**Detailed Interpretation of FastQC Metrics and Thresholds**

**Overview of FastQC Quality Control Modules**

FastQC is a commonly used tool for quality control of high-throughput sequencing data. Each report includes multiple analysis modules that evaluate different aspects of sequence quality [25†L135-L143]. Each module focuses on a specific statistic or plot, providing a result that is flagged as *“Pass,”* *“Warn,”* or *“Fail”* based on predefined thresholds [25†L149-L157]. Note that these thresholds assume data derived from a high-quality, randomly fragmented whole-genome library. For specialized libraries (e.g., RNA-Seq, ChIP-Seq), FastQC’s thresholds may be too stringent [25†L152-L159]. Below is a detailed description of each metric in FastQC—including the computation methods, influencing factors, and biological significance—followed by an explanation of default threshold rationales and suggestions for adjustment. We also include case analyses for different types of sequencing data (RNA-seq, DNA-seq, ChIP-seq, etc.).

**FastQC Metrics in Detail**

**1. Basic Statistics**

**Calculation & Content**: The Basic Statistics module provides summary information about the input FASTQ file, such as filename, file type (nucleotide or color-space), quality score encoding, total number of reads, read length range (or single value if fixed), and overall GC content [13†L133-L141][13†L155-L163]. If using Illumina CASAVA formats, it also reports how many reads were flagged as failing quality filters [13†L149-L158].

**Influencing Factors**: This module merely summarizes the input file; it is not generally affected by the sequencing process itself, though platforms and data types can cause variation (e.g., Phred+33 vs. Phred+64, variable read length, different GC content across species).

**Biological Significance**: Basic Statistics provides background context for interpreting the other modules. For instance, total reads reflect sequencing throughput; read length and quality encoding must match expectations; %GC can be compared to known genome-wide GC content. Note that Basic Statistics itself sets no warning or error thresholds (it never flags Warn or Fail) [13†L159-L163]. It is informational only, yet pivotal for understanding other metrics.

**2. Per base sequence quality**

**Calculation**: This module shows the distribution of quality scores at each base position across all reads [4†L169-L177]. FastQC typically presents this as a boxplot: the median is drawn as a red line, the box spans the 25th–75th percentile (IQR), and whiskers represent the 10th and 90th percentiles [4†L177-L185]. A blue line indicates the mean. The y-axis is the Phred quality score (the higher, the more reliable the base call), often overlaid with colored background segments indicating high (green), moderate (orange), and low (red) quality [4†L181-L189]. The x-axis spans the read length from the first to the last base, though FastQC may batch positions near the end to simplify plots if the read is long [9†L411-L419]. The Phred scale is defined as Q=−10log⁡10(p)Q = -10 \log\_{10}(p)Q=−10log10​(p), where ppp is the error probability—e.g., Q20 is ~1% error, Q30 ~0.1% error [9†L427-L434].

**Influencing Factors**: In Illumina sequencing, base quality often drops as the read progresses because of issues like:

1. **Signal decay**: Fluorescent intensity diminishes over cycles, reducing quality in the 3’ end [15†L130-L139].
2. **Phasing**: Accumulated cycle-to-cycle errors degrade synchronization [15†L145-L153].  
   Overclustering can also cause overlapping signals and overall lower quality [15†L158-L166]. Commonly, Read 2 has slightly lower quality than Read 1 due to additional cycles and reagent wear [9†L421-L425]. Any abrupt quality drop or a pronounced dip at a specific base often suggests machine malfunction or sample issues.

**Biological Significance**: The per-base quality plot helps detect consistent low-quality regions in the read. A typical pattern is slightly lower quality in the first few cycles, high in the middle, and a mild drop at the end—this is normal. Trimming the low-quality 3’ end can improve downstream results [15†L181-L189]. However, if major quality dips occur early or fluctuate drastically, investigate the sequencing or library prep. High-quality reads ensure reliable downstream alignment and variant calling. If large segments of reads show <Q20, the error rate becomes problematic, and additional measures (e.g., trimming/filtering) are recommended.

