Appendix A: Contingency Table, Mathematical Models and Likelihood Calculation

To study the effect of time since glacial retreat, distance between chronosequences, and soil depth on microbial community structure we used the number of samples belonging to each community type (group, G), identified through clustering (see the clustering subsection of the Statistical Analysis section and Figure 1), to construct a contingency table. Each row of this table corresponded to a combination of these factors (time, chronosequence, and depth). The columns corresponded to the different community groups. Each cell within a row contained a count representing the number of samples belonging to a certain community group that were collected from a particular time, chronosequence and depth. This resulted in a table with 24 rows (6 times x 2 chronosequences x 2 depths) and a number of columns equal to the number of groups resulting from the clustering step (Table 1). Groups that contained only 1-3 communities were regarded as outliers and were combined into one group. The rows of this contingency table were assumed to be independent samples from a multinomial distribution or multiple multinomial distributions depending on the model chosen to analyze the data as described below; see Agresti, [2002] for examples of such constructions.

Let $p_{ijkl}$ be the probability of obtaining a sample with a microbial community that belongs to community group $l$ and associated with the $i^{th}$ time, $j^{th}$ chronosequence, and $k^{th}$ depth (i.e., that occupies cell $ijkl$ of the contingency table). Also, let $x_{ijkl}$ be the count of microbial communities within cell $ijkl$, and $n_{ijk}$ (= 5) be the sample size corresponding to the $i^{th}$ time, $j^{th}$ depth and $k^{th}$ location. Finally, let $T$, $C$, $D$, and $G$ be the number of time periods (6), chronosequences (2), depths (2), and the number of soil microbial community groups, respectively. Then the multinomial probability distribution governing a row with count data
\(\hat{x}_{ijk} = \{x_{ijk1}, x_{ijk2}, \ldots, x_{ijk(G-1)}\}\) can be written in terms of the probability of observing the count data in that particular row, given the sample size \(n_{ijk}\) and the set of parameters \(\hat{p}_{ijk} = \{p_{ijk1}, p_{ijk2}, \ldots, p_{ijk(G-1)}\}\), as follows:

\[
Pr(\hat{x}_{ijk} | n_{ijk}, \hat{p}_{ijk}) = \left(x_{ijk1} x_{ijk2} \ldots x_{ijkG}\right)^{n_{ijk}} \prod_{l=1}^{G} p_{ijkl}^{x_{ijkl}} \tag{A1}
\]

where \(x_{ijkG} = n_{ijk} - \sum_{l=1}^{G-1} x_{ijkl}\), and \(p_{ijkG} = 1 - \sum_{l=1}^{G-1} p_{ijkl}\).

Equation (A1) represents the likelihood of the observed count data in that particular row, given the predetermined sample size and the true parameters governing that sample. Using the assumption that the rows are independent, the likelihood of the data presented in the contingency table can be written as follows:

\[
Pr(X | \bar{n}, P) = \prod_{i=1}^{I} \prod_{j=1}^{J} \prod_{k=1}^{G} Pr(\hat{x}_{ijk} | n_{ijk}, \hat{p}_{ijk}) \tag{A2}
\]

