



Running NVIDIA Parabricks on DNAnexus

Table of contents

What is NVIDIA Parabricks?

Finding the Parabricks tools on DNAnexus

Running the Parabricks FQ-to-BAM Pipeline using the GUI

Running the Parabricks FQ-to-BAM Pipeline using the CLI

List of Figures

Figure 0. Image Homepage

Figure 1. Image Tool List

Figure 2. Image Fq2bam Tool

Figure 3. Image Inputs And Outputs

Figure 4. Image Options

Figure 5. Image Interval Help

Figure 6. Image Select File

Figure 7. Image View Log

Figure 8. Image Terminal

Figure 9. Image View Inputs And Outputs

Figure 10. Image View Project Id

This guide shows how to run Parabricks on a compute instance on [DNAnexus](#) using both the GUI and the CLI.

What is NVIDIA Parabricks?

Parabricks is an accelerated compute framework that supports applications across the genomics industry, primarily supporting analytical workflows for DNA, RNA, and somatic mutation detection applications. With industry leading compute times, Parabricks rapidly converts a FASTQ file to a VCF using multiple, industry validated variant callers and also includes the ability to QC and annotate those variants. As Parabricks is based upon publicly available tools, results are easy to verify and combine with other publicly available data sets.

More information is available on the [Parabricks Product Page](#).

Detailed installation, usage, and tuning information is available in the [Parabricks user guide](#).

Finding the Parabricks tools on DNAnexus

In this section we will show how to find all the available Parabricks pipelines on DNAnexus.

Start on the DNAnexus homepage and click “Tools” from the toolbar at the top.

Projects ALL RESOURCES RGC








Any Name Any ID Any Creator Any Shared With Any Billed To

This will take you to the Tools Library, which shows all workflows you can run on DNAnexus. We can filter for just the Parabricks tools by clicking on “Name” and typing “Parabricks”. The list should look something like this:

Tools Library ALL TOOLS

Name: parabricks Any Category Any Type

Name ^

-  BAM-to-FQ Pipeline (Parabricks accelerated)
This pipeline uses GPU-accelerated software to convert BAM files to NGS FQ output at...
-  Bamsort Pipeline (Parabricks accelerated)
This pipeline uses GPU-accelerated software to sort BAM files.
-  DeepVariant Pipeline (Parabricks accelerated)
Call germline variants using a deep neural network analysis
-  DeepVariant Pipeline (Parabricks accelerated)
GPU accelerated germline analysis using DeepVariant
-  FQ-to-BAM Pipeline (Parabricks accelerated)
This pipeline uses GPU-accelerated software to convert NGS FQ output to BAM output...
-  Germline Pipeline (Nvidia Clara Parabricks accelerated)
Call germline variants using the exact same algorithms as the BWA-GATK4 germline va...
-  Mutectcaller (Parabricks accelerated)
Runs the Parabricks mutectcaller pipeline. See the Parabricks Docs: <https://docs.nvidi...>

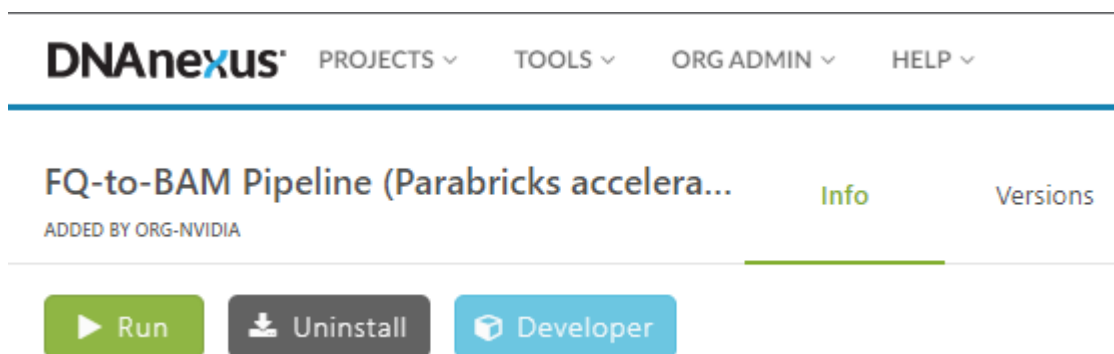
In this guide, we will run FQ-to-BAM as an example to show how to get started. All the workflows run in a similar way, so this information is transferable to any pipeline.

