DuEPublico
Duisburg-Essen Publications online

CANGS DB: a stand-alone web-based database tool for processing, managing and analyzing 454 data in biodiversity studies

Pandey, Ram Vinay; Nolte, Viola; Boenigk, Jens; Schlötterer, Christian

This text is provided by DuEPublico, the central repository of the University Duisburg-Essen.

This version of the e-publication may differ from a potential published print or online version.

DOI: https://doi.org/10.1186/1756-0500-4-227

URN: urn:nbn:de:hbz:464-20180912-135221-1

Link: https://duepublico.uni-duisburg-essen.de:443/servlets/DocumentServlet?id=47011

Legal notice:
© 2011 Pandey et al; 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.

Source: BMC Research Notes, 2011, 4:227; Published: 30 June 2011

Pandey et al. BMC Research Notes 2011, 4:227
http://www.biomedcentral.com/1756-0500/4/227

BMC Research Notes

RESEARCH ARTICLE Open Access

CANGS DB: a stand-alone web-based database tool for processing, managing and analyzing 454 data in biodiversity studies

Ram Vinay Pandey[1], Viola Nolte[1], Jens Boenigk[2] and Christian Schlötterer[1*]

Abstract

Background: Next generation sequencing (NGS) is widely used in metagenomic and transcriptomic analyses in biodiversity. The ease of data generation provided by NGS platforms has allowed researchers to perform these analyses on their particular study systems. In particular the 454 platform has become the preferred choice for PCR amplicon based biodiversity surveys because it generates the longest sequence reads. Nevertheless, the handling and organization of massive amounts of sequencing data poses a major problem for the research community, particularly when multiple researchers are involved in data acquisition and analysis. An integrated and user-friendly tool, which performs quality control, read trimming, PCR primer removal, and data organization is desperately needed, therefore, to make data interpretation fast and manageable.

Findings: We developed CANGS DB (Cleaning and Analyzing Next Generation Sequences DataBase) a flexible, stand alone and user-friendly integrated database tool. CANGS DB is specifically designed to organize and manage the massive amount of sequencing data arising from various NGS projects. CANGS DB also provides an intuitive user interface for sequence trimming and quality control, taxonomy analysis and rarefaction analysis. Our database tool can be easily adapted to handle multiple sequencing projects in parallel with different sample information, amplicon sizes, primer sequences, and quality thresholds, which makes this software especially useful for non-bioinformaticians. Furthermore, CANGS DB is especially suited for projects where multiple users need to access the data. CANGS DB is available at http://code.google.com/p/cangsdb/.

Conclusion: CANGS DB provides a simple and user-friendly solution to process, store and analyze 454 sequencing data. Being a local database that is accessible through a user-friendly interface, CANGS DB provides the perfect tool for collaborative amplicon based biodiversity surveys without requiring prior bioinformatics skills.

Background

Next generation sequencing technologies are delivering data at a hitherto unprecedented speed and dramatically reduced costs. In addition to genome sequencing and transcriptome profiling, ultra-deep sequencing of short amplicons offers an enormous potential in clinical studies [1] and in surveys of ecological diversity [2-4]. Typical biodiversity surveys include sequences from a diverse set of samples. An effective data analysis requires the ability to link additional data, such as time of collection and ecological variables, to the sequences. Furthermore, biodiversity surveys often require sequence information on different taxonomic levels. Hence, researchers need an analytical tool that provides the flexibility to handle different PCR primers.

Until now several tools have been developed, but none of them unite all of the requirements for a comprehensive tool. In the following we briefly introduce these tools, highlight their features, and discuss missing options.

1) RDP [5] is an online tool for sequence trimming and filtering. It provides an excellent taxonomic classifier, which is, however, limited to small ribosomal subunit gene sequences from bacteria and archea. Furthermore, it provides no option to store and manage data provided by the user. MOTHUR [6] combines read

* Correspondence: christian.schloetterer@vetmeduni.ac.at
[1]Institut für Populationsgenetik, Veterinärmedizinische Universität Wien, Veterinärplatz 1, Vienna, Austria
Full list of author information is available at the end of the article

BioMed Central
© 2011 Pandey et al; 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.

trimming and filtering capabilities along with rare-faction analyses. MOTHUR is a command line software and provides many useful utility commands for biodiversity studies but it does not offer a data storage option. CANGS [7] and CANGS DB rely on MOTHUR for rarefaction analyses. VAMPS [8] provides sequence trimming, filtering of low quality reads and taxonomic path assignment using the GAST pipeline. The user can upload data for visualization and analysis of microbial population structures. The limitation of VAMPS is a rigid sequence-processing pipeline that does not allow for user-defined options (e.g.: reads are only filtered allowing for ambiguities, it is not possible to define a size range for amplicon sizes, and quality scores of the sequence reads are not accounted for). Furthermore, it is not possible to store additional data about the sequences, such as ecological variables. Finally, the user cannot retrieve data according to user-defined criteria. PANGEA [9] allows for trimming of the barcodes and groups sequences according to the barcode. PANGEA has many useful features including clustering, classification, and comparison of microbial communities. While PANGEA uses a local database for classification, it is not designed to incorporate user-generated sequences into this database. Thus, data manipulation and organization of 454 data from multiple runs is not possible.

We developed CANGS DB (http://code.google.com/p/cangsdb/) as an integrated user-friendly database tool that can be easily installed on local computers and accessed through the internet by standard browsers. It offers a flexible, customizable sequence-processing pipeline where 454 sequences can be uploaded/downloaded and data can be manipulated via a user-friendly interface. A variety of tools are available in the CANGS DB web interface for the downstream analysis of stored 454 sequencing data. CANGS DB links external information, such as details about the collection site, time of the year and environmental variables, to the sequence information. This allows the user to extract sequences according to combinations of particular variables (e.g.: all sequences obtained from water samples with a given temperature). A demo of CANGS DB is running on http://i122mc100.vu-wien.ac.at/CANGSdb/

Construction and content

Database and web interface development

The CANGS DB is completely written in Perl and uses data stored in a relational database (MySQL). The relational database schema is shown in Figure 1. CANGS DB web interface is developed using the CGI.pm and runs of Apache (2.0.53) web server. The interaction between user interface and database is established by using DBI.pm and DBD::mysql.pm modules. CANGS DB can be run on Mac OS, Linux and other Unix-like systems.

