Analiza danych z wysokoprzepustowego sekwencjonowania

W4: genomy

sekwencjonowanie genomów

Zakończone projekty sekwencjonowania genomów

153 899
31 920
4 032
Genomes

111 840
26 450
3 672
Bacterial

1 509
649
177
Archaeal

40 550
4 821
183
Eukaryal

11.2018
11.2014
11.2012

sekwencjonowanie genomów

Trwające projekty sekwencjonowania genomów

25 559
17 128
15 513
Genomes

13 155	**197**	**12 207**
11 267	**344**	**5 436**
12 668	**252**	**2 593**
Bacterial	Archaeal	Eukaryal

11.2018

11.2014

11.2012

sekwencjonowanie genomów

Sekwencjonowanie nowych organizmów:

- poznanie genomów organizmów modelowych
- identyfikacja nowych mechanizmów adaptacji do środowiska
- śledzenie ewolucji genomów
- inżynieria genetyczna organizmów istotnych gospodarczo i przemysłowo
- bo możemy i nas stać!

Sekwencjonowanie genomów ludzkich:

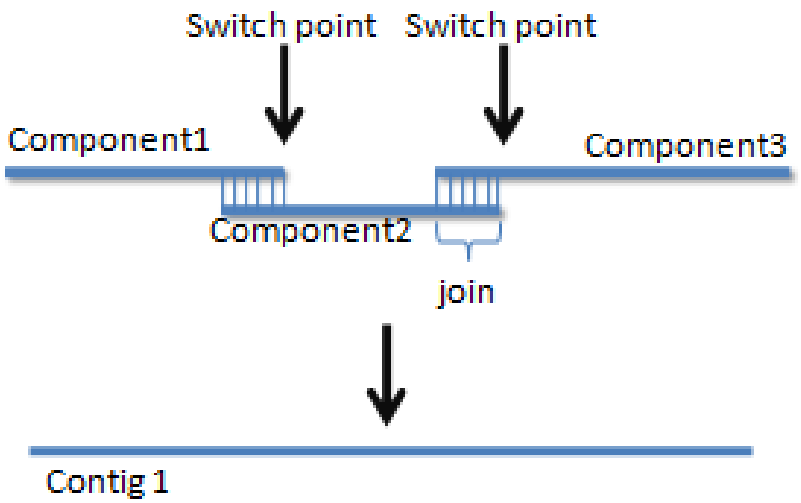
- poznanie różnic genetycznych na poziomie populacyjnym
- poznanie mechanizmów chorób genetycznych wielogenowych i sposobu ich dziedziczenia
- identyfikacja markerów chorobowych
- identyfikacja markerów prognostycznych w terapii

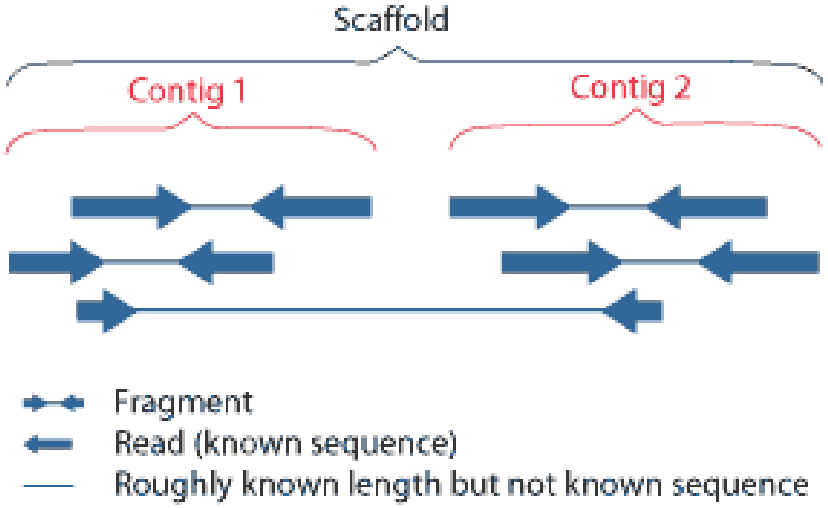
sekwencjonowanie genomów terminologia

BAC – Bacterial Artificial Chromosome – plazmid zdolny do przenoszenia insertów o długości 150-350 kpz

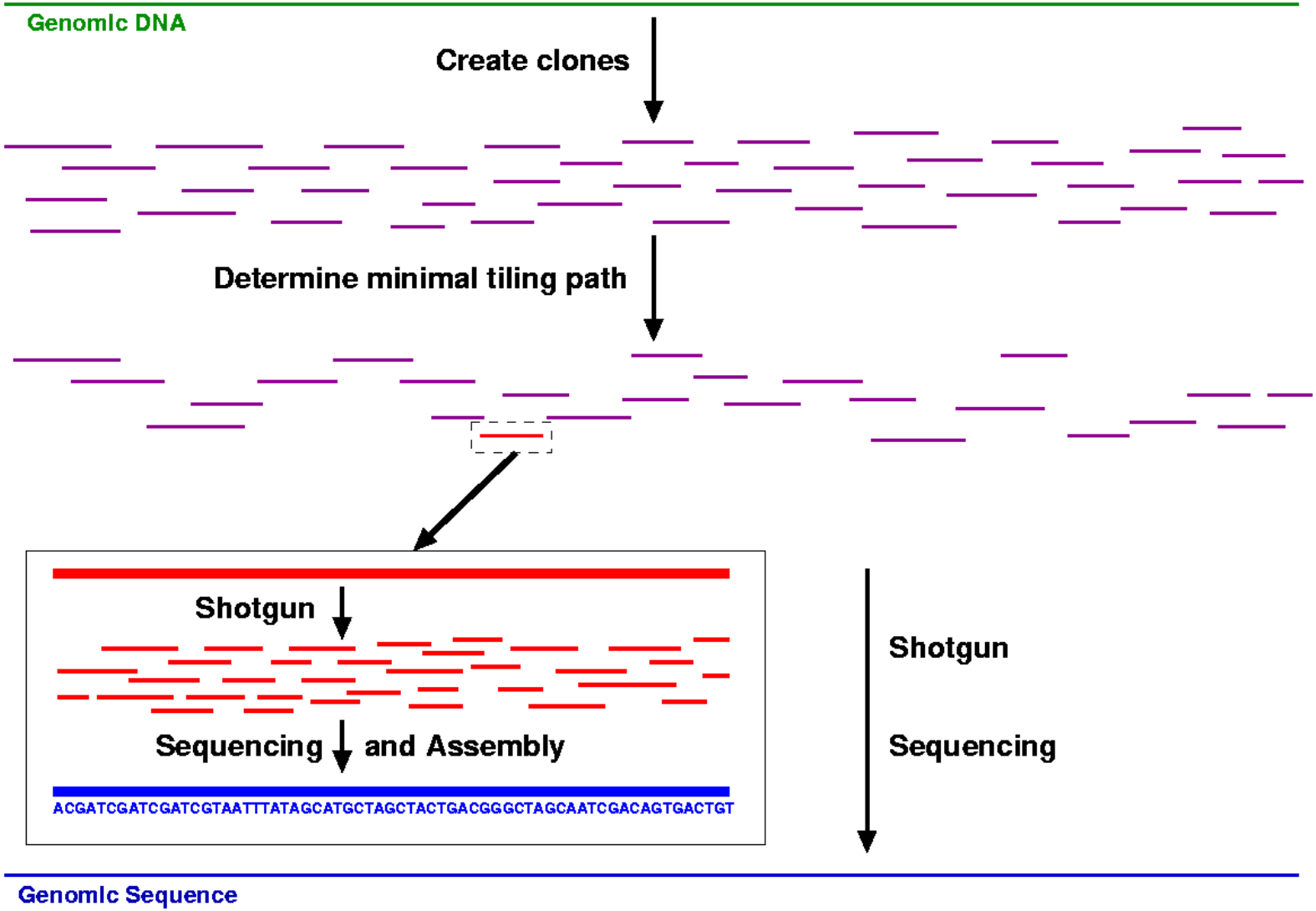
contig – ciągła sekwencje wygenerowana na podstawie nakładających cię na siebie fragmentów sekwencji uzyskanych z sekwencjonowania

scaffold – uporządkowany pod względem kolejności i orientacji zestaw contigów. Zazwyczaj zawiera przerwy w sekwencji.

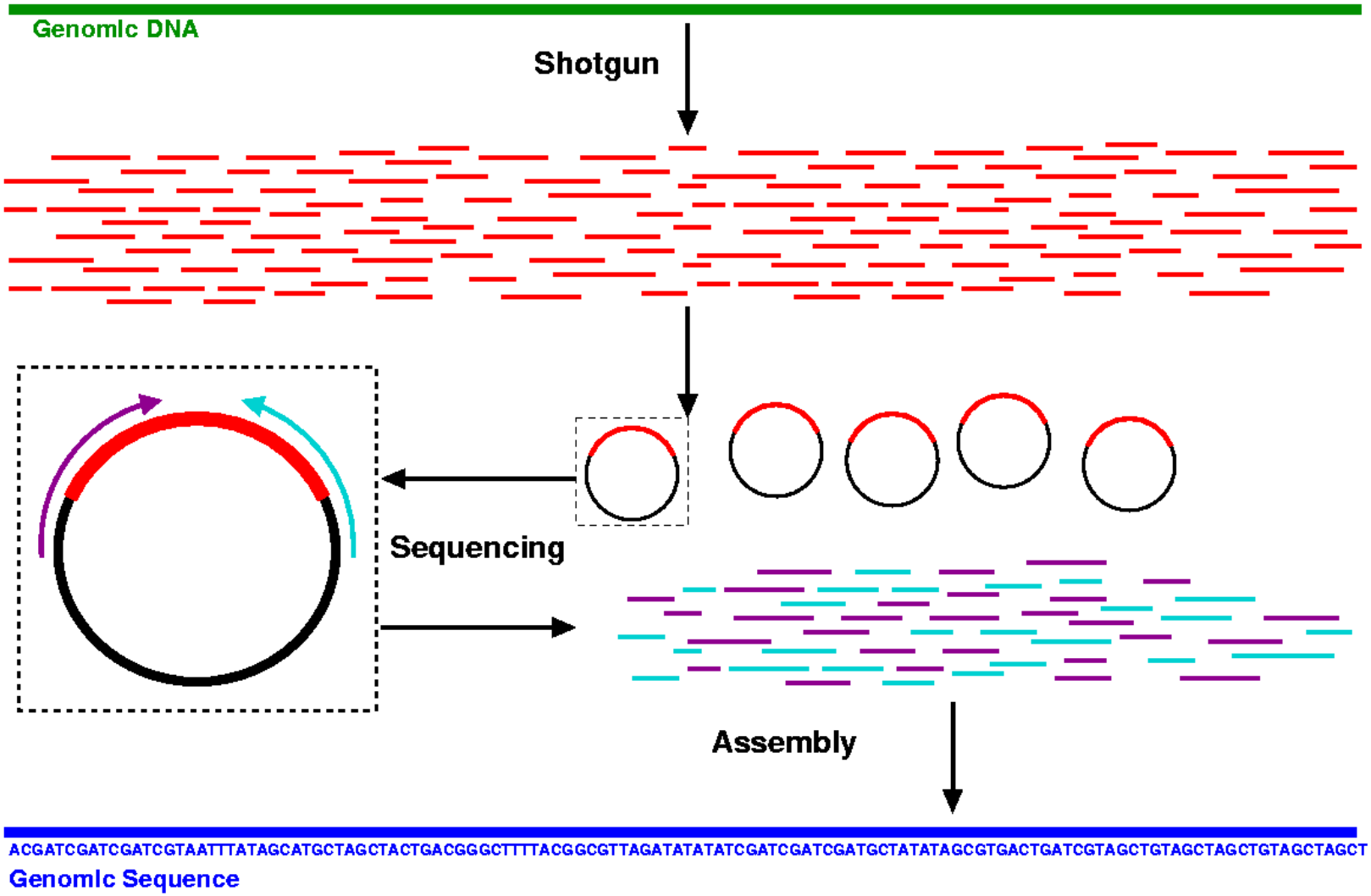




sekwencjonowanie genomów hierarchiczne – Human Genome Project



sekwencjonowanie genomów shotgun – Cellera



sekwencjonowanie genomów wysokoprzepustowe

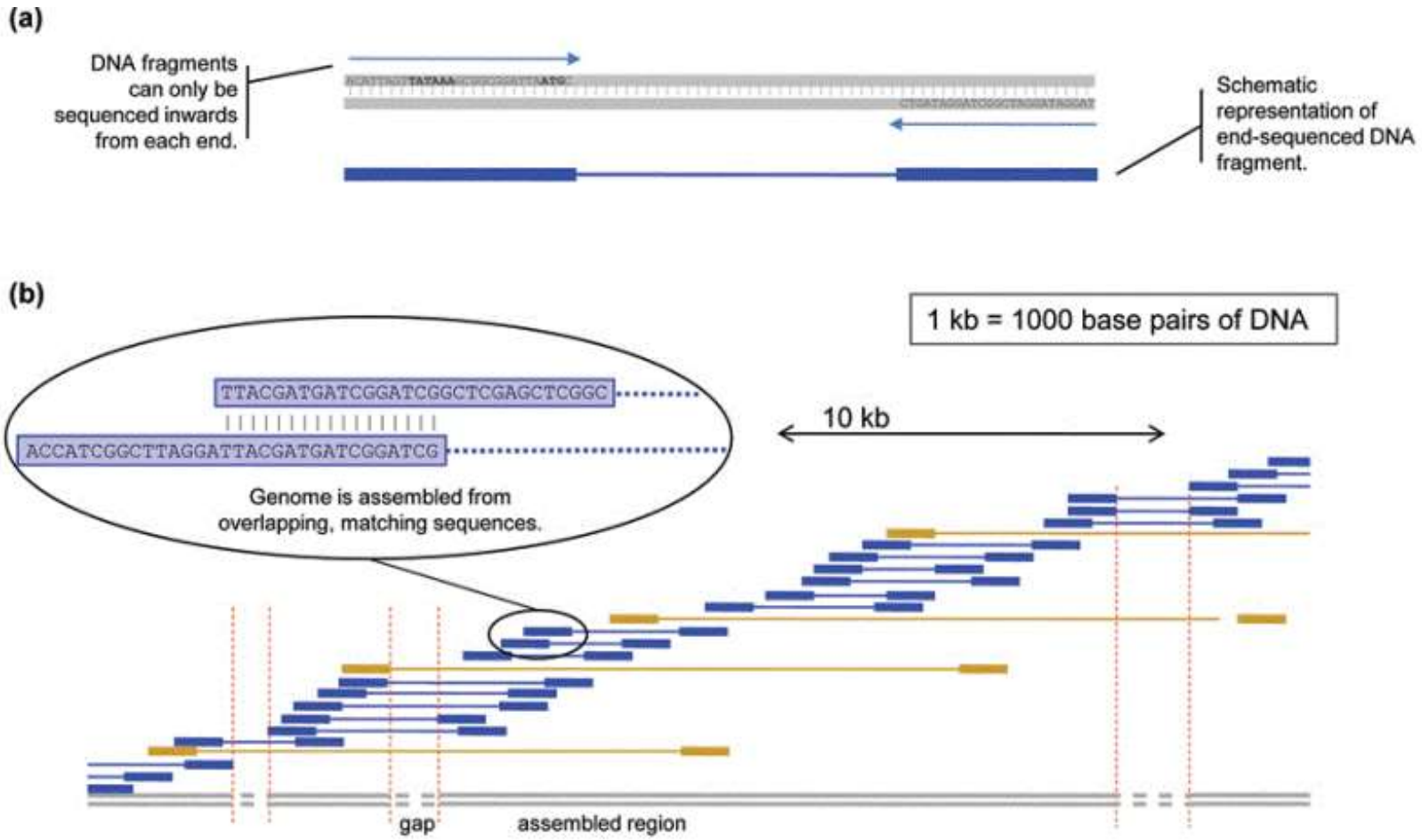
Zalety NGS:

- tanie
- ultra-wysokowydajne
- amplifikacja zamiast klonowania

Wady:

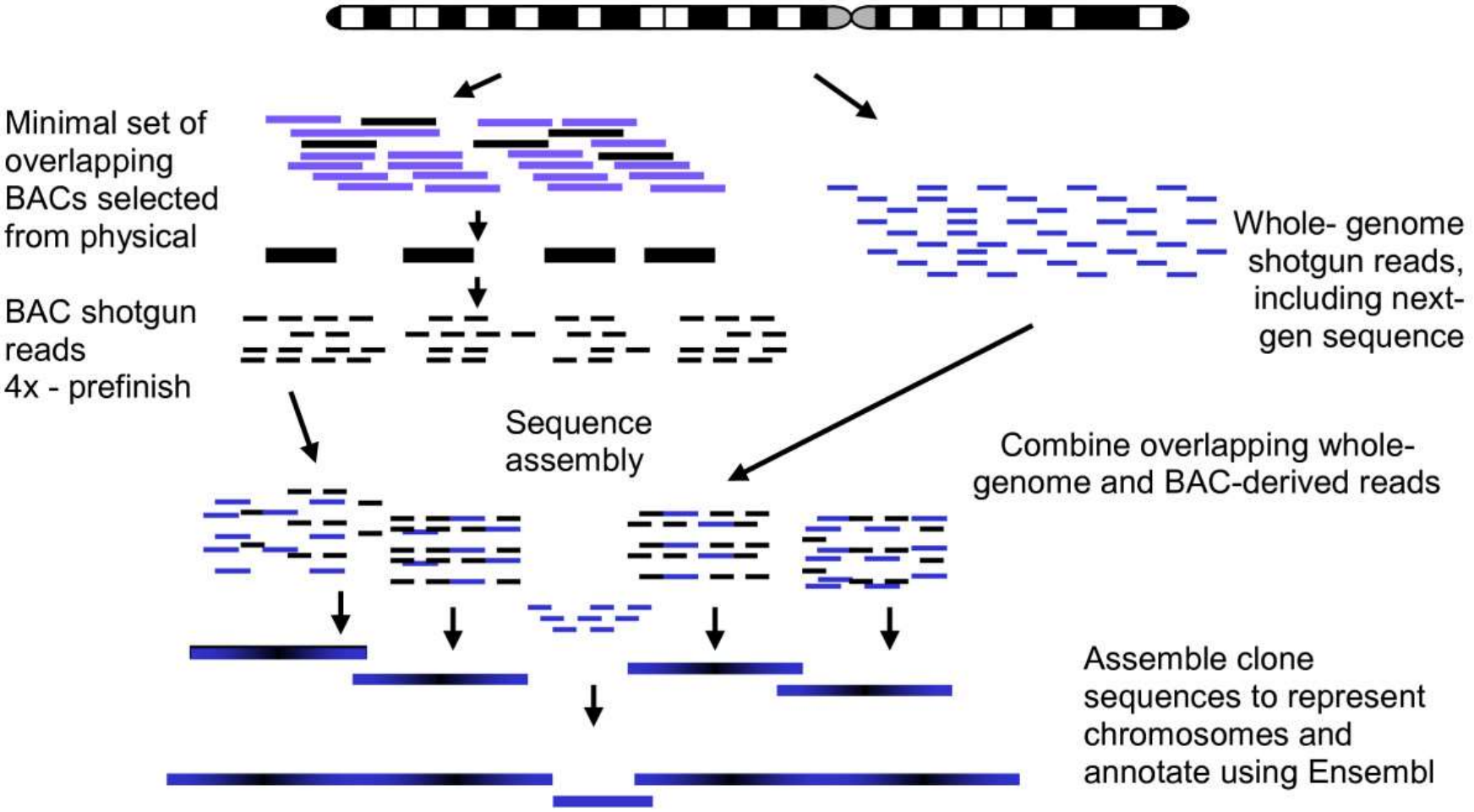
- krótkie odczyty
- ograniczony dystans pomiędzy paired-ends

sekwencjonowanie genomów shotgun (NGS)



