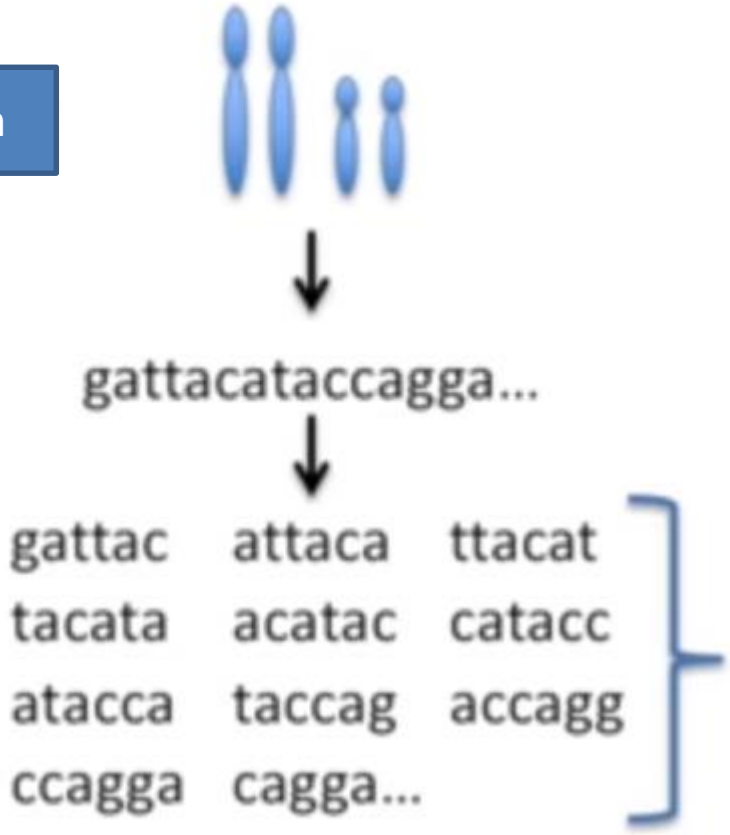
RNA-seq – Analiza danych

Alaingt odczytów do genomu

Genom



Dzielimy sekwencje genomu na krótkie fragmenty

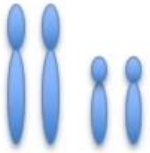
Genom

gattacataccagga...

gattac	attaca	ttacat
tacata	acatac	catacc
atacca	taccag	accagg
ccagga	cagga...	

Tworzymy index dla każdego fragmentu z informacją o jego lokalizacji w genomie

Genom



gattacataccagga...

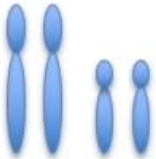
gattac attaca ttacat
tacata acatac catacc
atacca taccag accagg
ccagga cagga...

Tworzymy index dla każdego fragmentu z informacją o jego lokalizacji w genomie

Odczyt sekwencji

ACACGACGATGAG...

Genom



gattacataccagga...

gattac attaca ttacat
tacata acatac catacc
atacca taccag accagg
ccagga cagga...

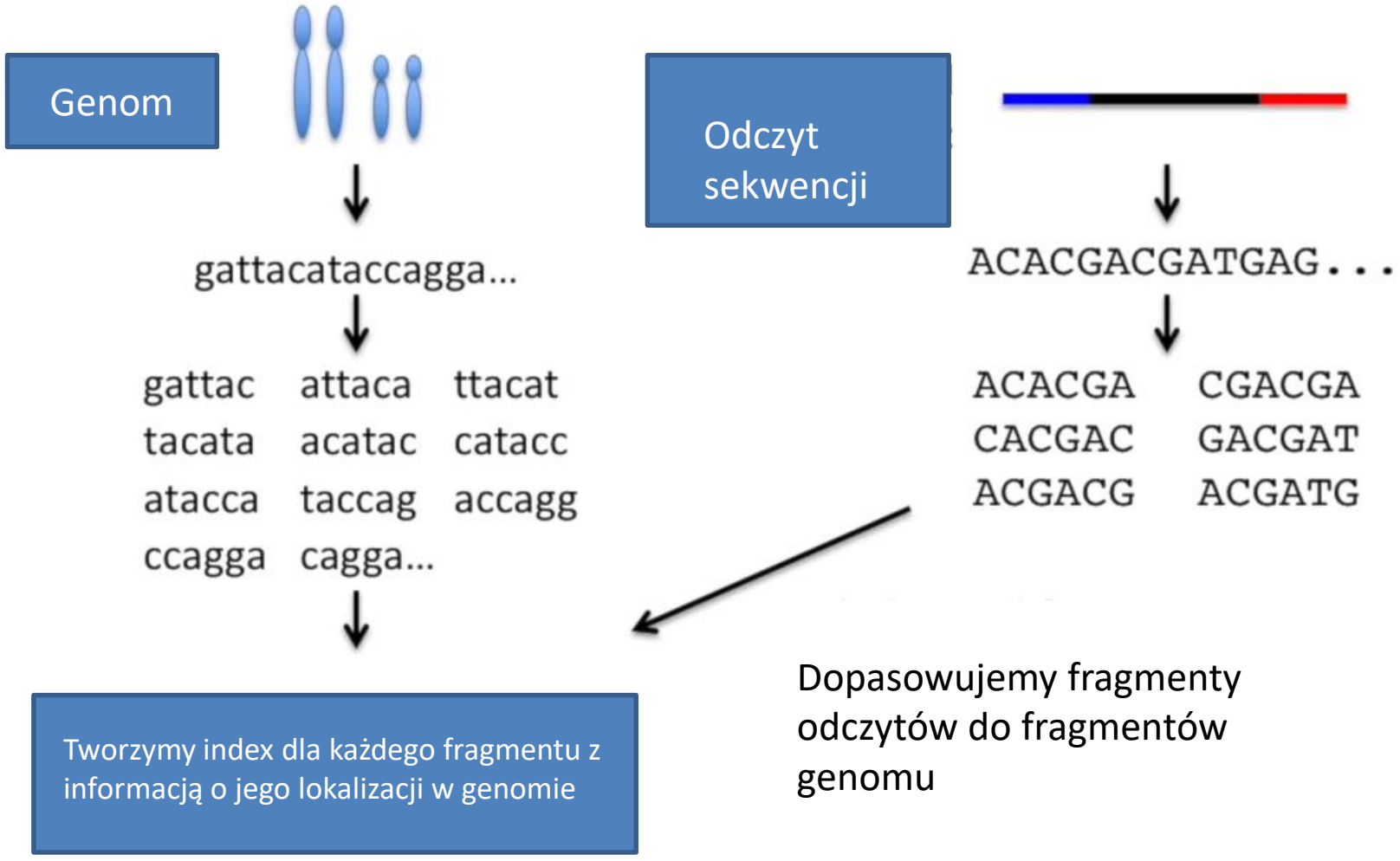
Tworzymy index dla każdego fragmentu z informacją o jego lokalizacji w genomie

