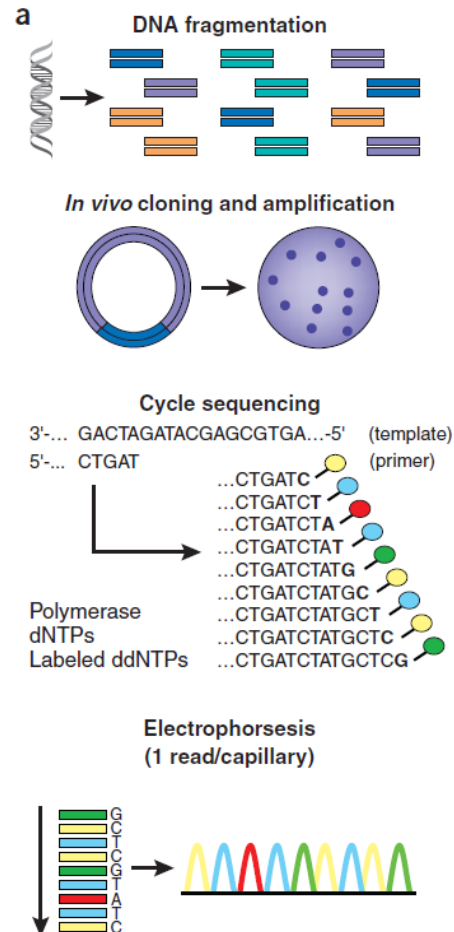


Next Generation Sequencing

- Sekwencjonowanie XXI
wielu

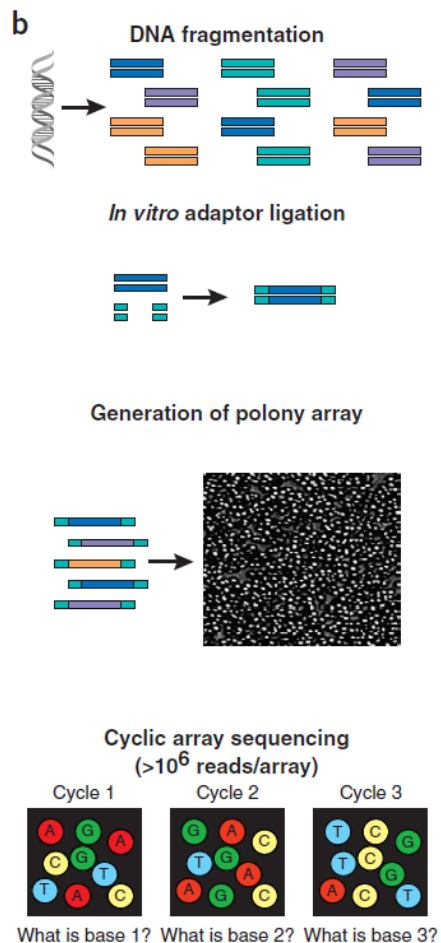
- Sekwencjonowanie pierwszej generacji (Sangera)
- NGS (obecnie)
 - Druga generacja metod sekwencjonowania
 - Sekwencjonowanie równoległe (równoczesne) wielu fragmentów
 - Ultra high-throughput sequencing (wysoka przepustowość)
- Co przyniesie przyszłość?
- Wyzwania związane z analizą danych NGS

Sekwencjonowanie pierwszej generacji (metoda Sanger)



- DNA ulega fragmentacji
- DNA jest wklinowywane do plazmidów i namnażane w kom. bakteryjnych
- Po izolacji plazmidy są sekwencjonowane w reakcji cyklicznego sekwencjonowania z użyciem ddNTP znakowanych fluorescencyjnie
- Sekwenator automatycznie czytuje wyniki na podstawie barwnika fluorescencyjnego

Metoda cyklicznych – macierzy (NGS)

