

1

2

3

4 **Lacer: accurate base quality score recalibration for improving variant calling from**
5 **next-generation sequencing data in any organism**

6 Jade C.S. Chung^a and Swaine L. Chen^{a,b,*}

7 ^aNational University of Singapore, Department of Medicine, Yong Loo Lin School of
8 Medicine, 1E Kent Ridge Road, NUHS Tower Block, Level 10, Singapore 119074; ^bGenome
9 Institute of Singapore, Infectious Diseases Group 3, 60 Biopolis Street, Genome, Level 6,
10 Singapore 138672.

11 *Corresponding author. Tel.: +6568088074. Email: slchen@gis.a-star.edu.sg.

12

13 **Running title:** Accurate base recalibration for all organisms

14 **Keywords:** base quality score recalibration, NGS, genetic variation

15

16 **Abstract**

17 Next-generation sequencing data is accompanied by quality scores that quantify sequencing
18 error. Inaccuracies in these quality scores propagate through all subsequent analyses; thus
19 base quality score recalibration is a standard step in many next-generation sequencing
20 workflows, resulting in improved variant calls. Current base quality score recalibration
21 algorithms rely on the assumption that sequencing errors are already known; for human
22 resequencing data, relatively complete variant databases facilitate this. However, because
23 existing databases are still incomplete, recalibration is still inaccurate; and most organisms do
24 not have variant databases, exacerbating inaccuracy for non-human data. To overcome these
25 logical and practical problems, we introduce Lacer, which recalibrates base quality scores
26 without assuming knowledge of correct and incorrect bases and without requiring knowledge
27 of common variants. Lacer is the first logically sound, fully general, and truly accurate base
28 recalibrator. Lacer enhances variant identification accuracy for resequencing data of human
29 as well as other organisms (which are not accessible to current recalibrators), simultaneously
30 improving and extending the benefits of base quality score recalibration to nearly all ongoing
31 sequencing projects. Lacer is available at: <https://github.com/swainechen/lacer>.

32 **Introduction**

33 Next-generation sequencing (NGS) has revealed broad insights into the prevalence and
34 function of genetic variation, especially with respect to single nucleotide polymorphisms
35 (SNPs) that influence human disease (Thorisson et al. 2005; Cotton et al. 2008; The 1000
36 Genomes Project Consortium 2012; The Cancer Genome Atlas Network 2012). Given the
37 scale of the data, even incremental increases in accuracy have an enormous impact. Accurate
38 SNP identification relies on accurate sequencing, which is quantified by base quality scores;
39 unfortunately, machine-reported qualities are inaccurate (Li et al. 2009b). Therefore,

40 recalibration of base quality scores improves downstream variant calling, mostly by
41 excluding false positive SNPs (Li et al. 2009b; DePristo et al. 2011).

42 Quality score recalibration is trivial if the status of every base (correct or error) is known; the
43 fraction of sequencing errors with a given quality score can be used to calculate the empirical
44 (recalibrated) quality. For real sequencing data, however, erroneous bases are of course not
45 already known. Intriguingly, all current recalibrators (Li et al. 2009b; DePristo et al. 2011;
46 Zook et al. 2012; Cabanski et al. 2012) are strongly based on this assumption that erroneous
47 bases are known; sequencing errors are identified as mismatches to a reference genome,
48 excluding sites of known variants (e.g., dbSNP (Sherry et al. 2001) for humans). This
49 assumption would be tenable if variant databases were complete, but this is also not the case
50 (The 1000 Genomes Project Consortium 2010), and the purpose of sequencing is often to
51 discover variants not present in existing databases. Furthermore, outside of humans and
52 several model organisms, variant databases are not available and thus recalibration is often
53 not done.

54 We have solved these logical and practical problems with a new algorithm for recalibrating
55 base quality scores, Lacer. By comparing Lacer with GATK (DePristo et al. 2011), the
56 current standard for base recalibration software, on four Illumina sequencing data sets
57 (*Escherichia coli*, human, macaque and marmoset), we show that Lacer more accurately
58 recalibrates base quality scores in the absence of complete (or any) variant knowledge,
59 enabling application to any organism. Lacer's more accurate recalibration in turn results in a
60 specific increase in the confidence of true variants (based on variant quality scores), leading
61 to more effective exclusion of false positive variants and more accurate variant calls. Finally,
62 we demonstrate that Lacer can be applied to a wide range of NGS data sets, including
63 different types of sequencing data (whole genome, exome, metagenome, and targeted
64 sequencing) and data from different sequencing platforms (Illumina, Ion Torrent, and 454).

65 **Results**

66 **The Lacer algorithm**

67 Lacer (summarized in **Fig. 1A**) uses linear algebra concepts instead of assuming that correct
68 and incorrect bases can be directly identified. Briefly, sequencing reads are mapped to a
69 reference, and bases are binned into sets based on consensus status and depth of coverage.
70 The sorting leverages the intuition that high coverage, consensus bases are likely to be
71 correct, whereas singly observed, nonconsensus bases at high coverage positions are likely
72 incorrect. Each set of bases defines a histogram of quality scores. The collection of
73 histograms is cast as a matrix and analyzed by singular value decomposition (SVD).
74 Assuming that correct and incorrect bases have a consistent distribution of quality scores and
75 that the percentage of incorrect bases in each set varies, SVD will extract information that can
76 be used to infer these distributions and the percentage of incorrect bases (for further details,
77 see **Methods**); importantly, no individual bases are assumed to be either correct or incorrect.
78 Finally, a Bayesian calculation based on the inferred aggregate frequency of incorrect bases
79 and the distributions of correct and incorrect quality scores yields recalibrated quality scores.

80 **Lacer matches a gold standard GATK recalibration on bacterial resequencing data** 81 **regardless of reference sequence**

82 We first compared Lacer with GATK, the current standard for base recalibration software, on
83 an Illumina resequencing data set for *Escherichia coli* UTI89 (Chen et al. 2006), for which a
84 complete reference sequence is available; in other words, for this data set we do know that all
85 mismatches are indeed errors. Lacer and GATK yielded nearly identical results on UTI89
86 when the UTI89 genome itself was used as a reference sequence (**Fig. 1B**). Consistent with
87 this, the quality profiles for error bases predicted by both Lacer and GATK were nearly
88 identical (**Fig. 1C**) and matched a third method (Quake (Kelley et al. 2010)) that identifies
89 sequencing errors using k-mer analysis (**Supplemental Fig. S1A**). GATK performs a second

90 order recalibration incorporating covariates (sequence context and cycle number); Lacer also
91 matched these well (**Supplemental Fig. S1B**). Therefore, in the presence of perfect
92 information about which bases are correct and incorrect (using the known reference
93 sequence), satisfying the assumptions made by GATK, Lacer and GATK provide concordant
94 recalibration, which we take as gold standard recalibration for this particular UTI89
95 resequencing data.

96 In general, there is not perfect information about mutations; to mimic this, we recalibrated the
97 same sequencing data using a different reference genome, *E. coli* MG1655 (which is ~98%
98 identical to UTI89). Lacer's overall and covariate recalibration still matched the gold
99 standard (GATK using the UTI89 reference sequence), but GATK (using the MG1655
100 reference sequence) reported a dramatically different recalibration, based on overall base
101 quality recalibration, second order recalibration, and concordance of error base quality score
102 profile with Quake (**Fig. 1B, Supplemental Fig. S1B,C**). This could not be fully mitigated
103 by providing GATK with a maximal variant database generated by marking SNP calls from
104 uncalibrated data as well as potentially mismapped reads (**Fig. 1B**), while Lacer's
105 recalibration was unaffected by exclusion of sites in this bootstrapped variant database
106 (**Supplemental Fig. S1D**). Therefore, Lacer provides an accurate recalibration independent
107 of an imperfect reference sequence and of knowledge of known variants, suggesting that
108 Lacer might effectively be used to recalibrate sequencing data for any organism.