Required programs are Bioperl [10], BLAST [11] for the similarity search in the taxonomy analysis tool, MAFFT [12] for pairwise distance calculation (MAFFT is used for pairwise alignment) in the rarefaction analysis tool, MOTHUR [6] for estimating the number of species (OTUs), MySQL [13] for data storage, update_blastdb.pl [14] for downloading the BLAST database on a local computer, and R [15] to plot rarefaction curves from MOTHUR output.

Processing the raw sequence data

CANGS DB uses the basic pipeline of CANGS [7], but has been modified to provide more flexibility and some additional features. CANGS DB processes raw sequences to provide the user with high quality 454 reads by trimming the adapter B, barcodes and PCR primers and filtering the low quality reads as described below. The sequence-processing pipeline of CANGS DB is highly flexible and user friendly. Each step in the pipeline can be modified on the interface (Figure 2). Moreover, CANGS DB is able to handle multiple primers of different length in the trimming step. Most importantly, filtering according to sequence length can be individually defined for each amplicon.

1. Removal of adapter B

based on the sequences of adapter B, as specified in the primer & barcodes input file, the 3'- end of each read is trimmed. It is possible to process only sequences with a perfect match to adapter B. Alternatively, a pattern search that allows for imperfection in adapter B can also be used, allowing more sequences to be recovered. CANGS DB also enables trimming multiple adapter B sequences of different lengths. Poly N tails at the 3' end of the 454 reads are removed before Adapter B trimming.

2. Filtering sequences with ambiguities

CANGS DB allows the removal of reads with one or more Ns (unknown bases).

3. Grouping of sequences according to bar codes

CANGS DB groups sequences based on the barcodes specified in the primer & barcodes input file. Sequences with the same barcode are grouped into one category. This step is skipped when only a single sample is processed. CANGS DB is designed to allow for barcodes of different length.

4. Removal of singletons

to ameliorate the problem of sequencing errors and chimeric sequences generated by jumping PCR, CANGS DB allows the user to remove low frequency variants from the data set. Sequences must be present in at least two different samples in the entire data set before trimming of the primers. This is more stringent than the former criterion implemented in CANGS [7], namely a sequence being present two or three times in the whole

Figure 1 The CANGS DB database Entity Relationship Data Model (ERDM). There are two static tables (data_upload_summary and statistics) and seven dynamic tables (sample_info_ table_number, primer_barcode_info_ table_number, raw_454_data_ table_number, redundant_sequence_table_number, nonredundant_sequence_pooled_ table_number, nonredundant_sequence_sample_ table_number, nonredundant_pooled2sample_ table_number), CANGS DB creates these dynamic set of tables with each new data set. In this figure the primary key (PK) is given in red color, Foreign key (FK) is given in green color and Primary+Foreign key (PFK) is given in blue color.

dataset. By conditioning on the presence in two independent data sets CANGS DB excludes chimeric sequences more efficiently, as a chimeric sequence could occur multiple times in a data set, in particular if the chimeric molecule has been generated during the early PCR steps. Note that several data sets could be combined to minimize the removal of true low frequency sequence variants.

5. Filtering sequences according to length threshold

CANGS DB removes sequence reads falling outside the size range specified in the primer & barcode input file.

6. Removal of PCR primers

forward and reverse PCR primers are specified in the primer & barcode input file and removed from the sequence. Only sequences with perfect identity to the specified PCR primers are processed. The 454

