Processing and analyzing experimental stochastic profiles

So you’ve measured your profiles… now what?

- Single-gene measurements (qPCR)
  - Use same approach from Monte-Carlo simulations to filter heterogeneities
- Global expression profiles (Illumina arrays)
  - Process raw array reads
  - Filter genes that are heterogeneously regulated
- Principle of filtering will be similar with RNA sequencing

Illumina array normalization

- Normalize each array to its average fluorescence

```
for i=1:2:size(ControlSamplings.data,2)-1
    ControlSamplings.data(:,i)=ControlSamplings.data(:,i)* ...
        mean(avgintensity)/mean(ControlSamplings.data(:,i));
    RangeSamplings(:,j)=ControlSamplings.data(:,i);
    j=j+1;
end

j=1;
for i=1:2:size(StochSamplings.data,2)-1
    StochSamplings.data(:,i)=StochSamplings.data(:,i)* ...
        mean(avgintensity)/mean(StochSamplings.data(:,i));
    RangeSamplings2(:,j)=StochSamplings.data(:,i);
    j=j+1;
end
```

Filter reliably detected genes

- Less stringent filtering to account for difficult preamplification protocol
  - Filter median p-value per gene (you decide the stringency)

```
median(pdetect(i,:)) >= pdetectthresh
```

  - Filter based on range of values detected (you decide the range)

```
max(RangeSamplings(i,:)/ min(RangeSamplings(i,:))) > reprodthresh
```

  - Ensure that each array has a non-zero fluorescence for a given gene

```
geomean(RangeSamplings(i,:)) == 0
```

Renormalize detected genes to array median

```
for i=1:2:size(StochSamplings.data,2)-1
    NormStochSamplings(:,k)=StochSamplings.data(:,i)* ...
        median(medianintensity)/median(StochSamplings.data(:,i));
    k=k+1;
end

k=1;
for i=1:2:size(ControlSamplings.data,2)-1
    NormControlSamplings(:,k)=ControlSamplings.data(:,i)* ...
        median(medianintensity)/
    median(ControlSamplings.data(:,i));
    k=k+1;
end
```

“Z-score” profiles

- Normalize each array to its geometric mean, normalize each gene to its geometric mean
  - Dividing by average in normal space
- Log transform the data
  - Brings the data into normal space
- Samples equal to mean will be 0, above > 0, below < 0

MATLAB implementation

```
for i=1:genlength
    Stochscalematrix(i,:)=ones(1,Stochsamplength).*geomean(NormStochSamplings);

Controlscalematrix(i,:)=ones(1,Controlsamplength).*geomean(NormControlSamplings);
end
Stochscaledsamplingstemp=NormStochSamplings./Stochscalematrix;
Controlscaledsamplingstemp=NormControlSamplings./Controlscalematrix;

for i=1:Stochsamplength
    Stochscalematrix2(:,i)=ones(genlength,
1).*geomean(Stochscaledsamplingstemp')';
end
for i=1:Controlsamplength
    Controlscalematrix2(:,i)=ones(genlength,
1).*geomean(Controlscaledsamplingstemp')';
end
ScaledStochSamplings=Stochscaledsamplingstemp./Stochscalematrix2;
ScaledControlSamplings=Controlscaledsamplingstemp./Controlscalematrix2;

% Log transform and extract genes with significant sampling variations
% based on F test
LogControlSamplings=log(ScaledControlSamplings);
LogStochSamplings=log(ScaledStochSamplings);
```

