CS273A

Lecture 1: Overview

MW 1:30-2:50pm in Clark **S361*** (behind Peet's)

Profs: Serafim Batzoglou & Gill Bejerano

CAs: Karthik Jagadeesh & Johannes Birgmeier

* Handful of lectures/primers elsewhere: track

Welcome / back !

Announcements

- http://cs273a.stanford.edu/ is up
  - Course guidelines, lecture slides, office hours, etc.

- Course communications via Piazza
  - Auditors please sign up too

- Pre-reqs: None (in Bio or CS).
  - Programming knowledge (any language) strongly recommended
  - Biologists will learn new things (code or audit)

- Grade (see website):
  - Two homework assignments
  - Project: in groups, ~½ quarter
  - Attendance

Announcements

- Three tutorials:
  - Topics:
    - Introductory Biology Primer
    - UCSC Genome Browser Tools (rec: bring your laptop!)
    - Introduction to Text Processing
  - Times/locations:
    - Friday 9/25, Fri 10/2, Fri 10/9 @2pm in Beckman B302
    - Follow website or Piazza for final times & locations

- Relationship to other genomics classes:
  - CS173: Very similar to CS273A. Cannot take both. Not given this year.
  - CS262: Winter qtr. Perfect follow-up to CS273A. Algorithmic focus.
  - CS374: Spring quarter. Advanced seminar in genomics.
  - Other genomics classes in BMI, HumBio, Biology, Genetics, Stats, etc.

- Lots of genomics research happening on campus
  - If you enjoy this class many labs would love to have you!

CS300:
9/28 Serafim
10/5 Gill

Motivation

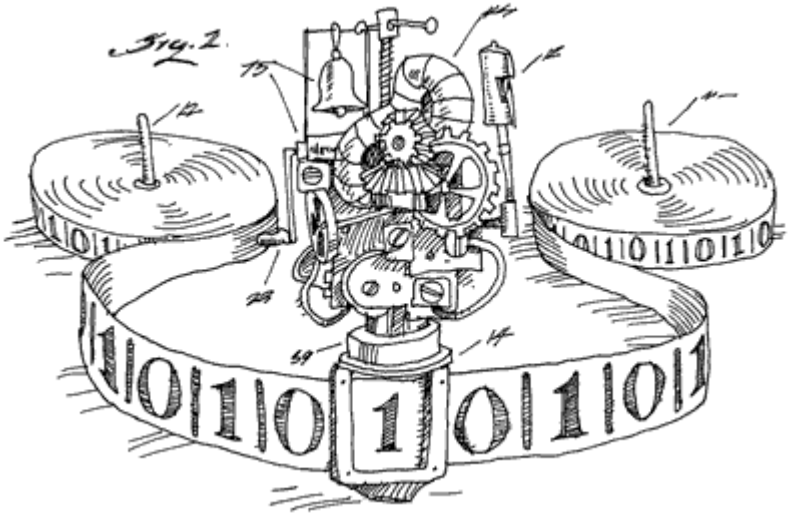
What will we study?

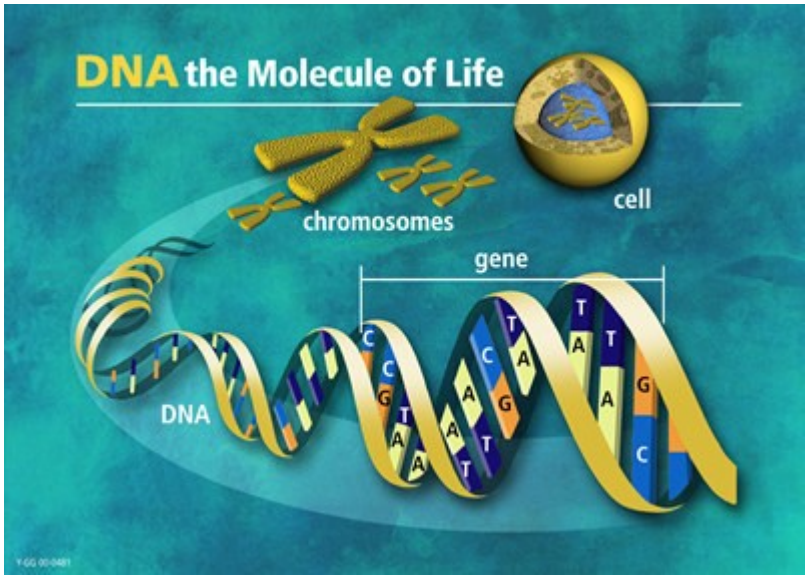
CS 273A: A Computational Tour of the Human Genome (BIOMEDIN 273A, DBIO 273A)

Introduction to computational biology through an informatic exploration of the human genome. Topics include: genome sequencing (technologies, assembly, personalized sequencing); functional landscape (genes, gene regulation, repeats, RNA genes, epigenetics); genome evolution (comparative genomics, ultraconservation, co-option). Additional topics may include population genetics, personalized genomics, and ancient DNA. Course includes primers on molecular biology, the UCSC Genome Browser, and text processing languages. Guest lectures from genomic researchers. No prerequisites. See http://cs273a.stanford.edu/.

Terms: Aut | Units: 3 | Grading: Letter or Credit/No Credit

The most amazing "Turing tape" in existence, your genome.






Genome Context
http://cs273a.stanford.edu [BejeranoFall15/16]

Organism – Cell - Genome

10^{13} different cells in an adult human.

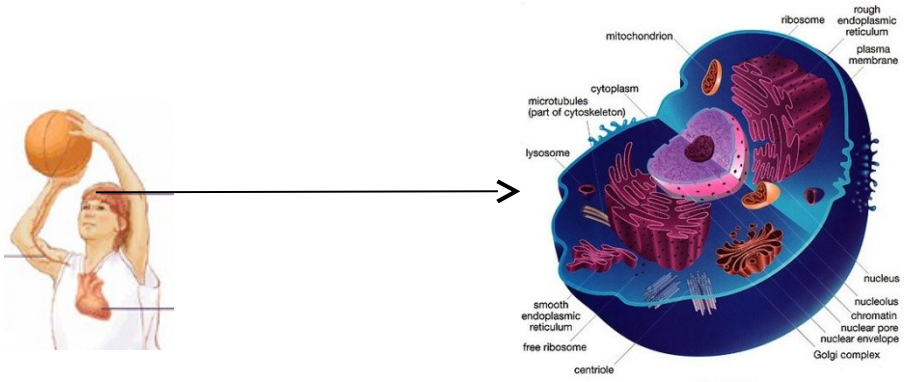
The cell is the basic unit of life.

DNA = linear molecule inside the cell that carries instructions needed throughout the cell's life ~ long string(s) over a small alphabet

Alphabet of four (nucleotides/bases) {A,C,G,T} Strings of length 10^4-10^{11}

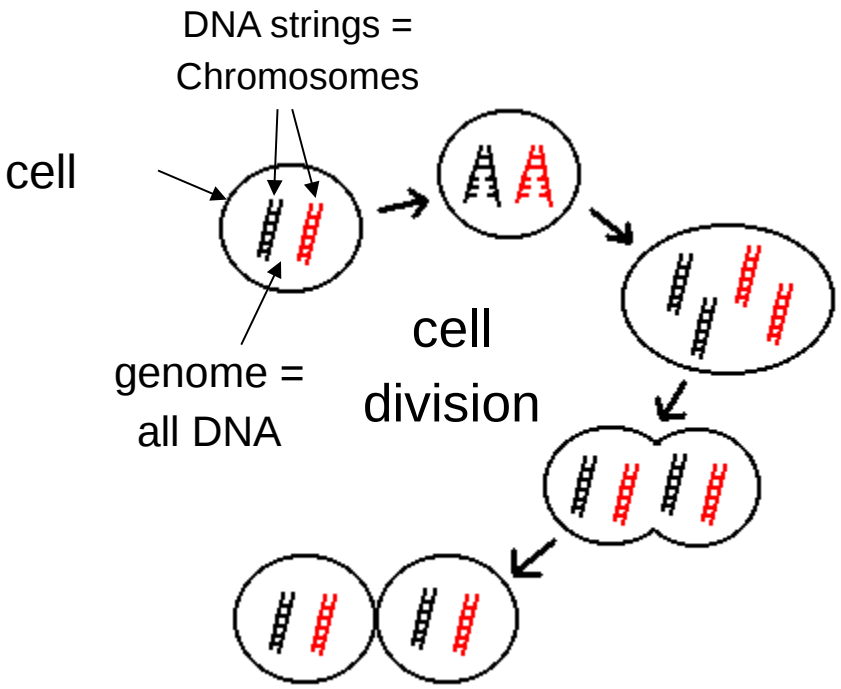
Genome: "instruction"

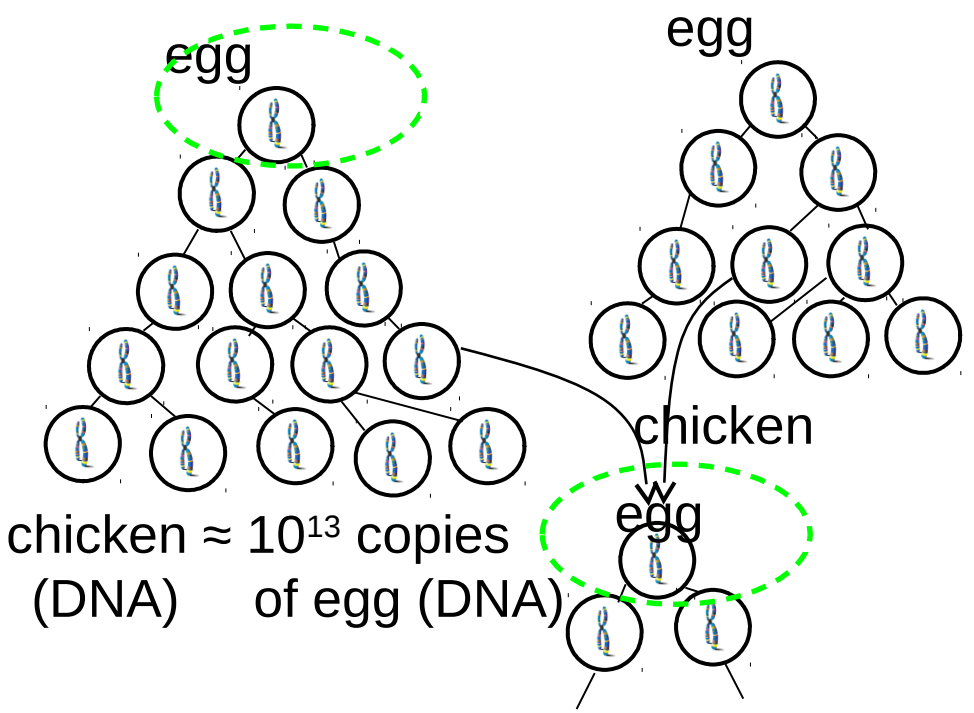
...ACGTACGACTGACTAGCATCGACTACGACTAGCAC...



One Cell, One Genome, One Replication

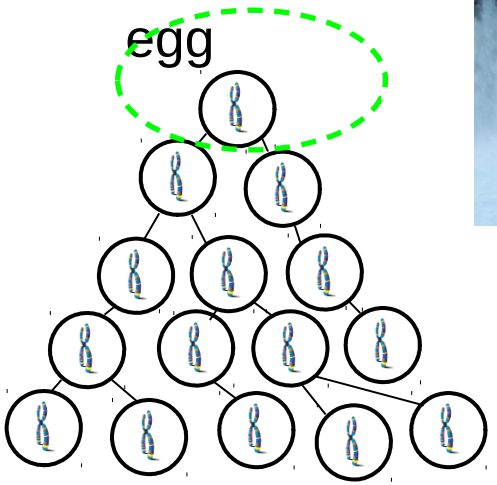
- Every cell holds a copy of all its DNA = its genome.
- The human body is made of ~10^{13} cells.
- All originate from a *single* cell through *repeated* cell divisions.