Let's start by clicking on FQ-to-BAM which will take us to the landing page for that tool.

Each tool has a page like this which includes information such as a README, instructions for running on the command line, and input/outputs for this specific tool.

Running the Parabricks FQ-to-BAM Pipeline using the GUI

Let's start by using the GUI to run FQ-to-BAM. Click "Run" in the top left corner.

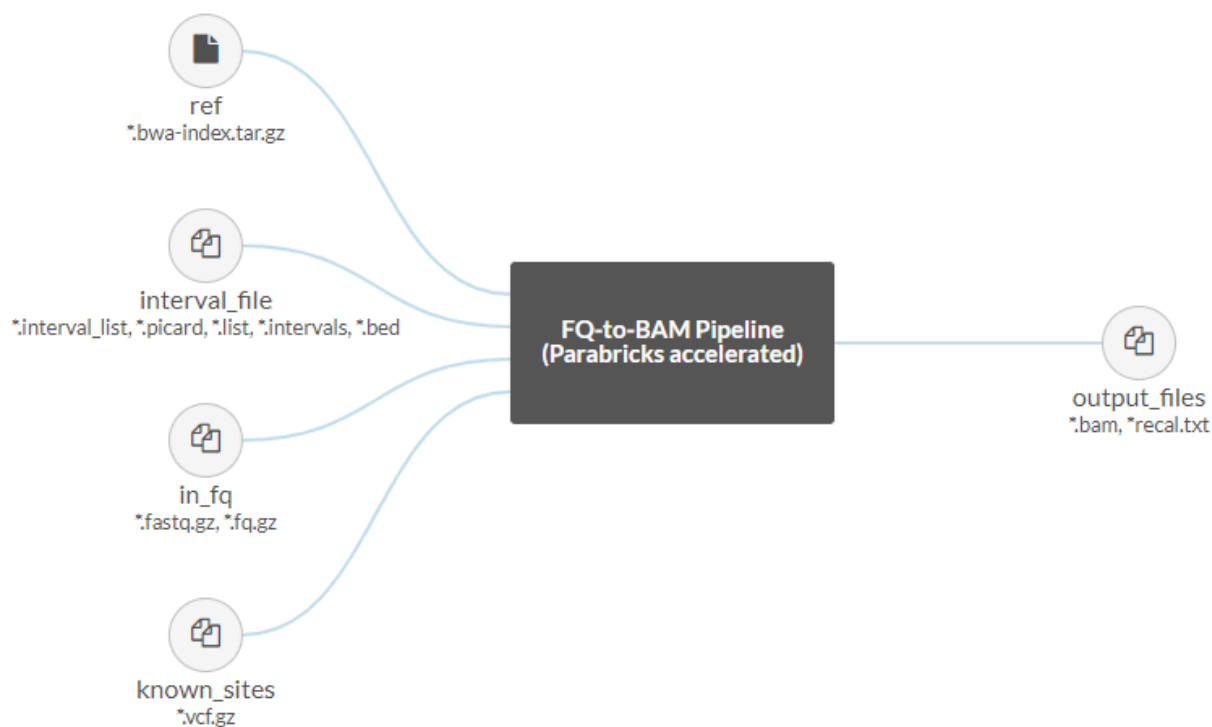


This will open a new page and prompt us to select a project that has data for this run. You can use any fastq and reference files that you like, or you can download Parabricks sample files from using:

```
$ wget -O parabricks_sample.tar.gz \  
"https://s3.amazonaws.com/parabricks.sample/parabricks_sample.tar.gz"
```

and upload them to DNAnexus as we have done for this tutorial. Note that the reference files must be zipped together in one folder.