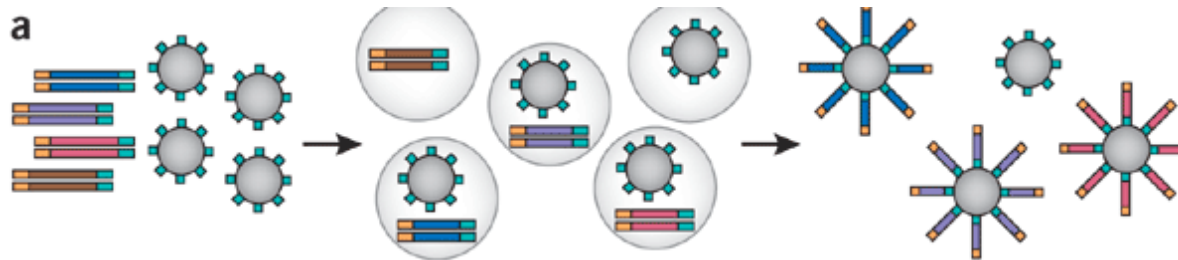
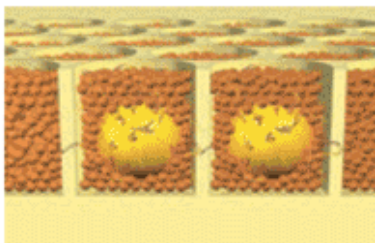
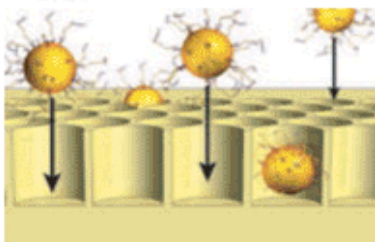
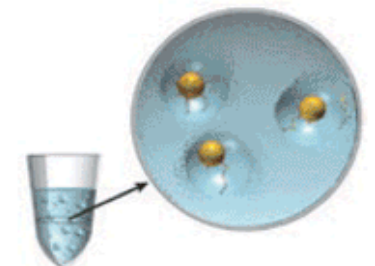
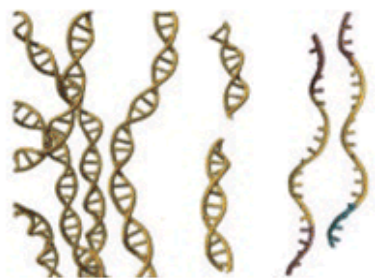


- DNA ulega fragmentacji
- Ligowane są adaptory (o znanej sekwencji)
- Tworzenie macierzy z bibliotekom klonów fragmentów DNA (kilka różnych protokołów)
- Cykliczna reakcja polimeryzacji z wykorzystaniem znakowanych nukleotydów
- Automatyczne sczytanie wyników poprzez cykliczną analizę obrazu macierzy

Dostępne platformy NGS

- **Illumina/Solexa**
- **ABI SOLiD**
- **Roche 454**
- Polonator
- HeliScope

PCR w kropli



Fragmenty z adaptorami są amplifikowane (PCR w kropli wody w oleju).

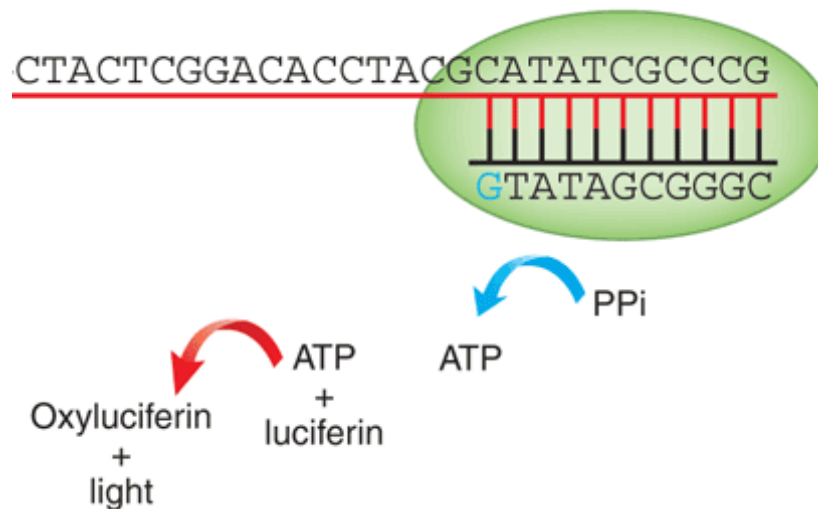
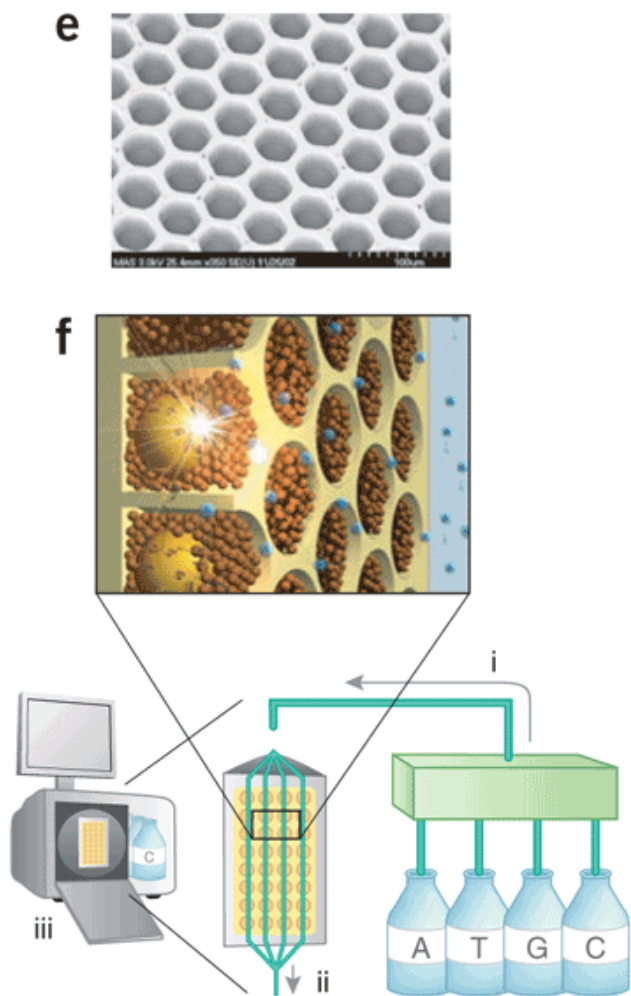
Jeden starter jest przyczepiony do powierzchni kropli.