where \(X\) is a matrix of count data with rows equal to \(\hat{x}_{ijk} = \{x_{ijk1}, x_{ijk2}, \ldots, x_{ijk(G-1)}\}\); \(P\) is a matrix of parameters with rows equal to \(\hat{p}_{ijk} = \{p_{ijk1}, p_{ijk2}, \ldots, p_{ijk(G-1)}\}\); and \(\bar{n} = \{n_{111}, n_{112}, n_{121}, n_{122}, \ldots, n_{611}, n_{612}, n_{621}, n_{622}\}\) is a vector of sample sizes associated with each row of the contingency table.
The parameter matrix $P$, and hence, the likelihood, depends on the model used to fit the data. Eight models were introduced to analyze this count data. The simplest model assumed no effect of time, chronosequence, or depth on the microbial community structure. This was equivalent to obtaining 24 independent samples from one multinomial distribution governed by one set of parameters equal to the number of columns minus one. We referred to this model as the *simple-null* model. The second model accounted for only the effect of time (*time-alone* model). Under this model, the rows of the contingency table represented independent samples from six multinomial distributions each of which corresponded to a time period and each of which had a number of parameters equal to the number of columns minus one. This model had six fold the number of parameters as the simple-null model aiming to account for more of the variability in the data. The third and fourth models accounted for the effect of the distance between the chronosequences (*chronosequence-alone*) and different soil depth (*depth-alone*), respectively. In these two cases the rows were assumed to be independent samples from two multinomial distributions (associated with each of the two chronosequences and two depths, respectively). The fifth model assumed that time since glacial retreat and soil depth (*time-depth* model) both influenced the microbial community structure. Under this model the rows of the contingency table were taken to be independent samples from twelve multinomial distributions each corresponding to a specific time-depth combination. Similar to the fifth model, the sixth model (*time-chronosequence*) and the seventh model (*chronosequence-depth*) accounted for the combined effect of time and chronosequence, and chronosequence and depth, respectively. Finally, the eighth model accounted for the combined effect of time, chronosequence, and depth. In this case all factors (time, chronosequence, and depth) were assumed important for explaining the variation in the microbial community structures of the observed samples. Hence, each row of
the contingency table was assumed to be an independent sample from a multinomial distribution specifically associated with that particular time-chronosequence-depth combination. The number of parameters of interest in this case was 24 times as many as those of the simple-null model. This last model was the most parameter rich and hereafter we referred to it as the saturated model.

In the following we give one example to show the procedure of evaluating the likelihood and the maximum likelihood estimates (MLEs) of the model parameters based on a sample in which there was no effect of time, location, or depth on the microbial community structure. A similar procedure can be applied for other models. Under this null model all rows of matrix $P$ are equal:

$$
\bar{p}_{111} = \bar{p}_{112} = \bar{p}_{121} = \bar{p}_{122} = \ldots = \bar{p}_{611} = \bar{p}_{612} = \bar{p}_{621} = \bar{p}_{622} = \bar{p} = \{p_1, p_2, \ldots, p_{G-1}\} \quad (A3)
$$

and hence, the model involves only $G-1$ parameters (the number of columns minus one). Based on this, equation (A1) can be rewritten to reflect this parameter structure as follows:

$$
Pr(X|\tilde{n}, P) = \prod_{i=1}^{T} \prod_{j=1}^{G-2} \prod_{k=1}^{G-2} Pr\left(\tilde{x}_{ijk} | n_{ijk}, \bar{p}\right) \quad (A4)
$$

To use equation A4 to calculate the likelihood of the data given this model we need to know the values of the unknown parameters, $\bar{p}$. To overcome this problem these parameters are replaced with their maximum likelihood MLEs. In this case it can be shown (see Bain and Engelhardt, 1991) that the MLE for any of these parameters of this simple-null model is:
Using these parameter estimates we can calculate the likelihood of the data \( L(\hat{n}, \hat{P}) \) as in equation (A6),

\[
L(\hat{n}, \hat{P}) = \prod_{i=1}^{P} \prod_{j=1}^{C} \prod_{k=1}^{D} Pr(\hat{x}_{ijk} | n_{ijk}, \hat{p})
\]  

(A6)

Table (A1) shows the parameters of interest and the MLEs for each of the eight models introduced in the Materials and Methods. The likelihood of the data given each of these models is then calculated using equation A6 and the estimated MLEs.
Appendix B: Step-wise model selection (Introduction to the likelihood ratio test and the bootstrap).

A step-wise model selection strategy was used to choose the model that best fit our data wherein nested models were sequentially compared using the likelihood ratio tests and the bootstrap. A nested model is defined to be a model resulting from constraining some of the parameters of a more general model (Burnham & Anderson, 2002). Eight possible models were identified that could be used to interpret the pattern observed in the contingency table. These models were the simple-null, time-alone, chronosequence-alone, depth-alone, time-depth, time-chronosequence, chronosequence-depth, and time-chronosequence-depth (saturated model). Each of these models allowed us to assess the effect of one or more of the factors of interest (time, chronosequence, and depth) on the resulting microbial community structure. Figure B1 introduces a directed graph that describes the nesting structure of these models. Note that the simple-null model is nested within all other models and that all models are nested within the saturated model.

