

Biologinių sistemų teorija

Kotryna Kvederavičiūtė
VU GMC DMTS
kotryna.kvederaviciute@gmail.com

Biologinių sistemų teorija

- Kurso formatas:
 - Teorinės paskaitos ir pratybos. Griežto atskyrimo nebus. Pirmos 2 savaitės bus gryna teorija.
- Kurso pobūdis:
 - Praktinis, t.y. reikės daugiau daryti, o ne tiesiog skaityti/klausyti.

Biologinių sistemų teorija

- Kurso temos:

- Sekoskaitos duomenų kilmė ir savybės
- Transkriptomika
- Genomo surinkimas ir sekų analizė

Biologinių sistemų teorija

- Kurso darbai:
 - Praktikos darbas iš RNA-seq analizės ir duomenų kokybės
 - Praktikos darbas iš genomo surinkimo ir sekų analizės
 - Savaitiniai testukai (testukų bus mažiau nei savaičių)
 - Egzaminas

Biologinių sistemų teorija

- Galutinio balo sandara:
 - Savaitiniai_testukai*1.5 + Pirmas_ND * 2.25 + Antras_ND * 2.25 + Egzaminas * 4 =10
- Formalūs reikalavimai egzaminui:
 - Atsiskaityti visus praktikos darbus (visi savaitiniai testai kartu yra vienas darbas) ir rašyti visus atsiskaitymus/egzaminą.
 - Atsiskaityti visus praktikos darbus bei egzaminą ne bet kaip, o surinkus ne mažiau kaip 40% maksimalaus įvertinimo iš kiekvieno (savaitiniai testai vertinami visi kartu).
 - Neatsiskaičius/nesurinkus reikiama balo iš tik vieno darbo (pirmas ND, antras ND) - galima pasitaisyti per pakartotinę sesiją. Neatsiskaičius/nesurinkus balų iš daugiau nei vieno darbo - kursas kartojamas.

Biologinių sistemų teorija

- Kurso prerekvizitai:

- Studentai turi orientuotis UNIX komandinėje eilutėje (bus dirbama su ja).
 - Studentai turi turėti bazines biologijos žinias (kas yra genai, genomai, transkriptai, genų raiška ir pan.)
 - P. s. dalis kurso medžiagos ir užduočių bus pateikiama anglų kalba.

Dėl literatūros

- Vieno vadovėlio nebus.
- Įvairioms temoms bus pateikiama (paprastai vienas ar keli iš nurodytų tipų):
 - Straipsniai
 - Video medžiaga
 - Lietuviški/angliški aprašai

Biologinių sistemų teorija

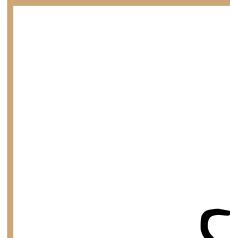
- Klausimai :)



Sequencing technologies

Topics

- Sequencing technologies
- Data formats

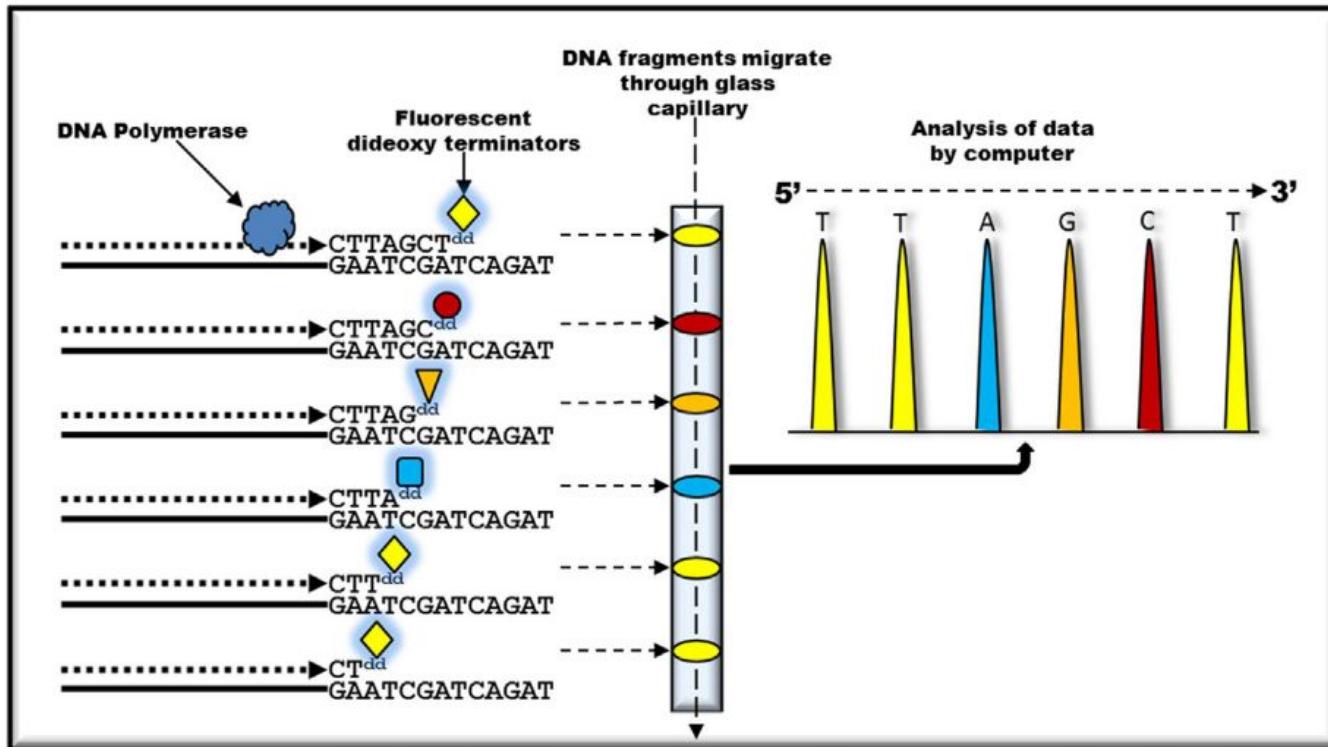


Sequencing technologies

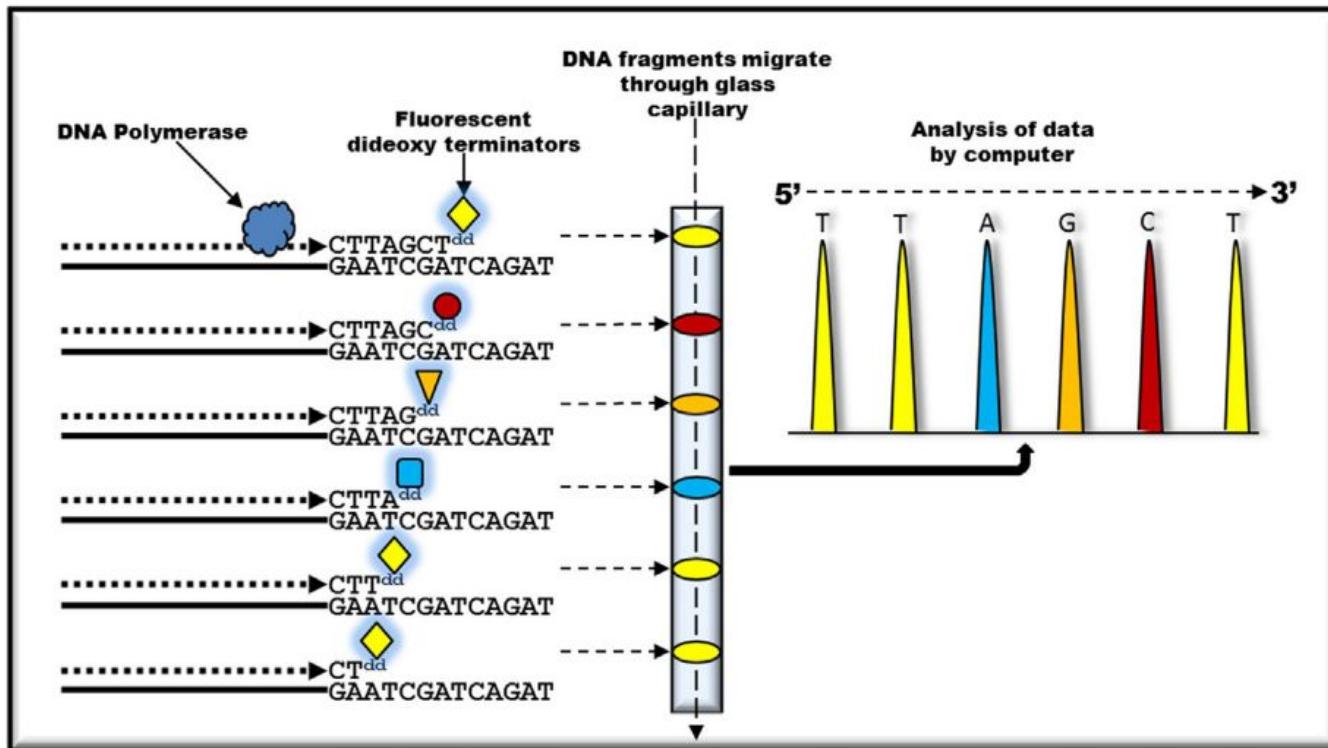
Sequencing

- DNA sequencing is the process of determining the sequence of nucleotide bases in a piece of DNA.
- Sequencing short fragments – simple.
- Getting full genome – complicated.

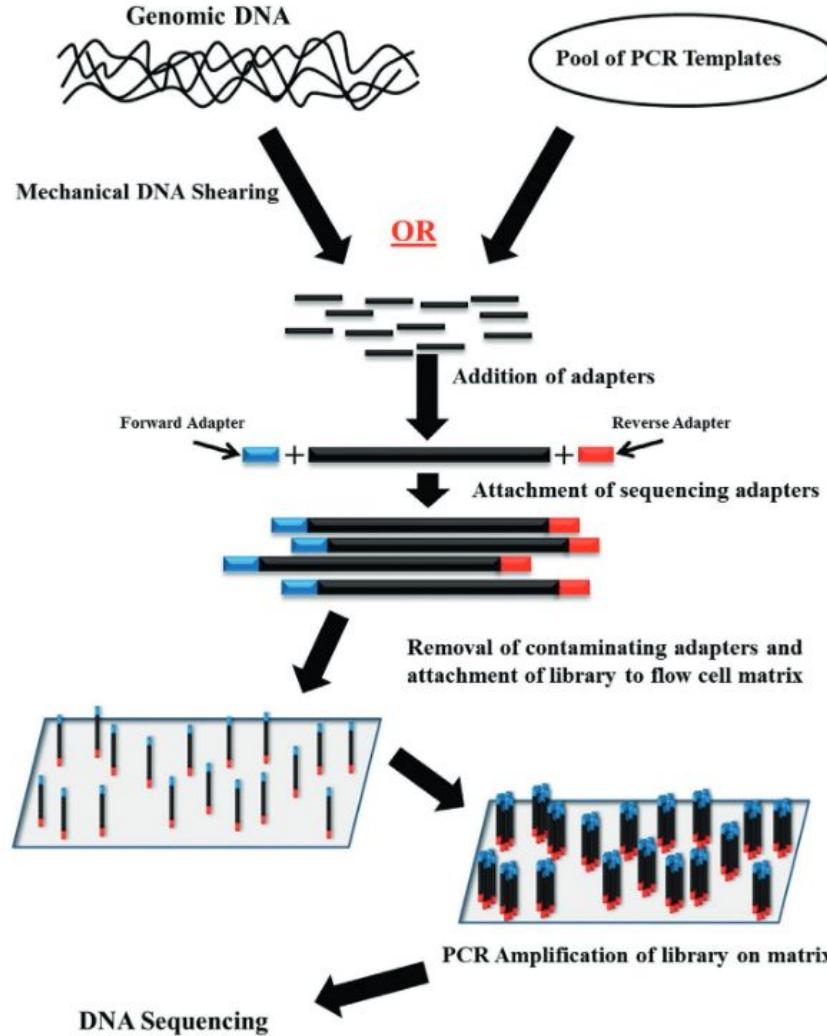
1st generations sequencing (Sanger method)



1st generations sequencing (Sanger method)



2nd generation sequencing (NGS)



2nd generation
sequencing (NGS)

2nd generation sequencing (NGS)

Single Index

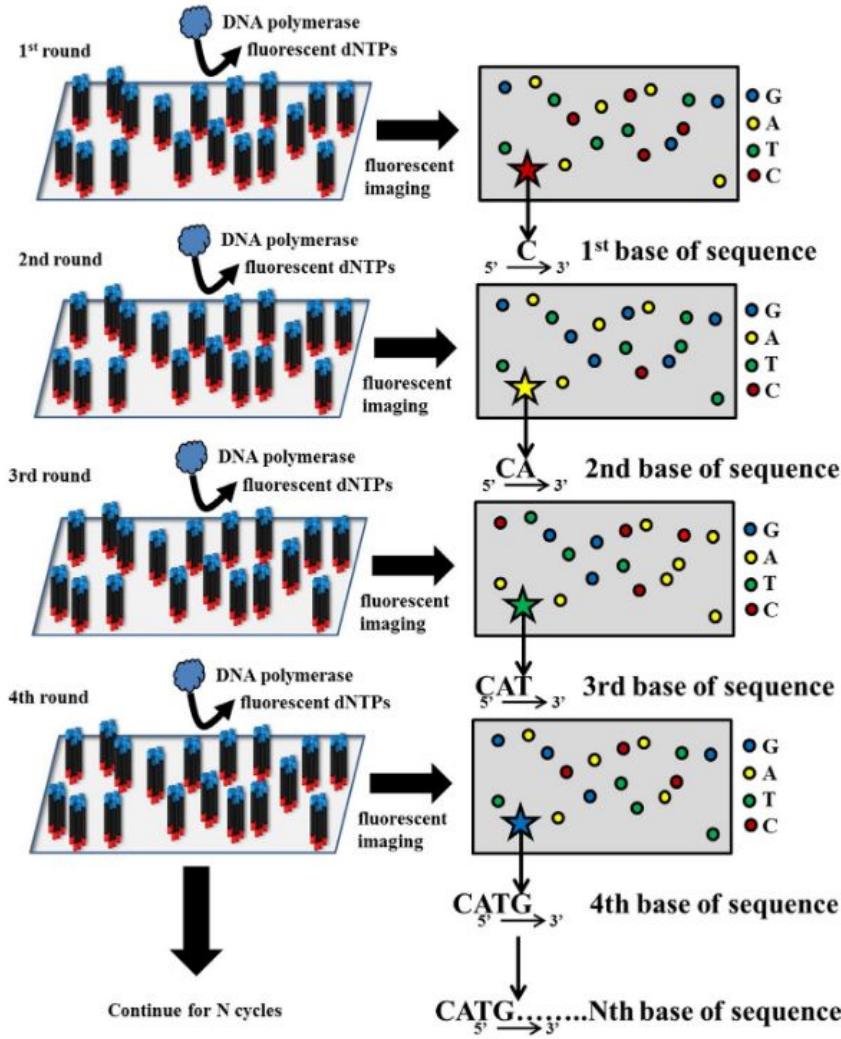


* Up to 12 forward indices allows 12 NGS libraries to be pooled and de-multiplexed

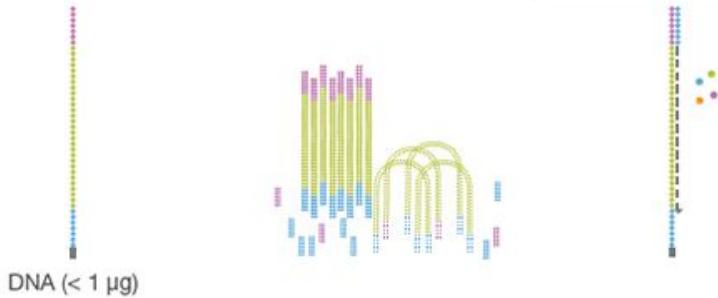
Dual Indices



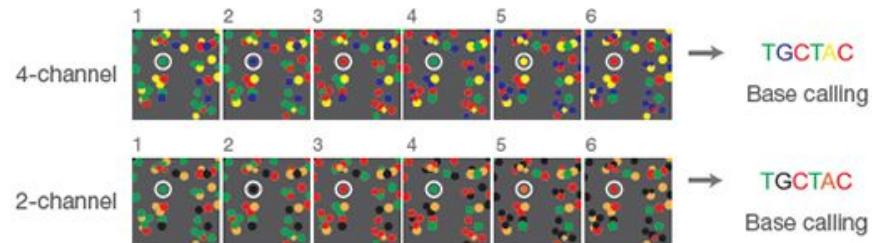
* Up to 12 forward indices and 8 reverse indices allows up to 96 NGS libraries to be pooled and de-multiplexed



2nd generation sequencing (NGS)



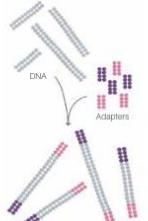
Imaging



2nd generation
sequencing (NGS)

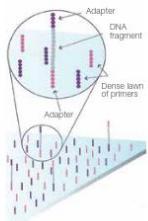
Illumina and bridge amplification

Figure 2: Prepare Genomic DNA Sample



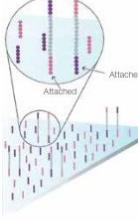
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.

Figure 3: Attach DNA to Surface



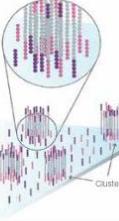
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

Figure 6: Denature the Double-Stranded Molecules



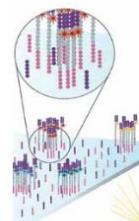
Denaturation leaves single-stranded templates anchored to the substrate.

Figure 7: Complete Amplification



Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

Figure 10: Determine Second Base



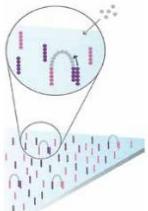
The next cycle repeats the incorporation of four labeled reversible terminators, primers, and DNA polymerase.

Figure 11: Image Second Chemistry Cycle



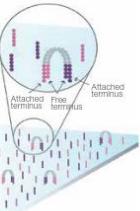
After laser excitation, the image is captured as before, and the identity of the second base is recorded.

Figure 4: Bridge Amplification



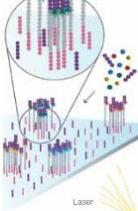
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

Figure 5: Fragments Become Double Stranded



The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

Figure 8: Determine First Base



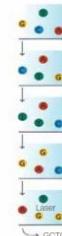
The first sequencing cycle begins by adding four labeled reversible terminators, primers, and DNA polymerase.

Figure 9: Image First Base



After laser excitation, the emitted fluorescence from each cluster is captured and the first base is identified.

Figure 12: Sequencing Over Multiple Chemistry Cycles



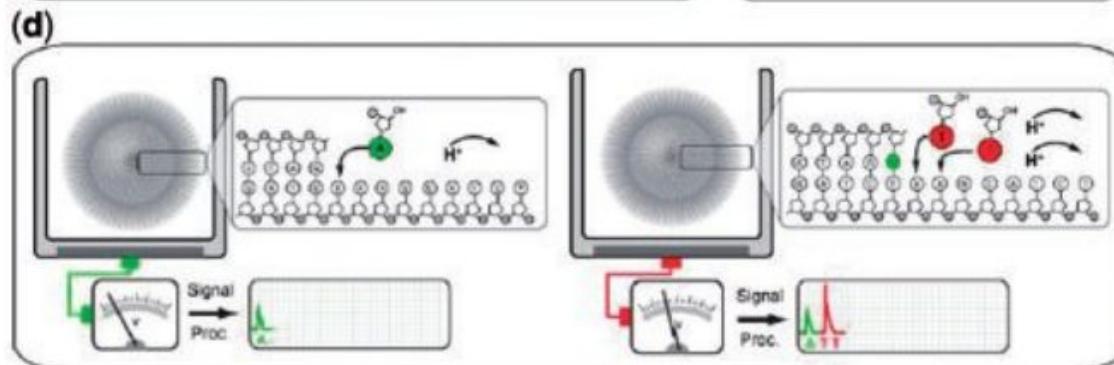
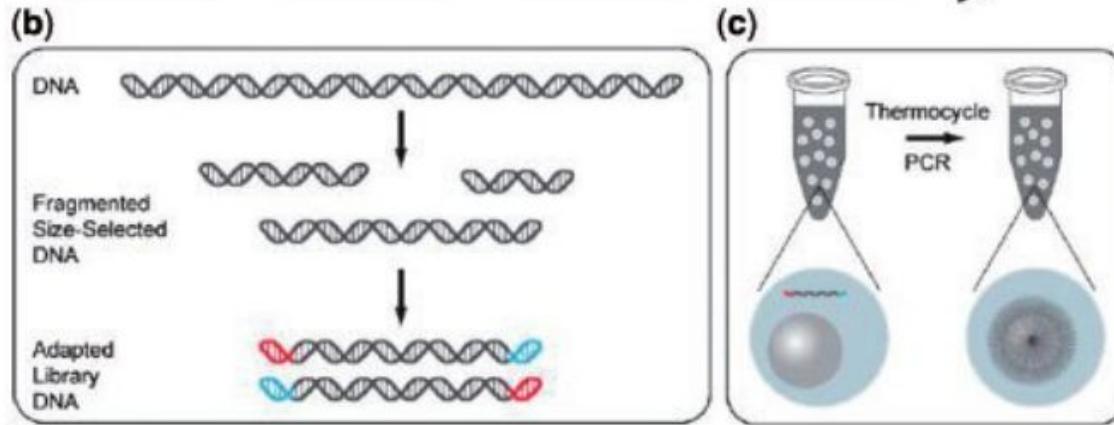
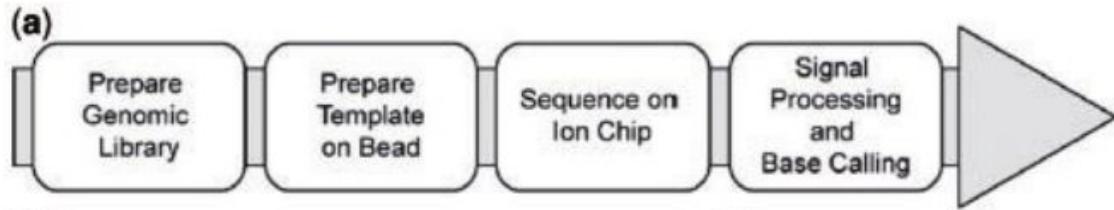
The sequencing cycles are repeated to determine the sequence of bases in a fragment, one base at a time.



The data are aligned and compared to a reference, and sequencing differences are identified.

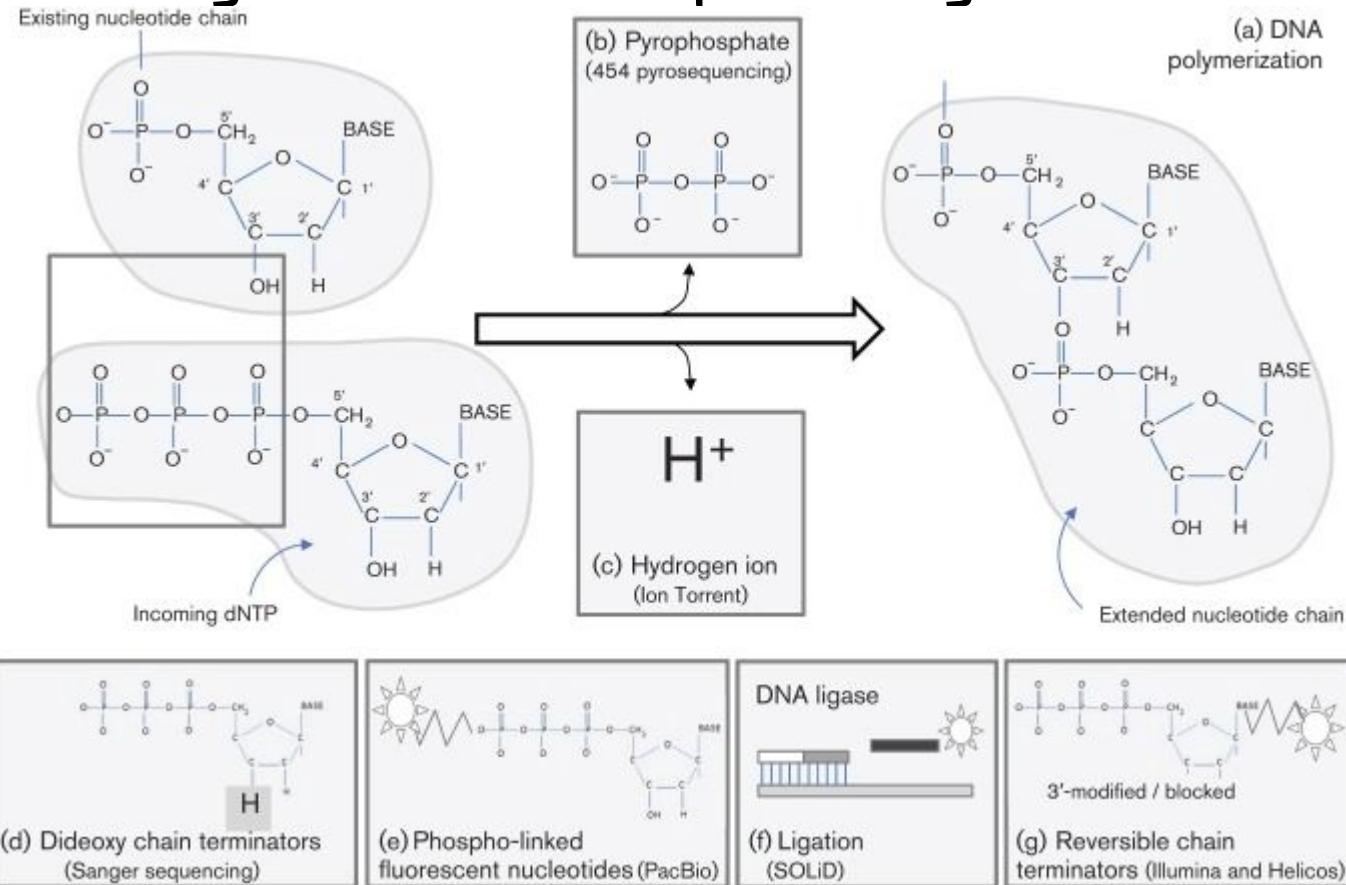
Source Illumina

Ion Torrent



SOURCE

2nd generation sequencing (NGS)



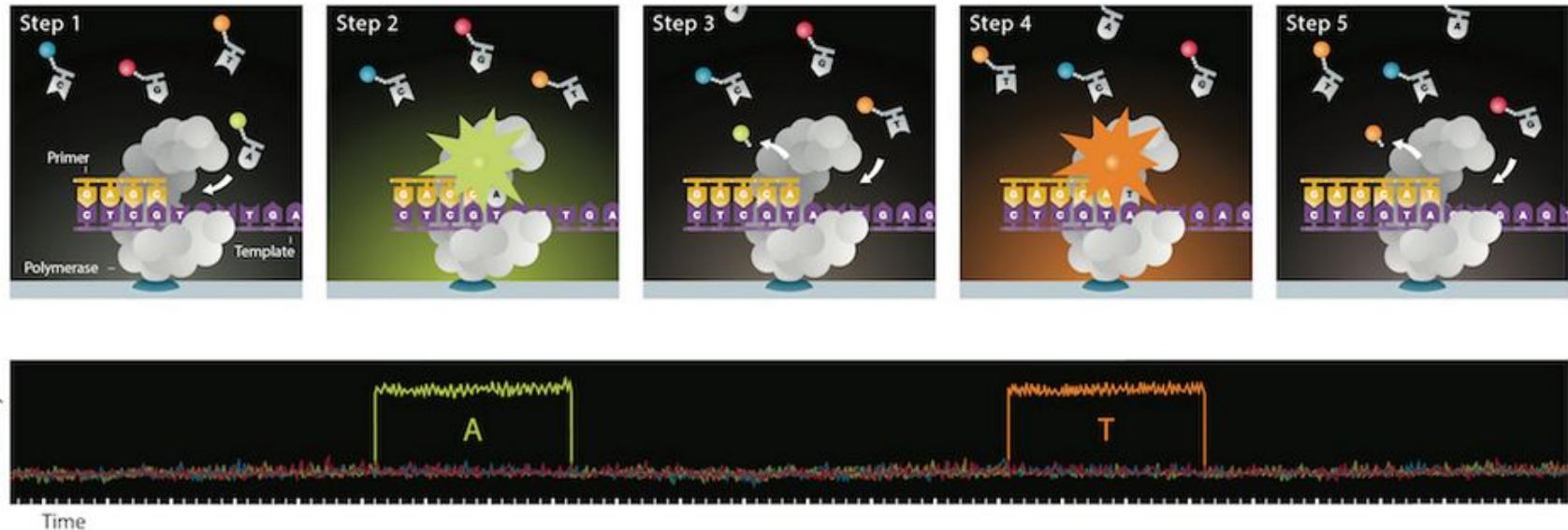
2nd Generations sequencing

Features:

- **Highly parallel:** many sequencing reactions take place at the same time
- **Micro scale:** reactions are tiny and many can be done at once on a chip
- **Fast:** because reactions are done in parallel, results are ready much faster
- **Low-cost:** sequencing a genome is cheaper than with Sanger sequencing
- **Shorter length:** reads typically range from 50-700 nucleotides in length

3^d Generation sequencing

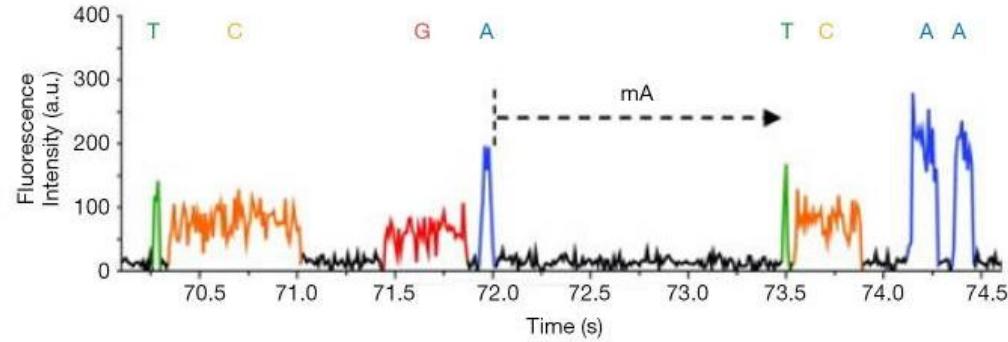
Pacific Bioscience



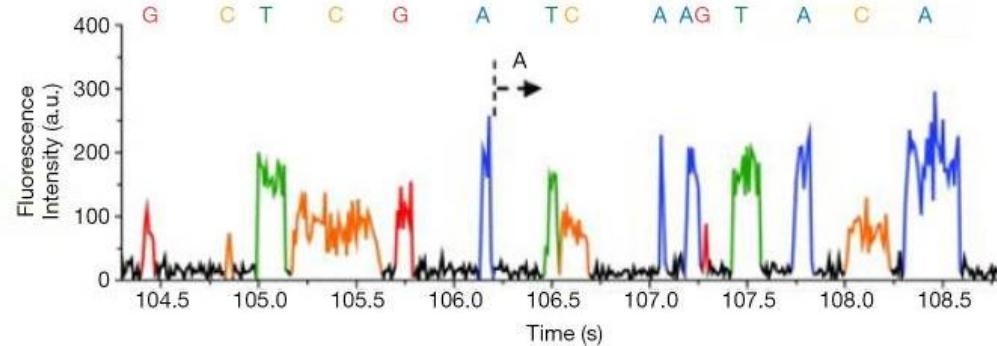
SOURCE

Pacific Bioscience: modifications

A



B



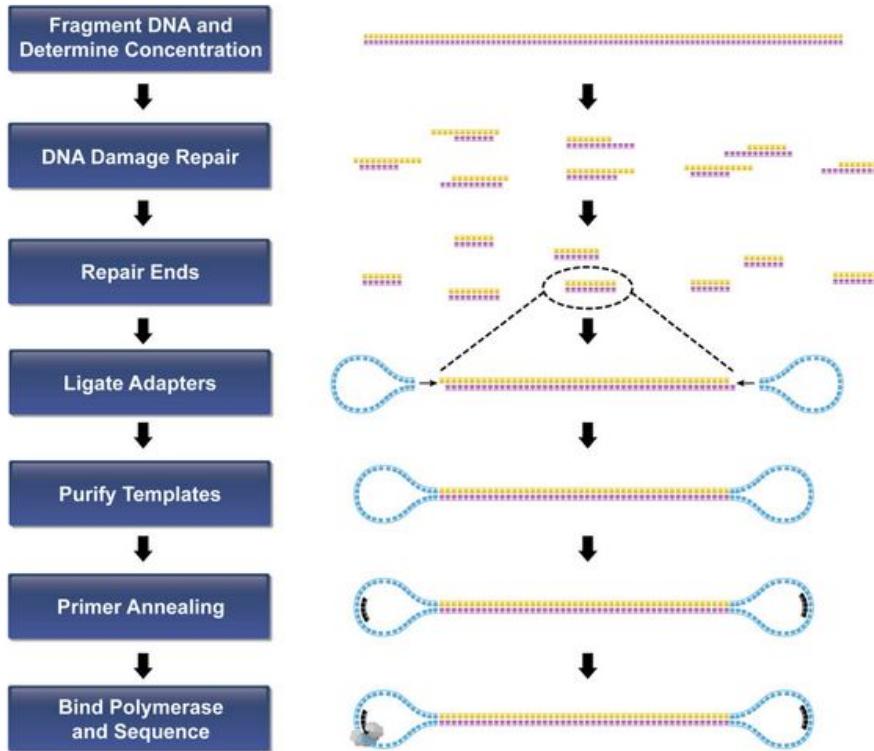
Source

Pacific Bioscience

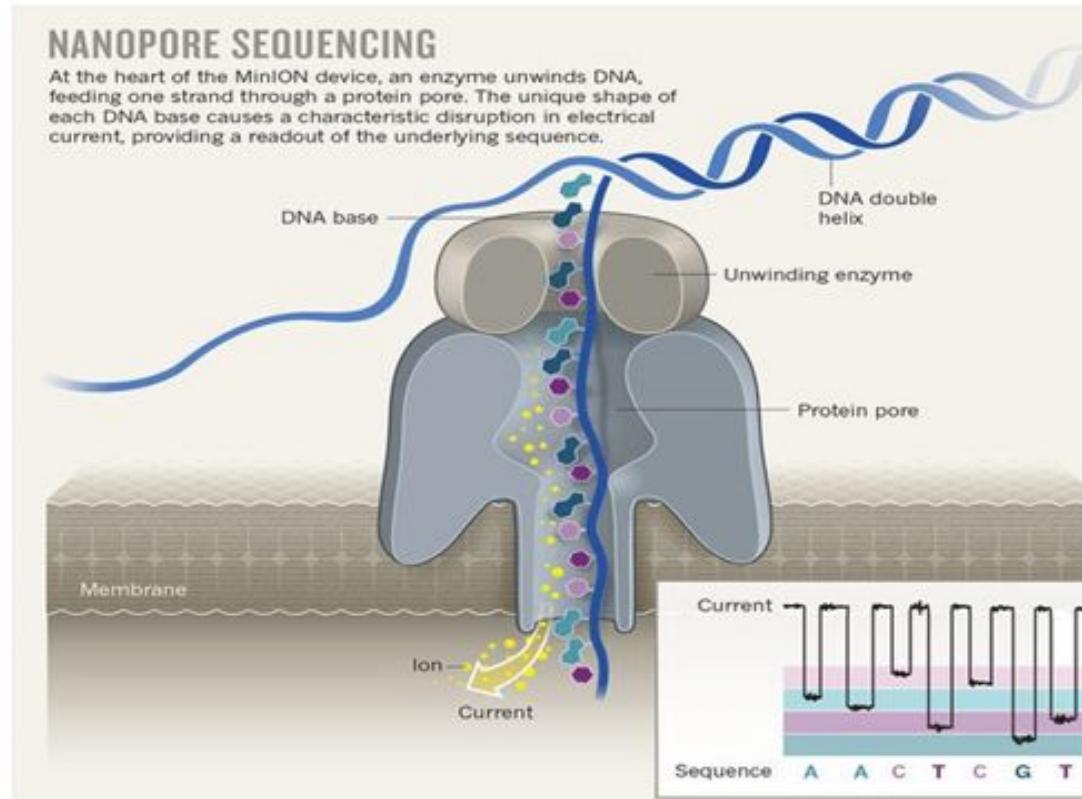


[Source](#)

Pacific Bioscience



MinION sequencing



Source

3^d Generation sequencing

- Single molecule sequencing
- No amplification
- Long reads/full length transcripts
- Real-time
- Portability
- Price :(

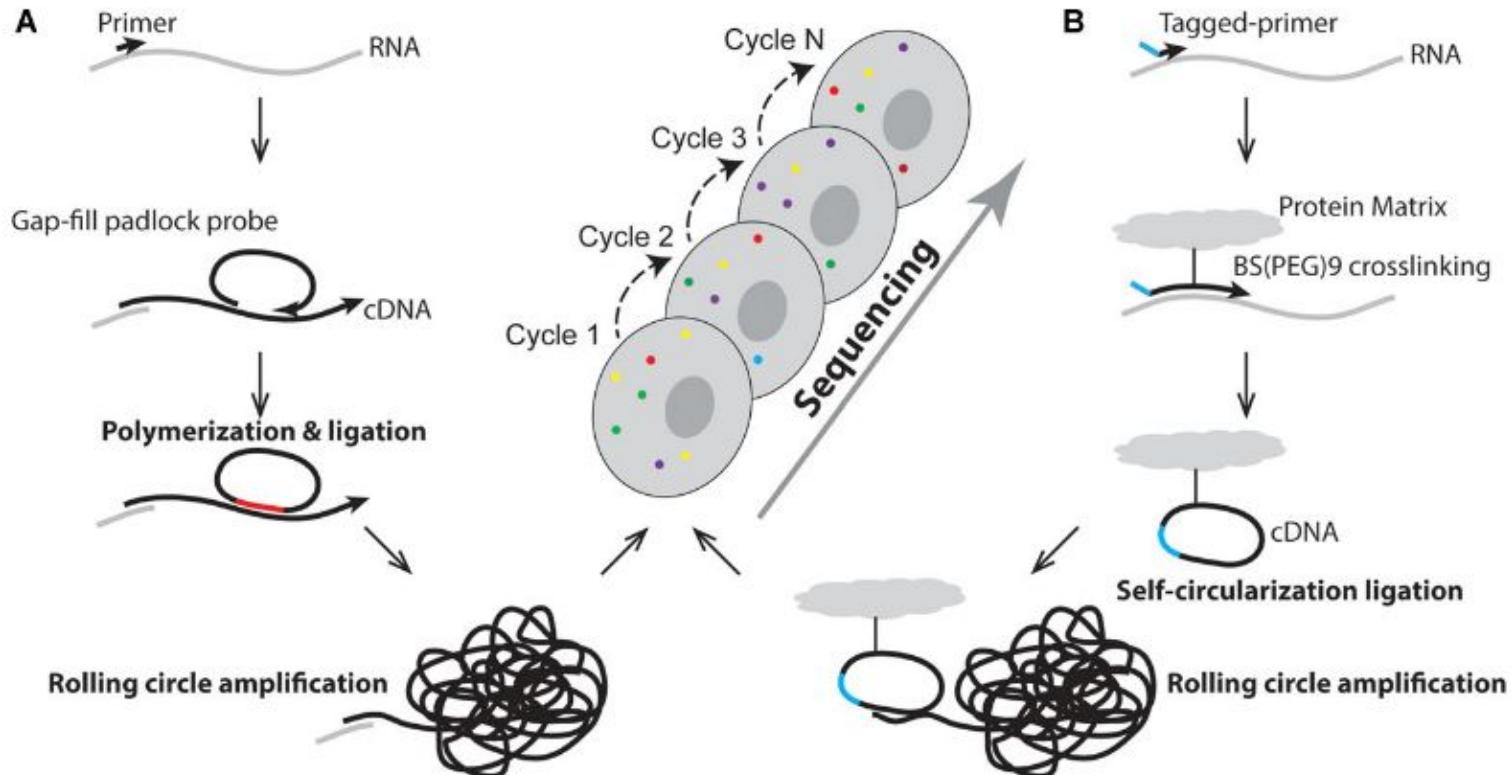
Small comparison



	Illumina (Hiseq 4000)	PacBio (Sequel)	Oxford Nanopore (MinION)
Read length	Up to 150 bp	10-15kb	Up to 900kb
Number of reads	2.5-5 Million	500 K	Up to 1 M
Processing time	<1-3.5 days	Up to 10 hours	~ 6 hours
Error rate	<1%	10-15%	5-15%
Cost per run	~\$3000	~\$850	\$500-\$900
Instrument price	\$900 K	\$350K	\$1K
Advantages	Highly accurate	Sequence long reads	Sequence long reads Portable device

[Source](#)

Fourth generation



Source

Some video (for better understanding)

- Sanger
- Illumina
- Ion Torrent
- Pacific Bioscience
- Oxford Nanopore
- For general understanding

File formats

File formats

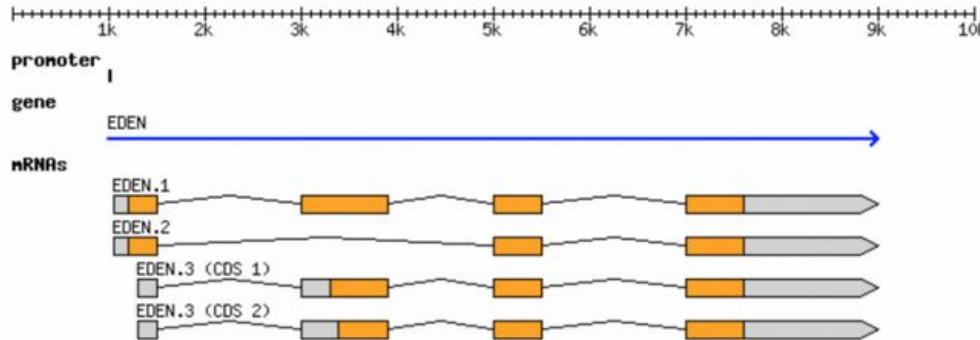
- FASTA
- FASTQ
- SAM/BAM
- GFF3
- GTF
- BED
- VCF
- etc.

Count matrix

GENE ID	KD.2	KD.3	OE.1	OE.2	OE.3	IR.1	IR.2	IR.3
1/2-SBSRNA4	57	41	64	55	38	45	31	39
A1BG	71	40	100	81	41	77	58	40
A1BG-AS1	256	177	220	189	107	213	172	126
A1CF	0	1	1	0	0	0	0	0
A2LD1	146	81	138	125	52	91	80	50
A2M	10	9	2	5	2	9	8	4
A2ML1	3	2	6	5	2	2	1	0
A2MP1	0	0	2	1	3	0	2	1
A4GALT	56	37	107	118	65	49	52	37
A4GNT	0	0	0	0	1	0	0	0
AA06	0	0	0	0	0	0	0	0
AAA1	0	0	1	0	0	0	0	0
AAAS	2288	1363	1753	1727	835	1672	1389	1121
AACS	1586	923	951	967	484	938	771	635
AACSP1	1	1	3	0	1	1	1	3
AADAC	0	0	0	0	0	0	0	0
AADACL2	0	0	0	0	0	0	0	0
AADACL3	0	0	0	0	0	0	0	0
AADACL4	0	0	1	1	0	0	0	0
AADAT	856	539	593	576	359	567	521	416
AAGAB	4648	2550	2648	2356	1481	3265	2790	2118
AAK1	2310	1384	1869	1602	980	1675	1614	1108
AAMP	5198	3081	3179	3137	1721	4061	3304	2623
AANAT	7	7	12	12	4	6	2	7
AARS	5570	3323	4782	4580	2473	3953	3339	2666
-----	-----	-----	-----	-----	-----	-----	-----	-----

[Source](#)

GFF3 file format



```
0 ##gff-version 3.2.1
1 ##sequence-region ctg123 1 1497228
2 ctg123 . gene      1000 9000 . + . ID=gene00001;Name=EDEN
3 ctg123 . TF_binding_site 1000 1012 . + . ID=tfbs00001;Parent=gene00001
4 ctg123 . mRNA      1050 9000 . + . ID=mRNA00001;Parent=gene00001;Name=EDEN.1
5 ctg123 . mRNA      1050 9000 . + . ID=mRNA00002;Parent=gene00001;Name=EDEN.2
6 ctg123 . mRNA      1300 9000 . + . ID=mRNA00003;Parent=gene00001;Name=EDEN.3
7 ctg123 . exon      1300 1500 . + . ID=exon00001;Parent=mRNA00003
8 ctg123 . exon      1050 1500 . + . ID=exon00002;Parent=mRNA00001,mRNA00002
9 ctg123 . exon      3000 3902 . + . ID=exon00003;Parent=mRNA00001,mRNA00003
10 ctg123 . exon     5000 5500 . + . ID=exon00004;Parent=mRNA00001,mRNA00002,mRNA00003
11 ctg123 . exon     7000 9000 . + . ID=exon00005;Parent=mRNA00001,mRNA00002,mRNA00003
12 ctg123 . CDS      1201 1500 . + 0 ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
13 ctg123 . CDS      3000 3902 . + 0 ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
14 ctg123 . CDS      5000 5500 . + 0 ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
15 ctg123 . CDS      7000 7600 . + 0 ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
16 ctg123 . CDS      1201 1500 . + 0 ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
17 ctg123 . CDS      5000 5500 . + 0 ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
18 ctg123 . CDS      7000 7600 . + 0 ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
19 ctg123 . CDS      3301 3902 . + 0 ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
20 ctg123 . CDS      5000 5500 . + 1 ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
21 ctg123 . CDS      7000 7600 . + 1 ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
22 ctg123 . CDS      3391 3902 . + 0 ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
23 ctg123 . CDS      5000 5500 . + 1 ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
24 ctg123 . CDS      7000 7600 . + 1 ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
```

GFF3 file format

The first line of a GFF3 file must be a comment that identifies the version.

Fields must be tab-separated.

Fields:

Seqid

source

Type

Start

End

Score

Strand

Phase

attributes

[Source](#)

GTF file format

Tab separated.