109 **Lacer accurately recalibrates human resequencing data without a perfect reference** 110 **sequence**

111 We therefore tested Lacer on human sequencing data for the CEPH individual NA12878 (The
112 1000 Genomes Project Consortium 2012), where no true gold standard recalibration is
113 available. As seen with the UTI89 data using an imperfect reference, Lacer recalibrated the
114 NA12878 chr1 data (mapped to the GRCh37.p13 reference genome) to higher quality scores

115 than GATK (**Fig. 2A**) and was insensitive to the provision of known variants (**Supplemental**
116 **Fig. S1E**). To definitively verify the independence of Lacer from known variants, we
117 restricted Lacer to only dbSNP sites and found that it produced a similar recalibration to that
118 obtained from the full data set (**Fig. 2A**). To reconcile the differences between the Lacer and
119 GATK recalibrations, we again used Quake as an independent method. The quality profiles
120 for error bases predicted by Lacer and Quake were similar; however, error bases identified by
121 GATK had consistently higher quality scores (**Supplemental Fig. S1F**), somewhat
122 resembling the quality profile of correct bases (**Fig. 2B (inset), Supplemental Fig. S1G**).

123 We suspected that the discrepancy between the GATK and Lacer recalibrations was because
124 GATK misclassified some correct bases; an incomplete variant database would result in bases
125 supporting novel variants being classified as incorrect, and in the extreme, without a variant
126 database, GATK indeed recalibrates to lower quality (**Fig. 2A**). We therefore examined bases
127 of a given quality score (for example, Q39 (arrow in **Fig. 2A**)) that did not match the
128 reference and that were not in the variant database. These are bases that would be called
129 errors by GATK. We extracted bases that supported these “errors” – bases from different
130 reads but mapping to the same position with the same nucleotide. The quality score
131 distribution of these supporting bases was similar to the quality score distribution of correct
132 bases as predicted by Lacer (**Fig. 2B**), suggesting that these bases were largely correct.
133 However, a fraction of the Q39 “errors” was not supported (“single”). This fraction was
134 relatively small for both the NA12878 chr1 data set and for UTI89 mapped to the MG1655
135 genome, but was ~90% for UTI89 mapped to the UTI89 genome (**Supplemental Fig. S2A**).
136 The increase in single bases was not due to generally lower coverage at these positions
137 (**Supplemental Fig. S2B**). By hypothesizing that single bases were the only true incorrect
138 bases, we could derive the difference in recalibrated quality scores between Lacer and GATK
139 (**Fig. 2C, Supplemental Fig. S2C**). In other words, amending error bases identified by

140 GATK to include only single bases and to exclude all supported bases (whose overall quality
141 profile was similar to that for correct bases) resulted in concordant recalibration with Lacer.
142 Given the results using both *E. coli* and human data, we conclude that Lacer provides a more
143 robust and more accurate recalibration regardless of organism.

144 **Lacer recalibration improves SNP calling on human sequencing data**

145 Using the GATK best practices pipeline (DePristo et al. 2011; Van der Auwera et al. 2013) to
146 call SNPs on NA12878 chr1, we next compared no base quality recalibration to Lacer and
147 GATK (**Table 1**). Recalibration with either Lacer or GATK resulted in ~250,000 final calls;
148 without calibration, an additional 45-57,000 SNPs were identified. SNPs excluded by Lacer
149 appeared to be of lower quality, based on the transition-transversion ratio (Ti/Tv) of 1.49 for
150 Lacer compared to 1.66 for GATK. The expected value for true positives is ~2. On a high
151 confidence, true positive SNP call set (NIST (Zook et al. 2014)), Lacer and GATK both
152 predicted ~7,000-8,000 unique SNPs each; the Ti/Tv for the unique Lacer SNPs was closer to
153 the value in the intersection and higher than the value for the unique GATK SNPs (**Fig. 2D**).
154 Therefore, recalibration with Lacer provides more effective exclusion of false positive SNPs
155 (the primary benefit of recalibration) than with GATK.

156 To discover why Lacer excluded more false positive SNPs, we examined the final variant
157 quality scores (VQS) of SNPs following variant quality score recalibration (VQSR). Lacer
158 recalibration produced overall higher VQS than GATK or uncalibrated SNP calls on the
159 high-confidence NIST SNPs (**Fig. 2E**). Notably, GATK resulted in generally lower VQS than
160 uncalibrated data. To verify that Lacer was not simply elevating the quality scores of all
161 bases, we took the set of final high and low quality SNP calls from GATK recalibration and
162 examined their VQS in the GATK, Lacer, and uncalibrated data. Lacer had increased the
163 VQS on high quality SNPs without affecting VQS for low quality SNPs, as expected, when
164 compared with GATK or no recalibration. (**Fig. 2F, Supplemental Fig. S2D**). Again, GATK,

165 in contrast, reduced the VQS specifically on high quality SNPs compared with uncalibrated
166 data, the opposite of what would be expected (**Supplemental Fig. S2E**). We noticed a
167 bimodal distribution in the VQS differences between GATK and uncalibrated or Lacer
168 recalibrated data (potentially due to bimodal base quality scores after GATK recalibration)
169 (**Supplemental Fig. S2F**). We therefore performed the entire analysis on another
170 chromosome (chr19) that didn't have this artifact and saw similar improved results for Lacer
171 (**Supplemental Fig. S3, Supplemental Table S1**). Thus, Lacer recalibration results in a
172 specific increase in the confidence of true SNPs without inflating the confidence of false
173 SNPs, as would be expected from accurate recalibration, while GATK has the opposite effect.

174 **Lacer recalibration of non-human sequencing data results in improved SNP calls**

175 To further show that Lacer can effectively recalibrate the sequencing data for any organism in
176 the absence of a perfect reference, we tested Lacer on two non-human primate data sets,
177 Tibetan macaque (*Macaca thibetana*) (Fan et al. 2014) and common marmoset (*Callithrix*
178 *jacchus*). Lacer recalibrated the Tibetan macaque chr1 data (mapped to the rhesus macaque
179 (*Macaca mulatta*) reference) to higher quality scores than GATK; given that known variants
180 in this organism are likely incomplete, this was similar to the result obtained with the human
181 data and the *E. coli* data mapped to an imperfect reference (**Fig. 3A, compare with Fig. 1B**
182 **and Fig. 2A**). Furthermore, error bases identified by GATK again had consistently higher
183 quality scores, and the quality profile resembled that of correct bases predicted by both
184 GATK and Lacer (**Supplemental Fig. S4A,B**). In the presence of a complete reference
185 sequence (calJac3 (The Marmoset Genome Sequencing and Analysis Consortium. 2014)),
186 Lacer and GATK again yielded concordant results on the common marmoset exome data
187 (**Fig. 3B**). To mimic the absence of perfect information about mutations, we recalibrated the
188 same sequencing data using the rheMac3 reference genome. Again Lacer's recalibration
189 matched the recalibration on the perfect data, but GATK's recalibration was very different

190 **(Supplemental Fig. S4C,D)**. Unfortunately, we could not independently verify the quality
191 score profiles of error bases predicted by Lacer using Quake due to the low coverage (<20×)
192 of these data sets.

193 We next used the GATK best practices pipeline and the quality score-aware SNP caller,
194 LoFreq (Wilm et al. 2012), to call SNPs on these data sets (**Table 1**). For the Tibetan
195 macaque chr1 data, recalibration with either Lacer or GATK resulted in ~1,300,000 final
196 calls; without calibration, an additional 180-190,000 SNPs were identified. Recalibration by
197 Lacer led to a slightly higher Ti/Tv in the higher call set than GATK; but the SNPs unique to
198 Lacer (~7,000) had a much higher Ti/Tv (1.70) than SNPs unique to GATK (~20,000, Ti/Tv
199 0.94) (**Fig. 3C**). For the common marmoset exome data, recalibration with either Lacer or
200 GATK resulted in ~2,400,000 final calls; without calibration, an additional ~4.6 million
201 SNPs were identified. Once again, SNPs unique to Lacer (~39,000) had a higher Ti/Tv ratio
202 (1.86) than SNPs unique to GATK (~54,000; 1.59) (**Fig. 3D**). Therefore, accurate
203 recalibration by Lacer provides substantial improvement to SNP calling without requiring a
204 variant database or a perfect reference sequence.

