Genomes and their structures

CB 2010/6010

William KM Lai

Learning objectives:

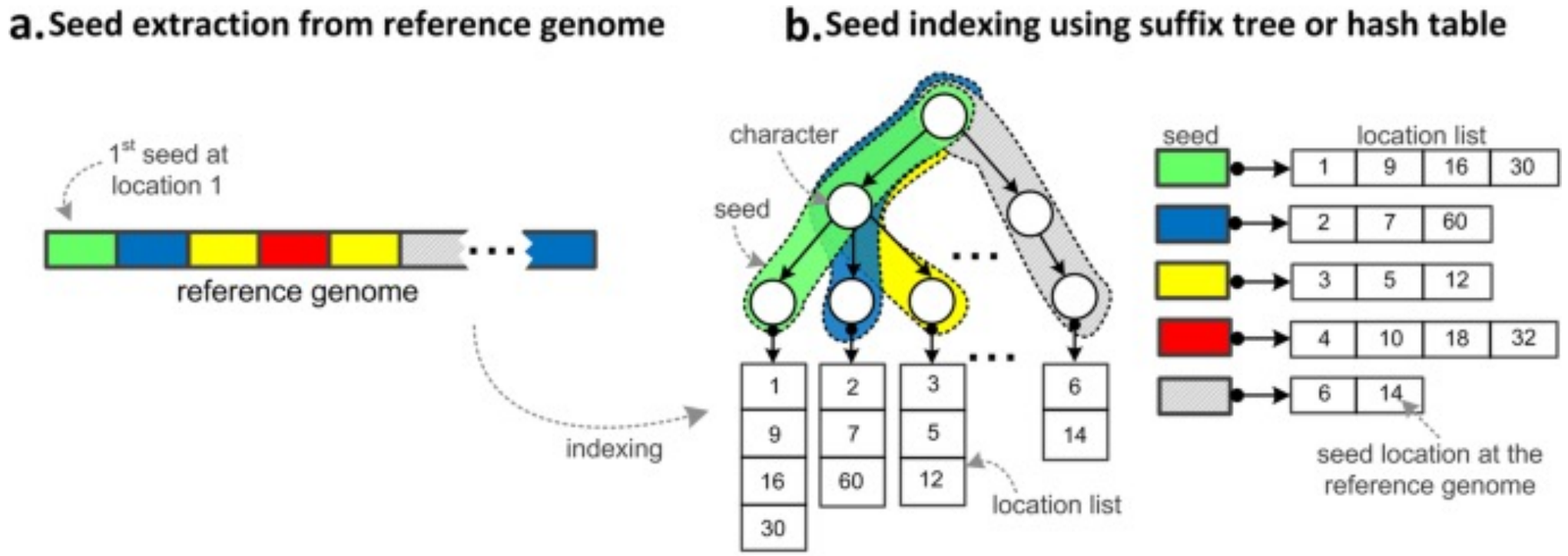
- Basics of DNA sequence aligners
- Overview of fundamentals of genome organization (i.e., chromosomes)
- Genome structure differences between organisms
- What is chromatin?
  - How chromatin structure effects functions
- What other elements are embedded in the genome?
  - Enhancers, origins of replication, centromeres, etc.
- Overview of common genomic assays
  - ChIP, ATAC, etc.

We've got our sequence read. Where does it go?

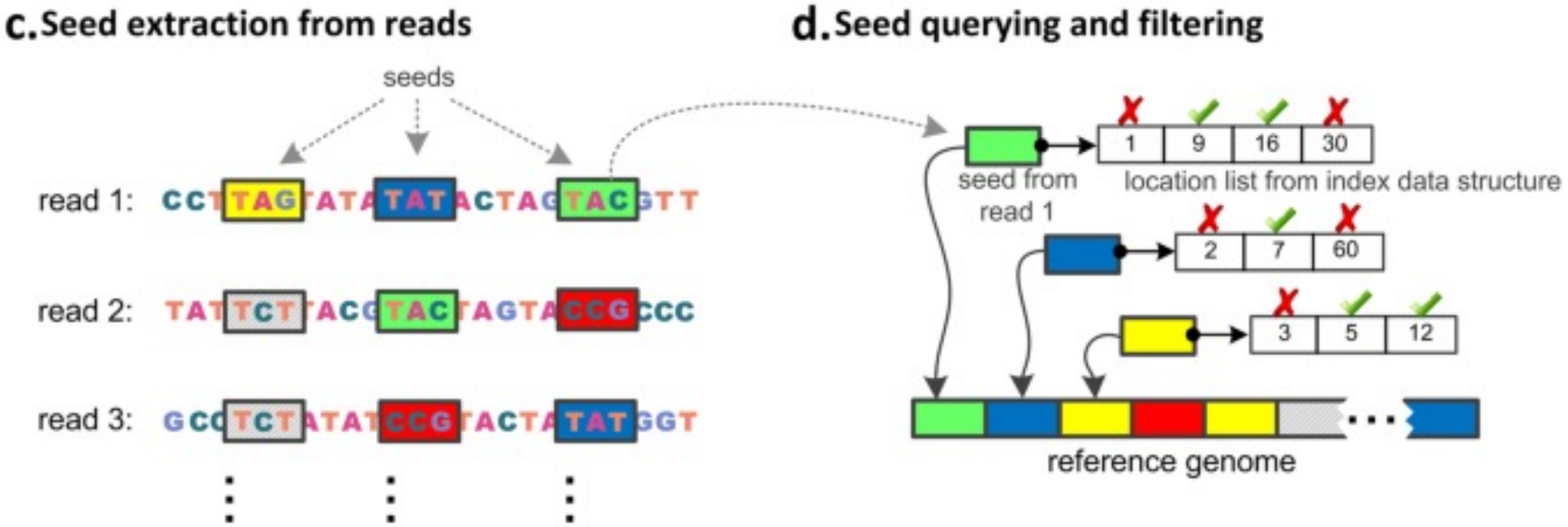
AACATCGTACGTCTAA

```
> grep "AACATCGTACGTCTAA" reference_genome.fa > location.txt
```

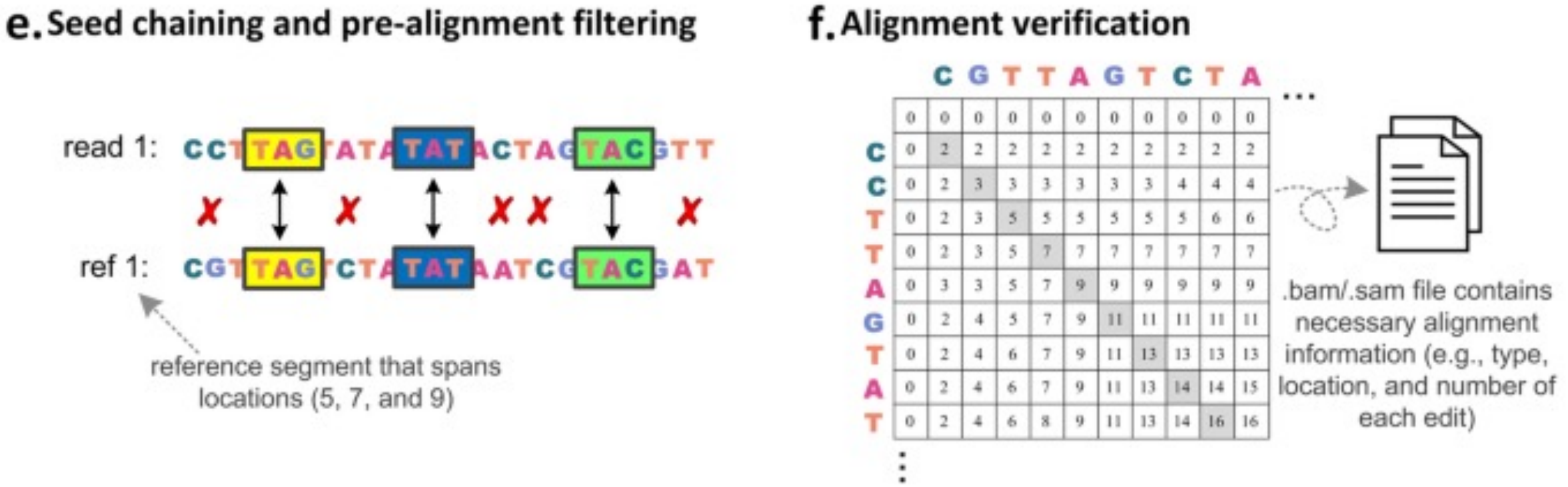
- Something like this 'works' but is super infeasible at scale
- For reference, top-tier sequencers can produce 20 billion reads a day and at places like BROAD are running 24/7
- We need true algorithmic approaches!



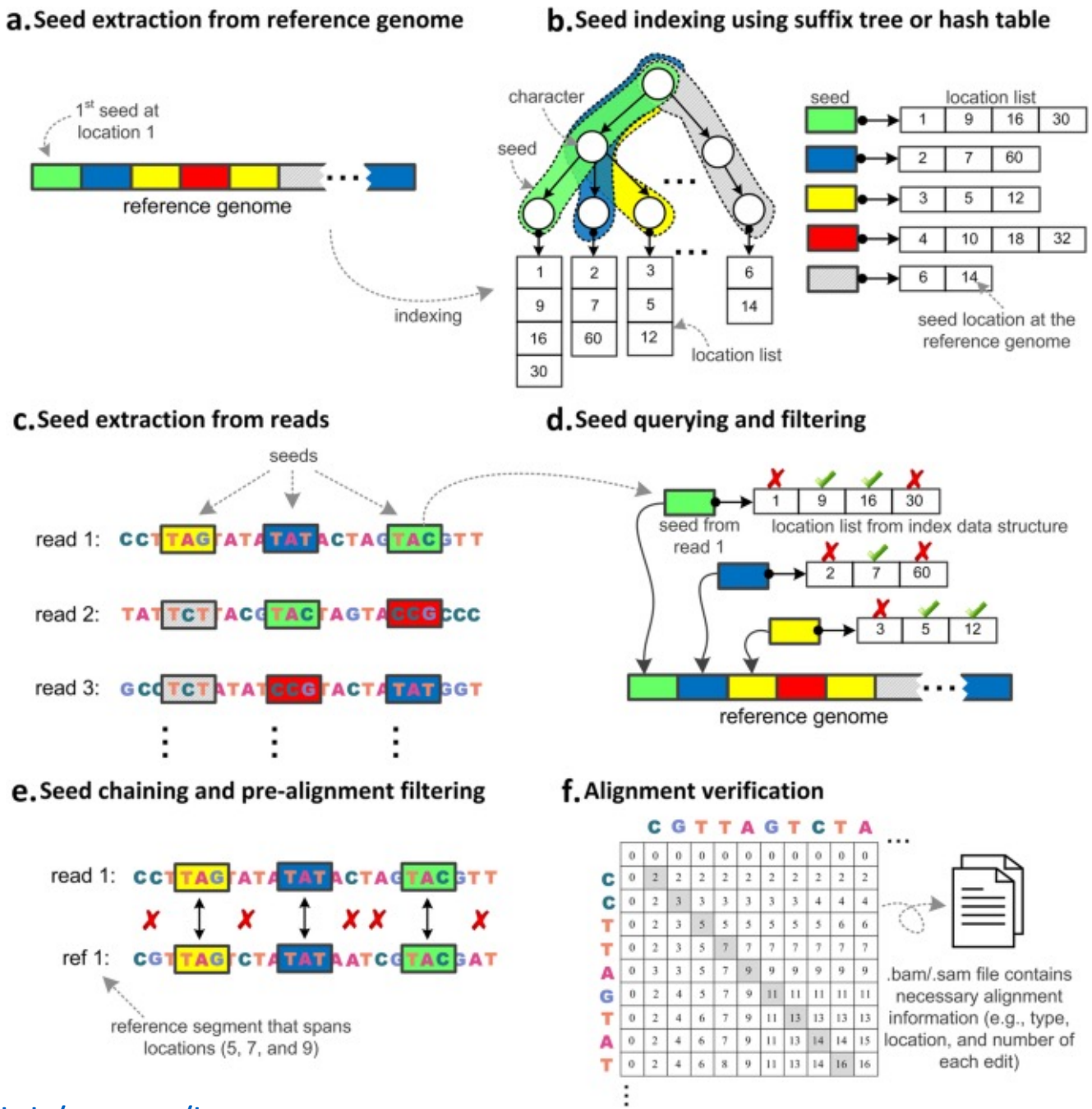
https://genomebiology.biomedcentral.com/articles/10.1186/s13059-021-02443-7



