Pyrobayes: an improved base caller for SNP discovery in pyrosequences

Aaron R Quinlan, Donald A Stewart, Michael P Strömberg & Gábor T Marth

Previously reported applications of the 454 Life Sciences pyrosequencing technology have relied on deep sequence coverage for accurate polymorphism discovery because of frequent insertion and deletion sequence errors. Here we report a new base calling program, Pyrobayes, for pyrosequencing reads. Pyrobayes permits accurate single-nucleotide polymorphism (SNP) calling in resequencing applications, even in shallow read coverage, primarily because it produces more confident base calls than the native base calling program.

The sequencing reads produced by the 454 Life Sciences pyrosequencers are the result of cyclical nucleotide tests in which ideally all nucleotides within a homopolymer (for example, AAA) are incorporated in a single test, and the light intensity signal observed in each cycle is proportional to the actual number of incorporated nucleotides[1]. In reality, the signal for a fixed number of incorporated bases varies substantially, and there is usually a nonzero signal even when no base is incorporated (**Supplementary Fig. 1a** online). This makes accurate base calling difficult and leads to nucleotide over-calls and under-calls that manifest as insertion and deletion errors[2–4]. Such errors often lead to misalignments that artificially inflate sequencing error estimates and cause the assignment of lower estimates of the base calls' accuracy (which we refer to as base quality) than warranted by their true accuracy (**Fig. 1**).

Accurate base qualities are crucial for resequencing applications in which true allelic variation must be distinguished from sequencing error. Reliable SNP calls can only be made if the base error rate for the called allele is substantially lower than the expected polymorphism rate. For example, in human studies for which the average pairwise polymorphism rate is on the order of 1 in 1,000 bp, no SNP call should be made from a single allele with a base quality lower than 30 (1 in 1,000 bp error rate). However, if most base calls in resequencing reads are well above such a threshold, SNPs can be detected with high confidence even in single-read coverage. Unfortunately, we found that the majority of the base qualities assigned by the native 454 base caller (version 1.0.52) were not sufficiently high for SNP calling in low-coverage conditions, as only 24% of the native 454 base calls were above 30 (**Fig. 2a**). However, we found that 454 reads can be called accurately, but the base qualities assigned by the native base caller underestimate the actual base accuracy (**Fig. 2b**). We developed a new base calling program, Pyrobayes, to produce more accurate (higher) base qualities and hence make more high-quality base calls in 454 pyrosequences.

Our base caller first determines the most likely number of incorporated bases from the measured incorporation signal for each nucleotide test. Our Bayesian strategy (**Supplementary Methods** and **Supplementary Fig. 1** online) requires 'data likelihoods', that is, the distribution of observed nucleotide incorporation signals for every possible homopolymer length. We estimated

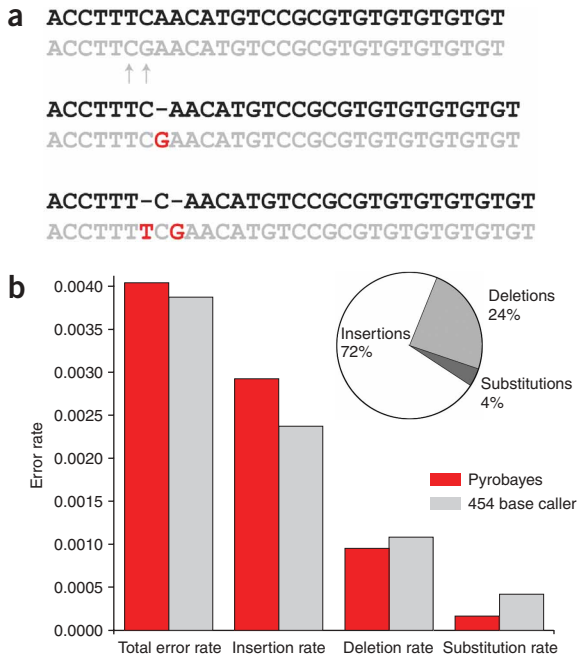

Figure 1 | Comparison of the error profiles of Pyrobayes and the native 454 base caller. (**a**) Illustration of the effects of calling too few or too many bases on the alignment of a read (gray) to the reference sequence (black). Top, too few thymines were called, resulting in two spurious mismatches (arrows) by misaligning the correctly called cytosine and the inserted guanine in the 454 read. Middle, the correct number of thymines was called, resulting in the correct read alignment of the single insertion error (red) in the 454 read. Bottom, too many thymines were called, resulting in the correct read alignment of the two base insertion errors (red) in the 454 read. (**b**) Base error rates for Pyrobayes and the native 454 base caller. The relative contribution of each error type based on Pyrobayes calls is shown in the pie chart.

Department of Biology, Boston College, 140 Commonwealth Avenue, Chestnut Hill, Massachusetts 02467, USA. Correspondence should be addressed to G.T.M. (marth@bc.edu).

