



Genomics in the Cloud

Book Club - Week 13

February 22, 2021

Agenda

- Chapter 12: Interactive Analysis in Jupyter Notebook
- Additional resources
- Open discussion



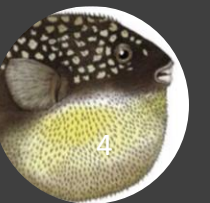
Chapter 12: Interactive Analysis in Jupyter Notebook

Genomics in the Cloud by Geraldine A. Van der Auwera and Brian D. O'Connor (O'Reilly). Copyright 2020 The Broad Institute, Inc. and Brian O'Connor, 978-1-491-97519-0.



Our guest
speaker

Dr. Joris
Vankerschaver



Jupyter notebook

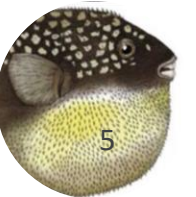
1.1 Hello Python

Let's try a basic Hello World example in Python.

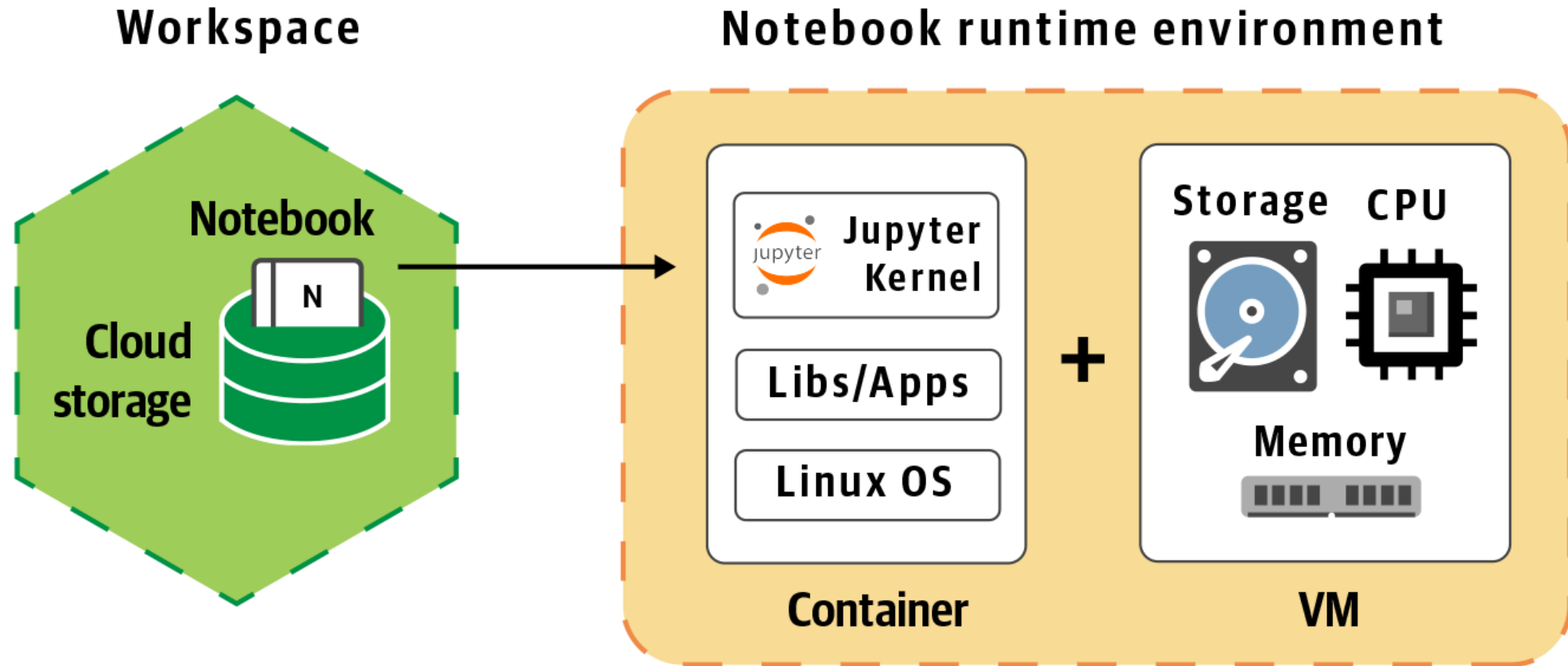
```
In [1]: print ("Hello World")
```

Hello World

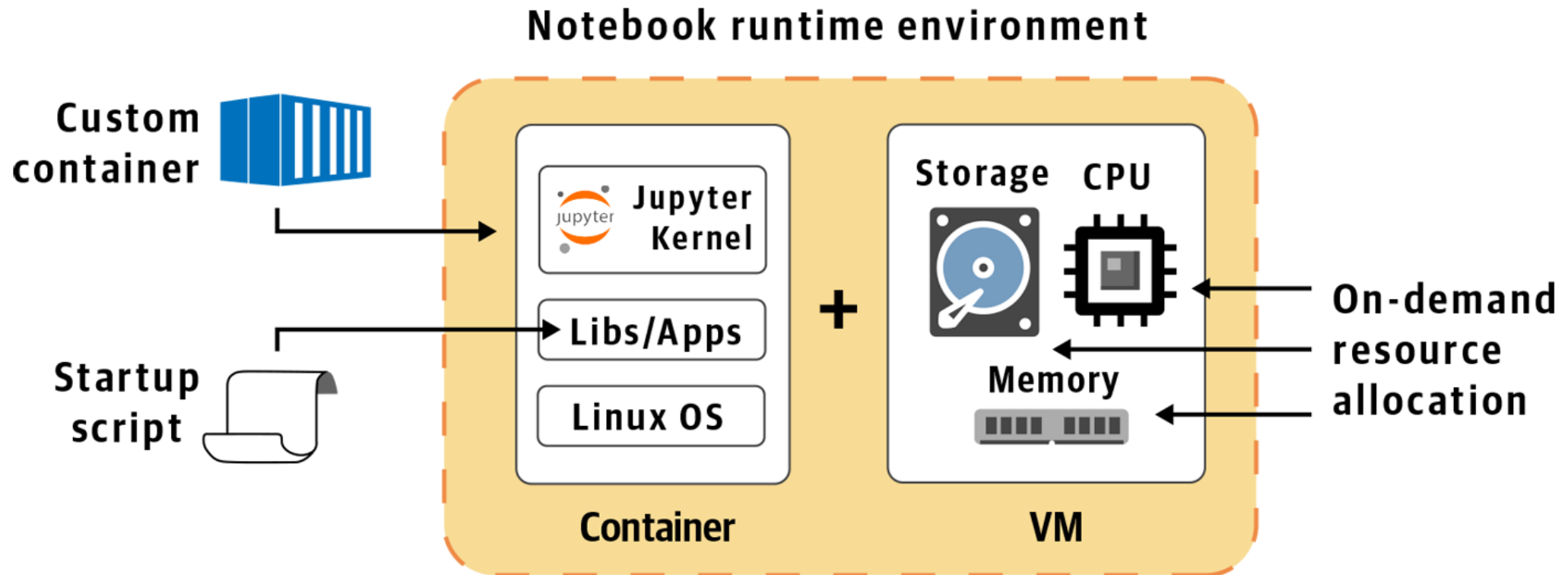
```
In [ ]: # Now you try adding a variable  
greeting =
```



