Whiteboard

CS273A

The Human Genome Source Code

Gill Lecture 5: Gene Regulation III, Repeats I

Mon, Wed 11:30 AM - 12:50, on Zoom*
Prof: Gill Bejerano
CA: Boyoung (Bo) Yoo
* Track class on Piazza

Announcements

- HW2 in/HW3 out today
- Anonymous feedback – mid term
  - We'll figure out a mechanism and solicit via piazza
  - Please share :)
- Welcome to e-mail me to discuss your science

Class Topics

good morning!

(0) Genome context:
cells, DNA, central dogma

(1) Genome content / genome function:
genes, gene regulation, epigenetics, repeats, SARS-CoV-2

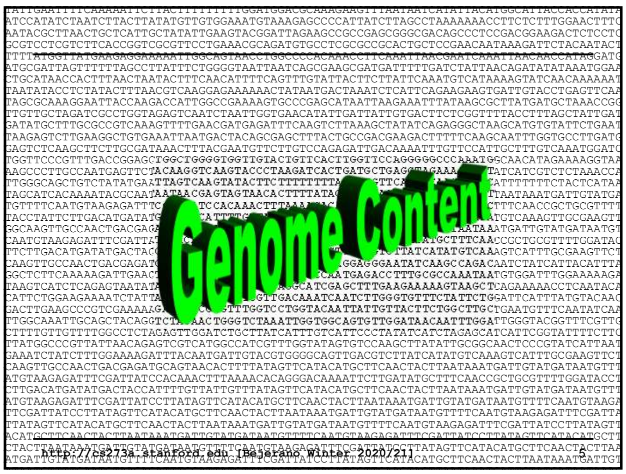
(2) Genome sequencing:
technologies, assembly/analysis, technology dependence

(3) Genome evolution:
evolution = mutation + selection, main forces of evolution:
Neutral evolution, Negative selection, Positive selection

(4) Population genomics:
Human migration, paternity testing, forensics, cryptogenomics

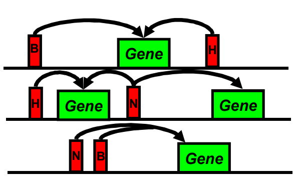
(5) Genomics of human disease:
personal genomics, GxE disease types, deep dive monogenics

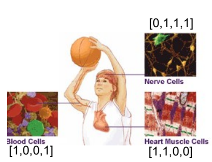
(6) Comparative Genomics :
Genomics of amazing animal adaptations, ultraconservation



Genes & Gene Regulation

- Gene = genomic substring that encodes HOW to make a protein (or ncRNA).
- Genomic switch = genomic substring that encodes WHEN, WHERE & HOW MUCH of a protein to make.





Transcription Regulation

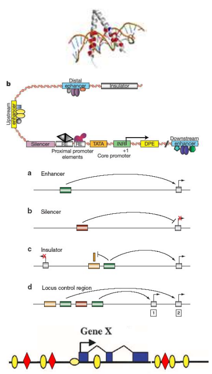
Conceptually simple:

1. The machine that transcribes (“RNA polymerase”)
2. All kinds of proteins and ncRNAs that bind to DNA and to each other to attract or repel the RNA polymerase (“transcription associated factors”).
3. DNA accessibility – making DNA stretches in/accessible to the RNA polymerase and/or transcription associated factors by un/wrapping them around nucleosomes.

(Distinguish DNA patterns from proteins they interact with)

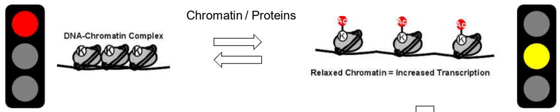
Cis-Regulatory Components

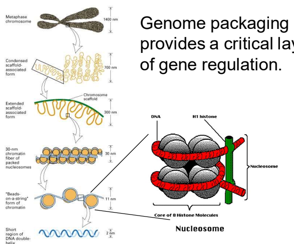
- **Low level** (“atoms”):
- Promoter motifs (TATA box, etc)
- Transcription factor binding sites (TFBS)
- **Mid Level**:
- Promoter
- Enhancers
- Repressors/silencers
- Insulators/boundary elements
- Locus control regions
- **High Level**:
- Epigenomic domains / signatures
- Gene regulatory domains
- Gene regulatory networks



Transcription & its regulation happen in open chromatin

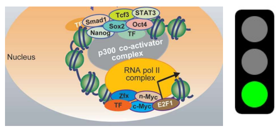
Nucleosomes, Histones, Transcription





Genome packaging provides a critical layer of gene regulation.

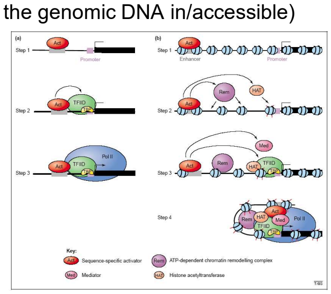
DNA / Proteins



Gene Activation / Repression via Chromatin Remodeling

A dedicated machinery opens and closes chromatin.

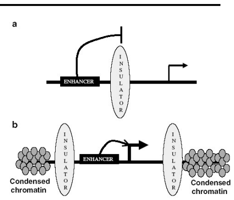
Interactions with this machinery turn genes and/or gene regulatory regions like enhancers and repressors on or off (by making the genomic DNA in/accessible)

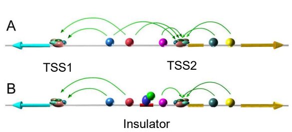


Insulators revisited

Insulators are DNA sequences that when placed between target gene and enhancer prevent enhancer from acting on the gene.

