analiza danych z wysokoprzepustowego sekwencjonowania

Dr Marek Żywicki (Marek.Zywicki@amu.edu.pl)

era pre-genomiczna

VS

era post-genomiczna

era pre-genomiczna

IDENTYFIKACJA GENÓW

Białko

Sekwencja

Gen

era pre-genomiczna

BADANIE POZIOMU EKSPRESJI

Białko:
- techniki immunologiczne
- techniki elektroforetyczne
- spektrometria mas

RNA:
- techniki elektroforetyczne
- qRT-PCR

pierwszy przełom - GENOM

1980 – pierwszy zsekwencjonowany wirus

1984 – sekwencja wirusa EBV

1997 – wstępna sekwencja *Escherichia coli* (5MB) oraz genom drożdży (*Sascharomyces cerevisiae*) – 7 lat, 100 ośrodków, 600 naukowców

2001 – dwie wstępne sekwencje genomowe człowieka (HUGO/HGP oraz CELERA)

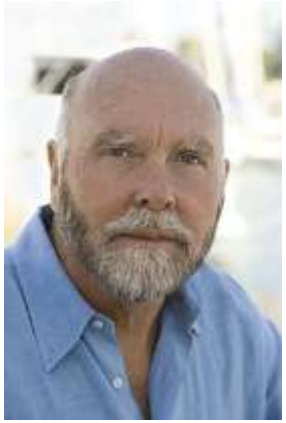
Human Genome Project

1990 – NIH przaznaje **3mld dolarów** na projekt zsekwencjonwania ludzkiego genomu w ciągu **15 lat**

1999 – pierwszy zsekwencjonowany chromosom (22)

2001 – zakończono 90% sekwencji, publikacja wyników

2003 – ostateczne ukończenie 99% sekwencji



“We will sequence HG in 3 years.”

- 1998 Craig Venter otwiera firmę „Celera”
- 2000: HGP & Celera ogłaszają pierwszy draft genomu

  (zapowiadają wspólną publikację)
- 2001: Publikacja wyników jednak osobno: HGP: Nature, 15 luty; Celera: Science, 16 luty
- 2007: Craig Venter publikuje swój diploidalny genom

  (pierwszy genom zsekwencjonowany w całości!)

2008: Start programu “The 1000 Genomes Project”, mającego na celu zbadanie różnorodności genetycznej człowieka

<u>Cel:</u> zsekwencjonować około 1000 genomów ludzkich w ciągu 3 lat!

era post-genomowa

znamy sekwencje nukleotydowe **wszystkich** ludzkich genów

wiemy że nie jesteśmy tak unikalni:

- 23 000 genów – podobnie jak mysz czy nicienie
- 98% podobieństwa sekwencji do myszy
- tylko 7% rodzin białkowych jest specyficzna dla kręgowców

Gdzie szukać różnic???

era post-genomowa

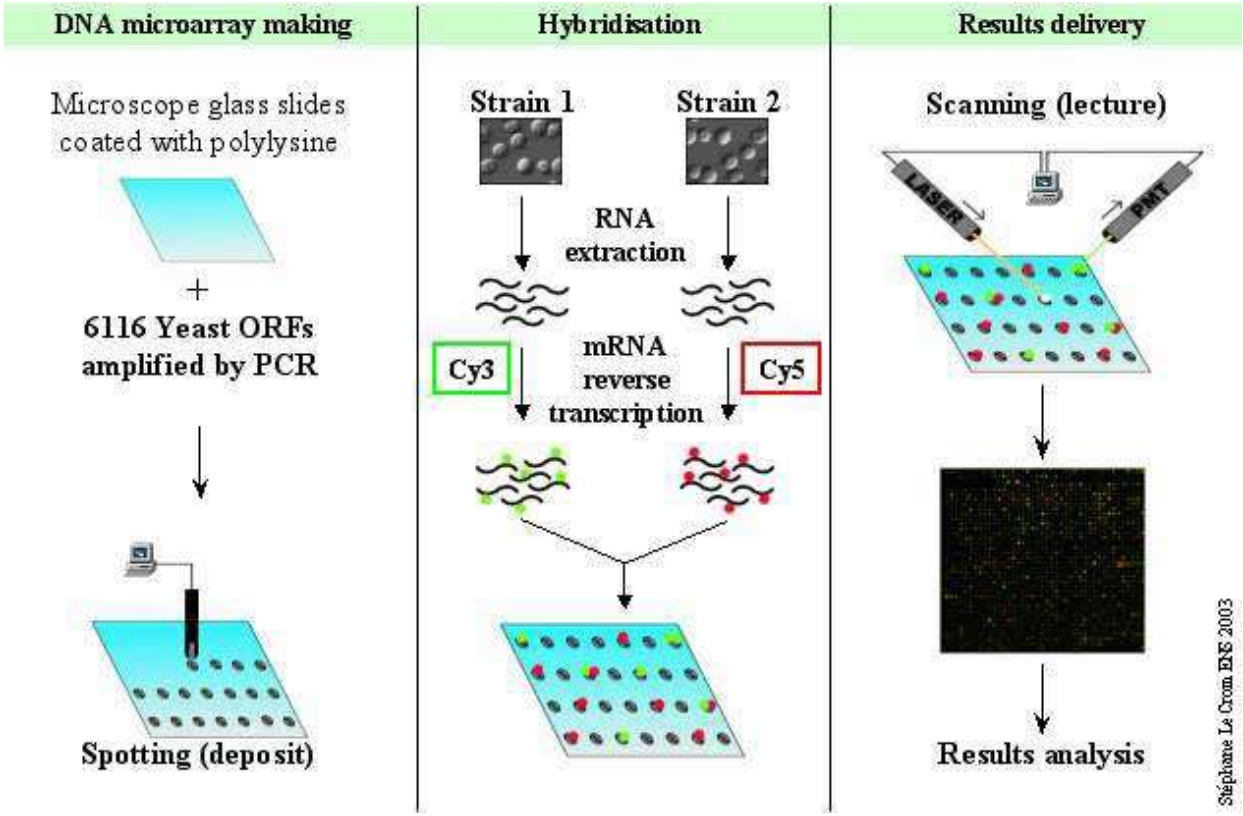
1 gen na raz to za mało!

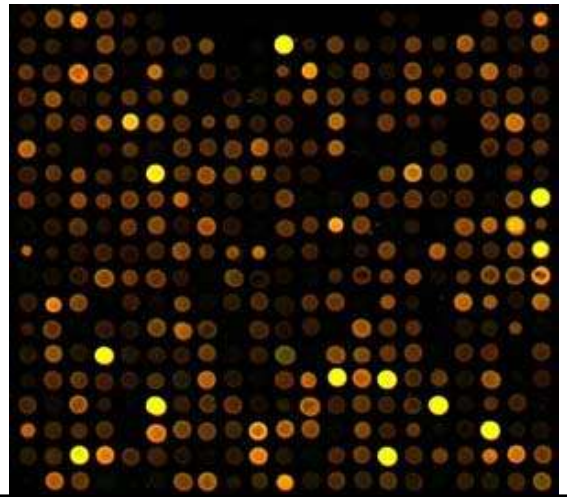
Musimy widzieć cały obraz!

Czas na techniki wysokoprzepustowe!

mikromacierze

odwrócony dot-plot + trochę elektroniki!





mikromacierze

- poziom ekspresji genów
- identyfikacja nowych genów
- alternatywne składanie genów
- analiza interakcji białek z kwasami nukleinowymi (CHIP-Chip)
- sekwencjonowanie
- ....

drugi przełom - sekwencjonowanie

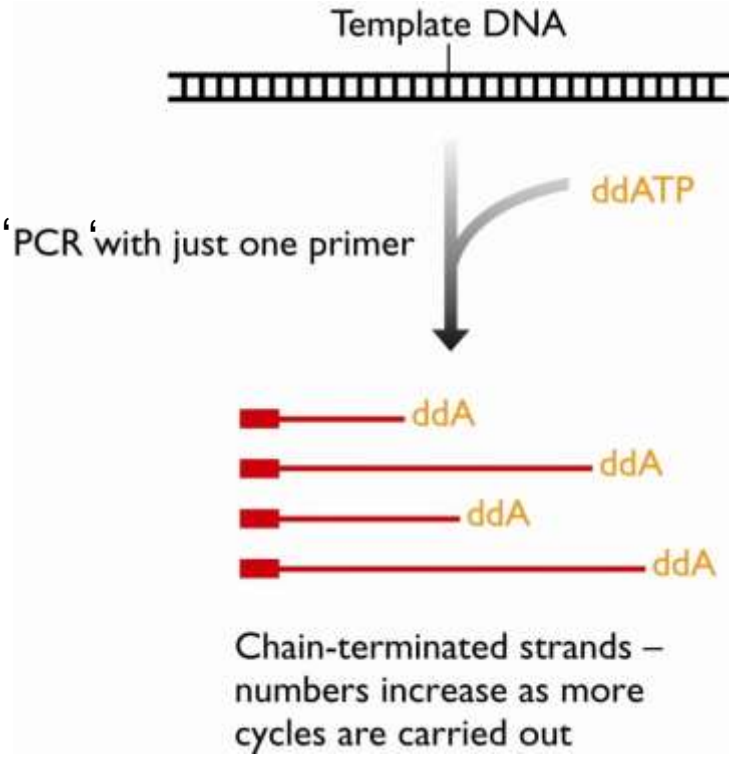
1990-2003: Human Genome Project

> 3mld USD, 13 lat, >3000 naukowców

2013: Next Generation Sequencing

1 000 USD, 1 dzień, 1 osoba

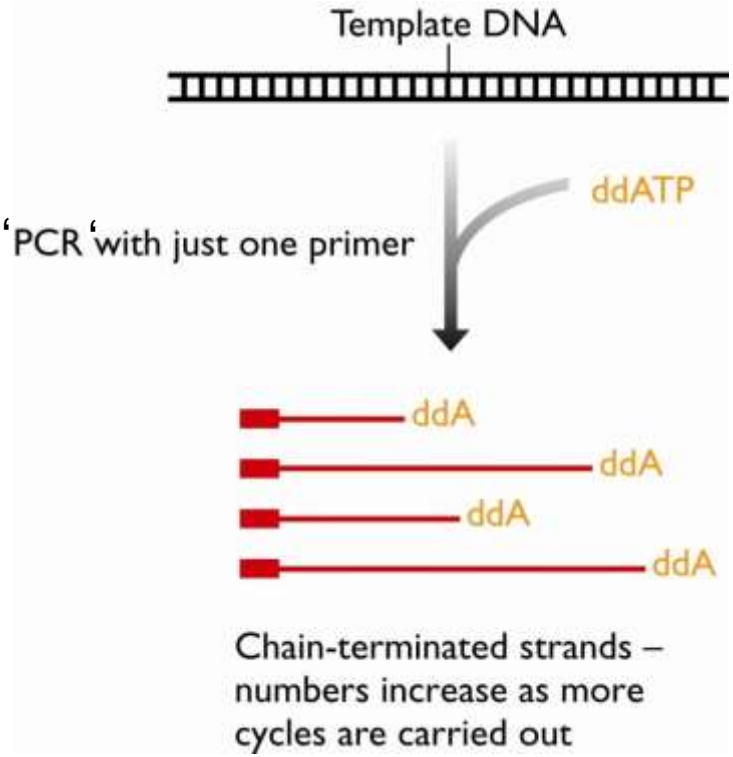
Cycle-Sequencing



Advantages

- ds DNA can be sequenced
- only little amounts of DNA required

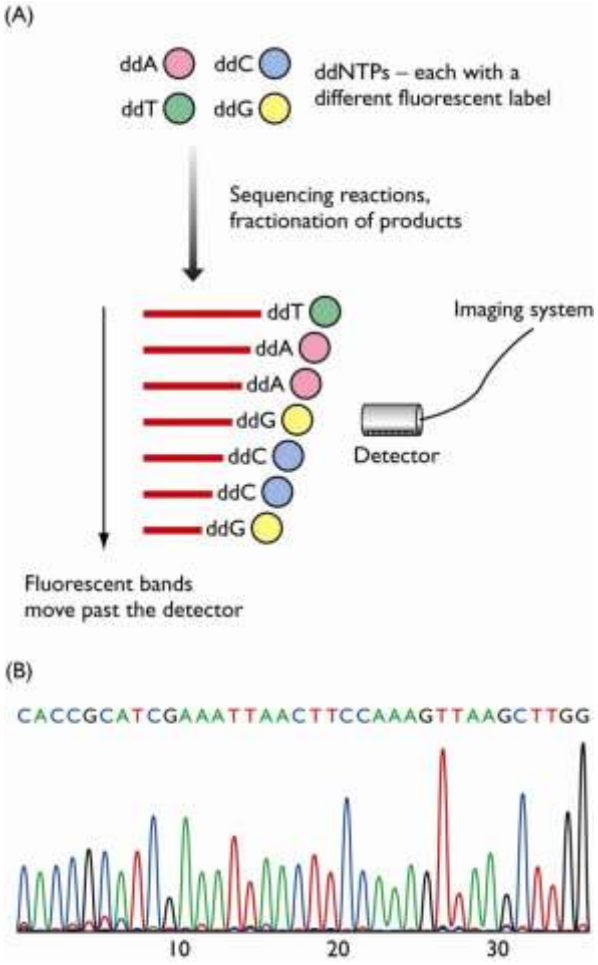
Cycle-Sequencing



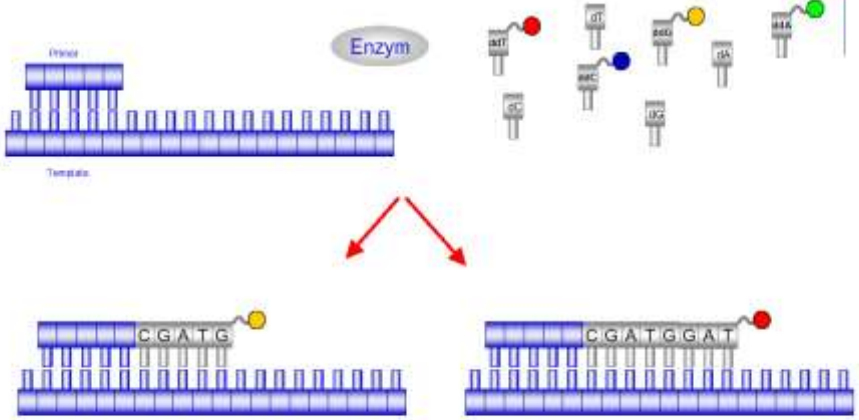
Advantages

- ds DNA can be sequenced
- only little amounts of DNA required

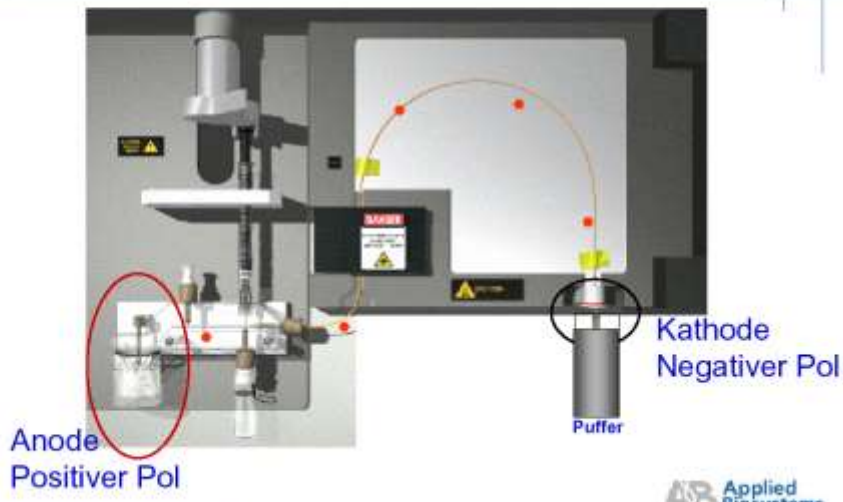
Automated fluorescence sequencing

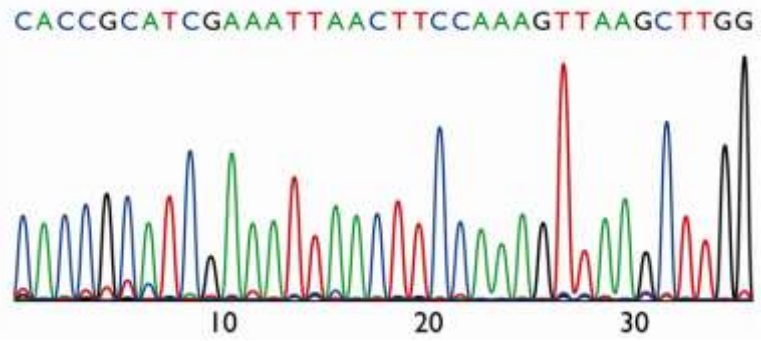


Dye termination reaction



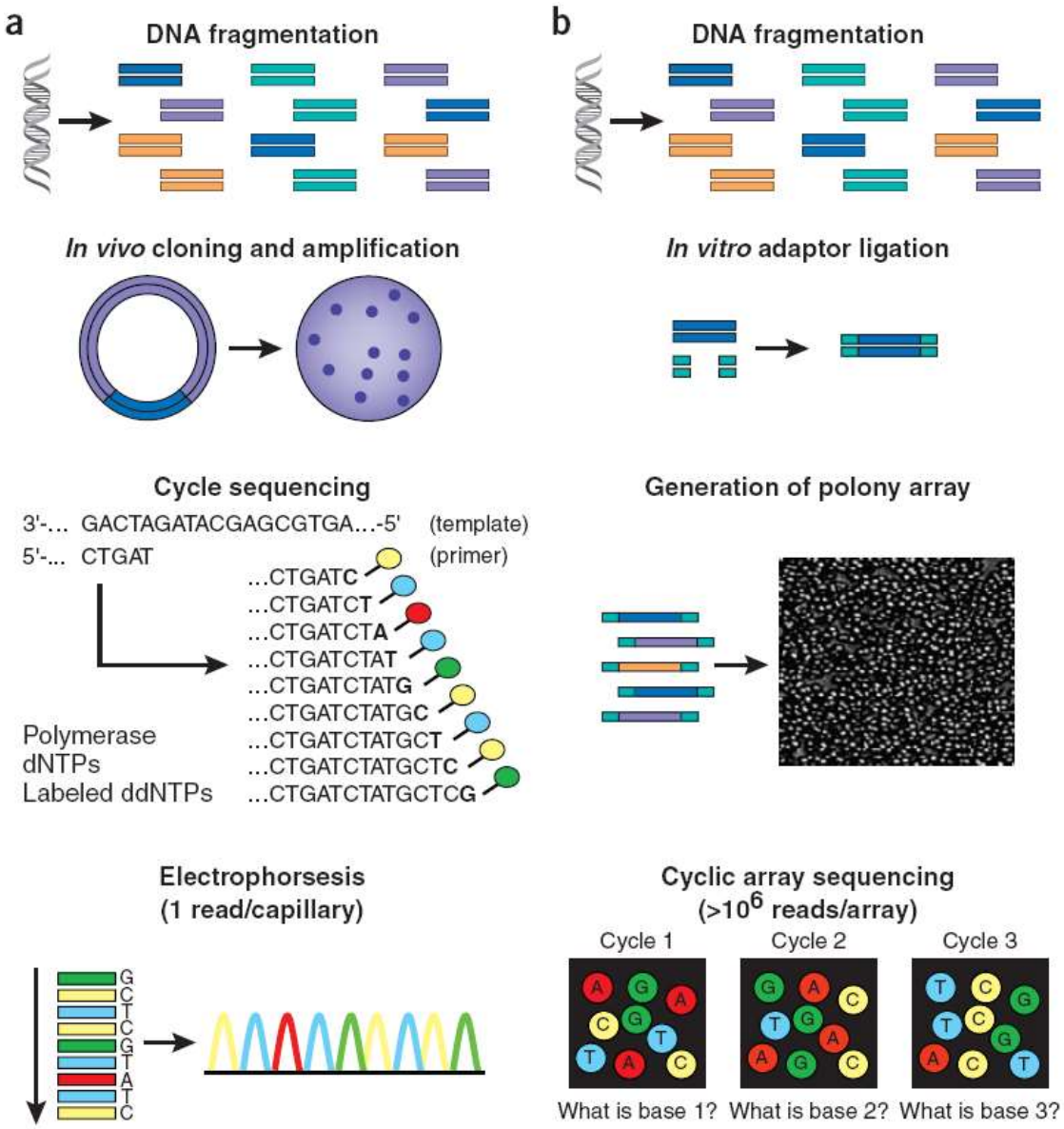
Capillary electrophoresis





Applied Biosystems

klucz do sukcesu - paralelizacja



Shendure J. & Ji H., *Nature Biotech* 26, 2008

Podstawowe zalety technik wysokoprzepustowego sekwencjonowania w porównaniu do techniki Sangera:

- konstrukcja biblioteki *in vitro*
- specyficzna amplifikacja DNA *in vitro*
- sekwencjonwanie na płytkach → paralelizacja!!!

