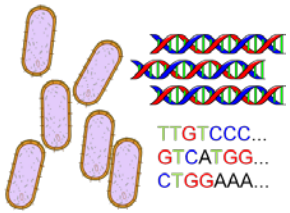
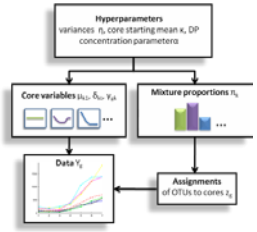
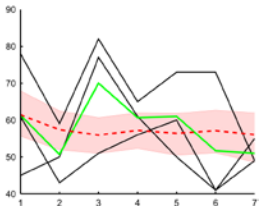
Computational Discovery of Shared Growth Trajectories from Time-Series Microbiome Data







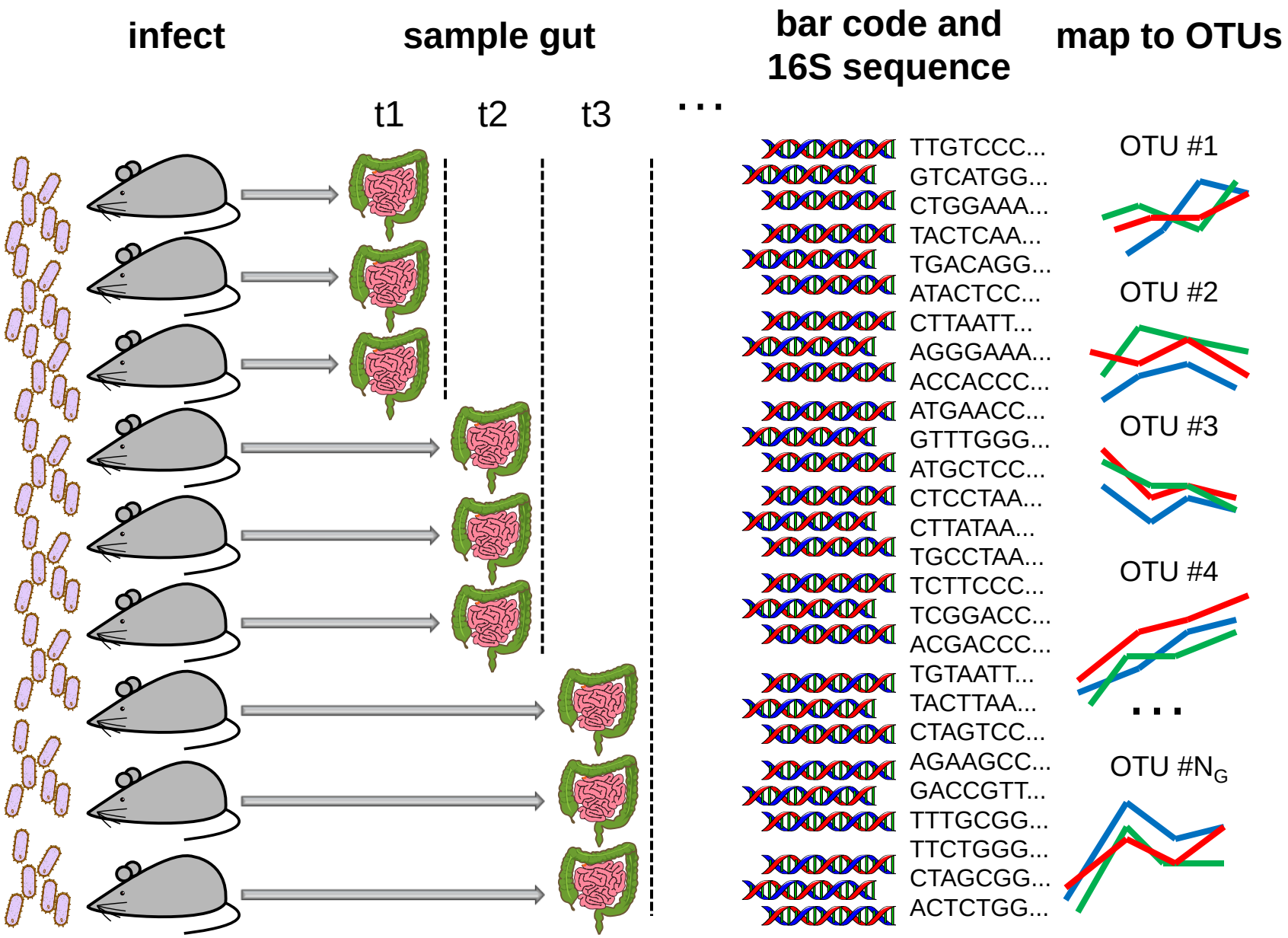
Georg Kurt Gerber, MD, PhD, MPH

Collaborators

- Lynn Bry[1]
- Colleen Cavanaugh[3]
- Guus Roeselers[2,3]
- Clara Belzer[1]

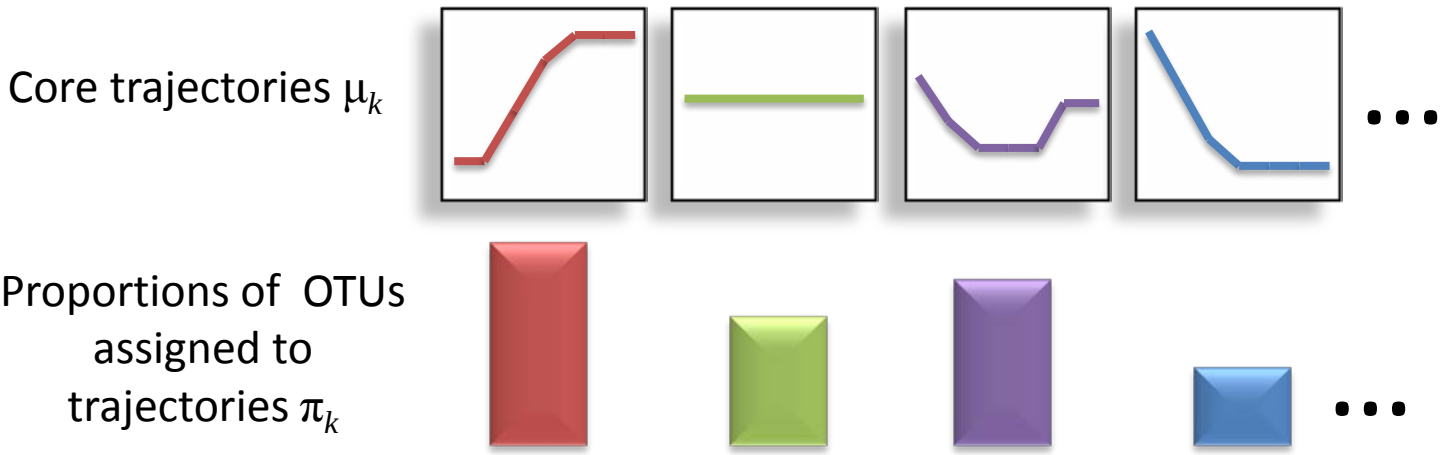
[1] Department of Pathology, Brigham and Women's Hospital, Harvard Medical School
[2] Institute of Water and Wetland Research, Radboud University Nijmegen
[3] Department of Organismic and Evolutionary Biology, Harvard University

