Tapestry - Feature #8700

Arvados job/pipeline generates GET-Evidence HTML report from VCF

03/15/2016 05:59 PM - Tom Clegg

Status:	Closed	**Start date:**	03/15/2016
Priority:	Normal	**Due date:**	
Assigned To:	Abram Connelly	**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:	Interpretation automation		

Description

Subtasks:

Task # 8705: Make sure VCF input files are processed properly	**Closed**
Task # 8706: Generate HTML from GET-Evidence JSON report	**Closed**
Task # 8859: Review #8700	**Resolved**

History

#1 - 03/15/2016 08:30 PM - Tom Clegg

- Subject changed from Arvados job/pipeline generates GET-Evidence HTML report to Arvados job/pipeline generates GET-Evidence HTML report from VCF

#2 - 03/15/2016 09:30 PM - Tom Clegg

- Story points set to 2.0

#3 - 03/15/2016 09:56 PM - Ward Vandewege

- Assigned To set to Abram Connelly

#4 - 03/22/2016 09:32 PM - Abram Connelly

- Status changed from New to In Progress

The VCF to GFF conversion is incomplete so I'm updating it. After a brief discussion with Ward and Brad, having some simple tests to make sure some known variants that are in ClinVar also appear through the GET-Evidence pipeline is probably a good idea.

#5 - 03/28/2016 09:57 PM - Abram Connelly

Still in the process of testing but the initial report creation and the HTML creation are now Arvados Git repositories. The 'refresh report' step still needs to be created as a separate Git repo and they all should be transferred to my GitHub account for easy public viewing.

#6 - 03/29/2016 08:38 PM - Abram Connelly

Sample report has been generated from the 'abe.vcf' file. The output collection can be seen at:

https://workbench.su92l.arvadosapi.com/collections/su92l-4zz18-306h2x9n4fvcug9

and a sample report can be seen at:

https://workbench.su92l.arvadosapi.com/collections/5ba11becdece9f5007d4474d843479b4+1882/get-evidence-report.html

The above was for pipeline instance su92l-d1hrv-ky6qzn44wd256zo from pipeline template su92l-p5p6p-2hdis7roemcaokr.

#7 - 03/31/2016 03:01 PM - Ward Vandewege

Abram Connelly wrote:

> Sample report has been generated from the 'abe.vcf' file. The output collection can be seen at:
>
> https://workbench.su92l.arvadosapi.com/collections/su92l-4zz18-306h2x9n4fvcug9
>
> and a sample report can be seen at:
>
> https://workbench.su92l.arvadosapi.com/collections/5ba11becdece9f5007d4474d843479b4+1882/get-evidence-report.html

> The above was for pipeline instance su92l-d1hrv-ky6qzn44wd256zo from pipeline template su92l-p5p6p-2hdis7roemcaokr.

Thanks Abram! Review comments:

- sorting on the Allele Freq column doesn't seem to work right: it's doing a numeric sort on the first few digits, instead of on the % (which is what I expected it would do).
- I think we need some sort of indication that data is being loaded. Looking at one of the reports, it takes several seconds for the data to load. Until that happens, the page looks strange and broken, and there's no indication that something is happening. The simplest solution is probably some text that says that the data is being loaded, which disappears when the table loads. Or even a spinning ball kind of indicator in addition.
- I believe Workbench sorts the files in a project by putting the most recently modified/created first. Perhaps if you do an extra 'touch' at the very end on the report html file in the html generation step, that one will be listed at the top. Is that hard to do?
- A process thing: please move the nice description you put on the review ticket to this ticket; we use a convention to have all information on the main ticket, not on review tickets. That way everything is in one place.

#8 - 03/31/2016 05:16 PM - Abram Connelly

As requested:

> The GET-Evidence pipeline is producing HTML reports for VCF and non-VCF (e.g. CGI-Var) data files.
>
> An example run with a data file that was converted to **VCF from 23andMe**:
>
> - https://workbench.su92l.arvadosapi.com/pipeline_instances/su92l-d1hrv-z9qiho889d52uid
> - https://workbench.su92l.arvadosapi.com/collections/e1fa7630e0762d1b22090fda681b827a+1882/get-evidence-report.html
>
> An example run with a date file that was originally a **CGI-Var file**:
>
> - https://workbench.su92l.arvadosapi.com/pipeline_instances/su92l-d1hrv-nhin4h27bqzjyvf
> - https://workbench.su92l.arvadosapi.com/collections/2b4703eb27aeb44eaaa320ae1ec67669+1925/get-evidence-report.html
>
> The pipeline template is su92l-p5p6p-2hdis7roemcaokr
>
> The three repositories for each leg of the pipeline can be found at:
>
> - https://github.com/abeconnelly/GETEvidenceReport
> - https://github.com/abeconnelly/GETEvidenceReportRefresh
> - https://github.com/abeconnelly/GETEvidenceReportHTML
>
> This closely resembles how the pipeline was setup. The initial report is generated with GETEvidenceReport, then refreshed with the latest GET-Evidence database by running GETEvidenceReportRefresh. Finally, the HTML is created with GETEvidenceReportHTML. The major differences from the previous incarnation of the GET-Evidence pipeline from this one are that this one is able to injects VCF and create a report and the final report resides completely in Arvados, rather than the GET-Evidence server.

- Sorting by allele frequency now functions properly. I parse the string 'Allele Freq.'. 'Unknown' is set to value 1.125 and other values are clamped to be in the range of [0,1]. This puts 'Unknown' Allele frequencies at the top or bottom of the list, depending on sort order.
- A spinner was added that disappears when the table is loaded.
- I've added some 'touch' statements to the pipeline but the workbench display order is still unchanged. From observation it looks to me like the order is directories first (alphabetically) then files (alphabetically).

A new sample run can be found at:

- https://workbench.su92l.arvadosapi.com/collections/478e56332cc341b028afdfa3848be664+2051/get-evidence-report.html

#9 - 03/31/2016 06:14 PM - Ward Vandewege

Thanks, much better now! Sorry about the touch thing, I'll have to double check that.

I updated the su92l-p5p6p-2hdis7roemcaokr template a little bit; the dataclass for some of the variables (INPUT_SAMPLE, GE_CONFIG and GETEV_LATEST) was set to 'Collection' but it had to be 'File'.

Other than that, I think this is good to go.

#10 - 04/14/2016 03:55 PM - Abram Connelly

- Status changed from In Progress to Closed