

Architecture of large projects in bioinformatics (ADP)

Genome browsers

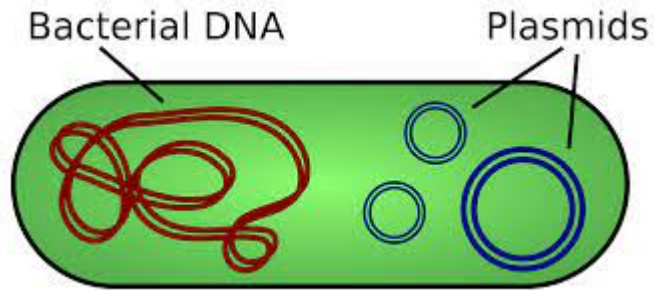
Łukasz P. Kozłowski

Warsaw, 2025

1. Data formats in bioinformatics,
2. Popular software libraries (BioPerl, BioPython)
3. Most important bioinformatics databases (UniProt, PDB, RefSeq, GenBank, ENA, InterPro, etc.)
4. Software licensing for scientific purposes. Free-software licensing. Patents.
5. Generic model Organism database (GMOD) project - assumptions, history and usage
6. Genome browsers, problem description and state of the solutions
7. Version control systems (CVS, SVN, git), and online collaboration and distribution platforms (github, sourceforge).
8. Software testing, automated testing frameworks.
9. Scientific workflow systems - taverna and galaxy. MyExperiment platform. Reproducible research.
10. Literate programming idea and sweave, markdown, software documentation.
11. Interactive scripting platforms, Rstudio, Jupyter.

1. Data formats in bioinformatics,
2. Popular software libraries (BioPerl, BioPython)
3. Most important bioinformatics databases (UniProt, PDB, RefSeq, GenBank, ENA, InterPro, etc.)
4. Software licensing for scientific purposes. Free-software licensing. Patents.
- 5. Generic model Organism database (GMOD) project - assumptions, history and usage**
- 6. Genome browsers, problem description and state of the solutions**
7. Version control systems (CVS, SVN, git), and online collaboration and distribution platforms (github, sourceforge)
8. Software testing, automated testing frameworks.
9. Scientific workflow systems - taverna and galaxy. MyExperiment platform. Reproducible research.
10. Literate programming idea and sweave, markdown, software documentation.
11. Interactive scripting platforms, Rstudio, Jupyter.

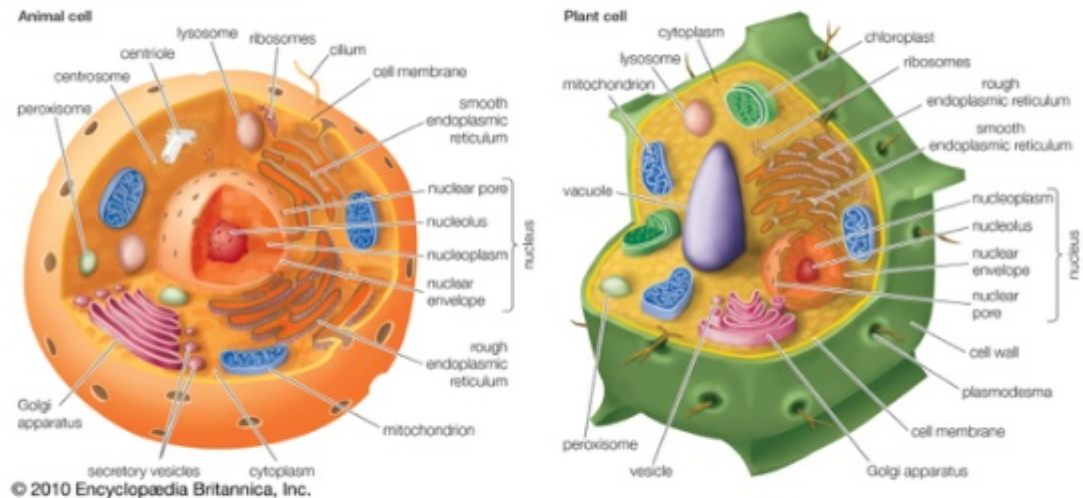
Genome browsers



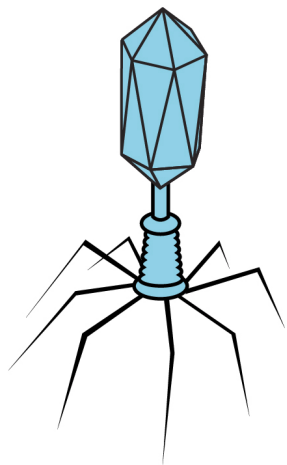
Eukaryotic genomes

Where is eukaryotic genome located?

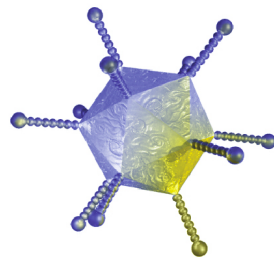
Typical animal cell and plant cell



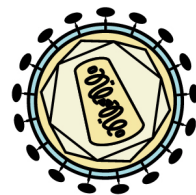
Nuclear genome, mitochondrial genome, chloroplast genome



Bacteriophage



Adenovirus



Human Immunodeficiency Virus

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

<https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi>

The Genetic Codes

The following genetic codes are described here:

- [1. The Standard Code](#)
- [2. The Vertebrate Mitochondrial Code](#)
- [3. The Yeast Mitochondrial Code](#)
- [4. The Mold, Protozoan, and Coelenterate Mitochondrial Code and the Mycoplasma/Spiroplasma Code](#)
- [5. The Invertebrate Mitochondrial Code](#)
- [6. The Ciliate, Dasycladacean and Hexamita Nuclear Code](#)
- [9. The Echinoderm and Flatworm Mitochondrial Code](#)
- [10. The Euplotid Nuclear Code](#)
- [11. The Bacterial, Archaeal and Plant Plastid Code](#)
- [12. The Alternative Yeast Nuclear Code](#)
- [13. The Ascidian Mitochondrial Code](#)
- [14. The Alternative Flatworm Mitochondrial Code](#)
- [16. Chlorophycean Mitochondrial Code](#)
- [21. Trematode Mitochondrial Code](#)
- [22. Scenedesmus obliquus Mitochondrial Code](#)
- [23. Thraustochytrium Mitochondrial Code](#)
- [24. Rhabdopleuridae Mitochondrial Code](#)
- [25. Candidate Division SR1 and Gracilibacteria Code](#)
- [26. Pachysolen tannophilus Nuclear Code](#)
- [27. Karyorelict Nuclear Code](#)
- [28. Condyllostoma Nuclear Code](#)
- [29. Mesodinium Nuclear Code](#)
- [30. Peritrich Nuclear Code](#)
- [31. Blastocrithidia Nuclear Code](#)
- [33. Cephalodiscidae Mitochondrial UAA-Tyr Code](#)

By default all transl_table in GenBank flatfiles are equal to id 1, and this is **not** shown. When transl_table is not equal to id 1, it is shown as a qualifier on the CDS feature.

[Click here to change format](#)

AUG

In rare cases, translation in eukaryotes can be initiated from codons other than AUG. A well documented case (including direct protein sequencing) is the GUG start of a ribosomal P protein of the fungus

Other examples can be found in the following references: [Peabody 1989](#); [Prats et al. 1989](#); [Hann et al. 1992](#); [Sugihara et al. 1990](#). The standard code currently allows initiation from UUG and CUG in addition to AUG.

[Back to top](#)

2. The Vertebrate Mitochondrial Code (transl_table=2)

AAs = FFLLSSSSYY**CCWLLLLLPPPHHQVRRRIIMMTTTTNNKKSS**VVVAAAAADDEEGGGG
Starts = -----**-----MMMM-----**--M-----
Base1 = TTTTTTTTTTTTTTTTCCCCCCCCCCCCCCCCCCAAAAAAAAAAAAAAAAAGGGGGGGGGGGGGGGGG
Base2 = TTTTCCCCAAAAGGGGTTCCTCCCAAAAGGGGTTCCTCCCAAAAGGGGTTCCTCCCAAAAGGGG
Base3 = TCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAG

[Click here to change format](#)

Differences from the Standard Code:

	Code	2		Standard
AGA	Ter	*	Arg	R
AGG	Ter	*	Arg	R
AUA	Met	M	Ile	I
UGA	Trp	W	Ter	*

Alternative Initiation Codons:

Bos: AUA

Homo: AUA, AUU

Mus: AUA, AUU, AUC

Coturnix, Gallus: also GUG ([Desjardins and Morais, 1991](#))

4. The Mold, Protozoan, and Coelenterate Mitochondrial Code and the Mycoplasma/Spiroplasma Code (transl_table=4)

AAs = FFLSSSSYY**CCWLLLLPPPHHQRRRIIIMTTTTNNKKSSRRVVVVAADDEEGGGG
Starts = --MM-----**-M-----MMMM-----M-----
Base1 = TTTTTTTTTTTTTTTTCCCCCCCCCCCCCAAAAAAAAAAAAAAGGGGGGGGGGGGGGGG
Base2 = TTTTCCCAAAAGGGGTTTTCCC AAAAGGGGTTTTCCC AAAAGGGGTTTTCCC AAAAGGGG
Base3 = TCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAG

[Click here to change format](#)

Differences from the Standard Code:

	Code 4	Standard
UGA	Trp W	Ter *

Alternative Initiation Codons:

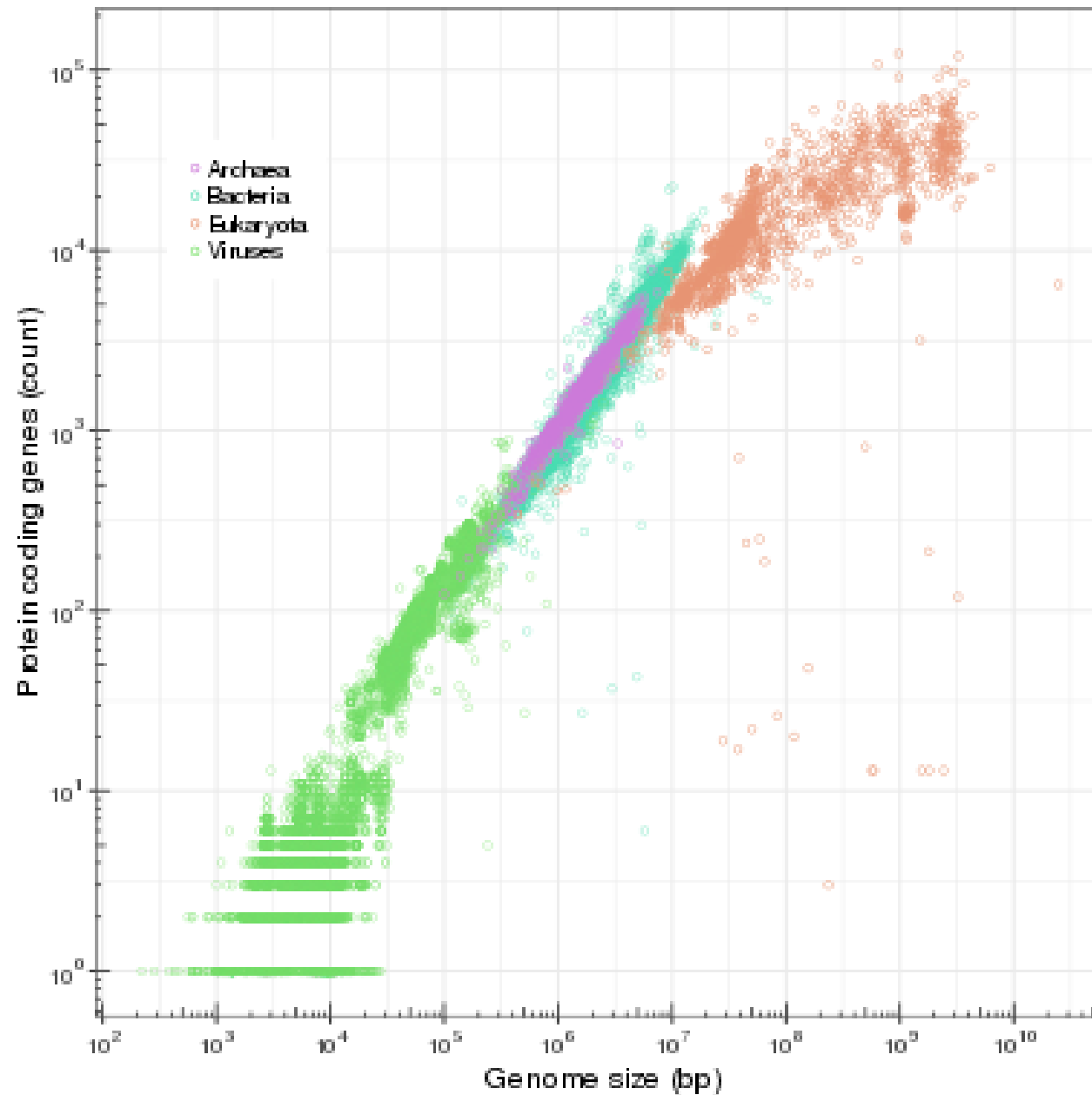
Trypanosoma: UUA, UUG, CUG

Leishmania: AUU, AUA

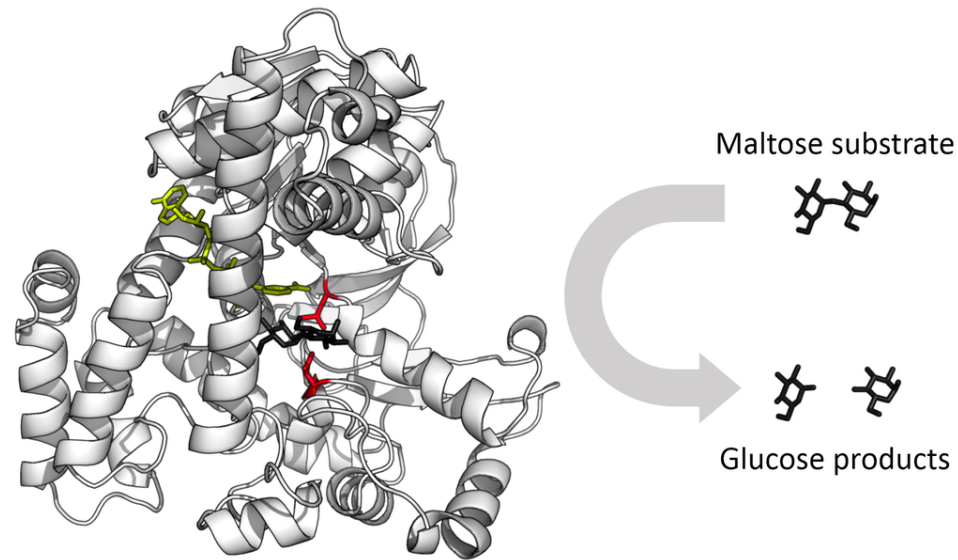
Tertrahymena: AUU, AUA, AUG

Paramecium: AUU, AUA, AUG, AUC, GUG, GUA(?)

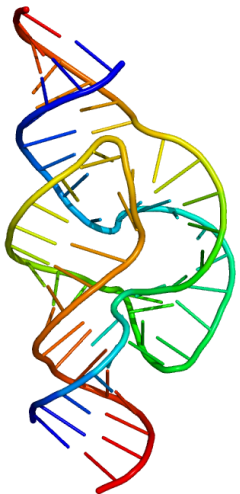
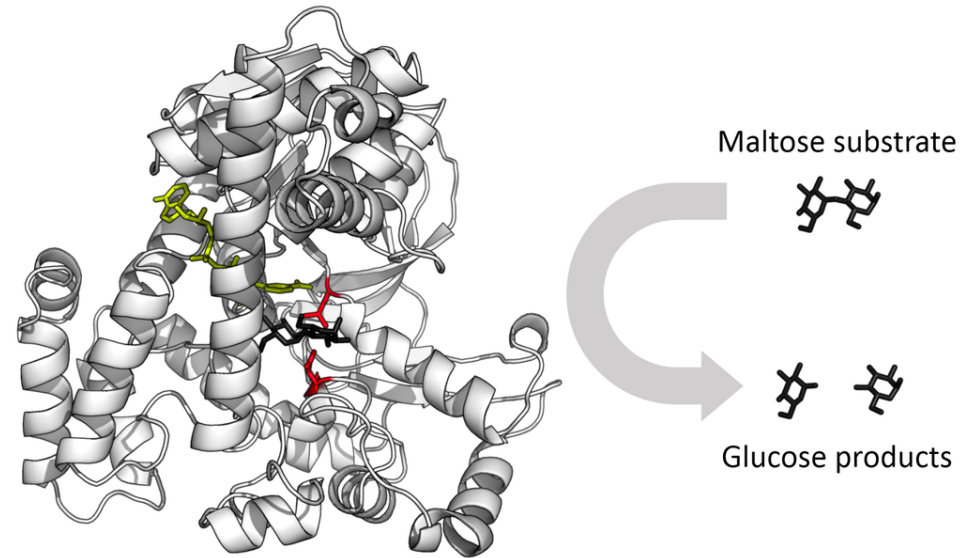
Genome size vs. protein count across NCBI genomes



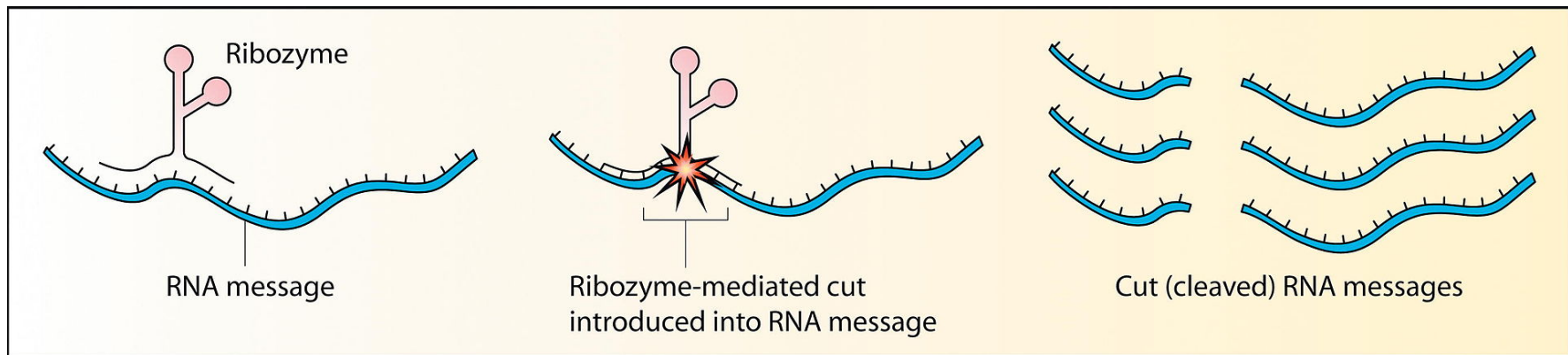
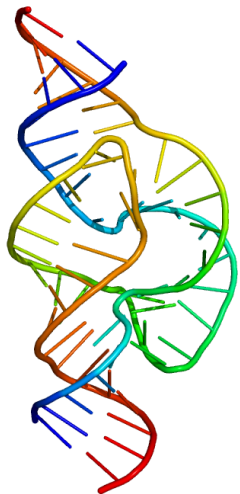
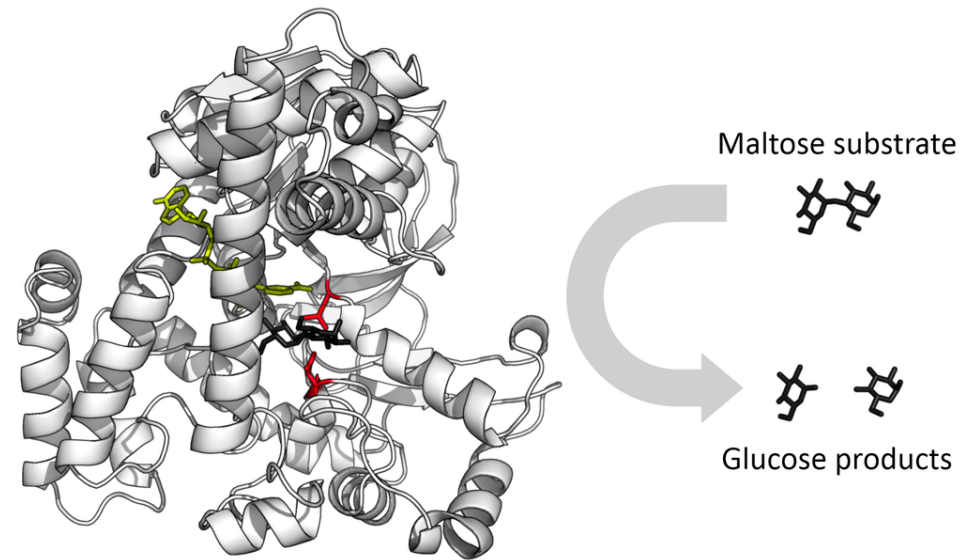
biological catalysts by accelerating chemical reactions

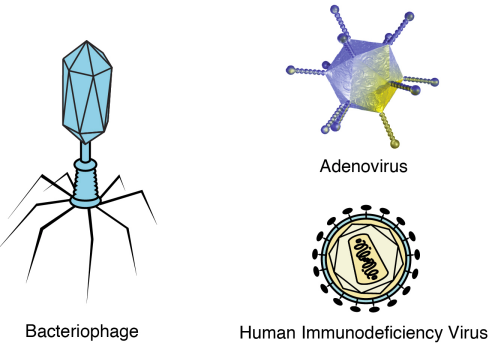


biological catalysts by accelerating chemical reactions



biological catalysts by accelerating chemical reactions

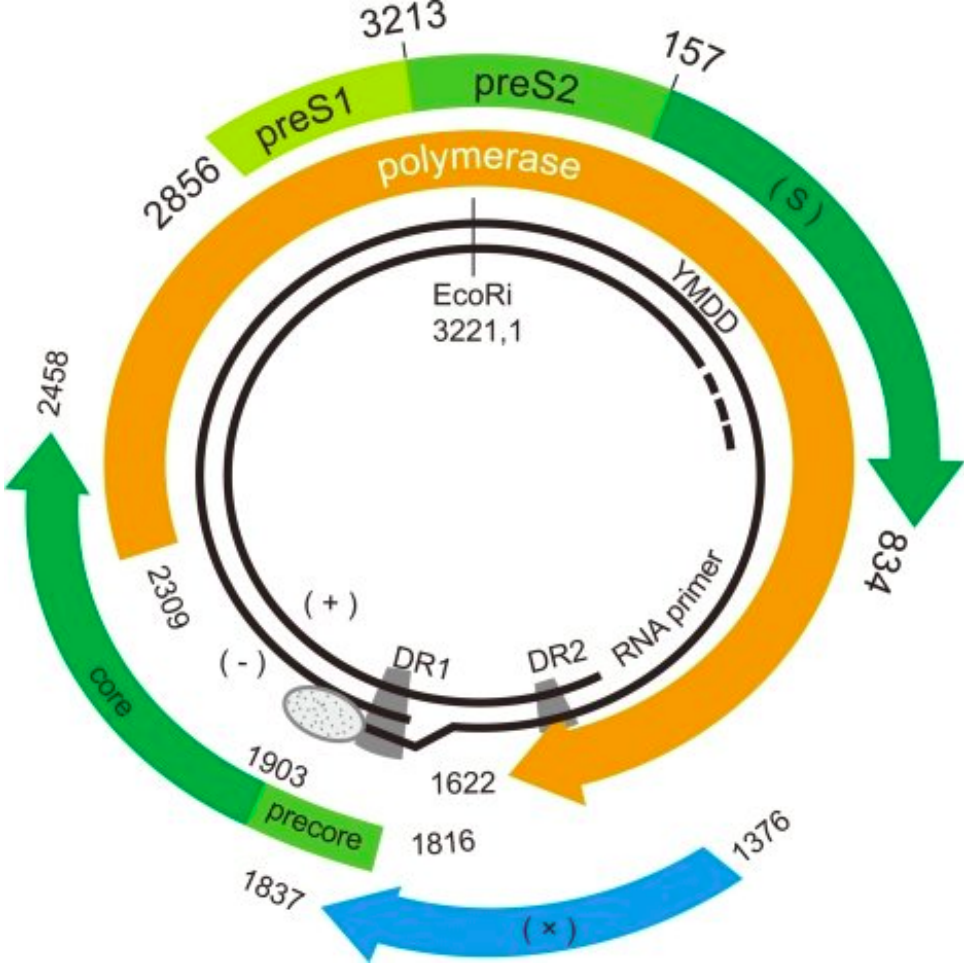




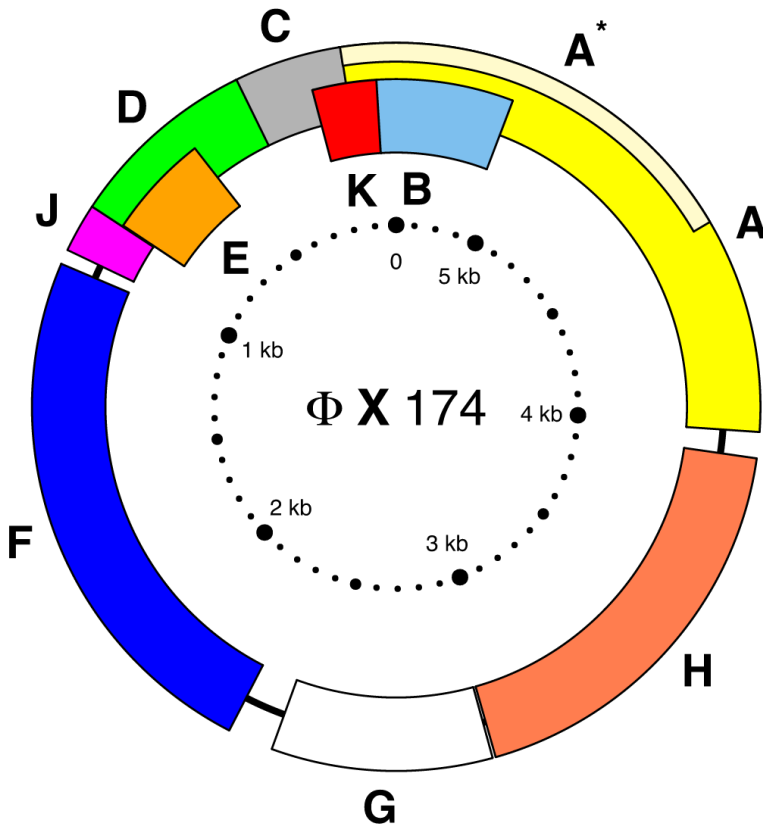
PROTEIN A: -- Gly - Lys - STOP

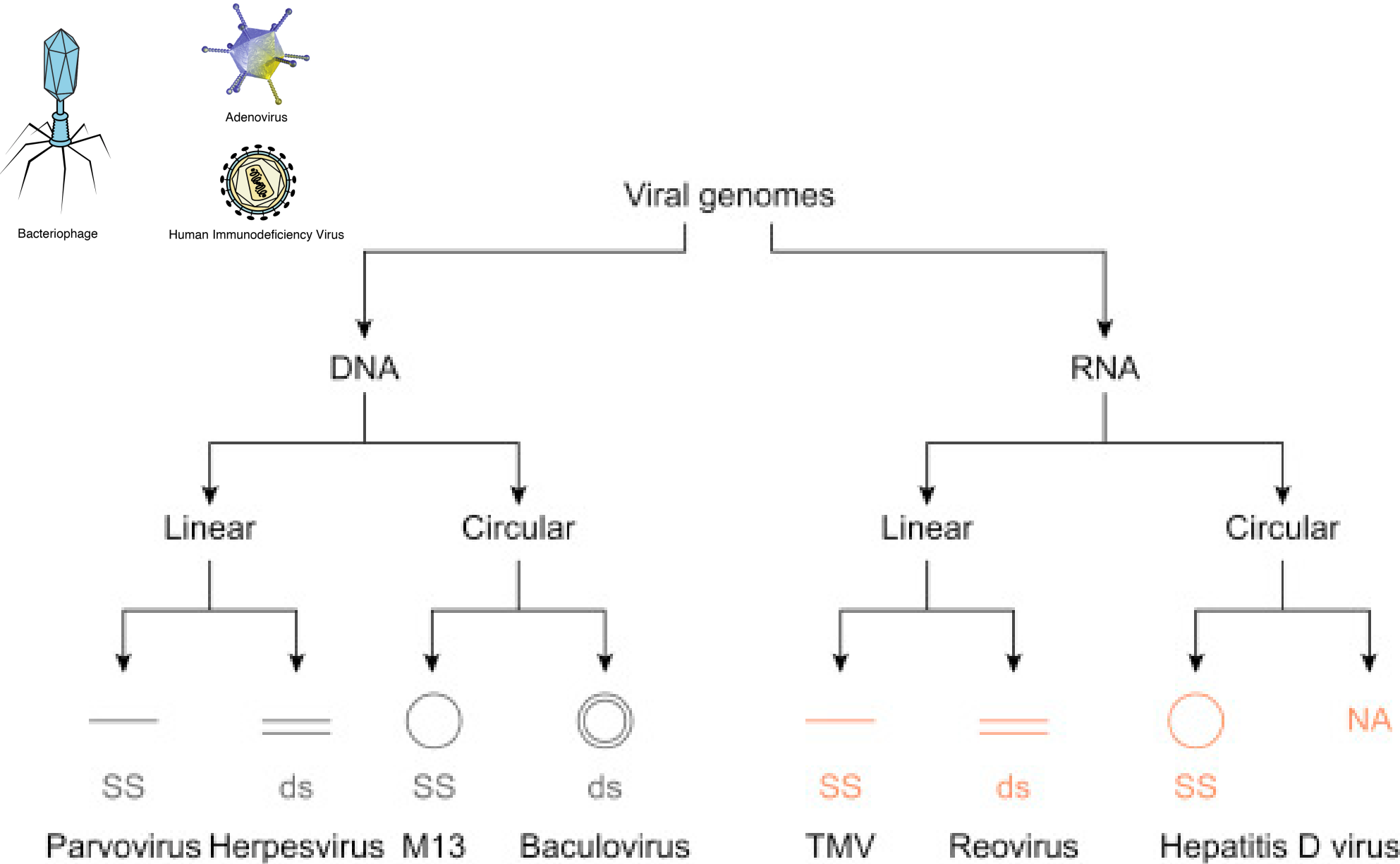
DNA: ... G G A A A A T G A G A A A A ...

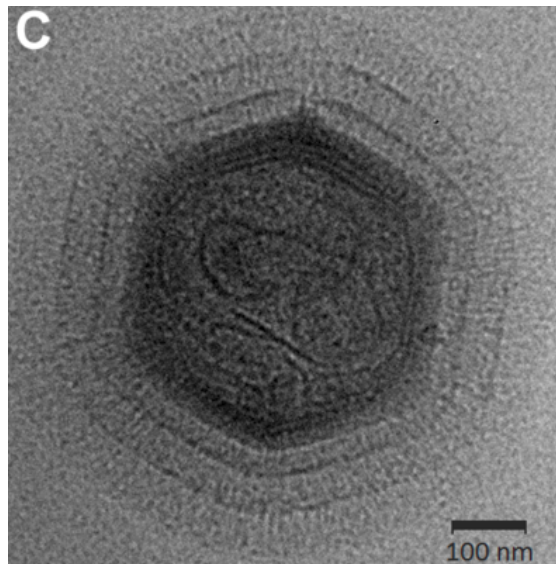
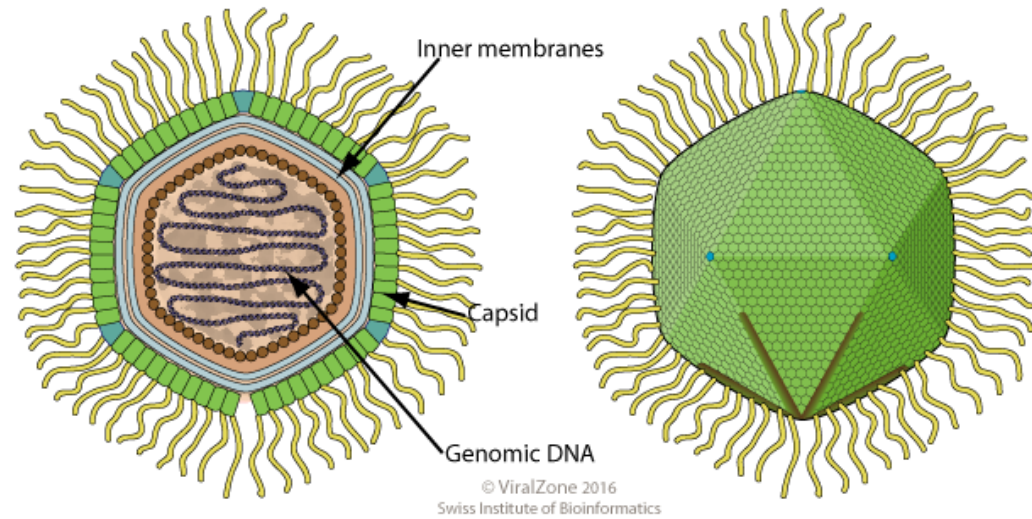
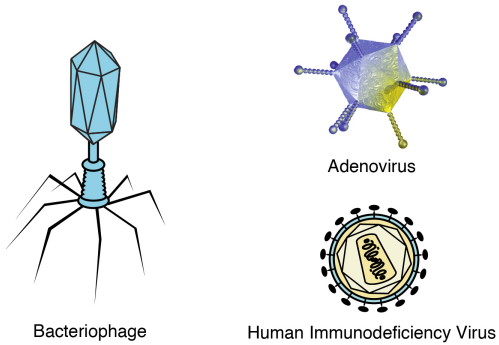
PROTEIN K: --- Glu - Asn - Glu - Lys ---
Met - Arg - Lys---



Hepatitis B virus



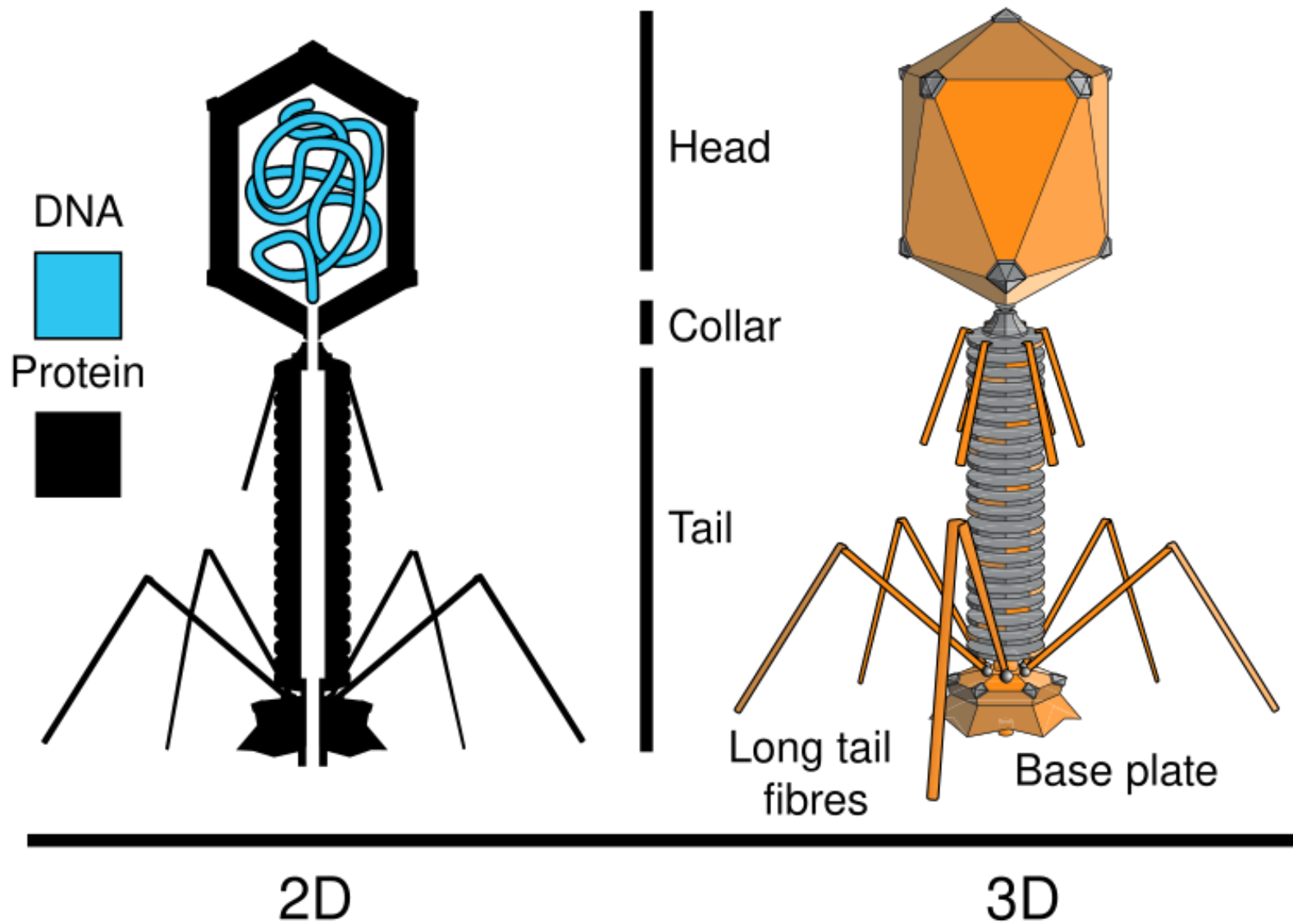


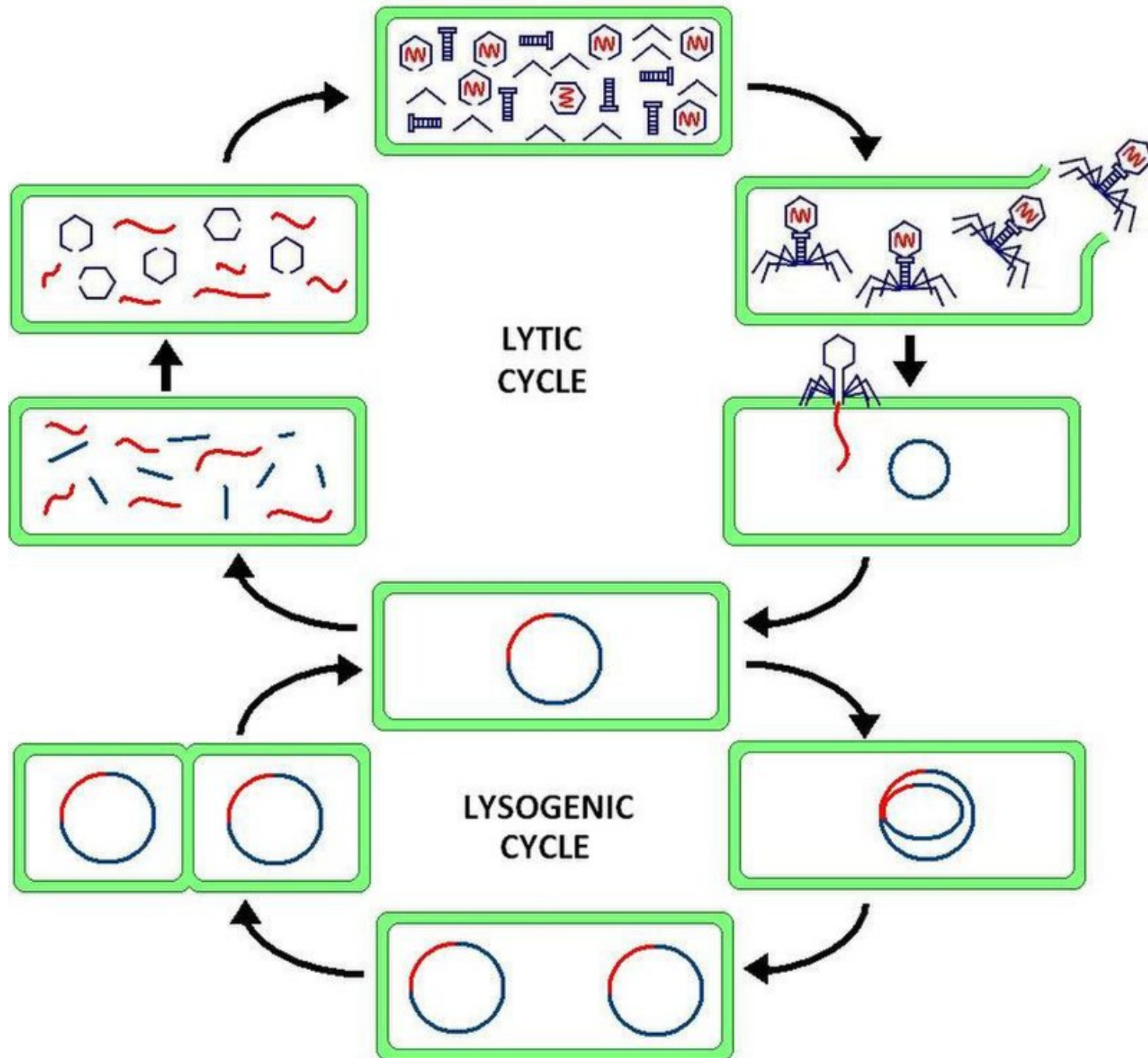


Mimivirus (host amoeba)

Genome: 1.2 MB

~1000 genes





Virophage

🌐 14 languages ▼

Article [Talk](#)

[Read](#) [Edit](#) [View history](#) [Tools](#) ▼

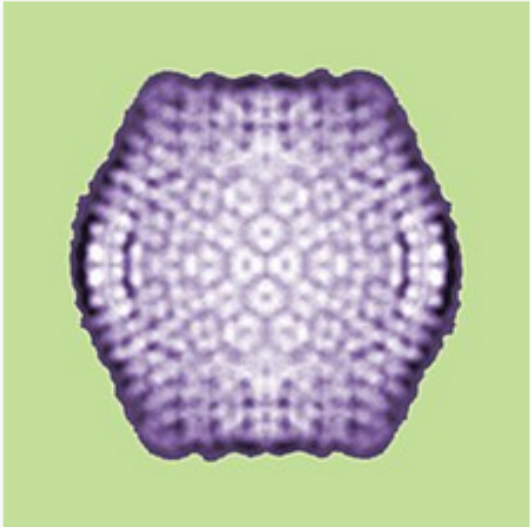
From Wikipedia, the free encyclopedia

Virophages are small, double-stranded DNA viral phages that require the [co-infection](#) of another virus. The co-infecting viruses are typically [giant viruses](#). Virophages rely on the viral replication factory of the co-infecting giant virus for their own replication. One of the characteristics of virophages is that they have a [parasitic](#) relationship with the co-infecting virus. Their dependence upon the giant virus for replication often results in the deactivation of the giant viruses. The virophage may improve the recovery and survival of the host organism. Virophages constitute the [class](#) *Virophaviricetes*.

