目录

第一部分

高通量测序平台介绍

第一章　Illumina®

Illumina 新一代测序技术可以高通量、并行对核酸片段进行深度测序，测序的技术原理是采用可逆性末端边合成边测序反应，首先在 DNA 片段两端加上序列已知的通用接头构建文库，文库加载到测序芯片 Flowcell 上，文库两端的已知序列与 Flowcell 基底上的 Oligo 序列互补，每条文库片段都经过桥式 PCR 扩增形成一个簇，测序时采用边合成边测序反应，即在碱基延伸过程中，每个循环反应只能延伸一个正确互补的碱基，根据四种不同的荧光信号确认碱基种类，保证最终的核酸序列质量，经过多个循环后，完整读取核酸序列。

1.1　测序原理

Illumina/Solexa Genome Analyzer 测序的基本原理是边合成边测序（SBS，Sequencing By Synthesis）。在 Sanger 等测序方法的基础上，通过技术创新，用不同颜色的荧光标记四种不同的 dNTP，当 DNA 聚合酶合成互补链时，每添加一种 dNTP 就会释放出不同的荧光，根据捕捉的荧光信号并经过特定的计算机软件处理，从而获得待测 DNA 的序列信息。

1.1.1　测序文库的构建（Library Construction）

首先准备基因组 DNA，然后将 DNA 随机片段化成几百碱基或更短的小片段，并在两头加上特定的接头（Adaptor）。如果是转录组测序，则文库的构建要相对麻烦些，RNA 片段化之后需反转成 cDNA，然后加上接头，或者先将 RNA 反转成 cDNA，然后再片段化并加上接头。片段的大小（Insert size）对于后面的数据分析有影响，可根据需要来选择。对于基因组测序来说，通常会选择几种不同的 insert size，以便在组装（Assembly）的时候获得更多的信息。

1.1.2 锚定桥接（Surface Attachment and Bridge Amplification）

Solexa 测序的反应在叫做 flow cell 的玻璃管中进行，flow cell 又被细分成 8 个 Lane，每个 Lane 的内表面有无数的被固定的单链接头。上述步骤得到的带接头的 DNA 片段变性成单链后与测序通道上的接头引物结合形成桥状结构，以供后续的预扩增使用。

1.1.3 预扩增（Denaturation and Complete Amplification）

添加未标记的 dNTP 和普通 Taq 酶进行固相桥式 PCR 扩增，单链桥型待测片段被扩增成为双链桥型片段。通过变性，释放出互补的单链，锚定到附近的固相表面。通过不断循环，将会在 Flow cell 的固相表面上获得上百万条成簇分布的双链待测片段。

1.1.4 单碱基延伸测序（Single Base Extension and Sequencing）

在测序的 flow cell 中加入四种荧光标记的 dNTP 、DNA 聚合酶以及接头引物进行扩增，在每一个测序簇延伸互补链时，每加入一个被荧光标记的 dNTP 就能释放出相对应的荧光，测序仪通过捕获荧光信号，并通过计算机软件将光信号转化为测序峰，从而获得待测片段的序列信息。从荧光信号获取待测片段的序列信息的过程叫做 Base Calling，Illumina 公司 Base Calling 所用的软件是 Illumina’s Genome Analyzer Sequencing Control Software and Pipeline Analysis Software。读长会受到多个引起信号衰减的因素所影响，如荧光标记的不完全切割。随着读长的增加，错误率也会随之上升。

1.1.5 数据分析（Data Analyzing）

这一步严格来讲不能算作测序操作流程的一部分，但是只有通过这一步前面的工作才显得有意义。测序得到的原始数据是长度只有几十个碱基的序列，要通过生物信息学工具将这些短的序列组装成长的 Contigs 甚至是整个基因组的框架，或者把这些序列比对到已有的基因组或者相近物种基因组序列上，并进一步分析得到有生物学意义的结果

1.2 NovaSeq™

https://www.illumina.com/systems/sequencing-platforms/novaseq/specifications.html