**Default Thresholds**: FastQC combines criteria: if **any** base position’s lower quartile (Q1) is below Q10 (~10% error) it triggers a warning; below Q5 triggers failure. Likewise, if the median is below Q25 it warns, below Q20 fails [21†L447-L454][23†L1-L4]. That is, if any position drifts into an unacceptably low range (≥25% of reads with Q<10 or median < Q20), it flags the module. These values align with the commonly accepted Q20 and Q30 cutoffs [9†L427-L434].

**Example**: In high-quality WGS data, most positions remain within high (green) territory except for a slight dip at the end; FastQC typically passes [15†L181-L189]. If instrument failure caused bad cycles, the quality plot would drop to the red zone and yield a Fail. This might require re-sequencing or discarding affected reads. In RNA-Seq, the second read’s tail often shows more pronounced declines, but as long as the median remains above Q20, it’s generally considered acceptable [9†L421-L425].

**3. Per tile sequence quality**

**Calculation**: This module inspects Illumina “tiles” on the flow cell. Each read ID contains flowcell coordinates (tile number, etc.). FastQC groups reads by tile for each cycle (base position) and computes average quality per tile, comparing it to the overall average [32†L7-L15]. A heatmap displays differences: blue indicates average or above, red below. Ideally, all tiles are near the same quality (mostly “blue”). Red “hotspots” highlight tiles with subpar quality [32†L13-L18].

**Influencing Factors**: Variation among tiles typically reflects local problems on the flow cell—e.g., bubbles, smudges, or debris. Overclustering can also exacerbate tile-to-tile differences [32†L37-L44]. A small red patch may be ignorable if it’s isolated or brief. But persistent, large areas of low quality across many cycles strongly suggest hardware or reagent failure [32†L39-L44].

**Biological Significance**: This primarily monitors sequencing hardware performance rather than any biological feature. It indicates potential localized issues. One could remove reads from problematic tiles, but it may only affect a small fraction, so many labs simply note the presence of tile problems and proceed. For the sequencing facility, recurrent tile issues could prompt instrument servicing [32†L37-L44].

**Default Thresholds**: FastQC flags if any tile’s mean Phred quality is notably lower than the global average (older doc mentions a 2–5 point difference for warning, >5–10 for failure) [32†L27-L35][20†L479-L487]. The exact default in FastQC 0.11.9 is warn=5 and error=10 difference [20†L479-L487]. If you see a Warn/Fail in this module, it means a tile or multiple tiles consistently underperform. Minor, sporadic anomalies can be tolerated; broad sustained issues should be investigated.

**Example**:

* **High-quality runs**: typically pass with an all-blue heatmap.
* **Local issues**: you might see a row of red tiles indicating partial obstructions on that section of the flow cell [32†L21-L29].
* **Single-tile glitch**: a small patch of red in a few cycles might be negligible.

**4. Per sequence quality scores**

**Calculation**: This plots the distribution of **mean** quality scores per read [9†L427-L434]. For each read, FastQC computes the average Phred quality across all bases, then displays a histogram or density curve of these averages. The x-axis is the average quality score, and the y-axis is the number (or frequency) of reads at that score. Typically, one expects most reads to have a high average, forming a peak on the right side. Low-quality runs or partial failures can create a secondary peak at lower values.

**Influencing Factors**: The average quality for each read depends on read length and the quality distribution across positions. If most bases remain high, the average is also high (like Q30). If there is significant tail drop or widespread low-quality cycles, more reads will shift toward lower means. For longer reads, the average tends to be a bit lower than for short reads, since more cycles can degrade at the end [9†L427-L434]. Mixed read lengths can broaden the distribution.

**Biological Significance**: This metric quickly reveals if a large fraction of entire reads are of poor quality. If there's a big secondary peak at Q20 or below, it indicates many reads have overall low quality. That might require more aggressive filtering. Ideally, the main peak is near Q30 or higher with no major sub-peaks [9†L427-L434].

**Default Thresholds**: FastQC flags based on where the **most frequently observed average** (the main peak) lies [9†L427-L434]. If it’s below Q27, that’s a warning; below Q20 is a fail [9†L427-L434]. Q27 ~0.2% error, Q20 ~1%. If the entire library’s average read quality is only around Q25, it issues a Warn; if it’s around Q20 or lower, it Fails. A bimodal distribution can also trigger warnings if the second peak has significant reads below Q20.

