Computational analyses used in stochastic profiling

Sameer Bajikar

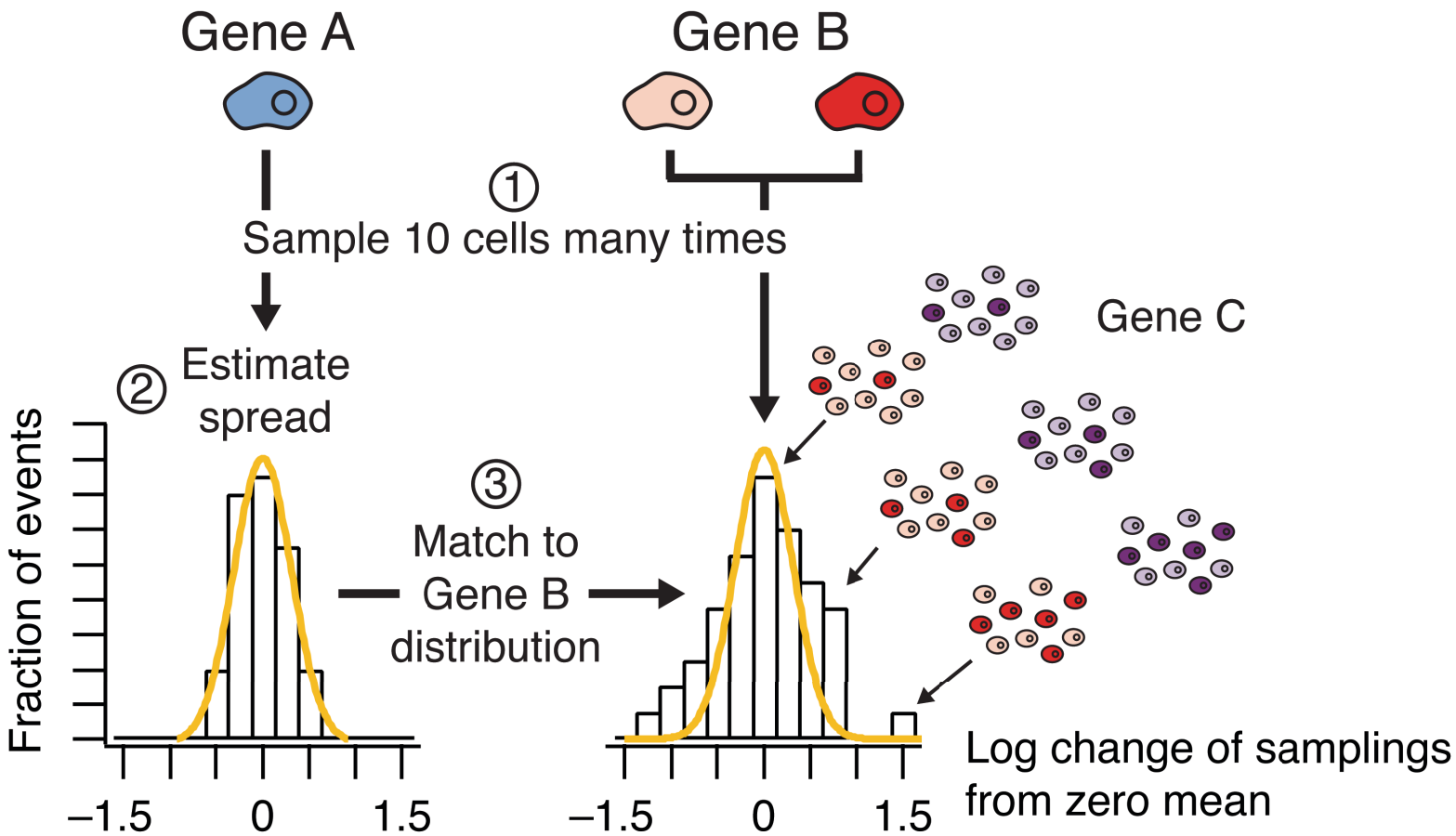
Stochastic profiling workshop

7/30/15

Objectives for lecture

- Understand the mathematical implementation of stochastic profiling
- Use simulations to identify sampling strategies for your heterogeneities of interest
- Understand how global profiles are processed and filtered
- Understand how heterogeneities can be parameterized

Theory behind stochastic profiling



Nat Meth **7**, 311-317 (2010)

Where did “10-cells” come from?

