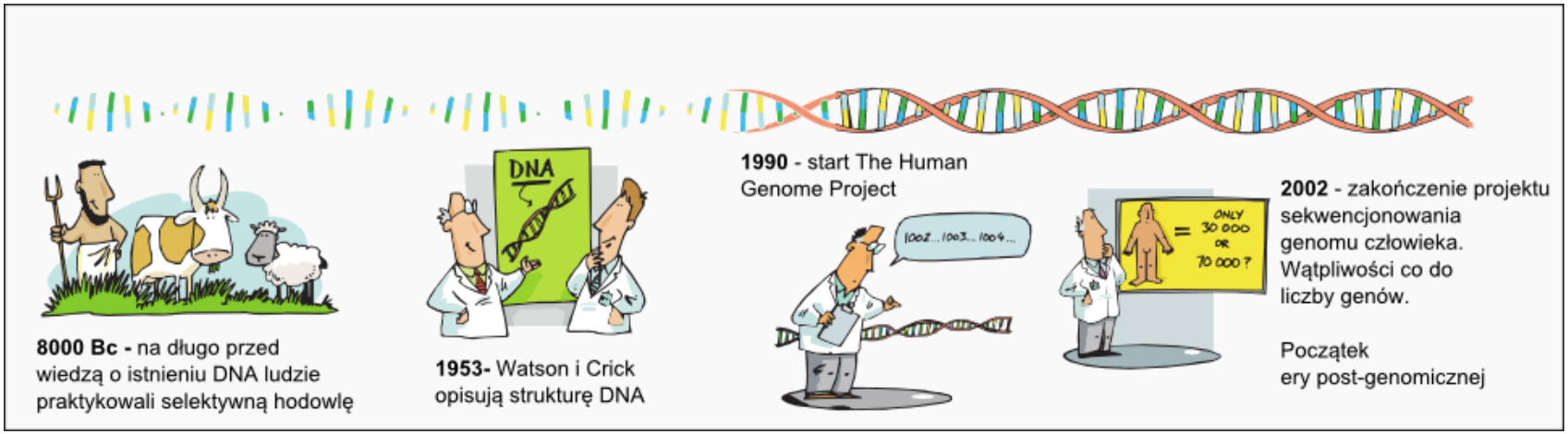
Inżynieria genetyczna (INGE)

Wykład 4 - sekwencjonowanie

Robert Nowak

2024L

Wstęp

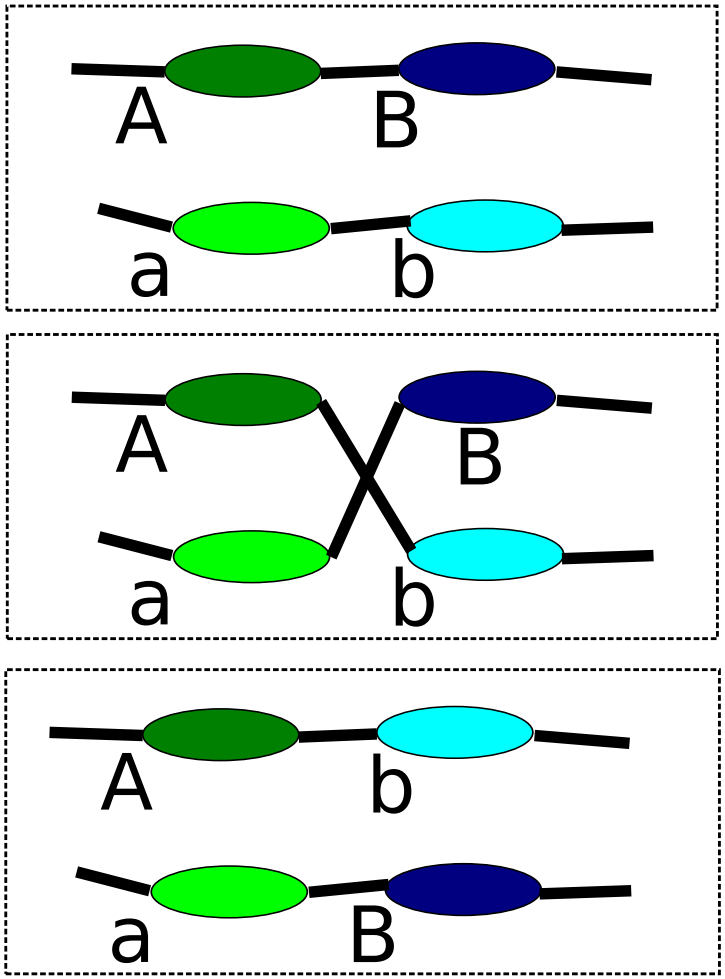


Genotyp informacja genetyczna danego osobnika, najczęściej łańcuch lub zbiór łańcuchów DNA (od 3500 par zasad u wirusów do $3 * 10^{11}$ bp u ameby)

Genom podstawowy materiał genetyczny (DNA jądrowy u eukariotów, $3 * 10^{9}$ u człowieka)

Mapa genetyczna i fizyczna

- Mapa genetyczna - mapa lokalizacji genów lub markerów na chromosomie powstała poprzez badanie osobników w krzyżówce testowej. Jednostka – centymorgany (cM).
- Mapa fizyczna - mapa powstała poprzez odczyt sekwencji. Jednostka - nt (nukleotydy) lub bp (pary zasad).



cM - prawdopodobieństwo rozdzielenia w jednym pokoleniu podczas rekombinacji (cross-over) wynosi 1%

Tworzenie mapy genetycznej

zlicza się cechy u osobników potomnych:

- markery są położone blisko na chromosomie $\rightarrow$ są dziedziczone wspólnie (małe prawd. ich rozdzielenia)
- markery są oddalone $\rightarrow$ są dziedziczone niezależnie (duże prawdopodobieństwo ich rozdzielenia)

Przykład, krzyżowanie `AB/ab` i `ab/ab`[a]

[a] krzyżówka z podwójną homozygotą recesywną upraszcza obliczenia

- jeżeli markery są odległe (niesprzężone), to

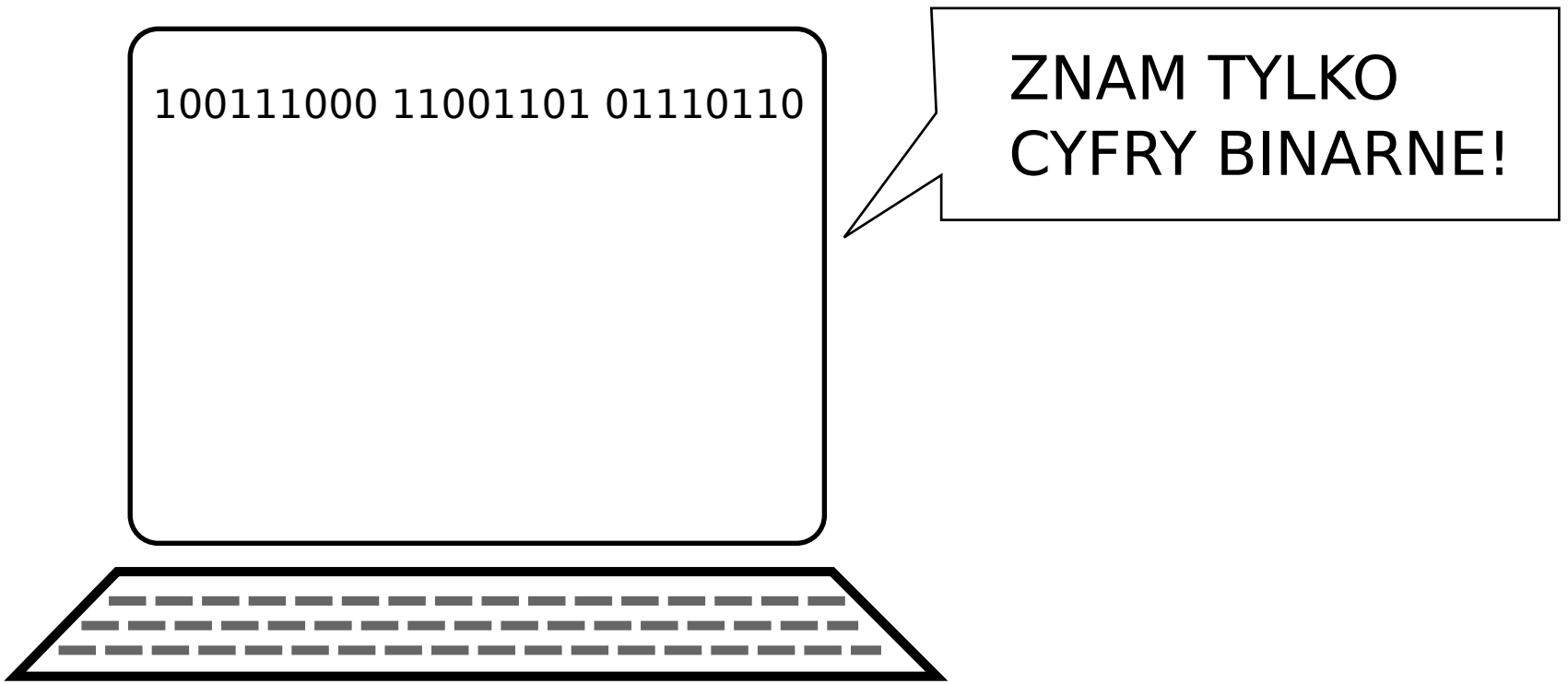
  $$P(AB/ab) = P(Ab/ab) = P(aB/ab) = P(ab/ab) = 0.25$$

- markery sprzężone zupełnie (w 100%)

  $$P(AB/ab) = P(ab/ab) = 0.5, P(aB/ab) = P(Ab/ab) = 0.0$$

- inne częstości oznaczają sprzężenie częściowe

Reprezentacja cyfrowa



Komputer może przetwarzać jedynie informację dostępną w formie cyfrowej

DNA, RNA, białko - reprezentacja cyfrowa

Opis biopolimerów:

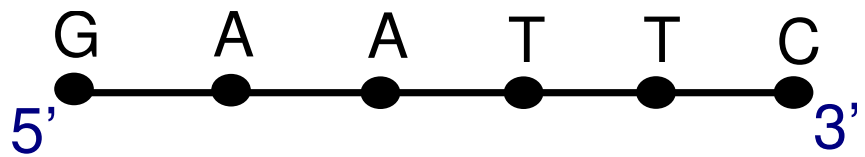
- struktura pierwszorzędowa - sekwencja symboli
- struktura drugorzędowa - uwzględnienie oddziaływania nukleotydów (DNA, RNA) lub aminokwasów (białka)
- struktura trzeciorzędowa - położenie atomów w 3D

Σ - alfabet, $\Sigma_{DNA} = \{A, C, G, T\}$

s - sekwencja nukleotydów $s = s_1 s_2 ... s_n$, gdzie

n - długość sekwencji s, oznaczana $|s|$,

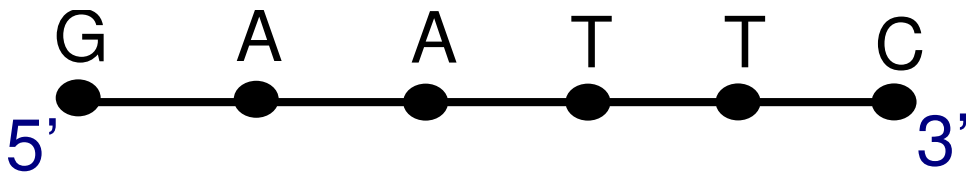
s_i - symbol (o indeksie i) sekwencji s, $s_i \in \Sigma$



Konwencja: ciągi piszemy od końca 5'.

