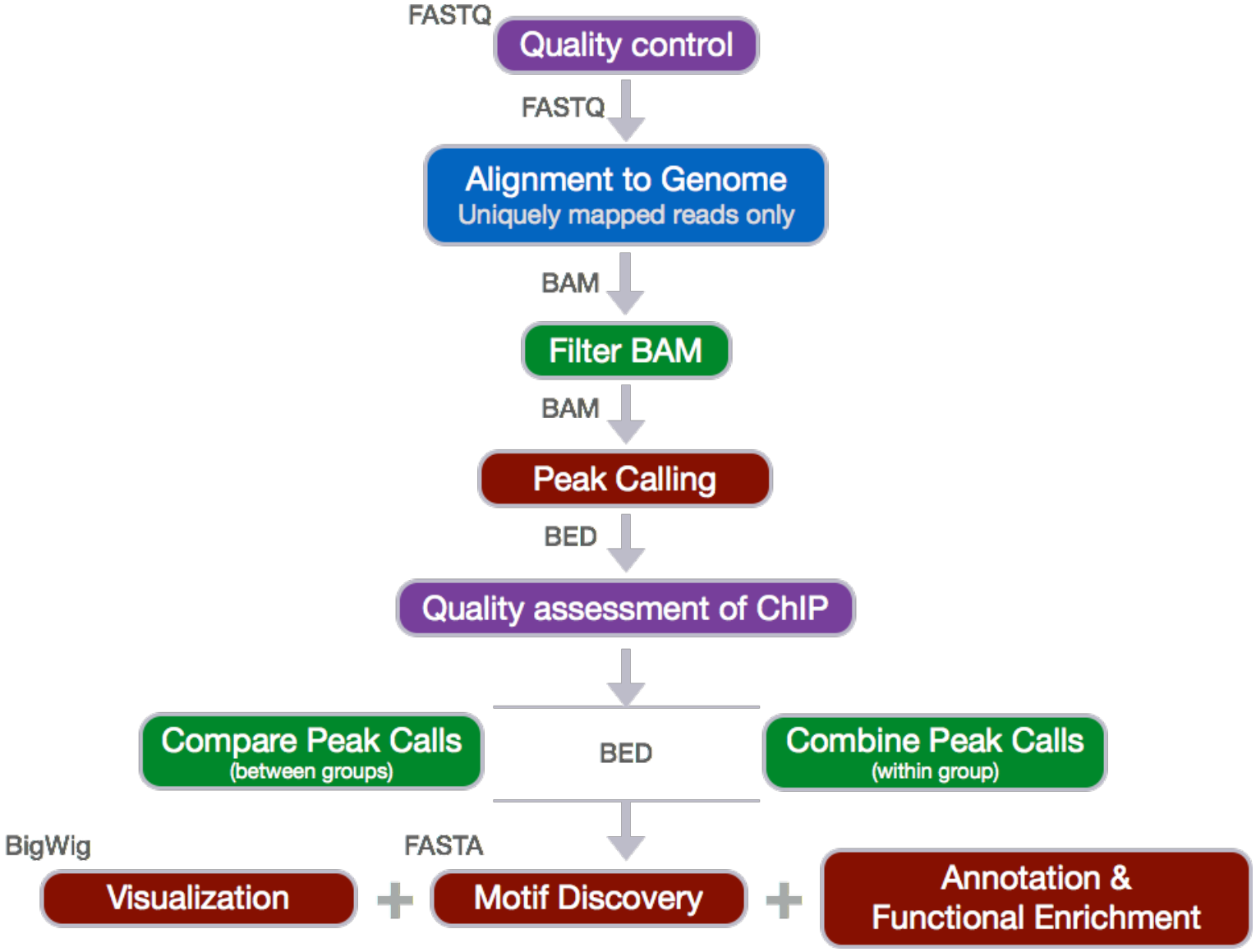
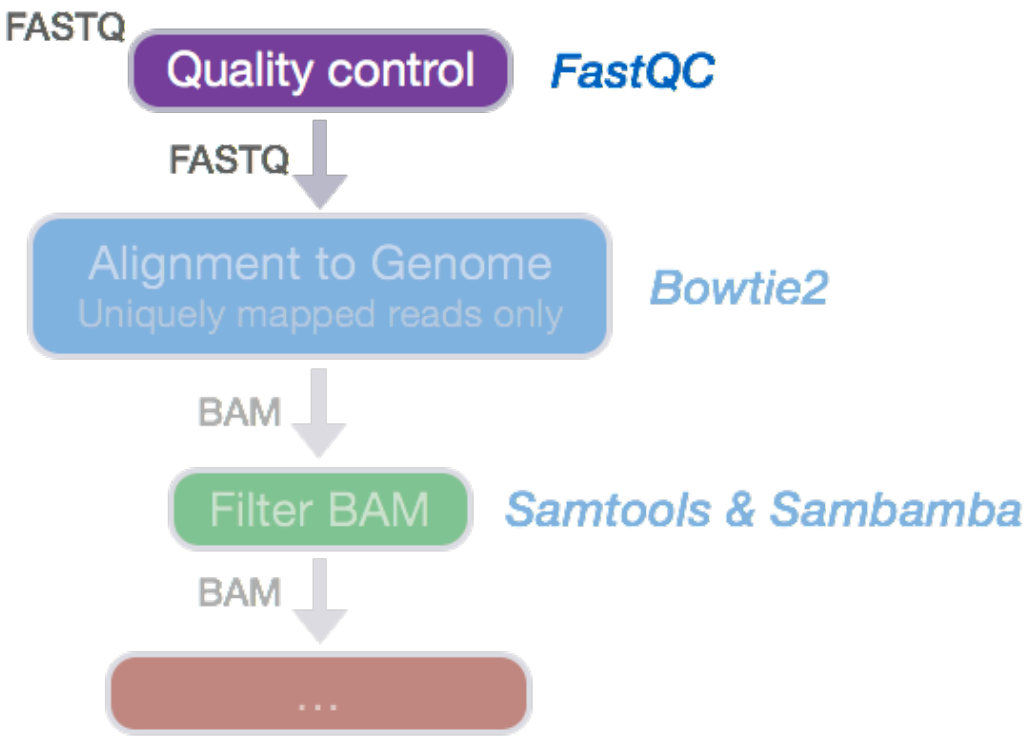
ChIP-seq Analysis
Workflow and Troubleshooting

