

Community structure measures for meta-genomics

István Albert

Bioinformatics Consulting Center
Penn State

BioStar: <http://www.biostars.org>

A screenshot of a web browser window showing the BioStar website. The title bar reads "A question and answer site for bioinformatics". The address bar shows the URL "http://biostar.stackexchange.com/". The page header includes a user profile for "Istvan Albert" with a badge count of 6,487, and navigation links for "Questions", "Tags", "Users", "Badges", "Unanswered", and "Ask Question". A sidebar on the right indicates "We're in Bootstrap Mode" with a message about relaxed reputation requirements. It also lists "All users can:" followed by a bulleted list of actions: Post Questions, Post Answers, Comment, Create Tags, Re-Tag Questions, and Vote. Below this is a link "Learn how you can help ». The main content area displays a list of recent questions:

- Bioinformatics Project Idea** (1 vote, 1 answer, 109 views) - posted 2h ago by Aleksandr Levchuk with 1,501 reputation.
- varscan somatic...syntax to run with redirect?** (0 votes, 0 answers, 11 views) - posted 3h ago by maziz with 87 reputation.
- Where can I find a list of human TFs?** (1 vote, 1 answer, 20 views) - posted 4h ago by avilella with 1,381 reputation.
- Mapping BIGG metabolic reaction IDs to KEGG pathway map IDs** (0 votes, 0 answers, 12 views) - posted 6h ago by bernie with 1 reputation.
- What are the most common stupid mistakes in bioinformatics?** (14 votes, 26 answers, 1k views) - posted 14h ago by ShellfishGene with 1 reputation.
- normalization of transcriptome 454** (1 vote, 0 answers, 13 views) - posted 15h ago by bdv with 197 reputation.

On the left side of the browser window, there is a vertical toolbar with icons for Gmail, Google, GitHub, and Bookmarks.

Question and Answer site for Bioinformatics

A question and answer site for bioinformatics

Gmail - installation sc... Welcome! — Bioinform... Peak Predictions and ... A question and answer... Google Docs - Home Your Dashboard - Git... + http://biostar.stackexchange.com/ Gmail Google GH NGS BioStar LionDB BCC Edda Bookmarks

BioStar

Questions Tags Users Badges Unanswered Ask Question

Recent Questions

active 0 featured hot week month

		votes	answers	views	Tags	Asked	By	Score
3	4	72	72	72	visualization rna-seq alternative-splicing	8m ago	Mark Rogers	1
14	28	1k	1k	1k	mistakes software-engineering	11m ago	GWW	770 ● 1 ● 7
2	1	34	34	34	rnaseq insert size selection	1h ago	bioquant	111 ● 3
1	1	26	26	26	bwa paired-end solid	3h ago	brentp	6,005 ● 5 ● 24
3	1	34	34	34	functional divergence	4h ago	avilella	1,466 ● 4 ● 10
0	2	29	29	29	varscan redirect samtools somatic	4h ago	Sean Davis	1,055 ● 2 ● 7
...
...

We're in Bootstrap Mode

Reputation requirements are relaxed while we grow the site.

All users can:

- Post Questions
- Post Answers
- Comment
- Create Tags
- Re-Tag Questions
- Vote

[Learn how you can help »](#)

Interesting Tags

Add

Ignored Tags

Add

Spread the word!

Holistic data analysis

- Try to lay out all steps first – worry about parameter settings later – knowing what the data looks like
- Too much planning is actually detrimental
- So is fretting about the details
- Start with the end result → I'd like to have my data laid out as a ...



JUST DO IT.



Topic: metagenomics

- the majority of microbial biodiversity cannot be captured by cultivation based methods
- metagenomics = the study of genetic materials recovered from environmental samples

Increased complexity

- We're interested in the bacterial species (memberships) present in the colonies
- Also interested in the relative abundances between these species
- We need to detect changes → we need to compare two bacterial communities

16S ribosomal protein

- highly conserved between different species of bacteria and archaea
- whereas the rest of genetic content varies greatly across species
- 16S RNA can be used for taxonomical classification

Two widely used approaches

Classifying sequences:

- by their similarity to reference sequences (phlyotyping)
- by their similarity to other sequences in the sample
(operations taxonomic units → OTU)

Taxonomy

- Placing a bacteria into a taxonomy is difficult
- Several competing groups – each maintain a separate taxonomical database
- Three widely used curated taxonomy outlines that contain significant conflicts with each other

Greengenes

Firefox

Gmail - Inbox - istvan.al... Analysis Report greengenes.lbl.gov - Ali...

http://greengenes.lbl.gov/cgi-bin/nph-index.cgi

Google

Gmail Google iGoogle Jen RSeek LionDB Silva Local BCC Th Calendar Mend Biostar Unfuddle W Bookmarks

green genes 16S rRNA gene database and workbench compatible with ARB greengenes.lbl.gov

Browse Export Slice Consensus Compare Search Probes Align Trim Download More Tools

Functions

- Home
- Browse
- Export
- Slice
- Consensus
- Compare
- Search
- Probe
- Align
- Trim
- Download
- Curate
- More Tools...

