

Automated Algorithms for Variation Detection in Diploid Resequencing Data

Daniel J. Richter¹, Mark J. Daly^{1,2}, David Altshuler^{1,3}, Stacey B. Gabriel¹

Keywords: resequencing, SNP discovery, variation detection

1 Introduction.

Existing software for automated SNP discovery in diploid resequencing data requires manual review to achieve high sensitivity and specificity. Thus, essentially all studies to date have required manual review of trace data, limiting throughput and introducing subjectivity into the analyses. We set out to develop an automated system for SNP discovery by creating a novel algorithm, PolyDhan, and by overlapping the results of our novel method with an existing method, PolyPhred [1].

This poster will focus on the novel algorithm we developed, how we combined its output computationally with PolyPhred to achieve superior overall performance, a summary of current results from our system, and future directions for algorithm development for automated variation detection in resequencing data.

2 Novel algorithm and method.

We developed a novel algorithm, PolyDhan. This algorithm does not depend on base calls; instead, raw sequence traces are aligned to an artificially generated reference trace. Following alignment, sequencing artifacts that meet several characteristic thresholds are detected and flagged. Then, in a process similar to the Neighborhood Quality Standard [2,3] (but not implemented using Phred [4] quality scores), positions within each trace are removed from evaluation if their local sequence neighborhood contains any flagged artifacts. High quality passing positions are then examined for polymorphisms by comparison among traces from different samples. A SNP is declared at a position if one or more individuals are polymorphic at the position, and the traces for those individuals pass appropriate heuristically defined quality thresholds.

The second step in our computational method is to combine the results of PolyPhred and PolyDhan. For each unique position, we tally the total number of SNP calls made by each automated method (a call at a position only counts once, regardless of the predicted number of variants at the position). Multiple calls may be made by the same program at the same position because sequencing is conducted bidirectionally, and because sequencing amplicons are designed in a tiled overlapping manner. Any site with more than one SNP call (either by the same program twice, or by two programs, or any combination thereof) is considered a putative high quality SNP. This method takes advantage of the fact that if the two methods are relatively independent, then the probability of both methods making a false positive call at the same position is relatively low.

3 Evaluation of the novel system.

¹ Broad Institute of MIT and Harvard, One Kendall Square, Building 300, Cambridge, MA, USA.

² Whitehead Institute, 9 Cambridge Center, Cambridge, MA, USA.

³ Harvard Medical School and Massachusetts General Hospital, Boston, MA, USA.

To determine rigorously and objectively the false positive and false negative rates, we evaluated performance on a large resequencing data set (500,000 sequencing traces representing 120 Mb of total sequence: 2.5 Mb in each of 48 individuals) by genotyping all SNPs discovered, as well as those SNPs already in dbSNP, as part of the HapMap ENCODE Project.

We find that both PolyPhred and PolyDhan perform extremely well on the data set, although each individual program misses some true SNPs, and the lower SNP quality categories in PolyPhred have high rates of false positive calls if not manually reviewed. The combination of the two programs, however, is able to produce very high quality results without human review: we detected over 3,000 novel SNPs (in addition to 2,500 already in dbSNP) with a false positive rate of 9.3%, and, in comparison to SNPs from dbSNP confirmed by genotyping, a false negative rate of 5.9%.

All data from the HapMap ENCODE Project are publicly available. Polymorphism and genotyping data are available on the HapMap website, <http://www.hapmap.org>. Sequencing traces are available at the NCBI Trace Archive, <http://www.ncbi.nlm.nih.gov/Traces/trace.cgi>.

4 Future directions for algorithm development.

To our knowledge, no previous evaluation of SNP discovery efforts has integrated SNPs discovered by resequencing with objective verification by genotyping and comparison to dbSNP. Thus, we now have available for the first time a large data set for evaluation of SNP discovery algorithms, and we are provided with an opportunity for continued development and testing.

Going forward, we would like to reduce the dependence of PolyDhan on explicitly determined thresholds, and move towards a more probabilistic system. We will define a set of test statistics to evaluate on trace data (for example, peak height ratios), and, using the existing ENCODE data as a training set, evaluate these test statistics for true SNPs sites and artifactual SNP sites. We can then compute the test statistics on novel data, evaluate them in the context of their distributions from the training data, and therefore produce usable measures of the likelihood that a potential novel SNP position is real.

References

- [2] Altshuler, D., Pollara, V.J., Cowles, C.R., Van Etten, W.J., Baldwin, J., et al. 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* 407:513-516.
- [4] Ewing, B. and Green, P. 1998. Base-calling of automated sequencer traces using *Phred*. II. Error probabilities. *Genome Research* 8:186-194.
- [3] Mullikin, J.C., Hunt, S.E., Cole, C.G., Mortimore, B.J., Rice, C.M., et al. 2000. An SNP map of human chromosome 22. *Nature* 407:516-520.
- [1] Nickerson, D.A., Tobe, V.O., Taylor, S.L. 1997. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Research* 25:2745-2751.