Microbiome time-course experiments



Modeling approach to microbiome time-series data

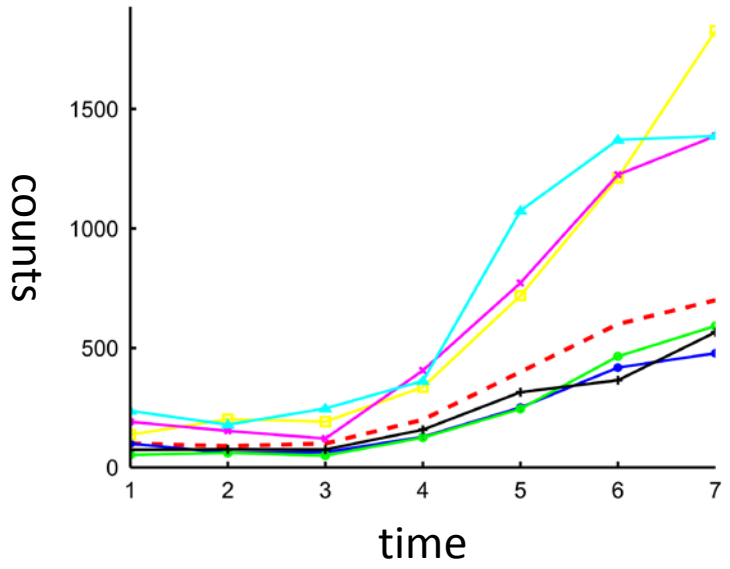
- Main assumption: there is a limited set of unknown "core" growth trajectories that characterize OTU dynamic response to infection



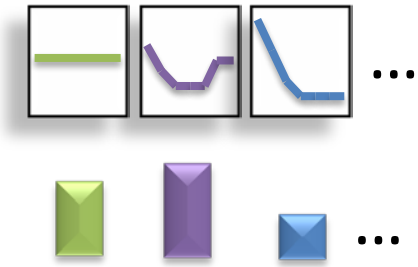
- Goal: simultaneously discover core trajectories and probabilistic assignments of OTUs to them

Modeling challenges

1. How many core trajectories?
2. Temporal dependences
3. Noisy counts w/ limited # of replicates
4. Individual OTU deviations from core trajectories



Challenge #1: How many core trajectories? Dirichlet Process Mixture Model

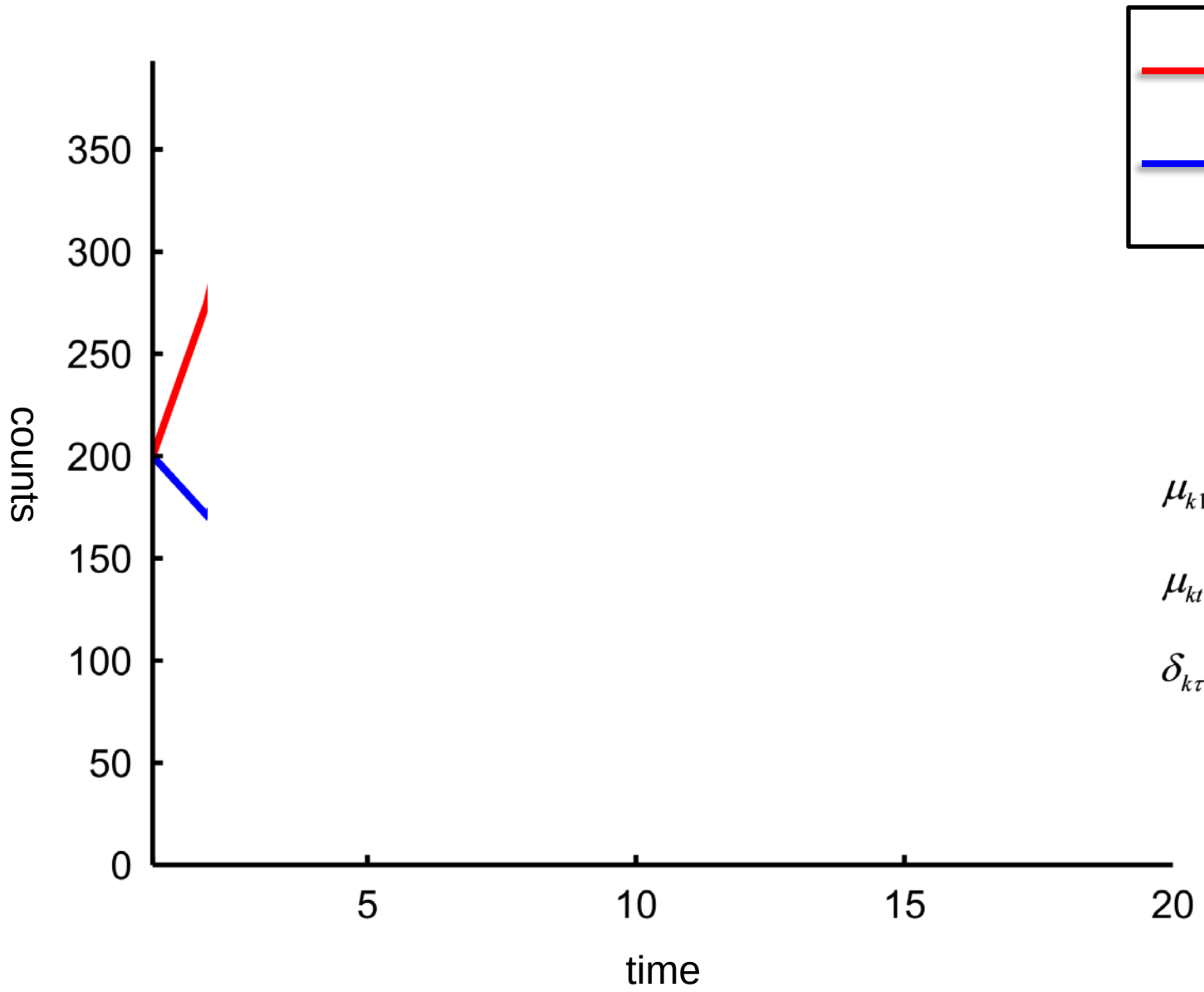


- Bayesian nonparametric technique
  - Potentially unlimited # of core trajectories
  - Infer distribution of trajectories from data
- Prior probability on core trajectory mixture proportions uses a “stick-breaking” construction
- Concentration parameter α controls uniformity of breaks

