**FastQC各項指標解讀與閾值設定**

**FastQC質控模組概述**

FastQC是一款常用的高通量測序數據質控工具，每份報告包含多個分析模組（analysis modules），用於評估序列資料的不同品質面向​[rtsf.natsci.msu.edu](https://rtsf.natsci.msu.edu/genomics/technical-documents/fastqc-tutorial-and-faq.aspx#:~:text=FastQC%2C%20written%20by%20Simon%20Andrews,to%20the%20information%20provided%20here)。每個模組針對特定統計或圖譜提供結果，並根據預設閾值標記\*「通過」（Pass）*、*「警告」（Warn）*或*「失敗」（Fail）\*​ [rtsf.natsci.msu.edu](https://rtsf.natsci.msu.edu/genomics/technical-documents/fastqc-tutorial-and-faq.aspx#:~:text=section%20for%20each%20FastQC%20module,that%20has%20a%20%E2%80%9CWarn%E2%80%9D%20or)。需要注意這些閾值假定資料來自高品質的全基因組隨機片段文庫，因此對於RNA-Seq、ChIP-Seq等特殊文庫可能過於嚴格​[rtsf.natsci.msu.edu](https://rtsf.natsci.msu.edu/genomics/technical-documents/fastqc-tutorial-and-faq.aspx#:~:text=sequence%20data,and%20consider%20what%20that%20results)。以下我們詳細介紹FastQC報告中的各項指標，包括其計算方法、影響因素、生物學意義，以及FastQC預設閾值的原理與調整建議。同時，我們將結合不同類型生物數據（如RNA-seq、DNA-seq、ChIP-seq等）的質控案例進行說明。

**各項FastQC指標詳解**

**1. Basic Statistics（基本統計）**

**計算方法與內容**：Basic Statistics模組提供對序列檔的基本資訊彙總，包括：檔名、檔案類型（是否為核苷酸序列或顏色空間數據）、品質值的編碼格式、讀取序列總數、序列長度範圍（若固定長度則只報告單一值），以及總體GC含量等​[mugenomicscore.missouri.edu](https://mugenomicscore.missouri.edu/PDF/FastQC_Manual.pdf#:~:text=analysed,There%20are); [mugenomicscore.missouri.edu](https://mugenomicscore.missouri.edu/PDF/FastQC_Manual.pdf#:~:text=analysis,Statistics%20never%20raises%20an%20error)。如果使用Illumina CASAVA模式，該模組也會列出被標記為低質量而過濾掉的讀數數量​[mugenomicscore.missouri.edu](https://mugenomicscore.missouri.edu/PDF/FastQC_Manual.pdf#:~:text=this%20for%20now,only%20one%20value%20is%20reported)。

**影響因素**：此模組只是輸入檔案的描述性統計，一般不受測序流程影響，但會因測序平台和資料類型而異。例如，不同平台可能使用不同品質值編碼（如Phred+33或Phred+64），read長度視實驗設計而定，%GC則與物種基因組或目標區域特性相關。

**生物學意義**：基本統計為質控提供背景資訊。讀數總數反映測序產出量，序列長度和品質編碼需確認符合預期，以便後續分析。%GC可與物種基因組平均GC含量比較，以初步檢查樣本組成是否合理。**注意**：Basic Statistics模組本身不設定任何警告或錯誤閾值（永遠不會標記Warn或Fail）​[mugenomicscore.missouri.edu](https://mugenomicscore.missouri.edu/PDF/FastQC_Manual.pdf#:~:text=%EF%82%B7%20,Statistics%20never%20raises%20an%20error)；它只是資訊摘要，但其結果有助於解讀其他指標。

**2. Per base sequence quality（每個鹼基的序列品質）**

**計算方法**：此模組顯示每個碱基位置上所有讀長中品質值的概況​[mugenomicscore.missouri.edu](https://mugenomicscore.missouri.edu/PDF/FastQC_Manual.pdf#:~:text=3,line%20is%20the%20median%20value)。FastQC以箱形圖（boxplot）方式呈現不同位置的品質分佈：中位數以紅線表示，黃盒代表25%-75%區間（IQR），上下鬚則是10%和90%分位數​[mugenomicscore.missouri.edu](https://mugenomicscore.missouri.edu/PDF/FastQC_Manual.pdf#:~:text=%EF%82%B7%20The%20central%20red%20line,red)。同時，以藍線標示該位置的平均品質​[mugenomicscore.missouri.edu](https://mugenomicscore.missouri.edu/PDF/FastQC_Manual.pdf#:~:text=%EF%82%B7%20The%20central%20red%20line,red)。縱軸為Phred品質分數（越高表示碱基辨識可靠度越高），背景色將品質分數區分為高（綠色）、中等（橙色）、低（紅色）三個區段​[mugenomicscore.missouri.edu](https://mugenomicscore.missouri.edu/PDF/FastQC_Manual.pdf#:~:text=The%20y,the%20end%20of%20a%20read)。x軸從讀長起始到末端，其中為節省繪圖對超長讀序，FastQC在後段將多個位置的數值集成在一起顯示（例如150bp讀長時，後半部可能每5個碱基做一個區間）​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=any%20base%20is%20less%20than,range%20for%2025th%20to%2075th)。品質分數的Phred計算公式為 $Q = -10 \log\_{10}(p)$，p為誤差機率。例如Q20表示1%的錯誤率，Q30約0.1%錯誤率​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=8,Image%2045%3A%20Figure%2046)。

**影響因素**：Illumina測序中，隨著循環數增加碱基品質往往下降，這是正常現象​[hbctraining.github.io](https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lessons/qc_fastqc_assessment.html#:~:text=)。主要原因包括：\*\*(1)**螢光信號衰減（signal decay）：讀長越往後，簇的螢光強度降低，導致3’端品質下降​**[**hbctraining.github.io**](https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lessons/qc_fastqc_assessment.html#:~:text=As%20sequencing%20progresses%20from%20the,as%20the%20sequencing%20run%20progresses)**；**(2)\*\*相位偏移（phasing）：循環累積誤差使同步性降低，也造成尾端品質降低​[hbctraining.github.io](https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lessons/qc_fastqc_assessment.html#:~:text=)。此外，過度聚簇（overclustering）會讓相鄰簇訊號重疊，整體品質變差​[hbctraining.github.io](https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lessons/qc_fastqc_assessment.html#:~:text=,scores%20across%20the%20entire%20read)。通常Read 2（配對讀的第二條）品質低於Read 1，因為Read 2經歷了更長的循環和可能的試劑疲勞​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=scores,Image%2044%3A%20Figure%2045)。任何突然的品質驟降或特定位置出現低品質峰，可能暗示儀器故障（如某循環曝光問題）或樣本問題。

**生物學意義**：每碱基品質圖有助於辨識讀長中是否存在一致性的低品質區段。如果大部分區段品質高且僅末端稍降，屬正常現象，可透過剪除尾端低質碱基改善下游分析​[hbctraining.github.io](https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lessons/qc_fastqc_assessment.html#:~:text=Now%20if%20we%20return%20back,facility%20is%20of%20good%20quality)。若早期循環就出現品質低值或品質劇烈波動，則需調查是否為測序流程問題或樣本DNA品質差所致。高品質的讀序可確保下游比對和解析的可靠性；相反，若大片段低品質（尤其中位數低於Q20），將導致較高錯誤率，需要進一步品質控制如剪裁或過濾。

**預設閾值**：FastQC對此模組設定了綜合判定標準：**任何位置**如果下四分位數（Q1）低於Q10（錯誤率~10%）則觸發警告，低於Q5則判定失敗；或者**任何位置**中位數低於Q25時警告，低於Q20時失敗​[github.com](https://github.com/s-andrews/FastQC/blob/master/Configuration/limits.txt#:~:text=quality_base_lower%20warn%2010)。換言之，若有碱基位置品質偏低（大約≥25%讀數品質<Q10）或整體中位品質跌破25，FastQC會亮黃燈警示；若嚴重到中位數<20或有大量讀數<Q5，則亮紅燈​[github.com](https://github.com/s-andrews/FastQC/blob/master/Configuration/limits.txt#:~:text=quality_base_lower%20warn%2010); [journals.plos.org](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0304158#:~:text=Warning%3A%20a%20warning%20is%20issued,base%20is%20less%20than%2020)。這些閾值反映實務中常用標準（例如Q20作為品質底線，Q30以上視為高質量）​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=8,Image%2045%3A%20Figure%2046)。在正常Illumina數據中，前5-10個碱基的中位品質常略低但通常仍高於Q20，3’末端品質則逐步降低屬正常現象​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=scores,Image%2044%3A%20Figure%2045)。只有當品質曲線明顯低於綠區（高品質）進入橙或紅區，才需特別關注​[mugenomicscore.missouri.edu](https://mugenomicscore.missouri.edu/PDF/FastQC_Manual.pdf#:~:text=The%20y,the%20end%20of%20a%20read)。

**案例**：對於高品質的DNA重測序（WGS）資料，此圖通常大部分區間在綠色高品質範圍，僅末端略降至橙色，FastQC會判定**通過**​[hbctraining.github.io](https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lessons/qc_fastqc_assessment.html#:~:text=Now%20if%20we%20return%20back,facility%20is%20of%20good%20quality)。若測序過程中出現儀器問題（如某幾個cycle故障），圖中可能出現對應位置品質劇跌到紅區，導致**失敗**標記，需要與測序平台聯繫排查​[hbctraining.github.io](https://hbctraining.github.io/Intro-to-rnaseq-hpc-salmon/lessons/qc_fastqc_assessment.html#:~:text=,contacted%20for%20resolution%2C%20if%20possible)。在RNA-Seq樣本中，整體品質趨勢類似，但有時Read 2末端品質下降更明顯；只要中位數未低於Q20，多數讀長仍可用於分析​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=scores,Image%2044%3A%20Figure%2045)。若品質偏低，可以在比對前先用**剪切工具**（例如Trimmomatic、fastp等）剪去低質區段以改善後續結果。總之，每碱基品質圖能直觀檢視讀長品質走勢，及早發現測序品質問題。

**3. Per tile sequence quality（每流水槽瓷磚的序列品質）**

**計算方法**：該模組針對Illumina讀長中的「tile」（流動槽上的小區域）計算品質值分佈​[bioinformatics.babraham.ac.uk](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/12%20Per%20Tile%20Sequence%20Quality.html#:~:text=This%20graph%20will%20only%20appear,one%20part%20of%20the%20flowcell)。Illumina每個讀序ID中包含flowcell座標資訊（如tile編號）；FastQC利用這些資訊將每一循環（base position）的讀序按tile分組，比較同一cycle中不同tile的平均品質是否有系統性差異​[bioinformatics.babraham.ac.uk](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/12%20Per%20Tile%20Sequence%20Quality.html#:~:text=This%20graph%20will%20only%20appear,one%20part%20of%20the%20flowcell)。結果以heatmap形式呈現：顏色由藍(冷)到紅(熱)，表示該tile相對於全體平均品質的偏差程度​[bioinformatics.babraham.ac.uk](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/12%20Per%20Tile%20Sequence%20Quality.html#:~:text=The%20plot%20shows%20the%20deviation,should%20be%20blue%20all%20over)。藍色表示該tile品質等於或高於該循環平均值，紅色表示低於平均​[bioinformatics.babraham.ac.uk](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/12%20Per%20Tile%20Sequence%20Quality.html#:~:text=The%20plot%20shows%20the%20deviation,should%20be%20blue%20all%20over)。理想情況下圖中應該「一片藍」，表示各tile品質均勻；若出現局部**熱點**（橘/紅色區塊），代表那些tile在該循環品質較差​[bioinformatics.babraham.ac.uk](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/12%20Per%20Tile%20Sequence%20Quality.html#:~:text=was%20at%20or%20above%20the,should%20be%20blue%20all%20over)。FastQC透過顏色強度顯示品質差距大小，以識別是否存在局部區域的測序問題。

