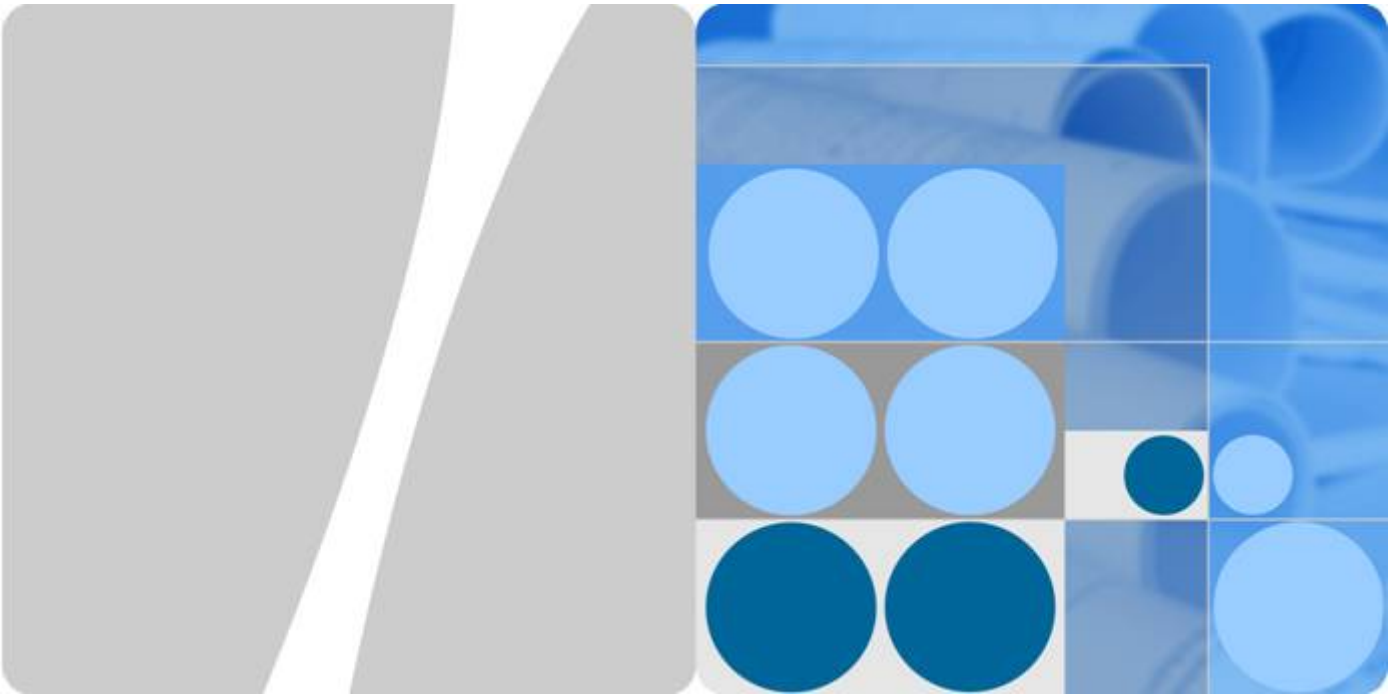


基因容器

快速入门

文档版本　01

发布日期　2020-09-28

华为技术有限公司

商标声明


和其他华为商标均为华为技术有限公司的商标。
本文档提及的其他所有商标或注册商标，由各自的所有人拥有。

注意

您购买的产品、服务或特性等应受华为公司商业合同和条款的约束，本文档中描述的全部或部分产品、服务或特性可能不在您的购买或使用范围之内。除非合同另有约定，华为公司对本文档内容不做任何明示或默示的声明或保证。

由于产品版本升级或其他原因，本文档内容会不定期进行更新。除非另有约定，本文档仅作为使用指导，本文档中的所有陈述、信息和建议不构成任何明示或暗示的担保。

目 录

1 入门必读

基因容器（GeneContainer Service，简称GCS）提供云端基因测序解决方案，支持DNA、RNA、液态活检等主流生物基因测序场景。

基因容器服务对GATK 4.0官方所推荐的最佳实践流程进行封装，让您能快速基于GATK最佳实践流程完成原数据分析。该流程为Broad Institute官方推荐流程，用于全基因组测序比对、去重、碱基校正以及突变检测，关于该流程的详细描述请参见Broad Institute官方文档。

