Sage v4.0

Background and motivation

In v6, we continue to build on the improvements made in 5.34. While 5.34 was an incremental release offering targeted improvements and fixes to obvious deficiencies, v6 includes comprehensive overhauls of core elements of Sage, while still not changing the high level workflow of Sage calling depicted below:

High level workflow

The holistic, high-level intent of these overhauls is to make Sage a more robust small variant caller.

For example, key changes include:

- BQR now handles sparsity in post-consensus base quals in targeted sequencing, handles different per-base qual distributions, and offers recalibrated quals per consensus type
- The read context generation for candidate variants is rewritten. It's easier to understand, and handles difficult sites like repeat context transitions better than the old logic
- Read counting conventions are overhauled, allowing some reads with a truncated core or with low quality errors in the core to still provide AD and qual support for a variant
- Per-tier minTumorQual soft filter thresholds are replaced by a probabilistic model that adapts to sequencing and site-specific depth, without needing to specify parameterisation
- MSI indels have their qual and jitter characteristics determined by a fitted model based on empirical sample-specific data, similar to BQR for SNVs

High level changes in v6

Category	Changes
BQR	• Round consensus BQ to canonical quals and 1 • Increase min mapping quality for BQR from 10 to 50, and change AD/AF site thresholds • Handle contexts with 0 identified errors • BQR by consensus type (including balanced vs standard for qual 40 Ultima) • Update error rate calculation to be robust to different qual distributions per-base
Read context & evidence counting	• New read context algorithm • Changes to read context annotation • New read counting conventions • Approximate ref matching logic for long inserts • Allow SNV/MNV support in softclips • Fragment sync prefers aligned over non-aligned base quals
Qual model	• Qual filter dynamically scales with depth and splits map and base qual tests • Explicit germline mode which incorporates empirical VAF implausibility into model

	• Novel indel qual is conditionally boosted • Changes to high depth mode • Modify edge penalty more in line with realised error rates, and facilitates partial core qual contribution • Use min base qual rather than mean base qual for MNVs • Indel qual for MSI repeats is based on microsatellite empirical qual • Jitter penalty model is overhauled
Filtering	• MED based on distance to soft clip instead of distance to end of read • MED uses max per-read edge distance for fragments with overlap • MED uses max edge distance across depth-supporting reads, not just ref • MAX_READ_EDGE_DISTANCE_PERC split between panel and non-panel regions • Fix dedupMnv bug • Filter if ALT average fragment length << REF average fragment length • Filter if portion of realigned read support for short indel is unusually high • Filter at candidate phase if only one distinct fragment provides support • Min fragment coordinate check is stepped and applies all the time • Tumor VAF filter now uses recalibrated qual • maxGermlineVAF filter in PANEL is conditionally relaxed

BQR changes

Change	Details
Round consensus BQ to canonical quals and 1	Use of unbucketed consensus qualities leads to many sparse buckets which can lead to inaccurate error estimation. We now round a consensus qual up to the nearest empirical BQ if within 1.5 points, or else rounded down to the nearest empirical BQ or 1.
Increase min MAPQ from 10 to 50	To ensure that BQR is not affected by poorly aligned reads the minMAPQ is raised to 50, BQR now requires min MAPQ >50 to count to BQR.
Split AD/AF site filtering thresholds for DUPLEX	New thresholds: DUAL/BALANCED: (AF<1% or AD<3) AND AF<7.5% OTHER: (AF<5% or AD<4) AND AF <12.5%
Handle contexts with 0 errors	Previously contexts with 0 errors were not adjusted. This caused issues for DUAL context which may have 0 error observations (due to very low error rates). Instead if there are 0 errors the recalibrated qual is set to the qual calculated with 0.5 errors, floored at the raw qual.
BQR by consensus type	Dual stranded observations have typically 2 orders of magnitude error empirical rates and should count more to quality. We therefore split BQR by consensus type: [DUAL/BALANCED, SINGLE, NONE]
Use only T0=BQ=40 bases for ultima	For ULTIMA, only high qual bases (T0=BQ=40) bases are considered to return the pre-sequencing empirical DNA damage rate. Lower T0/BQ SNV errors are likely due to flow errors on ultima sequencer.
Update error rate calculation	Previously for a qual 30 GTA>C error, our error rate was (qual 30 GTA>C count) / (qual 30 GTA>T count). However, this fails if the BAM has no qual 30 Ts. Instead, we do: (qual 30 GTA>C count) / (qual 30 GCA>C count) * (total GCA ref count) / (total GTA ref count)