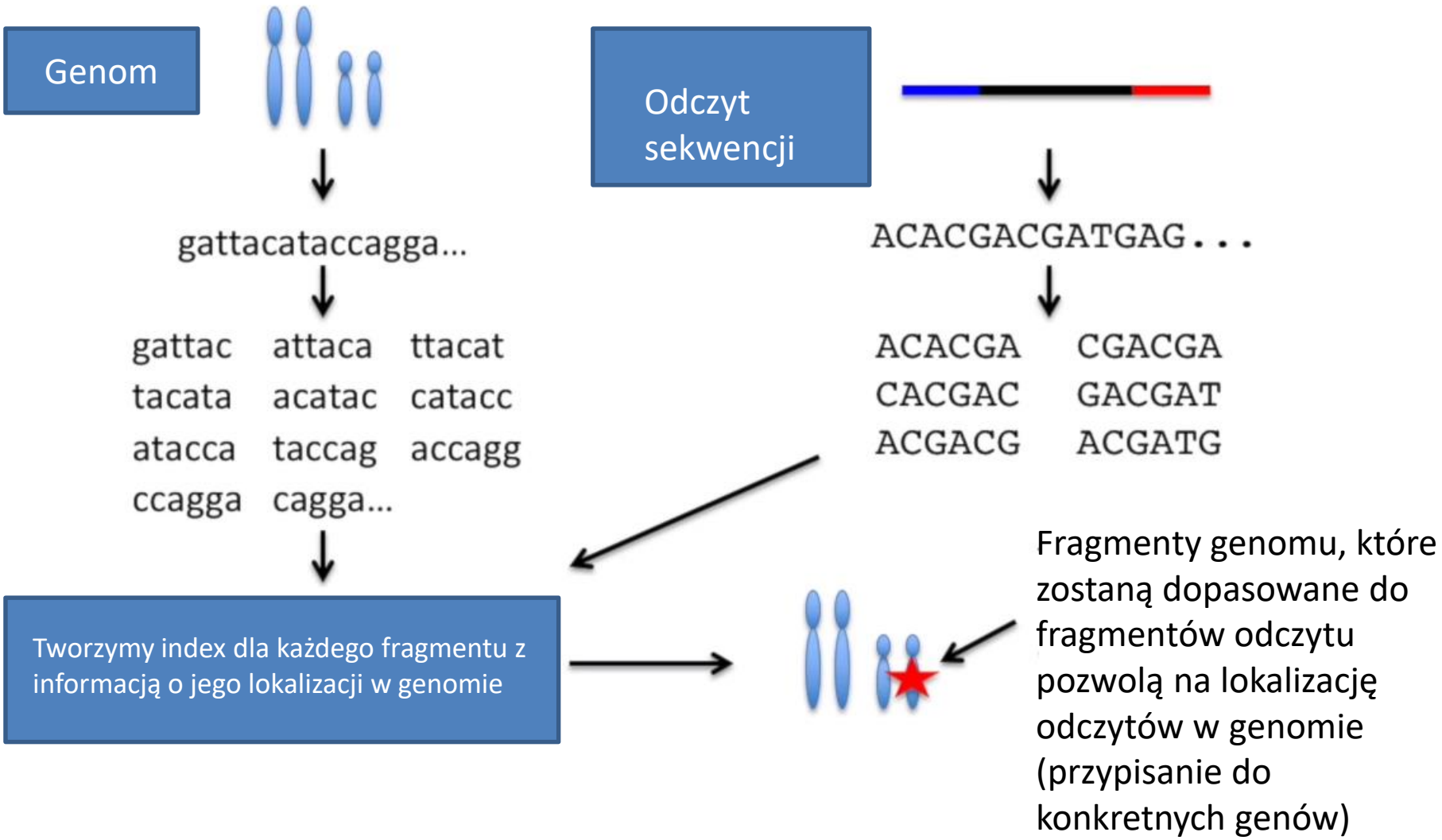
Odczyt sekwencji