Sekwencjonowanie

Sekwencjonowanie DNA - proces ustalania kolejności nukleotydów tworzących cząsteczkę DNA.



Proces składa się z:

1. odczyt sekwencji fragmentów
2. składanie (asemblacja)
3. wykańczanie

Poznawanie sekwencji genomów - historia

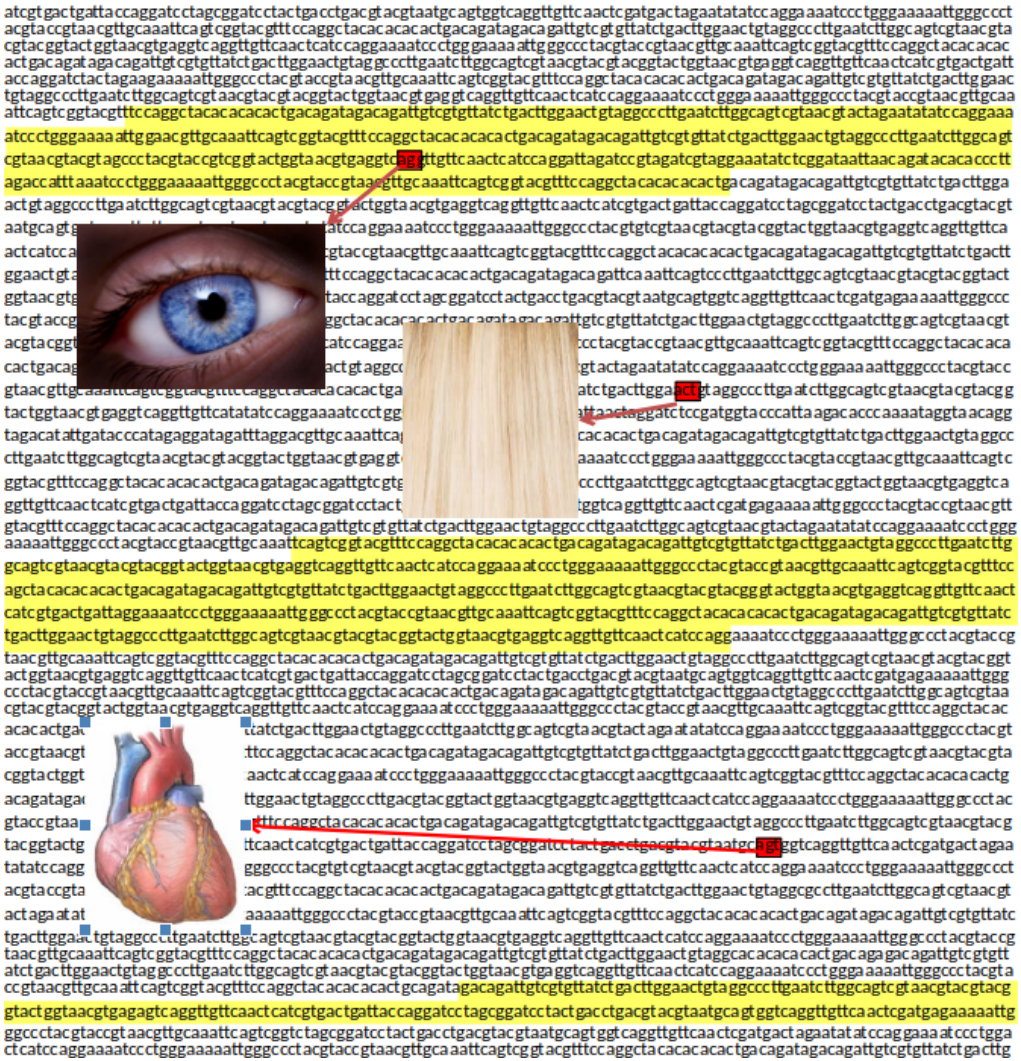
fag Φ174	5400 bp	1977
wirus	170 kbp	1984
bakteria	1.8 Mbp	1995
drożdże	12 Mbp	1996
człowiek	3.3 Gbp	2004

obecnie(2024) 410850 genomów: archea(4066), bakterie(353956), eukariotyczne(35158), wirusy(17670)

Genom referencyjny człowieka:

- GRCh38 - major release (2013)
- GRCh38 v.14 - bieżąca wersja (Luty 2022), 434 gaps
- udało się uzyskać pełny genom człowieka (31.03.2022), `doi:10.1126/science.abj6987`

Warianty genetyczne



Sekwencja jest polimorficzna, zawiera warianty, które mogą:

- determinować cechy fenotypowe: np. kolor oczu, kolor włosów
- mieć znaczenie medyczne: np. grupa krwi, metabolizm leków

warianty genetyczne dla człowieka - statystyki

Znane modyfikacje (`http://www.hapmap.org`)

- zmiana jednego symbolu (SNP) : $11 * 10^6$
- indel (do $10000nt$) : $3 * 10^6$
- re-aranżacje: $3 * 10^4$ (zajmują ok. 10% genomu)

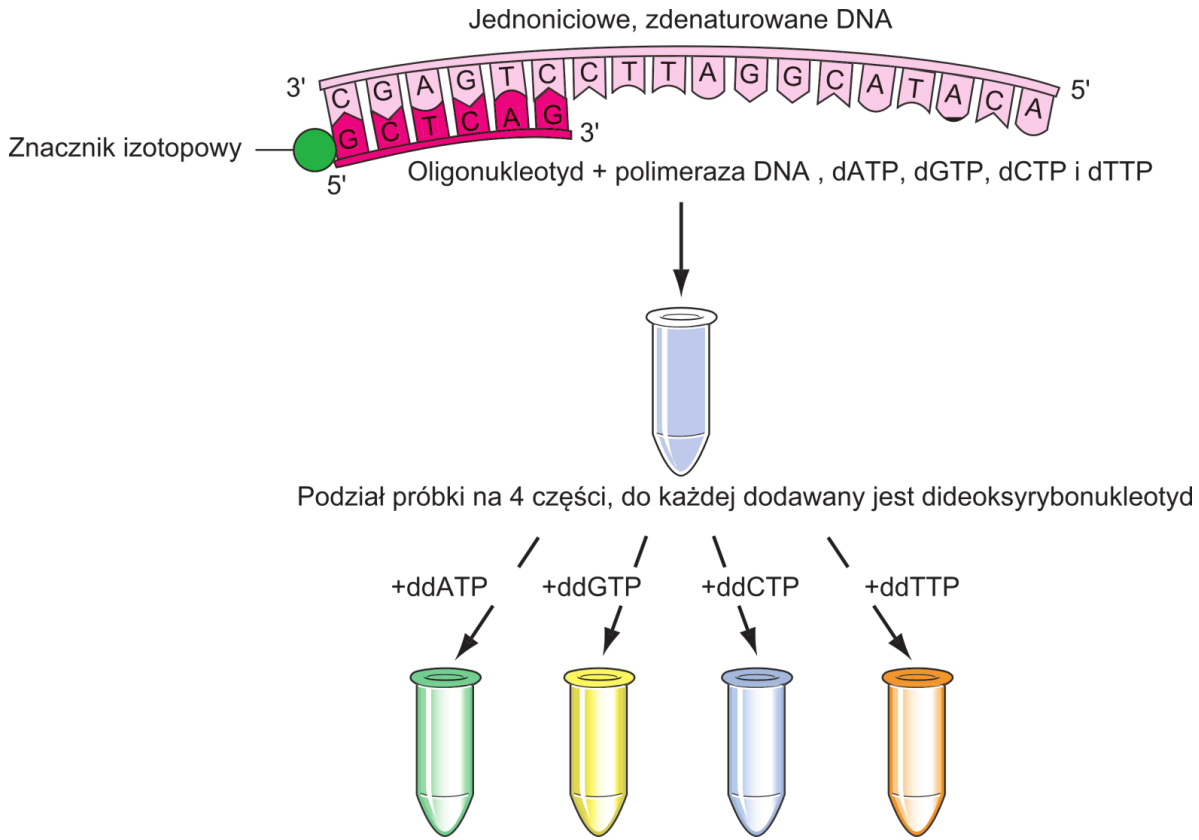
średnio dla zdrowego człowieka
(`http://www.1000genomes.org/`)

- SNP (wszystkich $3 * 10^6$), w sekwencjach kodujących:
  - wstawienie kodonu stop (nonsense) : 1057
  - utrata kodonu stop : 77
  - zmiana kodonu (missense) : $68 * 10^3$
  - bez zmiany kodonu (silent) : $60 * 10^3$
- small indel (do $10000nt$): $362 * 10^3$
- usunięcia: $16 * 10^3$
- insercje: 4775
- duplikacje: 407

