

# Evaluation Measures of ‘Marker Selection’ and ‘Genotype Membership’ in Large-Scale SNP Genotyping

**Eli Kaminuma**<sup>1</sup>  
eli@gsc.riken.jp

**Hiroshi Masuya**<sup>2</sup>  
hmasuya@gsc.riken.jp

**Hiromi Motegi**<sup>2</sup>  
hmotegi@gsc.riken.jp

**K. Ryo Takahasi**<sup>1</sup>  
kenzi@gsc.riken.jp

**Miki Nakazawa**<sup>1</sup>  
nakazawa@inplanta.jp

**Minami Matsui**<sup>1</sup>  
minami@riken.jp

**Yoichi Gondo**<sup>1</sup>  
gondo@gsc.riken.jp

**Tetsuo Noda**<sup>2</sup>  
tnoda@ims.u-tokyo.ac.jp

**Toshihiko Shiroishi**<sup>2</sup>  
tshirois@lab.nig.ac.jp

**Shigeharu Wakana**<sup>2</sup>  
swakana@gsc.riken.jp

**Tetsuro Toyoda**<sup>1</sup>  
toyop@gsc.riken.jp

<sup>1</sup> Functional Genomics Research Group, Genomic Sciences Center, RIKEN Yokohama Institute, 1-7-22 Suehiro, Tsurumi, Yokohama, Kanagawa 230-0045, Japan

<sup>2</sup> Functional Genomics Research Group, Genomic Sciences Center, RIKEN Tsukuba Institute, 3-1-1 Koyadai, Tsukuba, Ibaraki 305-0074, Japan

**Keywords:** Single Nucleotide Polymorphism, genotyping, genetic marker, evaluation measure

## 1 Introduction

Recently, Single Nucleotide Polymorphisms (SNPs) have been widely used as a new genetic marker for gene mapping and association studies. In large-scale functional genomics projects, a large amount of processing of SNP genotyping assays is required[4]. Most genotyping processes are implemented as high-throughput systems. However, in present technologies the process of allele discrimination requires subjective judgements by expert operators. Objective measures to evaluate genotyping results are necessary for avoiding subjective biases. In this report, we propose two useful evaluation measures for two dimensional fluorescent scatter plots of the TaqMan assays[3], which are a popular technology in SNP genotyping. They can be used in succeeding processes of genotyping. The first proposed measure is the Marker Selection Measure (MSM) to choose SNP markers with good distributions without depending on clustering results. The second proposed measure is the Individual Genotype Membership Measure (IGMM) which utilizes the membership probability of each genotype to provide beneficial information for gene mapping.

## 2 Method and Results

The first measure, MSM, quantifies the distance to an ideal distribution as SNP markers. First, we normalize scatter data during pre-processing (Fig.1A). Next, nonlinear probabilistic distribution of the scatter data is estimated by a Kernel method[1]. Then, the ideal probabilistic distribution for the SNP marker is set(Fig.1C). The difference between the ideal distribution and the estimated distribution of the scatter plot calculated by the multimodal overlap measure[2] is defined as the MSM. The second measure, IGMM, is defined as the membership probabilities of genotype classes. Currently, indistinct individuals remaining after allele discrimination are labelled ‘unknown’ by expert operators. The ‘unknown’ individuals may indicate intermediate fluorescent values of two genotypes or weak fluorescent values. The Mahalanobis distances for individuals based on genotype classes are first obtained. The IGMM is then calculated as the membership probability based on the distribution of squared Mahalanobis distances, which approximately follows the  $\chi^2$  distribution. For convenience, we defined the IGMM for four clusters by adding the negative control(NTC) of water, which indicates low fluorescent values around zero.

