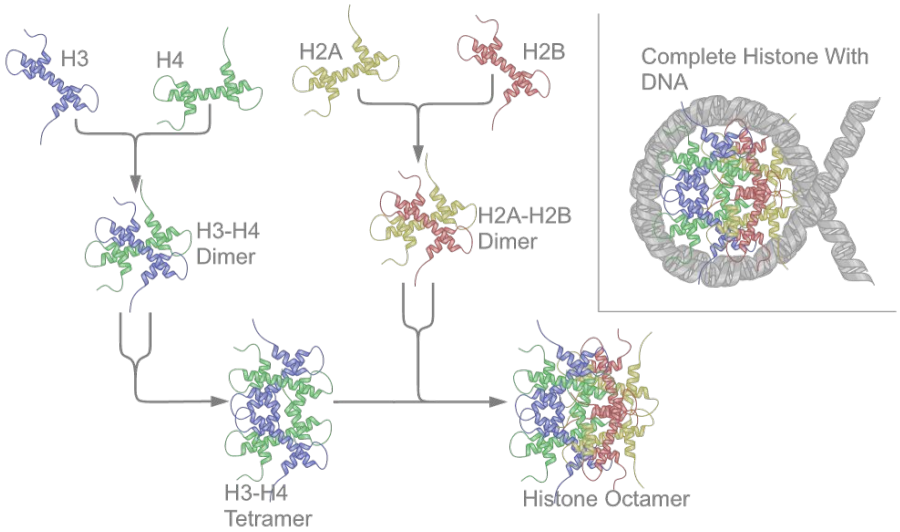
Eukaryotic Genome Structure and Features

Genome Biology (BIOL7263)
24Oct24

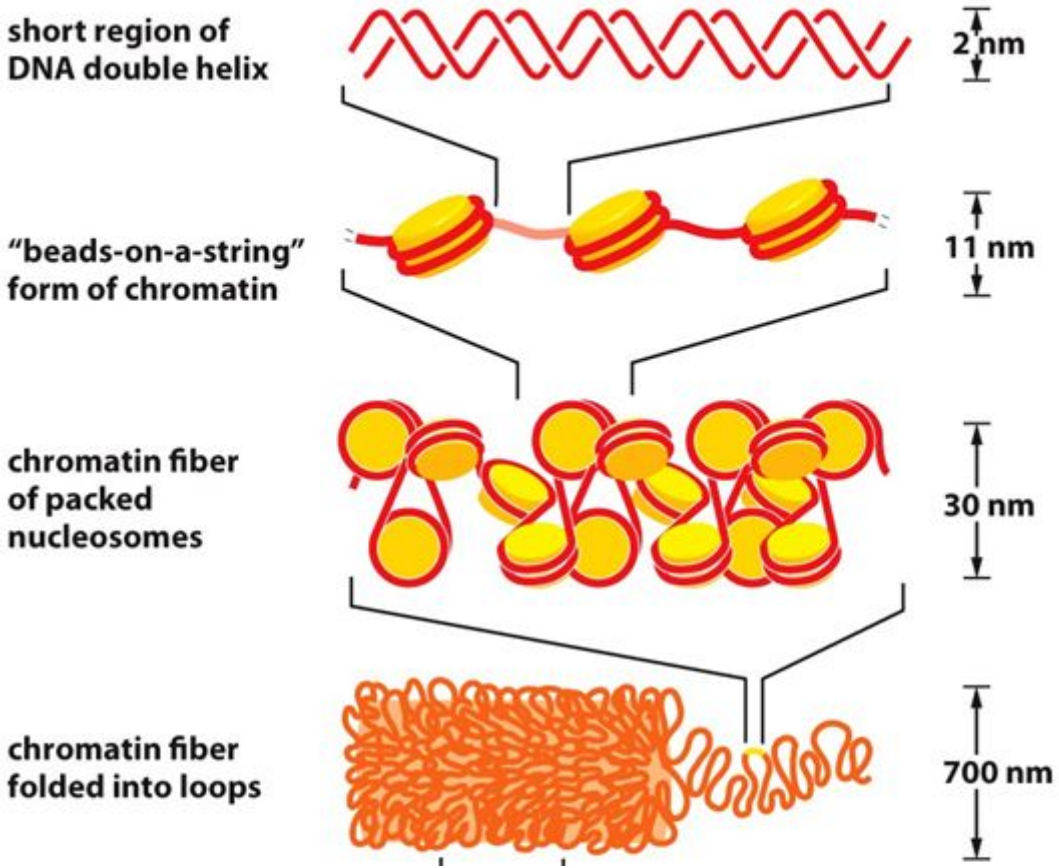
The genomes of Eukaryotes is organized into linear chromosomes

- Chromosomes are made up of chormatin - DNA and protein complex
- DNA is wrapped around histones composed of 8 protein molecules



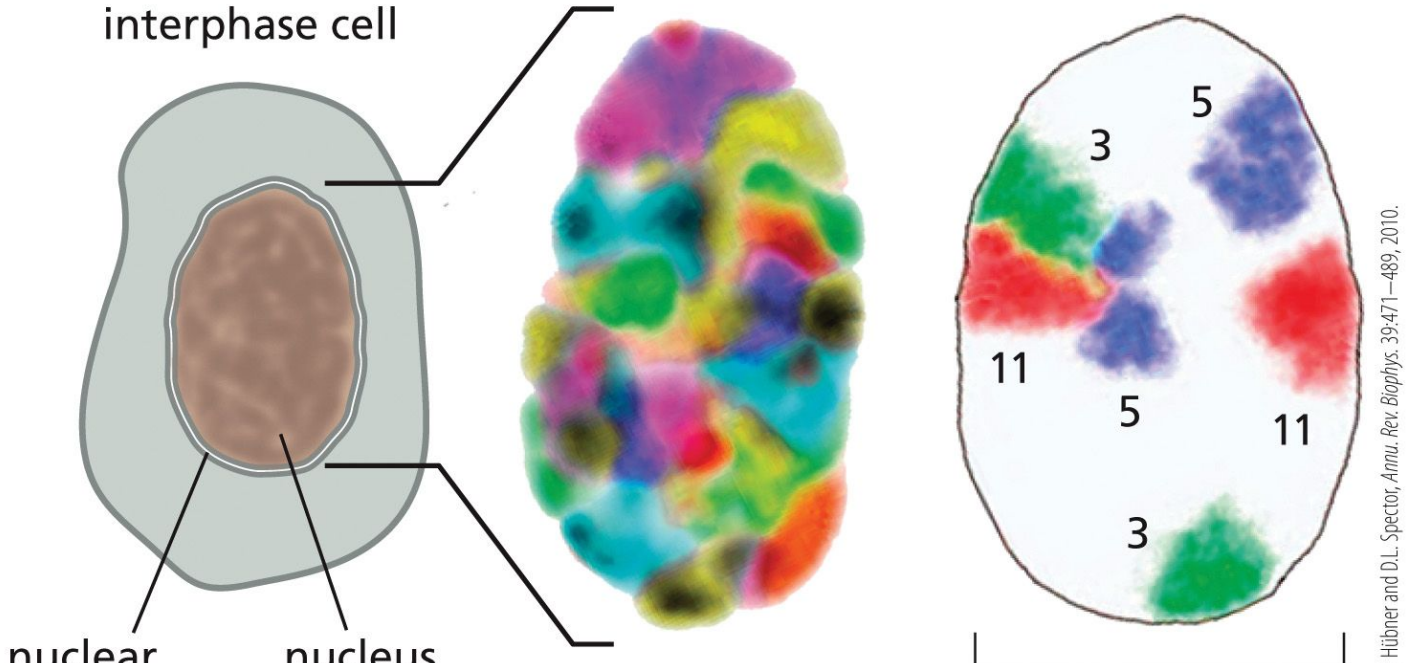
- Histone proteins are positively charges and highly conserved among organisms

These “beads-on-a-string” are further condensed into chromatin fibers

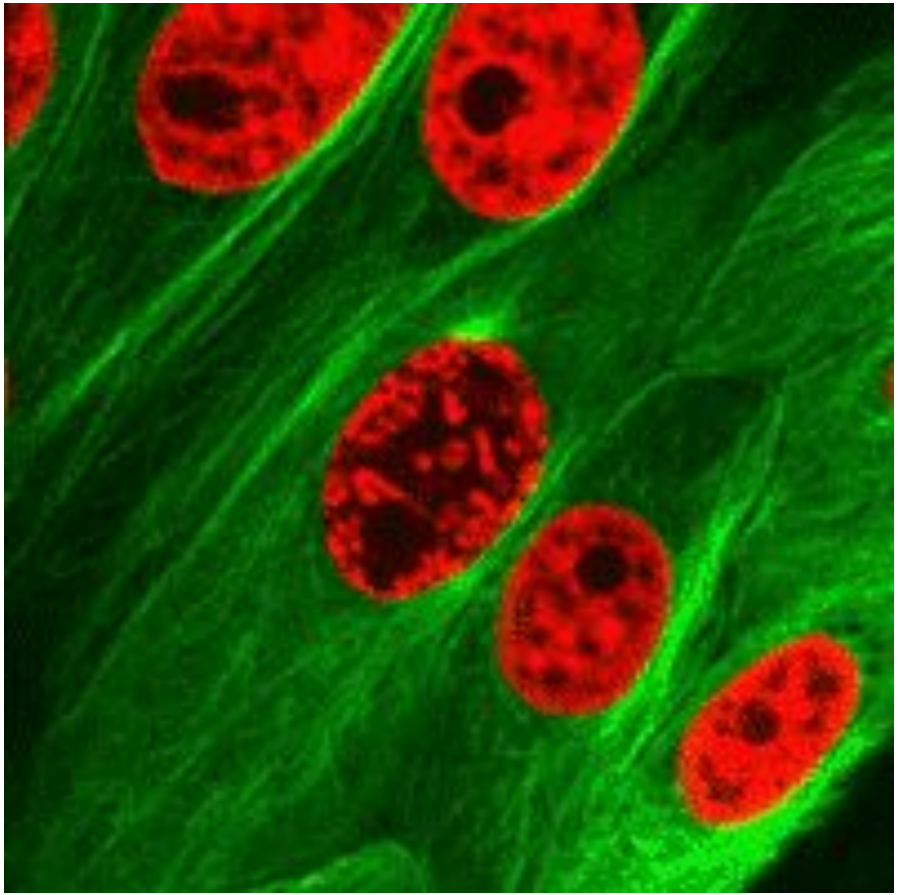


Chromatin is dynamic

- The packing and condensation of chromatin changes with **cell cycle stage** and **transcriptional state**.



Metaphase



- Chromosomes duplicated and highly condensed
- Essential step in the distribution of the genome to daughter cells

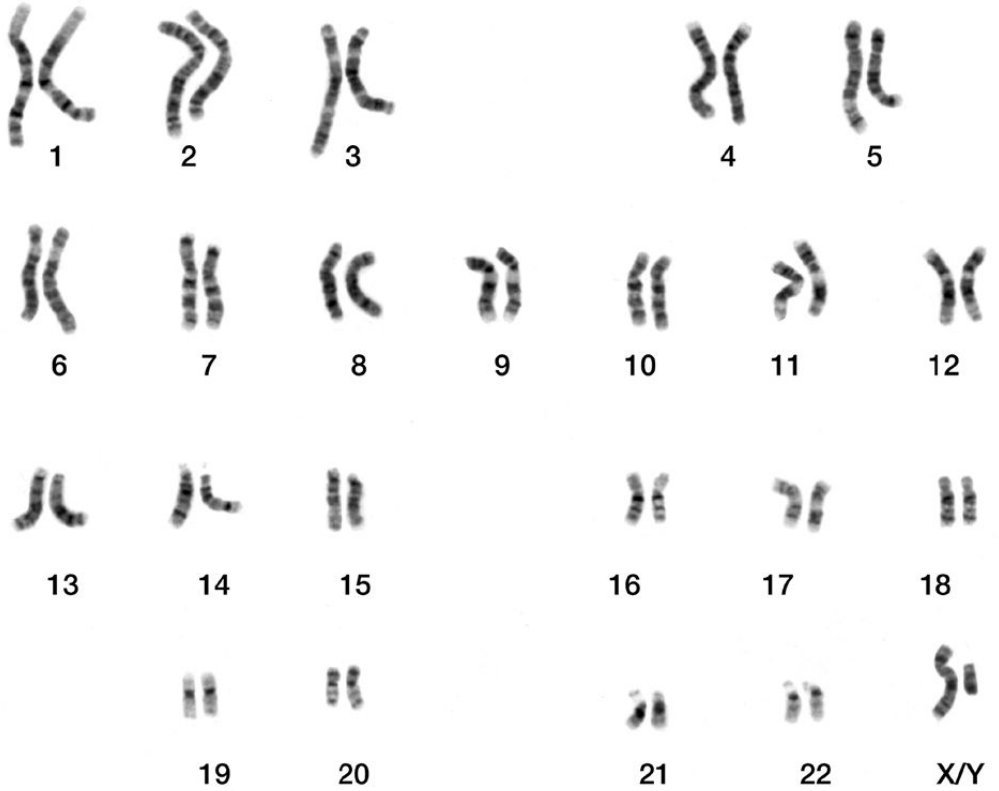
Metaphase chromosome banding and karyogram

