Use of Stratified Dirichlet to Estimate Multinomial Parameters in the Evidence of Skewness, with Application to SAGE

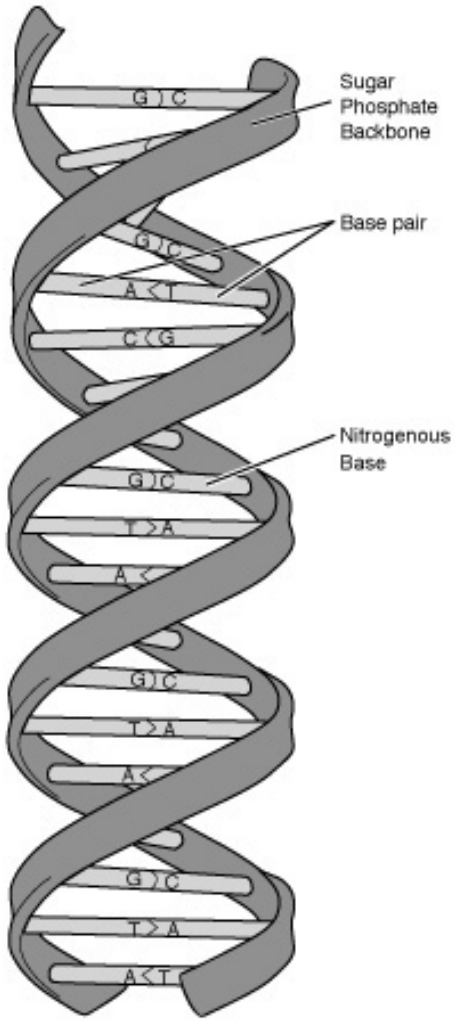
Jeffrey S. Morris

University of Texas, MD Anderson Cancer Center

Outline

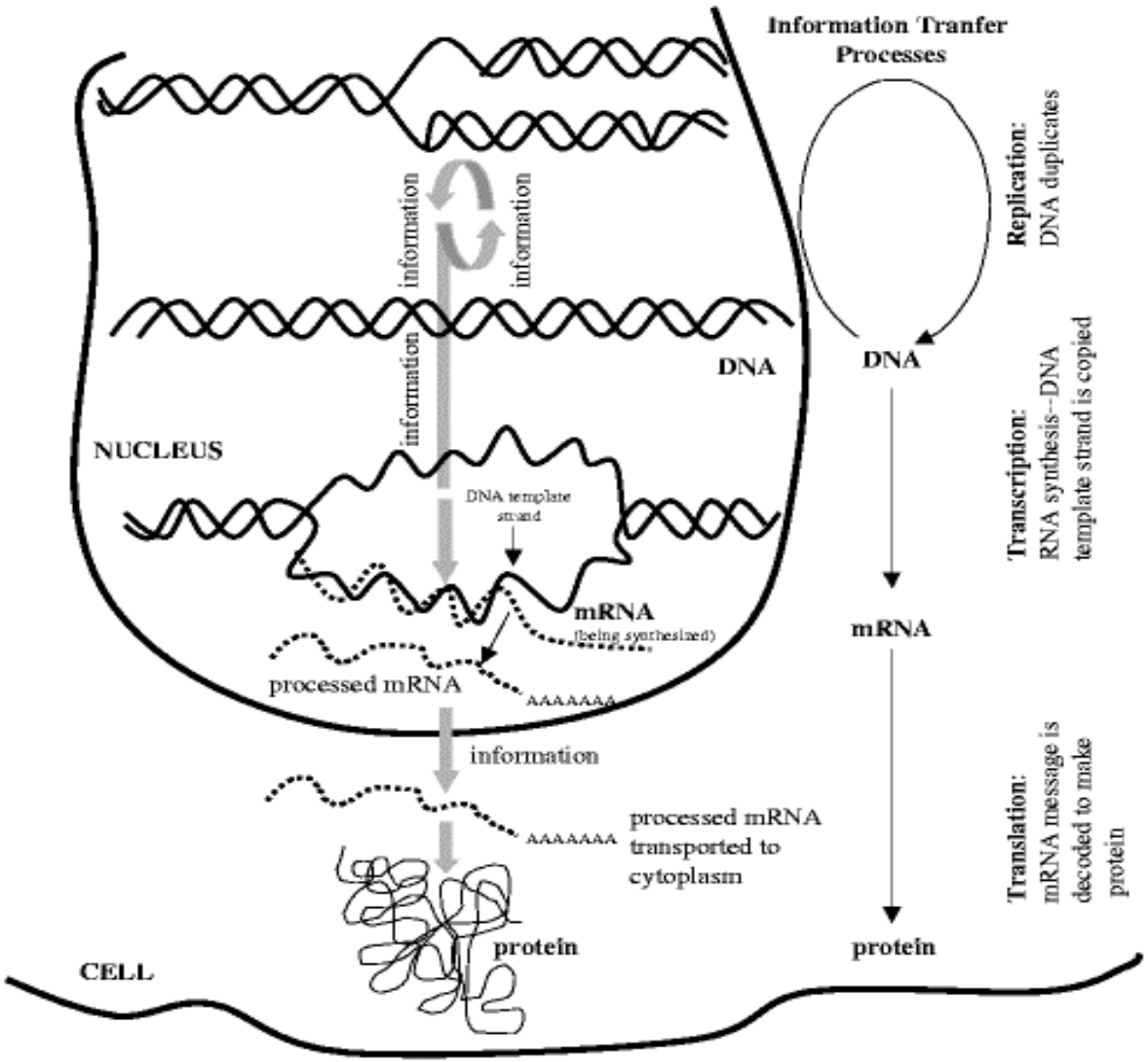
- Analysis of Gene Expression
- Statistical Model for SAGE data
- MLEs and Standard Bayesian Estimators
- Our method: Stratified Dirichlet Prior
- Simulation Study
- Conclusion

DNA



- DNA: blueprint for living organisms
- Double Helix
- “Rongs” of ladder are pairs of matching bases
- A=adenine $\Leftrightarrow$ T=thymine
  G=guanine $\Leftrightarrow$ C=cytosine
- Gene = sequence of DNA that codes for a protein

