基因组组装评估中的测序reads深度-GC含量（Depth-GC）滑窗分布图

在得到组装好的基因组序列之后，首先要做的就是使用多种方法来对组装结果的质量进行评估。将二代测序reads数据比对至组装结果中，通过统计基因组各位点的reads覆盖深度，并结合组装序列的碱基含量组成等一并分析，是一种常见的评估方法。除了对组装序列，这种方法还同时能对测序reads数据的质量也进行检查。

对于组装后的基因组序列，将组装基因组结果中的contigs/scaffolds序列分隔为一定长度的滑窗（如2000bp为一滑窗区间），统计每段滑窗内的各碱基占比，或者GC含量等，以查看组装结果中碱基分布是否均匀，基因组中存在多少比例的高GC区域等。据此，可在一定程度上推测物种基因组结构特征，组装中是否存在明显的错配，或判断测序数据中是否存在其他物种污染等。

对于测序reads序列，使用特定的序列比对软件，常见的如BWA，Bowtie2等，将二代测序reads序列比对至组装得到的基因组草图中。同样地，以滑窗的形式，统计基因组中各位点的测序reads覆盖深度（Depth），即可根据基因组中每个滑窗区域的测序深度评估测序质量，如是否存在数据随机性不好导致深度不均一的问题，以及判断存在PCR偏好性扩增的基因组结构区等。

将GC含量、测序深度统计二者相结合，在基因组中以滑窗的形式统计并绘制测序reads深度-GC含量（Depth-GC）分布图，还可检查在测序扩增时是否存在了明显的GC偏向等。

本文使用一细菌的基因组组装草图，及其组装前的Illumina测序reads数据（双端测序，PE150），展示这种基于比对的评估方法。测试数据、自定义程序脚本以及部分结果示例等，已上传自百度盘。

fastq格式的Illumina测序reads数据、基因组组装草图fasta文件（无提取码）：<https://pan.baidu.com/s/1yYQNxty5hBdcpoA8xkrQUg>

自定义程序脚本及重要的结果文件示例（提取码hc3l）：<https://pan.baidu.com/s/1WYccQluDZYR079HjkVvNPA>

## 测序reads比对及覆盖深度统计、碱基含量统计

本文示例使用网盘附件中的Illumina测序reads数据“Bacillus\_subtilis.filt\_R1.fastq.gz/Bacillus\_subtilis.filt\_R2.fastq.gz”，以及基因组组装草图fasta文件“Bacillus\_subtilis.scaffolds.fasta”，进行评估过程展示。

**测序reads比对及基因组单位点覆盖深度统计**

在本示例中，使用BWA（Burrows-Wheeler Aligner）软件，将测序reads比对至组装后的基因组草图序列中，得到比对结果sam文件。之后使用Samtools软件，将比对结果sam文件压缩为二进制的bam文件后（一是节省存储空间，二是提升数据处理效率），对bam文件中的比对结果进行排序操作，并统计每个位点的测序深度。

#建立 bwa/samtools 索引

bwa index -a bwtsw Bacillus\_subtilis.scaffolds.fasta

samtools faidx Bacillus\_subtilis.scaffolds.fasta

#bwa 比对 > bam

bwa mem -t 4 -M Bacillus\_subtilis.scaffolds.fasta Bacillus\_subtilis.filt\_R1.fastq.gz Bacillus\_subtilis.filt\_R2.fastq.gz | samtools view -@ 4 -bS > Bacillus\_subtilis.bam

#bam 排序

samtools sort -@ 4 Bacillus\_subtilis.bam > Bacillus\_subtilis.sort.bam

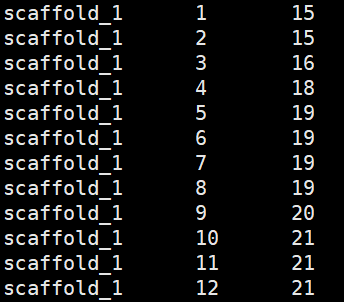
rm Bacillus\_subtilis.bam

#单碱基位点测序深度统计

samtools depth Bacillus\_subtilis.sort.bam > Bacillus\_subtilis.depth

结果文件中，“Bacillus\_subtilis.sort.bam”为排序后的比对结果文件。该文件中记录了每个测序reads的ID、其在基因组中的对齐位置、对齐程度、测序质量等众多详细信息。所谓“排序”，即根据测序reads与基因组中的碱基对齐位置，按这些碱基在基因组每条contigs/scaffolds序列中的先后顺序对比对结果进行排序。详情可参见bam文件格式说明（<http://boyun.sh.cn/bio/wp-content/uploads/2012/07/SAM1.pdf>）。