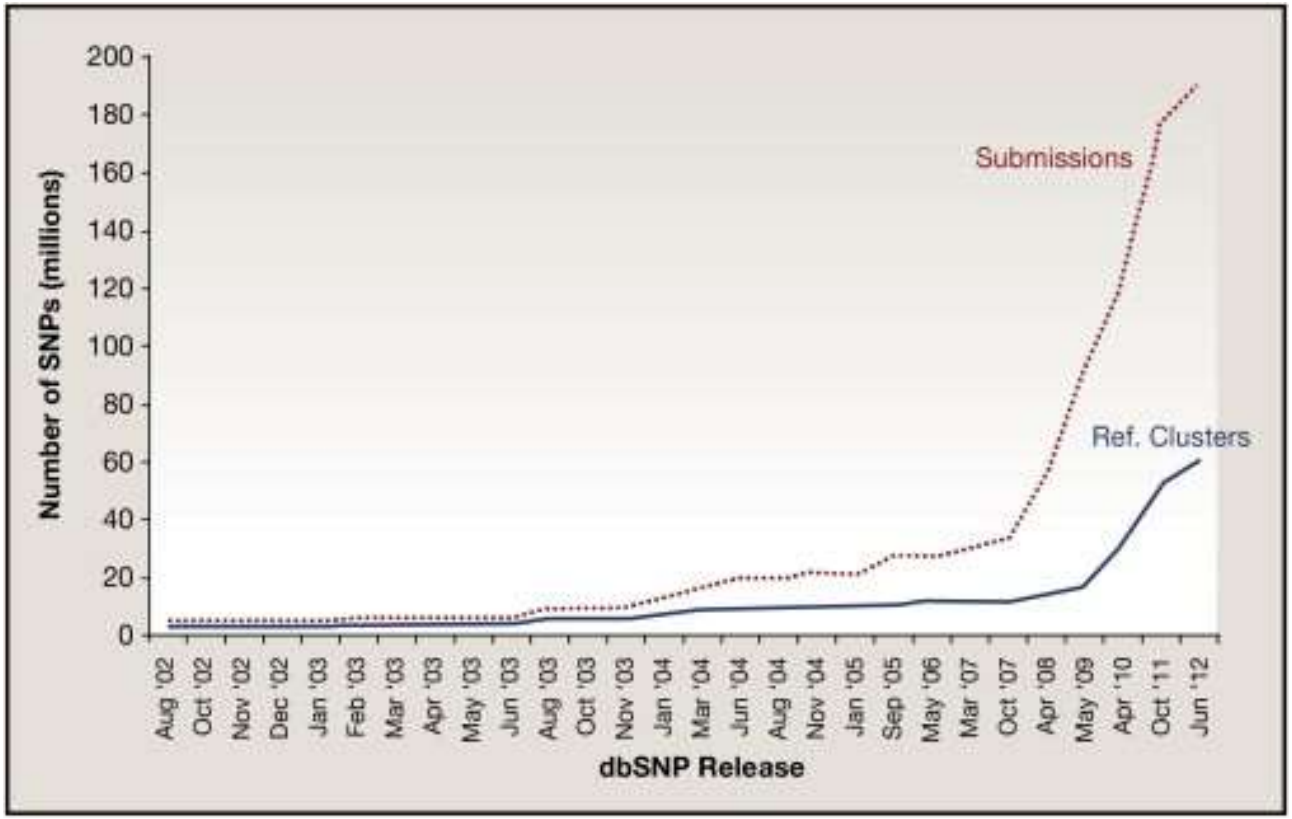
  (setki milionów odczytów w jednej reakcji!!!)
- immobilizacja reakcji na płytce → odczynniki można dostarczać jednocześnie do wszystkich reakcji
- bardzo niski koszt (50 - 10000 razy tańsze)

drugi przełom - sekwencjonowanie

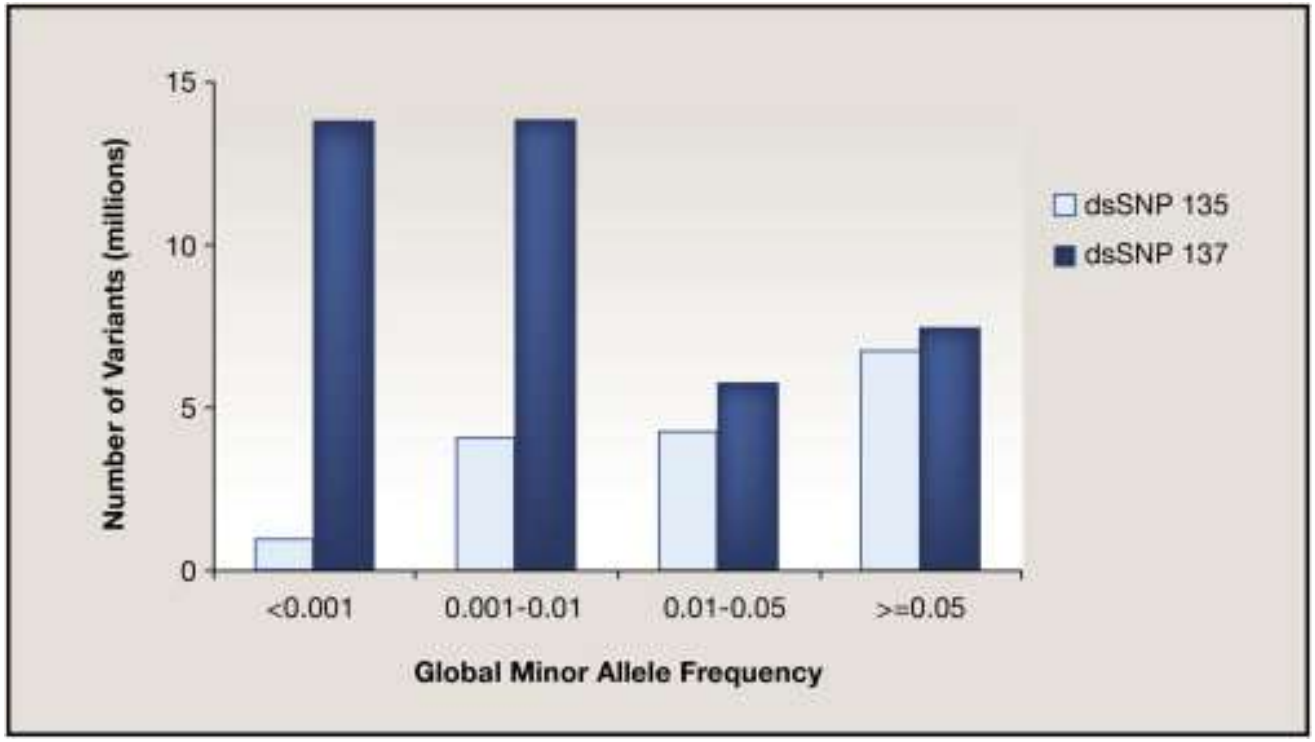
drugi przełom - sekwencjonowanie

- sekwencjonowanie genomów
- identyfikacja różnic genetycznych
- identyfikacja markerów chorobowych
- sekwencjonowanie transkryptomów
- poziom ekspresji genów
- identyfikacja nowych genów
- alternatywne składanie genów
- analiza interakcji białek z kwasami nukleinowymi (CHIP-Seq)
- analiza metylacji DNA
- analiza modyfikacji RNA
- analiza struktury drugorzędowej RNA
-....

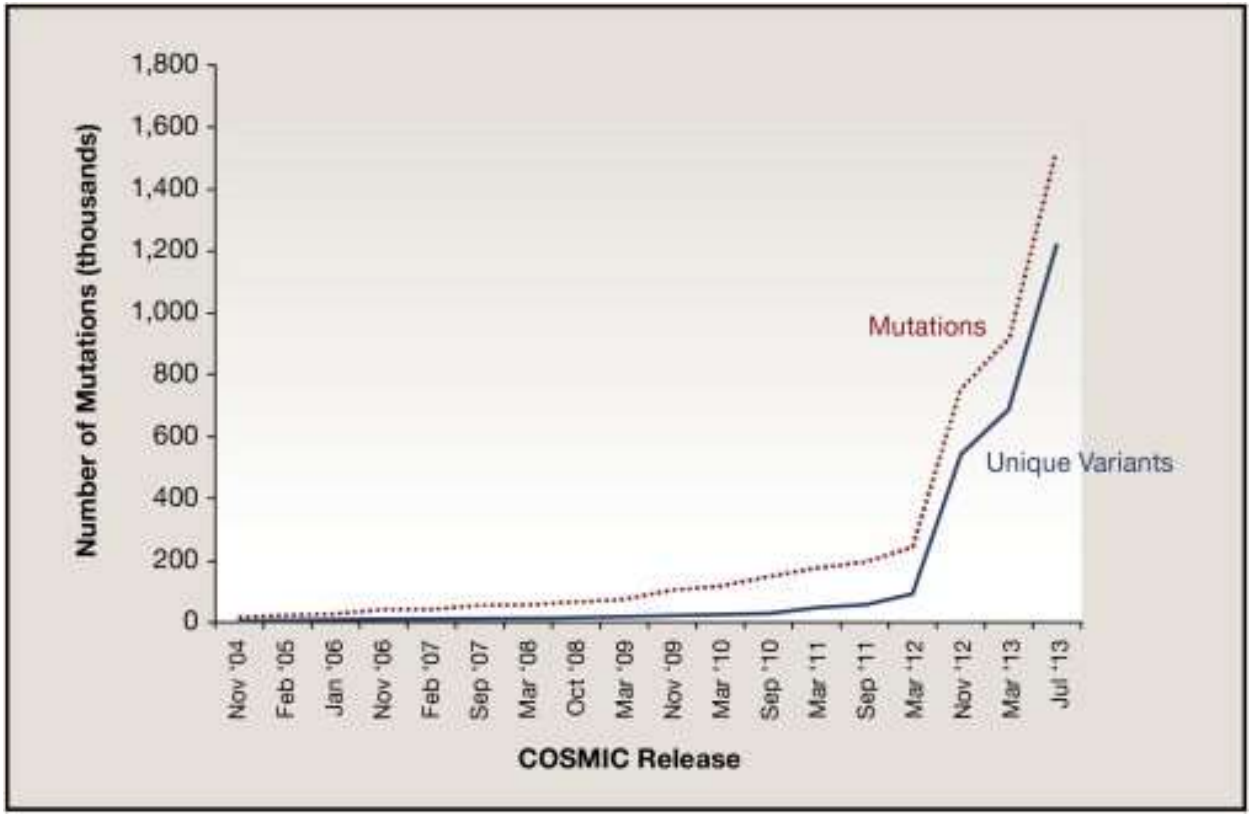
wyniki – różnorodnośc genetyczna



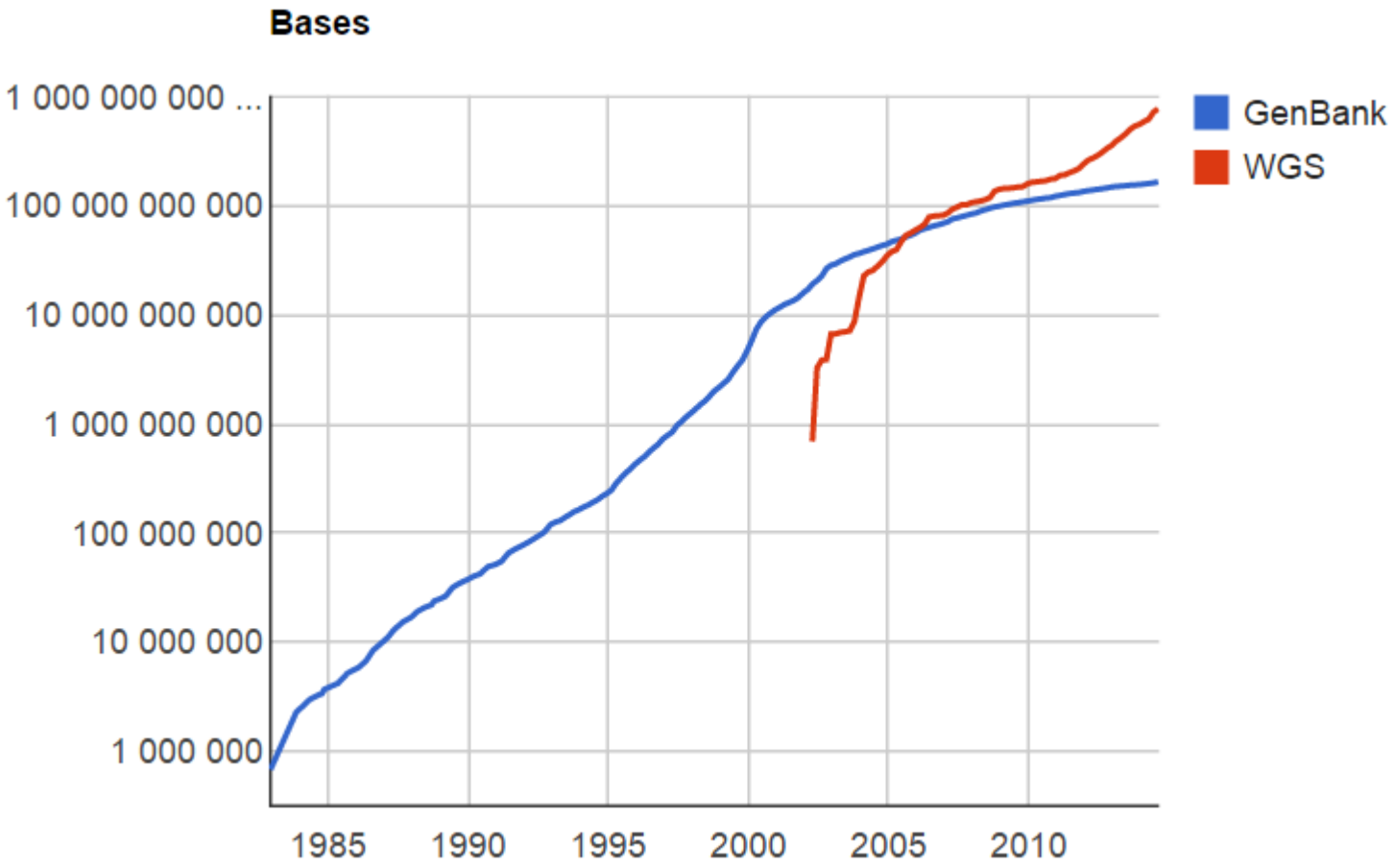
wyniki – różnorodnośc genetyczna



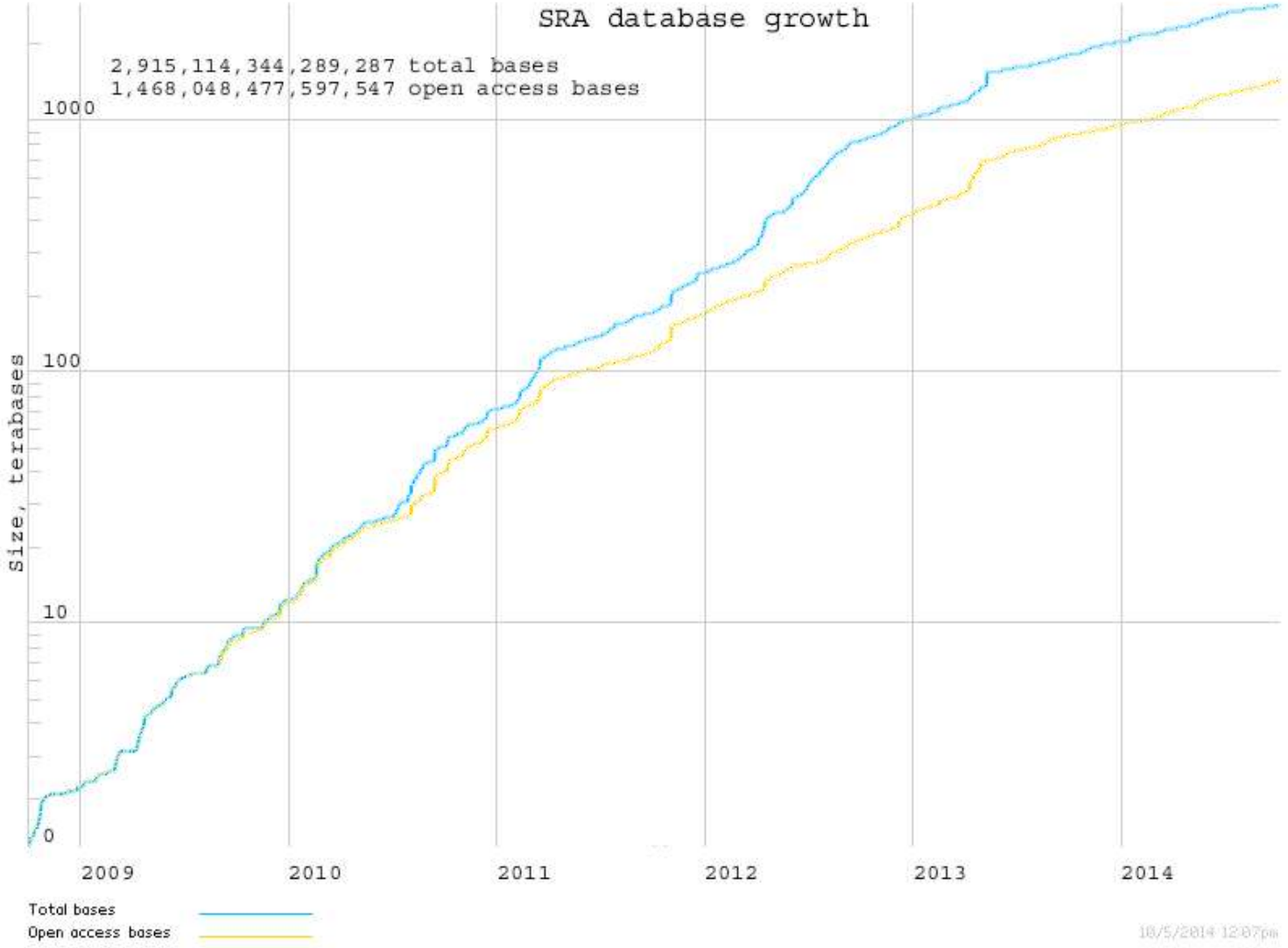
wyniki – różnorodnośc genetyczna



Olbrzymie ilości danych!!!
GenBank



Olbrzymie ilości danych!!!
NCBI Short Read Archive (SRA)



Olbrzymie ilości danych!!!
Bioinformatyka

Nowe narzędzia

Nowe podejście do analizy danych

Nowy poziom wydajności

Analiza danych z wysokoprzepustowego sekwncjonowania

Sekwencjonowanie:

- analiza jakości
- mapowanie
- składanie transkryptów oraz genomów
- różnicowa ekspresja genów
- identyfikacja alternatywnego składania mRNA
- koimmunoprecypitacja
- identyfikacja różnic genetycznych

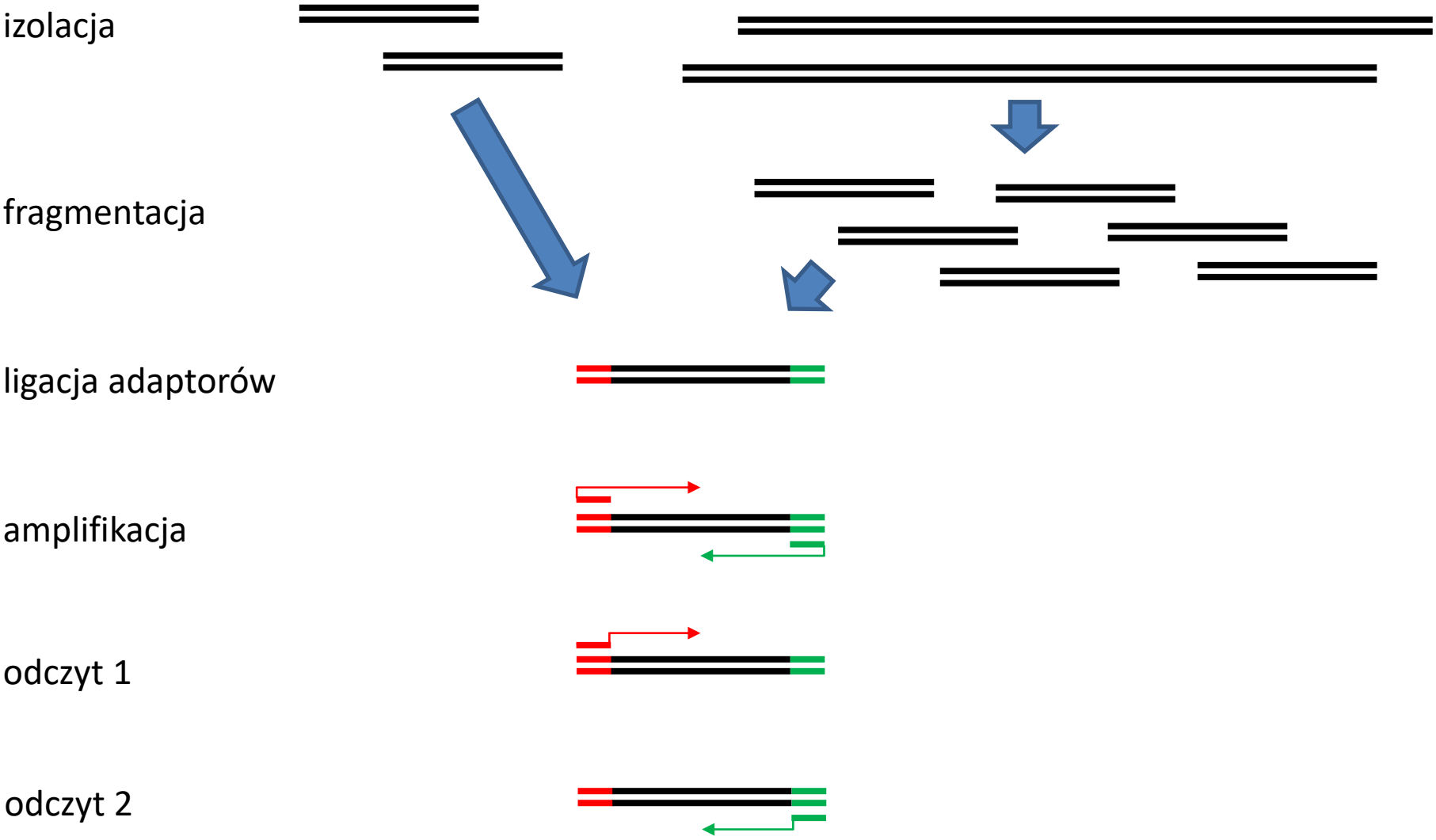
Etapy sekwencjonowania

1. Przygotowanie DNA

2. Amplifikacja

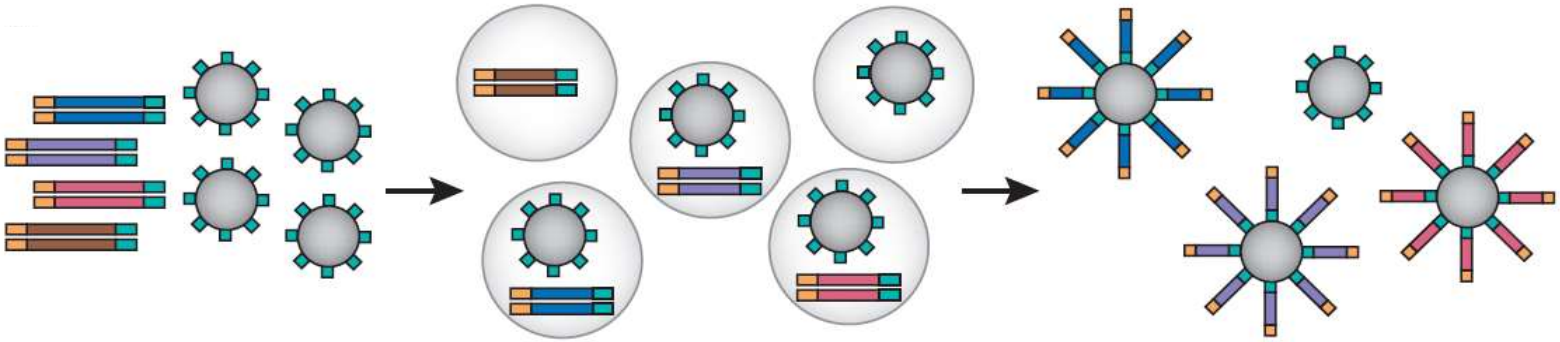
3. Odczyt sekwencji

1. Przygotowanie DNA



pirosekwencjonowanie (454)

