

# 全基因组/全外显子组体细胞变异 SNVs+indels 基础分析平台 V1.0

## 使用说明书

---

2021 年 7 月

# 目录

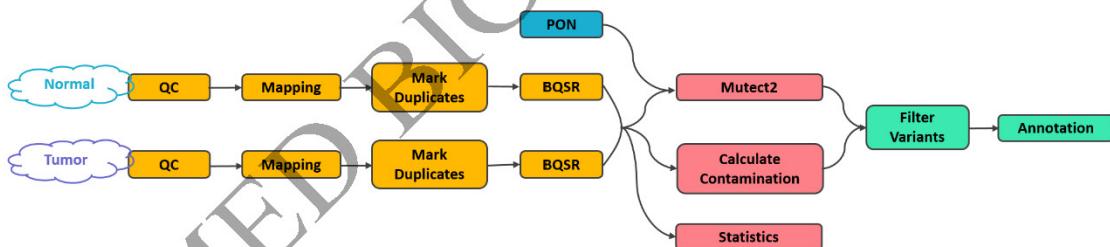
1. 系统简介.....	3
2. 系统常规操作.....	3
2.1 系统登录.....	3
2.2 用户注册.....	4
3. 功能介绍.....	5
3.1 系统介绍.....	5
3.1.1 简介.....	5
3.1.2 操作手册.....	6
3.1.3 联系我们.....	7
3.2 数据上传.....	7
3.2.1 文件上传系统.....	7
3.2.2 用户上传数据文件.....	10
3.3 标准数据分析.....	11
3.3.1 参数选择.....	11
3.3.2 数据分析结果.....	17

## 1. 系统简介

全基因组/全外显子组体细胞变异 SNVs+indels 基础分析平台 V1.0 是在精准医学研究发展的研究背景下，设计开发的全基因组/全外显子组体细胞突变位点检测分析平台系统。全基因组测序 (Whole Genome Sequencing, WGS) 是指使用二代测序技术 (NGS) 对基因组的所有区域进行测序。外显子测序(Whole Exome Sequencing, WES)是指利用序列捕获技术将全基因组外显子区域 DNA 捕捉并富集后，再进行高通量测序的基因组分析方法。

本系统支持单个样本或多个样本测序数据 fastq 文件体细胞突变分析，包括 tumor-only 和 tumor-normal 两种数据模式。用户可以使用自行上传的正常样本做背景参照 (PON)，也可以使用 GATK4 提供的公共参照 (PON) 进行体细胞突变位点检测分析。此外，当用户递交全外显子组测序分析任务时，可以使用自定义测序配套的外显子捕获区域用于过滤，默认使用 CCDS 提供的外显子区域。

本系统提供了全基因组/外显子组体细胞突变检测分析的标准化流程，整合了从原始数据、数据预处理、到分析结果下载与可视化查看等多项步骤成为一站式的自动化分析流程。同时，对于重要参数，用户可以使用默认推荐值或自定义选项，满足不同用户的个性化需求。

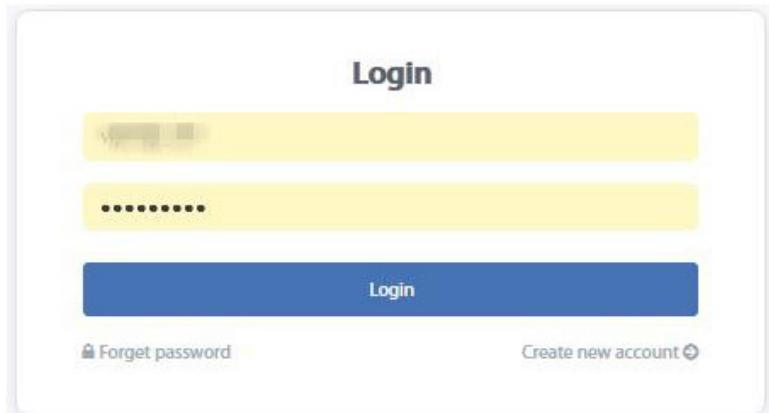


## 2. 系统常规操作

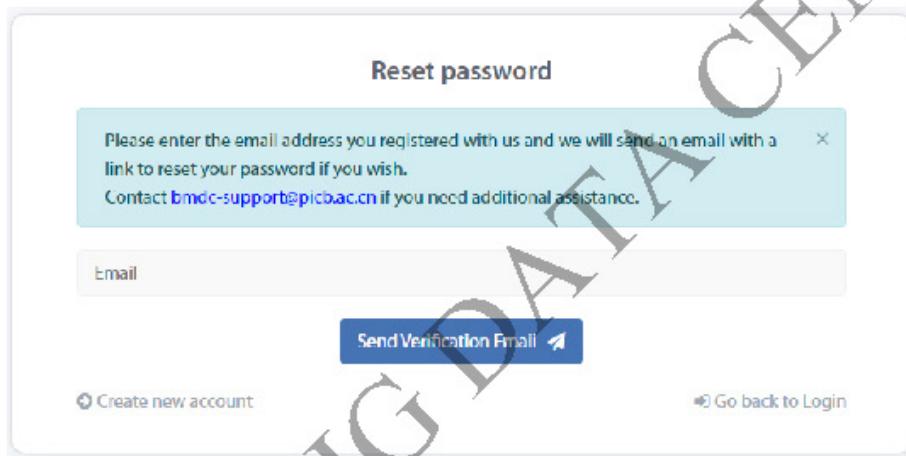
### 2.1 系统登录

第一步：访问 VIPMAP 网页；

第二步：若用户已在 NODE 网页或 VIPMAP 注册，可直接通过上图右上角 Sign in 进行登录。如下图，在登录页面输入账号密码，点击“Login”按钮。



若需要重新设置登录密码，点击左下角“Forgot password”进入下图页面，输入注册时使用的邮箱，点击下方按钮，等待系统发送确认邮件。



## 2.2 用户注册

若用户尚未注册，需要在登录页面点击“Create new account”进入用户注册界面，如下图：

## Register

\* required information

**Account information**

\* Email  \* This field is required.