**影響因素**：每tile品質差異通常反映測序流道上局部位置的問題。例如，暫時性的氣泡通過flowcell會導致該區域短暫品質下降，而持續性的熱點可能由於flowcell表面有污漬或塵粒導致該tile信號品質偏低​[bioinformatics.babraham.ac.uk](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/12%20Per%20Tile%20Sequence%20Quality.html#:~:text=Reasons%20for%20seeing%20warnings%20or,debris%20inside%20the%20flowcell%20lane)。此外，若整個flowcell過度聚簇載量過高，也會增加tile間品質變異​[bioinformatics.babraham.ac.uk](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/12%20Per%20Tile%20Sequence%20Quality.html#:~:text=Whilst%20warnings%20in%20this%20module,which%20persisted%20for%20several%20cycles)。通常，一兩個tile在個別cycle出現輕微偏低品質（如少數紅點）可能是小事件，可忽略；但若許多cycle裡同一tile皆品質偏差，或多個tile持續偏差，則顯示系統性問題​[bioinformatics.babraham.ac.uk](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/12%20Per%20Tile%20Sequence%20Quality.html#:~:text=have%20also%20observed%20that%20greater,which%20persisted%20for%20several%20cycles)。這類問題往往源自於測序儀硬體（例如某通道阻塞）或化學試劑在該區域失效等。

**生物學意義**：該圖主要協助檢測測序硬體層面的問題，對生物學本身沒有直接意義，更多是對技術質量的監控。如果發現明顯的tile品質熱點，說明部分讀序在那些位置可能可靠性較低。用戶可考慮將受影響tile的讀序在下游分析中剔除或標記，以免局部低品質數據影響結果（某些pipeline允許基於tile標記去除讀序）。然而，大多數情況下，tile品質問題只影響少量讀序且難以精確定位去除，因此更常用在測序中心的品質追蹤上。研究人員通常在確認tile品質問題不嚴重時繼續分析，同時向提供測序服務的單位反饋，以避免下次測序重複出現此類問題​[bioinformatics.babraham.ac.uk](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/12%20Per%20Tile%20Sequence%20Quality.html#:~:text=Whilst%20warnings%20in%20this%20module,which%20persisted%20for%20several%20cycles)。

**預設閾值**：FastQC對每tile品質採用相對差值判斷：若任何tile在任一循環的平均Phred分數比全體平均低超過某閾值則觸發警示。官方說明舊版以**低於平均超過2**分作為警告門檻，**超過5**分作為失敗門檻​[bioinformatics.babraham.ac.uk](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/12%20Per%20Tile%20Sequence%20Quality.html#:~:text=This%20module%20will%20issue%20a,that%20base%20across%20all%20tiles)；但實際程式中設定可能稍有不同（FastQC 0.11.9的config設定warn=5、error=10，表示相差5以上警告，10以上失敗）​[github.com](https://github.com/s-andrews/FastQC/blob/master/Configuration/limits.txt#:~:text=The%20per%20tile%20module%20tests,maximum%20phred%20score%20loss%20between)。總之，若看到該模組報*Warn*或*Fail*，代表存在tile品質明顯低於其他區域的情形，需要留意​[bioinformatics.babraham.ac.uk](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/12%20Per%20Tile%20Sequence%20Quality.html#:~:text=Reasons%20for%20seeing%20warnings%20or,debris%20inside%20the%20flowcell%20lane)。輕微且短暫的偏差通常影響有限，可不作嚴重處理​[bioinformatics.babraham.ac.uk](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/12%20Per%20Tile%20Sequence%20Quality.html#:~:text=have%20also%20observed%20that%20greater,which%20persisted%20for%20several%20cycles)；但大範圍持續偏差則應引起重視並可能要求重測該lane。

**案例**：在大多數高品質測序跑道中，此圖**不會出現**警告（全藍）。偶爾在流動槽邊緣tile可能品質稍低但未達警示門檻。若碰到某次測序FastQC報告該圖有多處紅色塊（Fail），如某lane有一整列tile品質偏低，可能是流動槽有擦拭痕或灰塵導致信號受阻​[bioinformatics.babraham.ac.uk](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/12%20Per%20Tile%20Sequence%20Quality.html#:~:text=Reasons%20for%20seeing%20warnings%20or,debris%20inside%20the%20flowcell%20lane)。這樣的案例應與技術人員聯繫，並在下游分析中过濾可疑讀序。對於單細胞測序等需要極高數據品質的實驗，更應謹慎檢查tile品質，以確保資料均勻可靠。

**4. Per sequence quality scores（每序列平均品質分數）**

**計算方法**：此模組統計每條讀序（read）的平均品質分數分佈情況​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=8,Image%2045%3A%20Figure%2046)。具體而言，對於每個讀序，先計算其所有碱基Phred分數的平均值，然後將所有讀序的平均品質分數繪製成直方圖或密度曲線​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=8,Image%2045%3A%20Figure%2046)。x軸為讀序的平均品質值，y軸為具有該平均品質值的讀序數量。該圖反映測序資料中讀序「整體」品質的概況​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=8,Image%2045%3A%20Figure%2046)。通常，我們期望大部分讀序的平均品質很高，集中在右側高分區域，只有少部分讀序平均品質較低。FastQC著重標示最常見的平均品質範圍（頻數最高的區段）。

**影響因素**：每條讀序的平均品質受到該讀序長度和在各個位置品質的綜合作用。如果測序過程順利，絕大多數讀序在全程維持高品質（除了末端少量下降），則平均品質分佈將尖峰出現在較高Phred值（例如Q30左右）。若測序有系統性問題，使大量讀序在某段出現低質量，例如試劑失效導致一批讀序後半部分品質很差，則那些讀序的平均值會大幅降低，拉出一個次峰或長尾分佈到較低品質區。讀序長度也有影響：**(1)** 長讀序往往平均品質略低（因為末端品質下降攤薄了平均值），而短讀序平均品質易偏高；**(2)** 如果資料混合集成不同讀長（如不同跑道合併），平均品質分佈可能寬一些。

**生物學意義**：此指標可快速評估有無大量**整條讀序**質量偏低的情況。如果圖形顯示有明顯的次要峰值在低品質區（例如Q20以下），意味著存在許多讀序的整體品質不佳​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=8,Image%2045%3A%20Figure%2046)。這可能由於樣本品質差或測序儀問題，需要考慮過濾掉這些讀序或改進實驗流程。理想情況下，大部分讀序平均品質應落在高品質區（Q30附近）且分佈單峰集中​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=8,Image%2045%3A%20Figure%2046)。生物學上，我們希望每條序列都可靠，以確保下游分析（例如變異檢測、定量）準確無偏。如果平均品質分佈良好，說明資料整體可信；若有明顯低質讀序群體，則那些讀序可能在比對或組裝中產生錯配，建議在前處理時過濾。

**預設閾值**：FastQC對每序列品質分佈設定閾值以檢測偏低品質的趨勢​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=8,Image%2045%3A%20Figure%2046)。具體來說，**最常見的讀序平均品質值**（直方圖最高峰所在的品質分）若低於Q27則給警告，低於Q20則標記失敗​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=8,Image%2045%3A%20Figure%2046)。Q27約對應0.2%的錯誤率，而Q20約1%錯誤率​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=8,Image%2045%3A%20Figure%2046)。因此若整體樣本中**主要讀序**的品質只達Q25左右，就會觸發Warn，提示平均每讀序錯誤率偏高；若竟然主要峰落在Q20或更低，說明大量讀序品質堪憂​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=8,Image%2045%3A%20Figure%2046)。通常良好的Illumina資料主要峰值在Q30以上，這一模組會Pass。值得注意的是，雙峰分佈（如一部分讀序高品質，一部分很低）有時也會引發警告，研究者應檢視低品質那部分讀序的佔比和原因，是局部污染還是普遍問題。

**案例**：在人類全基因組重測序中，由於讀序長度較長（150bp）且末端品質下降，讀序平均品質可能介於Q30左右，主峰位於高品質區，FastQC通常**通過**。在某些低質樣本（例如降解RNA測序）中，可能有大量讀序含較多錯讀，導致平均品質下降，出現一個偏低的次峰，FastQC會對此**警告**甚至**失敗**。研究人員可據此決定是否舍棄這批質量極差的讀序。RNA-Seq資料中，若文庫品質不均一，可能看到雙峰：一峰高品質（大部分讀序），另一小峰低品質（或對應某lane問題）。這提醒需要檢查測序過程或在比對前濾除低品質讀序以確保分析可靠。

**5. Per base sequence content（每個碱基位置的序列成分）**

**計算方法**：此模組繪製讀序中每個位置（碱基序號）上四種鹼基（A/T/G/C）的比例​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=9,%5BImage%2046%3A%20Figure%2047)。對於讀長的每一個位置，計算在所有讀序中該位置為A的比例、為C的比例，依此類推，得到四條隨位置變化的曲線，每條曲線代表一種鹼基含量百分比​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=9,%5BImage%2046%3A%20Figure%2047)。理論上，若序列來自隨機片段化的複雜文庫，每個位置的堿基組成應近似均等，四條線應彼此平行且約為25%​ [bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=9,%5BImage%2046%3A%20Figure%2047)。FastQC利用偏離程度來判斷序列成分是否異常：線條之間的偏差越大，表示某些位置存在鹼基偏好。

**影響因素**：理想情況下，無論讀序的哪個位置，A/T/G/C的比例都應接近樣本總體的平均碱基比例（通常整體基因組~25%各自，除非物種本身GC含量極端）。然而，**文庫構建過程**和**測序引物**可能引入系統性偏差：

* **隨機引物偏差**：RNA-Seq文庫常用隨機六聚體進行cDNA合成，在理論上應隨機，但實際上某些六聚體比其他更高效，導致讀序起始幾個碱基的成分不均​[sequencing.qcfail.com](https://sequencing.qcfail.com/articles/positional-sequence-bias-in-random-primed-libraries/#:~:text=The%20cause%20of%20this%20bias,primed%20by%20the%20random%20primers)。典型現象是RNA-Seq數據的前~10–12個碱基A/T含量高於G/C，或出現明顯波動​[sequencing.qcfail.com](https://sequencing.qcfail.com/articles/positional-sequence-bias-in-random-primed-libraries/#:~:text=)。這是因為六聚體隨機引物偏好性造成的，屬廣泛觀察到的現象​[sequencing.qcfail.com](https://sequencing.qcfail.com/articles/positional-sequence-bias-in-random-primed-libraries/#:~:text=You%20can%20clearly%20see%20the,a%20greater%20or%20lesser%20extent)。
* **接頭序列讀穿**：若片段長度短於讀長，讀到插入片段末端後開始讀接頭序列，接頭通常富含特定序列（如Illumina接頭常GC豐富）。當讀到接頭時，最後幾個循環的堿基組成會偏向接頭序列，導致末端碱基成分比例異常（例如某位置開始突然G/C比例上升）。
* **文庫構建酶切偏好**：某些基因組文庫（如轉座酶建庫）可能在特定位點切割效率不同，導致序列起點有偏好，例如Nextera文庫常在5'端表現出明顯的碱基偏斜​[dnatech.ucdavis.edu](https://dnatech.ucdavis.edu/faqs/why-does-fastqc-show-unexpectedly-high-sequence-duplication-levels-pcr-duplicates#:~:text=they%20used%20in%20high,Image%201)。
* **非隨機文庫**：若樣本不是全基因組隨機片段，如PCR擴增的產物，則各讀序起始可能相同或具有序列共性，會表現在前幾個碱基成分極不均勻。例如miRNA測序中，大部分讀序開始為T（因為很多成熟miRNA的5'末端為U），則該位置的碱基含量明顯偏斜，這是生物學特性造成的。

**生物學意義**：每碱基序列成分圖檢測文庫序列偏好性，可反映文庫準備中的系統性偏差或污染。**理想狀態**：四條線幾乎平行且重疊，說明序列隨機無明顯偏差。**偏差情況**：如果某些位置A與T或G與C曲線分離明顯，表示該位置存在AT或GC含量不平衡​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=9,%5BImage%2046%3A%20Figure%2047)。少量波動可能由統計誤差或基因組本身組成（如偏AT或偏GC基因組）造成，屬正常。如果看到**明顯的圖形模式**（如起始幾個碱基嚴重偏A/T），需判斷是否文庫構建步驟引起：例如RNA-Seq常見的起始偏差通常不影響下游分析結論​[sequencing.qcfail.com](https://sequencing.qcfail.com/articles/positional-sequence-bias-in-random-primed-libraries/#:~:text=In%20practice%20it%20seems%20that,mappability%20and%20other%20factors%20beyond)。但是，如果偏差持續整個讀序或出現在不該有的區段，可能暗示樣本混入了單一來源序列（如接頭污染或某高豐度外源DNA）。生物學上，嚴重的偏好性可能導致測序覆蓋不均（某些序列因偏好被過度或不足表示），影響定量和發現變異的準確性。然而許多偏差（如RNA-Seq前端的hexamer偏差）已被證實對整體定量影響不大​[sequencing.qcfail.com](https://sequencing.qcfail.com/articles/positional-sequence-bias-in-random-primed-libraries/#:~:text=In%20practice%20it%20seems%20that,mappability%20and%20other%20factors%20beyond)。因此此圖更多是質控提醒，用於識別異常情況並溯源其成因。

