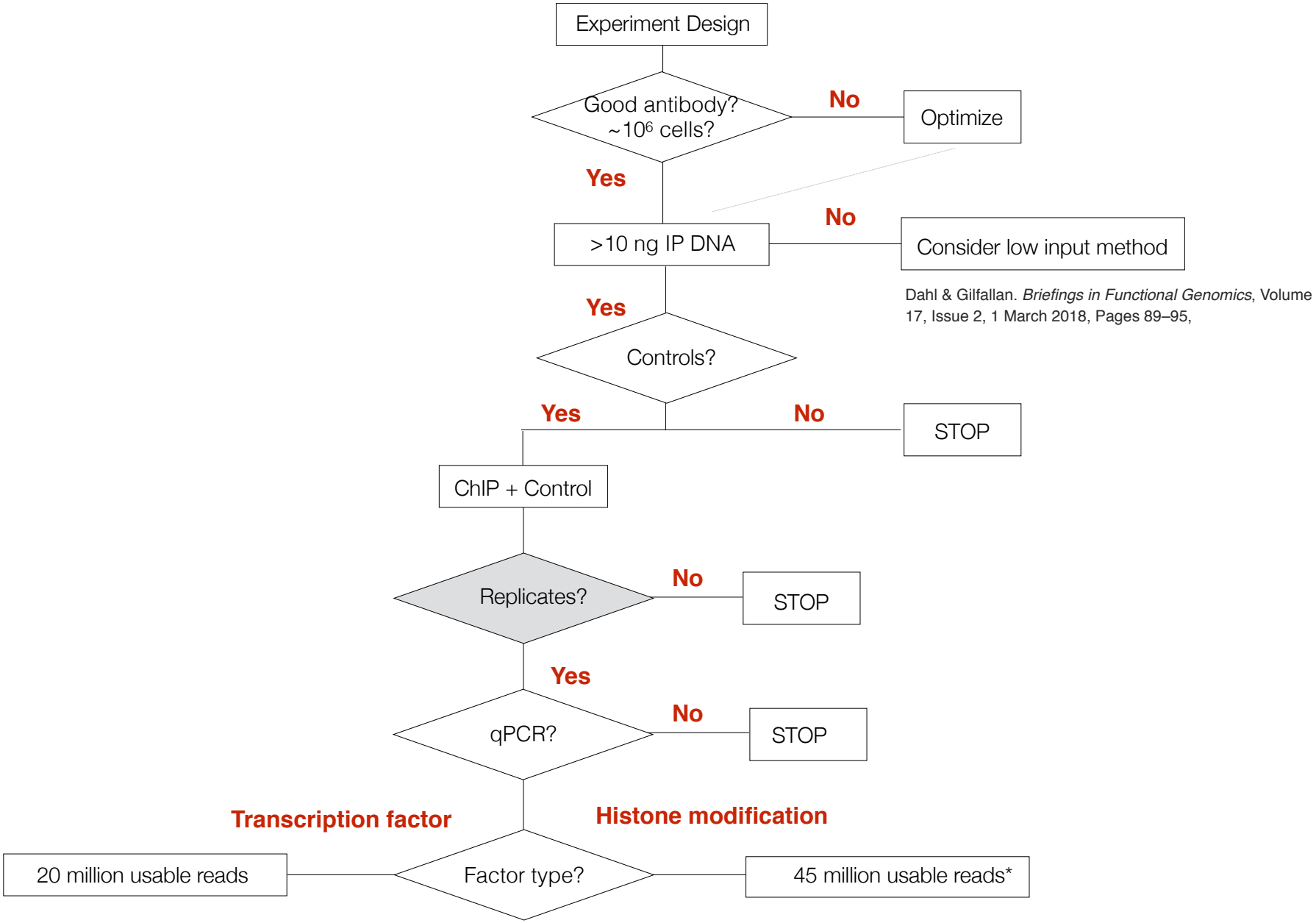
ChIP-seq Analysis
Workflow and Troubleshooting