Gilchrist M (2010) Mill Hill Essays

sekwencjonowanie genomów podejście hybrydowe



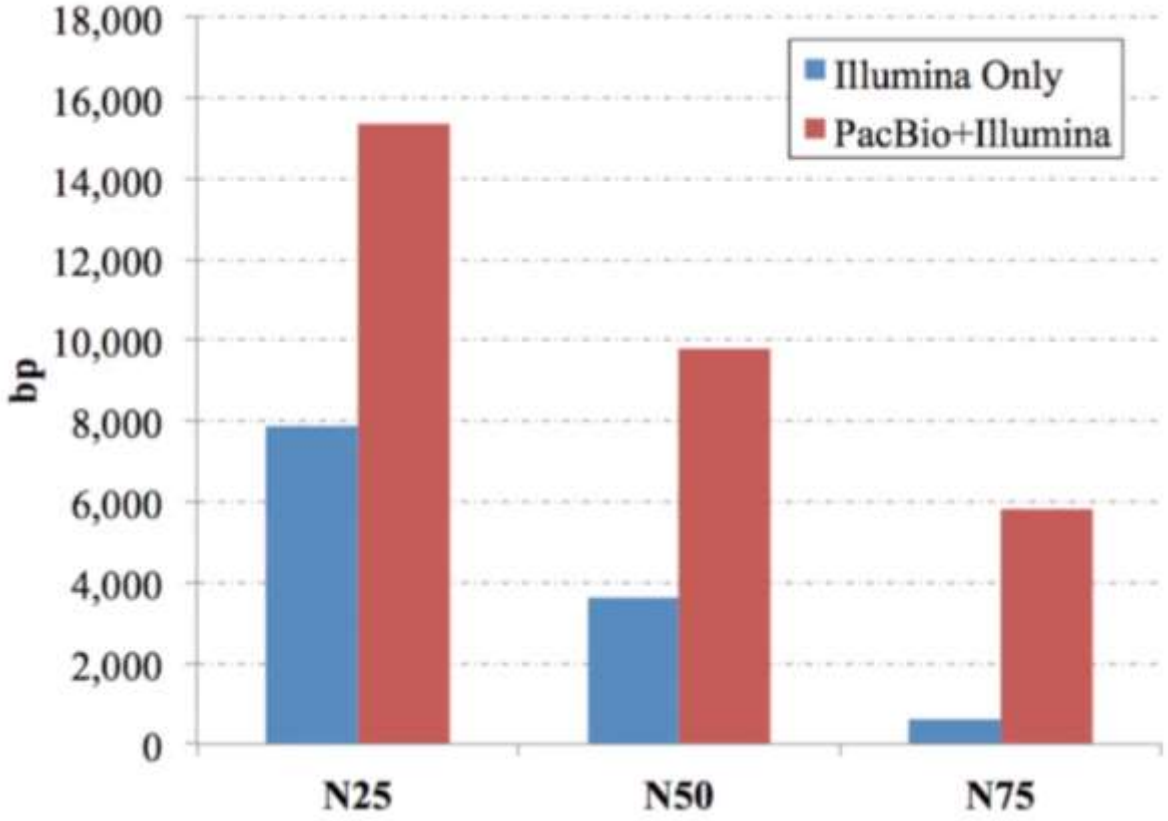
Archibald et al. (2010) BMC Genomics 11:438

sekwencjonowanie genomów alternatywy dla BAC

- Wprowadzenie długich odczytów jako rusztowania:
  - 454
  - PacBio
  - Oxford Nanopore

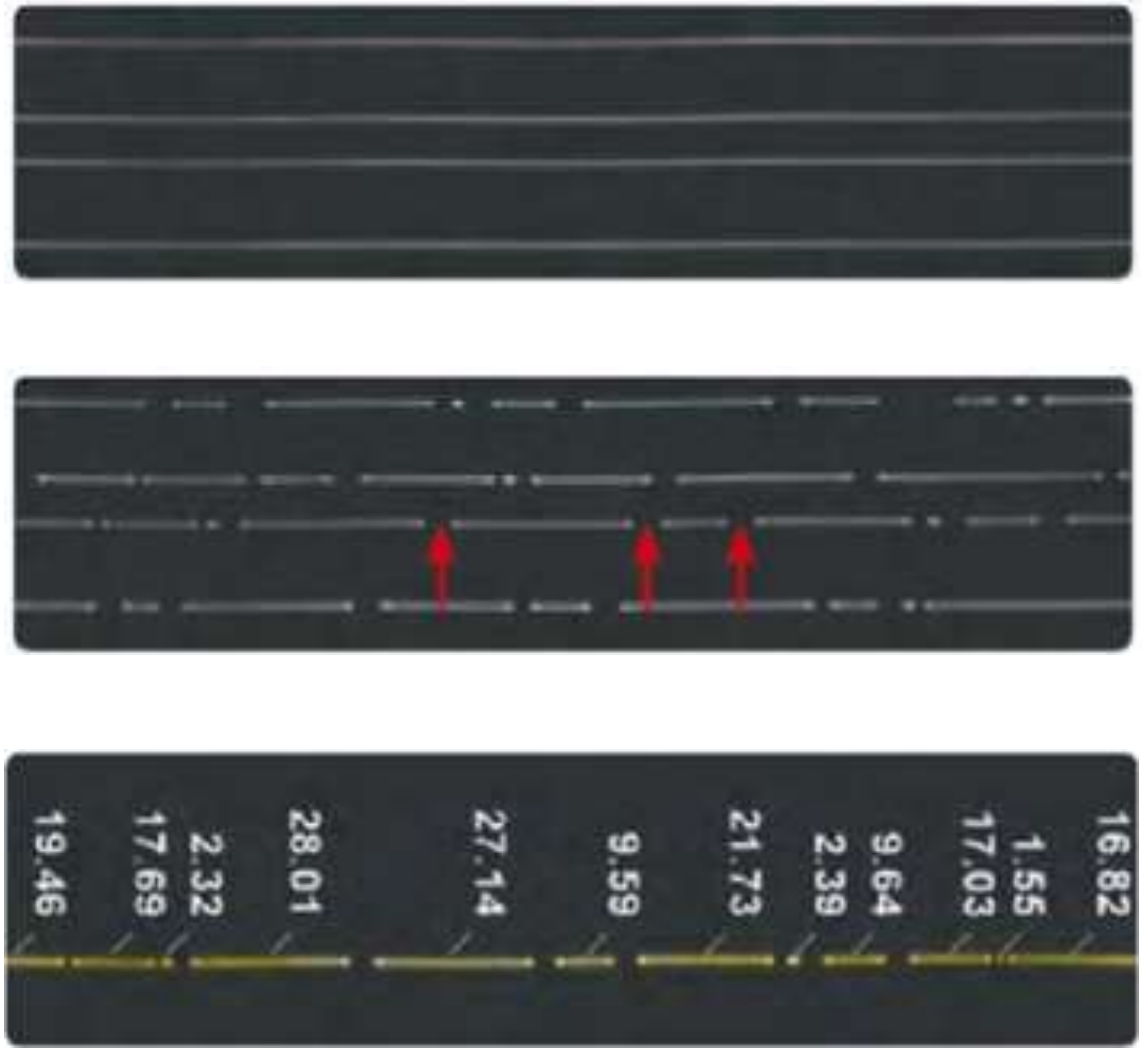
- Mapowanie optyczne genomu

sekwencjonowanie genomów podejście hybrydowe PacBio

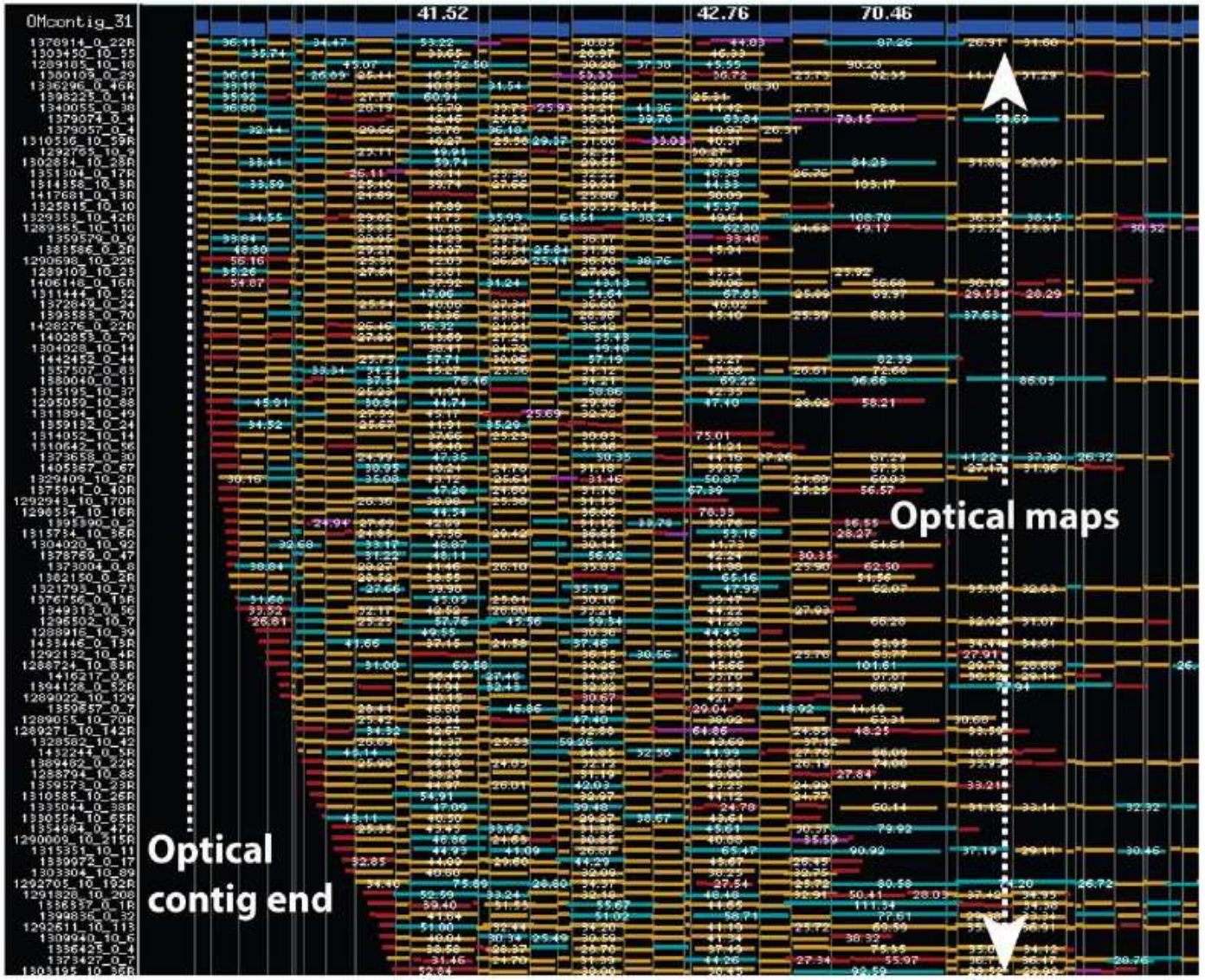


Mason C et al. (2012) http://www.pacificbiosciences.com

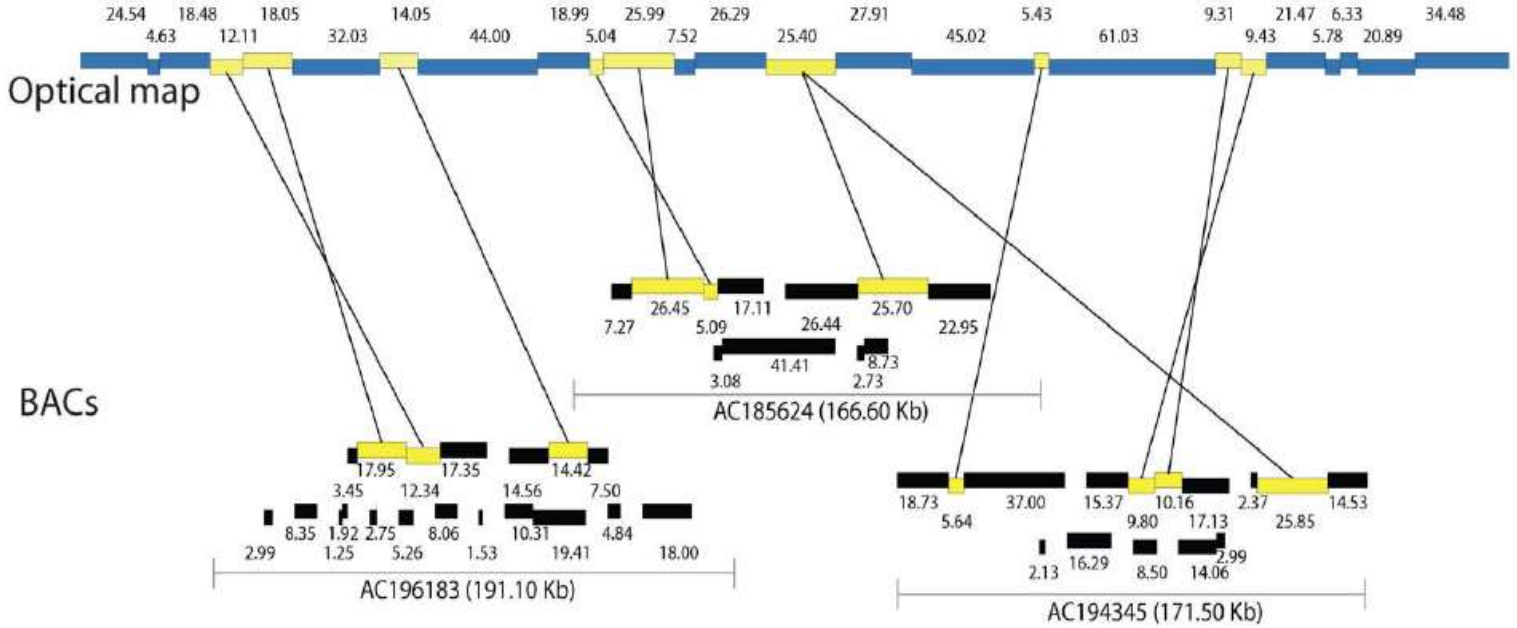
sekwencjonowanie genomów
mapowanie optyczne



sekwencjonowanie genomów
mapowanie optyczne



sekwencjonowanie genomów
mapowanie optyczne



sekwencjonowanie genomów wysokoprzepustowe

shotgun

- genomy bakteryjne
- małe genomy eukariotyczne
- duże genomy eukariotyczne z niską ilością sekwencji powtórzonych

hierarchiczne/hybrydowe/wspomagane

- duże genomy eukariotyczne o znacznej liczbie sekwencji powtórzonych (ssaki, rośliny)

składanie sekwencji *de novo*

proces rekonstrukcji oryginalnego DNA
z użyciem fragmentów sekwencji

strategie składania

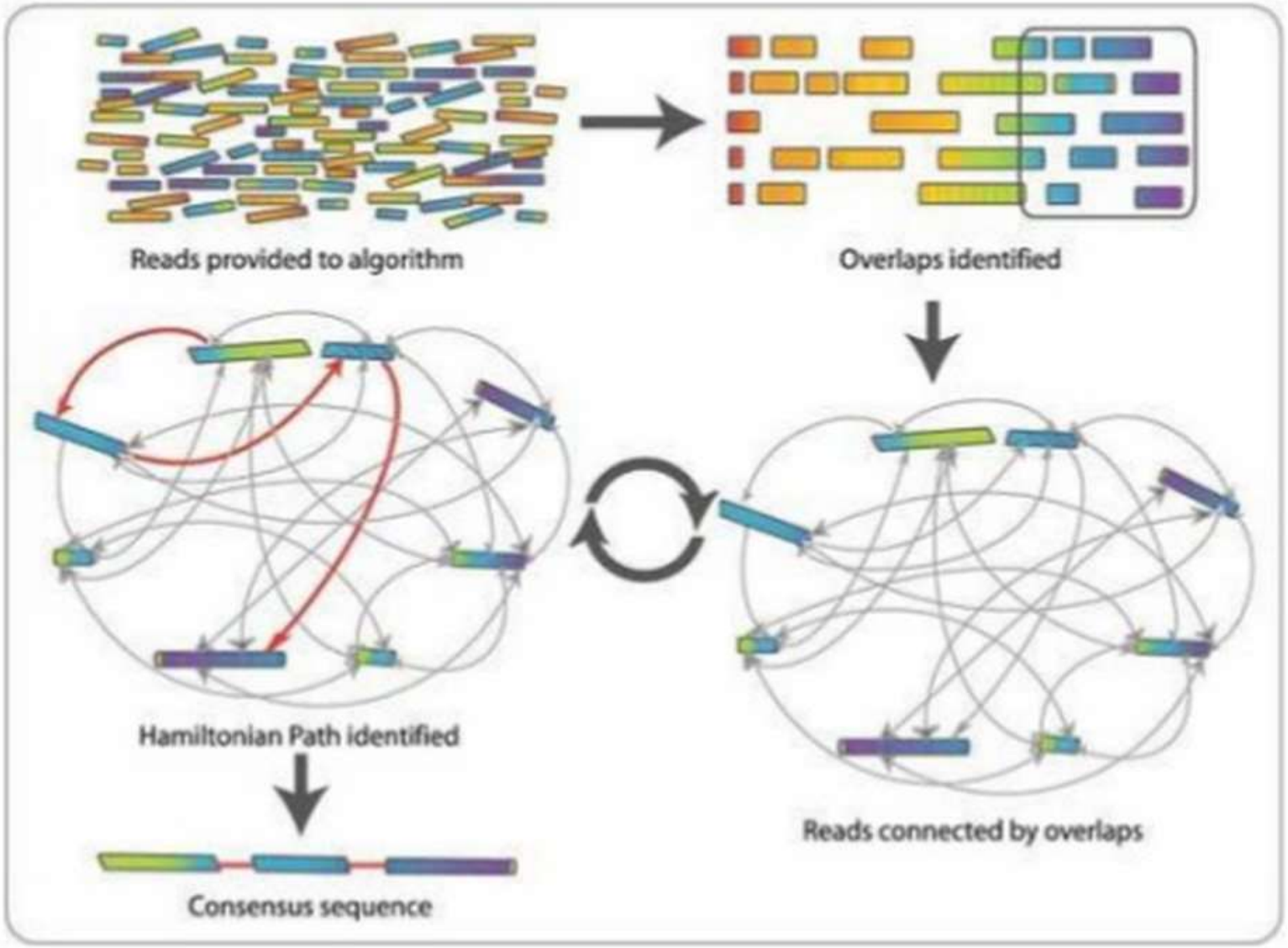
Wszystkie metody robią to samo
Różnią się skrótem który biorą

