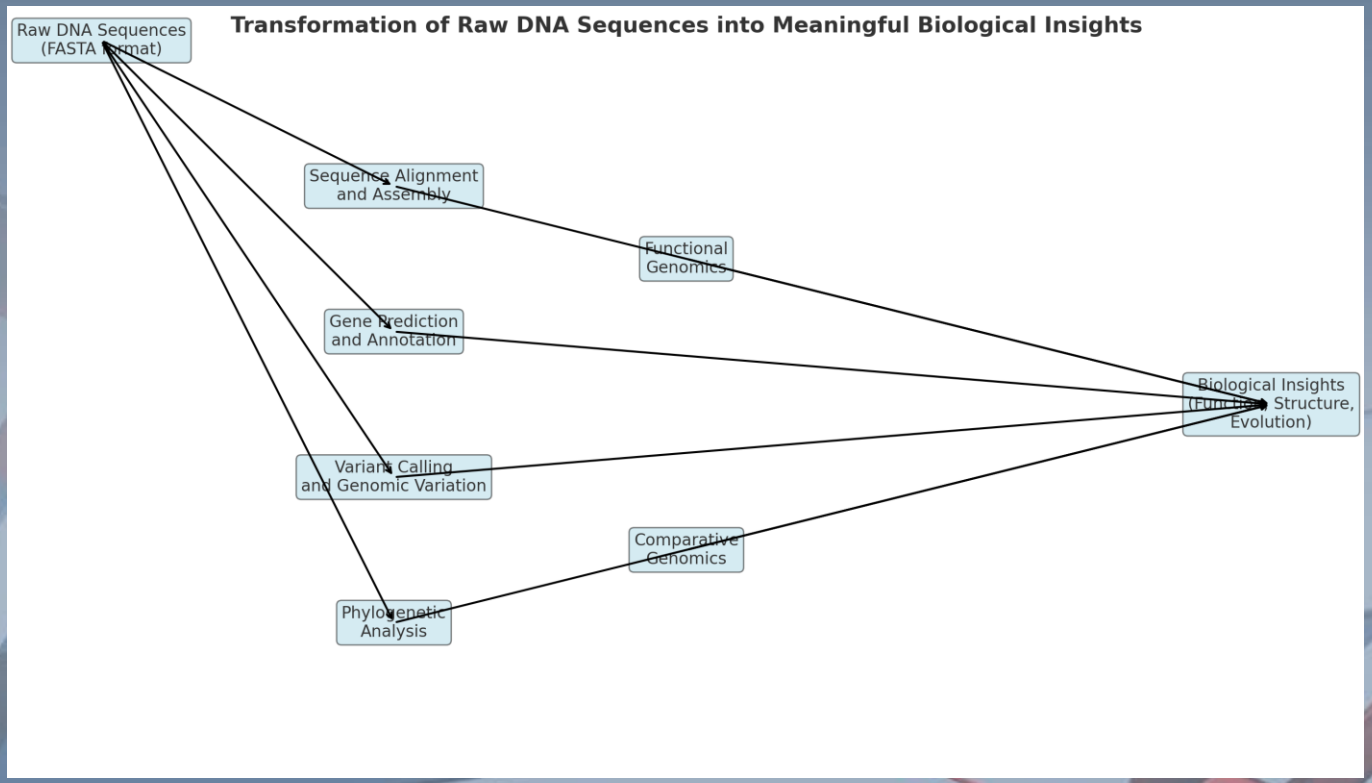
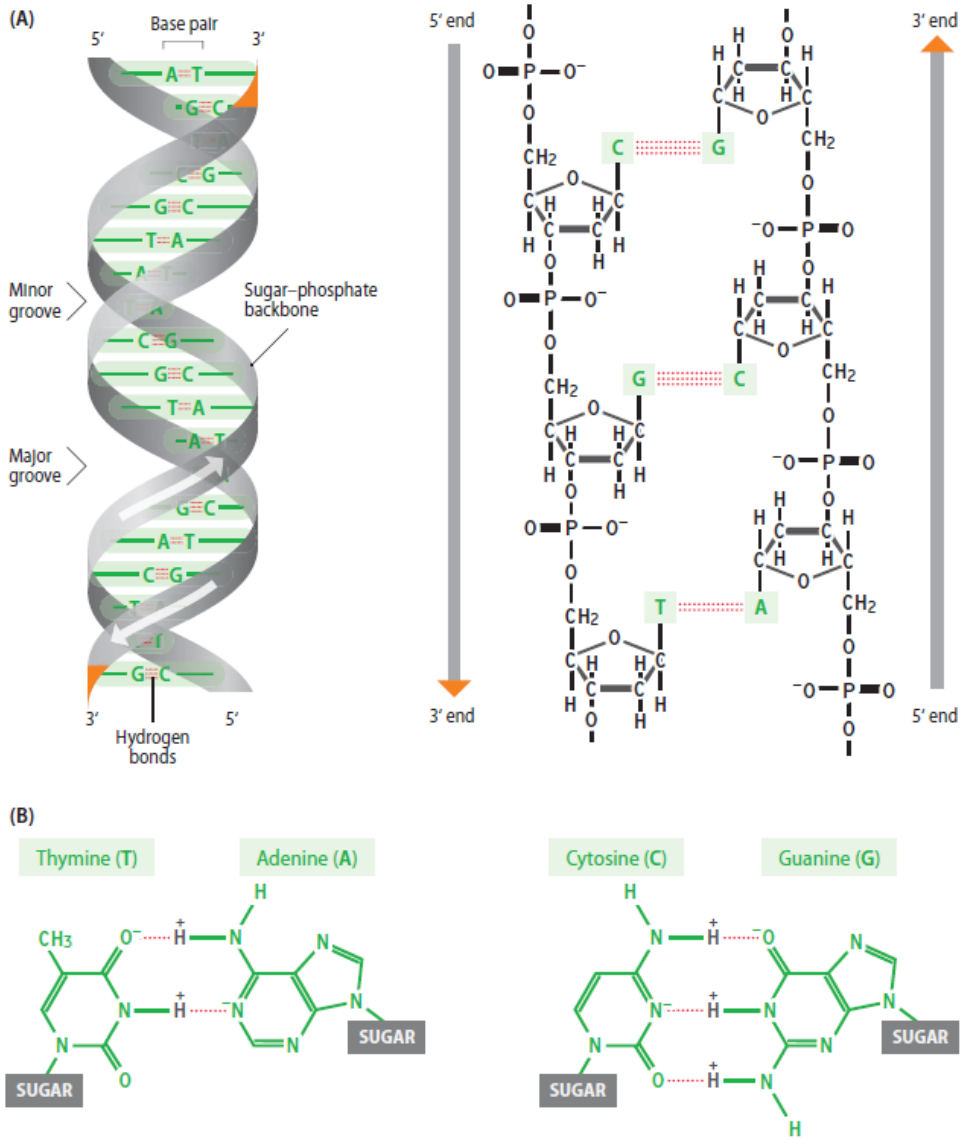
Analiza sekvenci DNA pomoću bioinformatičkih alata