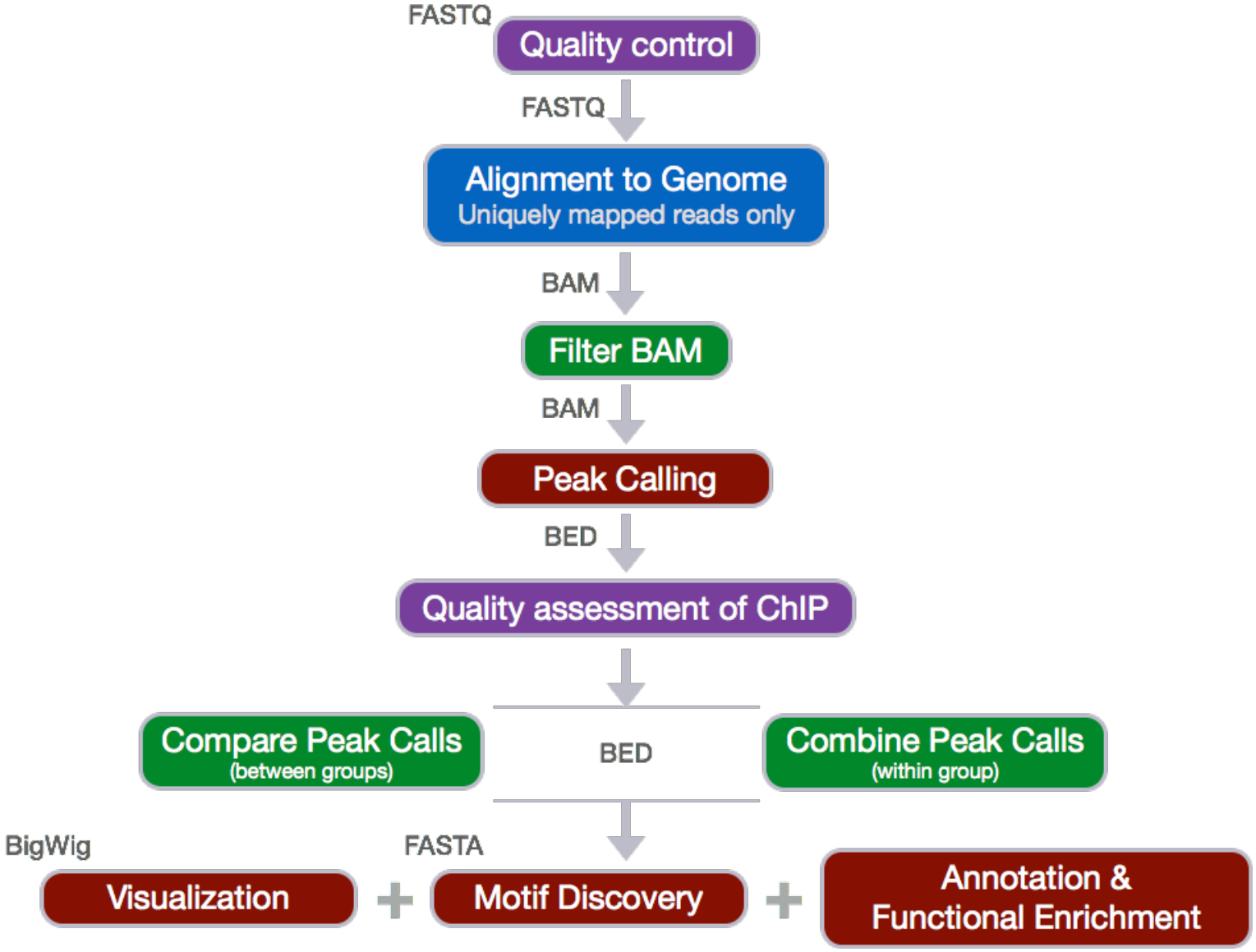
Before the sequencer…

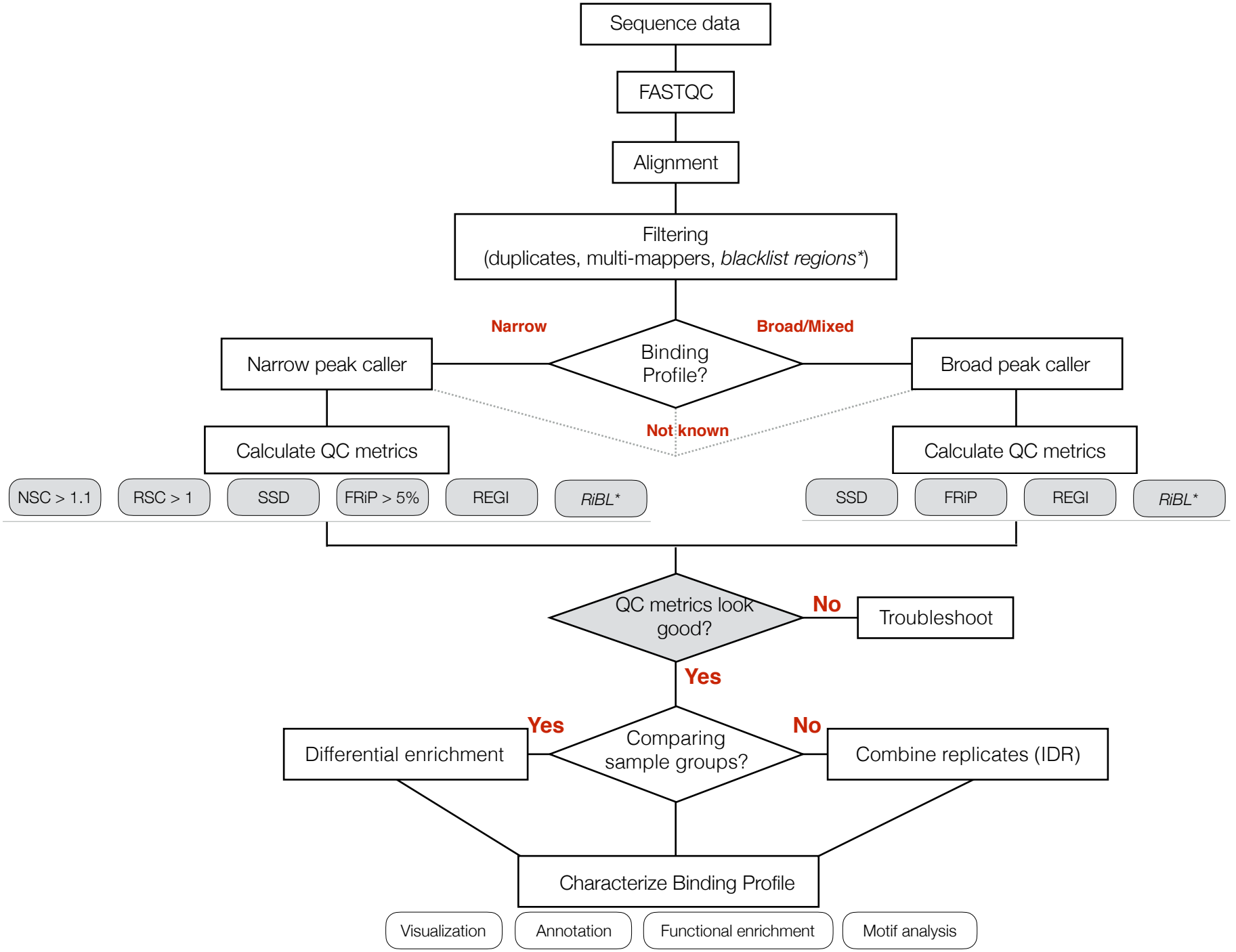


Dahl & Gilfallan. *Briefings in Functional Genomics*, Volume 17, Issue 2, 1 March 2018, Pages 89–95,

