FastQC - Comprehensive Reference Guide

Overview

FastQC is a quality control tool for high-throughput sequencing data. It provides a modular set of analyses to identify potential problems with raw sequence data before downstream analysis.

Applicable to: RNA-seq, DNA-seq (WGS/WES), ChIP-seq, ATAC-seq, BS-seq, and any Illumina/other NGS data

Website: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

What FastQC Analyzes

FastQC provides quality metrics across multiple modules:

1. **Basic Statistics** - File info, total sequences, read length
2. **Per-base sequence quality** - Phred quality scores at each position
3. **Per-tile sequence quality** - Quality variation across flowcell tiles
4. **Per-sequence quality scores** - Distribution of average read quality
5. **Per-base sequence content** - Nucleotide composition at each position
6. **Per-sequence GC content** - GC% distribution
7. **Per-base N content** - Proportion of uncalled bases
8. **Sequence length distribution** - Read length uniformity
9. **Sequence duplication levels** - Library complexity
10. **Overrepresented sequences** - Highly abundant sequences
11. **Adapter content** - Adapter contamination levels
12. **K-mer content** - Overrepresented k-mers (optional)

Understanding FastQC Status

Each module receives a status flag:

- **PASS (Green)**: Normal, meets expected criteria
- **WARN (Yellow)**: Slightly unusual, may be acceptable depending on experiment
- **FAIL (Red)**: Unusual, likely indicates a problem

  Critical: WARN/FAIL does not always mean bad data! Context matters based on:

  - Experiment type (RNA-seq vs WGS vs ChIP-seq)
  - Library prep method (random priming vs PCR-free)
  - Biological expectations (highly expressed genes, repetitive regions)

Module-by-Module Guide

1. Basic Statistics

What it shows:

- Filename
- File type (FASTQ format detected)
- Encoding (e.g., Sanger/Illumina 1.9)
- Total sequences
- Sequences flagged as poor quality
- Sequence length
- %GC

No pass/fail - informational only

What to check:

- Total sequences matches expected depth
- Encoding is correct (should be Sanger/Illumina 1.9 for modern data)
- GC% is in expected range for organism

---

2. Per-Base Sequence Quality

What it shows: Box-and-whisker plot of Phred quality scores at each base position

Phred Score Interpretation:

```
Q40: 99.99% accuracy (1 in 10,000 error)
Q30: 99.9% accuracy (1 in 1,000 error)
Q20: 99% accuracy (1 in 100 error)
Q10: 90% accuracy (1 in 10 error)
```

FastQC thresholds:

- **PASS**: Median quality 28 across all positions
- **WARN**: Median quality <28 for any position
- **FAIL**: Lower quartile <10 or median <20 for any position

Expected patterns:

Illumina sequencing:

- High quality (Q35-40) at read start
- Gradual decline toward 3' end (normal)
- Quality drop is expected and acceptable if median stays >28

What's concerning:

- Low quality (<Q20) in first 10-20 bases → Sequencing run issue
- Sudden quality drops (not gradual) → Specific cycle failure
- Very low quality throughout → Failed sequencing run

Action:

- Trim bases where median quality <20
- If quality poor throughout, consider re-sequencing

---

3. Per-Tile Sequence Quality

What it shows: Heatmap of quality deviation across flowcell tiles

Only available for: Illumina data with tile information in read headers

Interpretation:

- Blue = Better quality than average
- Red = Worse quality than average
- Uniform color = Good (even quality across flowcell)
- Patches of red = Specific tiles have problems

Causes of tile-specific issues:

- Bubbles in flowcell
- Dust particles
- Flowcell surface defects
- Imaging problems

Action:

- Minor variations normal (ignore)
- Large red patches → May need to filter reads from bad tiles
- Severe → Contact sequencing facility

---

4. Per-Sequence Quality Scores

What it shows: Distribution of average quality per read

Expected pattern:

- Single peak at high quality (Q35-40)
- Most reads should have average quality >28

Concerning patterns:

- Peak at low quality (<Q20) → Many poor reads
- Broad distribution → Mixed quality population
- Bimodal distribution → Two quality populations (investigate)