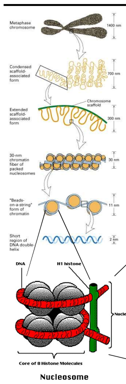
- Known insulators contain binding sites for a specific DNA binding protein (CTCF) that is involved in DNA 3D conformation.
- However, CTCF fulfills additional roles besides insulation. I.e, the presence of a CTCF site does not ensure that a genomic region acts as an insulator.



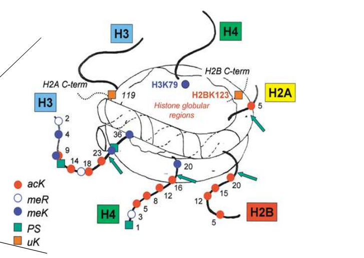


Epigenomics
The histone code

Histone Tails, Histone Marks

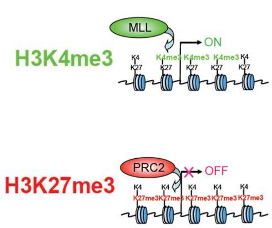


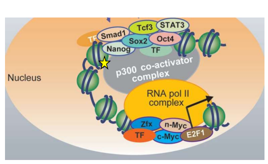
DNA is wrapped around nucleosomes.
Nucleosomes are made of histones.
Histones have free tails.
Residues in the tails are modified in specific patterns in conjunction with specific gene regulation activity.



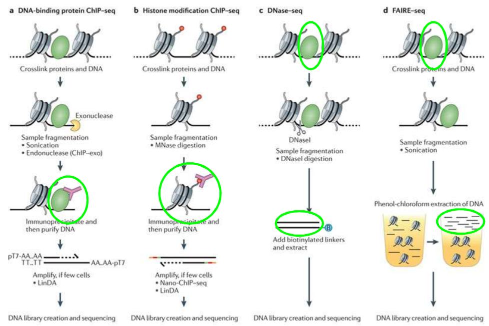
Histone Mark Correlation Examples

Active gene promoters are marked by H3K4me3
Silenced gene promoters are marked by H3K27me3
p300, a protein component of many active enhancers acetylates H3k27Ac.



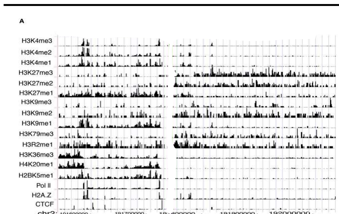


Measuring these different states

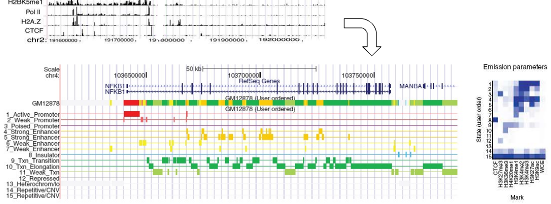


Note that the DNA itself doesn't change. We sequence different portions of it that are currently in different states (bound by a TF, wrapped around a nucleosome etc.)

Epigenomics: study all these marks genomewide

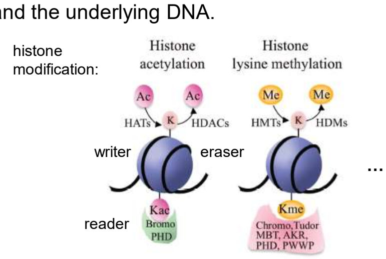


Translate observations into current genome state.



Histone Code Hypothesis

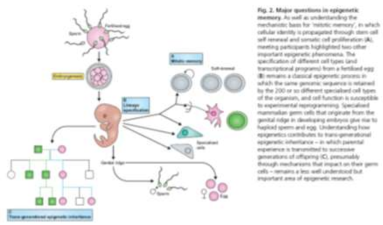
Histone modifications serve to recruit other proteins by specific recognition of the modified histone via protein domains specialized for such purposes, rather than through simply stabilizing or destabilizing the interaction between histone and the underlying DNA.



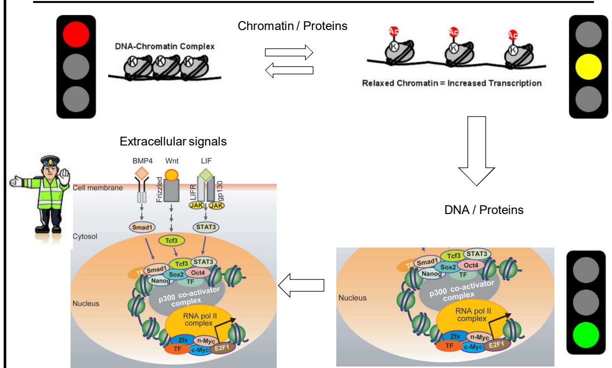
Epigenomics is not Epigenetics

Epigenetics is the study of heritable changes in gene expression or cellular phenotype, caused by mechanisms other than changes in the underlying DNA sequence

There are objections to the use of the term epigenetic to describe chemical modification of histone, since it remains unknown whether or not these modifications are heritable.



