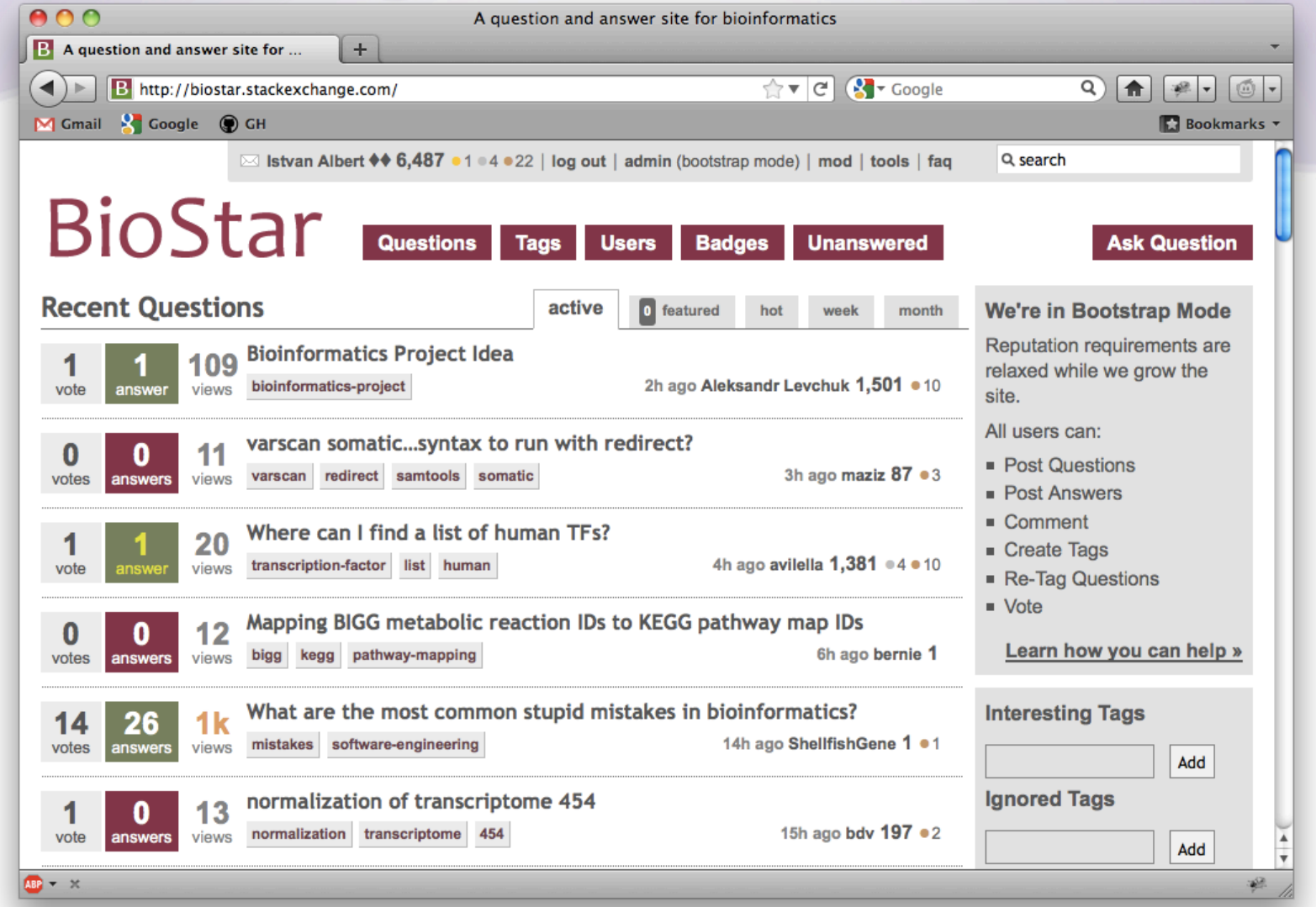
Community structure measures for meta-genomics

István Albert

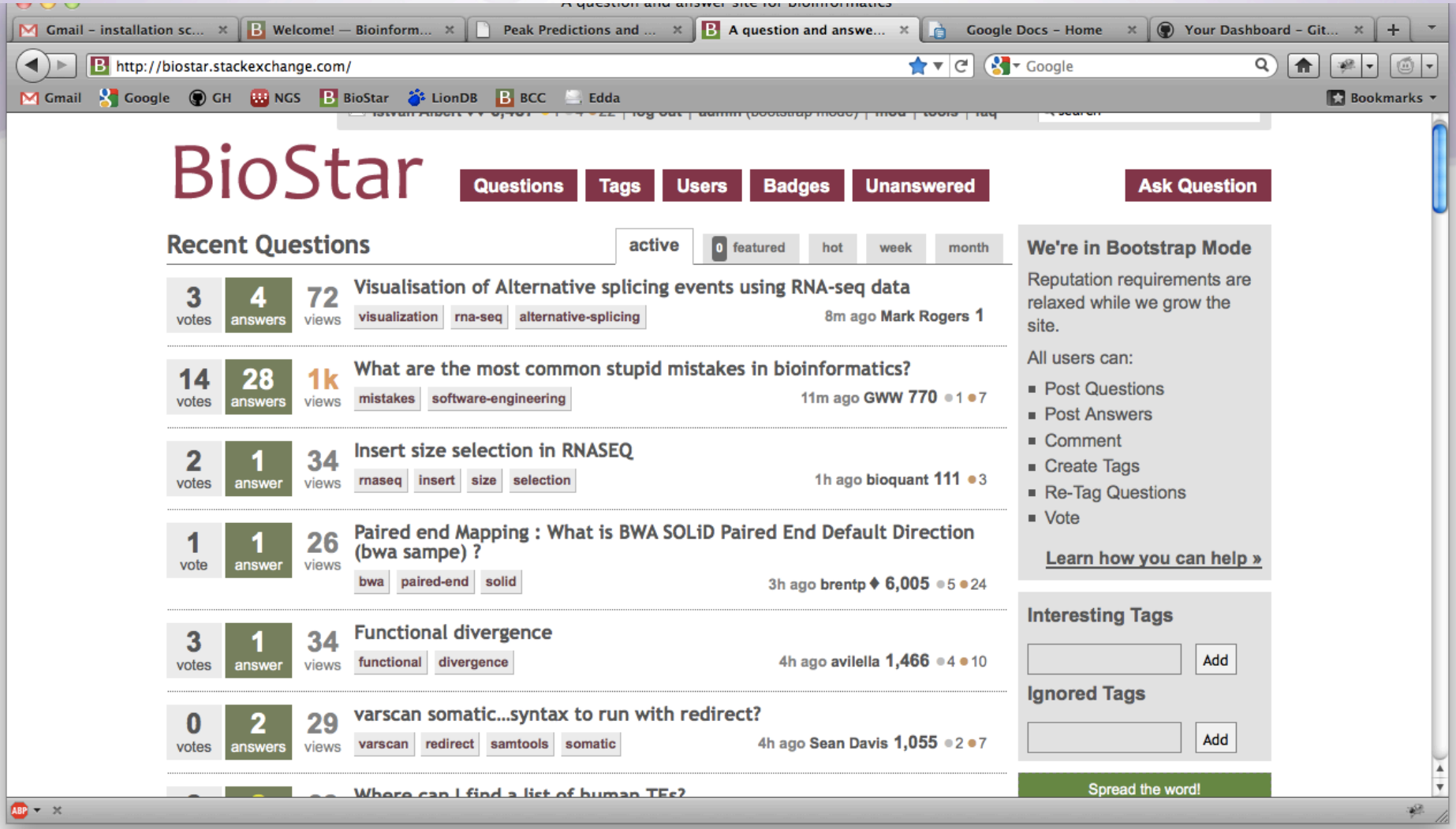
Bioinformatics Consulting Center

Penn State

BioStar: http://www.biostars.org



Question and Answer site for Bioinformatics


Gmail – installation sc...
Welcome! — Bioinform...
Peak Predictions and ...
A question and answe...
Google Docs – Home
Your Dashboard – Git...
http://biostar.stackexchange.com/
Google
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
3 votes 4 answers 72 views
Visualisation of Alternative splicing events using RNA-seq data
visualization rna-seq alternative-splicing
8m ago Mark Rogers 1
14 votes 28 answers 1k views
What are the most common stupid mistakes in bioinformatics?
mistakes software-engineering
11m ago GWW 770 ●1 ●7
2 votes 1 answer 34 views
Insert size selection in RNASEQ
rnaseq insert size selection
1h ago bioquant 111 ●3
1 vote 1 answer 26 views
Paired end Mapping : What is BWA SOLiD Paired End Default Direction (bwa sampe) ?
bwa paired-end solid
3h ago brentp ♦ 6,005 ●5 ●24
3 votes 1 answer 34 views
Functional divergence
functional divergence
4h ago avilella 1,466 ●4 ●10
0 votes 2 answers 29 views
varscan somatic...syntax to run with redirect?
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