TABLE 7.1 STAINING TECHNIQUES USED TO PRODUCE CHROMOSOME BANDING PATTERNS

Technique	Procedure	Banding pattern
G-banding	Mild proteolysis followed by staining with Giemsa	Dark bands are AT-rich Pale bands are GC-rich
R-banding	Heat denaturation followed by staining with Giemsa	Dark bands are GC-rich Pale bands are AT-rich
Q-banding	Stain with quinacrine	Dark bands are AT-rich Pale bands are GC-rich
C-banding	Denature with barium hydroxide and then stain with Giemsa	Dark bands contain constitutive heterochromatin (see Section 10.1)

Human Karyotype

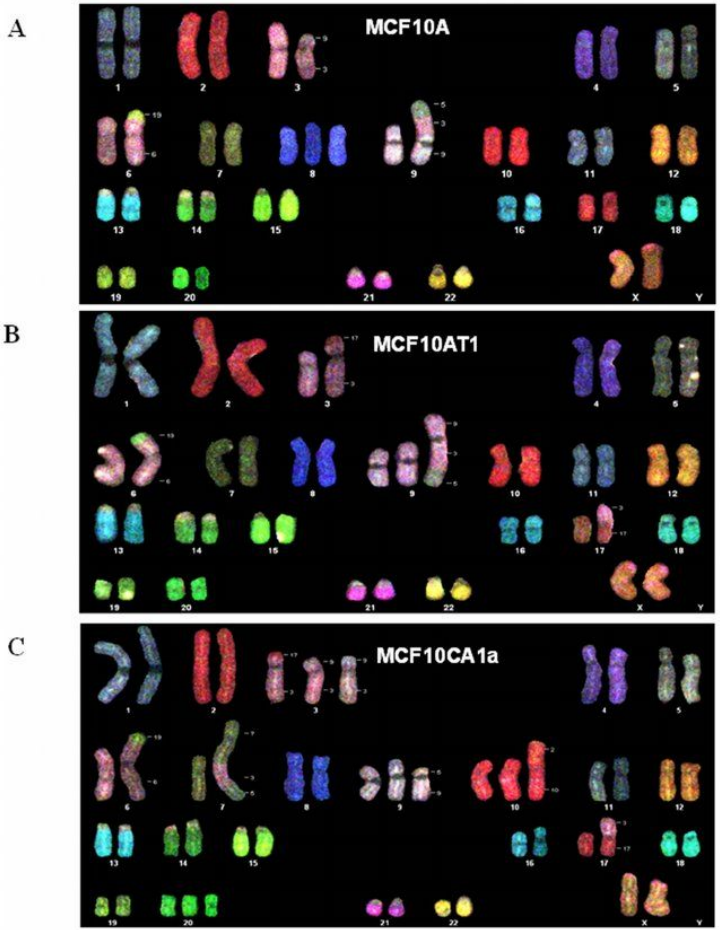
- G-band staining



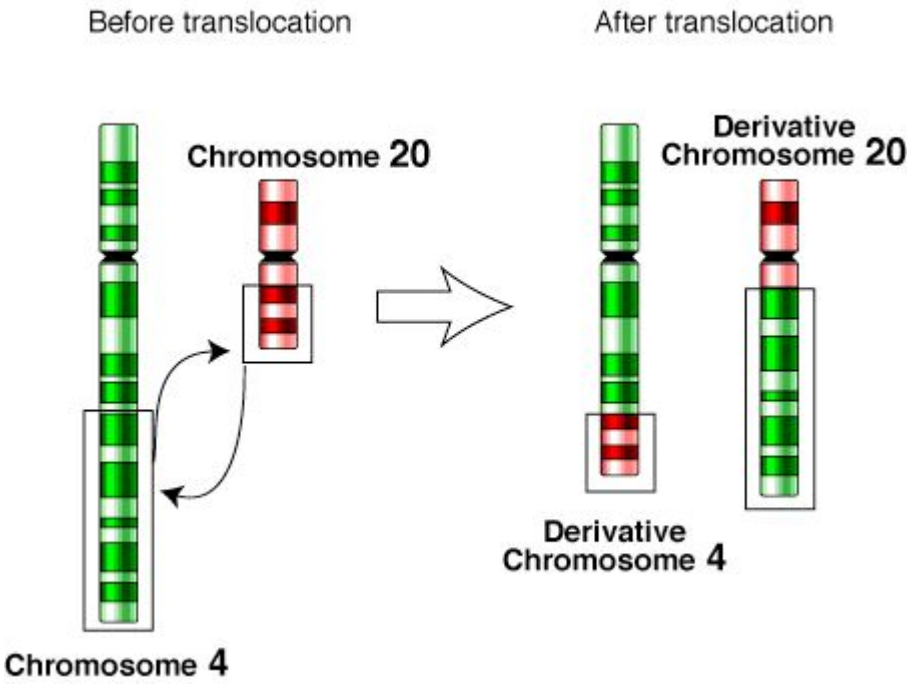
https://en.wikipedia.org/wiki/Karyotype#/media/File:DNA_human_male_chromosomes.gif

Karyotyping is an important diagnostic tool

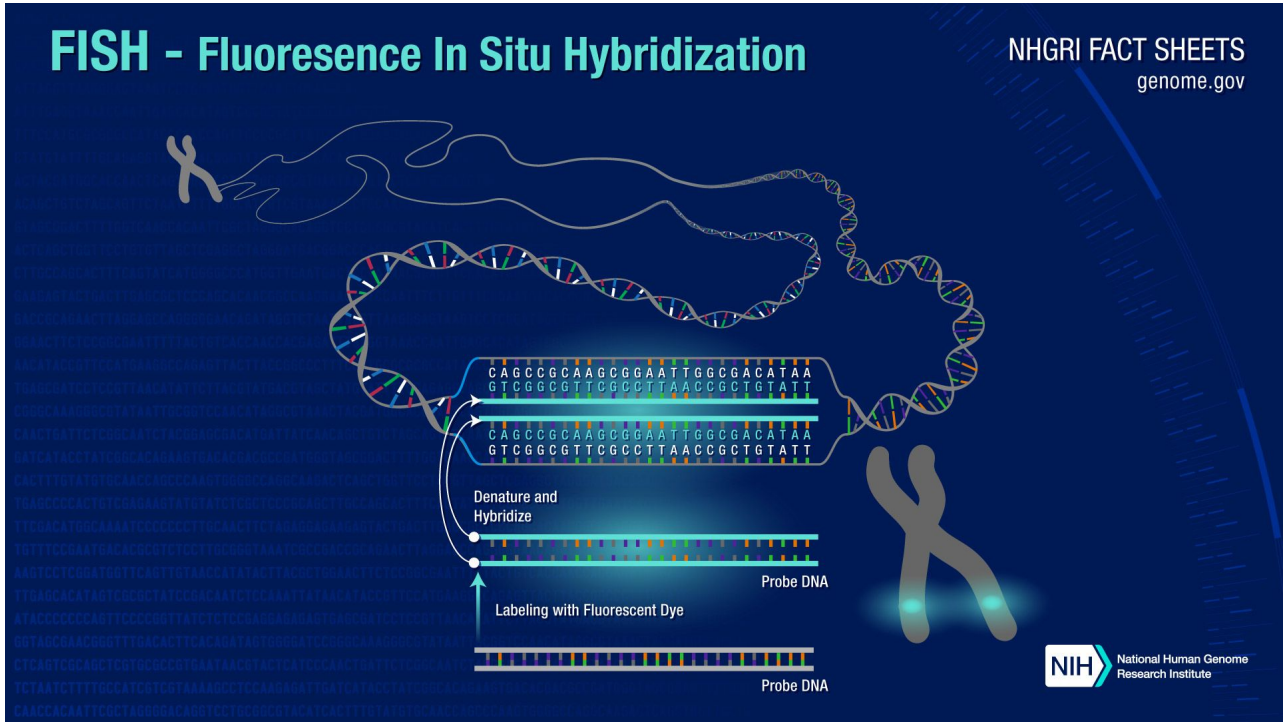
- Cancer progression often involves non-disjunction and abnormal chromosome numbers (Aneuploidy)



Karyotyping also can be used to detect translocation events

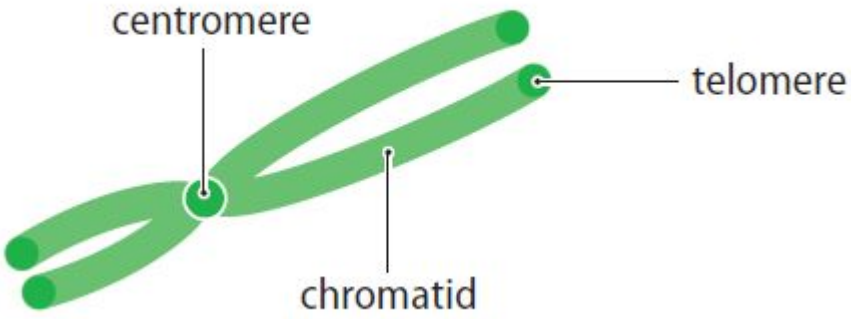


Fluorescence *in situ* hybridization (FISH) is an important tool for mapping specific sequences to chromosome structure



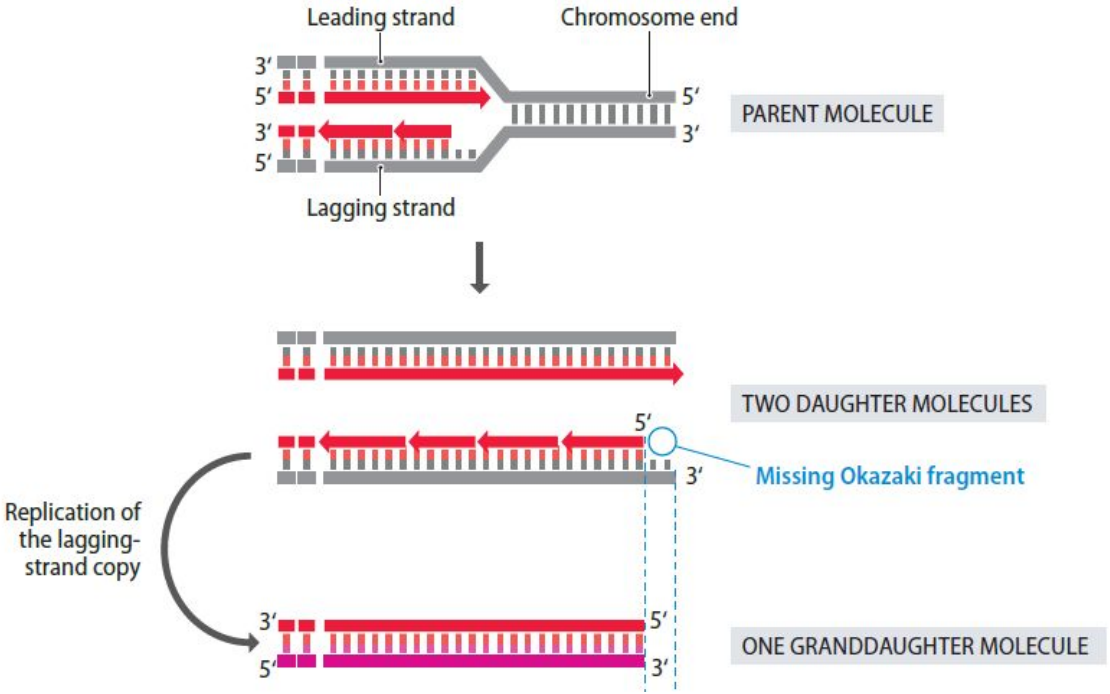
Metaphase chromosome features

- **Chromatid** - one copy of the duplicated chromosome
- **Telomere** - Highly repetitive sequences at the end of the chromosome that protect against degradation.
- **Centromere** - specialized DNA sequences and associated proteins essential for disjunction of the duplicated chromosomes.