Discovery [\[edit\]](#)

The first virophage was discovered in a [cooling tower](#) in Paris in 2008. It was discovered with its co-infecting giant virus, *Acanthamoeba castellanii* [mamavirus](#) (ACMV). The virophage was named [Sputnik](#) and its replication relied entirely on the co-infection of ACMV and its cytoplasmic replication machinery. Sputnik was also discovered to have an inhibitory effect on ACMV and improved the survival of the host. Other characterised virophages include Sputnik 2, Sputnik 3, [Zamilon](#) and [Mavirus](#).^{[2][3][4][5]}

Virophaviricetes

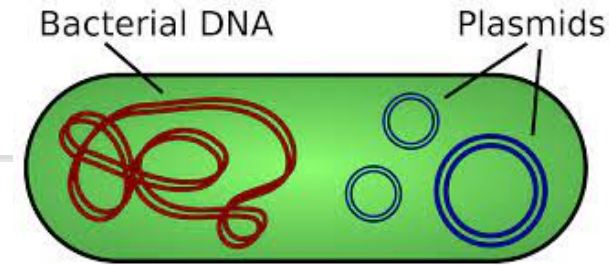


Sputnik virophage

Virus classification

(unranked):	Virus
Realm:	Varidnaviria
Kingdom:	Bamfordvirae
Phylum:	Bamfordviricota

Prokaryotic genomes

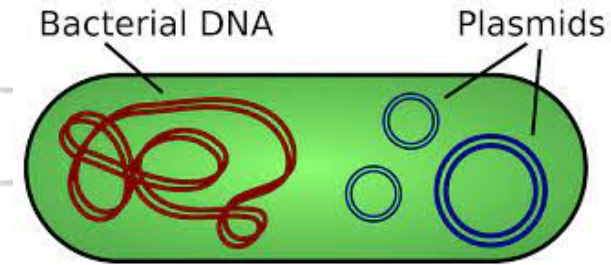
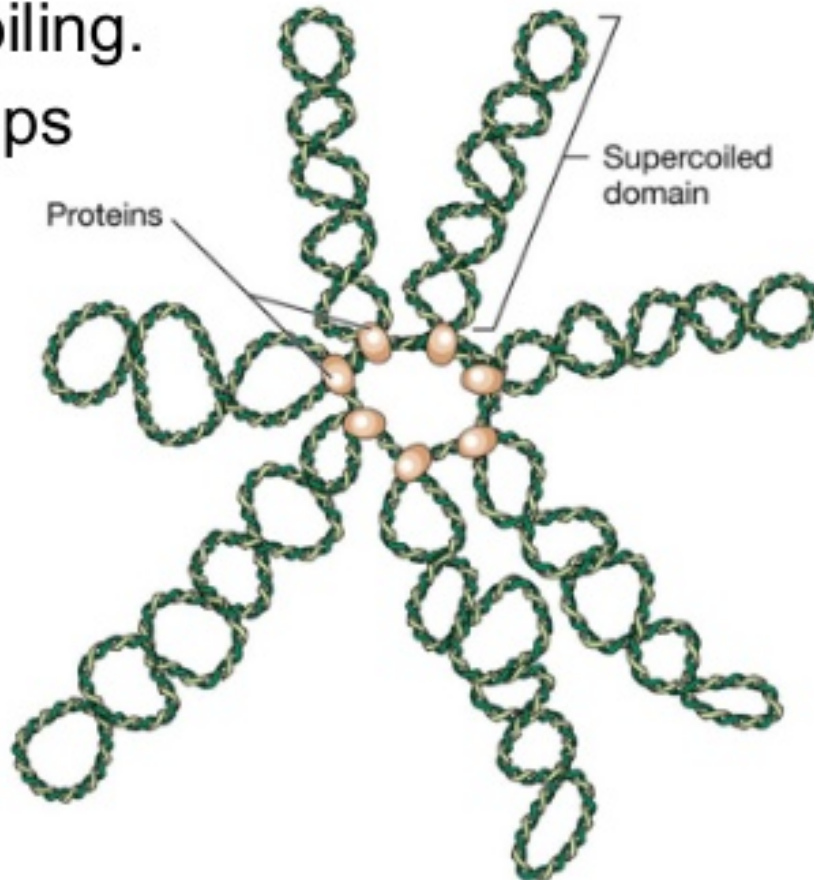


- Most genomes are:
 1. single molecule.
 2. one circular chromosome.
 3. double stranded DNA.
 4. some have small additional circular DNA that can replicate independently (**plasmid**).
- Genome is organized in a structure called (**Nucleoid**).

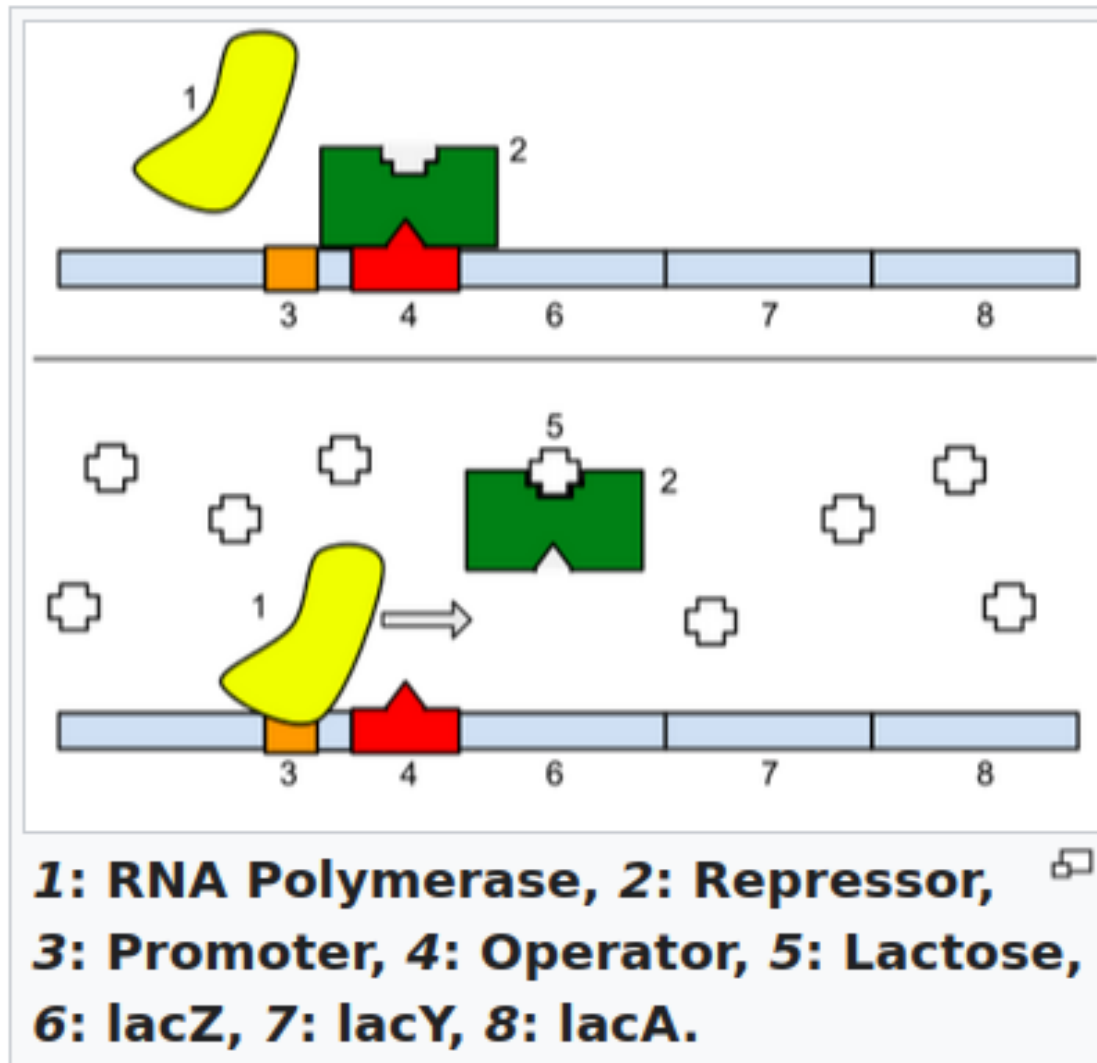
Prokaryotic genomes organization

Prokaryote genome organization through:

1. Nucleoid (protein DNA interaction).
2. Supercoiling.
3. DNA loops

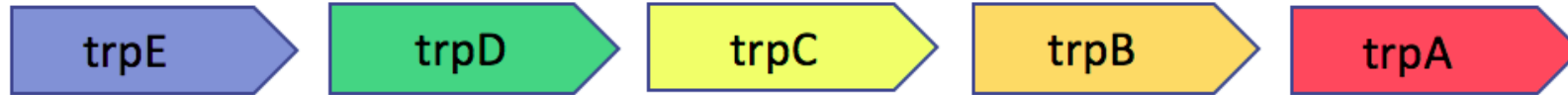


Operon is a functioning unit of DNA containing a cluster of genes under the control of a single promoter

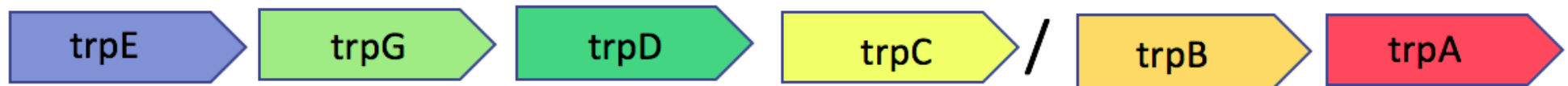


Operon is a functioning unit of DNA containing a cluster of genes under the control of a single promoter

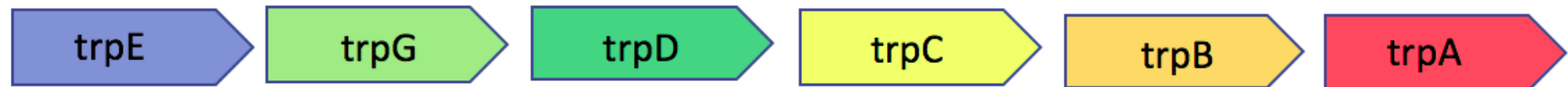
Escherichia coli



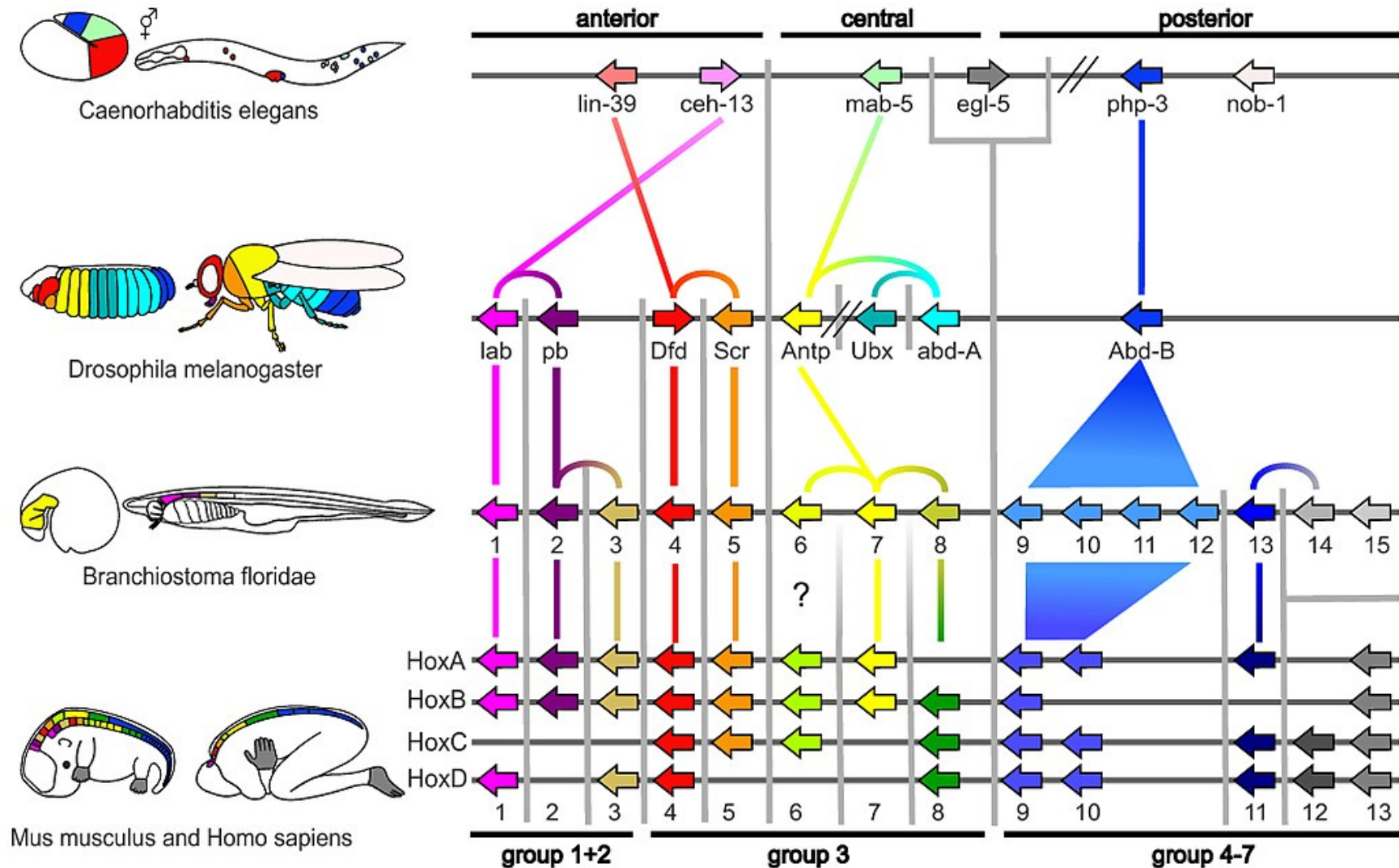
Haemophilus influenzae



Helicobacter pylori



Synteny conservation of blocks of order within two sets of chromosomes



Feature	Operon	Synteny
Definition	Cluster of co-transcribed, related genes	Conserved gene order across genomes
Functionality	Genes have related biological function	May or may not have related functions
Transcription	Transcribed as one mRNA (polycistronic)	Not necessarily co-transcribed
Scope	Within one genome	Across genomes (comparative context)
Example	lacZ-lacY-lacA in <i>E. coli</i>	dnaA-dnaN-recF conserved in multiple species
Used in	Gene regulation, pathway analysis	Evolutionary studies, genome structure comparison

Not all syntenic genes form operons, and

Not all operons are syntenic across species.

However, conserved operons across species can be a part of syntenic blocks. In fact, synteny can help detect candidate operons in poorly annotated genomes.

Synteny Detection Tools:

MCSanX – <http://chibba.pgml.uga.edu/mcscan2/>

Mauve – <http://darlinglab.org/mauve/mauve.html>

SynMap (CoGe) – <https://genomevolution.org/coge/SynMap.pl>

i-ADHoRe – <https://bioinformatics.psb.ugent.be/software/details/i-ADHoRe>

SyMAP – <https://sourceforge.net/projects/symap/>

GEvo (part of CoGe) – <https://genomevolution.org/CoGe/GEvo.pl>

LastZ (for pairwise synteny) – <https://github.com/lastz/lastz>

C-Sibelia – <https://github.com/bioinf/csibelia>

Operon Detection Tools:

DOOR2 – <http://csbl.bmb.uga.edu/DOOR/>

Operon-mapper – https://biocomputo.ibt.unam.mx/operon_mapper/

Rockhopper – <https://cs.wellesley.edu/~btjaden/Rockhopper/>

ProOpDB – <https://www.bioinformatics.org/proopdb/>

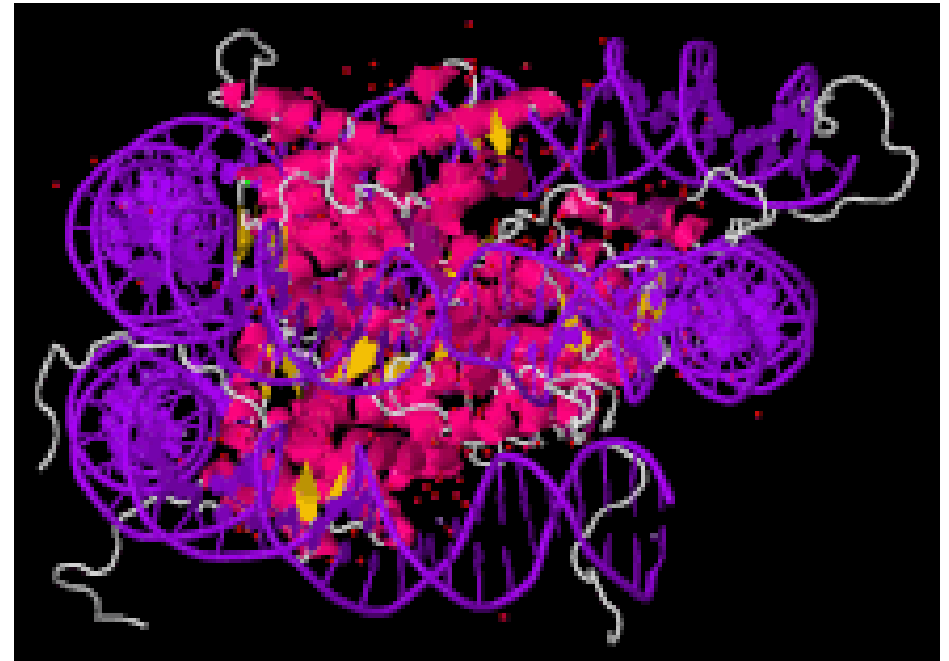
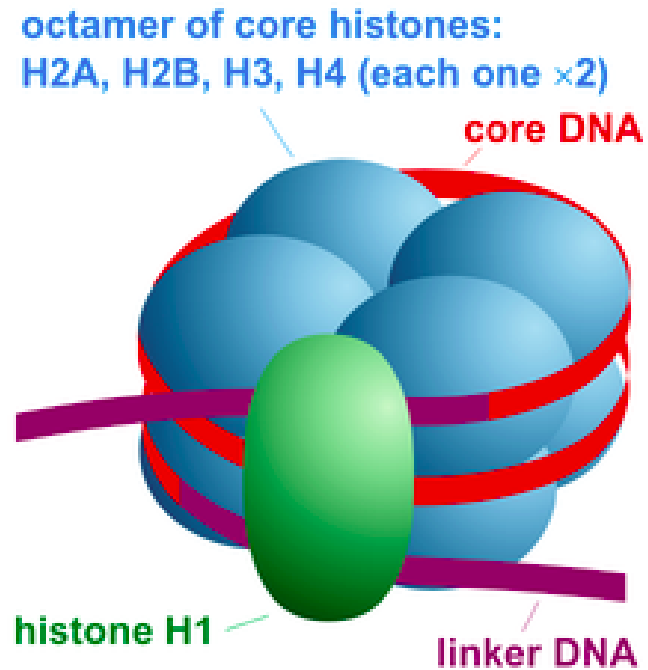
FgenesB – <http://www.softberry.com/berry.phtml>

OperPred – <https://ccb.nki.nl/software/operpred/>

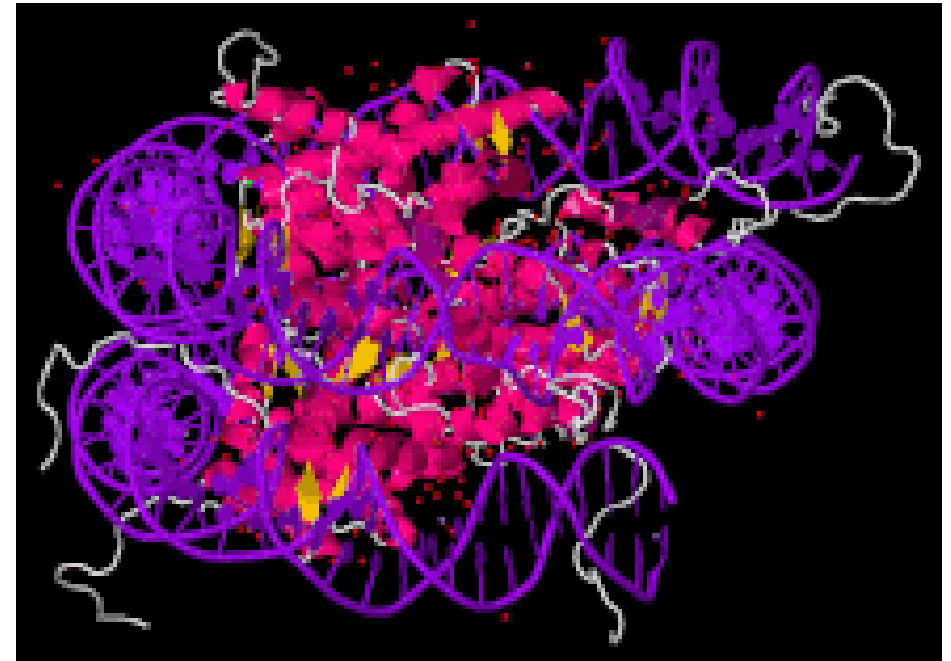
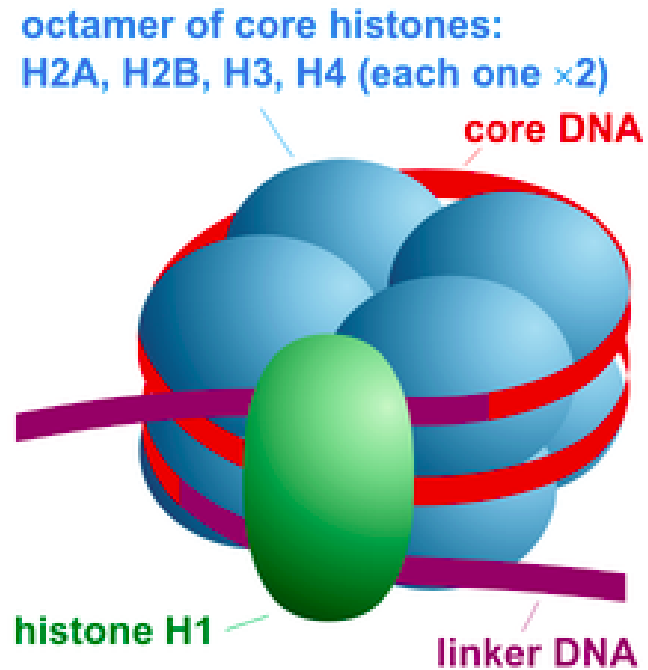
YOOOPER – <https://www.cmbi.uga.edu/software/yoooper.html>

nnpOperon – <https://github.com/bioinformatics-centre/nnpoperon>

DNA packing proteins



DNA packing proteins



Histone-organized chromatin in bacteria

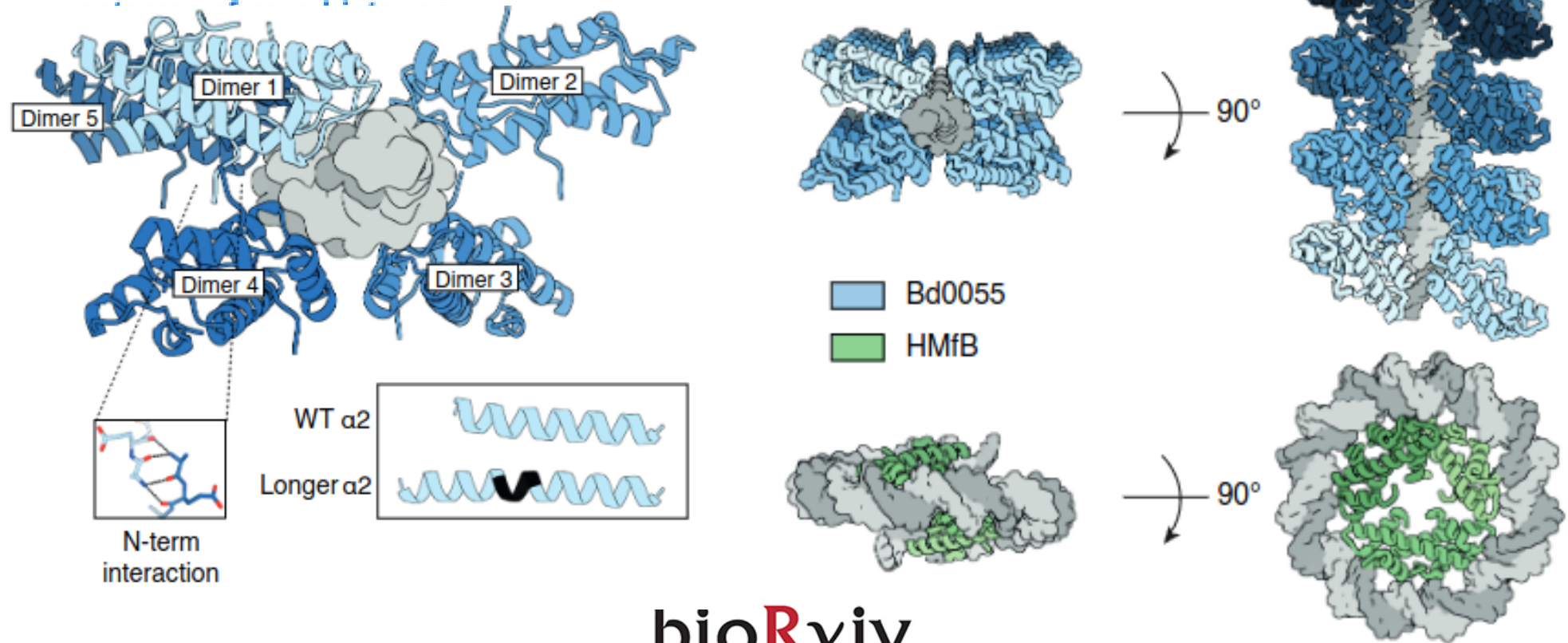
bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

Antoine Hoher, Shawn P. Laursen, Paul Radford, Jess Tyson, Carey Lambert, Kathryn M Stevens, Mathieu Picardeau, R. Elizabeth Sockett, Karolin Luger, Tobias Warnecke

doi: <https://doi.org/10.1101/2023.01.26.525422>

Posted January 26, 2023.

DNA packing proteins



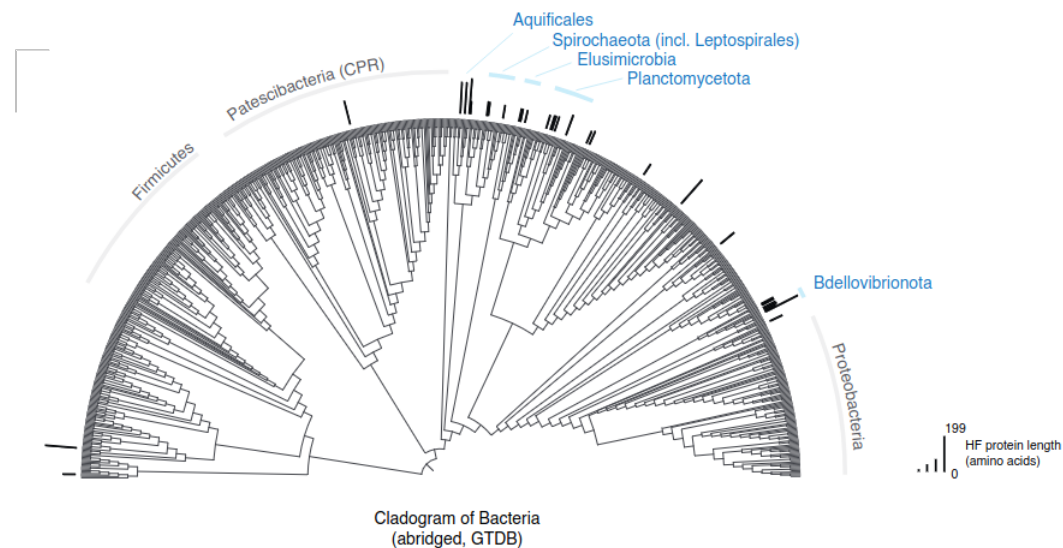
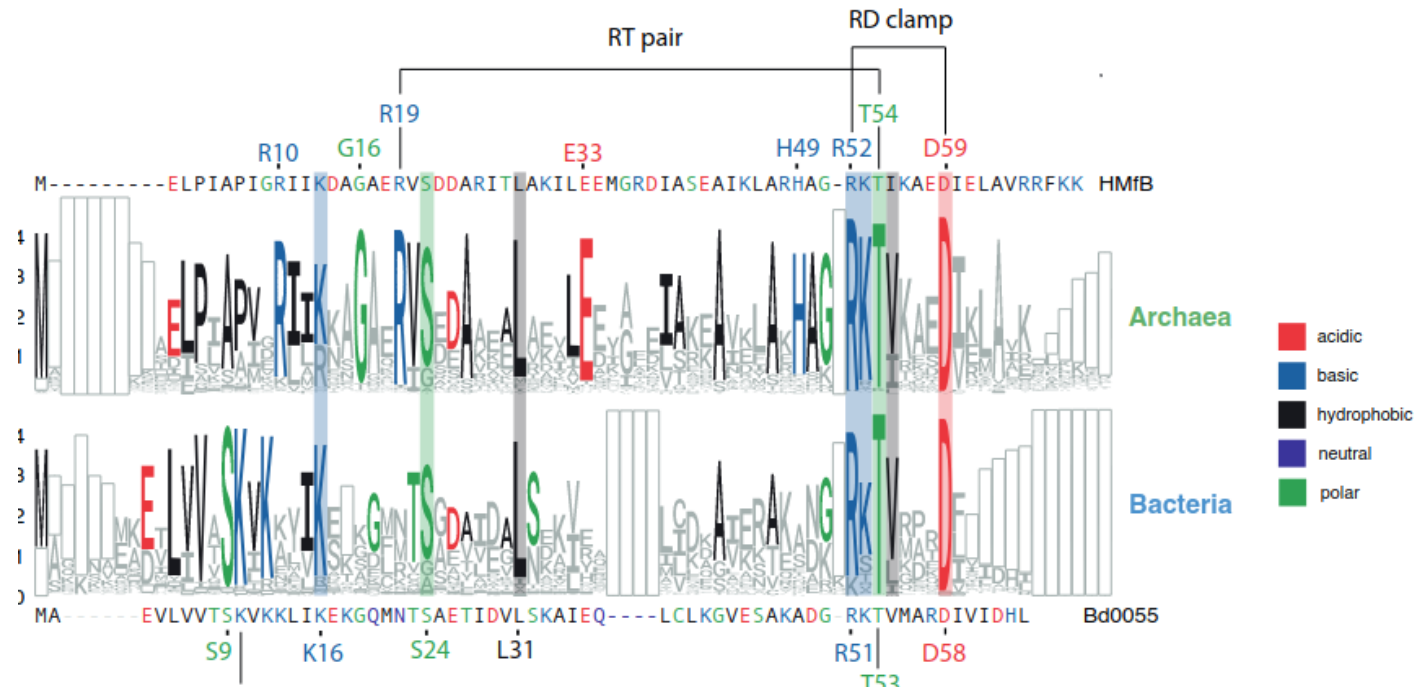
Histone-organized chromatin in bacteria

bioRxiv
THE PREPRINT SERVER FOR BIOLOGY

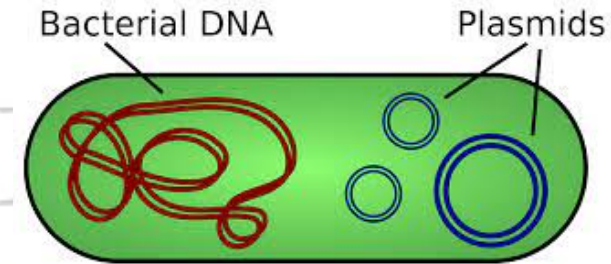
Antoine Hoher, Shawn P. Laursen, Paul Radford, Jess Tyson, Carey Lambert, Kathryn M Stevens, Mathieu Picardeau, R. Elizabeth Sockett, Karolin Luger, Tobias Warnecke
doi: <https://doi.org/10.1101/2023.01.26.525422>

Posted January 26, 2023.

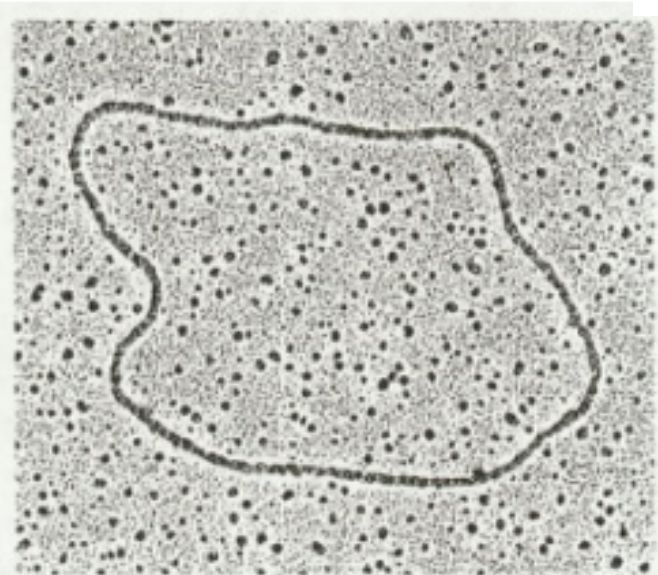
DNA packing proteins



Prokaryotic genomes organization

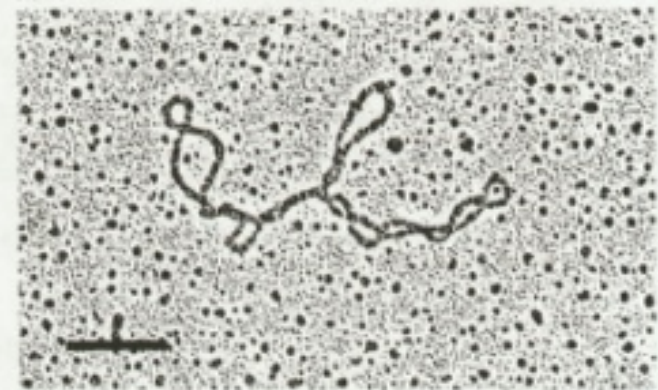


Relaxed circular
chromosome

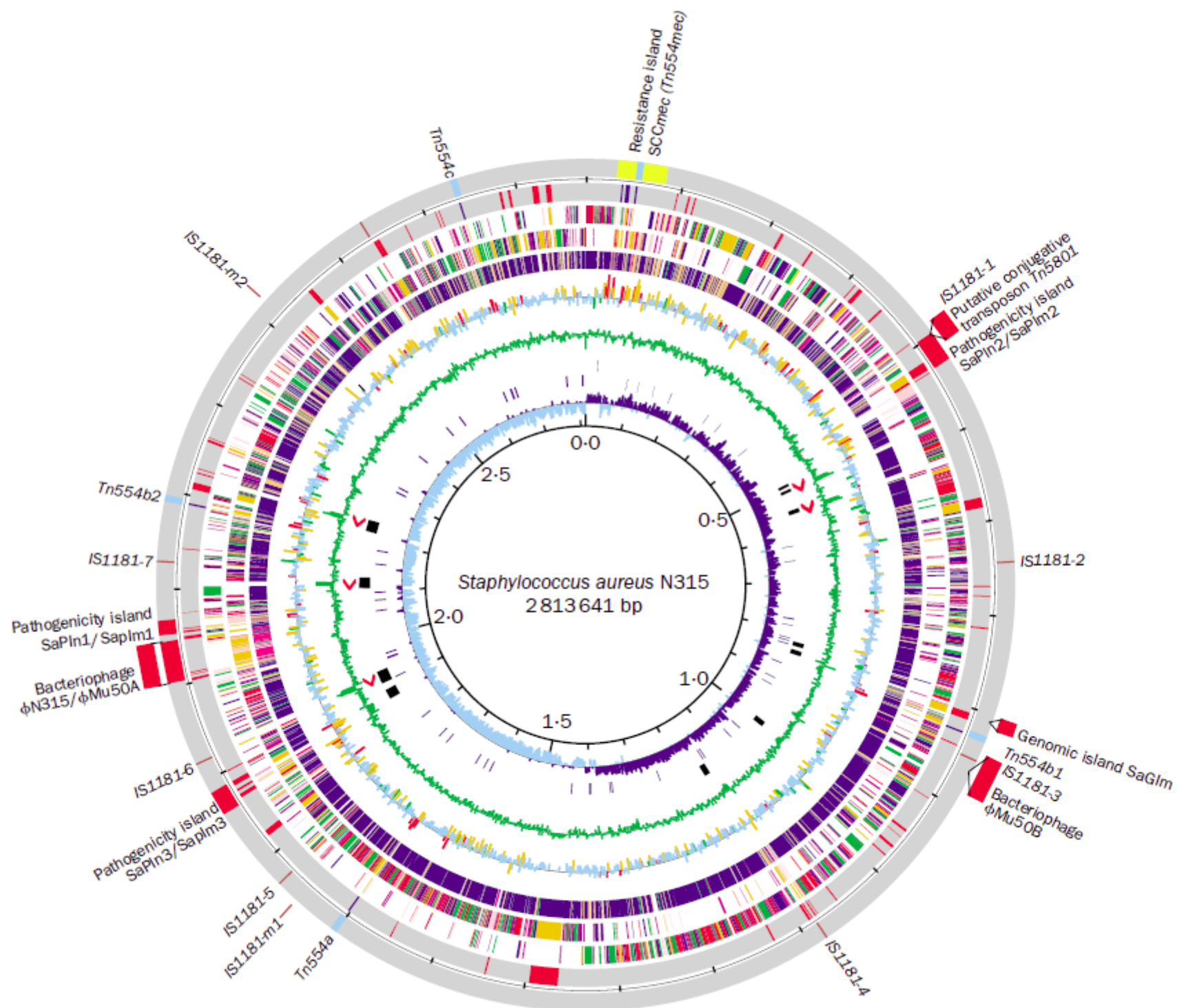


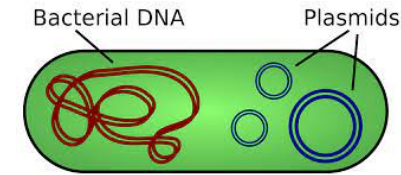
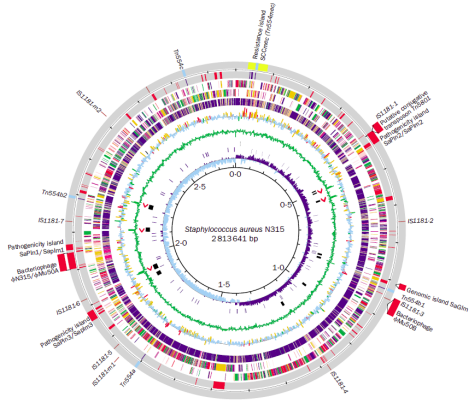
(a)

Supercoiled circular
chromosome



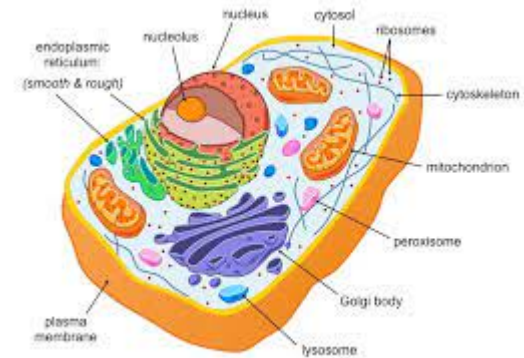
(b)





Organism type ⇅	Organism ⇅	Genome size (base pairs) ⇅		Approx. no. of genes ⇅	Note
Bacterium	<i>Nasuia deltocephalinicola</i> (strain NAS-ALF)	112,091	112 kB	137	Smallest known non-viral genome. Symbiont of leafhoppers. ^[59]
Bacterium	<i>Carsonella ruddii</i>	159,662	160 kB		An endosymbiont of psyllid insects
Bacterium	<i>Buchnera aphidicola</i>	600,000	600 kB		An endosymbiont of aphids ^[60]
Bacterium	<i>Wigglesworthia glossinidia</i>	700,000	700Kb		A symbiont in the gut of the tsetse fly
Bacterium – cyanobacterium	<i>Prochlorococcus</i> spp. (1.7 Mb)	1,700,000	1.7 MB	1,884	Smallest known cyanobacterium genome. One of the primary photosynthesizers on Earth. ^{[61][62]}
Bacterium	<i>Haemophilus influenzae</i>	1,830,000	1.8 MB		First genome of a living organism sequenced, July 1995 ^[63]
Bacterium	<i>Escherichia coli</i>	4,600,000	4.6 MB	4,288	^[64]
Bacterium – cyanobacterium	<i>Nostoc punctiforme</i>	9,000,000	9 MB	7,432	7432 open reading frames ^[65]
Bacterium	<i>Solibacter usitatus</i> (strain Ellin 6076)	9,970,000	10 MB		^[66]

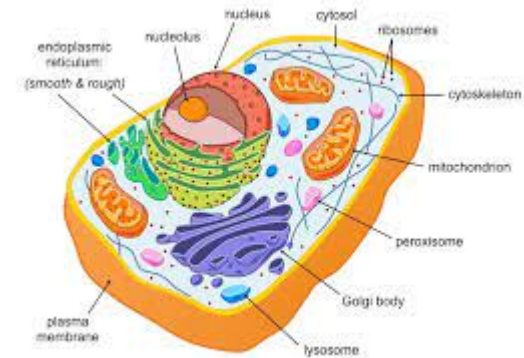
Eukaryotic genomes



- Genomes are:
 1. Genome is a double stranded DNA.
 2. Genome arranged in several linear packages called **chromosomes** through interactions with several proteins.