Jupyter notebook overview in Terra

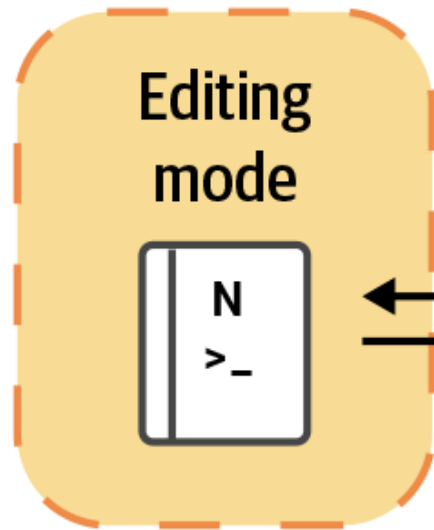


Customizing the notebook runtime

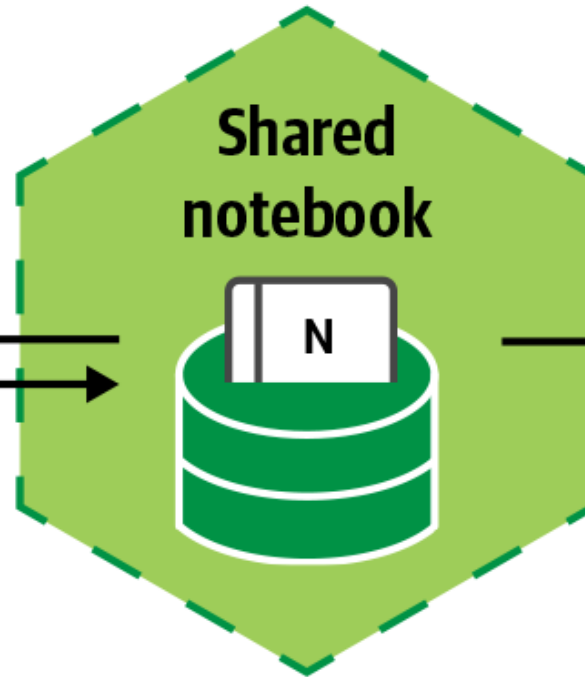


Shared notebook operation

Your runtime

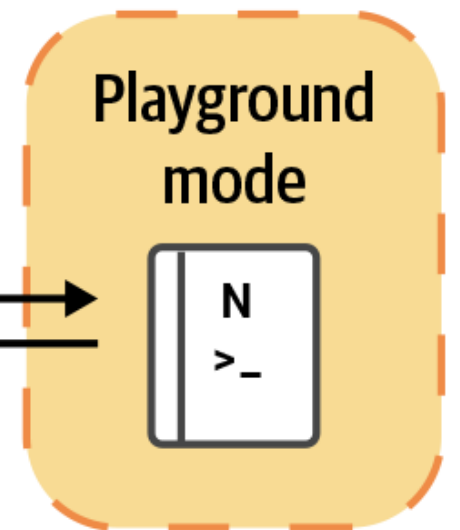


Open
Autosave








Open
x
No save

Their runtime






Notebooks in Terra


**Notebook Runtime**
RUNNING (\$0.10 hr)


[DASHBOARD](#) [DATA](#) **[NOTEBOOKS](#)** [WORKFLOWS](#) [JOB HISTORY](#) 


NOTEBOOKS Sort By:

Most Recently Updated 

[Create a New Notebook](#)


 Genomics-Notebook Last edited: Today

 Genomics-Notebook-executed Last edited: Today



Notebook runtime configuration

RUNTIME CONFIGURATION ✕

Create a cloud compute instance to launch Jupyter Notebooks or a Project-Specific software application.

ENVIRONMENT ⓘ

New Default (released on January 14): (GATK 4.1.4.1, Python 3.7.6, R 3.6.2) ▼

[What's installed on this environment?](#) Updated: Feb 25, 2020
Version: 0.0.13

COMPUTE POWER
Select from one of the default runtime profiles or define your own

Profile Default (Moderate) computer power ▼

CPUs 4 **Memory (GB)** 15 **Disk size (GB)** 50

COST: \$0.19 per hour

[DELETE RUNTIME](#) [CANCEL](#) [REPLACE](#)



View of installed packages

INSTALLED PACKAGES ← ×