Gene Regulation

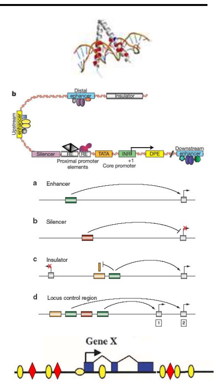


Cis-Regulatory Components

Low level ("atoms"):
- Promoter motifs (TATA box, etc)
- Transcription factor binding sites (TFBS)

Mid Level:
- Promoter
- Enhancers
- Repressors/silencers
- Insulators/boundary elements
- Locus control regions

High Level:
- Epigenomic domains / signatures
- Gene regulatory domains
- Gene regulatory networks

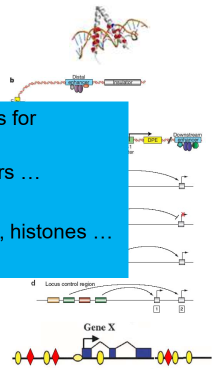


One genome to rule them all

Low level ("atoms"):
- Promoter motifs (TATA box, etc)
- Transcription factor binding sites (TFBS)

Mid Level:
- Promoter
- Enhancer
- Represso
- Insulators
- Locus cor

High Level:
- Epigenom
- Gene regulatory domains
- Gene regulatory networks



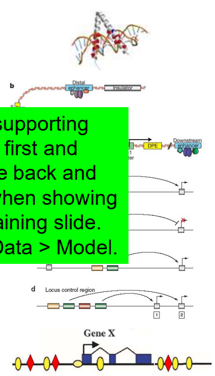
Note that the genome codes for everything we discussed:
Motifs, promoters, enhancers …
But also
RNA polymerase, ribosome, histones …
Everything

Remember: It's All Empirical Science

Low level ("atoms"):
- Promoter motifs (TATA box, etc)
- Transcription factor binding sites (TFBS)

Mid Level:
- Promoter
- Enhancer
- Represso
- Insulators
- Locus cor

High Level:
- Epigenomic domains / signatures
- Gene regulatory domains
- Gene regulatory networks

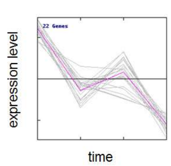


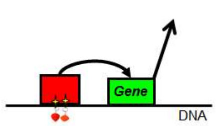
I show you model first and supporting data after. In reality it's data first and model after, or at least some back and forth between the two. Eg when showing gene in sito vs enhancer staining slide. But also everywhere else. Data > Model.

Inferring Gene Expression Causality

Measuring gene expression over time provides sets of genes that change their expression in synchrony.
- But who regulates whom?
- Some of the necessary regulators may not change their expression level when measured, and yet be essential.

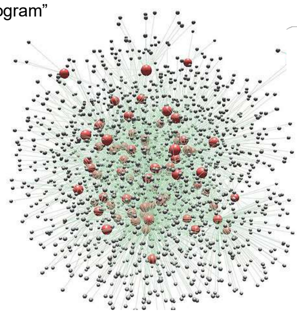
"Reading" enhancers can provide gene regulatory logic:
- If present(TF A, TF B, TF C) then turn on nearby gene X

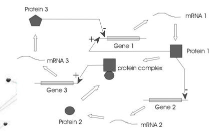




Obtain a network of all active genes & DNA

"Ridicilogram"





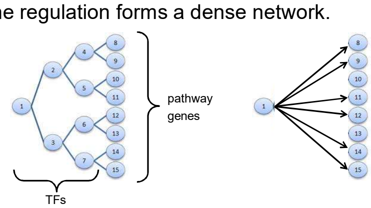
Now what?
(to be revisited)

Transcription Factors have Large "fan outs"

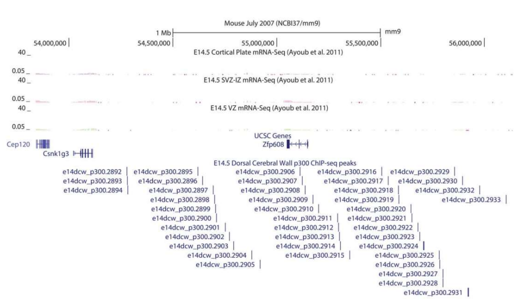
We could have had one TF regulate two TFS, each of which regulates two other TFs, etc. and each of those contributing to the regulation of a modest number of target genes (that do the real work).

Instead TFs reproducibly bind to thousands of genomic locations almost anywhere we've looked.

Gene regulation forms a dense network.



Some important genes have large "fan ins"



We are technically DONE with genome function

Biology – not that complicated!!

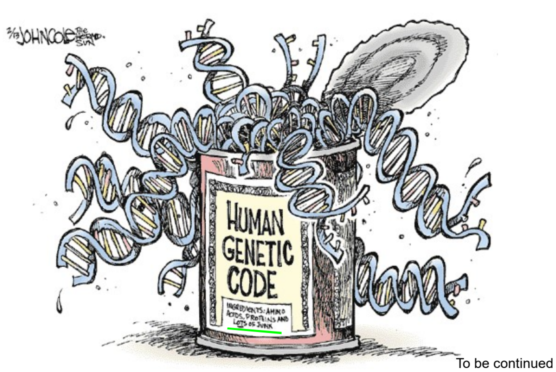
Functional part list

- In our genome:
  - Gene
    - Protein coding
    - Non coding / RNA genes
  - Gene regulatory elements
    - "Atomic" event: transcription factor binding site
    - Build up: promoters, enhancers, silencers, gene reg. domain
- "Around" our genome
  - Chromatin – open / closed
  - Epigenomic (and some epigenetic) marks

A Quiz!

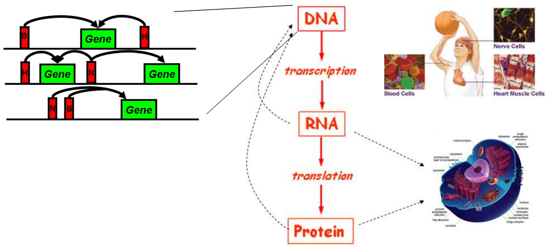
Let's quiz Gill on the board…

So Why This??



