# SGCaller: a program to call and review genotypes measured by sequencing

Carl Manaster, Ruta Valentonyte, Markus Teuber, Weiyue Zheng, Stefan Schreiber, and Jochen Hampe
*Christian-Albrechts-University Kiel, Kiel, Germany*

*BioTechniques 38:544-546 (April 2005)*

The genetic investigation of complex disorders requires single nucleotide polymorphism (SNP) genotyping in large patient cohorts. Although many specialized high-throughput methods exist, in a significant number of cases, no assay for a particular SNP can be designed due to either detection or chemistry-related problems. There are scarce published data on the frequency of such design failures (1); in our experience, this rate is in the range of 5%–25%, depending on the method and flanking sequence. SNPs used for simple haplotype tagging can tolerate some level of design failures, but those of particular functional relevance must be genotyped. For these SNPs, researchers may be prepared to use sequencing as an alternative method for genotyping for the following reasons: (*i*) the variations are often discovered through sequencing, which then represents a "default assay" for the SNP; (*ii*) in the case of highly polymorphic DNA regions, assay design may be impossible, and thus sequencing offers the opportunity to simultaneously call several SNP genotypes; (*iii*) sequencing equipment and technical know-how are widely available and do not require additional training or investment; (*iv*) in the case of difficult assays (as on genomically duplicated regions), sequencing offers additional confidence through the possibility to inspect the flanking sequence. For regions of very high variability, such as the HLA region, sequencing is an established genotyping method (2,3), and specialized commercial analysis software is available.

When "genotyping by sequencing" is performed, extracting genotype information is very cumbersome. There is scarce supportive software (except in specialized cases) because this is an unusual and underappreciated use of sequencing technology. PolyPhred (4) can automatically call genotypes at predefined sites through the use of manual tags in the ".polyphredrc" file. However, PolyPhred and Consed had a different and broader design scope than that of genotype calling. Because the interface is mainly designed to support contig assembly, the "vertical" assessment of hundreds of sequences at one specific base is difficult. GenBank® files cannot be directly processed, and the incorporation of manual call correction performed in Consed requires extra effort. Finally, a moderately complex set of programs and scripts in a UNIX environment needs to be maintained for the relatively straightforward task of genotype assessment at predefined sites.

Despite the existence of several automated methods for sequence trace analysis, sequence finishing still requires some user interaction. The same applies to sequencing genotypes—we think that an efficient and user-friendly way to review, assign, or correct genotype calls is essential. We therefore developed a program to perform largely automatic assignment of SNP genotypes based on annotated SNPs in a GenBank reference sequence (SNPs annotated with the "variation" feature key; file format specification provided at www.ncbi.nlm.nih.gov/Genbank). Sequences in the required GenBank format are available from various sources; they can be generated for a locus of interest through interfaces such as ENSEMBL (www.ensembl.org) or by the user using local sequence and SNP information with tools such as the free ARTEMIS (www.sanger.ac.uk/Software/Artemis/) sequence editor. SGCaller reads a collection of trace files in ABI or Staden (SCF) format, finds the variant position and tries to call the

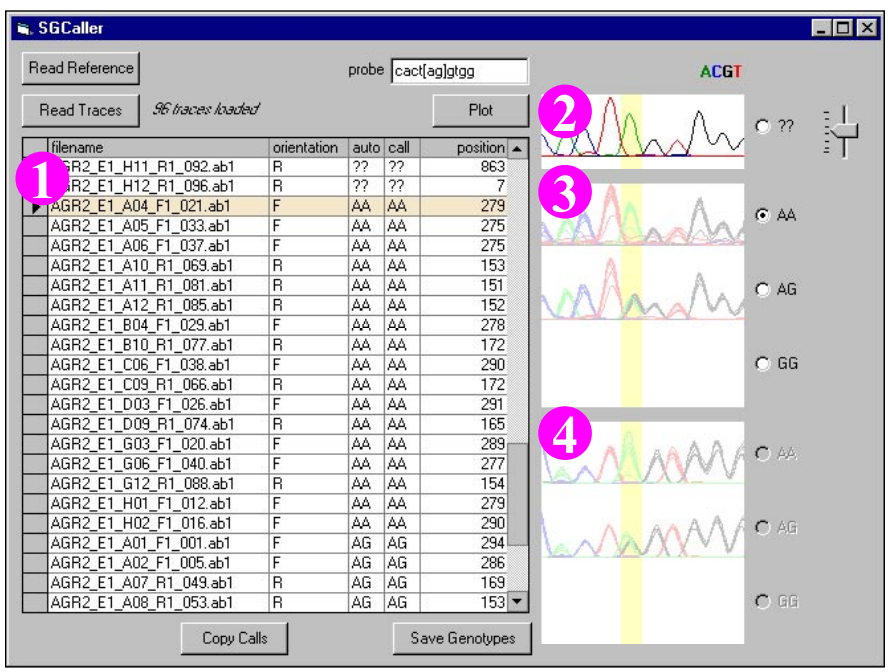


**Figure 1. The SGCaller user interface window.** The SGCaller window presents a list of trace files with the automatic calls. In detail, this window displays the list of trace files with the automatic calls [marked at (1) in the figure]; the trace for the selected file (2); already-called forward traces grouped by genotype (3); and genotype groups of reverse traces (4). In the trace groups, all relevant traces are plotted together on the same axes.