\* Password  \* This field is required.

\* Confirm password  Confirm password can not be empty.

**Personal information**

\* First name  Middle name  \* Last name

\* Organization

Department

Title  Phone

Staff

\* Country/Region  Province/State  City   
Afghanistan

**Create account** 

 [Forget password](#)  [Back](#)

### 3. 功能介绍

#### 3.1 系统介绍

##### 3.1.1 简介

点击左侧菜单栏中的“Introduction”，进入系统简介页面，用户可通过此页面了解本系统的功能和流程。

- Analysis**
- > RNASeq-DEG
- > RNA-Seq-ASE-Paean
- > scRNA-Seq
- WES/WGS-somatic**
- Introduction
- Start Analysis
- > WES/WGS-germline
- > Proteomics

### Introduction

**Summary**

Whole exome sequencing (WES) is a genomic technique for sequencing all of the protein-coding regions of genes in a genome.

For this models, we will provide a pipeline to identify somatic mutations from WES dataset.

**Pipeline**

This pipeline starts by reading in the raw fastq data, and return an annotated somatic mutation file.

[Next](#)

© 2021 Shanghai Institute of Nutrition and Health, CAS

### 3.1.2 操作手册

点击菜单栏“About”中的 Guidelines，进入操作手册下载页面，如下图：

- Analysis**
- > RNASeq-DEG
- > RNA-Seq-ASE-Paean
- > scRNA-Seq
- WES/WGS-somatic**

### Introduction

**Summary**

Whole exome sequencing (WES) is a genomic technique for sequencing all of the protein-coding regions of genes in a genome.

For this models, we will provide a pipeline to identify somatic mutations from WES dataset.

点击 WES/WGS-somatic guideline 下的“Download”按钮，下载对应的操作手册（pdf 文档），如下图：

**Guideline**

► Guidelines will continued to be released and updated. For more questions, please contact us directly by email or telephone.

scRNA-Seq guideline <a href="#">Download</a>	RNASeq-DEG guideline <a href="#">Download</a>	WES/WGS-somatic guideline <a href="#">Download</a>
---	--	---

### 3.1.3 联系我们

点击菜单栏“About”中的Contact us，进入“联系我们”页面，如下图：

The screenshot shows the VIPMAP platform's navigation bar with 'Home', 'Analysis', 'About', and 'Links'. The 'About' section is active, displaying 'Guidelines' and 'Contact us' buttons. The 'Contact us' button is highlighted in blue. To the left, there is a sidebar for 'Analysis' with options like 'RNASeq-DEG', 'RNA-Seq-ASE-Paean', 'scRNA-Seq', and 'WES/WGS-somatic'. The main content area shows an 'Introduction' section with a 'Summary' heading and some descriptive text about Whole exome sequencing (WES).

若用户对于页面使用、数据上传、结果获取与解读等有任何疑问或建议，可以通过下图页面中提供的邮箱、电话等信息同开发团队进行联系。

The screenshot shows the 'Contact Us' page with contact details: Phone (+86-21-54920457), Email (bioinfo@picb.ac.cn), and Address (Bio-Med Big Data Center, CAS-MPG Partner Institute for Computational Biology (PICB), Shanghai Institute of Nutrition and Health (SINH), Chinese Academy of Sciences, 320 Yueyang Rd., the Main Building 309, Shanghai 200031, China). It also includes a detailed map of Shanghai with the SINH building location marked, and a zoomed-in view of the surrounding area.

### 3.2 数据上传

#### 3.2.1 文件上传系统

用户通过主页面右下角的“SUBMIT DATA”按钮提交数据，进入文件管理系统（File Manager System），如下图：

**VIPMAP**  
Visualization Integrated Precision Medicine Analysis Platform

Home      Analysis      About      Links

**Analysis**

- > RNASeq-DEG
- > RNA-Seq-ASE-Pearson
- > scRNA-Seq
- > **WES/WGS-somatic**
  - Introduction
  - Start Analysis
- > WES/WGS-germline
- > Proteomics

**Introduction**

**Summary**

Whole exome sequencing (WES) is a genomic technique for sequencing all of the protein-coding regions of genes in a genome.

For this models, we will provide a pipeline to identify somatic mutations from WES dataset.

**Pipeline**

This pipeline starts by reading in the raw fastq data, and return an annotated somatic mutation file.