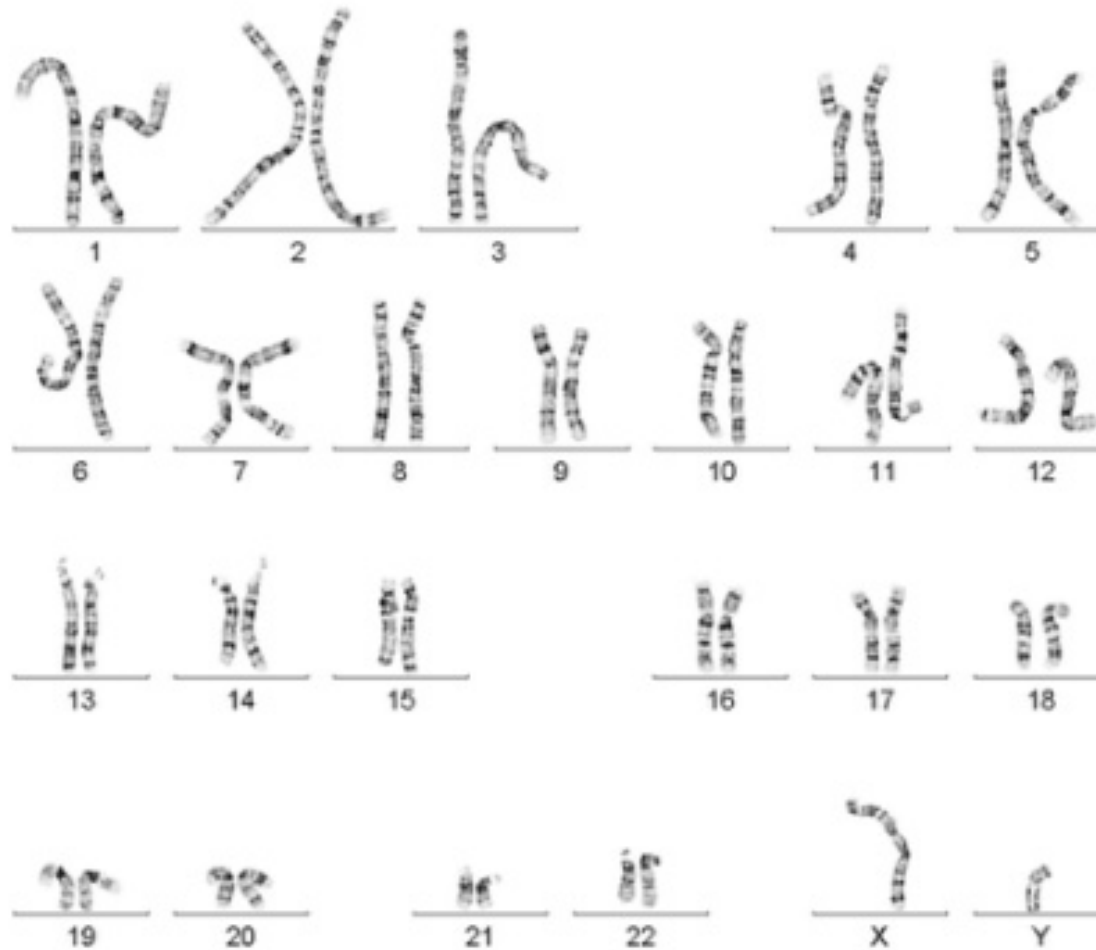
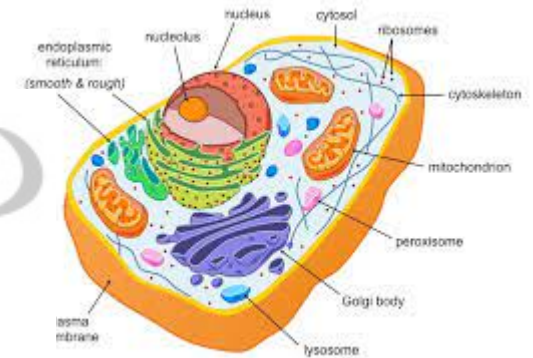
Eukaryotic genomes

- The chromosomes in their most condensed form are called metaphase chromosomes.
- The entire genome represented by metaphase chromosomes is called **karyotype**.
- The number of sets of chromosomes in a given eukaryotic cell is referred to as the level of ploidy.



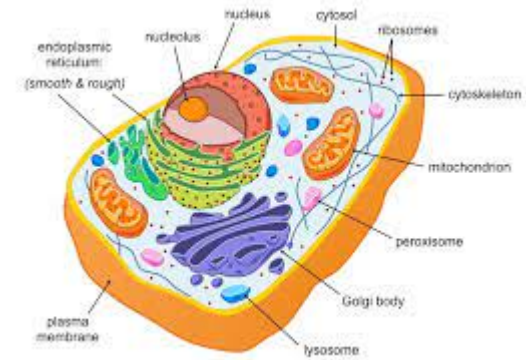
What is your level of ploidy?

The genome packaged into chromosomes



Karyotype

Ploidy

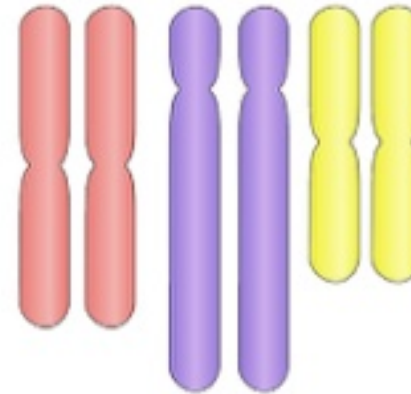


- **Haploid (1N):** cells that contain one set of chromosomes.
Which of human cells are haploid and what $1N = ?$

Haploid (N)

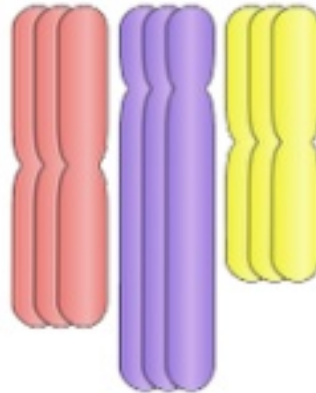


Diploid (2N)

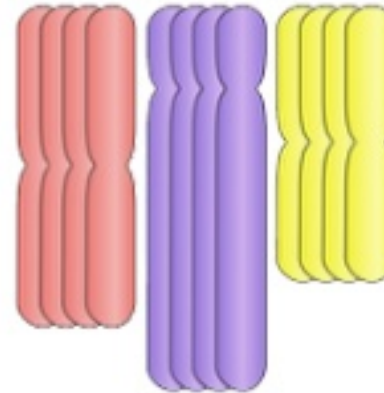


- **Diploid (2N):** cells that contain two sets of chromosomes.
Which of human cells are haploid and what $2N = ?$

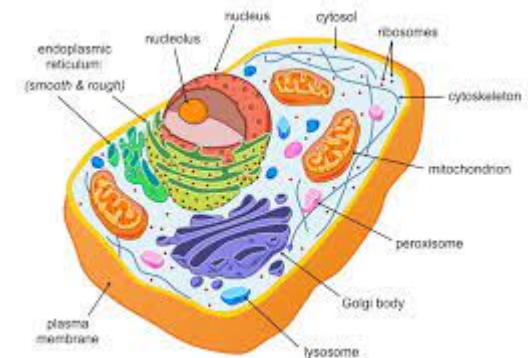
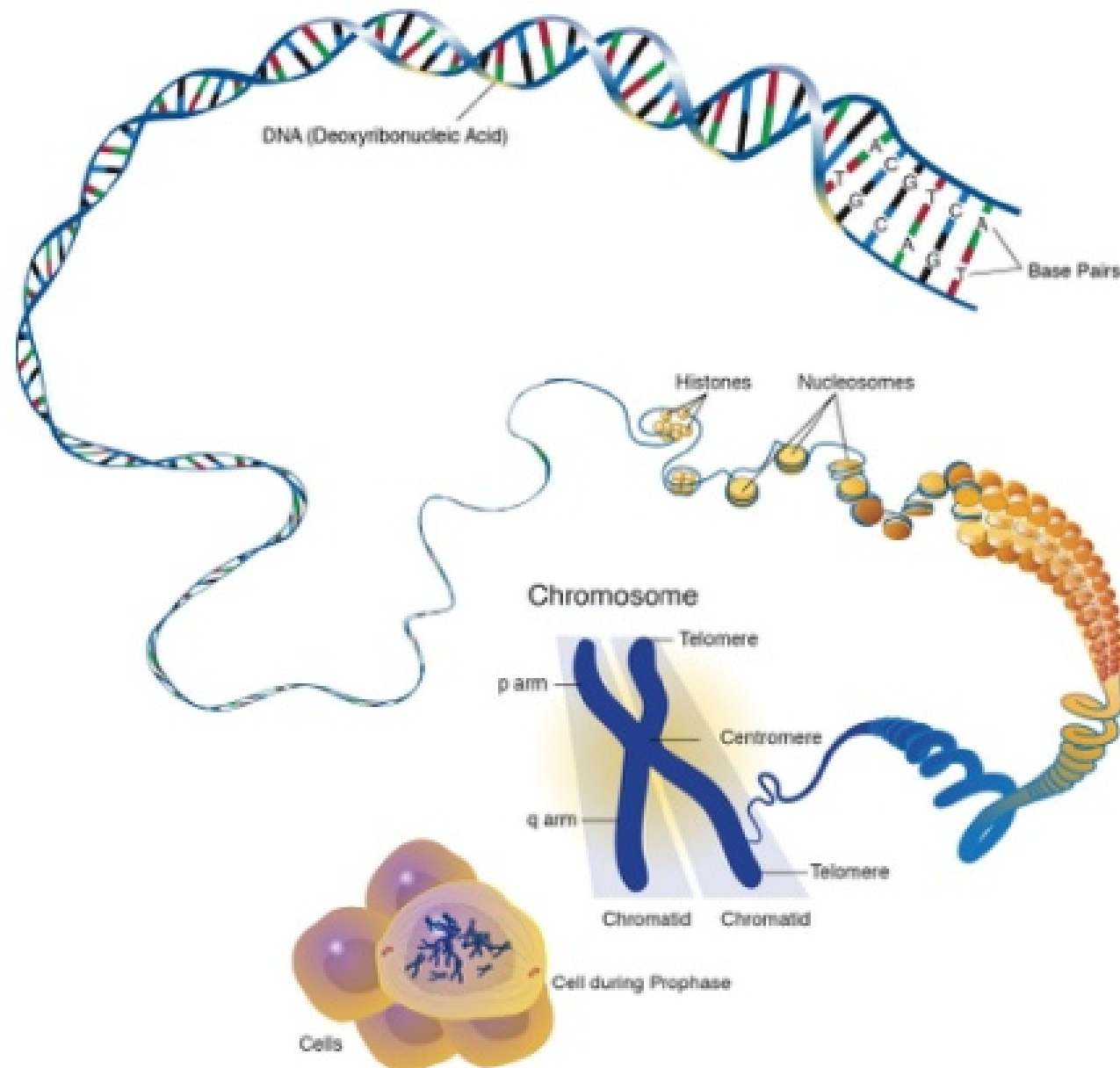
Triploid (3N)



Tetraploid (4N)



Eukaryotic genome packaging



Genome browsers

Organism type ⇅	Organism ⇅	Genome size (base pairs) ⇅		Approx. no. of genes ⇅	Note
Amoeboid	<i>Polychaos dubium</i> (<i>"Amoeba" dubia</i>)	670,000,000,000	670 GB		Largest known genome. ^[67] (Disputed) ^[68]
Plant	<i>Genlisea tuberosa</i>	61,000,000	61 MB		Smallest recorded flowering plant genome, 2014. ^[69]
Plant	<i>Arabidopsis thaliana</i>	135,000,000 ^[70]	135 MB	27,655 ^[71]	First plant genome sequenced, December 2000. ^[72]
Plant	<i>Populus trichocarpa</i>	480,000,000	480 MB	73,013	First tree genome sequenced, September 2006 ^[73]
Plant	<i>Fritillaria assyriaca</i>	130,000,000,000	130 GB		
Plant	<i>Paris japonica</i> (Japanese-native, pale-petal)	150,000,000,000	150 GB		Largest plant genome known ^[74]
Plant – moss	<i>Physcomitrella patens</i>	480,000,000	480 MB		First genome of a bryophyte sequenced, January 2008. ^[75]
Fungus – yeast	<i>Saccharomyces cerevisiae</i>	12,100,000	12.1 MB	6,294	First eukaryotic genome sequenced, 1996 ^[76]
Fungus	<i>Aspergillus nidulans</i>	30,000,000	30 MB	9,541	^[77]
Nematode	<i>Pratylenchus coffeae</i>	20,000,000	20 MB		^[78] Smallest animal genome known ^[79]
Nematode	<i>Caenorhabditis elegans</i>	100,300,000	100 MB	19,000	First multicellular animal genome sequenced, December 1998 ^[80]
Insect	<i>Drosophila melanogaster</i> (fruit fly)	175,000,000	175 MB	13,600	Size variation based on strain (175-180Mb; standard y w strain is 175Mb) ^[81]
Insect	<i>Apis mellifera</i> (honey bee)	236,000,000	236 MB	10,157	^[82]
Insect	<i>Bombyx mori</i> (silk moth)	432,000,000	432 MB	14,623	14,623 predicted genes ^[83]
Insect	<i>Solenopsis invicta</i> (fire ant)	480,000,000	480 MB	16,569	^[84]
Mammal	<i>Mus musculus</i>	2,700,000,000	2.7 GB	20,210	^[85]
Mammal	<i>Pan paniscus</i>	3,286,640,000	3.3 GB	20,000	Bonobo - estimated genome size 3.29 billion bp ^[86]

Genome browsers

Organism type ⇅	Organism ⇅	Genome size (base pairs) ⇅		Approx. no. of genes ⇅	Note
Amoeboid	<i>Polychaos dubium</i> ("Amoeba" dubia)	670,000,000,000	670 GB		Largest known genome. ^[67] (Disputed) ^[68]
Plant	<i>Genlisea tuberosa</i>	61,000,000	61 MB		Smallest recorded flowering plant genome, 2014. ^[69]
Plant	<i>Arabidopsis thaliana</i>	135,000,000 ^[70]	135 MB	27,655 ^[71]	First plant genome sequenced, December 2000. ^[72]
Plant	<i>Populus trichocarpa</i>	480,000,000	480 MB	73,613	First tree genome sequenced, September 2006. ^[73]
Plant	<i>Fritillaria assyriaca</i>	130,000,000,000	130 GB		
Plant	<i>Paris japonica</i> (Japanese-native, pale-petal)	150,000,000,000	150 GB		Largest plant genome known ^[74]
Plant – moss	<i>Physcomitrella patens</i>	480,000,000	480 MB		First genome of a bryophyte sequenced, January 2008. ^[75]
Fungus – yeast	<i>Saccharomyces cerevisiae</i>	12,100,000	12.1 MB	6,294	First eukaryotic genome sequenced, 1996 ^[76]
Fungus	<i>Aspergillus nidulans</i>	30,000,000	30 MB	9,541	^[77]
Nematode	<i>Pratylenchus coffeae</i>	20,000,000	20 MB		^[78] Smallest animal genome known ^[79]
Nematode	<i>Caenorhabditis elegans</i>	100,300,000	100 MB	19,000	First multicellular animal genome sequenced, December 1998 ^[80]
Insect	<i>Drosophila melanogaster</i> (fruit fly)	175,000,000	175 MB	13,600	Size variation based on strain (175-180Mb; standard y w strain is 175Mb) ^[81]
Insect	<i>Apis mellifera</i> (honey bee)	236,000,000	236 MB	10,157	^[82]
Insect	<i>Bombyx mori</i> (silk moth)	432,000,000	432 MB	14,623	14,623 predicted genes ^[83]
Insect	<i>Solenopsis invicta</i> (fire ant)	480,000,000	480 MB	16,569	^[84]
Mammal	<i>Mus musculus</i>	2,700,000,000	2.7 GB	20,210	^[85]
Mammal	<i>Pan paniscus</i>	3,286,640,000	3.3 GB	20,000	Bonobo - estimated genome size 3.29 billion bp ^[86]

Genome browsers

Organism type ↕	Organism ↕	Genome size (base pairs) ↕		Approx. no. of genes ↕	Note
Amoeboid	<i>Polychaos dubium</i> ("Amoeba" dubia)	670,000,000,000	670 GB		Largest known genome. ^[67] (Disputed) ^[68]
Plant	<i>Genlisea tuberosa</i>	61,000,000	61 MB		Smallest recorded flowering plant genome, 2014. ^[69]
Plant	<i>Arabidopsis thaliana</i>	135,000,000 ^[70]	135 MB	27,655 ^[71]	First plant genome sequenced, December 2000. ^[72]
Plant	<i>Populus trichocarpa</i>	480,000,000	480 MB	73,013	First tree genome sequenced, September 2006. ^[73]
Plant	<i>Fritillaria assyriaca</i>	130,000,000,000	130 GB		
Plant	<i>Paris japonica</i> (Japanese-native, pale-petal)	150,000,000,000	150 GB		Largest plant genome known ^[74]
Plant – moss	<i>Physcomitrella patens</i>	480,000,000	480 MB		First genome of a bryophyte sequenced, January 2008. ^[75]
Fungus – yeast	<i>Saccharomyces cerevisiae</i>	12,100,000	12.1 MB	6,294	First eukaryotic genome sequenced, 1996 ^[76]
Fungus	<i>Aspergillus nidulans</i>	30,000,000	30 MB	9,541	^[77]
Nematode	<i>Pratylenchus coffeae</i>	20,000,000	20 MB		^[78] Smallest animal genome known ^[79]
Nematode	<i>Caenorhabditis elegans</i>	100,300,000	100 MB	19,000	First multicellular animal genome sequenced, December 1998 ^[80]
Insect	<i>Drosophila melanogaster</i> (fruit fly)	175,000,000	175 MB	13,600	Size variation based on strain (175-180Mb; standard y w strain is 175Mb) ^[81]
Insect	<i>Apis mellifera</i> (honey bee)	236,000,000	236 MB	10,157	^[82]
Insect	<i>Bombyx mori</i> (silk moth)	432,000,000	432 MB	14,623	14,623 predicted genes ^[83]
Insect	<i>Solenopsis invicta</i> (fire ant)	480,000,000	480 MB	16,569	^[84]
Mammal	<i>Mus musculus</i>	2,700,000,000	2.7 GB	20,210	^[85]
Mammal	<i>Pan paniscus</i>	3,286,640,000	3.3 GB	20,000	Bonobo - estimated genome size 3.29 billion bp ^[86]

Genome browsers

Organism type ↕	Organism ↕	Genome size (base pairs) ↕		Approx. no. of genes ↕	Note
Amoeboid	<i>Polychaos dubium</i> ("Amoeba" dubia)	670,000,000,000	670 GB		Largest known genome. ^[67] (Disputed) ^[68]
Plant	<i>Genlisea tuberosa</i>	61,000,000	61 MB		Smallest recorded flowering plant genome, 2014. ^[69]
Plant	<i>Arabidopsis thaliana</i>	135,000,000 ^[70]	135 MB	27,655 ^[71]	First plant genome sequenced, December 2000. ^[72]
Plant	<i>Populus trichocarpa</i>	480,000,000	480 MB	73,013	First tree genome sequenced, September 2006. ^[73]
Plant	<i>Fritillaria assyriaca</i>	130,000,000,000	130 GB		
Plant	<i>Paris japonica</i> (Japanese-native, pale-petal)	150,000,000,000	150 GB		Largest plant genome known ^[74]
Plant – moss	<i>Physcomitrella patens</i>	480,000,000	480 MB		First genome of a bryophyte sequenced, January 2008. ^[75]
Fungus – yeast	<i>Saccharomyces cerevisiae</i>	12,100,000	12.1 MB	6,294	First eukaryotic genome sequenced, 1996 ^[76]
Fungus	<i>Aspergillus nidulans</i>	30,000,000	30 MB	9,541	^[77]
Nematode	<i>Pratylenchus coffeae</i>	20,000,000	20 MB		^[78] Smallest animal genome known ^[79]
Nematode	<i>Caenorhabditis elegans</i>	100,300,000	100 MB	19,000	First multicellular animal genome sequenced, December 1998 ^[80]
Insect	<i>Drosophila melanogaster</i> (fruit fly)	175,000,000	175 MB	13,600	Size variation based on strain (175-180Mb; standard y w strain is 175Mb) ^[81]
Insect	<i>Apis mellifera</i> (honey bee)	236,000,000	236 MB	10,157	^[82]
Insect	<i>Bombyx mori</i> (silk moth)	432,000,000	432 MB	14,623	14,623 predicted genes ^[83]
Insect	<i>Solenopsis invicta</i> (fire ant)	480,000,000	480 MB	16,569	^[84]
Mammal	<i>Mus musculus</i>	2,700,000,000	2.7 GB	20,210	^[85]
Mammal	<i>Pan paniscus</i>	3,286,640,000	3.3 GB	20,000	Bonobo - estimated genome size 3.29 billion bp ^[86]

Genome browsers

Organism type ↕	Organism ↕	Genome size (base pairs) ↕		Approx. no. of genes ↕	Note
Amoeboid	<i>Polychaos dubium</i> ("Amoeba" dubia)	670,000,000,000	670 GB		Largest known genome. ^[67] (Disputed) ^[68]
Plant	<i>Genlisea tuberosa</i>	61,000,000	61 MB		Smallest recorded flowering plant genome, 2014. ^[69]
Plant	<i>Arabidopsis thaliana</i>	135,000,000 ^[70]	135 MB	27,655 ^[71]	First plant genome sequenced, December 2000. ^[72]
Plant	<i>Populus trichocarpa</i>	480,000,000	480 MB	73,013	First tree genome sequenced, September 2006. ^[73]
Plant	<i>Fritillaria assyriaca</i>	130,000,000,000	130 GB		
Plant	<i>Paris japonica</i> (Japanese-native, pale-petal)	150,000,000,000	150 GB		Largest plant genome known ^[74]
Plant – moss	<i>Physcomitrella patens</i>	480,000,000	480 MB		First genome of a bryophyte sequenced, January 2008. ^[75]
Fungus – yeast	<i>Saccharomyces cerevisiae</i>	12,100,000	12.1 MB	6,294	First eukaryotic genome sequenced, 1996 ^[76]
Fungus	<i>Aspergillus nidulans</i>	30,000,000	30 MB	9,541	^[77]
Nematode	<i>Pratylenchus coffeae</i>	20,000,000	20 MB		^[78] Smallest animal genome known ^[79]
Nematode	<i>Caenorhabditis elegans</i>	100,300,000	100 MB	19,000	First multicellular animal genome sequenced, December 1998 ^[80]
Insect	<i>Drosophila melanogaster</i> (fruit fly)	175,000,000	175 MB	13,600	Size variation based on strain (175-180Mb; standard y w strain is 175Mb) ^[81]
Insect	<i>Apis mellifera</i> (honey bee)	236,000,000	236 MB	10,157	^[82]
Insect	<i>Bombyx mori</i> (silk moth)	432,000,000	432 MB	14,623	14,623 predicted genes ^[83]
Insect	<i>Solenopsis invicta</i> (fire ant)	480,000,000	480 MB	16,569	^[84]
Mammal	<i>Mus musculus</i>	2,700,000,000	2.7 GB	20,210	^[85]
Mammal	<i>Pan paniscus</i>	3,286,640,000	3.3 GB	20,000	Bonobo - estimated genome size 3.29 billion bp ^[86]

Genome browsers

Organism type ↕	Organism ↕	Genome size (base pairs) ↕		Approx. no. of genes ↕	Note
Amoeboid	<i>Polychaos dubium</i> ("Amoeba" dubia)	670,000,000,000	670 GB		Largest known genome. ^[67] (Disputed) ^[68]
Plant	<i>Genlisea tuberosa</i>	61,000,000	61 MB		Smallest recorded flowering plant genome, 2014. ^[69]
Plant	<i>Arabidopsis thaliana</i>	135,000,000 ^[70]	135 MB	27,655 ^[71]	First plant genome sequenced, December 2000. ^[72]
Plant	<i>Populus trichocarpa</i>	480,000,000	480 MB	73,013	First tree genome sequenced, September 2006. ^[73]
Plant	<i>Fritillaria assyriaca</i>	130,000,000,000	130 GB		
Plant	<i>Paris japonica</i> (Japanese-native, pale-petal)	150,000,000,000	150 GB		Largest plant genome known ^[74]
Plant – moss	<i>Physcomitrella patens</i>	480,000,000	480 MB		First genome of a bryophyte sequenced, January 2008. ^[75]
Fungus – yeast	<i>Saccharomyces cerevisiae</i>	12,100,000	12.1 MB	6,294	First eukaryotic genome sequenced, 1996 ^[76]
Fungus	<i>Aspergillus nidulans</i>	30,000,000	30 MB	9,541	^[77]
Nematode	<i>Pratylenchus coffeae</i>	20,000,000	20 MB		^[78] Smallest animal genome known ^[79]
Nematode	<i>Caenorhabditis elegans</i>	100,300,000	100 MB	19,000	First multicellular animal genome sequenced, December 1998 ^[80]
Insect	<i>Drosophila melanogaster</i> (fruit fly)	175,000,000	175 MB	13,600	Size variation based on strain (175-180Mb; standard y w strain is 175Mb) ^[81]
Insect	<i>Apis mellifera</i> (honey bee)	236,000,000	236 MB	10,157	^[82]
Insect	<i>Bombyx mori</i> (silk moth)	432,000,000	432 MB	14,623	14,623 predicted genes ^[83]
Insect	<i>Solenopsis invicta</i> (fire ant)	480,000,000	480 MB	16,569	^[84]
Mammal	<i>Mus musculus</i>	2,700,000,000	2.7 GB	20,210	^[85]
Mammal	<i>Pan paniscus</i>	3,286,640,000	3.3 GB	20,000	Bonobo - estimated genome size 3.29 billion bp ^[86]

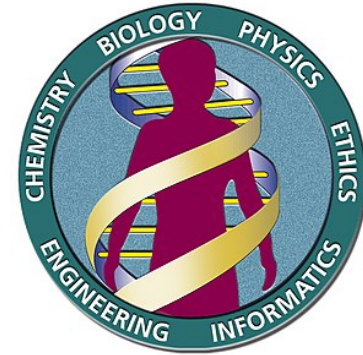
Organism type ⇅	Organism ⇅	Genome size (base pairs) ⇅		Approx. no. of genes ⇅	Note
Nematode	<i>Caenorhabditis elegans</i>	100,300,000	100 MB	19,000	First multicellular animal genome sequenced, December 1998 ^[80]
Insect	<i>Drosophila melanogaster</i> (fruit fly)	175,000,000	175 MB	13,600	Size variation based on strain (175-180Mb; standard y w strain is 175Mb) ^[81]
Insect	<i>Apis mellifera</i> (honey bee)	236,000,000	236 MB	10,157	^[82]
Insect	<i>Bombyx mori</i> (silkworm)	432,000,000	432 MB	14,623	14,623 predicted genes ^[83]
Insect	<i>Solenopsis invicta</i> (fire ant)	480,000,000	480 MB	16,569	^[84]
Mammal	<i>Mus musculus</i>	2,700,000,000	2.7 GB	20,210	^[85]
Mammal	<i>Pan paniscus</i>	3,286,640,000	3.3 GB	20,000	Bonobo - estimated genome size 3.29 billion bp ^[86]
Mammal	<i>Homo sapiens</i>	3,000,000,000	3 GB	20,000	<i>Homo sapiens</i> genome size estimated at 3.2 Gbp in 2001 ^{[87][88]} Initial sequencing and analysis of the human genome ^[89]
Bird	<i>Gallus gallus</i>	1,043,000,000	1.0 GB	20,000	^[90]
Fish	<i>Tetraodon nigroviridis</i> (type of puffer fish)	385,000,000	390 MB		Smallest vertebrate genome known estimated to be 340 Mb ^{[91][92]} – 385 Mb. ^[93]
Fish	<i>Protopterus aethiopicus</i> (marbled lungfish)	130,000,000,000	130 GB		Largest vertebrate genome known

Genome browsers

ADP

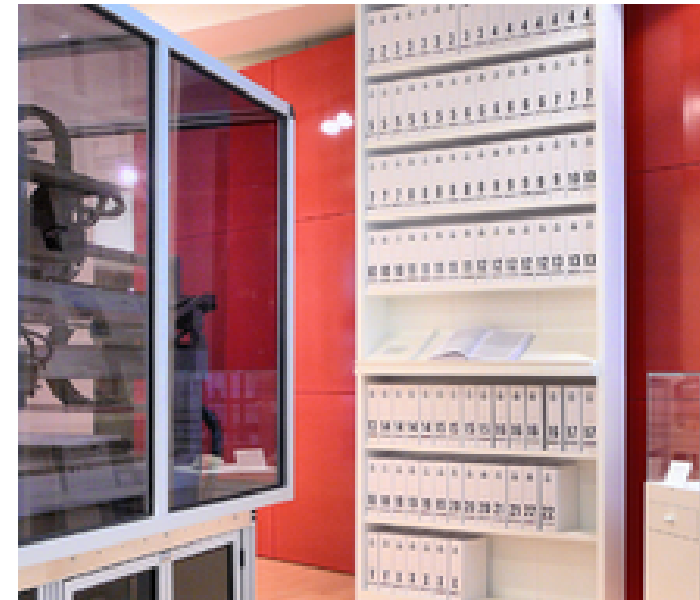
Organism type ⇅	Organism ⇅	Genome size (base pairs) ⇅		Approx. no. of genes ⇅	Note
Nematode	<i>Caenorhabditis elegans</i>	100,300,000	100 MB	19,000	First multicellular animal genome sequenced, December 1998 ^[80]
Insect	<i>Drosophila melanogaster</i> (fruit fly)	175,000,000	175 MB	13,600	Size variation based on strain (175-180Mb; standard y w strain is 175Mb) ^[81]
Insect	<i>Apis mellifera</i> (honey bee)	236,000,000	236 MB	10,157	^[82]
Insect	<i>Bombyx mori</i> (silk moth)	432,000,000	432 MB	14,623	14,623 predicted genes ^[83]
Insect	<i>Solenopsis invicta</i> (fire ant)	480,000,000	480 MB	16,569	^[84]
Mammal	<i>Mus musculus</i>	2,700,000,000	2.7 GB	20,210	^[85]
Mammal	<i>Pan paniscus</i>	3,286,640,000	3.3 GB	20,000	Bonobo - estimated genome size 3.29 billion bp ^[86]
Mammal	<i>Homo sapiens</i>	3,000,000,000	3 GB	20,000	<i>Homo sapiens</i> genome size estimated at 3.2 Gbp in 2001 ^{[87][88]} Initial sequencing and analysis of the human genome ^[89]
Bird	<i>Gallus gallus</i>	1,043,000,000	1.0 GB	20,000	^[90]
Fish	<i>Tetraodon nigroviridis</i> (type of puffer fish)	385,000,000	390 MB		Smallest vertebrate genome known estimated to be 340 Mb ^{[91][92]} – 385 Mb. ^[93]
Fish	<i>Protopterus aethiopicus</i> (marbled lungfish)	130,000,000,000	130 GB		Largest vertebrate genome known

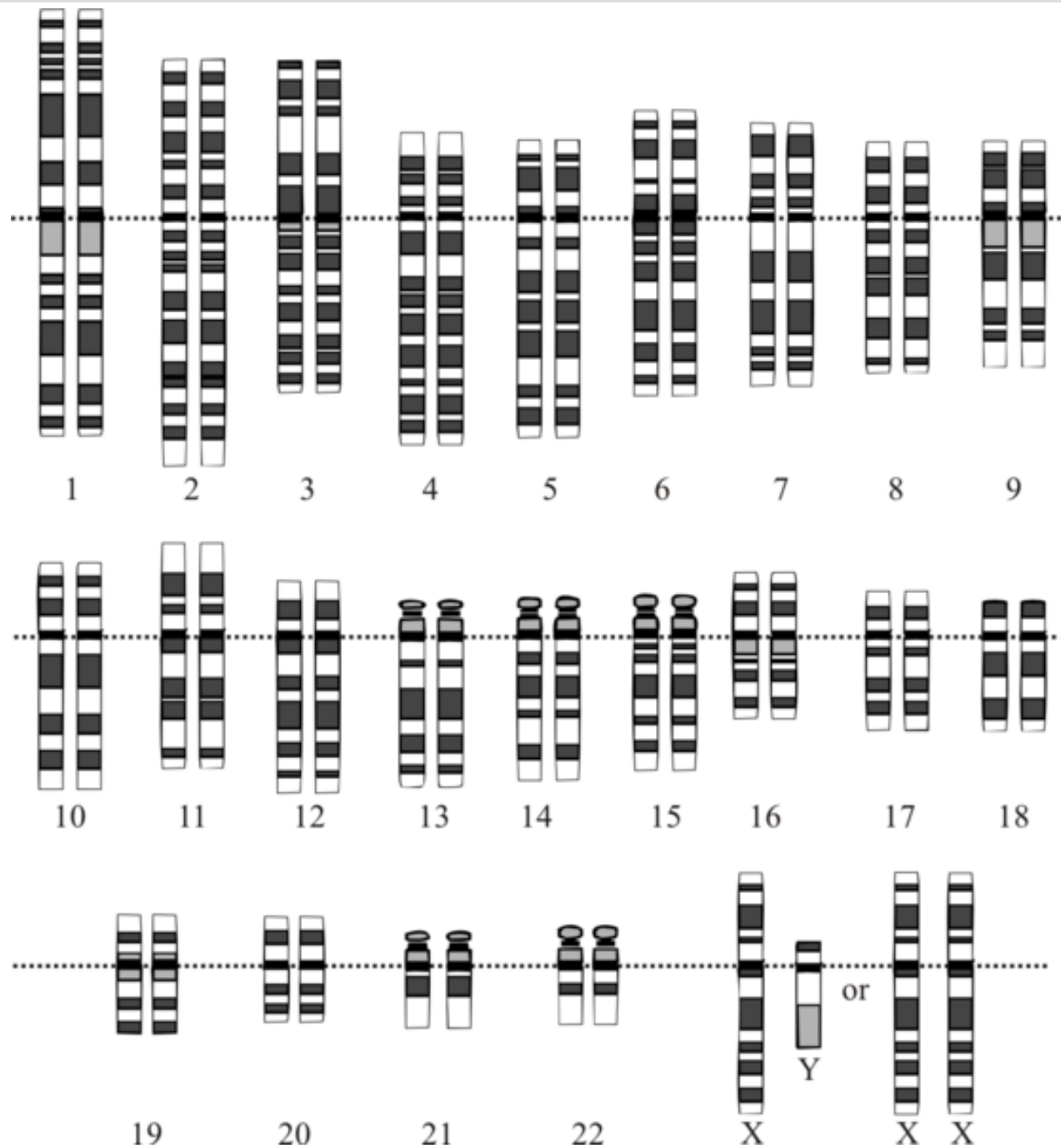
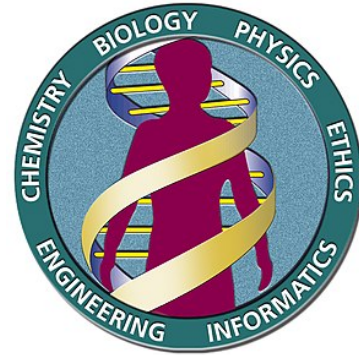
Human Genome Project

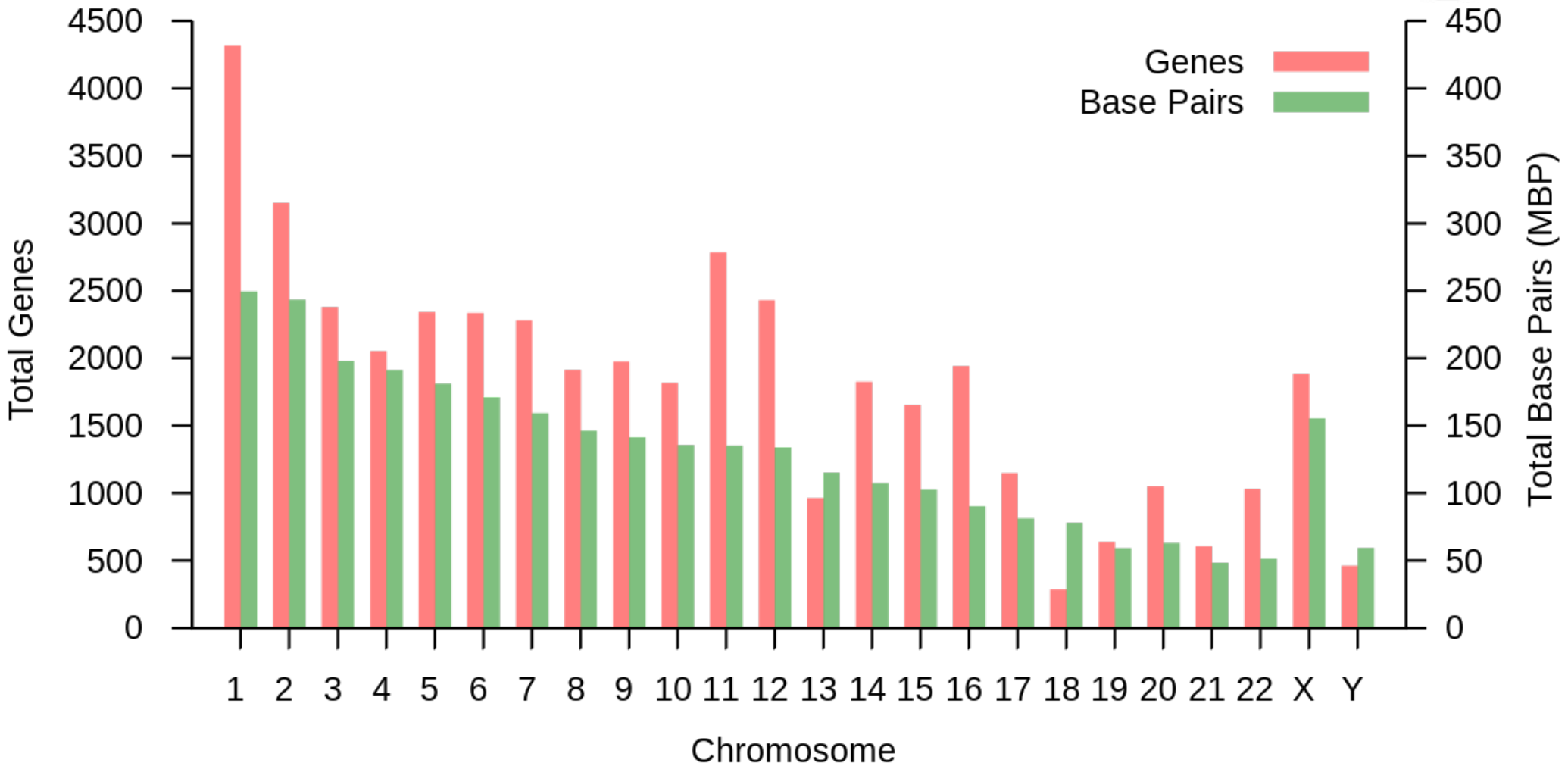
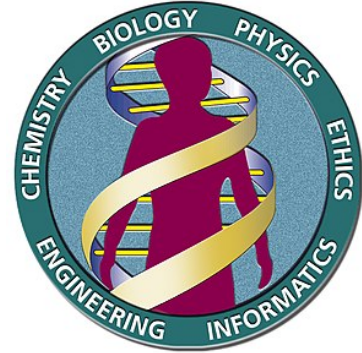


Started in 1990, finished 2003

Cost: \$2.7 billion

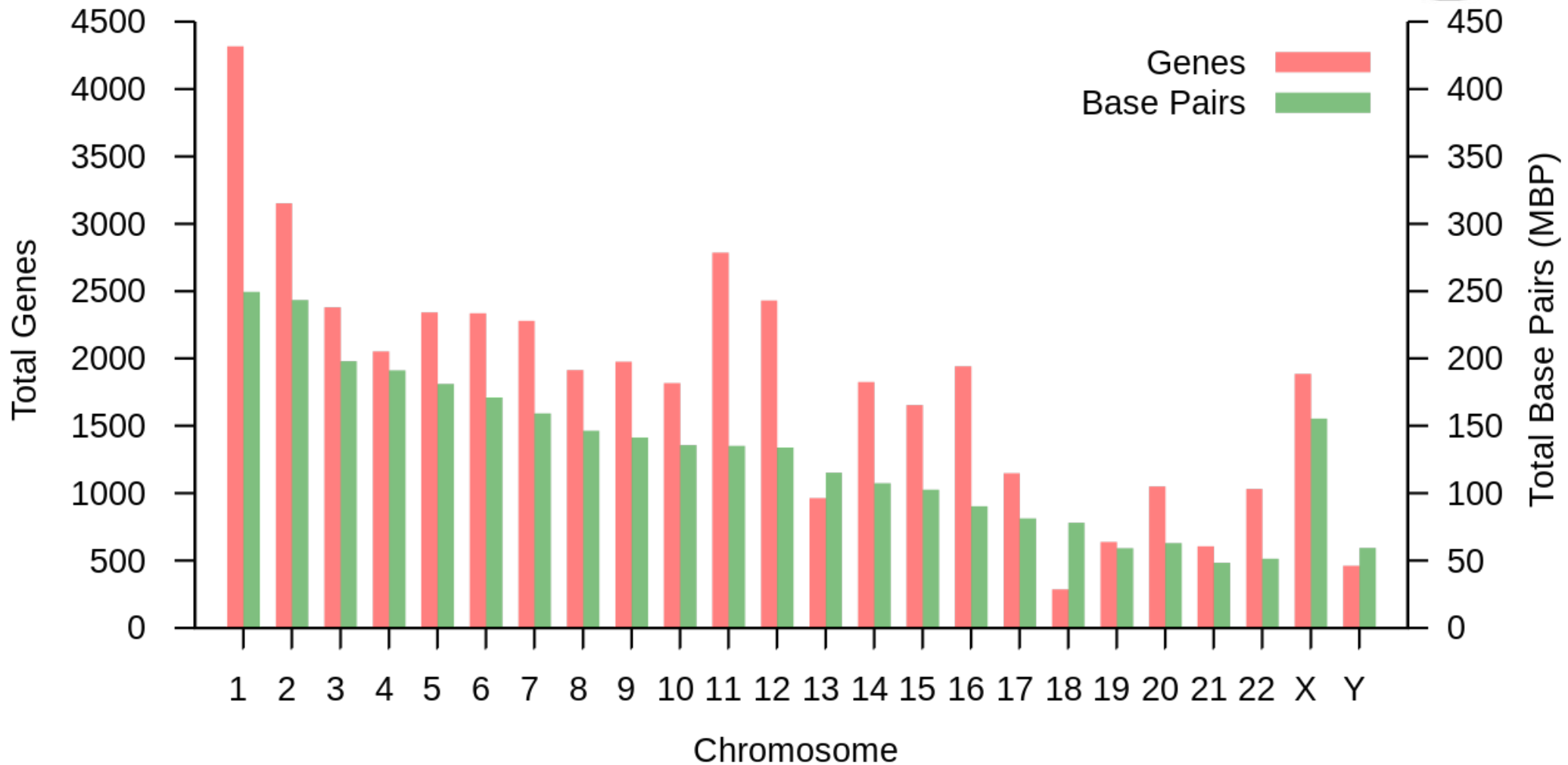
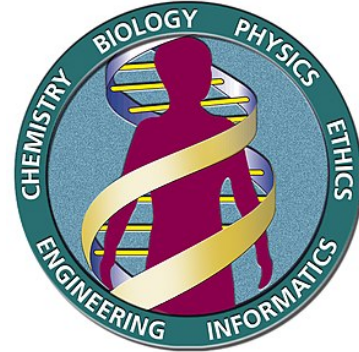






