Lightning - Story #13425

Create CWL pipeline for VCF + BED file to gVCF conversion

04/30/2018 09:56 PM - Abram Connelly

Status:	Closed	**Start date:**	05/18/2018
Priority:	Normal	**Due date:**	
Assigned To:	Abram Connelly	**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:			

Description

Encapsulate the vcfbed2homref tool to create a CWL pipeline that converts an input VCF + BED file to a gVCF file (VCF 4.2+ with homozygous reference calls) as output.

There should be one VCF file output that is indexed.

Subtasks:

Task # 13470: review #13425	**Closed**

History

#1 - 05/03/2018 09:35 PM - Abram Connelly

- Status changed from New to In Progress

#2 - 05/07/2018 11:35 PM - Abram Connelly

Using the GIAB files to test. Taken from:

- NA12878_HG001

Specifically:

- NA128178_HG001 BED file
- NA128178_HG001 VCF file

The GIAB VCF file is one big VCF with all chromosomes appearing in it (instead of split out, say), so the CWL will be developed with this in mind. Once we start testing with a Veritas genome we might need to change this to work on each chromosome independently with a consolidation step afterwards.

Currently the vcfbed2homref expected the bed file to be uncompressed and the VCF file to be compressed and indexed.

#3 - 05/08/2018 12:09 AM - Abram Connelly

Converting the GIAB file takes roughly ~20min (single-cpu, minimal memory) with a resulting file size ~170Mb.

A cursory use of tabix seems to work fine, properly reporting the lines that fall within the start and END tag.

The next steps are:

- Confirm that it works under the Docker image
- Create the actual CWL pipeline and test submission
- Push the resulting gVCF through a gVCF to FastJ CWL pipeline to make sure the resulting FastJ is correct

#4 - 05/08/2018 09:53 AM - Abram Connelly

For testing, I uploaded the 'release' directories available through the ftp site. I don't know why but the "Chinese trio" only has a directory labelled "son".

I found some other references to GIAB data in keep but I didn't find the VCF plus BED files I needed.

The files were downloaded locally then uploaded to keep (~5Gb):

```
wget -r ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release
arv-put --name GIAB-release --project-uuid su92l-j7d0g-vo06mmd9dg6f5bc
 ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/
```

I've put the GIAB "release" data under a project called reference. The Data collection itself is called GIAB-release.

I created the CWL along with a test YAML and submission script. The test YAML file includes the VCF and BED files in the keep collection as well as the FASTA reference file.

In keeping with our loose convention of CWL, the project is organized as:

```
cwl-version/preprocess/vcf-bed-to-gvcf
  |
  |--- cwl/
  |
  |--- src/
  |
  |--- yml/
  |
  |--- cwl-run/
```

Where cwl contains the CWL file, src contains any scripts that are needed for this pipeline, yml includes a test YAML file and cwl-run includes the submission script to submit the test CWL pipeline to Arvados.

The Docker image has been updated to include the vcfbed2homref executable.

The test data was run (NA12878 with it's BED file) successfully and stored in su92l-4zz18-1bw0ykm3cia0x2z.

Requiring the FastJ conversion be successful might be out of scope for this ticket. I think it's best to close this ticket with what's available now and check the FastJ conversion works separately. If the FastJ conversion is unsuccessful, either because of quirks in the input data, a bad conversion or some other issue, it's probably best to open a new ticket.

#5 - 05/08/2018 10:23 PM - Abram Connelly

In addition, there should be a CWL workflow that scatters on multiple VCF files provided. The scatter workflow should be tested NA12878 as well as one of the other datasets provided from GIAB.

Testing locally to make sure the gVCF can be converted to FastJ passes some simple checks so I'm happy to call this ticket closed after the scatter workflow has been created and tested.

#6 - 05/10/2018 07:54 AM - Abram Connelly

The gather steps dumps all files flat into the directory. If this is undesirable we can put them in subdirectories labelled with their names.

#7 - 05/15/2018 11:35 PM - Abram Connelly

There's a collection with two VCF + BED files converted to their respective gVCF files that can be found here that was made from the HG004 and NA12878 GIAB datasets.

#8 - 05/18/2018 03:41 PM - Jiayong Li

Nitpicking: I noticed in the gvcf you've created the symbolic alt allele is "<NON-REF>" (i.e, this is a hom-ref block). In contrast, GATK uses "<NON_REF>" as the symbolic alt allele. Note that there are no standards regarding this in vcf specs (https://samtools.github.io/hts-specs/VCFv4.2.pdf). In fact, freebayes uses "." as the symbolic alt allele. But it might be nice to conform to a well know tool (e.g GATK) for the ease of downstream processing.

#9 - 05/18/2018 07:20 PM - Abram Connelly

vcfbed2homref has been updated to now report "<NON_REF>" with a command line option to override the default if necessary. The arvados/l7g Docker container has been updated to reflect the change.

I ran the CWL pipeline with the new vcfbed2homref and a sample run of two GIAB datasets can be seen in the su92l-4zz18-co5isboy7rcxtwu collection.

#10 - 05/18/2018 08:48 PM - Jiayong Li

lgtm, please merge.

#11 - 05/21/2018 03:47 AM - Abram Connelly

- Status changed from In Progress to Closed