Workflow Orchestration for Genomics Analysis with Nextflow on the AWS Cloud

Biotech Blueprint Quick Start
Reference Deployment

December 2019

Paul Underwood and Shivansh Singh, Amazon Web Services

Visit our GitHub repository for source files and to post feedback, report bugs, or submit feature ideas for this Quick Start.

Contents

This Biotech Blueprint Quick Start was created by solutions architects with Amazon Web Services (AWS).

Biotech Blueprint Quick Starts are automated reference deployments that use AWS CloudFormation templates to build a preclinical, cloud-based infrastructure on the AWS Cloud, based on industry best practices, and to deploy informatics software from leading biotech companies into this biotech cloud environment.

Overview

This Quick Start reference deployment guide provides step-by-step instructions for deploying a genomics analysis environment on the AWS Cloud, using Nextflow to create and orchestrate analysis workflows and AWS Batch to run the workflow processes.

Nextflow is an open-source workflow framework and domain-specific language (DSL) for Linux, developed by the Comparative Bioinformatics group at the Barcelona Centre for Genomic Regulation (CRG). The tool enables you to create complex, data-intensive workflow pipeline scripts, and simplifies the implementation and deployment of genomics analysis workflows in the cloud.

This Quick Start is for teams or individuals who are responsible for managing informatics infrastructure and genomics analysis for a biotech company. Whether your company is new to the cloud or has some existing cloud infrastructure, the Biotech Blueprint Quick Start provides the informatics software your researchers need, and implements it by following industry best practices.

The underlying AWS CloudFormation templates that are launched are the same as those detailed in the AWS Genomics Workflows blog. In the blog, you'll find more information about Nextflow, its architecture, and other workflow alternatives like Cromwell and AWS Step Functions. For more information about Nextflow, see the Nextflow website.

Cost and licenses

You are responsible for the cost of the AWS services used while running this Quick Start reference deployment. There is no additional cost for using the Quick Start.

The AWS CloudFormation template for this Quick Start includes configuration parameters that you can customize. Some of these settings, such as instance type, will affect the cost of deployment. For cost estimates, see the pricing pages for each AWS service you will be using. Prices are subject to change.

> **Tip**   After you deploy the Quick Start, we recommend that you enable the AWS Cost and Usage Report to track costs associated with the Quick Start. This report delivers billing metrics to an S3 bucket in your account. It provides cost estimates based on usage throughout each month and finalizes the data at the end of the month. For more information about the report, see the AWS documentation.

Nextflow is free, open-source software that is distributed under the Apache 2.0 license.

Architecture

The Quick Start deploys Nextflow into the infrastructure set up by the Biotech Blueprint core Quick Start.

If you want to use an existing VPC or create a new VPC, follow the Genomics Workflows on AWS instructions instead.