205 **Discussion**

206 Lacer solves the two primary practical and logical problems with current recalibrators and is
207 thus the first and only way at present to correctly recalibrate most NGS data (including
208 exome, metagenome, targeted sequencing, and data across multiple sequencing platforms)
209 (**Supplemental Fig. S5**). Lacer (i) does not require a variant database, (ii) can tolerate an
210 imperfect reference sequence (Lacer may in future be extended to utilize k-mer analysis,
211 potentially eliminating the need for reference sequences entirely), and (iii) is more accurate
212 across a wider range of data sets (including human) than GATK, resulting in better

213 downstream variant calls. Lacer is the first algorithm to extend the increased accuracy of
214 variant calling from base quality recalibration to all non-human resequencing projects.

215 Current base recalibrators assume that any mismatch to a reference genome is a sequencing
216 machine error. In reality, these mismatches may also include true variants, PCR-amplification
217 errors, and data processing errors (e.g. alignment errors). These three classes of mismatches
218 arise from sources outside of the sequencing process itself; in other words, if there is a true
219 unknown variant, current recalibrators will treat those bases as errors despite the sequencing
220 machine having sequenced them correctly. A complete variant database will only prevent
221 misidentification true variants as error bases, and mapping scores could be used to minimize
222 the impact of alignment errors. However, errors introduced during PCR cannot be eliminated
223 by current recalibrators. In practice, we find that current recalibration algorithms significantly
224 (10- to 100-fold) underestimate empirical quality scores, especially among high quality bases,
225 even when provided with a variant database. Importantly, these high quality bases account for
226 the majority of the data (see, for example, the black bars in **Fig. S2F**). In contrast, the Lacer
227 algorithm directly extracts quality scores and aggregate error probabilities based on the
228 assumption that correct and incorrect bases have different quality score profiles. The
229 calculation of consensus bases effectively eliminates unknown true variants as a source of
230 error and seems to mitigate against mapping errors. Based on the bacterial and human
231 sequencing data, single (unsupported), non-consensus bases are the majority of true
232 sequencing errors; in other words, the majority of supported bases, even with only 2
233 supporting bases at high coverage ($>50\times$) positions, are actually correct. These bases
234 generally do not result in a SNP call by current algorithms, and we therefore hypothesize that
235 these bases may be due to PCR amplification or other unknown sources of error. Regardless,
236 Lacer is robust to all of these potential sources of error that affect current recalibrators, and in
237 so doing is the first recalibrator to effectively isolate and correct the errors introduced during

238 the sequencing process itself, which is precisely what the base quality scores should be
239 measuring.

240 Lacer specifically increases the confidence (or VQS) of high quality SNPs without affecting
241 the confidence of low quality SNPs. Intriguingly, GATK's recalibration resulted in a
242 *reduction* in VQS for high quality SNPs; again, incomplete variant knowledge leads to the
243 misclassification of error bases, a reduction in the empirical quality score, and the
244 concomitant decrease in VQS. Interestingly, these lowered VQS did not significantly impact
245 GATK's ultimate SNP calls from the human data, based on Ti/Tv ratios and comparison to
246 the high confidence NIST SNP set. This may be a result of the fact that GATK's best
247 practices pipeline is tuned for the identification of genetic variants in human sequencing data;
248 the availability of large human training data sets enables more effective prediction of SNPs.
249 The tuning of GATK for human data, however, limits its utility for non-human sequencing.
250 Indeed, when applied to non-human data sets, Lacer's consistent performance relative to
251 GATK is apparent; not only are false positive SNPs more effectively excluded, but unique
252 SNPs predicted from Lacer-recalibrated data also have quality metrics more consistent with
253 that of true positive SNPs. Use of the LoFreq SNP caller (instead of the GATK
254 HaplotypeCaller) for these data sets more effectively captures the benefit of accurate
255 recalibration, since LoFreq takes into account base quality scores for SNP prediction.
256 Furthermore, VQSR is not applicable to non-human data sets due to the lack of training data,
257 which substantially reduces the accuracy of SNP identification.

258 Currently, there are very few methods to objectively and computationally assess the accuracy
259 of variant calls; since Lacer uses the distribution of quality scores to calculate aggregate
260 numbers of correct and incorrect bases, it could conceivably accomplish this when combined
261 with the analysis of supporting bases. A further extension could enable the systematic
262 assessment of variant calling covariates to identify rigorous filtering criteria. Finally, future

263 SNP callers may gain additional power and accuracy by directly incorporating these insights
264 about quality scores.

265 **Methods**

266 **Lacer algorithm – theory**

267 We first assume that correct and incorrect bases have consistent but different reported quality
268 distributions. The quality score distribution of correct bases is denoted as $C = \{ c_j \}$ and that
269 of incorrect bases as $E = \{ e_j \}$. C and E are probability distributions over the set of assigned
270 quality scores and therefore satisfy $\sum_j c_j = \sum_j e_j = 1$. Considering a set s containing n bases, s

271 will have a fraction p correct bases and $(1 - p)$ incorrect bases. Since correct and incorrect
272 bases have consistent distributions, s is sampled from the quality score distribution:

$$273 \quad \{ q_j \} = \{ p c_j + (1 - p) e_j \}$$

274 also satisfying $\sum_j q_j = 1$. Clearly, as n increases, the sampling error of $\{ q_j \}$ decreases. Given
275 a collection of sets $S = \{ s_i \}$, each with n bases and containing a fraction p_i of correct bases, a
276 collection of quality profiles $Q = \{ q_{ij} \}$ can be created and represented as a matrix.

277 Crucially, for any given i :

$$278 \quad q_{ij} = \{ p_i c_j + (1 - p_i) e_j \}$$

279 Considering this as a vector, and given c_j and e_j , differences between q_{ij} can be parameterized
280 with one variable, p_i . Although p_i , c_j and e_j are *a priori* unknown, providing p_i are not all
281 identical, a singular value decomposition (SVD) can extract the covariance between the
282 individual quality score variations and enable deduction of p_i . Explicitly:

$$283 \quad q_{ij} = \{ e_j + p_i (c_j - e_j) \}$$

284 An SVD produces $c_j - e_j$ (defined up to a sign) as the first eigenvector, while p_i is related (up
285 to addition of a constant due to data centering) to the eigencoordinates in the first
286 eigendimension. Careful construction of the sets of bases $S = \{ s_i \}$ provides a reasonable
287 assurance that p_i spans close to the entire range from 0 to 1; this is done as described in the
288 main text by sorting bases by consensus status (including number of supporting bases) and
289 depth of coverage at that position. Thus, c_j and e_j can be directly calculated from the SVD
290 results as $\min(x_{Ij}) v_{Ij}$ and $\max(x_{Ij}) v_{Ij}$ (adjusted for centering), respectively, where x_{Ij} are the
291 coordinates in the first SVD dimension and v_{Ij} are the components of the first SVD
292 eigenvector.

