Bioinformatics Pipeline Frameworks

Stephen Kelly

NYU Langone Medical Center

November 13, 2017

Advantages

- Standardized format
- Modular
- Consistency
- Automate the Easy Stuff... More Easily!

snsxt

- https://github.com/NYU-Molecular-Pathology/snsxt
- http://snsxt.readthedocs.io/en/latest/snsxt.html

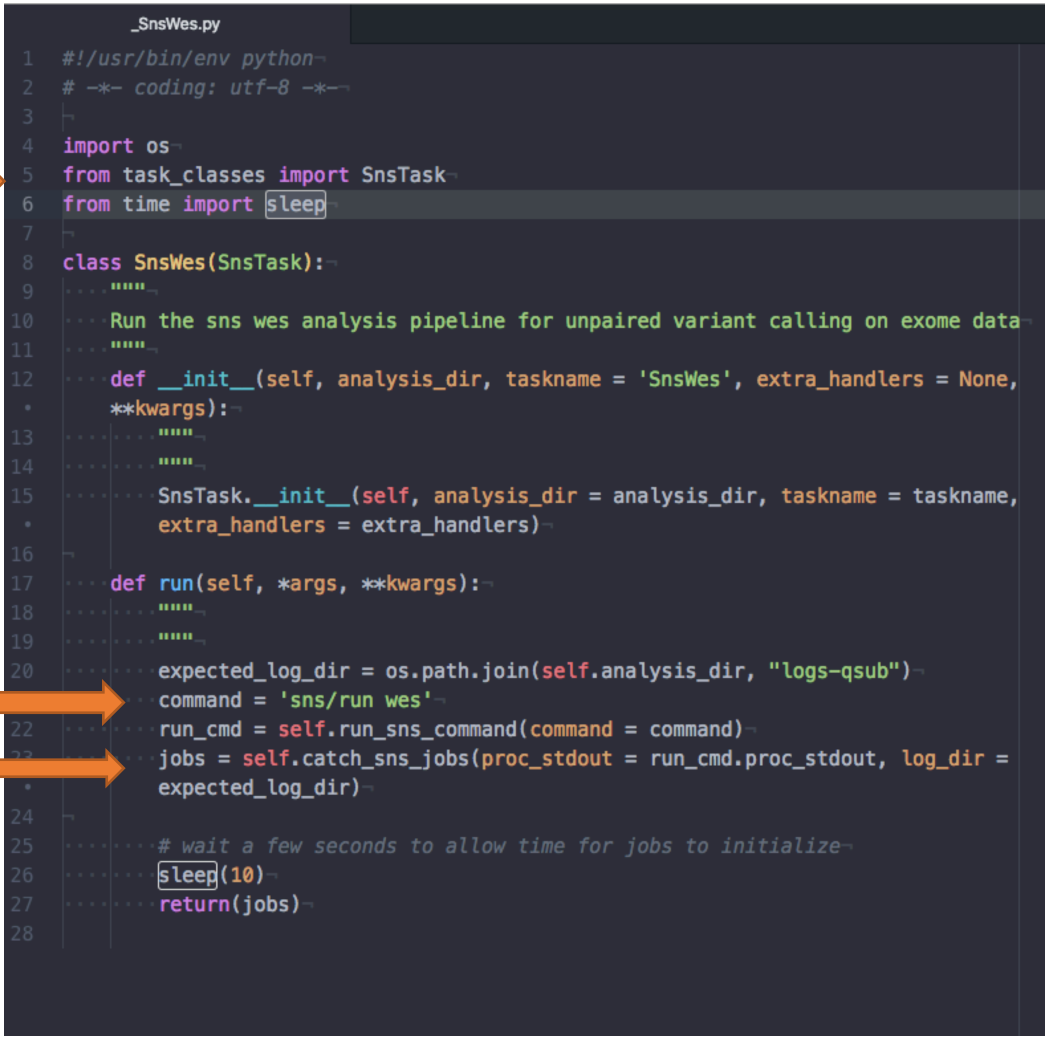
- Modular, object oriented, written in Python
- Easy wrapper around misc. analysis tasks
  - run Python code directly
  - run external scripts
- Easy to wrap entire pipelines
  - built-in wrapper around Igor’s sns pipeline (original purpose)
- **Tailored for easy usage on NYULMC phoenix HPC cluster**
  - modules: Python 2.7+, pandoc 1.13, batteries included; **clone & run**
  - `$ snsxt/run.py -d analysis_dir/ -f fastq_dir/`