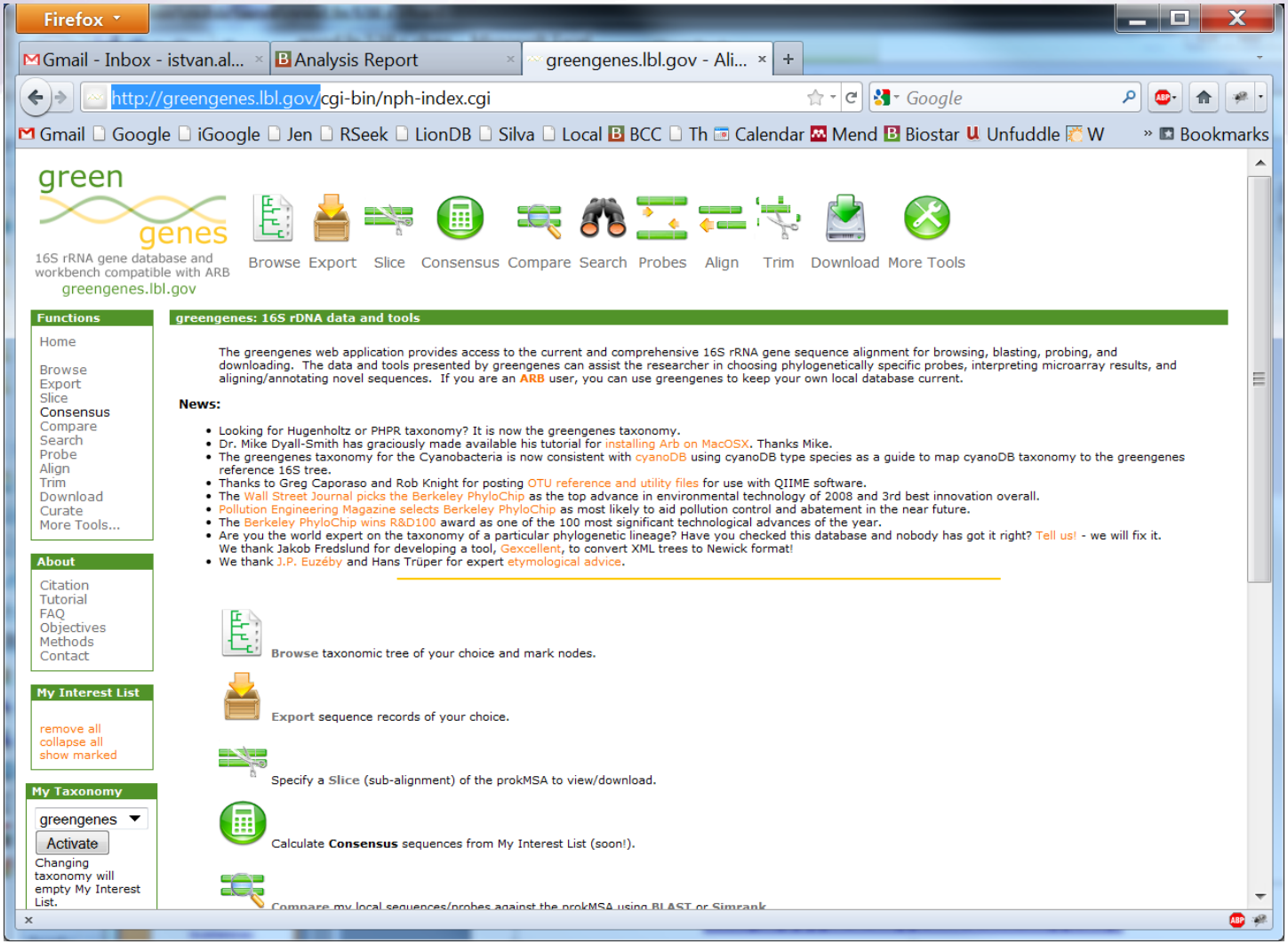
Classifying sequences:

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

Taxonomy

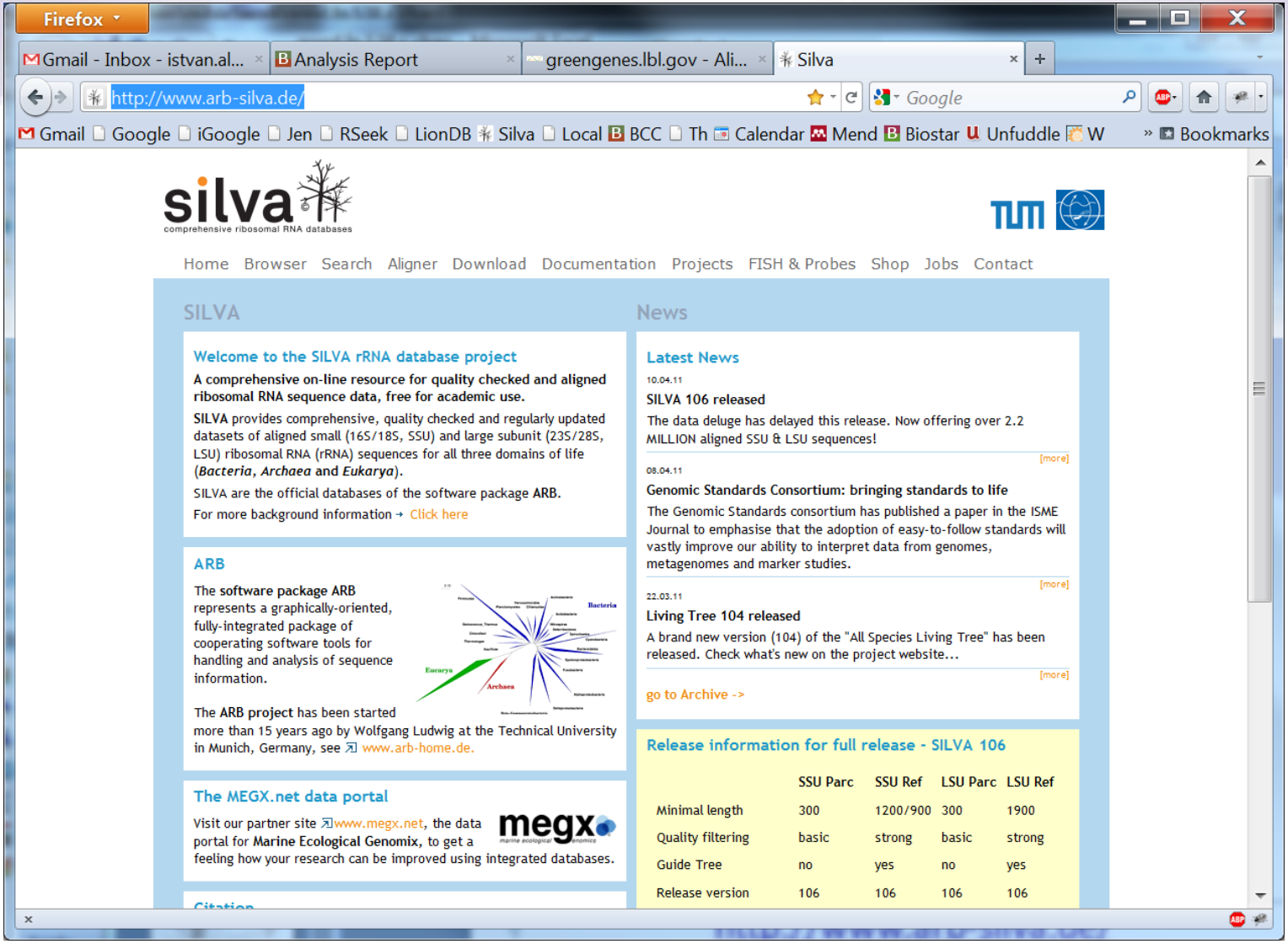
- Placing a bacteria into a taxonomy is difficult
- Several competing groups – each maintain a separate taxonomical database
- Three widely used curated taxonomy outlines that contain significant conflicts with each other