本文旨在帮助您了解基因容器的基本知识，解答您在使用基因容器中可能遇到的问题，帮助您快速使用基因容器服务开始您的基因测序数据分析之旅。

文档导读

基因容器的计算环境可以使用云容器实例（Cloud Container Instance，CCI）、云容器引擎（CloudContainer Engine，CCE）、SGE（Sun Grid Engine）集群、Cromwell引擎和病毒基因分析平台。

- 云容器实例，不需要关注底层计算资源的创建与维护，通过简单的配置即可快速部署容器负载。如果基因分析流程成熟稳定，建议您使用云容器实例环境，可以省去对资源的关注。基于云容器实例的快速入门，请参见**2 基于CCI执行gcs-grammar流程**。
- 云容器引擎，您需要创建管理集群及节点资源。云容器引擎使与云容器实例不同点在于环境的底层资源不同，其余基本相同。
- Cromwell引擎，Cromwell 是 Broad Institute 开发的工作流管理系统。通过 Cromwell 可以将 WDL（Workflow Description Language） 描述的 workflow 运行在CCI容器中。详细使用方法请参见**Cromwell引擎使用指南**。
- 病毒基因分析平台，使用基因容器分析病毒基因组，实时反馈分析结果，简单易用，请参见**3 病毒基因组分析平台入门**。

基本概念

- 镜像

  容器镜像是一个应用的快照。例如，一个容器镜像可以包含一个完整的Ubuntu操作系统环境，里面仅安装了用户需要的应用程序及其依赖文件。容器镜像用于创建容器，您也可以下载其他人已经创建好的镜像来使用。

- 容器

容器是对软件和其依赖环境的标准化打包，可以实现应用层面的隔离，并且可以运行在主流的操作系统上。

镜像类似于操作系统，而容器类似于虚拟机。它可以被启动、开始、停止、删除等操作，每个容器都是相互隔离的

- 流程

  基因测序流程包含测序过程所需工具的执行先后信息以及数据输入输出等定义。流程由至少一个工具组成。流程中的各个工具由其前后顺序关系形成数据流，前序工具为后序工具提供输入。

- 流程描述文件

  基因容器提供特定的描述语言，用于控制流程的详细步骤。基因容器的流程描述文件的编写请参见**流程语法参考**。

- 工具

  工具是生物信息软件的镜像封装，工具既可以编排入流程串联使用，也可以独立使用，同时用户可以制作自定义工具，这些工具都存放在工具仓库中。

- OBS

  对象存储服务（Object Storage Service，OBS）是华为云中基于对象的存储服务，可以为您提供海量、安全、高可靠、低成本的数据存储能力。

- 桶

  桶（Bucket）是OBS中存储对象的容器。对象存储提供了基于桶和对象的扁平化存储方式，桶中的所有对象都处于同一逻辑层级，去除了文件系统中的多层级树形目录结构。

  对象存储服务设置有三类存储类别，分别为：标准存储、低频访问存储、归档存储，从而满足客户业务对存储性能、成本的不同诉求。创建桶时可以指定桶的存储类别。桶的存储类别可以修改。

- SFS

  弹性文件服务（Scalable File Service，SFS）为您的弹性云服务器（Elastic Cloud Server，ECS）提供一个完全托管的共享文件存储，能够弹性伸缩至PB规模，具备可扩展的性能，为海量数据、高带宽型应用提供有力支持。

常见问题

1. 我不懂容器技术，可以使用基因容器服务吗?

   基因容器已将容器技术实现做了封装，提供了简单易用的控制台界面，您可以基于基因容器提供的控制台快速开始基因测序分析，无需额外了解容器或是云计算等相关技术。

2. 已有测序流程和工具迁移到基因容器服务上，方便吗?