**Next**

© 2021 Shanghai Institute of Nutrition and Health, CAS

进入文件管理系统界面后，用户可通过左上角查看目录结构，在页面左下角可以查看目录总大小的限制和当前使用量。

**File Manager System**

注意：大小超过100M的文件，请下载ftp客户端软件FileZilla，使用注册用户名和密码登录http://www.biosino.org:26007上传文件

**Delete**

File Name	Size	Time	Action
...			刪除
wes			刪除

**Upload Files** **New Folder**

Vipmap iMac  
620GB / 1TB Used

为方便管理不同批次的数据，用户可点击右侧的“New Folder”在文件管理系统中创建新的文件夹来存放待分析的数据。

在弹出窗口中输入自定义目录名称，点击“Save”按钮创建成功，如下图。

**New Folder**

Folder Name

**Cancel** **Save**

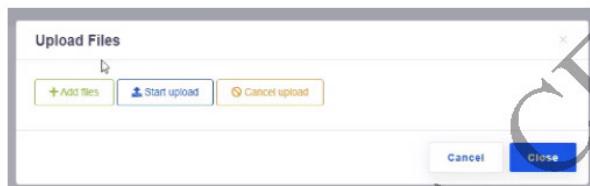
**Upload Files** **New Folder**

Size	Time	Action
		刪除

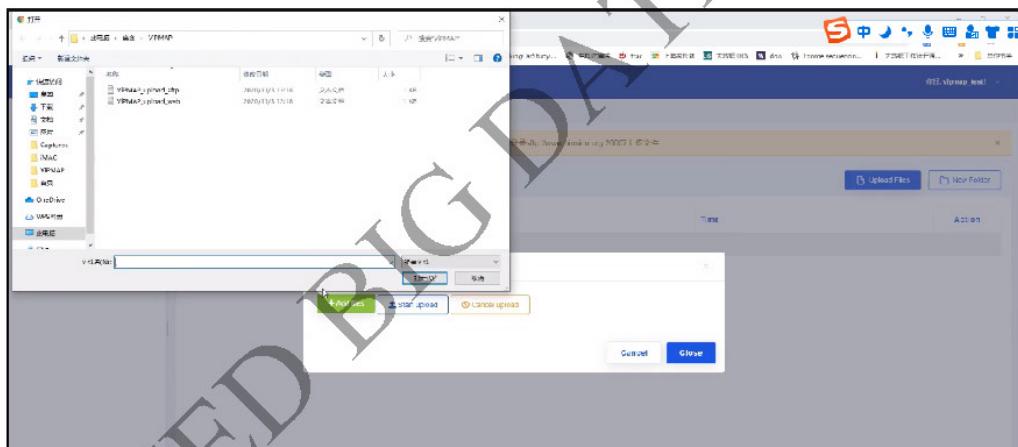
点击新创建好的文件夹名称，进入该目录。



点击“Upload Files”，进行网页版上传。

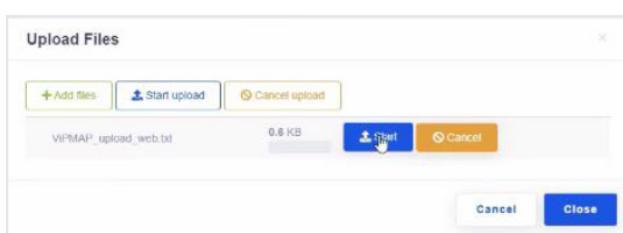


点击“Add files”按钮，在本地文件中选择需要上传的文件，如下图。



点击“Start”，开始上传文件，根据进度条的提示查看文件是否已上传成功。

如下图：



上传成功后，文件即可在目录中查看。如下图：

<input type="checkbox"/>	File Name	Size	Time	Action
<input type="checkbox"/>	VIPMAP_upload_sftp.txt	0.55 K	2020-11-03 16:49:56	
<input type="checkbox"/>	VIPMAP_upload_web.txt	0.55 K	2020-11-01 16:49:31	

如需要上传的待分析文件大小超过 100M，则可通过 ftp 客户端软件 FileZilla 或直接邮寄硬盘到开发团队。

打开 FileZilla 软件，输入主机名，和 FileManager 的用户名及密码，点击快速链接即可进入个人 FTP 目录。点击左侧窗口右键选中文件并完成上传，即可将本地文件上传至 FTP 指定目录中。如下图：

The screenshot shows the FileZilla client interface. A context menu is open over a file named 'VIPMAP\_upload\_web.txt' in the left-hand file browser. The menu options include '粘贴' (Paste), '剪切' (Cut), '删除' (Delete), '发送到' (Send To), '属性' (Properties), '发送' (Send), '另存为' (Save As), and '发送到' (Send To). Below the client window is a screenshot of the FileManager upload interface. The table shows the uploaded file 'VIPMAP\_upload\_web.txt' with a size of 0.55 K and a time of 2020-11-03 16:49:31.

<input type="checkbox"/>	File Name	Size	Time	Action
<input type="checkbox"/>	VIPMAP_upload_sftp.txt	0.55 K	2020-11-03 16:49:56	
<input type="checkbox"/>	VIPMAP_upload_web.txt	0.55 K	2020-11-03 16:49:31	

### 3.2.2 用户上传数据文件

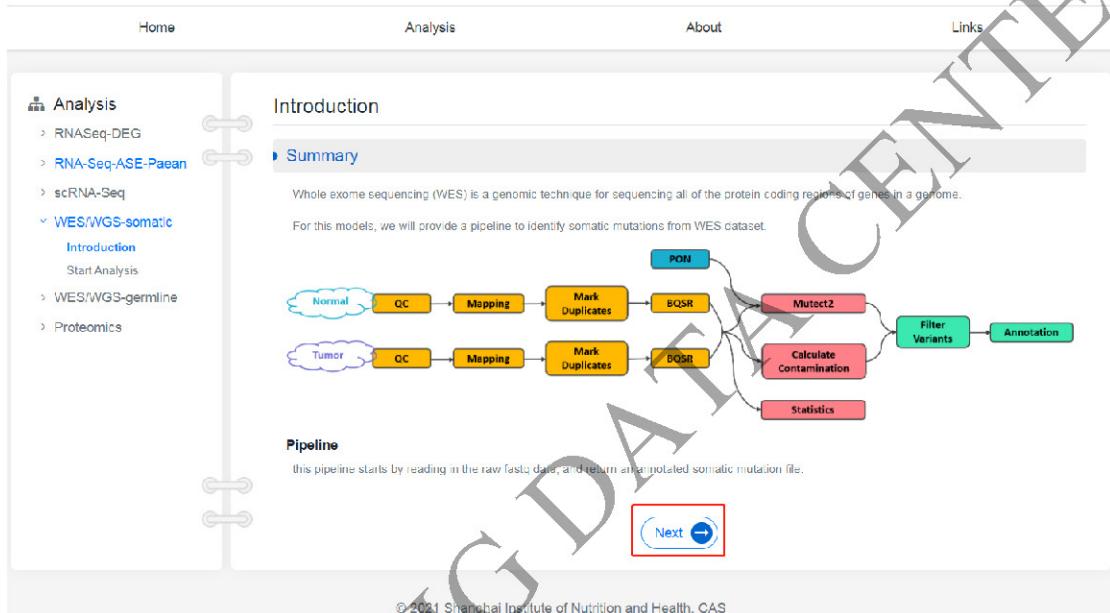
用户需要预先上传测序数据到文件管理系统，二代测序数据文件需要以“fastq.gz”、“fq.gz”、“fastq”或“fq”结尾，具体 fastq 文件格式请参考 <https://support.illumina.com/bulletins/2016/04/fastq-files-explained.html>。

