Generalized estimating equation modeling on correlated microbiome sequencing data with longitudinal measures

Emily Palmer

Journal Club, Oregon State University

March 16, 2021

Overview

Introduction

Correlation structure

- Taxonomic correlation structure of OTUs
- Correlation structure from longitudinal repeated measures
- Integrative Correlation Matrix

Microbiome Taxonomic Longitudinal Correlation model

- General GEE framework
- Two-part model

Introduction

Challenges of applying regression models on association studies of microbiome composition and environmental factors

- Many OTUs, potentially correlated
- Repeated Measures (longitudinal, other repeated measures)
- OTU data has excessive zeros

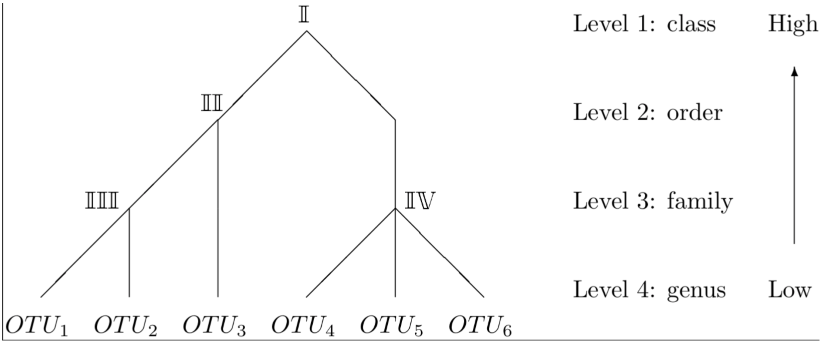
Goals

- Estimate correlations between multiple OTUs
- Incorporate correlations into models with longitudinal OTU measures
- Estimate predictors effects using GEEs
- Two-part Microbiome Taxonomic Longitudinal Correlation (MLTC) model

Correlation matrix of taxonomic structure - Assumptions

- Assume that OTUs that belong to the same taxa at some higher level have some correlation
- All OTUs will belong to same taxa at highest level, so there are $\binom{N}{2}$ possible correlations - infeasible to model
- Assume that two pairs of OTUs have the same correlation if the first common taxa of both pairs are identical

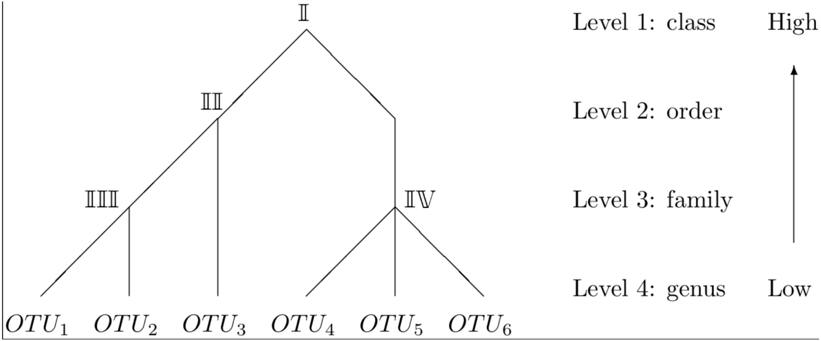
Example


𝕀
𝕀𝕀
𝕀𝕀𝕀
𝕀𝕍
OTU_1
OTU_2
OTU_3
OTU_4
OTU_5
OTU_6
Level 1: class
Level 2: order
Level 3: family
Level 4: genus
High
Low

Notation and Definitions

- I levels:
    - 1st taxonomic level is the level at which all observed N OTUs belong to the same taxon but not at one level lower
- M_i: number of taxa at level i $(M_1 = 1, M_I = N)$
- $t_{m_i,i}$: taxon at level i for $m_i = 1, \ldots, M_i$
- $n_{m_i,i}$: number of OTUs belonging to taxon $t_{m_i,i}$.
  $\mathbf{n}_i = (n_{1i}, \ldots, n_{M_i,i})$

Example



$M_1 = 1, M_2 = 2, M_3 = 3, M_4 = 6,$
$\boldsymbol{n}_1 = 6, \boldsymbol{n}_2 = (3,3), \boldsymbol{n}_3 = (2,1,3), \boldsymbol{n}_4 = (1,1,1,1,1,1)$
$\mathbb{I}$ represents correlation of same class different orders,
$\mathbb{II}$ correlation of same order different families,
$\mathbb{III}$, $\mathbb{IV}$ same family