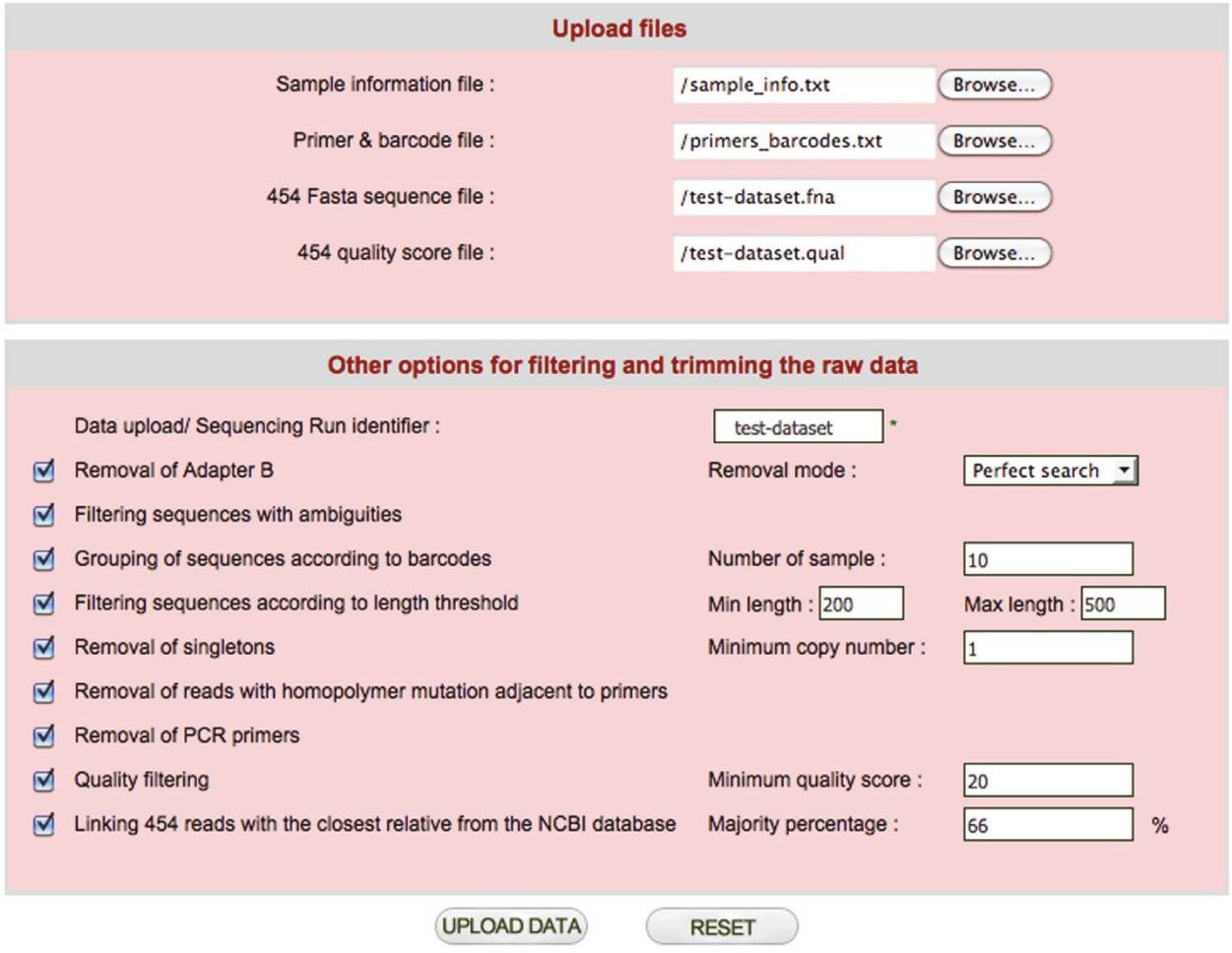


Figure 2 The 454 sequence trimming and filtering interface. Interface for trimming and quality filtering of new sequences to be added to the database.

sequencing process preferentially generates length variants in homopolymers. As homopolymers can be as short as two bases and the target sequence is frequently not known, we developed a special procedure to recognize such sequencing errors at the end of the PCR primer as described in [7].

7. Quality filtering

CANGS DB averages the quality values for each base in a read. Quality values are taken from the .qual file after the values corresponding to adapter B, bar code and primer bases have been removed in step 1, step 3 and step 6, respectively. Sequence reads with a quality value lower than the threshold specified in the user interface page will be discarded. Note that the quality filtering may result in new singletons, which remain in the data set, as the quality filtering is the last step in the analysis.

After trimming the sequence reads, CANGS DB creates a non-redundant sequence data set in order to reduce the computational burden for further analysis. In the non-redundant sequence data set each sequence variant is only represented once. Note that in this step indels are considered to be informative. Hence, two sequences differing only by an indel will be listed independently in the non-redundant data set. The frequency of each sequence in the non-redundant data set is included in the FASTA header. The output file contains non-redundant reads that are ranked by copy number in descending order.

Taxonomic classification assignment

CANGS DB uses the CANGS [7] taxonomy analysis pipeline to assign a taxonomic path to the newly sequenced 454 reads. CANGS taxonomy analysis pipeline does not rely on a pre-curated database for either 16S or 18S sequence like RDP classifier [5], thus it can handle sequences from any genomic region.

Utility and Discussion

Query interface

One of the unique features of CANGS DB is its powerful query interface. The query interface was designed with the diversity of 454 data (ecological or clinical) in mind that need to be stored and queried. In case of ecological survey sequencing data, users can retrieve subsets of the sequences in the data base according to user defined variables, such as time of the year, temperature, pH, sampling location etc. In case of clinical data, the user can retrieve sequences according to tissue type, experiment date, sampling date, time, and other related information.

Figure 3 shows how the data search could be customized:

1) According to data sets: it is possible to select any combination of data sets loaded in the database.

2) According to PCR primer ID

3) According to sample ID: the use of barcodes permits the sequencing of multiple samples in one experiment. This option permits the user to restrict the search to specific sample IDs.