ChIP-seq Workflow



Quality Checks: Raw Data



Quality Checks: Raw Data

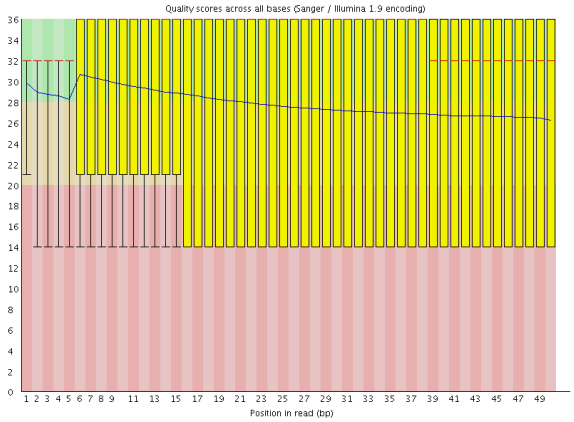
All NGS analyses require that the **quality of the raw data** is assessed prior to any downstream analysis.

The quality checks at this stage in the workflow include:

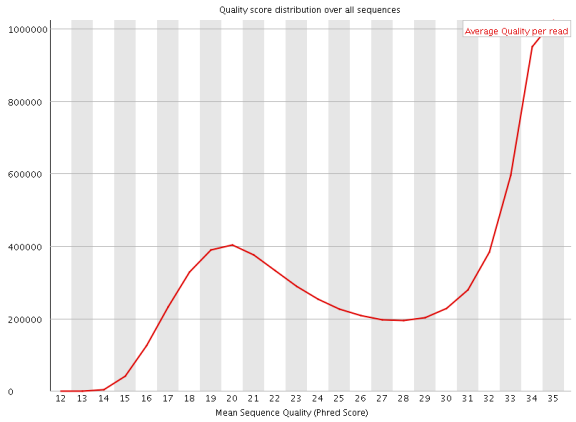
1. Checking the **quality of the base calls** to ensure that there were no issues during sequencing
2. Examining the reads to ensure their **quality metrics adhere to our expectations** for our experiment
3. Exploring reads for **contamination**