**Example**:

* **WGS** with 150 bp reads might have a main peak around Q30.
* **Poor runs** (degraded RNA or failing cycles) might produce a second peak <Q20, leading to a Fail.
* **RNA-Seq** can sometimes show slightly lower average quality for read 2; as long as the main peak remains >Q27, it passes.

**5. Per base sequence content**

**Calculation**: This module shows the proportion of A, T, G, C at each base position across all reads [9†L433-L440]. It draws four lines representing the percentage of reads that have A, C, G, or T at each position. Under a random fragmentation assumption, these lines should be roughly parallel and close to 25% each [9†L433-L440]. FastQC uses how far they deviate from uniformity to assess bias.

**Influencing Factors**:

* **Random primer bias**: In RNA-Seq libraries using random hexamers, the first ~10–12 bases often show uneven base composition [27†L43-L49]. This is widely observed and considered a known artifact rather than a severe error [27†L38-L46].
* **Adapter read-through**: If reads extend into adapter sequence, the tail region might show an overrepresentation of G/C if the adapter is GC-rich.
* **Enzyme/fragmentation bias**: e.g., Tn5-based Nextera libraries can have a distinct bias at read starts [30†L167-L170].
* **Non-random libraries**: PCR amplicons or targeted sequences might share identical starts, inflating certain base frequencies. For small RNA (e.g., miRNA-Seq), the first base is often biologically biased (e.g., U in many miRNAs).

**Biological Significance**: In a truly random library, each position should reflect the overall base composition with minimal skew [9†L433-L440]. Deviation can indicate library prep bias or contamination. A slight or early bias may be known and harmless (e.g., random hexamer-based RNA-Seq). However, if the entire read length shows large compositional skew, it suggests possible contamination or non-random sampling, which can lead to coverage bias in downstream analyses. Recognizing whether bias is expected or not is crucial.

**Default Thresholds**: FastQC checks if **any** position’s difference between A and T or G and C exceeds 10% (Warn) or 20% (Fail) [9†L433-L440]. For instance, if at base 5, A=35% and T=15%, that is a 20% difference, likely failing. For typical random libraries, this rarely happens unless there’s a real problem or special library type. Note that RNA-Seq often fails because of the well-known start bias, but that is often benign [25†L152-L159][27†L36-L44].

**Example**:

* **RNA-Seq**: Typically fails this module but is normal. The first 10–12 bases are skewed (e.g., A/T > C/G). Don’t be alarmed if it’s only at the start [27†L43-L49].
* **WGS**: If you see a fail, that suggests contamination or leftover adapters. A random genome library should be uniform.
* **miRNA**: The first base often strongly biased to U, so the module fails. That reflects true biology, not an error.

**6. Per sequence GC content**

**Calculation**: This calculates the GC percentage for each read (the fraction of G and C bases) and plots a histogram [9†L441-L448]. FastQC also generates a theoretical distribution (assuming random fragmentation with the overall average GC) and compares the observed distribution to see if it matches a roughly normal curve. Large deviations from the expected curve indicate anomalies [9†L441-L448].

**Influencing Factors**:

* **Sample contamination**: If multiple genomes with different GC contents are mixed, a **bimodal** or wide distribution can occur (e.g., human at ~40% plus bacterial at ~60%).
* **Library selection bias**: Certain methods (PCR or capture-based) might preferentially amplify mid-GC fragments, leading to a narrower or shifted GC distribution.
* **Extremely high or low GC** fragments might be underrepresented due to amplification difficulties.
* **Organism-specific**: Some eukaryotic genomes have complex GC domains, but typically an overall unimodal distribution is expected.

**Biological Significance**: GC content is a fundamental property. For a single organism, the read-level GC distribution should center near the genome’s overall GC [9†L441-L448]. Deviation implies contamination or strong library bias. Extreme GC bias can affect alignment and coverage, thus impacting variant detection or expression estimates. If the distribution is heavily skewed, investigate sample purity or library prep steps.

