

# Single-sample SNP models: genotyping, copy number estimation and signal calibration

Ralph C.A. Rippe<sup>1</sup>, Paul H.C. Eilers<sup>2</sup> and Jacqueline J. Meulman<sup>1</sup>

<sup>1</sup> Leiden University; <sup>2</sup> Erasmus Medical Center

More and more publications relate (clinical) outcomes to DNA composition. One specific DNA feature is the Single Nucleotide Polymorphism (SNP), a variation in two supposedly equal alleles (one for each chromosome). Applications are found in child, education and family studies (e.g. attachment, susceptibility), but also in tumor research as well as in plants and food.

Common (bio)statistical challenges in using SNP data are two-fold. First, one needs to determine SNP genotypes. Many researchers take SNP genotypes as given, while a whole range of statistical models is used to obtain them. Second, one can determine the number of alleles present at a given position on a chromosome. In healthy tissue, we find two allele copies, but random fluctuations as well as tumor tissue can result in deviations from the normal two alleles for whole regions of chromosomes. In my research I (we) have worked on both challenges. For genotype estimation, we fit log-concave densities on smooth 2D-histograms to obtain cluster probabilities for each SNP in one sample. Our approach is in strong contrast with common methods, resulting in (fierce) discussions on ethics, project cost management and return time for results. Comparisons with gold standard HapMap genotypes show adequate performance and illustrate additional possibilities. To estimate allele copy number profiles we use a segmentation method that is a so-called  $L_0$ -penalized smoother for single signals, with interesting extensions, for example in scatter-plot smoothing.

In the above models we use two signals, one for each of the two chromosomes, obtained from a chemical process. Unfortunately, this methodology does not provide perfect results. While working on the above challenges, we also found structural behavior of SNP signals over different samples. To improve the quality of the fluorescence measurements we model this behavior using an ANOVA-like model.