Chapter 2. Genome sequencing and assemble technology

Huaqin He

PPT slides and Message @ http://jxpt.fafu.edu.cn/meol/homepage/common/

Email: 1156743645@qq.com

Lecture 2.1 Genome sequencing technology

OUTLINES

1. Conventional sequencing technologies
   一代测序技术
2. 2nd generation sequencing technologies
   二代测序技术（NGS）
3. 3rd generation sequencing technologies
   三代测序技术

1. Conventional sequencing technologies

(by Ladder fragments separation)

1.1 一代测序技术

❶Maxam Gilbert's chemical DNA sequencing method

化学裂解测序法

❷Sanger's dideoxy-chain termination method

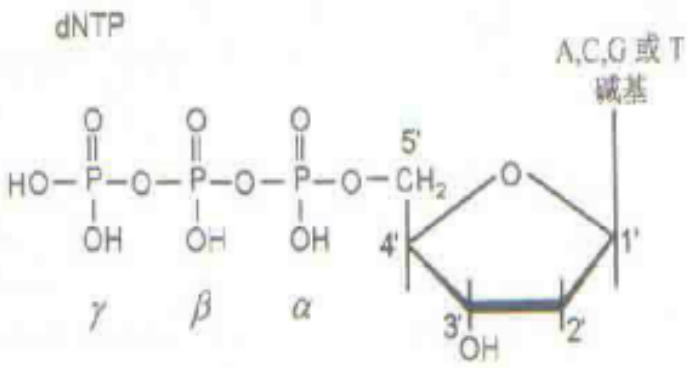
双脱氧链终止法

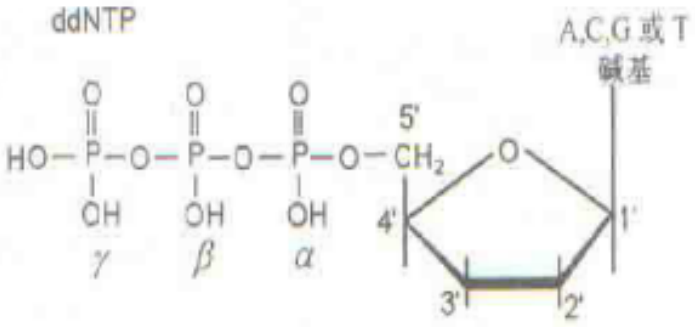
❸Dye-termination method

荧光标记链终止法

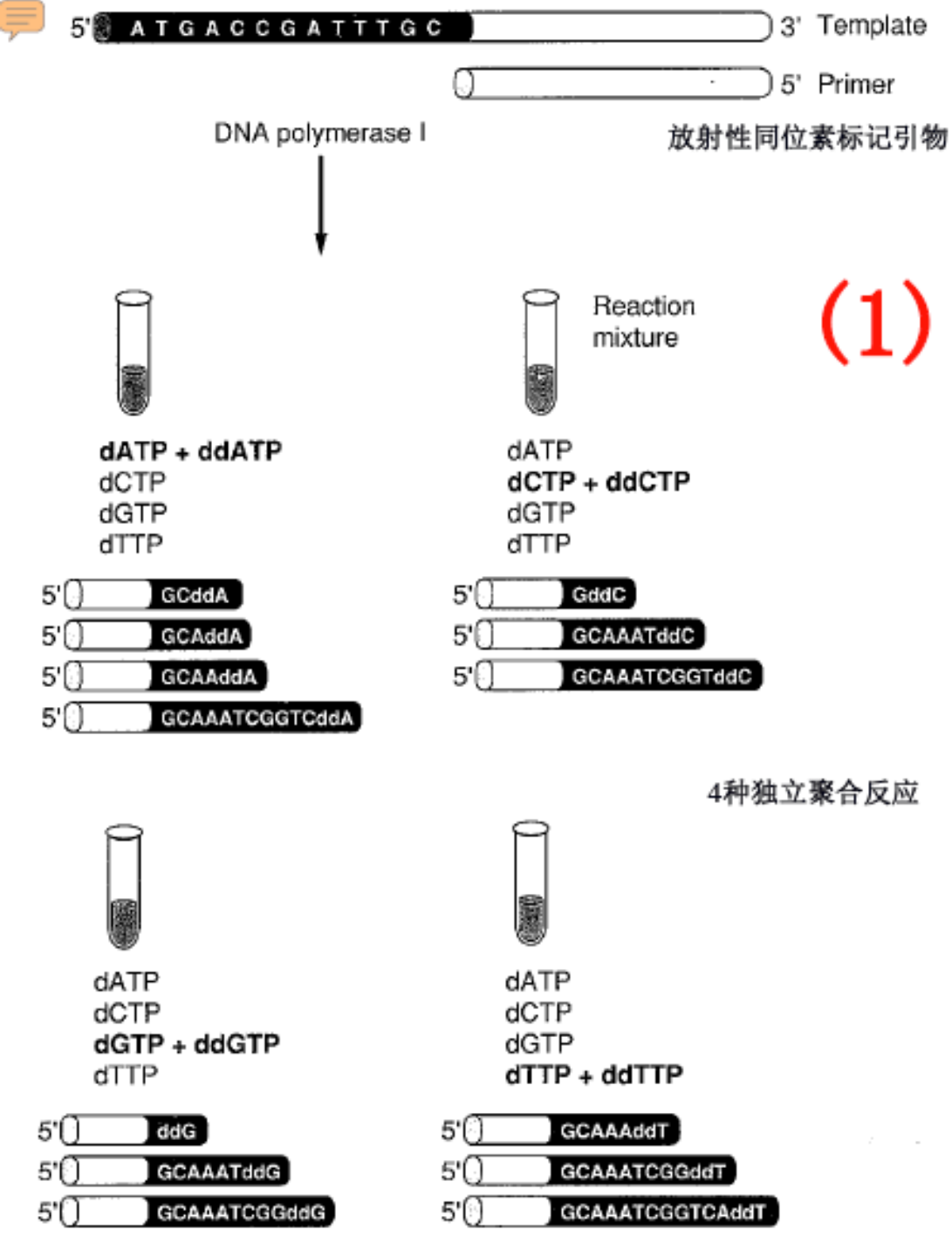
链终止法基本原理：

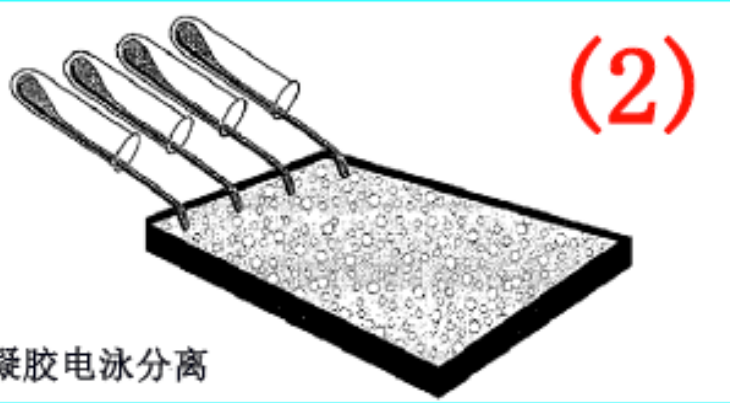
➢利用**DNA**聚合酶不能够区分**dNTP**和**ddNTP**的特性，使**ddNTP**参入到寡核苷酸链的**3'-**末端。因为**ddNTP 3'**不是**-OH**，不能与下一个核苷酸聚合延伸，从而终止**DNA**链的增长。

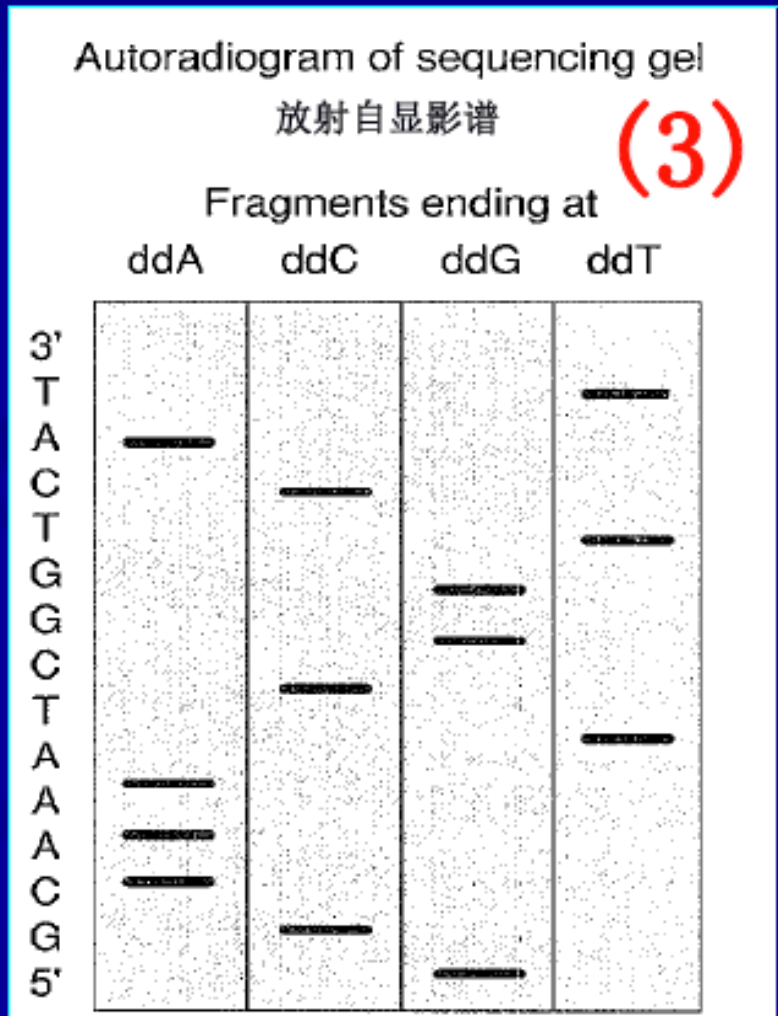




➢聚丙烯酰胺凝胶电泳可以区分长度只差一个核苷酸的**DNA**分子。

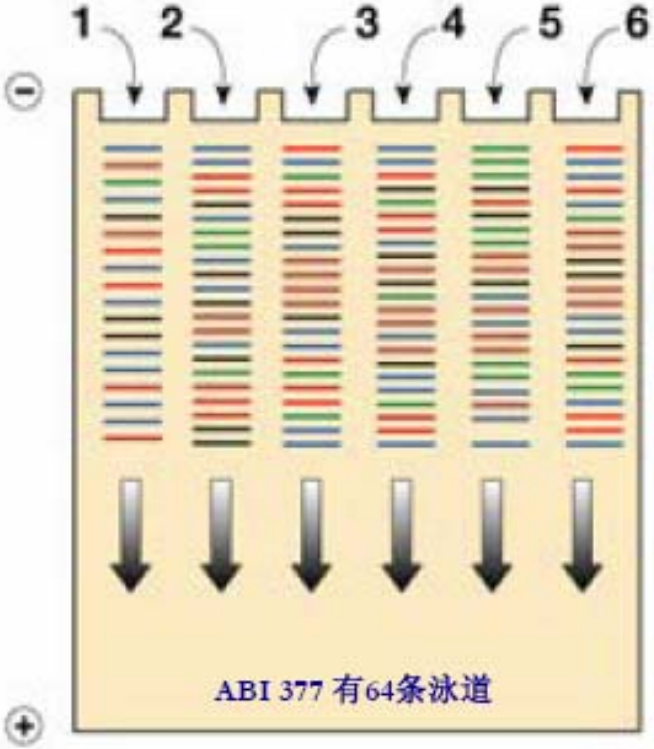

5' ATGACCGATTTGC 3' Template
5' Primer
DNA polymerase I
放射性同位素标记引物
(1)
Reaction mixture
dATP + ddATP
dCTP
dGTP
dTTP
5' GCddA
5' GCAddA
5' GCAAddA
5' GCAAATCGGTCddA
dATP
dCTP + ddCTP
dGTP
dTTP
5' GddC
5' GCAAATddC
5' GCAAATCGGTddC
4种独立聚合反应
dATP
dCTP
dGTP + ddGTP
dTTP
5' ddG
5' GCAAATddG
5' GCAAATCGGddG
dATP
dCTP
dGTP
dTTP + ddTTP
5' GCAAAddT
5' GCAAATCGGddT
5' GCAAATCGGTCAddT

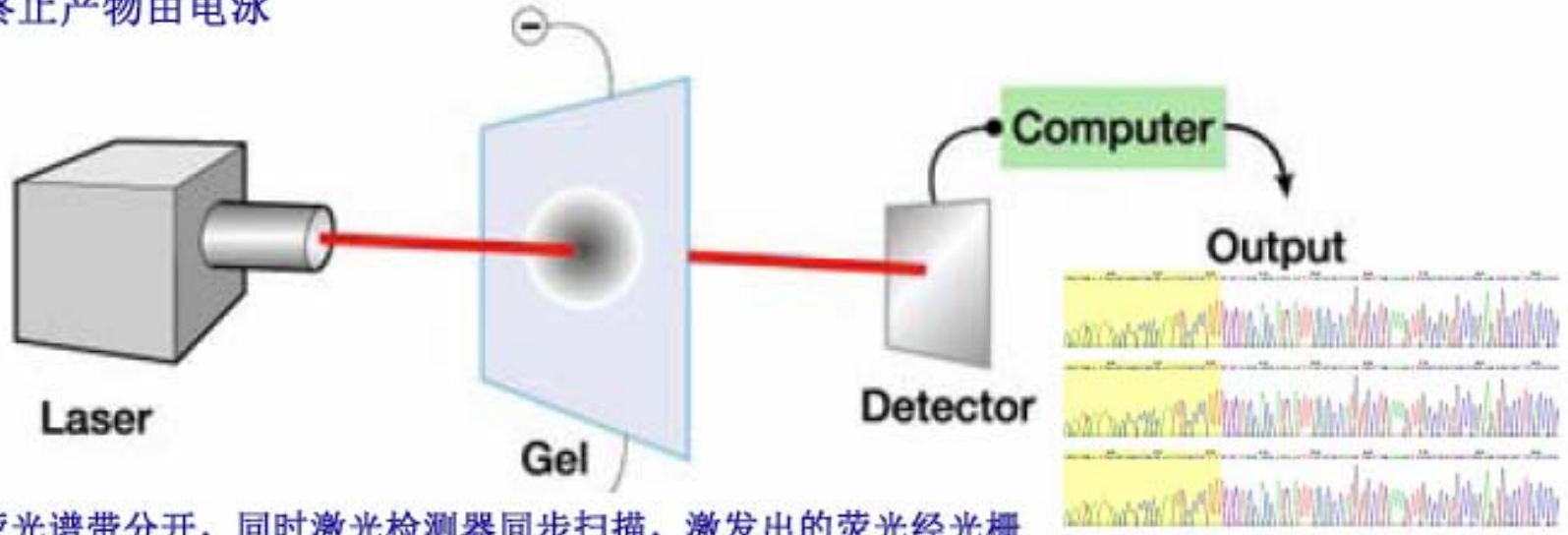

(2)
凝胶电泳分离


Autoradiogram of sequencing gel
放射自显影谱
(3)
Fragments ending at
ddA
ddC
ddG
ddT
3' T A C T G G C T A A A C G 5'

荧光测序仪

- Four fluorescently labeled terminators（4种荧光染料标记链终止核苷酸）
- One DNA polymerization reaction（1个聚合反应）
- Replication product separated by gel-electrophoresis（荧光标记链终止产物由电泳分离）





经电泳后各个荧光谱带分开，同时激光检测器同步扫描，激发出的荧光经光栅分光后打到CCD摄像机上同步成像，将信息输送给电脑进行分析和保存

1.2 Limitation of 1st sequencing technologies

一代测序技术的局限性

❶ Low throughput 低通量

- Time-consuming separation of chain-terminated fragment
- Hard to produce massively parallel system based electrophoresis separation

❷ High sequencing cost 高成本

- Complex sample preparation & handling
- High reagent consumption
- Difficult to miniaturize

Year 2000
(Celera Genomics)

Now

2. 2nd generation sequencing technologies

2.1 Methods of NGS 二代测序技术

❶ Pyrophosphate Sequencing Sequence by Synthesize
焦磷酸测序 （Roche 454）

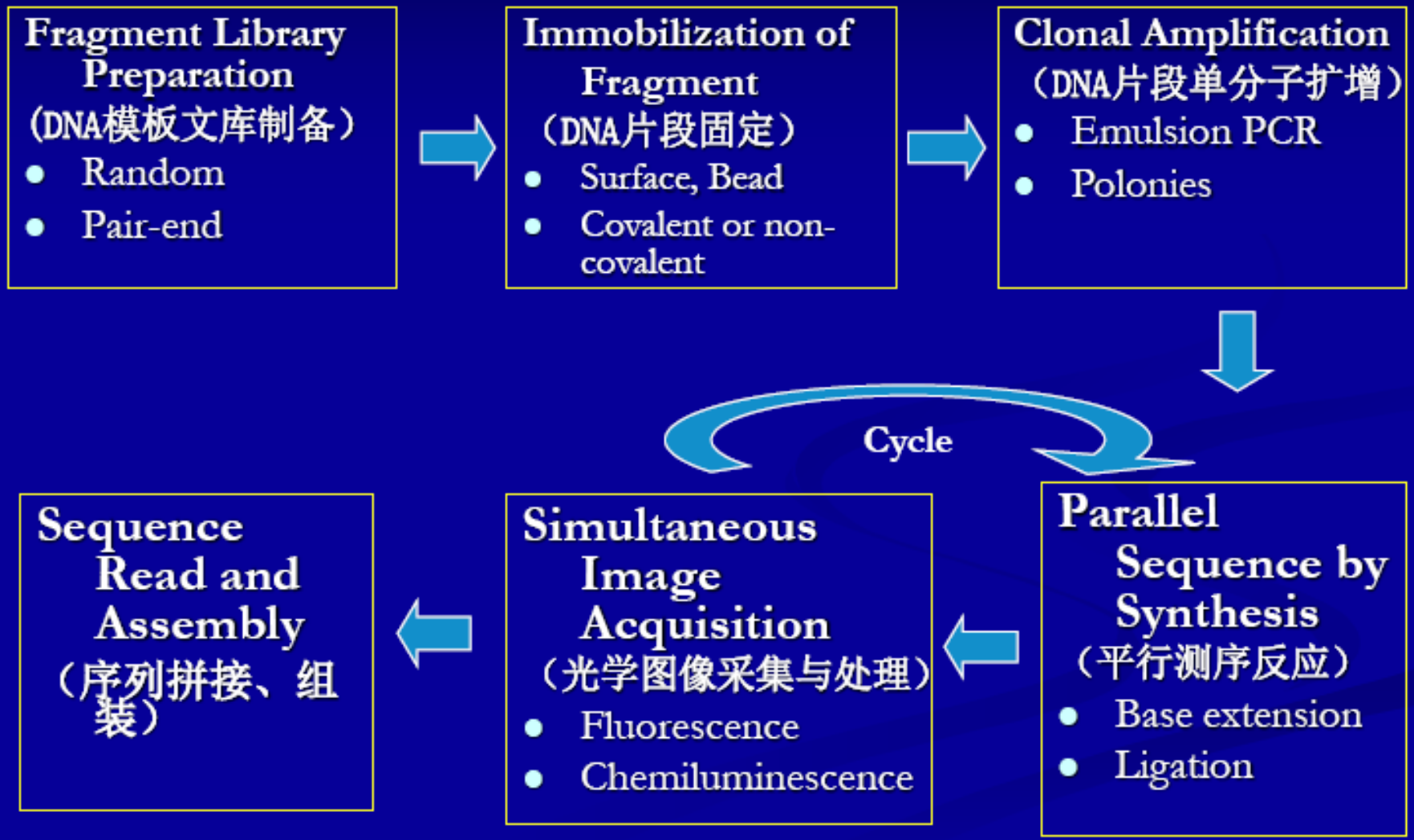
❷ Sequence by Synthesize
合成测序 （Illumina Solexa）

