

Genomic Services

SEESNP[®] PROPRIETARY ANALYSIS ADVANTAGES

Kevin McKernan, Jim Yang, Betty Woolf, Rey Sequerra, Kathy Makowski, Tom Tang
Agencourt Bioscience Corporation, A Beckman Coulter Company, Beverly, MA 01915, USA

Abstract

Understanding genetic variation and its relationship to drug response and disease susceptibility is becoming an increasingly important consideration in drug design and clinical trials. Since the completion of the Human Genome Project, the majority of human genes can be readily resequenced in diseased or drug-responsive populations. Typical studies involve the resequencing of 20–500 amplicons in 50 or more patients, generating thousands of sequence traces that need to be analyzed. To efficiently automate the analysis of such resequencing data, Agencourt Bioscience Corporation has developed novel algorithms that enhance the accuracy of automated procedures and maintain the sensitivity that is required for the diverse samples that are often present in cancer biopsies. In this article, we will discuss the common problems encountered in SNP resequencing and the innovative solutions developed at Agencourt.

Introduction

PolyPhred is currently the most popular software tool utilized for analyzing DNA sequencing reads for SNPs (1,2,3). Other programs such as Mutation Surveyor, NovoSNP (4), and Paracel have certain beneficial features that PolyPhred lacks. However, unlike PolyPhred, they do not have a Linux-compatible analysis program that can automatically interface with a database or sequencing pipeline, allowing the automated assembly of thousands of reads.

Since PolyPhred is the most scalable software tool available, Agencourt opted to design algorithms to enhance and correct PolyPhred's common errors. Many of the current criticisms of the software are related to its high false positive rate, which seems directly tied to its high false negative rate. In other words, any attempt to reduce the false positive rates with the current software has resulted in a comparable increase in the false negative rate.

After reviewing hundreds of amplicons, we were ultimately able to classify the majority of PolyPhred analysis errors into four general categories:

- A. G after A TaqFS incorporation bias.
- B. Low quality sequence at the ends of amplicons.
- C. Over-estimation of base accuracy after enzyme slippage due to homopolymer stretches.
- D. Insertions and deletions.

The remainder of this Application Note describes how we have addressed the first three problems outlined above.

Methods

A. Enzymatic Incorporation Bias

TaqFS has four characterized incorporation biases with fluorescent DyeTerminator sequencing nucleotides: G incorporation after A is significantly reduced, C incorporation after one or more T's is slightly enhanced, and A or T incorporation after G is slightly enhanced (5). These four incorporation biases are consistently reproducible and create high peak to peak variance in a DNA sequencing read.

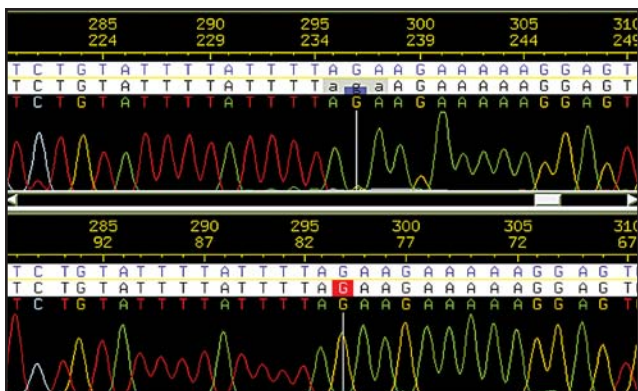


Figure 1. Forward read (upper) has G incorporation significantly reduced when incorporated after an A. The reverse read (lower) is presented with a C after a T (G after A when reverse complemented) that incorporates readily.

The reduced G after A is the most dramatic and problematic of the four incorporation biases. It is demonstrated in Figure 1, where the forward read G signal nearly disappears. The reverse read captures the mistake, emphasising the value of the bidirectional sequencing utilized at Agencourt. Nonetheless, this bias is still problematic when the artifact is present in a heterozygote base-call.

When forward and reverse reads generate different genotypes, PolyPhred resolves this discordance by selecting the genotype that is generated from the higher quality read. This function of PolyPhred creates false negatives because homozygote bases generate higher quality base-calls than heterozygote base-calls (Figure 2).