**Default Thresholds**: FastQC sums the difference between the observed and theoretical distribution. If more than 15% of reads deviate from the expected curve, it warns; over 30% fails [21†L463-L471]. A 15% deviation often signals moderate bias, while 30% suggests a severe issue. Nonetheless, certain real genomic features can cause mild deviations. The shape of the histogram matters—a second peak typically indicates contamination.

**Example**:

* **WGS**: A single peak around the known genome GC. If a second peak emerges, suspect contamination.
* **RNA-Seq**: Usually near the known transcriptome GC average, but sometimes slightly broadened or skewed due to highly expressed transcripts. Minor warning is often acceptable.
* **ChIP-Seq**: Possibly a small shift if targeted regions are GC- or AT-rich, but rarely a big fail unless the target regions are extremely biased.

**7. Per base N content**

**Calculation**: This shows the proportion of undetermined bases (“N”) at each read position [10†L449-L457]. For each position, FastQC calculates how many reads contain “N” at that position and plots it as a function of the base index.

**Influencing Factors**:

* **Instrument errors**: A failing cycle or insufficient signal can produce “N” calls at that specific base across many reads.
* **Overclustering**: Overlapping clusters reduce basecalling accuracy, increasing N calls.
* **Library or adapter issues**: Typically, real biological sequences do not contain “N.” If present, it’s due to sequencing uncertainty or residual ambiguous barcodes/adapters.

**Biological Significance**: High N content reduces usable information since N doesn’t match any reference base, hindering alignment or assembly. Normally, N% should be ~0 across positions. A slight increase (e.g., <1%) at the tail is manageable; a large spike indicates a major problem at a particular cycle.

**Default Thresholds**: If **any** position has >5% N, FastQC warns; >20% fails [10†L449-L457]. Even 5% is quite high for typical Illumina data. This suggests a serious glitch in that cycle or region. Usually, well-run data show a flat line near 0%.

**Example**:

* **Most Illumina**: 0% N or near-zero, passes.
* **Fail**: If an entire cycle was mis-called (e.g., imaging failure), you could see 100% N at that position, producing a massive spike. The run might require re-sequencing or dropping that cycle.

**8. Sequence Length Distribution**

**Calculation**: This module plots a histogram of read lengths [10†L455-L463]. On platforms with fixed read lengths (Illumina standard), one expects a single peak. For variable-length data (Nanopore, PacBio), a broader distribution is typical. If some reads were trimmed, the histogram may show a range rather than a single length.

**Influencing Factors**:

* **Sequencing design**: e.g., 2×150 bp Illumina run typically yields a uniform 150 nt length.
* **Quality trimming**: Tools like Trimmomatic remove low-quality or adapter-contaminated segments, producing variable read lengths.
* **Long-read platforms**: naturally generate wide length distributions.
* **Unexpected outcomes**: If you see multiple discrete peaks, it might indicate library contamination or a mix of different read sets.

**Biological Significance**: For Illumina-based workflows, consistent length is standard. Variation might signal partial read failure or pre-processing. If the entire library is heavily trimmed, many short reads can complicate alignment. For targeted amplicon sequencing (fixed amplicon size), a wide distribution suggests off-target products or library issues.

**Default Thresholds**: FastQC logic is straightforward: if there is more than one read length, it triggers a Warn; if any read has length 0, it fails [10†L455-L463]. Modern workflows commonly produce multi-length data (after adapter/quality trimming), so a Warn is typical and not necessarily bad. The Fail condition for zero-length is extremely rare unless there’s a file corruption.

**Example**:

* **Raw Illumina**: Usually uniform read length, passes.
* **Trimmed**: A range of lengths leads to a Warn.
* **Mixed data sets**: Combining 75 bp and 150 bp reads will produce two peaks.
* **Zero-length**: Possibly a faulty FASTQ conversion.

**9. Sequence Duplication Levels**

**Calculation**: This measures how often identical read sequences appear [10†L461-L469]. FastQC groups reads by sequence (ignoring quality scores) and counts how many times each unique sequence occurs (e.g., a sequence that appears 5 times is in the “5-duplication group”). The report plots duplication levels on the x-axis (occurrence count) versus the proportion of reads that fall into each duplication bin. FastQC also reports the percentage of reads remaining after deduplication (retaining only one copy of each unique sequence) [29†L69-L77].