❸ Sequence by Ligation
连接法测序 （ABI SOLiD）

2.2 Workflow of NGS 二代测序流程

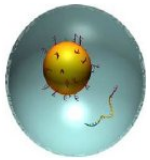
❶ Sample fragmentation
❷ Library preparation
❸ Clonal Amplification
❹ Sequencing reaction
❺ Data analysis

Workflow of 2nd-Gen Sequencing
二代测序流程

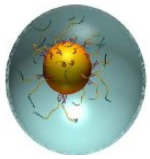


CASE: Protocol of Roche 454 Sequencing method

Library of double stranded DNA molecules

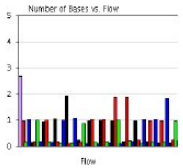


One DNA molecule per bead



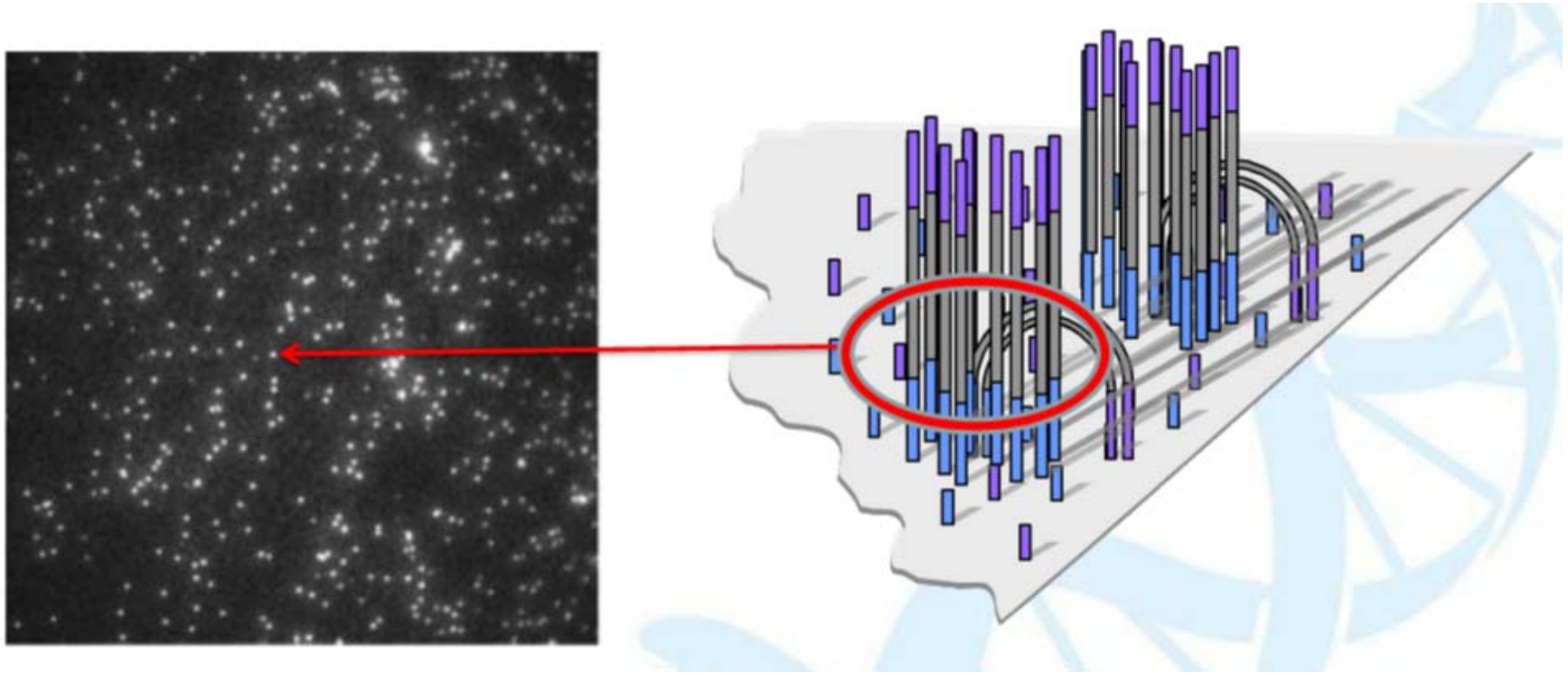
Clonal amplification of that single molecule to ~10 million copies

Independent sequencing of each bead



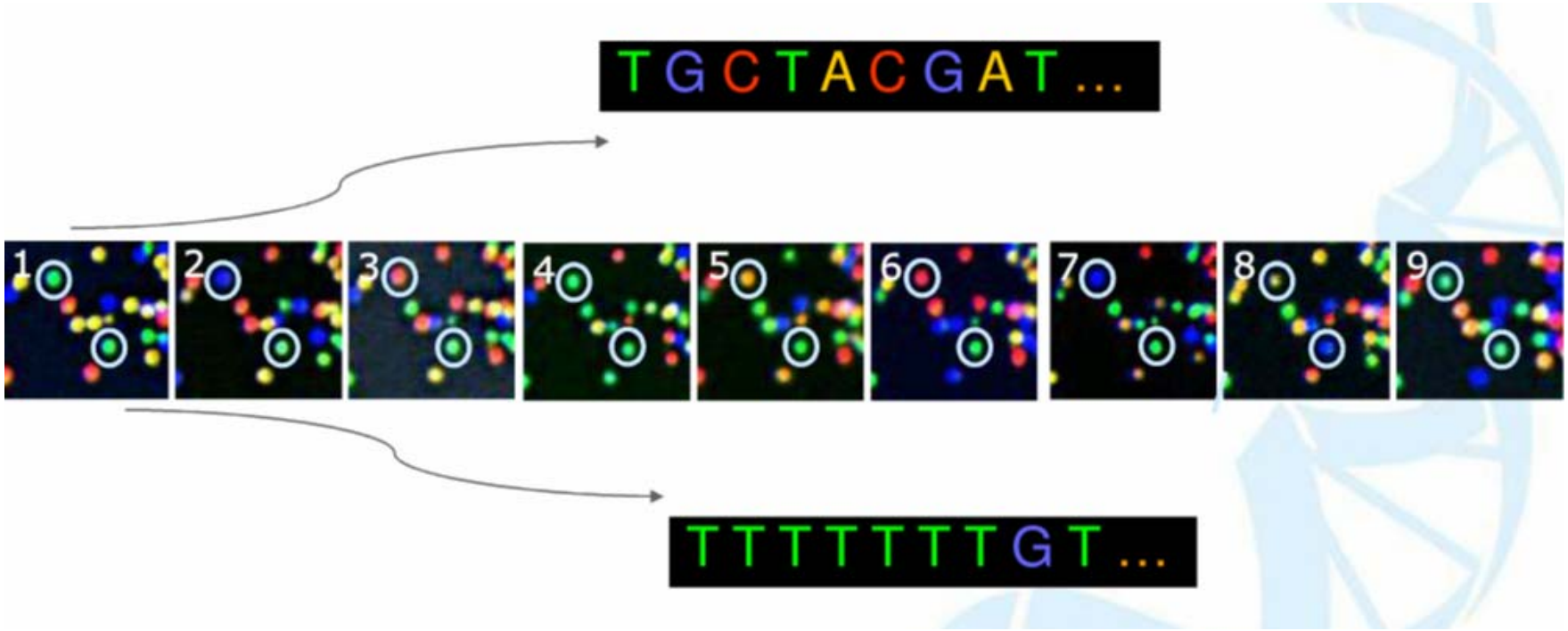
One Bead = One Read = One DNA molecule

CASE: Protocol of Solexa Sequencing technology



每个亮点代表每个基因簇上的信号

CASE: Protocol of Roche 454 Sequencing method



Base calling

CASE: Key technology in NGS

❶ 测序策略-循环芯片测序法 (cyclic-array sequencing).

❷ 所谓的循环芯片测序法，对布满DNA样品的芯片重复进行基于DNA的聚合酶反应（模板变性、引物退火杂交及延伸）以及荧光序列读取反应.

❸ 与传统测序法相比，循环芯片测序法具有操作更简易、费用更低廉的优势.

2.3 Limitation of 2nd sequencing technologies

公司	系统名	测序长度	优点	缺点
Roche/454	FLX System	200-700	读长长	单碱基重复错误；通量低；试剂贵
Illumina	HiSeq 2000/miSeq	2 x 150	通量非常高	速度慢
ABI/SOLiD	5500xl SOLiD	35-50	通量高	读长太短

Table 1 Price comparison of benchtop instruments and sequencing runs

Platform	List price	Approximate cost per run	Minimum throughput (read length)	Run time	Cost/Mb	Mb/h
454 GS Junior	$108,000	$1,100	35 Mb (400 bases)	8 h	$31	4.4
Ion Torrent PGM						
(314 chip)	$80,490[a,b]	$225[c]	10 Mb (100 bases)	3 h	$22.5	3.3
(316 chip)		$425	100 Mb[d] (100 bases)	3 h	$4.25	33.3
(318 chip)		$625	1,000 Mb (100 bases)	3 h	$0.63	333.3
MiSeq	$125,000	$750	1,500 Mb (2 × 150 bases)	27 h	$0.5	55.5

2.3 Limitation of 2nd sequencing technologies

二代测序技术的局限性

❶ Short Read 读长是二代的一大局限

- 50-100 base per read

❷Low throughput 通量不够高

❸ High sequencing cost 高成本

- Complex sample preparation

3. 3rd generation sequencing technologies

3.1 Methods of 3rd sequencing technologies

三代测序技术

❶ Single Molecule Real Time (SMRT)

单分子实时测序技术 (Pacific Biosciences 公司)

❷ Single Molecule Sequencing

单分子测序技术 (Helicos公司)

❸ Nanopore single molecule

纳米孔单分子测序技术 (Oxford Nanopore 公司)

3.2 Principle of 3rd sequencing technologies

❶ 单分子荧光测序：脱氧核苷酸用荧光标记，显微镜实时记录荧光的强度变化。当荧光标记的脱氧核苷酸被掺入DNA链时，荧光同时能在DNA链上探测到。当它与DNA链形成化学键时，其荧光基团就被DNA聚合酶切除，荧光消失。这种荧光标记的脱氧核苷酸不会影响DNA聚合酶的活性，并且在荧光被切除之后，合成的DNA链和天然的DNA链完全一样。

❷ 纳米孔测序法：借助电泳驱动单个分子逐一通过纳米孔来实现测序。由于纳米孔仅允许单个核酸聚合物通过，而ATCG单个碱基的带电性质不一样，通过电信号的差异就能检测出通过的碱基类别，从而实现测序。

3.2 Principle of SMRT

❶ Millions of zero-mode waveguides (ZMWs), embedded with only one set of enzymes and DNA template that can be detected during the whole process.

❷ During the reaction, the enzyme will incorporate the nucleotide into the complementary strand and cleave off the fluorescent dye previously linked with the nucleotide.

❸ The camera inside the machine will capture signal in a movie format in real-time observation

3.2 Principle of SMRT

❶ zero-mode waveguides (ZMWs): Pacific Biosciences公司发明，直径只有几十纳米的纳米孔。单分子的DNA聚合酶被固定在这个孔内。在这么小的孔内，DNA链周围的荧光标记的脱氧核苷酸有限，而且由于A, T, C, G这4种荧光标记的脱氧核苷酸非常快速地从外面进入到孔内又出去，它们形成了非常稳定的背景荧光信号。而当某一种荧光标记的脱氧核苷酸被掺入到DNA链时，这种特定颜色的荧光会持续一小段时间，直到新的化学键形成，荧光基团被DNA聚合酶切除为止。

❷ 共聚焦显微镜实时地快速地对集成在板上的无数的纳米小孔同时进行记录.

3.3 Application of 3rd sequencing technologies

❶ DNA测序：三代测序现在可以测几千个碱基，二代测序现在可以测到上百个碱基，一秒可以测10个碱基，速度是化学法测序的2万倍.

❷ RNA序列的直接测序。以RNA为模板复制DNA的逆转录酶也同样可以直接检测。RNA的直接测序，将大大降低体外逆转录产生的系统误差。

❸ 甲基化DNA序列的直接测序。DNA聚合酶复制A, T, C, G的速度不同。正常的C或者甲基化的C为模板，DNA聚合酶停顿的时间不同。根据不同的时间，可以判断模板的C是否甲基化。

3.4 Advantage and limitation of 3rd sequencing technologies

❶ Single molecule sequencing.

提高了样本的检测速度

❷ PCR is not needed before sequencing.

减少了样品准备时间，避免了PCR过程的错误

❸ The signal is captured in real time.

信号捕捉的实时性

3.4 Limitation of 3rd sequencing technologies

二代测序技术的局限性

❶ Low accuracy

三代读长超长，准确低，费用高

❷ Technology improving

Pac Bio的SMRT技术比较成熟，多次读取后准确率有所提高。基于纳米孔的技术还在起步阶段，不过灵活性上有优势.

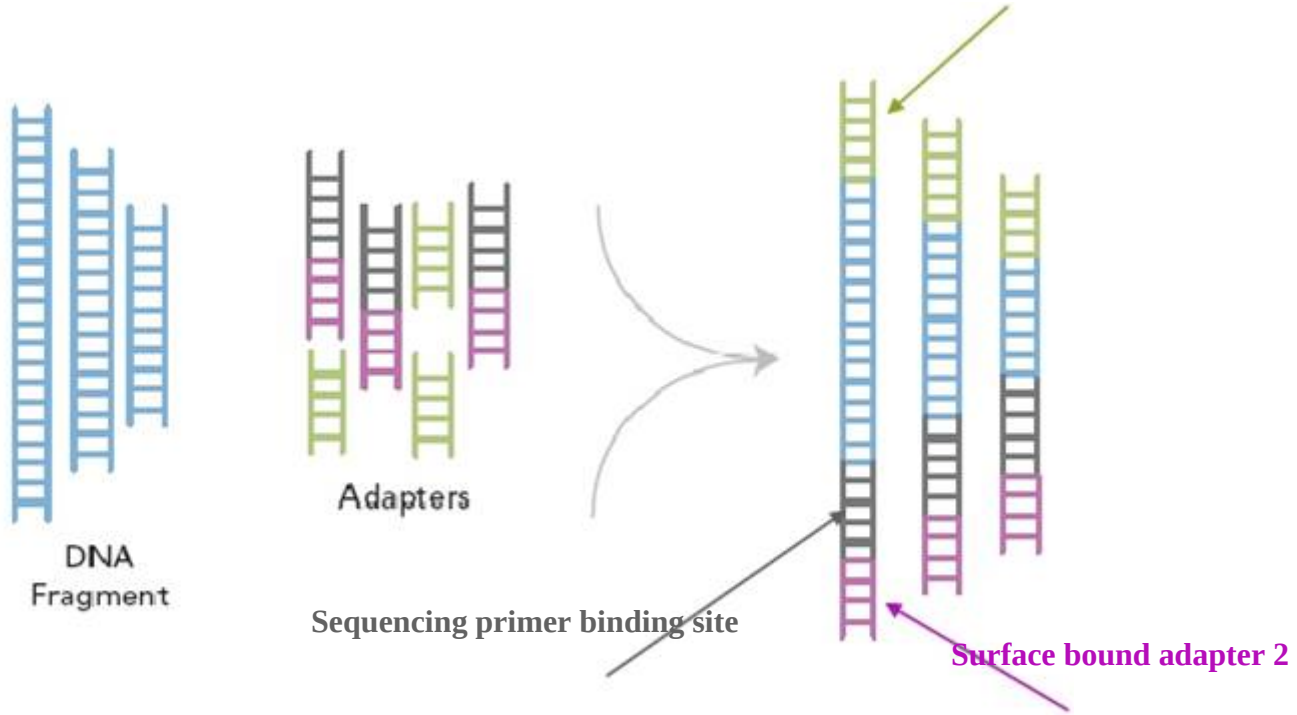
Summary:

	测序方法/平台	公司/公司网站	方法/酶	测序长度	每个循环的数据产出量	每个循环耗时	主要错误来源
第一代测序技术	Sanger/ABI3730 DNA Analyzer	Applied Biosystems www.appliedbiosystems.com	Sanger法/DNA聚合酶	1000 bp	56 Kb		
第二代测序技术	454/GS FLX Titanium Series	Roche www.roche-applied-science.com	焦磷酸测序法/DNA聚合酶	400 bp	400-600 Mb	10 h	插入、缺失
	Solexa/Illumina Genome Analyzer	Illumina www.illumina.com	边合成边测序/DNA聚合酶	2*75 bp	20.5-25 Gb	9.5 d	替换
	SOLiD/SOLiD 3 system	Applied Biosystems www.appliedbiosystems.com	连接酶测序/DNA连接酶	2*50 bp	10-15 Gb	6-7 d	替换
第三代测序技术	Heliscope/Helicos Genetic Analysis System	Helicos www.helicosbio.com	边合成边测序/DNA聚合酶	30-35 bp	21-28 Gb	8 d	替换
	SMRT	Pacific Biosciences www.pacificbiosciences.com	边合成边测序/DNA聚合酶	100000 bp			
	纳米孔单分子	Oxford Nanopore Technologies www.nanoporetech.com	电信号测序/核酸外切酶	无限长			

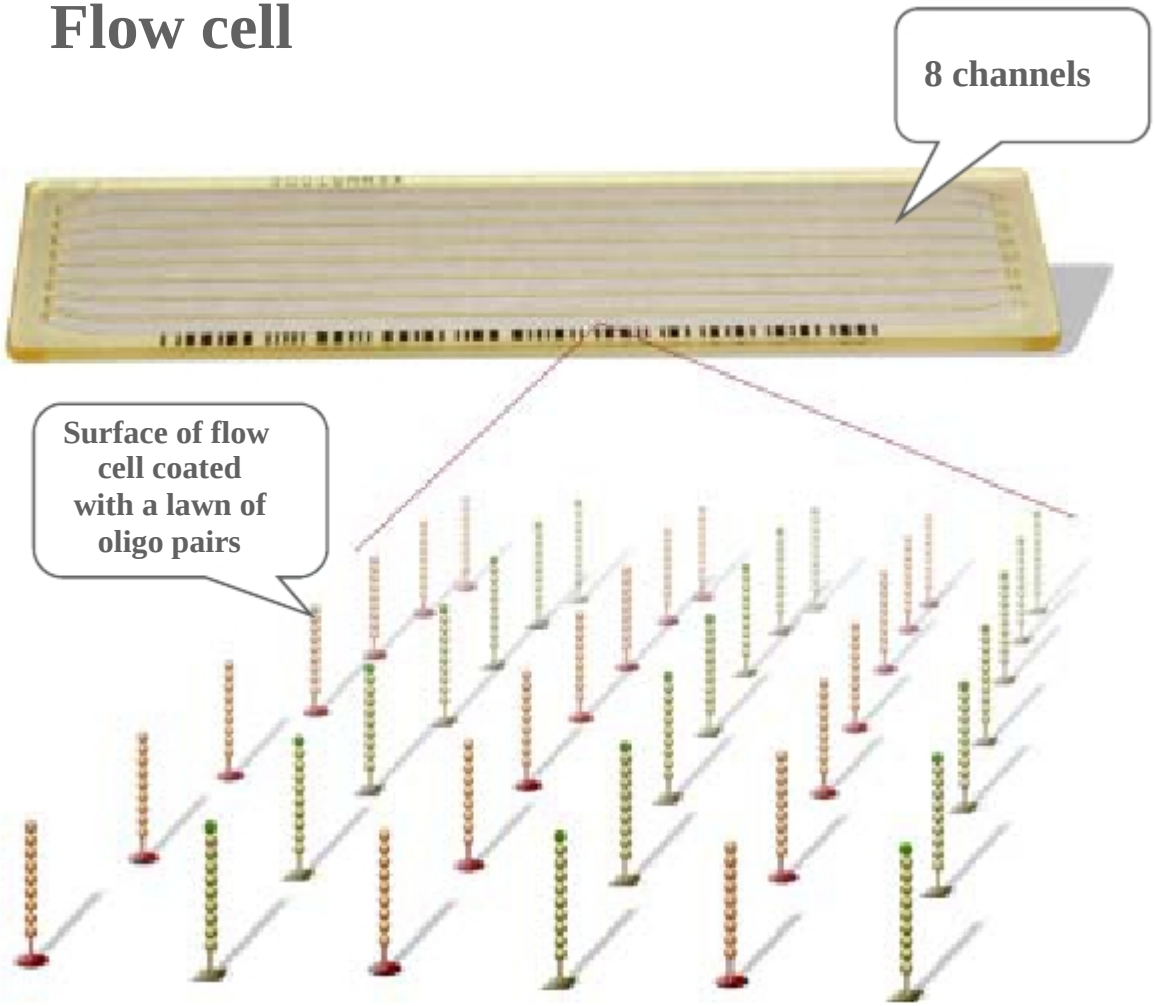
Thanks for your attentions!

