



Genomics in the Cloud

Book Club - Week 9

January 25, 2021

Agenda

- Chapter 8: Automating Analysis Execution with Workflows
- Additional resources
- Open discussion





Our guest speaker

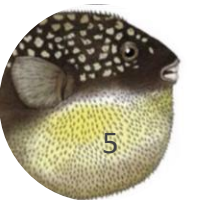
Ms. Lynn Langit

Chapter 8: Automating Analysis Execution with Workflows

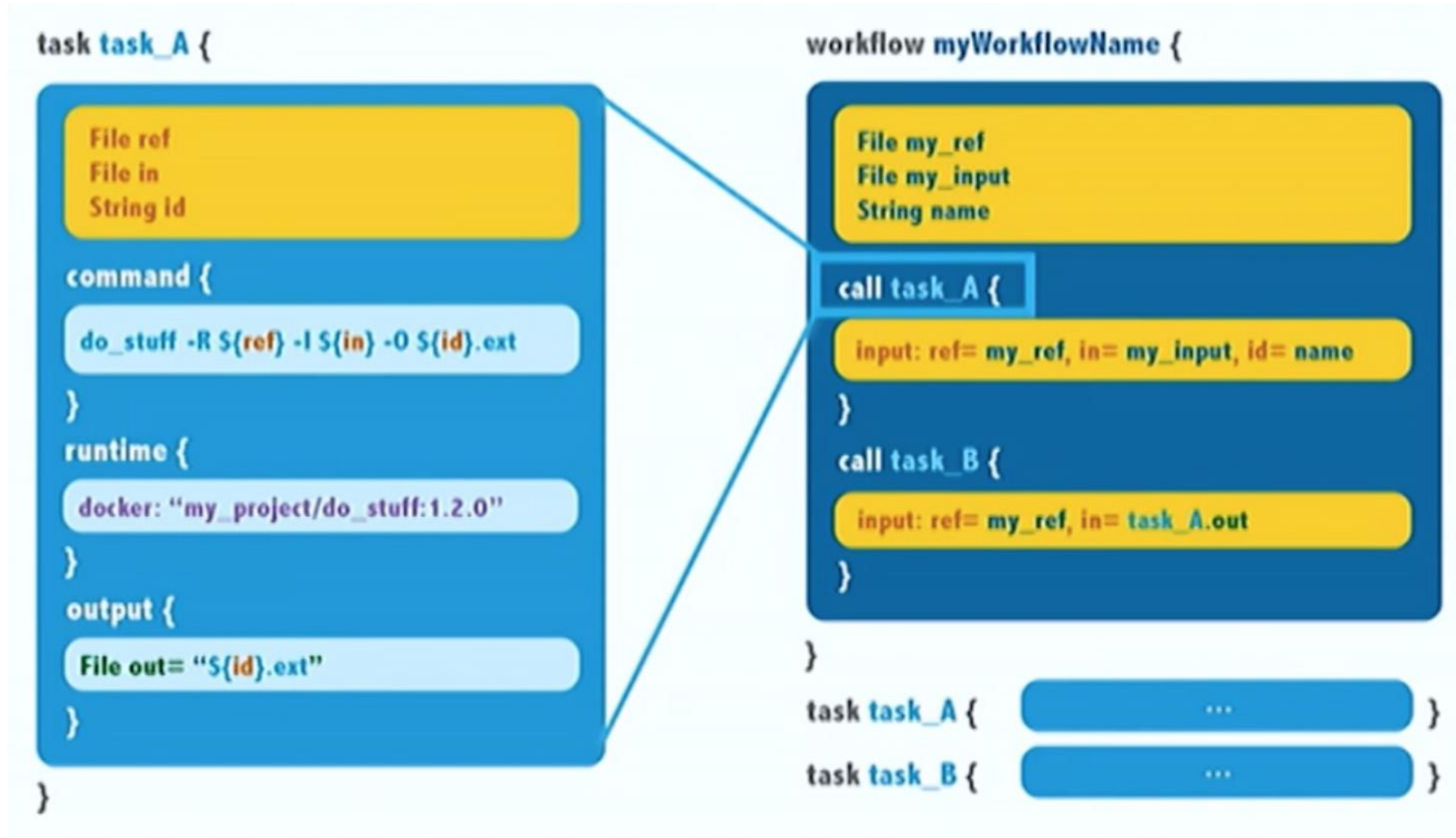
Genomics in the Cloud by Geraldine A. Van der Auwera and Brian D. O'Connor (O'Reilly). Copyright 2020 The Broad Institute, Inc. and Brian O'Connor, 978-1-491-97519-0.