Telomeres

- The replication of linear DNA molecules results in progressive shortening of the chromosomes



Telomeres

- The enzyme telomerase extends the chromosome end sequence by reverse transcribing an RNA template with repeating sequence of bases

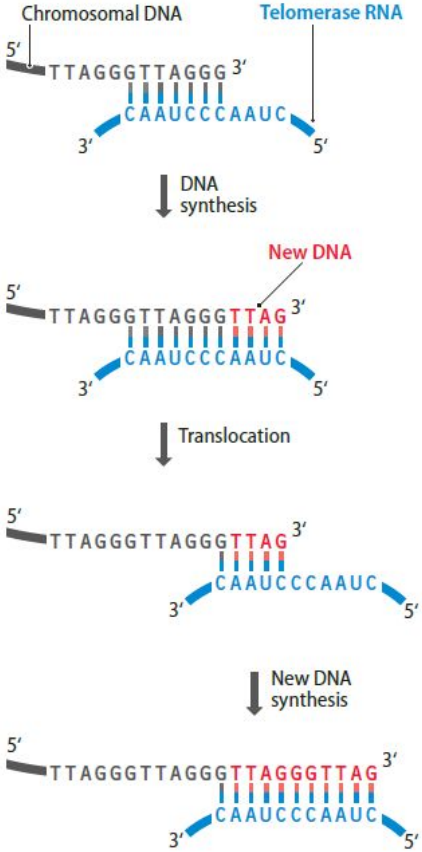
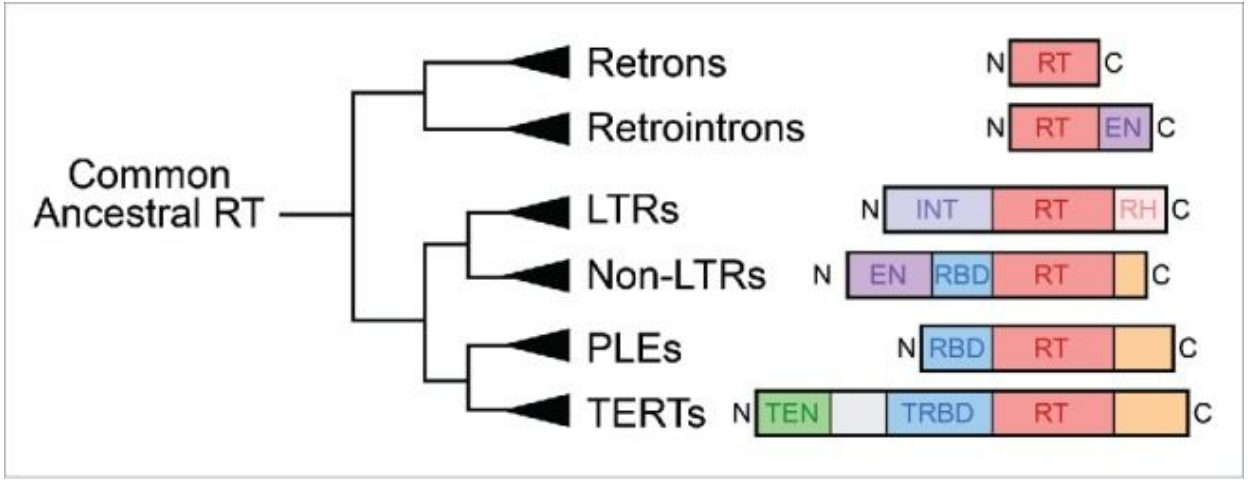


TABLE 15.3 SEQUENCES OF TELOMERE REPEATS AND TELOMERASE RNAS IN VARIOUS ORGANISMS

Species	Telomere repeat sequence	Telomerase RNA template sequence
Human	5′–TTAGGG–3′	5′–CUAACCCUAAC–3′
Oxytricha	5′–TTTTGGGG–3′	5′–CAAAACCCCAAAACC–3′
Tetrahymena	5′–TTGGGG–3′	5′–CAACCCCAA–3′

Telomeres

- Telomerase (TERT) may derive from reverse transcriptases carried by retrotransposons!



The central catalytic RT domain (red) is flanked by variable accessory domains, including endonuclease (EN, violet), integrase (INT, indigo), RNase-H (RH, pink), RNA binding domain (RBD, blue), and a thumb domain (orange). TERT contains a large N-terminal extension compromising of the DNA binding TEN (green) domain and TR binding domain (TRBD, blue).

https://pmc.ncbi.nlm.nih.gov/articles/PMC4993307/#s0003

The centromere

- Cohesin proteins bind and hold sister chromatids together
- Assembly point for the kinetochore [define]

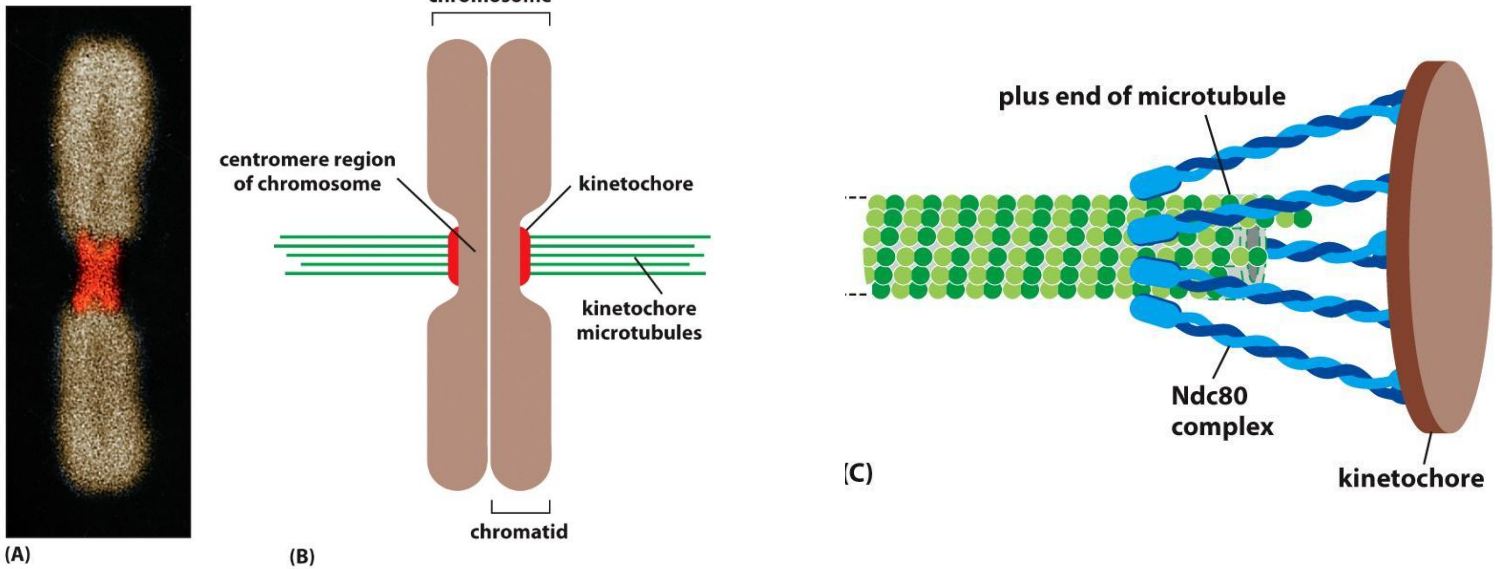
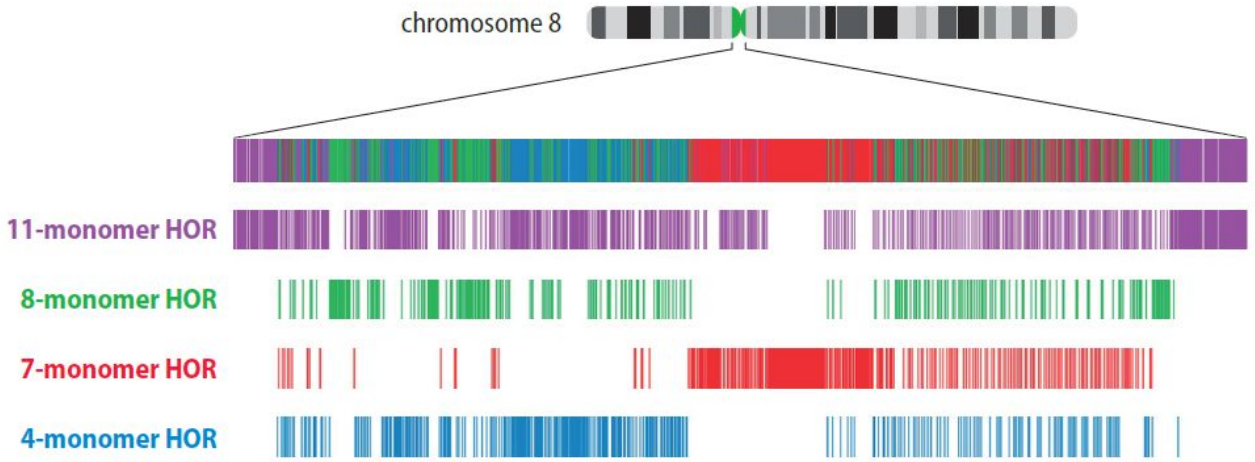


Figure 18-23 Essential Cell Biology, 4th ed. (© Garland Science 2014)

The centromere

- Centromere sequence is millions of bases long
- Repetitive alphoid DNA
  - 171 bp units
  - Arranged head-to-tail arrays in highly-order repeats (HORs)
  - Varying numbers of repeating units with location and among individuals
  - Impossible to resolve with short-read sequencing



Telomere-to-telomere sequence of a human genome

RESEARCH ARTICLE

HUMAN GENOMICS

The complete sequence of a human genome

Sergey Nurk[1]†, Sergey Koren[1]†, Arang Rhie[1]†, Mikko Rautiainen[1]†, Andrey V. Bzikadze[2], Alla Mikheenko[3], Mitchell R. Vollger[4], Nicolas Altemose[5], Lev Uralsky[6,7], Ariel Gershman[8], Sergey Aganezov[9]‡, Savannah J. Hoyt[10], Mark Diekhans[11], Glennis A. Logsdon[4], Michael Alonge[9], Stylianos E. Antonarakis[12], Matthew Borchers[13], Gerard G. Bouffard[14], Shelise Y. Brooks[14], Gina V. Caldas[15], Nae-Chyun Chen[9], Haoyu Cheng[16,17], Chen-Shan Chin[18], William Chow[19], Leonardo G. de Lima[13], Philip C. Dishuck[4], Richard Durbin[19,20], Tatiana Dvorkina[3], Ian T. Fiddes[21], Giulio Formenti[22,23], Robert S. Fulton[24], Arkarachai Fungtammasan[18], Erik Garrison[11,25], Patrick G. S. Grady[10], Tina A. Graves-Lindsay[26], Ira M. Hall[27], Nancy F. Hansen[28], Gabrielle A. Hartley[10], Marina Haukness[11], Kerstin Howe[19], Michael W. Hunkapiller[29], Chirag Jain[1,30], Miten Jain[11], Erich D. Jarvis[22,23], Peter Kerpedjiev[31], Melanie Kirsche[9], Mikhail Kolmogorov[32], Jonas Korlach[29], Milinn Kremitzki[26], Heng Li[16,17], Valerie V. Maduro[33], Tobias Marschall[34], Ann M. McCartney[1], Jennifer McDaniel[35], Danny E. Miller[4,36], James C. Mullikin[14,28], Eugene W. Myers[37], Nathan D. Olson[35], Benedict Paten[11], Paul Peluso[29], Pavel A. Pevzner[32], David Porubsky[4], Tamara Potapova[13], Evgeny I. Rogaev[6,7,38,39], Jeffrey A. Rosenfeld[40], Steven L. Salzberg[9,41], Valerie A. Schneider[42], Fritz J. Sedlazeck[43], Kishwar Shafin[11], Colin J. Shew[44], Alaina Shumate[41], Ying Sims[19], Arian F. A. Smit[45], Daniela C. Soto[44], Ivan Sović[29,46], Jessica M. Storer[45], Aaron Streets[5,47], Beth A. Sullivan[48], Françoise Thibaud-Nissen[42], James Torrance[19], Justin Wagner[35], Brian P. Walenz[1], Aaron Wenger[29], Jonathan M. D. Wood[19], Chunlin Xiao[42], Stephanie M. Yan[49], Alice C. Young[14], Samantha Zarate[9], Urvashi Surti[50], Rajiv C. McCoy[49], Megan Y. Dennis[44], Ivan A. Alexandrov[3,7,51], Jennifer L. Gerton[13,52], Rachel J. O'Neill[10], Winston Timp[8,41], Justin M. Zook[35], Michael C. Schatz[9,49], Evan E. Eichler[4,53]*, Karen H. Miga[11,54]*, Adam M. Phillippy[1]*

Since its initial release in 2000, the human reference genome has covered only the euchromatic fraction of the genome, leaving important heterochromatic regions unfinished. Addressing the remaining 8% of the genome, the Telomere-to-Telomere (T2T) Consortium presents a complete 3.055 billion–base pair sequence of a human genome, T2T-CHM13, that includes gapless assemblies for all chromosomes except Y, corrects errors in the prior references, and introduces nearly 200 million base pairs of sequence containing 1956 gene predictions, 99 of which are predicted to be protein coding. The completed regions include all centromeric satellite arrays, recent segmental duplications, and the short arms of all five acrocentric chromosomes, unlocking these complex regions of the genome to variational and functional studies.

T2T-CHM13 fills 8% of the reference which was incomplete

- ~200 Mbps new sequence
- 99 new protein coding genes
- Resolution of telomere and centromere sequences of each chromosome

Nurk, S., S. Koren, A. Rhie, M. Rautiainen, A. V. Bzikadze, A. Mikheenko, M. R. Vollger, N. Altemose, L. Uralsky, A. Gershman, S. Aganezov, et al. (2022). The complete sequence of a human genome. Science (New York, N.Y.) 376:44–53.