~20,000 genes

(~10,000 active in given tissue + splicing isoforms)



~20,000 genes

(~10,000 active in given tissue + splicing isoforms)

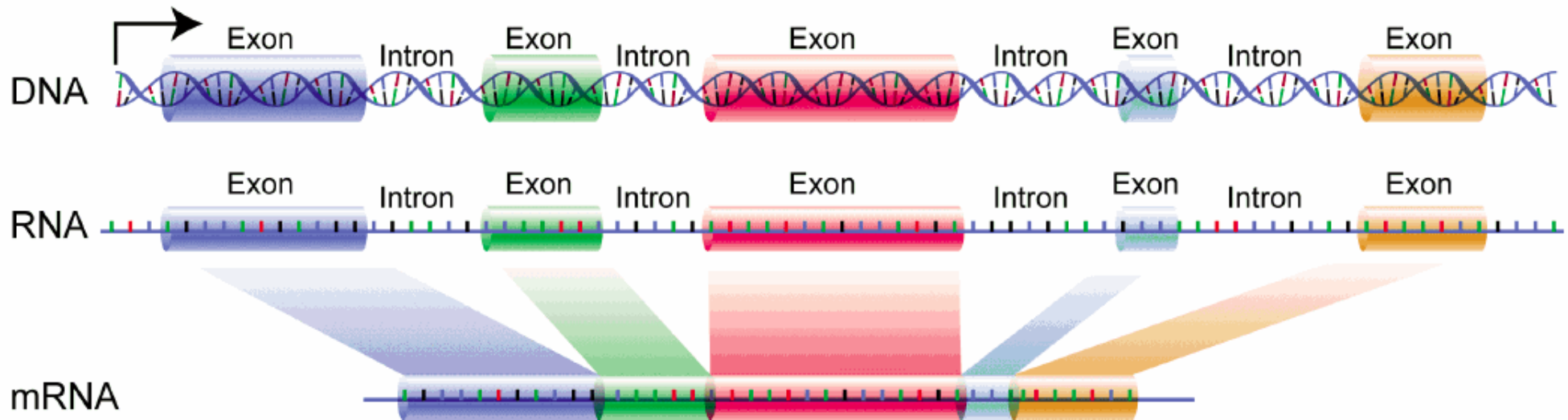
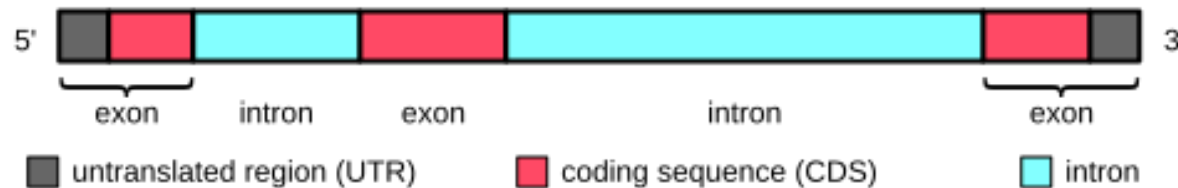
~20,000 genes

(~10,000 active in given tissue + **splicing isoforms**)



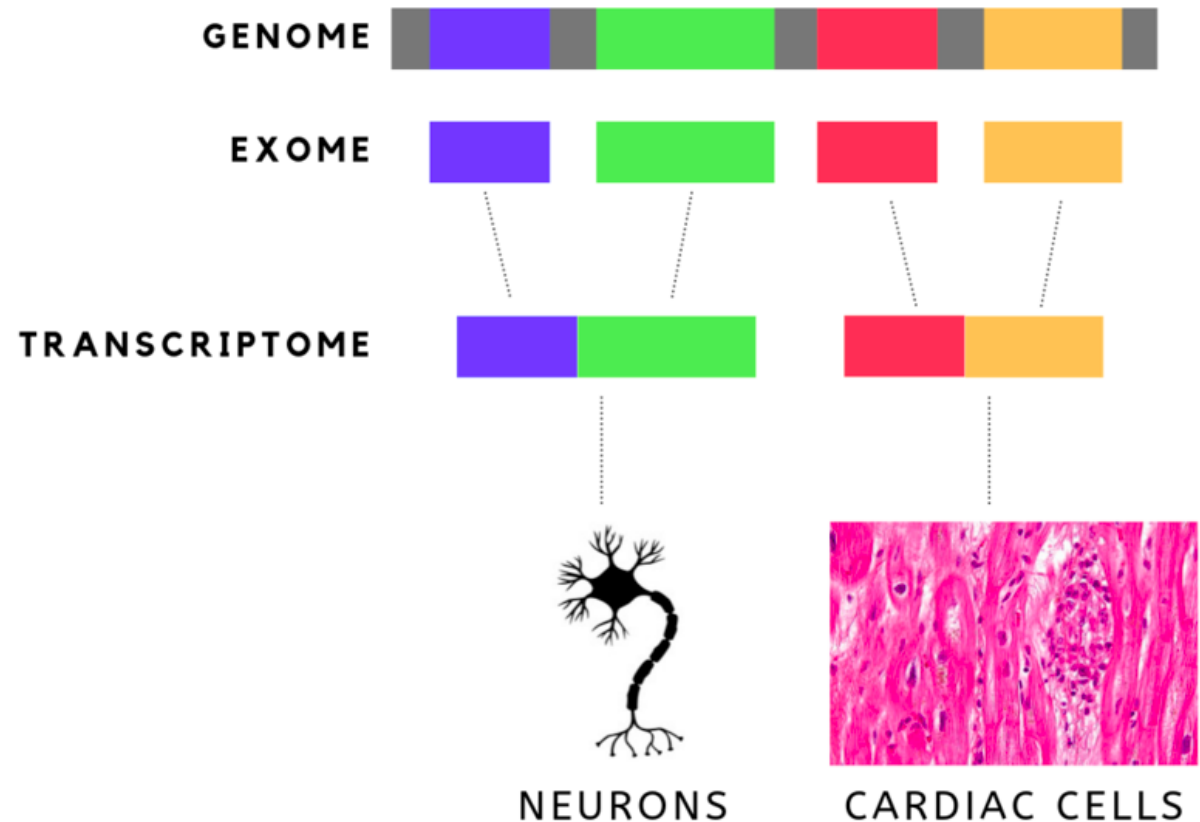
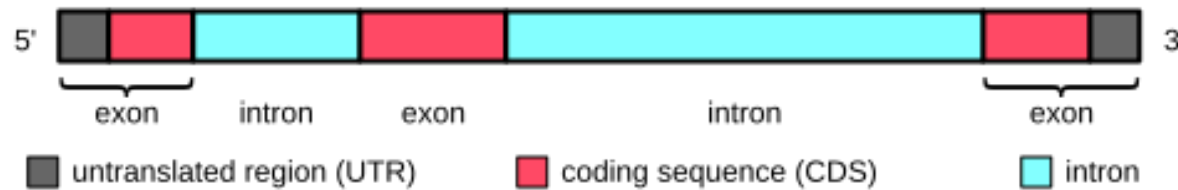
~20,000 genes

(~10,000 active in given tissue + **splicing isoforms**)



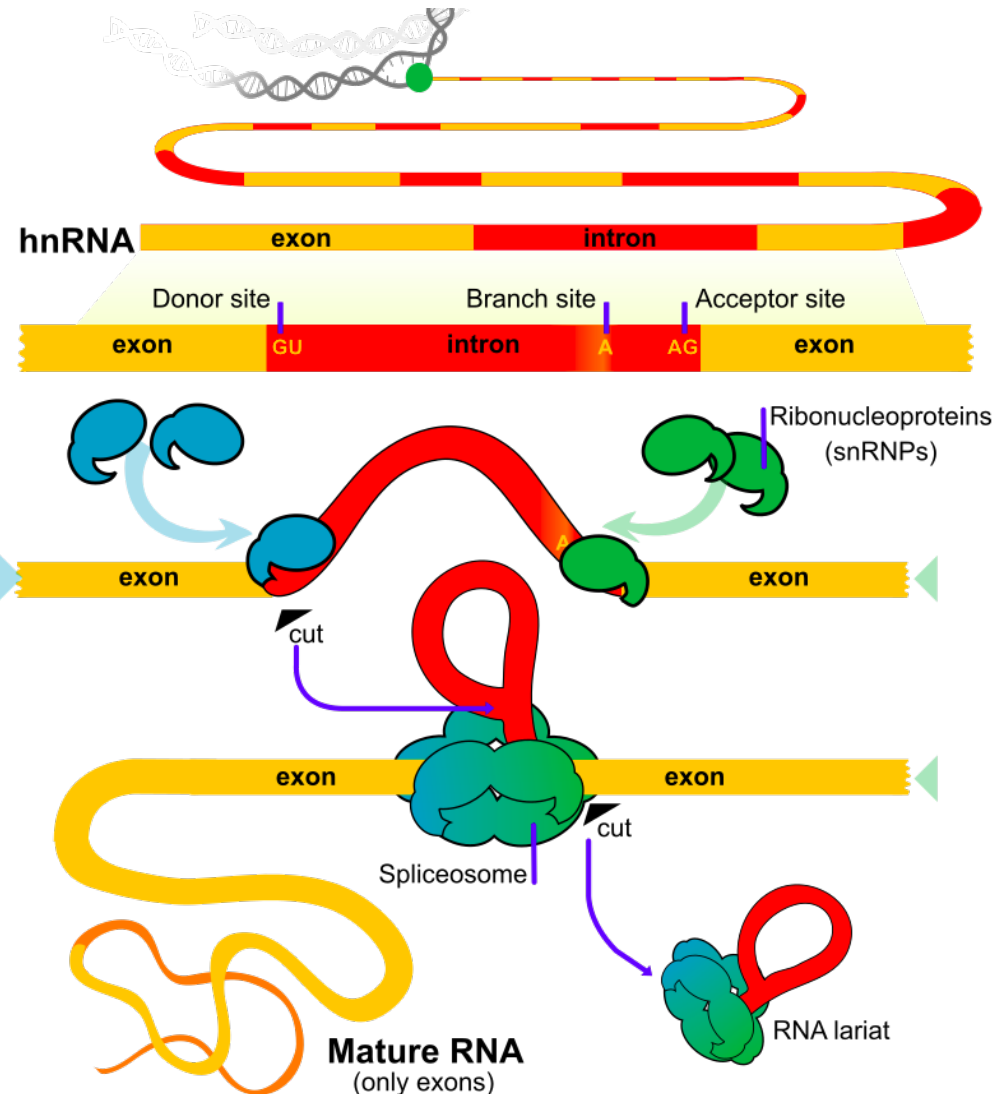
~20,000 genes

(~10,000 active in given tissue + **splicing isoforms**)

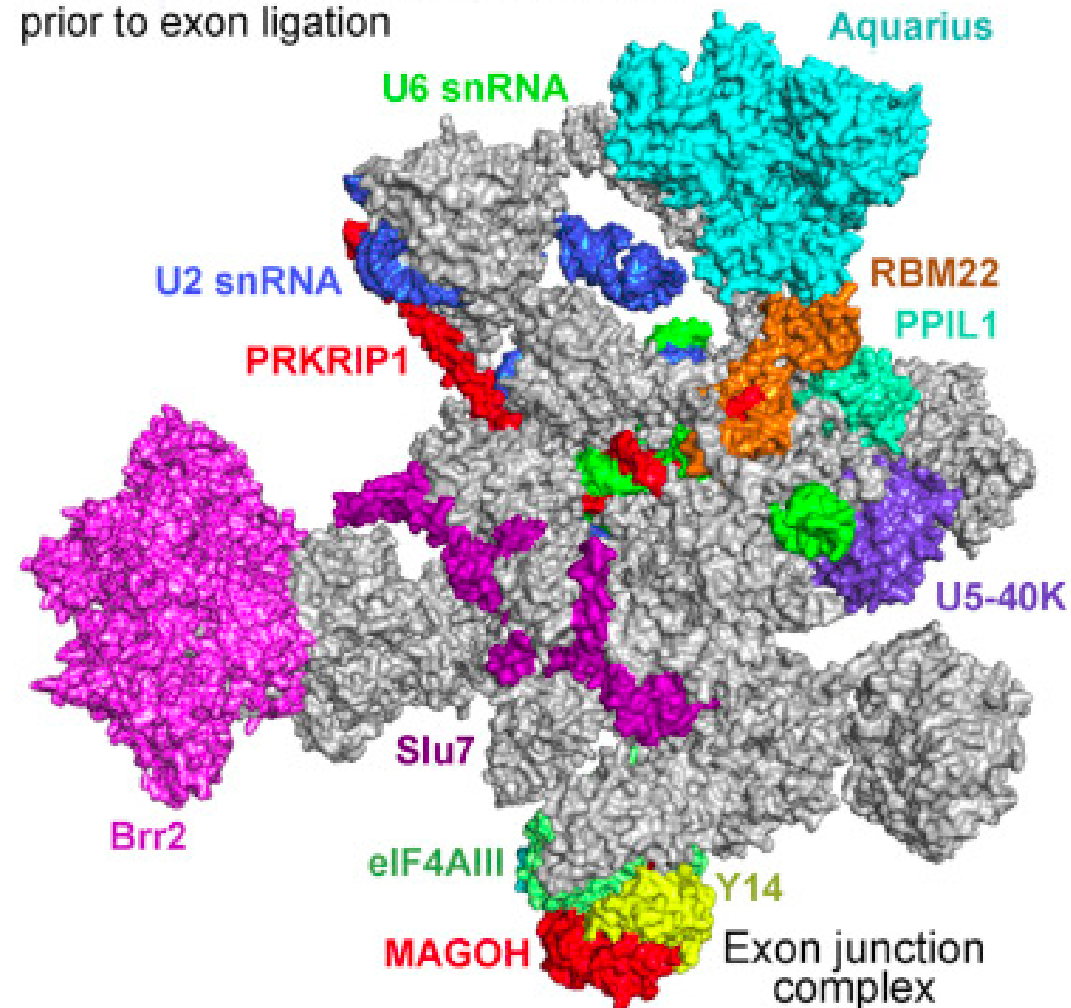


~20,000 genes

(~10,000 active in given tissue + **splicing isoforms**)



Structure of the human spliceosome prior to exon ligation



~20,000 genes

(~10,000 active in given tissue + splicing isoforms)

19354 MANE.GRCh38.v1.4.refseq_protein.faa

20598 UP000005640_9606.fasta

~20,000 genes

(~10,000 active in given tissue + splicing isoforms)

19354 MANE.GRCh38.v1.4.refseq_protein.faa

20598 UP000005640_9606.fasta

83975 UP000005640_9606_additional.fasta

93663 Homo_sapiens_RefSeq_NCBI.fasta

100025 chess3.1.3.CHM13.protein.fa

101026 catLiftOffGenesV1.protein.fa

106188 chess3.1.3.GRCh38.protein.fa

129662 GCF_009914755.1_T2T-CHM13v2.0_protein.faa

~20,000 genes

(~10,000 active in given tissue + splicing isoforms)

19354 MANE.GRCh38.v1.4.refseq_protein.faa

20598 UP000005640_9606.fasta

83975 UP000005640_9606_additional.fasta

93663 Homo_sapiens_RefSeq_NCBI.fasta

100025 chess3.1.3.CHM13.protein.fa

101026 catLiftOffGenesV1.protein.fa

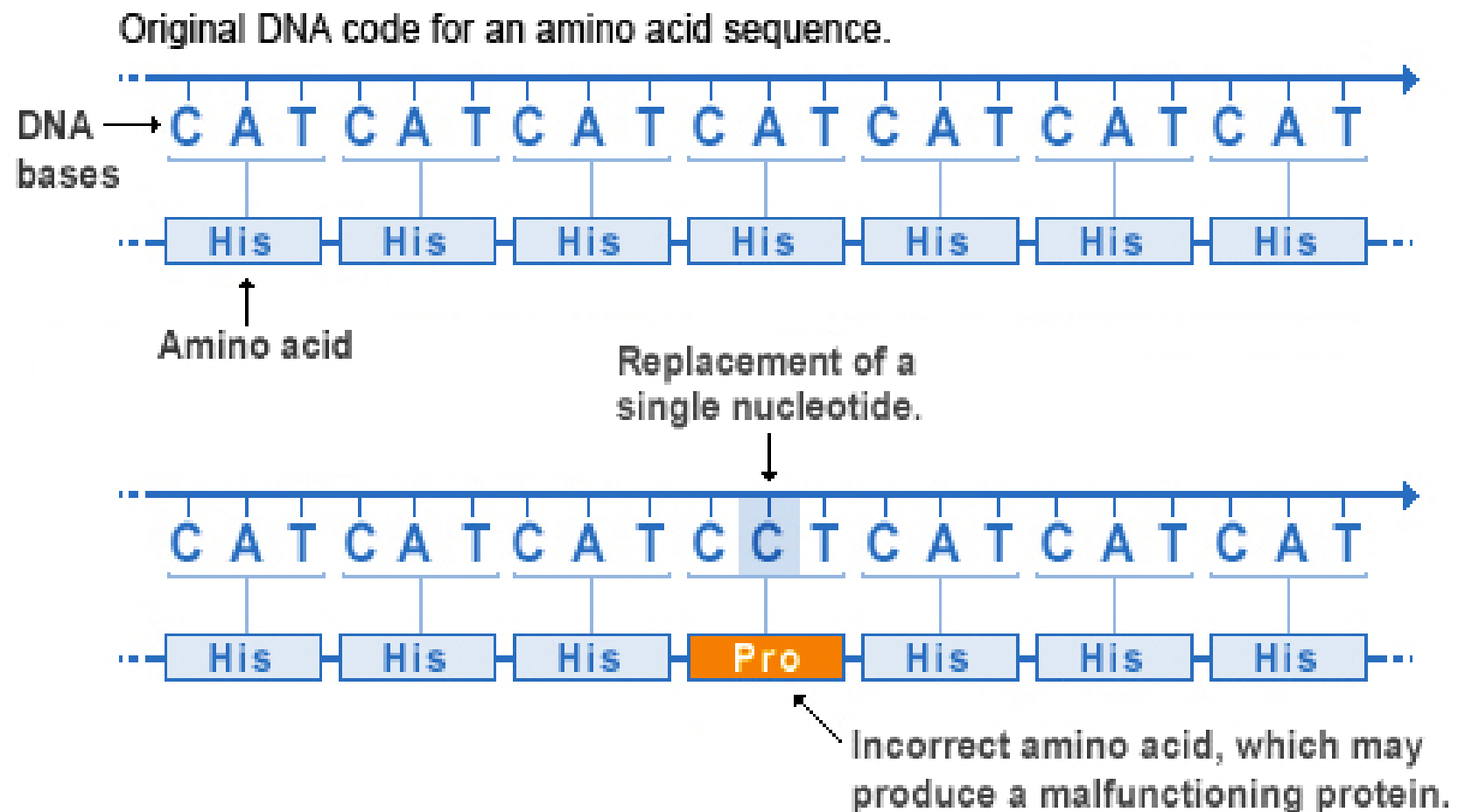
106188 chess3.1.3.GRCh38.protein.fa

129662 GCF_009914755.1_T2T-CHM13v2.0_protein.faa

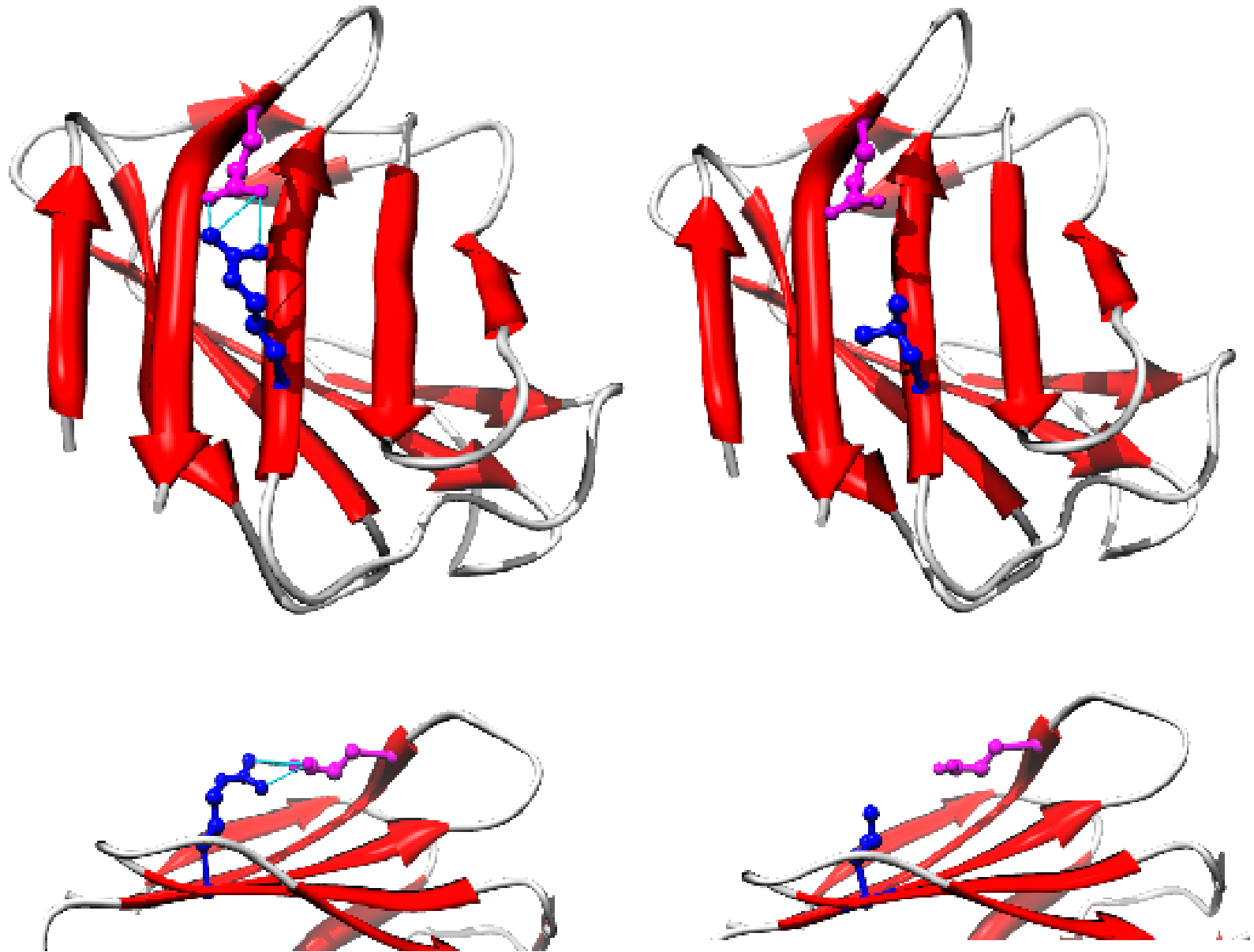
226232 uniprotkb_taxonomy_id_9606_2024_04_03.fasta

SNPs

Missense mutation

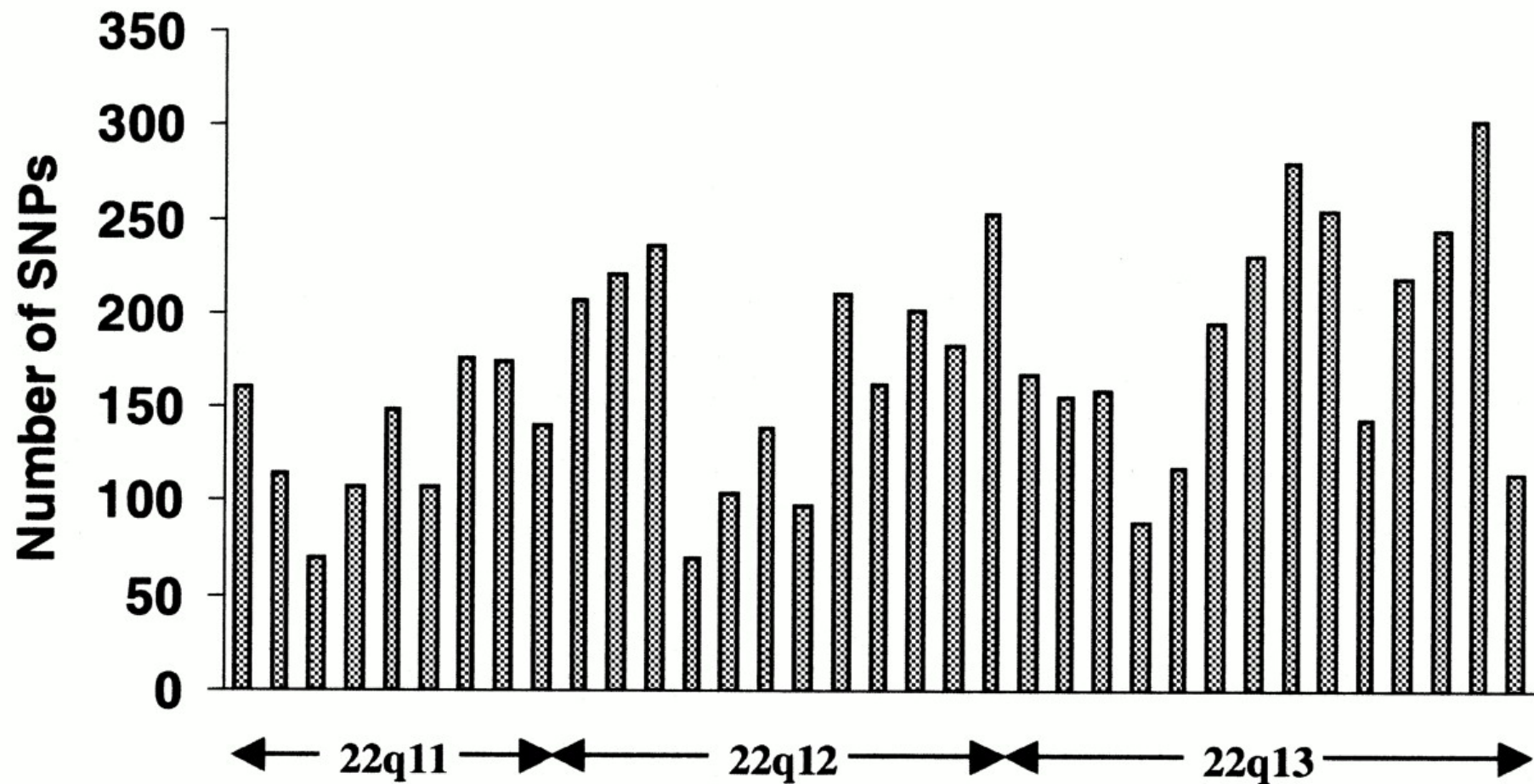


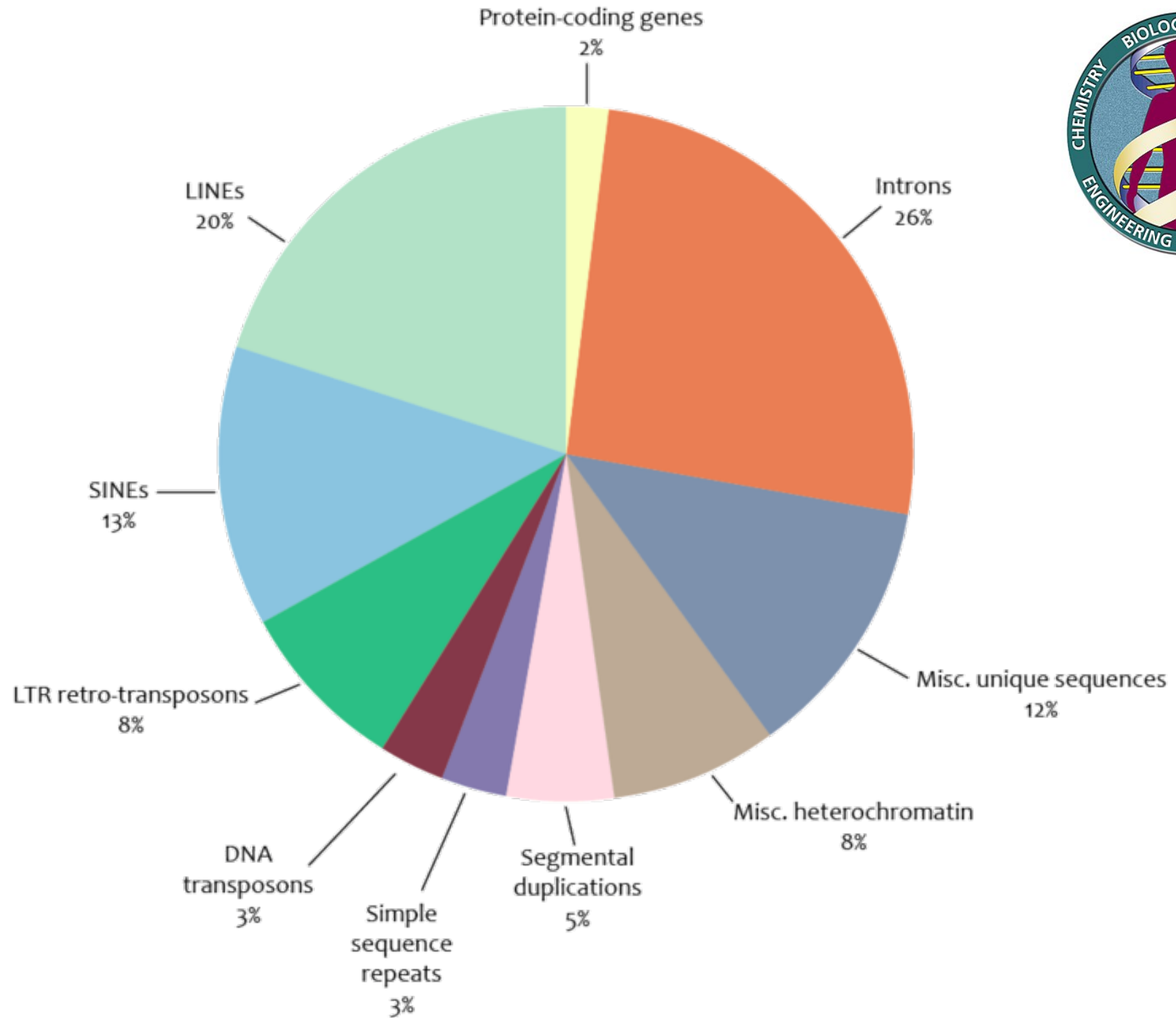
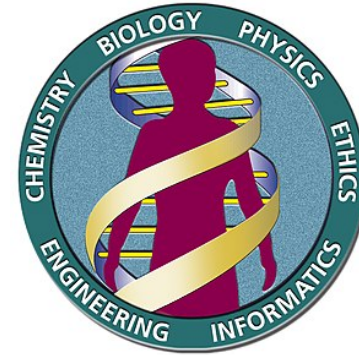
SNPs

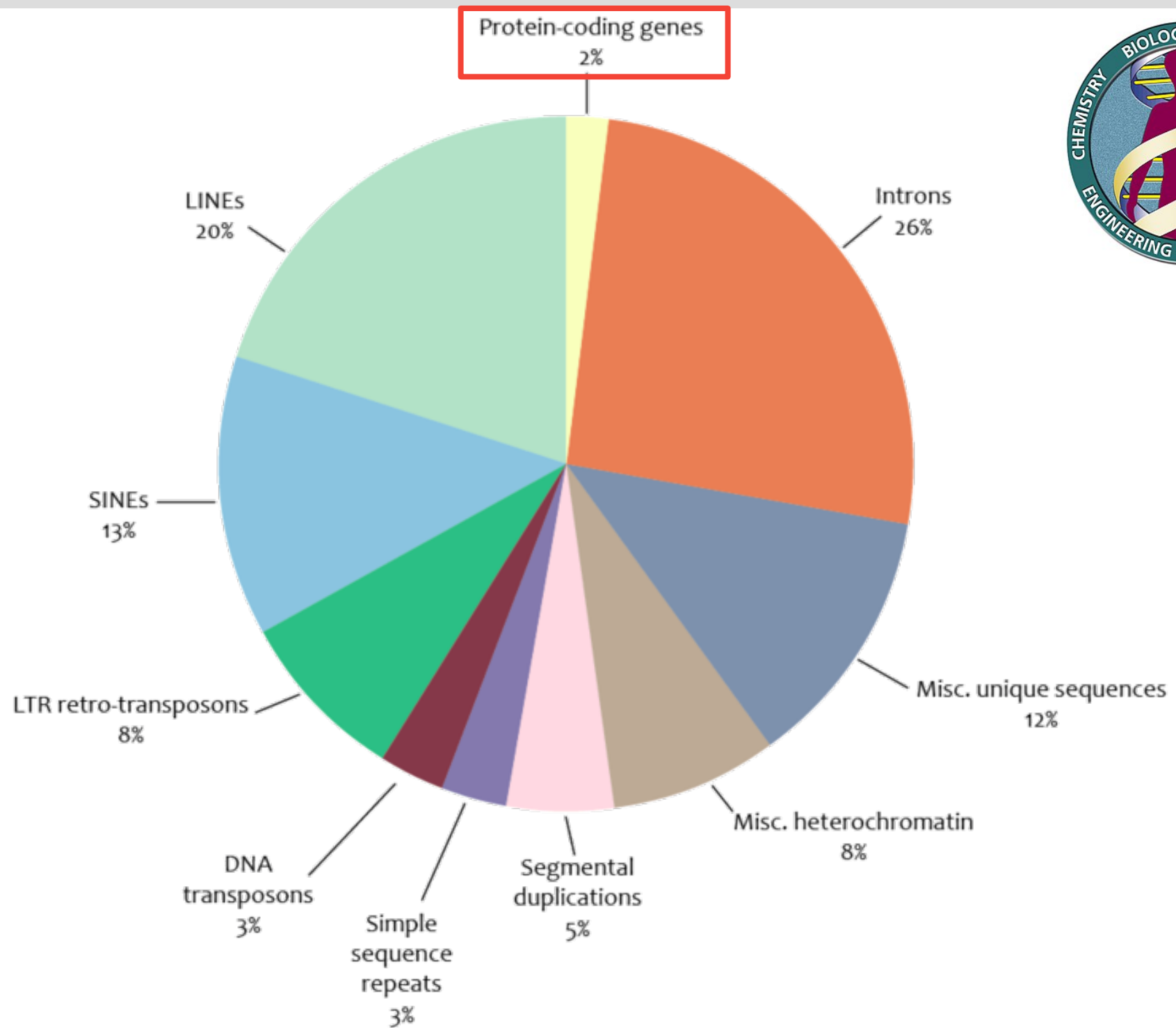
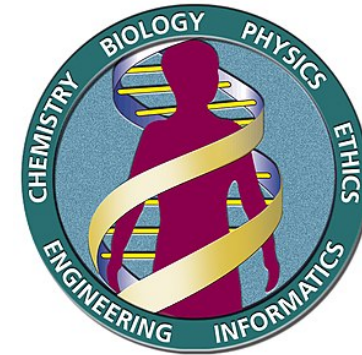


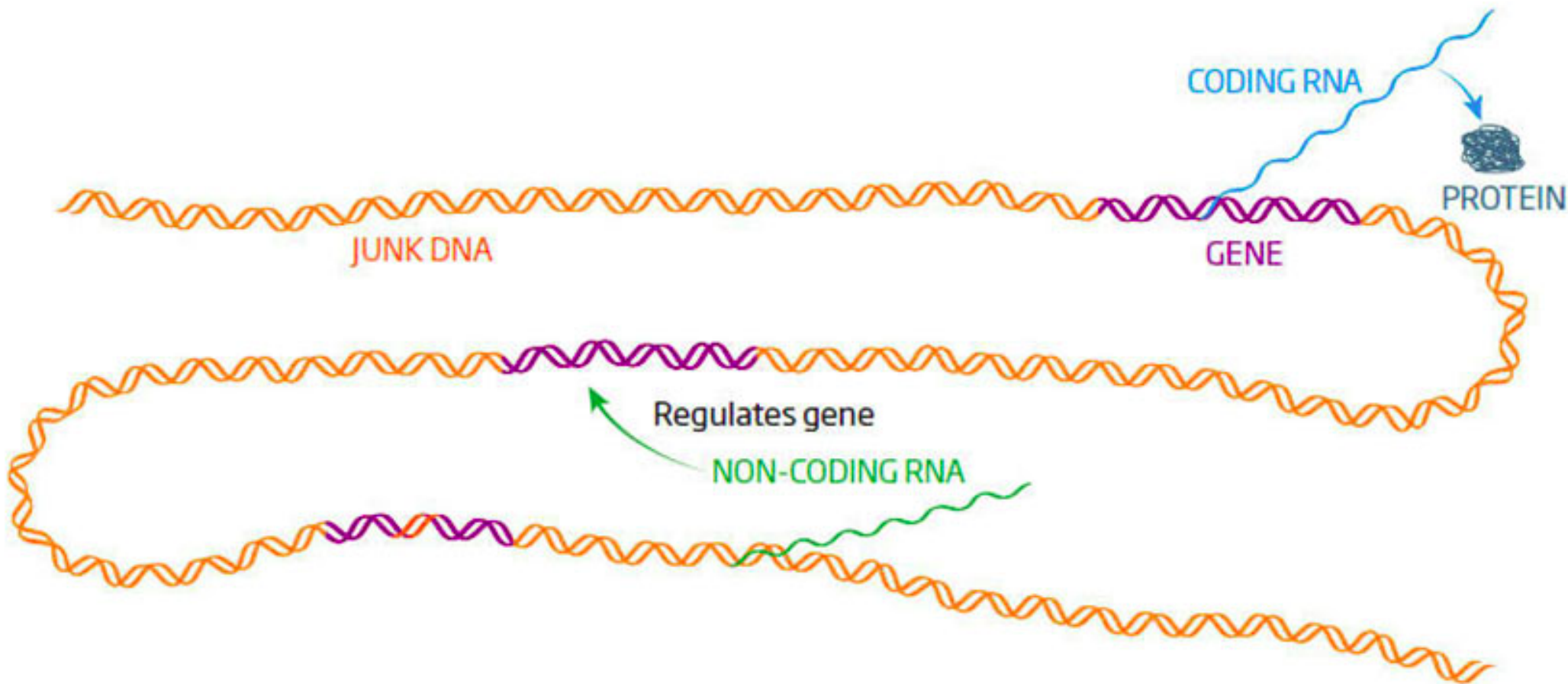
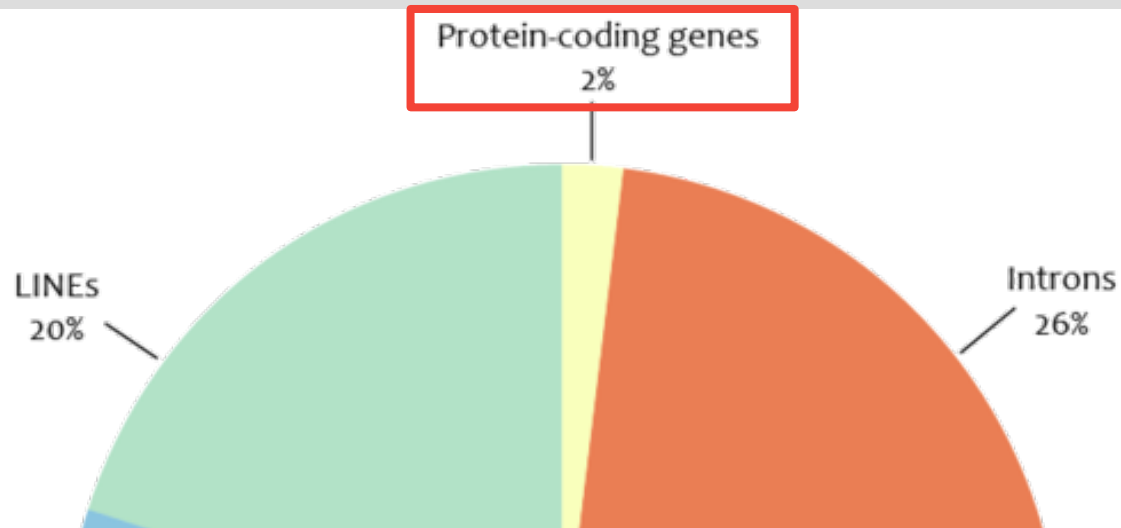
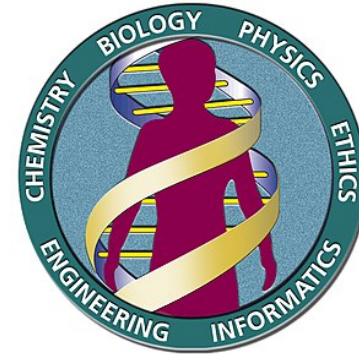
Wild type (left) and mutated (right) form of lamin A (pdb id: 1IFR). Normally, Arginine 527 (blue) forms salt bridge with glutamate 537 (magenta), but R527L substitution results in breaking this interaction (leucine has a nonpolar tail and therefore cannot form a static salt bridge)

