**Script organization on Github**

Date: September 3, 2020

***UV Lesion Script organization***

1. Binning UV lesion signal scripts
   1. UV lesion binning (100kb and 1Mb) script
   2. Rank normalization of UV lesion signal script
   3. UV lesion binning per chromatin state script
   4. UV lesion binning per gene
   5. UV lesion binning per enhancer

*Input files:*

1. IP64100vInput.fc.signal.bw and CPD100vInput.fc.signal.bw (Bigwig files)
2. UVlesion\_mastertable\_1Mb.csv
3. UVlesion\_mastertable\_100kb.csv
4. E017\_15\_coreMarks\_dense.bed.bgz (*from Roadmap*)
5. IMR90\_enhancer\_list\_EnhancerAtlas2.0.bed (*from EnhancerAtlas2.0 website*)

*Output files:*

1. 1Mb and 100kb binned files
2. UV lesions rank normalized to norm dist 1Mb.RData (**Used in other scripts**)
3. UV lesions rank normalized to norm dist 100kb.RData (**Used in other scripts**)
4. IP64\_binned\_to\_chrom15.RData (**Used in other scripts**)
5. CPD\_binned\_to\_chrom15.RData (**Used in other scripts**)
6. IP64\_FC\_per\_prot\_coding\_gene.RData (**Used in other scripts**)
7. CPD\_FC\_per\_prot\_coding\_gene.RData (**Used in other scripts**)
8. IP64\_FC\_per\_Enh.RData (**Used in other scripts**)
9. CPD\_FC\_per\_Enh.RData (**Used in other scripts**)
10. UV lesion sequencing analyses
    1. IP 6-4PP replicate correlation script
    2. Di-pyrimidine frequency in raw sequencing reads script
    3. Di-pyrimidine frequency boxplot script

***Genomic and Epigenetic Feature Script organization***

1. Genomic feature analyses
   1. RepeatMasker pre-processing script
   2. Enhancer target gene for IMR90 dataset EnhancerAtlas
   3. Genic feature enrichment in top UV susc regions scripts
   4. Enhancer element enrichment in top UV susc regions script
   5. Super enhancer enrichment in top UV susc regions script
   6. Genic feature and enhancer element enrichment boxplot script
   7. Repeat class enrichment in top UV susc regions script
   8. Repeat class boxplot script
   9. EnrichR - genes in top regions script
   10. EnrichR - enhancer target genes in top regions script
   11. EnrichR - genes with top UV ratio scores script
   12. EnrichR - enh target genes with top UV ratio scores script

*Input files:*

1. hg19.database.rmsk.EXTRA.COLS.csv (*From RepeatMasker website*)
2. EnhancerAtlas\_IMR90\_Enh\_gene\_interactions\_v2.0.txt (*from EnhancerAtlas2.0 website*)
3. IMR90\_enhancer\_list\_EnhancerAtlas2.0.bed (*from EnhancerAtlas2.0 website*)
4. Enrich\_function.RData (custom function)
5. UV lesions rank normalized to norm dist 100kb.RData (**Commonly used file**)
6. IMR90\_dbSUPER\_super\_enhancer\_regions.bed (*from dbSUPER website*)

*Output files:*

1. modified.RMSK.hg19.RData (**Used in other scripts in folder**)
2. Repeat Classes with subFamily list.RData (**Used in other scripts in folder**)
3. IMR90\_Enh\_prot\_coding\_gene\_interactions.RData (**Used in other scripts in folder**)
4. Epigenetic feature analyses scripts
   1. Epigenetic feature binning (1Mb bin)
   2. Histone mark & UV lesion correlation - FC and RankNorm script
   3. Euchr and Het scatterplot script
   4. PCA script

*Input files:*

1. Master\_Table\_FC\_1Mb.RData (Table with ALL features)
2. Master\_Table\_FC\_1Mb.No\_RankNorm.RData

*Output files:*

1. All binned histone files
2. Chromatin states analyses
   1. Rank normalize UV lesion signal in chromatin state script
   2. UV lesion signal in chrom state boxplot script
   3. Histone mark binning to chromatin state script
   4. Histone mark heatmap script
   5. Genic feature and enhancer enrich in chrom state script
   6. Genic feature and enhancer heatmap script
   7. Repeat class enrich in chrom state script
   8. Repeat class heatmap
   9. Di-nucleotide frequency of each chromatin state script

