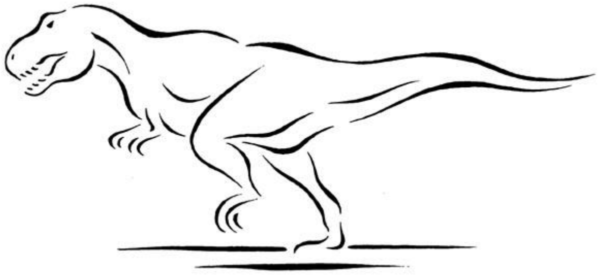


T-REX Software Documentation

T-RFLP Analysis Expedited (*T-REX*): Software for the

Processing and Analysis of T-RFLP data

http://trex.biohpc.org/

Please cite the program as:
Culman, S.W., Bukowski, R., Gauch, H.G., Cadillo-Quiroz, H., Buckley, D.H. 2009. *T-REX: Software for the Processing and Analysis of T-RFLP data.* BMC Bioinformatics 10:171

Last modified: June 29, 2009

Table of Contents

1. Overview

Despite increasing popularity and improvements in terminal restriction fragment length polymorphism (T-RFLP) and other molecular-based microbial community fingerprinting techniques, there are still some formidable barriers that plague the analysis of these datasets. Many steps are required to process raw data into a format ready for analysis and interpretation. These steps can be time-intensive, error-prone, and can introduce unwanted variability into the analysis.

We developed *T-REX* (**T-R**FLP analysis **EX**pedited), a free, web-based tool that was designed to address current obstacles in T-RFLP analysis.

T-REX **allows users to:**

- Label raw data with attributes related to the experimental design of the samples
- Determine a baseline threshold for identification of true peaks over noise
- Align T-RFs in all samples (bin T-RFs)
- Construct a two-way data matrix from labeled data and manipulate the matrix in a variety of ways
- Produce several measures of data matrix complexity, including the distribution of variance between main and interaction effects and sample heterogeneity
- Analyze a data matrix with the Additive Main Effects and Multiplicative Interaction Model (AMMI)

T-REX offers users a consolidated, flexible and rapid analysis of T-RFLP data.

2. T-REX at a glance

In this section, we briefly describe the general philosophy behind *T-REX* and its basic functions. These general ideas are presented in more detail in subsequent sections of this document.

To use *T-REX*, you need to open your web browser and navigate to trex.BioHPC.org. Typically, a *T-REX* session starts in the **Upload Data** page by the uploading of two files: the **raw data file** and the **label file**. The **raw data file** contains tabulated electropherogram peak information as exported from *GeneMapper©* or *PeakScanner™*. The **label file** associates each sample in the data file with a set of labels which describe the sample (e.g., treatment to which the sample belongs). *T-REX* is somewhat sensitive to the format of these files, so read section 6 carefully for detailed directions. Once these two files are uploaded to *T-REX*, the information is stored internally in the program's memory (more specifically – in SQL database) as a **project**. *T-REX* is all about accessing and processing this peak and label information in a variety of useful ways.

An important point is that, unless explicitly requested, *T-REX*'s operations **do not erase the uploaded peak data from memory**. Thus, **you never have to re-upload your data** in the middle of the analysis just because you want to re-run it with different parameters. *T-REX* will handle this task within the current session using the same data you uploaded in the beginning. If you are a **registered user**, your project will be stored on our server indefinitely (or until you choose to delete it) and you will be able to come back to it any time. If you are working as a **guest user**, your project will be there for the duration of a session only (i.e., it will be deleted after you close your browser or leave it inactive for more than an hour).

One of *T-REX*'s functions, which you may want to invoke right after uploading the data, is to examine and "weed through" the peaks and samples. *T-REX* helps you "eliminate" unwanted peaks either manually, via the **Sample Summary** page (section 8), or automatically using the **Filter Noise** function (section 10). In some cases, all peaks from a sample may be "eliminated", in which case the whole such sample will be marked as a **missing data** sample. When "eliminating" a peak, *T-REX* does not actually erase it from memory. Instead it just flags such peaks as **inactive**. If needed, you can revert it back to the **active** state using the same *T-REX* functions you used to "eliminate" or "deactivate" it. Only **active** peaks (i.e., the ones not marked as inactive) are taken into account in the downstream data analysis. When the data are first uploaded, all peaks are treated as active and it is up to you to weed through them, or you can just **skip this step** if you are confident it is not needed.

With the established set of **active peaks**, *T-REX* performs binning of fragment sizes. For example, if a peak from sample A has the size 63.2 and another peak from sample B is 63.1, chances are you would like to put them both into a bin of size 63bp. **In the language of T-REX, the size bins are referred to as T-RFs**. Assignment of T-RFs to peaks across all samples may be a nontrivial problem, and *T-REX* has two algorithms implemented to deal with this issue (see section 11 for details). The default is just to round-off each fragment size to the nearest integer, which defines the bin. This function is performed automatically after you first upload the data. Then, whenever the set of active peaks changes as a result of noise filtering or a manual change, the binning process is **automatically** repeated to make the T-RFs (bins) consistent with the current set of active peaks. This automatic T-RF update can be time-consuming and you may want to turn it off when you plan on performing frequent "weeding" or noise filtering operations. The button **Align T-RFs** gives you access to an option which disables automatic T-RF updates. Here you can also change the algorithm to be used for binning. **If you are happy with the default binning setup, you do not have to touch the Align T-RFs button at all.**

Samples are often organized into conceptually equivalent groups or replicates, referred to as **environments** in *T-REX*. If your samples are replicated, you can define the **environments** in the label file when uploading the data (see section 6.b). You can also easily change the defined **environments** in the **Environments** page after the data are uploaded (section 12).

Once the set of active peaks and environment assignments are configured to your satisfaction (note that in principle both these steps are optional), you can export the data using the **Export Labeled Data** page (section 9) for use in other statistical software packages. This function produces a tab-delimited file resembling the raw data file, but with labels and other specified information paired with each peak.

Alternatively, the current set of **active peaks** can be used to construct a data matrix in the **Data**

Matrix/ AMMI page. Using built-in filtering mechanisms, you can select samples and T-RFs to be included in this matrix. The newly constructed data matrix can be exported or carried over for further analyses. One of such analysis, the **Additive Main Effects and Multiplicative Interaction Model (AMMI)**, is built into *T-REX* and activated with a mouse click. The resulting tables, charts, and output files are than available for viewing and download.

Ordination results extracted from the Data Matrix may depend on a variety of factors, such as noise reduction threshold, environment (replicate) definitions, or filtering criteria applied during data matrix construction. *T-REX* offers a unique possibility of real-time analysis of these dependencies. For example, if having performed an AMMI analysis you suspect that the noise filtering threshold was initially set too high, you can always use the **Filter Noise** function with a different parameter and then re-run the analysis using the **Data Matrix/AMMI** function within the same session. **There is no need to re-upload the data and start a new project from scratch.**

3. About T-REX

T-REX was developed by Robert Bukowski under the guidance of Steve Culman, Hugh Gauch, Hinsby Cadillo-Quiroz and Dan Buckley at Cornell University, Ithaca, NY.

T-REX was written using Microsoft ASP.NET and MS SQL Server platforms on Windows Server 2003. It is hosted at the Computational Biology Service Unit (CBSU) at Cornell University.