https://genomebiology.biomedcentral.com/articles/10.1186/s13059-021-02443-7



https://genomebiology.biomedcentral.com/articles/10.1186/s13059-021-02443-7

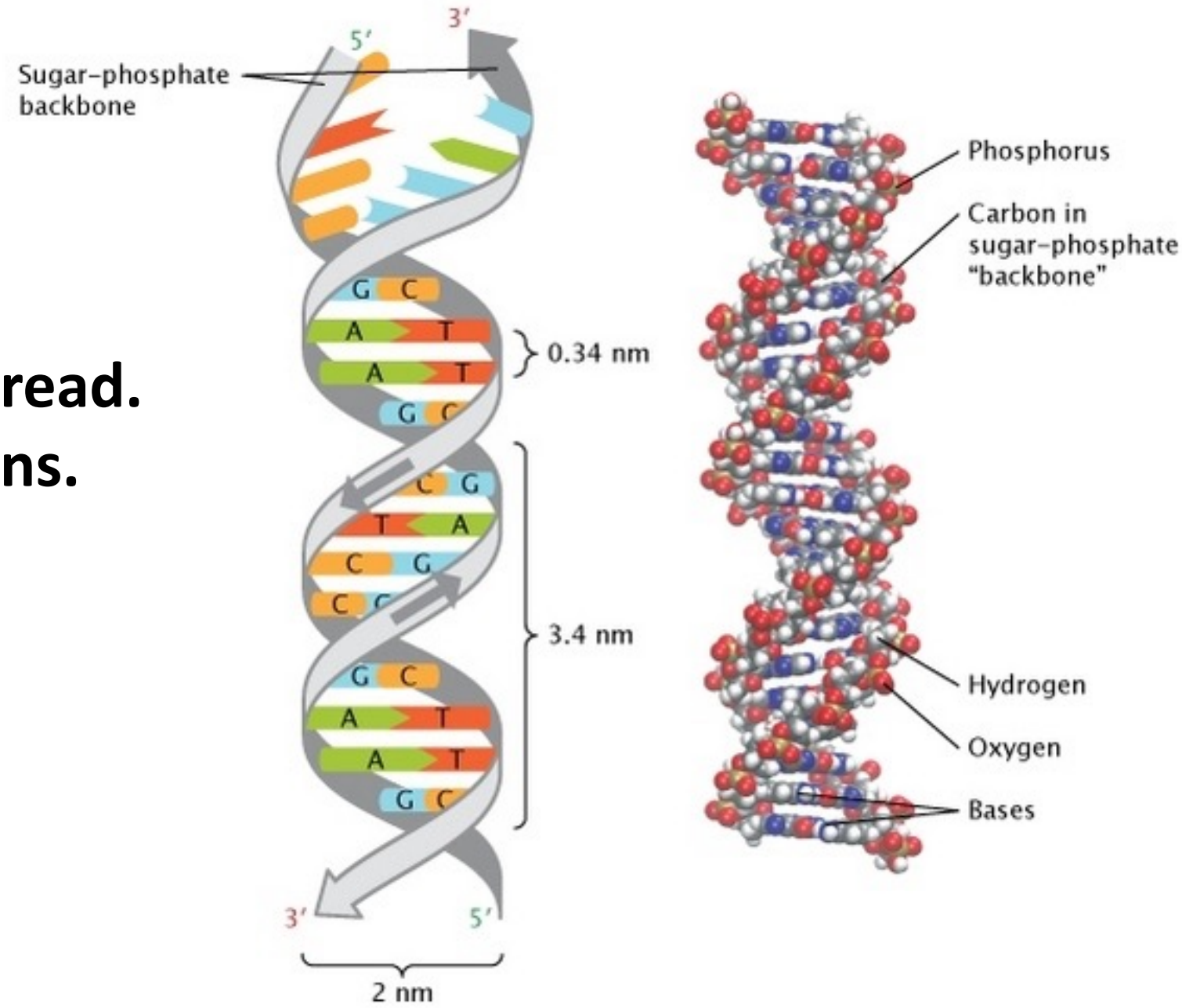


https://github.com/lh3/bwa/blob/master/bwamem.c

https://genomebiology.biomedcentral.com/articles/10.1186/s13059-021-02443-7

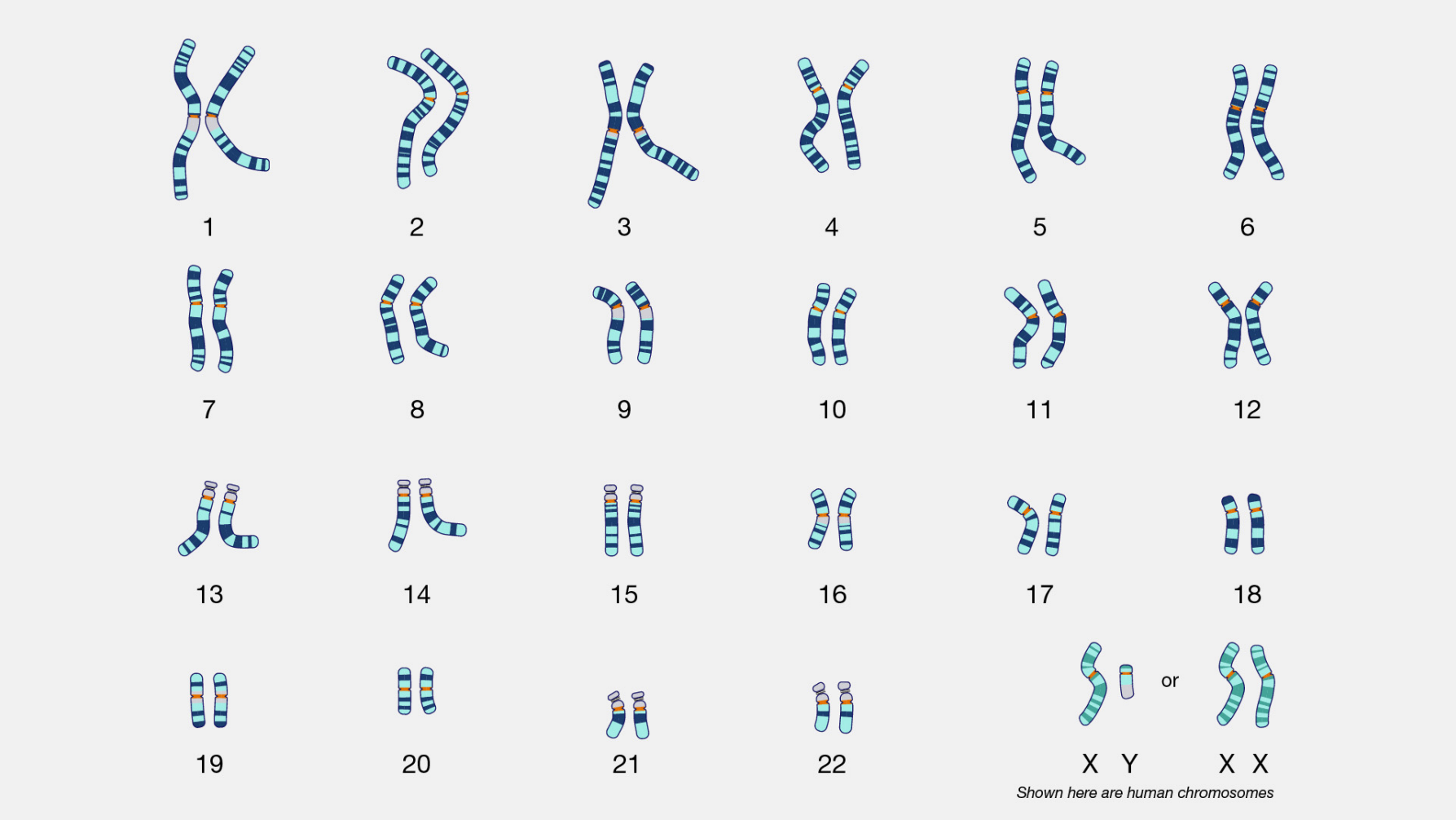
We've got our sequence read.
We know where it aligns.

Now what?

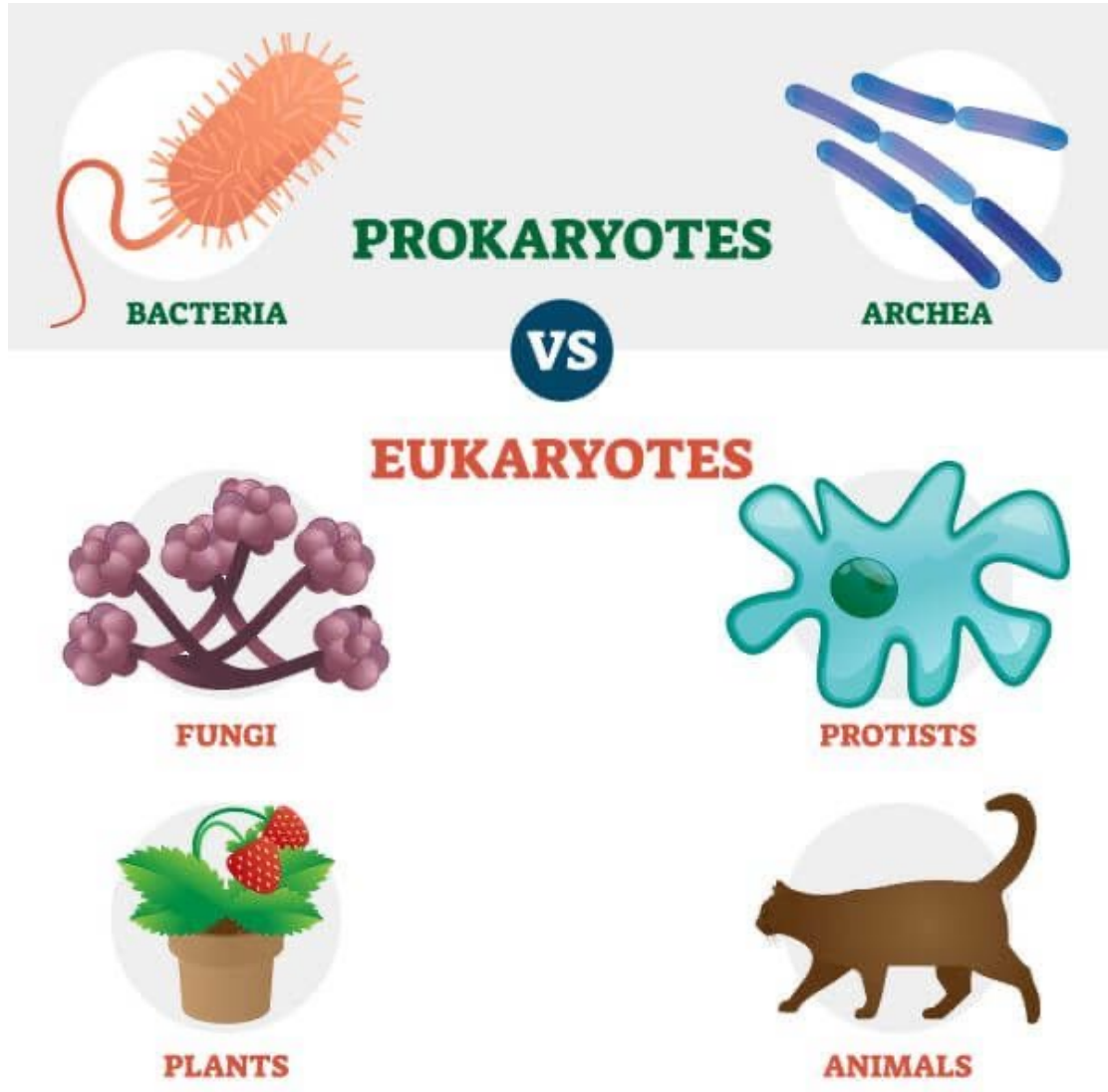


https://www.nature.com/scitable/content/ne0000/ne0000/ne0000/ne0000/104944953/73_1_2.jpg

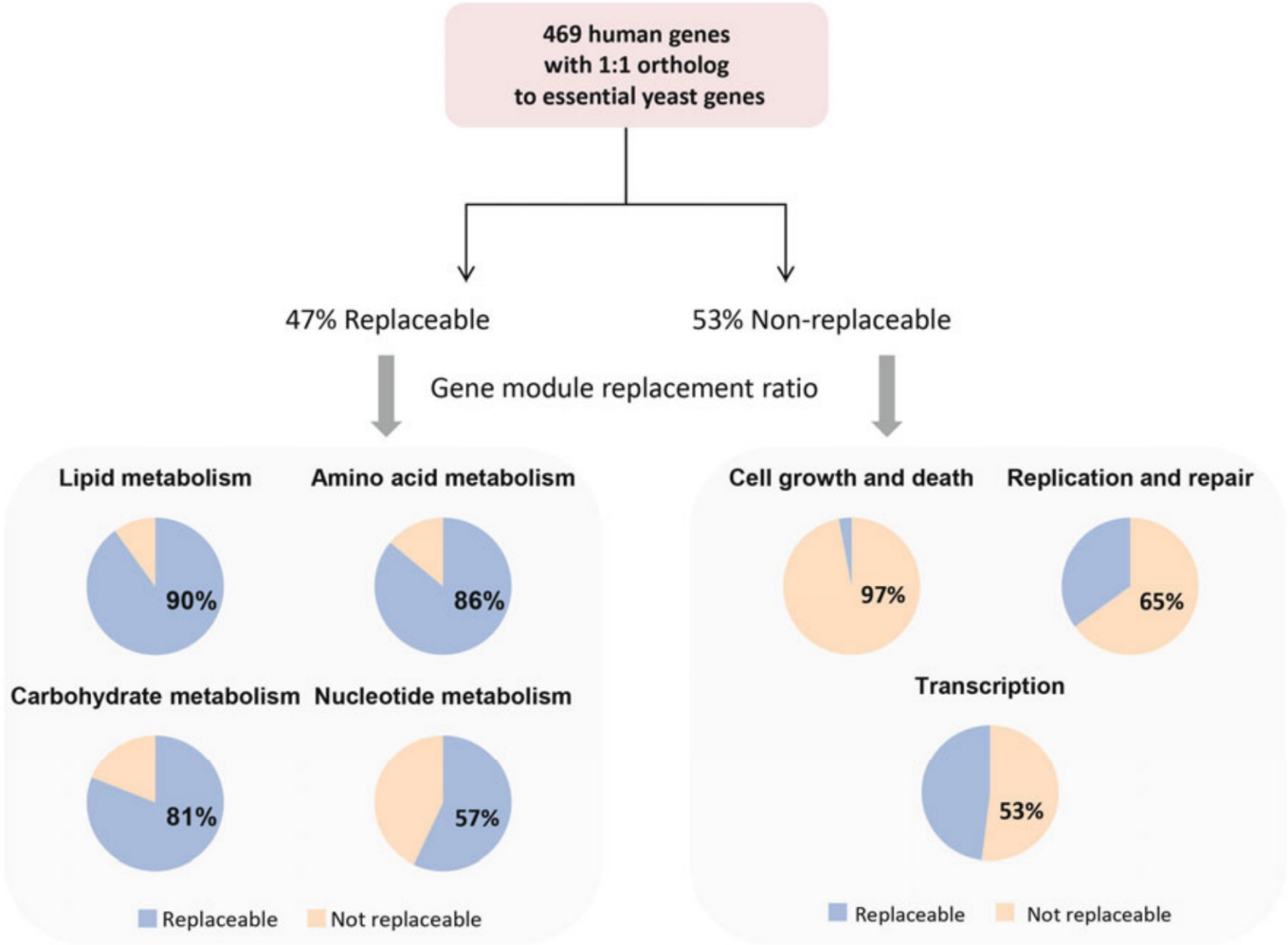
DNA in the nucleus is very well organized!



DNA in eukaryotes are typically organized into chromosomes


PROKARYOTES
BACTERIA
ARCHEA
VS
EUKARYOTES
FUNGI
PROTISTS
PLANTS
ANIMALS

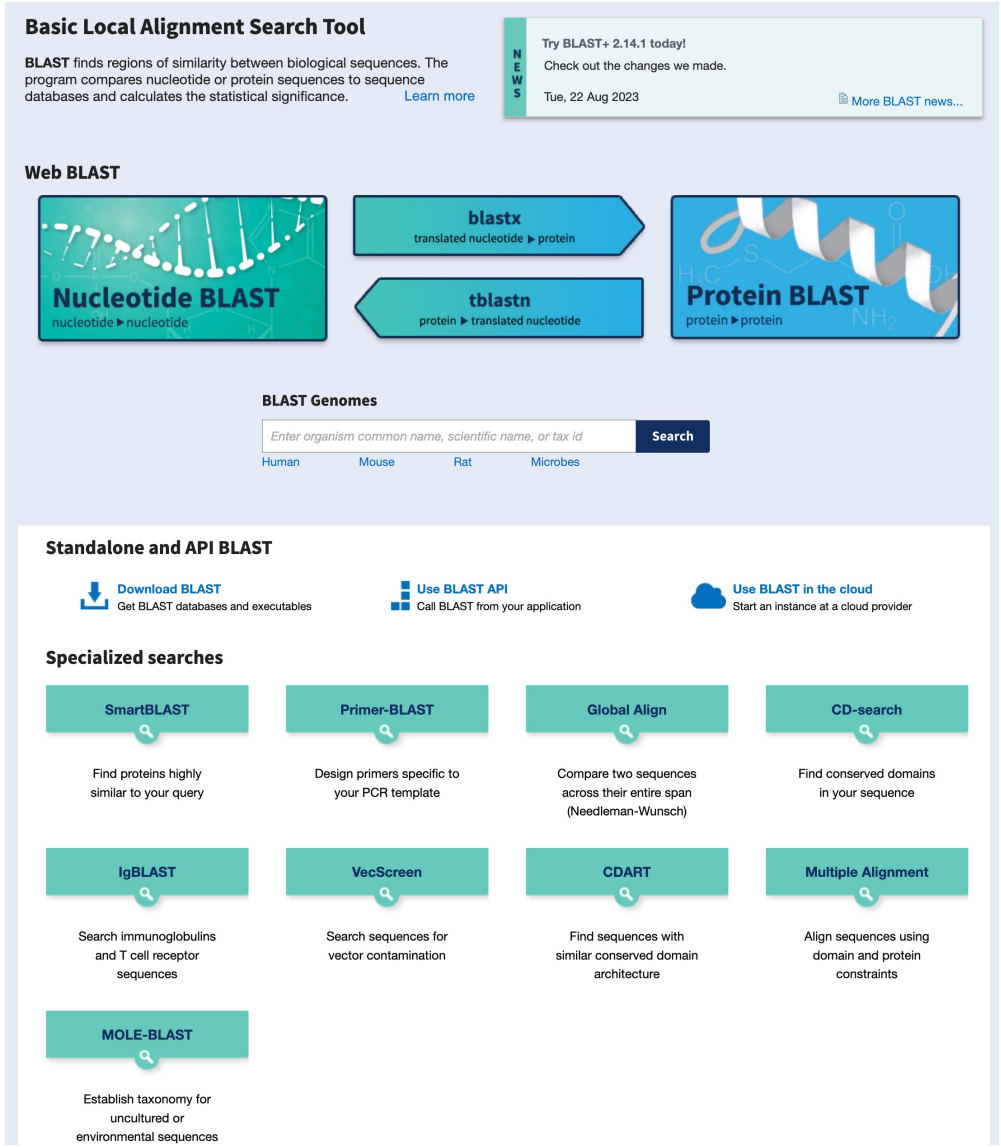
Sequence is VERY conserved across organisms

