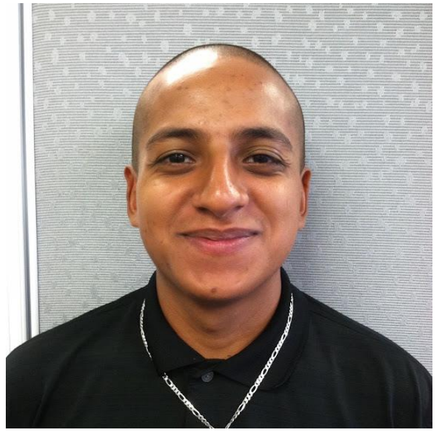
St. Jude Children's Research Hospital®
Finding cures. Saving children.

Introduction to ChIP Sequencing and QC

Sharon Freshour, PhD
St. Jude Children's Research Hospital
December 11, 2024

The DNB Bioinformatics Core Team



Cody Ramirez, PhD

Senior Bioinformatics Research Scientist
Core Director
Boston, Massachusetts

Antonia Chroni, PhD

Senior Bioinformatics Research Scientist
New York, New York

Asha Jacob Jannu, PhD

Bioinformatics Research Scientist
Indianapolis, Indiana

Sharon Freshour, PhD

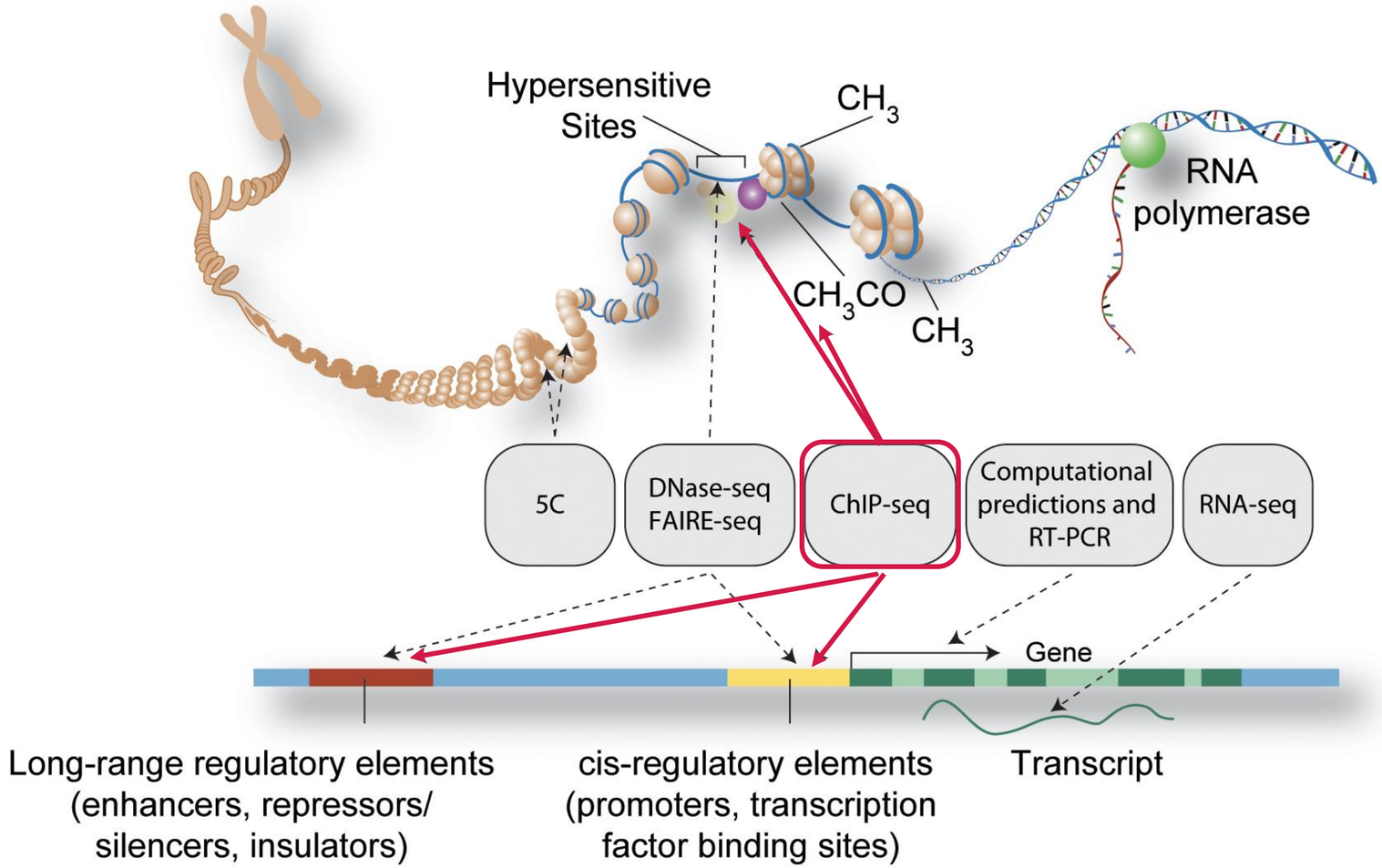
Bioinformatics Research Scientist
St. Louis, Missouri

Introduction to ChIP Sequencing and QC Workshop Overview

- ChIP-Seq overview
  - What is ChIP-Seq?
  - How is ChIP-Seq done?
  - What are other methods for profiling regulatory elements?
- Submitting samples for sequencing and analysis
- CAB’s ChIP-Seq pipelines
  - AutoMapper, QC, and Peak Calling
- Example of ChIP-Seq QC and Peak Calling report from CAB

ChIP Sequencing Overview

ChIP sequencing is a useful tool for understanding complex transcriptional regulation



Adapted from The ENCODE Project Consortium (2011). PLOS Biology

How does ChIP-Seq work?

- Uses a combination of chromatin immunoprecipitation (ChIP) and NGS (seq)
  - Antibody selection for proteins of interest
  - Next generation sequencing
  - Assays protein-DNA binding in vivo, across genome
- Complements gene expression profiling, DNA accessibility methods
- Caveats
  - Qualitative, not quantitative, profiles enrichment
  - Need quite a bit of material for standard ChIP-Seq
  - Heterogeneity can be hard to capture
  - Must have good antibodies for selection step

ChIP-Seq library preparation considerations

Must have sufficient starting material

- At minimum, 10^7 cultured cells recommended for single ChIP experiment

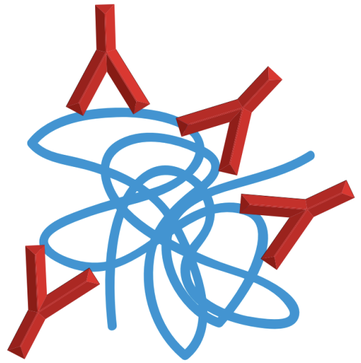
Success dependent on antibody selection

- Should be specific
- Can monoclonal or polyclonal

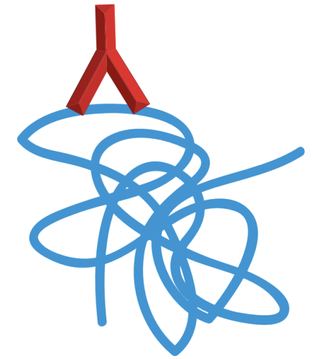
Two general ChIP procedures to choose from:

- Native ChIP (N-ChIP)
- Cross-linking ChIP (X-ChIP)

