Przydatność technologii Sekwencjonowania Nowej Generacji (NGS) w kolekcjach Banków Genów

Joanna Noceń

Kinga Smolińska

Marta Puchta

Kierownik tematu:

prof. dr hab. Jerzy H. Czembor

SEKWENCJONOWANIE

I generacji

metoda Sangera,
metoda chemicznej degradacji Maxama-Gilberta

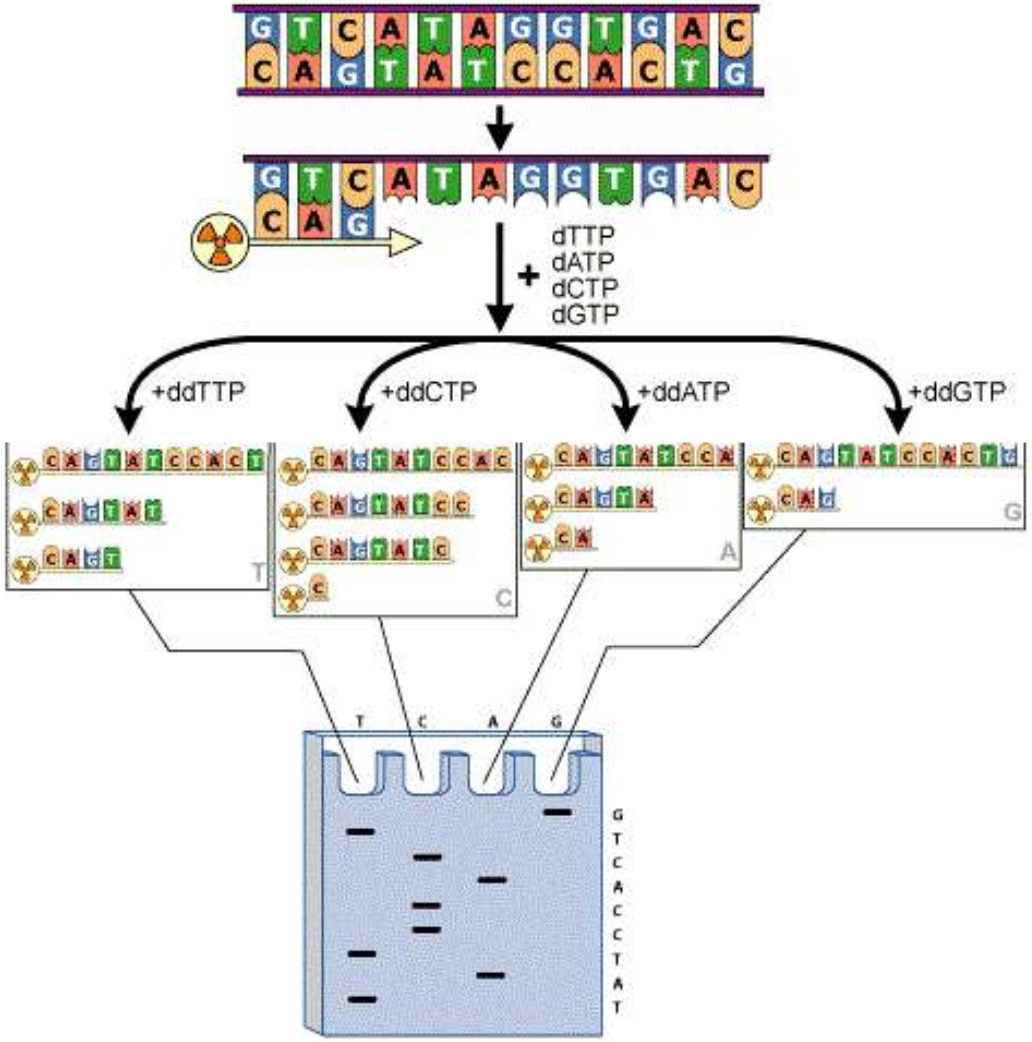
II generacji

metody NGS

III generacji

technologia SMS (ang. Single Molecule Sequencing)

Metoda Sangera



Technologia NGS … (ang. Next Generation Sequencing – Sekwencjonowanie Nowej Generacji)

- Metody sekwencjonowania II generacji proponują szereg zmian w porównaniu z metodami sekwecjonowania I genracji:

(+)

- Prostsze przygotowanie biblioteki
- Platformy do sekwencjonowania (oparte o macierze) pozwalają na sekwencjonowanie znacznie większej liczby odczytów
- Unieruchomienie sekwencjonowanych cząsteczek na jednej powierzchni umożliwia przeprowadzenie reakcji w małej objętości (mniejsze koszty sekwencjonowania)

(-)

- Krótsze odczyty
- Gorsza jakość i dokładność odczytów
- Trudna algorytmicznie analiza danych

Sekwencjonowanie Nowej Generacji - Platformy

illumina

SOLiD®

PACIFIC BIOSCIENCES®

Roche 454 SEQUENCING

ion torrent by life technologies™

Sequencing By Synthesis

- Pirosekwencjonowanie, przez syntezę z kaskadą 4 reakcji enzymatycznych
- Sekewencjonowanie mostkowe, przez syntezę z odwracalną terminacją
- Sekwencjonowanie typu Ion Torrent, przez syntezę z detekcją protonów na układzie scalonym
- Sekwencjonowanie przez ligację, fluorescencyjne znakowanie krótkich oligonukleotydów
- Nowe pomysły: sekwencjonowanie przez spektrometrię mas, sekwencjonowanie w nanoporach

illumina

Sekwenator MiSeq:

