**生信分析管理软件**

**需求说明**

# 范围

## 标识

1. 标题：智慧实验室建设软件需求规格说明；
2. 名词和缩略语：

缩略语见表1：

表1 缩略语

|  |  |  |
| --- | --- | --- |
| 序号 | 名称 | 本文档中的简称 |
|  | 生信分析管理软件 | 多组学数据管理 |

1. 本文档适用的系统：多组学数据管理软件。

## 系统概述

多组学数据管理软件内容包括自动化分析管理、生信工具管理、生信技术平台功能开发。

其中自动化分析管理作为测序下游生信分析调度平台，实现资源统一管理，分析任务统一调度，数据规范化分层管理。生信工具管理作为测序下游多组学工具的统一管理平台，工具可包括基因组学、转录组、外显子、宏基因组等多组学工具的管理，通过工具的统一管理，实现工具标准化、共享化，提升运维效率。可通过可视化方式管理、导入流程，实现全流程自动化。生信分析技术平台Variant calling（变异检测）是生物信息学中的重要分析流程，主要用于检测基因组中的单核苷酸变异（SNV）、插入缺失（Indel）等遗传变异。

# 需求描述

## 整体需求描述

本次项目建设的主要内容是开发一款多组学数据管理软件，为医务工作者和研究人员提供更加便捷、高效的生信自动化分析、可视化的工具管理及生信技术平台应用。具体包括地生信自动化分析、生信工具管理、生信技术平台三部分。

（1）生信自动化分析

①系统基于标准SGE调度方案的基础上结合业务流程支持多级别优先级调度；（1）

②针对分析任务提供可视化的管理界面；（2）

③分析流程所用的分析参数提供管理功能；（3）

④分析任务可视化，查看报告，查看日志。（4）

（2）生信工具管理

①应用导入及更新/安装；（5）

②应用展示；（6）

③开发者平台；（7）

（3）生信技术平台

①数据过滤、序列比对、变异位点检测、变异位点筛选、构建进化树；（8）

## 功能性需求

1. 提供可视化任务管理功能。（9）
2. 提供优先级管理。（10）
3. 分析任务可视化，查看报告，查看日志。（11）
4. 分析流程所用的分析参数提供管理功能。（12）
5. 应用导入、更新/安装，展示。（13）
6. 提供结核病生信分析技术平台（14）

## 非功能性需求

1. 界面要求：界面友好、操作便捷。（15）
2. 增加自动化程度，降低学习成本，具有可扩展性。（16）

# 数据处理需求详述

1. 自动化分析管理

作为测序下游生信分析调度平台，实现资源统一管理，分析任务统一调度，数据规范化分层管理。（17）

1. 生信工具管理

作为测序下游多组学工具的统一管理平台，工具可包括基因组学、转录组、外显子、宏基因组等多组学工具的管理，通过工具的统一管理，实现工具标准化、共享化，提升运维效率。平台可提供200+开源工具，可通过可视化方式管理、导入流程，实现全流程自动化。（18）

1. 生信分析技术平台

Variant calling（变异检测）是生物信息学中的重要分析流程，主要用于检测基因组中的单核苷酸变异（SNV）、插入缺失（Indel）等遗传变异。（19）

# 软件详细功能需求详述

## 自动化分析管理

作为测序下游生信分析调度平台，实现资源统一管理，分析任务统一调度，数据规范化分层管理。系统基于标准SGE调度方案的基础上结合业务流程支持多级别优先级调度。提供可视化操作界面，增加自动化程度，降低学习成本，具有可扩展性。



功能清单如下：

|  |  |  |
| --- | --- | --- |
| **序号** | **功能** | **详情** |
| 1 | 新增分析任务 | 支持自动化启动分析，通过对接测序仪实现下机数据自动分析。同时支持手动启动分析，支持分析流程、参数、metadata的设置。（20） |
| 2 | 分析任务管理 | 针对分析任务提供可视化的管理界面，包括任务的暂停、重跑、启动、删除等。（21） |
| 3 | 任务查看 | 提供批次、样本维度的任务查看，包括分析进度、耗时、资源利用等。（22） |
| 4 | 报告查看 | 提供分析包括的预览、下载等功能。（23） |
| 5 | 日志查看 | 提供分析日志、错误日志的在线查看。（24） |
| 6 | 优先级管理 | 针对分析任务提供流程、项目、样本三个维度的优先级调整功能（25） |
| 7 | 参数管理 | 针对分析流程所用的分析参数提供管理功能，包括参数的新增、修改、查询、应用等。方便标准化进行生信分析。（26） |
| 8 | 参考基因组管理 | 针对分析流程所用的参考基因组提供管理功能，包括基因组的新增、修改、查询、应用等。方便标准化进行生信分析。（27） |

