Setting up a portable RNA-Seq pipeline for CBIO - Support #25

Support # 24 (New): Setting up the NGI-RNAseq pipeline on UCT Hex

Configure NGI-RNAseq pipeline to run on hex

03/15/2018 02:55 PM - Katie Lennard

Status:	In Progress	**Start date:**	04/09/2018
Priority:	Normal	**Due date:**	
Assignee:		**% Done:**	0%
Category:		**Estimated time:**	0.00 hour
Target version:		**Spent time:**	12.60 hours

Description

Pertinent points for setup of NGI-RNAseq pipeline on UCT hex
*Main pipeline source code is at https://github.com/SciLifeLab/NGI-RNAseq
*Currently used pipeline source code however is at https://github.com/ewels/nf-core-RNAseq (this was kindly customized for us by the authors for easy configuration on hex and includes a config file 'uct_hex.config') so that this 'profile' can be called as a flag on the command line (further customization may be required following testing).
*Additional overview on NGI-RNAseq pipeline at https://scilifelab.github.io/courses/rnaseq/1711/slides/pipeline.pdf
*Software requirements will be met using Singularity - the image has been downloaded and stored here /scratch/DB/bio/singularity-containers/ngi-rnaseq.img using the command: singularity pull --name ngi-rnaseq.img docker://scilifelab/ngi-rnaseq
Note that the singularity image path has been specified in the aforementioned uct_hex.config file so no need to specify on job submission.

- First test: nextflow run SciLifeLab/NGI-RNAseq --help | ewels/nf-core-RNAseq

- Reference genomes and annotation files should be placed in /scratch/DB/bio/rna-seq (iGenomes GRCh37 has been pulled to /scratch/DB/bio/rna-seq/references/ from https://ewels.github.io/AWS-iGenomes/) and this location is referenced in our custom uct_hex.config file under the parameter igenomes_base = '/scratch/DB/bio/rna-seq/references'

  In order to download /scratch/DB/bio/rna-seq/references/ from https://ewels.github.io/AWS-iGenomes/ Andrew had to install aws tools on hex, which should be loaded as follows:
  module load python/anaconda-python-2.7
  aws configure
  You may then be prompted for a key and a security key (you need to register an aws account to get this, which is free but you still need to specify credit card details – see https://console.aws.amazon.com)

- For reproducibility please specify the pipeline version used when running the pipeline using the -r flag (e.g. –r 1.3.1)

- The basic run will look something like this:
  nextflow run ewels/nf-core-RNAseq --reads '/researchdata/fhgfs/katie/NGI-RNAseq-test/*_R{1,2}.fastq.gz' --genome GRCh37 --outdir /researchdata/fhgfs/katie/NGI-RNAseq-test/nextflow-output -profile uct_hex --email katie.viljoen@uct.ac.za

- Human RNAseq test data to be used: http://h3data.cbio.uct.ac.za/assessments/RNASeq/practice/ (downloaded to /researchdata/fhgfs/katie/NGI-RNAseq-test)

- First test run:

  qsub -I -q UCTlong -d pwd
  nextflow run ewels/nf-core-RNAseq --reads '/researchdata/fhgfs/katie/NGI-RNAseq-test/*_R{1,2}.fastq.gz' --genome GRCh37 --outdir /researchdata/fhgfs/katie/NGI-RNAseq-test/nextflow-output -profile uct_hex --email katie.viljoen@uct.ac.za

Subtasks:

Support # 30: Update Nextflow to version 0.27.6	**In Progress**

History

#1 - 03/15/2018 03:07 PM - Gerrit Botha

Katie just a note. The section singularity.cacheDir = "/scratch/DB/bio/singularity-containers in our config.txt is actually not being used at the moment. This is used for when we pull containers directly from Dockerhub or Quay.io and convert them to singularity containers on the fly. It does no harm to leave it as is.

#2 - 03/15/2018 03:53 PM - Katie Lennard

- Description updated

#3 - 03/15/2018 04:30 PM - Katie Lennard

- Description updated

#4 - 03/16/2018 12:07 PM - Katie Lennard

- Description updated

#5 - 04/06/2018 09:10 AM - Katie Lennard

- Description updated

#6 - 04/06/2018 03:38 PM - Katie Lennard

- Description updated

- Status changed from New to In Progress

#7 - 04/09/2018 11:08 AM - Gerrit Botha

Hi Katie,

Initially asked me if the info you are planning to share in this ticket should go in the ticket or on the wiki. Seeing how you have been working on this ticket and adding info in the description all the time I think it should now actually go to the Wiki. Please add a page to the Wiki where you document this config for now.

You might add challenges or issues relating the config as updates to this ticket.