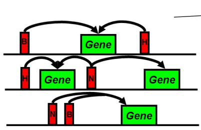
To be continued

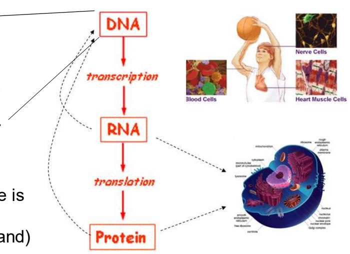
The Functional Genome



Type	# in genome
genes	20,000
ncRNA	20,000
cis elements	1,000,000

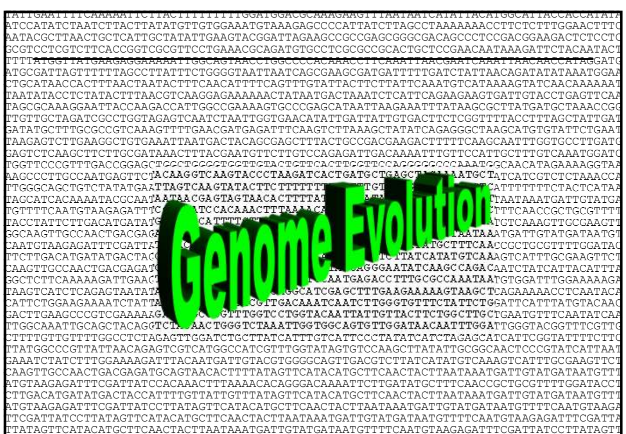
The Functional Genome





Corollary: most of the genome is devoid of function (which we understand)

Type	# in genome	% of genome
genes	20,000	2-3%
ncRNA	20,000	2%
cis elements	1,000,000	10-15%

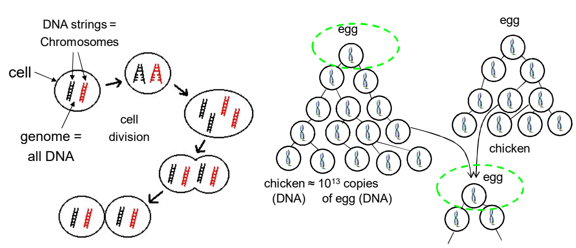


"Nothing in Biology Makes Sense
Except in the Light of Evolution"

Theodosius Dobzhansky

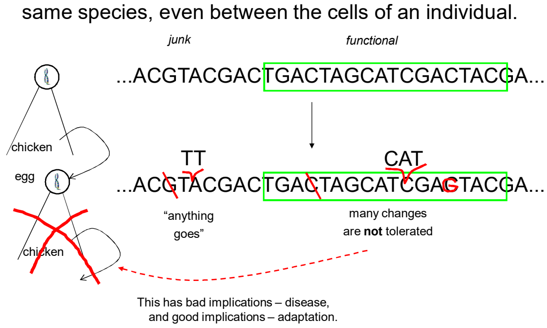
One Cell, One Genome, One Replication

Every cell holds a copy of all its DNA = its genome.
The human body is made of ~10^{13} cells.
All originate from a *single* cell through *repeated* cell divisions.

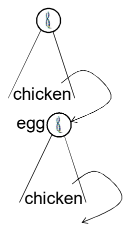


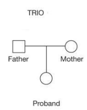
Every Genome is Different

DNA Replication is imperfect – between individuals of the same species, even between the cells of an individual.

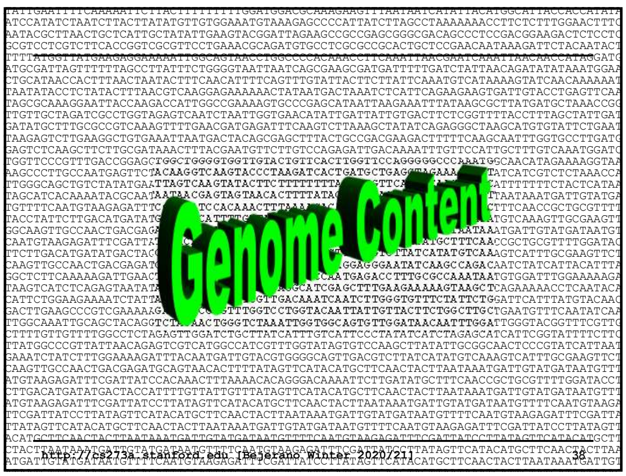


Human Mutation Rate

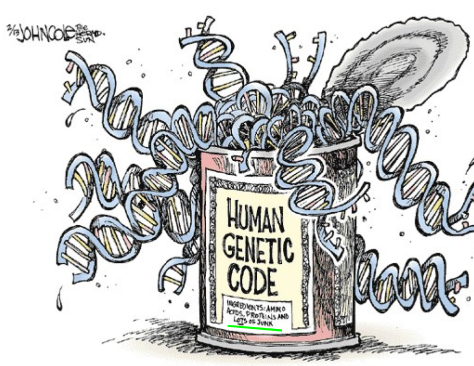




- Recent sequencing analysis suggests ~40-80 new mutations in a child that were not present in either parent: Roughly 1 mutation per replication!
- Mutations range from the smallest possible (single base pair change) to the largest – whole genome duplication (to be discussed).
- Selection does not tolerate *all* of these mutation, but it sure does tolerate many.



Why this cartoon?