Example sns task – sns wes pipeline

import base class

run external `sns` command

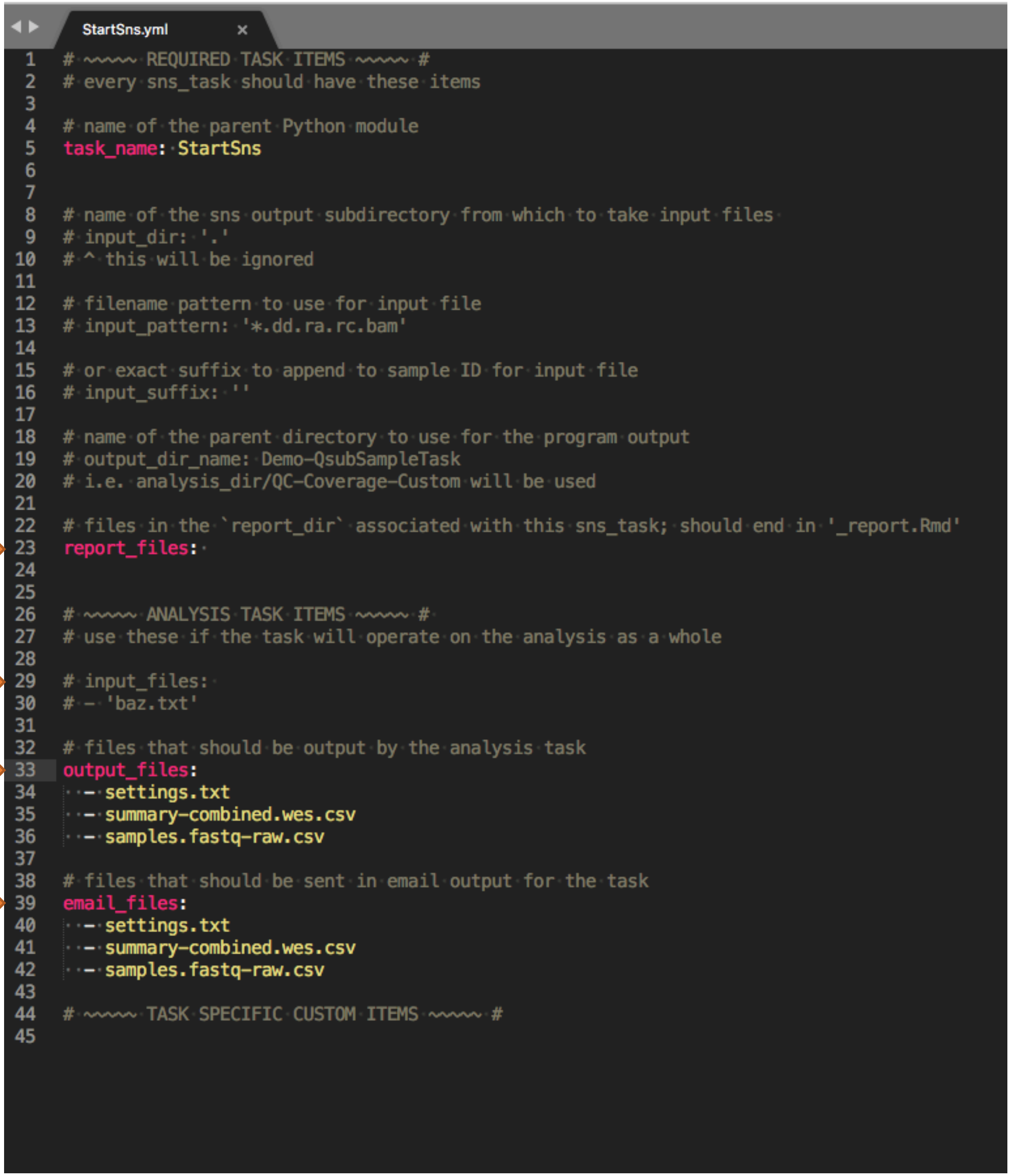
capture qsub jobs



_SnsWes.py

```python
#!/usr/bin/env python
# -*- coding: utf-8 -*-

import os
from task_classes import SnsTask
from time import sleep

class SnsWes(SnsTask):
    """
    Run the sns wes analysis pipeline for unpaired variant calling on exome data
    """
    def __init__(self, analysis_dir, taskname = 'SnsWes', extra_handlers = None,
    **kwargs):
        """
        """
        SnsTask.__init__(self, analysis_dir = analysis_dir, taskname = taskname,
        extra_handlers = extra_handlers)

    def run(self, *args, **kwargs):
        """
        """
        expected_log_dir = os.path.join(self.analysis_dir, "logs-qsub")
        command = 'sns/run wes'
        run_cmd = self.run_sns_command(command = command)
        jobs = self.catch_sns_jobs(proc_stdout = run_cmd.proc_stdout, log_dir =
        expected_log_dir)

        # wait a few seconds to allow time for jobs to initialize
        sleep(10)
        return(jobs)