The taxonomic structure matrix Γ

$$
\begin{array}{ccccccc}
 & OTU_1 & OTU_2 & OTU_3 & OTU_4 & OTU_5 & OTU_6 \\
OTU_1 & \mathbb{D} & \mathbb{IIII} & \mathbb{III} & \mathbb{I} & \mathbb{I} & \mathbb{I} \\
OTU_2 & \mathbb{IIII} & \mathbb{D} & \mathbb{III} & \mathbb{I} & \mathbb{I} & \mathbb{I} \\
OTU_3 & \mathbb{III} & \mathbb{III} & \mathbb{D} & \mathbb{I} & \mathbb{I} & \mathbb{I} \\
OTU_4 & \mathbb{I} & \mathbb{I} & \mathbb{I} & \mathbb{D} & \mathbb{IV} & \mathbb{IV} \\
OTU_5 & \mathbb{I} & \mathbb{I} & \mathbb{I} & \mathbb{IV} & \mathbb{D} & \mathbb{IV} \\
OTU_6 & \mathbb{I} & \mathbb{I} & \mathbb{I} & \mathbb{IV} & \mathbb{IV} & \mathbb{D}
\end{array}
$$

Finding the taxonomic structure matrix

- Create $I-1$ $N \times N$ block matrices

$$\boldsymbol{\Gamma}_i = \begin{pmatrix} \boldsymbol{B}_{1i} & & \\ & \ddots & \\ & & \boldsymbol{B}_{M_i,i} \end{pmatrix}$$

  For $m_i = 1, \ldots, M_i$, each block $\boldsymbol{B}_{1,i}$ is an $n_{m_i,i} \times n_{m_i,i}$ matrix, with diagonal entries $\mathbb{D}$ and off diagonal entries $\sum_{h=0}^{i-1} M_h + m_i$

- Create interim correlation after replacement at level i ($\boldsymbol{\Gamma}^{(i)}$)
  - For $i = 1$, $\boldsymbol{\Gamma}^{(1)} = \boldsymbol{\Gamma}_1$
  - For $i = 2, \ldots, I-1$, Replace the block diagonal entries of $\boldsymbol{\Gamma}^{(i-1)}$ with $\boldsymbol{B}_{m_i,i}$, but keep all other entries the same.
- Sort all elements from largest to smallest. Different ranks are the distinct correlations to estimate

Example

$$\boldsymbol{\Gamma}_1 = \begin{pmatrix} \mathbb{D} & \mathbb{1} & \mathbb{1} & \mathbb{1} & \mathbb{1} & \mathbb{1} \\ \mathbb{1} & \mathbb{D} & \mathbb{1} & \mathbb{1} & \mathbb{1} & \mathbb{1} \\ \mathbb{1} & \mathbb{1} & \mathbb{D} & \mathbb{1} & \mathbb{1} & \mathbb{1} \\ \mathbb{1} & \mathbb{1} & \mathbb{1} & \mathbb{D} & \mathbb{1} & \mathbb{1} \\ \mathbb{1} & \mathbb{1} & \mathbb{1} & \mathbb{1} & \mathbb{D} & \mathbb{1} \\ \mathbb{1} & \mathbb{1} & \mathbb{1} & \mathbb{1} & \mathbb{1} & \mathbb{D} \end{pmatrix}, \boldsymbol{\Gamma}_2 = \begin{pmatrix} \mathbb{D} & \mathbb{2} & \mathbb{2} & & & \\ \mathbb{2} & \mathbb{D} & \mathbb{2} & & & \\ \mathbb{2} & \mathbb{2} & \mathbb{D} & & & \\ & & & \mathbb{D} & \mathbb{3} & \mathbb{3} \\ & & & \mathbb{3} & \mathbb{D} & \mathbb{3} \\ & & & \mathbb{3} & \mathbb{3} & \mathbb{D} \end{pmatrix}$$

$$\boldsymbol{\Gamma}_3 = \begin{pmatrix} \mathbb{D} & \mathbb{4} & & & & \\ \mathbb{4} & \mathbb{D} & & & & \\ & & \mathbb{D} & & & \\ & & & \mathbb{D} & \mathbb{6} & \mathbb{6} \\ & & & \mathbb{6} & \mathbb{D} & \mathbb{6} \\ & & & \mathbb{6} & \mathbb{6} & \mathbb{D} \end{pmatrix}.$$