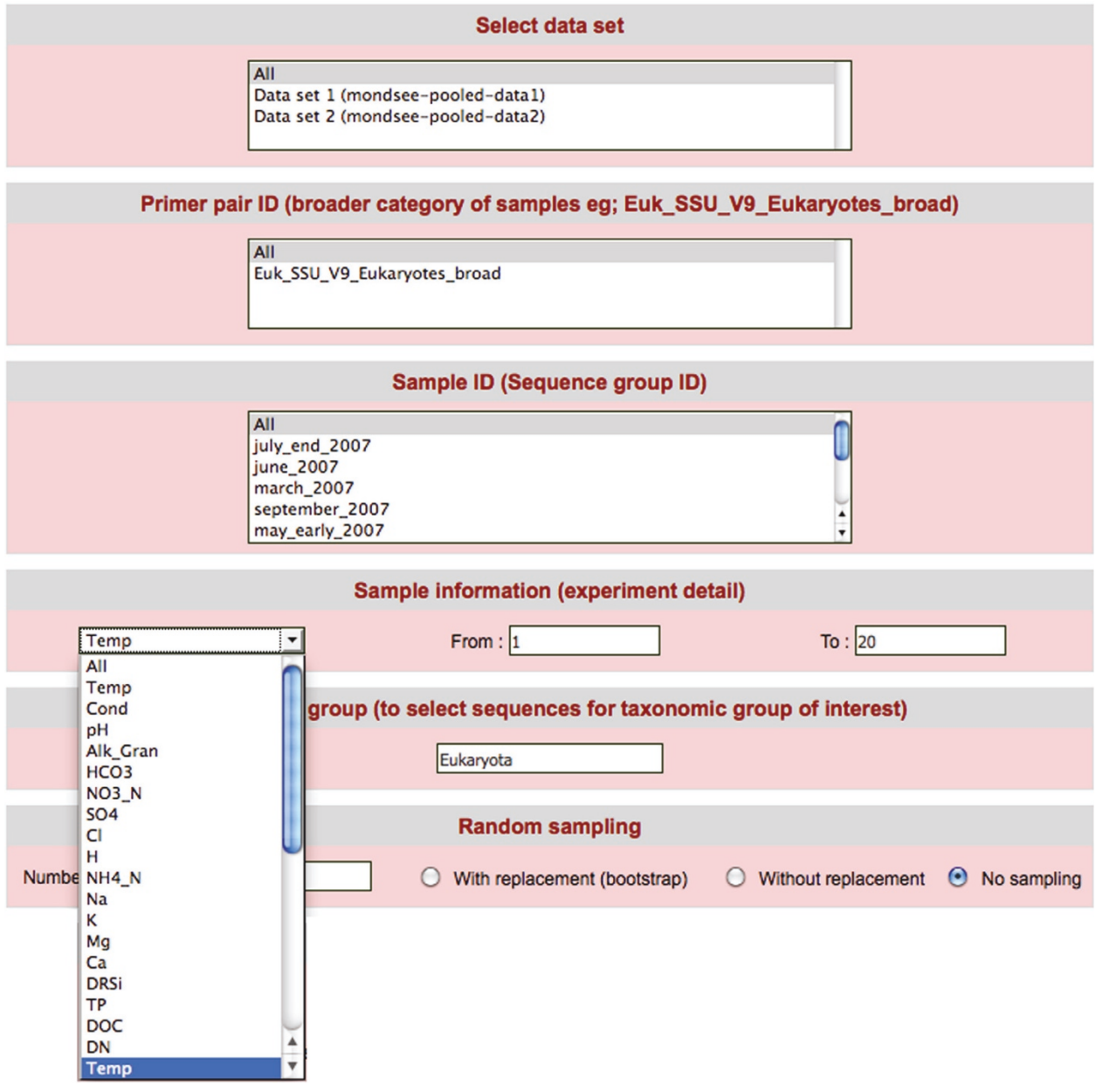


Figure 3 The CANGS DB data retrieval query interface.

4) According to external sample information: A drop down menu displays all external data categories (e.g.: temperature, pH, depth, conductivity etc). The user can then specify the desired range for each of the external data categories (e.g.: temperature range: 10-20° C)

5) According to taxonomic group

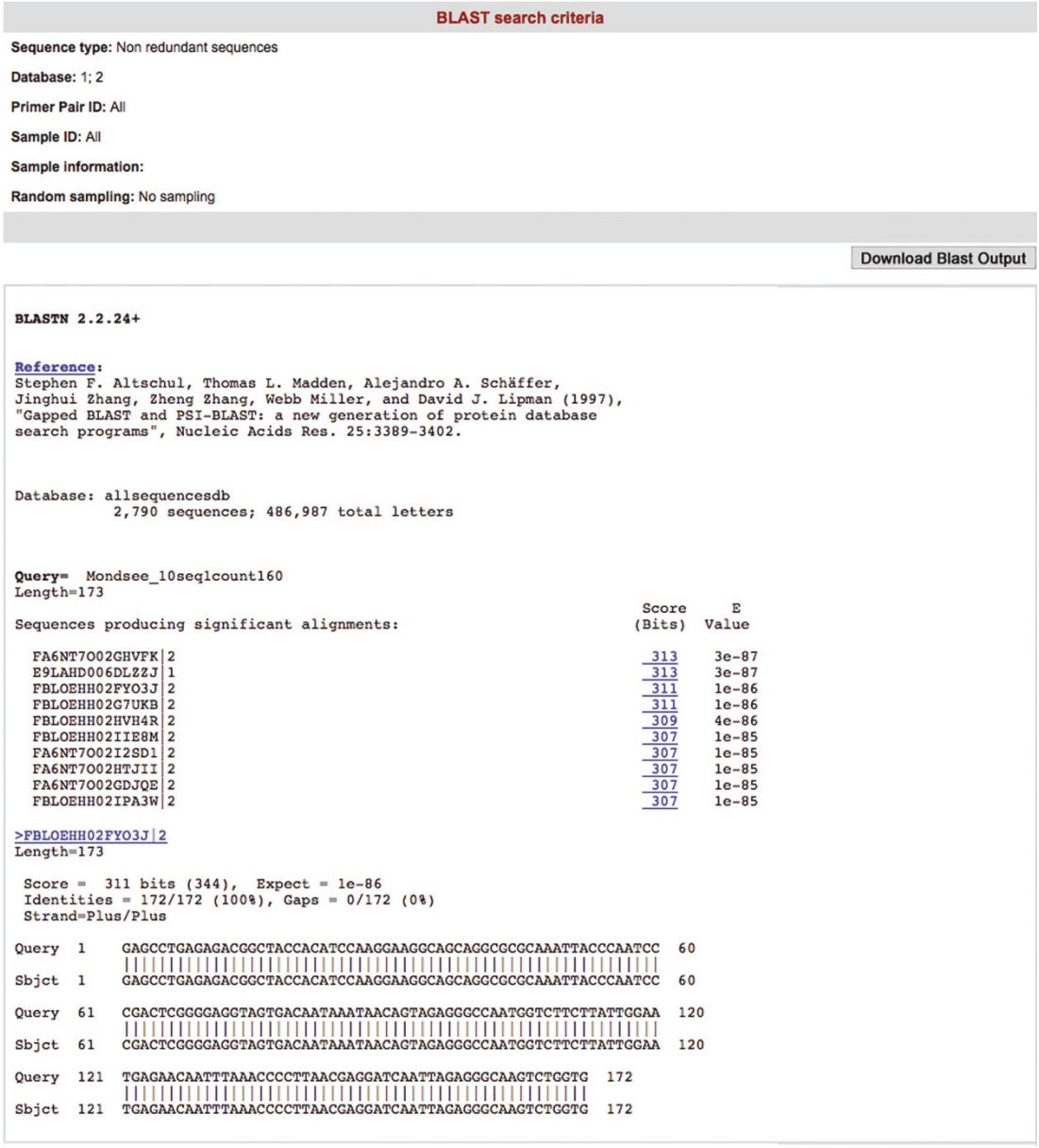
6) According to sample size: The user can select a sample size and obtain random samples from all selected data sets for downstream analysis. This is particularly helpful for analyses that are sensitive to different sample sizes.

The query interface is also integrated with the Blast and rarefaction interfaces so that these analyses can be performed on the subsequent query output.

Sequence Analysis Tools

CANGS DB provides several tools for the analysis of stored data.