- składanie zachłanne
- poszukiwanie nałożeń sekwencji
- grafy de Bruijn’a
- grafy ciągów
- „seed and extend”

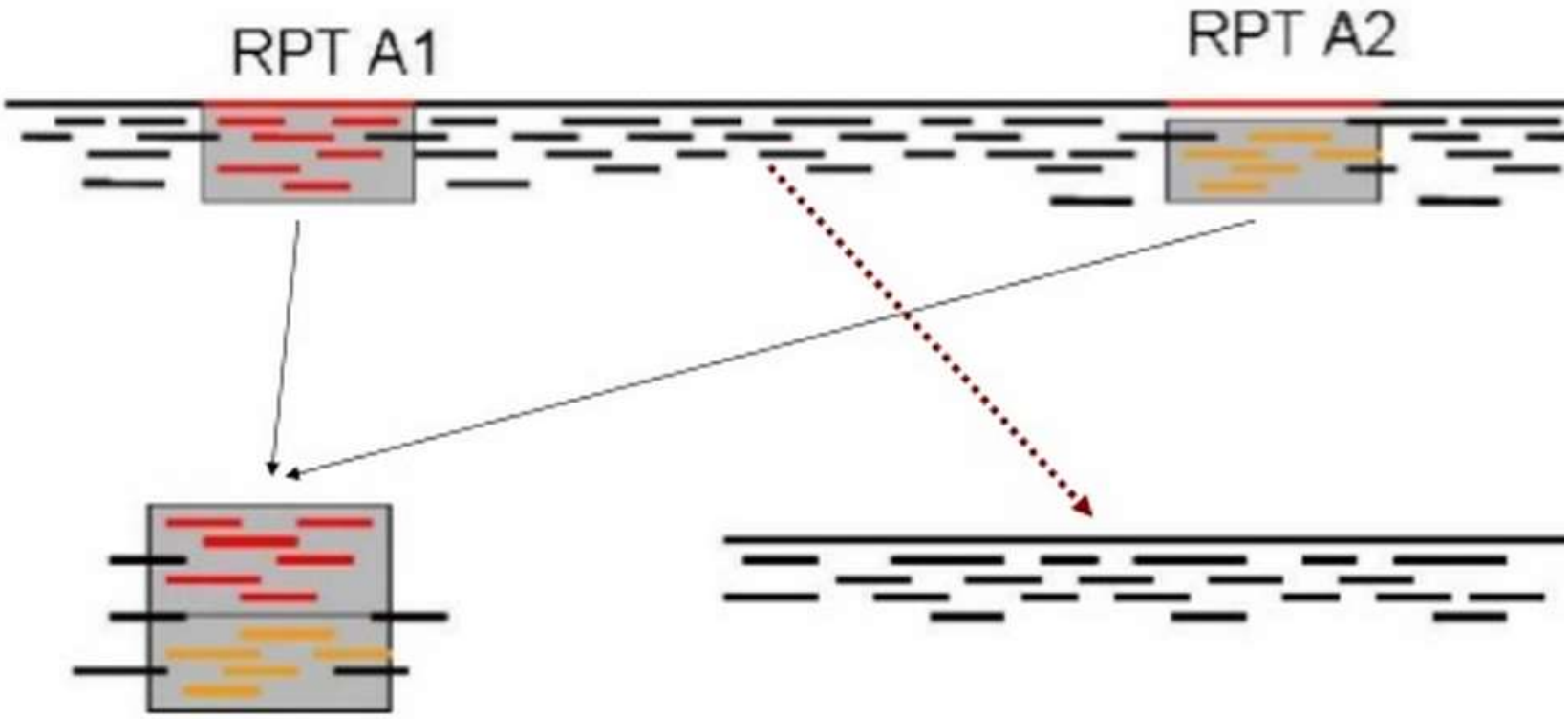
workflow

- Znaleźć wszystkie możliwe nałożenia pomiędzy odczytami
- Zbudować graf
- Uprościć graf
- Przejść graf

workflow

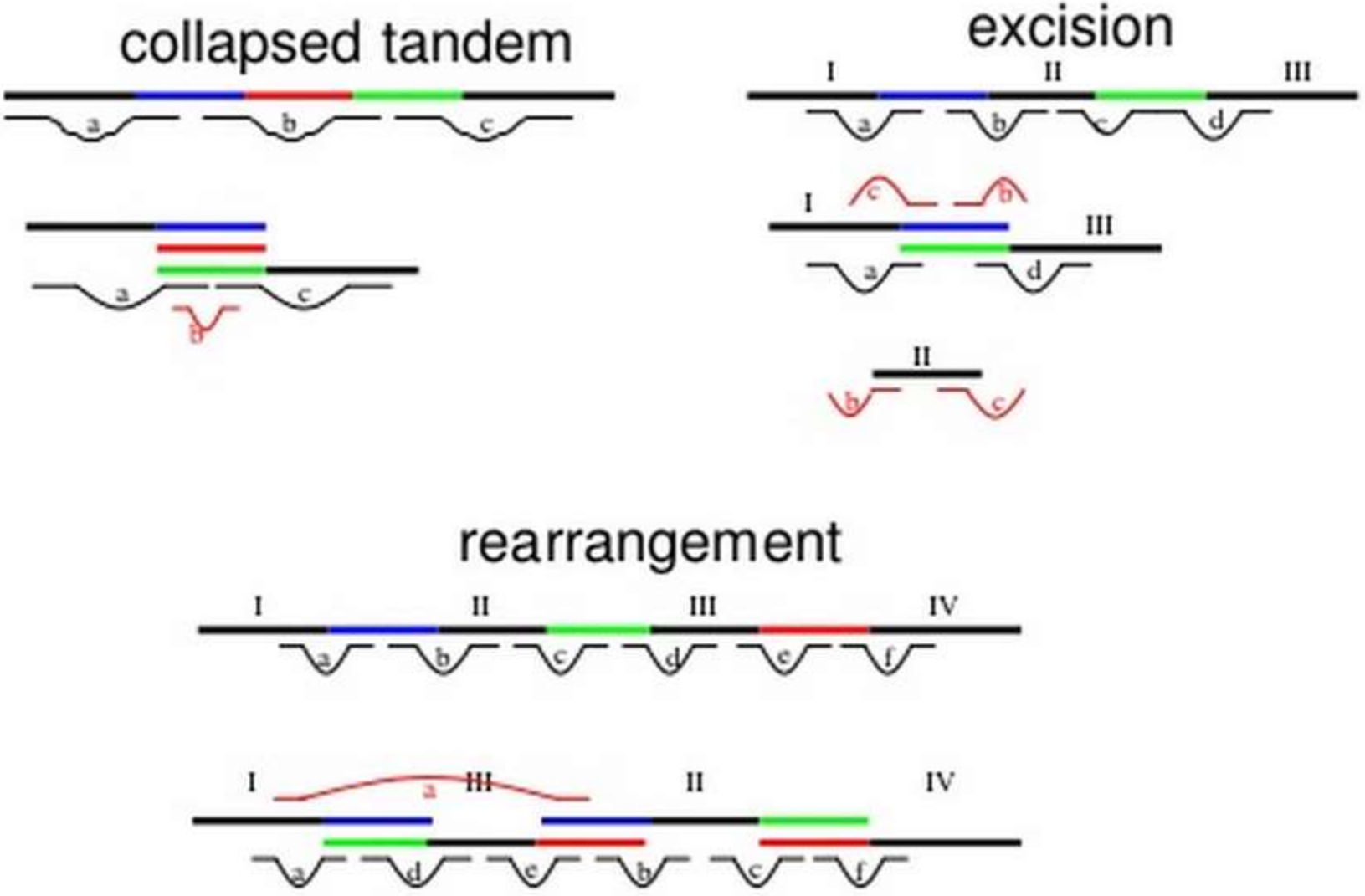

Reads provided to algorithm
Overlaps identified
Hamiltonian Path identified
Reads connected by overlaps
Consensus sequence

powtórzenia



The repeated element is collapsed into a single contig

powtórzenia



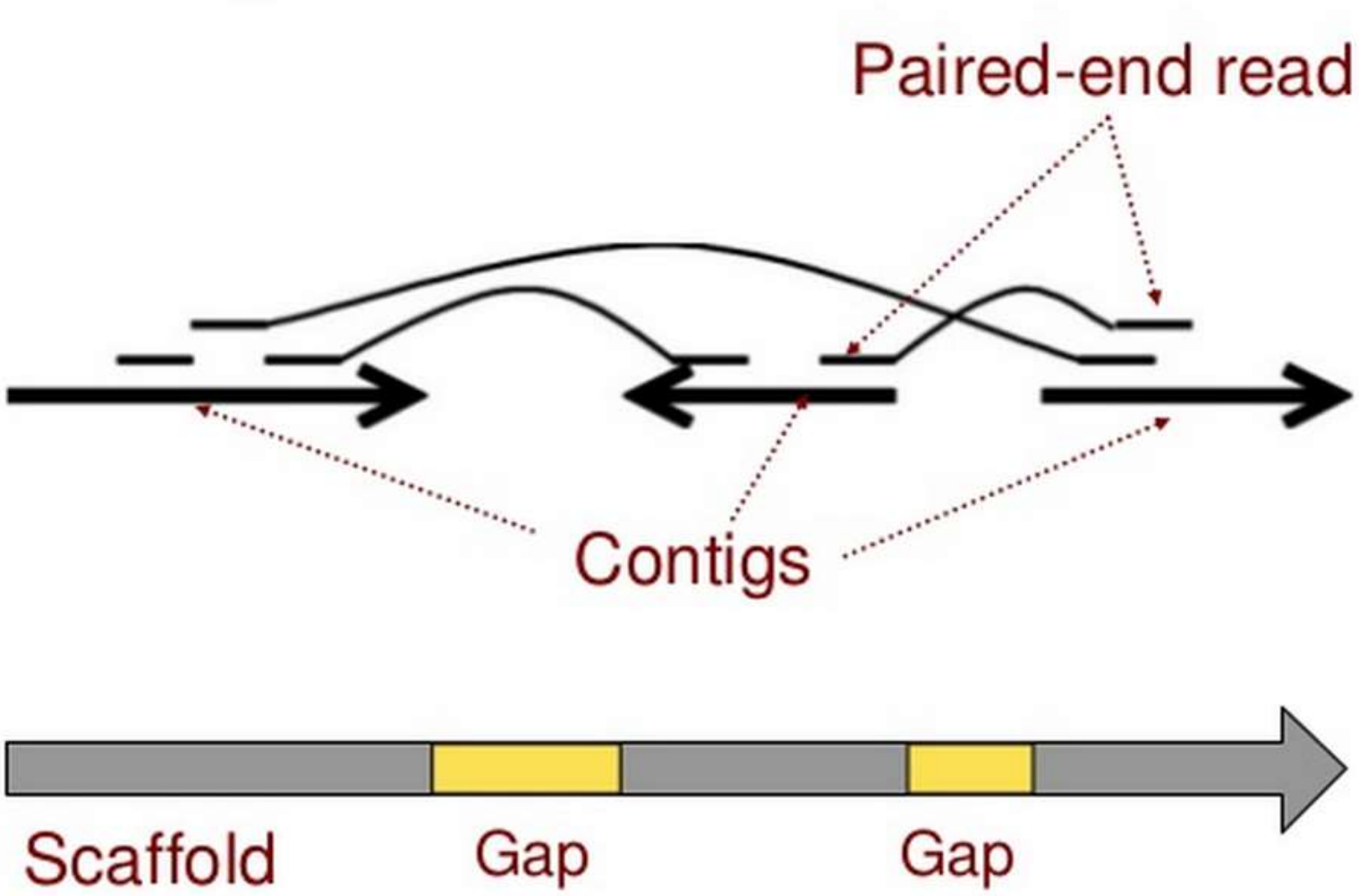
3 prawa powtórzeń

1. Niemożliwe jest rozwiązanie struktury powtórzenia o długości S dopóki nie posiadasz odczytów dłuższych niż S
2. Niemożliwe jest rozwiązanie struktury powtórzenia o długości S dopóki nie posiadasz odczytów dłuższych niż S
3. Niemożliwe jest rozwiązanie struktury powtórzenia o długości S dopóki nie posiadasz odczytów dłuższych niż S

odczyty paired-end

- bardzo pomocne podczas tworzenia scaffoldów
- mogą pochodzić z różnej długości fragmentów
- te z krótkich najczęściej są składane w ramach jednego kontigu – przyczyniają się do wydłużenia kontigu
- te z długich czasem dzielą się pomiędzy kontigi – jest to przesłanka do połączenie tych kontigów w scaffold

odczyty paired-end



miara sukcesu – N50

Przykład:

- mamy 10 kontigów o długościach: 1, 1, 3, 3, 5, 7, 8, 12, 16, 20
- Suma: 76 (tyle nukleotydów jest w naszym złożeniu)
- patrzymy „od góry” który kontig przekracza połowę sumy, czyli 38:

  20+16 = 36 *(<38)*
  20+16+12 = 48 *(>38)* **N50 = 12**

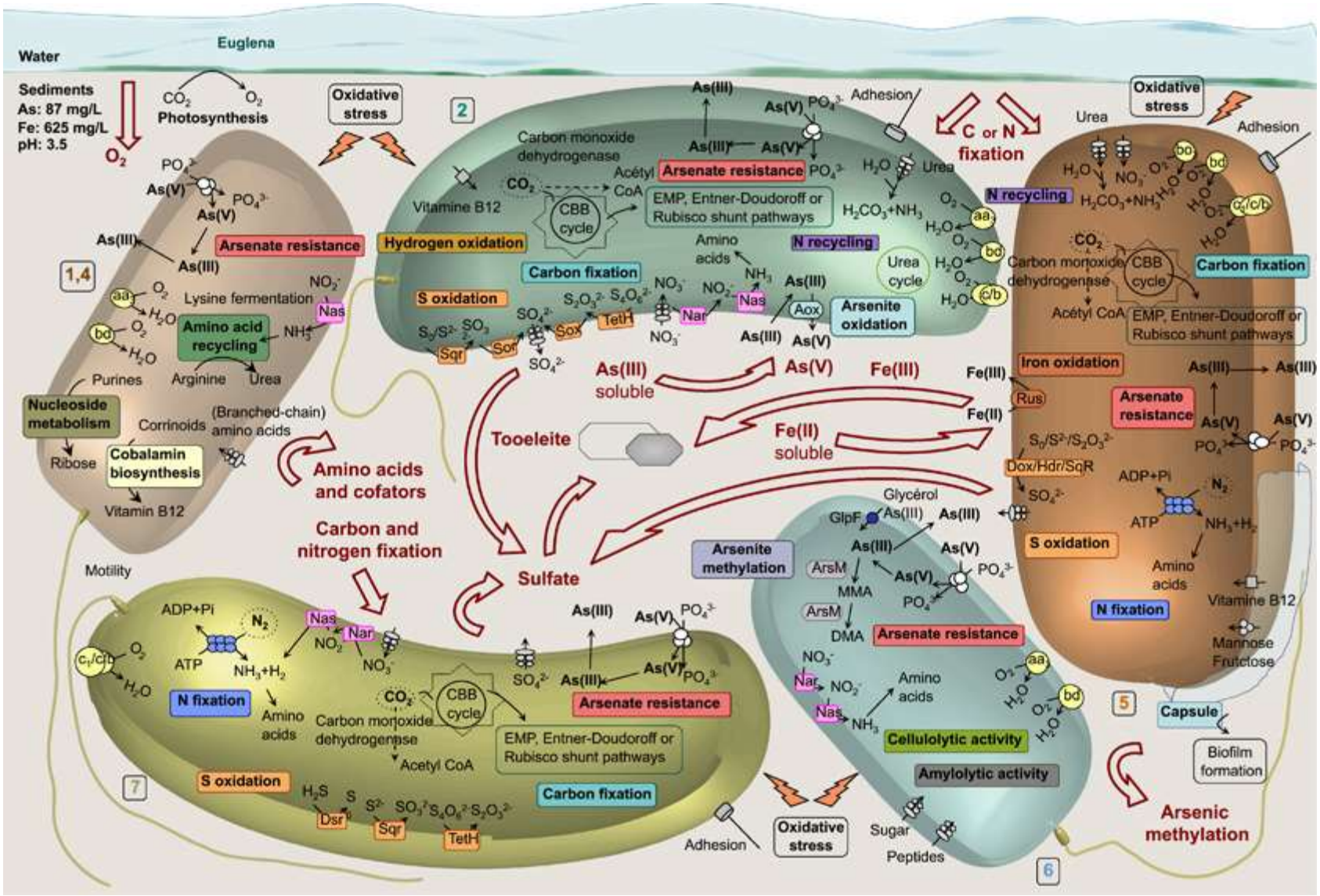
N50: długość kontigu, który wraz z krótszymi od niego zawiera połowę nukleotydów w złożeniu (połowę sumy długości wszystkich kontigów)

polecane narzędzia:

- Velvet
- Soap *de novo*
- aBYss
- Celera assembler
- Canu (PacBio, Nanopore)

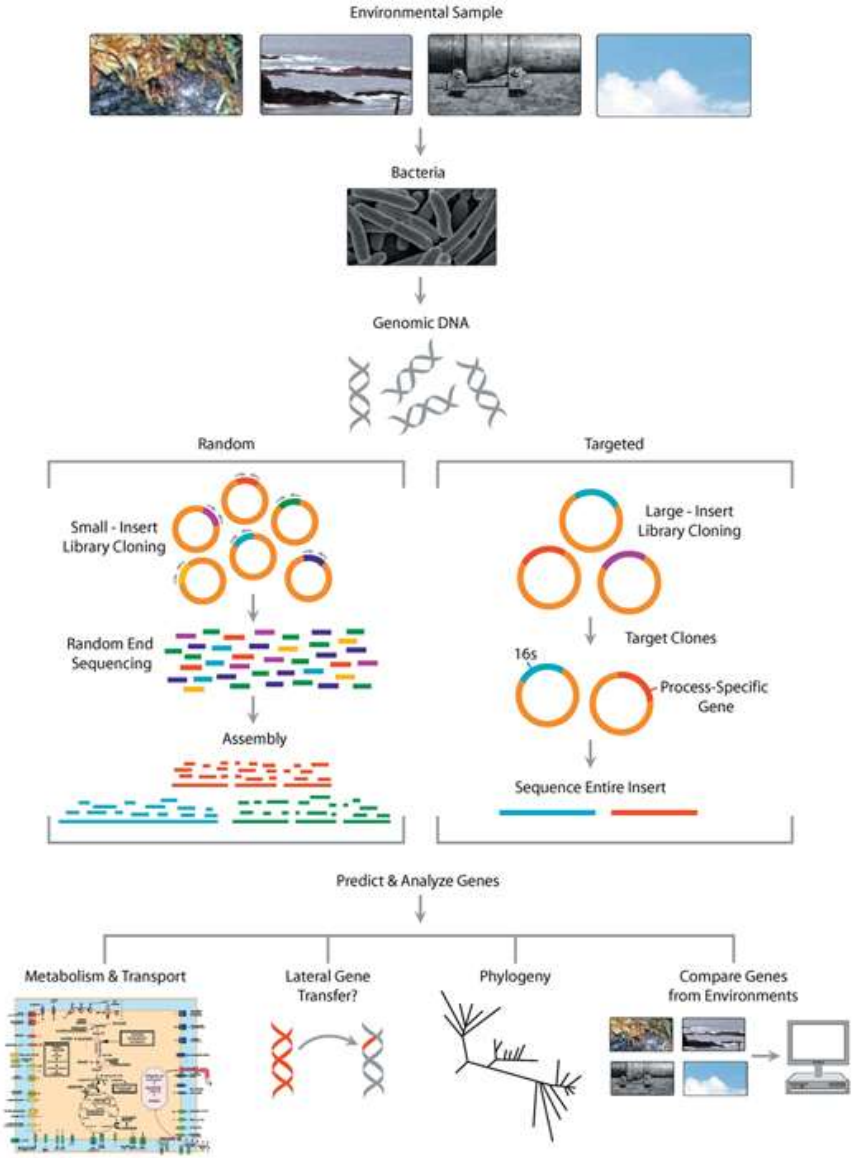
sekwencjonowanie metagenomów

Metabolic diversity among main microorganisms inside an arsenic-rich ecosystem revealed by meta- and proteo-genomics