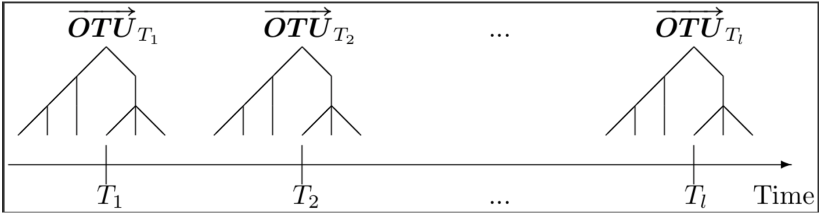
$$\boldsymbol{\Gamma}^{(3)} = \begin{pmatrix} \mathbb{D} & \mathbb{4} & \mathbb{2} & \mathbb{1} & \mathbb{1} & \mathbb{1} \\ \mathbb{4} & \mathbb{D} & \mathbb{2} & \mathbb{1} & \mathbb{1} & \mathbb{1} \\ \mathbb{2} & \mathbb{2} & \mathbb{D} & \mathbb{1} & \mathbb{1} & \mathbb{1} \\ \mathbb{1} & \mathbb{1} & \mathbb{1} & \mathbb{D} & \mathbb{6} & \mathbb{6} \\ \mathbb{1} & \mathbb{1} & \mathbb{1} & \mathbb{6} & \mathbb{D} & \mathbb{6} \\ \mathbb{1} & \mathbb{1} & \mathbb{1} & \mathbb{6} & \mathbb{6} & \mathbb{D} \end{pmatrix}$$

$\boldsymbol{\Gamma}$ can be represented by $(\boldsymbol{n}_1, \ldots, \boldsymbol{n}_l)$

The taxonomic structure matrix Γ

$$
\begin{array}{ccccccc}
 & OTU_1 & OTU_2 & OTU_3 & OTU_4 & OTU_5 & OTU_6 \\
OTU_1 & \mathbb{D} & \mathbb{IIII} & \mathbb{III} & \mathbb{I} & \mathbb{I} & \mathbb{I} \\
OTU_2 & \mathbb{IIII} & \mathbb{D} & \mathbb{III} & \mathbb{I} & \mathbb{I} & \mathbb{I} \\
OTU_3 & \mathbb{III} & \mathbb{III} & \mathbb{D} & \mathbb{I} & \mathbb{I} & \mathbb{I} \\
OTU_4 & \mathbb{I} & \mathbb{I} & \mathbb{I} & \mathbb{D} & \mathbb{IV} & \mathbb{IV} \\
OTU_5 & \mathbb{I} & \mathbb{I} & \mathbb{I} & \mathbb{IV} & \mathbb{D} & \mathbb{IV} \\
OTU_6 & \mathbb{I} & \mathbb{I} & \mathbb{I} & \mathbb{IV} & \mathbb{IV} & \mathbb{D}
\end{array}
$$

Correlation structure from longitudinal repeated measures


$\overrightarrow{OTU}_{T_1}$
$\overrightarrow{OTU}_{T_2}$
...
$\overrightarrow{OTU}_{T_l}$
T_1
T_2
...
T_l
Time

Types of correlations between pairs of time points

- Exchangeable
  - Assumes all correlations are equal to each other
- Toeplitz
  - Assumes time points with equal temporal distance have equal correlation
- Unstructured
  - Assumes each pair has a different correlations
  - Most complicated structure for correlation parameter estimation

Correlation structure matrix for the the same individual is denoted Ω_T

Example Correlation Matrices for 3 timepoints

Exchangable structure

$$\begin{array}{cccc} & T_1 & T_2 & T_3 \\ T_1 & \mathbb{D} & \mathbb{i} & \mathbb{i} \\ T_2 & \mathbb{i} & \mathbb{D} & \mathbb{i} \\ T_3 & \mathbb{i} & \mathbb{i} & \mathbb{D} \end{array}$$