Thinking of it now. Ticket #24 was probably sufficient as the base of the setup and updates on setup on Hex. Ticket #30 could probably stand on its own or rather be a subtask of ticket #24.

Cool work.

Gerrit

#8 - 04/09/2018 11:14 AM - Katie Lennard

Gerrit Botha wrote:

> Hi Katie,
>
> Initially asked me if the info you are planning to share in this ticket should go in the ticket or on the wiki. Seeing how you have been working on this ticket and adding info in the description all the time I think it should now actually go to the Wiki. Please add a page to the Wiki where you document this config for now.
>
> You might add challenges or issues relating the config as updates to this ticket.
>
> Thinking of it now. Ticket #24 was probably sufficient as the base of the setup and updates on setup on Hex. Ticket #30 could probably stand on its own or rather be a subtask of ticket #24.
>
> Cool work.
>
> Gerrit

Ok Gerrit, will do!

#9 - 04/10/2018 01:41 PM - Katie Lennard

Development branch with custom 'uct_hex' profile up and running but having issues with the singularity image specified (input files not found on hex even when specified with --reads flag). Need to pull devel branch singularity image with singularity pull --name nfcore-rnaseq-1.4.img docker://nfcore/rnaseq:1.4 but I can't do this - error: Could not obtain the container size, try using --size
ABORT: Aborting with RETVAL=255
Note: the singularity image apparently can also be downloaded dynamically for each run (~2GB in size) by specifying the following in the config file:
singularity {
enabled = true
}
process {

container = "docker://$wf_container"
}

#10 - 04/10/2018 02:22 PM - Gerrit Botha

So what happens if you do use the --size flag with a rough estimate of the container size?

Ok, try running things by pulling the docker file and converting to singularity on the fly. See how it goes.

Just something on the non-native singularity mounts. You need to check if the kernel on the Hex cluster allow for mounting of non-native singularity mounts. Otherwise you would need to modify the dockerfile to mount /researchdata.

#11 - 04/10/2018 05:33 PM - Katie Lennard

Gerrit Botha wrote:

> So what happens if you do use the --size flag with a rough estimate of the container size?
>
> Ok, try running things by pulling the docker file and converting to singularity on the fly. See how it goes.
>
> Just something on the non-native singularity mounts. You need to check if the kernel on the Hex cluster allow for mounting of non-native singularity mounts. Otherwise you would need to modify the dockerfile to mount /researchdata.

https://github.com/ewels/nf-core-RNAseq/issues/21

#12 - 04/11/2018 10:55 AM - Gerrit Botha

Hi Katie,

1. Can you please point me to the Docker / Singularity file on GitHub which you have an issue with?
2. Also please send me the instructions to run things on Hex so that I can reproduce your issue.

I will try and have a look at this later this afternoon otherwise tomorrow.

Regards,
Gerrit

#13 - 04/11/2018 11:11 AM - Katie Lennard

Hi Gerrit,

At the moment the main problem is that I can't download the development version of the Singularity image, which I think is a problem on their side because I can download the ngi-rnaseq one fine. So I'll get that first and test with that before you start troubleshooting. I've also asked Andrew what the current settings are for Singularity on hex regarding user defined bind points. Will keep you posted.

Thanks!

#14 - 04/11/2018 11:18 AM - Gerrit Botha

OK that is fine Katie.

Just on the overlay issue. Overlayfs is only available for kernel 3.18 or higher: https://wiki.archlinux.org/index.php/Overlay_filesystem

On Hex

```
gerrit@srvslshpc001:nextflow> uname -r
3.0.101-108.13-default
```

I do not think it is is possible to automount non-native Singularity mount points on Hex using Nextflow since the kernel < 3.18. You would need to change the Docker/Singularity file to manually mount /researchdata.

#15 - 04/11/2018 11:23 AM - Katie Lennard

Oh dear..Ok we'll have to get rebuild the image then I guess..See latest response from Phil on this
https://github.com/ewels/nf-core-RNAseq/issues/21

#16 - 04/11/2018 11:28 AM - Katie Lennard

Katie Lennard wrote:

> Oh dear..Ok we'll have to get rebuild the image then I guess..See latest response from Phil on this
> https://github.com/ewels/nf-core-RNAseq/issues/21

So we will need to make our own version of the container from nf-core/RNAseq - do you just need the dockerfile under nf-core/RNAseq for this?

#17 - 04/11/2018 11:41 AM - Gerrit Botha

Katie if this is the Docker file you are using https://github.com/ewels/nf-core-RNAseq/blob/master/Dockerfile you would need to find a way to add this line https://github.com/h3abionet/h3abionet16S/blob/master/dockerfiles/fastqc/Dockerfile#L36 to it. Maybe you can create a branch in the nf-core-RNAseq repos or fork it and make that small change in their Docker file and then use that branch/fork .

