5. Improving SNP genotyping results by introducing Gaussian mixture models for genotype calling

Christian Gieger, [1] **Maik Kschischo,** [2] **Rainer Kern,** [3] **Martin Steinhauser,** [4] **Ralf Tolle** [5]

Keywords: SNP, clustering, Gaussian mixture model, genotyping, quality control

1 Introduction.

Genetic diversity and its functional relevance is a main topic of current genomic research. Single nucleotide polymorphisms (SNPs) are used in genetic mapping, target gene association studies and in probing the causes of differential responses to drug treatment. SNP genotyping on an industrial scale relies on a few technologies, which are robust, amenable to automation and relatively cost-efficient. One of the genotyping assays which has been adopted in numerous labs is the template-directed dye-terminator incorporation assay with fluorescence polarization detection (FP-TDI) [2]. Here a genotyping primer is annealed next to the prospective polymorphic site and extended by the respective dye-terminator labeled base(s). Which dye-terminator was incorporated is then determined by measuring the decrease in fluorescence depolarization for the excitation/emission wavelengths corresponding to the fluorescent dyes used for labeling.

Raw data are analyzed by plotting the data for the two colors which are representing the two alleles of a SNP. Ideally, this XY-plot leads to distinct clusters representing the different genotypes. There are however cases, when data points cannot be easily assigned to an individual cluster [4]. Our approach is to apply Gaussian mixture (GM) models [1] to the clustering of SNP assay data in order to reduce human intervention in genotype calling to a minimum and at the same time to provide objective and quantitative criteria for the assessment of assay quality. We derived several quantities in the framework of Gaussian mixture models which provide figures of merit for the assignment of observations to clusters, thus vitally improving the whole process of primer extension SNP genotyping and opening a door to fully automated SNP genotyping on an industrial scale.

2 Methods.

A Gaussian mixture model assumes that each group (genotype cluster) of the data is generated by an underlying probability distribution. Let $\mathbf{y} = \{\mathbf{y_1}, ..., \mathbf{y_n}\}$ denote the independent multivariate observations to be clustered, with $\mathbf{y_i}$ being a d-dimensional vector ($i = 1, ..., n$). In our case we have $d = 2$ and the components of $\mathbf{y_i}$ correspond to the scaled values of the FP-TDI. Let K be the number of components in the data. Under the assumtion of

[1]Fraunhofer Institute for Algorithms and Scientific Computing, Schloss Birlinghoven, St. Augustin, Germany. E-mail: christian.gieger@scai.fhg.de

[2]University of Applied Sciences Koblenz, RheinAhrCampus, Remagen, Germany. E-mail: kschischo@rheinahrcampus.de

[3]LION bioscience AG, Heidelberg, Germany. E-mail: rainer.kern@lionbioscience.com

[4]LION bioscience AG, Heidelberg, Germany. E-mail: martin.steinhauser@lionbioscience.com

[5]PHENEX pharma, Heidelberg, Germany. E-mail: ralf.tolle@phenex-pharma.com

independent and identical distributed observations $\mathbf{y_i}$ one obtains for the likelihood $\mathcal{L}$ of the mixture model:

$$\mathcal{L}_{MIX}(\Theta_1, ...\Theta_K|\mathbf{y}) = \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k f_k(\mathbf{y_i}|\mu_k, \Sigma_k), \quad (1)$$

where $\Theta_k = (\mu_k, \Sigma_k, \pi_k)$ $(k = 1, ..., K)$ is the set of parameters for the kth cluster and f_k is the density of the kth Gaussion mixture component with mean μ_k and the covariance Σ_k. Here, π_k is the probability that an observation belongs to the kth cluster with $\sum_{k=1}^{K} \pi_k = 1$.

Maximization of the likelihood in Equation (1) provides estimates $\hat{\Theta}_k = (\hat{\mu}_k, \hat{\Sigma}_k, \hat{\pi}_k)$ for the unknown parameters of the mixture. The standard choice for fitting mixture models is the Expectation-Maximization (EM) algorithm [3] which proved to be efficient for our purposes. An important issue in mixture modeling is the selection of the number of clusters. Two models $\mathcal{M}_1$ and $\mathcal{M}_2$ can be compared by the Bayes factor, that is, the posterior odds ratio $P(\mathcal{M}_1|\mathbf{y})/P(\mathcal{M}_2|\mathbf{y})$ for one model against the other. We used the Bayesian Information Criterion $BIC = 2\log[\mathcal{L}(\mathbf{y}|\hat{\Theta})] - m(\mathcal{M})\log n$ to get a computationally efficient approximation of the Bayes factor. Here, $\mathcal{L}(\mathbf{y}|\hat{\Theta})$ is the maximized likelihood, $m(\mathcal{M})$ is the number of parameters of model $\mathcal{M}$.

The fit of a mixture model to an assay puts us in the position to quantify the reliability of (i) individual observations and (ii) the assay as a whole. Individual observations are unreliable, when the uncertainty about assigning the observation to a cluster is high. We quantify this by estimating the probability of an observation pertaining to one of the potential clusters by

$$p(k|\mathbf{y_i}) = \frac{\pi_k f_k(\mathbf{y_i}|\mu_k, \Sigma_k)}{\sum_{\nu}^{K} \pi_\nu f_\nu(\mathbf{y_i}|\mu_\nu, \Sigma_\nu)} \quad (2)$$

However, also the assay itself may be unreliable and the distinguishability of different clusters may be poor. This can be quantified by the overlap $\Psi_{k,l} = \int \sqrt{f_k(\mathbf{y}|\Theta_k) f_l(\mathbf{y}|\Theta_l)} d\mathbf{y}$, also known as Bhattacharyya distance. The overlap is bounded between 0 and 1. Two identical densities have overlap 1 while infinitely separated clusters have overlap 0. The smaller the overlap, the more distinguishable are the densities. In addition, the overlap allows to compute an upper and lower boundary for the probability to decide for the wrong cluster (Bayesian probability of error).

3 Results.

The applicability of GM models to the clustering of SNP assay data was tested on a total of 356 non-redundant SNP assays generated in the course of a genetic target gene validation study. In total, this experimental data set contains 118,016 data points and 178 different SNPs. We will discuss the results and benefits of using GM models.

References

[1] Banfield, J. D. and Raftery, A. E. 1993. Model based Gaussian and non- Gaussian clustering. *Biometrics* 49:803-821.

[2] Chen, X., Levine, L. and Kwok, P. Y. 1999. Fluorescence polarization in homogeneous nucleic acid analysis. *Genome Res.* 9:492- 498.

[3] Dempster, N., Laird, N. and Rubin, D. 1977. Maximum likelihood estimation from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B* 39:1-38.

[4] Grant, S. F. A., Steinlicht, S., Nentwich, U., Kern, R., Burwinkel, B. and Tolle, R. 2002. SNP genotyping on genome wide amplified DOP-PCR template. *Nucleic Acids Res.* 30:e125.