Once we've selected our project we are shown a graphical representation of the file inputs and outputs for this pipeline on the left side of the page:



The Parabricks FQ-to-BAM pipeline accepts a reference and input fastq pairs as required files, with the option to add interval and known indel files as well. The output will be a bam with option recall file.

Other options can be found on the right side of the page under "Analysis Inputs 2":

ANALYSIS SETTINGS
ANALYSIS INPUTS 2
APP SETTINGS

PBFQ2BAM ▾
Enable Batch OFF

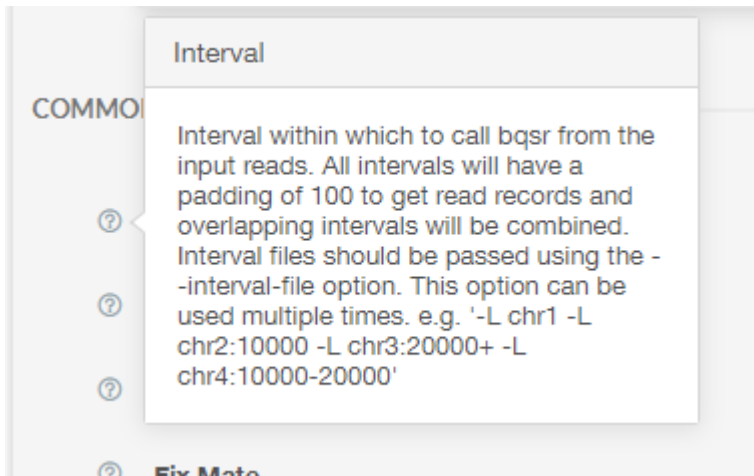
FQ-to-BAM Pipeline (Parabricks accelerated)
[? About this app](#)

*	BWA Reference Genome Index <small>*.bwa-index.tar.gz</small>	<input type="button" value="Select File"/>
?	Interval File <small>*.interval_list *.picard *.list *.intervals *.bed</small>	<input type="button" value="Select File (Array)"/>
*	FQ Read pairs <small>*.fastq.gz *.fq.gz</small>	<input type="button" value="Select File (Array)"/>
?	Known indel files <small>*.vof.gz</small>	<input type="button" value="Select File (Array)"/>

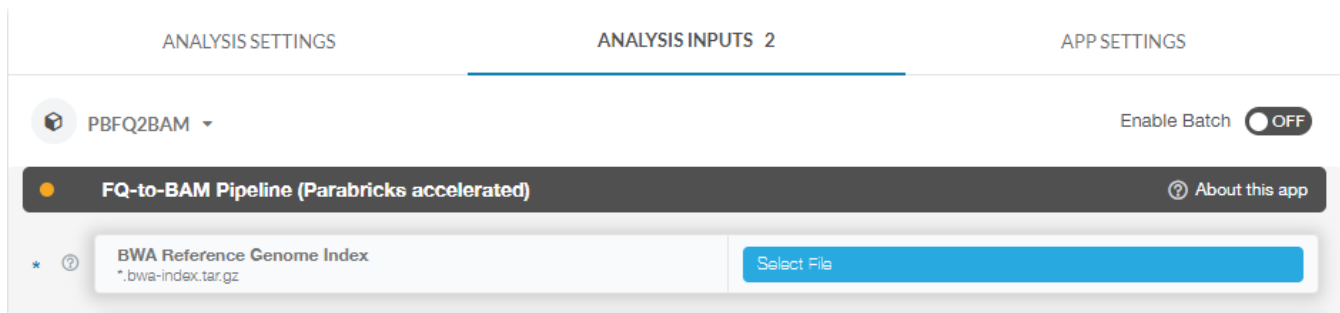
COMMON

?	Interval	<input type="text"/>
?	No Warnings	<input type="button" value="True"/> <input type="button" value="False"/>
?	Mark Dups Assume Sort Order Query Name	<input type="button" value="True"/> <input type="button" value="False"/>
?	Fix Mate	<input type="button" value="True"/> <input type="button" value="False"/>
?	CRAM	<input type="button" value="True"/> <input checked="" type="button" value="False (default)"/>
?	Optical Duplicate Pixel Distance	<input type="text"/>
?	Assume Mark Dups Picard Version 2.18.2	<input type="button" value="True"/> <input type="button" value="False"/>
?	Interval Padding	<input type="text"/>
?	BWA Mem Options	<input type="text"/>
?	Do not mark duplicates	<input type="button" value="True"/> <input type="button" value="False"/>