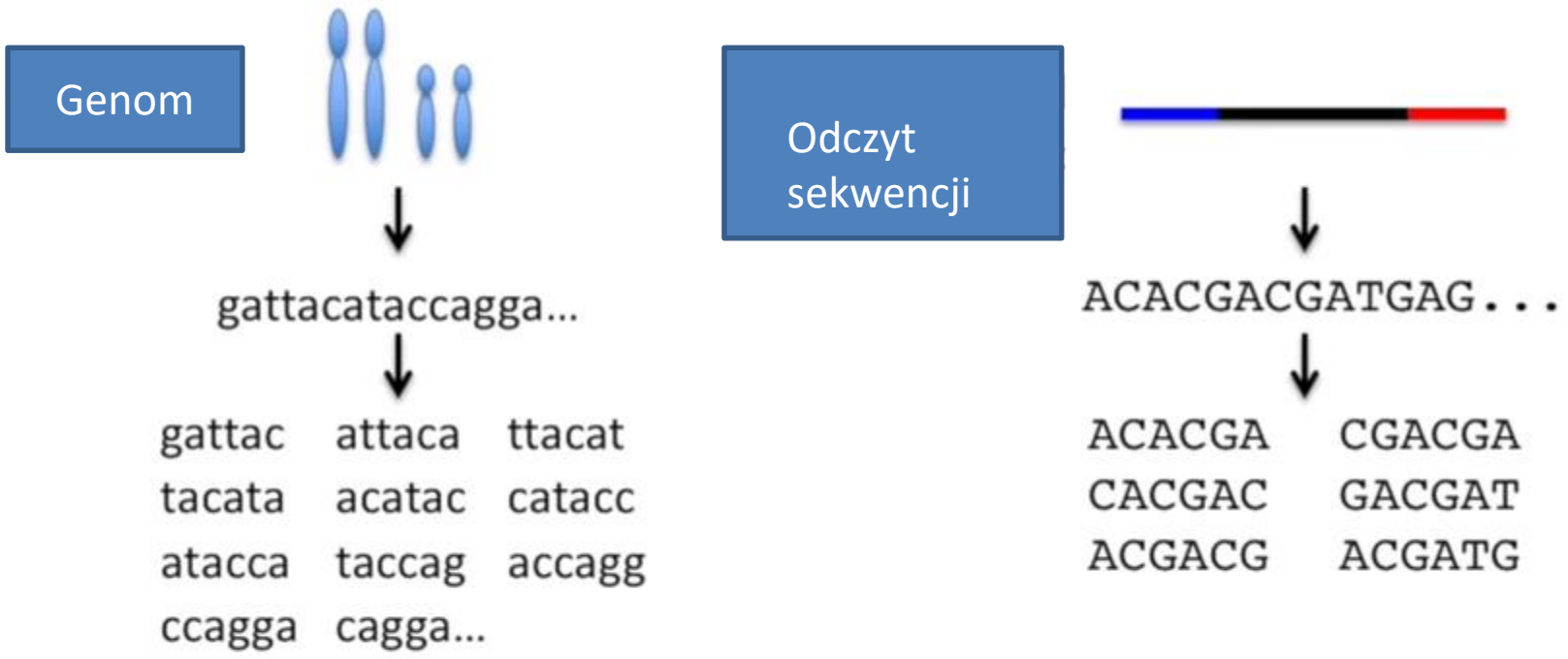
ACACGACGATGAG...