Używany przez 454, Polonator i SOLiD.

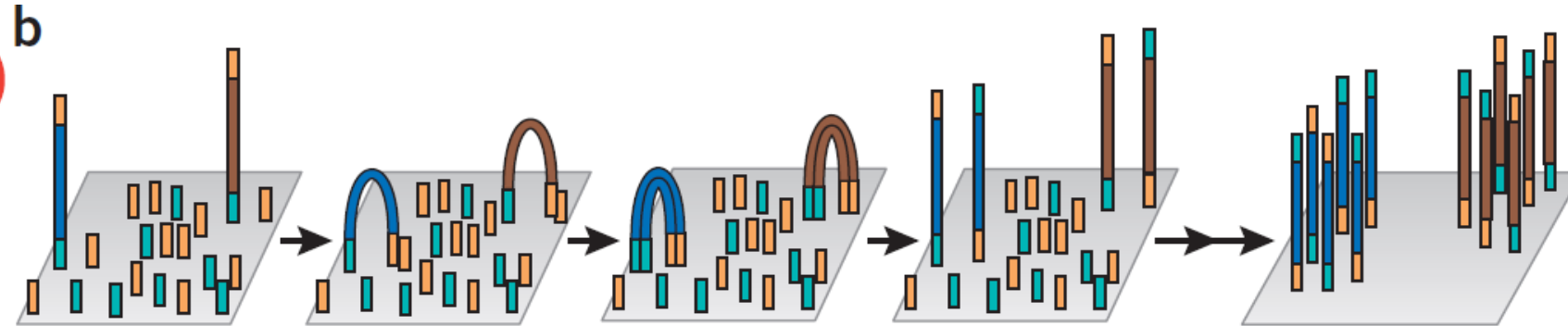
454 Sequencing

Statystyki:

- odczytaj długości 200-300 pz
- problem dokładności z homopolimerami
- 400 000 odczytów na run
- kosztuje 60 USD za Mb (1 000 000 pz)