**預設閾值**：FastQC將**任一位置**A和T的差異或G和C的差異作為判據：若差值超過10%則標記警告，超過20%則標記失敗​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=9,%5BImage%2046%3A%20Figure%2047)。例如，在某個循環位置，A%=35%、T%=15%（相差20%，且G≈C剩餘），這已達Fail標準。同樣地，如果A遠多於T或者C遠多於G超過閾值，也會報警​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=9,%5BImage%2046%3A%20Figure%2047)。這一閾值設定相對嚴格，因為在隨機文庫中各鹼基比例通常相差不會超過幾個百分點，因此>10%的偏差意味著明顯的不均衡。需要注意的是：RNA-Seq等資料**常**因起始偏差而超過此閾值（往往在前10bp差異>10%），因此FastQC幾乎總是對RNA-Seq的此模組給出*Fail*或*Warn*​[rtsf.natsci.msu.edu](https://rtsf.natsci.msu.edu/genomics/technical-documents/fastqc-tutorial-and-faq.aspx#:~:text=sequence%20data,and%20consider%20what%20that%20results)。用戶應結合實驗設計解讀：若知道是隨機引物導致且僅限開頭幾位，可在此情境下放寬對Fail的擔憂​[sequencing.qcfail.com](https://sequencing.qcfail.com/articles/positional-sequence-bias-in-random-primed-libraries/#:~:text=You%20can%20clearly%20see%20the,a%20greater%20or%20lesser%20extent)。反之，若整段讀序都存在巨大偏差或無明確原因的偏差，則應深入調查文庫準備問題或污染來源。

**案例**：

* **RNA-Seq**：常出現**FAIL**但屬正常情況。幾乎所有RNA-Seq文庫在首端約10–12個碱基顯示明顯的A/T或G/C偏差​[sequencing.qcfail.com](https://sequencing.qcfail.com/articles/positional-sequence-bias-in-random-primed-libraries/#:~:text=)。這是因隨機六聚體引物引起的序列偏差​[sequencing.qcfail.com](https://sequencing.qcfail.com/articles/positional-sequence-bias-in-random-primed-libraries/#:~:text=The%20cause%20of%20this%20bias,primed%20by%20the%20random%20primers)。研究表明此偏差不會嚴重影響基因表達定量​[sequencing.qcfail.com](https://sequencing.qcfail.com/articles/positional-sequence-bias-in-random-primed-libraries/#:~:text=In%20practice%20it%20seems%20that,mappability%20and%20other%20factors%20beyond)。因此對於RNA-Seq，看到該圖Fail可不必驚慌，但仍需確認曲線在之後趨於平行——也就是偏差只局限於序列起點部分[sequencing.qcfail.com](https://sequencing.qcfail.com/articles/positional-sequence-bias-in-random-primed-libraries/" \l ":~:text=You%20can%20clearly%20see%20the,a%20greater%20or%20lesser%20extent" \t "_blank)。
* **WGS DNA-Seq**：如果該圖顯示**警告或失敗**，通常意味有異常。全基因組隨機片段理應平坦。如果某WGS樣本顯示明顯鹼基偏差，可能是文庫污染了高單一序列（如接頭未剪或含有某高拷貝的載體片段）。例如，有報告顯示甲基化測序(PBAT)因引物造成5’端強烈序列偏斜​[sequencing.qcfail.com](https://sequencing.qcfail.com/#:~:text=QC%20Fail%20Sequencing%20Random%20priming,As%20a%20result)。對於DNA文庫，此模組Fail往往需要進一步排查實驗流程。
* **小RNA或定向文庫**：如miRNA測序，由於生物序列本身特性，開頭幾位可能嚴重偏某一鹼基（很多miRNA首位為U），FastQC會報Fail。但這是樣本真實特徵，不是技術錯誤。對策是理解此結果的生物學意義（如某核苷酸偏好）而非認為文庫失敗。

**6. Per sequence GC content（每序列GC含量分佈）**

**計算方法**：此模組計算每條讀序（read）的GC百分比（即該read中G和C所占比例），然後將所有讀序的GC含量分佈製成直方圖​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=10,%5BImage%2047%3A%20Figure%2048)。FastQC同時構建一個理論的GC含量分佈作為對照：假定隨機序列且整體GC含量等於樣本平均GC，則讀序GC分佈應近似**正態**分佈，以該平均值為中心​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=10,%5BImage%2047%3A%20Figure%2048)。FastQC將實測分佈與理論正態分佈進行比較。如果實測分佈與理論期望吻合較好，圖中藍色實線（實測）與灰色背景曲線（理論）重合；若偏離，則表示GC含量有異常模式。FastQC累積計算實測和期望之間的偏差量，用於判斷是否超出閾值​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=10,%5BImage%2047%3A%20Figure%2048)。

**影響因素**：讀序的GC含量由樣本本身序列組成決定。對於**全基因組隨機片段**，每條讀序相當於從基因組隨機取樣一定長度，因此各read GC%應圍繞基因組整體GC含量呈正態散佈（中心為基因組平均GC，高斯形狀）。偏離正態的情形常見於：

* **樣本混雜**：如果樣本包含來自不同GC含量的來源（例如基因組混合物，或污染了另一物種DNA），則GC分佈可能出現**雙峰或寬峰**。例如，人類DNA（約40% GC）中混入細菌DNA（例如60% GC），則一部分reads的GC落在60%左右，整體分佈呈雙峰。
* **文庫選擇偏好**：某些實驗對特定GC含量片段有選擇性。例如PCR可能偏好擴增中等GC片段，高GC片段可能擴增效率低導致不足，從而reads GC分佈窄化或偏移。再如捕獲實驗可能富集特定GC範圍的序列。
* **測序偏差**：極端GC片段在Illumina中可能測序效率較低（AT極高或GC極高的序列可能聚簇效果差），導致文庫中實際讀到的序列GC分佈與原樣本不同。
* **物種基因組本身**：有些基因組GC含量分佈並非單峰正態。例如，真核基因組可由高GC的外顯子與低GC的內含子組成，使隨機片段GC分佈出現某種*肩峰*。

**生物學意義**：GC含量是序列的基本組成特徵。**正常情況**：單一物種基因組文庫應呈現單峰分佈，峰位置約在該物種基因組平均GC附近​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=10,%5BImage%2047%3A%20Figure%2048)。**異常情況**：偏離期望分佈通常暗示：

* **污染或樣本混合**：多峰分佈強烈暗示樣本中存在不同來源DNA。例如環境樣本可能混有細菌與宿主DNA，不同峰對應不同GC特徵的生物。此時需甄別外源污染（如根據GC峰定位可能判斷出是否混入特定微生物）。
* **文庫偏差**：如果分佈明顯偏窄，可能代表某些序列在文庫構建中被偏愛或排除。例如，GC極端的序列沒被有效擴增，導致文庫GC範圍變窄。這可能影響全面性，比如在基因組或轉錄組測序中，GC偏差會導致覆蓋深度與GC含量相關聯，進而影響定量和組裝。  
  總的來說，極端的GC偏差會對**比對**和**變異檢測**產生影響（如高GC區域reads少，低GC區域reads少，造成假性的覆蓋坑谷）。若發現GC分佈異常，需要視情況決定：是樣本本身組成使然（如目標序列GC特殊）還是技術問題，需要補充實驗或過濾特定reads。

**預設閾值**：FastQC採用一個統計量：實測分佈相對理論分佈的偏離和。​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=10,%5BImage%2047%3A%20Figure%2048)中提到“偏離的reads比例超過一定閾值”作判定。具體地，FastQC將每一個GC含量值上實測頻數與理論頻數的差取絕對值累加，若\*\*偏離累積超過總reads的15%\*\*則警告，超過30%則失敗​[github.com](https://github.com/s-andrews/FastQC/blob/master/Configuration/limits.txt#:~:text=The%20per%20sequence%20GC%20content,tests%20the%20maximum%20deviation%20between)。簡言之，如果有很多reads的GC含量明顯偏離正態曲線，占比超過15%，就會亮黃燈​[github.com](https://github.com/s-andrews/FastQC/blob/master/Configuration/limits.txt#:~:text=The%20per%20sequence%20GC%20content,tests%20the%20maximum%20deviation%20between)；若幾乎有三分之一以上的reads落在不期望的GC區間，則亮紅燈​[github.com](https://github.com/s-andrews/FastQC/blob/master/Configuration/limits.txt#:~:text=The%20per%20sequence%20GC%20content,tests%20the%20maximum%20deviation%20between)。這一設定試圖量化“非隨機”程度。需要指出，在測序深度很高時，即便存在偏差，比例可能不大；反之在樣本基數小時，少量污染也可超出比例。用戶應結合分佈形狀判斷：例如雙峰即使各佔10%和90%，總偏差20%可能fail或warn，這顯示明確的混合來源。而若只是峰寬略寬導致>15%的偏離，可能只是基因組特性，不一定嚴重問題。

**案例**：

* **WGS**：應呈單峰。若FastQC報**Fail**，往往是明顯的異常，如樣本混有其他物種DNA，導致雙峰GC分佈。文獻案例顯示，某些腸道菌基因組混入人類DNA時，在FastQC GC圖會看到兩個峰，一個約40%（人體基因組），一個~50-60%（細菌）​[mugenomicscore.missouri.edu](https://mugenomicscore.missouri.edu/PDF/FastQC_Manual.pdf#:~:text=and%20there%20are%20no%20problems,which%20can%20spot%20problems%20which)。此時需考慮實驗流程（可能樣本未純化乾淨）並視分析目的決定是否分離讀序。
* **RNA-Seq**：轉錄組的GC分佈通常仍近似單峰，但由於轉錄本序列不均勻（例如高度表達的mRNA可能GC偏中值區，而低表達可能偏端），分佈形狀可能稍有偏斜或肩峰。然而FastQC的模型未考慮表達量差異，所以有時轉錄組FastQC的GC模組會**Warn**。通常如果只有輕微偏差（如峰位置符合已知轉錄組GC均值，只是曲線較寬），不影響下游分析。若出現雙峰，要留意是否rRNA或其他高GC/低GC片段富集。
* **ChIP-Seq**：ChIP針對基因組中特定位點富集，理論上讀序GC分佈應接近基因組背景，但實際可能偏向富集區域的GC特性。例如某轉錄因子傾向結合AT富含區域，則pull-down的片段AT含量高，reads GC偏低峰。若這差異足夠大，也可能引致FastQC警告。ChIP數據中看到GC偏移需判斷是否與靶點序列有關；若是，則屬生物學訊號；若無關，可能暗示背景雜質。

**7. Per base N content（每個碱基位置的N比例）**

**計算方法**：此模組顯示每一讀序位置上，未能確定的鹼基“N”出現的頻率​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=11,%5BImage%2048%3A%20Figure%2049)。對每個碱基位置，計算該位置為N的讀序比例，並繪製成隨位置的曲線​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=11,%5BImage%2048%3A%20Figure%2049)。理想情況下，測序儀對每個循環都能讀出明確的A/T/G/C，因此N含量應接近0%；如果某些位置上多條讀序出現N，那在圖上該位置的N百分比會升高。

**影響因素**：出現“N”通常表示測序儀在該循環無法判定明確的鹼基（信號強度低或混亂）。可能原因包括：

* **儀器問題**：某一cycle試劑或成像問題，導致普遍信號不佳，許多讀序在該cycle記錄為N。這種情況會表現在N含量圖上形成明顯“峰”，例如在第X位碱基處N比例突然上升。
* **過度聚簇**：簇太密導致信號疊加，使basecalling困難，有些讀序以N代替。
* **樣本因素**：理論上，生物學序列本身不會含N（N僅表示未知），但假如文庫製備過程中引入了N（例如接頭序列不明確或barcode decode失敗），也可能導致特定位置固定出現N。