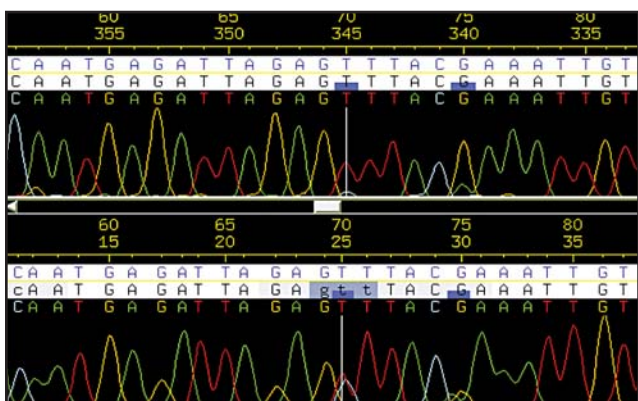


Figure 2. Reverse read (upper) demonstrates reduced G after A incorporation (C after T when reverse complemented). The forward read (lower) has no incorporation bias and as a result shows clean heterozygote base. Phred penalized the quality of this base as seen by the gray quality shading above the read (white is high quality, gray is low quality).

To address this issue, we have incorporated a proprietary algorithm to automatically screen the .poly files for all areas likely to be heterozygous and survey the areas for potential G after A artifacts.

If certain criteria are met, the algorithm inflates the quality scores of the heterozygote PHD files such that PolyPhred's source function will choose the heterozygote base over the false homozygote, thus reducing the false negatives.

B. Low Quality Sequence Coverage at the Ends of Amplicons

Sanger sequencing produces electrophoretic mobility artifacts in the first 80 bases of sequence. As a result, bidirectional sequences of 500 bp amplicons contain regions on the ends of the amplicons with a high quality read in one direction and a read with mobility artifacts in the other direction. The end-regions containing mobility artifacts will produce false positives under the PolyPhred parameters selected to maximize SNP discovery in the middle of the amplicon, where there is a high quality double-stranded sequence.

To optimize PolyPhred for the ends of the amplicons, we have implemented Read Pair Discordance (RPD) analysis. This requires Polyphred to be run multiple times with different score, rank and source settings to identify regions of the amplicon that produce discordant genotypes in their respective forward and reverse read directions. The ends of the amplicons tend to have a higher frequency of these discordant genotypes. The algorithm resolves these discordant genotypes using several sequence context rules described in sections A and C. As a result, a significant number of false positives can be resolved.

C. Base Accuracy Over-Estimation with Phred

Homopolymer stretches can cause problems with sequence accuracy due to polymerase slippage (6). In these situations, the sequence 3' to the homopolymer has falsely inflated quality values and cannot be trusted. The sequence chromatographs after the homopolymer generate peak structures that phred evaluates as high quality. This inflates PolyPhred's false positive rate (Figure 3).

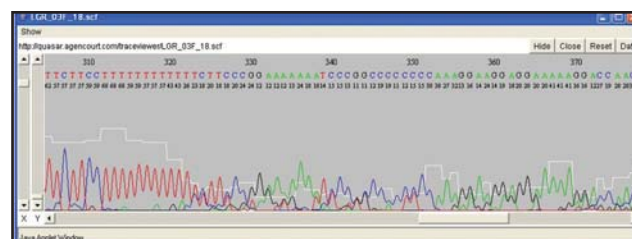


Figure 3. Quality scores after a homopolymer stretch are frequently above a Q20 but should not be trusted for SNPs.

Using an approach similar to the one described to correct G after A artifacts, we can also search the .poly files for homo-polymer stretches and modify the quality score. The algorithm sets the quality to zero for bases 3' to any homopolymer seven bases or longer. This prevents PolyPhred from making any false positive calls in these regions.

Results

Implementing the algorithms described above has dramatically improved our automated SNP-calling accuracy. The table below demonstrates the results of employing these modifications. The results are reported for the entire amplicon (all) or just the exonic regions of the amplicons (typically the higher quality, middle portion of the amplicons). The use of PolyPhred alone produces unacceptable results. However, when PolyPhred is used with the Agencourt algorithms, the accuracy of the SNP calls is dramatically improved.

PolyPhred Version	Number of Reads	Region Covered	% Accuracy (True Positives/Automated SNP Call)
4.26	2	All	43 ± 17
4.27	2	All	68 ± 34
4.27	2	Exon	81 ± 20
Agencourt	4	All	81 ± 25
Agencourt	4	Exon	96 ± 9

Conclusions and Discussion

To accelerate the data analysis process, we have developed a reviewing interface that integrates with our Oracle database. This interface assists the reviewers by: 1) marking the exonic regions of the amplicons and reporting any SNPs that result in amino acid changes; 2) remapping the amplicon SNP coordinates to the base coordinate on the mRNA; 3) annotating Pfam domains so a user can quickly assess if the SNP is in a critical protein domain; 4) referencing dbSNP to determine if the SNP has been previously reported; 5) displaying the likelihood of heterozygosity; 6) providing SNP frequency in the population; and 7) providing a three-tier manual reviewing and version control system. The optional manual reviewing system records and tracks edits to any SNP calls or score change made by the three individual manual reviewers.