Read context & evidence counting changes

Change	Details
New read context algorithm	New algorithm for constructing core: • Add 2 bases left (2 bases includes variant pos if the variant is an indel). If leftmost base is part of a repeat of 1-5 lengthˇ with 3+ repeat count then extend to end of repeat. Do this again if this repeat transitions to a new repeat of 6+ count. Break out of repeat if necessary. If last base is non-matching extend to first matching base • Maximally right align variant + add 2 bases right. If rightmost base is part of a repeat of 1-5 length with 3+ repeat count then extend to end of repeat. Do this again if this repeat transitions to a new repeat of 6+ count. Break out of repeat if necessary. If last base is non-matching extend to first matching base Notes: • additional read contexts as candidates IF there are at least max(25% max support,3) reads with FULL support for that read context, as per current logic • For a variant created off a right softclip, maximally left align the variant before creating the read context. If this results in a variant with the same ref/alt/pos/core/flanks as another non-right softclip variant, they will be deduped (e.g. by being considered the same AltContext with the same associated RefContext) For ref context: • Get the ref bases from the aligned core start to the aligned core end
Changes to read context annotation	New VCF annotation called RC_INFO, which contains a delimited set of info, in order: • AlignmentStart (i.e. aligned pos of left flank start) • VarIndex (index of variant, where 0 is left flank start) • LeftFlank • Core • RightFlank • Cigar (of whole read context, including flanks) Other details are stored internally, including: • A ‘ref core’ used for determining REF matching • The core aligned start/end positions (calculated using AlignmentStart and Cigar) • The read homology length for an indel in a repeat context, used to determine right alignment
READ counting conventions	**New read counting rules** • FULL = matches CORE + at least 1 FULL FLANK (i.e. same as 5.34 FULL+PARTIAL) • PARTIAL_CORE = matches on partial CORE with 1 FULL FLANK. The read must reach at least 1 base past the maximal left and right alignment of the variant. Tumor evidence must also fully cover all repeats in the core, plus one base padding • CORE = matches CORE but high qual errors in at least one flank • REF = matches REF context on the core, with partial cores tolerated (no realignment is attempted) • REALIGNED = NONE status when looking L->R but FULL or PARTIAL when looking R->L (internally, the realignment algo finds the read index of the core end position, walks back the length of the core to get the start read index, and checks this against the variant core)

	• SIMPLE_ALT = Matches the base(s) of a variant for an SNV/MNV, or has the expected Cigar element for an indel. Only assessed for germline evidence, and is not assessed inside softclips • NONE = any other read that is considered for counting **Reads to consider for counting (i.e. that provides depth support):** • For an SNV/MNV: must cover all variant bases • For germline evidence for indels: Any read that reaches at least 1 base past the maximal left and right alignment of the variant • For tumor evidence for indels: The above, plus fully covering all repeats in the core, with 1 base padding NOTE: for assessing realignment for a delete of N bases, a read can start as late as variant pos+N **RNA evidence:** • Splice events (i.e. 'N' element in the read cigar) are permitted in the CORE and such reads are counted as PARTIAL_CORE as long as the normal PARTIAL_CORE rules are met. The read is not considered truncated for the purpose of MED. **Depth annotation rules:** • DP = FULL + PARTIAL + CORE + REF + REALIGNED + NONE • AD = FULL + PARTIAL + CORE + REALIGNED • AF = AD/DP SIMPLE_ALT contributes in the germline VAF filter, as does germline jitter for indels of >10 bases, but these aren't considered in the AD or AF values in the VCF. **QUAL annotation rules** • ABQ is based on adjusted base quals for FULL, PARTIAL, CORE and REALIGNED • Variant qual is the phred score associated with TQP, capped at 200 • Hotspot exception for low mappability regions uses total alt BQ instead of RABQ • Special exception to max_germline_vaf for low base qual is also removed **Mismatch allowances:** • FLANK: Allow 3 low qual mismatches in each flank, otherwise downgraded to CORE • CORE: Allow ROUND_UP(trVariantCoreLength/8) low qual mismatches (where the longest repeat is trimmed to 2 repeat counts). For SNV & MNVs the variant base(s) cannot be an error. For INDELs the 1st base of difference between ref and alt contexts cannot be an error, looking either L->R or R->L (specifically, for an insert this means the key bases are the last inserted base and the first base of difference between alt and ref, and for a delete the key bases are the variant pos and the first base of difference between alt and ref). Additionally, the last base of homology and first non-homologous base cannot be an error (either L->R or R->L).

