Outline of the materials and methods for HIV project

1) DNAse-seq

Files were processed with a modified Nextflow (Tommaso et al., 2017), BICF ChiP-seq Analysis Workflow version 1.0.0 (Barnes et al., 2019). Briefly, we used trimgalore version 0.4.1 (Martin, 2011) on the raw reads to remove reads shorter than 35bp or with phred quality scores less than 20 and then aligned trimmed reads to the human reference genome (GRCh38) using default parameters in BWA samse version 0.7.12 (Li and Durbin, 2009). The aligned reads were subsequently filtered for quality and uniquely, mappable reads were retained for further analysis using SAMtools version 1.3 (Li et al., 2009) and Sambamba version 0.6.6 (Tarasov et al., 2015). Relaxed peaks were called using MACS version 2.1.0 (Zhang et al., 2008) with the following parameters: -p 1e-2 --nomodel --shift -100 --extsize 200 --keep-dup all -B --SPMR. Peaks that overlap at least 50% between replicates were retained.

2) MNase-seq

Files were processed with a modified Nextflow (Tommaso et al., 2017), BICF ChiP-seq Analysis Workflow version 1.0.0 (Barnes et al., 2019). Briefly, we used trimgalore version 0.4.1 (Martin, 2011) on the raw reads to remove reads shorter than 35bp or with phred quality scores less than 20 and then aligned trimmed reads to the human reference genome (GRCh38) using default parameters in BWA sampe version 0.7.12 (Li and Durbin, 2009). The aligned reads were subsequently filtered for quality and uniquely mappable reads were retained for further analysis using SAMtools version 1.3 (Li et al., 2009) and Sambamba version 0.6.6 (Tarasov et al., 2015), and bedtools version 2.26.0 (Quinlan and Hall, 2010) bamtobed converted the bed file to tagalign. Peaks were called with iNPS version 1.2.2 (Chen et al., 2014) and filtered for a -log10(Pvalue\_of\_peak) of less than 0.05.

3) Chip-seq

The Nextflow (Tommaso et al., 2017) BICF ChiP-seq Analysis Workflow version 1.0.0 (Barnes et al., 2019) proccessed all ChiP-seq files, merging separate experiments as technical replicates. Briefly, reads were trimmed with trimgalore version 0.4.1 (Martin, 2011) (parameters: -q 25 --illumina --gzip --length 35), aligned with bwa aln (-q 5 -l 32 -k 2) and then bwa samse (standard parameters) version 0.7.12 (Li and Durbin, 2009), sorted and indexed with SAMtools version 1.3 (Li et al., 2009)(-F 1804 -q 30), and duplicates removed with Sambamba version 0.6.6 (Tarasov et al., 2015) (standard options). Bam files were converted to tagAlign with bedtools version 2.26.0 bamtobed (Quinlan and Hall, 2010), after which, samples were checked for quality control using deeptools version 2.5.0.1 multiBamSummary, plotCorrelation, plotCoverage, and plotFingerprint (all standard protocols) (Ramírez et al., 2016) and cross correlation analysis with phantompeakqualtools version 1.2 (Kharchenko et al., 2008 and Landt et al., 2012). Peaks were called with MACS2 version 2.1.0-20151222 (Zhang et al., 2008), using the predominant fragment length from the cross correlation analysis as --extsize (other parameters: -p 1e-2 --nomodel --shift 0 --keep-dup all -B --SPMR). Consensus peaks were called (bedtools version 2.26.0; Quinlan and Hall, 2010) and annotated (library ChIPseeker in R; Yu et al., 2015 and R Core Team, 2014) if at least 2 replicates or pseudoreplicates contained a peak.

4) RNAseq

Files were processed with the BICF RNA-seq Analysis workflow version 0.5.5. Briefly, reads with phred quality scores less than 20 and less than 35 bp after trimming were removed from further analysis using trimgalore version 0.4.1 (Martin, 2011). Quality-filtered reads were then aligned to the human reference genome (GRCh38) using the HISAT version 2.0.1 (Pertea et al., 2016) aligner using default settings and marked duplicates using Sambamba version 0.6.6 (Tarasov et al., 2015). Aligned reads were quantified to coding sequences of known transcripts using ‘featurecount’ version 1.4.6 (Liao et al., 2014) per gene ID against Gencode version 25.