Maybe for this it is good to ask Phil for suggestions on what would be the best practice? We do not want to be out of date in regards to the main branch if we branch or fork out.

#18 - 04/12/2018 05:04 PM - Gerrit Botha

Hi Katie,

I did the following.

1. Made a fork of the stable RNASeq pipeline to here: https://github.com/uct-cbio/RNAseq . I have added you as contributor to the repos so you will be able to make changes witout having to make pull requests.

2. Then I pulled the code to bst.cbio.uct.ac.za. bst.cbio.uct.ac.za has Docker installed so we can build Docker images on the machine and convert them to Singularity. Just check the machine only has around 80GB of space, so clean up images/files you are not using so that we do not run out of space.

3. I switched to the dev branch because Phil place the Hex nextflow conflig in there. We can maybe merge everything later to the main branch and then try to keep up to date with everything that is going on in nf-core/RNASeq. Switching to dev

```
gerrit@bst:~/code/UCT-CBIO-RNAseq$ git branch -a
* master
  remotes/origin/HEAD -> origin/master
  remotes/origin/dev
  remotes/origin/master
gerrit@bst:~/code/UCT-CBIO-RNAseq$ git checkout dev
Branch dev set up to track remote branch dev from origin.
Switched to a new branch 'dev'
You have new mail in /var/mail/gerrit
gerrit@bst:~/code/UCT-CBIO-RNAseq$ git branch -a
* dev
  master
  remotes/origin/HEAD -> origin/master
  remotes/origin/dev
  remotes/origin/master
```

1. Then I editted the Docker file and added the lines (see https://github.com/uct-cbio/RNAseq/blob/dev/Dockerfile#L11 and https://github.com/uct-cbio/RNAseq/blob/dev/Dockerfile#L12

```
RUN mkdir -p /researchdata/fhgfs
RUN mkdir -p /scratch
```

1. I committed the code back to the dev branch on GitHub

2. Then I build the Docker image

```
gerrit@bst:~/code/UCT-CBIO-RNAseq$ docker build --tag uct-cbio-rnaseq .
```

1. And made the Singularity image

```
docker run -v /var/run/docker.sock:/var/run/docker.sock -v /home/gerrit/scratch/singularity-containers/:/outpu
t --privileged -t --rm singularityware/docker2singularity uct-cbio-rnaseq
```

1. I copied it over to Hex and tested the mounts. Note, uct-cbio-rnaseq.img is soft linked to uct-cbio-rnaseq-2018-04-12-64b02180ccd0.img.

```
gerrit@srvslshpc605:singularity-containers> singularity exec /scratch/DB/bio/singularity-containers/uct-cbio-r
naseq.img ls /scratch/
DB                     abaqus-613       arossgillespie      build-uvcdat  dnyangahu              gerrit     h
sadiq        jbergh       kmcdermott    paul.nicol    sa_utilities  sea         tsalie
Freesurfer             akoch            ayton               carine        eskom                  ghrism001  i
natarajan    jmugo        lmbjas002     researchdata  sancoop       sgarnett    uvcdat
HumGen_Resources       alewis           beegfs-client.conf  coct          fnindo                 gsmith     j
ambler       kbnmat001    nmathai       rndroj001     sclaassen     splunk-bak  vshekhar
Structures_181_linux64  arecibo-scratch  beegfs-mounts.conf  dharris       geog_static_wrf3pt7n8  gventer    j
ason.hlozek  kelsey.jack  opt_exp_soft  root          sdalvie       tcarr       ztimol
```

```
gerrit@srvslshpc605:singularity-containers> singularity exec /scratch/DB/bio/singularity-containers/uct-cbio-rnaseq.img ls /researchdata/fhgfs/
TBMseq          apinska           cbio              dmsnic001       gventer     hpc05  hpc13  hpc21  hpc29  hpc37
                jason.hlozek      lamech.mwapagha   mgglor001  nrmjar001     rtaylor       sumir.panji
a               arecibo-scratch   celia.vdmerwe     elssam003       gzxeph001   hpc06  hpc14  hpc22  hpc30  hpc38
                jlxdef001         lerato.majara     mkuttel     nthomford      sancoop       susan.miller
adamwest        arghavan          chigorimbo.N      emeline.cadier  hksale001   hpc07  hpc15  hpc23  hpc31  hpc39
                jonathan.ipser    lgdlet001         mlhmic002  nyaria_exome   sea           timothy
aesterhuizen    bchmar018         cissarow          emma.rocke      hlnman006   hpc08  hpc16  hpc24  hpc32  hpc40
                katie             lmbjas002         mlnleo005  oldtem001      serena.illig  tsewell
akoch           bmongwane         clinton.moodley   eragumika       hpc01       hpc09  hpc17  hpc25  hpc33  hpc_hu
mgen_scratch    kevin.sack        lmllin001         mmaoyi     psych_gen       shkzay003     wlsath001
alecia.naidu    bpb9              crrlen001         gerrit          hpc02       hpc10  hpc18  hpc26  hpc34  ihalo
                kmwaikono         mamana            mnynol006  ptgmat003       sinaye        wrg
alewis          brianwillis       djmmou001         gjackson        hpc03       hpc11  hpc19  hpc27  hpc35  imane
                krksam004         mario.jonas       mskset001  rdctak001       snxmon002     ynegishi
andani.mulelu   carine            dmatten           grskir002       hpc04       hpc12  hpc20  hpc28  hpc36  irnara
001             krtale002         melnel000         nglhan001  rndroj001       sprtim002
```