Polyclonal antibodies



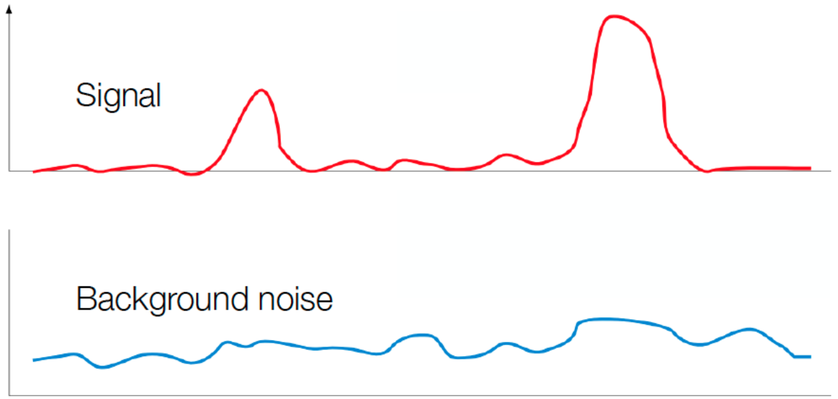
Monoclonal antibodies



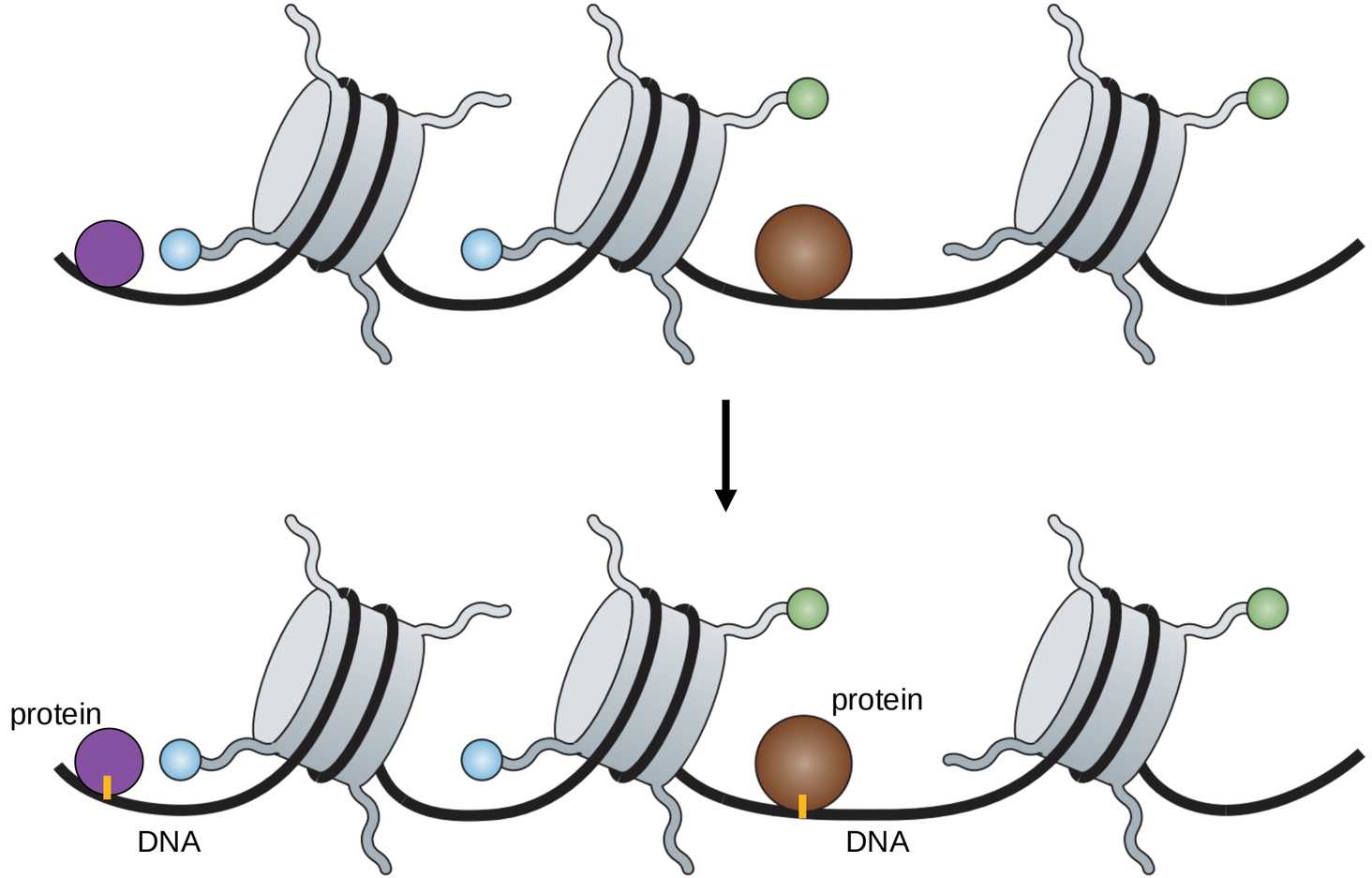
Credit: www.abcam.com/chip

Controls are necessary for ChIP-Seq experiments

- Noise in ChIP experiments not uniform
  - Affected by chromatin confirmation, local biases, mappability
- Need to model background noise to distinguish true peaks
- Input controls are necessary to estimate noise
  - Cross-linked, fragmented DNA without antibody enrichment
  - Recommended one for every immunoprecipitation done
  - If constraints, one per sample group can be sufficient
- Isotype (IgG) controls can also be used
  - Immunoprecipitation with an isotype-matched control
  - Similar to experimental antibody, but non-specific binding
- Can also use positive and negative controls, qPCR to check success
  - Positive: check signal in with expected protein binding
  - Negative: check lack of signal in non-enriched region

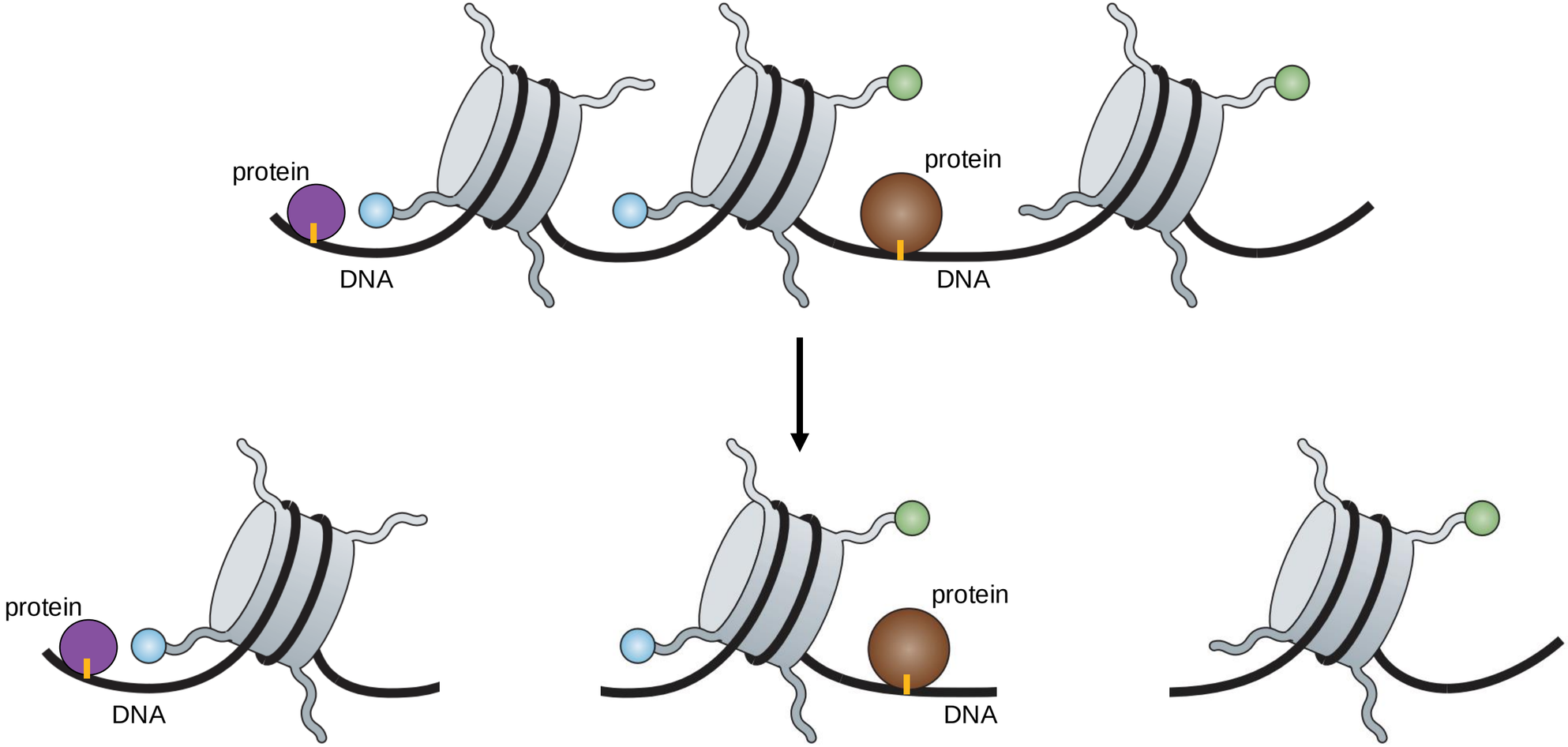


Cross-linking proteins and DNA is (often) the first step for ChIP-Seq library prepartion