Telomere-to-telomere sequence of a human genome

One weird trick to sequencing the entire genome:

Complete hydatidiform mole (CHM)

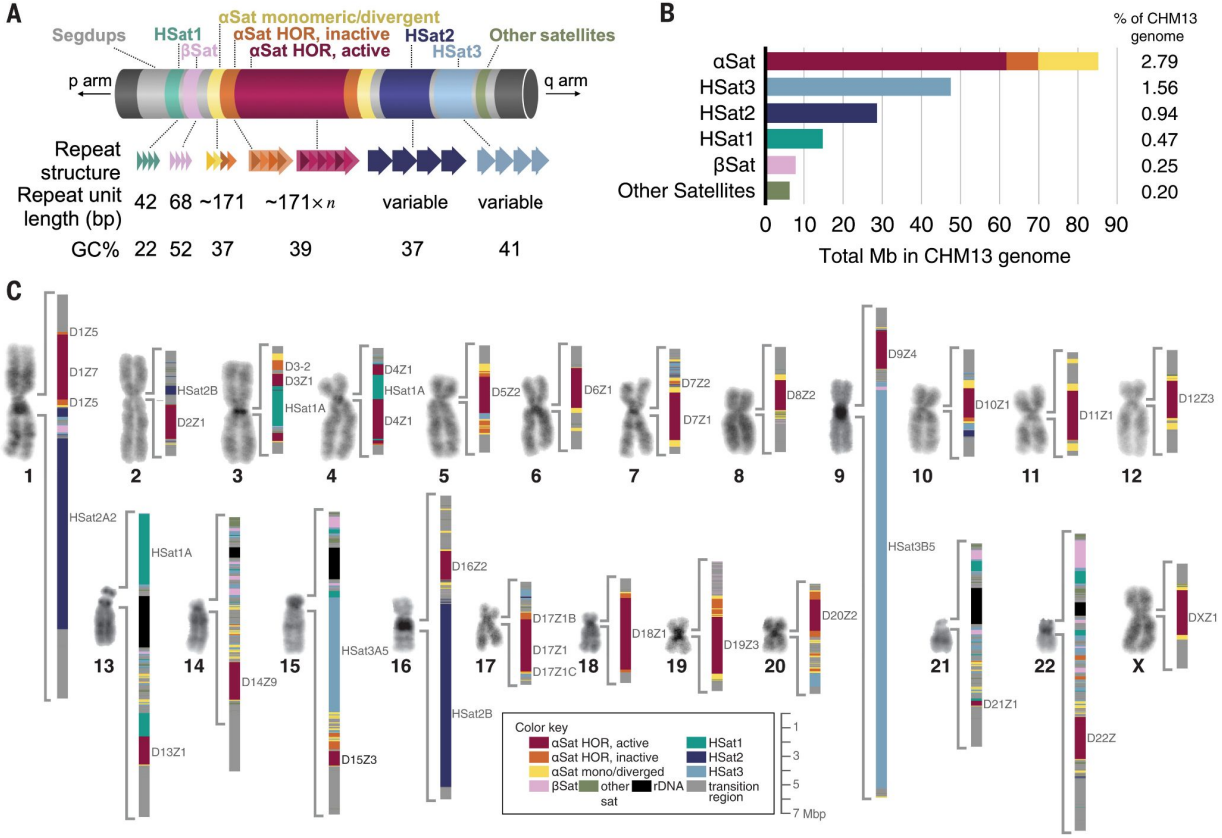
- non-viable fertilized egg, missing the maternal nucleus
- duplication of the paternal complement post-fertilization produces a perfectly homozygous genome.

Then:

- 30x PacBio HiFi
- 120x Nanopore ultralong reads
- 100x illumina
- 70x illumina HiC

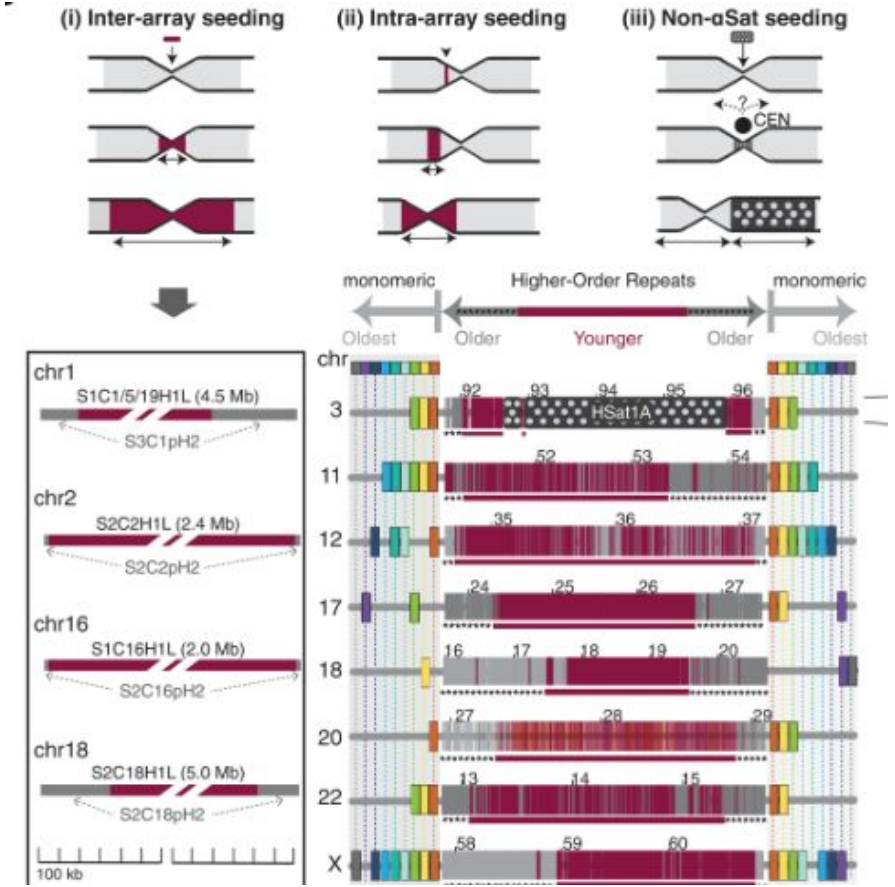
Telomere-to-telomere sequence of a human genome

- Resolves sequence of the centromere
- Reveals centromeric sequences are expanding through tandem duplication



Telomere-to-telomere sequence of a human genome

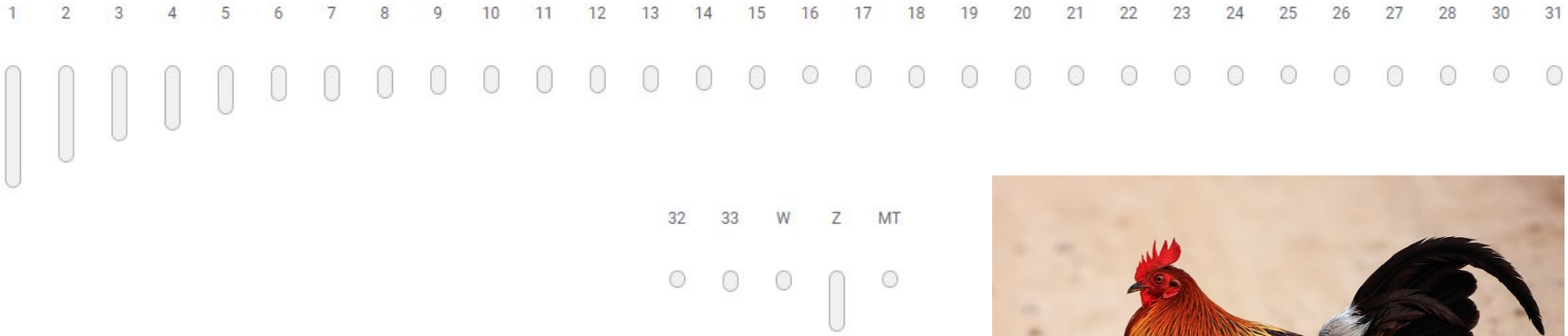
- Resolves sequence of the centromere
- Reveals centromeric sequences are expanding through tandem duplication
- Youngest sequences are the ones interacting with kinetochore proteins



Altemose, N., G. A. Logsdon, A. V. Bzikadze, P. Sidhwani, S. A. Langley, G. V. Caldas, S. J. Hoyt, L. Uralsky, F. D. Ryabov, C. J. Shew, M. E. G. Sauria, et al. (2022). Complete genomic and epigenetic maps of human centromeres. Science (New York, N.Y.) 376:eabl4178.

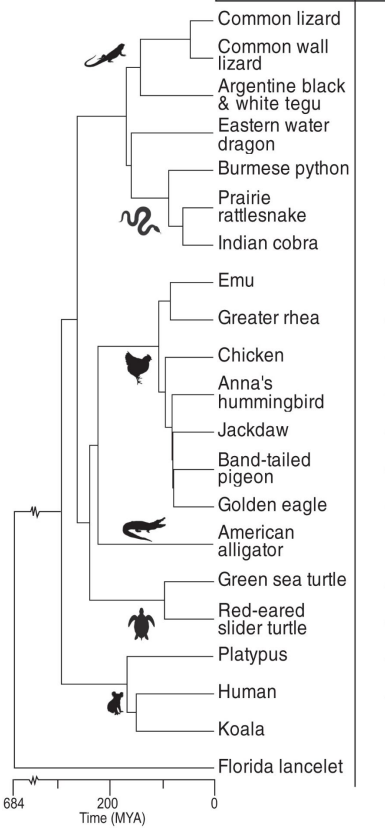
Chromosomal Diversity

- Microchromosomes - < 20 Mb in length - birds, fish, reptiles and amphibians
- Gene-dense, low repetitive sequence content, and have high rates of recombination.



Chromosomal Diversity

- Microchromosomes may be the ancestral state and macrochromosomes a derived state in mammals
- How and why these changes have occurred remains unresolved

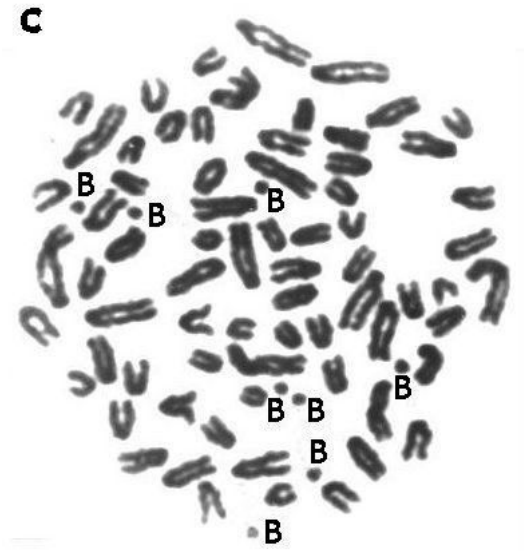


Common name	Chr number (n)	Macro assembled (n) % of genome	Micro assembled (n) % of genome	Genome size (Gb)
Common lizard	19	11 (68.7%)	8 (22.5%)	1.34
Common wall lizard	19	15 (87.9%)	4 (9.7%)	1.48
Argentine black & white tegu	19	5 (77.5%)	14 (15.8%)	1.93
Eastern water dragon	?	6 (82.8%)	19 (15.8%)	1.79
Burmese python	18	8 (84.4%)	10 (12.7%)	1.4
Prairie rattlesnake	19	8 (86.4%)	11 (10.4%)	1.3
Indian cobra	19	8 (80.6%)	11 (13.7%)	1.67
Emu	40	9 (72.8%)	31 (26.2%)	1.18
Greater rhea	40	9 (73.2%)	21 (25.8%)	1.15
Chicken	39	9 (75.3%)	24 (23.3%)	1.05
Anna's hummingbird	37	9 (71.6%)	26 (27.5%)	1.05
Jackdaw	40	10 (75.4%)	21 (24%)	1.03
Band-tailed pigeon	40	9 (74.2%)	21 (23%)	1.06
Golden eagle	31	13 (71.4%)	14 (28%)	1.19
American alligator	16	5 (68.2%)	10 (30.6%)	2.14
Green sea turtle	28	11 (77.7%)	17 (17.8%)	2.11
Red-eared slider turtle	25	11 (82.2%)	14 (17%)	2.11
Platypus	26	26 (97.6%)	0	1.81
Human	23	23 (97.8%)	0	3.21
Koala	8	8 (98.7%)	0	3.15
Florida lancelet	19	0	19 (94.2%)	0.51

https://www.pnas.org/doi/10.1073/pnas.2112494118

Chromosomal Diversity

- B chromosomes
  - Non-essential, atypical chromosomes found in some individuals
  - Usually fragments of other chromosomes and may contain genes including rRNAs

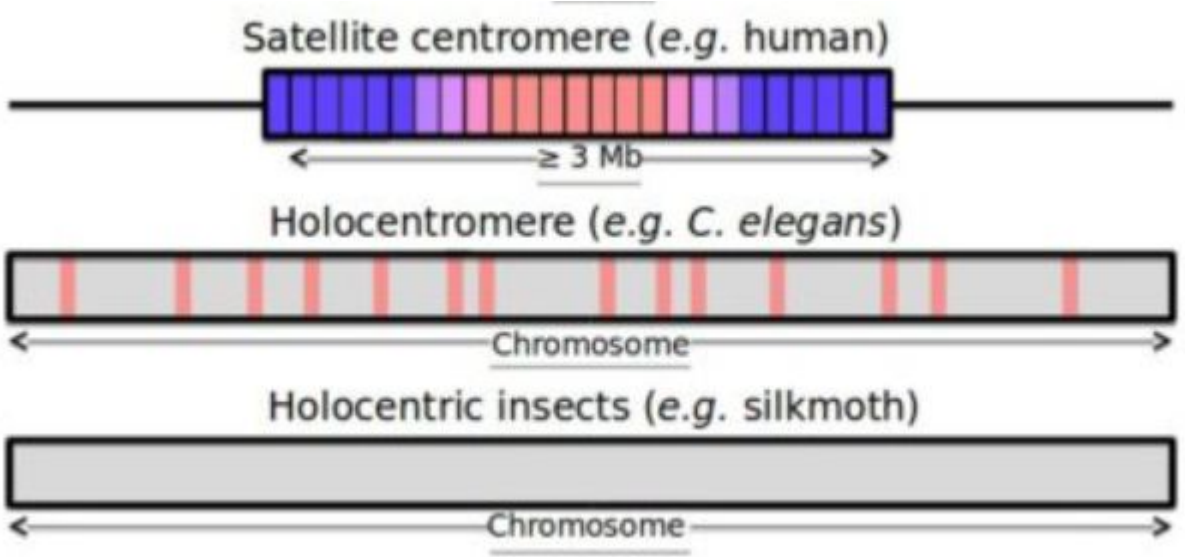


Roe deer karyotype

https://molecularcytogenetics.biomedcentral.com/articles/10.1186/1755-8166-4-22

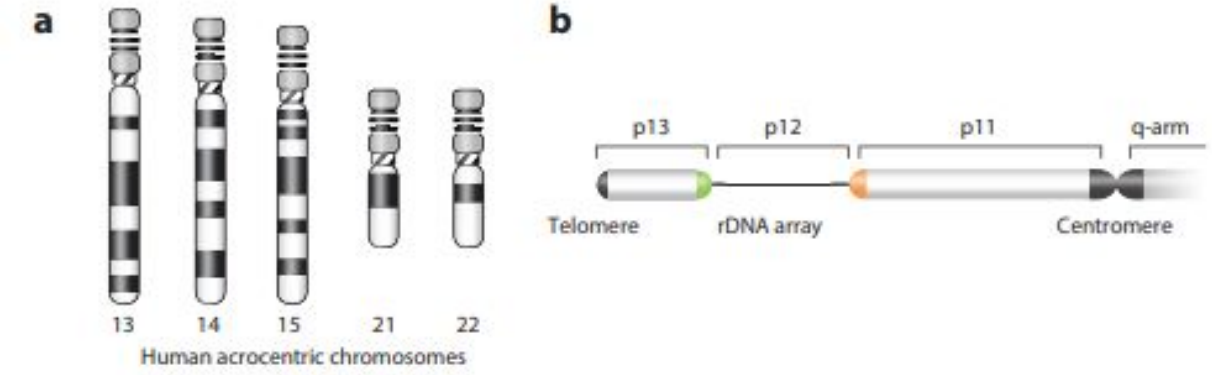
Chromosomal Diversity

- Holocentric chromosomes - *C. elegans*, some insects, some plants - multiple centromere-like structures distributed along the chromosomes or entirely absent.