	NovaSeq 5000 and 6000 Systems		NovaSeq 6000 System	
Flow Cell 类型	S1	S2	S3	S4
2×50 bp	$\approx$ 167 Gb	280 $\sim$ 333 Gb	NA[1]	NA

	NovaSeq 5000 and 6000 Systems		NovaSeq 6000 System	
2×100 bp	≈ 333 Gb	$560 \sim 667$ Gb	NA	NA
2×150 bp	≈ 500 Gb	$850 \sim 1000$ Gb	≈ 2000 Gb	≈ 3000 Gb

表 1.1: NovaSeq™ 通量

	NovaSeq 5000 and 6000 Systems		NovaSeq 6000 System	
Flow Cell 类型	S1	S2	S3	S4
	≈ 1.6 B	$2.8 \sim 3.3$ B	≈ 6.6 B	≈ 10 B

表 1.2: NovaSeq™ 通过筛选标准的 Reads 数目

	NovaSeq 6000 System		
Flow Cell 类型	S2		
读长	2×50 bp	2×100 bp	2×150 bp
Q30	$\geq 85\%$ bp	$\geq 80\%$ bp	$\geq 75\%$ bp
运行时间	19h	29h	40h

表 1.3: NovaSeq™ 测序质量和运行时间

	NovaSeq 5000 and 6000 Systems		NovaSeq 6000 System	
Flow Cell 类型	S1	S2	S3	S4
人类全基因组[2]	≈ 8	≈ 16	≈ 32	≈ 48
人类全外显子组[3]	≈ 66	≈ 132	NA	NA
转录组[4]	≈ 66	≈ 132	NA	NA

[1]NA：Not Applicable

[2]全基因组测序每样本测序数据量 ≥ 120 Gb，测序深度 30×。

[3]全外显子组测序每样本采用 2×75bp 测序，数据量 ≥ 50 M。

[4]转录组测序每样本采用 2×50bp 测序，数据量 ≥ 50 M。

	NovaSeq 5000 and 6000 Systems	NovaSeq 6000 System

表 1.4: NovaSeq™ 不同用途的最大支撑样本量

第二章　Thermo Fisher Scientific

2.1　Ion PGM™

2.2　Ion Proton™

主要参数如下：

指标	参数（Ion PI™Chip ）	参数（Ion PII™Chip ）
通量	10G	20× （人类）
读长	≈ 200bp	
Reads 数目	60 ∼ 80M	
运行时间	2 ∼ 4h	

第三章　Pacific Biosciences

Pacific Biosciences（PacBio®）公司的测序平台主要是基于单分子实时测序技术（SMRT，Single Molecular Real Time Sequencing）。现在已经商业化的测序平台包括 PacBio Sequel 和 PacBio RS II。

3.1　测序原理

SMRT 技术的基本原理是：DNA 模板被聚合酶捕获后，四种不用荧光标记的 dNTP 随机进入检测区域与聚合酶/模板复合体结合。与模板匹配的碱基生成的化学键的时间远远长于其他碱基停留的时间。因此统计不同荧光信号存在的时间长短，即可区分与 DNA 模板结合的碱基。通过统计四种荧光信号与时间的关系图，即可测定 DNA 模板序列。

3.1.1　ZMWs：Zero-Mode Waveguides

ZMWs 可是光仅仅照亮固定着一个 DNA 聚合酶/DNA 模板复合体的小孔的底部。ZWM 是一个直径只有 10∼ 15nm 的孔，远小于检测激光的波长（数百纳米），因此，当激光打在 SMW 底部时，激光无法穿过，而是在 ZMW 底部发生衍射，只能照亮很小的区域。DNA 聚合酶以及 DNA 聚合酶捕获的 DNA 模板就被固定在这一区域。只有在这个区域内，碱基携带的荧光基团才能被激活而被检测到，大幅地降低了背景荧光干扰。每个 ZWM 只固定一个 DNA 聚合酶。当一个 ZMW 结合的 DNA 模板数目不是一时，该 ZWM 所产生的测序结果会在后续数据分析时被过滤掉，由此保证每个可用的 ZMW 都是一个单独的 DNA 合成体系。15 万个 ZMW 聚集在一个芯片上，成为一个 SMRT Cell。PacBio RS II 测序仪一个流程内可以同时完成 8 个 SMRT Cell 的测序，产生约 3.2Gb 的数据。