WDL Workflow Components

- Language
 - DSL – **WDL** ‘widdle’ script language
 - Parser – **womtool** WDL parser
- Job Manager – **cromwell** job runner (run | server)
- Execution env – **cloud**, desktop, HPC...
- Dev env – code editor, **VSCode** w/WDL plugin



WDL Workflow Language Concepts



WDL Workflow **Run** Concepts

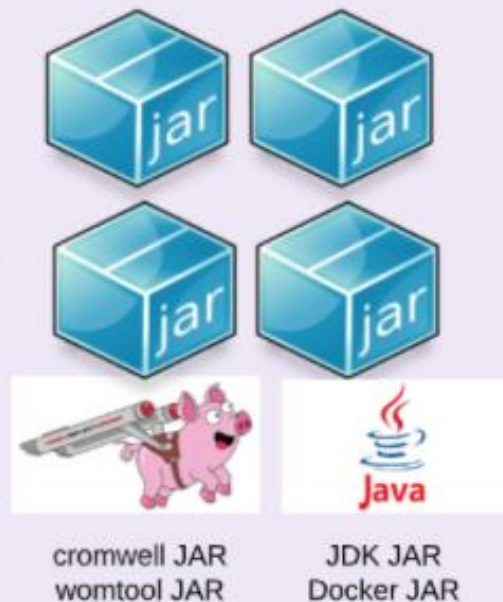
- **Workflow** Job Execution GUID
 - Each run produces a folder of output
 - stdout, stderr and by task name
 - “SUCCEEDED” is good
- **Tasks** must be optimized
 - Use best container practices and | or GATK
 - Container registries
- **Variables**
 - Are strongly typed
 - Are fully qualified (wf.task.var | wf.var)
 - No globals



Compute



+



}

All
JARS



Container Registry
Task Container



Base WDL
Container

Optional Component

Docker Hub



(Optional)
Task Container,
i.e. GATK, etc...