$$\pi'_j \mid \alpha \sim \text{Beta}(1, \alpha)$$

$$\pi_j = \pi'_j \prod_{l=1}^{j-1} (1 - \pi'_l)$$

Challenge #2: temporal dependence



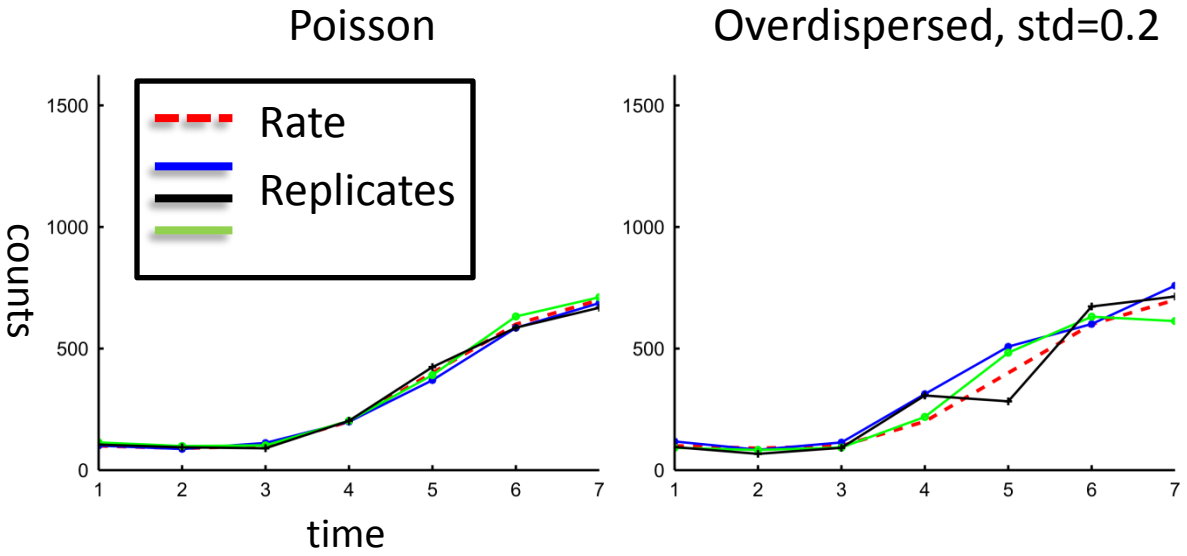
Random walk

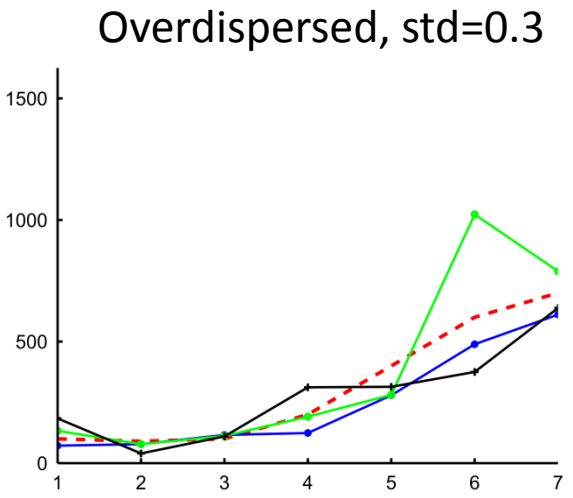
$$\mu_{k1} \sim \text{Gamma}(\kappa_\mu / \eta_\mu, \eta_\mu)$$

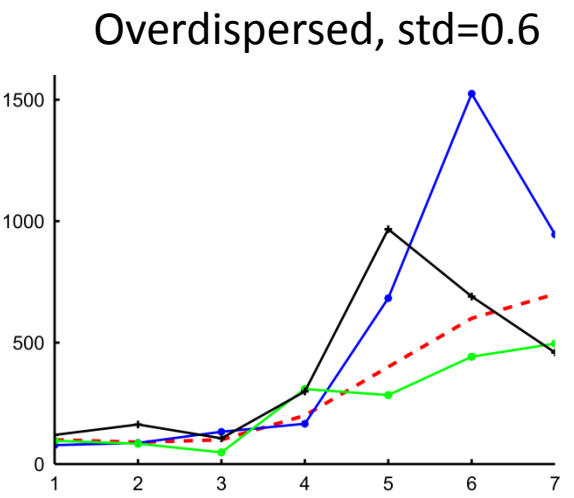
$$\mu_{kt} = \mu_{k1} \prod_{\tau=2}^{t} \delta_{k\tau}$$

$$\delta_{k\tau} \sim \text{Gamma}(1/\eta_{\delta t}, \eta_{\delta t})$$

Challenge #3: noisy count data w/ limited # of replicates







Poisson

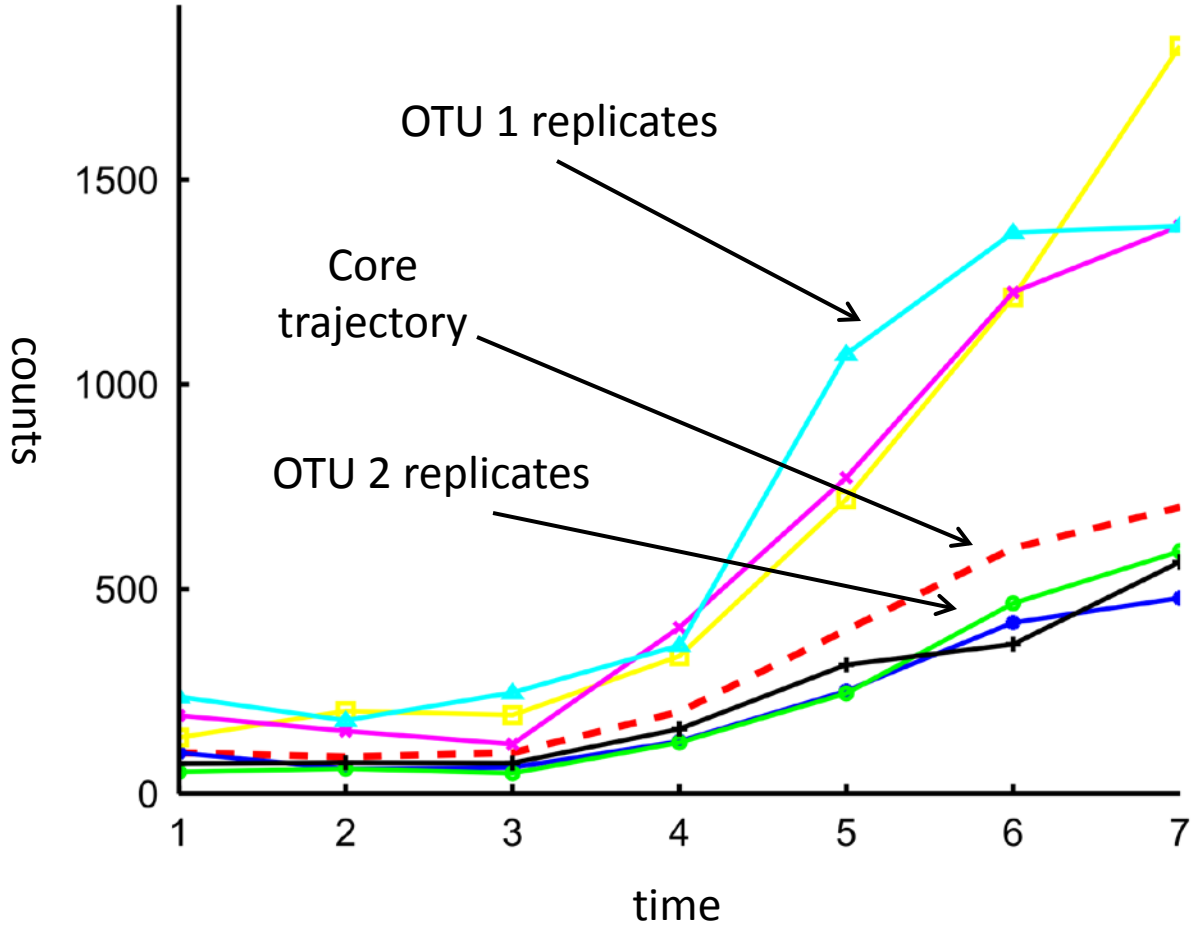
$$P(Y_{gtr}) = \frac{\mu_{kt}^{Y_{gtr}} e^{-\mu_{kt}}}{Y_{gtr}!}$$