3.1.2 phospholinked nucleotides

SMRT 测序的领一个核心技术是荧光基团标记在核苷酸 3‘端的磷酸上。在 DNA 合成过程中，3’端的磷酸键随着 DNA 链的延伸被切断，标记物被弃去，减少了 DNA 合成的空间位阻，维持 DNA 链连续合成，延长了测序读长。

而在第二代测序技术中，荧光基团标记在 DNA 链的 5‘端，在合成过程中，荧光标记物保留在 DNA 链上，随着 DNA 链的延伸会产生三维空间阻力，导致 DNA 链延长到一定程度后出现错读，这是限制二代测序读长的原因之一。

SMRT 测序最大限度地保持了聚合酶的活性，更接近天然状态的聚合酶反应体系。在实时监控系统下，DNA 链以每秒 10 个碱基的速度合成，从建库到测序，整个过程可在两天内完成。

3.1.3 优点

超长读长

无需模板扩增

直接检测表观修饰位点

直接测转录本

较高但完全随机的测序错误

基本原理：DNA 聚合酶和模板结合，4 色荧光标记 4 种碱基，经过 Watson 配对后不同的碱基加入，会发出不同光，根据光的波长与峰值可判断进入的碱基类型。这个 DNA 聚合酶是实现超长读长的关键之一，读长主要跟酶的活性保持有关，主要受激光对它的损伤的影响。PacBio 还在不断优化聚合酶的性能，比如给聚合酶加上免受激光影响的保护基团等，进一步地提高读长，提高测序质量和通量。

和其他基本测序技术一样，在反应管中进行的是大规模平行的多分子反应，怎样在其中进行单分子反应检测？周围有大量的荧光标记的游离碱基，怎样将反应信号与周围游离碱基的强大荧光背景区别出来？

通过一个物理现象解释：ZMW（zero-mode waveguides，零模波导孔）。例如微波炉壁上可看到有很多密集的小孔。小孔直径有考究，如果直径大于微波波长，能量就会穿透面板泄露。如果孔径小于波长，能量不会辐射外部，可起保护作用。

同理，在一个反应管（SMRTCell：单分子实时反应孔）中有许多这样的圆形纳米小孔，即 ZMW（零模波导孔），外径 100 多纳米，比检测激光波长小（数百纳米），激光从底部打上去后不能穿透小孔进入上方溶液区，能量被限制在一个小范围（体积 20X

10-21 L）里，正好足够覆盖需要检测的部分，使得信号仅来自这个小反应区域，孔外过多游离核苷酸单体依然留在黑暗中，将背景降到最低。

单个 ZMW 底部固定有一个结合了模板 DNA 的聚合酶，当加入测序反应试剂后，每个碱基配对合成后会发出相应的光并被检测。一个 SMRTCell 中有 15 万个 ZMW，每个孔中有一个单分子 DNA 链在高速合成，如众星闪烁。原始检测数据的结果，每合成一个碱基即显示为一个脉冲峰，每分钟 >100 个碱基的速度，配上高分辨率的光学检测系统，就能实时检进行检测。

关键点之二：荧光标记位点。这是影响测序长度的非常关键的因素。

二代测序都标记在 5‘端甲基上，在合成过程中，荧光标记物保留在 DNA 链上，随 DNA 链的延伸会产生三维空间阻力导致 DNA 链延长到一定程度后会出现错读。这是 NGS 的测序读长仅能达到 100 多 bp 到 200bp 的一个原因。

PacBio 平台的碱基荧光标记在 3‘端磷酸键。在 DNA 合成过程中正确的碱基进入时，在 3’端磷酸键的标记是会随磷酸键断裂自动被打断，标记物被弃去，亦即合成的 DNA 链不带荧光标记，和天然的 DNA 链合成产物一致，可以达到很长的读长。