About

- Citation
- Tutorial
- FAQ
- Objectives
- Methods
- Contact

My Interest List

- remove all
- collapse all
- show marked

My Taxonomy

- greengenes
- Activate
- Changing taxonomy will empty My Interest List.

greengenes: 16S rDNA data and tools

The greengenes web application provides access to the current and comprehensive 16S rRNA gene sequence alignment for browsing, blasting, probing, and downloading. The data and tools presented by greengenes can assist the researcher in choosing phylogenetically specific probes, interpreting microarray results, and aligning/annotating novel sequences. If you are an ARB user, you can use greengenes to keep your own local database current.

News:

- Looking for Hugenholtz or PHPR taxonomy? It is now the greengenes taxonomy.
- Dr. Mike Dyall-Smith has graciously made available his tutorial for [installing ARB on Mac OSX](#). Thanks Mike.
- The greengenes taxonomy for the Cyanobacteria is now consistent with [cyanoDB](#) using cyanoDB type species as a guide to map cyanoDB taxonomy to the greengenes reference 16S tree.
- Thanks to Greg Caporaso and Rob Knight for providing [OTU reference and utility files](#) for use with QIIME software.
- The Wall Street Journal picks the Berkeley PhyloChip as the top advance in environmental technology of 2008 and 3rd best innovation overall.
- Pollution Engineering Magazine selects Berkeley PhyloChip as most likely to aid pollution control and abatement in the near future.
- The Berkeley PhyloChip wins R&D100 award as one of the 100 most significant technological advances of the year.
- Are you the world expert on the taxonomy of a particular phylogenetic lineage? Have you checked this database and nobody has got it right? [Tell us!](#) - we will fix it. We thank Jakob Fredslund for developing a tool, [Gexcellent](#), to convert XML trees to Newick format!
- We thank J.P. Euzéby and Hans Truper for expert [etymological advice](#).

Browse taxonomic tree of your choice and mark nodes.

Export sequence records of your choice.

Specify a Slice (sub-alignment) of the prokMSA to view/download.

Calculate Consensus sequences from My Interest List (soon!).

Compare my local sequences/probes against the prokMSA using BLAST or Simeank

<http://greengenes.lbl.gov/>

SILVA

Firefox ▾

Gmail - Inbox - istvan.al... Analysis Report greengenes.lbl.gov - Ali... Silva

http://www.arb-silva.de/ Google

Gmail Google iGoogle Jen RSeek LionDB Silva Local BCC Th Calendar Mend Biostar Unfuddle W Bookmarks

silva comprehensive ribosomal RNA databases

TUM

Home Browser Search Aligner Download Documentation Projects FISH & Probes Shop Jobs Contact

SILVA

Welcome to the SILVA rRNA database project
A comprehensive on-line resource for quality checked and aligned ribosomal RNA sequence data, free for academic use.

SILVA provides comprehensive, quality checked and regularly updated datasets of aligned small (16S/18S, SSU) and large subunit (23S/28S, LSU) ribosomal RNA (rRNA) sequences for all three domains of life (*Bacteria*, *Archaea* and *Eukarya*).

SILVA are the official databases of the software package ARB.
For more background information → [Click here](#)

ARB
The software package ARB represents a graphically-oriented, fully-integrated package of cooperating software tools for handling and analysis of sequence information.

The ARB project has been started more than 15 years ago by Wolfgang Ludwig at the Technical University in Munich, Germany, see [www.arb-home.de](#).

The MEGX.net data portal
Visit our partner site [www.megx.net](#), the data portal for Marine Ecological Genomics, to get a feeling how your research can be improved using integrated databases.

Citation

News

Latest News

10.04.11
SILVA 106 released
The data deluge has delayed this release. Now offering over 2.2 MILLION aligned SSU & LSU sequences! [\[more\]](#)

08.04.11
Genomic Standards Consortium: bringing standards to life
The Genomic Standards consortium has published a paper in the ISME Journal to emphasise that the adoption of easy-to-follow standards will vastly improve our ability to interpret data from genomes, metagenomes and marker studies. [\[more\]](#)

22.03.11
Living Tree 104 released
A brand new version (104) of the "All Species Living Tree" has been released. Check what's new on the project website... [\[more\]](#)

[go to Archive ->](#)

Release information for full release - SILVA 106

	SSU Parc	SSU Ref	LSU Parc	LSU Ref
Minimal length	300	1200/900	300	1900
Quality filtering	basic	strong	basic	strong
Guide Tree	no	yes	no	yes
Release version	106	106	106	106