If you're new to AWS or don't have a strong VPC architecture already, we recommend that you first use the Biotech Blueprint core Quick Start to set up the landing zone and VPC environment shown in Figure 1 for future AWS usage. This environment is automatically configured for identity management, access control, encryption key management, network configuration, logging, alarms, partitioned environments (for example, to separate preclinical, clinical, and management processes), and built-in compliance auditing, to help meet your security and compliance requirements. For more information about the core architecture, see the Biotech Blueprint core Quick Start.
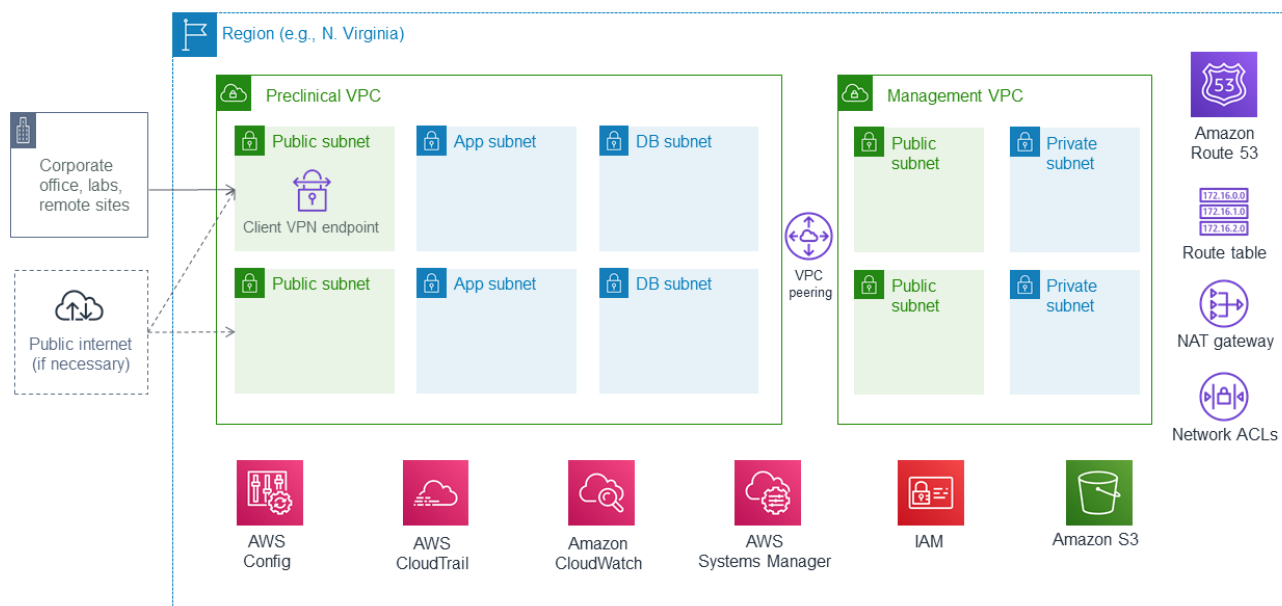
Figure 1: VPC architecture for Biotech Blueprint core template on AWS

This Quick Start deploys Nextflow into the preclinical (research) VPC illustrated in Figure 1. Figure 2 shows the Nextflow deployment in more detail.
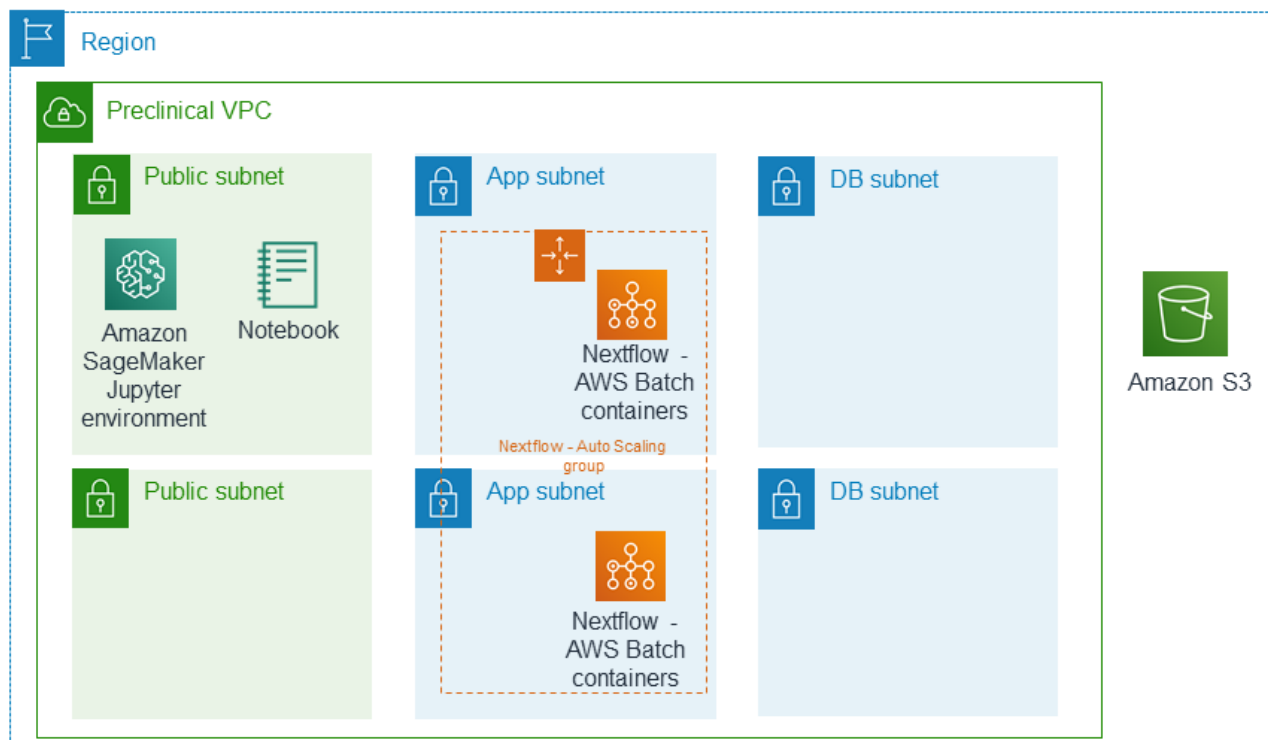
Figure 2: Quick Start architecture for Nextflow on AWS