The mounts look OK now.

Maybe you can now give the image a try and if I need to add something else to the container you now have instructions on how to get to a Singularity image from a Dockerfile.

Regards,
Gerrit

#19 - 04/12/2018 05:28 PM - Katie Lennard

Thanks Gerrit, much appreciated - I'll go through this and do some testing on the new image.

#20 - 04/16/2018 03:55 PM - Katie Lennard

I've tested the custom build singularity image described above. This specific repository (https://github.com/nf-core/RNAseq) that has been forked to create https://github.com/uct-cbio/RNAseq has an issue with picking up the uct_hex.conf file (even though it is there) that doesn't occur in the author's personal github site at https://github.com/ewels/nf-core-RNAseq. Will get in touch with him to try resolve this

#21 - 04/16/2018 04:31 PM - Katie Lennard

Gerrit I see even though you noticed that the uct_hex.conf was only on the dev branch it looks like the master branch may have gotten pulled? Because our current UCT repo https://github.com/uct-cbio/RNAseq doesn't have the uct_hex profile config. Phil has now also included it on the master branch he says but we would have to pull to update our branch - should we do that and then just add the lines:

RUN mkdir -p /researchdata/fhgfs
RUN mkdir -p /scratch

back to the docker file - would that work? (and then specify the custom built singularity image on hex with the -with singularity flag)..

#22 - 04/16/2018 04:44 PM - Gerrit Botha

Hi Katie,

Everything is on the dev branch: https://github.com/uct-cbio/RNAseq/tree/dev . So should things not work? You can also push the code from there into our master branch if you really want it in the master.

Regards,
Gerrit

#23 - 04/16/2018 04:50 PM - Katie Lennard

Aha, yes you're right - if we specify -r dev with the run (using our repo fork) it does pick up the uct_hex profile. For general updates from the original NGI-RNAseq I guess we can just pull and update as necessary (and then re-edit only docker file? I'm assuming the singularity image that you built only needs to happen once even if we update the rest?)

#24 - 04/16/2018 05:05 PM - Gerrit Botha

For now I do not think we will make changes on the main code. So yes, all that we would need to do is merge the main branch of NGI-RNAseq into our repos. Because our type of changes we are minor there would not be conflicts when we do the merge. If there is we can sort it over Redmine or Slack.

I do however think is that we should merge the docker file and hex config into our master branch. Because that is the two files that makes our repos different to NGI-RNAseq. Will chat to you over Slack about that and then we can later report back to Redmine on what our final structure is.

#25 - 04/17/2018 02:26 PM - Gerrit Botha

Hi Katie.

It seems like I found the issue why it was not submitting to PBS. The executer setting should have been in the process configuration part. See here.

I then restarted the job.

It seems that you need to remove the old version of the code or specify the correct version of the branch you are using. Otherwise it still uses your old repo code in the run. So I did a

```
rm -rf /home/gerrit/.nextflow/assets/uct-cbio/
```

Then restarted.

```
/opt/exp_soft/cbio/nextflow/nextflow -log /researchdata/fhgfs/gerrit/rnaseq/nextflow.log run uct-cbio/RNAseq
-r dev --reads "/researchdata/fhgfs/gerrit/rnaseq/reads/*_R{1,2}.fastq.gz" --genome GRCh37 -profile uct_hex -w
ith-singularity /scratch/DB/bio/singularity-containers/uct-cbio-rnaseq.img --outdir /researchdata/fhgfs/gerrit
/rnaseq/nf-outdir -w /researchdata/fhgfs/gerrit/rnaseq/nf-workdir --email gerrit.botha@uct.ac.za
```