“Bacillus\_subtilis.depth”为使用Samtools统计得到的基因组中每个碱基位点的测序覆盖深度。其内容如下。第一列，组装基因组中每条contigs/scaffolds序列ID；第二列，该序列中每个碱基位点，数值代表了该碱基在该序列中的位置，仅展示有reads覆盖度的位点；第三列，每个碱基的reads覆盖深度，或者理解为该位点的碱基被测序测到的频数。通过碱基位点测序深度统计文件中的内容，我们即可知哪些位点出现了PCR偏好性扩增等信息。



补充信息：将测序reads回比至组装结果中，还可借助变异检测软件（如GATK等），检测基因组中的“SNP位点”，即 “自身call SNP”，这一步仿照重测序分析中的“call SNP”思路，可用于反过来评估测序reads质量。由于这里不是重测序分析，基因组序列本来就是由测序reads组装得到的，那么当测序reads回比至自己的组装结果中时，若检测出SNP位点，则考虑两种情况。对于单倍体物种来讲，理论上就不应该检测出SNP位点，检测出的变异位点肯定就是由测序错误导致的碱基不匹配等情况。而对于二倍体或多倍体物种来讲，除了碱基的测序错误外，还可能是两条或多条同源染色体间杂合位点所致。若对于非单倍体物种，检测出的变异类型为“纯合SNP”，则可以推测就是由测序错误所致；若为“杂合SNP”，还需考虑可能是由两条或多条同源染色体间碱基组成为杂合的情况所致了。

**基因组碱基含量及测序reads覆盖度的滑窗统计**

原始的单碱基位点测序深度统计文件以单个的碱基作为展示，因此查看起来可能并不方便，特别是在较大的基因组中。这时候，通常考虑以滑窗的形式，展示统计结果。所谓“滑窗”，即滑动窗格，简单地理解就是将基因组序列划分为一定长度大小的区间，这每一段序列区间就是“滑窗”。然后统计这一段区间序列内测序深度的平均值，即可作为测序reads在这段滑窗中的平均覆盖深度。

同样地，对于每段滑窗序列，同时再统计这段序列中的各碱基组成含量，常见统计值即为GC含量的百分比。

本示例使用自己编写的python3脚本（位于网盘附件，“depth\_base\_stat.py”），读入上述得到的单碱基位点测序深度统计文件“Bacillus\_subtilis.depth”，以及基因组组装草图fasta文件“Bacillus\_subtilis.scaffolds.fasta”，按指定滑窗大小进行统计之后，输出结果。

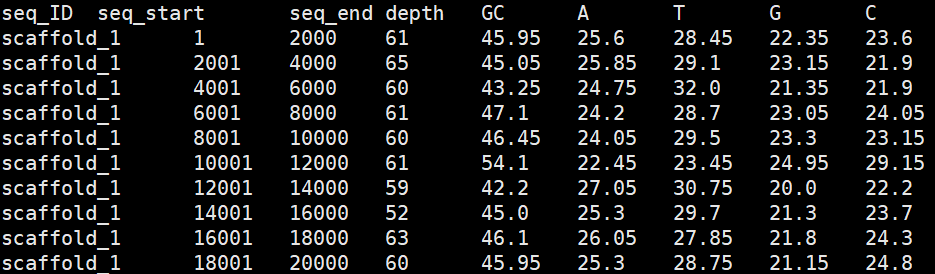
#测序深度、基因组碱基含量滑窗统计（这里设置 2000bp 为一滑窗长度）

#python3 depth\_base\_stat.py -h

python3 depth\_base\_stat.py -g Bacillus\_subtilis.scaffolds.fasta -d Bacillus\_subtilis.depth -s depth\_base.stat.txt -l 2000

输出文件“depth\_base.stat.txt”内容如下。第一列（seq\_ID），组装基因组中每条contigs/scaffolds序列ID；第二列（seq\_start）和第三列（seq\_end），各滑窗在contigs/scaffolds序列中的起始/终止位置；第四列（depth），每段滑窗序列的平均测序深度；第五列及之后（GC、A、T、G、C），每段滑窗序列内的GC含量百分比、A/T/G/C四种碱基含量百分比。

与上述得到的单碱基位点测序深度统计文件“Bacillus\_subtilis.depth”相比，该文件虽然内容缩减，单可读性增强，更容易我们发现基因组中的不均一扩增测序区域，以及高GC区域等。



补充信息：上述内容首先使用samtools depth对单个碱基位点进行测序深度统计，再通过自己写的脚本将其转化为滑窗统计结果的样式。此外还可使用Bedtools软件，通过指定一个包含基因组序列ID以及区间长度的bed文件，能够直接基于排序后的bam结果文件以滑窗的形式统计处每段滑窗序列的测序深度信息。这里不再阐述，有兴趣可了解下。

## 测序reads深度-GC含量（Depth-GC）分布图

接下来，我们可以根据上述统计结果文件“depth\_base.stat.txt”中的内容，将每段滑窗序列的测序reads深度及碱基含量的关系较直观地展示出来，便于评估测序深度、碱基组成分布，检查在测序扩增时是否存在了明显的GC偏向，以及测序reads中是否存在污染情况等。