This Quick Start sets up the following environment in the preclinical VPC:

- A custom Nextflow container image in a private Docker repository provided by an Amazon Elastic Container Service (Amazon ECS) repository. You can also specify your own container image.

- In the public subnet, an optional Jupyter notebook in Amazon SageMaker that is integrated with an AWS Batch compute environment to provide an example of the Nextflow pipeline with AWS Identity and Access Management (IAM) permissions to submit jobs and read log output from Amazon CloudWatch.

- In the private application subnets, an AWS Batch compute environment for managing Nextflow job definitions and queues, and for running Nextflow jobs. AWS Batch containers have Nextflow installed and configured, in an Auto Scaling group. This compute environment contains a high-priority and default-priority queue and a Nextflow head node (a container with Nextflow installed, which submits workflow jobs).

  The default-priority queue uses Spot Instances, and the high-priority queue uses On-Demand Instances. When you deploy the Quick Start, you can set the minimum and maximum CPUs for the compute environments, and specify a maximum bid for Spot Instances.

- Because there are no databases required for Nextflow, this Quick Start does not deploy anything into the private database subnets created by the Biotech Blueprint.

- An S3 bucket to store your Nextflow workflow scripts, input and output files, and working directory. There is no reason all of these assets need to be in the same bucket. For example, if you already have your input files in an existing bucket, you can leave them there. You just need to make sure that the IAM roles associated with the Amazon SageMaker notebook and AWS Batch environment have IAM permissions to those S3 locations. Figure 3 illustrates how you can keep your workflow scripts separate from your input/output files.

- An IAM role for the Nextflow head node job that allows it to access AWS Batch.

Workflow orchestration

Figure 3 shows the Nextflow workflow orchestration environment that is based on the Quick Start architecture.
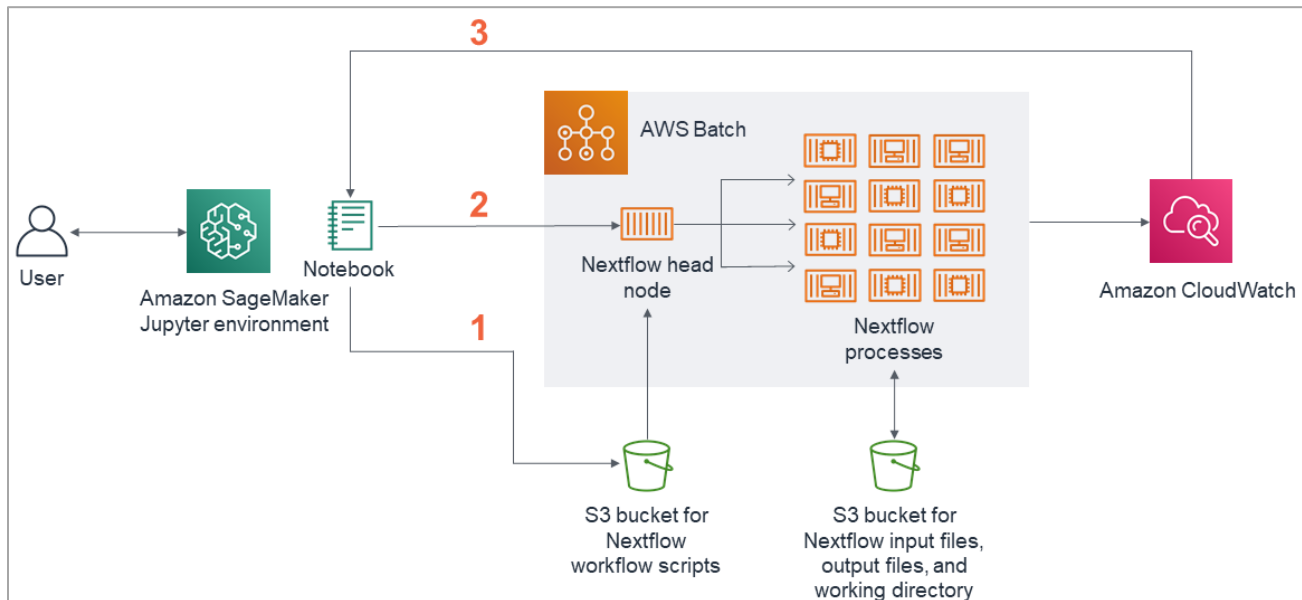
Figure 3: Nextflow workflow orchestration