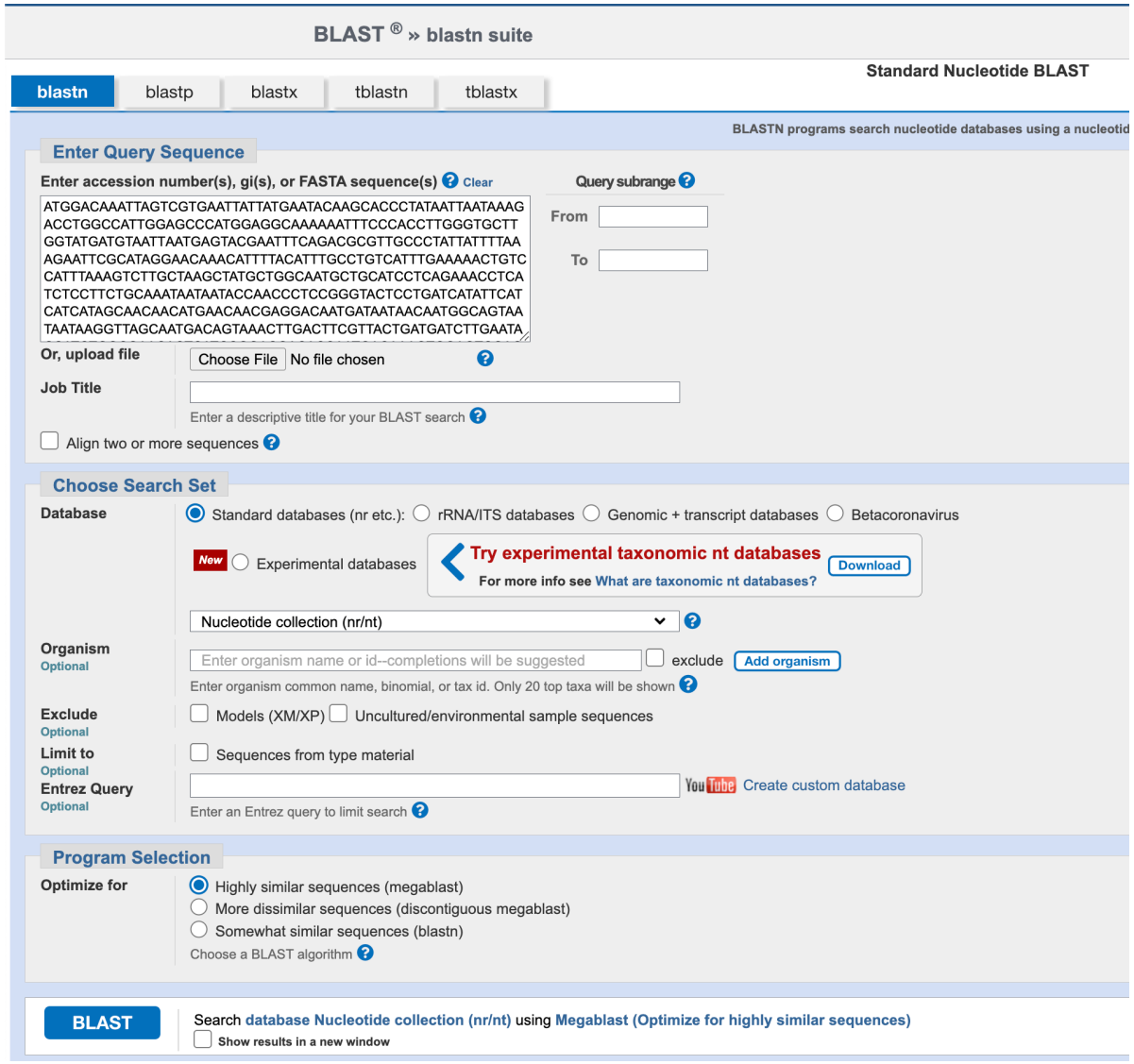


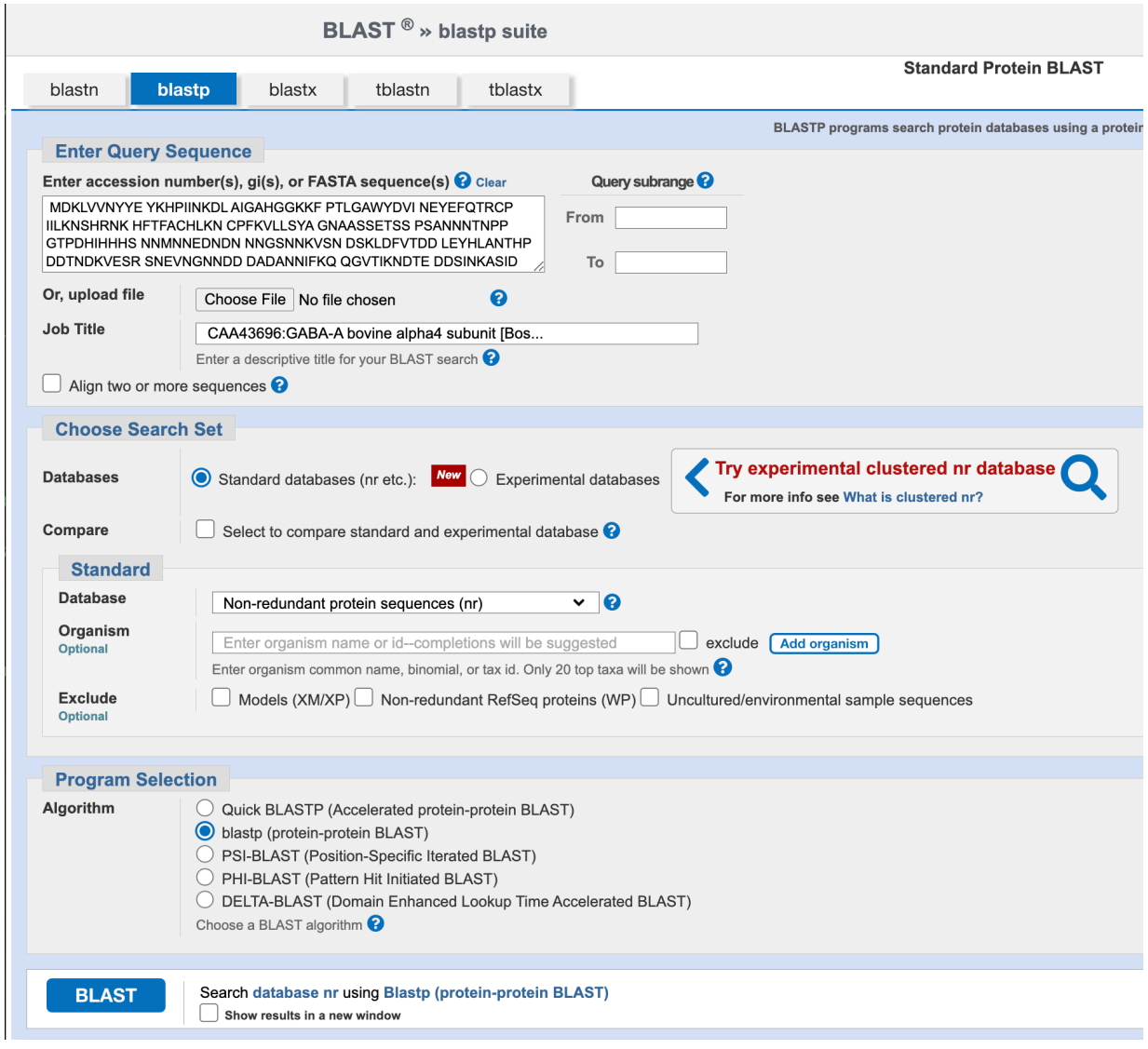
Kim et al. Journal of Microbiology (2020)

Basic Local Alignment Search Tool (BLAST)

https://blast.ncbi.nlm.nih.gov/Blast.cgi




BLAST ® » blastn suite
Standard Nucleotide BLAST
blastn
blastp
blastx
tblastn
tblastx
BLASTN programs search nucleotide databases using a nucleotid
Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s) Clear
ATGGACAAATTAGTCGTGAATTATTATGAATACAAGCACCCTATAATTAATAAAG
ACCTGGCCATTGGAGCCCATGGAGGCAAAAAATTTCCCACCTTGGGTGCTT
GGTATGATGTAATTAATGAGTACGAATTTCAGACGCGTTGCCCTATTATTTTAA
AGAATTCGCATAGGAACAAACATTTTACATTTGCCTGTCATTTGAAAAACTGTC
CATTTAAAGTCTTGCTAAGCTATGCTGGCAATGCTGCATCCTCAGAAACCTCA
TCTCCTTCTGCAAATAATAATACCAACCCTCCGGGTACTCCTGATCATATTCAT
CATCATAGCAACAACATGAACAACGAGGACAATGATAATAACAATGGCAGTAA
TAATAAGGTTAGCAATGACAGTAAACTTGACTTCGTTACTGATGATCTTGAATA
Query subrange
From
To
Or, upload file
Choose File No file chosen
Job Title
Enter a descriptive title for your BLAST search
Align two or more sequences
Choose Search Set
Database
Standard databases (nr etc.): rRNA/ITS databases Genomic + transcript databases Betacoronavirus
New Experimental databases
Try experimental taxonomic nt databases
For more info see What are taxonomic nt databases?
Download
Nucleotide collection (nr/nt)
Organism
Optional
Enter organism name or id--completions will be suggested
exclude
Add organism
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown
Exclude
Optional
Models (XM/XP) Uncultured/environmental sample sequences
Limit to
Optional
Sequences from type material
Entrez Query
Optional
Create custom database
Enter an Entrez query to limit search
Program Selection
Optimize for
Highly similar sequences (megablast)
More dissimilar sequences (discontiguous megablast)
Somewhat similar sequences (blastn)
Choose a BLAST algorithm
BLAST
Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)
Show results in a new window


BLAST ® » blastp suite
Standard Protein BLAST
blastn
blastp
blastx
tblastn
tblastx
BLASTP programs search protein databases using a protein
Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s) Clear
MDKLVVNYYE YKHPIINKDL AIGAHGGKKF PTLGAWYDVI NEYEFQTRCP
IILKNSHRNK HFTFACHLKN CPFKVLLSYA GNAASSETSS PSANNNTNPP
GTPDHIHHHS NNMNNEDNDN NNGSNNKVSN DSKLDFVTDD LEYHLANTHP
DDTNDKVESR SNEVNGNNDD DADANNIFKQ QGVTIKNDTE DDSINKASID
Query subrange
From
To
Or, upload file
Choose File No file chosen
Job Title
CAA43696:GABA-A bovine alpha4 subunit [Bos...
Enter a descriptive title for your BLAST search
Align two or more sequences
Choose Search Set
Databases
Standard databases (nr etc.): New Experimental databases
Try experimental clustered nr database
For more info see What is clustered nr?
Compare
Select to compare standard and experimental database
Standard
Database
Non-redundant protein sequences (nr)
Organism
Optional
Enter organism name or id--completions will be suggested
exclude
Add organism
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown
Exclude
Optional
Models (XM/XP) Non-redundant RefSeq proteins (WP) Uncultured/environmental sample sequences
Program Selection
Algorithm
Quick BLASTP (Accelerated protein-protein BLAST)
blastp (protein-protein BLAST)
PSI-BLAST (Position-Specific Iterated BLAST)
PHI-BLAST (Pattern Hit Initiated BLAST)
DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)
Choose a BLAST algorithm
BLAST
Search database nr using Blastp (protein-protein BLAST)
Show results in a new window

BLAST ® » blastp suite » results for RID-EVWMKN7A01R

Home | Recent Results | Saved Strategies | Help

Edit Search | Save Search | Search Summary

How to read this report? | BLAST Help Videos | Back to Traditional Results Page

Job Title	CAA43696:GABA-A bovine alpha4 subunit [Bos...
RID	EVWMKN7A01R *Search expires on 08-30 21:28 pm* Download All
Program	BLASTP Citation
Database	nr See details
Query ID	lcl|Query_642224
Description	unnamed protein product
Molecule type	amino acid
Query Length	1468
Other reports	Distance tree of results Multiple alignment MSA viewer

Filter Results

Organism *only top 20 will appear* exclude

Type common name, binomial, taxid or group name

+ Add organism

Percent Identity to

E value to

Query Coverage to

Filter | Reset

Compare these results against the new Clustered nr database BLAST

Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments Download | Select columns | Show 100

select all *100 sequences selected* GenPept | Graphics | Distance tree of results | Multiple alignment | MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
chromatin-remodeling ATPase CHD1 [Saccharomyces cerevisiae S288C]	Saccharomyces cerevisiae S288C	3036	3036	100%	0.0	100.00%	1468	NP_011091.1
chromatin-remodeling ATPase CHD1 [Saccharomyces cerevisiae]	Saccharomyces cerevisiae	3026	3026	100%	0.0	99.59%	1468	PTN16185.1
BAM_G0013330.mRNA.1.CDS.1 [Saccharomyces cerevisiae]	Saccharomyces cerevisiae	3026	3026	100%	0.0	99.59%	1468	CAI4407594.1
ANM_collapsed_G0016510.mRNA.1.CDS.1 [Saccharomyces cerevisiae]	Saccharomyces cerevisiae	3026	3026	100%	0.0	99.66%	1468	CAI6639774.1
Chd1p [Saccharomyces cerevisiae YJM993]	Saccharomyces cerevisiae YJM993	3026	3026	100%	0.0	99.66%	1468	AHY75720.1

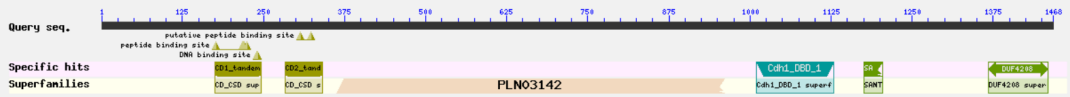
Descriptions | Graphic Summary | Alignments | Taxonomy

hover to see the title *click to show alignments* *Show Conserved Domains*

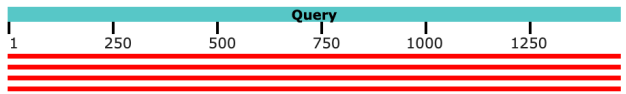
Alignment Scores < 40 | 40 - 50 | 50 - 80 | 80 - 200 | >= 200

100 sequences selected

Putative conserved domains have been detected, click on the image below for detailed results.



Distribution of the top 100 Blast Hits on 100 subject sequences

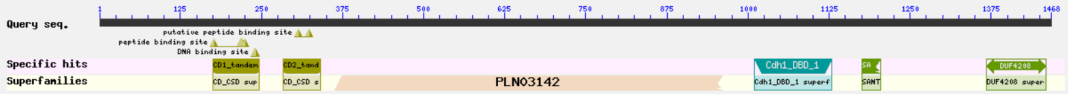


Descriptions | **Graphic Summary** | Alignments | Taxonomy

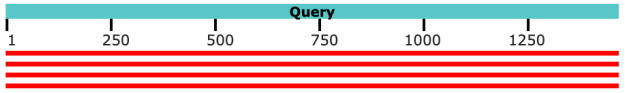
hover to see the title | *click to show alignments* | *Show Conserved Domains* | Alignment Scores < 40 | 40 - 50 | 50 - 80 | 80 - 200 | >= 200

100 sequences selected

Putative conserved domains have been detected, click on the image below for detailed results.