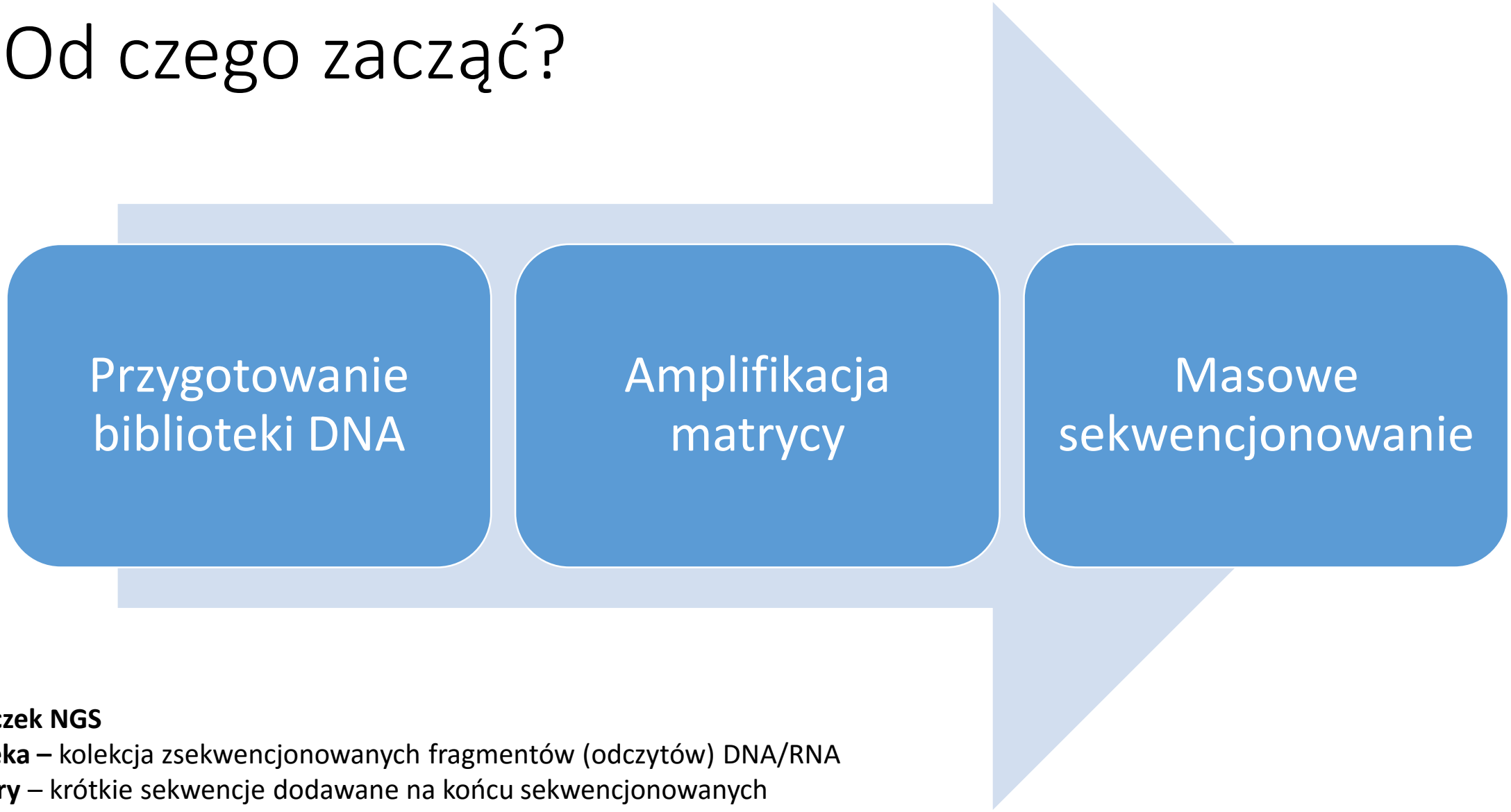
- Technologia SBS (*ang. Sequencig By Synthesis*) sekwencjonowanie mostkowe
- Integracja etapów- amplifikacja, sekwencjonowanie i wstępna analiza danych- w jednym urządzeniu
- Odczyty do 2x 300pz
- 15Gb produktu w trakcie jednego cyklu
- Prosta obsługa
- Wysoka dokładność odczytów

Zastosowanie MiSeq:

- ✓ Sekwencjonowanie celowane
- ✓ Metagenomika
- ✓ Sekwencjonowanie małych genomów
- ✓ Celowana ekspresja genów
- ✓ Sekwencjonowanie amplikonów
- ✓ Identyfikacja różnic w poziomie ilości kopi
- ✓ Rearanżacje chromosomalne

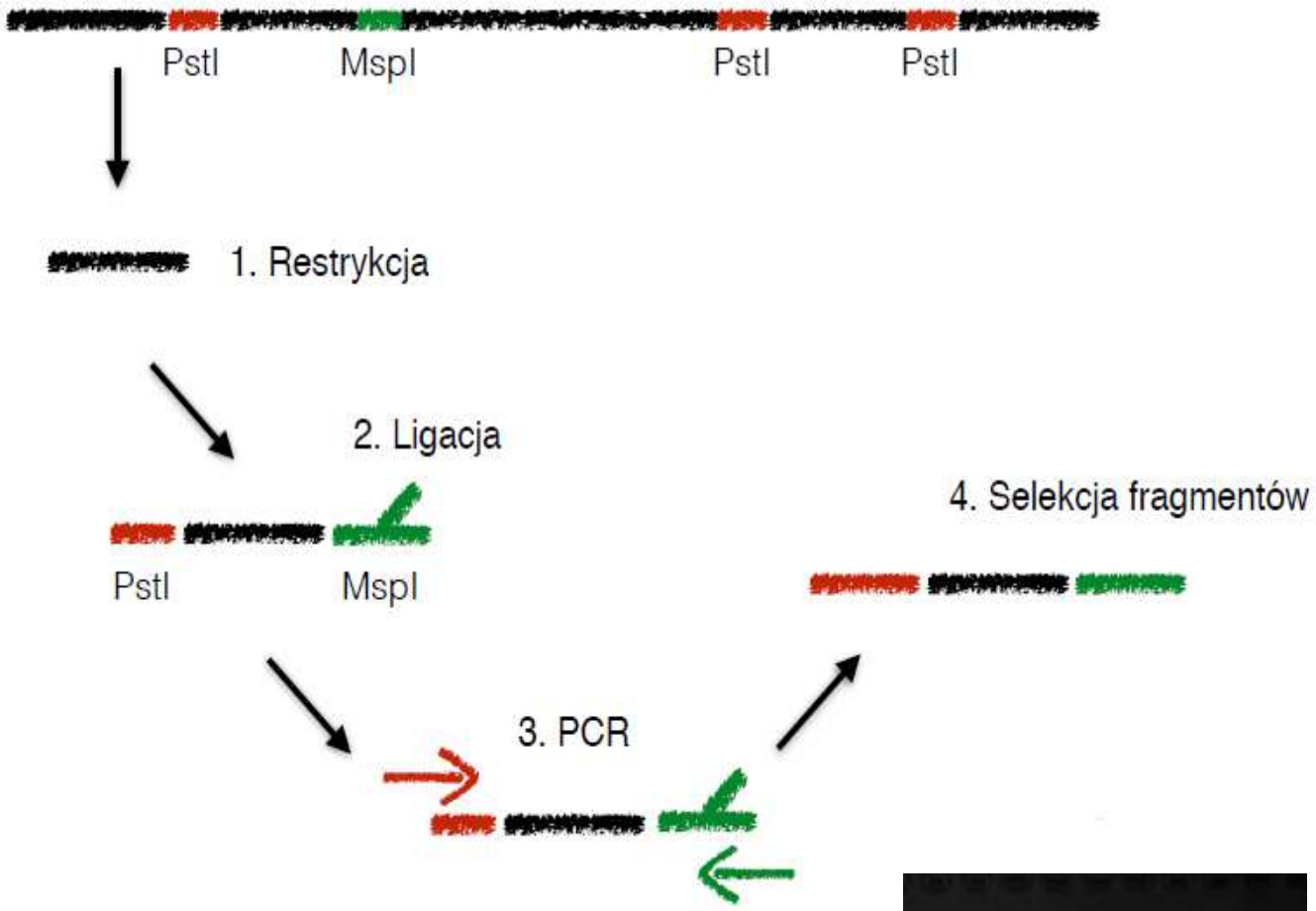
Sekwenator MiSeq firmy Illumina został zakupiony w ramach Programu Wieloletniego IHAR-PIB
Obszar 1; Zadanie 1.2; Temat ,,Charakterystyka i diagnostyka molekularna wybranych zasobów genowych roślin uprawnych i towarzyszących im chwastów" finansowanego przez Ministerstwo Rolnictwa i Rozwoju Wsi.