Here's how the workflow orchestration works:

1. A Jupyter notebook in Amazon SageMaker stages Nextflow workflow scripts in an S3 bucket. It does this with a simple Amazon S3 sync command that moves the entire contents of your workflow directory into a folder in Amazon S3.

2. The Jupyter notebook submits the head node to AWS Batch through the AWS Batch application programming interface (API) for Python. That head node copies the folder in Amazon S3 with your workflow scripts onto the head node and runs the Nextflow executable. Nextflow then submits additional AWS Batch jobs for each Nextflow process that begins.

3. The Jupyter notebook monitors the Amazon CloudWatch Logs log stream for the head node output. You can see additional output files and logs by looking in the Nextflow working directory in Amazon S3.

After you deploy the Quick Start, follow the directions in step 4 to learn how to run a sample Nextflow script.

Planning the deployment

Specialized knowledge

This Quick Start assumes familiarity with genomics analysis and the Nextflow tool.

This deployment guide also requires a moderate level of familiarity with AWS services such as AWS Batch, Amazon CloudWatch, and Amazon S3. If you're new to AWS, visit the Getting Started Resource Center and the AWS Training and Certification website for materials and programs that can help you develop the skills to design, deploy, and operate your infrastructure and applications on the AWS Cloud.

AWS account

If you don't already have an AWS account, create one at https://aws.amazon.com by following the on-screen instructions. Part of the sign-up process involves receiving a phone call and entering a PIN using the phone keypad.

Your AWS account is automatically signed up for all AWS services. You are charged only for the services you use.

Deployment options

We recommend that you set up an informatics infrastructure for this Quick Start that follows best practices for high availability, security, and compliance from AWS and the industry.

This Quick Start provides the following deployment options:

- The easiest way to set up this infrastructure is to use the Biotech Blueprint core Quick Start. Follow the instructions in the Biotech Blueprint core Quick Start deployment guide to build your environment.

- If you want to deploy Nextflow into a new VPC, but don't want the advantages of the Biotech Blueprint core architecture, follow the "all-in-one" deployment instructions in the AWS Genomics blog and launch the template.

- If you want to deploy Nextflow into your own existing infrastructure, make sure that you have a VPC with at least two private subnets and at least one public subnet. You can specify your VPC and subnet IDs when you launch the Quick Start. You should consider the guidance in the AWS Genomics blog, but use this no VPC template, so that it does not create a new VPC.

Deployment steps

Step 1. Sign in to your AWS account

1. Sign in to your AWS account at https://aws.amazon.com with an IAM user role that has the necessary permissions. For details, see Planning the deployment earlier in this guide.

2. Use the Region selector in the navigation bar to choose the AWS Region where you want to deploy Nextflow. This should match the AWS Region you selected when you set up your biotech blueprint infrastructure.

   > **Note**   This Quick Start currently supports the following AWS Regions:
   >
   > `us-east-1` – US East (N. Virginia)
   >
   > `us-west-2` – US West (Oregon)
   >
   > `eu-west-1` – EU (Ireland)

3. Make sure that at least one Amazon EC2 key pair exists in your AWS account in the Region where you are planning to deploy the Quick Start. Make note of the key pair name. You'll be prompted for this information during deployment. To create a key pair, follow the instructions in the AWS documentation.

   If you're deploying the Quick Start for testing or proof-of-concept purposes, we recommend that you create a new key pair instead of specifying a key pair that's already being used by a production instance.

Step 2. Deploy the Biotech Blueprint core Quick Start

If you haven't already deployed the Biotech Blueprint core Quick Start, do so now.

