CytoPreparePercentage

Questions

How to prepare the percentages for a heatmap?
What is the meaning of the color scale?

Definitions

Tab Delimited, Multiple Sample (TDMS) File Format

NAME	GenBank	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5
	Treatment	placebo	placebo	placebo	placebo	placebo
	Patient ID	101	102	103	104	105
	Gender	F	F	M	M	M
A1	AA757429	0.246457	0.226136	-0.27992	0.278945	0.416614
A2	AA159571	-0.32316	0.007096	0.354105	-0.32402	-0.40812
A3	AI289227	0.202585	-0.09618	-0.32919	-0.38246	-0.29769
A4	AA521292	-0.36587	0.253069	0.359742	-0.14589	0.032514
A5	AI344061	0.033734	0.341731	0.371982	-0.11216	-0.48229
A6	AA705237	0.104424	-0.1359	0.425171	0.275234	-0.07836
A7	AA485782	-0.02516	0.047777	-0.36558	-0.15541	0.401895
A8	H81821	0.024793	-0.10033	0.137568	-0.10322	0.287513
A9	T48524	0.483648	0.429508	-0.23373	0.389819	0.204742
A10	W24076	-0.03916	-0.00601	0.479826	0.086263	0.338572
A11	AA936275	0.265952	0.067	0.027316	-0.09708	0.463795
A12	AA917076	-0.21308	0.229778	-0.18721	0.052883	0.119697
A13	T41038	-0.43006	0.057859	0.253943	-0.46023	-0.34688

Matrix values are percentages
Lines are populations of cells
Columns are samples
Samples are arranged as groups

Main formula

transform, center, then scale

$$display_{ij} = \frac{transform(percent_{ij}) - center_{ij}}{scale_{ij}}$$

Typical uses

Usual "absolute difference" use:
log transform
overall centering

"Scaled discovery" use:
log transform with soft floor
overall centering using median
standardize range using percentiles

"Supervised" use:
log transform with soft floor
overall centering using median
standardize average within group dispersion

"Reference group" option:
discovery or supervised use then
fix the reference group level

"Summary" use:
discovery or supervised use
reduce to central tendancy of each group

"Noisy differences" use:
discovery (or supervised) use then
sort values within group

"Robust noisy differences" use:
discovery or supervised use
sort values within group
substitute values using one gaussian per group

Step benefits

transform:
* compress overall dynamical range
* make dynamical range similar between populations
* soft floor avoids over dispersion of lowest values

centering:
* remove base line for each population
* allow comparing differences within each population
* color scale relates differences to the base line

standardize:
* relate differences within each population to the same standard
* stretch or squeeze differences in proportion to the standard
* color scale relates differences to the same standard
* within group dispersion standard enlights statistical significance
* wgds avoids confusing within and between group dispersions

sort within group:
* visual check that values between groups overlap or not

substitute:
* reshape data in a robust approach trying to weaken outliers
* grouping is kept between population, but sample order is lost

reference group:
* set a group as base line; any further scaling should not change it
* define the value for this group

summarize:
* give a synthetic view, possibly with dispersion

Steps

- 1/ transform
  - none
  - log2
  - asinh
  - **option: soft floor**
- 2/ centering
  - to: overall.line, ref.group, line.but.ref
  - method: trimmed mean, median
- 3/ standardize
  - discovery = no grouping
    - to: overall dispersion = 1
    - to: low-high percentile range = 1
  - **supervised = groups defined**
    - to: average within group dispersion = 1
  - method: dispersion = sd or mad
- **4/ sort within group**
  - imputation: NA = median
- **5/ substitute**
  - overall rank
  - overall gaussian
  - gaussian within group
- **6/ within group summarized value**
  - median
  - median plus percentiles