Setting up a portable 16S rDNA pipeline for CBIO - Feature #54

DADA2 workflow as an additional option for processing 16S data

05/17/2018 02:05 PM - Gerrit Botha

Status:	New	**Start date:**	05/17/2018
Priority:	Normal	**Due date:**	
Assignee:		**% Done:**	0%
Category:		**Estimated time:**	0.00 hour
Target version:		**Spent time:**	37.00 hours

Description

As mentioned in the description of this project non OTU picking methods such as DADA2 seems to be the choice of many research groups currently. If we managed to find time to setup an additional pipeline using DADA2 processing it would be great.

We have however got DADA2 installed on Hex so researhers can at least use it as is if needed.

History

#1 - 05/17/2018 02:09 PM - Gerrit Botha

Katie and Samson,

Chris has made their DADA2 Nextflow pipeline available here: https://github.com/HPCBio/dada2-Nextflow . Katie if you are keen to test it out let me know. I think we need to make some modifications to get it running on Hex and they also hard coded seems to be all paths.

Gerrit

#2 - 05/17/2018 02:20 PM - Katie Lennard

Hi Gerrit,

Yes I'm keen to help with this and test on the WISH dataset.

#3 - 05/17/2018 02:52 PM - Gerrit Botha

Hi Katie,

I've forked the HPCBIo repos to here: https://github.com/uct-cbio/16S-rDNA-dada2-pipeline . I've given you write permissions to the repos. So lets work and modify that.

Gerrit

#4 - 05/17/2018 03:37 PM - Samson Kilaza

Yeah, I have tried to run the tutorial dataset online (https://benjjneb.github.io/dada2/tutorial.html) using R on my local machine> I could go to the end. On hex we shall need the dada2 formatted reference files for taxonomic annotation (at genus and species level)

Samson

#5 - 05/22/2018 01:54 PM - Katie Lennard

Gerrit I see the Nextflow implementation of DADA2 currently only consists of one .nf file, no configs etc. Shall I try structure if more like the RNAseq Nextflow pipeline?

#6 - 05/22/2018 02:12 PM - Gerrit Botha

Hi Katie,

I think it is definitely better to keep the config and code separate as the RNASeq Nextflow pipeline. For this exercise however I think we should just test it as is first. Once you are happy things run we can make a better arrangement.

Let me know if you need help.

Gerrit

#7 - 06/04/2018 04:42 PM - Katie Lennard

The DADA2 Nextflow pipeline has now been customized for use on UCT hex.
Major updates from originally shared pipeline (https://github.com/HPCBio/dada2-Nextflow)

1. Create separate files for configuration settings (nextflow.config, uct_hex.config) and the main script
2. Include more detailed README and install instructions with UCT and CBIO logos; include help message, detailed log info and email report (emails not currently sending)
3. Build R packages required by pipeline into a singularity container (first docker, then singularity) >The docker image was taken from here https://github.com/joey711/dada2docker and edited to include a user-defined bind point (/researchdata/fhgfs/) on hex; in addition to the dada2 package, all other R packages required by the nextflow dada2 pipeline were included in the docker file (https://github.com/kviljoen/16S-rDNA-dada2-pipeline/blob/dev/Dockerfile) >The docker image was built as follows (from the bst server where docker is installed, can't do on hex): docker build https://github.com/kviljoen/dada2docker.git#dev:base > docker images will list all docker images available on bst >The docker image was converted to a singularity image using the command docker run -v /var/run/docker.sock:/var/run/docker.sock -v /home/katie/h3abionet16S/singularity-containers/:/output --privileged -t --rm singularityware/docker2singularity 29acb43fd7ab >The singularity image was copied via rsync to hex at /scracth/DB/bio/singularity-containers/ and tested using singularity exec (all libraries can be loaded successfully)
4. This version of the pipeline can be found here https://github.com/kviljoen/16S-rDNA-dada2-pipeline/tree/dev

#8 - 06/07/2018 10:46 AM - Katie Lennard

The pipeline is now up and running. Several features were added to make the pipeline more user-friendly:

1. Include tag for each process so easier to see which processes are running on hex when using qstat
2. Customize each process's resource limits (as in RNAseq pipeline - setup to retry if not enough memory assigned in the first place)
3. Build report email, which includes parameters set, success/fail, time spent etc.
4. Change original single file (dada2.nf) to main.nf with parameters in separate config file: main config file with user-defined parameters = nextflow.config; cluster-specific config = uct_hex.config, which can be specified using --profile uct_hex
5. Expand README with links to 'installation' and 'running the pipeline'
6. All R package requirements now loaded via singularity image

Still to do:

1. Run in parallel where possible (i.e. one job/sample)
2. Unit testing, need to do some research for this (https://travis-ci.org/ ?)
3. Run on full WISH dataset (see below)
4. For QC: create one report file per sample (not one aggregate file for all F and another for all R reads)
5. Include step to rerun QC after filtering/trimming step
6. Compare results from dada2 pipeline to uparse pipeline for WISH dataset

Have run on WISH dataset - after some resource setting optimisation in the base.config file, most steps are now fine except 'AlignAndGenerateTree' which takes extremely long (> 20h). I further noticed with the WISH dataset (/scratch/researchdata/cbio/immun/project03/raw/) that a few of the samples (which were included in a separate run after the main run) consist of 300bp reads instead of the usual ~250bp. This is likely due to a different sequencing facility that did not remove Illumina technical sequences such as barcodes/linkers? This led to strange artifacts in the resulting data since dada2 does not have sophisticated primer stripping functionality. Therefore its best to strip primers before the dada2 pipeline (currently testing cutadapt for this)

#9 - 06/07/2018 11:03 AM - Samson Kilaza

Good progress Katie, I will make a try at some point.

Samson

Katie Lennard wrote:

> The pipeline is now up and running. Several features were added to make the pipeline more user-friendly:
>
> 1. Include tag for each process so easier to see which processes are running on hex when using qstat
> 2. Customize each process's resource limits (as in RNAseq pipeline - setup to retry if not enough memory assigned in the first place)
> 3. Build report email, which includes parameters set, success/fail, time spent etc.
> 4. Change original single file (dada2.nf) to main.nf with parameters in separate config file: main config file with user-defined parameters = nextflow.config; cluster-specific config = uct_hex.config, which can be specified using --profile uct_hex
> 5. Expand README with links to 'installation' and 'running the pipeline'
> 6. All R package requirements now loaded via singularity image
>
> Still to do:
>
> 1. Run in parallel where possible (i.e. one job/sample)
> 2. Unit testing, need to do some research for this (https://travis-ci.org/ ?)
> 3. Run on full WISH dataset

#10 - 06/22/2018 02:51 PM - Katie Lennard

The following updates have been made:

1. *Run in parallel where possible (i.e. one job/sample*) - done
2. *Run on full WISH dataset (see below)* - done
3. *For QC: create one report file per sample (not one aggregate file for all F and another for all R reads)* - done - now uses fastQC and multiQC on

a per-sample basis (with multiqc for all sample summary) both before and after filterandtrim step.
4. *Include step to rerun QC after filtering/trimming step* - done, see 4.
5. *Compare results from dada2 pipeline to uparse pipeline for WISH dataset* - basic comparison done, compares well to uparse pipeline in terms of broader taxonomic composition; report to follow.

Most up to date pipeline is available at https://github.com/uct-cbio/16S-rDNA-dada2-pipeline