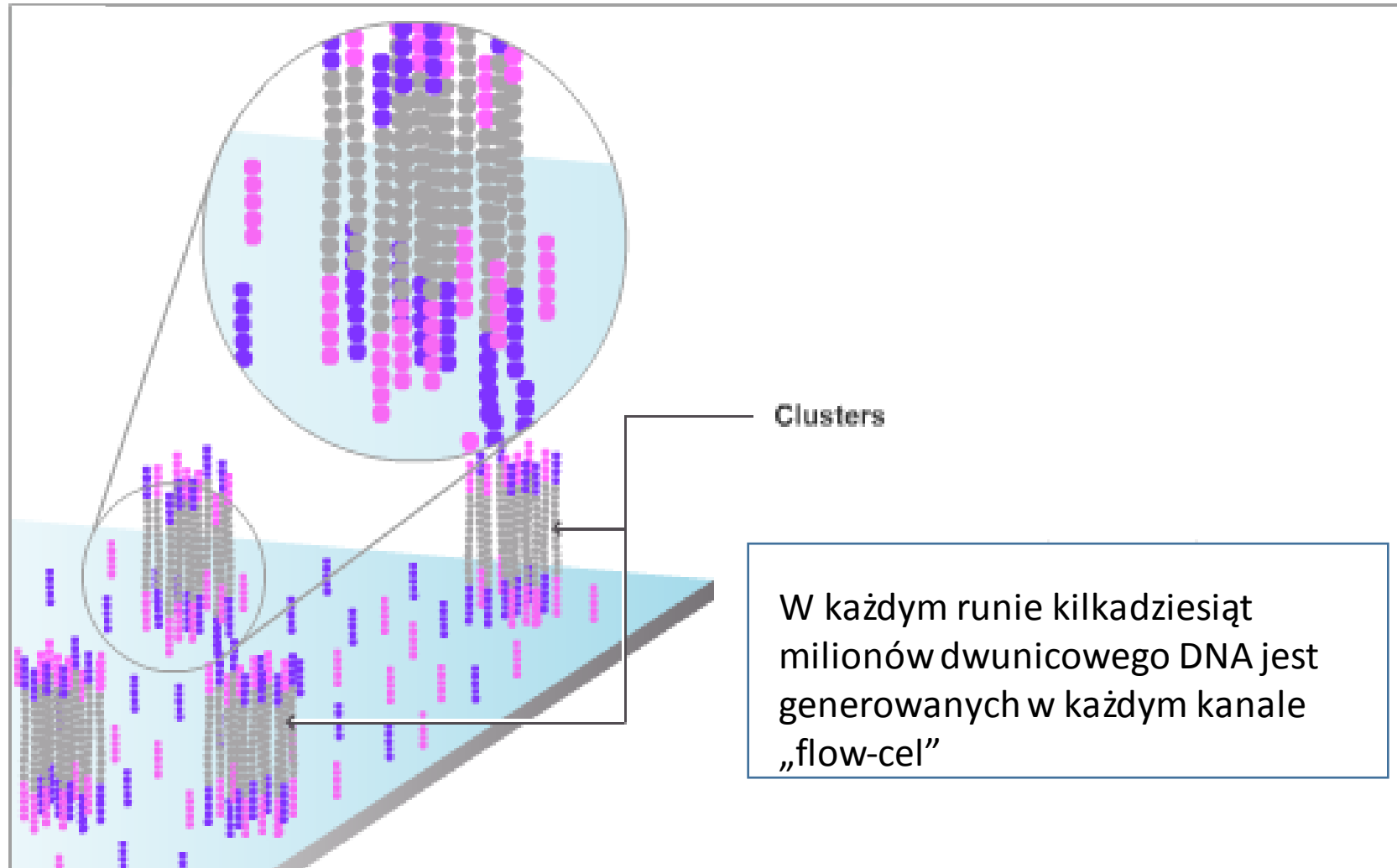


Bridge PCR

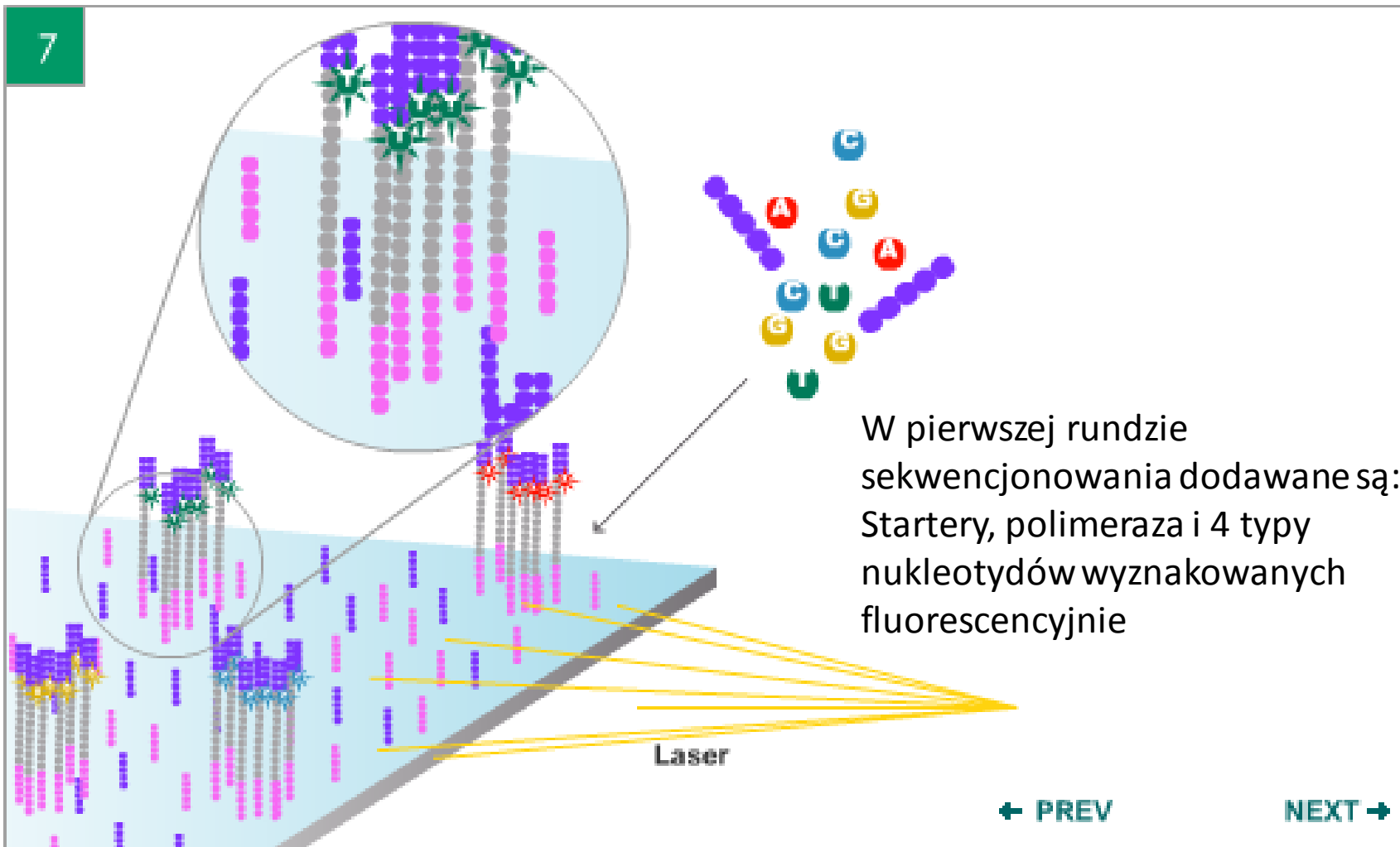


- Fragmenty DNA ligowane są do adaptorów na macierzy szklanej
- Macierz zawiera dwa typy cząsteczek które są komplementarne do obydwu adaptorów przłączonych do fragmentu DNA
- Amplifikacja cykliczna (PCR) – Bridge PCR
- Powstają klastry – grupy klonów wyjściowych fragmentów DNA
- Wykorzystywana przez Illumina/Solexa