Toeplitz structure

$$\begin{array}{cccc} & T_1 & T_2 & T_3 \\ T_1 & \mathbb{D} & \mathbb{i} & \mathbb{ii} \\ T_2 & \mathbb{i} & \mathbb{D} & \mathbb{i} \\ T_3 & \mathbb{ii} & \mathbb{i} & \mathbb{D} \end{array}$$

Combining longitudinal and sample correlation

When both longitudinal and sample correlations exist, the repeated measure correlation matrix is all combinations of time points and repeated samples

$$
\begin{array}{ccccccc}
 & (T_1, S_1) & (T_2, S_1) & (T_3, S_1) & (T_1, S_2) & (T_2, S_2) & (T_3, S_2) \\
(T_1, S_1) & \mathbb{D} & \mathring{\imath} & \mathring{\imath} & \mathring{\imath}\mathring{\imath} & \mathring{\imath}\mathring{\imath}\mathring{\imath} & \mathring{\imath}\mathring{\imath}\mathring{\imath} \\
(T_2, S_1) & \mathring{\imath} & \mathbb{D} & \mathring{\imath} & \mathring{\imath}\mathring{\imath}\mathring{\imath} & \mathring{\imath}\mathring{\imath} & \mathring{\imath}\mathring{\imath}\mathring{\imath} \\
(T_3, S_1) & \mathring{\imath} & \mathring{\imath} & \mathbb{D} & \mathring{\imath}\mathring{\imath}\mathring{\imath} & \mathring{\imath}\mathring{\imath}\mathring{\imath} & \mathring{\imath}\mathring{\imath} \\
(T_1, S_2) & \mathring{\imath}\mathring{\imath} & \mathring{\imath}\mathring{\imath}\mathring{\imath} & \mathring{\imath}\mathring{\imath}\mathring{\imath} & \mathbb{D} & \mathring{\imath} & \mathring{\imath} \\
(T_2, S_2) & \mathring{\imath}\mathring{\imath}\mathring{\imath} & \mathring{\imath}\mathring{\imath} & \mathring{\imath}\mathring{\imath}\mathring{\imath} & \mathring{\imath} & \mathbb{D} & \mathring{\imath} \\
(T_3, S_2) & \mathring{\imath}\mathring{\imath}\mathring{\imath} & \mathring{\imath}\mathring{\imath}\mathring{\imath} & \mathring{\imath}\mathring{\imath} & \mathring{\imath} & \mathring{\imath} & \mathbb{D}
\end{array}
$$

Incorporating taxonomic structure with repeated measures

$\mathbf{\Omega}$ with dimension L, for $a, b = 1, \ldots, N$,

$$\mathbf{\Omega}(\Gamma_{ab}) = \begin{pmatrix} \rho_{(\Gamma_{ab},\Omega_{11})} & \cdots & \rho_{(\Gamma_{ab},\Omega_{1L})} \\ \vdots & \ddots & \vdots \\ \rho_{(\Gamma_{ab},\Omega_{L1})} & \cdots & \rho_{(\Gamma_{ab},\Omega_{LL})} \end{pmatrix}.$$

$$\boldsymbol{R} = \begin{pmatrix} \mathbf{\Omega}^{11} & \cdots & \mathbf{\Omega}^{1N} \\ \vdots & \ddots & \vdots \\ \mathbf{\Omega}^{N1} & \cdots & \mathbf{\Omega}^{NN} \end{pmatrix}$$

Dimension of $R = (N \times L) \times (N \times L)$
Diagonals of $R = \rho(\mathbb{D}, \mathbb{D})$ are 1, off-diagonals need to be estimated

Example R

For two correlated OTUs and two repeated measures at different time points

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{\rho}_{(\mathbb{D},\mathbb{D})} & \boldsymbol{\rho}_{(\mathbb{D},\mathbb{i})} & \boldsymbol{\rho}_{(\mathbb{I},\mathbb{D})} & \boldsymbol{\rho}_{(\mathbb{I},\mathbb{i})} \\ \boldsymbol{\rho}_{(\mathbb{D},\mathbb{i})} & \boldsymbol{\rho}_{(\mathbb{D},\mathbb{D})} & \boldsymbol{\rho}_{(\mathbb{I},\mathbb{i})} & \boldsymbol{\rho}_{(\mathbb{I},\mathbb{D})} \\ \boldsymbol{\rho}_{(\mathbb{I},\mathbb{D})} & \boldsymbol{\rho}_{(\mathbb{I},\mathbb{i})} & \boldsymbol{\rho}_{(\mathbb{D},\mathbb{D})} & \boldsymbol{\rho}_{(\mathbb{D},\mathbb{i})} \\ \boldsymbol{\rho}_{(\mathbb{I},\mathbb{i})} & \boldsymbol{\rho}_{(\mathbb{I},\mathbb{D})} & \boldsymbol{\rho}_{(\mathbb{D},\mathbb{i})} & \boldsymbol{\rho}_{(\mathbb{D},\mathbb{D})} \end{pmatrix}$$

