

Software

Open Access

## XplorSeq: A software environment for integrated management and phylogenetic analysis of metagenomic sequence data

Daniel N Frank

Address: Department of Molecular, Cellular, and Developmental Biology, Mucosal and Vaccine Research Program Colorado, University of Colorado, Boulder, CO, USA 80309, USA

Email: Daniel N Frank - [daniel.frank@colorado.edu](mailto:daniel.frank@colorado.edu)

Published: 7 October 2008

Received: 10 July 2008

BMC Bioinformatics 2008, 9:420 doi:10.1186/1471-2105-9-420

Accepted: 7 October 2008

This article is available from: <http://www.biomedcentral.com/1471-2105/9/420>

© 2008 Frank; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

### Abstract

**Background:** Advances in automated DNA sequencing technology have accelerated the generation of metagenomic DNA sequences, especially environmental ribosomal RNA gene (rDNA) sequences. As the scale of rDNA-based studies of microbial ecology has expanded, need has arisen for software that is capable of managing, annotating, and analyzing the plethora of diverse data accumulated in these projects.

**Results:** XplorSeq is a software package that facilitates the compilation, management and phylogenetic analysis of DNA sequences. XplorSeq was developed for, but is not limited to, high-throughput analysis of environmental rRNA gene sequences. XplorSeq integrates and extends several commonly used UNIX-based analysis tools by use of a Macintosh OS-X-based graphical user interface (GUI). Through this GUI, users may perform basic sequence import and assembly steps (base-calling, vector/primer trimming, contig assembly), perform BLAST (Basic Local Alignment and Search Tool; [1-3]) searches of NCBI and local databases, create multiple sequence alignments, build phylogenetic trees, assemble Operational Taxonomic Units, estimate biodiversity indices, and summarize data in a variety of formats. Furthermore, sequences may be annotated with user-specified meta-data, which then can be used to sort data and organize analyses and reports. A document-based architecture permits parallel analysis of sequence data from multiple clones or amplicons, with sequences and other data stored in a single file.

**Conclusion:** XplorSeq should benefit researchers who are engaged in analyses of environmental sequence data, especially those with little experience using bioinformatics software. Although XplorSeq was developed for management of rDNA sequence data, it can be applied to most any sequencing project. The application is available free of charge for non-commercial use at <http://vent.colorado.edu/phyloware>.

### Background

The recent explosions in culture-independent studies of environmental DNA sequences ("metagenomics") and automated DNA sequencing capabilities have prompted the creation of numerous software applications designed to aid the analysis of an avalanche of sequence data. How-

ever, many of the commonly used, freely available applications require some facility with the UNIX/Linux operating system and/or specialized scripting languages to either manipulate files in batch or pipe data between applications. As automated DNA sequencing and sequence analysis has become commonplace in laborato-

ries that do not specialize in bioinformatics, need has arisen for the development of powerful, yet simple-to-use, software.

XplorSeq, written for the Macintosh OS X operating system, provides a graphical user interface (GUI) that integrates the use of several popular UNIX-based DNA sequence analysis applications. A number of additional features have been incorporated in order to track, annotate, and analyze sequence information in a manner conducive to high-throughput metagenomics. Implementation of a common GUI (Fig. 1) eliminates the need for the user to possess the kind of knowledge generally restricted to bioinformaticists and computer professionals: UNIX/Linux command shells (tsh, csh, or bash) and associated scripting languages. By presenting a unified GUI, XplorSeq, simplifies sequence analysis projects by allowing the user to focus on science, rather than the details of disparate software.

XplorSeq was developed for rapid compilation and analysis of rDNA clone libraries, but should be applicable to any sequencing project (computer hardware may, however, limit the scale of projects). Although several commercial and non-commercial software packages implement some of the same basic functionalities as XplorSeq, the development of XplorSeq was motivated by the absence of GUI-based software designed specifically for high-throughput, batch analysis of rDNA sequences, such as arise from culture-independent metagenomic studies. Specifically, the extant software could not accommodate the phylogenetic orientation of analyses and sequence annotations that are most useful for metagenomics. In contrast, XplorSeq implements several domain-specific software tools (e.g. for state of the art phylogenetic tree inference, OTU clustering, biodiversity estimates) that are not available in general-purpose DNA analysis packages. Many published studies, from a variety of laboratories engaged in metagenomics, have used XplorSeq, and thereby established its stability, ease-of-use, and capabilities [4-29]. The software is freely available for non-commercial use at <http://vent.colorado.edu/phyloware>.

## Implementation

XplorSeq is written in Objective-C using the Cocoa application framework (Apple Inc.). Releases are compiled for the OS X operating system (current versions require OS 10.4.x or 10.5.x) as universal binaries, which run natively on Macintosh computers with Intel or PowerPC microprocessors. Similar to Cocoa, the architecture of XplorSeq is based on the Model-View-Controller (MVC) design pattern. XplorSeq is multi-threaded and can adjust its operation to accommodate multiple shared-memory microprocessors.

The rationale for implementing XplorSeq as a standalone Macintosh application involved 1) desire for a highly responsive, feature-rich graphical output, thus precluding web-based applications; 2) recognition that the BSD-Unix operating system underpinning OS X would allow leveraging of existing open-source software; 3) observation that many computer novices (an intended audience for this software) were more comfortable with OS X than other operating systems; and 4) the maturity, stability, and support inherent in the Cocoa application framework.

Third-party software packages and plugin executables (sortx and biodiv) were written in C and C++. When possible, compiled executables are incorporated directly into the XplorSeq application bundle (essentially, a hidden directory structure) so that users can install and operate XplorSeq without the need for local compilation or extensive configuration. Full implementation of XplorSeq requires separate installation of phred and phrap (obtained at <http://www.phrap.org>).

## Results and discussion

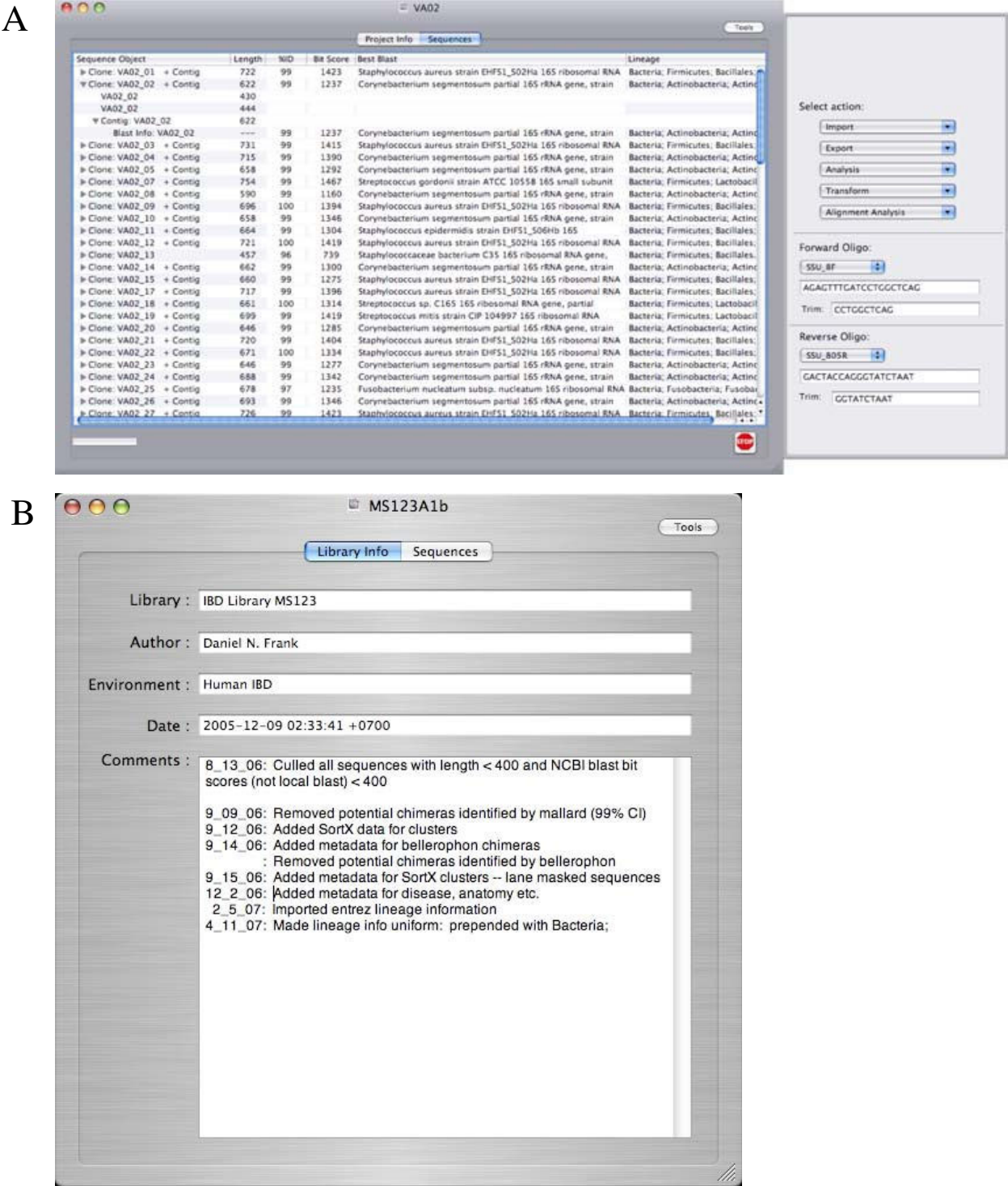
The following sections outline the data structures and analytic tools that form the basis of the XplorSeq workflow.