**生物學意義**：N含量高直接影響下游分析，因為N作為不確定鹼基，無法準確比對到參考序列，並可能中斷拼接。高比例的N會降低有效信息量。正常樣本中N比例應趨近於0%；任何明顯上升都應引起注意。**小幅N增加**（如<1%）在尾端可能出現，反映測序末端信號變弱，這通常可以透過剪除尾端數據解決。**嚴重N峰值**則明顯異常，意味著該cycle可能失敗或存在諸如試劑氣泡等問題。生信分析時遇到很多N可能需要過濾或做N的填補處理。不過N含量圖主要是檢查測序流程問題的技術指標，與樣本生物學特性無關（因為生物序列不應有N）。

**預設閾值**：FastQC對N含量設定硬閾值：**任何位置**N比例>5%即給警告，>20%即標記失敗​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=11,%5BImage%2048%3A%20Figure%2049)。因此，只要有一個碱基位置超過1/20的讀序是N，便觸發Warn；若超過1/5為N則Fail​

[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=11,%5BImage%2048%3A%20Figure%2049)。這反映了對未知碱基的嚴格要求——理論上連5%都已相當高，說明很多序列在那一位無法確定​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=11,%5BImage%2048%3A%20Figure%2049)。一般資料幾乎所有位置N% = 0，如有上升通常也不會到5%，除非測序明顯有問題。

**案例**：絕大多數Illumina資料**通過**此模組，因為N含量通常極低。有些低品質run，可能看到末幾位逐漸有N出現，例如最後5個碱基N含量上升到1-2%，FastQC尚不至警告，但提醒我們尾端品質差​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=7,not%20uniform%2C%20it%20starts%20out)。如果碰到某測序報告第50個碱基N含量達10%（Warn）或更高，極可能是該cycle出了技術問題。曾有案例報導一整個Flowcell在某cycle成像失敗，導致所有讀序第N位都是N，那在N含量圖上會是一個100%的尖峰（肯定Fail）。這種情況下數據須謹慎處理（例如將該位當作缺失位，不用于分析，或考慮重新測序）。總之，任何N含量超標都表示技術瑕疵，應了解其範圍和原因，以決定對後續分析的影響範圍。

**8. Sequence Length Distribution（序列長度分佈）**

**計算方法**：此模組統計所有讀序的長度，繪製讀長分佈直方圖​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=12,Image%2049%3A%20Figure%2050)。對於固定讀長的平台（如Illumina常見的固定讀長），理論上所有讀序長度相等，圖上會呈現單一的值；對於長度可變的資料（如PacBio或經過剪裁的reads），則會出現一個範圍的分佈。FastQC在報告中顯示最短和最長讀序長度（Basic Statistics已列出）之外，還提供此圖以觀察是否存在異常的長度模式​[mugenomicscore.missouri.edu](https://mugenomicscore.missouri.edu/PDF/FastQC_Manual.pdf#:~:text=analysis,only%20one%20value%20is%20reported); [bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=12,Image%2049%3A%20Figure%2050)。

**影響因素**：讀序長度主要由測序儀讀取循環數和前處理決定。Illumina常設定讀長（如2×150bp），因此原始資料一般都是固定長度。如果在fastq中看到不同長度，通常是因為**質控剪裁**或**接頭截斷**：一些讀序在測序過程因品質過低被終止（稀少情況），或經過剪枝軟體（trimmer）切除了低質或接頭序列，導致短於原始長度。對於長讀長技術（Nanopore/PacBio），讀長本身固有可變，FastQC同樣能顯示其長度分佈，例如數百至數千鹼基的區間範圍。如果讀長分佈呈現出多峰，比如一部分reads極短（可能為接頭二聚體或明顯失敗reads），則值得注意。

**生物學意義**：讀長分佈主要反映技術狀況而非生物特性。在Illumina實驗中，我們期望讀長一致。如果發現大部分reads為150bp，卻有相當部分reads只有50-100bp，這意味著文庫中有不少reads被過早終止或後處理剪短，可能源自於大量低品質區段或接頭污染需要剪除。這對下游**比對**和**組裝**有影響：過短reads可能對基因組的唯一性比對造成困難。對於一些應該固定長度的實驗（如定長的擴增子），如果發現長度分佈寬泛，可能表示非預期的片段出現（如非目標產物）。

**預設閾值**：FastQC對序列長度的檢查比較簡單：**如果所有序列未保持相同長度**則給警告；**如果存在長度為0的序列**則報錯​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=12,Image%2049%3A%20Figure%2050)。也就是說，只要檔案中reads長度不全一致，FastQC形式上會標示一個Warn（提示有多種長度）；實際這通常不嚴重，除非長度差異很大。Fail條件只有序列長度為0發生時——那表示明顯的錯誤（空序列），正常情況不應存在​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=12,Image%2049%3A%20Figure%2050)。由於現代實驗中常對reads進行質控處理（剪除接頭和低質部分）導致長度不一，因此對很多下游分析者而言，看到Sequence Length Distribution模組報Warn並不意外。這主要是FastQC提醒我們資料非均一，需留意原因。

**案例**：

* **Illumina原始數據**：若無任何處理，理論上應**Pass**（所有reads長度相等）。實際中，某些儀器在run末尾可能產生少量過短reads（可能由於簇過早消失），但佔比極低通常不影響。FastQC仍可能標Warn（因為有不同長度存在）。
* **經過Trimmomatic/fastp剪裁**：常見情況是**Warn**。因為剪裁去除了接頭或低質末端，不同reads被剪成不同長度。例如剪裁後reads在50-150bp都有分佈，FastQC報Warn但這是預期結果，不是問題。
* **混合資料**：如果將不同實驗（不同讀長）合併分析，FastQC也會看到多個長度峰。比如合併75bp和150bp的reads，圖上兩峰，這也是Warn。然而這提醒用戶在分析時需區分處理不同長度reads，或至少知道來源不同以免影響均一性。
* **異常情況**：Fail幾乎不該發生，除非輸入fastq格式有誤。有報告提及在某些錯誤轉換的檔案裡可能出現零長度條目，此時應修正檔案。

**9. Sequence Duplication Levels（序列重複水平）**

**計算方法**：此模組衡量序列在讀序集合中的重複程度​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=13,%5BImage%2050%3A%20Figure%2051)。FastQC會將**完全相同**的讀序序列視為重複（不考慮品質分數，只看序列本身），統計每個唯一序列出現的次數。例如一條序列出現5次則歸為「5重覆」一類。接著，繪製一個圖：x軸為重複度（出现次数），y軸為該重複度的讀序數佔總reads的比例​[sequencing.qcfail.com](https://sequencing.qcfail.com/articles/libraries-can-contain-technical-duplication/#:~:text=In%20this%20plot%20you%20can,library%20were%20to%20be%20deduplicated)。通常圖中包括“一次出現的序列”佔比、出現2次的佔比、…直到高重複（FastQC會綜合顯示>10次以上歸為一類）。另外，FastQC報告一個「序列去重後剩餘%」，表示如果移除所有重複（僅保留唯一序列），仍保留的reads比例​[sequencing.qcfail.com](https://sequencing.qcfail.com/articles/libraries-can-contain-technical-duplication/#:~:text=from%20sequences%20which%20occur%20different,library%20were%20to%20be%20deduplicated)。這個百分比低意味著有大量重複reads。

**影響因素**：**理論上**，對於理想的隨機文庫（如WGS），每個read序列出現一次的比例應非常高，因為基因組龐大且片段覆蓋不會完全重複​[sequencing.qcfail.com](https://sequencing.qcfail.com/articles/libraries-can-contain-technical-duplication/#:~:text=In%20this%20plot%20you%20can,library%20were%20to%20be%20deduplicated)。然而，實際中重複reads可能來自：

* **技術重複（PCR duplicates）**：文庫製備經PCR擴增，多次擴增同一模板會產生多條相同序列的讀序。PCR循環數越多，過擴增的風險越高，技術重複比例上升​[seqanswers.com](https://www.seqanswers.com/forum/bioinformatics/bioinformatics-aa/13952-what-might-cause-the-sequence-duplication-levels-failures-in-fastqc-report#:~:text=What%20might%20cause%20the%20,sequencing%20%28very%20high%20fold)。這在ChIP-Seq、目標捕獲等涉及擴增的實驗中特別要注意。
* **測序深度過高（over-sequencing）**：即樣本複雜度有限但測序量過大，導致**生物學重複**。例如，對一個小基因組（病毒、細菌）做超高深度測序，最終所有可能的片段都讀取多次​[sequencing.qcfail.com](https://sequencing.qcfail.com/articles/libraries-can-contain-technical-duplication/#:~:text=requires%20sufficient%20read%20depth%20to,an%20indication%20of%20technical%20duplication)。這不一定是文庫準備問題，而是因為序列空間有限，重複讀序在所難免。極端例子：將PhiX噬菌體文庫跑滿一個lane，所有reads都來自同一小基因組，理論上每個可能片段都覆蓋數千次​[sequencing.qcfail.com](https://sequencing.qcfail.com/articles/libraries-can-contain-technical-duplication/#:~:text=requires%20sufficient%20read%20depth%20to,an%20indication%20of%20technical%20duplication)，FastQC的duplication圖會顯示非常高的重複比例，但這屬於**預期的生物學重複**而非技術問題。
* **序列富集偏向**：轉錄組或ChIP等測序中，coverage不均一，某些區域讀序極為密集。對RNA-Seq，高表達基因會產生許多片段落在同一區域，即使斷片起點不同，也有可能部分片段序列完全相同（尤其是短reads情況下）​[sequencing.qcfail.com](https://sequencing.qcfail.com/articles/libraries-can-contain-technical-duplication/#:~:text=The%20observation%20of%20duplication%20is,of%20duplication%20plots%20is%20very)。ChIP-Seq中熱門結合位點周圍片段堆積，也可能產生一些重複reads。這些重複有部分是**生物訊號**（真實高豐度片段），但大量重複也可能反映文庫複雜度不足。
* **隨機巧合**：在數據量很大時，即使沒有系統原因，不相關的兩個片段恰巧序列相同也有微小概率。但對於Illumina 100bp以上reads，這概率極低，可以忽略。

**生物學意義**：重複讀序的存在意味著**有效覆蓋度**降低（相同片段多次測序不提供新信息）。對**技術重複**而言，重複的reads並非獨立證據，會**高估**定量（如計數突變或算覆蓋時重複算多了）。因此下游分析通常會**標記或去除PCR duplicates**（特別在DNA重測序做變異檢測時）。然而對**生物學重複**（如高表達片段）則不應盲目去重，否則會丟失真正的豐度訊號​[seqanswers.com](https://www.seqanswers.com/forum/bioinformatics/bioinformatics-aa/13952-what-might-cause-the-sequence-duplication-levels-failures-in-fastqc-report#:~:text=What%20might%20cause%20the%20,sequencing%20%28very%20high%20fold)。FastQC的重複水平圖提供一個總覽：高重複比例提醒我們文庫可能**複雜度不足**。例如，如果去重後僅剩下50%甚至更少的reads，說明一半以上的測序是浪費在重覆片段上​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=high%20level%20of%20duplication%20is,%5BImage%2050%3A%20Figure%2051)。這可能需要提高文庫複雜度（減少PCR循環或增加起始物料）或在分析中考慮這點（如標記duplicates）。也需區分**實驗類型**：全基因組測序希望低重複，而目標序列測序（如16S擴增子）本身所有reads幾乎都是重複（同一菌株的相同序列反覆測），那是預期的，不代表數據不好。