**Table 1. Comparison of SGCaller and TaqMan Genotypes for 12 SNPs in 47 Individuals**

| | Number of Genotypes Called | |
|---|---|---|
| SNP | SGCaller | TaqMan |
| rs9268430 | 43 | 46 |
| rs9268431 | 42 | 46 |
| AL034394_C20283T | 42 | 46 |
| AL034394_C12339T | 44 | 45 |
| AL034394_T10552C | 41 | 46 |
| BTNL2_i4_A239T | 34 | 46 |
| rs2076530 | 47 | 43 |
| DRAp-127 | 41 | 46 |
| hCV1702535[a] | 43 | 45 |
| AGR_CP-176 | 43 | 45 |
| AGR_CP-197 | 44 | 47 |
| AGR_CP-299 | 43 | 47 |
| Automatic Call Rate | 90% | 97% |

[a]Two genotypes of the hCV1702535 assay were discarded due to mismatch between sequencing and genotyping results. The overview window [marked as (3) in Figure 1] indicated poor sequence quality and led to the removal of the genotypes. Other calls were identical for all investigated genotypes.

genotype for each, and presents them for the user to confirm and enhance the calling. SGCaller determines each file's read direction from its name. After aligning sample traces to the reference, SGCaller calculates the total area under all four peaks for the base to be genotyped. Peak boundaries from the trace file are used; traces are not normalized to one another. By default, any trace where one peak has 80% or more of the total area under all four peaks is considered to be a homozygote. Traces with two peaks, each of which has 40% or more of the total area under all four peaks, are considered heterozygotes. Changing these thresholds tunes the calling specificity. The "plot" button refreshes the "overplotted" trace summaries (Figure 1) to provide visual feedback on such changes.

Automatic calling relies on high-quality sequence, and the software makes neither an extraordinary effort to automatically call poor sequence, nor does it assign quality scores. This is a deliberate design decision because robust genotype assignments can only be made in high-quality sequence. Only sequences that exactly match the reference for the eight bases surrounding the SNP are automatically called. Because SGCaller's call rate is limited, it leaves questionable traces for an interactive "user call." Peak forms vary depending on local sequence context and read direction, so it is useful to see what well-formed peaks from easily called traces look like when trying to determine a questionable call. For this reason, SGCaller presents already-called traces, grouped by call and read direction, below the selected trace (Figure 1). The trace groups also serve as a quick check of the program's automatic calls. This grouped overplotting visualization makes it easy to see if any heterozygotes have been miscalled as homozygotes. After review, the calls are saved to a tab-delimited text file.

SGCaller is written in Visual Basic and runs on Microsoft® Windows® XP, 2000, and NT. It does not require a database, but we have also developed a database-enabled version that works with our previously published genetic LIMS system (5). The program places no limits on sequence length, number of trace files, or the number of SNPs per read.

We have evaluated the performance of the automatic genotype assignment using the default settings for 12 SNPs in comparison to TaqMan® fluorescent genotyping technology (Applied Biosystems, Foster City, CA, USA); Supplementary Table S1 (available online at the *BioTechniques*' web site at www.BioTechniques.com) lists the sequencing primers and TaqMan primers and probes, while the genotyping results are shown in Table 1. Two automatic genotype calls of SGCaller at hCV1702535 differed from the TaqMan call. This difference due to poor sequence quality was spotted in the overplotted overview, which also serves as a quality feedback. All other genotype assignments were identical between the two methods. The automatic call rate of SGCaller was 90%; that of TaqMan was 97%.

In our experience, many positional cloning experiments encounter some difficult SNPs. We have used this software in many such cases with a significant improvement in productivity and data reliability. We anticipate that this program will serve to improve the efficiency of "genotyping by sequencing" for difficult and functionally crucial SNPs. The application binaries, sources, user guide, and sample data are available online at www.mucosa.de/sgcaller/.

## ACKNOWLEDGMENTS

*This study was supported by the German National Genome Research Network (NGFN, DHGP), the Competence Network IBD of the BmBF, the European Commission (FP5 grant), the German Research Council (DFG For423, HA3091/1-1) to S.S. and J.H., and an SUR grant from IBM to S.S.*

## COMPETING INTERESTS STATEMENT

*The authors declare no competing interests.*

## REFERENCES

1. **Yuryev, A., J. Huang, M. Pohl, R. Patch, F. Watson, P. Bell, M. Donaldson, M.S. Phillips, et al.** 2002. Predicting the success of primer extension genotyping assays using statistical modeling. Nucleic Acids Res. *30*:e131.
2. **Rajalingam, R., P. Ge, and E.F. Reed.** 2003. A novel sequencing-based typing system for HLA-DQA1 alleles. Hum. Immunol. *64*:S98.
3. **Santamaria, P., M.T. Boyce-Jacino, A.L. Lindstrom, J.J. Barbosa, A.J. Faras, and S.S. Rich.** 1992. HLA class II "typing": direct sequencing of DRB, DQB, and DQA genes. Hum. Immunol. *33*:69-81.
4. **Nickerson, D.A., V.O. Tobe, and S.L. Taylor.** 1997. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. Nucleic Acids Res. *25*:2745-2751.
5. **Hampe, J., A. Wollstein, T. Lu, H.J. Frevel, M. Will, C. Manaster, and S. Schreiber.** 2001. An integrated system for high throughput TaqMan based SNP genotyping. Bioinformatics *17*:654-655.

Received 10 October 2004; accepted 21 January 2005.

*Address correspondence to Jochen Hampe, National Genotyping Platform, Institute of Clinical Molecular Biology, Christian-Albrechts-Universität Kiel, Schittenhelmstr. 12 24105 Kiel, Germany. e-mail: J.Hampe@mucosa.de*