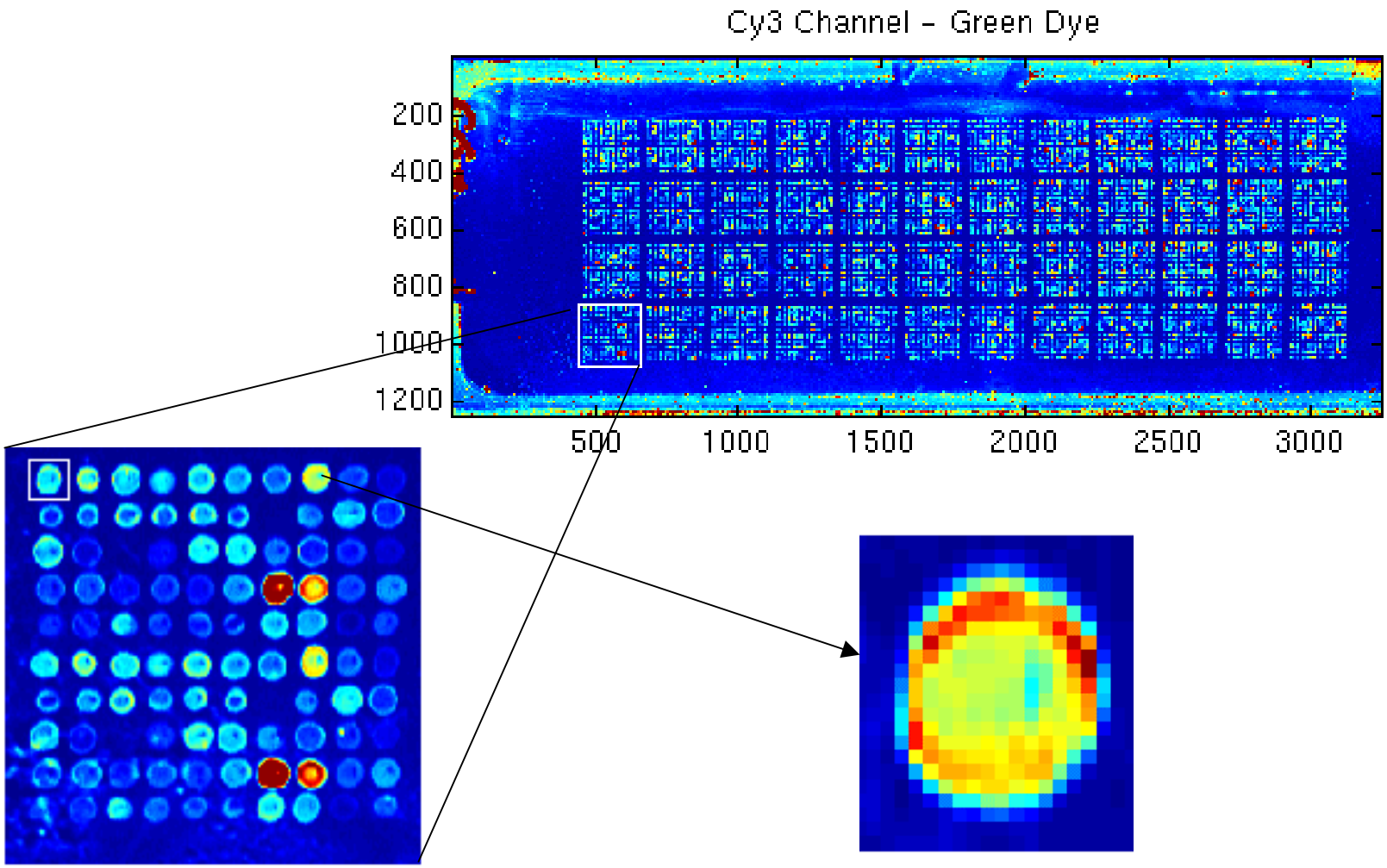
Gene Expression



cDNA Microarrays

- Method to quantify gene expression in samples of cells.
- cDNA for large number of specific genes spotted in an array on a glass slide.
- Fluorescent dye attached to mRNA transcripts in tissue samples.
- Tissue sample run over array, mRNA ‘sticks’ to cDNA.
- Gene expression levels quantified as intensity of fluorescent dye, obtained by shining laser over array.

cDNA Microarrays



SAGE

- Serial Analysis of Gene Expression
- Another method to quantify gene expression levels in samples of cells
- Open system
  - Can potentially reveal expression levels of all genes: “unbiased” and “comprehensive”
  - Microarrays are closed, since they only tell you about the genes spotted on the array

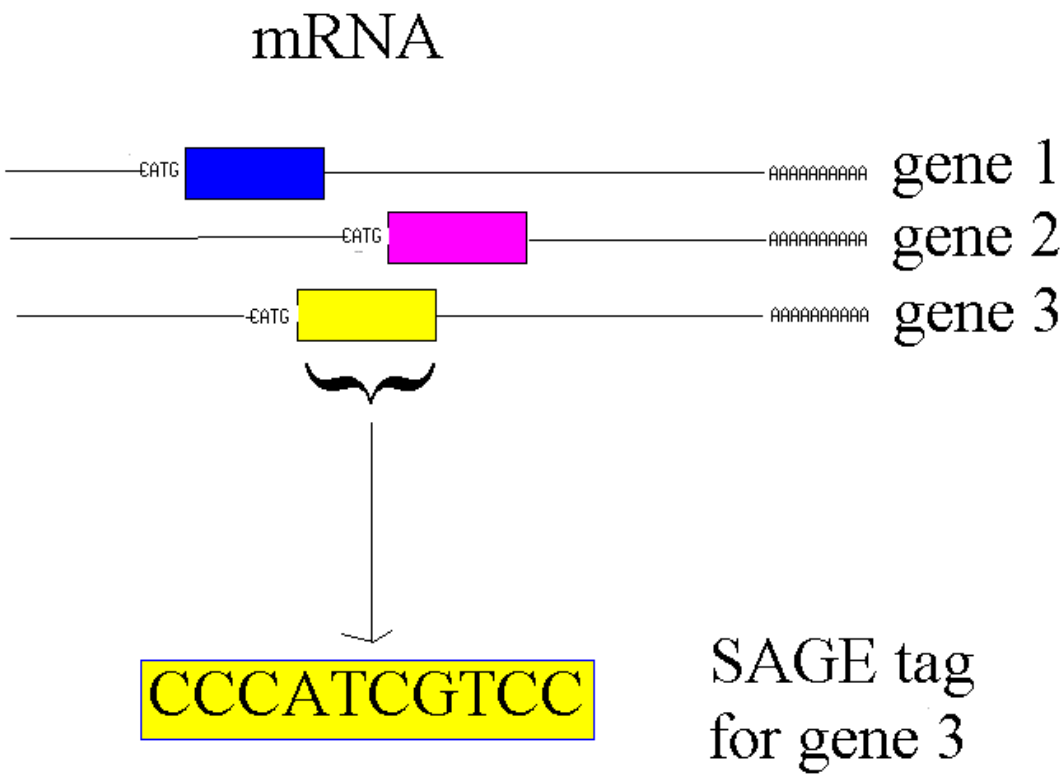
Summary of SAGE Procedure

- Sample of *n* mRNA transcripts selected from tissue (often *n* ~ 10,000 - 100,000)

- Specific 10-base region sequenced for each transcript -- called a ***tag***

- Counts of each unique tag measures the expression level of corresponding gene.

SAGE tags


mRNA
CATG
AAAAAAAAAA
gene 1
CATG
AAAAAAAAAA
gene 2
CATG
AAAAAAAAAA
gene 3
CCCATCGTCC
SAGE tag
for gene 3

SAGE Data

TAG	COUNT		TAG	COUNT		TAG	COUNT
CCCATCGTCC	1286		CACTACTCAC	245		TTCACTGTGA	150
CCTCCAGCTA	715		ACTAACACCC	229		ACGCAGGGAG	142
CTAAGACTTC	559		AGCCCTACAA	222		TGCTCCTACC	140
GCCCAGGTCA	519		ACTTTTTCAA	217		CAAACCATCC	140
CACCTAATTG	469		GCCGGGTGGG	207		CCCCCTGGAT	136
CCTGTAATCC	448		GACATCAAGT	198		ATTGGAGTGC	136
TTCATACACC	400		ATCGTGGCGG	193		GCAGGGCCTC	128
ACATTGGGTG	377		GACCCAAGAT	190		CCGCTGCACT	127
GTGAAACCCC	359		GTGAAACCCT	188		GGAAAACAGA	119
CCACTGCACT	359		CTGGCCCTCG	186		TCACCGGTCA	118
TGATTTCACT	358		GCTTTATTTG	185		GTGCACTGAG	118
ACCCTTGGCC	344		CTAGCCTCAC	172		CCTCAGGATA	114
ATTTGAGAAG	320		GCGAAACCCT	167		CTCATAAGGA	113
GTGACCACGG	294		AAAACATTCT	161		ATCATGGGGA	110

Statistical Model for SAGE Data

- X_i = counts for gene i

$$\underline{\mathbf{X}} = (X_1, X_2, \ldots, X_k)^T$$

  – k = total # of expressed genes in biological sample (~20,000-100,000)

- $\underline{\mathbf{X}} \sim$ Multinomial$(n, \underline{\pi})$, $\underline{\pi} = (\pi_1, \pi_2, \ldots, \pi_k)^{\mathrm{T}}$

$$f(\mathbf{X}) = n! \prod_{i=1}^{k} \frac{\pi_i^{X_i}}{X_i!}$$

- π_i=relative expression of gene i in population

$$\textbf{Relative Frequency Constraint}: \sum_{i=1}^{k} \pi_{\mathrm{i}} = 1$$

Maximum Likelihood Estimation

- Maximum likelihood estimators of π_i:

$$\hat{\pi}_{i,\mathrm{MLE}} = X_i / n$$

- Optimality Properties:
  - Unbiased, nothing beats it for large enough n
- Optimal for SAGE data?
  - n ~10,000 to 100,000 -- large sample?
  - $n \sim k$, so sample size not *really* large