293 Once c_j and e_j are known, a Bayesian calculation determines the relationship between
294 reported and empirical quality. Empirical quality can be recorded as:

$$295 \quad q_{\text{empirical}} = -10 \log_{10}(P(\text{error}|q_{\text{reported}}))$$

296 By Bayes theorem:

$$297 \quad P(\text{error}|q_{\text{reported}}) = P(q_{\text{reported}}|\text{error}) P(\text{error}) / P(q_{\text{reported}})$$

298 Tabulation of the overall distribution of quality scores in the sequencing data set gives
299 $P(q_{\text{reported}})$. $P(q_{\text{reported}}|\text{error})$ is given by $E = \{ e_j \}$. Importantly, since each set s_i contains n
300 bases and the SVD provides p_i , the total number of error bases can be calculated as
301 $\sum_i n(1 - p_i)$ (without classifying individual bases as correct or incorrect). Division by the
302 total number of bases then yields $P(\text{error})$.

303 **Sequencing data and data processing**

304 **Human data sets.** The sequencing data set for the CEPH individual NA12878 (sequenced on
305 an Illumina Genome Analyzer II and aligned to the GRCh37.p13 human reference genome)
306 was publically available from the 1000 Genomes database (The 1000 Genomes Project

307 Consortium 2012). The high-confidence variant call set used for benchmarking was available
308 at the NCBI Genome in a Bottle ftp site (Zook et al. 2014):
309 NISTIntegratedCalls_14datasets_131103_allcall_UGHapMerge_HetHomVarPASS_VQSRv
310 2.18_all_nouncert_excludesimplerep_excludesegdups_excludedecoy_excludeRepSeqSTRs_n
311 oCNVs.vcf. The BAM file for the NA12878 exome data set (sequenced on the Illumina
312 Genome Analyzer Iix and aligned to the NCBI36 human reference genome using Maq) was
313 available from the 1000 Genomes Project database (The 1000 Genomes Project Consortium
314 2012). The BAM file for the TruSeq Amplicon Cancer Panel targeted sequencing data
315 (sequenced on an Illumina MiSeq and aligned to the NCBI36 human reference genome) was
316 available from the BaseSpace Public Data repository
317 (<https://basespace.illumina.com/run/358358>).

318 Sequencing data were aligned to the corresponding reference genomes using BWA (Li and
319 Durbin 2009). SAMtools (Li et al. 2009a) was utilized for alignment, sorting, and indexing of
320 sequencing reads. The aligned data was then processed following the GATK (version 2.8-1)
321 best practices workflow (DePristo et al. 2011; Van der Auwera et al. 2013). For the NA12878
322 whole genome sequencing data set, a variant database for GATK's recalibration was
323 generated by including variant positions from dbSNP (BUILD_ID 137, reference
324 GRCh37.p10), Mills and 1000 Genomes gold standard indels (reference GRCh37.p10), and
325 1000 Genomes Phase 1 indels (reference GRCh37.p10), which were downloaded from the
326 GATK resource bundle (version 2.5). For the NA12878 exome and TruSeq Amplicon Cancer
327 Panel targeted sequencing data sets, a variant database for GATK's recalibration was
328 generated using the GATK resource bundle for the NCBI36 reference genome. Blank VCF
329 files for all data sets were generated by including mock VCF headers and fields alone.

330 **Non-human data sets.** For the UTI89 Illumina sequencing data set, UTI89 was grown in LB
331 broth overnight at 37°C with agitation. Genomic DNA was purified using the Wizard

332 genomic DNA purification kit (Promega). DNA was sheared using the Covaris S2x and size
333 selected on a Pippin Prep (Sage Science) to isolate 400-500 bp fragments. Illumina
334 sequencing libraries were constructed using the Illumina TruSeq kit using manufacturers
335 protocols, then sequenced on a HiSeq 2000 with 2×76 paired end reads. For the UTI89 Ion
336 Torrent sequencing data set, UTI89 was grown in LB broth overnight at 37°C with agitation.
337 Genomic DNA was purified using the Wizard genomic DNA purification kit (Promega). Ion
338 Torrent sequencing libraries were constructed using the Ion Xpress Plus Fragment Library kit
339 (Life Technologies) according to the manufacturer's protocols, then sequenced on an Ion 316
340 chip. The MG1655 454 data set (sequenced on a Roche 454 GS FLX Titanium System) was
341 available from the SRA database (SRX255226; www.ncbi.nlm.nih.gov/sra). The UTI89 and
342 MG1655 sequencing data were aligned to the *E. coli* UTI89 (NC_007946) or *E. coli* MG1655
343 (NC_000913) reference genomes. The SRS017191 human metagenome (sequenced on an
344 Illumina Genome Analyzer Iix and aligned to the *Bacteriodes vulgatus* ATCC 8482
345 (NC_009614) reference genome) data set was available from the Human Microbiome Project
346 database (The NIH HMP Working Group 2009). The Tibetan macaque (*Macaca thibetana*)
347 data set (sequenced on an Illumina HiSeq 2000) was available from the SRA database
348 (SRX373102); sequences were trimmed to equal lengths and aligned to the *Macaca mulatta*
349 (rheMac3; rhesus macaque) reference genome. The common marmoset (*Callithrix jacchus*)
350 exome data set (sequenced on an Illumina HiSeq 2000 and aligned to the calJac3 (*Callithrix*
351 *jacchus*; The Marmoset Genome Sequencing and Analysis Consortium. 2014) or rheMac3
352 reference genomes) was available from the SRA database (SRX375642).

353 The sequencing data were aligned to the corresponding reference genomes using BWA.
354 SAMtools was utilized for alignment, sorting, and indexing of sequencing reads. With the
355 exception of the Tibetan macaque and marmoset data sets, which were processed following
356 the GATK best practices workflow, all aligned data were recalibrated directly. A variant

357 database for GATK's recalibration was generated by including variants called by SAMtools,
358 LoFreq (Wilm et al. 2012), and where indicated, positions showing poor mappability.

359 **Supporting base analysis**

360 The SAMtools Perl API and custom Perl scripts were used to perform pileups and extract
361 nonconsensus bases, their corresponding quality, and the same data from supporting bases
362 (identical bases from different reads) at positions outside of dbSNP from the aligned
363 NA12878 data. Using custom R scripts, histograms of supporting base qualities were plotted
364 and compared to the Lacer predicted correct and error base quality profiles. For these, the
365 target base quality was excluded; for example, an analysis of Q39 bases utilized only quality
366 score histograms for Q6-Q38 and Q40. The fraction of single (unsupported) bases (out of the
367 total number of nonconsensus, non-VCF bases; denoted $f_{Q;\text{single}}$) was plotted for each quality
368 score, and converted to a recalibration amendment as $-10\log_{10}(f_{Q;\text{single}})$ for each quality score
369 Q from 30-40.

370 **Quake analysis**

371 Quake (version 0.3) (Kelley et al. 2010) was run on the FASTQ file for UTI89 using a k-mer
372 of 13 and the --log option to output the bases and associated qualities that were corrected.
373 Quake was similarly run on a downsampled (to 1%) FASTQ file of reads mapping to chr1 or
374 chr19 of NA12878 using a k-mer of 15. The histogram of quality scores for bases corrected
375 by Quake was taken as the error profile of bases identified by Quake.

376 **Program availability**

377 Lacer is available at <https://github.com/swainechen/lacer>.

378 **Acknowledgments**

379 This research was supported by the National Research Foundation, Prime Minister's Office,
380 Singapore under its NRF Research Fellowship Scheme (NRF Award No. NRF-RF2010-10)

381 and the Genome Institute of Singapore (GIS)/Agency for Science, Technology and Research
382 (A*STAR). We thank See Ting Leong, Dawn Choi and Xiaoan Ruan for the preparation and
383 sequencing of the UTI89 Ion Torrent sequencing library. We thank Andreas Wilm, Niranjan
384 Nagarajan, Shyam Prabhakar, and members of the Chen lab for helpful discussions and
385 suggestions regarding the algorithm and manuscript. J.C.S.C. and S.L.C. designed the
386 algorithm, performed the analyses, and wrote the paper.