```

External config



```yaml
# ~~~~~ REQUIRED TASK ITEMS ~~~~~ #
# every sns_task should have these items

# name of the parent Python module
task_name: StartSns


# name of the sns output subdirectory from which to take input files
# input_dir: '.'
# ^ this will be ignored

# filename pattern to use for input file
# input_pattern: '*.dd.ra.rc.bam'

# or exact suffix to append to sample ID for input file
# input_suffix: ''

# name of the parent directory to use for the program output
# output_dir_name: Demo-QsubSampleTask
# i.e. analysis_dir/QC-Coverage-Custom will be used

# files in the `report_dir` associated with this sns_task; should end in '_report.Rmd'
report_files: 


# ~~~~~ ANALYSIS TASK ITEMS ~~~~~ #
# use these if the task will operate on the analysis as a whole

# input_files:
# - 'baz.txt'

# files that should be output by the analysis task
output_files:
  - settings.txt
  - summary-combined.wes.csv
  - samples.fastq-raw.csv

# files that should be sent in email output for the task
email_files:
  - settings.txt
  - summary-combined.wes.csv
  - samples.fastq-raw.csv

# ~~~~~ TASK SPECIFIC CUSTOM ITEMS ~~~~~ #

```

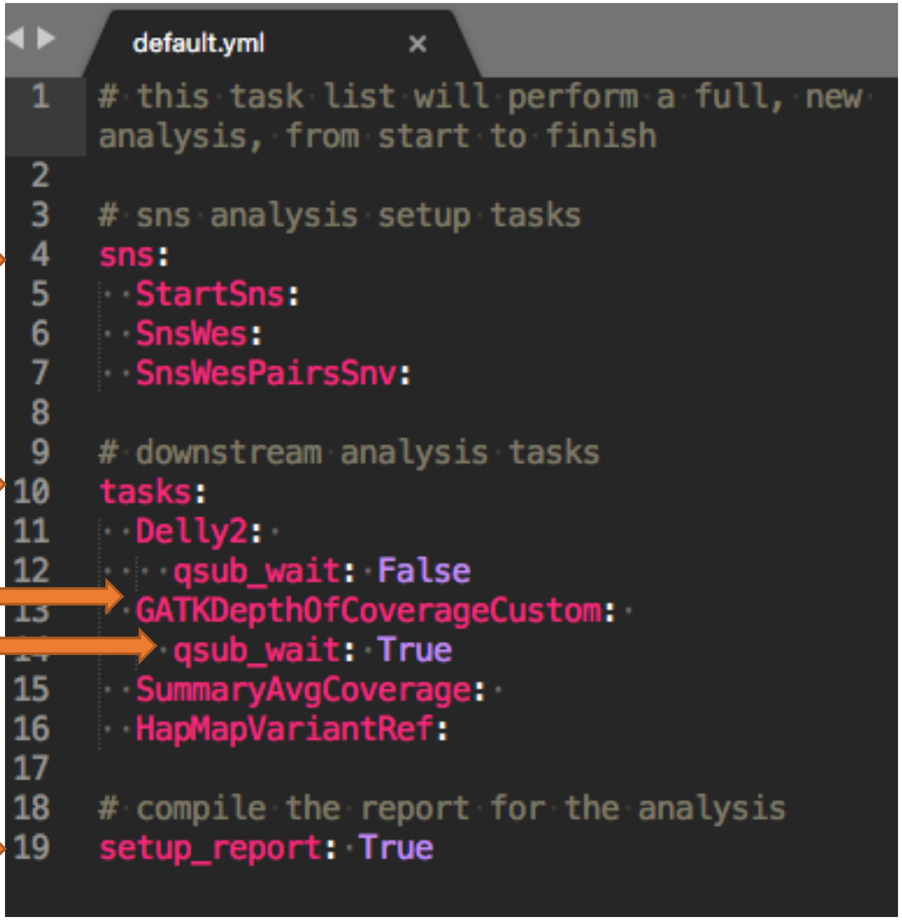
files to use for reporting

expected input files

expected output files

files to send in email output

YAML Task List



```yaml
# this task list will perform a full, new analysis, from start to finish

# sns analysis setup tasks
sns:
  StartSns:
  SnsWes:
  SnsWesPairsSnv:

# downstream analysis tasks
tasks:
  Delly2: 
    qsub_wait: False
  GATKDepthOfCoverageCustom: 
    qsub_wait: True
  SummaryAvgCoverage: 
  HapMapVariantRef:

# compile the report for the analysis
setup_report: True
```