Distribution of the top 100 Blast Hits on 100 subject sequences

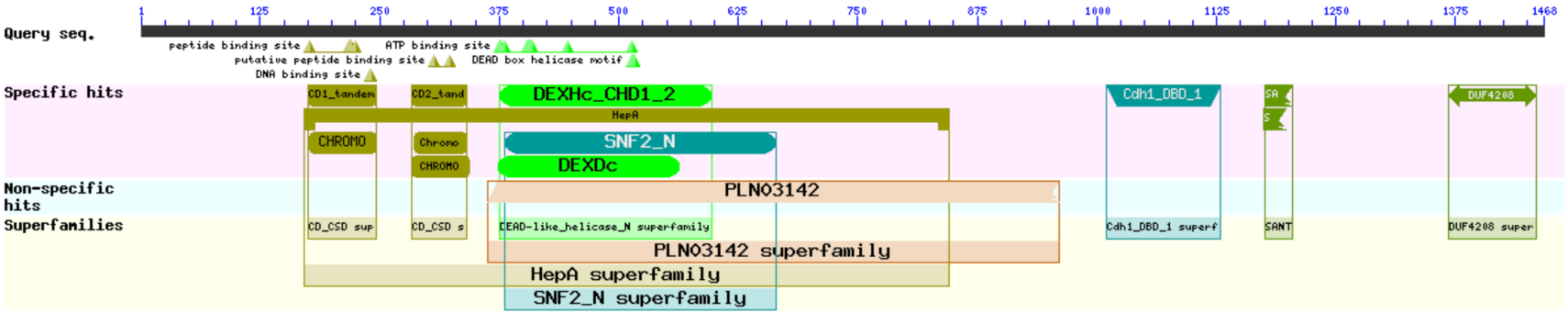


Conserved domains on [lcl|Query_214074]

View: Standard Results

Local query sequence

Graphical summary | Zoom to residue level | show extra options »

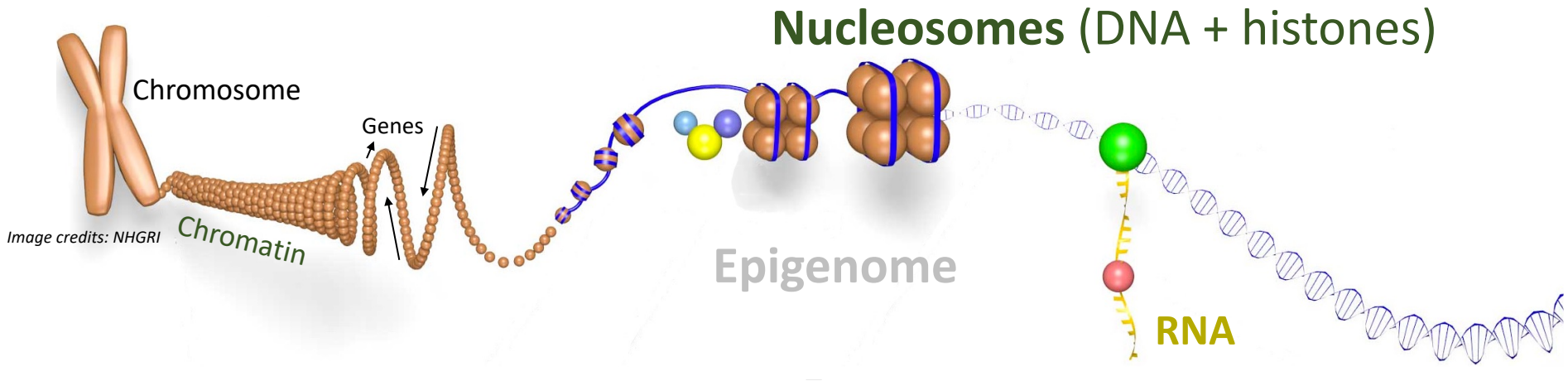


Search for similar domain architectures | Refine search

List of domain hits

Name	Accession	Description	Interval	E-value
[+] PLN03142	PLN03142	Probable chromatin-remodeling complex ATPase chain; Provisional	363-961	4.30e-160
[+] DEXHc_CHD1_2	cd17993	DEXH-box helicase domain of the chromodomain helicase DNA binding proteins 1 and 2, and ...	375-597	2.08e-142
[+] HepA	COG0553	Superfamily II DNA or RNA helicase, SNF2 family [Transcription, Replication, recombination and ...	171-845	1.01e-120
[+] SNF2_N	pfam00176	SNF2 family N-terminal domain; This domain is found in proteins involved in a variety of ...	380-665	1.12e-87
[+] Cdh1_DBD_1	pfam18196	Chromodomain helicase DNA-binding domain 1; This domain can be found in chromodomain helicase ...	1010-1129	1.65e-48
[+] DUF4208	pfam13907	Domain of unknown function (DUF4208); This domain is found at the C-terminus of ...	1368-1460	2.13e-36
[+] CD2_tandem_ScCHD1_like	cd18664	repeat 2 of the paired tandem chromodomains of yeast chromodomain helicase DNA-binding protein ...	283-341	2.19e-32
[+] CD1_tandem_CHD1_yeast_like	cd18665	repeat 1 of the paired tandem chromodomains of yeast chromodomain helicase DNA-binding protein ...	175-247	3.01e-29
[+] DEXDc	smart00487	DEAD-like helicases superfamily;	372-565	1.52e-26
[+] Chromo	pfam00385	Chromo (CHRomatin Organisation MOdifier) domain;	286-341	3.02e-11
[+] CHROMO	smart00298	Chromatin organization modifier domain;	284-343	1.58e-09
[+] SANT_TRF	cd11660	Telomere repeat binding factor-like DNA-binding domains of the SANT/myb-like family; Human ...	1175-1204	3.01e-07
[+] CHROMO	smart00298	Chromatin organization modifier domain;	176-247	1.37e-04
[+] SANT	smart00717	SANT SWI3, ADA2, N-CoR and TFIIIB" DNA-binding domains;	1174-1198	7.18e-03

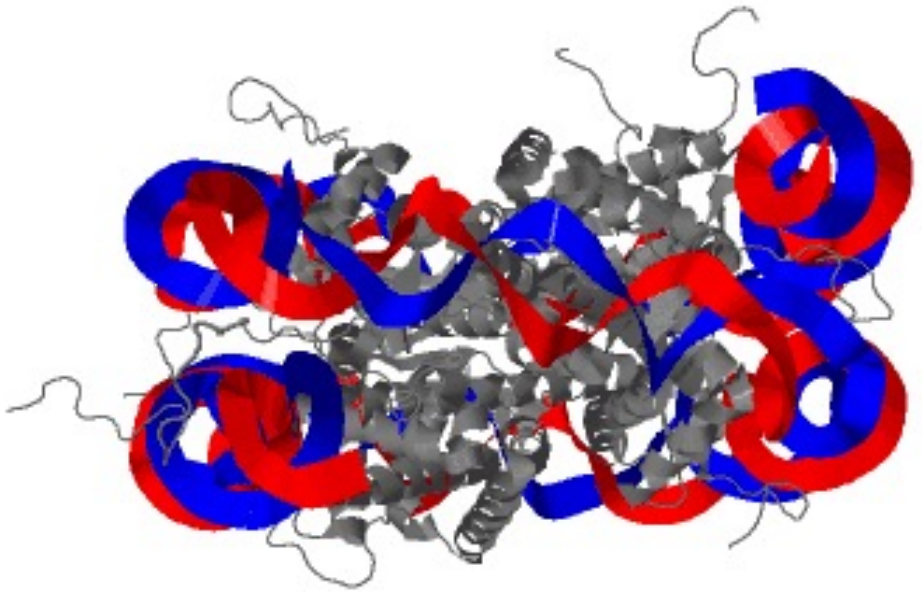
DNA is VERY well organized



Nucleosome core particle

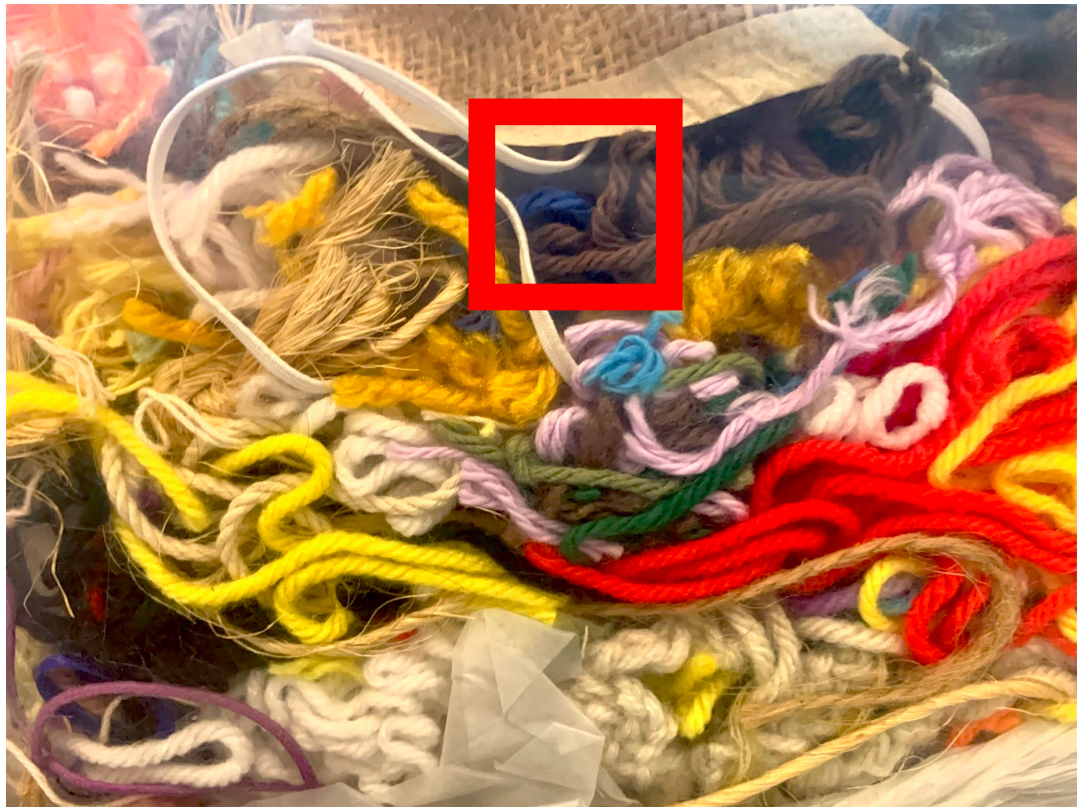
Consists of:

- 147 base pairs of DNA
- 8 histone proteins
  - 2 H2A
  - 2 H2B
  - 2 H3
  - 2 H4

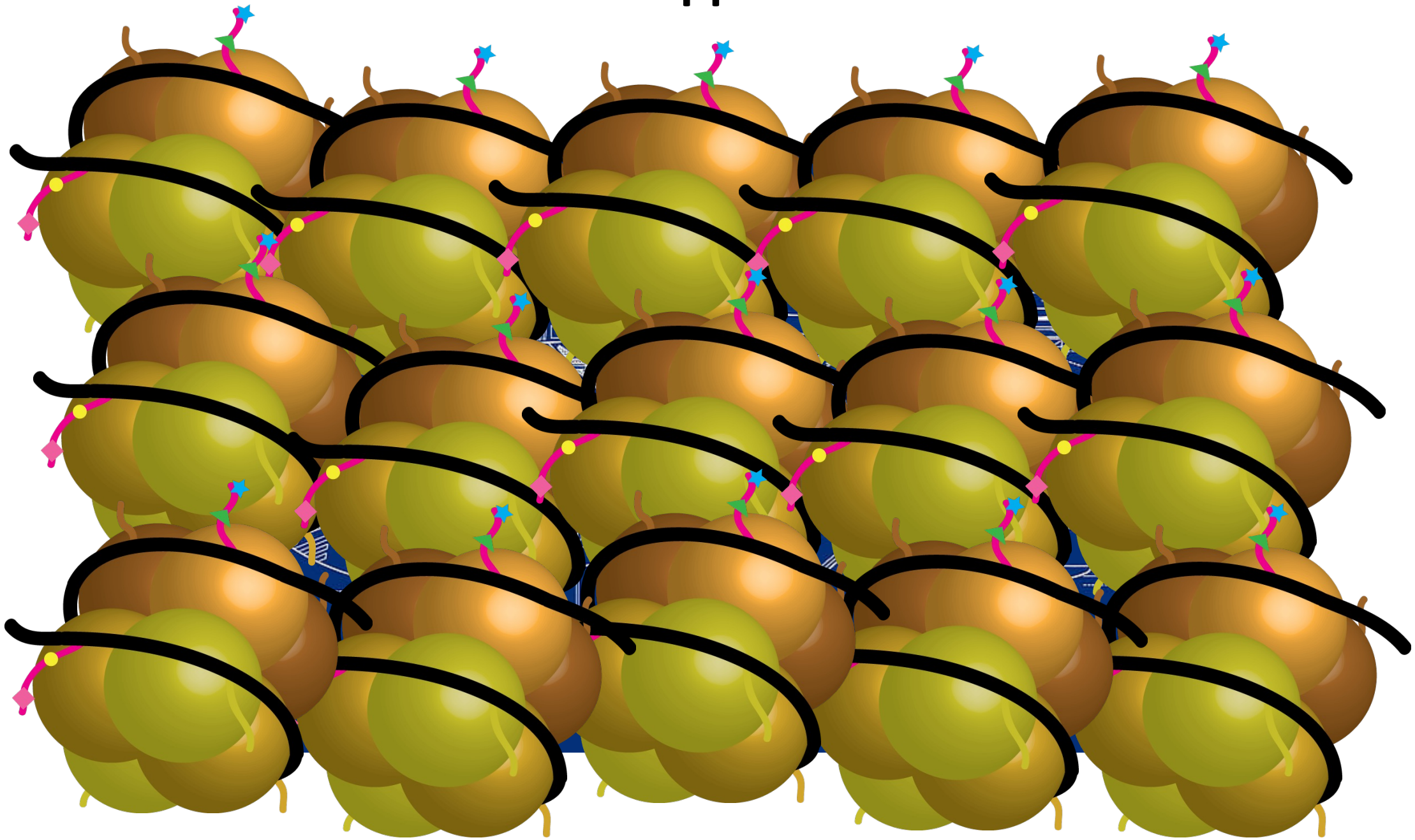


How does a nucleosome core particle affect the genome?

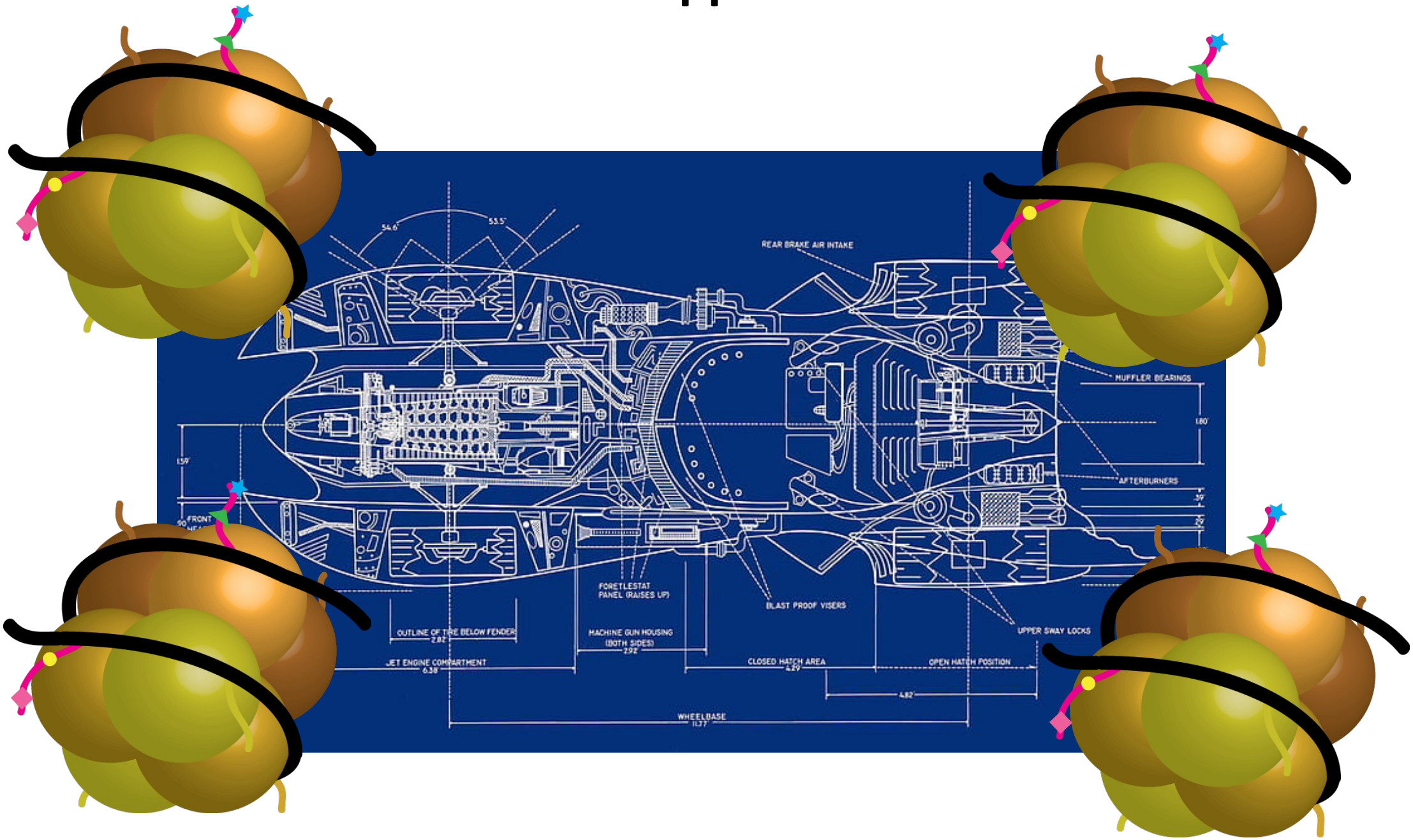
Where's the blue yarn?



