Arvados - Story #17049

Make a single chromosome run of the demo

10/28/2020 02:44 PM - Peter Amstutz

Status:	Resolved		**Start date:**	
Priority:	Normal		**Due date:**	
Assigned To:	Jiayong Li		**% Done:**	0%
Category:			**Estimated time:**	0.00 hour
Target version:				

Description

https://github.com/arvados/arvados-tutorial

Run 5 samples on su92l, on a single chromosome.

Related issues:

Related to Arvados Epics - Story #17001: WGS Demo Works in WB2 and Is Updated...	**New**	**01/01/2022**	**02/28/2022**

History

#1 - 10/28/2020 02:44 PM - Peter Amstutz

- Related to Story #17001: WGS Demo Works in WB2 and Is Updated to Reflect WB2 Features added

#2 - 11/05/2020 08:50 PM - Jiayong Li

- Tracker changed from Bug to Story

#3 - 12/01/2020 10:03 PM - Jiayong Li

- Description updated

#4 - 12/15/2020 10:29 PM - Jiayong Li

- Status changed from New to Feedback

- Description updated

branch 17049-make-singlechrom, commit d2f6d47da38e96de4597bb58ba1193e261af8fb1

I finished a run for chr19 input fastqs for 5 samples.
https://workbench.su92l.arvadosapi.com/container_requests/su92l-xvhdp-3dqeqodgi1usk0r

I encountered a gatk bug that the memory usage spikes when writing result vcf in the haplotypecaller step, if the region is outside of chr19. (And in general, in regions when very few reads are placed.) I had to schedule 14G ram machines for that step to complete it. (See failure https://workbench.su92l.arvadosapi.com/container_requests/su92l-xvhdp-v0sq9csiz1p94nx)

Also I had to upload a different docker image for curii/clinvar-report to keep in order to finish the gvcf-to-vcf step (see failure https://workbench.su92l.arvadosapi.com/container_requests/su92l-xvhdp-zulp4y3e9f7tv59). The new docker image is https://workbench.su92l.arvadosapi.com/collections/su92l-4zz18-kl40n4746vmjnw2

Input fastqs here https://workbench.su92l.arvadosapi.com/collections/su92l-4zz18-j1jotbx4uqckkec
The run for generating the chr19 input fastqs is https://workbench.su92l.arvadosapi.com/container_requests/su92l-xvhdp-ilo79fgj728afkl

#5 - 12/16/2020 03:05 PM - Peter Amstutz

So the goal of this was to have a relatively lightweight pipeline that still had all the pieces of the whole genome pipeline.

I think this still needs a little work to be lightweight --

I see dozens of haplotype caller instances that are each running for 1-2 minutes. Can we consolidate that?

Same for selectvariants/basecalibrator/applyBSQR. Can we either (a) not scatter or (b) use the RunInSingleContainer feature.

#6 - 12/28/2020 10:11 PM - Jiayong Li

branch 17049-make-singlechrom, commit d2f6d47da38e96de4597bb58ba1193e261af8fb1

I changed the fullintervallist to reflect chr19 only calling, hopefully this run is better.

https://workbench.su92l.arvadosapi.com/container_requests/su92l-xvhdp-1l4x28f6q1rn955

#7 - 01/14/2021 10:50 PM - Jiayong Li

- Status changed from Feedback to Resolved

commit d147d1d1fafeeea06bd09d9479337b0f5aab43b0

Added comments for the singe chrom chr19 yml.

#7 - 01/14/2021 10:50 PM - Jiayong Li

- Status changed from Feedback to Resolved

commit d147d1d1fafeeea06bd09d9479337b0f5aab43b0