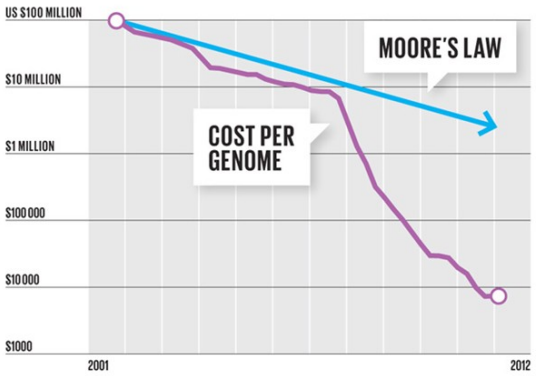
Talk about code reuse

- The same genome "runs" hundreds, likely thousands of very different cell types.
- Nature vs. Nurture – If life is like a river, our genome provides the boat…



human ≈ 10^{13} cells

Integrated Development Environment



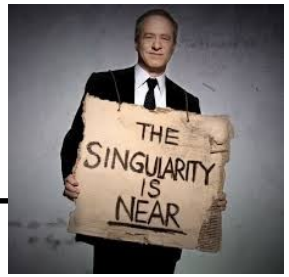
Read

Understand

Write

Bye bye natural selection
Bye bye human race …

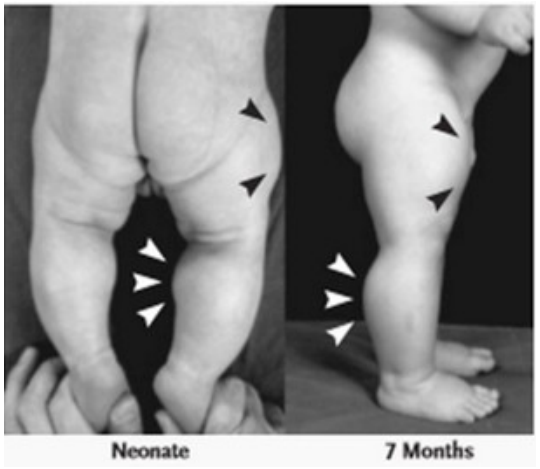
The Singularity

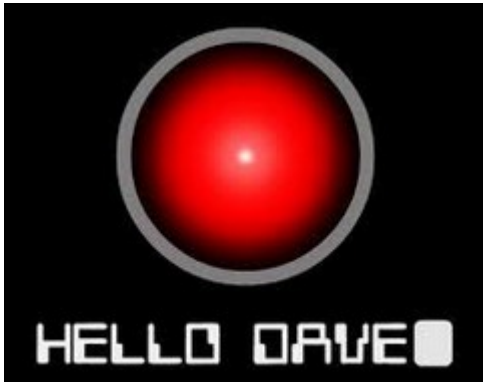


Singularity – (more) sentient non homo sapiens

The Singularity can come from one of:

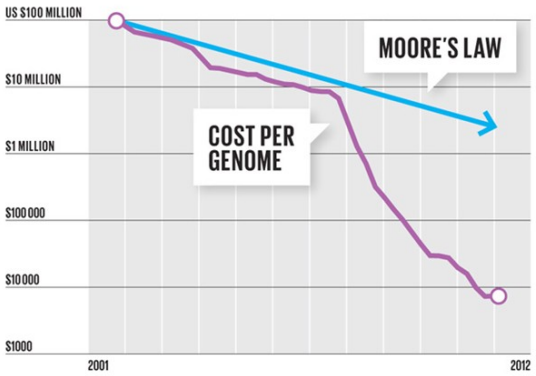
1. AI – Artificial Intelligence
2. Cyborg – Part wo/man part machine
3. **AH – Artificial Humanity**





single letter change
in the human genome

Integrated Development Environment



Read

Write

Understand

“Sending man to the moon is the easy bit, getting him back is the tough one. Similarly, getting the [human genome] sequence is the easy bit, understanding what the sequence means is the difficult bit.”

Sydney Brenner

The Biggest Challenge in Genomics…

… is <u>computational</u>:

How does this

encode *this*

GCTTTGCGCCGTCAAA
GTCTTGAAGGCTGTGA
TCAAGCTTCTTGCGAT
CCCGTTTGACCGGAGC
CTTGCCAATGAGTTCT
CAGCTGTCTATATGAA

Program

Output

This “coding” question has <u>profound</u> implications for our lives

The Biggest Challenge in Genomics…

… is <u>computational</u>:

How does this

encode *this*

Program

Bugs

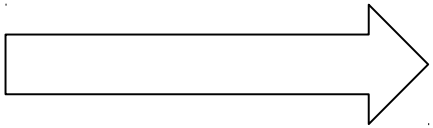
Output

What genomic mutations predispose us to <u>disease</u>?

The Biggest Challenge in Genomics…

… is <u>computational</u>:

How does this encode *this*



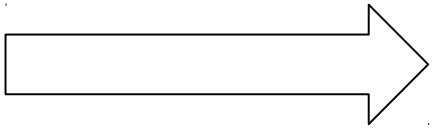
Program Bugs Patching

What genomic mutations determine our <u>drug response</u>?

The Biggest Challenge in Genomics…

… is <u>computational</u>:

How does this encode *this*



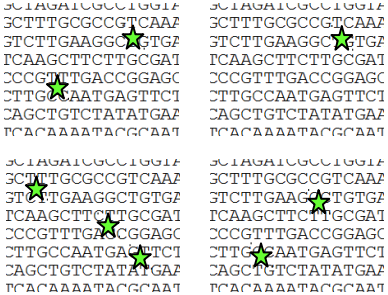
Program Bugs Debugging

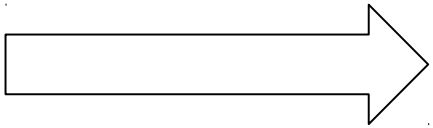
We are learning to alter our genome… But what to alter?

The Biggest Challenge in Genomics…

… is computational:

How does this encode *this*





Program

Output

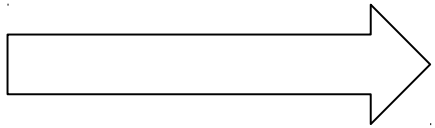
Where did we come from? How are we different from each other?

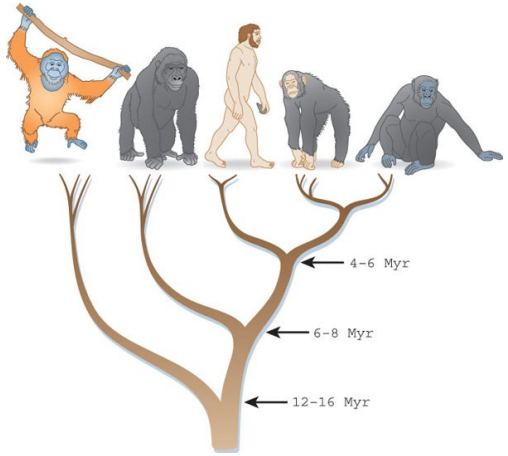
The Biggest Challenge in Genomics…

… is <u>computational</u>:

How does this

encode *this*





Program

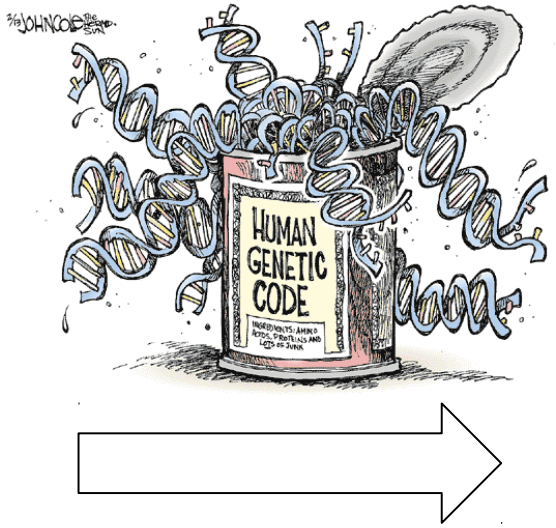
Output

What in our genomes make us different from other species?

The Biggest Challenge in Genomics…

… is <u>computational</u>:

How does this



encode *this*

Program

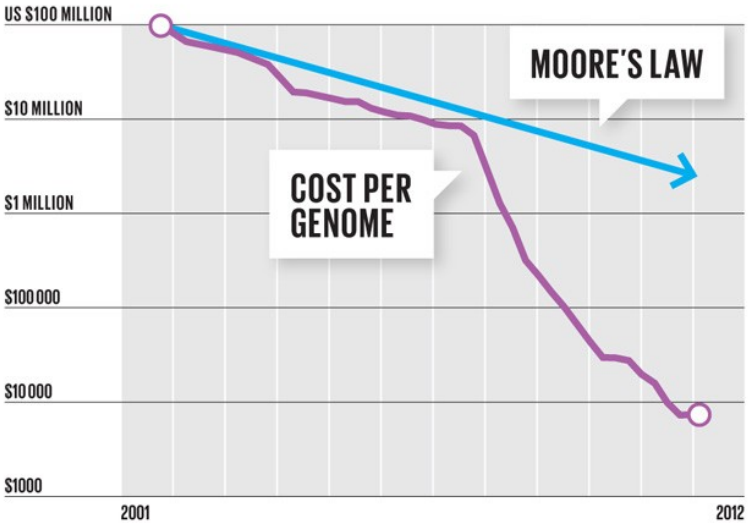
Output

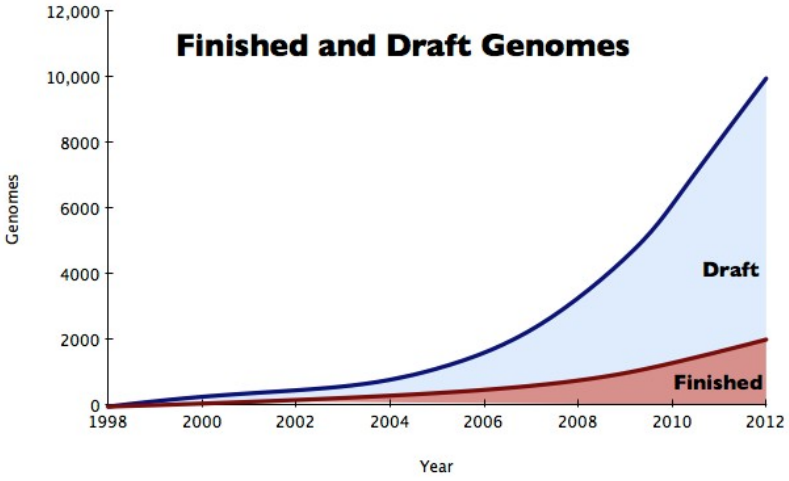
Why is our genome full of “memory leaks” and cruft?

Why Genomics?

- Rewriting the book on our understanding of life
- Growing impact on everybody's life
  - It is starting to save lives
- Genomics is an information/computational science
  - You can save lives from your keyboard
- This century is owned by Genomics
- "There is gold in them thar hills"
  - That gold can be *yours*

Genomics is affecting multiple fields of CS

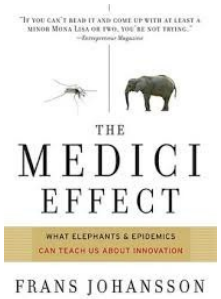




Storage
Compression
Architecture
Databases
HCI
etc.

Even if you do not want to be a genomicist, some of the most exciting challenges in your field may be at the interface with Genomics.

Most exciting things in Science happen at the interface of very different fields.



Computational Genomics

Genomics is three related fields bundled under one name:

- Technology development – build devices
- Functional genomics – do experiments
- Computational genomics – interpret results

Roles of computational biology (genomics):

- Summarize current experiment
- Discover the most exciting hypothesis / next experiment
- Develop new computational methods

CS273A focuses more on discovery.

CS262 on methods.

Why understand the why first?

Theoretical CS studies the hardness of questions.

A question is as hard as its easiest solution.

A lot of focus is put on how to answer questions.

In genomics (an empirical science) the temporal order is:

- What to ask and why?
- Can available/acquirable data answer it and if so how?
- Got data. Computed. What does my answer mean?

Advice: Reject the “us” (CS) and “them” (BIO) dichotomy

- Read what you need
- Develop your own taste for questions and answers

 Be a better methods developer, discoverer

Field Goals


PHASE TWO: INTERPRETATION
I THINK I FOUND A CORNER PIECE.
3 BILLION PIECES
GENOME

Class Goals

- Meet your genome (learn to surf, learn the surf)
- Understand genomic tools (theory, applications)
- DIY (pose questions, write & run tools, understand answers)

Class Topics

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

(1) Genome content / genome function:
genes, gene regulation, repeats, epigenetics

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

(3) Genome evolution:
evolution = mutation + selection, modes of evolution, comparative genomics, ultraconservation, exaptation

(4) Population genomics:
Tracking human migration patterns via neutral evolution

(5) Genomics of human disease:
disease susceptibility, cancer genomics, personal genomics

(6) Genome “output” (organism) evolution:
Evolutionary developmental biology (“evo-devo”)


Genome Content

Genomes, Genes & Proteins

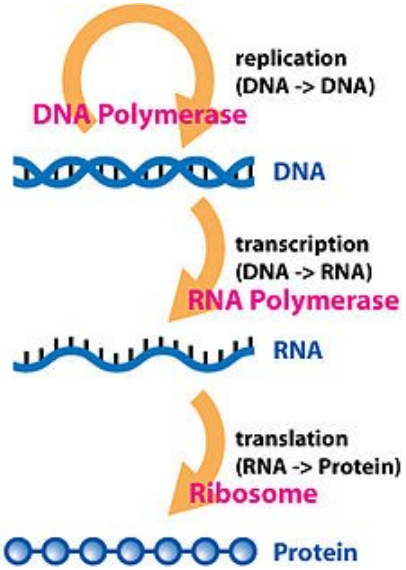
The most visible instructions in our genome are Genes.

Genes explain exactly HOW to synthesize any protein.

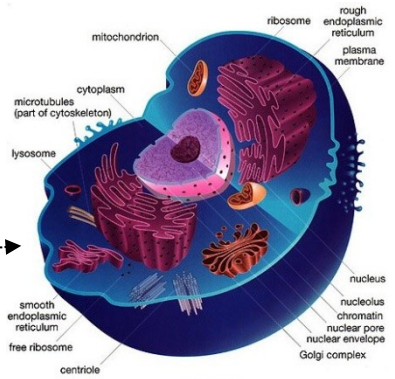
Proteins are the work horses of every living cell.

Genome: gene

...ACGTACGACTGACTAGCATCGACTACGACTAGCAC...