New Default (released on January 14): (GATK 4.1.4.1, Python 3.7.6, R 3.6.2) ▼

Updated: Feb 25, 2020
Version: 0.0.13

Installed packages Python ▼

Package	Python ✓	Version
lazy-object-proxy	R	1.4.3
pandocfilters		1.4.2
googleapis-common-protos	Tools	1.51.0
biopython		1.72
tf-estimator-nightly		1.14.0.dev2019030115
ipython-genutils		0.2.0



Installing GATK and IGV

```
pip3 install igv-jupyter
```

```
jupyter serverextension enable --py igv --sys-prefix
```

```
jupyter nbextension install --py igv --sys-prefix
```

```
jupyter nbextension enable --py igv --sys-prefix
```



Compute power

PREVIEW (READ-ONLY)

 EDIT

 PLAYGROUND MODE



COMPUTE POWER
Select from one of the default runtime profiles or define your own

Profile

Custom ▼

CPU

4 ▼

Memory (GB)

15 ▼

Disk size (GB)

50

Startup script

`gs://genomics-in-the-cloud/v1/scripts/install_GATK_4130_with_igv.`

☐ Configure as Spark cluster

COST: \$0.19 per hour



"Hello World" in Jupyter

Cell 1: Run the basic Hello World in Python

```
In [1]: print("Hello World!")  
Hello World!
```

Cell 5: Import the `rpy` package and activate the notebook extension

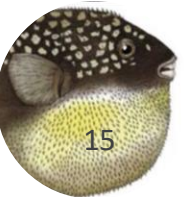
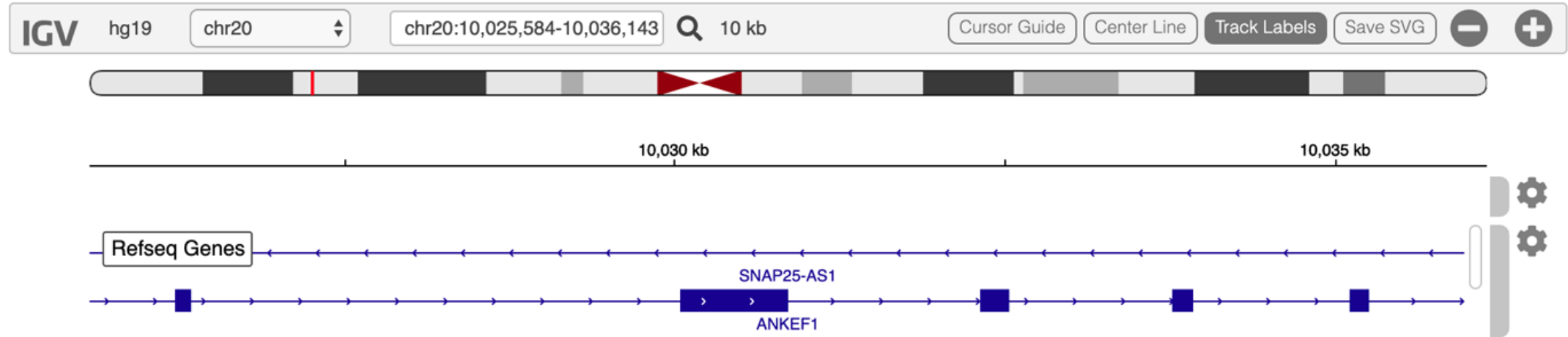
```
In [5]: import rpy2  
%load_ext rpy2.ipython
```