Fields:

reference sequence

source

method

start position

stop position

score

strand

phase

group

[Source](#)

BAM/SAM file format

SAM

```
@HD      VN:1.0  SO:coordinate
@SQ      SN:chr20          LN:64444167
@PG      ID:TopHat        VN:2.0.14          CL:/srv/dna_tools/tophat/tophat -N 3 --read-edit-dist 5 --read-realign-edit-dist 2 -i 50 -I 5000 --max-coverage-intron 5000 -M -o out /data/user446/mapping_tophat/index/chr20 /data/user446/mapping_tophat/L6_18_GTGAAA_L007_R1_001.fastq
HWI-ST1145:74:C101DACXX:7:1102:4284:73714    16      chr20    190930  3      100M    *      0      0
                                                CCGTGTAAAGGTGGATCGGGCACCTCCCAGCTAGGCTTAGGGATTCTAGTGGCCTAGGAAATCCAGCTAGTCCTGTCTCAGTCCCCCTCT
C      BBDCCDDCCDDDDCDDDDCDCCDBC?DDDDDDDDDDDDCCDCDDDDDDDDCCCCEDDC?DDDDDDDDDDDDDDDDDBDHFFFFDC@?
AS:i:-15         XM:i:3  X0:i:0  XG:i:0  MD:Z:55C20C13A9 NM:i:3  NH:i:2  CC:Z:=  CP:i:55352714  HI:i:0
HWI-ST1145:74:C101DACXX:7:1114:2759:41961    16      chr20    193953  50     100M    *      0      0
                                                TGCTGGATCATCTGGTTAGTGGCTTCTGACTCAGAGGACCTCGTCCCTGGGCAGTGGACCTTCAGTGATTCCCTGACATAAGGGGCATGGACGA
G      DCDDDDDEDDDDDDCDDDDDDCCCDDCDDEEC>DFFEJJJJJIGJJJJIHGBHHGJIJJJJJJGJJJJIJJJJJJHJJJJJJHHHHHFFFFFCCC
AS:i:-16         XM:i:3  X0:i:0  XG:i:0  MD:Z:60G16T18T3 NM:i:3  NH:i:1
HWI-ST1145:74:C101DACXX:7:1204:14760:4030    16      chr20    270877  50     100M    *      0      0
                                                GGCTTATTGGTAAAAAAGGAATAGCAGATTAATCAGAAATTCCACCTGGCCAGCAGCACCAACCAGAAAGAAGGGAAAGAACAGGAAAAACCA
C      DDDDDDDDDCCDDDDDDDDDEEEEEEEFFFEFFEGHHHFGDJJIHJJIIJJIIIIGGFJJHIIIIJJJJJJIGHHFAHGFHJHFGGHFFFDD@BB
AS:i:-11         XM:i:2  X0:i:0  XG:i:0  MD:Z:0A85G13  NM:i:2  NH:i:1
HWI-ST1145:74:C101DACXX:7:1210:11167:8699    0       chr20    271218  50     50M4700N50M    *      0
0       GTGGCTCTTCCACAGGAATGTTGAGGATGACATCCATGTCTGGGTGCACTTGGGTCTCCGAAGCAGAACATCCTCAAATATGACCTCTG
```

SAM

- SAM – Sequence Alignment Map.
- Text format
- Specific structure
- Uses a lot of space (memory)

SAM specification

Field	Regular expression	Range	Description
QNAME	[^ \t\n\r]+		Query pair NAME if paired; or Query NAME if unpaired ²
FLAG	[0-9]+	[0,2 ¹⁶ -1]	bitwise FLAG (Section 2.2.2)
RNAME	[^ \t\n\r@=]+		Reference sequence NAME ³
POS	[0-9]+	[0,2 ²⁹ -1]	1-based leftmost POSition/coordinate of the clipped sequence
MAPQ	[0-9]+	[0,2 ⁸ -1]	MAPping Quality (phred-scaled posterior probability that the mapping position of this read is incorrect) ⁴
CIGAR	([0-9]+[MIDNSHP])+ *		extended CIGAR string
MRNM	[^ \t\n\r@]+		Mate Reference sequence NaMe; “=” if the same as <RNAME> ³
MPOS	[0-9]+	[0,2 ²⁹ -1]	1-based leftmost Mate POSition of the clipped sequence
ISIZE	-? [0-9]+	[-2 ²⁹ ,2 ²⁹]	inferred Insert SIZE ⁵
SEQ	[acgtnACGTN.=]+ *		query SEQuence; “=” for a match to the reference; n/N/. for ambiguity; cases are not maintained ^{6,7}
QUAL	[!--]+ *	[0,93]	query QUALity; ASCII-33 gives the Phred base quality ^{6,7}
TAG	[A-Z][A-Z0-9]		TAG
VTYPE	[AifZH]		Value TYPE
VALUE	[^\t\n\r]+		match <VTYPE> (space allowed)

SAM specification



A

Header	Format version	Coordinate system	Sequence length	RNEXT: Ref. name of the mate/next fragment	TLEN: observed template length	QUAL: ASCII of Phred-scaled base quality
Alignment sorting order @HD @SQ	VN:1.1 SN:test	SO:coordinate LN:97	Sequence length	RNEXT: Ref. name of the mate/next fragment	TLEN: observed template length	QUAL: ASCII of Phred-scaled base quality
Sequence name r1 r2 ...	0 16	chr10 chr10	27 12 30 20 75M	*	0 0 0 0 0	TGTGTTCTTGC GTT CTCTATCATCAACATTGTG
QNAME: Read or Query name	Bitwise FLAG	RNAME: Reference seq name	POS starting position of the read	MAPQ: Mapping quality	CIGAR String	PNEXT: Position of the mate/next fragment
						SEQ: fragment sequence

B

Source

SAM tags

- Tags are meaningful.
- You can decode tags here:
 - <https://broadinstitute.github.io/picard/explain-flags.html>

SAM resources online

- [Here](#)

BAM

- Binary version of SAM
- Does not use as much space as SAM
- Is used to store mapped data
- May be used to store raw reads
- Possible to convert to SAM or FASTQ

SAM/BAM

- We will use these:
 - To store reads (just theoretically. IonTorrent produces BAM files, but we will not use raw IonTorrent data)
 - After mapping (to store the results)
 - After QA/QC of mapped data

FASTA

```
SKLSSESSITLSTDSSSEQVNLKSNSGNYQVRAMSADDFFDLPMVENGAFLKVNANSFAVSLKSTLFASSTDEAKQILTG  
VNLCFEGNSLKSAATDGHLAVLDLNQNVIASETNPEINNLSEKLEVTPLSRSLRELERFLSGCKSDSEISCFYDQQGFVF  
ISSGQIITTRTLGDGNYPNQNQLIPDQFSNQLVLDDKKYFIAALERIAVLAEQHNNVVKISTNKELQILNISADAQDLGSGS  
ESIPIKYDSEDIQIAFNSRYLLEGKIIETNTILLKFNAPTTPAIFTPNDETNFVYLMPVQIRS  
>WP_011124158.1 phosphoribosylformylglycinamide synthase subunit PurL [Prochlorococcus marinus]  
MIDISFNRLIKSKDYLVDYNEVLAALKEGLTKADYIEICNRLRKSPRNTELGMFVGMSEHCYRNRSRLLSNPTTGKN  
ILVGPGENAGVVDLGEQRALFKIESHNHPSAIEPFQGAATGVGGILRDFTMGARPIALLNSLRFGPLDTPINVGLLEG  
VVGIAHYGNCVGPTVGVGEAFDKSYGNPLVNAMALGIMETKEIVCSGAKGIDFPVIVGSGTGRDMGGAASFASSEL  
TQASIDDRPAVQVGDPLEKGLIEGCLEAFKTGYVVAQDMAAGLTCSCEMAAKGGVGIELDDLDLVPPAREKNMTAYEF  
LLSESQERMLFVVEPGEELIMNKRKWLQAKVVGKVLEENIVRVIHEKQIVVNLNPDLAEDTPVNKHILLEEPGFI  
KDHWKWDETSLPQDSVSIKGVHFKENNTILDWNQIIRLLDDPTIASKRWVYQVQDNNTIAPGIGDAALIRLREIEN  
QNQNSNRGIAAVVDCPNRWALDPERGAAVAEAEARNISNCVGAKPLAVTDNLNFSSPEDPIGYWQLAKACKGLSKACVV  
LETPTVGGNVSLVNETALPQNLKQPQPTPVGMIGLIQDINSATRQGWKDQGDQIYLLGTSIDPSILNNQNISLAATSY  
LENIYGLKTGRPLIDLEFEKVLVQLFLRESIANNLKSAHDISDGGLVIGLAESCISSGLGIECNLPEIDNRLDKLLFAE  
GGSRVLVSVSPNINIHKNSLNNFNIANSEQISFNYLGTVDNKYFQININQTKIIDLGVNEITNKFERSIPRINSTIV  
S  
>WP_011124159.1 MULTISPECIES: amidophosphoribosyltransferase [Prochlorococcus]  
MCGIVGIVSNQNIQKIQYDSSLLOHQRGQDSTGIATMEGSVFHLHKSKQGVKEAYRTDRMRILTGNIGIGHVRYATSGEA  
HREDEAQPFYVNAPYGIILVHNGNLNTRELEKEFLSIDRRHTNSSDTEMLLNVATEIQSENHYSSLSPHEHIFSAISS  
LHKRVEGTYSIAIMAGYGLFARPDYGIPLVKGRLVLDNGKTEWIVASESLVLENNDFDVRDVVPGEAIFIISNEGEF  
YSKQCADNPQLPCSFYEVYLARPDSVMNGISVYEARLARPGDYLAKTIQEQTSGEIDVMPIDSSSRPSAMQVARQLGL  
EYREGFFKNRYVGRFTIMPQQAQRKKSVRQLNAMSSEFKGNVLLVDDSRGTTSRREVQMAKLAGANKVFTSAAPP  
IRFPHVYGINMPSKMELIAHKDSIDQIQTLLADLGTVYQKISDLEQSILKGSQVKNLDSCFCNGEYVTGKVTEEYLAWVG  
SKYSS  
>WP_011124160.1 MULTISPECIES: DNA topoisomerase 4 subunit A [Prochlorococcus]  
MAKERQPISLHQEMQRSYLEYAMSIVVGRALPDARDGLKPVQRRLYAMHEGLTPERPFRKCARVVGDLCKYHGD  
QAVYDALVRLVQSFSSRYPVLDGHGNFGSIDDPPAAMRYTETRLAPIANQALLNEIDDKTVDFSQNFDGSSQQEPDVLP  
QLPFLVNLNGCSGIAVGMATNIPPHNINEIINGLIALIKKPELSEDCKLIEIIPGDPFTGGEIILSNGIKETYLKKGKGSIP  
MRGITHIEEINPGKGKHKRKKGIVITELPYQVNKAWEIKLADLVNNSKIDGIADIRDESDRDMRIVIELRRDIDHEIVK  
DTLYQKTNLQNNFSATLLALVNGQPKQLSKRLLIITLEYRELTIRKRTKLNKLQTLARLEILEGLIKALKNKKVINVLI  
ENAKDSIDAKIKIMSELKLNKEQSEGILSMPRLKLNLETQGLYNEANDLKILEKLNREILENRSQQLSEMVNELKLLKK  
KFGSQRKTKLVEGGDELVAERNANLRPNAELRKKAFAESLPKDGYLIQDDQVKILRPQILTKLNLTESCLLGEGPVPT  
RLLWPIAKQPKILAITNHGKIALLKWEAGSQPGQLNRFLPSGLEGEKITNLLPLNTENLSLGLSTDGRFKRTNLNEI  
IDISGRATTIQLKLDGVFLKSALLCPLNGHLLVVTVNIGRIIKLKINEEISIPLMGKLAQGSVMIKLFPGENIIGALTKEK  
ENCNLLISEKGTVIKHSLSKIKSEKGKLDGEIGISFKDKNNQDKRLIQTYNAKQLVGIKTNQGKHGRVSSNQIEKLKFN  
EEQKKAFNLEKNELLEKIPLIEENSYN  
>WP_011124161.1 MULTISPECIES: tetratricopeptide repeat protein [Prochlorococcus]  
MKNYKKIYWLICIIINTLNFVINKPSHAYIPNIYSPNPVKLIDTSIGIILTASEYIKYQGTKEAIGLAKLAIISLNPKEIE  
LWIIILARAQLSNKLEALISIERAKNINPNPIPIWLFTKASIEMOMGEIQLAINSINKCLKIEKKNSNAYFLGNQAKLQ
```

FASTA

- Header line and information lines
- Header starts with “>”
- DNA/protein sequences
- Multiple sequences in one file

FASTA

- We will use FASTA to :
 - Store reference sequence
 - Store contigs, scaffolds, etc.

FASTQ

```
@read1
AGCTTATCCTCTGCTACCCCCGGGTTAGCGCACTTGATGTATTACAGC
+
BA1@CC7CBCCC9C8;B2@>C?B@B@B3=9?@B1 : AB7B?B8B?B6B. 7.
@read2
TTGGGCAGGGATCTCCAGAACATATGGATGTGATCCACACAGCATTCTGC
+
?>?B@) <?@, AA7A@C<C?=@@B;+ ) ?B5* @2=@+=BB, =B6C>AB@B24
@read3
TATGCTCAAGAACGGGCTGATGAGTTGGTGTTCACGATATCACTGCCTC
+
A3AB: B1 : B; 9/0BBCBB<BB@AA0?BB9: BB<A@BB@7@6@<A@@@<3
```

FASTQ

- Text file
- Used to store NGS data
- Stores sequence and quality scores
- 4 rows per sequence:
 - Header (starts with @) (header is meaningful)
 - Sequence
 - Comment
 - Quality scores

FASTQ header (Illumina)

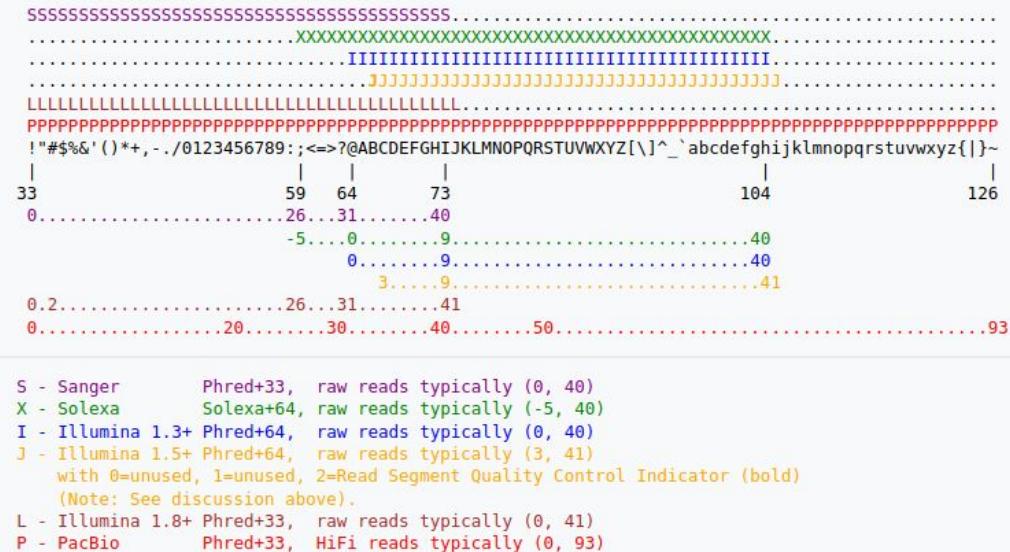
Element	Requirements	Description
@	@	Each sequence identifier line starts with @
<instrument>	Characters allowed: a–z, A–Z, 0–9 and underscore	Instrument ID
<run number>	Numerical	Run number on instrument
<flowcell ID>	Characters allowed: a–z, A–Z, 0–9	
<lane>	Numerical	Lane number
<tile>	Numerical	Tile number
<x_pos>	Numerical	Run number on instrument
<y_pos>	Numerical	X coordinate of cluster
<read>	Numerical	Read number. 1 can be single read or Read 2 of paired-end
<is filtered>	Y or N	Y if the read is filtered (did not pass), N otherwise
<control number>	Numerical	0 when none of the control bits are on, otherwise it is an even number. On HiSeq X systems, control specification is not performed and this number is always 0.
<sample number>	Numerical	Sample number from sample sheet

FASTQ comment

- Normally just “+”.
- May include sample ID (in processed data)

FASTQ QUALITY

PHRED score



Phred quality score	Probability that the base is called wrong	Accuracy of the base call
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%

FASTQ

- We will use FASTQ to:
 - Store original reads
 - Store cleaned data/reads

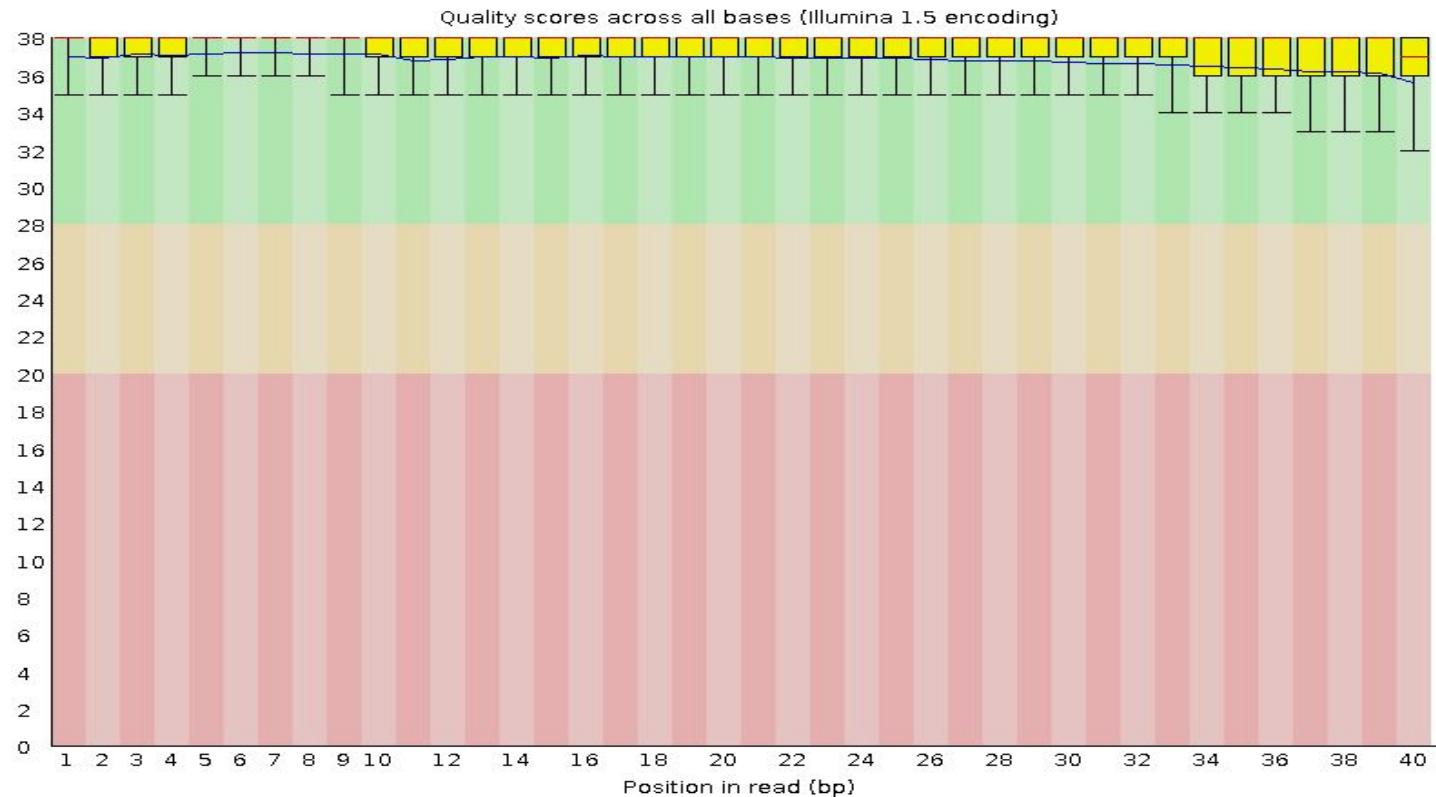
FASTQ QUALITY ASSESSMENT

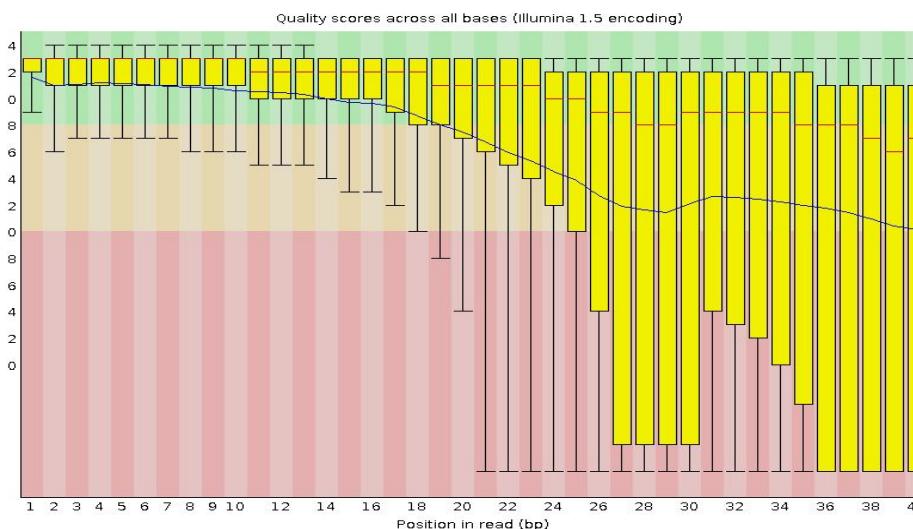
Measurements

Measurements

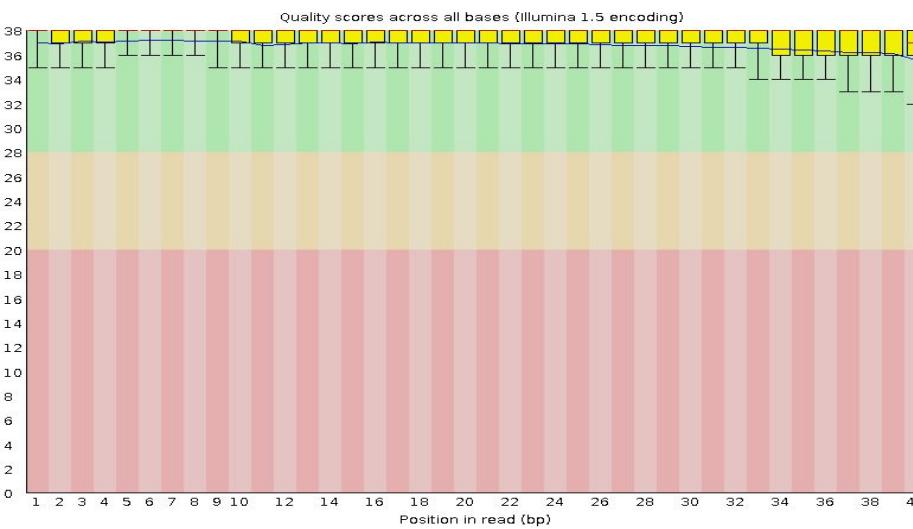
- Quality per position
- Average quality
- Length distribution
- Composition
- GC content

Per base quality





- Average sample (needs some processing)
- Good sample



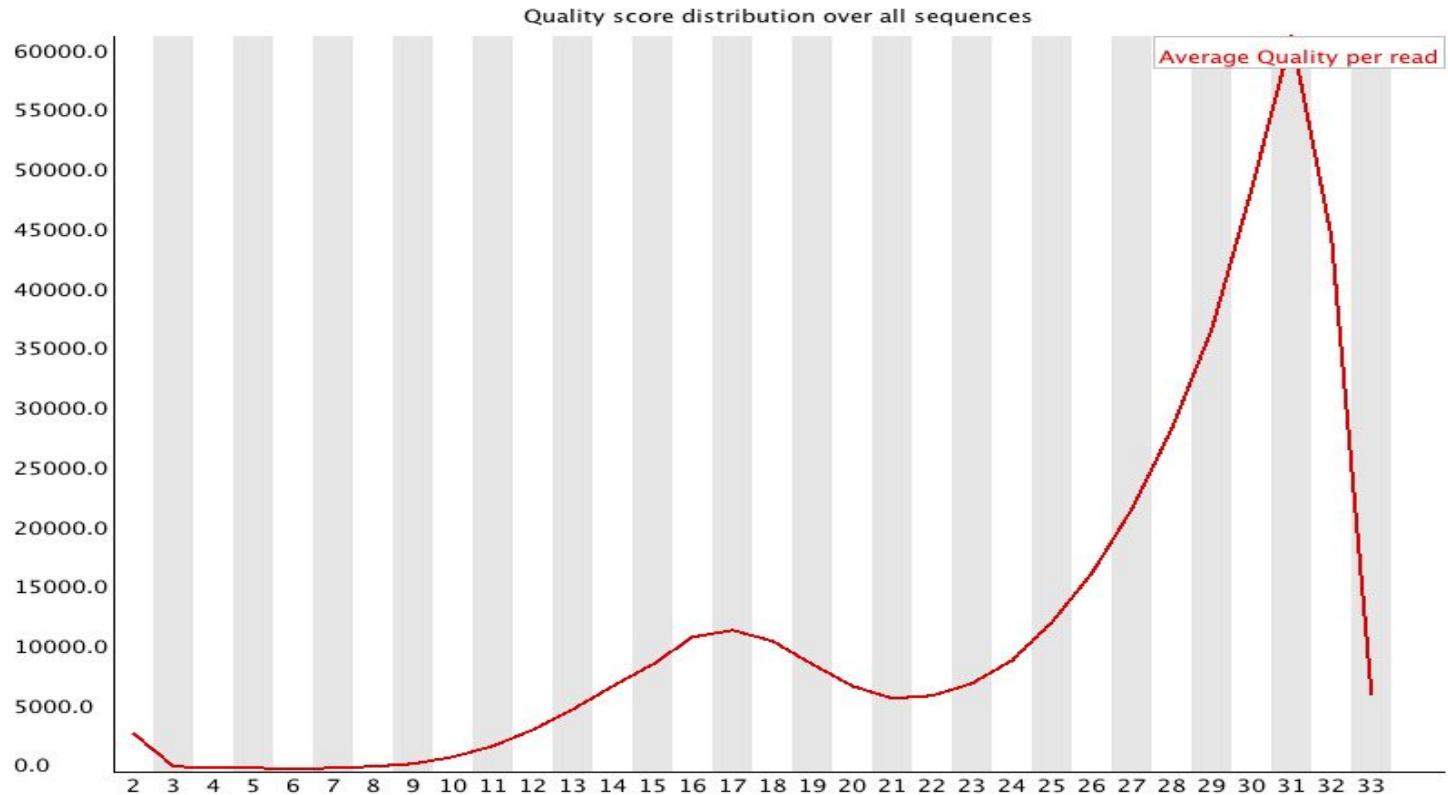
Per base quality

- Some statistics are calculated for each base:
 - Red line - median
 - Yellow box - interquartile range
 - Blue line - average quality

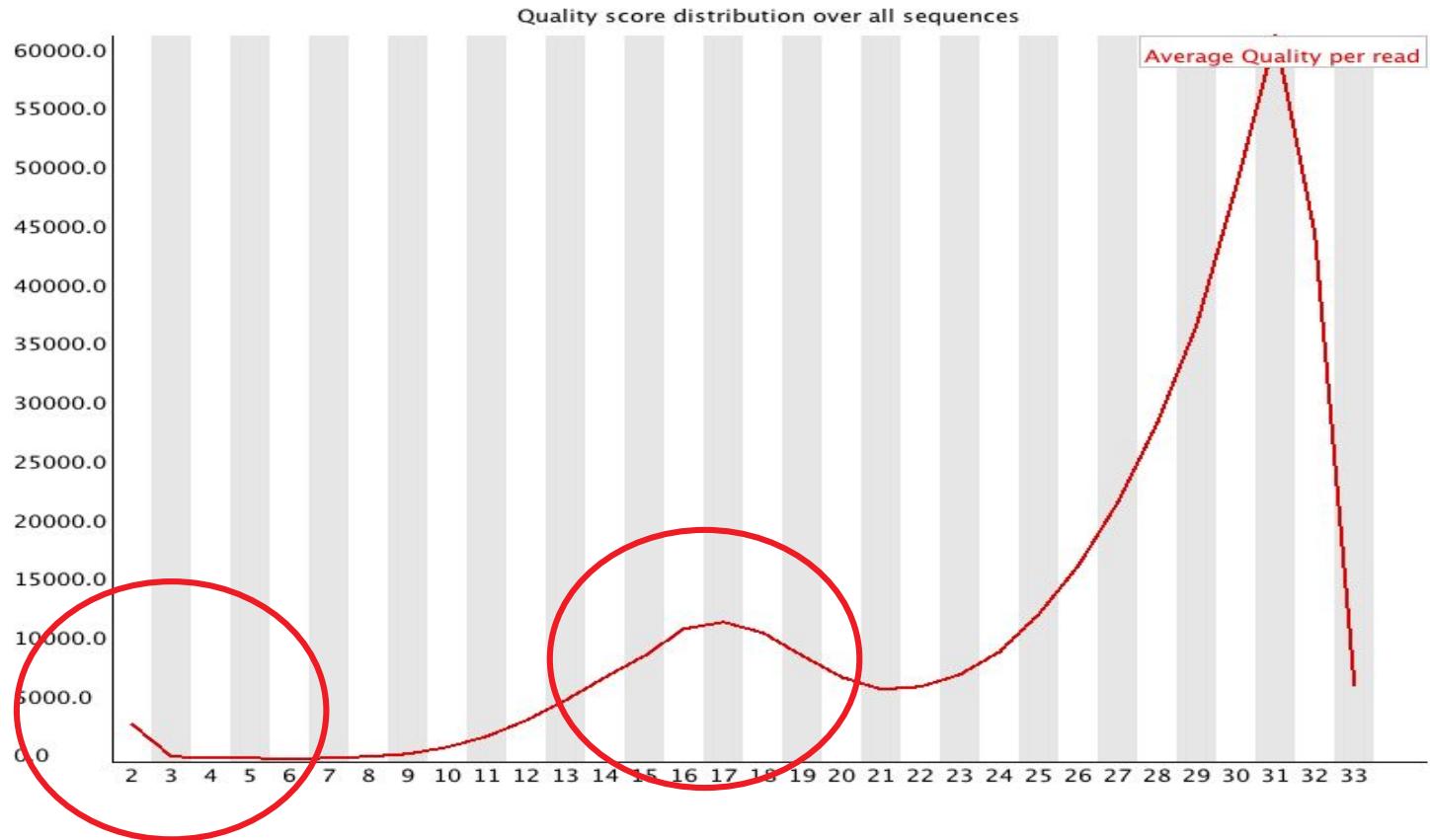
Per base quality

- How to understand/evaluate?
 - Green zone - good
 - Red zone - not so good
- What is important?
 - It is normal that the quality drops at the end of the reads.
 - Make sure that quality scoring system is set correctly (mostly this is not a problem)

Per sequence quality



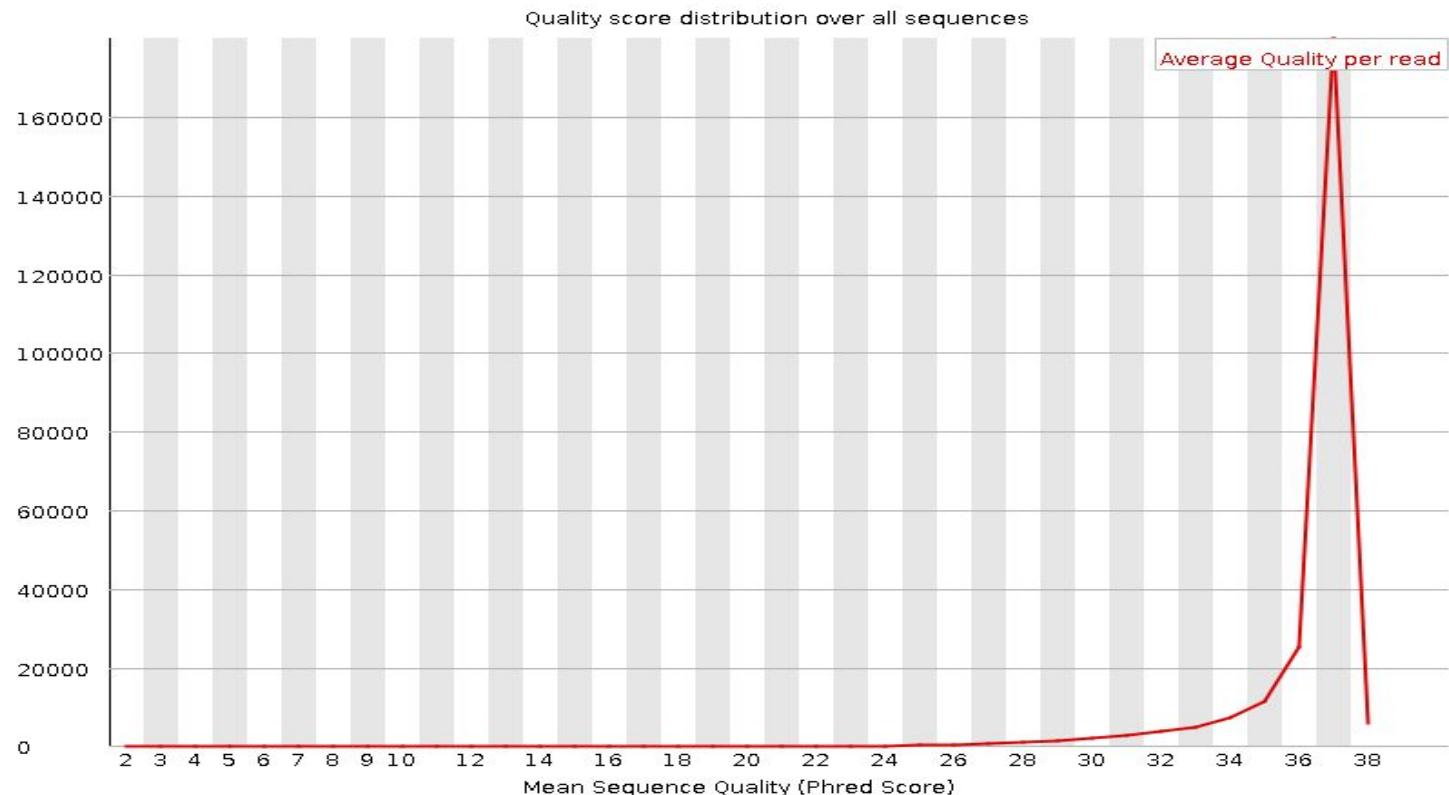
Per sequence quality



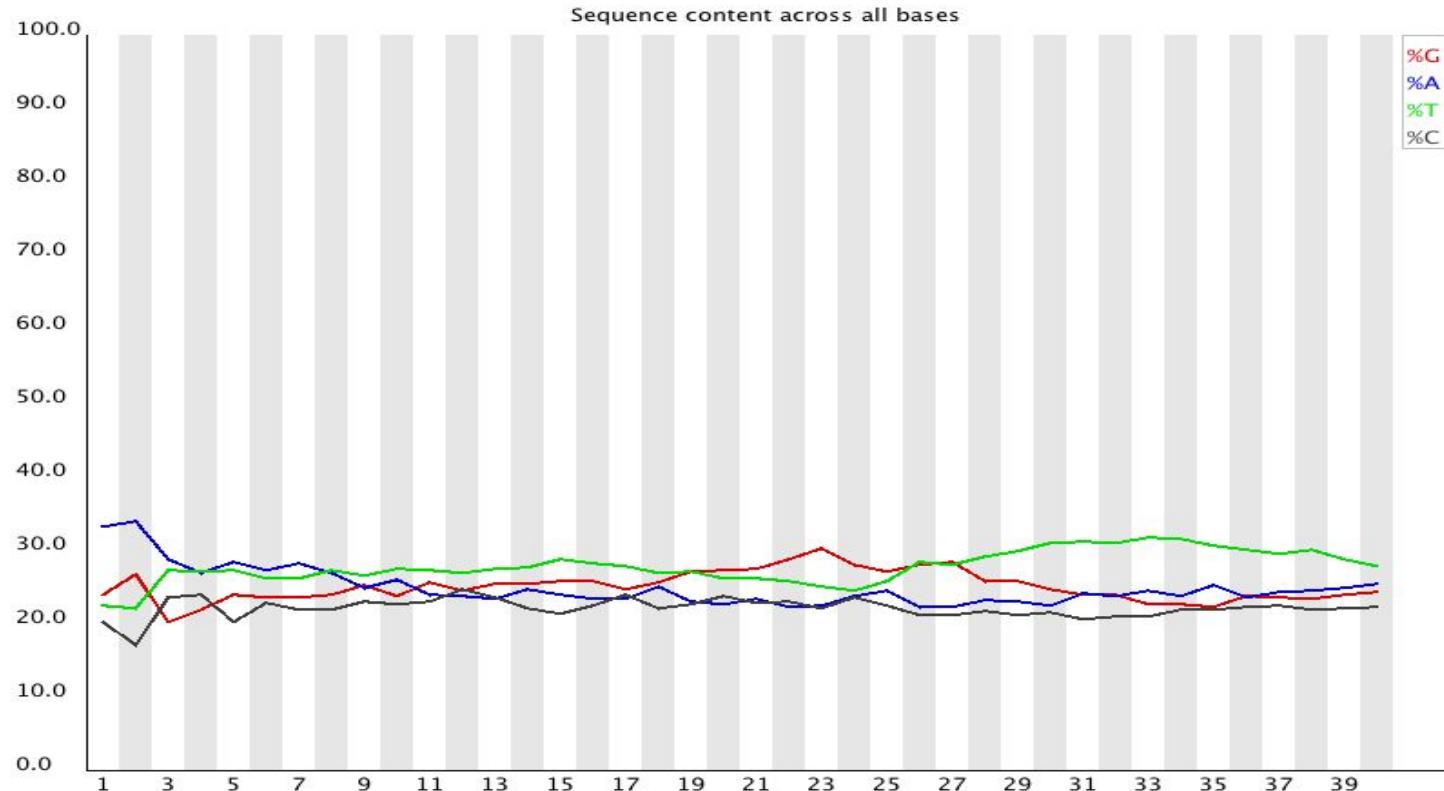
Per sequence quality

- Average sequence quality
- You can identify if all sequences are similar (based on quality)

Per sequence quality - good



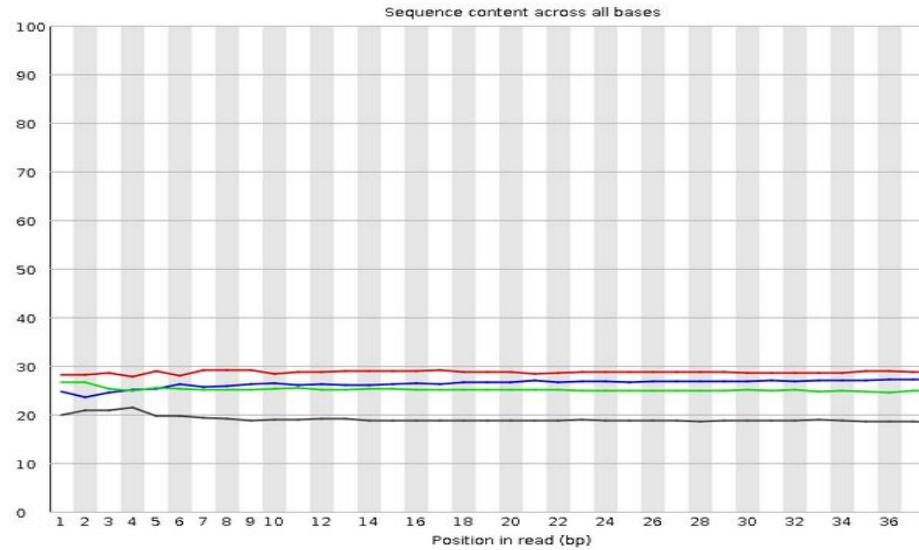
Per base sequence content



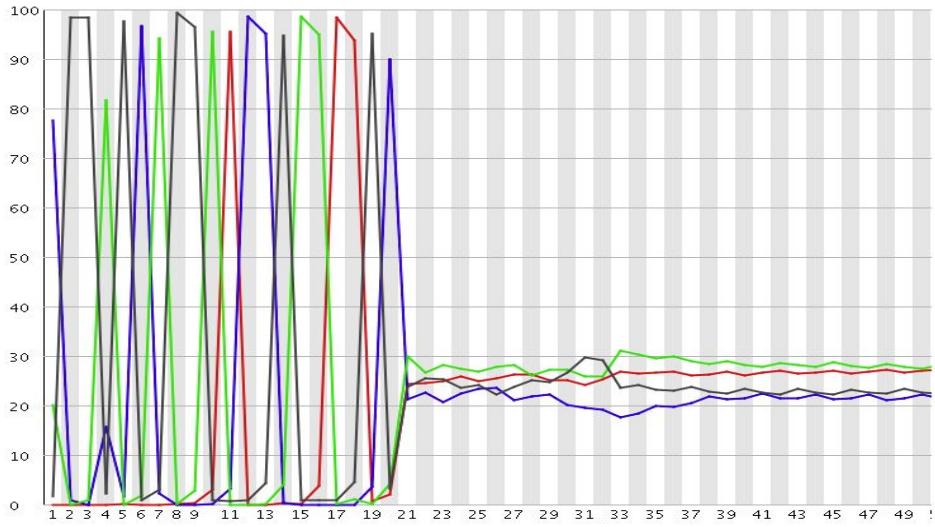
Per base sequence content

- Nucleotide frequency per positions
- Any ideas for the good/bad results?

Per base sequence content - good



GOOD

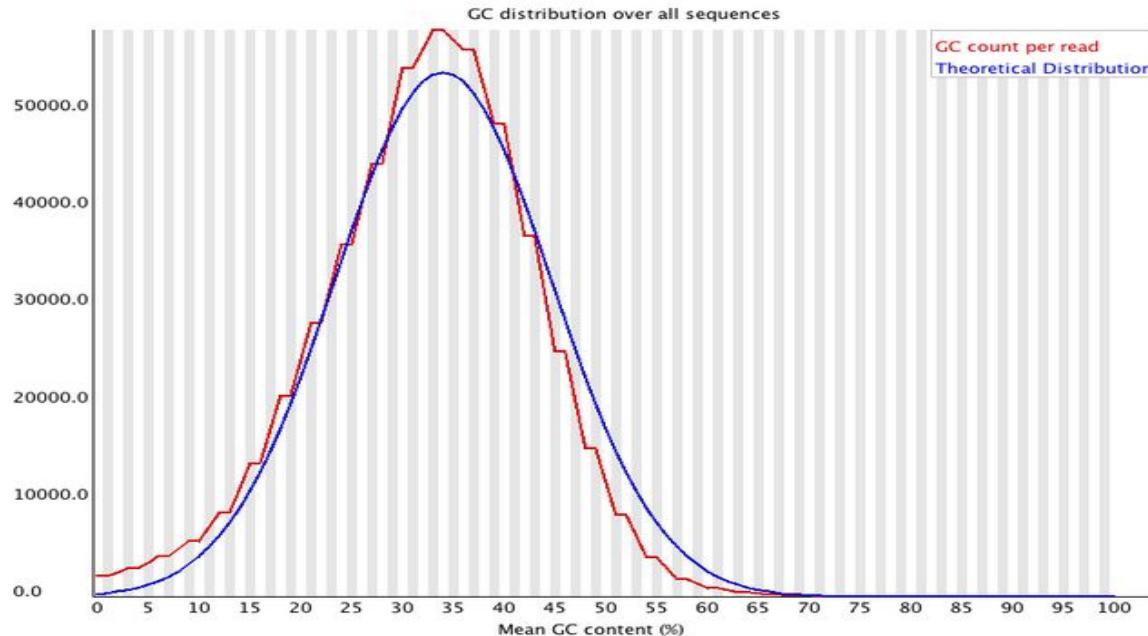


Suspicious (bad in most cases)

Per base sequence content - bad

- Important:
 - Specific experiments may require/give different content and this may be normal.
 - Straight lines are normal for genomic sequences; small 5' bias is normal for transcriptomics; general bias is normal for amplicon; etc.

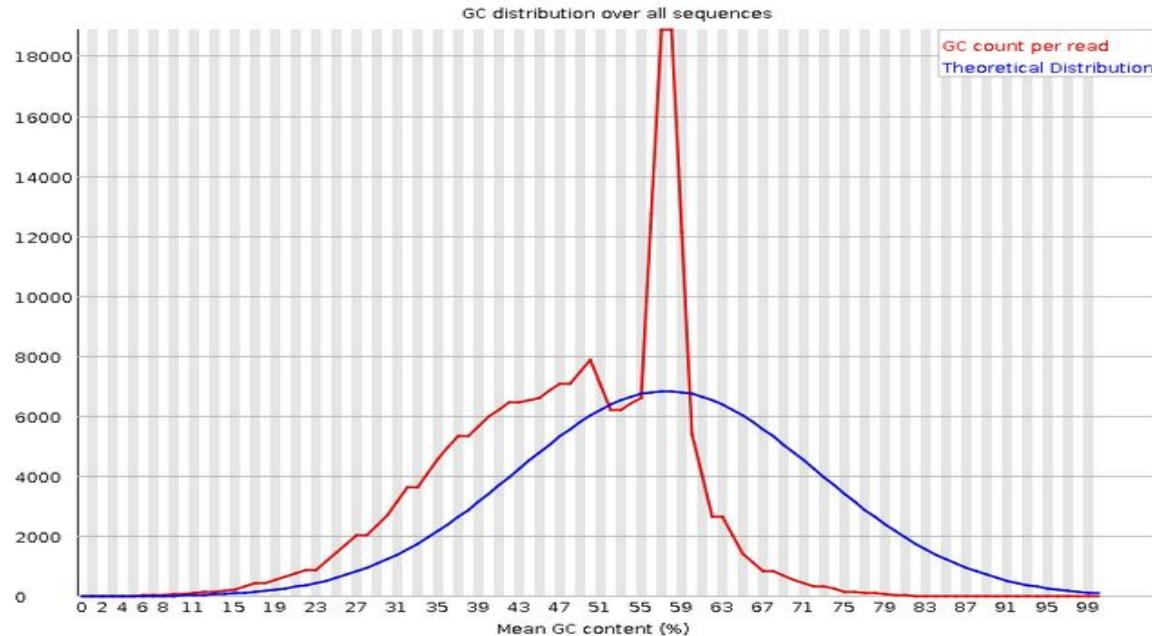
GC content



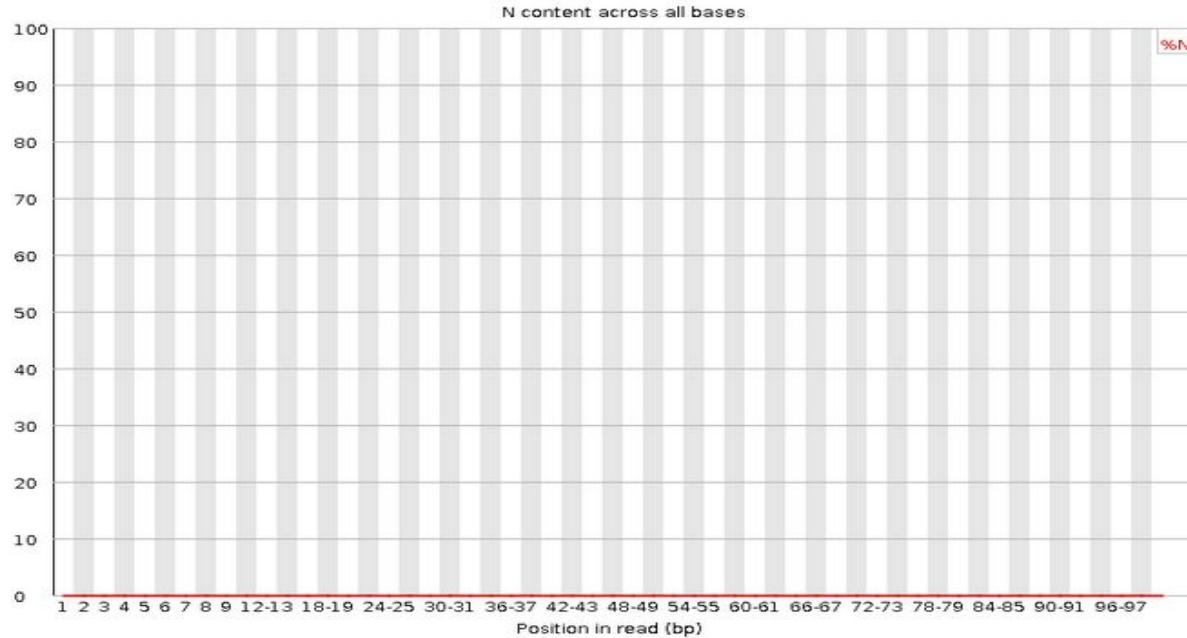
GC content

- Should be normal distribution
- You have to know your organisms GC content
- Non-standard (normal) distribution may be due to:
 - Specificity of the organisms/workflow
 - Errors during library preparation
 - Contamination

GC content



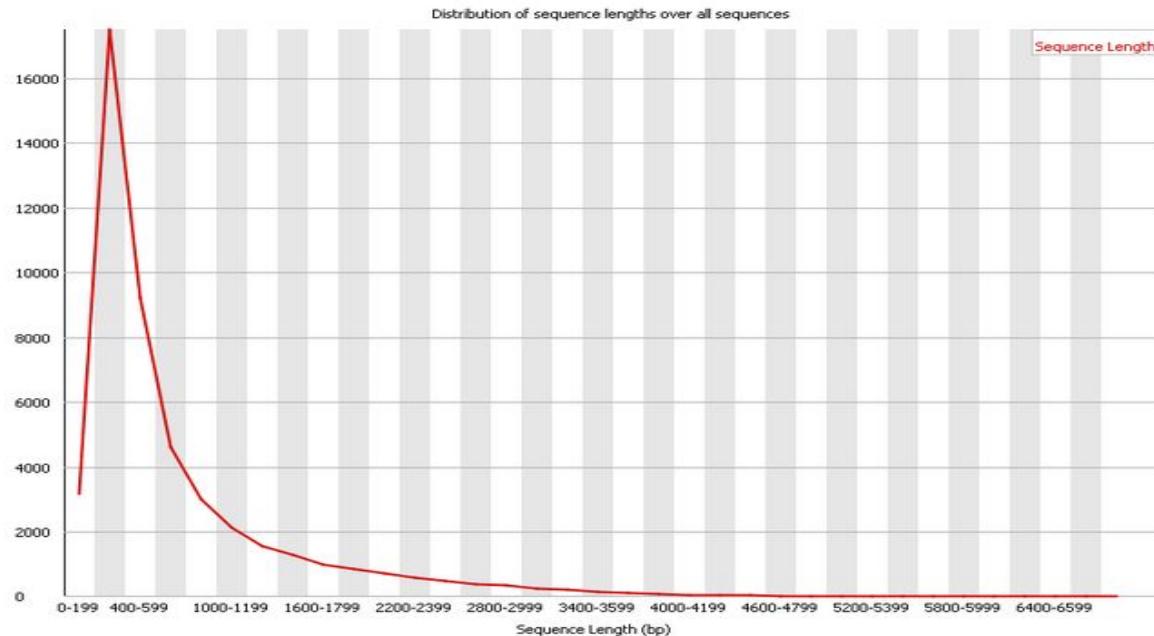
Per base N content



Per base N content

- There should be no in high quality data
- Sometimes is observed at the end of the read – just trim them

Sequence length distribution



Sequence length distribution

- Displays sequence length distribution
- Interpretations depends on the platform/preparations of the samples

Hyperrepresented sequences

- Some sequences may be observed much more often than others
- This may be due to biological reasons or due to contamination..