Step 3. Launch the Quick Start

If you chose to deploy the Biotech Blueprint Informatics Service Catalog when you deployed the Biotech Blueprint core Quick Start, you can see Nextflow available to launch in the AWS Service Catalog. In the AWS Service Catalog console, choose **Nextflow** from the products list, and then choose **Launch product**. Follow the instructions in the AWS documentation to complete the launch.

Or, if you opted out of Service Catalog by setting the **Create Informatics Catalog?** (`pLaunchServiceCatalogPortfolio`) parameter to **No** when you deployed the Biotech Blueprint core Quick Start, follow the steps in this section to deploy Nextflow.

> **Notes**   The instructions in this section reflect the older version of the AWS CloudFormation console. If you're using the redesigned console, some of the user interface elements might be different.
>
> You are responsible for the cost of the AWS services used while running this Quick Start reference deployment. There is no additional cost for using this Quick Start. For full details, see the pricing pages for each AWS service you will be using in this Quick Start. Prices are subject to change.

1. Launch the AWS CloudFormation template into your AWS account.

<div align="center">

Deploy • Nextflow

[Deploy Nextflow into your AWS account](#)

</div>

The deployment takes about 10 minutes to complete.

2. On the **Select Template** page, keep the default setting for the template URL, and then choose **Next**.

3. On the **Specify Details** page, change the stack name if needed. Review the parameters (described in the following tables) for the template. Provide values for the parameters that require input. For all other parameters, review the default settings and customize them as necessary.

    When you finish reviewing and customizing the parameters, choose **Next**.

aws

[View template](#)

Required parameters:

Parameter label (name)	Default	Description
S3 bucket name (`S3BucketName`)	*Requires input*	The name of the S3 bucket you want to use to store analysis results and staging your Nextflow workflow scripts. This can be an existing or new bucket. The bucket name can include numbers, lowercase letters, uppercase letters, and hyphens, but should not start or end with a hyphen.
Existing bucket? (`ExistingBucket`)	Yes	Set this parameter to **No** if the S3 bucket you specified in the previous parameter doesn't exist, and you want the Quick Start to create a new bucket for storing analysis results.
S3 Nextflow script folder prefix (`S3NextflowScript Prefix`)	scripts	The parent folder in the S3 bucket that will contain Nextflow workflow scripts (such as `*.nf`, `*.config`, and other files). The Quick Start will create this folder for you.
S3 Nextflow metadata folder prefix (`S3NextflowPrefix`)	_nextflow	(Optional) Parent folder in the Nextflow metadata bucket for metadata folders. Used only if the Nextflow metadata bucket is the same as the data bucket.
S3 Nextflow log folder prefix (`S3LogsDirPrefix`)	logs	(Optional) Folder in the Nextflow metadata bucket (under the {Nextflow Prefix}, if needed) for session cache and logs.
S3 Nextflow working directory folder prefix (`S3WorkDirPrefix`)	runs	(Optional) Folder in the Nextflow metadata bucket (under the {Nextflow Prefix}, if needed) that contains workflow intermediate results.

Tags:

Parameter label (name)	Default	Description
Owner of this environment (used for tagging) (`OwnerTag`)	owner	Your name. The Quick Start uses this string to tag the resources it launches. For example, the notebook will have the name owner-nextflow-genomics in SageMaker.
Name for this environment (used for tagging) (`NameTag`)	nextflow	Your environment. The Quick Start uses this string to tag the resources it launches.
Purpose for this environment (used for tagging) (`PurposeTag`)	genomics	The purpose. The Quick Start uses this string to tag the resources it launches.

Notebook options

Parameter label (name)	Default	Description
Include JuptyerLab environment (`DeployNotebook`)	Yes	Set this parameter to **No** if you don't want to deploy a Jupyter notebook to provide an example of the Nextflow workflow orchestration environment.
Notebook instance size (`NotebookInstanceSize`)	ml.t2.medium	The SageMaker instance type to use for the Jupyter notebook. The Quick Start provides a selection of instance types that are optimized for different machine learning (ML) use cases. For more information about ML instance type specifications and prices, see the SageMaker webpage.
Subnet notebook should run in? (`NotebookSubnet`)	/BB/Networking/ VPC/Research/ Subnet/DMZ/A	The ID of the public subnet you want to use for the Jupyter notebook nodes. This must be a subnet of the VPC you specify with the VPC ID parameter.
Storage capacity for notebook (`NotebookStorageSpace`)	50	The amount of Amazon Elastic Block Store (Amazon EBS) storage to use for the Jupyter notebook instance, in GiB.