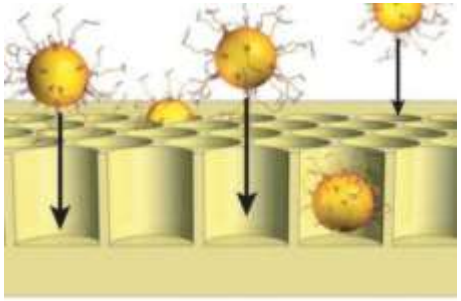
PCR w emulsji:



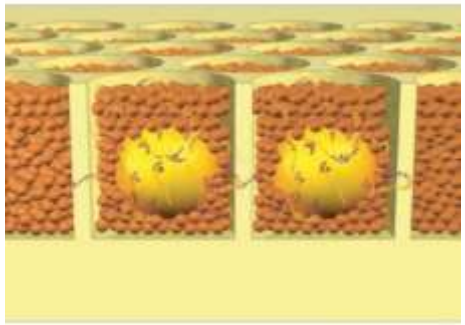
- DNA z dołączonymi adaptorami jest mieszany z kulkami (28 µm) posiadającymi jeden z primerów do PCR zakotwiczony na powierzchni
- stosunek: poniżej jednej cząsteczki DNA / kulkę
- tworzona jest emulsja wodna w cieczy oleistej
- faza wodna zawiera: DNA, kulki, biotynylowany drugi primer, dNTP, polimerazę DNA
- prowadzona jest reakcja PCR
- kulki które posiadały matrycę w swojej kropli i teraz posiadają zakotwiczone zamplifikowane DNA są wyławiane za pomocą streptawidyny
- denaturacja, w wyniku której otrzymujemy kulki z jednoniciowym zamplifikowanym DNA

pirosekwencjonowanie (454)

- kulki z DNA są umieszczane na płytce PicoTiterPlate (PTP) (> 400.000 studzienek)

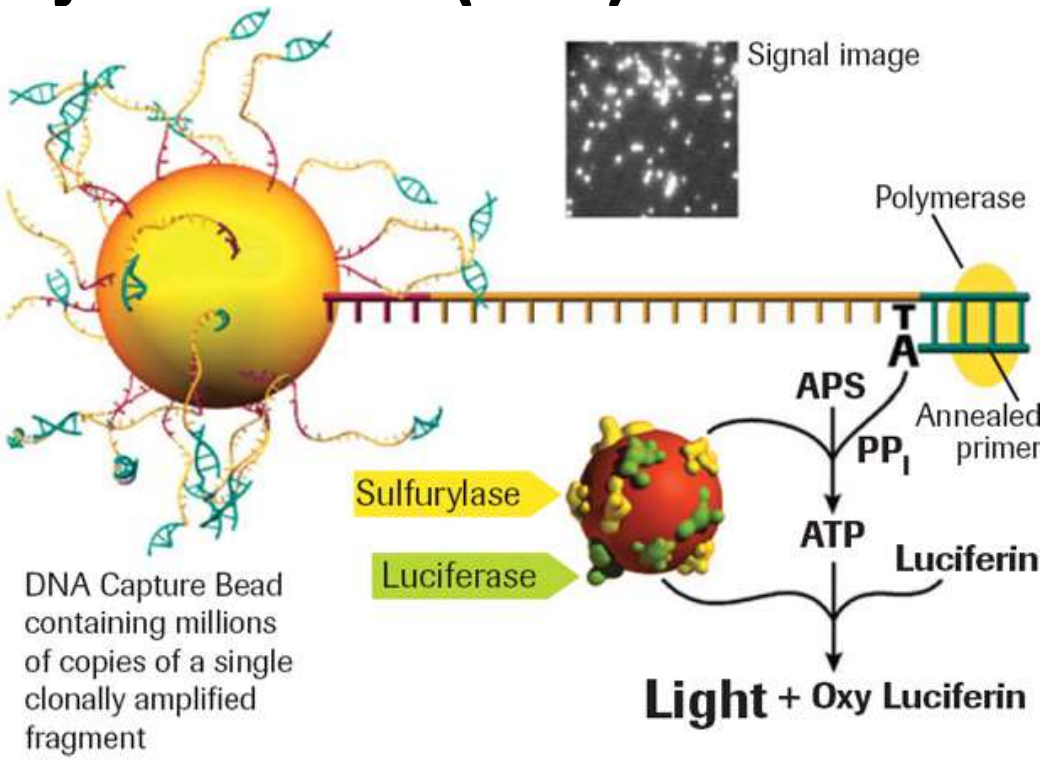


- studzienki zalewane są roztworem zawierającym:
  - polimerazę DNA oraz odczynniki niezbędne do przeprowadzenia reakcji sekwencjonowania
  - kuleczkami zawierającymi enzymy niezbędne do wzbudzenia sygnału: sulfurylazą oraz lucyferazą



pirosekwencjonowanie (454)

Sekwencjonowanie:



- hybrydyzacja startera
- płytka umieszczana jest przed kamerą
- płytka jest przepłukiwana kolejnymi nukleotydami w ustalonym porządku
- w studzienkach, gdzie dany nukleotyd jest komplementarny następuje jego przyłączenie przez polimerazę DNA – **Sekwencjonowanie przez syntezę!!!**
- podczas przyłączenia każdego nukleotydu następuje uwolnienie PPi -> utworzenie ATP przez sulfurylazę -> utlenienie lucyferyny przez lucyferazę z użyciem ATP
- skutkuje to emisją światła o natężeniu proporcjonalnym do liczby przyłączonych nukleotydów
- nie wykorzystane nukleotydy usuwane są przez enzym apyrazę

pirosekwencjonowanie (454)

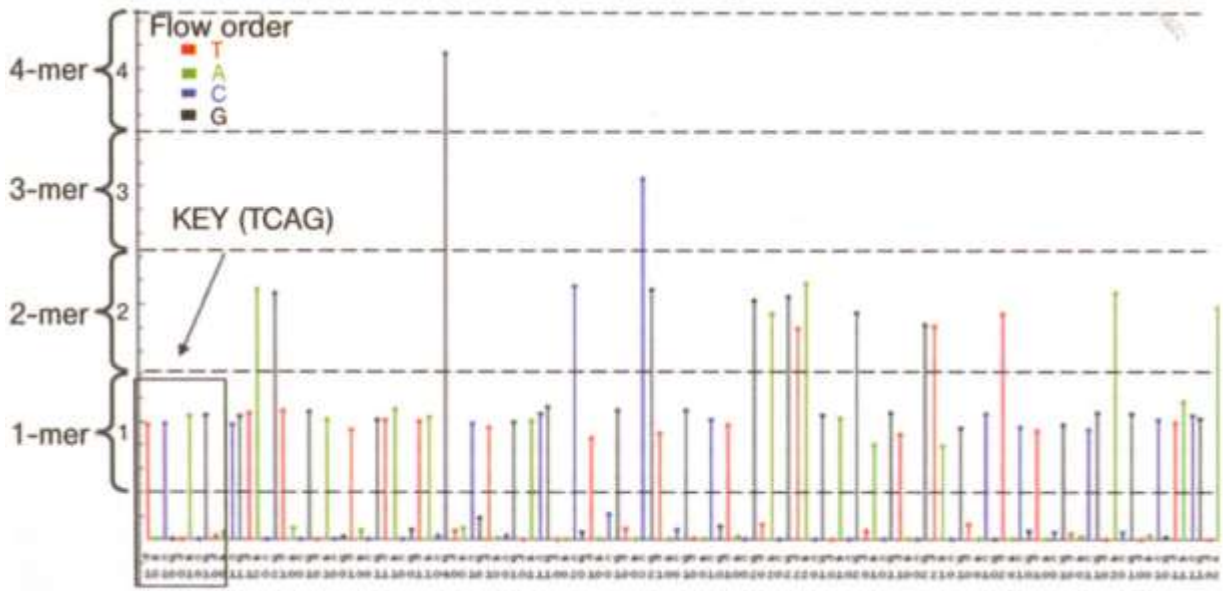


Figure 4.4 Signal flowgram and base calling.

Ograniczenia:

- Dokładność sekwencjonowanie traktów homopolimerowych (AAAAA, TTTTTT itp.) jest ograniczona: zależność między ilością światła a liczbą włączonych nukleotydów jest liniowa wyłącznie do 8 nt
- najczęstsze błędy: insercje i delecje
- dokładność: 99,5 % przy długości odczytu do 200 pz (Sanger: 99,999 % przy 1000 pz)

pirosekwencjonowanie (454)

Zalety:

- **KOSZT**: w momencie wprowadzenia koszt sekwencjonowania genomu człowieka 1000 razy niższy od HGP i 100 razy niższy od Celery
- Długość odczytów: na początku od 100-400 nt; w 2013: do 1000 nt
- Ilość odczytów: na początku około 400 000 / płytkę; teraz: 1 000 000 / płytkę

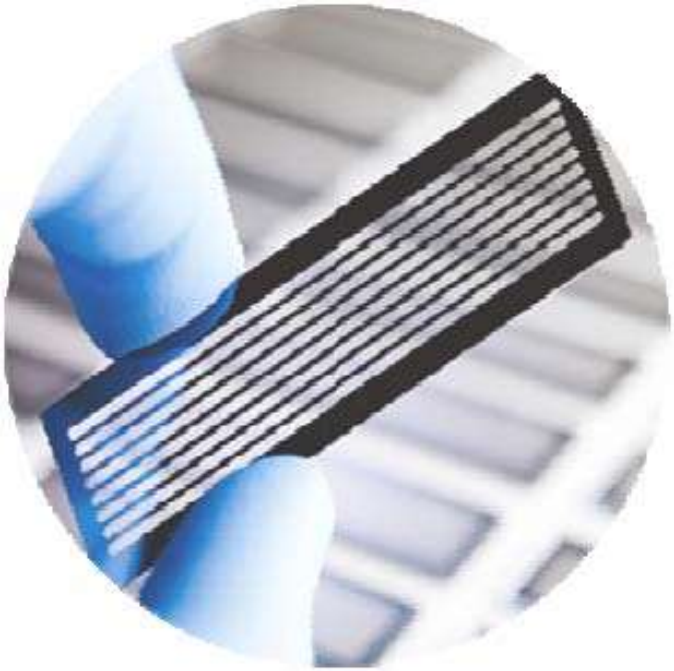
Czerwiec 2007: zsekwencjonowanie genomu Jamesa Watsona za pomocą 454:

- Pierwszy genom ludzki zsekwencjonowany „nie-Sangerem”
- 2 miesiące
- koszt < 1 miliona dolarów

pirosekwencjonowanie (454) podsumowanie

- sekwencjonowanie przez syntezę
- amplifikacja klonalna prowadzona w emulsji
- DNA immobilizowane na kulkach
- w każdym cyklu sekwencjonowania dodawany jeden typ nukleotydu, nieblokowany
- trakty homonukleotydowe odczytywane przez wzrost intensywności sygnału
- błędy sekwencjonowania: insercje i delecje, dokładność 99,5 %
- długie odczyty (do 1 000 nt)
- dużo odczytów w 1 reakcji w porównaniu z Sangerem, jednak zdecydowanie mniej niż w innych technologiach NGS – do 1 000 000
- koszt zdecydowanie niższy niż Sanger, jednak drożej niż inne technologie NGS

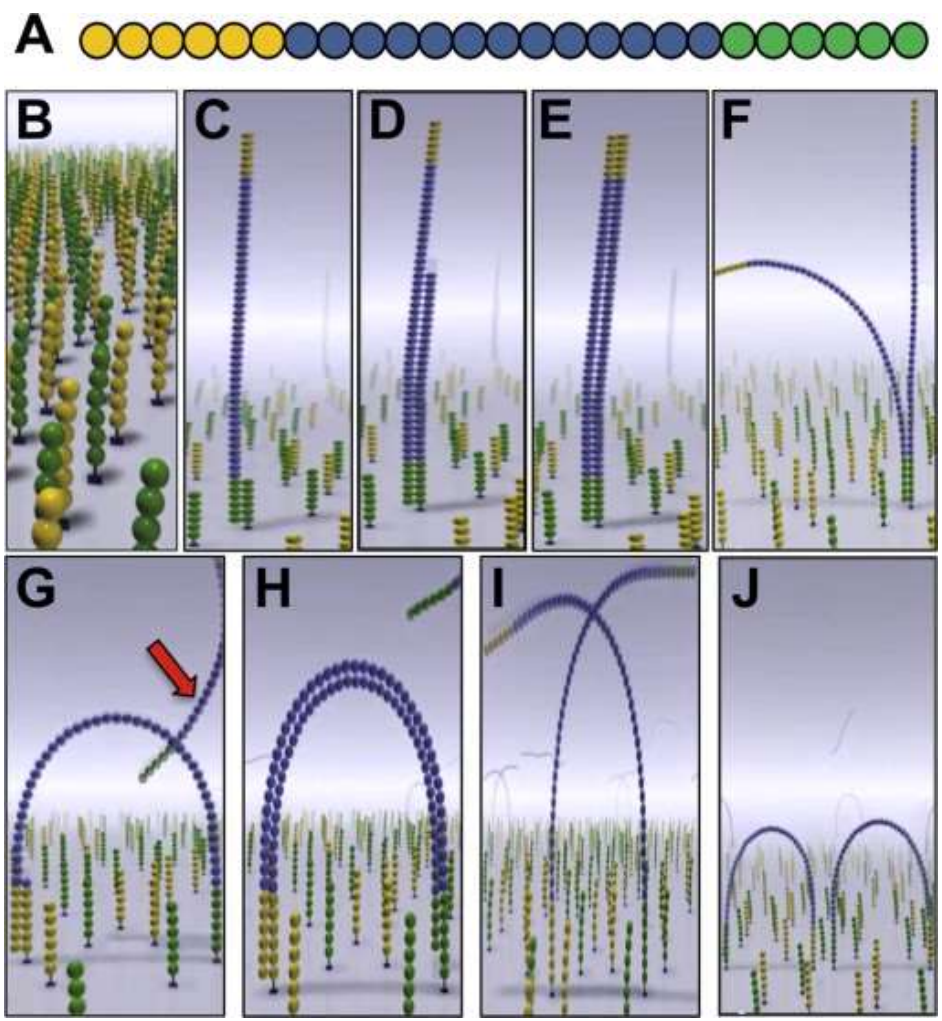
Illumina



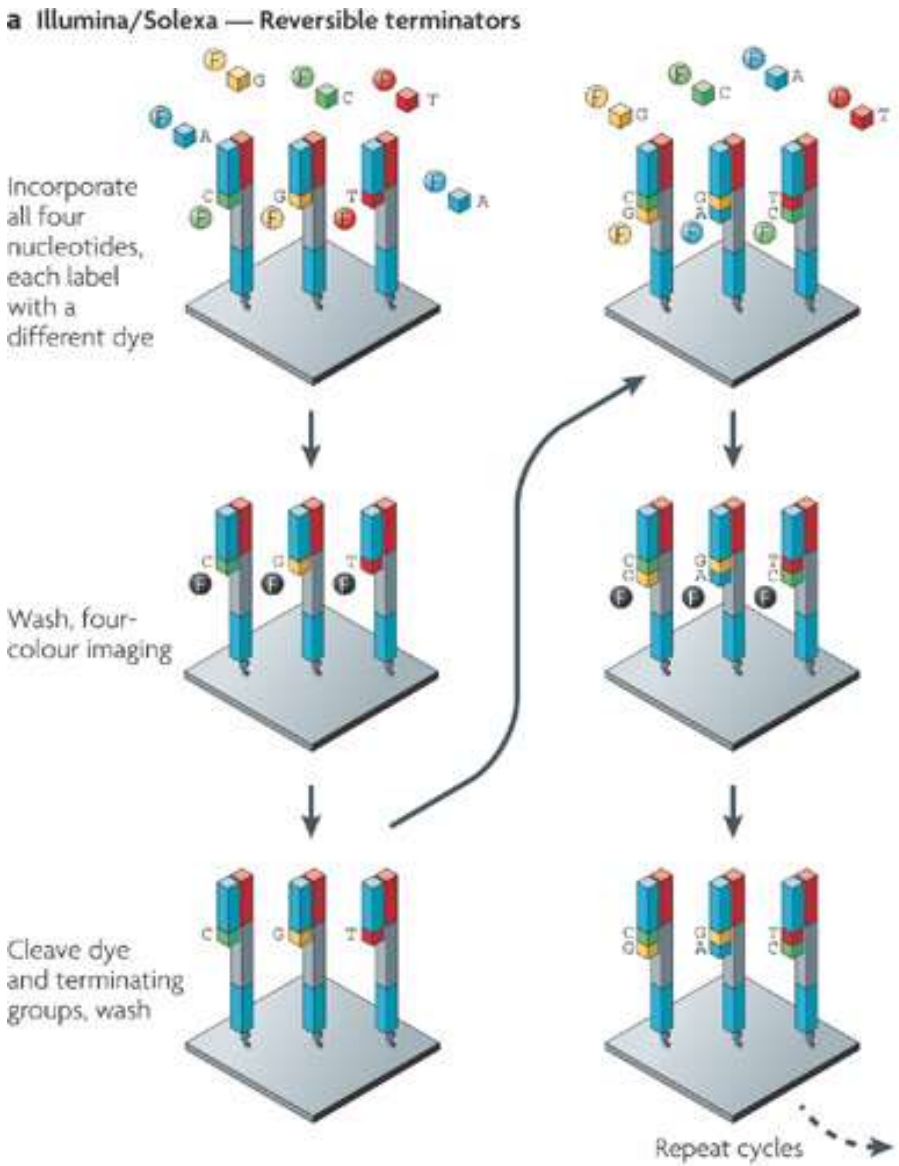
Illumina

amplifikacja klonalna przez „mostkowy” PCR:

- zdenaturowany DNA nanoszony jest na płytkę
- DNA hybrydyzuje do starterów związanych kowalentnie z płytką
- wydłużanie łańcucha przez polimerazę – powstanie nici kowalentnie związanej z podłożem
- denaturacja i hybrydyzacja „mostkowa” do sąsiedniego startera na płytce
- kopiowanie „mostka” -> denaturacja -> ponowne formowanie „mostka”
- 35 powtórzeń tworzy gęsty klaster około 2000 cząsteczek DNA
- nici komplementarne są usuwane
- końce 3’ są blokowane aby zapobiec niespecyficznej inicjacji

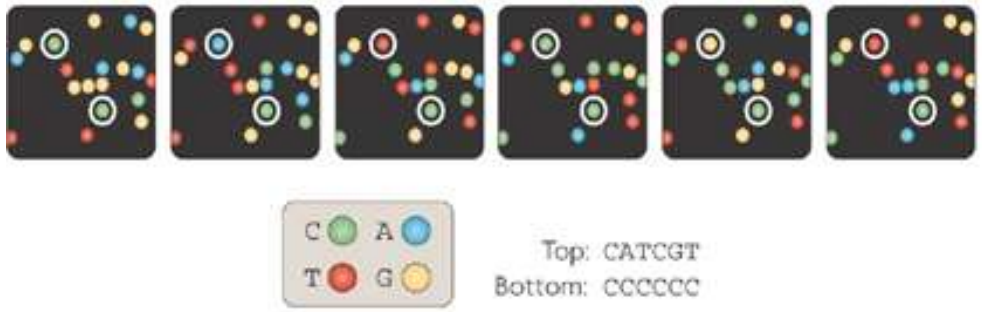


Illumina



sekwencjonowanie:

- przyłączenie startera
- dostarczenie wyznakowanych fluorescencyjnie i zablokowanych dNTP oraz ich przyłączenie przez polimerazę do komplementarnych klastrów
- zdjęcie
- usunięcie fluoroforów i reszt blokujących
- powtórzenie cyklu



Illumina

Ograniczenia:

- krótkie odczyty: na początku 35-50 nt; 2016: do 2 x 300 nt
- dokładność: 98,5 – 99,5 %
- dominujące błędy: substytucje

Illumina

Zalety:

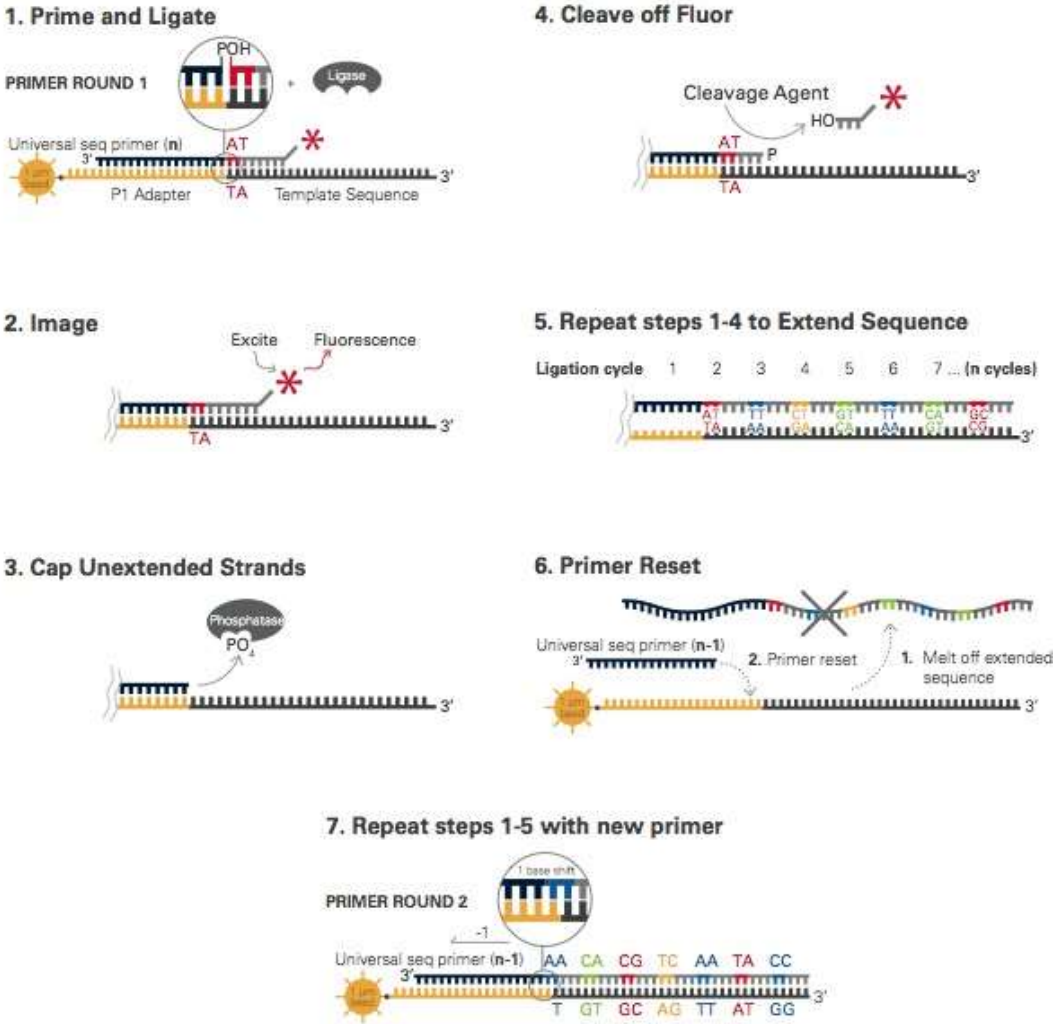
- dokładne odczyty sekwencji homopolimerowych
- możliwość wykonywania odczytów typu „paired-end”
- ilość odczytów: 3 miliardy (3 x 10^9) dla pojedynczych końców lub 6 miliardów dla dwóch końców (do 600 Gb)
- bardzo niski koszt