Skewness in SAGE data

From Colon Cancer SAGE Libraries
(Velculescu, et al. 1999)

Copies/cell	% of genes	% of mass
≤5	89.9%	23%
5-50	9.2%	30%
50-500	0.8%	27%
500-5000	0.1%	20%

- Small number of "abundant" genes
- Large number of "scarce" genes

- Distribution of multinomial probabilities π strongly skewed right
- Note: Many genes will be missing in the SAGE sample (i.e. will have X_i=0)

Maximum Likelihood Estimation

- SAGE: Many missing genes (with $X_i=0$)
  - We ***know*** that, for these genes:

$$\boxed{\hat{\pi}_{i,\mathrm{MLE}} = 0 < \pi_i}$$

- Relative frequency constraint ($\Sigma\pi_i=1$) implies:

$$\boxed{\sum_{i:X_i>0} \hat{\pi}_{i,MLE} > \sum_{i:X_i>0} \pi_i}$$

- In given data set, MLE:
  - Underestimates π_i for genes with zero counts
  - Overestimates π_i for genes small nonzero counts

Simple Example

- Population:
  - One abundant gene with $\pi_0 = 0.50$
  - Fifty scarce genes with $\pi_i = 0.01$ for i=1, ..., 50
  - Sample n=20 mRNA transcripts
- For scarce genes:
  - If $X_i = 0$ $\Rightarrow$ $\hat{\pi}_{i,\mathrm{MLE}} = 0$
  - If $X_i = 1$ $\Rightarrow$ $\hat{\pi}_{i,\mathrm{MLE}} = 0.05$

Bayesian Estimation

- Likelihood: $f(\underline{\mathbf{X}}|\underline{\pi})$
- Prior distribution: $f(\underline{\pi})$
- Bayes Theorem gives Posterior distribution:
  - $f(\underline{\pi}|\underline{\mathbf{X}}) = f(\underline{\mathbf{X}}|\underline{\pi}) * f(\underline{\pi})/f(\underline{\mathbf{X}})$
- Posterior mean $E(\underline{\pi}|\underline{\mathbf{X}})$ often used as estimate for $\underline{\pi}$
- **Conjugate prior**: $f(\underline{\pi})$ and $f(\underline{\pi}|\underline{\mathbf{X}})$ have same distributional form.

Simple Dirichlet Prior

- Dirichlet prior conjugate for Multinomial

$$\underline{\pi} \sim \text{Dirichlet}(\theta, \ldots, \theta)$$

$$\underline{\pi}|\underline{\mathbf{X}} \sim \text{Dirichlet}(\theta + X_1, \ldots, \theta + X_k)$$

- Posterior mean:

$$\hat{\pi}_{i,\text{DIR}} = \mathrm{E}\left(\pi_i \mid \mathbf{X}\right) = \frac{X_i + \theta}{n + k * \theta}$$

$$= \left(\frac{n}{n + k\theta}\right) * \frac{X_i}{n} + \left(\frac{k\theta}{n + k\theta}\right) * \frac{1}{k}$$