5) TTseq

Files were processed a modified version of the BICF RNA-seq Analysis workflow version 0.5.5. Briefly, reads with phred quality scores less than 20 and less than 35 bp after trimming were removed from further analysis using trimgalore version 0.4.1 (Martin, 2011). Quality-filtered reads were then aligned to the human reference genome (GRCh38) using the HISAT version 2.0.1 (Pertea et al., 2016) aligner using default settings and marked duplicates using Sambamba version 0.6.6 (Tarasov et al., 2015). Aligned reads were quantified to the entire annotated transcript region using ‘featurecount’ version 1.4.6 (Liao et al., 2014) per gene ID against Gencode version 25.

6) Hi-C

Reads were pooled per library an ran through the standard Hi-C pipeline using homer version 4.10.4 (Heinz et al., 2010). Briefly, reads were trimmed with homerTools trim -3 GATC -mis 0 -matchStart 20 -min 20, mapped to human reference genome (GRCh38, canonical) with bowtie2 version 2.2.8 (Langmead and Salzberg, 2012), and converted to tag directory (makeTagDirectory -genome hg38 -checkGC -restrictonSite GATC). Matrices are normalized with analyzeHiC (standard protocol), and tads/loops identified (findTADsAndLoops.pl -res 3000 -window 15000 -genome hg38 -p badRegions.bed). The USCS gap and genomicSuperDups were combined to make badRegions file. Tads and Loops from the 2 libraries were combined with merge2Dbed.pl

7) DamID-seq

Reads were trimmed with cutadapt version 1.9.1 (Martin, 2011), attaching GATC to the adapter sequence (-a GATCCTCGGCCGCGACC -g ^GGTCGCGGCCGAGGATC), and mapped with bowtie2 version 2.3.2 (Langmead and Salzberg, 2012). Sam files were converted to bam and sorted with SAMtools version 1.3 (Li et al., 2009), and coverted to bed format with bedtools version 2.26.0 bamToBed (Quinlan and Hall, 2010). Reads that did not cross over a GATC site were kept (custom script) and filtered again if they touch a GATC site by extending the GATC site by 1 bp in each direction and using bedtools intersect version 2.26.0 (Quinlan and Hall, 2010). Broad peaks were called with MACS2 version 2.1.0-20151222 (-g hs --keep-dup all -bw 300 -qvalue 0.05 -mfold 5 50 --broad --broad-cutoff 0.1) (Zhang et al., 2008) and filtered for a qvalue less than 0.05 and a signalValue greater than 2.

8) BHIVE expression and integration

BHIVE data was processed with the B-HIVE for single provirus transcriptomics docker container (https://github.com/gui11aume/BHIVE\_for\_single\_provirus\_transcriptomics).

9) chromHMM with extra files

We implemented ChromHMM (Ernst et al. 2012), which uses a multivariate Hidden Markov Model (HMM) to calculate the probabilistic nature of a multistate model and the biological nature of the state of chromatin at that location, in order to discover chromatin states in Jurkat cells using epigenomics information derived from 7 individual ChIP-seq marks (H3K27me3, H3K4me3, H3K27ac, H3K4me1, H3K36me3, H3K79me3, and H3K9me3) known as state emissions. Filtered bam files, with mapping read duplicates removed, for each of the 7 histones, were individually converted to binary bin files with chromHMM BinarizeBam V1.19 (Ernst et al., 2012). We first obtained the state emissions of 15 different chromatin states defined as described on the Roadmap epigenomics project (<https://egg2.wustl.edu/roadmap/web_portal/chr_state_learning.html>) on the basis of the observed data for the above 7 histone modifications. We used a core 15-state model for our analyses since it captured all the key interactions between the chromatin marks, and because larger numbers of states (e.g., “expanded 18-state model”) did not apparently capture sufficiently distinct interactions. To *de novo* generate the core 15-state model in Jurkat cells we compared the relative abundance of the state emissions in Jurkat cells with known chromatin states for the 3 ENCODE cell lines most genetically and phenotypically linked with Jurkats (**E115**: Dnd41 T cell leukaemia, **E116**: GM1282878 lymphoblastoid, **E123**: K562 T cell leukaemia) (as shown in Figure X). In order to assign biologically meaningful mnemonics to the 15 chromatin states, we used the ChromHMM package to compute the overlap and neighborhood enrichments of each chromatin state relative to various types of functional annotations including the ChromHMM buildt in refseq annotations of: 1) CpG islands, 2) genes, 3)exons, 4)introns, 5) transcription start sites (TSS), 2-Kb windows around TSS (TSS flanking), transcription termination sites (TTS), and 2-Kb windows around TTS (TTS flanking) based on the GENCODE v27 annotation, 3) Zinc finger genes (ZNF genes) obtained from XXXX, and 4) typical enhancers and super enhancers obtained as described in the “Enhancer” section below or above.