**基因组Depth-GC分布图简介**

这里首先使用自己编写的R脚本（位于网盘附件，“depth\_GC\_plot.r”），读入“depth\_base.stat.txt”统计结果文件，绘制Depth-GC散点密度分布图。使用该示例图简介其在基因组组装评估中的用途。

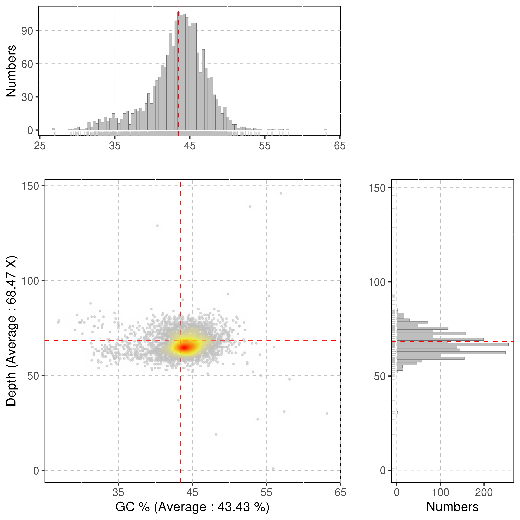
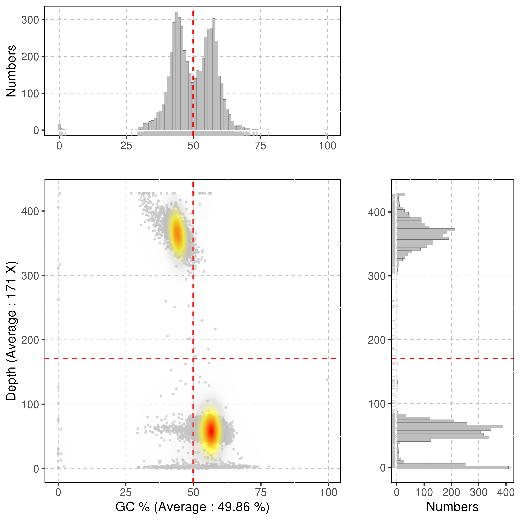
#depth\_GC 散点图

#Rscript depth\_GC\_plot.r -h

Rscript depth\_GC\_plot.r -i depth\_base.stat.txt -o depth\_GC

输出做图结果“depth\_GC.png”，本示例默认样式见下图左图。对于散点图，横坐标为GC含量，纵坐标为reads覆盖深度；对于两侧的直方图，分别为GC含量、测序深度的滑窗频数分布；红色虚线为平均值，高密度点分布区域使用了更易于区分的颜色。对于下图右图，来自于其它处的数据，展示在这里方便加深理解，是一个存在污染的测序样本的组装结果的Depth-GC评估图。

该图可用于评估测序深度均匀度，测序物种的GC分布特征；例如，查看基因组中高GC的区域是否存在测序深度偏低的问题等。同时亦可检验组装结果中是否存在非目标物种的污染，由于每个物种的GC含量不同，同一个物种的GC含量会呈现集中分布。如果多数的点集中分布在一个比较窄的范围内，则表明样本不存在污染；如果分布在多个区域，则表明样本中可能存在其他物种的污染。若存在污染时，可根据点分布的集中区判断污染程度，例如根据GC分布大致判断污染物种的数量，或者根据污染序列的测序reads覆盖深度来大致推测污染的比例等。可对以下两张评估图进行比较查看，方便加深理解。

**R包ggplot2绘制基因组Depth-GC分布图方法展示**

以下所阐述的R作图（ggplot2作图）命令，是对该图绘制方法的简单描述。大家有兴趣可参考以下命令并作更改，自定义样式做图展示。

我们加载ggplot2包，读入上文中的基因组测序深度-碱基含量滑窗统计结果文件“depth\_base.stat.txt”，绘制散点图展示测序深度与GC含量的关系。

#导入 ggplot2 包

library(ggplot2)

#读取文件

depth\_base <- read.delim('depth\_base.stat.txt')