Cell 6: Run the R Hello World with single-line R interpretation using `%R`

```
In [6]: %R print ("Hello World!")  
[1] "Hello World!"  
Out[6]: array(['Hello World!'], dtype='<U12')
```



IGV in Jupyter



Adding data to IGV

```
import igv

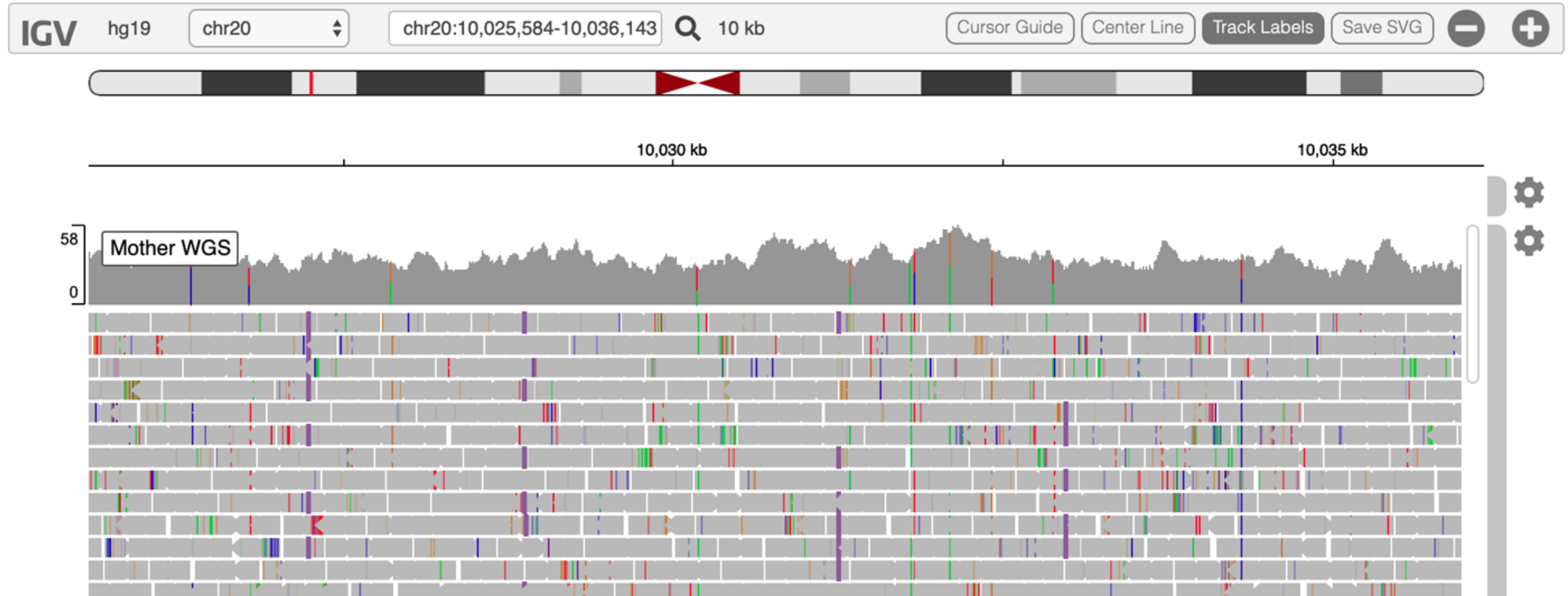
IGV_explore = igv.Browser(
    {"genome": "hg19",
     "locus": "chr20:10,025,584-10,036,143"})

IGV_explore.show()
```

```
IGV_explore.load_track(
    {"name": "Mother WGS",
     "url": GERM_DATA + "/bams/mother.bam",
     "indexURL": GERM_DATA + "/bams/mother.bai",
     "format": "bam"})
```



IGV in Jupyter



Adding an access token to view private data

```
! gcloud auth print-access-token > token.txt
```

Do not share your access token with anyone!

```
with open("token.txt") as token_file:  
    token = token_file.readline()
```

```
IGV_explore.load_track(  
    { ...  
      "token": token})
```