**預設閾值**：FastQC對重複水平設定的判斷基於「去重後剩餘%」：若非唯一序列（即重複reads）佔總reads超過20%，則*Warn*；超過50%，則*Fail*​ [bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=high%20level%20of%20duplication%20is,%5BImage%2050%3A%20Figure%2051)。等價地，唯一序列少於80%時警告，少於50%時失敗。這些閾值對WGS較合理——一般全基因組文庫唯一序列應>80%。但對於轉錄組、ChIP等富集序列實驗，20%-50%的重複可能習以為常​[biostar.galaxyproject.org](https://biostar.galaxyproject.org/p/27279/index.html#:~:text=%2A%20ChIP,specific%20reads)。因此看到Warn/Fail需結合實驗性質解讀：在WGS可能意味著過度PCR或文庫有問題；在ChIP-Seq 50% duplicates可能正常​[biostar.galaxyproject.org](https://biostar.galaxyproject.org/p/27279/index.html#:~:text=%2A%20ChIP,specific%20reads)。FastQC本身無法分辨技術或生物重複，但其預設閾值是假定高重複通常不期望。

**案例**：

* **全基因組重測序（WGS）**：良好文庫通常>80%唯一reads（<20%重複），FastQC此模組**通過**或僅輕微Warn。如果發現Fail（如只剩40%唯一），多半是文庫問題（PCR過度或樣本量不足導致高重複）。需在分析時去重，並反思文庫構建過程。
* **RNA-Seq**：經常**警告或失敗**。由於RNA表達差異巨大，常見現象是少數高豐度轉錄本佔據大量reads。例如有報告稱，RNA-Seq資料中最表達最高的10個基因可能佔到50%的reads​[dnatech.ucdavis.edu](https://dnatech.ucdavis.edu/faqs/why-does-fastqc-show-unexpectedly-high-sequence-duplication-levels-pcr-duplicates#:~:text=assumed%20from%20read%20coverage%20data,In%20addition)。這導致FastQC計算的重複率非常高（因為許多reads序列可能重疊在這些轉錄本上）​[dnatech.ucdavis.edu](https://dnatech.ucdavis.edu/faqs/why-does-fastqc-show-unexpectedly-high-sequence-duplication-levels-pcr-duplicates#:~:text=it%20is%20not%20uncommon%20that,babrah)。FastQC只能逐條read看，無法知道這些重複reads其實對應不同位置的同一基因，因此會高估重複程度​[dnatech.ucdavis.edu](https://dnatech.ucdavis.edu/faqs/why-does-fastqc-show-unexpectedly-high-sequence-duplication-levels-pcr-duplicates#:~:text=data.%20All%20DNA,the%20ten%20most%20highly%20expressed)。對RNA-Seq，70%甚至更高的「表觀重複」也不罕見​[dnatech.ucdavis.edu](https://dnatech.ucdavis.edu/faqs/why-does-fastqc-show-unexpectedly-high-sequence-duplication-levels-pcr-duplicates#:~:text=assumed%20from%20read%20coverage%20data,In%20addition)。不過這不應直接用PCR duplicates的標準去處理；一般RNA-Seq不在FASTQ層面去重，而是在比對到轉錄組後再評估。**因此，RNA-Seq樣本看到Fail需認識到這可能是正常的生物學現象**，建議使用更進階的工具（如Fastp、HTStream考慮配對信息）來估算實際重複​[dnatech.ucdavis.edu](https://dnatech.ucdavis.edu/faqs/why-does-fastqc-show-unexpectedly-high-sequence-duplication-levels-pcr-duplicates#:~:text=consequence%20FASTQC%20tends%20to%20generate,enzymatic%20reactions%2C%20all%20have%20some); [dnatech.ucdavis.edu](https://dnatech.ucdavis.edu/faqs/why-does-fastqc-show-unexpectedly-high-sequence-duplication-levels-pcr-duplicates#:~:text=it%20is%20not%20uncommon%20that,babrah)。
* **ChIP-Seq**：高重複也**常見**。一個簡單ChIP-Seq實驗，FastQC報告50-60%重複是可能的​[biostar.galaxyproject.org](https://biostar.galaxyproject.org/p/27279/index.html#:~:text=%2A%20ChIP,specific%20reads)。因為特定結合位點富集片段，反覆測序那些區域。分析ChIP時通常在比對後根據位點濃度和峰解析來處理重複（可允許一定水平重複但過高可能影響peak calling）。FastQC的Fail提醒我們文庫獨立片段有限，但ChIP的目的本就找重複的peak。實務上，如果重複率極高（>80%），則可能peak以外幾乎沒有覆蓋，潛在問題是輸入DNA量太少或PCR擴增偏好性強，需要更多實驗重複。
* **擴增子測序**：例如16S微生物組或目標基因擴增子，所有reads來源有限的PCR產物，重複必定極高（甚至>90%序列都是重覆的短片段）。FastQC此模組會Fail，但這完全預期，因為我們**刻意**擴增了那些序列。此時不應將Fail視為數據質量問題。

**10. Overrepresented sequences（過度代表序列）**

**計算方法**：FastQC會檢查整個read集合中出現次數異常高的序列（read）​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=14.%20A%20normal%20high,%5BImage%2051%3A%20Figure%2052)。它掃描所有讀序並計算每個唯一序列的出現頻率，列出佔比超過一定閾值的序列作為「過度代表」列表​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=14.%20A%20normal%20high,%5BImage%2051%3A%20Figure%2052)。報告中會顯示這些序列的序列本身、出現次數和百分比，並嘗試標註其可能來源（例如比對常見接頭或rRNA等庫，若匹配則註明“Adapter”或某基因等）。通常情況下，我們預期沒有任何單一序列占比明顯，因為理想文庫應該極其多樣，每條read都不同​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=14.%20A%20normal%20high,%5BImage%2051%3A%20Figure%2052)。因此，任何超常頻率的序列都值得關注。

**影響因素**：出現過度代表序列的原因通常有：

* **接頭或接頭二聚體**：測序接頭殘留是*最常見*的過度序列。若文庫插入片段過短或存在接頭二聚體（兩接頭直接連接的假插入），測序時很快就讀完整DNA進入接頭序列。結果大量讀序包含接頭片段甚至純接頭序列，這幾乎一模一樣，因而高頻​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=15,%5BImage%2052%3A%20Figure%2053)。FastQC常將此類序列標註為Illumina接頭。
* **重複的PCR引物序列**：有時文庫構建使用了特定引物，若引物不完全去除，也可能在reads中反覆出現。
* **高豐度天然序列**：在某些樣本中，可能存在極高拷貝數的分子。例如：
  + **RNA-Seq**中如果rRNA去除不完全，殘留rRNA序列將非常高頻（因rRNA在總RNA中佔比極高）。此時某些rRNA片段read可能出現成千上萬次。FastQC可能將其標為某rRNA基因的片段。
  + **病原測序**中，如果樣本裡某病毒極高載量，也會出現大量相同或高度相似的reads。
  + **低複雜度序列**：例如poly-A尾巴（AAAA...）可能在轉錄組或DNA末端出現，雖不同read長度可能不同，但如果某read幾乎全是A，也可能被檢出頻率高。
* **其他污染**：環境或實驗過程引入的污染，如蘿蔔DNA片段、實驗室常見載體序列等。如果污染片段集中一種序列，也會高頻出現。

**生物學意義**：過度代表序列通常意味著文庫**多樣性降低**或**存在污染/殘留**。在理想隨機文庫中，每條read的序列應當獨一無二或至少非常少重複，因此出現比例超過0.1%以上的單一序列都是異常​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=14.%20A%20normal%20high,%5BImage%2051%3A%20Figure%2052)。生物學上，如果這些序列對應**真實高豐度分子**，則反映樣本組成（例如樣本被rRNA主導）；這可能需要在實驗上解決（如加強rRNA去除）或者在分析時剔除，以免影響對低豐度信號的檢測。若屬**技術污染**（如接頭），則這些序列不具有生物意義，必須在分析前清理，不然會嚴重干擾下游（例如比對時幾乎全是接頭序列無法正確比對，或浪費資源）。總而言之，過度序列列表提供了具體的可疑序列，研究者應檢查其來源。如果是接頭/引物殘留，可透過**修剪工具**將其剪除。如果是生物學原因（如某轉錄本極高表達），則可在分析中考慮這些序列對下游的影響（或在必要時在下游分析中降權重）。

**預設閾值**：FastQC預設將\*\*任何佔比超過0.1%**的序列列入報告清單​**[**bv-brc.org**](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=significant%2C%20or%20that%20the%20library,%5BImage%2051%3A%20Figure%2052)**。同時，若發現**>0.1%**的序列存在，報告整體標記Warn；若有序列**>1.0%\*\*佔比，則標記Fail​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=significant%2C%20or%20that%20the%20library,%5BImage%2051%3A%20Figure%2052)。0.1%對應於在100萬reads中出現1000次，相對已是非常高的重複頻率。1%以上更加明顯（每百條reads至少一條是一模一樣的序列）​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=significant%2C%20or%20that%20the%20library,%5BImage%2051%3A%20Figure%2052)。因此，只要有任何read序列超過這些門檻，FastQC就認為文庫多樣性可能有問題。這一標準對大多數基因組/轉錄組是合理的——正常情況不應有單一序列這麼多次。但對於預期有高重複的實驗，如擴增子，小心解讀：例如16S擴增子序列重複非常常見，此模組可能列出那些擴增子序列且標Fail，但那實際上是我們感興趣的序列而非污染。

**案例**：

* **接頭污染**：最典型例子，FastQC列出ILLUMINA.adapters相關序列，高度重複。很多初學者的FastQC報告中，Overrepresented序列常標示為接頭序列，佔比可能數％甚至更多（尤其是讀長遠超過片段長度時）。這時FastQC多半**Fail**該模組。解決辦法是在**比對前**用專門工具去除接頭序列（Adapter trimming）。如案例顯示，如果>10% reads含接頭未剪，FastQC直接Fail​ [bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=15,%5BImage%2052%3A%20Figure%2053)，數據需要清理。
* **rRNA序列**：在一份mRNA-Seq資料中，如果rRNA去除不完全，FastQC可能在過度序列中列出一些18S或28S rRNA片段。這些序列可能各自佔幾十分之一百分比，但累計不少。報告通常**Warn**。研究者看到具體序列是rRNA時，就能判斷需不需要在分析時把那些rRNA reads排除，或在實驗上改進rRNA去除。
* **重複的引物或條碼**：有時錯誤的文庫製備可能導致讀序開頭就是某固定序列（除了接頭以外）。如果該序列長度足夠被FastQC當成一整個read出現多次，也會上榜。例如某擴增實驗所有reads都帶相同開頭10bp且讀長短，那些reads本身可能都相似。這情況下整個模組肯定Fail，但實際上這可能是設計使然（例如所有reads前面那段是序列標籤）。
* **小RNA高豐度**：miRNA-Seq中，一條特定的miRNA序列如果極高表達，可能上榜。舉例而言，miR-21在某細胞中非常豐富，以致其序列佔總reads的2%。FastQC將其列為過度序列並Fail。但這其實是生物結果（真實的miRNA豐度），分析時並不會將其移除，只是要知道文庫被幾個主要miRNA主導是正常的。

**11. Adapter Content（接頭序列含量）**

**計算方法**：此模組專門檢測測序讀序中接頭（adapter）序列的殘留情況​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=15,%5BImage%2052%3A%20Figure%2053)。FastQC內建常用接頭序列庫（依據用戶指定的文庫類型，如Illumina TruSeq接頭等），透過尋找這些接頭的子序列（k-mer）來判斷在讀序的不同位置是否含有接頭。報告以一張累積圖表示：x軸為讀序位置，y軸為**累積比例**，代表從該位置開始檢測到接頭序列的reads比例​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=15,%5BImage%2052%3A%20Figure%2053)。曲線通常隨讀長增加而上升：在讀序末端越往後，累積越多reads出現接頭。直觀來說，如果很多片段比讀長短，讀到某長度後接頭出現頻率升高。FastQC將不同接頭序列分別繪線（如測多種接頭則多條線），通常Illumina文庫只會有一條主要接頭曲線。