VPC options:

Parameter label (name)	Default	Description
VPC ID (`VpcId`)	/BB/Networking/ VPC/Research	The ID of your existing VPC (e.g., vpc-0343606e) to use for the Nextflow environment.
Private subnet 1 ID (`AppSubnetA`)	/BB/Networking/ VPC/Research/ Subnet/App/A	The ID of the first subnet in your existing VPC for setting up the AWS Batch compute environment (e.g., subnet-a0246dcd). We recommend using a private subnet.
Private subnet 2 ID (`AppSubnetB`)	/BB/Networking/ VPC/Research/ Subnet/App/B	The ID of the second subnet in your existing VPC for setting up the AWS Batch compute environment (e.g., subnet-b58c3d67). We recommend using a private subnet.

AWS Batch configuration options:

Parameter label (name)	Default	Description
Spot bid % (`SpotBidPercentage`)	100	The bid percentage set for the AWS Batch managed compute environment with Spot Instances. Specify a value between 1 and 100.
		The bid percentage specifies the maximum percentage that a Spot Instance price can be when compared with the On-Demand price for that instance type before instances are launched. For example, if you set this parameter to 20, the Spot price must be below 20% of the current On-Demand price for that EC2 instance. You always pay the lowest (market) price and never more than your maximum percentage. For more information, see the Spot Instance pricing webpage.
Default min vCPU (`DefaultCEMinvCpus`)	0	The minimum number of virtual CPUs for the default-priority AWS Batch compute environment. Leaving this at zero means that AWS Batch will remove any compute resources, when there are no jobs on the queue.
Default max vCPU (`DefaultCEMaxvCpus`)	100	The maximum number of virtual CPUs for the default AWS Batch compute environment.
High-priority min vCPU (`HighPriorityCEMin vCpus`)	0	The minimum number of virtual CPUs for the high-priority AWS Batch compute environment. Leaving this at zero means that AWS Batch will remove any compute resources, when there are no jobs on the queue.
High-priority max vCPU (`HighPriorityCEMax vCpus`)	100	The maximum number of virtual CPUs for the high-priority AWS Batch compute environment.
Starting size for scratch EBS volumes (`StartingScratch VolumeSize`)	200	The starting EBS volume size for the scratch space, which provides auto-expanding storage to support unpredictable runtime demands. For more information, see the Genomics Workflows on AWS documentation.
Location to mount scratch point (`ScratchMountPoint`)	—	The path for the scratch mount point in the instance. Leave this parameter blank to use the auto-expanding EBS volume capability described at https://docs.opendata.aws/genomics-workflows/core-env/create-custom-compute-resources/.

Nextflow options:

Parameter label (name)	Default	Description
Nextflow container image	*Requires input*	(Optional) The location of your existing container image for Nextflow, if you have one (e.g., "/nextflow:latest").

Parameter label (name)	Default	Description
(`NextFlowContainer Image`)		If you leave this setting blank, the Quick Start will build a container for you.

AWS Quick Start configuration:

> **Note**   We recommend that you keep the default settings for the following two parameters, unless you are customizing the Quick Start templates for your own deployment projects. Changing the settings of these parameters will automatically update code references to point to a new Quick Start location. For additional details, see the AWS Quick Start Contributor's Guide.

Parameter label (name)	Default	Description
Quick Start S3 bucket name (`QSS3BucketName`)	/BB/QuickStart/QS S3BucketName	The S3 bucket you created for your copy of Quick Start assets, if you decide to customize or extend the Quick Start for your own use. The bucket name can include numbers, lowercase letters, uppercase letters, and hyphens, but should not start or end with a hyphen.
Quick Start S3 key prefix (`QSS3KeyPrefix`)	/BB/QuickStart/QS S3KeyPrefix	The S3 key name prefix used to simulate a folder for your copy of Quick Start assets, if you decide to customize or extend the Quick Start for your own use. This prefix can include numbers, lowercase letters, uppercase letters, hyphens, and forward slashes.