Run output

```
N E X T F L O W  ~  version 0.28.0
Pulling uct-cbio/RNAseq ...
 downloaded from https://github.com/uct-cbio/RNAseq.git
Launching `uct-cbio/RNAseq` [nauseous_lovelace] - revision: 4030eeff38 [dev]
===================================
 nfcore/RNAseq  ~  version 1.5dev
===================================
Run Name       : nauseous_lovelace
Reads          : /researchdata/fhgfs/gerrit/rnaseq/reads/*_R{1,2}.fastq.gz
Data Type      : Paired-End
Genome         : GRCh37
Strandedness   : None
Trim R1        : 0
Trim R2        : 0
Trim 3' R1     : 0
Trim 3' R2     : 0
Aligner        : STAR
STAR Index     : /scratch/DB/bio/rna-seq/references/Homo_sapiens/Ensembl/GRCh37/Sequence/STARIndex/
GTF Annotation : /scratch/DB/bio/rna-seq/references/Homo_sapiens/Ensembl/GRCh37/Annotation/Genes/genes.gtf
BED Annotation : /scratch/DB/bio/rna-seq/references/Homo_sapiens/Ensembl/GRCh37/Annotation/Genes/genes.bed
Save Reference : No
Save Trimmed   : No
Save Intermeds : No
Max Memory     : 128 GB
Max CPUs       : 16
Max Time       : 10d
Output dir     : /researchdata/fhgfs/gerrit/rnaseq/nf-outdir
Working dir    : /researchdata/fhgfs/gerrit/rnaseq/nf-workdir
Container      : /scratch/DB/bio/singularity-containers/uct-cbio-rnaseq.img
Pipeline Release: dev
Current home   : /home/gerrit
Current user   : gerrit
Current path   : /home/gerrit
R libraries    : false
Script dir     : /home/gerrit/.nextflow/assets/uct-cbio/RNAseq
Config Profile : uct_hex
E-mail Address : gerrit.botha@uct.ac.za
=========================================
[warm up] executor > pbs
[warm up] executor > local
[bf/522312] Submitted process > get_software_versions
[d3/454a7e] Submitted process > workflow_summary_mqc
[6d/be0bd7] Submitted process > fastqc (sample38)
[e7/3097be] Submitted process > fastqc (sample39)
[77/40b2ac] Submitted process > trim_galore (sample38)
[30/0c36f2] Submitted process > trim_galore (sample39)
```

You will see some jobs have the local and some the pbs executer. I've investigated the get_software_versions and workflow_summary_mqc jobs are being run locally. All others are send to the queue.

Check the queue

```
qstat
```

```
Job id                    Name             User            Time Use S Queue
```

```
------------------------ ---------------- --------------- -------- - -----
1833463.srvslshpc001       STDIN            gerrit          01:23:17 R UCTlong
1837921.srvslshpc001       STDIN            gerrit          00:02:24 R UCTlong
1838670.srvslshpc001       sample38         gerrit          00:02:13 R UCTlong
1838671.srvslshpc001       sample39         gerrit          00:02:13 R UCTlong
1838672.srvslshpc001       sample38         gerrit          00:04:04 R UCTlong
1838673.srvslshpc001       sample39         gerrit          00:03:52 R UCTlong
```

The jobs you see is a fastqc run for sample38 and sample38. Also a trim_galore run for sample38 and sample39. Would be nice to tag it better.

Eventually the trim_galore jobs failed because of memory restrictions. We need to assign more cores to the job (because of the 1 core / 2GB RAM requirement on Hex). I've added the maxRetries = 4 (see here) to our uct_hex.config. It should overwrite the base.config.

I then did another delete of what we have in the nextflow repos and did a restart.

```
rm -rf /home/gerrit/.nextflow/assets/uct-cbio/
```

```
/opt/exp_soft/cbio/nextflow/nextflow -log /researchdata/fhgfs/gerrit/rnaseq/nextflow.log run uct-cbio/RNAseq
-r dev --reads "/researchdata/fhgfs/gerrit/rnaseq/reads/*_R{1,2}.fastq.gz" --genome GRCh37 -profile uct_hex -w
ith-singularity /scratch/DB/bio/singularity-containers/uct-cbio-rnaseq.img --outdir /researchdata/fhgfs/gerrit
/rnaseq/nf-outdir -w /researchdata/fhgfs/gerrit/rnaseq/nf-workdir --email gerrit.botha@uct.ac.za
```