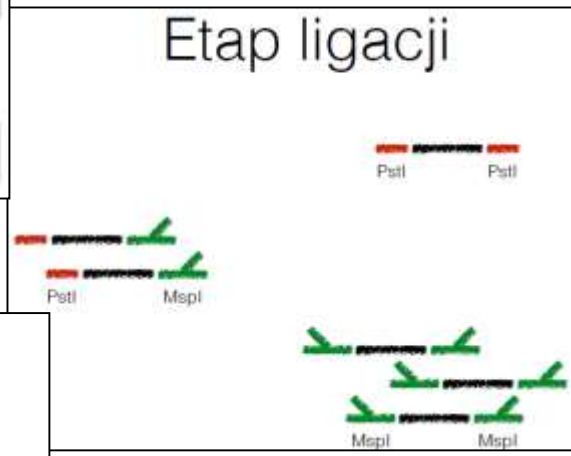
Od czego zacząć?

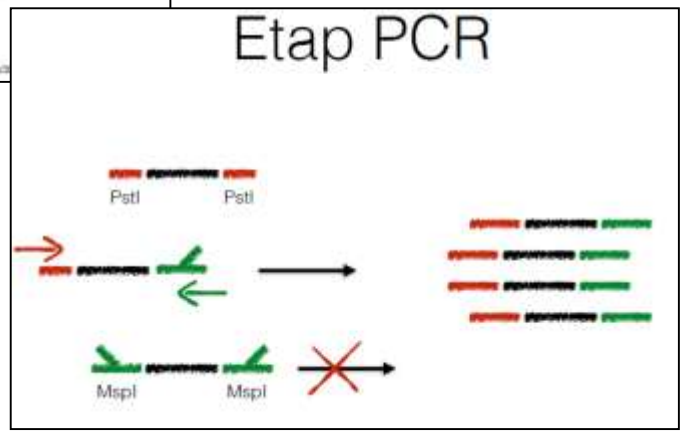


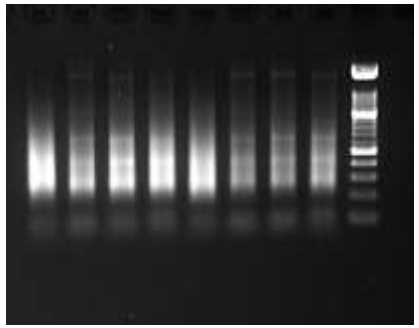
Słowniczek NGS
Biblioteka – kolekcja zsekwencjonowanych fragmentów (odczytów) DNA/RNA
Adaptery – krótkie sekwencje dodawane na końcu sekwencjonowanych fragmentów. Bardzo często usuwane są już przez sekwenator, jednak niekiedy trzeba usuwać je samodzielnie.