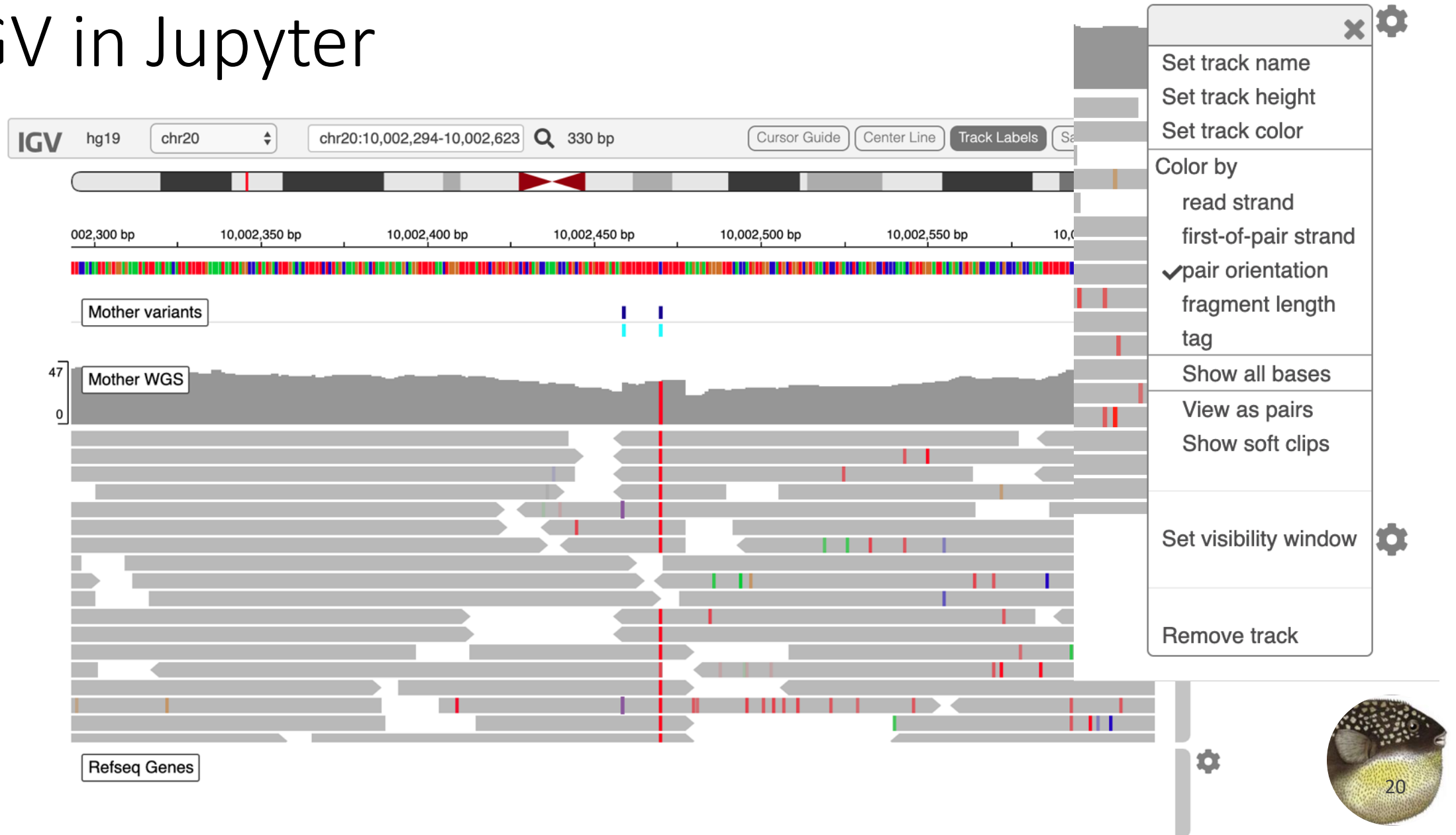
Running a basic GATK command

```
! gatk HaplotypeCaller \  
  -R {GERM_DATA}/ref/ref.fasta \  
  -I {GERM_DATA}/bams/mother.bam \  
  -O sandbox/mother_variants.200k.vcf.gz \  
  -L 20:10,000,000-10,200,000
```