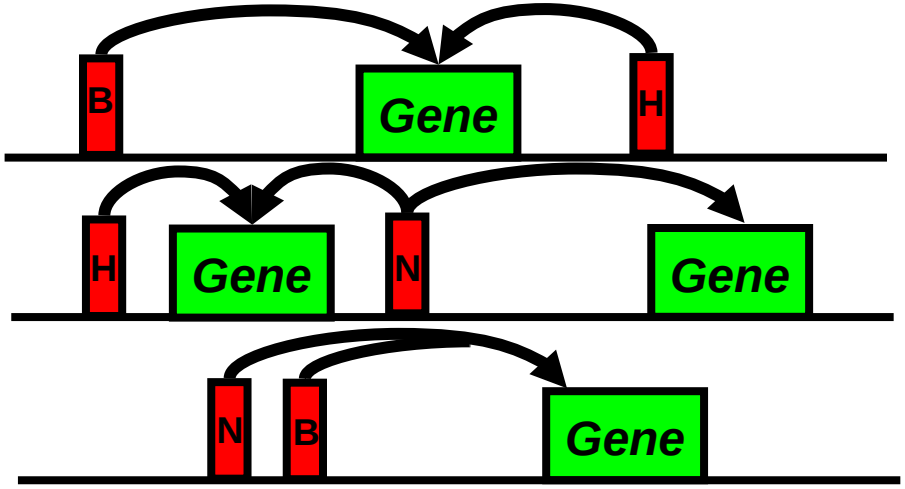
protein

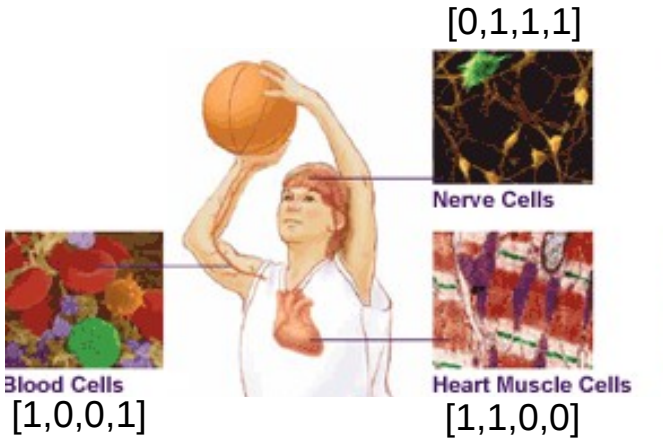


cell

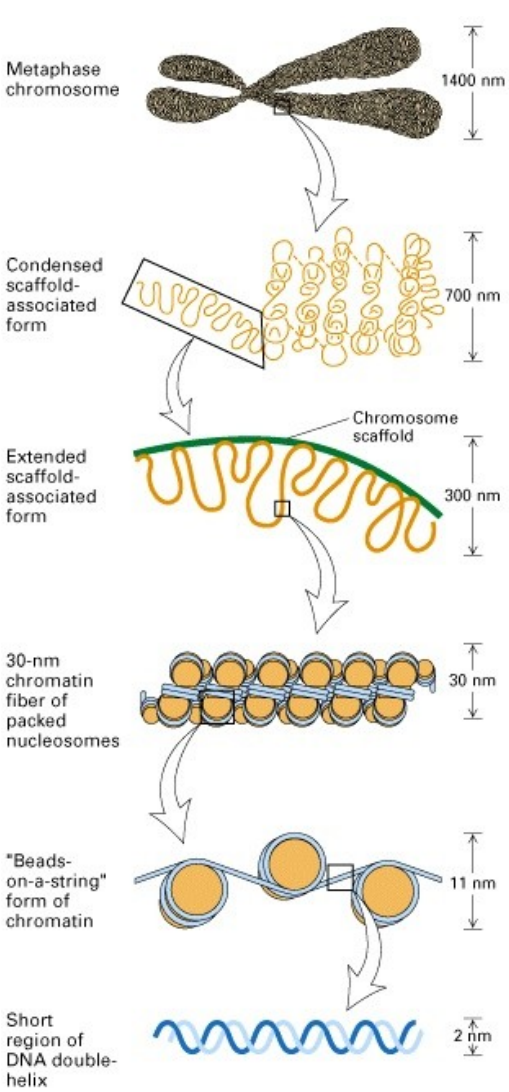
Genes & Gene Regulation

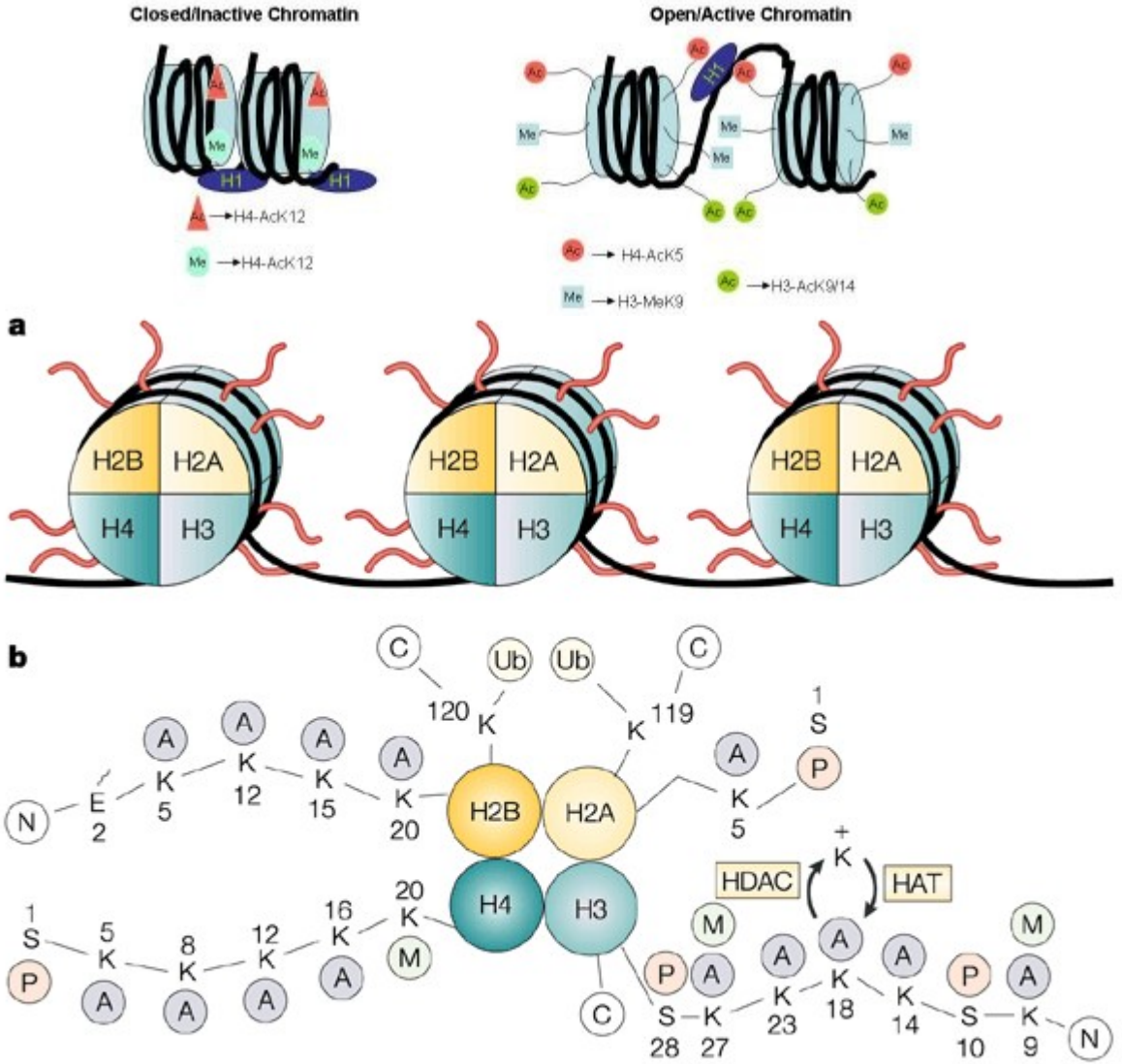
- Human genome encodes 20-25,000 genes (2% genome), >1,000,000 genomic switches that control genes (>10%).
- Gene = genomic substring that encodes HOW to make a protein.
- Genomic switch = genomic substring that encodes WHEN, WHERE & HOW MUCH of a protein to make.



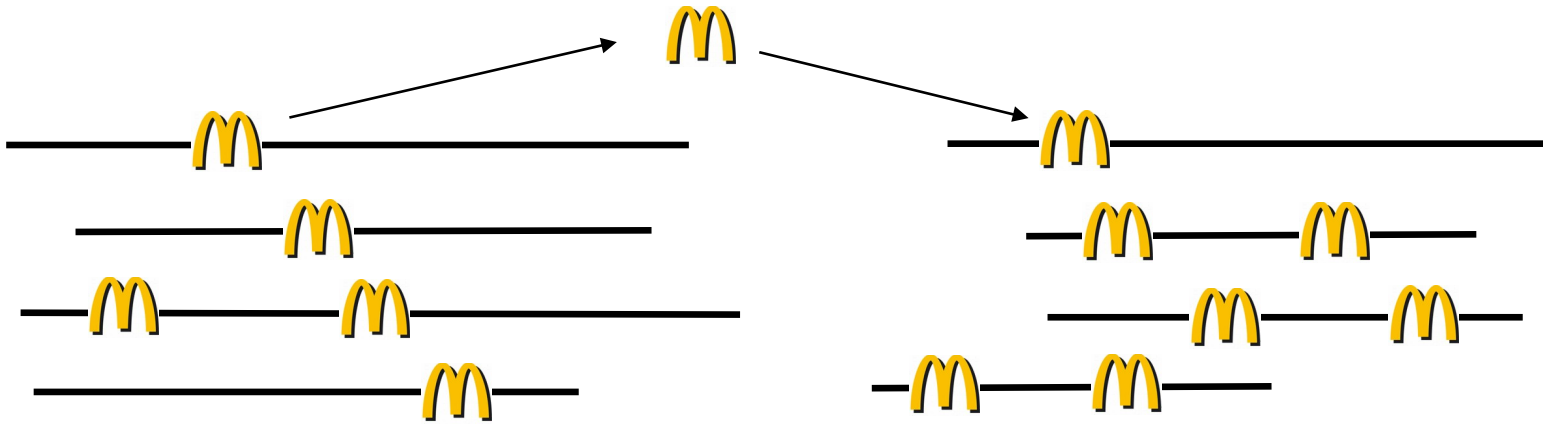


Epigenomics: transient writing “on” the genome





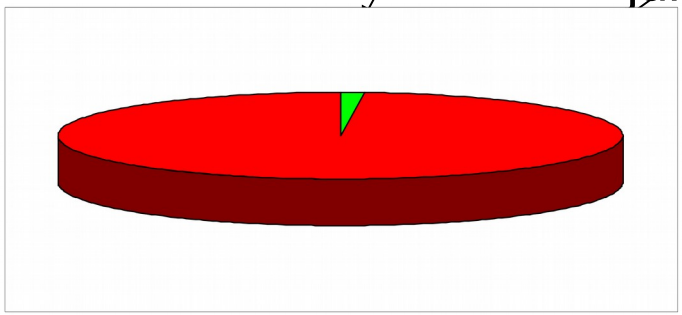
Repeats / Mobile Elements ("selfish/junk DNA")



Human Genome: $3*10^9$ letters

2% known function

>50% junk



Genome: conceptual part list

Genome:

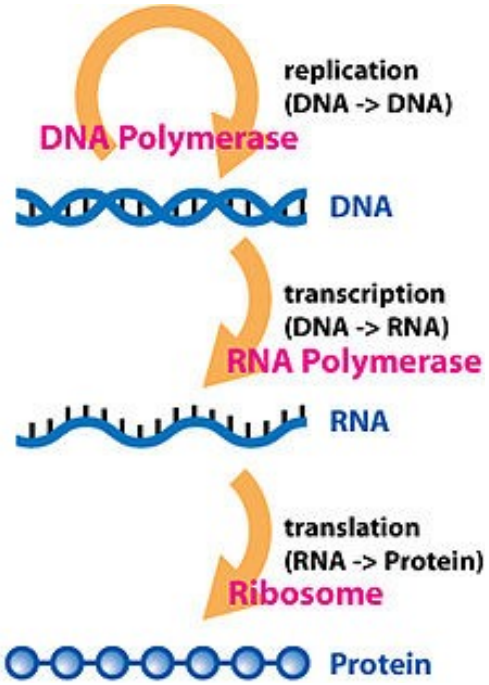
“instruction”

...ACGTACGACTGACTAGCATCGACTACGACTAGCAC...

A copy (actually two) in every cell.

