Universiteit
Leiden
The Netherlands

Advanced statistical tools for SNP arrays : signal calibration, copy number estimation and single array genotyping

Rippe, R.C.A.

Citation
Rippe, R. C. A. (2012, November 13). *Advanced statistical tools for SNP arrays : signal calibration, copy number estimation and single array genotyping*. Retrieved from https://hdl.handle.net/1887/20118

Version:	Not Applicable (or Unknown)
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/20118

Note: To cite this publication please use the final published version (if applicable).

Cover Page

Universiteit Leiden

The handle http://hdl.handle.net/1887/20118 holds various files of this Leiden University dissertation.

Author: Rippe, Ralph Christian Alexander
Title: Advanced statistical tools for SNP arrays : signal calibration, copy number estimation and single array genotyping
Issue Date: 2012-11-13

Summary

Single Nucleotide Polymorphisms (SNPs) are small variations in DNA, in single nucleotides. This is the official name when one or both alleles (one for each chromosome) vary in 1% or more of the population. Many of these polymorphisms are innocent, but in worse cases they can constitute tumor development. For this (and other) reasons research on the presence, form and effects of these SNPs was performed on an increasing scale in recent years. It is important to know which combination of alleles (the genotype) is present and whether there are deviations from the normal number of two alleles. The latter deviations are known as copy number variations.

Different manufacturers deliver platforms for SNP analysis. In all cases a process called selective hybridization is applied: probes selectively react (hybridize) to specific positions on the genome. Fluorophores are attached to specific molecules to label them and use the resulting fluorescence signals to determine allele A and B concentrations. Some manufacturers use one fluorophore (Affymetrix), others use two (Illumina). The amount of observed fluorescence represent the (relative) dose of the alleles. High signal for A and low for B indicates genotype AA, and vice versa for the BB genotype. Equal signals, thus equal doses, indicate genotype AB. Hence the total allele dosage in healthy DNA equals 2.

It practice it appears that these signals have some structural properties. If in one sample the signal for a single SNP is relatively bright compared to that of other SNPs, an equivalent proportionality also extends to other samples. This, combined with differences between samples and given the signal differences relating to allele doses (genotypes), implies that the aforementioned properties can be modeled. A model that does so, SCALA (Signal Calibration Algorithm) is introduced in this thesis. It estimates the systematic effectlevel for each SNP and corrects for it: calibration. SCALA distinguishes two variants: 1) estimating effects that allow for calibration independent of

genotypes (genotype-free calibration) and 2) estimating effects conditional on genotype, hence allowing for genotype-specific calibration. The latter approach requires known genotypes in order to be applied. The estimated effects are stable within one type of chip: they hold for new arrays and hence allow for calibration. Applying calibration with the estimated effects from one of the variants reduces variance, while dose ratios remain intact and chromosomal regions with deviations from dose 2 are clearer.

Determining genotypes from allele dosage is not completely trivial. Here, we look at allele dose ratio for all SNPs within one individual. The main advantage is that we never suffer from cluster imbalance due to low minor allele frequencies. After transformation of the signals for A and B we observe data scatter with three distinct clusters, representing the genotypes AA, AB and BB. The method is of semi-parametric nature, hence not making assumptions on the cluster shapes. The idea is that the data are transformed into a 2-dimensional histogram. On the obtained counts we fit a logconcave density model for each of the three clusters. This way, the model is also insensitive to increases of the number of observations: the histogram has unchanged dimensions. The model itself is also highly efficient.

The described approach is a so-called "single array" method and is introduced in this thesis. It contrasts with mainstream methods that are "multi-array". In case of genotyping that means that each individual SNP is genotyped in a set of arrays, based on the same dosage ratio. Hence the results and their quality depend on the amount of available observations, i.e. the number of arrays. In practice the proposed single array method has at least equivalent performance when comparing both branches to HapMap genotypes. There is no gold standard, but HapMap is a database that provides derived genotypes — from a few algorithms — for a set of reference arrays. Furthermore, working with single arrays allows for genotyping of SNPs that would be otherwise undetermined.

The previously mentioned signal calibration is also useful is case of low quality arrays. After calibration we observe strongly improved cluster separation, and therefore decreased calling uncertainty, for different platforms.

In unhealthy DNA variations on the total allele dosage 2 can be found.

Dosage for alleles A and B can also be 0, 1, 3 or more, resulting e.g. in genotypes 0, A0, 0B, AAA or ABB. When estimating these dosage profiles along positions on the chromosome a signal smoother is used that irons out small differences due to signal variation. The usability of this profile depends on the smoother used. A smoother that penalizes the number of changes between segments (a penalty with L_0 norm) finds only those changes, and nothing else, while penalizing the size of the changes also picks up noise.

Here, reference profiles with healthy DNA, thus dosage 2 can be used. However, in practice it is sometimes hard to obtain these samples. A comparison to VEGA shows that using an L_0 norm in the penalty gives reliable results even without reference data. The penalty can also be used in smoothed scatterplots; by using it in one direction we obtain smoother approximations per segment.

An extension of this penalty to a model for allelic imbalance seems obvious. However, in practice it is not: it requires additional steps, thereby increasing uncertainty. A first attempt is given by fitting a mixture of distributions on histograms per segment.

Some directions are left untouched. For example, extrapolations of the techniques to tetraploid DNA (e.g. in potatoes, leek, roses and some species of fish, like salmon. Another interesting approach is to jointly model genoypes, copy numbers and calibration parameters. Replacing the hard-called genotypes by their fuzzy counterparts (the genotype probabilities) may further increase model effectiveness.