诺禾致源 元基因组交付目 录说明手册 (V5.0)



目录

(注: 单击即可跳转至相应文档的详细说明)

01. CleanData ——【测序数据预处理结果】	3
novototal.QCstat.info.xls ——【QC结果基本信息表】	3
total.QCstat.info.xls —— 【 QC结果详细信息表】	4
total.*.NonHostQCstat.info.xls ——【去除宿主后的结果的详细信息表】	5
` Sample —— 【各样品对应的质控结果,文件夹以样品名称来命名】	6
*.fq1.gz ——【 质控后 read1 的 FASTQ 文件 】	6
*.fq2.gz ——【 质控后 read2 的 FASTQ 文件 】	6
*.nohost.fq1.gz ——【当存在宿主基因组序列时,去除宿主后 read1 的 FASTQ 文件】	7
*.nohost.fq2.gz ——【当存在宿主基因组序列时,去除宿主后 read2 的 FASTQ 文件】	7
*.qual.png ——【 Clean Data的碱基质量分布图】	7
` *.base.png ——【 Clean Data 的碱基类型分布图】	7
` 01 CleanDataReadMe ndf	8

|-- 01. CleanData ——【测序数据预处理结果】

|-- novototal.QCstat.info.xls ——【QC结果基本信息表】

该文件即对应的是结题报告中的数据预处理统计表,可以用 excel 打开该文件,在该文件中,各列所代表的含义如下:

列数	列标题	
1	Sample	样品名称
2	InsertSize(bp)	建库时的插入片段大小,单位为 bp
3	SeqStrategy	测序的策略,若为 150:150 即代表采用的是 Pair-end 测序,测序 reads 长度为 150bp
4	RawData	RawData 的数据量, 单位为 M,
5	CleanData	CleanData 数据量,单位为 M
6	Clean_Q20	CleanData 的 Q20
7	Clean_Q30	CleanData 的 Q30
8	Clean_GC(%)	CleanData 碱基的 GC 含量
9	Effective(%)	CleanData 占 RawData 的百分比

|-- total.QCstat.info.xls —— 【 QC 结果详细信息表】

可以用 excel 打开该文件,其相比于 novototal.QCstat.info.xls 而言,多出了 RawReads、Low_Q、N_num、Adapter、Duplication、Poly 这几列,各列所代表的含义如下:

列标题	说明
Sample	样品名称
InsertSize(bp)	建库时的插入片段大小,单位为 bp
SeqStrategy	测序的策略, 若为 150:150 即代表采用的是 Pair-end 测序, 测序 reads 长度为 150bp
RawData	RawData 的数据量,单位为 M,
RawReads(#)	原始下机的 Reads 数目
Low_Q	去除含低质量碱基超过一定比例的 reads 序列总长,单位为 M
N_num	去除含 N 碱基达到一定比例的 reads 序列总长,单位为 M
Adapter	去除与 Adapter 之间 overlap 超过一定阈值的 reads 序列总长,单位为 M
Duplication	去除的重复 reads 序列总长,单位为 M
Poly	代表去除的 Poly reads 的 reads 序列总长,单位为 M
CleanData	CleanData 数据量,单位为 M
Clean_Q20	CleanData 的 Q20
Clean_Q30	CleanData 的 Q30
	Sample InsertSize(bp) SeqStrategy RawData RawReads(#) Low_Q N_num Adapter Duplication Poly CleanData Clean_Q20

14	Clean_GC(%)	CleanData 碱基的 GC 含量
15	Effective(%)	CleanData 占 RawData 的百分比

|-- total.*.NonHostQCstat.info.xls ——【去除宿主后的结果的详细信息表】

可以用 excel 打开该文件,其相比于 total.QCstat.info.xls 而言,多出了 NonHostData 这一列,各列所代表的含义如下:

列数	列标题	说明
1	Sample	样品名称
2	InsertSize(bp)	建库时的插入片段大小,单位为 bp
3	SeqStrategy	测序的策略,若为 150:150 即代表采用的是 Pair-end 测序,测序 reads 长度为 150bp
4	RawData	RawData 的数据量,单位为 M,
5	RawReads(#)	原始下机的 Reads 数目
6	Low_Q	去除含低质量碱基超过一定比例的 reads 序列总长,单位为 M
7	N_num	去除含 N 碱基达到一定比例的 reads 序列总长,单位为 M

8	Adapter	去除与 Adapter 之间 overlap 超过一定阈值的 reads 序列总长,单位为 M
9	Duplication	去除的重复 reads 序列总长,单位为 M
10	Poly	代表去除的 Poly reads 的 reads 序列总长,单位为 M
11	CleanData	CleanData 数据量,单位为 M
12	Clean_Q20	CleanData 的 Q20
13	Clean_Q30	CleanData 的 Q30
14	Clean_GC(%)	CleanData 碱基的 GC 含量
15	Effective(%)	CleanData 占 RawData 的百分比
16	NonHostData	去除宿主后,剩余数据量 <mark>,</mark> 单位为 M

`-- Sample —— 【各样品对应的质控结果,文件夹以样品名称来命名】

关于 FASTQ 文件格式介绍,请参考结题报告中的常见数据格式说明文档。

关于 FASTQ 文件格式介绍,请参考结题报告中的常见数据格式说明文档。

|-- *.nohost.fq1.gz ——【当存在宿主基因组序列时,去除宿主后 read1 的 FASTQ 文件】

关于 FASTQ 文件格式介绍,请参考结题报告中的常见数据格式说明文档。

|-- *.nohost.fq2.gz ——【当存在宿主基因组序列时,去除宿主后 read2 的 FASTQ 文件】

关于 FASTQ 文件格式介绍,请参考结题报告中的常见数据格式说明文档。

|-- *.qual.png ——【 Clean Data 的碱基质量分布图】

在该图中,图中横轴为 reads 上的碱基位置,从中间隔开,前后代表的是 PE reads, 图中纵轴为该 reads 位置上碱基质量的分布

坐标轴	标题	说明
X轴	Position along reads	reads 上的碱基位置,从中间隔开,前后代表的是 PE reads
Y轴	Quality Value	该 reads 位置上碱基质量的分布

`-- *.base.png ——【 Clean Data 的碱基类型分布图】

在该图中,右上角的图例中有各个碱基所代表的颜色说明,图中横轴为 reads 上的碱基位置,中间黄线前后则代表的是 PE reads,图中纵轴代表该 reads 位置上某个碱基的比例。

坐标轴	标题	说明
X轴	Position along reads	reads 上的碱基位置,中间黄线前后则代表的是 PE reads
Y轴	Percent Value	该 reads 位置上某个碱基的比例,右上角的图例中有各个碱基所代表的颜色说明

`-- 01.CleanData--ReadMe.pdf ——【 01.CleanData/交付结果目录说明】