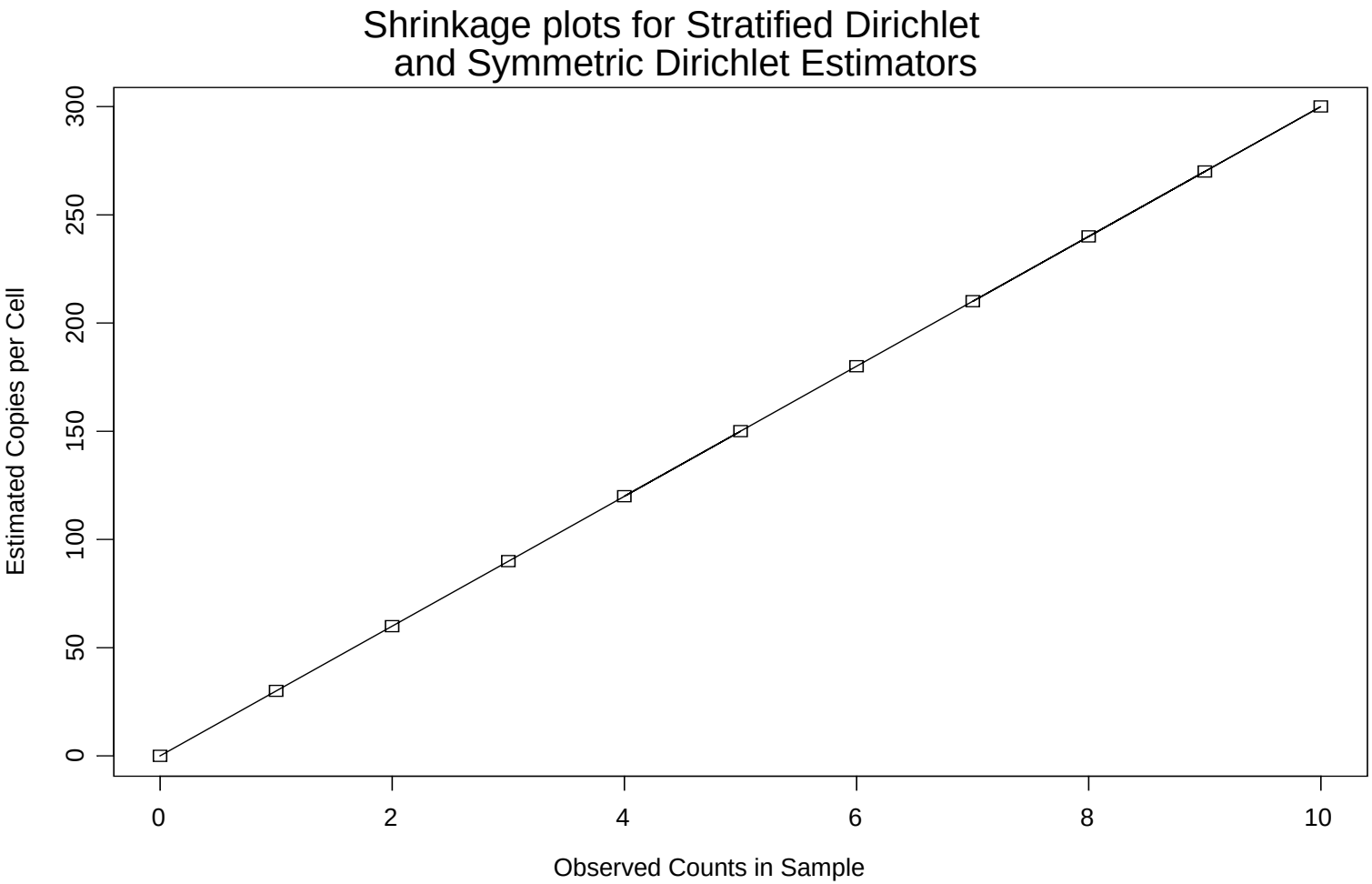
Example Revisited

$\hat{\pi}_{i,\mathrm{DIR}}$: Bayesian estimator assuming Dir(1) prior

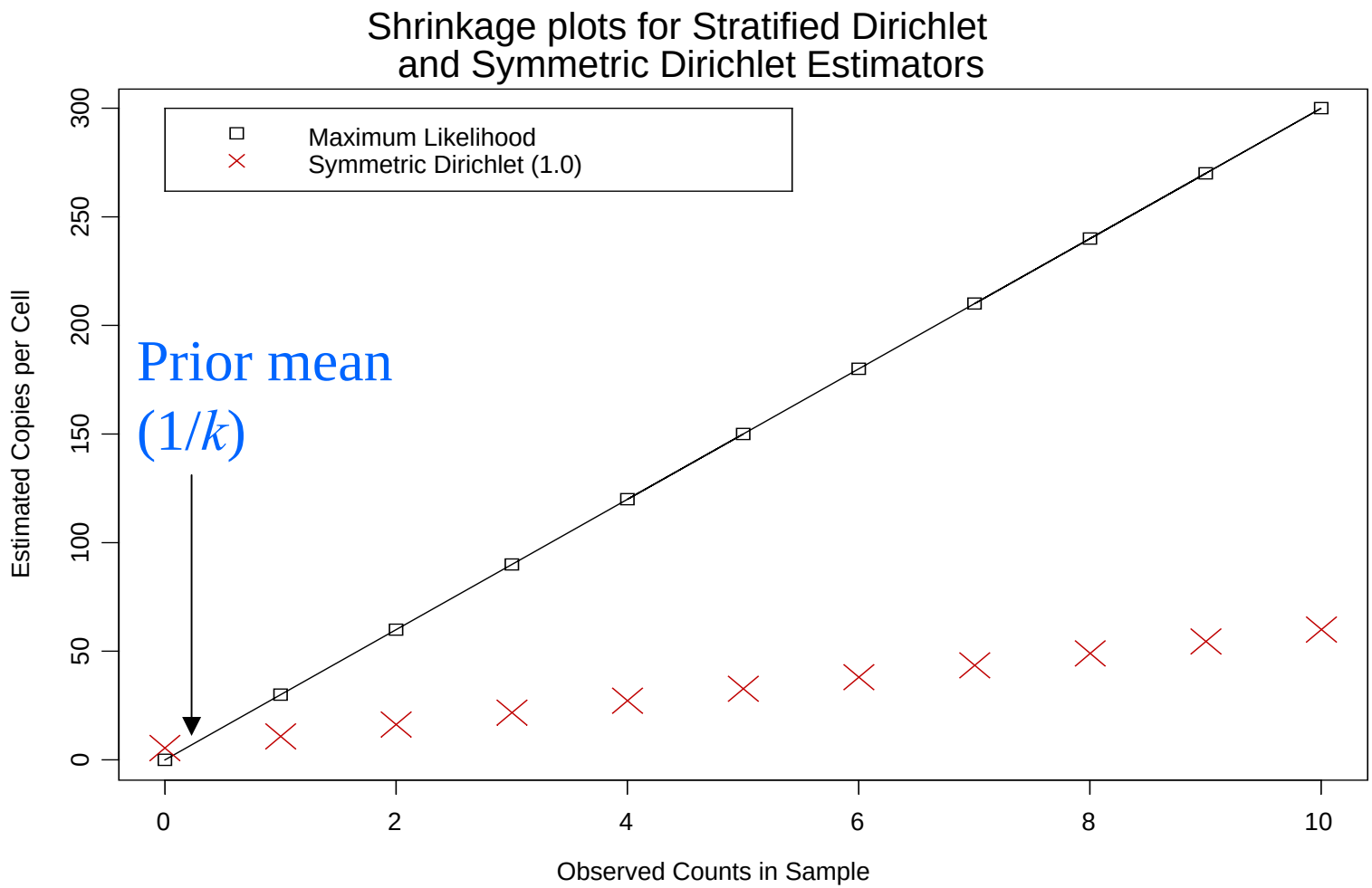
	$\hat{\pi}_{i,\mathrm{DIR}}$	$\hat{\pi}_{i,\mathrm{MLE}}$
• Scarce Gene *i*:		
$X_i = 0$:	$1/71 = \mathbf{0.014}$	$0/20 = \mathbf{0.000}$
$X_i = 1$:	$2/71 = \mathbf{0.028}$	$1/20 = \mathbf{0.050}$
• Abundant Gene:		
$X_0 = 10$:	$11/71 = \mathbf{0.15}$	$10/20 = \mathbf{0.50}$

- Improves estimation of scarce species
- Induces severe bias in abundant species

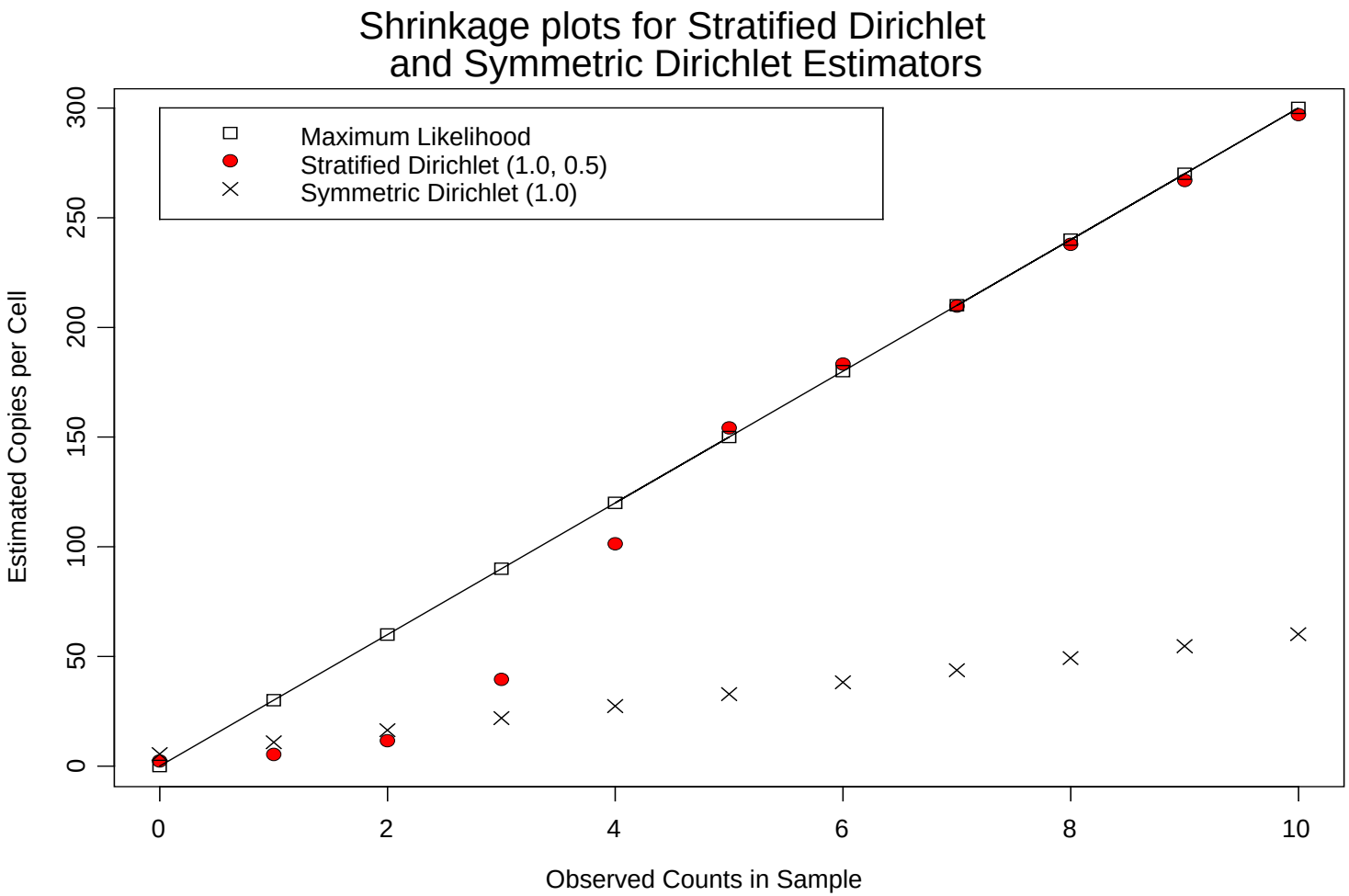
Shrinkage Curves



Shrinkage Curves



Shrinkage Curves



Stratified Dirichlet Prior

- **Idea**: Partition genes into 2 classes:
  - ***Scarce*** and ***Abundant***, different priors for each
- Tricky to do -- must introduce new parameters:
  - λ_i = indicator that gene *i* is in abundant class
  - $\pi^* = \Sigma\, \lambda_i\, \pi_i$ = total mass for abundant class
  - q_i = proportion of abundant/scarce mass for gene *i*
    - Abundant (λ_i=1): $q_i = \pi_i/\pi^*$
    - Scarce (λ_i=0): $q_i = \pi_i/(1-\pi^*)$
- NOTE: $\pi_i = q_i\pi^*$ or $q_i(1-\pi^*)$