Greengenes



http://greengenes.lbl.gov/

SILVA

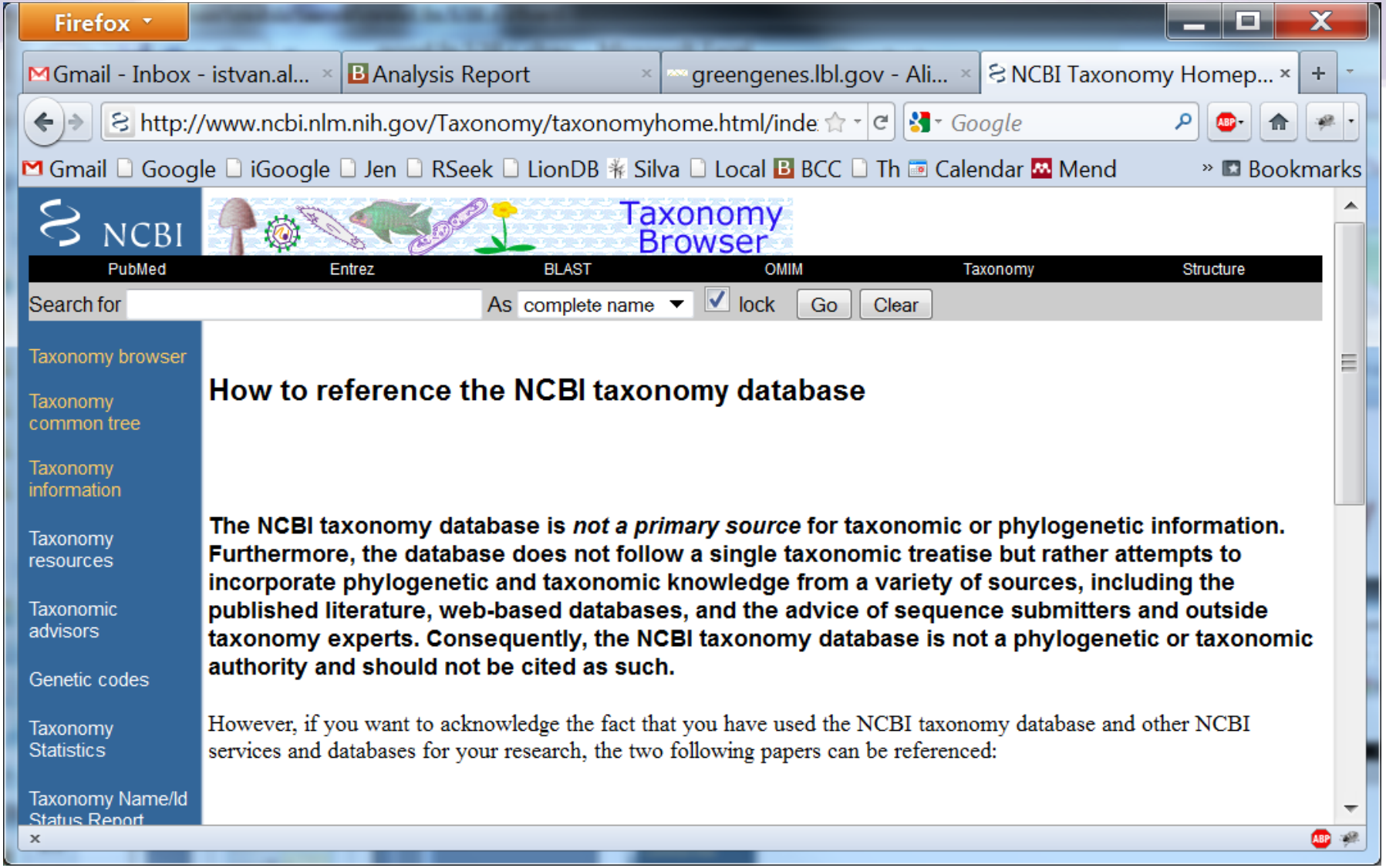


http://www.arb-silva.de/

RDP

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

NCBI taxonomy



One common phylotyping 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

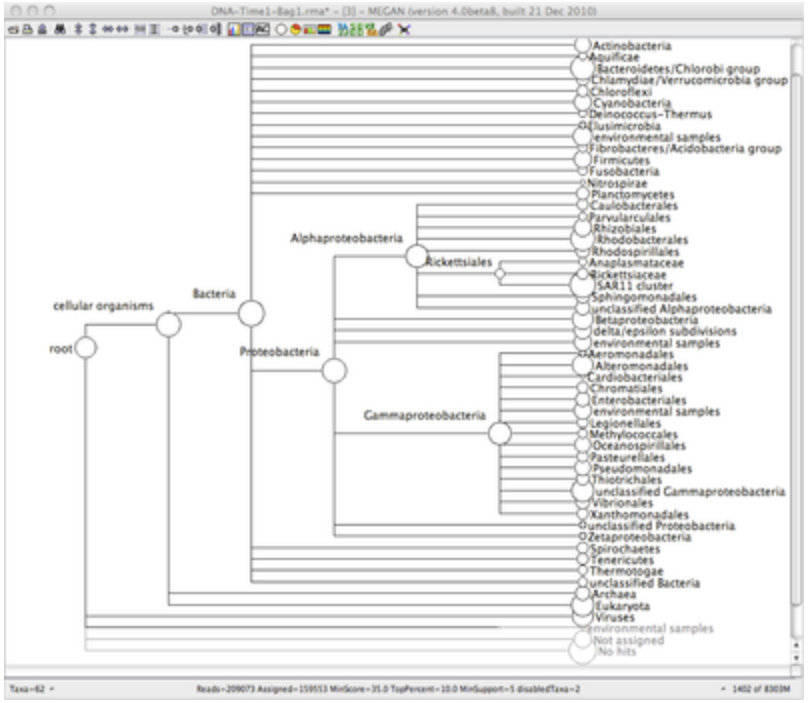
- Align against a hand curated, prealigned representative selection (NAST algorithm).

- This needs far fewer resources

MEGAN

- http://ab.inf.uni-tuebingen.de/software/megan/

- Graphical user interface – nice visualizations



Pitfalls in phyloytping 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 phylotyping 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