Contains:

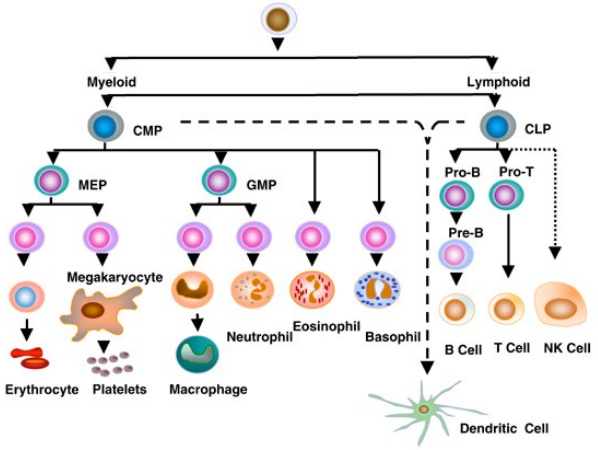
1. Genes
2. Gene regulation sequences
3. Repetitive DNA
4. Transient marks
   on genome packing material



Bottom line

- The Genome is really simple
  - Genes
  - Gene regulation

- The Genome is really fascinating
  - One code used in many contexts
  - Lots of code reuse
  - Output is breathtaking



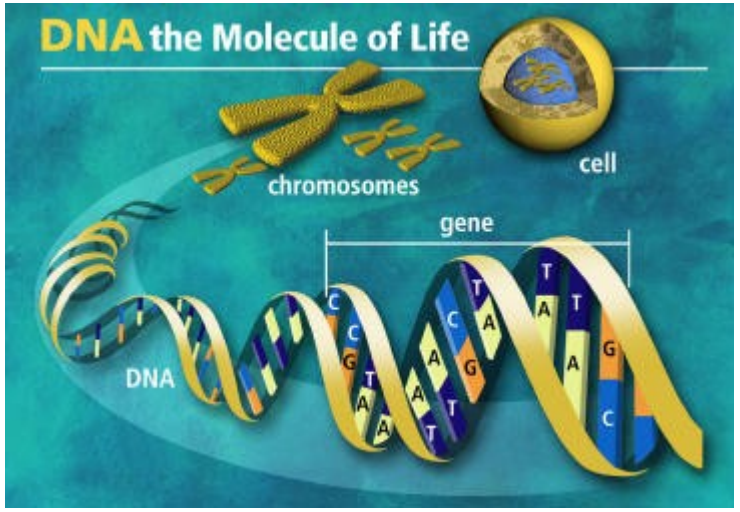
- Biology is a vast and deep sea
  - Humanity's biological knowledge is shallow
  - Dive anywhere and you quickly reach the frontier of human knowledge

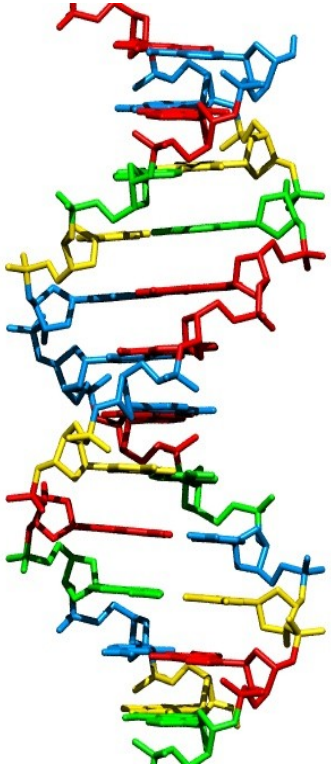

Genome Sequencing

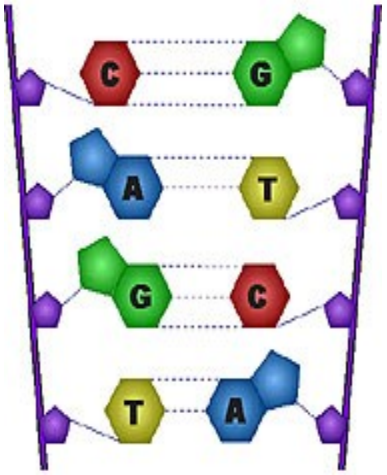
DNA sequencing

Genomes are awesome.

Let's get'em.



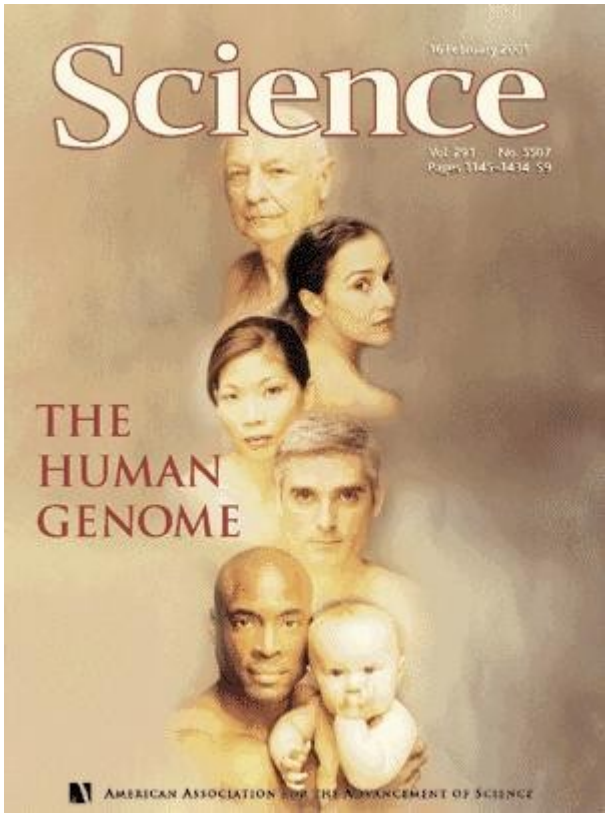




…ACGTGACTGAGGACCGTG
CGACTGAGACTGACTGGGT
CTAGCTAGACTACGTTTTA
TATATATATACGTCGTCGT
ACTGATGACTAGATTACAG
ACTGATTTAGATACCTGAC
TGATTTTAAAAAAATATT…

Genomic Drama

Graduate student Jim Kent wrote most of the software code used to assemble the working draft of the human genome sequence in this garage office. Photo: Don Harris



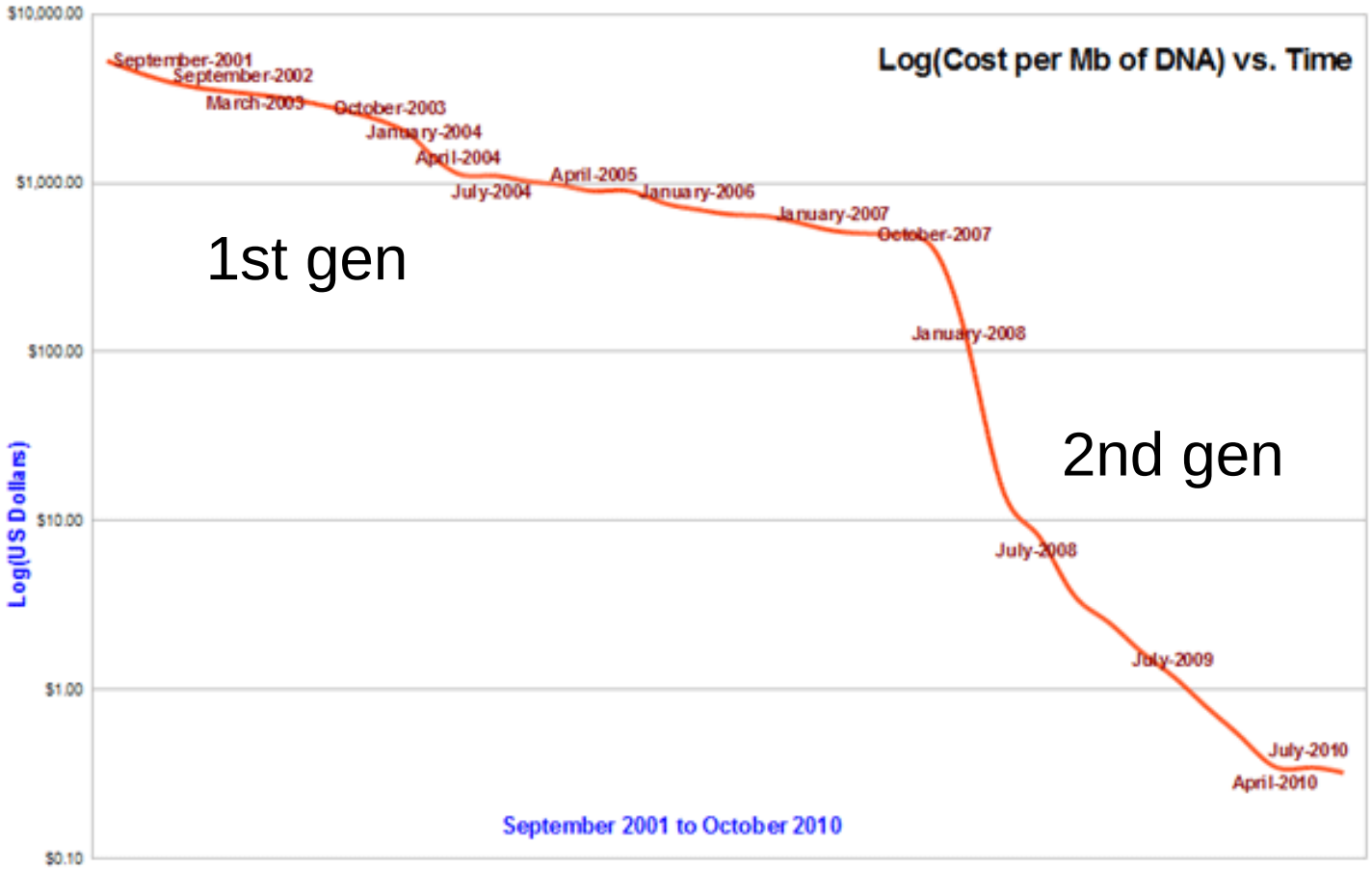
HGC

Celera

Getting the “blueprint of life”

2001

DNA sequencing costs



1st Genome Assembly

Some Terminology

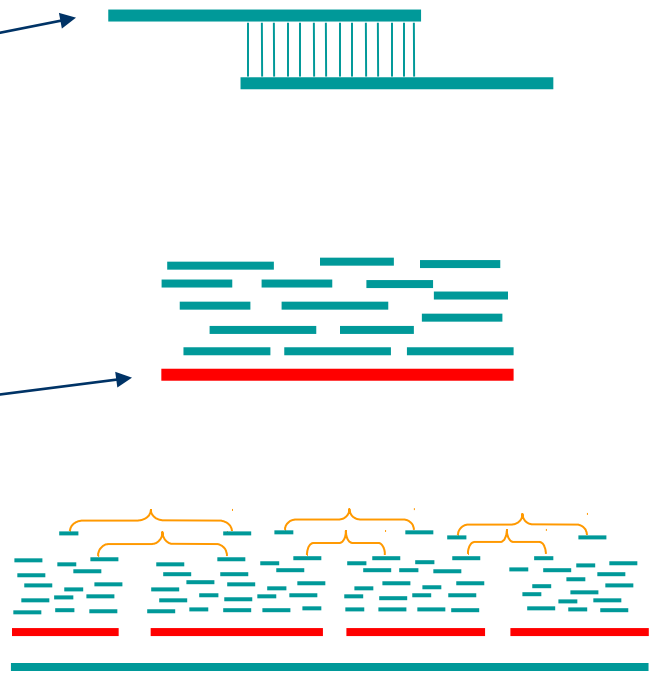
read a 500-900 long word that comes out of sequencer

mate pair a pair of reads from two ends of the same insert fragment

contig a contiguous sequence formed by several overlapping reads with no gaps

supercontig (scaffold) an ordered and oriented set of contigs, usually by mate pairs

consensus sequene sequence derived from the multiple alignment of reads in a contig



..ACGATTACAATAGGTT..