<http://www.arb-silva.de/>

RDP

Firefox ▾

Gmail - Inbox - istvan.al... Analysis Report greengenes.lbl.gov - Ali... Ribosomal Database Pr... +

http://rdp.cme.msu.edu/ Google Bookmarks

Gmail Google iGoogle Jen RSeek LionDB Silva Local BCC Th Calendar Mend Biostar Unfuddle W Bookmarks

RDP HOME | ABOUT | ANNOUNCEMENTS | CITATION | CONTACTS | RESOURCES | RELATED SITES | TUTORIALS

RIBOSOMAL DATABASE PROJECT

BROWSERS | CLASSIFIER | LIBCOMPARE | SEQMATCH | PROBE MATCH | TREE BUILDER | PYRO | TAXOMATIC | SEQCART | ASSIGNGEN

RDP Release 10, Update 26 :: Mar 28, 2011 :: 1,613,063 16S rRNAs

The Ribosomal Database Project (RDP) provides ribosome related data and services to the scientific community, including online data analysis and aligned and annotated Bacterial and Archaeal small-subunit 16S rRNA sequences.

Cite RDP's NAR article ↗

RDP Release 10 brings two major changes to the RDP:

- RDP10 provides new **Bacterial** and **Archaeal** alignments with several significant enhancements over the previous RDP 9 alignments.
- Use of the *Infernal* secondary-structure based aligner that provides better support for short partial sequences and handles certain sequencing artifacts in a more intuitive manner.

Explore our online analysis tools:

- myRDP
- BROWSERS
- CLASSIFIER
- LIB COMPARE
- SEQ MATCH
- PROBE MATCH
- TREE BUILDER
- PYROSEQUENCING PIPELINE
- ASSIGNMENT GENERATOR
- TAXOMATIC

HOVER over any tool item in the menu to see a brief popup description of its features;

CLICK on the tool menu item to begin working with it.

Be sure to view the video tutorials and visit each tool's help file to use our site to your fullest advantage.

Sponsors:

National Science Foundation Office of Biological and Environmental Research National Institutes of Health

myRDP login

RDP News

04/07/2011 Latest MIMARKS (formerly MIENS) standards released 'The Genomic Standards Consortium: Bringing Standards to Life for Microbial Ecology' is now published online in ISMEJ

04/04/2011 RDP MIMARKS GoogleSheet RDP developed the MIMARKS GoogleSheet to help you manage the metadata for your samples.

03/28/2011 RDP 10, update 26 released Release 10.26 contains 1,613,063 aligned 16S rRNA sequences.

02/17/2011 RDP 10, update 25 released Release 10.25 contains 1,545,680 aligned 16S rRNA sequences.

02/16/2011 NCBI has announced phasing out of the SRA Sequence Read Archive being phased out by NCBI due to budget constraints

01/13/2011 RDP 10, update 24 released

<http://rdp.cme.msu.edu/>

NCBI taxonomy

Firefox ▾

Gmail - Inbox - istvan.al... Analysis Report greengenes.lbl.gov - Ali... NCBI Taxonomy Home... +

http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/inde... Google

Gmail Google iGoogle Jen RSeek LionDB Silva Local BCC Th Calendar Mend Bookmarks

NCBI Taxonomy Browser

PubMed Entrez BLAST OMIM Taxonomy Structure

Search for As complete name lock Go Clear

Taxonomy browser
Taxonomy common tree
Taxonomy information
Taxonomy resources
Taxonomic advisors
Genetic codes
Taxonomy Statistics
Taxonomy Name/Id Status Report

How to reference the NCBI taxonomy database

The NCBI taxonomy database is *not a primary source* for taxonomic or phylogenetic information. Furthermore, the database does not follow a single taxonomic treatise but rather attempts to incorporate phylogenetic and taxonomic knowledge from a variety of sources, including the published literature, web-based databases, and the advice of sequence submitters and outside taxonomy experts. Consequently, the NCBI taxonomy database is not a phylogenetic or taxonomic authority and should not be cited as such.