Protokół ddRad



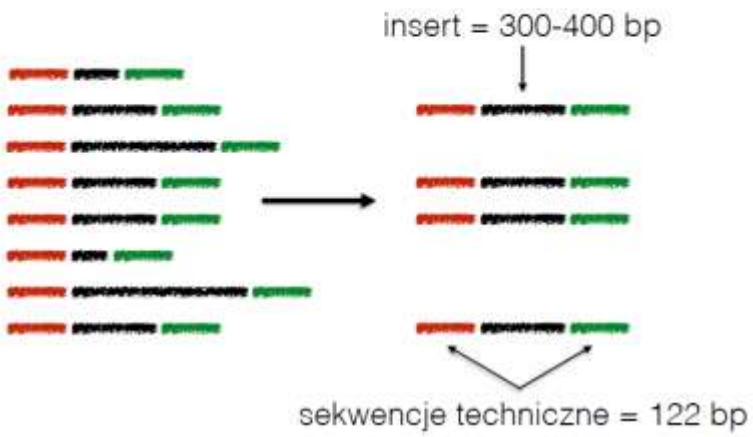






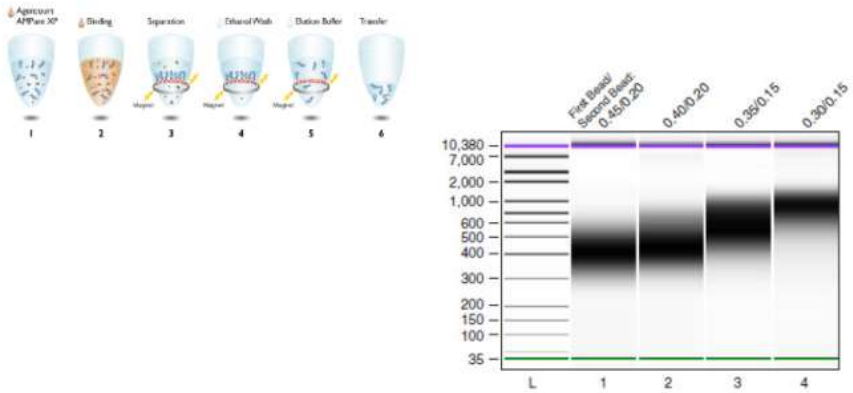
Wg dr Tomasza Suchana

Selekcja fragmentów



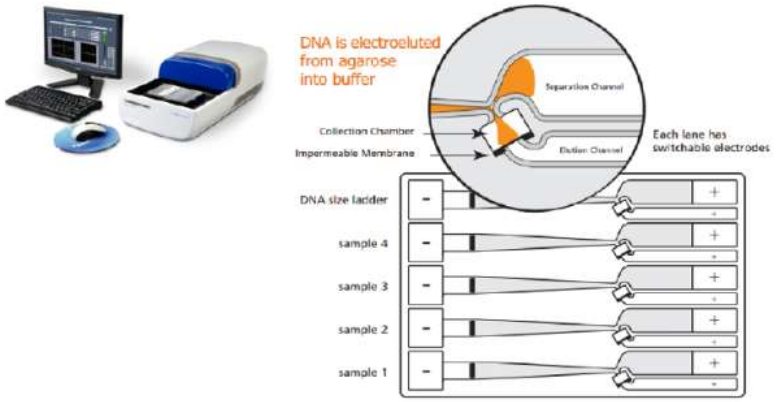
Selekcja fragmentów