2nd Gen: Next Generation (re)Sequencing

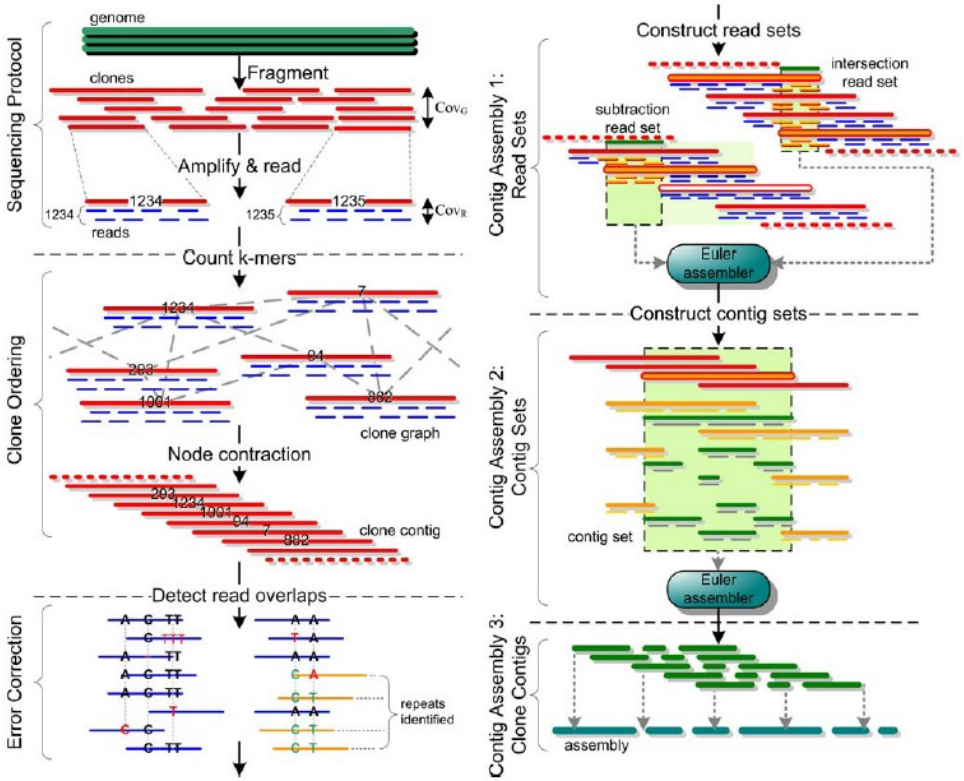


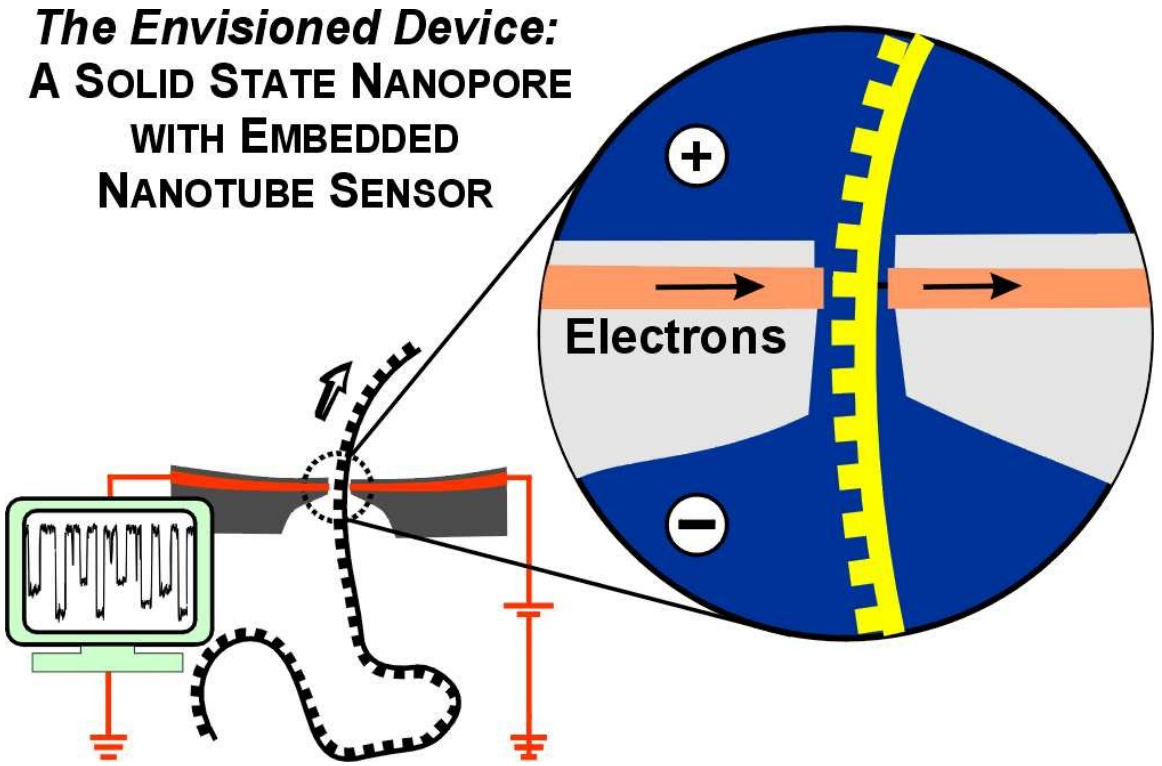
Figure 1. Sequencing protocol and assembly methodology. Reads are obtained in a hierarchical sequencing protocol with high genome-clone coverage and low clone-read coverage. From the *k*-mer content of each clone we construct a clone graph whose edge weights reflect the likely clone proximities, and from this our clone ordering algorithm determines the clone contigs. Next, we find all putative read overlaps by only looking in nearby clones and perform error correction. In three stages of contig assembly we 1) create *read sets* via set operations that consist of reads from multiple overlapping clones within small clone subregions and assemble using *Euler*, 2) combine contigs resulting from the previous stage in clone-sized *contig sets* for assembly, and 3) use a scalable assembler to merge entire clone contigs.
doi:10.1371/journal.pone.0000484.g001

Output = massive amounts of short, lower quality reads.

New Technologies + New Algorithms = New Opportunities

3rd Gen: cost effective, long reads

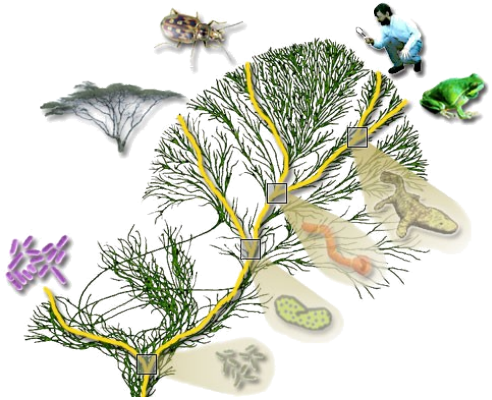
Just one example:



Output: very long reads of 10,000-100,000 basepairs each.

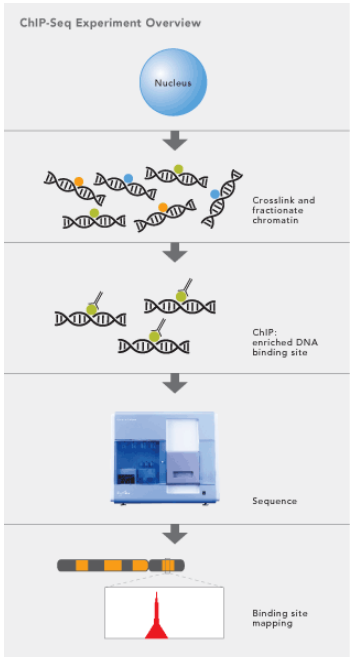
Sequence “anything” you like. In a lab. Trivial assembly.

Genomes, sequences everywhere

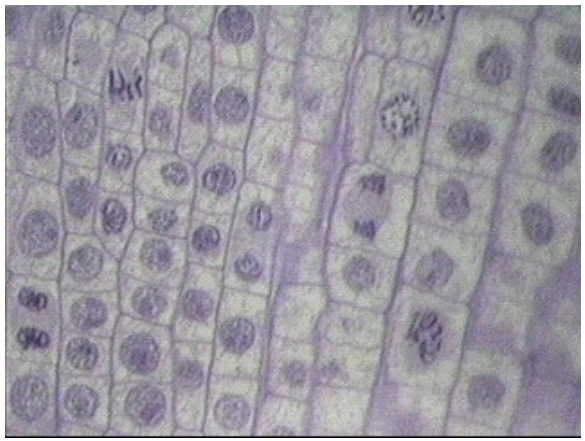


100 million species

7 billion
individuals



or sequence just
an active portion



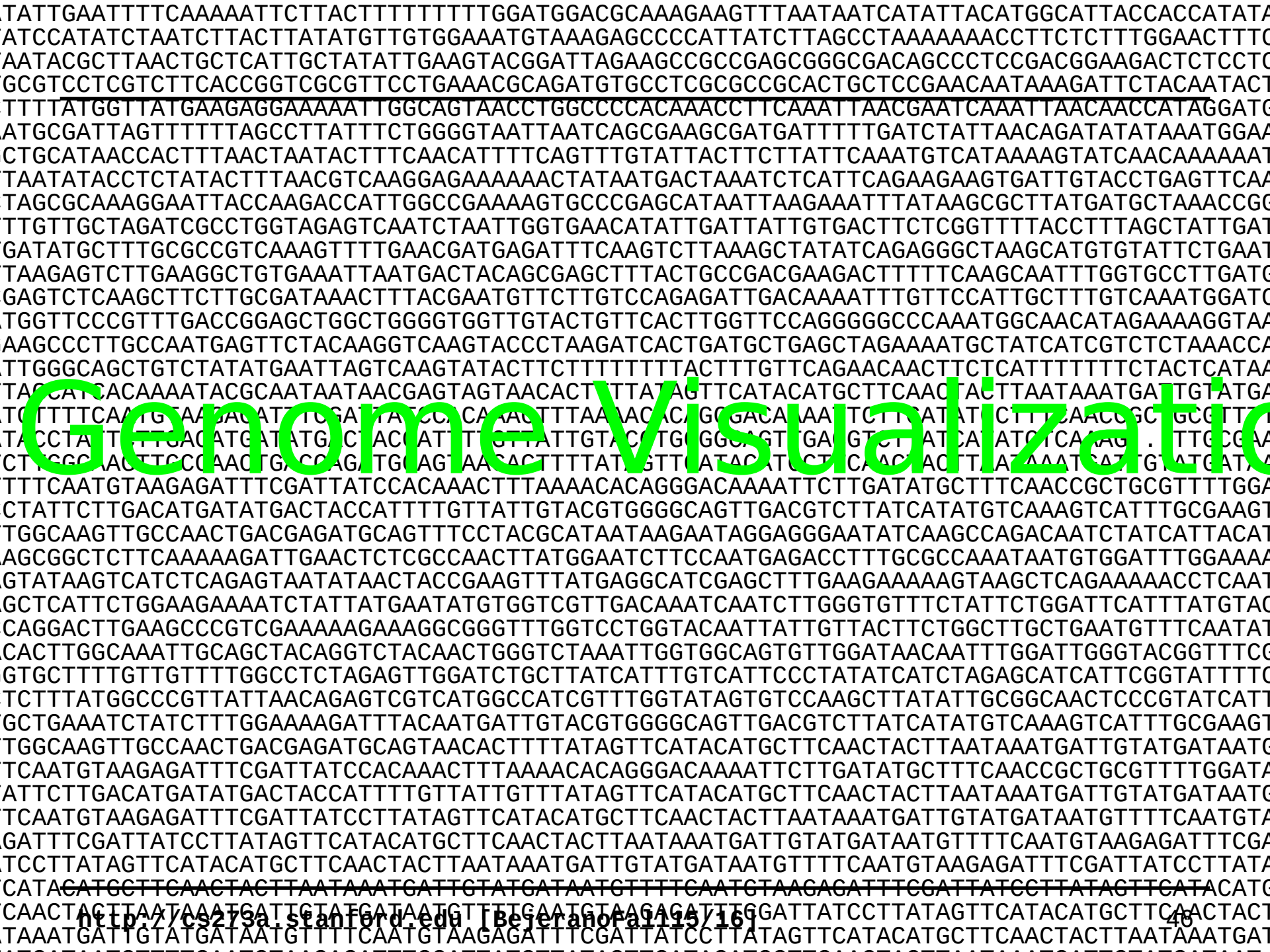
10^{13} cells
in a human

We could never sequence enough

The age of Omics:

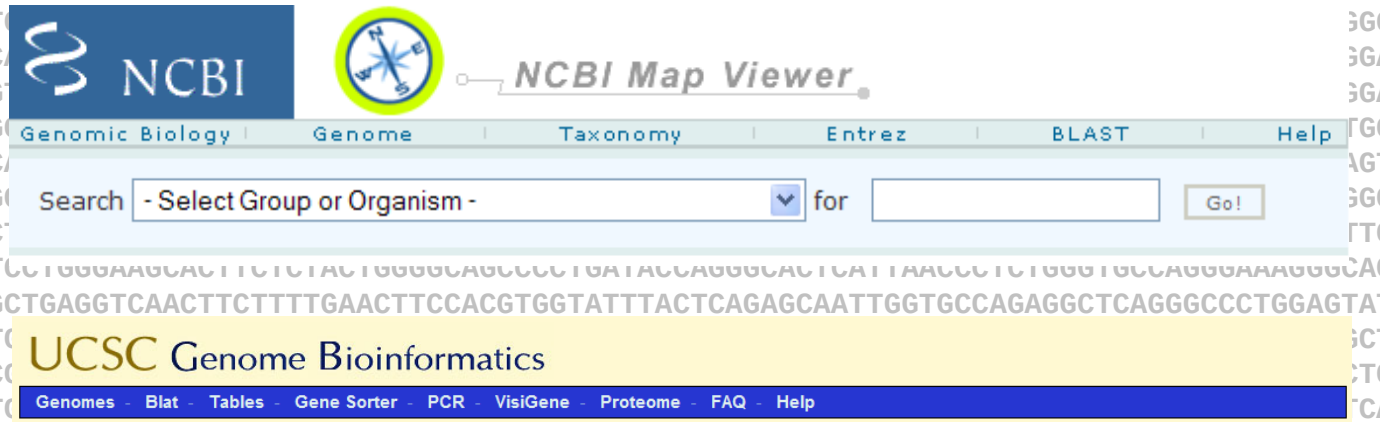
- "We are drowning in a sea of data,
  yet we are dying of thirst for knowledge.“

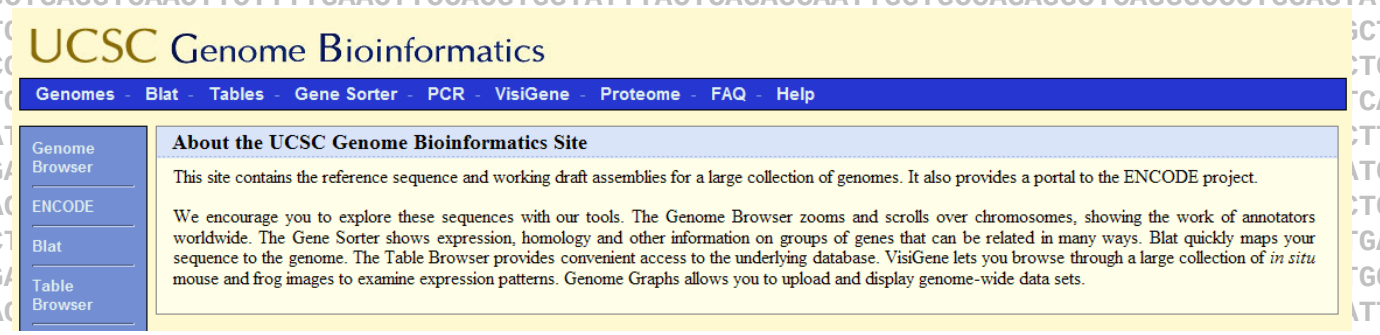
- "Data is not information,
  information is not knowledge,
  knowledge is not understanding,
  understanding is not wisdom."