Struktura molekule DNA

- povezana s funkcijom nositelja genetske informacije u živim organizmima



- struktura dvostruke zavojnice
  - dva isprepletena lanca koji se drže zajedno vodikovim vezama između komplementarnih nukleotidnih baza (parovi adenina s timinom i parovi citozina s gvaninom)

- specifično sparivanje baza u DNA (A-T i C-G)
  - ključno je za vjerni prijenos genetskih informacija s jedne generacije na drugu
  - omogućuje točnu replikaciju DNA, osiguravajući da svaka nova stanica dobije identičnu kopiju genetske informacije

- slijed nukleotida duž molekule DNA
  - kodira genetske informacije koje određuju karakteristike organizma
  - geni su specifične sekvence nukleotida koji kodiraju proteine ili funkcionalne RNA molekule, koje su bitne za razne biološke procese

- genetički kod
  - skup pravila prema kojima se informacije kodirane u DNA prevode u proteine
  - redoslijed nukleotida u DNA određuje redoslijed aminokiselina u proteinu, koji zauzvrat određuje strukturu i funkciju proteina

- struktura kromosoma
  - u eukariotskim stanicama DNA je organizirana u kromosome koji se sastoje od dugih niti DNA omotanih oko proteina koji se nazivaju histoni
  - pomaže u zbijanju DNA i reguliranju pristupa genetskim informacijama pohranjenim u molekuli DNA

Genetički kod

- strukturiran na način da svaki od 64 mogućih tripleta RNA (kodona) sastavljenih od kombinacija 4 nukleotida (adenin, uracil, gvanin, citozin) odgovara jednoj od 20 biološki aktivnih aminokiselina, uz dodatna 3 „stop” kodona koji terminiraju translaciju proteina (Crick, 1968.)

UUU Phe	UCU Ser	UAU Tyr	UGU Cys
UUC	UCC	UAC	UGC
UUA Leu	UCA	UAA Stop	UGA Stop
UUG	UCG	UAG	UGG Trp
CUU Leu	CCU Pro	CAU His	CGU Arg
CUC	CCC	CAC	CGC
CUA	CCA	CAA Gln	CGA
CUG	CCG	CAG	CGG
AUU Ile	ACU Thr	AAU Asn	AGU Ser
AUC	ACC	AAC	AGC
AUA	ACA	AAA Lys	AGA Arg
AUG Met	ACG	AAG	AGG
GUU Val	GCU Ala	GAU Asp	GGU Gly
GUC	GCC	GAC	GGC
GUA	GCA	GAA Glu	GGA
GUG	GCG	GAG	GGG

