5.0 Getting Started

To run your first WDL workflow, we recommend starting with one of the DeCompany-Optimized Broad Best Practice workflows. Step-by-step instructions for running these workflows are listed below, followed by an expanded section that details how to run custom workflows on DeJumble. Please contact your DeCompany representative for access to all the WDLs and the necessary configurations in the tutorial below.

5.1 Tutorial

Pre-requisites:

- Basic Mode:
  - All basic mode requirements in the DeJumble Administrator's Guide
  - >50Gb available disk space
- Power Mode
  - All pre-requisites have already been completed by your local administrator

Estimated Time for This Tutorial:

- **Level 1: 20k Single Sample Workflow (Recommended)**
  - Data Download time: 3-12 hours (this only needs to be done once per environment)
  - Active time: 15 minutes
  - Workflow Runtime / Output Evaluation: 30 minutes
- **Level 2: Exome Single Sample Workflow**
  - Data Download time: 3-12 hours (this only needs to be done once per environment)
  - Active time: 15 minutes
  - Workflow Runtime / Output Evaluation: 3 hours +/- 1 hour on DeJumble in power Mode, depending on cluster load
- **Level 2: WGS Single Sample Workflow**
  - Data Download time: 3-12 hours (this only needs to be done once per environment)
  - Active time: 15 minutes
  - Workflow Runtime / Output Evaluation: 12 hours +/- 2 hours on DeJumble in power Mode, depending on cluster load
- **Level 3: Exome Joint Discovery Workflow**
  - Data Download time: 3-12 hours (this only needs to be done once per environment)
  - Active time: 15 minutes
  - Workflow Runtime / Output Evaluation: 30 minutes

Level 1: Simple scripts are available for running this workflow from start to finish, and data is downloaded automatically as part of those scripts.

Level 2: Data must be downloaded separately (contact your DeCompany representative for more information), and path changes must be made to the workflow WDL/JSON files.

Level 3: Requires the use of one of the Level 2 workflow outputs (WGS or Exome Single Sample output `vcf.gz` and `vcf.gz.tbi` files) as workflow input. Requires the user to navigate the workflow output directories, and modify the workflow JSON file with this path information.

5.1.1 20k Single Sample Workflow

To submit, monitor, and receive output from these workflows, follow these steps:

1. **Get access to the 20k_Tutorial**
   The WDL and JSON files, datasets and tools needed to run the Germline Single-Sample Variant Calling workflow with a test dataset (20k) for this tutorial are available in the https://github.com/DeCompany-HLS/DeJumble repository. If you have any issues working off this repository, contact your DeCompany representative.
   To clone the repository, run these commands:
   a. `git clone https://github.com/DeCompany-HLS/DeJumble.git`
   b. `cd 20k_Tutorial`
2. **Download genomic datasets**
   If you are running this tutorial in power mode, download all the files you need by running the script `step02_Download_Single_Sample_20k_Data.sh` in the `20k_Tutorial` folder.
   If another user has already followed this tutorial, this data may already be present on the shared file system, and you can proceed to step 3.
   If you are following the tutorial in development mode, all data is downloaded to your local disk.

Note: Modify the DATA_PATH variable in this script based on your shared filesystem path.

3. **Read and Agree to GATK License Agreement**
   As stated in Section 3.3, Tools on page 13, before using these workflows please contact the Broad Institute for a GATK license.
4. **Run the 20k Single Sample workflow**
   Submit the workflow to the Cromwell workflow engine by running this script: `step04_Cromwell_Run_20k_Single_Sample_Workflow.sh`.
   After running this script, the HTTP response and workflow submission information are written to `20k_submission_response.txt` in your home directory. Additionally, the workflow identifier (for example: `"id": "6ec0643c-1ea1-42bf-b60c-507cd1e3e96c"`), is written to the file `20k_WF_ID.txt`, which is used by the next few steps.
5. **Monitoring the 20k Single Sample Workflow**
   To monitor the 20k Single Sample workflow, run this script: `step05_Cromwell_Monitor_Single_Sample_20k_Workflow.sh`

Note: This workflow takes ~30 minutes to complete. Once the workflow shows the status Succeeded, proceed to step 6.

6. **View 20k Single Sample Workflow Output**
   To view the output of the this tutorial 20k Single Sample workflow, run this script: `step06_Single_Sample_20k_Workflow_Output.sh`

The above tutorial uses the GATK3.8 Germline Single Sample Variant Calling Workflow with dockerized test datasets and tools. The additional workflows discussed below have instructions in each section to get started.