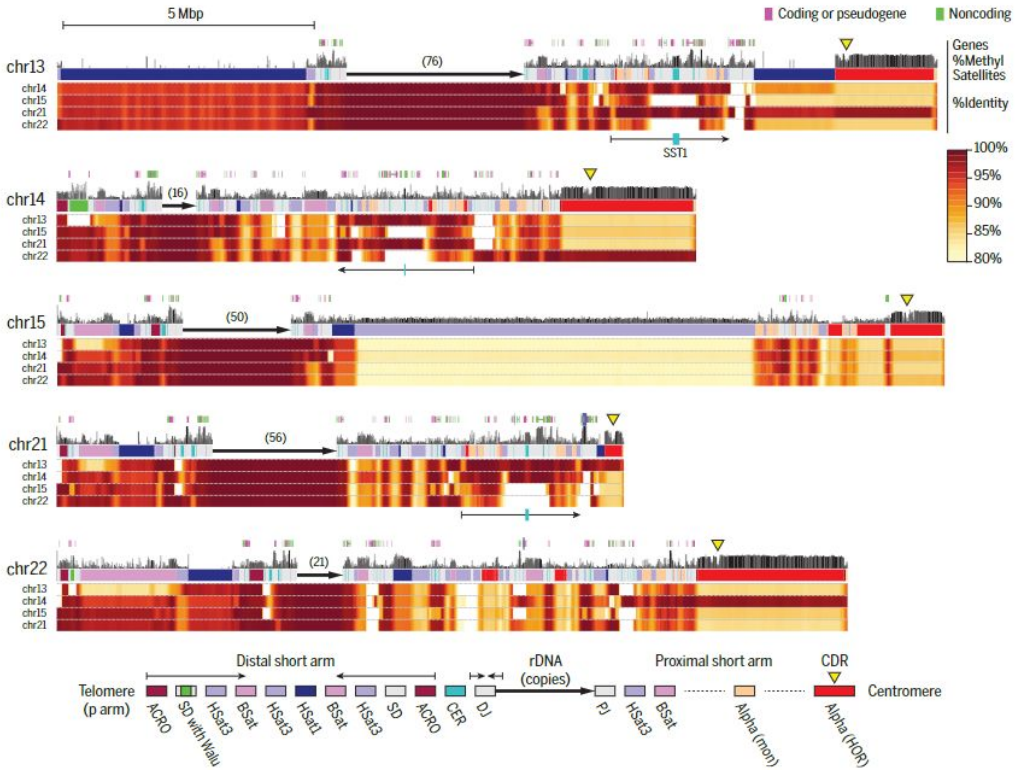
Chromosomal Diversity

- Acrocentric chromosomes - chromosomes with centromere located close to the end of the chromosome
- In humans the short arm sequences are highly repetitive and high rates of heterologous recombination among chromosomes



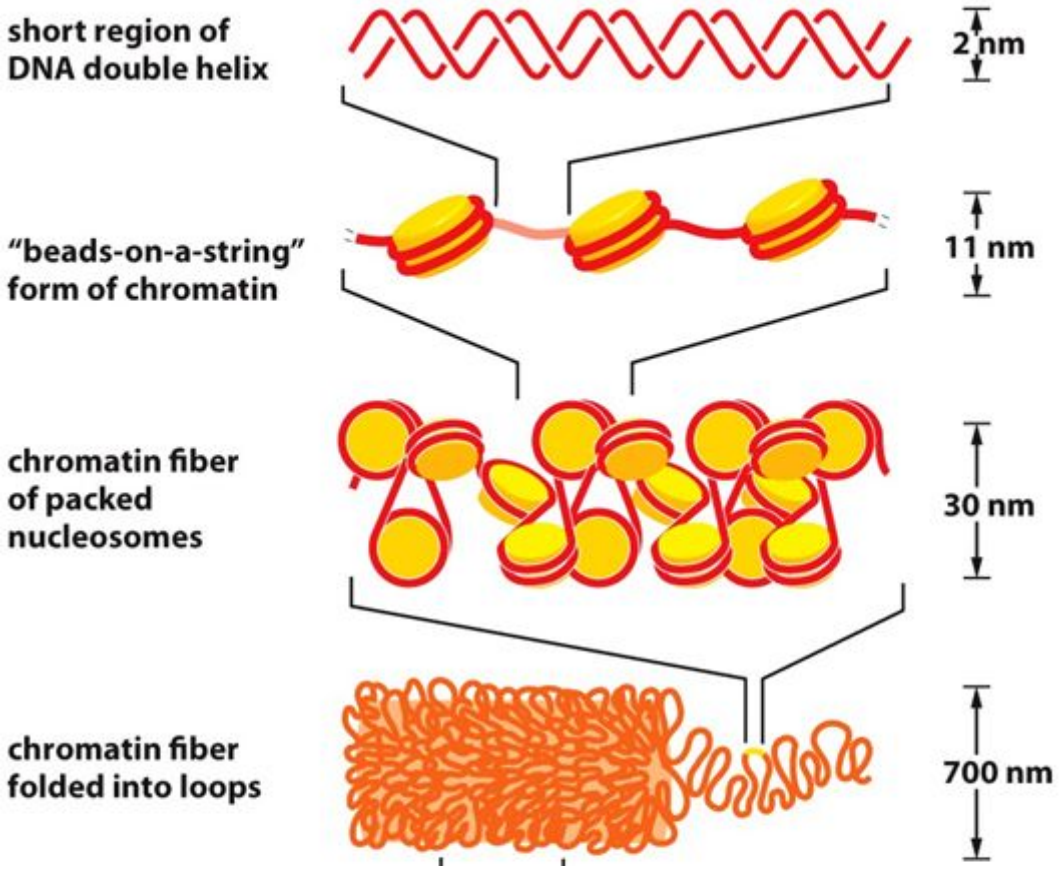
Chromosomal Diversity

- In humans the short arm sequences are highly repetitive and high rates of heterologous recombination among chromosomes
- rDNA copy numbers highly variable and among populations and individuals



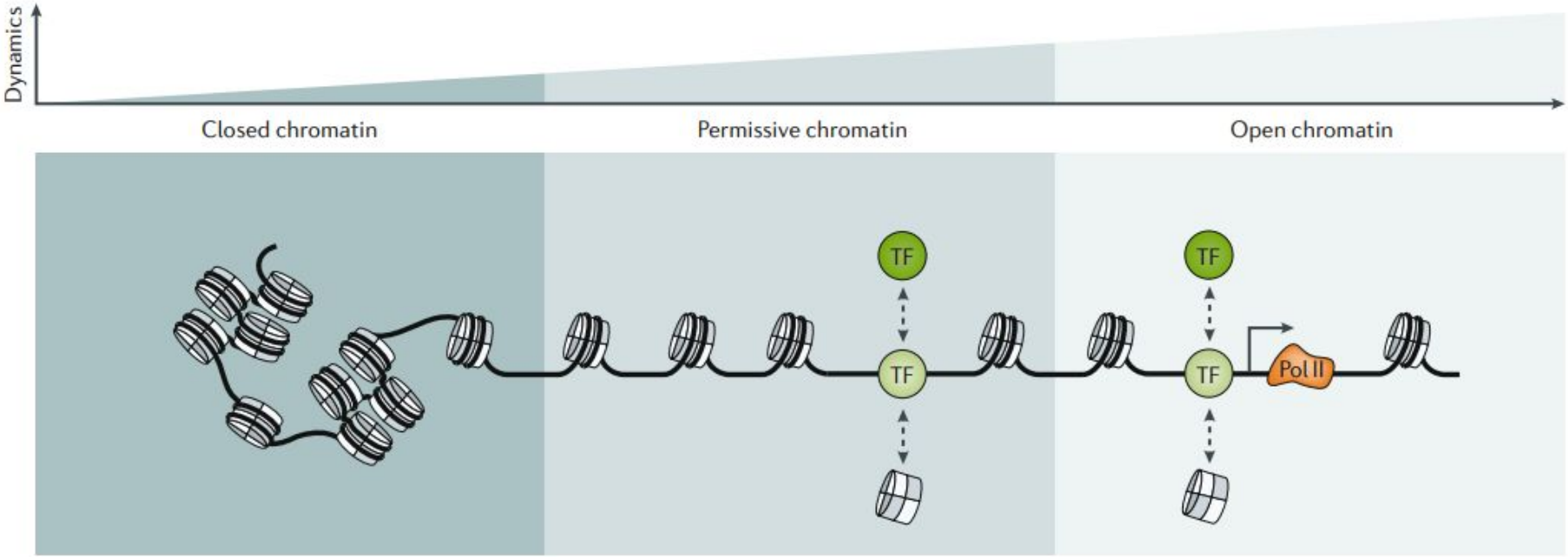
Chromatin state

- Chromatin must be “opened” for gene expression to proceed



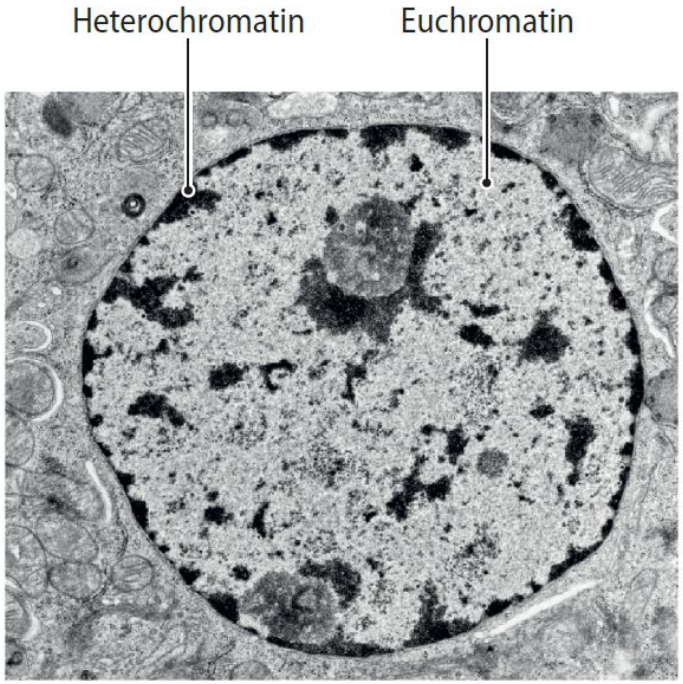
Chromatin state

- Chromatin must be “opened” for gene expression to proceed



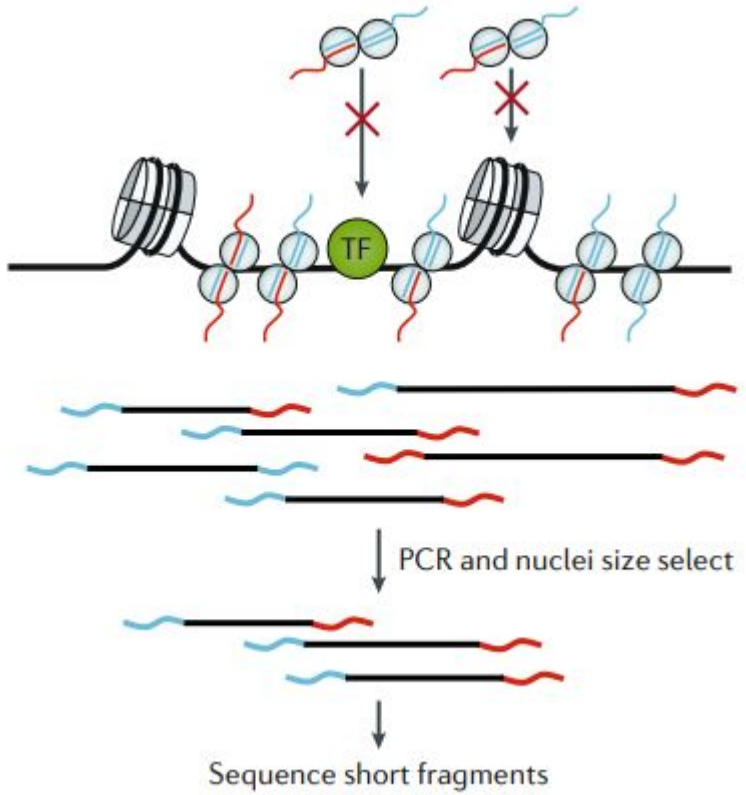
Chromatin state

- Tightly condensed **heterochromatin** in transcriptionally inactive
- Less sense **euchromatin** is the site of active transcription