Quality Checks: Raw Data

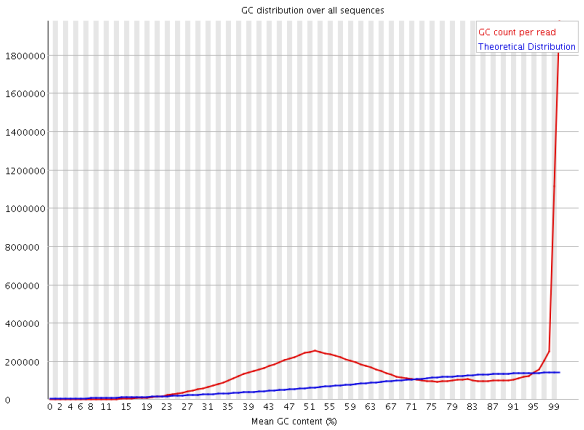
Per base sequence quality



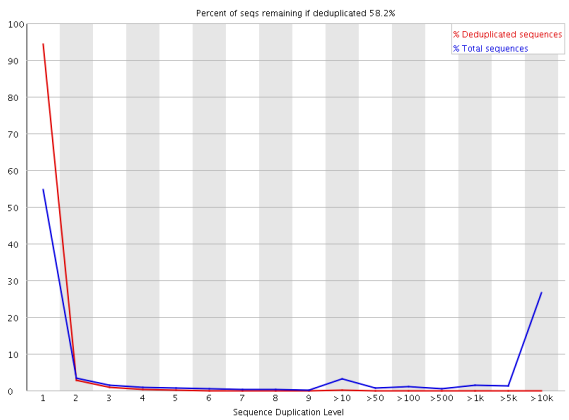
Per sequence quality scores



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

- Over-represented sequences more than 1-2%, unless expected based on experimental design

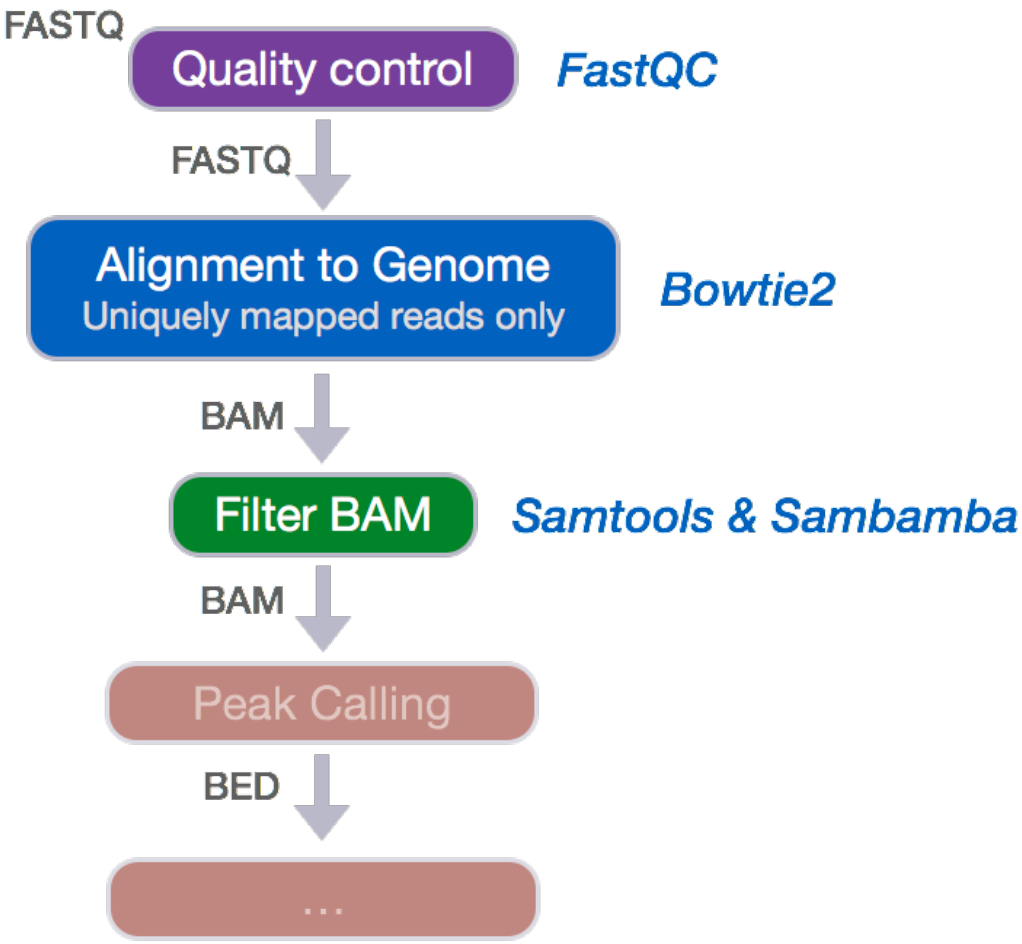
  - contaminating sequences: adapters, vector

