Analysis of your ChIP-Seq data: Collaborating with the CCRIFX Core

Best practices to achieve meaningful results

Monday, November 18, 2013

Acknowledgements

This presentation was developed by the CCRIFX core of the Advanced Biomedical Computing Center operated by SAIC-Frederick. Funding for CCRIFX comes from the CCR Bioinformatics Initiative (YT-10-060).

CCRIFX

Natalie Abrams, PhD

Yvonne Edwards, PhD

Parthav Jailwala, MS

Li Jia, MS

Manjula Kasoji, MS

Nathan Starner, BS

Anand S Merchant, MD, PhD

Eric Stahlberg, PhD

ABCC

Brian Luke, PhD

Uma Mudunuri, MS

Bob Stephens, PhD

Ming Yi, PhD

Jack Collins, PhD

CCR experts

Maggie Cam, PhD

Sean Davis, MD, PhD

Max Lee, PhD

Peter FitzGerald, PhD

David Goldstein, PhD

Sequencing Facility

Yongmei Zhao, MS

Bao Tran, MS

What is the CCR Bioinformatics (CCRIFX) core?

- http://ccrifx.cancer.gov
- Commenced support efforts in January 2011
- Collaborative support for > 100 data projects
  - Design2Delivery
  - NGS data, microarray, metabolomics, data mining
- Education and training
- Workflow/pipeline development
- No additional cost to collaborate with the core

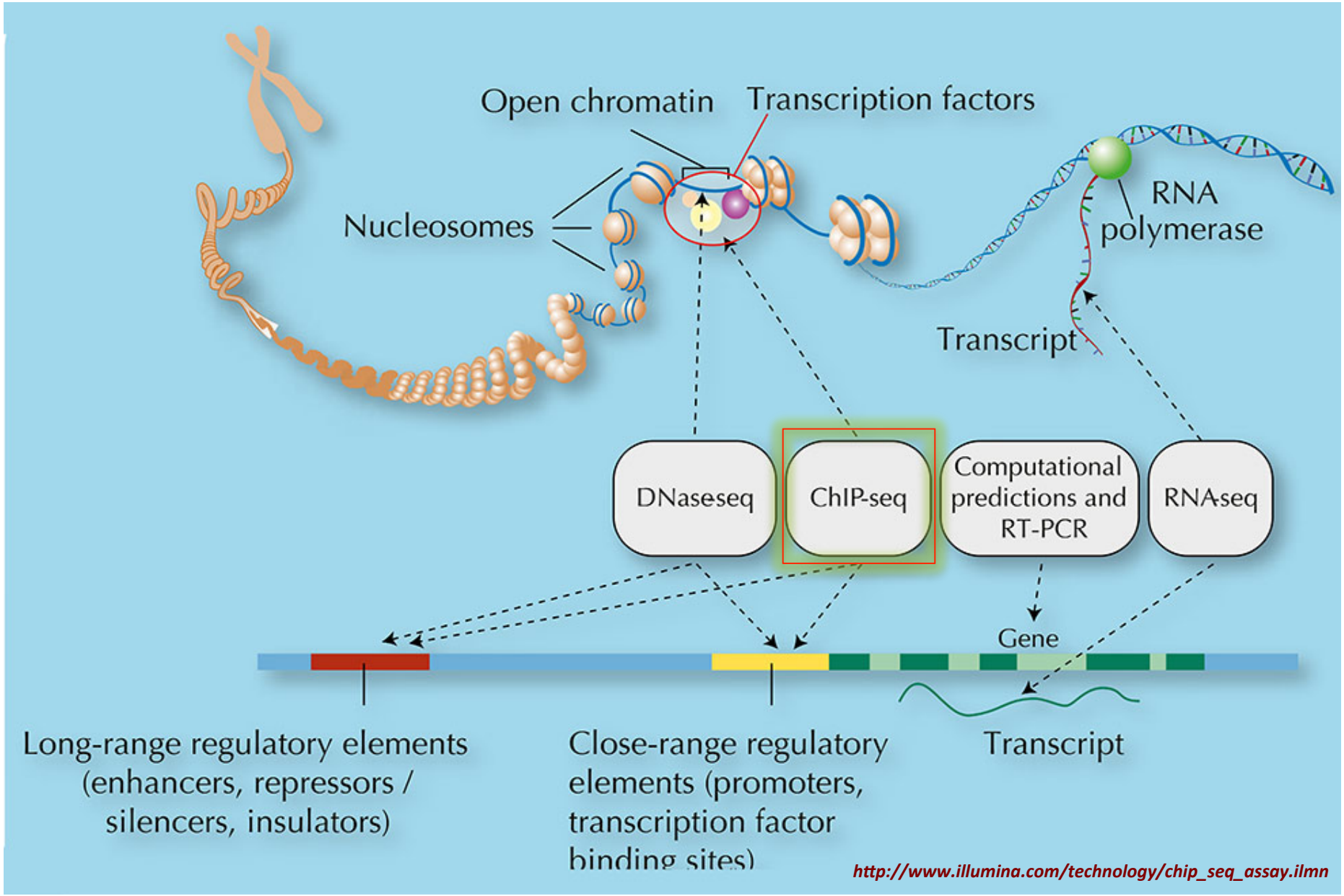
Outline

- **Introduction**
  - ✧ Challenges with ChIP-Seq
  - ✧ Snapshot of ENCODE recommendations

