**Bulk RNA-seq FastQC 模組調整**

## Ref(Share): https://chatgpt.com/share/680499ed-1b60-8002-abf0-0dbf20ae2f7b

## Ref: https://chatgpt.com/g/g-p-67b46db7124c8191900f83011a09eeaa-charlene-fastqc/c/680488c2-2944-8002-bd30-0a3a75c74d3c

\*\*Illumina 平臺 Homo sapiens Bulk RNA-seq FastQC 模組調整建議：\*\*以下針對各 FastQC 模組提供建議的啟用狀態（ignore 0=啟用，1=忽略）及 warn/error 閾值調整，並說明理由與文獻依據：

| **模組名稱** | **建議設定<br>(ignore, warn, error)** | **調整理由** | **參考資料** |
| --- | --- | --- | --- |
| **Sequence Duplication Levels** (duplication) | ignore 0（啟用） ⚠️ warn：30％ ❌ error：20％ | Bulk RNA-seq 常出現高序列重複率，因為少數高豐度轉錄本會產生大量重複讀序​[dnatech.ucdavis.edu](https://dnatech.ucdavis.edu/faqs/why-does-fastqc-show-unexpectedly-high-sequence-duplication-levels-pcr-duplicates#:~:text=data%20which%20are%20often%20dominated,nt%20of%20the%20first%20100%2C000)​[bioinformatics-core-shared-training.github.io](https://bioinformatics-core-shared-training.github.io/RNAseq_November_2020_remote/html/Analysis_of_bulk_RNA-seq_data_Main_Reference_Document.pdf#:~:text=Natalie%20van%20Dis,have%20gone%20wrong%20with%20this)。FastQC 原預設在僅剩 <50% 唯一序列時即標記失敗，對 RNA-seq 過嚴​[rtsf.natsci.msu.edu](https://rtsf.natsci.msu.edu/genomics/technical-documents/fastqc-tutorial-and-faq.aspx#:~:text=When%20sequencing%20RNA%20there%20will,is%20expected%20in%20this%20case)。實務上單端 RNA-seq 讀長資料 60–70% 重複屬正常範圍，可忽略其警告​[bioinformatics-core-shared-training.github.io](https://bioinformatics-core-shared-training.github.io/RNAseq_November_2020_remote/html/Analysis_of_bulk_RNA-seq_data_Main_Reference_Document.pdf#:~:text=Natalie%20van%20Dis,have%20gone%20wrong%20with%20this)。因此提高容忍度：僅當唯一序列低於約30%（重複率>70%）時給警告，低於20%（重複率>80%）時視為失敗，平衡檢出低庫存複雜度樣本的需求​[bioinformatics-core-shared-training.github.io](https://bioinformatics-core-shared-training.github.io/RNAseq_November_2020_remote/html/Analysis_of_bulk_RNA-seq_data_Main_Reference_Document.pdf#:~:text=Natalie%20van%20Dis,have%20gone%20wrong%20with%20this)​[pmc.ncbi.nlm.nih.gov](https://pmc.ncbi.nlm.nih.gov/articles/PMC11263697/#:~:text=match%20at%20L1645%20the%20number,i%29%20the)。 | [dnatech.ucdavis.edu](https://dnatech.ucdavis.edu/faqs/why-does-fastqc-show-unexpectedly-high-sequence-duplication-levels-pcr-duplicates#:~:text=data%20which%20are%20often%20dominated,nt%20of%20the%20first%20100%2C000)​[bioinformatics-core-shared-training.github.io](https://bioinformatics-core-shared-training.github.io/RNAseq_November_2020_remote/html/Analysis_of_bulk_RNA-seq_data_Main_Reference_Document.pdf#:~:text=Natalie%20van%20Dis,have%20gone%20wrong%20with%20this)​[rtsf.natsci.msu.edu](https://rtsf.natsci.msu.edu/genomics/technical-documents/fastqc-tutorial-and-faq.aspx#:~:text=When%20sequencing%20RNA%20there%20will,is%20expected%20in%20this%20case)​[pmc.ncbi.nlm.nih.gov](https://pmc.ncbi.nlm.nih.gov/articles/PMC11263697/#:~:text=match%20at%20L1645%20the%20number,i%29%20the) |
| **K-mer Content** (kmer) | ignore 1（停用） *（預設停用）* | RNA-seq 文庫的隨機引物和高豐度轉錄本會導致特定 k-mer 富集，但多為正常現象而非污染​[rtsf.natsci.msu.edu](https://rtsf.natsci.msu.edu/genomics/technical-documents/fastqc-tutorial-and-faq.aspx#:~:text=the%20six%20most%20biased%20kmers,seq)。FastQC 的 k-mer 模組難以解讀，經常將來源於真實高表達基因的 k-mer 偏差標記為警訊​[rtsf.natsci.msu.edu](https://rtsf.natsci.msu.edu/genomics/technical-documents/fastqc-tutorial-and-faq.aspx#:~:text=the%20six%20most%20biased%20kmers,seq)。其他模組（如 Overrepresented 序列、Adapter Content）已足夠檢測實際污染序列，因此建議維持停用以避免誤警訊。必要時可另行啟用並提高統計閾值以減少假警報。 | [rtsf.natsci.msu.edu](https://rtsf.natsci.msu.edu/genomics/technical-documents/fastqc-tutorial-and-faq.aspx#:~:text=the%20six%20most%20biased%20kmers,seq) |