### Data organization and GUI architecture

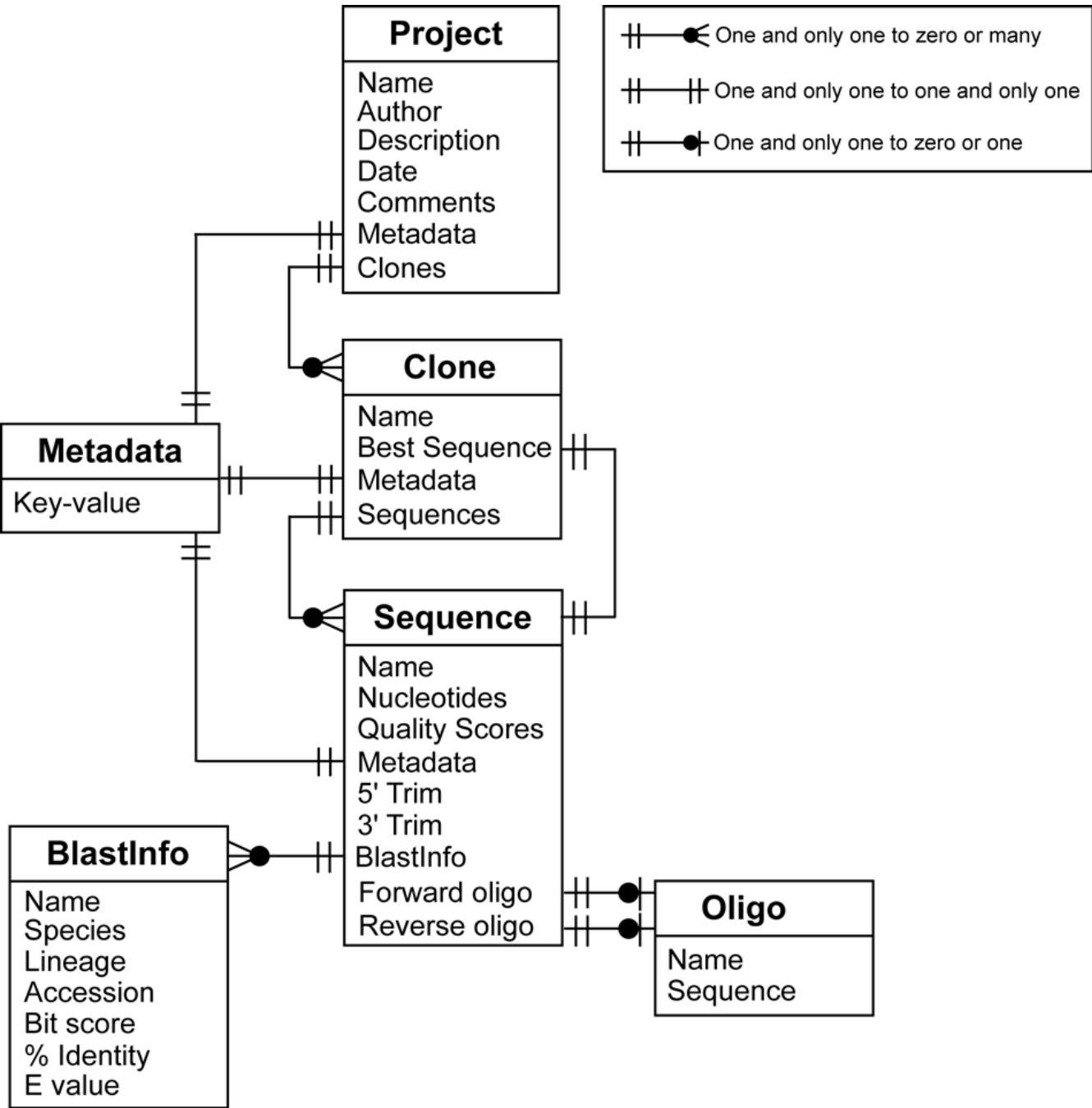
XplorSeq uses a document-based approach for project data management in which multiple sequences and their associated data are stored and accessed in a single file. As a project evolves, sequences may be added, deleted, amended and analyzed as needed. XplorSeq does not enforce a highly constrained analytic schema and thereby grants the user more autonomy in designing and implementing an analysis plan than typically is possible in a hard-wired software pipeline.

Data are organized as a hierarchy of data objects (Fig 2), which represent the model layer in XplorSeq's Model-View-Controller design pattern architecture. Each data object class (e.g. Sequence, Project) is associated with a window (the view layer) that is used to inspect data associated with particular instances of the class.

The top layer data object is the "Project", which stores all other data and is synonymous with the document as a whole. Hence, the main XplorSeq window (Fig 1) is the Project Inspector window. Projects organize and manage lists of "Clones", which represent individual cloned genes or PCR amplicons. Clones, in turn, manage groups of "Sequences" which map to unique DNA sequences. Sequences can be imported directly (e.g. as polished GenBank sequences), read from DNA sequencer traces, or assembled from other sequence objects ("contigs"). For



**Figure 1**  
**Main XplorSeq window.** Screen shots of XplorSeq main window. (A) Listing of imported sequences, contigs, and blast data. Import, export, data analysis, and data transformation options are presented in menus within the tool drawer adjacent to the main window. (B) Project-associated data fields. Comments box provides space for recording details of analysis.



**Figure 2**  
**Organization of basic XplorSeq Data structures.**

each sequence analyzed by BLAST, XplorSeq creates a "BlastInfo" object that summarizes pertinent blast output data: identity and phylogenetic lineage of the sequence's closest BLAST hit, BLAST statistics, etc. Each Clone ranks its constituent Sequence objects based on BLAST bit-score and the "Best Sequence" (i.e. that with the highest bit-score) serves as a proxy for the entire Clone. "Oligo"

objects encapsulate data that describe oligonucleotide sequences used in construction of clone libraries.

One of XplorSeq's more important design goals is to provide users with an extremely flexible and simple means of annotating data objects with meta-data consisting of user-specified information. XplorSeq implements meta-data