Action:

- If peak <Q20: Filter low-quality reads or re-sequence
- If broad distribution: Check if specific tile/lane issue

---

5. Per-Base Sequence Content

What it shows: %A, %T, %G, %C at each base position

Theory (random DNA):

- A%, T%, G%, C% should each be ~25%
- A% should equal T% (complementary)
- G% should equal C% (complementary)
- Should be stable across read length

Experiment-specific expectations:

Experiment Type	Expected Pattern	Why
RNA-seq (random hexamer)	Bias in first 10-15bp	Hexamer priming bias - NORMAL
RNA-seq (poly-A)	Some 3' A-bias	Poly-A tail fragments - NORMAL
WGS (random frag)	Flat across read	Truly random
WGS (PCR-free)	Flat across read	No bias
WGS (Nextera)	Bias in first 5-10bp	Transposase bias - NORMAL
ChIP-seq	May have bias	Protein binding preference - can be NORMAL
ATAC-seq	First 10bp bias	Tn5 transposase bias - NORMAL
Amplicon-seq	FAIL	PCR primers - EXPECTED
BS-seq	Different pattern	C→T conversion - EXPECTED

Action:

- RNA-seq: Ignore FAIL in first 10-15bp
- WGS: Should PASS; if not, investigate contamination
- Know your experiment: Some bias is expected!

---

6. Per-Sequence GC Content

What it shows: Distribution of GC% across all reads vs. theoretical normal

Species-specific GC content:

Organism	Expected GC%
Human	40-42%
Mouse	42%

Organism	Expected GC%
Drosophila	42%
C. elegans	35%
Zebrafish	36%
E. coli	50%
Yeast	38%
Arabidopsis	36%

Expected pattern:

- Single peak near species-specific GC%
- Approximately normal distribution

Concerning patterns:

Sharp peak(s) within main distribution:

- Adapter dimers
- Overrepresented sequences
- rRNA contamination (RNA-seq)

Broad or bimodal distribution:

- Contamination with different organism
- Mixed sample
- DNA and RNA mix

Shifted peak (far from expected):

- Wrong organism
- Heavy contamination
- Very biased library

Experiment-specific expectations:

Experiment Type	Expected
WGS	Single peak at genome GC%
WES	May be shifted (exons have different GC than genome average)
RNA-seq	May be shifted (transcriptome GC genome GC)
ChIP-seq	May be shifted if protein binds GC-rich/poor regions
ATAC-seq	Usually matches genome, open chromatin not GC-biased

Action:

- Know expected GC for your organism
- Sharp peaks → Check overrepresented sequences
- Broad distribution → Check for contamination

---

7. Per-Base N Content

What it shows: % of reads with uncalled base (N) at each position

Expected: <1% N content, ideally <0.1%

Interpretation:

- **PASS**: N% <5% at any position
- **FAIL**: N% 5% at any position

Causes of high N content:

- Sequencing chemistry issues
- Low signal intensity (poor cluster density)
- Flowcell problems
- Overclustering

Action:

- Low N% (<1%): Normal, acceptable
- Moderate N% (1-5%): May be OK depending on application
- High N% (>5%): Consider re-sequencing

---

8. Sequence Length Distribution

What it shows: Distribution of read lengths

Expected patterns:

Illumina (fixed length):

- All reads exactly same length
- Single peak (e.g., all 150bp)

After quality trimming:

- Range of lengths (trimmed to different extents)
- Should have minimum length threshold applied

PacBio/Nanopore (variable):

- Broad distribution
- Expected for long-read platforms

Interpretation:

- **PASS**: All reads same length (or expected distribution)
- **WARN**: Multiple peaks (check if expected)
- **FAIL**: Very different from expected

Concerning:

- Many very short reads (<20bp) → Adapter dimers or over-trimming

- Unexpected variation in fixed-length data

---

9. Sequence Duplication Levels

What it shows: Degree to which sequences appear multiple times

Theory: In random sequencing, duplicates should be rare

Reality varies by experiment:

Experiment Type	Expected Duplication	Interpretation
WGS	<20%	>50% = Low complexity or PCR bias
WES	20-40%	Higher than WGS due to targeted regions
RNA-seq	50-70%	High expression genes = many identical reads - NORMAL
ChIP-seq	30-50%	Enrichment creates duplicates - can be NORMAL
ATAC-seq	40-60%	Open regions enriched - can be NORMAL
Amplicon-seq	>90%	PCR products - EXPECTED
Small RNA-seq	>80%	Limited diversity - EXPECTED

FastQC limitation:

- FastQC samples only first 100,000 sequences
- May not represent whole library
- Underestimates duplication in very deep sequencing

When duplication is concerning:

WGS >50%:

- Low library complexity
- Too many PCR cycles
- Low input DNA amount
- Consider PCR-free libraries

RNA-seq >80%:

- Extremely low complexity
- Very low RNA input
- Check library prep quality

Action:

- Know what's normal for your experiment type
- WGS: Consider UMI-based deduplication
- RNA-seq: 50-70% is normal, don't worry!

---

10. Overrepresented Sequences

What it shows: Sequences appearing in >0.1% of total reads

FastQC attempts to identify source:

- Adapter sequences (from adapter database)
- rRNA (for RNA-seq)
- Other known contaminants

Common sources by experiment:

RNA-seq:

- rRNA (18S, 28S, 5S, 5.8S) → rRNA depletion failed
- Highly expressed genes (actin, GAPDH) → Expected!
- Mitochondrial RNA → Common, usually acceptable

WGS/WES:

- Adapters → Need trimming
- Repetitive elements → Can be normal
- Contamination → Investigate

ChIP-seq/ATAC-seq:

- Mitochondrial DNA → Common, may need filtering
- Repetitive regions → Can be expected

Action:

- Adapters: Trim before alignment
- rRNA >30%: Consider better rRNA depletion
- Highly expressed genes (RNA-seq): Normal
- Unknown sequences: BLAST to identify

---

11. Adapter Content

What it shows: % of reads containing adapter sequences at each position

Why adapters appear:

- Short insert size (read longer than fragment)
- Read-through into adapter
- Adapter dimers (very short/no insert)

Interpretation:

- **<5%**: Minimal contamination
- **5-10%**: Moderate, should trim
- **>10%**: Significant, definitely trim

Expected patterns:

Good library:

- Minimal adapter at read start
- May increase toward 3' end (normal)

Short inserts:

- Adapter appears mid-read and increases
- Common in small RNA-seq (expected)

Adapter dimers:

- High adapter content from read start
- Very short or no insert

Common adapters:

- Illumina Universal
- Illumina Small RNA 3'
- Nextera Transposase
- SOLiD Small RNA

Action:

- Always trim adapters before alignment
- If >10%, check library prep quality
- Adapter dimers → Improve size selection

---

12. K-mer Content

What it shows: Overrepresented k-mers (short sequences) across read positions

When enabled: Not run by default, use `--kmers` flag

Interpretation:

- Flags positionally-biased k-mers
- Can indicate adapter contamination
- Can indicate sequence-specific bias

Action:

- Usually redundant with other modules
- More sensitive to subtle biases
- Check if other modules also flag issues

Workflow-Specific Interpretation

RNA-seq Expectations

Commonly FAIL (but normal):

- Per-base sequence content (hexamer bias)
- Sequence duplication levels (50-70% normal)
- Overrepresented sequences (highly expressed genes)

Should PASS:

- Per-base sequence quality (after position 15)
- Per-sequence quality scores
- GC content (within ±10% of transcriptome GC)
- Adapter content <5% (after trimming)

Red flags:

- 30% rRNA (poor depletion/selection)
- High intergenic reads (indicates gDNA contamination - check with alignment QC)
- Very low complexity (>80% duplication)

WGS Expectations

Should PASS:

- Per-base sequence content (unless Nextera)
- Sequence duplication levels (<20%)
- Per-base sequence quality
- GC content (match genome)

May WARN/FAIL:

- Nextera libraries: First 5-10bp content bias (normal)

Red flags:

- 50% duplication (low complexity)
- GC content shifted or bimodal (contamination)
- High adapter content (poor size selection)