A software program cannot be universally successful if it fails to account for the diverse and relatively frequent spectrum of polymorphisms and context-related sequencing artifacts that occur.

PolyPhred can be an extremely valuable and high-throughput tool for SNP discovery if it is applied in a heuristically-controlled manner to account for such artifacts.

The core component of PolyPhred that calculates the likelihood of heterozygosity by evaluating multi-component signals is well tested, easily automated, and highly reliable. The input quality files can be corrected for known sequencing artifacts, enabling the implementation of a very powerful automated analysis system.

The final frontier for automated SNP discovery is to be able to identify which reads contain heterozygous insertions or deletions, and then to resolve their genotypes. Trace subtraction tools (7) have shown promise addressing this challenge and we are currently investigating these tools to offer our customers in the future.

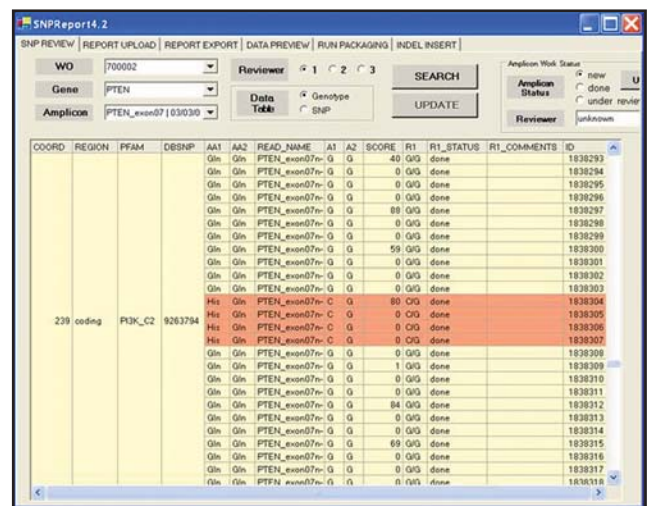


Figure 4. Display of the SeeSNP database viewing package allows easy identification of non-synonymous SNPs and previously discovered SNPs in dbSNP.

References

1. DA Nickerson, VO Tobe, and SL Taylor. "PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing," *Nucleic Acids Research*, July 1997; 25: 2745–2751.
2. B Ewing, L Hillier, M Wendl, and P Green. "Base-calling of automated sequencer traces using *phred*. I. Accuracy assessment," *Genome Research*, Mar 1998; 8: 175–185.
3. B Ewing and P Green. "Base-calling of automated sequencer traces using *phred*. II. Error probabilities," *Genome Research*, Mar 1998; 8: 186–194.

4. S Weckx, J Del-Favero, R Rademakers, L Claes, M Cruts, P De Jonghe, C Van Broeckhoven, and P De Rijk. "novoSNP, a novel computational tool for sequence variation discovery," Genome Research, Mar 2005; 15: 436–442.
5. C Korch and H Drabkin. "Improved DNA sequencing accuracy and detection of heterozygous alleles using manganese citrate and different fluorescent dye terminators," Genome Research, Jun 1999; 9: 588–595.
6. AB Kotlyar, N Borovok, T Molotsky, L Fadeev, and M Gozin. "In vitro synthesis of uniform poly(dG)–poly(dC) by Klenow exo-fragment of polymerase I," Nucleic Acids Research, Jan 2005; 33: 525–535.
7. JK Bonfield, C Rada, and R Staden. "Automated detection of point mutations using fluorescent sequence trace subtraction," Nucleic Acids Research, Jul 1998; 26: 3404–3409.

All trademarks are the property of their respective owners.



Innovate **Automate**
SIMPLIFY

Agencourt Bioscience Corporation, A Beckman Coulter Company • 800-361-7780 • www.agencourt.com
500 Cummings Center, Suite 2450 • Beverly, Massachusetts 01915

Worldwide Offices:

Australia (61) 2 9844-6000 **Canada** (905) 819-1234 **China** (86) 10 6515 6028 **Eastern Europe, Middle East, North Africa** (41) 22 994 07 07
France 01 49 90 90 00 **Germany** (31) 10 470 79 26 **Hong Kong** (852) 2814 7431/2814 0481 **Italy** 02-953921 **Japan** 03-5404-8359
Mexico (55) 560-57770 **Netherlands** (31) 10 470 79 26 **Singapore** (65) 6339 3633 **South Africa, Sub-Saharan Africa** (27) 11-805-2014/5 **Spain** 91 3836080
Sweden 08-564 85 900 **Switzerland** 0800 850 810 **Taiwan** (886) 2 2378 3456 **Turkey** 90 216 309 1900 **U.K.** 01494 441181 **U.S.A.** 1-978-867-2600