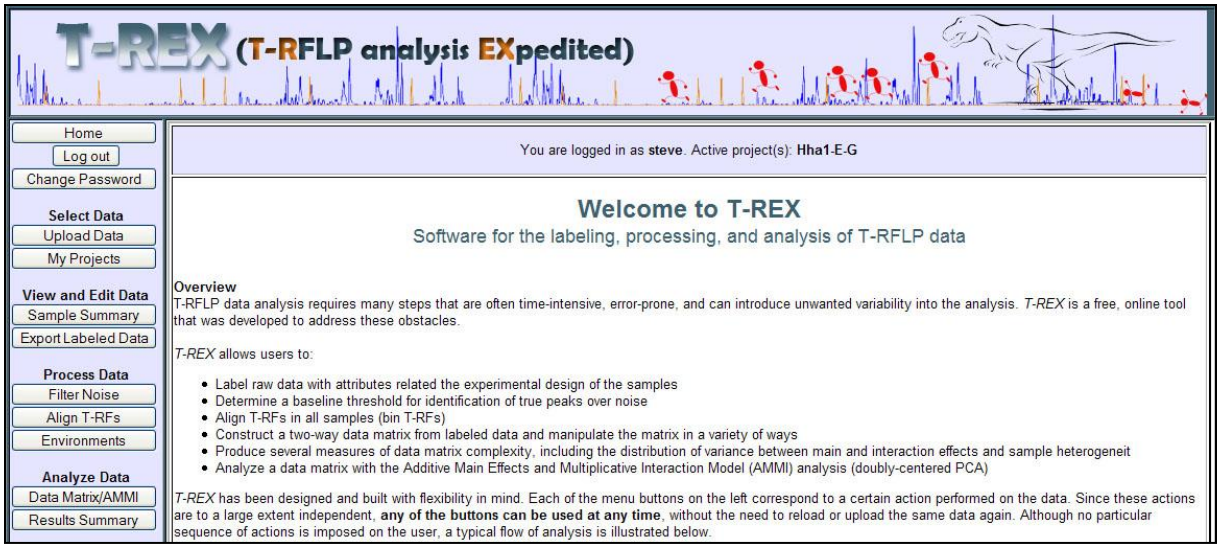
T-REX can be found at the web address: http://trex.biohpc.org/. The access to the program is free and requires only a web browser and an internet access.

In the future, the source code of T-REX will be available under the GNU GPL License.

Funding for *T-REX*'s development was made possible by the National Science Foundation's Biogeochemistry and Biocomplexity IGERT grant, and by the Microsoft Corporation.

4. Home Page

The **Home** page outlines the program’s features and introduces the user to the typical flow of analysis (Figure 1). Tabs on the left of the page direct the user to different functions of the program.



You are logged in as **steve**. Active project(s): **Hha1-E-G**

Welcome to T-REX

Software for the labeling, processing, and analysis of T-RFLP data

Overview

T-RFLP data analysis requires many steps that are often time-intensive, error-prone, and can introduce unwanted variability into the analysis. *T-REX* is a free, online tool that was developed to address these obstacles.

T-REX allows users to:

- Label raw data with attributes related the experimental design of the samples
- Determine a baseline threshold for identification of true peaks over noise
- Align T-RFs in all samples (bin T-RFs)
- Construct a two-way data matrix from labeled data and manipulate the matrix in a variety of ways
- Produce several measures of data matrix complexity, including the distribution of variance between main and interaction effects and sample heterogeneit
- Analyze a data matrix with the Additive Main Effects and Multiplicative Interaction Model (AMMI) analysis (doubly-centered PCA)

T-REX has been designed and built with flexibility in mind. Each of the menu buttons on the left correspond to a certain action performed on the data. Since these actions are to a large extent independent, **any of the buttons can be used at any time**, without the need to reload or upload the same data again. Although no particular sequence of actions is imposed on the user, a typical flow of analysis is illustrated below.

Figure 1. Screenshot of *T-REX* Home Page.

5. Guests and Registered Users

You can work as a guest or, if you have an account, as a registered user. Registered users can have up to 25 unique projects saved on the server at any time. Saved projects can later be concatenated, renamed, or deleted. Guests are allowed everything except storing their data on the server.

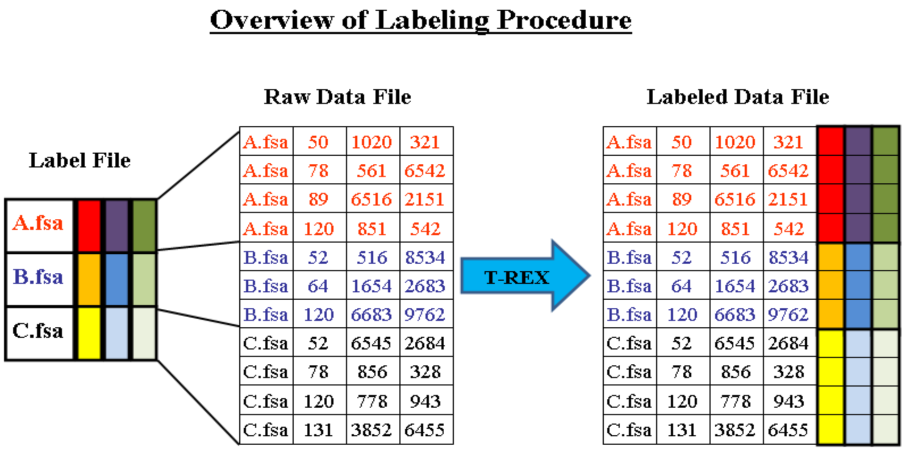
If you would like to register, please email your request to cbsu@cornell.edu with a short explanation of the nature of your research.

6. Upload Data

The first step in using *T-REX* is to upload and label the data. This process happens simultaneously and requires two files:

- **Raw data file**. This is the tabulated file that is exported from GeneMapper™, Peak Scanner™, or similar size-calling software that contains the peak information for a set of samples.
- **Label file**. This contains a set of labels/attributes that describe each sample and often correspond to factors in the experimental design.

A user should identify the raw data file and the label file by selecting the 'Browse' button to locate the saved files. A new project is created when a user uploads and labels data. Registered users should specify a unique name for this project as it will be stored on the server, and can later be modified. Alternatively, a registered user can upload new raw data to a specified existing project. The name of the active project and the registered user is displayed in the blue box under the *T-REX* header icon. Once uploaded, the data will be labeled, i.e., each peak in the raw data file will have the corresponding labels attached, as illustrated in Figure 2.



Label File

A.fsa	🟥	🟪	🟩
B.fsa	🟧	🟦	🟩
C.fsa	🟨	🟦	⬜

Raw Data File

A.fsa	50	1020	321
A.fsa	78	561	6542
A.fsa	89	6516	2151
A.fsa	120	851	542
B.fsa	52	516	8534
B.fsa	64	1654	2683
B.fsa	120	6683	9762
C.fsa	52	6545	2684
C.fsa	78	856	328
C.fsa	120	778	943
C.fsa	131	3852	6455

Labeled Data File

A.fsa	50	1020	321	🟥	🟪	🟩
A.fsa	78	561	6542	🟥	🟪	🟩
A.fsa	89	6516	2151	🟥	🟪	🟩
A.fsa	120	851	542	🟥	🟪	🟩
B.fsa	52	516	8534	🟧	🟦	🟩
B.fsa	64	1654	2683	🟧	🟦	🟩
B.fsa	120	6683	9762	🟧	🟦	🟩
C.fsa	52	6545	2684	🟨	🟦	⬜
C.fsa	78	856	328	🟨	🟦	⬜
C.fsa	120	778	943	🟨	🟦	⬜
C.fsa	131	3852	6455	🟨	🟦	⬜

Figure 2. Illustration of labeling procedure in *T-REX*.