Dzielimy sekwencje odczytu na krótkie fragmenty

ACACGA CGACGA
CACGAC GACGAT
ACGACG ACGATG

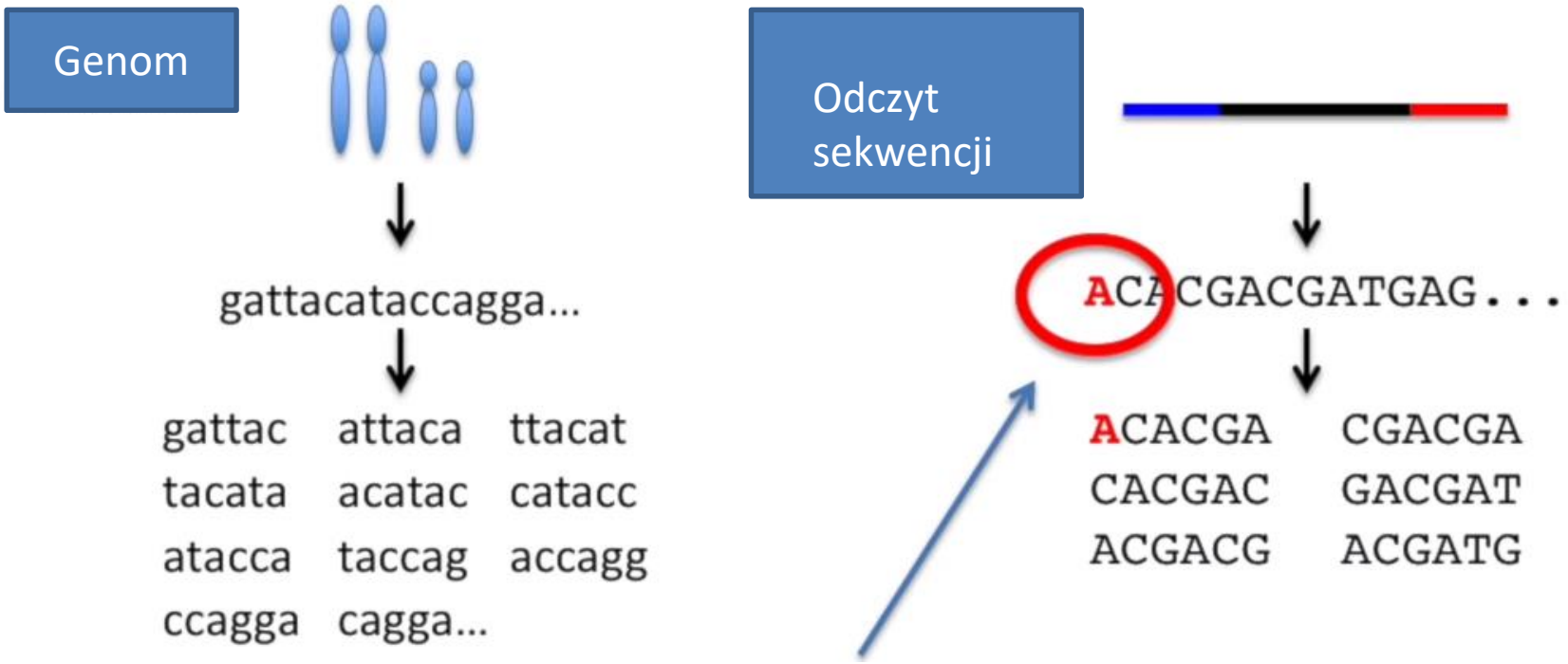

Genom
gattacataccagga...
gattac attaca ttacat
tacata acatac catacc
atacca taccag accagg
ccagga cagga...
Odczyt sekwencji
ACACGACGATGAG...
ACACGA CGACGA
CACGAC GACGAT
ACGACG ACGATG
Tworzymy index dla każdego fragmentu z informacją o jego lokalizacji w genomie
Dopasowujemy fragmenty odczytów do fragmentów genomu


Genom
gattacataccagga...
gattac attaca ttacat
tacata acatac catacc
atacca taccag accagg
ccagga cagga...
Tworzymy index dla każdego fragmentu z informacją o jego lokalizacji w genomie
Odczyt sekwencji
ACACGACGATGAG...
ACACGA CGACGA
CACGAC GACGAT
ACGACG ACGATG
Fragmenty genomu, które zostaną dopasowane do fragmentów odczytu pozwolą na lokalizację odczytów w genomie (przypisanie do konkretnych genów)



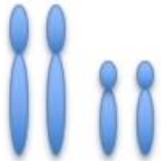
Ale dlaczego dzielimy sekwencje na krótkie odcinki?

Takie podejście pozwala nam na dopasowanie odczytów nawet jeżeli ich sekwencja nie pokrywa się w 100% z sekwencją genomu



Wyobraźmy sobie, że ten nukleotyd nie pasuje do genomu referencyjnego (mutacja)

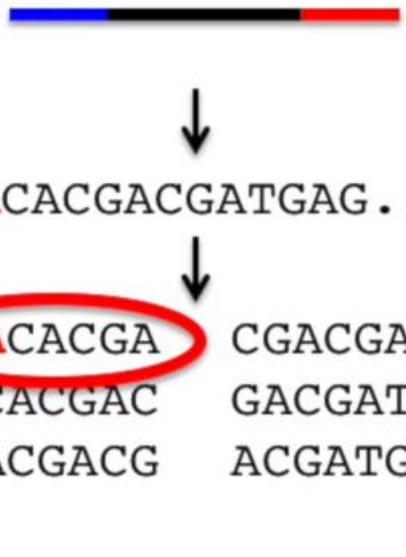
Genom



gattacataccagga...

gattac attaca ttacat
tacata acatac catacc
atacca taccag accagg
ccagga cagga...

Odczyt sekwencji

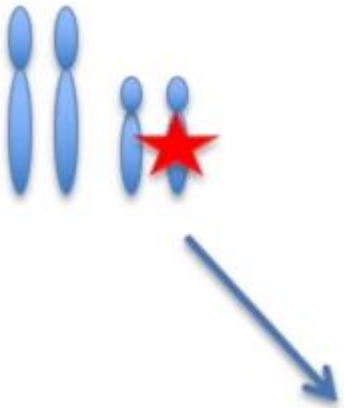


W takim wypadku ten fragment nie zostanie dopasowany do żadnego indexu, ale pozostałe fragmenty odczytu zostaną dopasowane i będziemy w stanie wywnioskować z jakiego fragmentu genomu pochodzi odczyt

Zliczenie odczytów

- Jeżeli znamy pozycję odczytu w genomie możemy go przyporządkować do konkretnego genu

- Zliczamy odczyty uzyskane dla każdego genu



Out[5]:

	#Replicon Name	Replicon Accession	Start	Stop	Strand	GeneID	Locus	Locus tag	Protein product	Length	Protein name
0	Un	NW_018745756.1	1541410	1552853	-	110653968	LOC110653968	-	XP_021665476.1	1717	cellulose synthase A catalytic subunit 1
1	Un	NW_018745803.1	845147	864233	-	110656796	LOC110656796	-	XP_021669453.1	1082	cellulose synthase A catalytic subunit 3
2	Un	NW_018745803.1	845616	852163	-	110656797	LOC110656797	-	XP_021669455.1	1082	cellulose synthase A catalytic subunit 3
3	Un	NW_018745803.1	845616	852163	-	110656797	LOC110656797	-	XP_021669456.1	1082	cellulose synthase A catalytic subunit 3
4	Un	NW_018745803.1	845616	851763	-	110656797	LOC110656797	-	XP_021669454.1	1081	cellulose synthase A catalytic subunit 3
5	Un	NW_018745803.1	858009	864233	-	110656796	LOC110656796	-	XP_021669452.1	1082	cellulose synthase A catalytic subunit 3

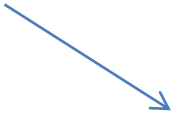
Geneid	bam/flower.bam	bam/stem.bam	bam/leaf.bam
LOC109343272	41	189	410
LOC109343320	9	19	14
LOC109343262	0	0	0
LOC109343339	0	0	0
LOC109343296	4	0	0
LOC109343328	94	2	0
LOC109343288	35	5	0
LOC109343349	0	0	0
LOC109343304	0	2	0
LOC109343312	4	0	0
LOC109343390	1449	186	17

Jeżeli zliczymy odczyty otrzymamy taką tabelę

Geneid	bam/flower.bam	bam/stem.bam	bam/leaf.bam
LOC109343272	41	189	410
LOC109343320	9	19	14
LOC109343262	0	0	0
LOC109343339	0	0	0
LOC109343296	4	0	0
LOC109343328	94	2	0
LOC109343288	35	5	0
LOC109343349	0	0	0
LOC109343304	0	2	0
LOC109343312	4	0	0
LOC109343390	1449	186	17

Dla organizmu, który posiada 30+ tysięcy genów, tabela będzie miała 30+ tysięcy wierszy. Organizmy wyższe (ssaki rośliny) mają 20-60 tyś. genów (roślinny mają przeważnie więcej genów od ssaków)

Geneid	bam/flower.bam	bam/stem.bam	bam/leaf.bam
LOC109343272	41	189	410
LOC109343320	9	19	14
LOC109343262	0	0	0
LOC109343339	0	0	0
LOC109343296	4	0	0
LOC109343328	94	2	0
LOC109343288	35	5	0
LOC109343349	0	0	0
LOC109343304	0	2	0
LOC109343312	4	0	0
LOC109343390	1449	186	17



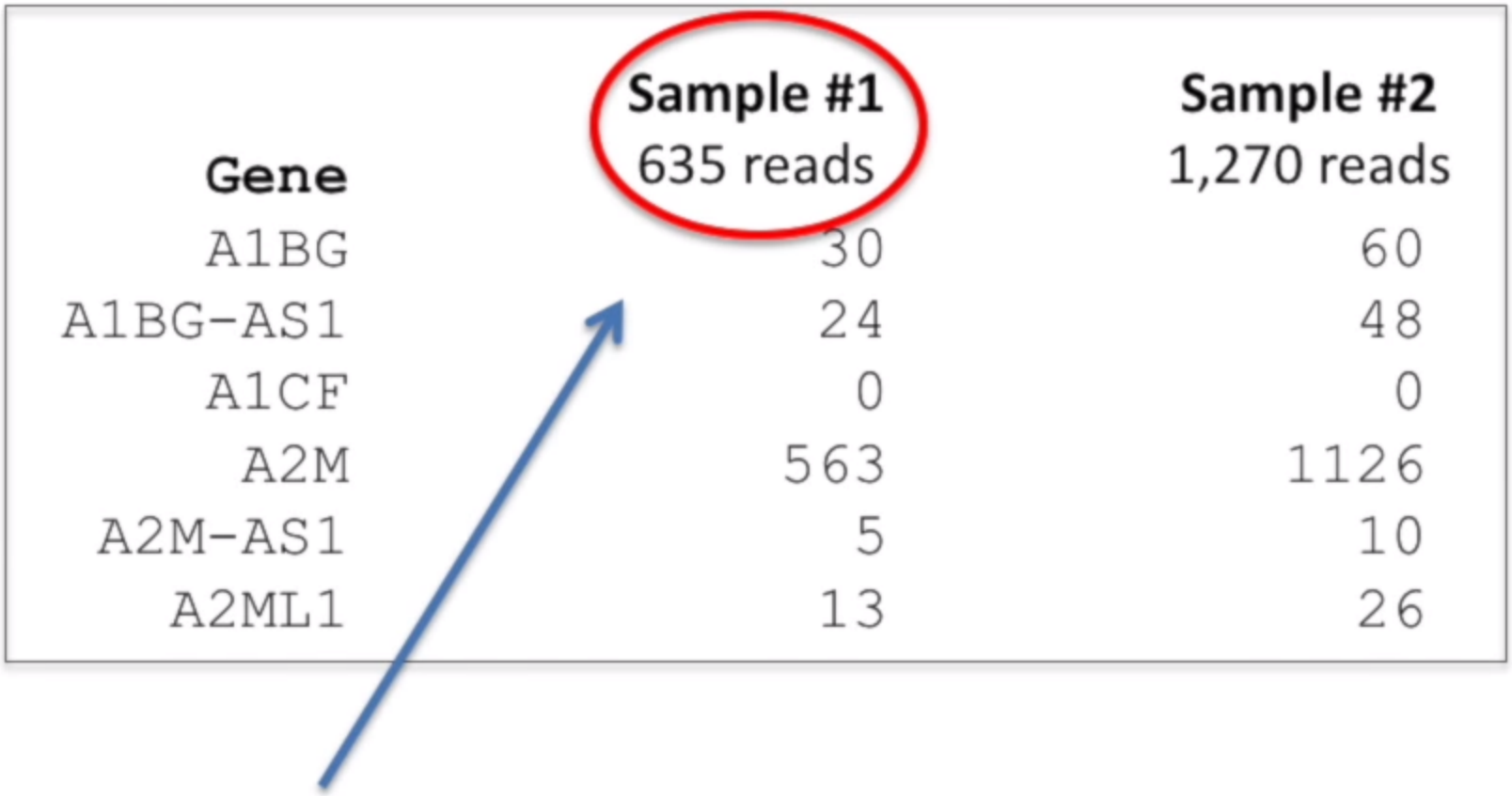
Pozostałe kolumny zawierają liczby odczytów dla próbek, które były sekwencjonowane

