Sequence analysis

MACGT: multi-dimensional automated clustering genotyping tool for analysis of microarray-based mini-sequencing data

David C. Walley¹, Ben W. Tripp¹, Young C. Song¹, Keith R. Walley¹ and Scott J. Tebbutt¹,*
¹James Hogg iCAPTURE Centre for Cardiovascular and Pulmonary Research, St. Paul’s Hospital, University of British Columbia, Vancouver, V6Z 1Y6, Canada

Received on January 13, 2006; revised on February 16, 2006; accepted on March 2, 2006
Advance Access publication March 7, 2006
Associate Editor: Martin Bishop

ABSTRACT
Summary: Multi-dimensional Automated Clustering Genotyping Tool (MACGT) is a Java application that clusters complex multi-dimensional vector data derived from single nucleotide polymorphism (SNP) genotyping experiments using mini-sequencing based microarray chemistries such as arrayed primer extension (APEX). Spot intensity output files from microarray experiments across multiple samples are imported into MACGT. The datasets can include four channels of intensity data for each spot, replica spots for each SNP probe and multiple probe types (APEX and allele-specific APEX probes) on both DNA strands for each SNP. MACGT automatically clusters these multi-dimensionality datasets for each SNP across multiple samples. Incorporation of additional array datasets from known samples that have previously validated SNP genotype calls allows unknown samples to be automatically assigned a genotype based on the clustering, along with numerical measures of confidence for each genotype call. Calling accuracy by MACGT exceeds 98% when applied to genotyping data from APEX microarrays, and can be increased to >99.5% by applying thresholds to the confidence measures.

Availability: MACGT is open source and is freely available (under a GNU General Public License) from the iCAPTURE Centre web site, http://www.mrl.ubc.ca/who/who_bios_scott_tebbutt.shtml.

Contact: stebbutt@mrl.ubc.ca

Supplementary information: Additional information, including Supplementary Figure S1, test data and a user’s manual, is available from the iCAPTURE Centre web site (see above).

1 INTRODUCTION

Single nucleotide polymorphisms (SNPs) are common DNA sequence variations that occur when a single base pair in the genome sequence is altered (Wang et al., 1998). Combinations of SNPs in association with complex environmental factors are hypothesized to have a major impact in disease susceptibilities and outcomes (Risch and Merikangas, 1996). Determination of the base sequence at a specific SNP site is called genotyping. An ideal genotyping technique would be accurate, scalable, automated, fast and inexpensive. One such genotyping protocol is arrayed primer extension (APEX—Kurg et al., 2000), a mini-sequencing method combining advantages of a highly parallel microarray with the discriminatory power of the Sanger dideoxy terminator sequencing chemistry.

Open source software tools that fully automate the genotype calling part of the microarray image analysis process have not been forthcoming for APEX. Algorithms have been developed for the widely used Affymetrix GeneChip-based system, which relies solely on the discriminatory power of nucleic acid hybridization to generate the genotyping signals (Cutler et al., 2001; Huentelman et al., 2005; Liu et al., 2003). For APEX chemistries, Genorama proprietary software (www.asperbio.com) is able to detect all four colours of fluorescence emitted from the dyes used in an APEX experiment, and then automatically call the base(s) incorporated at a particular probe spot. However, the scoring algorithm treats all probes equally and considerable inspection of the original array data may be required to make a final genotype call (Kaminski et al., 2005; Kurg et al., 2000). Gemignani et al. (2002) developed a simple matrix-score based algorithm that takes Genorama base calling data for multiple probe types [APEX and allele-specific (AS) APEX probes] and calls the most likely genotype, but this still requires considerable manual inspection (Gemignani et al., 2002).

We recently developed SNP Chart—an APEX visualization and quality control tool (Tebbutt et al., 2005). SNP Chart generates visual patterns of spot intensity values from multiple channels, from a multiple probe set specific for a given SNP, easily interpretable as a specific genotype. The advantage over existing array data display methods is that one can easily look at an entire multiple probe set for a specific SNP, which can be more informative than looking at individual probes separately. In SNP Chart, the process of genotype calling is time-consuming and has user-subjectivity issues, because it is done manually (Tebbutt et al., 2004, 2005). As mentioned above, the ideal genotyping technique must be fast, with automated calling. Creating such an automated genotype calling protocol is a difficult task, since computers lack proficiency in...
recognizing visual patterns of multiple probe datasets. Our goal has been to implement a novel tool, providing accurate genotypes for many samples in a significantly shorter amount of time. We present here Multi-dimensional Automated Clustering Genotyping Tool (MACGT)—an open source, freely available software that is capable of automated genotype calling of the complex, multiple channel datasets derived from microarray-based mini-sequencing platforms such as APEX.