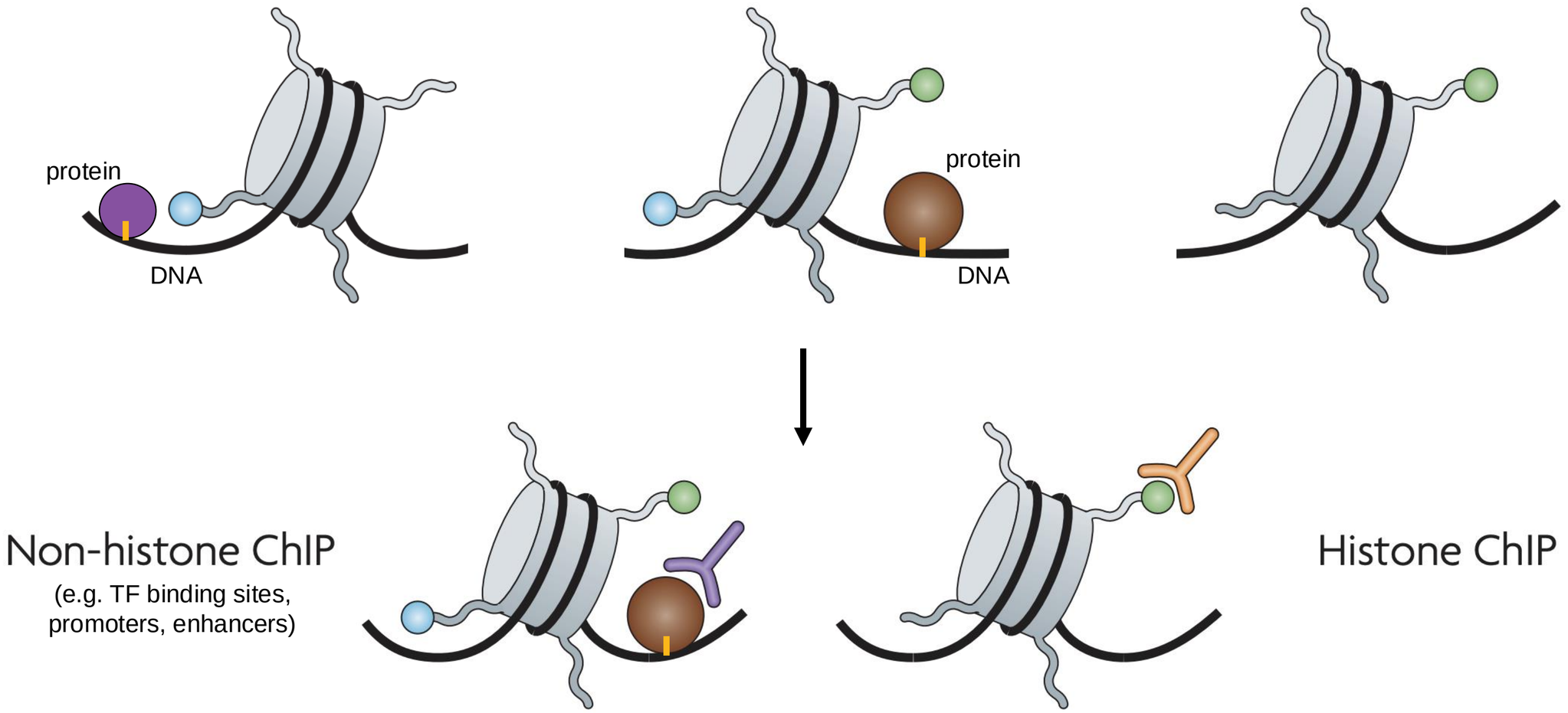
Adapted from doi.org/10.1038/nrg2641

Next step for ChIP-Seq preparation is fragmenting DNA



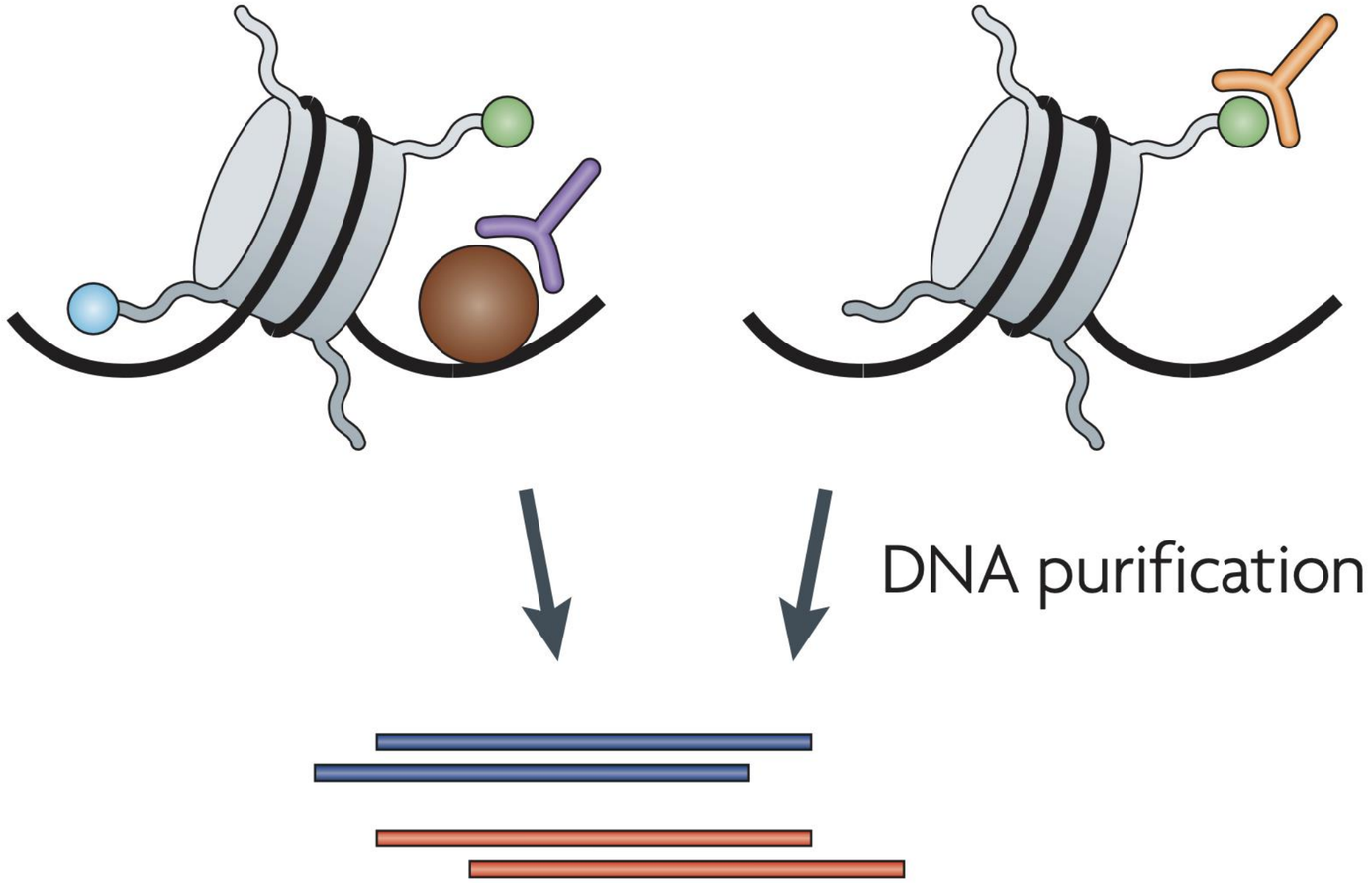
Adapted from doi.org/10.1038/nrg2641

Fragmented, protein-bound DNA is immunoprecipitated using specific antibodies



Adapted from doi.org/10.1038/nrg2641

Cross-linking is reversed and DNA is purified for sequencing

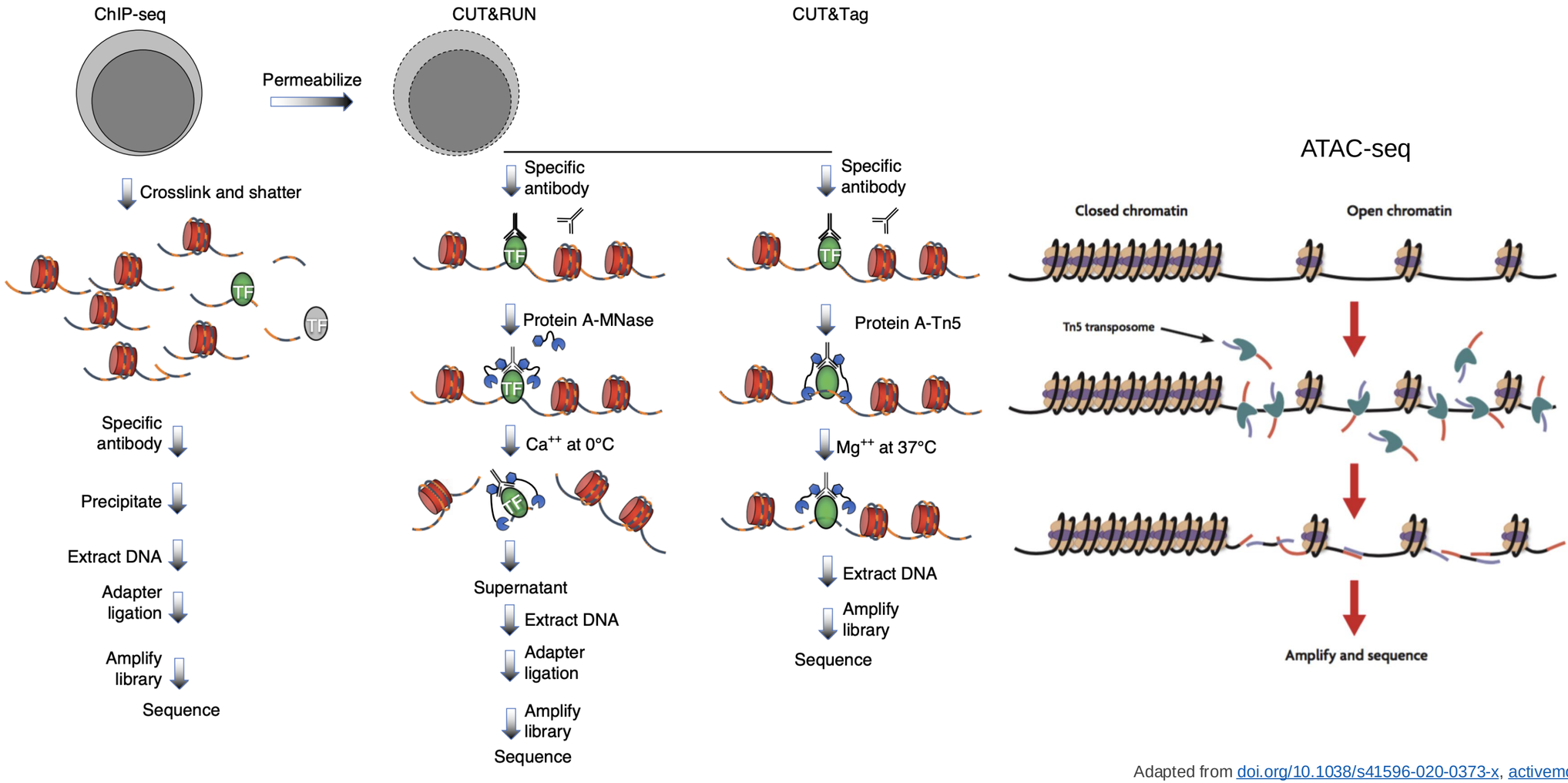


Adapted from doi.org/10.1038/nrg2641

ChIP-PCR vs. ChIP-chip vs. ChIP-Seq

ChIP-PCR	ChIP-chip	ChIP-Seq
• Targeted regions of genome • Regions known beforehand • Cheaper • More time efficient • qPCR can allow quantitative comparisons • qPCR can confirm ChIP successful	• Whole genome (but can profile specific regions) • Microarray-based • 30-100 bp resolution typically • Requires ~a few micrograms DNA • Useful for broad binding	• Whole genome • Next-generation sequencing • Single nucleotide resolution • Only requires ~10 – 50 ng of DNA • Becoming more cost effective • Useful for sharp binding

Alternative methods to ChIP-Seq



Adapted from doi.org/10.1038/s41596-020-0373-x, activemotif.com/blog-atac-seq

Submitting Samples for Sequencing and Analysis

Submitting ChIP-Seq samples for Hartwell sequencing