**Influencing Factors**:

* **PCR duplicates**: Library prep can over-amplify certain fragments [28†L39-L43].
* **Over-sequencing**: Very high coverage of a small genome leads to many reads that are truly identical (biological duplication), e.g., a phiX library measured with 10,000× coverage inevitably sees repeated fragments [29†L79-L87].
* **Coverage unevenness**: In RNA-Seq, highly expressed transcripts yield many identical reads. Similarly, ChIP-Seq peaks can produce repeated reads from the same region.
* **Random coincidence**: With large data sets, the chance of two different fragments being identical is low for >100 bp reads, so mostly duplicates reflect real or technical replication.

**Biological Significance**: Duplicate reads do not provide additional coverage information; they can inflate coverage metrics or expression levels if not accounted for [28†L39-L43]. In whole-genome sequencing, high duplication rates often indicate poor library complexity or over-amplification. However, in RNA-Seq or ChIP-Seq, many duplicates arise from true biological signals (e.g., highly expressed genes or strong binding sites). Thus, you typically remove or mark PCR duplicates but handle biologically duplicated fragments differently [28†L39-L43]. This metric highlights library complexity issues: if post-dedup retention is only 50%, half of the reads are duplicates—potentially a big waste of sequencing capacity [10†L463-L468].

**Default Thresholds**: FastQC classifies duplication levels based on the percentage of unique reads. If >20% are duplicates (i.e., <80% unique), it warns; if >50% duplicates (<50% unique), it fails [10†L463-L469]. This is aimed at WGS logic. For transcriptome or amplicon data, duplication can be inherently high; thus, Fail or Warn might be expected [33†L114-L118]. FastQC cannot distinguish between PCR duplicates and real biological duplication.

**Example**:

* **WGS**: Typically low duplicates; often >80% unique (Pass). If <50% unique, suspect over-amplification or extremely deep coverage.
* **RNA-Seq**: Often warns/fails because a few transcripts dominate, leading to many identical reads. This may not reflect actual library prep issues but rather real expression differences [30†L153-L160].
* **ChIP-Seq**: 50–60% duplication is common due to peak enrichment [33†L114-L118]. A Fail can be normal. Extremely high duplication (~80%+) might indicate insufficient starting material or over-amplification.
* **Amplicon**: e.g., 16S—practically all reads may be identical or from a few variants. Expect a massive duplication rate, not necessarily a problem for targeted studies.

**10. Overrepresented sequences**

**Calculation**: FastQC looks for any **individual read sequence** that appears at an unusually high frequency [10†L470-L477]. It tallies all reads and calculates the proportion for each unique sequence, listing those exceeding a threshold. The report identifies the sequence, its frequency, and attempts to match it to known adapters or contaminants (e.g., matches Illumina adapter or rRNA). Ideally, no single sequence should dominate a random library [10†L470-L477].

**Influencing Factors**:

* **Adapter or adapter dimer**: Short inserts might result in quickly reading into adapter, producing identical or nearly identical reads [10†L479-L485].
* **Primer contamination**: If library prep primers remain, they might show up repeatedly.
* **Highly abundant natural sequences**:
  + **rRNA** in RNA-Seq if depletion was incomplete, producing high frequencies of certain rRNA fragments.
  + **Viral sequences** in infection studies with high viral load.
  + **Low-complexity** motifs like poly-A tails.
* **Other contaminants**: e.g., vector or plasmid sequences if introduced accidentally.

**Biological Significance**: Overrepresented sequences often indicate reduced library diversity or contamination. In a truly random library, no single read should exceed 0.1% of total [10†L470-L477]. If the cause is real biology (like extremely abundant miRNA or rRNA), it may be expected; if it’s technical (adapters, primer dimers), it must be removed before mapping.

**Default Thresholds**: Any sequence above 0.1% is listed. If any exceed 0.1%, you get a Warn; if >1%, you get a Fail [10†L473-L477]. 0.1% means that sequence appears 1000 times in 1 million reads, which is already quite high for a single read. 1% is more extreme [10†L473-L477]. This criterion suits random libraries. For amplicon data, single sequence dominance can be expected, so interpret carefully.