AMPure XP



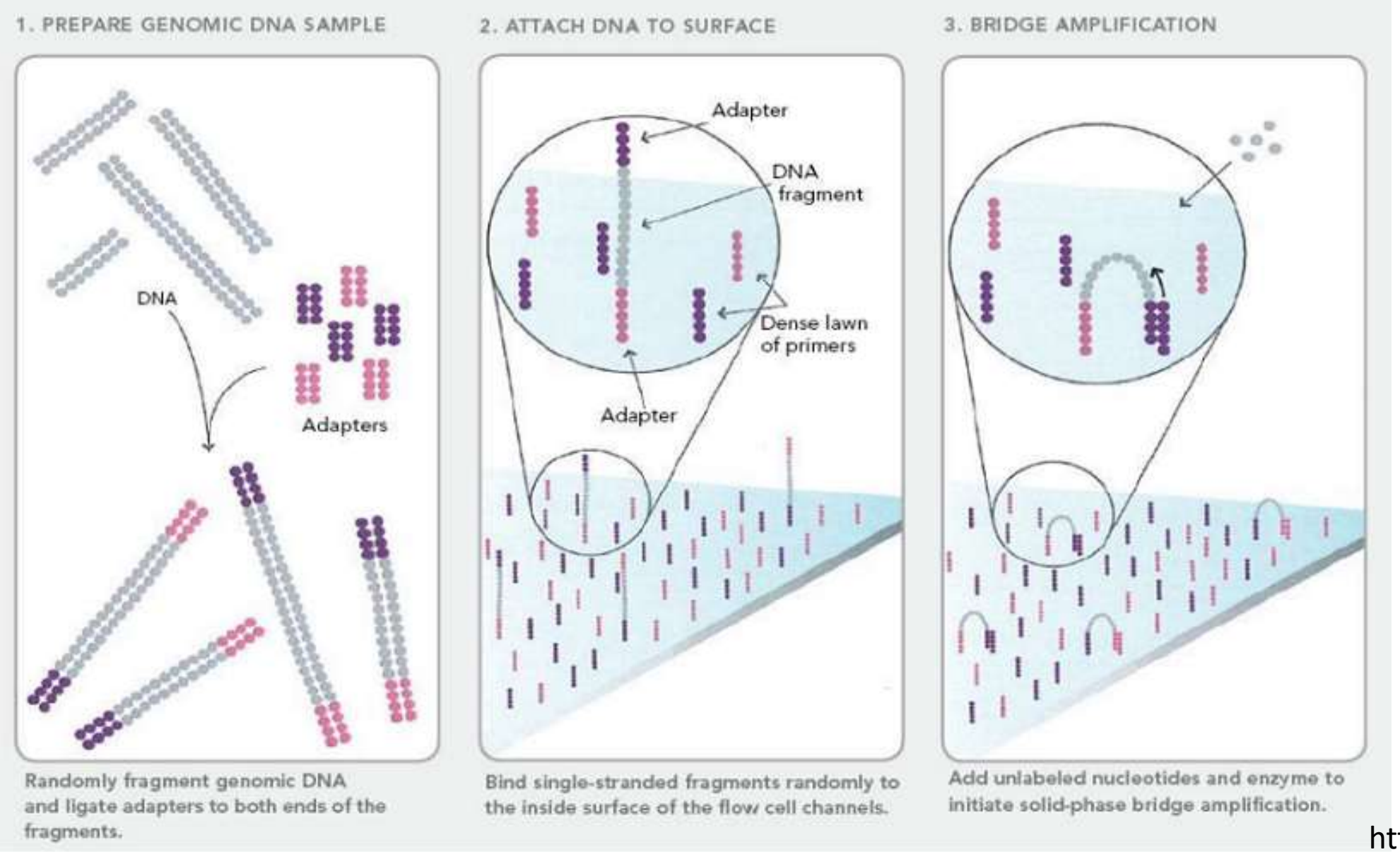
Selekcja fragmentów

Pippin Prep

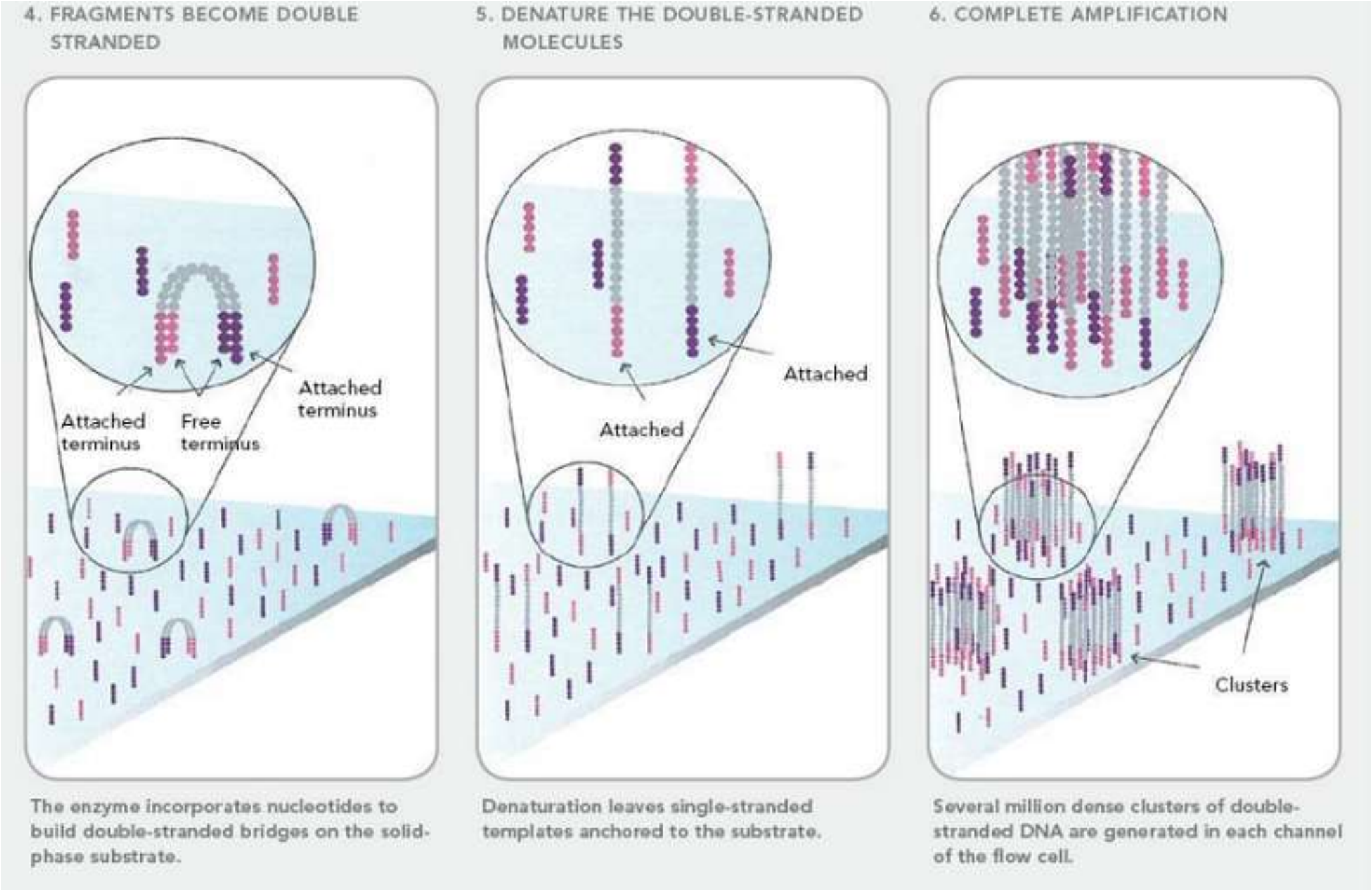


Wg dr Tomasza Suchana

Sekwencjonowanie w Technologii Illumina (SBS)