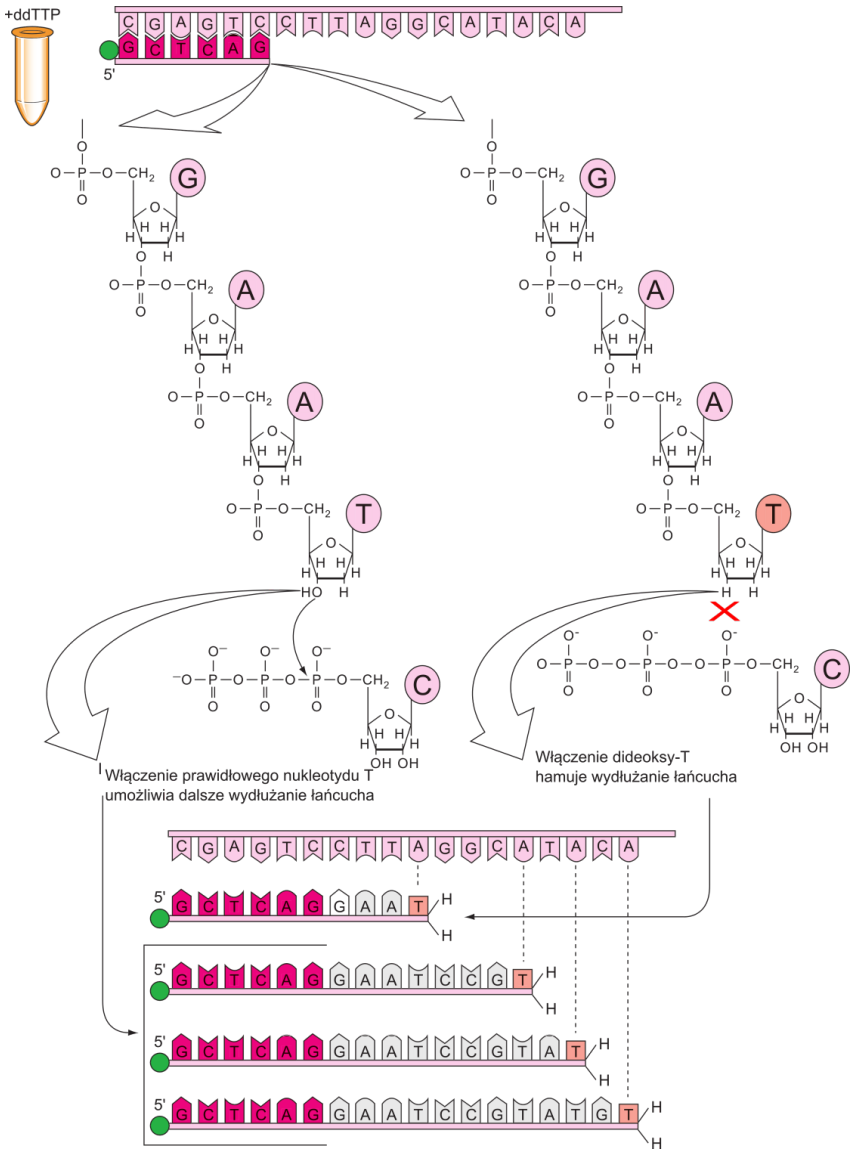
Sekwencjonowanie metodą Sangera

Metoda Sangera (1) - metoda terminacji łańcucha

Matryca DNA jest denaturowana i dodawany jest znakowany izotopem starter, polimeraza DNA i nukleotydy (dNTP)

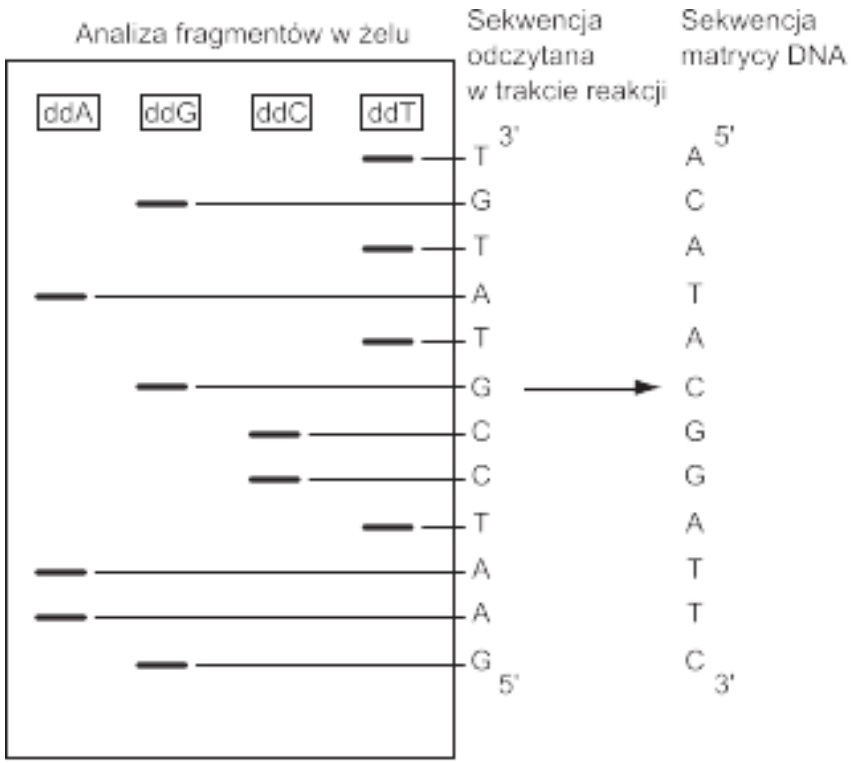


Metoda Sangera (2)



- w trakcie syntezy DNA przez polimerazę dd-nukleotydy są włączane do nowej nici tak jak d-nukleotydy
- brak grupy 3'OH uniemożliwia dalsze wydłużanie DNA
- każda próbka z ddT produkuje serię różnej długości fragmentów zakończonych T, co odpowiada A na nici matrycowej

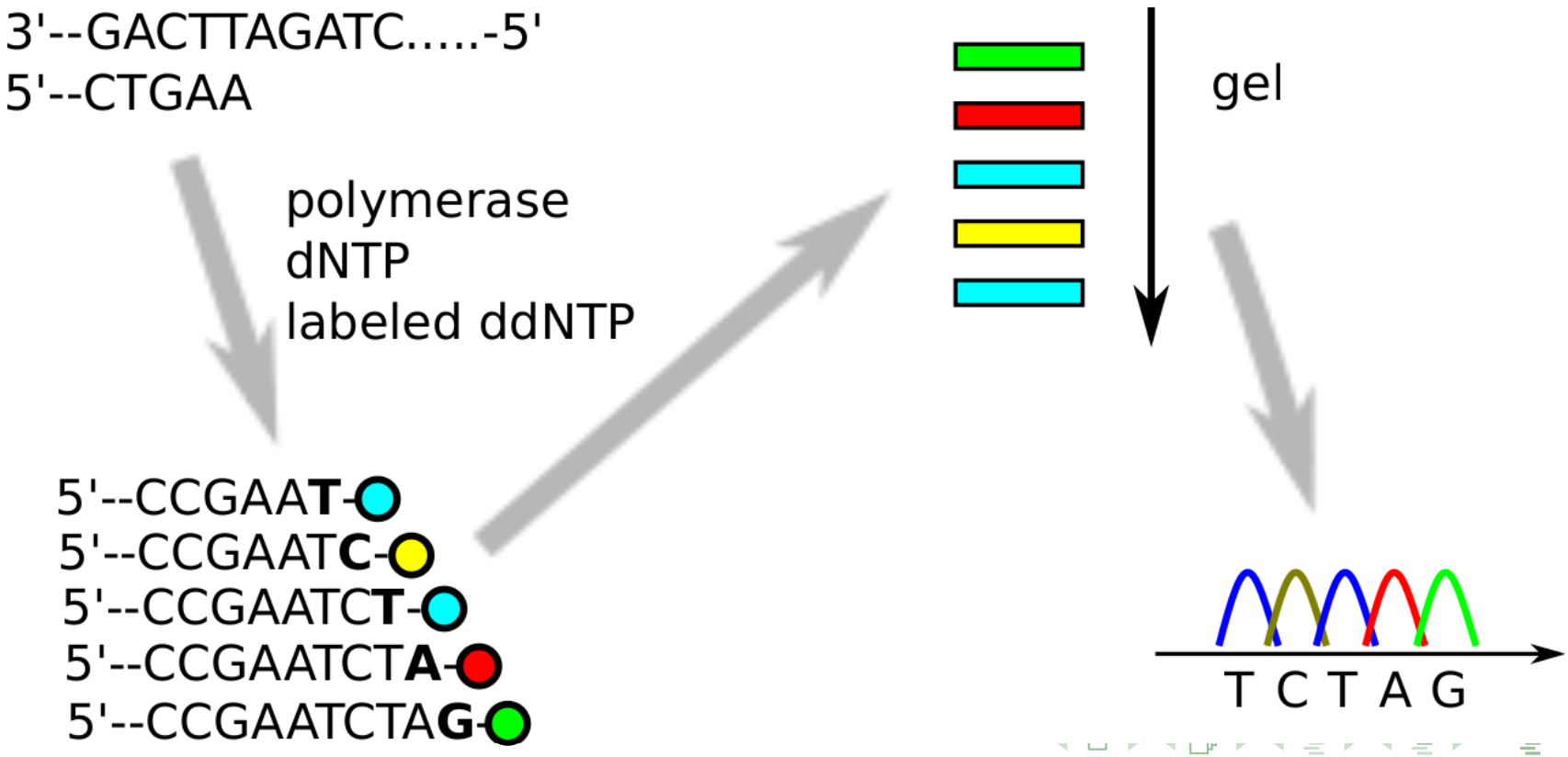
Metoda Sangera (3)



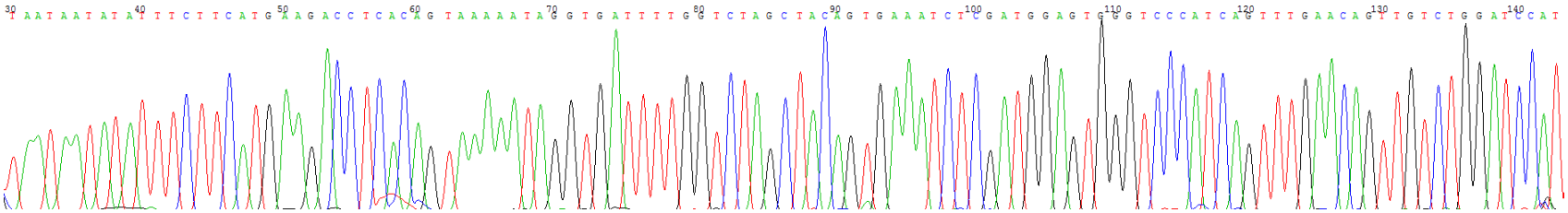
- do rozdziału fragmentów używana jest elektroforeza w żelu poliakrylamidowym
- długość fragmentu odpowiada specyficznym dd-nukleotydom
- sekwencja od 5' do 3' syntezowanej nici jest czytana od dołu żelu

Automatyczne sekwencjonowanie metodą Sangera

- dd-nukleotydy ze znacznikiem fluorescencyjnym
- pojedyncza reakcja i ta sama linia żelu (kapilara)
- sygnał rejestrowany jako chromatogram fluorescencji



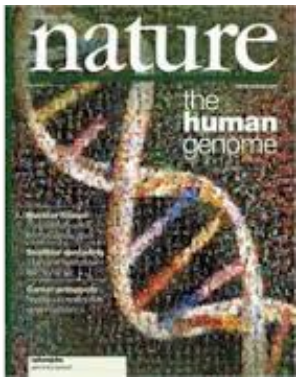
Automatyczna metoda Sangera (2)