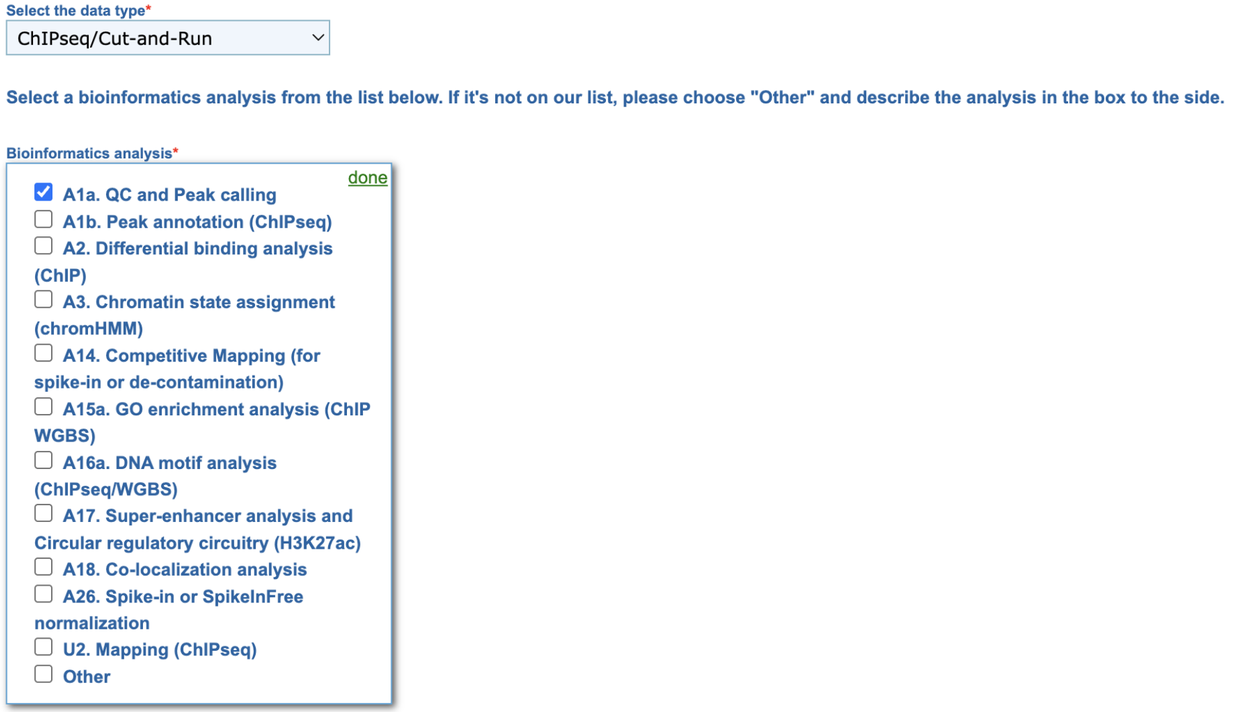
- Submit request for Hartwell sequencing via SRM
- Fill out template spreadsheet to submit (or manually file out table in SRM)
  - Sample info for each well of 96 well plate
- Hartwell pipeline is optimized for 10 ng input DNA per sample, 1 ng input accommodated, 100 pg attempted (submit in 52 uL volume)
- Receive email with SRM order confirmation with QR for sample label
- After sequencing, receive email with fastq file paths, basic sequencing QC

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	**Please use this excel to upload samples for Genome Sequencing Service in SRM2 System**															
3	**Well Location**	**Sample Name**	**Is this a Human Derived Sample?**	**SJ Tissue Bank #**	**SJUID**	**Alternative # (lab, cooperative group, etc.)**	**Submission Material**	**Xenograft?**	**Application**	**Illumina Sequencer**	**Run Type**	**Read Length**	**Molecules Sequenced**	**Reference Genome**	**Please specify Reference Genome**	**User Comments**
4	**A01**								ChIP-seq	NovaSeq	Single End	50 bps	Default for Selected Application			
5	**B01**															
6	**C01**															
7	**D01**															
8	**E01**															
9	**F01**															