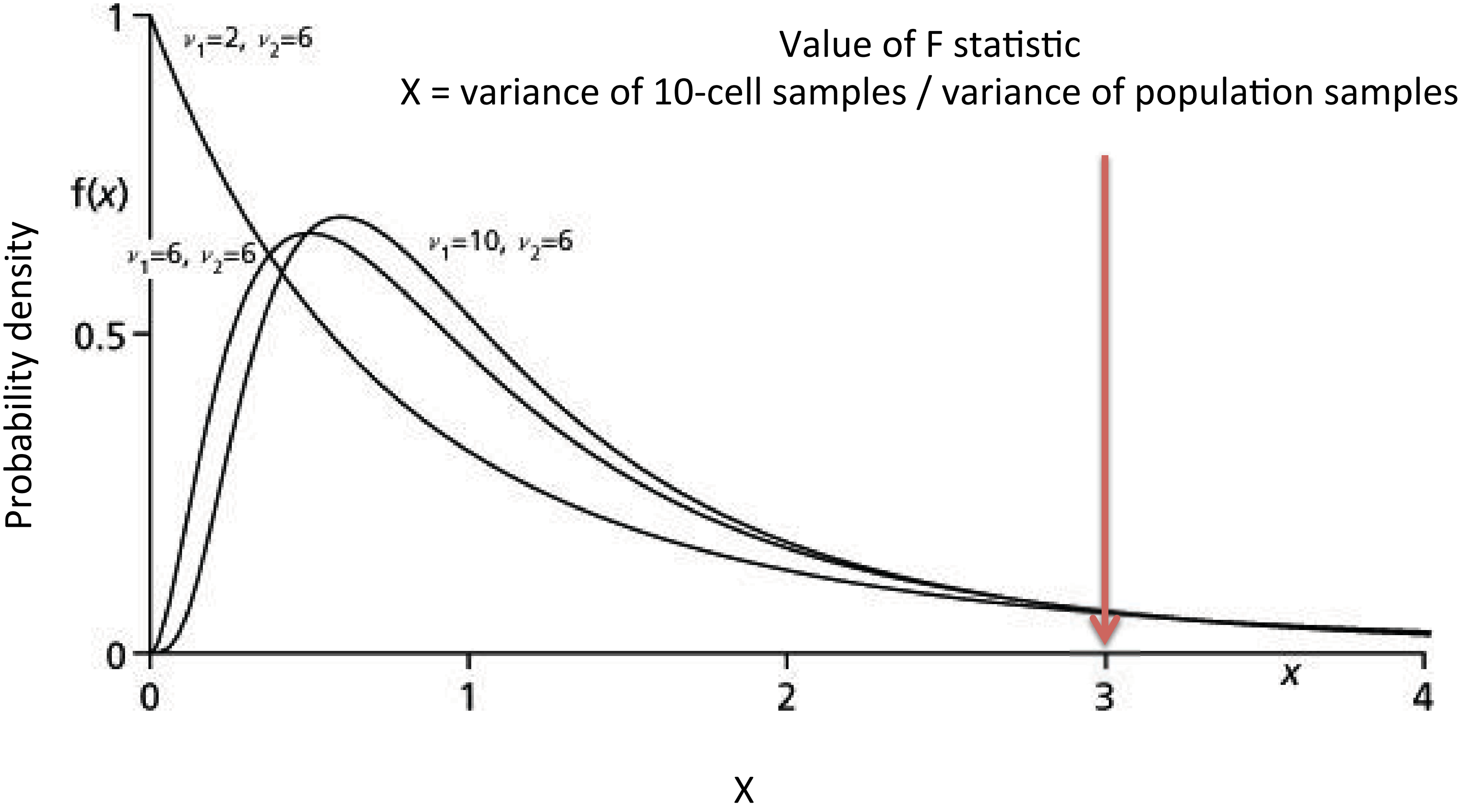
Identifying significant fluctuations

- Need to separate true biological variation from technical, sample variation
- Compare the variance per gene in stochastic samples to control, population-level samples

F-test

- Statistical test to determine whether two samples have different variation
  - Null hypothesis: variances are equal
- F statistic is the ratio of the two sample variations
- Degrees of freedom are the number of samples in each data set