However, if you want to acknowledge the fact that you have used the NCBI taxonomy database and other NCBI services and databases for your research, the two following papers can be referenced:

One common phlyotyping workflow

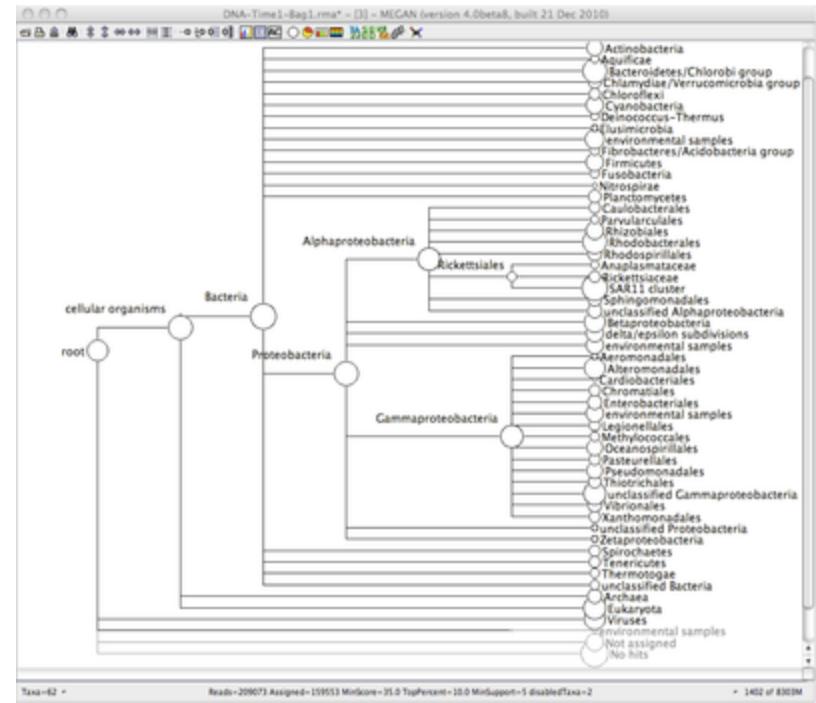
- Run the blast aligners with the reads against the NCBI bacterial database (can be very time consuming)
- Use MEGAN – Metagenome Analyzer to process the results

For 16S RNA there is faster way

- Align against a hand curated, prealigned representative selection (NAST algorithm).
- This needs far fewer resources

MEGAN

- <http://ab.inf.uni-tuebingen.de/software/megan/>
- Graphical user interface – nice visualizations



Pitfalls in phylootyping approaches

The advantage of a phylotype based methods is that it places a label onto each sequence

- Yet the same species may have very different phenotypes
- Same phenotypes may actually belong to different lineages
- Nonetheless overall it works well for taxonomical classification

OTU based approaches

- Clustering based – sequences are clustered by their similarity (must be conserved regions 16S RNA)
- We (YOU) choose a percent similarity level that can range from 0 → 100% at which to merge sequences

Pitfalls of OTU based approaches

- No consistent method for converting between the thresholds used to define OTUs and taxonomic levels
- The distances within a taxonomic group are not evenly distributed
- Clustering is computationally intensive

The methods are slowly merging

- Phylotyping software get more OTU based functionality
- OTU based software get more phylotyping functionality

The **mothur** package

- Primarily OTU based but it has phlyotyping functionality built in

<http://www.mothur.org/>

Exceedingly well documented
with binaries for every platform

download it for your computer

Comparing bacterial communities: A and B

We always have an incomplete sample of a large community

- What is the overlap between A and B
- Is B a subset of A?
- If membership of A and B are identical are their abundances the same?
- What if A was sampled at a higher ratio than B

OTU based calculators for single communities

Each calculator gives us a small window into one particular property of the dataset

- Community richness
- Community evenness
- Community diversity
- OTU number extrapolation

OTU based calculators for multiple communities

- Shared community richness
- Similarity in community membership
- Similarity in community structure

Community richness – alpha diversity

- Chao estimator

Based on what we see
how many microbes are really there

$$S_{chao1} = S_{obs} + \frac{n_1(n_1 - 1)}{2(n_2 + 1)}$$

- Many other estimators: ACE, jackknife etc

Community diversity – dominance

- Berger-Parker index

largest abundance / total number of individuals

$$d = \frac{n_{max}}{N}$$

- Many other estimators: Simpson, Shannon etc

Firefox

Gmail - Life Sciences Di... Analysis Reports — Bio... NCBI Taxonomy Home... mothur Welcome! — Bioinfor... +

http://bcc.bx.psu.edu/reports.html#demo-reports

Gmail Google iGoogle Jen RSeek LionDB Silva Local BCC Th Calendar Mend Biostar Unfuddle W Bookmarks

Analysis Examples

The Analysis Examples are freely accessible and demonstrate the summaries and plots produced by our pipeline implemented with the [Mothur Software](#) and other tools.

- [DSGR run 1, stool sample](#) sequenced on February 4, 2011
- [DSGR run 1, soil sample](#) sequenced on February 4, 2011
- [Microbial diversity in the deep sea](#) based on the [sogin data analysis](#).
- [Bacterial community variation in human body](#) based on the [costello data analysis](#).