## 生信工具管理模块

作为测序下游多组学工具的统一管理平台，工具可包括基因组学、转录组、外显子、宏基因组等多组学工具的管理，通过工具的统一管理，实现工具标准化、共享化，提升运维效率。平台可提供200+开源工具，可通过可视化方式管理、导入流程，实现全流程自动化。



功能清单如下：

|  |  |  |
| --- | --- | --- |
| **序号** | **功能** | **详情** |
| 1 | 应用导入 | 针对已有应用可直接通过docker导入系统，从而实现流程的标准化、自动化管理。（28） |
| 2 | 应用更新/安装 | 针对平台新发布的应用，可支持离线状态的应用导入，从而快速方便的获取最新应用。（29） |
| 3 | 应用展示 | 展示应用的整体介绍、说明文档、常规问题等。（30） |
| 4 | 应用使用统计 | 统计各个分析流程的使用情况，监控使用许可。（31） |
| 5 | 应用流程设计 | 选择已有的开源工具，通过托拉拽的可视化界面进行流程分步设计，支持参数、结果的自定义配置。（32） |
| 6 | 开发者平台 | 提供notebook开发者平台，内置R、python等软件，可直接进行代码级个性化分析。（33） |

## 生信技术平台

流程名称：variant calling

功能点：

1.数据过滤

软件接受二代测序平台生成的双端读取（fastq或fastq.gz格式），使用sickle（v 1.33）进行质量控制。该软件可以去除低质量的reads,而且还可以保证双端测序的完整性。我们默认保留碱基质量≥20，长度≥20的reads进行后续分析。（34）

2.序列比对

使用 bowtie2 （v 2.3.3）将保留的测序读数映射到参考基因组（M.tuberculosis NC\_000962.3）上，然后使用 SAMtools (v 1.6)计算映射结果，包括映射读数、平均深度、覆盖度。当平均深度值≥20且覆盖度≥95%时将比对质量>30序列进行变异位点检测。（35）

3.变异位点检测

使用VarScan（v 2.4.0）用于识别SNP和Indel。默认情况下，软件使用内部检测标准如下：碱基质量≥20、映射质量≥30、深度≥20、突变等位基因频率 ≥75%，但这些参数可以由用户修改。（36）

4.变异位点筛选

软件筛选位于特定基因组区域（PE/PPE基因家族、插入元件和重复区域）变异位点，这些区域在某些生物体（如结核分枝杆菌）中具有重要的功能和进化意义。（37）

5.构建进化树

筛选出样本间差异位点，根据参考菌株的序列碱基生成新的序列，使用Fasttree（v 2.1.9）构建进化树。（38）

**技术参数要求**

1、生信分析调度

△1.1分析任务管理，支持按照批次维度和样本维度的任务查询，支持单个及批量任务的新增、修改、查看、暂停、取消、删除等操作。支持批量分析、多样本单分析等多种分析方案。（39）

1.2列表展示：支持分析任务按批次或样本级别展示，包括包括芯片号、lane、分析状态、交付状态、日志、分析开始时间、分析完成时间、耗时等字段（40）

1.3分析资源配置，支持基于流程级别的资源匹配。（41）

1.4分析优先级管理，支持项目级、样本级的优先级配置，以满足不同的分析需求。（42）

2、工具管理

△2.1应用市场提供200+生信工具，支持已发布的应用查找、介绍及报告展示等（43）

2.2工具升级，已安装应用的在线和离线升级，包括升级日志、升级进度查看（44）

2.3工具使用记录，已安装应用的使用记录（45）

△2.4参数管理，针对不同流程的参数进行精细化管理，支持不同版本的参数管理（46）

2.5参考基因组管理，提供基础模板进行参考基因组管理，创建成功的参考基因组可直接应用于流程；（47）

3、生信工具集成

△3.1支持根据提供的结核分枝杆菌全基因组序列分析代码进行流程编写、docker环境封装和上线（48）

3.2支持结核分枝杆菌全基因组序列分析报告展示、结果预览和下载等（49）

3.3支持其他可选流程分析、开源工具分析（50）