6.a. File Formats

T-REX requires specific data file formats. The raw data and label files must be in tab-delimited format. The **Raw Data File** must **NOT** contain a header line and should look similar to this:

B,1	A.fsa	51.22	3647	73224	1364
B,2	A.fsa	57.04	109	714	1430
B,3	A.fsa	62.18	113	847	1486
B,1	B.fsa	51.09	3246	66742	1350
B,2	B.fsa	55.26	77	570	1396
B,3	B.fsa	65.86	84	1371	1512
B,1	C.fsa	50.91	3852	77118	1349
B,2	C.fsa	58.25	54	390	1429
B,3	C.fsa	61.01	127	1235	1459

The columns should correspond to those of a typical Genemapper™ or Peak Scanner™ exported file, with columns representing (from left to right): Dye/Sample Peak, Sample File Name, Size, Height, Area, and Data Point (since the Data Point column is not actually used by T-REX, the values in this column may be arbitrary). Entries in the first column may contain quotation marks and spaces, for example, B,3 would be equivalent to “B,3” or “B, 3”, but must not contain tab characters, as these characters are used as column separators. If the number of entries in a line is not equal to 6, such a line will be ignored (not uploaded).

The **Label File** should contain a header line and the first entry **MUST** be called **FileName**. The file should be similar to this:

FileName	Date	Site	Depth	Replicate
A.fsa	1	C	1	1
B.fsa	1	C	1	2
C.fsa	1	C	3	1

Column headings in the label file can be up to 20 characters and must not contain spaces. Up to 10 attributes (columns) are allowed in the label file, and should represent some aspect of the experimental design. Column headings **Replicate** and **Environment** have special meaning: if present, these headings will be used to organize samples into groups of replicates (see below).

If the header line is not included in the Label File, generic column headings label1, label2, etc. will be assigned. These headings can be changed later using the **Sample Summary** page. All file names in the label file must be unique. Example raw data and label files are available for download on the **Upload Data** page.

Below is a checklist of the required formats for both file types:

<u>Raw Data File</u>

- ✓ Tab-delimited text file
- ✓ No column headings
- ✓ File names (second column) at most 50 characters long
- ✓ Six columns total
- ✓ Entries "Dye,Peak" (e.g., "B,5") must not contain tabs
- ✓ No empty entries (i.e., no two consecutive TAB characters)
- ✓ No TAB characters at the end of any line
- ✓ No blank lines (in particular, be sure to remove any blank lines from the end of the file)

<u>Label Data File</u>

- ✓ Tab-delimited text file
- ✓ Column headings included with the first entry called **FileName**
- ✓ Column headings and entries contain no spaces and are no more than 20 characters long (except for entries in **FileName** column which may me up to 50 characters long)
- ✓ Column headings **Replicate** and **Environment** have special meanings
- ✓ There should be at least one column (label) different from **FileName**, **Environment**, or **Replicate**
- ✓ All sample names are unique

- ✓ No empty entries (i.e., no two consecutive TAB characters)
- ✓ No TAB characters at the end of any line
- ✓ No blank lines (in particular, be sure to remove any blank lines from the end of the file)

Do not use long names for the raw data file and the labels file! The length of each file's name should be less than 50 characters. Also, if you are working as a guest, the combined length of both file names should be less than 50 characters.

6.b. Defining Replicates

T-REX performs several functions that take advantage of information provided by replicated data (i.e., samples that are conceptually identical, or belong to the same environment/treatment). Replicated samples are organized into groups called **environments**. All samples belonging to a given **environment** are treated as replicates of one another. In *T-REX*, each environment has a unique identifier—a positive integer. At any given time, each sample belongs to only one environment and the corresponding environment identifier is displayed in the **Env** column on **Sample Summary** page. In the case where data are not replicated, each environment will contain only one sample—that is every sample will be assigned a unique environment (integer).

Defining Replicates during data upload

By default, when uploading the data, *T-REX* will automatically assign samples to environments based on all labels, i.e., two samples will be considered belonging to the same environment if they have the same sets of labels. This default behavior can be changed if the Label File contains a header line and one of the two special columns: **Environment** or **Replicate**. The first option is likely most useful for experiments with simple experimental designs. Using the word **Environment** as a column header in the label file will designate all samples with the same value in this **Environment** as replicates. The second approach is potentially more flexible for complex experimental designs. With the word **Replicate** as a column header in the label file, the program will designate replicates as samples whose attributes in the label file are all identical, **except** the value in the **Replicate** column. If both **Environment** and **Replicate** are found in label file, the program will assign replicates by the **Environment** attribute. If neither the **Environment** nor **Replicate** column is provided in the labels file (or if this file does not have a header line at all), the default behavior is expected.

Defining Replicates in the Sample Summary Page

Since replication can occur at multiple scales (e.g., analytical, field), manual manipulation was made possible to allow the user more flexibility in the analysis of a T-RFLP dataset. Replicates can be defined at any time by manually entering in numbers (positive integers) in the **Env** column in the **Sample Summary** page. All samples with the same numbers will be treated as replicates. When all numbers are assigned, clicking the 'Submit' button on top of the **Env** column will change any previous designations to the newly defined replicates.

Defining Replicates using the Environments page

Replicates can be defined automatically based on label comparison using the **Environments** page. By checking checkboxes, the user can define a set of labels which determine an environment. Two samples will be considered replicates (i.e., belonging to the same environment) if they have identical sets of checked label values.

6.c. Missing Data

Missing data occurs when one or more samples are omitted from the analysis. Missing data can result from multiple scenarios. First, a researcher could throw a sample out after the visual inspection of the electropherogram showed that the run was of too poor quality to be meaningful. Genemapper™ software can also omit poor-quality samples from the exported Genemapper™ file. (In this scenario, the sample will not be present in the raw data file.) Missing data can also arise from noise filtering procedure which can eliminate whole samples from down-stream analyses.

T-REX is able to appropriately deal with all possible cases of missing data. In scenarios when there is a discrepancy regarding samples between the raw data file and label file the program will enter and record these non-matched samples, but mark them as missing data. More specifically, if the label file contains a file name that is not represented in the raw data file, the corresponding sample will be entered into the system and marked as missing data. Likewise, if the raw data file contains file names not in the label file, the samples corresponding to these missing names will be marked as missing data and zeroes will be added as labels. In an extreme case when a label file is not supplied, all samples in the raw data file will be marked as missing data. Users have the option of manually supplying the labels and removing the missing data mark from the **Sample Summary** page.

Missing data are stored in the active project, but do not affect any procedures or manipulations of data. They are marked in red and can be viewed in the **Sample Summary** page by clicking ‘Show Details’.