- Use Monte-Carlo simulation to identify optimal sample conditions for heterogeneities of interest
  - Simulate random selection of individual cells within an *n*-cell sample
  - Compare distribution of samples to control (homogeneous) stochastic profiles

Mathematical implementation of filter

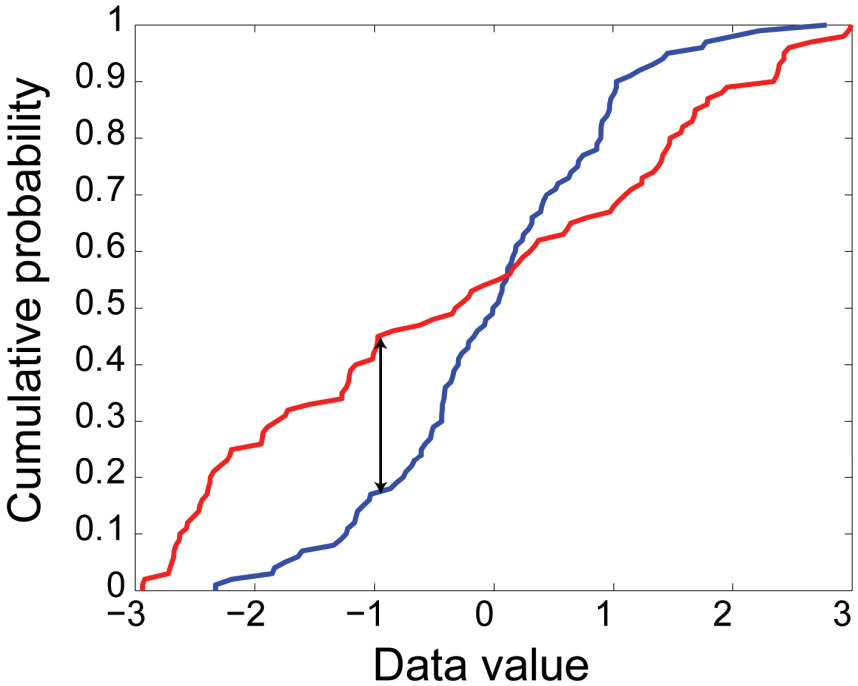
- Considerations
  - Unbiased
  - Robust with few samples
  - Computationally efficient

Kolomogorov-Smirnov Test (KS Test)

- Non-parametric statistical method to test equality of probability distributions of two samples
  - Can also test whether data follow a given distribution
- Calculates the maximum difference between the given cumulative distribution functions (cdf) of the data (or hypothetical distribution)

