Variant Calling with GATK

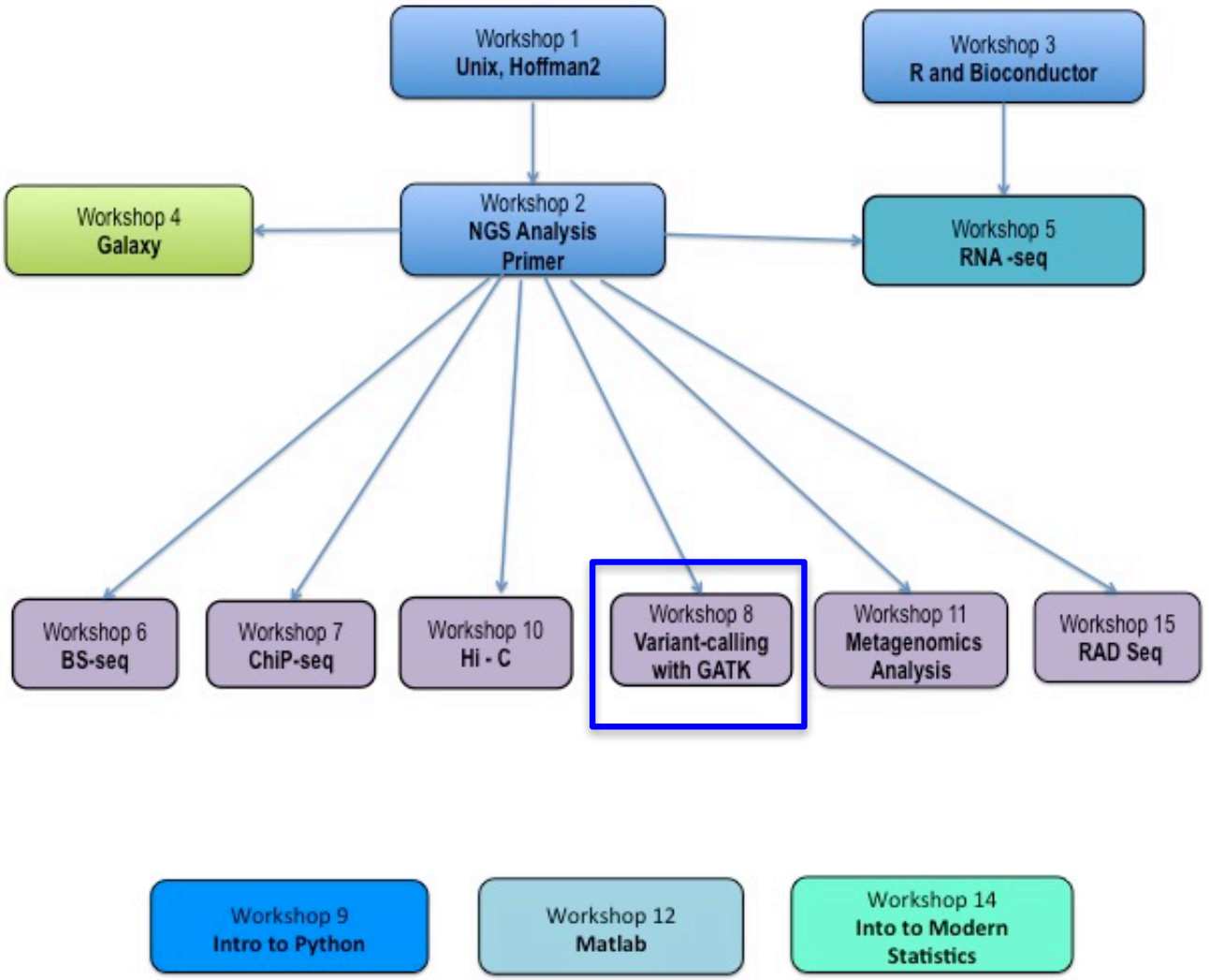
May 16-18 2017

Michael Weinstein (michael.weinstein@ucla.edu)

(covering for Sorel Fitz-Gibbon, sorel@ucla.edu)

I will be presenting materials from MIT and Harvard's The Broad Institute BroadE workshops. Further materials and videos of their presentations are available at the following Broad website:
https://software.broadinstitute.org/gatk/documentation/presentations.php

UCLA Collaboratory Workshops



http://qcb.ucla.edu/collaboratory/workshops/

Variant Calling with GATK- Day 1

- Hoffman2 setup
- Laptop/local computer setup
- GATK Primer
  - GATKwr17-01-Intro_to_Variant_Discovery.pdf
- Start GATK Tutorial, if time
  - Variant_Discovery_Tutorial.pdf

Variant Calling with GATK- Day 1

- Hoffman2 setup
- Laptop/local computer setup
- GATK Primer
  - GATKwr17-01-Intro_to_Variant_Discovery.pdf
- GATK Tutorial
  - Variant_Discovery_Tutorial.pdf

Request an interactive shell on hoffman and copy over the workshop materials to your hoffman directory…

- Log on to hoffman2, e.g.

```
ssh joebruin@hoffman2.idre.ucla.edu
```

- request an interactive shell

```
qrsh –l i,time=3:00:00,mem=4g
```

- When you have an interactive shell, think about where you want to keep your work, move there and then copy the gatkWorkshop directory from the below path. If you don't move anywhere, your files will be in your hoffman home directory, which is fine.

```
cp -pr /u/nobackup/galaxy/collaboratory/sorel/gatkWorkshop ./
```

-p preserve the timestamp information for the file

-r copy recursively the entire contents

Move into the workshop directory and look at what's there

```
cd gatkWorkshop/
ls
```

You should see:
1702 commands.txt gatk_profile slides.pdf

The "**1702**" directory is from the Broad Institute (https://software.broadinstitute.org/gatk/documentation/presentations.php), although the original 1702 directory at the Broad has additional files which I've deleted since we're not using them.