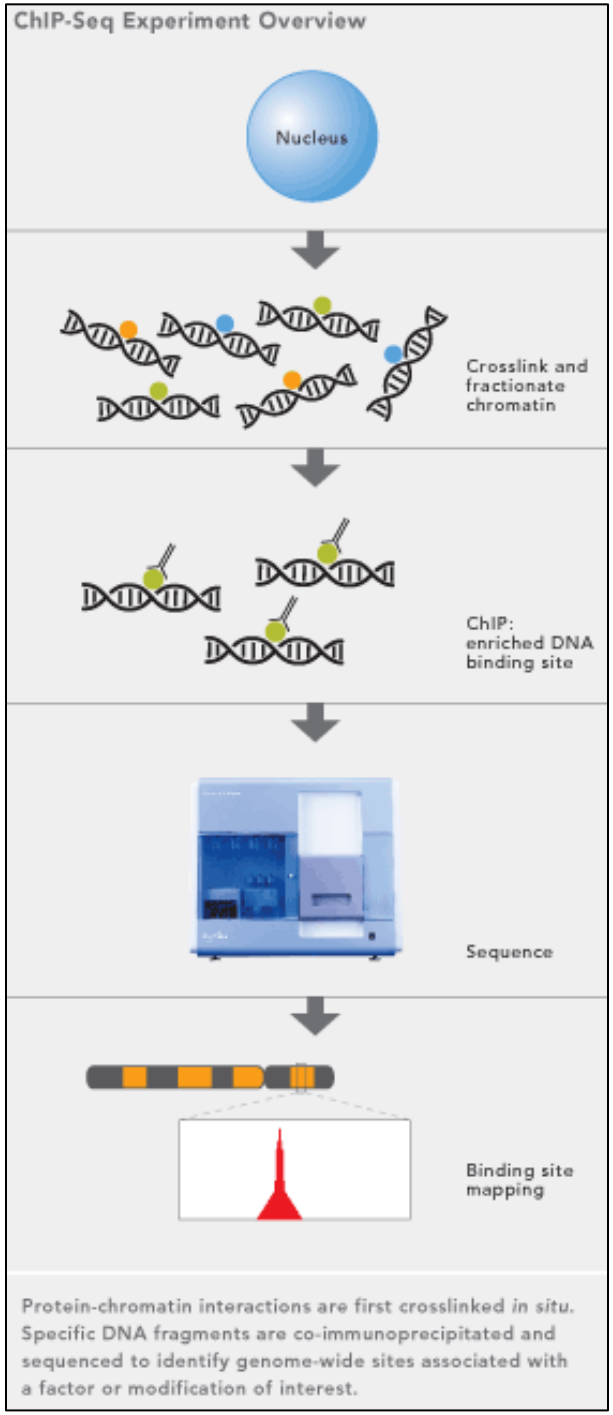
- **Consultations with the CCRIFX CORE**
  - ✧ Practical design considerations
  - ✧ Data/ metadata for the analysis

- **Examples of ChIP-Seq data** – QC metrics
- **Enhanced support** – enrichment, integration, methods
- **Summary of Proposed Guidelines for ChIP-Seq Assessment**

Targeting Different Elements in Biological Space using NGS



http://www.illumina.com/technology/chip_seq_assay.ilmn



Critical phases	Challenges across phases
Experimental Phase	▪ Variations in conditions ▪ Non-Specificity of antibodies ▪ Low IP efficiency ▪ Absence of Replicates/ Controls
Sequencing Phase	▪ Library construction ▪ Sequencing chemistry ▪ Instrumentation ▪ Depth of coverage
Analysis Phase	▪ Different protein classes require different analysis ▪ Parameter optimization ▪ Comparisons across experiments

Snapshot of ENCODE Recommendations

Really good antibody to start with!

EXPERIMENTAL DESIGN GUIDELINES

- At least 2 replicates
- Input Control for each condition
- Reproducibility
- Library complexity
- Adequate Sequencing depth to capture events across genome

DATA QUALITY ASSESSMENT

- Metrics at every stage possible to assess quality of experiment
- Cross-correlation for stranded reads
- Irreproducible Discovery Rate (IDR) for peak concordance in replicates

DATA REPORTING GUIDELINES

- Minimal Information for Chip-seq Experiment (MICE)
- Analysis Details
- High-throughput sequencing data

Important to Confirm that the Antibody Works!

BIOMEDICINE: NATURE NEWS | 1 AUGUST 2013 | VOL 500 | NATURE | 15

NIH mulls rules for validating key results

US biomedical agency could enlist independent labs for verification

"Having a third party validate every batch would be a fabulous thing," says Peter Park, a computational biologist at Harvard Medical School.

He notes that the consortium behind ENCODE — a project aimed at identifying all the functional elements in the human genome — tested more than 200 antibodies targeting modifications to proteins called histones and found that more than 25% failed to target the advertised modification.

Consultation with the CCRIFX Core

The Quest for a Successful ChIP-Seq –

- **The recommendation is to approach the CCRIFX Core prior to or during the ChIP-Seq experimental design process**
- **A practical goal is to maximize site discovery by**
  - Optimizing IP and
  - Sequencing deeply, within reasonable expense constraints
- **Goals of data quality assessment**
  - Identification of most suitable analytical approaches
  - Defining outcomes

Advantage of collaborating with CCRIFX

- Better understand research question
- Suggest viable experimental designs
- Guide on use of technologies
- Assist in data tracking and transfer of very large datasets
- Perform analysis and explain results
- Deliver figures and methods for manuscripts, presentations, posters
- Maximize the research impact of your invested effort