Reading before bedtime

- Some most common “fails”:
 - <https://sequencing.qcfail.com/articles/>

Quality control

- If you see any problems in your data – you have to fix it.
- HOW?

FastQC program

FASTQC program

- Download:

<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

- Video link: <https://www.youtube.com/watch?v=bz93Re0v87Y>

- User guide:

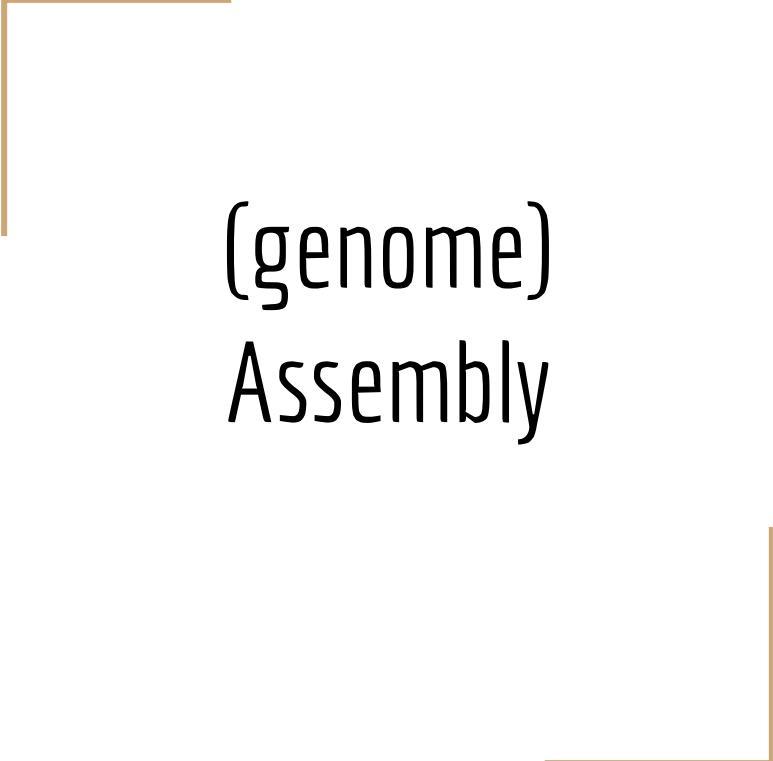
<https://biof-edu.colorado.edu/videos/dowell-short-read-class/day-4/fastqc-manual>

Are these all data formats? NO :)



Resources

- <https://doi.org/10.1016/B978-0-12-803077-6.00005-9> and
<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4527835/#pcbi.1004393.s005> (figures, if not indicated by the picture)



(genome) Assembly

Assembly



Lecture outline

- What is a genome assembly and why do we need it?
- Terminology and basic principles
- OLC and de Bruijn graph
- Galaxy resources for genome assembly (and other things)
- Limitations

Genome assembly

- **Assembly** is the process of turning shorter **reads** into longer **contigs**.



Why?

- General interest
- Variant calling
- Evolutionary aspects
- Disease analysis/treatment
- Etc.



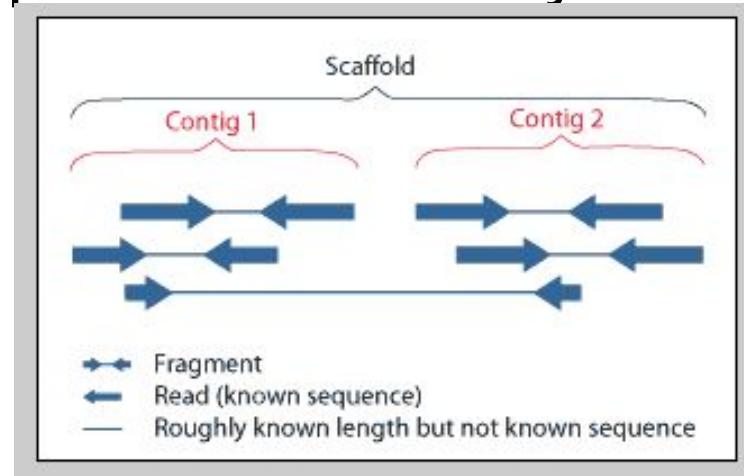
Source

Assembly types

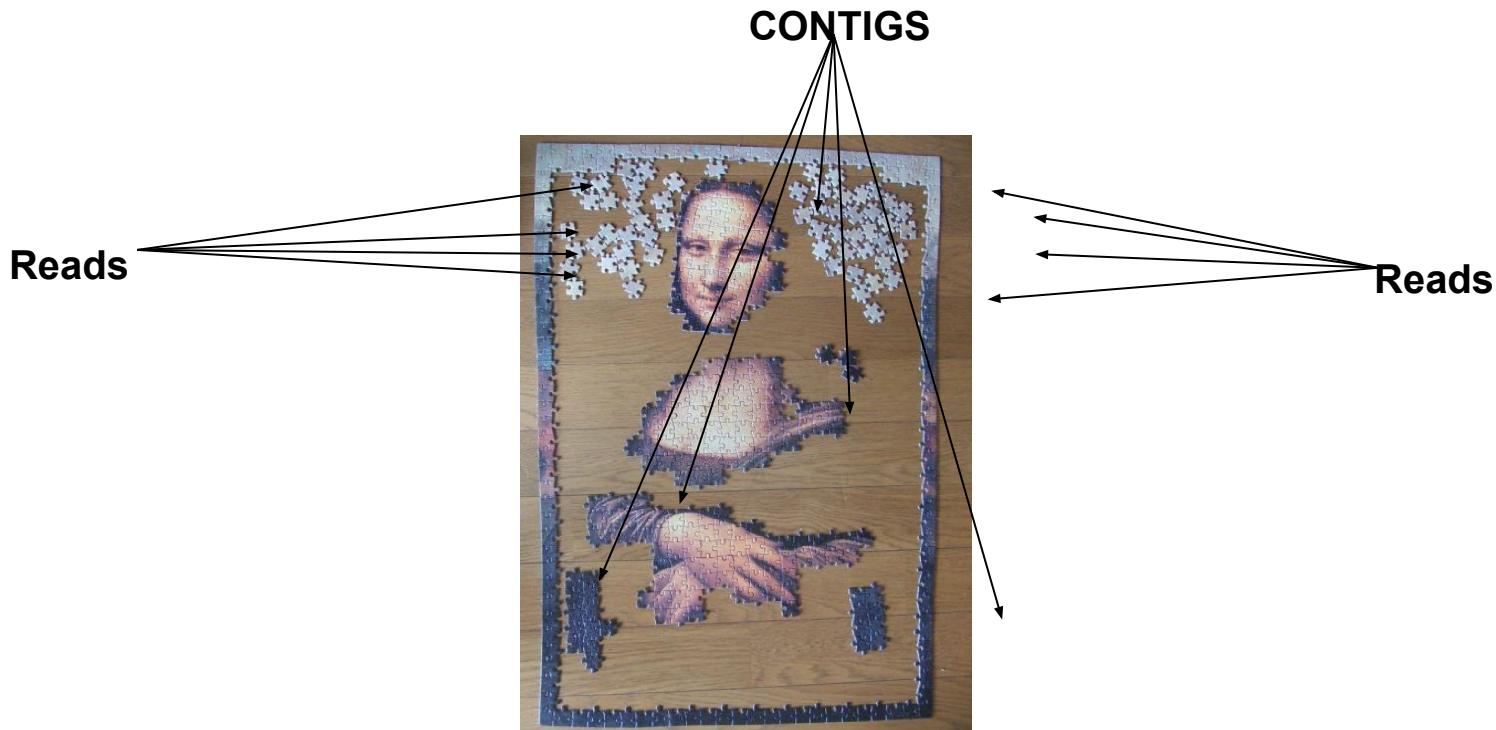
- Genome assembly
- Transcriptome assembly
- Meta-genome assembly
- Meta-transcriptome assembly

Genome assembly (terminology)

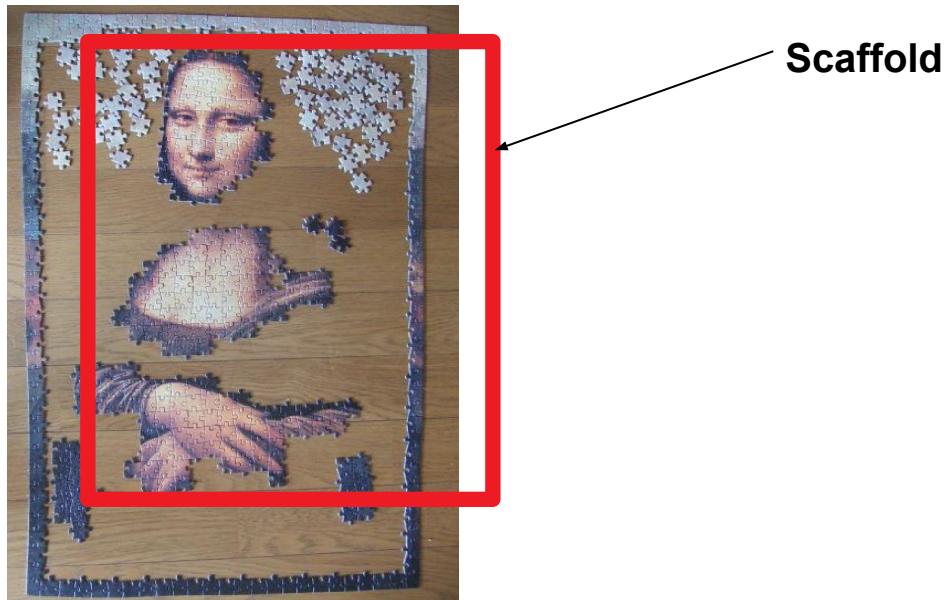
- Read („nuskaitymas“) - any sequence that comes out of the sequencer
- Contig (?) - gap-less assembled sequence
- Scaffold („pastolis“) - sequence which may contain gaps (N)
- K-mer - any
- sequence of length k

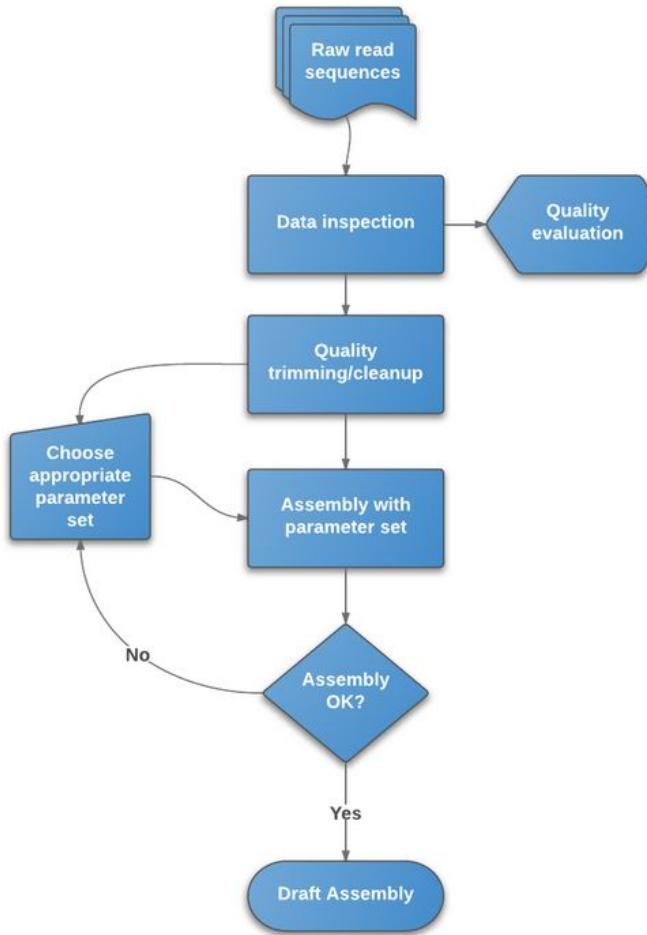


Genome assembly (terminology)



Genome assembly (terminology)





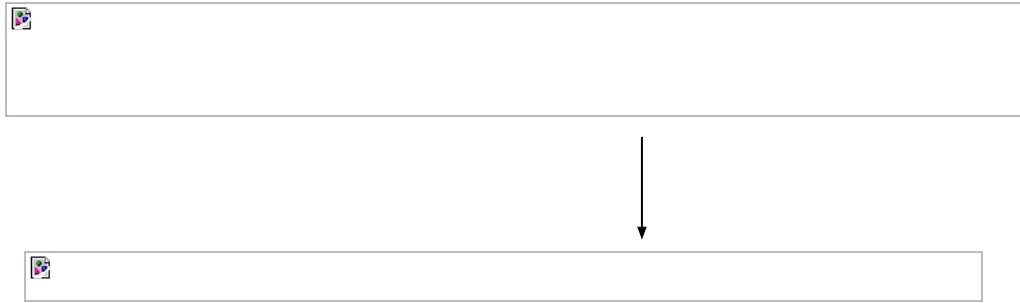
Basic workflow

Why it is complicated?

. Reads to be assembled:

- ATATCGAAGAA
- GGAAATACATTATTAA
- GTACAAACATA
- AGAATAAGGAAAG
- CATAGAAGGAGGAAA
- AGGAAAGATGGCATTAA

. How to assemble? What is optimal assembly?



Notice:

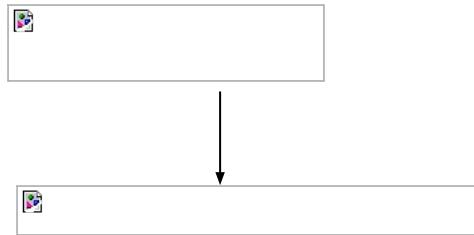
Contigs may represent different chains
Contigs are not ordered

. Reads to be assembled

- ATATCGAAGAA
- GGAAATACATTATTAA
- GTACAAACATA
- AGAATAAGGAAAG
- CATAAGAAGGAGGAAA
- AGGAAAGATGGCATTAA

Any alternative assemblies?

• Alternative assembly (one read was removed)



Assemblers decide which variant of assembly to use.

. Why it is hard to assemble
repetitive fragments?



[Listen, if you don't know this song](#)

. We have reads

- TROLLOLLOOTROLLOL
- LOOTROLLLOLOTROLLOLL
- ROLLOLLOTROLLOLLO
- OLLOLLOOTROLLLOLOTROLLO

Alternative assemblies



- TROLLOLLOOTROLLOLLOOTROLO
- ROLLOLLOOTROLLOLLOTROLLOLLOOTROLLO
LLOOTROLLOLLO
- OLLOLLOOTROLLOLLOOTROLLOLLOTROLLOL
LOOTROLLOLLOOTROLLOLLO

• Problems in genome assembly

- Only 4 nucleotides – high amount of repetitive fragments
- Reads are short
- System bias
- “Biology” (diploids, natural variants and other)

Long reads

- Polymerase read – whole sequence with SMRTbells
- Subreads – sequences without SMRTbell
- Circular consensus read, CCS – consensus of many subreads
- Read of insert, ROI – CSS
- 1D reads - Template and complement sequences (MinION)
- 2D reads – Consensus between template and

Basic steps

- Prepare the library (not your job, but..)
- Do sequencing
- Do data QC
- Assemble contigs with specific parameters
- Make scaffolds
- Check quality. If not happy, restart from step 2
- Finalize the genome/annotate

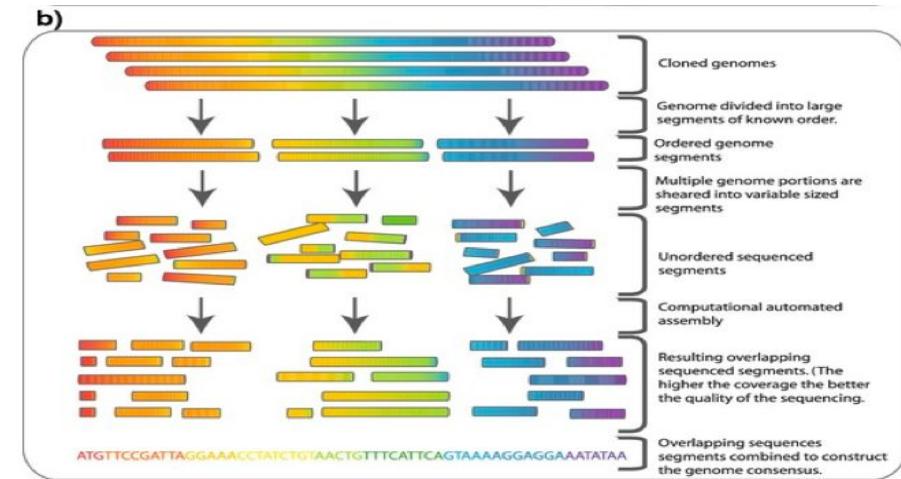
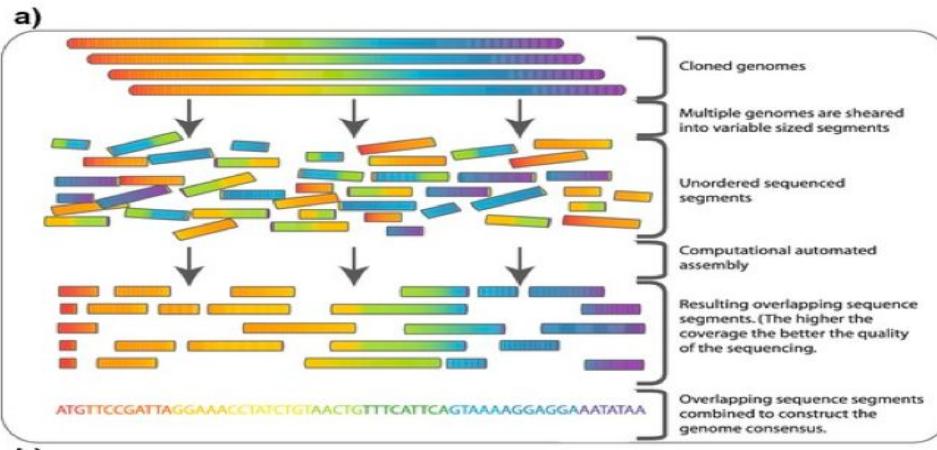
DNA extraction and QC

- **Sample collection**
 - Collect DNA sample from a high-quality source.
 - Sample should be free from contaminants, degradation, and other sources of DNA damage.
- **DNA extraction**
 - Use optimized protocol for the sample type and downstream sequencing technology.
 - Use column-based methods, bead-based methods, or phenol-chloroform extraction.
 - Yield high-quality DNA that is free from contaminants and inhibitors.
- **DNA quantification**
 - Determine concentration and purity using a spectrophotometer or fluorometer.
 - Concentration should be in the range required for the sequencing technology being used.
 - Assess purity by measuring the ratio of absorbance at 260 nm and 280 nm.
- **Quality control**
 - Assess DNA quality using gel electrophoresis or a bioanalyzer.
 - Confirm that DNA is of high quality and free from degradation or contaminants.
 - Assess quality of sequencing library using a bioanalyzer or qPCR.

Select sequencing platform

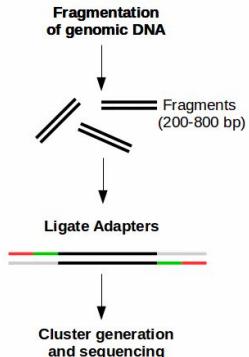
- **Genome size:**
 - Selection of the appropriate sequencing platform depends on the size of the genome
 - Short-read sequencing technologies like Illumina are suitable for smaller genomes
 - Long-read sequencing technologies like PacBio or Oxford Nanopore are better for larger and more complex genomes
- **Sequencing depth and budget**
 - Sequencing depth and budget are important factors to consider
 - Short-read technologies may be more cost-effective for higher sequencing depth, while long-read technologies may be more suitable for lower sequencing depth
 - Illumina sequencing is typically less expensive than long-read sequencing but may require higher sequencing depth
- **Sequencing error rate**
 - Sequencing error rate can affect genome assembly accuracy and completeness
 - Long-read technologies have higher error rates but can generate longer reads for resolving complex regions of genome
 - Short-read technologies have lower error rates and can generate higher coverage
- **Assembly method**
 - Choice of sequencing platform also depends on the assembly method
 - Long-read technologies may be more effective for de novo assembly and resolving complex regions of genome
 - Short-read technologies may be sufficient for reference-based assembly
- **Availability of ref. Genome**
 - The availability of a high-quality reference genome can impact the sequencing platform selection
 - A hybrid sequencing approach that combines short and long-read technologies may be suitable for improving assembly quality and completeness

Shotgun sequencing

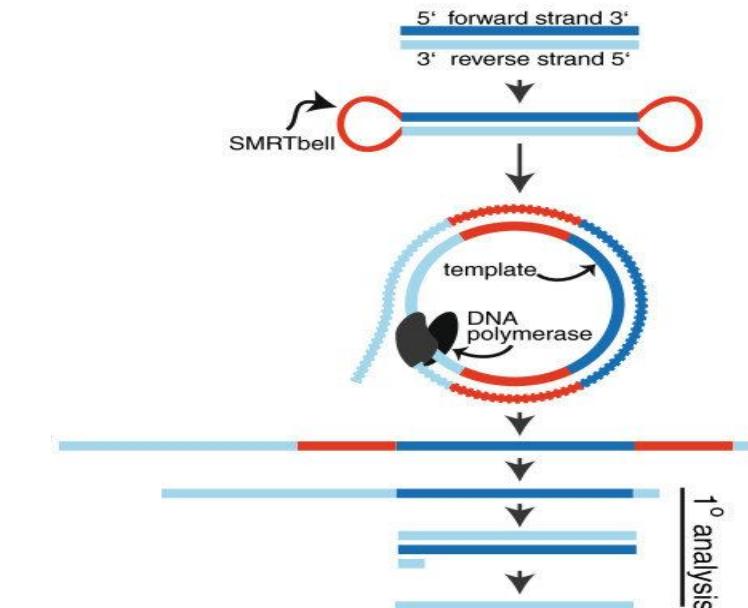
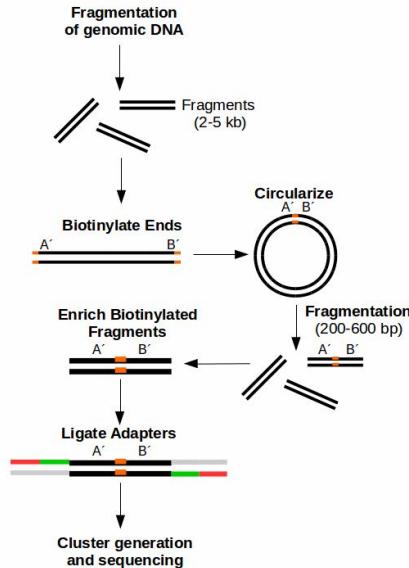


Genome assembly: libraries

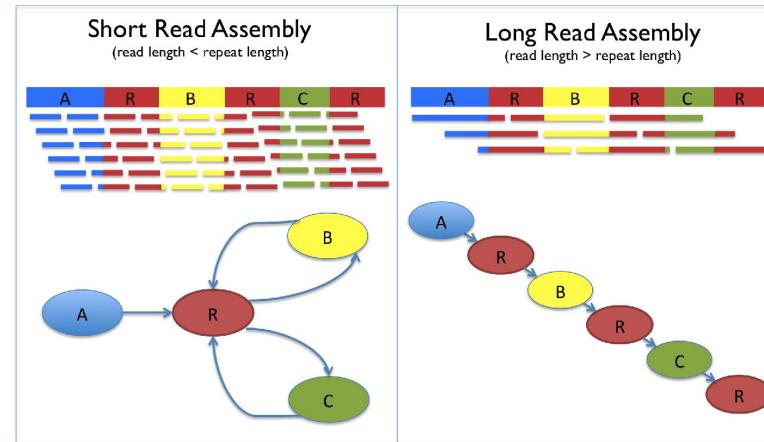
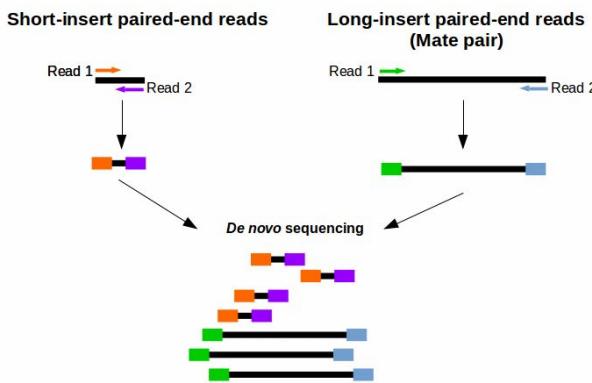
Paired-End Sequencing
(Short-insert paired-end reads)



Mate Pair Sequencing



Genome assembly: libraries

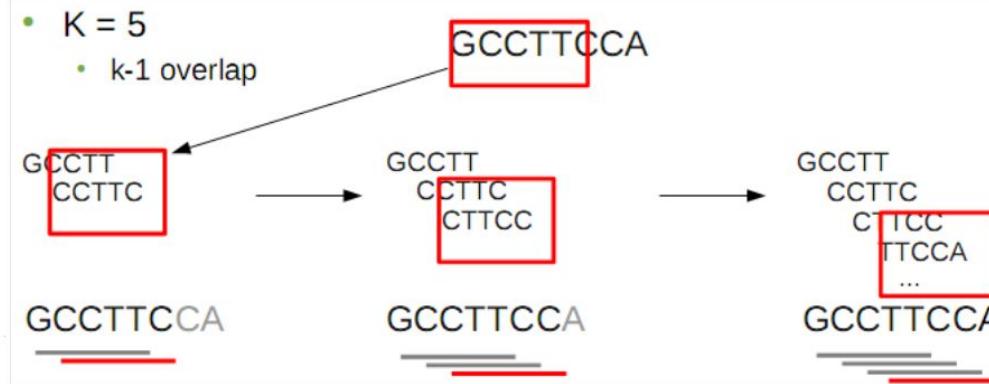


Assembly variants

- De novo
- Reference based
- A mix of de novo and ref. based (hybrid)

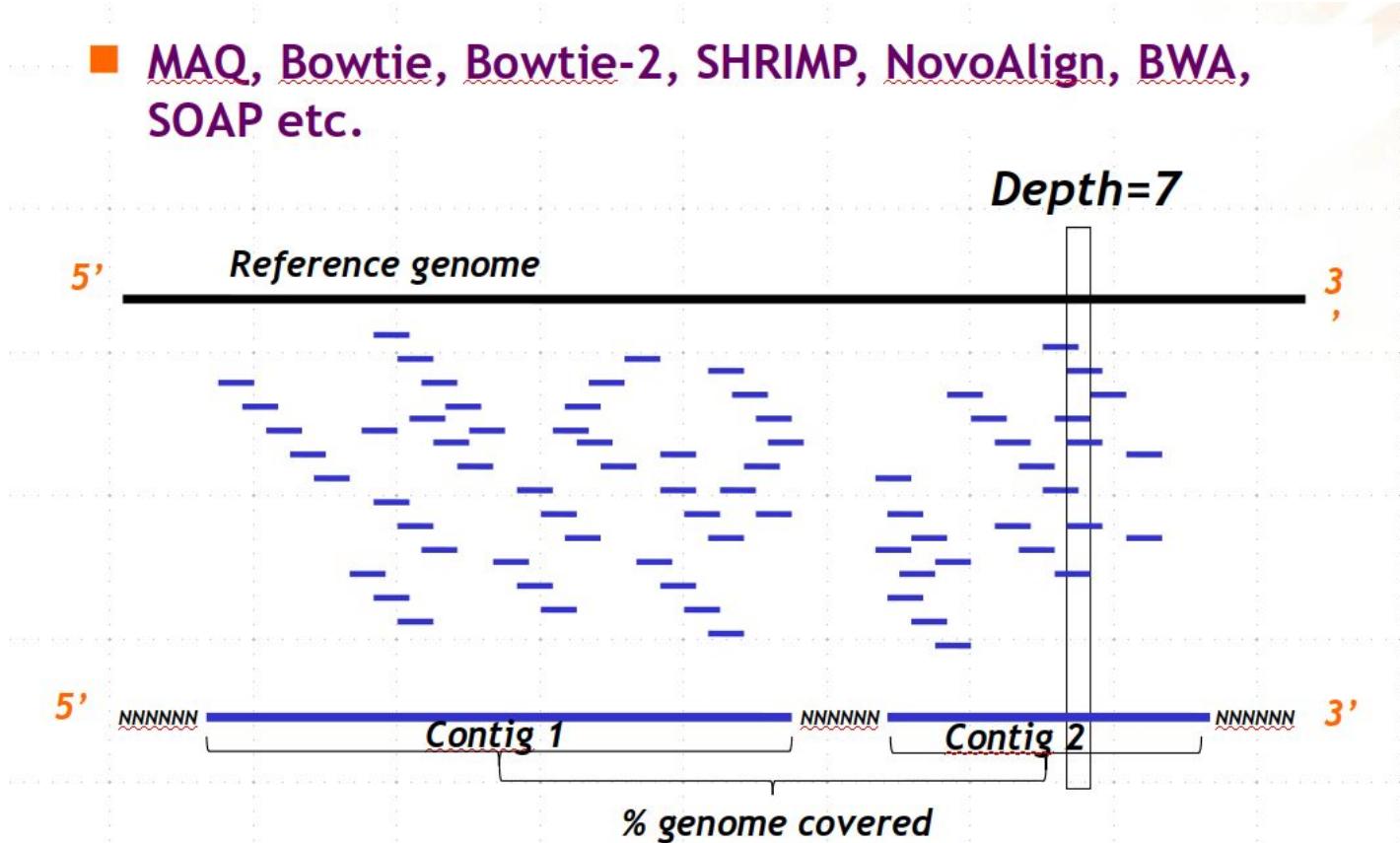
De novo assembly

- Velvet, SPAdes, ABySS, SOAP-denovo etc.
- Mostly Bruijn graph methods:
 - Reads are further broken into shorter sequence (k-mers)
 - A k-mer is then used as node to assemble overlapping nodes
- K-mer: size of the word k

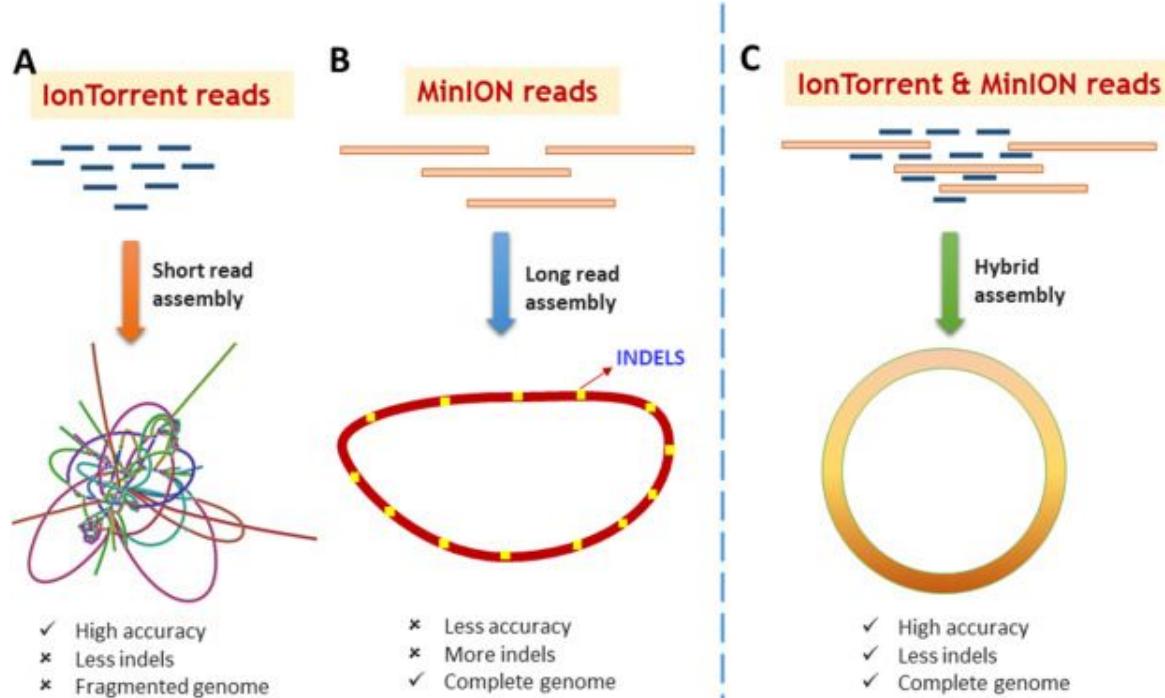


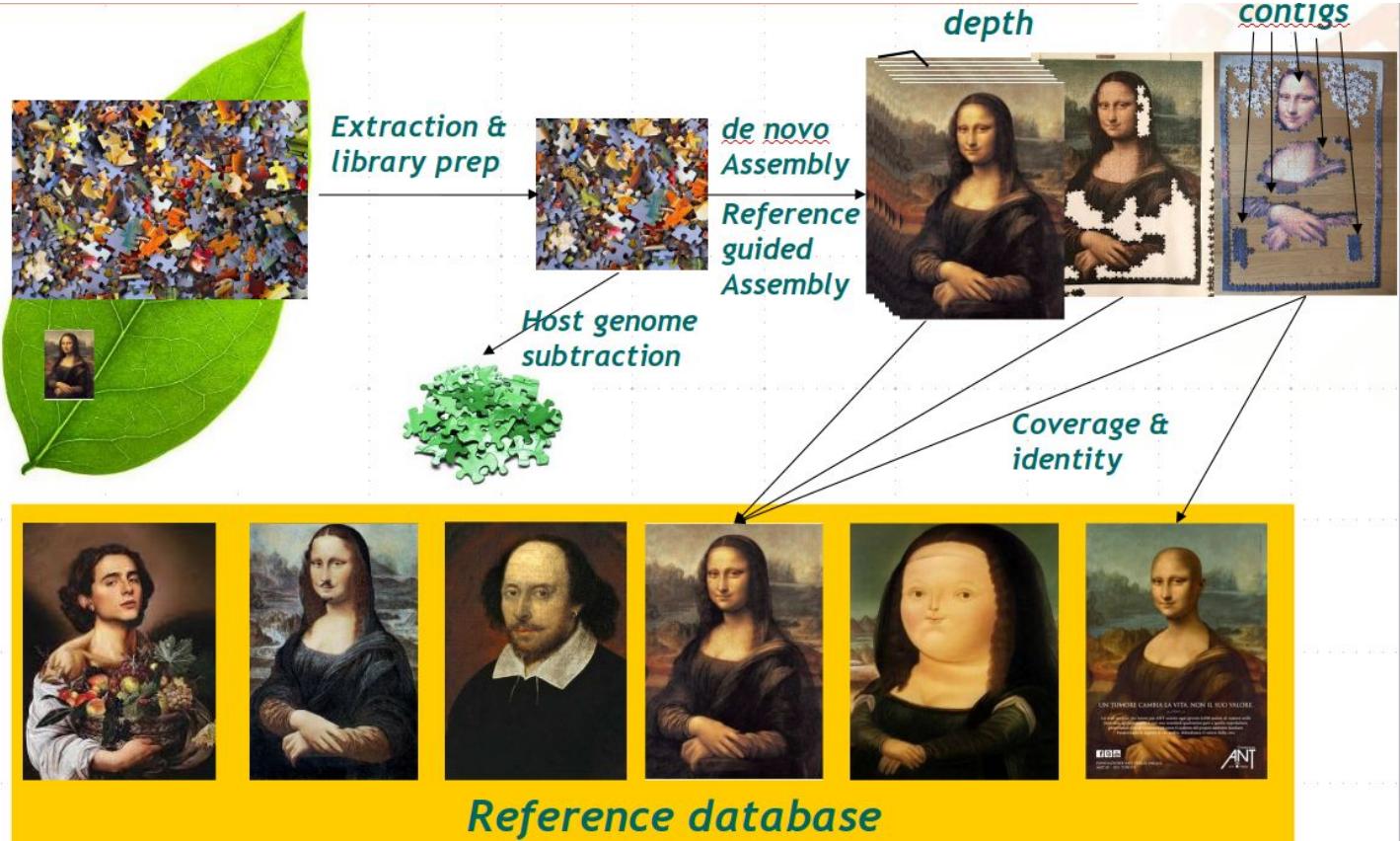
Reference based assembly

- MAQ, Bowtie, Bowtie-2, SHRIMP, NovoAlign, BWA, SOAP etc.

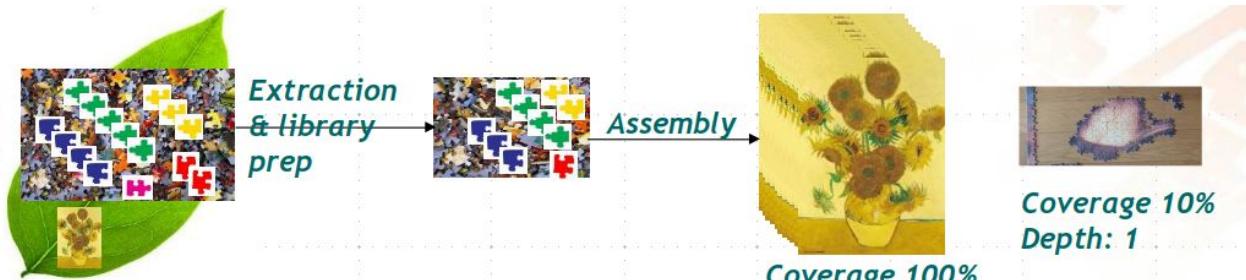


Hybrid assembly

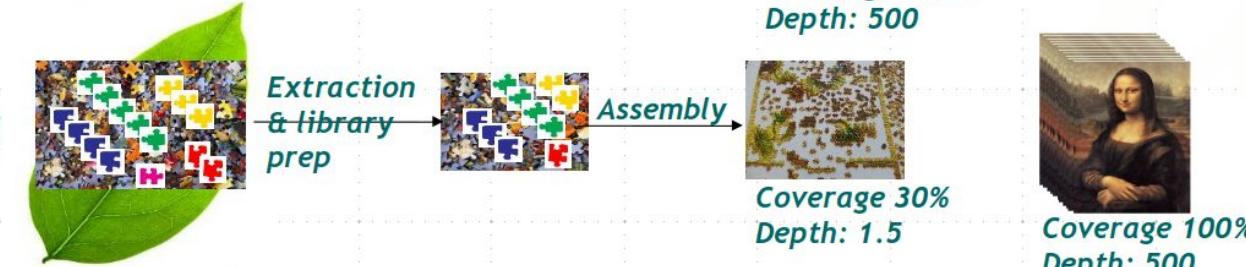




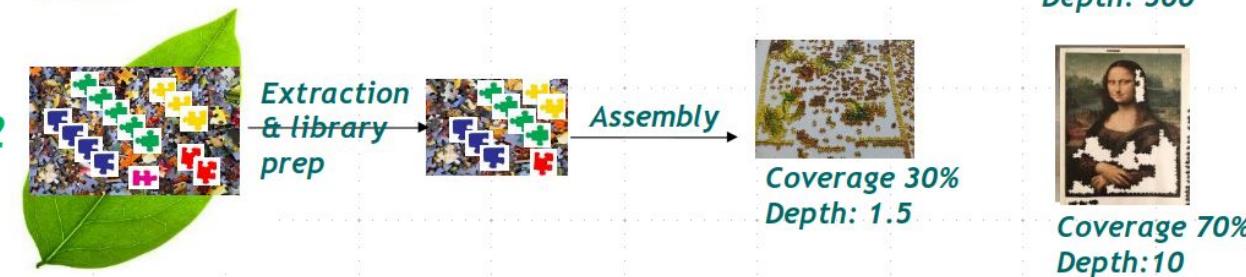
Control



Sample 1



Sample 2



Coverage 10%
Depth: 1



Coverage 100%
Depth: 500



Coverage 70%
Depth: 10

How to evaluate the quality of your assembly?

- 3Cs rule:
- Contiguity
- Correctness
- Completeness

Contiguity + Correctness + Completeness = 3Cs = CCC

How to evaluate assembly?

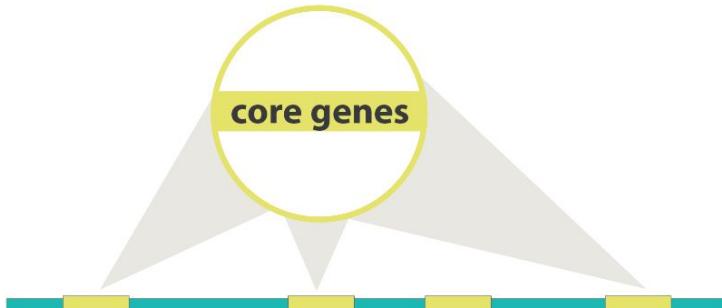
- If you have a reference genome, you can use special programs (quast and other)
- If reference genome is not present:
 - – N50
 - – NG50
 - – BUSCO/CEGMA

How to evaluate the quality of your assembly?