DSGR

- [DSGR run 1, stool sample](#) sequenced on February 4, 2011
- [DSGR run 1, soil sample](#) sequenced on February 4, 2011

Brantley Lab

- [Bisley Watershed at All Depths](#) requested by Laura Liermann, sequenced on Nov 11, 2010; 1 file, 7 barcoded samples
- [Sverefjell volcano in Svalbard](#) requested by Laura Liermann, sequenced on Nov 11, 2010; 1 file, 5 barcoded samples

Harwill Lab

- [COPD datafile](#) requested by Laura Weyrich, sequenced on Feb 7, 2011; 1 file, 6 barcoded samples

Baums Lab

- [French Reef Corals](#) requested by Jennifer Boulay, sequenced on Sept 2, 2010; 1 file, 7 barcoded samples

Cantorna Lab

- [Mutant study – CANTIHO](#) requested by Jot Hui Ooi, 1 file, 4 barcoded samples
- [Diet study – CANTIV](#) requested by Jot Hui Ooi, 1 file, 4 barcoded samples

Pugh Lab

- [CTCF Binding Domains](#) requested by Kashturi Kannan, sequenced on Nov 1, 2010;
- [Reb 1 Binding Domains](#) requested by Ho Sung Rhee, sequenced on Feb 5, 2009;

Reese Lab

- [mRNA Alignment](#) requested by Jason Miller, sequenced on Dec 4, 2010; 8 barcoded samples

Quick search

Enter search terms or a module, class or function name.

Go

index

- <http://bcc.bx.psu.edu/> → Analysis Examples

Running through an example

Requirements:

- Use the datasets in **day7/meta**
- The mothur software is installed in:
day7/meta/mothur

Run it as: **./mothur**

- If you are having trouble accessing it copy the **mothur** executable next to the data

Example dataset

From the paper:

"Microbial diversity in the deep sea and the underexplored 'rare biosphere'", **PNAS, 2006**

The first to use pyrosequencing technology to sequence 16S rRNA gene tags.

mothur command line (a bit like R)

~/docs/web/bcc-web/source/course/ppt/week5

mothur v.1.12.3
Last updated: 8/5/2010

by
Patrick D. schloss

Department of Microbiology & Immunology
University of Michigan
pschloss@umich.edu
<http://www.mothur.org>

When using, please cite:

Schloss, P.D., et al., Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl Environ Microbiol, 2009. 75(23):7537-41.

Distributed under the GNU General Public License

Type 'help()' for information on the commands that are available

Type 'quit()' to exit program

C:\cygwin\home\bin\mothur.exe

mothur >

You can type commands into **mothur** even better create a text file with commands and run those commands with **mothur**

Start Page commands.txt X

```
1
2 # redirect reporting into the same file
3 set logfile(name=report.txt)
4
5
6
7
8
9
10
```

Command Output

Ready



CP1252

Ln: 6 Col: 1

Python



All commands generate a log that you can look at later.
This command redirects the log in to a known file. Note the command structure.

commands.txt X

report.txt

```
1
2 # redirect reporting into the same file
3 set.logfile(name=report.txt)
4
5 # generate a summary on the fasta file
6 summary.seqs(fasta=sogin.fasta)
7
8
9
10
11
12
13
14
15
16
17
18
19
```

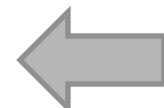
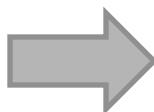
report.txt (Z:\src\public\work) - Komodo Edit 5.2

File Edit Code View Project Toolbox Tools Window Help

commands.txt report.txt X

```
28
29 mothur > setlogfile(name=report.txt)
30
31 mothur > summary.seqs(fasta=sogin.fasta)
32
33          Start    End    NBases  Ambigs  Polymer
34 Minimum:      1      50      50       0       2
35 2.5%-tile:   1      57      57       0       3
36 25%-tile:    1      60      60       0       3
37 Median:      1      61      61       0       3
38 75%-tile:    1      64      64       0       4
39 97.5%-tile:  1      71      71       0       5
40 Maximum:     1     100     100       0      31
41 # of Seqs:   222291
42
43 Output File Name:
44 sogin.fasta.summary
45
46
```

Ready CP1252 Ln: 1 Col: 1 Text



The way **mothur** works

1. Commands generate a readable output to the screen
2. The same output goes to the log file
3. Each command usually creates one or more result files that we can either load in later stages or plot (analyze)

Quality filtering

- We skip this step since the data is published and is quality filtered (and the authors did not provide the original sequence quality data)
- See the **trim.seq** command

File Edit Code View Project Toolbox Tools Window Help

commands2.txt

report.txt

commands1.txt X

```
1
2 # redirect reporting into the same file
3 set logfile(name=report.txt)
4
5 # generate a summary on the fasta file
6 summary.seqs(fasta=sogin.fasta)
7
8 # keep only unique sequences
9 unique.seqs(fasta=sogin.fasta)
10
11 # check the summary for the unique sequences
12 summary.seqs(fasta=sogin.unique.fasta)
13
14
15
16
17
18
19
20
```