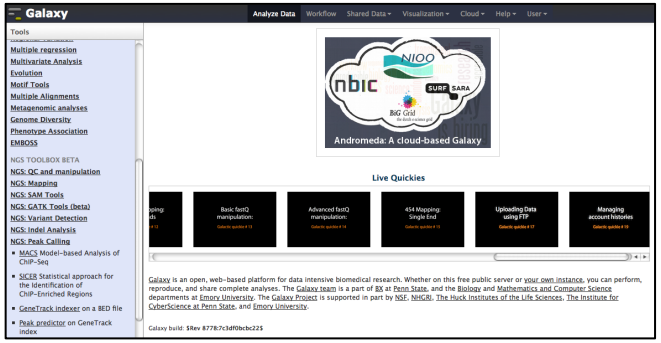
ChIP-SEQ DATA ANALYSIS

Most widely used data analysis tools

- MACS (**Model-based Analysis for ChIP-Seq**)
  - Widely used method for peak calling
- HOMER **(Hypergeometric Optimization of Motif EnRichment)**
  - Suite of tools for motif discovery and ChIP-Seq analysis
- SICER (**Statistical approach for the Identification of ChIP-Enriched Regions**
  - Clustering approach for identification of domains from histone modification
- CCRIFX Wiki - **https://ccrod.cancer.gov/confluence/display/ifx1/Differential+ChIP-seq+binding**

GUI tools/packages for ChIP-Seq analysis

Galaxy Server



Partek

- Genomatix NGS Suite (available through CCR OSTR)

Checklist for a ChIP-seq experiment

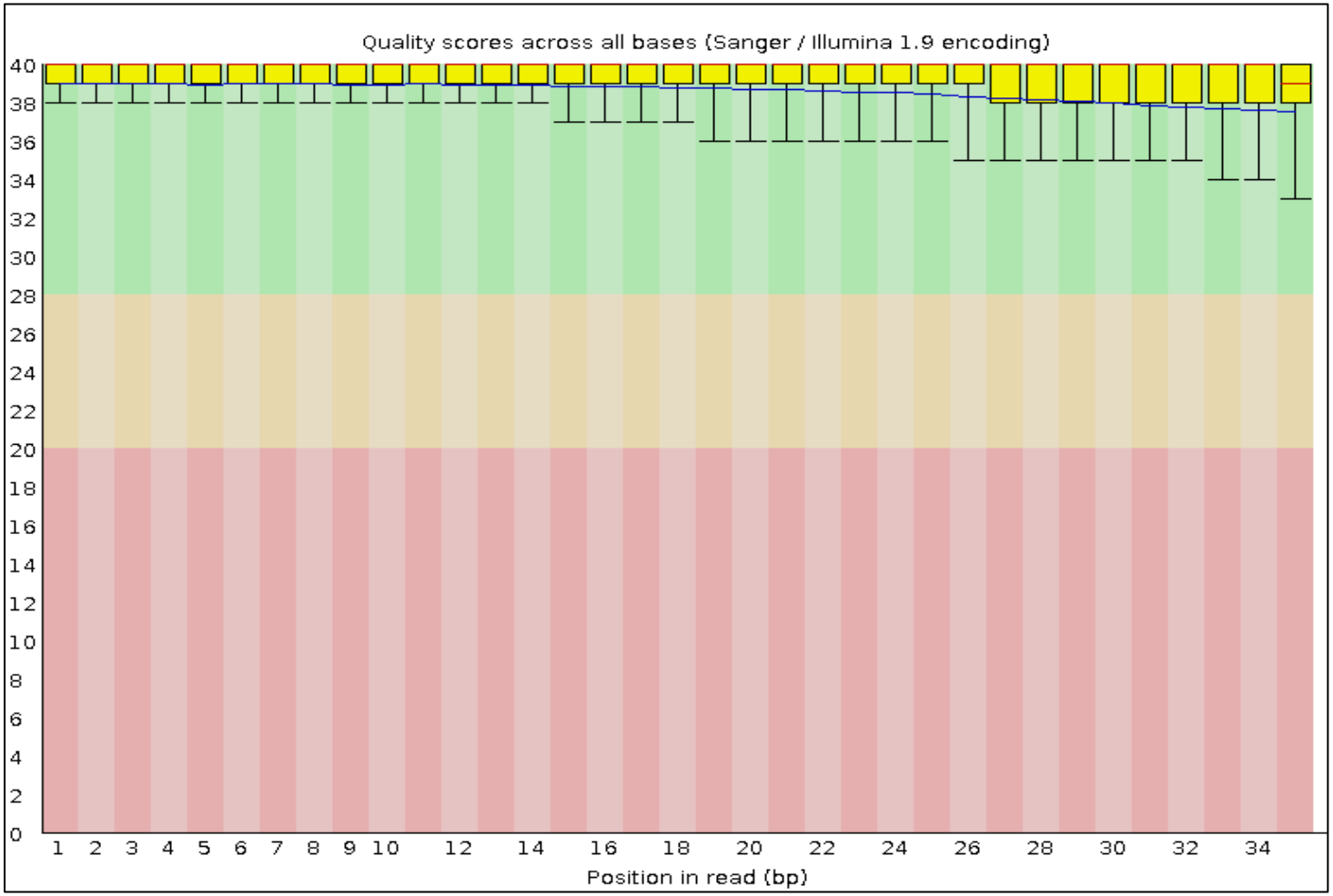
1. **Raw data** – format (.fastq, .bam); location (server, hard drive); SF CSAS#, if applicable
2. **Sample Metadata**
   a. Species and strains: human, mouse, etc.; Type: tissue, cell lines, etc.
   b. Total number of Samples; Experimental Groups; Replicates
3. **Native ChIP or cross-linking ChIP?**
   a. Antibody validation (immunoblots, IP assays, etc.)
   b. Good antibody specificity and IP efficiency? (Yes/No)
4. **Library Metadata**
   a. Sample prep/enrichment and Library construction protocols
   b. Sequencing platforms and instruments used
   c. Libraries: Type (PE/SE); Sequence read length
5. **Analysis-related specifications**
   a. Preferred reference genome (hg19, mm10, etc.), and annotation (Refseq, UCSC)
   b. Integrated analysis with other data types (microarray, RNA-Seq)