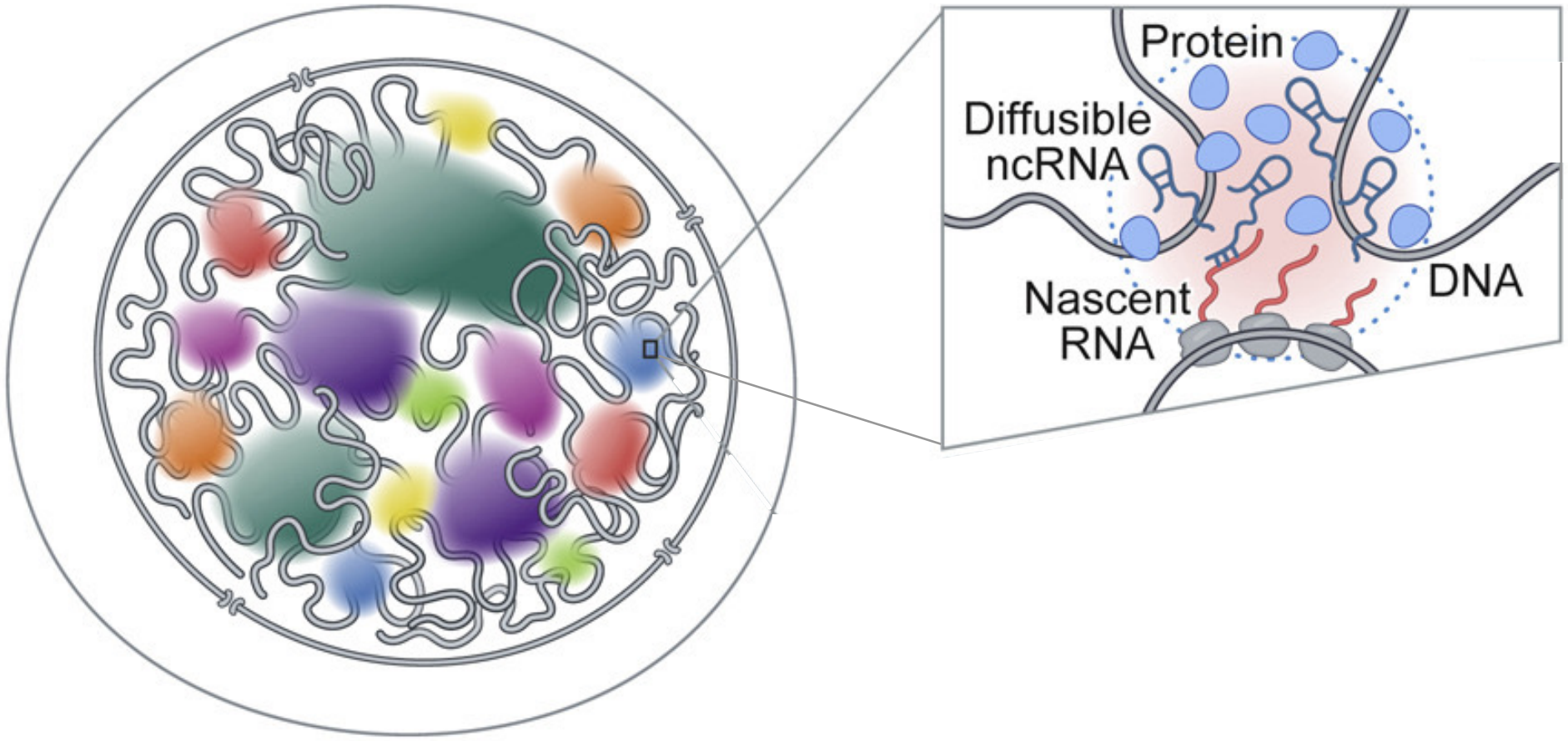
What are we supposed to build?



What are we supposed to build?

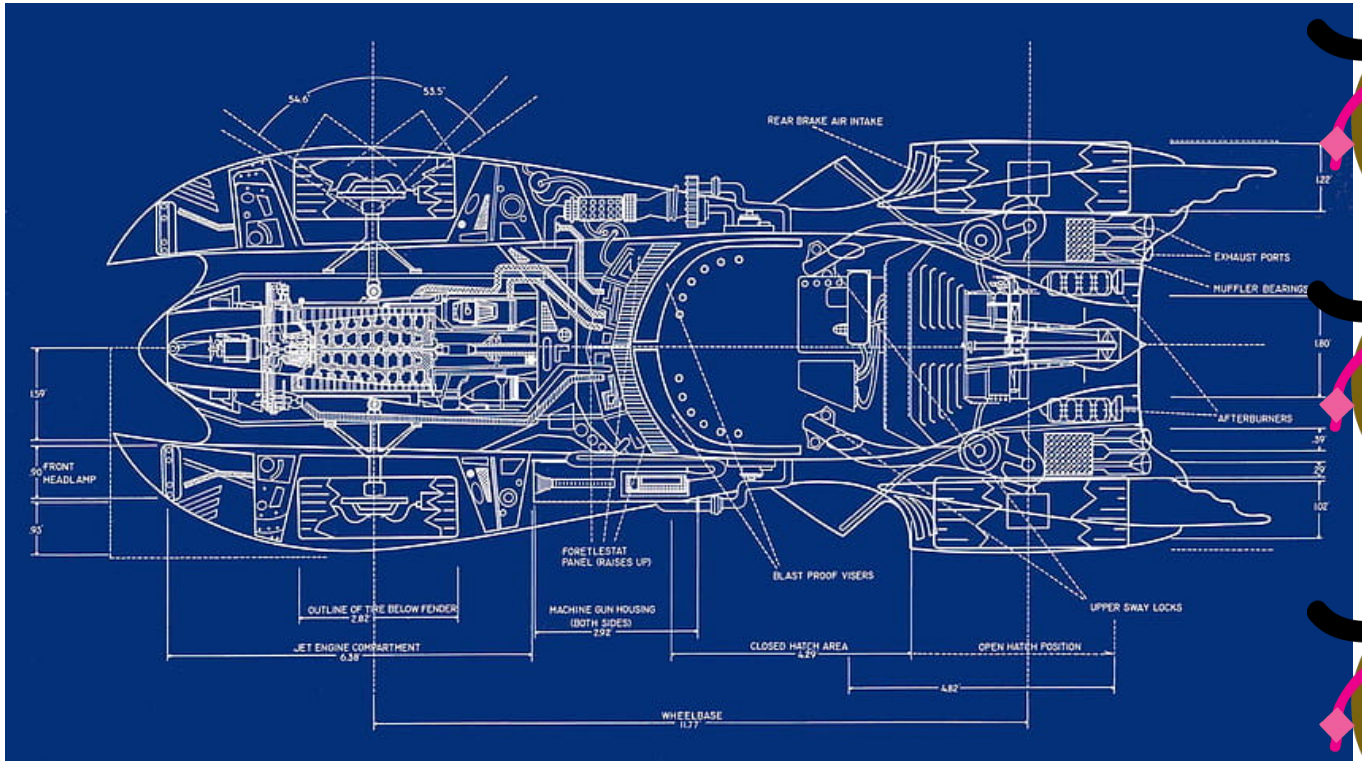


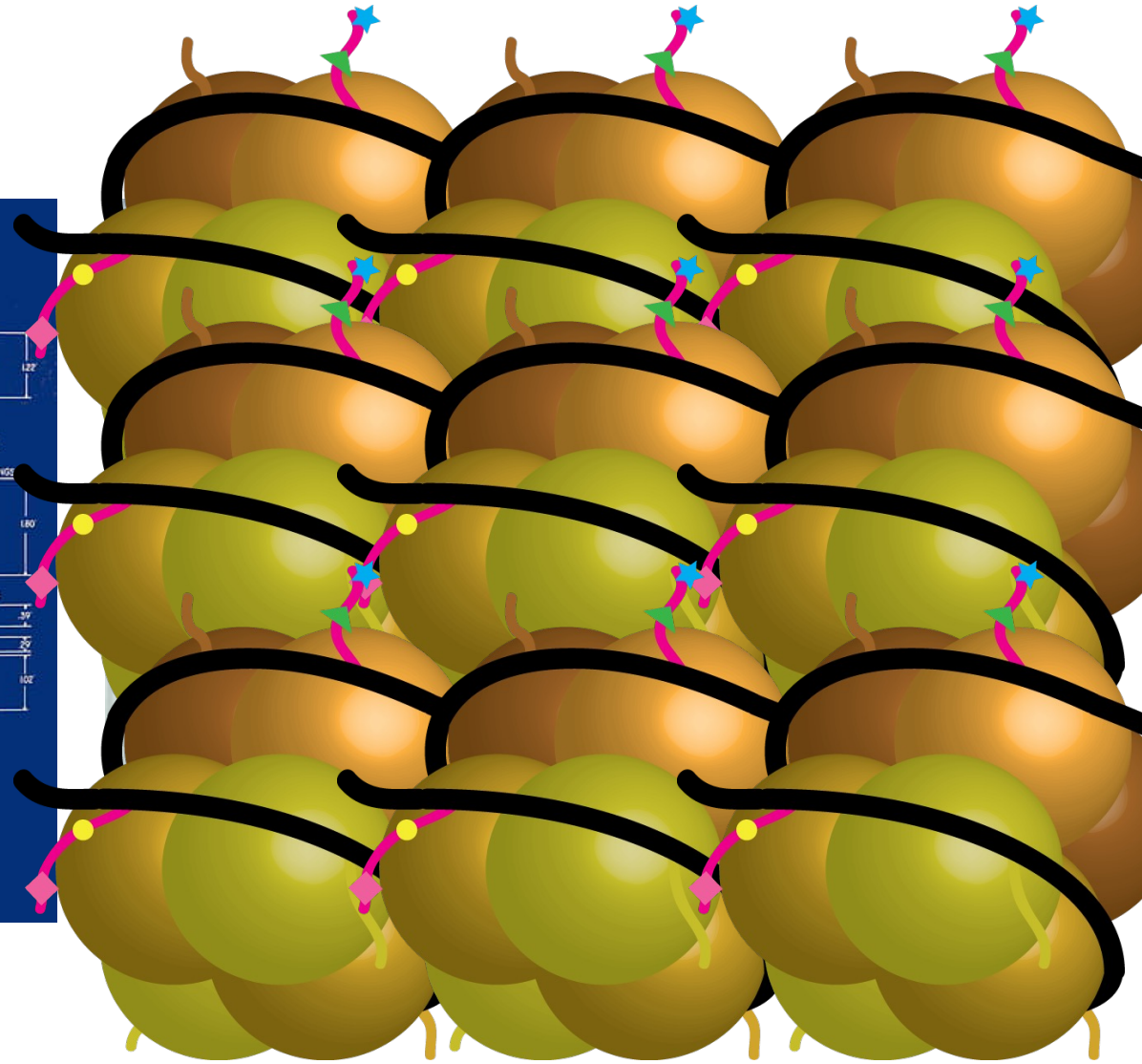
DNA is compacted in precise ways



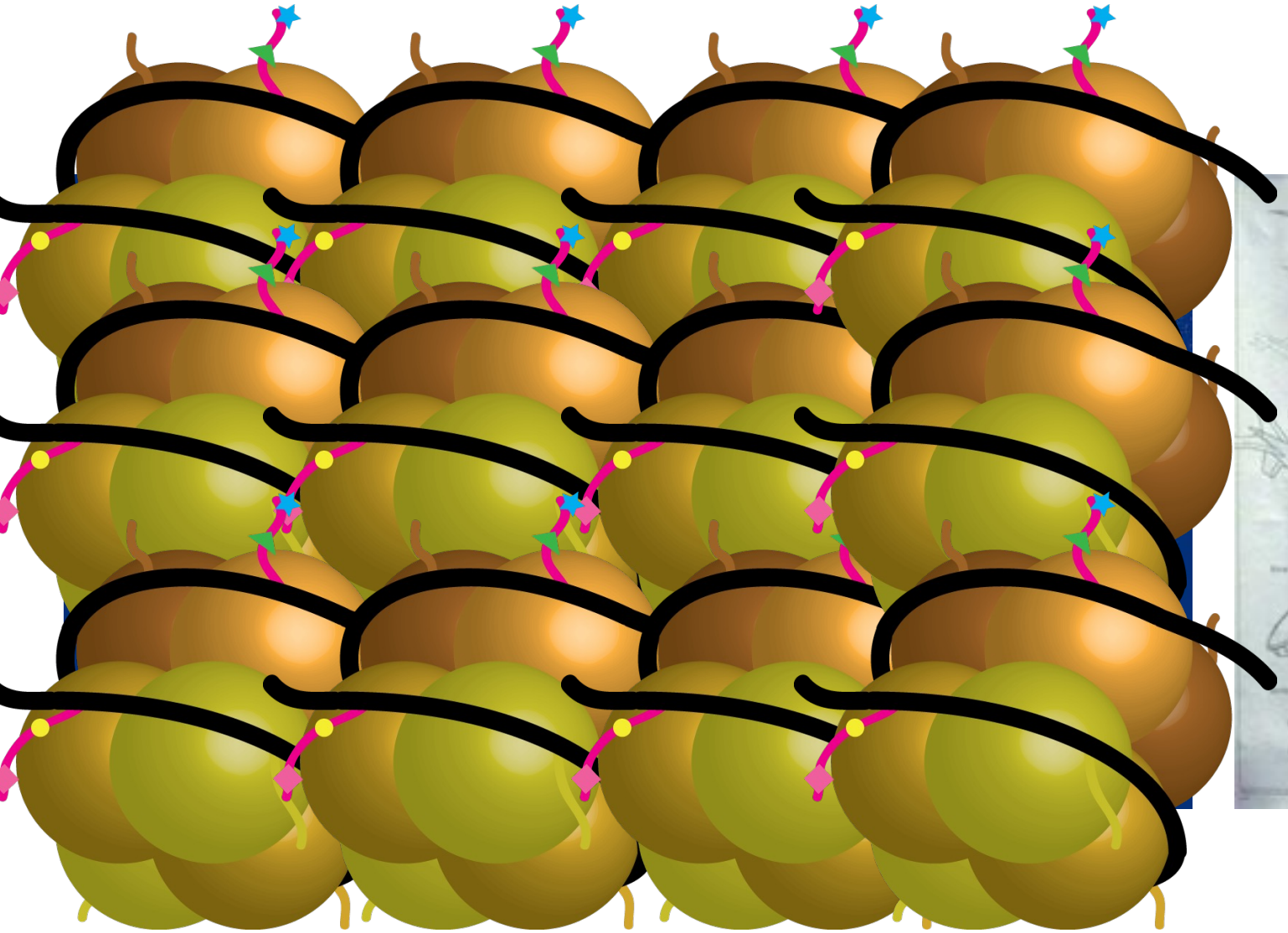
Quinodoz et al. Cell (2021)

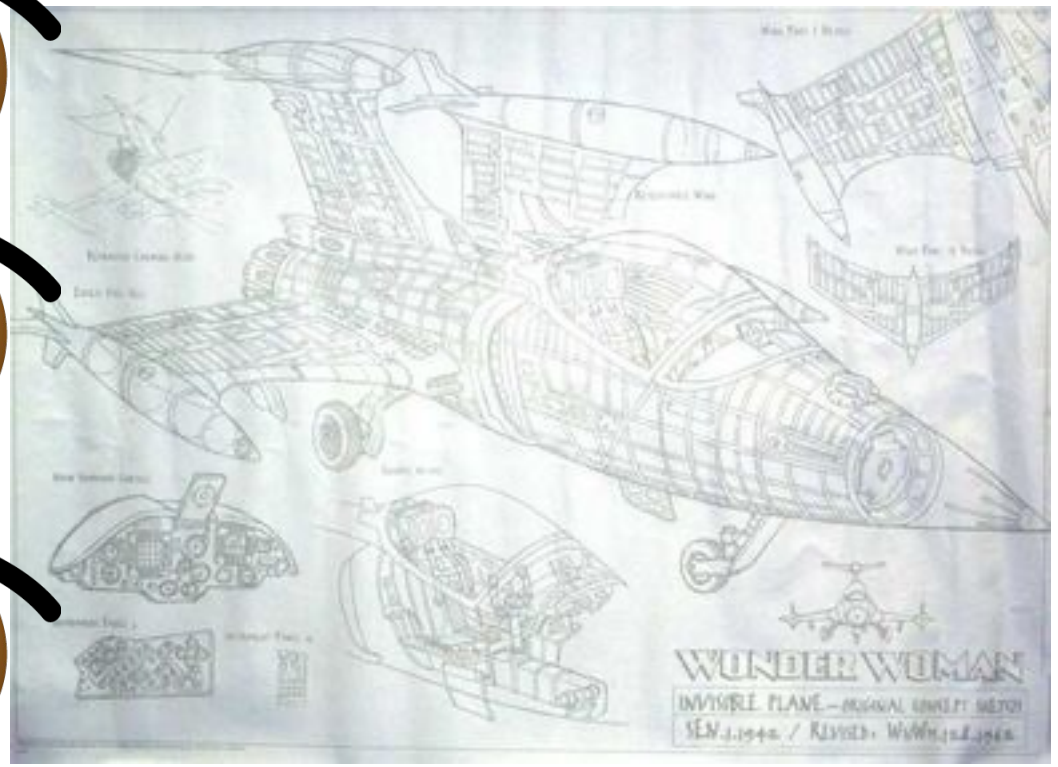
How does the DNA in our genomes create the variety of cells we have?