SNPs





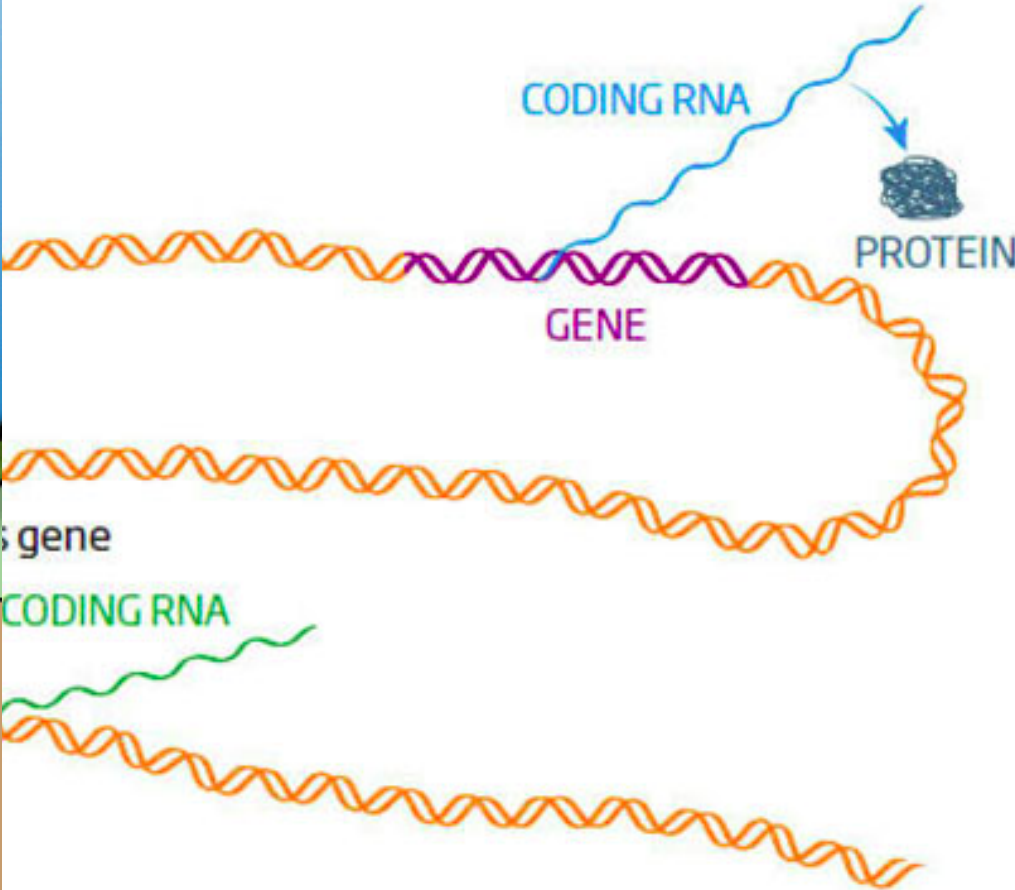


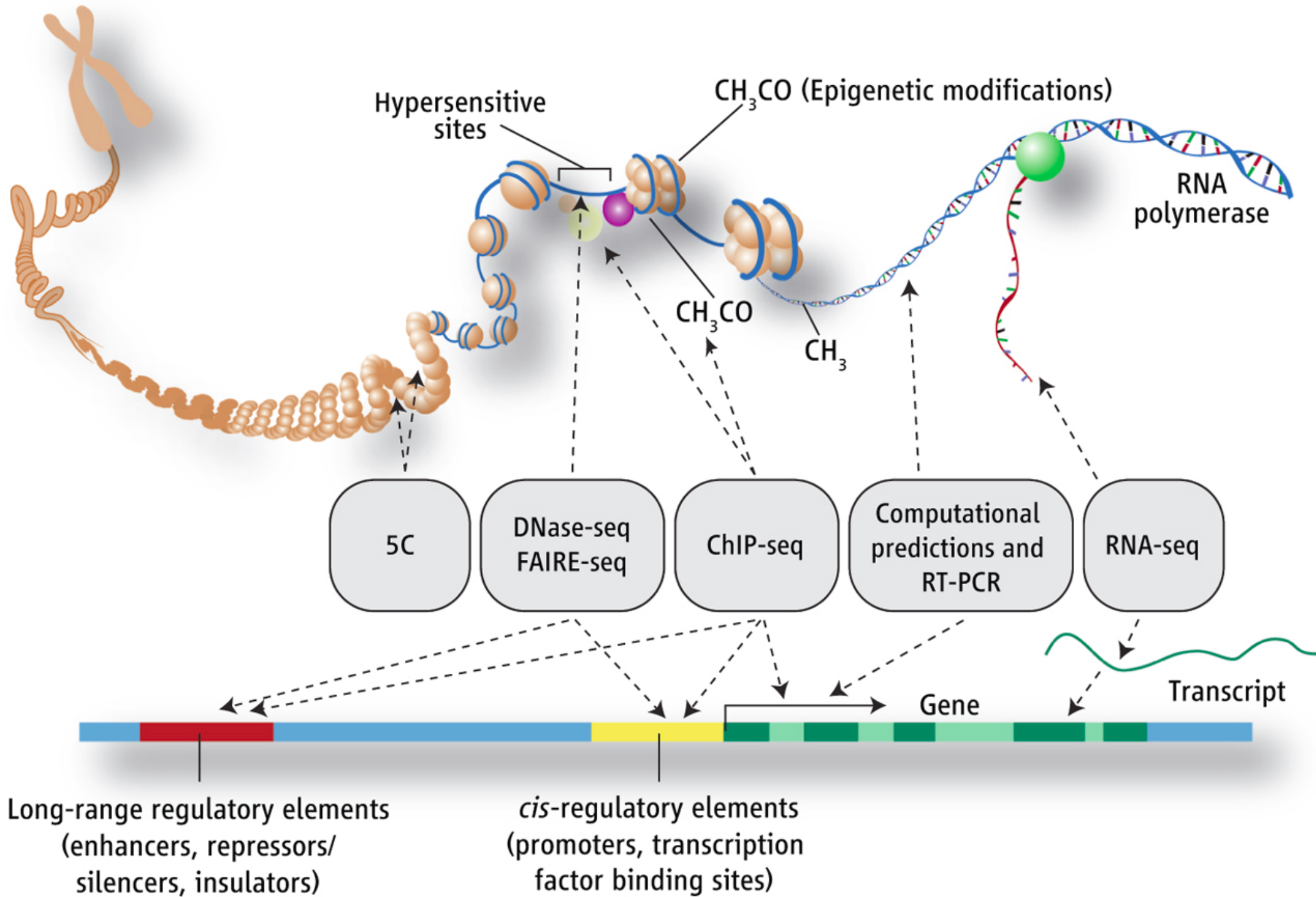


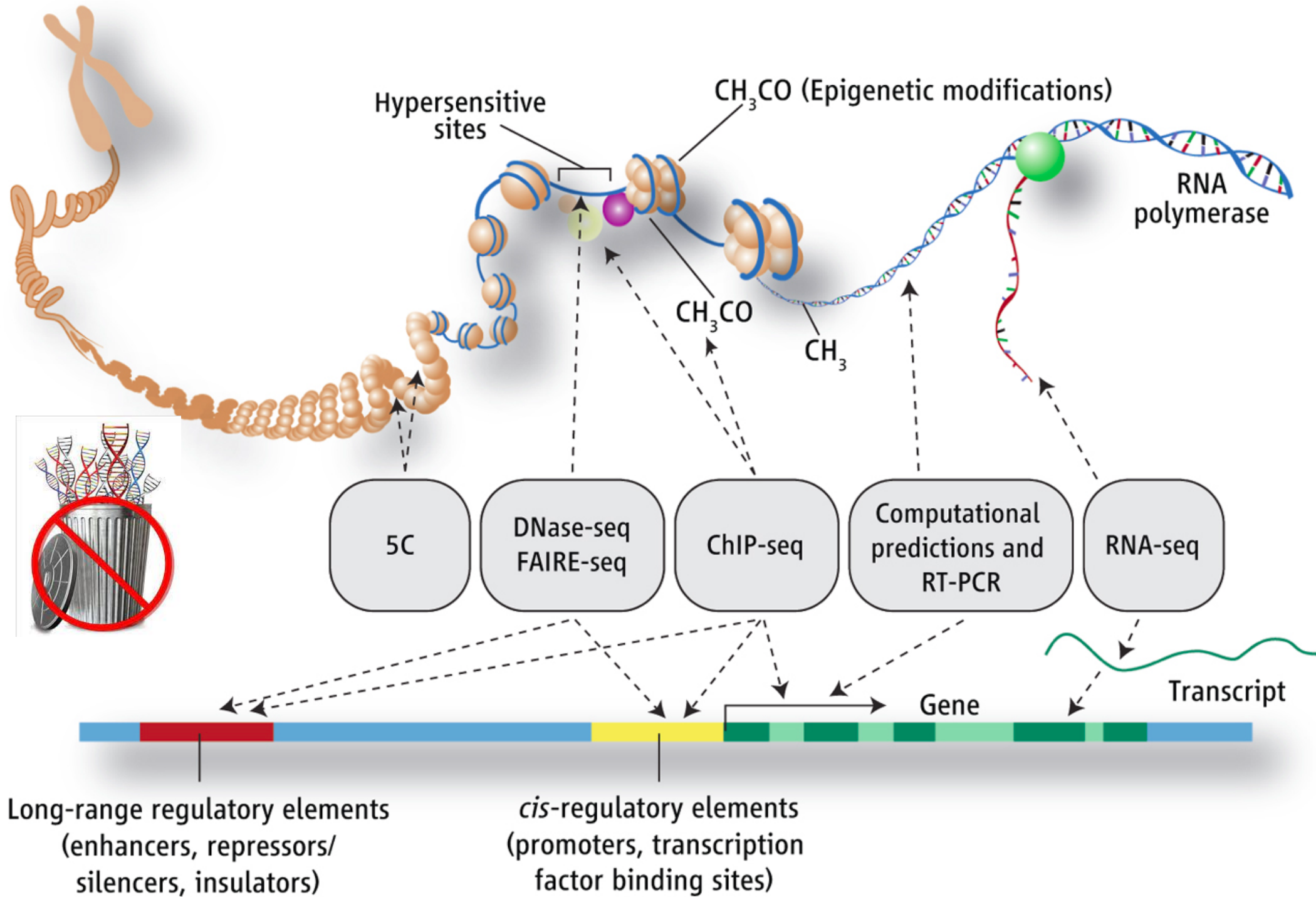
Just because you don't understand,
you can't call us "Junk"!

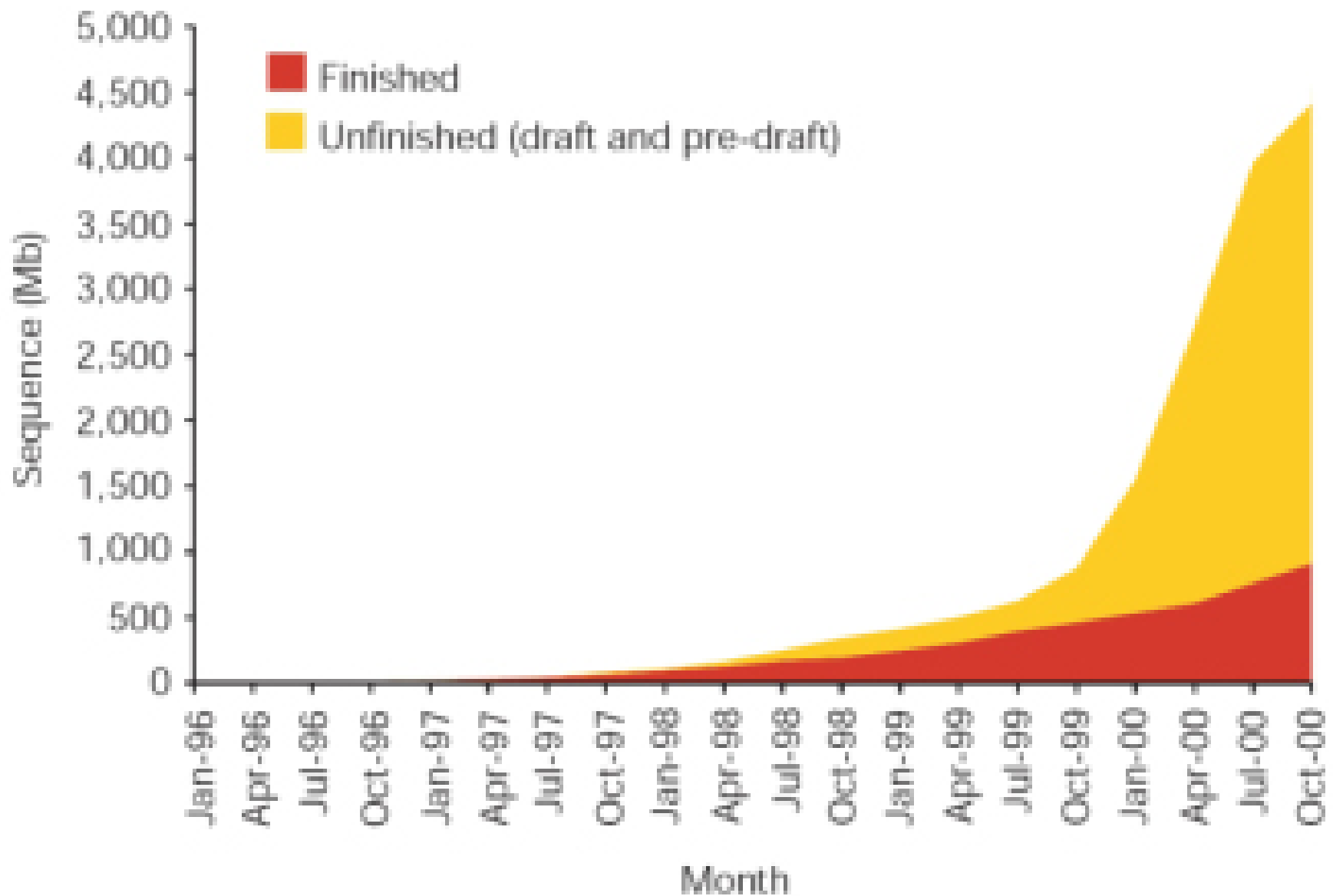


ing genes







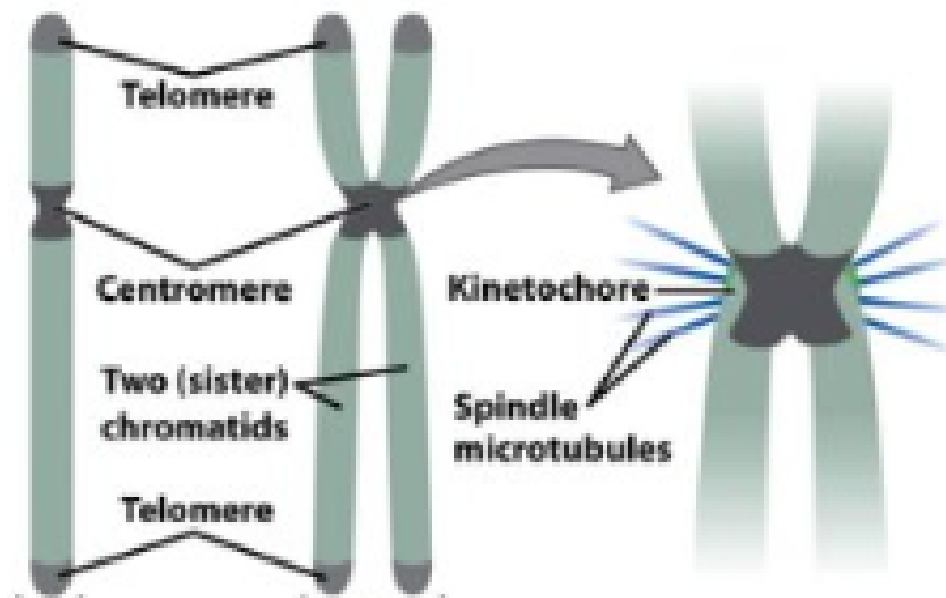


Completeness: 92%



Metaphase chromosomes have:

1. Two arms (short p, long q) and separated by centromere.
2. Centromere (region holds the arms).
3. telomeres (ends of the chromosome).





bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

[HOME](#) | [ABOUT](#) | [S](#)
| [CHANNELS](#)

bioRxiv posts many COVID19-related papers. A reminder: they have not been formally peer-reviewed and should not guide health-related behavior or be reported in the press as conclusive.

[Previous](#)

Posted May 27, 2021

New Results

The complete sequence of a human genome

Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, Alla Mikheenko, Mitchell R. Vollger, Nicolas Altemose, Lev Uralsky, Ariel Gershman, Sergey Aganezov, Savannah J. Hoyt, Mark Diekhans, Glennis A. Logsdon, Michael Alonge, Stylianos E. Antonarakis, Matthew Borchers, Gerard G. Bouffard, Shelise Y. Brooks, Gina V. Caldas, Haoyu Cheng, Chen-Shan Chin, William Chow, Leonardo G. de Lima, Philip C. Dishuck, Richard Durbin,

[Download PDF](#)

[Supplementary M](#)

[Data/Code](#)

Telomere-to-Telomere (T2T) consortium

Elton D. Davis, Peter Korpaljev, Melanie Hirsche, Mikhail Komogorov, Jonas Kordon, Milinn Kremitzki, Heng Li, Valerie V. Maduro, Tobias Marschall, Ann M. McCartney, Jennifer McDaniel, Danny E. Miller, James C. Mullikin, Eugene W. Myers, Nathan D. Olson, Benedict Paten, Paul Peluso, Pavel A. Pevzner, David Porubsky, Tamara Potapova, Evgeny I. Rogaev, Jeffrey A. Rosenfeld, Steven L. Salzberg, Valerie A. Schneider, Fritz J. Sedlazeck, Kishwar Shafin, Colin J. Shew, Alaina Shumate, Yumi Sims, Arian F. A. Smit, Daniela C. Soto, Ivan Sovic, Jessica M. Storer, Aaron Streets, Beth A. Sullivan, Francoise Thibaud-Nissen, James Torrance, Justin Wagner, Brian P. Walenz, Aaron Wenger, Jonathan M. D. Wood, Chunlin Xiao, Stephanie M. Yan, Alice C. Young, Samantha Zarate, Urvashi Surti, Rajiv C. McCoy, Megan Y. Dennis, Ivan A. Alexandrov, Jennifer L. Gerton, Rachel L. O'Neill, Winston Timp, Justin M. Zook, Michael C. Schatz

COVID-19 SAI
medRxiv and

Subject Area

Genomics

Subject Areas

All Articles



bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

HOME | ABOUT | S
| CHANNELS

Search

bioRxiv posts many COVID19-related papers. A reminder: they have not been formally peer-reviewed and should not guide health-related behavior or be reported in the press as conclusive.

← Previous

Posted May 27, 2021

New Results

The complete sequence of a human genome

Sergey Nurk, Sergey Koren, Arang Rhie, Mikko Rautiainen, Andrey V. Bzikadze, Alla Mikheenko, Mitchell R. Vollger, Nicolas Altemose, Lev Uralsky, Ariel Gershman, Sergey Aganezov, Savannah J. Hoyt, Mark Diekhans, Glennis A. Logsdon, Michael Alonge, Stylianos E. Antonarakis, Matthew Borchers, Gerard G. Bouffard, Shelise Y. Brooks, Gina V. Caldas, Haoyu Cheng, Chen-Shan Chin, William Chow, Leonardo G. de Lima, Philip C. Dishuck, Richard Durbin,

Download PDF

Supplementary M

Data/Code

Telomere-to-Telomere (T2T) consortium

Elton D. Calvis, Peter Korpejlov, Melanie Kirsche, Mikhail Komogorov, Jonas Konecny, Milinn Kremitzki, Heng Li, Valerie V. Maduro, Tobias Marschall, Ann M. McCartney, Jennifer McDaniel, Danny E. Miller, James C. Mullikin, Eugene W. Myers, Nathan D. Olson, Benedict Paten, Paul Peluso, Pavel A. Pevzner, David Porubsky, Tamara Potapova, Evgeny I. Rogaev, Jeffrey A. Rosenfeld, Steven L. Salzberg, Valerie A. Schneider, Fritz J. Sedlazeck, Kishwar Shafin, Colin J. Shew, Alaina Shumate, Yumi Sims, Arian F. A. Smit, Daniela C. Soto, Ivan Sovic, Jessica M. Storer, Aaron Streets, Beth A. Sullivan, Francoise Thibaud-Nissen, James Torrance, Justin Wagner, Brian P. Walenz, Aaron Wenger, Jonathan M. D. Wood, Chunlin Xiao, Stephanie M. Yan, Alice C. Young, Samantha Zarate, Urvashi Surti, Rajiv C. McCoy, Megan Y. Dennis, Ivan A. Alexandrov, Jennifer L. Gerton, Rachel L. O'Neill, Winston Timp, Justin M. Zook, Michael C. Schatz

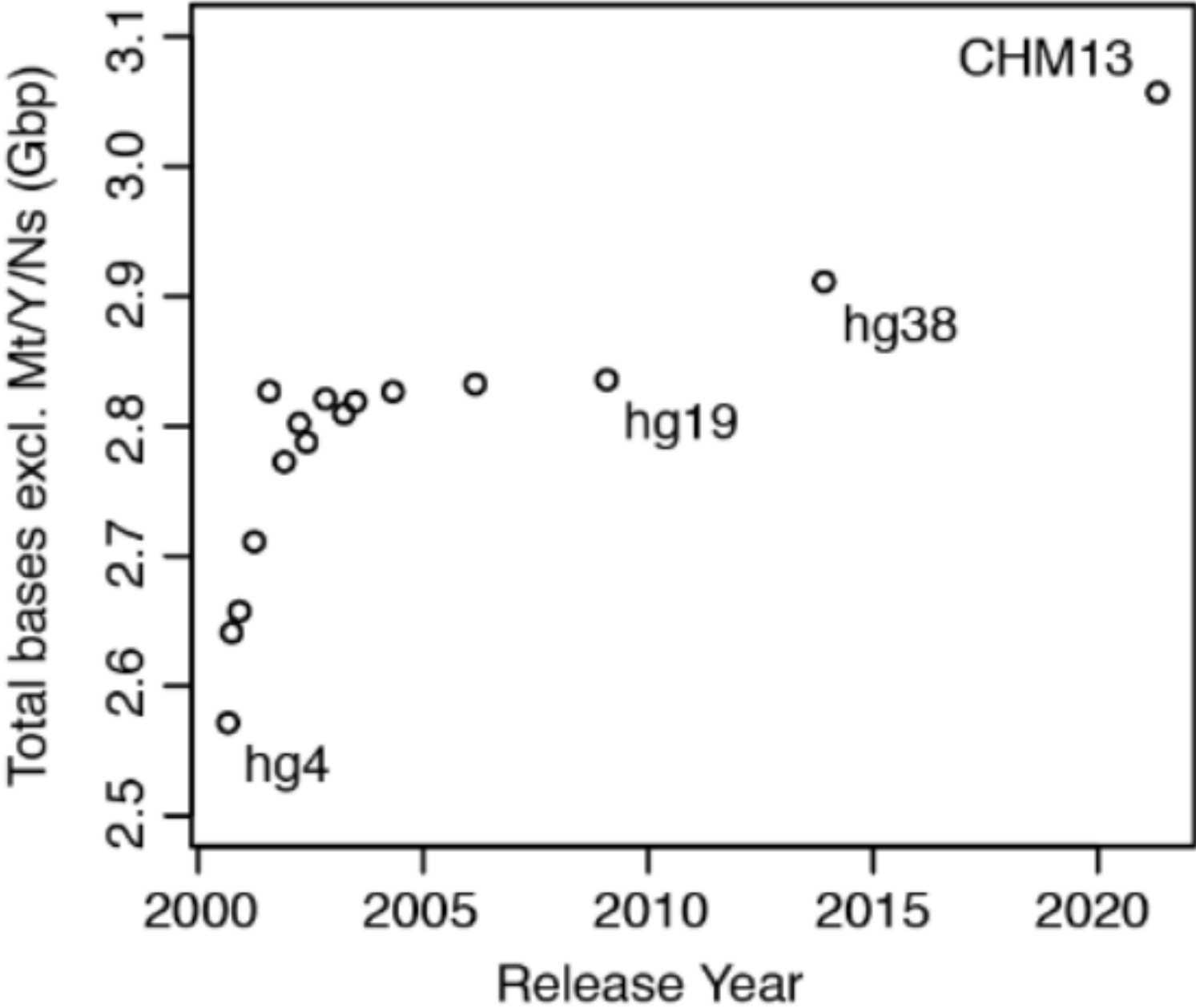
COVID-19 SAI
medRxiv and

Subject Area

Genomics

Subject Areas

All Articles



Telomere-to-telomere consortium

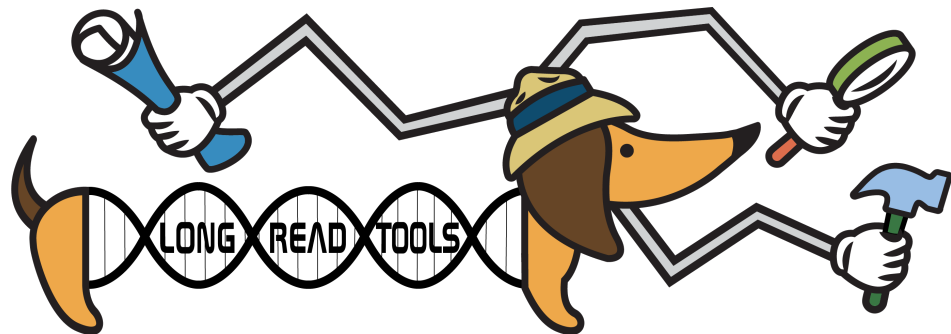
Introduction

We have sequenced the CHM13hTERT human cell line with a number of technologies. Human genomic DNA was extracted from the cultured cell line. As the DNA is native, modified bases will be preserved. The data includes 30x [PacBio HiFi](#), 120x coverage of [Oxford Nanopore](#), 70x [PacBio CLR](#), 50x [10X Genomics](#), as well as [BioNano DLS](#) and [Arima Genomics HiC](#). Most raw data is available from this site, with the exception of the PacBio data which was generated by the University of Washington/PacBio and is available from [NCBI SRA](#).

Telomere-to-telomere consortium

Introduction

We have sequenced the CHM13hTERT human cell line with a number of technologies. Human genomic DNA was extracted from the cultured cell line. As the DNA is native, modified bases will be preserved. The data includes 30x [PacBio HiFi](#), 120x coverage of [Oxford Nanopore](#), 70x [PacBio CLR](#), 50x [10X Genomics](#), as well as [BioNano DLS](#) and [Arima Genomics HiC](#). Most raw data is available from this site, with the exception of the PacBio data which was generated by the University of Washington/PacBio and is available from [NCBI SRA](#).



2.3 Mb

Nanopore sequencing provides the **longest read** lengths, from 500 bp to the current record of 2.3 Mb [16], with 10–30-kb genomic libraries being common. Feb 7, 2020

<https://genomebiology.biomedcentral.com/articles>

Opportunities and challenges in long-read sequencing data ...

We have sequenced the GM12878 cell line with a number of technologies. Human genomic DNA was extracted from the cultured cell line. As the DNA is native, modified bases will be preserved. The data includes 30x **PacBio HiFi**, 120x coverage of **Oxford Nanopore**, 70x **PacBio CLR**, 50x **10X Genomics**, as well as **BioNano DLS** and **Arima Genomics HiC**. Most raw data is available from this site, with the exception of the PacBio data which was generated by the University of Washington/PacBio and is available from **NCBI SRA**.

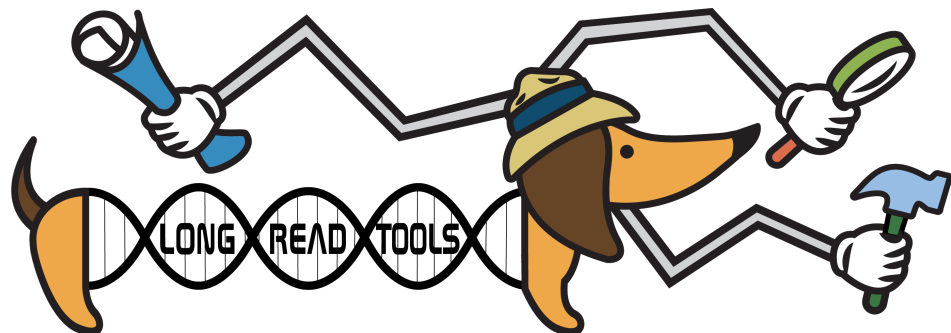


Table 1. Comparison of GRCh38 and T2T-CHM13 human genome assemblies

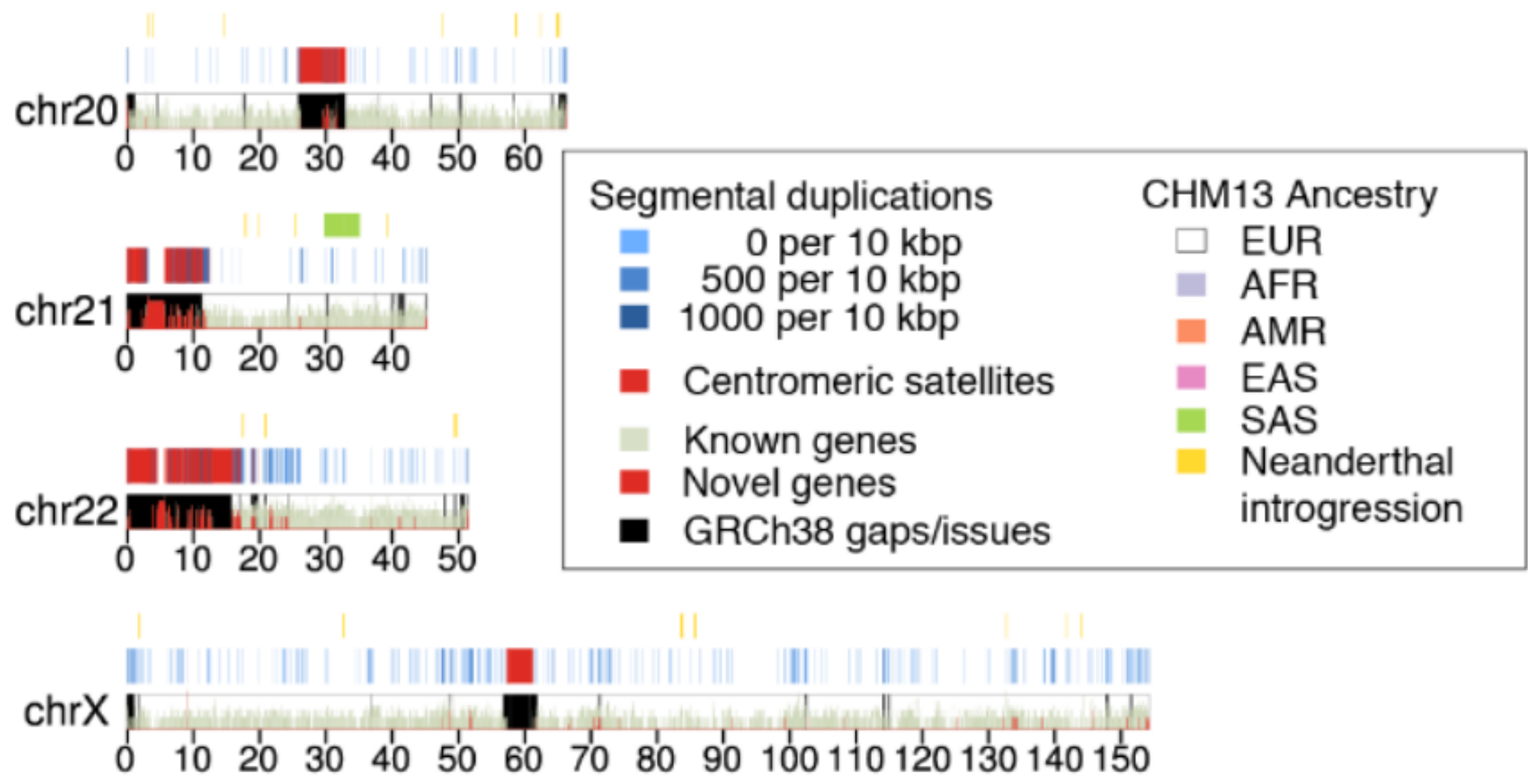
Summary	GRCh38p13	CHM13v1.1	±%
Assembled bases (Gbp)	2.92	3.05	+4.5%
Unplaced bases (Mbp)	11.42	0	-100.0%
Gap bases (Mbp)	120.31	0	-100.0%
# Contigs	949	24	-97.5%
Ctg NG50 (Mbp)	56.41	154.26	+173.5%
# Issues	230	46	-80.0%
Issues (Mbp)	230.43	8.18	-96.5%
Gene Annotation			
# Genes	60,090	63,494	+5.7%
protein coding	19,890	19,969	+0.4%
# Exclusive genes	263	3,604	
protein coding	63	140	
# Transcripts	228,597	233,615	+2.2%
protein coding	84,277	86,245	+2.3%
# Exclusive transcripts	1,708	6,693	
protein coding	829	2,780	

Segmental duplications (SDs)

% SDs	5.00%	6.61%	
SD bases (Mbp)	151.71	201.93	+33.1%
# SDs	24097	41528	+72.3%

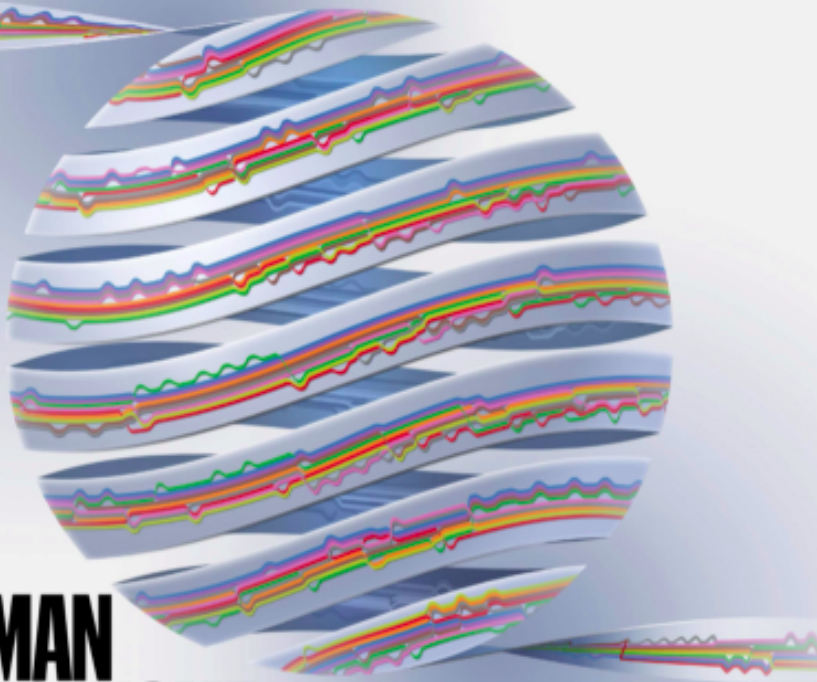
RepeatMasker

% Repeats	50.03%	53.94%	
Repeat bases (Mbp)	1,516.37	1,647.81	+8.7%
LINE	626.33	631.64	+0.8%
SINE	386.48	390.27	+1.0%
LTR	267.52	269.91	+0.9%
Satellite	76.51	150.42	+96.6%
DNA	108.53	109.35	+0.8%
Simple repeat	36.5	77.69	+112.9%
Low complexity	6.16	6.44	+4.6%
Retroposon	4.51	4.65	+3.3%
rRNA	0.21	1.71	+730.4%



The international journal of science / 11 May 2023

nature



HUMAN PANGENOME

Data from 47 individuals combine to create reference resource that reflects human diversity

Considered position

The debate over priorities for autism research

Predictive power

Four steps to fashion a science of human-algorithm behaviour

Screen test

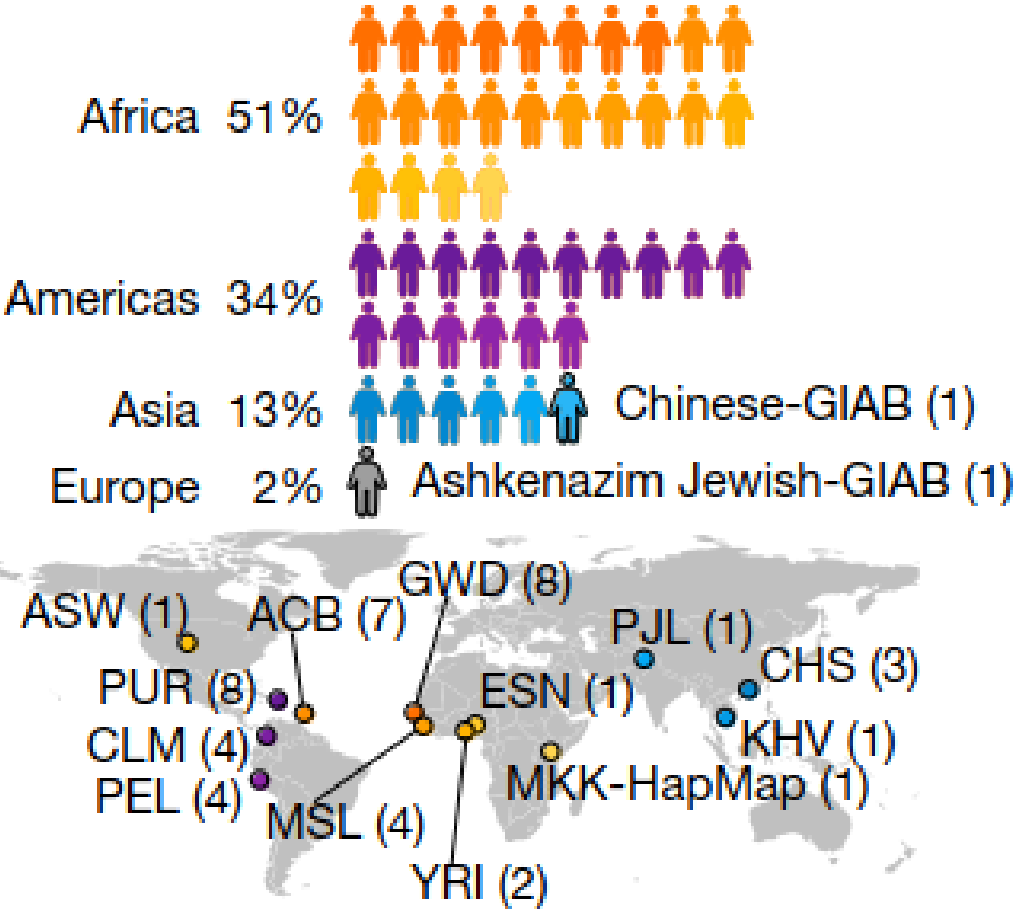
MicroLEDs induced to self-assemble into next-generation display

Vol. 617 No. 7960
nature.com

Human pangenome

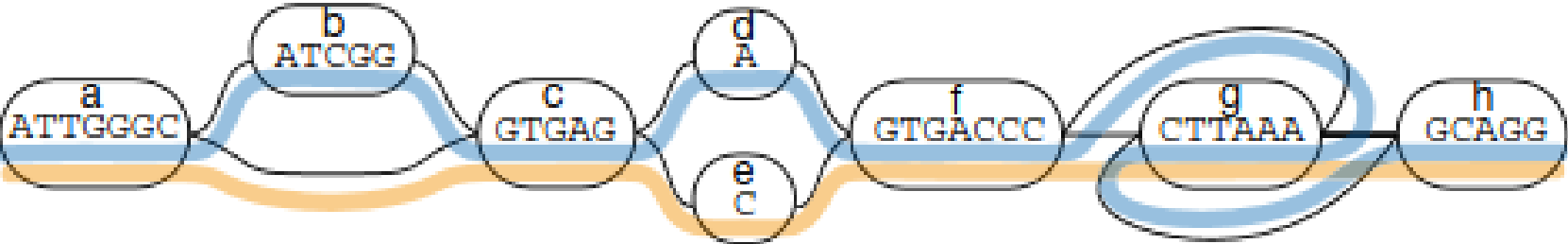
The human reference genome has been the backbone of human genomics since the release of the draft sequence in 2001. But it has its limitations: one genome cannot hope to capture the diversity of the human species. In this week's issue, the Human Pangenome Reference Consortium presents the first draft human pangenome, which **combines genetic material from 47 genetically diverse individuals to provide a more complete picture of the human genome**. Two additional papers use this new resource to study regions of the genome that include repetitive DNA: one looking at segmental duplication

<https://doi.org/10.1038/s41586-023-05896-x>



a

ATTGGGCATCGGGTGAGAGTGACCCTTTAAGGCAGG
ATTGGGC-----GTGAGCGTGACCCCTTAAAGCAGG



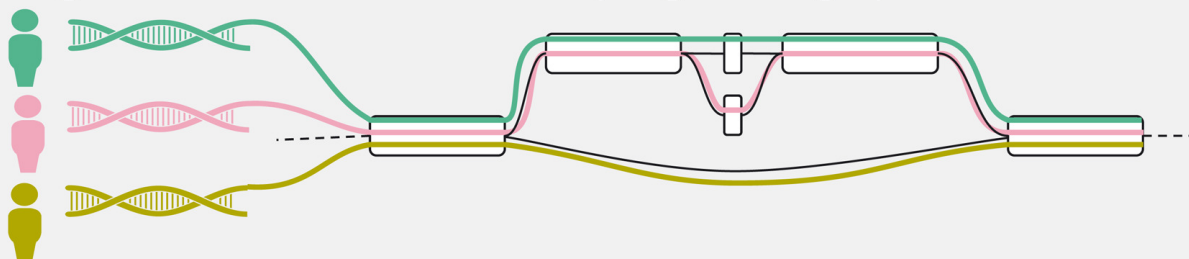
Pangenomics enables genotyping of known structural variants in 5202 diverse genomes

JOUNI SIRÉN ^{ID}, JEAN MONLONG ^{ID}, XIAN CHANG ^{ID}, ADAM M. NOVAK ^{ID}, JORDAN M. EIZENGA ^{ID}, CHARLES MARKELLO ^{ID}, JONAS A. SIBBESEN ^{ID},
GLENN HICKEY ^{ID}, PI-CHUAN CHANG ^{ID}, [...], AND BENEDICT PATEN ^{ID} [+10 authors](#) [Authors Info & Affiliations](#)