You will now see that jobs are being resubmitted on retries.

```
...
...
[warm up] executor > pbs
[warm up] executor > local
[bf/522312] Submitted process > get_software_versions
[d3/454a7e] Submitted process > workflow_summary_mqc
[6d/be0bd7] Submitted process > fastqc (sample38)
[e7/3097be] Submitted process > fastqc (sample39)
[77/40b2ac] Submitted process > trim_galore (sample38)
[30/0c36f2] Submitted process > trim_galore (sample39)
[77/40b2ac] NOTE: Process `trim_galore (sample38)` terminated with an error exit status (143) -- Execution is
retried (1)
[30/0c36f2] NOTE: Process `trim_galore (sample39)` terminated with an error exit status (143) -- Execution is
retried (1)
[2e/477977] Re-submitted process > trim_galore (sample38)
[f7/518efc] Re-submitted process > trim_galore (sample39)
[2e/477977] NOTE: Process `trim_galore (sample38)` terminated with an error exit status (143) -- Execution is
retried (2)
[8a/30a52a] Re-submitted process > trim_galore (sample38)
[f7/518efc] NOTE: Process `trim_galore (sample39)` terminated with an error exit status (143) -- Execution is
retried (2)
[4f/02000b] Re-submitted process > trim_galore (sample39)
```

Will let you know how if it completes and goes on to the next steps.

Gerrit

#26 - 04/19/2018 01:11 PM - Gerrit Botha

I've been looking a few things over the last 2 days. Just did not get time to record it.

The maxRetries setting did work in terms of updating the submissions script with a request for more cores. However the setting

```
clusterOptions = { "-M $params.email -m abe -l nodes=1:ppn=1:series600" }
```

overwrites what Nextflow automates in setting ppn. You will for example see this in your header.

```
#!/bin/bash
#PBS -N sample38
#PBS -o /researchdata/fhgfs/gerrit/rnaseq/nf-workdir/ac/20eb832bb2dcf2353330bc8a35733e/.command.log
#PBS -j oe
#PBS -q UCTlong
#PBS -l nodes=1:ppn=2
#PBS -l walltime=08:00:00
#PBS -l mem=16gb
#PBS -M gerrit.botha@uct.ac.za -m abe -l nodes=1:ppn=1:series600
```

Now the only way for a job to recognise Nextflow's ppn request is to ignore the hardcoded setting of setting `ppn and modify the setting to

```
clusterOptions = { "-M $params.email -m abe" }
```

This causes a problem because we loose the series600 flag. I've contacted Andrew to know how to get about that and is it possible to add that flag in another way. He replied and said that flag is being added by a wrapper script. So very specific to the Hex cluster so that jobs go to the correct series. Not native to PBS. He however removed that requirement for us (or everyone on the cluster).

I submitted and trim_galore jobs complete. It started doint allignment with STAR however complains about not able to access the reference.

```
/scratch/DB/bio/rna-seq/references/Homo_sapiens/Ensembl/GRCh37/Sequence/STARIndex -- Cause: java.nio.file.Acce
ssDeniedException: /
scratch/DB/bio/rna-seq/references/Homo_sapiens/Ensembl/GRCh37/Sequence/STARIndex
```

I've asked Kated to open up permissions and will then resume the run.

Katie something I notice. If I do a

```
nextflow pull nf-core/RNAseq -r dev
```

And start my run it is still using the dev branch version of my older commit.

So I'm still doing

```
rm -rf /home/gerrit/.nextflow/assets/uct-cbio/
```

And restart/resume my run so that it uses the latest version of my code.

#27 - 04/20/2018 12:53 PM - Gerrit Botha

The STAR job on sample38 failed again.

```
PBS Job Id: 1841542.srvslshpc001
Job Name:   sample39
Exec host:  srvslshpc603/56+srvslshpc603/54+srvslshpc603/53+srvslshpc603/52+srvslshpc603/31+srvslshpc603/30+sr
vslshpc603/29+srvslshpc603/28+srvslshpc603/27+srvslshpc603/26+srvslshpc603/25+srvslshpc603/24+srvslshpc603/23+
srvslshpc603/22+srvslshpc603/21+srvslshpc603/20
job deleted
Job deleted at request of root@srvslshpc001
Job 1841542.srvslshpc001 terminated as it used too much RAM (36.9 GB) for the core assignment. Please resubmit
 with ppn=19. Please read the section on Memory control under http://hpc.uct.ac.za/index.php/hex-3/ if this ha
ppen repeatedly
```

I can coninue increasing maxRetries to get the job trhough but the best sollution is to get things configures so that it complies to the 2GB RAM per core ratio required on Hex.

Let me get some of the process configurations we have here https://github.com/uct-cbio/RNAseq/blob/dev/conf/base.config and move it to here https://github.com/uct-cbio/RNAseq/blob/dev/conf/uct_hex.config . I will then reconfigure the ppn requirements per job in uct_hex.config it will overwrite what we have in base.config.

#28 - 04/20/2018 04:15 PM - Gerrit Botha

I've added STAR requirements to use 40 cores on the first run attempt. With 40 cores we will be able to access 80GB RAM which was the default in the base.conf. It is strange the so much RAM is required becuase they note here https://github.com/alexdobin/STAR#hardwaresoftware-requirements 30GB RAM is required for human.