Automatyczna metoda Sangera - sekwencjonowanie pierwszej generacji

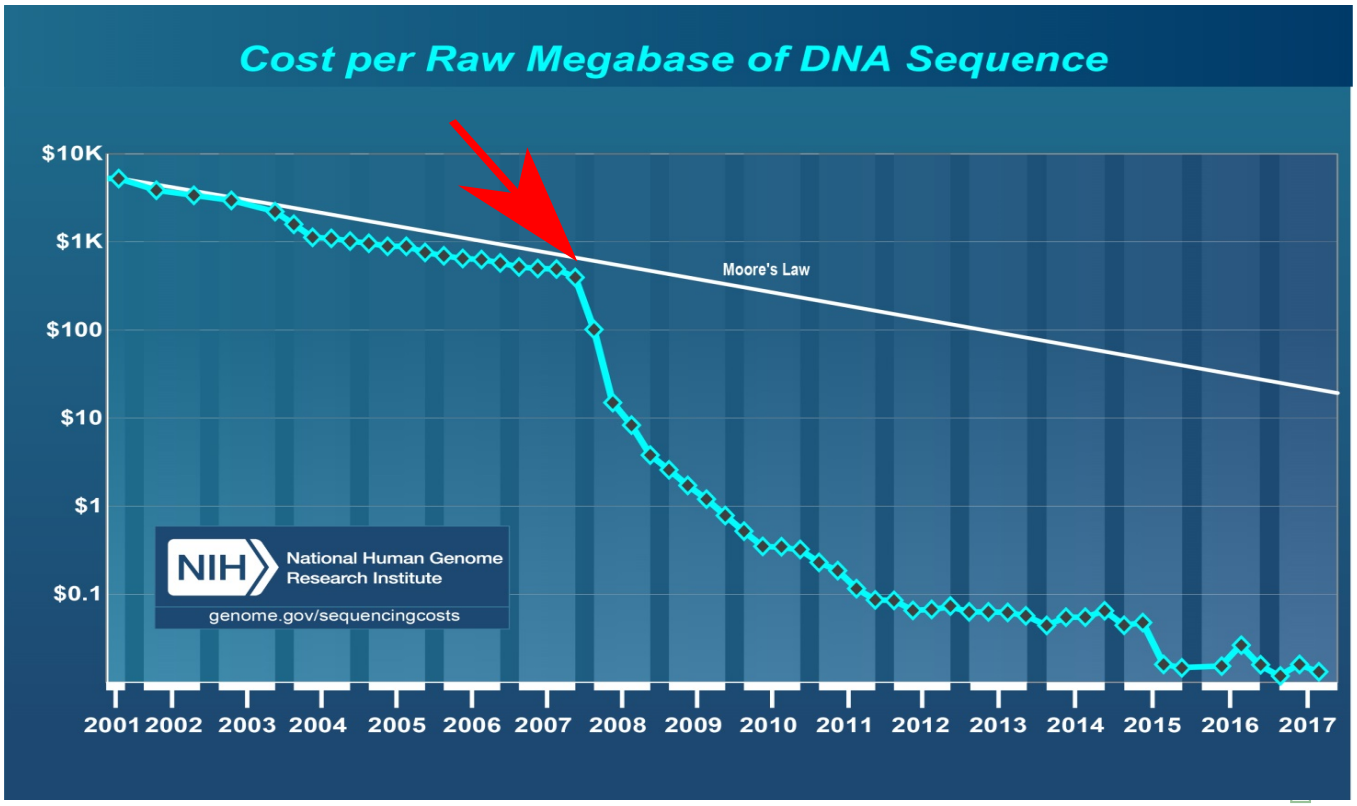
- pozwala odczytać sekwencje 2000 bp na reakcję
- dokładność 99.999%
- zdominowała sekwencjonowanie na 20 lat
- użyta do wygenerowania sekwencji genomu ludzkiego
- niska przepustowość i wysokie koszty

Human Genome Project, 2001, 2.7 Mld$

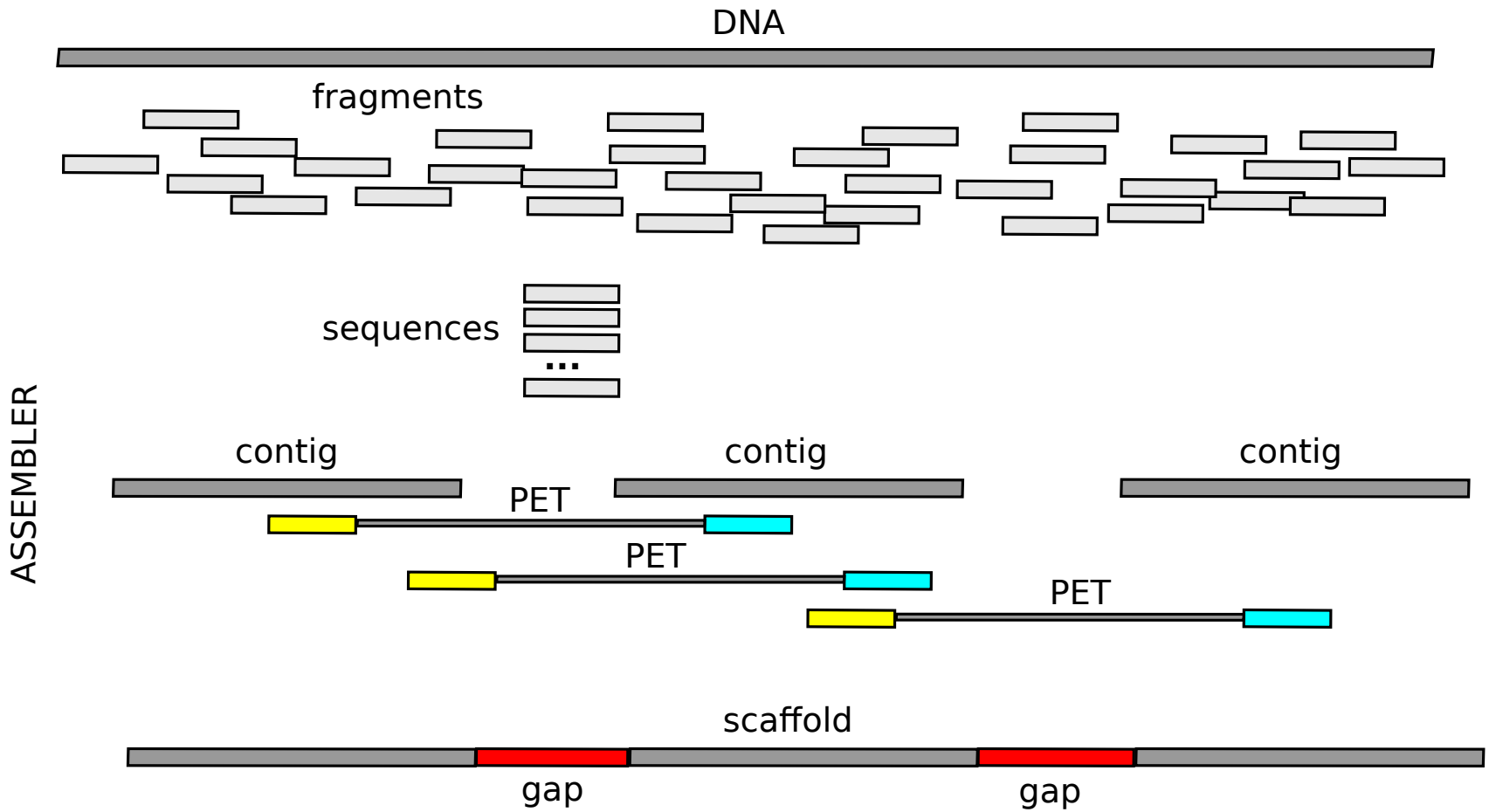


Odczyt sekwencji metodą shotgun (WGS)

- podział łańcucha DNA na losowe fragmenty
- odczytywanie sekwencji fragmentów
- odtwarzanie sekwencji łańcucha na podstawie nakładających się sekwencji fragmentów



Odczyt sekwencji metodą shotgun (WGS)

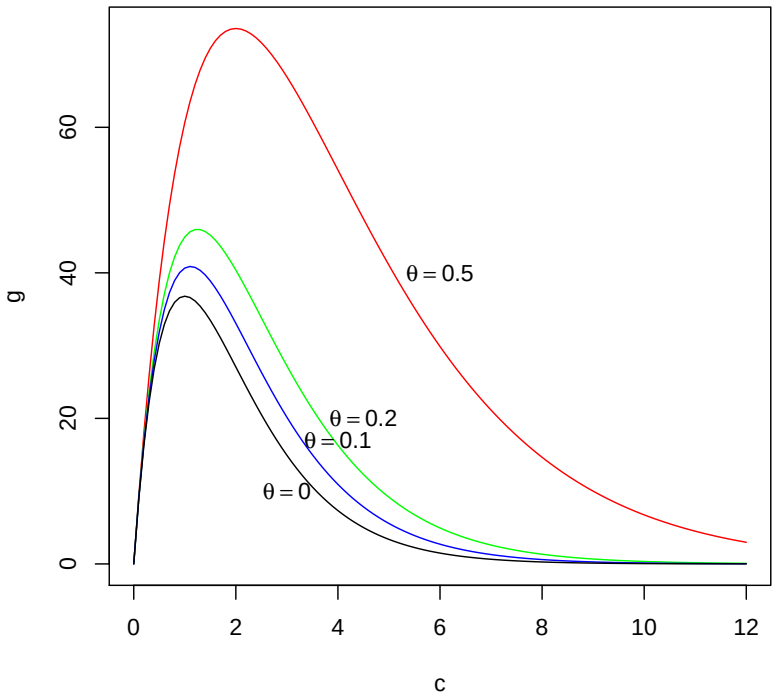


WGS, nieciągłość wyniku – niepełne pokrycie

G długość genomu, L długość fragmentu, N liczba fragmentów
T liczba nukleotydów wymagana do wykrycia pokrywania

$$g = Ne^{-\frac{LN}{G}(1-\theta)}$$

gdzie g - liczba przerw

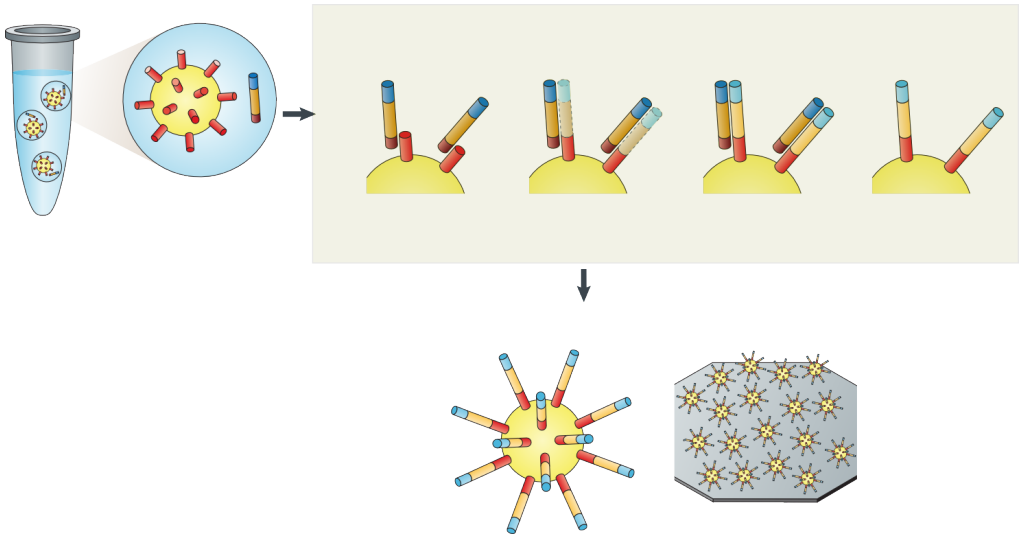


- $\theta = \frac{T}{L}$
- $c = \frac{LN}{G}$, nadmiarowość sekwencjonowania

człowiek 3Gbp, fragmenty 100bp, c $= 30x$, $P(g) \approx 10^{-4}$, $N \approx 10^9$, plik $360GB$