#ggplot2 散点图

depth\_GC <- ggplot(depth\_base, aes(GC, depth)) +

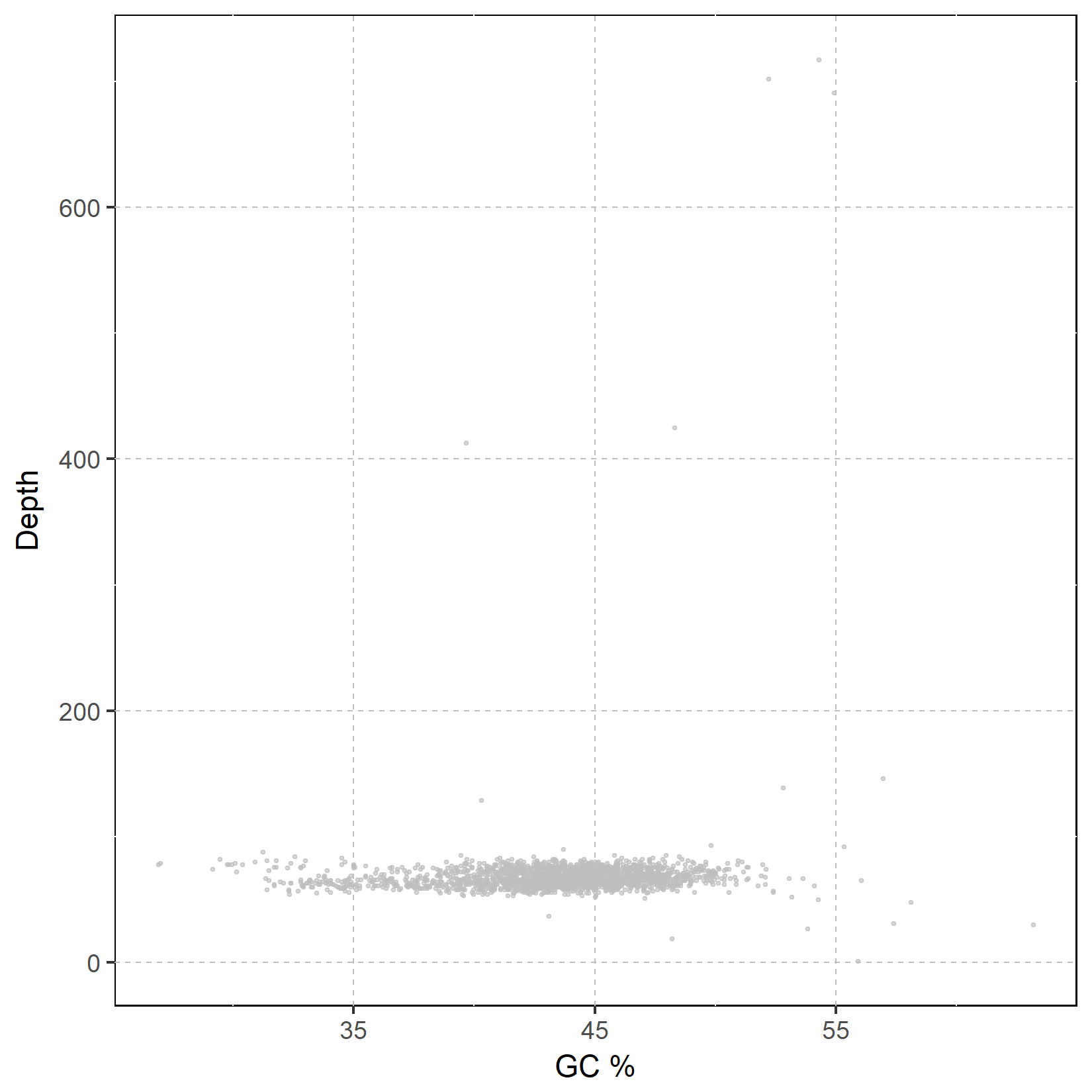
geom\_point(color = 'gray', alpha = 0.6, pch = 19, size = 0.5) +

theme(panel.grid.major = element\_line(color = 'gray', linetype = 2, size = 0.25), panel.background = element\_rect(color = 'black', fill = 'transparent')) +

labs(x = 'GC %', y = 'Depth')

ggsave('depth\_GC\_plot.png', depth\_GC, width = 5.5, height = 5.5)

直接以输入文件中的GC列为横坐标，depth列为纵坐标，绘制散点图（geom\_point()）。见要设置点的颜色为灰色并半透明，设置背景框样式，添加坐标轴标签。结果如下。



由于PCR的偏好性扩增等影响，导致基因组中个别序列出现了非特异性扩增，测序深度远大于平均值。由于这些序列所占比例极少，我们在作图时可考虑将其过滤掉。

#统计平均测序深度（百分比）

depth\_mean <- round(mean(depth\_base$depth), 2)

depth\_base <- subset(depth\_base, depth <= 3 \* depth\_mean)

统计平均测序深度，并屏蔽掉3倍于平均测序深度的滑窗区间。使用剩余的滑窗序列的统计信息进行作图展示。

#统计平均测序深度（百分比）

depth\_mean <- round(mean(depth\_base$depth), 2)

depth\_base <- subset(depth\_base, depth <= 3 \* depth\_mean)

#统计平均 GC 含量（百分比）

GC\_mean <- round(mean(depth\_base$GC), 2)

#ggplot2 散点图

depth\_GC <- ggplot(depth\_base, aes(GC, depth)) +

geom\_point(color = 'gray', alpha = 0.6, pch = 19, size = 0.5) +

theme(panel.grid.major = element\_line(color = 'gray', linetype = 2, size = 0.25), panel.background = element\_rect(color = 'black', fill = 'transparent')) +