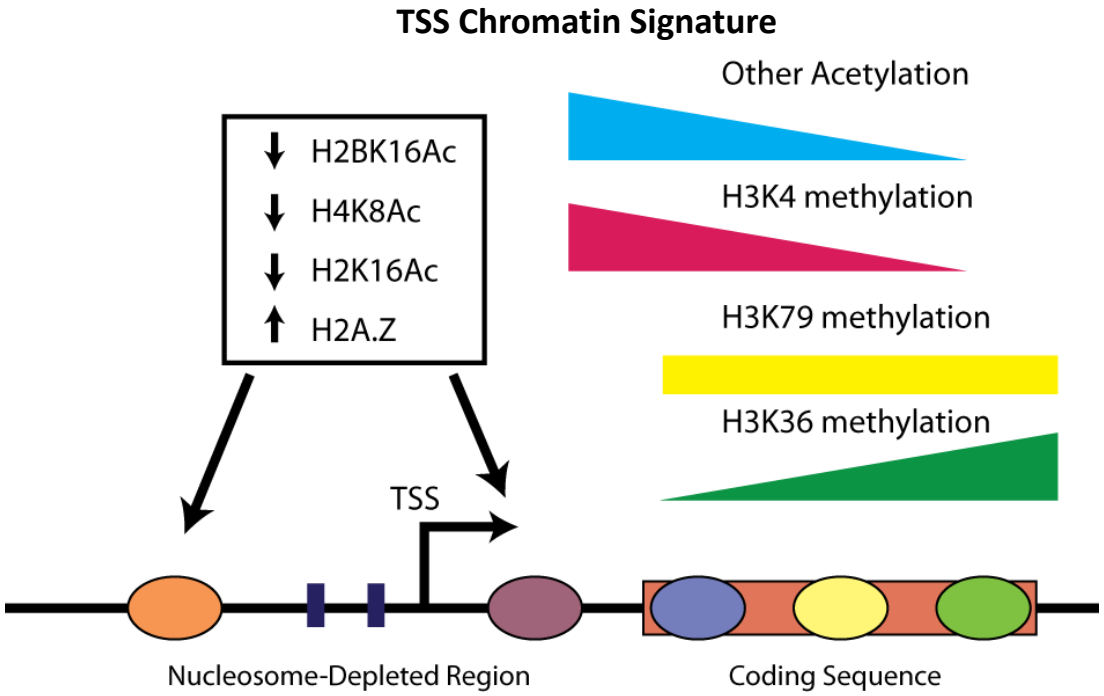
How does the DNA in our genomes create the variety of cells we have?

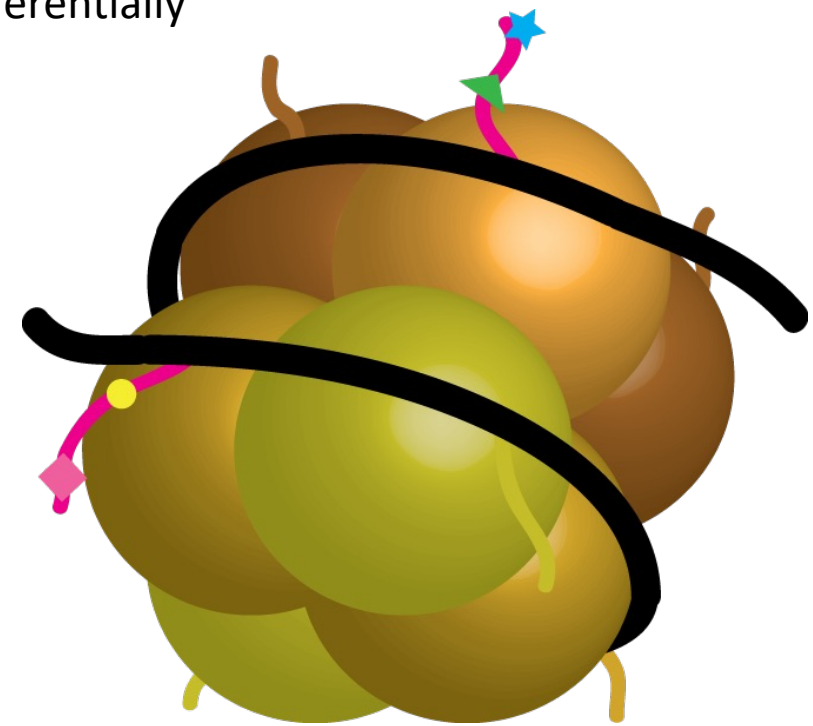




Nucleosome core particles are information dense!

- ‘Histone code’ – Strahl and Allis (2000)
  - H3K9ac – Critical for recruitment of TFIID at IFN-β
  - H3K36me3 – methylated by Set2, marker of “transcriptional elongation”
  - H3K27me3 – associated with heterochromatin, bound by Polycomb preferentially
  - etc.

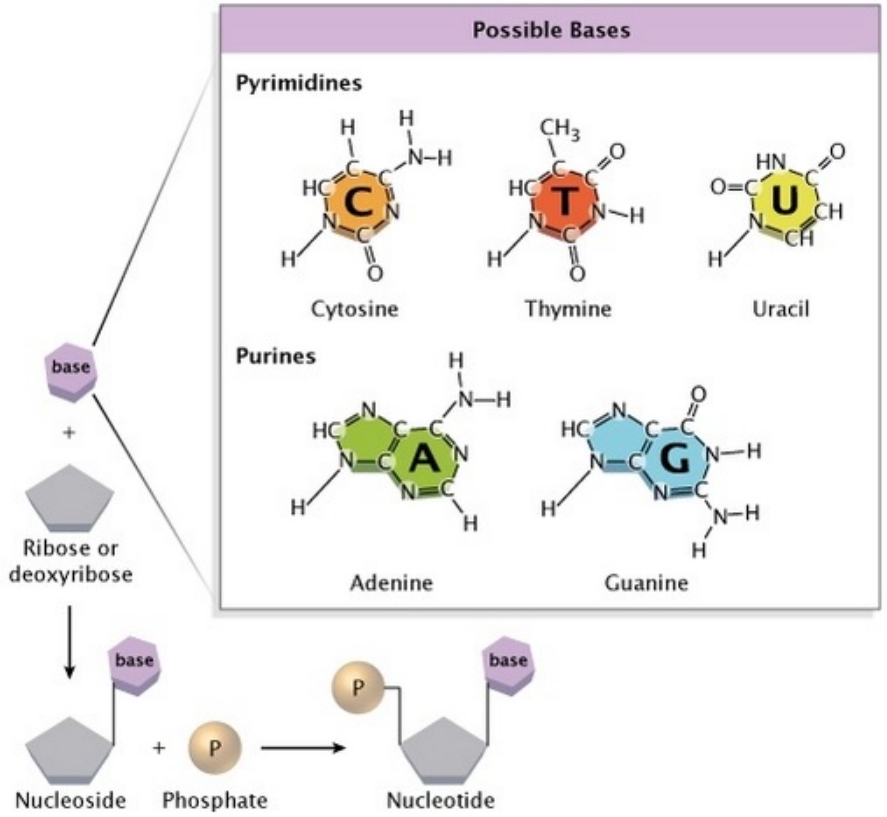


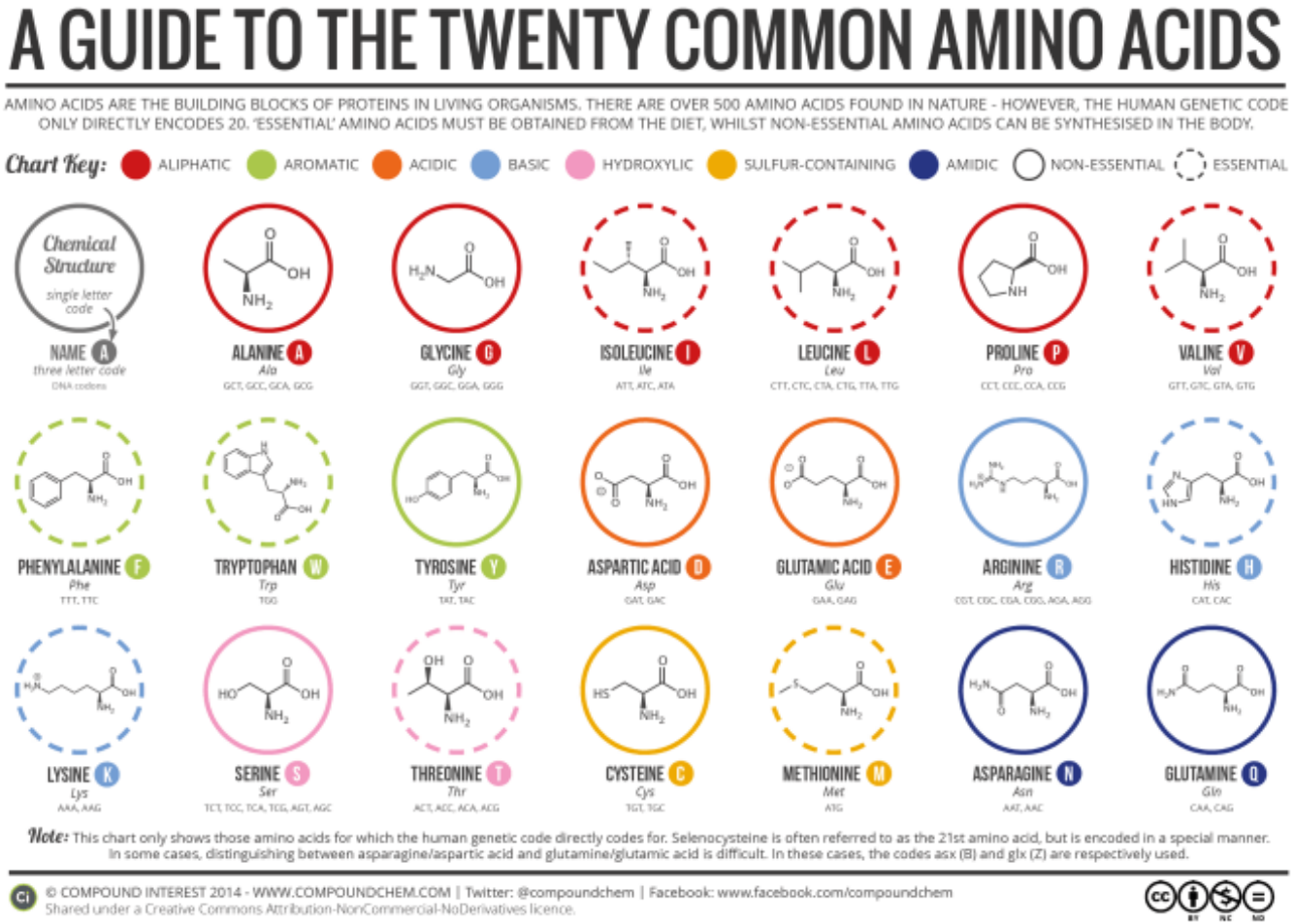


Adapted from OJ Rando (2007). “Global Patterns of Histone Modification." Current Patterns in Genetics & Development 17 (2)**: 94-99.**

From Lecture 4

Historical context: there is something else in the cell with potentially greater information density





https://www.nature.com/scitable/content/ne0000/ne0000/ne0000/ne0000/104573606/70_1_2.jpg

https://i0.wp.com/www.compoundchem.com/wp-content/uploads/2014/09/20-Common-Amino-Acids-v3.png?ssl=1

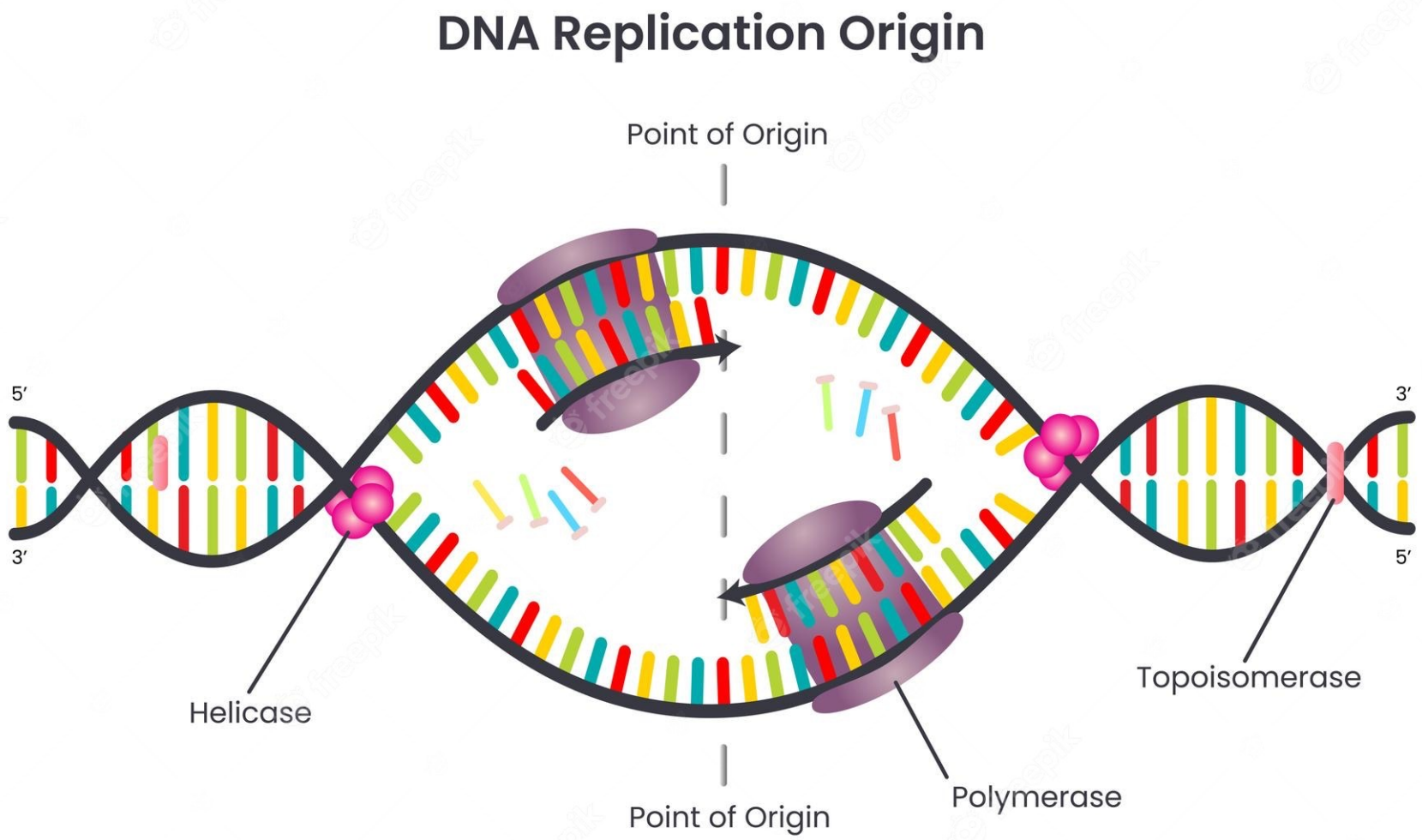
Genomic features

- There are a LOT of regulatory components
  - e.g., Codon code, 3D structure, histone protein modifications, etc.