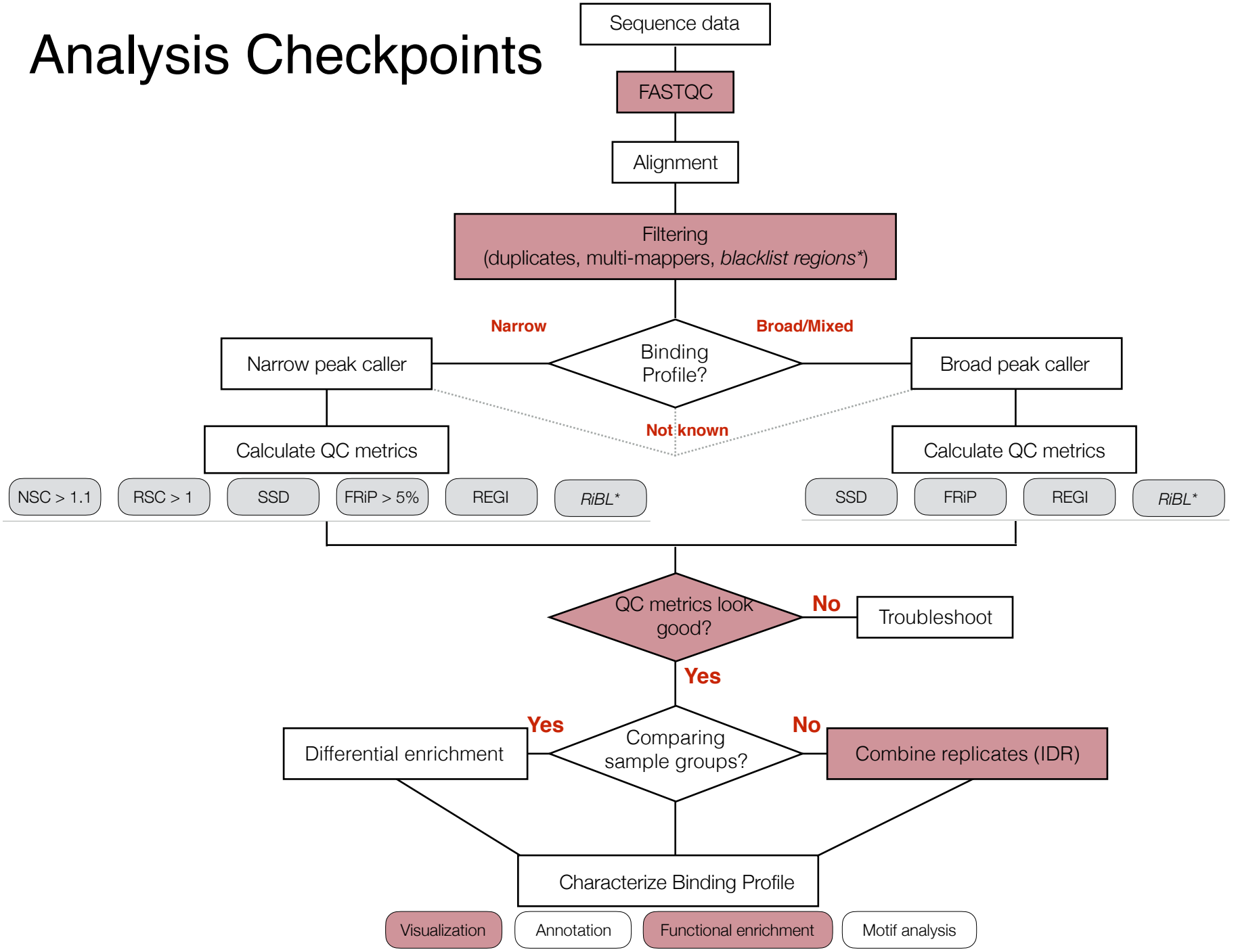
After sequencing…

ChIP-seq Workflow

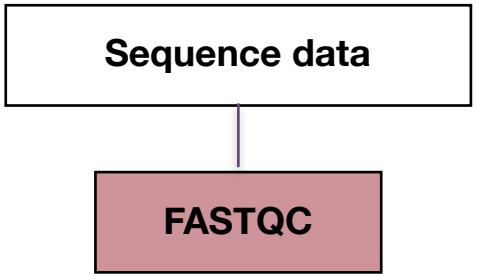



Sequence data
FASTQC
Alignment
Filtering
(duplicates, multi-mappers, *blacklist regions**)
Narrow
Broad/Mixed
Binding Profile?
Narrow peak caller
Broad peak caller
Not known
Calculate QC metrics
Calculate QC metrics
NSC > 1.1
RSC > 1
SSD
FRiP > 5%
REGI
*RiBL**
SSD
FRiP
REGI
*RiBL**
QC metrics look good?
No
Troubleshoot
Yes
Yes
Comparing sample groups?
No
Differential enrichment
Combine replicates (IDR)
Characterize Binding Profile
Visualization
Annotation
Functional enrichment
Motif analysis

Analysis Checkpoints



Quality Checks: Raw Data

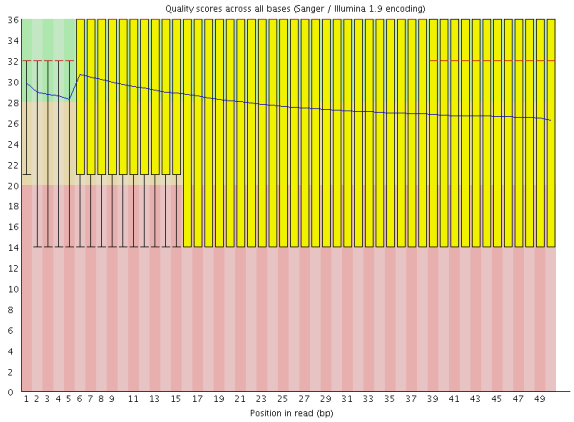


The quality checks at this stage in the workflow include:

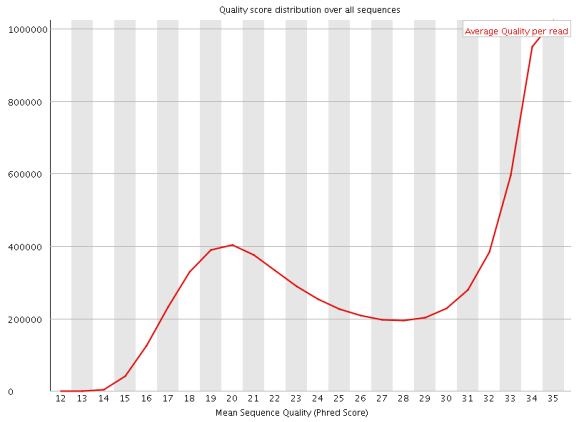
1. Checking the **quality of the base calls** to ensure that there were no issues during sequencing
2. Examining the reads to ensure their **quality metrics adhere to our expectations** for our experiment
3. Exploring reads for **contamination**

Quality Checks: Raw Data

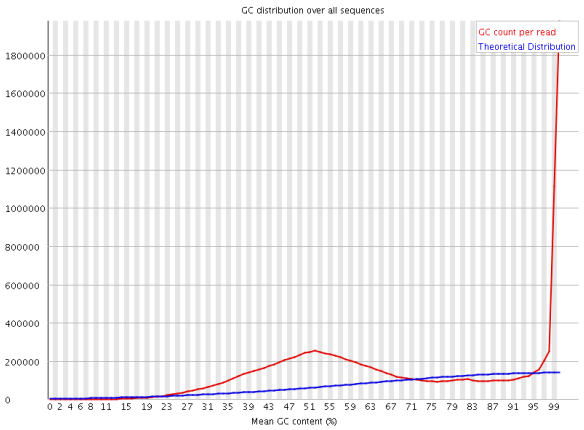
Per base sequence quality



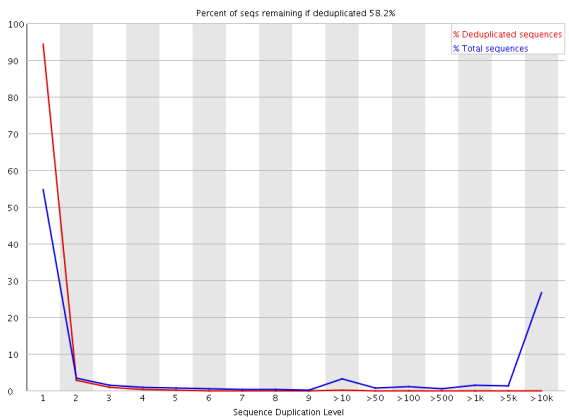
Per sequence quality scores



Per sequence GC content



Sequence Duplication Levels



Overrepresented sequences

Sequence	Count	Percentage	Possible Sou
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGG	1838931	25.27061807325126	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGA	22246	0.3057048740042707	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAG	19143	0.26306340029954844	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAGG	14083	0.1935288025084125	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGAAAAAAAAAAAAAAAAAAAAAA	13303	0.18281003051689354	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAAAAAA	12912	0.17743690250576033	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAAAAAAAAAAAAAAAAA	11561	0.1588714397358345	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGACG	11428	0.15704375169112678	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGAAAAAAAAAAAAAAAAAA	11152	0.1532509554479739	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGACGG	9922	0.1363482765382709	No Hit
GGGGGGGGGGGGGGGGGGGGGGGAAAAAAAAAAAAAAAAAAAAAAAA	9693	0.1332013550176839	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAA	9340	0.12835042359075288	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAAAAA	8557	0.11759042555311268	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAAAA	8156	0.11207987738824203	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAGGGG	7471	0.10266659685722856	No Hit
GGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGGAGGGG	7294	0.10023426013607618	No Hit

Quality Checks: Raw Data

Troubleshooting raw data quality problems:

- Low sequence quality reads

  - loss of signal in later cycles, technical problems with sequencer

- Unexpected %GC for organism

  - contaminating sequences: different species, adapters, vector

- High level of sequence duplications

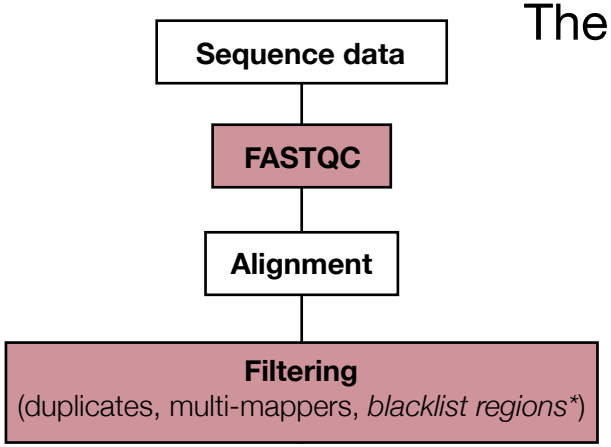
  - low complexity library, too many cycles of PCR amplification / too little starting material, biological duplicates

- Over-represented sequences can be biologically significant or represent bias

  - sequences that represent binding sites

  - contaminating sequences: adapters, vector

Quality Checks: Aligned Data



The quality checks at this stage in the workflow include:

1. Checking the **total percent of reads aligning** to the genome
2. Determine the **percent of duplicate reads**
3. Determining the **percent uniquely mapping** reads
4. Identify percent of reads **mapping in blacklist regions**
5. Checking percent of paired-end reads that are **properly paired**

Quality Checks: Aligned Data

Troubleshooting aligned data quality problems:

- Low percentage (< 70%) of **total reads** aligned

  - poor quality reads, contaminating sequences, inappropriate alignment parameters chosen, poor quality reference genome

- Low percentage (< 60%) of **uniquely** aligning reads

  - high number of multi-mappers, high number of duplicates (20-30% duplication rate)

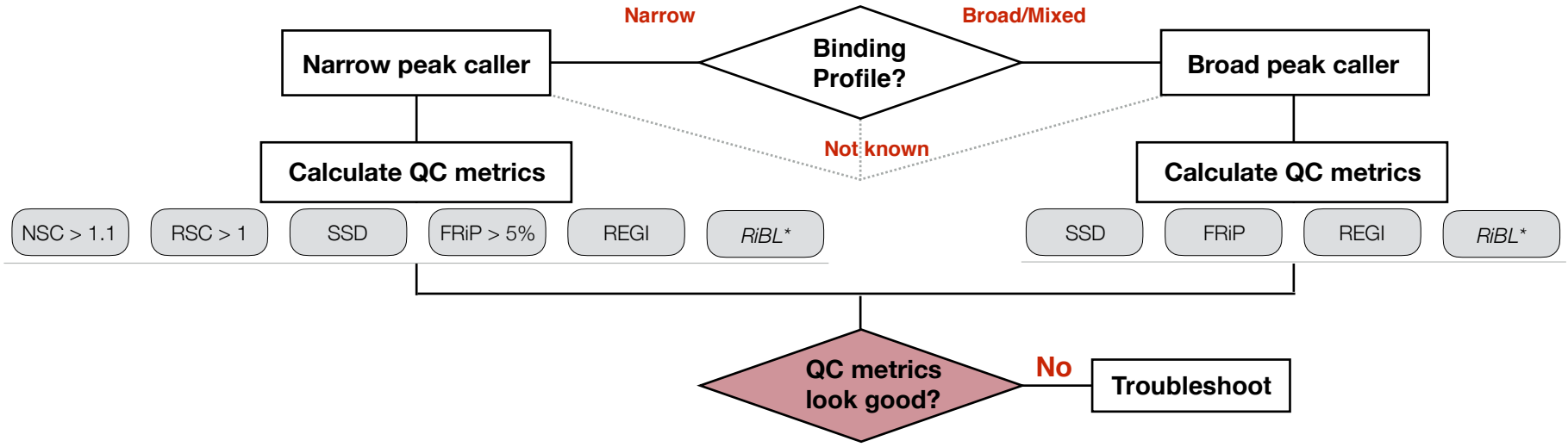
- High percentage (> 10%) of reads mapping in **blacklist regions**