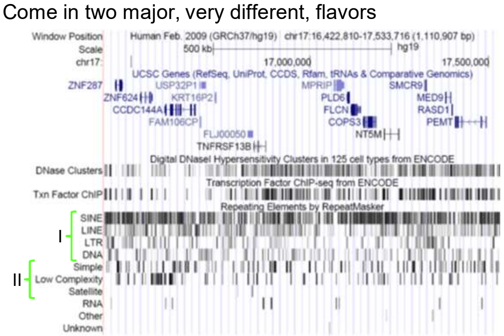
Genome Composition

The functional genome takes about 20% of the genome.
The remaining 80% is far from homogeneous…

http://cs273a.stanford.edu [Bejerano Winter 2020/21] 40

Sequences that repeat many times in the genome

- Take up cumulatively a whooping *half* of the genome
- Come in two major, very different, flavors



http://cs273a.stanford.edu [Bejerano Winter 2020/21] 41

I. Interspersed Repeats / TEs

Transposable elements are pieces of genetic information that somehow manage to multiply themselves and move around in the genome.

[Adapted from Lunter]

http://cs273a.stanford.edu [Bejerano Winter 2020/21] 42

I. Interspersed Repeats / TEs

Transposable elements are pieces of genetic information that somehow manage to multiply themselves and move around in the genome.

History: First suspected in 1940 from work by Barbare McClintock on genomic instability in maize. Existence of transposable elements was proven experimentally in 1970s. She received Nobel prize in 1983.

[Adapted from Lunter]

I. Interspersed Repeats / TEs

Transposable elements are pieces of genetic information that somehow manage to multiply themselves and move around in the genome.

History: First suspected in 1940 from work by Barbare McClintock on genomic instability in maize. Existence of transposable elements was proven experimentally in 1970s. She received Nobel prize in 1983.

Four classes of transposable elements live in our genome:

- DNA transposons
- LINEs (long interspersed nuclear elements), retroposons
- SINEs (short interspersed nuclear elements), non-autonomous retroposons
- Retroviruses and retrovirus-like LTR (long terminal repeat) retrotransposons

[Adapted from Lunter]

Classes of Interspersed Repeats

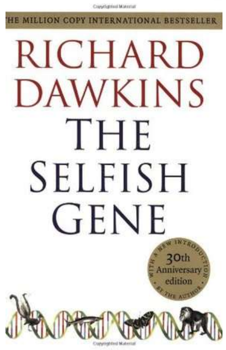
Classes of interspersed repeat in the human genome

Element	Transposition	Structure	Length	Copy number	Fraction of genome
LINEs	Autonomous	ORF1 ORF2 *(pol)* AAA	1–5 kb	20,000–40,000	21%
SINEs	Nonautonomous	A B AAA	100–300 bp	1,500,000	13%
Retrovirus-like elements	Autonomous	*gag* *pol* *(env)*	6–11 kb	450,000	8%
	Nonautonomous	*gag*	1.5–3 kb		
DNA transposons	Autonomous	*transposase*	2–3 kb	300,000	3%
	Nonautonomous		80–3000 bp		

- LINEs and SINEs were first distiguished by their length. Turned out to have different 'lifestyle' and are now distinguished by that.
- DNA transposons and retrotransposons code for *transposase* (or related *integrase*). Insert double-stranded DNA into host genome.
- LINE retroposons and retrovirus-like retrotransposons code for *reverse transcrip-tase*. Go through intermediate RNA phase.

From: Nature, Feb. 2001

The Selfish Gene



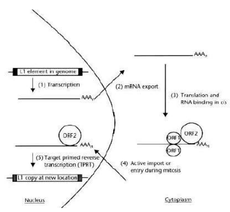
- What do you really share with your ancestors?
- What is the physical(..) manifestation of bloodline?
- Is the dog wagging its tail or is it the other way around?
- (How) is genome engineering changing this fundamental equation?

LINE & SINE Elements

LINEs have

- their own (pol II) promotors,
- two ORFs coding for protein,
- 3' binding site for ORF2 protein,
- poly-A tail

Act in cis, i.e. proteins coded by LINE bind to own mRNA.

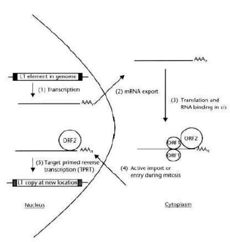


LINE & SINE Elements

LINEs have

- their own (pol II) promotors,
- two ORFs coding for protein,
- 3' binding site for ORF2 protein,
- poly-A tail

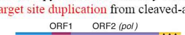
Act in cis, i.e. proteins coded by LINE bind to own mRNA.



After translation and binding to own mRNA, the LINE element:

From: EHG R 53

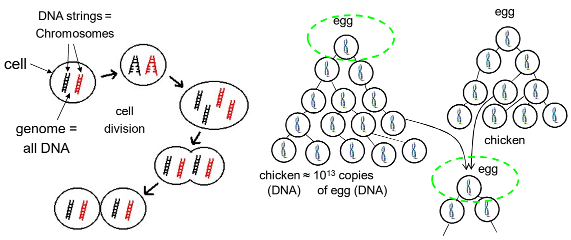
- Gets transported back into nucleus;
- Cleaves host DNA, preferentially at `TT|AAAA`;
- Transcribes a DNA copy from RNA directly into genome. New copy is flanked by a 7–20 bp target site duplication from cleaved-and-repaired host DNA.

LINEs	Autonomous		1–5 kb	20,000–40,000	21%
SINEs	Nonautonomous		100–300 bp	1,500,000	13%

Genomic Transmission

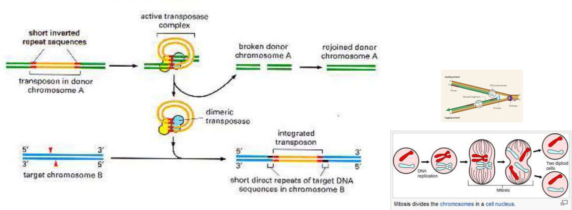
For repeat copies to accumulate through human generations they must make it into the germline cells (eggs & sperms).