*Input files:*

1. CPD\_binned\_to\_chrom15.RData (**from UV Lesion scripts**)
2. IP64\_binned\_to\_chrom15.RData (**from UV Lesion scripts**)
3. E017\_15\_coreMarks\_dense.bed.bgz (*from Roadmap*)
4. IMR90\_enhancer\_list\_EnhancerAtlas2.0.bed (*from EnhancerAtlas2.0 website*)
5. IMR90\_dbSUPER\_super\_enhancer\_regions.bed (*from dbSUPER website*)

*Output files:*

1. Ranked\_UV\_lesions\_chrom\_states.RData
2. Ranked\_CPD\_boxplot\_stats\_for\_chrom\_states.RData (**Used in other scripts in folder**)
3. Ranked\_64PP\_boxplot\_stats\_for\_chrom\_states.RData (**Used in other scripts in folder**)
4. Ranked\_64PP\_CPD\_difference\_boxplot\_stats\_for\_chrom\_states.RData (**Used in other scripts in folder**)
5. E017\_summary\_StateMean\_GenomeMean.csv (**Used in other scripts in folder**)
6. Genic.EnhAltas\_15ChromState\_Enrchment.RData (**Used in other scripts in folder**)
7. RepeatClass\_15ChromState\_Enrichment.RData (**Used in other scripts in folder**)
8. IMR90\_chrom\_states\_with\_DiPy\_freq.RData (**Used in other scripts in folder**)

***3D Genome Modeling Script organization***

1. Whole chromosome radial position scatterplot script
2. 3D genome modeling scripts
   1. Chrom3D workflow batch script
   2. Test for optimal TAD overlap with top region script
   3. IMR TAD ID list for top regions script
   4. Batch script for coloring genome on Sherlock
   5. TAD nuclear distance boxplot script

*Input files:*

1. IMR91\_chrom\_radial\_positions.csv (1kb)
2. UV lesions rank normalized to norm dist.RData (**Commonly used file**)
3. UV lesions rank normalized to norm dist 100kb.RData (**Commonly used file**)
4. IMR90\_inter\_intra\_chr\_w\_LADs.gtrack

*Output files:*

1. TAD\_stats\_with\_variable\_overlap\_100kb.csv

***Melanoma Mutation Script organization***

1. MELA-AU pre-processing scripts
   1. Subset MELA-AU dataset - only essential SSM info script
   2. Extracting donor IDs for cutaneous mela with UV sig script
   3. SSM dataset – cutaneous mela with UV sig only

*Input files:*

1. Donor ID files from melanoma sequencing dataset
   1. Donor\_ID\_list\_cutaneous\_MELA\_with\_UV\_sig.csv
   2. donor.MELA-AU.tsv.gz
2. hg19\_1Mb\_DNA\_stats.RData
3. Whole genome mutation rate scripts
   1. Genome melanoma mutation rate - 100kb and1Mb
   2. Mutation rate vs UV lesion abundance scatterplot script
   3. Mutation rate analysis of top susc. regions

*Input files:*

1. MELA\_AU\_cutaneous\_SSM.RData (**large file size**)

*Output files:*

1. All\_SSM\_rates\_1Mb\_hg19.RData (Output file)
2. All\_SSM\_rates\_100kb\_hg19.RData (Output file)
3. Gene and enhancer mutation rate scripts
   1. Measurement of gene mutation rates
   2. Measurement of enhancer mutation rates
   3. Gene mutation rate analysis of top regions script
   4. Enhancer mutation rate analysis of top regions script

*Input files:*

1. IMR90\_enhancer\_list\_EnhancerAtlas2.0.bed (Input file)
2. UV lesions rank normalized to norm dist 100kb.RData (Input file)
3. Gene\_list\_64PP\_dmg\_repair\_mutR.RData (Input file)
4. Gene\_list\_CPD\_dmg\_repair\_mutR.RData (Input file)
5. Enh\_Total\_XR\_and\_mutR\_dataset.RData (Input file)