Illumina
podsumowanie

- sekwencjonowanie przez syntezę
- amplifikacja klonalna prowadzona na płytce poprzez „mostkowy” PCR
- DNA immobilizowane na płytce
- w każdym cyklu sekwencjonowania dodawany są wszystkie nukleotydy, blokowane
- trakty homonukleotydowe odczytywane są podczas kolejnych cyklów
- błędy sekwencjonowania : substytucje, dokładność 98,5 %
- krótkie odczyty (do 2 x 300 nt)
- najwięcej odczytów w 1 reakcji – do 3/6 miliardów (6 000 000 000)
- bardzo niski koszt przy dużych próbkach

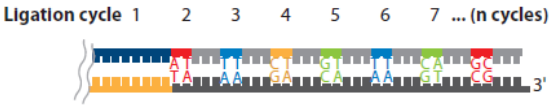
SOLID

Sequencing by Oligonucleotide Ligation and Detection



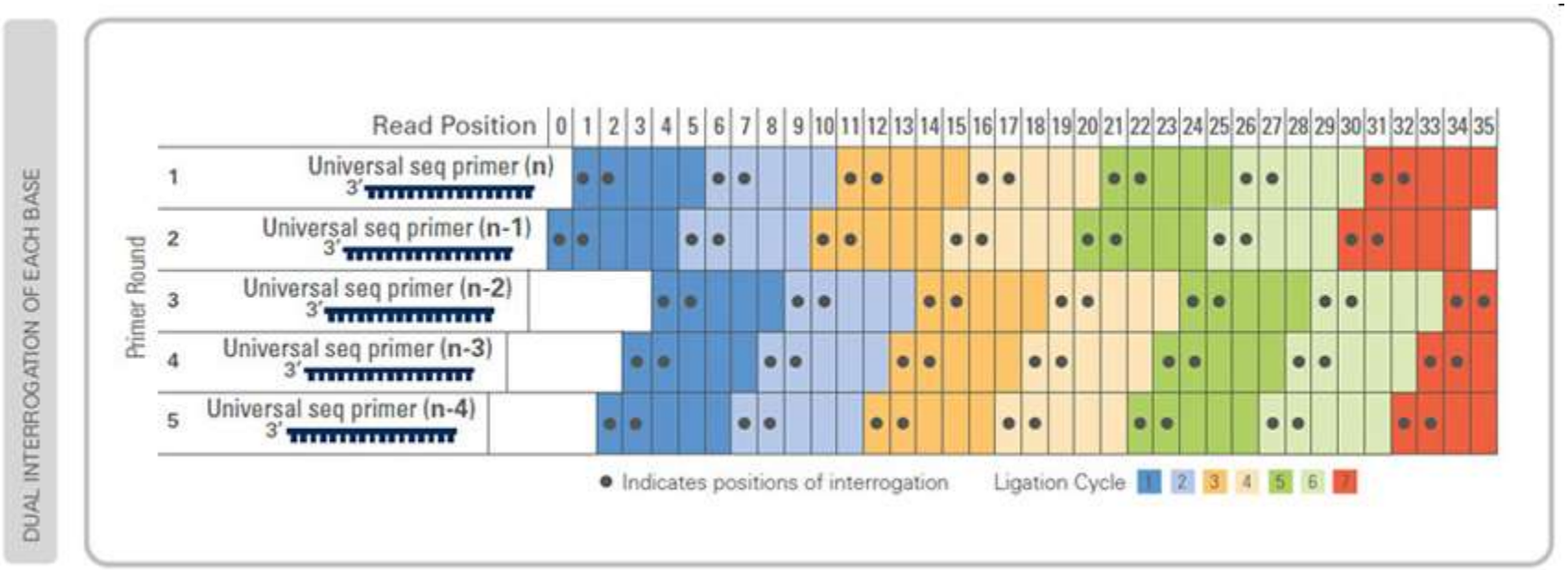
- starter sekwencyjny hybrydyzuje do adaptera a mieszanina 4 znakowanych fluorescencyjnie **2nt-prób** konkuruje o przyłączenie do startera

- ligacja, detekcja, cięcie – powtórzone w cyklach



SOLID

Sequencing by Oligonucleotide Ligation and Detection



- powtarzamy SOLID ze starterami kolejno: **n-2, n-3, n-4**

SOLID

Sequencing by Oligonucleotide Ligation and Detection

- odczyty: nawet do 100 bp, łącznie do **60 Gb/przebieg**
- koszt: około 4,000 USD
- kodowanie dwunukleotydowe – 99,94 % dokładności odczytu – przydatne w analizie SNP
- niska „mapowalność” odczytów

Semiconductor Sequencing for Life™

life technologies™ | ion torrent

Ion Torrent

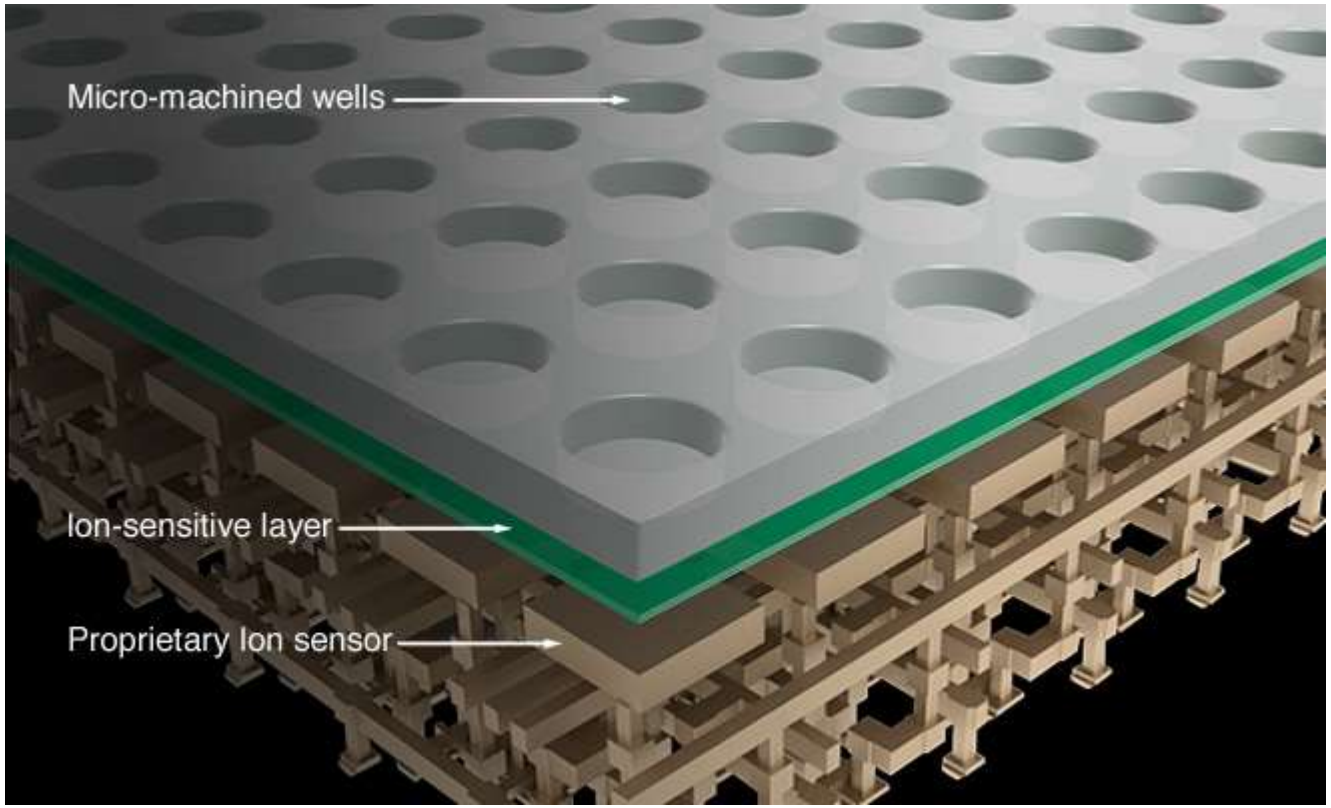
- zamiast fluorescencji odczytuje zmiany pH
- każdy dołek fazy stacjonarnej to mini-pHmetr, który odnotowywuje lokalne zmiany pH powstające podczas emisji protonu z przyłączanego nukleotydu!
- mikrodołki = DNA pływa w jednym rodzaju nt
  włączenie nt = uwolnienie jonu wodorowego = uaktywnienie sensora jonowego = zaszła reakcja
- nie ma problemu z homopolimerami

- koszt: 350 USD/ przebieg
- ilość: 100 Mbp/przebieg
- długość: do 200 bp
- 10 MB - >1GB / przebieg

http://www.youtube.com/embed/yVf2295JqUg

Semiconductor Sequencing for Life™

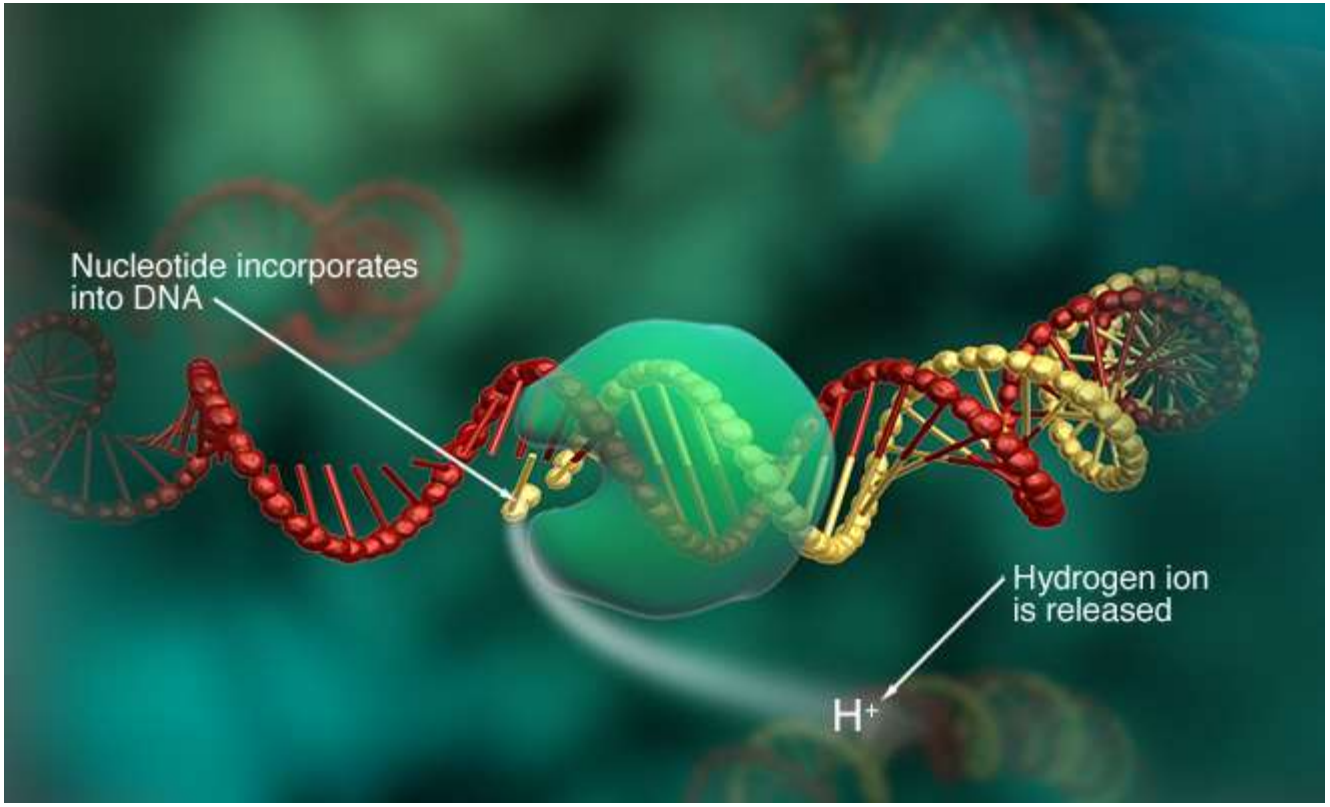
Ion Torrent



Semiconductor Sequencing for Life™

life technologies™ | ion torrent

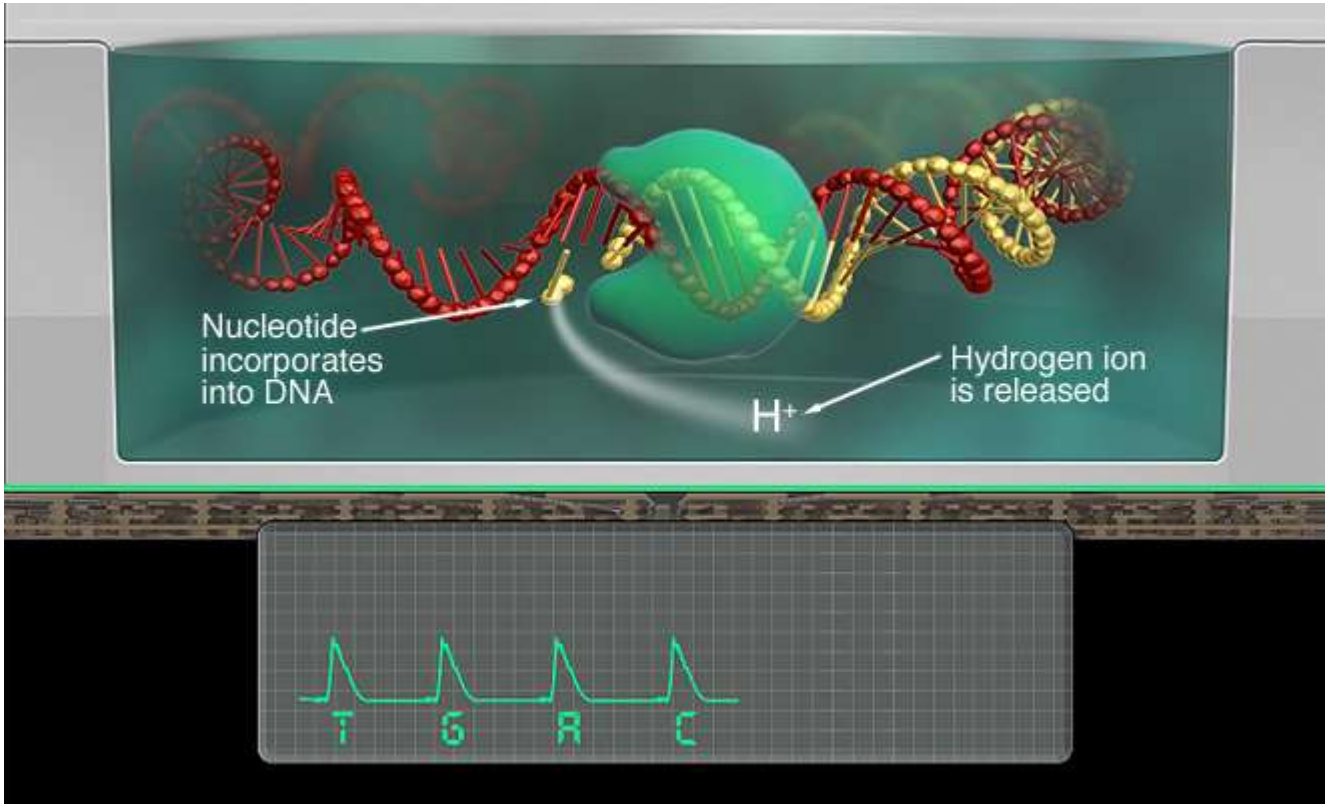
Ion Torrent



Semiconductor Sequencing for Life™

life technologies™ | ion torrent

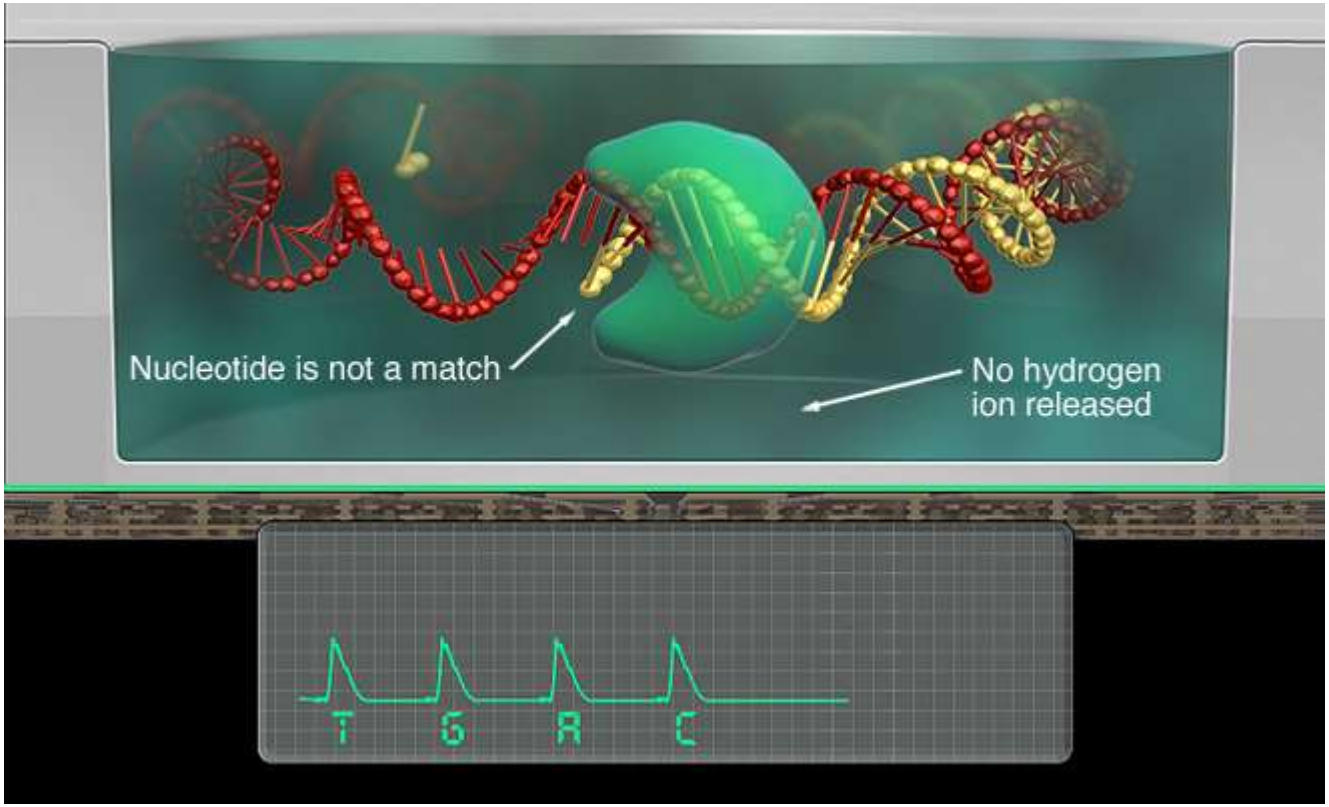
Ion Torrent



Semiconductor Sequencing for Life™

life technologies™ | ion torrent

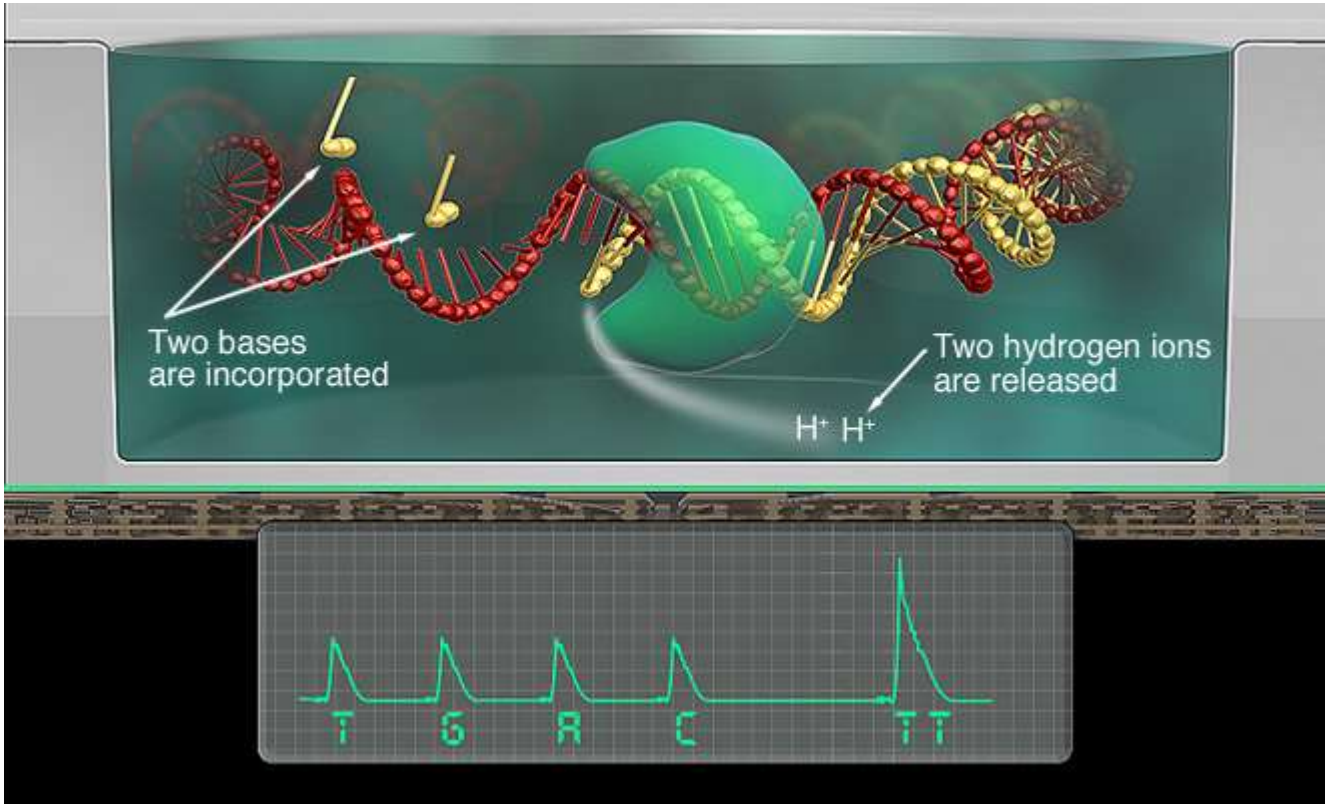
Ion Torrent



Semiconductor Sequencing for Life™

life technologies™ | ion torrent

Ion Torrent



Semiconductor Sequencing for Life™

life technologies™ | ion torrent

Ion Torrent

Zalety:

- Średnia długość odczytów: do 200-400 nt
- Stosunkowo tanio dla małych projektów
- Niewielkie i tanie urządzenia

Wady:

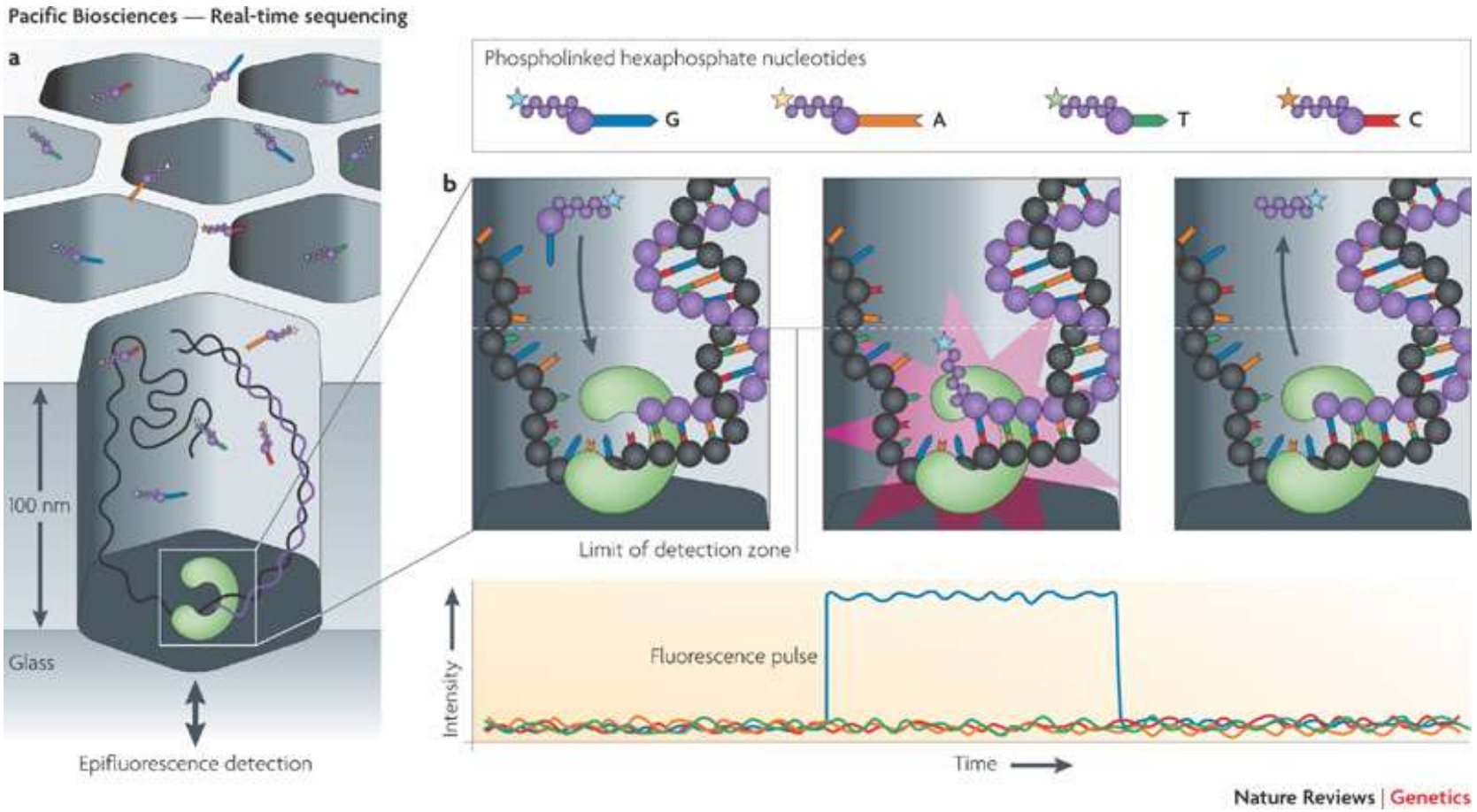
- Niska jakość, jednak można ją „podrasować”

III generacja

sekwencjonowanie w czasie rzeczywistym

z pojedynczej cząsteczki DNA

PacBio SMRT sequencing



PacBio SMRT sequencing

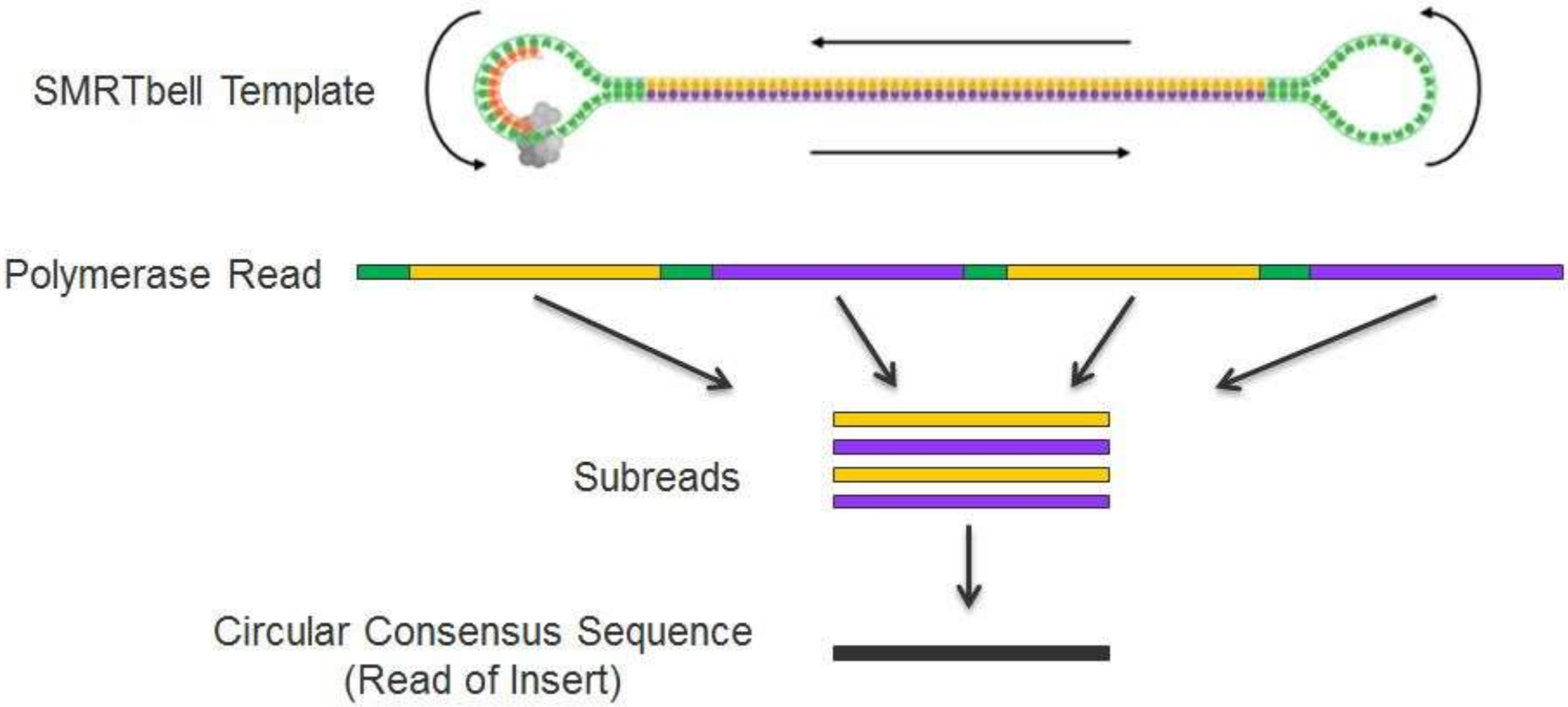
Zalety:

- Bardzo długie odczyty: do 25 kb !!!
- Stosunkowo tanio
- Brak amplifikacji
- Odczyt w czasie rzeczywistym – możliwość dowolnego prowadzenia reakcji

Wady:

- Niska jakość, około 85% dokładności, jednak można ją „podrasować”

PacBio SMRT sequencing



2015 – nanopores

TECHNOLOGY

For more technology stories,

Reader of genomes

Sequence DNA in seconds

DNA sequencing can now be done on a device that plugs into your computer like a memory stick

Duncan Graham-Rowe

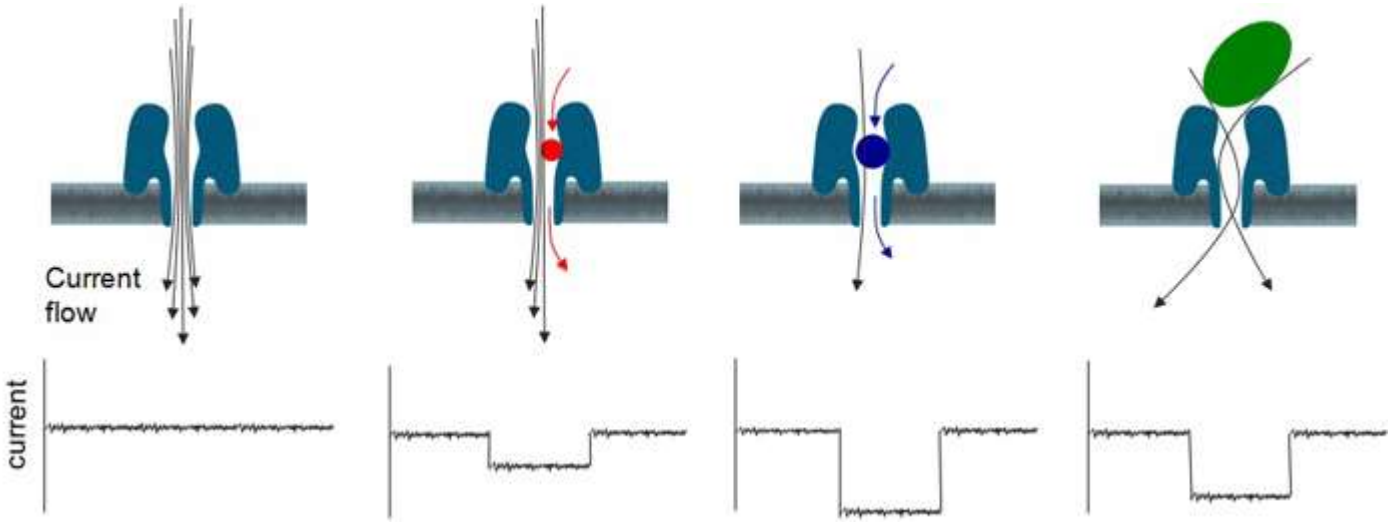
IT LOOKS like an ordinary USB memory stick, but a little gadget that can sequence DNA while plugged into your laptop could

Genome Biology and Technology (AGBT) conference in Marco Island, Florida.

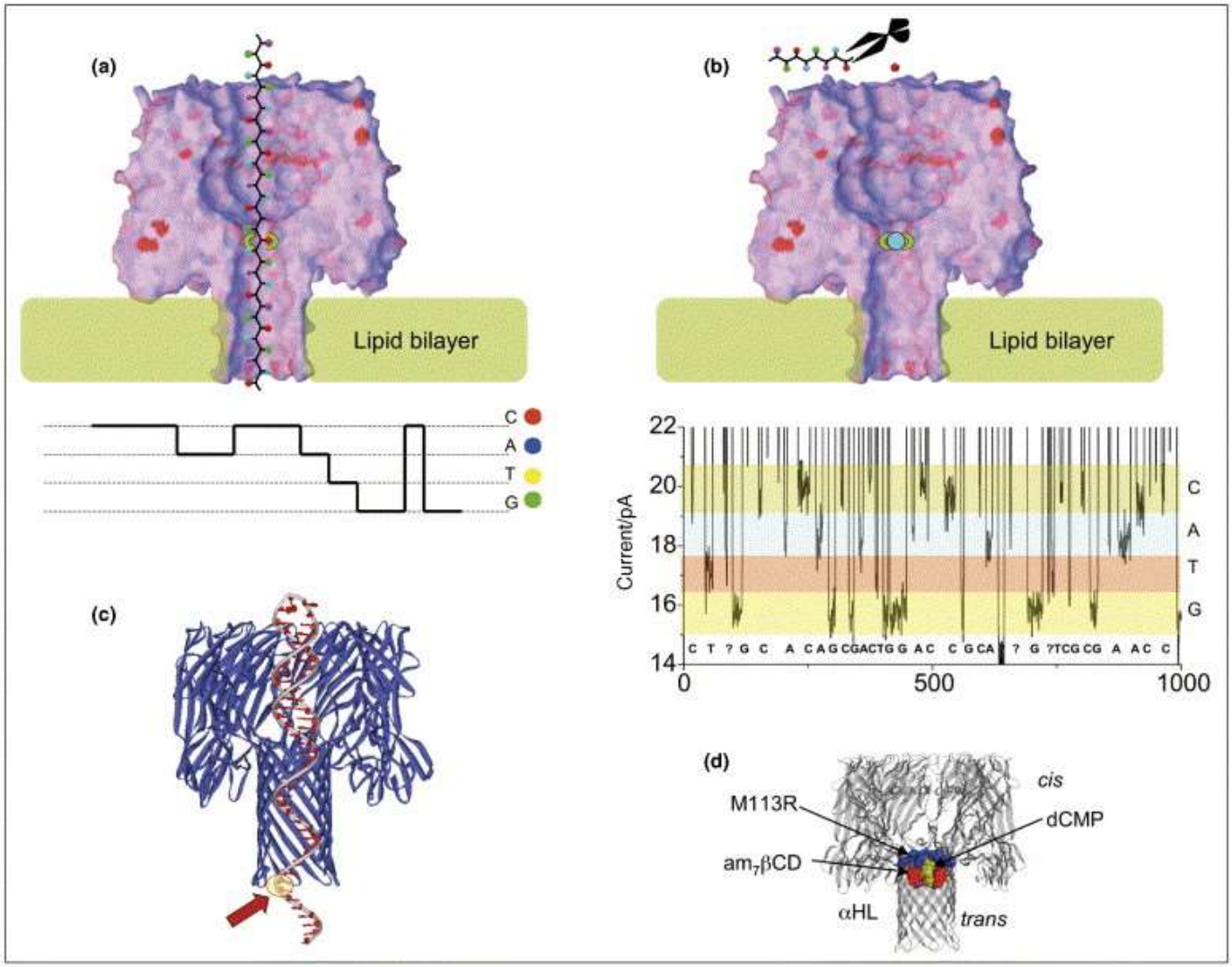
This was done as a proof of principle. "Phi X was the first genome ever to be sequenced," says the University of Birmingham, UK, and author of the blog Pathogens: Genes and Genomes. It shows the technology works, he says. "If you can sequence this genome you should be able to sequence larger genomes."

Oxford Nanopore is also building a larger device, called GridION, for lab use. Both GridION