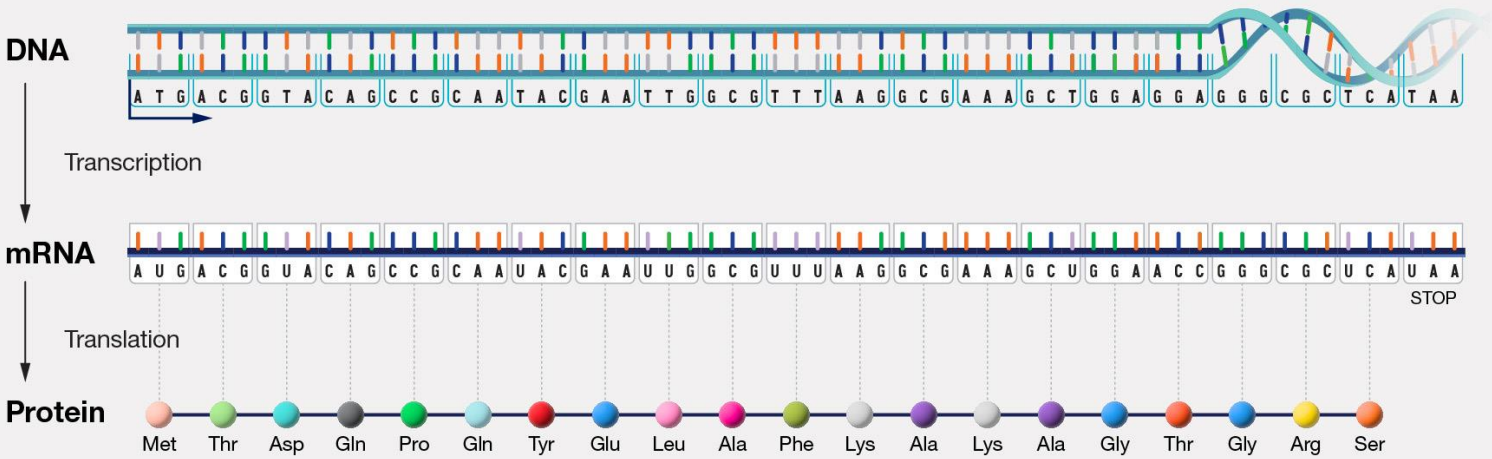
T.A. Brown - Genomes 4-Garland Science (2018)

Središnja dogma molekularne biologije

Francis Crick (1958)

- DNA sadrži upute za stvaranje proteina, koje kopira RNA
- RNA zatim koristi upute za stvaranje proteina

DNA -> RNA -> bjelančevine



https://www.genome.gov/sites/default/files/media/images/2022-05/Central-dogma.jpg

Sekvenciranje DNA

- proces određivanja točnog redoslijeda nukleotida unutar molekule DNA

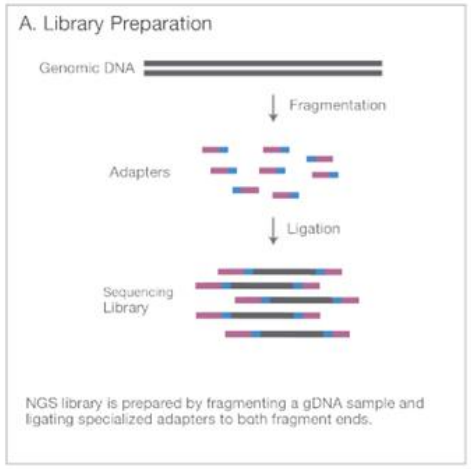
https://www.novogene.com/eu-en/wp-content/uploads/sites/7/2021/02/novogene-blog-ngsbeginnersguide-20210203.jpg

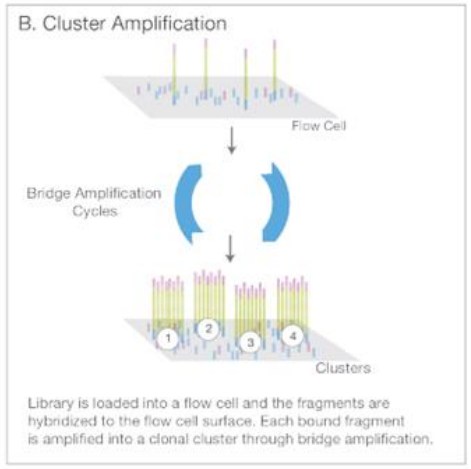
Sekvenciranje sljedeće generacije

engl. *Next Generation Sequencing*, NGS

1. Priprema uzorka

- ekstrakcija DNA iz stanica organizma
- fragmentacija ekstrahirane DNA u manje dijelove, obično duge stotine do tisuće parova baza.
- priprema knjižnice (engl. *library preparation*) - kratke sekvence adaptera dodaju se na krajeve fragmenata DNA kako bi se fragmenti vezali na platformu za sekvenciranje



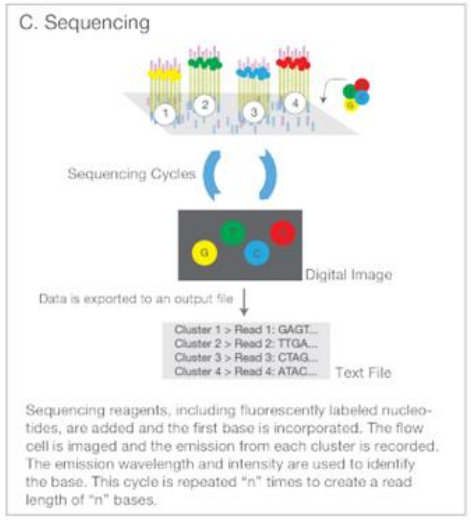


2. stvaranje klastera

- fragmenti DNA pričvršćuju se na protočnu stanicu (engl. *flow-cell*) i umnožavaju dajući nakupine (engl. *clusters*) identičnih sekvenci