INPUT Data



Cloud Storage
INPUT DATA
bucket



INPUTS



INPUTS

+



Cloud Storage
INPUT CONFIG
bucket



WDL



CONFIG



OUTPUT Data



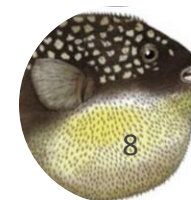
Cloud Storage
OUTPUT DATA
bucket



Job logs



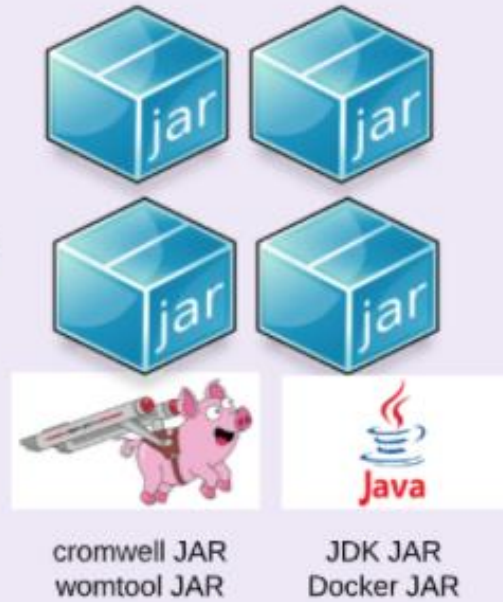
OUTPUTS



Compute

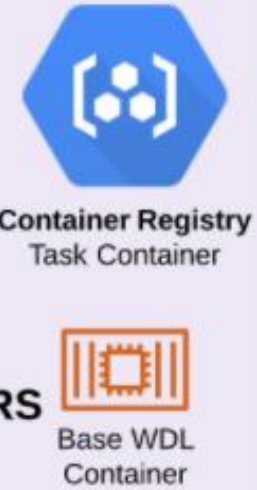


+



}

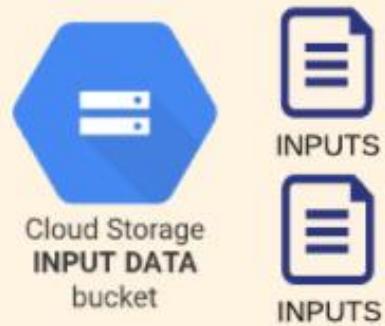
All
JARS



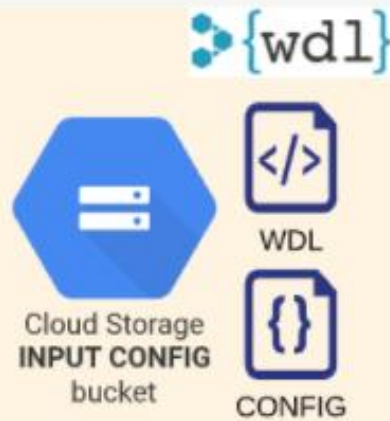
Optional Component



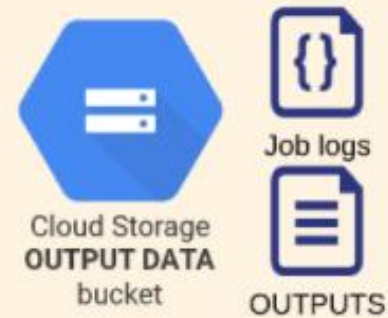
INPUT Data



+



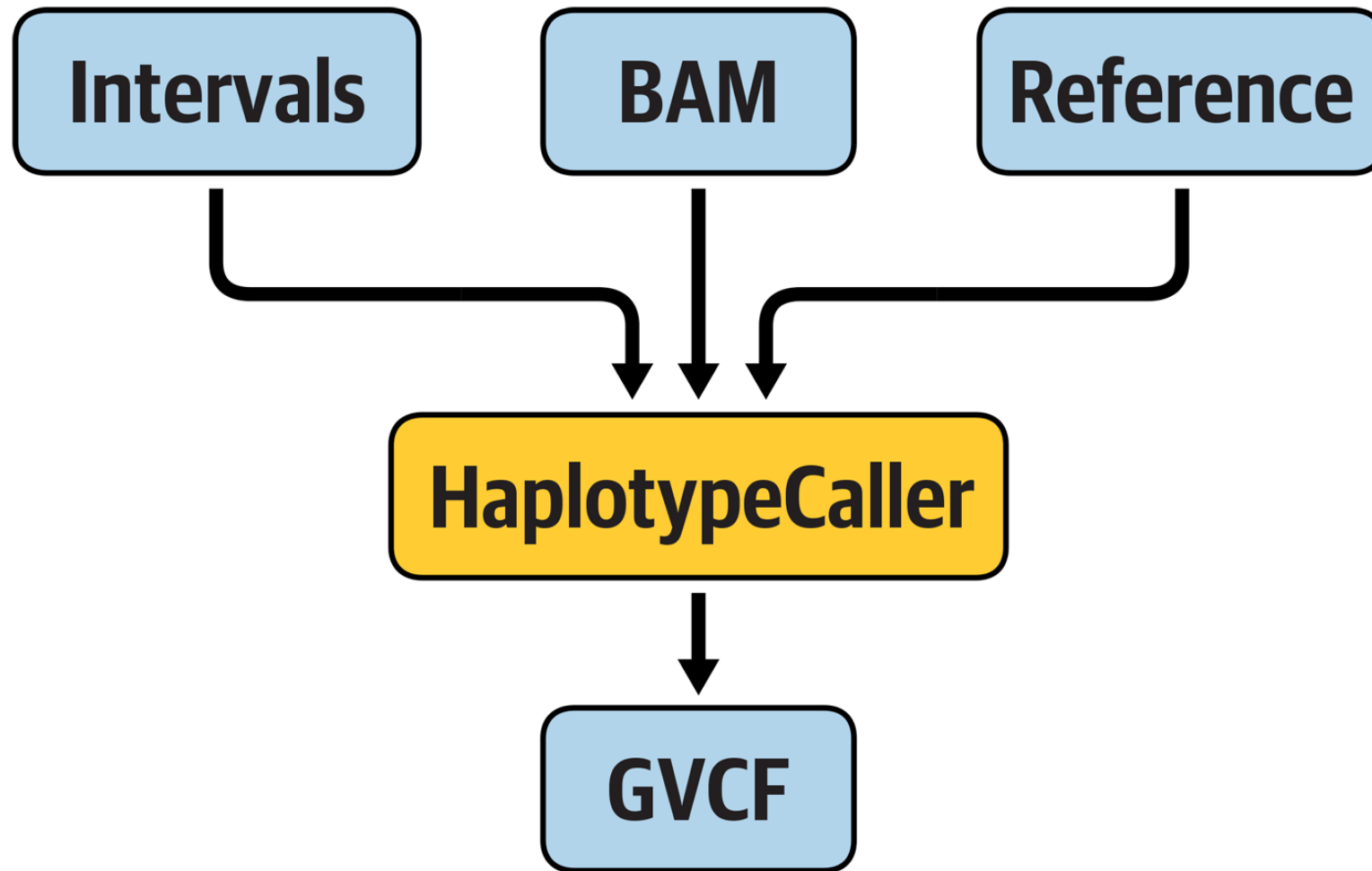
OUTPUT Data



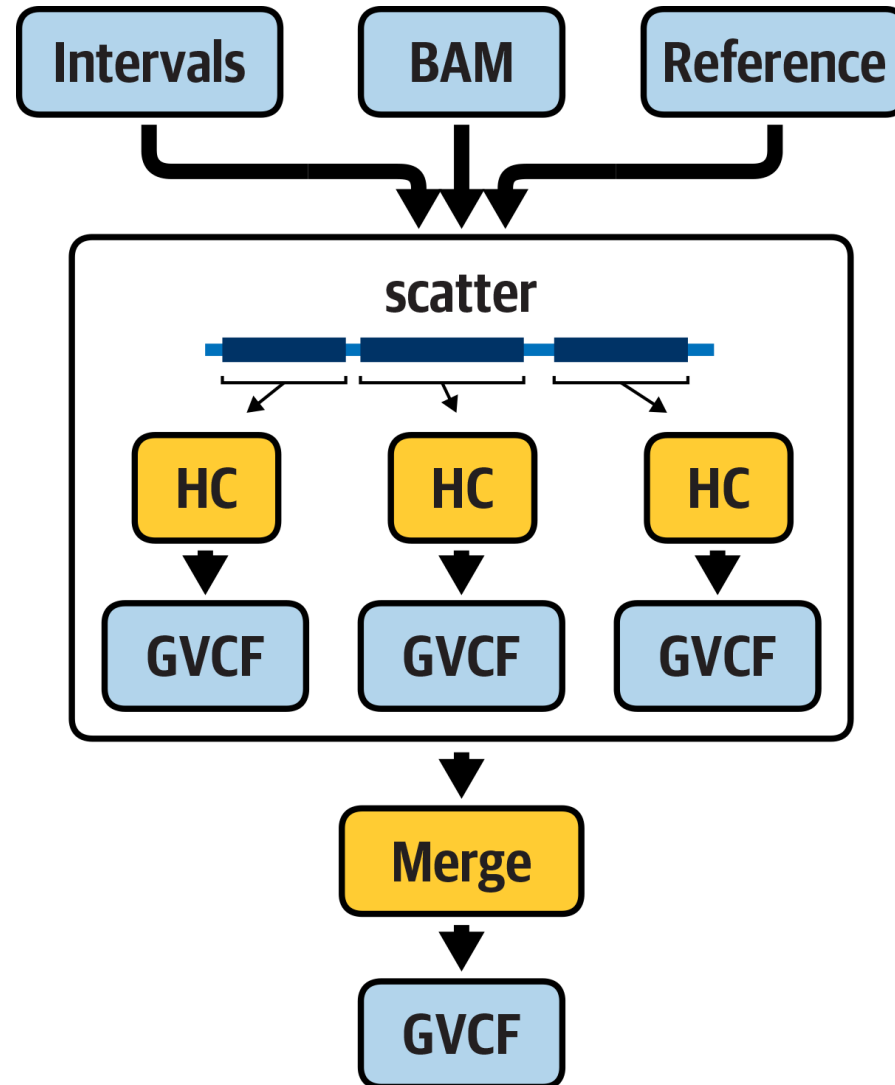
Demos

WDL on GCP

Example workflow with HaplotypeCaller



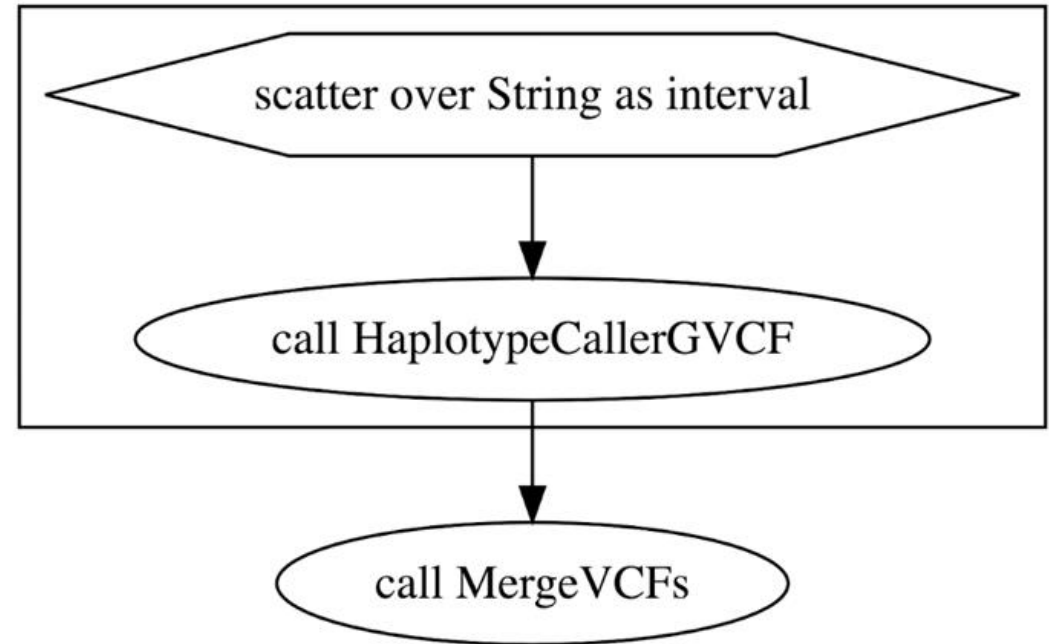
Parallel execution of HaplotypeCaller

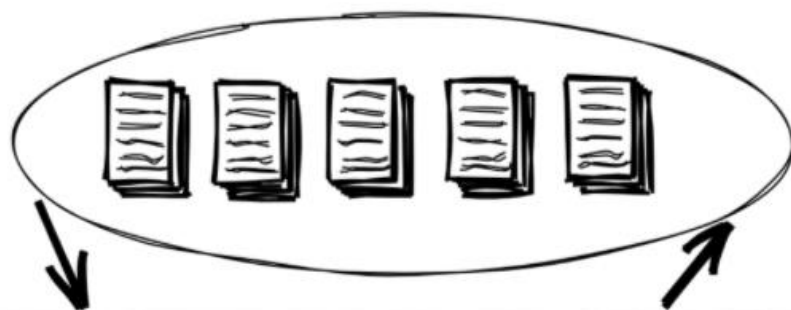


Visualizing the workflow graph

```
1 digraph ScatterHaplotypeCallerGVCF {  