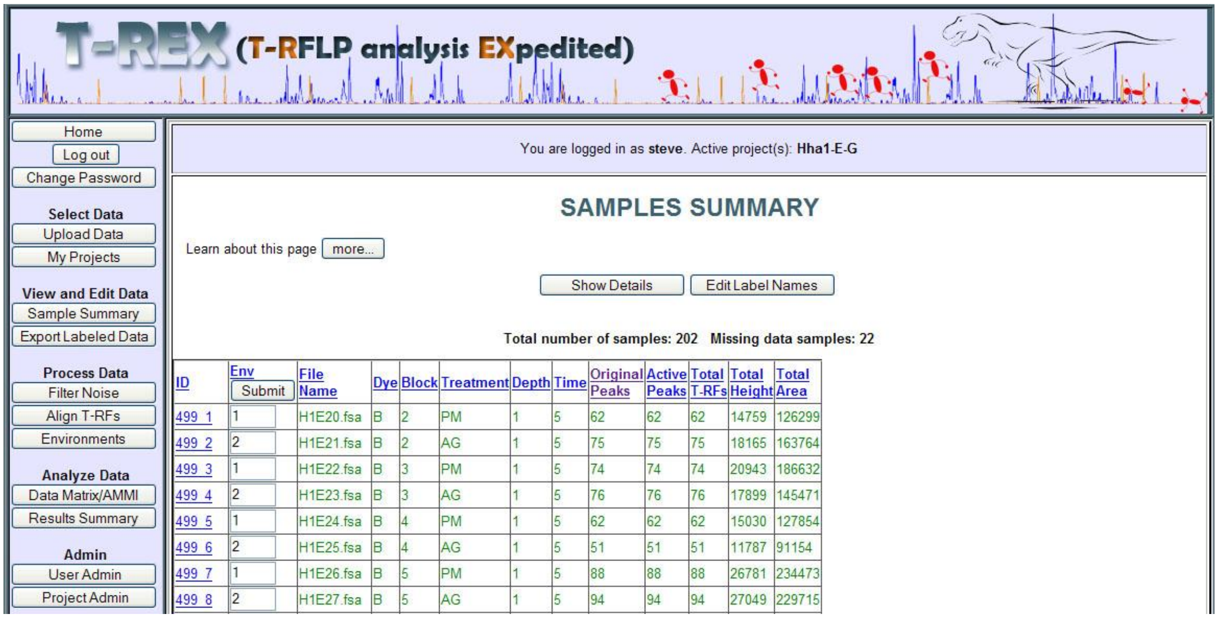
7. My Projects

This function is available only to registered users. Once a project is created, it can be renamed, merged, or deleted in the **My Projects** page. Users can also come back to pre-existing projects and load them using this page for further manipulation. Merging two or more projects is possible only if the label names are consistent across all projects being combined.

8. Sample Summary Page

The **Sample Summary** page is synonymous to the home page of a particular project (Figure 3). All samples are consolidated to show the total number of peaks, total peak height and peak area, as well as the properties relating to the experimental factors assigned in the labeling procedure. The identifier of the environment a given sample belongs to is shown in the **Env** column. The **Sample Summary** page also shows users the effect of the noise elimination procedure on the number of active peaks.

Individual samples can be viewed, edited, and even removed from the analysis in the **Sample Details** page, accessible by selecting an individual sample's **ID** in the **Sample Summary** page. Once viewing an individual sample, the user will see a reconstructed graphic of the electropherogram as well as tabulated individual peak properties. Users are able to manipulate labels, deactivate individual peaks of that sample, or mark the entire sample as missing data within the project in this page.



You are logged in as **steve**. Active project(s): **Hha1-E-G**

SAMPLES SUMMARY

Learn about this page

Total number of samples: 202   Missing data samples: 22

ID	Env	File Name	Dye	Block	Treatment	Depth	Time	Original Peaks	Active Peaks	Total T-RFs	Total Height	Total Area
499 1	1	H1E20.fsa	B	2	PM	1	5	62	62	62	14759	126299
499 2	2	H1E21.fsa	B	2	AG	1	5	75	75	75	18165	163764
499 3	1	H1E22.fsa	B	3	PM	1	5	74	74	74	20943	186632
499 4	2	H1E23.fsa	B	3	AG	1	5	76	76	76	17899	145471
499 5	1	H1E24.fsa	B	4	PM	1	5	62	62	62	15030	127854
499 6	2	H1E25.fsa	B	4	AG	1	5	51	51	51	11787	91154
499 7	1	H1E26.fsa	B	5	PM	1	5	88	88	88	26781	234473
499 8	2	H1E27.fsa	B	5	AG	1	5	94	94	94	27049	229715

Figure 3. Screenshot of *T-REX* Samples Summary page.

Below are specific definitions related to the **Sample Summary** page:

- **ID** is a unique identifier of a sample, generated when raw data are uploaded to the program, creating a new project. Clicking on ID will allow you to view and edit details of that individual sample.
- **Env** is an integer number denoting the **environment** (group of replicates) a given sample belongs to. If a set of samples have the same environment value, they will be treated as replicates. This field can be edited manually by entering values and clicking the 'Submit' button in the column header.
- **Original Peaks** designate the total number of peaks originally present in the uploaded

raw data file.

- **Active Peaks** is the number of peaks which have not been filtered out as noise or manually marked as inactive.
- **Total T-RFs** is the number of T-RFs present in the sample. If different from **Active Peaks**, the T-RFs are not up to date and **Align T-RFs** procedure should be run. In such a case, a star ("*") is added to the number in the **Total T-RFs** field.
- **Total Height** and **Total Area** are the cumulative peak height and areas, respectively, summed over all active peaks.

Additional details related to the samples can be viewed by clicking on 'Show Details' button:

- **Removed Peaks** is the total number of peaks from the raw data file that have been made inactive manually or as a result of noise elimination function.
    - **Noise** indicates how many peaks were deactivated in the **Filter Noise** procedure.
    - **Manually** indicates how many peaks were deactivated manually by the user.

Samples that have had no peaks deactivated (i.e., **Original Peaks** equal **Active Peaks**) are shown in **green**. Samples with some of the peaks inactive are shown in **brown**. Samples marked as missing data (i.e., all peaks made inactive) are shown in **red**. Missing data samples can be displayed or hidden using the 'Show/Hide Details' toggle button.

9. Export Labeled Data

The **Export Labeled Data** page was designed for users who want to take advantage of *T-REX*'s rapid labeling and/or data processing procedures, but analyze their data with another software program. After data are uploaded and labeled, users can export the labeled data directly, or can manipulate the data before exporting. The **Sample Summary** page indicates the current status of a project and will reflect the exact details of the data that will be exported.

Labeled data is exported as a simple text file with columns separated by a specified separator (tab-delimited by default). Using the checkboxes provided on the **Export Labeled Data** page, the user controls which of the data columns should be exported. These may include the peak and labels information uploaded initially as well as additional information generated by T-REX, such as T-RF (or bin size) a given peak belongs to, peak status (active/inactive), or the unique Sample ID. Once exported by clicking on the ‘Export’ button, the labeled data file can be uploaded via an http link.

10. Filter Noise

Determining "true peaks" (i.e., distinguishing peaks from background fluctuations in fluorescence) is often a major challenge in T-RFLP data analysis, as the baseline threshold can dramatically affect the community fingerprint and downstream analyses. A common procedure is to apply a researcher-determined baseline threshold across all samples to delineate true peaks from noise (Abdo et al., 2006; Blackwood et al., 2003; Dunbar et al., 2001). However, this threshold is often subjectively determined and may not be the most appropriate approach. Since the number of spurious peaks in a sample may increase when the amount of PCR product analyzed increases, the amount of noise relative to signal in a sample may be a methodological artifact. When DNA concentrations vary from sample to sample, determining true peaks from noise based on variability within each sample (Abdo et al., 2006), rather than a value across all samples is usually more appropriate.

T-REX uses the approach outlined by Abdo et al. (2006) to find true peaks and eliminate background noise. True peaks are identified as those whose height (or area) exceeds the standard deviation (assuming zero mean) computed over all peaks within a sample and multiplied by a factor (to be specified in the box provided). The procedure is then reiterated with the peaks which were not identified as true ones. The iterations continue until no new true peaks are found.

The filtering of peaks can be based on standard deviations of peak height or area and may be applied to all samples or just selected samples in the active project. Users should select an appropriate (fluor-specific) standard deviation multiplier based on the original electropherograms and results of the filtering procedure. The program allows for rapid manipulation of the multiplier and subsequent reviewing of results in the **Samples Summary** page if a user wants to determine an appropriate multiplier empirically. At any time the filtering procedure can be cleared and the data reverted to their original state with the ‘Clear filtering’ button.

11. Align T-RFs

The size (in base pairs) of every T-RF is determined by referencing the T-RF with an internal size standard. However, T-RFs can be improperly sized due to differences in fragment migration, purine content, and fluorophores (Kaplan and Kitts, 2003; Marsh, 2005). These analytical errors in determining fragment length (T-RF drift) are often corrected for by aligning peaks manually (Blackwood et al., 2003), aligning them automatically (Dunbar et al., 2001; Smith et al., 2005), or simply ignoring them and treating them as analytical error.

