

Readme (base observed data)
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@

The EpiATLAS observed dataset contains 337 complete epigenomes. Each includes:

** ChIP-Seq enrichment data for six histone modifications (H3K27ac, H3K4me1, H3K4me3, H3K36me3, H3K27me3, H3K9me3)

** RNA-Seq quantification (gene and transcript level)

** Genome-wide DNA methylation measurements

In addition, 1942 epigenomes are available with a partial set of these assays. In total, the collection contains 7539 processed datasets. All datasets were processed using the same standardized pipelines to ensure consistency across samples.

Data Processing Pipelines

=====

ChIP-seq data was processed using the following container pipeline:
https://github.com/IHEC/integrative_analysis_chip/releases

WGBS data was processed using the following container pipeline:
<https://github.com/heathsc/gemBS/releases/tag/v3.5.0>

RNA-seq data was processed using the following container pipeline:
<https://github.com/IHEC/grape-nf/release>

Analysis File Nomenclature

=====

ihec.<analysis_type>.<pipeline&version>.<epirr_id>.<uuid>.<file extensions>

where:

analysis type is one of: chipseq, wgb, rna-seq

epirr_id includes the epirr minor version

uuid-auniquealpha-numeric string (lower-case)

Example:

- ihec.chipseq.ihec-chipseq-containerv1.1.4.IHECRE00000001.4.8e4a8d6c-26e0-4b8d-aa08-7875973212d4.*

- ihec.rna-seq.ihec-grapenf-containerv1.1.0.IHECRE00000001.4.5e6cc904-0c49-4a82-8bf0-1f727f9f37fa.*

- ihec.wgb.ihec-gembs-containerv3.5.0.IHECRE00000001.4.40e2b84e-18fc-4f52-925f-c35af2f9e611.*

ChIP-Seq outputs

=====

For each histone mark, the following files are provided:

Peak calls (MACS2, p-value threshold 0.01):

*pval0.01.500K.narrowPeak.gz (unfiltered)
*pval0.01.500K.bfilt.narrowPeak.gz (ENCODE blacklist-filtered)

Signal tracks:

*fc.signal.bigwig (fold enrichment)
*pval.signal.bigwig (-log10 p-value signal)

Quality control reports:

*qc.html
*qc.json

Alignment and coverage files (read sequence information removed for privacy):

*noseq.bam (alignment without sequences)
*raw.bigwig (raw coverage signal)

Matching input control files:

*ctl_noseq.bam
*ctl_raw.bigwig

RNA-Seq outputs

=====

Outputs provide gene- and transcript-level quantification based on STAR aligner results, bigWig coverage files for both strands (excluding and including multimapped fragments); QC/stats report from grape_nf and RSEM model details.

*genes.results
*isoforms.results

*Unique.minusRaw.bw
*Unique.plusRaw.bw
*UniqueMultiple.minusRaw.bw
*UniqueMultiple.plusRaw.bw

*report.pdf
*stats.json

*model
*theta
*cnt

RNA-Seq data includes both mRNA and total RNA experiments. Also note that while majority of data was generated using strand-specific protocol, some datasets were generated with non-stranded protocol and for non-stranded data only one coverage bigWig file is available.

DNA methylation outputs

=====

DNA methylation data includes GemBS outputs containing coverage and fractional methylation of cytosine methylation in 3 possible contexts
*cpg.bed.gz
*chg.bed.gz
*chh.bed.gz

Output signal tracks for both strands
*gembs_neg.bw
*gembs_pos.bw

Complete GemBS QC/analysis reports:
*gemBS_reports.tar.gz

Note the EpiATLAS includes data from both standard Whole-genome bisulfate sequencing (WGBS) protocol (standard) as well as PBAT-based measurements.

Notes

=====

All BAM files have sequence information removed to preserve donor anonymity.
bigWig files can be directly visualized in genome browsers (e.g., IGV, UCSC).
Filtered peak files (bfilt) are recommended for most downstream analyses.
For file collection for every analysis data sets we also provide a file containing the corresponding md5 check sums.

For full methodological details, see the Methods section of the EpiATLAS manuscript and information presented in the descriptions of the individual pipelines (see links above).

The base and post-processed data and metadata are available for the download from:
<https://ihec-epigenomes.org/epiatlas/data/>

questions: info@ihec-epigenomes.org
==

mbilenky@bcgsc.ca