	Jitter counting: • Reads may also count to LENGTHENED or SHORTENED jitter while still counting to a non-alt category above (e.g. if ref is 8xT and alt is 9xT, a 8xT read is REF and also SHORTENED). • The jitter routine is rewritten in an iterative manner, handling realignment and with LENGTHENED and SHORTENED jitter using a similar level of strictness to the new FULL (specifically, one low qual mismatches is tolerated in the padding region of the core, and unlimited low qual mismatches in the FLANK)
Approximate ref matching logic for long inserts	When assessing germline evidence for inserts of >10 bases, we run a supplemental routine where we assign PARTIAL_CORE if the number of matching bases post-variant index against RC is >=2 more than against ref. This makes us more robust to errors (even high qual ones) in exceptionally long cores when the support is still clearly evident.
Allow SNV/MNV support in softclips	If a variant exists in a read's softclip it is still eligible to provide alt support (in evidence phase) if the read matching rules are satisfied. In fragment sync, we clip a read's 3' S element if beyond 5' edge of mate, then remove the ignoreSoftClipAdapter check.

Qual Model

Change	Details
Qual filter as F(depth)	We want to move away from static qual thresholds that don't scale with depth. To do this, we will extend the idea of the p-score AF filter to QUAL. For each variant, set: • strongSupport = FULL + PARTIAL_CORE + REALIGNED read counts • PerReadModifiedBaseQual = Avg(ModifiedBaseQual) across FULL + PARTIAL_CORE + REALIGNED, floored at 15 (20 for MSI indels in probable MSI samples) • TQP = upper p-score of Bin(n=DP, k=strongSupport, p=$10^{-\text{PerReadModifiedBaseQual}/10}$) If pScore is above the threshold then the variant is minTumorQual due to inadequate aggregate base qualities: • HOTSPOT: 1e-2 -> 20 phred score • PANEL: 1e-5 -> 50 phred score • HIGH_CONFIDENCE: 1e-8 -> 80 phred score • LOW_CONFIDENCE: 1e-14 -> 140 phred score We also use aggregate mapping information to assess if the variant actually belongs here, versus plausibly belonging elsewhere or otherwise being artefactual. To start with, define the following (AED = average edge distance): • PerReadModifiedMapQual = Avg(ModifiedMapQual) across FULL+PARTIAL_CORE+REALIGNED supporting reads (for this, we change fixed mapq penalty from 15 to 0) • MQDiffPenalty = 2 * (AMQ[all] - AMQ[alt]) if AMQ[all] > AMQ[alt], else 0 (this penalty is ignored for HLA variants with avg alt mapq > 40) • RSBPenalty = If conditions used to apply RSB filter are satisifed, this is the phred score associated with the p-value of observed RSB. Otherwise, 0 • AEDPenalty = 10 * AD * log10(AED[all] / max(AED[alt], 1)) if AED[alt] / AED[all] < 0.66, else 0 (this penalty is ignored for HLA variants) • MSPenalty = 3 * RC_REPC if len(RC_REPS) > 1 and len(RC_REPS) * RC_REPC >= 15, else 0. This penalty is capped at 18 for indels and 24 for other variants Many of these concepts and thresholds are already used in other parts of Sage. The MSPenalty only applies to cores with a non-homopolymer MaxRepeat, since these are empirically associated with more error phone contexts, particularly at repeat transitions.