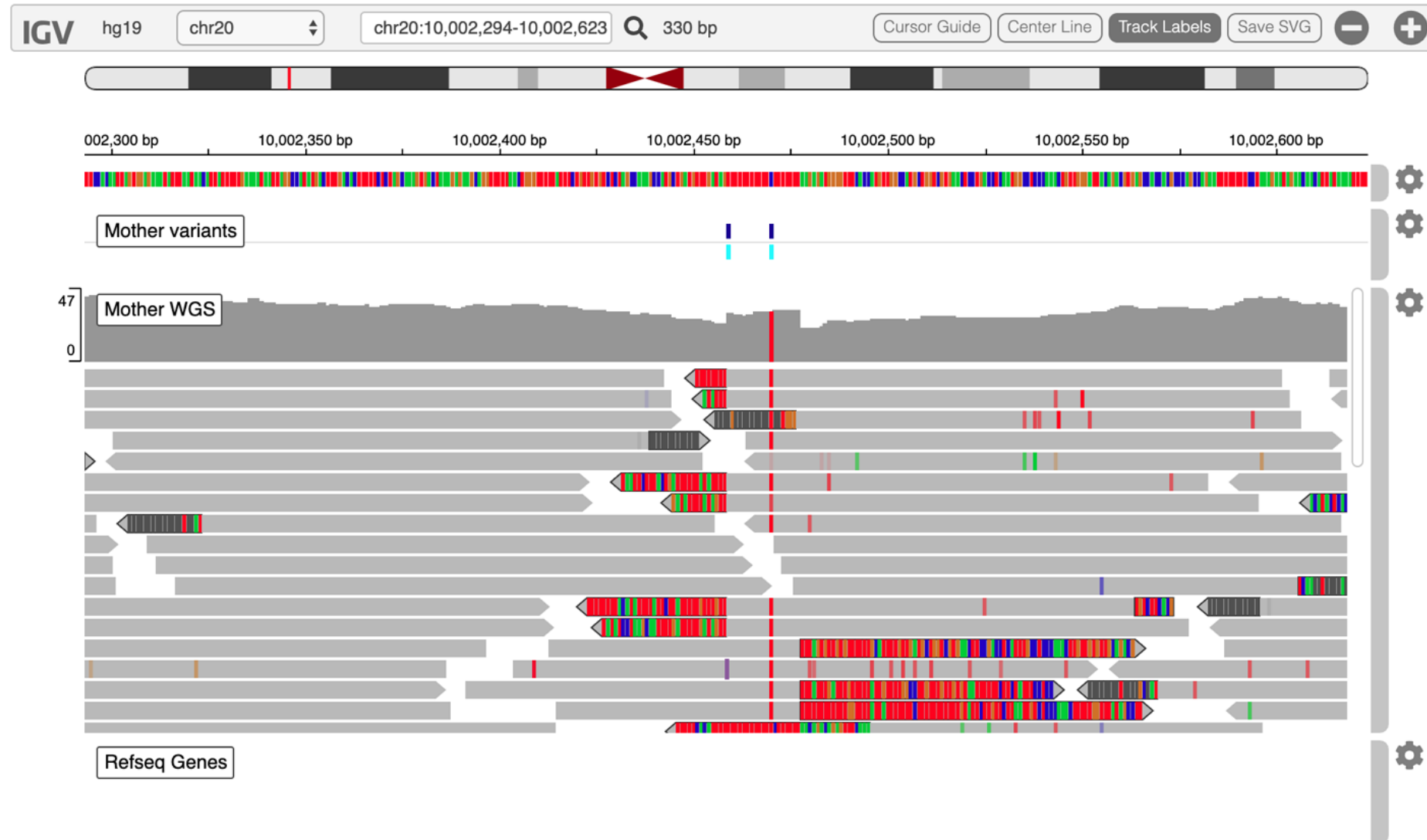
```
! gatk HaplotypeCaller \  
  -R {GERM_DATA}/ref/ref.fasta \  
  -I {GERM_DATA}/bams/mother.bam \  
  -O sandbox/motherHCdebug.vcf.gz \  
  -bamout sandbox/motherHCdebug.bam \  
  -L 20:10,002,000-10,003,000
```