Equally true for any genomic mutation.



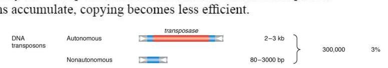
DNA Transposons

Transposons move by a cut-and-paste mechanism.



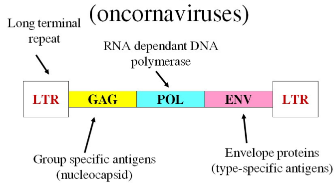
Multiply when excising themselves during mitosis, when DNA repair mechanisms can recover removed portion from newly duplicated strand.

Work in trans, i.e. gene gets translated, then transposase looks for "itself" in genome. Recognises itself by 10 – 30 bp stretch, so often binds to inactive transposon. Result: mutations accumulate, copying becomes less efficient.

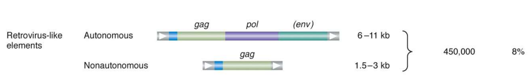


Retrovirus-like Elements

The viral genome (oncornaviruses)

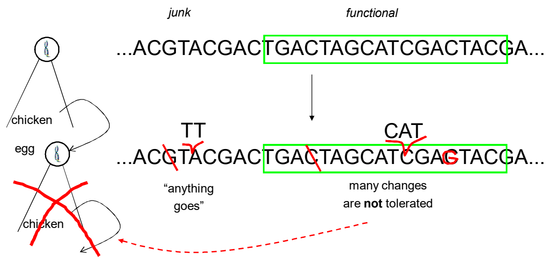


All three genes - GAG, POL, ENV - required for replication



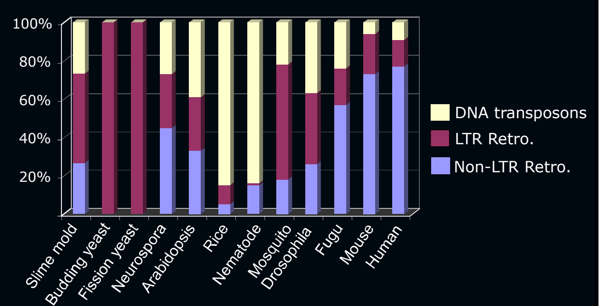
Every Genome is Different

DNA Replication is imperfect – between individuals of the same species, even between the cells of an individual.



This has bad implications – disease, and good implications – adaptation.

TE composition and assortment vary among eukaryotic genomes



Feschotte & Pritham 2006

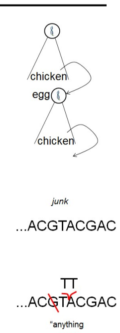
Repeats: mostly neutral

Most repeat events/instances are neutral.

Ie, a repeat instance is dropped in a new place, and joins the rest of the neutral DNA, gradually decaying over time.

Many repeat copies are "dead as a duck" on arrival at their new location (eg 5' truncation).

Some instances may be active (spawn new instances) for a while, but when an active copy is hit by a mutation – the host is not affected, the instance is inactivated and decays away.

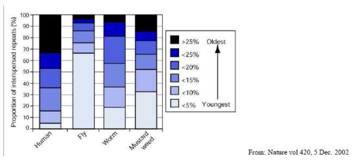


Repeat Ages

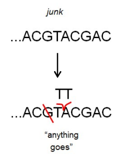
Activity of transposable elements

Activity varies greatly per organism:

- Humans: Rather quiet, ≈ 50 active LINEs, no or very few active DNA transposons, no LTRs through to be active.
- Mice: ≈ 3000 active LINEs, many active DNA transposons, many active LTRs.
- Maize: Genome size doubled in last ≈ 3 Myr because of transposon insertions.



From: Nature vol 420, 5 Dec. 2002



In fruitfly, most TEs have few mutations (relative to concensus = ancestor): young.

In human DNA, there are relatively few young transposable elements.

INTERSPECIES VARIATION IN GENOME SIZE WITHIN VARIOUS GROUPS OF ORGANISMS

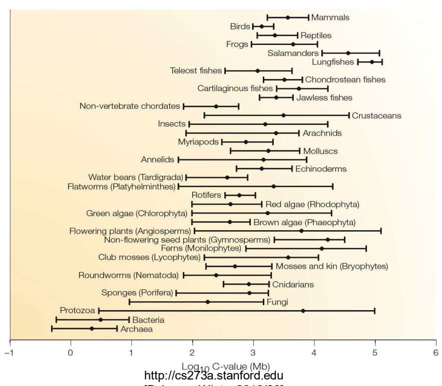
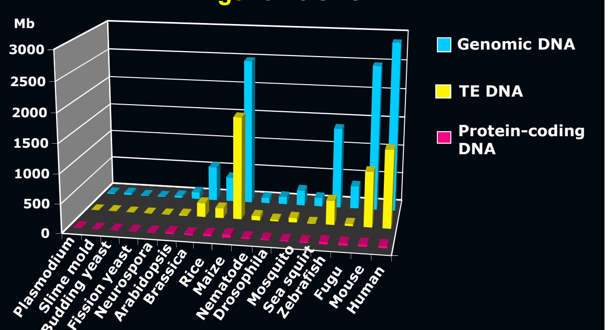


Figure from Ryan Gregory (2005)

The amount of TE correlate positively with genome size



Feschotte & Pritham 2006

The proportion of protein-coding genes decreases with genome size, while the proportion of TEs increases with genome size

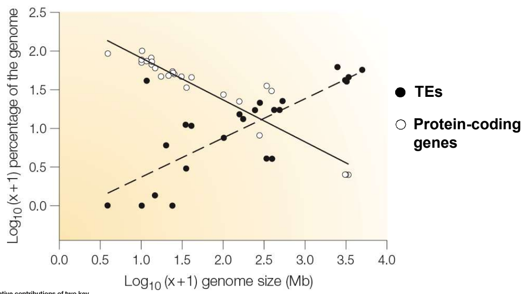


Figure 1 | **The relative contributions of two key components of eukaryotic genomes.** The relationships between haploid genome size and the percentage of the genome that consists of protein-coding genes (white circles) and transposable elements (black circles) are shown. The data are based on species that have been the subject of large-scale sequencing studies. Larger genomes contain proportionately fewer genes and more transposable elements than small genomes. A $\log_{10}(x+1)$ transformation was used because some tiny genomes contain no recognizable transposable elements.