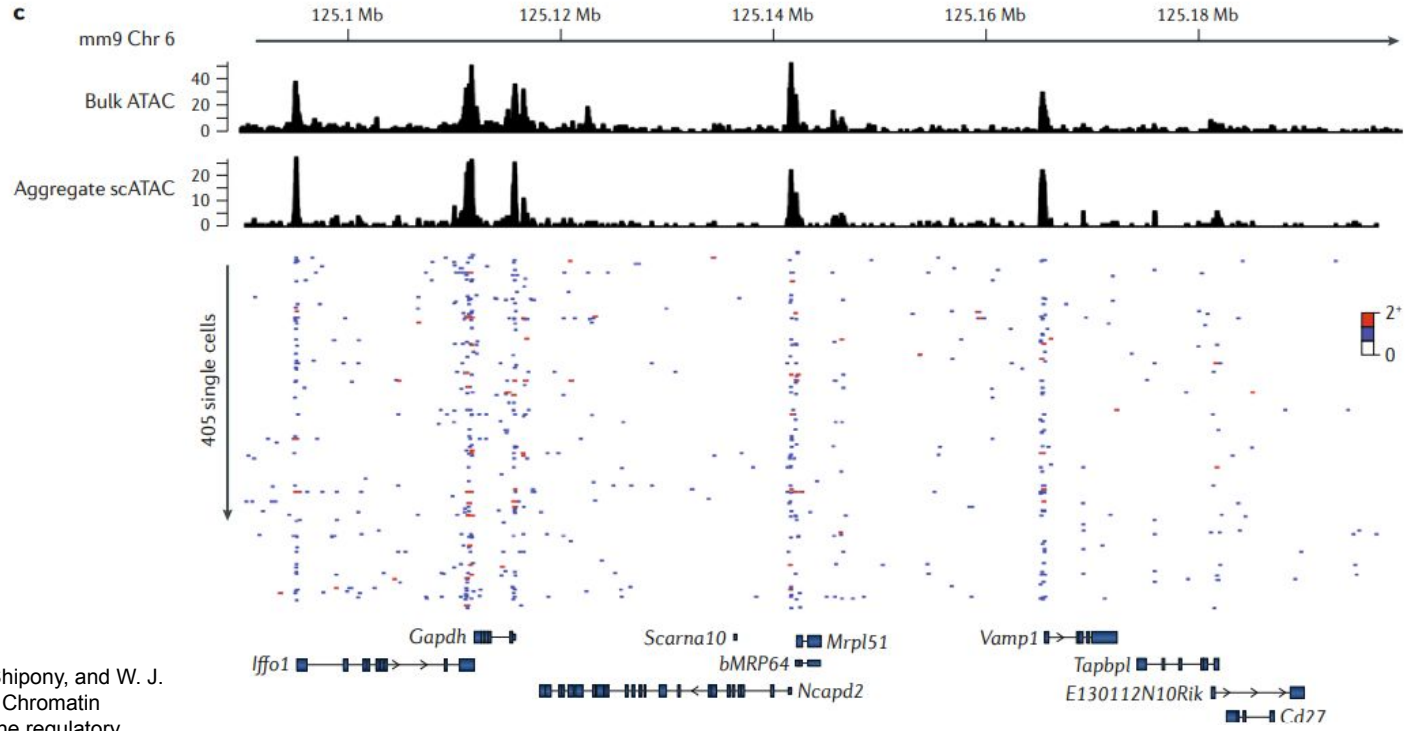


Mapping open chromatin can yield insights into transcriptional regulation

- ATAC-Seq - Assay for Transposase-Accessible Chromatin with Sequencing is a widely employed approach
- Transposase enzyme simultaneously cleaves and ligates sequencing adapters to accessible DNA

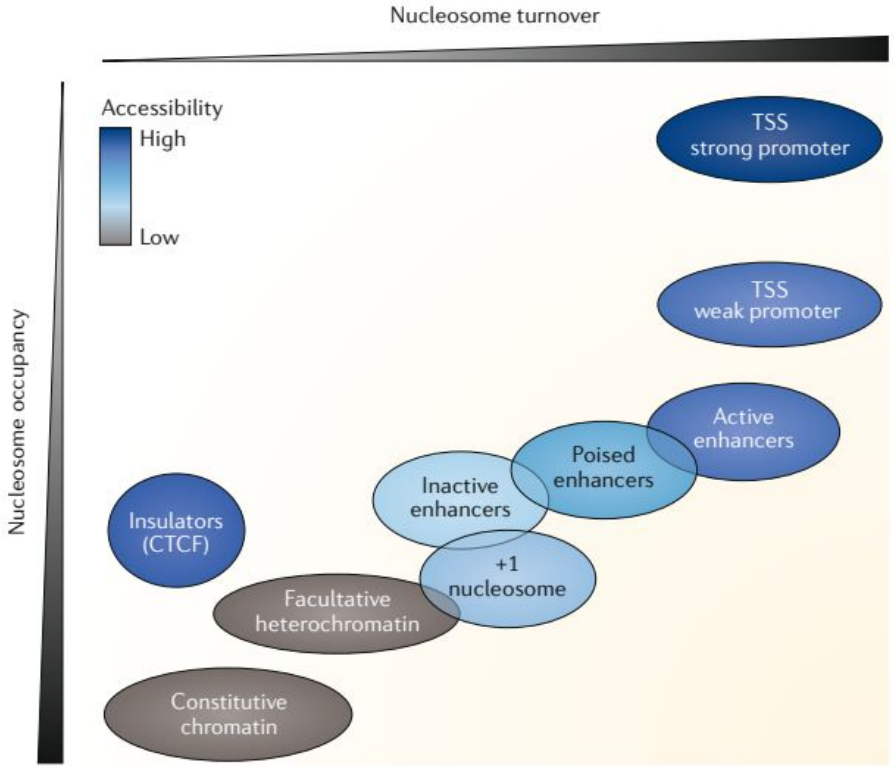


ATAC-Seq provides maps of accessible regions of the genome (in the cells or tissues sampled)



Klemm, S. L., Z. Shipony, and W. J. Greenleaf (2019). Chromatin accessibility and the regulatory epigenome. Nature reviews. Genetics 20:207–220.

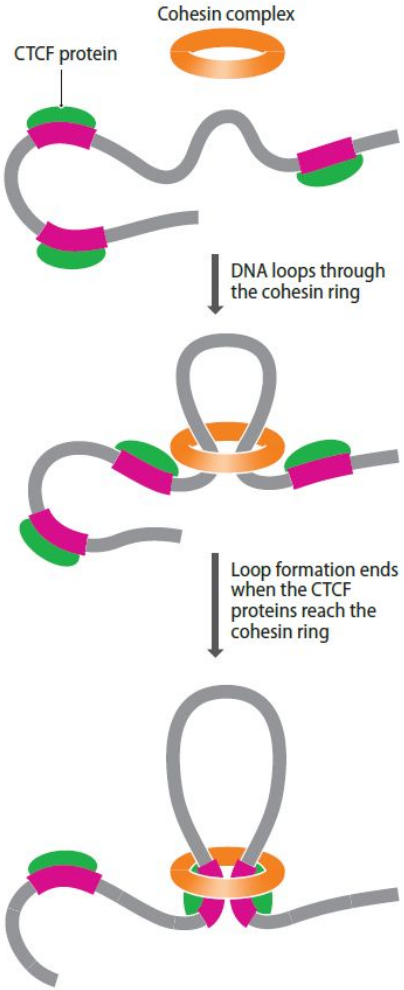
ATAC-Seq is especially good at identifying specific elements



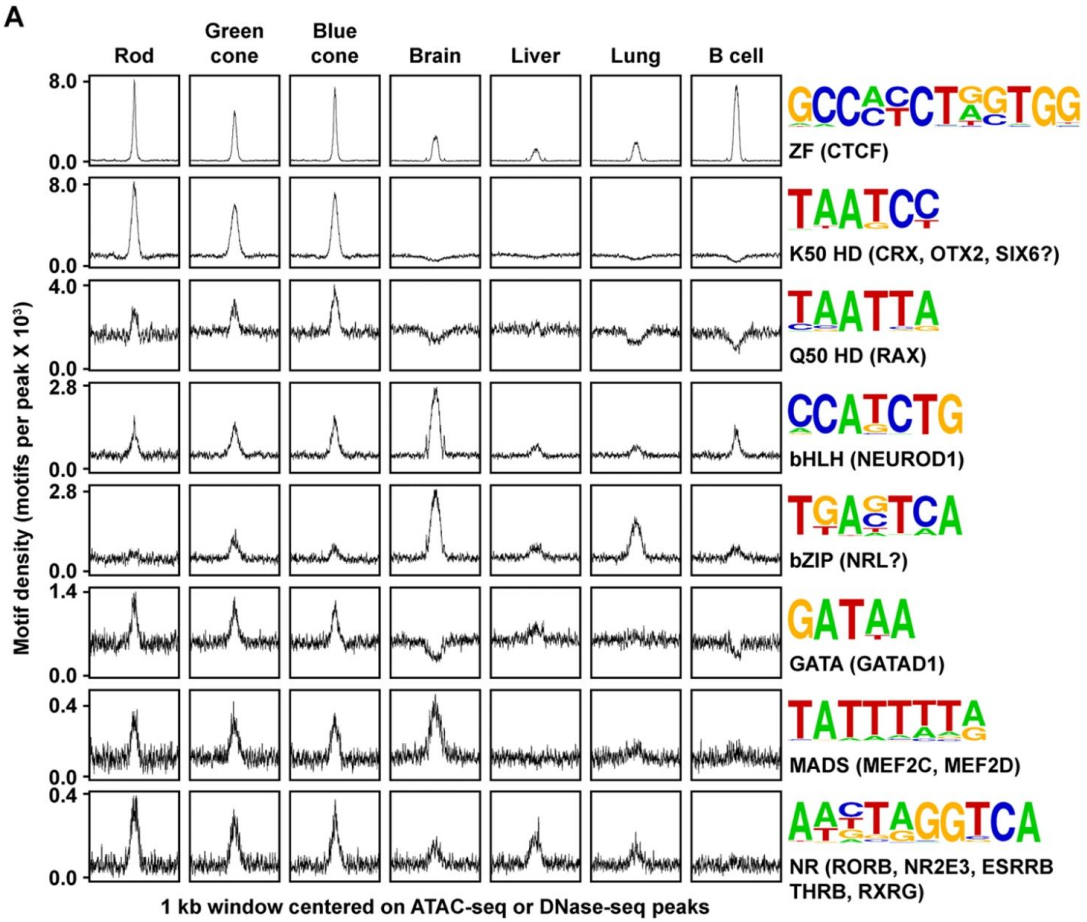
CTFC sites define topological domains in the genome

- Regulatory interactions (activation, silencing) occur within loops.

- Such interactions generally do not occur between loops

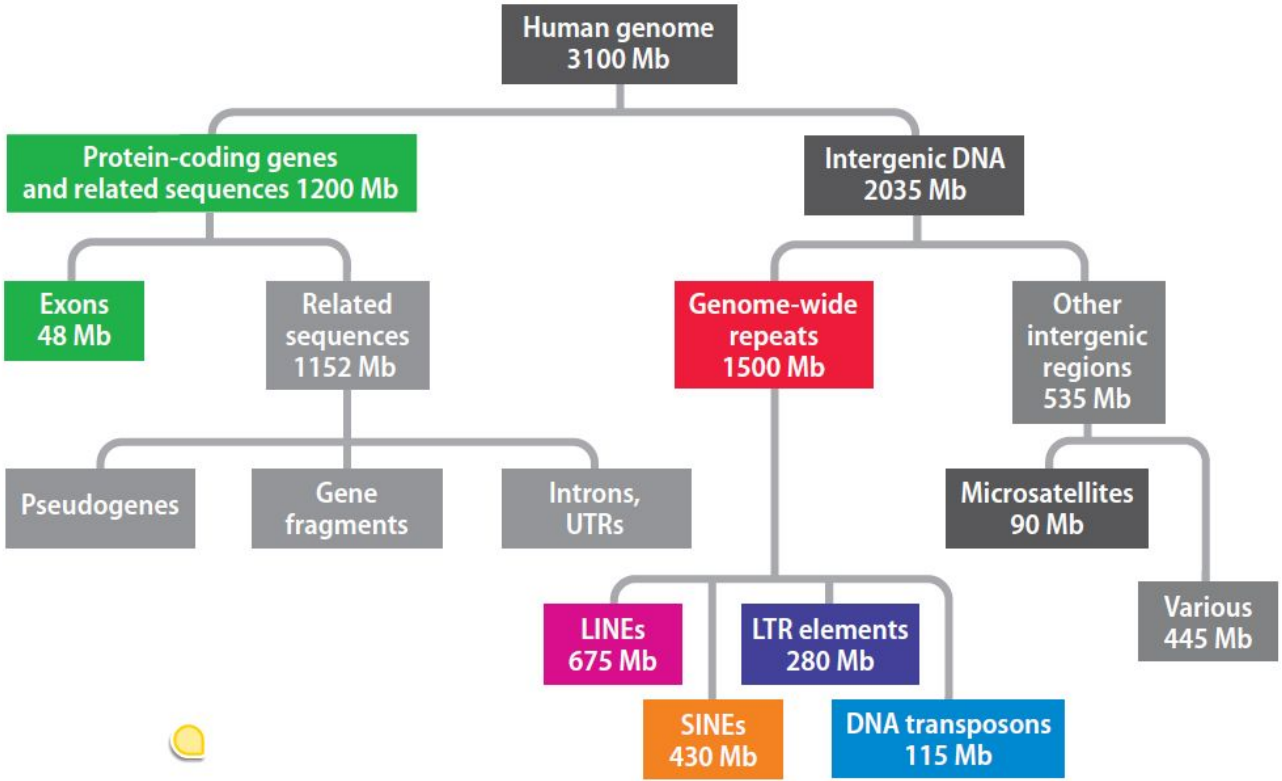


The sequences of ATAC-seq peaks can be used to infer important transcription factors



Hughes, A. E. O., J. M. Enright, C. A. Myers, S. Q. Shen, and J. C. Corbo (2017). Cell Type-Specific Epigenomic Analysis Reveals a Uniquely Closed Chromatin Architecture in Mouse Rod Photoreceptors. Scientific reports 7:43184.