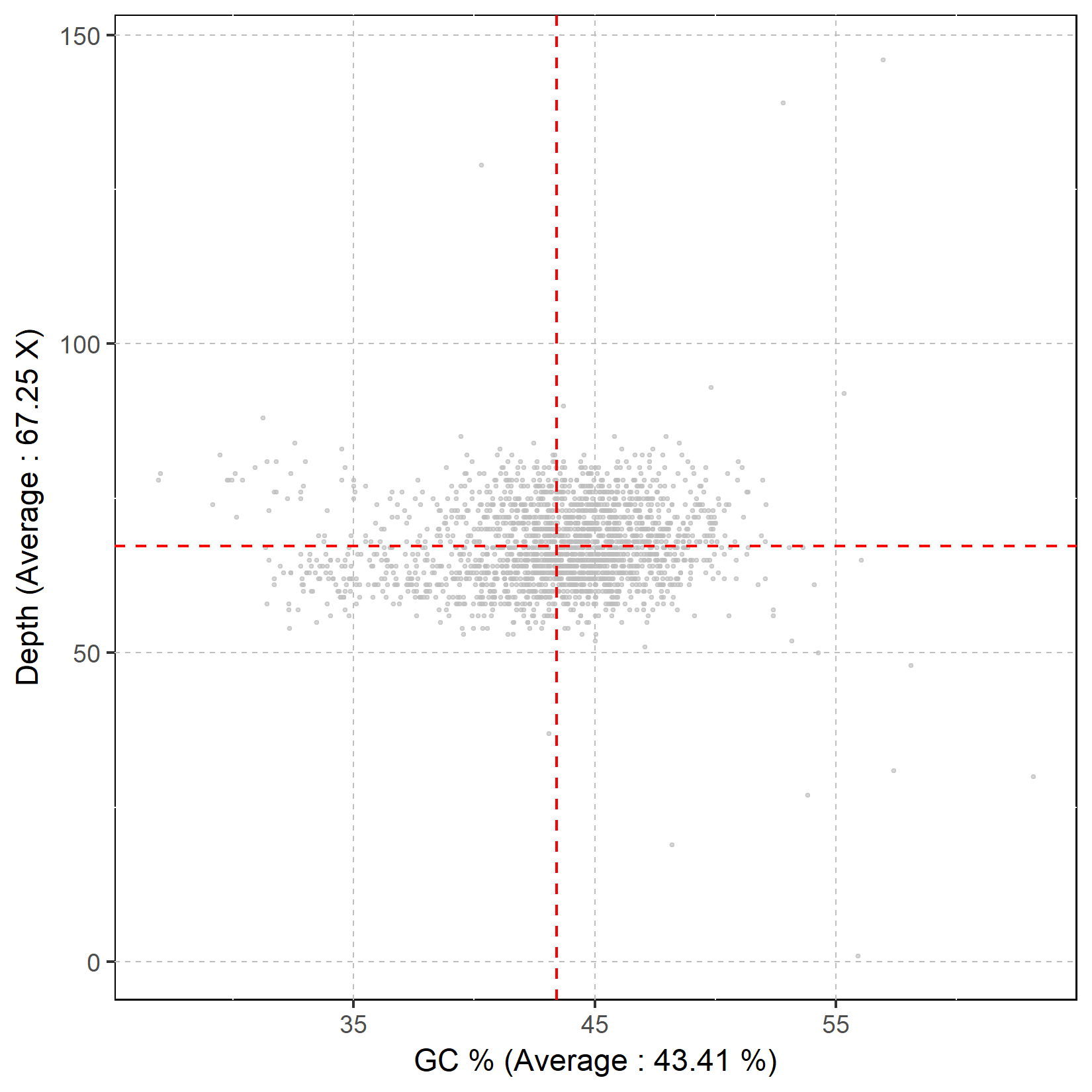
geom\_vline(xintercept = GC\_mean, color = 'red', lty = 2, lwd = 0.5) +

geom\_hline(yintercept = depth\_mean, color = 'red', lty = 2, lwd = 0.5) +

labs(x= paste('GC % (Average :', GC\_mean, '%)'), y = paste('Depth (Average :', depth\_mean, 'X)'))

ggsave('depth\_GC\_plot.png', depth\_GC, width = 5.5, height = 5.5)

一并统计GC含量的平均值，并在使用屏蔽3倍于平均测序深度的滑窗区间后的数据作图时，将平均测序深度、GC含量信息添加在作图结果中（geom\_vline()、geom\_hline()、labs()）。输出图片如下。



对于正常的组装基因组序列来讲，无论GC含量还是测序深度，都会在某一区域大量集中。对于上图来讲，虽然能够看到散点主要分布的区域，但效果不很明显。因此我们可考虑在图中点的集中区域，使用易于区分的颜色展示点的密度。