Gregory, Nat Rev Genet 2005

Repeats: not just neutral

So far we treated all repeat proliferation events as neutral.

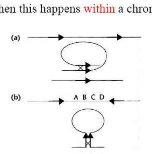
While the majority of them appear to be neutral, this is certainly not the case for all repeat instances.

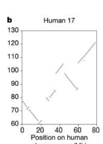
And because there are so many repeat instances even a small fraction of all repeats can be a big set compared to other types of elements in the genome. (Eg, 1% of ½ the genome is still a lot)

Transposable elements: Effect on genome

High copy number of transposable elements provide many opportunities for unequal homologous recombination.

When this happens within a chromosome, leads to deletions or inversions.





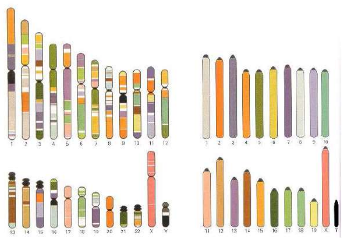
EHG T 622; Nature Feb. 2001

Direct evidence:

- Existence of solo-LTRs, result of recombination between two LTRs flanking one (or two) LTR-retrotransposon(s). EHG T 622
- 20% of Alus have no flanking target-site repeats. CW Schmid, Nuc Acids Res 1998 26(20) 4541

Transposable elements: Effect on genome

When unequal homologous recombination occurs *between* chromosomes, chromosome rearrangements occur.



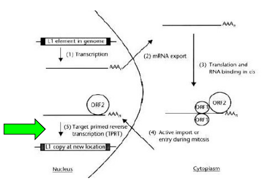
From: Alberts et al., The Cell, after Nature vol 420, 5 Dec. 2002

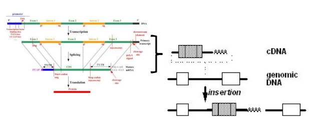
(Right: mouse chromosomes. Left: human chromosomes, colored according to which mouse chromosome region correspond to)

Repeats & Retroposed Genes

Remember how LINEs reverse transcribe copies of themselves back into the genome? How they sometimes reverse transcribe SINEs “by mistake”? Well, they also grab m/ncRNAs and reverse transcribe them into the genome!

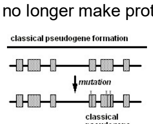
Retrogenes (“retrotranscribed”):
Protein coding RNA that was reverse transcribed and inserted back into the genome.
The RNA can be grabbed at any stage (partial/full transcript, before/during/after all introns are spliced).

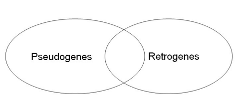




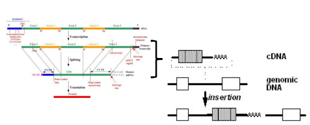
Retroposed Genes & Pseudogenes

Pseudogenes (“dead genes”):
Genomic sequences that resemble (originated from) genes that no longer make proteins.





Retrogenes (“retrotranscribed”):
Protein coding RNA that was reverse transcribed and inserted back into the genome.
The RNA can be grabbed at any stage (partial/full transcript, before/during/after all introns are spliced).



Repeat Insertions Can “Break Things”

1: Brain Dev. 2007 Mar;29(2):105-8. Epub 2006 Dec 18.

The first reported case of Menkes disease caused by an Alu insertion mutation.

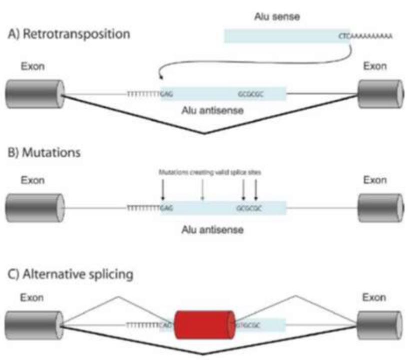
Gu Y, Kodama H, Watanabe S, Kikuchi N, Ishitsuka I, Ozawa H, Fujisawa C, Shiga K.

Department of Health Policy, National Research Institute for Child Health and Development, 2-10-1 Okura, Tokyo, Japan. gyh@nch.go.jp

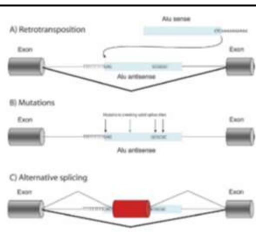
We present the first reported case of Menkes disease caused by an Alu element insertion mutation that interfered with splicing regulatory elements. A whole young AluYa5a2 element, which was 382-bp long, was identified within exon 9 of the ATP7A gene, and all of exon 9 was aberrantly skipped in the cDNA, resulting in severely truncated proteins. To confirm whether the aberrant skipping resulted in Alu insertion, an exonic splicing enhancer finder was used. The Alu element created two new high-score exonic splicing enhancer sequences in the mutation located near the site of the insertion. Exon 9, which encodes the first and second transmembrane domains, is necessary for the normal function of the ATP7A protein.

