Arvados - Task #15997

Story # 15996 (Resolved): human WGS fastq to gvcf variant calling with GATK4 CWL (with report)

Review

01/06/2020 05:32 PM - Peter Amstutz

Status:	Resolved	**Start date:**	01/29/2020
Priority:	Normal	**Due date:**	
Assigned To:	Nico César	**% Done:**	0%
Category:		**Estimated time:**	0.00 hour
Target version:	2020-09-09 Sprint		

Description

History

#1 - 01/29/2020 03:20 PM - Peter Amstutz

- Remaining (hours) set to 0.0

- Start date set to 01/29/2020

- Status changed from New to Resolved

#2 - 01/29/2020 03:20 PM - Peter Amstutz

- Estimated time set to 0.00 h

- Status changed from Resolved to New

#3 - 06/25/2020 10:26 PM - Jiayong Li

1. Mostly looks good to me. There're a lot of white spaces through out workflow, would look better to customers if we remove them.
2. There is style inconsistency for invoking requirement.
For example, in wgs-processing-wf.cwl,

```
  - class: SubworkflowFeatureRequirement
```

In gather-vcf.cwl,

```
  InlineJavascriptRequirement: {}
```

These two styles are both valid, but having both present could be confusing to customers. The latter style is recommented.
4. In gatk-haplotypecaller-with-interval.cwl, $(runtime.outdir) is not necessary, since it's the default output directory.
5. GATK uses java run time environment, and max heap size can be correlated with ram requirement by using $(runtime.ram).
For example, in gatk-splitintervals.cwl, we have

```
  ResourceRequirement:
    ramMin: 5000
```

Max heap size can be specified as

```
  "-Xmx$(runtime.ram)M"
```

Similarly, in samtools-sort.cwl, '2G' can be specified as '$(runtime.ram)M'. The advantage of this is when resource requirement changes, we don't have to modify the cwl arguments.
6. Resource requirement need adjustment, for example, in mark-duplicates.cwl, we have

```
  ResourceRequirement:
    ramMin: 20000
```

But max heap size is "-Xmx8G". There is a lot of unused ram here.
7. InitialWorkDirRequirement is used for when input file needs to be modified, like samtools-index.cwl. Using it when unnecessary could cause confusion. For example, I don't think it's necessary to use it in fastqc.cwl and gatk-applyBSQR-with-interval.cwl. There might be others.
8. In gatk-wf-with-interval.cwl and scatter-gatk-wf-with-interval.cwl, variable "knownsites1" could be simply called "knownsites" (in

bwamem-gatk-report-wf.cwl, it's called "knownsites"). Variable naming could cause confusion.

#4 - 06/25/2020 10:41 PM - Sarah Zaranek

Thanks Jiayong

1. For white space, I try to use it for separating major sections like inputs/outputs/etc - I think it really helps with parsing the code. You can see something similar here:
https://www.commonwl.org/user_guide/20-software-requirements/index.html. I think I am consistent between files when the same white spacing. I will try to check that though.

2. I can fix that to the later.

3. I don't see a 3 :)

4. Will fix that. Good catch.

5. I think we have to be careful since the RAM is being used for other things so if we set heap to be the same size it might ultimately freak out on us since we don't have swapping set up. I can check in with Tom/Peter about this. For samtools-sort, we have to be careful since that is per thread, so we have to divide by threads being used hence why is is smaller since it is multi-core machine. I will check in and try perhaps to do something smarter there. Ultimately, I think it is nice to do this in a more automatic way - I just don't want to over prescribe the RAM.

6. I can check for mark duplicates as well.

7. OK, I will check out and fix those initial work directory unnecessary uses.

8. Yup, I intended to maybe put in multiple knownsites but ended up sticking with 1 of them - so I will take away that 1.

#5 - 06/26/2020 04:00 PM - Jiayong Li

setting max heap to $(runtime.ram) is usually ok, since it is equal to the ramMin you specify in the cwl. if ramMin is 13000, arvados will take into account the overhead, and assign a node with at least ~15000 ram. since max heap is 13000, it will never exceed total ram.
however, this still needs an experiment run to make sure it actually works. the usually workflow is to run crunchstat summary on all jobs, adjust resource, and then rerun to confirm. if you're under time pressure, maybe skip this and make a ticket that does this later.

#6 - 07/03/2020 01:06 PM - Sarah Zaranek

Finished --
1. Misunderstood, Jiayong meant white space not spacing, fixed (I think) all extra white spacing at end of lines (hopefully!)
8. Fixed this.

In the works --
2. Fixed for 1 file and will work on replacing it with InlineJavascriptRequirement: {} today
4. In gatk-haplotypecaller-with-interval.cwl, $(runtime.outdir) is not necessary, since it's the default output directory. --- Need to look into this and fix it
5. I will up the Java requirements but for now I think I will not do the automatic thing just for time concerns. I will circle back and fix it later.
7. Fixed this for 1 case and will fix the rest...

Also Done --

- Finished annotating rest of files
- Swapped out fastq docker file for "standard" ones from biocontainers

To do:

- Swap out bwa docker file for standard from biocontainers if it will let me bgzip
- Swap out reference to one without alt except has the decoys -- https://lh3.github.io/2017/11/13/which-human-reference-genome-to-use
  Note: if you want to include alts you need and want to do it correctly -- you need to use bwakit not just bwa-mem --
  https://github.com/lh3/bwa/blob/master/README-alt.md
- Need to regenerate reference dictionaries,etc needed for GATK --
- Download rest of the set of fastqs
- Rerun to make sure this still works

#7 - 07/03/2020 01:07 PM - Sarah Zaranek

Oh - and updated the format for the report to collect variants by "type".

#8 - 07/07/2020 12:45 AM - Sarah Zaranek

In the works --
DONE -- 2. Fixed for 1 file and will work on replacing it with InlineJavascriptRequirement: {} today
DONE --- 4. In gatk-haplotypecaller-with-interval.cwl, $(runtime.outdir) is not necessary, since it's the default output directory. --- Need to look into this and fix it
DONE --7. Fixed this for 1 case and will fix the rest...

Still to do 5. I will up the Java requirements but for now I think I will not do the automatic thing just for time concerns. I will circle back and fix it later.

Currently downloading fastqs, need to download 10 at time to not hit downloading limits so it will take abit of time.

#9 - 08/12/2020 03:42 PM - Peter Amstutz

- Assigned To changed from Peter Amstutz to Nico César

#10 - 08/12/2020 04:10 PM - Nico César

- Status changed from New to In Progress

#11 - 08/17/2020 08:17 PM - Nico César

Good news: I successfully ran the tutorial from webshell. I'm making some extra comments/changes in the doc

#12 - 08/17/2020 09:18 PM - Nico César

- Status changed from In Progress to Resolved

I think on my side we can close this ticket, if there is something to review I can re-open it.