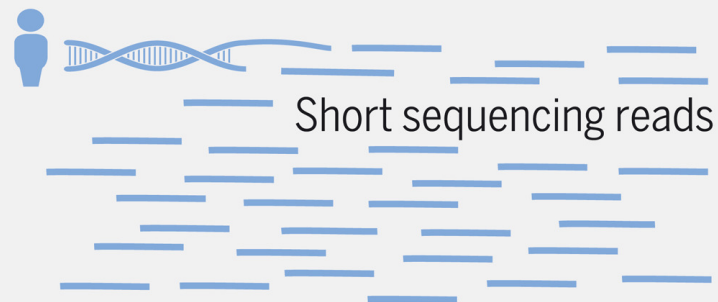
~167,000 structural variants

Catalogs from
long-read studies

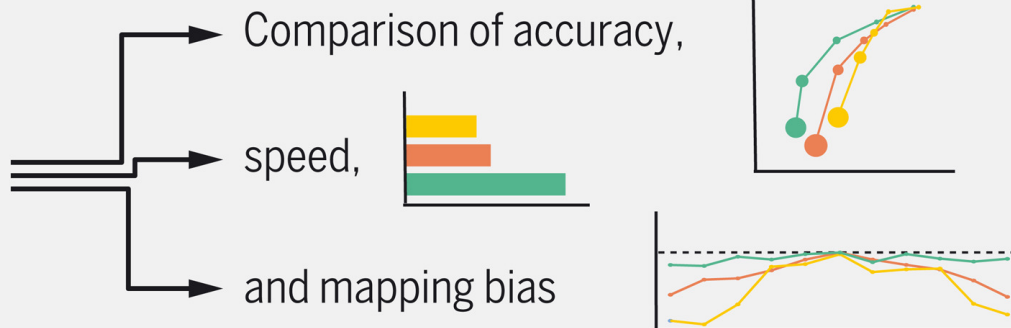
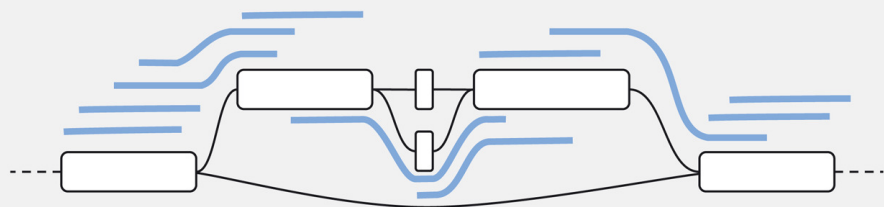
Reference structural variant
pangenome graph



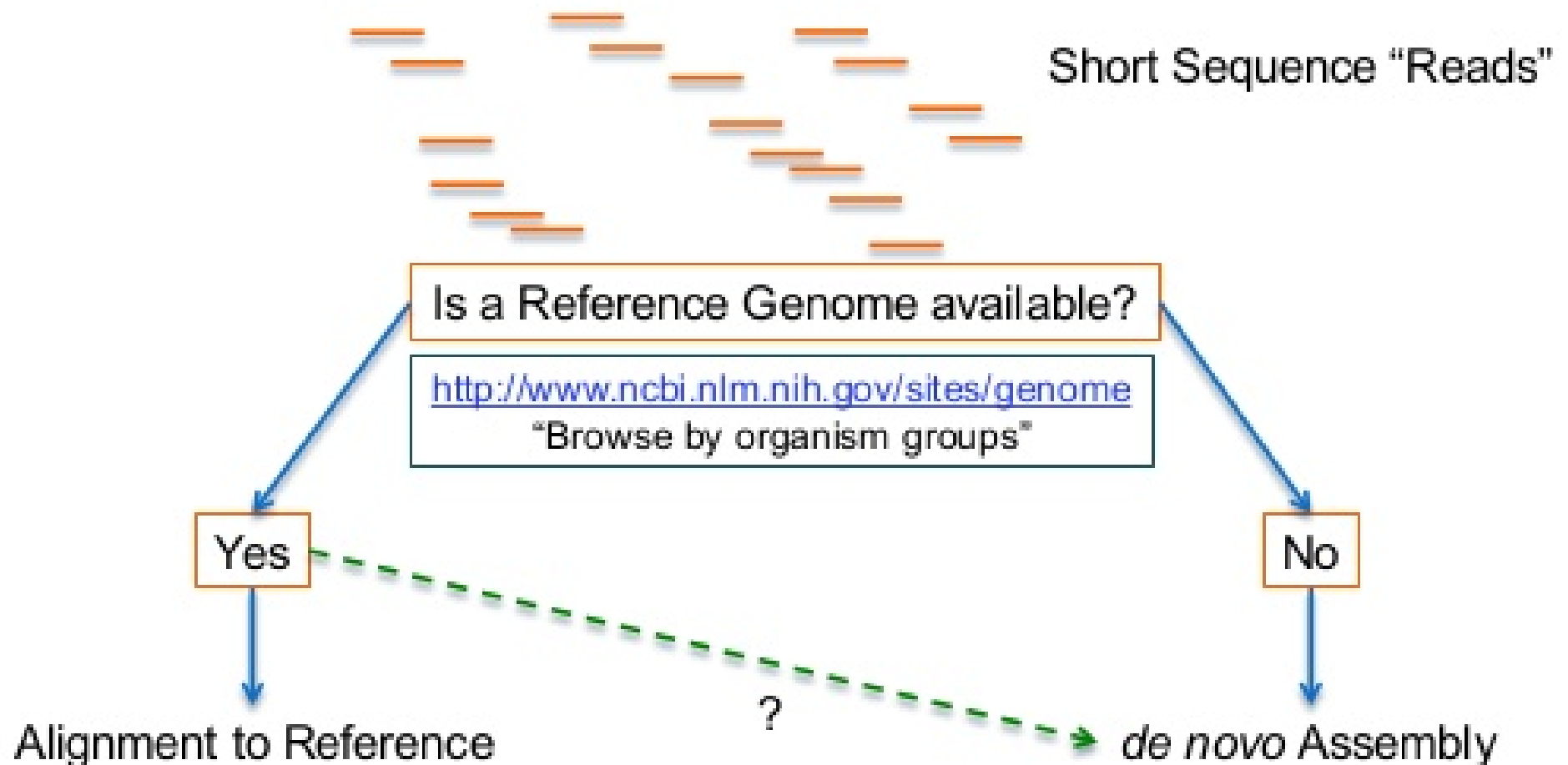
5202 genomes



Map reads to pangenome with Giraffe



Alignment versus *De Novo* Assembly



Sequence file formats

- Next gen sequence file formats are based on the commonly used

FASTA format

>sequence_ID and optional comments

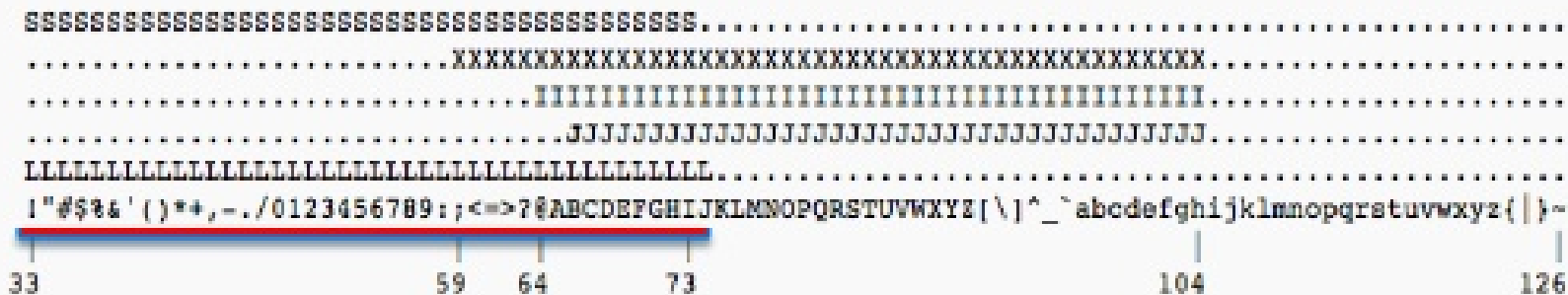
ATTCCGGTGCGGTGCGGTGCTGCCGTGCCGGTGC

TTCGAAATTGGCGTCAGT

- The Phred quality scores per base were added to form the FASTQ format

Fastq Quality values

From http://en.wikipedia.org/wiki/FASTQ_format



- S - Sanger Phred+33, raw reads typically (0, 40)
- X - Solexa Solexa+64, raw reads typically (-5, 40)
- I - Illumina 1.3+ Phred+64, raw reads typically (0, 40)
- J - Illumina 1.5+ Phred+64, raw reads typically (3, 40)
with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
(Note: See discussion above).
- I - Illumina 1.8+ Phred+33,** raw reads typically (0, 41)

$Q = -10 \log_{10} P$

or

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

Phred Quality Score	Probability of incorrect base call
10	1 in 10
20	1 in 100
30	1 in 1000
40	1 in 10000
50	1 in 100000

BBBBCCCC?<A?BC?7@@@??????DBBA@@@A@@@
The highest base quality score in this sequence: 'D'=(68-33)=35

If base quality = 35

$P=10^{-35/10} = 0.00032$ (or 1/3200 incorrect)

Quality scores are normally expected up to 40 in a Phred scale.
ASCII characters <<http://en.wikipedia.org/wiki/ASCII>>

Running FastQC

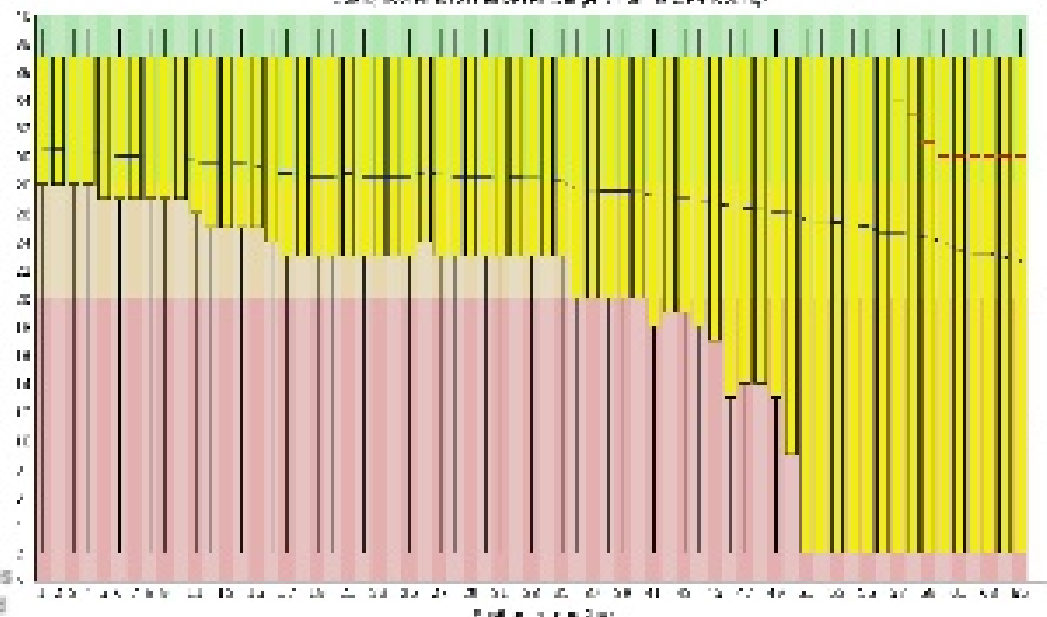
Open FastQC program

Open in browser:
fastqc_report.html



Per base sequence quality

Quality scores are shown as boxes (larger = better quality)



p-value = 0.0001

p-value = 0.001

p-value = 0.01

p-value = 0.05

Short Read Alignment Software

- BFAST
- BLASTN
- BLAT
- **Bowtie**
- **BWA**
- **ELAND**
- GNUMAP
- GMAP and **GSNAP**
- **MAQ**
- mrFAST and mrsFAST
- MOSAIK
- **Novoalign**
- RUM
- SHRiMP
- SOAP
- SpliceMap
- SSAHA and SSAHA2
- STAR
- **TopHat**
- **~20 more...**

```
ls /usr/local/bio_apps/
```

http://en.wikipedia.org/wiki/List_of_sequence_alignment_software

Issues of Consideration for Alignment Software

- Types of reads

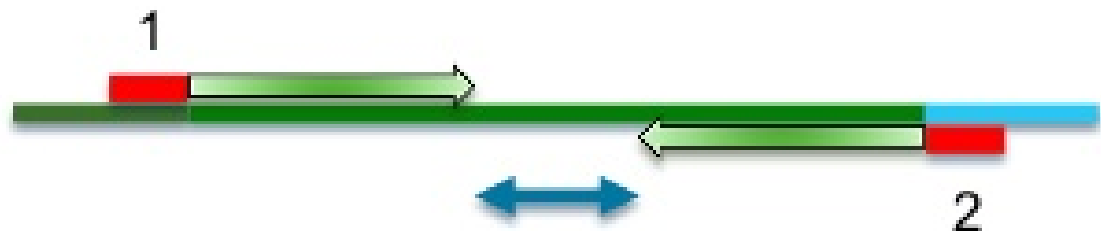
- Single-end

e.g. SRR036642.fastq



- Paired-end

e.g. SRR027894_1.fastq, SRR027894_2.fastq



Mean, Standard Deviation of Inner Distance

Most commonly used alignment file formats

- SAM (sequence alignment map)

Unified format for storing alignments to a reference genome

- BAM (binary version of SAM)

Compressed SAM file, is normally indexed

SAM/BAM format (sequence alignment map): Most commonly used alignment file formats

Unified format for storing alignments to a reference genome
BAM is a compressed SAM file, normally indexed

<http://picard.sourceforge.net/explain-flags.html>

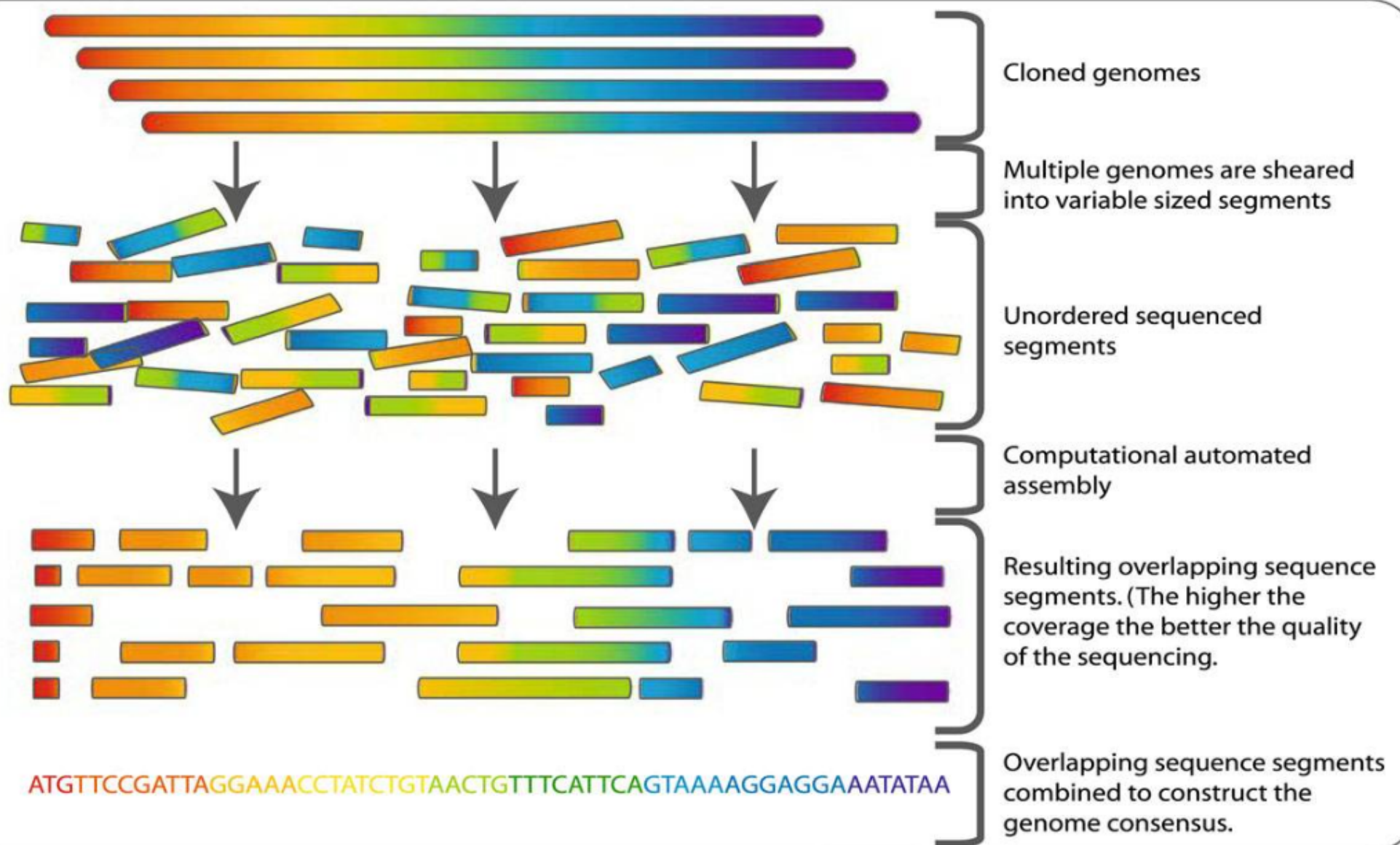
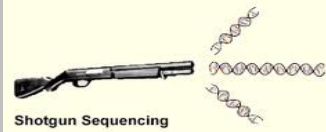
QNAME	FLAG	RNAME	POSITION	MAPQ	CIGAR	MRNM	MPOS	TLEN
ERR004054.7965693	147	1	925615	29	37M	=	925487	-165
ERR004054.9540692	99	1	925777	29	37M	=	925878	138
ERR004054.9540692	147	1	925878	23	37M	=	925777	-138

SEQ	QUAL	OPT
TGATCCGCCCACCTCAGCCTCCCAAAGTGCTGGGATT	?A1@@>:C@C46EAGB@CEAC@F@@BAABDABBD?=:	XT:A:R NM:i:0 SM:i:0
GATTCTCCTGCCTCAGCCTCCCGAGTAGCTGGGACTG	<@?7BADBAFCCBDAEDBBDBB>AE>=E@BF577B@D	XT:A:R NM:i:0 SM:i:0
GGCCAGGCCGGTCTCGAACTCCTGACCTCAGGCGATT	<?:DA@@@7A@AEA>D?>D?@DAC>BCAFDB@=B=>:	XT:A:U NM:i:0 SM:i:23

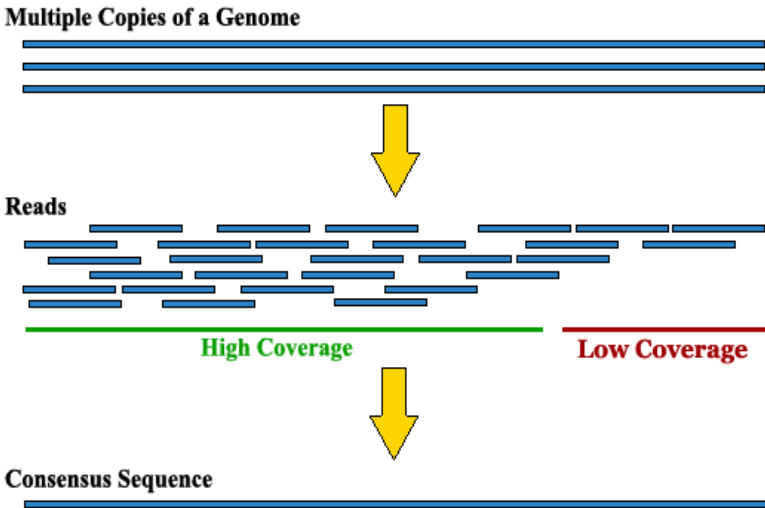
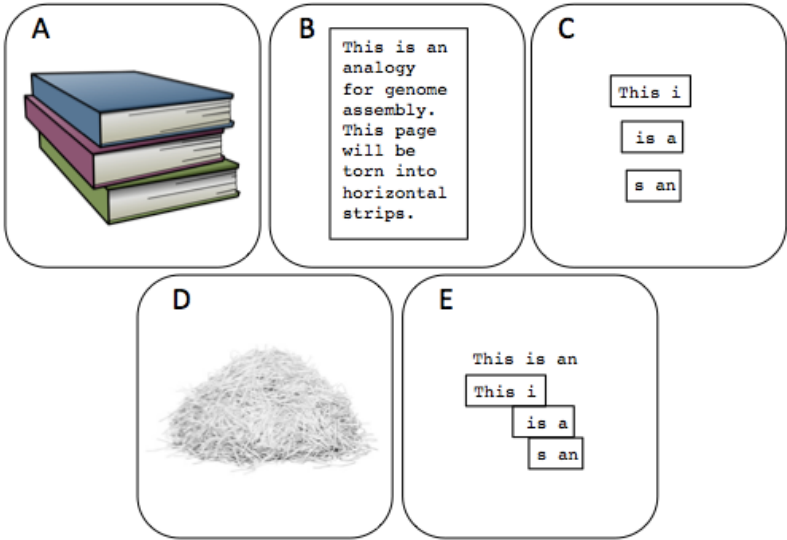
Col	Field	Description
1	QNAME	Query template/pair NAME
2	FLAG	bitwise FLAG
3	RNAME	Reference sequence NAME
4	POS	1-based leftmost POSition/coordinate of clipped sequence
5	MAPQ	MAPPing Quality (Phred-scaled)
6	CIGAR	extended CIGAR string
7	MRNM	Mate Reference sequence Name ('=' if same as RNAME)
8	MPOS	1-based Mate POSition
9	TLEN	Inferred Template LENGTH (insert size)
10	SEQ	query SEquence on the same strand as the reference
11	QUAL	query QUALity (ASCII-33 gives the Phred base quality)
12+	OPT	variable OPTional fields in the format TAG:VTYPE:VALUE

Flag	Chr	Description
0x0001	p	the read is paired in sequencing
0x0002	P	the read is mapped in a proper pair
0x0004	u	the query sequence itself is unmapped
0x0008	U	the mate is unmapped
0x0010	r	strand of the query (1 for reverse)
0x0020	R	strand of the mate
0x0040	1	the read is the first read in a pair
0x0080	2	the read is the second read in a pair
0x0100	s	the alignment is not primary
0x0200	f	the read fails platform/vendor quality checks
0x0400	d	the read is either a PCR or an optical duplicate

Introduction to sequencing



Introduction to sequencing



	ATGGCATTGCAA
	TGGCATTGCAATTTG
	AGATGGTATTG
Reads	GATGGCATTGCAA
	GCATTGCAATTTGAC
	ATGGCATTGCAATTT
	AGATGGTATTGCAATTTG
Consensus	
Sequence	AGATGGCATTGCAATTTGAC

Sequencing platforms

Platform	Illumina MiSeq	Ion Torrent PGM	PacBio RS	Illumina GAIIx	Illumina HiSeq 2000
Instrument Cost*	\$128 K	\$80 K**	\$695 K	\$256 K	\$654 K
Sequence yield per run	1.5-2Gb	20-50 Mb on 314 chip, 100-200 Mb on 316 chip, 1Gb on 318 chip	100 Mb	30Gb	600Gb
Sequencing cost per Gb*	\$502	\$1000 (318 chip)	\$2000	\$148	\$41
Run Time	27 hours***	2 hours	2 hours	10 days	11 days
Reported Accuracy	Mostly > Q30	Mostly Q20	<Q10	Mostly > Q30	Mostly > Q30
Observed Raw Error Rate	0.80 %	1.71 %	12.86 %	0.76 %	0.26 %
Read length	up to 150 bases	~200 bases	Average 1500 bases**** (C1 chemistry)	up to 150 bases	up to 150 bases
Paired reads	Yes	Yes	No	Yes	Yes
Insert size	up to 700 bases	up to 250 bases	up to 10 kb	up to 700 bases	up to 700 bases
Typical DNA requirements	50-1000 ng	100-1000 ng	~1 µg	50-1000 ng	50-1000 ng

* All cost calculations are based on list price quotations obtained from the manufacturer and assume expected sequence yield stated.

** System price including PGM, server, OneTouch and OneTouch ES.

*** Includes two hours of cluster generation.

**** Mean mapped read length includes adapter and reverse strand sequences. Subread lengths, i.e. the individual stretches of sequence originating from the sequenced fragment, are significantly shorter.

Sequencing platforms

Platform	Illumina MiSeq	Ion Torrent PGM	PacBio RS	Illumina GAIIx	Illumina HiSeq 2000
Instrument Cost*	\$128 K	\$80 K**	\$695 K	\$256 K	\$654 K
Sequence yield per run	1.5-2Gb	20-50 Mb on 314 chip, 100-200 Mb on 316 chip, 1Gb on 318 chip	100 Mb	30Gb	600Gb
Sequencing cost per Gb*	\$502	\$1000 (318 chip)	\$2000	\$148	\$41
Run Time	27 hours***	2 hours	2 hours	10 days	11 days
Reported Accuracy	Mostly > Q30	Mostly Q20	<Q10	Mostly > Q30	Mostly > Q30
Observed Raw Error Rate	0.80 %	1.71 %	12.86 %	0.76 %	0.26 %
Read length	up to 150 bases	~200 bases	Average 1500 bases**** (CI chemistry)	up to 150 bases	up to 150 bases
Paired reads	Yes	Yes	No	Yes	Yes
Insert size	up to 700 bases	up to 250 bases	up to 10 kb	up to 700 bases	up to 700 bases
Typical DNA requirements	50-1000 ng	100-1000 ng	~1 µg	50-1000 ng	50-1000 ng

* All cost calculations are based on list price quotations obtained from the manufacturer and assume expected sequence yield stated.

** System price including PGM, server, OneTouch and OneTouch ES.

*** Includes two hours of cluster generation.

**** Mean mapped read length includes adapter and reverse strand sequences. Subread lengths, i.e. the individual stretches of sequence originating from the sequenced fragment, are significantly shorter.

 **GenomeMedicine**

Search for

Home

Articles

Authors

Reviewers

About this journal

My Genome Medicine

Subscriptions

Top

Musings

Musings

The \$1,000 genome, the \$100,000 analysis?

Elaine R Mardis

Highly accessed

Technology used

Generation	Company	Platform	Approx. Read Length (nt)
First	ABI/Life Technologies	3730xl	600 - 1000
Next	Roche/454	Genome Sequencer FLX Titanium	300 - 1000
Next	Illumina	HiSeq 2000	36 - 100
Next	ABI/Life Technologies	5500xl SOLiD System	50 - 75



One Genome – One Large Scaffold – One Week

Microbial Genome Assembly				
Organism	<i>S. pneumoniae</i>	<i>E. coli</i>	<i>T. thermophilus</i>	<i>C. jejuni</i>
Number of Chromosomes	1	1	2	1
Large Scaffolds	1	1	2	1
Genome Size	2.2 Mb	4.6 Mb	2.1 Mb	1.6 Mb
N50 Scaffold Size	2.2 Mb	4.6 Mb	1.9 Mb	1.6 Mb
N50 Contig Size	26.1 kb	57.1 kb	10.5 kb	153.8 kb
Genome Coverage	99.6%	100%	100%	99.3%
Oversampling	25x	15x	33x	33x
Number of GS FLX Runs	1/4	1/4	1/4	1/4
Number of GS Junior Runs	2	2	2	2

Genome Assembly	
Organism	<i>S. cerevisiae</i>
Number of Chromosomes	17
Large Scaffolds	17
Genome Size	12.1 Mb
N50 Scaffold Size	79-1,523 kb
Chromosome Size	86-1,532 kb
N50 Contig Size	53.1 kb
Genome Coverage	99.5%
Oversampling	23x
Number of Runs	1

Technology used

Generation	Company	Platform	Approx. Read Length (nt)
First	ABI/Life Technologies	3730xl	600 - 1000
Next	Roche/454	Genome Sequencer FLX Titanium	300 - 1000
Next	Illumina	HiSeq 2000	36 - 100
Next	ABI/Life Technologies	5500xl SOLiD System	50 - 75



One Genome – One Large Scaffold – One Week

Microbial Genome Assembly				
Organism	<i>S. pneumoniae</i>	<i>E. coli</i>	<i>T. thermophilus</i>	<i>C. jejuni</i>
Number of Chromosomes	1	1	2	1
Large Scaffolds	1	1	2	1
Genome Size	2.2 Mb	4.6 Mb	2.1 Mb	1.6 Mb
N50 Scaffold Size	2.2 Mb	4.6 Mb	1.9 Mb	1.6 Mb
N50 Contig Size	26.1 kb	57.1 kb	10.5 kb	153.8 kb
Genome Coverage	99.6%	100%	100%	99.3%
Oversampling	25x	15x	33x	33x
Number of GS FLX Runs	1/4	1/4	1/4	1/4
Number of GS Junior Runs	2	2	2	2

Genome Assembly	
Organism	<i>S. cerevisiae</i>
Number of Chromosomes	17
Large Scaffolds	17
Genome Size	12.1 Mb
N50 Scaffold Size	79-1,523 kb
Chromosome Size	86-1,532 kb
N50 Contig Size	53.1 kb
Genome Coverage	99.5%
Oversampling	23x
Number of Runs	1



Prokazyme



**Imperial College
London**



Total budget ~ 1 mln €



HU7BLKY01 = Unamplified DNA, pools of 4 old and 4 new phages

Provided fasta file contains 89130 reads

The average size of sequence is 358.55 (min=39, max=1220)
32 Mbps in total (GC content 0.61)
N50 399

HU7BLKY02 = phi amplified DNA pool of 8 old phages

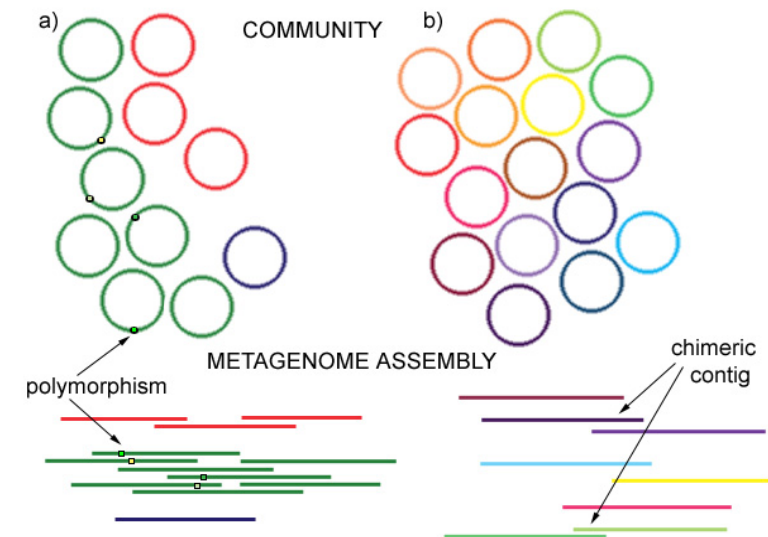
Provided fasta file contains 117186 reads

The average size of sequence is 429.21 (min=39, max=847).
50.3 Mbps in total (GC content 0.44).
N50 482

HU7BLKY03 = phi amplified DNA pool of 8 new phages

Provided fasta file contains 108722 reads

The average size of sequence is 418.04 (min=39, max=859).
45,5 Mbps in total (GC content 0.58).
N50 479



-aim to discover, characterise and bring on the market new thermophilic enzymes:

- **Strand-displacing DNA polymerase**
- **RNA/DNA ligase**
- **Protelomerase**
- **Lysozyme**
- **Nucleases**
- **SSB protein**

- genomes come from Island (hot springs, geysers)

- IIMCB will do bioinformatics (detecting interesting genes, fold recognition, domains, models etc.)

- people from Tricity will do experimental part

- 4 commercial firms will bring the product on the market

Expected outcome:

Firms: new, better products, patents

Academics: publications (after patenting and commercialization of the products)

General strategy of assembling a genome *de novo*

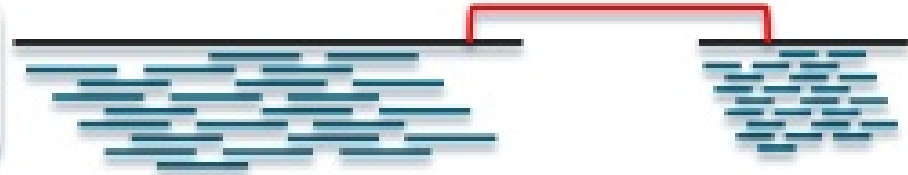
Pre-process short reads
(trim, quality filter...)

```
@SRR1000000.23890 8042AABXX110103:4:1:20154:2830/1
CTTGCTAGGCTATGCCTACACTGGGCTTGTGACATATCTCTGCACTGATCACCAGTTGATATAACTCTTCTCT/
+
HHHHHHHHHHHHHHHHHHHHFFHHFFHHHFBHHEFFFFHHFFHHEHEEHFHDFDDOADDFFECBEDD78EBDDC@DCE+EDF
@SRR1000000.23891 8042AABXX110103:4:1:2263:2831/1
AATGCCTATDGCCECCCAATTCTGCTTGTCCAAAGAGTTGCCAGCCCCAAAACCTCCTCACTTC
+
HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHFFHHFFBHBEB?CEBCC70AA5D0:,<1<+@1-9??
```

Assemble sequences into
contigs

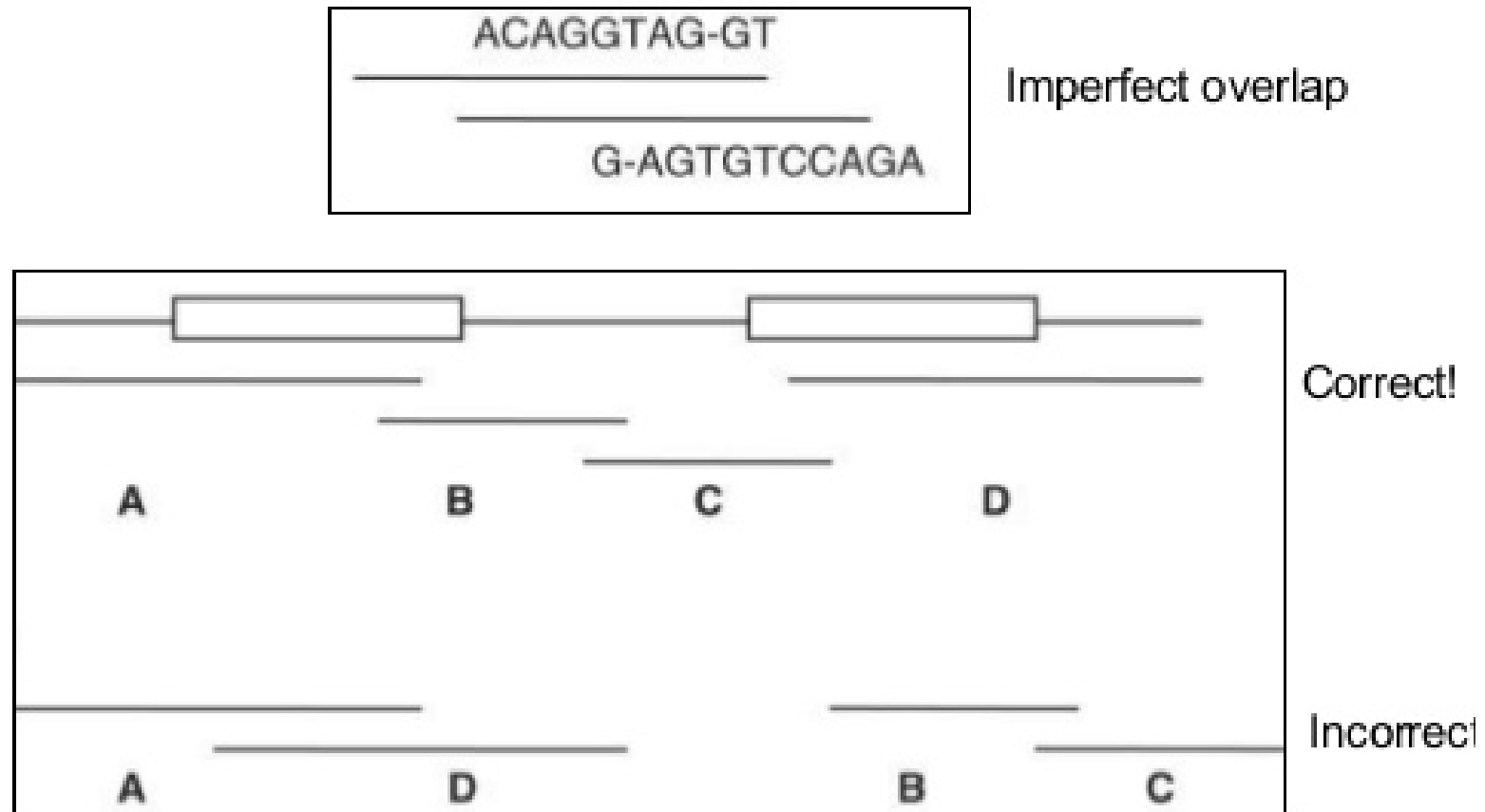


Order contigs into
scaffolds



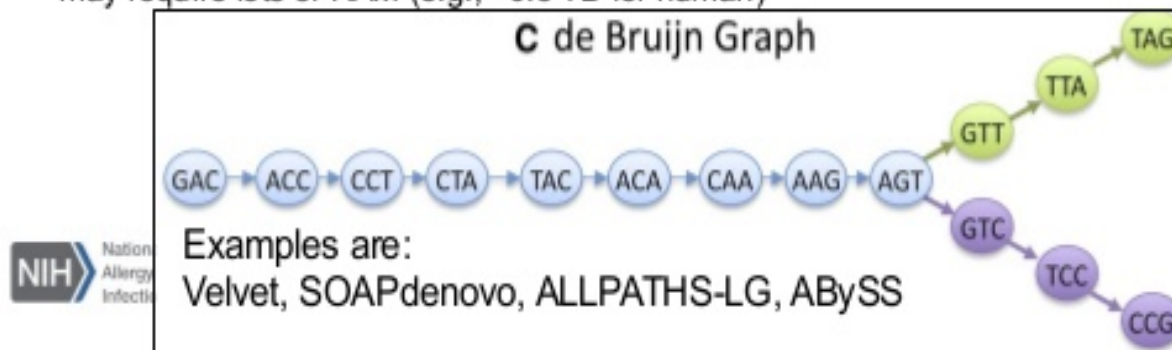
Annotate genome

Imperfect Overlap Between Reads Can Lead to Incorrect Assembly in the Greedy Approach



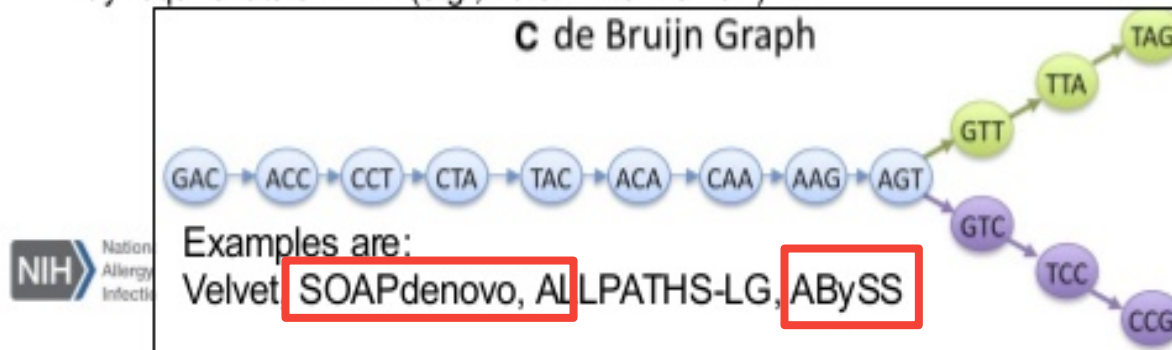
De Bruijn Graph

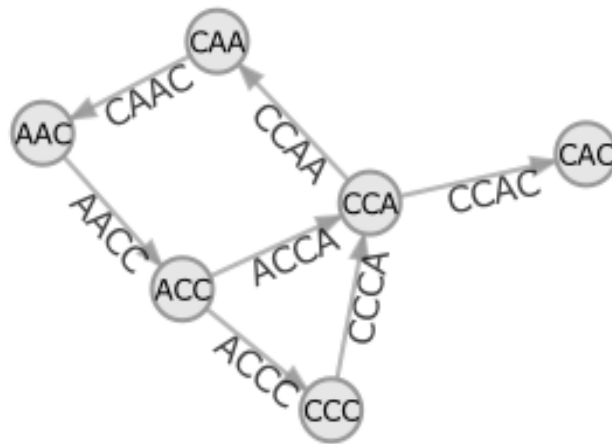
- It breaks reads into **successive** k-mers and the graph maps the k-mers
- Each k-mer is a node and edges are drawn between each k-mer in a read.
- Repeat sequences create a fork in the graph; alternative sequences create a bubble.
- The k-mer size can only be determined by “trial and error”.
- A small value of K will create a complex graph but a large value of K may miss small overlaps. A good starting point would be a k-mer size that is 2/3 the size of the read
- Good for short reads or small genomes. With long reads and/or large genomes, may require lots of RAM (e.g., ~0.5 TB for human)



De Bruijn Graph

- It breaks reads into **successive** k-mers and the graph maps the k-mers
- Each k-mer is a node and edges are drawn between each k-mer in a read.
- Repeat sequences create a fork in the graph; alternative sequences create a bubble.
- The k-mer size can only be determined by “trial and error”.
- A small value of K will create a complex graph but a large value of K may miss small overlaps. A good starting point would be a k-mer size that is 2/3 the size of the read
- Good for short reads or small genomes. With long reads and/or large genomes, may require lots of RAM (e.g., ~0.5 TB for human)





de Bruijn graph with $k(\text{mer}) = 4$

ACCC
CCCA
CCAA
CAAC
AACC
ACCA
CCAC

Reads

ACCCAACCAC

Consensus

For simplicity, reverse complement edges and nodes are ignored.

Edges – unique subsequences of length k within reads

Nodes – common subsequences of length $k - 1$

Edge connects two nodes if the suffix of the source node shares an exact match of length $k - 2$ with the prefix of the destination node,

For example, the edge "CCAA" connects the nodes "CCA" and "CAA" because there is an overlap of "CA"

Evaluating the assembly

- Genome assembly results:
 - contig size and number of contigs produced
 - scaffold size and number
 - N50 and N90
- Coverage
- GC Content
- Genome annotation
 - repeats analysis and annotation
 - protein-coding gene annotation (including gene structure prediction and gene function annotation)
 - non-coding RNA gene annotation (including annotation of microRNA, tRNA, rRNA, and other ncRNA)
 - transposon and tandem repeats annotation
- Comparative genomics and evolution (chromosome structure, conserved gene families)

Evaluating the assembly

Basic statistics

N50 the length of the shortest contig such that the sum of contigs of equal length or longer is at least 50% of the total length of all contigs.