| **Per base N content** (n\_content) | ignore 0（啟用） ⚠️ warn：5% ❌ error：20% | Bulk RNA-seq 中每個位置的 N 比例應接近 0%，Illumina 序列儀通常對每個循環都有明確鹼基呼叫，因此任何明顯的 N 峰值都表示測序過程出現問題​[rtsf.natsci.msu.edu](https://rtsf.natsci.msu.edu/genomics/technical-documents/fastqc-tutorial-and-faq.aspx#:~:text=Per%20base%20N%20content)。預設閾值 (任一位置 N 超過5%警告、20%錯誤) 已足以檢測異常情況，RNA-seq 無特殊原因調整此模組，維持原設定即可。 | [rtsf.natsci.msu.edu](https://rtsf.natsci.msu.edu/genomics/technical-documents/fastqc-tutorial-and-faq.aspx#:~:text=Per%20base%20N%20content) |
| **Overrepresented Sequences** (overrepresented) | ignore 0（啟用） ⚠️ 列出閾值：0.1% ❌ error：1% | Bulk RNA-seq 中常見的過度代表序列包括接頭鏈接序列或核糖體RNA片段等​[training.galaxyproject.org](https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html#:~:text=Ideally%20Illumina%20sequence%20data%20should,Nextera%20dapater%20has%20been%20detected)。FastQC 預設將佔比≥0.1%的序列列出，單一序列佔比≥1%標記錯誤。RNA-seq 樣本若有少量高豐度序列（例如多聚A尾巴或殘餘 rRNA），可能出現警告但屬正常​[dnatech.ucdavis.edu](https://dnatech.ucdavis.edu/faqs/why-does-fastqc-show-unexpectedly-high-sequence-duplication-levels-pcr-duplicates#:~:text=analyses%20because%20it%20only%20works,end%20sequencing)。建議保留此模組以發現明顯污染，例如當某序列佔讀長超過1%時需特別注意（可能未移除接頭或rRNA污染）。由於此閾值已能捕捉潛在問題且過嚴可能漏報，小幅調整意義不大，因此維持預設，但對於已知的RNA序列偏高情況可不視為失敗，而是結合生物背景判讀​[dnatech.ucdavis.edu](https://dnatech.ucdavis.edu/faqs/why-does-fastqc-show-unexpectedly-high-sequence-duplication-levels-pcr-duplicates#:~:text=analyses%20because%20it%20only%20works,end%20sequencing)。 | [dnatech.ucdavis.edu](https://dnatech.ucdavis.edu/faqs/why-does-fastqc-show-unexpectedly-high-sequence-duplication-levels-pcr-duplicates#:~:text=analyses%20because%20it%20only%20works,end%20sequencing)​[training.galaxyproject.org](https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html#:~:text=Ideally%20Illumina%20sequence%20data%20should,Nextera%20dapater%20has%20been%20detected) |
| **Per base sequence quality** (quality\_base) | ignore 0（啟用） ⚠️ Q1<10；中位數<25 ❌ Q1<5；中位數<20 | Illumina 平臺讀長的品質多在高品質區域，預設閾值對 RNA-seq 亦適用：中位數品質若低於Q20顯示嚴重劣化​[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)。Bulk RNA-seq 首幾個鹼基因隨機引物多樣性低，品質得分稍低屬常見，但通常不致低於上述閾值​[training.galaxyproject.org](https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html#:~:text=Figure%203%3A%20Adapter%20Content)（Illumina 通常在前5–7個鹼基後品質上升​[training.galaxyproject.org](https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html#:~:text=It%20is%20normal%20with%20all,The%20quality%20of%20reads%20on)）。因此不需特別調整此模組閾值；保持預設設定可確保偵測明顯的測序品質問題，同時對正常的些微品質下降不過度警示。 | [training.galaxyproject.org](https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html#:~:text=%28red%29) |
| **Per base sequence content** (sequence) | ignore 0（啟用） ⚠️ warn：20% ❌ error：40% | RNA-seq 常因隨機六聚體引物產生**序列偏好**：前10–12個鹼基的A/T/G/C比例明顯不均，但這是正常的技術偏差​[bioinformatics.babraham.ac.uk](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/4%20Per%20Base%20Sequence%20Content.html#:~:text=It%27s%20worth%20noting%20that%20some,or%20error%20in%20this%20module)​[bioinformatics.babraham.ac.uk](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/4%20Per%20Base%20Sequence%20Content.html#:~:text=of%20random%20hexamers%20or%20through,are%20inherently%20biased%20in%20their)。