Sekwencjonowanie przez syntezę

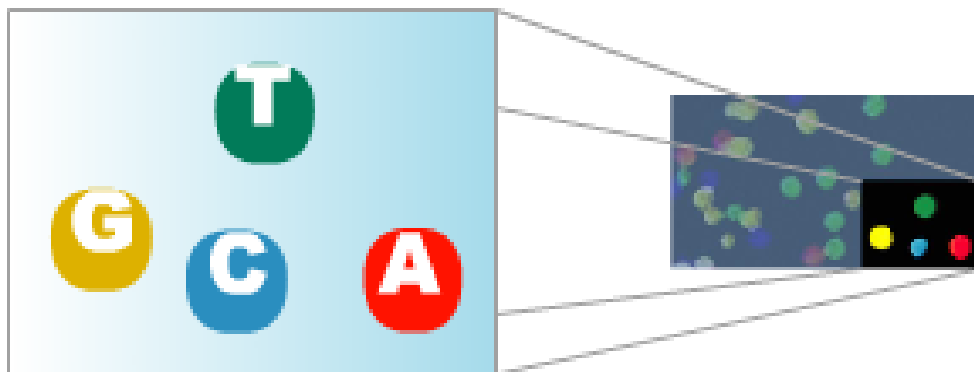


Sekwencjonowanie przez syntezę



Sekwencjonowanie przez syntezę

8

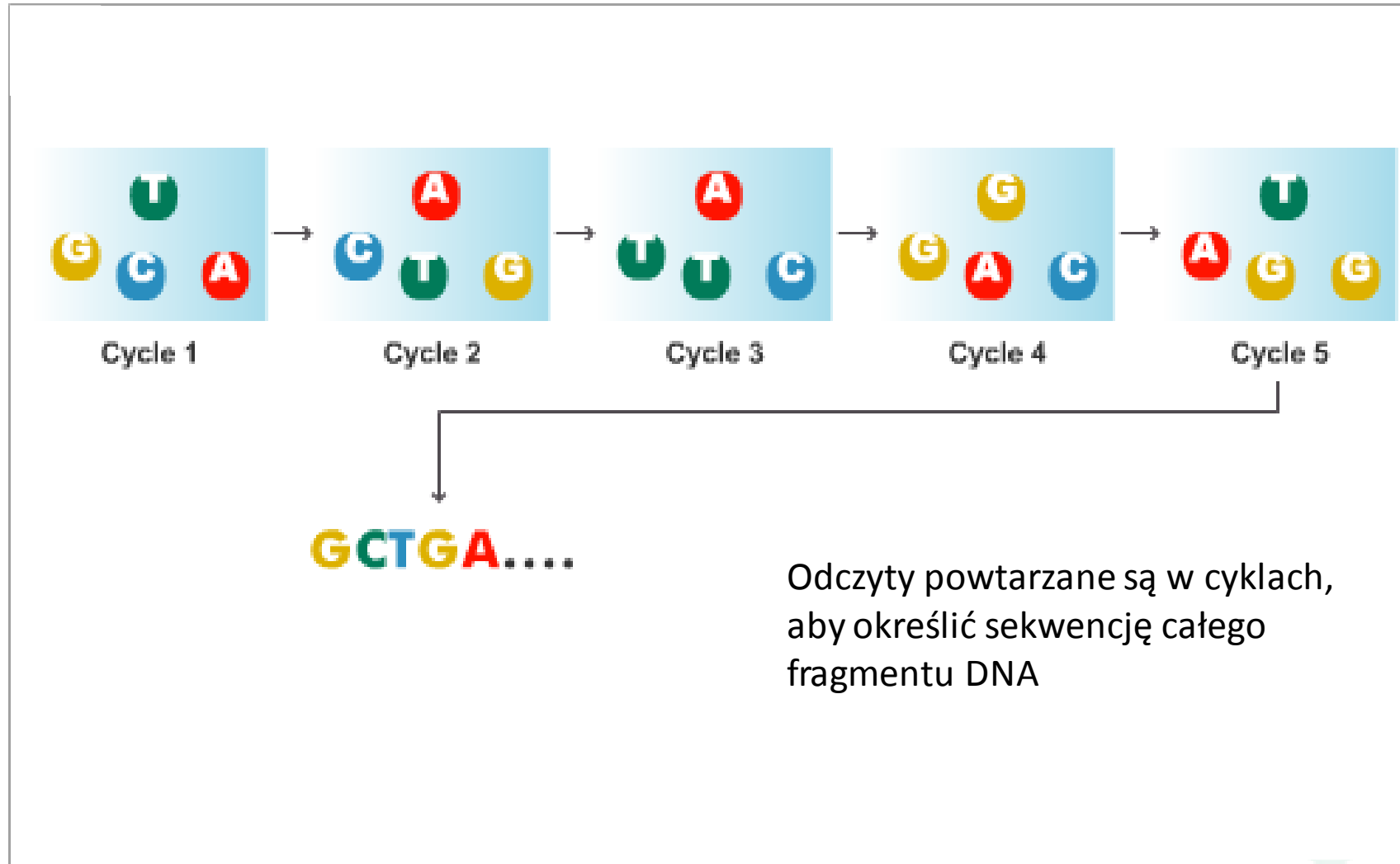


1. Robione jest zdjęcie po pierwszym cyklu

2. Fluorochromy są usuwane

3. Usuwane są blokery reakcji na końcu 3'

Sekwencjonowanie przez syntezę



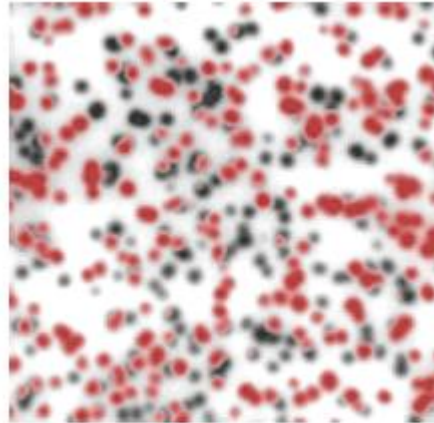
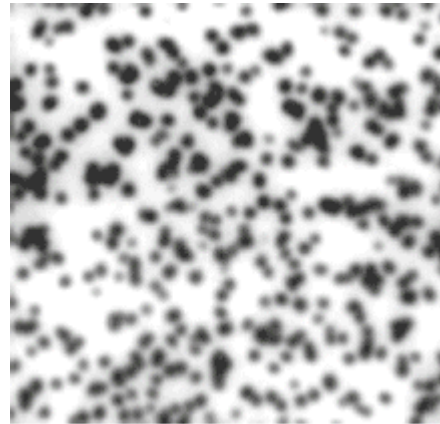
Statystyki:

odczytaj długości do 36 pz

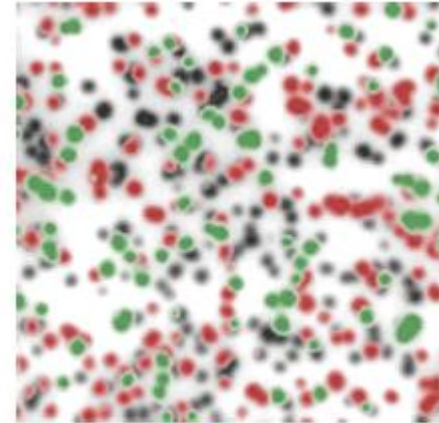
wskaźniki błędów 1-1,5%

kilka milionów „clastrów” na ścieżkę (8 ścieżek
na płytce)

