# Novel algorithm for automated genotyping of real time data

Callegaro A.(1), Spinelli R.(2,3), Battaglia C.(2,3), Caristina L.(2), Cenzuales S.(4), Beltrame L.(2,3), Bicciato S.(1)

(1) Department of Chemical Engineering Processes – University of Padova – Padova (2) Array Technology Group, Interdisciplinary Center for Biomolecular Studies and Industrial Applications (CISI) – University of Milano – Milano(3) Laboratoty of Biochemistry, Department of Sciences and Biomedical Technology – University of Milano – Milano (4) Servizio trasfusionale, Ospedale Sacco e Polo Universitario , Milano

## Motivation

A set of over 3 million putative single Nucleotide Polymorphism (SNPs) are now available for disease association studies, eventually replacing the current RFLP and microsatellite (STRP) linkage analysis screening sets. Many different technological platforms are available for allelic discrimination and, among them, Taq Man chemistry employing 5' nuclease allelic discrimination is one of the most widely diffused. Each Taq Assay employs two allelic specific probes for each SNPs, commonly VIC or FAM dyes. Although technological advancements in assay design and application allows monitoring SNPs at a high-throughput rate with this technology, the allele calling procedure of any single sample still requires the manual intervention of an expert operator for adjustment of SNP-dependent thresholds, signal selection, and quality revision. Usually the genotyping is calculated at the end of the amplification process resulting in an end point analysis. To enable the genotyping of multiple SNPs in several different samples we developed an algorithm for the automatic allele-calling from real time data.

## Methods

The proposed algorithm for allelic discrimination is composed of three major steps, namely i) sample clustering, ii) cluster assignment (call) and iii) quality assessment of genotype call, and uses all fluorescent data obtained at the end of each amplification cycle. Specifically, samples are first clustered into k (k=2, 3) clusters based on the fluorescent signal differences of the two dyes (i.e., d=FAM-VIC) at the time of maximum difference variance. Clustering is assessed using the clustering around medoids partitioning technique. Clusters are then labeled using group differences (d=FAM-VIC) at the time of maximum difference variance. The heterozygote class is identified by comparison with the cluster of the blank samples, given the assumption that the fluorescent signals of both probes in the heterozygote group is comparable to those of blank samples. Homozygote groups are identified by comparison to the heterozygote or blank cluster. For each sample, the quality of the allele assignment is assessed evaluating the silhouette width s(i). Silhouette width, intended as an index of sample clustering strength, is defined as the average dissimilarity between sample i and all other points of its class as compared to all observations in the neighbor clusters. In particular, observations with a large s(i) (close to 1) are very well clustered, a small s(i) indicates an observation lying between two clusters, while observations with a negative s(i) are misclassified. The average silhouette width over all samples quantifies the quality of the entire SNP clustering.

## Results

The algorithm has been tested for the genotyping of 7 different SNPs obtained using Taq Man Assay on demand (ABI) genotyping assays on DNA health individuals. The assays have been

carried out on iCycler instrument (BIORAD). The genotype of all DNA samples has been confirmed by conventional sequencing technique. The autocalling algorithm has been able to correctly identify all possible allelic combinations (e.g., one, two or three populations) without requiring any training set or the definition of a predefined allelic frequency. Each single genotype call is associated to an accuracy estimation parameter s(i), quantifying the classification confidence. In addition, since genotype call is obtained at the time of maximum difference variance, the method is robust also in the presence of extremely noisy data. As compared to sequencing data, the real time data set, composed of 7 SNPs and 14 samples, was classified with an overall efficiency of 100%. The algorithm is threshold independent and takes advantages of all real time signals. This work was supported by grants MIUR-FIRB RBNE01TZZ8 and MIUR-FIRB RBNE01HCFK and by CISI funds.

**Contact email:** silvio.bicciato@unipd.it