Overdispersed Poisson

$$P(Y_{gtr}) = \frac{(\mu_{kt}\varepsilon_{kgtr})^{Y_{gtr}} e^{-\mu_{kt}\varepsilon_{kgtr}}}{Y_{gtr}!}$$

$$P(\varepsilon_{kgtr}) = \text{Gamma}(1/\eta_\varepsilon, \eta_\varepsilon)$$

Challenge #4: OTU deviation from core trajectory

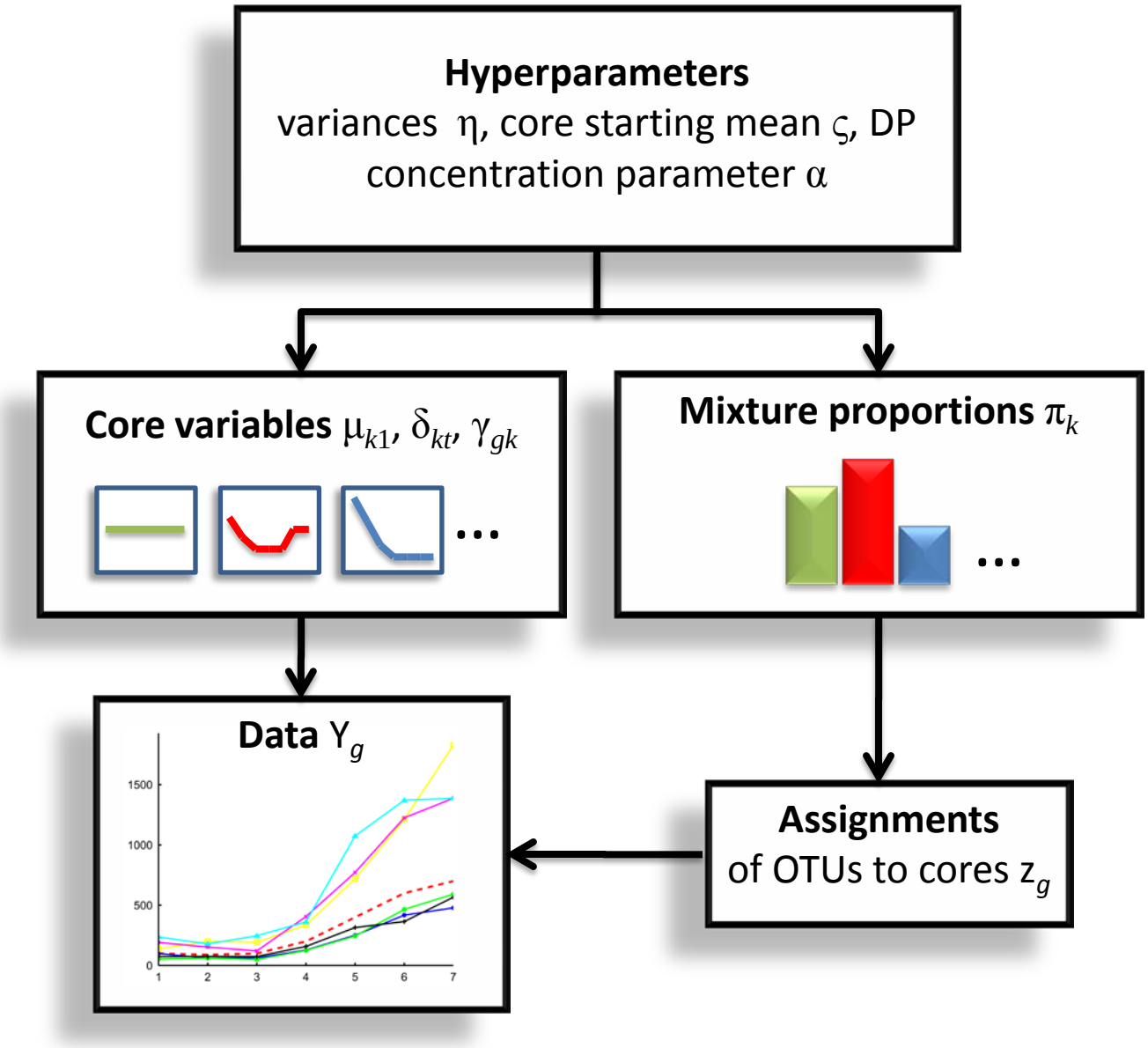


$$P(Y_{gtr}) = \frac{(\mu_{kt}\gamma_{kg}\varepsilon_{kgtr})^{Y_{gtr}} e^{-\mu_{kt}\gamma_{kg}\varepsilon_{kgtr}}}{Y_{gtr}!}$$

$$P(\gamma_{kg}) = \text{Gamma}(1/\eta_\gamma, \eta_\gamma)$$

$$P(\varepsilon_{kgtr}) = \text{Gamma}(1/\eta_\varepsilon, \eta_\varepsilon)$$

Model schematic



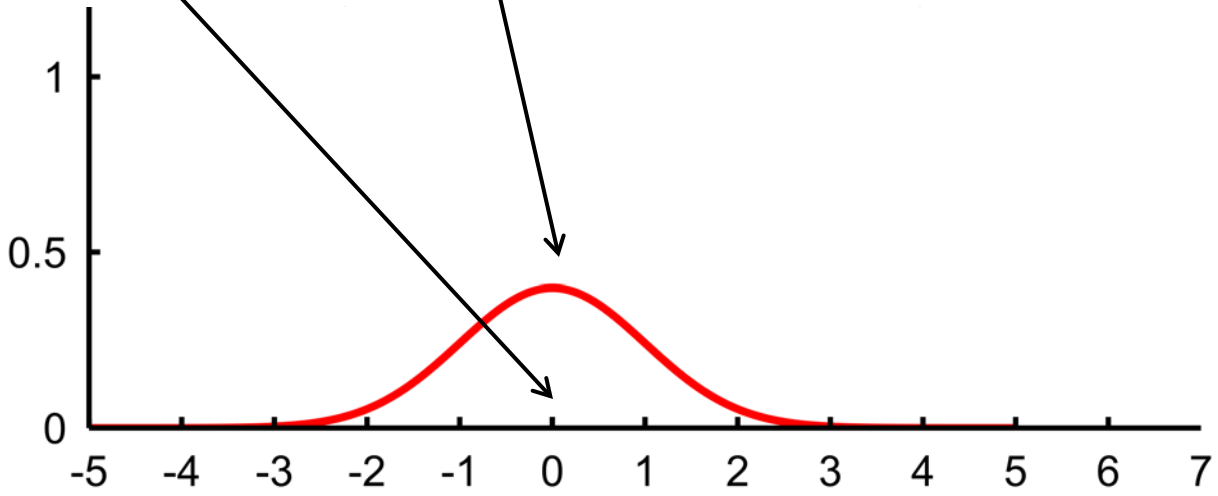
Bayesian model inference

- Estimate probability distribution of variables θ (hyperparameters, core trajectories, OTU specific deviations, OTU assignments, etc.) given the data Y