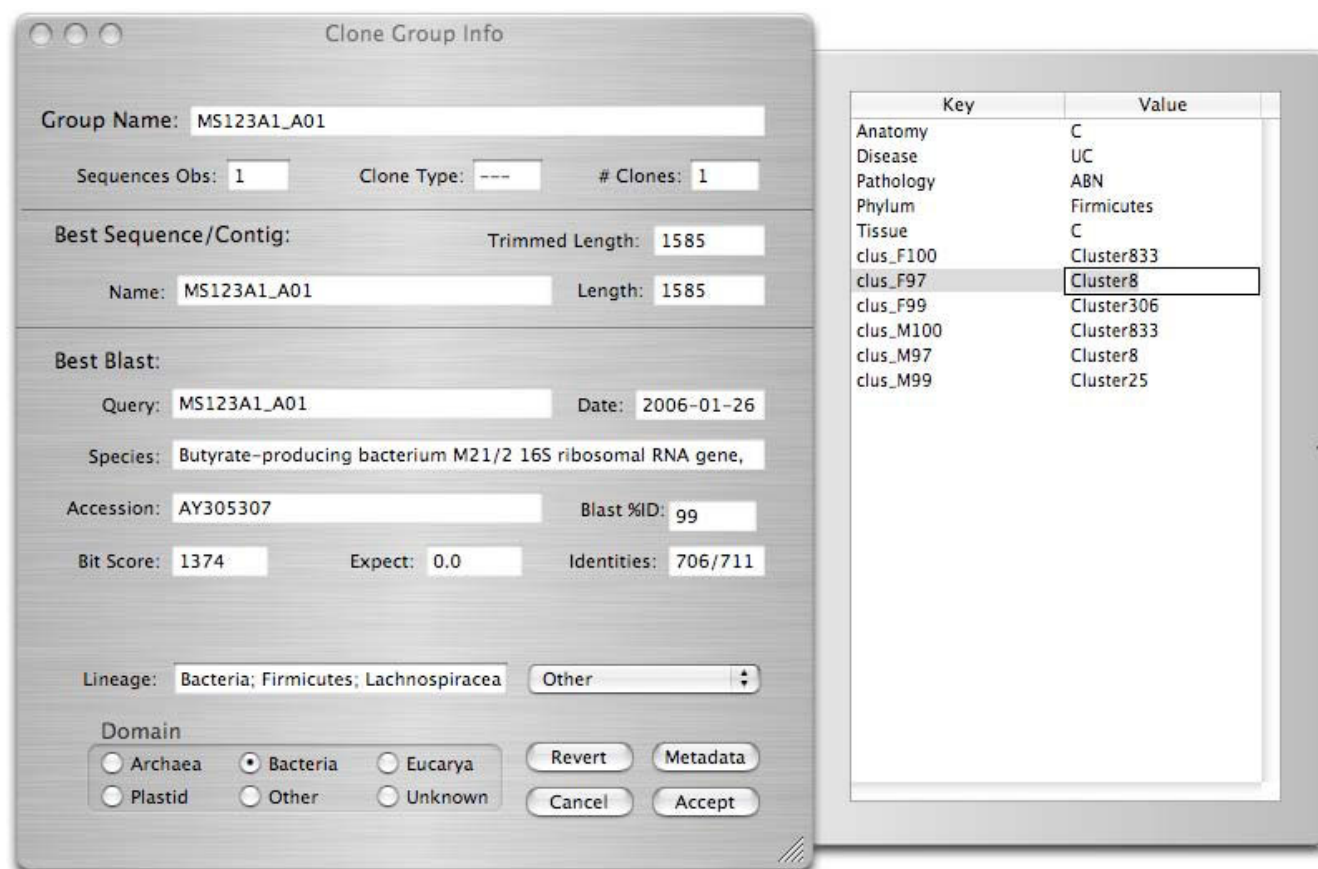
through two strategies: 1) hard-coded data objects (i.e. "BlastInfo" and "Oligo" objects) and 2) a customizable "Metadata" object. Metadata objects are implemented as key-value dictionaries (e.g. hashes, maps) that can be linked to Project, Clone, and Sequence data objects. Users specify keys and values for particular objects through import of tab-delimited spreadsheets in which the first column designates the names of objects (i.e. sequence/clone names) and subsequent columns specify values (keyed by column headings) to attach to the objects. Alternatively, object inspector windows provide GUI-based features in which to display and edit meta-data (c.f., Figs. 3, 4, and 5).

### Data display and control of data processing

A project's Clone, Sequence, and BlastInfo objects are displayed in the Project Inspector window (Fig. 1A), which functions as the main XplorSeq window. Data are

arranged hierarchically to reflect nesting of data structures. For each Clone, a summary of its best blast hit, which includes the taxa name, percent sequence identity and bit-score is displayed in the main XplorSeq window. The phylogenetic lineage of the top blast hit can be imported (through either an entrez idfetch query or import of tab-delimited data) to provide information about the taxonomic placement of a clone.

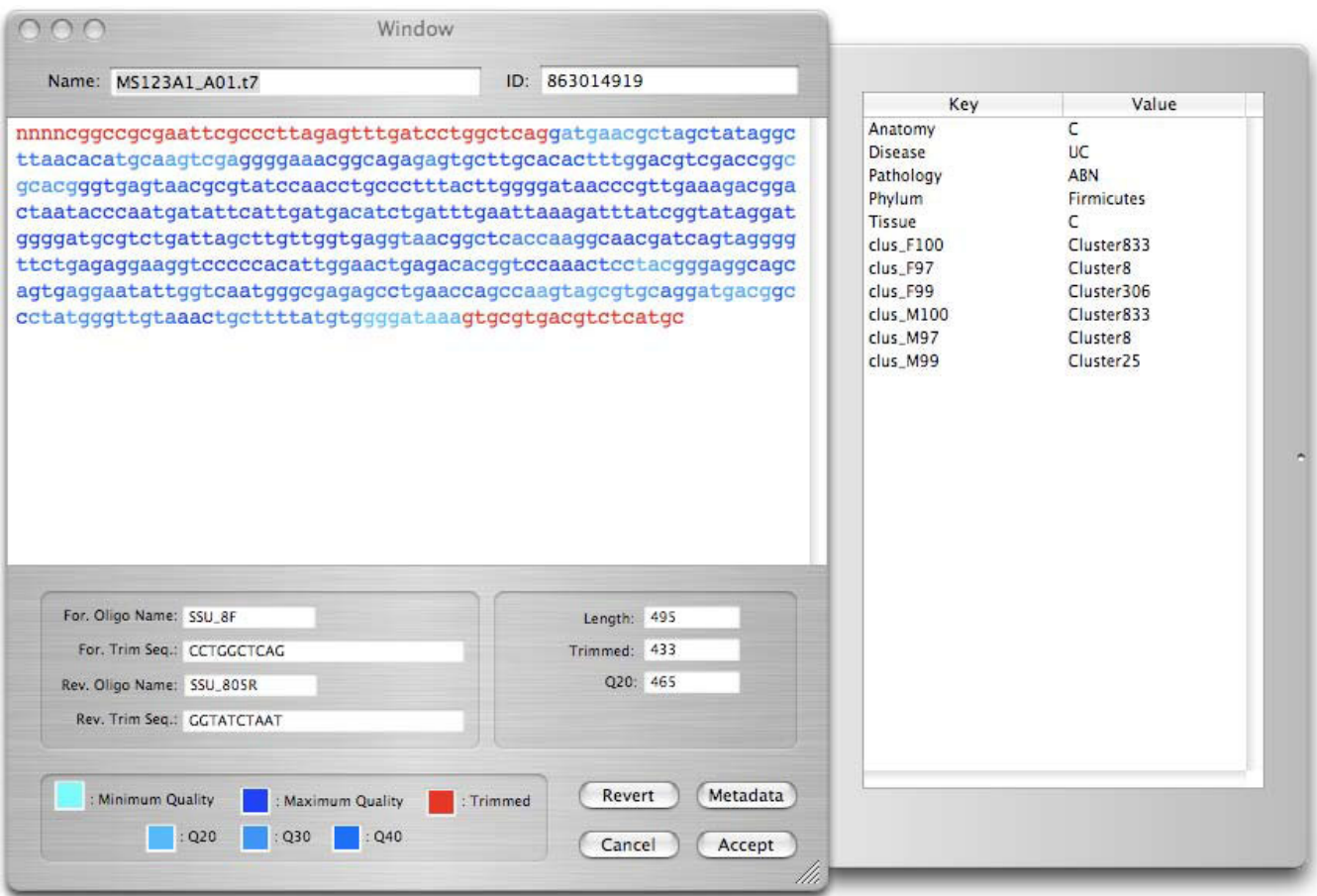
The user controls all steps of data processing by selecting objects to be acted upon and then choosing a function from menu items presented in the tool panel that extends from the main XplorSeq window (Fig 1A). Methods to import, export, and analyze data are accessible through these menus. Below the menus lie controls through which oligonucleotides used to generate PCR libraries can be designated, if relevant to the project; entries in the oligo menu can be modified through a preferences dialog.



**Figure 3**

**Clone inspector window.** Summarizes information associated with a selected clone, a group of sequences (both individual reads and contigs) from the same amplified or cloned gene. The main window summarizes the top BLAST hit for the clone (i.e. the sequence or contig BLAST hit with the highest bit-score). The phylogenetic lineage of the clone can be assigned at the bottom of the main window. The drawer to the right of the window presents the contents of the meta-data dictionary associated with this clone.