(1) **Blast analysis tool** (Figure 4): it performs a BLAST search of any user provided query sequences against data stored in the CANGS DB MySQL database. This analysis provides the user with an estimate about



BLAST search criteria

Sequence type: Non redundant sequences

Database: 1; 2

Primer Pair ID: All

Sample ID: All

Sample information:

Random sampling: No sampling

Download Blast Output

```
BLASTN 2.2.24+

Reference:
Stephen F. Altschul, Thomas L. Madden, Alejandro A. Schäffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database
search programs", Nucleic Acids Res. 25:3389-3402.


Database: allsequencesdb
           2,790 sequences; 486,987 total letters


Query=  Mondsee_10seqlcount160
Length=173
                                                                  Score     E
Sequences producing significant alignments:                      (Bits)  Value

  FA6NT7002GHVFK|2                                                 313    3e-87
  E9LAHD006DLZZJ|1                                                 313    3e-87
  FBLOEHH02FYO3J|2                                                 311    1e-86
  FBLOEHH02G7UKB|2                                                 311    1e-86
  FBLOEHH02HVH4R|2                                                 309    4e-86
  FBLOEHH02IIE8M|2                                                 307    1e-85
  FA6NT7002I2SD1|2                                                 307    1e-85
  FA6NT7002HTJII|2                                                 307    1e-85
  FA6NT7002GDJQE|2                                                 307    1e-85
  FBLOEHH02IPA3W|2                                                 307    1e-85

>FBLOEHH02FYO3J|2
Length=173

 Score =  311 bits (344),  Expect = 1e-86
 Identities = 172/172 (100%), Gaps = 0/172 (0%)
 Strand=Plus/Plus

Query  1    GAGCCTGAGAGACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTACCCAATCC  60
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  1    GAGCCTGAGAGACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTACCCAATCC  60

Query  61   CGACTCGGGGAGGTAGTGACAATAAATAACAGTAGAGGGCCAATGGTCTTCTTATTGGAA  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  61   CGACTCGGGGAGGTAGTGACAATAAATAACAGTAGAGGGCCAATGGTCTTCTTATTGGAA  120

Query  121  TGAGAACAATTTAAACCCCTTAACGAGGATCAATTAGAGGGCAAGTCTGGTG  172
            ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  121  TGAGAACAATTTAAACCCCTTAACGAGGATCAATTAGAGGGCAAGTCTGGTG  172
```

Figure 4 An example output of the Blast Analysis tool. An example output of the Blast analysis tool in CANGS DB. It shows BLAST hits for the query sequences against the selected data sets stored in CANGS DB.

the abundance of the queried sequences in the databank. Most important, by specifying the queried dataset this tool allows the user to ask specific questions (e.g.: does this sequence occur in samples collected in May?). Additionally, the user can retrieve the abundance of each hit sequence by clicking on it (Figure 5)

(2) **Taxonomy analysis** [7]: this tool classifies the query 454 reads by assessing their similarity to taxonomic entries in the NCBI database (Figure 6). This analysis requires the nucleotide preformatted BLAST database from ftp://ftp.ncbi.nih.gov/blast/db/ to be installed, which is done using the perl program "*update_blastdb.pl*" [14].

(3) **Rarefaction analysis tool** [7]: this tool is used for estimating the species richness in given pooled sequences (Figure 7). CANGS DB also generates a rarefaction plot based on the MOTHUR output (Figure 8).

Editing of sequence information

CANGS DB provides a user-friendly interface to update existing information, add new information or delete information. CANGS DB also allows users to edit the taxonomic path for any stored sequences. CANGS DB does not overwrite the edited taxonomic path; rather it keeps the edited taxonomic path in an additional column. This option is particularly helpful if multiple users work on the same data set, as it permits experts to correct the automated species assignation by CANGS DB and these changes are traceable for all users.

Re-processing stored 454 sequences

CANGS DB provides a unique feature (Figure 9), which enables users to combine raw 454 sequences uploaded into CANGS DB database into a single dataset, which can then be trimmed and filtered. This feature will be especially useful when the same samples are sequenced in different 454 runs/plates.

Deleting dataset

CANGS DB provides option to delete any uploaded data set.

Evaluation of the database

CANGS DB has been designed to provide maximum flexibility for the user. To demonstrate the efficiency of CANGS DB sequence processing pipeline, we processed and analysed 454 sequences previously deposited in the NCBI database [NCBI: SRA008706.2]. This data set consists of 447,909 reads from the 18S rRNA gene obtained from 10 temporal freshwater samples. Applied to our example data set, the CANGS DB sequence-processing pipeline eliminated approximately 37% of all sequences (Table 1), leaving a total of **281,003 (~63%)** sequences for downstream analyses. CANGS DB took 2.5 hours to process this data set using a Macintosh OS X version 10.6.4 with a single processor. If the user skips the removal of singletons then it takes only 20 minutes to process the same data set. When including the CANGS Taxonomic assignment pipeline the total processing time increases to 22 hours.

Download

Sequence information for BLAST hit: FBLOEHH02FYO3J

S.No.	Data set	Sequence ID	Sample ID	Abundance	Sequence	Taxonomic path (CANGS)	Taxonomic path (Corrected)
1	mondsee-pooled-data2	FBLOEHH02H3744	july_end_2007	4	GAGCCTGAGAGACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTACCCAATCCCGACTCGGGGAGGTAGTGACAATAAATAACAGTAGAGGGCCAATGGTCTTCTTATTGGAATGAGAACAATTTAAACCCCTTAACGAGGATCAATTAGAGGGCAAGTCTGGTGA	Eukaryota; environmental samples; environmental samples; Cryptomonadaceae; environmental samples; uncultured.	
2	mondsee-pooled-data2	FBLOEHH02FYO3J	october_2007	2	GAGCCTGAGAGACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTACCCAATCCCGACTCGGGGAGGTAGTGACAATAAATAACAGTAGAGGGCCAATGGTCTTCTTATTGGAATGAGAACAATTTAAACCCCTTAACGAGGATCAATTAGAGGGCAAGTCTGGTGA	Eukaryota; environmental samples; environmental samples; Cryptomonadaceae; environmental samples; uncultured.	
3	mondsee-pooled-data2	FBLOEHH02J1LBK	may_early_2007	1	GAGCCTGAGAGACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCAAATTACCCAATCCCGACTCGGGGAGGTAGTGACAATAAATAACAGTAGAGGGCCAATGGTCTTCTTATTGGAATGAGAACAATTTAAACCCCTTAACGAGGATCAATTAGAGGGCAAGTCTGGTGA	Eukaryota; environmental samples; environmental samples; Cryptomonadaceae; environmental samples; uncultured.	