Geneid	bam/flower.bam	bam/stem.bam	bam/leaf.bam
LOC109343272	41	189	410
LOC109343320	9	19	14
LOC109343262	0	0	0
LOC109343339	0	0	0
LOC109343296	4	0	0
LOC109343328	94	2	0
LOC109343288	35	5	0
LOC109343349	0	0	0
LOC109343304	0	2	0
LOC109343312	4	0	0
LOC109343390	1449	186	17

Geneid	bam/flower.bam	bam/stem.bam	bam/leaf.bam
LOC109343272	41	189	410
LOC109343320	9	19	14
LOC109343262	0	0	0
LOC109343339	0	0	0
LOC109343296	4	0	0
LOC109343328	94	2	0
LOC109343288	35	5	0
LOC109343349	0	0	0
LOC109343304	0	2	0
LOC109343312	4	0	0
LOC109343390	1449	186	17

Ostatnią rzeczą do wykonania przed analizą statystyczną jest normalizacja danych

Normalizacja musi zostać wykonana ponieważ każda próbka ma inną liczbę odczytów. Może to być związane z ilością dorzuconych danych w czasie filtrowania (niektóre próbki mogą posiadać więcej odczytów z niskim QS) oraz ilością fragmentów DNA zligowanych do „flow cell”. Niektóre próbki mogą zawierać więcej fragmentów od innych próbek.



Gene	Sample #1 635 reads	Sample #2 1,270 reads
A1BG	30	60
A1BG-AS1	24	48
A1CF	0	0
A2M	563	1126
A2M-AS1	5	10
A2ML1	13	26

Próbka 1 ma 635 odczytów

Gene	Sample #1 635 reads	Sample #2 1,270 reads
A1BG	30	60
A1BG-AS1	24	48
A1CF	0	0
A2M	563	1126
A2M-AS1	5	10
A2ML1	13	26

Próbka 2 ma 1270 odczytów,
2x więcej od próbki 1

Gene	Sample #1 635 reads	Sample #2 1,270 reads
A1BG	30	60
A1BG-AS1	24	48
A1CF	0	0
A2M	563	1126
A2M-AS1	5	10
A2ML1	13	26

Nie oznacza to, że wszystkie geny w próbce 2 miały 2x wyższą ekspresję niż w próbce. Oznacza to natomiast, że odczyty z próbki 2 miały wyższy QS i większa ich liczba przeszła wstępne filtrowanie danych oraz/lub że na „flow cell” próbki 2 znajdowało się więcej fragmentów DNA niż na „flow cell” próbki 1

Gene	Sample #1 635 reads	Sample #2 1,270 reads
A1BG	30	60
A1BG-AS1	24	48
A1CF	0	0
A2M	563	1126
A2M-AS1	5	10
A2ML1	13	26

Gene	Sample #1 635 reads	Sample #2 1,270 reads
A1BG	30	60
A1BG-AS1	24	48
A1CF	0	0
A2M	563	1126
A2M-AS1	5	10
A2ML1	13	26

W związku z tym musimy musimy powiązać ilość odczytów uzyskaną dla każdego genu z ogólną liczbą odczytów uzyskanych dla danej próbki (biblioteki)

Najprostszym rozwiązaniem byłoby podzielenie liczby oczytów dla każdego genu przez łączną liczbę odczytów dla próbki (biblioteki)

Takie podejście ma jednak wiele wad i w praktyce stosuje się bardziej wyrafinowane metody normalizacji bibliotek

Metody normalizacji

- RPKM (Reads per Kilobase Milion)
- FPKM (Fragments per Kilobase Milion)
- TPM (Transcripts per Milion)

Metody normalizacji

- Stosowane do normalizacji (zniwelowania różnic w):
  - „Głębokości sekwencjonowania – ogólna liczba odczytów (część Milion)
    - Bardziej głębokie sekwencjonowanie (więcej fragmentów na „flow cell”) będzie generowało więcej odczytów
  - Długości poszczególnych genów (część Kilobase)
    - Dłuższe geny będą miały więcej odczytów