Illumina Solexa Sequencing

Sample Prep - Resequencing



Flow cell



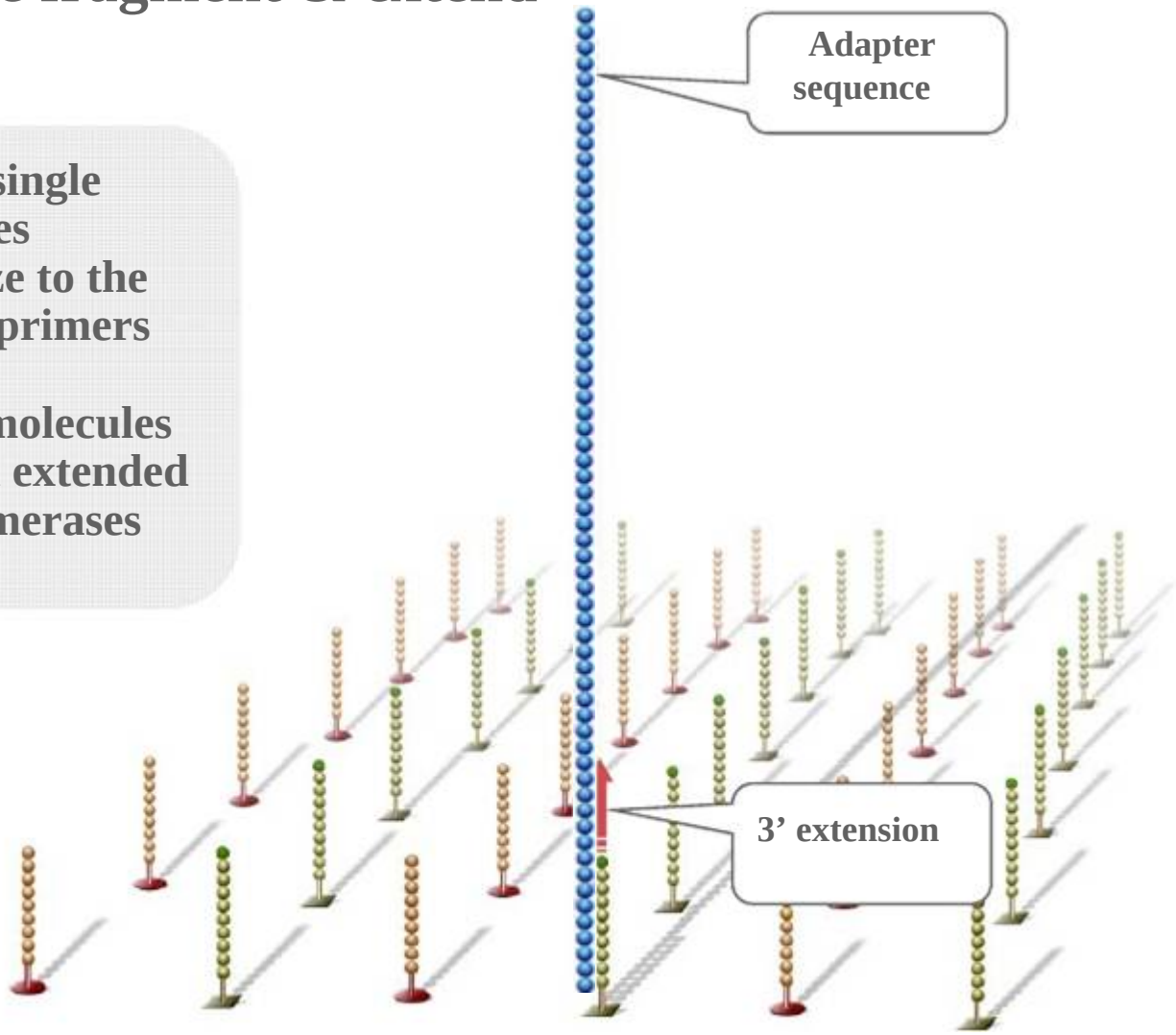
Key to the simplified workflow

- Clonal clusters are generated in a contained environment (need no clean rooms)
- Sequencing also performed in the flow cell on the generated clusters

Cluster generation: Hybridize fragment & extend

> 50 M single molecules hybridize to the lawn of primers

Bound molecules are then extended by polymerases

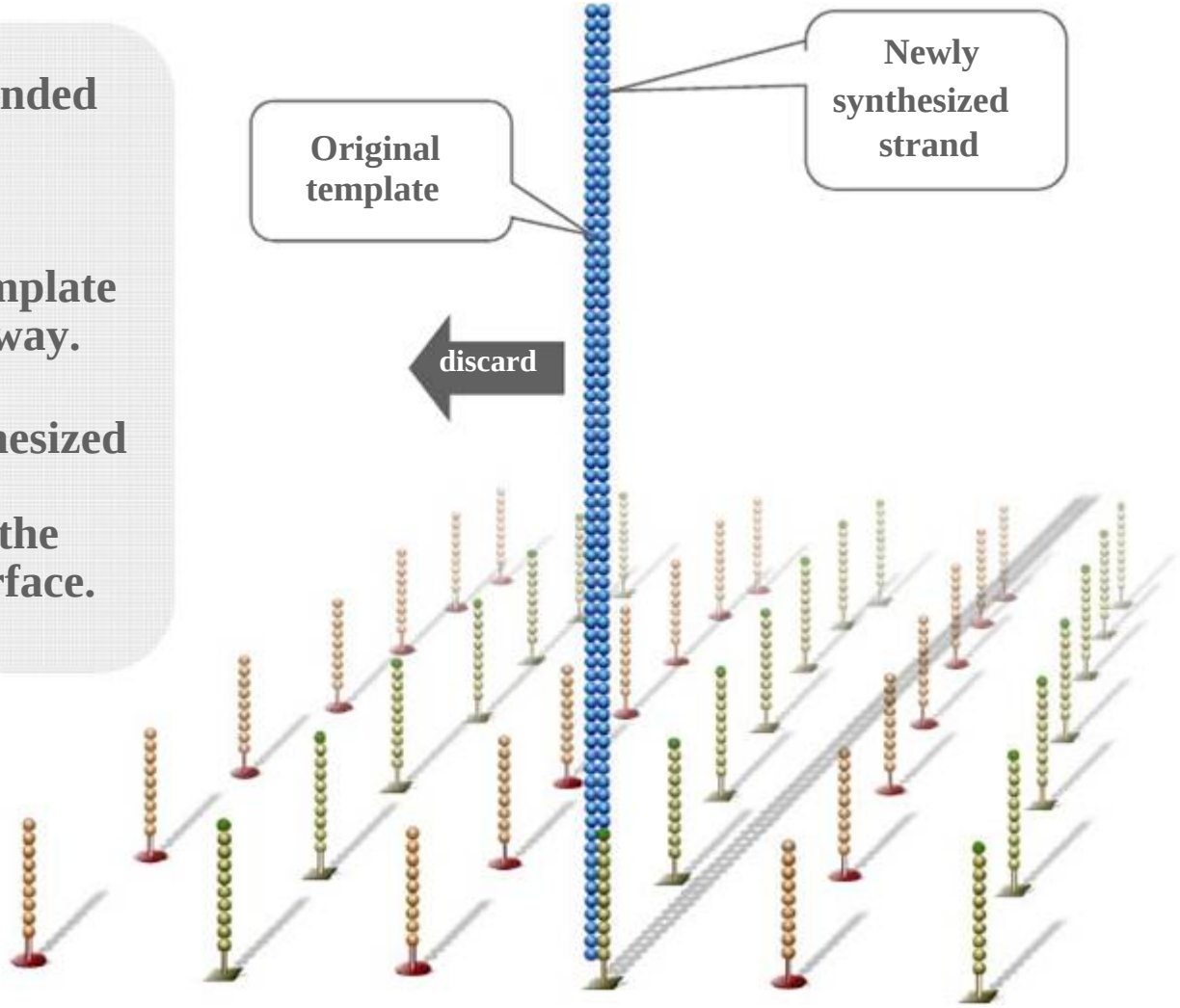


Cluster generation:
Denature double-stranded DNA

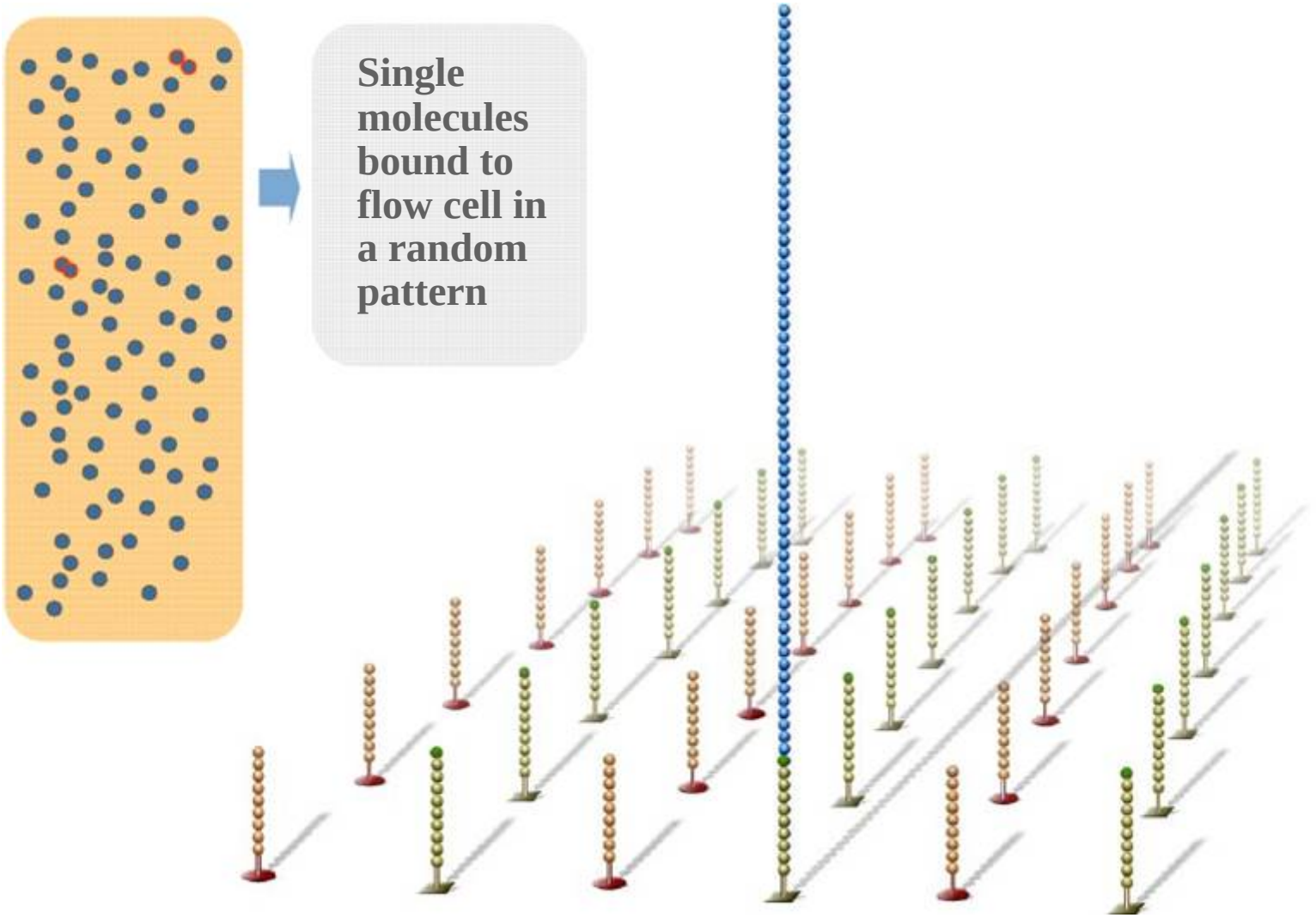
Double-stranded molecule is denatured.

Original template is washed away.

Newly synthesized covalently attached to the flow cell surface.



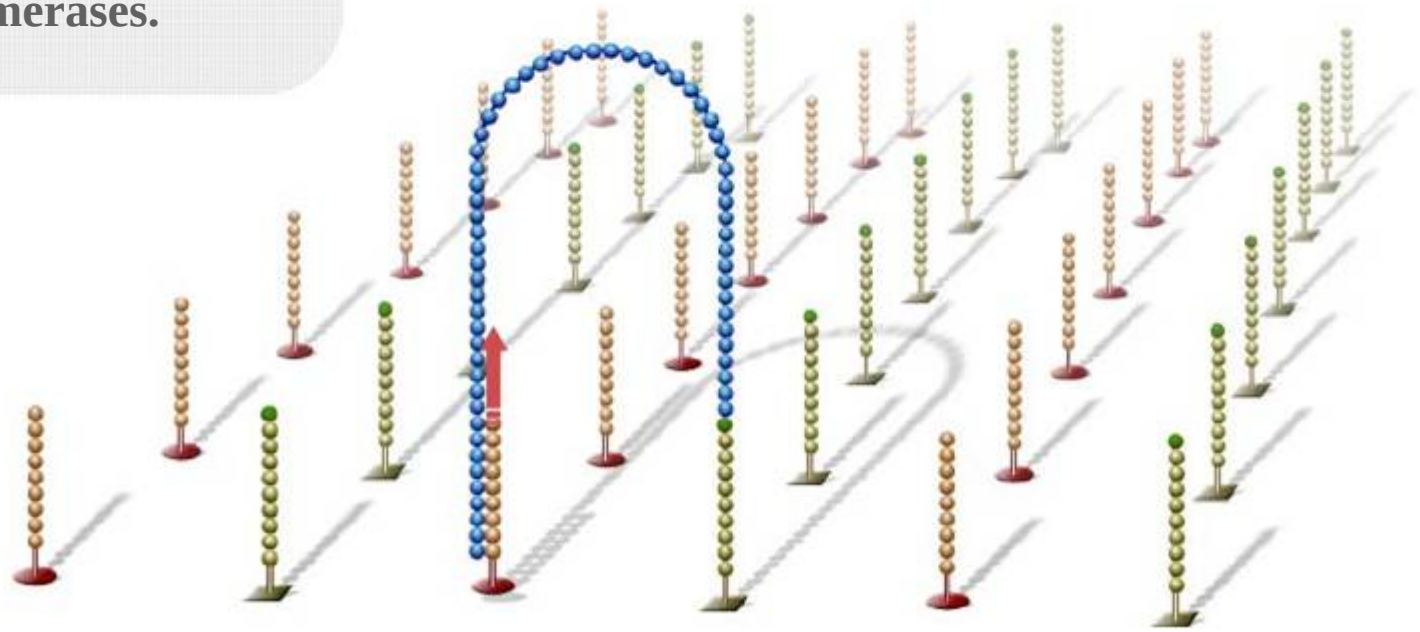
Cluster generation: Covalently bound spatially separated single molecules



Cluster generation: Bridge amplification

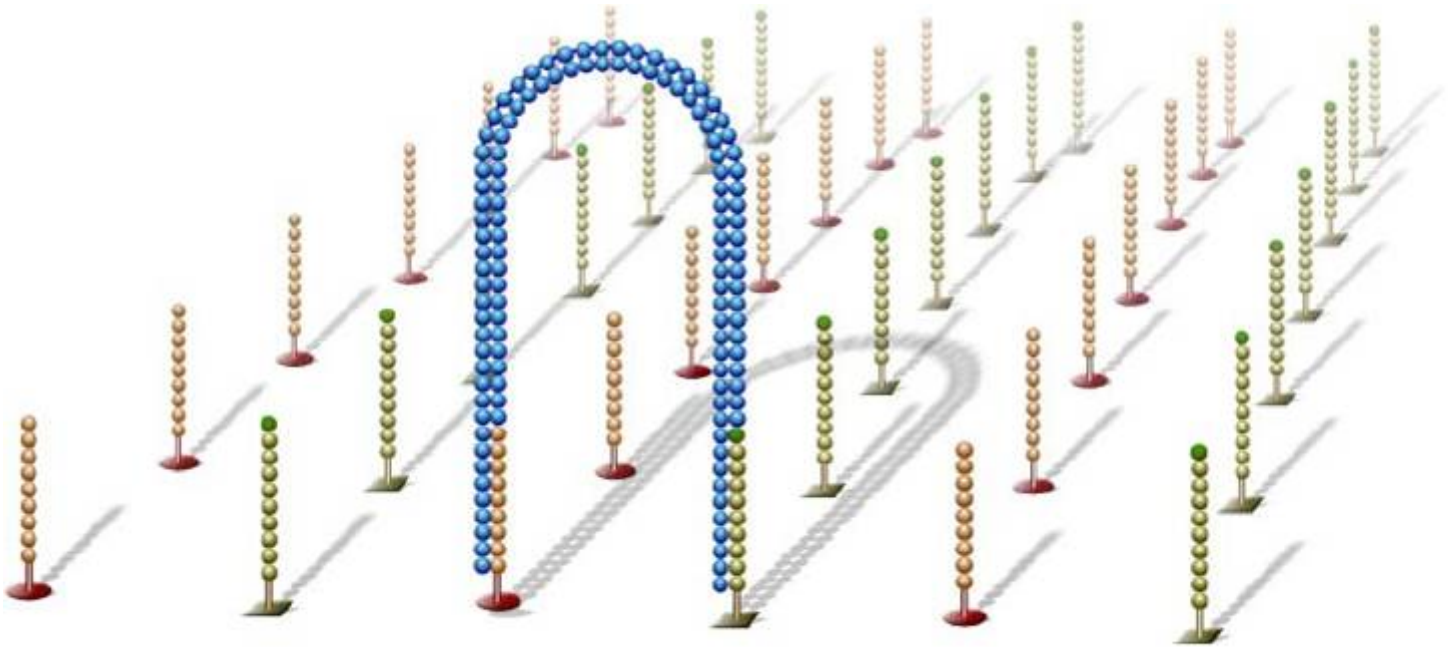
Single-strand flips over to hybridize to adjacent primers to form a bridge.

Hybridized primer is extended by polymerases.