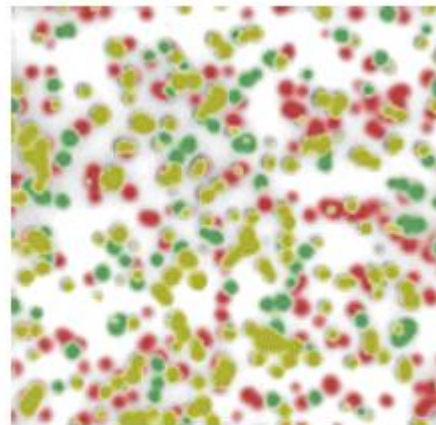
kosztować 2 USD za Mb (1 000 000 pz)



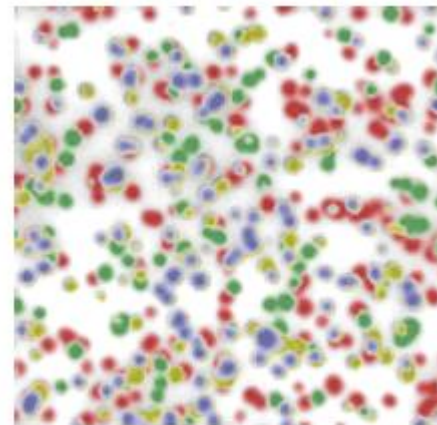
T



T G



T G C



T G C A

Konwencjonalne sekwencjonowanie

- Pozwala na sekwencjonowanie fragmentów do 1000 pz, z dokładnością 99,999%. Koszty sekwencjonowania metodą Sanger w ujęciu sekwencjonowania genomu metodą „strzelby genowej” wynoszą ok 0,50 \$ za kbp (1000 pz).

Jakość sekwencji

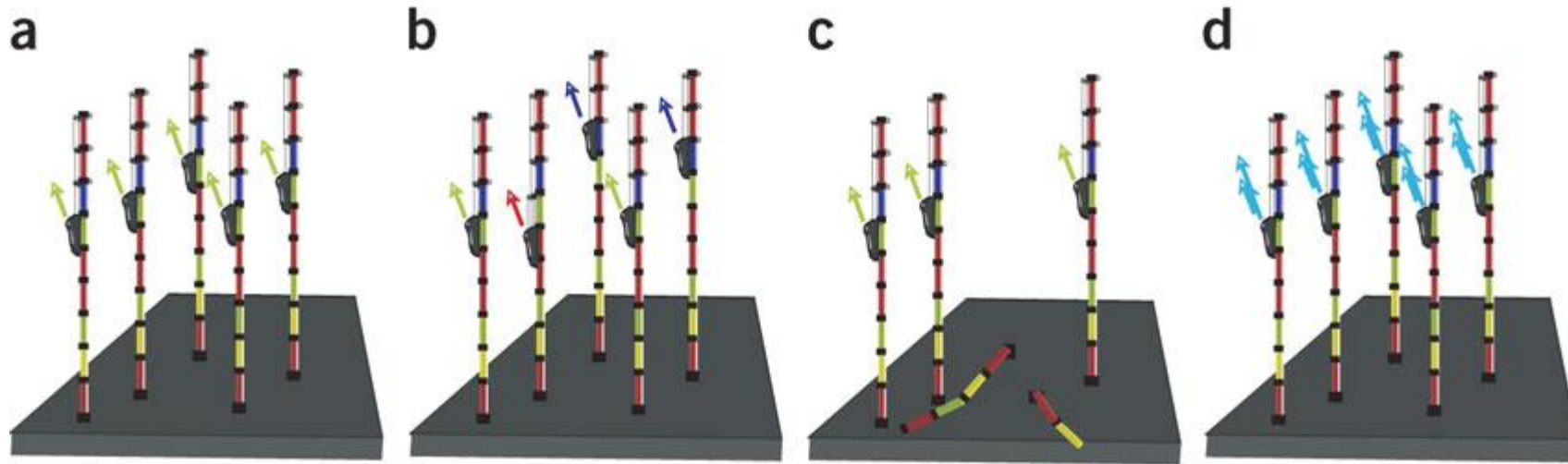
- W większości przypadków jakość odczytu jest najniższa na końcach, a najwyższa w rejonie środkowym odczytywanej sekwencji
- Wykorzystanie miary jakości sekwencji (specjalna jednostka)
 - "Przycinanie" odczytów
 - Usuwanie końców niskiej jakości
 - Oznaczanie konsensusu podczas składania sekwencji
 - „variant calling” – detekcja mutacji SNP
- Ogólnie rzecz biorąc, nowsze podejścia dają większe ilości sekwencji, które są krótsze i mają niższą jakość
 - Sekwencjonowanie nowej generacji ma współczynnik błędów około 1% lub więcej