Quality Checks: Raw Data

Raw Data QC Goals:

- **Identify sequencing problems and determine whether there is a need to contact the sequencing facility**
- Identify over-represented contaminating sequences
- Gain insight into library complexity
- Ensure organism is properly represented by %GC content

Quality Checks: Aligned Data



Quality Checks: Aligned Data

Evaluating the **quality of the aligned data** can give important information about the quality of the library. The quality checks at this stage in the workflow include:

1. Checking the total percent of reads aligning to the genome
2. Examining the total number of reads aligning to each sample
3. Determining the percent uniquely mapping reads
4. Checking percent of paired-end reads that are properly paired

Quality Checks: Aligned Data

Troubleshooting aligned data quality problems:

- Low percentage (< 70%) of reads aligned

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

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

  - low number of total reads aligning, high number of multi-mappers, high number of duplicates due to over-amplification

- Large differences in sequencing depth between samples

  - library prep / sequencing

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

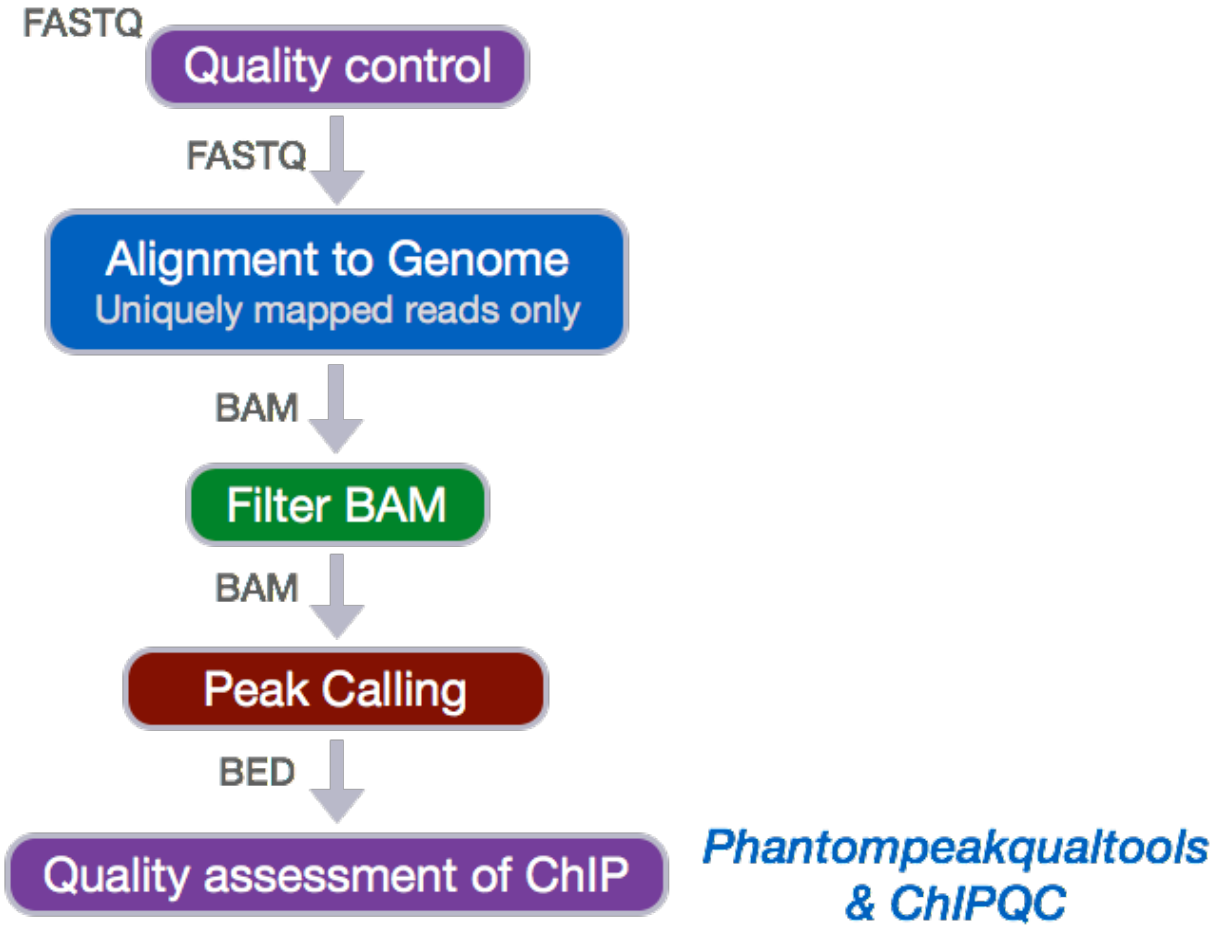
  - poor quality reads

Quality Checks: Aligned Data

Aligned Data QC Goals:

- **Ensure the library depth and percentage of reads mapping to each sample is similar**
- Evaluate read mapping metrics for multi-mapping and duplicates before filtering
- Identify poor alignment parameters or low quality library

Quality Checks: Peak Calling



Quality Checks: Peak Calling

Evaluating the **quality of the peak calls** can give important information about the **signal to noise ratio**, and the distribution of signal. The quality checks at this stage in the workflow include:

1. Using cross-correlation to compute NSC and RSC metrics
2. Evaluate the distribution of reads across the genome, within specific genomic regions/features, and within known artefact regions
3. Evaluate the distribution of reads across the whole genome

Metrics based on cross correlation

- **Normalized strand cross-correlation coefficient (NSC):**
  - Minimum value: 1
  - Critical threshold: 1.1
- **Relative strand cross-correlation coefficient (RSC):**
  - Minimum value: 0
  - Critical threshold: 1
- Low scores indicate low signal to noise
  - Failed ChIP, poor sequence quality (leading to mismapping), inadequate sequencing depth
  - OR factor only binds a few sites