Figure 5 An example output of the abundance of BLAST Hits. An example output of the abundance of each BLAST hit in the CANGS DB. This feature provides information about the abundance and turnover of species among samples.

Download Taxonomy Output

Taxonomic path assigned by CANGS pipeline

Sequence ID	Sequence	Taxonomic path
Mondsee_10seq1count160	GAGCCTGAGAGACGGCTACCACATCCAAGGAAGGCAGCAG GCGCGCAAATTACCCAATCCCGACTCGGGGAGGTAGTGAC AATAAATAACAGTAGAGGGCCAATGGTCTTCTTATTGGAA TGAGAACAATTTAAACCCCTTAACGAGGATCAATTAGAGG GCAAGTCTGGTGC	Eukaryota; Cryptophyta; Cryptomonadales; Cryptomonadaceae; Cryptomonas.
Mondsee_10seq2count145	GAGCCTGAGAAATGGCTACCACATCCAAGGAAGGCAGCAG GCGCGTAAATTACCCAATCCTGACACAGGGAGGTAGTGAC AATAAATAACAATGCCGGGCTTTTTTAAGTCTGGCACTTG GAATGAGAACAATTTAAATCCCTTATCGAGGATCAATTGG AGGGCAAGTCTGGTGC	Eukaryota; stramenopiles; Chrysophyceae; Chromulinales; Chromulinaceae; Uroglena.
GFKCM5307H5E2D_311495	GCTTTTACCGATTGAGTGTTCAGGTGAGTCTTCTTGATAG TGTCAAAACTAAAATTAAACAAATCTTAGCACTTAGAGGA AAAAGAAGTCGTAACAAGGTTTCC	Eukaryota; Alveolata; Ciliophora; Intramacronucleata; Oligohymenophorea; Peritrichia; Vorticellidae; Vorticella.
GFKCM5307IP41M_298827	GCTCCTACCGATTGAATGGTCCGGTGAGAACGTCGGATCA CGCGCATGTCGCCTTCACGGGCAACTCGTGCGAGAGAAGT TGTTCAAACCTTACCATTTAGAGGAAGGAGAAGTCGTAAC AAGGTTTCC	Eukaryota; environmental samples.
GFKCM5306GV6QW_552016	GCTCCTACCGATTTCGAGTGGTCCGGTGAATCTTTCGGAC TGCAAGTGGTTCCGTGCCACGCGCGGAAAGTCAGGTAAAT CTTATCACTTAGAGGAAGGAGAAGTCGTAACAAGGTTTCC	Eukaryota; Alveolata; Ciliophora; Intramacronucleata; Spirotrichea; Oligotrichia; Strombidiidae; Strombidium.

Figure 6 An example output of the Taxonomy Analysis tool. An example output of the Taxonomy Analysis from CANGS pipeline. It is a tabular output, the columns from the left to the right are 1) query sequence ID 2) query sequence 3) taxonomic path of the closest NCBI sequence.

Rarefaction analysis criteria

Sequence type: Non redundant sequences

Database: 1; 2

Primer Pair ID: All

Sample ID:

Environmental variable:

Random sampling: No sampling

Clustering method: furthest

Sequence frequency to write OTUs: 500

Pairwise distance cutoff: 0.03

Download Rarefaction plot

Download Mothur Output

numsampled	unique	lci	hci	0.01	lci	hci	0.02	lci	hci	0.03	lci	hci
1	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	
500	482.7510		474.9698	490.5322		479.0130		470.6428		487.3832		471.5060
1000	930.8060		916.5153	945.0967		917.2770		902.1210		932.4330		890.7790
1500	1345.2800		1325.2454	1365.3146		1316.3180		1294.5804		1338.0556		1263.3700
2000	1726.5390		1702.4350	1750.6430		1678.2630		1652.9156		1703.6104		1594.0090
2500	2074.1870		2046.7244	2101.6496		2005.7240		1978.4868		2032.9612		1886.6260
3000	2389.4000		2361.0876	2417.7124		2299.0260		2269.1152		2328.9368		2144.2690
3500	2671.6350		2641.9636	2701.3064		2559.2590		2529.0628		2589.4552		2370.0380
4000	2922.2290		2892.2319	2952.2261		2788.6690		2759.1222		2818.2158		2565.5740
4500	3140.7310		3112.0462	3169.4158		2987.5840		2959.8409		3015.3271		2734.2210
5000	3328.2640		3303.0372	3353.4908		3156.9510		3132.5297		3181.3723		2875.8020
5500	3484.6950		3465.0055	3504.3845		3298.5870		3279.4461		3317.7279		2993.0050
6000	3610.1000		3600.2732	3619.9268		3411.6410		3401.4798		3421.8022		3086.2000
6165	3645.0000		3645.0000	3645.0000		3443.0000		3443.0000		3443.0000		3112.0000

Figure 7 An example output of the Rarefaction Analysis tool. An example output of the rarefaction analysis generated by MOTHUR program.