- For paired-end data: large number of reads **not properly paired**

  - poor quality reads

Even if your samples meet the suggested thresholds, **always filter** *for (duplicates), multi-mappers and blacklist regions.*

Quality Checks: Peak Calling



Quality Checks: Peak Calling

The quality checks at this stage in the workflow include:

1. Evaluate **degree of enrichment** (FRiP, coverage plots, SSD)
2. Evaluating **signal-to-noise** using strand cross-correlation based metrics (NSC, RSC, fragment length)
3. Evaluate enrichment **within specific genomic regions**/features, and within **known artifact regions**
4. **Compare and contrast** measures described above based on **thresholds** and/or **what you anticipate** for the binding profile

Quality Checks: Peak Calling

Troubleshooting ChIP quality problems:

- Low FRiP (< 5%), Low SSD scores

  - poor enrichment due to poor antibody, the protein of interest binds few regions

- Low **signal-to-noise** (RSC < 0.8, NSC < 1.1), bad fragment length estimates, low diversity of depth

  - IP did not work, high cross-reactivity of antibody with other proteins

- No enrichment in anticipated **genomic features**

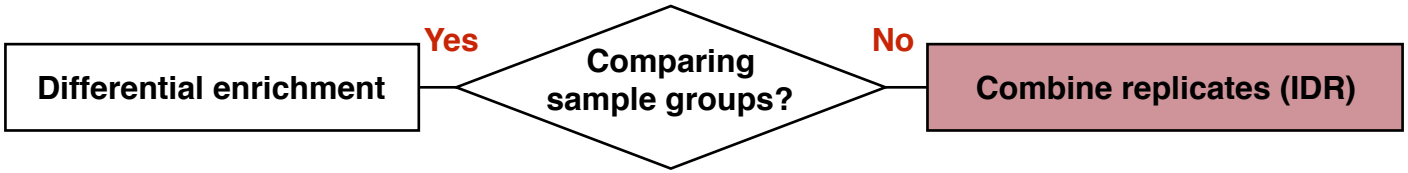
  - experiment did not work, other regions identify interesting novel behaviors

- High RiBL (> 10%)

  - repeat regions or other artifact regions displaying artificial enrichment, this can drive up the SSD score

Evaluate these metrics collectively for each sample.

Quality Checks: Handling Replicates



The quality checks at this stage in the workflow include:

1. Using bedtools to perform a crude **comparison of peak calls** between replicates

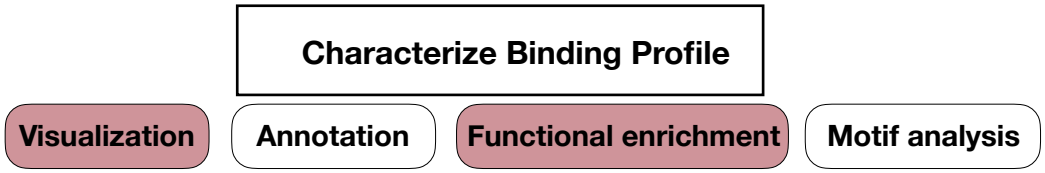
2. **Statistical evaluation** of reproducibility between replicates using the IDR pipeline

Quality Checks: Handling Replicates

Replicate Evaluation Goals:

- **Ensure that there is concordance in the peaks being called across all replicates in a sample group**
- For IDR analysis use a more liberal threshold for peak calling to increase the search space
- Focus on peaks that meet the IDR < 0.05 threshold for downstream analysis (visualization and functional analysis)
- Pseudo-replicate analysis for reporting in a publication

Quality Checks: Characterizing the binding profile



The quality checks at this stage in the workflow include:

1. Using deepTools to **visualize the data** and evaluate enrichment in specific regions.

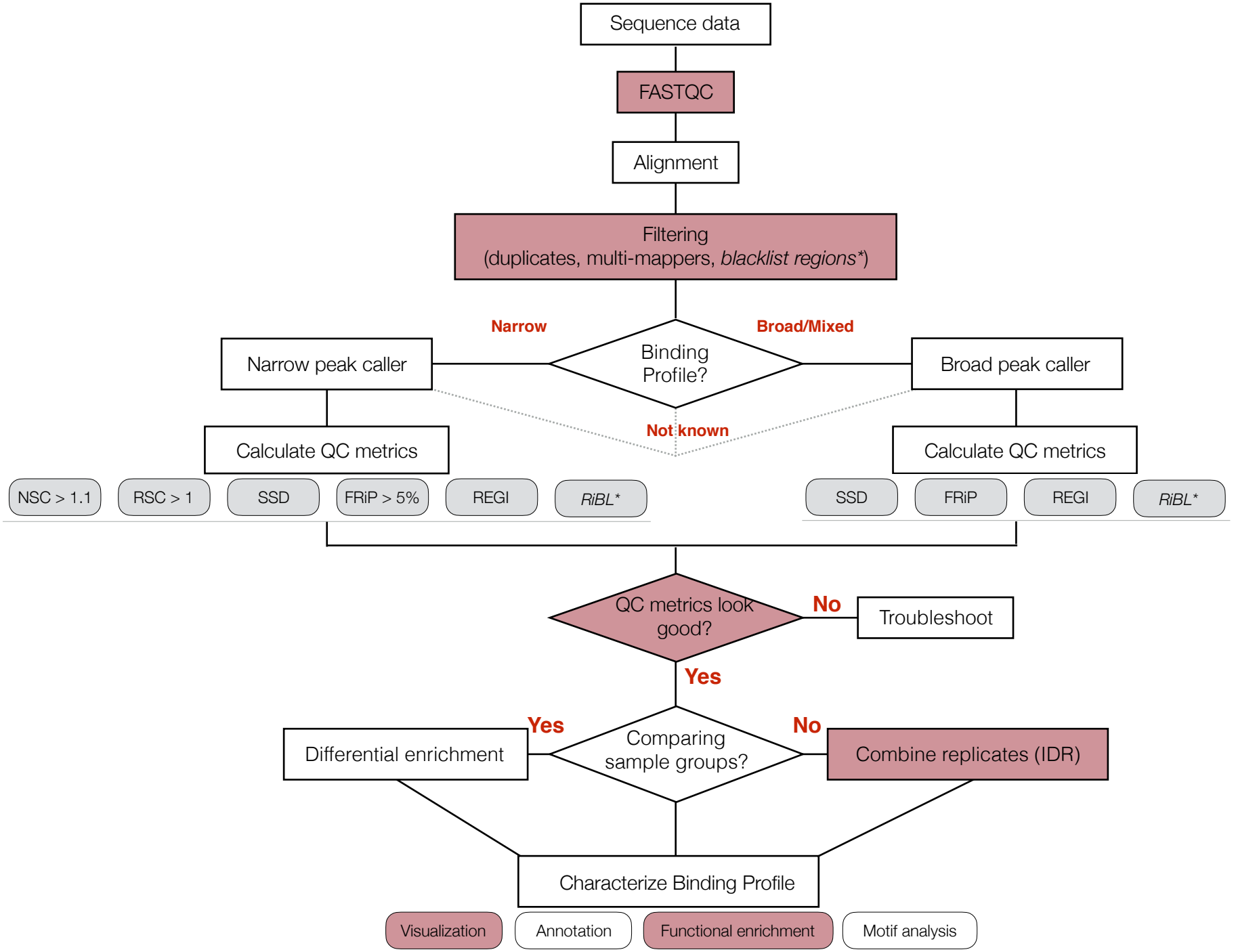
2. Look at **specific regions of interest** individually (anticipated target genes). Overlay relevant public datasets (using IGV or a genome viewer).

3. Evaluate target genes and assess **functional analysis** to see if there is any biological relevance.

Quality Checks: Characterizing the binding profile

Troubleshooting problems:

- Do we see enrichment associated with genomic features we anticipated?
- Does the visual inspection validate what we know/identified from statistical analysis?
- Do these target genes collectively represent specific pathway(s)? Is there significant over-representation of certain biological processes?
- Is this all relevant based on what I know about my protein of interest?


Sequence data
FASTQC
Alignment
Filtering
(duplicates, multi-mappers, *blacklist regions**)
Narrow
Broad/Mixed
Binding Profile?
Narrow peak caller
Broad peak caller
Not known
Calculate QC metrics
Calculate QC metrics
NSC > 1.1
RSC > 1
SSD
FRiP > 5%
REGI
*RiBL**
SSD
FRiP
REGI
*RiBL**
QC metrics look good?
No
Troubleshoot
Yes
Yes
Comparing sample groups?
No
Differential enrichment
Combine replicates (IDR)
Characterize Binding Profile
Visualization
Annotation
Functional enrichment
Motif analysis

ChIP-seq Workflow