Gen	Pow 1	Pow 2	Pow 3
A (2 kpz)	10	12	30
B (4kpz)	20	25	60
C (1 kpz)	5	8	15
D (10kpz)	0	0	1

Gen	Pow 1	Pow 2	Pow 3
A (2 kpz)	10	12	30
B (4kpz)	20	25	60
C (1 kpz)	5	8	15
D (10kpz)	0	0	1

Gen	Pow 1	Pow 2	Pow 3
A (2 kpz)	10	12	30
B (4kpz)	20	25	60
C (1 kpz)	5	8	15
D (10kpz)	0	0	1

Normalizacja RPKM – krok 1 normalizacja względem wielkości biblioteki

Gen	Pow 1	Pow 2	Pow 3
A (2 kpz)	10	12	30
B (4kpz)	20	25	60
C (1 kpz)	5	8	15
D (10kpz)	0	0	1
suma	35	45	106

Normalizacja RPKM – krok 1 normalizacja względem wielkości biblioteki

Gen	Pow 1	Pow 2	Pow 3
A (2 kpz)	10	12	30
B (4kpz)	20	25	60
C (1 kpz)	5	8	15
D (10kpz)	0	0	1
suma	35	45	106
Dziesiątki odczytów	3,5	4,5	10,6

Dla celów tego przykładu policzymy dziesiątki odczytów, ale w jednostce RPKM liczymy miliony odczytów

Dla jednostki RPKM został wybrany przelicznik milion, aby liczby „dobrze wyglądały”

Normalizacja RPKM – krok 1 normalizacja względem wielkości biblioteki

Gen	Pow 1	Pow 2	Pow 3
A (2 kpz)	10	12	30
B (4kpz)	20	25	60
C (1 kpz)	5	8	15
D (10kpz)	0	0	1
suma	35	45	106
Dziesiątki odczytów	3,5	4,5	10,6

Otrzymaliśmy „per Milion” przelicznik dla każdego powtórzenia

Normalizacja RPKM – krok 1 normalizacja względem wielkości biblioteki

Gen	Pow 1	Pow 2	Pow 3
A (2 kpz)	10	12	30
B (4kpz)	20	25	60
C (1 kpz)	5	8	15
D (10kpz)	0	0	1
suma	35	45	106
Dziesiątki odczytów	3,5	4,5	10,6

Gen	Pow 1 RPM	Pow 2 RPM	Pow 3 RPM
A (2 kpz)	2,86	2,67	2,83
B (4kpz)	5,71	5,56	5,66
C (1 kpz)	1,43	1,78	1,42
D (10kpz)	0,00	0,00	0,09

Normalizacja RPKM – krok 2 normalizacja względem długości genu

Gen	Pow 1 RPM	Pow 2 RPM	Pow 3 RPM
A (2 kpz)	2,86	2,67	2,83
B (4kpz)	5,71	5,56	5,66
C (1 kpz)	1,43	1,78	1,42
D (10kpz)	0,00	0,00	0,09

Gen	Pow 1 RPKM	Pow 2 RPKM	Pow 3 RPKM
A (2 kpz)	1,43	1,33	1,42
B (4kpz)	1,43	1,39	1,42
C (1 kpz)	1,43	1,78	1,42
D (10kpz)	0,00	0,00	0,01

Podsumowanie - RPKM

Przed

Gen	Pow 1	Pow 2	Pow 3
A (2 kpz)	10	12	30
B (4kpz)	20	25	60
C (1 kpz)	5	8	15
D (10kpz)	0	0	1

Normalizacja:

1) Względem wielkości biblioteki
2) Względem długości genu

Po

Gen	Pow 1 RPKM	Pow 2 RPKM	Pow 3 RPKM
A (2 kpz)	1,43	1,33	1,42
B (4kpz)	1,43	1,39	1,42
C (1 kpz)	1,43	1,78	1,42
D (10kpz)	0,00	0,00	0,01

RPKM i FPKM – dwie podobne jednostki

- RPKM – stosowany dla sekwencjonowania „single end RNAseq”
  - Jeden odczyt odpowiada jednemu fragmentowi DNA
- FPKM – stosowany dla sekwencjonowania „pair end RNAseq”
  - Dwa odczyty dla jednego fragmentu DNA (ale nie zawsze czasem dla jednego fragmentu mamy 1 odczyt – problem z QS)
  - FPKM bierze pod uwagę fragmentu i zapobiega podwójnemu zliczeniu odczytów pochodzących z tego samego fragmentu

TPM (Transcript per Milion)

Normalizacja TPM – krok 1
normalizacja względem długości genu

Gen	Pow 1	Pow 2	Pow 3
A (2 kpz)	10	12	30
B (4kpz)	20	25	60
C (1 kpz)	5	8	15
D (10kpz)	0	0	1

Gen	Pow 1 RPK	Pow 2 RPK	Pow 3 RPK
A (2 kpz)	5	6	15
B (4kpz)	5	6,25	15
C (1 kpz)	5	8	15
D (10kpz)	0	0	0,1