Contiguity



Completeness



Correctness

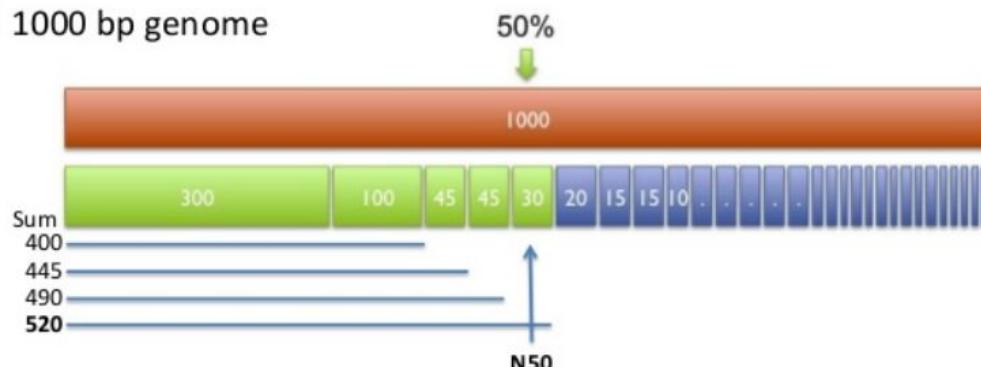
Real

T	G	C	A	G	T	A	A	C	G	A	T	T
T	G	C	A	A	T	A	A	C	T	A	T	T

Assembly

N50

50% of the genome is in contigs as large as the N50 value

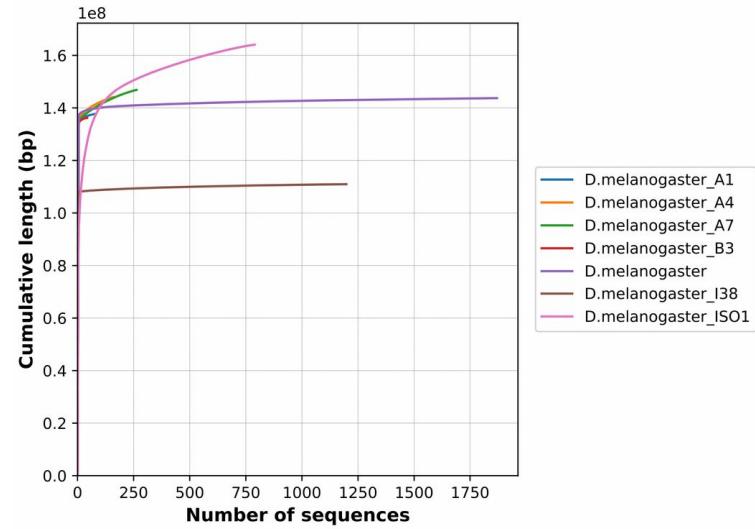
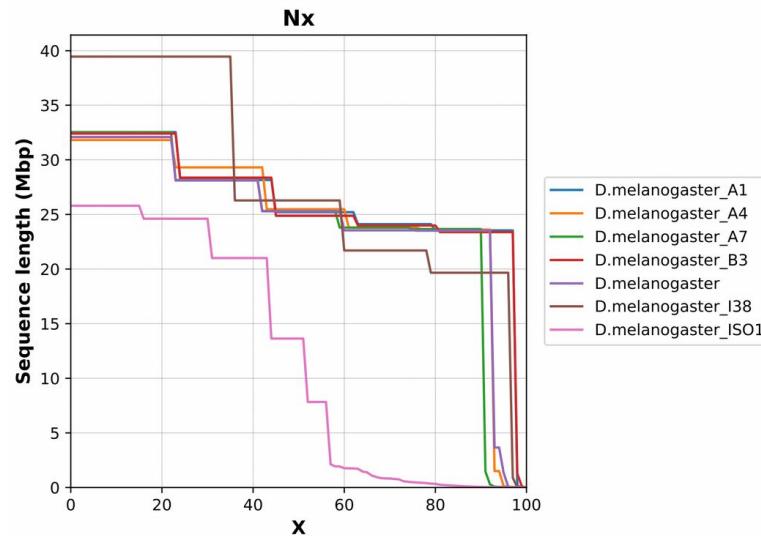


Courtesy of Michael Schatz, CSHL

NG50

- Identical to N50, but “real” genome size is used
- (not length of all contigs)

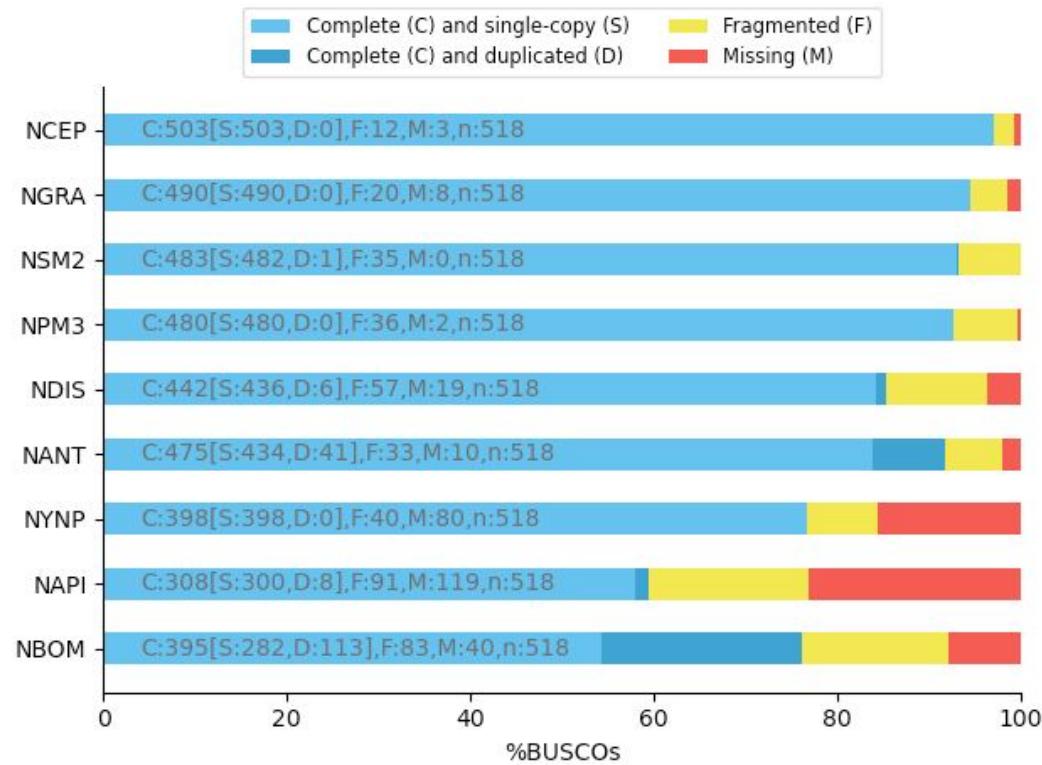
Nx and cumulative lengthcurve



BUSCO/CEGMA

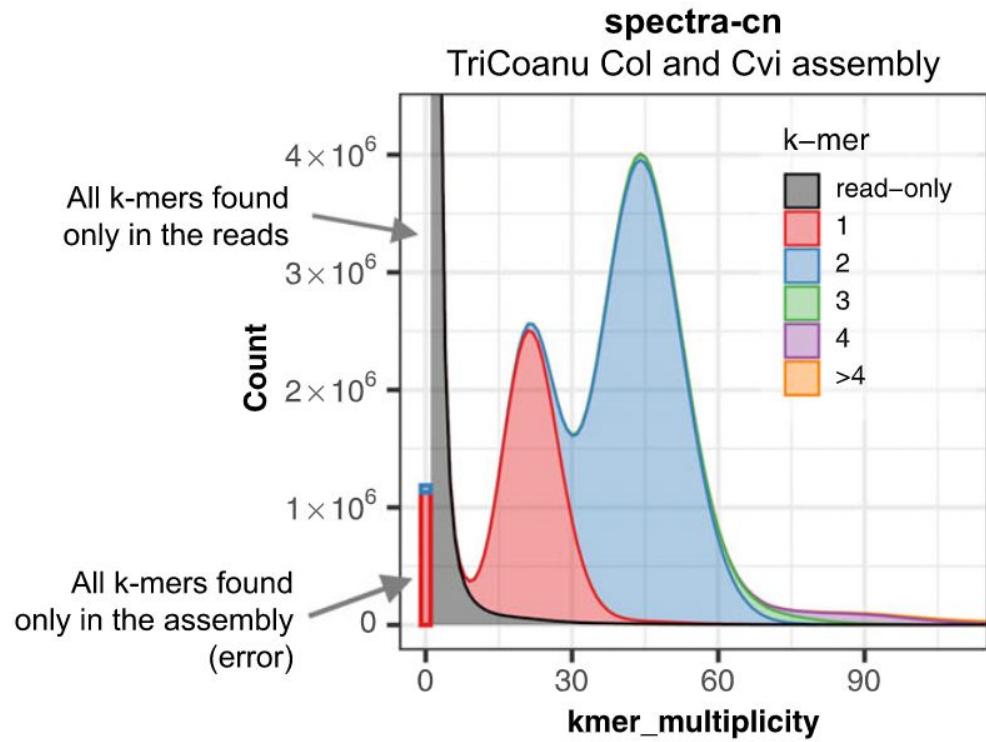
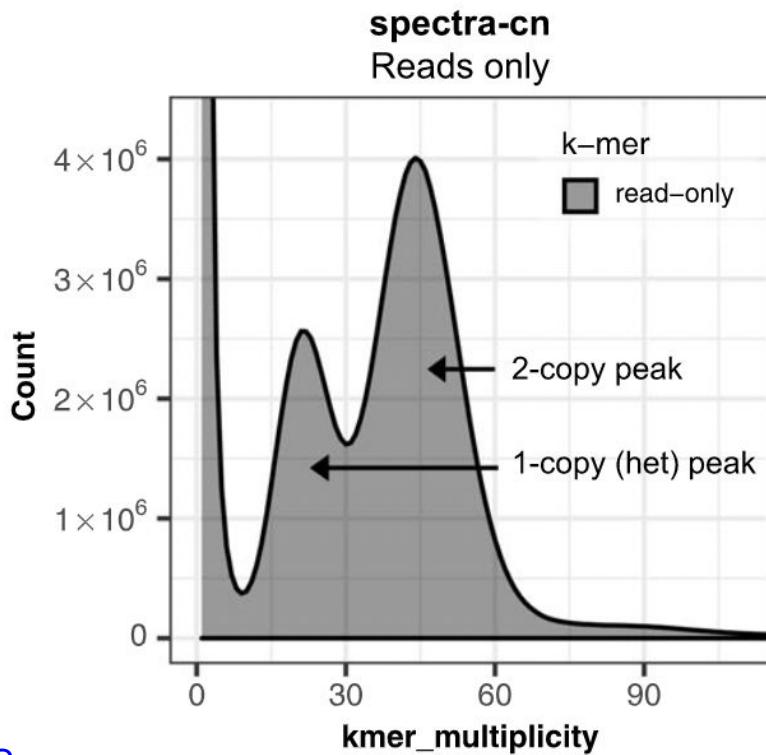
- Is it possible to make a list of genes that are common to some taxonomic groups?
- BUSCO – conserved genes list. All genes of the list are found in 90% of species of specific taxonomic group

BUSCO/CEGMA



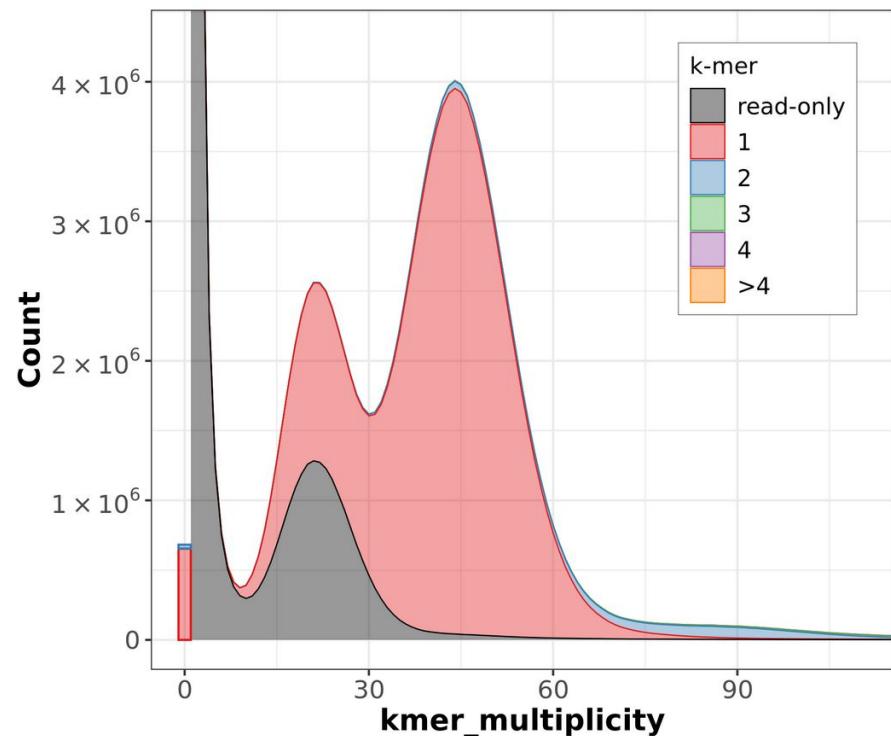
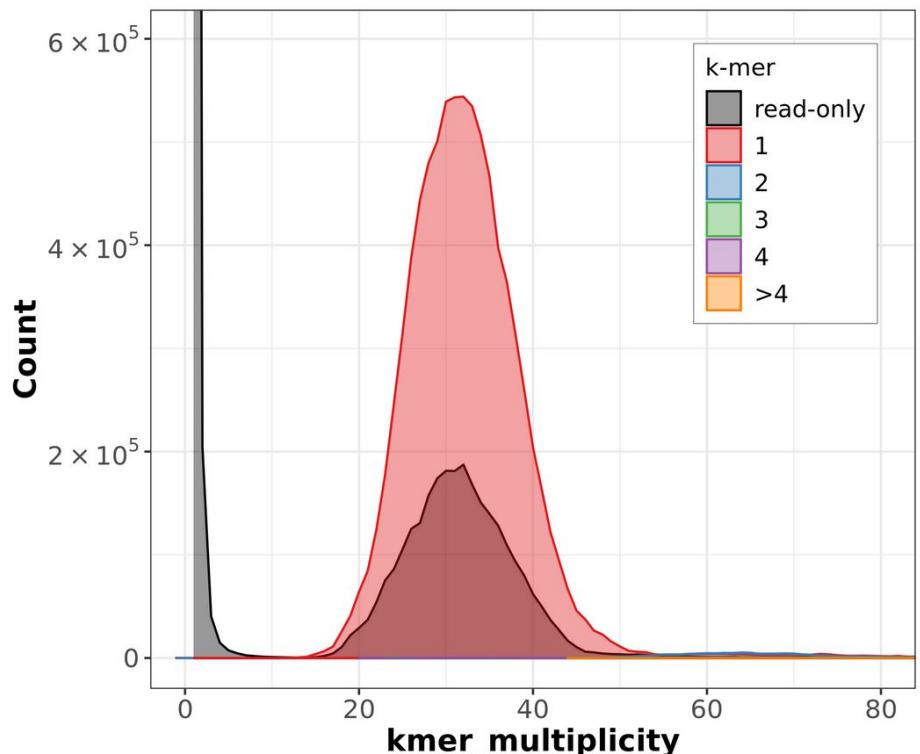
[Source](#)

K-mer content



[Source](#)

K-mer content (homo versus hetero)



[Source](#)

Correctness

Proportion of the assembly that is free from mistakes

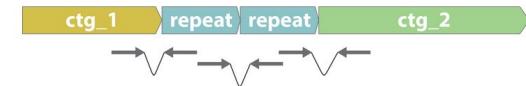
- Indels / SNPs
- Mis-joins
- Repeat compressions
- Unnecessary duplications
- Rearrangements

Assembly	T	A	C	A	G	T	A	A	C	G	A	T	T
R1	T	G	T	A	-	T	A	A	C	T	A	T	T
R2	T	G	T	A	-	T	A	A	C	T	A	T	T
R3	T	G	T	A	-	T	A	A	C	T	A	T	T
R4	T	A	T	A	-	T	A	A	C	T	A	T	T
R5	T	G	T	A	A	T	A	A	C	C	A	T	T

Correct assembly



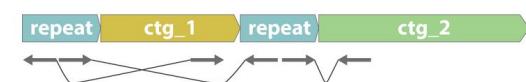
Correct assembly



Collapsed repeat



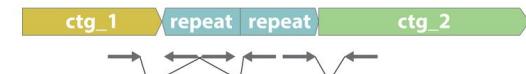
Rearrangement



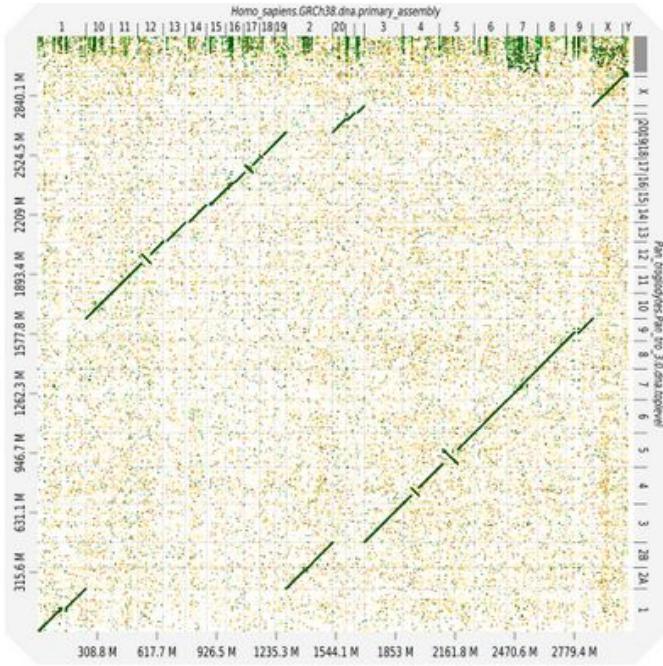
Expanded repeat



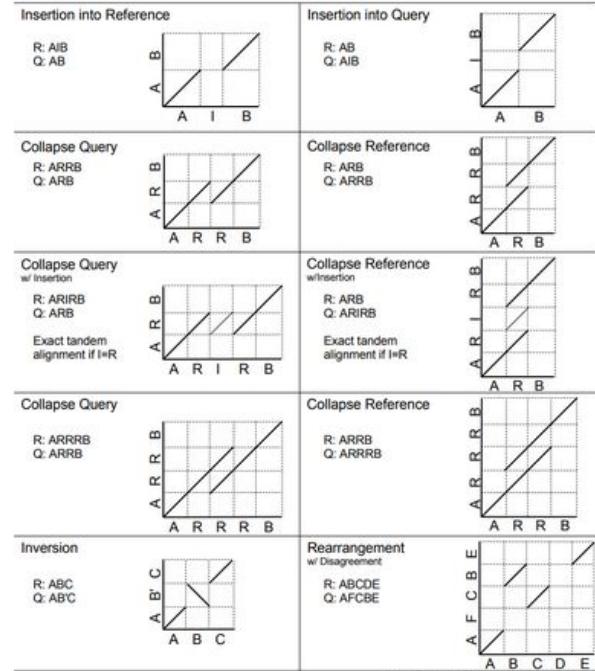
Inversion



Other things to test



QUAST
MUMmer
GenomeScope
Etc.



Michael Schatz: mschatz [at] umiacs.umd.edu

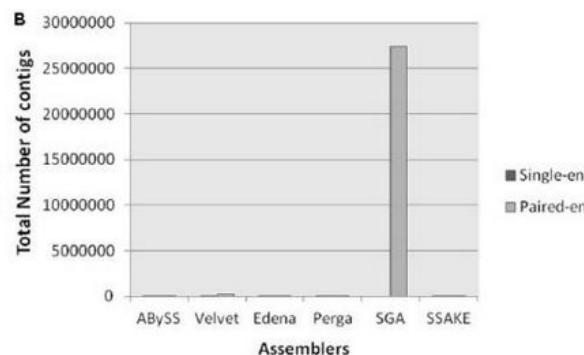
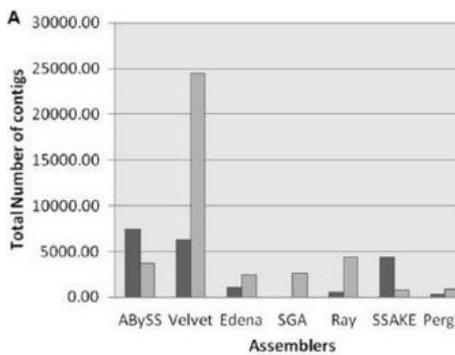
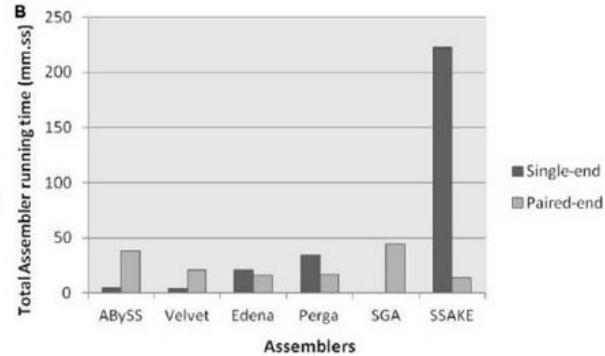
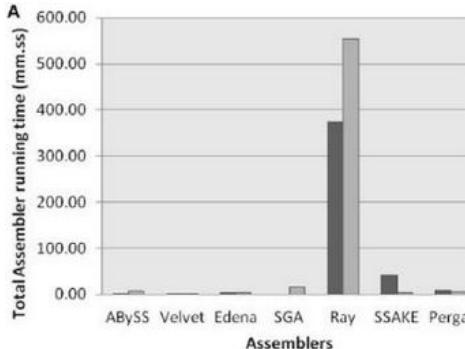
Why evaluation is not trivial?

- There is no trivial total order (i.e. ranking) between assemblies.
- > 2 independent criteria to optimize (e.g., total length, and average size of assembled sequences)
- Would you rather have an assembly with good coverage and short contigs, or an assembly with mediocre coverage and long contigs ?

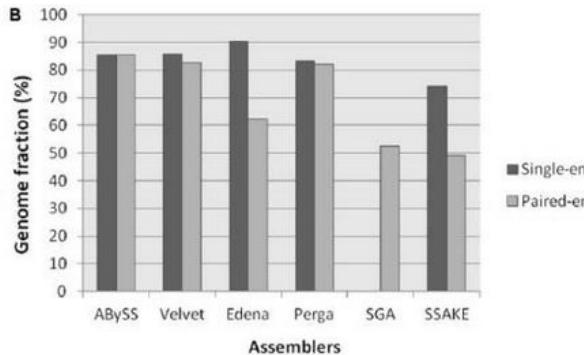
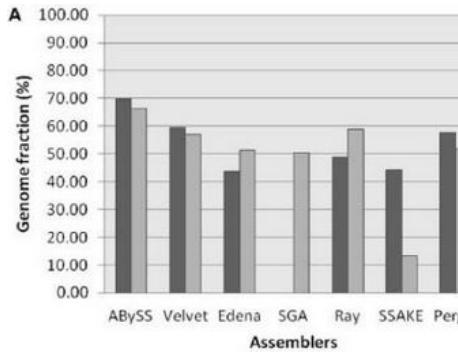
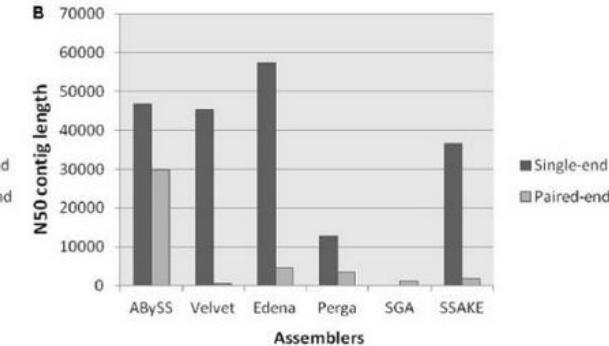
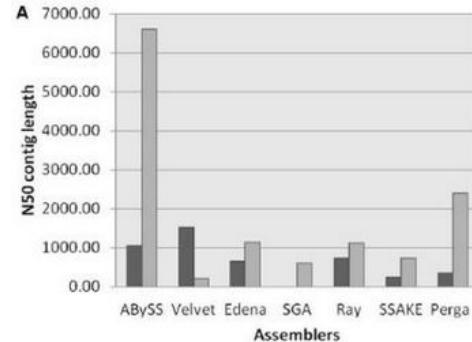
How genome is assembled? (Programs)

Assembler	Method	Err. corr.	Remarks
Euler	de Bruijn	pre-assembly	Pioneer
Velvet	de Bruijn	in-assembly	Still in use:)
ABySS	de Bruijn	in-assembly	Used widely
SOAPdenovo	de Bruijn	in-assembly	Used widely
Newbler, celera	String	in-assembly	Long reads
Ray	hybrid	in-assembly	Parallel short/long reads
SGA	String	pre-assembly	Compressed, promising
Masucra		pre-assembly	Mix-ed assembly
Trinity		pre-assembly	Transcriptome
Etc..			

Which assembler is the best?



Which assembler is the best?



Which assembler is the best?

- „In summary, in case of paired-end and single-end prokaryotic genomes, **ABySS** efficiently produced genome assembly and consumed less amount of time but consumed high amount of memory, whereas **Velvet** proved to be a time-efficient and memory-efficient program for only single-end data sets. **Edena** was a memory-efficient program for both types of data sets, and **SGA** was also a memory-efficient program, but it is only available for paired-end data.

How genome is assembled? (algorithms)

- OLC methods (overlap – layout – consensus)
- DBG methods (de Bruijn graph)

How genome is assembled? (algorithms)

- OLC:
 - Arachne
 - Cap3
 - Phusion
 - Newbler

How genome is assembled? (algorithms)

- DBG:
 - Velvet
 - ABySS
 - SOAPdenovo

OLC

• Overlap

Say $l = 3$

Look for this in Y ,
going right-to-left

X: CTCTAGGCC
Y: TAGGCCCTC

X: CTCTAGGCC

Y: TAGGCCCTC

Extend to left; in this case, we
confirm that a length-6 prefix
of Y matches a suffix of X

X: CTCTAGGCC

Y: TAGGCCCTC

Found it

We're doing this for *every pair* of input strings

OLC

- Overlap
- – It is not so primitive
- – Suffix trees are used to make/calculate Overlaps
- – Or dynamic programming
- – Or both (suffix trees and dynamic programming are used together)

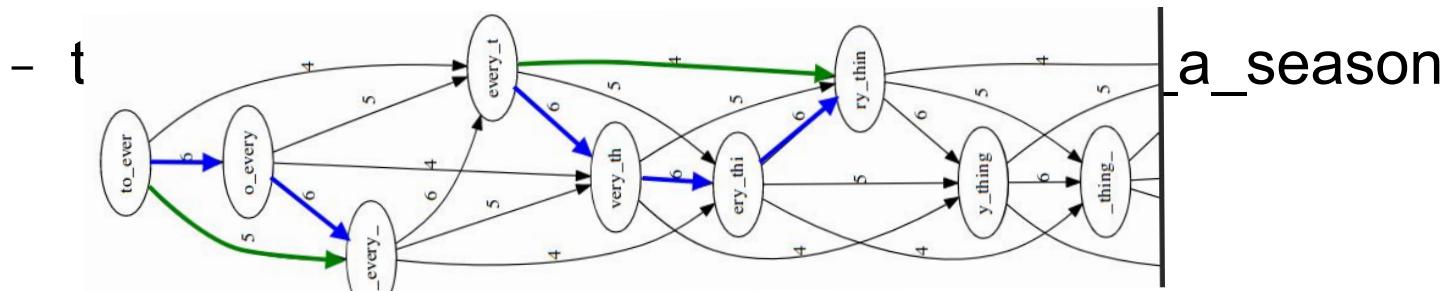
OLC

- Overlap
 - Time/computer power consuming step

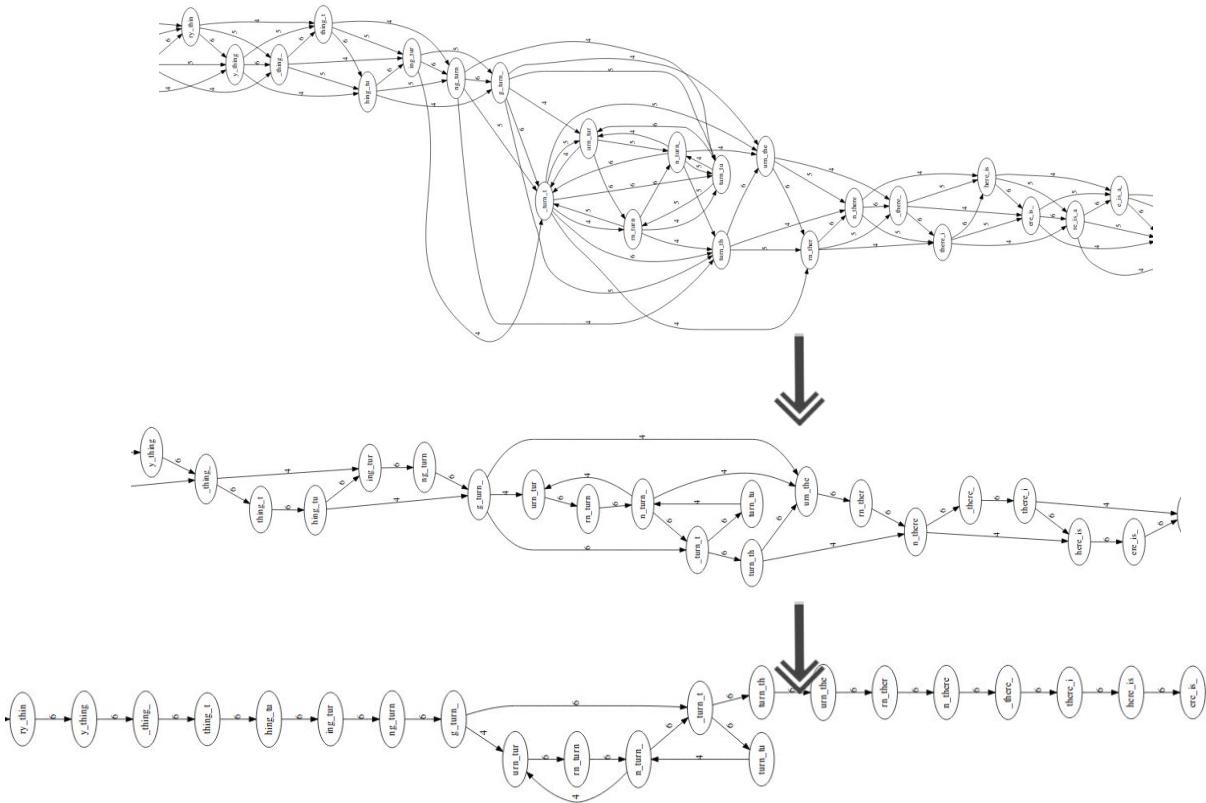
OLC

- Layout

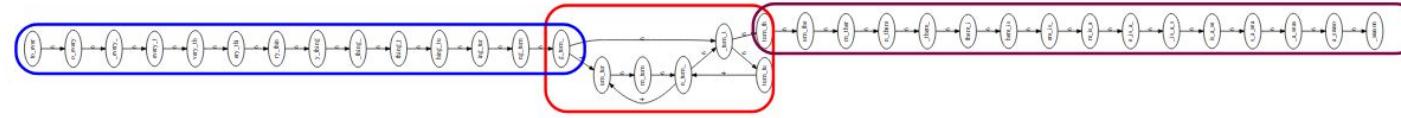
- How to make contigs? Put all overlaps into graph
- and.. rearrange it.
- Input “sequence”:



OLC



OLC



Contig 1
to_every_thing_turn_

Contig 2
turn_there_is_a_season

Unresolvable repeat

OLC

• Consensus

TAGATTACACAGATTACTGA TTGATGGCGTAA CTA
TAGATTACACAGATTACTGACTTGATGGCGTAAACTA
TAG TTACACAGATTATTGACTTCATGGCGTAA CTA
TAGATTACACAGATTACTGACTTGATGGCGTAA CTA
TAGATTACACAGATTACTGACTTGATGGCGTAA CTA

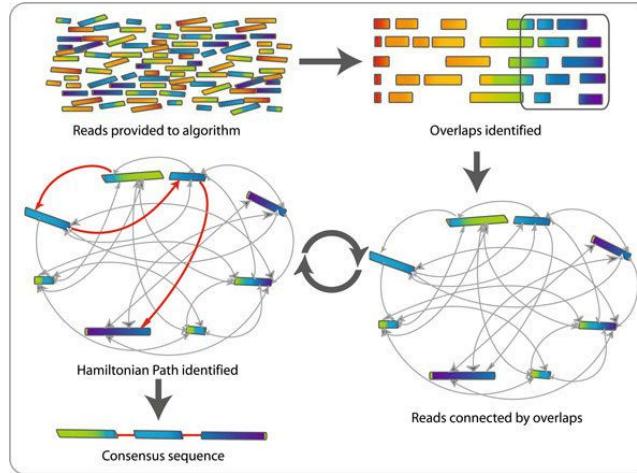


Take reads that make up a contig and line them up

Take *consensus*, i.e. majority vote

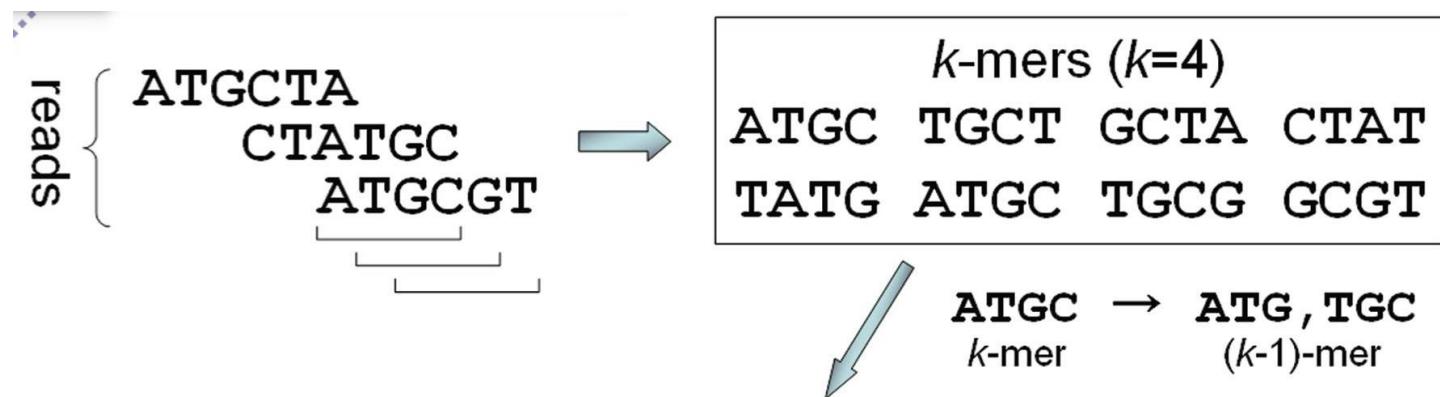
OLC

- Cons:
 - Overlaps step takes a lot of time
 - Overlay graph is huge



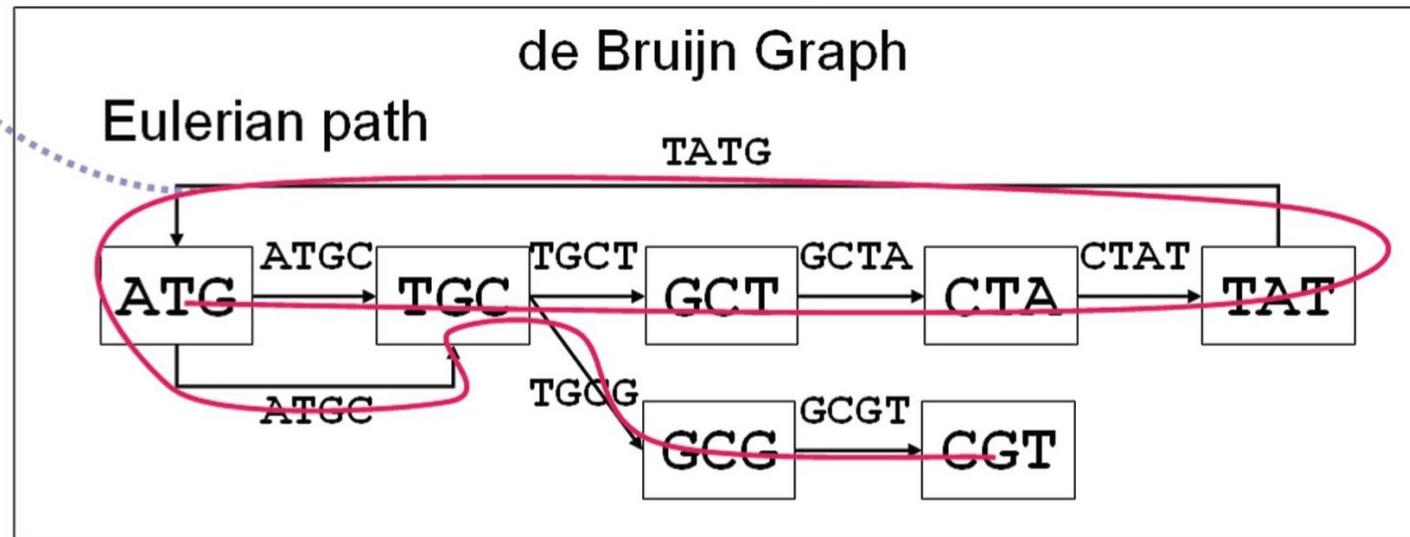
DBG

- Uses k-mers
- Uses k-1mers (left and right)



DBG

- Take each k-1-mer and make a graph:



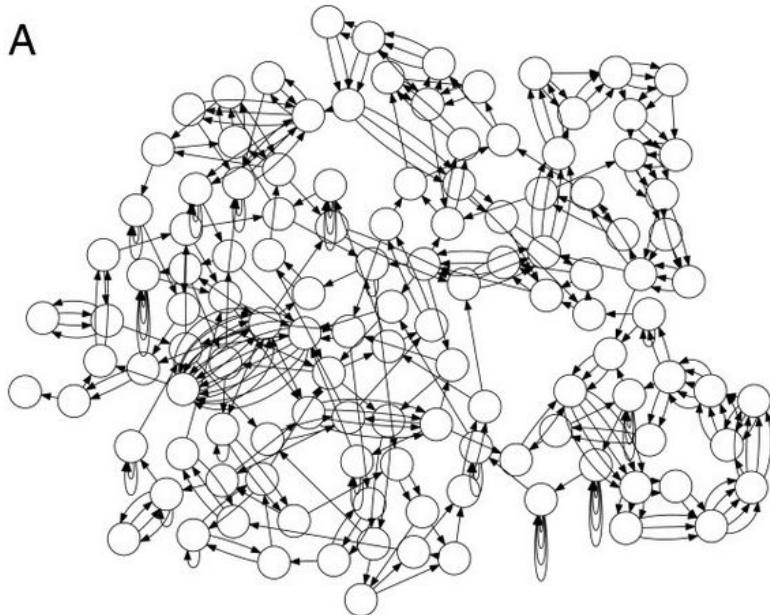
How to assemble genome?

Find Eulerian path.

In graph theory, an Eulerian trail (or Eulerian path) is a trail in a finite graph which visits every edge exactly once. Similarly, an Eulerian circuit or Eulerian cycle is an Eulerian trail which starts and ends on the same vertex. (wikipedia)

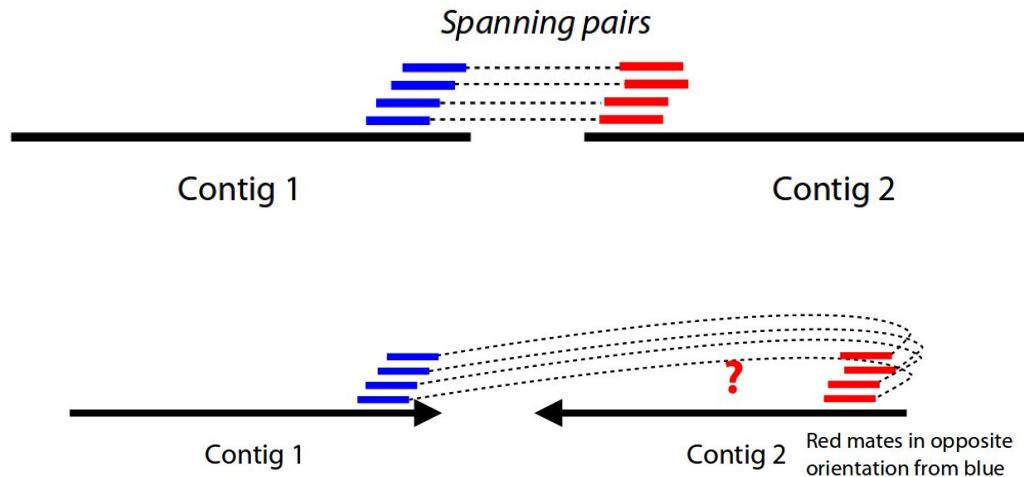
And final result is:

A



Scaffolds

- Orientations of individual contigs in genome space:
 - PE data:



What is next?

- Finalize genome (in many cases you have to do additional sequencing)
- Genome annotation

Some toy exercises:

- . You have reads: TACAGT, CAGTC, AGTCA, CAGA
 - How many k -mers are in these reads (including duplicates), for $k = 3$?
 - How many distinct k -mers are in these reads ? ($k=2,3,5$)
 - It appears that these reads come from the (toy) genome TACAGTCAGA. What is the largest k such that the set of k -mers in the genome is exactly the set of k -mers in these reads ?
 - For any value of k , what is a mathematical relation between N , the number of k -mers (incl. duplicates)in a sequence, and L , the length of that sequence? $N = L - k + 1$

Some toy exercises (answers)

- . You have reads: TACAGT, CAGTC, AGTCA, CAGA
 - How many k -mers are in these reads (including duplicates), for $k = 3$? **12**
 - How many distinct k -mers are in these reads ? ($k=2,3,5$) **7, 7, 4**
 - It appears that these reads come from the (toy) genome TACAGTCAGA. What is the largest k such that the set of k -mers in the genome is exactly the set of k -mers in these reads ? **3**
 - For any value of k , what is a mathematical relation between N, the number of k -mers (incl. duplicates)in a sequence, and L, the length of that sequence ?

Toy exercise II

Here are two assemblies, aligned to the same reference :



- For each, compute the following metrics :
 - Total size of the assembly, N50, NG50 (bp)
 - Coverage (%)
 - Which one is better than the other ?

Toy exercise II (solutions)

Here are two assemblies, aligned to the same reference :



- For each, compute the following metrics :
 - Total size of the assembly (19 bp, 18 bp), N50 (6 bp, 9 bp), NG50 (6 bp, 5 bp)
 - Coverage (%) (90, 90)
 - Which one is better than the other ? (I would say first one)

Literature

- <https://www.youtube.com/playlist?list=PL2mpR0RYFQsBiCWVJSvVAO3OJ2t7DzoHA>
- Comparison of the two major classes of assembly algorithms: overlap layout consensus and de-bruijn-graph (Li et al., 2011)
- <http://www.langmead-lab.org/teaching-materials/>

-

B. Assembly algorithms

Eulerian path-based algorithms

- Greedy algorithms
- Hybrid algorithms

C. Challenges in de novo assembly

- Repeats and complex regions
- Heterozygosity
- Polyploidy

IV. Reference-based assembly

A. Mapping reads to a reference genome

- Single nucleotide polymorphism (SNP) calling
- Structural variant (SV) calling

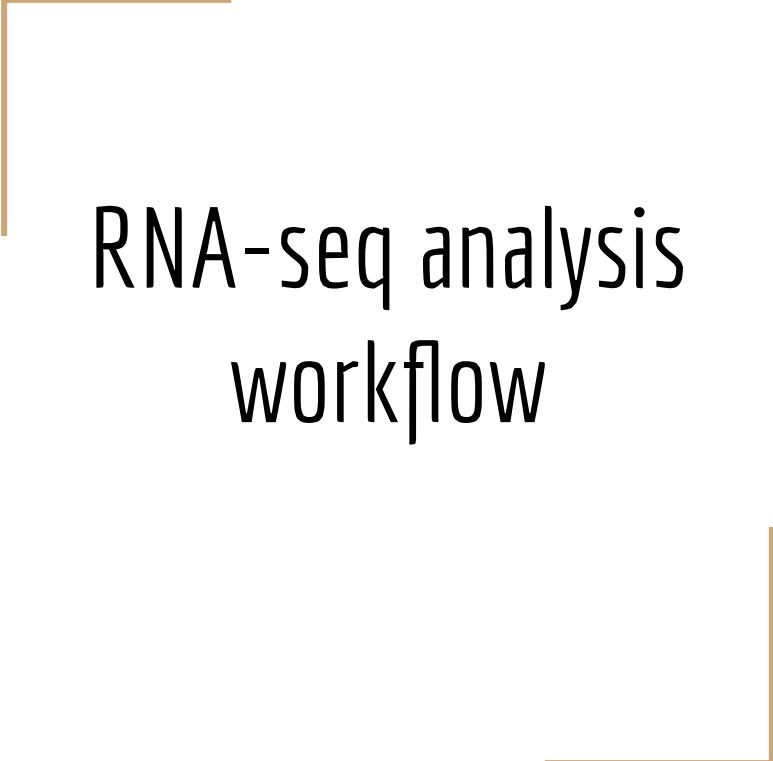
B. Improving reference-based assembly

- Gap filling
- Polishing

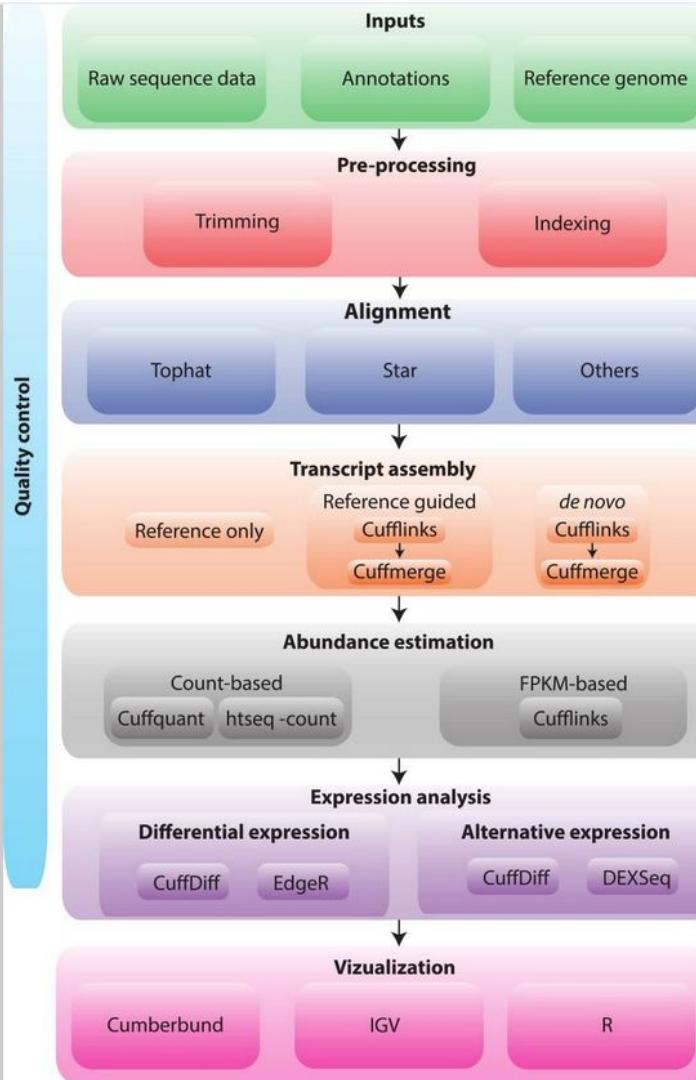
V. Hybrid assembly

A. Advantages of hybrid assembly

- Combining strengths of de novo and reference-based approaches
- Better coverage and accuracy



RNA-seq analysis workflow



RNA-seq analysis: data preparation/analysis

RNA-seq analysis flow

Each experiment will require some adaptation, but there are some general rules/steps:

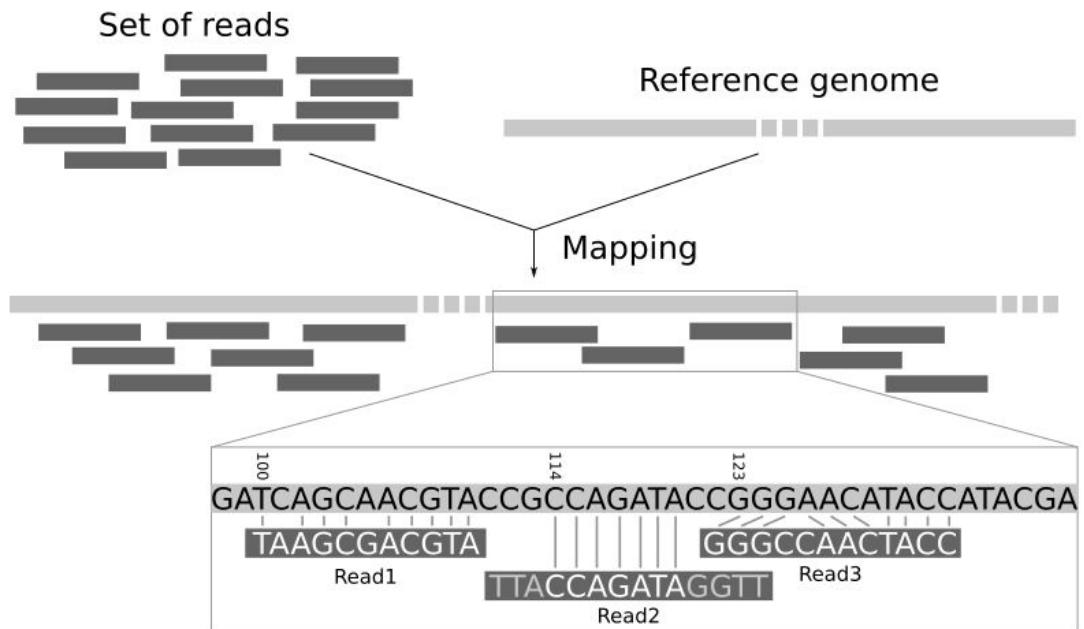
1. Get the data and do QA/QC
2. Assemble/Align reads and do QA/QC
3. Use specific tools (depends on the aim)
4. Do statistical analysis (R, Matlab, etc.)
5. Summarize, visualize and present

Data QA/QC

- Get the data (normally some kind of ftp download)
- Check integrity (md5 if available)
- Make a backup (on a different physical disk)
- Ask for: sequencing machine name and software (name, version, pipeline) used to prepare the data (data submission/publications)
- Check quality (clean if needed)
- Useful tools: FASTQC, trimmomatic, cutadapt, SeqPurge, etc.

Mapping

- What is mapping?
- Why mapping is important?
- How to do mapping?



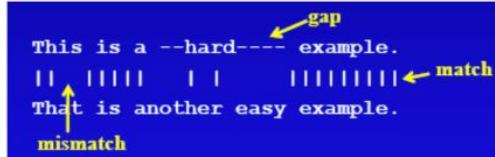
Mapping

- Mapping
 - A mapping is a region where a read sequence is placed.
 - A mapping is regarded to be correct if it overlaps the true region.
- Alignment
 - An alignment is the detailed placement of each base in a read.
 - An alignment is regarded to be correct if each base is placed correctly.

What is Pairwise Sequence Alignment?

Sequence alignment is a method of arranging biological sequences to identify regions of similarity. The similarity being identified, may be a result of functional, structural, or evolutionary relationships between the sequences.

Pairwise sequence alignment is one form of a sequence alignment technique, where we compare only two sequences. This process involves finding the optimal alignment between the two sequences, scoring based on their similarity (how similar they are) or distance (how different they are), and then assessing the significance of this score.