（笔者疑问：是不是 NGS 改用 5‘端标记，就能实现延长读长？

答：首先，荧光标记在 3‘端磷酸键是 PacBio 的专利。其它公司没法做。荧光标记位点仅仅是影响读长的一个重要因素之一，PacBio 的单分子实时测序反应是最接近天然状态的聚合酶反应体系，最大限度地保持了聚合酶的活性。NGS 测序反应原理不尽相同，有的是焦磷酸测序反应，除聚合酶外有多种酶参与测序反应，要兼顾多种酶的活力并容非一件易的事；有的是通过添加保护基团来控制碱基的加入和检测，通过淬灭试剂来消除背景荧光和保护基团，这些都增加了测序反应体系本身的复杂性，此外，NGS 每加入一种碱基或一个碱基后都需要清洗步骤清除没有反应的多余反应物及反应产生的次级产物，这都影响了聚合酶的合成进程。）

关键点之三：时空段概念

合成过程中，每次进入一个碱基，原始数据会实时地产生一个脉冲峰，每两个相邻的脉冲峰之间有一定的距离，也就是有一个时间段的概念。这个距离的长短与模板上碱基是否存在修饰有关，如果有碱基修饰，就像开车经过路障时，通过速度会减慢，导致两个相邻峰之间距离加大。根据这个距离的变化，可以判断模板相应位点是否出现碱基修饰，并且结果是实时的。甲基化就是一种主要的碱基修饰，PacBio 技术不仅可以提供序列信息，还可提供实时信息了解模板修饰的情况，用于甲基化等碱基修饰研究。

第四章　Oxford Nanopore

The Single-Molecule Nanopore DNA Sequencing

第五章　Helicos Biosciences

Helicos Biosciences Corporation 已是过去式。

True Single-Molecule Sequencing

第二部分

全基因组测序

第六章　预处理和测序质量控制

6.1 Base calling

Next-generation sequencing platforms are dramatically reducing the cost of DNA sequencing. With these technologies, bases are inferred from light intensity signals, a process commonly referred to as base-calling. Thus, understanding and improving the quality of sequence data generated using these approaches are of high interest.

6.1.1 bcl2fastq Conversion Software

`https://support.illumina.com/downloads/bcl2fastq_conversion_software_184.html`

6.2 Read quality control

Next-generation sequencing (NGS) technologies have been widely used in life sciences. However, several kinds of sequencing artifacts, including low-quality reads and contaminating reads, were found to be quite common in raw sequencing data, which compromise downstream analysis. Therefore, quality control (QC) is essential for raw NGS data.

6.2.1 Picard

`http://broadinstitute.github.io/picard/`

6.2.2 FastQC

`http://www.bioinformatics.babraham.ac.uk/projects/fastqc/`

6.2.3 NGS QC Toolkit

http://nipgr.res.in/ngsqctoolkit.html

6.2.4 SAMtools

http://www.htslib.org/

6.2.5 FASTX-Toolkit

http://hannonlab.cshl.edu/fastx_toolkit/

6.2.6 BAMStats

6.3 Error correction

6.4 Duplicate read removal

6.5 Adapter trimming

6.5.1 Trimmomatic

Trimmomatic 是使用 java 编写的去除 Illumina 高通量测序数据接头的程序 [1]。最新程序可以通过以下链接获取：http://www.usadellab.org/cms/index.php?page=trimmomatic

使用简介

先按照官方说明，来两段代码吧：

第一段

```
java -jar trimmomatic-0.35.jar PE -phred33
    input_forward.fq.gz input_reverse.fq.gz
    output_forward_paired.fq.gz output_forward_unpaired.fq.gz
    output_reverse_paired.fq.gz output_reverse_unpaired.fq.gz
    ILLUMINACLIP:TruSeq3-PE.fa:2:30:10
    LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
```

这段代码，对 Illumina 测序平台双端测序（PE，Paired End）的数据文件 input_forward.fq.gz 和 input_reverse.fq.gz 进行了以下操作：