**影響因素**：接頭序列出現在reads中有幾種情況：

* **插入片段短於讀長**：如果文庫中DNA片段長度分布有一部分小於測序讀長（例如建庫產物長度100bp，但跑150bp讀序），那麼對於那些短片段，讀完插入後接著就會讀到接頭序列。片段越短，接頭出現在read中的位置越靠前。因此Adapter Content圖會在接近片段長度的位置開始上升。這是常見原因，如pair-end測序中第二條read經常讀入接頭因為插入不夠長。
* **接頭二聚體（adapter dimer）**：建庫時如果有接頭沒有接上DNA插入而自連形成接頭-接頭小片段，就可能被擴增並測序。讀這種片段時，開頭就是接頭序列本身，導致整條read幾乎都是接頭。如果接頭二聚體比例不低，Adapter Content圖在read一開始就高企。
* **測序索引錯誤分離**：有時混樣本上機需要用接頭條碼（index），如果實驗設計或解碼錯誤，接頭序列可能出現在讀序中（例如沒分離好index read混入主read）。不過這通常不體現在Adapter Content圖，而是在Overrepresented sequences中被檢出。
* **接頭類似序列**：極少數情況，樣本本身的序列中可能包含與接頭序列相同的motif。如果該motif頻繁出現在特定位置，也可能被算法當成接頭。但FastQC對adapter檢測比較嚴格，誤報機率低。

**生物學意義**：接頭序列本身是人工序列，殘留在reads中沒有生物學意義，但對下游分析**危害大**。如果不移除接頭，這些reads在與參考基因組比對時可能**完全不對齊**（因為接頭與基因組不匹配），或錯對到其他位置（若接頭序列碰巧匹配某基因組區段，產生假比對）。另外，在**裝配**時，接頭序列會形成虛假的重疊，干擾拼接。Adapter Content圖揭示我們是否需要在分析前剪除接頭：理想情況應趨近0（曲線幾乎不抬升），如果顯示有明顯的接頭累積，就表示**必須進行接頭修剪**。一般規則是，只要有>5-10%的reads受接頭污染，就應該清理​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=15,%5BImage%2052%3A%20Figure%2053)。

**預設閾值**：FastQC設定當**任意位置**有超過5%的讀序含接頭序列時給警告，超過10%時標記失敗​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=15,%5BImage%2052%3A%20Figure%2053)。也就是說，如果接頭累積曲線在讀序末端達到5%，則Warn；若達到10%，Fail​ [bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=15,%5BImage%2052%3A%20Figure%2053)。這相當嚴格但合理，因為即使只有5%reads有接頭，對後續比對也可能造成上百萬條reads對不上參考，需要處理。特別地，FastQC報告中若Adapter Content模組出現黃或紅標誌，通常意味著需要執行**接頭剪除**程序。值得一提，FastQC會根據文庫類型選擇搜尋相應的接頭序列（例如small RNA文庫還會找接頭序列，不同於標準Illumina接頭），以減少漏報。

**案例**：

* **標準Illumina文庫**：假設片段長度為200bp配對讀150bp×2，那麼第二條讀序幾乎總是讀不進接頭（因為兩端加起來300bp大於插入200bp，確定會重疊接頭）。FastQC Adapter Content圖通常在Read2末端**上升明顯**。如果在150bp處達到比如20%，那Fail。對策是在比對前用工具剪去這些接頭尾巴。Read1則可能無接頭或很少（除非插入<150）。
* **小片段文庫**：如ChIP-Seq片段有時很短（50-100bp），150bp讀序很早就進入接頭，Adapter Content曲線也會更早升起甚至中段就超10%。這時FastQC肯定Fail，但也提醒我們很多reads後半部分沒有效信息。剪掉接頭後，實際可用的read長度可能只有前50-100bp。
* **無接頭**：某些情況例如定長擴增子且讀長不超片段長度，則Adapter Content圖應該完全平緩為0，FastQC**通過**。
* **接頭污染嚴重**：曾有案例文庫構建失敗，產生大量接頭二聚體。結果測序資料中絕大多數reads都是接頭序列本身。FastQC的Adapter Content一開始就是100%，很明顯Fail。這種情況下有效數據幾乎沒有，需要重新建庫。

**FastQC預設閾值設定原理與對質控的影響**

FastQC使用\*\*「紅-黃-綠」**標誌體系標示各模組結果是否符合預期品質。這些預設閾值（Pass/Warn/Fail判定標準）源自開發者對高通量資料常見品質特徵的經驗設定，主要針對**高品質全基因組隨機文庫\*\*調校​[rtsf.natsci.msu.edu](https://rtsf.natsci.msu.edu/genomics/technical-documents/fastqc-tutorial-and-faq.aspx#:~:text=sequence%20data,and%20consider%20what%20that%20results)。原則上，FastQC選取閾值時考慮了測序錯誤率和生物學合理性，例如：Phred品質低於Q20（錯誤率>1%）通常被視為不合格區，於是將每碱基品質模組Fail閾值設在中位數Q20、下四分位Q5​ [github.com](https://github.com/s-andrews/FastQC/blob/master/Configuration/limits.txt#:~:text=quality_base_lower%20warn%2010)；又如單一序列佔比超1%極不尋常，於是Overrepresented序列Fail閾值定為1%​ [github.com](https://github.com/s-andrews/FastQC/blob/master/Configuration/limits.txt#:~:text=overrepresented%20warn%200)。下表彙總了各主要模組的FastQC默認警告/錯誤閾值及其意義​[github.com](https://github.com/s-andrews/FastQC/blob/master/Configuration/limits.txt#:~:text=quality_base_lower%20warn%2010); [github.com](https://github.com/s-andrews/FastQC/blob/master/Configuration/limits.txt#:~:text=The%20per%20sequence%20quality%20module,the%20phred%20score%20which%20is)：

* **Per base sequence quality**：中位數 < Q25 或 下四分位 < Q10 警告；中位數 < Q20 或 下四分位 < Q5 錯誤​[github.com](https://github.com/s-andrews/FastQC/blob/master/Configuration/limits.txt#:~:text=quality_base_lower%20warn%2010)。（基於Phred分數對應的錯誤率，Q201%錯誤作為最低容許，Q250.3%為警示線。此確保整體測序誤差率維持在較低水準。）
* **Per sequence quality scores**：主要峰值 < Q27 警告；< Q20 錯誤​[github.com](https://github.com/s-andrews/FastQC/blob/master/Configuration/limits.txt#:~:text=most%20frequently%20observed)。（Q27約0.2%錯誤率，即若最常見的平均讀序質量不到Q27，表示讀序整體質量偏低。）
* **Per base sequence content**：任意位置A與T或G與C差異 >10% 警告；>20% 錯誤​[github.com](https://github.com/s-andrews/FastQC/blob/master/Configuration/limits.txt#:~:text=The%20per%20base%20sequence%20content,module%20tests%20the%20maximum%20deviation)。（假定隨機序列應各碱基均勻，10%偏差已非常顯著，20%極端。）
* **Per sequence GC content**：累積偏離>15%讀數 警告；>30% 錯誤​[github.com](https://github.com/s-andrews/FastQC/blob/master/Configuration/limits.txt#:~:text=The%20per%20sequence%20GC%20content,tests%20the%20maximum%20deviation%20between)。（偏離比例指reads GC分佈相對理論分佈的總體差異佔比。）
* **Per base N content**：任意位置N >5% 警告；>20% 錯誤​[github.com](https://github.com/s-andrews/FastQC/blob/master/Configuration/limits.txt#:~:text=For%20the%20N%20module%20the,on%20the%20percentage%20of%20Ns)。
* **Sequence length distribution**：存在不同長度 警告；存在0長度 錯誤​[github.com](https://github.com/s-andrews/FastQC/blob/master/Configuration/limits.txt#:~:text=The%20sequence%20length%20module%20tests,binary%2C%20so%20the%20values%20here)。
* **Sequence duplication levels**：非唯一reads >20% 警告；>50% 錯誤​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=high%20level%20of%20duplication%20is,%5BImage%2050%3A%20Figure%2051)。（或等價唯一序列少於80%/50%。）
* **Overrepresented sequences**：任一序列 >0.1% 警告；>1% 錯誤​[github.com](https://github.com/s-andrews/FastQC/blob/master/Configuration/limits.txt#:~:text=overrepresented%20warn%200)
* **Adapter content**：任意位置>5% reads含接頭 警告；>10% 錯誤​[github.com](https://github.com/s-andrews/FastQC/blob/master/Configuration/limits.txt#:~:text=The%20adapter%20module%27s%20warnings%20and,errors%20are%20based%20on%20the)。
* **Per tile quality**：任意tile平均質量低於全局平均>~5（Phred） 警告；>~10 錯誤​[github.com](https://github.com/s-andrews/FastQC/blob/master/Configuration/limits.txt#:~:text=The%20per%20tile%20module%20tests,maximum%20phred%20score%20loss%20between); [bioinformatics.babraham.ac.uk](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/12%20Per%20Tile%20Sequence%20Quality.html#:~:text=This%20module%20will%20issue%20a,that%20base%20across%20all%20tiles)。

這些預設閾值的**設定原理**在於快速標記明顯偏離常態的情況，以提示用戶潛在問題。例如品質值模組以Q20作底限是因為1%錯誤率通常被視為可接受上限，再低則錯誤過多難以容忍​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=8,Image%2045%3A%20Figure%2046)。GC分佈模組允許15%以內的累積偏差，是給予生物多樣性一定浮動範圍，但30%則表示重大偏差，極可能來源混雜​[github.com](https://github.com/s-andrews/FastQC/blob/master/Configuration/limits.txt#:~:text=The%20per%20sequence%20GC%20content,tests%20the%20maximum%20deviation%20between)。Base content模組10%/20%的閾值則是基於隨機序列不應出現那麼大的AT/GC差，除非有系統性偏好。這些閾值多是經驗和實踐反覆調整的結果，以在靈敏檢出異常與避免誤報之間取得平衡。

**對質控的影響**：FastQC的紅/黃標誌並不是絕對判決，而是建議用戶“此處值得留意”​[rtsf.natsci.msu.edu](https://rtsf.natsci.msu.edu/genomics/technical-documents/fastqc-tutorial-and-faq.aspx#:~:text=shotgun%20DNA%20sequencing,of%20sequencing%20that%20was%20run)。當模組標記Warn/Fail時，應促使研究者檢視該指標異常的原因：

* 某些情況下，**Fail確實表示需要干預**。例如Adapter Content fail幾乎明確告訴我們需要做接頭剪除​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=15,%5BImage%2052%3A%20Figure%2053)；Per base quality fail（中位<20）表示讀序末端品質太差，建議剪掉或丟棄那些reads。
* 有些Warn可能**問題不大但提醒注意**。如Per base quality黃標（中位跌破Q25但仍高於Q20）時，多半尾端品質下降，解決方案是視下游需求決定是否修剪橙色區段。
* 另外一些Fail/Warn則**需結合背景判斷**：例如RNA-Seq的Per base sequence content fail是預料中的偏差​[rtsf.natsci.msu.edu](https://rtsf.natsci.msu.edu/genomics/technical-documents/fastqc-tutorial-and-faq.aspx#:~:text=sequence%20data,and%20consider%20what%20that%20results)；ChIP-Seq的duplication fail（50%以上重複）也可能正常​[biostar.galaxyproject.org](https://biostar.galaxyproject.org/p/27279/index.html#:~:text=%2A%20ChIP,specific%20reads)。FastQC本身無法了解實驗設計，因此給出一般性標記，科學家須根據樣本特性決定是否屬假警報。正如官方教程所言：“Warn/Fail並不一定表示測序跑道真正失敗，而是提醒研究者針對該結果，結合樣本和測序型別斟酌其意義”​[rtsf.natsci.msu.edu](https://rtsf.natsci.msu.edu/genomics/technical-documents/fastqc-tutorial-and-faq.aspx#:~:text=shotgun%20DNA%20sequencing,of%20sequencing%20that%20was%20run)。  
  因此FastQC在質控中的作用更多是**診斷提示**：紅黃燈提示我們該檢查那些指標，找到原因並採取相應措施或在後續分析中修正。