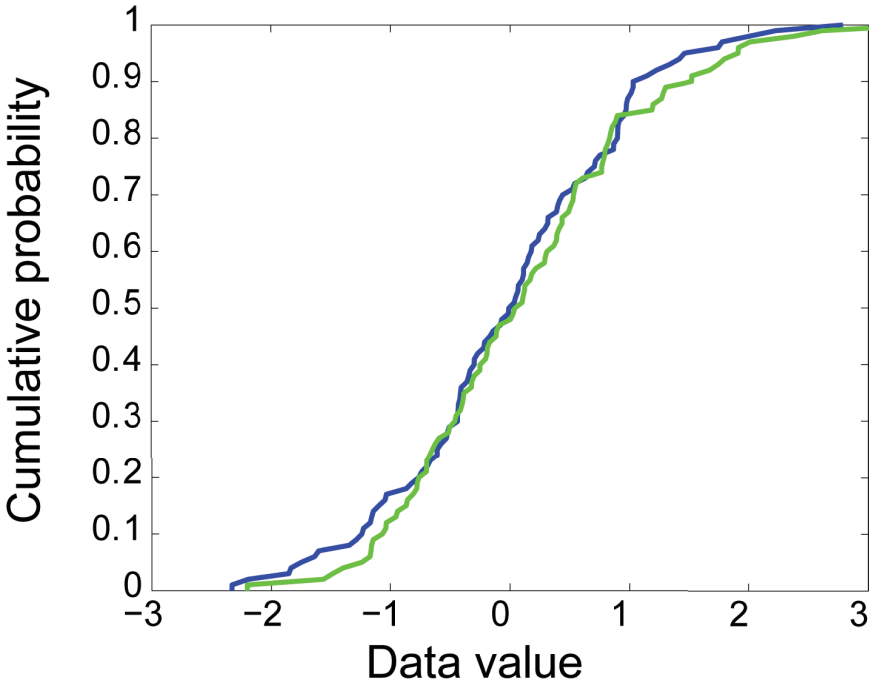
Kolomogrov-Smirnov Test

Normal vs. Uniform



Test statistic = 0.28
p << 0.05

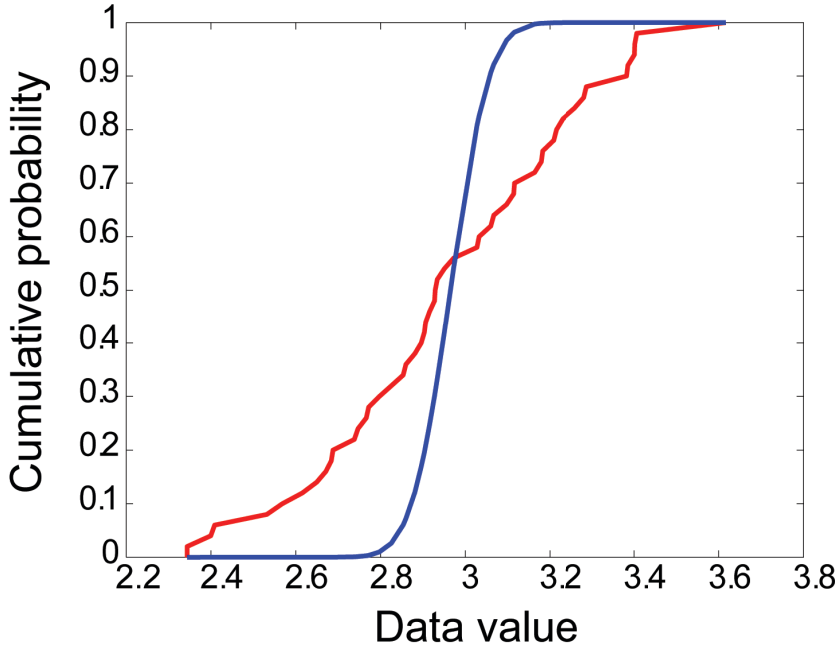
Normal vs. Normal



Test statistic = 0.07
p > 0.05

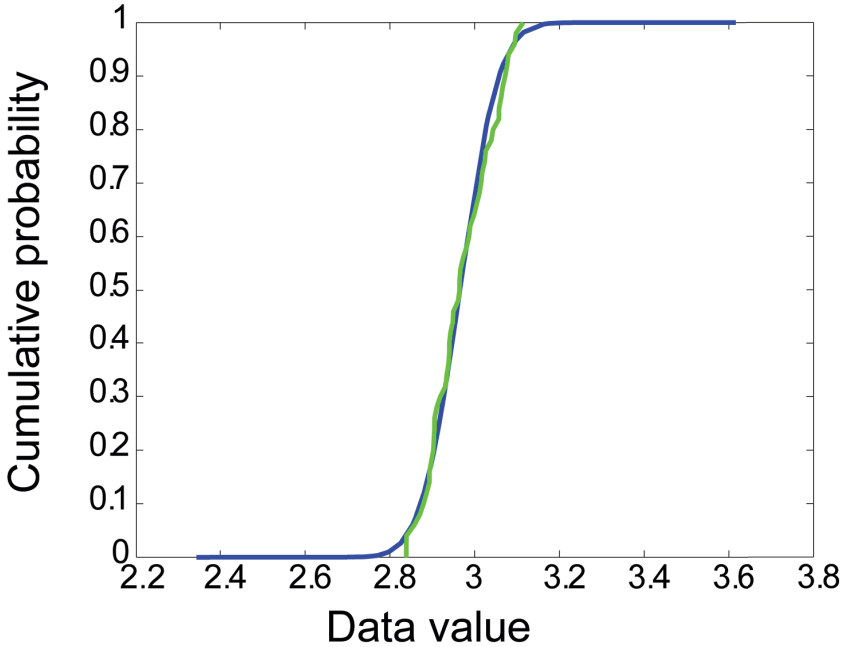
Using KS Test in Stochastic Profiling

Homogenous reference
vs. Heterogenous regulation



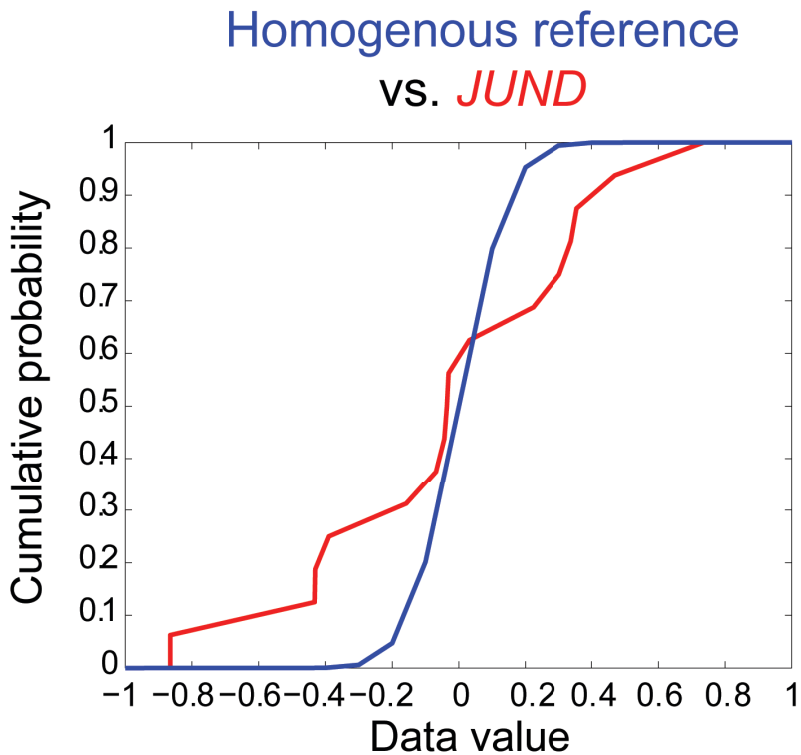
Test statistic = 0.32
p << 0.05

Homogenous reference
vs. Homogenous regulation

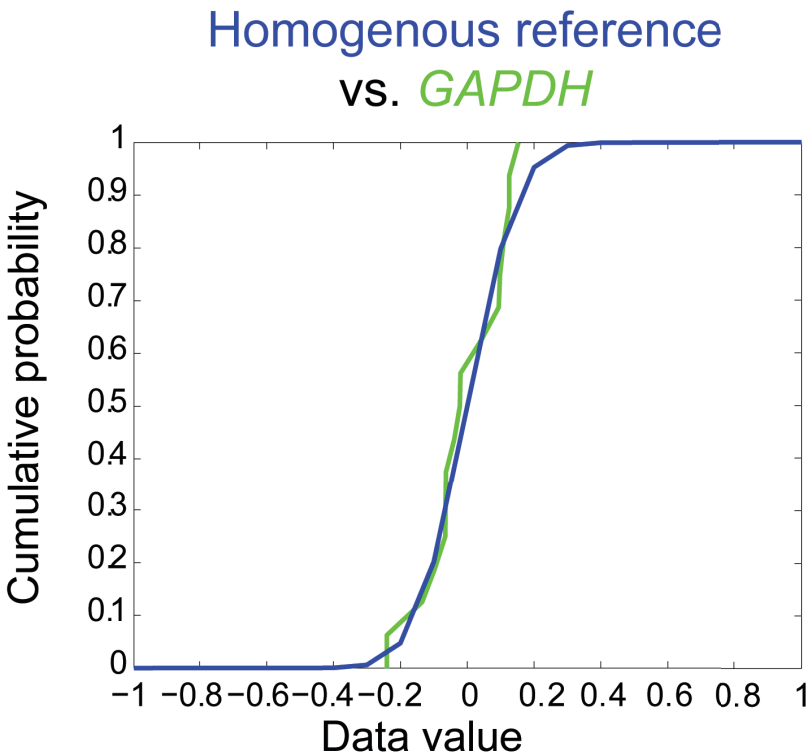


Test statistic = 0.08
p >> 0.05

Using KS Test in Stochastic Profiling



Test statistic = 0.34
$p < 0.05$



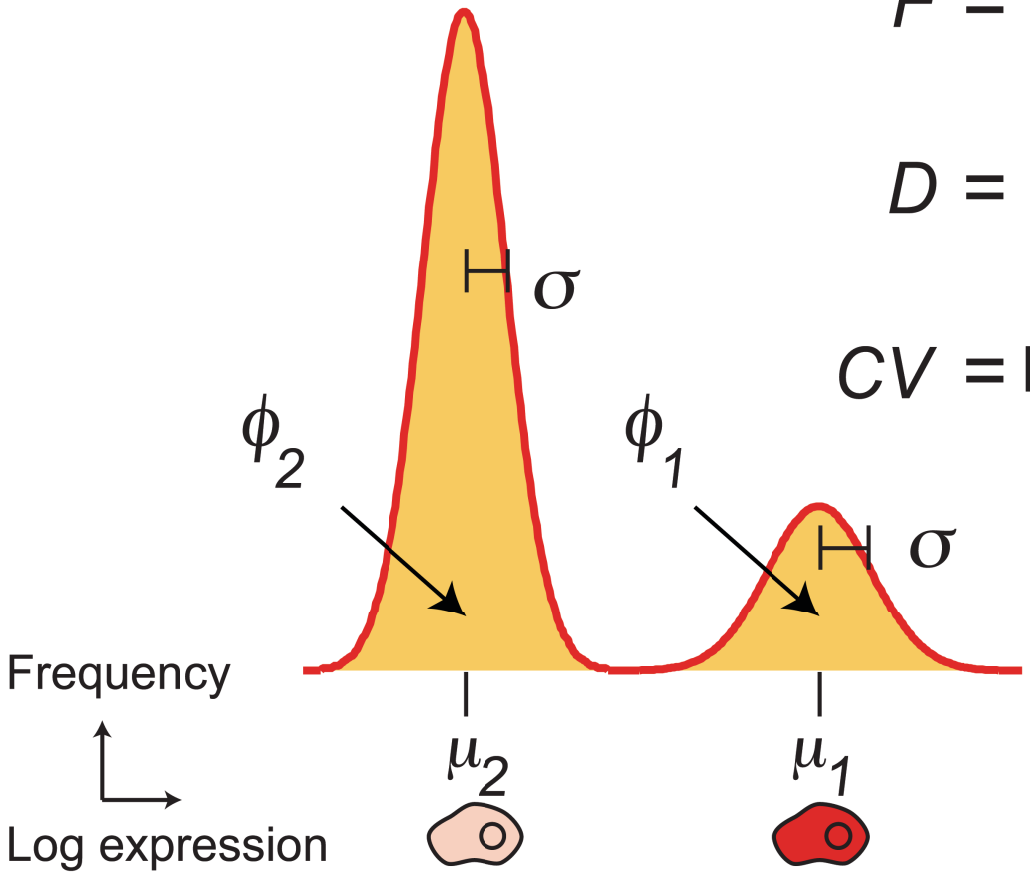
Test statistic = 0.16
$p >> 0.05$

Dichotomous gene expression model