10) Enhancers/SuperEnhancers

TTseq contains information on transient transcription which we used to identify possible enhancers and super enhancers. TT-seq also allows the transcription to be identified as the forward or reverse strand. For this purpose, enhancers are defined as regions of the genome that are bi-directionally transcribed, and not in a gene or its promoter or an annotated linc-RNA. A 4 state hidden markov model on TT-seq all data (not separated by strand), TT-seq forward strand, and TT-seq reverse strand identified regions of the genome that are actively transcribed. Four states were chosen over a 2 state model because there were various amounts of transcription found in the genome; genic regions were easily identified, but low expression intergenic regions could not be identified with 2 states. Thus, 3 of the 4 states were coded for transcribed, and 1 state was labeled as untranscribed. A database of possible enhancers was created by identifing regions of the genome that were identified as transcribed for all data, and also overlapped with regions that were forward and reverse transcribed. Also added to the possible enhancer list were regions where there were overalpping forward and reverse transcription, but not identified as transcribed in all data. From this list, protein coding genes with 2 kb upstream and downstream were removed. Next, annotated regions from RNAseq with an fpkm greater than 1 for all replicates and not protein-coding were removed. This final list of 20,943 regions are purported enhancers.

Super enhancers were identified using Rose v0.1 (Warren et al., 2013 and Loven et al., 2013), stitching together a 12,500 bp distatnce, exluding 2500 bp from TSS. Purported enhancer regions from TT-seq were used as previouly identified enhancer regions. Merged, filtered, and read mapped duplicates removed bam files of histone marks, H3K27ac, H3K4me3, and H3K4me1 were used to rank the possible enhancers. Since not all enhancers are identifiable with the 3 histone marks, TT-seq bam files were also used to identify super enhancers, and the output was filtered again for transcribed, annotated regions of the genome (E.G. genes and linc-RNAs). The final super enhancer database contained 767 merged regions, of which 360 were identified with H3K27ac alone, 436 identified with H3K4me3 alone, 301 identified with H3K4me1 alone, and 115 identified with TT-seq alone.

The 767 super enhancer regions were removed from the 20,943 purported enhancer regions leaving 18,357 possible enhancer regions. Of these regions, 701 overlap with H3K27ac peaks, 262 overlap with H3K4me3 peaks, 702 overlap with H3K4me1 peaks, and 1301 overlap with forward and reverse TT-seq transcribed regions with bidirectional transcription. The final merged regions contain 2180 enhancers.

11) machine learning

In order to study the immediate landscape surround HIV insertion and its possible effects on expression, we looked at 2kb region, in 200 bp increments around the sites. Rpkm values of each 200 bp bin (20 bins in total) were calculated for the 7 histone marks (H3K27ac, H3K4me3, H3K4me1, H3K36me3, H3K79me3, H3K9me3, and H3K27me3), RNA-seq, MNase-seq, DNase-seq, and TT-seq using RPKM.py (reference ???). Discrete values for chromHMM states and lamin subcompartment states were also noted for each 200 bp region. Jeon fill this in?????

References:

Bailey T. L., M. Bodén, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble. 2009. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Research 37: W202-W208. doi:10.1093/nar/gkp335

Spencer D. Barnes, Holly Ruess, Jeremy A. Mathews, Beibei Chen, & Venkat S. Malladi. (2019, May 16). BICF ChIP-seq Analysis Workflow (Version publish\_1.0.6). Zenodo. http://doi.org/10.5281/zenodo.3238425

The ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. Nature 489(7414):57-74. doi: 10.1038/nature11247.

Ewels P., Magnusson M., Lundin S. and Käller M. 2016. MultiQC: Summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32(19): 3047–3048. doi:10.1093/bioinformatics/btw354

Heinz S, Benner C, Spann N, Bertolino E et al. Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol Cell 2010 May 28;38(4):576-589.

Hnisz et al., 2016 Science. 2016 Mar 25;351(6280):1454-1458. doi: 10.1126/science.aad9024

Langmead B, Salzberg S. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012, 9:357-359.