#添加颜色密度

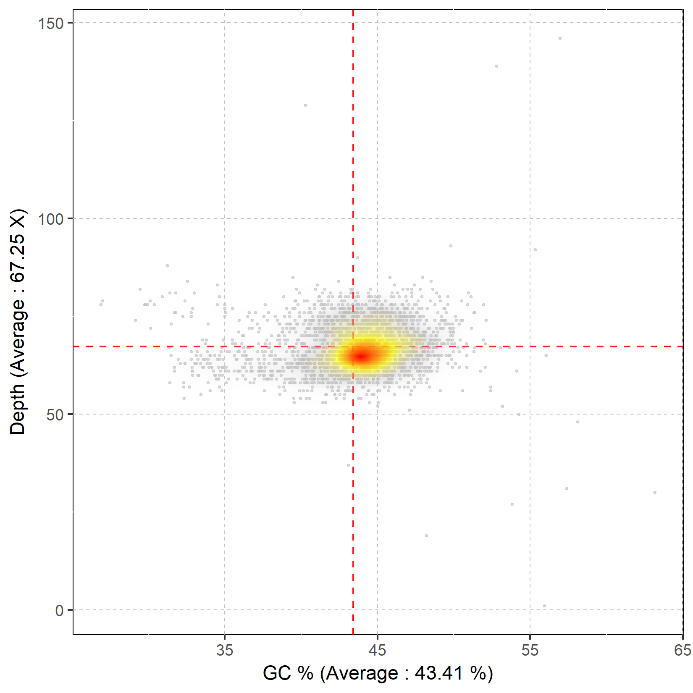
depth\_GC <- depth\_GC +

stat\_density2d(aes(fill = ..density.., alpha = ..density..), geom = 'tile', contour = FALSE, n = 500, show.legend = FALSE) +

scale\_fill\_gradientn(colors = c('transparent', 'gray', 'yellow', 'red'))

ggsave('depth\_GC\_plot.png', depth\_GC, width = 5.5, height = 5.5)

在前述作图结果“depth\_GC”样式的基础上，使用stat\_density2d()命令，根据散点在图中的密度分布继续添加颜色密度标记，并使用scale\_fill\_gradientn()指定颜色梯度，根据散点分布密度由低到高在图中标记为不同颜色。做图结果如下。



到这里，基因组Depth-GC密度分布图基本上就算完成了。通过该图可以获得我们想要得知的信息，判断组装基因组的质量，以及测序质量等。

此外，还可考虑统计滑窗GC含量，或者滑窗测序reads覆盖度，绘制二者的分布直方图，添加在散点图的两侧，以更直观展示数据分布情况。

#depth 深度直方密度图

depth\_hist <- ggplot(depth\_base, aes(depth)) +

geom\_histogram(binwidth = (max(depth\_base$depth) - min(depth\_base$depth))/100, fill = 'gray', color = 'gray40', size = 0.1) +

geom\_rug(color = 'gray', alpha = 0.6) +

theme(panel.grid.major = element\_line(color = 'gray', linetype = 2, size = 0.25), panel.background = element\_rect(color = 'black', fill = 'transparent')) +