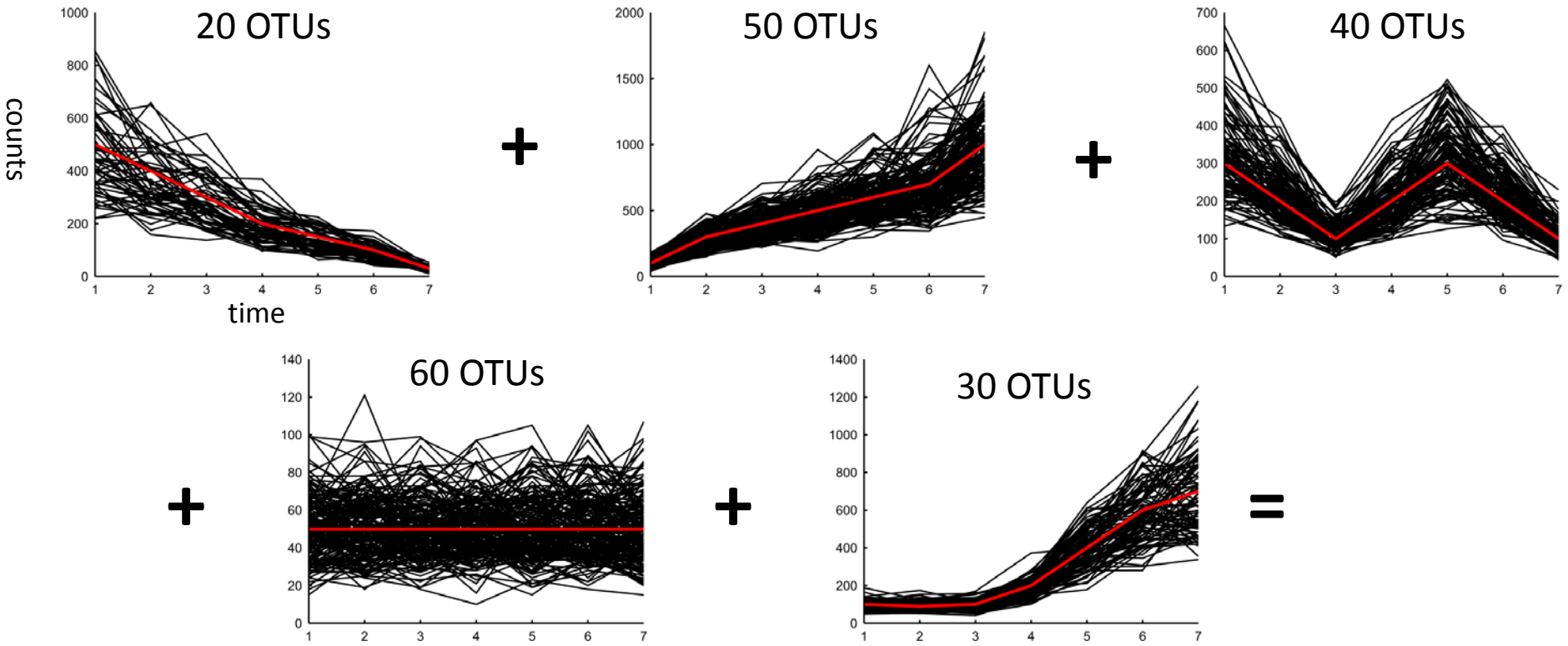
$$P(\theta \mid Y) = P(Y \mid \theta)\ P(\theta) / P(Y)$$

posterior likelihood prior

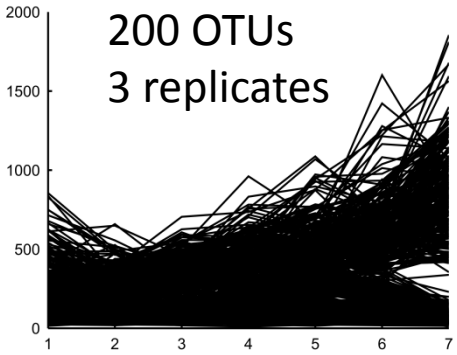
- We use Gibbs sampling with Adaptive Rejection Monte Carlo sampling and/or variable augmentation steps for hyper