Nucleic Acid for Prep and Sequencing
User Made Libraries to be Sequenced

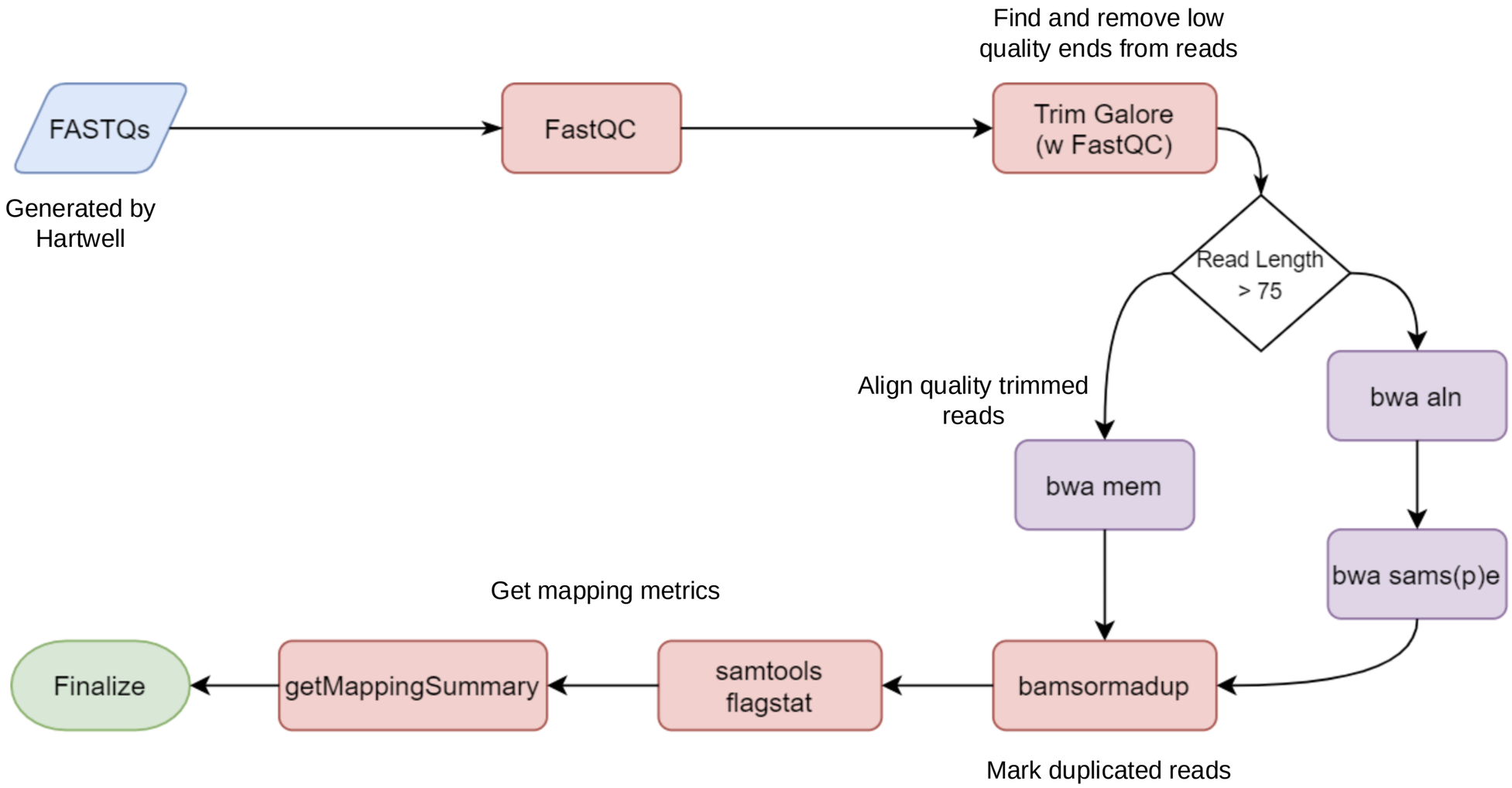
Requesting ChIP-Seq analysis from Center for Applied Bioinformatics (CAB)

- CAB AutoMapper for ChIP-Seq runs automatically after Hartwell sequencing completed
- Receive email when AutoMapper submission is started
- Receive email when AutoMapper run is completed
- AutoMapper pipeline does not include analysis, only alignment
- Must submit new SRM request for analysis
- QC and Peak Calling is standard analysis
- Additional analysis considered customized

