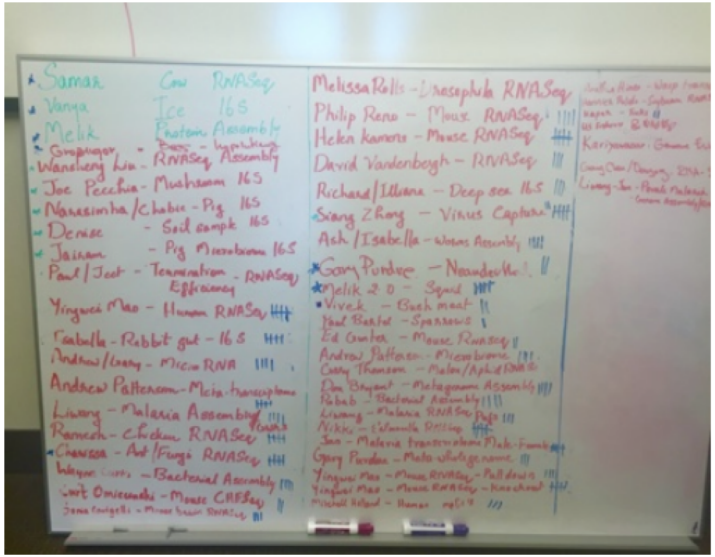
The Unexpected Challenges of Reusing a Pipeline

Aswathy Sebastian

Bioinformatics Consulting Center
Huck Institute of Life Sciences
Penn State

Bioinformatics Consulting Center

- Headed by Dr.Istvan Albert.
- Provide Bioinformatics analysis services.
- Manage and distribute the sequencing data produced by Penn State Sequencing Facilities.
- Collaborators - Penn State Faculty, Northeast Fishery Center, U.S. Fish & Wildlife.

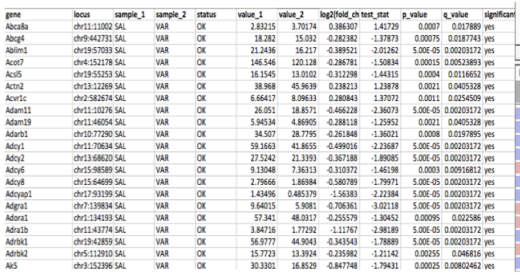


Common requests

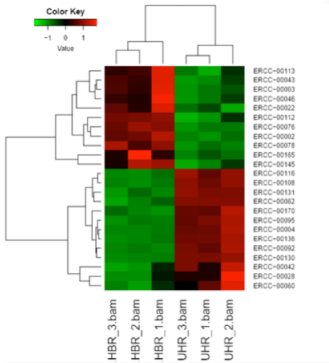
Can you do an analysis like ….

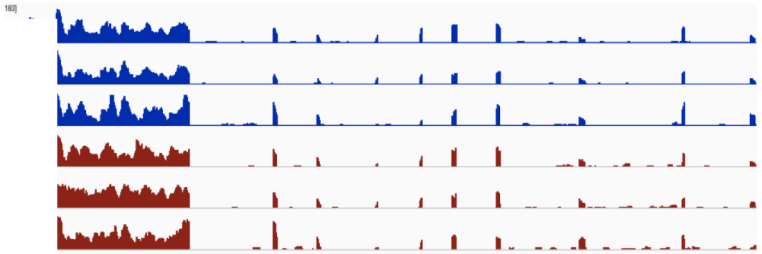
Can you make a plot like …

Can you use the tool published in ..



