UGI\_H3K36me3\_ChIP\_foldChange.bin100bp.RPGC.ratio.filtered.bdg |
| RTX\_H3K36me3 | RTX\_UGI\_H3K36me3\_ChIP\_foldChange.bin100bp.RPGC.ratio.filtered.bdg |

References

Buske, O. J., Hoffman, M. M., Ponts, N., Le Roch, K. G., & Noble, W. S. (2011). Exploratory analysis of genomic segmentations with Segtools. *BMC Bioinformatics*, *12*(1), 415. https://doi.org/10.1186/1471-2105-12-415

Chan, R. C. W., Libbrecht, M. W., Roberts, E. G., Bilmes, J. A., Noble, W. S., & Hoffman, M. M. (2017). Segway 2.0: Gaussian mixture models and minibatch training. *Bioinformatics*, *34*(4), 669–671. https://doi.org/10.1093/bioinformatics/btx603

ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. https://doi.org/10.1038/nature11247

Hayden, K. E., Strome, E. D., Merrett, S. L., Lee, H.-R., Rudd, M. K., & Willard, H. F. (2013). Sequences Associated with Centromere Competency in the Human Genome. *Molecular and Cellular Biology*, *33*(4), 763–772. https://doi.org/10.1128/MCB.01198-12

Hoffman, M. M., Buske, O. J., & Noble, W. S. (2010). The Genomedata format for storing large-scale functional genomics data. *Bioinformatics (Oxford, England)*, *26*(11), 1458–1459. https://doi.org/10.1093/bioinformatics/btq164

Hoffman, M. M., Buske, O. J., Wang, J., Weng, Z., Bilmes, J. A., & Noble, W. S. (2012). Unsupervised pattern discovery in human chromatin structure through genomic segmentation. *Nature Methods*, *9*(5), 473–476. https://doi.org/10.1038/nmeth.1937

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95.

Kuhn, R. M., Haussler, D., & Kent, W. J. (2013). The UCSC genome browser and associated tools. *Briefings in Bioinformatics*, *14*(2), 144–161. https://doi.org/10.1093/bib/bbs038

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., … Kellis, M. (2015). Integrative analysis of 111 reference human epigenomes. *Nature*, *518*(7539), 317–330. https://doi.org/10.1038/nature14248

Layer, R. M., Pedersen, B. S., DiSera, T., Marth, G. T., Gertz, J., & Quinlan, A. R. (2018). GIGGLE: a search engine for large-scale integrated genome analysis. *Nature Methods*, *15*(2), 123–126. https://doi.org/10.1038/nmeth.4556

Mei, S., Qin, Q., Wu, Q., Sun, H., Zheng, R., Zang, C., … Liu, X. S. (2017). Cistrome Data Browser: a data portal for ChIP-Seq and chromatin accessibility data in human and mouse. *Nucleic Acids Research*, *45*(D1), D658–D662. https://doi.org/10.1093/nar/gkw983

Ramírez, F., Ryan, D. P., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A. S., … Manke, T. (2016). deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research*, *44*(W1), W160-5. https://doi.org/10.1093/nar/gkw257

Shu, X., Liu, M., Lu, Z., Zhu, C., Meng, H., Huang, S., … Yi, C. (2018). Genome-wide mapping reveals that deoxyuridine is enriched in the human centromeric DNA. *Nature Chemical Biology*, *14*(7), 680–687. https://doi.org/10.1038/s41589-018-0065-9

Zerbino, D. R., Achuthan, P., Akanni, W., Amode, M. R., Barrell, D., Bhai, J., … Flicek, P. (2018). Ensembl 2018. *Nucleic Acids Research*, *46*(D1), D754–D761. https://doi.org/10.1093/nar/gkx1098