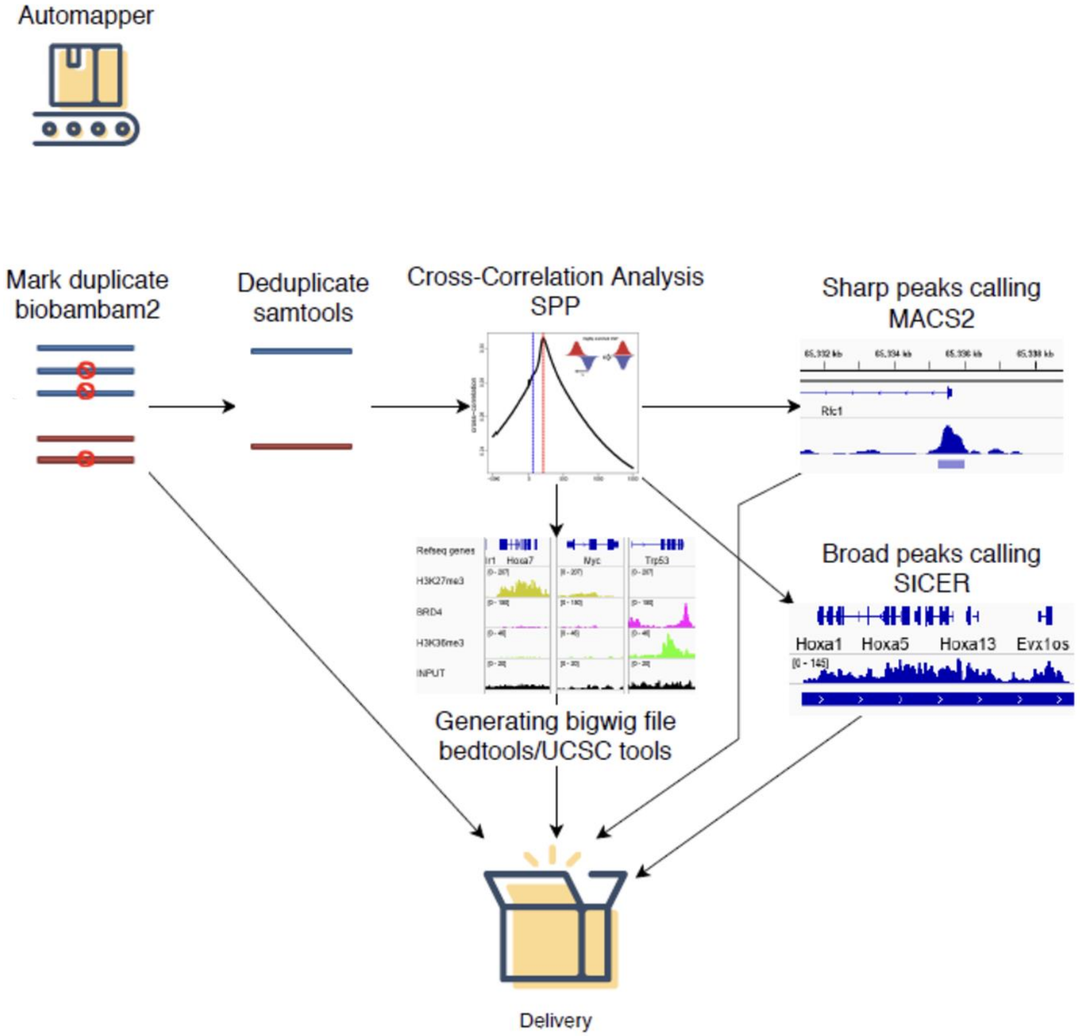


CAB ChIP-Seq AutoMapper Pipeline, QC, and Peak Calling

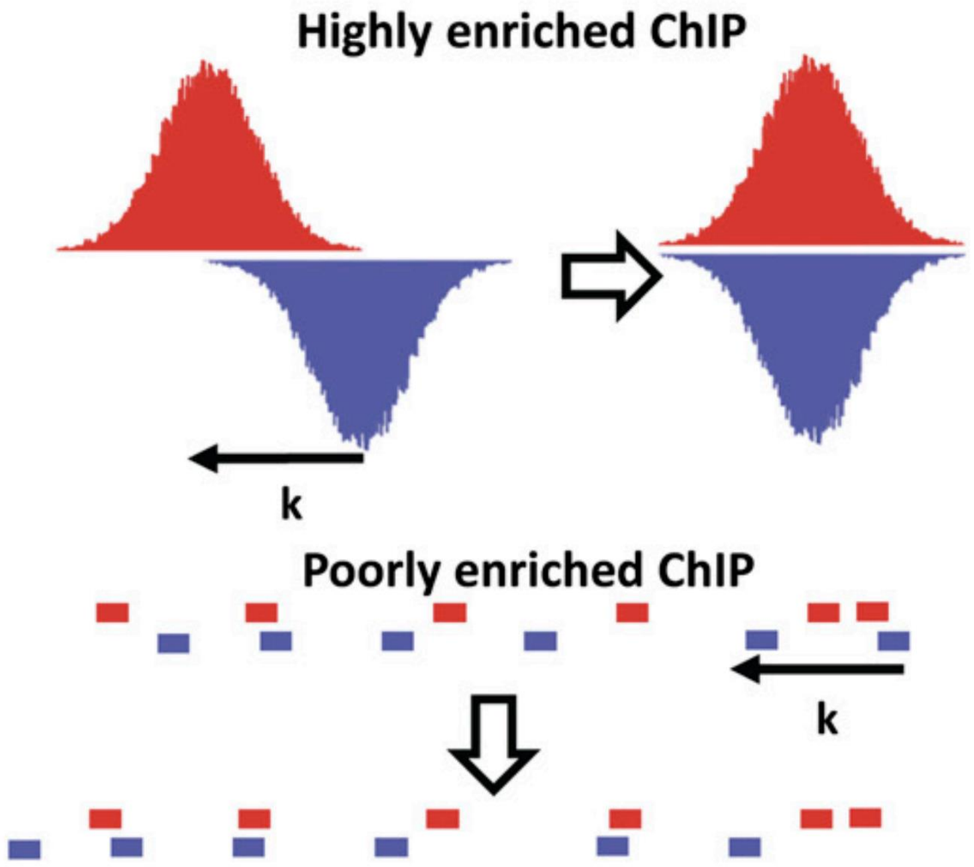
CAB AutoMapper pipeline

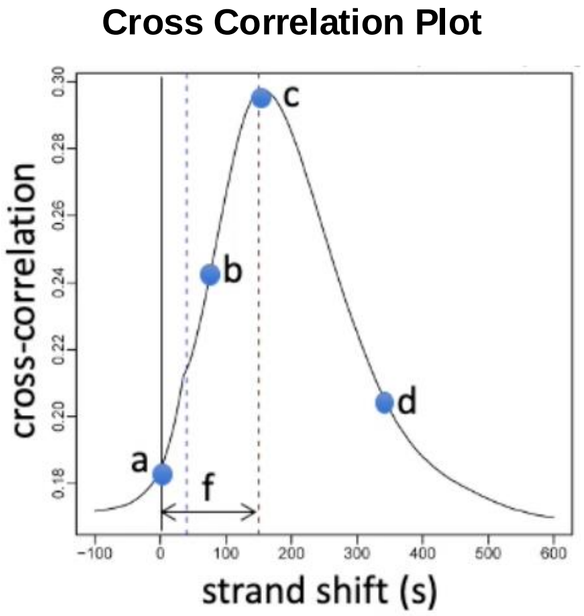


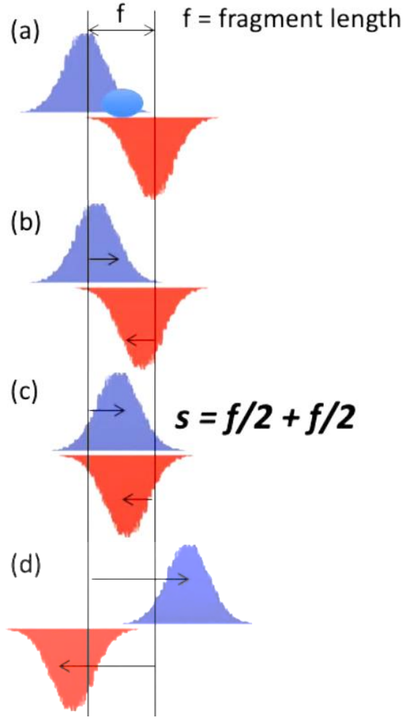
CAB QC and Peak Calling



Cross-correlation analysis indicates quality of sequencing, fragment length, binding sites

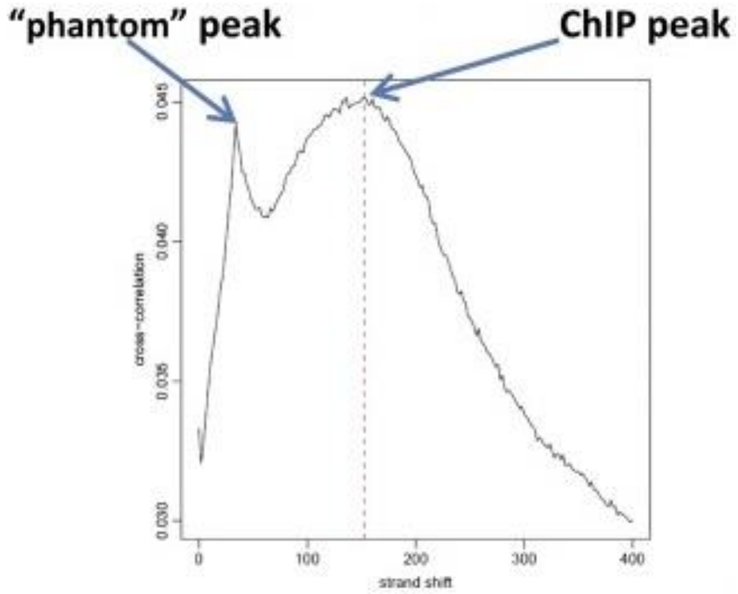


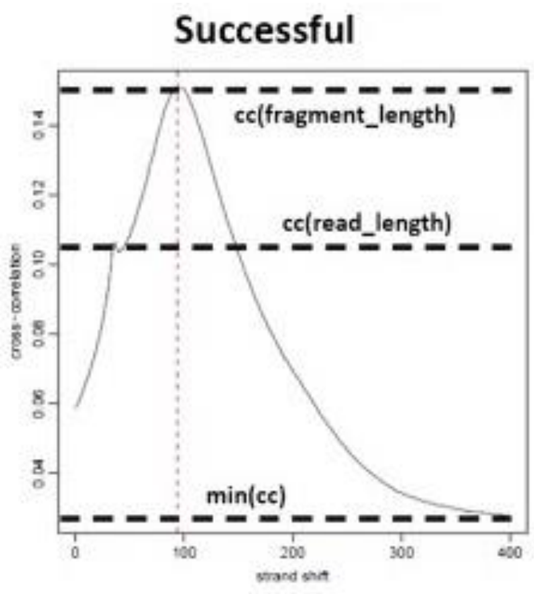


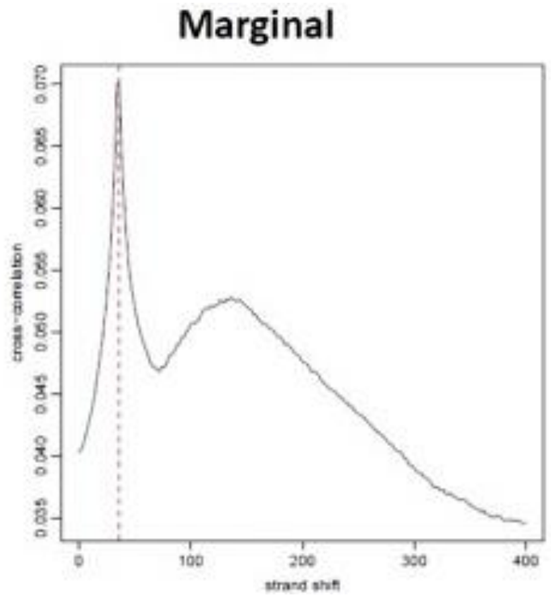


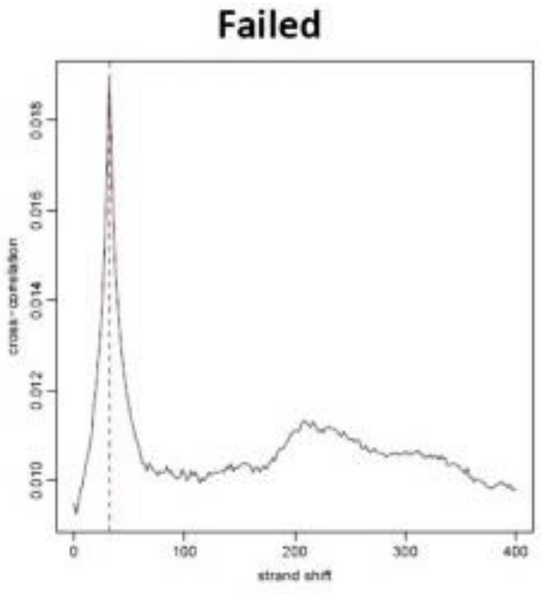
Adapted from doi.org/10.1101/gr.136184.111