Our step-wise procedure to select the best model began with the simple-null model and progressed towards the saturated model. In each of the resulting cases we calculated a log-likelihood test statistic. For example, to compare the time-alone model to the simple-null model we used the following equation to calculate the log-likelihood test statistic,

\[ -2 \ln(A_T^N) = -2 \ln \left( \frac{L(\hat{\pi}_T, \hat{\pi}_N)}{L(\hat{\pi}_T)} \right) \]  

(B1)
where $\Lambda_T^N = \frac{L(\hat{\pi}_N)}{L(\hat{\pi}_T)}$ is the ratio of the likelihood of the simple-null model, $N$, to the likelihood of the time-alone model, $T$. $\ln \left( \frac{L(\hat{\pi}_N)}{L(\hat{\pi}_T)} \right)$ represents the natural logarithm of this ratio. A description of these models, their parameterization, and the method used to calculate the associated likelihoods are presented in Appendix A and in the Materials and Methods section. Intuitively it is clear that if the two models were comparable in terms of explaining the data, then there would be no improvement in the fit if the more complex model (the time-alone) was used. Hence, we would expect the likelihood ratio to be close to one, and the log likelihood ratio to be nearly zero. Accordingly, we can conclude that incorporating information about the sites does not help explain a significant part of the variation. In this particular case, such a conclusion indicates that the microbial community structure did not change along the chronosequences. If the more complex model better fits the data than the simple-null, then the likelihood of the more complex model would be larger than that of the simple-null and the ratio would be less than one, and, hence, the $-2 \log$-likelihood [the log-likelihood test statistic] would be positive. The better the fit the smaller the likelihood ratio and the larger the log-likelihood test statistic will be.

To determine the p-values for the different test statistics we used a bootstrap strategy to construct the null distribution (Efron & Tibshirani, 1997). This approach was devised to overcome some of the limitations of the classical asymptotic approaches (Bickel & Doksum, 2001). We used the null model to generate a 1000 data replicates (simulated contingency tables) then calculated the log-likelihood test statistic using both the simple-null and the site-model for each of these data sets. This resulted in a distribution of the log-likelihood test statistic given that the null model is true. To test the significance of the difference in the fit of the two models we compared the test statistic calculated using the actual data to the null distribution. This allowed
us to find the proportion of times that the simulated test statistics were larger than or equal to the calculated test statistic. This proportion represented a bootstrap p-value. In our case, a difference was considered to be significant if the p-value was less than a Bonferroni-corrected, significance level of 0.05.

At first we compared the simple-null model to the time-alone, chronosequence-alone, and depth-alone models in three separate tests. The model with the smallest p-value was then chosen as the new “null” model to be compared to other models it was nested in. For example, let’s assume that the time-alone model had the largest statistically significant log-likelihood test statistic, in the second step the time-alone model would be compared against time-chronosequence and time-depth, respectively. Favoring one of these models over the time-alone model highlights an interaction effect between the two factors contributing to that model. For example, choosing the time-depth model indicates that the microbial community structure changes significantly over time occurs, but that this change differs between the soil layers. The third and final step was to compare the most saturated model to the model chosen in the second step. Choosing the most saturated model means that all factors influence the microbial community structure.

An R™ suite of functions were developed to calculate the log-likelihood test statistics and build the null distributions utilized in this step-wise model selection procedure. These functions can be found at http://www.webpages.uidaho.edu/~zabdo/Software_files/Schuette2009/Schuette2009.htm.
Appendix C: Akaike’s Information Criterion (AIC)

One of the drawbacks of the step-wise model selection procedure is the need to correct for multiple testing (Appendix B, materials and methods). Another disadvantage is the need to traverse through a network of models before we come to the best model. Accordingly, the choice of the best model may depend on which branch we traverse first; choosing a different branch might result in a different optimum model. The limitations can be overcome using the AIC that allows for the comparison of all models all at once (Burnham & Anderson, 2002). To do so we first calculate an AIC score for each model using equation (C1):

\[
AIC_m = -2\ln \left( L(\vec{n}, \vec{\hat{P}}_m) \right) + 2K
\]