Metrics based on read distribution

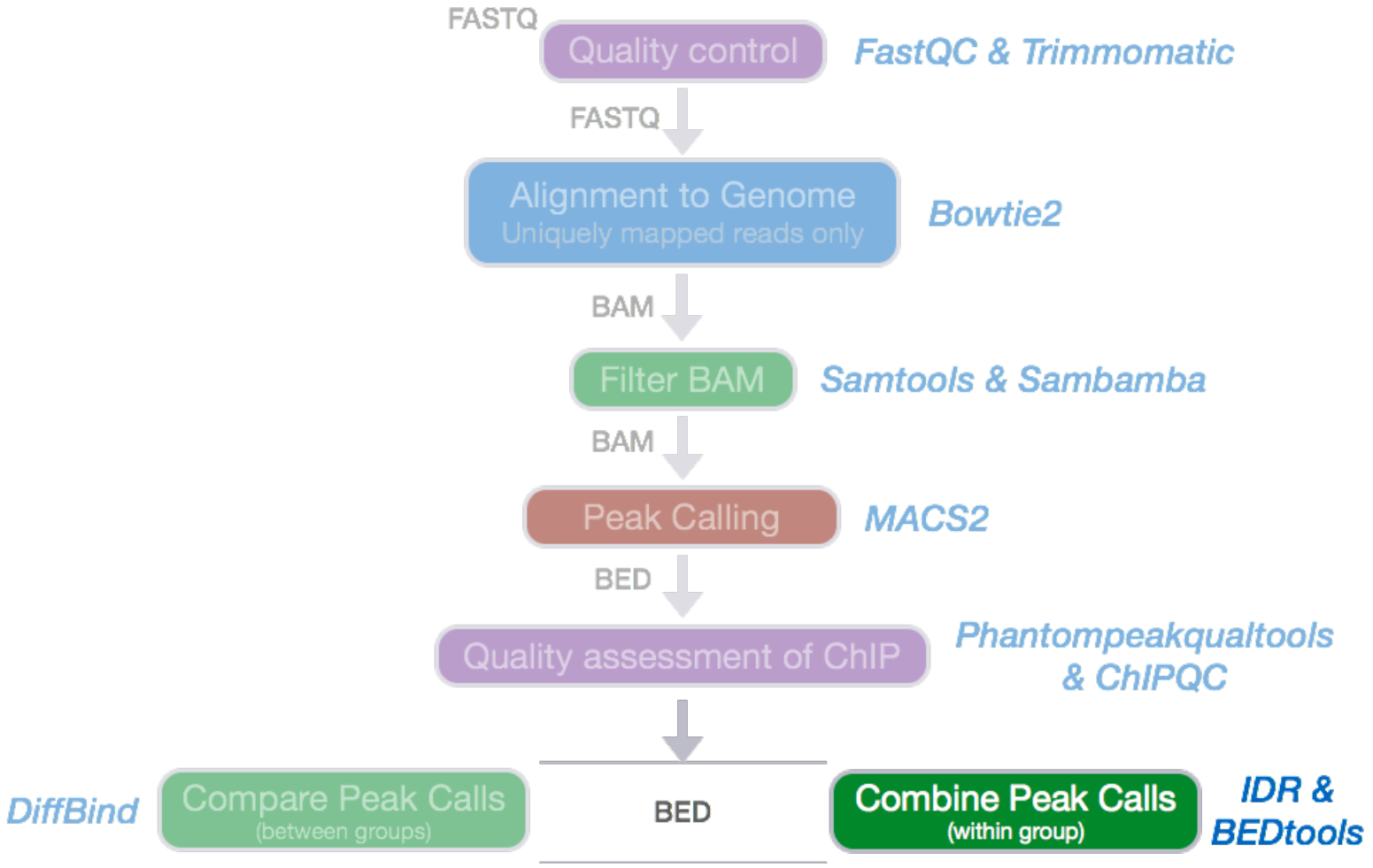
- **Fraction of reads in peaks (FRiP):**
  - For a typical TF experiment expect > 5%; can vary with protein of interest
- **Sum of standard deviations (SSD):**
  - Higher numbers are better; blacklist regions and ChIP enrichment are sensitive to this measure
- **Reads mapping at specific locations:**
  - Genomic features: look for enrichment in the context of what you know about your protein
  - RiBL: High percentage is > 10%. May want to filter these out before peak calling

Quality Checks: Peak Calling

ChIP QC Goals:

- **Ensure that you have enrichment and that your IP worked**
- Identify samples that consistently show measures below the acceptable thresholds for further troubleshooting
- Consider the protein of interest, the type of binding profile and the anticipated binding locations when evaluating quality metrics

Quality Checks: Handling Replicates



Quality Checks: Handling replicates

Evaluating the **concordance in peak calls across biological replicates**. The quality checks at this stage in the workflow include:

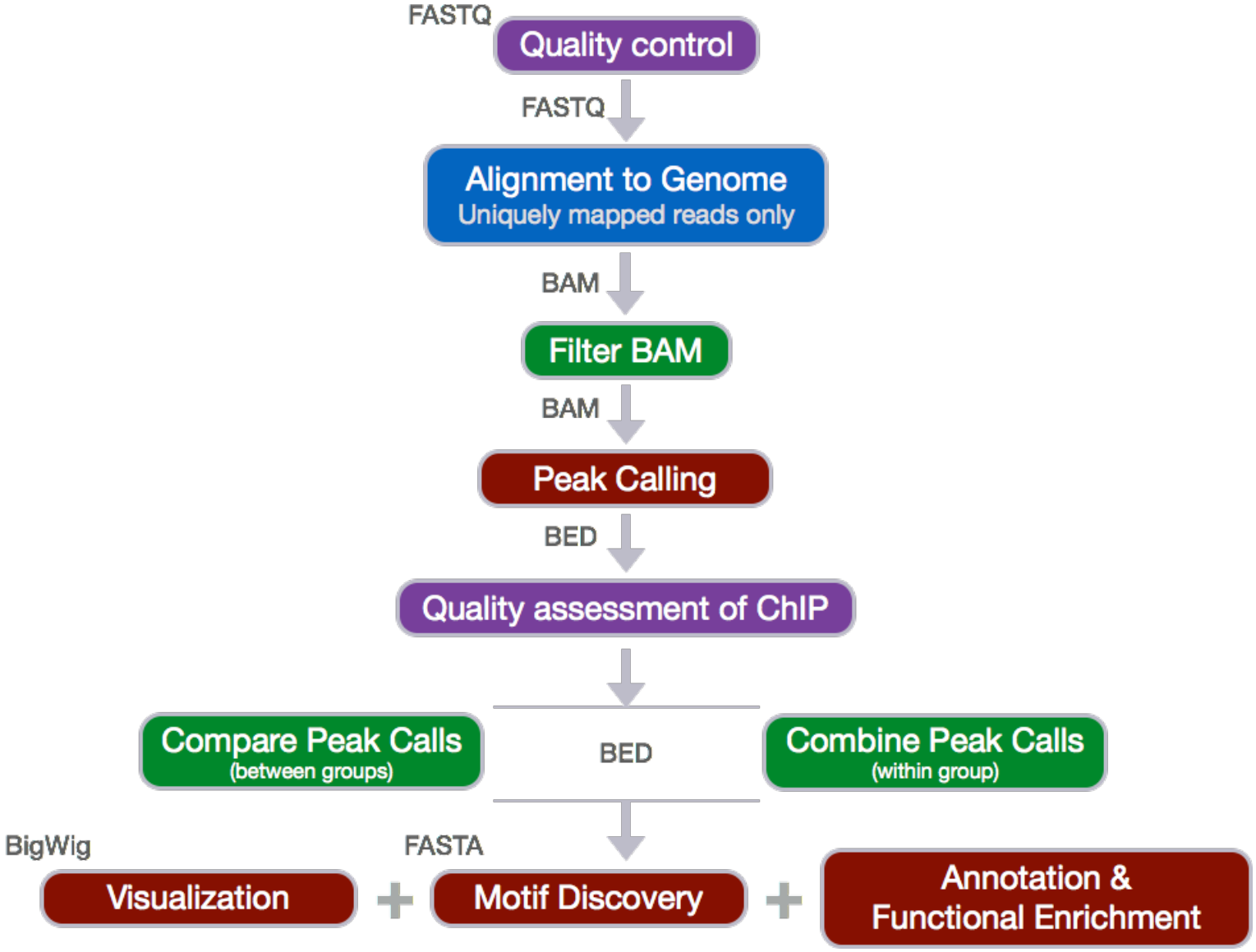
1. Using bedtools to perform a crude comparison of peak calls between replicates
2. Statistical evaluation of reproducibility between replicates using the IDR pipeline

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 to increase the search space and the total number (p-value instead of q-value
  - Focus on peaks that meet the IDR < 0.05 threshold for downstream analysis (visualization and functional analysis)
  - Pseudo-replicate analysis is useful for reporting in a publication

Quality Checks: Downstream analysis



Quality Checks: Annotation and functional analysis

Use nearest gene analysis to annotate peaks with gene annotations and use gene list as input to tools for functional analysis. The quality checks at this stage in the workflow include asking appropriate questions:

1. Do these genes collectively represent specific pathway(s)? Is there significant over-representation of certain biological processes?
2. Are these pathways relevant based on what I know about my protein of interest?

Quality Checks: Qualitative Assessment

Evaluating **enrichment patterns in specific regions of interest**.

The quality checks at this stage in the workflow include:

1. Using deepTools to generate signal profile plots and heatmaps
2. Looking at specific regions of interest and overlaying relevant public datasets (using IGV or a genome viewer)

Quality Checks: Qualitative Assessment

Qualitative Assessment Goals:

- **Use visual inspection to ensure that there is convincing read density in the regions of interest**
- Use read density to validate high confidence peaks that we have identified computationally
- Create publication quality tracks for specific genomic regions that are biologically meaningful (can focus on the strongest replicate to avoid redundancy)