FastQC 預設閾值嚴格（任一位置A對T或G對C差異>20%即Fail），導致幾乎所有RNA-seq樣本此模組皆Fail​[bioinformatics.babraham.ac.uk](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/4%20Per%20Base%20Sequence%20Content.html#:~:text=of%20random%20hexamers%20or%20through,are%20inherently%20biased%20in%20their)。為避免正常偏差被誤判，我們將警告閾值放寬至20%、錯誤閾值放寬至40%，容許前端合理範圍內的序列成分偏差。同時若超出此範圍（例如出現異常序列構成偏差，可能是接頭污染或實驗問題），仍會觸發警示。​[bioinformatics.babraham.ac.uk](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/4%20Per%20Base%20Sequence%20Content.html#:~:text=of%20random%20hexamers%20or%20through,are%20inherently%20biased%20in%20their)​[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) | [bioinformatics.babraham.ac.uk](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/4%20Per%20Base%20Sequence%20Content.html#:~:text=It%27s%20worth%20noting%20that%20some,or%20error%20in%20this%20module)​[bioinformatics.babraham.ac.uk](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/4%20Per%20Base%20Sequence%20Content.html#:~:text=of%20random%20hexamers%20or%20through,are%20inherently%20biased%20in%20their)​[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) |
| **Per sequence GC content** (gc\_sequence) | ignore 0（啟用） ⚠️ warn：20% ❌ error：40% | Bulk RNA-seq 讀長的 GC 含量分布未必符合理論正態分布，因為不同轉錄本的 GC 含量可能有偏重，導致實測分布比理論值更寬或更窄​[rtsf.natsci.msu.edu](https://rtsf.natsci.msu.edu/genomics/technical-documents/fastqc-tutorial-and-faq.aspx#:~:text=What%20to%20look%20for%3A%20For,Seq%20data%20yet%20FastQC%20still)​[rtsf.natsci.msu.edu](https://rtsf.natsci.msu.edu/genomics/technical-documents/fastqc-tutorial-and-faq.aspx#:~:text=assignment%20can%20be%20ignored,was%20narrower%20than%20the%20theoretical)。高品質 RNA-seq 資料仍可能因此被 FastQC 給予 Warn​[rtsf.natsci.msu.edu](https://rtsf.natsci.msu.edu/genomics/technical-documents/fastqc-tutorial-and-faq.aspx#:~:text=assignment%20can%20be%20ignored,was%20narrower%20than%20the%20theoretical)。為了不對正常變異發出過多警告，建議放寬此模組允許的分布差異（如Warn提高至與理論分布差異20%，Error提高至40%）。如此僅當GC分布明顯異常（例如雙峰極端偏移，可能暗示樣本混雜或偏倚）時才標記問題​[rtsf.natsci.msu.edu](https://rtsf.natsci.msu.edu/genomics/technical-documents/fastqc-tutorial-and-faq.aspx#:~:text=distribution%20deviates%20too%20far%20from,was%20narrower%20than%20the%20theoretical)。 | [rtsf.natsci.msu.edu](https://rtsf.natsci.msu.edu/genomics/technical-documents/fastqc-tutorial-and-faq.aspx#:~:text=distribution%20deviates%20too%20far%20from,was%20narrower%20than%20the%20theoretical)​[rtsf.natsci.msu.edu](https://rtsf.natsci.msu.edu/genomics/technical-documents/fastqc-tutorial-and-faq.aspx#:~:text=assignment%20can%20be%20ignored,was%20narrower%20than%20the%20theoretical) |
| **Per sequence quality scores** (quality\_sequence) | ignore 0（啟用） ⚠️ 中心峰值<Q27 ❌ 中心峰值<Q20 | 每個讀序的平均品質分數分布主要取決於測序平台性能。Illumina Bulk RNA-seq 通常平均品質集中在高分區域（Q30左右），若此分布峰值低於Q27表示整體品質偏低​[github.com](https://github.com/s-andrews/FastQC/blob/master/Configuration/limits.txt#:~:text=The%20per%20sequence%20quality%20module,the%20phred%20score%20which%20is)​[github.com](https://github.com/s-andrews/FastQC/blob/master/Configuration/limits.txt#:~:text=quality_sequence%20warn%2027)；低於Q20則顯示資料品質嚴重不佳。由於RNA-seq與DNA-seq在此模組無顯著差異，維持預設Warn/Fail閾值即可。這確保當整體讀序品質明顯下降時會被檢出，但不會對正常範圍內的品質變動過敏感。 | [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 tile sequence quality** (tile) | ignore 1（停用） | 該模組檢測流動槽上局部區域(tile)的品質不均，一般僅在儀器問題（如氣泡、污漬）導致某些tile持續低品質時才有意義​[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)​[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)。現代Illumina流程若整體數據通過QC，少量tile品質偏差對下游分析影響很小​[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)。由於使用者對此問題通常無法在事後補救，且輕微的tile偏差可忽略​[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)，建議停用此模組以簡化報告。如有重大品質區域性問題，通常也會反映在每個鹼基品質模組中。 | [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)​[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) |