We attempted two experiments for the proposed measures. Inbred lines of C57BL/6 and DBA/2 of *M. musculus* were used for individuals. Seventy-six SNP markers were selected over all mouse

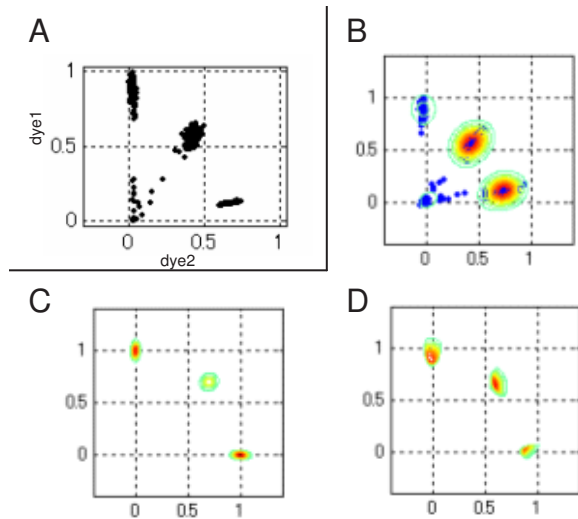


Figure 1: A: original scatter data; B: nonlinearly estimated probability distribution; C: ideal probability distribution; D: overlap probability distribution

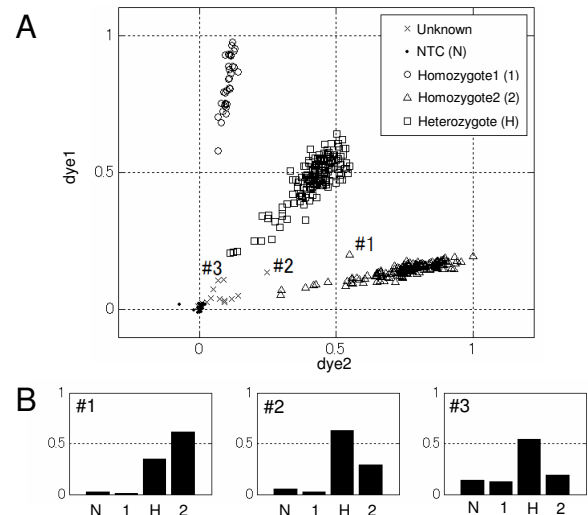


Figure 2: A: An example of individuals classified by expert operators; B: IGMM for individuals of #1, #2 and #3 in the upper scatter plot.

chromosomes. Each SNP marker had 384 data points, consisting of 376 individuals and 8 NTC points. In the first experiment, we subjectively ordered 76 SNP markers by the criterion of good separation and aggregation in the scatter plots. The Spearman correlation coefficient between the order by MSM and subjective order was 0.43, which was highly significant ( $P=3.5 \times 10^{-5}$ ). Thus MSM could provide a ranking of SNPs equivalent to subjective judgement. In the second experiment, we examined the number of 'unknown' individuals by applying IGMM. Fig.2A illustrates the scatter plot of D12SNP2 classified by an expert operator. The symbol  $\times$  denotes an 'unknown' individual. Three individuals on the figure were sampled as #1, #2, #3. IGMMs for the three individuals were calculated as shown in Fig.2B. All 'unknown' individuals were re-labelled as one of genotypes except where the membership probability of NTC was at a maximum. The total number of 'unknown' individuals was reduced by a mean value of 44% among the 76 SNPs.

### 3 Discussion

We proposed two useful measures of MSM in marker selection during the first process, and IGMM in allele discrimination during the second process. The experiment results indicated the effectiveness of MSM and IGMM. The silhouette measure[5] has been used for similar purposes. However, the silhouette measure does not consider membership probabilities and cannot deal with unknown individuals without forced assignment to a genotype class. IGMM can reduce the number of unknown individuals while avoiding those problems. Our proposed measures would save operators from problems with allele discrimination. In future works, incorporating IGMM into gene mapping to estimate chromosomal locations of causal genes needs to be considered.

### Acknowledgments

We thank Risa Shoji and genotyping staff for technical support in the SNP genotyping assays.

### References

- [1] Hastie T., Tibshirani R., Friedman J.H., *The Elements of Statistical Learning*, Springer, 2001.
- [2] Kil, D.H. and Shin, F.B., *Pattern Recognition and Prediction with Applications to Signal Characterization*, AIP Press, 1996.
- [3] Livak, K.J., Allelic discrimination using fluorogenic probes and the 5' nuclease assay, *Genet. Anal.*, 14:143-149, 1999.
- [4] Masuya H. et al., Development and implementation of a database system to manage a large-scale mouse ENU-mutagenesis program, *Mammalian Genome*, 15(5):404-411, 2004.
- [5] Rousseeuw, P.J., Silhouette: A graphical aid to the interpretation and validation of cluster analysis, *J. of Computational and Applied Mathematics*, 20:53-65, 1987.