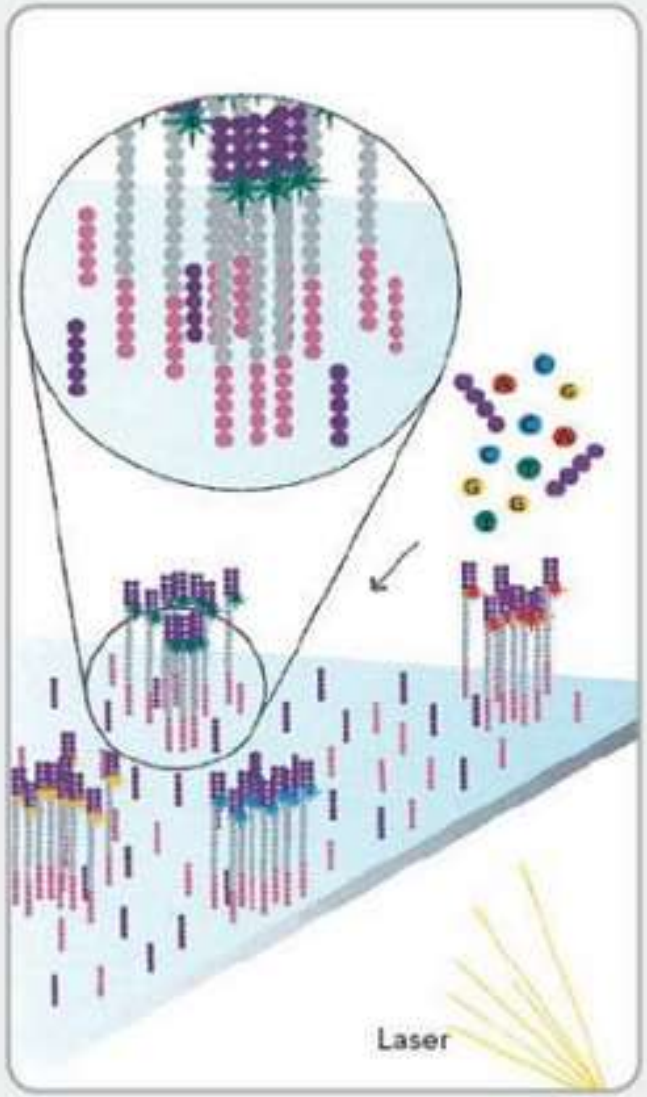


https://www.illumina.com/



https://www.illumina.com/

7. DETERMINE FIRST BASE

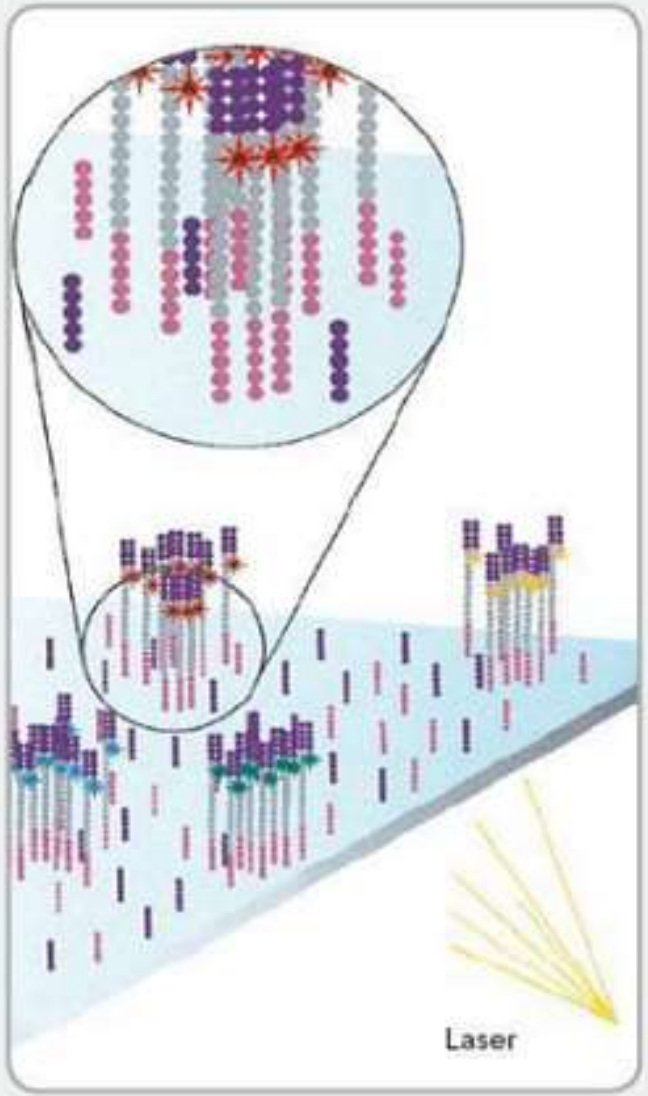


First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.

8. IMAGE FIRST BASE

After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.

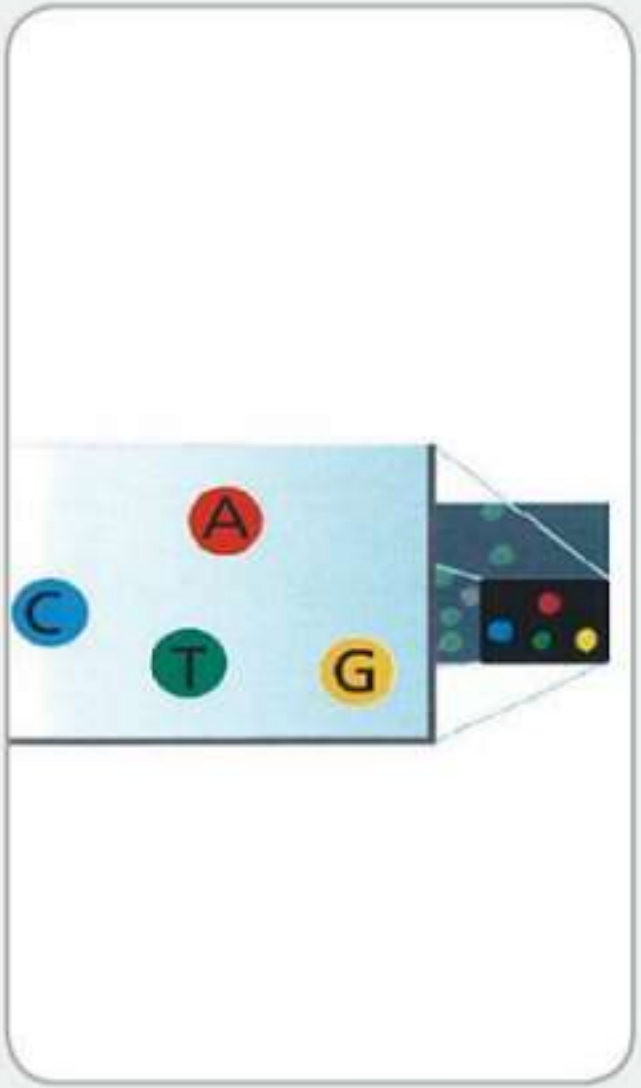
9. DETERMINE SECOND BASE



Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.