The file **commands.txt** has all of the command lines we'll use in the workshop. You can use it to cut and paste as needed. But make sure you pay attention to what you're cutting and pasting, and ask about any parts of the commands you don't understand. Take a look…

```
more commands.txt
```

The file **slides.pdf** has the powerpoint slides for all three days of the workshop. Note, they frequently refer to Broad presentation files which we will cover at least in part.

The file **gatk_profile** has commands to setup your environment on hoffman2 to efficiently use GATK, samtools and picard. Take a look…

```
more gatk_profile
```

You can activate these by running…

```
source gatk_profile
```

If you want these to be loaded every time you log in you can put the lines into your **$HOME/.bash_profile** (note the "." at the beginning of that file name).

Explore inside 1702, and move into the data directory

```
cd 1702
```

```
ls
```

data_bundle docs presentations

```
ls docs
ls presentations
ls data_bundle
```

Move into the data directory for the rest of the workshop

```
cd data_bundle/data
```

```
ls
```

You should see
bams gvcfs inputVcfs outputs ref sandbox

Test GATK

Cut & Paste only from commands.txt, not the pdf

Type the following to get the main GATK help page

```
java –Xmx1g -jar /u/nobackup/galaxy/collaboratory/apps/gatk/GenomeAnalysisTK.jar  \
--help
```

Note, GATK is written in java and must be run with the java –jar command.
The back slash at the end of the first line extends the command to the next line.

But this is unwieldy, so in the gatk_profile a shortcut was setup.
(export GATK=/u/nobackup/galaxy/collaboratory/apps/gatk/GenomeAnalysisTK.jar)

```
java –Xmx1g –jar $GATK  --help
```

GATK commands all follow this format…

Java –jar path/to/GenomeAnalysisTK.jar
-Xmx1g = allocate 1Gb memory to this job
-T toolName, in this case it's CountReads

--help can be added after any tool name to get parameter options for that tool or without a tool name to get general options for GATK.

If you get this error, you are probably not in an interactive shell...

```
Error occurred during initialization of VM
Could not reserve enough space for object heap
Error: Could not create the Java Virtual Machine.
Error: A fatal exception has occurred. Program will exit.
```

Test Setup

Type your first gatk command. CountReads is a command to count the number of reads in a bam file.

```
java –Xmx1g –jar $GATK  -T CountReads –R ref/ref.fasta –I bams/father.bam
```

Note, GATK commands all follow this format **AND** they almost always require a reference seqeunce to be specified, even commands that don't seem to need it.

Java –jar path/to/GenomeAnalysisTK.jar
–Xmx1g = allocate 2Gb memory to this job
–T toolName, in this case it's CountReads
–R reference.fasta

How many reads are in the father.bam file?

Variant Calling with GATK- Day 1

- Hoffman2 setup
- Laptop/local computer setup
- GATK Primer
  - GATKwr17-01-Intro_to_Variant_Discovery.pdf
- GATK Tutorial
  - Variant_Discovery_Tutorial.pdf

Set up local computer – laptop or desktop

- Make sure you have a working, somewhat recent version of IGV.
- If not, get it from

http://software.broadinstitute.org/software/igv/download

Set up local computer – laptop or desktop

Use cyberduck, filezilla, scp or another method to transfer **a copy** of the gatkWorkshop directory to your local computer.

- Think about where you put the directory on hoffman, get it from there.
- **Think** about where you want to put these files.
  - If you're using a collaboratory computer, you can put them in a directory/folder on the desktop. Please remember to delete them after the final day of the workshop.

e.g. if you're joebruin and you put the gatkWorkshop directory in your hoffman home directory, you could use this command to copy it into your current local location.

```
scp –pr joebruin@dtn2.hoffman2.idre.ucla.edu:gatkWorkshop ./
```

Variant Calling with GATK- Day 1

- Hoffman2 setup
- Laptop/local computer setup
- GATK Primer
  - GATKwr17-01-Intro_to_Variant_Discovery.pdf
- GATK Tutorial
  - Variant_Discovery_Tutorial.pdf

Variant Calling with GATK- Day 1

A brief introductory lecture following slides in **GATKwr17-01-Intro_to_Variant_Discovery.pdf**

- open it on your local computer from gatkWorkshop/1702/presentations/

This will just be a quick review as you should be familiar with much of this material from the pre-requisite workshops.

These pages/topics will be **skipped**

- Unmapped BAMs instead of FASTQ
- RNAseq
- Somatic SNVs, Indels and CNV.
- Pipelines

Variant Calling with GATK- Day 1

- Hoffman2 setup
- Laptop/local computer setup
- GATK Primer
  - GATKwr17-01-Intro_to_Variant_Discovery.pdf
- If time allows, start GATK Tutorial (see day2 slides)
  - Variant_Discovery_Tutorial.pdf