4. On the **Options** page, you can specify tags (key-value pairs) for resources in your stack and set advanced options. When you're done, choose **Next**.

5. On the **Review** page, review and confirm the template settings. Under **Capabilities**, select the two check boxes to acknowledge that the template will create IAM resources and that it might require the capability to auto-expand macros.

6. Choose **Create** to deploy the stack.

7. Monitor the status of the stack. When the status is **CREATE_COMPLETE**, the workflow orchestration environment with Nextflow is ready.

8. Use the URLs displayed in the **Outputs** tab for the stack to view the resources that were created.

Step 4. Run a sample Nextflow script

1. When deployment is complete, open the Amazon SageMaker console at
    https://console.aws.amazon.com/sagemaker/.

2. In the navigation pane, choose **Notebook instances**.
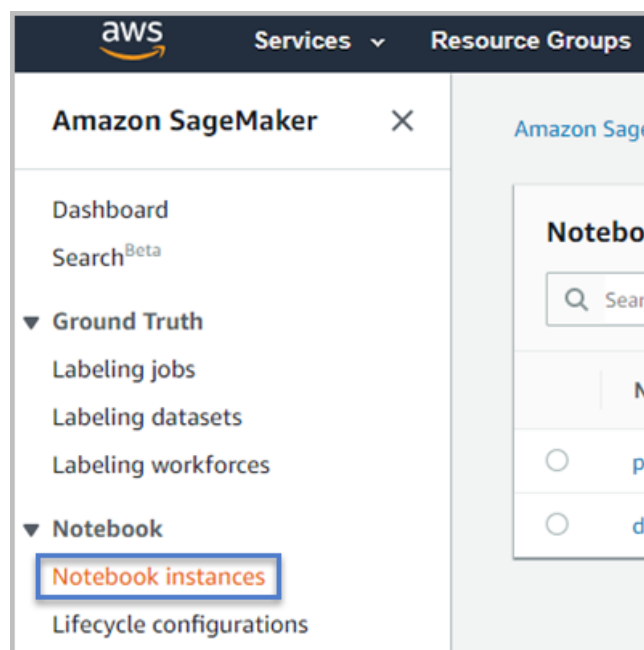
Figure 4: Choosing notebook instances in Amazon SageMaker

The console displays notebook instances that match the tags you provided through the
`OwnerTag`, `NameTag`, and `PurposeTag` parameters in step 3.
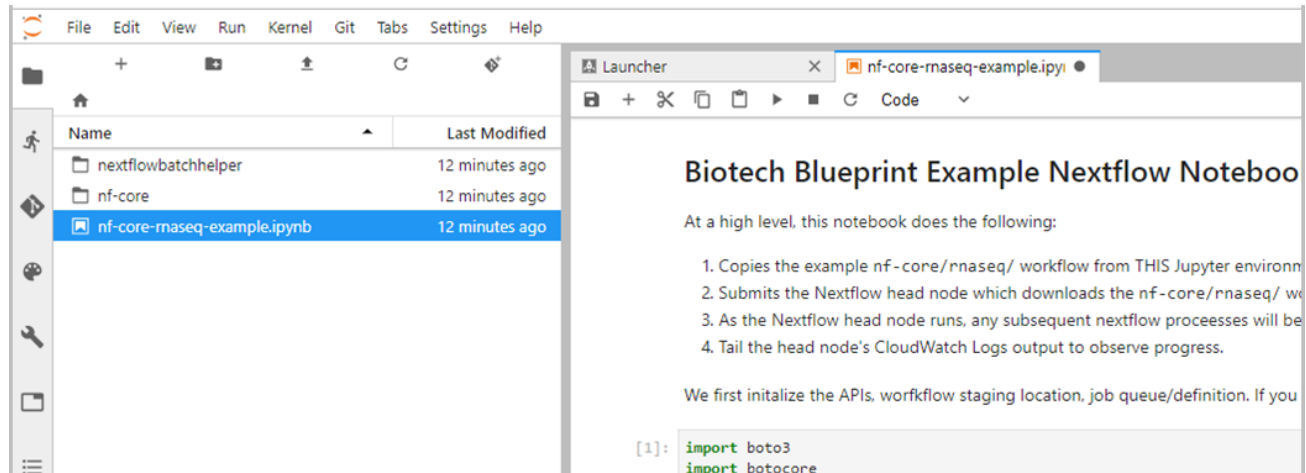
3. Choose **Create notebook instance**.
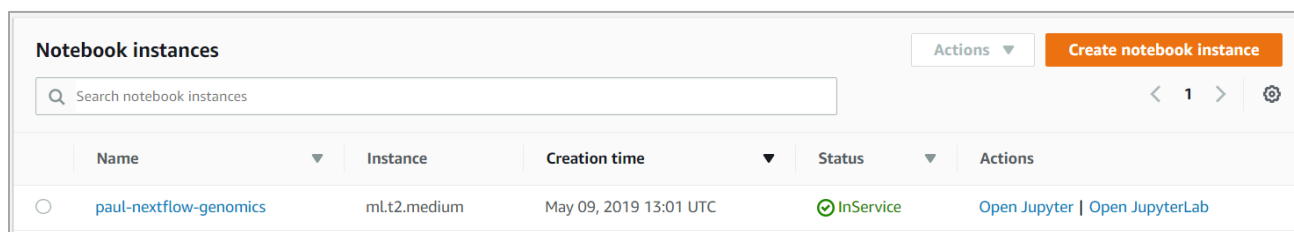
Figure 5: Creating a notebook instance