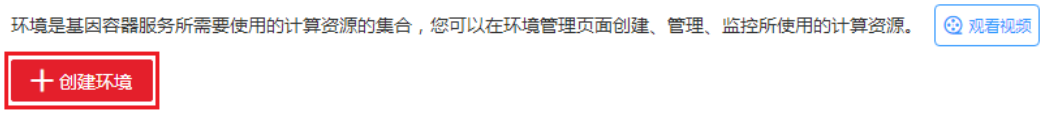
   方便。基因容器提供了工具仓库，用于存储自有工具，此外，您可基于基因容器提供的图形化编辑器，通过界面上拖拽的方式，快速创建测试流程，从而将已有测试流程迁移至基因容器服务上。

2 基于 CCI 执行 gcs-grammar 流程

本文通过创建CCI环境并执行gcs-grammar流程，向您展示GCS的使用流程。

创建环境

在所有操作前，请先完成环境准备，请单击“环境管理 > 创建环境”，根据界面提示填写参数。



在“填写环境信息”界面，参数按如下填写，填写完成后，单击“下一步”，在“详情”页确认后，单击“提交”，完成CCI环境的创建。

- 环境类型

  基因容器的环境由云容器实例CCI、云容器引擎CCE和SGE集群。本示例使用的是CCI提供的环境。

- 默认环境

  是否设置为默认环境。如果只有一个环境，默认只能选择“是”。

- 关联OBS存储

  OBS存储用于存储分析前后产生的数据，包括原始基因数据、流程执行中间数据及执行结果数据。

  选择OBS存储，如果没有OBS桶请单击“创建OBS存储”创建，然后单击刷新，再选择OBS存储。

- 命名空间选择

  命名空间是对于同一用户下的云容器实例的逻辑划分，适用于用户中存在多个团队或项目的场景。

  可以选择“已有命名空间”或“新建命名空间”。建议为此环境选择独立命名空间，保证资源的隔离独立，本示例使用新建命名空间。

  选择企业项目，该参数针对企业用户使用。如需使用该功能，请联系客服申请开通。企业项目是一种云资源管理方式，企业项目管理服务提供统一的云资源按项目管理，以及项目内的资源管理、成员管理，默认项目为default。请从下拉列表中选择所在的企业项目。更多关于企业项目的信息，请参见**《企业管理用户指南》**。

选择命名空间所在的虚拟私有云。虚拟私有云用于构建隔离的、私密的虚拟网络环境，这样在命名空间中容器能够运行于一个隔离的网络环境。如果没有创建虚拟机私有云，请单击“创建虚拟私有云”，完成创建后单击刷新按钮再选择。

- 访问密钥

  上传访问密钥，用于访问OBS存储桶时鉴权。若您在本地没有访问密钥，请前往**管理访问秘钥**新增并下载访问密钥。

查看环境

待环境创建完成，可以单击环境名称查看环境信息。

执行测序流程

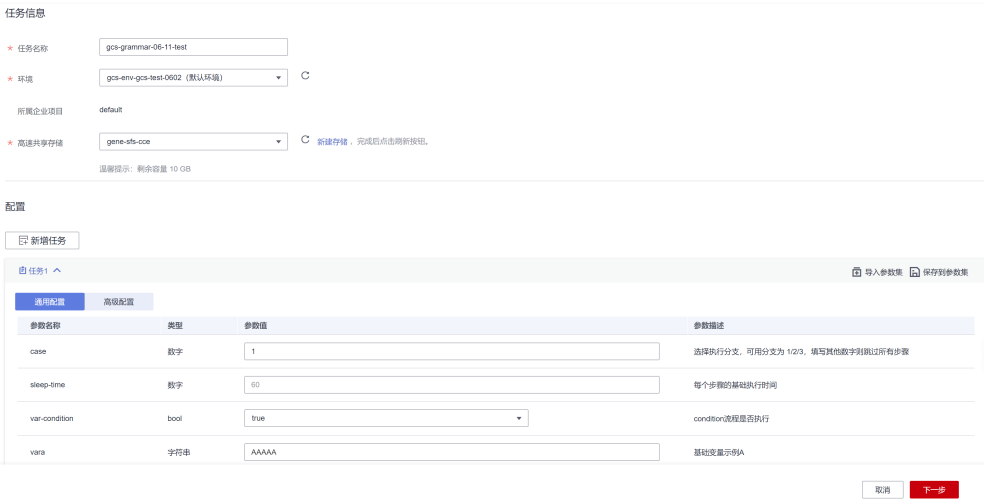
基因容器提供了gcs-grammar示例流程，通过该流程您可以了解到GCS的编排语法及能力。