F-distribution

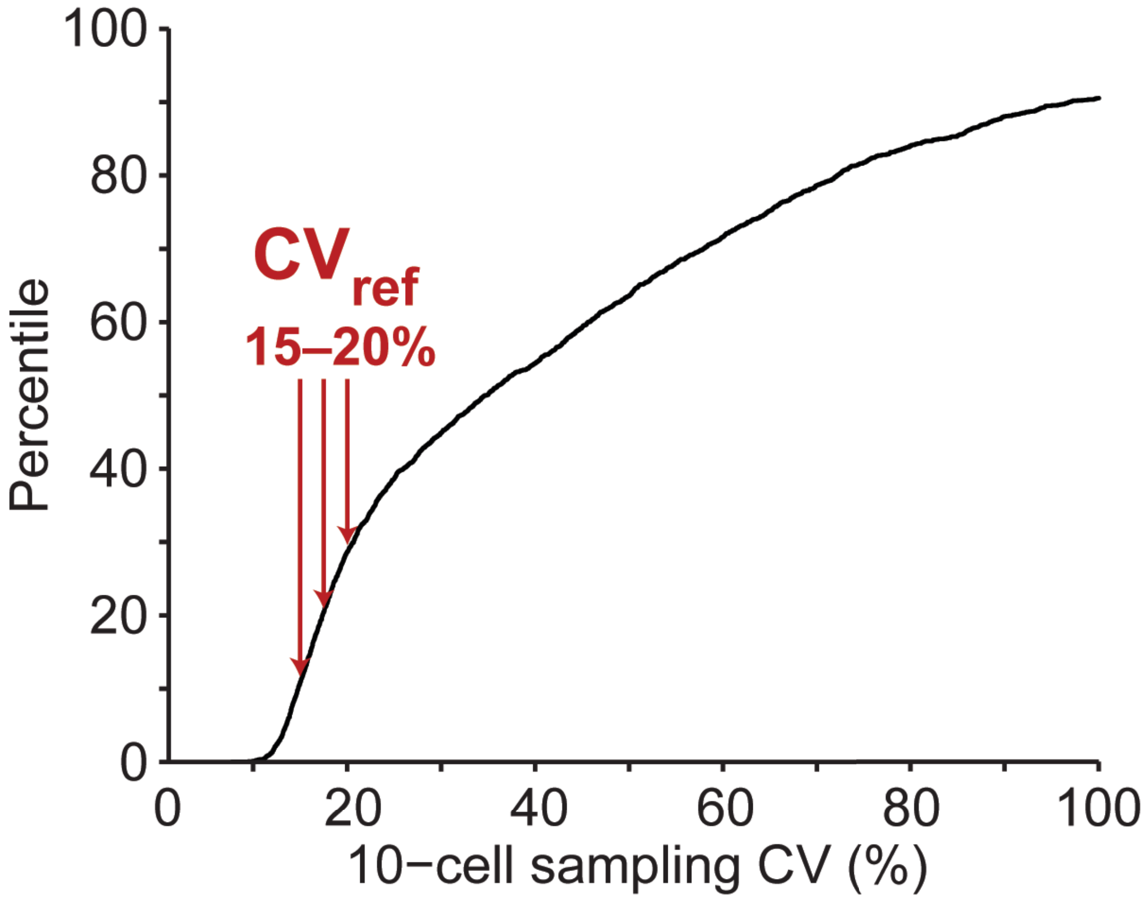


Value of F statistic
X = variance of 10-cell samples / variance of population samples

MATLAB implementation

```
for i=genelist:-1:1
    psampvar(i)=1-fcdf(var(LogStochSamplings(i,:))/
var(LogControlSamplings(i,:)), ...
        size(LogStochSamplings,2)-1,size(LogControlSamplings,2)-1);
end
sortpsampvar = sort(psampvar);
i=1;
while sortpsampvar(i) < i/length(psampvar)*FDRval
    i=i+1;
end
pcrit=sortpsampvar(i-1);
for i=genelist:-1:1
    if psampvar(i) > pcrit
        LogControlSamplings(i,:)=[];
        LogStochSamplings(i,:)=[];
        GeneNames(i)=[];
    end
end
```

Determine reference CV



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

MATLAB filtering of global profiles

```
[h,p(i)]=kstest(SortedStochSamplings(i,:)’,
    [SortedStochSamplings(i,:)'
    normcdf(SortedStochSamplings(i,:)',0,sqrt(log(CVref^2+1)))],
    0.05,'unequal');
```

- p < p_critical – gene is scored as heterogeneously regulated

Processing RNA sequencing data

- Read normalization – divide raw counts for each transcript by the total reads per lane, multiply this number by 1 million (transcripts per million, TPM)
- Working now to test if array heterogeneity filtering is appropriate and valid for RNA sequencing results

Exercise and Discussion

- Using the data in **Sampling_example.txt** and **Control_example.txt**, pre-filtering code **StochProfMicroarrayFilt.m**, analysis code **StochProfAnalysis.m** to observe the effect of using different reference CVs on the final geneset