| **Sequence Length Distribution** (sequence\_length) | ignore 0（啟用） *(關閉Warn提示)* | Bulk RNA-seq 資料經品質/接頭修剪後，讀長長度通常不一致，FastQC 對**不同長度序列**預設即給警告​[bioinformatics.babraham.ac.uk](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/7%20Sequence%20Length%20Distribution.html#:~:text=Warning)。然而，長度變異在剪除低品質尾端或接頭後是正常現象​[bioinformatics.babraham.ac.uk](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/7%20Sequence%20Length%20Distribution.html#:~:text=Common%20reasons%20for%20warnings)。因此建議仍查看此模組的長度分布圖以了解讀長分佈情形，但忽略其自動警告（可在配置中關閉Warn提示）。只有在出現長度為0的讀序（理論上不應存在）時才視為錯誤​[bioinformatics.babraham.ac.uk](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/7%20Sequence%20Length%20Distribution.html#:~:text=This%20module%20will%20raise%20a,are%20not%20the%20same%20length)。此調整允許我們關注實質問題，同時不因正常的讀長變化誤判。 | [bioinformatics.babraham.ac.uk](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/7%20Sequence%20Length%20Distribution.html#:~:text=This%20module%20will%20raise%20a,are%20not%20the%20same%20length)​[bioinformatics.stackexchange.com](https://bioinformatics.stackexchange.com/questions/22601/issues-with-adapter-trimming-trim-galore#:~:text=ngs%20,Per%20base%20sequence%20content) |
| **Adapter Content** (adapter) | ignore 0（啟用） ⚠️ warn：5% ❌ error：10% | Bulk RNA-seq 文庫通常經過片段大小選擇以避免插入片段過短，因此理論上不應有明顯接頭序列殘留。然而，若讀長較長而部分插入片段偏短，讀到尾端可能出現接頭序列​[training.galaxyproject.org](https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html#:~:text=Ideally%20Illumina%20sequence%20data%20should,Nextera%20dapater%20has%20been%20detected)。FastQC 預設在≥5%讀序含接頭時警告，≥10%時錯誤。此標準適用於RNA-seq：少量接頭出現（<5%）屬於長讀長下的正常情況​[training.galaxyproject.org](https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html#:~:text=Ideally%20Illumina%20sequence%20data%20should,Nextera%20dapater%20has%20been%20detected)，超過閾值則提示需要進行接頭修剪處理。建議維持預設值並在報告中檢視各位置接頭含量圖，以確認是否需要進一步的資料清理​[training.galaxyproject.org](https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html#:~:text=Ideally%20Illumina%20sequence%20data%20should,Nextera%20dapater%20has%20been%20detected)​[training.galaxyproject.org](https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html#:~:text=,Image%3A%20Adapter%20Content)。 | [training.galaxyproject.org](https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html#:~:text=Ideally%20Illumina%20sequence%20data%20should,Nextera%20dapater%20has%20been%20detected)​[training.galaxyproject.org](https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/quality-control/tutorial.html#:~:text=,Image%3A%20Adapter%20Content) |

\*\*說明：\*\*以上調整平衡了 FastQC 自帶標準與 Bulk RNA-seq 實際特性，使正常的 RNA-seq 偏差不致產生誤判，同時確保真正異常的情況會被標示警示​[bioinformatics.babraham.ac.uk](https://www.bioinformatics.babraham.ac.uk/projects/fastqc/Help/3%20Analysis%20Modules/4%20Per%20Base%20Sequence%20Content.html#:~:text=of%20random%20hexamers%20or%20through,are%20inherently%20biased%20in%20their)​[dnatech.ucdavis.edu](https://dnatech.ucdavis.edu/faqs/why-does-fastqc-show-unexpectedly-high-sequence-duplication-levels-pcr-duplicates#:~:text=analyses%20because%20it%20only%20works,end%20sequencing)。這些建議參考了官方指南（如 ENCODE）、大型專案經驗及社群共識，以提高 QC 判讀對 Bulk RNA-seq 資料的適用性。