若数据为外显子组测序，请上传外显子区域文件，由于文件较小，可通过网页进行文件上传。网页上可默认使用 cds 文件作为外显子区域进行过滤 ([ftp://ftp.ncbi.nlm.nih.gov/pub/CCDS/current\\_human/CCDS.current.txt](ftp://ftp.ncbi.nlm.nih.gov/pub/CCDS/current_human/CCDS.current.txt))，但不同公司使用的外显子捕获区域不同，尽量使用测序数据配套外显子捕获区域。请注意

使用 hg38 版本的 bed 文件，bed 文件至少包含“染色体”、“起始位置”、“终止位置”三列，并以“tab 键”进行分隔，BED 文件中起始坐标从 0 开始计数，具体可参考 <https://bedtools.readthedocs.io/en/latest/content/general-usage.html>。

### 3.3 标准数据分析

用户可以通过“Introduction”页面的“Next”或 WES/WGS-somatic 下的“Start Analysis”进入任务递交页面，如下图：



#### 3.3.1 参数选择

The screenshot shows the 'WES/WGS-somatic' task configuration page. The sidebar menu is identical to the previous screenshot. The main area has a title 'WES/WGS-somatic' and a 'Add Task' button. Below it is a 'Task List' table with columns: 'RunName', 'R1(tumor)', 'R2(tumor)', 'R1(normal)', and 'R2(normal)'. Each column has a 'Select' button. Underneath the table is a 'Quality Control' section with a 'Trimomatic' link. A red box highlights the 'Start Analysis' button in the sidebar.

本流程支持 tumor-only、tumor-normal 两种模式。

tumor-only 模式为用户仅使用肿瘤(tumor)数据鉴定体细胞变异位点，缺失匹配的正常(normal)样本对结果影响较大，不推荐使用。此模式下，用户需要填写

样本名称(RunName)、肿瘤样品左端文件(R1(tumor))和肿瘤样品右端文件(R2(tumor)), 如下图:

#### WES/WGS-somatic

Add Task		Task List			
选择文件		未选择任何文件			
RunName	R1(tumor)	R2(tumor)	R1(normal)	R2(normal)	
test	Select test.tum... <span style="color:red;">*</span>	Select test.tum... <span style="color:red;">*</span>	Select	Select	<span style="color:blue;">+</span> <span style="color:blue;">-</span>

tumor-normal 模式为一个肿瘤样本与正常样本相匹配。需要用户将配对的肿瘤(tumor)和正常(normal)测序文件名填写在同一样本名称(RunName)下, 如下图:

#### WES/WGS-somatic

Add Task		Task List			
选择文件		未选择任何文件			
RunName	R1(tumor)	R2(tumor)	R1(normal)	R2(normal)	
test	Select test.tum... <span style="color:red;">*</span>	Select test.tum... <span style="color:red;">*</span>	Select test.nor... <span style="color:red;">*</span>	Select test.nor... <span style="color:red;">*</span>	<span style="color:blue;">+</span> <span style="color:blue;">-</span>

用户可以选择仅选择正常样本用于 PON 的构建, 即肿瘤(tumor)信息为空。但至少需要一组肿瘤(tumor)样品用于鉴定体细胞突变, 如下图:

#### WES/WGS-somatic

Add Task		Task List			
选择文件		未选择任何文件			
RunName	R1(tumor)	R2(tumor)	R1(normal)	R2(normal)	
test1	Select	Select	Select test.nor... <span style="color:red;">*</span>	Select test.nor... <span style="color:red;">*</span>	<span style="color:blue;">+</span> <span style="color:blue;">-</span>
test2	Select test.tum... <span style="color:red;">*</span>	Select test.tum... <span style="color:red;">*</span>	Select	Select	<span style="color:blue;">+</span> <span style="color:blue;">-</span>

若用户递交少量样本, 可直接通过网页在线表格填写, 点击 + 添加新的样本, 样本名称不允许重复, 且样本名称为字母、数字或“\_”的组合。用户也可以通过 - 删除数据, 如下图:

## WES/WGS-somatic

Add Task		Task List			
选择文件 未选择任何文件		上传	下载模版		
RunName	R1(tumor)	R2(tumor)	R1(normal)	R2(normal)	
test	Select test.tum... *	Select test.tum... *	Select test.nor... *	Select test.nor... *	
test2	Select	Select	Select	Select	
test3	Select	Select	Select	Select	

若用户同本示例一样需要大批量运行一批数据，可通过任务提交页中的“下载模板”下载名为《somatic\_template.xlsx》的模板文件。同在线提交一样填写表格，若肿瘤样本的左端测序数据 R1(tumor)选择的是/wes/test 文件夹下的 test\_data.tumor.R1.fq.gz 则 R1(tumor)填充为/wes/test/test\_data.tumor.R1.fq.gz，如下图：

## WES/WGS-somatic

The screenshot shows a Windows file save dialog box. The file name is 'somatic\_template.xlsx' and the save type is 'XLSX 文件 (\*.xlsx)'. The dialog box is overlaid on a web browser window titled 'WES/WGS-somatic' which displays a task submission form. The 'Download Template' button in the browser is highlighted with a red box.