3. sekvenciranje sintezom

- nukleotidi se dodaju jedan po jedan, a svaki dodatak detektira se signalom (npr. fluorescencijom) koji identificira nukleotid ugrađen na svakom položaju



4. analiza podataka

- određivanje očitanih baza (engl. *base calling)* - signali otkriveni tijekom sekvenciranja prevode se u niz nukleotida (A, T, C, G)
- sastavljanje očitanih sekvenci (engl. *reads*) - kratke sekvence sastavljaju se u duže sekvence, usklađujući ih s referentnim genomom ako je dostupan ili ih sklapaju *de novo* ako referenca nije dostupna

https://www.illumina.com/content/dam/illuminamarketing/documents/products/illumina_sequencing_introduction.pdf

FASTQ file

- nakon završetka sekvenciranja, slijed nukleotida i ocjena kvalitete svake očitane sekvence (engl. *reads)* pohranjuju se u formatu tekstualne datoteke FASTQ za daljnje korake analize

```
@EAS139:136:FC706VJ:2:5:1000:12850 1:Y:18:ATCACG
ACGTGTACACAGTCAGATGATAGCAGATAGGAAAT
+
BBBBCCCC?<A?BC?7@@??????DBBA@@@@A@
```

- linija 1 počinje sa '@' znakom, a nakon nje je identifikator sekvence i opcionalan opis
  - Specific to Illumina Sequences
  - @EAS139:136:FC706VJ:2:5:1000:12850 1:Y:18:ATCACG - FASTQ Sequence Start
  - @EAS139:136:FC706VJ:2:5:1000:12850 1:Y:18:ATCACG - Instrument Name
  - @EAS139:136:FC706VJ:2:5:1000:12850 1:Y:18:ATCACG - Run ID
  - @EAS139:136:FC706VJ:2:5:1000:12850 1:Y:18:ATCACG - Flowcell ID
  - @EAS139:136:FC706VJ:2:5:1000:12850 1:Y:18:ATCACG - Lane Number
  - @EAS139:136:FC706VJ:2:5:1000:12850 1:Y:18:ATCACG - Tile Number
  - @EAS139:136:FC706VJ:2:5:1000:12850 1:Y:18:ATCACG - X-coordinate of cluster
  - @EAS139:136:FC706VJ:2:5:1000:12850 1:Y:18:ATCACG - Y-coordinate of cluster
  - @EAS139:136:FC706VJ:2:5:1000:12850 1:Y:18:ATCACG - read Number (Paired 1/2)
  - @EAS139:136:FC706VJ:2:5:1000:12850 1:Y:18:ATCACG - Filtered
  - @EAS139:136:FC706VJ:2:5:1000:12850 1:Y:18:ATCACG – Control Number
  - @EAS139:136:FC706VJ:2:5:1000:12850 1:Y:18:ATCACG – Index Sequence

https://www.slideshare.net/rjorton/ngs-data-formats-and-analyses

- linija 2 je sama sekvenca (niz slova)
- linija 3 počinje s '+' znakom, a opcionalno iza njega može biti isti identifikator sekvence (i bilo koji opis) ponovo
- linija 4 predstavlja vrijednosti kvalitete za sekvencu iz linije 2; broj znakova mora biti jednak broju slova u sekvenci

$P = 10^{(-Q/10)}$

Q40 = 1 in 10,000 error rate

$Q = -10 \times \log_{10}(P)$

Q40 is not the maximum quality score

Q-Symbol	Q-ASCII	Q-Score	P-Error
I	73	40	0.00010
H	72	39	0.00013
G	71	38	0.00016
F	70	37	0.00020
E	69	36	0.00025
D	68	35	0.00032
C	67	34	0.00040
B	66	33	0.00050
A	65	32	0.00063
@	64	31	0.00079
?	63	30	0.00100
>	62	29	0.00126
=	61	28	0.00158
<	60	27	0.00200
;	59	26	0.00251
:	58	25	0.00316
9	57	24	0.00398
8	56	23	0.00501
7	55	22	0.00631
6	54	21	0.00794
5	53	20	0.01000
4	52	19	0.01259
3	51	18	0.01585
2	50	17	0.01995
1	49	16	0.02512
0	48	15	0.03162
/	47	14	0.03981
.	46	13	0.05012
-	45	12	0.06310
,	44	11	0.07943
+	43	10	0.10000
*	42	9	0.12589
)	41	8	0.15849
(	40	7	0.19953
'	39	6	0.25119
&	38	5	0.31623
%	37	4	0.39811
$	36	3	0.50119
#	35	2	0.63096
"	34	1	0.79433
!	33	0	1.00000