Ready



CP1252

Ln: 20 Col: 1

Python

Work on unique sequences - mothur will keep track how many times a sequence was seen

commands2.txt

report.txt X

commands1.txt *

```
46
47 mothur > unique.seqs(fasta=sogin.fasta)
48
49 Output File Names:
50 sogin.unique.fasta
51 sogin.names
52
53
54 mothur > summary.seqs(fasta=sogin.unique.fasta)
55
56          Start    End    NBases  Ambigs  Polymer
57 Minimum:      1      50      50       0        2
58 2.5%-tile:   1      57      57       0        3
59 25%-tile:    1      60      60       0        3
60 Median:      1      62      62       0        4
61 75%-tile:    1      65      65       0        4
62 97.5%-tile: 1      73      73       0        6
63 Maximum:     1     100     100       0       31
64 # of Seqs:   21907
65
```



compare to the value before

We skipping a few steps that are not suited for in class exercise

- Running the alignments, distances and clustering
- The following slides show the commands but we should not do them in class as it may take a while.
- The **dist** folder contains this precomputed dataset
- Check the content of the **commands1.txt** and **commands2.txt** to see all the commands
- We also renamed **sogin.unique.filter.fasta --> good.fasta**

For this you need the silva reference files → 800MB

This can take a bit more - depending on your computer and number of processors - with a very fast computer with 12 processors it took about five minutes

report.txt (Z:\src\public\work) - Komodo Edit 5.2

File Edit Code View Project Toolbox Tools Window Help

commands2.txt report.txt commands1.txt

```
297 1823
298 1765
299 1841
300 Some of you sequences generated alignments that eliminated to
301 It took 53 secs to align 21907 sequences.
302
303
304 Output File Names:
305 sugin.unique.align
306 sugin.unique.align.report
307 sugin.unique.flip.accnos
308
309
310 mothur > filter.seqs(fasta=sugin.unique.align, vertical=T)
311 Creating Filter...
312 100
313 200
314 300
315 400
316 500
```

Ready CP1252 Ln: 49 Col: 20 Text

commands1.txt (Z:\src\public\work) - Komodo Edit 5.2

File Edit Code View Project Toolbox Tools Window Help

commands2.txt report.txt commands1.txt X

```
22 # rename sogen.unique.filter.fasta to good.fasta
23 system(cp sogen.unique.filter.fasta good.fasta)
24
25 # compute the distance matrix
26 dist.seqs(fasta=good.fasta, cutoff=0.10, processors=12)
27
28 # read the distance matrix and cluster the results
29 read.dist(column=good.dist, name=sogen.names)
30
31 # run the clustering method
32 cluster()
33
34
35
36
37
38
39
40
41
```

Ready CP1252 Ln: 24 Col: 1 Python

As a result we compute the similarity matrix that interrelates all sequences

commands2.txt

report.txt X

commands1.txt

```
1006
1007 It took 140 to calculate the distances for 21907 sequences.
1008
1009 mothur > read.dist(column=good.dist, name=sogin.names)
1010 ****#*****#*****#*****#*****#*****#*****#*****#*****#
1011 Reading matrix:   | | | | | | | | | | | | | | | | | | | | | |
1012 ****#*****#*****#*****#*****#*****#*****#*****#*****#
1013 It took 1 secs to read
1014
1015 mothur > cluster()
1016
1017 Output File Names:
1018 good.fn.sabund
1019 good.fn.rabund
1020 good.fn.list
1021
1022 It took 4 seconds to cluster
1023
1024 mothur > quit()
1025
```