点击“选择文件”选择填好的表格，再点击“上传”文件，根据上传的 excel 表格自动填充好文件表格，用户如需要调整，也可直接在网页上操作增删、修改输入文件，如下图：

### WES/WGS-somatic

用户上传表格后会，系统自动检测表格中的文件是否在文件管理系统，若文件不存在，则报错，如下图：

目前，体细胞分析流程质控步骤使用 Trimmomatic 进行数据过滤，控制数据质量及左右两端测序数据的匹配。使用 BWA mem 将序列比对到参考基因组上，

如下图。参考基因组默认使用 GATK4 官方提供的 hg38 参考基因组进行比对，若用户需要其他物种或人的其他版本进行比对，可与我们联系。

The screenshot shows the GATK4 Somatic Mutation Calling interface. Under 'Quality Control', 'Method' is set to 'Trimmomatic'. Under 'Mapping', 'Species' is set to 'Homo Species (Human)' and 'version' is set to 'hg38(GRCh3)'. A note says 'Or if you want to use your own genome please contact us'. Under 'Method', 'BWA' is selected.

在鉴定体细胞突变时(Somatic Mutation Calling)，用户需要选择是否及使用哪个文件做 PON。PON 文件有如下两个特点：(1) 都来自正常的样本。(2) 目的是为了减少 variant calling analysis 中的假阳性结果，提高分析的准确性。使用与肿瘤技术尽可能相似的正常基因(相同的外显子组或基因组制备方法，测序技术等等)是非常重要的。对于多少个样本可以构成 PON，并没有明确的说法(即使是小的 PON 也比没有 PON 好)，但 GATK 根据经验给的建议是最少使用 40 个样本来构建。

对于人类样本，通常最好使用自有的 normal 数据集构建 PON。因为测序仪系统错误通常是相同的，用户可使用本流程的默认选项“Public Panel”，使用 GATK4 官方公共的 PON，1000g\_pon.hg38.vcf.gz 来自千人基因组健康人的血液样本，如下图：

The screenshot shows the 'Somatic Mutation Calling' section. 'Method' is set to 'GATK4'. Under 'Panel of Normal(PON)', 'Public Panel' is selected. A note says 'Or if you want to use your own normal samples to make a PON please contact us, and it is usually only recommended if you have more than 40 samples.' There are also options for 'Making a PON' and 'None'.

用户可以选择“Making a PON”使用上传的 normal 样本建 PON。用户选择后，会自动选择所有上传的 normal 样本用于建 PON，用户可以通过复选框去除不想用于建 PON 的样本，如下图：

### Somatic Mutation Calling

Method  GATK4

Panel of Normal(PON)  Public Panel  Making a PON  None

RunName	normal
test1	<input checked="" type="checkbox"/>
test2	<input checked="" type="checkbox"/>
test3	<input checked="" type="checkbox"/>
test4	<input checked="" type="checkbox"/>
test5	<input checked="" type="checkbox"/>

若用户在使用 mutect2 分析体细胞突变时，不想使用 PON，可选择“None”，如下图：

Somatic Mutation Calling

Method  GATK4

Panel of Normal(PON)  Public Panel  Making a PON  None

Or if you want to use your own normal samples to make a PON please contact us, and it is usually only recommended if you have more than 40 samples.

分析全外显子组测序时，使用外显子组区域进行过滤及统计，默认使用 cds 区域进行过滤（“Statistics”-“Protein Coding Regions CCDS”），如下图：

Statistics

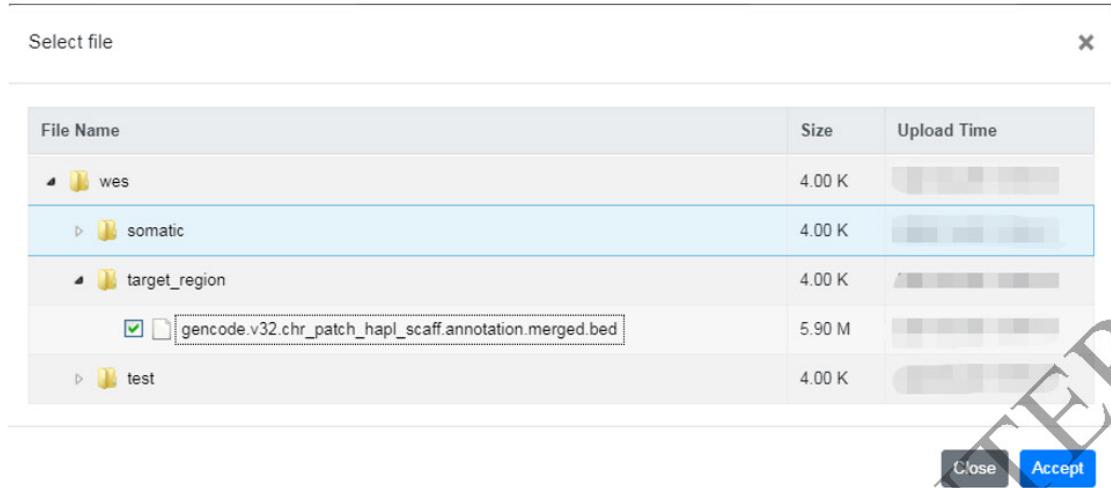
Regions file   Protein Coding Regions CCDS  upload your own target regions  WGS

因各公司使用的外显子捕获区域不同，建议用户通过“upload your own target regions”选择并上传配套的外显子组区域文件，如下图：

Statistics

Regions file   Protein Coding Regions CCDS  upload your own target regions  WGS

Upload:



### Statistics

Regions file ?  Protein Coding Regions CCDS  upload your own target regions ?  WGS  
 Upload:  gencod...