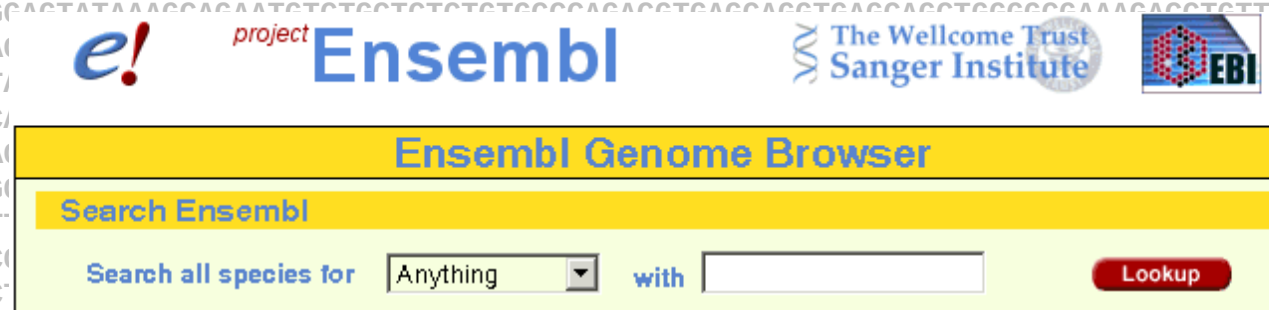

Genome Visualizatio

Portals to the Human Genome

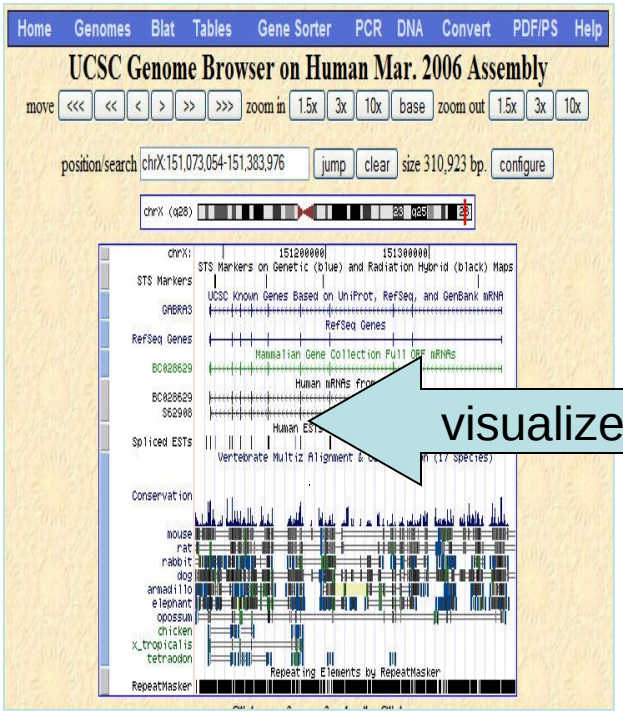
Human Genome = three billion ($3*10^9$) basepairs:







Genome Browser Database



visualize

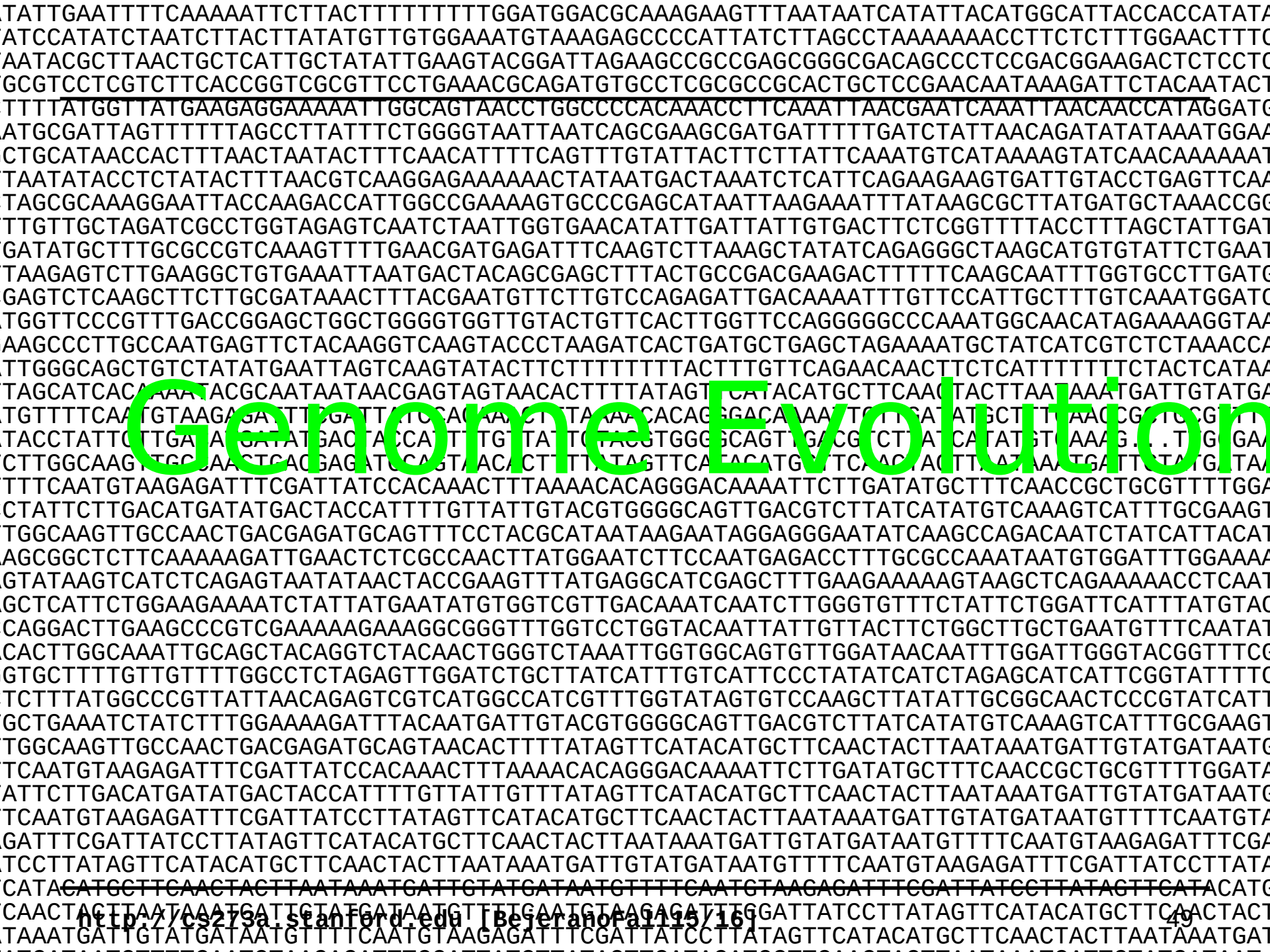
search & download

Underlying Database (MySQL)

#chrom	chromStart	chromEnd	name	class	func	
chr17	38451241	38451242	rs28897699	single	coding-nonsynon,untranslated,cd	
chr17	38451354	38451355	rs8176316	single	untranslated,intron	
chr17	38452072	38452073	rs8176315	single	untranslated,intron	
chr17	38452146	38452147	rs8176314	single	untranslated,intron	
chr17	38452200	38452201	rs9896925	single	untranslated,intron	
chr17	38452527	38452528	rs8176313	single	untranslated,intron	
chr17	38452648	38452649	rs8176312	single	untranslated,intron	
chr17	38452664	38452665	rs8176311	single	untranslated,intron	
chr17	38452703	38452704	rs8068463	single	untranslated,intron	
chr17	38453219	38453220	rs4438367	single	coding-synon,untranslated,cds-r	
chr17	38453227	38453228	rs28897698	single	coding-nonsynon,untranslated,cd	
chr17	38453438	38453439	rs8176310	single	untranslated,intron	
chr17	38453634	38453635	rs4793190	single	untranslated,intron	
chr17	38453634	38453635	rs7406700	single	untranslated,intron	
chr17	38453818	38453819	rs4792971	single	untranslated,intron	
chr17	38453895	38453896	rs8176308	single	untranslated,intron	
chr17	38454062	3845	rs4792972	single	untranslated,intron	
			rs7405512	single	untranslated,intron	
			176307	single	untranslated,intron	
			176306	single	untranslated,intron	
chr17	38454889	3845	rs8176305	single	untranslated,intron	
chr17	38454957	38454958	rs8176304	single	untranslated,intron	
chr17	38454990	38454991	rs8176303	single	untranslated,intron	
chr17	38455227	38455228	rs3092988	single	untranslated,intron	
chr17	38455362	38455363	rs8176302	single	untranslated,intron	
chr17	38455548	38455549	rs8176301	single	untranslated,intron	
chr17	38455820	38455821	rs8176300	single	untranslated,intron	
chr17	38455966	38455967	rs8176299	single	untranslated,intron	
chr17	38456104	38456105	rs12949768	single	untranslated,intron	
chr17	38456115	38456116	rs11079053	single	untranslated,intron	

Primary table: positions, names, etc.

Auxiliary table: related data


Genome Evolution

Genome Evolution

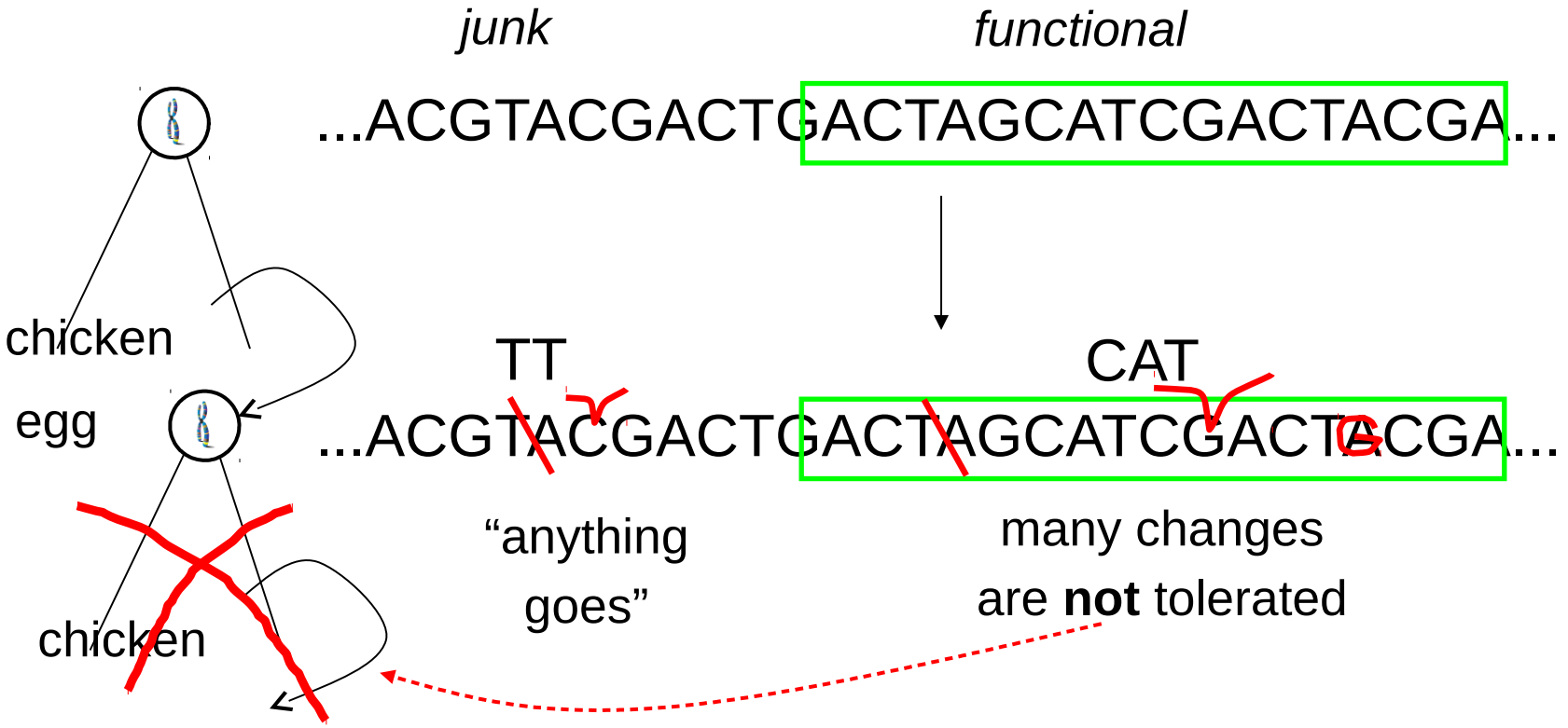
So far we've treated the genome as immutable.

But boy is it **alive**.

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

Every Genome is Different

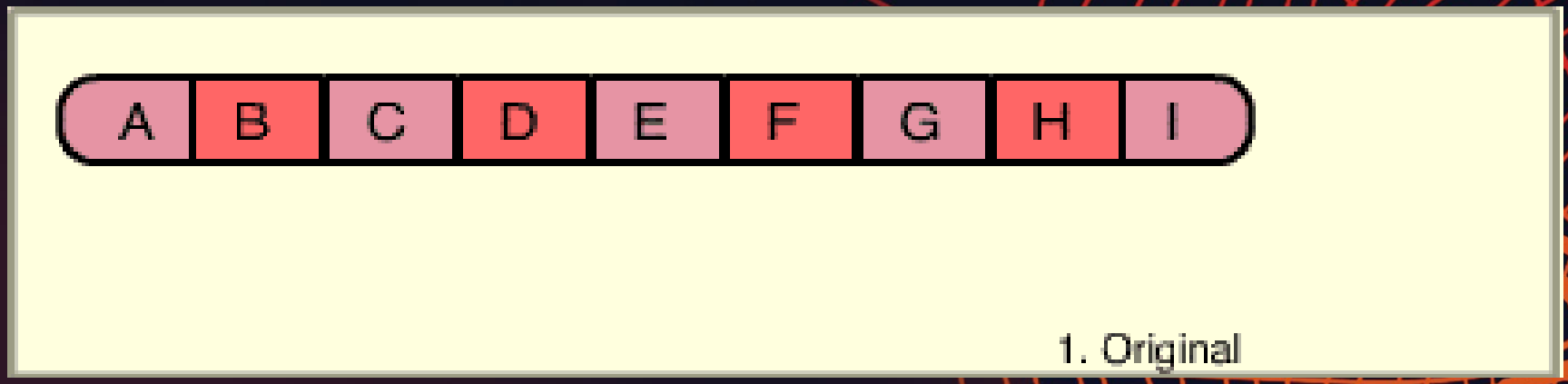
DNA Replication is imperfect – between individuals of the same species, even between the cells of an individual.