Here you can see inputs that are not files, for example boolean and integer inputs. Clicking the question mark next to each option will pull up a dialogue box explaining how to use each option. For example, clicking the question mark next to “Interval” results in the following:



For this tutorial, we will click on “Select File” to select our reference zip file:



We will do the same for our fastq pairs. At this point you can set any other options we’d like, however we will leave the default values for everything else for the sake of simplicity in this tutorial.

Now that we have our files selected, we can click “Start Analysis” in the top right corner. This takes us to a page where we can monitor the status of our job. Let’s click on “View Log” and watch as the job runs.



It should take a few minutes for the job to start, and a few more minutes for the job to run to completion.

When the job is done we can check the logs by clicking on View Log. At the bottom of the log we can see the Parabricks terminal output and the confirmation text that the job successfully completed:

```
[PB Info 2023-Jan-11 19:54:09] -----  
[PB Info 2023-Jan-11 19:54:09] ||                Parabricks accelerated Genomics Pipeline                ||  
[PB Info 2023-Jan-11 19:54:09] ||                Version 4.0.0-1                ||  
[PB Info 2023-Jan-11 19:54:09] ||                Marking Duplicates, BQSR                ||  
[PB Info 2023-Jan-11 19:54:09] -----  
[PB Info 2023-Jan-11 19:54:09] progressMeter - Percentage  
[PB Info 2023-Jan-11 19:54:19] 52.3      9.23 GB  
[PB Info 2023-Jan-11 19:54:29] 100.0     0.00 GB  
[PB Info 2023-Jan-11 19:54:29] BQSR and writing final BAM: 20.034 seconds  
[PB Info 2023-Jan-11 19:54:29] -----  
[PB Info 2023-Jan-11 19:54:29] ||      Program:                Marking Duplicates, BQSR                ||  
[PB Info 2023-Jan-11 19:54:29] ||      Version:                4.0.0-1                ||  
[PB Info 2023-Jan-11 19:54:29] ||      Start Time:            Wed Jan 11 19:54:09 2023                ||  
[PB Info 2023-Jan-11 19:54:29] ||      End Time:              Wed Jan 11 19:54:29 2023                ||  
[PB Info 2023-Jan-11 19:54:29] ||      Total Time:                20 seconds                ||  
[PB Info 2023-Jan-11 19:54:29] -----
```

```
Done with Parabricks analysis. Preparing output for upload...  
adding output: file-GKzV7xj0Py5FK6KK66jQ7P0v  
finished creating output!
```

You can click on View all Inputs/Outputs to see the output files as well as the input arguments:

Inputs

BWA Reference Genome Index (ref)
[sample_data.bwa-index.tar.gz](#)
FQ Read pairs (in_fq)
[sample_1.fq.gz](#)
[sample_2.fq.gz](#)
CRAM (cram)
false

Outputs

output_files
[sample.bam](#)

Congratulations! We have successfully run a Parabricks job on DNAnexus.

Running the Parabricks FQ-to-BAM Pipeline using the CLI

For users who prefer to use the terminal as opposed to a GUI, that option exists as well, provided you have the [DNAnexus SDK](#) installed. We can use the following command to run FQ-to-BAM with the same data we used in the previous section:

```
$ dx run fq2bam \ -iref=<project-id:reference-file-id> \ -iin_fq=<project-id:fastq1-file-id> \ -iin_fq=<project-id:fastq2-file-id>
```

For this we need the ID for the project and files that we plan to use. One way to get these is to go to the GUI, click on the file, and copy the ID from the right sidebar:



Once we have our project and file IDs ready, we can run the command and it should come up in the Monitor tab for the project just like using the GUI.

© Copyright 2024, Nvidia.. PDF Generated on 06/05/2024