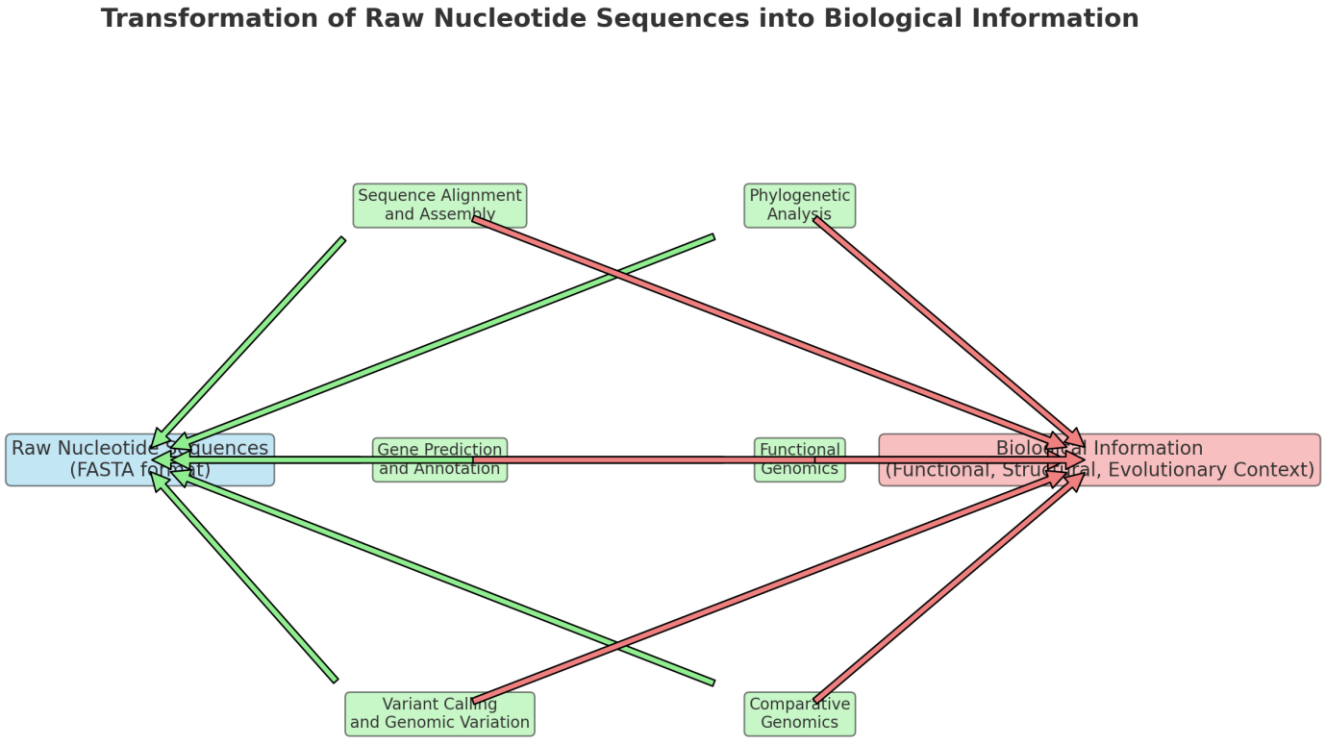
FASTA datoteka

- sadrži retke nukleotidnih sekvenci (adenin [A], timin [T], citozin [C] i gvanin [G]) kojima prethodi redak zaglavlja koji počinje simbolom ">"