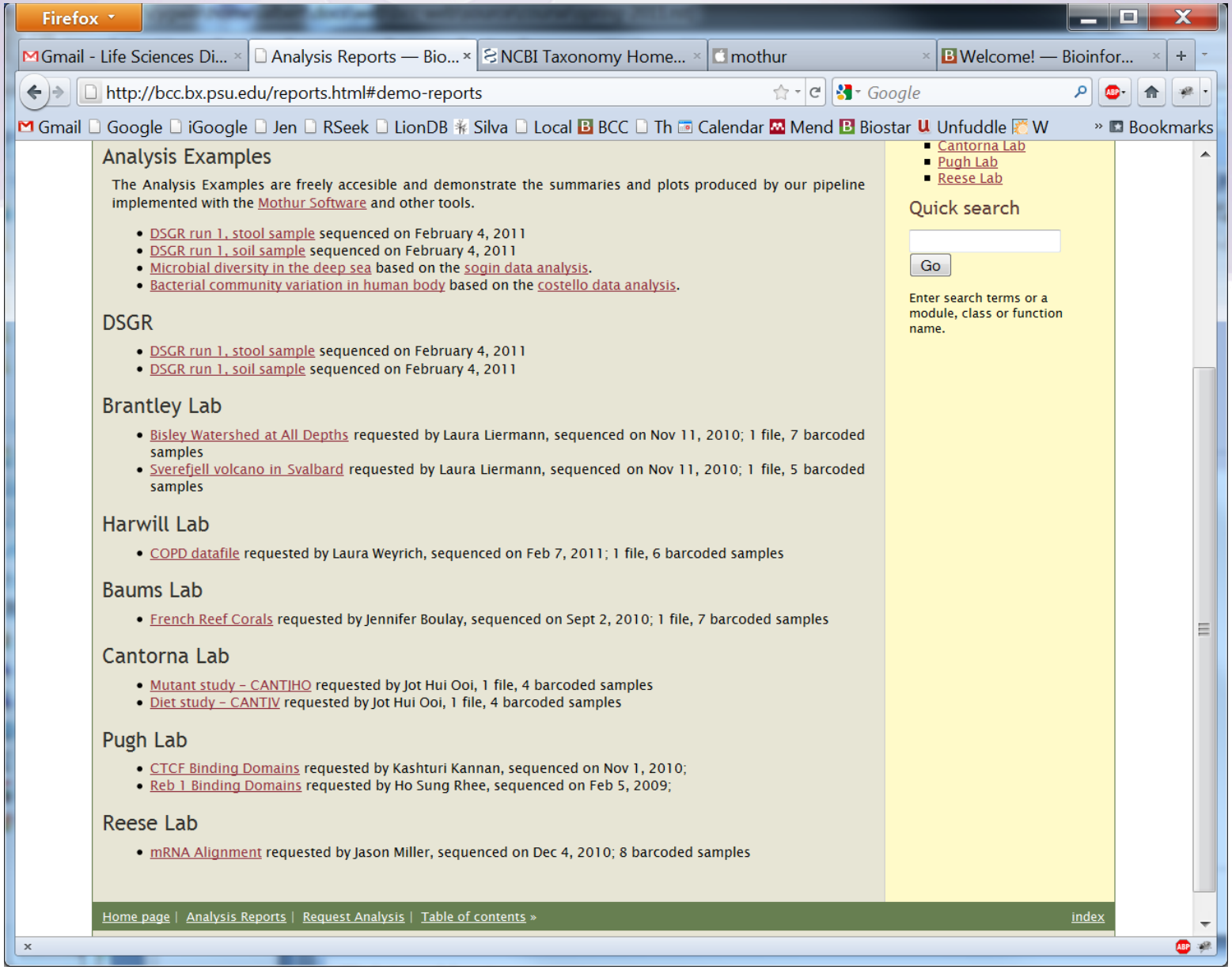
Community diversity – dominance

- Berger-Parker index

largest abundance / total number of individuals

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

- Many other estimators: Simpson, Shannon etc



- 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**

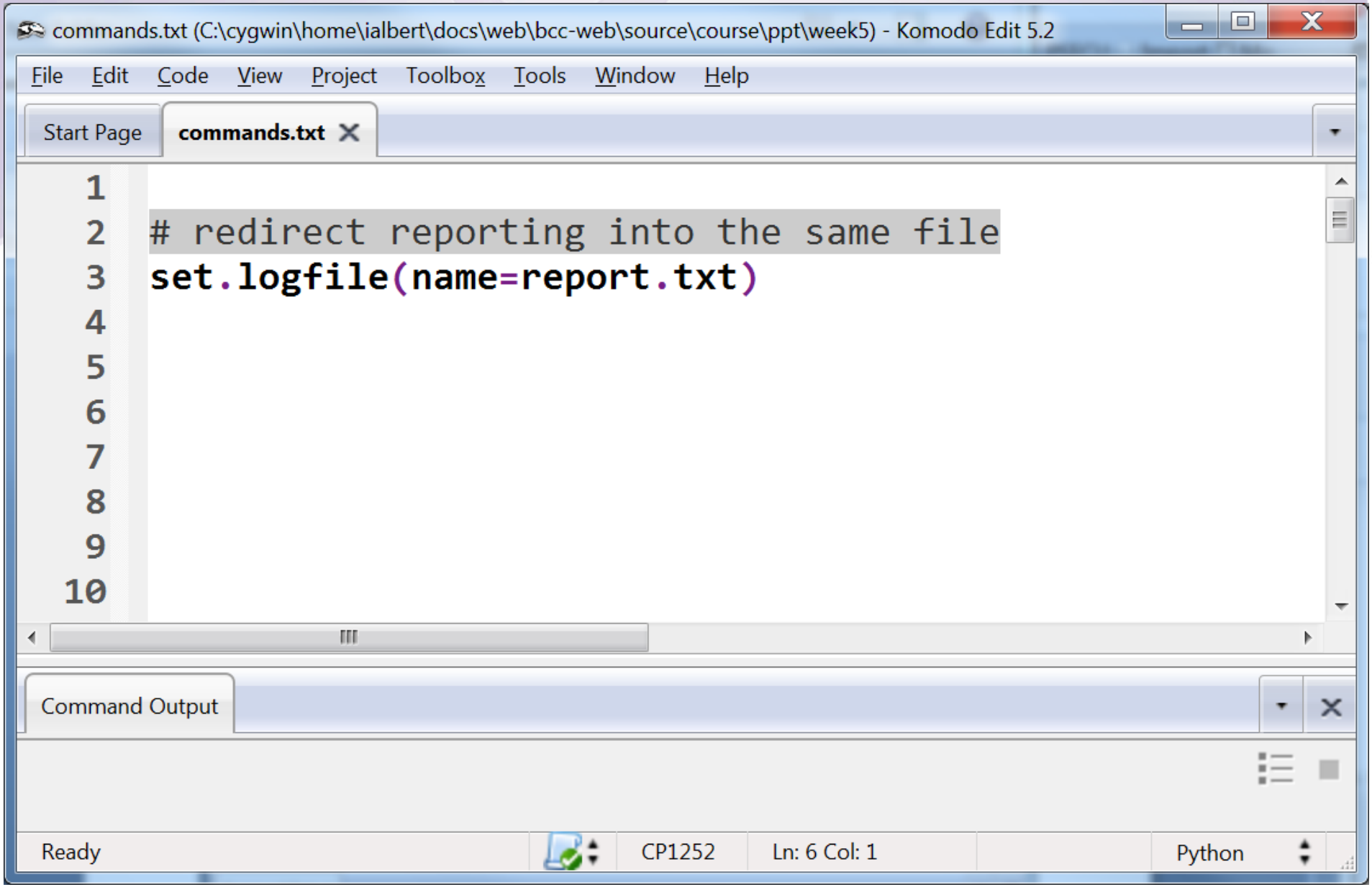
The first to use pyrosequencing technology to sequence 16S rRNA gene tags.

mothur command line (a bit like R)

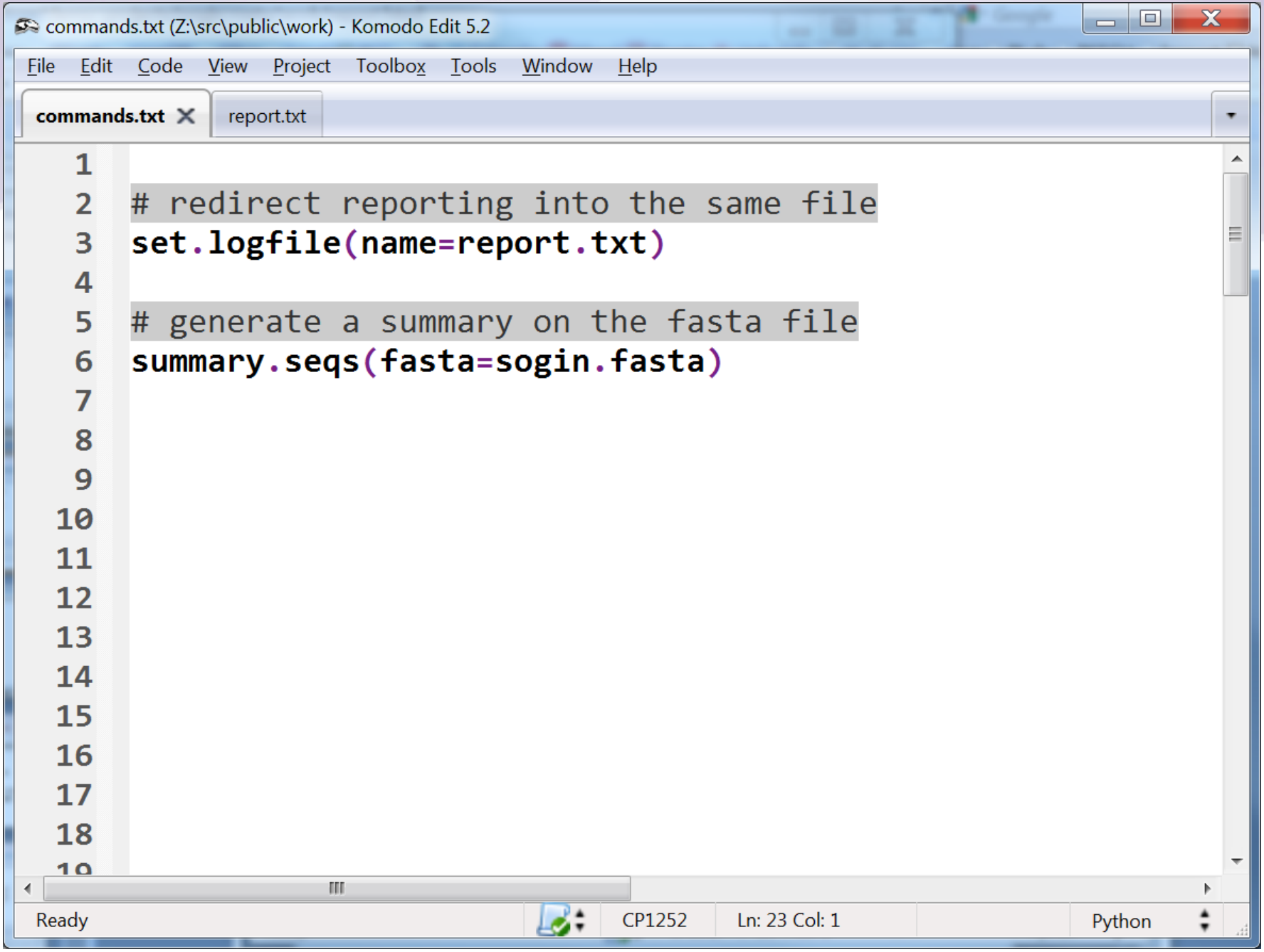
~/docs/web/bcc-web/source/course/ppt/week5

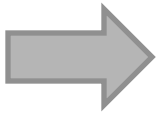
```
mothur v.1.12.3
Last updated: 8/5/2010

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 so
ftware 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**



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


commands.txt (Z:\src\public\work) - Komodo Edit 5.2
File Edit Code View Project Toolbox Tools Window Help
commands.txt
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
Ready
CP1252
Ln: 23 Col: 1
Python

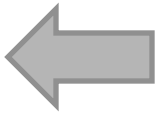


report.txt (Z:\src\public\work) - Komodo Edit 5.2

File Edit Code View Project Toolbox Tools Window Help

commands.txt | report.txt