**Example**:

* **Adapter contamination**: The most frequent cause. If the adapter appears at >1% it triggers Fail. You should trim adapters [10†L479-L485].
* **rRNA**: Overrepresented rRNA fragments can appear in transcriptome libraries if ribosomal depletion was incomplete, leading to a Warn.
* **High-abundance transcript**: e.g., a single miRNA might account for 2% of reads. FastQC fails, but that might be biologically correct.
* **Severe contamination**: If a single adapter-dimer sequence is 50% of reads, the library is almost entirely adapter contamination—serious QC failure.

**11. Adapter Content**

**Calculation**: This module specifically checks for the presence of adapter sequences [10†L479-L485]. FastQC has an internal database of common adapters (Illumina TruSeq, etc.). It searches for k-mers matching these adapters at each read position and plots a cumulative percentage of reads containing adapter at or after that position [10†L479-L485]. Typically, the curve rises toward the read end if many fragments are shorter than the read length, resulting in reading into the adapter.

**Influencing Factors**:

* **Insert length < read length**: If the average fragment is 100 bp but the read is 150 bp, the last part of the read will often be adapter.
* **Adapter dimer**: If dimers (adapter–adapter with no insert) are present, you get reads that are basically all adapter.
* **Index misassignment**: Not as commonly reflected in this plot, though occasionally short barcodes appear.
* **Similar sequences**: Rarely, the sample itself might contain sequences resembling the adapter.

**Biological Significance**: Adapters are artificial sequences that do not carry biological information and can severely interfere with alignment and assembly. The module helps determine whether you need to trim. Ideally, the curve remains near zero. If it climbs significantly, it indicates many reads contain adapter segments [10†L479-L485].

**Default Thresholds**: If at any position in the read >5% of reads contain adapter, it warns; >10% fails [10†L479-L485]. Even 5% contamination can be problematic for alignment. Typically, if you see an orange or red flag here, adapter trimming is recommended.

**Example**:

* **Standard Illumina**: With a 2×150 bp run but 200 bp inserts, read 2 often extends into adapter for short inserts. The adapter content curve might rise significantly near the end, leading to a Fail if >10%. Trimming is essential.
* **Short fragments**: e.g., ChIP-Seq with <100 bp fragments read by 150 bp length often shows adapter content rising early. Must trim.
* **No adapter**: Amplicon libraries where the product length is longer than the read. The curve is zero, so it passes.
* **Severe contamination**: If many reads are just adapter dimers, the curve is near 100% from the beginning.

**FastQC Default Thresholds: Rationale & Impact on QC**

FastQC employs a “red-yellow-green” flag system to mark whether each module meets expected quality standards. These defaults (Pass/Warn/Fail) stem from the developers’ experience with typical high-quality random libraries [25†L152-L159]. Essentially, they rely on widely accepted error-rate cutoffs and normal distribution assumptions (e.g., Q20 as a baseline, random GC distribution, minimal overrepresented sequences) [25†L152-L159].

**Key threshold rationales**:

* **Phred-based modules** (Per base quality, Per sequence quality): Q20 (~1% error) is considered the minimal acceptable; Q25–30 is comfortable. FastQC sets Fail if median < Q20 or if the major peak < Q20, etc. [21†L447-L454].
* **Base content**: A difference >10% is quite large for a random library, >20% is extreme.
* **GC distribution**: A cumulative deviation of 15% (Warn) or 30% (Fail) indicates substantial departure from expected single-peak normal [21†L463-L470].
* **Sequence duplication**: <80% unique triggers a Warn, <50% is a Fail, presuming it’s random WGS [10†L463-L469].
* **Overrepresented sequences**: >0.1% is flagged, >1% is a Fail [21†L437-L445].
* **Adapter**: >5% or 10% contamination triggers Warn or Fail, respectively [21†L499-L507].
* **Per tile**: Typically ~5 difference in mean quality from global average is Warn, 10 difference is Fail [32†L27-L35].

These cutoffs are not final verdicts but alerts. FastQC’s documentation emphasizes that any Warn/Fail simply means “this merits attention,” not necessarily that the data are unusable [25†L155-L162]. Users must examine whether the flagged condition is truly problematic or expected for the library type.