	Then MQHeuristic = PerReadModifiedMapQual – 25 - MQDiffPenalty – RSBPenalty – AEDPenalty – MSPenalty. If MQHeuristic < 0 (HC/LC) or < -6 (PANEL/HOTSPOT), the variant is minTumorQual due to weakness in site quality characteristics. If both the p-score and MQHeuristic conditions are satisfied, we have an adequate level of base support and don't see systemic site-side issues, so the variant avoids minTumorQual.
Explicit germline mode which incorporates empirical VAF implausibility into model	If the '-germline' config is passed, tumor/normal based filtering will not be applied, including the hard filter that applies to soft filtered variants with >3 reference AD. Additionally, we penalise the TQP of variants that have a statistically unlikely VAF for a legitimate het/hom germline variant. Specifically, hetP = lower p-score of Bin(n=DP, k=AD, p=0.4). 0.4 is used instead of 0.5 because some variants such as MSI indels will be expected to lose some proportion of their AD due to jitter. If hetP < TQP, we divide TQP by max(hetP, 1e-3). The 1e-3 is used to floor hetP, since at sufficiently low values of hetP a germline copy number event or subclonality becomes a more plausible explanation.
Novel indel qual is conditionally boosted	To boost our sensitivity in calling novel indels not part of a microsatellite (since these variants cannot benefit from recalibration like SNVs/MNVs can), we unwind the 12 point fixed base qual penalty for the purpose of TQP if: • The variant is a non-MSI indel • The variant is in PANEL or HOTSPOT tiers • The variant has no tumor jitter • Average NM of non-alt tumor reads is <0.75 (i.e. it's a clean site)
Changes to high depth mode	High depth mode now has only the following effects: • Reads with pre-recalibration base qual < 30 are not assessed (this rule does not apply to MSI indels since a lot of MSI indels will have empirical quals < 30) • Reads that have discordant or unmapped mates are not assessed • Different default MSI parameters are used with a higher expected rate of jitter (which can also affect if a sample falls back to the defaults)
Modify edge penalty to be more in line with realised error rates, and facilitate partial core qual contribution	We will change the distanceFromReadEdge penalty to allow partial core reads to provide non-0 qual. Specifically, we calculate distanceFromReadEdge as distance from read edge to variant base (SNV), closest variant base (MNV), or closest of the required bases for PARTIAL_CORE (indel). We then subtract: • 15 pts if distanceFromReadEdge = 0 • 5 pts if distanceFromReadEdge = 1 • 0 pts if distanceFromReadEdge > 1. This more closely reflects observed biases.
Use min base qual rather than mean base qual for MNVs	If any of the bases in a potential MNV are systematically low quality, we have lower confidence in the MNV as a whole, and would prefer to call the high quality base(s) as SNVs. So, we use the minimum base's qual for MNVs rather than the average base qual.
Indel qual for MSI repeats is based on microsatellite empirical qual	In 5.34, SAGE used the average base qual over the variant core as the qual contribution for indels. This is now capped for indels in microsatellites of ref repeat length >=4, using an empirical estimate of error rate for each repeat context, repeat length and indel length. For each of 7 MSI repeat contexts {A/T;C/G;AT/TA;CG/GC;AC/CA/GT/TG;3-5mers}, a 6 parameter model {optimal_scale(rep=4); optimal_scale(rep=5); optimal_scale(rep=6); scale_fit_gradient; scale_fit_intercept; microsatellite_skew} is fit (see Microsatellite Jitter Model for details) which specifies {scale, skew} for each repeat length. Specifically: • For repeat lengths 4-6, scale = optimal_scale(rep) and skew = microsatellite_skew • For repeat lengths 7-15, scale = scale_fit_intercept + scale_fit_gradient * rep and skew = microsatellite_skew The capped per-read base qual for (signed) indel length x becomes: -10*log10(modified_asymmetric_laplace(x, scale, skew)), capped at 40. If abs(x) > 5, this model is not used, and we fall back to the old base qual approach.

	If the sample-specific jitter params call more jitter on aggregate than the defaults (with sample-specific skew), we fall back to the defaults and flag the sample as a probable MSI sample. Specifically, we consider the difference in computed base qual between the sample-specific and default params for each valid repeat unit / repeat count / signed indel length, weighted by the count of reads for each permutation. Indels with a computed MSI error rate > 1e-4 do not have the 12pt fixed base qual penalty applied for modified per-read base qual, and are not assessed for minAvgBaseQual filter.
Jitter penalty model is overhauled	The current jitter model, which penalise variants that are potentially sequencing or PCR amplification artefacts, is too simplistic and doesn't handle some sites well (e.g. a real two-alt site where the repeat counts are within one of each other). To address this, we replace it with a probabilistic approach (more details in Microsatellite Qual Plan). Consider a variant where the longest repeat in the core (by repeat count) has numUnits=N. Compare the count of FULL reads to SHORTENED: 1. If SHORTENED > FULL, get modelled error rate for +1 jitter with numUnits = N-1. Suppose this value is e. If FULL/(FULL+SHORTENED) < 2*e or p-value of observed counts > 0.00025, FULL is considered noise of SHORTENED and the variant is soft-filtered. If >0.05, we hard filter the variant 2. Do the same if LENGTHENED > FULL, although this time e is the modelled error rate for –1 jitter with numUnits = N+1 3. If min(LENGTHENED, SHORTENED) >= FULL, do a similar test with e=mean(long_e, short_e), combining long and short jitter for the p-value calculation 4. Otherwise the variant is processed. If LENGTHENED is considered noise from FULL as defined in rule 1, scale variant qual by (FULL+LENGTHENED)/FULL. Same goes for SHORTENED, using the noise calculation in rule 2. The maximum qual boost a variant can receive is +30%. In summary, a repeat count that is less frequent than adjacent repeat count(s) may be considered as noise. If so, it is soft-filtered. Alternatively, a repeat count that is more frequent than adjacent repeat count(s) may have its qual scaled up. NOTE 1: If we need a modelled error rate for a repeat count of length < 4, use 1e-4 NOTE 2: We floor the modelled error rate at 0.04 for PANEL/HOTSPOT variants with a trinucleotide repeat in the core that aren't themselves a non-trinucleotide indel