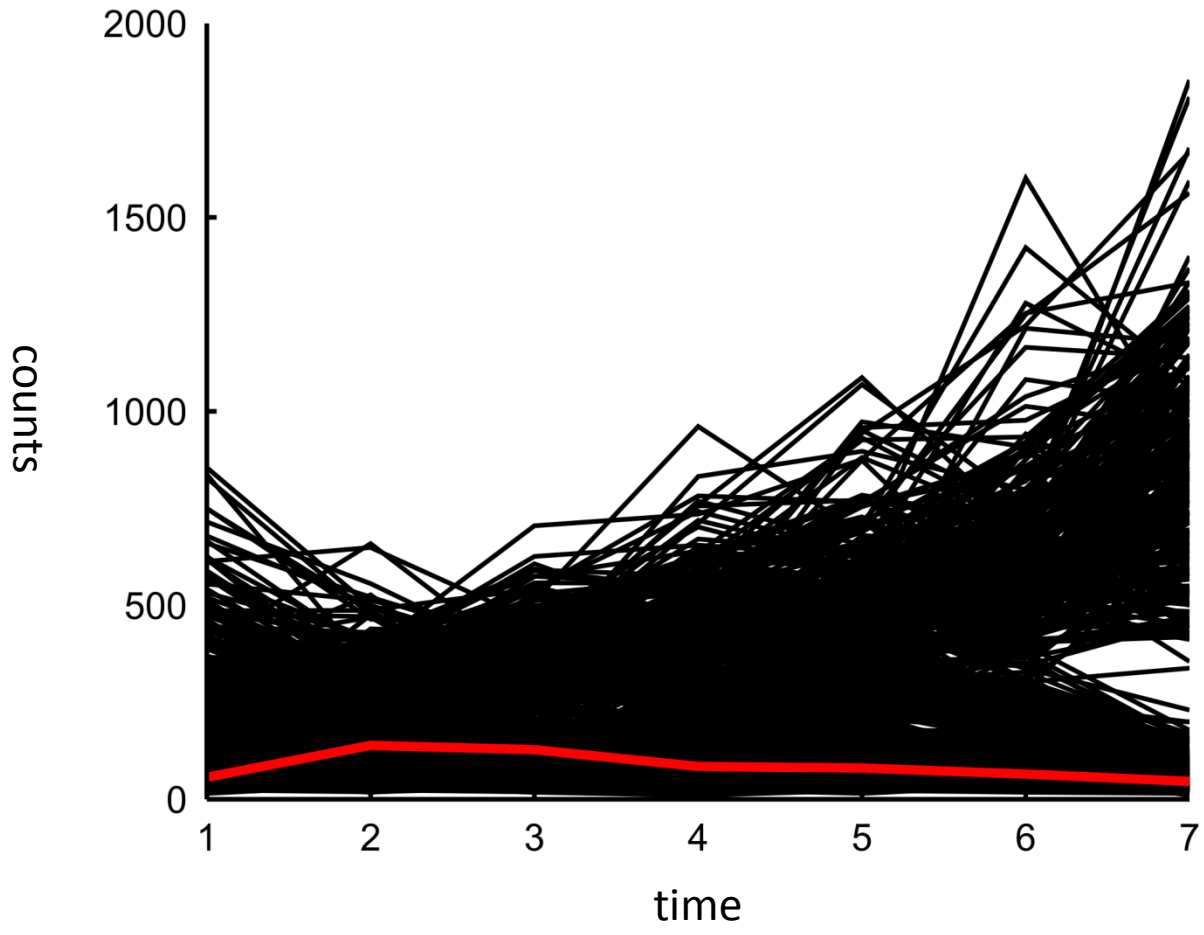
Synthetic data experiment



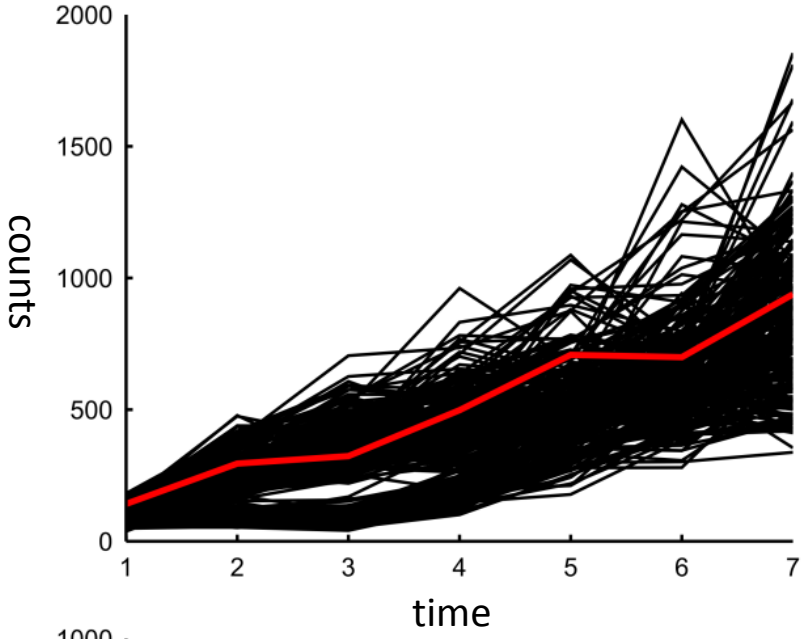
Input to algorithm

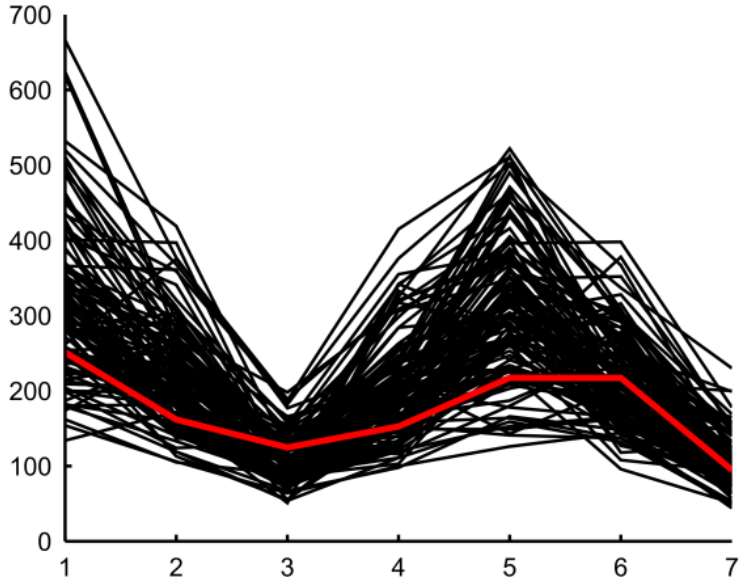


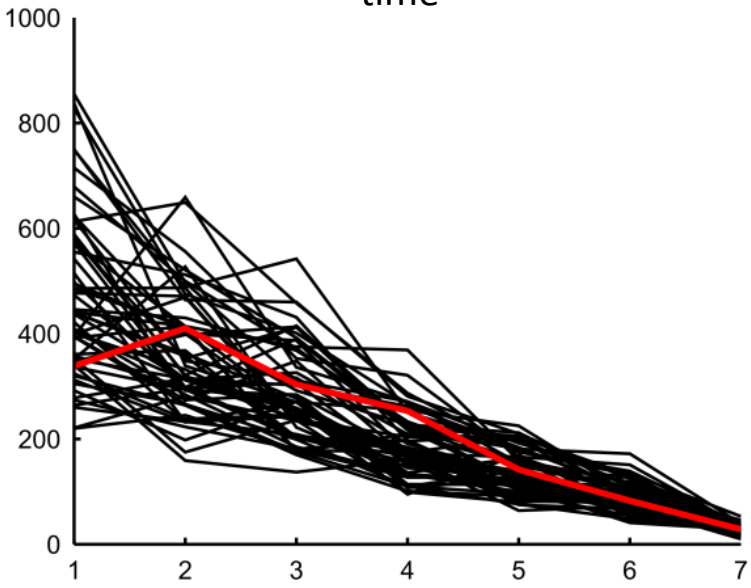
Sample #1

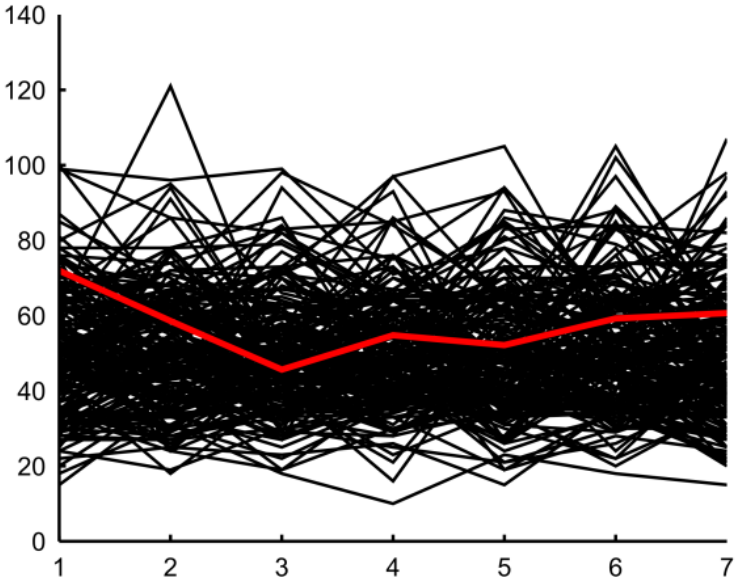


Sample #2

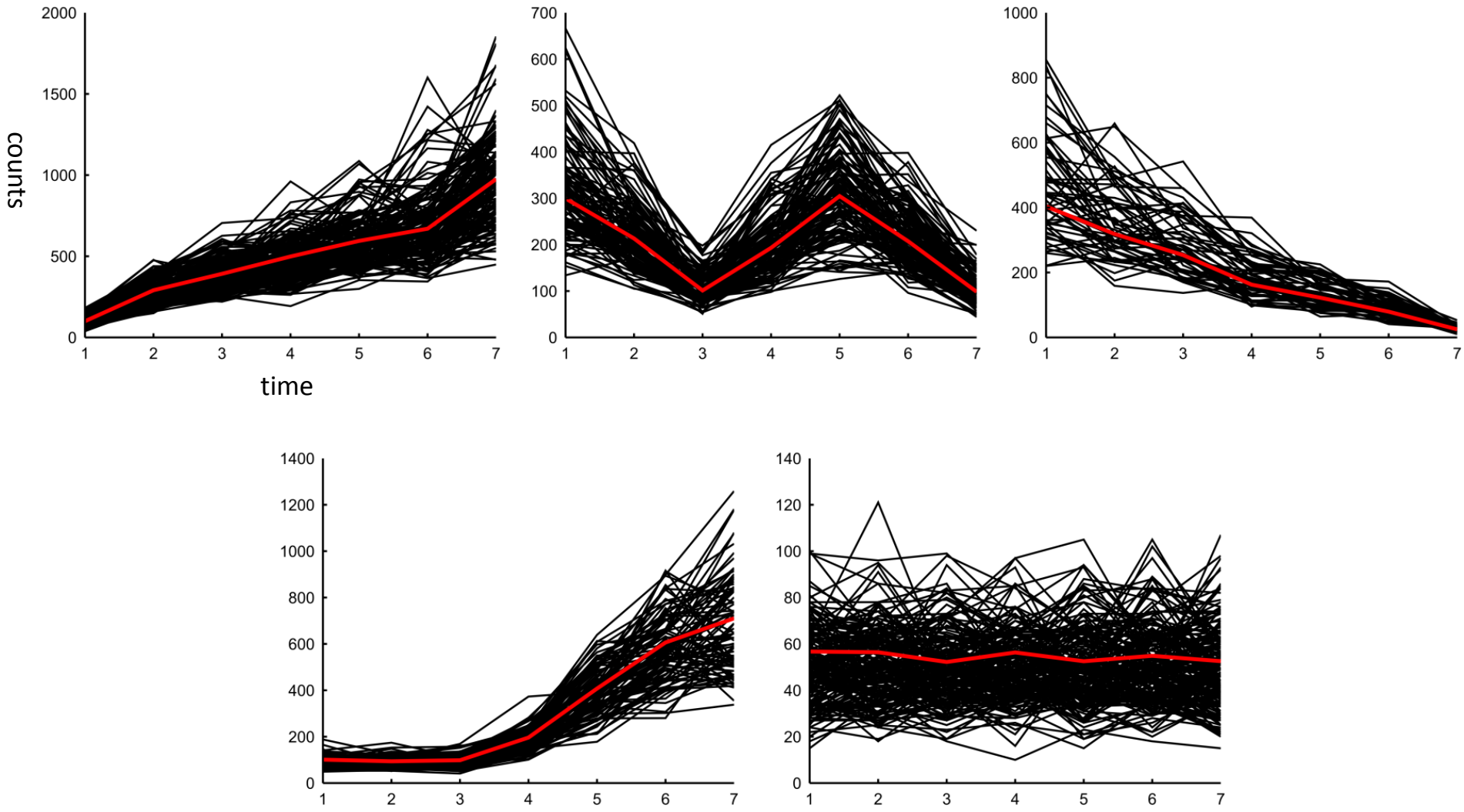

counts
time
2000
1500
1000
500
0
1
2
3
4
5
6
7


700
600
500
400
300
200
100
0
1
2
3
4
5
6
7

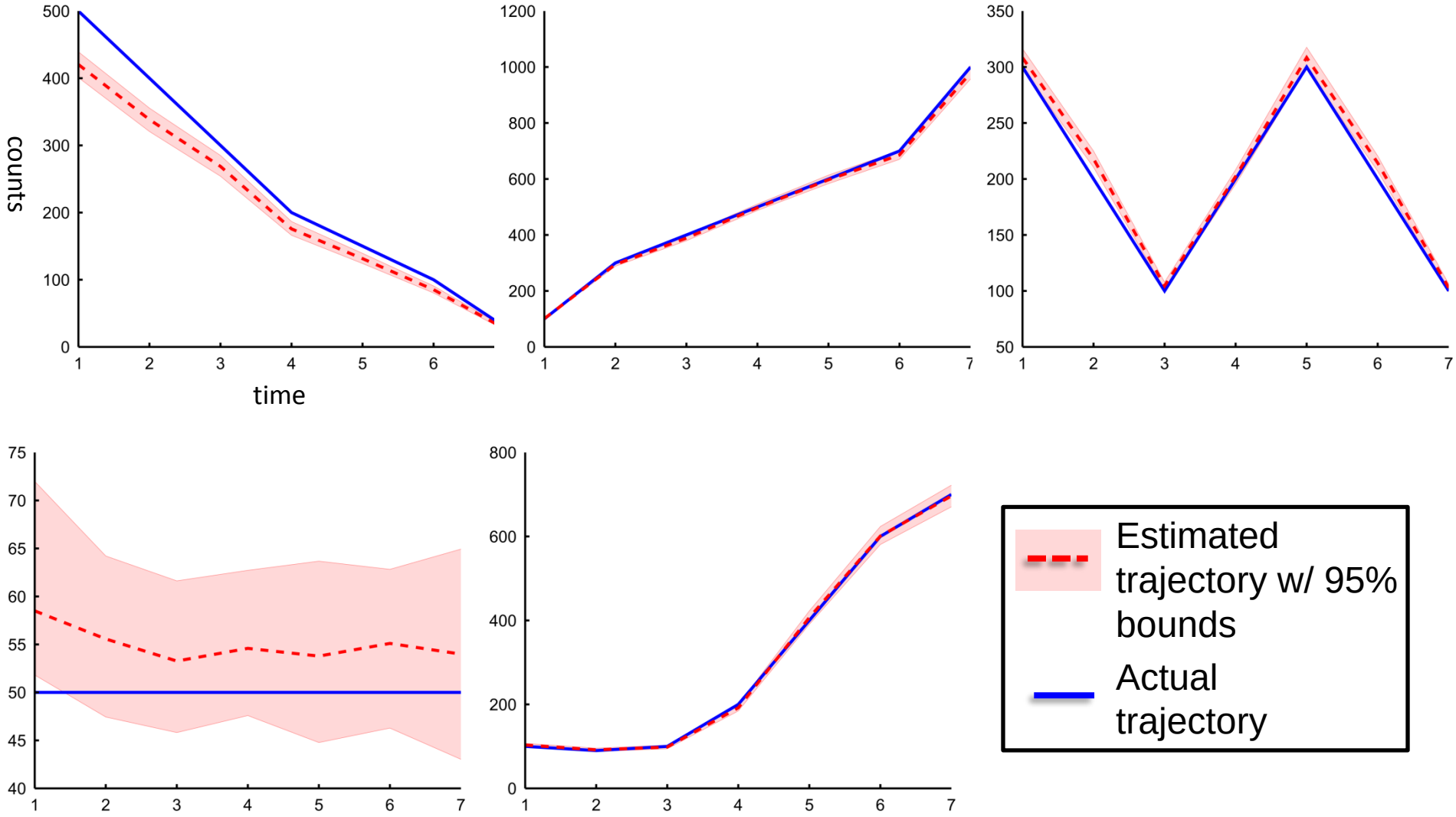

1000
800
600
400
200
0
1
2
3
4
5
6
7


140
120
100
80
60
40
20
0
1
2
3
4
5
6
7

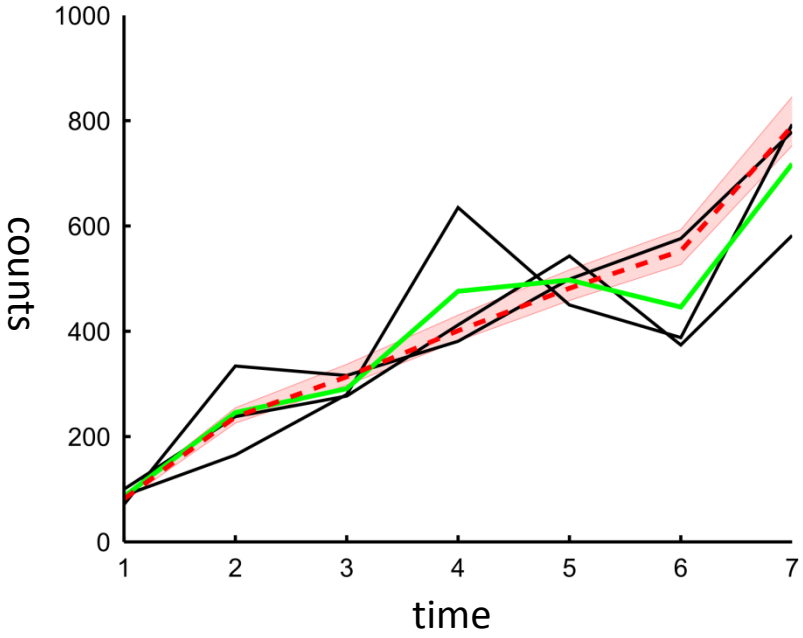
Sample #250