T-REX performs alignment of the current set of **active peaks**. The peaks flagged as inactive are always ignored by the alignment procedure. By default, the alignment is done **automatically** whenever a set of active peaks is established or whenever it changes. It happens first after the data is uploaded, at which point all peaks are considered active. Then, each time a peak (or set of peaks) is manually put in inactive state or filtered out as noise, the peak alignment is automatically repeated from scratch. If the default alignment settings are acceptable, the user is not required to take any special action. Otherwise, the **Align T-RFs** page can be used to switch to a different alignment algorithm and/or to turn off the automatic alignment (see below).

T-REX offers two approaches for an automated alignment of peaks which can be configured via the **Align T-RFs** page. The default approach rounds every T-RF up/down to the nearest integer size in base pairs. The second option models the approach taken by the software program *T-Align* (Smith et al., 2005). Briefly, the peak corresponding to the smallest fragment size is identified across all samples and tagged. Peaks within the size range given by the (user-specified) **clustering threshold** are then identified and grouped into a T-RF. The next smallest-fragment-size peak across all samples not falling into the first T-RF is identified and tagged. Peaks within the specified clustering threshold are identified and grouped with the second T-RF. This process continues until all peaks are grouped into T-RFs. The process leads to averaging of the positions of the peaks belonging to a given TRF, resulting in an average **T-RF position**. In this sense, a T-RF can be thought of as a "peak size averaged over samples". Since each peak is assigned to one T-RF (whose average position is very close to this peak's actual peak position), it is tempting to use the terms "peak" and "T-RF" as synonyms. However, it is important to understand that this association can be made only after the peaks have been aligned and the T-RFs have been found. Besides the average **T-RF position**, each T-RF is also assigned a **T-RF ID** – an integer which (unlike the average position) uniquely identifies this T-RF among the others. Both the average **T-RF position** (denoted simply as **T-RF**) and the **TRF ID** are shown in the **Sample Details** section of the **Sample Summary** page. Most likely, the T-RF ID is of no particular concern to the user. It is used mostly by *T-REX* itself to properly keep track of T-RFs.

It should be noted that both alignment approaches allow for more than one peak from the same sample to be assigned to the same T-RF (in the case of the *T-Align*-style algorithm, this default behavior may be changed by checking the box "At most one peak per plot included in each TRF"). This creates an ambiguity later on during the Data Matrix construction, since it is not immediately clear which peak should be considered to represent the sample for this T-RF. *T-REX* resolves this ambiguity simply by assuming that smaller peaks in a T-RF are most likely noise and selecting the largest (by height or area) peak as a representative. Allowing multiple peaks from the same sample to be grouped in one T-RF provides additional assurance (besides noise filtering) that noise peaks will not be included in the Data Matrix.

T-REX allows users to manipulate datasets with multiple fluors. Whichever the alignment method used, it is always applied separately within groups of (active) peaks with different fluor colors. For example, if samples in a project contain peaks associated with both blue and green fluors, the blue peaks will be aligned separately from the green ones. Therefore, each T-RF will always contain peaks of only one color, regardless of the T-RF's size. In particular, one may end up with two different T-RFs, both corresponding to the same size in base pairs, but to different colors. These two T-RFs will be treated as distinct by the downstream Data Matrix construction and AMMI analysis routines.

Since the automatic T-RF update (realignment) is somewhat time-consuming, it can be turned off when frequent changes of peak status are expected, for example, if the user plans on interactive adjustment of the noise filtering factors. The **Align T-RFs** page gives you access to an option which disables automatic T-RF updates. In such a case, after the peak operations are completed, peak alignment must be restored manually by clicking on the "Save settings and find T-RFs now" button on the **Align T-RFs** page. Otherwise, a warning will appear on the information bar on top of the page informing the user that the T-RFs are absent or outdated (i.e., not consistent with the current state of active peaks). The button mentioned above has to be used whenever any settings have been changed on the **Align T-RFs** page.

12. Environments

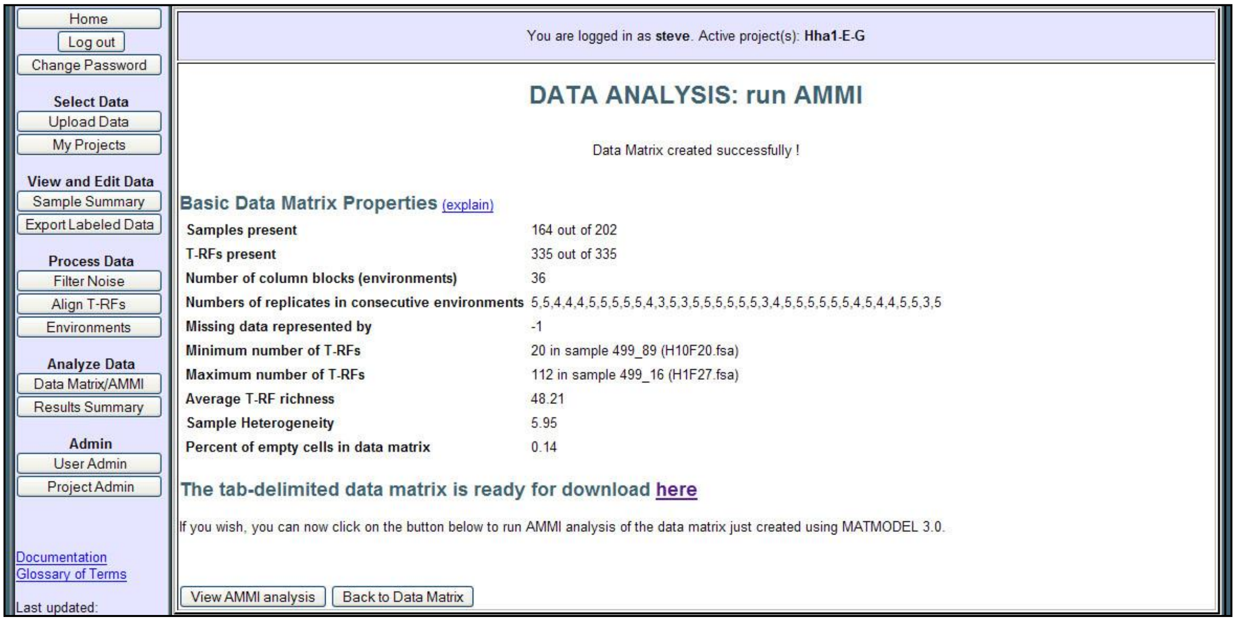
The **Environments** page allows users to rapidly classify samples into environments based on the given labels. This approach is especially useful when replication in an experiment occurred at multiple scales (e.g., analytical, field) and a user wants to compare results based on these different ways of defining replication. Users can assign and/or reassign replicated samples into environments by using the checkboxes to define the set of labels that determine an environment (check as many as you want). Samples will be considered replicates (i.e., belonging to the same environment) if they have identical sets of checked label values. Once all the appropriate boxes are checked, clicking the "Define Environments" button will redefine environments accordingly. Un-checking all the boxes and clicking on the button will group all samples together in one environment.

Users can view the currently defined environments by downloading a tab-delimited file provided via an html link on **Environments** page or by viewing the **Env** column in the **Samples Summary** page.

13. Data Matrix

Because of the complexity associated with T-RFLP and other microbial community datasets, multivariate statistical analyses are typically performed on these data to summarize the complex relationships of the microbial communities with their environments. Raw T-RFLP data exported from Genemapper™, Peak Scanner™, or similar size-calling software is typically in a tabulated or listed format, where one column contains all the records for each variable (i.e., one column for all T-RF sizes, one column for all peak heights, etc.). However, these data often need to be formatted into a two-way data matrix, where rows are indexed by the T-RFs and columns – by samples (or *vice versa*). Such a two-way format is required by many multivariate-focused statistical software packages. Assuming the researcher employed randomization on the lab bench to minimize bias throughout the entire analysis, the process of data matrix construction can be time-intensive and error-prone.