gene	locus	sample_1	sample_2	status	value_1	value_2	log2(fold_ch	test_stat	p_value	q_value	significant
Abca8a	chr11:110026	SAL	VAR	OK	2.83215	3.70174	0.386307	1.41729	0.0007	0.0178893	yes
Abcg4	chr9:442731	SAL	VAR	OK	18.282	15.032	-0.282382	-1.37873	0.00075	0.0187743	yes
Ablim1	chr19:57033	SAL	VAR	OK	21.2436	16.217	-0.389521	-2.01262	5.00E-05	0.00203172	yes
Acot7	chr4:152178	SAL	VAR	OK	146.546	120.128	-0.286781	-1.50834	0.00015	0.00523893	yes
Acsl5	chr19:55253	SAL	VAR	OK	16.1545	13.0102	-0.312298	-1.44315	0.0004	0.0116652	yes
Actn2	chr13:12269	SAL	VAR	OK	38.968	45.9639	0.238213	1.23878	0.0021	0.0405328	yes
Acvr1c	chr2:582674	SAL	VAR	OK	6.66417	8.09633	0.280843	1.37072	0.0011	0.0254509	yes
Adam11	chr11:10276	SAL	VAR	OK	26.051	18.8571	-0.466228	-2.36073	5.00E-05	0.00203172	yes
Adam19	chr11:46054	SAL	VAR	OK	5.94534	4.86905	-0.288118	-1.25952	0.0021	0.0405328	yes
Adarb1	chr10:77290	SAL	VAR	OK	34.507	28.7795	-0.261848	-1.36021	0.0008	0.0197895	yes
Adcy1	chr11:70634	SAL	VAR	OK	59.1663	41.8655	-0.499016	-2.23687	5.00E-05	0.00203172	yes
Adcy2	chr13:68620	SAL	VAR	OK	27.5242	21.3393	-0.367188	-1.89085	5.00E-05	0.00203172	yes
Adcy6	chr15:98589	SAL	VAR	OK	9.13048	7.36313	-0.310372	-1.46198	0.0003	0.00916812	yes
Adcy8	chr15:64699	SAL	VAR	OK	2.79666	1.86984	-0.580789	-1.79971	5.00E-05	0.00203172	yes
Adcyap1	chr17:93199	SAL	VAR	OK	1.43496	0.485379	-1.56383	-2.22384	5.00E-05	0.00203172	yes
Adgra1	chr7:139834	SAL	VAR	OK	9.64015	5.9081	-0.706361	-3.02118	5.00E-05	0.00203172	yes
Adora1	chr1:134193	SAL	VAR	OK	57.341	48.0317	-0.255579	-1.30452	0.00095	0.022586	yes
Adra1b	chr11:43774	SAL	VAR	OK	3.84716	1.77292	-1.11767	-2.98189	5.00E-05	0.00203172	yes
Adrbk1	chr19:42859	SAL	VAR	OK	56.9777	44.9043	-0.343543	-1.78889	5.00E-05	0.00203172	yes
Adrbk2	chr5:112910	SAL	VAR	OK	15.7723	13.3924	-0.235982	-1.21142	0.00255	0.046816	yes
Ak5	chr3:152396	SAL	VAR	OK	30.3301	16.8529	-0.847748	-1.79431	0.00025	0.00802462	yes





Nature Methods

Benjamin J Callahan, Paul J McMurdie, Michael J Rosen, Andrew W Han,Amy Jo A Johnson , Susan P Holmes

DADA2: High-resolution sample inference from Illumina amplicon data

Nature Methods volume 13, pages 581–583 (2016)

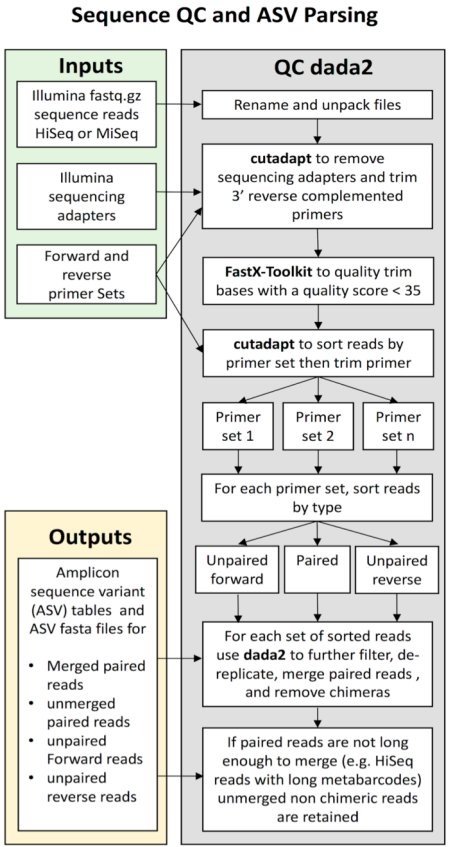
- DADA2 - Divisive Amplicon Denoising Algorithm.
- Models and corrects Illumina-sequenced amplicon sequences.
- Resolve sequences at 1 base pair difference.
- Implemented in CRUX-Anacapa pipeline.

CRUX - Anacapa Pipeline

- Assigns taxonomy to metabarcoded sequences.
- Uses existing tools to process the data.
- Good documentation, github repo.
- https://github.com/limey-bean/Anacapa
- Detailed explanation of the steps in analysis.

Excellent !!

My hope and dream - just run it and get the results



Reality - List of tools to be installed

1. OBITools (requires python 2.7)
2. ecoPCR
3. Cutadapt v1.16
4. FastX toolkit
5. Bowtie2
6. BLAST+ (version 2.6.0)
7. entrez_qiime
8. Muscle v 3.8.31
9. R v 3.4.2
10. ggplot2
11. plyr
12. dplyr
13. seqRFLP
14. Reshape2
15. Tibble
16. devtools
17. Matrix
18. mgcv
19. readr
20. stringr
21. vegan
22. plotly
23. otparse
24. ggrepel
25. cluster
26. phyloseq
27. genefilter
28. impute
29. Biostrings
30. dada2 Version 1.6

Let’s install - #1

Python:

Install an old version of python - python 2.7

Python 2.7 is supported only until 2020.