Cluster generation: Bridge amplification

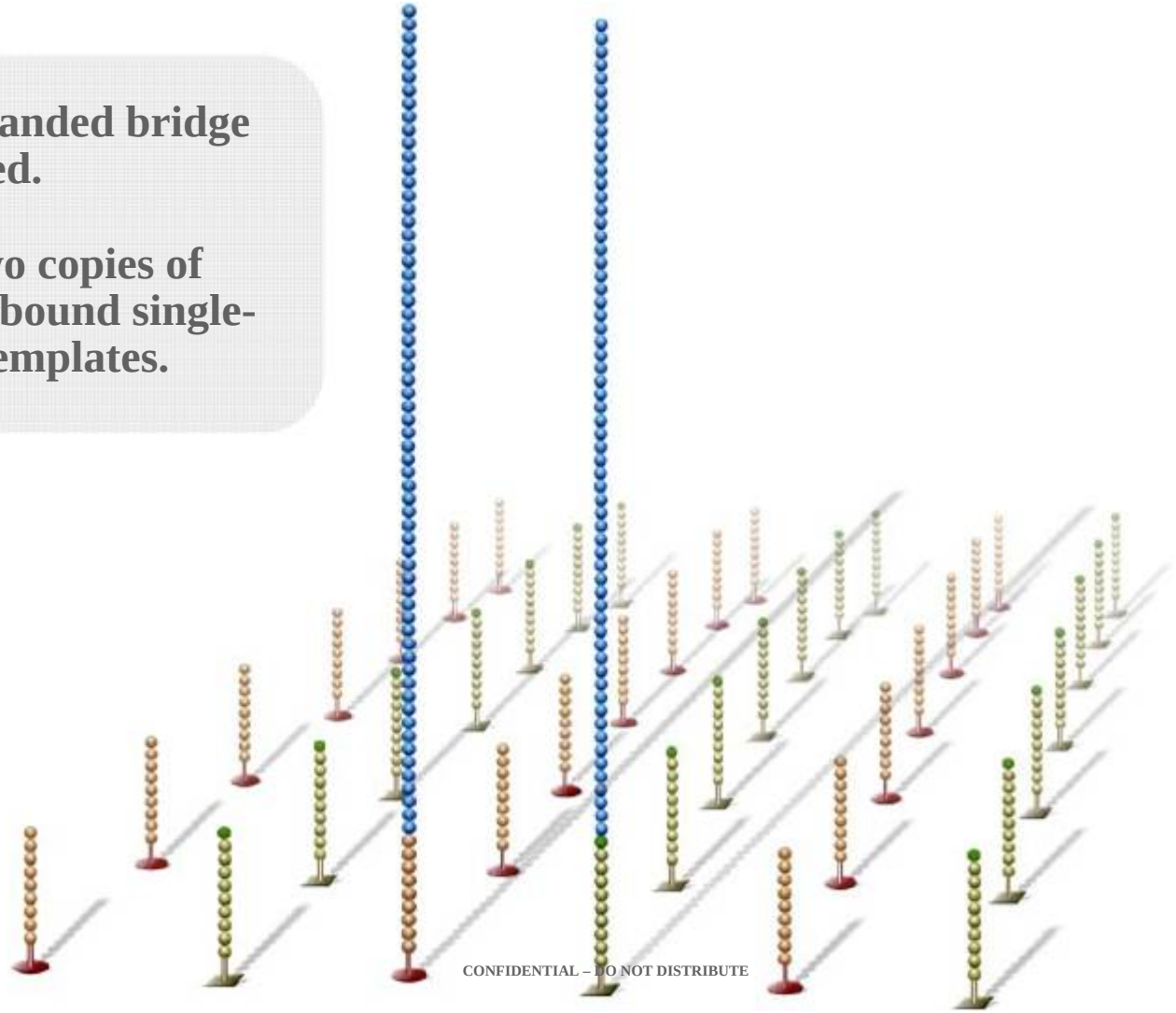
double-stranded bridge is formed.



Cluster generation: Bridge amplification

Double-stranded bridge is denatured.

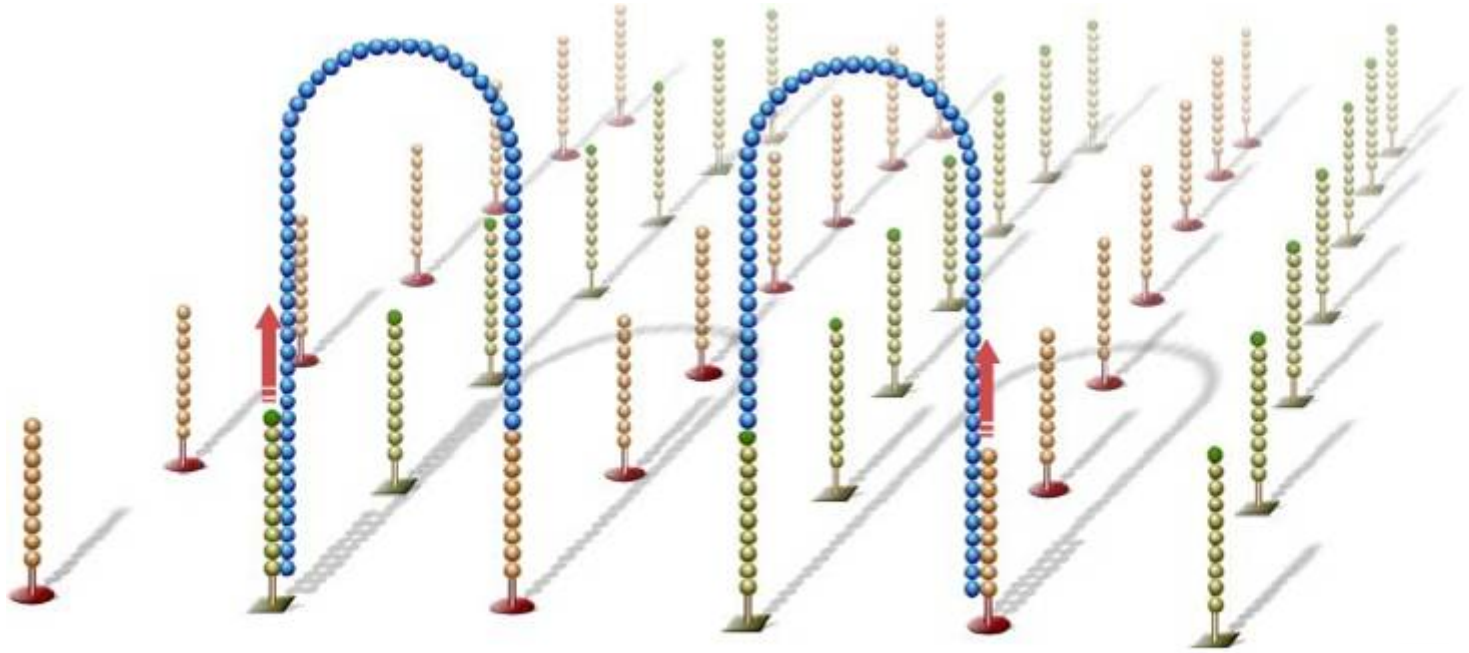
Result: Two copies of covalently bound single-stranded templates.



Cluster generation: Bridge amplification

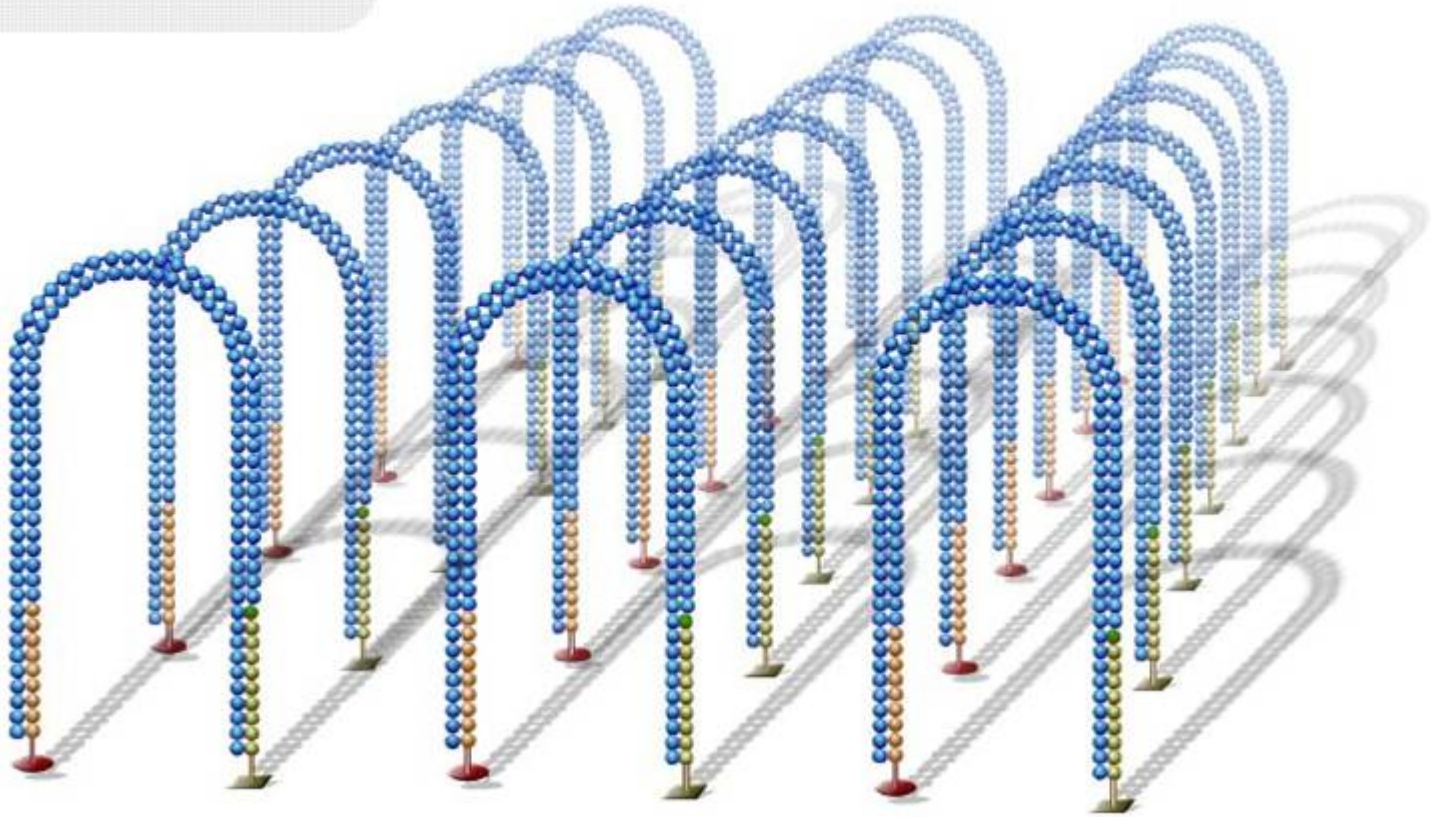
Single-strands flip over to hybridize to adjacent primers to form bridges.

Hybridized primer is extended by polymerase.



Cluster generation: Bridge amplification

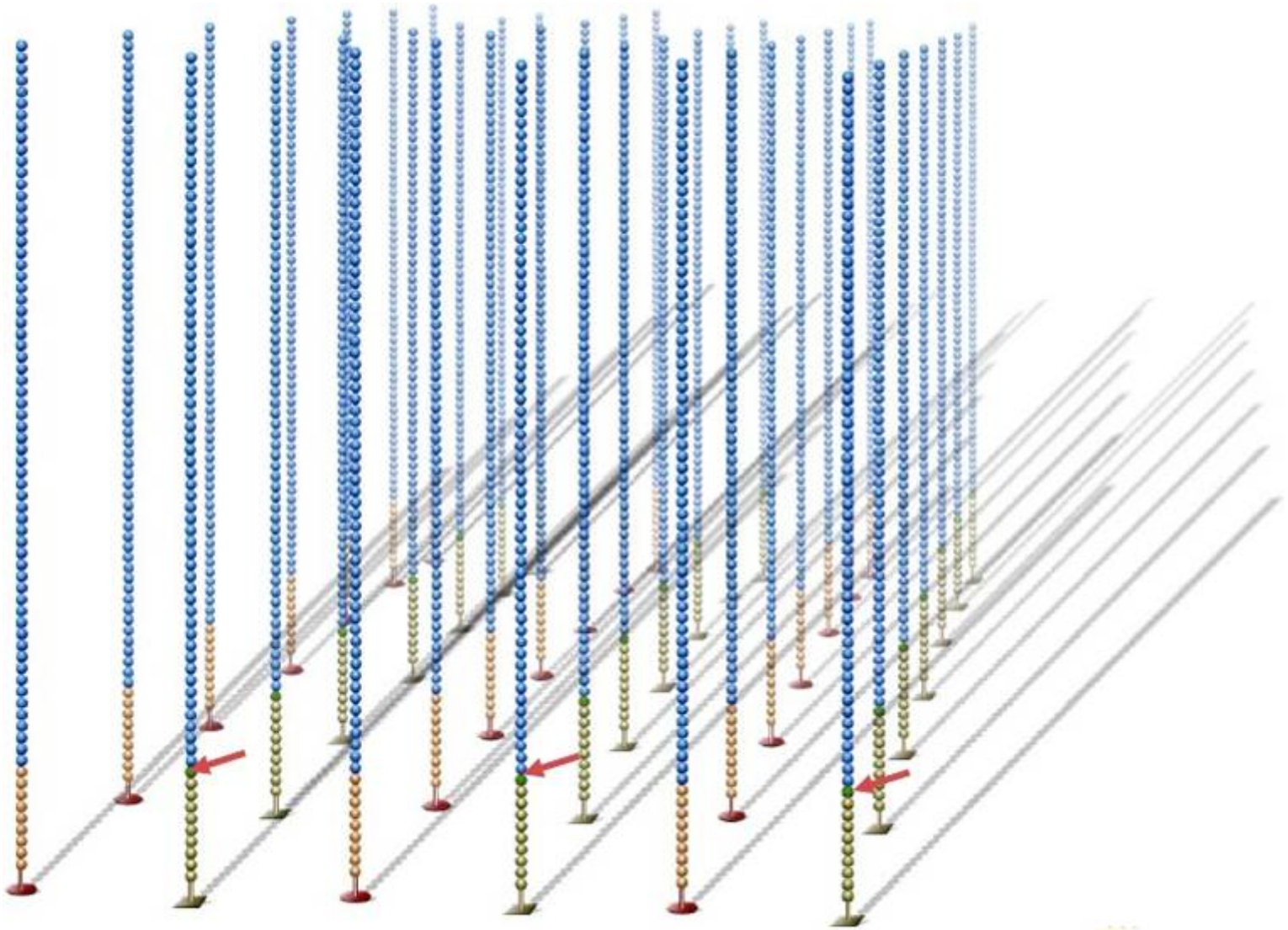
Bridge amplification cycle repeated till multiple bridges are formed



Cluster generation

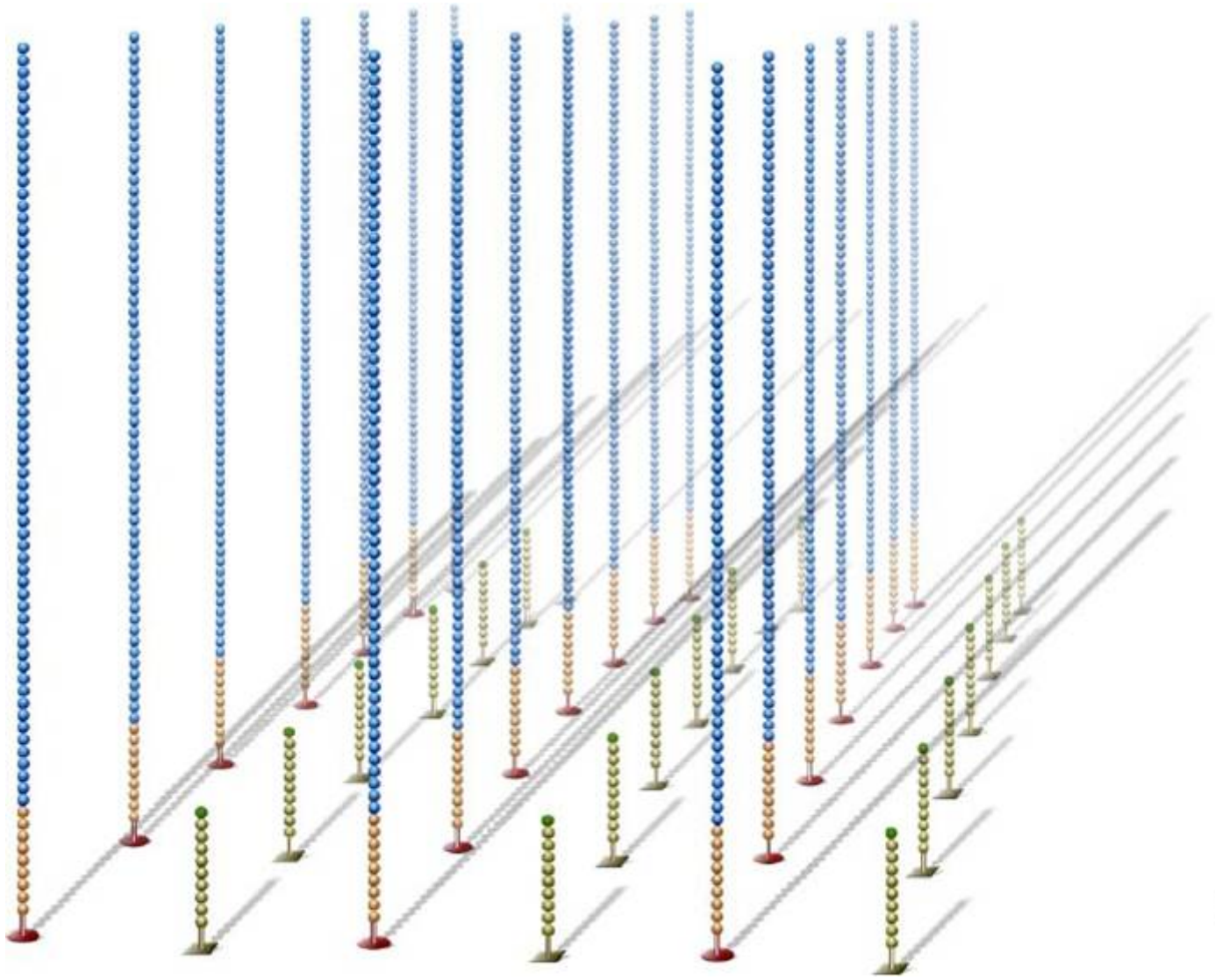
dsDNA bridges denatured.

Reverse strands cleaved and washed away.



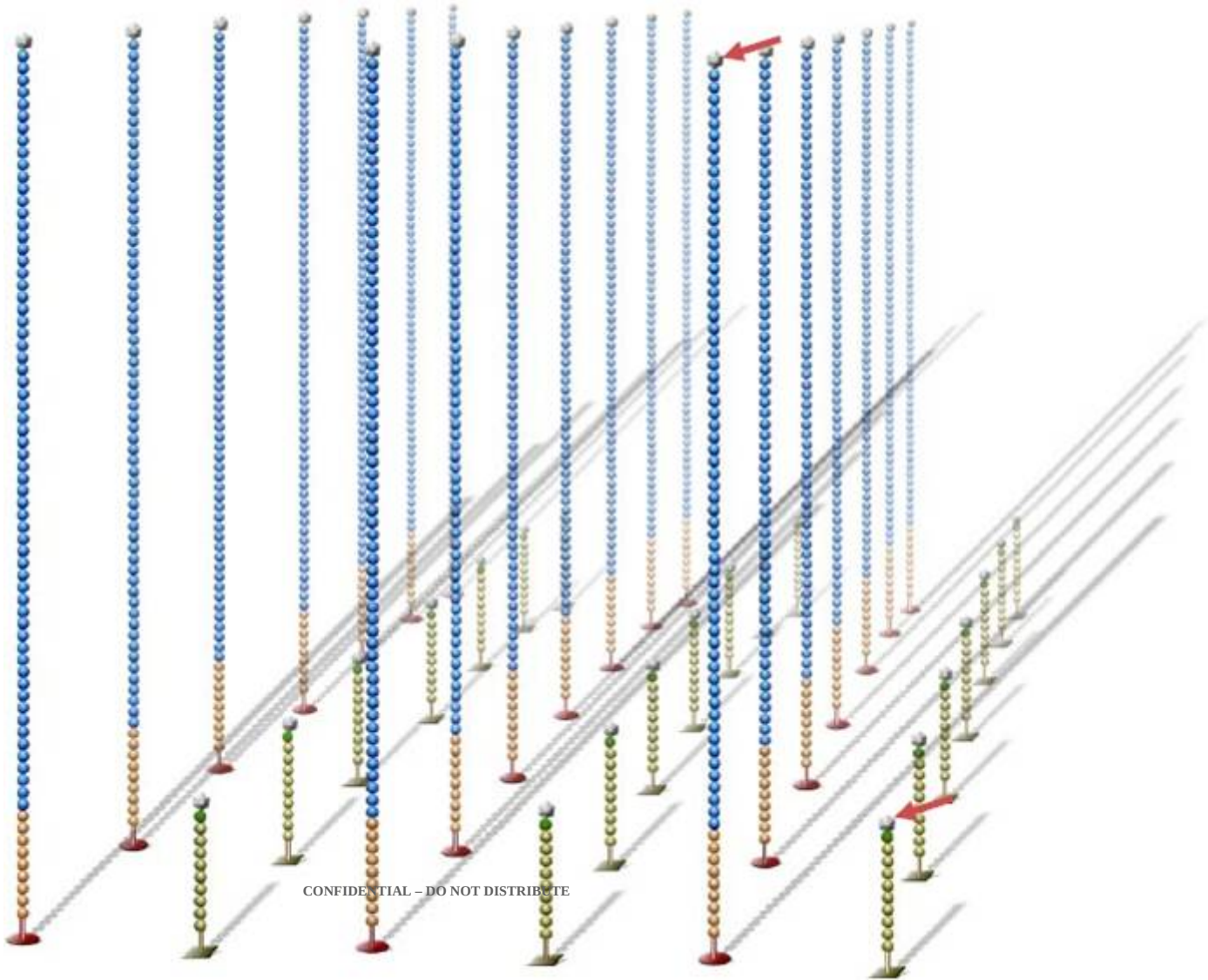
Cluster generation

… leaving
a cluster
with forward
strands only.