This launches a new JupyterLab environment for the notebook instance.

4. From the left sidebar, choose **nf-core-rnaseq-example.ipynb**.

Figure 6: JupyterLab interface for Nextflow notebook

The relevant variables, like job queue name, S3 workflow bucket, and prefix, are prepopulated by the Quick Start template.

```python
import boto3
import botocore
import time
batchClient = boto3.client('batch')
s3Client = boto3.resource('s3')
logClient = boto3.client('logs')

workflowBucket = 'devspacepaul'
workflowFolderPrefix = 'nfs'
jobQueueName = 'default-55173c40-725a-11e9-ac24-0297b9e8ab64'
headNodeJobDef = 'nextflow'
```

Figure 7: Prepopulated variables

5. Step through the notebook to see the `nf-core/rnaseq` pipeline kick off. As the pipeline begins to process, you will see the Nextflow head node output begin to arrive from CloudWatch.

Figure 8: Output from the Nextflow head node

You can see the resulting output in the S3 bucket and folder you specified with the **S3 bucket name** (`S3BucketName`) and **S3 Nextflow log folder prefix** (`S3LogsDirPrefix`) parameters in [step 3](#).

For more information about using the JupyterLab interface, see the [documentation](#) on the JupyterLab website.

FAQ

Q. I encountered a CREATE_FAILED error when I launched the Quick Start.

A. If AWS CloudFormation fails to create the stack, we recommend that you relaunch the template with **Rollback on failure** set to **No**. (This setting is under **Advanced** in the AWS CloudFormation console, **Options** page.) With this setting, the stack's state will be retained and the instance will be left running, so you can troubleshoot the issue.

> **Important**   When you set **Rollback on failure** to **No**, you will continue to incur AWS charges for this stack. Please make sure to delete the stack when you finish troubleshooting.

For additional information, see Troubleshooting AWS CloudFormation on the AWS website.

Q. I encountered a size limitation error when I deployed the AWS CloudFormation templates.

A. We recommend that you launch the Quick Start templates from the links in this guide or from another S3 bucket. If you deploy the templates from a local copy on your computer or from a non-S3 location, you might encounter template size limitations when you create the stack. For more information about AWS CloudFormation limits, see the AWS documentation.

Send us feedback

To post feedback, submit feature ideas, or report bugs, use the **Issues** section of the GitHub repository for this Quick Start. If you'd like to submit code, please review the Quick Start Contributor's Guide.

Additional resources

AWS resources

- Getting Started Resource Center
- AWS General Reference
- AWS Glossary

AWS services

- AWS CloudFormation

- [Amazon EBS](#)

- [Amazon EC2](#)

- [IAM](#)

- [Amazon VPC](#)

Nextflow documentation

- [Nextflow website](#)

- [Nextflow documentation](#)

- [Genomics Workflows and Nextflow on AWS](#)

Other Quick Start reference deployments

- [AWS Quick Start home page](#)

Document revisions

Date	Change	In sections
December 2019	Initial publication	—