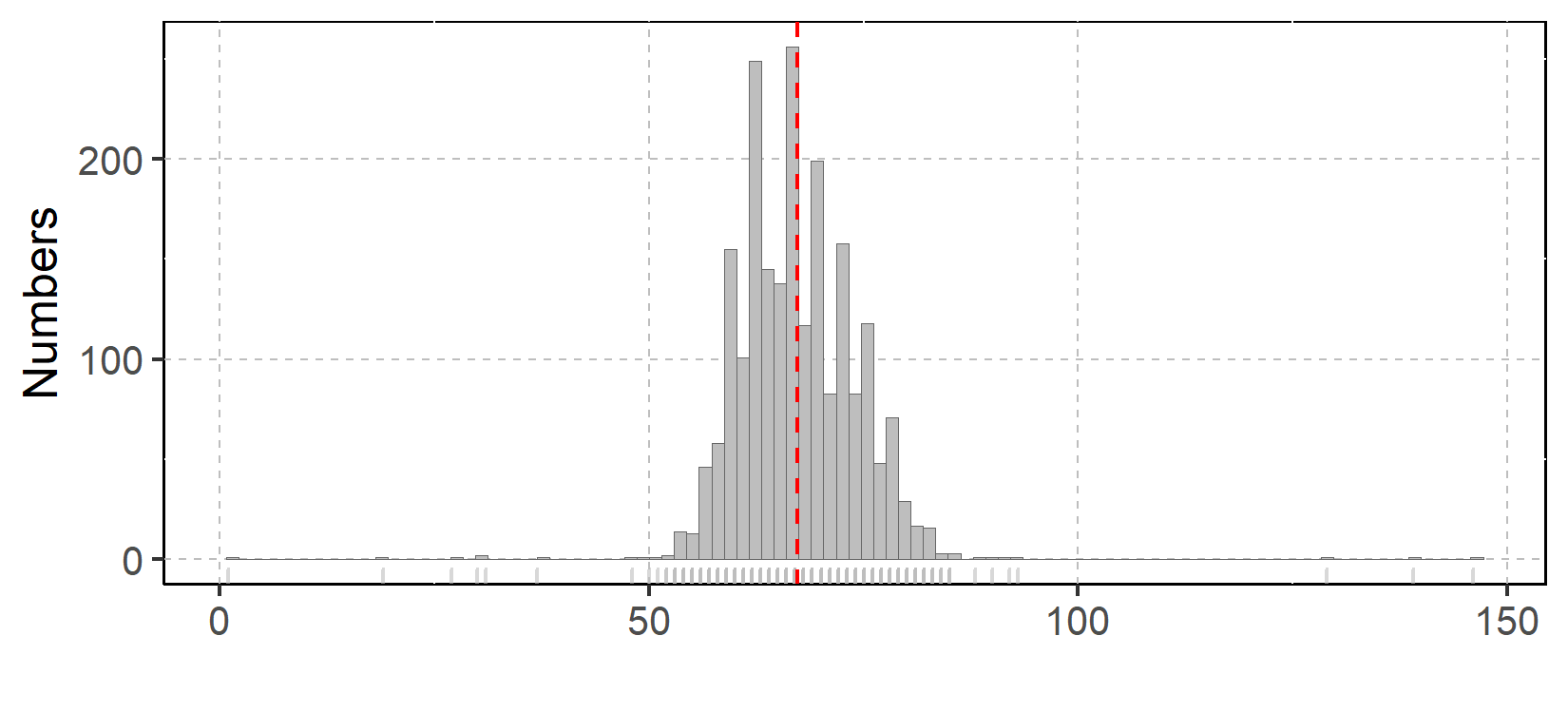
theme(axis.line = element\_line(color = 'black', size = 0.3), axis.text = element\_text(size = 10), axis.title = element\_text(size = 12)) +

labs(x = '', y = 'Numbers') +

geom\_vline(xintercept = depth\_mean, color = 'red', lty = 2, lwd = 0.5)

ggsave('depth\_hist\_plot.png', depth\_hist, width = 5.5, height = 2.5)

geom\_histogram()用于绘制频数直方图，并使用binwidth参数划分为100个频数统计窗格（即无论测序深度怎样分布，均根据深度频数划分为100个“柱子”展示）。额外使用geom\_rug()在频数直方图的下方添加地毯图，作为展示每个基因组滑窗序列测序深度的位置。输出结果如下。



同样地，对于GC分布直方图。

#GC 含量直方密度图

GC\_hist <- ggplot(depth\_base, aes(GC)) +

geom\_histogram(binwidth = (max(depth\_base$GC) - min(depth\_base$GC))/100, fill = 'gray', color = 'gray40', size = 0.1) +

geom\_rug(color = 'gray', alpha = 0.6) +

theme(panel.grid.major = element\_line(color = 'gray', linetype = 2, size = 0.25), panel.background = element\_rect(color = 'black', fill = 'transparent')) +