387

388 **Figure Legends**

389 **Figure 1: Lacer and recalibration of *E. coli* sequencing data.** (A) Workflow of the Lacer
390 algorithm. For details, see main text and **Methods**. (B) Empirical (recalibrated) quality scores
391 plotted against reported (uncalibrated) quality scores for the UTI89 sequencing data from
392 GATK, using UTI89 or MG1655 as a reference sequence (with or without excluding variant
393 positions based on poorly-mapped positions and variant calls from SAMtools and LoFreq)
394 (w/ or w/o VCF, respectively); and from Lacer, using UTI89 or MG1655 as reference and
395 without excluding variants. Lacer recalibration using UTI89 or MG1655 as a reference
396 matches the gold standard GATK recalibration using UTI89 as the reference. GATK
397 recalibration using MG1655 as the reference results in generally lower recalibrated quality
398 scores which are not fully rescued by provision of a variant database (w/ VCF). (C) Quality
399 score distribution of correct and erroneous bases following recalibration of UTI89 sequencing
400 data by GATK (with UTI89 as a reference) or Lacer (with MG1655 as a reference). The
401 distributions generated by Lacer using the MG1655 reference match those generated by
402 GATK using the UTI89 reference.

403 **Figure 2: Recalibration of human sequencing data.** (A) Recalibration of the NA12878
404 chr1 sequencing data using GATK, with or without excluding known variants based on
405 dbSNP, Mills and 1000 Genomes gold standard indels, and 1000 Genomes Phase 1 indels (w/
406 or w/o VCF, respectively), or using Lacer, w/o VCF or with dbSNP variant positions only
407 (VCF positions only). The Lacer recalibration results in generally higher recalibrated quality
408 scores compared with GATK, and these are not rescued by provision of a variant database
409 (w/ VCF). Recalibration using only VCF positions by Lacer approaches the Lacer
410 recalibration using the full data set. (B) Quality score distributions of error (red) and correct
411 (green) bases defined by Lacer recalibration and of supporting bases (black) for all non-
412 reference, non-VCF bases with a quality score of 39 (Q39) for the NA12878 chr1 data set.

413 All Q39 bases themselves have been excluded from this graph. The distribution for the Q39
414 supporting bases is similar to the distribution for the correct bases. Inset, quality score
415 distributions for error (blue) and correct (yellow) bases defined by GATK recalibration of the
416 NA12878 chr1 data set. The quality score distribution for error bases is enriched for high
417 quality scores and somewhat resembles the distribution for correct bases. Only data for
418 quality scores greater than 25 are shown. The full distribution is shown in **Supplemental Fig.**
419 **S1G. (C)** Correction of GATK recalibration using supporting base analysis. Dark gray bars
420 represent the difference between Lacer and GATK recalibration for each quality score. Light
421 gray bars represent quality difference from the GATK recalibration assuming only single
422 (unsupported), non-reference, non-VCF bases are true errors. **(D)** Venn diagram of Lacer-
423 and GATK-recalibrated SNP calls for NA12878 chr1 identified by GATK HaplotypeCaller
424 that passed recalibration by the GATK VariantRecalibrator (VQSR) and that also matched
425 the high-confidence NIST call set. Total numbers of SNP calls (black) in each category and
426 their associated Ti/Tv ratios (red) are shown. SNP calls unique to Lacer recalibration have an
427 aggregate Ti/Tv closer to the intersection than the SNP calls unique to GATK recalibration.
428 **(E)** Comparison of VQS for SNPs identified in the uncalibrated, Lacer-recalibrated, or
429 GATK-recalibrated NA12878 chr1 data which matched the NIST call set. Lacer recalibrated
430 data results in overall higher VQS than that from uncalibrated (red) and GATK recalibrated
431 (green) data on these high confidence true positive SNPs. GATK recalibrated data results in
432 overall lower VQS than uncalibrated (gray) data on these same SNPs. **(F)** Difference in VQS
433 between the Lacer- and GATK-recalibrated call sets for NA12878 chr1 for all high quality
434 (PASS; green) or low quality (LowQual; red) SNPs in the GATK-recalibrated call set. Lacer
435 recalibrated data results in a mild increase in VQS on low quality SNPs (as called by GATK
436 recalibrated data), but this is small compared to the increase in VQS for high quality SNPs.

437 **Figure 3: Recalibration of non-human primate sequencing data.** Recalibration of (A)
438 Tibetan macaque chr1 sequencing data using Lacer or GATK, with or without excluding
439 known variants based on SAMtools and LoFreq (w/ or w/o VCF, respectively). GATK
440 recalibration produces lower quality scores that are only partially increased by exclusion of
441 known variants, and (B) common marmoset exome sequencing data using Lacer or GATK,
442 with or without excluding known variants based on SAMtools and LoFreq (w/ or w/o VCF,
443 respectively), and the calJac3 or rheMac3 reference genomes. Lacer recalibration using
444 calJac3 or rheMac3 as a reference matches the gold standard GATK recalibration using
445 calJac3 as the reference. GATK recalibration using rheMac3 as the reference results in
446 generally lower recalibrated quality scores which are not fully rescued by provision of a
447 variant database.

448 Venn diagram of Lacer- and GATK-recalibrated final SNP calls for (C) Tibetan macaque
449 chr1 and (D) common marmoset exome data, as identified by LoFreq. Total numbers of SNP
450 calls (black) in each category and their associated Ti/Tv ratios (red) are shown. SNP calls
451 unique to Lacer recalibration have an aggregate Ti/Tv closer to the intersection than the SNP
452 calls unique to GATK recalibration for both data sets.

453 **Table Legends**

454 **Table 1:** Impact of base quality score recalibration by Lacer and GATK on SNP calls from the NA12878 chr1 (GATK HaplotypeCaller; filtered
 455 by VQSR), Tibetan macaque chr1 (LoFreq), and common marmoset exome (LoFreq) data sets.

456

Call Set	NA12878 chr1				Tibetan macaque chr1				Common marmoset exome			
	No. of SNPs		Ti/Tv		No. of SNPs		Ti/Tv		No. of SNPs		Ti/Tv	
	Lacer	GATK	Lacer	GATK	Lacer	GATK	Lacer	GATK	Lacer	GATK	Lacer	GATK
Raw reads, all calls	305,220	305,220	2.00	2.00	1,463,349	1,463,349	1.62	1.62	7,089,824	7,089,824	2.06	2.06
Unique to raw read calls	2,902	12,859	0.70	0.86	29,771	34,390	0.57	0.66	20,276	69,222	1.05	1.36
Unique to +recal/+LRA	2,187	3,173	1.01	1.34	40,328	66,202	0.68	0.65	62,459	229,667	1.17	1.17
+recal/+LRA, all calls	304,505	295,534	2.01	2.07	1,473,906	1,495,161	1.62	1.59	7,132,007	7,250,269	2.05	2.03
Filtered	57,303	44,853	1.49	1.66	204,625	212,664	0.60	0.58	4,763,917	4,867,640	1.95	1.92
Final call set	247,202	250,681	2.16	2.16	1,269,281	1,282,497	1.91	1.89	2,368,090	2,382,629	2.27	2.26
Unique to final call set	12,030	15,509	1.84	1.82	7,080	20,296	1.70	0.94	39,237	53,776	1.86	1.59