Python 3 is available since late 2008.

R Package:

otparse - package not available for R.3.4.2 !

Spent time on google search

Solution

Typo! optparse is the package

install.packages(“optparse”)

Let’s install - #2

entrez_qiime.py - a script included in the pipeline.

No Module Found : cogent.parse.ncbi.taxonomy

Easy to fix - pip install cogent

But … another error!

Value error: cogent/align/_comapre.pyx doesn’t match any files.

My Fix

DONT_USE_PYREX=1 pip install -U cogent

Spent about 30 minutes to figure this out.

Undocumented dependencies

Pandas - Which version?

Python Data Analysis Library

pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

pandas is a NumFOCUS sponsored project. This will help ensure the success of development of *pandas* as a world-class open-source project, and makes it possible to donate to the project.

A Fiscally Sponsored Project of

NUMFOCUS

OPEN CODE = BETTER SCIENCE

VERSIONS

Release
0.23.0 - May 2018
download // docs // pdf

Development
0.24.0 - 2018
github // docs

Previous Releases
0.22.0 - download // docs // pdf
0.21.1 - download // docs // pdf
0.21.0 - download // docs // pdf
0.20.3 - download // docs // pdf
0.19.2 - download // docs // pdf

Let’s Run it

No such file or directory:

‘/home/aswathy/src/anacapa_db/scripts/BCC_default_cutoff.sh’

But the file is present !

Program hangs? Why?

```
12S
Running Dada2 inline

Running dada2 on paired reads
0 n
No index, query, or output file specified!
Bowtie 2 version 2.2.8 by Ben Langmead (langmea@cs.jhu.edu, www.cs.jhu.edu/~langmea)
Usage:
  bowtie2 [options]* -x <bt2-idx> {-1 <m1> -2 <m2> | -U <r>} [-S <sam>]

  <bt2-idx>  Index filename prefix (minus trailing .X.bt2).
             NOTE: Bowtie 1 and Bowtie 2 indexes are not compatible.
  <m1>       Files with #1 mates, paired with files in <m2>.
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
  <m2>       Files with #2 mates, paired with files in <m1>.
             Could be gzip'ed (extension: .gz) or bzip2'ed (extension: .bz2).
```

Hi Aswathy,

Ok two things. 1) Thanks for finding the superfluous bowtie2 call in that script. 2) Bowtie2 may cause you troubles in the classifier step. Letmeknow if that happens and I will help you get through it.

The script published in GitHub was never tested and never worked !

Cannot find a tool!

```
Traceback (most recent call last):
  File "/home/aswathy/src/anacapa_db/scripts/blca_from_bowtie.py", line 310, in <module>
    proc = subprocess.Popen([muscle_path, '-quiet', '-clw'], stdout=subprocess.PIPE, stdin=subprocess.PIPE)
  File "/home/aswathy/miniconda3/envs/py2/lib/python2.7/subprocess.py", line 390, in __init__
    errread, errwrite)
  File "/home/aswathy/miniconda3/envs/py2/lib/python2.7/subprocess.py", line 1025, in _execute_child
    raise child_exception
OSError: [Errno 2] No such file or directory
```

The error is due to the fact that the script cannot find muscle. We need fix that particular bug,

Fix - Copy the tool into a specific directory.

Similar issues - Tech support

1. How can I get 'otparse' package? I have R.3.4.2 and when trying to install I get the message that this package is not available.
2. What does anaconda/python-2-4.2 in dependencies list mean? I have conda environment with python 2.7 and biopython installed. Do I need anything else?

You can just delete the 16S directory. You can download both CRUX databases from https://drive.google.com/drive/folders/0BycoA83WF7aNOEFFV2Z6bC1GM1E?usp=sharing. I would select the <metabarcode> filtered library. After you uncompress it you will need to delete everything in the directory name except <metabarcode>.

So were you able to build the CRUX ref lib?!?! Our 12S lib was made with the MiFish primers as recently as December so it is still pretty up to date, but it would be exciting if everything worked to plan on your machine.

Please find a replacement script for run_dada2.sh. I also updated it on the github page. Thanks again for helping us find problems. To the best of my knowledge you may be the first person to attempted to Anacapa this on a linux machine. Hopefully there will not be too many more bugs.

I have optparse installed, I am trying to run 'anacapa_QC_dada2.sh' but it seems to be hanging at some point.

I am not sure if it has to do with the specifications in the config file. Attached is the stdout/err (runlog.txt) along with the config file.