The **Data Matrix/ AMMI** page allows users to first construct a two-way data matrix and, second run the AMMI model on this data matrix. Data matrix construction involves six steps. **The first step** requires that all peaks be assigned to particular T-RFs (aligned). Unless specified otherwise by the user, this assignment is performed automatically, so typically no action is required here. If the page detects that the T-RFs are not up to date, it will ask the user to use the **Align T-RFs** function before proceeding. **In the second step**, the user is given a chance to re-define the assignment of samples to environments, if the current assignment is not satisfactory for any reason. **The third step** allows users to specify which type of data (presence/absence, peak height, or peak area) to use for data matrix construction, and if these data should be averaged across replicates and/or relativized. In the event that a sample has more than one peak belonging to a given T-RF, only the peak with the largest value of height or area (depending on what was chosen as the data type) is included in the Data Matrix**. The fourth step** allows users to select which samples should be included in the data matrix and subsequent analysis. Users have the option of selecting all samples, or just those with specific dye colors and/or specific labels. One can also omit samples with poor peak representation or too small total peak height and/or area. **The fifth step** allows rare T-RFs to be omitted from the Data Matrix. This step represents a final quality control process to be placed on the data matrix. Clicking on 'Create Data Matrix' in the **sixth step** will take the user to another page where a data matrix in tab-delimited format is ready, as well as output on basic data matrix properties, such as total samples and T-RFs present, maximum and minimum, and average number (richness) of T-RFs across samples, and sample heterogeneity (Figure 4). At this point the user is able to export this data matrix for analysis with another software package, or continue with the AMMI analysis by clicking 'View AMMI Analysis' (see section 14 below).



You are logged in as **steve**. Active project(s): **Hha1-E-G**

DATA ANALYSIS: run AMMI

Data Matrix created successfully !

Basic Data Matrix Properties (explain)

Samples present	164 out of 202
T-RFs present	335 out of 335
Number of column blocks (environments)	36
Numbers of replicates in consecutive environments	5,5,4,4,4,5,5,5,5,5,4,3,5,3,5,5,5,5,5,5,3,4,5,5,5,5,5,5,4,5,4,4,5,5,3,5
Missing data represented by	-1
Minimum number of T-RFs	20 in sample 499_89 (H10F20.fsa)
Maximum number of T-RFs	112 in sample 499_16 (H1F27.fsa)
Average T-RF richness	48.21
Sample Heterogeneity	5.95
Percent of empty cells in data matrix	0.14

The tab-delimited data matrix is ready for download here

If you wish, you can now click on the button below to run AMMI analysis of the data matrix just created using MATMODEL 3.0.

Figure 4. Screenshot of T-REX Create Data Matrix/ Run AMMI page.

13.a. Basic Data Matrix Properties

Below are the calculations for some of the Data Matrix properties:

- Average T-RF richness is the average number of T-RFs in a the dataset; average T-RF richness = sum of total number of T-RFs in each sample/ number of samples
- Sample Heterogeneity = [(total number of T-RFs in a dataset) / (average T-RF richness in the environments)] - 1. Sample or environment heterogeneity is also known as beta diversity, as defined by Whitaker (1972). McCune and Grace (2002) state as a rule of thumb for ecological datasets, beta diversities less than 1 are rather low and greater than 5 are very high. Culman et al. (2008) found that T-RFLP datasets with high beta diversities

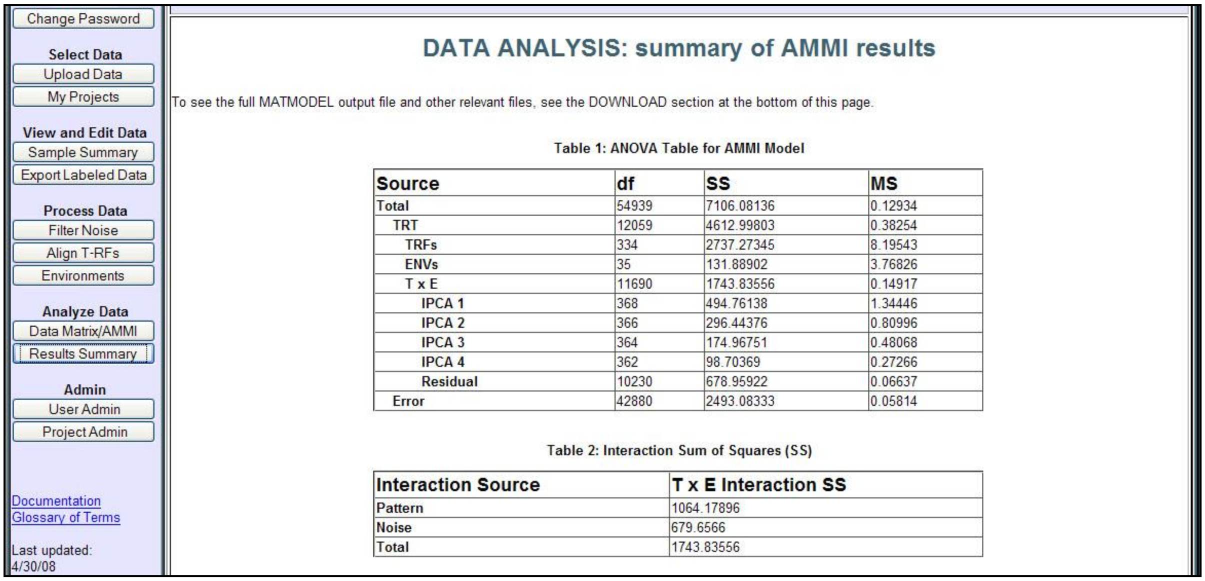
(≥ 2) should likely employ NMS analyses, while datasets with low beta diversities (< 1) should use theoretical criteria to selection an appropriate ordination analysis.

- Percent of Empty Cells in Data Matrix = (total number of cells in data matrix - sum of every sample richness/ (total number of cells in data matrix)

14. AMMI Analysis

The Additive Main Effects and Multiplicative Interaction Model (AMMI), also known as doubly-centered PCA, has been demonstrated to be an empirically robust and theoretically advantageous ordination analysis for T-RFLP data (Culman et al., 2008). The model uses analysis of variance (ANOVA) to first partition the variation into main effects and interactions, and then applies PCA to the interactions to create interaction principal components axes (IPCAs). *T-REX* interfaces with MATMODEL 3.0 (Gauch, 2007) to run the AMMI analysis.

Selecting 'Run AMMI Analysis' after the data matrix has been constructed will take the user to another page where four output tables summarize the ANOVA results and a number of output files are available to download (Figures 5 and 6). Explanations of the tables and their calculations are presented below.



DATA ANALYSIS: summary of AMMI results

To see the full MATMODEL output file and other relevant files, see the DOWNLOAD section at the bottom of this page.

Table 1: ANOVA Table for AMMI Model

Source	df	SS	MS
Total	54939	7106.08136	0.12934
TRT	12059	4612.99803	0.38254
TRFs	334	2737.27345	8.19543
ENVs	35	131.88902	3.76826
T x E	11690	1743.83556	0.14917
IPCA 1	368	494.76138	1.34446
IPCA 2	366	296.44376	0.80996
IPCA 3	364	174.96751	0.48068
IPCA 4	362	98.70369	0.27266
Residual	10230	678.95922	0.06637
Error	42880	2493.08333	0.05814

Table 2: Interaction Sum of Squares (SS)

Interaction Source	T x E Interaction SS
Pattern	1064.17896
Noise	679.6566
Total	1743.83556

Figure 5. Screenshot of tables 1 and 2 produced with output from the AMMI analysis in *T-REX*

Table 1 reports the results of the analysis of variance (ANOVA) on the two-way data matrix, where df = degrees of freedom, SS = sum of squares, and MS = mean square. All subsequent tables are based on calculations from this ANOVA table.