RECEIVED 3 OCTOBER 2007; ACCEPTED 4 DECEMBER 2007; PUBLISHED ONLINE 13 JANUARY 2008; DOI:10.1038/NMETH.1172

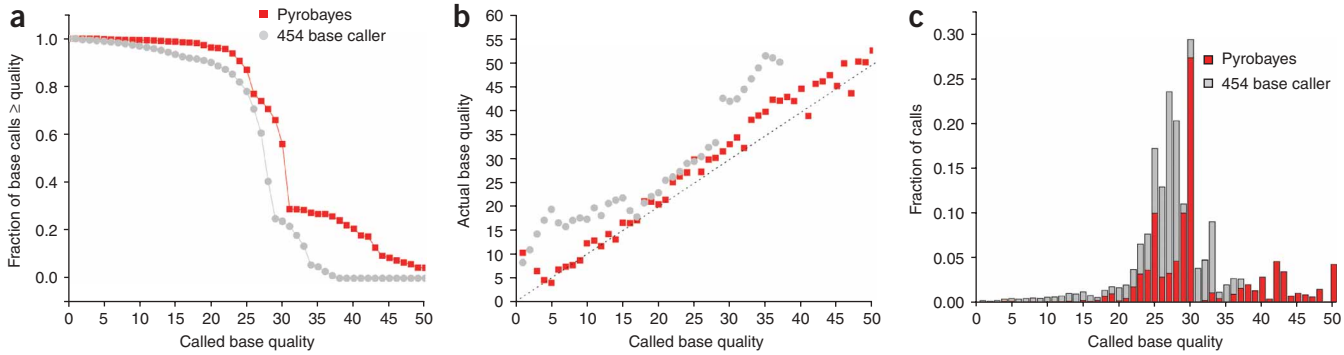

Figure 2 | Comparison of the base qualities assigned by Pyrobayes and the native 454 base caller. (**a**) The cumulative distribution of base qualities assigned by each program. (**b**) Comparison between assigned base quality and the base quality calculated from measured base accuracy. A value of 50 was assigned when no errors were found. (**c**) The distribution of base calls according to base quality.

these by collecting shotgun resequencing data with the 454 Life Sciences GS20 instrument from a finished mouse bacterial artificial chromosome (BAC) clone and extrapolating to higher homopolymer lengths for which few or no examples could be found (**Supplementary Fig. 1a** and **Supplementary Fig. 2** online). For 'prior probabilities', we used the relative frequency of homopolymer lengths tabulated from several different reference genome sequences. We found that these frequencies were consistently different from the theoretical expectation that they are proportional to $1/4^n$, where n is the homopolymer length (**Supplementary Fig. 1b**). In the software we used a single distribution because the frequencies are very similar across all eukaryotic genomes we considered. Using data likelihoods and prior distributions, we determined the 'Bayesian posterior probability' of the correct number of bases given the measured incorporation signal (**Supplementary Fig. 1c**). The called base sequence was produced by concatenating the most likely number of bases for every consecutive incorporation test. The base quality assigned to each base is the probability that the base in question is not an over-call. We found it also useful to call one extra base, as long as the presence of that base is above a minimum probability (see below).

We compared the Pyrobayes and native base calling accuracy in 299,654 reads from the inbred reference (*iso-1*) strain of *Drosophila melanogaster* (**Supplementary Methods**). The overall base accuracy (**Fig. 1b**) was quite high for both Pyrobayes and the native base caller (99.60% versus 99.61%). Notably, 96% of all sequencing errors were insertions or deletions. The Pyrobayes insertion error rate was higher (0.29% versus 0.24%), but its deletion rate was lower (0.09% versus 0.10%). Most importantly for SNP discovery, the Pyrobayes substitution error rate was 60% lower (0.017% versus 0.042%) than that of the native base caller. A large fraction (74%) of the base calling errors was shared between the two methods. Characteristically, 86% of the errors solely made by Pyrobayes were insertions whereas 82% of the unique 454 base caller errors were deletions or substitutions (**Supplementary Fig. 3**). The Pyrobayes base qualities corresponded substantially better to the actual base accuracy than the native base qualities (**Fig. 2b**), and therefore our base qualities were typically higher (**Fig. 2c**). For example, 56% of the Pyrobayes base calls were assigned base qualities of 30 or higher, as compared to 24% of the native base calls (**Fig. 2a,c**). Additionally, Pyrobayes produced base qualities up to 50, whereas the highest native base quality was 38.

We investigated the effect of our higher overall base qualities on SNP detection. First, we searched for single-base-pair differences between the 454-sequenced *iso-1* reads and the *iso-1* reference sequence. We expected few true polymorphisms as these sequences were from the same inbred *D. melanogaster* strain, and the overall accuracy of the *D. melanogaster* genome reference sequence is very high. Therefore, SNPs discovered in this comparison estimate the false positive SNP rate. This rate was 1.22/10,000 bp using the native base calls, but only 0.97/10,000 bp using the Pyrobayes base calls. It is important to consider that the false SNP discovery rate depends on the polymorphism rate in the resequenced organism. For example, in *D. melanogaster*, where the pairwise polymorphism rate is ~1/200 bp (ref. 5), our results corresponded to a false SNP discovery rate of 1.9%.