```
28
29 mothur > set.logfile(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
```



Ready CP1252 Ln: 1 Col: 1 Text

The way **mothur** works

1. Commands generate a readable output to the screen

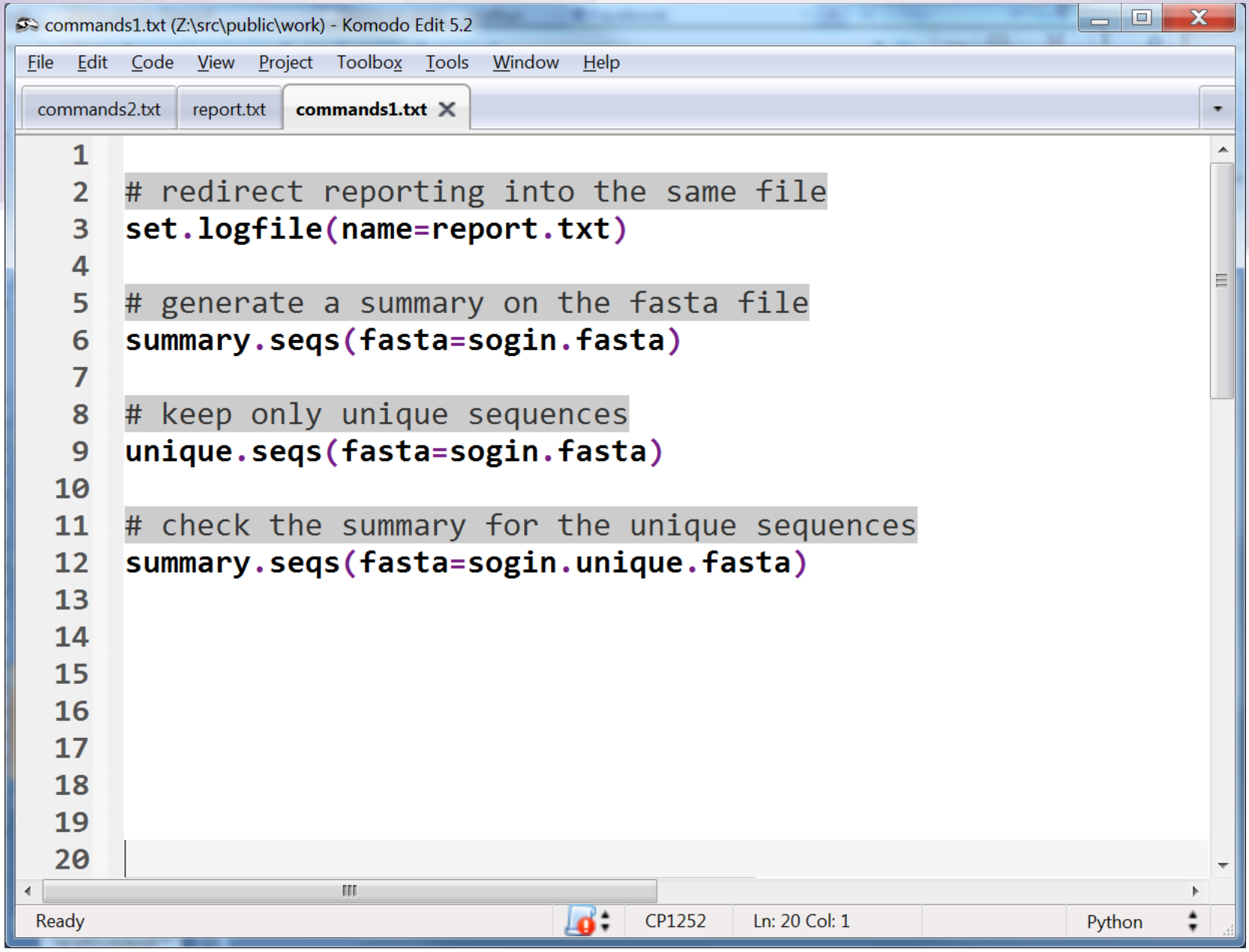
2. The same output goes to the log file

3. Each command usually creates one or more result files that we can either load in later stages or plot (analyze)

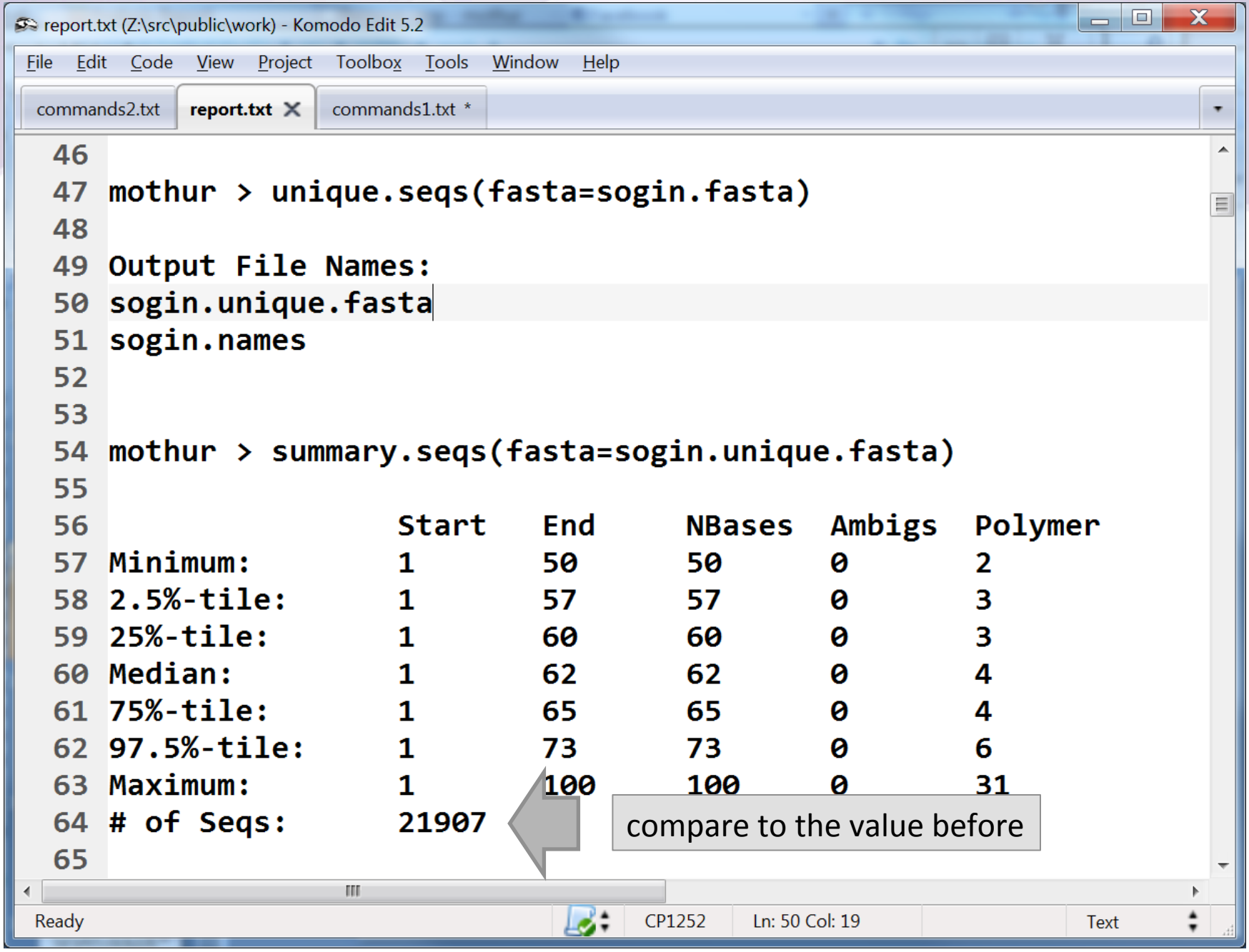
Quality filtering

- 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



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


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 *
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
Ready
CP1252
Ln: 50 Col: 19