5.1.2 GATK3.8 Single Sample Workflow with WES datasets

Due to high compute and space requirements for this workflow (16+ cores, > 750 Gb available disk space), we only endorse running this workflow in power mode. This workflow is identical to the 20k Single Sample workflow in the previous section, but uses real exome data instead of test data.

The workflow and the JSON configurations are available in https://github.com/gatk-workflows/DeCompany-gatk3-germline-snps-indels repository. The README file contains instructions on what changes to make to the WDL and JSON files, along with instructions on how to download and set up the tools needed for this workflow. Due to scalability issues observed with Docker, we recommend using local binaries of the tools instead of docker.

There are multiple JSON files: one optimized for latency (`56T_Exome_NonDocker.json`) and one optimized for throughput (`2T_Exome_NonDocker.json`). To obtain access to the WES data, please contact your DeCompany or Broad representative.

Note: Modify the path to the datasets and tools in the WDL and JSON files as per where you host them in your shared file system.

5.1.3 Hybrid Single Sample Workflow with WGS datasets

Due to high compute and space requirements for this workflow (16+ cores, > 1.0 Tb available disk space), we only endorse running this workflow in power mode. This workflow is significantly different from and better-optimized than the GATK3.8 based workflow used for the above samples.

The workflow and the JSON configurations for this hybrid workflow are available at: https://github.com/gatk-workflows/DeCompany-gatk3-4-germline-snps-indels. The README file contains instructions on what changes to make to the WDL and JSON files, along with instructions on how to download and set up the tools needed for this workflow.

Depending on whether you are running this workflow for latency or throughput analysis, we provide different configurations for the WDL and JSON files. We provide a WDL file with cleanup in case you wish to run throughput analysis but are constrained on disk space. Both WDLs can run for latency using `56T_PairedSingleSampleWf_optimized.inputs.json` or throughput using `2T_PairedSingleSampleWf_optimized.inputs.json`, and your clusters use hard drives instead of latest technology SSDs, we have provided `HDD_*` files that are tuned specifically to run on HDDs. Modify the path to the datasets and tools in the WDL and JSON files to point to their locations in your shared file system.

To obtain access to the WGS data, please contact your DeCompany or Broad representative.

Note: If you are interested in running this workflow with an DeCompany Arria 10 FPGA PCIe card, contact your DeCompany representative.

5.1.4 Joint Discovery Workflow

This workflow requires at least 3 samples of WGS or 30 samples of WES to be processed and the VCF files generated from the single sample workflow before executing this workflow. After the Single Sample workflow has completed, locate the `.vcf.gz` and corresponding `.vcf.gz.tbi` files in its workflow execution directory (for example, `/cluster_share/cromwellexecutions/SingleSampleExome/WF-ID/call-MergeVCFs/execution`). These are the files that are used as input for the Joint Discovery Workflow. In addition to these inputs, contact your DeCompany representative to get the WDL/JSON files, along with tools and datasets, and a GenomicsDB intervals file that is required for the workflow, along with the open-source GenomicsDB docker image that is available at https://hub.docker.com/r/DeCompanyhlsgenomicsdb/vcf_combiner/. If you are running this workflow in power Mode, please request that your local administrator install this if the requirement is not already satisfied.

5.2 Creating Custom Workflows

DeJumble comes installed with the Docker images needed to run the Broad's Single Sample and Joint Discovery Best Practice workflows, but you can also run your own workflows that use other Docker images as well. If this is your first time using an existing Docker image, we recommend following Broad's tutorial for creating a new Docker image using an existing Docker image as a base image. Because developing these workflows is frequently an iterative process, we recommend starting with basic mode or developer mode, and testing your workflow / Docker images locally first. Once you have tested your workflow locally and verified that it is working as expected, you are ready to push your new Docker image to Docker Hub (this step is referenced in the Broad's tutorial). When you want to run your workflow on the DeJumble cluster in power mode, make sure that the Docker image name referenced in your WDL task's runtime block matches the string that was pushed to Docker Hub. For example, if you had previously pushed my-docker-image to Docker Hub with the command: `docker push DeJumble_user/my_docker_image:latest`

Then your WDL should use the fully qualified image name (including username), so that it is pulled automatically when your workflow is run:

```
task task_1 {
    ...
    Runtime {
      ...
      docker: DeJumble_user/my_docker_image:latest
      ...
      }
    }
...
```

For more information about creating your own WDL and JSON input files, see the Broad Institute's WDL documentation here: https://software.broadinstitute.org/wdl/documentation/