start a new `sns` analysis

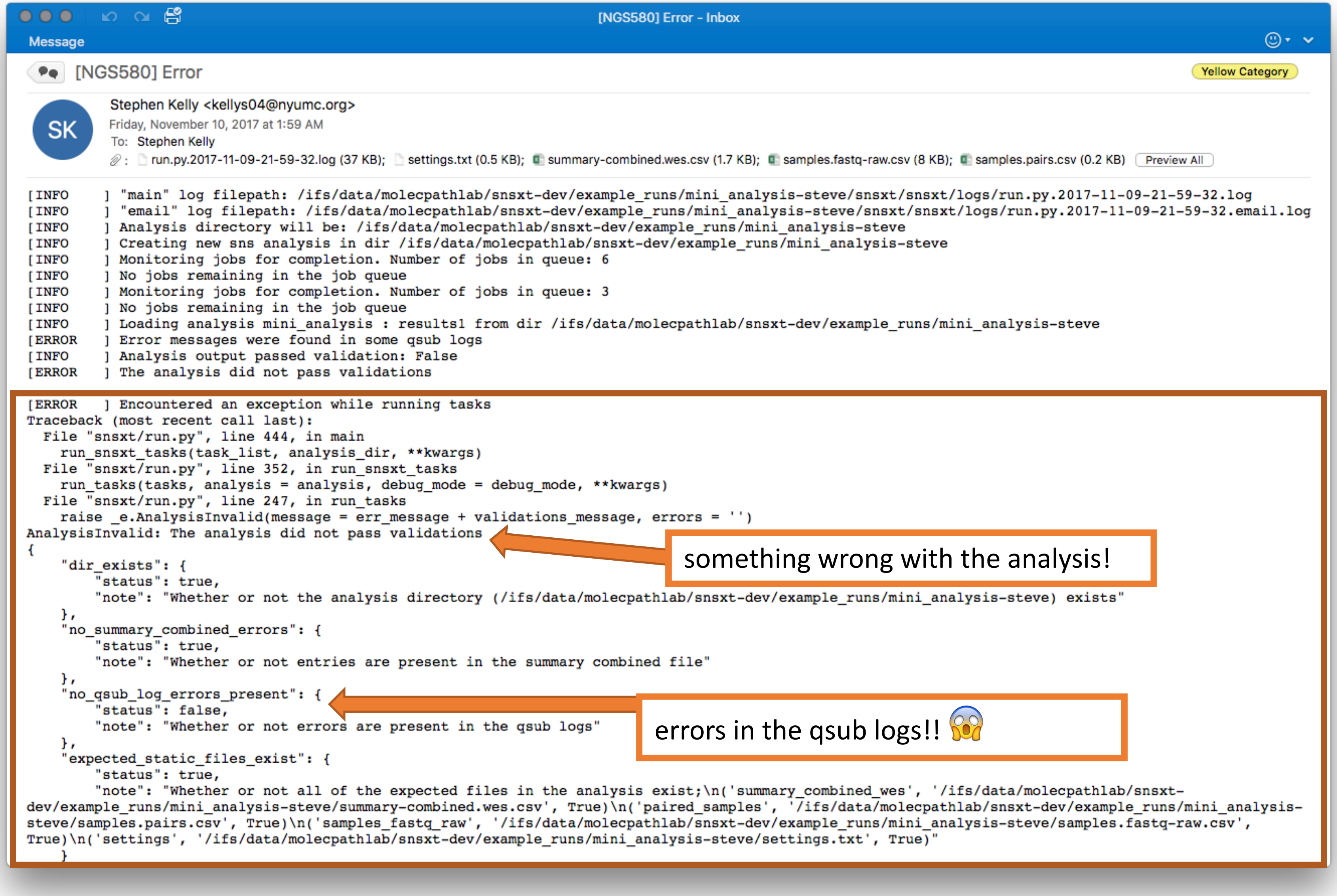
downstream tasks

task name = Python class name

extra args for the task’s `run()` method

compile report for the analysis

Email output – oops it broke!