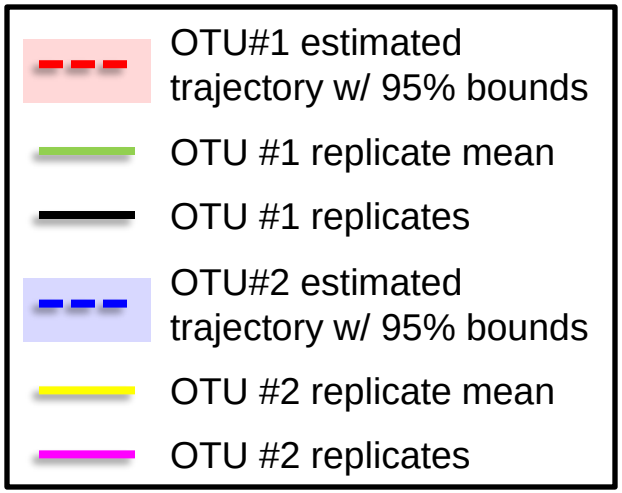

counts
time
2000
1500
1000
500
0
700
600
500
400
300
200
100
0
1000
800
600
400
200
0
1400
1200
1000
800
600
400
200
0
140
120
100
80
60
40
20
0
1
2
3
4
5
6
7

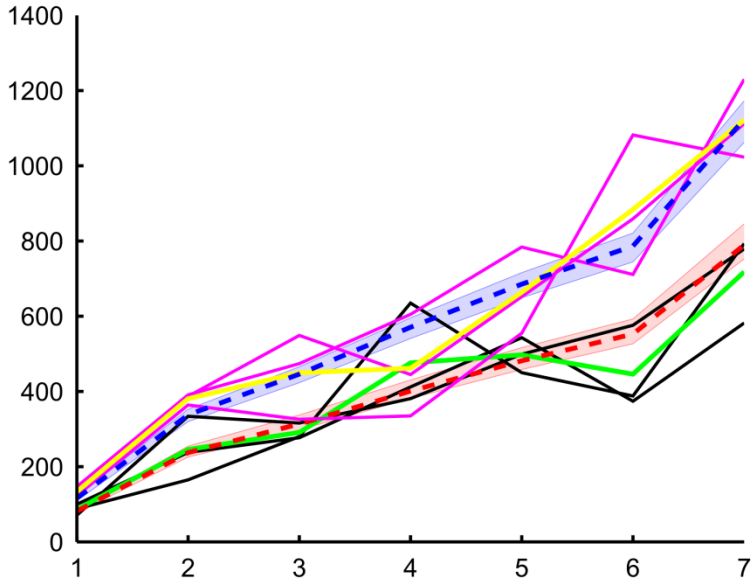
Recovered core trajectories

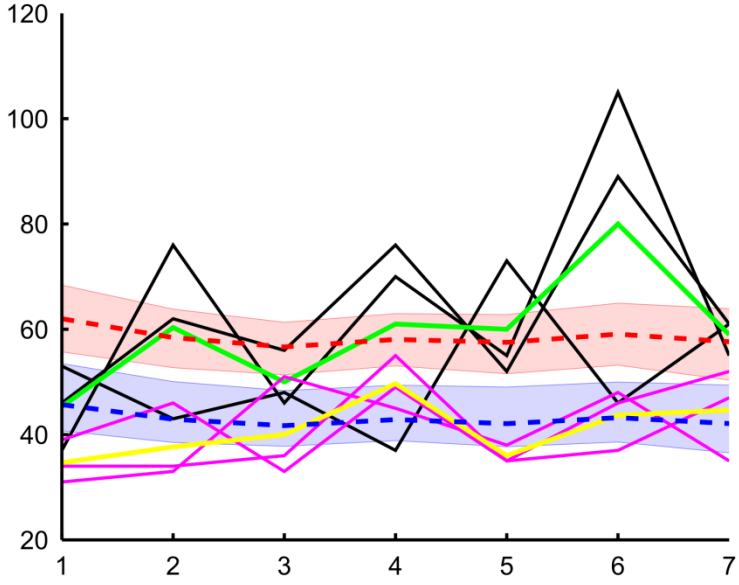


Estimated OTU trajectories









Summary/future work

- Developed novel model for analyzing microbiome time-series data
  - Automatically discovers core growth trajectories shared among OTUs using nonparametric Bayesian approach
  - Addresses challenges of overdispersed data, time dependences, OTU specific deviations and uncertainty in # of core growth trajectories
- Implemented inference algorithms in Java and demonstrated accurate recovery of trajectories from noisy simulated data
- Future work
  - Apply to *Citrobacter* infection time-course data
  - Use model to estimate # of unseen OTUs in samples
  - Handle sequencing error
  - Incorporate phylogenetic or functional information

Funding acknowledgements

- Pilot feasibility grant, Harvard Digestive Diseases Center (P30-DK034854; Cavanaugh, Bry)
- NIH/NICHD (R01-HD061916; Bry, Mantis)
- BWH Pathology Department Residency Training Program