$$F = \frac{\phi_1}{\phi_1 + \phi_2}$$

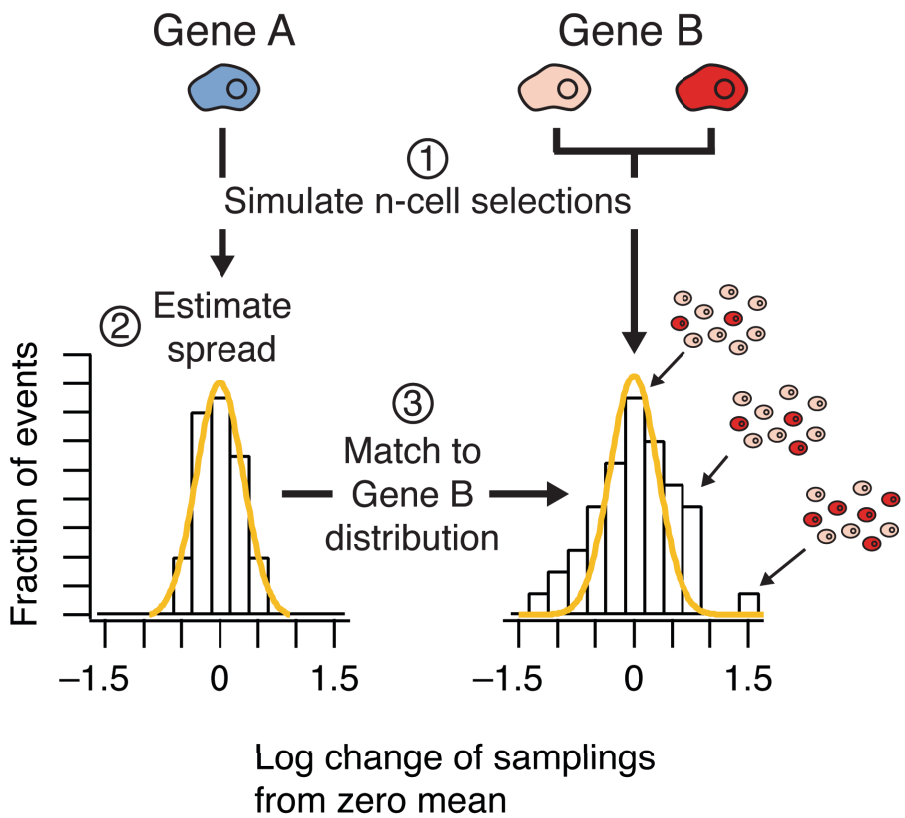
$$D = \frac{\mu_1}{\mu_2}$$

$$CV = \mathrm{Exp}(\sigma^2 + 1)$$



PNAS **111**, E626-35 (2014); *Nat Meth* **7**, 311-317 (2010)

Monte-Carlo simulations of stochastic profiling



Normal distribution with mean = sample mean, CV = reference CV

Monte-Carlo simulation parameters

- Number of cells per sample
- Number of total samples
- Expression fraction (F)
- Fold separation between populations (D)
- Expression spread (CV_{test})
- Reference spread ($CV_{reference}$)

MATLAB Implementation

```
for m =1:NUMBER_OF_TOTAL_SAMPLES
    genemeastemp=0;  % Simulated heterogeneous profiles
    refmeastemp=0; % Reference distribution
    controlmeastemp=0; % Simulated homogenous profiles
    hititer = binornd(NUMBER_OF_CELLS, F); % Randomly select number of
                                            cells from the upper population
    for n=1:hititer
        genemeastemp=genemeastemp+lognrnd(log(D),sqrt(log(CVtest^2+1)));
    end
    for n=hititer+1:NUMBER_OF_CELLS
        genemeastemp=genemeastemp+lognrnd(0,sqrt(log(CVtest^2+1)));
    end
    for n = 1:NUMBER_OF_CELLS
        refmeastemp=refmeastemp+lognrnd(0,sqrt(log(CVref^2+1)));
        controlmeastemp=controlmeastemp+lognrnd(0,sqrt(log(CVtest^2+1)))
    end
    genemeas(m) = genemeastemp;
    refmeas(m) = refmeastemp;
    controlmeas(m) = controlmeastemp;
end
```