Normalizacja TPM – krok 2 normalizacja względem wielkości biblioteki

Gen	Pow 1 RPK	Pow 2 RPK	Pow 3 RPK
A (2 kpz)	5	6	15
B (4kpz)	5	6,25	15
C (1 kpz)	5	8	15
D (10kpz)	0	0	0,1
Suma RPK	15	20,25	45,1
Dziesiątki RPK	1,5	2,025	4,51

Standardowo dzielimy przez milion, ale w narzym przykładzie ponownie podzielimy przez 10

Normalizacja TPM – krok 2 normalizacja względem wielkości biblioteki

Gen	Pow 1 RPK	Pow 2 RPK	Pow 3 RPK
A (2 kpz)	5	6	15
B (4kpz)	5	6,25	15
C (1 kpz)	5	8	15
D (10kpz)	0	0	0,1
Suma RPK	15	20,25	45,1
Dziesiątki RPK	1,5	2,025	4,51

Gen	Pow 1 TPM	Pow 2 TPM	Pow 3 TPM
A (2 kpz)	3,33	2,96	3,33
B (4kpz)	3,33	3,09	3,33
C (1 kpz)	3,33	3,95	3,33
D (10kpz)	0,00	0,00	0,02

RPKM vs TPM

RPKM

Gen	Pow 1 RPKM	Pow 2 RPKM	Pow 3 RPKM
A (2 kpz)	1,43	1,33	1,42
B (4kpz)	1,43	1,39	1,42
C (1 kpz)	1,43	1,78	1,42
D (10kpz)	0,00	0,00	0,01

TPM

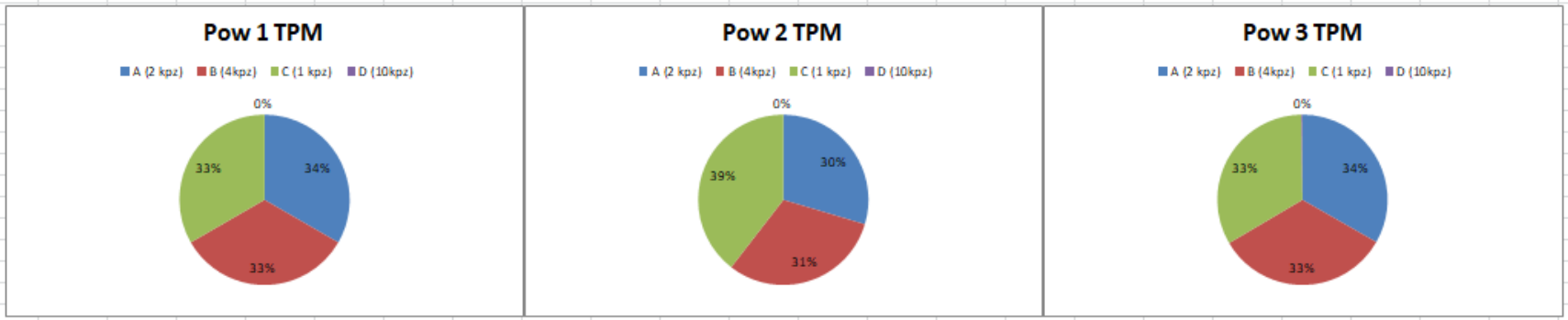
Gen	Pow 1 TPM	Pow 2 TPM	Pow 3 TPM
A (2 kpz)	3,33	2,96	3,33
B (4kpz)	3,33	3,09	3,33
C (1 kpz)	3,33	3,95	3,33
D (10kpz)	0,00	0,00	0,02

RPKM vs TPM

Gen	Pow 1 RPKM	Pow 2 RPKM	Pow 3 RPKM
A (2 kpz)	1,43	1,33	1,42
B (4kpz)	1,43	1,39	1,42
C (1 kpz)	1,43	1,78	1,42
D (10kpz)	0,00	0,00	0,01
suma	4,29	4,50	4,25

Gen	Pow 1 TPM	Pow 2 TPM	Pow 3 TPM
A (2 kpz)	3,33	2,96	3,33
B (4kpz)	3,33	3,09	3,33
C (1 kpz)	3,33	3,95	3,33
D (10kpz)	0,00	0,00	0,02
suma	10,00	10,00	10,00

RPKM vs TPM



Gen	Pow 1 TPM	Pow 2 TPM	Pow 3 TPM
A (2 kpz)	3,333	2,96	3,326
B (4kpz)	3,33	3,09	3,33
C (1 kpz)	3,33	3,95	3,33
D (10kpz)	0,00	0,00	0,02
suma	10,00	10,00	10,00

RPKM vs TPM

Gen	Pow 1 RPKM	Pow 2 RPKM	Pow 3 RPKM
A (2 kpz)	1,43	1,33	1,42
B (4kpz)	1,43	1,39	1,42
C (1 kpz)	1,43	1,78	1,42
D (10kpz)	0,00	0,00	0,01
suma	4,29	4,50	4,25

W przypadku RPKM trudniej jest określić proporcję wszystkich odczytów ponieważ dla każdej próbki suma odczytów jest inna 1,43 określa inną proporcję w każdej próbce (bo suma odczytów w każdej próbce jest inna)

TPM

- TPS jest wykorzystywane ponieważ pozwala na łatwe określenie jaka proporcja wszystkich odczytów mapuje się do konkretnego genu (w każdej próbce)
- Ponieważ RNAseq opiera się na porównaniu względnej poporcji odczytów, ta miara wydaje się być najbardziej odpowiednia