**閾值調整建議**：由於FastQC預設閾值並非適用所有情境，必要時可以進行調整：

* **根據測序類型調整**：對於已知會偏離預期的實驗，使用者可以**忽略或放寬**某些模組的評價。例如mRNA-Seq幾乎總是fail序列成分，這屬正常，可選擇忽略該模組的Fail或在心中調高容忍閾值（如允許前10bp差異達15-20%）​[sequencing.qcfail.com](https://sequencing.qcfail.com/articles/positional-sequence-bias-in-random-primed-libraries/#:~:text=)。同理，小RNA或Amplicon實驗中，重複率閾值應放寬，否則所有樣本都報Fail。
* **自訂FastQC設定**：FastQC允許修改配置檔limits.txt來改變閾值​[github.com](https://github.com/s-andrews/FastQC/blob/master/Configuration/limits.txt#:~:text=The%20per%20base%20quality%20filter,values%2C%20one%20for%20the%20value); [github.com](https://github.com/s-andrews/FastQC/blob/master/Configuration/limits.txt#:~:text=quality_base_median%20warn%2025)。高級用戶可調整其中數值（例如將sequence成分的warn從10%改20%）以符合特殊資料需求。然而，一般質控分析更常見做法是保留預設但**理解**何時不必驚慌。
* **使用其他工具**：某些指標FastQC方法學有限，例如重複率在單端層面估計不準確​[dnatech.ucdavis.edu](https://dnatech.ucdavis.edu/faqs/why-does-fastqc-show-unexpectedly-high-sequence-duplication-levels-pcr-duplicates#:~:text=data.%20All%20DNA,the%20ten%20most%20highly%20expressed)。可考慮使用如**Fastp**、**MultiQC**或**HTStream**等更現代工具，這些工具有時提供自動化判斷或針對配對讀更精確的計算​[dnatech.ucdavis.edu](https://dnatech.ucdavis.edu/faqs/why-does-fastqc-show-unexpectedly-high-sequence-duplication-levels-pcr-duplicates#:~:text=consequence%20FASTQC%20tends%20to%20generate,enzymatic%20reactions%2C%20all%20have%20some)。如果FastQC總報一些預期內的Fail讓使用者困惑，升級工具可能比調閾值更有益。
* **報告上下文解釋**：在質控報告中，建議對每個Warn/Fail進行註解說明其可能的原因及是否需行動。這實際上是**人工調整閾值的思維過程**：例如在RNA-Seq QC報告中註明「序列成分偏差由隨機引物造成，可忽略」，相當於在該情境下自訂了一個寬鬆標準。

總之，FastQC的預設閾值為一般樣本提供了有科學根據的參考範圍​[journals.plos.org](https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0304158#:~:text=Warning%3A%20a%20warning%20is%20issued,base%20is%20less%20than%2020)。但在實踐中，我們應避免機械地依賴Pass/Fail，建議將其作為**質控輔助手段**，結合理解實驗特性來調整判斷標準。必要時，可透過更改設定或二次檢視數據（如比對後再評估duplicates）來得到更適合的評價。

**不同類型測序數據的質控案例分析**

各類測序實驗由於樣本性質和文庫構建方式不同，FastQC指標的表現也有所差異。以下結合RNA-seq、DNA-seq、ChIP-seq等數據，說明如何解讀FastQC結果並進行針對性質控。

**RNA-Seq數據**

**特點**：轉錄組測序通常以隨機六聚體引物逆轉錄，文庫可能存在5’端序列偏差和高動態範圍的片段覆蓋。FastQC在RNA-Seq資料上常見的flag有：

* *Per base sequence content*：**幾乎必定Fail**。原因是前10–12個碱基有明顯不均勻的A/T/G/C比例，這是隨機引物偏差所致​[sequencing.qcfail.com](https://sequencing.qcfail.com/articles/positional-sequence-bias-in-random-primed-libraries/#:~:text=)。例如常看到A和T高於25%、G和C偏低的圖形。這並非測序錯誤，而是已知的文庫現象​[sequencing.qcfail.com](https://sequencing.qcfail.com/articles/positional-sequence-bias-in-random-primed-libraries/#:~:text=The%20cause%20of%20this%20bias,primed%20by%20the%20random%20primers)。研究顯示此偏差對基因定量影響不大​[sequencing.qcfail.com](https://sequencing.qcfail.com/articles/positional-sequence-bias-in-random-primed-libraries/#:~:text=In%20practice%20it%20seems%20that,mappability%20and%20other%20factors%20beyond)。質控建議：此Fail可接受，不需要特別處理（無需剪除序列開頭），但務必確認偏差僅限序列起點且隨後趨於隨機。如果整段都有明顯偏差則另當別論。
* *Sequence Duplication Levels*：**通常Warn/Fail**。RNA-Seq中因基因表達差異，部分轉錄本產生極大量reads，FastQC報告重複率可高達50-70%​ [dnatech.ucdavis.edu](https://dnatech.ucdavis.edu/faqs/why-does-fastqc-show-unexpectedly-high-sequence-duplication-levels-pcr-duplicates#:~:text=assumed%20from%20read%20coverage%20data,In%20addition)。需要理解這多是生物學重複（同一基因不同片段可能序列重複）而非PCR重複。質控建議：不在FASTQ階段盲目去重，以免丟失真實生物信號。可在比對後再用工具標記PCR duplicates。對於FastQC的Fail，可通過MultiQC等將多個樣本比較，確認皆為類似模式則問題不大。也可採用fastp等工具評估實際duplicate情況​[dnatech.ucdavis.edu](https://dnatech.ucdavis.edu/faqs/why-does-fastqc-show-unexpectedly-high-sequence-duplication-levels-pcr-duplicates#:~:text=data.%20All%20DNA,the%20ten%20most%20highly%20expressed)。
* *Overrepresented sequences*：RNA-Seq若rRNA去除不充分，常見過度序列是rRNA片段。亦或是接頭序列（若未剪乾淨）。質控建議：檢視FastQC列出的序列，如果是接頭則做**剪除**，如果是rRNA可在比對時排除或重做文庫去除rRNA。
* *Adapter content*：轉錄組片段長度通常大於讀長，不應有大量接頭。若Adapter Content出現上升，表明樣本片段較短（可能為小RNA碎片或者degrade碎片），需要剪掉接頭部分再分析。一般mRNA-Seq若文庫構建正常，此圖應Pass或輕微Warn。
* 其他指標如*Per base quality*通常表現正常趨勢（高->低），*GC content*接近物種mRNA平均GC，一些偏差可由高表達基因構成差異造成但通常不嚴重。

**實例**：某人類mRNA-Seq（2×150bp）資料，FastQC結果：前10bp序列成分Fail（A約30%, T20%, G/C各25%），序列品質末端下降到Q25左右但均在可接受範圍，GC含量分佈與理論值吻合（峰在~48%），重複序列Fail（剩餘只有50%唯一）。過度序列列出了一條Illumina接頭序列（佔0.5%）和幾條來源於18S rRNA的序列。對策：接頭序列雖只有0.5%但還是剪除以免干擾，比對時也會排除對rRNA的比對或者在定量時忽略這些reads。序列成分和重複的Fail理解為RNA文庫正常現象而非錯誤，因此不特別處理。經處理剪除接頭後重新FastQC，Adapter Content通過，其餘指標趨於一致。下游分析（差異表達）沒有因這些FastQC的Fail項受到負面影響。這說明對RNA-Seq，FastQC更多是提示我們注意文庫特性而不是嚴重錯誤。

**DNA WGS（全基因組重測序）**

**特點**：隨機片段化的基因組DNA應該在各指標上接近理想狀態，也是FastQC閾值設定的主要依據​[rtsf.natsci.msu.edu](https://rtsf.natsci.msu.edu/genomics/technical-documents/fastqc-tutorial-and-faq.aspx#:~:text=sequence%20data,and%20consider%20what%20that%20results)。對WGS數據的FastQC解讀：

* *Per base sequence quality*：WGS通常讀長固定、品質較高，圖表現為起始稍低隨即高企，末端略降但多半在Q20以上。FastQC一般**全綠通過**。若看到異常（如中段突然掉至低品質），懷疑測序過程有問題。需檢查是否儀器有警報或該區段reads應剪除。
* *Per base sequence content*：隨機基因組reads應四線平行。如果FastQC在WGS上報告警告，可能意味著污染。例如曾有WGS資料顯示末端T含量激增，後證實那幾個碱基其實是接頭遺留片段高T序列導致。故WGS若此模組Warn/Fail，大多需要尋找技術原因（接頭沒剪或GC偏性的污染DNA）。
* *GC content*：人類或一般物種基因組GC峰單一。若雙峰或明顯偏移，要高度懷疑混入不同物種DNA。WGS的GC Fail通常就是樣本不純所致。對此要麼重新準備樣本，要麼在分析時分離reads來源（例如metagenomics approach）。
* *Duplication*：理想WGS庫PCR循環少，覆蓋深度適中，重複率很低（剩餘>80%唯一）。FastQC幾乎都Pass。如果重複Warn甚至Fail，可能因為**過深測序**：e.g.小基因組高深度，或**過擴增**：樣本DNA量低PCR cycle過多。對過深測序，可接受Fail但要在變異分析時標記duplicates避免統計偏差​[sequencing.qcfail.com](https://sequencing.qcfail.com/articles/libraries-can-contain-technical-duplication/#:~:text=from%20sequences%20which%20occur%20different,library%20were%20to%20be%20deduplicated)。對過擴增，建議下次文庫優化。
* *Overrepresented sequences & Adapter*：WGS文庫理應無明顯特定序列高豐度。若報警，多是接頭序列（需剪除）或其他試劑污染（如PhiX spike-in序列如果未濾掉，在文庫中也可達到閾值）。Adapter Content如出現上升，更需處理剪接頭。

**實例**：某細菌基因組重測序，高覆蓋(100×)。FastQC顯示Quality圖完美（全高品質），Base content平坦，GC含量圖有輕微肩峰：主峰在51%，但在30-35%處有一個次峰佔總reads約10%，導致GC模組Warn。過度序列列出了多條佔比0.2-0.5%的reads，經BLAST發現對應於宿主雜質DNA片段。此說明樣本提取混入了宿主DNA，約佔一成reads。處理方法：用**分類軟體**將宿主reads過濾後再進行組裝。重複序列Fail（唯一序列只有20%）：這是預期的，因100×深度對3Mbp基因組已極大過採，每個片段讀到多次​[sequencing.qcfail.com](https://sequencing.qcfail.com/articles/libraries-can-contain-technical-duplication/#:~:text=requires%20sufficient%20read%20depth%20to,an%20indication%20of%20technical%20duplication)。因此重複Fail並不影響我們對數據質量的評估，組裝時會自動利用覆蓋信息。這案例體現：WGS的FastQC Fail多半是真有需要處理的問題（如污染），除非能解釋為實驗設計（如深度高導致重複高）。

**ChIP-Seq數據**

**特點**：ChIP-Seq富集基因組中特定區段（如轉錄因子結合位點），不是對全基因組的均勻抽樣。因此讀序品質通常良好，但序列分佈、覆蓋均勻性與WGS不同。FastQC在ChIP-Seq上可能引發的一些注意：

* *Sequence duplication levels*：**往往警告甚至失敗**。因為特定peak區域的片段被擴增測序多次，reads重疊度高。FastQC報告唯一序列比例常在50%左右甚至更低​[biostar.galaxyproject.org](https://biostar.galaxyproject.org/p/27279/index.html#:~:text=%2A%20ChIP,specific%20reads)。這不全是PCR artifact，而是真實富集使然。質控策略：檢查peak區重複是否自然產生，通常對ChIP-Seq，分析時會允許一定程度重複，但若全域重複極高（>80%）仍是文庫不良的信號。對於Fail，可以在Mapping後再過濾**多重比對**reads、合併重複計算，以得到實際有效讀序數。
* *Per base sequence content*：ChIP樣本若抗體無偏好，應接近隨機。但是某些結合位點序列可能存在偏好（例如富含AT的motif），這可能讓pull-down的所有片段在中間某區域呈現偏AT或GC。然FastQC的此模組看的是read各位置同位比較，除非片段長度很一致導致reads某固定位置對齊motif，否則一般不會在未比對的reads中顯示特異圖案。因此ChIP-Seq通常該模組Pass，除非文庫構建有別的問題（如前述接頭或隨機偏差）。
* *GC content*：可能稍有偏移，但通常不至於Fail。舉例，如果目標結合序列GC低於全基因組平均，reads整體GC分佈可能略向低移，但峰應該相對接近基因組均值。Fail只會在極端情況如pull-down片段全部局限於極AT富集區域。多數轉錄因子結合位點散布基因組，不會造成劇烈全局GC偏差。
* *Overrepresented sequences*：ChIP的過度序列可能有兩類：一類是接頭/引物（技術原因，同WGS）；另一類是**非常強的peak**。如果某轉錄因子高度偏好一個特定序列且覆蓋狹窄，可能很多reads幾乎相同，該序列就可能上榜。然而一般來說，不同片段序列還是會有變化（即使都在同一區，也可能不同位置或不同DNA鏈），因此完全相同的read不至於佔比那麼高，除非peak區極短和測序超深。
* 其他如*Quality分佈*多半正常，*N含量*正常。Sequence length有時ChIP會剪裁掉接頭導致變長度，也不影響分析。