MATLAB Implementation

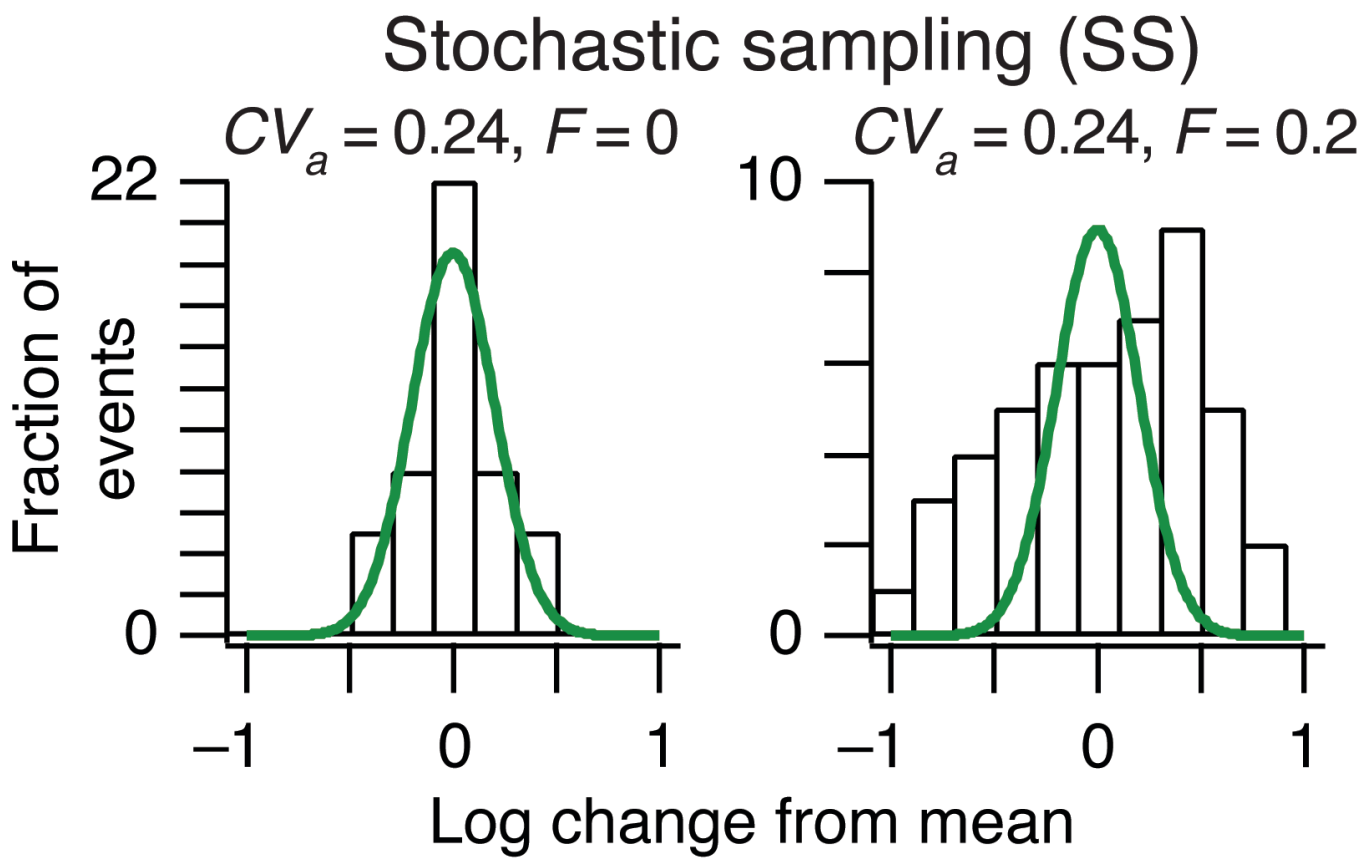
- Test heterogeneous profiles versus a normal distribution with same mean and reference CV

```
[h_heterogeneity,p_heterogeneity,s_heterogeneity]=kstest(log(genemeas)',
[log(genemeas)' normcdf(log(genemeas)',
mean(log(genemeas)),std(log(refmeas)))],0.05,'unequal');
```

- Test control homogenous profiles versus a normal distribution with same mean and reference CV

```
[h_control,p_control, s_control]=kstest(log(controlmeas)',
[log(controlmeas)'
normcdf(log(controlmeas)',mean(log(controlmeas)),std(log(refmeas)))],
0.05,'unequal');
```

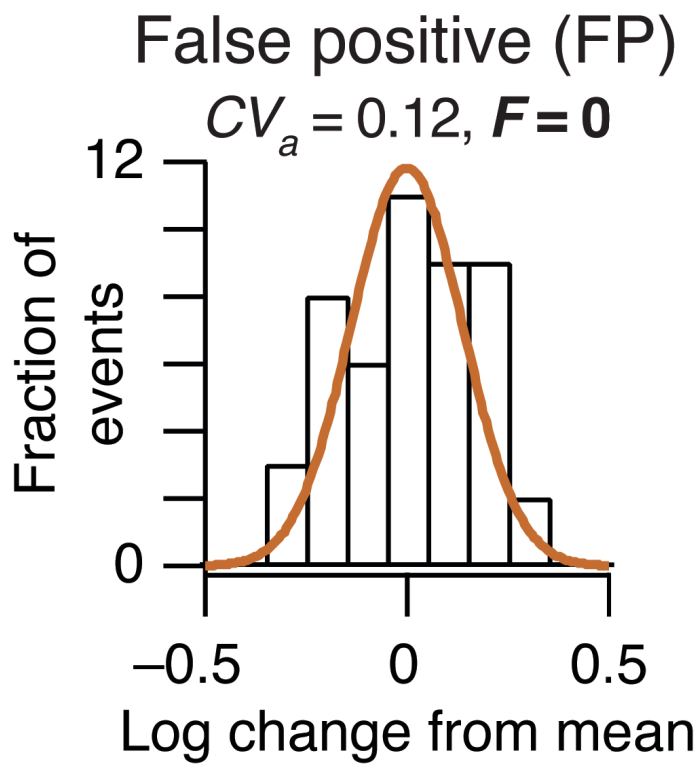
Successful sampling



Nat Meth **7**, 311-317 (2010)