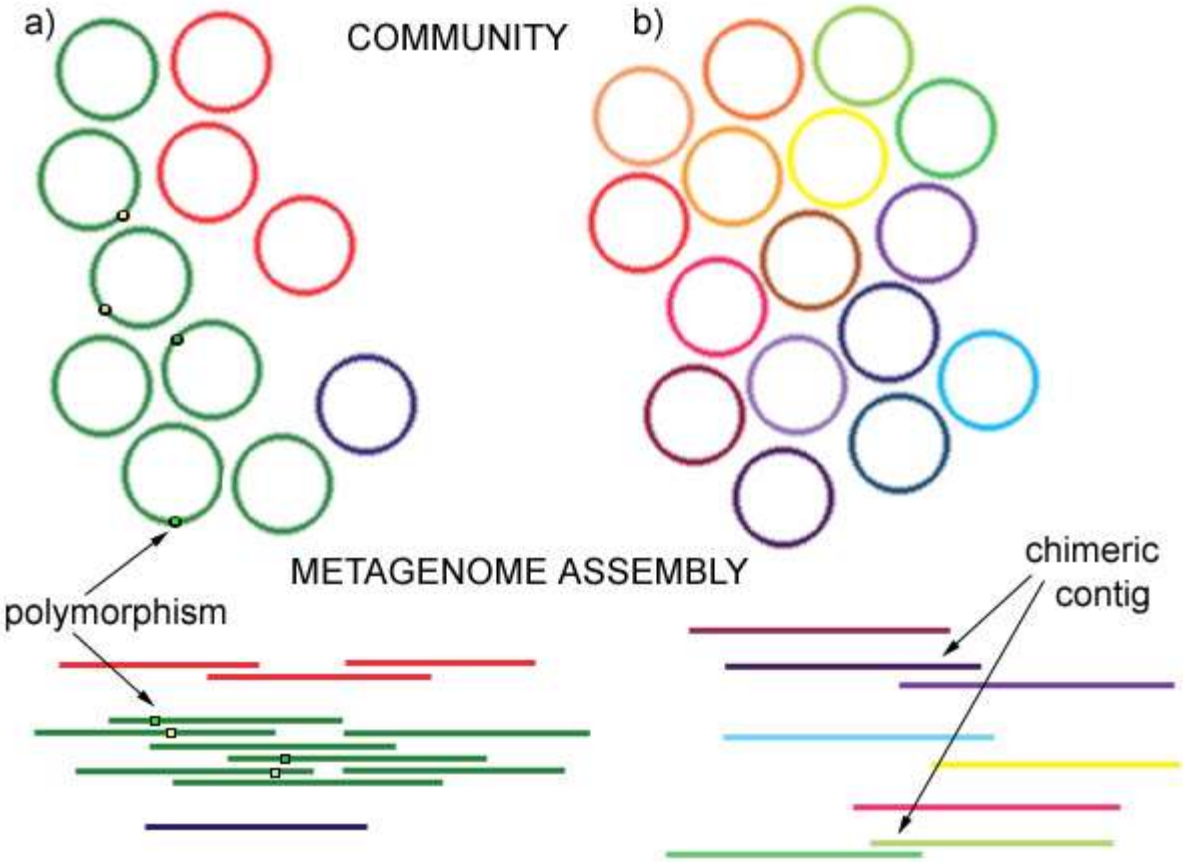


Bertin NP. (2011) ISME 5:1735

sekwencjonowanie metagenomów



sekwencjonowanie metagenomów



http://wiki.biomine.skelleftea.se

sekwencjonowanie metagenomów

Vol 464|4 March 2010|doi:10.1038/nature08821 nature

ARTICLES

A human gut microbial gene catalogue established by metagenomic sequencing

Junjie Qin[1]*, Ruiqiang Li[1]*, Jeroen Raes[2,3], Manimozhiyan Arumugam[2], Kristoffer Solvsten Burgdorf[4], Chaysavanh Manichanh[5], Trine Nielsen[4], Nicolas Pons[6], Florence Levenez[6], Takuji Yamada[2], Daniel R. Mende[2], Junhua Li[1,7], Junming Xu[1], Shaochuan Li[1], Dongfang Li[1,8], Jianjun Cao[1], Bo Wang[1], Huiqing Liang[1], Huisong Zheng[1], Yinlong Xie[1,7], Julien Tap[6], Patricia Lepage[6], Marcelo Bertalan[9], Jean-Michel Batto[6], Torben Hansen[4], Denis Le Paslier[10], Allan Linneberg[11], H. Bjørn Nielsen[9], Eric Pelletier[10], Pierre Renault[6], Thomas Sicheritz-Ponten[9], Keith Turner[12], Hongmei Zhu[1], Chang Yu[1], Shengting Li[1], Min Jian[1], Yan Zhou[1], Yingrui Li[1], Xiuqing Zhang[1], Songgang Li[1], Nan Qin[1], Huanming Yang[1], Jian Wang[1], Søren Brunak[9], Joel Doré[6], Francisco Guarner[5], Karsten Kristiansen[13], Oluf Pedersen[4,14], Julian Parkhill[12], Jean Weissenbach[10], MetaHIT Consortium†, Peer Bork[2], S. Dusko Ehrlich[6] & Jun Wang[1,13]

To understand the impact of gut microbes on human health and well-being it is crucial to assess their genetic potential. Here we describe the Illumina-based metagenomic sequencing, assembly and characterization of 3.3 million non-redundant microbial genes, derived from 576.7 gigabases of sequence, from faecal samples of 124 European individuals. The gene set, ~150 times larger than the human gene complement, contains an overwhelming majority of the prevalent (more frequent) microbial genes of the cohort and probably includes a large proportion of the prevalent human intestinal microbial genes. The genes are largely shared among individuals of the cohort. Over 99% of the genes are bacterial, indicating that the entire cohort harbours between 1,000 and 1,150 prevalent bacterial species and each individual at least 160 such species, which are also largely shared. We define and describe the minimal gut metagenome and the minimal gut bacterial genome in terms of functions present in all individuals and most bacteria, respectively.

sekwencjonowanie metagenomów

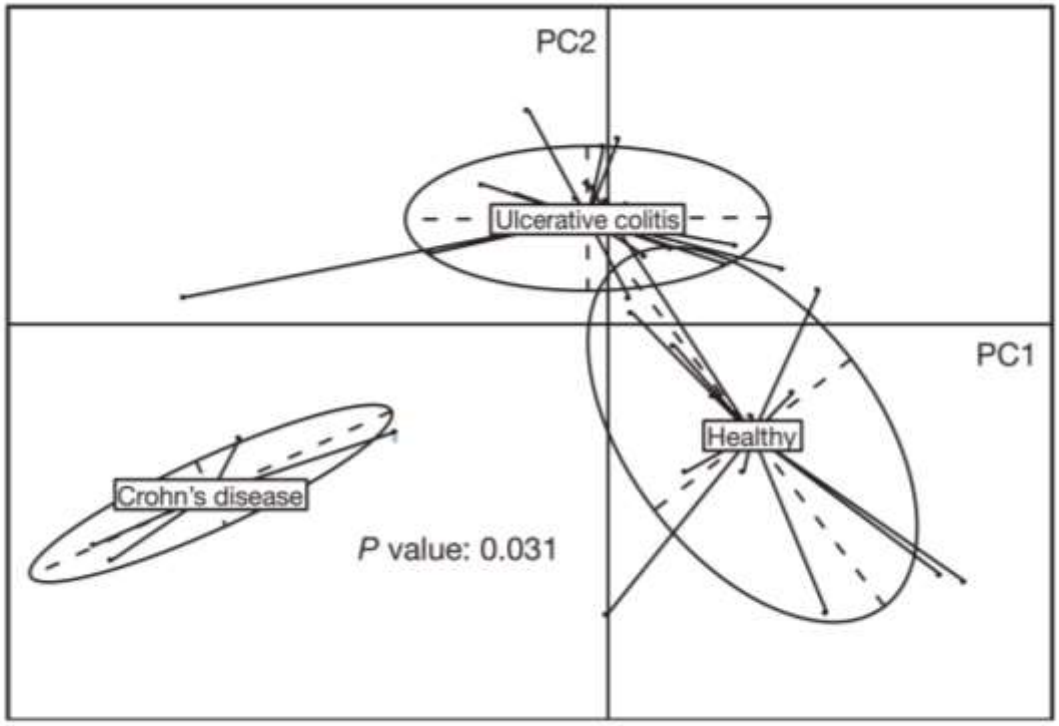
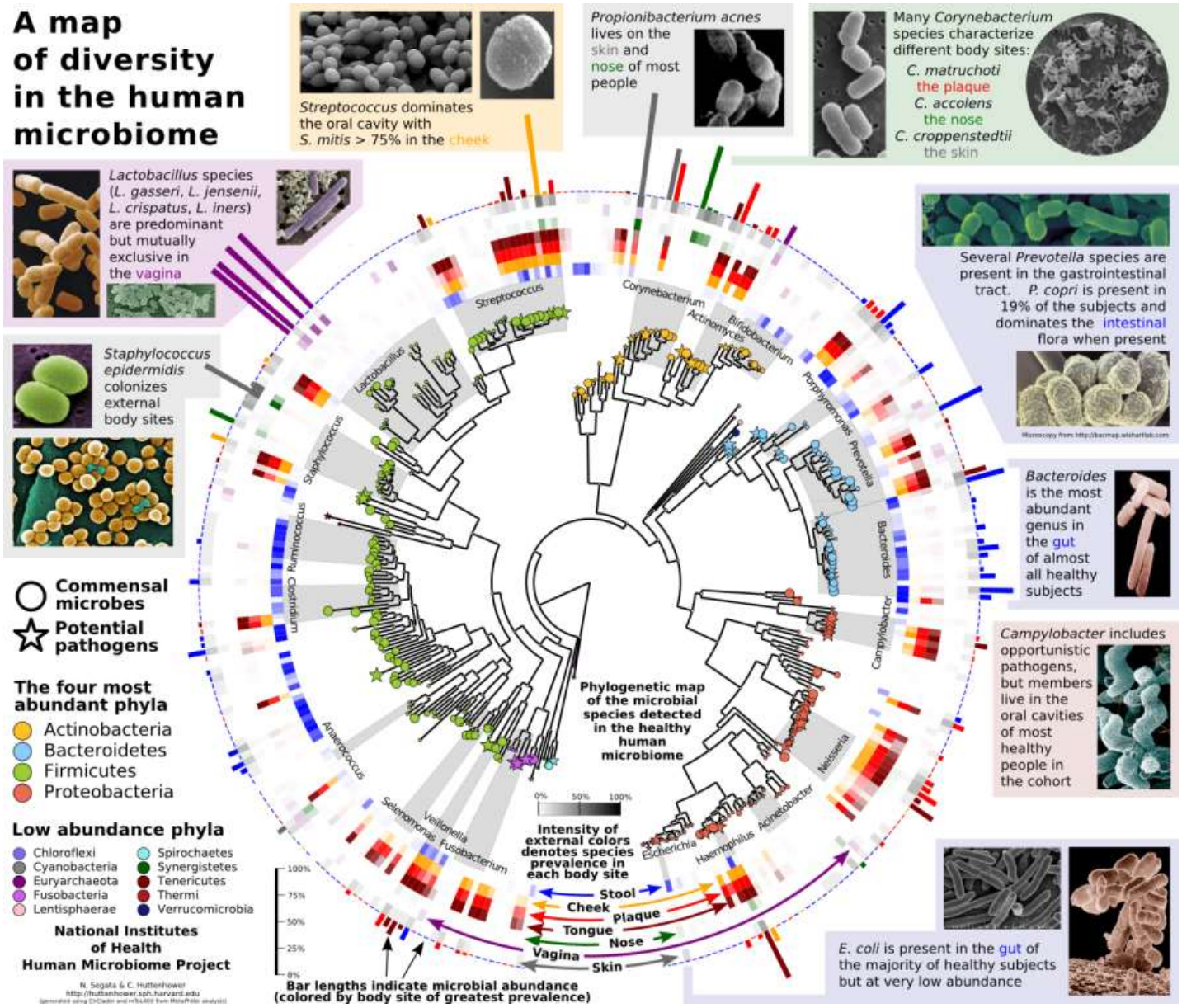


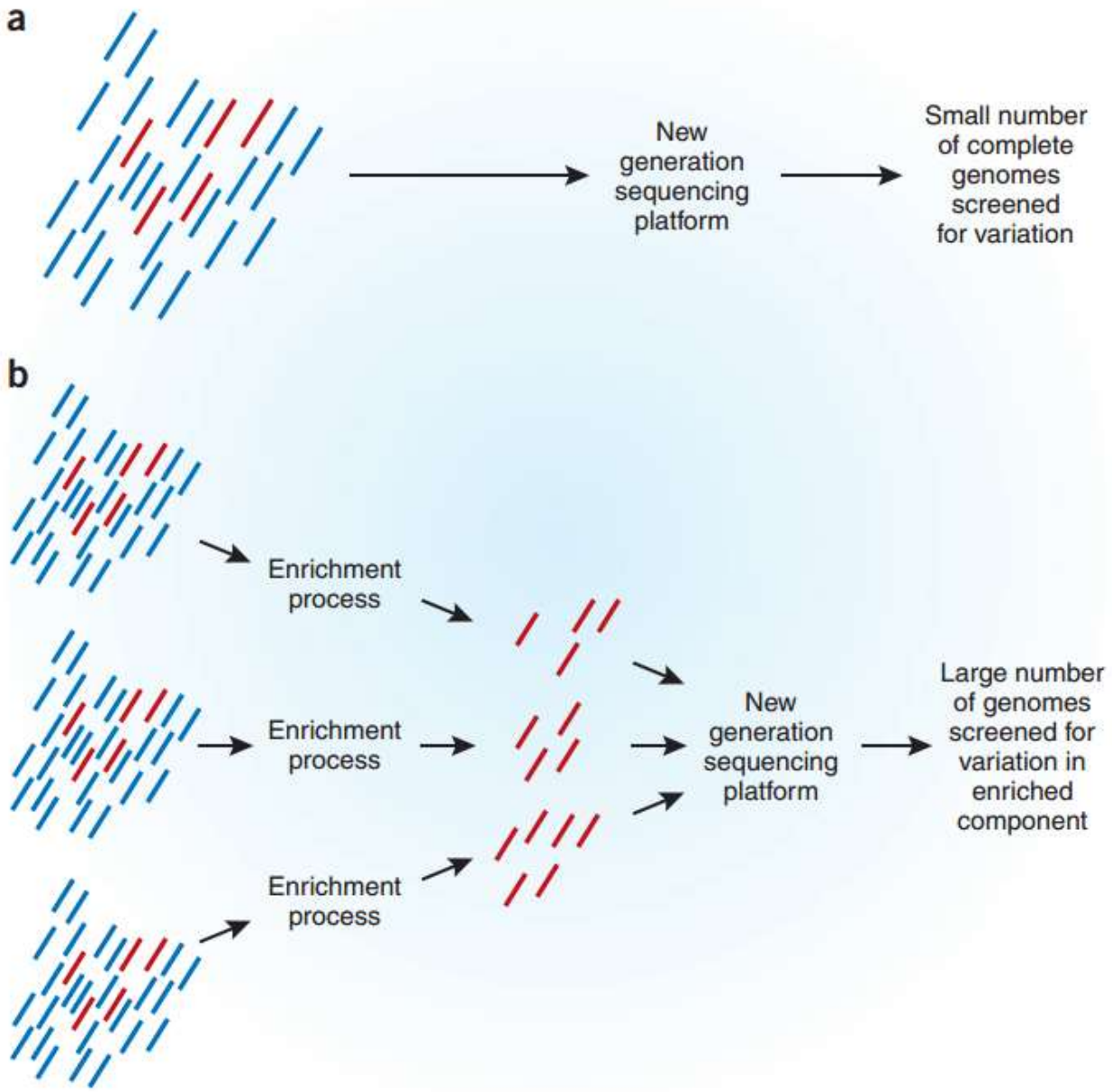
Figure 4 | Bacterial species abundance differentiates IBD patients and healthy individuals. Principal component analysis with health status as

sekwencjonowanie metagenomów



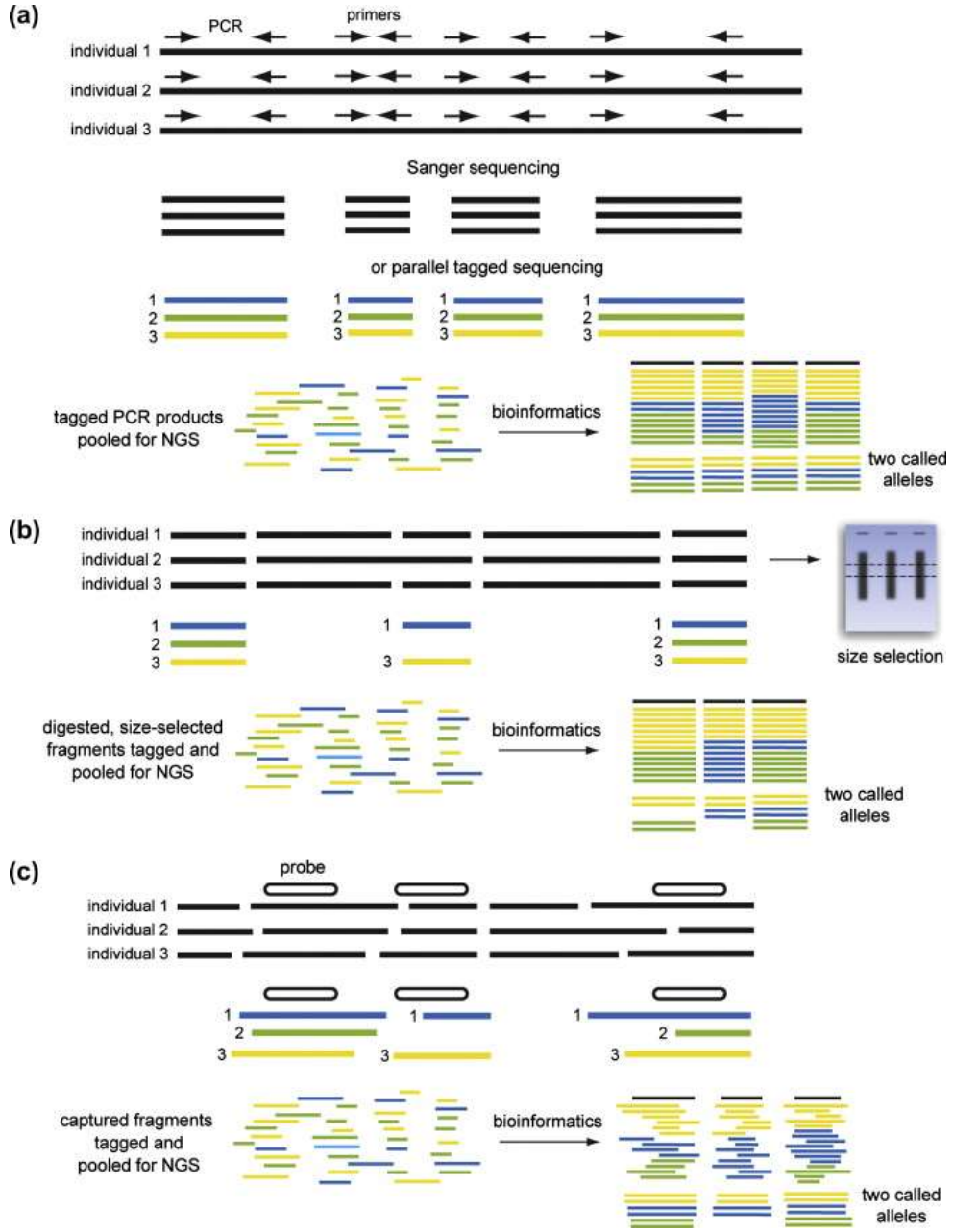
Human Microbiome Project

resekwencjonowanie



Stratton M (2008) Nat. Biotech 26:65

resekwencjonowanie



McCormack JE (2011) Mol Phyl Evol

resekwencjonowanie

Zalety:

- pozwala na uzyskanie tych samych informacji co sekwencjonowanie *de novo*
- wymaga znacznie mniejszego pokrycia sekwencji odczytami
- tańsze i szybsze

Wady:

- Składanie sekwencji wymaga genomu referencyjnego
- nie nadaje się do szybko zmieniających się genomów oraz organizmów o silnym zróżnicowaniu genetycznym w populacji

resekwencjonowanie

Resekwencjonowanie fragmentów genomów:

- pozwala na studiowanie istotnych fragmentów genomu w dużej skali (setki/tysiące osobników)
- najczęściej stosowane w identyfikacji zmienności genetycznej zasocjowanej z chorobami (rak, schorzenia genetyczne)

Rodzaje bibliotek zawierających fragmenty genomu:

- biblioteki egzomowe
- nakierowane na rejony uwikłane w schorzenia genetyczne
- nakierowane na uzupełnienie przerw w sekwencjach genomowych
- nakierowane na geny istotne dla analiz filogenetycznych

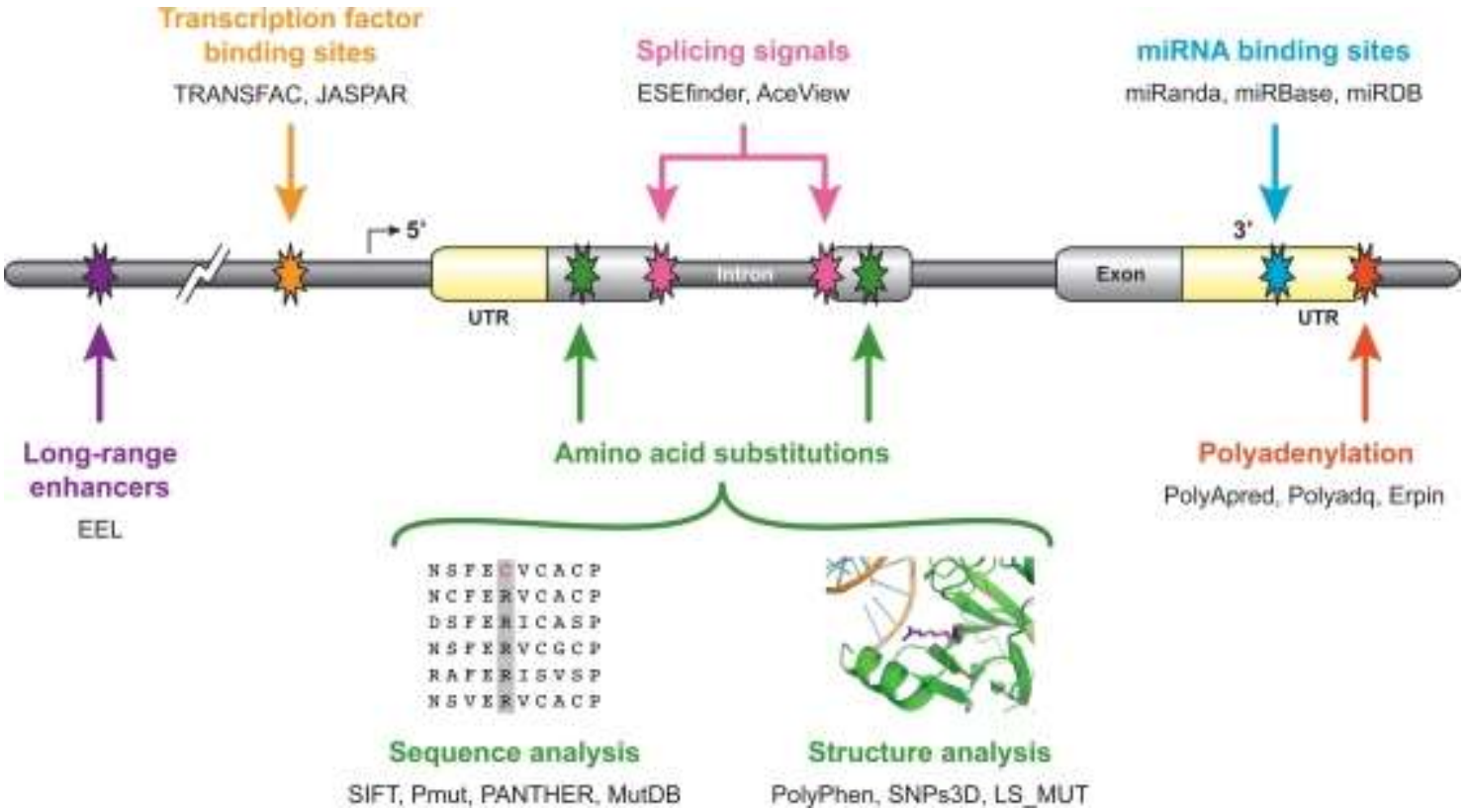
analiza wariacji

- SNP, krótkie delecje i insercje (pojedyncze odczyty, paired-ends)
- rearanżacje chromosomowe (odczyty paired-ends, mate-pair)
- wariacje liczby kopii (pojedyncze odczyty, paired-ends)

Podstawy:

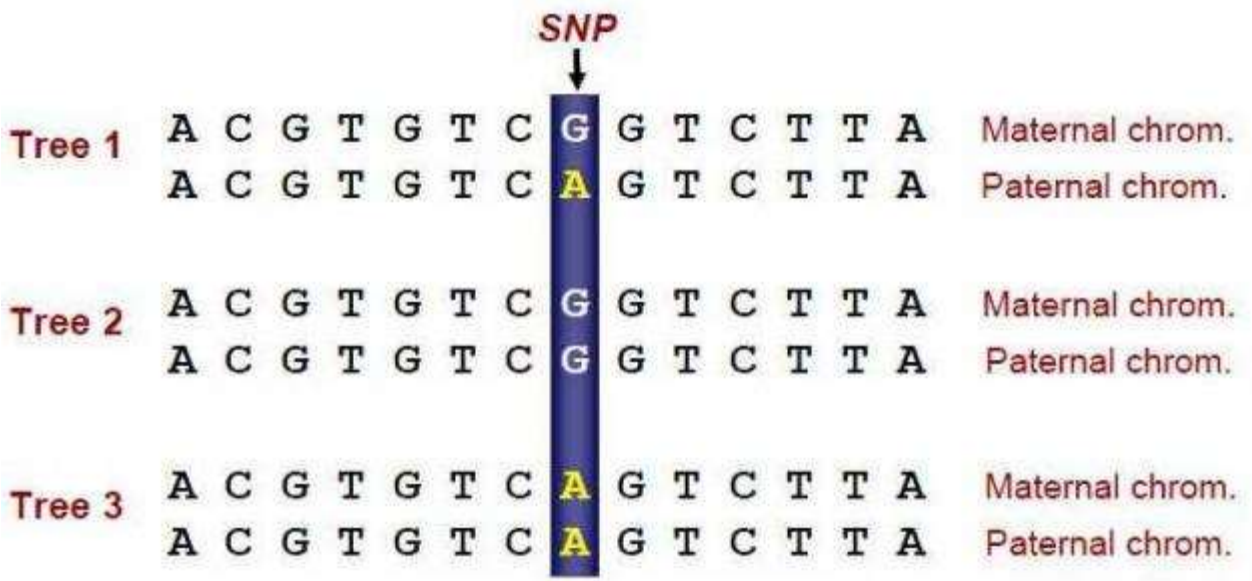
- allele – wersje genu występująca w populacji (A, a)
- genotyp – zestawy alleli danego genu występujące w populacji (AA, Aa, aa)