[NGS580] Error

SK

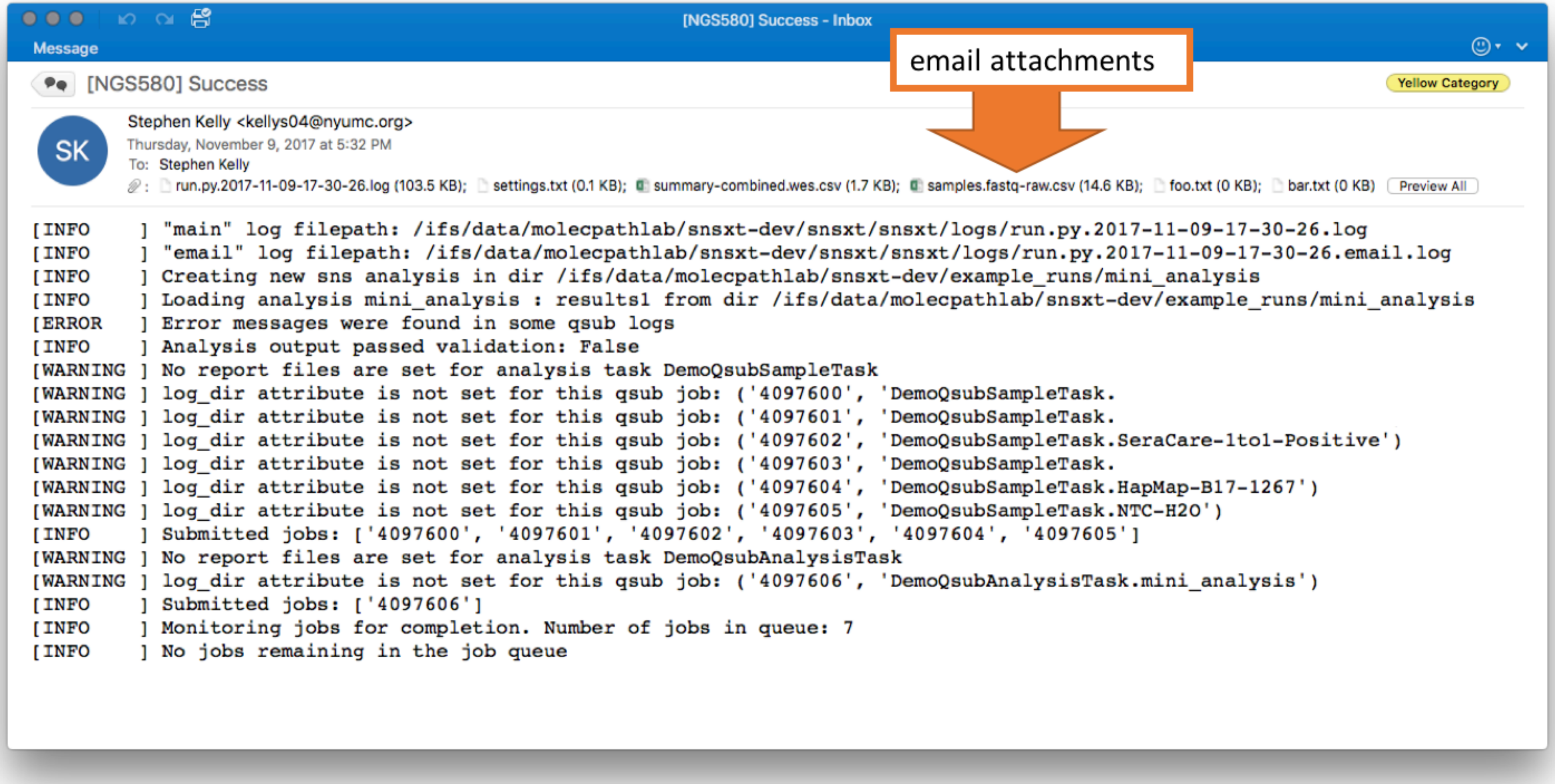
Stephen Kelly <kellys04@nyumc.org>
Friday, November 10, 2017 at 1:59 AM
To: Stephen Kelly
: run.py.2017-11-09-21-59-32.log (37 KB); settings.txt (0.5 KB); summary-combined.wes.csv (1.7 KB); samples.fastq-raw.csv (8 KB); samples.pairs.csv (0.2 KB)

```
[INFO    ] "main" log filepath: /ifs/data/molecpathlab/snsxt-dev/example_runs/mini_analysis-steve/snsxt/snsxt/logs/run.py.2017-11-09-21-59-32.log
[INFO    ] "email" log filepath: /ifs/data/molecpathlab/snsxt-dev/example_runs/mini_analysis-steve/snsxt/snsxt/logs/run.py.2017-11-09-21-59-32.email.log
[INFO    ] Analysis directory will be: /ifs/data/molecpathlab/snsxt-dev/example_runs/mini_analysis-steve
[INFO    ] Creating new sns analysis in dir /ifs/data/molecpathlab/snsxt-dev/example_runs/mini_analysis-steve
[INFO    ] Monitoring jobs for completion. Number of jobs in queue: 6
[INFO    ] No jobs remaining in the job queue
[INFO    ] Monitoring jobs for completion. Number of jobs in queue: 3
[INFO    ] No jobs remaining in the job queue
[INFO    ] Loading analysis mini_analysis : results1 from dir /ifs/data/molecpathlab/snsxt-dev/example_runs/mini_analysis-steve
[ERROR   ] Error messages were found in some qsub logs
[INFO    ] Analysis output passed validation: False
[ERROR   ] The analysis did not pass validations

[ERROR   ] Encountered an exception while running tasks
Traceback (most recent call last):
  File "snsxt/run.py", line 444, in main