Filtering changes

Change	Details
MED based on distance to soft clip instead of distance to end of read	When a variant is always seen very close to a softclip, it probably means the variant is an artefact of an incorrect mapping, or otherwise an alternate realignment is possible that makes the variant disappear. While this will cause us to soft filter genuine variants near SVs, the frequency of these is very low compared to the frequency of artefacts we will now filter.
MED uses max per-read edge distance for fragments with overlap	MED is intended to be the max edge distance on a per-read basis. So when a fragment overlaps a variant with both reads, calculate the per-read edge distance (i.e. min of left and right edge distances) and take the max of these.
MED uses max edge distance across depth-supporting reads, not just ref	This is needed to keep MED sufficiently sensitive as a check given REF support is now more sparingly given.
MAX_READ_EDGE_DISTANCE_PERC split between panel and non-panel regions	This makes MED more targeted, given that panel regions tend to be of higher quality. We are now using 0.2 for panel and 0.33 for non-panel regions.
Fix dedupMnv bug	A dedupMnv was fixed which prevented SAGE from deduping a 3 base MNV against a 2 base subset MNV.
Filter if ALT average fragment length << REF average fragment length	If there is a short SV in a fairly homologous region, this can be mapped without an SV with a series of small variants instead, which should be identifiable via the alt vs ref fragment length distribution

	Thus, if max(alt supporting fragment length) < mean(non-alt supporting fragment length) and (0.5 * max(alt supporting fragment length) / mean(non-alt supporting fragment length)) ^ AD < 1e-4, we filter the variant. In this calculation, we only consider fragment lengths of length < 1000 (our existing chimeric threshold) We could easily extend this to unexpectedly high fragment lengths, too.
Filter if portion of realigned read support for short indel is unusually high	For a short indel, we expect the majority of read support to not require realignment. Otherwise, the variant is probably an artefact of a nearby real indel plus a tolerated low-quality error. Thus if a variant has >70% of AD from realigned reads, and is not an indel of length > 10, we filter the variant.
Filter at candidate phase if only one distinct fragment provides support	In Sage v3.4, we require that two distinct fragments provide tumor support, but only require two distinct *reads* (which can be from the same fragment) in candidate phase. Now, we require two distinct fragments in candidate phase as well. This filters out some common artefact variants from further processing.
Min fragment coordinate check is stepped and applies all the time	The min fragment coordinates check now applies all the time (even to hotspots, and even outside high depth mode). However instead of always requiring 3 distinct coordinates, the required number now scales with AD: • AD <= 2 doesn't have a requirement • 3 <= AD <= 4 requires 2 distinct coordinates • AD >= 5 requires 3 distinct coordinates
Tumor VAF filter now uses recalibrated qual	The probabilistic minTumorVAF test now uses recalibrated qual, making the filter more permissive in cases where BQR adjusts base quals up
maxGermlineVAF filter in PANEL is conditionally relaxed	For PANEL tier variants not in a long repeat (i.e. RC_REPC < 10) our prior expectation of a marginal variant tends more towards somatic rather than germline or artefactual. Thus we conditionally increase our maxGermlineVAF filter from 4% to 5%. Additionally, if only one low quality germline read (with below 25 recalibrated base qual) then we also tolerate up to min(10%, tumorAF/3) germline VAF.