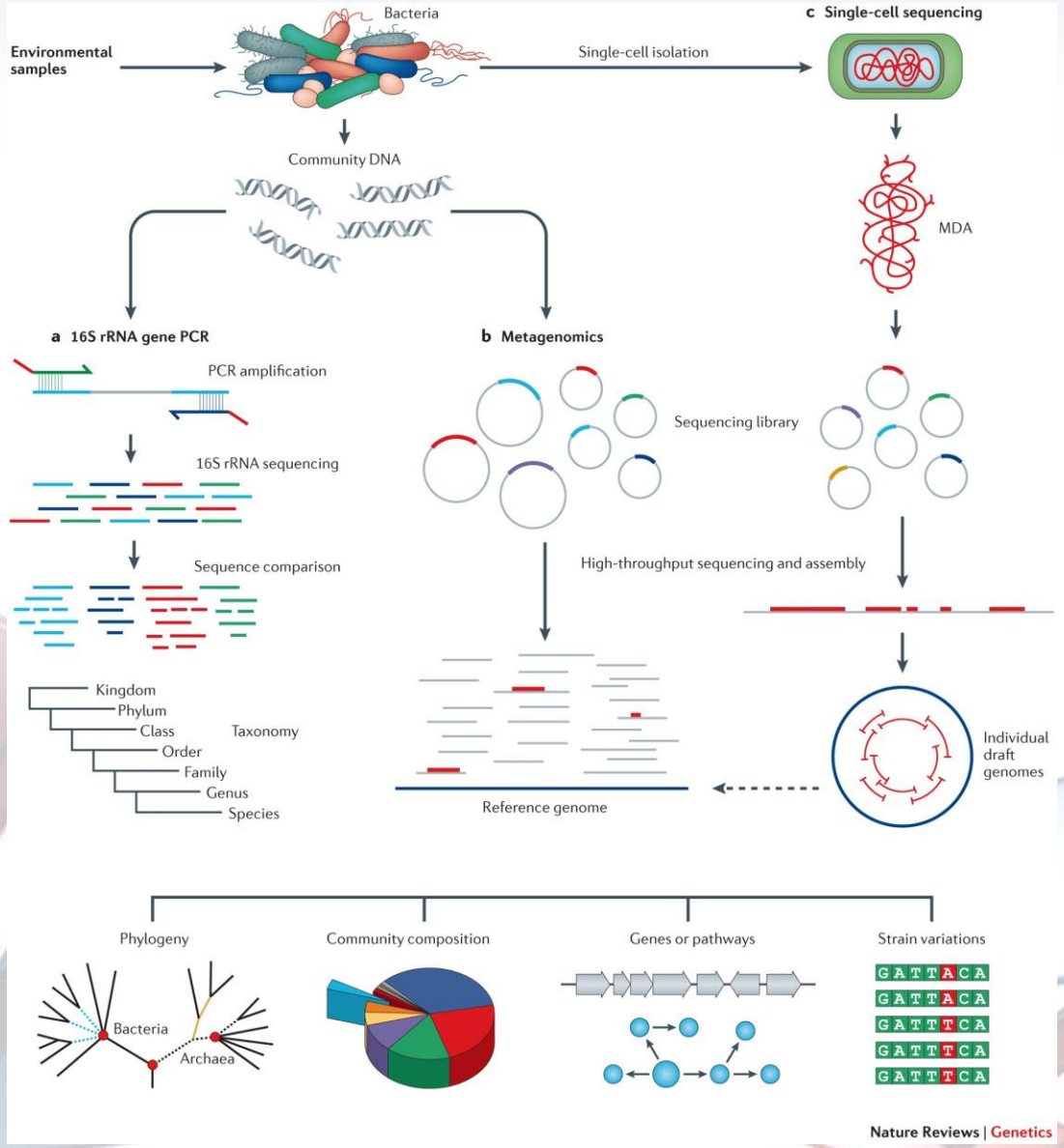
```
>10001f
GTGTGGTAAAAAAGTAGGTAAGAGCAGTTAAAATGCAGAATTCCACGTTCAGTTGACTAAAATTCCTGAGTCAGTCGACTGTAAGTAGAACCCGTGAGTTTTTAACATGCATTCAGTTGAC
>10001r
CTGAGTTAGTCGACTGAATTTCATAGTAAAATCTCTCGGGTTCAGCTTAGTAAACAGGAAAAACTGAGATATTCAATTGAGTTTCATATTAAAAGCTCTCTGTTTCGGCTCAGTCGACTGA
>10002f
TTCCGGGTTGTTACATTCTCCCTTCCTTAAGAGATTTCGTCGTCGAAATCTCTAAGGATAAAAAAAATGAGCTAGATATTTTCATCCAATATCTACACCCAACTACTCAATCTTCAACCATT
>10002r
ATGATGTTTGAAGCATGCATATAGAAGTCATCATTAGGAATATTTGCATTCCTTTGGATACATAGCATGGGCATTATTGCATAAAAGGTGGTCAATTAAATGAGTGGACTGCCTATCTTTT
>10003f
AATAATTTGCTCAATACCGATTTTTCTTAAGAGTTTCAACATTCATTTCTTAAGATTTACTATAAATGGTTACACACTGCTTATACTGATTTTCACTTCTCTATTGAAGTTCATATTCATT
>10003r
AACTTTTGAAATTTTCATTTAACGACAAATAATTTAAATTTGTTTAGCAAATGAAATTGAAATTGGATATTGGAATCTAATTCAAATGTGAATGTAAGGATAAGTGAAGTGAGAAAACTAA
>10004f
AGGATCGAAAGTAAATTGGCAACGGCTGTTCATAAATTAAAATAGAATAATCAACTAAAATTTATAAATTTAGAATTAGAACACAAAAAAGCACAGTACCTTTTCTTTTCAAGTTTCAGGG
>10004r
ATCTTCCCAATGCAAAATCGCGATACAACAAAAGAAAAAAAAACAAACCAAGGCGAAGCACCACCCACCAGCGATCAACGATGCCTCATGACTGATTGGCGGGTAAGCGTATGTGGCTGCG
>10005f
TGGGCCGATAAGTGTGAGGAAGCATTCCAAGGATTGAAACACTTATTGACTAATGCACCTATGTTAGTTGTTCCAGAAGGGAATCTGGACTTAGTAATGGATACAAATGCTAGTGGAACTG
>10005r
AAATCTTTCCACACAACTATCATTGCCAACACTGCATTTTGACATTTTCGGCTAAGTGCATCTGCAACATTATTGTCTTTTCCTGGAGTGTGATCAATACAAAAATCATAAGTGGCCATCA
>10006f
TTTAACTAATTTTATCTCAAAATTTTATGTGTATACATTTTTATATCATGGTTTATATCTTTATACTATATTATATACACATATATGTATGTATGTATGTATTTCAATAATTTGATTTTTA
>10006r
CATACATATATATTATTTTAAATCGAAAAATAAATATTTTTATTACAAAAACAATAGTACATACACATACACATTTTGATATATATATATATATATATTATGTGTGTGGTTACTCATTATT
>10007f
TTCGGATAGCCTTTTTTTGGGTGTGAAACTATCCGTTGATGGCTTACAGCGGTCAATGCCGTAGAACTCACTGAGACCTTCGTTTTGACTGGTTATGGCTCCCGTAACTCTAAAAGGGGCA
>10007r
GTTACGGGAGCCATAACCAGTCAAAACGAAGGTCTCAGTGAGTTCTACGGCATTGACCGCTGTAAGCCATCAACGGATAGTTTCACACCCAAAAAAGGGCTATCCGAACCCCAATCGGGTC
>10008f
TGACGAGGCCGGTAGCGGCGTGTAGTGGCGGCCGTGGTTGGTCAAACGTTGATGAAGTGGGTTCGGTTTTGTTCGCTGAAGAGTGTTGCGTGATGCTATAACTACTGTTCTTGTGAGGGAG
>10008r
ATTTAAATCGAATTAAAAATTCAATTTAAACCCCAAAATTCGACGGTCTCAGAATCCGTCCAAACCAGCATCATCAAGTCCAGAATTTCAATACGAACCTAGCGGAGACAACCGCGTGTAA
>10009f
GGCTACGCCGATGATATAAGGCAATTTTATTTATTCATCACTTTGGTGCTGATATTCTCTTTGGAGATTTTTACTGTTGATTGATGATTAGGGATCGTCATTTTATTCATGTGGTCCAAAA
>10009r
CAAAGCCACTGCCTACACATGTGAGTCAAAAATATGTACATATCAGAGTCTGAATAATAACAGAGTAGCGGAATAGTGGACAAATAAAAGCGCTATCTATCAGCATAGCAGATAATGTCCA
>10010f
AGGTTACTTTACAAAATATAAAACTAAATTTCCTCTTTTAAAACAAAATTCCAAGTCACCTTAAAATGAGTACTATTATAATAAATTTGAAAAAGAATTACTTTAAAATTTAAACATTTAA
>10010r
```

