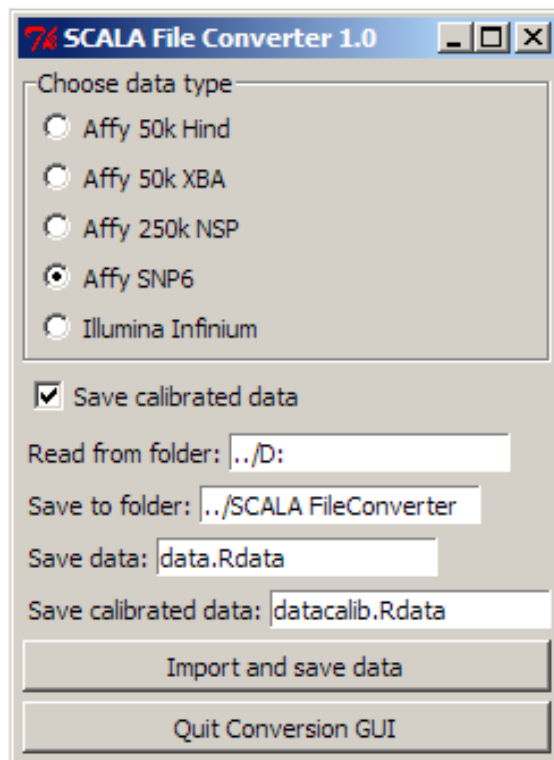


Convert SNP platform data

SCALA File Converter

Version 1.0.00 for R



Author:
Ralph C.A. Rippe

Maintainer:
Ralph C.A. Rippe

Contents

1	File Conversion	2
1.1	Platforms	2
1.2	File formats	2
1.3	File locations	3
2	Calibration	4

Chapter 1

File Conversion

SNP fluorescence signals can be obtained from various platforms like Affymetrix, Illumina and Agilent (see HapMap, 2007). For any given platform we can transform the output (from CEL-files or other files) into signal x for allele b and y for allele b for a single sample. Obtaining signals for multiple samples then leads to matrices X and Y , where the samples are in the columns and SNPs are in the rows. x and y contain signals that are averaged over all probes. Each allele now has a single intensity value.

Note that this software is developed to work only with R, versions 2.7.2 and above. If you encounter any problems, please notify the author.

1.1 Platforms

This tool takes the raw datafiles from Illumina (Xraw and Yraw) or Affymetrix (CEL-files) as input and returns the X and Y data matrices for all samples available in the selected folder (section 1.3). Internally, a platform-specific mapping is used to transform the raw input signals to x and y signals. Currently supported platforms are Affymetrix (50k Hind and XBA, 250k NSP and SNP6) and Illumina (Infinium). Suggestions for other platforms can be requested at the author (email to rrippe /at/ fsw.leidenuniv.nl and state your platform type and subtype, as well as a publicly available download location, if possible).

1.2 File formats

The resulting signals are stored as internal R objects and saved to a chosen file at a chosen location. Data files contain several objects, including

- $[X,Y]$, the signal matrices of size $n \times m$, where n is the number of SNPs and m is the number samples
- $[fnames]$ the included sample names from the input folder, matching m
- $[chr]$, of length n , the associated chromosome allocations for the n SNPs
- $[pos]$, of length n , the SNP position on the associated chromosome.

1.3 File locations

At GUI startyp, the GUI doesn't show and initialize completely until the input and output-folder are chosen (Figure 1.1. After these are selected, they show in the GUI, but cannot be changed *via* the GUI! After full initialization, the filenames to save the (calibrated, see chapter 2) data to can be chosen. Just type in the desired filename and *hit* [Enter] afterwards. Skipping [Enter] will not save the chosen filename!