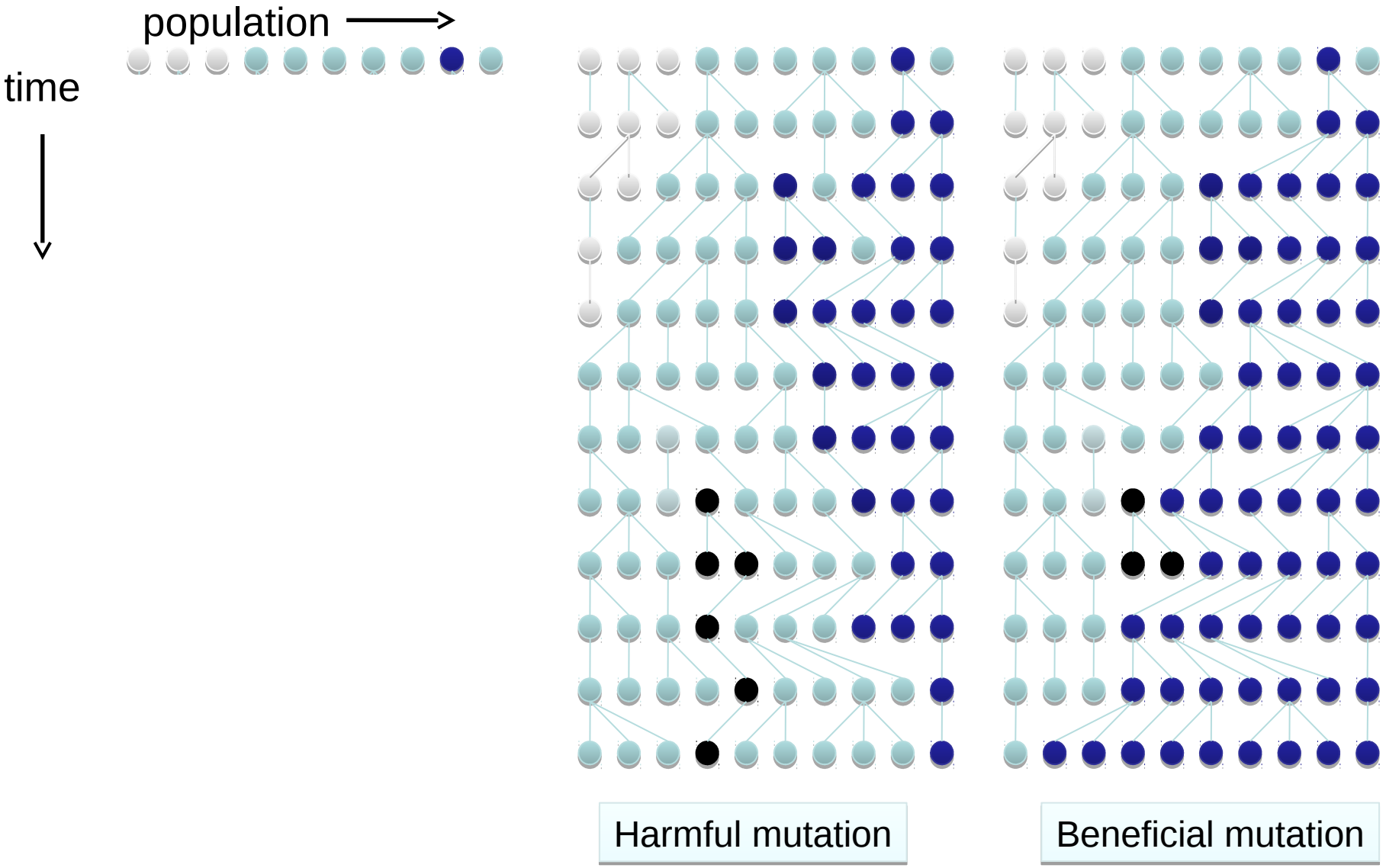
This has bad implications – disease, and good implications – evolution.

Genome mutation types: anything you can do to a string

- Deletion
- Inversion
- Translocation
- Duplication

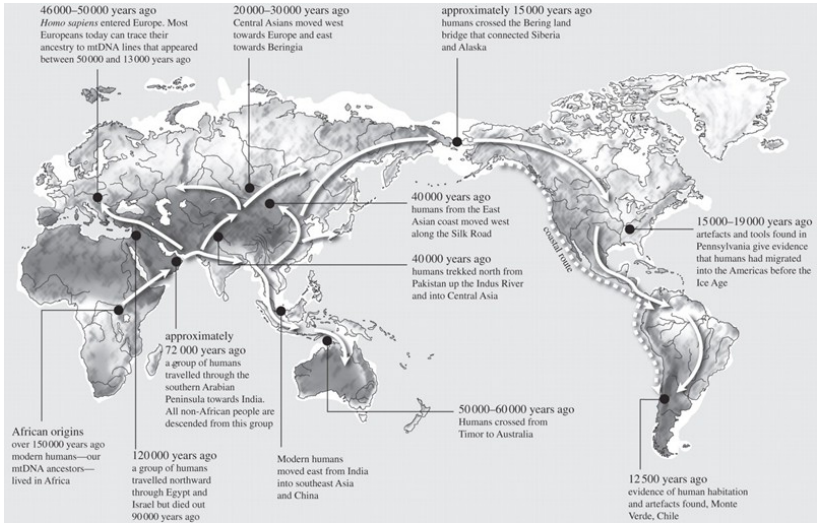


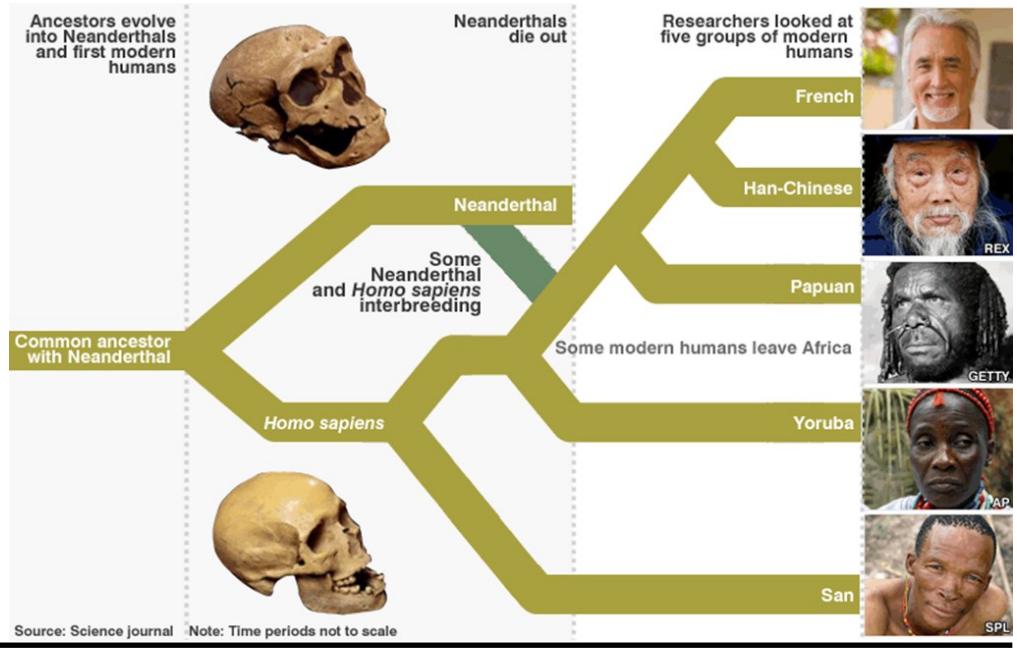
Modes of evolution = Mutation + Selection