IGV in Jupyter



IGV in Jupyter – paired orientation

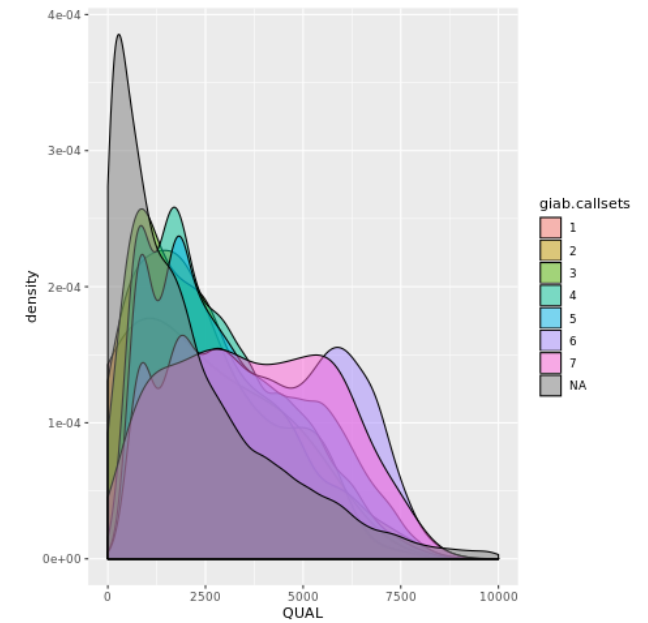
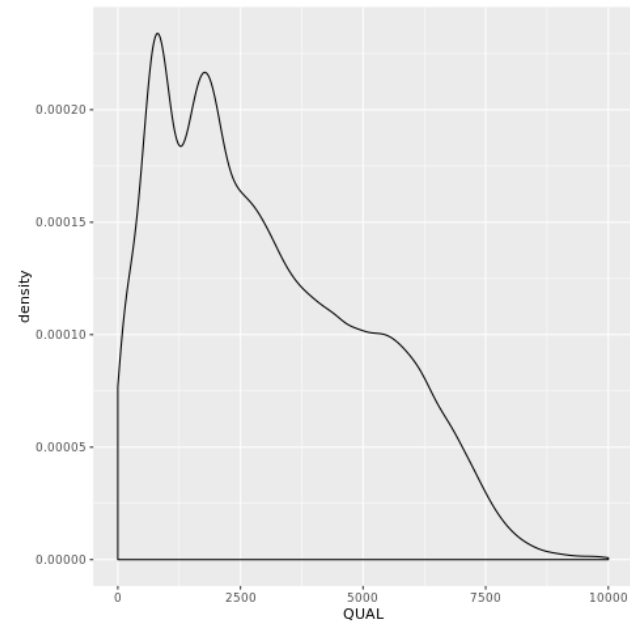
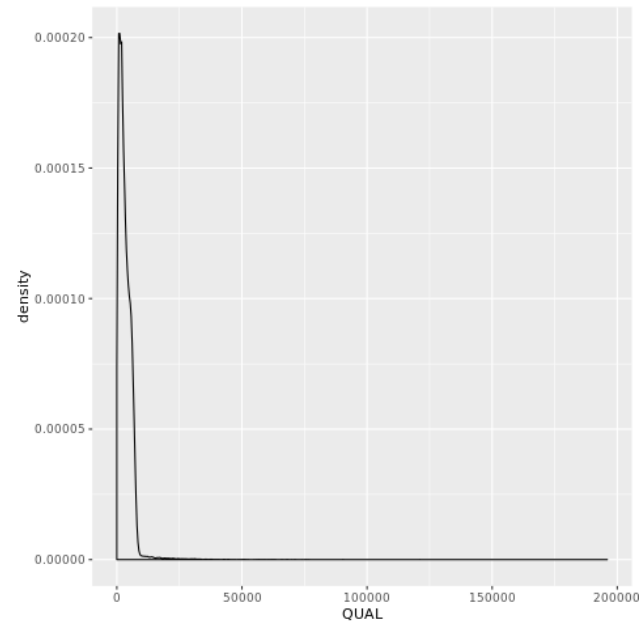