PMID: 17178205 [PubMed - indexed for MEDLINE]

Repeat Insertions Can “Make Things”



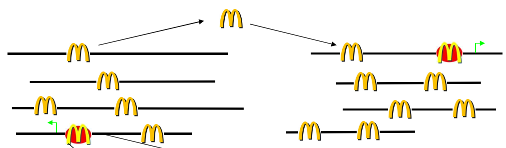
Any Sequence Can Become Functional



Random mutation (especially in a large place like our genome) can create functional DNA elements out of neutrally evolving sequences.

So is there anything special about a piece of DNA from a repetitive origin that takes on a new function?

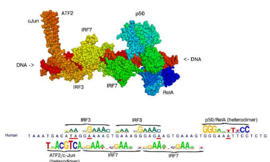
Regulatory elements from Mobile Elements



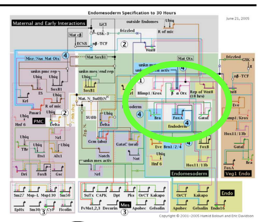
Co-option event, probably due to favorable genomic *context*

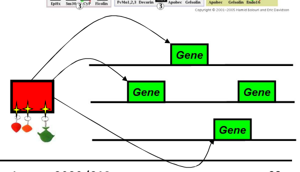
[Yass is a small town in New South Wales, Australia.]

Britten & Davidson Hypothesis: Repeat to Rewire!

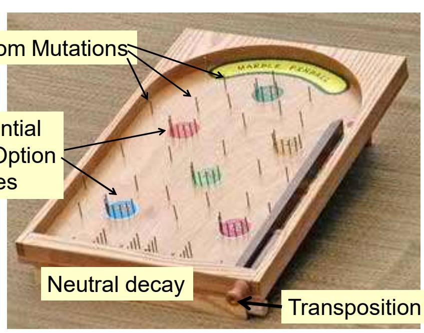


Enhancer structure reminder

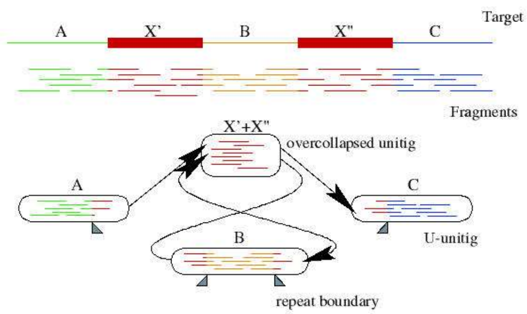




The Road to Co-Option



Assemby Challenges



Inferring Phylogeny Using Repeats

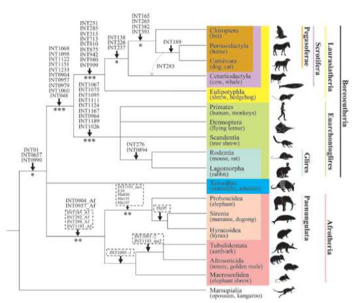
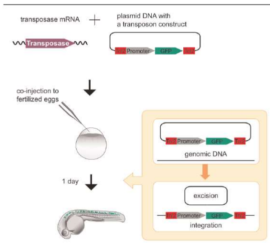


Fig. 2. An interordinal mammalian phylogeny reconstructed by our retroposon insertion analysis. Downward arrows denote insertions of retroposons into each lineage. Locus INT283, denoted by a dashed arrow, supports the monophyly of Cetartiodactyla, Perissodactyla, and Carnivora. The loci surrounded by dashed lines in Afrotheria were identified in our previous study (22). Asterisks below the branches denote that the monophylies are statistically significant (*, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$).

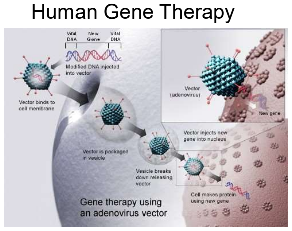
[Nishihara et al, 2006]

Transposons as Genetics Engineering Tools

The medaka fish *Tol2* element is an autonomous transposon that encodes a fully functional transposase. The transposase protein can catalyze transposition of a transposon construct that has 200 and 150 base pairs of DNA from the left and right ends of the *Tol2* sequence, respectively. These sequences contain essential terminal inverted repeats and subterminal sequences. DNA inserts of fairly large sizes (as large as 11 kilobases) can be cloned between these sequences without reducing transpositional activity. The *Tol2* transposon system has been shown to be active in all vertebrate cells tested thus far, including zebrafish, *Xenopus*, chicken, mouse, and human. In this review I describe and discuss how the *Tol2* transposon is being applied to transgenic studies in these vertebrates, and possible future applications.



Human Gene Therapy



Repeats: fun conspiracy theories

1. Repeats wreck so much havoc in the genome, by inserting themselves, deleting segments between instances and more – they make the genome feel like a “rolling sea”. Maybe it is because of them that enhancers evolved to work irrespective of distance and orientation?

2. When the last active copy of a repeat dies, all instances of the repeat are now decaying. Wait long enough and they lose resemblance to each other. Look in 200My and you never know they belonged to the same repeat family. So… if half the genome is recognizable as repetitive now, how much of the genome originated from repeats? Most of it?

Repeats: fun conspiracy theories

3. If repeats do significantly accelerate the rate of creation of novel functional (gene/regulation) elements – how many functional elements today came from repeats (including old ones we no longer can recognize as such)? Most?

4. Is that why our genome “tolerates” these elements?

5. You make a conspiracy theory…

6. You think of ways* to solve one!

* Computationally. Evolution is mostly computational business.