Text

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 -->** to **good.fasta**

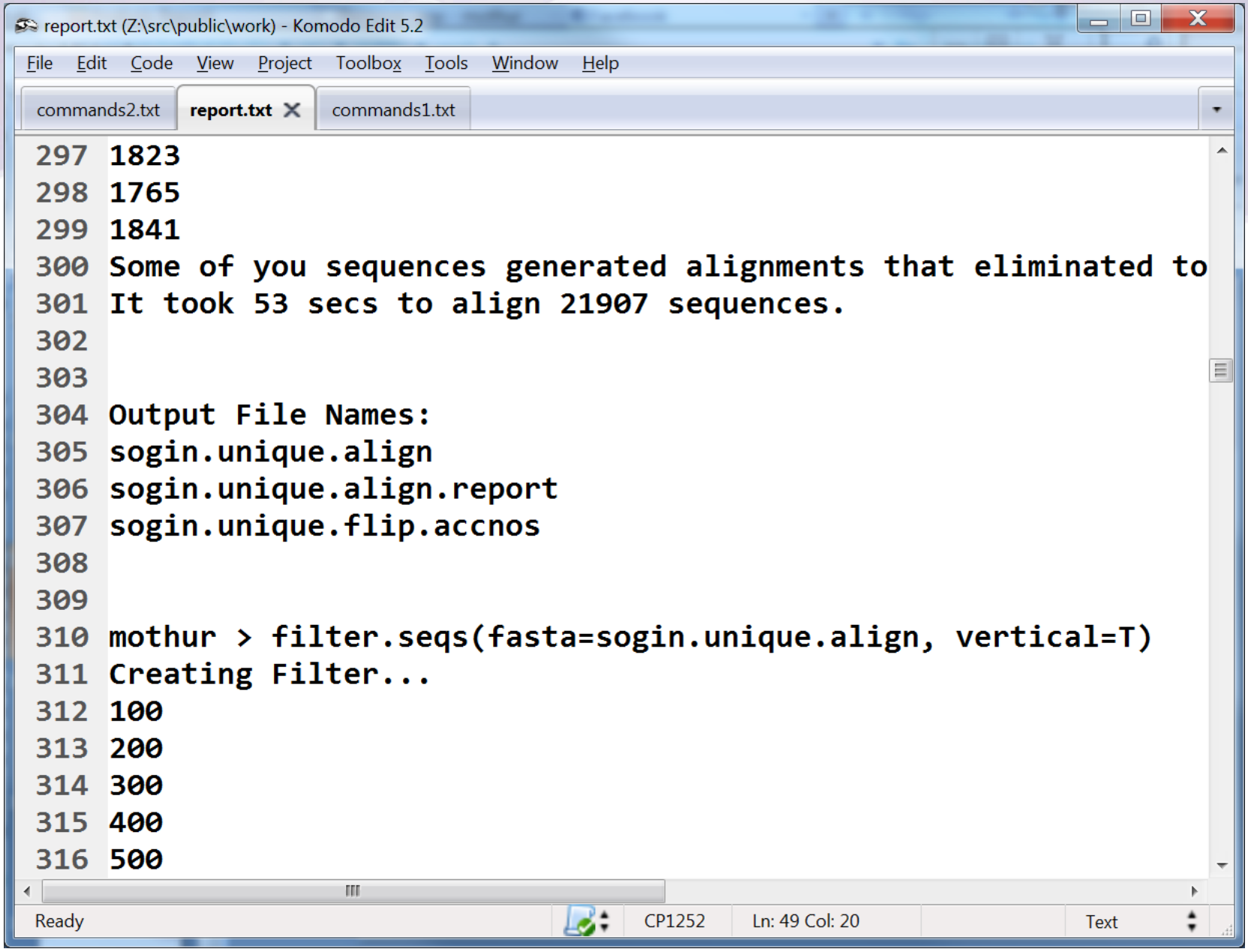
commands1.txt* (Z:\src\public\work) - Komodo Edit 5.2

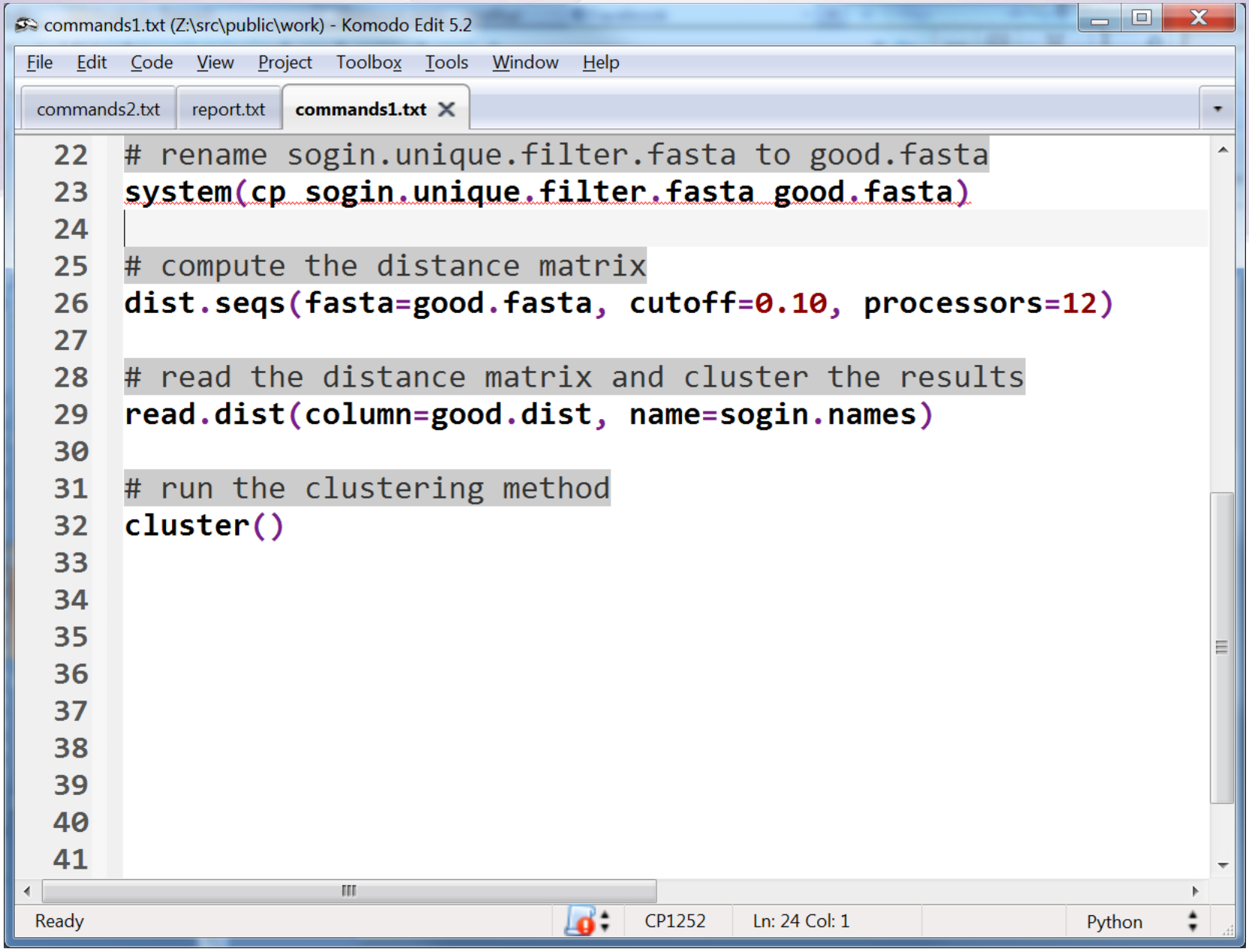
File Edit Code View Project Toolbox Tools Window Help

commands2.txt | report.txt | commands1.txt *

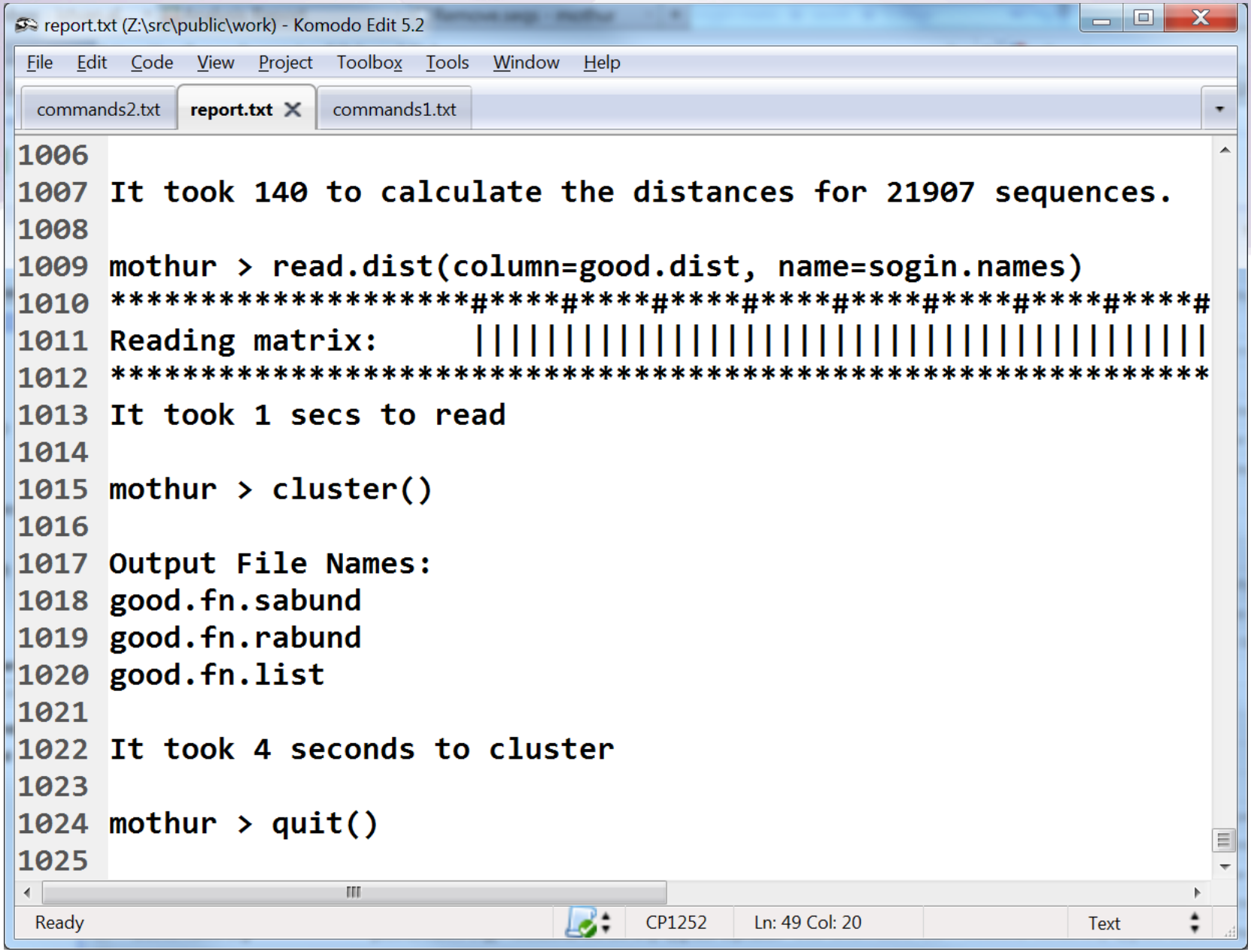
```
10
11 # check the summary for the unique sequences
12 summary.seqs(fasta=sogin.unique.fasta)
13
14 # align against the silva reference
15 # put it on a single line (not as shown here)
16 align.seqs(candidate=sogin.unique.fasta,
17         template=silva.bacteria.fasta, processors=12)
18
19 # filter the resulting alignment