Stratified Dirichlet Prior

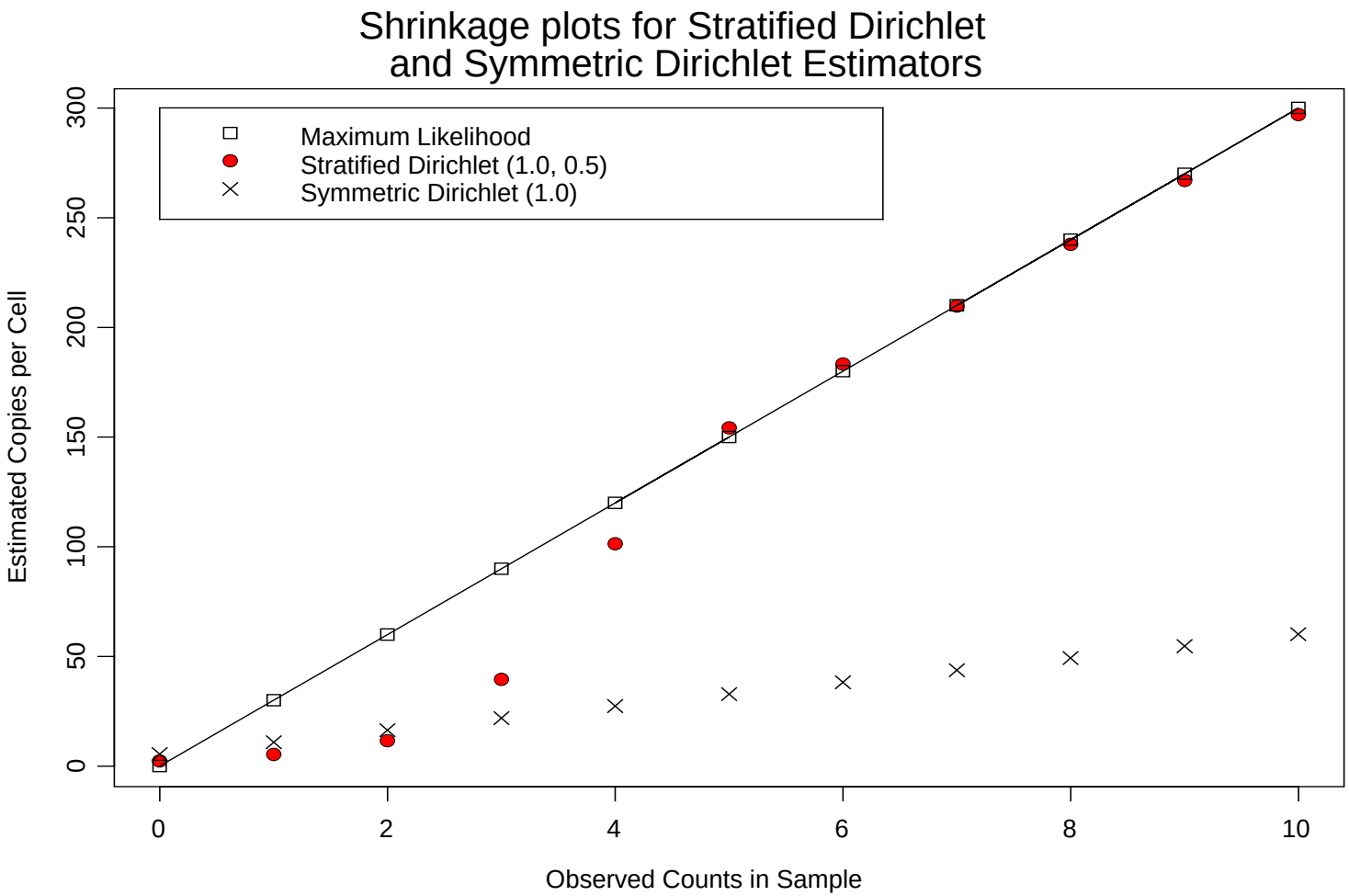
- $\underline{q}_A$ = Dirichlet ($\theta_{A,\ \ldots,}\ \theta_A$)
  $\underline{q}_S$ = Dirichlet ($\theta_{S,\ \ldots,}\ \theta_S$)
- $\lambda_i \sim$ Bernoulli (P)
- $P \sim$ Beta (a_P, b_P)
- $\pi^* \sim$ Beta (a_{π^*}, b_{π^*})
- NOTE: Prior imposes nonlinear shrinkage
  - Steals from ‘scarce’ species, leaves ‘abundant’ alone.

Fitting the Model

- Prior not conjugate, posterior distribution not available in closed form.
- Must simulate from posterior distribution using MCMC methods.
- Generate 10,000 samples $\underline{\pi}_j$ from f($\underline{\pi}$|$\underline{\mathbf{X}}$)

$$\hat{\pi}_{i,\mathrm{STRAT}} = \text{mean value of } \pi_i \text{ taken over } 10{,}000 \text{ samples}$$

Shrinkage Curves



Simulation Study: SAGE data

- Simulation done to compare 3 methods performance in estimating $\underline{\pi}$ in SAGE.
- SAGE samples generated from "true population" of mRNA transcripts (k=45,984)
  - 100 datasets with n = 10,000
  - 100 datasets with n = 50,000
- Since true $\underline{\pi}$ known, can evaluate estimators

Comparing Estimators

- Squared error for gene *i*, data set *j*:

$$\boxed{\mathrm{SE}_{ij} = \left(\hat{\pi}_{ij} - \pi_i\right)^2}$$

- Mean squared error (MSE) for gene *i*:

$$\boxed{\mathrm{MSE}_i = \frac{1}{100}\sum_{j=1}^{100}\left(\hat{\pi}_{ij} - \pi_i\right)^2}$$

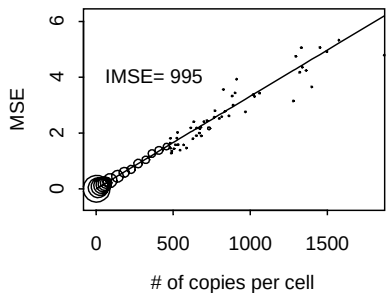
- Integrated MSE:

$$\boxed{\mathrm{IMSE} = \sum_{i=1}^{k}\mathrm{MSE}_i}$$

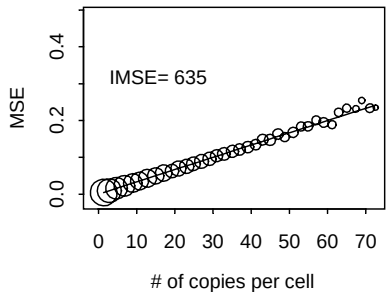
Simulation Results (n=*10,000*)