Sekwencjonowanie drugiej generacji

Klonalna amplifikacja fragmentów, emulsyjny PCR

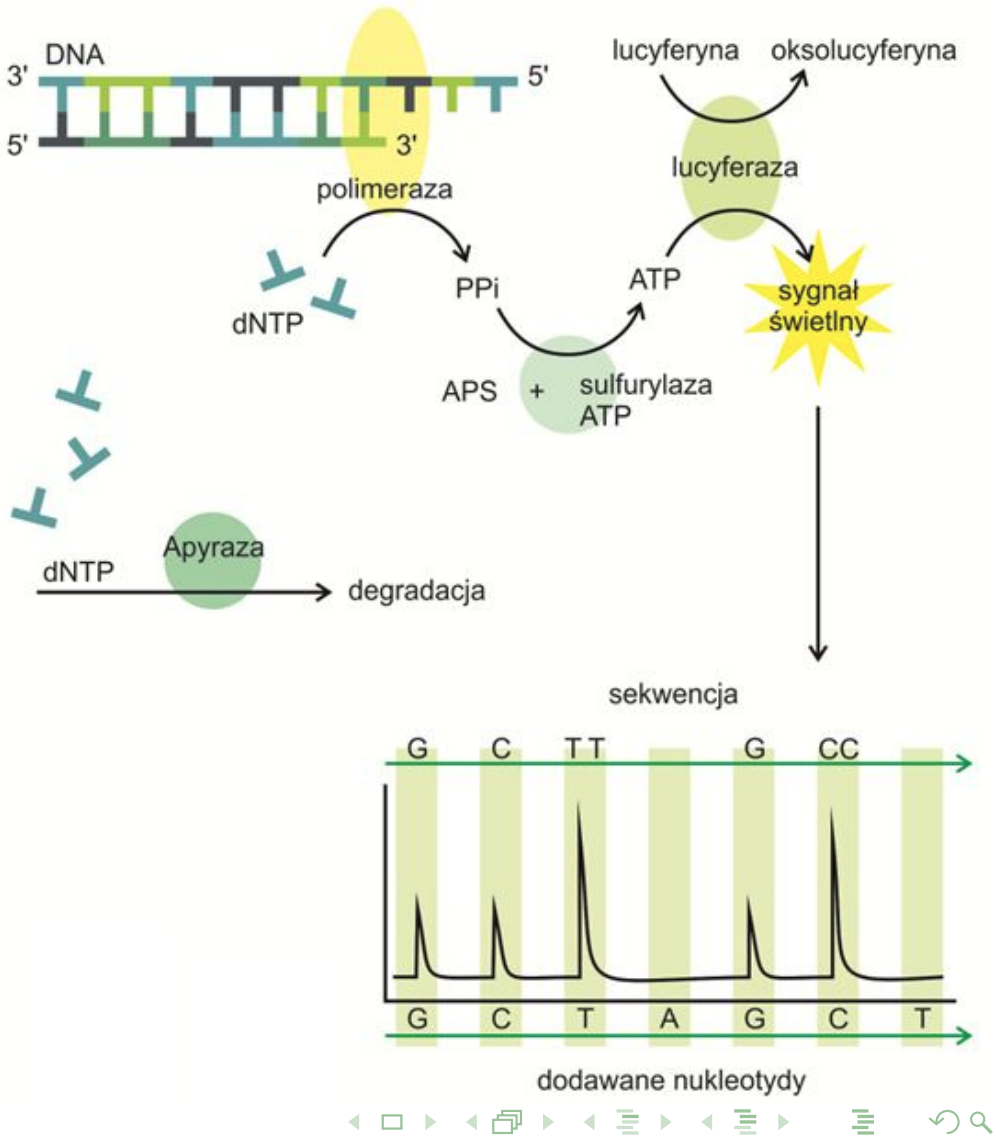


- reaktory: krople w oleju
- kropla zawiera: kulkę, startery, dNTP, matrycę, polimeraza
- jedna cząsteczka DNA na kulkę

- produkty PCR związane do kulek są wtłaczane do studzienek reakcyjnych
- stosowana w urządzeniach: Roche 454, SOLID, Ion Torrent

Piro-sekwencjonowanie, platforma 454

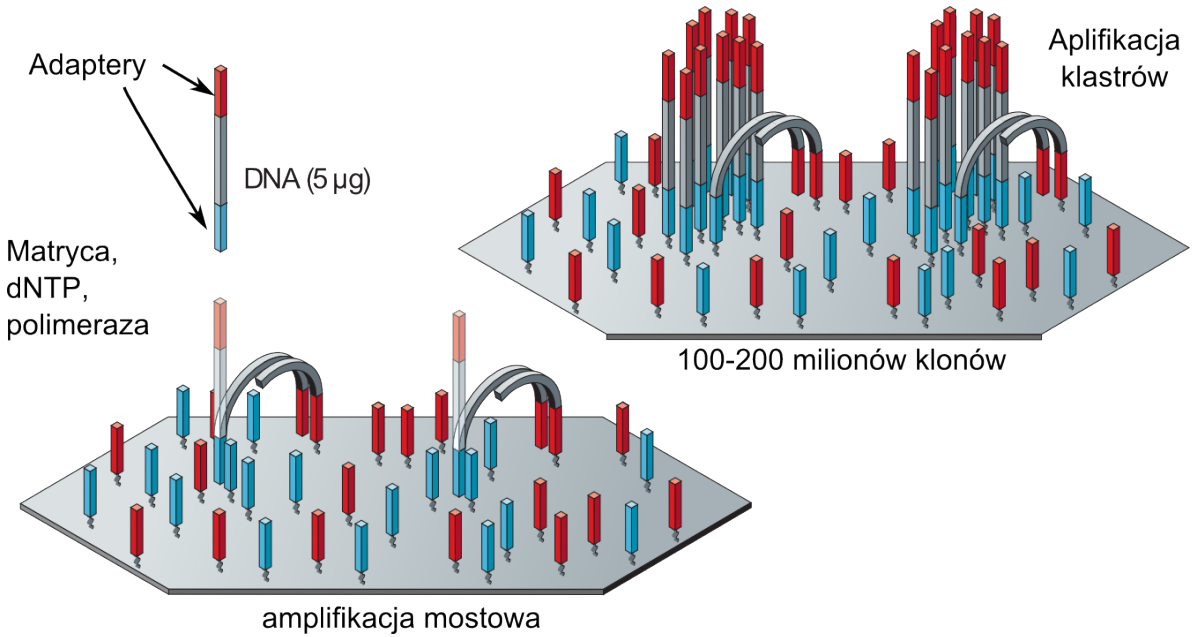
- przeprowadzana jest polimeryzacja z dATP, później dCTP, dGTP, dTTP
- sulfurylaza wychwytuje pirofosforany (PPi) i generuje ATP
- lucyferaza generuje sygnał świetlny
- siła sygnału zależna od ilości zużytego ATP



Platforma Roche 454

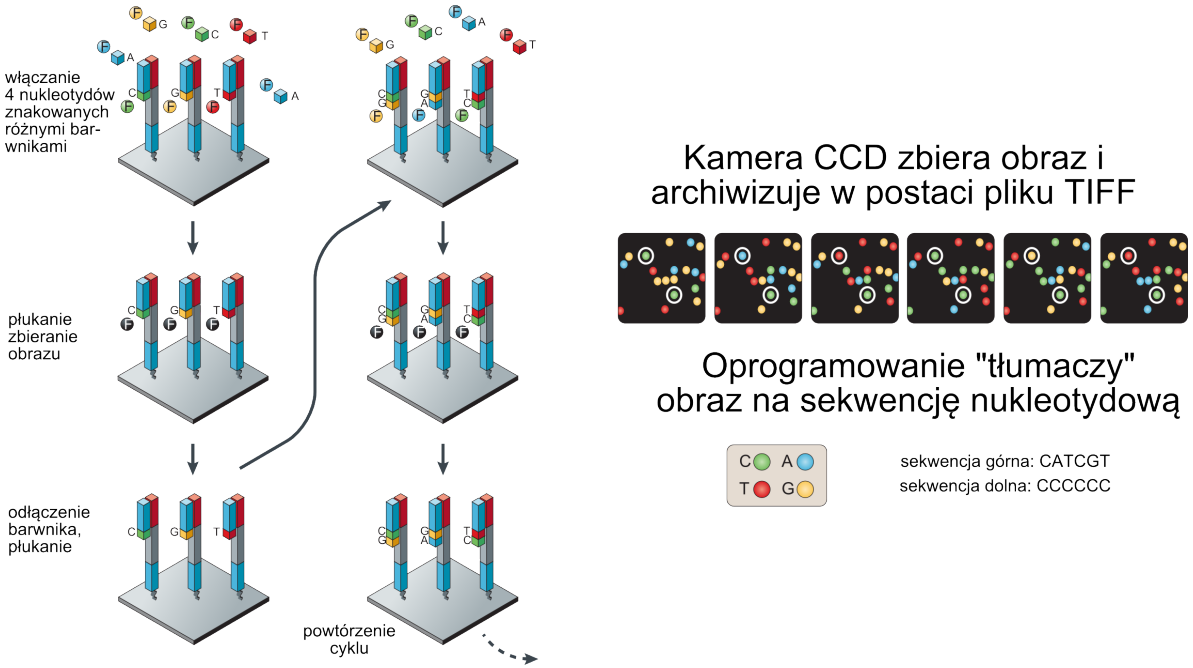
- pierwszy sekwenator drugiej generacji
- początek sprzedaży w 2005 r, koniec produkcji w 2016 r
- sekwencjonowanie $\approx 10x$ tańsze niż metodą Sangera
- $\approx 300x$ bardziej wydajne

Klonalna amplifikacja fragmentów, mostowy PCR, Illumina



- biblioteka losowych fragmentów z adapterami wiązana dwoma końcami do powierzchni płytki
- każdy klaster powiela jeden fragment