- 去除测序接头（ILLUMINACLIP:TruSeq3-PE.fa:2:30:10）
- 去除开头的 3 个或低质量碱基（LEADING:3）
- 去除结尾的 3 个或低质量碱基（TRAILING:3）
- 以四个碱基为一个窗口,当平均碱基质量低于 15 时,去除（SLIDINGWINDOW:4:15）
- 去除短语 36 个碱基的 Reads（MINLEN:36）

其输出文件为：

- output_forward_paired.fq.gz：经清理后，input_forward.fq.gz 文件剩余的配对的 Reads。
- output_forward_unpaired.fq.gz：经清理后，input_forward.fq.gz 文件剩余的未配对的 Reads。
- output_reverse_paired.fq.gz：经清理后，input_reverse.fq.gz 文件剩余的配对的 Reads。
- output_reverse_unpaired.fq.gz：经清理后，input_reverse.fq.gz 文件剩余的未配对的 Reads。

第二段

```
java -jar trimmomatic-0.35.jar SE -phred33
    input.fq.gz
    output.fq.gz
    ILLUMINACLIP:TruSeq3-SE:2:30:10
    LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36
```

这一段代码，和上一段代码完成了相同的工作，不过，其输入文件为 SE 测序数据（input.fq.gz）

使用说明

Trimmomatic 可以用来处理 Illumina 双端测序和单端测序产生的 FASTQ 文件。其中 FASTQ 文件可以采用 Phred-33 或 Phred-64 碱基质量表示方法。Trimmomatic 支撑以 gzip 和 bzip2 压缩的 FASTQ 文件。

其支持的命令包括：

命令	说明
ILLUMINACLIP	去除 Illumina 测序平台的接头
SLIDINGWINDOW	以窗口的形式从 5 端扫描 reads。如果平均碱基质量低于给定值，则切断 Reads
MAXINFO	综合考虑 Read 长度和碱基质量，以使 Read 的评分最高
LEADING	去掉开头的几个低于阈值的碱基
TRAILING	去掉结尾的几个低于阈值碱基
CROP	截取开头的指定数目的碱基
HEADCROP	去除开头指定数目的碱基
MINLEN	去掉小于指定长度的 reads
AVGQUAL	去掉评价碱基质量小于指定值的 reads
TOPHRED33	将碱基质量表示方法转化为 Phred-33
TOPHRED64	将碱基质量表示方法转化为 Phred-64

```
java -jar <path to trimmomatic jar> SE [-threads <threads>]
    [-phred33 | -phred64] [-trimlog <logFile>]
    <input> <output> <step 1> ...

java -classpath <path to trimmomatic jar>
    org.usadellab.trimmomatic.TrimmomaticSE
    [threads <threads>] [-phred33 | -phred64] [-trimlog <logFile>]
    <input> <output> <step 1> ...

java -jar <path to trimmomatic.jar> PE
    [-threads <threads] [-phred33 | -phred64]
    [-trimlog <logFile>]
    [-basein <inputBase> | <input 1> <input 2>]
    [-baseout <outputBase> | <unpaired output 1>
```

```
        <paired output 2> <unpaired output 2>
    <step 1> ...

java -classpath <path to trimmomatic jar>
    org.usadellab.trimmomatic.TrimmomaticPE
    [threads <threads>] [-phred33 | -phred64]
    [-trimlog <logFile>]
    [-basein <inputBase> | <input 1> <input 2>]
    [-baseout <outputBase> | <paired output 1> <unpaired output 1>
        <paired output 2> <unpaired output 2>
    <step 1> ...
```

6.5.2 cutadapt

6.5.3 NGS QC Toolkit

6.5.4 BCL2FASTQ Conversion Software

6.5.5 FASTX-Toolkit

http://hannonlab.cshl.edu/fastx_toolkit/

6.6 Read Clustering

6.7 k-mer counting

6.7.1 Jellyfish

6.8 Depth of Coverage

6.9 Variant recalibration

6.10 Ion AmpliSeq™

参考文献

[1] Anthony M. Bolger, Marc Lohse, and Bjoern Usadel. Trimmomatic: a flexible trimmer for illumina sequence data. Bioinformatics, 30(15):2114, 2014.