MLE

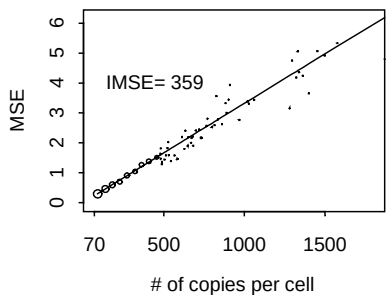
All Species



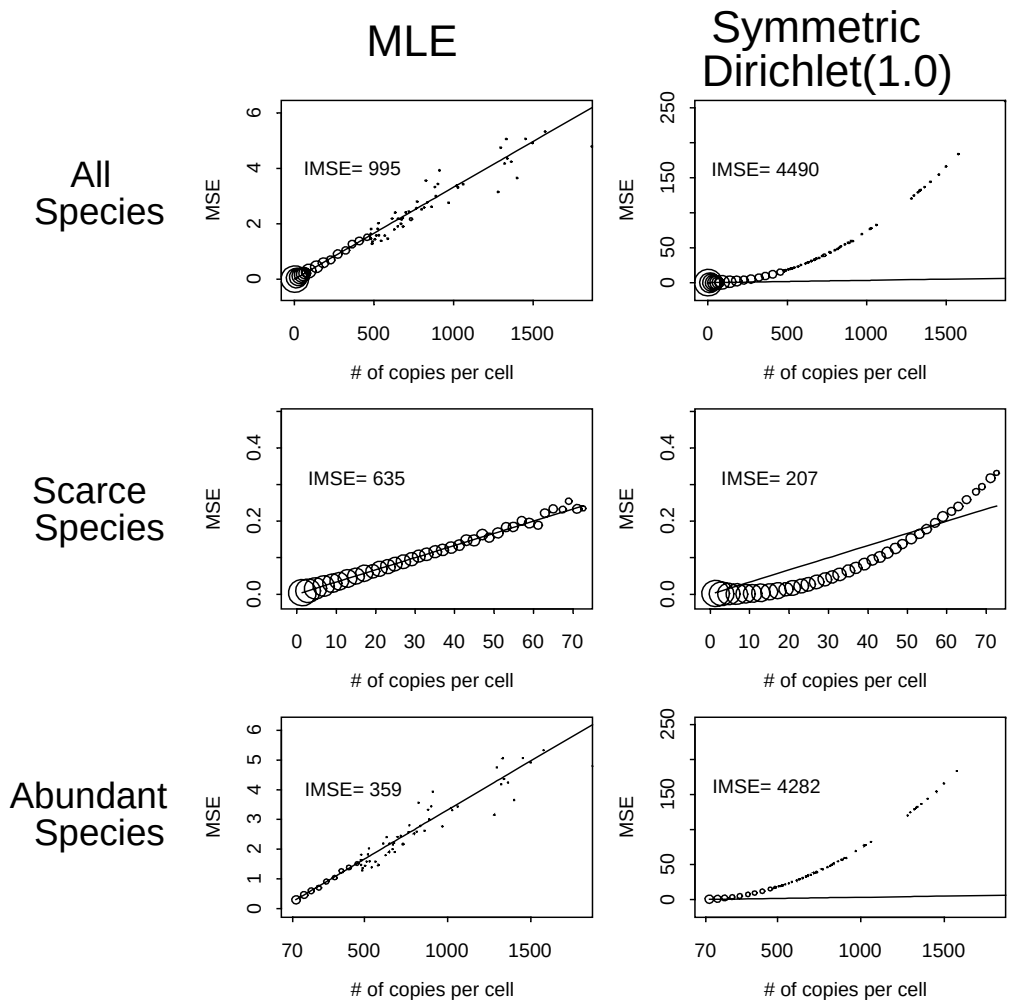
Scarce Species



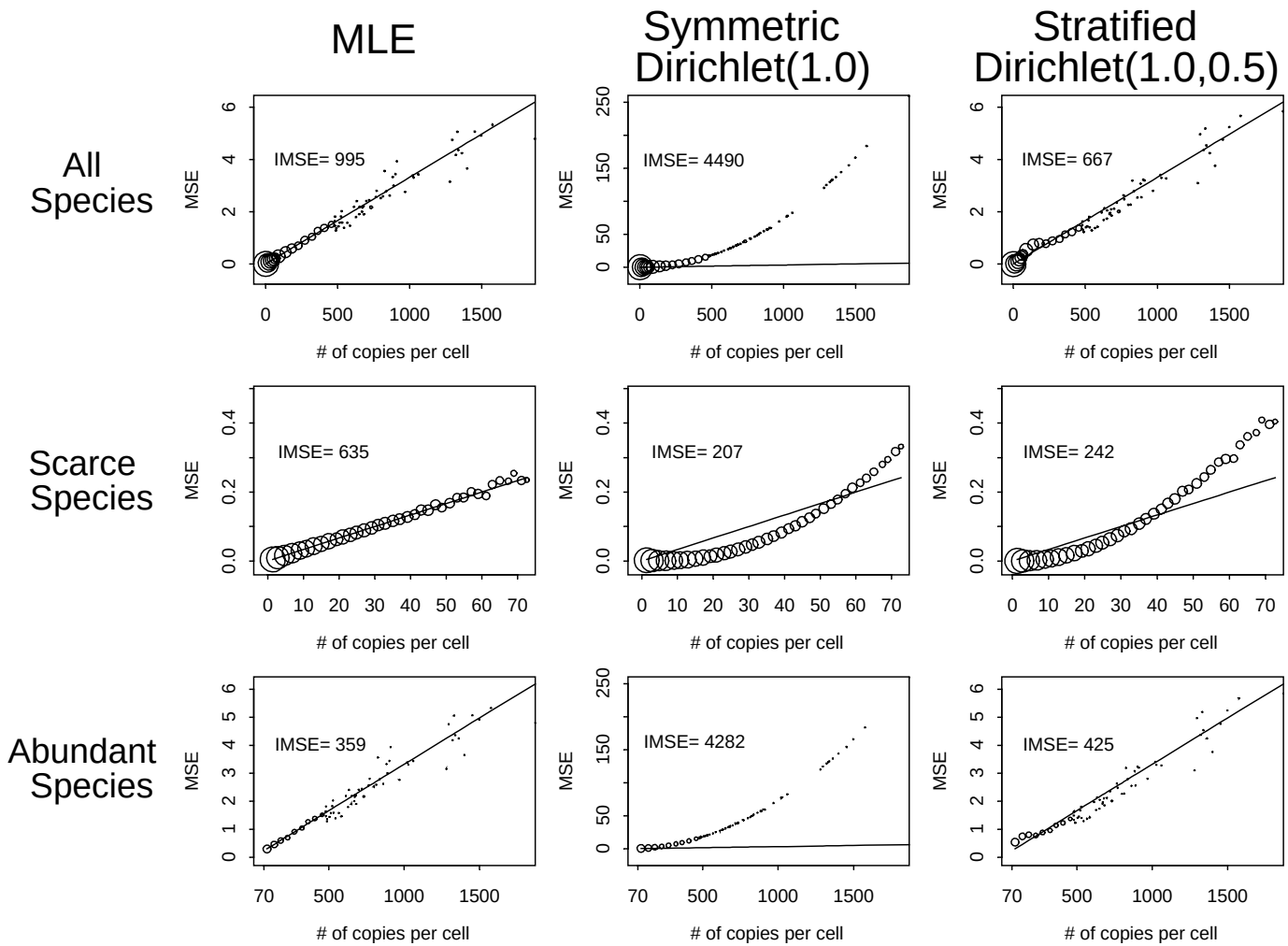
Abundant Species



Simulation Results (n=*10,000*)



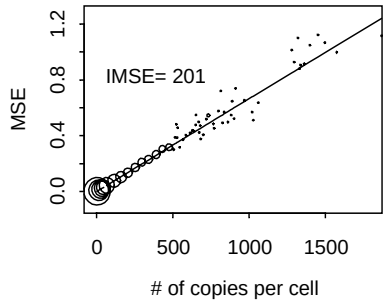
Simulation Results (n=*10,000*)



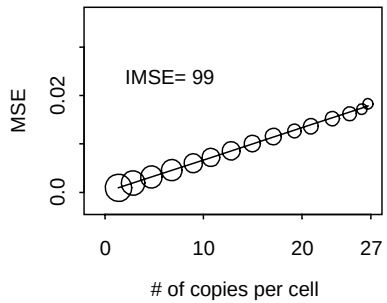
Simulation Results (n=*50,000*)