该流程的执行顺序以及包含工具，可以通过单击”基因测序 > 示例流程 > gcs-grammar”查看流程详细信息。

执行测序流程请按如下步骤操作。

步骤1 单击“开始分析”，开始配置流程参数。流程参数中包括“任务信息”、“配置”和“高级设置”。

- 任务信息：包括任务名称、环境和存储选择。“高速共享存储”对应的是文件存储服务SFS，用于存储流程中间数据。如果下拉列表中没有选项，请单击“导入存储”在弹窗中选择，如果弹窗中无可选项，请单击创建“文件存储创建”，文件存储需要与CCI命名空间在同一个虚拟私有云。
- 配置：包括“通用配置”和“高级配置”，请根据界面提示信息完成参数配置。
- 高级设置：可选项，包括“超时时间”、“批次名称”、“优先级”和“订阅消息通知”，使用默认值即可。



步骤2 单击“下一步”，在“执行预览”页面，确认配置信息。

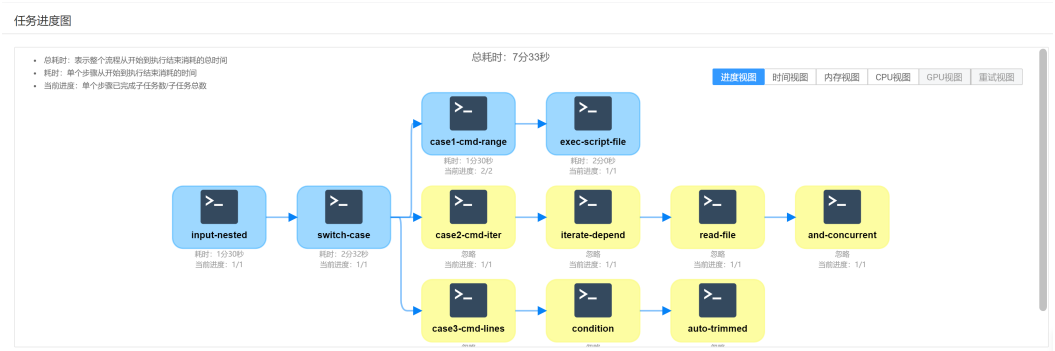
步骤3 确认无误后，请单击“开始”即可完成流程的创建。

----结束

查看分析结果

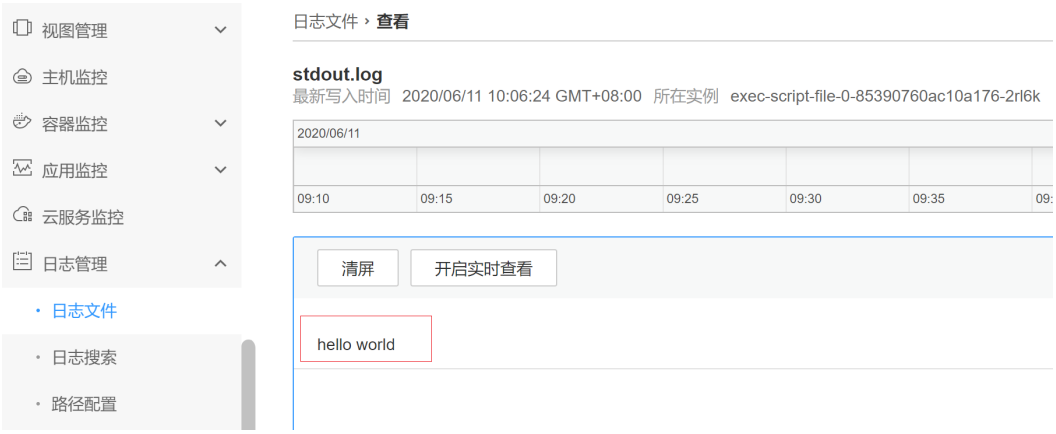
测序流程执行时间需要数小时，具体时长与环境资源类型、环境资源大小、处理数据大小等相关。使用测试数据，预计要20分钟。