Contig size (bp)

3000

2000 ← N50

1200

800

600 ← N90

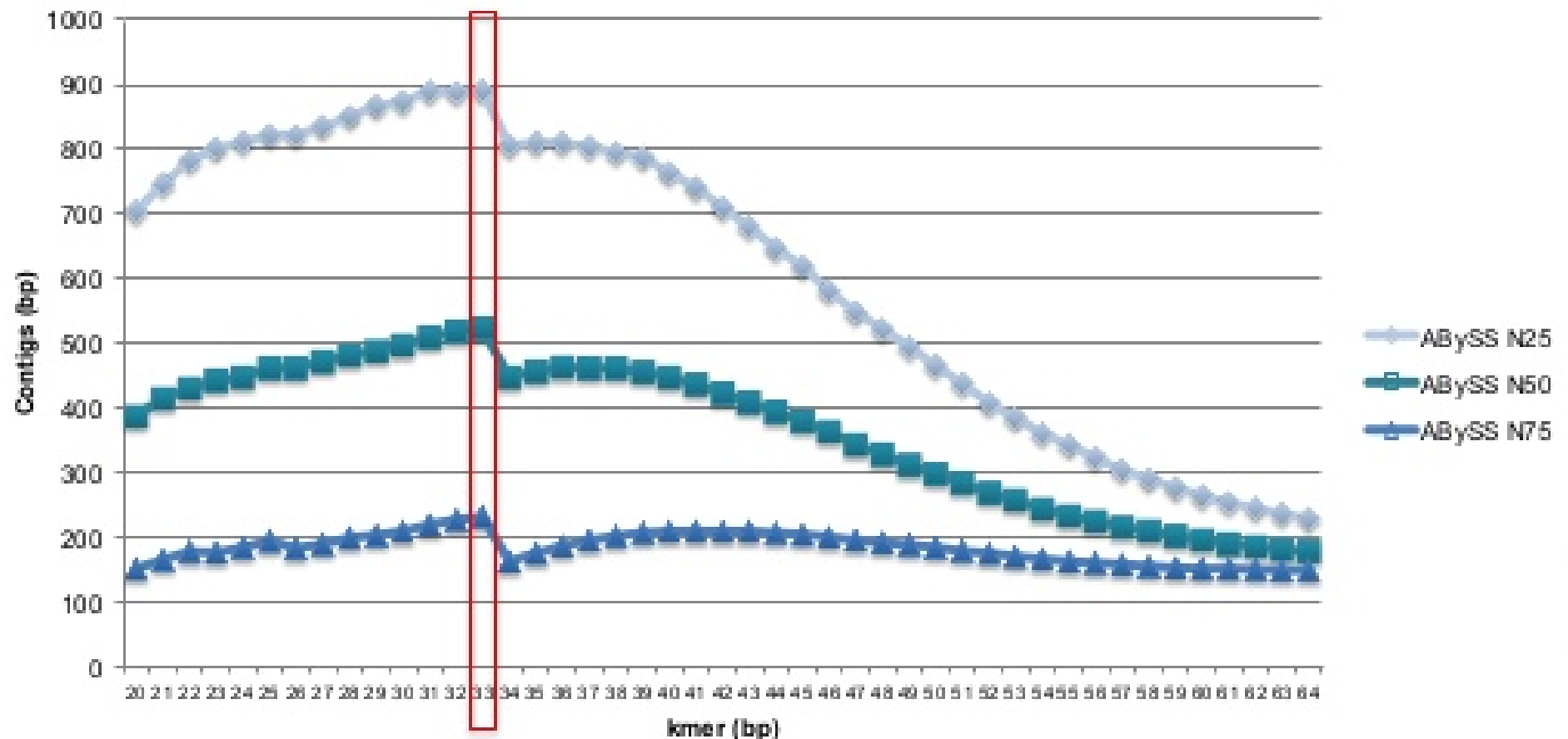
400

Total: **8000**

N90 = the length of the shortest contig such that the sum of contigs of equal length or longer is at least 90% of the total length of all contigs.

To Determine Optimal kmer Size, Try Many

Effect of kmer Length on Contig Length in ABySS



*This will vary based on dataset (genome, read length, etc.)

A Practical Comparison of *De Novo* Genome Assembly Software Tools for Next-Generation Sequencing Technologies

Wenyu Zhang, Jiajia Chen, Yang Yang, Yifei Tang, Jing Shang, Bairong Shen*

A Practical Comparison of *De Novo* Genome Software Tools for Next-Generation Sequencing Technology

arXiv.org > q-bio > arXiv:1301.5406

Wenyu Zhang, Jia

Quantitative Biology > Genomics



Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species

Keith R. Bradnam (1), Joseph N. Fass (1), Anton Alexandrov (36), Paul Baranay (2), Michael Bechner (39), İnanç Birol (33), Sébastien Boisvert¹⁰, (11), Jarrod A. Chapman (20), Guillaume Chapuis (7,9), Rayan Chikhi (7,9), Hamidreza Chitsaz (6), Wen-Chi Chou (14,16), Jacques Corbeil (10,13), Cristian Del Fabbro (17), T. Roderick Docking (33), Richard Durbin (34), Dent Earl (40), Scott Emrich (3), Pavel Fedotov (36), Nuno A. Fonseca (30,35), Ganeshkumar Ganapathy (38), Richard A. Gibbs (32), Sante Gnerre (22), Élénie Godzaridis (11), Steve Goldstein (39), Matthias Haimel (30), Giles Hall (22), David Haussler (40), Joseph B. Hiatt (41), Isaac Y. Ho (20), Jason Howard (38), Martin Hunt (34), Shaun D. Jackman (33), David B Jaffe (22), Erich Jarvis (38), Huaiyang Jiang (32), et al. (55 additional authors not shown)

(Submitted on 23 Jan 2013)

Background - The process of generating raw genome sequence data continues to become cheaper, faster, and more accurate. However, assembly of such data into high-quality, finished genome sequences remains challenging. Many genome assembly tools are available, but they differ greatly in terms of their performance (speed, scalability, hardware requirements, acceptance of newer read technologies) and in their final output (composition of assembled sequence). More importantly, it remains largely unclear how to best assess the quality of assembled genome sequences. The Assemblathon competitions are intended to assess current state-of-the-art methods in genome assembly. **Results** - In Assemblathon 2, we provided a variety of sequence data to be assembled for three vertebrate species (a bird, a fish, and snake). This resulted in a total of 43 submitted assemblies from 21 participating teams. We evaluated these assemblies using a combination of optical map data, Fosmid sequences, and several statistical methods. From over 100 different metrics, we chose ten key measures by which to assess the overall quality of the assemblies. **Conclusions** - Many current genome assemblers produced useful assemblies, containing a significant representation of their genes, regulatory sequences, and overall genome structure. However, the high degree of variability between the entries suggests that there is still much room for improvement in the field of genome assembly and that approaches which work well in assembling the genome of one species may not necessarily work well for another.

A Practical Comparison of *De Novo* Genome Software Tools for Next-Generation Sequencing Technology

arXiv.org > q-bio > arXiv:1301.5406

Wenyu Zhang, Jia Quantitative Biology > Genomics



Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species

Keith R. Bradnam (1), Joseph N. Fass (1), Anton Alexandrov (36), Paul Baranay (2), Michael Bechner (39), İnanç Birol (33), Sébastien Boisvert¹⁰, (11), Jarrod A. Chapman (20), Guillaume Chapuis (7,9), Rayan Chikhi (7,9), Hamidreza Chitsaz (6), Wen-Chi Chou (14,16), Jacques Corbeil (10,13), Cristian Del Fabbro (17), T. Roderick Docking (33), Richard Durbin (34), Dent Earl (40), Scott Emrich (3), Pavel Fedotov (36), Nuno A. Fonseca (30,35), Ganeshkumar Ganapathy (38), Richard A. Gibbs (32), Sante Gnerre (22), Élénie Godzaridis (11), Steve Goldstein (39), Matthias Haimel (30), Giles Hall (22), David Haussler (40), Joseph B. Hiatt (41), Isaac Y. Ho (20), Jason Howard (38), Martin Hunt (34), Shaun D. Jackman (33), David B Jaffe (22), Erich Jarvis (38), Huaiyang Jiang (32), et al. (55 additional authors not shown)

(Submitted on 23 Jan 2013)

Background - The process of generating raw genome sequence data continues to become cheaper, faster, and more accurate. However, assembly of such data into high-quality, finished genome sequences remains challenging. Many genome assembly tools are available, but they differ greatly in terms of their performance (speed, scalability, hardware requirements, acceptance of newer read technologies) and in their final output (composition of assembled sequence). More importantly, it remains largely unclear how to best assess the quality of assembled genome sequences. The Assemblathon competitions are intended to assess current state-of-the-art methods in genome assembly. **Results** - In Assemblathon 2, we provided a variety of sequence data to be assembled for three vertebrate species (a bird, a fish, and snake). This resulted in a total of 43 submitted assemblies from 21 participating teams. We evaluated these assemblies using a combination of optical map data, Fosmid sequences, and several statistical methods. From over 100 different metrics, we chose ten key measures by which to assess the overall quality of the assemblies. **Conclusions** - Many current genome assemblers produced useful assemblies, containing a significant representation of their genes, regulatory sequences, and overall genome structure. However, the high degree of variability between the entries suggests that there is still much room for improvement in the field of genome assembly and that approaches which work well in assembling the genome of one species may not necessarily work well for another.



<http://cnag.bsc.es/>

de novo genome assembly (dnGASP) using data from 2nd generation sequencing platforms.

A sister project dubbed RGASP3 (the third incarnation of the RNA-Seq Genome Annotation Assessment Project) is focused on evaluating RNASeq read alignment

A Practical
Software
Technique

Wenyu Zhang

Assembler	Contigs			Scaffolds		
	Num	N50 (kb)	E-size (kb)	Num	N50 (kb)	E-size (kb)
ABYSS	40,451	14.4	20.9	34,298	18	27
Allpaths-LG	Could not run: incompatible library types					
CABOG	22,107	23.5	34.2	1,191	1,125	1,367
MSR-CA	21,885	32.4	46.9	2,551	1,246	1,528
SGA	Program crashed: cause unclear					
SOAPdenovo	15,957	57.1	78.2	5,800	1,374	1,608
Velvet	Program crashed: insufficient memory (256 GB)					

...necessarily work well for another.



), Guillaume (34), Dent ve Goldstein affe (22),

me
rer read

d for three
p data,
current
bility
/ not



<http://cnag.bsc.es/>

de novo genome assembly (dnGASP) using data from 2nd generation sequencing platforms.

A sister project dubbed RGASP3 (the third incarnation of the RNA-Seq Genome Annotation Assessment Project) is focused on evaluating RNASeq read alignment

A Practical
Software
Technique

Wenyu Zhang

Assembler	Contigs			Scaffolds		
	Num	N50 (kb)	E-size (kb)	Num	N50 (kb)	E-size (kb)
ABYSS	40,451	14.4	20.9	34,298	18	27
Allpaths-LG	Could not run: incompatible library types					
CABOG	22,107	23.5	34.2	1,191	1,125	1,367
MSR-CA	21,885	32.4	46.9	2,551	1,246	1,528
SGA	Program crashed: cause unclear					
SOAPdenovo	15,957	57.1	78.2	5,800	1,374	1,608
Velvet	Program crashed: insufficient memory (256 GB)					

...necessarily work well for another.



<http://cnag.bsc.es/>

de novo genome assembly (dnGASP) using data from 2nd generation sequencing platforms.

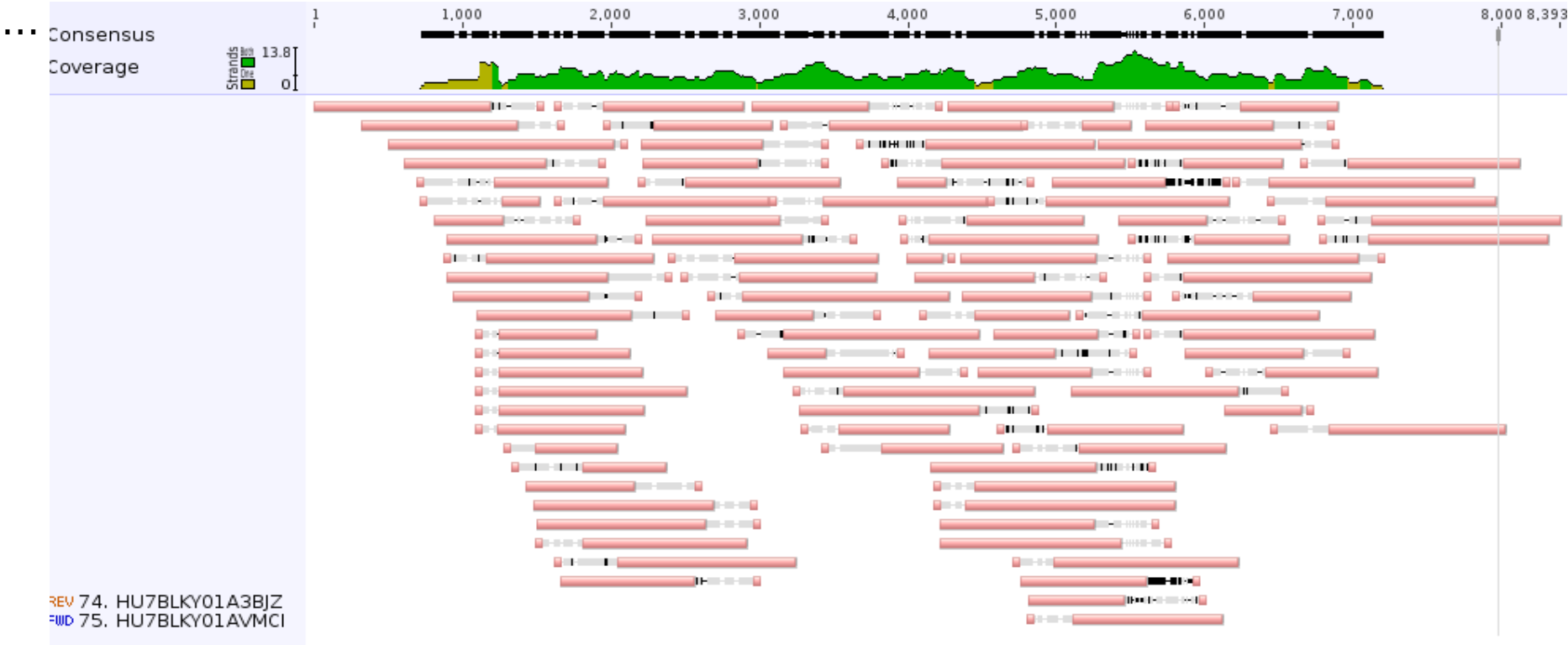
A sister project dubbed RGASP3 (the third incarnation of the RNA-Seq Genome Annotation Assessment Project) is focused on evaluating RNASeq read alignment

), Guillaume (34), Dent ve Goldstein affe (22),
me
rer read
d for three p data, current bility / not

Finally – genome assembly results

- 1. Velvet
- 2. Edena
- 3. Abyss
- 4. AMOS
- 5. Ray
- 6. Geneius©
- 7. SOAPdenovo
- 8. CABOG
- 9. SGA

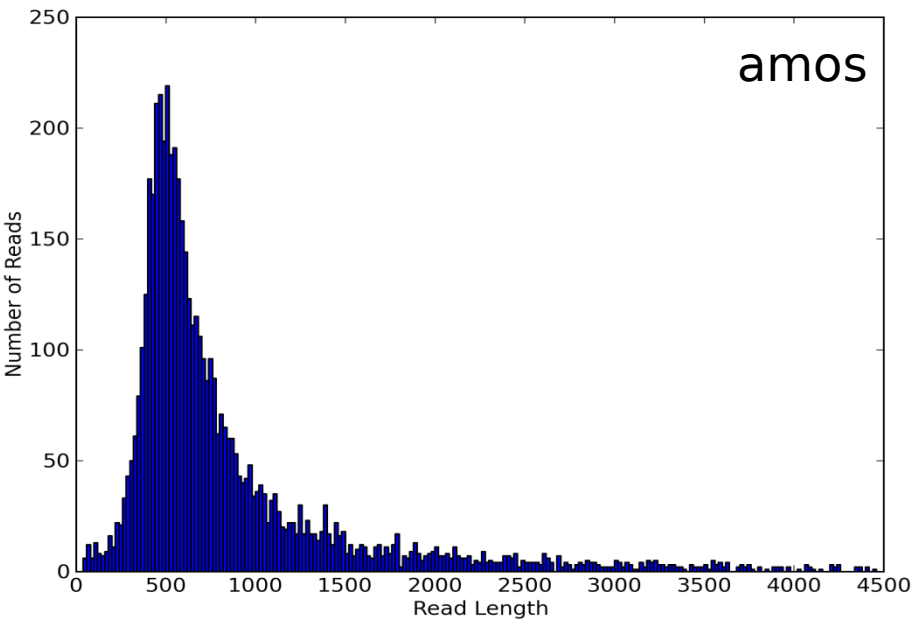
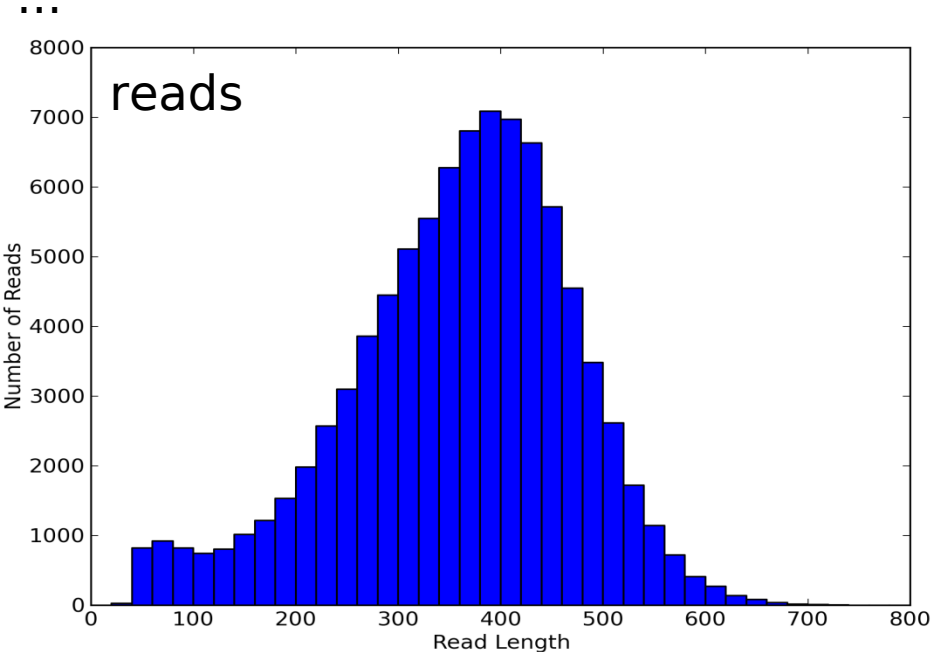
HU7BLKY01 4 old + 4 new genomes					Global average <u>contig</u> depth: 6.22		
Program	Optimal k-mer length	N50	Contig numbers	Avg <u>contig</u> size	Max length	Proteins > 60 aa (stop <u>splitted</u>)	Proteins > 60 aa (CDS)
ABYSS	89 (21-95)	502	9224	368	5737	29420	13861
AMOS	<u>minimus</u>	1127	5342	946	25172	49832	18534
Ray	27 (21-91)	403	9623	324	74610	28033	14250
<u>Geneious©</u>	SW	2232	3457	1344	79370	46666	12488



Finally – genome assembly results

- 1. Velvet
- 2. Edena
- 3. Abyss
- 4. AMOS
- 5. Ray
- 6. Geneius©
- 7. SOAPdenovo
- 8. CABOG
- 9. SGA

HU7BLKY01 4 old + 4 new genomes					Global average <u>contig depth</u> : 6.22		
Program	Optimal k-mer length	N50	<u>Contig numbers</u>	<u>Avg contig size</u>	Max length	Proteins > 60 aa (<u>stop splitted</u>)	Proteins > 60 aa (CDS)
ABYSS	89 (21-95)	502	9224	368	5737	29420	13861
AMOS	<u>minimus</u>	1127	5342	946	25172	49832	18534
Ray	27 (21-91)	403	9623	324	74610	28033	14250
<u>Geneious©</u>	SW	2232	3457	1344	79370	46666	12488

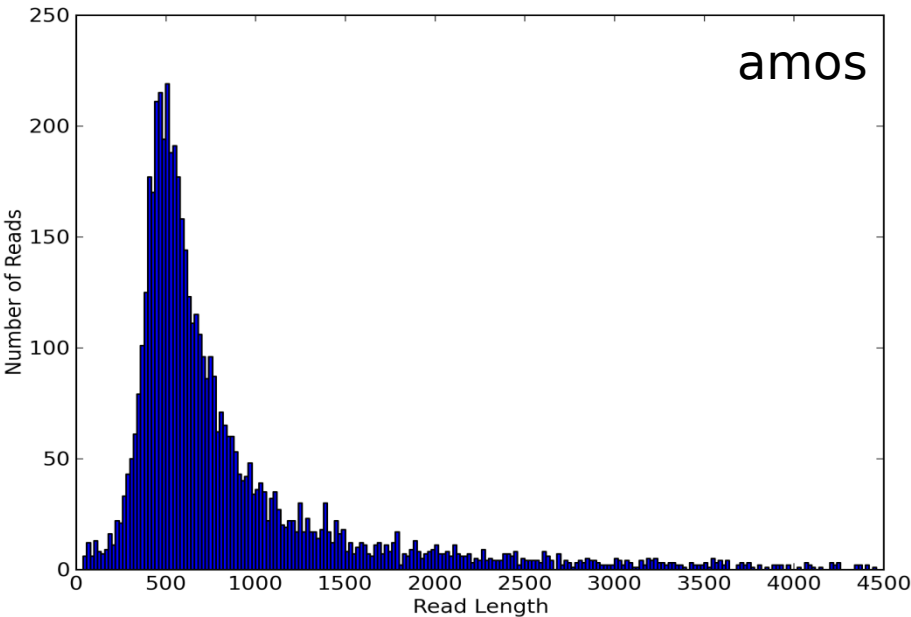
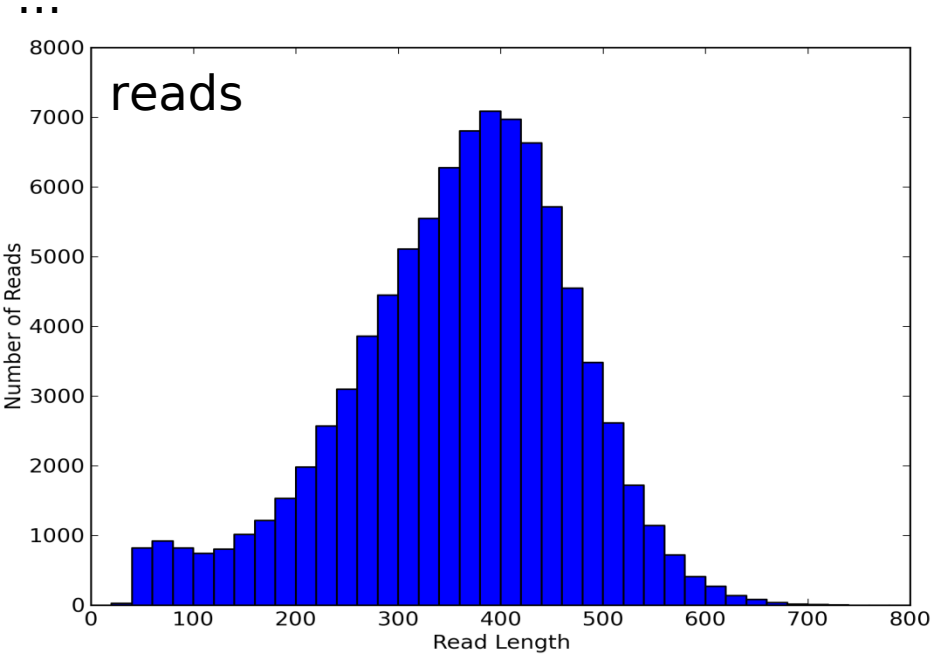


reads vs. contigs

Finally – genome assembly results

- 1. Velvet
- 2. Edena
- 3. Abyss
- 4. AMOS
- 5. Ray
- 6. Geneius©
- 7. SOAPdenovo
- 8. CABOG
- 9. SGA

HU7BLKY01 4 old + 4 new genomes						Global average contig depth: 6.22	
Program	Optimal k-mer length	N50	Contig numbers	Avg contig size	Max length	Proteins > 60 aa (stop splitted)	Proteins > 60 aa (CDS)
ABYSS	89 (21-95)	502	9224	368	5737	29420	13861
AMOS	<u>minimus</u>	1127	5342	946	25172	49832	18534
Ray	27 (21-91)	403	9623	324	74610	28033	14250
<u>Geneious©</u>	SW	2232	3457	1344	79370	46666	12488



reads vs. contigs

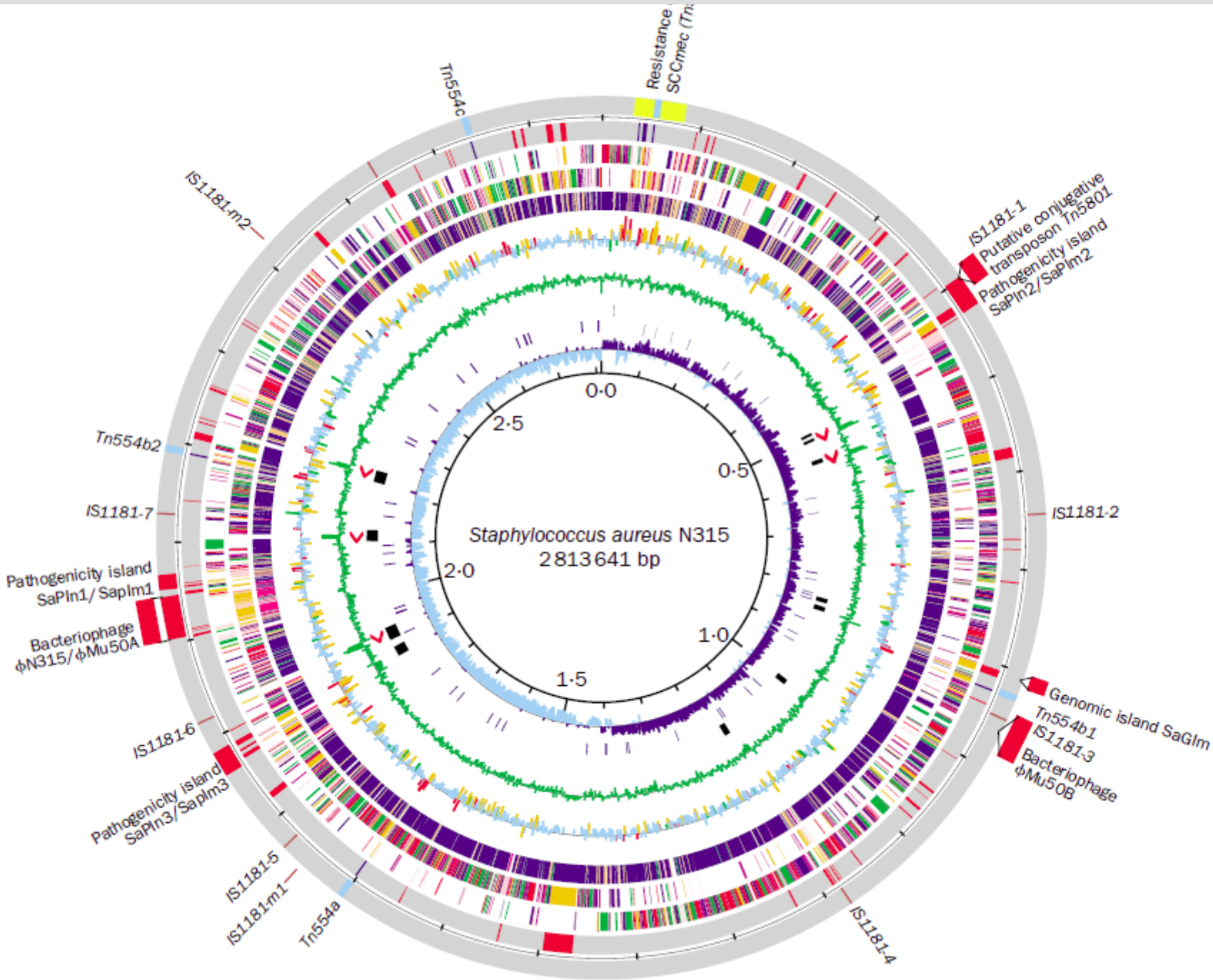
What matters for genome assembly?

1. Assembly program (try to use at least 2-3 with different options)
2. Read length (the longer the better)
3. Coverage (~20-30 is a must)

New putative genes were detected

Strand-displacing DNA polymerase	62
RNAseH	38
Nucleases	69
Helicase	9
RecA	33
PGPR	1

... will be investigated in Gdansk by our co-workers

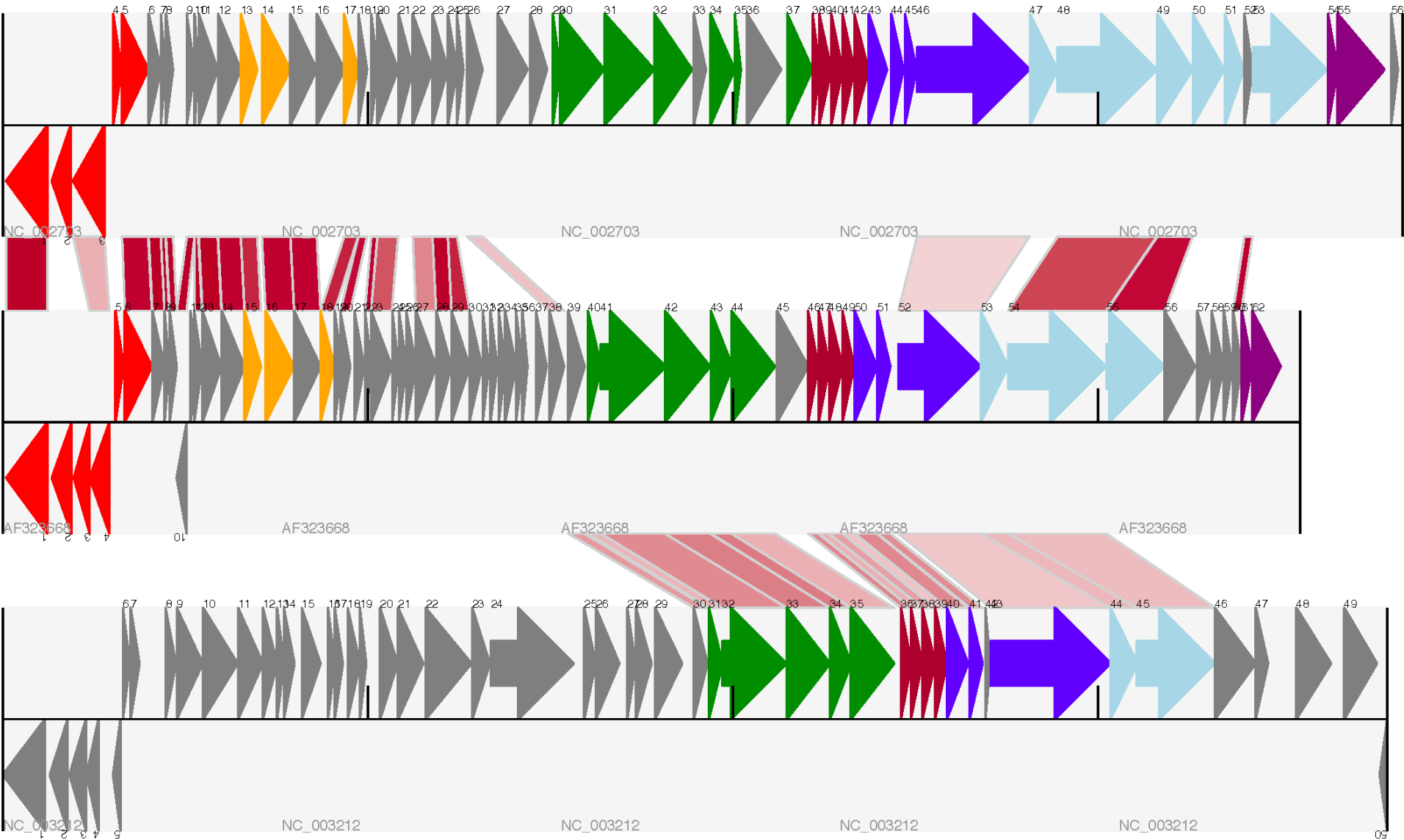




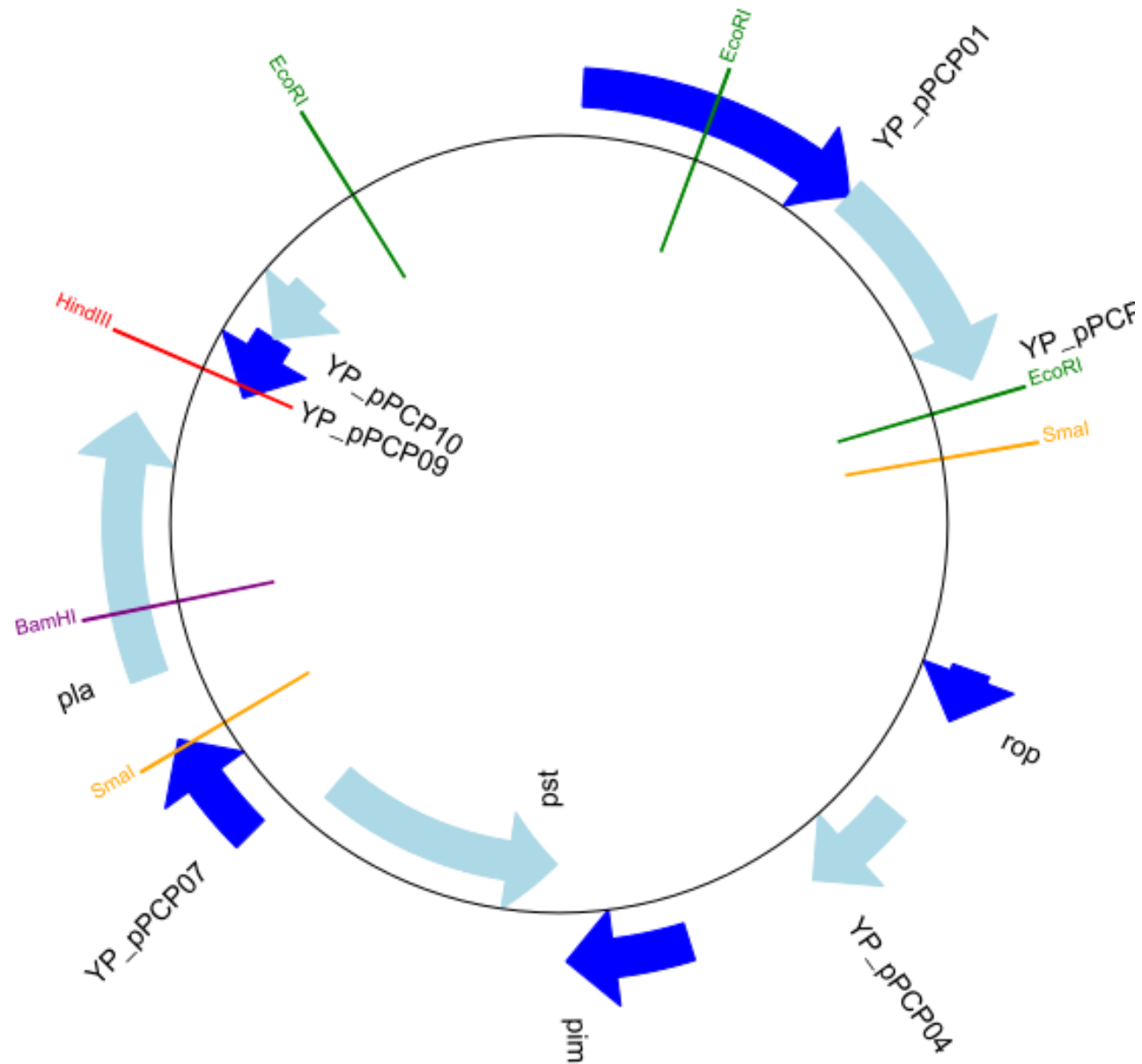
GenomeDiagram

**[http://biopython.org/DIST/docs/tutorial/
Tutorial.html#sec336](http://biopython.org/DIST/docs/tutorial/Tutorial.html#sec336)**

GenomeDiagram



GenomeDiagram







UNIVERSITY OF CALIFORNIA
SANTA CRUZ Genomics
Institute



Genome Browser

[Home](#) [Genomes](#) [Genome Browser](#) [Tools](#) [Mirrors](#) [Downloads](#) [My Data](#) [Projects](#) [Help](#) [About Us](#)



Our tools

- **Genome Browser**
interactively visualize genomic data
- **COVID-19 Research**
use the SARS-CoV-2 genome browser and explore coronavirus datasets
- **BLAT**
rapidly align sequences to the genome
- **Table Browser**
download data from the Genome Browser database
- **Variant Annotation Integrator**
get functional effect predictions for variant calls
- **Data Integrator**
combine data sources from the Genome Browser database
- **Genome Browser in a Box (GBiB)**
run the Genome Browser on your laptop or server
- **In-Silico PCR**
rapidly align PCR primer pairs to the genome
- **LiftOver**
convert genome coordinates between assemblies
- **Track Hubs**
import and view external data tracks
- **REST API**
returns data in JSON format

[More tools...](#)

<https://genome.ucsc.edu/>




Genome Browser Gateway

[Home](#)
[Genomes](#)
[Genome Browser](#)
[Tools](#)
[Mirrors](#)
[Downloads](#)
[My Data](#)
[Projects](#)
[Help](#)
[About Us](#)

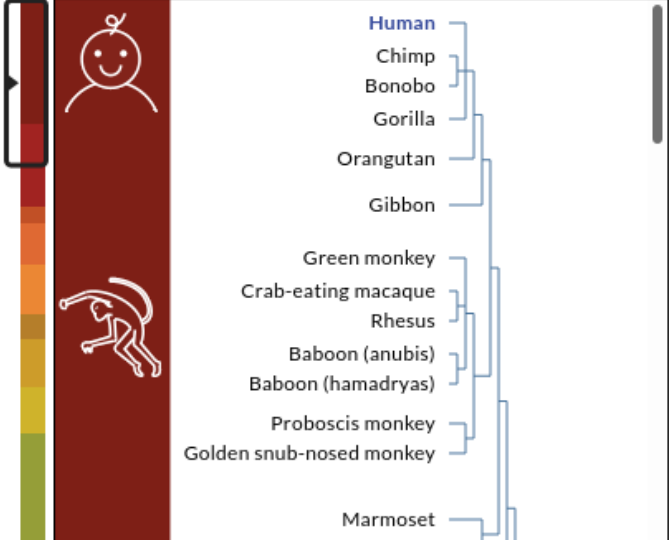
Browse/Select Species

POPULAR SPECIES

 Human
  Mouse
  Rat
  Zebrafish
  Fruitfly
  Worm
  Yeast

[Can't find a genome assembly?](#)

REPRESENTED SPECIES



- Human
- Chimp
- Bonobo
- Gorilla
- Orangutan
- Gibbon
- Green monkey
- Crab-eating macaque
- Rhesus
- Baboon (anubis)
- Baboon (hamadryas)
- Proboscis monkey
- Golden snub-nosed monkey
- Marmoset

Find Position

Human Assembly

Dec. 2013 (GRCh38/hg38)

Position/Search Term

Enter position, gene symbol or search terms

Current position: chrX:15,560,138-15,602,945 [🔗](#)

GO 

Human Genome Browser - hg38 assembly [view sequences](#)

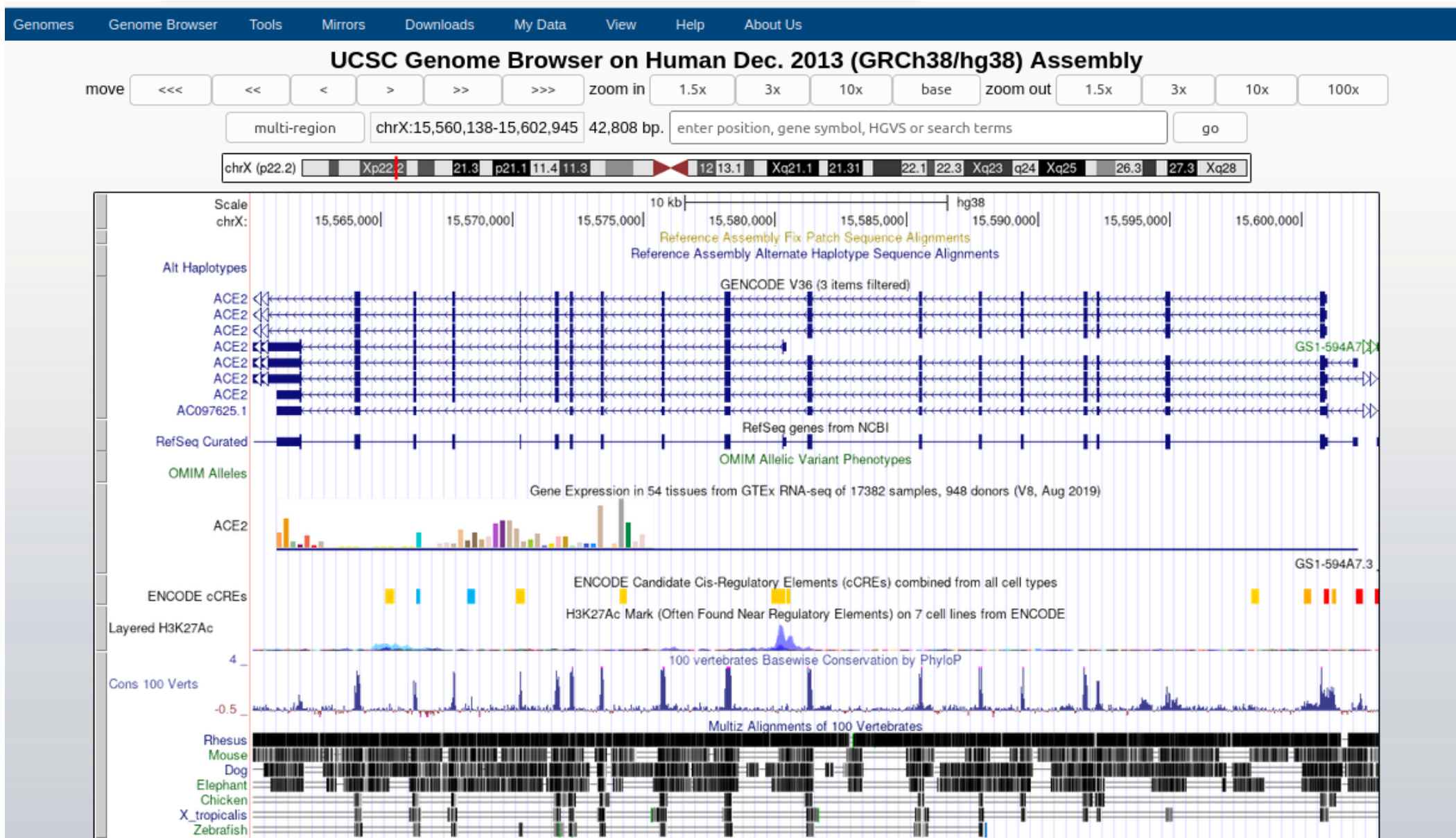
UCSC Genome Browser assembly ID: hg38
 Sequencing/Assembly provider ID: Genome Reference Consortium Human GRCh38.p12 (GCA_000001405.27)
 Assembly date: Dec. 2013 initial release; Dec. 2017 patch release 12
 Assembly accession: [GCA_000001405.27](#)
 NCBI Genome ID: [51](#) (Homo sapiens (human))
 NCBI Assembly ID: [5800238](#) (GRCh38.p12, GCA_000001405.27)
 BioProject ID: [PRJNA31257](#)