457 VQSR: variant quality score recalibration; recal: recalibration; LRA: local realignment.

458 **References**

- 459 Cabanski CR, Cavin K, Bizon C, Wilkerson MD, Parker JS, Wilhelmsen KC, Perou
460 CM, Marron JS, Hayes DN. 2012. ReQON: a Bioconductor package for recalibrating quality
461 scores from next-generation sequencing data. *BMC Bioinformatics* **13**: 221.
- 462 Chen SL, Hung CS, Xu J, Reigstad CS, Magrini V, Sabo A, Blasiar D, Bieri T, Meyer
463 RR, Ozersky P, et al. 2006. Identification of genes subject to positive selection in
464 uropathogenic strains of *Escherichia coli*: a comparative genomics approach. *PNAS* **103**,
465 5977-5982.
- 466 Cotton RG, Auerbach AD, Axton M, Barash CI, Berkovic SF, Brookes AJ, Burn J, Cutting
467 G, den Dunnen JT, Flicek P, et al. 2008. The Human Variome Project. *Science* **322**, 861-862.
- 468 DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del
469 Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and
470 genotyping using next-generation DNA sequencing data. *Nat Genet* **43**, 491-498.
- 471 Fan Z, Zhao G, Li P, Osada N, Xing J, Yi Y, Du L, Silva P, Wang H, Sakate R, et al. 2014.
472 Whole-genome sequencing of Tibetan macaque (*Macaca thibetana*) provides new insight
473 into the macaque evolutionary history. *Mol Biol Evol* **31**, 1475-1489.
- 474 Kelley DR, Schatz MC, Salzberg SL. 2010. Quake: quality-aware detection and correction of
475 sequencing errors. *Genome Biol* **11**, R116.
- 476 Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler
477 transform. *Bioinformatics* **25**:1754-1760.
- 478 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin
479 R; 1000 Genome Project Data Processing Subgroup. 2009a. The Sequence Alignment/Map
480 format and SAMtools. *Bioinformatics* **15**, 2078-2079.
- 481 Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, Wang J. 2009b. SNP detection for
482 massively parallel whole-genome resequencing. *Genome Res* **19**, 1124-1132.

483 Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001.
484 dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**, 308-311.

485 The NIH HMP Working Group. 2009. The NIH Human Microbiome Project. *Genome Res*
486 **19**, 2317-2323.

487 The 1000 Genomes Project Consortium. 2010. *Nature* **467**, 1061-1073.

488 The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from
489 1,092 human genomes. *Nature* **491**, 56-65.

490 The Cancer Genome Atlas Network. 2012. Comprehensive molecular characterization of
491 human colon and rectal cancer. *Nature* **487**, 330-337.

492 The Marmoset Genome Sequencing and Analysis Consortium. 2014. The commonest
493 marmoset genome provides insight into primate biology and evolution. *Nat Genet* **46**, 850-
494 857.

495 Thorisson GA, Smith AV, Krishnan L, Stein LD. 2005. The International HapMap Project
496 Web site. *Genome Res* **15**, 1592-1593.

497 Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A,
498 Jordan T, Shakir K, Roazen D, Thibault J, et al. 2013. From FastQ data to high-confidence
499 variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc in Bioinform*
500 **43**, 11.10.1-11.10.33.

501 Wilm A, Aw PP, Bertrand D, Yeo GH, Ong SH, Wong CH, Khor CC, Petric R, Hibberd
502 ML, Nagarajan N. 2012. LoFreq: a sequence-quality aware, ultra-sensitive variant caller for
503 uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic*
504 *Acids Res* **40**, 11189-11201.

505 Zook JM, Samarov D, McDaniel J, Sen SK, Salit M. 2012. Synthetic spike-in standards
506 improve run-specific systematic error analysis for DNA and RNA sequencing. *PLoS ONE* **7**,
507 e41356.

508 Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, Salit M. 2014.
509 Integrating human sequence data sets provides a resource of benchmark SNP and indel
510 genotype calls. *Nat Biotechnol* **32**, 246-251.

511 **Supplemental Information**

512 **Supplemental Figure S1: Lacer is insensitive to a variant database and an imperfect**

513 **reference sequence.** (A) Quality score distributions for error bases in the UTI89 data set

514 (using UTI89 as a reference) as predicted by Quake on the unmapped data (black) or by

515 Lacer (red) and GATK (green) on the mapped data. When a perfect reference sequence is

516 used (UTI89 resequencing data mapped to the UTI89 genome), the quality distribution for

517 error bases is concordant among Lacer, GATK, and Quake. (B) Histogram of deviations of

518 covariate quality scores for each data set indicated from the covariate quality scores derived

519 from the GATK “gold standard” recalibration of the UTI89 data set using the UTI89 genome

520 as the reference. For each covariate (context and cycle number), the difference in recalibrated

521 quality between GATK (using the UTI89 genome as a reference) and a comparison

522 recalibration was calculated. A histogram of all of these values is plotted here for GATK

523 (using the MG1655 genome as a reference and w/ VCF; blue), Lacer (using the MG1655

524 genome as a reference; red), and Lacer (using the UTI89 genome as a reference; black) as the

525 comparison recalibrations. Lacer is unaffected by changing the reference sequence, and the

526 histogram of deviations is centered around zero, indicating no systematic bias in covariate

527 recalibration. GATK using the MG1655 genome as a reference produces systematically lower

528 recalibrated covariate quality scores, seen as a bias towards a tail of positive values. (C)

529 Quality score distributions for error bases in the UTI89 data set (using MG1655 as a

530 reference) as predicted by Quake on the unmapped data (black) or by Lacer (red) and GATK

531 (w/ VCF; green) on the mapped data. Quake and Lacer produce concordant quality

532 distributions for error bases. GATK produces a quality distribution that is biased towards

533 high quality bases. **(D)** Lacer recalibration of the UTI89 data set using MG1655 (w/ or w/o
534 VCF) as a reference. GATK recalibration of the UTI89 data set using UTI89 as a reference
535 (w/o VCF) is shown for comparison. Provision of a variant database has minimal effect on
536 the Lacer recalibration. **(E)** Lacer recalibration of the NA12878 chr1 data set (downsampled
537 by region to 1Mbp) with and without excluding known variant positions. The provision of a
538 database of known variants has little effect on Lacer recalibration. **(F)** Quality score
539 distributions for error bases predicted by Quake on the unmapped NA12878 chr1 data set
540 (black) or by Lacer (red) and GATK (green) recalibration on the mapped data. Quake and
541 Lacer produce concordant quality distributions for error bases. GATK produces a quality
542 distribution that is biased towards high quality bases. **(G)** Quality score distributions for error
543 (blue) and correct (yellow) defined by GATK recalibration of the NA12878 chr1 data set.
544 The quality score distribution for error bases is enriched for high quality scores and
545 somewhat resembles the distribution for correct bases.

546 **Supplemental Figure S2: Validation of Lacer recalibration.** **(A)** Fraction of unsupported
547 “single” non-reference, non-VCF bases with quality scores ≥ 30 . The fraction of single bases
548 for UTI89 mapped to the UTI89 genome (circles) as a reference is ~ 0.9 . The fraction of
549 single bases for the same UTI89 data set mapped to the MG1655 genome (triangles) is much
550 lower, ~ 0.2 . The fraction of single bases for NA12878 chr1 mapped to GRCh37.p13
551 (squares) is also low, ~ 0.1 . **(B)** Histograms of coverage depth at all non-reference, non-
552 VCF bases for the UTI89 (red) and NA12878 chr1 (green) data sets. **(C)** Correction of GATK
553 recalibration using supporting base analysis for GATK recalibration on the UTI89 data set
554 (UTI89 (left) and MG1655 (right) used as the reference sequence). Dark gray bars represent
555 the difference between Lacer and GATK recalibration for each quality score. Light gray bars
556 represent the quality difference from the GATK recalibration assuming only single
557 (unsupported) non-reference, non-VCF bases are true errors. **(D)** Difference in VQS between

558 the Lacer-recalibrated and uncalibrated call sets for all high quality (PASS; green) or low
559 quality (LowQual; red) SNPs in the GATK-recalibrated call set for NA12878 chr1. On high
560 quality SNPs, Lacer recalibration results in higher VQS (positive quality difference), while
561 on low quality SNPs, the quality difference is centered about zero. **(E)** Difference in VQS
562 between the GATK-recalibrated and uncalibrated call sets for all high quality (PASS; green)
563 or low quality (LowQual; red) SNPs in the GATK-recalibrated call set. GATK recalibrated
564 data results in generally lower VQS for all SNPs compared with uncalibrated data, but the
565 decrease is greater for high quality SNPs than for low quality SNPs, contrary to expectation.
566 **(F)** Distribution of overall base quality scores for the uncalibrated (black), Lacer-recalibrated
567 (red), and GATK-recalibrated (green) NA12878 chr1 data. GATK recalibration results in a
568 bimodal distribution that is not seen in the uncalibrated or Lacer-recalibrated data.