(C1)

where \( AIC_m \) is the AIC score of model \( m \); \( L(\vec{n}, \vec{\hat{P}}_m) \) is the likelihood of the data given that we are using model \( m \) that has \( K \) parameters, a vector of sample sizes \( \vec{n} \), and the matrix of MLEs of model parameters, \( \vec{\hat{P}}_m \). For the simple-null model \( K \) is equal to the number of identified community types minus one (equals to five in the High Arctic soil-sample case). The best model is one that minimizes the AIC score. An intuitive explanation of the AIC is based on noticing the two parts forming this score. The first part is the likelihood of the data given the MLEs of the parameters. This part describes the fit of the proposed model. The second part represents an adjustment proportional to the number of parameters of the model that penalizes for too many parameters.
Table A1: Parameters and ML estimates for each of the eight models. We calculated the likelihood of the data given each of these models by using these ML estimates in equation (A6).

| Model                | Row Parameters                                                                 | Per cell MLE                                                                 |
|----------------------|--------------------------------------------------------------------------------|------------------------------------------------------------------------------|
| Simple-null          | $\vec{p} = \{p_1, p_2, \ldots, p_{G-1}\}$                                    | $\hat{p}_i = \frac{\sum\sum\sum x_{ijkl}}{\sum\sum\sum n_{ijk}}$            |
| Time-alone           | $\vec{p}_i = \{p_{i1}, p_{i2}, \ldots, p_{i(G-1)}\}$                         | $\hat{p}_{il} = \frac{\sum\sum\sum x_{ijkl}}{\sum\sum\sum n_{ijk}}$         |
| Chronosequence-alone | $\vec{p}_j = \{p_{j1}, p_{j2}, \ldots, p_{j(G-1)}\}$                         | $\hat{p}_{jl} = \frac{\sum\sum\sum x_{ijkl}}{\sum\sum\sum n_{ijk}}$         |
| Depth-alone          | $\vec{p}_k = \{p_{k1}, p_{k2}, \ldots, p_{k(G-1)}\}$                         | $\hat{p}_{kl} = \frac{\sum\sum\sum x_{ijkl}}{\sum\sum\sum n_{ijk}}$         |
| Time-chronosequence  | $\vec{p}_{ij} = \{p_{ij1}, p_{ij2}, \ldots, p_{ij(G-1)}\}$                  | $\hat{p}_{ijl} = \frac{\sum x_{ijkl}}{\sum n_{ijk}}$                        |
| Time-depth           | $\vec{p}_{ik} = \{p_{ik1}, p_{ik2}, \ldots, p_{ik(G-1)}\}$                  | $\hat{p}_{ikl} = \frac{\sum x_{ijkl}}{\sum n_{ijk}}$                        |
| Chronosequence-depth | $\vec{p}_{jk} = \{p_{jk1}, p_{jk2}, \ldots, p_{jk(G-1)}\}$                  | $\hat{p}_{jkl} = \frac{\sum x_{ijkl}}{\sum n_{ijk}}$                        |
| Saturated            | $\vec{p}_{ijk} = \{p_{ijk1}, p_{ijk2}, \ldots, p_{ijk(G-1)}\}$               | $\hat{p}_{ijkl} = \frac{x_{ijkl}}{n_{ijk}}$                                 |
**Figure Legend**

Fig.B1: Depiction of nested models. Simple-null is nested within time-alone, chronosequence-alone, and depth-alone. Time-alone is nested within time-depth and time-chronosequence; chronosequence-alone is nested within time-chronosequence and chronosequence-depth; and depth-alone is nested within time-depth and chronosequence-depth. Note that the simple-null model is nested within all other models, while all models are nested within the saturated model.
FIGURE B1

Simple-Null

Time-alone

Depth-alone

Chronosequence-alone

Time-depth

Time-chronosequence

Depth-chronosequence

Saturated

Direction of model selection