

Automating Image Generation for Variant Call Review

In this exercise we will start with a VCF file and generate an igv batch script to make screenshots for each variant. After creating the batch file, we will run it from the igv tools menu and examine the resulting screenshots.

Data

Data for this exercise is located in the “data/snps” folder. It is derived from whole genome alignments from the 1000 genomes pilot project. The genome assembly used for these alignments is human “hg18”.

Overview

To create the batch file we need to transform a vcf file that looks like this

```
## Various header lines
#CHROM    POS    ID      REF    ALT    ...
1        63635328  .      G      A      ...
1        63647603  .      C      T      ....
...
```

into an igv batch file

```
goto 1:63635328
sort base
snapshot
goto 1:63647603
sort base
snapshot
...
```

Before beginning examine the variant.vcf file in a text editor. Our strategy will be to first remove the header lines, then use the chromosome and position information in the first 2 columns to generate “goto” statements. The sort and snapshot commands are then easily added. There are many ways to accomplish this, for this exercise we will use the unix utilities “grep” and “awk

Step by Step

Open a terminal window

Create a directory to dump images. This directory can be placed anywhere, but the instructions below assume it has been created in the user home directory.

mkdir ~/images

"cd" to the data/snps directory

Use grep to remove the header lines

grep -v ^# variants.vcf

Notes: the "-v" option causes grep to invert the match, i.e. lines not matching the pattern are output. The ^ symbol indicates that the line must start with the #, as opposed to finding # anywhere in the line.

The output from the grep command above should remove all header lines. Now pipe this output to awk and use the "print" statement to format and output the batch commands.

grep -v ^# variants.vcf | awk '{print "goto " \$1 ":" \$2 "\nsort base\nsnapshot"}'

The output looks like an igv batch file! It will work as is, but the filenames will be auto-generated from the genomic location. We can do a little better by generating our own file names incorporating the ref and alt columns.

grep -v ^# variants.vcf | awk '{print "goto " \$1 ":" \$2 "\nsort base\nsnapshot g." \$1 "_" \$2 "_" \$4 ">" \$5 ".png"}'

Now let's pipe the result to a file (use the up-arrow key to avoid retyping the command)

grep -v ^# variants.vcf | awk '{print "goto " \$1 ":" \$2 "\nsort base\nsnapshot g." \$1 "_" \$2 "_" \$4 ">" \$5 ".png"}' > igv.batch

If we run the script as-is it will dump the images into the <user home> directory. To specify a directory for the images edit the igv.batch in a plain text editor and add

the following line at the top (replace <user name> with your user name). NOTE: do not use textedit.

snapshotDirectory /Users/<user name>/images

Start IGV and load the bam file NA12878.SLX.sample.bam from the file menu (File > Load from file...)

Run the batch file from the tools menu (Tools > Run Batch Script)

Upon completion examine the snapshots in the "images" directory