False positive sampling

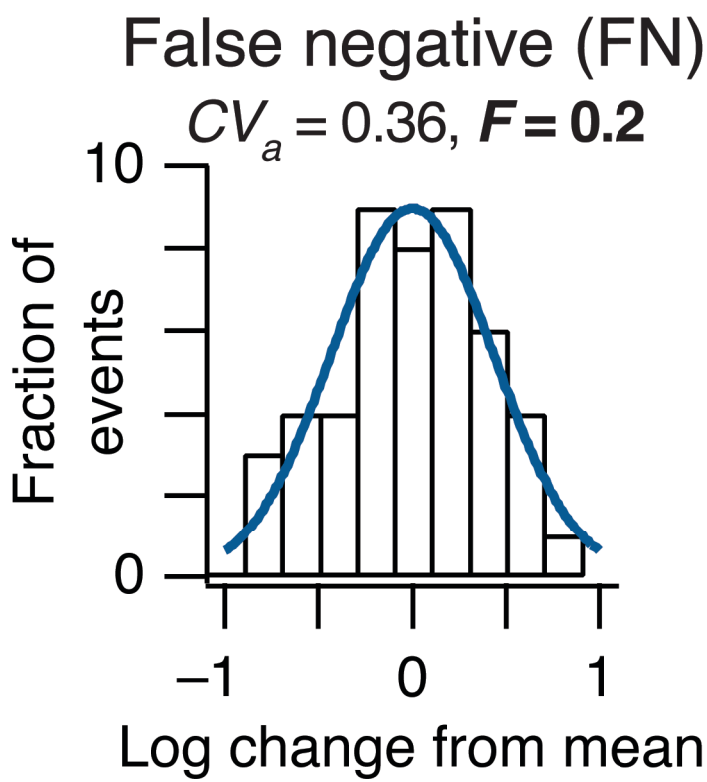
- Homogenous fluctuations are more variable than the reference distribution



Nat Meth **7**, 311-317 (2010)