- But less than 1% of the genome codes for genes
- What else is in there?

- Origins of Replication
- Centromeres
- Telomeres
- Enhancers
- Much, much more!

Origins of Replication



Where are human replication origins?

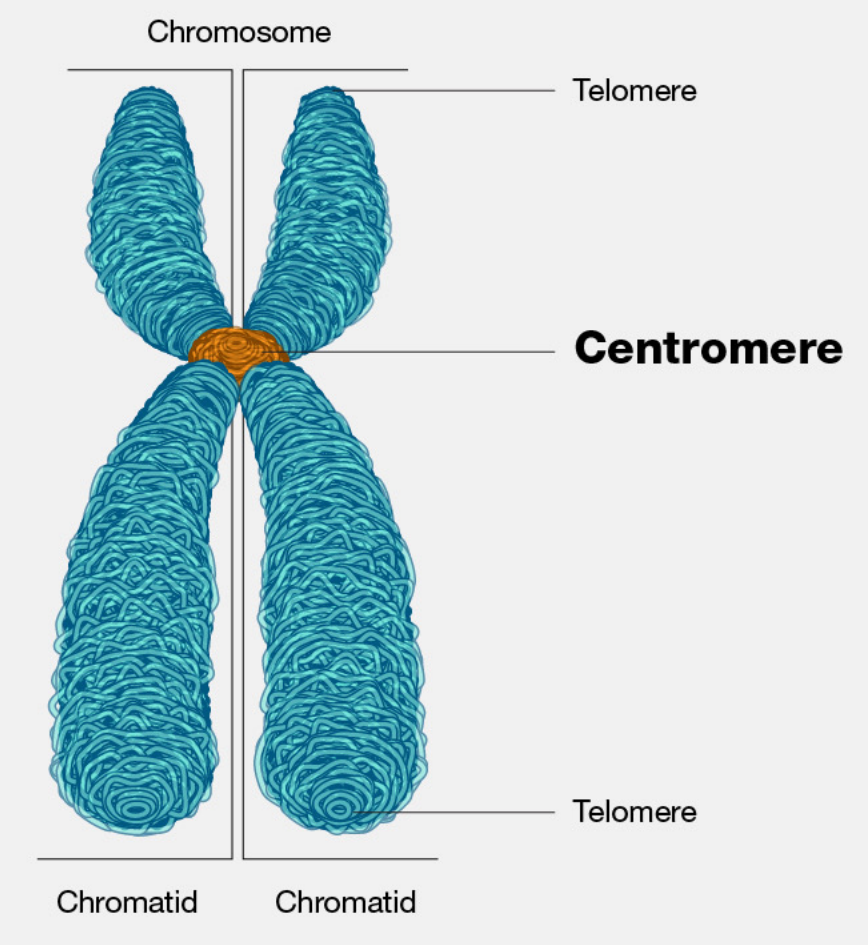
https://www.freepik.com/premium-vector/dna-replication-point-origin-replication-bubble-vector-diagram_30806078.htm

Telomeres

We know where these are!

But…

We don’t know what’s going on genomically



In humans, theses are 10-15,000 base pairs long on each end and get SHORTER every time your cells divide

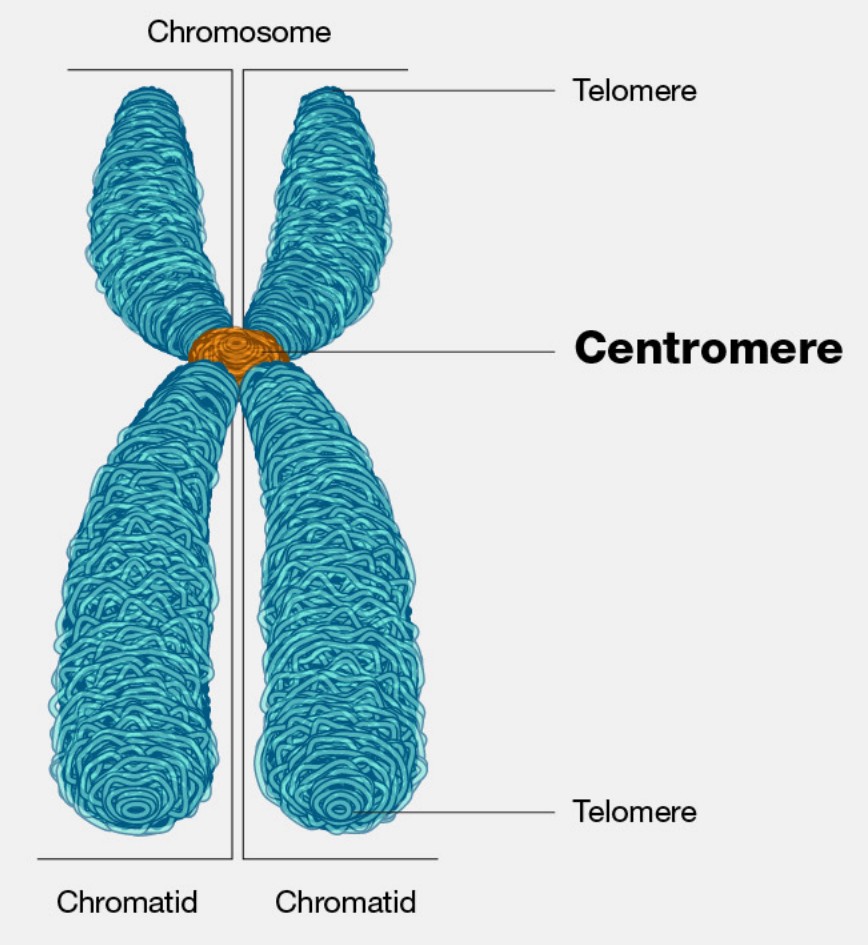
https://www.genome.gov/genetics-glossary/Telomere

Centromeres

We know where these are!

But…

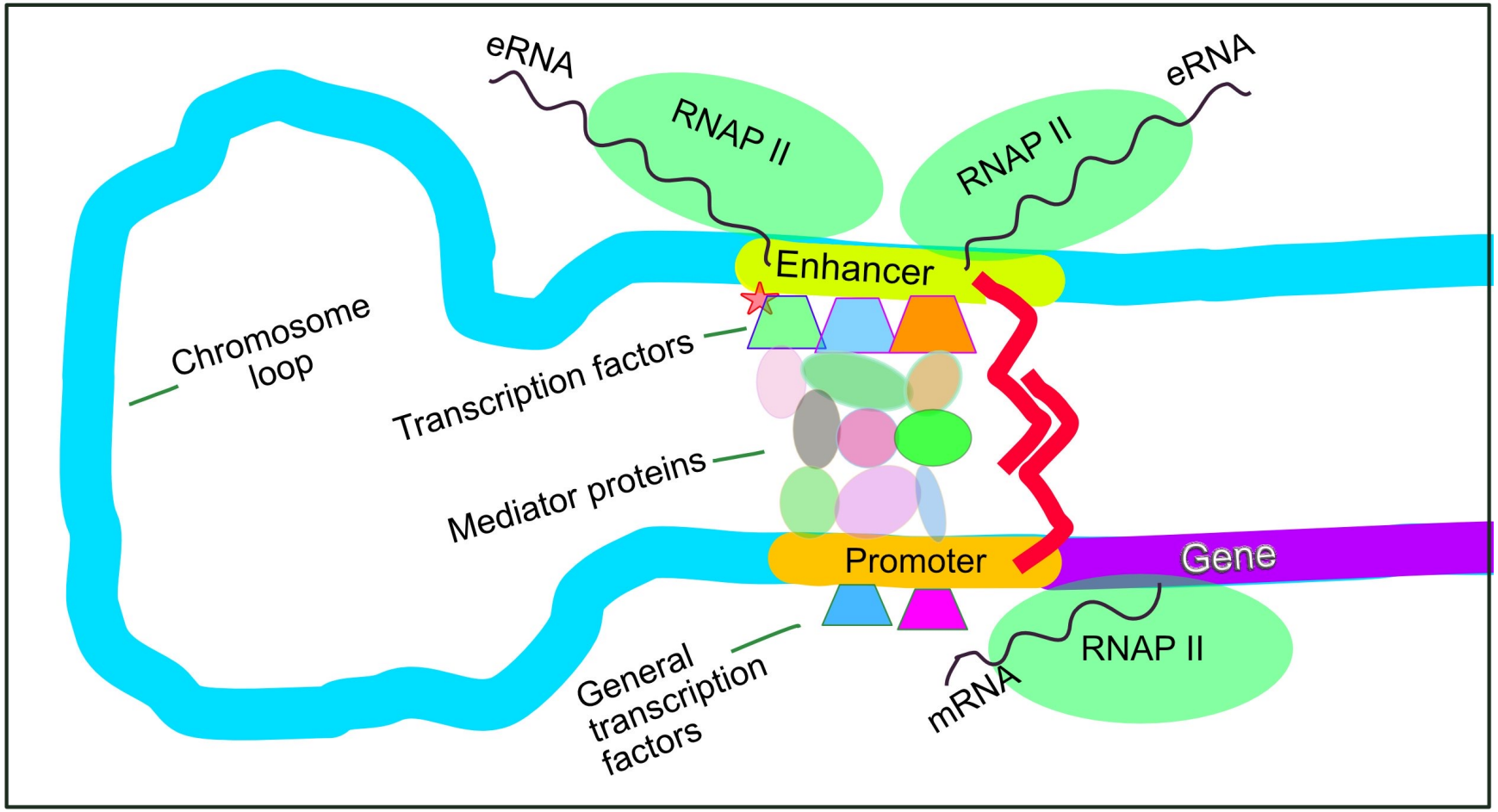
We don’t know what’s going on genomically



In humans, composed of 171 bp of repeating DNA (often 1,000’s of repeats)

https://www.genome.gov/genetics-glossary/Telomere

Enhancers



We DON'T know where these are…
We also don't know how they work!

fighting words at scientific conferences

https://en.wikipedia.org/wiki/Enhancer_%28genetics%29#/media/File:Regulation_of_transcription_in_mammals.jpg

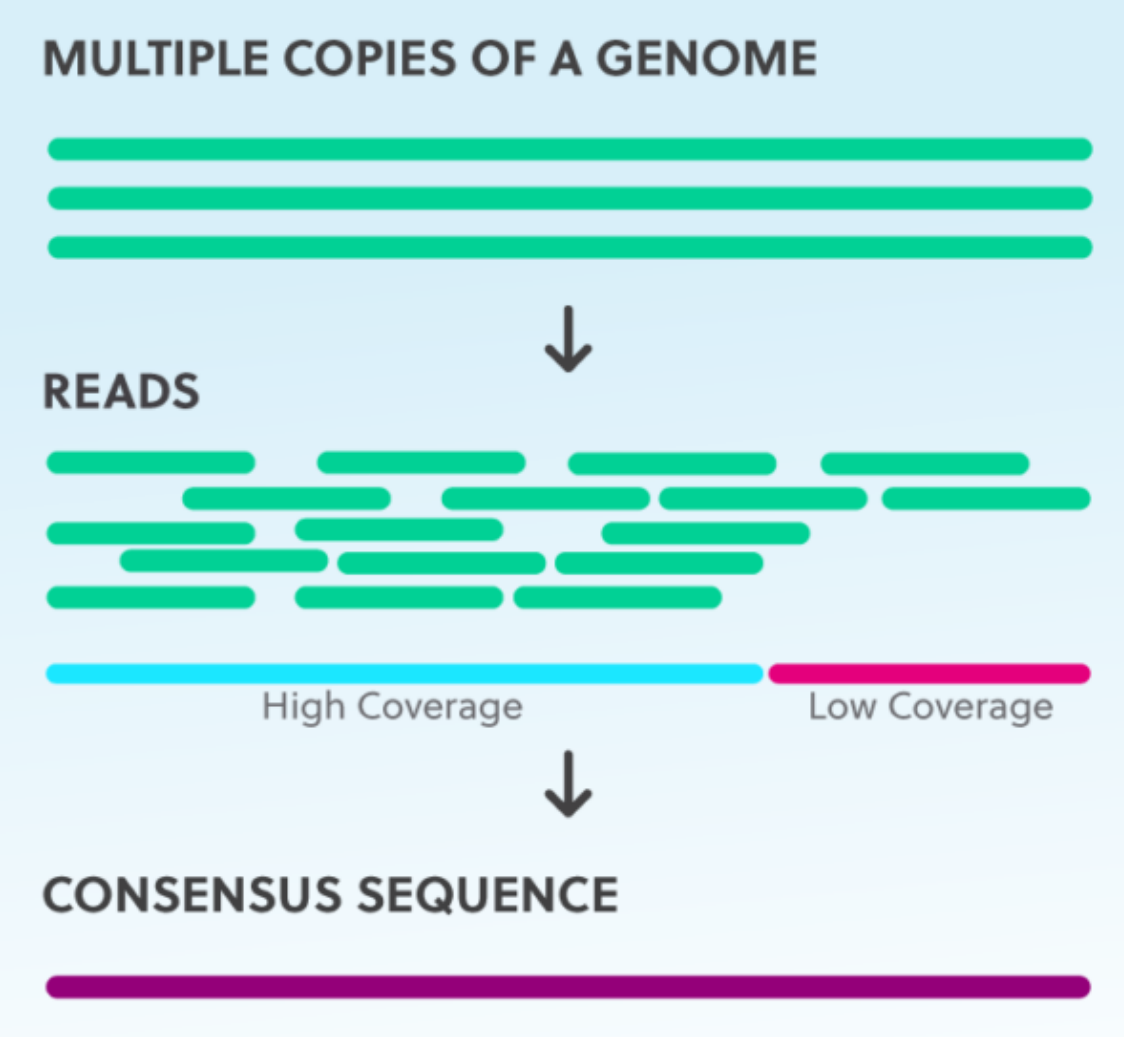
Genomic Assays (or how do we measure this?)

- Whole Genome Sequencing (WGS)
- ChIP-seq
- ATAC-seq
- RNA-seq (covered later)

Other ‘omics (not covered here)

- Proteomics – what are the proteins doing?
- Metabolomics – what are the small molecules doing
- Transcriptomics – what is the RNA doing?
- Microbiomics – what is the bacteria doing?
  - Did you know you’re composed of 1:10, human to bacteria?