569 **Supplemental Figure S3: Recalibration of NA12878 chr19.** **(A)** Recalibration of the
570 NA12878 chr19 sequencing data using GATK, with or without excluding known variants
571 based on dbSNP, Mills and 1000 Genomes gold standard indels, and 1000 Genomes Phase 1
572 indels (w/ or w/o VCF, respectively), or using Lacer, w/o VCF or with dbSNP variant
573 positions only (VCF positions only). GATK recalibration produces lower quality scores that
574 are only partially increased by exclusion of known variants. Lacer recalibration using only
575 known variant positions results in relatively similar recalibration to Lacer using the entire
576 data set. **(B)** Distribution of quality scores for error bases predicted by Quake (black) on the
577 unmapped data or by Lacer (red) and GATK (green) on the mapped data. Quake and Lacer
578 produce concordant quality distributions for error bases. GATK produces a quality
579 distribution that is biased towards high quality bases. **(C)** Distribution of overall base quality
580 scores for the uncalibrated (black), Lacer-recalibrated (red), and GATK-recalibrated (green)
581 data. None of the data sets have a bimodal distribution of base quality scores for chr19. **(D)**
582 Venn diagram of Lacer- and GATK-recalibrated SNP calls identified by GATK

583 HaplotypeCaller on chr19 that passed recalibration by the GATK VariantRecalibrator and
584 that also matched the high-confidence NIST call set. Total numbers of SNP calls in each
585 category and their associated Ti/Tv ratios are shown. SNP calls unique to Lacer recalibration
586 have an aggregate Ti/Tv closer to the intersection than the SNP calls unique to GATK
587 recalibration. **(E)** Comparison of VQS for SNPs identified in the uncalibrated, Lacer-
588 recalibrated, or GATK-recalibrated data which matched the NIST call set. Lacer recalibrated
589 data results in overall higher VQS than that from uncalibrated (red) and GATK recalibrated
590 (green) data on these high confidence true positive SNPs. GATK recalibrated data results in
591 overall lower VQS than uncalibrated (gray) data on these same SNPs. **(F)** Difference in VQS
592 between the Lacer-recalibrated and GATK-recalibrated call sets for all high quality (PASS;
593 green) or low quality (LowQual; red) SNPs in the GATK-recalibrated call set. Lacer
594 recalibrated data results in a mild increase in VQS on low quality SNPs (as called by GATK
595 recalibrated data), but this is small compared to the increase in VQS for high quality SNPs.
596 **(G)** Difference in VQS between the GATK-recalibrated and uncalibrated call sets for all high
597 quality (PASS; green) or low quality (LowQual; red) SNPs in the GATK-recalibrated call set.
598 GATK recalibrated data results in generally lower VQS for all SNPs compared with
599 uncalibrated data, but the decrease is greater for high quality SNPs than for low quality SNPs,
600 contrary to expectation. **(H)** Difference in VQS between the Lacer-recalibrated and
601 uncalibrated call sets for all high quality (PASS; green) or low quality (LowQual; red) SNPs
602 in the GATK-recalibrated call set. On high quality SNPs, Lacer recalibration results in higher
603 VQS (positive quality difference), while on low quality SNPs, the quality difference is
604 centered about zero.

605 **Supplemental Figure S4: Recalibration of non-human primate sequencing data.** Quality
606 score distributions for error (red) and correct (green) defined by: **(A)** GATK recalibration
607 (excluding known variants) or **(B)** Lacer recalibration of the Tibetan macaque chr1

608 sequencing data, and (C) GATK recalibration (excluding known variants) or (D) Lacer
609 recalibration of the common marmoset exome sequencing data. For both data sets, the quality
610 score distributions for error bases defined by GATK are enriched for high quality scores and
611 somewhat resemble the distributions for correct bases. The quality profiles for Lacer were
612 similar to those observed in previous data sets (c.f. **Fig. 1c**).

613 **Supplemental Figure S5: Recalibration of different NGS data sets using Lacer.**

614 Empirical (recalibrated) quality scores plotted against reported (uncalibrated) quality scores
615 for: (A) NA12878 chr1 exome (NCBI36 reference) sequencing data (Illumina GAIIx).
616 Recalibrated quality scores are shown for each read group individually; (B) SRS017191
617 human metagenome (*Bacteriodes vulgatus* ATCC 8482 reference) sequencing data (Illumina
618 GAIIx); (C) TruSeq Amplicon Cancer Panel (NCBI36 reference) targeted sequencing data
619 (Illumina MiSeq); (D) UTI89 (UTI89 or MG1655 reference) sequencing data (Ion Torrent);
620 (E) MG1655 (MG1655 or UTI89 reference) sequencing data (Roche 454). Recalibration by
621 Lacer (w/o VCF) and GATK (w/ or w/o VCF) is shown. In the absence of a perfect reference,
622 GATK recalibration produces lower quality scores compared with Lacer recalibration; these
623 lower quality scores are only partially increased by exclusion of known variants.
624 Furthermore, Lacer recalibration (using a perfect or imperfect reference) matches GATK's
625 recalibration when a perfect reference is provided. Lacer is therefore more accurate across a
626 wide range of NGS data sets without the requirement of a variant database.

627

628 **Supplemental Table S1:** Impact of base quality score recalibration by Lacer and GATK on