Ready



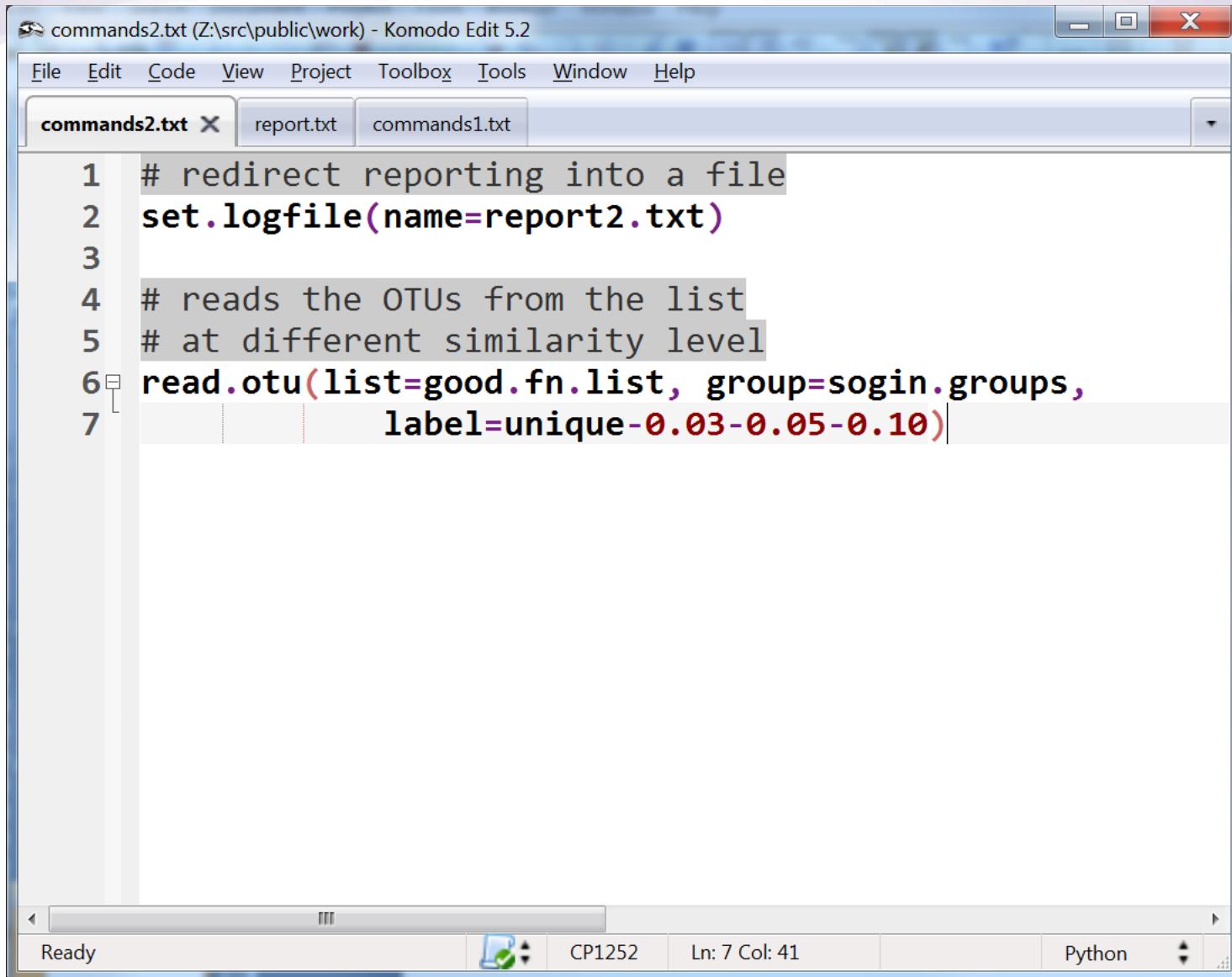
CP1252

Ln: 49 Col: 20

Text

We can now head onto doing analysis

Rarefaction curves



The screenshot shows a window titled "commands2.txt (Z:\src\public\work) - Komodo Edit 5.2". The menu bar includes File, Edit, Code, View, Project, Toolbox, Tools, Window, and Help. The tab bar shows three tabs: "commands2.txt" (active), "report.txt", and "commands1.txt". The code editor displays the following Python script:

```
1 # redirect reporting into a file
2 set.logfile(name=report2.txt)
3
4 # reads the OTUs from the list
5 # at different similarity level
6 read.otu(list=good.fn.list, group=sogin.groups,
7           label=unique-0.03-0.05-0.10)
```

The status bar at the bottom indicates "Ready", "CP1252", "Ln: 7 Col: 41", and "Python".

report2.txt (Z:\src\public\work) - Komodo Edit 5.2

File Edit Code View Project Toolbox Tools Window Help

commands2.txt report.txt commands1.txt report2.txt X

```
30
31 mothur > read.otu(list=good.fn.list, group=sogin.groups, labe
32 unique
33 0.03
34 0.05
35 0.10
36
37 Output File Names:
38 good.fn.112R.rabund
39 good.fn.115R.rabund
40 good.fn.53R.rabund
41 good.fn.55R.rabund
42 good.fn.137.rabund
43 good.fn.138.rabund
44 good.fn.FS312.rabund
45 good.fn.FS396.rabund
46 good.fn.shared
47
```

A separate
outputfile for each
group

File Edit Code View Project Toolbox Tools Window Help

commands2.txt X report.txt commands1.txt report2.txt

```
1 # redirect reporting into a file
2 set.logfile(name=report2.txt)
3
4 # reads the OTUs from the list
5 # at different similarity level
6 read.otu(list=good.fn.list, group=sogin.groups,label=unique-0.03-0.05-0.10)
7
8 # compute the rarefaction curves
9 rarefaction.single(calc=chao, label=unique-0.03-0.05-0.10, freq=0.10)
```