Bioinformatika

- integrira biologiju, računalnu znanost i informacijsku tehnologiju za analizu i interpretaciju bioloških podataka
- transformira neobrađene nukleotidne sekvence (engl. *raw data*) u složenu kompoziciju bioloških informacija, omogućujući znanstvenicima da dešifriraju funkcionalni, strukturni i evolucijski kontekst genoma
- kroz sofisticirane algoritme i računalne alate, pretvara sekvence u smisleni prikaz uputa za rast, razvoj, preživljavanje i reprodukciju organizma



Uloga bioinformatike u analizi sekvenci DNA i sastavljanju genoma

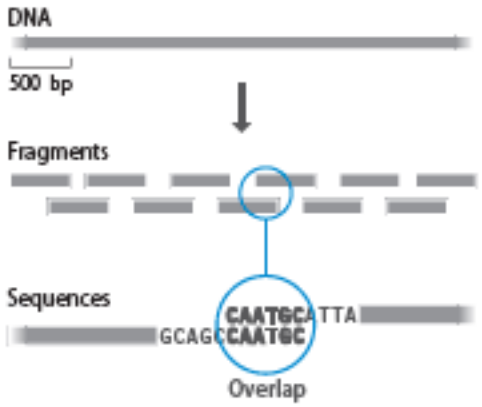


https://media.springernature.com/full/springer-static/image/art%3A10.1038%2Fnrg3785/MediaObjects/41576_2014_Article_BFnrg3785_Fig1_HTML.jpg

Sklapanje genoma (engl. *sequence assembly*)

Kako se mnoštvo kratkih očitanih sekvenci generiranih metodama NGS sastavlja u sekvencu cijelog genoma?

- kontinuirana sekvenca genoma se sastavlja traženjem preklapanja između slijeda nukleotida zbirke pojedinačnih fragmenata
- proces se provodi pomoću specijaliziranog softvera i algoritama koji analiziraju preklapajuća područja sekvenci kako bi ih poravnali i spojili u veće susjedne sekvence, poznate kao kontig sekvence (engl. *contig*) - djelomično sastavljen niz od nekoliko očitanih sekvenci



T.A. Brown - Genomes 4-Garland Science (2018)