2   #rankdir=LR;  
3   compound=true;  
4   # Links  
5   CALL_HaplotypeCallerGVCF -> CALL_MergeVCFs  
6   SCATTER_0_VARIABLE_interval -> CALL_HaplotypeCallerGVCF  
7   # Nodes  
8   CALL_MergeVCFs [label="call MergeVCFs"]  
9   subgraph cluster_0 {  
10    style="filled,solid";  
11    fillcolor=white;  
12    CALL_HaplotypeCallerGVCF [label="call HaplotypeCallerGVCF"]  
13    SCATTER_0_VARIABLE_interval [shape="hexagon" label="scatter over String as interval"]  
14  }  
15 }  
16
```

Engine: dot Format: svg ☐ Show raw output [Share](#)





Data Lake



WDL, JSON



Docker (GATK)



Daemon



cromwell -> WDL



Hypervisor



Life Sciences API



VMs

Pipeline and params

Job/Container Controller

VM/Cluster Controller

Compute Cluster



Additional resources

- **Open source courses**

- Learn-WDL and YouTube playlist
 - <https://github.com/openwdl/learn-wdl>
 - <https://www.youtube.com/playlist?list=PL4Q4HssKcxYv5syJKUKRrD8Fbd-CnxTM>
- Terra WDL documentation
 - <https://support.terra.bio/hc/en-us/sections/360007274612-WDL-Documentation>
- gcp-for-bioinformatics
 - <https://github.com/lynnlangit/gcp-for-bioinformatics>



Additional resources (continued)

- **WDL resources**

- OpenWDL community
 - <https://openwdl.org/>
- WARP
 - <https://support.terra.bio/hc/en-us/articles/360050981492-Introducing-WARP-A-collection-of-cloud-optimized-workflows-for-biological-data-processing-and-analysis>
- WDL spec
 - <https://github.com/openwdl/wdl/blob/main/versions/1.0/SPEC.md>

- **miniwdl resources**

- <https://miniwdl.readthedocs.io>
- https://github.com/openwdl/learn-wdl/tree/master/6_miniwdl_course





Thank you for joining us today!

Next week: Chapter 9

Next meeting: February 1, 2021