Table 2 reports the interaction sums of squares (SS). If the data are replicated, the interaction total will be decomposed into interaction pattern (or signal) and interaction noise. If data are not replicated an N/A will be displayed in the cells. With replicated data, Table 2 gives an estimate of the degree to which the interactions (differential responses of T-RFs to the samples) are meaningful signal vs. idiosyncratic noise. Higher percentages of interaction pattern reflect greater similarity among replicates (samples in the same environment). It is possible to get a negative value for the interaction pattern. This occurs when the predicted interaction noise exceeds the interaction total, and indicates that the interaction term is probably mostly noise.

Table 2 Calculations: The interaction noise SS was estimated by multiplying the interaction degrees of freedom (df) by the mean squared error (MSE). The interaction signal SS was estimated by subtracting the interaction noise SS from the interaction (total) SS.

Table 2. Calculations for Interaction SS

Interaction Source	**T x E Interaction SS**	**Percent of Total Interaction**
Interaction Pattern	(T x E) SS - Interaction Noise SS	(Interaction Pattern SS/ Interaction Total SS) x 100
Interaction Noise	df (T x E) x MS Error	(Interaction Noise SS/ Interaction Total SS) x 100
Interaction Total	(T x E) SS	Total = 100%

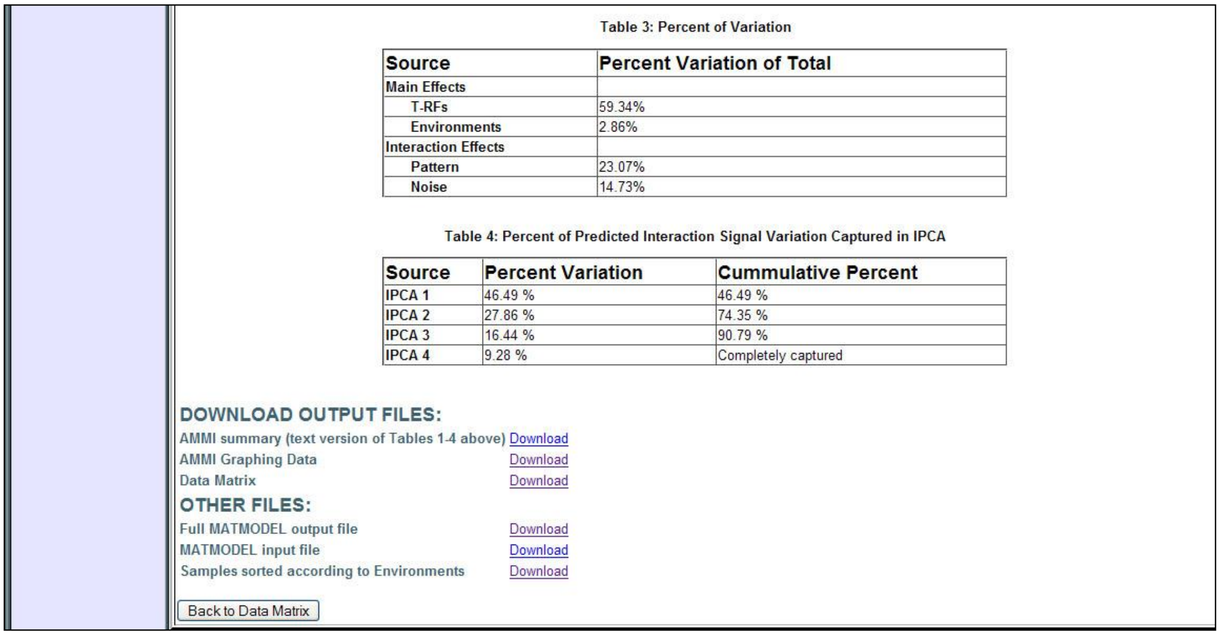


Table 3: Percent of Variation

Source	Percent Variation of Total
Main Effects	
T-RFs	59.34%
Environments	2.86%
Interaction Effects	
Pattern	23.07%
Noise	14.73%

Table 4: Percent of Predicted Interaction Signal Variation Captured in IPCA

Source	Percent Variation	Cummulative Percent
IPCA 1	46.49 %	46.49 %
IPCA 2	27.86 %	74.35 %
IPCA 3	16.44 %	90.79 %
IPCA 4	9.28 %	Completely captured

DOWNLOAD OUTPUT FILES:
AMMI summary (text version of Tables 1-4 above) Download
AMMI Graphing Data Download
Data Matrix Download
OTHER FILES:
Full MATMODEL output file Download
MATMODEL input file Download
Samples sorted according to Environments Download
Back to Data Matrix

Figure 6. Screenshot of the tables 3 and 4 produced with output from the AMMI analysis in *T-REX*

Table 3 reports the percent of variation from each source in the model. Main effects variation is composed of variation from T-RFs and Environments. Likewise, with replicated data interaction effects are composed of interaction pattern and interaction noise. Culman et al. (2008) found that T-RFLP datasets typically exhibit large variation from T-RFs and small amounts of variation from Environments. Variation due to interaction effects can have a considerable range, and reflect how similar or dissimilar the microbial communities are. They concluded that ANOVA could be used as a tool to objectively measure microbial community dissimilarity across multiple datasets. Note that relativized peak height and relativized peak area will yield variation from Environments equal to 0.

Table 3 Calculations: The percent of variation from each source in the ANOVA (T-RF, E, and TxE) was calculated by dividing that source's sum of squares (SS) by the treatment SS and multiplying by 100.

Table 3. Calculations for Percent of Variation

Source	Percent Variation of Total
Main Effects	
T-RFs	(T-RFs SS/ TRT SS) x 100
Environments	(Environments SS/ TRT SS) x 100
Interaction Effects	(Interaction Total SS/ TRT SS) x 100 [with non-replicated data]
Pattern	(Interaction Pattern SS/ TRT SS) x 100 [with replicated data]
Noise	(Interaction Noise SS/ TRT SS) x 100 [with replicated data]

Table 4 reports the percent of predicted interaction signal variation captured in the first four IPCAs. This is useful when research objectives determine interaction pattern (signal) is of primary interest. Note that this should not be confused as the percent of ***total*** variation captured by the IPCAs. This calculation is possible, if you substitute the Interaction Pattern SS with TRT SS in Table 4. See Culman et al. (2008) for more details regarding these concepts.

These calculations can determine how much of this variation is captured/ recovered in the graphed axes.

Table 4. Calculations for Percent of Predicted Interaction Signal

Source	Percent Variation of Predicted Interaction Signal	Cumulative Percent of Predicted Interaction Signal
IPCA 1	(IPCA 1 SS/ Interaction Pattern SS) x 100	Percent variation IPCA1
IPCA 2	(IPCA 1 SS/ Interaction Pattern SS) x 100	Percent variation IPCA1 + IPCA2
IPCA 3	(IPCA 1 SS/ Interaction Pattern SS) x 100	Percent variation IPCA1 + IPCA2 + IPCA3
IPCA 4	(IPCA 1 SS/ Interaction Pattern SS) x 100	Percent variation IPCA1 + IPCA2 + IPCA3 + IPCA4

Note: If Cumulative Percentage exceeds 100%, 'complete captured' will be displayed, indicating that all of the predicted signal has been selectively recovered. At this point, percentages greater than 100% will also contain predicted interaction noise and subsequent axes will not be displayed.

15. Download Output Files

T-REX provides several files available for download at the bottom of the **AMMI Analysis** page. Table 5 outlines these files, with more information below.

Table 5. *T-REX* files available for download.