Illustrations of ChIP-Seq Analysis

Determining data suitability using QC plots

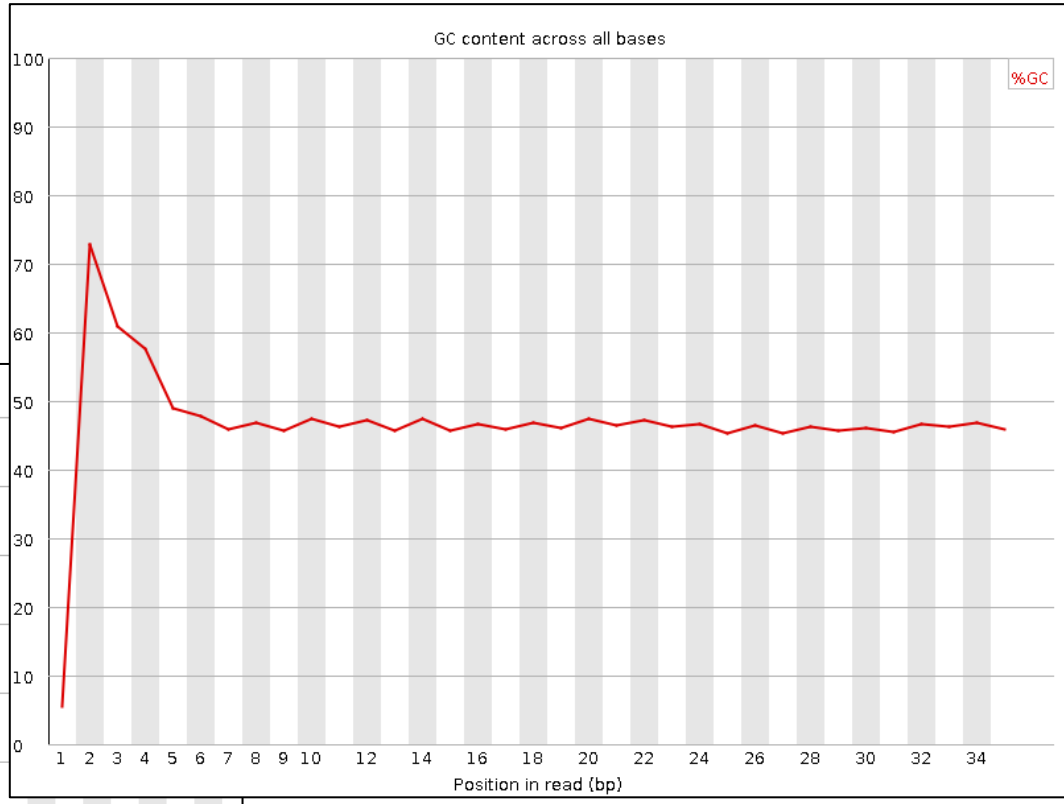
Features –

1. Qscore > 30 (green zone)
2. Less spread/bp

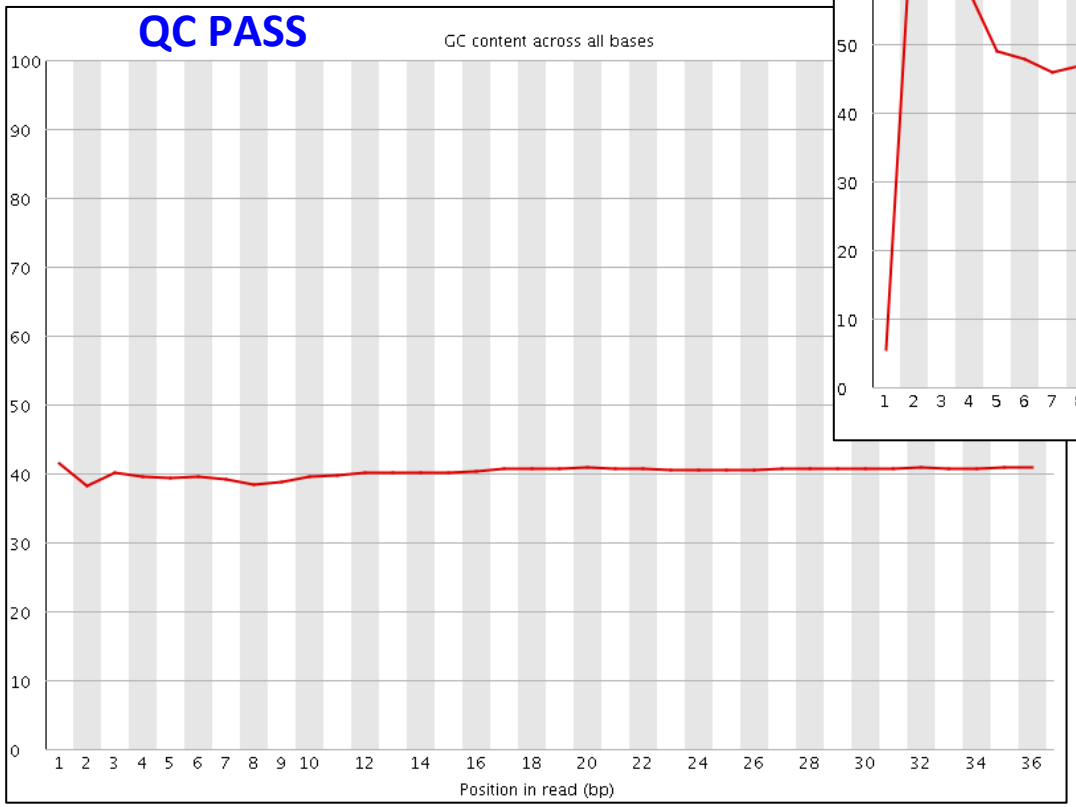


Per-base GC Content Distribution Plot

QC WARNING*



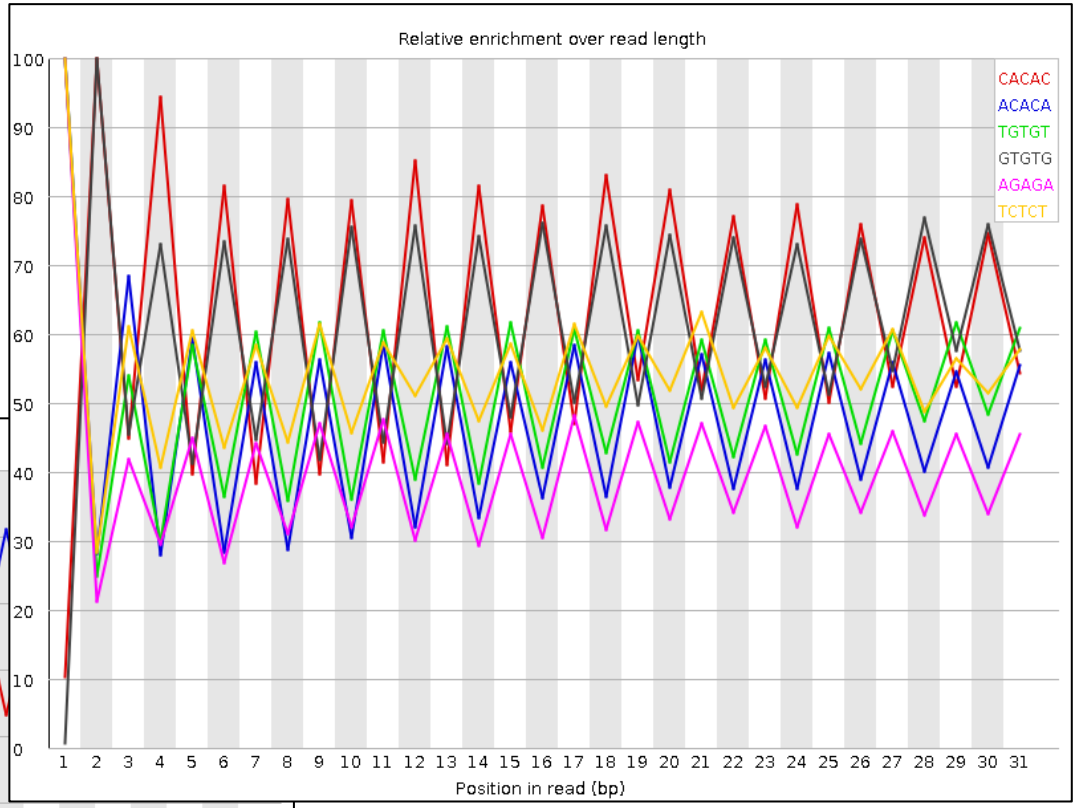
QC PASS



** Not a criteria to trim reads just based on this plot, but in combination with other QC parameters provides a basis to improve read quality*