**Adjusting thresholds**:

* **Based on library type**: For RNA-Seq, ignoring sequence content fails is common; for small RNA, ignoring duplication fails is typical [27†L36-L44][30†L153-L160].
* **Customizing FastQC**: Advanced users can modify the limits.txt config file to relax or tighten thresholds [20†L441-L459]. Alternatively, one can keep default flags but interpret them in context.
* **Other tools**: Tools like fastp or MultiQC might provide more refined QC analyses [30†L147-L155].
* **Explain flags in reporting**: Many labs provide commentary when delivering QC results, e.g., “The Per base sequence content is flagged due to random hexamer bias, which is normal for RNA-Seq.”

In summary, these thresholds are grounded in empirical norms for typical Illumina data. In practice, interpret them with knowledge of the experiment. If a module fails for reasons consistent with known library characteristics (e.g., small RNA or amplicon data), it may not actually be a problem [23†L1-L4].

**Case Studies for Different Sequencing Data Types**

**RNA-Seq**

**Features**: Uses random hexamer priming, has a wide dynamic range of transcript abundances. Common FastQC flags:

* **Per base sequence content**: Almost always fails due to ~10–12 bases of biased composition from random hexamer usage [27†L36-L44]. This is well-documented and usually harmless [27†L59-L67].
* **Sequence duplication**: Often warns/fails. Highly expressed genes produce many identical reads. This is mostly biological duplication, not PCR artifact [30†L153-L160]. Usually, do not remove duplicates at the FASTQ stage; handle them after alignment if needed.
* **Overrepresented sequences**: If incomplete rRNA depletion, rRNA fragments may dominate. Or there might be some adapter contamination.
* **Adapter content**: Typically minimal unless the inserts are very short (e.g., degraded RNA).

**Example**: Human mRNA-Seq (2×150 bp). FastQC results:

* Fails Per base sequence content for the first ~10 bp.
* Fails duplication (only 50% unique) because of a few highly expressed genes.
* Overrepresented sequences: 0.5% adapter, some 18S rRNA fragments.  
  **Response**: Trim away adapter, accept the initial base-composition bias, and do not forcibly remove duplicates. After trimming, everything else might pass. The data remain viable for differential expression analysis.

**DNA WGS (Whole-Genome Sequencing)**

**Features**: A random genomic DNA library is what FastQC thresholds were mainly designed for [25†L152-L159].

* **Per base sequence quality**: Typically high across the read, passing easily. Any abrupt dip can indicate major issues.
* **Per base sequence content**: Should be flat. A fail implies contamination or leftover adapter.
* **GC content**: Usually a single peak near the known genome GC. Double peaks strongly suggest a mixture.
* **Duplication**: Usually low if library complexity is good. >50% duplication would be a red flag unless extremely deep coverage or a small genome.
* **Overrepresented sequences/Adapter**: WGS libraries should have none or minimal. If flagged, suspect leftover adapters or a known spike-in.

**Example**: Bacterial WGS at 100× coverage.

* High quality (Per base quality passes).
* GC distribution has a second shoulder ~10% of reads with ~30–35% GC, leading to a warn. BLAST finds those reads are host contamination.
* Duplication fails (only 20% unique). That’s because 100× is high coverage for a 3 Mb genome, so repeated coverage is expected.

**Action**: Filter out contamination, keep the rest. Recognize that duplication fail is due to over-sequencing. Everything else is normal.

**ChIP-Seq**

**Features**: Targets specific binding sites, so coverage is not uniform. Key considerations:

* **Sequence duplication**: Commonly high because of peak enrichment [33†L114-L118]. A fail might still be normal if those duplicates represent real binding hotspots.
* **Per base sequence content**: Usually random enough if the antibody doesn’t have strong sequence preference; rarely triggers fails.
* **GC content**: Slight shifts are possible if the binding sites are GC- or AT-rich, but usually not dramatic.
* **Overrepresented sequences**: If extremely narrow peaks, you might get repeated identical reads. Usually, partial variation prevents one exact read from dominating though.

**Example**: H3K4me3 ChIP-Seq with moderate library size:

* Duplication ~50% (Warn).
* GC distribution close to the genome average, passes.
* Overrepresented sequence includes a small fraction of Illumina adapter.