The basis of sequence alignment lies with the scoring process, where the two sequences are given a score on how similar (or different) they are to each other. The pairwise sequence alignment algorithms require a scoring matrix to keep track of the scores assigned. The scoring matrix assigns a positive score for a match, and a penalty for a mismatch.

Three basic aspects are considered when assigning scores. They are,

1. **Match value**—Value assigned for matching characters
2. **Mismatch value**—Value assigned for mismatching characters
3. **Gap penalty**—Value assigned for spaces

Mapping

- Problems:
 - Reads are short (30-500 bp)
 - Genome is big (3G *H. sapiens*)
 - Genome may have introns (eukaryotes + rna-seq)
 - Genome may have repeats
 - Time
 - Memory

How to choose a mapper?

- Alignment algorithm
- Filtering/parameters
- Multimapping
- INDELS, clipping
- Splicing

Widely used mappers

- BWA (genomic, no splicing)
- Bowtie2 (genomic, no splicing)
- hisat2 (splicing+)
- STAR (splicing+)

Quantification

aligned read:
start: 113217600 end: 113217650



GTF

```
chr1 unknown exon 113217048 113217252 . + . gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
chr1 unknown exon 113217048 113217351 . + . gene_id "MOV10";p_id "P5535";transcript_id "NM_020963"
chr1 unknown exon 113217470 113217671 . + . gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
chr1 unknown CDS 113217535 113217671 . + 0 gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
chr1 unknown start_codon 113217535 113217537 . + gene_id "MOV10";p_id "P5535";transcript_id "NM_001130079"
```



feature type

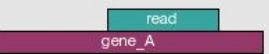
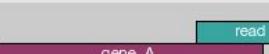
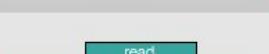


feature

Output of counting = A count matrix, with genes as rows and samples are columns

Quantification: things to consider

- Multimapping reads
- Reads overlapping several fragments
- Isoforms quantification (maybe different methods?)
- Mapping quality

	union	intersection _strict	intersection _nonempty
 A single read (green) overlaps a single gene (purple). This is a simple case of union.	gene_A	gene_A	gene_A
 A single read (green) overlaps two genes (purple). This is a case where intersection strict fails.	gene_A	no_feature	gene_A
 A single read (green) overlaps two genes (purple), but the genes overlap each other. This is a case where intersection nonempty fails.	gene_A	no_feature	gene_A
 Two reads (green) overlap two genes (purple). Both genes are assigned to both reads.	gene_A	gene_A	gene_A
 A single read (green) overlaps two genes (purple and blue). This is a case of ambiguity.	gene_A	gene_A	gene_A
 Two reads (green) overlap two genes (purple and blue). One gene is assigned to both reads.	ambiguous	gene_A	gene_A
 Two reads (green) overlap two genes (purple and blue). Both genes are assigned to both reads.	ambiguous	ambiguous	ambiguous

[Source](#)

Quantification programs

- FeatureCounts
- HTSeq-counts
- stringtie

Quantification: output

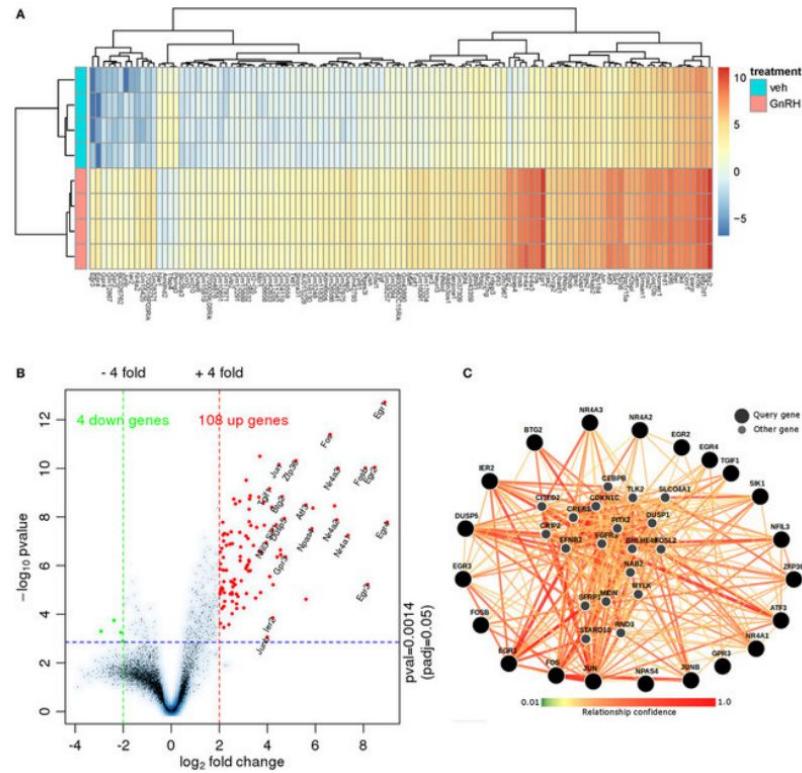
- An expression matrix (or a file that may be converted to an expression matrix) containing gene counts/abundance/similar data

Reference based quantification

- Reference genome is used often (if available)
- Gene annotations may help us
- Gene annotations have specific file format(s): gtf/gff/etc

Data analysis

- DE analysis
- New transcripts identification
- New splicing
- Biological “meaning”
- Etc.



DOI: 10.3389/fendo.2018.00034

To sum up

RNA-seq data bioinformatics analysis:

1. Get data (output - FASTQ files)
2. Do data QC/QA (input FASTQ, output FASTQ)
3. Map data, do QC/QA (input FASTQ, output SAM/BAM)
4. Quantify data (input BAM, output matrix)
5. Perform statistical analysis (input matrix)

GUI programs for NGS

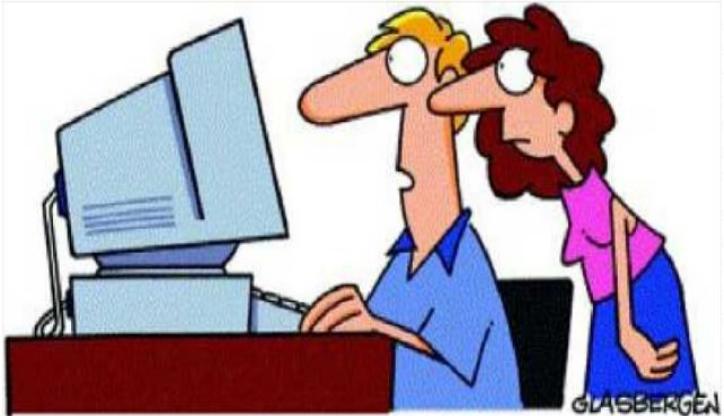
Lecture layout

- Bioinformatics workflow management systems
- What is GALAXY?
- Main principles of GALAXY
- ChIP-seq with GALAXY

Most popular systems

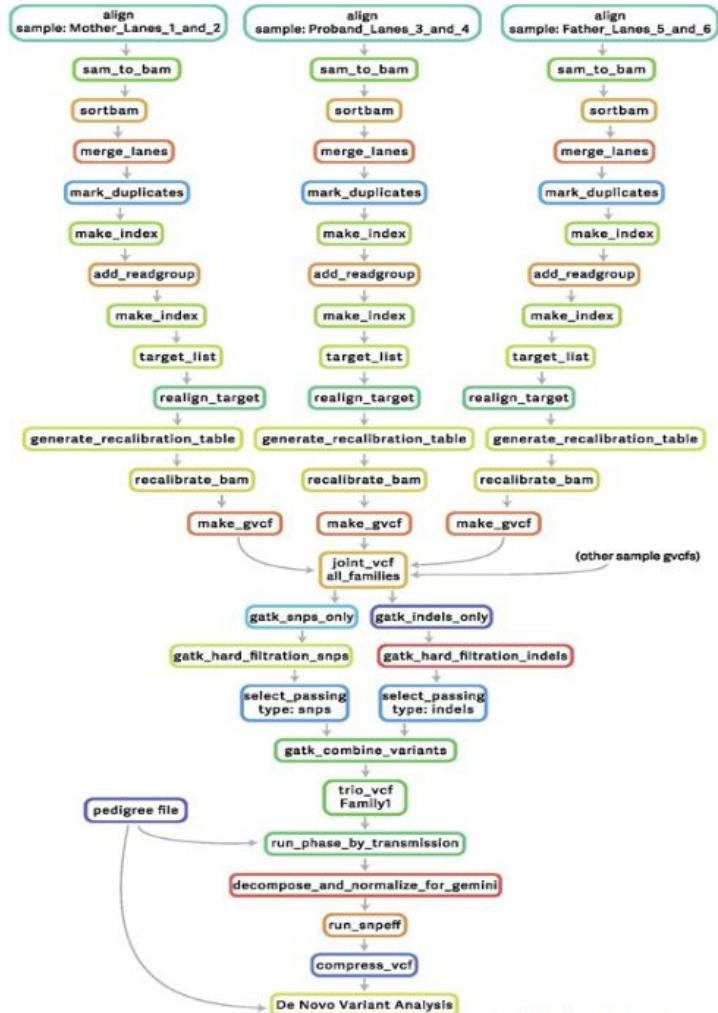
- GALAXY <https://usegalaxy.org/>)
- KNIME <https://www.knime.com/>)
- Taverna <https://taverna.incubator.apache.org/>)
- Yabi <https://ccgapps.com.au/yabi/login/?next=/yabi/>)
- UGENE <http://ugene.net/>)
- etc.

Why such systems are useful?

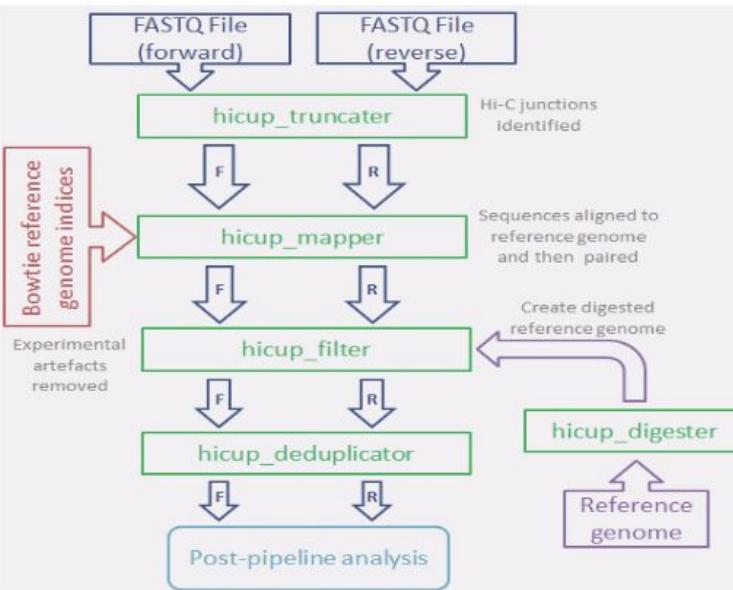


"The computer says 3 need to upgrade my fa rain to be compatible with mrcnarray data analysis."

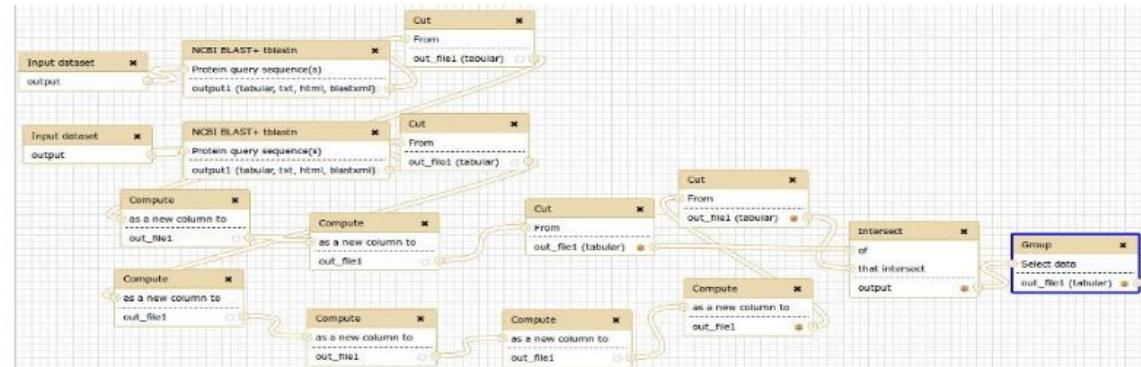
Leung, 2002



Front-end UI



<https://www.bioinformatics.babraham.ac.uk/projects/hicup/flow.png>



<http://www.bioinf.uni-freiburg.de/Galaxy/workflow-blast-902x356.jpg>

Alternatives

- Scripts
- Make (Make utility)
- Implicit convention frameworks (Snakemake, Nextflow)
- Explicit frameworks
- Configuration frameworks
- Class-based frameworks

How to choose a pipeline?

- Requirements of the users
- Requirements of the developer(s)
- Reusability
- Adaptation to cloud computing/parallelisation

GALAXY



<https://www.nasa.gov/image-feature/an-infrared-view-of-the-m81-galaxy>

GALAXY

Galaxy

Analyze Data Workflow Visualize ▾ Shared Data ▾ Help ▾ User ▾

Using 0%

Tools

search tools

Get Data

Send Data

Collection Operations

Expression Tools

GENERAL TEXT TOOLS

Text Manipulation

Filter and Sort

Join, Subtract and Group

Datamash

GENOMIC FILE MANIPULATION

FASTA/FASTQ

FASTQ Quality Control

SAM/BAM

BED

History

search datasets

Unnamed history

1 deleted

1.67 GB

This history is empty. You can load your own data or get data from an external source

Galaxy 101
an introduction to Galaxy tutorial

Galaxy Training Network

Tweets by @galaxyproject

Galaxy Project
@galaxyproject

Doc ▾ Training ▾ Updates ▾ data types ▾ citing

The screenshot displays the Galaxy web application interface. At the top, there's a dark header bar with the 'Galaxy' logo, navigation links like 'Analyze Data', 'Workflow', 'Visualize', 'Shared Data', 'Help', 'User', and a grid icon. A progress bar on the right shows 'Using 0%'. The left sidebar contains a 'Tools' section with a search bar, followed by categorized lists: 'Get Data', 'Send Data', 'Collection Operations', 'Expression Tools', 'GENERAL TEXT TOOLS', 'Text Manipulation', 'Filter and Sort', 'Join, Subtract and Group', 'Datamash', 'GENOMIC FILE MANIPULATION', 'FASTA/FASTQ', 'FASTQ Quality Control', 'SAM/BAM', and 'BED'. The main content area features a central text block about Galaxy's purpose, a 'Galaxy 101' tutorial thumbnail, and a 'Tweets' section from the Galaxy Project's Twitter account. The right sidebar shows an empty history panel with a message encouraging users to load their own data.

What is Galaxy?

- WEB page
- Simple GUI
- HTS (and other data) analysis
- Open Source (academic free license)

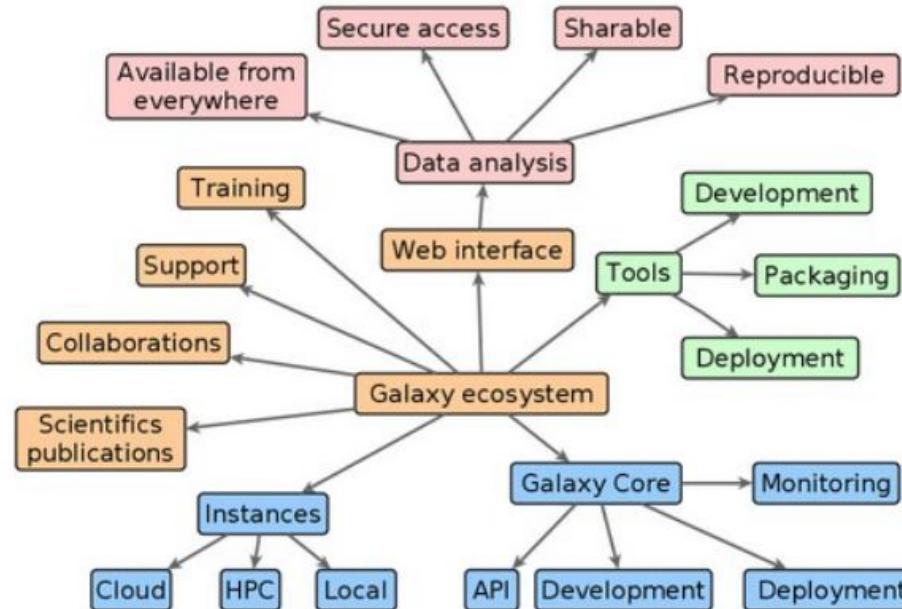
What is Galaxy?

- WEB page
- Reproducibility
- HTS (and other data) analysis
- Open Source (academic free license)
- Accessibility
- Simple GUI
- Transparency

What is Galaxy?

Galaxy ecosystem

Administrators All Developers Users



Main Galaxy page

Galaxy

Analyze Data Workflow Visualize Shared Data Help User Using 0%

Tools  

search tools 

Get Data
Send Data
Collection Operations
Expression Tools
GENERAL TEXT TOOLS
Text Manipulation
Filter and Sort
Join, Subtract and Group
Datamash
GENOMIC FILE MANIPULATION
FASTA/FASTQ
FASTQ Quality Control
SAM/BAM
BED

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.


Galaxy 101
an introduction to Galaxy tutorial
Galaxy Training Network

Tweets by @galaxyproject

Galaxy Project
@galaxyproject
Doc + Training Updates: data types, citing

History    

search datasets 

Unnamed history
1 deleted
1.67 GB  

This history is empty. You can load your own data or get data from an external source

Galaxy usage

The screenshot shows the Galaxy web interface. At the top, there's a navigation bar with links for Analyze Data, Workflow, Shared Data, Visualization, Help, and User. Below the navigation bar, a main content area displays a welcome message about Galaxy being an open source platform for data intensive biomedical research. It includes links to start here, help resources, and the Tool Shed. A central banner for BioStars (GALAXY EXPLAINED) features the text "Want help? Get answers." Below the banner, logos for various institutions are displayed: Penn State, Johns Hopkins, Oregon Health & Science, TACC, and CyVerse. To the right, a "History" panel shows an empty history with a note about loading data. On the left, a sidebar titled "Tools" lists numerous bioinformatics tools categorized under "Text Manipulation", "Data mash", "Convert Formats", "Filter and Sort", "Join, Subtract and Group", "Fetch Alignments/Sequences", "NGS: QC and manipulation", "NGS: DeepTools", "NGS: Mapping", "NGS: RNA Analysis", "NGS: SAMtools", "NGS: BamTools", "NGS: Picard", "NGS: VCF Manipulation", "NGS: Peak Calling", "NGS: Variant Analysis", "NGS: RNA Structure", "NGS: Du Novo", "NGS: Gemini", "NGS: Assembly", "NGS: Chromosome Conformation", and "NGS: Mothur".

Tools

Main

History

Tools

Tools



search tools

[Get Data](#)

[Send Data](#)

[Collection Operations](#)

[Expression Tools](#)

[GENERAL TEXT TOOLS](#)

[Text Manipulation](#)

[Filter and Sort](#)

[Join, Subtract and Group](#)

[Datamash](#)

[GENOMIC FILE MANIPULATION](#)

[FASTA/FASTQ](#)

[FASTQ Quality Control](#)

[SAM/BAM](#)

[BED](#)

[VCF/BCF](#)

[Nanopore](#)

[Convert Formats](#)

[Lift-Over](#)

[COMMON GENOMICS TOOLS](#)

[Operate on Genomic Intervals](#)

[Fetch Sequences/Alignments](#)

[GENOMICS ANALYSIS](#)

History/Files

The screenshot shows a mobile application interface for managing datasets. At the top, there's a navigation bar with a back arrow, a search icon, and a user profile icon. Below the bar, the word "History" is displayed in large letters. To the right of "History" are icons for refresh, add, delete, and settings. A search bar with the placeholder "search datasets" is positioned below the navigation. The main content area is titled "Unnamed history". It shows a summary: "1 deleted" and "1.67 GB". To the right of this summary are two small icons: a heart and a speech bubble. A blue callout box with an information icon contains the text: "This history is empty. You can load your own data or get data from an external source".

Work zone

Galaxy is an open source, web-based platform for data intensive biomedical research. If you are new to Galaxy start here or consult our help resources. You can install your own Galaxy by following the tutorial and choose from thousands of tools from the Tool Shed.



Tweets by @galaxyproject

Galaxy Project
@galaxyproject
Doc + Training Updates: data types, citing
@gxytraining tutorials, Galaxy on @Windows, Account
Creation Help, #Metatranscriptomics, & #UseGalaxy
for #everyone galaxyproject.org/news/2019-12-g...

Hub, and Training -
Training Data Types
Galaxy
A comprehensive
joumey to define data types
in Galaxy

Cite GTN Tutorials
Thanks to Helena Blazquez,
every GTN tutorial now
includes a 'Citing this'
section, including a
BibTeX citation for the
tutorial.

Documentation on how
to run Galaxy on Windows
has been greatly
enhanced by Tom
Rognes to use the
Windows Subsystem for
Linux.

Windows

View on Twitter

Embed

Tools menu

HISAT2 as an example

HISAT2 A fast and sensitive alignment program
(Galaxy Version 2.1.0+galaxy5)

☆ Favorite

Versions

Options

Source for the reference genome

Use a built-in genome

Built-in references were created using default options

Select a reference genome

A. mellifera 04 Nov 2010 (Amel_4.5/apiMeI4) (apiMeI4)

If your genome of interest is not listed, contact the Galaxy team

Is this a single or paired library

Single-end

FASTA/Q file



No fastqsanger, fastqsanger.gz, fastqsanger.bz2 or fasta dataset av...



Must be of datatype "fastqsanger" or "fasta"

Specify strand information

Unstranded

'F' means a read corresponds to a transcript. 'R' means a read corresponds to the reverse complemented counterpart of a transcript. With this option being used, every read alignment will have an XS attribute tag: '+ means a read belongs to a transcript on '+' strand of genome. '-' means a read belongs to a transcript on '-' strand of genome. (-rna-strandness)

[Summary Options](#)



[Advanced Options](#)



Job Resource Parameters

Use default job resource parameters

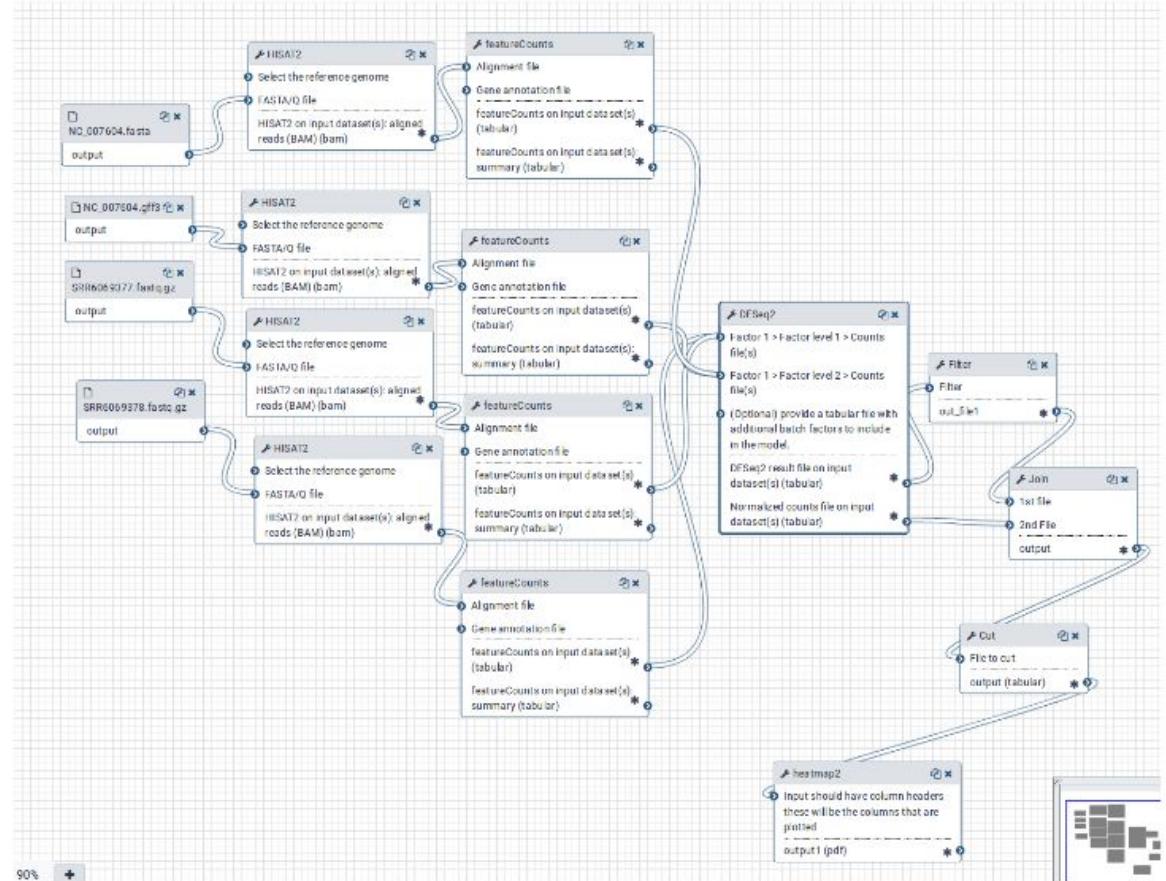
Execute

Introduction

What is HISAT?

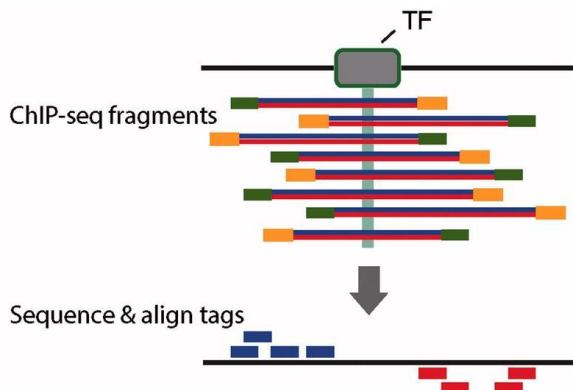
Workflows

- 1) Extracted from the history
- 2) Built manually
- 3) Imported

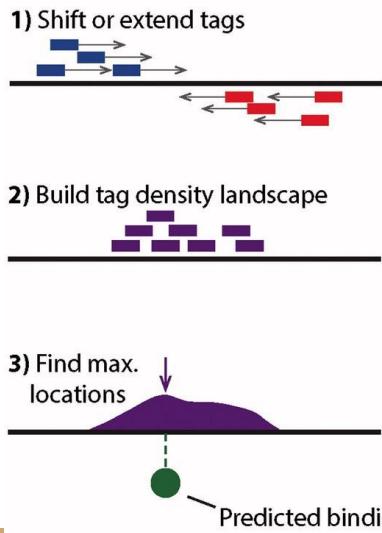


Why Workflows?

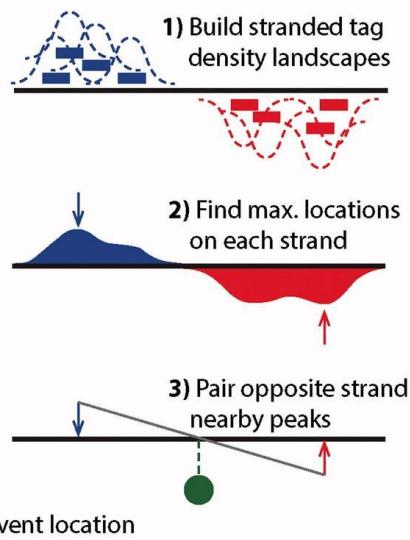
- Re-run the same analysis on different input data sets
- Change parameters before re-running a similar analysis
- Make use of the workflow job scheduling: jobs are submitted as soon as their inputs are ready
- Create sub-workflows: a workflow inside another workflow
- Share workflows for publication and with the community



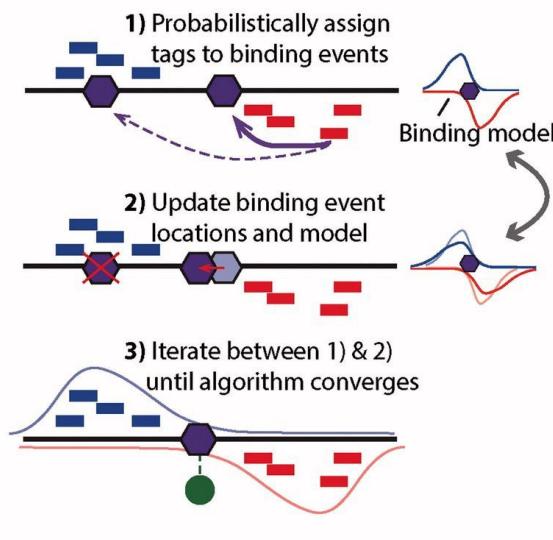
Peak-finding

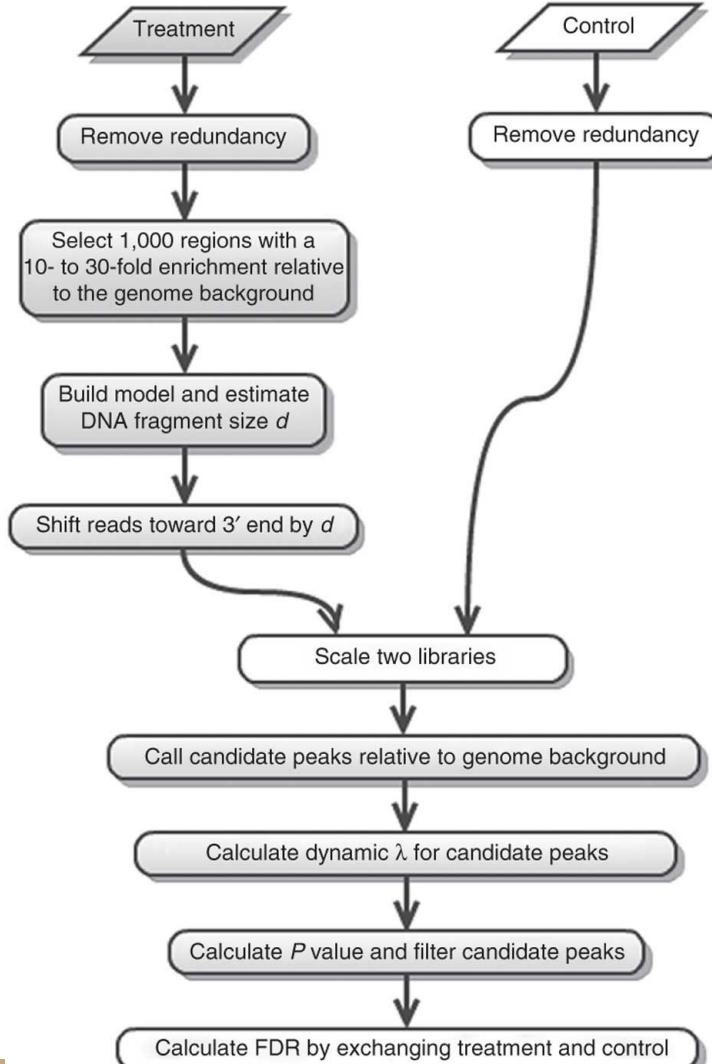


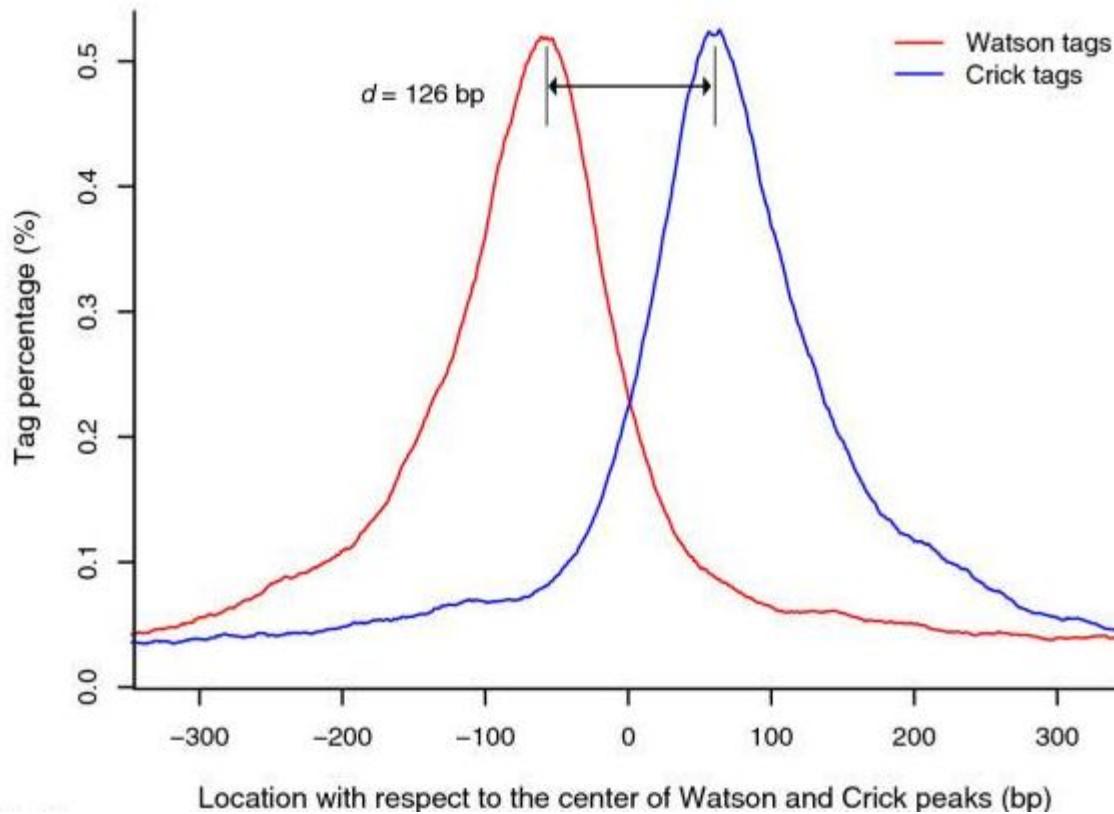
Peak-pairing

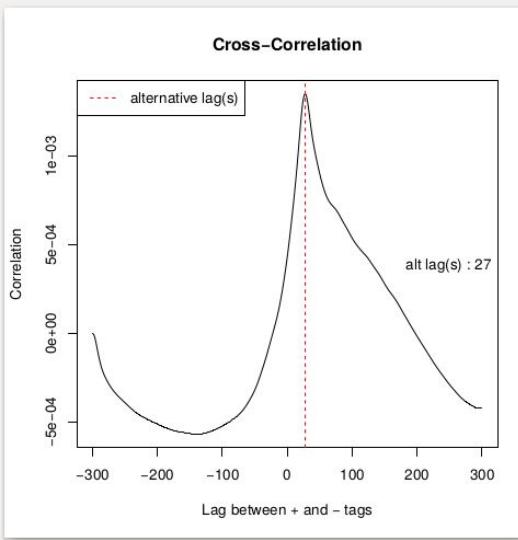
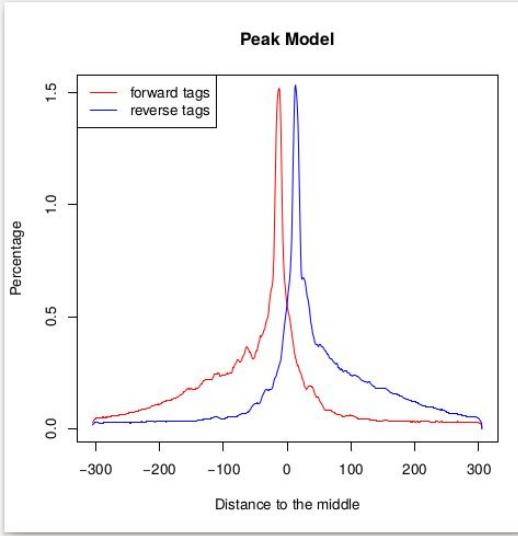


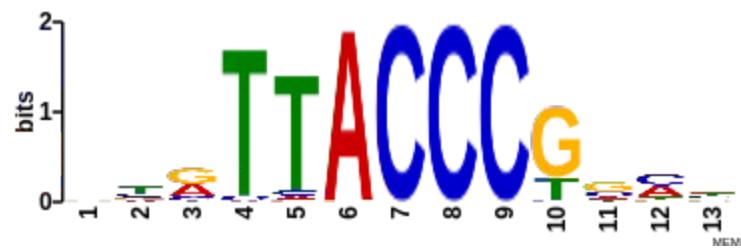
Probabilistic binding detection

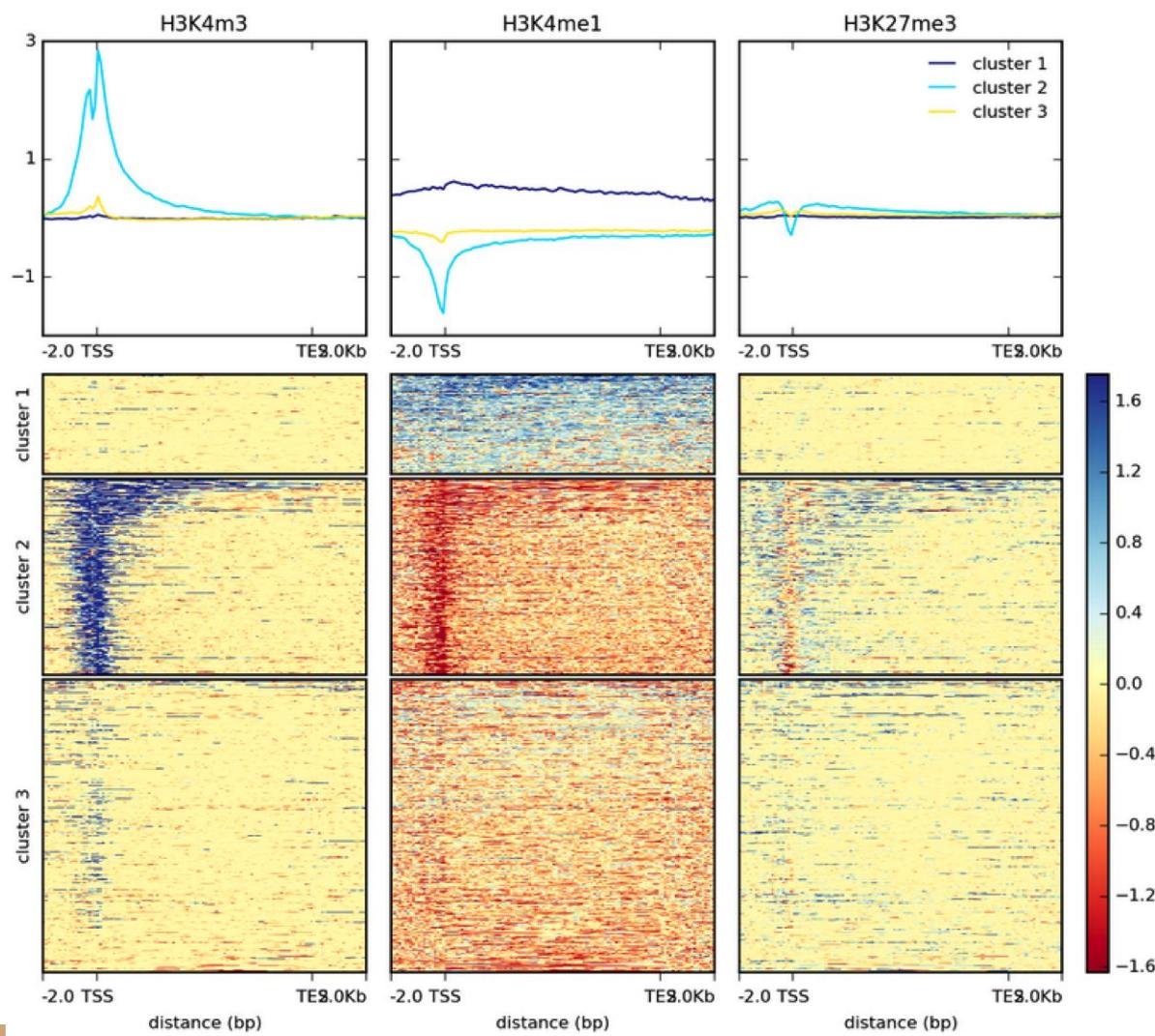










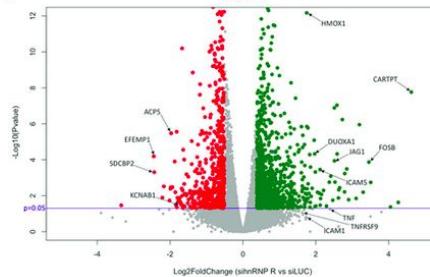




“Biological”
insights

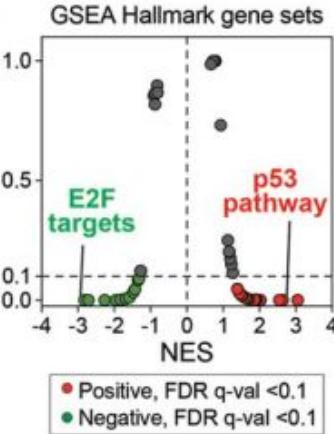
Small remainder

- Until now you learned:
 - Download/get raw sequencing reads
 - Evaluated raw data quality and clean up data
 - Map data (or know methods that skip this)
 - Quantify data
 - Perform differential expression analysis

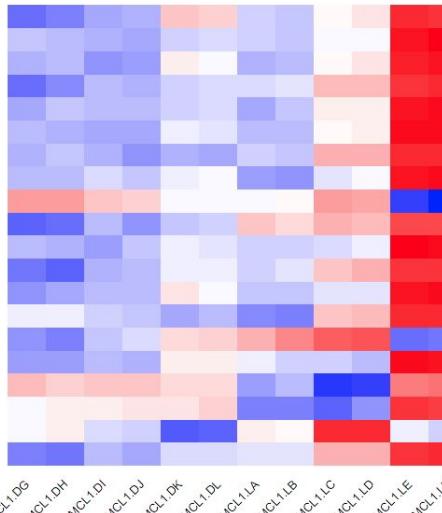
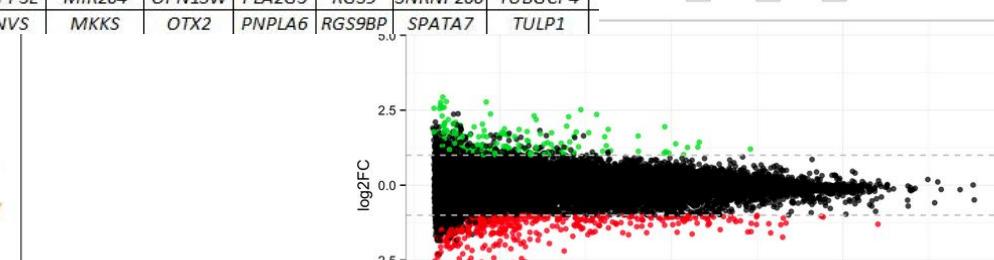
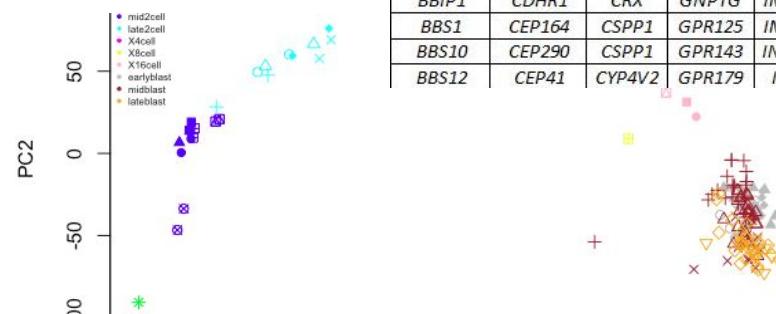


Your result “today”

B



ABCA4	BBS5	CERKL	DHX38	GRM6	JAG1	MVK	PCDH15	PRCD	RLBP1	TEAD1
ABCC6	BBS7	CHM	DRAM2	GRN	KCNJ13	MYO7A	PCYT1A	PRDM13	ROM1	TEK
ABHD12	BBS9	CIB2	DTHD1	GUCA1A	KCNV2	NDP	PDE6A	PROM1	RP1	TIMM8A
ACBD5	BEST1	CLN3	EFEMP1	GUCA1B	KCTD7	NEK2	PDE6B	PRPF3	RP1L1	TIMP3
ADAM9	C1QTNF5	CLN5	ELOVL4	GUCY2D	KIAA1549	NEUROD1	PDE6C	PRPF31	RP2	TMEM126A
ADAMTS18	C21orf2	CLN6	EMC1	HARS	KIF11	NMNAT1	PDE6G	PRPF4	RP9	TMEM231
AFG3L2	C2ORF71	CLN8	ERCC6	HMX1	KIZ	NPHP1	PDE6H	PRPF6	RPE65	TMEM237
AHI1	C5orf42	CLRN1	EYS	IDH3B	KLHL7	NPHP3	PDZD7	PRPF8	RPGR	TMEM67
AIPL1	C8orf37	CNGA1	FAM161A	IFT122	LCA5	NPHP4	PEX1	PRPH2	RPGRIP1	TOPORS
ALMS1	CA4	CNGA3	FLVCR1	IFT140	LRAT	NR2E3	PEX10	RAB28	RPGRIP1L	TPP1
ARL13B	CABP4	CNGB1	FOXF2	IFT172	LRIT3	NRL	PEX14	RAX2	RS1	TREX1
ARL2BP	CACNA1F	CNGB3	FSCN2	IFT27	LRP5	NYX	PEX16	RBP3	SAG	TRIM32
ARL6	CACNA2D4	CNNM4	FZD4	IFT43	LZTFL1	OAT	PEX19	RBP4	SDCCAG8	TRPM1
ASRG1L	CAPN5	COL11A1	GDF6	IFT80	MAK	OCA2	PEX2	RCBTB1	SEMA4A	TSPAN12
ATF6	CC2D2A	COL2A1	GNAT1	IFT88	MAPKAPK3	OFD1	PEX5	RD3	SLC24A1	TTC21B
ATXN7	CDH23	COL9A1	GNAT2	IKBKG	MERTK	OPA1	PEX6	RDH12	SLC25A46	TTC8
BBIP1	CDH3	CRB1	GNB1	IMPDH1	MFN2	OPA3	PEX7	RDH5	SLC45A2	TTLL5
BBIP1	CDHR1	CRX	GNPTG	IMPG1	MFRP	OPN1LW	PHYH	REEP6	SLC4A5	TPPA
BBS1	CEP164	CSPP1	GPR125	IMPG2	MFSD8	OPN1MW	PITPNM3	RGR	SLC7A14	TUB
BBS10	CEP290	CSPP1	GPR143	INPP5E	MIR204	OPN1SW	PLA2G5	RGS9	SNRNP200	TUBGCP4
BBS12	CEP41	CYP4V2	GPR179	INVS	MKKS	OTX2	PNPLA6	RGS9BP	SPATA7	TULP1



Today

ABCA4	BBS5	CERKL	DHX38	GRM6	JAG1	MVK	PCDH15	PRCD	RLBP1	TEAD1	L	
ABCC6	BBS7	CHM	DRAM2	GRN	KCNJ13	MYO7A	PCYT1A	PRDM13	ROM1	TEK		
ABHD12	BBS9	CIB2	DTHD1	GUCA1A	KCNV2	NDP	PDE6A	PROM1	RP1	TIMM8A		
ACBD5	BEST1	CLN3	EFEMP1	GUCA1B	KCTD7	NEK2	PDE6B	PRPF3	RP1L1	TIMP3		
ADAM9	C1QTNF5	CLN5	ELOVL4	GUCY2D	KIAA1549	NEUROD1	PDE6C	PRPF31	RP2	TMEM126A		
ADAMTS18	C21orf2	CLN6	EMC1	HARS	KIF11	NMNNAT1	PDE6G	PRPF4	RP9	TMEM231		
AFG3L2	C20orf71	CLN8	ERCC6	HMX1	KIZ	NPHP1	PDE6H	PRPF6	RPE65	TMEM237		
AHI1	C5orf42	CLRN1	EYS	IDH3B	KLHL7	NPHP3	PDZD7	PRPF8	RPGR	TMEM67		
AIPL1	C8orf37	CNGA1	FAM161A	IFT122	LCA5	NPHP4	PEX1	PRPH2	RPGRIPI1	TOPORS		
ALMS1	CA4	CNGA3	FLVCR1	IFT140	LRAT	NR2E3	PEX10	RAB28	RPGRIPI1L	TPP1		
ARL13B	CABP4	CNGB1	FOXF2	IFT172	LRIT3	NRL	PEX14	RAX2	RS1	TREX1		
ARL2BP	CACNA1F	CNGB3	FSCN2	IFT27	LRP5	NYX	PEX16	RBP3	SAG	TRIM32		
ARL6	CACNA2D4	CNNM4	FZD4	IFT43	LZTFL1	OAT	PEX19	RBP4	SDCCAG8	TRPM1		
ASRGL1	CAPN5	COL11A1	GDF6	IFT80	MAK	OCA2	PEX2	RCBTB1	SEMA4A	TSPAN12		
ATF6	CC2D2A	COL2A1	GNAT1	IFT88	MAPKAPK3	OFD1	PEX5	RD3	SLC24A1	TTC21B		
ATXN7	CDH23	COL9A1	GNAT2	IKBKG	MERTK	OPA1	PEX6	RDH12	SLC25A46	TTC8		
BBIP1	CDH3	CRB1	GNB1	IMPDH1	MFN2	OPA3	PEX7	RDH5	SLC45A2	TTLL5		
BBIP1	CDHR1	CRX	GNPTG	IMPG1	MFRP	OPN1LW	PHYH	REEP6	SLC4A5	TTPA		
BBS1	CEP164	CSPP1	GPR125	IMPG2	MFSD8	OPN1MW	PITPNM3	RGR	SLC7A14	TUB		
BBS10	CEP290	CSPP1	GPR143	INPP5E	MIR204	OPN1SW	PLA2G5	RGS9	SNRNP200	TUBGCP4		
BBS12	CEP41	CYP4V2	GPR179	INVS	MKKS	OTX2	PNPLA6	RGS9BP	SPATA7	TULP1		

Biological analysis

- What is common between those genes?
- What biochemical pathways are regulated by them?
- Is there any enrichment of pathways/function?