K-mer (repeat) profile

QC WARNING*



Post-repeat removal

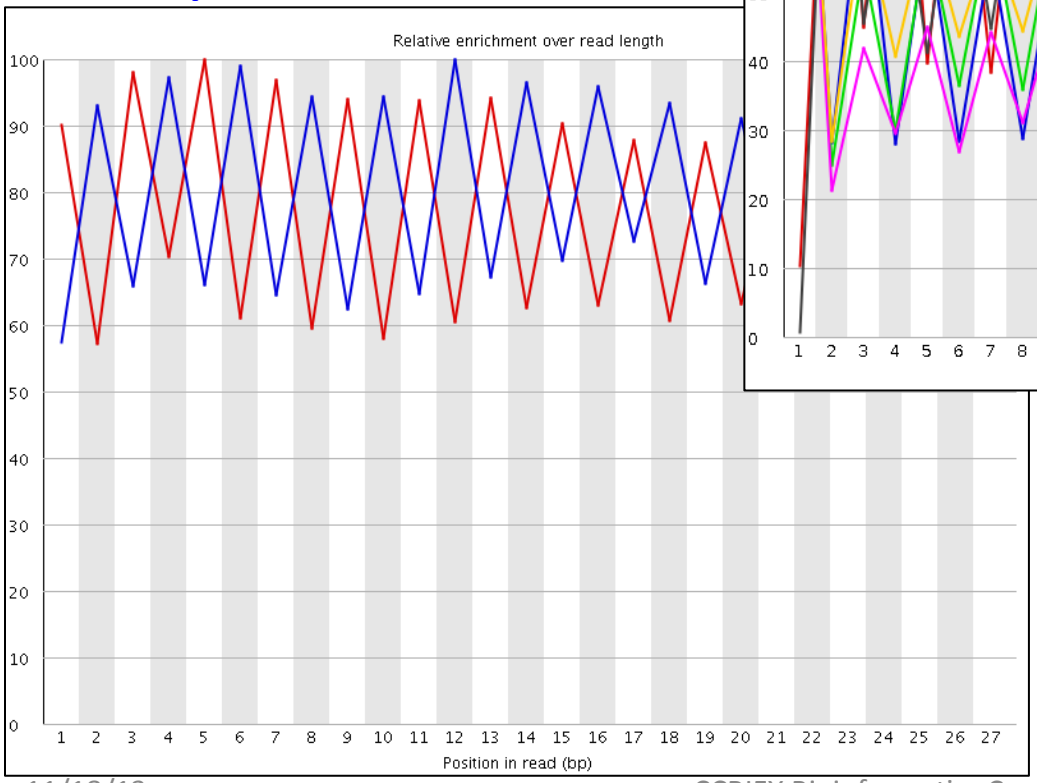
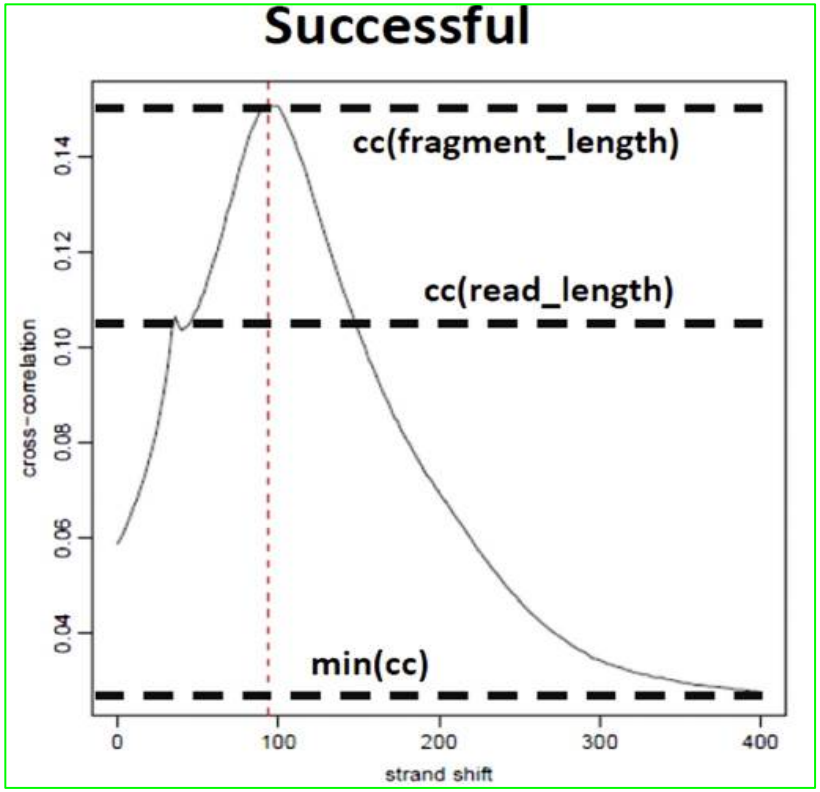
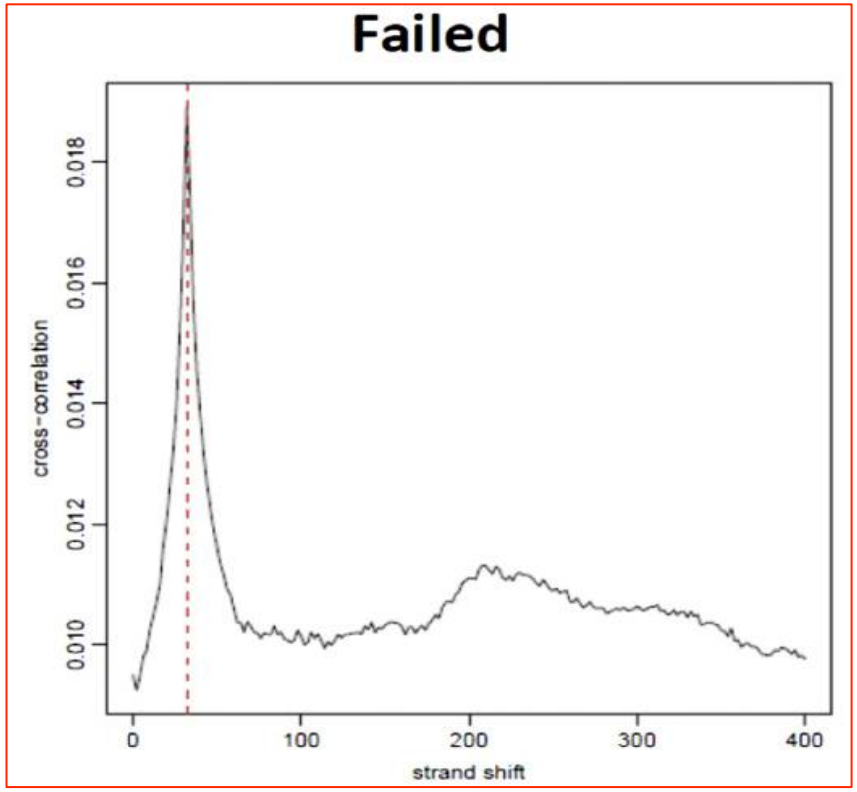


Figure from ENCODE publication on strand correlation





$$NSC = \frac{cc(fragment\ length)}{min(cc)}$$

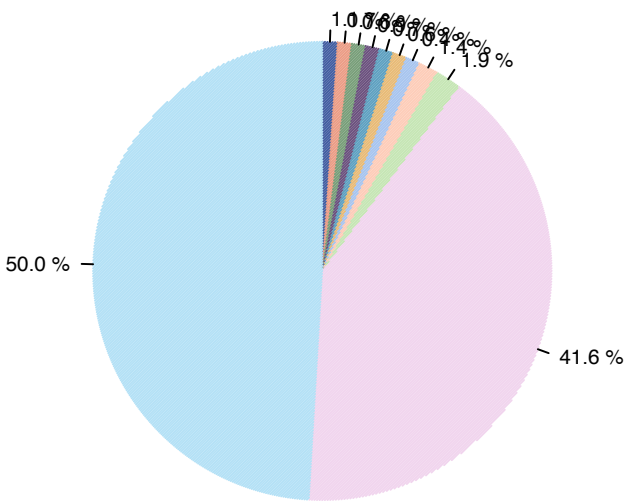
$$RSC = \frac{cc(fragment\ length) - min(cc)}{cc(read\ length) - min(cc)}$$