## 2 PROGRAM OVERVIEW

MACGT is fully automated genotyping software written in the Java language. MACGT is platform independent: it can be run on any operating system—including Windows and Linux—that can implement the Java run-time environment. In mini-sequencing methodologies such as APEX, each probe in the APEX microarray (classic APEX probes and AS APEX probes for both DNA strands) reacts with sample DNA to provide data (spot intensity measurement) in four different colour-coded channels (corresponding to the four bases of DNA—A,C,G and T). The different channels across multiple probes for each SNP are called vectors. MACGT undertakes multi-dimensional clustering genotyping analysis, because the information gathered from multiple vectors (or dimensions) is used to determine the genotype of the sample for a particular SNP. Using multiple dimensions provides more robustness to the data (Tebbutt et al., 2004).

MACGT has four major functionalities—data import and clustering, genotype calling with numerical confidence scores, data export and cluster visualization in 2D and 3D. Multi-channel spot intensity data, with SNP-specific oligonucleotide probe information, are imported from genotyping microarray experiments as text (tab-delimited) files (please see Supplementary information for format). A file of validated genotypes can also be imported, for samples that have corresponding APEX experimental data in the import set. The user then selects a data export ‘Report’ type, chooses an optional ‘Graphical Plots’ function for cluster visualization and has the ability to select various ‘Data Conditioning’ options (in our experience the ‘Normalized Groups of 4, Plot NNs’ option, with default numerical values, works well). MACGT goes through a number of calculation steps (outlined in summary in Supplementary information) generating clusters based on the spot intensity variances within the multi-dimensional dataset for each SNP. Genotypes are assigned to each cluster based on the validated genotype file provided during data import, and automated genotype calling of unknown samples is carried out.

If requested, a ‘Graphical Plot’ is generated for each SNP, displaying multiple information types (please see Supplementary Figure S1 as an example). In these plots, the multi-dimensional dataset for each sample is reduced to a single colour-coded point symbol. Automatically called genotypes for the samples are displayed, along with numerical confidence scores.

The ‘Report’ function of MACGT allows the called genotypes to be downloaded to a comma separated values file (genotype calling is independent of the ‘Graphical Plots’ option—the latter being a useful visualization of the clustering for troubleshooting purposes). Three export formats are available: ‘Verbose’ retrieves all data, including original channel intensities for each spot, data calculation steps for genotype clustering and assignment, and final genotype calls with associated quality scores; ‘Concise’ delivers the final genotype calls with associated quality scores and ‘Details’ delivers the genotypes and some of the data calculations.

To evaluate the accuracy of MACGT, we imported APEX data from a total of 35 independent experiments that genotyped Coriell DNA (http://coriell.umdnj.edu/) and negative control samples on our genomic control APEX microarray chip (Tebbutt et al., 2004; and S.J. Tebbutt, unpublished data). For 94 of the SNPs on this microarray, we had validated Coriell genotypes available to us that had been determined by other research groups and that showed all three genotypes across the Coriell DNA samples used. We randomly selected half of the Coriell samples as ‘Validated Genotype’ and we tested the accuracy of MACGT-calling for the remaining Coriell samples. Of the 1809 genotypes called that were not in the MACGT training (validated) set, we found 1776 to be concordant to the independent genotype data, representing an accuracy rate of 98.2% (at a call rate of 100%). Importantly, MACGT assigns numerical confidence scores to each genotype called. The most useful of these is the ‘Fit’ score. Applying a threshold to the ‘Fit’ such that any genotype call having a fit of zero becomes a non-call, results in the accuracy increasing to 99.6% (call rate drops to 90.7%). All datasets and MACGT graphical plots are available in Supplementary information. The data files freely downloadable from our website would allow potential users of MACGT to reduce the validation/training set by selecting just one or two examples of each genotype for each SNP, as long as the array intensity data for these samples were manually inspected beforehand for quality (e.g. using SNP Chart). Initial genotyping of reference samples is required for generating datasets corresponding to the validated genotype samples, necessitating that the genotyping method be independent of assay to assay variability. In our previously published work (Tebbutt et al., 2004) we showed that our APEX platform is robust to batch to batch variability, specifically by normalizing our channel intensity data to the signals from positive control probes.

In summary, MACGT is a clustering tool that accurately calls genotypes from complex, multi-dimensional datasets associated with robust microarray-based mini-sequencing platforms. MACGT is fully automated, requiring no manual inspection of the original microarray image data, and provides numerical measures of confidence for each genotype called, as well as an optional and highly effective graphical plot of the multi-dimensional data clustering for visualization and troubleshooting purposes.

## ACKNOWLEDGEMENTS

The authors thank Jian Ruan for laboratory technical assistance. This research was supported by the National Sanitarium Association (Canada), AllerGen NCE, the BC Lung Association, the Heart and Stroke Foundation of British Columbia and Yukon, CIHR and MSFHR.

Conflict of Interest: none declared.

## REFERENCES