Biological analysis

- How to answer those questions? (from previous slide)



- Databases search?
- Articles?
- Big notebook?



GENEONTOLOGY
Unifying Biology

What is GO?

- GO – Gene ontologies
- Molecular functions
- Biological processes
- Cellular component

GO example

mitochondrion

Term Information

Data health 

Accession GO:0005739
Name mitochondrion
Ontology cellular_component
Synonyms mitochondria

Alternate IDs None

Definition A semiautonomous, self replicating organelle that occurs in varying numbers, shapes, and sizes in the cytoplasm of virtually all eukaryotic cells. It is notably the site of tissue respiration. Source: GOC:giardia, ISBN:0198506732

Comment Some anaerobic or microaerophilic organisms (e.g. *Entamoeba histolytica*, *Giardia intestinalis* and several *Microsporidia* species) do not have mitochondria, and contain mitochondrion-related organelles (MROs) instead, called mitosomes or hydrogenosomes, very likely derived from mitochondria. To annotate gene products located in these mitochondrial relics in species such as *Entamoeba histolytica*, *Giardia intestinalis* or others, please use GO:0032047 'mitosome' or GO:0042566 'hydrogenosome'. (See PMID:24316280 for a list of species currently known to contain mitochondrion-related organelles.)

History See term [history for GO:0005739](#) at QuickGO

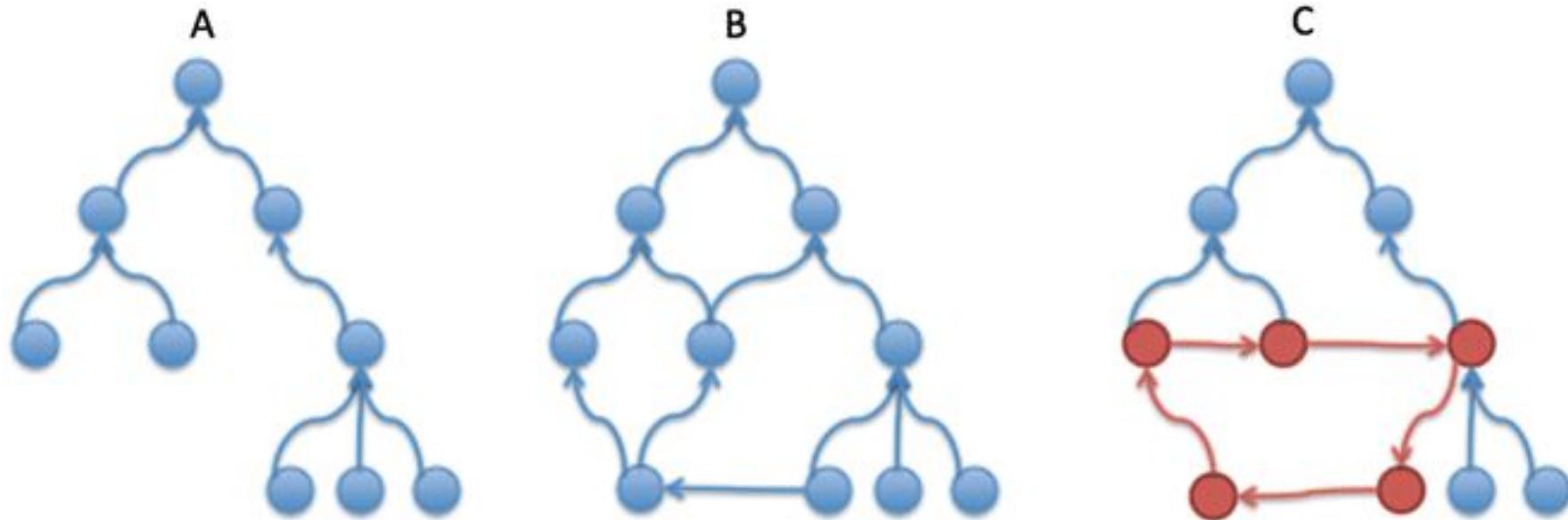
Subset goslim_aspergillus
goslim_flybase_ribbon
goslim_drosophila
goslim_chembl
goslim_plant
goslim_generic
goslim_mouse
goslim_agr
goslim_candida
goslim_yeast
goslim_pir

Related [Link](#) to all **genes and gene products** annotated to mitochondrion.

[Link](#) to all direct and indirect **annotations** to mitochondrion.

[Link](#) to all direct and indirect **annotations download** (limited to first 10,000) for mitochondrion.

GO structure



GO classes

Table 1

A selection of relationship types commonly used in bio-ontologies

Relationship type	Informal meaning	Examples
part_of	The standard relation of parthood.	A brain is part_of a body.
derives_from	Derivation holds between distinct entities when one succeeds the other across a temporal divide in such a way that a biologically significant portion of the matter of the earlier entity is inherited by the latter.	A zygote derives_from a sperm and an ovum.
has_participant	A relation that links processes to the entities that participate in them.	An apoptotic process has_participant a cell.
has_function	A relation that links material entities to their functions, e.g. the biological functions of macromolecules.	An enzyme has_function to catalyse a specific reaction type.

Applications

- Support the structured annotation of data in a database
- Ontologies can serve as a rich source of vocabulary for a domain of interest
- It is possible to annotate data to the most specific applicable term but then to examine large-scale data in aggregate for patterns at the higher level categories

Limitations

- They are good at representing statements that are either true or false (categorical), but they cannot elegantly represent knowledge that is vague, statistical or conditional
- There are pragmatic limits to ensure the scalability of the reasoning tools

Gene ID	Gene Ensembl ID	GO ID
2	ENSDARG00000059215	GO:0005922 GO:0016021 ...
5	ENSDARG00000043963	GO:0005515
12	ENSDARG00000059187	GO:0005922 GO:0016021 ...
14	ENSDARG00000059183	GO:0005874 GO:0043234 ...
16	ENSDARG00000059168	GO:0016021 GO:0016020 ...
17	ENSDARG00000059164	GO:0016021 GO:0016020 ...
19	ENSDARG00000043982	GO:0005874 GO:0043234 ...
20	ENSDARG00000020785	GO:0005102 GO:0030903 ...
21	ENSDARG00000043973	GO:0005634 GO:0046872 ...
26	ENSDARG00000044013	GO:0005634 GO:0003677 ...
29	ENSDARG00000044016	GO:0016459 GO:0005524 ...
31	ENSDARG00000059041	GO:0000166 GO:0046872 ...
32	ENSDARG00000018477	GO:0000166 GO:0046872 ...
33	ENSDARG00000006332	GO:0007242 GO:0035091 ...
34	ENSDARG00000009020	GO:0004930 GO:0001584 ...
35	ENSDARG00000032261	GO:0004484 GO:0003676 ...
37	ENSDARG00000043902	GO:0016021 GO:0045211 ...
38	ENSDARG00000014057	GO:0016021 GO:0045211 ...
39	ENSDARG00000033489	GO:0006512 GO:0006464 ...
42	ENSDARG00000003751	GO:0016740 GO:0004674 ...
43	ENSDARG00000019747	GO:0006694 GO:0004769 ...
44	ENSDARG00000034076	GO:0004872
45	ENSDARG00000015201	GO:0006464 GO:0016740 ...
48	ENSDARG00000021509	GO:0008270 GO:0046872 ...

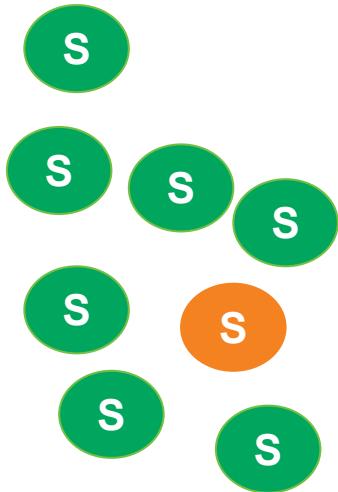
What is next?

Enrichment analysis



[Source](#)

Skittles

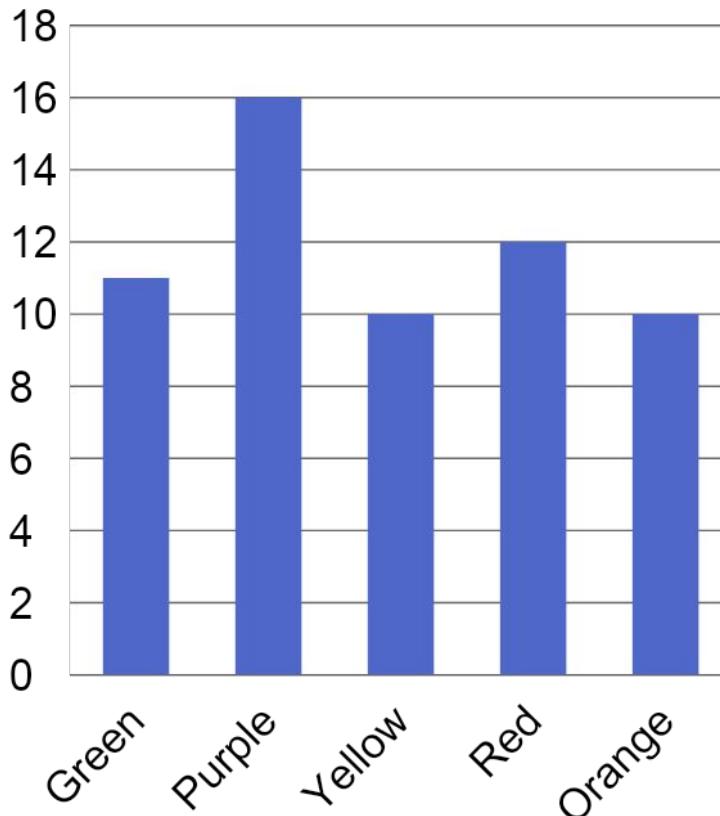


We took some skittles from the bag (blindly) and got such set.
Do you have any ideas about the content of the bag? (colors distribution)

Standard skittles bag

Total # of each Color

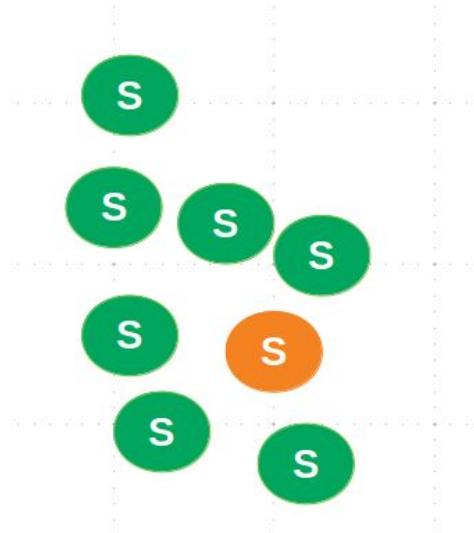
Color	Count	%
GREEN	11	19
PURPLE	16	27
YELLOW	10	17
RED	12	20
ORANGE	10	17



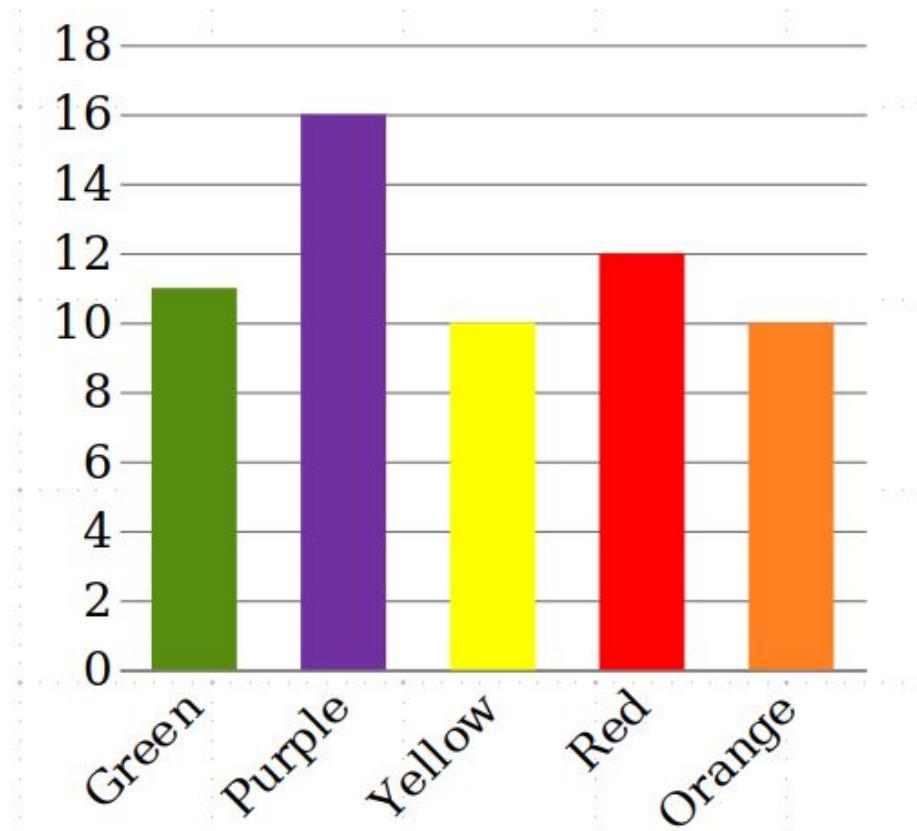
From Kristin Taylor slides (found on the internet)

Question

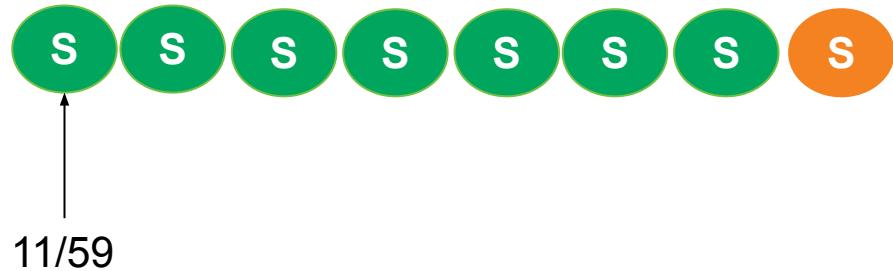
- Is our skittles bag



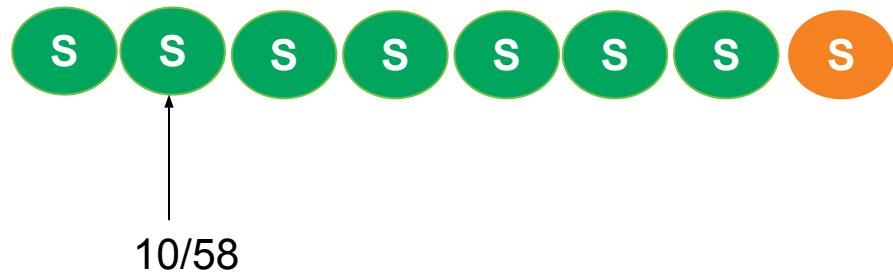
Is skittles order important here?



Calculate probability



Calculate probability

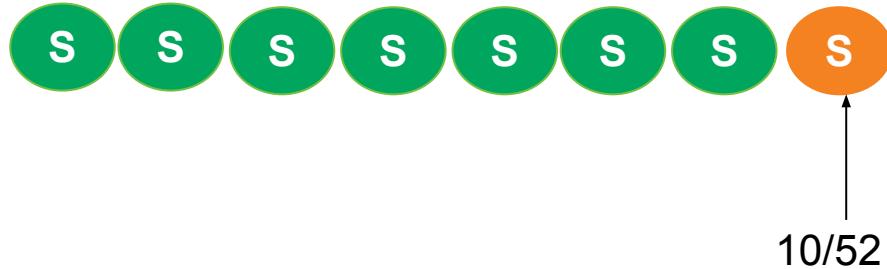


Calculate probability



Continue..

Calculate probability



What is the probability of



- $11/59 * 10/58 * 9/57 * 8/56 * 7/55 * 6/54 * 5/53 * 10/52 = 1.860227e-07$

What is the probability of



- $11/59 * 10/58 * 9/57 * 8/56 * 7/55 * 6/54 * 5/53 * 10/52 = 1.860227e-07$

Any problems?

Any problems?



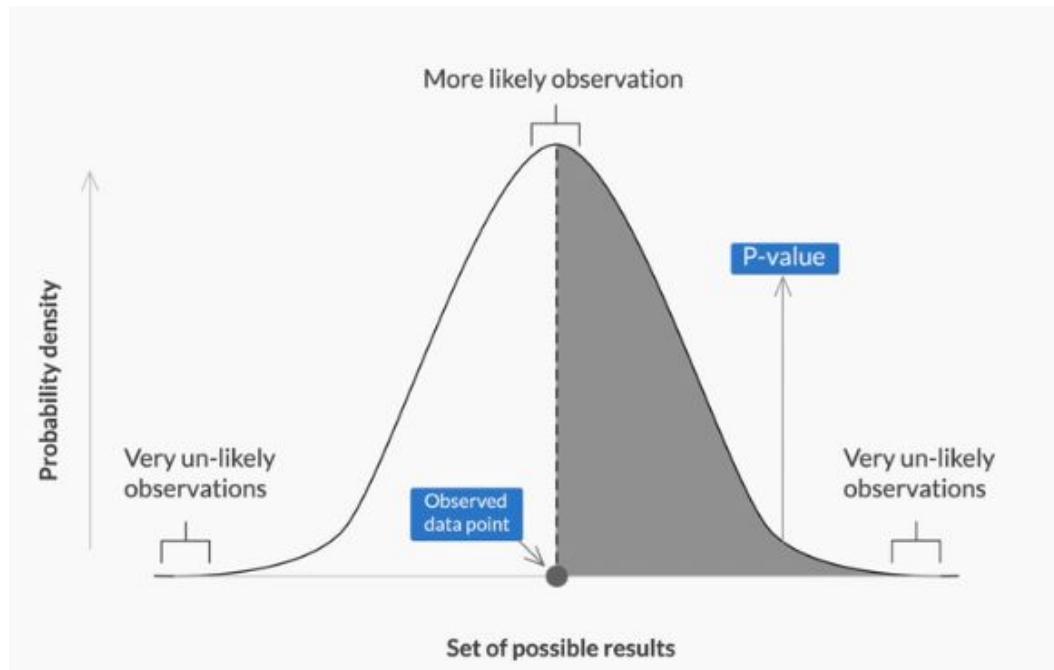
What is a probability?

0.000001488182

What is the P-value here?

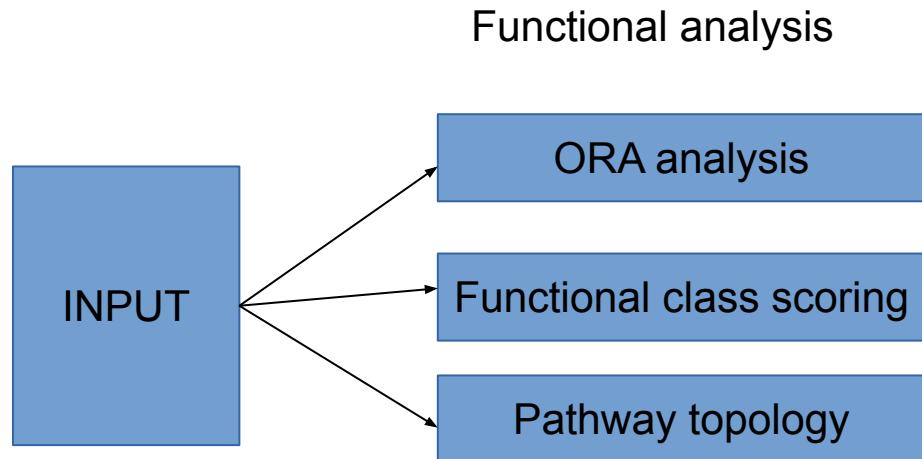
0.0012

P-value 0.05



Source

Other analysis



ORA

- Pros: simple, powerful, less input needed, easy to understand
- Cons: Background assumptions, data loss, data dependency (which is not true), takes specific number of genes, many false positive

FCS

- Pros: better in accuracy. Uses entire list of genes
- Cons: ignores interactions, analysis are individual, many false positives

Pathway topology

- Pros: Considers each gene's role, position, magnitude and interactions
- Cons: takes more time, not applicable to all data

Main tools for the analysis

WEB based:

- GOrilla - <http://cbl-gorilla.cs.technion.ac.il/>
- ShinyGO - <http://bioinformatics.sdbstate.edu/go/>
- EnrichR - <https://maayanlab.cloud/Enrichr/>

Local (R) tools:

- clusterProfiler
- enrichR

Other interesting tools:

- KEGG - <https://www.genome.jp/kegg/pathway.html>
- REACTOME - <https://reactome.org/>

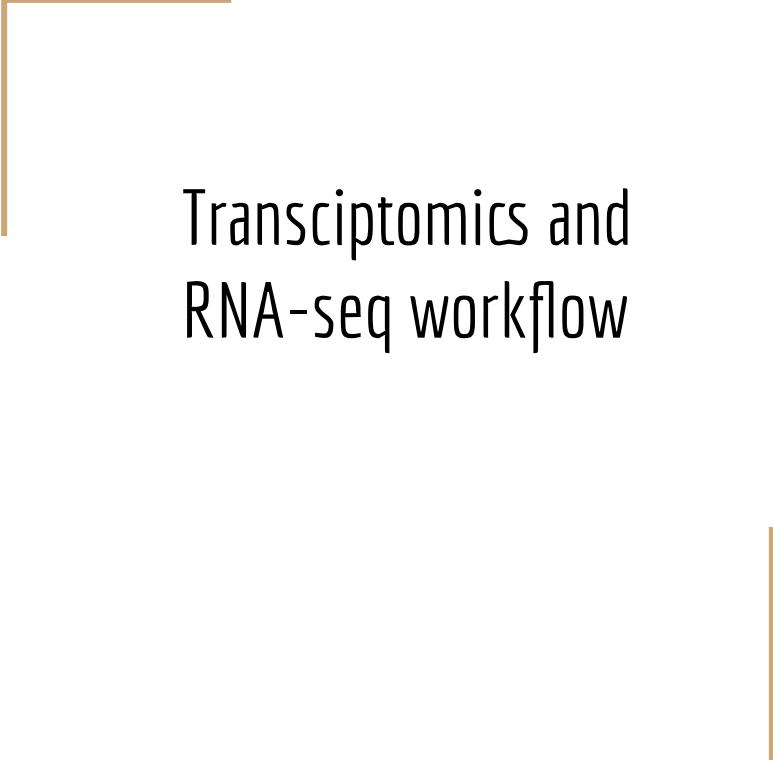
Figures and information source

- Volcano plot
- Heatmap
- PCA
- MA
- Profile
- Genes
- GO IDs
- GO book

RNA-seq workflow

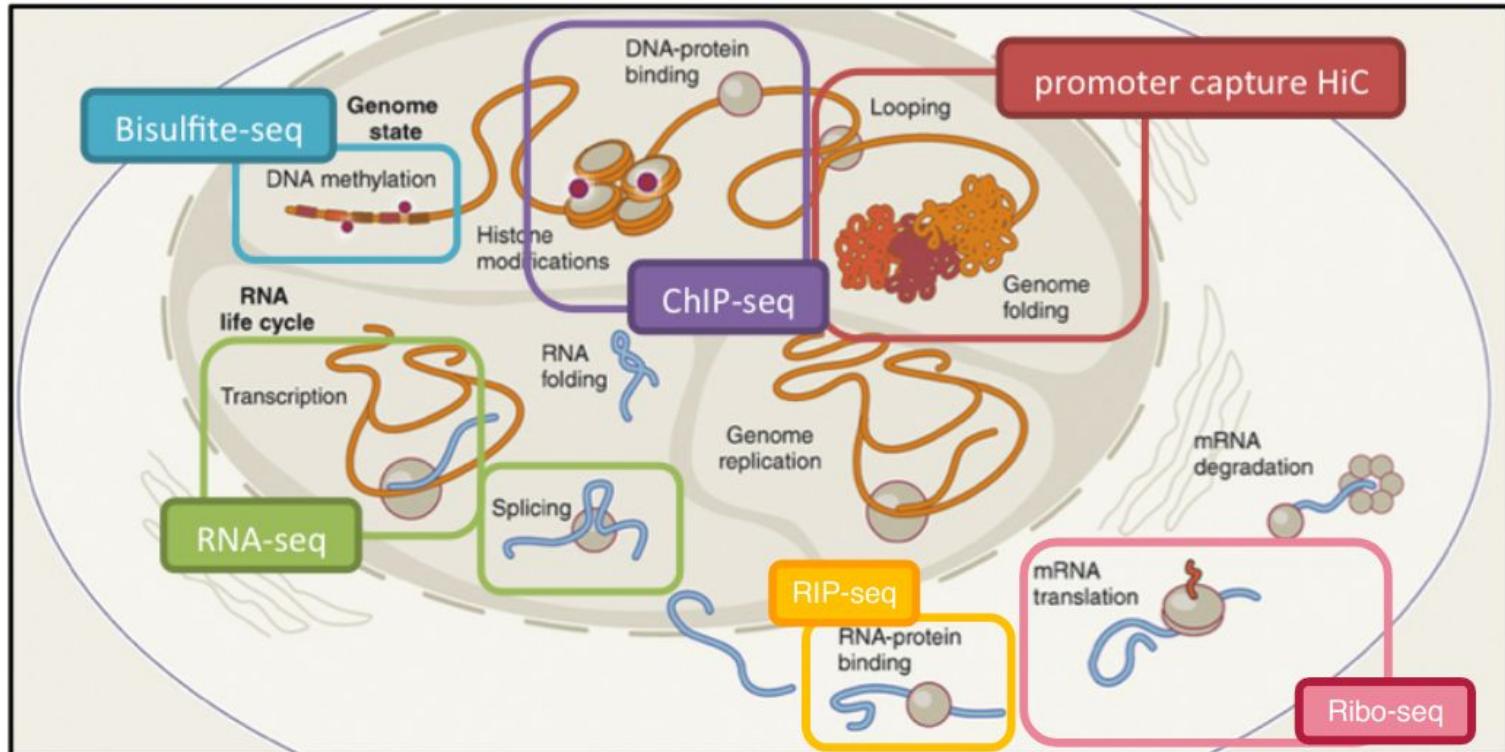
Topics

- RNA-seq workflow

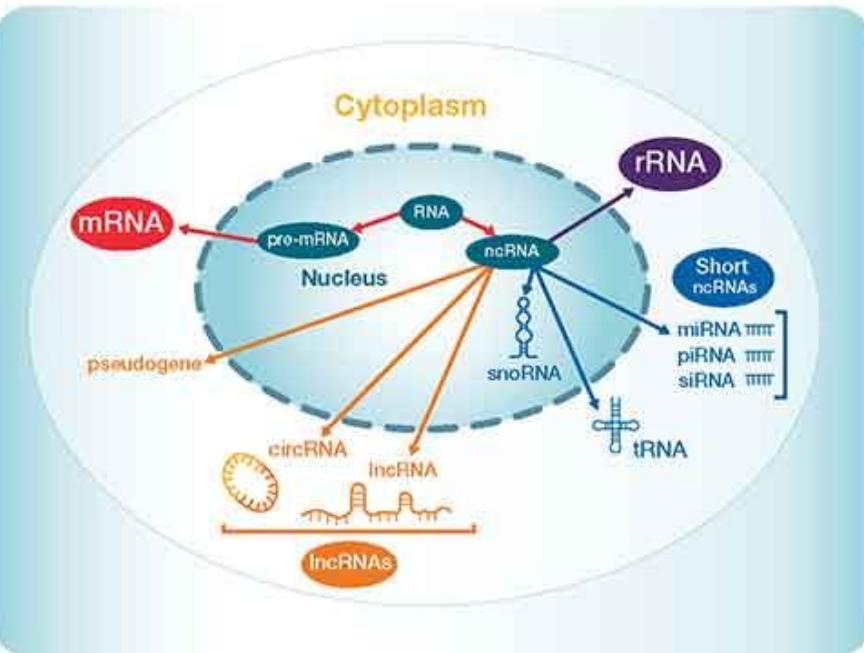


Transciptomics and RNA-seq workflow

Other “omics” (from NGS)



The CELL



RNA type	Abbreviation	Function
Messenger RNA	mRNA	Codes for protein
Ribosomal RNA	rRNA	Translation
Transfer RNA	tRNA	Translation
Small nuclear RNA	snRNA	Splicing and other functions
Small nucleolar RNA	snoRNA	Nucleotide modification of RNAs
Small Cajal body-specific RNA	scaRNA	Type of snoRNA; nucleotide modification of RNAs
Long noncoding RNA	lncRNA	Regulation of gene transcription; epigenetic regulation
MicroRNA	miRNA	Gene regulation
Piwi-interacting RNA	piRNA	Transposon defense; maybe other functions
Small interfering RNA	siRNA	Gene regulation

Source

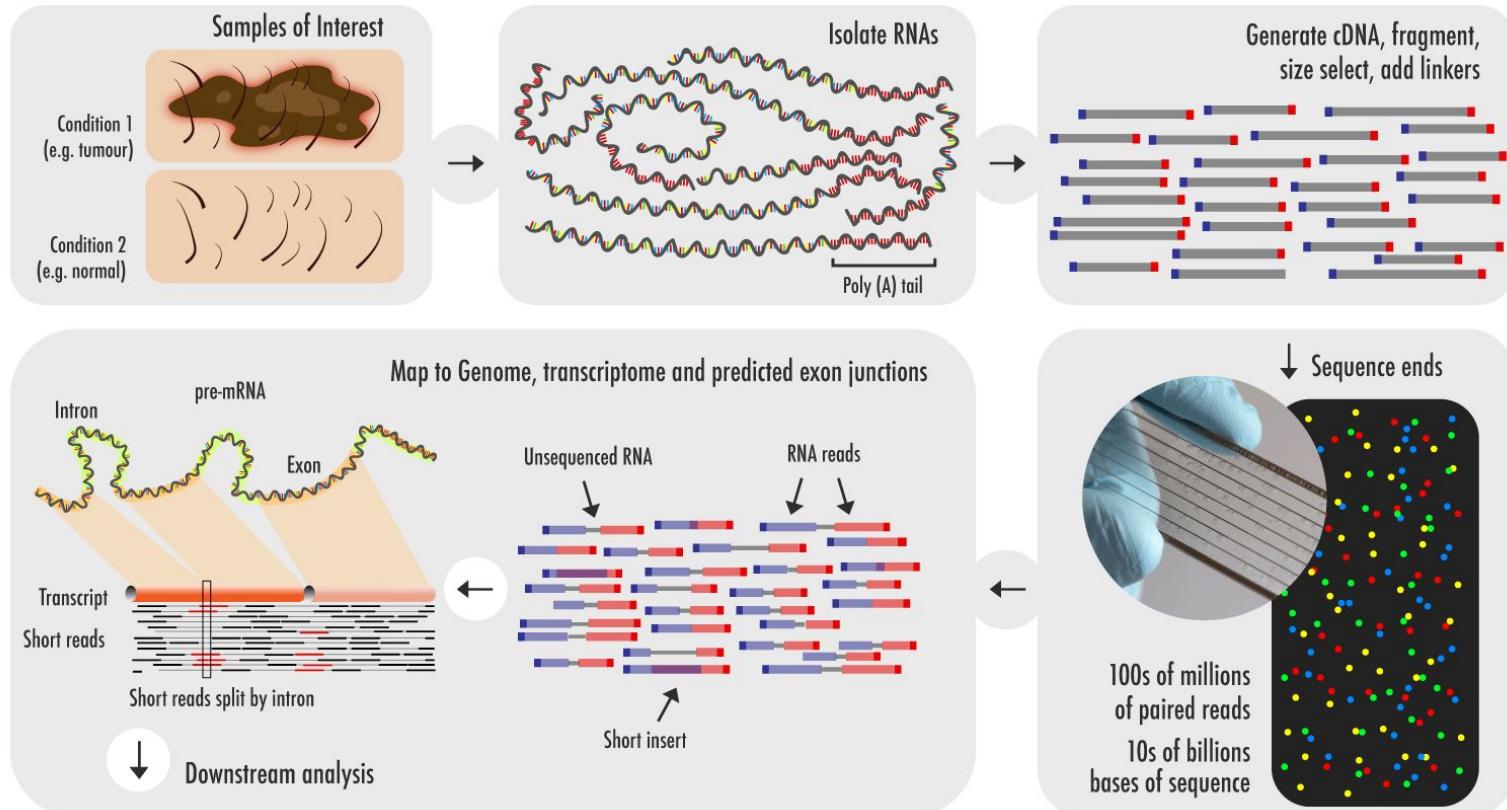
RNA-seq “Questions”

- Gene expression and differential expression analysis
- Alternative expression analysis
- Transcripts discovery
- Condition/Allele specific expression
- Etc.

RNA sequencing experimental considerations and workflow

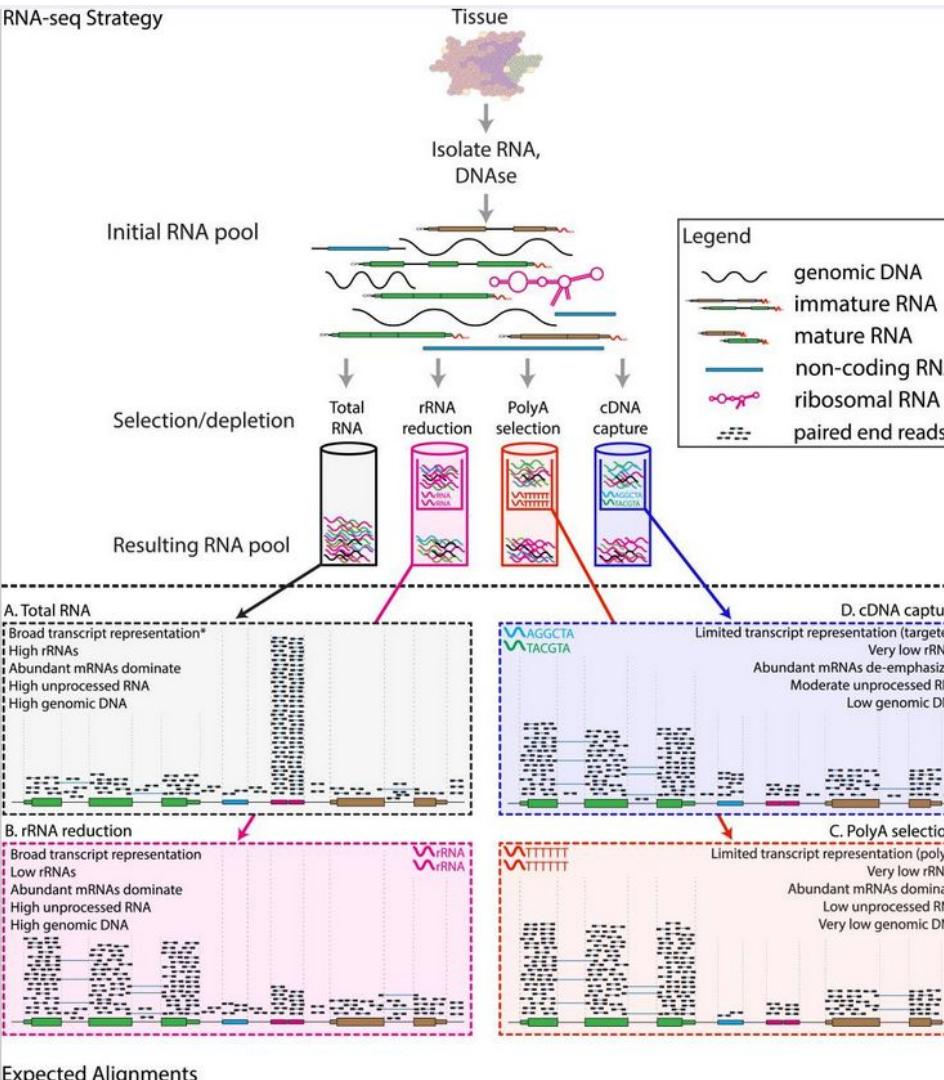
- Area of research
- Sample source
- Time dependence
- Coverage

RNA-seq workflow



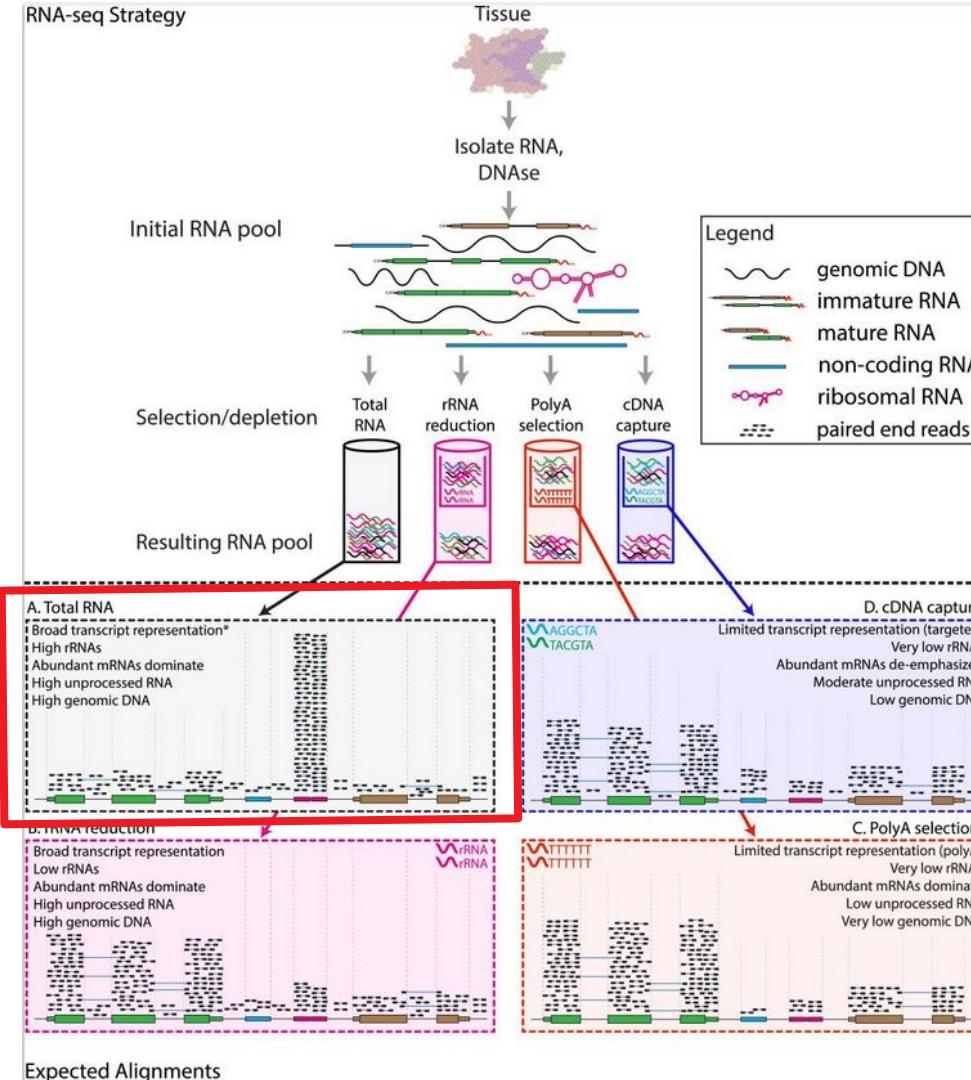
Wet-lab

- RNA Isolation
- RNA selection/depletion
- cDNA synthesis
- (sequencing)



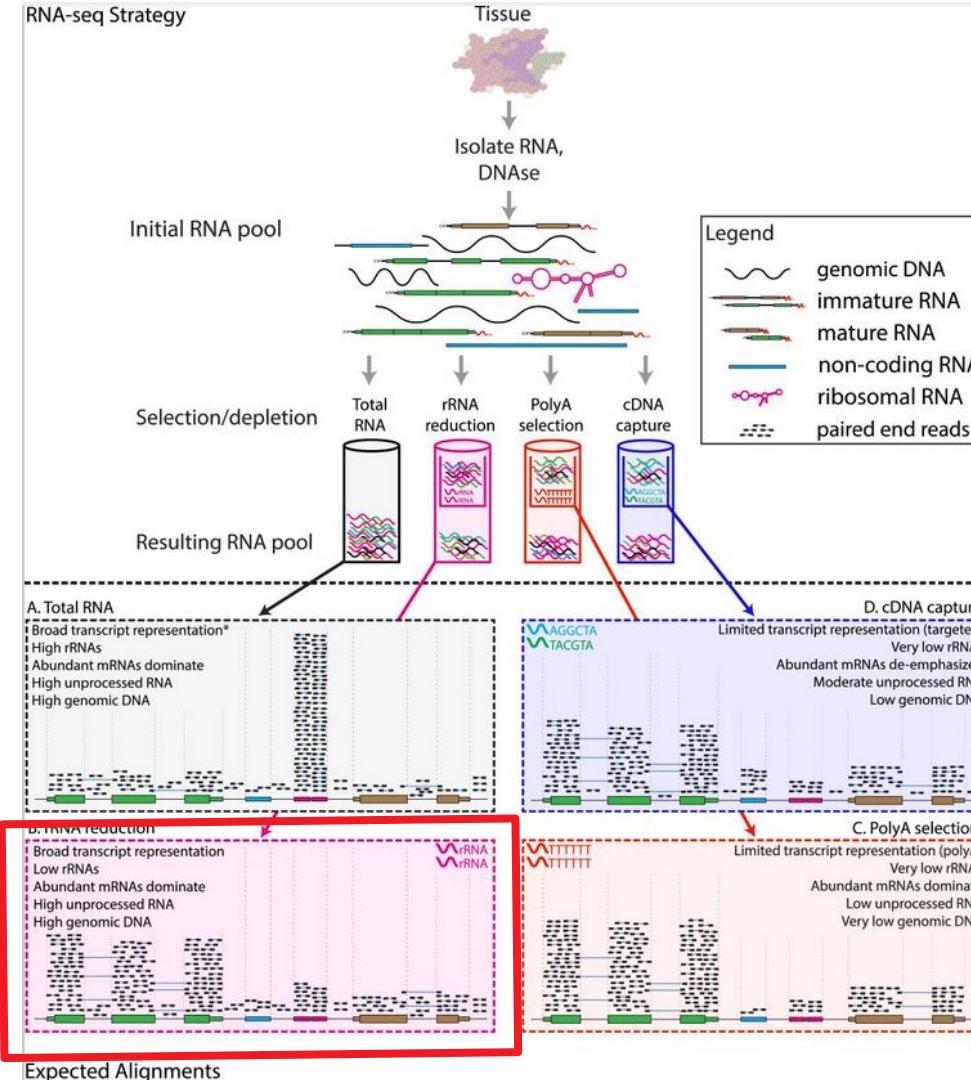
RNA-seq library fragmentation and size selection strategies

RNA-seq Strategy



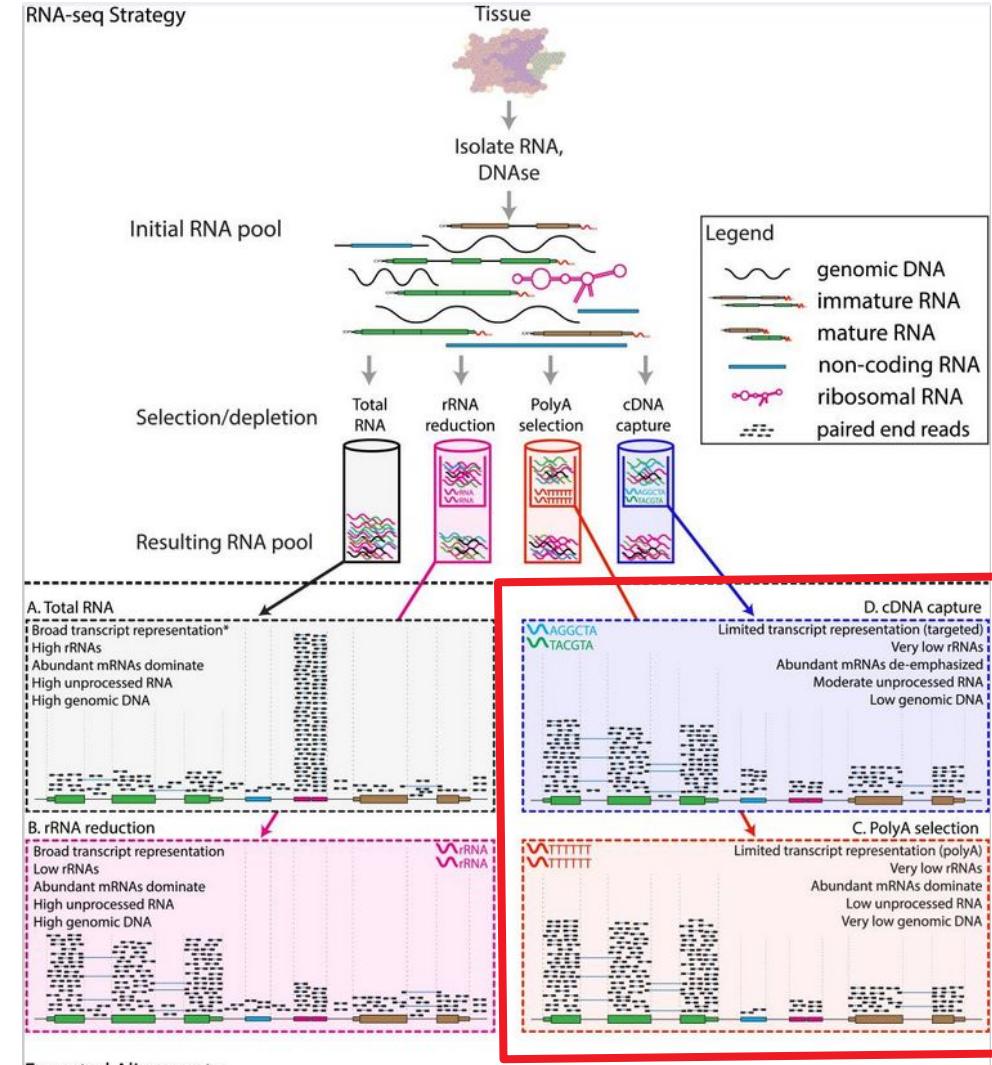
RNA-seq library fragmentation and size selection strategies

RNA-seq Strategy



RNA-seq library fragmentation and size selection strategies

RNA-seq Strategy



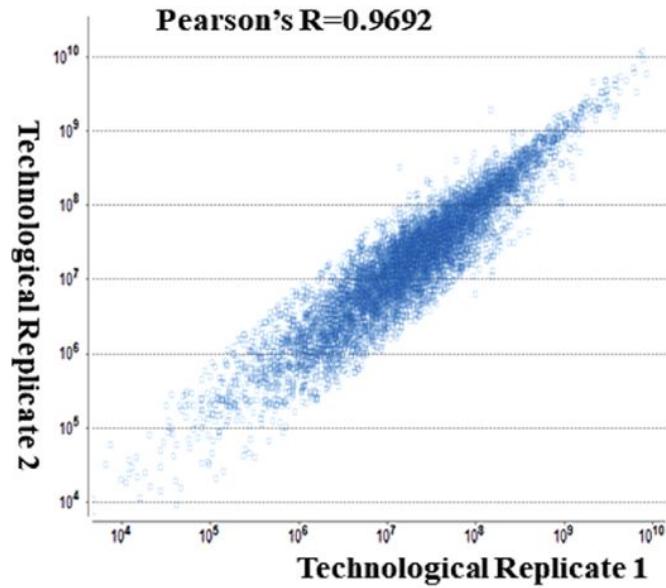
RNA-seq library fragmentation and size selection strategies

RNA-seq library fragmentation and size selection strategies

Strategy	Type of RNA	Ribosomal RNA content	Unprocessed RNA content	Genomic DNA content	Isolation method
Total RNA	All	High	High	High	None
PolyA selection	Coding	Low	Low	Low	Hybridization with poly(dT) oligomers
rRNA depletion	Coding, noncoding	Low	High	High	Removal of oligomers complementary to rRNA
RNA capture	Targeted	Low	Moderate	Low	Hybridization with probes complementary to desired transcripts

Replicates

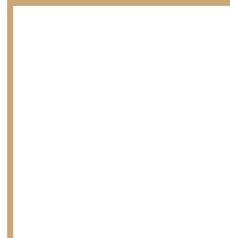
- Technical
 - Sequencing same sample many times (individual lanes, individual flowcells..)
- Biological
 - Multiple isolates from same phenotype/condition/treatment



DOI: 10.1002/pmic.201700408

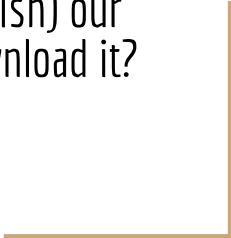
Small reading

RNA-seq



Data sources

Where should we put (publish) our
data and from where to download it?