Stephen G. Landt, Georgi K. Marinov, Anshul Kundaje, et al. 2012 22: 1813-1831 Genome Res.; doi:10.1101/gr.136184.111

Pie chart with distribution of peaks across different genomic regions

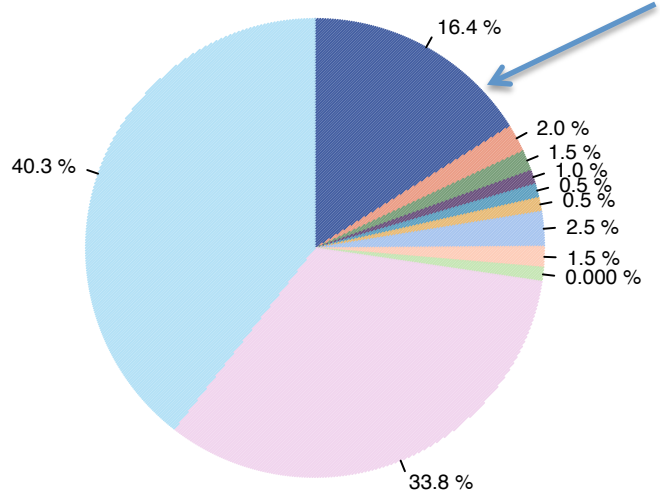
Output from CEAS (Cis-regulatory Element Annotation System)

Genome



- Promoter (<=1000 bp): 1.1 %
- Promoter (1000–2000 bp): 0.7 %
- Promoter (2000–3000 bp): 0.6 %
- Downstream (<=1000 bp): 0.9 %
- Downstream (1000–2000 bp): 0.7 %
- Downstream (2000–3000 bp): 0.6 %
- 5'UTR: 0.4 %
- 3'UTR: 1.4 %
- Coding exon: 1.9 %
- Intron: 41.6 %
- Distal intergenic: 50.0 %