    run_snsxt_tasks(task_list, analysis_dir, **kwargs)
  File "snsxt/run.py", line 352, in run_snsxt_tasks
    run_tasks(tasks, analysis = analysis, debug_mode = debug_mode, **kwargs)
  File "snsxt/run.py", line 247, in run_tasks
    raise _e.AnalysisInvalid(message = err_message + validations_message, errors = '')
AnalysisInvalid: The analysis did not pass validations
{
    "dir_exists": {
        "status": true,
        "note": "Whether or not the analysis directory (/ifs/data/molecpathlab/snsxt-dev/example_runs/mini_analysis-steve) exists"
    },
    "no_summary_combined_errors": {
        "status": true,
        "note": "Whether or not entries are present in the summary combined file"
    },
    "no_qsub_log_errors_present": {
        "status": false,
        "note": "Whether or not errors are present in the qsub logs"
    },
    "expected_static_files_exist": {
        "status": true,
        "note": "Whether or not all of the expected files in the analysis exist;\n('summary_combined_wes', '/ifs/data/molecpathlab/snsxt-
dev/example_runs/mini_analysis-steve/summary-combined.wes.csv', True)\n('paired_samples', '/ifs/data/molecpathlab/snsxt-dev/example_runs/mini_analysis-
steve/samples.pairs.csv', True)\n('samples_fastq_raw', '/ifs/data/molecpathlab/snsxt-dev/example_runs/mini_analysis-steve/samples.fastq-raw.csv',
True)\n('settings', '/ifs/data/molecpathlab/snsxt-dev/example_runs/mini_analysis-steve/settings.txt', True)"
    }
```