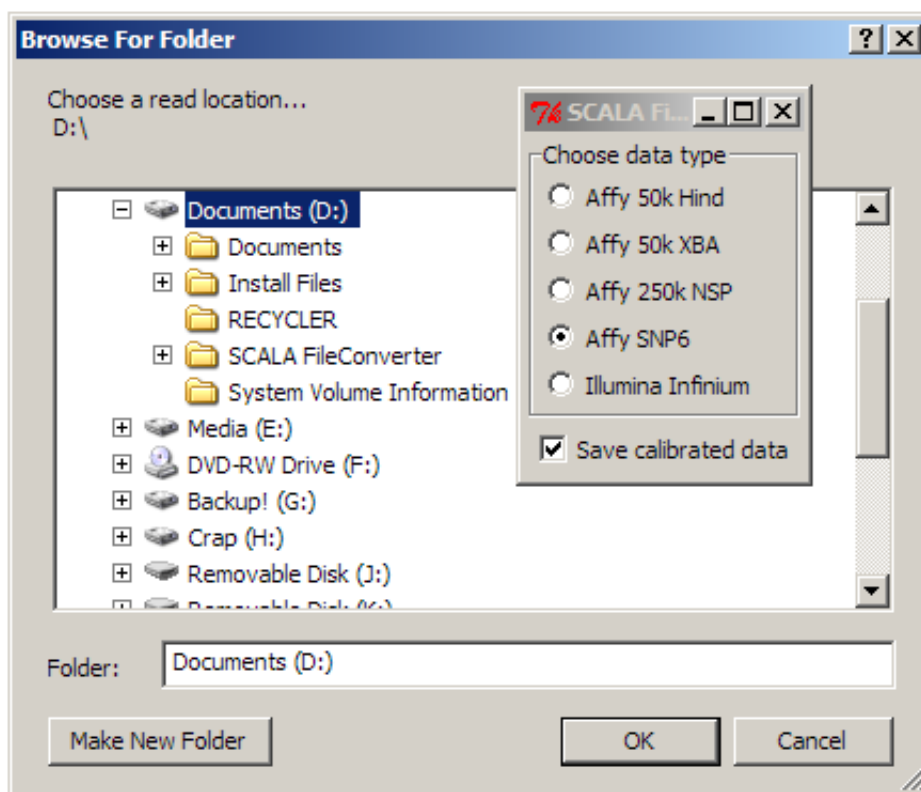


Figure 1.1: .

If all settings are as desired, hit Import and Save data.

Chapter 2

Calibration

Parameters (α and Γ , see Rippe et al, 2010) that represent the relative strength of each SNP are used to calibrate the noise in x and y . We distinguish two types of calibration (see Rippe et al, accepted for publication). The file convertor can do only the genotype-free approach, since genotype information may not (yet) be available. This calibration is performed by taking $x^* = x_{ij}/10_i^\alpha$ from the global model.

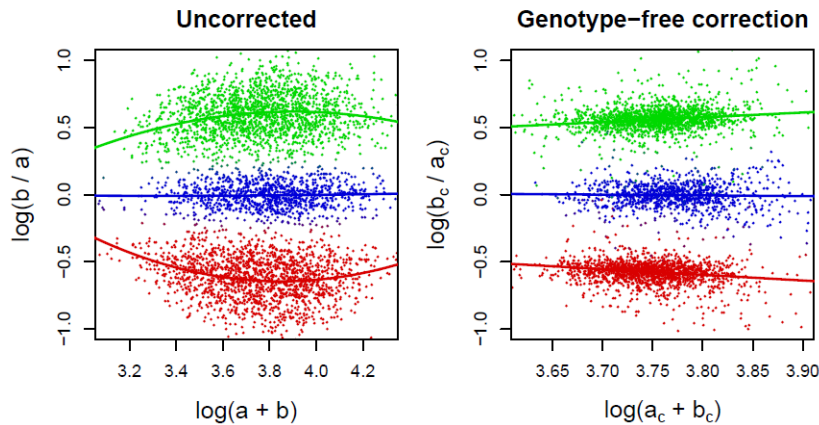


Figure 2.1: Calibration results. Left: signals before calibration. Right: signals after genotype-based calibration. Note the changed range on the x -axis, due to the calibration.

Figure 2.1 illustrates the results for an Affymetrix 50k Hind sample retrieved from the HapMap database.

Calibration can also be beneficial to detection of CNV and LOH, since the signal noise is strongly condensed, as is illustrated in Figure 2.2 and Figure 2.3.