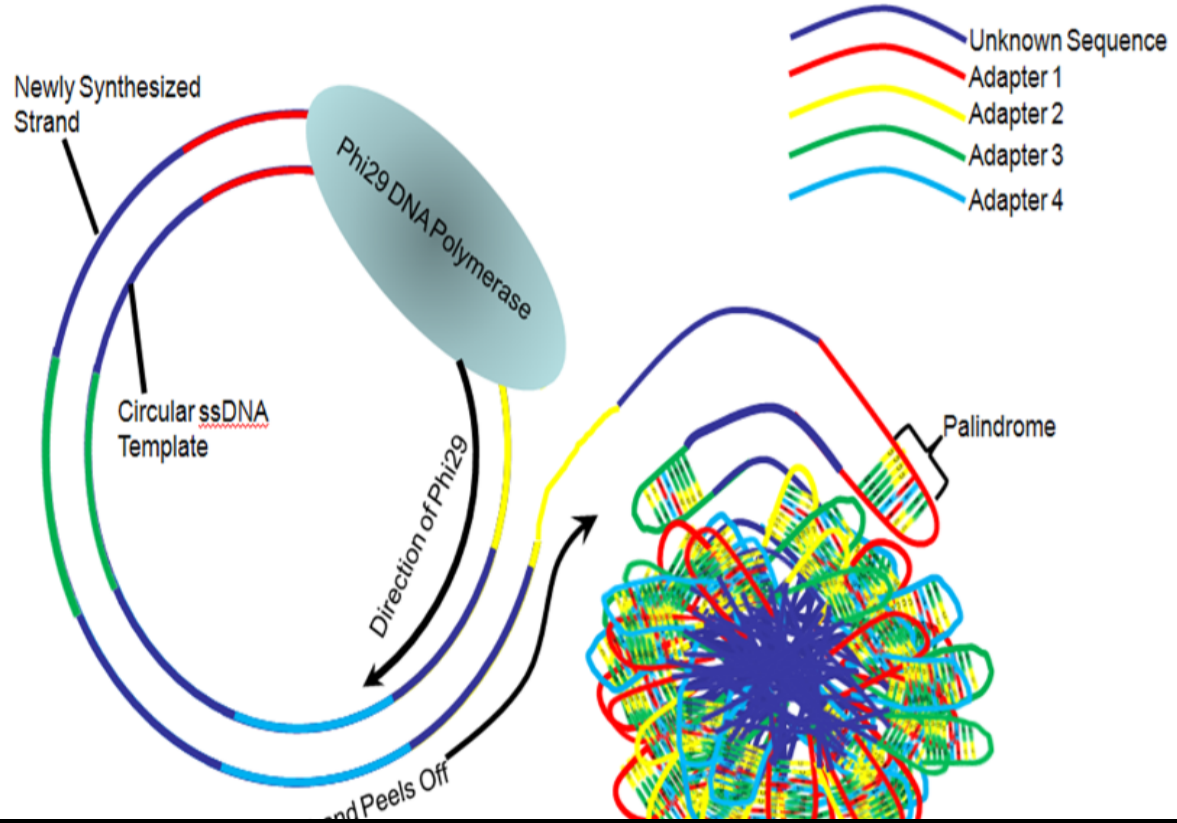
Sekwencjonowanie metodą Illumina



- do syntezy używane usuwalne ddNTP wyznakowane fluorescencyjnie
- iteracyjnie skanuje się a następnie usuwa ddNTP

Amplifikacja metodą BGI

- do powielania wykorzystuje koliste DNA zawierające interesującą sekwencję
- synteza tworzy długą nić, która następnie jest wykorzystywana podczas odczytu sekwencji



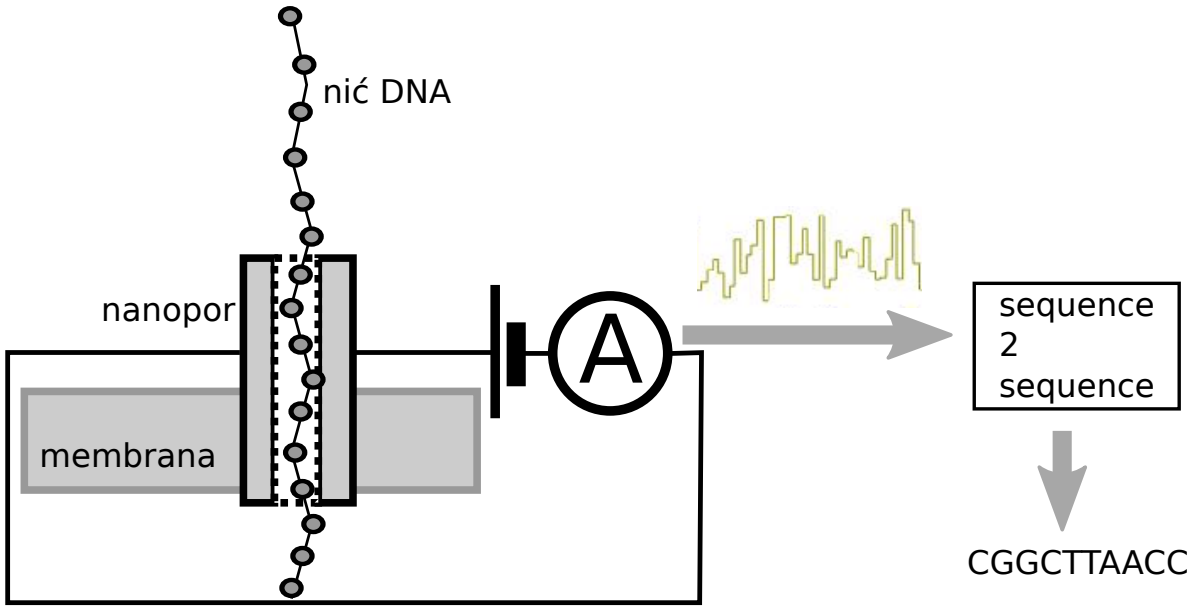
Illumina

- sekwencjonowanie $\approx 100x$ tańsze niż Roche 454
- $\approx 10x$ bardziej wydajne
- obecnie (2024) dominuje na rynku

Przykład: Illumine HiSeq

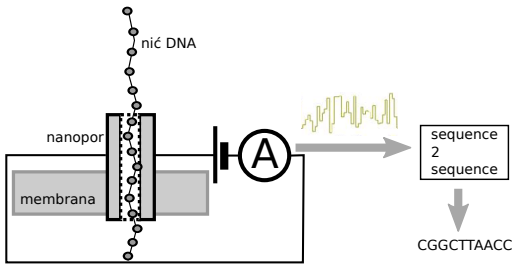
Sekwencjonowanie trzeciej generacji

Pomiar pola elektrycznego w nanoporach



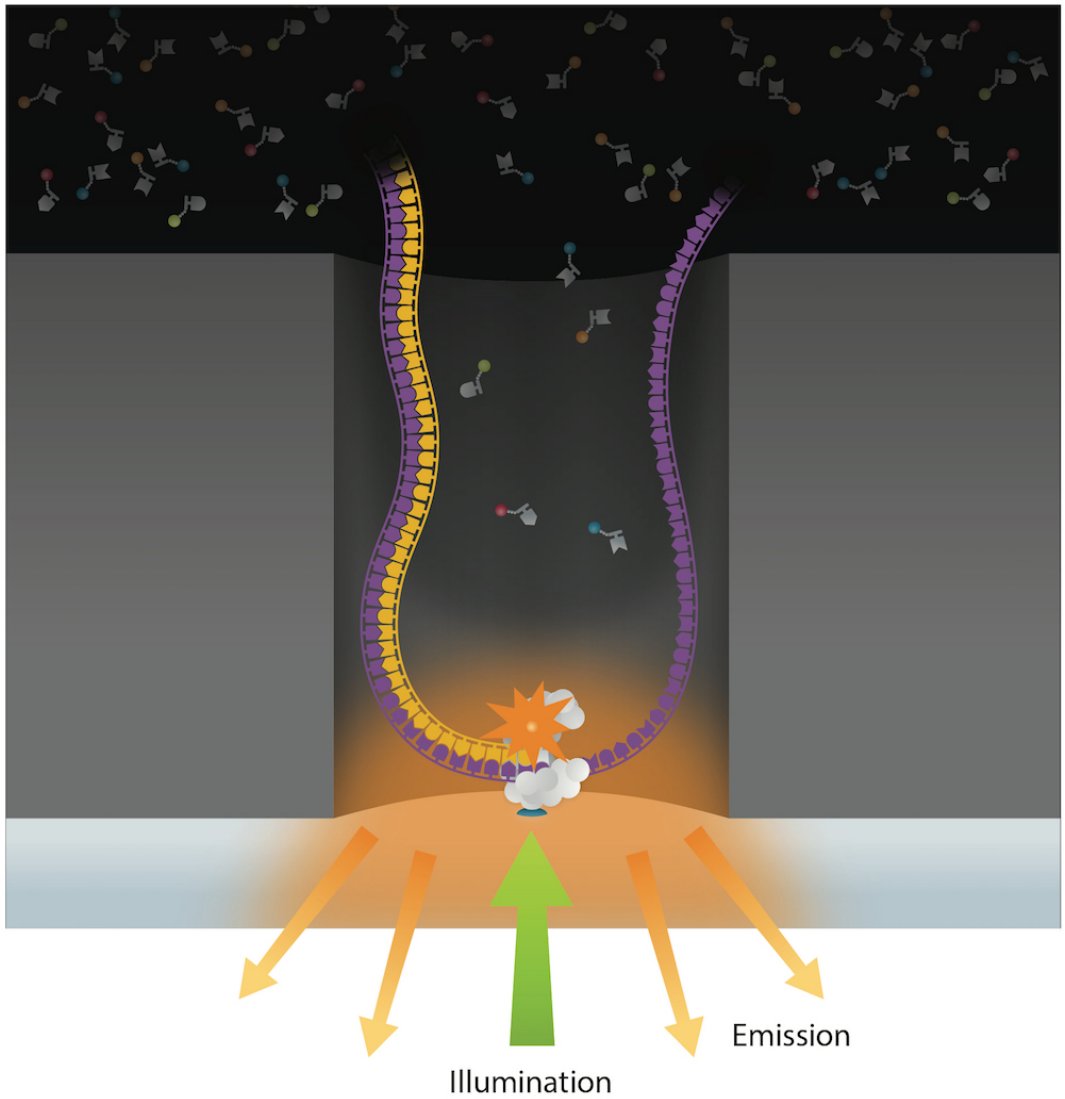
- używany Oxford Nanopore
- długie odczyty
- 15% błędów

Basecalling - poprawa jakości wyników



- wejście - szereg czasowy, próbki prądu
- metoda - analiza próbek dla kilku sąsiednich symboli
  - ukryte modele markowa
  - sztuczne sieci neuronowe
- osiągnięcia - błąd zredukowany do 3% (Guppy)

Synteza pojedynczych molekuł w czasie rzeczywistym



- reakcja w nano-komorach
- wyznakowane dNTP gdy są przyłączane, emitują światło
- używany PacBio