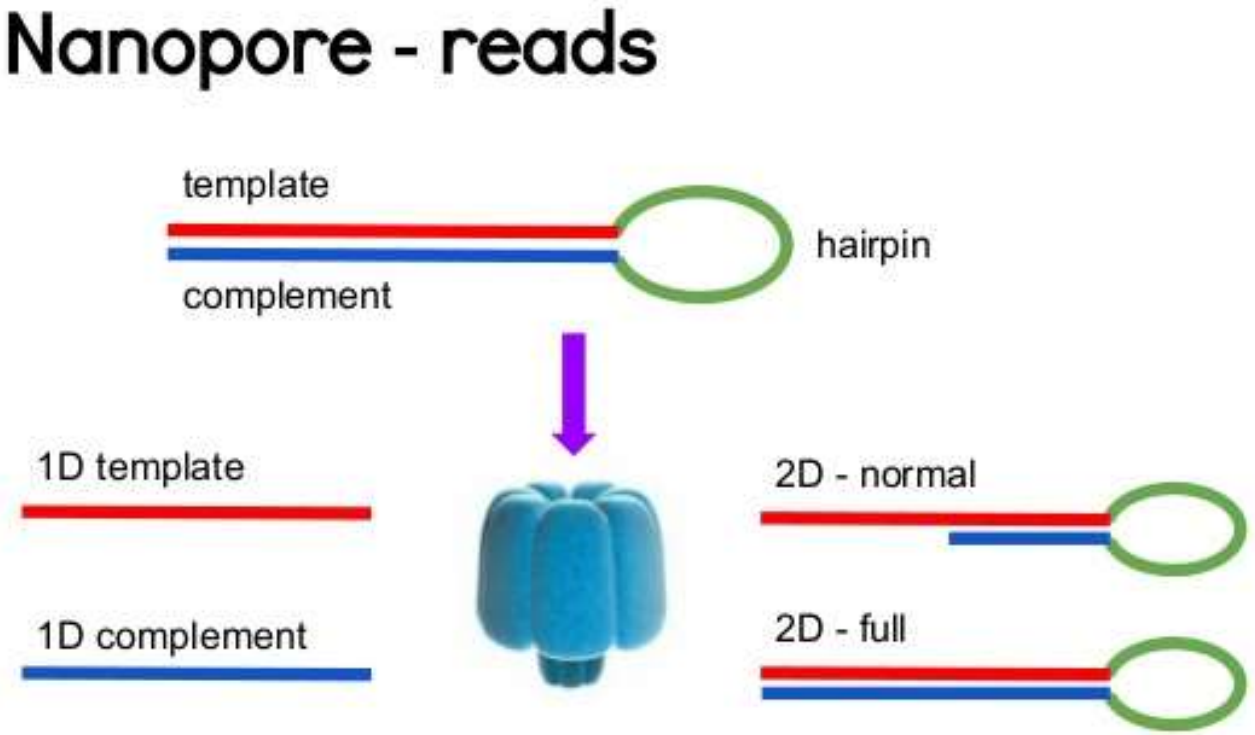
2015 – nanopore



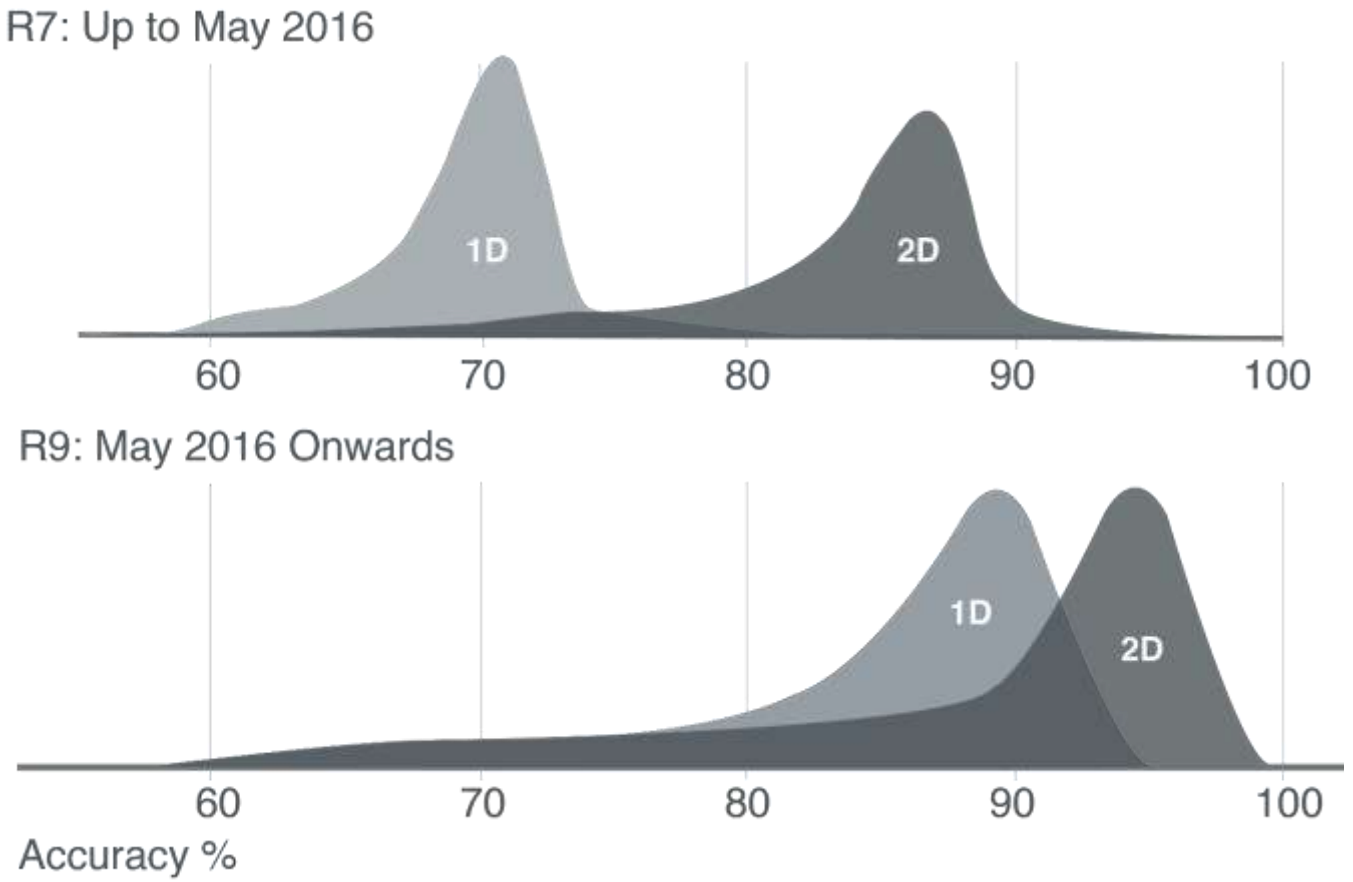
2015 – nanopore



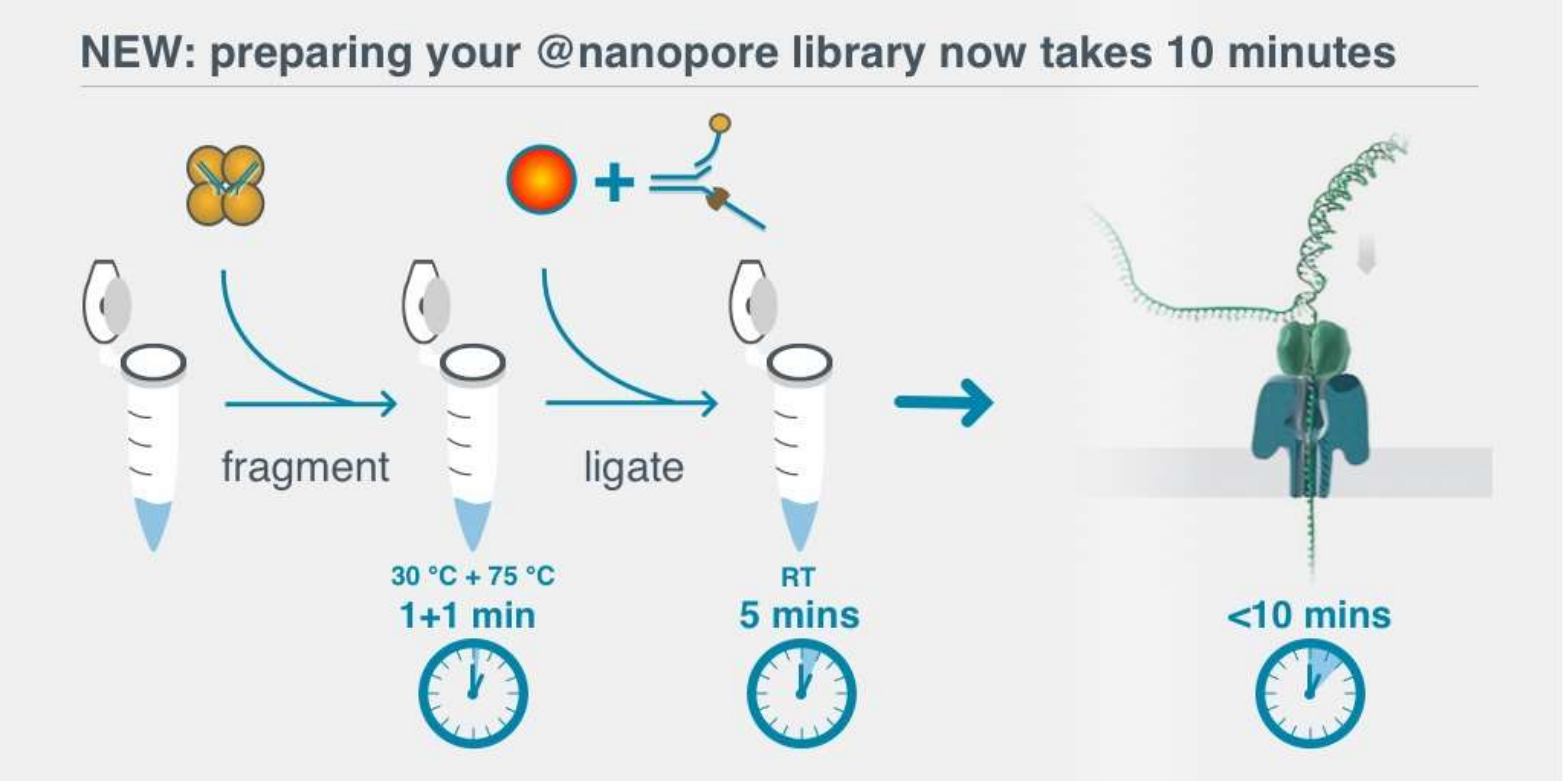
2015 – nanopore



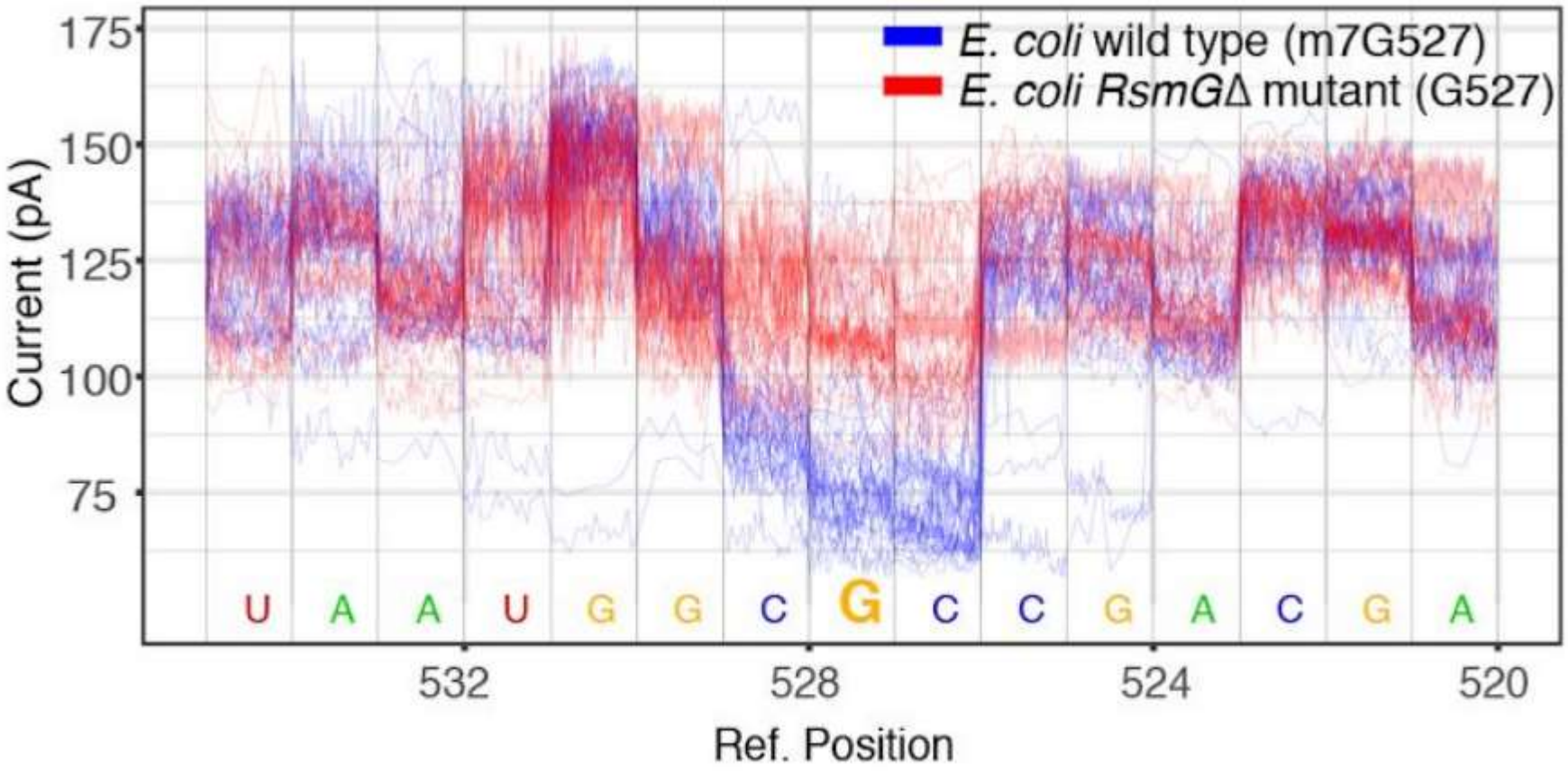
2016 – MinION


R7: Up to May 2016
1D
2D
60
70
80
90
100
R9: May 2016 Onwards
1D
2D
60
70
80
90
100
Accuracy %

2016 – MinION

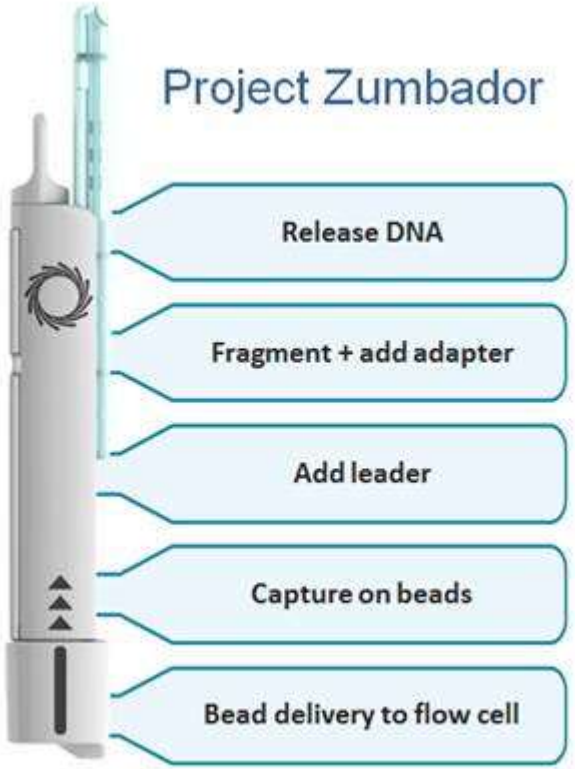

NEW: preparing your @nanopore library now takes 10 minutes
fragment
ligate
30 °C + 75 °C
1+1 min
RT
5 mins
<10 mins

2017 – Direct RNA sequencing



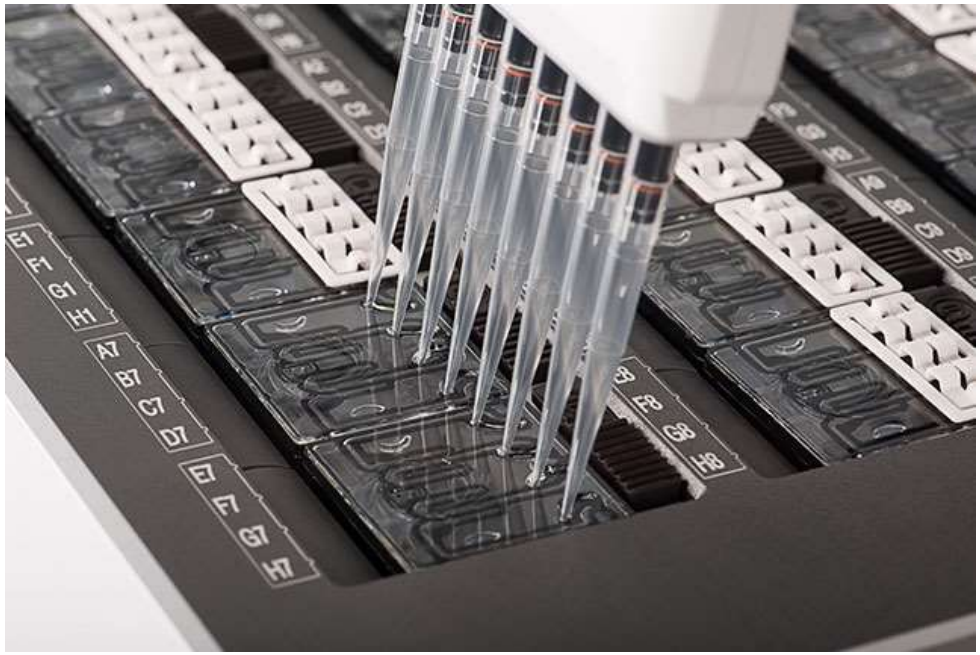
2017 – Voltrax

2018 – Zumbador



2017 – GridION X5

2017 – PromethION



2018 – SmidgION

2018 – Flongle

analiza danych

analiza:

- profilowanie ekspresji genów
- identyfikacja wariantów splicingowych
- identyfikacja zmienności genetycznej
- epigenetyka
- ….

jakość

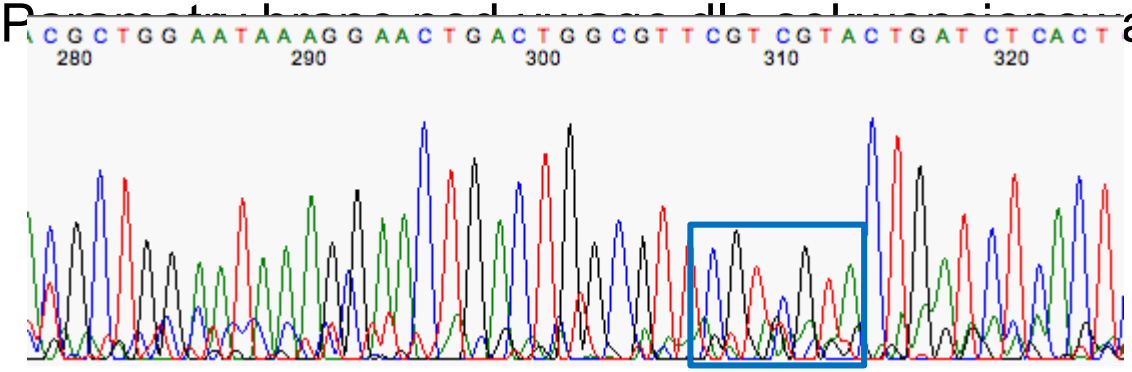
co wpływa na jakość zestawu odczytów?

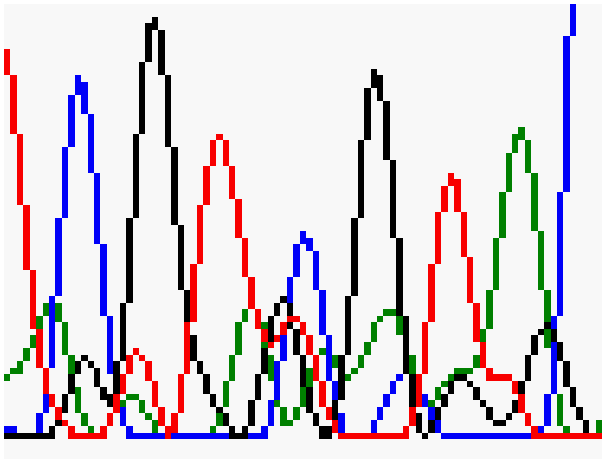
- błędny odczyt danej pozycji
- ominięcie/dodanie pozycji
- preferencje odnośnie sekwencji bogatych/ubogich w GC
- obecność i odczyt traktów homopolimerowych
- niska jakość zasad na 3’ końcu
- duplikacje klonalne
- zanieczyszczenia obcym DNA/RNA
- zanieczyszczenia organellowym DNA/RNA
- artefakty (fragmenty adaptorów, wektorów, brak insertów itp.)

błędy odczytu

To, co nas interesuje to **prawdopodobieństwo** błędnego odczytania danej pozycji

Parametry brane pod uwagę dla sekwencjonowania metodą Sangera:





- odstępy pomiędzy pikami
- dystans do najbliższego „N”
- stosunek wysokości pików dla nukleotydów zgłoszonych i niezgłoszonych w ramach okna 7 nt wokół analizowanej pozycji

Prawdopodobieństwo błędu oznaczane jest poprzez porównanie wartości powyższych parametrów do empirycznie zbadanych częstości występowania błędów dla nukleotydów o danych parametrach

błędy odczytu

Przykładowe parametry dla Illuminy:

- profile intensywności sygnału z poszczególnych punktów na płytce
- profile intensywności tła
- inne, w zależności od wersji

Wraz z każdą zmianą w chemii, płytkach lub instrumentach używanych do sekwencjonowania oznaczany jest nowy model parametrów przewidywania prawdopodobieństwa błędu odczytu

punktacja Phred

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

Q - jakość danego nukleotydu

P - prawdopodobieństwo błędnego odczytu danego nukleotydu

$$P = 10^{\frac{Q}{-10}}$$

punktacja Phred

Quality Value	Error Probability	Probability Called Base is Correct	Description
10	0.1	0.9	error rate of 1 in 10
20	0.01	0.99	error rate of 1 in 100
30	0.001	0.999	error rate of 1 in 1000
40	0.0001	0.9999	error rate of 1 in 10000

dla pozycji o Q=27 jakie jest prawdopodobieństwo błędu?

P = 10^(Q/-10) = 10^(27/-10) = 0,001995262

kodowanie punktacji Phred

2 pliki FASTA z kodowaniem liczbowym:

```
>seq1
ACGTACGTACGATAGACGACGA
>seq2
CCAGGAGATACGACGTAGCATCAA
```

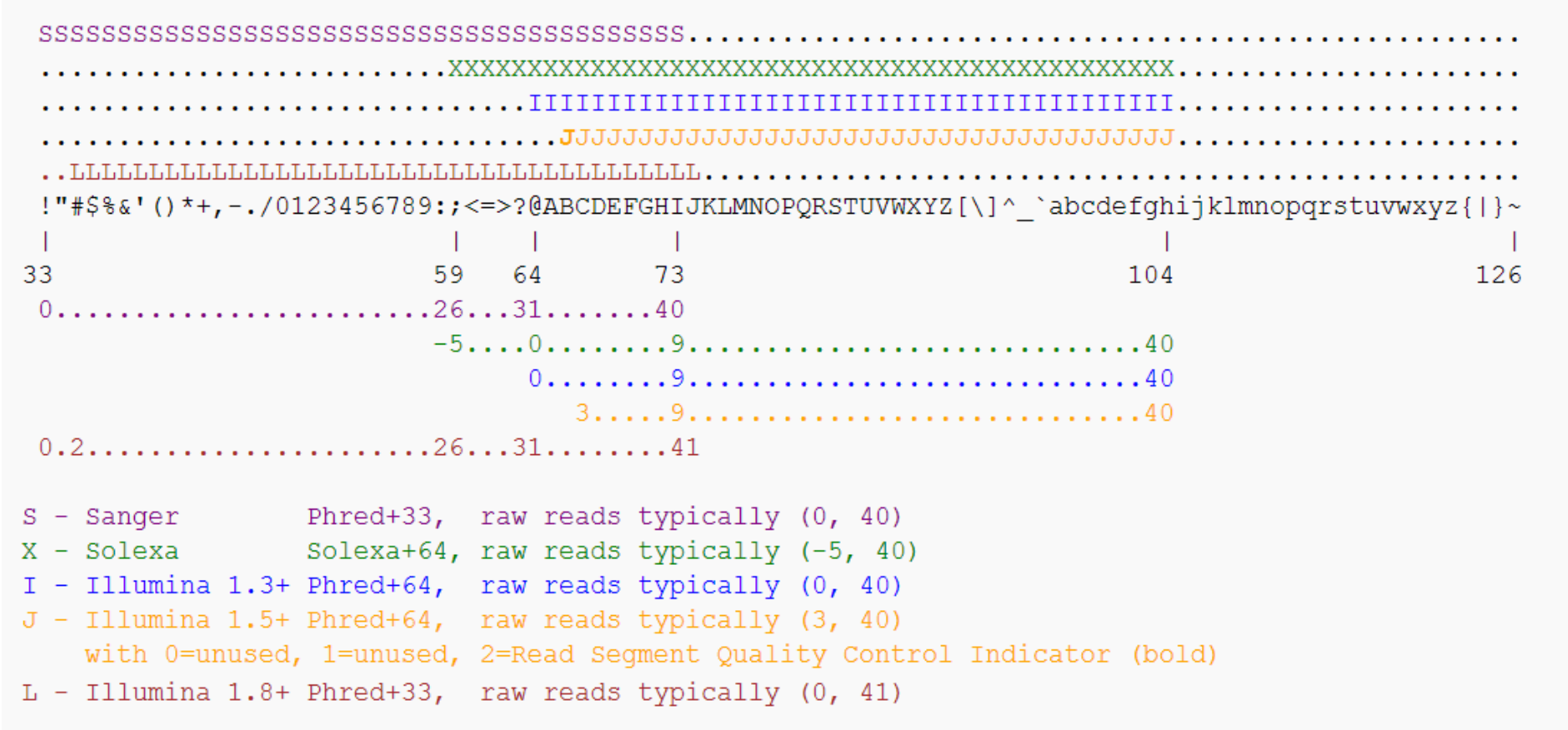
```
>seq1
30 30 40 37 36 38 39 36 37 37
35 30 30 40 37 36 38 39 36 37
37 35
>seq2
30 30 30 40 37 36 38 39 36 37
30 36 38 39 36 37 37 35 37 35
40 37
```

kodowanie punktacji Phred

1 plik FASTQ z kodowaniem ANSI:

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+SEQ_ID
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

kodowanie ANSI punktacji Phred



co jeszcze w formacie FASTQ?