Update is here: https://github.com/uct-cbio/RNAseq/commit/8006f8d1c232f886d92ba836c01c0dbccc3b282e

Restarted run

```
rm -rf /home/gerrit/.nextflow/assets/uct-cbio/
```

```
/opt/exp_soft/cbio/nextflow/nextflow -log /researchdata/fhgfs/gerrit/rnaseq/nextflow.log run uct-cbio/RNAseq
-r dev --reads "/researchdata/fhgfs/gerrit/rnaseq/reads/*_R{1,2}.fastq.gz" --genome GRCh37 -profile uct_hex -w
ith-singularity /scratch/DB/bio/singularity-containers/uct-cbio-rnaseq.img --outdir /researchdata/fhgfs/gerrit
/rnaseq/nf-outdir -w /researchdata/fhgfs/gerrit/rnaseq/nf-workdir --email gerrit.botha@uct.ac.za -resume
```

#29 - 04/20/2018 05:19 PM - Gerrit Botha

The core settings were stuck to ppn=16. This was because the default max core settings were set to 16. Added a default section to https://github.com/uct-cbio/RNAseq/commit/89204179ff8981c51f2f5aeaf9fcac877a21a179 which are now in line with what we have in terms of max resources on Hex. Mem settings will however be ignored by PBS.

Restarted run.

#30 - 04/23/2018 11:42 AM - Gerrit Botha

STAR alignment on sample38 and 39 completed successfully. It now however failed on running Picard MarkDuplicates. Cannot figure out from PBS mails (did not get any) or Nexflow log why it really failed. Will just restart the job with -resume and try to debug it from there.

#31 - 04/23/2018 01:03 PM - Gerrit Botha

OK the MarkDuplicates process ran out of memory. I added a section here: https://github.com/uct-cbio/RNAseq/commit/718fb1aff1f374cb3c6e2af0fe976d353c544737 so that it starts with a default of allowable mem for the job of 16GB RAM on Hex.

This actually works

```
nextflow pull uct-cbio/RNAseq -r dev
```

and then doing a rerun on the code.

I previously updated nf-core/RNAseq and that was why it was not working.

Started run

```
...
=========================================
[warm up] executor > pbs
[6d/be0bd7] Cached process > fastqc (sample38)
[99/9c4655] Cached process > trim_galore (sample38)
[e7/3097be] Cached process > fastqc (sample39)
[2a/e2a59e] Cached process > trim_galore (sample39)
[warm up] executor > local
[03/84facf] Cached process > star (sample39)
[76/f6cc96] Cached process > star (sample38)
          Passed alignment > star (sample38)   >> 97.36% <<
[aa/862601] Cached process > markDuplicates (sample38AlignedByCoord.out)
          Passed alignment > star (sample39)   >> 97.57% <<
[9e/7d43da] Cached process > featureCounts (sample38AlignedByCoord.out)
[8b/b294f7] Cached process > preseq (sample38AlignedByCoord.out)
[04/594c1c] Cached process > stringtieFPKM (sample38AlignedByCoord.out)
[b4/a8f0b3] Cached process > rseqc (sample38AlignedByCoord.out)
[a1/b6a570] Cached process > genebody_coverage (sample38AlignedByCoord.out)
[bd/997e86] Cached process > stringtieFPKM (sample39AlignedByCoord.out)
[e8/b55c4e] Cached process > featureCounts (sample39AlignedByCoord.out)
[b9/85c12e] Cached process > rseqc (sample39AlignedByCoord.out)
[41/5fc1a2] Cached process > preseq (sample39AlignedByCoord.out)
[21/9b5792] Cached process > genebody_coverage (sample39AlignedByCoord.out)
[23/6c7a3f] Cached process > merge_featureCounts (sample38AlignedByCoord.out_gene.featureCounts)
[cc/ef2e0b] Cached process > dupradar (sample38Aligned.sortedByCoord.out.markDups)
[f6/77092f] Submitted process > get_software_versions
[5e/0e2729] Submitted process > workflow_summary_mqc
[50/552125] Submitted process > markDuplicates (sample39AlignedByCoord.out)
```

It would be easier to track errors if we are able to rename the PBS jobs according to the naming given by Nextflow.

Lets see if it now gets past markDuplicates.

#32 - 04/23/2018 05:46 PM - Gerrit Botha

The run completed successfully.