**實例**：以H3K4me3的ChIP-Seq為例，這種組蛋白修飾富集在轉錄起始位附近，形成相對狹窄而強的peaks。FastQC結果可能顯示約50%重複reads（Warn）​[biostar.galaxyproject.org](https://biostar.galaxyproject.org/p/27279/index.html#:~:text=%2A%20ChIP,specific%20reads)。這是因為每個peak區域數百bp內堆積了成千上萬reads，必然有許多序列重覆。下游處理通常在peak calling時容忍這點，或在對比輸入時消除背景。若重複率更高如80%，可能樣本褪跨幾個超級巨大的peak，暗示pull-down特異性差（可能寡頭效應，需要降低輸入細胞數或減少PCR擴增）。此外，此ChIP資料的GC分佈與基因組幾乎一致，序列成分圖平坦，說明抗體沒有造成偏序列捕獲，FastQC其他模組都Pass。過度序列列出一條Illumina接頭序列（0.2%），提示仍需剪除接頭。整體而言ChIP-Seq的FastQC需要特別關注duplication，這關係到文庫複雜度和peak可信度。

**其他數據類型簡述**

* **小RNA-Seq**：讀長很短（~20-30nt），通常已將接頭剪掉。Base quality一般高。主要挑戰在於**序列成分**和**重複**：小RNA（例如miRNA）有強保守序列特徵，特定位點鹼基偏差明顯，FastQC序列成分模組一定Fail，但這是生物學訊號。重複率極高，因為某些miRNA可佔到數十%的reads，導致duplication Fail和大量過度序列（每條高豐度miRNA序列都上榜）。這些情況屬正常，質控時重點是確保接頭全剪除且沒有異常N或品質問題。
* **外顯子捕獲/目標序列**：由於只捕獲基因組特定片段，可能產生**中等重複率**（部分區域覆蓋深）。GC分佈可能偏離全基因組均值（因外顯子往往GC較高）。FastQC可能Warn duplication和GC。但只要偏差在預期範圍（如外顯子GC一般高於基因組5-10%），問題不大。需注意接頭和品質，如前述一般處理即可。
* **Amplicon序列**：極端情況，所有reads基本來自相同的PCR產物，FastQC會在序列成分、重複、過度序列全部Fail。然而這在情理之中——因為資料並非隨機文庫，而是一兩種序列反覆讀取。質控時主要檢查品質值是否OK、接頭是否清除即可。其他模組Fail可忽略。

**支持文獻與官方文件**

FastQC的各項指標設計與判讀基於扎實的經驗與業界共識，已有多份官方文件和學術資料支撐：

* **FastQC官方手冊與説明**：Babraham Bioinformatics提供了FastQC各模組的詳細説明​[mugenomicscore.missouri.edu](https://mugenomicscore.missouri.edu/PDF/FastQC_Manual.pdf#:~:text=%EF%82%B7%20The%20central%20red%20line,red); [bioinformatics.babraham.ac.uk](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/12%20Per%20Tile%20Sequence%20Quality.html#:~:text=The%20plot%20shows%20the%20deviation,should%20be%20blue%20all%20over)。其中對圖表的繪製方式和閾值均有闡述，例如品質箱線圖如何解讀​[mugenomicscore.missouri.edu](https://mugenomicscore.missouri.edu/PDF/FastQC_Manual.pdf#:~:text=%EF%82%B7%20The%20central%20red%20line,red)、每tile品質熱圖的成因​[bioinformatics.babraham.ac.uk](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/12%20Per%20Tile%20Sequence%20Quality.html#:~:text=Reasons%20for%20seeing%20warnings%20or,debris%20inside%20the%20flowcell%20lane)等。配置檔limits.txt給出了預設Warn/Fail閾值​[github.com](https://github.com/s-andrews/FastQC/blob/master/Configuration/limits.txt#:~:text=quality_base_lower%20warn%2010); [github.com](https://github.com/s-andrews/FastQC/blob/master/Configuration/limits.txt#:~:text=The%20per%20sequence%20quality%20module,the%20phred%20score%20which%20is)。這些官方資料是理解FastQC指標計算方法和預設門檻的首要依據。
* **學術論文和報告**：雖然FastQC本身未正式發表論文，但許多文獻提及其應用並討論了指標的合理性。例如Bullard等人（2010）探討了RNA-Seq數據的偏差，指出隨機hexamer引發的前端序列偏好​[sequencing.qcfail.com](https://sequencing.qcfail.com/articles/positional-sequence-bias-in-random-primed-libraries/#:~:text=The%20cause%20of%20this%20bias,primed%20by%20the%20random%20primers)；Assoum et al.的研究引用了FastQC品質閾值（Q20, Q25）作為數據篩選標準​[pmc.ncbi.nlm.nih.gov](https://pmc.ncbi.nlm.nih.gov/articles/PMC4230301/#:~:text=Example%20plot%20of%20one%20QC,combined%20into%20an%20html%20document)。還有一些教程文章深入解析了FastQC報告，例如Haynes與Kelly的RNA-Seq教程強調了尾端品質下降和隨機引物偏差是正常的​[rtsf.natsci.msu.edu](https://rtsf.natsci.msu.edu/genomics/technical-documents/fastqc-tutorial-and-faq.aspx#:~:text=shotgun%20DNA%20sequencing,of%20sequencing%20that%20was%20run)。
* **社群經驗分享**：生物資訊論壇和博客提供了大量案例分析和解讀建議。例如SeqAnswers/BioStars上討論了FastQC duplication在RNA-Seq的誤報現象​[dnatech.ucdavis.edu](https://dnatech.ucdavis.edu/faqs/why-does-fastqc-show-unexpectedly-high-sequence-duplication-levels-pcr-duplicates#:~:text=assumed%20from%20read%20coverage%20data,nt%20of%20the%20first%20100%2C000)；QCFail專案網站匯總了測序中常見偏差，包括隨機引物造成的序列成分問題​[sequencing.qcfail.com](https://sequencing.qcfail.com/articles/positional-sequence-bias-in-random-primed-libraries/#:~:text=You%20can%20clearly%20see%20the,a%20greater%20or%20lesser%20extent)、技術重複的診斷​[sequencing.qcfail.com](https://sequencing.qcfail.com/articles/libraries-can-contain-technical-duplication/#:~:text=The%20observation%20of%20duplication%20is,of%20duplication%20plots%20is%20very)等，這些文章證實了FastQC指標背後的科學道理，並提醒何時屬正常情況。
* **工具比較**：近年一些現代QC工具（如Fastp）發表時，也討論了FastQC的局限，例如FastQC對雙端重複的低估​[dnatech.ucdavis.edu](https://dnatech.ucdavis.edu/faqs/why-does-fastqc-show-unexpectedly-high-sequence-duplication-levels-pcr-duplicates#:~:text=data.%20All%20DNA,the%20ten%20most%20highly%20expressed)。這些討論間接支持FastQC相應指標需要謹慎解讀，並非FastQC錯誤而是分析方法使然。

綜上所述，大量證據支持了FastQC各指標的設計初衷和科學基礎。例如，Phred品質作為測序準確度衡量已成行業標準​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=8,Image%2045%3A%20Figure%2046)；GC含量分佈異常提示污染這一點在實踐中屢有應驗​[bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=10,%5BImage%2047%3A%20Figure%2048)；RNA-Seq起始偏差問題在文獻中被充分記錄且證明對差異分析影響不大​[sequencing.qcfail.com](https://sequencing.qcfail.com/articles/positional-sequence-bias-in-random-primed-libraries/#:~:text=In%20practice%20it%20seems%20that,mappability%20and%20other%20factors%20beyond)。官方文件和社群經驗都強調：FastQC旗標應視為質控警示而非硬性判決，需融合對實驗的理解進行判讀​[rtsf.natsci.msu.edu](https://rtsf.natsci.msu.edu/genomics/technical-documents/fastqc-tutorial-and-faq.aspx#:~:text=shotgun%20DNA%20sequencing,of%20sequencing%20that%20was%20run)。適當調整閾值和參考多種資料來源，可以更好地運用FastQC進行高通量測序數據的質量控制，保證後續生物學分析的可靠性。

**參考文獻：**

1. Andrews, S. *FastQC: A Quality Control tool for High Throughput Sequence Data*. Babraham Bioinformatics (2010). [rtsf.natsci.msu.edu](https://rtsf.natsci.msu.edu/genomics/technical-documents/fastqc-tutorial-and-faq.aspx#:~:text=section%20for%20each%20FastQC%20module,that%20has%20a%20%E2%80%9CWarn%E2%80%9D%20or); [rtsf.natsci.msu.edu](https://rtsf.natsci.msu.edu/genomics/technical-documents/fastqc-tutorial-and-faq.aspx#:~:text=shotgun%20DNA%20sequencing,of%20sequencing%20that%20was%20run)
2. Positional sequence bias in random primed libraries – *QCFail Sequencing* (2013). [sequencing.qcfail.com](https://sequencing.qcfail.com/articles/positional-sequence-bias-in-random-primed-libraries/#:~:text=You%20can%20clearly%20see%20the,a%20greater%20or%20lesser%20extent); [sequencing.qcfail.com](https://sequencing.qcfail.com/articles/positional-sequence-bias-in-random-primed-libraries/#:~:text=In%20practice%20it%20seems%20that,mappability%20and%20other%20factors%20beyond)
3. FastQC Manual, Babraham Bioinformatics (v0.11.9). [mugenomicscore.missouri.edu](https://mugenomicscore.missouri.edu/PDF/FastQC_Manual.pdf#:~:text=%EF%82%B7%20The%20central%20red%20line,red); [bv-brc.org](https://www.bv-brc.org/docs/tutorial/fastq_utilities/fastq_utilities.html#:~:text=7,not%20uniform%2C%20it%20starts%20out)
4. Babraham Bioinformatics – FastQC Help Documentation. [bioinformatics.babraham.ac.uk](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/12%20Per%20Tile%20Sequence%20Quality.html#:~:text=The%20plot%20shows%20the%20deviation,should%20be%20blue%20all%20over); [bioinformatics.babraham.ac.uk](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/12%20Per%20Tile%20Sequence%20Quality.html#:~:text=Reasons%20for%20seeing%20warnings%20or,debris%20inside%20the%20flowcell%20lane)
5. *FastQC default limits configuration* (GitHub repository). [github.com](https://github.com/s-andrews/FastQC/blob/master/Configuration/limits.txt#:~:text=quality_base_lower%20warn%2010); [github.com](https://github.com/s-andrews/FastQC/blob/master/Configuration/limits.txt#:~:text=The%20per%20sequence%20GC%20content,tests%20the%20maximum%20deviation%20between)
6. Haynes, B. & Kelly, J.R. *Assessing FASTQC results*. Harvard FAS Informatics (2017). [rtsf.natsci.msu.edu](https://rtsf.natsci.msu.edu/genomics/technical-documents/fastqc-tutorial-and-faq.aspx#:~:text=section%20for%20each%20FastQC%20module,that%20has%20a%20%E2%80%9CWarn%E2%80%9D%20or); [rtsf.natsci.msu.edu](https://rtsf.natsci.msu.edu/genomics/technical-documents/fastqc-tutorial-and-faq.aspx#:~:text=shotgun%20DNA%20sequencing,of%20sequencing%20that%20was%20run)
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). [dnatech.ucdavis.edu](https://dnatech.ucdavis.edu/faqs/why-does-fastqc-show-unexpectedly-high-sequence-duplication-levels-pcr-duplicates#:~:text=assumed%20from%20read%20coverage%20data,nt%20of%20the%20first%20100%2C000); [dnatech.ucdavis.edu](https://dnatech.ucdavis.edu/faqs/why-does-fastqc-show-unexpectedly-high-sequence-duplication-levels-pcr-duplicates#:~:text=it%20is%20not%20uncommon%20that,babrah)
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).​