- $\rho_{(\mathbb{D},\mathbb{D})} = 1$
- $\rho_{(\mathbb{D},\mathbb{i})}, \rho_{(\mathbb{I},\mathbb{D})}$ correlation between two time points and two OTUs
- $\rho_{(\mathbb{I},\mathbb{i})}$ correlation from different OTU and different time points

Introduction to MTLC:

MTLC:

- Estimate predictor effects
- Estimate correlation coefficients between OTUs, longitudinal measures and other repeated measures
- Perform hypothesis testing of predictor effects

Generalized estimating equation framework

- $\boldsymbol{y}_k = (y_{k1}, \ldots, y_{kJ_k})$ clusters, length J_k for $k = 1, \ldots, K$
- $\mathbf{x}_{kj}$ the vector of covariates with length p, $j = 1, \ldots, J_k$
- $\boldsymbol{\mu}_k = (\mu_{k1}, \ldots, \mu_{kJ_k})$ mean of $\boldsymbol{y}_k$
- Each observation y_{kj}

$$g(\mu_{kj}) = \mathbf{x}'_{kj}\boldsymbol{\beta}$$

- Conditional variance of y_{kj}

$$\mathrm{Var}(y_{kj}|\boldsymbol{x}_{kj}) = v(\boldsymbol{\mu}_{kj})\phi$$

v is the variance function depending on the distribution of y_{kj}, ϕ is dispersion parameter

cont.

- Estimate β by solving the generalized estimating equation

$$U(\beta) = \sum_{k=1}^{K} \boldsymbol{D}'_k \boldsymbol{V}_k^{-1}(\boldsymbol{y}_k - \boldsymbol{\mu}_k) = 0$$

- $\boldsymbol{D}_k = \frac{d\boldsymbol{\mu}_k}{d\boldsymbol{\beta}}$, $\boldsymbol{V}_k = \boldsymbol{A}_k^{1/2}\boldsymbol{R}_k(\boldsymbol{\rho})\boldsymbol{A}_k^{1/2}$,
- $\boldsymbol{A}_k = diag(\mu_{k1}\phi, \ldots, \mu_{kJ_k}\phi)$ $\boldsymbol{\rho}$ collection of all correlation coefficients in $\boldsymbol{R_k}$
- $\boldsymbol{R}_k(\boldsymbol{\rho})$ is the working correlation matrix following correlation structure R

cont.

- $\phi, \boldsymbol{\rho}$ also need to be estimated

$$\hat{\phi} = \frac{1}{\sum_{k=1}^{K} J_k - p} \sum_{k=1}^{K} \sum_{j=1}^{J_k} e_{kj}^2$$

  where e_{kj} is the Pearson residual

- $\hat{\boldsymbol{\rho}}$ is estimated as a function of ϕ and e_{kj}, depending on the correlation structure R
- Iterative - switch between estimating $\boldsymbol{\beta}$ from fixed value of $\hat{\phi}$ and $\hat{\boldsymbol{\beta}}$ and estimating ϕ and $\boldsymbol{\rho}$ for a fixed value of $\hat{\boldsymbol{\beta}}$