如果用户分析全基因组数据，请选择“WGS”，如下图：

### Statistics

Regions file ?  Protein Coding Regions CCDS  upload your own target regions ?  WGS

注释 vcf 文件默认使用 ANNOVAR 进行注释，选择完成后，点击“Submit”递交任务，如下图：

### Annotation

Method  ANNOVAR

## 3.3.2 数据分析结果

### 3.3.2.1 任务进程

任务提交后可以在“Task List”中搜索、查看任务状态。体细胞突变流程中，单个样品名称对应一条记录。批量递交的任务会生成一个统一的批次号“Batch ID”、每一个任务对应独立的任务号“Task ID”，如下图：

## WES/WGS-somatic

Task List							
Task ID	Batch ID	Run Name	Start time	Status	Status time	Consuming	Action
w2011170001	201117163525	test	2020-11-17 16:35:25	● Analysis done	2020-11-18 09:38:42	17小时3分	

用户可以通过任务号、批次号或时间检索任务记录，进行任务号(Task ID)和批次号(Batch ID)及进度(Status)的查看。每个任务可以通过点击 进入任务结果详情页，每个任务页面内也包含了任务号、批次号、起始时间和任务状态，如下图。

任务状态包括“Analysis is in preparation”、“Start Analysis”、“Contamination Calculation Done”、“Mutect2 Done”、“Variants Filtering Done”和“Analysis done”。任务失败将显示“Analysis Error”。任务递交成功后，页面将显示“Analysis is in preparation”，数据全部拷贝到输入文件夹下后显示“Analysis is in preparation”，后端任务开始时，任务状态栏显示“Start Analysis”，“Contamination Calculation Done”、“Mutect2 Done”、“Variants Filtering Done”为体细胞突变流程分析过程中的任务状态，具体可参考 somatic 起始页上的流程图。任务完成后，状态显示“Analysis done”。

## WES/WGS-somatic

Task List							
Task ID	Batch ID	Run Name	Start time	Status	Status time	Consuming	Action
WES20092500007	200925135426	test	2020-09-25 13:54:27	● Start Analysis	2020-09-25 13:55:22	0分	

## WES/WGS-somatic

[View Result](#)

Start Analysis

TaskNo : WES20092500007

Batch : 200925135426

StartTime : 2020-09-25 13:54:27

### 3.3.2.2 任务参数

任务结果详情页中，除了上述信息外，还显示了在线提交时使用的所有参数信息(Parameters)，如下图：

#### - Parameters

RunName	R1(tumor)	R2(tumor)	R1(normal)	R2(normal)
test	/wes/somatic/test.tumor.R1.fq.gz	/wes/somatic/test.tumor.R2.fq.gz	/wes/somatic/test.normal.R1.fq.gz	/wes/somatic/test.normal.R2.fq.gz

Quality Control

Method Trimmomatic

Mapping

Species Homo sapiens Version hg38(GRCh38)

Method BWA

Somatic Mutation Calling

Method GATK4

Panel of Normal(PON) Public Panel

Statistics

Regions file Protein Coding Regions(CCDS)

Annotation

Method ANNOVAR

任务成功后，结果界面除任务状态及参数详情外，还将显示结果信息“Analysis Result”。结果可以查看数据的测序质量、覆盖度及测序深度、vcf 及注释文件下载。

#### 3.3.2.3 数据质量控制

“Quality Control”中的表格包含文件名、总 reads 数(Total\_reads)、总碱基数(Total\_bases)、序列长度(Sequence\_length)、GC 含量(GC\_pct)、Q20 比例(Q20\_pct)和 Q30 比例(Q30\_pct)。

表格分别统计了左右两端序列的质量情况。以肿瘤样本(tumor)样本的左端(R1)文件为例：{RunName}.tumor.R1 为原始下机数据的质量信息，{RunName}.tumor.trim.R1 为质控后数据的质量信息。

表头	说明
FileName	样本名称
Total_reads	序列 reads 数
Total_bases	序列碱基数
Sequence_length	序列长度
GC_pct	GC 含量，DNA 4 种碱基中，鸟嘌呤和胞嘧啶所占的比率
Q20_pct	通过质控过滤正确率达到 99% 的碱基所占的比例
Q30_pct	通过质控过滤正确率达到 99.9% 的碱基所占的比例

备注：Q20 一般要求>90%， Q30 一般要求>85%。

- Analysis Result

Quality Control

FileName	Total_reads	Total_bases	Sequence_length	GC_pct	Q20_pct	Q30_pct
test.normal.R1	21277723	3191658450	150	52	97.16	92.64
test.normal.R2	21277723	3191658450	150	51	96.88	91.93
test.normal.trim.R1	20374463	3056014126	36-150	52	97.17	92.66
test.normal.trim.R2	20374463	3056046646	36-150	51	96.88	91.92
test.tumor.R1	80182116	12027317400	150	52	97.48	93.15
test.tumor.R2	80182116	12027317400	150	52	96.93	91.76
test.tumor.trim.R1	78478877	11771581685	36-150	52	97.49	93.16
test.tumor.trim.R2	78478877	11771625587	36-150	52	96.93	91.75

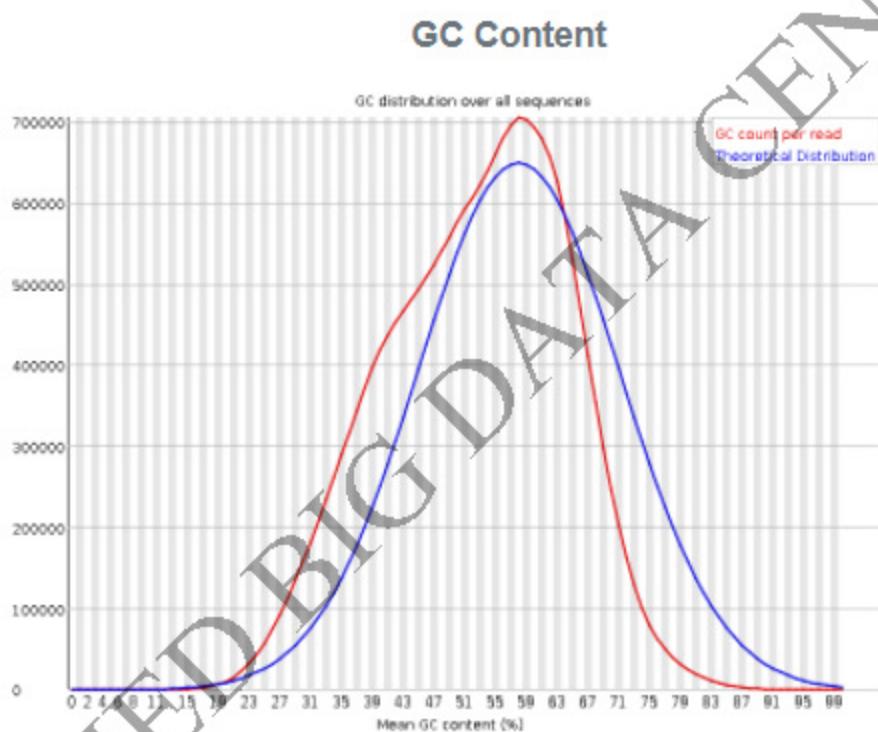
网页在线展示了 Per base sequence quality 和 Per Sequence GC content 两个统计图，下面主要对这两种图做说明。