**Figure 4**  
**Sequence inspector window.** Displays information associated with a selected sequence. Nucleotides are color-coded to represent the quality scores of individual nucleotides (the legend at the bottom of the window shows the meaning of the colors). Primers used to amplify the gene are summarized on the lower left. Basic summary statistics (length, trimmed length, and number of nts > Q20) are presented in the lower right. The drawer to the right of the window presents the contents of the meta-data dictionary associated with this sequence.

Project specific meta-data can be recorded in several text fields presented in the Project window, under the "Project Info" tab. An editable text box is presented in which the user can enter comments, for instance details specific to a project (Fig 1B).

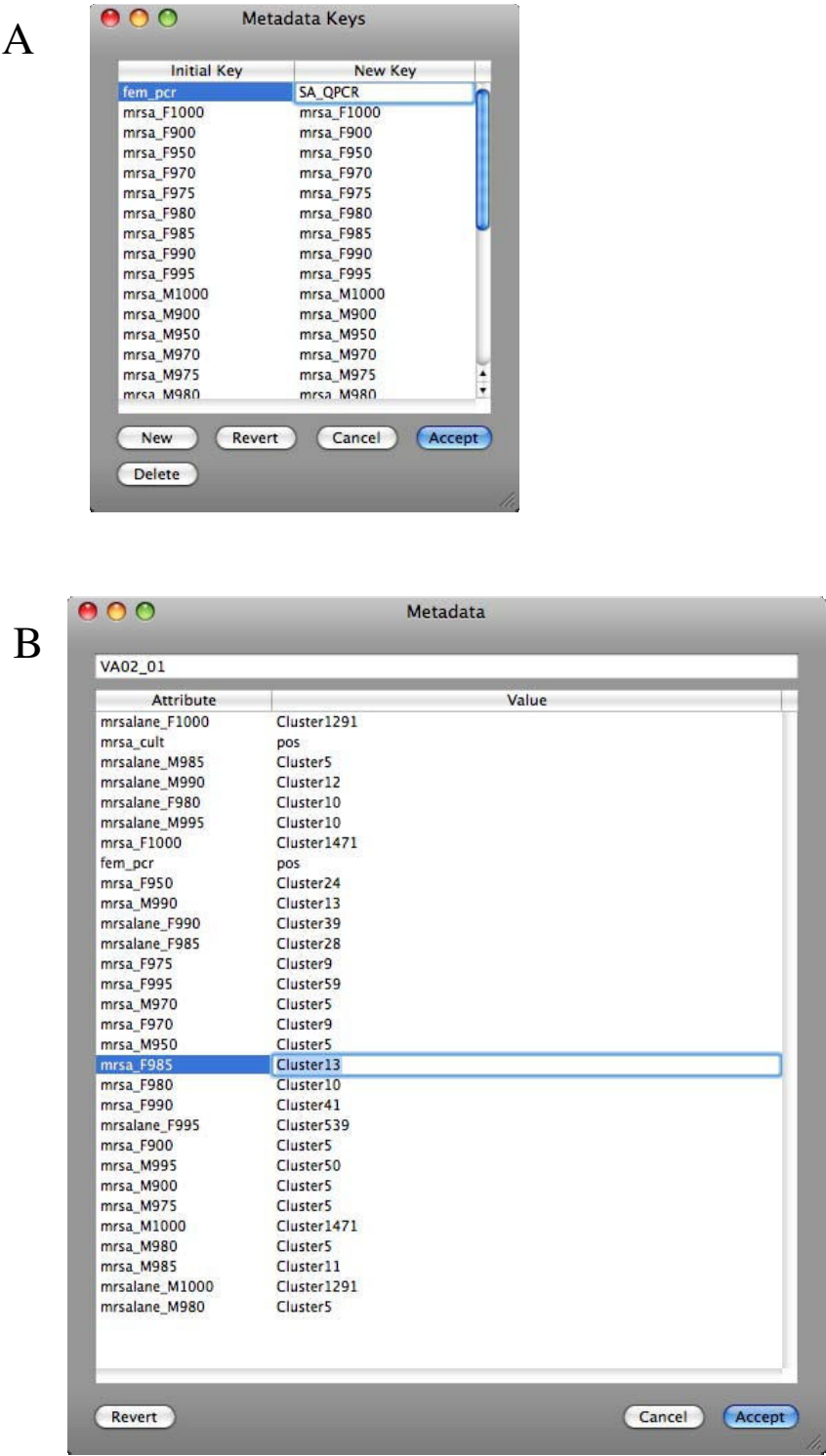
By double-clicking on an entry in the Project window, the user can display and edit more detailed information associated with that entry. For example, the Clone Inspector window (Fig. 3) summarizes the content of the Clone Object, including its top BLAST hit sequence and corresponding BlastInfo Object. The phylogenetic lineage and domain of the Clone Object can be set through the controls at the bottom of the window. A panel extending from the Clone Inspector window presents user-specified meta-data associated with the Clone Object.

DNA sequences, including contig sequences, are displayed in a Sequence Inspector window (Fig. 4). Individ-

ual nucleotides are color-coded to represent quality scores generated by the base-calling software (shades of blue) or trimmed sequences (red). Basic sequence information, such as primer sequences and trimmed sequence length, is displayed in a set of text fields at the bottom of the Sequence Inspector window. Similar to Clone Objects, sequence specific meta-data can be viewed through a panel that extends from the inspector window.

**Tools for data import and analysis**

Table 1 summarizes the key functions available through XplorSeq. A user can perform base-calling on chromatograms generated by automated DNA sequencers, (through tracetuner [30] or phred, [31,32]); trim away vector, primer, and poor quality sequences; assemble contiguous sequences (TIGR\_Assembler [33] or phrap, [31]); search for homologous sequences in sequence databases (blastcl3 or blastall; [1-3]); format BLAST databases (formatdb), perform multiple sequence alignments (clus-



**Figure 5**  
**Metadata editors.** A) Editing all keys for meta-data in project. B) Editing values associated with keys for a particular sequence object.

**Table 1: Summary of XplorSeq functionality**

Function	Description <sup>1</sup>
<b>Import</b>	
Chromatogram...	Import DNA chromatograms (.esd, .scf, .abi etc.): phred
PHD...	Import DNA sequences in Phd format
Contig...	Import DNA sequences and quality scores in FastA format
Blast...	Parse Blast records
FastA...	Import DNA sequences in FastA format
XplorSeq Library...	Import XplorSeq document
Phylogenetic Lineage...	Import phylogenetic lineage information from entrez
Metadata...	Import metadata in key-value format
<b>Export</b>	
Sequences...	Export DNA sequences in variety of formats
FastA + Qual...	Export DNA sequences and quality scores
Blast Info...	Export summary of Blast records
Cluster Table...	Enumerate OTUs belonging to groups of sequences
OTU Diversity...	Calculate OTU richness for set of sequences
Quality Scores...	Export summary of quality scores
Blast Accession #'s...	Export accession numbers of top Blast hits
Sequin Script...	Export data in format for Genbank submission (sequin)
Blast Database...	Create a Blast database (formatdb)
XML File...	Export data in XML format
Metadata...	Summarize and export metadata
Placeholder Tree...	List selected sequences in Newick format
<b>Analyze</b>	
Basecall->Blast...	Pipe data from chromatogram through Blast analysis
Contig->Blast...	Pipe data from contig assembly to Blast analysis
Basecall...	Perform base calling (phred or ttuner)
Contig...	Perform contig assembly (phrap or TIGR_Assembler)
Blast NCBI...	Blast query of Genbank
Blast Local...	Blast query of local blast database
Get Entrez Lineage Info.	Download entrez phylogenetic lineage information (idfetech)
Align...	Perform multiple sequence alignment (clustal)
Biodiversity (biodiv)...	Calculates biodiversity indices with random resampling (biodiv)
XplorSeq Doc Difference...	Generate differences between two XplorSeq documents
<b>Transform</b>	
Edit Sequence Names...	Alter names of sequences
Edit Lineage Names...	Edit phylogenetic lineage information
Edit Metadata...	Edit metadata associated with sequence
Edit Metadata Keys...	Edit all metadata keys in document
Group...	Group sequences and contigs
UnGroup...	Ungroup sequences and contigs
Clean...	Delete blast information, contigs
Sort...	Sort records in document
Set Oligos...	Associate primer sequences with sequence objects
Trim...	Trim sequences based on quality score and primer
UnTrim.	Remove trimming information
Rev.-Complement	Reverse complement sequence
DNA -> RNA	Convert DNA sequence to RNA sequence
RNA -> DNA	Convert RNA sequence to DNA sequence
UPPER CASE	Convert sequence to upper case
lower case	Convert sequence to lower case
<b>Alignment Analysis</b>	
OTU clustering	Cluster Operational Taxonomic Units (sortx)
Clearcut NJ Tree...	Fast neighbor joining trees (clearcut)
Phylip distance matrix...	Calculate distance matrix (dnadist)
Phylip NJ Tree...	Calculate Neighbor joining or UPGMA trees (neighbor)



