**Base Quality Score Recalibration（BQSR）について**

**1. 概要**

Base Quality Score Recalibration（BQSR）は、シーケンサーが出力したベースクオリティ（Base Quality; Q値）の系統的な誤差を補正し、実際の誤り率に整合させることを目的とした手法である。  
この処理は、変異検出（Variant Calling）の前段階で行われ、リードごとの信頼度を正確に反映させることで、SNP検出の感度を向上させる役割を担う。

**2. 背景：ベースクオリティとは**

シーケンサーが出力するFASTQファイルには、各塩基に対してPhredスケール（塩基の正しく読み取れている確率を対数スケールで表現するための指標）で表されたベースクオリティ（Q値）が付与されている。（ゲノム多様性解析p.15,16参照）

これは、呼び出された塩基が誤っている確率 を対数変換したものであり、

で定義される。

例えば、Q=20は誤り率1/100、Q=30は1/1000、Q=40は1/10000を意味する。  
しかし、このQ値はシーケンサー内部アルゴリズムに基づく理論値であり、実際の誤り率と完全には一致しない。  
装置特性、リード長、塩基種、周囲の配列コンテキスト、リード位置などに依存した系統的偏りが存在する。

**3. 入力データと基本原理**

BQSRの入力には以下の3種が用いられる：

| **入力ファイル** | **内容** | **役割** |
| --- | --- | --- |
| BAM | アライン済みリード（FASTQ由来のBaseQを保持） | 実際の観測値 |
| 参照ゲノム（FASTA） | 正しい塩基配列の基準 | 誤りの判定基準 |
| 既知SNP（VCF） | 既知の多型位置情報 | 真の変異を除外するため |

FASTQは直接使用されないが、各塩基のBaseQ情報はBAMに保持されているため、BAMファイルさえあれば実行可能である。

BQSRは、BAM中の既知SNP（フィルタリング済み）を除いた各塩基について

* その理論的な誤り確率（Q値由来）と
* 参照配列との実際の一致／不一致率  
  を比較する。

**4. 統計的補正の仕組み**

各塩基は、以下のような共変量（covariates）に基づいて分類される：

* 元のベースクオリティ（Q）
* ベース種（A/T/G/C）
* リード内位置（cycle）
* 周囲配列コンテキスト（例：前後1塩基）
* マッピング位置の特性

これらの条件ごとに、

* 観測された総塩基数
* 参照と一致しなかった塩基数  
  を集計し、観測誤り率を算出する。

たとえば、Q=30・base=C・context=GC・cycle=85 の条件下で、  
100,000塩基中120個が不一致なら観測誤り率は 0.0012 であり、  
対応する実測Phred値は

となる。  
このように、観測誤り率から理論的Q値との差を推定し、上記の共変数などを説明変数に機械学習を行い、補正モデル（recalibration table）を作成する。

この過程で生成されるファイルが recal\_data.table であり、  
その後 ApplyBQSR によりBAM内の各BaseQが再計算・置換される。

**5. Variant Callingにおける意義**

Variant Caller（例：GATK HaplotypeCaller）は、各リード上の塩基に付与されたベースクオリティ（BaseQ）を誤り確率 に変換し、それを尤度モデルに組み込んで「真の塩基」を推定する。  
A・C・G・T のいずれかが真の塩基であると仮定し、それぞれの仮説のもとで観測された塩基が得られる確率を以下のように表す：

ここで

εはベースクオリティ Q から求められる誤り確率である。

BaseQ が高いほど誤読の確率が小さく、観測された塩基の信頼性が高いことを意味する。

この尤度は各リードで計算され、すべてのリードの情報を統合して、各塩基（あるいは遺伝型）仮説の尤度が評価される。（どの塩基がその座位において尤もらしいか）

Variant Caller はこれらの尤度を比較し、参照配列と異なる塩基を仮定したモデルの方が統計的に有意に支持される場合に、その位置を「変異（SNP）」として呼び出すというような流れを採用している。

Base Quality（BaseQ）は、各塩基が誤って読み取られる確率（誤り確率 ε = 10⁻^(Q/10)）として尤度モデルに直接組み込まれるため、その値の正確さは変異検出の信頼性に直結する。  
BaseQ が実際のエラー率と乖離している場合、以下のような影響を及ぼす。

* **過小評価されている場合（実際よりも低いQ値が与えられている）**  
  誤り確率 ε が実際よりも大きく見積もられ、真の塩基であっても「誤読かもしれない」として尤度が過小評価される。  
  その結果、参照配列と異なる塩基を支持する証拠の重みが弱まり、真のSNPであっても検出されにくくなる（偽陰性 false negative の増加）。
* **過大評価されている場合（実際よりも高いQ値が与えられている）**  
  誤り確率 ε が過小に見積もられ、観測された不一致が「誤読ではあり得ないと過剰に信頼される。  
  その結果、実際にはランダムエラーによる不一致であっても、モデルが変異仮説を強く支持してしまい、誤ってSNPとして検出される（偽陽性 false positive の増加）。

→つまり、BaseQを正しい値に校正しておくBQSRのプロセスは偽陽性や偽陰性を防ぐための重要な工程である。

**6. まとめ**

Base Quality Score Recalibration（BQSR）は、シーケンサーが出力するベースクオリティスコアを実測誤り率に基づいて再キャリブレーションする処理。  
BAM内の各塩基について、既知SNPを除いた参照との一致・不一致を統計的に評価し、  
ベース種・リード位置・配列コンテキストごとの誤差傾向を学習することで、Q値をより現実的な信頼度へと補正する。

この補正により、Variant Callerの確率モデルが実際のデータ特性に一致し、結果としてSNP検出の精度（感度・特異度）がともに向上する。