Now I am at the classifier step and the error message says I do not have 16S_taxonomy.txt file.
Since I have only 12S in my dataset I have created only 12S specific reference using crux.sh

bash ~/src/crux_db/crux.sh -n 12S -f GTCGGTAAAACTCGTGCCAGC -r CATAGTGGGGTATCTAATCCCAGTTTG -s 80 -m 280 -o ./12S -d ~/src/crux_db -l

What should I do at this point? Do I need to create crux libraries for all markers though I do not have them in my dataset? Could you advice me?

Took about 2 weeks to successfully install the pipeline.

Takeaways if you are running a pipeline

Implementing a method may seem to be clear and easy, but in reality it is always harder than it looks. Lot harder.

Plan ahead to have enough time to install and run an analysis.

Don’t panic at the sight of error messages. It’s all part of the process and you learn a lot from it.

Have patience.

Takeaways if you are developing a pipeline

Test the pipeline properly, with multiple settings and on multiple platforms.

Must use exact versions - which might be non-recommended ones. Now what?

It is not just enough to have the code published, but it needs to be in a way that people can run it.

Bioinformatics Recipes

https://github.com/biostars/biostar-engine

A web application to allow scientists to create, execute, document and share data analysis scripts with each other.

We call these scripts **recipes.**

A recipe

- Can be any pipeline
- Can be shared and modified
- Comes with test data and results
- Can be run easily by life scientists

A recipe example

4 Data | 4 Recipes | 1 Results | Selection

Recipe: Taxonomic Classification

Run Recipe or View Results

View Code | Copy Recipe | Edit Description

Help

Centrifuge is a microbial classification software that enables rapid, accurate, and sensitive labeling of reads and quantification of species.

Generated Interface

Sequencing Data Directory:
Sequencing Run 2 - Trimmed and Merged
Multiple paired-end sequencing reads.

Sample description:
Samplesheet for Fish Metabarcode Testdata
The sample sheet that describes the data

Library layout:
Single end
Specify the library type.

Reference Sequences:
Fish Sequences by Accession Number
The sequences to classify against

Minimum Length:
150
The minimum length of the match to accept the classification.

Report cutoff:
1
The minimal sum for each row in the final classification.

Run | Back

Generated Script

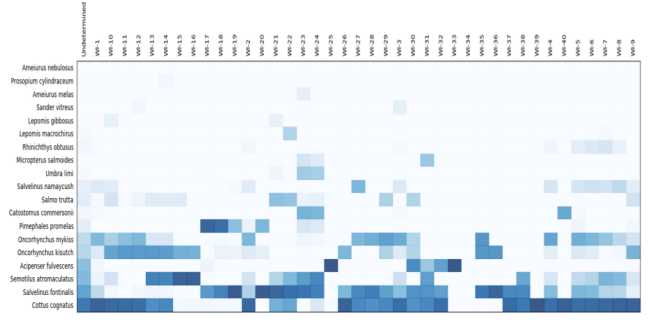
```
# The input directory for the data
DDIR=$(dirname Fish Metabarcode Testdata)

# The reference directory to classify against.
REFERENCE=Fish Sequences by Accession Number

# The sum of each row needs to be above this value.
CUTOFF=1

# The minimal hit length for classification.
HITLEN=150

# The list of files in the directory.
SHEET=Samplesheet for Fish Metabarcode Testdata

# Library type of input reads.
LIBRARY=SE

# Output generated while running the tool.
RUNLOG=runlog/runlog.txt

# Wipe clean the runlog.
```

Interface JSON

```
{
  reads:

  {
    value: Fish Metabarcode Testdata
    label: Sequencing Data Directory
    help: Multiple paired-end sequencing reads.
    source: PROJECT
    type: FASTQ
    display: DROPDOWN
  }

  sheet: {
    value: Samplesheet for Fish Metabarcode Testdata
    label: Sample description
    help: The sample sheet that describes the data
    source: PROJECT
    type: CSV
    display: DROPDOWN
  }
```

Script Template

```
# The input directory for the data
DDIR=$(dirname {{reads.value}})

# The reference directory to classify against.
REFERENCE={{reference.value}}

# The sum of each row needs to be above this value.
CUTOFF={{cutoff.value}}

# The minimal hit length for classification.
HITLEN={{hitlen.value}}

# The list of files in the directory.
SHEET={{sheet.value}}

# Library type of input reads.
LIBRARY={{library.value}}

# Output generated while running the tool.
RUNLOG=runlog/runlog.txt
```



Summary

- Reproducibility crisis exists not only at the analysis and interpretation level but also at implementing existing methods (computational reproducibility).

- To achieve computational reproducibility information about code, software and hardware requirements as well as implementation details are needed.

- Others should be able to run the methods easily to independently reproduce the results and/or to extend the methods to answer similar problems.