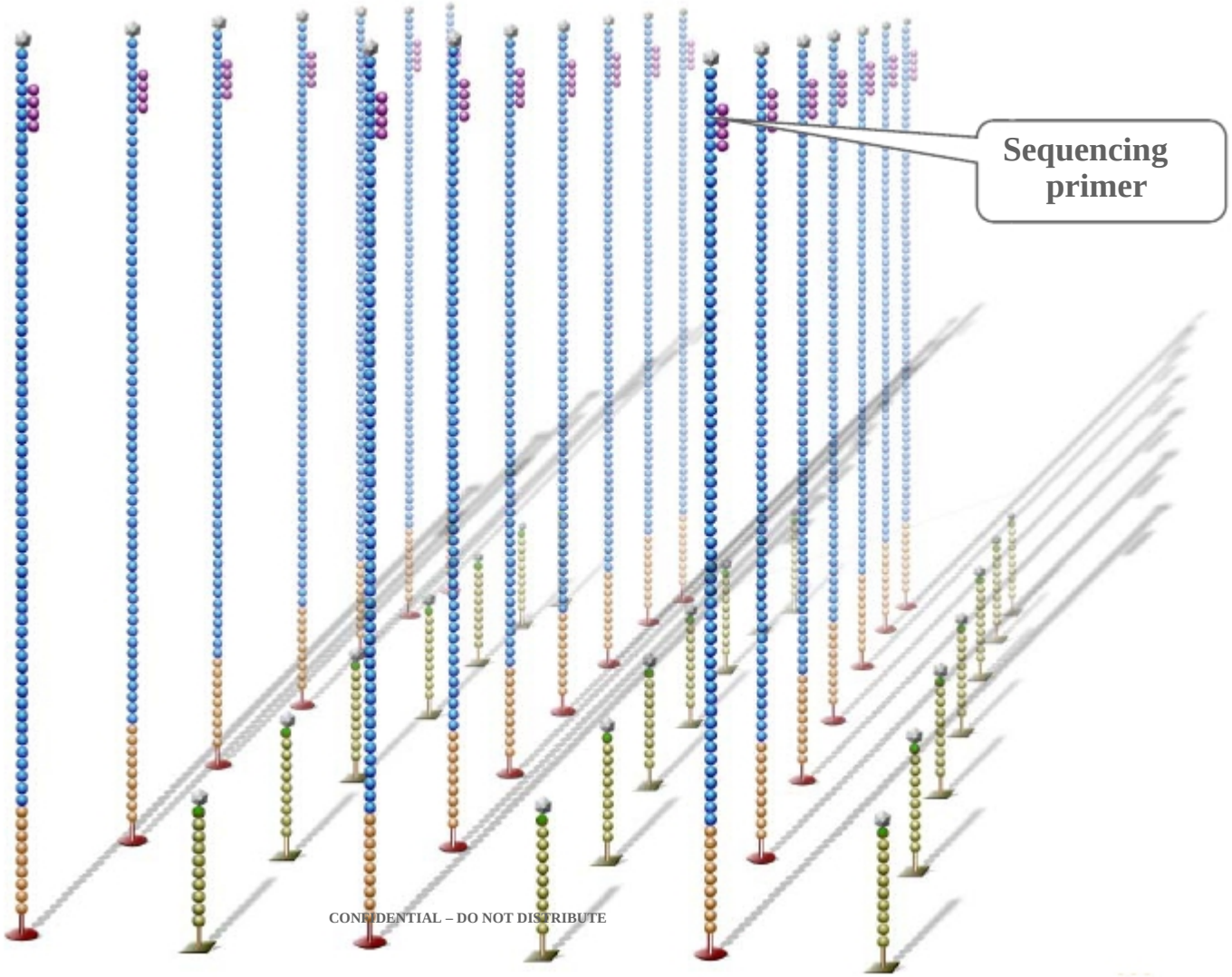
Cluster generation

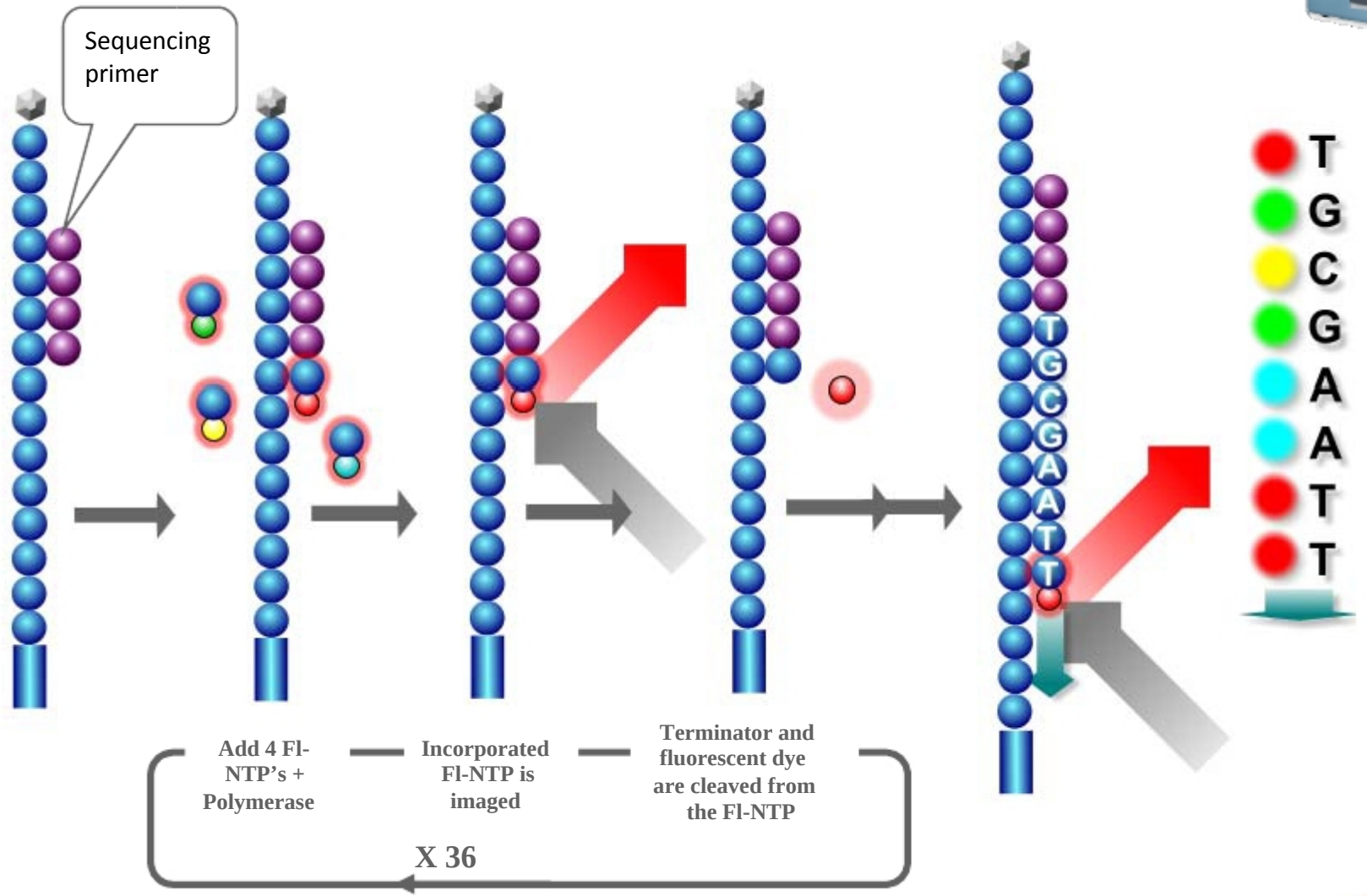
Free 3’ ends are blocked to prevent unwanted DNA priming.



Sequencing

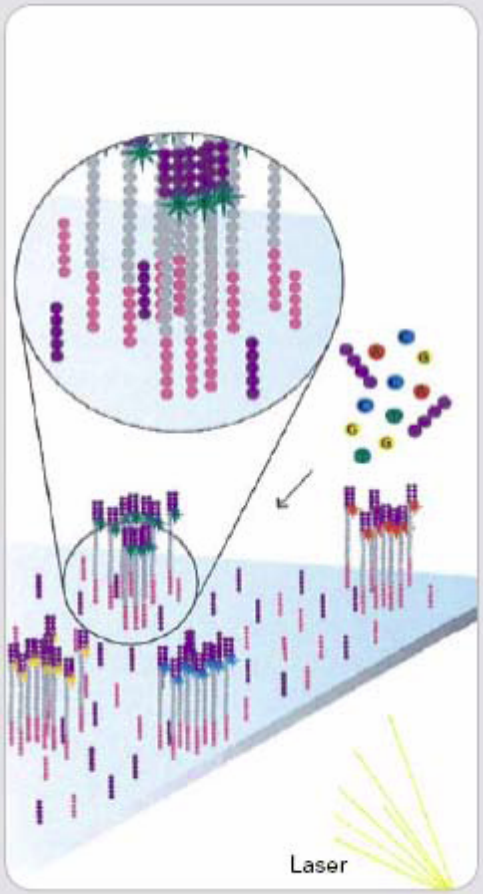
Sequencing primer is hybridized to adapter sequence.



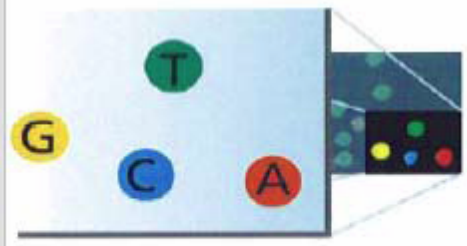

Sequencing primer
T
G
C
G
A
A
T
T
Add 4 Fl-NTP's + Polymerase
Incorporated Fl-NTP is imaged
Terminator and fluorescent dye are cleaved from the Fl-NTP
X 36