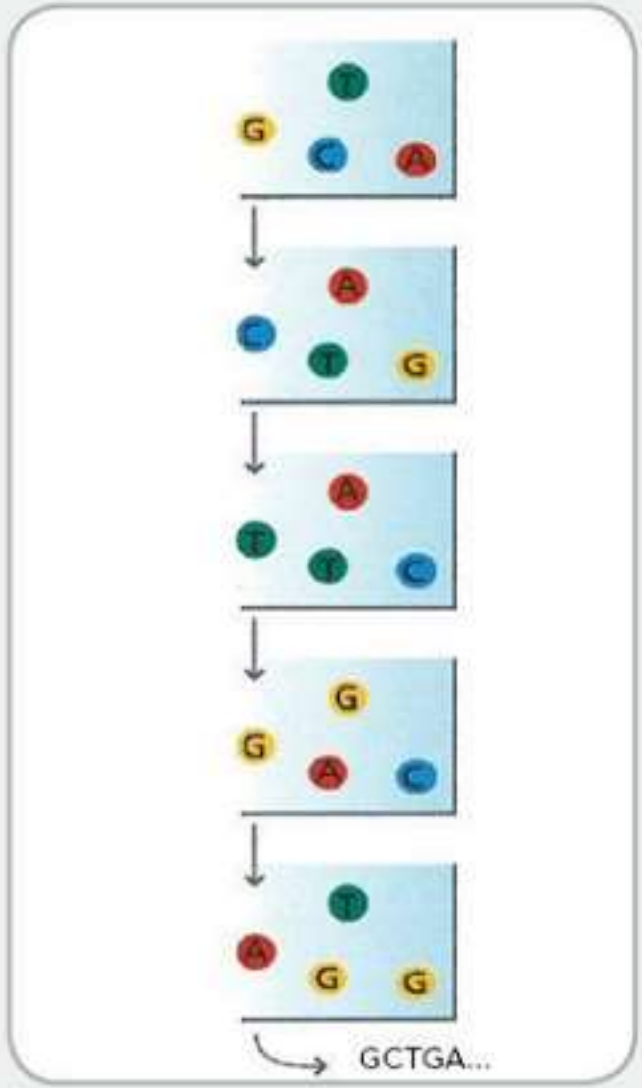
https://www.illumina.com/

10. IMAGE SECOND CHEMISTRY CYCLE



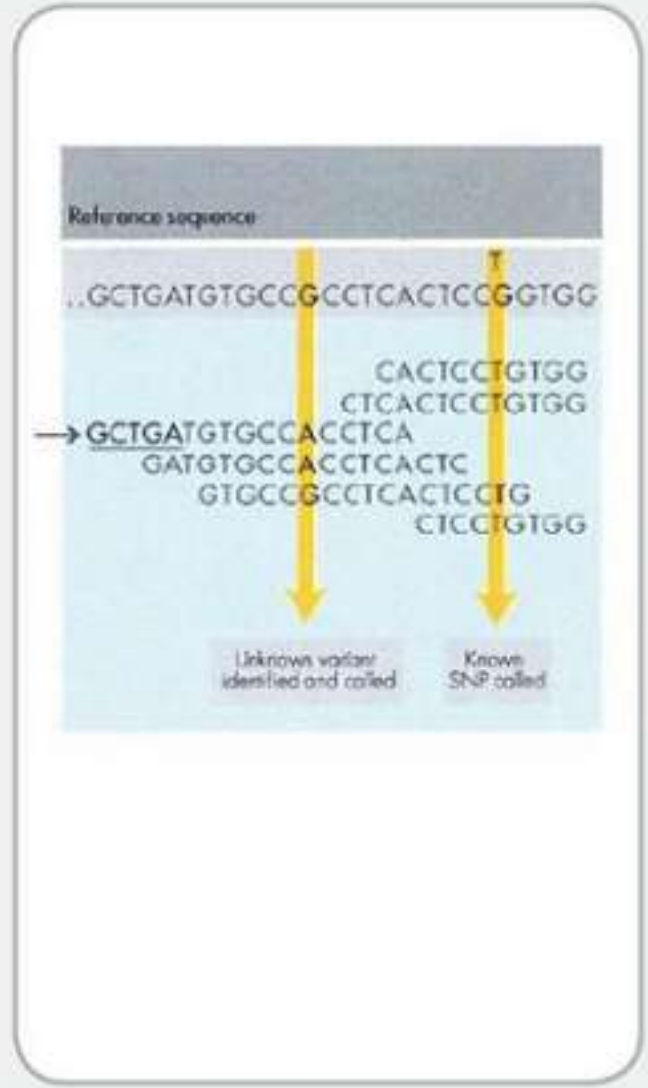
After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

11. SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES



Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at time.

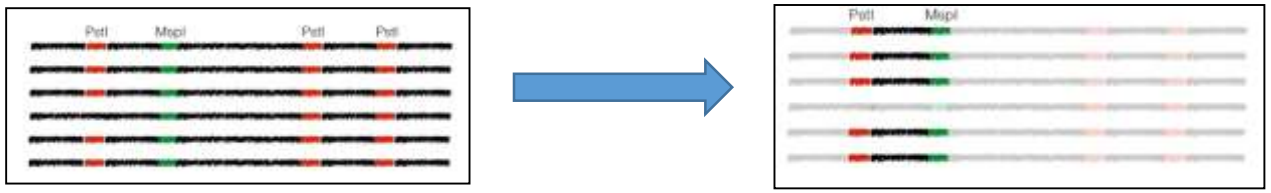
12. ALIGN DATA



Align data, compare to a reference, and identify sequence differences.

https://www.illumina.com/

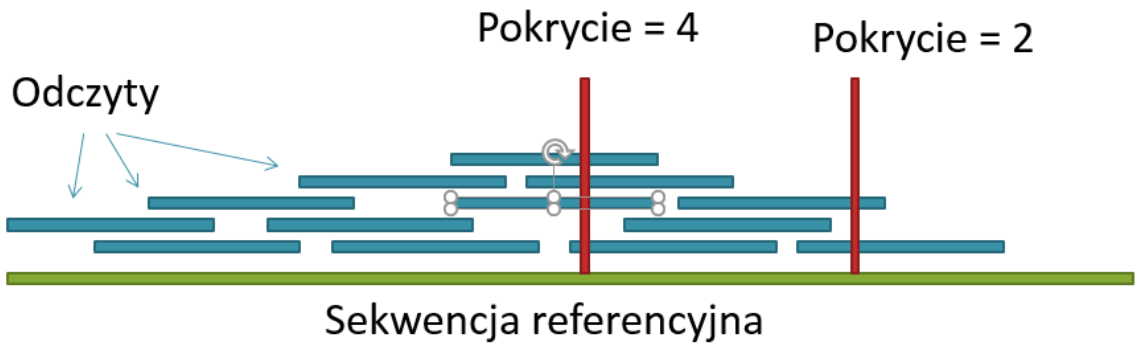
Analiza bioinformatyczna



Słowniczek NGS

Odczyty sparowane (ang. paired reads) – Szczególnie przydatne do mapowania fragmentów genomu z sekwencjami powtórzonymi

Pokrycie – liczba zmapowanych odczytów, przypadających na daną pozycję w sekwencji referencyjnej



Wg dr Wioleta Drobik-Czwarno

Pojedynczy rekord formatu FASTQ

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

>gnl|SRA|SRR191637.1221 HWUSI-EAS1600:IB_HJ1:8_5_2010:0:5:21:1153:5906
CGTACNNNNAATAGTTTAAACTGTTTGGTTCGTATG

One channel quality score

30 30 30 27 30 2 2 2 2 27 28 28 24 27 28 32 32 32 32 32
32 32 32 32 32 32 32 32 32 32 32 32 32 32 32 32

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1,000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%

Podstawowe etapy analizy NGS

1. Kontrola jakości surowych danych (**format fastq**)
   - Jakość odczytów, jakość par zasad w odczytach
2. Mapowanie do genomu referencyjnego:
   - Indeksowanie genomu referencyjnego
   - Mapowanie **– format fastq > SAM**
   - Zmiana formatu **SAM na BAM**
3. Obróbka pliku BAM: sortowanie, indeksowanie, formatowanie
4. Wykrywanie wariantów (generujemy plik **VCF**):
   - SNP – polimorfizm pojedynczego nukleotydu
   - INDEL – krótkie delecje i insercje
   - Warianty strukturalne (np. CNV)
5. Dalsze kroki zależnie od celu analizy

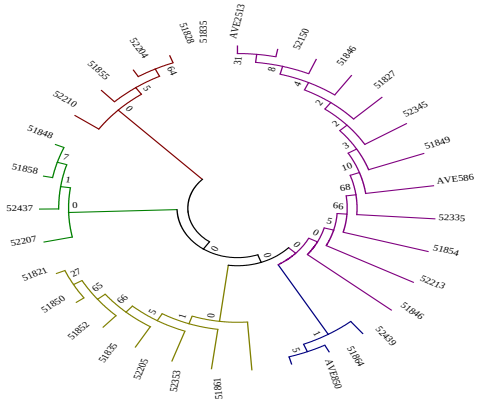
Dotychczasowe osiągnięcia KCRZG

Wykonawca tematu-dr Maja Boczkowska, Kierownik tematu- prof. dr hab. Jerzy H Czembor