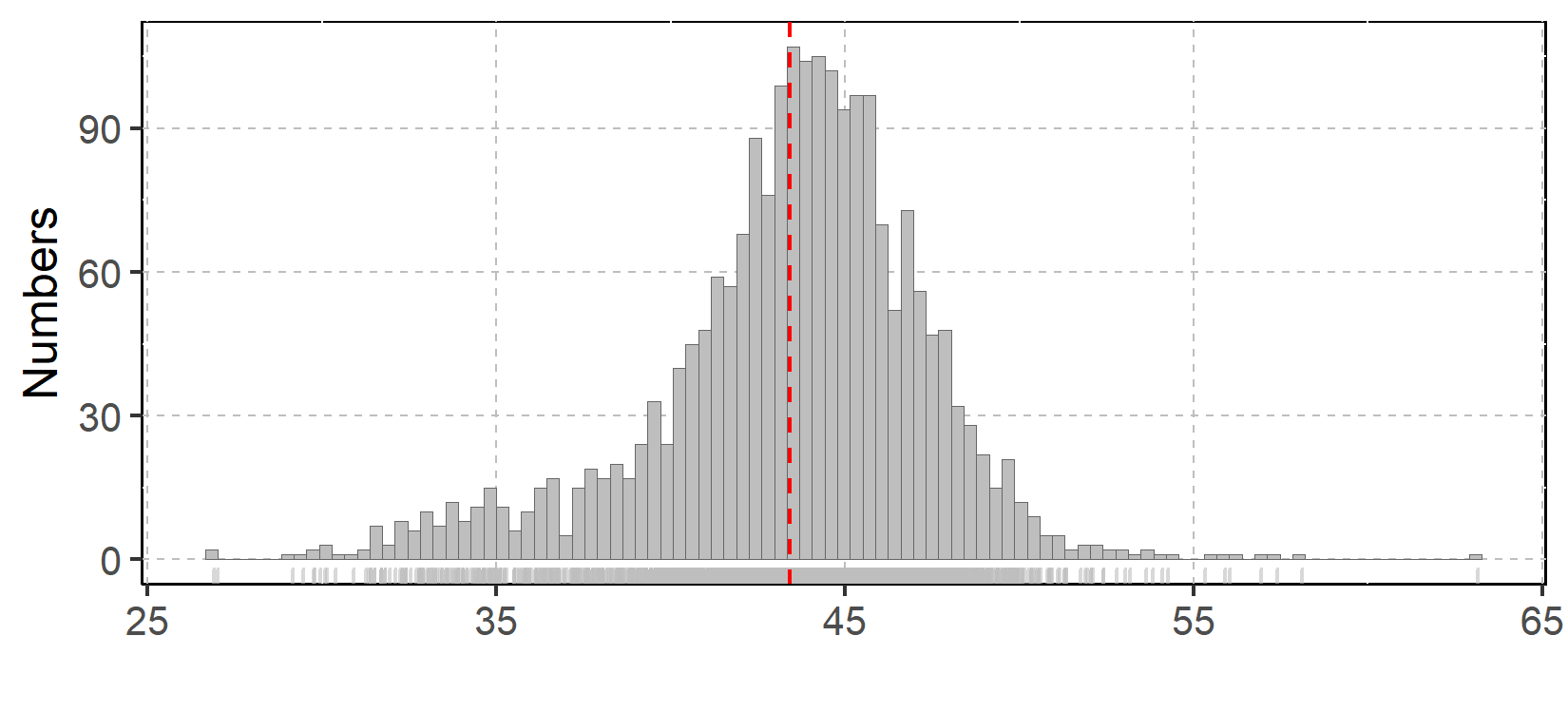
theme(axis.line = element\_line(color = 'black', size = 0.3), axis.text = element\_text(size = 10), axis.title = element\_text(size = 12)) +

labs(x = '', y = 'Numbers') +

geom\_vline(xintercept = GC\_mean, color = 'red', lty = 2, lwd = 0.5)

ggsave('GC\_hist\_plot.png', GC\_hist, width = 5.5, height = 2.5)

参数描述同上述，输出结果如下。



最后可使用grid包，将上述已完成的散点图与两个频数直方图组合在一起，得到最终的结果。

#导入 grid 包

library(grid)

#此处首先将上述 depth 深度直方密度图旋转 90 度

depth\_hist <- depth\_hist + coord\_flip()

#grid 组合图形

png('depth\_GC\_all\_plot.png', width = 4000, height = 4000, res = 600, units = 'px')

grid.newpage()

pushViewport(viewport(layout = grid.layout(3, 3)))

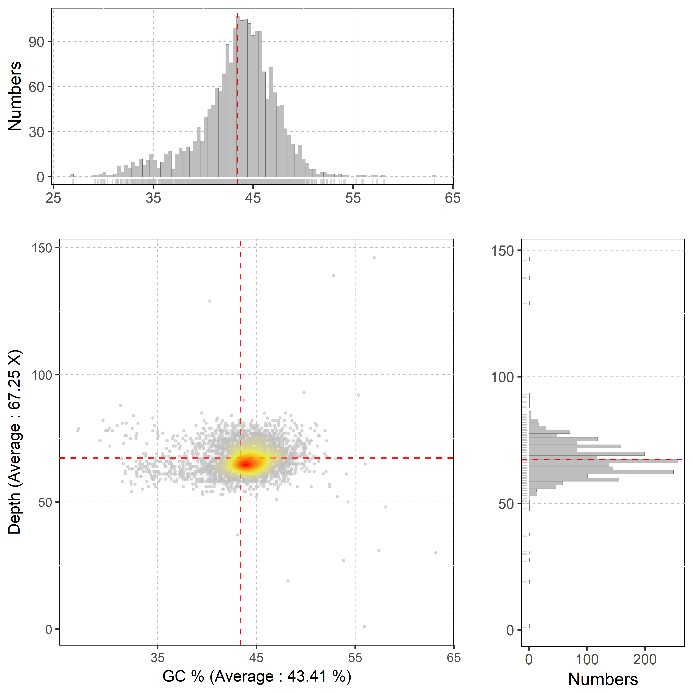
print(depth\_GC, vp = viewport(layout.pos.row = 2:3, layout.pos.col = 1:2))

print(GC\_hist, vp = viewport(layout.pos.row = 1, layout.pos.col = 1:2))

print(depth\_hist, vp = viewport(layout.pos.row = 2:3, layout.pos.col = 3))

dev.off()

组合后输出的图片结果如下。



大家有兴趣还可将两侧的直方图替换为箱线图，以展示数据分布等，这里就不再过多阐述了。