**Table 1: Summary of XplorSeq functionality** (Continued)

Phylip seqboot...	Generate bootstrap replicates of alignment (seqboot)
Phylip consense...	Generate consensus of multiple trees (consense)
RAXML...	Generate Maximum Likelihood tree (raxmlHPC)

<sup>1</sup>Command line executables are listed in parentheses.

talW; [34-36]); define Operational Taxonomic Units (sortx; D.N. Frank, unpublished); calculate distance matrices (phylip dnadist [37]); construct phylogenetic trees (clearcut, [38]; phylip neighbor, [37]; RAXML, [39]); and calculate a variety of biodiversity indices (biodiv; D.N. Frank, unpublished). By integrating software tools into a single application, XplorSeq facilitates seamless workflow through the sequence analysis process. Many of the key analytic steps are piped together, thus automating data processing. All data analysis operations may be executed in batch so that multiple sequences can be processed with minimal user intervention.

XplorSeq facilitates batch BLAST [1-3] analyses of DNA sequences through both networked and local searches of nucleotide databases. Local BLAST searches require properly formatted sequence databases, which may be downloaded from NCBI <http://ncbi.nlm.nih.gov> or created by use of the formatdb executable (through XplorSeq or the command line). XplorSeq dispatches sequences to the appropriate client software (blastcl3 or blastall; [1-3]) and then parses the resulting output file into BlastInfo Objects.

XplorSeq provides graphical interfaces for several commonly used command-line phylogenetic tools are shown in Fig. 6. Phylip modules for distance matrix calculation, neighbor-joining/UPGMA trees, data-set bootstrapping, and consensus tree generation are accessible through these interfaces [37]. Similarly, fast neighbor-joining and maximum-likelihood phylogenetic tree inference are available (clearcut [38] and RAXML [39], respectively). In each case, following selection of the appropriate options, sequence data are exported in the required format, the analysis performed, and results imported (if appropriate) into XplorSeq. For tasks that may take prolonged time to execute (e.g. RAXML [39]), a file listing the command-line invocation of the program is saved along with all data required for the task; these files can be transferred to another computer for subsequent execution of the scripted command.

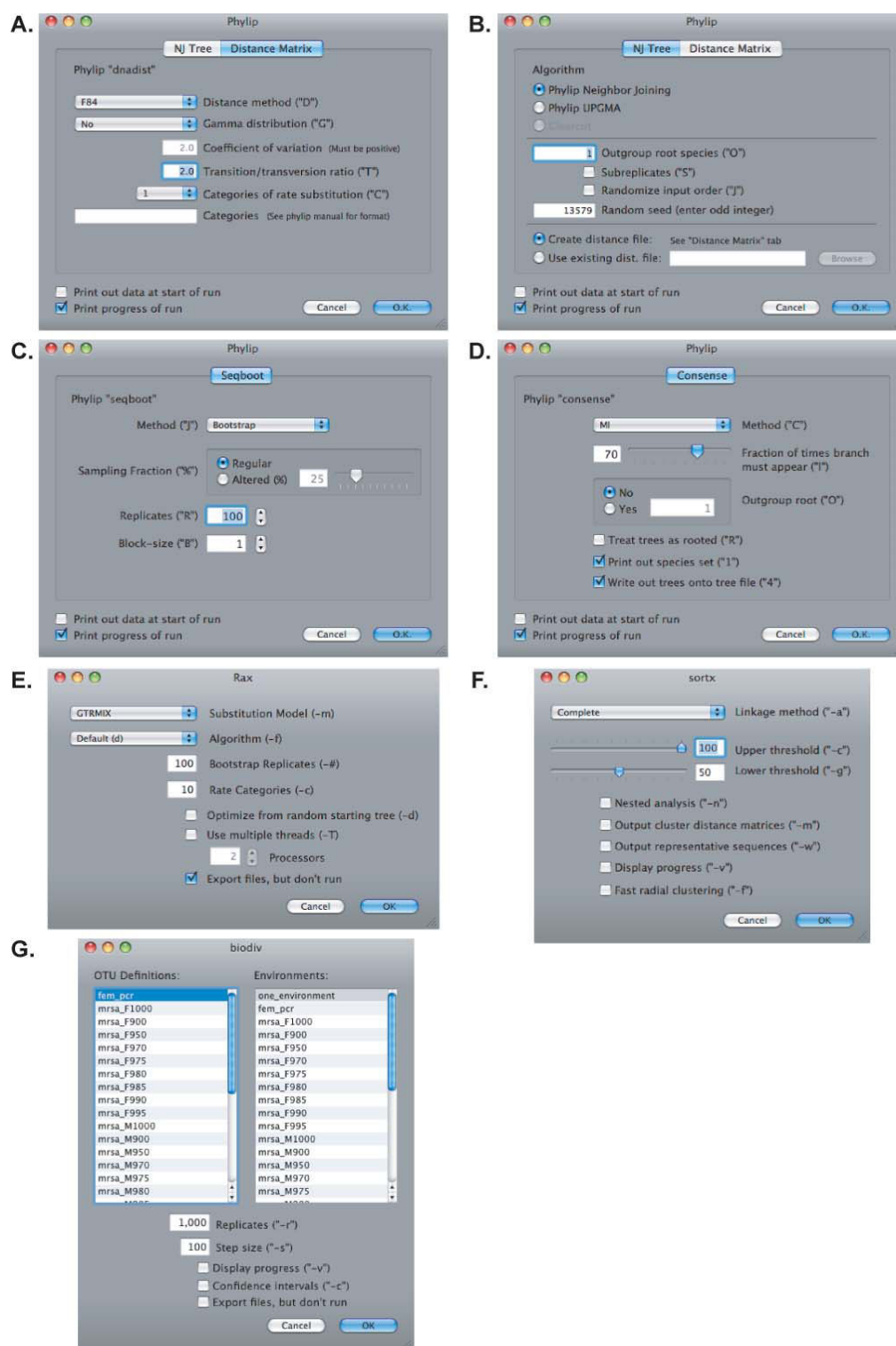
OTU clustering is implemented through the program sortx, which was written in tandem with XplorSeq (Fig. 6F). Sortx uses a fast radial clustering algorithm to bin aligned sequences based on uncorrected pairwise sequence distances (%ID). Clusters can be assembled based on furthest-, mean-, or nearest-neighbor rules. Following cluster formation, sortx selects a representative

sequence for each cluster, which maximizes both pairwise similarity to other cluster members and sequence length (simply choosing the sequence with minimum pairwise distance could select for short, but well-conserved sequences, which would not necessarily be representative of the cluster). Finally, the user can select a range of pairwise sequence distance thresholds by which to assemble OTUs in order to create multiple data sets at different phylogenetic depths.