Sequencing reaction and imaging

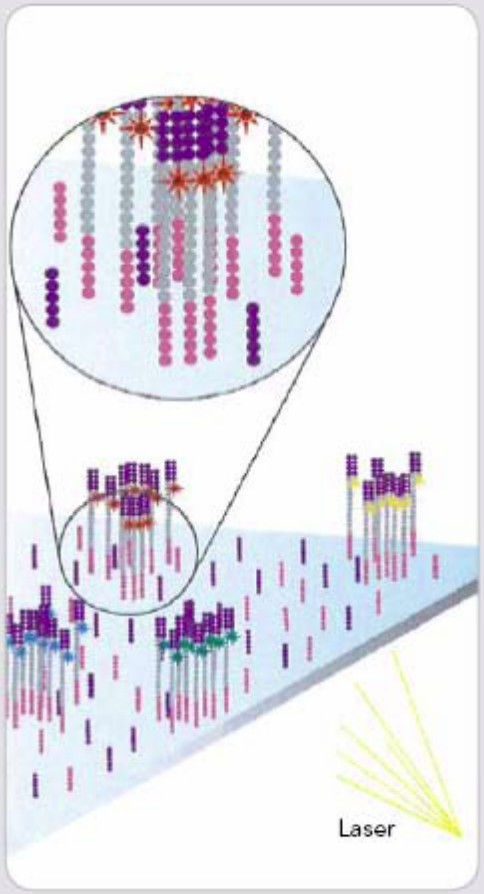
7. DETERMINE FIRST BASE



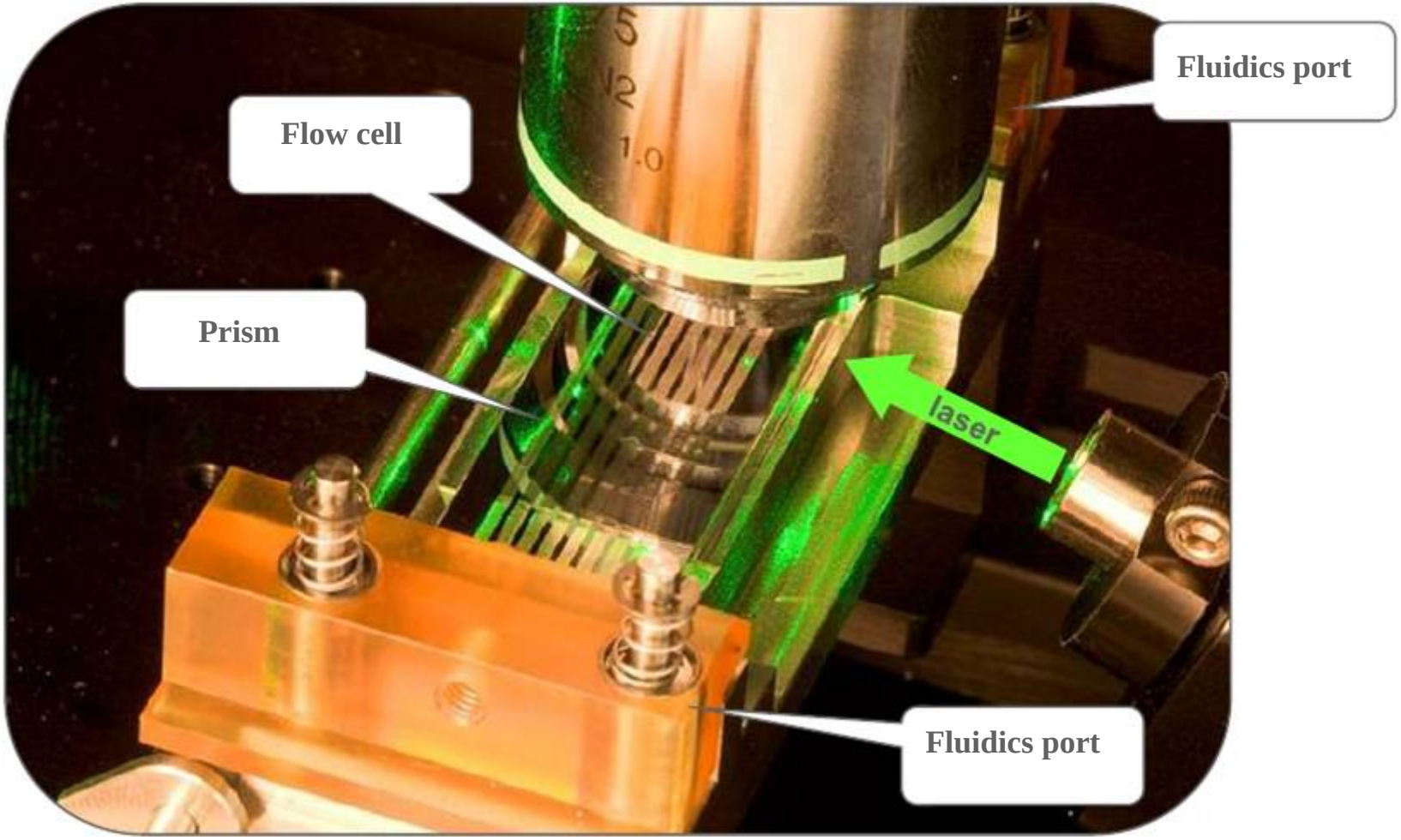
8. IMAGE FIRST BASE



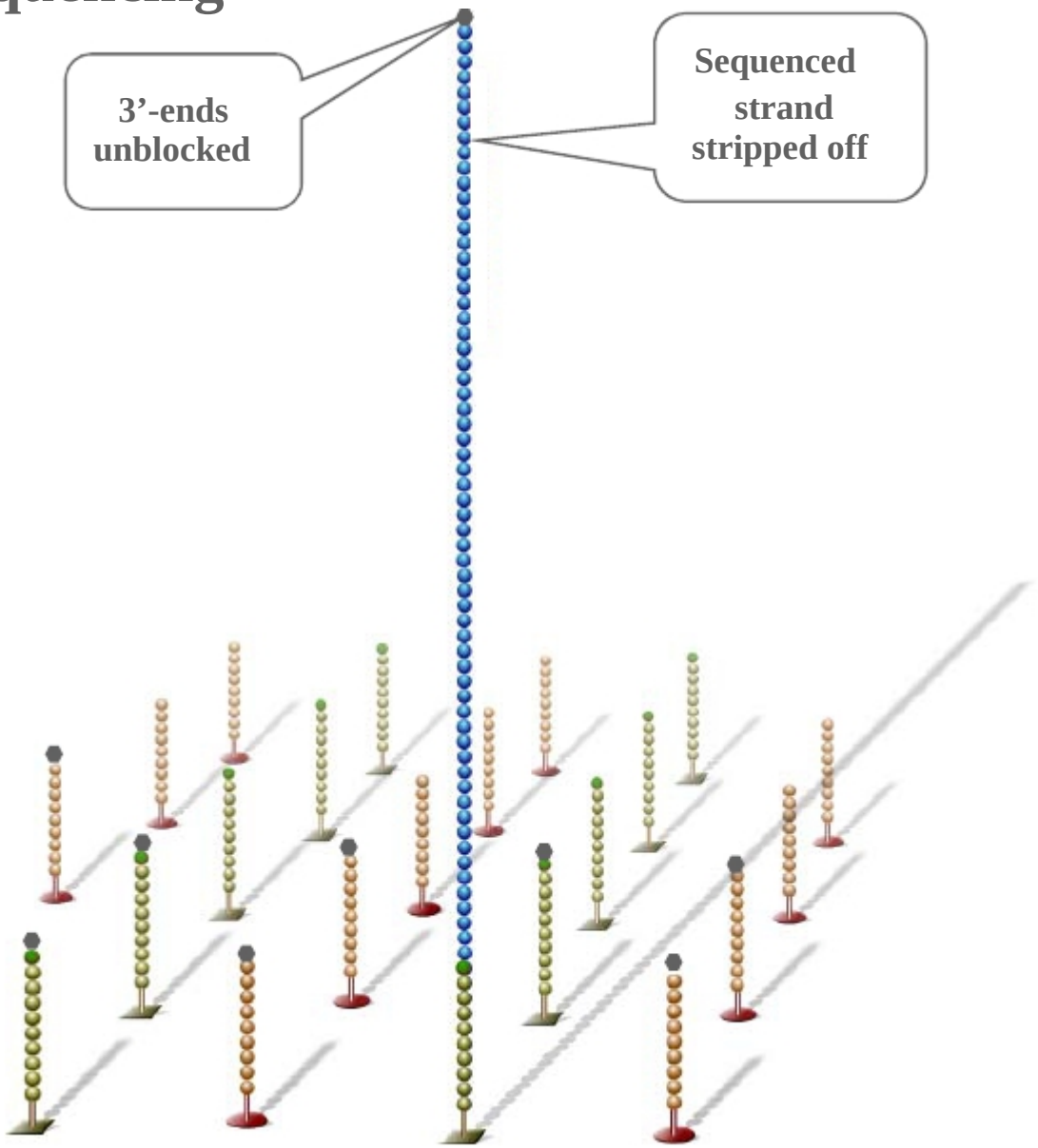
9. DETERMINE SECOND BASE



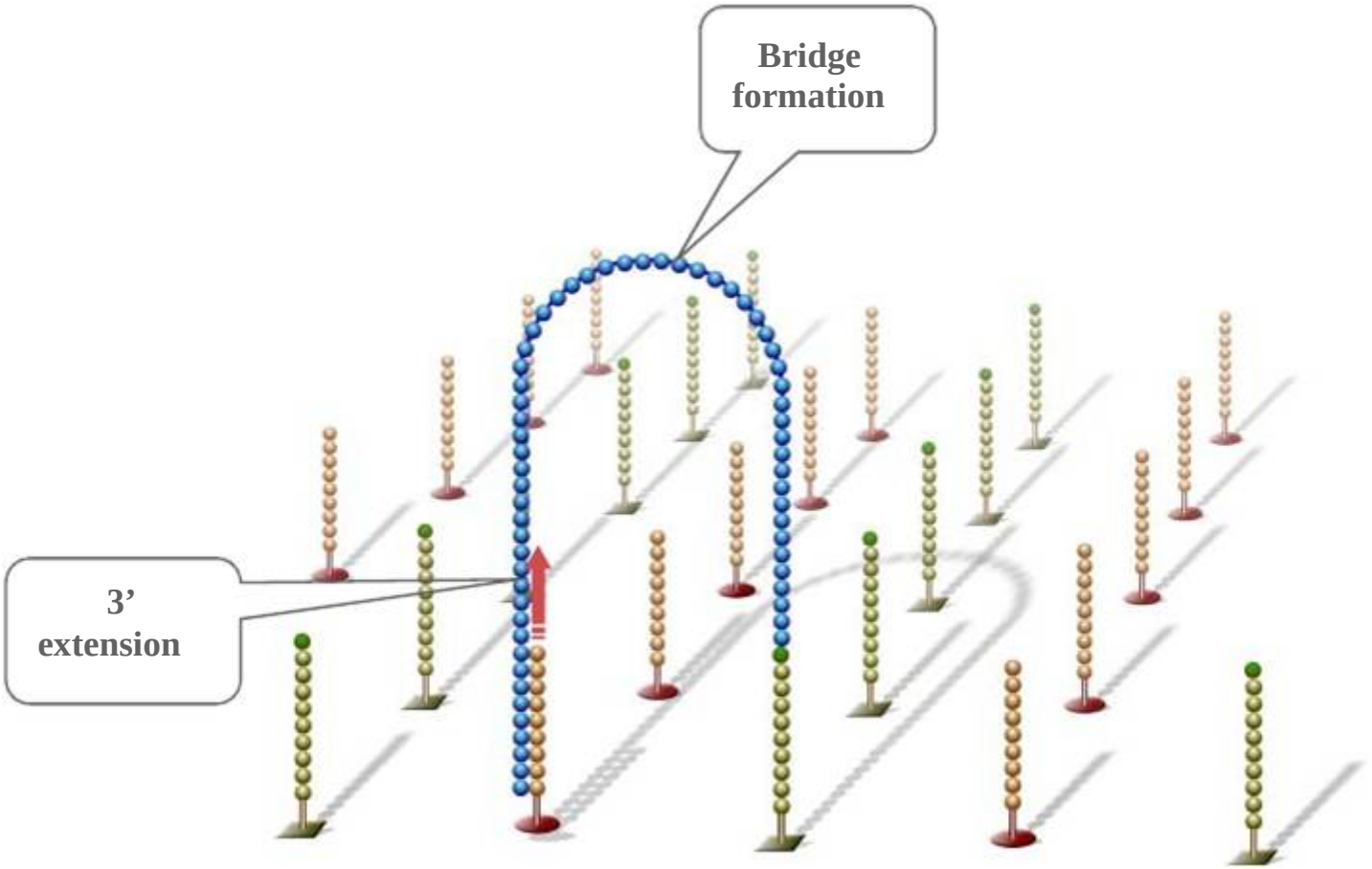
Flow cell imaging
Total Internal Reflection Fluorescence