something wrong with the analysis!

errors in the qsub logs!! 😱

Email output – yay it worked!



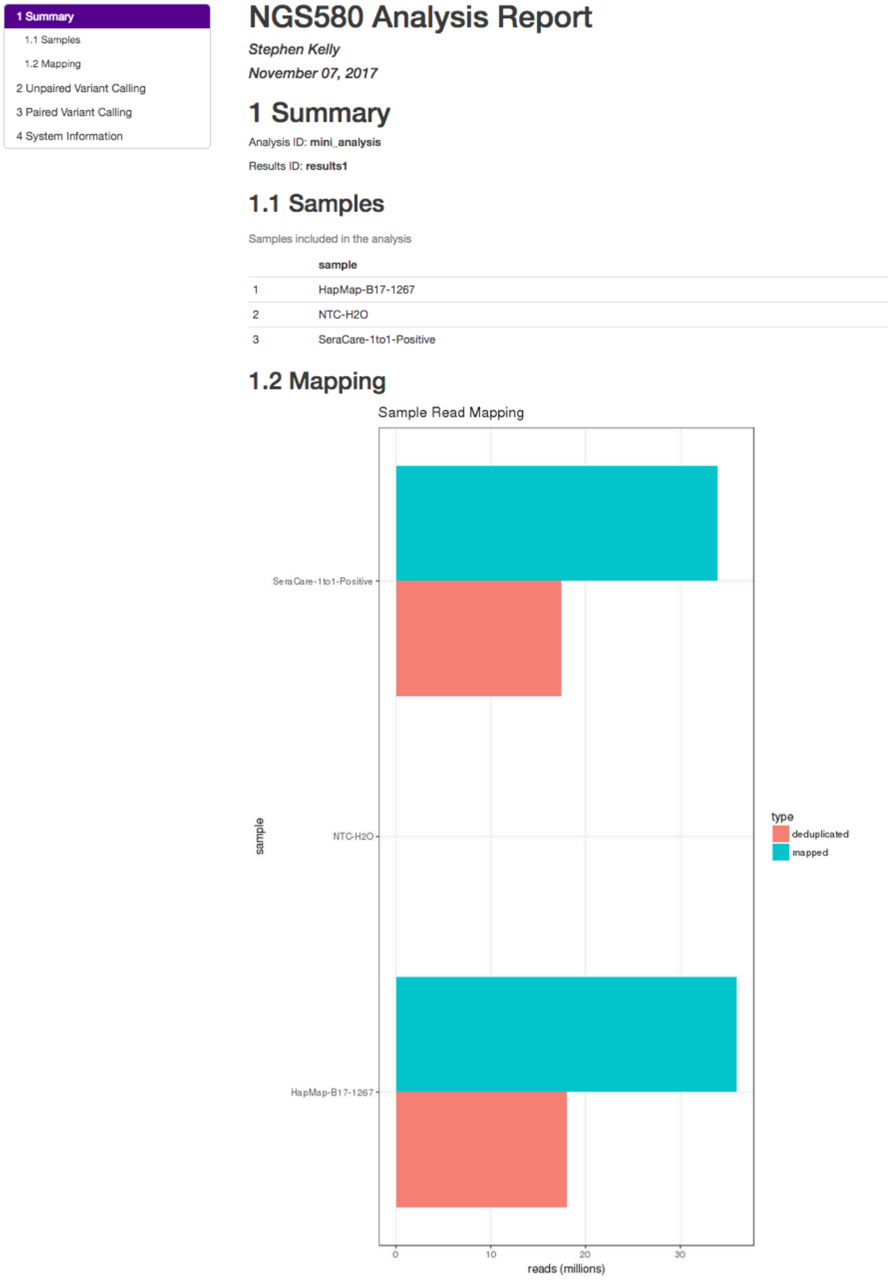
[NGS580] Success

Stephen Kelly <kellys04@nyumc.org>
Thursday, November 9, 2017 at 5:32 PM
To: Stephen Kelly
run.py.2017-11-09-17-30-26.log (103.5 KB); settings.txt (0.1 KB); summary-combined.wes.csv (1.7 KB); samples.fastq-raw.csv (14.6 KB); foo.txt (0 KB); bar.txt (0 KB)

```
[INFO    ] "main" log filepath: /ifs/data/molecpathlab/snsxt-dev/snsxt/snsxt/logs/run.py.2017-11-09-17-30-26.log
[INFO    ] "email" log filepath: /ifs/data/molecpathlab/snsxt-dev/snsxt/snsxt/logs/run.py.2017-11-09-17-30-26.email.log
[INFO    ] Creating new sns analysis in dir /ifs/data/molecpathlab/snsxt-dev/example_runs/mini_analysis
[INFO    ] Loading analysis mini_analysis : results1 from dir /ifs/data/molecpathlab/snsxt-dev/example_runs/mini_analysis
[ERROR   ] Error messages were found in some qsub logs
[INFO    ] Analysis output passed validation: False
[WARNING ] No report files are set for analysis task DemoQsubSampleTask
[WARNING ] log_dir attribute is not set for this qsub job: ('4097600', 'DemoQsubSampleTask.
[WARNING ] log_dir attribute is not set for this qsub job: ('4097601', 'DemoQsubSampleTask.
[WARNING ] log_dir attribute is not set for this qsub job: ('4097602', 'DemoQsubSampleTask.SeraCare-1to1-Positive')
[WARNING ] log_dir attribute is not set for this qsub job: ('4097603', 'DemoQsubSampleTask.
[WARNING ] log_dir attribute is not set for this qsub job: ('4097604', 'DemoQsubSampleTask.HapMap-B17-1267')
[WARNING ] log_dir attribute is not set for this qsub job: ('4097605', 'DemoQsubSampleTask.NTC-H2O')
[INFO    ] Submitted jobs: ['4097600', '4097601', '4097602', '4097603', '4097604', '4097605']
[WARNING ] No report files are set for analysis task DemoQsubAnalysisTask
[WARNING ] log_dir attribute is not set for this qsub job: ('4097606', 'DemoQsubAnalysisTask.mini_analysis')
[INFO    ] Submitted jobs: ['4097606']
[INFO    ] Monitoring jobs for completion. Number of jobs in queue: 7
[INFO    ] No jobs remaining in the job queue
```

Reporting

- Modular, extensible report framework
  - R Markdown + pandoc
- custom report per pipeline task
- parent report imports & compiles all child report docs