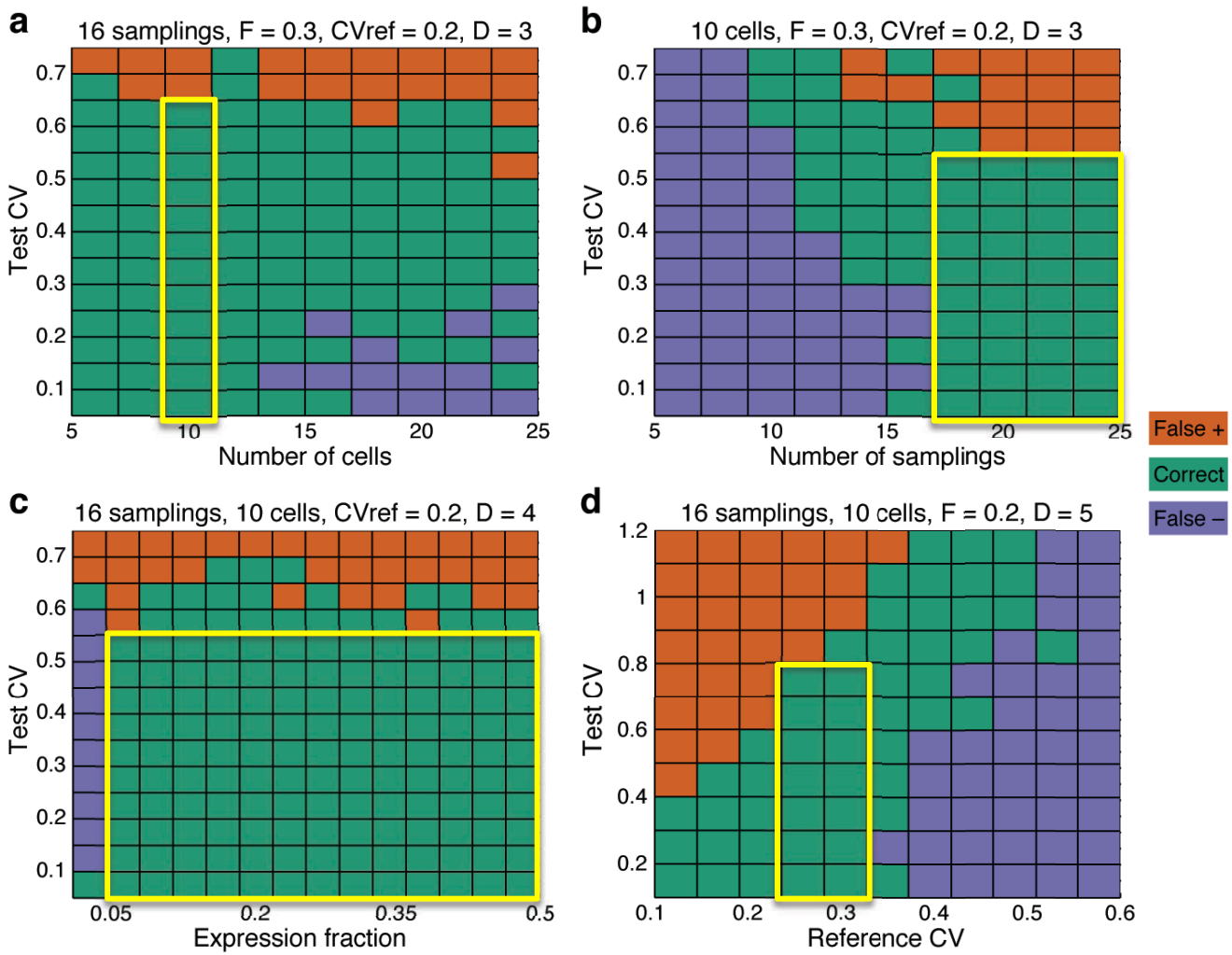
False negative sampling

- Heterogeneous fluctuations are not more variable than the reference distribution



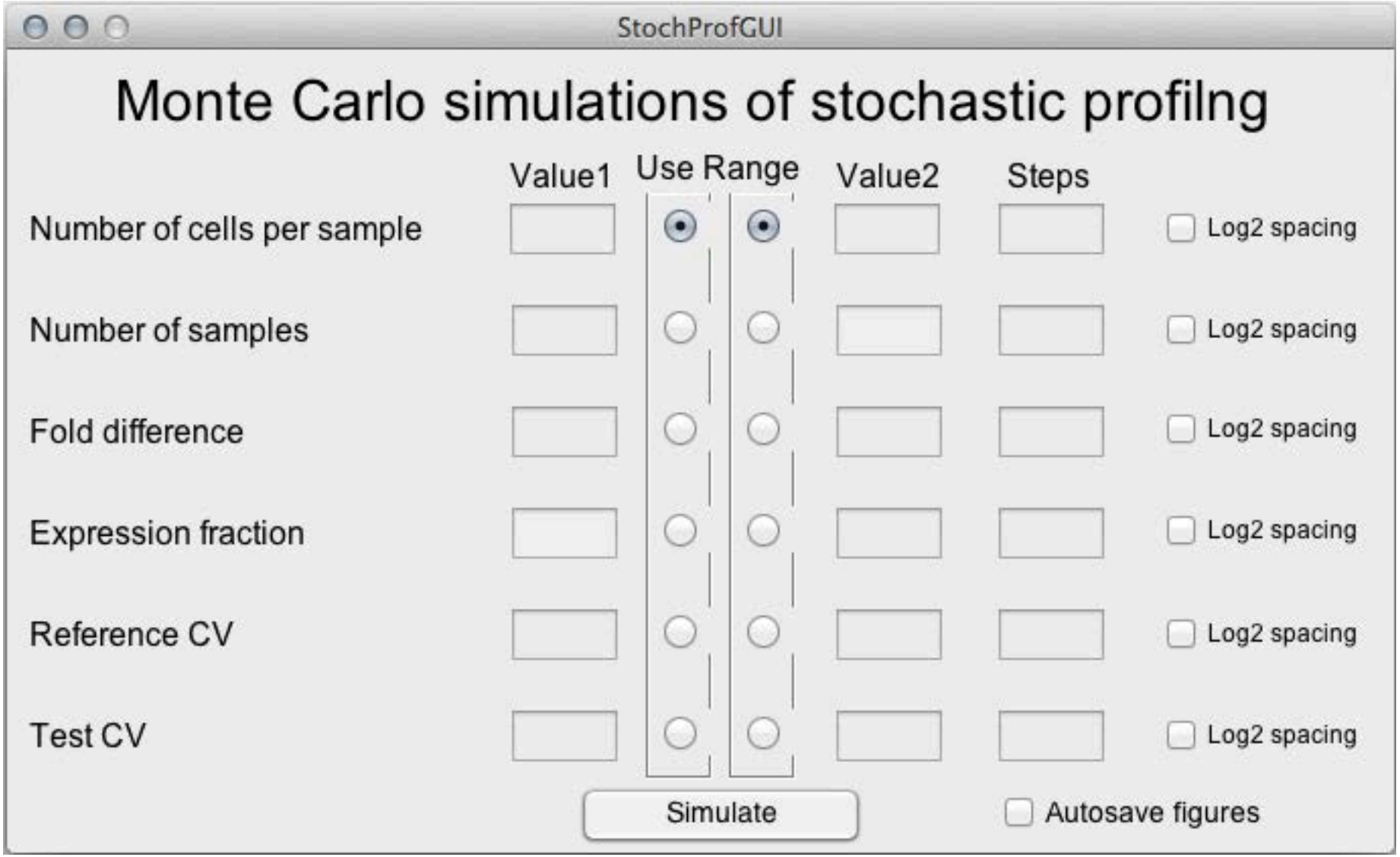
Nat Meth **7**, 311-317 (2010)

Using Monte-Carlo simulations to guide experiments



Nat Protoc **8**, 282-301 (2013)

Monte-Carlo simulation GUI



Exercises and Discussion

- Use the MATLAB GUI **StochProfGUI.m** to identify:
  - What number of cells is ideal for a very-rare (F < 1%) heterogeneity