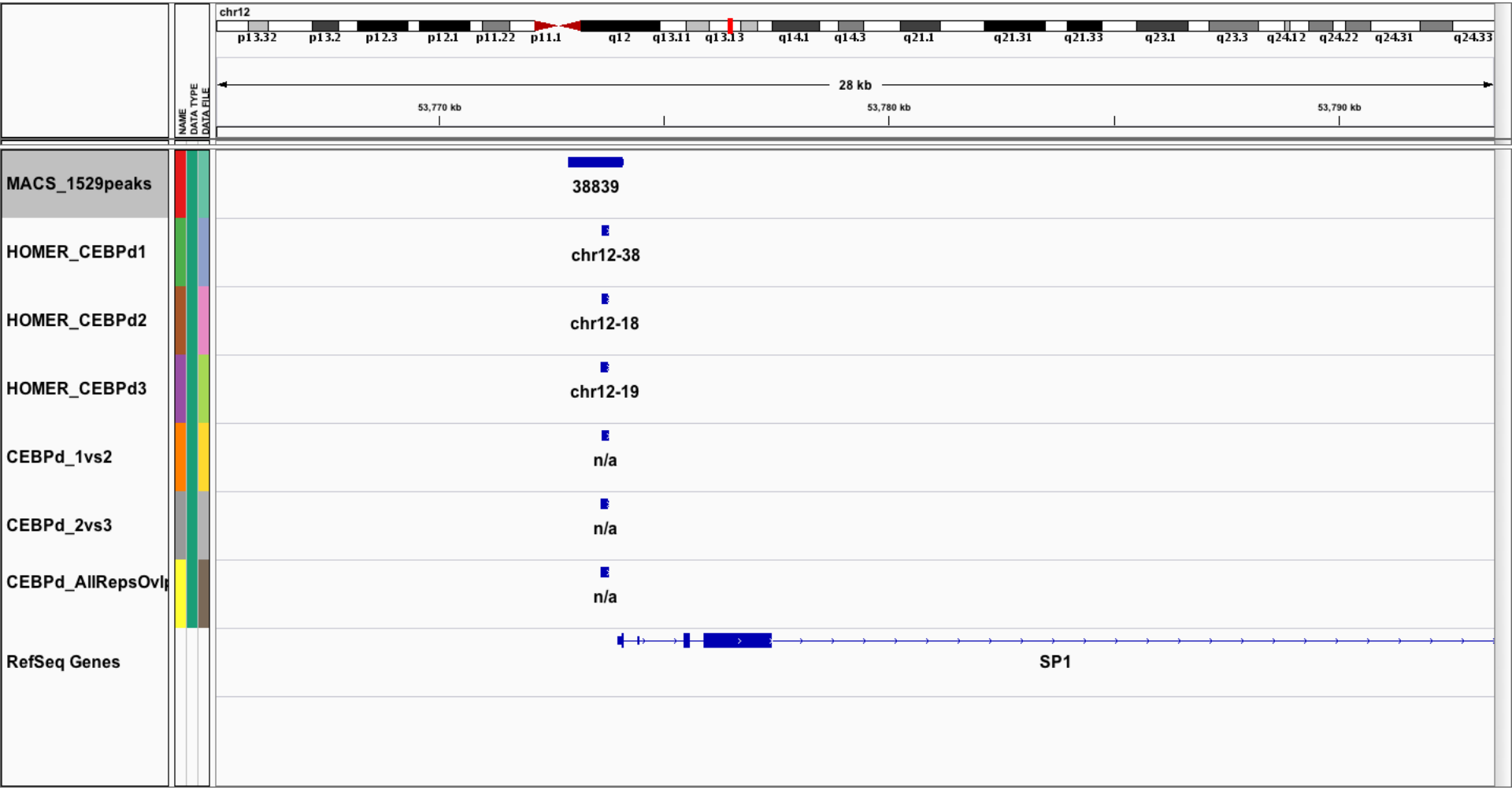
ChIP



- Promoter (<=1000 bp): 16.4 %
- Promoter (1000–2000 bp): 2.0 %
- Promoter (2000–3000 bp): 1.5 %
- Downstream (<=1000 bp): 1.0 %
- Downstream (1000–2000 bp): 0.5 %
- Downstream (2000–3000 bp): 0.5 %
- 5'UTR: 2.5 %
- 3'UTR: 1.5 %
- Coding exon: 0.000 %
- Intron: 33.8 %
- Distal intergenic: 40.3 %

Enrichment in promoter region

Visualization of BED files on IGV



SP1 Promoter Region

Proposed Guidelines for Assessing a ChIP-Seq Experiment

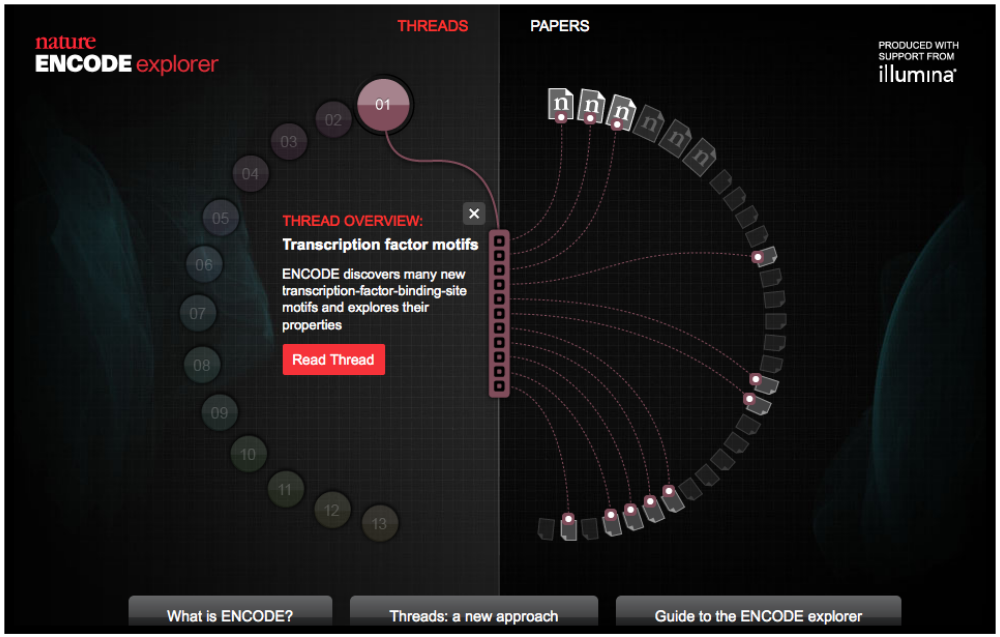
	Pre-ChIP-Seq Experimental Index		Post-ChIP-Seq Quality Index
1	Biological Replicates		Number of uniquely mapped reads
2	Documented ChIP-Valid Antibody		Significant statistics (FDR p-value; Fold-change)
3	Specific Sample Characteristics (Cell Line/Tissue; Culture Conditions; Comparisons; Multiple factors)		a. QC outputs b. Motif enrichment b. Browser Visualizations
4	Controls – Input/IgG or both		ENCODE Metric 1 - Non-Redundancy Fraction (NRF)
5	Available ChIP-PCR data/ Positive Controls		ENCODE Metric 2 – Strand cross-correlation coefficients (NSC/RSC)
6	Other corroborative evidence (e.g. publications)		ENCODE Metric 3 - Irreproducibility Discovery Rate (IDR)