Hypothesis testing

- From GEE theory $\hat{\boldsymbol{\beta}}$ is asymptotically normally distributed with mean $\boldsymbol{\beta}$ and variance $\boldsymbol{V}_{\boldsymbol{\beta}} = (\Sigma_{k=1}^{K}\boldsymbol{D}_{\boldsymbol{k}}'\boldsymbol{V}_{\boldsymbol{k}}^{-1}\boldsymbol{D}_{\boldsymbol{k}})^{-1}\{\Sigma_{k=1}^{K}\boldsymbol{D}_{\boldsymbol{k}}'\boldsymbol{V}_{\boldsymbol{k}}^{-1}\mathrm{Cov}(\boldsymbol{y}_{\boldsymbol{k}})\boldsymbol{V}_{\boldsymbol{k}}^{-1}\boldsymbol{D}_{\boldsymbol{k}}\}(\Sigma_{k=1}^{K}\boldsymbol{D}_{\boldsymbol{k}}'\boldsymbol{V}_{\boldsymbol{k}}^{-1}\boldsymbol{D}_{\boldsymbol{k}})^{-1}$
- Wald test statistic for testing $H_0 : \boldsymbol{C}\boldsymbol{\beta} = \boldsymbol{c}$

$$\boldsymbol{W} = (\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{c})'(\boldsymbol{C}\hat{\boldsymbol{V}}_{\beta}\boldsymbol{C}')^{-1}(\boldsymbol{C}\hat{\boldsymbol{\beta}} - \boldsymbol{c})$$

- $W \xrightarrow{d} \chi^2_{(q)}$ where q is the rank of $\boldsymbol{C}$

Estimating predictors effects on OTUs

- Two part model - two separate GEE models
  - Convert quantitative OTU observations to binary outcomes indicating prevalence of OTU in each observation - assessing predictor effects on OTU prevalence
  - Relative abundance of non-zero observation - assume RA follows normal distribution after log transformation - predictor effects on positive RA
  - Combine test statistics from two models for overall predictor effects

OTU GEE Model

- Assume each OTU observation y_{kj} follows a mixture of Bernoulli and log-normal distribution.
- OTU prevalence: $y_{kj}^{(0)}$ follows a Bernoulli distribution with $P(Y_{kj}^{(0)} = 1) = \mu_{kj}^{(0)}$
- Log-transform positive RAs: $y_{kj}^{(+)}$ follows a normal distribution

$$F(y) = \begin{cases} 1 - \mu_{kj}^{(0)} & y = 0 \\ 1 - \mu_{kj}^{(0)} + \mu_{kj}^{(0)} \Phi(\log y) & y > 0 \end{cases}$$

GEE Model for OTUs

- For OTU prevalence, use logit link function

$$\log \frac{\mu_{jk}^{(0)}}{1-\mu_{jk}^{(0)}} = \boldsymbol{x}'_{kj}\boldsymbol{\beta}^{(0)}$$

- For Log-transform RA, use identiy link function

$$\mu_{jk}^{(+)} = \boldsymbol{x}'_{kj}\boldsymbol{\beta}^{(0)}$$

- Use GEE framework to find parameter estimates $\hat{\boldsymbol{\beta}}^{(0)}$ and $\hat{\boldsymbol{\beta}}^{(+)}$

Hypothesis testing

- Test if the predictors have effects on either the prevalence of OTUs or the quantitative amount of RA,

$$H_0 : \boldsymbol{C}^{(0)}\boldsymbol{\beta}^{(0)} = \boldsymbol{c}^{(0)} \text{ and } H_0 : \boldsymbol{C}^{(+)}\boldsymbol{\beta}^{(+)} = \boldsymbol{c}^{(+)}$$

- Calculate Wald test statistics $W^{(0)}$ and $W^{(+)}$
- Cauchy combination test

$$W_{MTLC} = 0.5tan[(0.5 - p^{(0)})\pi] + 0.5tan[(0.5 - p^{(+)})\pi] \xrightarrow{d} Cauchy(0, 1)$$

Estimating correlation coefficients

- Estimated values of correlation coefficients $\hat{\boldsymbol{\rho}}^{(0)}$ and $\hat{\boldsymbol{\rho}}^{(+)}$ may be different.
- When Pearson correlations are available to compute, the GEE estimates are similar.

Discussion

- MTLC accounts for taxonomic correlation structure and longitudinal correlation structure
- MTLC has accurate Type I error, unbiased estimation of model parameters and robust power performance
- Correlation estimation is consistent
- Does not put a constraint on range of correlation coefficient
- Recommend using a subset of OTUs as model is time consuming when $N > 1000$.

Thank you!

References

Chen B, Xu W (2020)
Generalized estimating equation modeling on correlated microbiome sequencing data with longitudinal measures
PLoS Comput Biol 16(9): e1008108.
https://doi.org/10.1371/journal.pcbi.1008108.