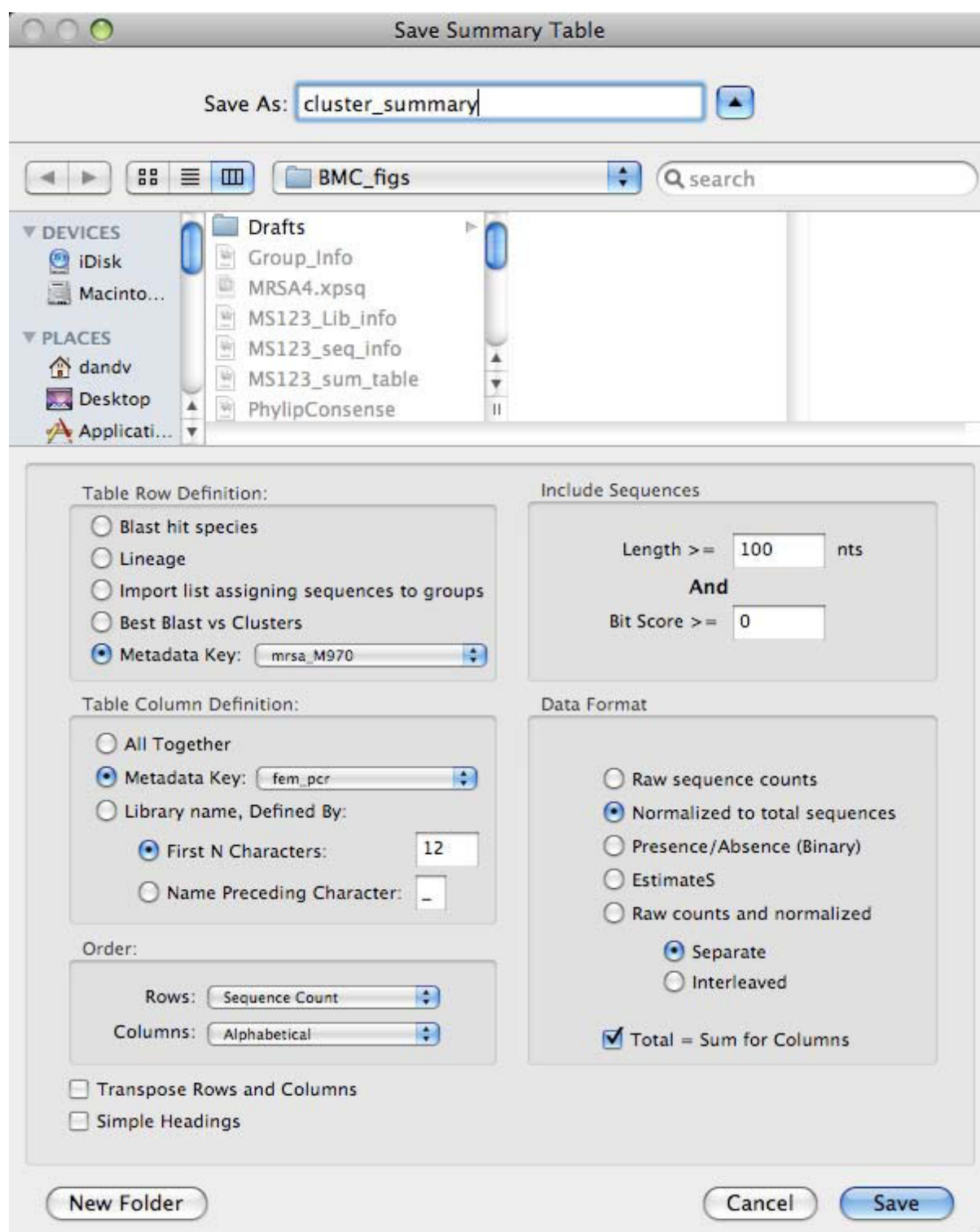
Estimates of biodiversity indices (species richness, diversity, evenness) can be reported through either of two modes. First, the export function OTU Diversity... reports basic calculations of commonly used indices ( $S_{obs}$ ,  $S_{chao1}$ ,  $C_{ACE}$ , Good's coverage, Shannon diversity; [40]) for a set of selected sequences. Alternatively, the same biodiversity estimates can be made in a more thorough manner through execution of the analysis function Biodiversity (biodiv)..., which invokes the program biodiv, a standalone command-line tool built in conjunction with XplorSeq. As shown in Fig. 6G, the user selects OTUs definitions for the selected sequences through choice of meta-data options. To compare indices between different groups of sequences, the user can also select multiple "environments" by which to differentiate the sequence subsets; biodiv then performs separate analyses for each designated environment. Biodiv performs random resampling of OTUs and calculates collector's curves and associated biodiversity indices as a function of sampling effort [41]. Biodiv also reports rarefied biodiversity indices, based on resampling, with 95% confidence intervals for each type of environment [41].

**Tools for data export and transformation**

XplorSeq provides several features by which to export sequences and data for further analyses. Data reports are generated in tab-delimited format for import into spreadsheets. A particularly useful export feature tabulates sequence abundances and/or prevalences on the basis of user-defined terms that define how data are categorized and enumerated. For instance, as shown in Fig. 7, the user has opted to create a table in which rows are defined by meta-data that specify OTUs and columns are specified by meta-data that organize samples by the results of a PCR assay; alternatively, similar sequences can be grouped by BLAST results, or phylogenetic lineage. The cells of the output tables enumerate the abundance or presence/absence of sequences for a given row and column (Additional file 1).

**Figure 6**

**Analyses of aligned sequences.** XplorSeq provides GUI-based access to several command-line programs used for phylogenetic analysis of multiple sequence alignments, including A-D) several commonly used programs from the phylip package [37]; E) RAxML for maximum-likelihood phylogenetic inference; F) sortx, for rapid clustering of sequences into OTUs; and G) biodiv for estimation of biodiversity indices through resampling statistics.

**Figure 7**

**Tabulation of sequence abundance/prevalence.** The Summary Table dialog provides multiple means of tabulating sequence data. In this example, rows are defined by the values of a meta-data key (e.g. 97% OTUs), while columns are defined by the value associated with another meta-data key (e.g. PCR results). The "Data Format" panel specifies that sequence counts in each column are normalized to the total of each column.

Preparation of sequence data for submission to public databases can be tedious and error-prone, despite the availability of tools such as NCBI's sequin software tool. XplorSeq can facilitate the submission process by automating the organization of sequences and their annotations into a form suitable for input into sequin (Fig. 8). The user can modify the output of the sequin script export function to include or exclude particular information and to tailor GenBank descriptors and/or features (e.g. Locus, Definition, Molecule fields).

For detailed phylogenetic studies, sequences assembled by XplorSeq can be exported in formats suitable for input into other software packages, such as ARB [42], DOTUR [43], NAST [44], SILVA [45], or UniFrac [46]. Users can choose sequences to be exported by selecting particular objects or through filters for sequence length, bit-score, type of sequence (i.e. contig vs. raw sequence), or other meta-data. Finally, "placeholder" Newick-formatted phylogenetic trees can be exported in order to select specific subsets of sequences in tree visualization software such as ARB. In this format, taxa belonging to user-defined subsets (categorized, for instance, through specific meta-data) are placed in clades and assigned artificial branch-lengths of 0.1; inter-group distances are set to 0.9, which produces a well-demarcated organization of sequences into assigned groups. Import of such a tree into ARB allows rapid, graphical marking of different sets of taxa (Fig. 9), the results of which are propagated to other "real" trees in ARB.

### Performance issues

The data analysis tools employed by XplorSeq (e.g. phred, phrap, blast, sortx, biodiv) are invoked in separate threads, so that multiple tasks can be run in parallel. During these steps, feedback to the GUI is limited in order to minimize the overhead of performing analyses through XplorSeq. Typical latencies associated with XplorSeq contribute no more than an additional ~10% to the total elapsed time required for an analysis, relative to the same task performed through command line execution of the underlying 3<sup>rd</sup> party software (Table 2). This overhead is due primarily to data import/export and parsing of results, which add value to the functionalities provided by command line utilities. Consequently, XplorSeq should not place any undue limitations on the use of command-line software to process and analyze sequence data, beyond those inherent in the computer system and/or software tool being used. In other words, tasks that can be performed on a particular Macintosh system through the command line should also be feasible through XplorSeq.

The XplorSeq file format, implemented using the Cocoa software framework, does not significantly add to the size

of data files. For example, an XplorSeq file containing an alignment of 250,000 rRNA sequences (1585 nucleotides per sequence) requires 396 megabytes (MB) of storage, compared to 381 MB for a fasta formatted file storing the same alignment. The addition of metadata, such as blast results, increases file size in roughly linear proportion to the size of parsed input text.

Although the use of a flat file format by XplorSeq greatly simplifies data storage and transfer, it does require that all data be read into memory before being manipulated. For relatively large files, bottlenecks are apparent primarily in tasks requiring import and export of data. Thus, sequence analysis projects are likely to be limited as much by computer hardware (e.g., quantity of random access memory and bus speeds) and the performance of underlying 3<sup>rd</sup> party software tools as by XplorSeq. However, the XplorSeq environment is scalable from laptops to more advanced workstations (e.g., 8-core/32 GB systems currently available) so its capabilities can be expanded as need arises and hardware evolves.