下图为碱基质量箱线图。横轴代表位置，纵轴为质量得分。黄色柱状是 25%-75% 区间，中间红线是中位数，error bar 是 10%-90% 区间，蓝线是平均数。分值越高代表质量越好。背景图划分 y 轴为高质量（绿色）、可接受质量（橙色）和低质量（红色）。若任一位置的下四分位数低于 10 或中位数低于 25，FastQC 报告显示“WARN”；若任一位置的下四分位数低于 5 或中位数低于 20，报“FAIL”。

用户可以通过 FileName 后的下拉菜单控制选择想要查看的样本的数据质量。



下图为 reads 平均 GC 含量分布图。横轴表示 GC 含量，纵轴表示不同 GC 含量对应的 read 数，蓝线是理论分布（正态分布，通过从所测数据计算并构建理论分布），红色是实际情况，两个比较接近判为好的。曲线形状的偏差往往是由文库的污染或是部分 reads 构成的子集有偏差 (overrepresented reads)；形状接近正态分布但偏离理论分布的情况提示我们可能有系统偏差；如果出现两个或多个峰值，表明测序数据里可能有其他来源的 DNA 序列污染，或者有接头序列的二聚体污染。偏离理论分布的 reads 超过 15% 时，FastQC 报告显示 "WARN"；偏离理论分布的 reads 超过 30% 时，报 "FAIL"。



### 3.3.2.4 数据质量控制

“Mapping”中的表格包括样本名 (Name)，比对上的 Reads 及比例 (Mapped\_Reads)、正确配对的 reads 数量及比例(Properly\_Mapped\_Reads)、平均测序深度 (Average\_Depth)、覆盖度(Coverage)，如下图：

表头	说明
Name	样本名，癌症样本名为(RunName).tumor、正常样本名为(RunName).normal
Mapped_Reads	比对上的 Reads 及比例 (总体比对率)

---

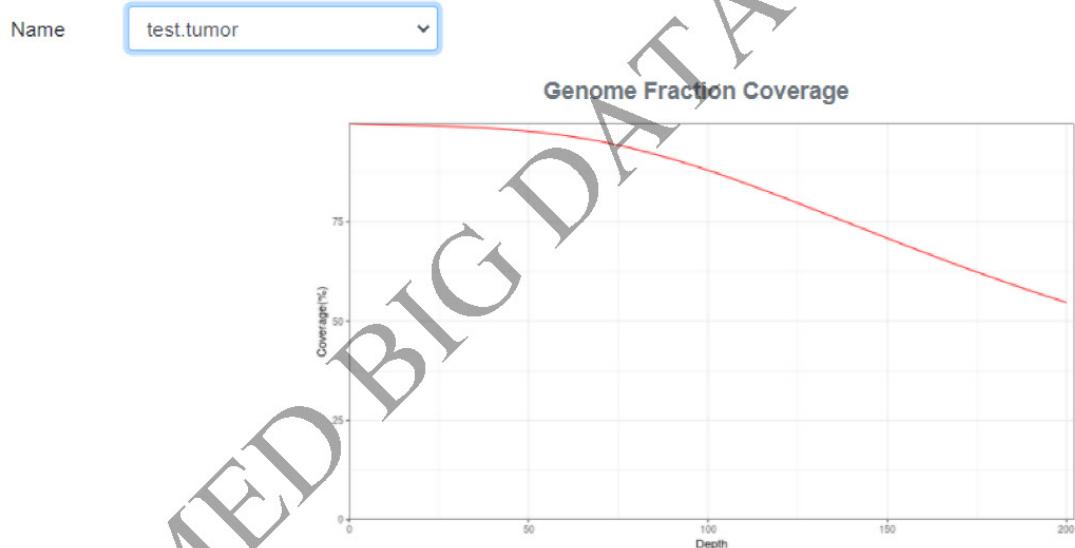
Properly_Mapped_Reads	正确配对的 reads 数量及比例。比对到同一条参考序列，并且两条 reads 之间的距离符合设置的阈值
Average_Depth	平均测序深度
Coverage	覆盖度

---

#### Mapping

Name	Mapped_Reads	Properly_Mapped_Reads	Average_Depth	Coverage
test.normal	40698587(99.29%)	40366634(98.48%)	72.65	99.59%
test.tumor	156761435(99.28%)	153475512(97.20%)	277.31	99.73%

下图为测序深度-覆盖度关系图，横坐标为测序深度、纵坐标为覆盖度。可以查看给定测序深度下，外显子组或基因组的覆盖程度。使用配对样本时，用户可以通过 Name 后的下拉菜单选择肿瘤或正常样本，从而查看相应覆盖度随测序深度变化曲线。



#### 3.3.2.5 肿瘤突变位点检测

体细胞突变结果文件下载包括 somatic 突变分析的 vcf 文件及 annovar 注释的 txt 文件，如下图：

##### Somatic Mutations

- [vcf-format file of mutations](#)
- [Annotated file by ANNOVAR](#)

VCF，即 Variant Call Format，是用于描述 SNP, INDEL 和 SV(structural variation calls)结果的文件。下文简要介绍一下 vcf 中的字段信息，具体可参考