WES Expectations

Similar to WGS but:

- GC content may differ from genome (exons often GC-rich)

- Higher duplication OK (20-40%) due to targeted capture
- More position-specific bias from capture probes

Red flags:

- 60% duplication (over-amplified)
- Very uneven GC distribution (poor capture)

ChIP-seq / ATAC-seq Expectations

May FAIL (can be normal):

- Sequence duplication (30-60% due to enrichment)
- GC content (if protein binds GC-rich/poor regions)

Should PASS:

- Per-base sequence quality
- Adapter content <5%

Red flags:

- Very high mitochondrial content (>50% - poor enrichment)
- Very low complexity (poor enrichment specificity)

BS-seq (Bisulfite Sequencing) Expectations

Will FAIL (expected):

- Per-base sequence content (C→T conversion)
- GC content (loss of C)

Should PASS:

- Per-base quality
- Sequence length

Special considerations:

- Expect lower complexity (C→T reduces diversity)
- Adapter content especially important to trim

When to Run FastQC

During Workflow

```
Raw FASTQ → FastQC → Assess quality → Trimming (if needed) → FastQC → Alignment
```

Pre-alignment (always):

- Identify quality issues
- Determine if trimming needed
- Check for contamination

Post-trimming (if trimmed):

- Verify improvement
- Confirm adapters removed
- Ensure reads not over-trimmed

Post-alignment (optional):

- Usually not needed (alignment QC is different)
- Can check unmapped reads separately

Quality Thresholds Guide

Universal Standards

Metric	Excellent	Good	Acceptable	Poor	Action
Per-base quality	>35	>28	>20	<20	Trim if <20
Per-sequence quality	>35	>30	>25	<25	Filter or re-sequence
Adapter content	<1%	<5%	<10%	>10%	Always trim
N content	<0.1%	<1%	<5%	>5%	May need re-sequencing

Experiment-Specific Duplication Thresholds

Experiment	Excellent	Good	Acceptable	Concerning
WGS	<10%	<20%	<30%	>50%
WES	<20%	<30%	<40%	>60%
RNA-seq	<40%	40-60%	60-80%	>80%
ChIP-seq	<30%	30-50%	50-70%	>70%
Small RNA	N/A	N/A	>80%	<80% (too low!)

Common Issues and Solutions

Issue: Low Quality Scores (<Q20)

Causes:

- Overclustering on flowcell

- Insufficient sequencing reagents
- Flowcell age/damage
- Instrument calibration issues

Solutions:

- Trim low-quality bases
- If severe, re-sequence
- Contact sequencing facility

Issue: High Adapter Content

Causes:

- Short insert sizes (smaller than read length)
- Poor size selection during library prep
- Adapter dimers

Solutions:

- Trim adapters (use cutadapt, Trim Galore, etc.)
- Improve size selection in future preps
- Filter very short reads after trimming

Trimming tools:

- Cutadapt
- Trim Galore
- Trimmomatic
- fastp

Issue: High Duplication

Diagnosis: Check if appropriate for experiment type

If concerning:

For WGS:

- Use UMI-based deduplication (if UMIs present)
- Mark duplicates with Picard
- Consider PCR-free library prep next time

For RNA-seq:

- 50-70% is normal
- 80% check library prep quality
- Consider UMI-based quantification

Issue: Contamination (Unusual GC, Overrepresented Seqs)

Diagnosis:

1. Check overrepresented sequences
2. BLAST unknown sequences
3. Compare GC to expected

Common contaminants:

- PhiX (Illumina spike-in) ~45% GC
- E. coli ~50% GC
- Adapters
- Vectors (plasmids)

Solutions:

- Filter contaminating sequences
- Align to contaminant reference and discard
- Screen with FastQ Screen

Issue: Tile-Specific Quality Problems

Diagnosis: Check per-tile quality heatmap

Solutions:

- Minor: Ignore
- Severe: Filter reads from bad tiles using read headers
- Contact facility if widespread

Technical Details

How FastQC Works

Sampling strategy:

- Analyzes subsets of reads (not all)
- Defaults: First 100,000 sequences for duplication
- Can be changed with --limit_bases or --extract