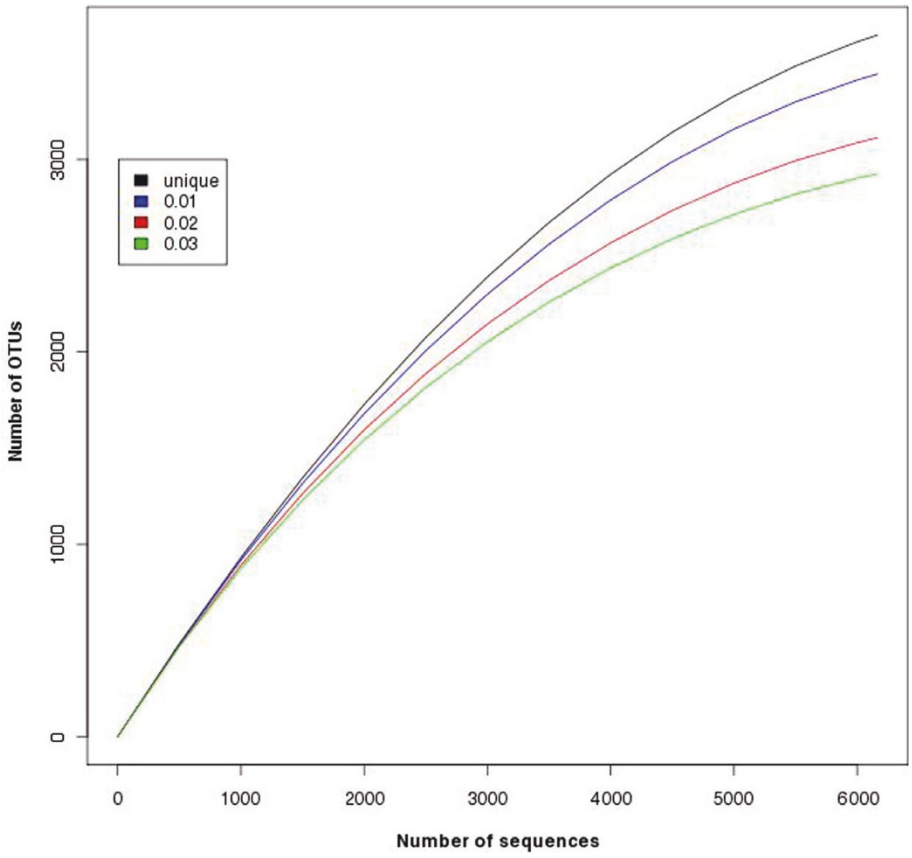


Figure 8 An example output of the Rarefaction plot from Rarefaction Analysis tool. An example output of the rarefaction plot created based on MOTHUR output.

Reproducibility

In order to increase transparency and reproducibility of the results, CANGS DB prints a log file for all processed raw data. This includes the parameter and summary statistics used for each step taken in sequence processing along with discarded sequence identities. Furthermore, stored sequences can be downloaded from CANGS DB according to the user-defined criteria.

Future Directions

The current version of CANGS DB is only compatible for Unix operating systems; however, we plan to make it PC compatible. Additionally we will integrate more downstream analysis tools in the CANGS DB web interface.

Conclusion

CANGS DB is a user-friendly and stand-alone database tool for processing, analyzing and managing the high throughput sequencing data from 454 amplicon resequencing projects. CANGSDB is very easy to use; it could be installed and used on any local UNIX based computer to handle individual as well as multiple sequencing projects in collaboration. It provides full-fledged flexibility with various options in raw sequence processing and analysis. CANGS DB provides a very powerful data retrieval interface, which enables researchers to retrieve information on samples, primers and barcodes from any individual data set or from a combination of data sets. It also provides an interface to update sample information and taxonomic classifications assigned by CANGS taxonomy analysis pipeline as well as delete any data set. The tool can be downloaded at http://code.google.com/p/cangsdb/.

Availability & requirements

Project name: CANGS DB –Cleaning, Analyzing and Managing 454 sequences.

Availability: http://code.google.com/p/cangsdb/

Operating System: Mac OS X, Linux and any other UNIX like system

Programming language: Perl 5.10.0

Other requirements: BioPerl, R, BLAST, MAFFT, MOTHUR, MySQL 5.1, Apache, CGI, DBI.pm, DBD::mysql.pm

License: GNU General Public License.