**Conclusion**: The duplication warning is expected because ChIP peaks cause local coverage inflation. Just remove adapter contaminants if needed.

**Other Data Types (Brief Notes)**

* **Small RNA-Seq**: Very short inserts (20–30 nt). Per base composition fails due to strong base bias (e.g., many miRNAs start with U). Duplication is extremely high if a few miRNAs dominate. Overrepresented sequences might list top abundant miRNAs. All are normal for small RNA.
* **Exome/Targeted capture**: Because only exons are captured, duplication may be somewhat high, GC distribution might shift a bit if exons differ in GC from the genome.
* **Amplicon**: If the entire library is one or few amplicons, you effectively see the same read many times—overrepresented sequences and duplication will almost certainly fail. This is expected.

**Supporting Literature and Official Documentation**

FastQC metrics and thresholds are grounded in extensive empirical evidence and widely referenced best practices:

* **FastQC official manual**: Babraham Bioinformatics documentation explains each module in detail [4†L177-L185][32†L13-L18]. Configuration files (limits.txt) specify default thresholds [21†L447-L454][21†L471-L478].
* **Scientific papers**: While FastQC itself does not have a dedicated publication, many studies reference or incorporate its metrics. For instance, Bullard et al. (2010) discuss random hexamer bias in RNA-Seq [27†L43-L49]. Others mention Q20/Q30 standards [9†L427-L434]. Various tutorials (Haynes & Kelly, 2017) clarify typical results and interpretation [25†L155-L162].
* **Community experience**: Forums like SeqAnswers and BioStars host extensive discussions on duplication in RNA-Seq [30†L153-L161], random primer base biases [27†L38-L46], and so on.
* **Tool comparisons**: Publications on new QC tools (e.g., fastp) often contrast their features with FastQC [30†L147-L155], highlighting limitations in duplication estimation for single-end data, reaffirming the need for contextual interpretation of FastQC results.

Collectively, these sources validate the principles behind each FastQC metric—e.g., the ubiquitous use of Phred scores [9†L427-L434], detection of GC anomalies as contamination evidence [9†L441-L448], or how random hexamer priming can skew base composition for RNA-Seq [27†L36-L44][27†L59-L67]. The official documentation and user forums stress that a Warn/Fail is not a rigid condemnation but a prompt to investigate anomalies [25†L155-L162]. Adjusting thresholds or ignoring specific fails for certain library types is standard practice.

**References** (as retained from the original text structure):

1. Andrews, S. *FastQC: A Quality Control tool for High Throughput Sequence Data*. Babraham Bioinformatics (2010).
2. Positional sequence bias in random primed libraries – *QCFail Sequencing* (2013).
3. FastQC Manual, Babraham Bioinformatics (v0.11.9).
4. Babraham Bioinformatics – FastQC Help Documentation.
5. *FastQC default limits configuration* (GitHub repository).
6. Haynes, B. & Kelly, J.R. *Assessing FASTQC results*. Harvard FAS Informatics (2017).
7. Chen, S. *et al.* (2018). **fastp**: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*, **34**(17): i884–i890. (Discusses improvements over FastQC).
8. Duplication in sequencing libraries – DNA Tech Core FAQ (UCDavis).
9. Bullard, J.H. *et al.* (2010). Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC Bioinformatics*, **11**:94. (Notes on random hexamer bias).
10. Encode Consortium (2012). *ENCODE quality metrics guidelines* – Discusses sequence quality, duplicates, and complexity in NGS data. (General best practices aligning with FastQC metrics).

**Conclusion**:  
FastQC provides a suite of metrics to evaluate various facets of next-generation sequencing data quality. Each module’s thresholds are rooted in established Phred error rate standards and typical random library properties. In real applications, you must interpret flags in light of the specific library type and experimental context. Some fails (e.g., early base content bias in RNA-Seq) are expected artifacts and do not invalidate the data. Others (e.g., heavy adapter contamination, suspicious GC distributions, large N spikes) require further investigation or corrective actions (e.g., trimming, contamination removal). By combining FastQC results with knowledge of the library construction and sample biology, researchers can effectively manage sequencing QC and ensure reliable downstream analyses.