MLE

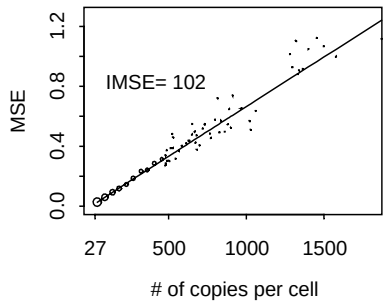
All Species



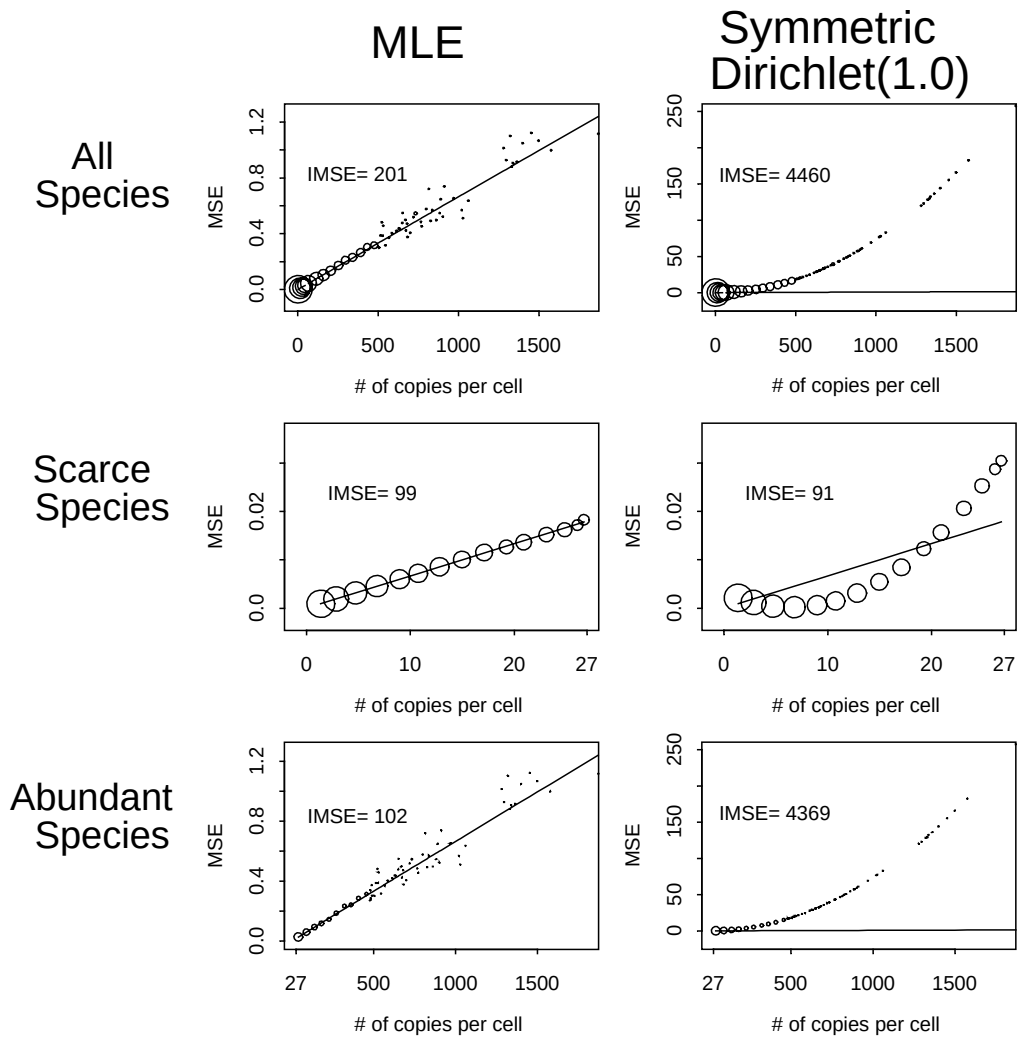
Scarce Species



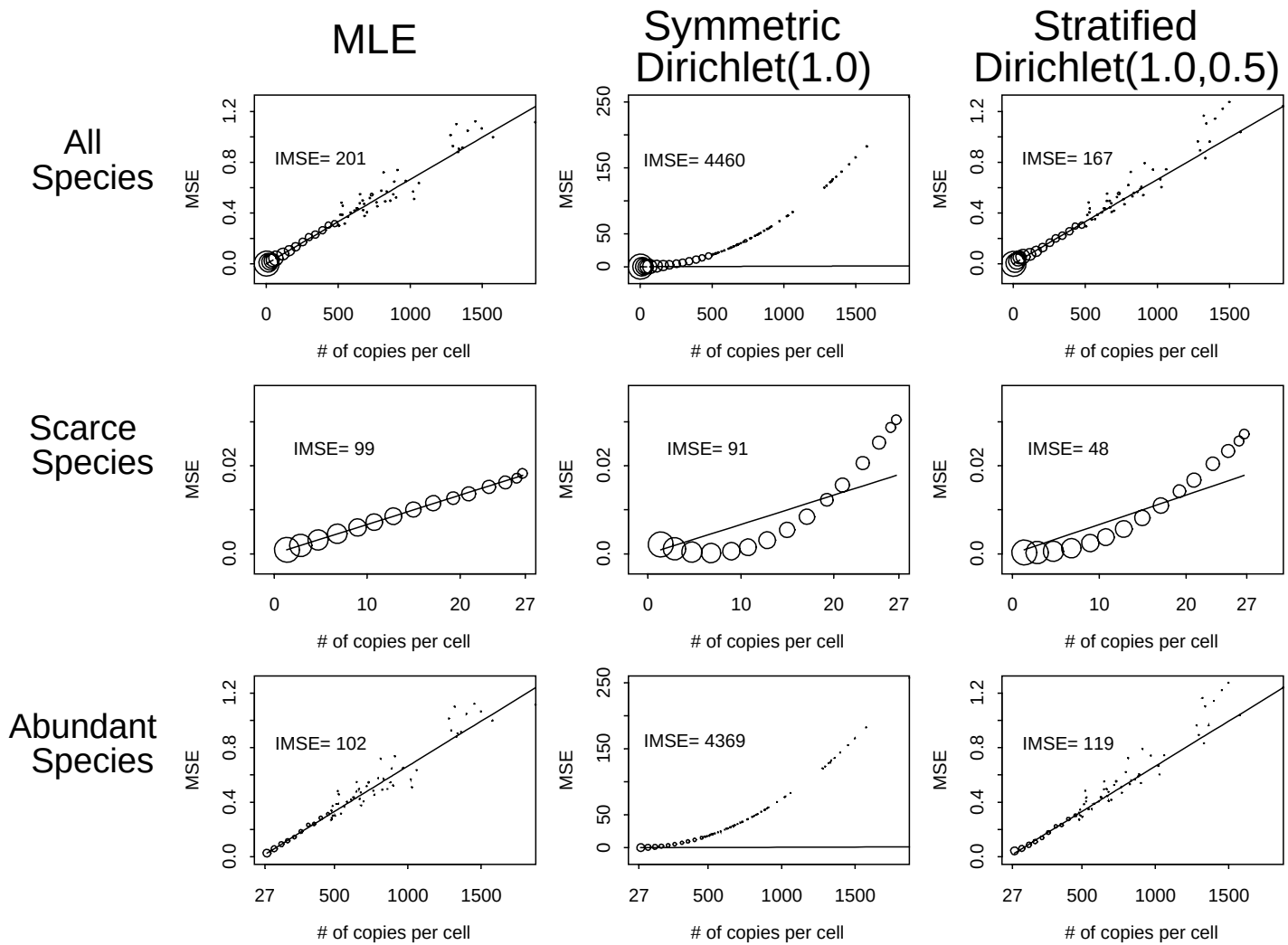
Abundant Species



Simulation Results (n=*50,000*)



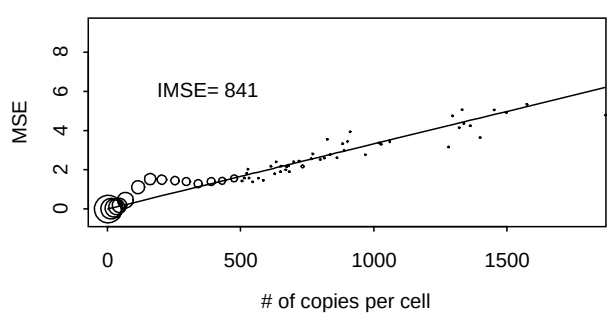
Simulation Results (n=*50,000*)