20 filter.seqs(fasta=sogin.unique.align, vertical=T)
21
22 # rename sogin.unique.filter.fasta to good.fasta
23 system(cp sogin.unique.filter.fasta good.fasta)
24
25
26
27
28
```

For this you need the silva reference files → 800MB
This can take a bit more - depending on you 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 sogin.unique.align
306 sogin.unique.align.report
307 sogin.unique.flip.accnos
308
309
310 mothur > filter.seqs(fasta=sogin.unique.align, vertical=T)
311 Creating Filter...
312 100
313 200
314 300
315 400
316 500
Ready
CP1252
Ln: 49 Col: 20
Text

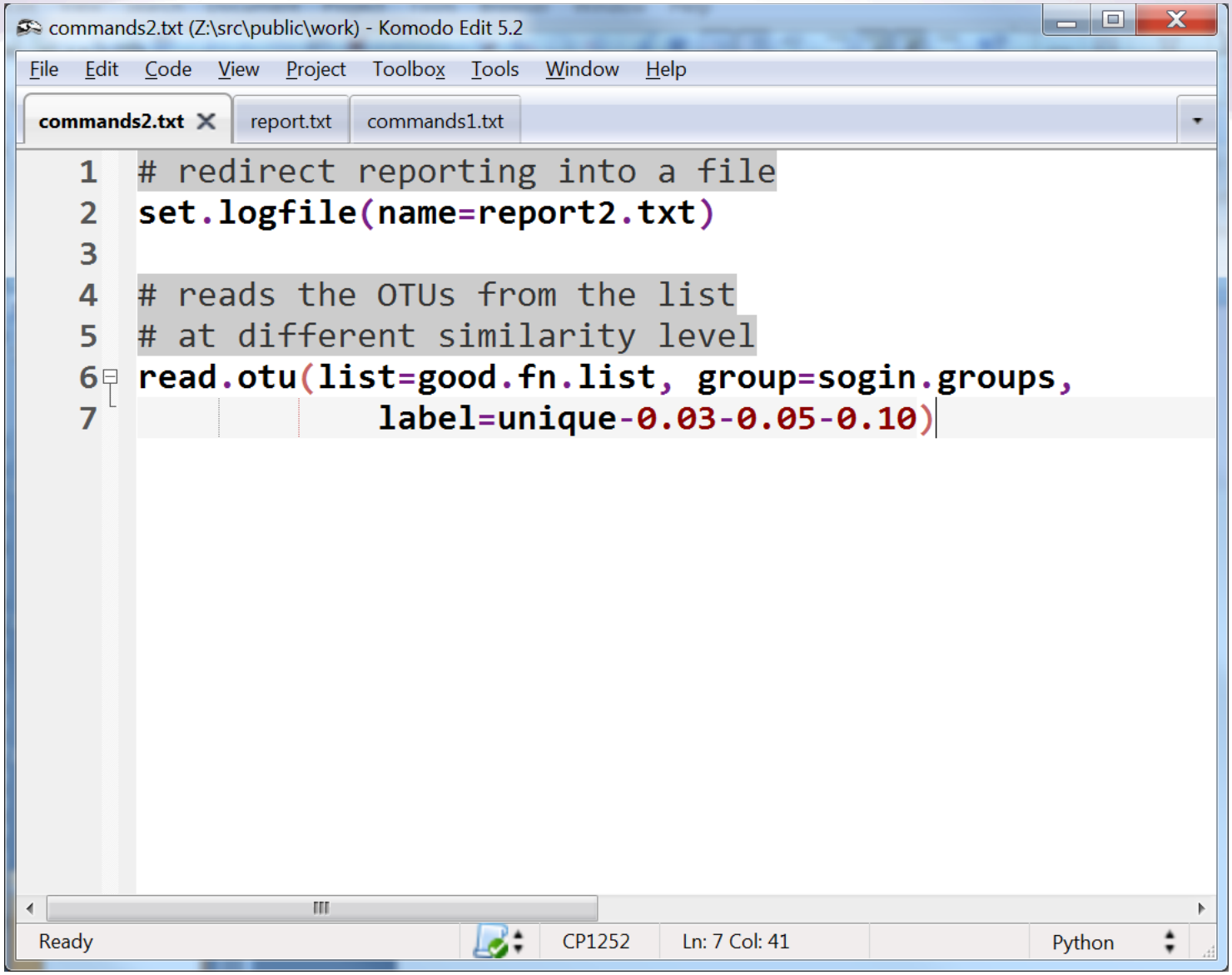


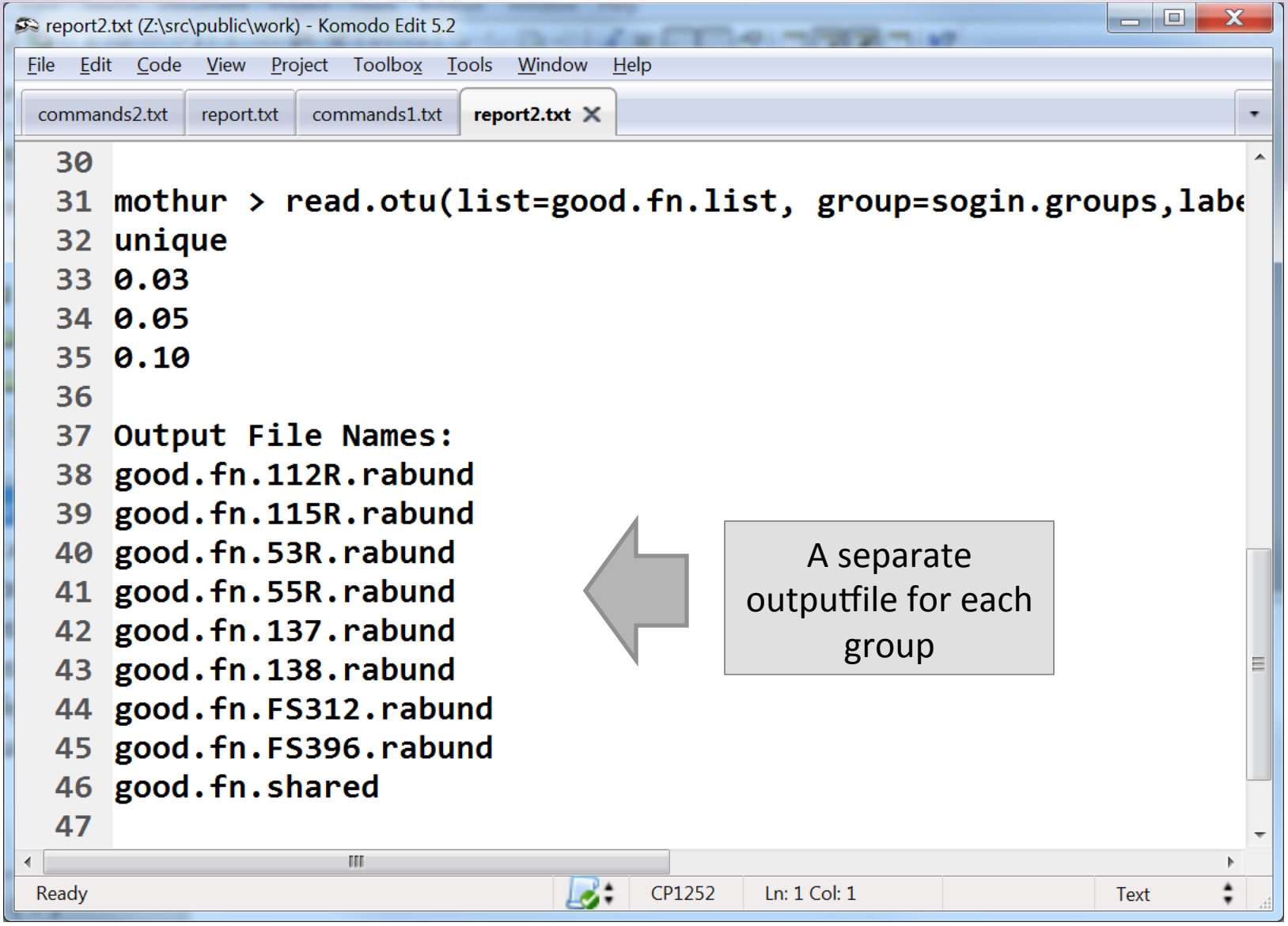
As a result we compute the similarity matrix that interrelates all sequences