File Type	**File Extension**	**Recommended Application**
Essential Files:		
AMMI summary	.mm_sum	word processor
AMMI Graphing Data	.mm_grph	spreadsheet
Data Matrix	.matrx	spreadsheet
Transposed Data Matrix	.tmatrx	spreadsheet
Other Files:		
MATMODEL output file	.mm_out	word processor
MATMODEL input file	.mm_in	word processor
Environments Assigned to Samples	.env	spreadsheet
Labeled Data (list format)	.label	spreadsheet
All Files:		
Zipped folder containing all files	.zip	compatible .zip extractor

1) AMMI summary file (.mm_sum) is a text file containing Tables 1 – 4.

2) AMMI Graphing Data (.mm_grph) is the file that contains the graphing output generated from the AMMI4 (AMMI analysis with 4 axes) analysis by MATMODEL. The first line reports the total number of T-RFs, Environments (ENVs), replicates (REPs), if any, and the grand mean. Since the AMMI analysis is an integrated, dual analysis, scores for both Environments and T-RFs are reported. Environment scores are reported first and include the following columns:

- Column/row—the sequential number of the Environment/ T-RF
- Environment/T-RF—the name of the Environment or T-RF representing the corresponding row of data
- Mean—the arithmetic average of the column/row (Env/T-RF). Because of problems associated with PCR bias, this value is often not of interest in the context of T-RFLP analysis. For example, relativizing peak height and area will create approximately equal means across all Environments. However, the mean of T-RFs (in particular) may be of interest, depending on the research goals and methods.

- IPCA 1—the first interaction principal component; typically, IPCA1 vs. IPCA 2 is graphed in an AMMI analysis of T-RFLP data
- IPCA 2—the second interaction principal component
- IPCA 3—the third interaction principal component
- IPCA 4—the fourth interaction principal component
- Attributes of sample/s from label file that correspond to the environment. These columns will be included only if the environments are determined by a subset of labels.

3) Data Matrix (.matrx) is a tab-delimited data matrix with T-RFs as rows and Environments as columns. The first row contains the sample ID and the second row contains the corresponding File Name. The data matrix is followed by all the selected parameters used to create the data matrix and then by the basic data matrix properties (see section 13.1).

4) Transposed Data Matrix (.tmatrx) is the transposed data matrix.
5) MATMODEL output file contains the full output from the MATMODEL program. See the MATMODEL 3.0 (Gauch, 2007) for more details.

6) MATMODEL input file is the data input file constructed by *T-REX* to run the AMMI analysis in MATMODEL. This file is generated in case the user wants to further manipulate fields and run MATMODEL independently of *T-REX*. See the (Gauch, 2007) for more details.

7) Environments Assigned to Samples is a tab-delimited file that lists which samples belong to which environments. This file can be referenced when interpreting the AMMI Graphing Data file and other files.

8) Label File is a tab-delimited file of the labeled data. If the uploaded data in a project have been processed, it will reflect those manipulated data. This file will only be available if the "Export Labeled Data" procedure has been run.

9) The zipped folder contains all seven above files in a compressed format for ease of storage, archiving, and/or convenience. Extractors for zipped files (.zip) are natively built into nearly all current operating systems.

16. Results Summary

The **Results Summary** page reports the results of relevant basic data matrix properties and summarizes the results of the AMMI analysis in one place. The 'T-RF Abundance table' reports the number of samples (samples present) and percentage of samples (% of samples present) that each T-RF occurs. All generated output files are also available for download at this page.

17. Definitions and Relevant Calculations

Active peak: a peak that is stored in a project and will be included in processing and analysis of the dataset.

Average T-RF richness is the average number of T-RFs in a the dataset; average T-RF richness = sum of total number of T-RFs in each sample/ number of samples

Environment: a set of samples assumed (for the purpose of statistical data analysis) by the user to represent the same or similar experimental conditions and therefore being treated as replicates of one another.

Env (in **Samples Summary** page) is an integer number denoting the **environment** (group of replicates) which a given sample belongs to. If a set of samples have the same environment value, they will be treated as replicates.

ID or **Sample ID** is a unique identifier of a sample, generated when raw data are uploaded to the program, creating a new project.
Inactive peak: a peak that is stored in a project, but will not but included in processing and analysis of the dataset; such peaks result from a manual change or from automatic noise elimination operations.

Interaction Principal Components—IPCs or IPCAs; constructed axes capturing the greatest interaction signal in the AMMI analysis; synonymous to principal components in PCA

Label file. This contains a set of labels/attributes that describe each sample and often correspond to factors in the experimental design.

Missing data sample: a sample with all peaks labeled inactive; a sample can designated as missing data as a result of peak filtering or can be marked as such manually.

Original Peaks (in **Samples Summary** page) refer to the total number of peaks originally present in the uploaded raw data file.

Peak: An entry point from an electropherogram, equivalent to a single line in a raw data file. It is characterized by the name of the file it is recorded in, the position (length of a fragment in nucleotide base pairs), height, and area.

Percent of Empty Cells in Data Matrix = (total number of cells in data matrix - sum of every sample richness) ÷ (total number of cells in data matrix)

Raw data file. This is the file that is exported in Genemapper™, Peak Scanner™, or similar size-calling software that contains the peak information for a set of samples.

Replicate: an individual sample belonging to an environment.

Removed Peaks (in **Sample Summary** page): the total number of peaks from the original raw data file which have been flagged as inactive.

Sample: an electrophoretic run; a set of peaks sharing the same file name in the raw data file.

Sample Heterogeneity = [(total number of T-RFs in a dataset) / (average T-RF richness in the environments)] - 1. Sample or environment heterogeneity is also known as beta diversity, as defined by Whittaker (1972).

T-RF: a group of active peaks from different samples with similar positions (base pair sizes)

References

Abdo, Z., U.M.E. Schuette, S.J. Bent, C.J. Williams, L.J. Forney, and P. Joyce. 2006. Statistical methods for characterizing diversity of microbial communities by analysis of terminal restriction fragment length polymorphisms of 16S rRNA genes. Environmental Microbiology 8:929-938.

Blackwood, C.B., T. Marsh, S.-H. Kim, and E.A. Paul. 2003. Terminal restriction fragment length polymorphism data analysis for quantitative comparison of microbial communities. Applied and Environmental Microbiology 69:926-932.

Culman, S.W., H.G. Gauch, C.B. Blackwood, and J.E. Thies. 2008. Analysis of T-RFLP data using Analysis of Variance and Ordination Methods: A Comparative Study. Journal of Microbiological Methods 75:55-63.

Dunbar, J., L.O. Ticknor, and C.R. Kuske. 2001. Phylogenetic Specificity and Reproducibility and New Method for Analysis of Terminal Restriction Fragment Profiles of 16S rRNA Genes from Bacterial Communities. Applied and Environmental Microbiology 67:190-197.

Gauch, H.G. 2007. MATMODEL Version 3.0: Open source software for AMMI and related analyses. Release 3.0. Crop and Soil Sciences, Cornell University, Ithaca, NY.

Kaplan, C.W., and C.L. Kitts. 2003. Variation between observed and true Terminal Restriction Fragment length is dependent on true TRF length and purine content. Journal of Microbiological Methods 54:121-125.

Marsh, T.L. 2005. Culture-independent microbial community analysis with terminal restriction fragment length polymorphism. Methods in Enzymology 397:308-329.

McCune, B., and J.B. Grace. 2002. Analysis of Ecological Communities MjM Software Design, Gleneden Beach, OR.

Smith, C.J., B.S. Danilowicz, A.K. Clear, F.J. Costello, B. Wilson, and W.G. Meijer. 2005. T-Align, a web-based tool for comparison of multiple terminal restriction fragment length polymorphism profiles. FEMS Microbiology Ecology 54:375-380.

Whittaker, R.H. 1972. Evolution and measurement of species diversity. Taxon 21:213-251.