Characteristics of a ChIP-Seq Experiment that is…

Good	Sub-optimal	Poor
• ChIP-validated Ab • Efficient IP • At least 2 experimental replicates • Input DNA control • Excellent sample quality and library complexity • Sufficient sequencing depth (>10x)	• Ab and IP issues • No replicates • Low quantity/quality of sample • QC Flags • Insufficient depth of sequencing	• Poor IP (Ab binding) • No Input • No replicates • Unbalanced experimental conditions (e.g. treatment, time)

Useful ENCODE Links

- Quality Metrics: http://encodeproject.org/encode/qualityMetrics.html#spreadsheets
- Data Coordination Center: http://encodeproject.org/encode/otherTerms.html
- Software Tools: http://genome.ucsc.edu/encode/encodeTools.html#metrics
- NATURE ENCODE Explorer: http://www.nature.com/encode/#/threads
- NPG Reporting Checklist: http://www.nature.com/authors/policies/checklist.pdf



Get in touch with us!

CCRIFX Website:
http://ccrifx.cancer.gov/apps/site/default

CCRIFX Staff:
http://ccrifx.cancer.gov/apps/staff/

CCRIFX Contact:
ccrifx_support@mail.nih.gov

Mon-Wed & Friday - Building 37, Room 1123
Building 41, Room B620

Thursday (NCI-Frederick) - Building 430, First Floor

Office Hours (all locations): Fridays 9:30am -11:30am