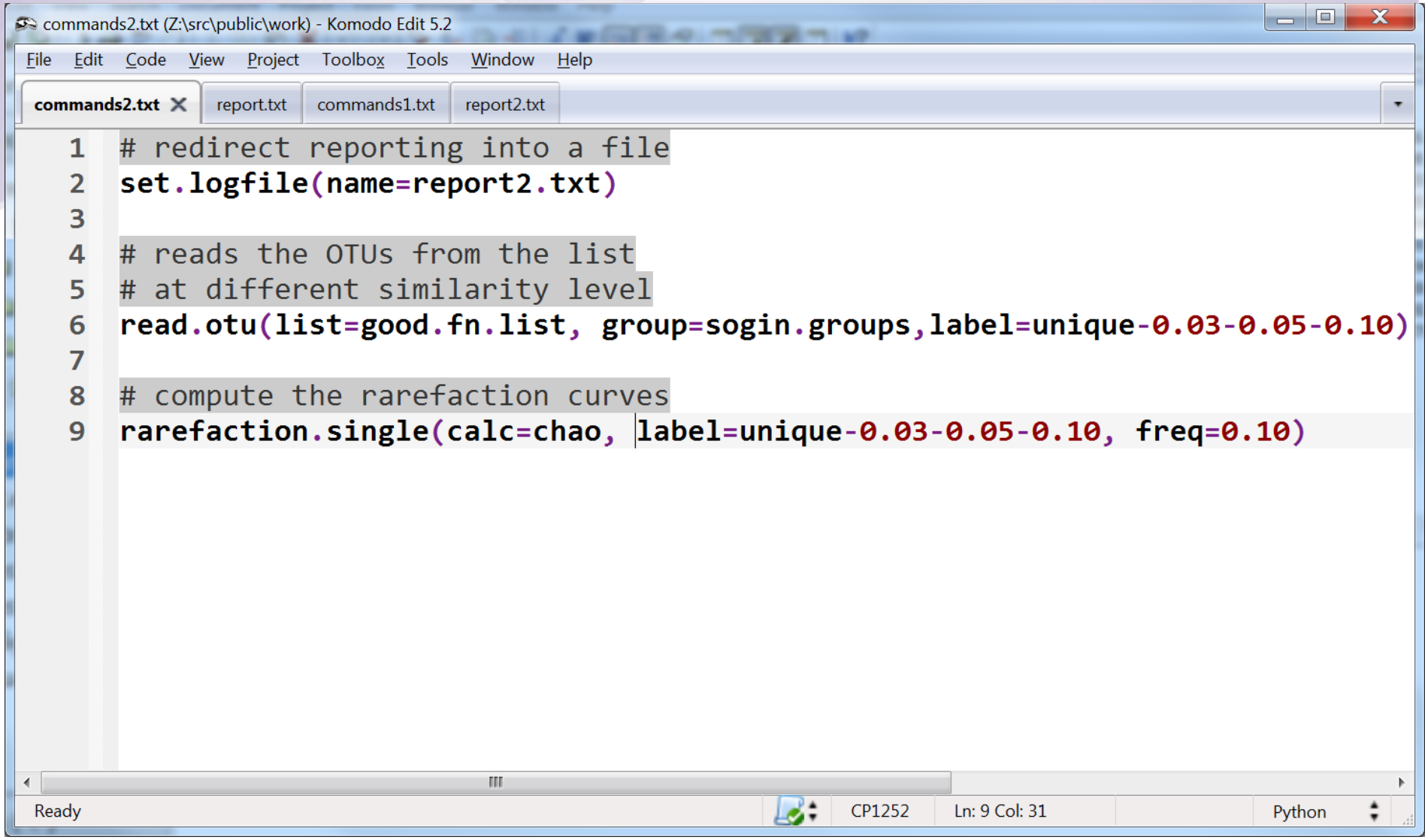


We can now head onto doing analysis

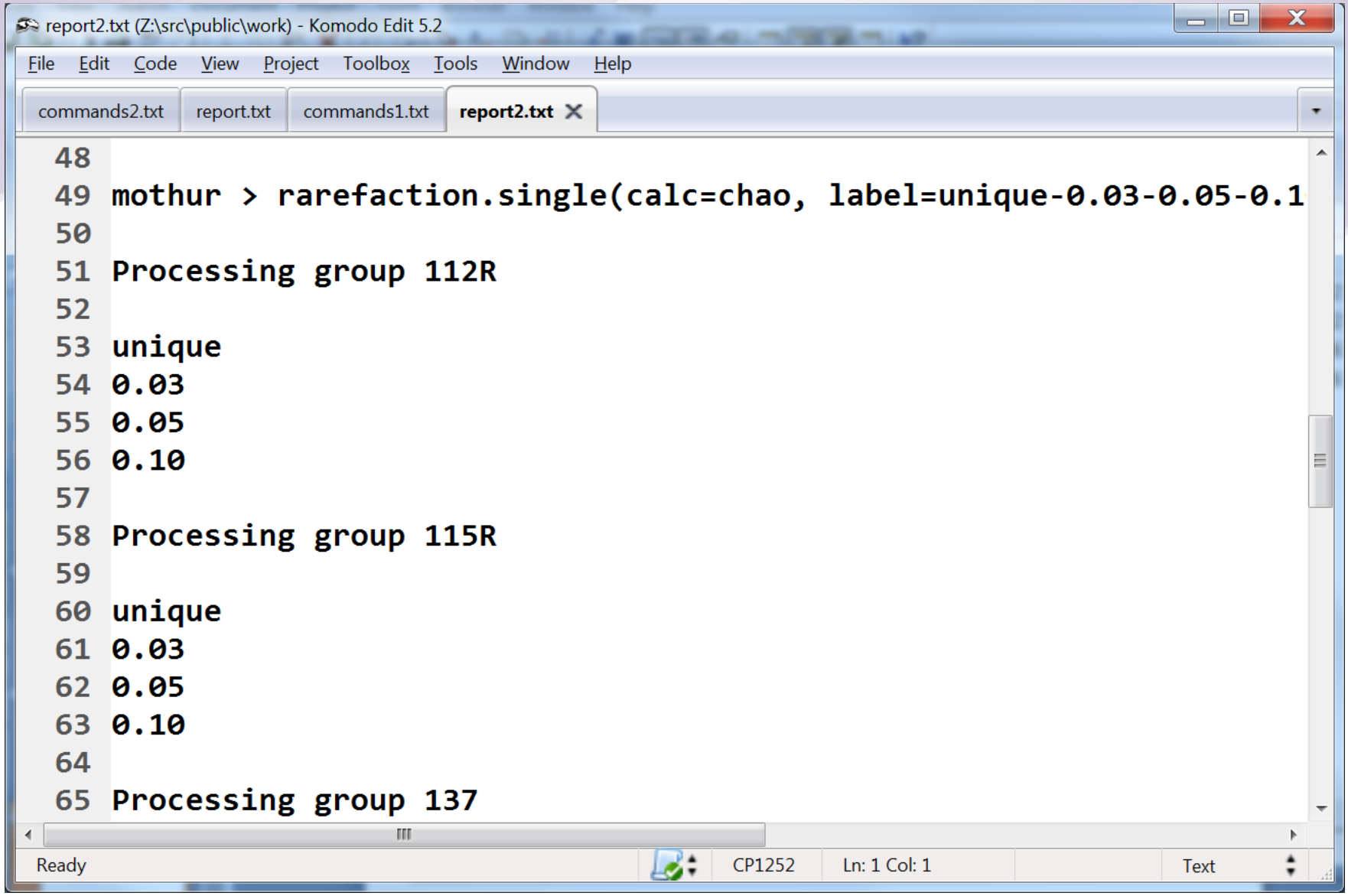
Rarefaction curves



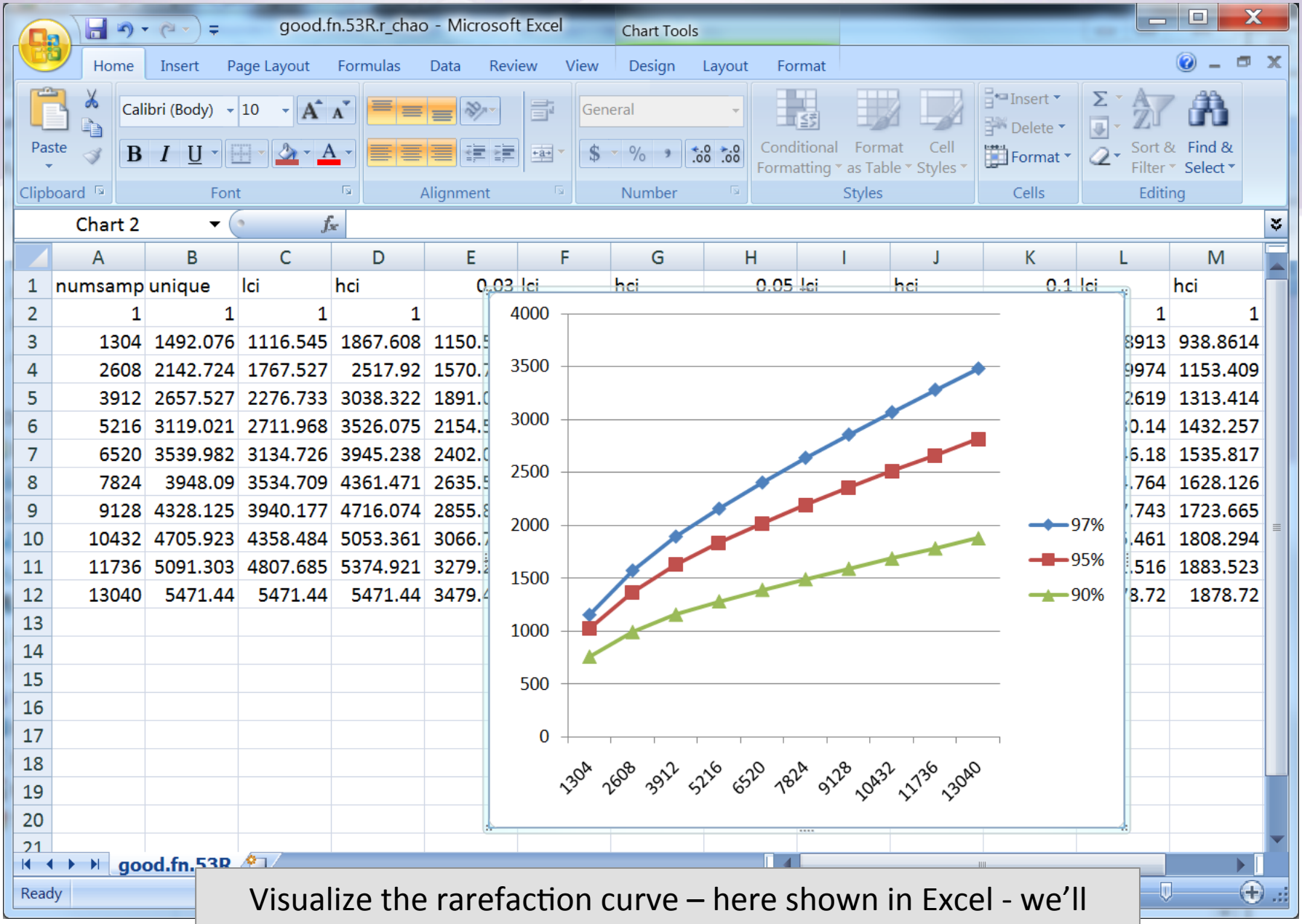

report2.txt (Z:\src\public\work) - Komodo Edit 5.2
File Edit Code View Project Toolbox Tools Window Help
commands2.txt
report.txt
commands1.txt
report2.txt
30
31 mothur > read.otu(list=good.fn.list, group=sogin.groups,labe
32 unique
33 0.03
34 0.05
35 0.10
36
37 Output File Names:
38 good.fn.112R.rabund
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
Ready
CP1252
Ln: 1 Col: 1
Text
A separate outputfile for each group



Generate the rarefaction curves



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