Ready



CP1252

Ln: 9 Col: 31

Python

Generate the rarefaction curves

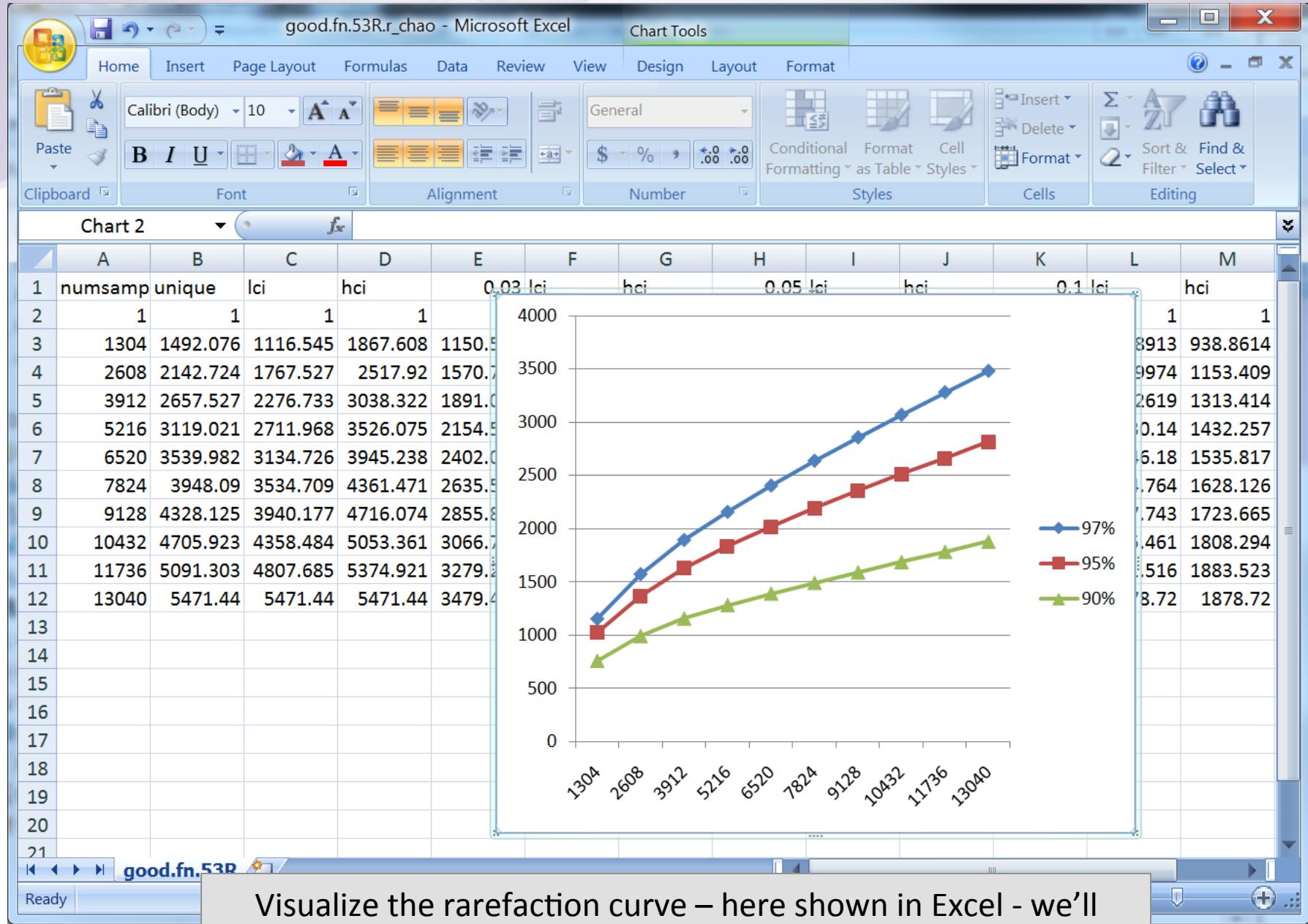
report2.txt (Z:\src\public\work) - Komodo Edit 5.2

File Edit Code View Project Toolbox Tools Window Help

commands2.txt report.txt commands1.txt report2.txt

```
48
49 mothur > rarefaction.single(calc=chao, label=unique-0.03-0.05-0.1
50
51 Processing group 112R
52
53 unique
54 0.03
55 0.05
56 0.10
57
58 Processing group 115R
59
60 unique
61 0.03
62 0.05
63 0.10
64
65 Processing group 137
```

A different file for each group → file extension **r_chao**



Visualize the rarefaction curve – here shown in Excel - we'll learn how to do this in R