To estimate SNP calling error rates directly, we also sequenced an inbred *D. melanogaster* isolate from Malawi with a single 454 run. In the alignments of the 454 reads base called with Pyrobayes we found 1,118 SNP candidates at or above the Polybayes SNP probability[6] cutoff value of 0.7. The validation rate for these candidates was 93% (1,036 of 1,118). The corresponding 7% false positive SNP rate observed in this experiment is a composite effect of false SNP calls, emulsion PCR errors before 454 sequencing and the usual artifacts associated with capillary sequence validation experiments[7]. We also estimated that we missed 14.8% of the SNPs (**Supplementary Methods**). We repeated the SNP discovery experiment in the alignments processed with the native 454 base caller: the false positive rates were identical, but twice as many (30.0%) SNPs were missed.

The primary cause of spurious substitution errors in 454 reads is the erroneous alignment of a base under-call followed by an over-call (or vice versa) as a base substitution (**Fig. 1a** and **Supplementary Fig. 3d**). Our alignment algorithm, Mosaik (**Supplementary Methods**), uses gap penalties that properly align reads in such situations. Additionally, we found that calling more bases in homopolymer runs often also improves the alignment (**Fig. 1a**). Eliminating spurious base errors resulting from alignment artifacts leads to assignment of higher base qualities. Higher base qualities increase SNP calling sensitivity.

The cost of tending toward calling more bases in homopolymer runs is a slightly increased insertion rate (**Fig. 1b**) even though the extra called bases are typically assigned very low base qualities. This is a logical choice for SNP discovery applications. However, it is not

yet clear what effect such extra called bases will have for *de novo* sequence assembly of 454 reads.

A natural, although undesirable consequence of having to determine homopolymer length from a single incorporation signal is that the likelihood of over-calling error increases with every consecutive nucleotide. Accordingly, the first called base in a homopolymer run is assigned the highest base quality, and the last called base, the lowest (**Supplementary Fig. 4a** online). This introduces an unintended directionality for the base qualities in the sequence alignment (**Supplementary Fig. 4b**). Clearly, it is not possible for the base calling program to resolve this ambiguity within the standard base quality framework defined by the Phred[8,9] base calling program. Consequently, one must rely on alignment and SNP calling software to account for this phenomenon.

We also evaluated base calling accuracy on the new 454 Life Sciences FLX sequencing machine model using two sequencing runs from the K12 strain of *Escherichia coli* and found that both base callers underestimate the FLX base accuracy (**Supplementary Fig. 5** online). The primary reason for this is that the overall error rate of the FLX machine (0.12%) was much lower than that of the GS20 (0.40%). Although the fact that the Pyrobayes base qualities were much closer to the actual accuracy suggests that our calibration procedure is robust, there is clearly a need to recalibrate our method for the FLX and future models.

The increased accuracy of our base qualities will likely permit more sensitive biological studies using the 454 machines. Although our data only illustrate this directly for low-coverage, survey-type applications, statistical fluctuations[10] will result in regions of shallow read depth even in deeper nominal coverage. The ability to call SNPs in such regions without a substantial loss of accuracy will permit more complete analyses of whole-genome alignments. Pyrobayes can process a single sequencing run in under 2 min. Pyrobayes and Mosaik are freely available for nonprofit use at http://bioinformatics.bc.edu/marthlab/Software.

Note: Supplementary information is available on the Nature Methods website.

ACKNOWLEDGMENTS

This work was supported by a grant from the US National Human Genome Research Institute (R01 HG003698) to G.T.M. We thank E. Mardis and the 454 production group at the Washington University Genome Sequencing Center for generating the sequence data used in this work, and A. Clark at Cornell University for providing access to the *D. melanogaster* reads.

AUTHOR CONTRIBUTIONS

A.R.Q., software and algorithm development and data analysis; D.A.S., data fitting and parameter estimation for Bayesian data likelihoods; M.P.S., alignment algorithm development. A.R.Q. and G.T.M. designed the experiment and wrote the manuscript.

Published online at http://www.nature.com/naturemethods/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Margulies, M. *et al. Nature* **437**, 376–380 (2005).
2. Girard, A., Sachidanandam, R., Hannon, G.J. & Carmell, M.A. *Nature* **442**, 199–202 (2006).
3. Thomas, R.K. *et al. Nat. Med.* **12**, 852–855 (2006).
4. Velicer, G.J. *et al. Proc. Natl. Acad. Sci. USA* **103**, 8107–8112 (2006).
5. Hoskins, R.A. *et al. Genome Res.* **11**, 1100–1113 (2001).
6. Marth, G.T. *et al. Nat. Genet.* **23**, 452–456 (1999).
7. Quinlan, A.R. & Marth, G.T. *Nat. Methods* **4**, 192 (2007).
8. Ewing, B. & Green, P. *Genome Res.* **8**, 186–194 (1998).
9. Ewing, B., Hillier, L., Wendl, M.C. & Green, P. *Genome Res.* **8**, 175–185 (1998).
10. Lander, E.S. & Waterman, M.S. *Genomics* **2**, 231–239 (1988).