Here was the rest of the screen logs.

```
....
[50/552125] NOTE: Process `markDuplicates (sample39AlignedByCoord.out)` terminated with an error exit status (
143) -- Execution is retried (1)
[35/2fce04] Re-submitted process > markDuplicates (sample39AlignedByCoord.out)
[58/a9b05a] Submitted process > dupradar (sample39Aligned.sortedByCoord.out.markDups)
[5b/e3791e] Submitted process > multiqc (sample39_R1)
[02/ec042d] Submitted process > output_documentation (sample39_R1)
[nfcore/RNAseq] Sent summary e-mail to gerrit.botha@uct.ac.za (mail)
[nfcore/RNAseq] Pipeline Complete
```

The ---outdir is also now populated by tool and pipeline stats. You will see these folders

Hi Katie,

I suggest that you now try and rerun. We can then inspect all the results in the output dir and try to make sense of it.

Some additional things that might needs some work.

1. Not all of the processes have been configured to complete on their first run. These would probably need to be resubmitted one or two times (this will be done automatically by Nextflow). If you can please note those processes that fail. I will then later reconfigure uct-hex.config for those.

2. The nextflow script send some additional mails using sendmail it seems that the ports are closed on the compute nodes so those mails do not go through. You can still do the run but I'm just going to mail Andrew to check if he can open the ports.

Do your run based on mine below.

```
nextflow pull uct-cbio/RNAseq -r dev
```

```
/opt/exp_soft/cbio/nextflow/nextflow -log /researchdata/fhgfs/gerrit/rnaseq/nextflow.log run uct-cbio/RNAseq
-r dev --reads "/researchdata/fhgfs/gerrit/rnaseq/reads/*_R{1,2}.fastq.gz" --genome GRCh37 -profile uct_hex -w
ith-singularity /scratch/DB/bio/singularity-containers/uct-cbio-rnaseq.img --outdir /researchdata/fhgfs/gerrit
/rnaseq/nf-outdir -w /researchdata/fhgfs/gerrit/rnaseq/nf-workdir --email gerrit.botha@uct.ac.za -resume
```

Gerrit

#33 - 04/24/2018 03:18 PM - Gerrit Botha

Andrew has enabled sendmail to send emails from the compute nodes. I tested

```
gerrit@srvslshpc602:~> echo "Subject: sendmail test" | /usr/sbin/sendmail -v gerrit.botha@uct.ac.za
Mail Delivery Status Report will be mailed to <gerrit>.
```

When I get a mail it is from gerrit@srvslshpc613.uct.ac.za, so some routing is going on but the mail is being send so would be OK.

#34 - 04/25/2018 10:51 AM - Katie Lennard

- File Gmail - [nfcore_RNAseq] Successful_ mad_ptolemy.pdf added

- File nextflow.log.txt added

Test run completed successfully! Input and output at /researchdata/fhgfs/katie/NGI-RNAseq-test/ (permissions open to cbio group)
Email report and nextflow.log attached. There seems to have been only 1 retry - for markDuplicates:

Apr-24 14:45:18.269 [Task monitor] INFO nextflow.processor.TaskProcessor - [a5/d4109c] NOTE: Process markDuplicates (sample40AlignedByCoord.out) terminated with an error exit status (143) -- Execution is retried (1)

#35 - 04/25/2018 11:55 AM - Gerrit Botha

HI Katie,

This is good news, thanks for testing a complete run from beginning to end.

Also , thanks for opening up permissions and including the logs.

I have a specific section for the MarkDuplicates process included already. As you mentioned it failed only on one sample and was then resubmitted because it needed >16GB RAM but <32GB RAM. This is fine. If we find that get a higher proportion of samples that fail on our next runs I will consider increasing the ppn on the first attempt for this process.

For now

1. Can you go through the results and make sense of the whole protocol. We need to understand most of the settings / reports before we run it on the real data. Maybe you can create a new ticket specifically for that. I can help with this but if you can start in the mean time.
2. I see that our repos is now quite behind the nf-core. I'm going to check how we can update ours and decide on a protocol for future updates.

Is it OK if I rename the repos from RNASeq to RNASeq-pipeline on the uct-cbio GitHub repos. I just want to keep a consistent naming for pipelines and non-pipelines withing the organisation. If I change the naming all that you would need to do is change the origin naming in you git config.

Regards,
Gerrit

#36 - 04/25/2018 12:57 PM - Katie Lennard

Thanks Gerrit,

Will go through the results and see at which steps we might want to adjust parameters. Yes it's fine to change the repo name to be consistent with other pipelines thanks.

K

#37 - 05/17/2018 02:54 PM - Gerrit Botha

Katie I've renamed the repos to https://github.com/uct-cbio/RNAseq-pipeline . In your Git config you just need to modify the origin path.

Files

Gmail - [nfcore_RNAseq] Successful_ mad_ptolemy.pdf	115 KB	04/25/2018	Katie Lennard