As a rough guide to system requirements, Table 2 presents benchmark comparisons of common XplorSeq tasks performed on a laptop and a workstation. To open the 250,000 sequence XplorSeq alignment file described above, for example, requires ~130 or ~12 seconds, on the laptop or workstation, respectively. Test datasets of  $1 \times 10^6$  25-nucleotide sequences can be manipulated with similar lag times. More fully annotated files containing ~50,000 16S rRNA sequences, each of length ~1000 nucleotides, have been used routinely on a laptop with little performance degradation [17,23,27]. However, projects with > 50,000 annotated sequences likely will require higher performance workstations to insure the responsiveness customary to users of GUI-based software. Alternatively, sequence data can be spread across multiple files, each of which maintains information generated from a particular experiment (e.g. a PCR amplicon library). More complex data storage strategies, such as the use of application-specific databases may be implemented in the future if they do not compromise the XplorSeq philosophy of ease of software installation and use.

### Conclusion

Although XplorSeq was developed to expedite the phylogenetic analysis of ribosomal RNA (rRNA) gene libraries, it should prove useful in any sequencing project, particularly ones facilitated by batch analysis of multiple clones. Moreover, any UNIX-based DNA sequence analysis tool that can be ported to Mac OSX can be readily incorporated into XplorSeq. Suggestions for the addition of other modules to the XplorSeq package are most welcome.

**Scripts**

Locus: [clone]

Organism: Uncultured [lineage\_2] [clone]

Clone: [clone]

Definition: Uncultured [domain adj] clone [clone] [SSU], partial sequence

Note: Partial taxonomy: [lineage\_all]

Source:

Lineage: [lineage\_all]

Molecule: RNA ☒ Environmental Samples

MolType: Ribosomal RNA

The following tokens can be used to provide clone-specific information:

[clone] = Clone name (Group or Sequence name)  
 [domain noun] = Domain name noun, specified in Group Info Window (e.g. Bacterium)  
 [domain adj] = Domain name adjective, specified in Group Info Window (e.g. bacterial)  
 [lineage\_all] = Full lineage specified in Group Info Window  
 [lineage\_1] = First term listed in lineage information  
 [lineage\_2] = Second term listed in lineage information  
 [lineage\_last] = Last term listed in lineage information  
 [SSU] = 12S/16S/18S ribosomal RNA gene (determined by [domain])  
 [LSU] = 23S/28S ribosomal RNA gene (determined by [domain])

**Sequences**

☒ Best Sequences  
☐ All Sequences  
☒ Remove gaps

Cancel

Export

**Oligos**

☒ Forward Name: SSU\_8F Sequence: AGAGTTTGATCCTGGCTCAG

☒ Reverse Name: SSU\_805R Sequence: GACTACCAGGTATCTAAT

**Example**

```
>VA02_01 [lineage=Bacteria; Firmicutes; Bacillales; Staphylococcus.] [clone=VA02_01]
[organism=Uncultured Firmicutes VA02_01] [note=Partial taxonomy: Bacteria;
Firmicutes; Bacillales; Staphylococcus.] Uncultured bacterial clone VA02_01 16S
ribosomal RNA, partial sequence [molecule=RNA] [moltype=Ribosomal RNA]
[environmental-sample=] [fwd-pcr-primer-name=SSU_8F] [fwd-pcr-primer-
seq=AGAGTTTGATCCTGGCTCAG] [rev-pcr-primer-name=SSU_805R] [rev-pcr-primer-
seq=GACTACCAGGTATCTAAT]
CAGCAGACGCTTCCTCTCTCATCTTACCGCCGACCGCTCAGTAACACCTGCATAA
```

Example

**Figure 8**

**Sequin script export.** Scripted export of sequence data for GenBank submission through Sequin. Data associated with each sequence can be manipulated in order to tailor the level of detail that will go into Sequin.



**Input of placeholder tree into ARB.** Sequences were exported from XplorSeq using the "Placeholder Tree" option. Sequences were split into two categories on the basis of associated meta-data (in this example, results of a PCR screen). Here, the user has selected all taxa belonging to the "neg" group. Because ARB propagates taxa markings between trees, placeholder trees can be used to graphically organize and manipulate groups of sequences that aren't necessarily related.

**Table 2: Execution times of commonly used software: comparison of XplorSeq with command line implementation**

Program	Execution Time (sec.) <sup>1</sup>				Task	Sequence Data
	System A <sup>2</sup>		System B <sup>3</sup>			
	XplorSeq <sup>4</sup>	Command Line	XplorSeq <sup>4</sup>	Command Line		
phred	50.5 (5.3) <sup>1</sup>	51.0 (2.9) <sup>1</sup>	23.3 (0.8) <sup>1</sup>	19.3 (0.5) <sup>1</sup>	Basecall	768 .esd files.
phrap	167.0 (2.4)	153.3 (0.8)	30.5 (1.2)	28.8 (1.5)	Contig	384 pairs of reads.
blastall	368.0 (2.4)	345.3 (0.8)	228.5 (1.2)	221.3 (2.4)	Local blast	24 1585-mers
XplorSeq	69.0 (8.1)	na	21.2 (0.4)	na	Import fasta	250,000 1585-mers
XplorSeq	130.8 (10.8)	na	12.2 (0.4)	na	Open XplorSeq file	250,000 1585-mers
XplorSeq	160.8 (14.4)	na	11.7 (0.5)	na	Save XplorSeq file	250,000 1585-mers
XplorSeq	60.5 (0.8)	na	46.3 (1.4)	na	Import fasta	1,000,000 25-mers
XplorSeq	216.3 (11.2)	na	28.0 (0.6)	na	Open XplorSeq file	1,000,000 25-mers
XplorSeq	34.3 (3.8)	na	16.5 (0.5)	na	Save XplorSeq file	1,000,000 25-mers

<sup>1</sup>Elapsed time of execution: Mean (St. Dev) seconds.

<sup>2</sup>Executed on a 2 GHz Intel Core Duo MacBook Pro. 1 GB 667 MHz DDR2 SDRAM. Mac OSX version 10.5.4.

<sup>3</sup>Executed on a workstation with 2 × 3 GHz Quad-Core Intel Xeon processors. 8 GB 800 MHz DDR2 FB-DIMM. Mac OS X Server version 10.5.4

<sup>4</sup>Elapsed time includes export and import of data.



## Availability and requirements

Project name: XplorSeq

Project home page: <http://vent.colorado.edu/phyloware>

Operating system: Macintosh OS X (currently requires 10.4.x or 10.5.x)

Programming language: Cocoa/Objective-C, C, C++

Other requirements: phred and phrap are available at <http://www.phrap.org>

License: Daniel N. Frank. Free for non-commercial use

Any restrictions to use by non-academics: Contact corresponding author. Users are requested to notify the corresponding author when XplorSeq is cited.

## Additional material

### Additional file 1

Example of sequence enumeration table.

Click here for file

[<http://www.biomedcentral.com/content/supplementary/1471-2105-9-420-S1.pdf>]

## Acknowledgements

The author wishes to thank Prof. Norman R. Pace for encouragement and support, Charles E. Robertson for invaluable mentorship in software engineering and web-site support and Laura Baumgartner, J. Kirk Harris, Jeffrey Walker and members of the Pace laboratory for extensive software testing and feedback. Both anonymous reviewers are thanked for their constructive feedback.