629 SNP calls from NA12878 chr19 (GATK HaplotypeCaller).

630

Fig. 1

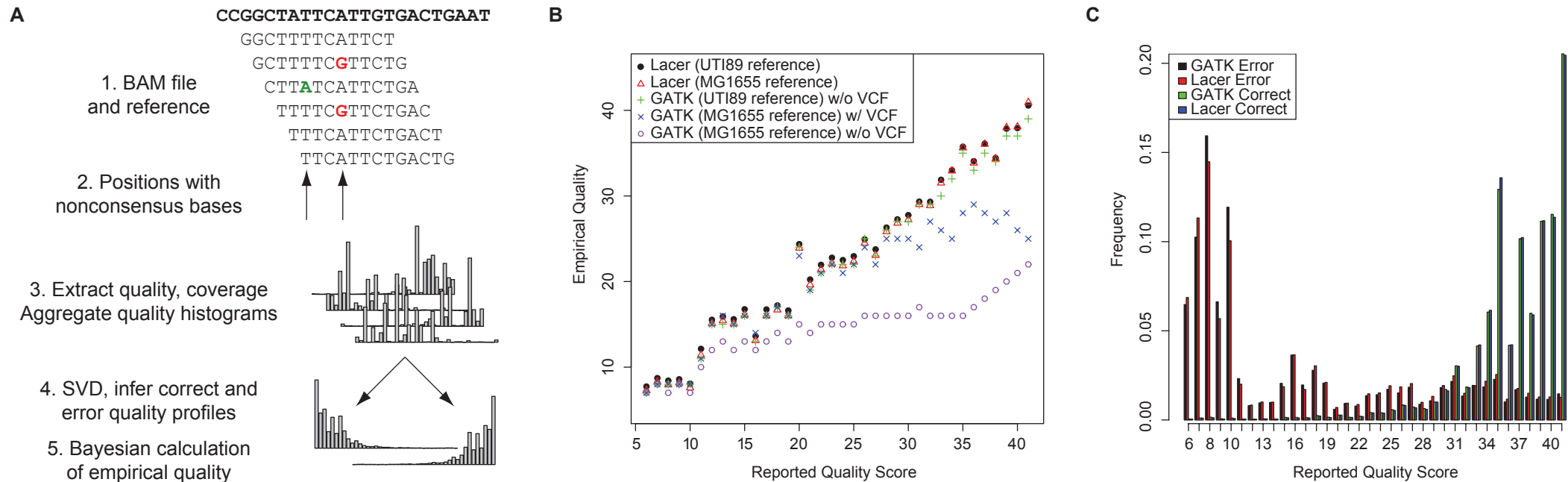
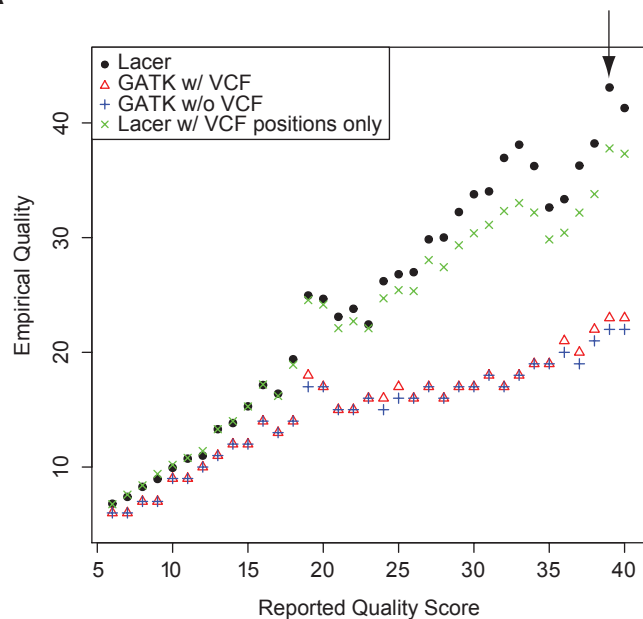
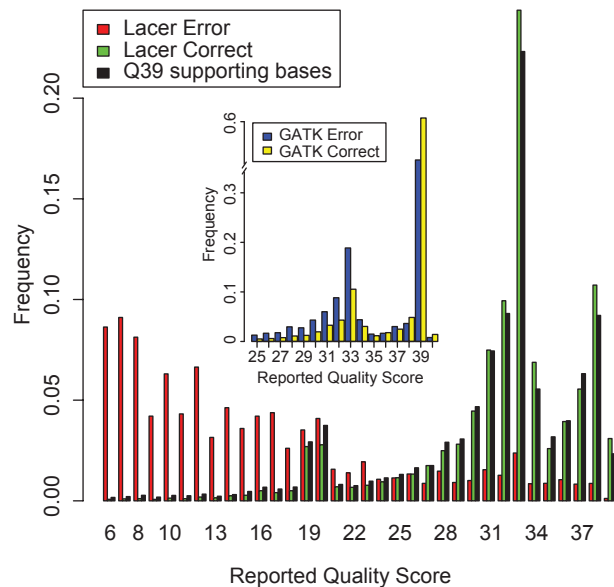


Fig. 2

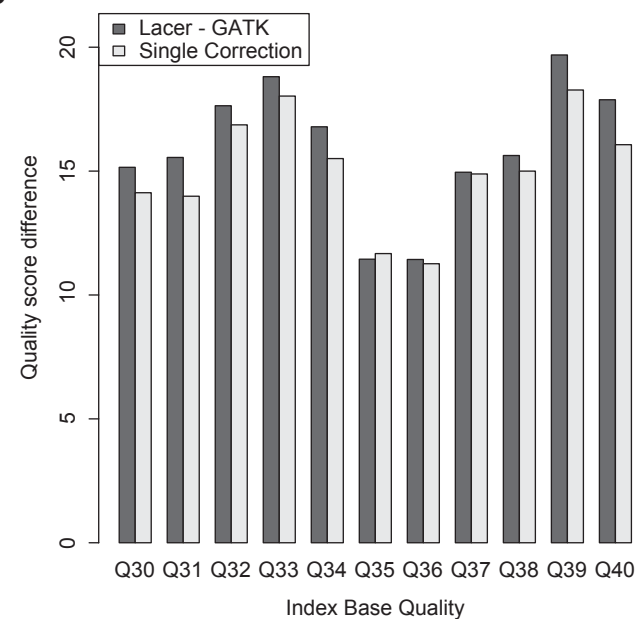
A



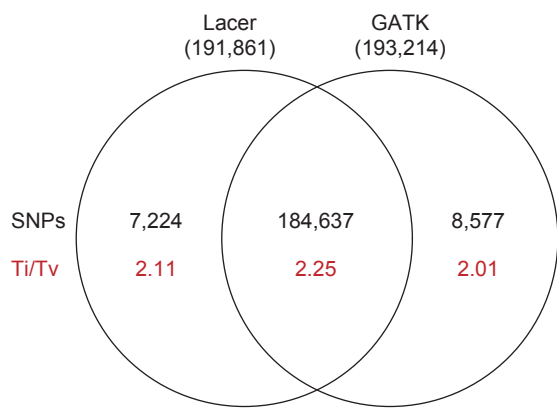
B



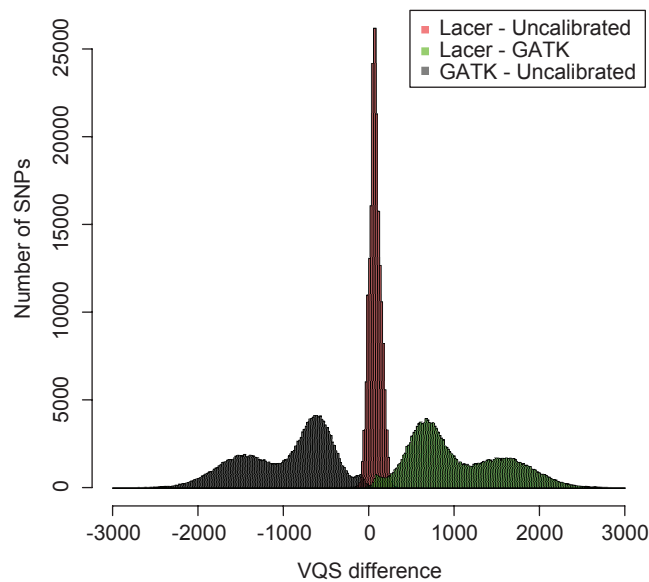
C



D



E



F

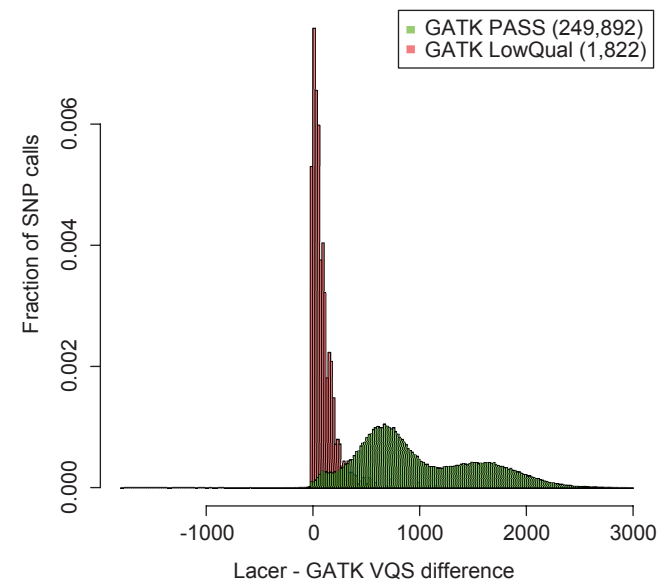


Fig. 3 bioRxiv preprint doi: <https://doi.org/10.1101/130732>; this version posted April 27, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