Exporting variants of interest

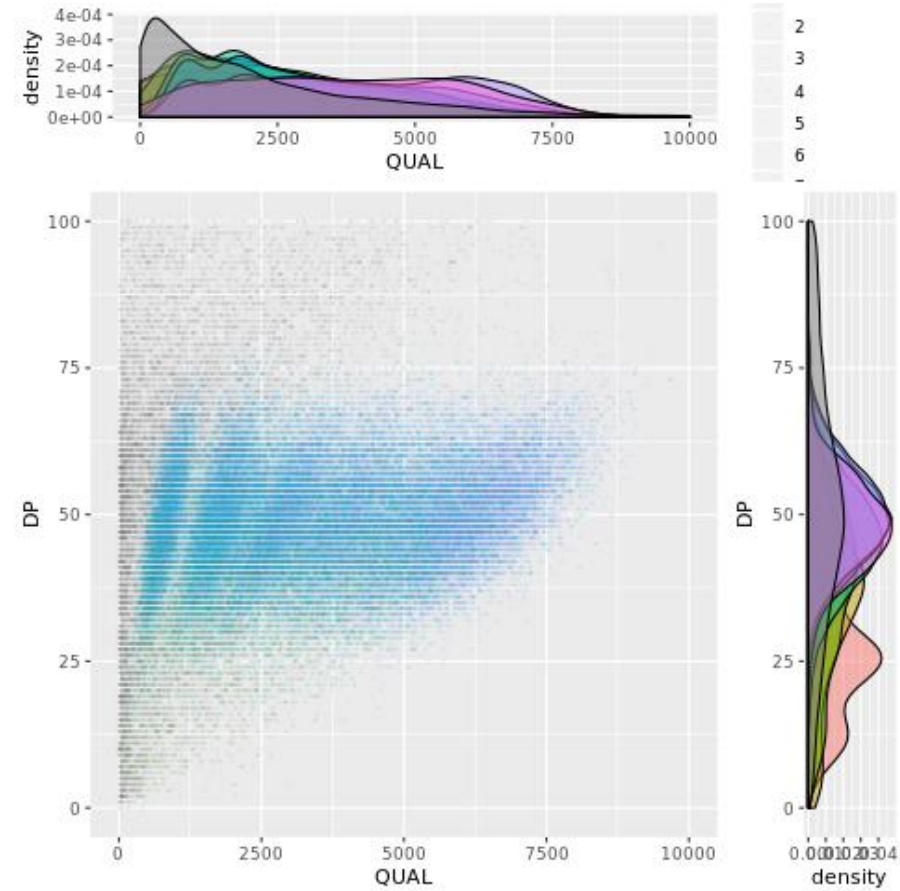
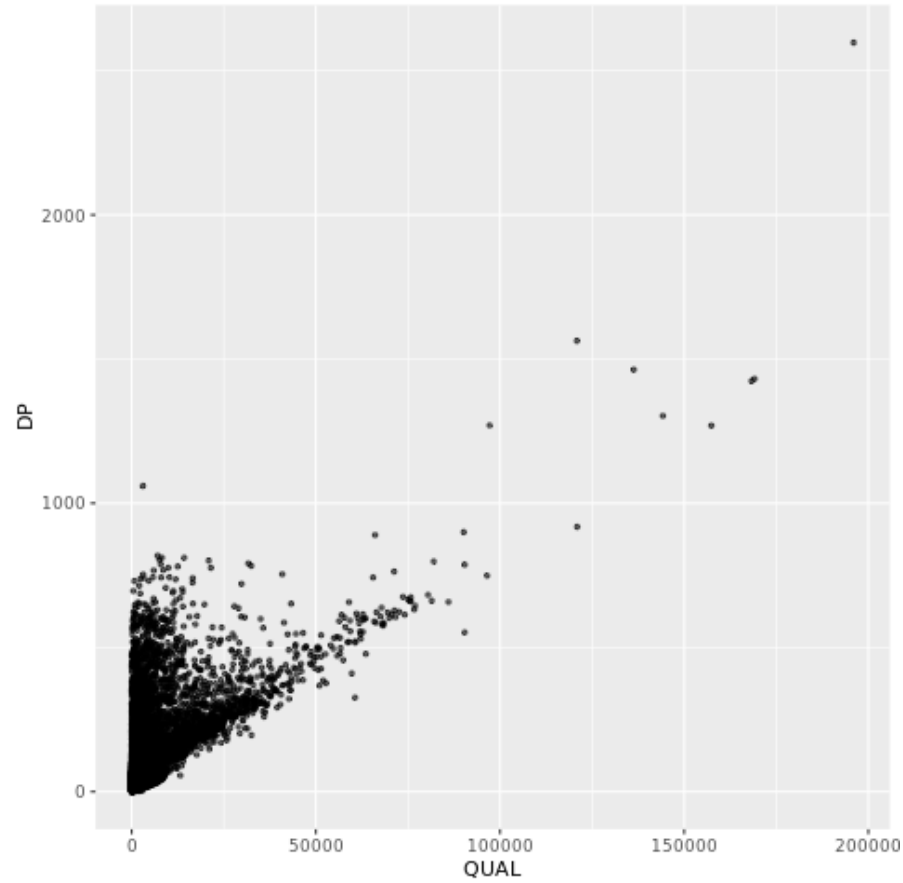
```
! Gatk VariantsToTable \  
  -V {GERM_DATA}/vcfs/motherSNP.giab.vcf.gz \  
  -F CHROM -F POS -F QUAL \  
  -F BaseQRankSum -F MQRankSum -F ReadPosRankSum \  
  -F DP -F FS -F MQ -F QD -F SOR \  
  -F giab.callsets \  
  -GF GQ \  
  -O sandbox/motherSNP.giab.txt
```



QUAL plots



QUAL vs DP // Scatter plot flanked by marginal density plots



Additional resources

- Jupyter documentation
 - <https://jupyter.org/>
 - <https://github.com/igvteam/igv-jupyter>
 - <https://github.com/QuantStack/ipyigv> (alternative igv.js wrapper for Jupyter)
- Hail resources
 - <https://hail.is>
 - <https://app.terra.bio/#workspaces/help-gatk/Hail-Notebook-Tutorials>
 - [https://app.terra.bio/#workspaces/amp-t2d-op/2019 ASHG Reproducible GWAS-V2](https://app.terra.bio/#workspaces/amp-t2d-op/2019%20ASHG%20Reproducible%20GWAS-V2)
- Terra blog (latest features)
 - <https://terra.bio/blog/>



A detailed illustration of a pufferfish, likely a species of pufferfish, shown in profile. The fish has a dark, mottled pattern on its upper body and a lighter, more uniform pattern on its lower body. Its mouth is open, revealing small, sharp teeth. The background is a light, textured surface.

Thank you for joining us today!

Next week: Chapter 13

Next meeting: March 1, 2021