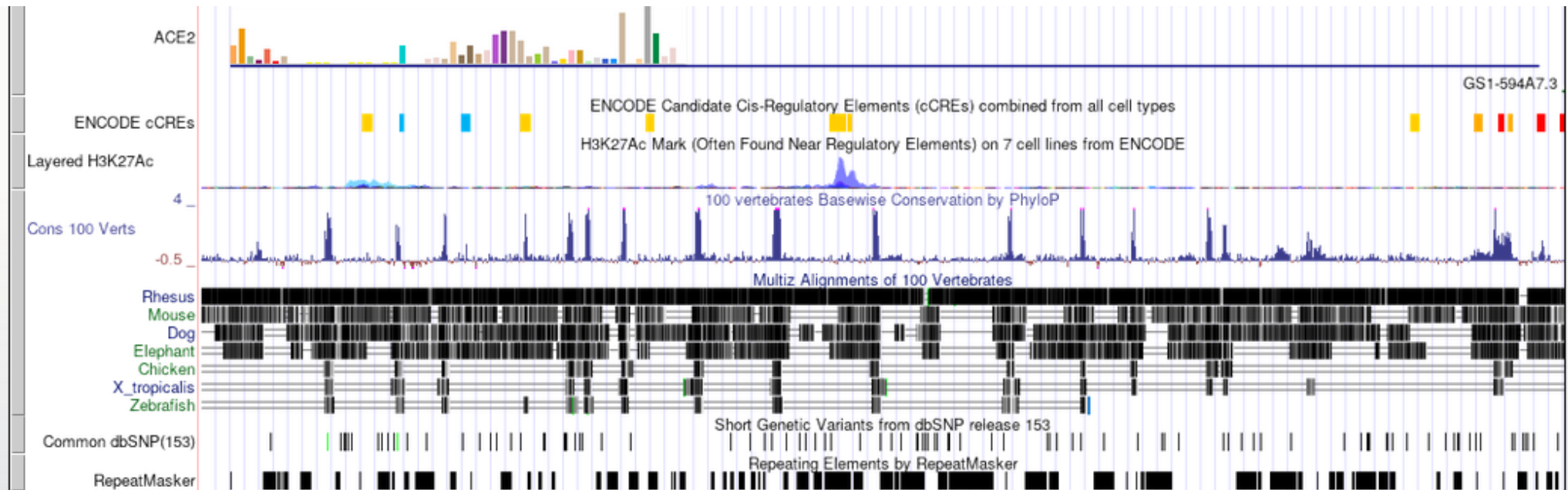
Search the assembly:

- **By position or search term:** Use the "position or search term" box to find areas of the genome associated with many different attributes, such as a specific chromosomal coordinate range; mRNA, EST, or STS marker names; or keywords from the GenBank description of an mRNA. [More information](#), including sample queries.
- **By gene name:** Type a gene name into the "search term" box, choose your gene from the drop-down list, then press "submit" to go directly to the assembly location associated with that gene. [More information](#)



Homo sapiens
(Graphic courtesy of CBSE)





move start move end

Click on a feature for details. Click+shift+drag to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position. Press "?" for keyboard shortcuts.

track search default tracks default order hide all add custom tracks track hubs configure reverse resize refresh

collapse all

Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.

expand all

Mapping and Sequencing

refresh

Base Position

dense

P12 Fix Patches

pack

P12 Alt Haplotypes

dense

P12 Assembly

hide

Centromeres

hide

P12 Chromosome Band

hide

Clone Ends

hide

Exome Probesets

hide

18 FISH Clones

hide

P12 Gap

hide

P12 GC Percent

hide

GRC Incident

hide

Hg19 Diff

hide

Hg19 Mapping

hide

P12 INSDC

hide

LRG Regions

hide

Mappability...

hide

P12 RefSeq Acc

hide

Restr Enzymes

hide

Scaffolds

hide

Short Match

hide

STS Markers

hide

Genes and Gene Predictions

refresh

Regulation

refresh

ENCORE

cCREs

dense

P12 ENCODE

Regulation...

show

P12 CpG Islands...

hide

GeneHancer

hide

Hi-C and Micro-C

hide

ORegAnno

hide

RefSeq Func

hide

Comparative Genomics

refresh

Conservation

Full

Cons 30 Primates

hide

Primate Chain/Net

hide

Placental Chain/Net

hide

Vertebrate Chain/Net

hide

Variation

refresh

dbSNP 153

pack

All SNPs(141)

hide

All SNPs(142)

hide

Common SNPs(142)

hide

Common SNPs(144)

hide

Flagged SNPs(142)

hide

Flagged SNPs(144)

hide

Mult. SNPs(142)

hide

Mult. SNPs(144)

hide

1000 Genomes Trios

hide

1000G Ph3 Vars

hide

All SNPs(144)

hide

All SNPs(146)

hide

All SNPs(147)

hide

All SNPs(150)

hide

All SNPs(151)

hide

Common SNPs(141)

hide

Common SNPs(146)

hide

Common SNPs(147)

hide

Common SNPs(150)

hide

Common SNPs(151)

hide

dbVar Common Struct Var...

hide

DGV Struct Var

hide

Flagged SNPs(141)

hide

Flagged SNPs(146)

hide

Flagged SNPs(147)

hide

Flagged SNPs(150)

hide

Flagged SNPs(151)

hide

Genome In a Bottle

hide

gnomAD Variants...

hide

Mult. SNPs(141)

hide

Mult. SNPs(146)

hide

Mult. SNPs(147)

hide

Mult. SNPs(150)

hide

Mult. SNPs(151)

hide

Platinum Genomes

hide

Repeats

refresh

P12 RepeatMasker

dense

Interrupted Rpts

hide

Microsatellite

hide

RepeatMasker Viz.

hide

Segmental Dups

hide

Self Chain

hide

P12 Simple Repeats

hide

P12 WM + SDust

hide

refresh

<https://genome.ucsc.edu/>

NCBI National Center for Biotechnology Information

Genome Data Viewer

GDV supports the exploration and analysis of *NCBI-annotated* and selected non-NCBI annotated eukaryotic genome assemblies. Currently, over 1220 assemblies are available.

Switch view

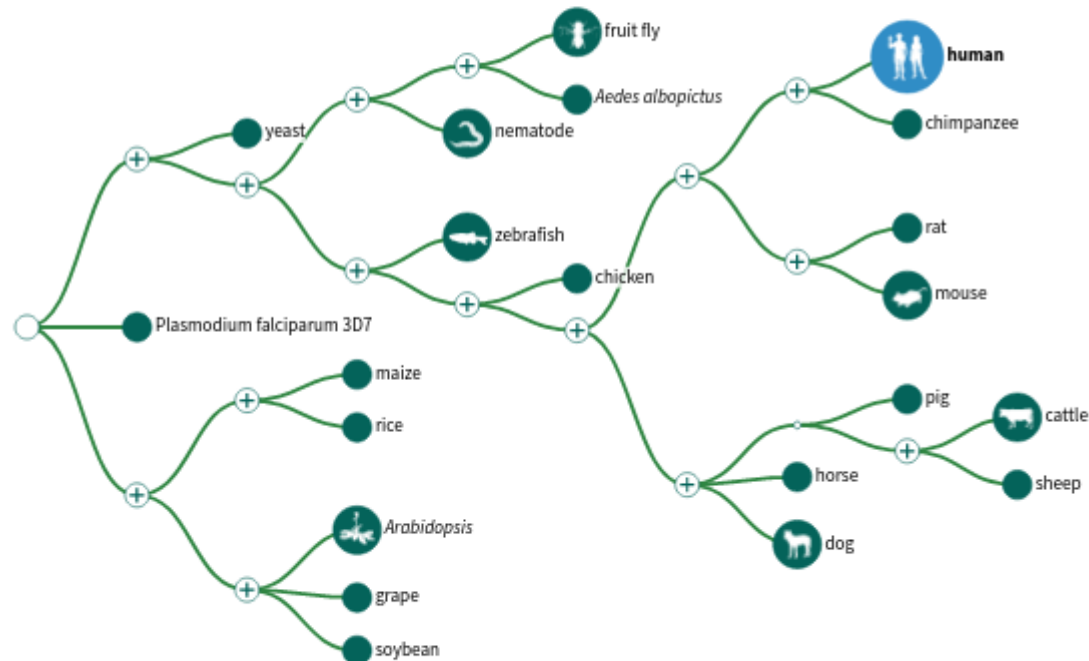


Search organisms

Homo sapiens (human)

To view more organisms in the tree, click on nodes that have '+' signs. Press and hold the '+' to expand and reveal all the subgroups. Or, search for an organism using the search box above.

New! Click on Switch view at the top to see another way of navigating genomes.



Homo sapiens (human)



Search in genome

Location, gene or phenotype



Examples: TXLNA, chr1:32178000-32200000, DNA repair

Assembly

GRCh38.p13

Browse genome

BLAST genome

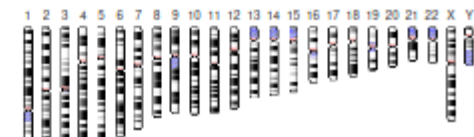
Download via NCBI Datasets

Assembly details

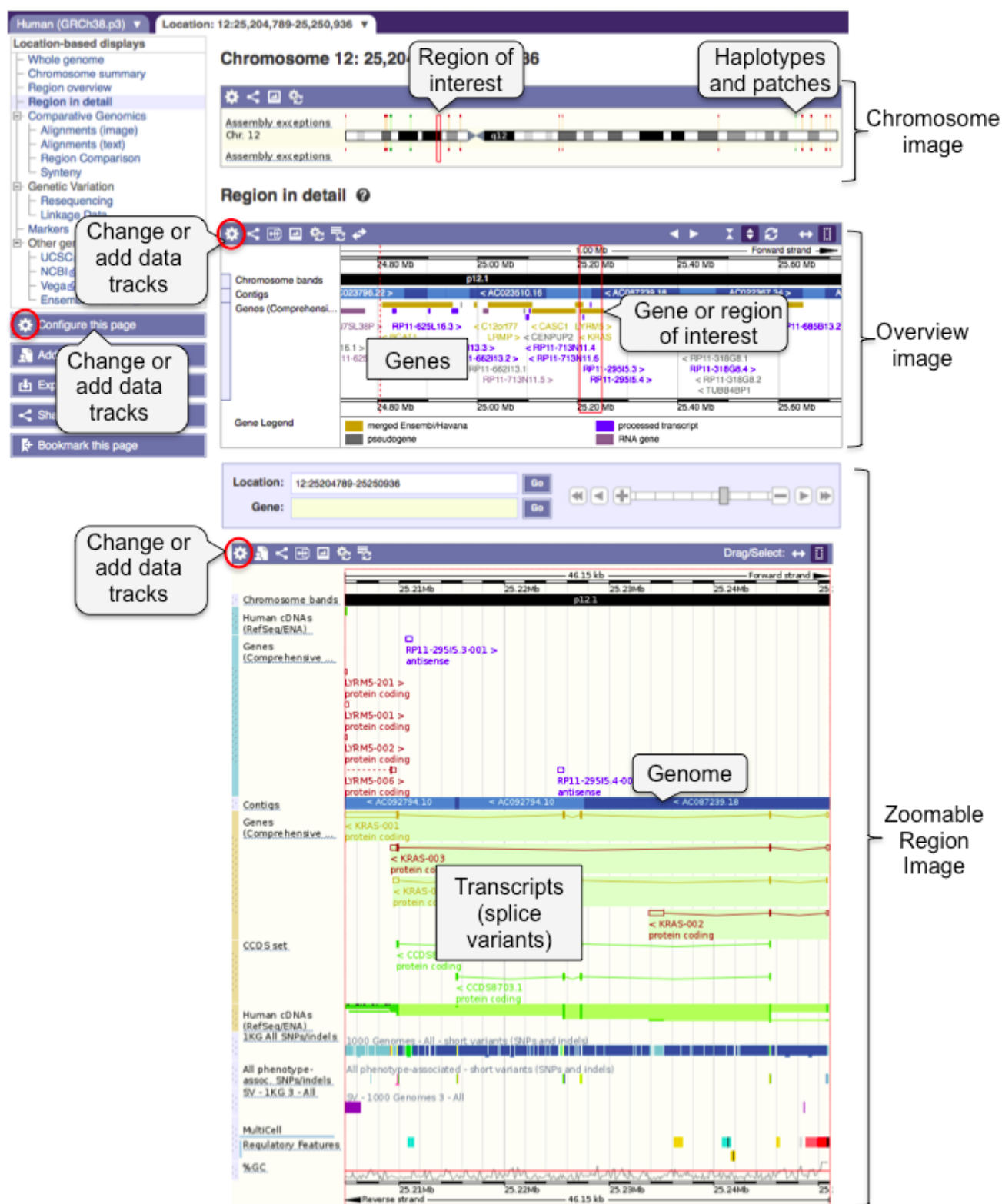
Name GRCh38.p13
RefSeq accession [GCF_000001405.39](#)
GenBank accession [GCA_000001405.28](#)
Submitter Genome Reference Consortium
Level Chromosome
Category Reference genome
Replaced by GCF_000001405.25

Annotation details

Annotation Release 109
Release date May 16, 2021



<http://www.ensembl.org/>



Genome Data Viewer

GDV supports the exploration and analysis of [NCBI-annotated](#) and selected non-NCBI annotated eukaryotic genome assemblies. Currently, over 1220 assemblies are available.

Switch view

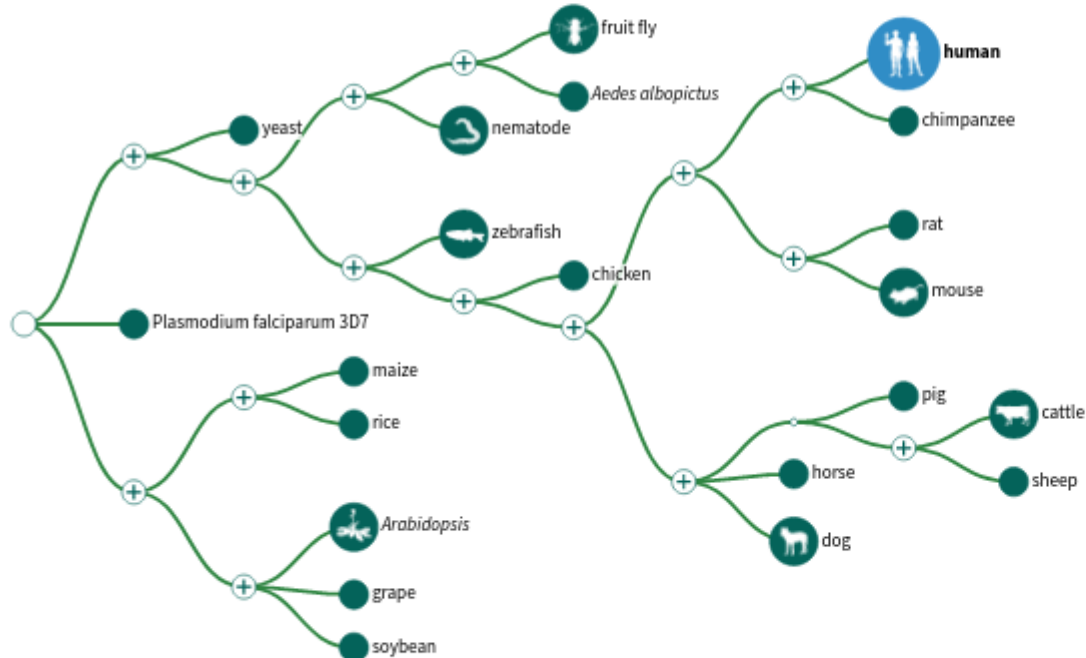


Search organisms

Homo sapiens (human)

To view more organisms in the tree, click on nodes that have '+' signs. Press and hold the '+' to expand and reveal all the subgroups. Or, search for an organism using the search box above.

New! Click on Switch view at the top to see another way of navigating genomes.



Homo sapiens (human)



Search in genome

Location, gene or phenotype


Examples: [TXLNA](#), [chr1:32178000-32200000](#), [DNA repair](#)

Assembly

GRCh38.p13



Browse genome

BLAST genome

Download via NCBI Datasets

Assembly details

Name GRCh38.p13
RefSeq accession [GCF_000001405.39](#)
GenBank accession [GCA_000001405.28](#)
Submitter Genome Reference Consortium
Level Chromosome
Category Reference genome
Replaced by GCF_000001405.25

Annotation details

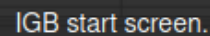
Annotation Release 109
Release date May 16, 2021

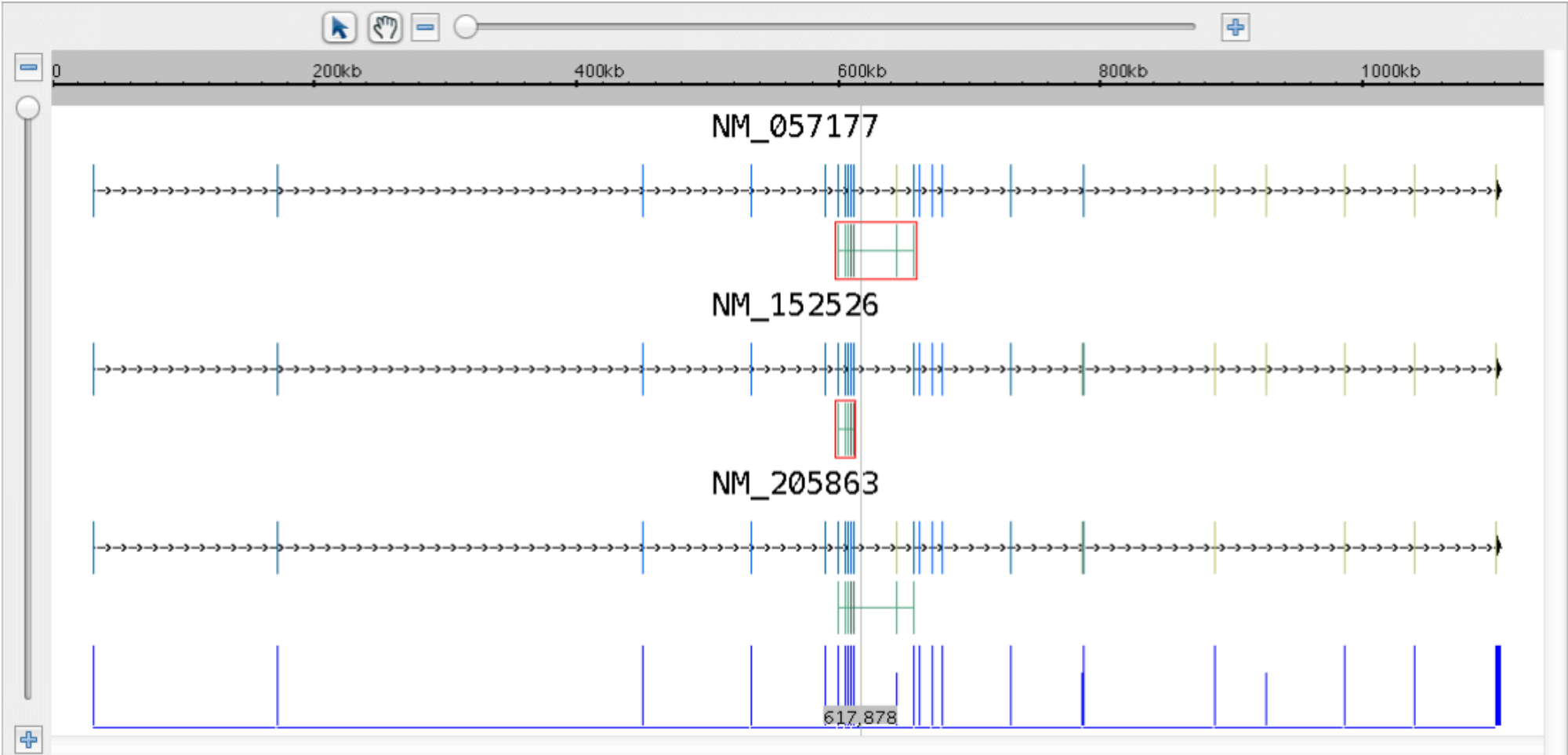


<https://www.ncbi.nlm.nih.gov/genome/gdv/>

Standalone programs

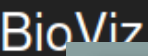
Fork me on Bitbucket





Properties InterProScan

InterPro name	PDZ	PDZ
application	HMMSmart	HMMSmart
library	SMART	SMART
URL	http://www.ebi.ac.uk/interpro/entry/IPR001478	http://www.ebi.ac.uk/interpro/entry/IPR001478
InterPro accession	IPR001478	IPR001478
InterPro description	PDZ domain	PDZ domain
start	599838	599838
end	658674	612021
length	58836	12183
num_spans	3	2
aa_start	211	211


[About](#)
[Help](#)
[Developers](#)
[IGB Genomes](#)
[Apps for IGB](#)
[BioViz Connect](#)

[Install IGB](#)




Integrative
Genomics
Viewer

[Home](#)
[Downloads](#)
[Documents](#)

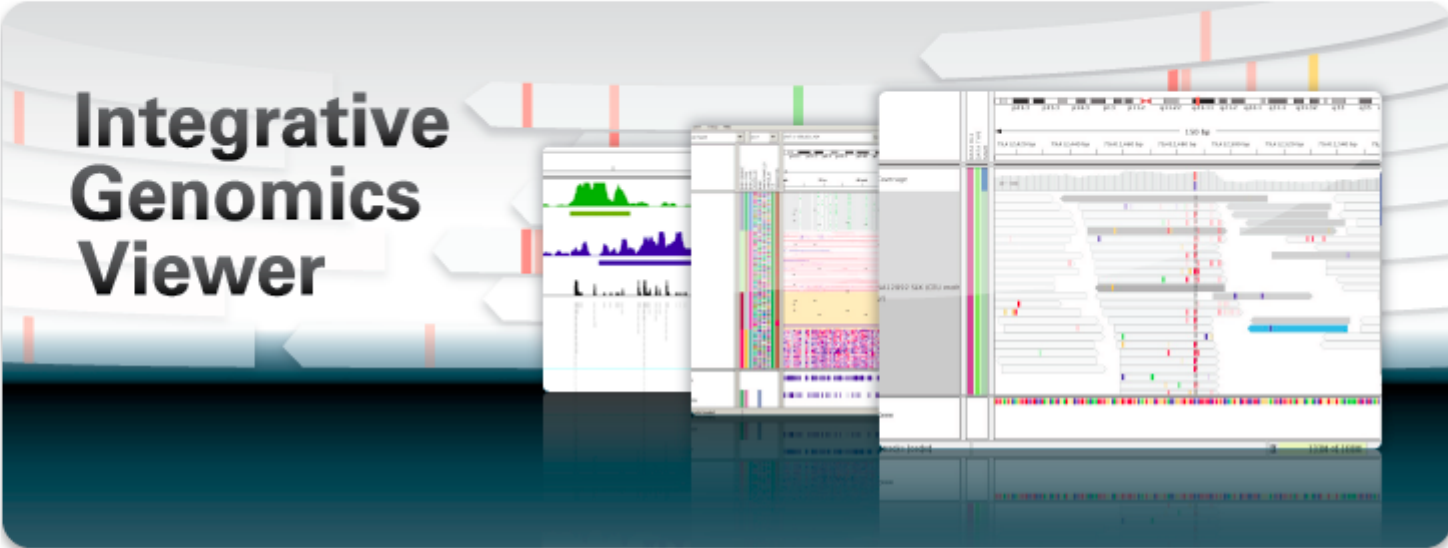
[+ IGV User Guide](#)
[↳ Tutorial Videos](#)
[+ File Formats](#)
[↳ Hosted Genomes](#)
[↳ FAQ](#)
[+ Release Notes](#)
[↳ Credits](#)

[@ Contact](#)

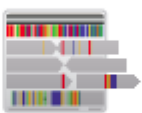
Search website

© 2013-2018
Broad Institute
and the Regents of the
University of California

Home



Overview



The **Integrative Genomics Viewer (IGV)** is a high-performance, easy-to-use, interactive tool for the visual exploration of genomic data. It supports flexible integration of all the common types of genomic data and metadata, investigator-generated or publicly available, loaded from local or cloud sources.

IGV is available in multiple forms, including:

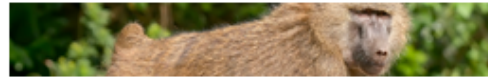
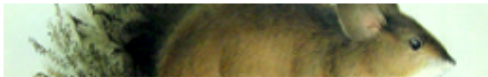
- the original **IGV** - a Java desktop application,
- IGV-Web** - a web application,
- igv.js** - a JavaScript component that can be embedded in web pages (for developers)

Citing IGV

To cite your use of IGV in your publication, please reference one or more of:

James T. Robinson, Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, Jill P. Mesirov. [Integrative Genomics Viewer](#). *Nature Biotechnology* 29, 24–26 (2011). (Free PMC article [here](#)).

Helga Thorvaldsdóttir, James T. Robinson, Jill P. Mesirov. [Integrative Genomics Viewer \(IGV\): high-performance genomics data visualization and exploration](#). *Briefings in Bioinformatics* 14, 178-192 (2013).

☐ Human☐ Chimp☐ Gorilla☐ Gibbon☐ Baboon☐ Rhesus☐ Marmoset☐ Cow☐ Rabbit☐ Dog☐ Mouse☐ Rat☐ Opossum☐ Chicken☐ Zebrafish☐ Spotted Gar

Please take our very brief user survey to help us plan new features for the GTEx Portal: <http://bit.ly/2021gtexsurvey>



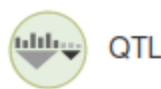
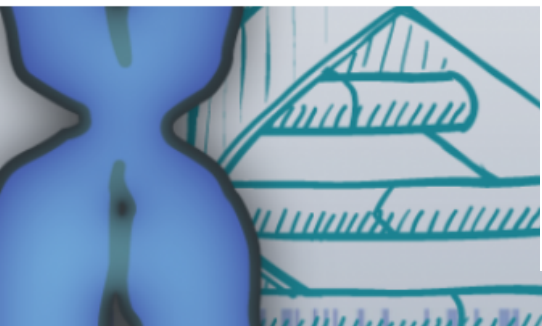
Resource Overview

Current Release (V8)
[Tissue & Sample Statistics](#)
[Tissue Sampling Info \(Anatomogram\)](#)
[Access & Download Data](#)
[Release History](#)
[How to cite GTEx?](#)
The Genotype-Tissue Expression (GTEx) project is

Explore GTEx

-  Browse
- Browse and search all data by gene
- Browse and search all data by variant
- [By Tissue](#) Browse and search all data by tissue
- [Histology Image Viewer](#) Browse and search GTEx histology images

Please take our very brief user survey to help us p



QTL

- | | |
|----------------------------------|--|
| Locus Browser | Visualize QTLs by gene in the Locus Browser |
| Locus Browser VC | Visualize QTLs by variant in the Locus Browser VC (Variant Centric) |
| IGV Browser | Visualize tissue-specific eQTLs and coverage data in the IGV Browser |
| eQTL Dashboard | Batch query eQTLs by gene and tissue |
| eQTL Calculator | Test your own eQTLs |

Resource Overview

Current Release (V8)

- [Tissue & Sample Statistics](#)
- [Tissue Sampling Info \(Anatomogram\)](#)
- [Access & Download Data](#)
- [Release History](#)
- [How to cite GTEx?](#)

The Genotype-Tissue Expression (GTEx) project is

Explore GTEx

 Browse

Browse and search all data by gene

Browse and search all data by variant

[By Tissue](#) Browse and search all data by tissue

[Histology Image Viewer](#) Browse and search GTEx histology images



eGTEx

[H3K27ac, m6A, WGBS](#)

Browse H3K27ac ChIP-seq, m6A methylation, and WGBS DNA methylation data in IGV Browser



Biobank

[Access Biospecimens](#)

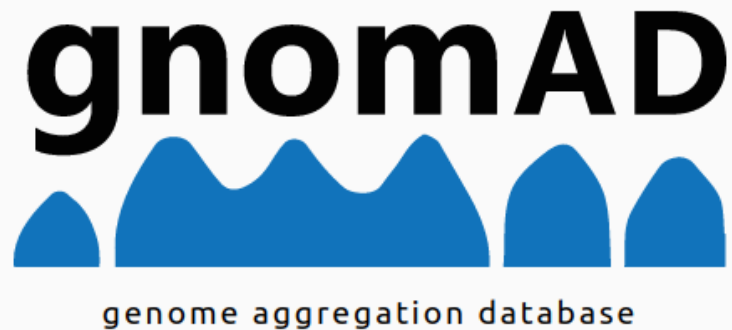
Search and request available GTEx biospecimens

gnomAD browser

gnomAD v2.1.1 ▾ Search

About News Downloads Terms Publications Contact Changelog Help

We want to hear about how you use gnomAD and your wish list! Please take 5 minutes to fill out [our user survey](#).



gnomAD v2.1.1 ▾

Search by gene, region, or variant

Please note that gnomAD v2.1.1 and v3.1.1 have substantially different but overlapping sample compositions and are on different genome builds. For more information, see ["Should I switch to the latest version of gnomAD?"](#)

Examples

- Gene: [PCSK9](#)
- Transcript: [ENST00000302118](#)
- gnomAD v2.1.1 variant: [1-55516888-G-GA](#)

<https://gnomad.broadinstitute.org/>

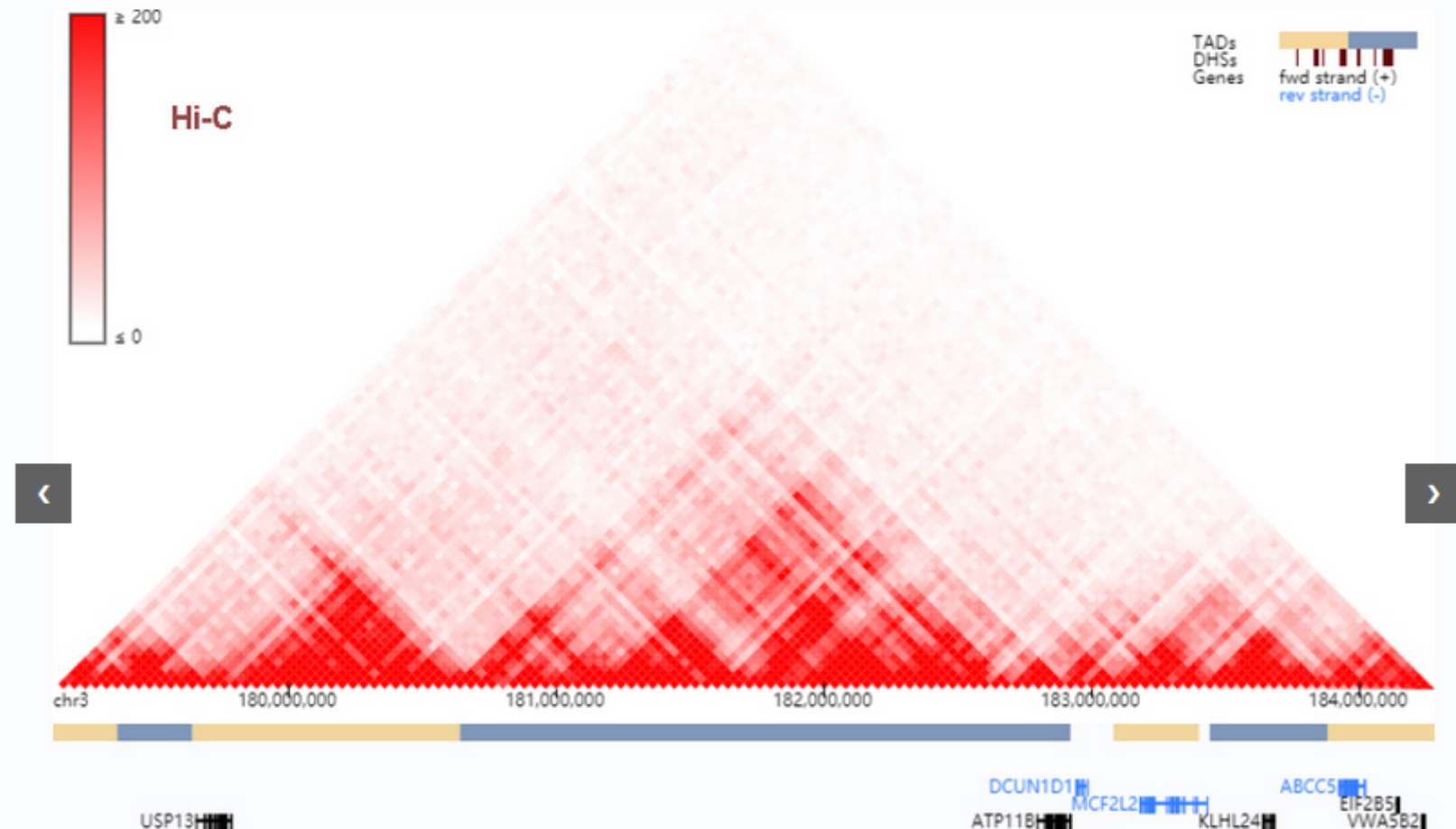


Welcome to YUE Lab
Computational and Functional Genomics/Epigenomics



Intro	Hi-C	Virtual 4C	ChIA-PET & Similar	Capture Hi-C	Compare Hi-C	Inter-chrom	Tutorial	Download
-------	------	------------	--------------------	--------------	--------------	-------------	----------	----------

Welcome to 3D Genome Browser, where you can join 50,000 other users from over 100 countries to explore chromatin interaction data, such as Hi-C, ChIA-PET, Capture Hi-C, PLAC-Seq, and more. Up till Dec. 2017, we have been visited more than 500,000 times!



<http://3dgenome.fsm.northwestern.edu/>



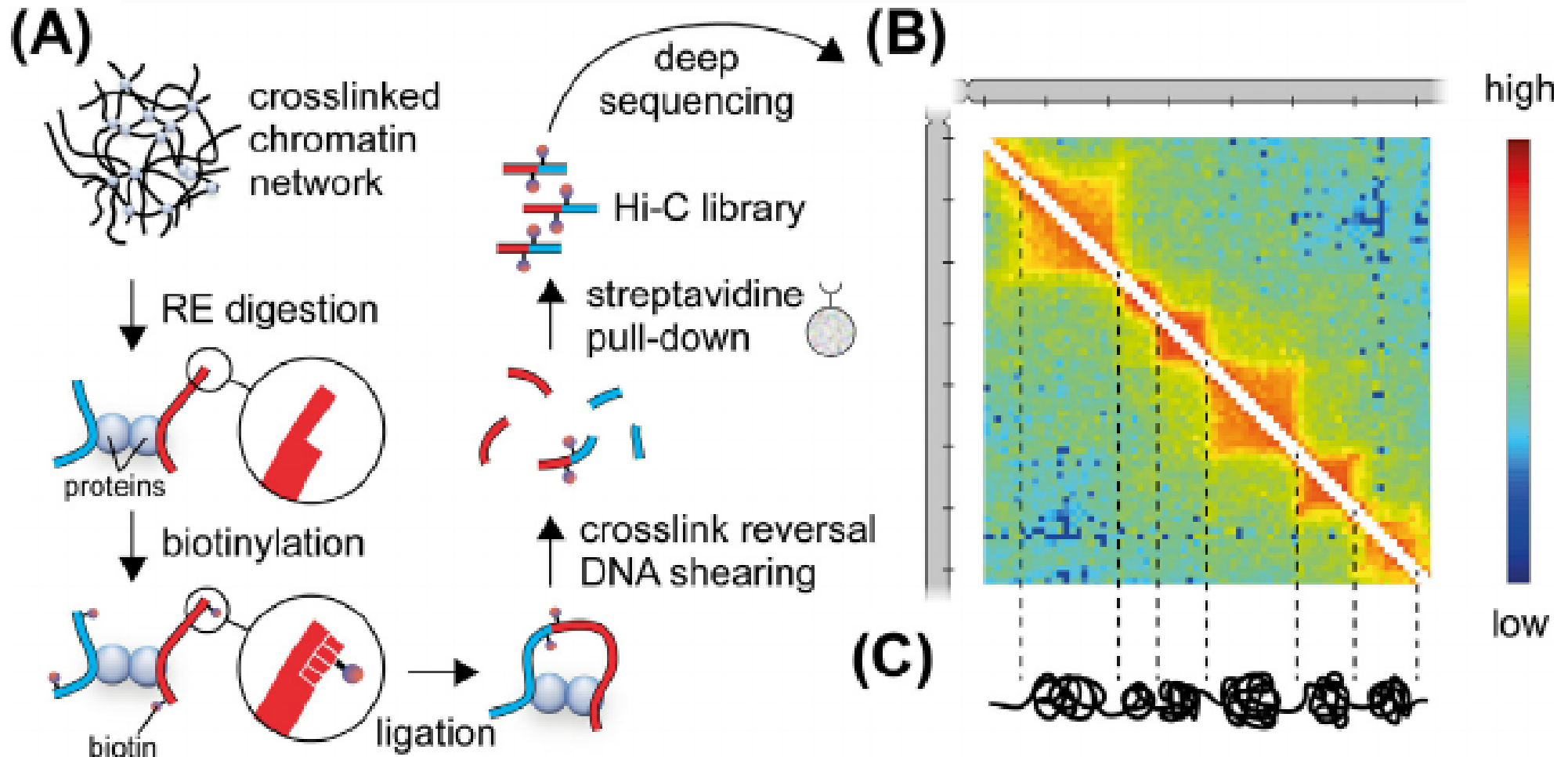
Welcome to YUE Lab

Computational and Functional Genomics/Epigenomics



Intro	Hi-C	Virtual 4C	ChIA-PET & Similar	Capture Hi-C	Compare Hi-C	Inter-chrom	Tutorial	Download
-------	------	------------	--------------------	--------------	--------------	-------------	----------	----------

Welcome to 3D Genome Browser, where you can join 50,000 other users from over 100 countries to explore chromatin interaction data, such as Hi-C, ChIA-PET, Capture Hi-C, PLAC-Seq, and more. Up till Dec. 2017, we have been visited more than 500,000 times!



<http://3dgenome.fsm.northwestern.edu/>

Page **Discussion**

Read

[View source](#)[View history](#)

Welcome to the **Generic Model Organism Database** project, a collection of open source software tools for managing, visualising, storing, and disseminating genetic and genomic data.

Get Started

Read the [GMOD overview](#) for the big picture, or visit [GMOD Components](#) for a comprehensive list of GMOD tools. If GMOD looks promising for your needs, consider attending the next [GMOD community meeting](#).

Get Support

GMOD support is available from several different sources. [Support](#) introduces each support option (this web site, [GMOD Mailing Lists](#), [Training and Outreach](#) activities (including [GMOD Schools](#)), and the [GMOD Help Desk](#)) and offers guidance on which one is the most appropriate for your question.

Get Involved

As an open source project GMOD relies on the [donation of time and software](#) by groups and individuals. Contribution of new tools, adoption of existing ones, and [improving the documentation](#) are all welcome. [Existing](#) and potential users are encouraged to provide feedback via [mailing lists](#) or the [help desk](#). You can also attend project [meetings](#).

Popular GMOD Tools

GMOD is Social

Follow [The Tweet of GMOD](#)

[@gmodproject](#)



Join other [GMOD users](#) on
[LinkedIn](#)

GMOD News

[Add](#)

[JBrowse2 v1.1.0 Released](#)
[Prospecting for Proposals for GSoC 2021](#)
[Codefest 2020](#)
[Tripal 3.5 Released](#)
[JBrowse 2 release](#)
[Prospecting for Proposals for GSoC 2020](#)
[Call for PAG Abstracts](#)
[New GMOD Server](#)
[GMOD-JBrowse 2016 Survey](#)
[GCC2016](#)

Navigation

[GMOD Home](#)
[Software](#)
[Categories / Tags](#)
[View all pages](#)

Documentation

[Overview](#)
[FAQs](#)
[HOWTOs](#)
[Glossary](#)

Community

[GMOD News](#)
[Training / Outreach](#)
[Support](#)
[GMOD Promotion](#)
[Meetings](#)
[Calendar](#)


Page [Discussion](#)

Read

[View source](#)
[View history](#)
 Search


JBrowse 2: New super-fast genome
annotation viewer



GBrowse: Genome annotation viewer



Galaxy: Data analysis & integration



Chado: Biological database schema



JBrowse: Super-fast genome annotation
viewer



BioMart: Data mining system



WebApollo: browser-based annotation editor



MAKER: Genome annotation pipeline



GBrowse_syn: Synteny viewer



Tripal: Chado web interface



InterMine: Data warehousing



CMap: Comparative map viewer



Pathway Tools: Metabolic, regulatory
pathways



Canto: literature annotation tool



on

Add

C 2021

C 2020

Thank you for your time
and
See you at the next lecture

Any other
questions & comments

lukaskoz@mimuw.edu.pl

Credits & links:

<https://www.slideshare.net/hhalhaddad/281-lec7-genomeorganization>

<https://www.slideshare.net/bcbbslides/mapping-and-de-novo-assembly>

<https://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi>

<https://en.wikipedia.org/wiki/Genome>

https://en.wikipedia.org/wiki/Bacterial_genome

https://en.wikipedia.org/wiki/Human_genome