Liao Y1, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30:923-30. doi: 10.1093/bioinformatics/btt656

Rasko Leinonen,1,\* Hideaki Sugawara,2 and Martin Shumway3, on behalf of the International Nucleotide Sequence Database Collaboration. 2011. The Sequence Read Archive. Nucleic Acids Res. 39: D19–D21.

Kent W. J., A. S. Zweig, G. Barber, A. S. Hinrichs, and D. Karolchik. BigWig and BigBed: enabling browsing of large distributed data sets. Bioinformatics 26(17): 2204-2207. doi:10.1093/bioinformatics/btq351

Kharchenko P. K., M. Y. Tolstorukov, and P. J. Park. 2008. Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol 26(12): 1351-1359. doi:10.1038/nbt.1508

Landt S. G., G. K. Marinov, A. Kundaje, et al. 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome Res 9: 1813-31. doi:10.1101/gr.136184.111

Li H., and R. Durbin. 2009. Fast and accurate short read alignment with Burrows-Wheeler Transform. Bioinformatics 25: 1754-60. doi:10.1093/bioinformatics/btp324

Li H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. The Sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25: 2078-9. doi:10.1093/bioinformatics/btp352

Machanick P., and T. L. Bailey. 2011. MEME-ChIP: motif analysis of large DNA datasets. Bioinformatics 27(12): 1696-1697. doi:10.1093/bioinformatics/btr189

Mansour et al., 2014. Science. 2014 Dec 12;346(6215):1373-7. doi: 10.1126/science.1259037

Marcel, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17(1):10-12. doi:10.14806/ej.17.1.200

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet. journal, 17(1), pp-10.

Mihaela Pertea, Daehwan Kim, Geo M Pertea, Jeffrey T Leek & Steven L Salzberg. 2016. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nature Protocols 11: 1650–1667.

Quinlan, A. R., and I. M. Hall. 2010. BEDTools: a flexible suite of utilities for comparing genomic feautures. Bioinformatics 26(6): 841-842. doi:10.1093/bioinformatics/btq033

R Core Team 2014. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL:http://www.R-project.org/.

Ramírez, F., D. P. Ryan, B. Grüning, V. Bhardwaj, F. Kilpert, A. S. Richter, S. Heyne, F. Dündar, and T. Manke. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. Nucleic Acids Research 44: W160-165. doi:10.1093/nar/gkw257

Reeder, J. E., Y-T Kwak, R. P. McNamara, C. V. Forst, I. D'Orso. 2015. HIV Tat controls RNA Polymerase II and the epigenetic landscape to transcriptionally reprogram target immune cells. eLife 2015;4:e08955 doi: 10.7554/eLife.08955.

Ross-Innes C. S., R. Stark, A. E. Teschendorff, K. A. Holmes, H. R. Ali, M. J. Dunning, G. D. Brown, O. Gojis, I. O. Ellis, A. R. Green, S. Ali, S. Chin, C. Palmieri, C. Caldas, and J. S. Carroll. 2012. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. Nature 481: 389-393. doi:10.1038/nature10730

Stark R., and G. Brown. 2011. DiffBind: differential binding analysis of ChIP-Seq peak data. http://bioconductor.org/packages/release/bioc/vignettes/DiffBind/inst/doc/DiffBind.pdf. doi:10.18129/B9.bioc.DiffBind

Tarasov, A., A. J. Vilella, E. Cuppen, I. J. Nijman, and P. Prins. 2015 Sambamba: fast processing of NGS alignment formats. Bioinformatics 31(12): 2032-2034. doi:10.1093/bioinformatics/btv098

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., and Notredame, C. 2017. Nextflow enables reproducible computational workflows. Nature biotechnology, 35(4), 316.

Weizhong Chen, Yi Liu, Shanshan Zhu, Christopher D. Green, Gang Wei & Jing-Dong Jackie Han. 2014. Improved nucleosome-positioning algorithm iNPS for accurate nucleosome positioning from sequencing data. Nature 5: 4909.

Yu G., L. Wang, and Q. He. 2015. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. Bioinformatics 31(14): 2382-2383. doi:10.1093/bioinformatics/btv145

Zhang Y., T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu. 2008. Model-based Analysis of ChIP-Seq (MACS). Genome Biol 9: R137. doi:10.1186/gb-2008-9-9-r137

Anaconda (Anaconda Software Distribution, https://anaconda.com)