Również identyfikator sekwencji ma znaczenie

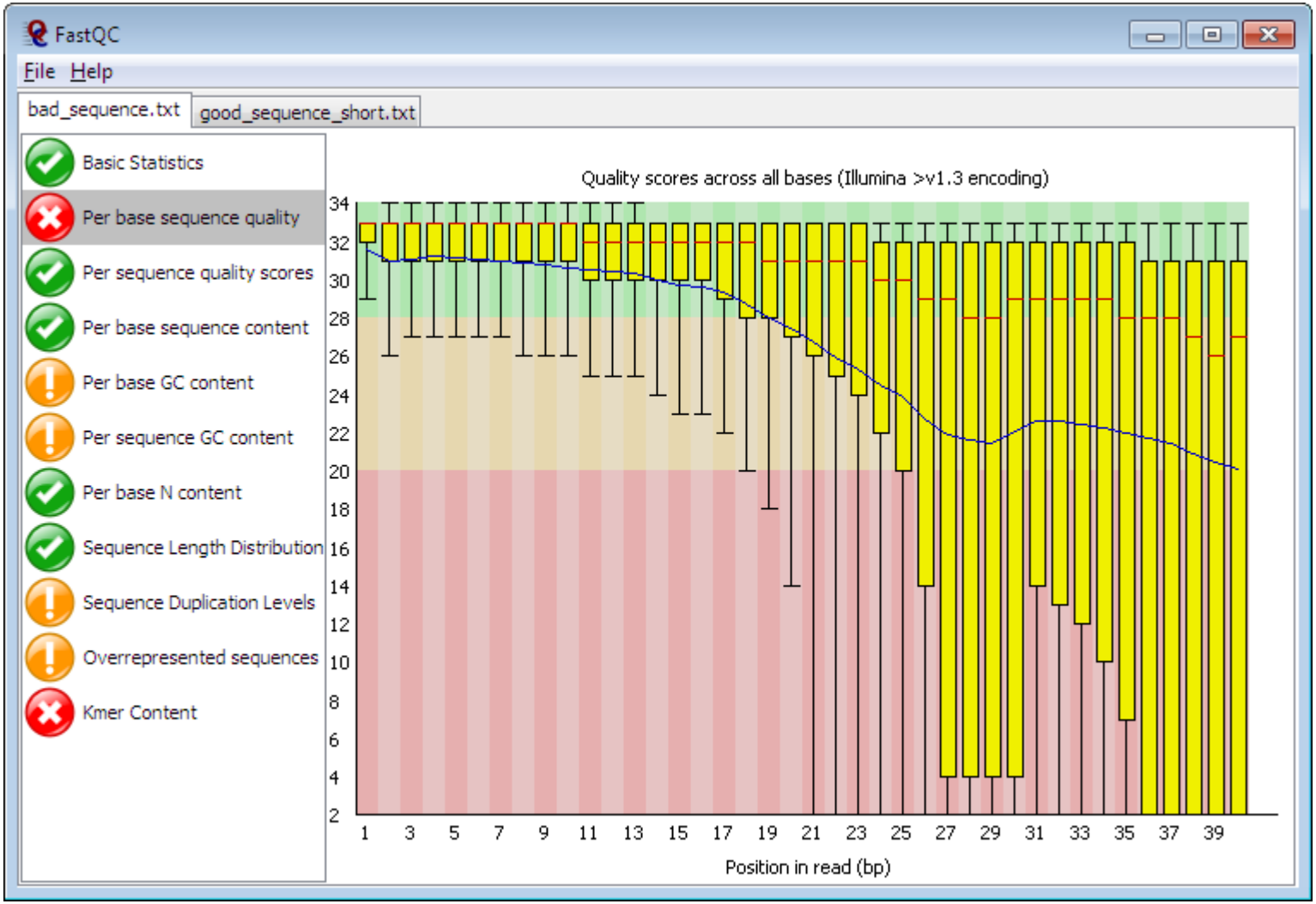
Przykład dla Illuminy:

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
```

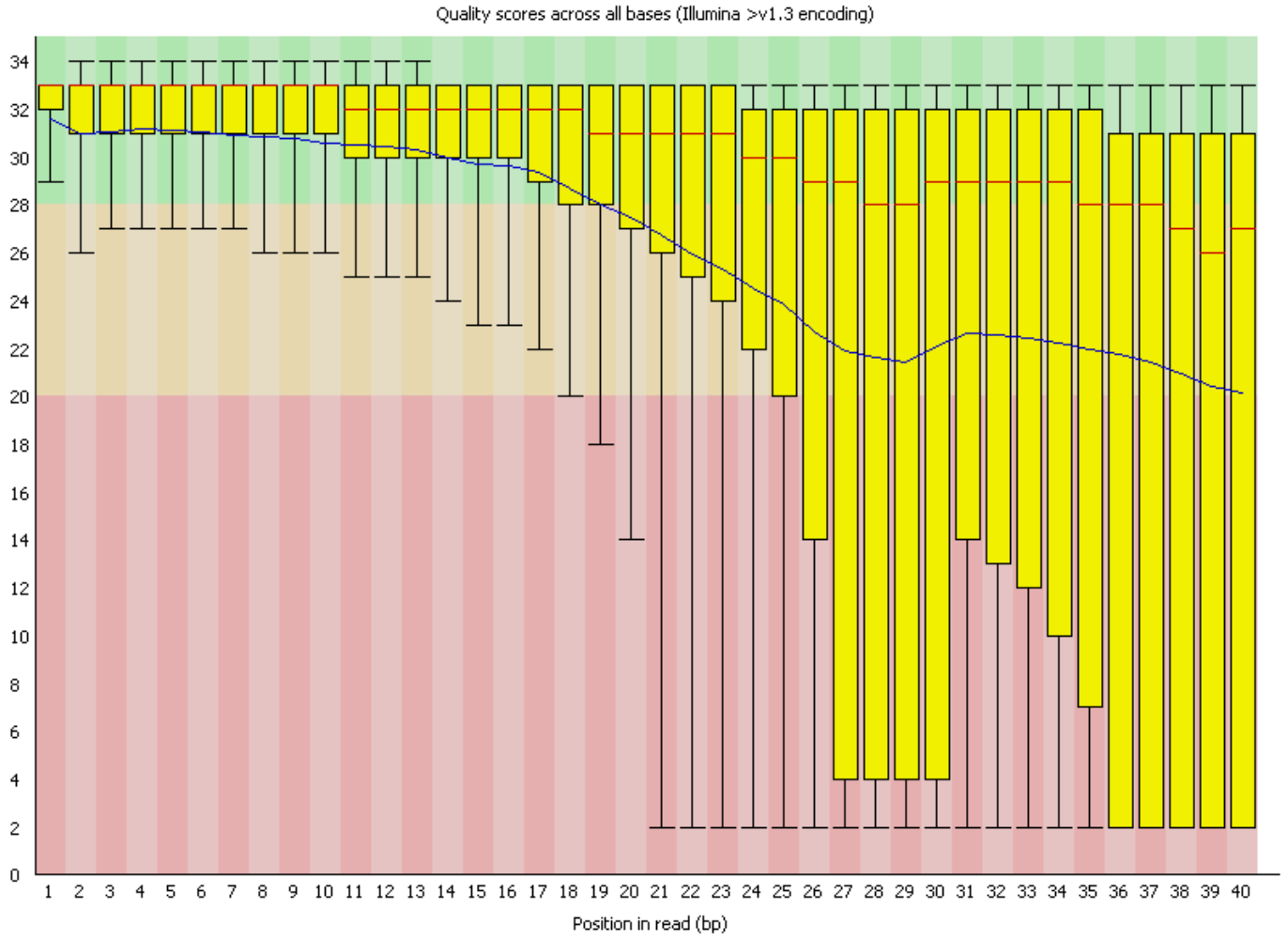
EAS139	unikalna nazwa instrumentu
136	identyfikator eksperymentu
FC706VJ	identyfikator płytki
2	numer linii płytki
2104	numer segmentu w obrębie linii
15343	koordynat 'x' odczytywanego punktu
197393	koordynat 'y' odczytywanego punktu
1	kolejność w parze (1/2) – dla odczytów z dwóch końców (paired-end)
Y	Y jeśli odczyt jest kiepskiej jakości i nie spełnia minimalnych kryteriów, N jeśli jest dobry
18	0 jeśli brak znaczników odnośnie kontroli jakości, bądź liczba parzysta jeśli są
ATCACG	sekwencja indeksu (przy multipleksowaniu próbek)

analiza jakości zestawu odczytów

FastQC

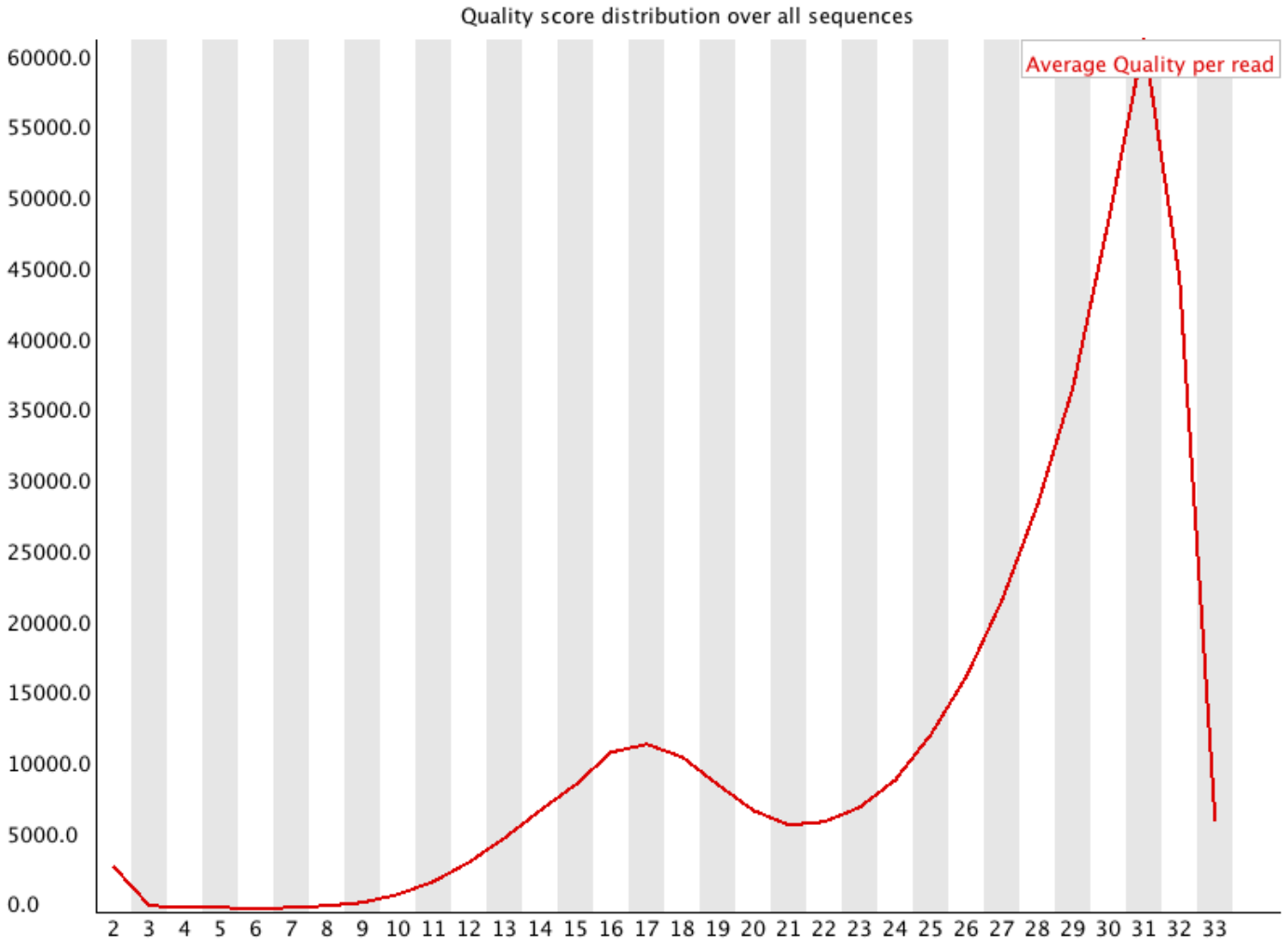


per base sequence quality



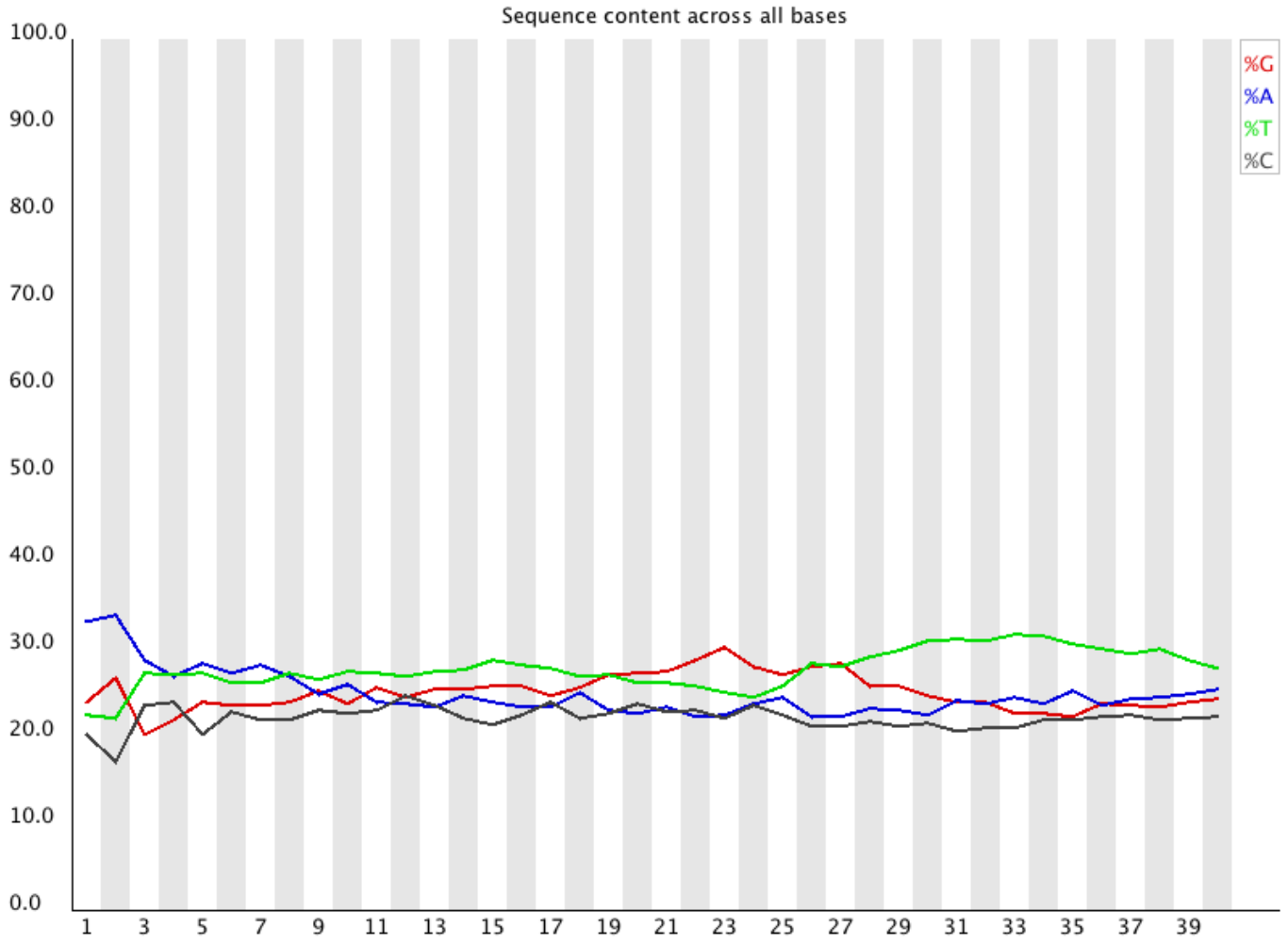
rozkład jakości zasad w poszczególnych pozycjach odczytów

per sequence quality scores



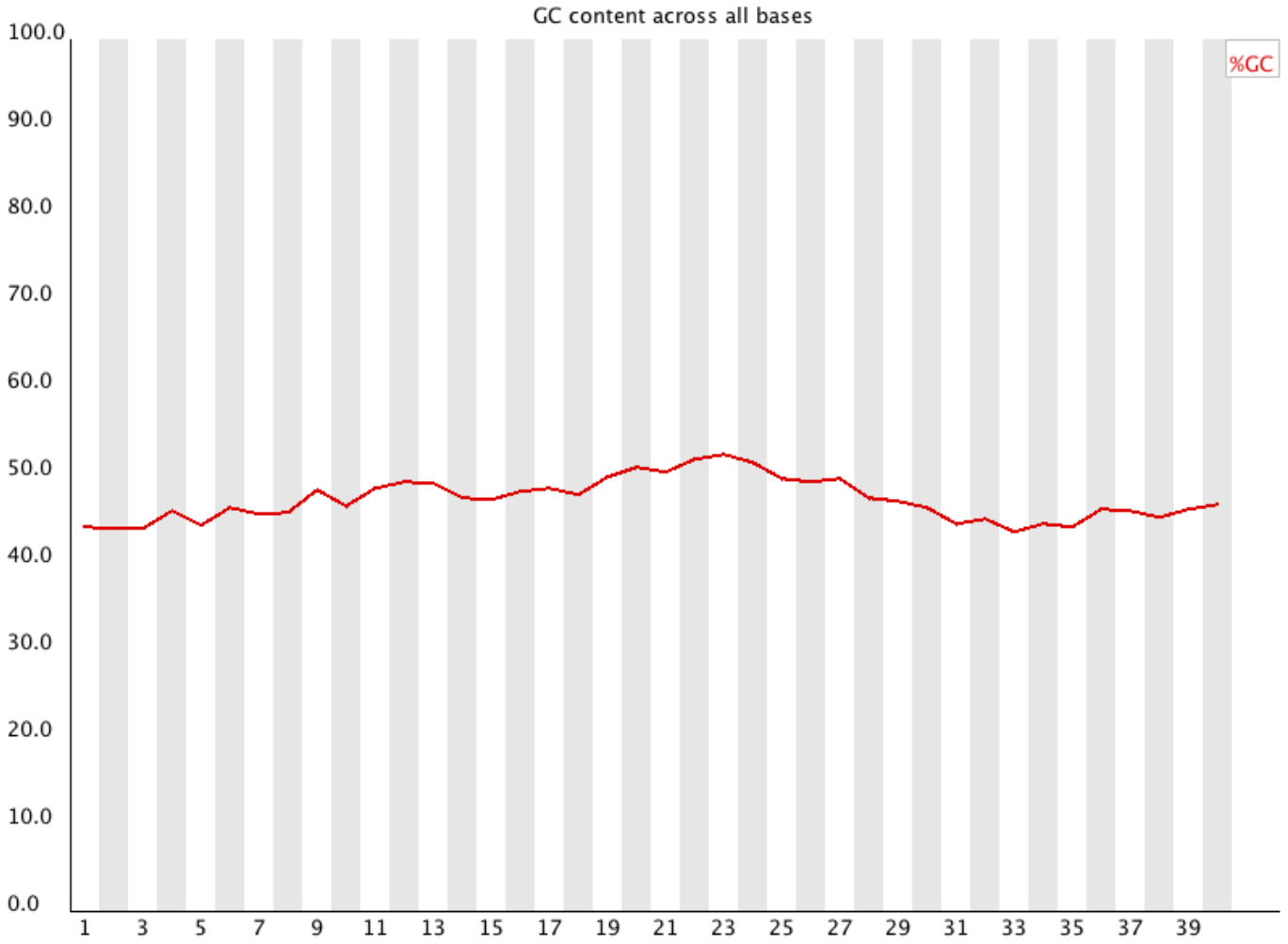
rozkład średnich jakości odczytów

per base sequence content



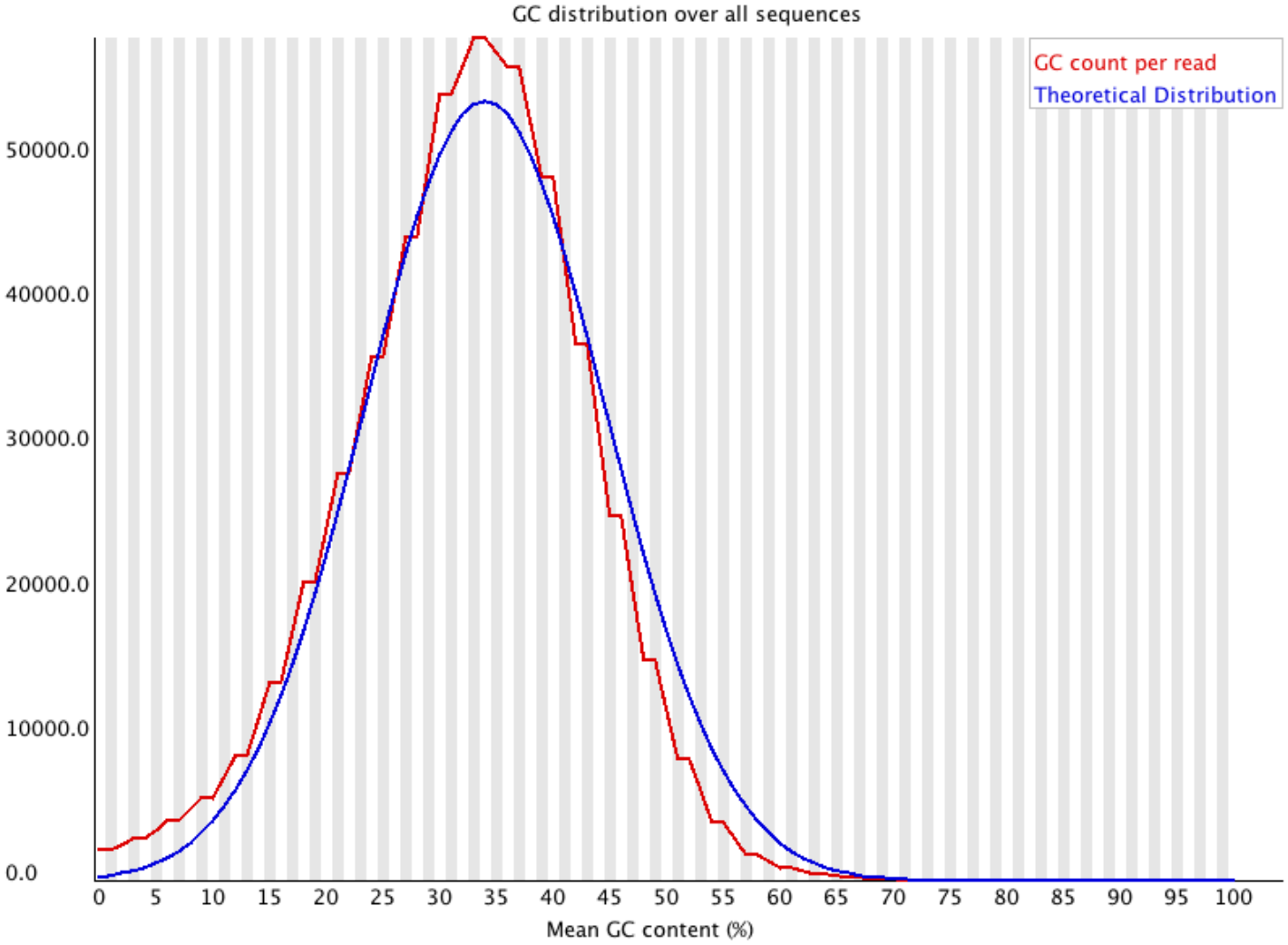
sprawdzenie preferencji sekwencyjnych wzdłuż odczytów

per base GC content



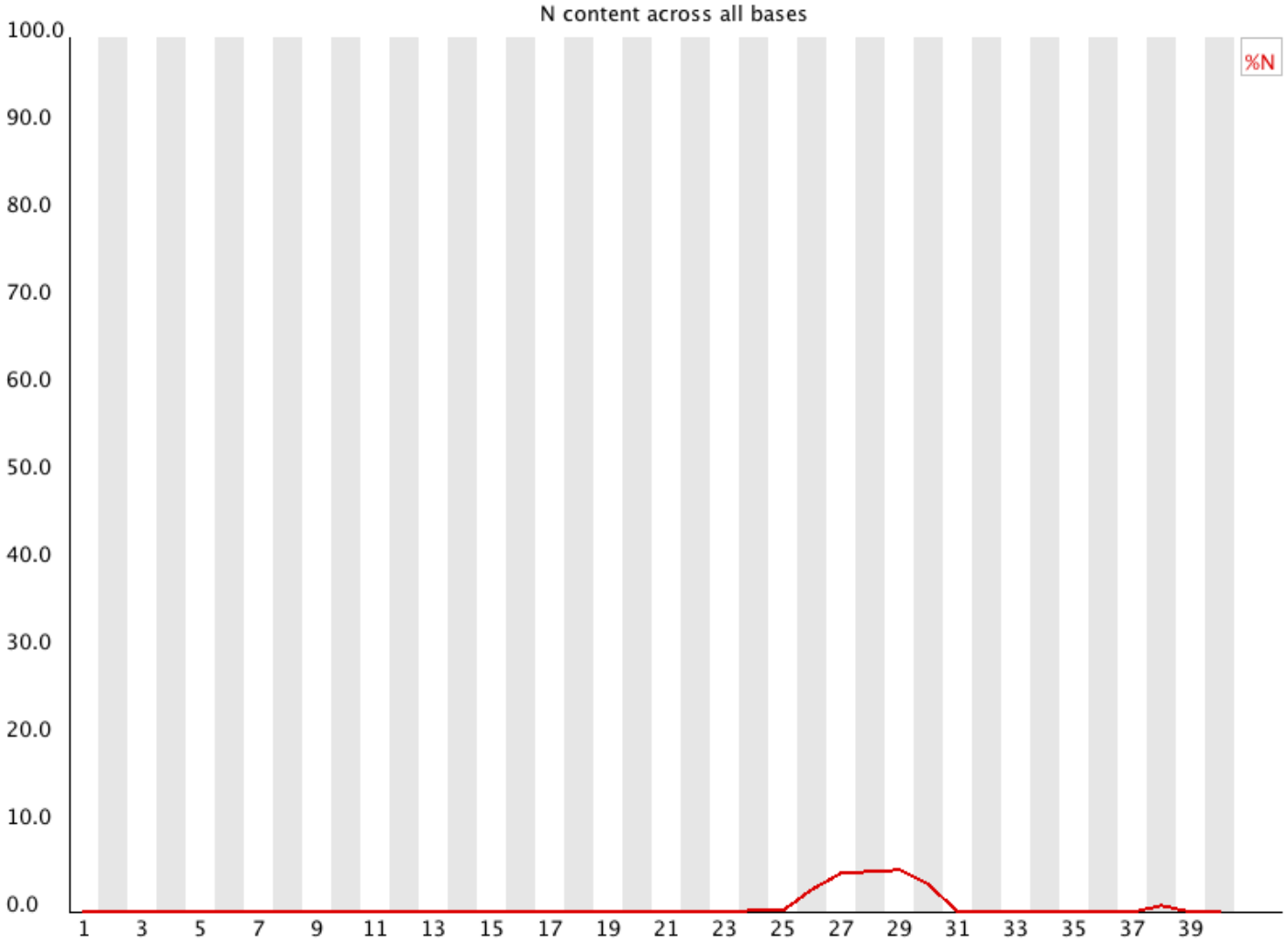
sprawdzenie preferencji w zawartości GC wzdłuż odczytów

per sequence GC content



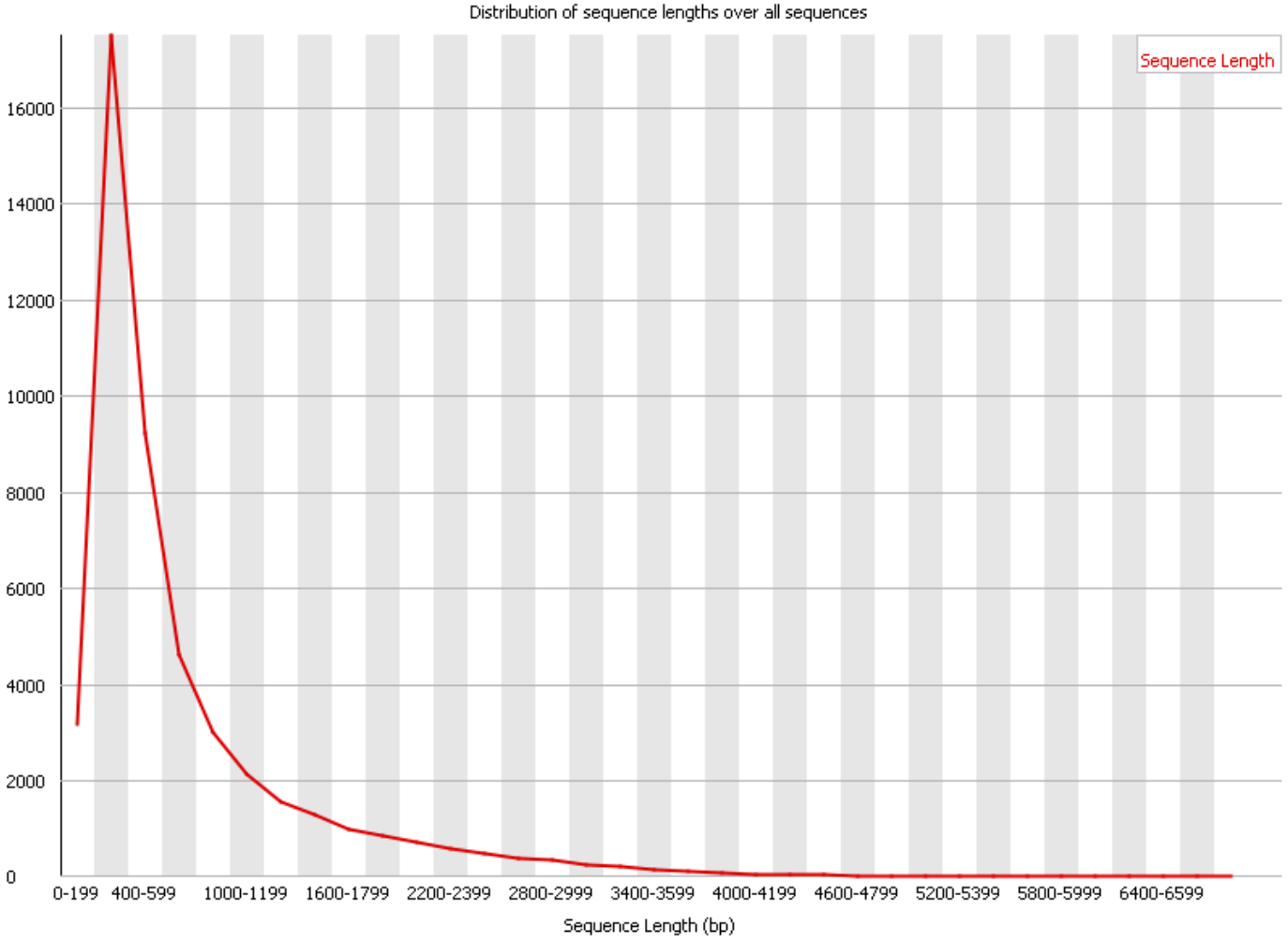
sprawdzenie rozkładu zawartości GC w odczytach

per base N content



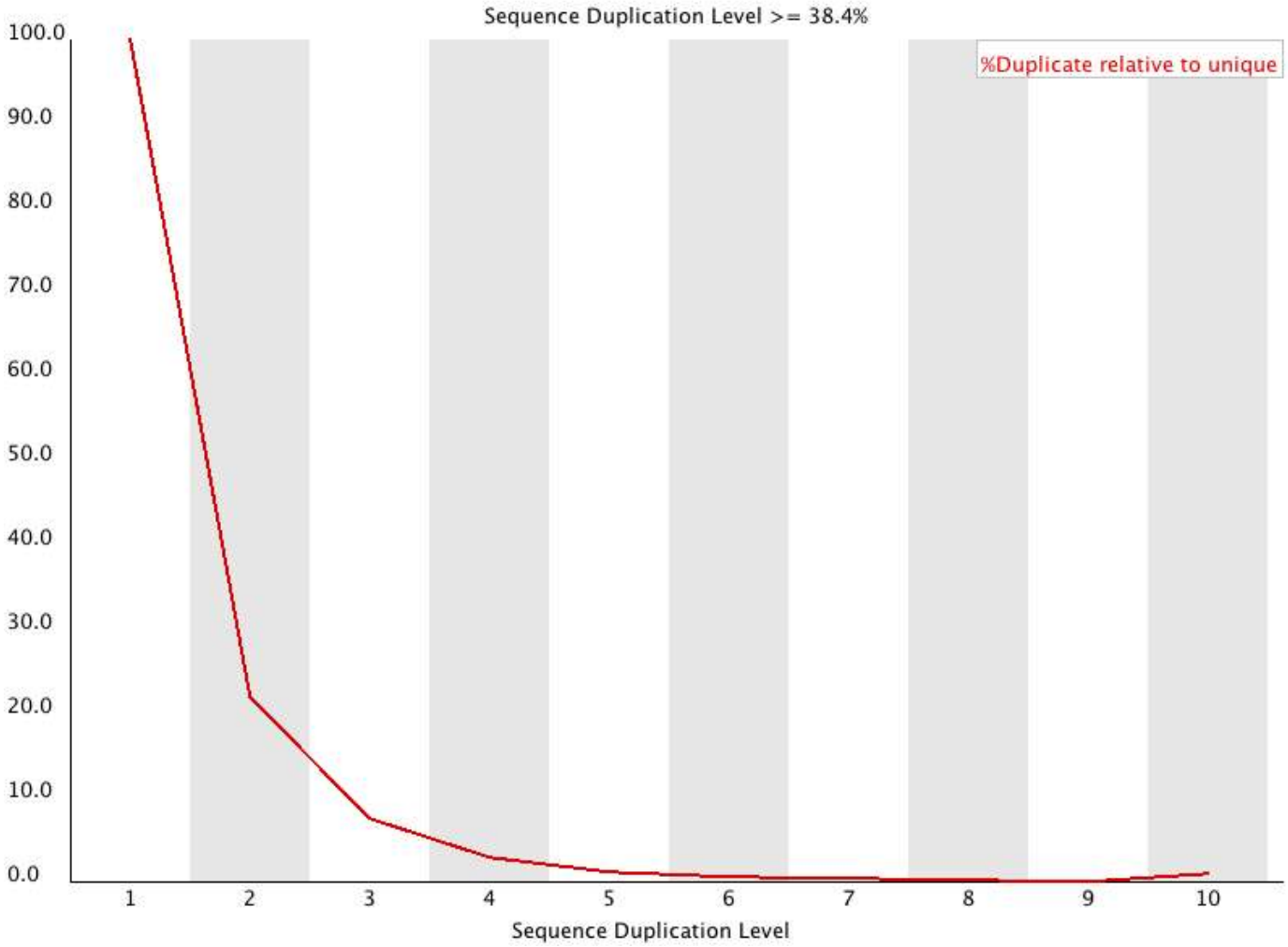
sprawdzenie rozkładu zawartości N wzdłuż odczytów

sequence length distribution



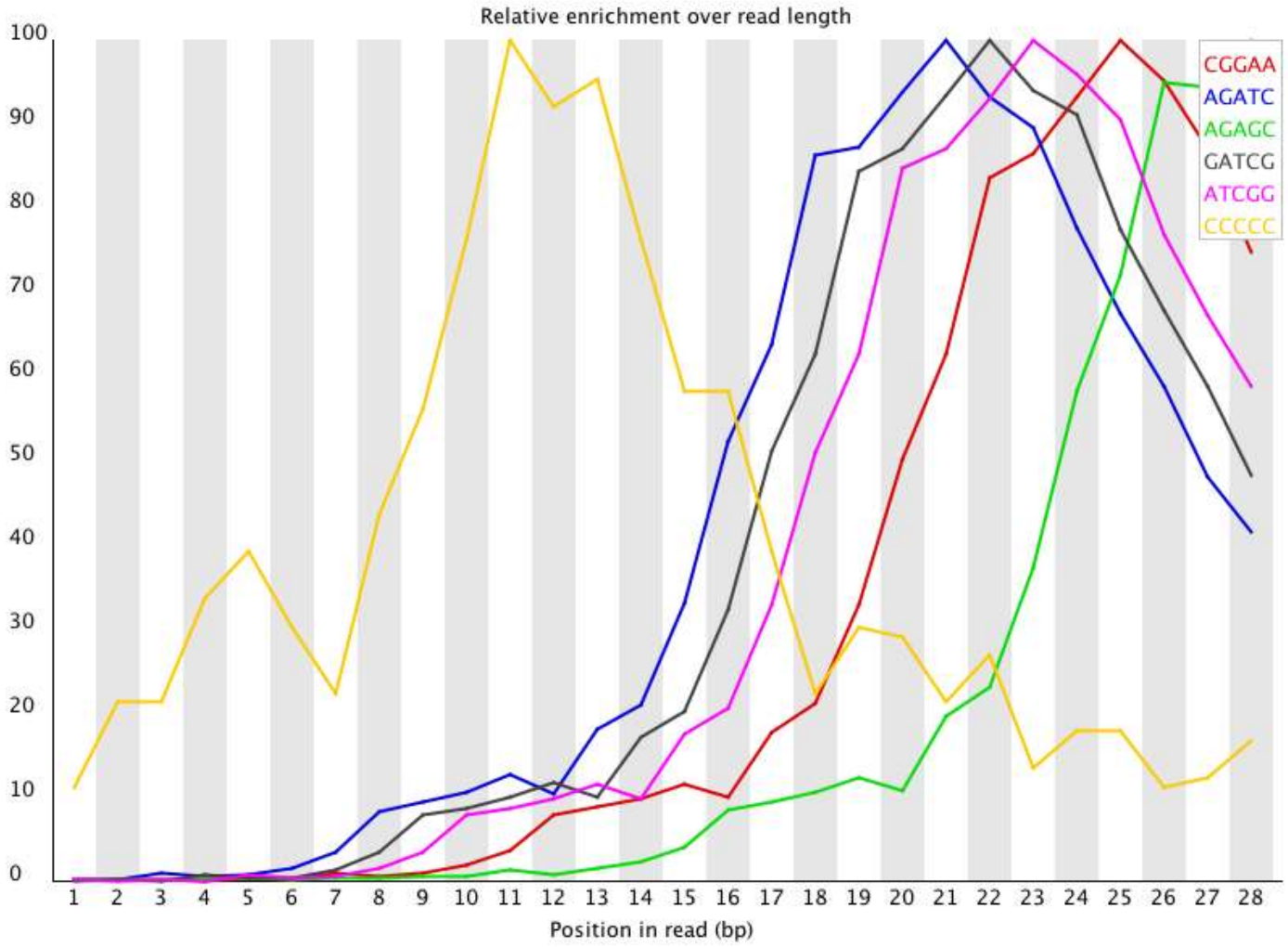
sprawdzenie rozkładu długości odczytów

sequence duplication levels



sprawdzenie rozkładu ilości duplikatów

overrepresented sequences and k-mers



analiza dokładnych i niedokładnych duplikatów

przygotowanie zestawu odczytów

1. Demultipleksowanie zestawu odczytów

2. Skracanie odczytów
   - usuwanie adaptorów
   - usuwanie końców o niskiej jakości
   - usuwanie ogonów poli-A / poli-T

3. Filtrowanie odczytów
   1. na podstawie średniej / minimalnej jakości
   2. na podstawie zawartości informacyjnej sekwencji
   3. usuwanie duplikatów

4. Korekcja błędów specyficznych dla platformy

demultipleksowanie

Dwa sposoby wstawiania znaczników:

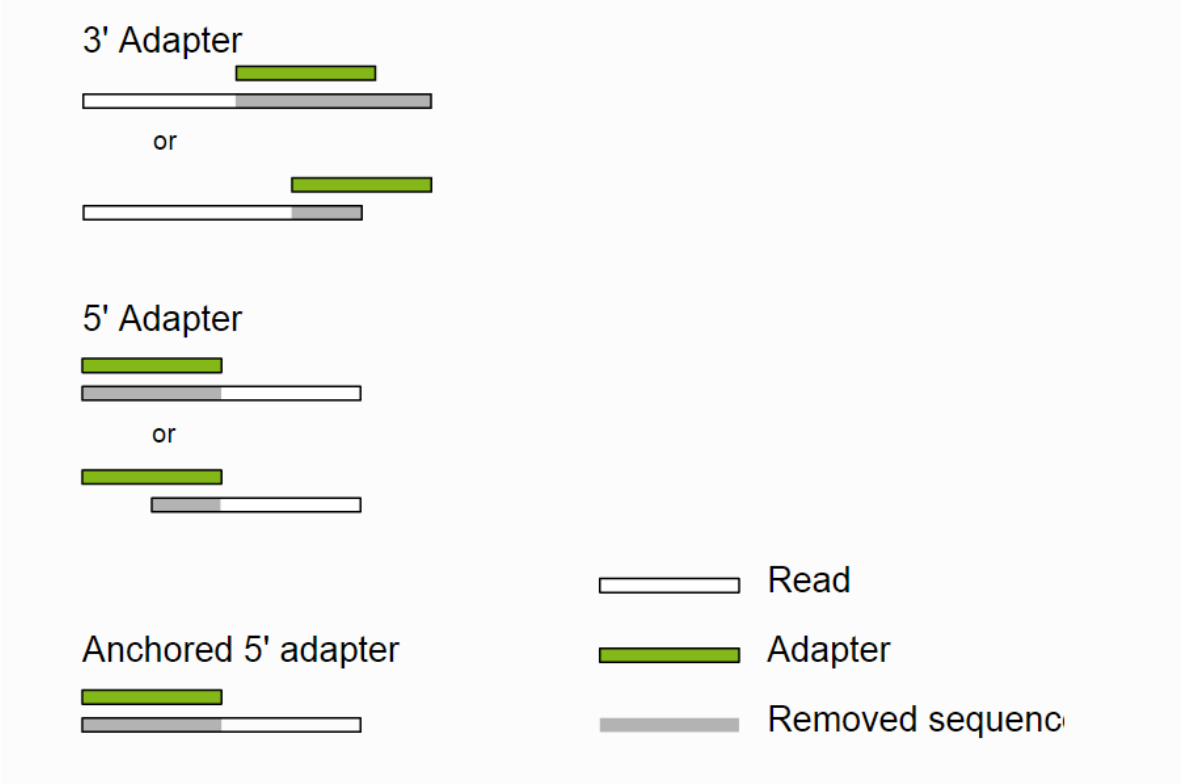
1. Systemowe znaczniki odczytywane za pomocą osobnej reakcji

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

2. „Domowe” znaczniki dodawane do adaptorów

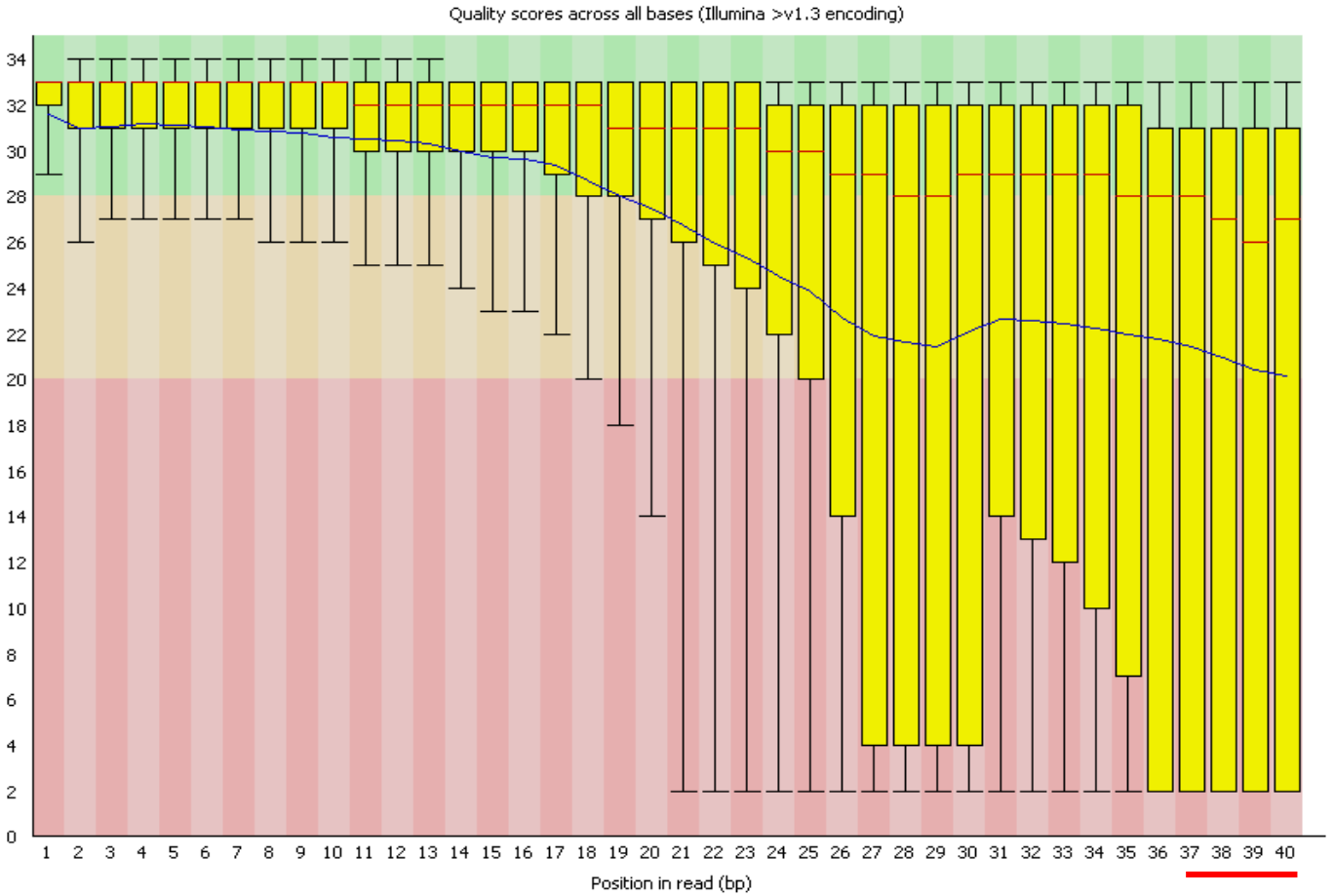
```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

usuwanie adaptorów



rekomendowanmy program: cutadapt

usuwanie końców o niskiej jakości



filtrowanie odczytów o niskiej jakości

Średnia jakość

punkty Q w odczycie	Średnie Q	Spodziewana liczba błędów
140 x Q35 + 10 x Q2	33	6.4 !
150 x Q25	25	0.5

Dla Q = 2 spodziewana liczba błędów P = 0,5

Dystrybucja jakości !

Filtrowanie odczytów w których mniej niż 80% zasad ma Q > 20 (15-25)