Biological databases

- NCBI
- ENSEMBL
- UNIPROT
- Etc.

NCBI (www.ncbi.nlm.nih.gov)

NCBI Resources How To

All Databases Search

COVID-19 is an emerging, rapidly evolving situation.

Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS)

NCBI Home

Resource List (A-Z)

- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

About the NCBI | Mission | Organization | NCBI News & Blog

Submit
Deposit data or manuscripts into NCBI databases

Download
Transfer NCBI data to your computer

Learn
Find help documents, attend a class or watch a tutorial

Develop
Use NCBI APIs and code libraries to build applications

Analyze
Identify an NCBI tool for your data analysis task

Research
Explore NCBI research and collaborative projects

Popular Resources

- PubMed
- Bookshelf
- PubMed Central
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

NCBI News & Blog

Introducing MicroBiGG-E, a browser for microbial AMR genes and other stress and resistance elements
25 Jan 2021
The Pathogen Detection project now

Allele Frequency Aggregator (ALFA)
Release 2 is available!
22 Jan 2021
We are excited to announce the NCBI Allele Frequency Aggregator (ALFA)

NCBI on YouTube: RAPT and BLAST+ on the Cloud, SARS-CoV-2 genome data in Datasets
15 Jan 2021
It's time we do another roundin' up what's

NCBI SRA/BioProject/BioSample

SRA SRA mouse Search Help

Create alert Advanced

COVID-19 is an emerging, rapidly evolving situation.

[Public health information \(CDC\)](#) | [Research information \(NIH\)](#) | [SARS-CoV-2 data \(NCBI\)](#) | [Prevention and treatment information \(HHS\)](#)

Access Summary ▾ 20 per page ▾ Send to: ▾ **Filters:** [Manage Filters](#)

Source **Search results** Items: 1 to 20 of 1436695 << First < Prev Page 1 of 71835 Next > Last >>

Type [RNA-Seq of *mus musculus*: old male muscle: resistant training](#)

exome (17,041)
genome (44,788)

Library Layout [RNA-Seq of *mus musculus*: old male muscle](#)

paired (852,323)
single (584,372)

Platform [RNA-Seq of *mus musculus*: Young male muscle](#)

ABI SOLID (3,258)
BGISEQ (3,900)
Capillary (44,771)
Complete Genomics (9)
Helicos (1,377)
Illumina (1,354,056)
Ion Torrent (7,607)
LS454 (19,607)
Oxford Nanopore (740)
PacBio SMRT (1,370)

Strategy [Targeted Deep Sequencing of NIH3T3 Cell](#)

EpiGenomics (84,648)
Exome (210,132)
Genome (91,607)
RNaseq (3,840)
other (1,046,468)

Data in Cloud [Targeted Deep Sequencing of NIH3T3 Cell](#)

CS 14/12/2021

Top Organisms [Tree]

Mus musculus (1218994)
mouse gut metagenome (80179)
Homo sapiens (55258)
gut metagenome (20418)
mouse metagenome (14444)
All other taxa (47402)
More...

Top Bioprojects

Epigenomics projects for the... (392)
Production ENCODE functional... (288)
Mouse ENCODE epigenomic data (248)
Mouse ENCODE functional geno... (175)
Mouse ENCODE transcriptome d... (98)
Production ENCODE epigenomic... (18)

Search in related databases

Database	Access		all
	public	controlled	
BioSample	1,222,638	13	1,222,651
BioProject	28,678	4	28,682
dbGaP			15
GEO Datasets	899,248		899,248

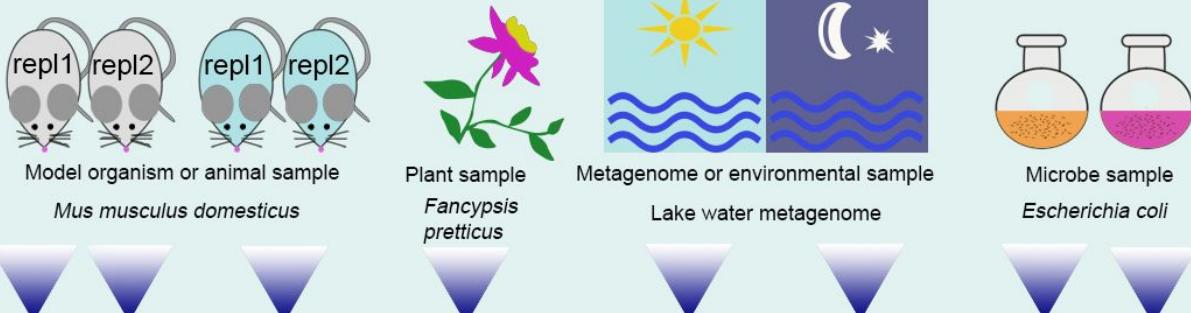
Find related data

BioProject and BioSample

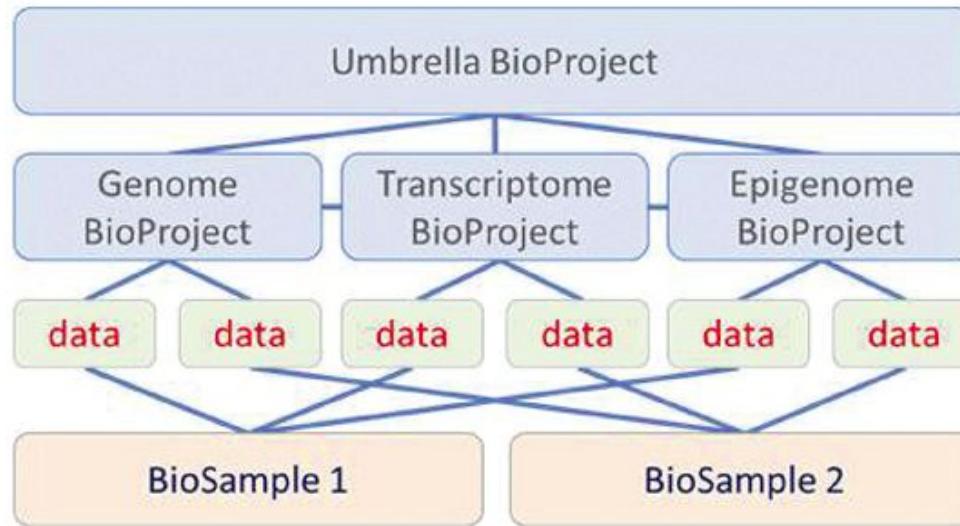
BioProject:
description
of research
project

BioSample:
description
of biological
samples

Project title:	Transcriptome analysis of hepatotoxicity induced by botulin in mice	Transcriptome of flowering plant	Metagenome of chlorophyll-containing microbiome in Norwegian lake	Mapping and manipulating E. coli transcriptome using antibiotics
Sample type:	Model organism or animal sample	Plant sample	Metagenome or environmental sample	Microbe sample
Organism:	<i>Mus musculus domesticus</i>	<i>Fancypsis preticus</i>	Lake water metagenome	<i>Escherichia coli</i>
Sample name:	Cntr1 Cntr2	Botulin	Pooled	Light Dark



NCBI SRA/BioProject/BioSample



NCBI SRA/BioProject/BioSample

- And many different IDs.
- As well as
- BioProject (PRJN),
- BioSample(SAMN),
- etc.

Accession Prefix	Accession Name	Definition	Example
SRA	SRA submission accession	The submission accession represents a virtual container that holds the <u>objects</u> represented by the other five accessions and is used to track the submission in the archive.	Since the SRA accession number is an artificial packaging construct, there is no example available since the SRA accession number has no specific response page
SRP	SRA study accession	A Study is an <u>object</u> that contains the project metadata describing a sequencing study or project. Imported from BioProject.	HTML
SRX	SRA experiment accession	An Experiment is an <u>object</u> that contains the metadata describing the library, platform selection, and processing parameters involved in a particular sequencing experiment.	HTML
SRR	SRA run accession	A Run is an <u>object</u> that contains actual sequencing data for a particular sequencing experiment. Experiments may contain many Runs depending on the number of sequencing instrument runs that were needed.	HTML
SRS	SRA sample accession	A Sample is an <u>object</u> that contains the metadata describing the physical sample upon which a sequencing experiment was performed. Imported from BioSample.	HTML
SRZ	SRA analysis accession	An analysis is an <u>object</u> that contains a sequence data analysis BAM file and the metadata describing the sequence analysis.	

ENSEMBL (<https://www.ensembl.org/index.htm>)

The screenshot shows the Ensembl homepage with a dark blue header containing the Ensembl logo and navigation links: BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs | Blog.

The main content area has four sections:

- Tools**: A link to All tools.
- BioMart >**: Description: Export custom datasets from Ensembl with this data-mining tool.
- BLAST/BLAT >**: Description: Search our genomes for your DNA or protein sequence.
- Variant Effect Predictor >**: Description: Analyse your own variants and predict the functional consequences of known and unknown variants.

Below these sections is a search bar labeled "Search" with a dropdown menu set to "All species" and a text input field containing "for". A "Go" button is to the right of the input field. Below the search bar is an example query: "e.g. BRCA2 or rat 5:62797383-63627669 or rs699 or coronary heart disease".

The bottom section contains two columns:

- All genomes**: A dropdown menu labeled "-- Select a species --". Below it is a "Pig breeds" section with a small pig icon and text: "Pig reference genome and 12 additional breeds". A link "View full list of all species" is at the bottom of this column.
- Favourite genomes**: A list with three entries:
 - Human**: An icon of a classical statue, text: "GRCh38.p13", and a link "Still using GRCh37?".
 - Mouse**: An icon of a mouse, text: "GRCm38.p6".
 - Zebrafish**: An icon of a zebrafish, text: "GRCz11".

DATA SOURCES (for NGS)

- Basic databases:
 - NCBI (<https://www.ncbi.nlm.nih.gov/>)
 - ENA (<https://www.ebi.ac.uk/ena/browser/home>)
 - NCBI GEO (<https://www.ncbi.nlm.nih.gov/geo/>)

DATA SOURCES (for NGS)

- Basic databases:
 - NCBI (<https://www.ncbi.nlm.nih.gov/>)
 - ENA (<https://www.ebi.ac.uk/ena/browser/home>)
 - NCBI GEO (<https://www.ncbi.nlm.nih.gov/geo/>)
- Basic tools:
 - wget/curl
 - ftp/GLOBUS client
 - SRA-tools
 - Edirect (for FASTA files)

SRA-toolkit

- Available from <https://github.com/ncbi/sra-tools/wiki>
- Command line tool
- Allows you to download/predownload FASTQ files from NCBI databases
- **Most important command for us:**
 - `fastq-dump --gzip --split-files SRR_ID`

EDirect

- Entrez Direct: E-utilities
- Available from: <https://www.ncbi.nlm.nih.gov/books/NBK179288/>
- Command line tool
- Contains several tools

EDirect

- Tools:
 - ESearch: Search a text query in a single Entrez database.
 - ESummary: Retrieve document summaries for each UID.
 - EFetch: Retrieve full records for each UID.
 - EPost: Upload a list of UIDs for later use.
 - ELink: Retrieve UIDs for related or linked records, or LinkOut URLs.
 - EInfo: Retrieve information and statistics about a single database.
 - ESpell: Retrieve spelling suggestions for a text query.
 - ECitMatch: Search PubMed for a series of citation strings.
 - EGQuery: Search a text query in all Entrez databases and return the number of results for the query in each database.

EDirect

- We will use (to get reference genomes):
 - efetch
 - epost

To sum up

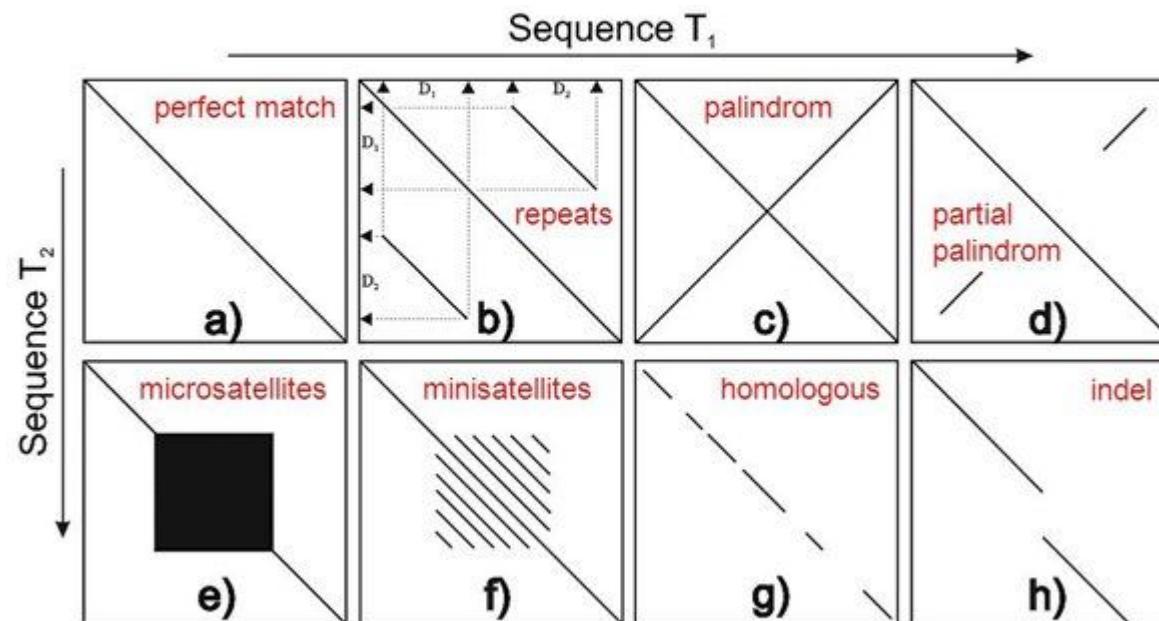
- Before publishing, data should be deposited to a public repository:
 - GEO (processed data) and NCBI SRA (raw data)
 - (or) ENA
 - Other specific databases
- Published data may be accessed:
 - NCBI GEO or NCBI SRA
 - ENA
 - Other specific databases
- Raw data (from sequencing facilities) may be downloaded using:
 - Web interfaces (if system allows this)
 - FTP clients (like FileZilla)
 - Special programs

Dviejų sekų lyginimas

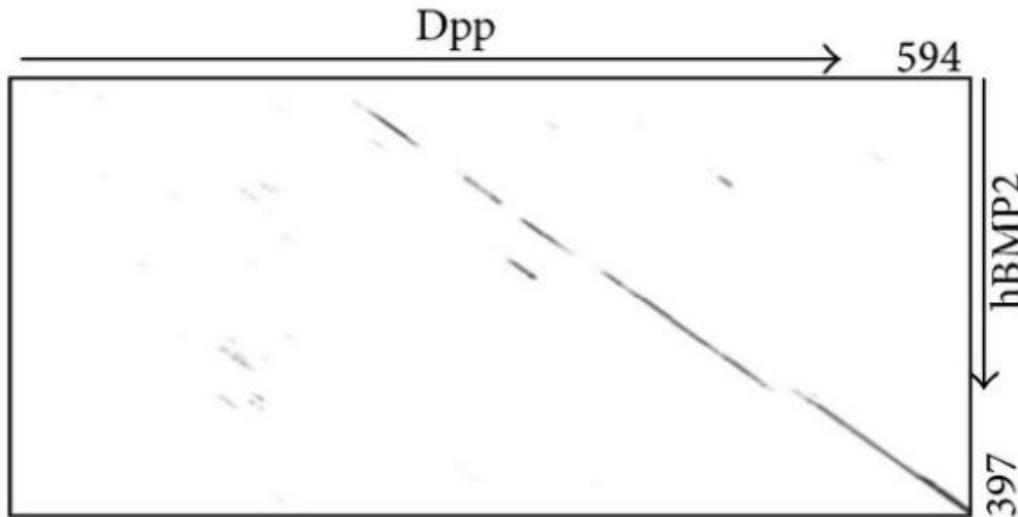
Taškinės matricos

	G	C	T	A	G	T	C	A <th>G</th> <th>A</th> <th>T</th> <th>C</th> <th>T</th> <th>G</th> <th>A</th> <th>C</th> <th>G</th> <th>C</th> <th>T</th> <th>A</th>	G	A	T	C	T	G	A	C	G	C	T	A	
G	*				*				*				*								
A			*			*				*				*							
T		*			*					*				*							*
G	*				*					*				*							
G	*				*					*				*							
T		*			*						*			*							
C	*				*						*				*						
A			*			*					*				*						
C	*				*						*				*						
A			*			*					*				*						
T			*			*					*				*						
C	*				*						*				*						
T			*			*					*				*						
G	*				*						*				*						
C	*				*						*				*						
C	*				*						*				*						
G	*				*						*				*						
C	*				*						*				*						

Taškinės matricos: rezultatų variantai



Kokie yra taškinių matricų privalumai ir trūkumai?



Negrafiniai sekų palyginiai

Sekų palyginiai

- Sekų palyginys (angl. *alignment*) yra metodika, kurios tikslas yra identifikuoti DNR/RNR ar baltymų sekų fragmentus, kurie galėtų dalintis struktūriniais, funkciniais ir evoliuciniais ryšiais.
- Sekų palyginių tipai: globalūs/lokalūs; poriniai/daugybiniai.
- Sekų palyginių pritaikymas: sekų lyginimas, sekų paieška, sekų savybių identifikavimas, sekų šeimų aprašymas, filogenetinė analizė, trimačių struktūrų spėjimas ir t.t.

Lokalūs ir globalūs palyginiai

input
string

HEAGAWGHEEAHGE^GA
PAWHEAEHE

Global alignment

HEAGAWGHEEAHGE^GA
---| - | | - | - | | -- | - | |
---P-AW-H-EA--E-HE

Local alignment

AWGHEEAH
|| - | | | |
AW-HEAEH

Palyginio įvertis

- Paprasta skaičiavimo sistema
 - $N^*(\text{įvertis už sutapimą}) + M^*(\text{įvertis už nesutapimą}) + Z^*(\text{įvertis už tarpą})$
- Afininė tarpo kaina
 - $N^*(\text{įvertis už sutapimą}) + M^*(\text{įvertis už nesutapimą}) + X^*\text{įvertis už tarpo atidarymą} + Q^*\text{įvertis už tarpą}$

N - sutapimų skaičius

M - nesutapimų skaičius

Z - tarpų skaičius

X - tarpo atidarymo įvertis

Q - tarpų skaičius

Globalūs palyginiai

Needleman-Wunsch algoritmas

Needleman, Saul B. & Wunsch, Christian D. (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins". *Journal of Molecular Biology*. **48** (3): 443–53. [doi:10.1016/0022-2836\(70\)90057-4](https://doi.org/10.1016/0022-2836(70)90057-4). PMID 542032

$$\text{score} = \max \begin{cases} F(i-1, j-1) + s(x_i, y_i) \\ F(i, j-1) - \text{gap penalty} \\ F(i-1, j) - \text{gap penalty} \end{cases}$$

$s(x_i, y_i)$ may be +1 or -2
depending on match/mismatch

NW matrica (I)

	G	T	C	G	A	C	G	C	A
0	-2	-4	-6	-8	-10	-12	-14	-16	-18
G	1	-1	-3	-5	-7	-9	-11	-13	-15
T	-1	2	0	-2	-4	-6	-8	-10	-12
A	-3		1	-1	-1	-3	-5	-7	-9
G	-5	-2	-1	2	0	-2	-2	-4	-6
G	-7	-4	-3	0	1	-1	-1	-3	-5
G	-9	-6	-5	-2	-1	0	0	-2	-4
	↑	↑	↑	↑	↑	↑	↑	↑	↑

Sequence 1

GTAGGGCACGAATGACA

Sequence 2

GTCGACGCA

Match Score

1

-1

-2

Mismatch Score

↑

↑

Gap Score

↑

Compute Optimal Alignment

Clear Path

Custom Path

NW matrica (II)

	G	T	C	G	A	C	G	C	A
0	-2	-4	-6	-8	-10	-12	-14	-16	-18
G	1	-1	-3	-5	-7	-9	-11	-13	-15
T	-1	2	0	-2	(empty)	-6	-8	-10	-12
A	-3	0	1	-1	-1	-3	-5	-7	-9
G	-5	-2	-1	2	0	-2	-2	-4	-6
G	-7	-4	-3	0	1	-1	-1	-3	-5
G	-9	-6	-5	-2	-1	0	0	-2	-4
	↑	↑	↑	↑	↑	↑	↑	↑	↑

Sequence 1

GTAGGGCACGAATGACA

Sequence 2

GTCGACGCA

Match Score

1



-1



-2



Compute Optimal Alignment

Clear Path

Custom Path

Mismatch Score

1



-1



Gap Score

NW matrica (III)

	G	T	C	G	A	C	G	C	A
0	-2	-4	-6	-8	-10	-12	-14	-16	-18
G	1	-1	-3	-5	-7	-9	-11	-13	-15
T	-1	2	0	-2	-4	-6	-8	-10	-12
A	-3	0	0	-1	-1	-3	-5	-7	-9
G	-5	-2	-1	2	0	-2	-2	-4	-6
G	-7	-4	-3	0	1	-1	-1	-3	-5
G	-9	-6	-5	-2	-1	0	0	-2	-4
	↑	↑	↑	↑	↑	↑	↑	↑	↑

Sequence 1

GTAGGGCACGAATGACA

Sequence 2

GTCGACGCA

Match Score

1

-1

-2

Mismatch Score

↑

↑

Gap Score

↑

Compute Optimal Alignment

Clear Path

Custom Path

NW matrica (IV)

	G	T	C	G	A	C	G	C	A
0	-2	-4	-6	-8	-10	-12	-14	-16	-18
G	1	-1	-3	-5	-7	-9	-11	-13	-15
T	-1	2	0	-2	-4	-6	-8	-10	-12
A	-3	0	1	-1	-1	-3	(empty)	-7	-9
G	-5	-2	-1	2	0	-2	-2	-4	-6
G	-7	-4	-3	0	1	-1	-1	-3	-5
G	-9	-6	-5	-2	-1	0	0	-2	-4
	↑	↑	↑	↑	↑	↑	↑	↑	↑

Sequence 1

GTAGGGCACGAATGACA

Sequence 2

GTCGACGCA

Match Score

1

-1

-2

Mismatch Score

↑

↑

Gap Score

↑

Compute Optimal Alignment

Clear Path

Custom Path

NW matrica (V)

	G	T	C	G	A	C	G	C	A
0	-2	-4	-6	-8	-10	-12	-14	-16	-18
G	1	-1	-3	-5	-7	-9	-11	-13	-15
T	-1	2	0	-2	-4	-6	-8	-10	-12
A	-3	0	1	-1	-1	-3	-5	-7	-9
G	-5	-2	-1	2	0	-2	-2	-4	-6
G	-7	-4	-3	0	1	(empty)	-1	-3	-5
G	-9	-6	-5	-2	-1	0	0	-2	-4
	↑	↑	↑	↑	↑	↑	↑	↑	↑

Sequence 1

GTAGGGCACGAATGACA

Sequence 2

GTCGACGCA

Match Score

1

-1

-2

Mismatch Score

↑

↑

Gap Score

↑

Compute Optimal Alignment

Clear Path

Custom Path

NW matrica (VI)

	G	T	C	G	A	C	G	C	A
0	-2	-4	-6	-8	-10	-12	-14	-16	-18
G	1	-1	-3	-5	-7	-9	-11	-13	-15
T	-1	2	0	-2	-4	-6	-8	-10	-12
A	-3	0	1	-1	-1	-3	-5	-7	-9
G	-5	-2	-1	2	0	-2	-2	-4	-6
G	-7	(empty)	-3	0	1	-1	-1	-3	-5
G	-9	-6	-5	-2	-1	0	0	-2	-4
	↑	↑	↑	↑	↑	↑	↑	↑	↑

Sequence 1

GTAGGGCACGAATGACA

Sequence 2

GTCGACGCA

Match Score

1

-1

-2

Mismatch Score

↑

↑

Gap Score

↑

Compute Optimal Alignment

Clear Path

Custom Path

NW palyginio atstatymas (I)

	G	T	C	G	A	C	G	C	A	
0	-2	-4	-6	-8	-10	-12	-14	-16	-18	
G	-2	1	-1	-3	-5	-7	-9	-11	-13	-15
T	-4	-1	2	0	-2	-4	-6	-8	-10	-12
A	-6	-3	0	1	-1	-1	-3	-5	-7	-9
G	-8	-5	-2	-1	2	0	-2	-2	-4	-6
G	-10	-7	-4	-3	0	1	-1	-1	-3	-5
G	-12	-9	-6	-5	-2	-1	0	0	-2	-4
C	-14	-11	-8	-5	-4	-3	0	-1	1	-1
A	-16	-13	-10	-7	-6	-3	-2	-1	-1	2
C	-18	-15	-12	-9	-8	-5	-2	-3	0	0
G	-20	-17	-14	-11	-8	-7	-4	-1	-2	-1
A	-22	-19	-16	-13	-10	-7	-6	-3	-2	-1
A	-24	-21	-18	-15	-12	-9	-8	-5	-4	-1
T	-26	-23	-20	-17	-14	-11	-10	-7	-6	-3
G	-28	-25	-22	-19	-16	-13	-12	-9	-8	-5
A	-30	-27	-24	-21	-18	-15	-14	-11	-10	-7
C	-32	-29	-26	-23	-20	-17	-14	-13	-10	-9
A	-34	-31	-28	-25	-22	-19	-16	-15	-12	-9

NW palyginio atstatymas (II)

G T C G - - - A C G - - - - C A
 G T A G G G C A C G A A T G A C A

	G	T	C	G	A	C	G	C	A	
0	-2	-4	-6	-8	-10	-12	-14	-16	-18	
G	-2	1	-1	-3	-5	-7	-9	-11	-13	-15
T	-4	-1	2	0	-2	-4	-6	-8	-10	-12
A	-6	-3	0	1	-1	-1	-3	-5	-7	-9
G	-8	-5	-2	-1	2	0	-2	-2	-4	-6
G	-10	-7	-4	-3	0	1	-1	-1	-3	-5
G	-12	-9	-6	-5	-2	-1	0	0	-2	-4
C	-14	-11	-8	-5	-4	-3	0	-1	1	-1
A	-16	-13	-10	-7	-6	-3	-2	-1	-1	2
C	-18	-15	-12	-9	-8	-5	-2	-3	0	0
G	-20	-17	-14	-11	-8	-7	-4	-1	-2	-1
A	-22	-19	-16	-13	-10	-7	-6	-3	-2	-1
A	-24	-21	-18	-15	-12	-9	-8	-5	-4	-1
T	-26	-23	-20	-17	-14	-11	-10	-7	-6	-3
G	-28	-25	-22	-19	-16	-13	-12	-9	-8	-5
A	-30	-27	-24	-21	-18	-15	-14	-11	-10	-7
C	-32	-29	-26	-23	-20	-17	-14	-13	-10	-9
A	-34	-31	-28	-25	-22	-19	-16	-15	-12	-9

Kuris geresnis?

Seka A: GATTACA

Seka B: GTCGACGCA

Palyginys A

G	T	C	G	A	C	G	C	A
G	A	T	T	A	C	-	-	A

Palyginys B

G	-	T	C	G	A	C	G	C	A
G	A	T	T	-	A	C	-	-	A

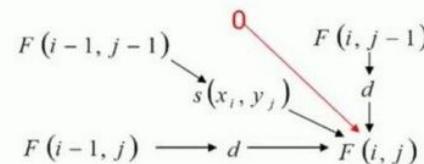
Lokalūs palyginiai

Smith-Waterman algoritmas

Smith, Temple F. & Waterman, Michael S. (1981). ["Identification of Common Molecular Subsequences"](#) (PDF). *Journal of Molecular Biology*. **147** (1): 195–197. [CiteSeerX 10.1.1.63.2897](#). doi:[10.1016/0022-2836\(81\)90087-5](#). PMID [7265238](#)

$$F(0,0) = 0$$

$$F(i, j) = \max \begin{cases} F(i-1, j-1) + s(x_i, y_j) \\ F(i-1, j) + d \\ F(i, j-1) + d \\ 0 \end{cases}$$



SW matrica (I)

S		A_1	A_2	C_3	G_4	A_5	C_6	T_7	A_8	A_9	G_{10}	C_{11}	A_{12}	
	0	0	0	0	0	0	0	0	0	0	0	0	0	
A_1	0	1	1	0	0	1	0	0	1	1	0	0	1	
A_2	0	1	2	0	0	1	0	0	1	2	0	0	1	
T_3	0	0	0	1	0	0	0	1	0	0	1	0	0	
C_4	0	0	0	1	0	0	1	0	0	0	0	2	0	
G_5	0	0	0	0	0	2	0	0	0	0	1	0	1	
G_6	0	0			0	1	1	0	0	0	0	1	0	0
T_7	0	0	0	0	0	0	0	0	1	0	0	0	0	0
A_8	0	1	1	0	0	1	0	0	2	1	0	0	1	
C_9	0	0	0	2	0	0	2	0	0	1	0	1	0	
G_{10}	0	0	0	0	3	1	0	1	0	0	2	0	0	
A_{11}	0	1	1	0	1	4	2	0	2	1	0	1	1	
A_{12}	0	1	2	0	0	2	3	1	1	3	1	0	2	

Sequence a :

AATCGGTACGAA

Sequence b :

AACGACTAAGCA

Scoring in s :

Match 1



Mismatch -1



Gap -2



SW matrica (II)

S		A_1	A_2	C_3	G_4	A_5	C_6	T_7	A_8	A_9	G_{10}	C_{11}	A_{12}
	0	0	0	0	0	0	0	0	0	0	0	0	0
A_1	0	1	1	0	0	1	0	0	1	1	0	0	1
A_2	0	1	2	0	0	1	0	0	1	2	0	0	1
T_3	0	0	0	1	0	0	0	1	0	0	1	0	0
C_4	0	0	0	1	0	0	1	0	0	0	0	2	0
G_5	0	0	0	0	2	0	0	0	0	0	1	0	1
G_6	0	0	0	0	1	1	0	0	0	0	1	0	0
T_7	0	0	0	0	0	0	0	1	0	0	0	0	0
A_8	0	1	1	0	0	1	0	0	2	1	0	0	1
C_9	0	0	0	0	0	0	2	0	0	1	0	1	0
G_{10}	0	0	0	0	3	1	0	1	0	0	2	0	0
A_{11}	0	1	1	0	1	4	2	0	2	1	0	1	1
A_{12}	0	1	2	0	0	2	3	1	1	3	1	0	2

Sequence a :

AATCGGTACGAA

Sequence b :

AACGACTAAGCA

Scoring in s :

Match 1



Mismatch -1



Gap -2



SW matrica (III)

S		A_1	A_2	C_3	G_4	A_5	C_6	T_7	A_8	A_9	G_{10}	C_{11}	A_{12}
	0	0	0	0	0	0	0	0	0	0	0	0	0
A_1	0	1	1	0	0	1	0	0	1	1	0	0	1
A_2	0	1	2	0	0	1	0	0	1	2	0	0	1
T_3	0	0	0	1	0	0	0	1	0	0	1	0	0
C_4	0	0	0	1	0	0	1	0	0	0	0	2	0
G_5	0	0	0	0	2	0	0	0	0	0	1	0	1
G_6	0	0	0	0	1	1	0	0	0	0	1	0	0
T_7	0	0	0	0	0	0	0	1	0	0	0	0	0
A_8	0	1	1	0	0	1	0	0	2	1	0	0	1
C_9	0	0	0	2	0	0	2	0	0	1	0	1	0
G_{10}	0	0	0	0	3	1	0	1	0	0	2	0	0
A_{11}	0	1	1	0	1	4	0	2	1	0	1	1	1
A_{12}	0	1	2	0	0	2	3	1	1	3	1	0	2

Sequence a :

AATCGGTACGAA

Sequence b :

AACGACTAAGCA

Scoring in s :

Match 1

Mismatch -1

Gap -2

SW palyginiai

S		A_1	A_2	C_3	G_4	A_5	C_6	T_7	A_8	A_9	G_{10}	C_{11}	A_{12}
	0	0	0	0	0	0	0	0	0	0	0	0	0
A_1	0	1	1	0	0	1	0	0	1	1	0	0	1
A_2	0	1	2	0	0	1	0	0	1	2	0	0	1
T_3	0	0	0	1	0	0	0	1	0	0	1	0	0
C_4	0	0	0	1	0	0	1	0	0	0	0	2	0
G_5	0	0	0	0	2	0	0	0	0	0	1	0	1
G_6	0	0	0	0	1	1	0	0	0	0	1	0	0
T_7	0	0	0	0	0	0	0	0	1	0	0	0	0
A_8	0	1	1	0	0	1	0	0	2	1	0	0	1
C_9	0	0	0	2	0	0	2	0	0	1	0	1	0
G_{10}	0	0	0	0	3	1	0	1	0	0	2	0	0
A_{11}	0	1	1	0	1	4	2	0	2	1	0	1	1
A_{12}	0	1	2	0	0	2	3	1	1	3	1	0	2

Galimi keli alternatyvūs geriausi rezultatai (tieka NW, tiek SW)

GCATAGATGC
*** | *****
GCAAAGATGC

S	G ₁	T ₂	A ₃	T ₄	G ₅	G ₆	C ₇	A ₈	A ₉	A ₁₀	G ₁₁	A ₁₂	T ₁₃	G ₁₄	C ₁₅
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A ₁	0	0	0	1	0	0	0	0	1	1	1	0	1	0	0
T ₂	0	0	1	0	2	0	0	0	0	0	0	0	0	2	0
G ₃	0	1	0	0	0	3	1	0	0	0	0	1	0	0	3
A ₄	0	0	0	1	0	1	2	0	1	1	1	0	2	0	1
C ₅	0	0	0	0	0	0	0	3	1	0	0	0	0	1	0
A ₆	0	0	0	1	0	0	0	1	4	2	1	0	1	0	0
T ₇	0	0	1	0	2	0	0	0	2	3	1	0	0	2	0
G ₈	0	1	0	0	0	3	1	0	0	1	2	2	0	0	3
C ₉	0	0	0	0	0	1	2	2	0	0	0	1	1	0	1
A ₁₀	0	0	0	1	0	0	0	1	3	1	1	0	2	0	0
T ₁₁	0	0	1	0	2	0	0	0	1	2	0	0	0	3	1
A ₁₂	0	0	0	2	0	1	0	0	1	2	3	1	1	1	2
G ₁₃	0	1	0	0	1	1	2	0	0	0	1	4	2	0	1
A ₁₄	0	0	0	1	0	0	0	1	1	1	1	2	5	3	1
T ₁₅	0	0	1	0	2	0	0	0	0	0	3	6	4	2	0
G ₁₆	0	1	0	0	0	3	1	0	0	0	1	1	4	7	5
C ₁₇	0	0	0	0	0	1	2	2	0	0	0	0	2	5	8

ATG _CATAGATGC
*** * | *****
ATGGCAAAGATGC

S	G ₁	T ₂	A ₃	T ₄	G ₅	G ₆	C ₇	A ₈	A ₉	A ₁₀	G ₁₁	A ₁₂	T ₁₃	G ₁₄	C ₁₅
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
A ₁	0	0	0	1	0	0	0	0	1	1	1	0	1	0	0
T ₂	0	0	1	0	2	0	0	0	0	0	0	0	0	2	0
G ₃	0	1	0	0	0	3	1	0	0	0	1	0	0	3	1
A ₄	0	0	0	1	0	1	2	0	1	1	1	0	2	0	1
C ₅	0	0	0	0	0	0	3	1	0	0	0	0	1	0	2
A ₆	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
T ₇	0	0	1	0	2	0	0	0	2	3	1	0	0	2	0
G ₈	0	1	0	0	0	3	1	0	0	1	2	2	0	0	3
C ₉	0	0	0	0	0	1	2	2	0	0	0	1	1	0	1
A ₁₀	0	0	0	1	0	0	0	1	3	1	1	0	2	0	0
T ₁₁	0	0	1	0	2	0	0	0	1	2	0	0	0	3	1
A ₁₂	0	0	0	2	0	1	0	0	1	2	3	1	1	1	2
G ₁₃	0	1	0	0	1	1	2	0	0	0	1	4	2	0	1
A ₁₄	0	0	0	1	0	0	0	1	1	1	1	2	5	3	1
T ₁₅	0	0	1	0	2	0	0	0	1	2	3	0	0	3	1
A ₁₆	0	0	0	0	2	0	1	0	0	1	2	3	1	1	2
G ₁₇	0	1	0	0	0	0	3	1	0	0	0	1	1	4	7

Galimi keli rezultatai (SW)

Sequence *a*: TGC GGTTGGAATTCCA
 Sequence *b*: TGC GGAACCTTTCCA
 Scoring in *s*: Match 1 Mismatch -1 Gap -3

<i>S</i>		T ₁	G ₂	C ₃	G ₄	G ₅	A ₆	A ₇	C ₈	C ₉	T ₁₀	T ₁₁	T ₁₂	T ₁₃	C ₁₄	C ₁₅
		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
T ₁	0	1	0	0	0	0	0	0	0	0	1	1	1	1	0	0
G ₂	0	0	2	0	1	1	0	0	0	0	0	0	0	0	0	0
C ₃	0	0	0	3	0	0	0	0	1	1	0	0	0	0	1	1
G ₄	0	0	1	0	4	1	0	0	0	0	0	0	0	0	0	0
G ₅	0	0	1	0	1	5	2	0	0	0	0	0	0	0	0	0
T ₆	0	1	0	0	0	2	4	1	0	0	1	1	1	1	0	0
T ₇	0	1	0	0	0	0	1	3	0	0	1	2	2	2	0	0
G ₈	0	0	2	0	1	1	0	0	2	0	0	0	1	1	1	0
G ₉	0	0	1	1	1	2	0	0	0	1	0	0	0	0	0	0
A ₁₀	0	0	0	0	0	0	3	1	0	0	0	0	0	0	0	0
A ₁₁	0	0	0	0	0	0	1	4	1	0	0	0	0	0	0	0
T ₁₂	0	1	0	0	0	0	0	1	3	0	1	1	1	1	0	0
T ₁₃	0	1	0	0	0	0	0	0	0	2	1	2	2	2	0	0
C ₁₄	0	0	0	1	0	0	0	0	1	1	1	0	1	1	3	1
C ₁₅	0	0	0	1	0	0	0	0	1	2	0	0	0	0	2	4
A ₁₆	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	1

	T ₁	G ₂	C ₃	G ₄	G ₅	A ₆	A ₇	C ₈	C ₉	T ₁₀	T ₁₁	T ₁₂	T ₁₃	C ₁₄	C ₁₅	A ₁₆
	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	0	1	0	0	0	0	0	0	0	0	1	1	1	1	0	0
	0	0	2	0	1	1	0	0	0	0	0	0	0	0	0	0
	0	0	0	3	0	0	0	0	1	1	0	0	0	0	1	1
	0	0	1	0	4	1	0	0	0	0	0	0	0	0	0	0
	0	0	1	0	5	2	0	0	0	0	0	0	0	0	0	0
	0	0	1	0	1	5	2	0	0	0	0	0	0	0	0	0
	0	0	1	0	4	1	0	0	0	0	0	0	0	0	0	0
	0	0	1	0	1	5	2	0	0	0	0	0	0	0	0	0
	0	0	1	0	1	5	2	0	0	0	0	0	0	0	0	0
	0	1	0	0	0	2	4	1	0	0	1	1	1	1	0	0
	0	1	0	0	0	0	1	3	0	0	0	1	2	2	0	0
	0	0	2	0	1	1	0	0	2	0	0	0	0	1	1	0
	0	0	1	1	1	2	0	0	0	1	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	3	1	0	0	0	0	1
	0	0	0	0	0	0	0	0	1	4	1	0	0	0	0	1
	0	0	0	0	0	0	0	0	1	4	1	0	0	0	0	1
	0	1	0	0	0	0	0	0	1	3	0	1	1	1	1	0
	0	1	0	0	0	0	0	0	0	2	1	2	2	2	0	0
	0	0	2	0	1	1	0	0	2	0	0	0	0	1	1	0
	0	0	1	1	1	2	0	0	0	0	1	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	3	1	0	0	0	1
	0	0	0	0	0	0	0	0	1	4	1	0	0	0	0	1
	0	0	0	0	0	0	0	0	1	4	1	0	0	0	0	1
	0	1	0	0	0	0	0	0	1	3	0	1	1	1	1	0
	0	1	0	0	0	0	0	0	0	2	1	2	2	2	0	0
	0	0	2	0	1	1	0	0	2	0	0	0	0	1	1	0
	0	0	1	1	1	2	0	0	0	0	1	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	3	1	0	0	0	1
	0	0	0	0	0	0	0	0	1	4	1	0	0	0	0	1
	0	0	0	0	0	0	0	0	1	4	1	0	0	0	0	1
	0	1	0	0	0	0	0	0	1	3	0	1	1	1	1	0
	0	1	0	0	0	0	0	0	0	2	1	2	2	2	0	0
	0	0	2	0	1	1	0	0	2	0	0	0	0	1	1	0
	0	0	1	1	1	2	0	0	0	0	1	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	3	1	0	0	0	1
	0	0	0	0	0	0	0	0	1	4	1	0	0	0	0	1
	0	0	0	0	0	0	0	0	1	4	1	0	0	0	0	1
	0	1	0	0	0	0	0	0	1	3	0	1	1	1	1	0
	0	1	0	0	0	0	0	0	0	2	1	2	2	2	0	0
	0	0	2	0	1	1	0	0	2	0	0	0	0	1	1	0
	0	0	1	1	1	2	0	0	0	0	1	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	3	1	0	0	0	1
	0	0	0	0	0	0	0	0	1	4	1	0	0	0	0	1
	0	0	0	0	0	0	0	0	1	4	1	0	0	0	0	1
	0	1	0	0	0	0	0	0	1	3	0	1	1	1	1	0
	0	1	0	0	0	0	0	0	0	2	1	2	2	2	0	0
	0	0	2	0	1	1	0	0	2	0	0	0	0	1	1	0
	0	0	1	1	1	2	0	0	0	0	1	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	3	1	0	0	0	1
	0	0	0	0	0	0	0	0	1	4	1	0	0	0	0	1
	0	0	0	0	0	0	0	0	1	4	1	0	0	0	0	1
	0	1	0	0	0	0	0	0	1	3	0	1	1	1	1	0
	0	1	0	0	0	0	0	0	0	2	1	2	2	2	0	0
	0	0	2	0	1	1	0	0	2	0	0	0	0	1	1	0
	0	0	1	1	1	2	0	0	0	0	1	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	3	1	0	0	0	1
	0	0	0	0	0	0	0	0	1	4	1	0	0	0	0	1
	0	0	0	0	0	0	0	0	1	4	1	0	0	0	0	1
	0	1	0	0	0	0	0	0	1	3	0	1	1	1	1	0
	0	1	0	0	0	0	0	0	0	2	1	2	2	2	0	0
	0	0	2	0	1	1	0	0	2	0	0	0	0	1	1	0
	0	0	1	1	1	2	0	0	0	0	1	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	3	1	0	0	0	1
	0	0	0	0	0	0	0	0	1	4	1	0	0	0	0	1
	0	0	0	0	0	0	0	0	1	4	1	0	0	0	0	1
	0	1	0	0	0	0	0	0	1	3	0	1	1	1	1	0
	0	1	0	0	0	0	0	0	0	2	1	2	2	2	0	0
	0	0	2	0	1	1	0	0	2	0	0	0	0	1	1	0
	0	0	1	1	1	2	0	0	0	0	1	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	3	1	0	0	0	1
	0	0	0	0	0	0	0	0	1	4	1	0	0	0	0	1
	0	0	0	0	0	0	0	0	1	4	1	0	0	0	0	1
	0	1	0	0	0	0	0	0	1	3	0	1	1	1	1	0
	0	1	0	0	0	0	0	0	0	2	1	2	2	2	0	0
	0	0	2	0	1	1	0	0	2	0	0	0	0	1	1	0
	0	0	1	1	1	2	0	0	0	0	1	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	3	1	0	0	0	1
	0	0	0	0	0	0	0	0	1	4	1	0	0	0	0	1
	0	0	0	0	0	0	0	0	1	4	1	0	0	0	0	1
	0	1	0	0	0	0	0	0	1	3	0	1	1	1	1	0
	0	1	0	0	0	0	0	0	0	2	1	2	2	2	0	0
	0	0	2	0	1	1	0	0	2	0	0	0	0	1	1	0
	0	0	1	1	1	2	0	0	0	0	1	0	0	0	0	0
	0	0	0	0	0	0	0	0	0	0	3	1	0	0	0	1
	0	0	0	0	0	0	0	0	1	4	1	0	0	0	0	1
	0	0	0	0	0	0	0	0	1	4	1	0	0	0	0	1
	0	1	0	0	0	0	0	0	1	3	0	1	1	1</td		

Daugybiniai sekų
palyginiai (MSA,
multiple sequence
alignment)

Ką daryti, kai sekų
daugiau negu dvi?