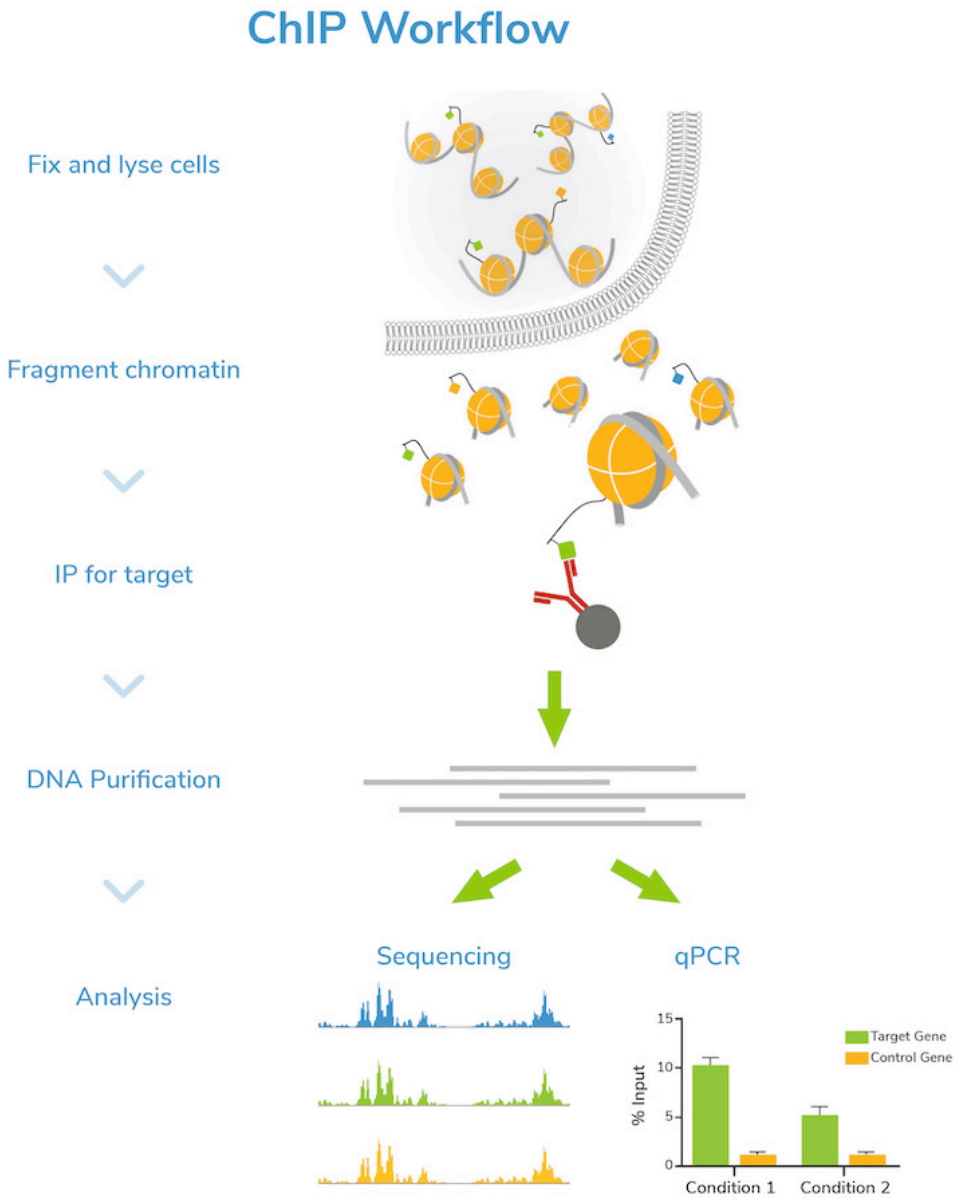
Whole Genome Sequencing (WGS)



- 30X coverage considered gold standard for data quality
- This and microarray-based technology serve as the foundation for most genome-wide association studies (GWAS)

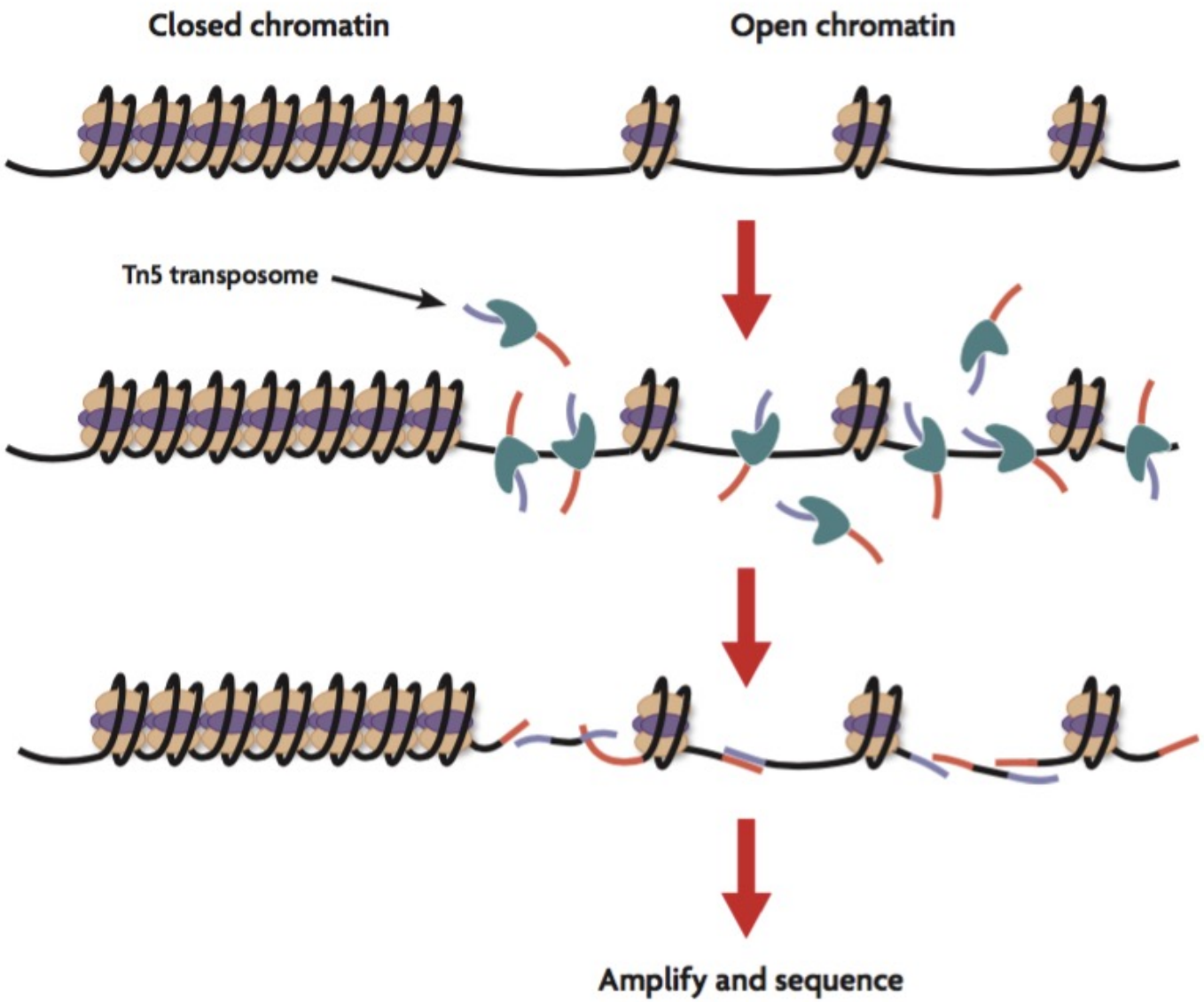
https://irepertoire.com/wp-content/uploads/2020/05/NGS-Considerations-Sequencing-Workflow.png

Chromatin Immunoprecipitation (ChIP)



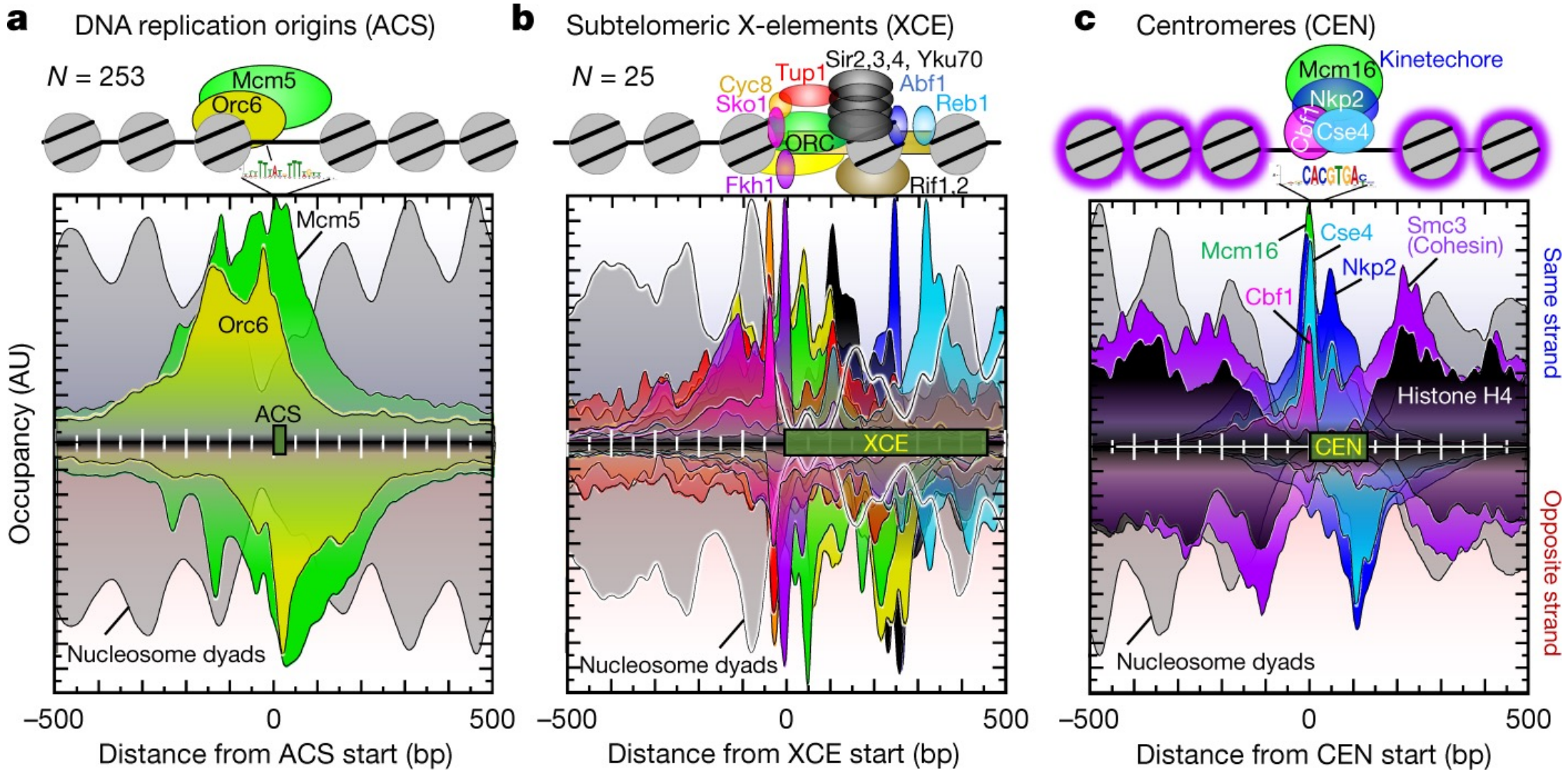
https://www.epicypher.com/content/Blog/Figures/ChIP-blog-figure-1.jpg

Transposase accessible chromatin (ATAC-seq)



https://www.activemotif.com/uploads/images/web_site/services-hp-dec2015/open-chromatin.png

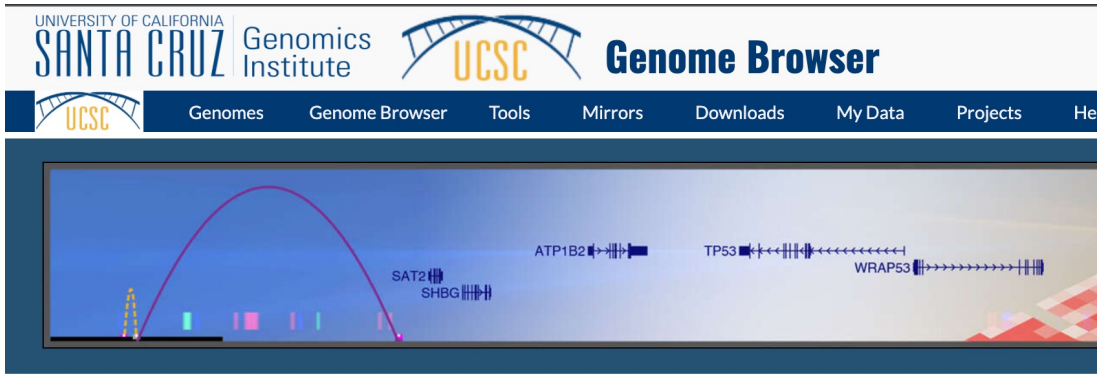
What do Origins/Telomeres/Centromeres look like with these technologies?



Rossi et al. Nature (2021)

UCSC Genome Browser

https://genome.ucsc.edu/



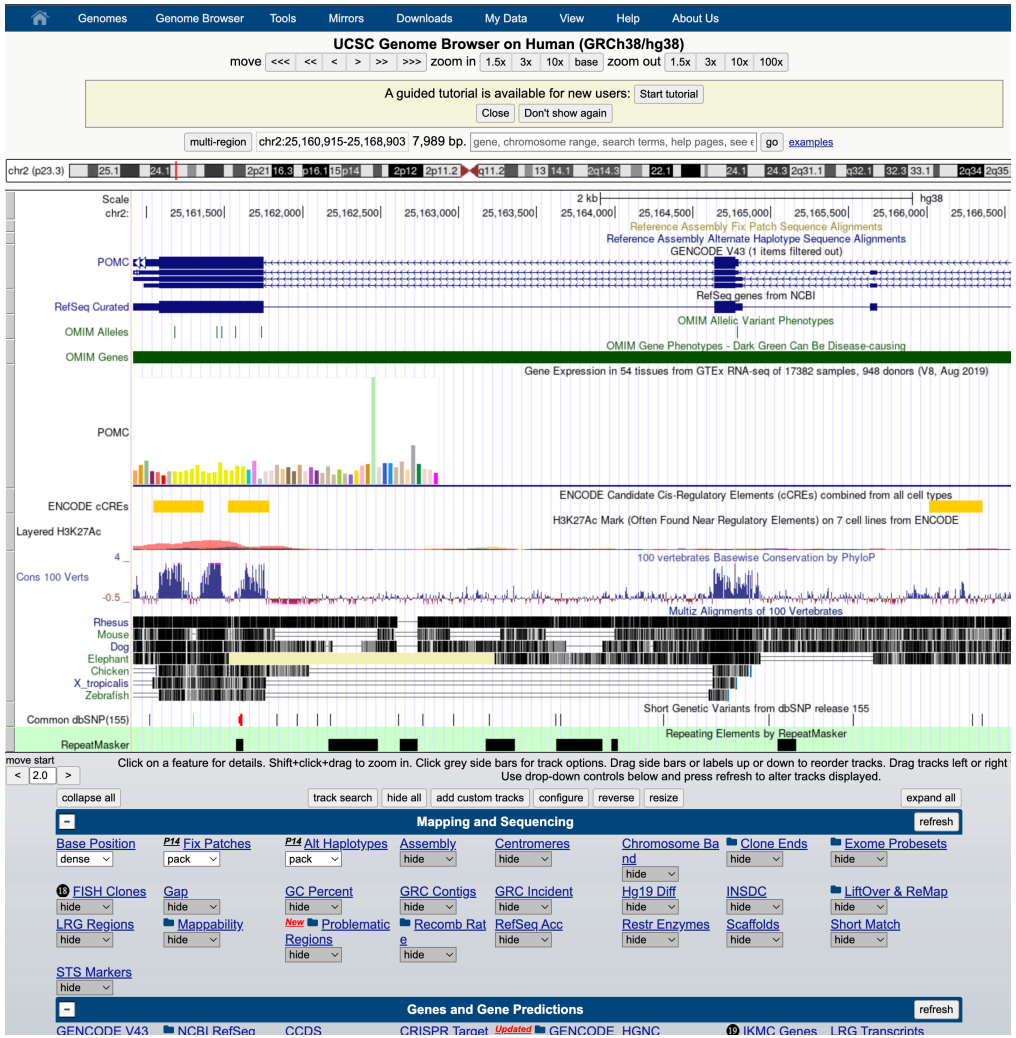
Tools

hg38

hg19

mm39

- **Genome Browser** - Interactively visualize genomic data
- **BLAT** - Rapidly align sequences to the genome
- **In-Silico PCR** - Rapidly align PCR primer pairs to the genome
- **Table Browser** - Download and filter data from the Genome Browser
- **LiftOver** - Convert genome coordinates between assemblies
- **REST API** - Returns data requested in JSON format
- **Variant Annotation Integrator** - Annotate genomic variants
- **More tools...**



Random Notes (things I didn't cover but are cool to know):

- Archaea are as genetically divergent from bacteria as we are