Sekwenatory PacBio i Oxford Nanopore

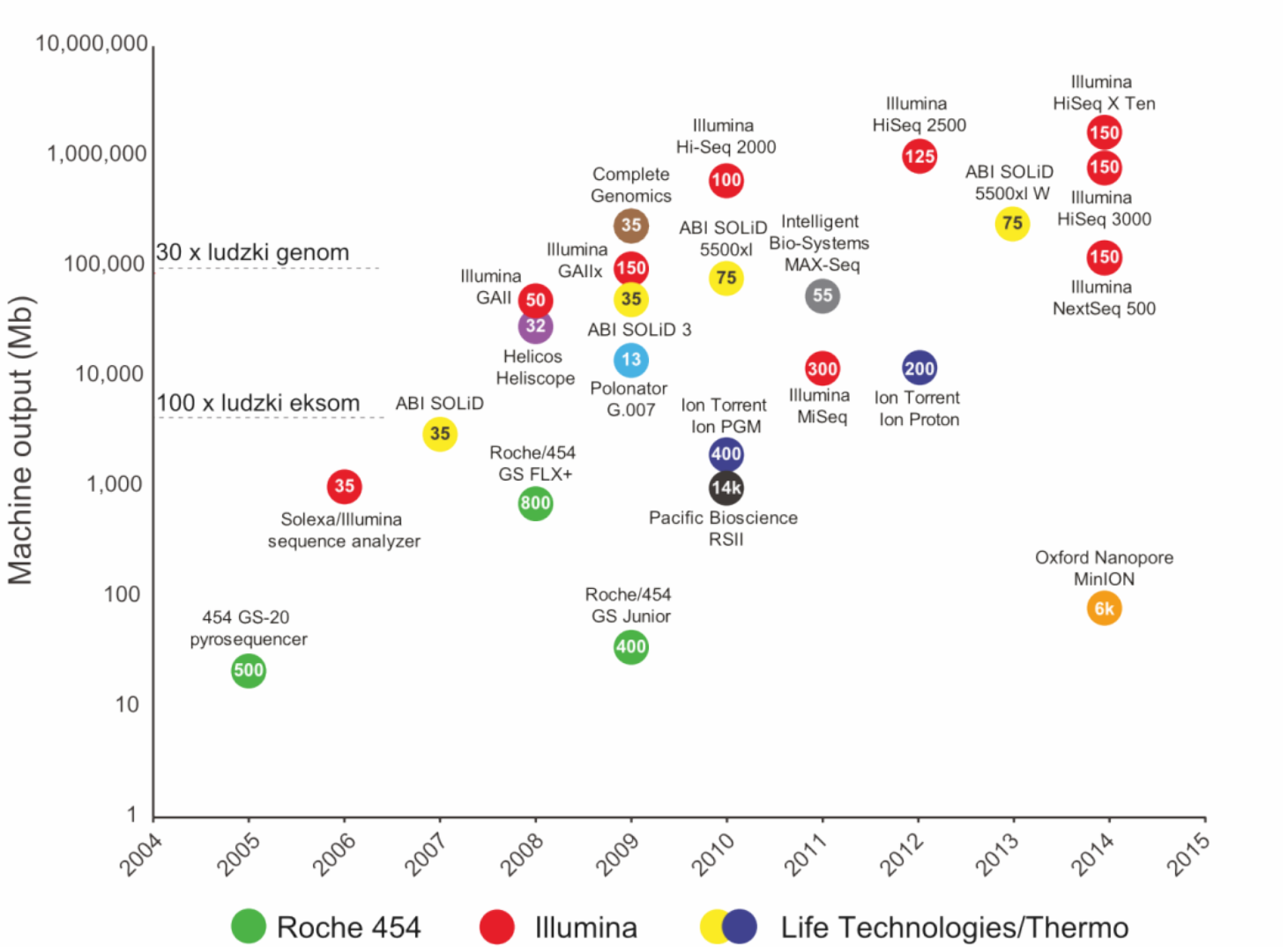
Podsumowanie

Sekwencjonowanie DNA - urządzenia

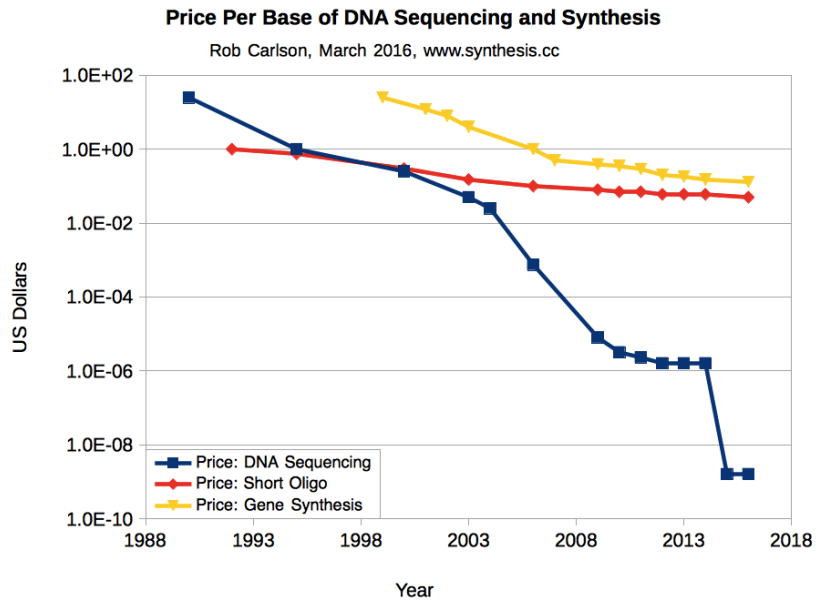
nazwa	długość odczytu	koszt *Mbp*	dokładność
pierwszej generacji			
Sanger	2000bp	2400$	99.999%
drugiej generacji			
Roche 454	500bp	10$	99.9%
Ilumina HiSeq	200bp	0.07$	98%
BGISEQ-500	200bp	0.05$	98%
drugiej generacji miniaturowe			
Ilumina, MiSeq	150bp	7$	98%
trzeciej generacji			
PacBio RS	10kbp	10$	85%
Nanopore	20kbp	2$	80%

wzrost 21% rocznie, 4.5 mld. $ (2013), 6.6 mld. $ (2016)

Sekwenatory nowej generacji



Koszty sekwencjonowania



- obecnie (2024) koszt odczytu to 600$ (BGI)
- $\sim 2 * 10^9$ do 2025 (25% populacji), w tym duże projekty np. 1+ Million Genomes in EU,
- jeden genom to 360GB odczytów (FASTQ)
- jeden genom to plik 15GB

Opis sekwencji DNA i RNA

International Union of Pure and Applied Chemistry (IUPAC)

A	adenine	R	G A (purine)
C	cytosine	Y	T C (pyrimidine)
G	guanine	K	G T (keto)
T	thymine	M	A C (amino)
U	uracil	S	G C
		W	A T
B	G T C	N	A G C T (any)
D	G A T		
H	A C T		
V	G C A		

Format danych FASTA, FASTQ

FASTA: nagłówek (1 linia), sekwencja (po 70 znaków)

```
>AB000263 |acc=AB000263|descr=Homo sapiens mRNA for ...|len=368
ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCC
CCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGC
CTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGCCGGGCCCCTCATAGGAGAGG
AAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCGCCGGGACAGAATGCC
CTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGAATGCTCACGCAAG
TTTAATTACAGACCTGAA
```

FASTQ: nagłówek (1 linia), sekwencja (po 70 znaków), jakość każdego symbolu

```
@SEQID description
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SEQID description
IIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

Format FASTQ, jakość

Jakość (Phred quality score) obliczana na podstawie prawdopodobieństwa, że symbol jest błędny (e):

$$Q = -10 \log_{10} e$$

zapisywana jako znak ASCII o kodzie $X = Q + 64$, kod 64 to @

Q	e	dokładność	znak ASCII
10	0.1	90%	**J**
20	0.01	99%	**T**
30	0.001	99.9%	**^**
40	0.0001	99.99%	**h**
50	0.00001	99.999%	**r**
60	0.000001	99.9999%	**|**

Asemblery DNA

Algorytmy do składania sekwencji (asemblacja)

Algorytmy (grafowe) przekształcające ciąg pod-sekwencji (o losowych przesunięciach) nazywanych „odczytami” na zbiór sekwencji (nazywanych kontigami lub kontigami sekwencyjnymi) i zbiór przerw.

Typowe algorytmy:

- graf pokrycia (overlap-layout-consensus, OLC)
- tzw. graf de Bruijn-a (De Bruijn graph, DBG)

Wyjściem nie jest to jedna sekwencja (jeden kontig), ponieważ występują:

- sekwencje powtarzające się dłuższe niż odczyt (OLC) lub rząd grafu (DBG)
- błędy odczytu, błędy sekwencjonowania

Dostępnych jest ok. 50 asemblerów.

Asembler, algorytm OLC

- obliczanie podobieństwa (analiza wszystkich par odczytów)
- konstrukcja grafu
- znajdowanie ścieżki Hamiltona w grafie (problem NP-zupełny), więc heurystyki

Ścieżka Hamiltona- przechodzi przez każdy wierzchołek grafu dokładnie raz

Złożoność: $O(|V|^k)$, gdzie k - maksymalna ilość krawędzi związanych z grafem

aplikacje[1]: Celera, Arachne, CAP, PCAP, Newbler, CABOG, Edena, Shorty

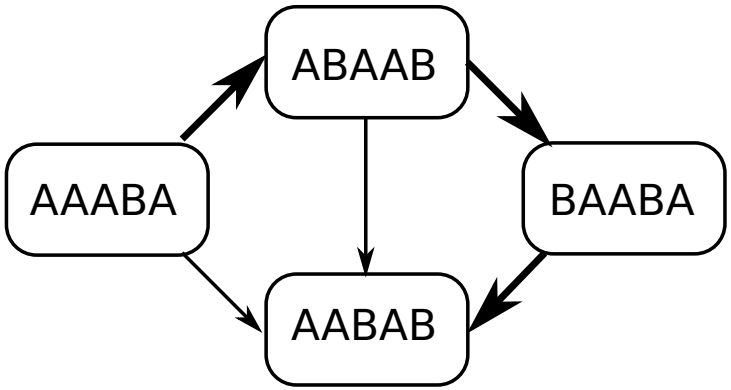
[1] J.Miller and oth, Assembly algorithms for next-generation sequencing data

Algorytm OLC - przykład

4 odczyty, każdy l = 5, brak błędów odczytu, brane podobieństwo z oceną > 2,

a) AAABA
b) ABAAB
c) BAABA
d) AABAB

-	**a**	**b**	**c**	**d**
a	-	3	2	4
b	-	-	4	3
c	-	-	-	4
d	-	-	-	-