analiza wariacji SNV/SNP

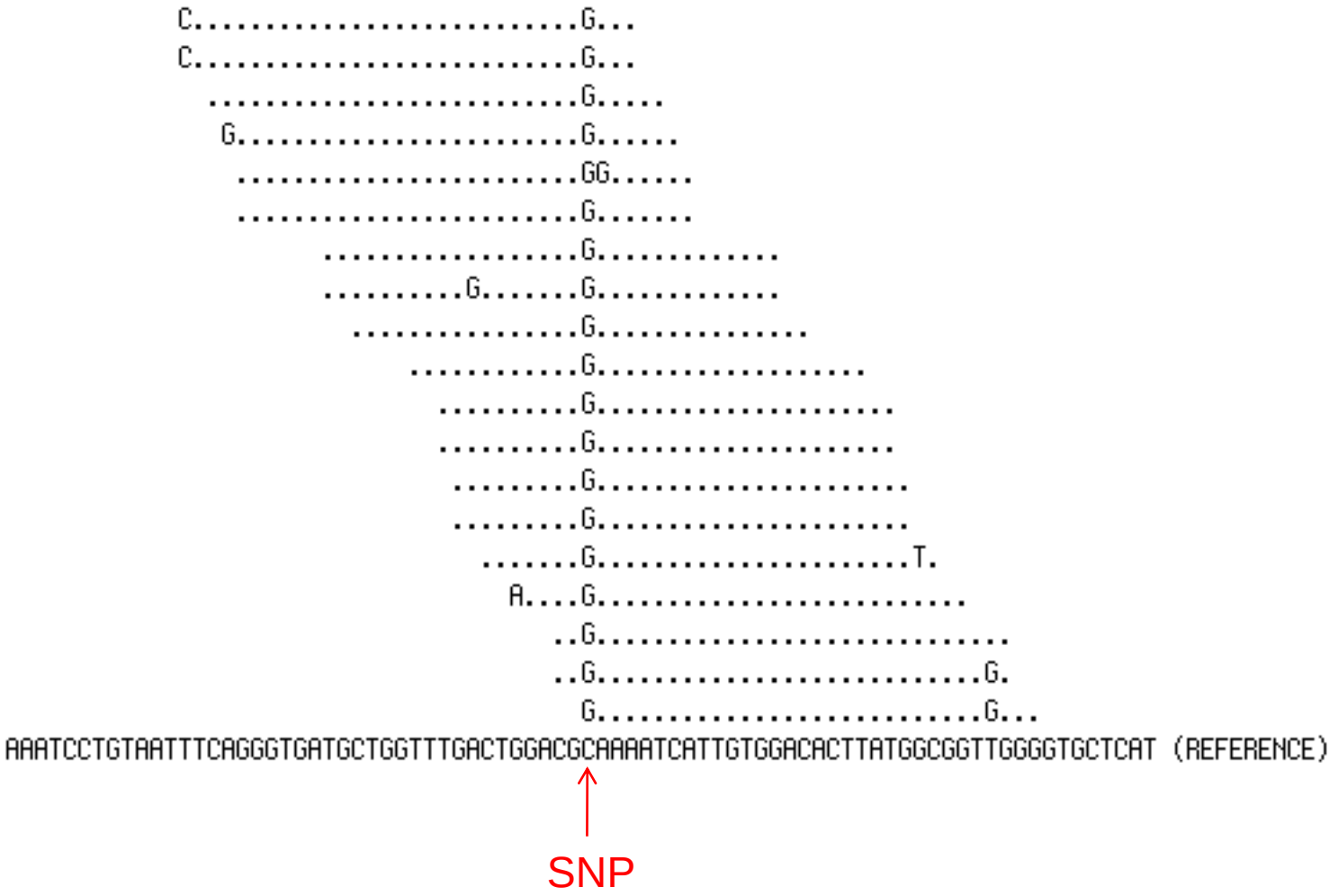


Lee W (2009) Hum Genet 126:481

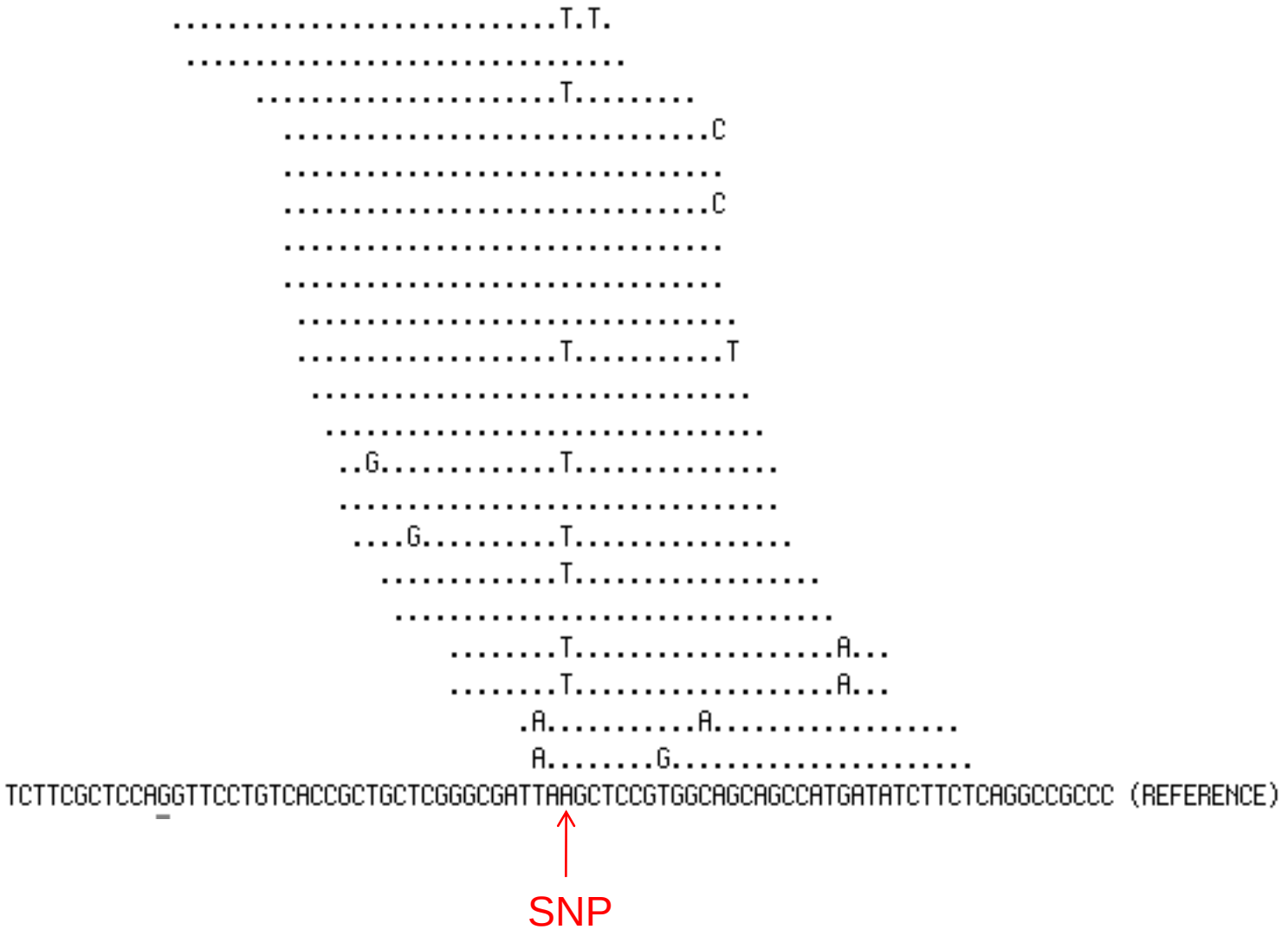
analiza wariacji SNV/SNP



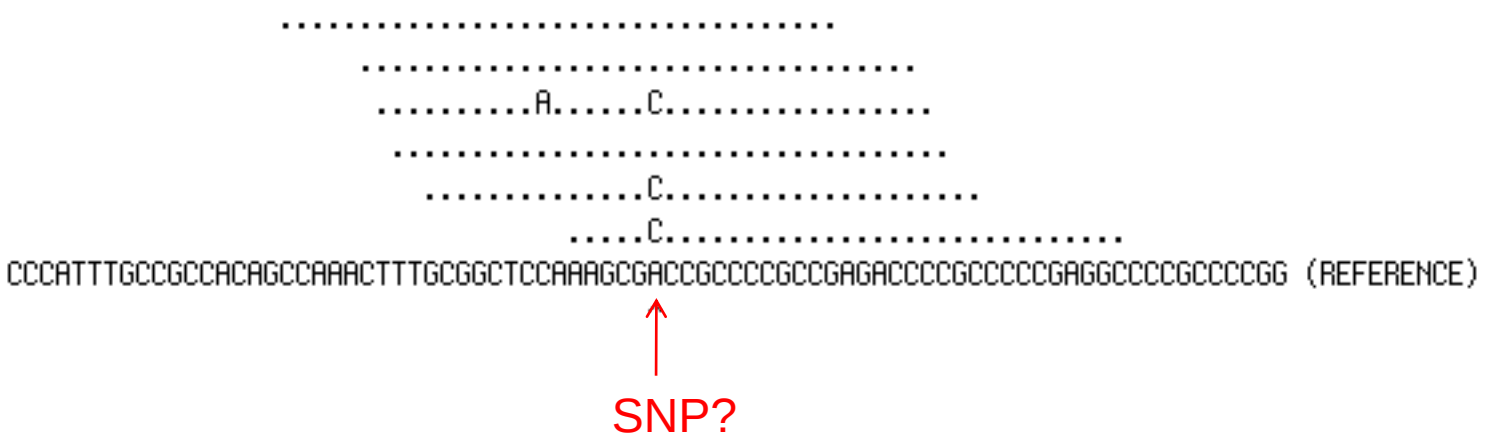
analiza wariacji SNV/SNP



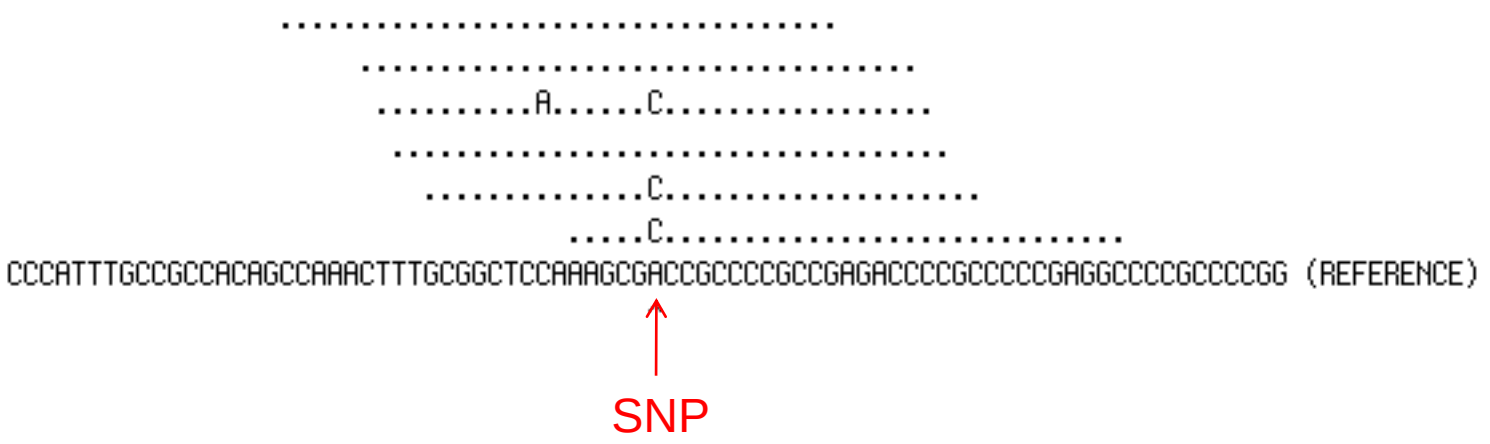
analiza wariacji SNV/SNP



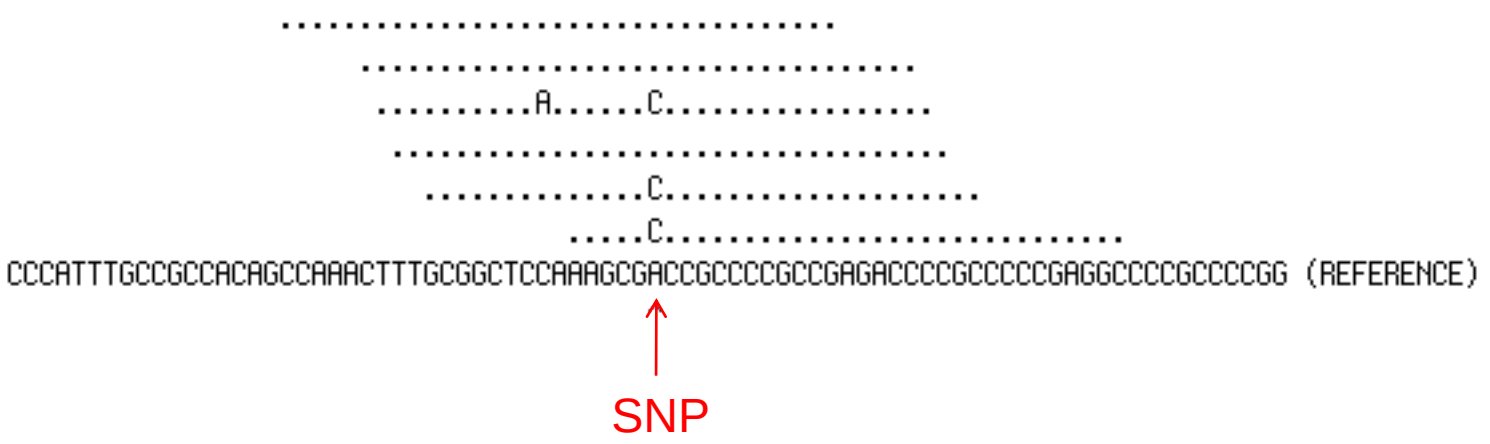
analiza wariacji SNV/SNP



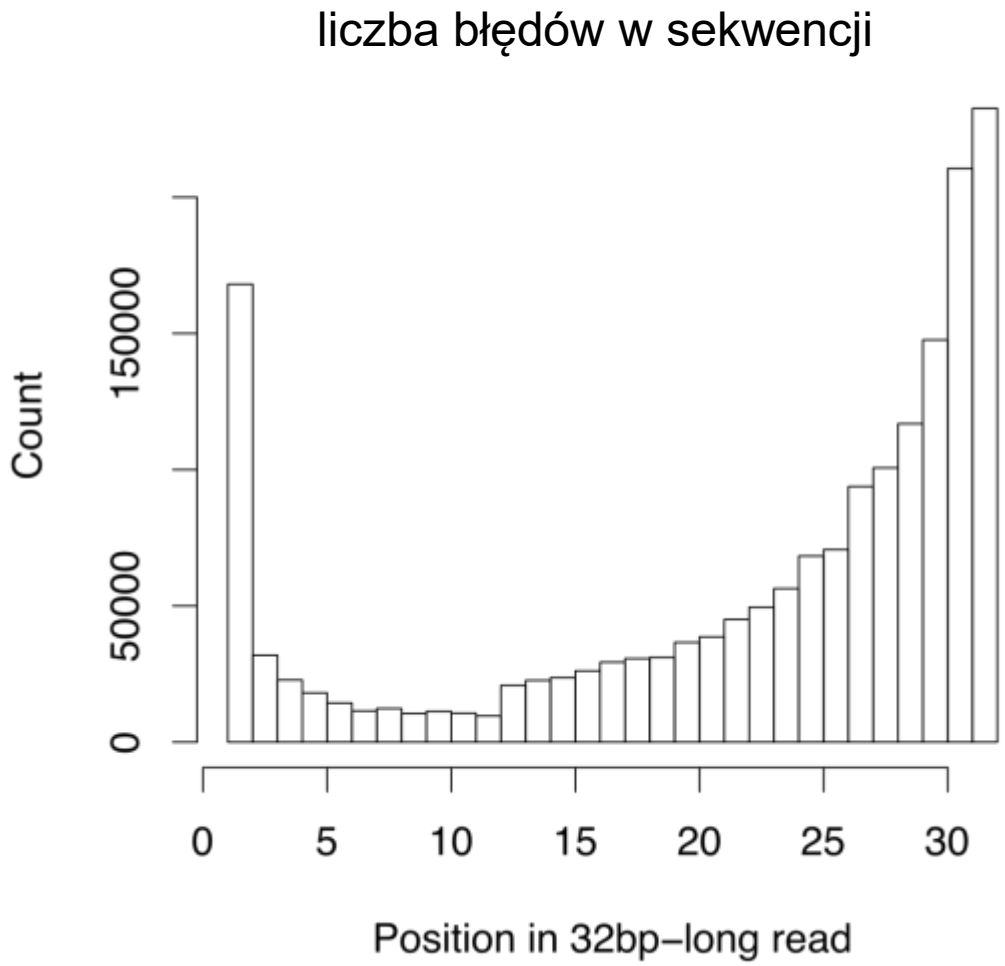
analiza wariacji SNV/SNP



analiza wariacji SNV/SNP



analiza wariacji SNV/SNP



genotypowanie

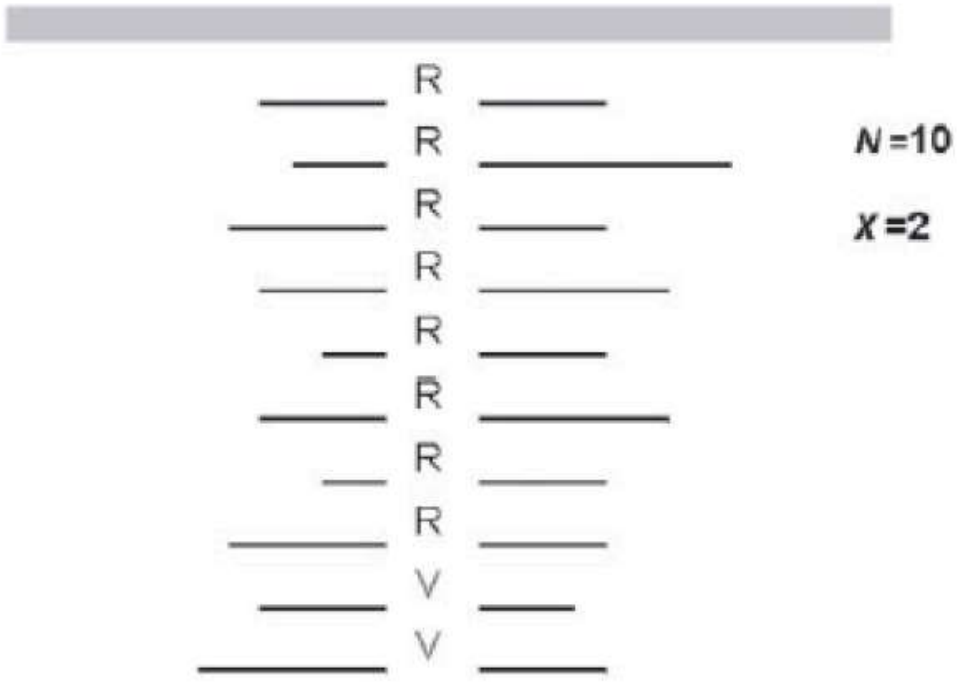


Fig. 1. Schematic of 10 aligned next-generation sequencing reads (R = reference nucleotide, V = variant nucleotide) for a single base position. *N* is read depth. *X* is variant count.

Cel: rozróżnić genotypy RR, RV i VV

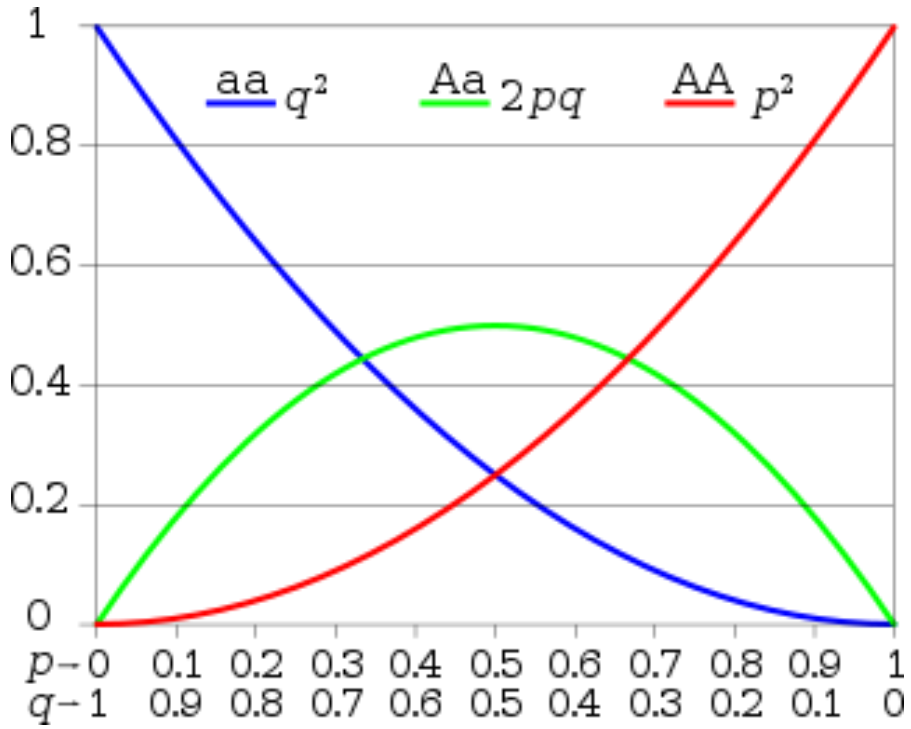
podejścia

- filtrowanie według jakości nukleotydów a następnie według zadanych progów:
  - jeśli % wariacji jest mniejszy niż próg 1 : RR
  - jeśli % wariacji jest większy niż próg 1 a mniejszy niż próg 2 : RV
  - jeśli % wariacji jest większy niż próg 2 : VV
- klasyfikator Bayesa
- Maksymalizacja oczekiwanych wariacji (expectation-maximalization) – optymalizacja parametrów aby maksymalizowały prawdopodobieństwo wariacji w obserwowanych danych

„state of the art” : Bayes

klasyfikator bayesa

- zakładamy równowagę Hardy'ego-Weinberg'a:
  - częstości alleli i genotypów w populacji są stałe w przypadku braku presji ewolucyjnej



klasyfikator bayesa

- zakładamy równowagę Hardy'ego-Weinberg'a:
  - częstości alleli i genotypów w populacji są stałe w przypadku braku presji ewolucyjnej
- Punkt wyjściowy: oczekiwane częstości i częstości błędów sekwencjonowania
- obliczamy wstępne prawdopodobieństwo dla genotypów RR, RV, VV
- szukamy :

$$P(G|D) = \frac{P(D|G)P(G)}{P(D)} = \frac{P(D|G)\,P(G)}{\sum_{i=1}^{n} P(D|G_i)\,P(G_i)}$$

częstości:
G = wstępne dla testowanego genotypu
D = obserwowane dla danych

- używamy P(G|D) jako końcowego prawdopodobieństwa dla P(G)

SAMTOOLS mpileup

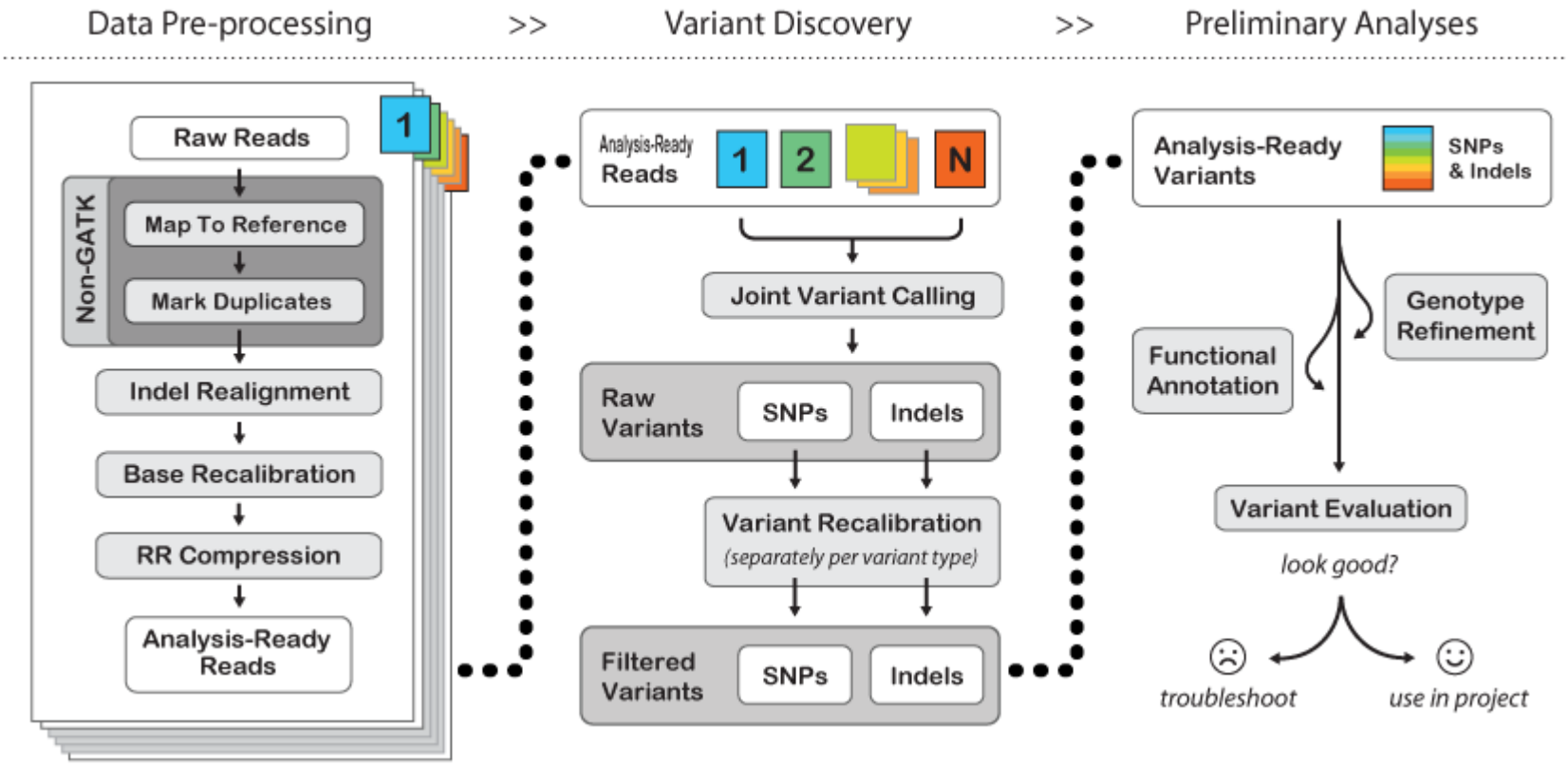
- pileup: dystrybucja zasad wywołanych dla danej pozycji
- ./, odpowiadają dopasowaniom do referencji na nici +/-
- ACGTN/acgtn odpowiadają niedopasowaniom na nici +/-
- ^ oraz $ opisują początek oraz koniec odczytu
- </> odpowiadają intronom na nici +/-
- +[0-9][ACGTN] odpowiada insercjom
- -[0-9][ACGTN] odpowiada delecjom

```
chr15 40328595 g 116 ....,.,,.,..,,,,......,...,..,,,,.,,,,.,.,.,,,,...,.,.....,......,.,..,..,,.,,,,.,,,,,,,,.,,.,..,.,,,...,.,...,,.,..,.,.,,...
chr15 40328596 g 116 .$.$..$,.,,.,..,,,,....$..,...,..,,,,.,,,,.,.,.,,,...,.,......,.,..,..,,.,,,,.,,,,,,,,.,,..,.,,,,...,.,...,,.,..,.,.,
chr15 40328597 t 112 .,G,,G,GG,,,,GG..,.GG,G.,,,,.,,,G,G,G,,,GGG,G,.GGG,G,.G,G.,,.,,,,G,,,,,,,.,,,.G,G,,,...,.,G..,,.,..,G,,....,..
chr15 40328598 g 113 .,.,,.,..,,,,.....,....,..,,,,.,,,,.,.,.,,,,...,.,.....,.,..,..,,.,,,,.,,,,,,,,.,,..,.,,,...,.,...,,.,..,.,.,,....,..
chr15 40328599 c 114 .,.,,.,..,,,,.....,....,..,,,,.,,,,.,.,.,,,,...,.,.....,.,..,..,,.,,,,.,,,,,,,,.,,..,.,,,...,.,...,,.,..,.,.,,....,..
...
chr15 40328696 A 71 ..,........,,..,.,..,.,........,,,,,.,,,,,,,,,.,,.,.,,.,,,,,......,,.........,.,,,,,,,., DDDC;DDADJIF@JFJ#CGGJBB
chr15 40328697 T 71 .$.,........,,..,.,..,.,.......,,,,,.,,,,,,,,,.,,.,.,,.,,,,,......,,.........,.,,,,,,,., C<FD5CECEJGE>JF
chr15 40328698 A 70 .,........,,..,.,..,.,........,,,,,.,,,,,,,,,.,,.,.,,.,,,,,......,,.........,.,,,,,,,., C@D>EEDEHGEAIFJ#EHIJHBI
chr15 40328699 A 70 .,........,,..,.,..,.,........,,,,,.,,,,,,,,,.,,.,.,,.,,,,,......,,.........,.,,,,,,,., D@D>DDEDHGECJEJ#:FIJGCI
chr15 40328700 G 70 .$,$........,,..,.,..,.,.......,,,,,.,,,,,,,,,.,,.,.,,.,,,,,......,,.........,.,,,,,,,., A@D@DC;DHHEEJFJ
chr15 40328701 C 68 .$......,,..,.,..,.,........,,,,,.,,,,,,,,,.,,.,.,,.,,,,,......,,.........,.,,,,,,,., >@B8CDHHEEHEJ#=EJIEAJIJ
chr15 40328702 C 69 .$.$.$>><<>><<>.<>><<>>>>><<<>><<<<<<><><<><<>>><<>>>>><<<<<<<><^~.^~. A4$CDHHECHEJ#AG
chr15 40328703 T 66 >><<>><<>.<>><<>>>>><<<>><<<<<<><><<><<>>><<>>>>><<<<<<<><.. CDHHECHEJ#EGJIG=JGJIGJG
chr15 40328704 G 66 >><<>><<>.<>><<>>>>><<<>><<<<<<><><<><<>>><<>>>>><<<<<<<><.. CDHHECHEJ#HGJIG=JGJIGJG
chr15 40328705 A 66 >><<>><<>.<>><<>>>>><<<>><<<<<<><><<><<>>><<>>>>><<<<<<<><.. CDHHECHEJ#CGJIG=JGJIGJG
chr15 40328706 A 66 >><<>><<>.<>><<>>>>><<<>><<<<<<><><<><<>>><<>>>>><<<<<<<><.. CDHHECHEJ#BGJIG=JGJIGJG
chr15 40328707 A 66 >><<>><<>.<>><<>>>>><<<>><<<<<<><><<><<>>><<>>>>><<<<<<<><.. CDHHECHEJ#DGJIG=JGJIGJG
```