Processing:

1. Detects FASTQ format and quality encoding
2. Calculates statistics per module
3. Applies thresholds for PASS/WARN/FAIL
4. Generates HTML report and ZIP data

Speed:

- ~1-2 minutes per FASTQ file
- Scales with file size
- Parallelizes multiple files with --threads

Output Files

HTML Report:

- Interactive visualization
- All modules with plots
- Summary table
- ~500KB-2MB size

ZIP File:

- Raw data in text format
- `fastqc_data.txt`: All statistics
- `summary.txt`: PASS/WARN/FAIL table
- `Images/`: PNG plots
- Used by MultiQC and other tools

Command Line Usage

```
# Basic usage
fastqc sample.fastq.gz

# Multiple files
fastqc sample1.fq.gz sample2.fq.gz

# Specify output directory
fastqc -o output_dir/ sample.fq.gz

# Use multiple threads
fastqc -t 4 sample1.fq.gz sample2.fq.gz

# Quiet mode (suppress progress)
fastqc -q sample.fq.gz

# Extract ZIP automatically
fastqc --extract sample.fq.gz

# Custom limits
fastqc --limits limits.txt sample.fq.gz
```

```
# Skip specific modules
fastqc --nogroup sample.fq.gz  # Disable read grouping
```

Integration with Other Tools

MultiQC

Aggregates multiple FastQC reports:

- Parses `*_fastqc.zip` files
- Creates comparative plots
- Generates single HTML report
- Useful for batch analysis

FastQ Screen

Complements FastQC for contamination screening:

- Aligns reads to multiple reference genomes
- Quantifies contamination sources
- More sensitive than GC content alone

Trim Galore

Wrapper around Cutadapt and FastQC:

- Trims adapters and low quality
- Automatically runs FastQC before/after
- Convenient for standard workflows

Troubleshooting

FastQC Won't Run

Error: "Couldn't find specified file"

- Check file path is correct
- Ensure file exists and readable
- Check permissions (chmod)

Error: "Unable to determine encoding"

- File may not be valid FASTQ
- Check format with: `head -n 20 file.fq`
- Ensure quality scores are present

Error: "Out of memory"

- Very large files (>50GB)
- Increase Java heap: `fastqc -Xmx4g file.fq`
- Or split file into chunks

HTML Report Won't Open

Causes:

- Browser compatibility (use Chrome/Firefox)
- JavaScript disabled
- File corrupted during transfer
- Network location issues

Solutions:

- Copy file locally
- Enable JavaScript
- Try different browser
- Re-run FastQC if corrupted

Results Look Wrong

All modules FAIL:

- Not a FASTQ file
- Wrong encoding detected
- Severely corrupted data

Check:

```
# Verify FASTQ format
head -n 20 file.fq.gz | gunzip

# Expected:
@sequence_id
ACTG...
+
IIII...
```

Unexpected patterns:

- Know your experiment type
- Some ”failures” are expected
- Context matters

Quality Decision Tree

```
Start
  ↓
Per-base quality >20 across read?
    NO → Trim low quality regions
    YES → Continue
        ↓
Adapter content >5%?
    YES → Trim adapters
    NO → Continue
        ↓
Duplication appropriate for experiment?
    NO → Investigate library prep
    YES → Continue
        ↓
GC content matches expectation?
    NO → Check for contamination
    YES → Continue
        ↓
Proceed to alignment
```

Best Practices

Always

Run FastQC on raw data before any processing
Check all modules, not just overall status
Know what's normal for your experiment type
Trim adapters if >5% content
Keep logs and reports for reproducibility

Never

Assume FAIL means unusable data
Ignore context of experiment type
Skip quality control to save time
Delete raw data before QC
Trust quality metrics from sequencing facility alone

References

- **FastQC Documentation**: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

- **FastQC Tutorial**: https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/
- **MultiQC**: https://multiqc.info/
- **Phred Quality Scores**: Ewing et al. Genome Research 1998

Document Version: 2.0
Last Updated: January 2026
Applicable to: All NGS platforms and experiment types