```
A  A  A  B  A  -  -  -  -
-  -  A  B  A  A  B  -  -
-  -  -  B  A  A  B  A  -
-  -  -  -  A  A  B  A  B
```

wynik: **AAABAABAB**

Asembler, algorytm DBG

Aplikacje wykorzystują podgraf grafu de Bruijna:

- nie wymagają oceny wszystkich par odczytów,
- wymagają wyznaczenia ścieżki Eulera (problem rozwiązywany wielomianowo).

Ścieżka Eulera- przechodzi przez każdą krawędź grafu dokładnie raz

Złożoność: algorytm Fleury'ego, $O(|E|^2)$, algorytm Hierholzer'a $O(|E|)$

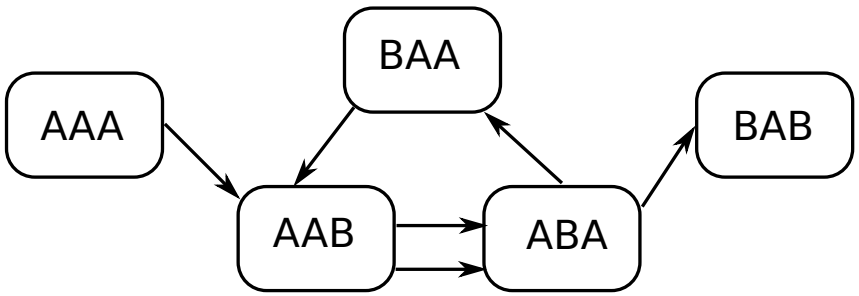
aplikacje[2]: Euler, Velvet, ABySS, AllPaths, SOAPdenovo

[2] J.Miller and oth, Assembly algorithms for next-generation sequencing data

Algorytm DBG - przykład

4 odczyty, każdy l = 5, graf de Bruijna rzędu 4,

a) `AAABA`: `AAAB`, `AABA`
b) `ABAAB`: `ABAA`, `BAAB`
c) `BAABA`: `BAAB`, `AABA`
d) `AABAB`: `AABA`, `ABAB`



wynik: `AAABAABAB`

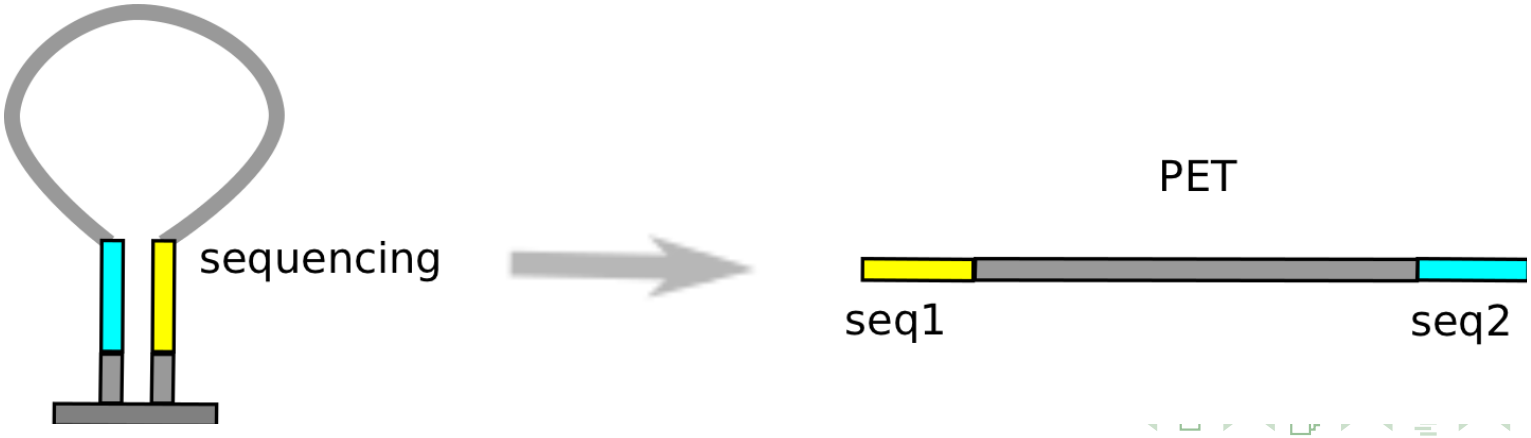
Kontigi sekwencyjne i kontigi fizyczne (ang. *scaffold)*

Kontig, kontig sekwencyjny - ciągła sekwencja uzyskana ze złożenia nachodzących na siebie fragmentów

Powody uzyskiwania dziur:

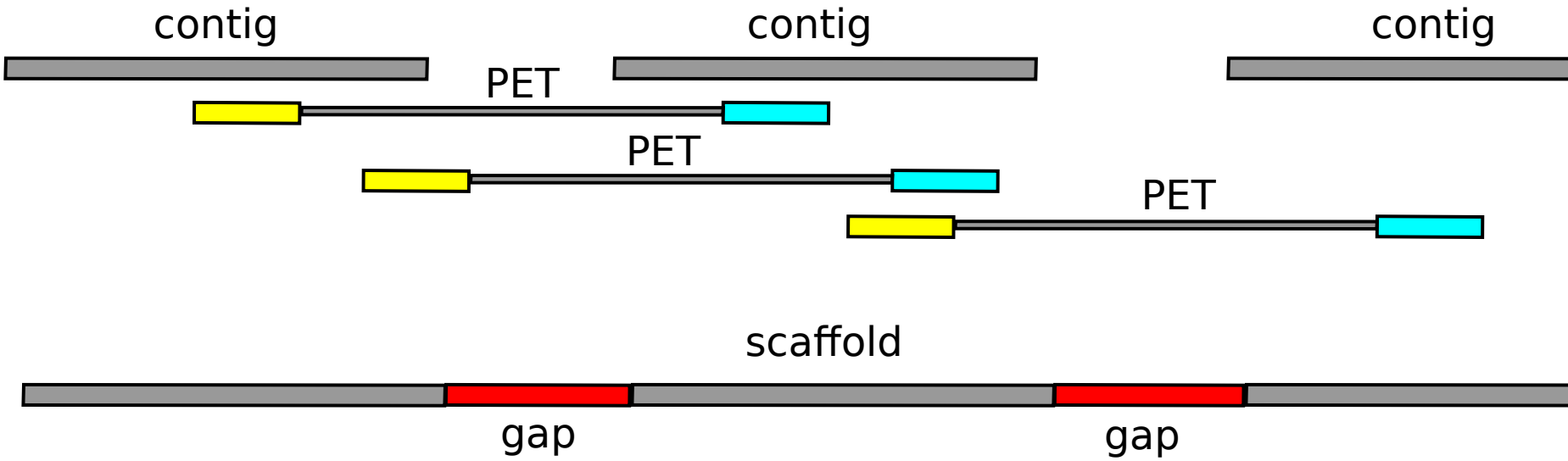
- sekwencje powtarzające się dłuższe niż odczyty
- niepełne pokrycie badanej cząsteczki losowymi fragmentami

Dodatkowa informacja pozwala ustalić orientację kontigów i je uszeregować, uzyskujemy kontig fizyczny



PET (paired-end tag) - sekwencje sparowanych końców

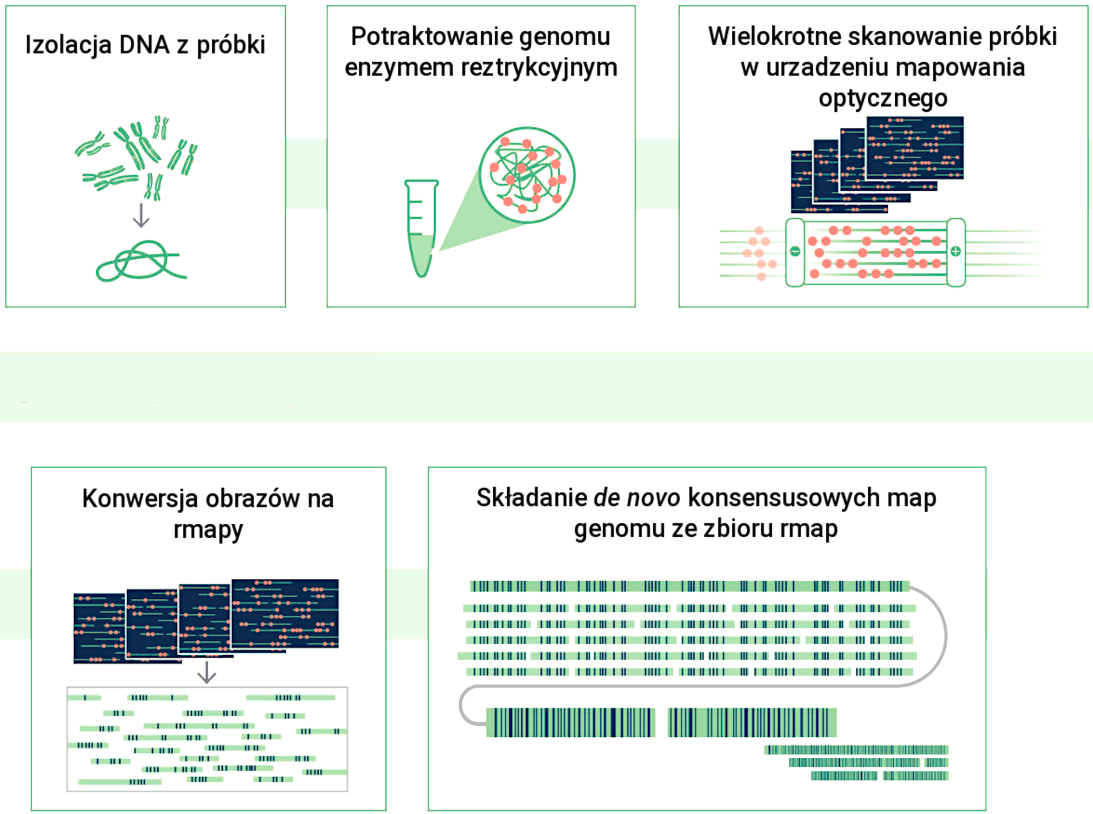
- znana długość (np. 3000 nt)
- znane sekwencje na obu końcach



kontig fizyczny (ang. scaffold) - sekwencja (niekoniecznie ciągła) uzyskana za pomocą sekwencji sparowanych końców

Wykorzystywane algorytmy przeszukiwania przestrzeni z ograniczeniami (CSP – Constraint satisfaction problems)

Mapowanie optyczne (optical mapping)



- położenie markerów (sekwencji rozpoznawanych przez enzymy restrykcyjne)
- jeden odczyt (rmapa), fragment 150 kbp – 2 Mbp

Składanie sekwencji - podsumowanie

- uwzględnienie błędnych sekwencji zwiększa złożoność problemu
- stosuje się dodatkowe odczyty (inny rodzaj doświadczenia, sparowane końce) do łączenia kontigów
- problemem są długie sekwencje powtarzające się (dłuższe niż odczyt)
- techniki sekwencjonowania 2 generacji dostarczają dużej ilości odczytu
- istniejące algorytmy wymagają dużej ilości pamięci (64GB) i mają wiele parametrów

Dziękuję