Genetic features of the eukaryotic nuclear genome



The number of protein coding genes is surprisingly small!

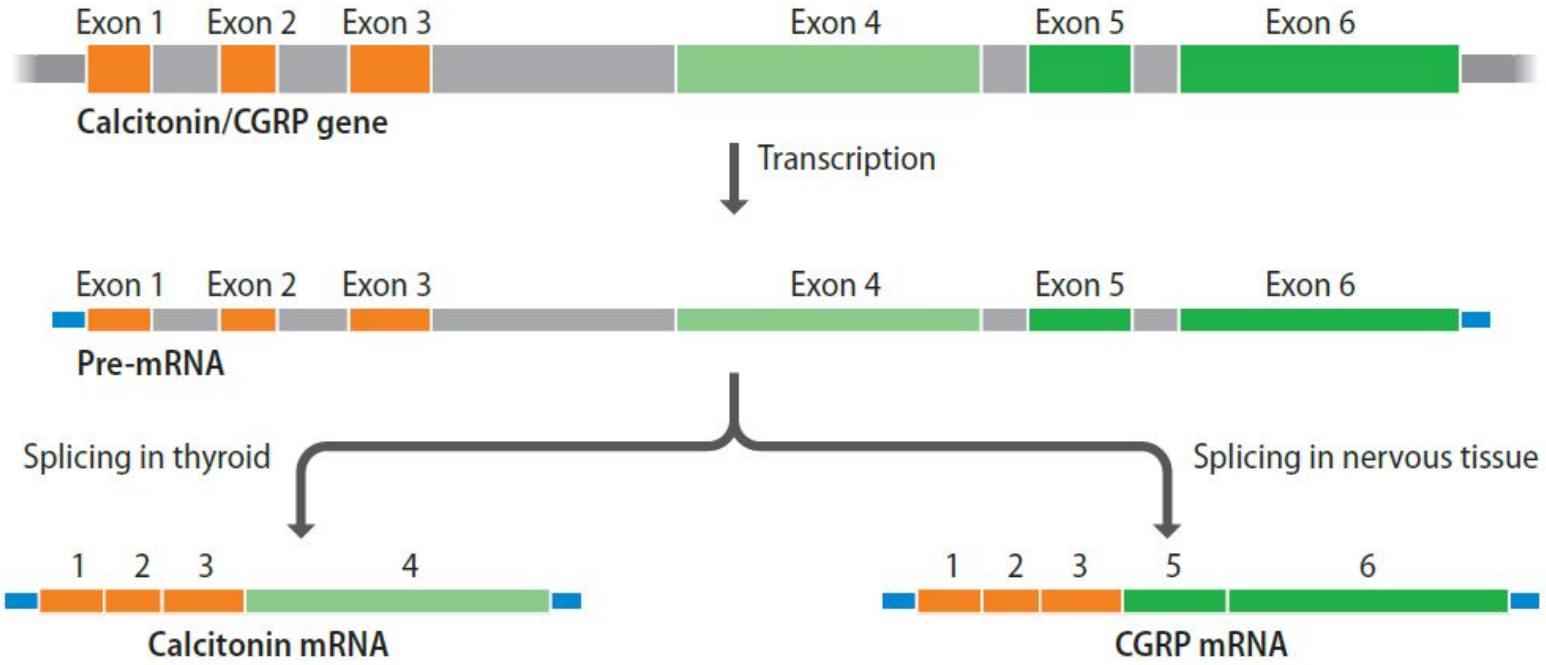
TABLE 7.2 NUMBERS OF PROTEIN-CODING GENES FOR VARIOUS EUKARYOTES

Species	Protein-coding genes
Saccharomyces cerevisiae (budding yeast)	6600
Schizosaccharomyces pombe (fission yeast)	5145
Caenorhabditis elegans (nematode worm)	20,191
Arabidopsis thaliana (plant)	27,655
Drosophila melanogaster (fruit fly)	13,968
Oryza sativa (rice)	37,960
Gallus gallus (chicken)	16,878
Homo sapiens (human)	20,442

Data taken from Ensembl release 104, Ensembl Plants release 51, and Ensembl Fungi release 51.

Alternative splicing increases diversity

- In humans 20,442 genes → 78,120 proteins
- ~75% of all human protein coding genes have splice variants



Pseudogenes

- Sequence of nucleotides that resembles a genuine gene but which does not specify a functional RNA or protein
  - Duplicated pseudogenes
    - often when gene duplication occurs one member may accumulate mutations and pseudogenize
  - Unitary pseudogene
    - single copy gene, true loss-of-function
    - e.g. *L-gulono-γ-lactone oxidase (gulo)* in Haplorhini primates
  - Processed pseudogenes
    - reverse transcription and reintegration of mRNA sequence.
    - Most common in human genome

Genes are unequally distributed in the genome

- In humans, ranges from:
  - 41.2 genes per Mb in Chr19
  - 6.3 in Chr13
  - just 3 per Mb in the Y chromosome
- Genes are also unevenly distributed within chromosomes:

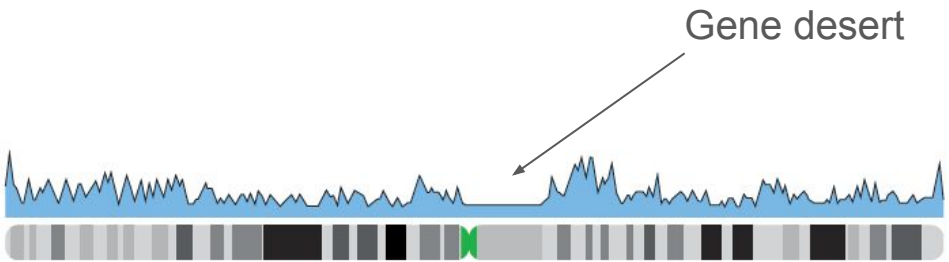
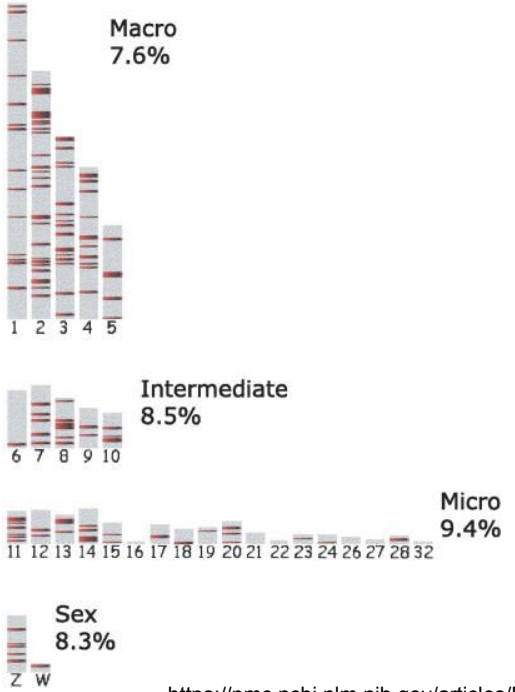


Figure 7.11 Gene density along human chromosome 1.

Gene deserts in chicken



https://pmc.ncbi.nlm.nih.gov/articles/PMC540279/

Protein coding genes are a very small portion of genome

- In humans, protein coding genes make up 48 Mb or ~1.5% of total sequence!
- Much of the genome (48.5%) is composed of **interspersed repeat sequences**

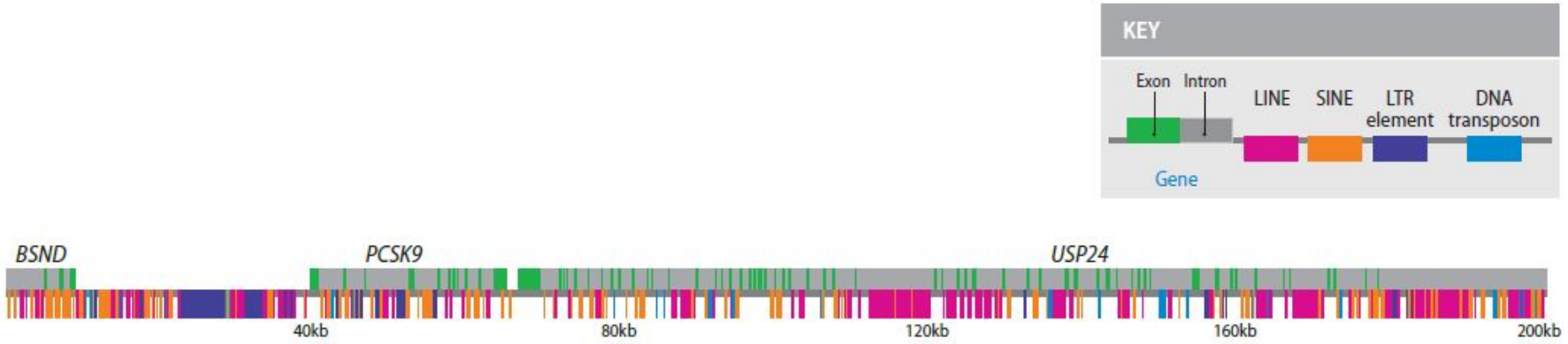


Figure 7.12 A 200 kb segment of the human genome. The segment runs from nucleotide position 55,000,000 to position 55,200,000 of chromosome 1. Within the genes, exons are shown as green boxes and introns as gray boxes. Data taken from the UCSC Genome Browser hg38 assembly.

Interspersed repeat sequences are primarily composed of transposable elements

- **Transposon** - A genetic element that can move from one position to another in a DNA molecule.

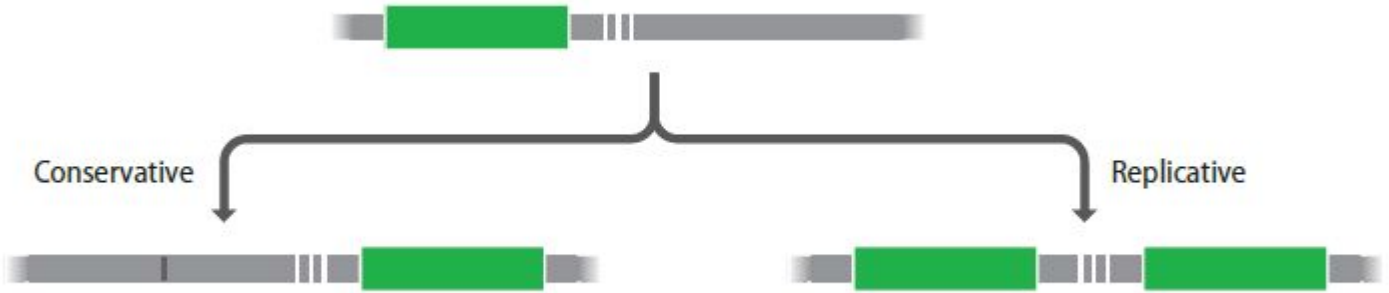


Figure 9.13 Conservative and replicative transposition.

Interspersed repeat sequences are primarily composed of transposable elements

- **Retrotransposition** - transposition via an RNA intermediate

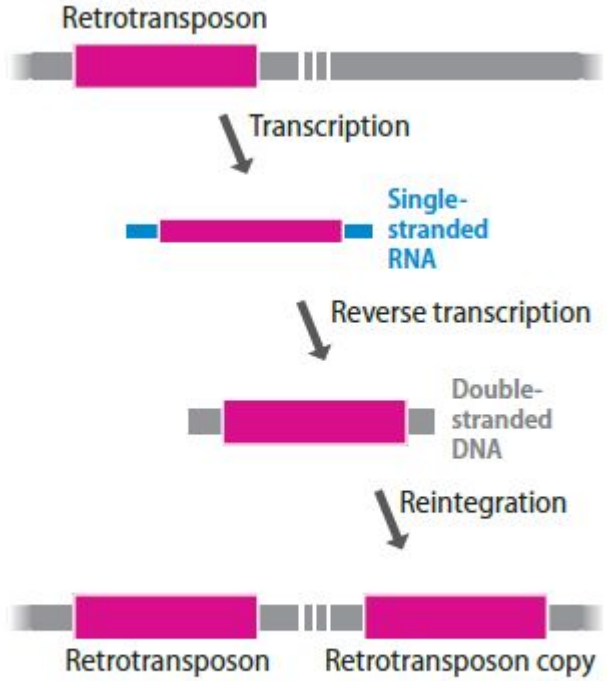


Figure 9.14 Retrotransposition.