*Output files:*

1. Gene\_mutation\_table\_ALL\_CtoT\_v2.RData (**large file size**)
2. IMR90\_Enh\_mutation\_table\_ALL\_CtoT.RData (Output file)
3. Cancer driver gene list scripts
   1. Dataset preparation script
   2. Heatmap – Cancer driver genes with top 64PP ratio scores
   3. Heatmap – Cancer driver genes with top CPD ratio scores
   4. Heatmap – Enhancer driver genes with top 64PP ratio scores
   5. Heatmap – Enhancer driver genes with top CPD ratio scores

*Input files:*

1. Gene\_list\_64PP\_dmg\_repair\_mutR.RData (Input file)
2. Gene\_list\_CPD\_dmg\_repair\_mutR.RData (Input file)
3. IMR90\_Enh\_prot\_coding\_gene\_interactions.RData (Input file)
4. IP64\_FC\_per\_Enh.RData (Input file)
5. CPD\_FC\_per\_Enh.RData (Input file)
6. Enh\_Total\_XR\_and\_mutR\_dataset.RData (Input file)
7. MELA\_driver\_list\_with\_mut\_freq\_info.RData (Input file)

*Output files:*

1. Prot\_coding\_genes\_w\_repair\_mutR\_UVratio\_info.RData (Output file)
2. Enhancers\_w\_repair\_mutR\_UVratio\_info.RData (Output file)

***Nucleotide Excision Repair Script organization***

1. Whole genome repair analyses scripts - 1Mb
   1. Binning XR-seq bw to 1Mb bin size
   2. Merging replicates
   3. Averaging plus and minus strand
   4. Master table with all time points
   5. Cumulative repair vs UV lesion abundance scatterplot script

*Input files:*

1. XR seq files (**not included**)

*Output files:*

1. All processed and binned XRseq files
2. IP64\_lesion\_XRseq\_mastertable.RData (1Mb)
3. CPD\_lesion\_XRseq\_mastertable.RData (1Mb)
4. Chromatin state repair analysis scripts
   1. Binning XR-seq bw to chromatin state bin size
   2. Averaging plus and minus strand
   3. Master table with all time points
   4. Cumulative repair in chromatin state boxplots

*Input files:*

1. E017\_15\_coreMarks\_dense.bed.bgz (*from Roadmap*)
2. Master\_Table\_FC\_1Mb.RData (Table with ALL features)

*Output files:*

1. All processed and binned XRseq files
2. XR64PP\_PLUS&MINUS\_ALL\_TIME\_POINTS\_core15.RData
3. XRcpd\_PLUS&MINUS\_ALL\_TIME\_POINTS\_core15.RData
4. Gene repair analysis scripts
   1. Subset uniquely mappable XRseq reads
   2. Measure XRseq signal in protein coding genes
   3. Averaging plus and minus strand
   4. Cumulative repair in protein coding genes
   5. Gene repair analysis of top regions

*Input files:*

1. ENCODE24mer\_mappability\_score1.RData (***Data from ENCODE***)
2. UV lesions rank normalized to norm dist 100kb.RData (**Commonly used file**)
3. Gene\_list\_64PP\_dmg\_repair\_mutR.RData (**Commonly used file**)
4. Gene\_list\_CPD\_dmg\_repair\_mutR.RData (**Commonly used file**)

*Output files:*

1. Total\_XR64\_strand\_mean\_repair\_Prot\_genes.RData
2. Total\_XRCPD\_strand\_mean\_repair\_Prot\_genes.RData
3. Enhancer repair analysis scripts
   1. Measure unique XRseq signal in enhancer regions
   2. Average plus & minus strand repair signals
   3. Enhancer repair analysis of top susc. regions (Figure 6B-C)

*Input files:*

1. IMR90\_enhancer\_list\_EnhancerAtlas2.0.bed (*from EnhancerAtlas2.0 website*)
2. Enh\_Total\_XR\_and\_mutR\_dataset.RData (**Previously consolidated in other script**)

*Output files:*

1. Total\_XR64\_IMR90\_Enh.RData

Total\_XRCPD\_IMR90\_Enh.RData