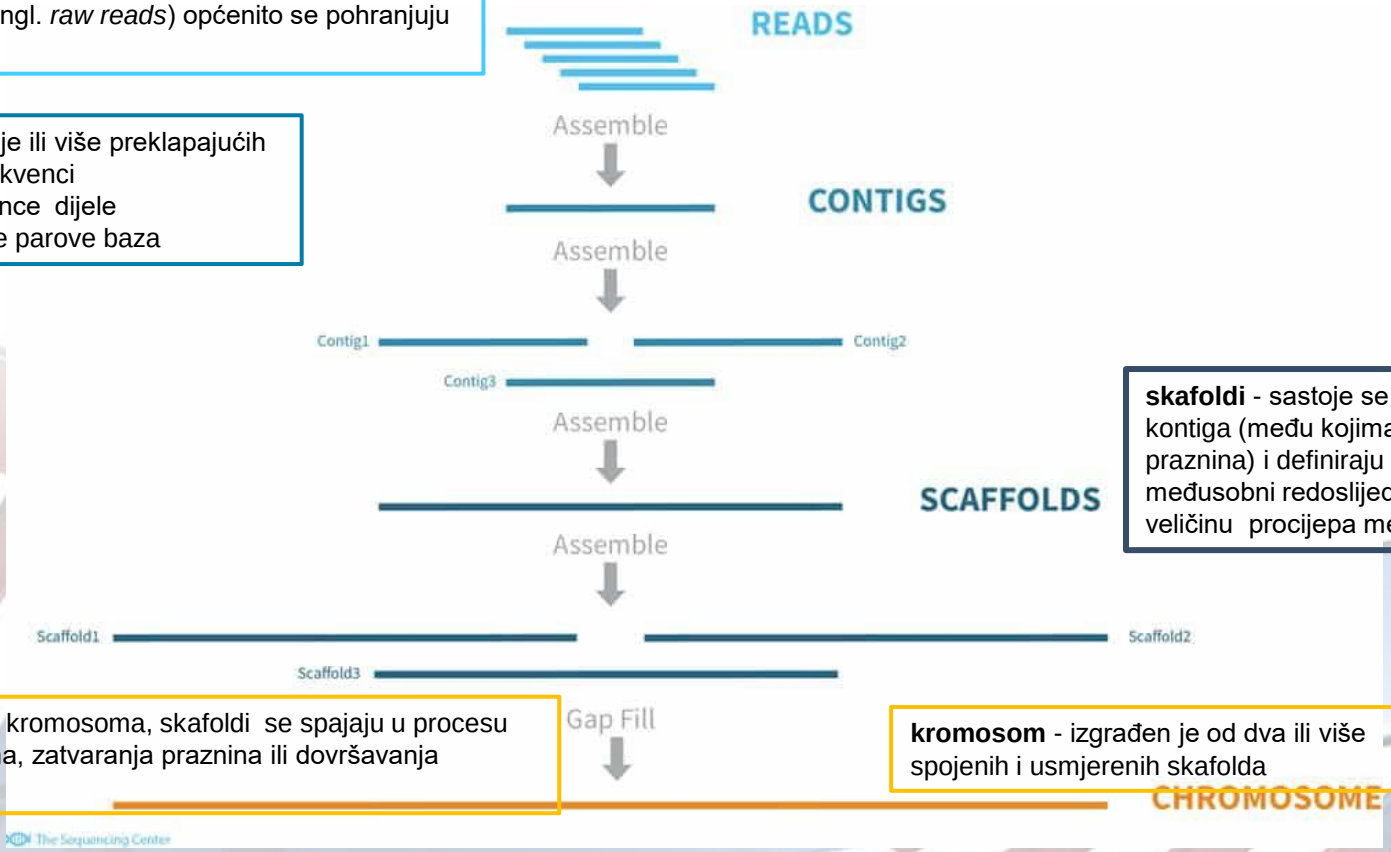
- konsenzus sekvenca - reprezentativni slijed nukleotida u kojem je svaki nukleotid onaj koja se najčešće pojavljuje na tom mjestu u različitim sekvencama

de novo sklapanje genoma

- metoda za konstruiranje genoma iz velikog broja (kratkih ili dugih) fragmenata DNA, bez prethodnog znanja o ispravnom slijedu ili redoslijedu tih fragmenata

očitane sekvence DNA (engl. *reads*)
- fragmenti DNA koja generiraju sekvenatori - obično se kreću u veličini od 35 do 1000 bp (˝kratka čitanja˝) ili "duga čitanja" u veličini od 1000 do 500 000 bp
- neobrađena očitanja (engl. *raw reads*) općenito se pohranjuju u FastQ datoteke

kontigi - građeni od dvije ili više preklapajućih usmjerenih sekvenci
- očitane sekvence dijele podskup ili sve parove baza



skafoldi - sastoje se od više kontiga (među kojima često ima praznina) i definiraju njihov međusobni redoslijed, orijentaciju i veličinu procijepa među njima

- u koraku sastavljanja kromosoma, skafoldi se spajaju u procesu popunjavanja praznina, zatvaranja praznina ili dovršavanja genoma

kromosom - izgrađen je od dva ili više spojenih i usmjerenih skafolda

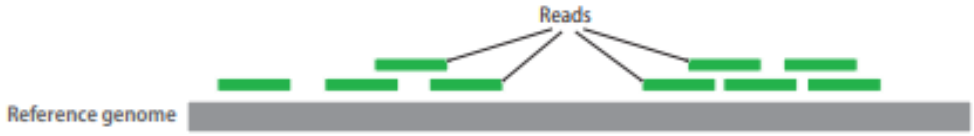
https://thesequencingcenter.com/knowledge-base/de-novo-assembly/

Sastavljanje genoma *de novo*

- metoda za konstruiranje genoma iz velikog broja (kratkih ili dugih) fragmenata DNA, bez prethodnog znanja o ispravnom slijedu ili redoslijedu tih fragmenata

- odnosi se na sekvenciranje novog genoma gdje ne postoji referentna sekvenca dostupna za poravnanje

- pristup referentnog genoma se može koristiti za sekvenciranje *de novo* ako je vrsta čiji se genom sekvencira povezana s drugim vrstama čiji su genomi već sastavljeni
- postojeća sekvenca koristi se kao referentni genom za sastavljanje sekvenci tražeći regije identičnosti ili sličnosti sekvenci
- kratke sekvence novog genoma bit će sastavljenu u kontige prije usporedbe s referencom, kako bi se povećao stupanj točnosti na temelju preklapanja



T.A. Brown - Genomes 4-Garland Science (2018)