执行过程可以通过“执行结果”页面查看，可查看内容包括“任务进度图”、“流程事件”、“任务事件”、“日志”、“输入”、“输出”和“监控”等。



gcs-grammar流程在exec-script-flie步骤会执行echo hello world命令，可以在日志中看到该步骤输出的日志。选中“日志”标签，找到exec-script-flie的输出日志文件。

单击“查看”，跳转到应用运维服务中，可以查看到stdout.log输出了hello world。



清理环境

流程执行完毕后，测序所使用的CCI计算资源会自动释放，无需担心产生多余的费用。

3 病毒基因组分析平台入门

病毒基因组分析平台基于rampart实时读取分析病毒数据，并且提供web界面查看病毒分析的结果。

说明

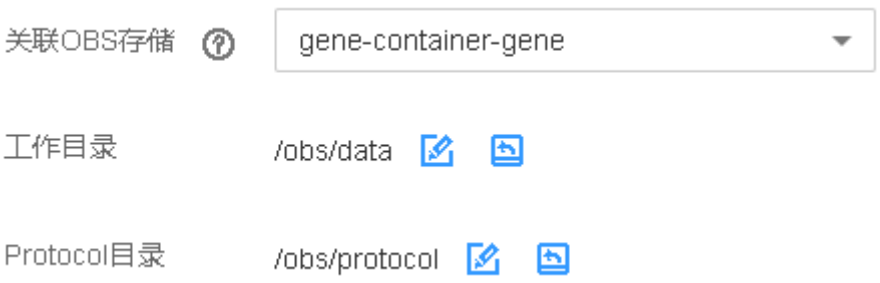
仅内测用户可以使用。

创建环境

单击“环境管理 > 创建环境”，根据界面提示填写参数。

- 环境类型
  选择“病毒基因组分析平台”。
- 关联OBS存储
  OBS存储用于存储分析数据，您需要在分析前将数据上传到OBS中，分析过程中会读取OBS中的数据。
  病毒基因组分析平台使用的镜像中默认带有示例数据和协议， 在“/root”目录下。
  - 工作目录：rampart的启动目录。单击 选择OBS桶中需要分析的数据目录，该目录会挂载到容器的“/obs”目录下，如图3-1所示。
  - Protocol目录：存放启动rampart使用协议的目录。单击 选择OBS桶中需要分析的数据目录，该目录会挂载到容器的“/obs”目录下，如图3-1所示。
  - 清除结果重跑：是否清楚之前分析结果，重新分析样本。

图 3-1 目录选择



- 命名空间

  病毒基因组分析平台底层使用CCI作为计算资源，此处选择CCI的命名空间。如果您在CCI中没有可用的命名空间，或不想使用已有命名空间，请单击“创建命名空间”，创建命名空间步骤请参见**命名空间**。

- 访问秘钥

  单击 单击上传 ，在弹出的对话框中上传已下载的访问密钥（AK/SK），单击“确认”。若没有访问密钥，请前往“我的凭证” 的**管理访问密钥**页面新增并下载访问密钥。

- 公网ELB实例

  访问rampart Web页面需要通过公网ELB实例。如没有公网ELB实例，请单击“创建增强型ELB实例”创建。

- 引擎规格

  底层CCI引擎的规格。

单击“下一步”，确认信息，然后单击“创建”。

查看分析结果

环境创建后，您可以查看分析过程和结果，复制下图中的访问地址，使用浏览器访问。

下图是rampart Web平台，您可以查看分析过程和结果。