Interpreting cross-correlation plots, relative stranded correlation (RSC) values









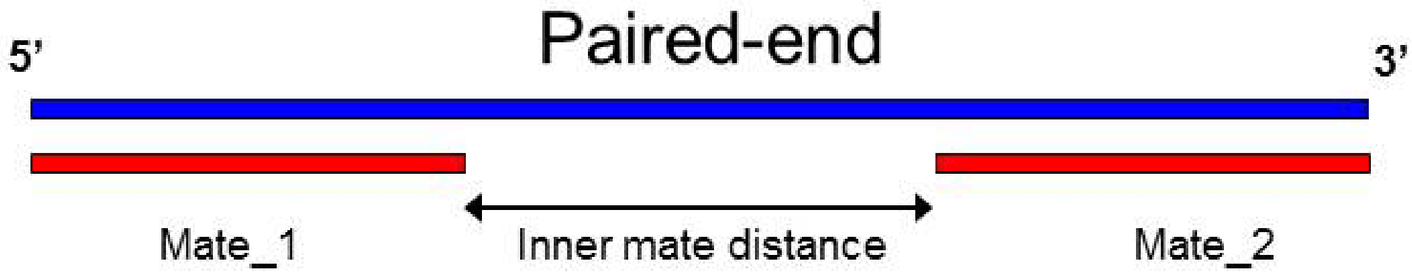
Qtag	-2	-1	0	1	2
RSC	0, 0.25	0.25, 0.5	0.5, 1	1, 1.5	≥1.5

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

Adapted from doi.org/10.1101/gr.136184.111

Cross-correlation analysis is only relevant for single-end sequencing

- If sequencing was paired end, ignore cross-correlation results
- Results are meaningless
- Paired reads will have mate read on opposite strand
- Separated by distance of ~average fragment length



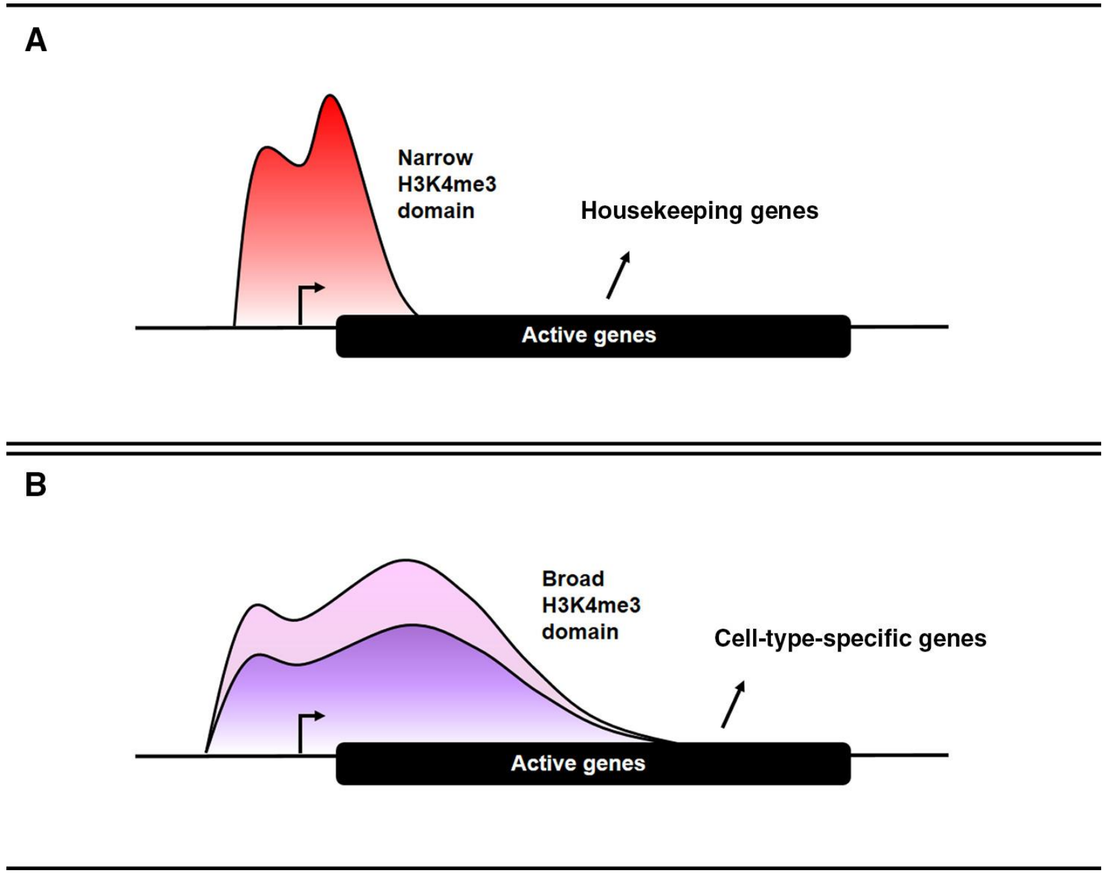
Common quality control metrics for ChIP-seq

- Mapping rate > 80%
- Duplication rate <= 30%
- Total number of unique reads (per ENCODE):
  - > 10M for narrow peaks (point-source data)
  - > 20M for broad peaks
- Fragment size > 100bp
- Qtag > 0 (single-end data only)
  - Wave pattern in cross-correlation plot, RSC invalid
- Visualize data with IGV
  - Check clear peaks
  - Known markers

Qtag	-2	-1	0	1	2
RSC	0, 0.25	0.25, 0.5	0.5, 1	1, 1.5	≥1.5

CAB pipeline can call broad or narrow peaks

- Peak calling identifies regions with enriched protein-DNA interactions
- Narrow peak calling is done with **MACS2**
- Broad peak calling is done with **SICER**
- If known, can tell CAB broad or narrow
- Or choose auto and pipeline will choose method
- Important to use correct method for peak calling