GATK

- korzysta z algorytmu MapReduce (Google, Yahoo!)
- nieograniczona paralelizacja, gdyż identyfikacja SNP i genotypowanie odbywa się dla każdego locus niezależnie
- może analizować oraz łączyć wiele próbek
- używa uniwersalnego formatu bam

GATK


Data Pre-processing
>>
Variant Discovery
>>
Preliminary Analyses
Raw Reads
1
Non-GATK
Map To Reference
Mark Duplicates
Indel Realignment
Base Recalibration
RR Compression
Analysis-Ready Reads
Analysis-Ready Reads
1
2
N
Joint Variant Calling
Raw Variants
SNPs
Indels
Variant Recalibration
(separately per variant type)
Filtered Variants
SNPs
Indels
Analysis-Ready Variants
SNPs & Indels
Functional Annotation
Genotype Refinement
Variant Evaluation
look good?
troubleshoot
use in project

realignment

wiele porównań do referencji ≠ wielokrotne porównanie

```
ref      aggttttataaaac----aattaagtctacagagcaacta
sample   aggttttataaaacAAATaattaagtctacagagcaacta
read1    aggttttataaaac****aa**A**taa
read2     ggttttataaaac****aa**A**taaTt
read3        ttataaaacAAATaattaagtctaca
read4           **C**aaa**T****aattaagtctacagagcaac
read5            aa**T****aattaagtctacagagcaact
read6             **T****aattaagtctacagagcaacta
```

```
ref      aggttttataaaac----aattaagtctacagagcaacta
sample   aggttttataaaacAAATaattaagtctacagagcaacta
read1    aggttttataaaacAAATaa
read2     ggttttataaaacAAATaatt
read3        ttataaaacAAATaattaagtctaca
read4              cAAATaattaagtctacagagcaac
read5               AATaattaagtctacagagcaact
read6                Taattaagtctacagagcaacta
```

realignment

wiele porównań do referencji ≠ wielokrotne porównanie

ref aggttttataaaacAAAAaattaagtctacagagcaacta
sample aggttttataaaacAAA-aattaagtctacagagcaacta
read1 aggttttataaaacAA-Aaattaagtctacagagcaacta
read2 aggttttataaaacA-AAaattaagtctacagagcaacta
read3 aggttttataaaac-AAAaattaagtctacagagcaacta
consensus aggttttataaaac**AAAA**aattaagtctacagagcaacta

Rozwiązania:

- realignment rejonów zawierających liczne wyspy/SNP (GATK)
- wyznaczenie prawdopodobieństwa że pozycja jest źle dopasowana i usunięcie problematycznych rejonów z analizy (SAMTOOLS)

rekalibracja wariacji

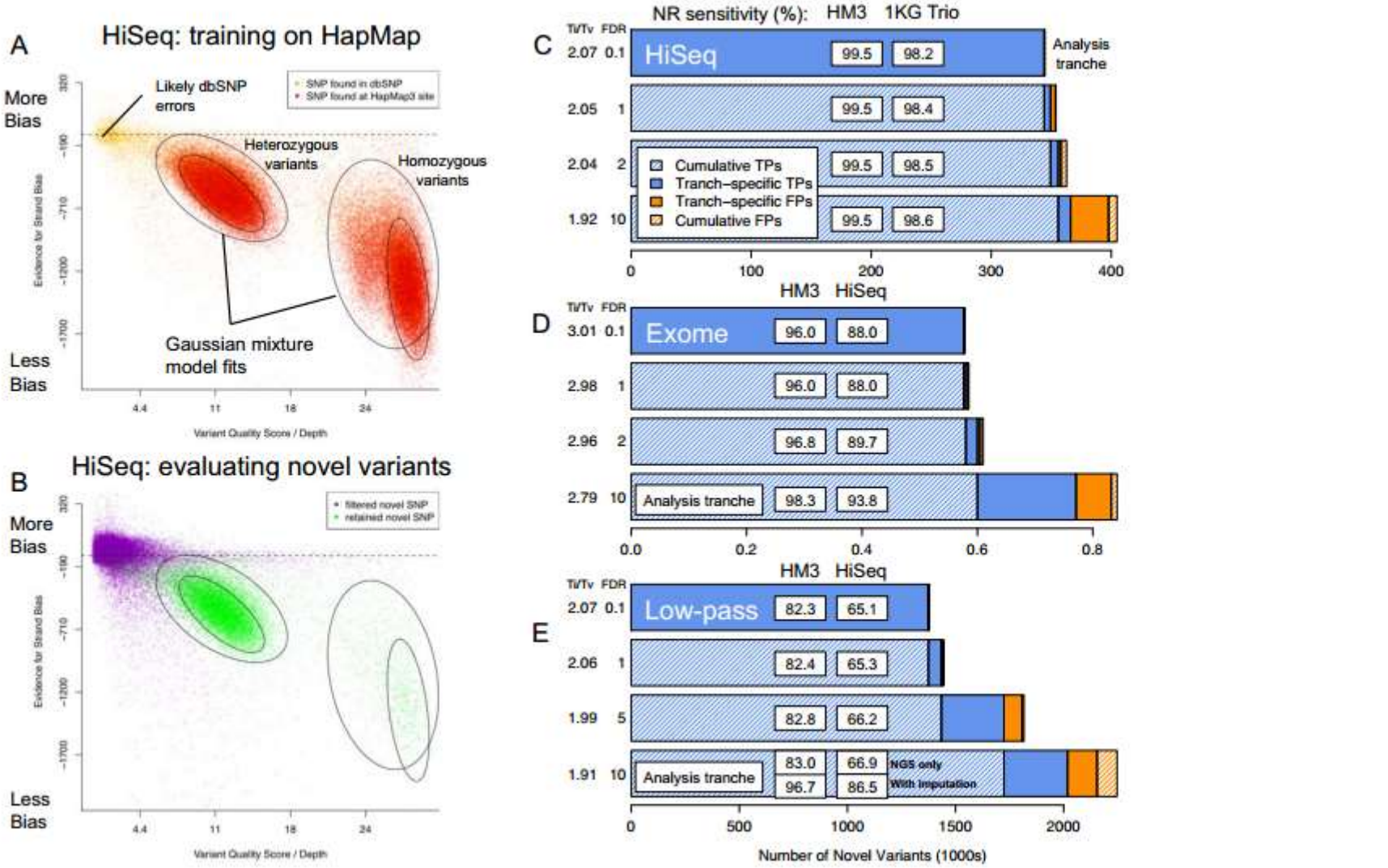
Dlaczego:

- programy zwracają dużo fałszywych wariacji
- liczba faktycznych wariacji genetycznych zależy od jakości i ilości konkretnych danych

Jak:

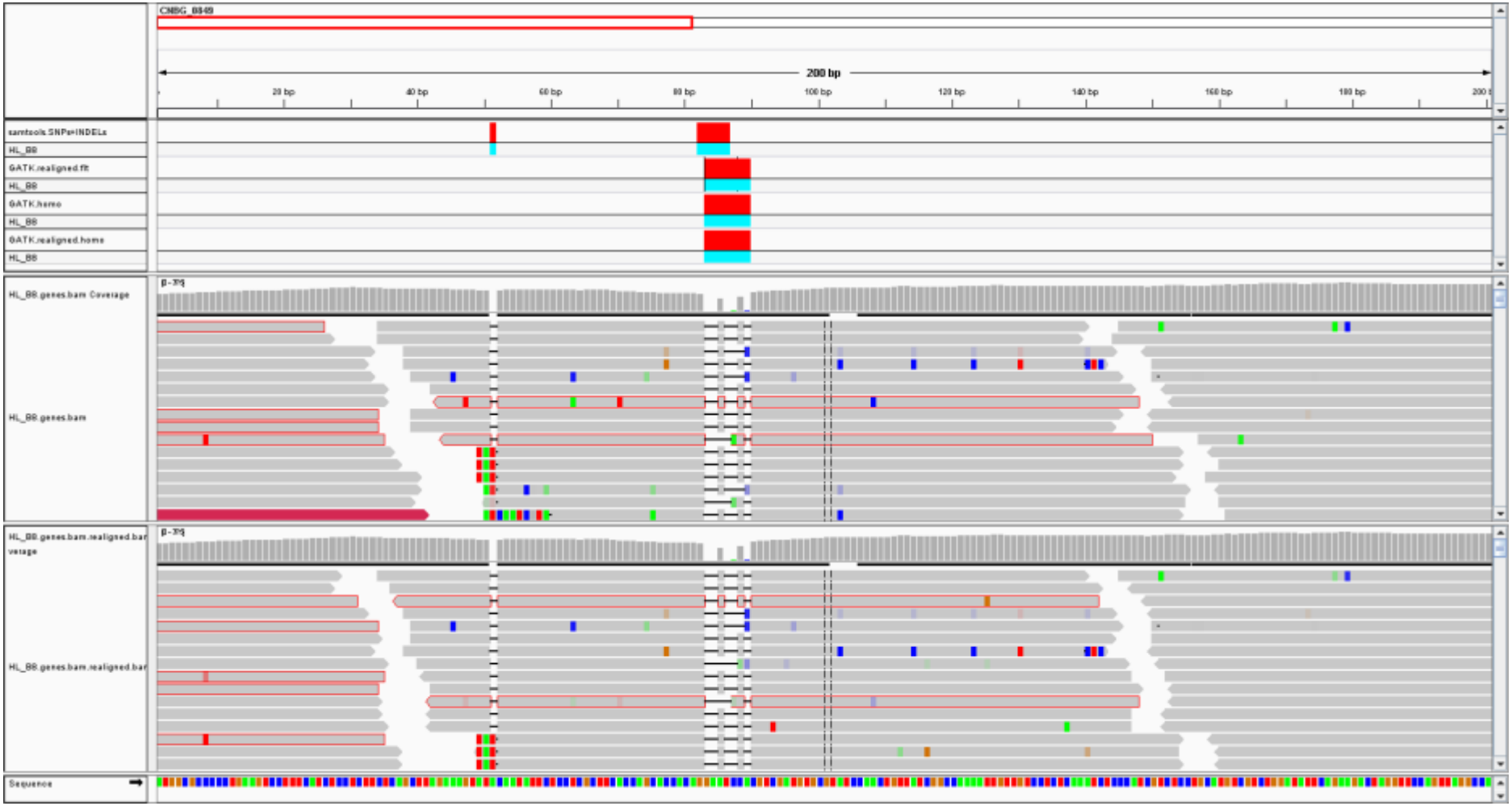
- znane wariacje służą jako wzorzec do wyznaczenia prawdopodobieństwa dla nowych wariacji (np. z HapMap)

rekalibracja wariacji



DePristo et al. Nature Genetics. 2011

analiza wariacji
krótkie insercje/delecje (indels)

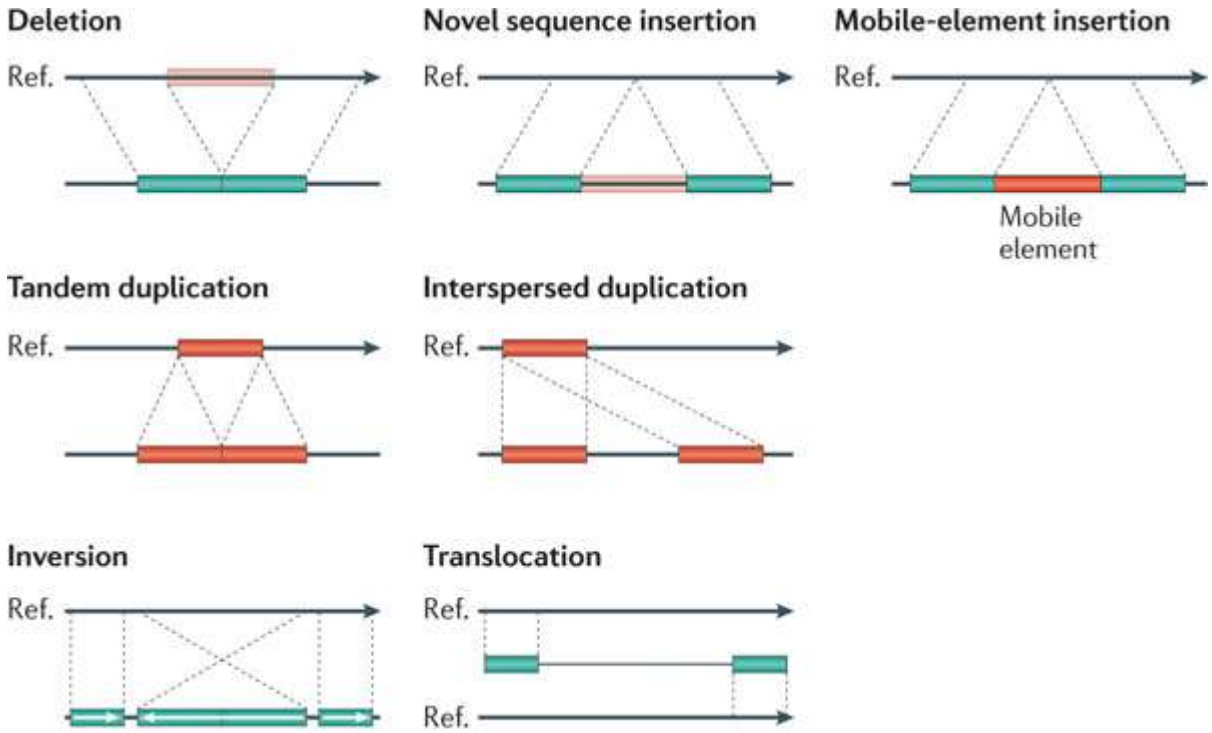


analiza wariacji SNP/indels

Narzędzia:

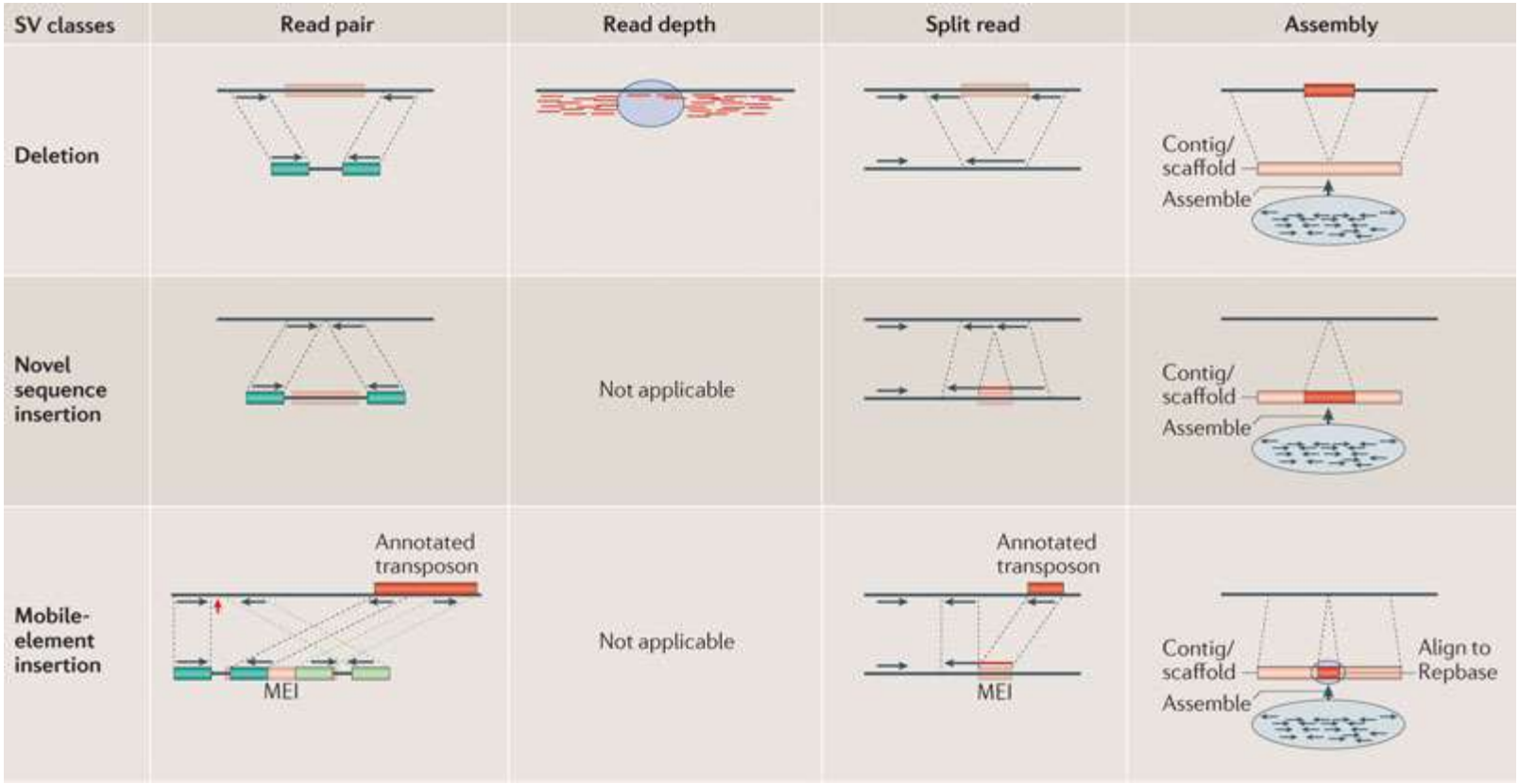
1. GATK Unified Genotyper
2. Komenda MPILEUPz pakietu SAMTOOLS

analiza wariacji
rearanżacje chromosomalne



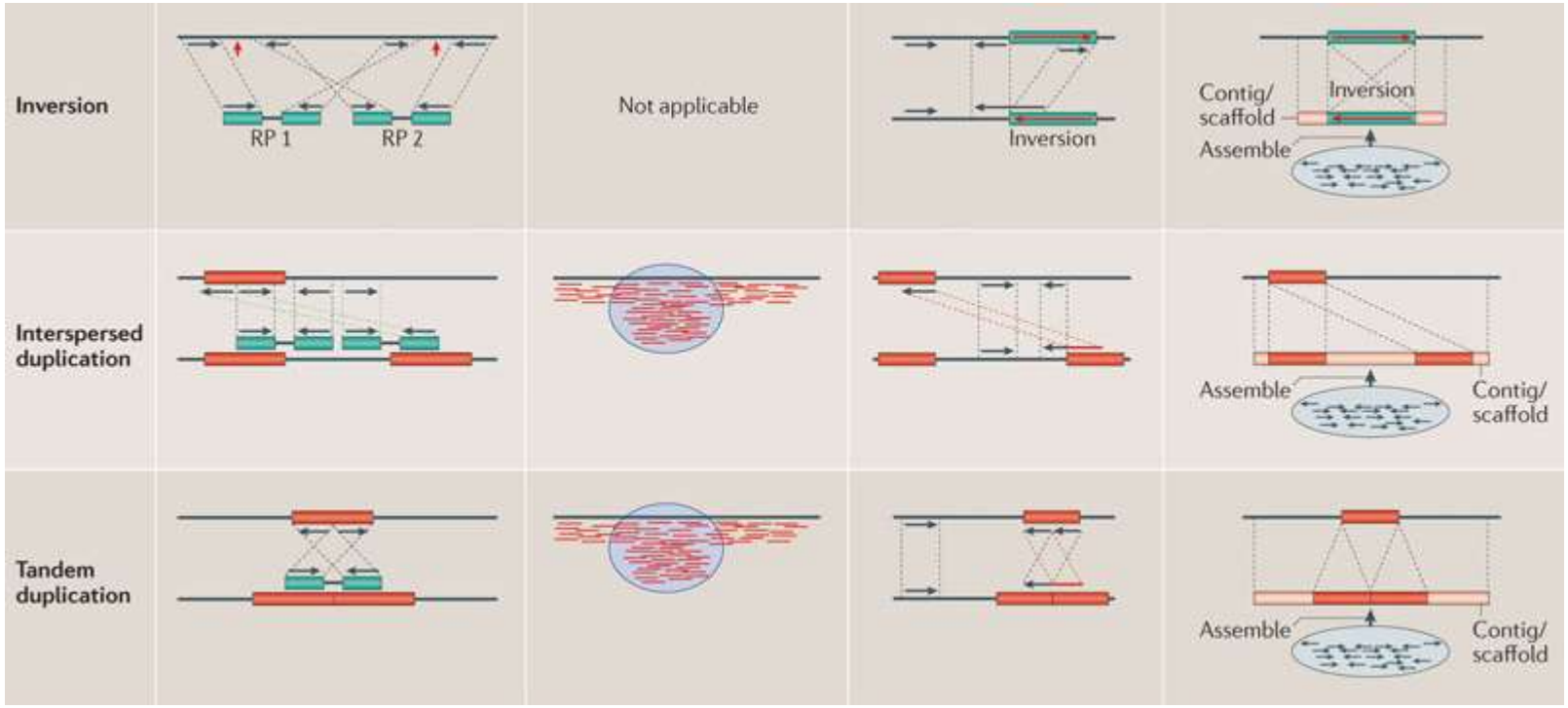
Alkan C (2011) Nat Rev Genet 12:363

analiza wariacji
rearanżacje chromosomalne



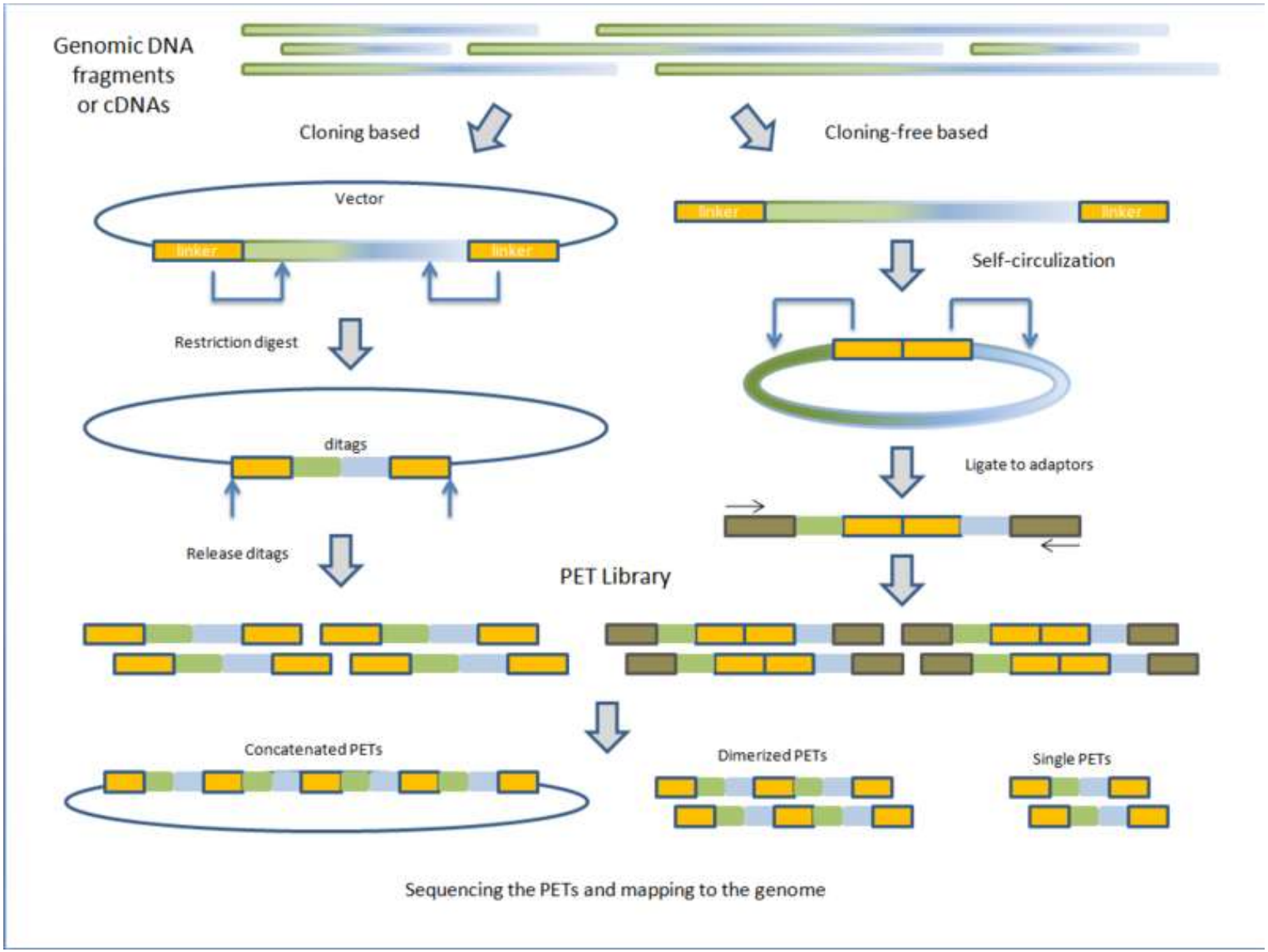
Alkan C (2011) Nat Rev Genet 12:363

analiza wariacji
rearanżacje chromosomalne



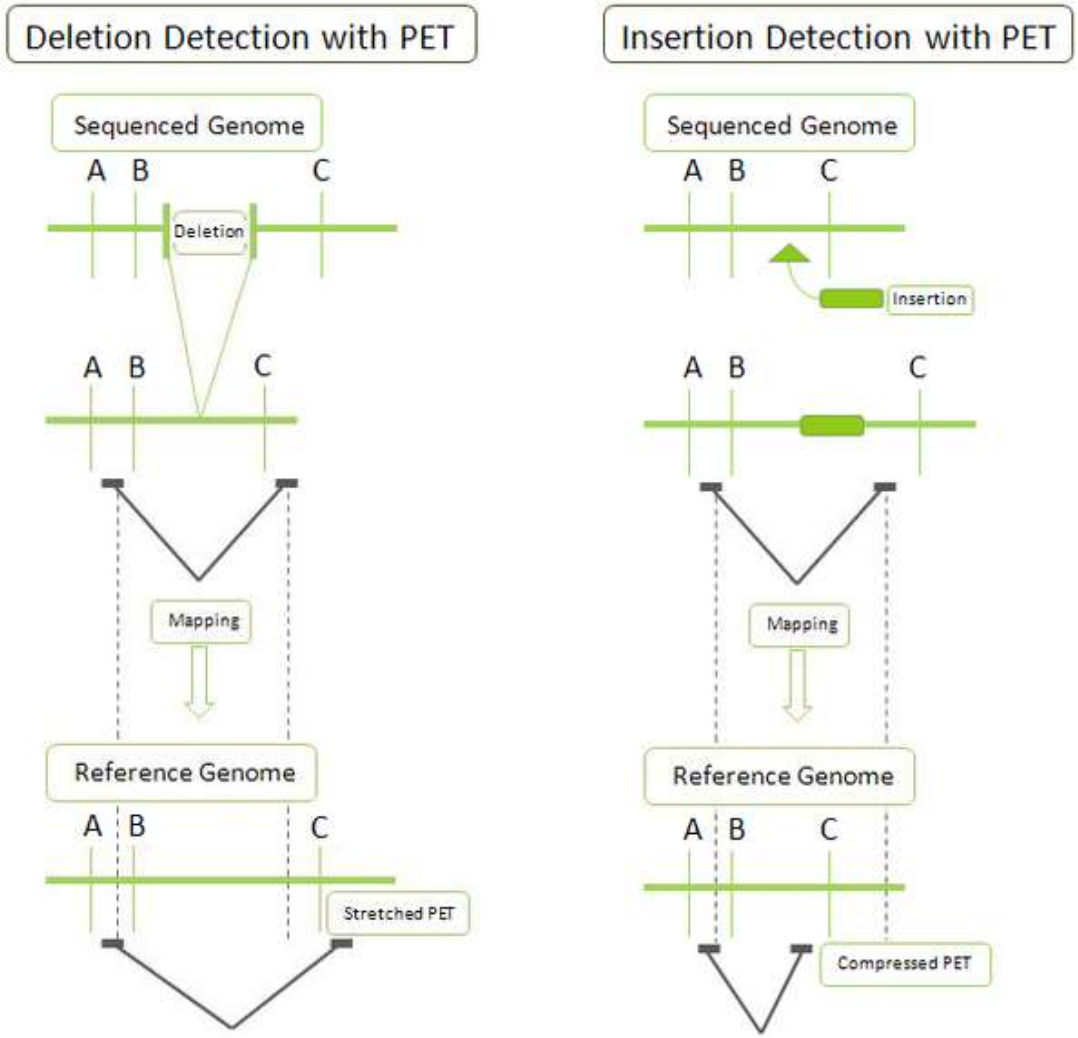
Alkan C (2011) Nat Rev Genet 12:363

analiza wariacji
rearanżacje chromosomalne – PET sequencing



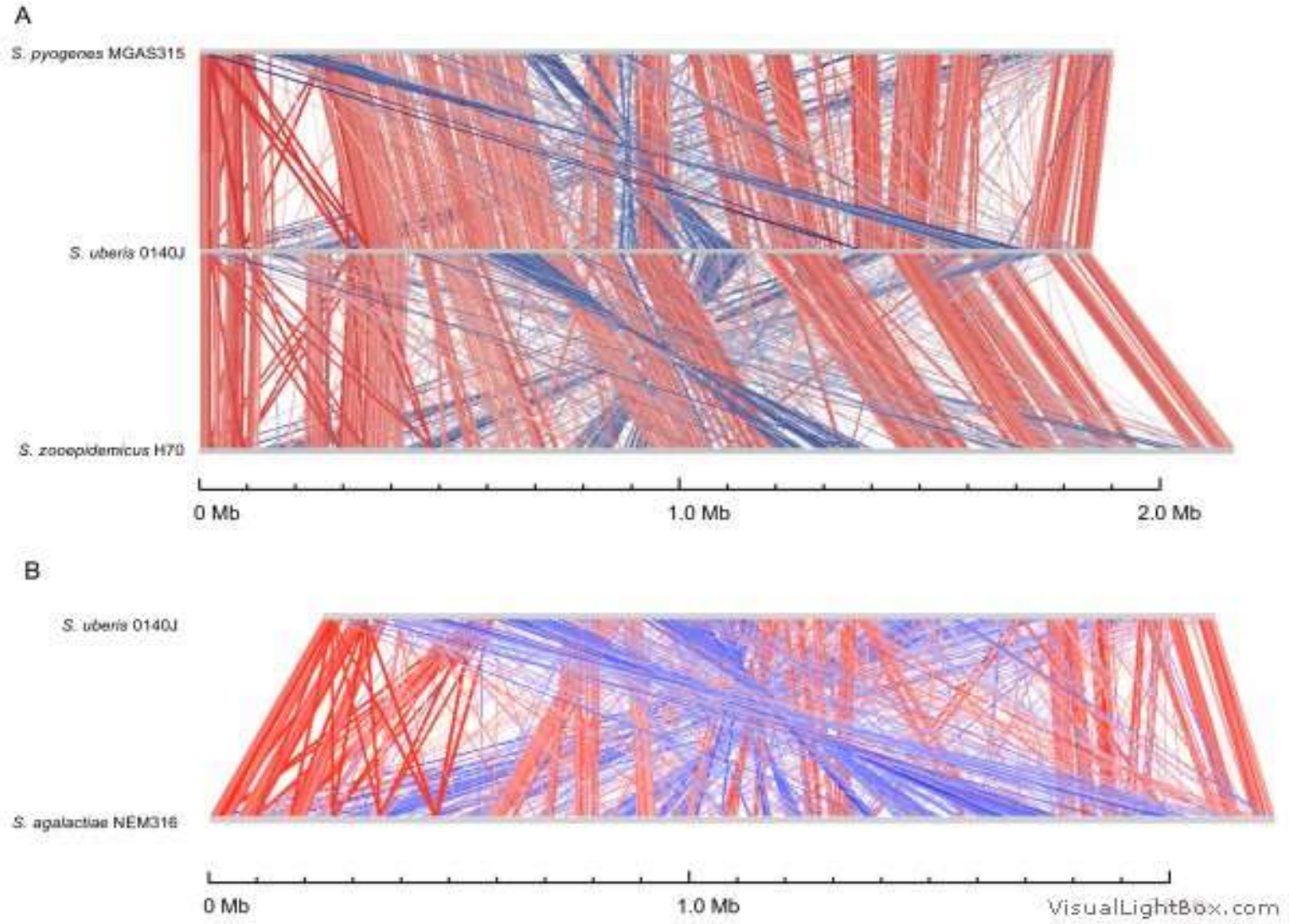
http://wikipedia.org

analiza wariacji
rearanżacje chromosomalne – PET sequencing



http://wikipedia.org

analiza wariacji
rearanżacje chromosomalne



analiza wariacji

ARTICLE

doi:10.1038/nature10983

The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups

Christina Curtis[1,2]†*, Sohrab P. Shah[3,4]*, Suet-Feung Chin[1,2]*, Gulisa Turashvili[3,4]*, Oscar M. Rueda[1,2], Mark J. Dunning[2], Doug Speed[2,5]†, Andy G. Lynch[1,2], Shamith Samarajiwa[1,2], Yinyin Yuan[1,2], Stefan Gräf[1,2], Gavin Ha[3], Gholamreza Haffari[3], Ali Bashashati[3], Roslin Russell[2], Steven McKinney[3,4], METABRIC Group‡, Anita Langerød[6], Andrew Green[7], Elena Provenzano[8], Gordon Wishart[8], Sarah Pinder[9], Peter Watson[3,4,10], Florian Markowetz[1,2], Leigh Murphy[10], Ian Ellis[7], Arnie Purushotham[9,11], Anne-Lise Børresen-Dale[6,12], James D. Brenton[2,13], Simon Tavaré[1,2,5,14], Carlos Caldas[1,2,8,13] & Samuel Aparicio[3,4]

The elucidation of breast cancer subgroups and their molecular drivers requires integrated views of the genome and transcriptome from representative numbers of patients. We present an integrated analysis of copy number and gene expression in a discovery and validation set of 997 and 995 primary breast tumours, respectively, with long-term clinical follow-up. Inherited variants (copy number variants and single nucleotide polymorphisms) and acquired somatic copy number aberrations (CNAs) were associated with expression in ~40% of genes, with the landscape dominated by *cis*- and *trans*-acting CNAs. By delineating expression outlier genes driven in *cis* by CNAs, we identified putative cancer genes, including deletions in *PPP2R2A*, *MTAP* and *MAP2K4*. Unsupervised analysis of paired DNA–RNA profiles revealed novel subgroups with distinct clinical outcomes, which reproduced in the validation cohort. These include a high-risk, oestrogen-receptor-positive 11q13/14 *cis*-acting subgroup and a favourable prognosis subgroup devoid of CNAs. *Trans*-acting aberration hotspots were found to modulate subgroup-specific gene networks, including a TCR deletion-mediated adaptive immune response in the 'CNA-devoid' subgroup and a basal-specific chromosome 5 deletion-associated mitotic network. Our results provide a novel molecular stratification of the breast cancer population, derived from the impact of somatic CNAs on the transcriptome.