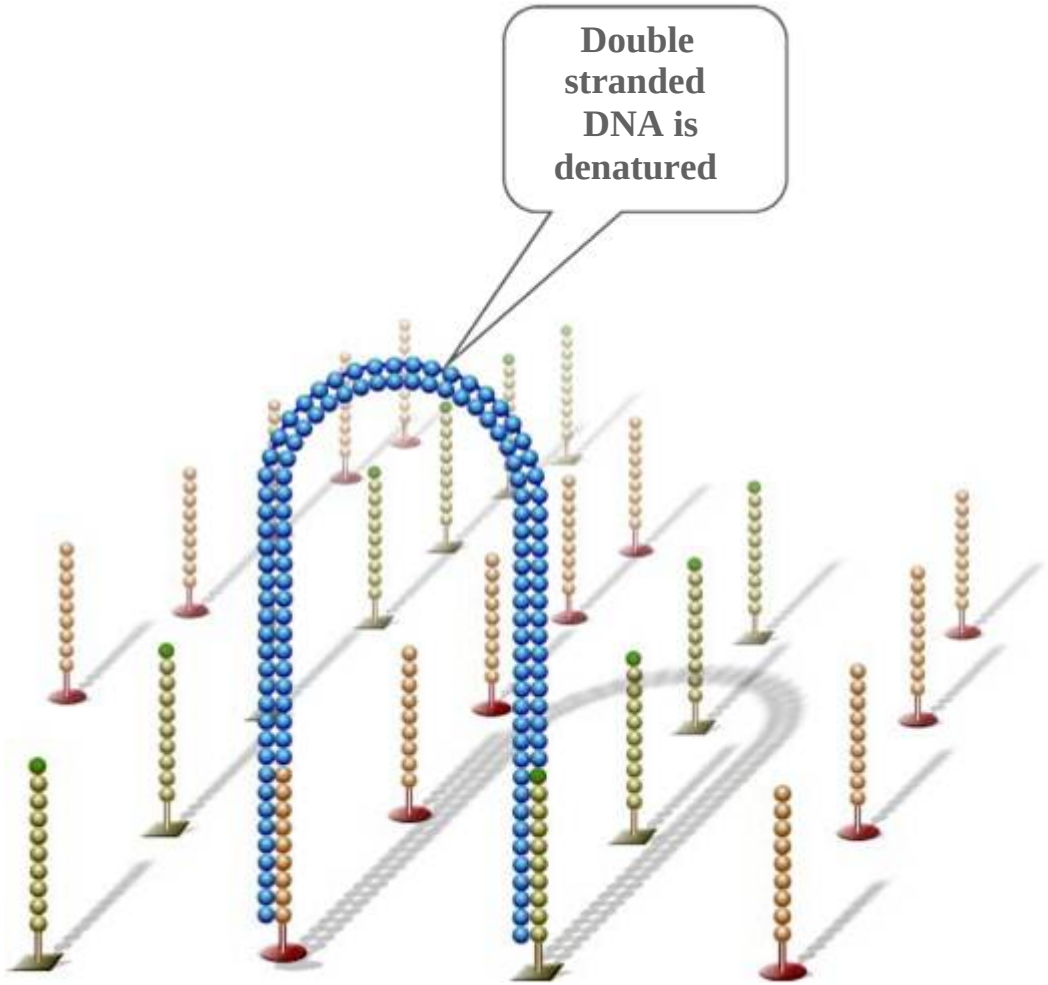
Paired end sequencing



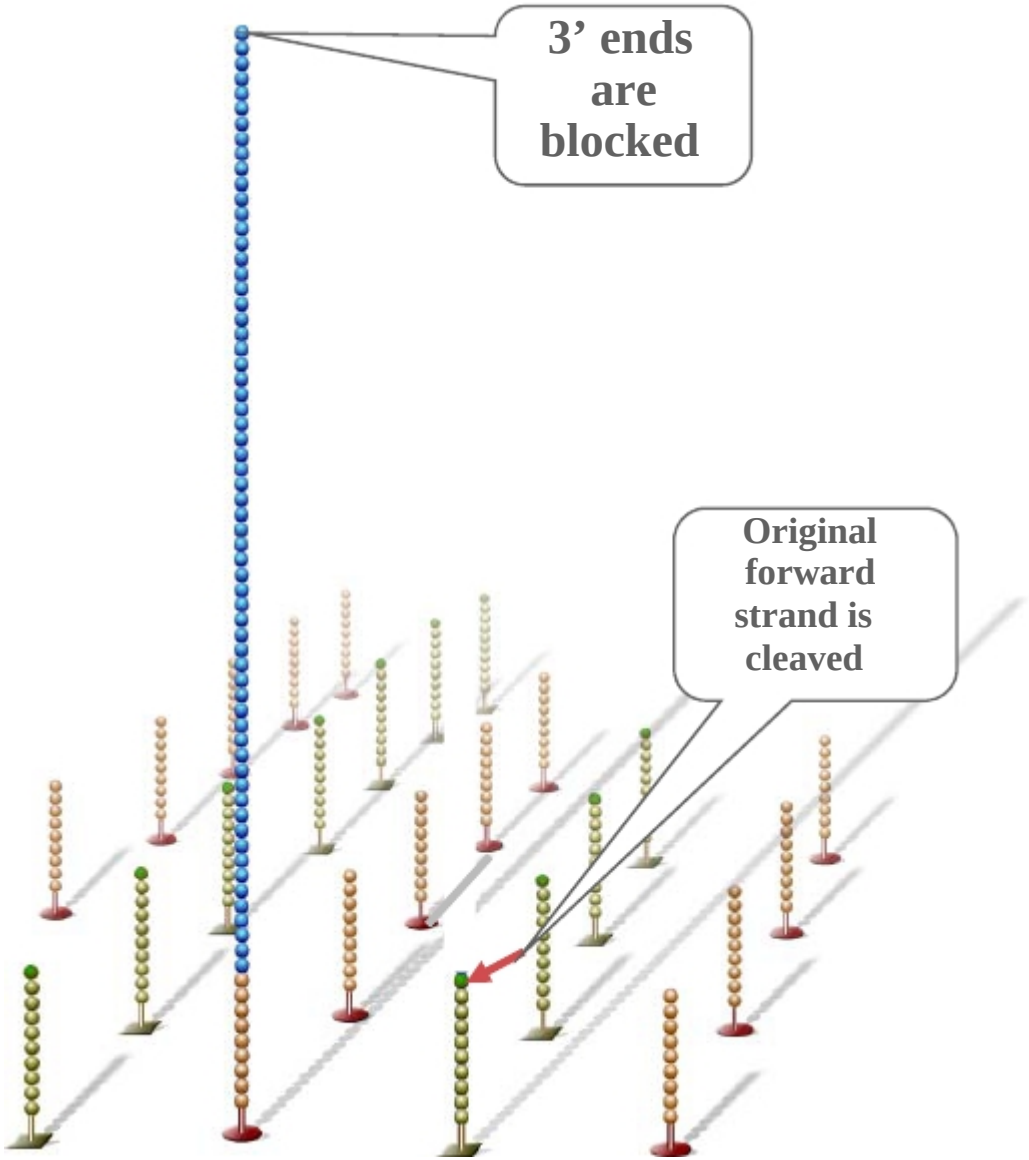
Paired end sequencing



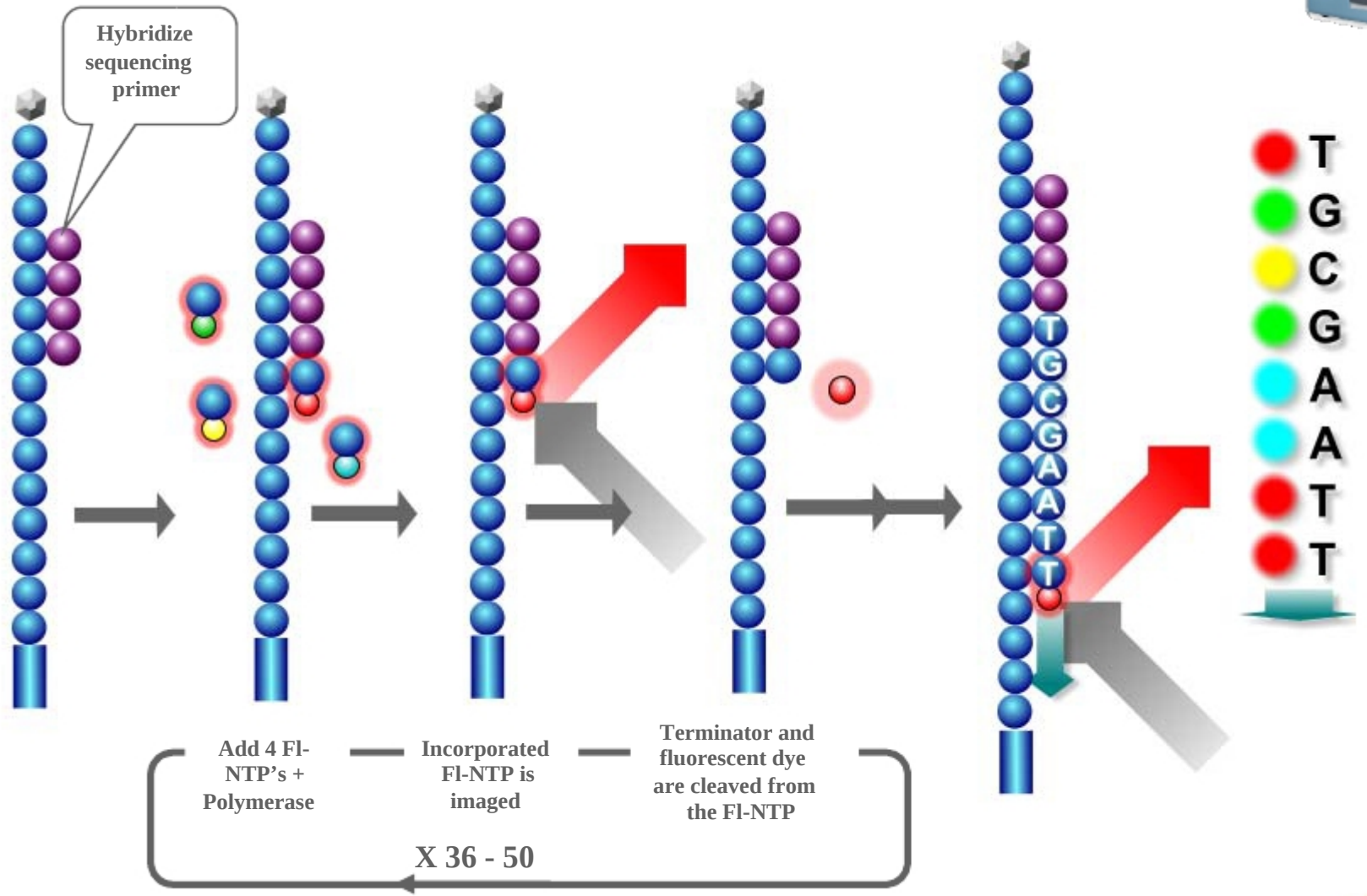
Paired end sequencing



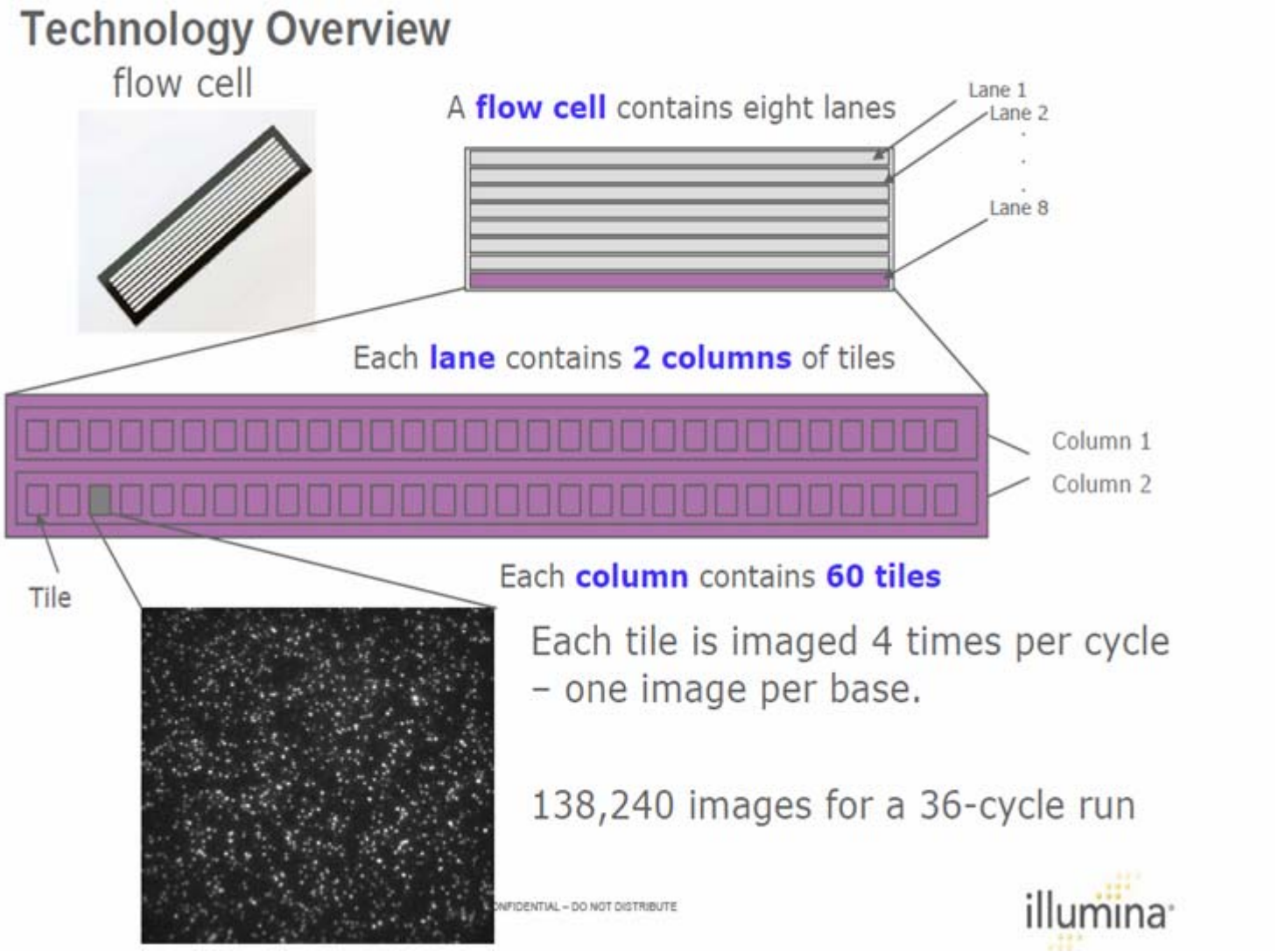
Paired end sequencing



Sequencing reverse strand



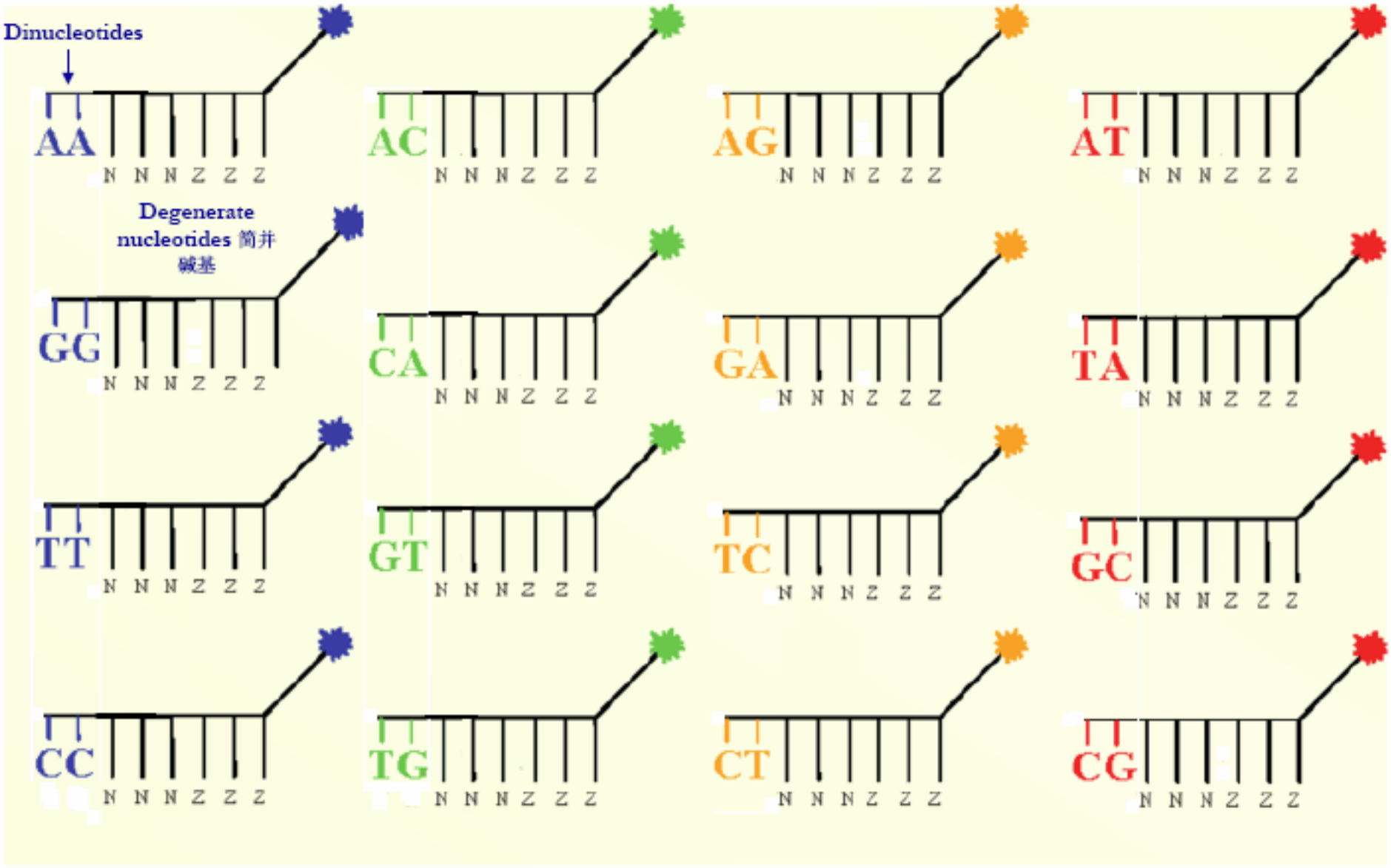
Flow cell in GAIIx



36cycle * 4time * 60tile * 2column * 8lane = 138240images

Fluorescent Oligo Octamer Probes

4组荧光双碱基编码寡核苷酸探针

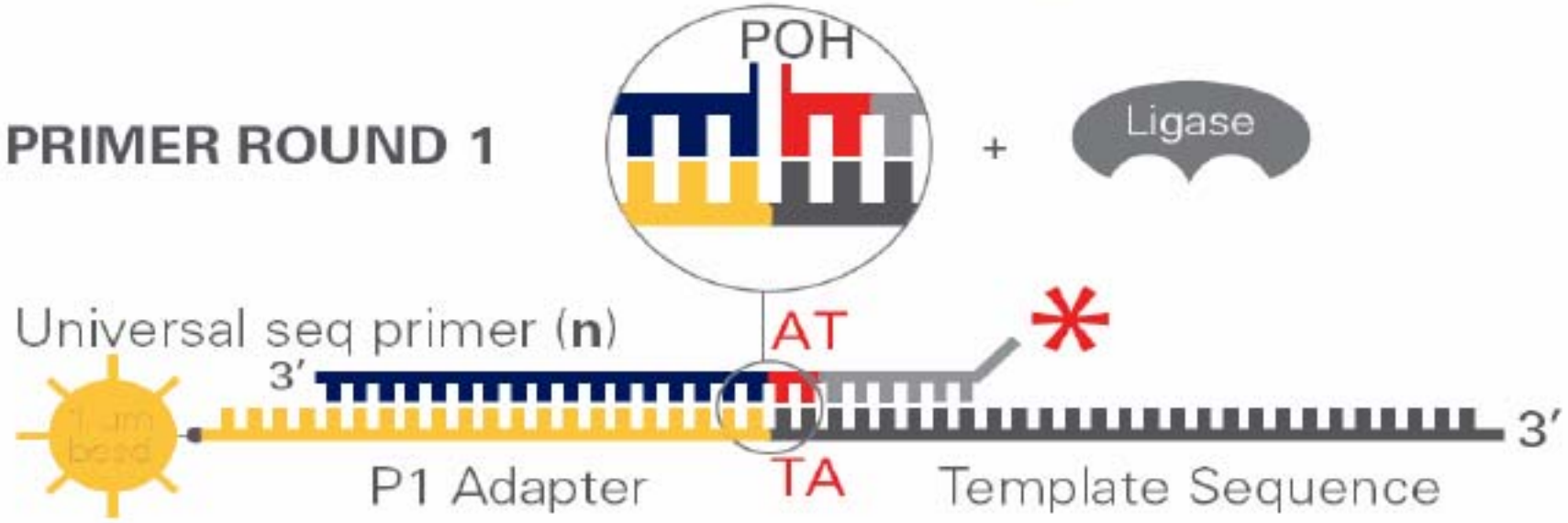


Sequence by Ligation

连接寡核苷酸探针

1. Prime and Ligate

Hybridization and ligation of a specific oligo whose 1st & 2nd bases match that of the template

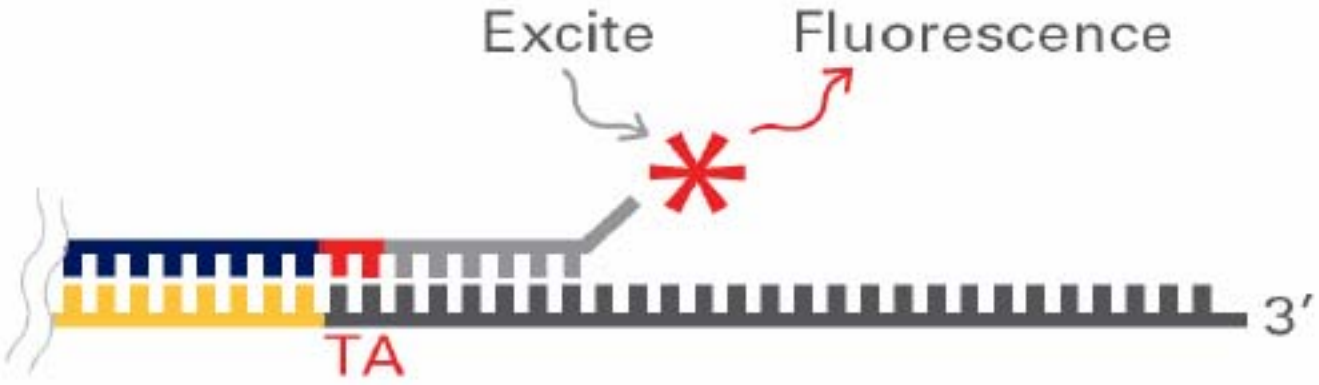


Sequence by Ligation

Detection of the specific fluorescence

成像

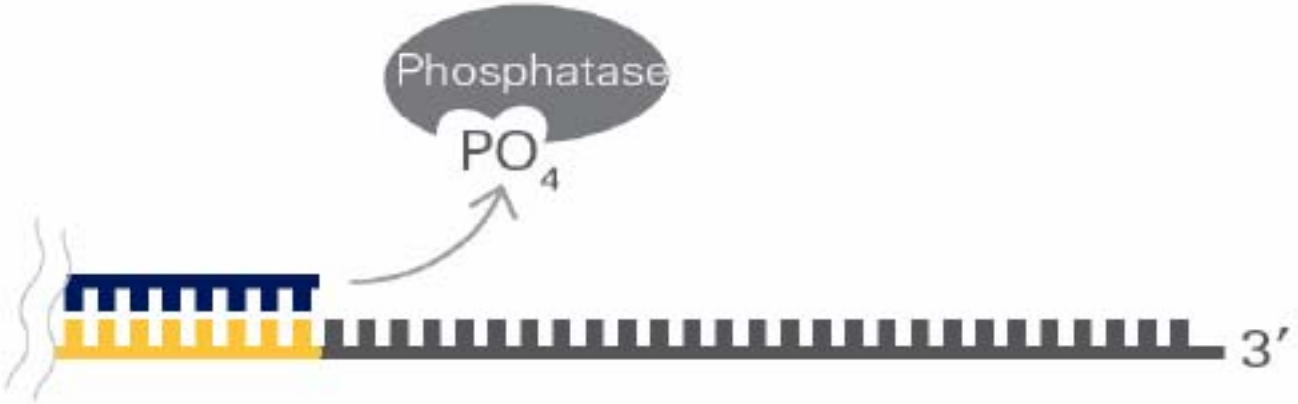
2. Image



Sequence by Ligation

保护未连接链

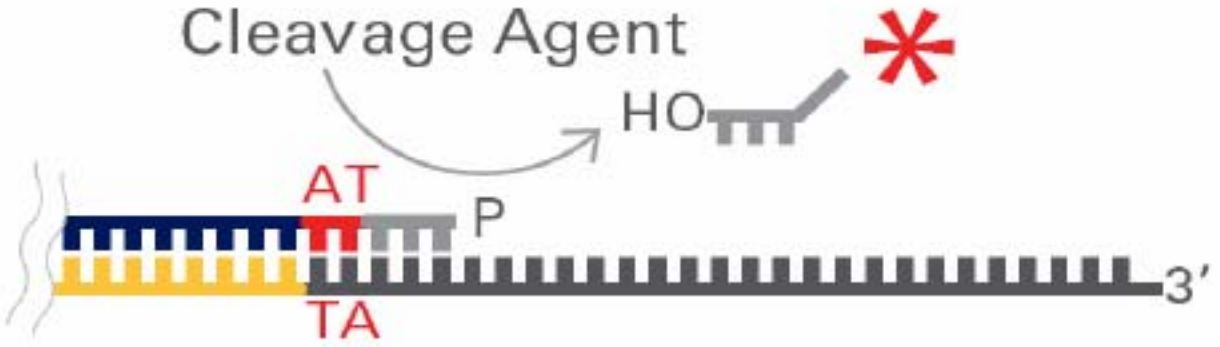
3. Cap Unextended Strands



Sequence by Ligation

去除荧光标记

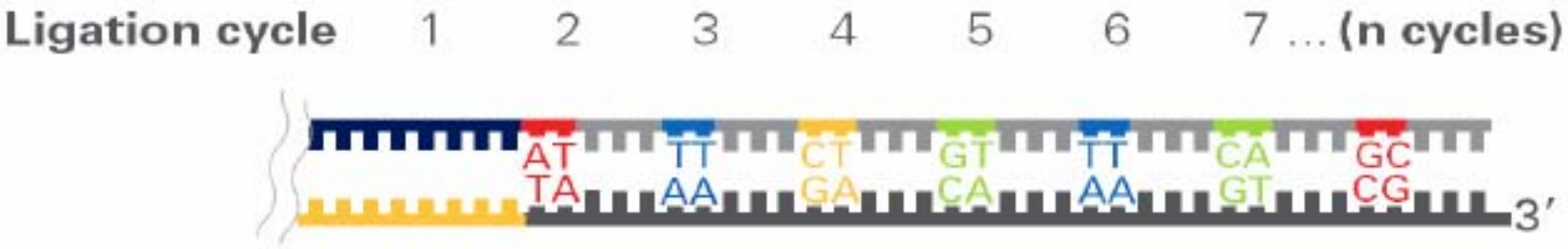
4. Cleave off Fluor Cleavage of all bases to the 5' of base 5