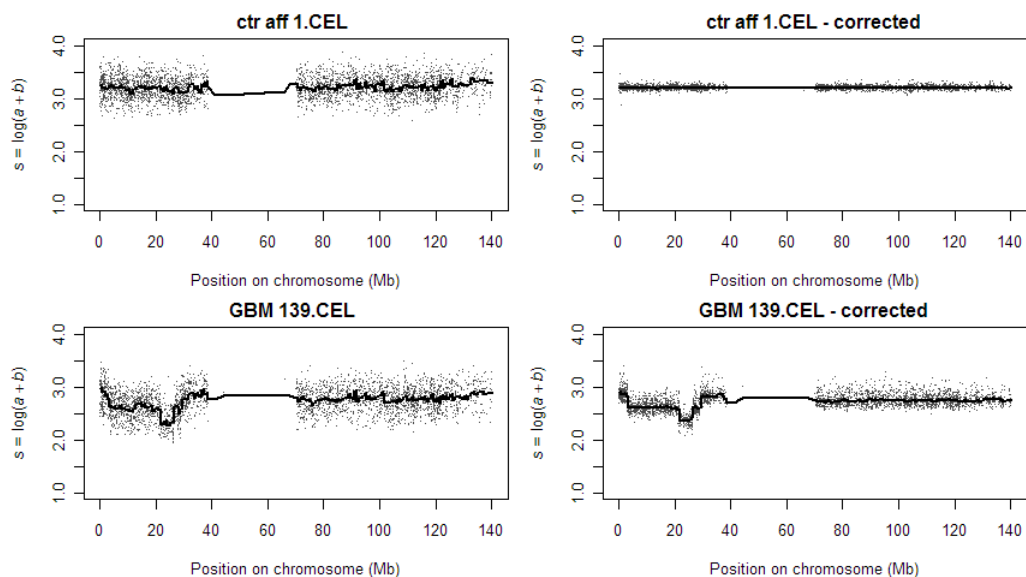


Figure 2.2: Calibration results for CNV in a clean and tumor sample from Affymetrix 250k NSP.

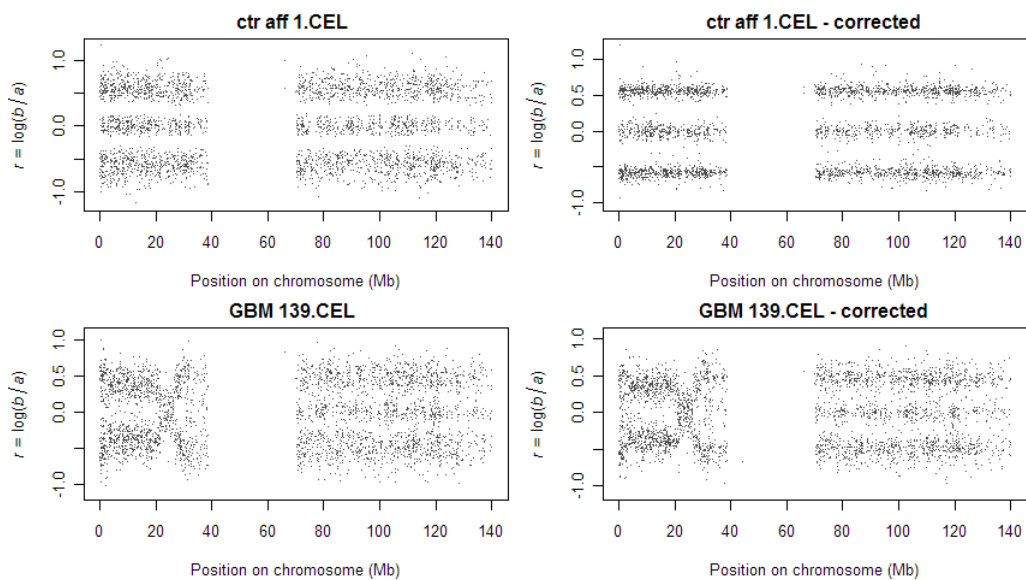


Figure 2.3: Calibration results for LOH in a clean and tumor sample from Affymetrix 250k NSP.