Long terminal repeat (LTR) elements

- ~8% of human genome up to 75% of maize genome
- Derive from retroviruses and encode
  - *gag* - capsid proteins
  - *pol* - reverse transcriptase
- May produce virus-like particles
- Many are “decayed” with loss of function mutations in genes

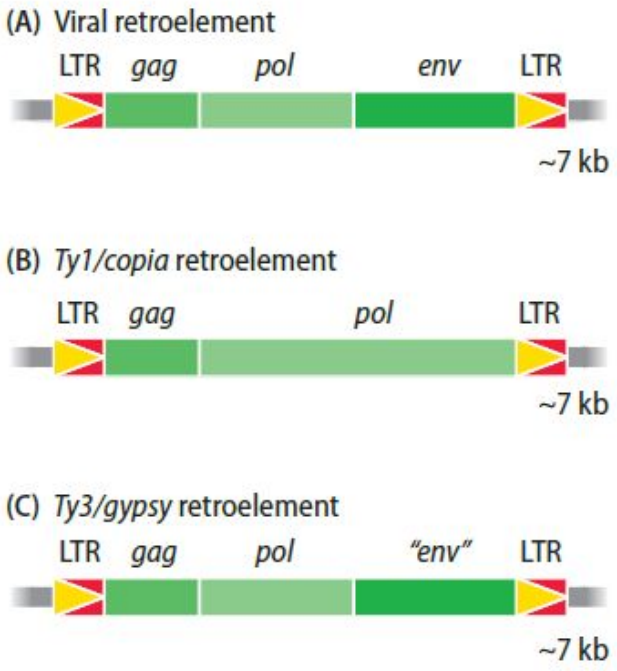
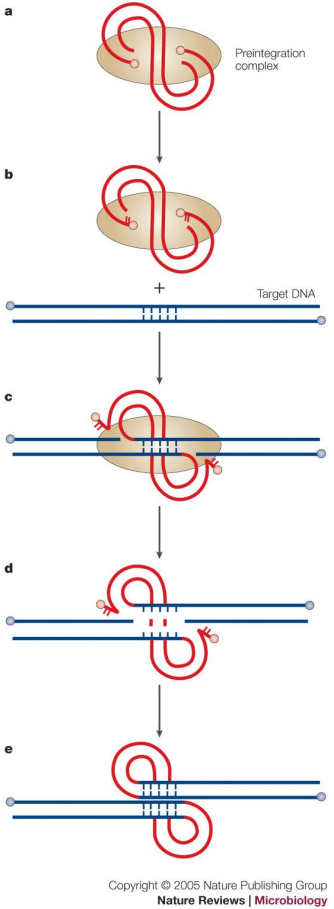


Figure 9.15 Genome structures for LTR

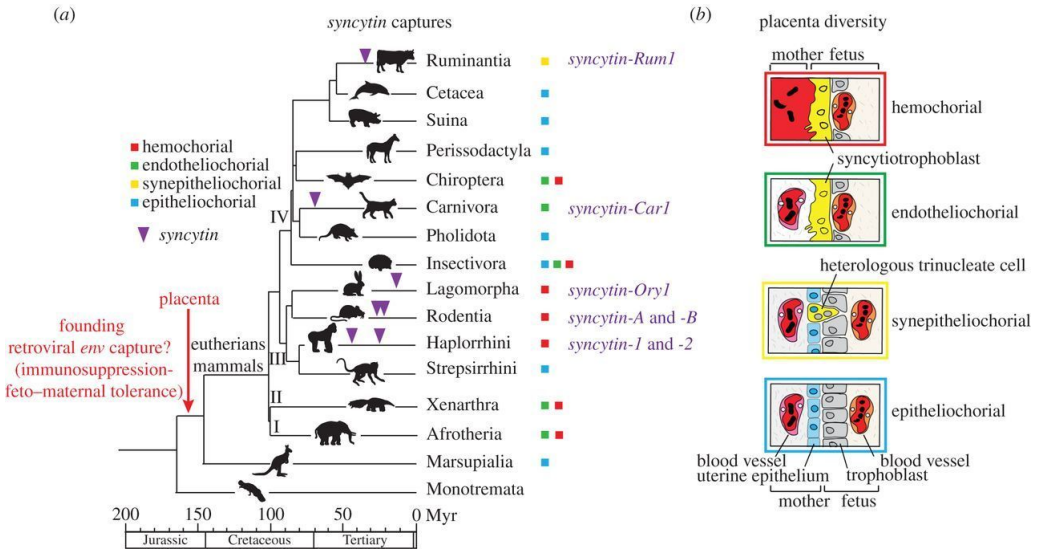
Long terminal repeat (LTR) elements

- ~8% of human genome up to 75% of maize genome
- Derive from retroviruses and encode
  - *gag* - capsid proteins
  - *pol* - reverse transcriptase
- May produce virus-like particles
- Many are “decayed” with loss of function mutations in genes
- LTR sequences mediate integration



Syncytins are derived from retroviruses

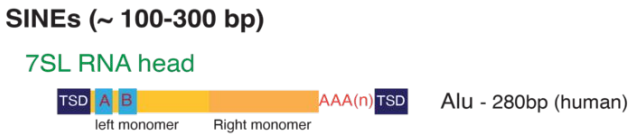
- Syncytins are proteins essential for placental development and implantation in mammals
- Evolved from endogenous retroviral envelope (*env*) gene

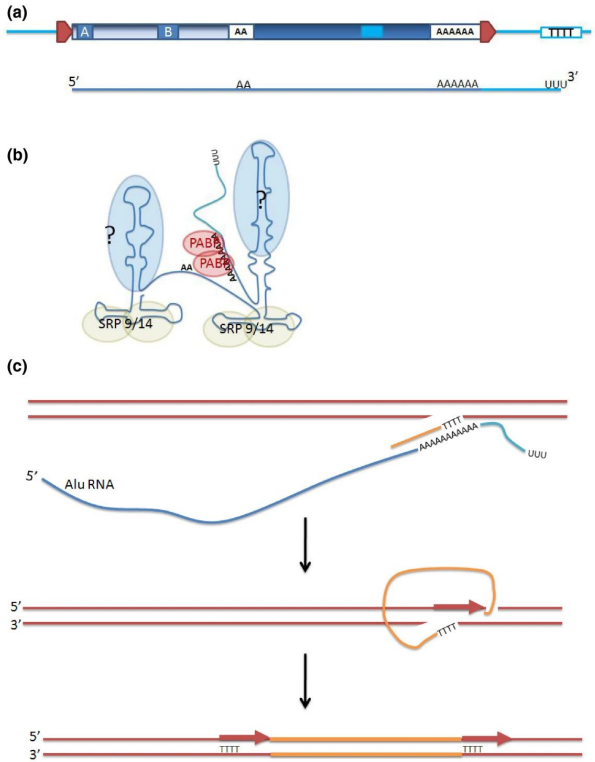


https://royalsocietypublishing.org/doi/full/10.1098/rstb.2012.0507

Short interspersed nuclear elements (SINEs)

- 1.7 million copies in human genome (~14%)
- Nearly 10% of human genome *Alu* elements
  - ~280 bp
  - Derives from 7SL RNA gene
  - Requires other elements (LINEs) for expression and replication.
  - Provide homology across genome and may drive high levels of recombination
  - Varied effect on gene expression and function

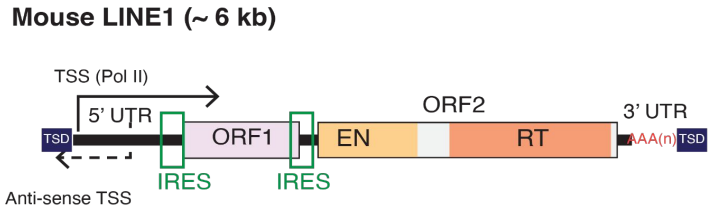




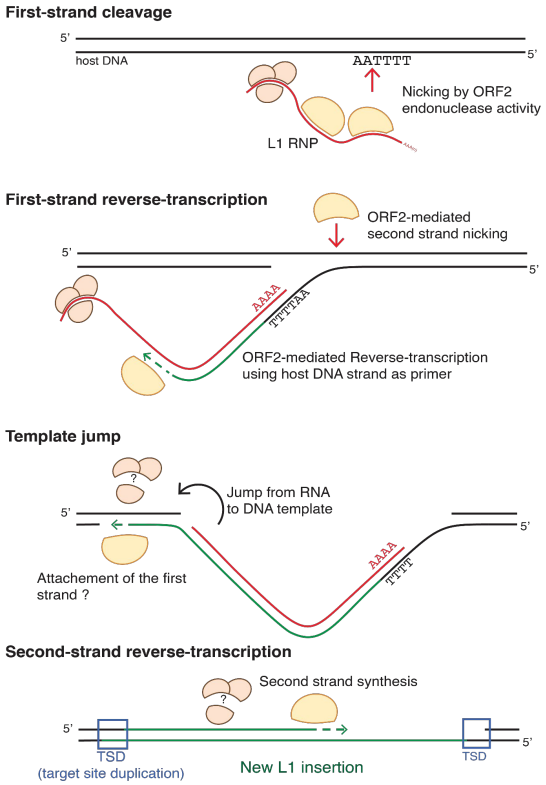
https://genomebiology.biomedcentral.com/articles/10.1186/gb-2011-12-12-236

Long interspersed nuclear elements (LINEs)

- ~6 kb in length, but many truncated
- >20% of human genome
- *gag* and *pol* elements, but not LTRs

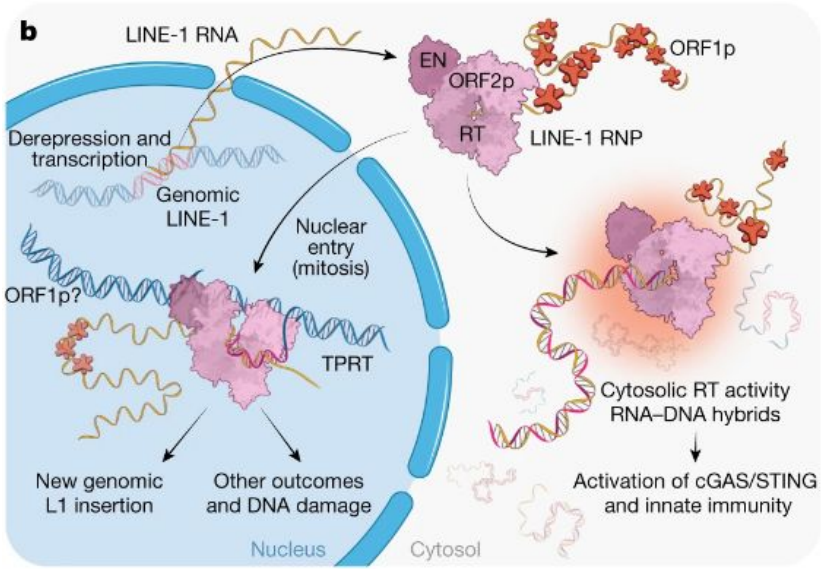


EN: Endonuclease domain. **RT:** Reverse transcriptase domain.
IRES: Internal ribosomal entry site **TSD**: Tandem site duplication.
A and **B**: RNA pol III promoter box



Long interspersed nuclear elements (LINEs)

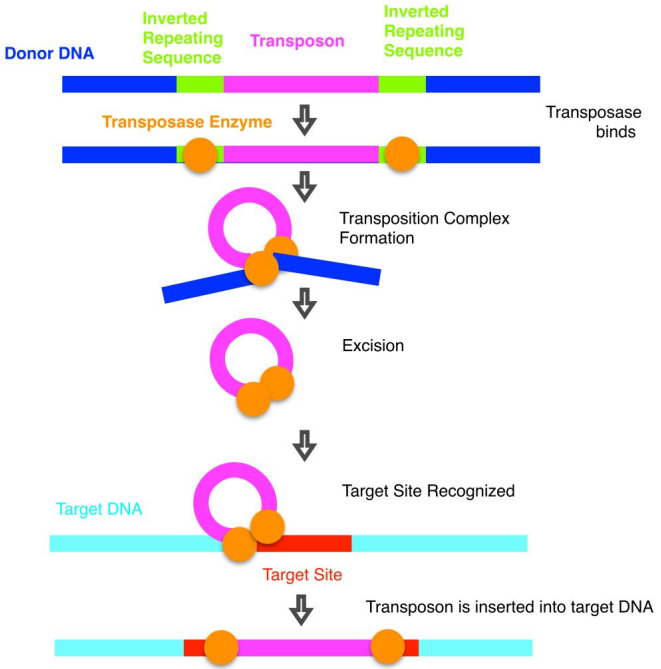
- ~6 kb in length, but many truncated
- >20% of human genome
- *gag* and *pol* elements, but not LTRs
- Activation of LINE-1 expression implicated in disease processes



https://www.nature.com/articles/s41586-023-06947-z

DNA transposons

- 3.7% of human genome, most in active
- More abundant in plants
- First described by Barbara McClintock in 1944



https://en.wikipedia.org/wiki/DNA_transposon#/media/File:Cut_and_Paste_mechanism_of_transposition.svg

Microsatellites

- ~3% of human genome
- Very short (<13 bp) tandemly repeated sequences
- Produced through polymerase slippage during DNA replication
- Highly variable - lengths are different among all individuals
- Major mechanism of “DNA fingerprinting”