Phred Quality Score – jednostka „jakości” sekwencji

$$q = -10 \log_{10}(p)$$

- p = prawdopodobieństwo błędnego odczytu zasady
- Jeżeli $p = 0.01$ (1% szans na błąd) to q wynosi 20
- Jeżeli $p = 0.00001$ (dokładność 99.999%) to q wynosi 50
- Wartości Phred (q) są zaokrąglane do najbliższej liczby całkowitej

Główne czynniki obniżające jakość odczytów

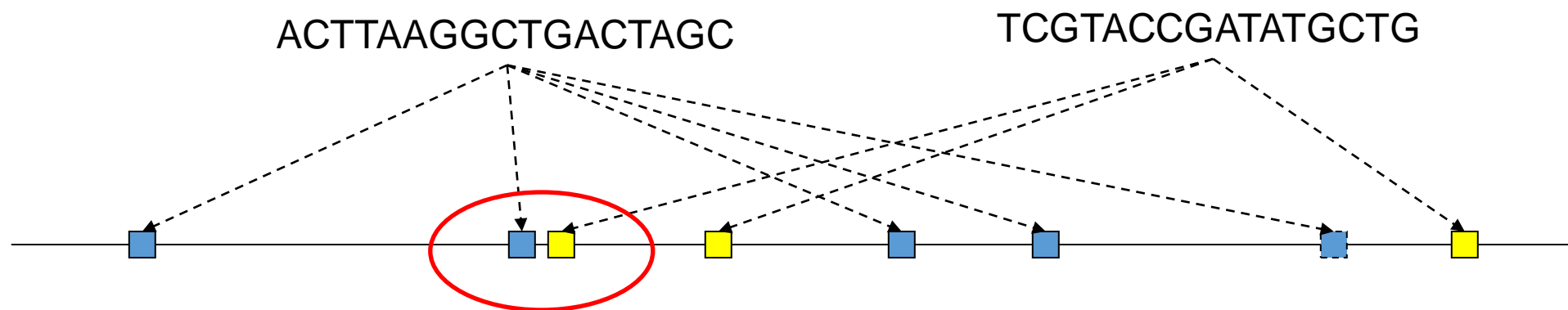


Porównanie istniejących metod

	Feature generation	Sequencing by synthesis
454	Emulsion PCR	Polymerase (pyrosequencing)
Solexa	Bridge PCR	Polymerase (reversible terminators)
SOLiD	Emulsion PCR	Ligase (octamers with two-base encoding)
Polonator	Emulsion PCR	Ligase (nonamers)
HeliScope	Single molecule	Polymerase (asynchronous extensions)

Cost per megabase	Cost per instrument	Paired ends?	1° error modality	Read-length
~\$60	\$500,000	Yes	Indel	250 bp
~\$2	\$430,000	Yes	Subst.	36 bp
~\$2	\$591,000	Yes	Subst.	35 bp
~\$1	\$155,000	Yes	Subst.	13 bp
~\$1	\$1,350,000	Yes	Del	30 bp

Długości odczytów i parowanie (pair-edn)



- Krótkie odczyty są problematyczne, ponieważ krótkie sekwencje mogą mapować się w wielu miejscach
- Rozwiązanie #1 sekwencjonować dłuższe fragmenty
- Rozwiązanie #2 sekwencjonować dwa końce dłuższej cząsteczki

Sekwencjonowanie trzeciej generacji

- Sekwencjonowanie pojedynczej cząsteczki
 - Sekwencjonowanie bez amplifikacji DNA
 - Helicos HeliScope
 - Pacific Biosciences SMRT
- Dłuższe odczyty
 - Roche/454 > 400bp
 - Illumina/Solexa > 150bp
 - Pacific Bioscience > 1000 bp i pojedyncza cząsteczka

Zastosowanie NGS

- Sekwencjonowanie genomów
- Targetowe reseqwencjonowanie fragmentów genomów
- Genotypowanie
- RNAseq
- Analizy metylacji DNA

Wyzwania w analizie danych

- Wyznaczanie zasad (base calling)
- Mapowanie do genomu referencyjnego
- Asambling de novo (składanie genomu de novo)