MSA (ir panašių) programų tipai

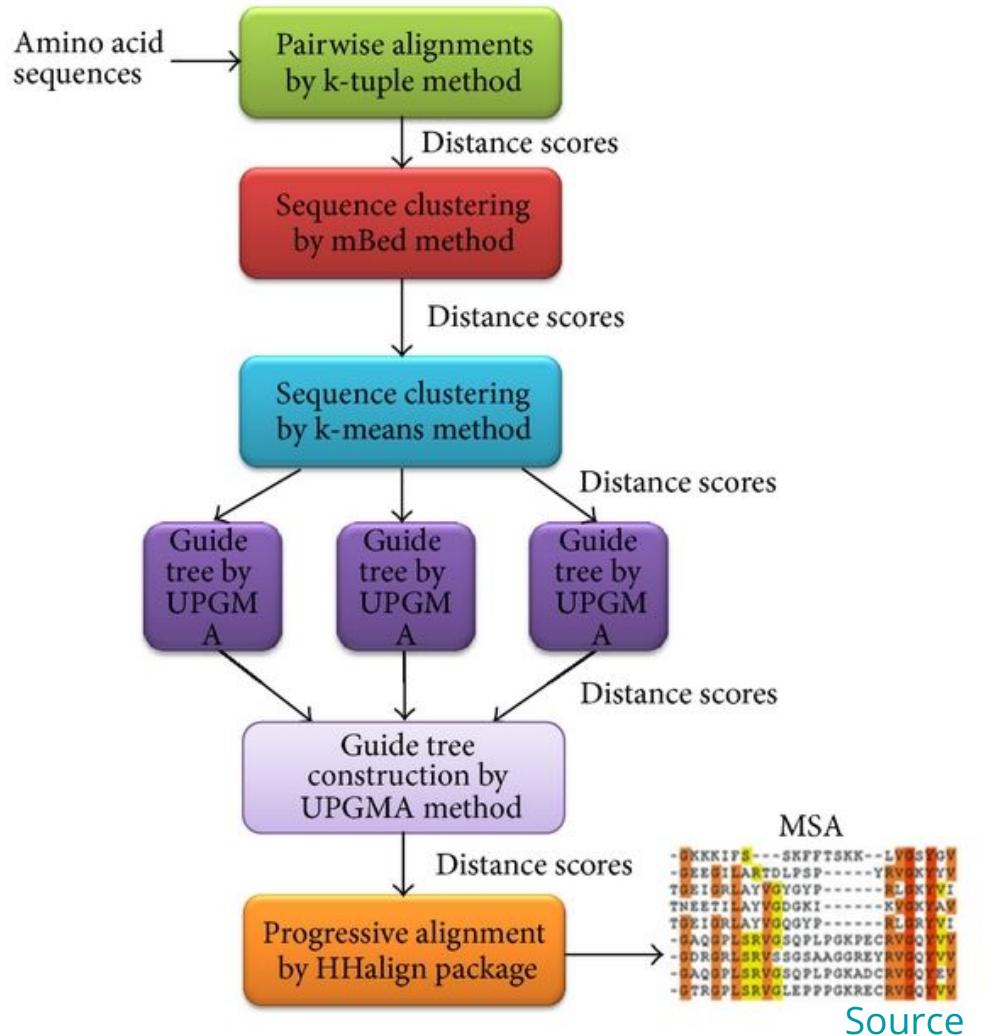
Table 1

Types of multiple sequence alignment and corresponding algorithms.

Types of MSA alignment	MSA algorithms
Pairwise alignment	Needleman-Wunsch, k-mer, k-tuple, and Smith-Waterman algorithms.
Progressive alignment	Clustal Omega, ClustalW, MAFFT, Kalign, Probalign, MUSCLE, Dialign, ProbCons, and MSAProbs.
Iterative progressive alignment	PRRP, MUSCLE, DIALIGN, SAGA, and T-COFFEE.
Homology search tools	BLAST, PSI-BLAST, and FASTA.
Structure incorporating alignment	3D-COFFEE, EXPRESSO, and MICAlign.
Motif alignment	PHI-Blast, GLAM2.
Short-read alignment	Bowtie, Maq, and SOAP.

Clustal Omega algoritmas

1. Poriniai palyginiai pagal k-tuple metodą
2. Sekų grupavimas pagal mBed ir k-means metodus
3. Medžio vedlio konstravimas
4. Palyginio kūrimas su HHalign



MSA

Q5E940_BOVIN -----MPREDRATWKSNSYFLKI1QLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
 RLA0_HUMAN -----MPREDRATWKSNSYFLKI1QLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
 RLA0_MOUSE -----MPREDRATWKSNSYFLKI1QLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
 RLA0_RAT -----MPREDRATWKSNSYFLKI1QLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
 RLA0_CHICK -----MPREDRATWKSNSYFLKI1QLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
 RLA0_RANSY -----MPREDRATWKSNSYFLKI1QLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--SALE
 Q7ZUG3_BRARE -----MPREDRATWKSNSYFLKI1QLLDDYPKCFIVGADNVGSKQMQTIRLSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
 RLA0_ICTPU -----MPREDRATWKSNSYFLKI1QLLDDYPKCFIVGADNVGSKQMQTIRLSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
 RLA0_DROME -----MYRENKAANKAQYIFIKVVELFDEFPKCFIVGADNVGSKQMONIRSLRGL-AVVLMGKNTMMRKAIRGHLENN--PQLE
 RLA0_DICDI -----MSGAG-SKRKKLFIEKATKLFTTYDKMIVAEADFYGSQLOKIRKSIRGI-GAVLMGKKIMIRKVIRDLADSK--PELD
 Q54LP0_DICDI -----MSGAG-SKRKNVFIIEKATKLFTTYDKMIVAEADFYGSQLOKIRKSIRGI-GAVLMGKKIMIRKVIRDLADSK--PELD
 RLA0_PLAF8 -----MAKLSKQKKQMYIEKTLSSLQYDVKCIVGADNVGSKQMLRIRTLALKNLQAV--POI
 RLA0_SULAC -----MIGLAVTTKKIAKKVDEVAELTEKLTHHTKTLIIANIEGFPADKLHEIRKKLRGK-ADIKVTKNNLFNIALKNAG----YDK
 RLA0_SULTO -----MRIMAVITQERKIAKKVIEEVKEELQKREYHTTIIANIEGFPADKLHDIRKKMREGM-AEIKVTKNTLFIAAKNAG----LDVS
 RLA0_SULSO -----MKRLALALKQRKVASWKLEEVEKELTELIKNSNTILIGLGNLEGFPADKLHEIRKKLRGK-AEIKVTKNTLFIAAKNAG----IDIE
 RLA0_AERPE MSVVSLVGQMYKREKDIPENKTLMRELELFSKHRVVFADLTGTFVVDYRVRKWLWKK-YPMHMVAKRRIILRAMKAAGLE--LDDN
 RLA0_PYRAE -MMLAIGKRRYYRTRQY PARKVVISEATELLQKPYVYFLFDLHGLSRLIRLHEYRYRLRRY-GVVIKIKPTLFKIAFTKVYGG--IPAE
 RLA0_METAC -----MAEERRHHTEIPQWKKDEIENIKELIQSHKVFGMVGIEGLLATKMKIRRDLKDY-AVLKVSRNTLTERALNQLG----ETIP
 RLA0_METMA -----MAEERRHHTEIPQWKKDEIENIKELIQSHKVFGMVRIEGILATKICKIRRDLKDY-AVLKVSRNTLTERALNQLG----ESIP
 RLA0_ARCFU -----MAAVRGS--PPEYXVRRAVEEIKRMISSSKPVVAVISPRNVPAGQMOXIRREFRGK-AEIKVVKNTLLERALDALG--GDYL
 RLA0_METKA MAVKAKGOPPSGYEPKVAEWKRREVYKEKELMDYEENGLVVDLEGIPAPQOEIRAKLRERDYIIRMSRNTLMRJAEEKKLER--PELE
 RLA0_METTH -----MAHVAEWKKKEVOELNDLIGYEVVGVIANLADIPAROLQKMRQTLDSD-ALIRMSKXILISLALEKAGREL--ENVD
 RLA0_METTL -----MITAESEHKIAPWKIEEVNKLKELLKNGQIVALVDMMEVPARQOEIRDKIR-GTMILKMSRNTLIERAIKEVAEETGNPEFA
 RLA0_METVA -----MIDAKSEHKIAPWKIEEVNALKELLKSANVILIDMMEVPARQOEIRDKIR-DQMTLKMSRNTLIKRAVEEEVVAEETGNPEFA
 RLA0_METJA -----METKVKAHVAPWKEEYVTLKGLIKSKPVVAVIYDMDVYAPQOEIRDKIR-DKVYKLKMSRNTLIIRALKEAAEELNNPKLA



BLAST: principai ir naudojimas

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.

[Learn more](#)

N
E
W
S

BLAST+ 2.13.0 is here!

Starting with this release, we are including the blastn_vdb and tblastn_vdb executables in the BLAST+ distribution.

Thu, 17 Mar 2022 12:00:00 EST

[More BLAST news...](#)

Web BLAST



BLAST Genomes

Enter organism common name, scientific name, or tax id

Search

Human

Mouse

Rat

Microbes

Kaip pasirinkti BLAST variantą?

Basic BLAST

Choose a BLAST program to run.

[nucleotide blast](#)

Search a **nucleotide** database using a **nucleotide** query
Algorithms: blastn, megablast, discontiguous megablast

[protein blast](#)

Search **protein** database using a **protein** query
Algorithms: blastp, psi-blast, phi-blast

[blastx](#)

Search **protein** database using a **translated nucleotide** query

[tblastn](#)

Search **translated nucleotide** database using a **protein** query

[tblastx](#)

Search **translated nucleotide** database using a **translated nucleotide** query

Paprasta BLAST paieška

Standard Protein BLAST

blastn blastp blasts tblastn tblastx

Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
 From To

Or, upload file No file selected.

Job Title Enter a descriptive title for your BLAST search

Align two or more sequences

BLAST results will be displayed in a new format by default
You can always switch back to the Traditional Results page.


Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism Optional: Enter organism name or id—completions will be suggested exclude

Exclude Optional: Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample sequences

Program Selection

Algorithm:

- Quick BLASTP (Accelerated protein-protein BLAST)
- blastp (protein-protein BLAST)
- PSI-BLAST (Position-Specific Iterated BLAST)
- PHI-BLAST (Pattern Hit Initiated BLAST)
- DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

BLAST Search database nr using Blastp (protein-protein BLAST)
 Show results in a new window

Čia yra help'as.
Naudokitės juo

BLAST rezultatai

BLAST® » blastp suite » results for RID-6C2NWAK0014

Home Recent Results Saved Strategies Help

[Edit Search](#)

[Save Search](#) [Search Summary](#)

[How to read this report?](#) [BLAST Help Videos](#) [Back to Traditional Results Page](#)

Job Title **sp|P01013|**

RID **6C2NWAK0014** Search expires on 03-10 17:22 pm [Download All](#)

Program **BLASTP** [Citation](#)

Database **nr** [See details](#)

Query ID **P01013.1**

Description RecName: Full=Ovalbumin-related protein X; AltName: Full=G ...

Molecule type amino acid

Query Length 232

Other reports [Distance tree of results](#) [Multiple alignment](#) [MSA viewer](#) [?](#)

Filter Results

Organism only top 20 will appear exclude

Type common name, binomial, taxid or group name

[+ Add organism](#)

Percent Identity

 to

E value

 to

Query Coverage

 to

[Filter](#)

[Reset](#)

[Descriptions](#)

[Graphic Summary](#)

[Alignments](#)

[Taxonomy](#)

Sequences producing significant alignments

[Download](#)

[Manage Columns](#)

Show

100

[?](#)

select all 100 sequences selected

[GenPept](#) [Graphics](#) [Distance tree of results](#) [Multiple alignment](#)

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
<input checked="" type="checkbox"/>	RecName: Full=Ovalbumin-related protein X; AltName: Full=Gene X protein [Gallus gallus]	481	481	100%	1e-171	100.00%	P01013.1
<input checked="" type="checkbox"/>	ovalbumin-related protein X [Gallus gallus]	477	477	100%	8e-168	98.71%	NP_001263315.1
<input checked="" type="checkbox"/>	ovalbumin-related protein X isoform X3 [Gallus gallus]	477	477	100%	2e-167	98.71%	XP_015137661.1
<input checked="" type="checkbox"/>	ovalbumin-related protein X isoform X2 [Gallus gallus]	476	476	100%	3e-167	98.71%	XP_015137660.1
<input checked="" type="checkbox"/>	ovalbumin-related protein X isoform X1 [Gallus gallus]	478	478	100%	5e-167	98.71%	XP_015137659.1
<input checked="" type="checkbox"/>	hypothetical protein CIB84_014810 [Bambusicola thoracicus]	442	442	100%	4e-155	91.81%	POI21443.1

BLAST rezultatai

- 4 dalys:
 - Aprašas
 - Grafinė santrauka
 - Palyginiai
 - Taksonominė informacija
- Galima modifikuoti
- Galima atsiųsti pas save

BLAST rezultatai

Lentelė

Surastų sekų pavadinimai

Surastų sekų statistiniai
jverčiai

Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
<input checked="" type="checkbox"/> RecName: Full=Ovalbumin-related protein X; AltName: Full=Gene X protein [Gallus gallus]	481	481	100%	1e-171	100.00%	P01013.1
<input checked="" type="checkbox"/> ovalbumin-related protein X [Gallus gallus]	477	477	100%	8e-168	98.71%	NP_001263315.1
<input checked="" type="checkbox"/> ovalbumin-related protein X isoform X3 [Gallus gallus]	477	477	100%	2e-167	98.71%	XP_015137661.1
<input checked="" type="checkbox"/> ovalbumin-related protein X isoform X2 [Gallus gallus]	476	476	100%	3e-167	98.71%	XP_015137660.1
<input checked="" type="checkbox"/> ovalbumin-related protein X isoform X1 [Gallus gallus]	478	478	100%	5e-167	98.71%	XP_015137659.1
<input checked="" type="checkbox"/> hypothetical protein CIB84_014810 [Bambusicola thoracicus]	442	442	100%	4e-155	91.81%	POI21443.1
<input checked="" type="checkbox"/> ovalbumin-related protein X [Meleagris gallopavo]	441	441	100%	1e-152	90.95%	XP_019469333.1
<input checked="" type="checkbox"/> ovalbumin-related protein X isoform X2 [Colurnix japonica]	428	428	100%	1e-148	88.36%	XP_015709962.2
<input checked="" type="checkbox"/> ovalbumin-related protein X isoform X1 [Colurnix japonica]	428	428	100%	3e-148	88.36%	XP_015709960.2
<input checked="" type="checkbox"/> ovalbumin-related protein X-like [Phasianus colchicus]	429	429	100%	6e-148	89.66%	XP_031445131.1

Aktyvios nuorodos

Svarbu! Visas „mėlynas“
tekstas yra nuorodos į kažkur :)

BLAST rezultatai

Lentelė

- Coverage (padengimas) – maksimumas yra 100%. Kuo arčiau 100%, tuo geriau.
- Perc. Identity (identiškumas) - maksimumas yra 100%. Kuo arčiau 100%, tuo geriau. Skirkite ir nemaišykite su panašumu.
- Total score - sum of alignment scores of all segments from the same database sequence that match the query sequence (calculated over all segments)
- Max score - highest alignment score (bit-score) between the query sequence and the database sequence segment

BLAST rezultatai

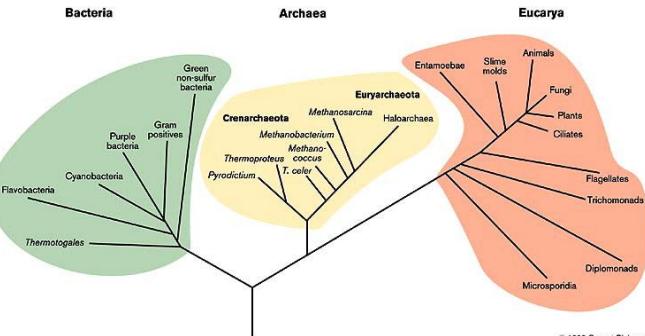
Lentelė

- **E-value** - The Expect value (E) is a parameter that describes the number of hits one can "expect" to see by chance when searching a database of a particular size. It decreases exponentially as the Score (S) of the match increases. Essentially, the E value describes the random background noise. For example, an E value of 1 assigned to a hit can be interpreted as meaning that in a database of the current size one might expect to see 1 match with a similar score simply by chance.
- **SVARBU: trumpos sekos turės didelį E-value tiesiog dėl skaičiavimo specifikų.**

Filogenetika

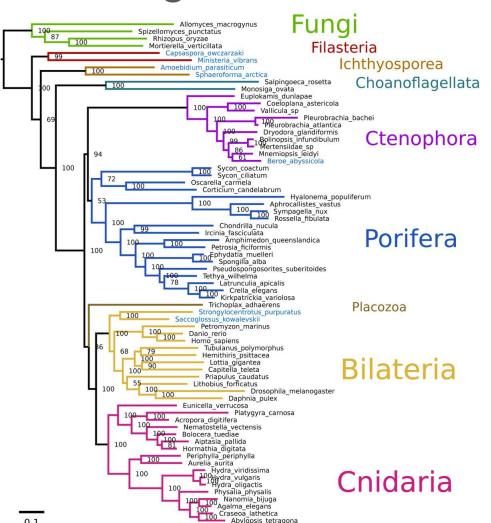
Filogenetika

(Graikiškai: phylon = rūšis and genetic = gimimas)



Filogenetika – tyrinėjimų sritis, kurios tikslas aptikti rūšių evoliucinius sąryšius.

Filogenezė – bendru protėviu grindžiamas organizmu evoliucinis sąryšis.



Homologija

Homologai

Ortologai

Pro-ortologai

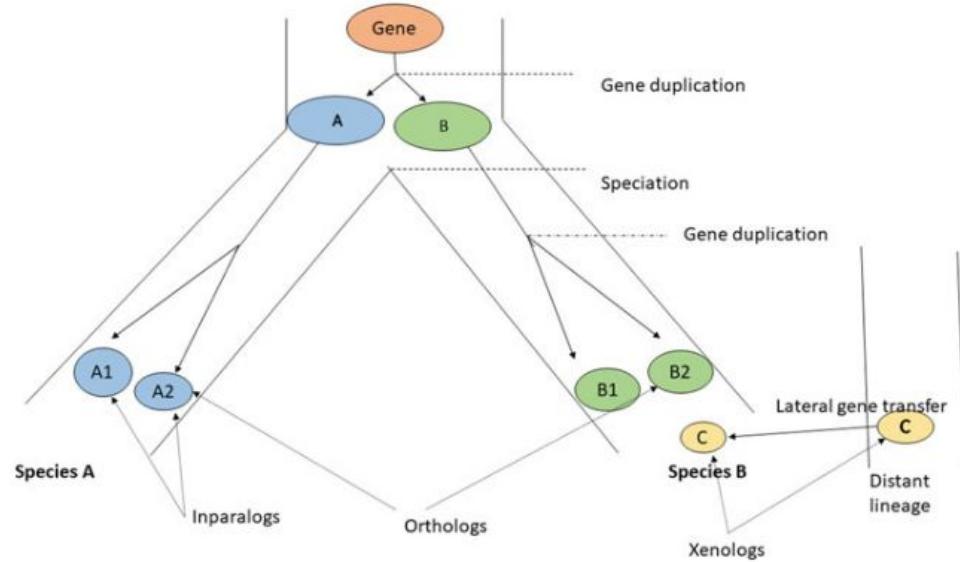
Paralogai

Inparalogai

Outparalogai

Ksenologai

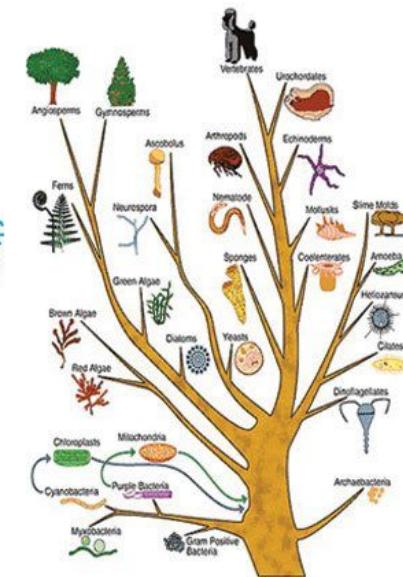
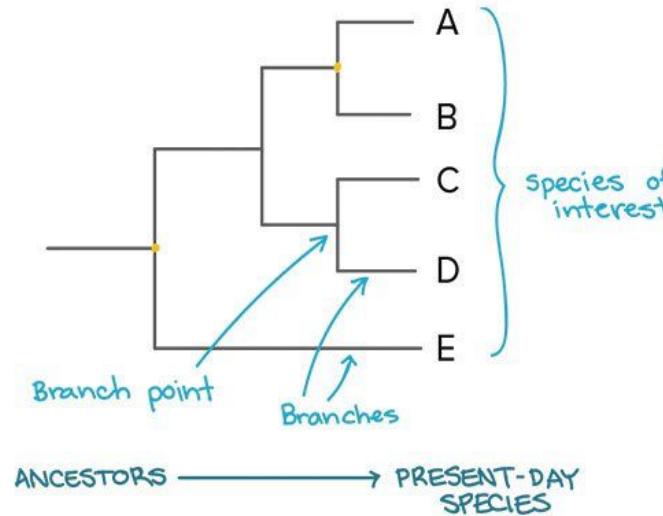
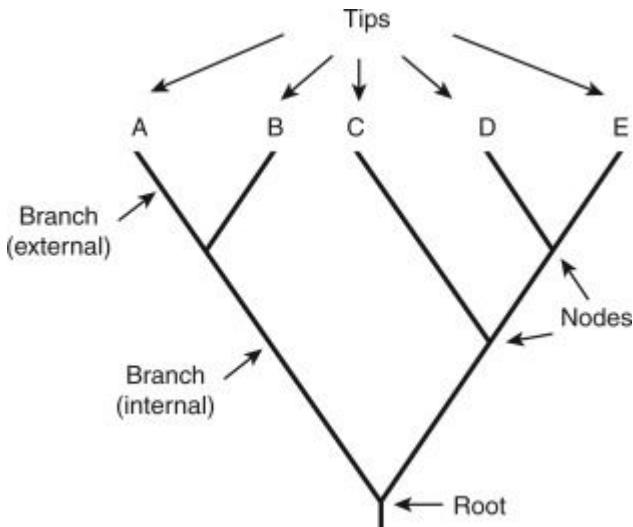
Daliniai homologai



[Source](#)

Filogenetinio medžio dalys

Filogenetinių medžių dalys (I)



source

Filogenetinių medžių dalys (II)

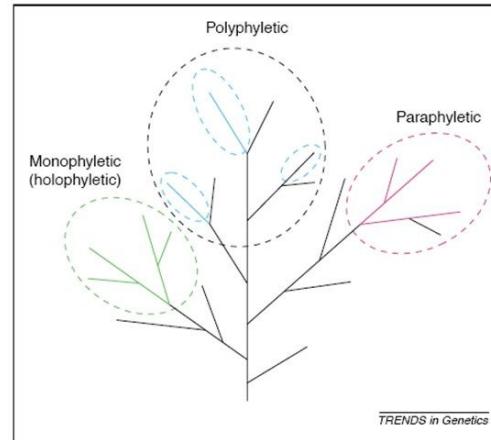
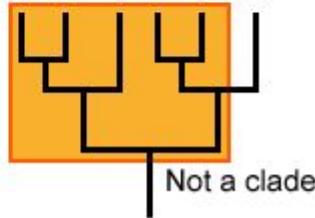
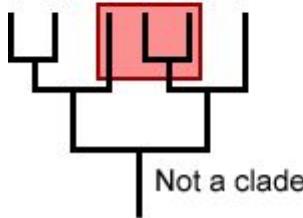
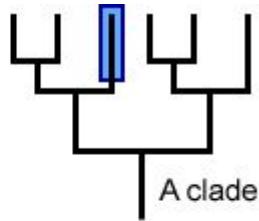
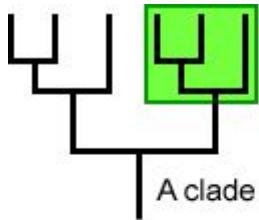


Fig. 1. Trees are about groups: monophyletic (holophyletic), paraphyletic and 'polyphyletic'.

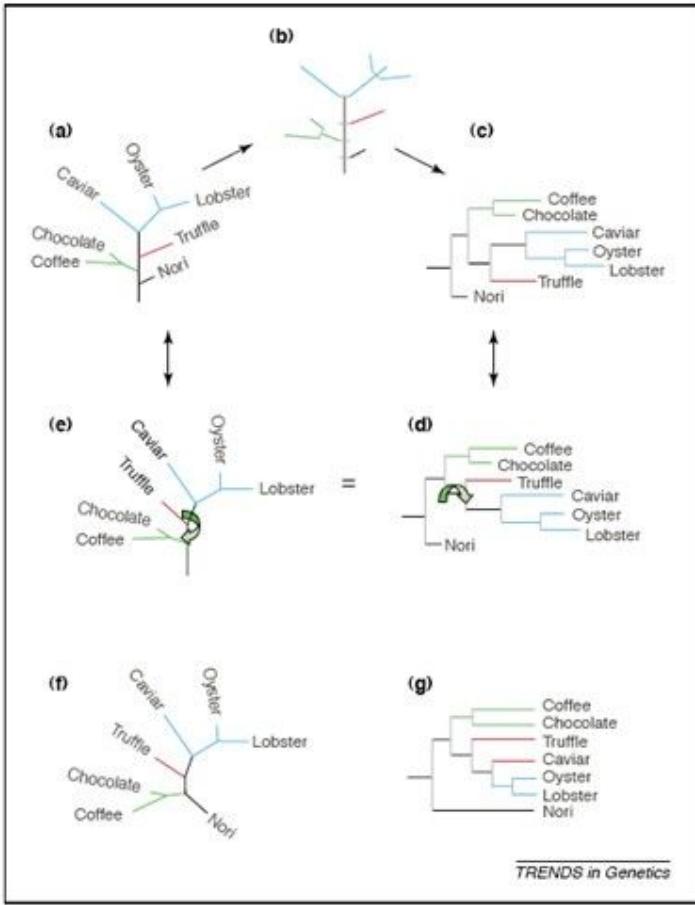
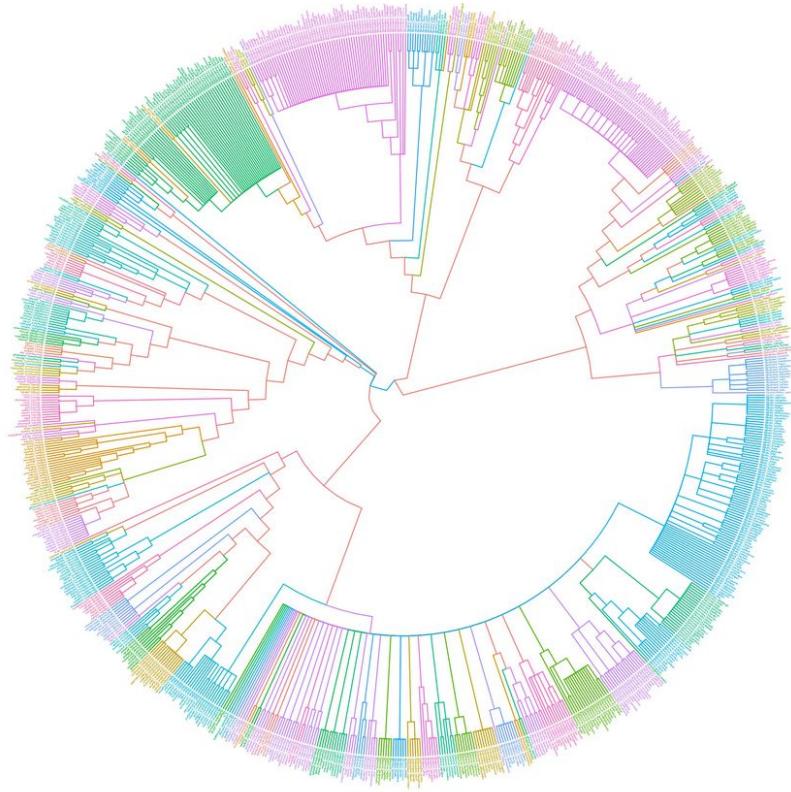
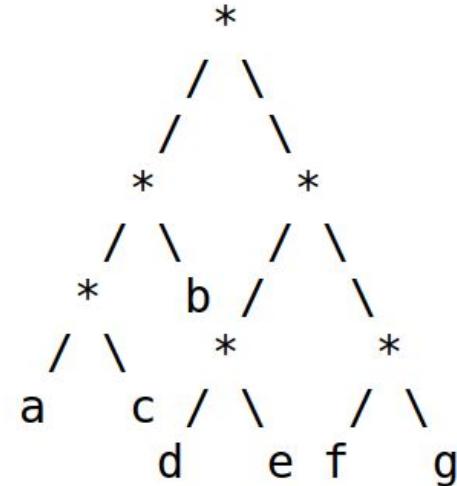
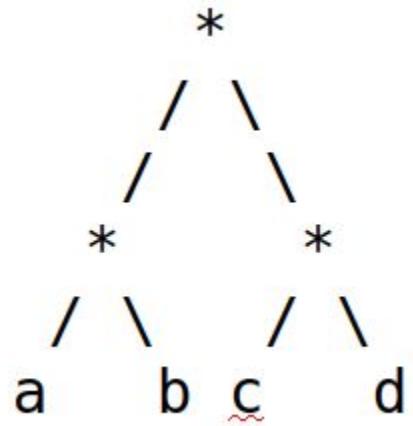


Fig. 2. Phylogenetic tree styles. All these trees have identical branching patterns. The only differences are (f), which is unrooted. (g) is a cladogram, so the branch lengths are right justified and not drawn to scale (i.e. they are not proportional to estimated evolutionary difference).

Kaip kompiuteriu užrašyti filogenetinius medžius?



Newick formatas



Newick formatas

O kaip atrodys medis:

((One:0.2,Two:0.3):0.3,(Three:0.5,Four:0.3):0.2):0.3,Five:0.7):0.0

Newick formatas

```
      +-+ One
      +---+
      |   +---+ Two
      +---+
      |   |   +----+ Three
      |   +--+
      |       +---+ Four
      +
      +-----+ Five
```

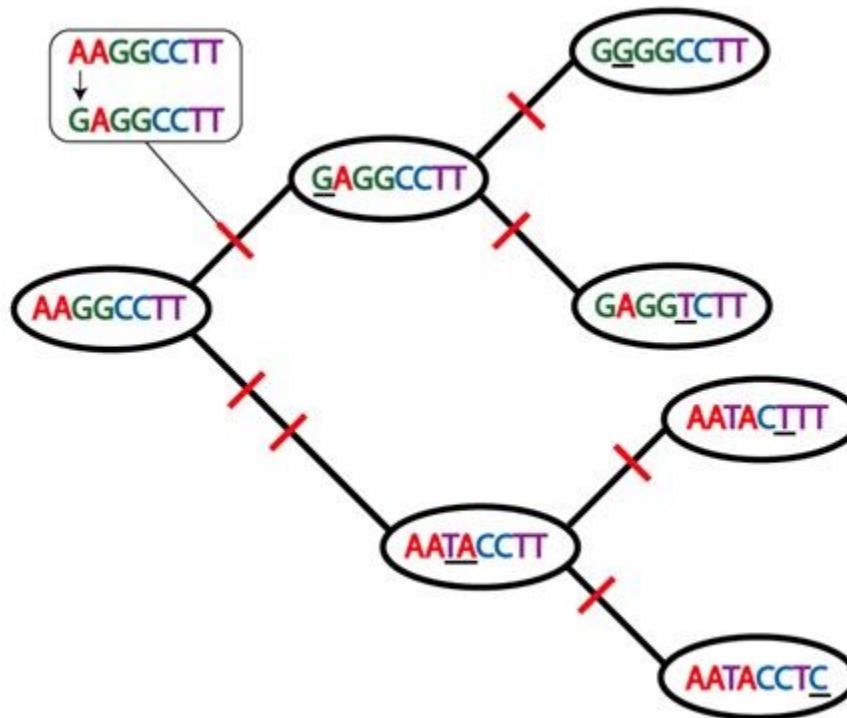
Filogenetinės analizės atlikimo žingsniai

1. Homologų ar molekulinių markerių paieška.
2. Daugybinio palyginio kūrimas.
3. Evoliucinio modelio parinkimas.
4. Medžio kūrimas.
5. Medžio įvertinimas – patikimumo analizė.
6. Rezultatų interpretavimas.

Filogenetinės analizės atlikimo žingsniai

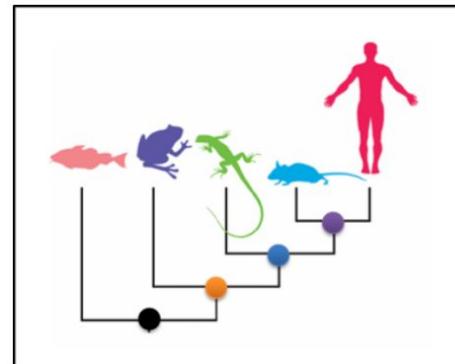
1. Homologų ar molekulinių markerių paieška (BLAST, HMMER, literatūros analizė).
2. Daugybinio palyginio kūrimas (muscle, mafft ir kt.).
3. Evoliucinio modelio parinkimas (modelFinder ir pan.).
4. Medžio kūrimas (IQ-TREE, MrBayes, RAxML).
5. Medžio įvertinimas – patikimumo analizė (bootstrap, jackknife).
6. Rezultatų interpretavimas.

Kaip analizė siejasi su medžiu?

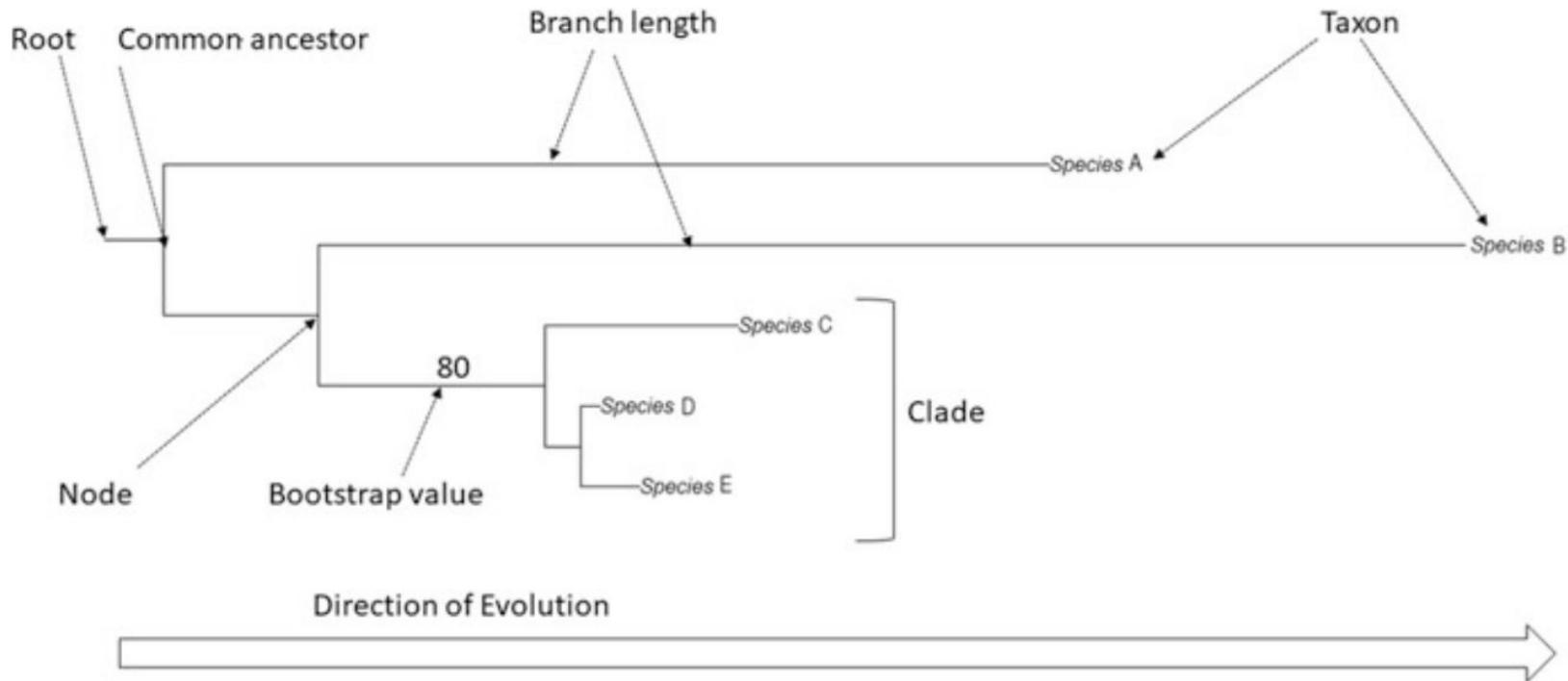


[Source](#)

Filogenetinių medžių interpretavimas

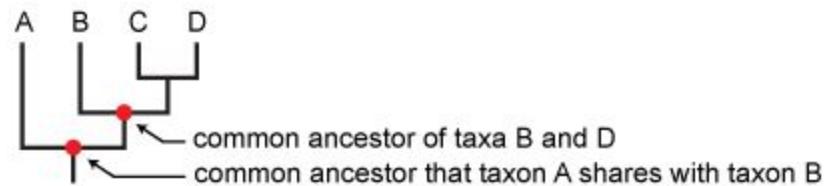


Filogenetinis medis



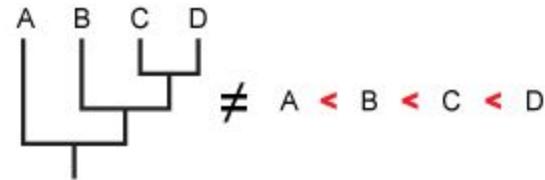
Dažniausios klaidos

Rūšys, kurių pavadinimai filogenetiniame medyje yra greta vienas kito visada yra artimesnės nei rūšys, kurių pavadinimai yra ne greta.



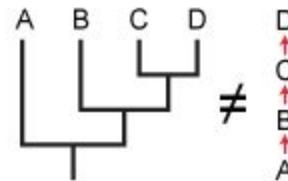
Dažniausios klaidos

Rūšys, kurios yra medžio viršuje/arčiau kažkurių šono yra labiau išsivysčiusios.

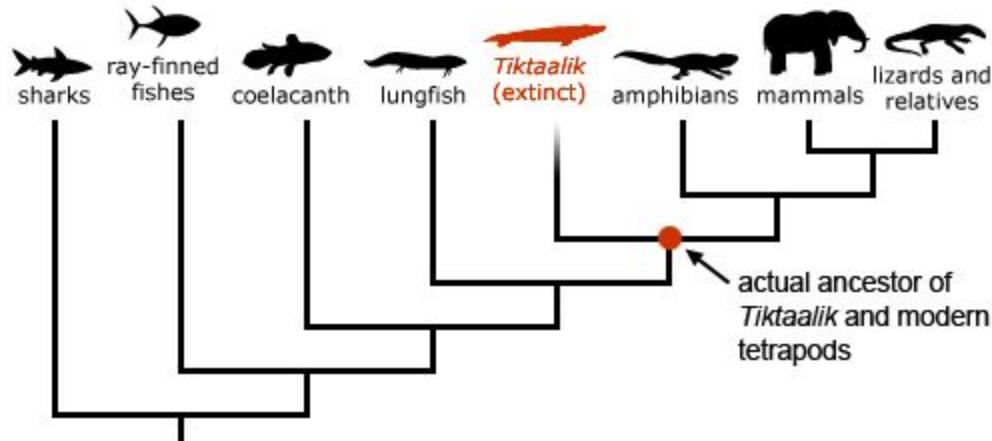


Dažniausios klaidos

Rūšys, kurios yra arčiau medžio pagrindo ar arčiau kairės pusės atspindi kitų medyje atvaizduotų rūšių protėvius.



Dažniausios klaidos



WPGMA metodas
(paprasčiausiu klasterizavimo medžių
pavyzdys - aptarta per pratybas)

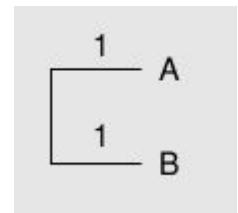
Klasterizavimo metodai - WPGMA metodas

$$d_{uk} = \frac{d_{(A,B)k} + d_{Ck}}{2}$$

WPGMA (I)

	A	B	C	D	E
B	2				
C	4	4			
D	6	6	6		
E	6	6	6	4	
F	8	8	8	8	8

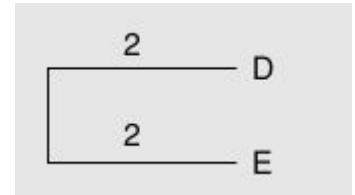
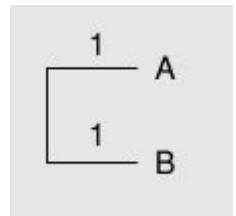
N = 6



WPGMA (II)

$$d_{(AB)C} = (d_{AC} + d_{BC})/2 = 4$$
$$d_{(AB)D} = (d_{AD} + d_{BD})/2 = 6$$
$$d_{(AB)E} = (d_{AE} + d_{BE})/2 = 6$$
$$d_{(AB)F} = (d_{AF} + d_{BF})/2 = 8$$

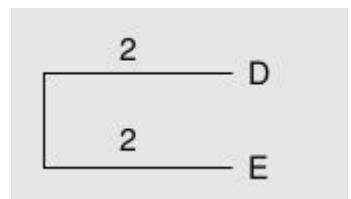
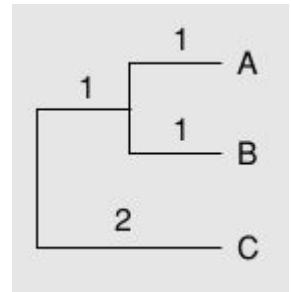
	(AB)	C	D	E
C	4			
D	6	6		
E	6	6	4	
F	8	8	8	8



WPGMA (III)

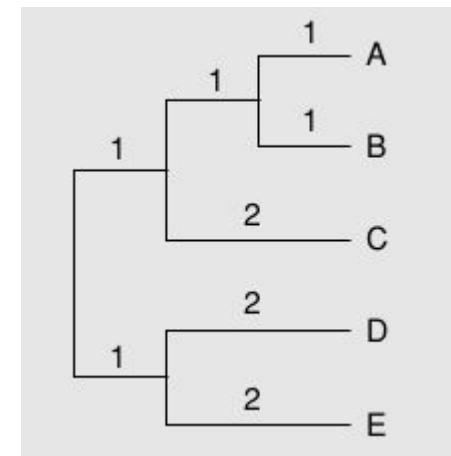
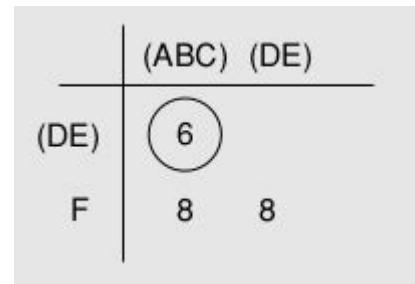
$$d_{(DE)(AB)} = (d_{D(AB)} + d_{E(AB)})/2 = 6$$
$$d_{(DE)C} = (d_{DC} + d_{EC})/2 = 6$$
$$d_{(DE)F} = (d_{DF} + d_{EF})/2 = 8$$

	(AB)	C	(DE)
C	4		
(DE)	6	6	
F	8	8	8



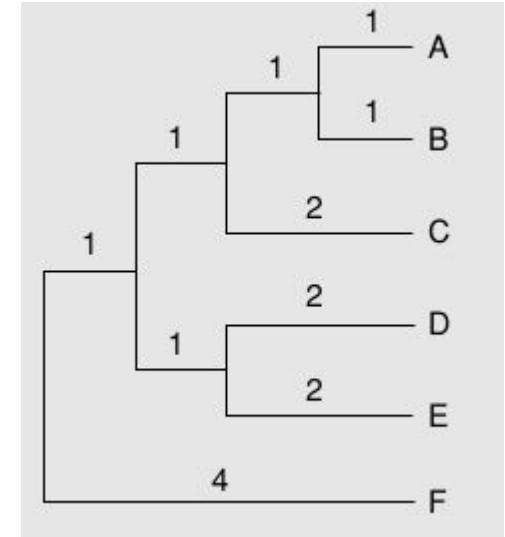
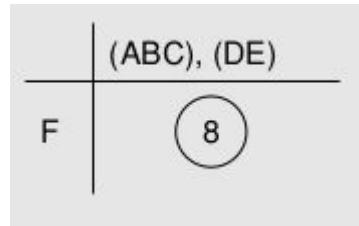
WPGMA (IV)

$$d_{(ABC)(DE)} = (d_{(AB)(DE)} + d_{C(DE)})/2 = 6$$
$$d_{(ABC)F} = (d_{(AB)F} + d_{CF})/2 = 8$$



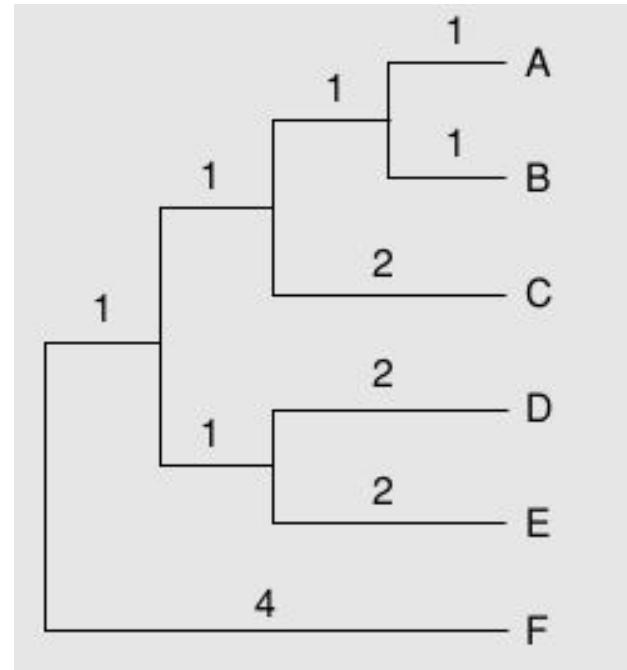
WPGMA (V)

$$d_{(ABCDE)F} = (d_{(ABC)F} + d_{(DE)F})/2 = 8$$



WPGMA (VI)

Galutinis medis (toks pat kaip priešitoje skaidrėje):



Klasterizavimo metodai

- Pliusai:
 - Labai greiti
 - Paprasti suvokti
 - Lengvai programuojami
- Minusai:
 - Labai jautrus nevienodam mutavimo greičiui
 - Dabar nebenaudojamas evoliucinei analizei

Filogenetinių medžių konstravimo metodai

- Metodai paremti atstumais
 - Klasterizavimas (UPGMA, WPGMA)
 - Minimalios evoliucijos, kaimynų jungimo
- Metodai paremti didžiausio tikėtinumo principu
- Metodai paremti šykštumu