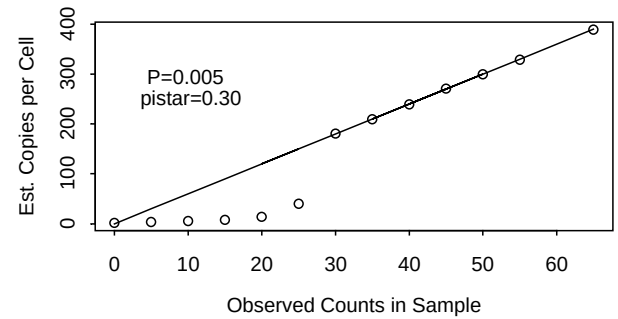


Prior Selection

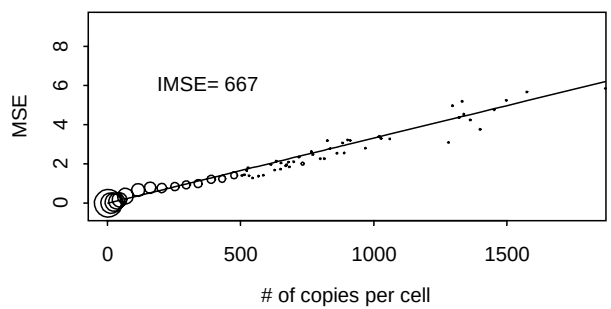
MSE by Species | Shrinkage Curves

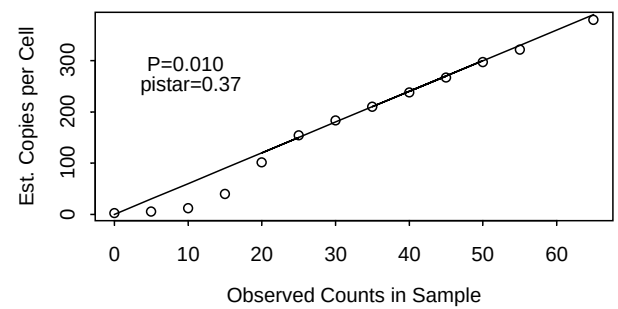
Stratified Dirichlet (1.00,0.01)



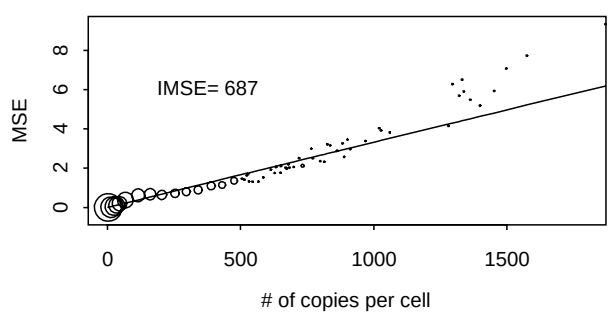


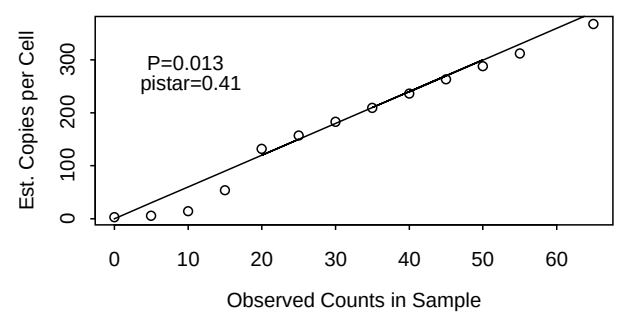
Stratified Dirichlet (1.00,0.50)





Stratified Dirichlet (1.00,1.00)

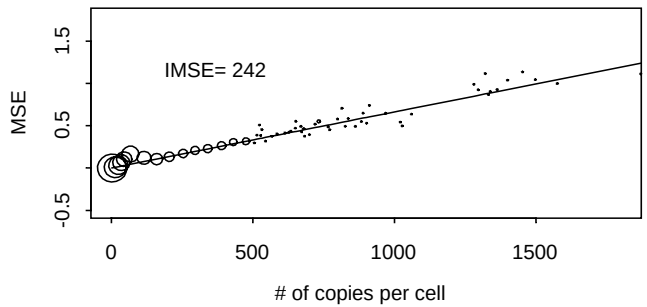


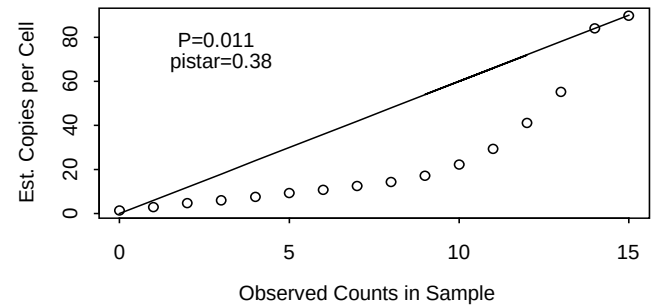


Prior Selection

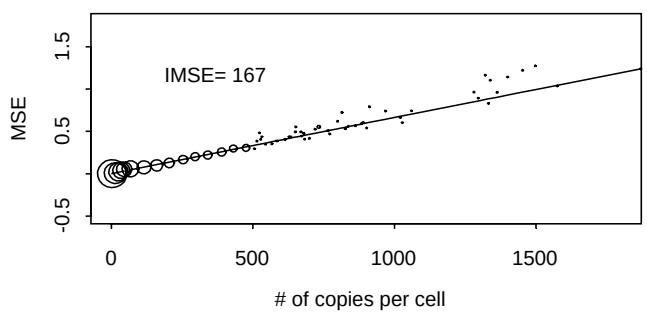
MSE by Species | Shrinkage Curves

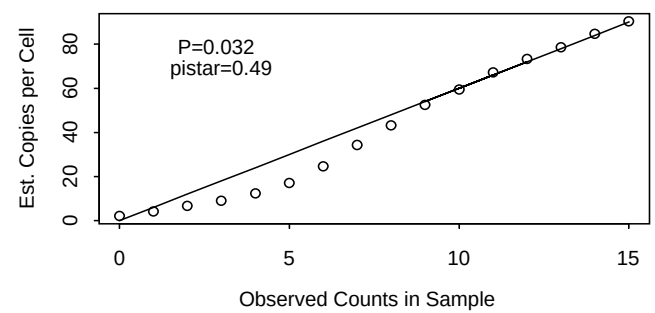
Stratified Dirichlet (1.00,0.01)



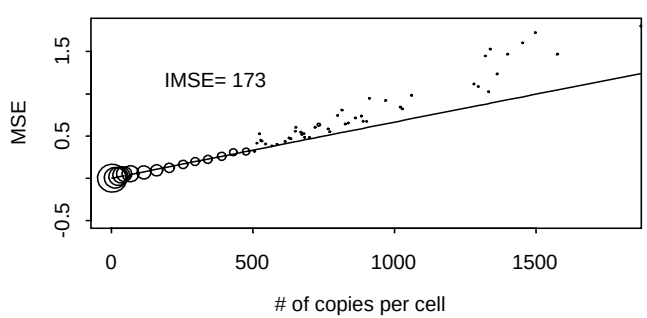


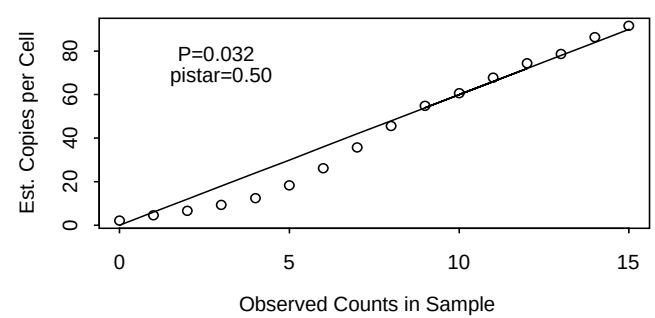
Stratified Dirichlet (1.00,0.50)





Stratified Dirichlet (1.00,1.00)





Summary of Simulation Results

- Comparing MLE and Simple Dirichlet:
  - Dirichlet outperformed MLE for scarce genes
  - Dirichlet much worse for abundant genes
- Comparing MLE and Stratified Dirichlet
  - Stratified Dirichlet beat MLE for scarce genes
  - Comparable for abundant genes
  - Stratified Dirichlet had lower IMSE overall

Conclusions

- SAGE: Method for quantifying gene expression
  - Yields multinomial data
    - Many scarce, few abundant genes
    - Sample size ~ number of genes
  - MLE can be improved upon by taking advantage of known prior information
  - Can get better estimators of relative gene expression profiles using Stratified Dirichlet