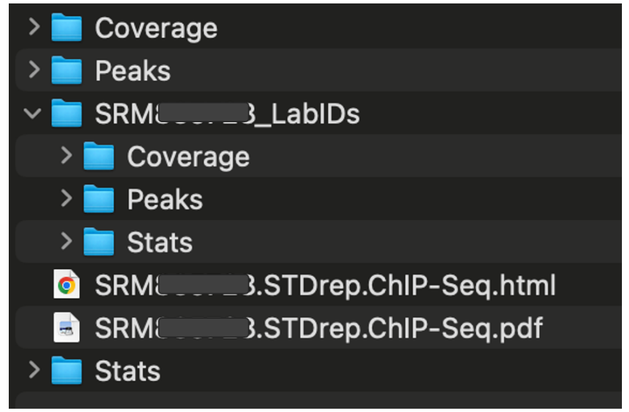


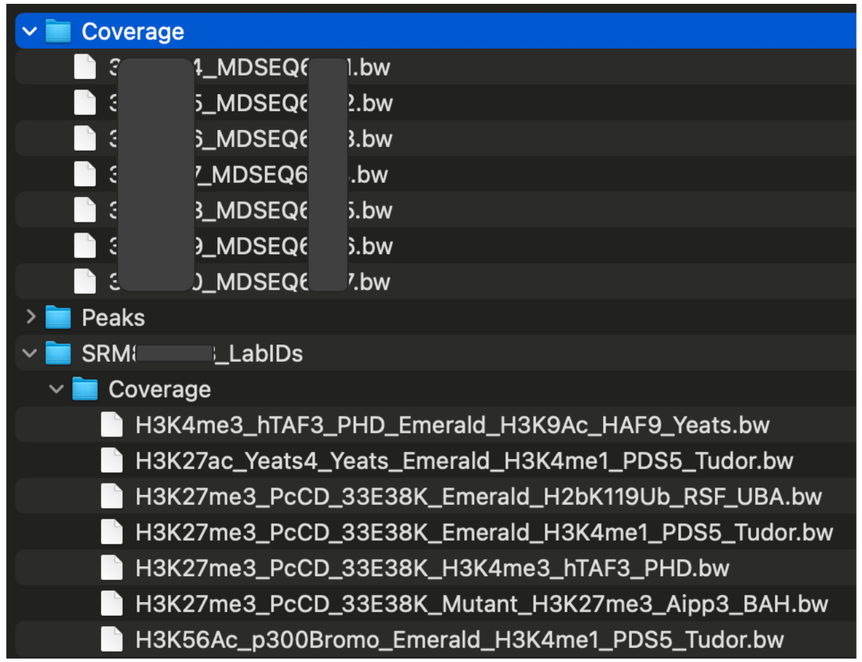
Adapted from doi.org/10.1111/febs.15219

ChIP-Seq Results Folders/Files and Report

Folder structure and file types for QC and peak calling results returned by CAB

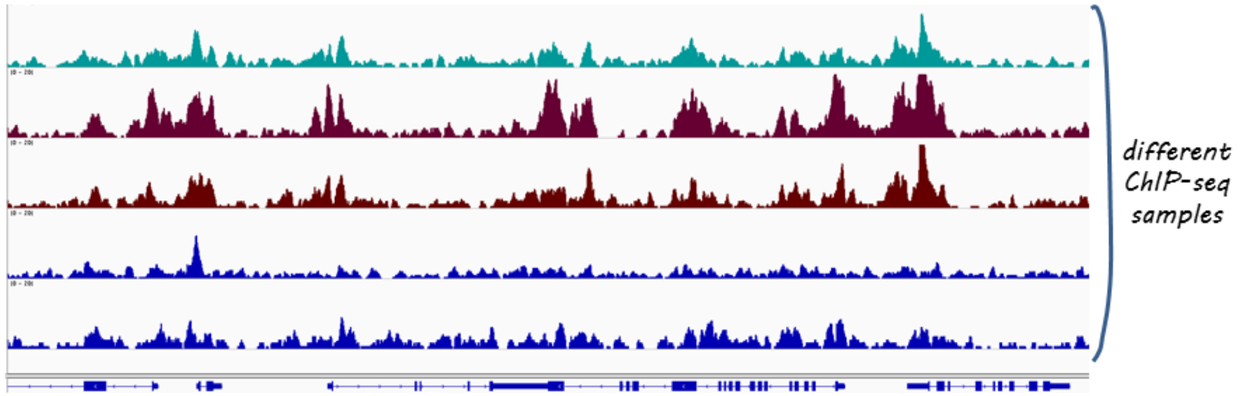
- Four folders: Coverage, Peaks, Stats, SRM#_LabIDs
- Two files: HTML report, PDF report
- Within SRM#_LabIDs: Coverage, Peaks, Stats
- SRM#_LabIDs contains files labeled by ChIP targets
- Coverage folders will contain BigWig files
  - Use these to visualize data in IGV, etc.
- Peaks folders will contain peak calling results
  - Use these to visualize peaks in IGV, etc.
  - Contain “filter” version of each peak files
  - Filter files remove ENCODE “blacklist” regions
- Stats folders will contain four QC-related files
  - Mapping metrics, QC metrics
  - Cross-correlation analysis table
  - Cross-correlation plot





Common options for visualizing data

- Integrative Genomics Viewer (IGV) - download desktop app or use web browser
- UCSC Genome Browser – select reference genome, upload your files (custom tracks)
- St. Jude ProteinPaint – select reference genome, upload your files
- Files to view often includes BigWig, broadPeak, narrowPeak, bam files with expression data (e.g. RNA-Seq)



Adapted from hbctraining.github.io/Intro-to-ChIPseq

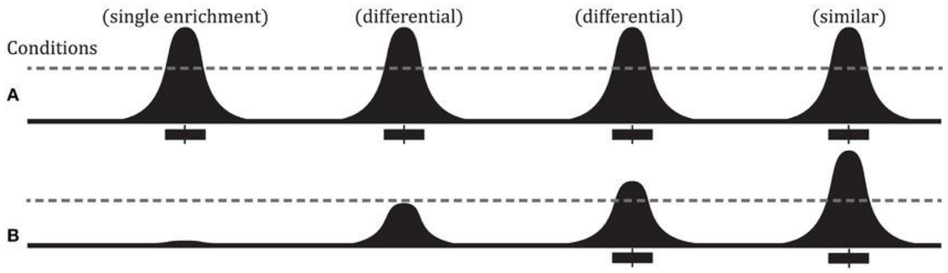
Example ChIP-Seq report from CAB

- Interactive HTML report, static PDF report
- HTML, PDF examples of reports available on CAB’s wiki page
- Walk through HTML example
- Download HTML to follow along here

Additional ChIP-Seq Analyses from CAB

Examples of other analyses that can be requested from CAB: Differential binding

- Additional analyses are considered customized or collaborative
- Require additional SRM requests
- For example:
  - Differential binding site analysis
  - Chromatin state assignment
- Differential binding site analysis can identify differences in enrichment peaks across conditions



Select the data type*

ChIPseq/Cut-and-Run

Select a bioinformatics analysis from the list below. If it's not on our list, please choose "Other" and describe the analysis in the box to the side.

Bioinformatics analysis*

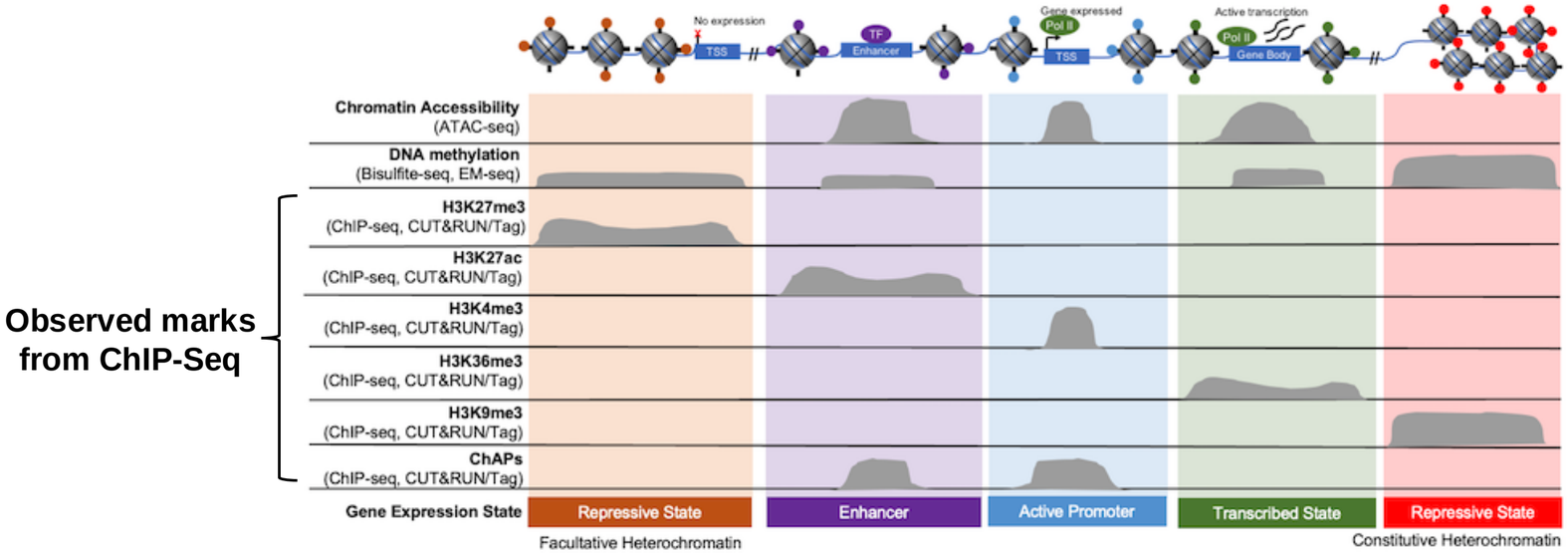
done

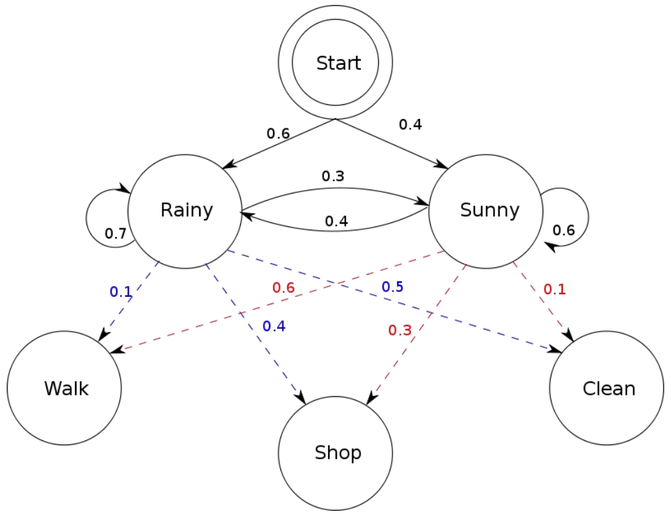
- A1a. QC and Peak calling
- A1b. Peak annotation (ChIPseq)
- A2. Differential binding analysis (ChIP)
- A3. Chromatin state assignment (chromHMM)
- A14. Competitive Mapping (for spike-in or de-contamination)
- A15a. GO enrichment analysis (ChIP WGBS)
- A16a. DNA motif analysis (ChIPseq/WGBS)
- A17. Super-enhancer analysis and Circular regulatory circuitry (H3K27ac)
- A18. Co-localization analysis
- A26. Spike-in or SpikeInFree normalization
- U2. Mapping (ChIPseq)
- Other

Adapted from doi.org/10.3389/fgene.2015.00169

Examples of other analyses that can be requested from CAB: Chromatin state assignment

- Chromatin state assignment uses ChromHMM to predict state
  - Based on Hidden Markov Model (HMM), models presence or absence of chromatin marks, annotates genome
- Hidden Markov Models are probabilistic, predict “outcomes” based on observable parameters
- Includes “hidden” states that influence outcome but aren’t “observable”
- In this case, “outcome” is chromatin state, predicted based on observed chromatin mark patterns





Credit: epicypher.com

Questions?