

CCRIFX ChIP-Seq PROJECT DESCRIPTION

Project Number: CCRIFX-

Project Title:

Requestor:

Requestor Lab PI:

Address:

Additional Investigators to include:

Bioinformatics contact:

Biowulf area:

Completed Request(s):

Current Active Request(s):

Rationale/Significance/Summary:

Scientific questions or goals

- *E.g. which genes are regulatory targets of the TF?*

Overall Design/Approach

- *E.g. ChIP-Seq profiling and identification of mutation-specific TF binding events*
- *E.g. Integration with microarray expression data*

Impact

-

Other priority considerations

- Event(s) in the near future that makes this request time-sensitive:
 - Site visits, manuscripts, presentations, meetings, etc.
- The availability of data to be analyzed: sequencing in progress, available now, public sources
- Role of the project:
 - Publishable experiment
 - Exploratory
 - Confirmatory

Known risks and limitations

Original description

Experimental design and metadata

- **Series and Samples** (feel free to attached a GEO template form instead)
 - Species and strains: *e.g. human, mouse (Mus musculus), rat, yeast*
 - Strains: *e.g. C57BL/6*
 - Source (tissue, cell lines, cells): *e.g. ES-derived neural progenitor cells*
 - Extracted material/molecule: genomic DNA

- Overall design: e.g. *Examination of 2 different histone modifications in 2 cell types*
- Total number of
 - Samples - ?
 - Groups - ?
 - Replicates (biological and technical) - ?
 - Sample names
 - Processed file names: e.g. *H3K4me2.aligned.txt, H3K4me2.peaks.txt*
 - Raw sequencing file names: e.g. *H3K4me2.peaks.txt, 080716_BI-EAS46_0001_209DH_L1.fastq*
- ChIP-Seq antibody: e.g. *H3K4me2 (Millipore, 07-030, lot 122116)*
- Native ChIP or cross-linking ChiP?
- Antibody and IP efficiency validation (immunoblots, IP assays, ChIP-PCR, etc.)
 - Good antibody specificity? (Yes/No)
 - Good IP efficiency? (Yes/No)
- Do you expect narrow sequence-specific peaks (e.g. transcription factors) or broad peaks (e.g. histones)?
- Diversity of the population, intra individual variability of the sample, clonal heterogeneity, similarity to the reference if known

- **Protocols** (feel free to attached a GEO submission template form instead)
- Library construction protocol: *Illumina TruSeq 2.0*
- Library strategy: *ChIP-Seq*
- Sequencing platforms and instruments used: *GAllx, Illumina HiSeq*
- Libraries: *SE; 1 library per lane*
- Sequence read length: *36bp*

- **Data**
- Sample metadata spreadsheet
- Read location and format: *Illumina fastq, Casava BAM files*
- Type of data: *Illumina reads*
- File checksum: *yes/no*
- SAIC-F CSAS#
- Preferred reference genome: *hg19, mm9, mm10*
- Preferred reference annotation: *Refseq, UCSC, Ensembl*
- SF – Base calling: *RTA 1.8.70.0; Alignment: Illumina Casava 1.8.2*
- For public data mining requests:
 - Types of data: *RNA-Seq, ChIP-Seq, microarray*
 - Types of studies: *diseases, cancer subtypes, databases*
 - GEO datasets: *GEO IDs, paper PubMed IDs*

Data processing pipeline/Analysis details

- **Data processing pipeline used by CCRIFX**
 - Preferred analytical applications – standard
 - Software
 - *MACS, SICER, Homer, Genomatix, DiffBind*
 - Major steps in the workflow or procedure that will be followed
 - Data QC and visualization using

- *FastQC, FastQC_Report and Genomatix*
 - *IGV, UCSC*
 - *Post-alignment QC using Genomatix*
 - Peak calling (vs. respective input samples)
 - Comparison of ChIP-Seq samples at peak-level
 - *e.g. KD vs. WT samples -> BED files*
 - Motif finding (for transcription factors)
 - Known motifs at the individual factor level
 - *de novo* motifs
 - Comparison with public data
 - Analysis of genes and pathways associated with peaks
 - Enrichment analysis using *IPA, GSEA, GO*
 - Data integration between ChIP-Seq and microarray data
 - Comparison of genelists
 - Up/down-regulated genes vs. gene lists associated with peaks
 - Network analysis using *IPA, GeneGo MetaCore*
 - Genome features of interest: *TF binding sites, target genes, pathways*
- **Results expected from CCRIFX**
 - Files, figures and formats
 - Peak lists, gene lists including distance to TSS and gene classification, binding sites, overlap and unique data for each mutant
 - List of methods used with brief descriptions
 - Education and training
 - Automated workflows
 - Submission to public repositories: *NCBI GEO, SRA, GenBank*
 - Access to intermediate files on the Helix/Biowulf project directory
 - Investigator's username on Helix:

Deferred tasks

- Related analyses on the same data that are too big to fit into the current request
- Related analyses on that require additional data
- Requests for continuous support

Publications

- Chen et al., **Systematic evaluation of factors influencing ChIP-seq fidelity**. *Nature* 2012
- Landt et al., **ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia**. *Genome Res.* 2012 Sep; 22(9): 1813-31
- Liu et al., Q&A: **ChIP-seq technologies and the study of gene regulation**. *BMC Biology* 2010, 8:56
- Kidder et al., **ChIP-Seq: technical considerations for obtaining high-quality data**. *Nature Immunology* 12, 918–922 (2011)
- Marinov GK, Kundaje A, Park PJ, Wold BJ. Large-Scale Quality Analysis of Published ChIP-seq Data. G3 (Bethesda). 2013 Dec 17. pii: g3.113.008680v1.