- **Identyfikacja filogenetyczna owsa:**

Barkoding (DNA chloroplastowe):

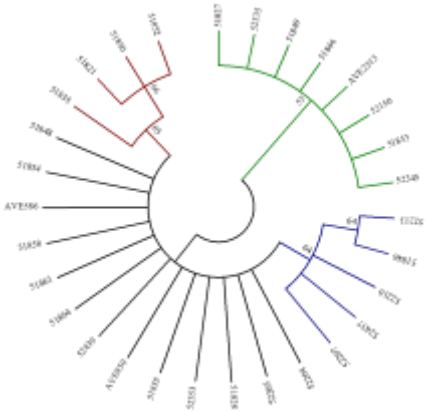
✓ matK
✓ trnH-psbA
✓ trnL-trnF
✓ psbK-psbI
✓ atpF-atpH



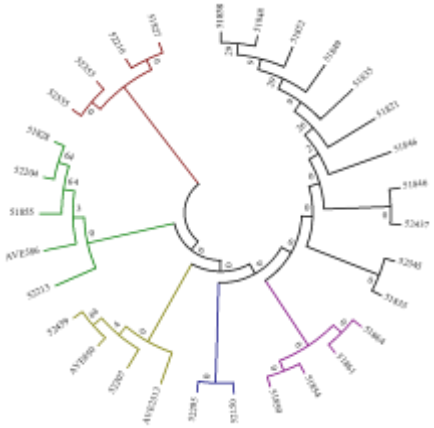
Rysunek 1. Drzewo filogenetyczne utworzone na podstawie regionu *trnH-psbA*

- **Analiza zróżnicowania genetycznego roślin z gatunku Avena (ISSR, SSR, AFLP, morfologia, cechy użyteczne rolniczo, SRAP)**

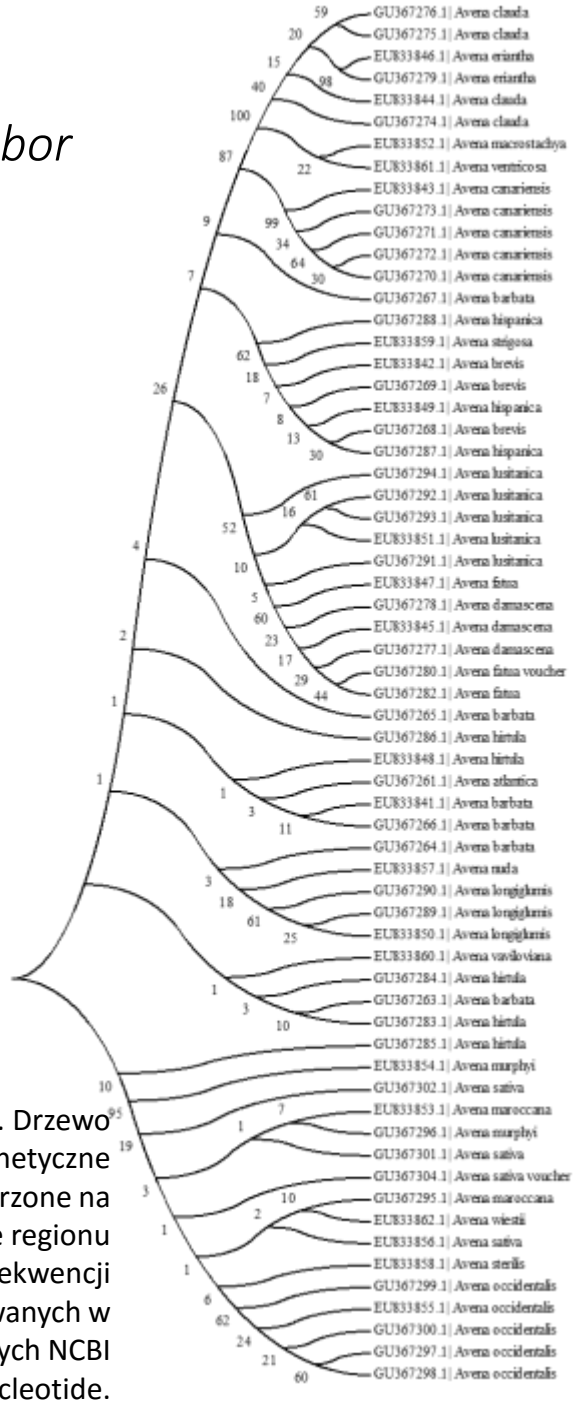
- **Trwają prace nad walidacją sekwenatora Illumina MiSeq oraz pierwsze próby sekwencjonowania regionów barkodowych**



Rysunek 2 Drzewo filogenetyczne utworzone na podstawie regionu *psbK-psbI*



Rysunek 3 Drzewo filogenetyczne utworzone na podstawie regionu *atpF-atpH*



Rysunek 4. Drzewo filogenetyczne utworzone na podstawie regionu *matK* dla sekwencji zdeponowanych w bazie danych NCBI nucleotide.

Publikacje:

- ***Boczkowska M., Łapiński B., Kordulasińska I., Dostatny D. F., Czembor J. H. 2016.*** *Promoting the Use of Common Oat Genetic Resources through Diversity Analysis and Core Collection Construction. PLoS ONE 11(12): e0167855.*
- ***Boczkowska M., Wolko B., Dostatny D. F., 2016.*** *Morphological, isoenzymatic and ISSRs-based description of diversity of eight sand oat (Avena strigosa Schreb.) landraces. Genet. Resour. Crop Evol.; pp. 1-14.*
- ***Boczkowska, M., & Onyśk, A. 2016.*** *Unused genetic resources: a case study of Polish common oat germplasm. Annals of Applied Biology DOI: 10.1111/aab.12289*

Dalsze prace …

- Analiza SNP metodą ddRAD-Seq w obrębie gatunku *Avena* oraz innych roślin upranych
- Wykorzystanie protokół ddRAD-Seq do:
  - Badań nad rozwojem ewolucyjnym
  - Analizach populacyjnych
  - Identyfikacji nowych markerów SSR w systemie MiSeq
- Wdrażanie nowych technik przygotowywania bibliotek z użyciem sekwenatora MiSeq Illumina
- Szukanie zmienności w akcesjach z gatunku *Avena macrostachya*

KONIEC

Zapraszamy do współpracy!!!