1 Summary
1.1 Samples
1.2 Mapping
2 Unpaired Variant Calling
3 Paired Variant Calling
4 System Information

NGS580 Analysis Report

Stephen Kelly

November 07, 2017

1 Summary

Analysis ID: **mini_analysis**

Results ID: **results1**

1.1 Samples

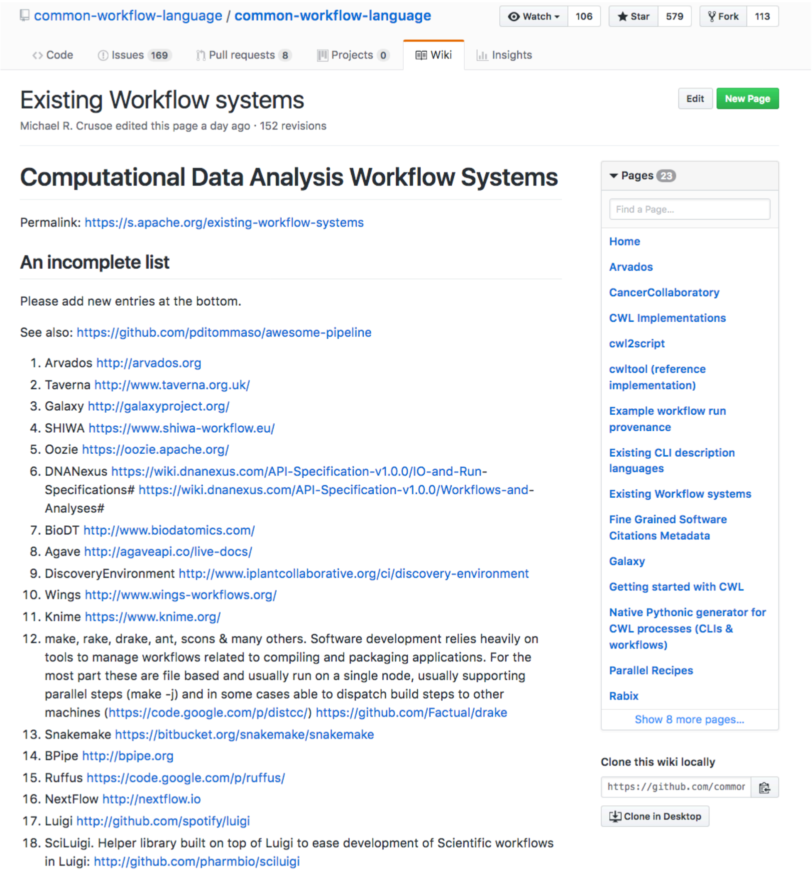
Samples included in the analysis

	sample
1	HapMap-B17-1267
2	NTC-H2O
3	SeraCare-1to1-Positive

1.2 Mapping

Sample Read Mapping

Other Frameworks


common-workflow-language / common-workflow-language
Watch 106
Star 579
Fork 113
Code
Issues 169
Pull requests 8
Projects 0
Wiki
Insights
Edit
New Page
Pages 23
Find a Page...
Home
Arvados
CancerCollaboratory
CWL Implementations
cwl2script
cwltool (reference implementation)
Example workflow run provenance
Existing CLI description languages
Existing Workflow systems
Fine Grained Software Citations Metadata
Galaxy
Getting started with CWL
Native Pythonic generator for CWL processes (CLIs & workflows)
Parallel Recipes
Rabix
Show 8 more pages...
Clone this wiki locally
https://github.com/commor
Clone in Desktop

Existing Workflow systems

Michael R. Crusoe edited this page a day ago · 152 revisions

Computational Data Analysis Workflow Systems

Permalink: https://s.apache.org/existing-workflow-systems

An incomplete list

Please add new entries at the bottom.

See also: https://github.com/pditommaso/awesome-pipeline

1. Arvados http://arvados.org
2. Taverna http://www.taverna.org.uk/
3. Galaxy http://galaxyproject.org/
4. SHIWA https://www.shiwa-workflow.eu/
5. Oozie https://oozie.apache.org/
6. DNANexus https://wiki.dnanexus.com/API-Specification-v1.0.0/IO-and-Run-Specifications# https://wiki.dnanexus.com/API-Specification-v1.0.0/Workflows-and-Analyses#
7. BioDT http://www.biodatomics.com/
8. Agave http://agaveapi.co/live-docs/
9. DiscoveryEnvironment http://www.iplantcollaborative.org/ci/discovery-environment
10. Wings http://www.wings-workflows.org/
11. Knime https://www.knime.org/
12. make, rake, drake, ant, scons & many others. Software development relies heavily on tools to manage workflows related to compiling and packaging applications. For the most part these are file based and usually run on a single node, usually supporting parallel steps (make -j) and in some cases able to dispatch build steps to other machines (https://code.google.com/p/distcc/) https://github.com/Factual/drake
13. Snakemake https://bitbucket.org/snakemake/snakemake
14. BPipe http://bpipe.org
15. Ruffus https://code.google.com/p/ruffus/
16. NextFlow http://nextflow.io
17. Luigi http://github.com/spotify/luigi
18. SciLuigi. Helper library built on top of Luigi to ease development of Scientific workflows in Luigi: http://github.com/pharmbio/sciluigi

- Advantages:
  - might have more features
- Disadvantages:
  - how to implement?
  - how to use?
  - how to apply to our Bioinformatic pipelines?

https://github.com/common-workflow-language/common-workflow-language/wiki/Existing-Workflow-systems

Conclusion

- snsxt's advantages
  - already works on NYULMC phoenix HPC out of the box
  - no installation needed, single module adjustment
    - `module load python/2.7`
  - standardized format to wrap existing pipelines & analysis output
  - builds off of our current code base
  - extensible for new pipelines