<https://gatk.broadinstitute.org/hc/en-us/articles/360035531692-VCF-Variant-Call-Format> 和 [https://docs.gdc.cancer.gov/Data/File\\_Formats/VCF\\_Format](https://docs.gdc.cancer.gov/Data/File_Formats/VCF_Format), 如下图:

```
##fileformat=VCFv4.2
##FILTER=<ID=FAIL,Description="Fail the site if all alleles fail but for different reasons.">
##FILTER=<ID=PASS,Description="Site contains at least one allele that passes filters">
##FILTER=<ID=base_qual,Description="alt median base quality">
...
##FORMAT=<ID=AD,Number=R,Type=Integer,Description="Allelic depths for the ref and alternate alleles in the sample">
##FORMAT=<ID=AF,Number=A,Type=Float,Description="Allele fractions of alternate allele">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Approximate read depth (read coverage)">
##FORMAT=<ID=F1R2,Number=R,Type=Integer,Description="Count of reads in F1R2 pair">
##FORMAT=<ID=F2R1,Number=R,Type=Integer,Description="Count of reads in F2R1 pair">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=PGT,Number=1,Type=String,Description="Physical phasing haplotype information">
##FORMAT=<ID=PID,Number=1,Type=String,Description="Physical phasing ID informative">
##FORMAT=<ID=PL,Number=G,Type=Integer,Description="Normalized, Phred-scaled likelihoods for genotypes (0-255)">
##FORMAT=<ID=PS,Number=1,Type=Integer,Description="Phasing set (typically the positive strand)">
##FORMAT=<ID=SB,Number=4,Type=Integer,Description="Per-sample component statistic">
...
##MutectVersion=2.2
##contig=<ID=chr1,length=248956422>
##contig=<ID=chr2,length=242193529>
...
##filtering_status=These calls have been filtered by FilterMutectCalls to label false positives
##normal_sample=Patient65_TPD_NB.normal
##source=FilterMutectCalls
##source=Mutect2
##tumor_sample=Patient65_TPD_NB.tumor
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT Pa
chr1 37762037 . G A . PASS AS_FilterStatus=SITE;AS_SE
chr1 40462711 . A G . PASS AS_FilterStatus=SITE;AS_SE
...
```

VCF 文件由两个主要部分组成:

头文件以‘##’为前缀，通常包含 fileformat、fileDate、reference 等信息。

变异记录为主体部分，记录了每个样品每个位点处的基因分型信息。

主体部分每列的含义：

- 1) CHROM - chromosome: 参考基因组标识。
- 2) POS - position: 变异位点相对于参考基因组所在的位置(1-based)。如果变异大于一个碱基则为第一个碱基的位置。
- 3) ID - identifier: 变异的 id 号；通常若在 dbSNP 中有该 SNP 的 rs 号，则会在此行给出。否则，默认使用‘.’。
- 4) REF - reference base(s): 参考序列碱基，必须是 A,C,G,T,N 其中的一种
- 5) ALT - alternate base(s): 表示 variant 的 Allele，若多个，则使用逗号分隔，(变异所支持的碱基类型及碱基数量)这里的碱基类型和碱基数量，对于 SNP 来说是单个碱基类型的编号，而对于 Indel 来说是指碱基个数的添加或缺失，以

及碱基类型的变化

- 6) QUAL - quality: 表示 Phred 质量值, 用来表示 ALT 的可靠性
- 7) FILTER - filter status: 表示是否通过过滤。PASS 表示该位点通过过滤, 否则表示没有通过。如果这一栏是一个“.”的话, 就说明没有进行过任何过滤。
- 8) INFO - additional information: 表示的是变异描述信息。以<key>=[,data]格式, 并使用分号分隔的形式, 其中很多的注释信息在 VCF 文件的头部注释中给出。
- 9) FORMAT: 可选的扩展, 例如 GT:AD:AF:DP:F1R2:F2R1:SB。该部分是主体部分, 表示基因型信息的多个标签, 这些标签之间以冒号分割, 其对应的值位于第 10 列, 同样以冒号分割, 表示第一个样品的基因型结果。
- 10) SAMPLES: 表示样本信息, 各个 Sample 的值, 由 BAM 文件中的@RG 下的 SM 标签所决定, 这些值对应着第 9 列的各个格式, 不同格式的值用冒号分开, tumor-only 模式下第 10 列为 tumor 信息, tumor-normal 模式第 10 列为 normal 信息、第 11 列为 tumor 信息。

---

标签	含义
GT	此位点样品的基因型 (genotype), 两个数字中间用‘/’分开, 这两个数字表示双倍体的样本的基因型。0 表示参考基因组的碱基类型; 1 表示 ALT 碱基类型的第一个碱基; 2 表示 ALT 碱基类型的第二个碱基。 0/0 表示 sample 中该位点为纯合的, 和 ref 一致; 0/1 表示 sample 中该位点为杂合的, 有 ref 和 variant 两个基因型; 1/1 表示 sample 中该位点为纯合的, 和 variant 一致。
AD	样品中每一种 allele 的 reads 覆盖度, 以“,”分隔, 前者对应 ref 基因型, 后者对应 variant 基因型。
AF	表示在 tumor 中 allele 的频率
DP	此位点的测序深度。

---

GT 此位点样品的基因型 (genotype), 两个数字中间用‘/’分开, 这两个数字表示双倍体的样本的基因型。0 表示参考基因组的碱基类型; 1 表示 ALT 碱基类型的第一个碱基; 2 表示 ALT 碱基类型的第二个碱基。

0/0 表示 sample 中该位点为纯合的, 和 ref 一致;

0/1 表示 sample 中该位点为杂合的, 有 ref 和 variant 两个基因型;

1/1 表示 sample 中该位点为纯合的, 和 variant 一致。

---

AD 样品中每一种 allele 的 reads 覆盖度, 以“,”分隔, 前者对应 ref 基因型, 后者对应 variant 基因型。

---

AF 表示在 tumor 中 allele 的频率

---

DP 此位点的测序深度。

---