## References

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *J Mol Biol* 1990, **215**(3):403-410.
- Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1990, **18**(17):3389-3402.
- Wheeler DL, Barrett T, Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, et al.: **Database resources of the National Center for Biotechnology Information.** *Nucleic Acids Res* 2007:D5-12.
- Ley RE, Backhed F, Turnbaugh P, Lozupone CA, Knight RD, Gordon JI: **Obesity alters gut microbial ecology.** *Proc Natl Acad Sci USA* 2005, **102**(31):11070-11075.
- McManus CJ, Kelley ST: **Molecular survey of aeroplane bacterial contamination.** *J Appl Microbiol* 2005, **99**(3):502-508.
- Papineau D, Walker JJ, Mojszis SJ, Pace NR: **Composition and structure of microbial communities from stromatolites of Hamelin Pool in Shark Bay, Western Australia.** *Appl Environ Microbiol* 2005, **71**(8):4822-4832.
- Spear JR, Walker JJ, McCollom TM, Pace NR: **Hydrogen and bioenergetics in the Yellowstone geothermal ecosystem.** *Proc Natl Acad Sci USA* 2005, **102**(7):2555-2560.
- Spear JR, Walker JJ, Pace NR: **Hydrogen and primary productivity: Inference of biogeochemistry from phylogeny in a geothermal ecosystem.** In *Geothermal Biology and Geochemistry in Yellowstone National Park* Edited by: Inskeep WP, McDermott TR. Bozeman, MT: Thermal Biology Institute, Montana State University; 2005:113-128.
- Walker JJ, Spear JR, Pace NR: **Geobiology of a microbial endolithic community in the Yellowstone geothermal environment.** *Nature* 2005, **434**:1011-1014.
- Baumgartner LK, Reid RP, Dupraz C, Decho AW, Buckley DH, Spear JR, Przekop KM, Visscher PT: **Sulfate reducing bacteria in microbial mats: changing paradigms, new discoveries.** *Sedimentary Geology* 2006, **185**:131-145.
- Dalby AB, Frank DN, St Amand AL, Bendele AM, Pace NR: **Culture-independent analysis of indomethacin-induced alterations in the rat gastrointestinal microbiota.** *Appl Environ Microbiol* 2006, **72**(10):6707-6715.
- Ley RE, Harris JK, Wilcox J, Spear JR, Miller SR, Bebout BM, Maresca JA, Bryant DA, Sogin ML, Pace NR: **Unexpected diversity and complexity of the Guerrero Negro hypersaline microbial mat.** *Appl Environ Microbiol* 2006, **72**(5):3685-3695.
- Rawls JF, Mahowald MA, Ley RE, Gordon JI: **Reciprocal gut microbiota transplants from zebrafish and mice to germ-free recipients reveal host habitat selection.** *Cell* 2006, **127**(2):423-433.
- Salmassi TM, Walker JJ, Newman DK, Leadbetter JR, Pace NR, Hering JG: **Community and cultivation analysis of arsenite oxidizing biofilms at Hot Creek.** *Environ Microbiol* 2006, **8**(1):50-59.
- Spear JR, Walker JJ, Pace NR: **Microbial ecology and energetics in Yellowstone hot springs.** *Yellowstone Science* 2006, **14**(1):17-24.
- Turnbaugh PJ, Ley RE, Mahowald MA, Magrini V, Mardis ER, Gordon JI: **An obesity-associated gut microbiome with increased capacity for energy harvest.** *Nature* 2006, **444**(7122):1027-1031.
- Frank DN, St Amand AL, Feldman RA, Boedeker EC, Harpaz N, Pace NR: **Molecular-phylogenetic characterization of microbial community imbalances in human inflammatory bowel diseases.** *Proc Natl Acad Sci USA* 2007, **104**(34):13780-13785.
- Harris JK, De Groote MA, Sagel SD, Zemanick ET, Kapsner R, Penvari C, Kaess H, Deterding RR, Accurson FJ, Pace NR: **Molecular identification of bacteria in bronchoalveolar lavage fluid from children with cystic fibrosis.** *Proc Natl Acad Sci USA* 2007, **104**(51):20529-20533.
- Lee L, Tin S, Kelley ST: **Culture-independent analysis of bacterial diversity in a child-care facility.** *BMC Microbiol* 2007, **7**(1):27.
- Spear JR, Barton HA, Robertson CE, Francis CA, Pace NR: **Microbial community biofabrics in a geothermal mine adit.** *Appl Environ Microbiol* 2007, **73**(19):6172-6180.
- Walker JJ, Pace NR: **Phylogenetic Composition of Rocky Mountain Endolithic Microbial Ecosystems.** *Appl Environ Microbiol* 2007, **73**(11):3497-3504.
- Feazel LM, Spear JR, Berger AB, Harris JK, Frank DN, Ley RE, Pace NR: **Eucaryotic Diversity in a Hypersaline Microbial Mat.** *Appl Environ Microbiol* 2008, **74**(1):329-332.
- Frank DN, Pace NR: **Gastrointestinal microbiology enters the metagenomics era.** *Curr Opin Gastroenterol* 2008, **24**(1):4-10.
- Frank DN, Wysocki A, Specht-Glick DD, Rooney A, Feldman RA, St Amand AL, Pace NR, Trent JD: **Microbial diversity in chronic open wounds determined by culture-independent molecular methods.** *Wound Repair and Regeneration* 2008 in press.
- Isenbarger TA, Finney M, Rios-Velazquez C, Handelsman J, Ruvkun G: **Miniprimer PCR, a new lens for viewing the microbial world.** *Appl Environ Microbiol* 2008, **74**(3):840-849.
- Ley RE, Hamady M, Lozupone C, Turnbaugh PJ, Ramey RR, Bircher JS, Schlegel ML, Tucker TA, Schrenzel MD, Knight R, et al.: **Evolution of Mammals and Their Gut Microbes.** *Science* 2008.
- Peterson DA, Frank DN, Pace NR, Gordon JI: **Metagenomic approaches for defining the pathogenesis of inflammatory bowel diseases.** *Cell Host Microbe* 2008, **3**(6):417-427.
- Sahl JW, Schmidt R, Swanner ED, Mandernack KW, Templeton AS, Kieft TL, Smith RL, Sanford WE, Callaghan RL, Mitton JB, et al.: **Sub-surface Microbial Diversity in Deep-Granitic-Fracture Water in Colorado.** *Appl Environ Microbiol* 2008, **74**(1):143-152.
- Turnbaugh PJ, Backhed F, Fulton L, Gordon JI: **Diet-induced obesity is linked to marked but reversible alterations in the mouse distal gut microbiome.** *Cell Host Microbe* 2008, **3**(4):213-223.
- Denisov GA, Arehart AB, Curtin MD: **A system and method for improving the accuracy of DNA sequencing and error probability estimation through application of a mathematical**

model to the analysis of electropherograms. Edited by 6681186 USP 2004.

31. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities.** *Genome Res* 1998, **8(3)**:186-194.
32. Ewing B, Hillier L, Wendl MC, Green P: **Base-calling of automated sequencer traces using phred. I. Accuracy assessment.** *Genome Res* 1998, **8(3)**:175-185.
33. **TIGR Assembler** [<http://www.jcvi.org/cms/research/software/>]
34. Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, Higgins DG, Thompson JD: **Multiple sequence alignment with the Clustal series of programs.** *Nucleic Acids Res* 2003, **31(13)**:3497-3500.
35. Higgins DG, Thompson JD, Gibson TJ: **Using CLUSTAL for multiple sequence alignments.** *Methods Enzymol* 1996, **266**:383-402.
36. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22(22)**:4673-4680.
37. Felsenstein J: **PHYLIP (Phylogeny Inference Package) version 3.6.** Distributed by the author. Department of Genome Sciences, University of Washington, Seattle. 2005.
38. Sheneman L, Evans J, Foster JA: **Clearcut: a fast implementation of relaxed neighbor joining.** *Bioinformatics* 2006, **22(22)**:2823-2824.
39. Stamatakis A, Ludwig T, Meier H: **RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees.** *Bioinformatics* 2005, **21(4)**:456-463.
40. Schloss PD, Handelsman J: **The last word: books as a statistical metaphor for microbial communities.** *Annu Rev Microbiol* 2007, **61**:23-34.
41. Magurran AE: **Measuring Biological Diversity.** Malden, USA: Blackwell Publishing; 2003.
42. Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadukumar, Buchner A, Lai T, Steppi S, Jobb G, et al.: **ARB: a software environment for sequence data.** *Nucleic Acids Res* 2004, **32(4)**:1363-1371.
43. Schloss PD, Handelsman J: **Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness.** *Appl Environ Microbiol* 2005, **71(3)**:1501-1506.
44. DeSantis TZ Jr, Hugenholtz P, Keller K, Brodie EL, Larsen N, Piceno YM, Phan R, Andersen GL: **NAST: a multiple sequence alignment server for comparative analysis of 16S rRNA genes.** *Nucleic Acids Res* 2006:W394-399.
45. Pruesse E, Quast C, Knittel K, Fuchs B, Ludwig W, Peplies J, Glockner FO: **SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB.** *Nucleic Acids Res* 2007, **35(21)**:7188-7196.
46. Lozupone C, Knight R: **UniFrac: a new phylogenetic method for comparing microbial communities.** *Appl Environ Microbiol* 2005, **71(12)**:8228-8235.

Publish with **BioMed Central** and every scientist can read your work free of charge

"BioMed Central will be the most significant development for disseminating the results of biomedical research in our lifetime."

Sir Paul Nurse, Cancer Research UK

Your research papers will be:

- available free of charge to the entire biomedical community
- peer reviewed and published immediately upon acceptance
- cited in PubMed and archived on PubMed Central
- yours — you keep the copyright

Submit your manuscript here:  
[http://www.biomedcentral.com/info/publishing\\_adv.asp](http://www.biomedcentral.com/info/publishing_adv.asp)