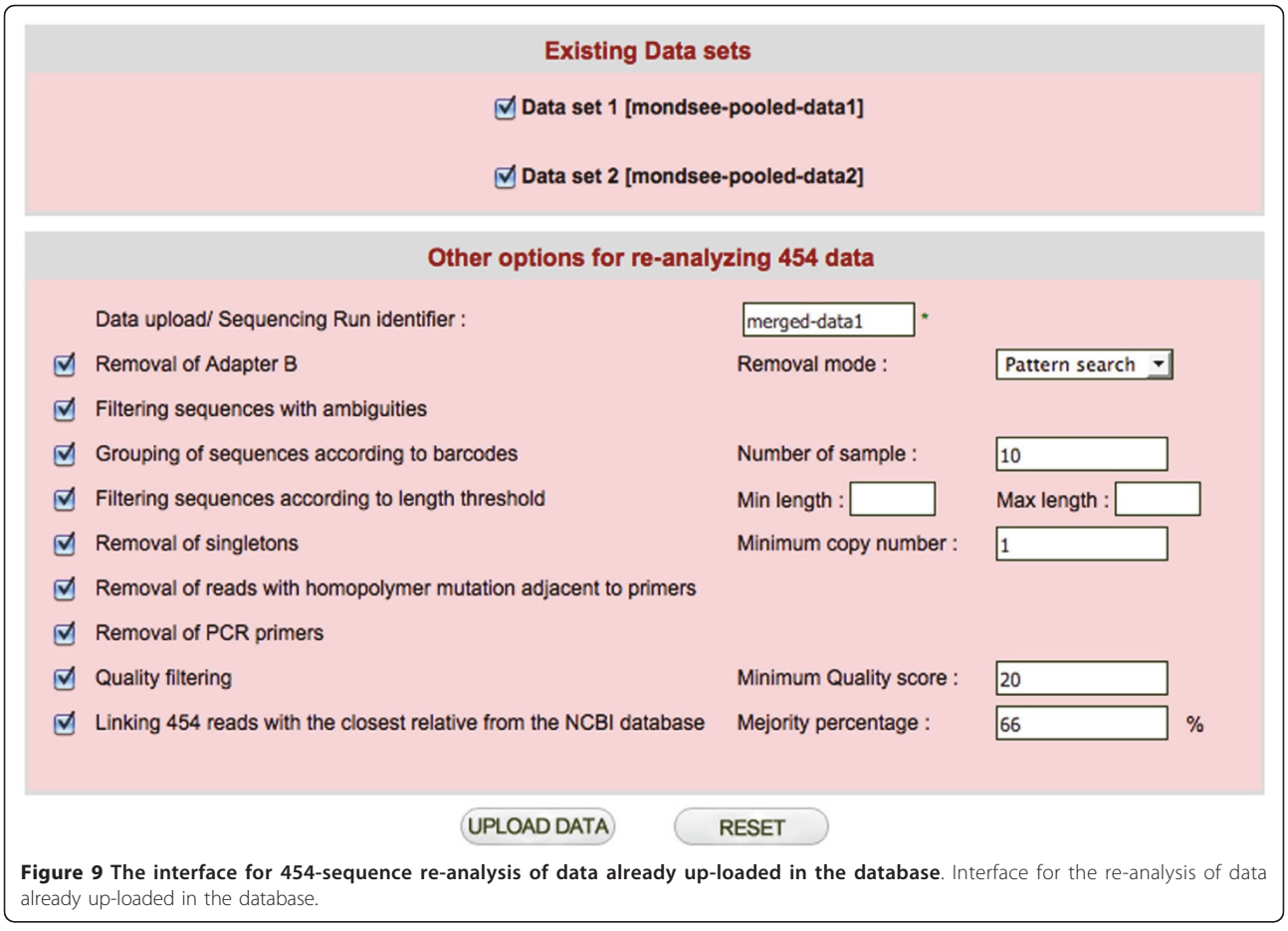


Figure 9 The interface for 454-sequence re-analysis of data already up-loaded in the database. Interface for the re-analysis of data already up-loaded in the database.

Table 1 Number of reads eliminated at different steps of the CANGS DB sequence processing pipeline

Order of steps	Steps	Total no. of sequences	No. of sequences considered	No. of sequences discarded
1	Removal of Adapter B	447,909	373,116	74,793
2	Filtering sequences with ambiguities	373,116	357,926	15,190
3	Removal of singletons	357,926	311,425	46,501
4	Grouping of sequences according to bar codes	311,425	306,042	5,383
5	Filtering sequences according to length threshold	306,042	305,884	158
6	Removal of PCR primers	305,884	282,053	23,831
7	Quality filtering	282,053	281,003	1,050
	Total Sequences	447,909	281,003	166,906

Any restrictions to use by non-academics: license needed.

Acknowledgements
We are thankful to members of the Institut für Populationsgenetik for helpful discussion and special thanks to Pablo Orozco-terWengel and Raymond Tobler for helpful discussion and comments. This work was supported by FWF grants (No. P19467-B11) to CS.

Author details
[1]Institut für Populationsgenetik, Veterinärmedizinische Universität Wien, Veterinärplatz 1, Vienna, Austria. [2]Allgemeine Botanik, Universität Duisburg-Essen, D-45117, Essen Germany.

Authors' contributions

JB and CS designed the study. RVP analyzed and wrote the code. RVP designed and developed the database and web interface. RVP wrote the draft of the manuscript and VN, CS, JB and RVP revised it. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Received: 13 December 2010 Accepted: 30 June 2011
Published: 30 June 2011

References

1. Thomas RK, Nickerson E, Simons JF, Jänne PA, Tengs T, Yuza Y, Garraway LA, LaFramboise T, Lee JC, Shah K, *et al*: **Sensitive mutation detection in heterogeneous cancer specimens by massively parallel picoliter reactor sequencing.** *Nature medicine* 2006, **12(7)**:852-855.
2. Huber JA, Mark Welch DB, Morrison HG, Huse SM, Neal PR, Butterfield DA, Sogin ML: **Microbial population structures in the deep marine biosphere.** *Science* 2007, **318(5847)**:97-100.
3. Nolte V, Pandey RV, Jost S, Medinger R, Ottenwälder B, Boenigk J, Schlötterer C: **Contrasting seasonal niche separation between rare and abundant taxa conceals the extent of protist diversity.** *Mol Ecol* 2010, **19(14)**:2908-2915.
4. Medinger R, Nolte V, Pandey RV, Jost S, Ottenwälder B, Schlötterer C, Boenigk J: **Diversity in a hidden world: potential and limitation of next-generation sequencing for surveys of molecular diversity of eukaryotic microorganisms.** *Mol Ecol* 2010, **19(Suppl. 1)**:32-40.
5. **RDP (Ribosonmal Database Project).** [http://pyro.cme.msu.edu/].
6. Schloss PD, Handelsman J: **Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness.** *Applied and Environmental Microbiology* 2005, **71(3)**:1501-1506.
7. Pandey RV, Nolte V, Schlötterer C: **CANGS: a user-friendly utility for processing and analyzing 454 GS-FLX data in biodiversity studies.** *BMC Res Notes* 2010, **11;3**:3.
8. **VAMPS: Visualization and Analysis of Microbial Population Structures.** [http://vamps.mbl.edu/index.php].
9. Giongo A, Crabb DB, Davis-Richardson AG, Chauliac D, Mobberley JM, Gano KA, Mukherjee N, Casella G, Roesch LF, Walts B, Riva A, King G, Triplett EW: **PANGEA: pipeline for analysis of next generation amplicons.** *ISME J* 2010, **4(7)**:852-861.
10. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JG, Korf I, Lapp H, *et al*: **The Bioperl toolkit: Perl modules for the life sciences.** *Genome Research* 2002, **12(10)**:1611-1618.
11. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *Journal of Molecular Biology* 1990, **215(3)**:403-410.
12. Katoh K, Kuma K, Toh H, Miyata T: **MAFFT version 5: improvement in accuracy of multiple sequence alignment.** *Nucleic Acids Research* 2005, **33(2)**:511-518.
13. **MySQL.** [http://www.mysql.com].
14. **Update_blastdb.pl.** [http://www.ncbi.nlm.nih.gov/BLAST/docs/update_blastdb.pl].
15. **R.** [http://cran.r-project.org/].

doi:10.1186/1756-0500-4-227
Cite this article as: Pandey *et al.*: CANGS DB: a stand-alone web-based database tool for processing, managing and analyzing 454 data in biodiversity studies. *BMC Research Notes* 2011 **4**:227.

Submit your next manuscript to BioMed Central and take full advantage of:

- **Convenient online submission**
- **Thorough peer review**
- **No space constraints or color figure charges**
- **Immediate publication on acceptance**
- **Inclusion in PubMed, CAS, Scopus and Google Scholar**
- **Research which is freely available for redistribution**

Submit your manuscript at
www.biomedcentral.com/submit

BioMed Central