Sequence by Ligation

重复连接反应七次

5. Repeat steps 1-4 to Extend Sequence

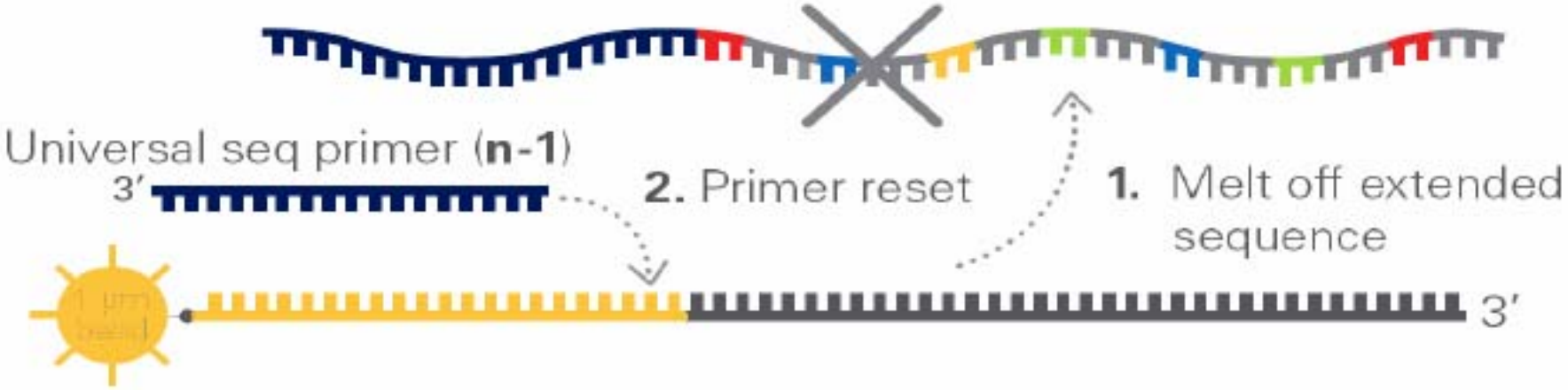


Sequence by Ligation

重启引物

6. Primer Reset

Primer and all ligated portions are melted from the template and discarded

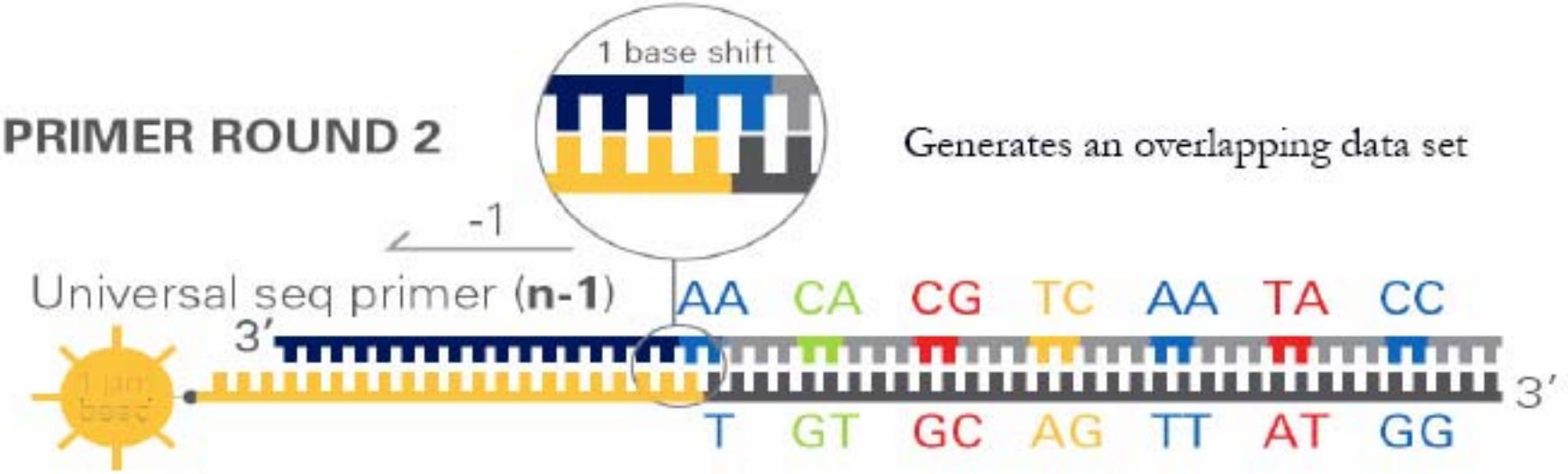


New initial primer is used that is N-1 in length

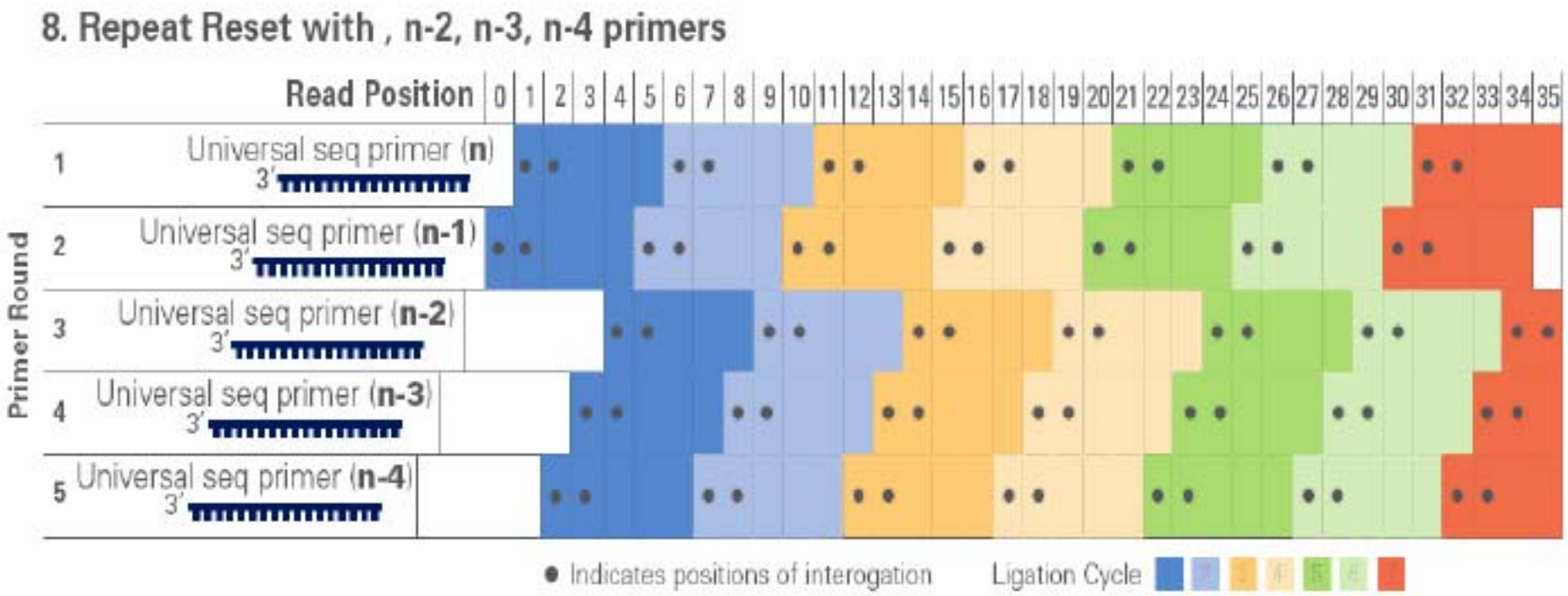
Sequence by Ligation

重启后循环连接反应

7. Repeat steps 1-5 with new primer

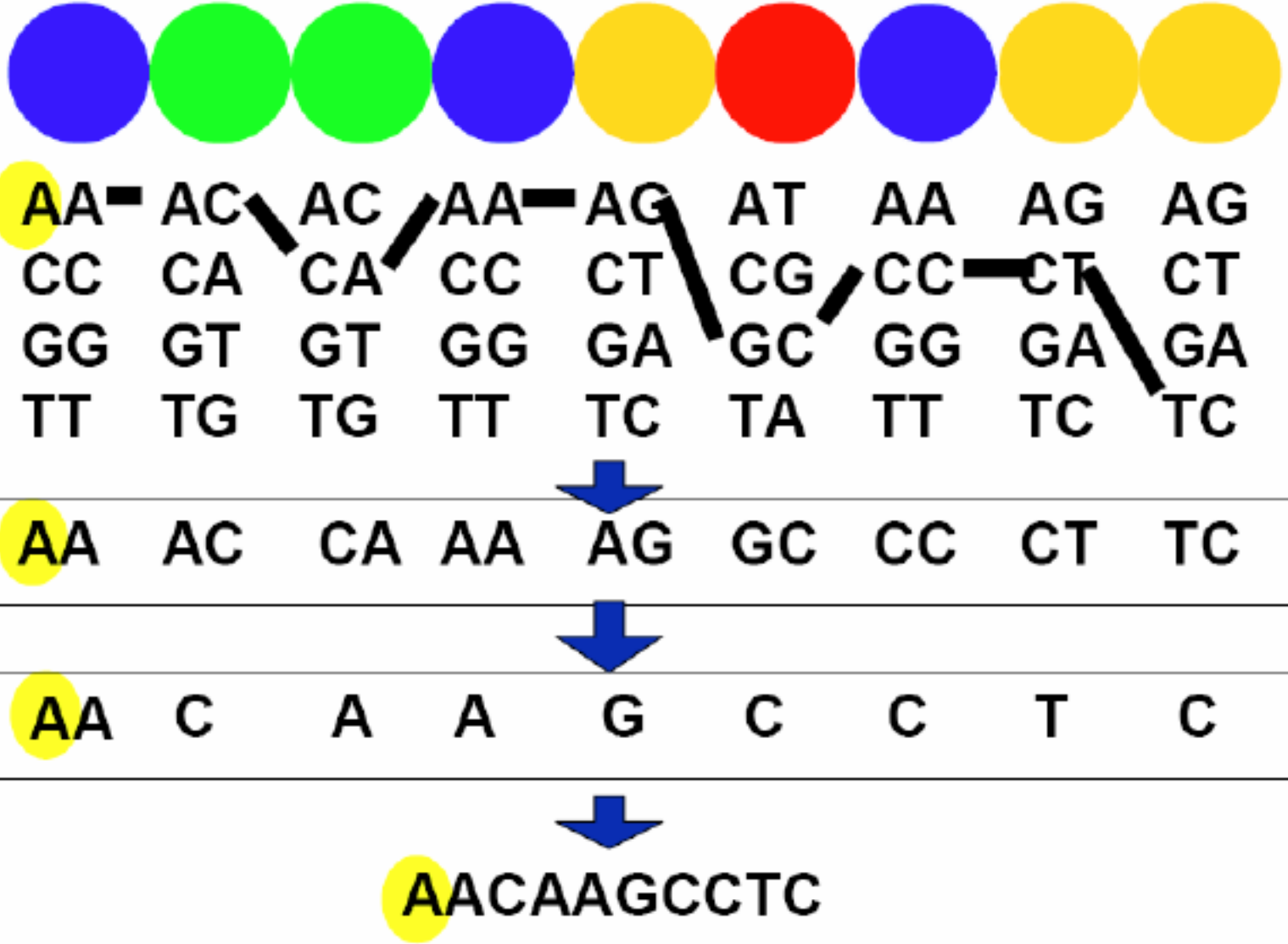


Full Sequence Coverage
模板片段覆盖完毕



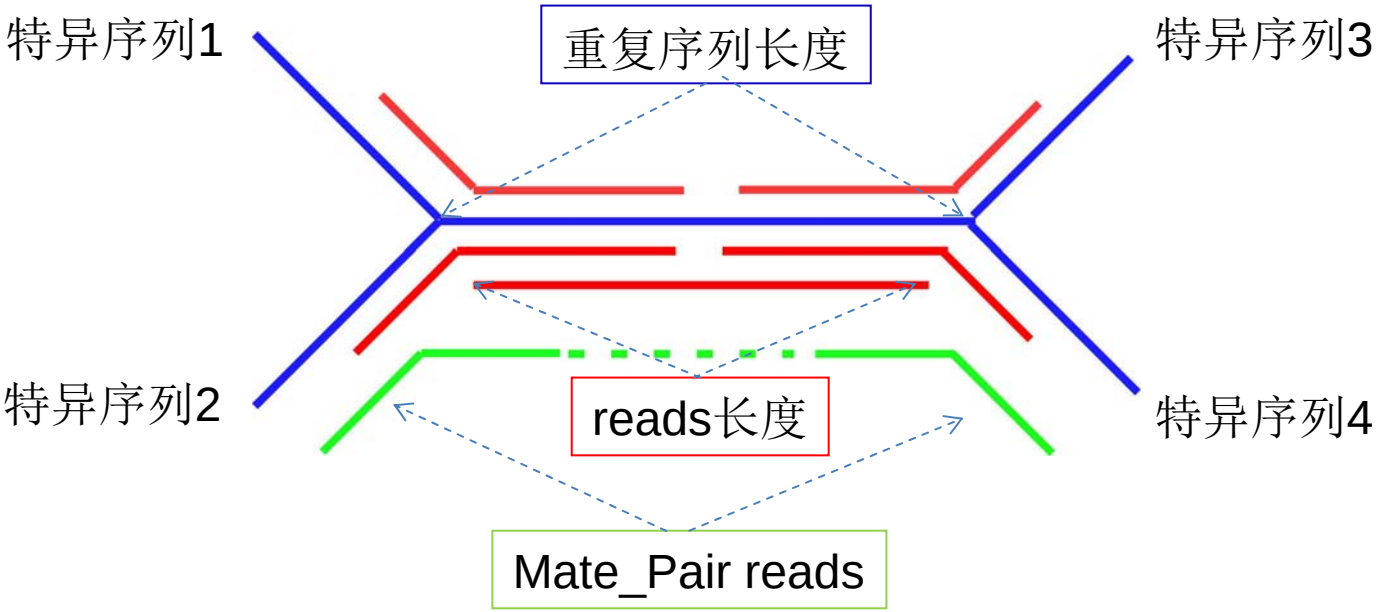
Di-Base Encoding Color Space

双碱基编码颜色空间解读

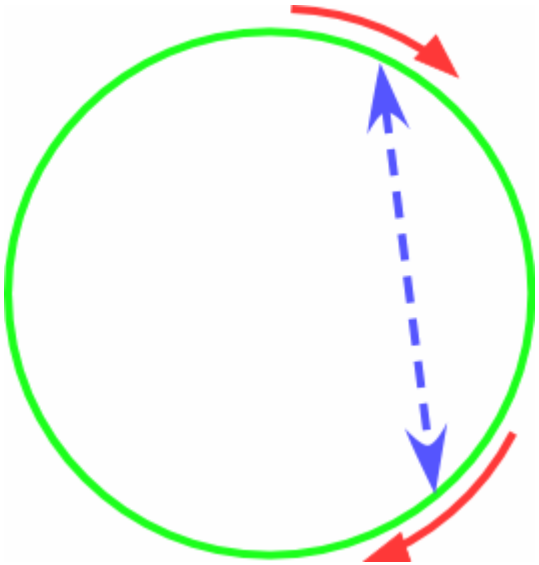


Mate_Pair的意义

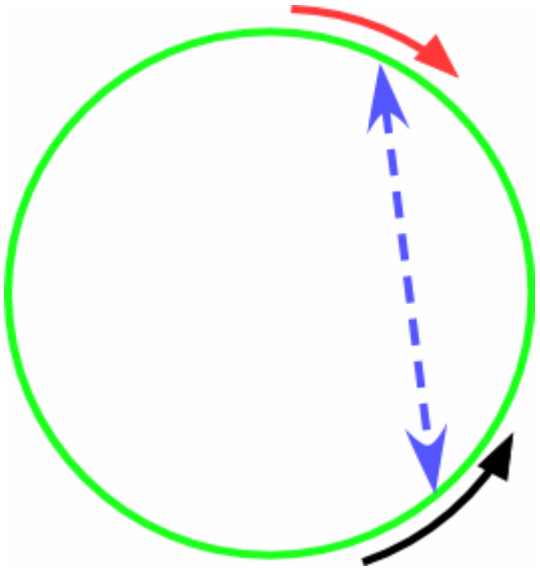
- 重复序列长度超过read的长度，拼接将产生分支，无法延续，从而形成断点
- 解决方法：
  - ✓ 选择长read的测序仪器
  - ✓ 构建大片段Mate_Pair样品库进行测序



正向重复

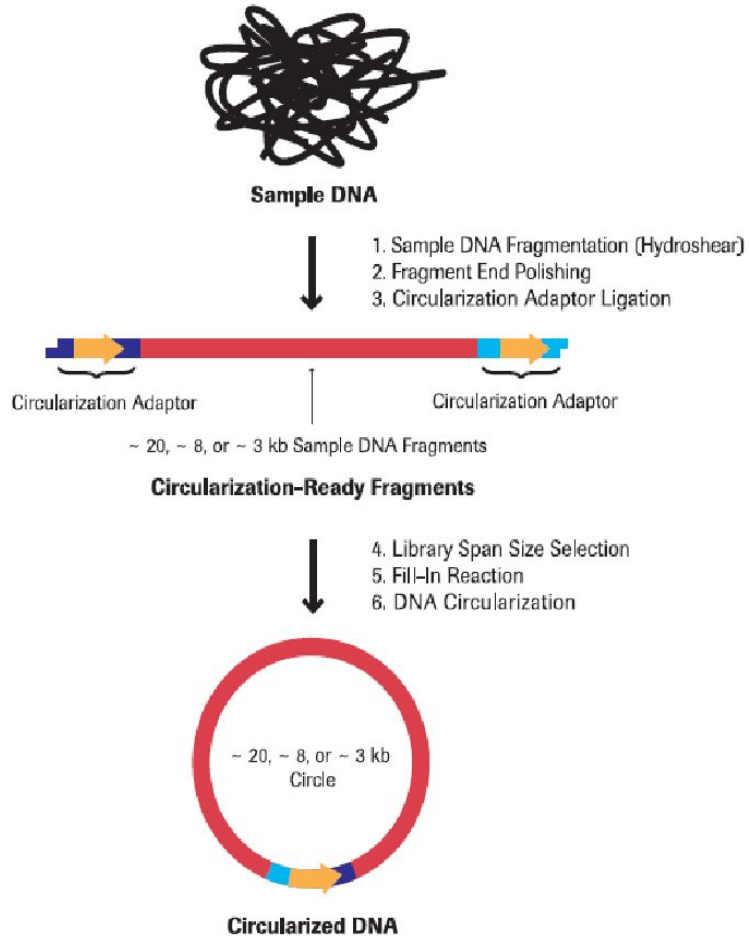


反向重复



- 正向重复导致环化脱落
- 反向重复导致序列翻转

GS FLX Titanium Long Span Paired End Overview



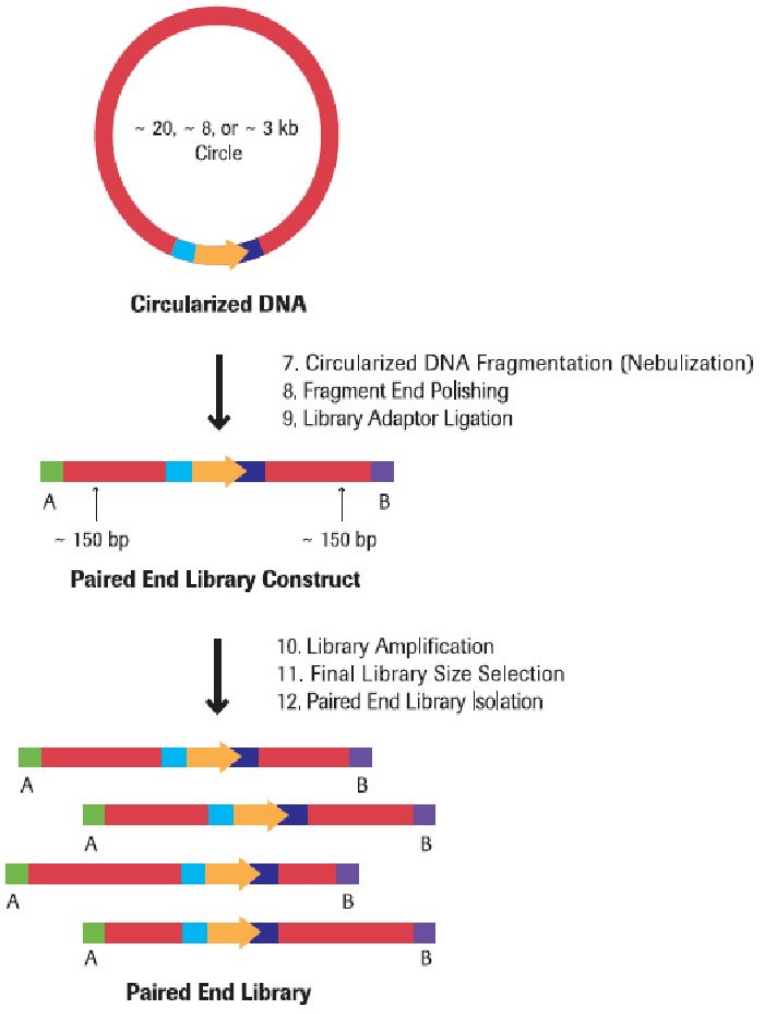
High molecular weight genomic DNA is sheared to desired size of either 20kb, 8kb, or 3 kb span distance

Circularization adaptors containing a loxP target sequence are ligated onto fragment ends.

Cre-mediated recombination circularizes fragments

n

GS FLX Titanium Long Span Paired End Overview

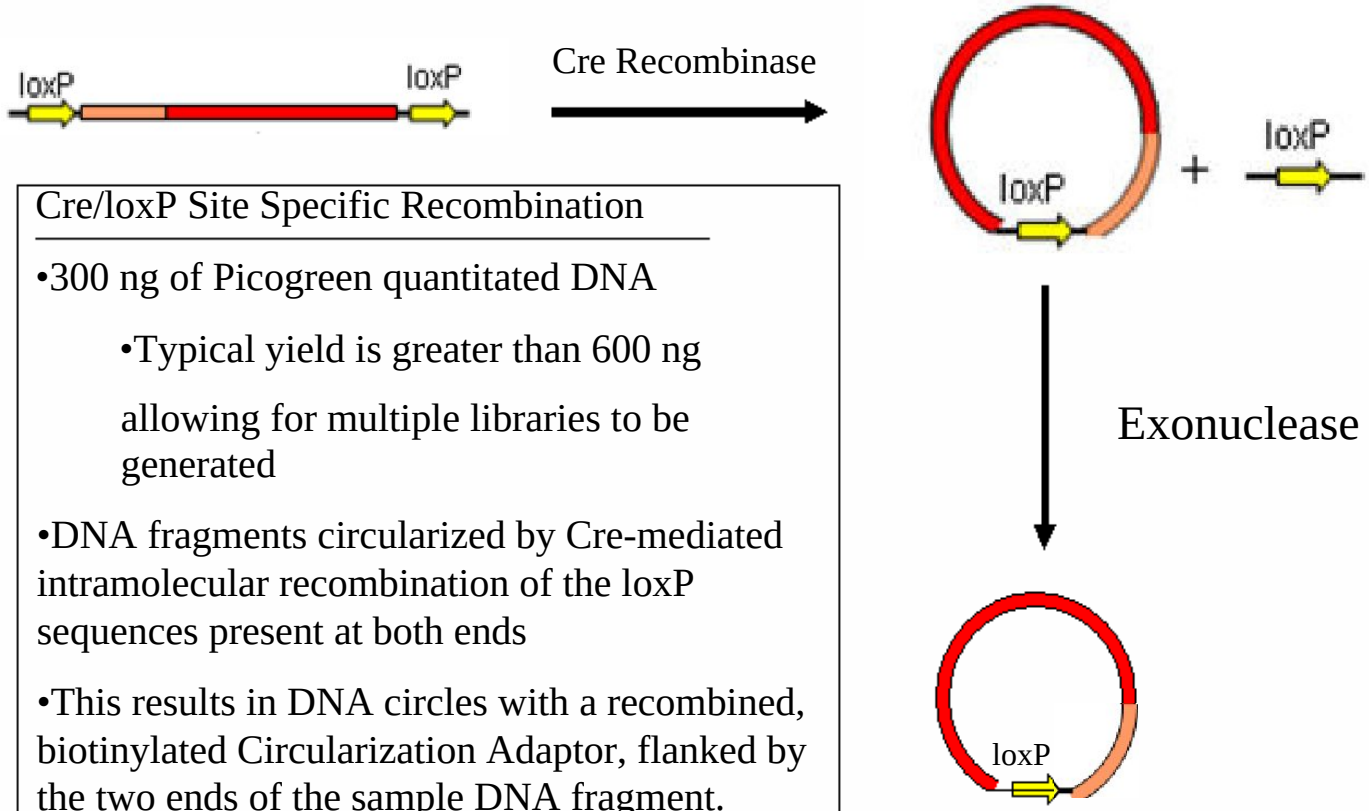


- Ligation of 454 sequencing adaptors.
- Adaptors required for emPCR and sequencing

Amplification and sequencing with the GS FLX Titanium System

n

GS FLX Titanium Long Span Paired End Workflow Day 3:



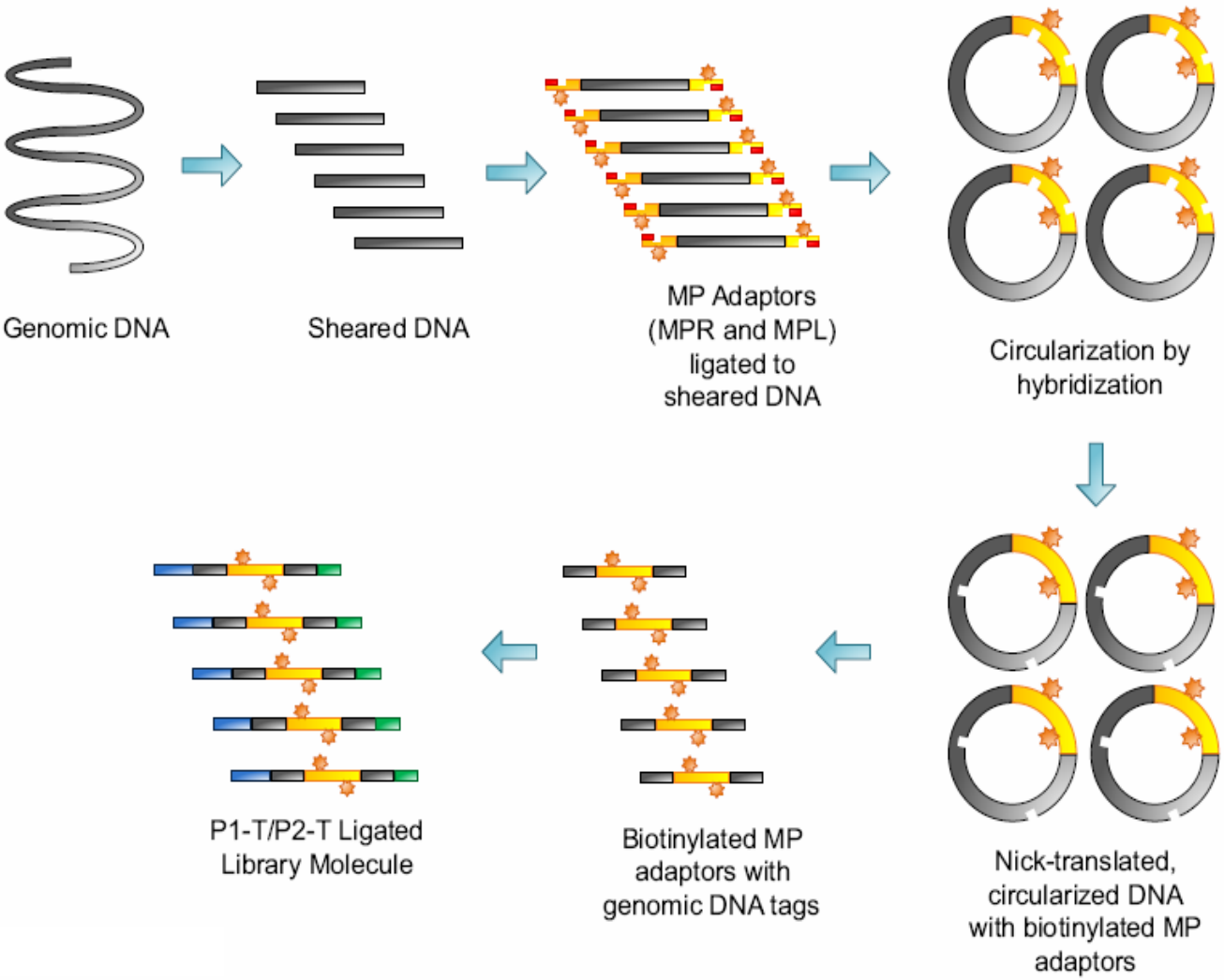
Cre/loxP Site Specific Recombination

- 300 ng of Picogreen quantitated DNA
  - Typical yield is greater than 600 ng

  allowing for multiple libraries to be generated
- DNA fragments circularized by Cre-mediated intramolecular recombination of the loxP sequences present at both ends
- This results in DNA circles with a recombined, biotinylated Circularization Adaptor, flanked by the two ends of the sample DNA fragment.
- Exonuclease digests excised loxP adaptor

www.roche-applied-science.com

Solid Mate-Pair Library



谢 谢 ！