PopulationGenomic

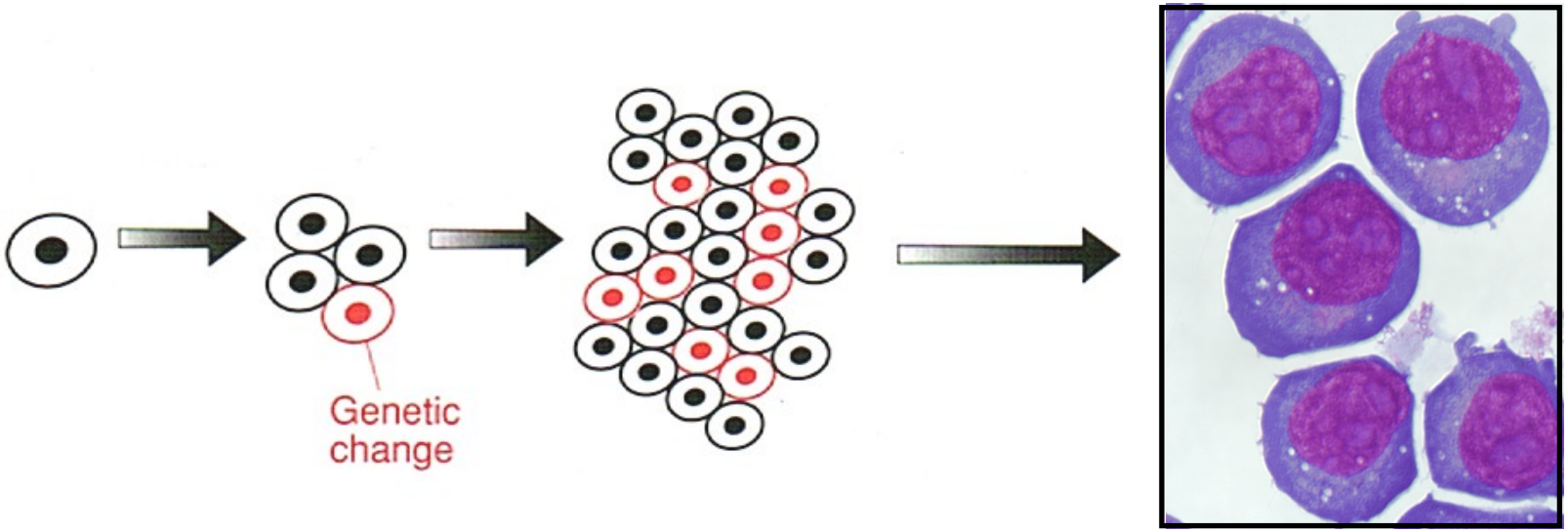
From neutral evolution alone





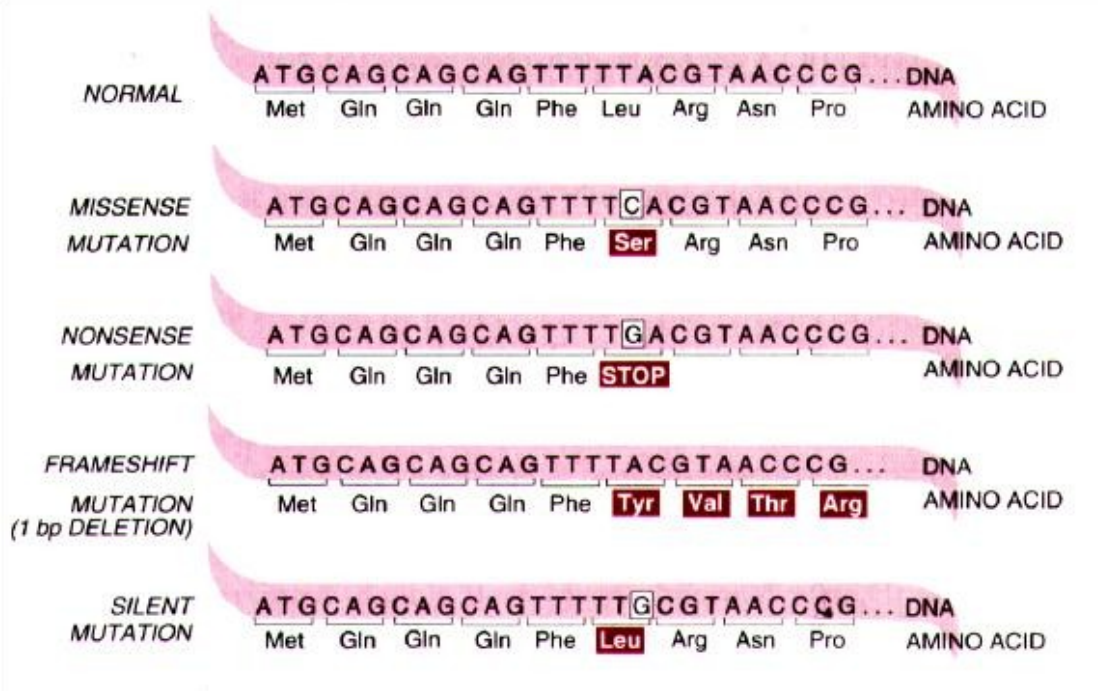

Genomics&Diseas

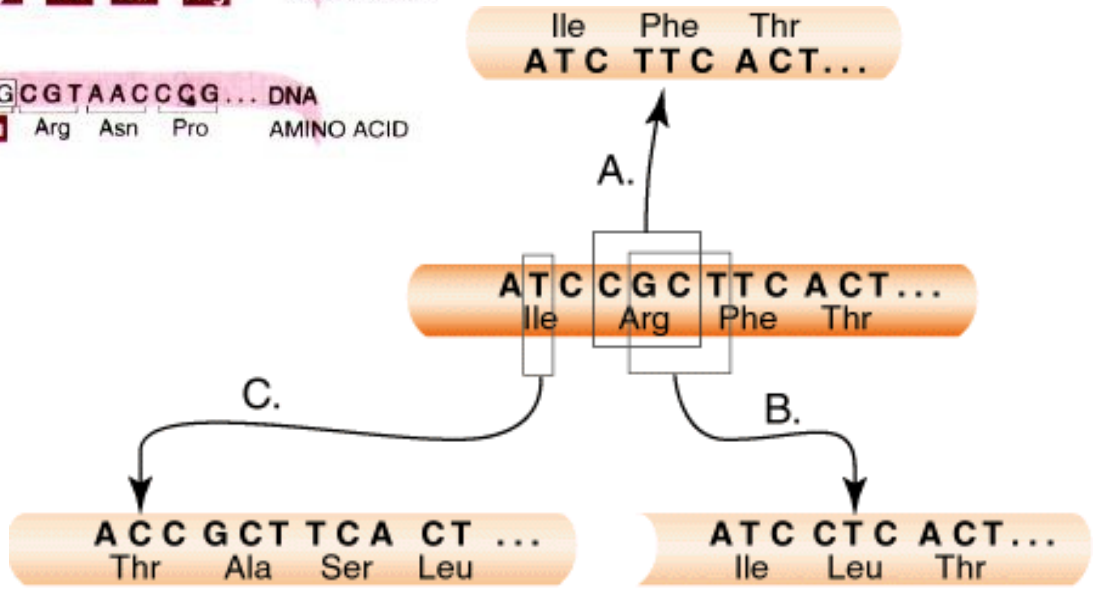
Cancer is a disease of the genome



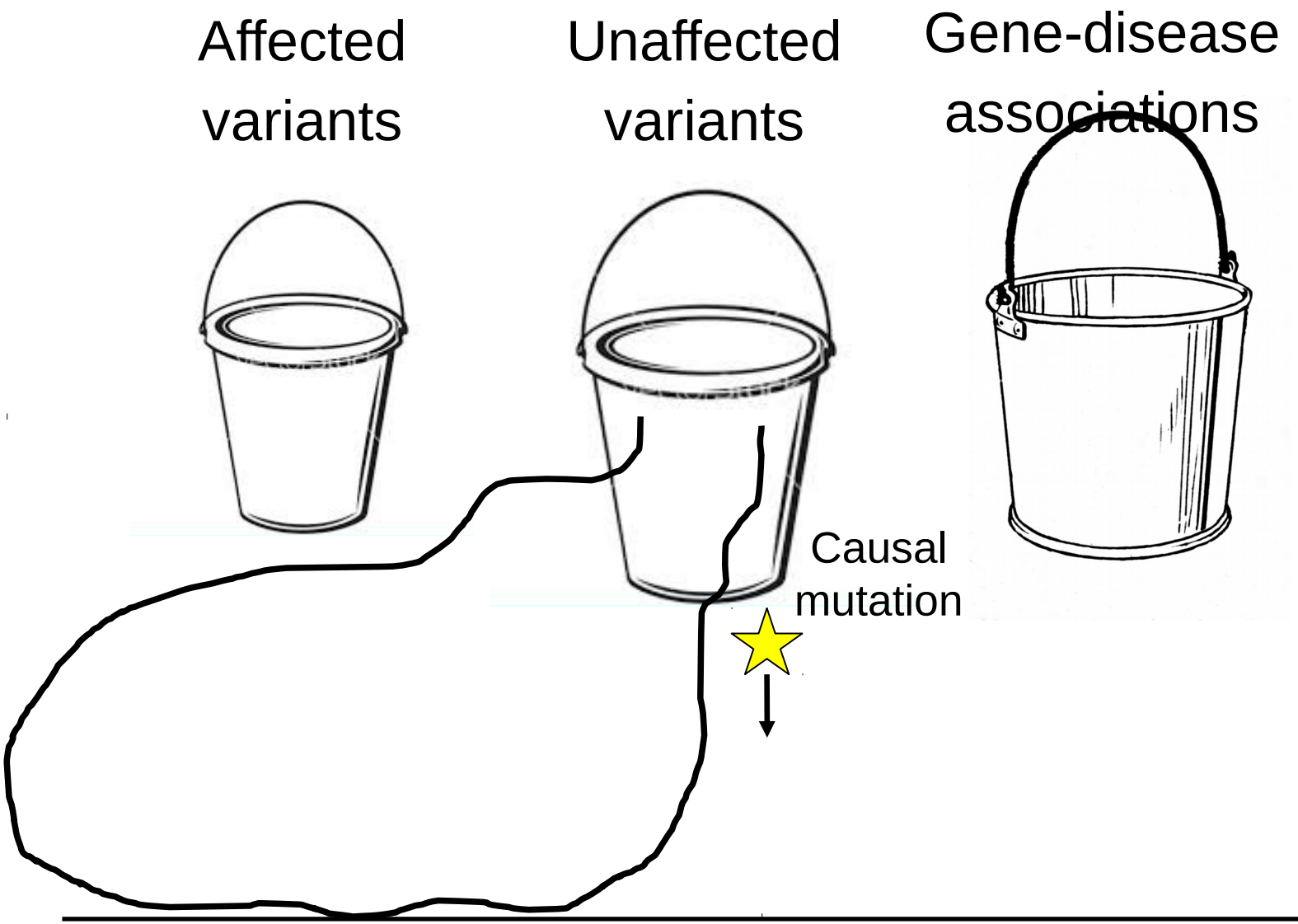
that makes a cell veer off plan
(whatever its particular plan is)
and start dividing uncontrollably
ultimately throwing the organism off balance

Single Base Changes Can Be Detrimental

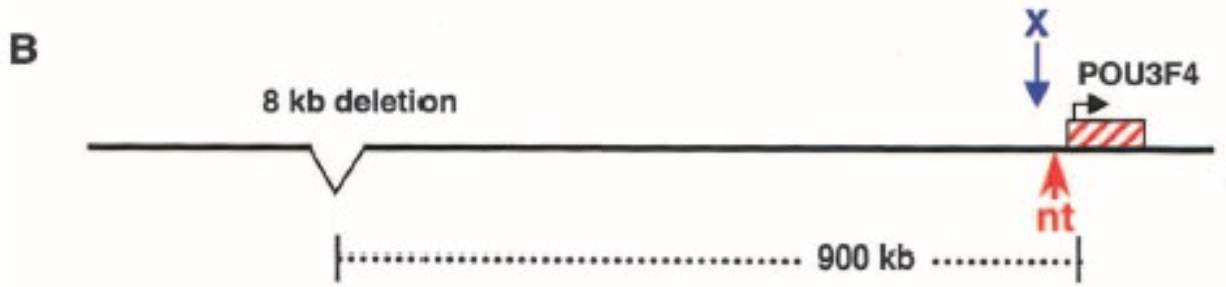




Mendelian Diseases



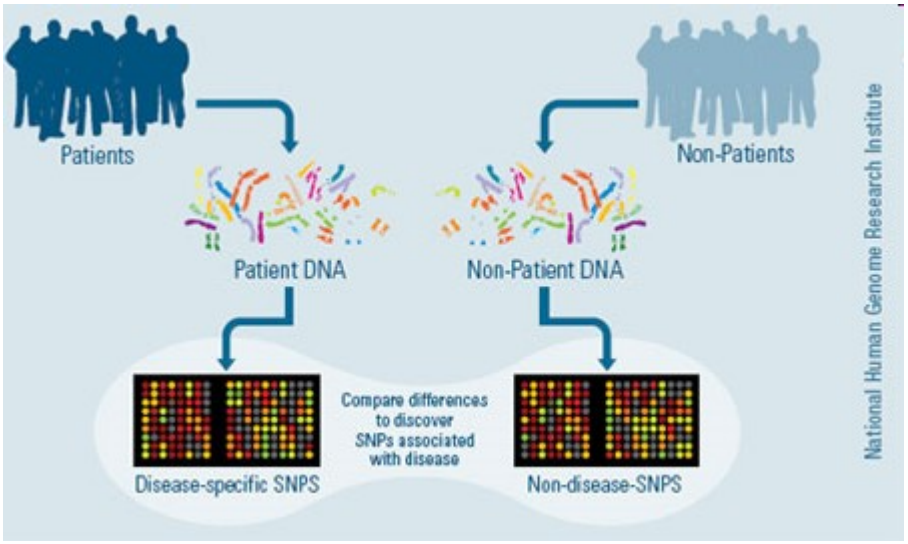
“Non-coding” mutations can be detrimental

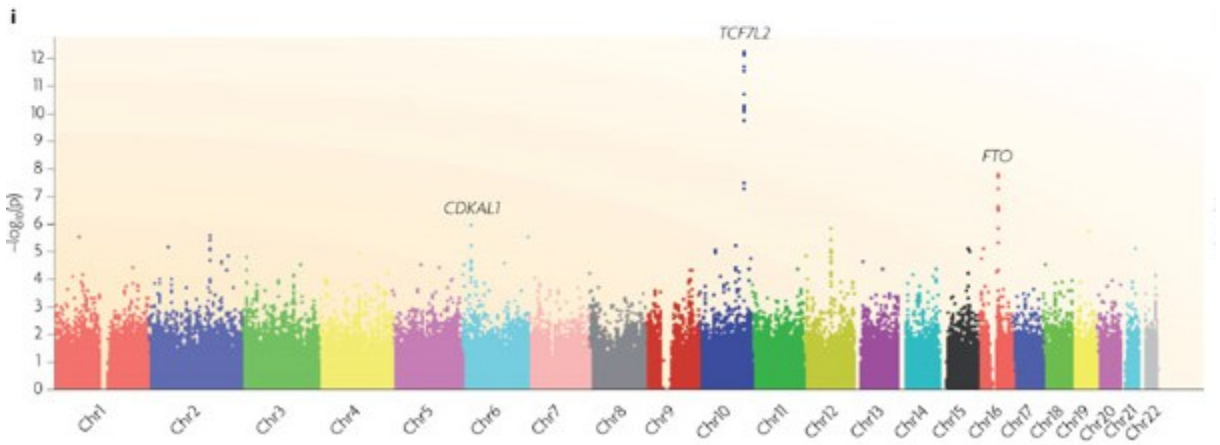


retinal expression. *B*, The human *POU3F4* deafness locus. The microdeletion of an 8-kb region located 900 kb upstream of the gene contains a conserved noncoding sequence, the loss of which leads to congenital deafness. The mouse slf inversion breakpoint X leaves the neural tube

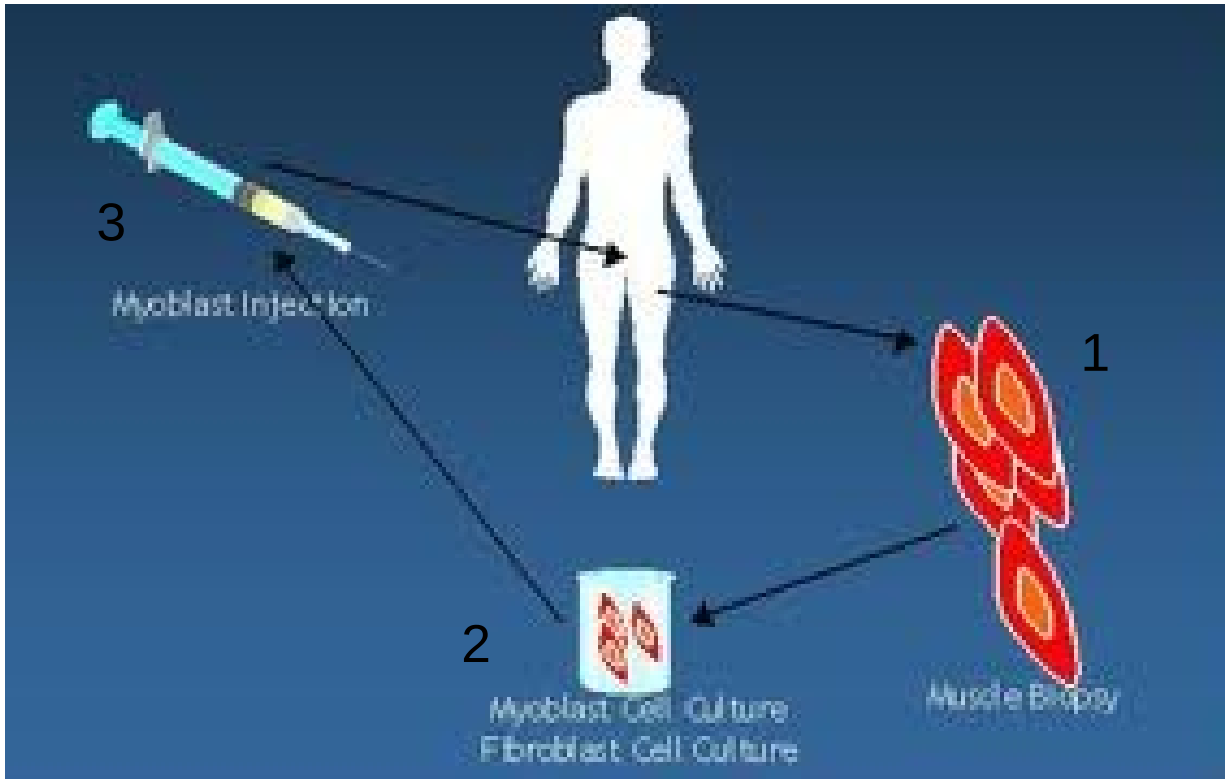
[de Kok et al, 1996]

Finding Disease Loci





Gene/Cell Therapy: Curing Genomics Defects



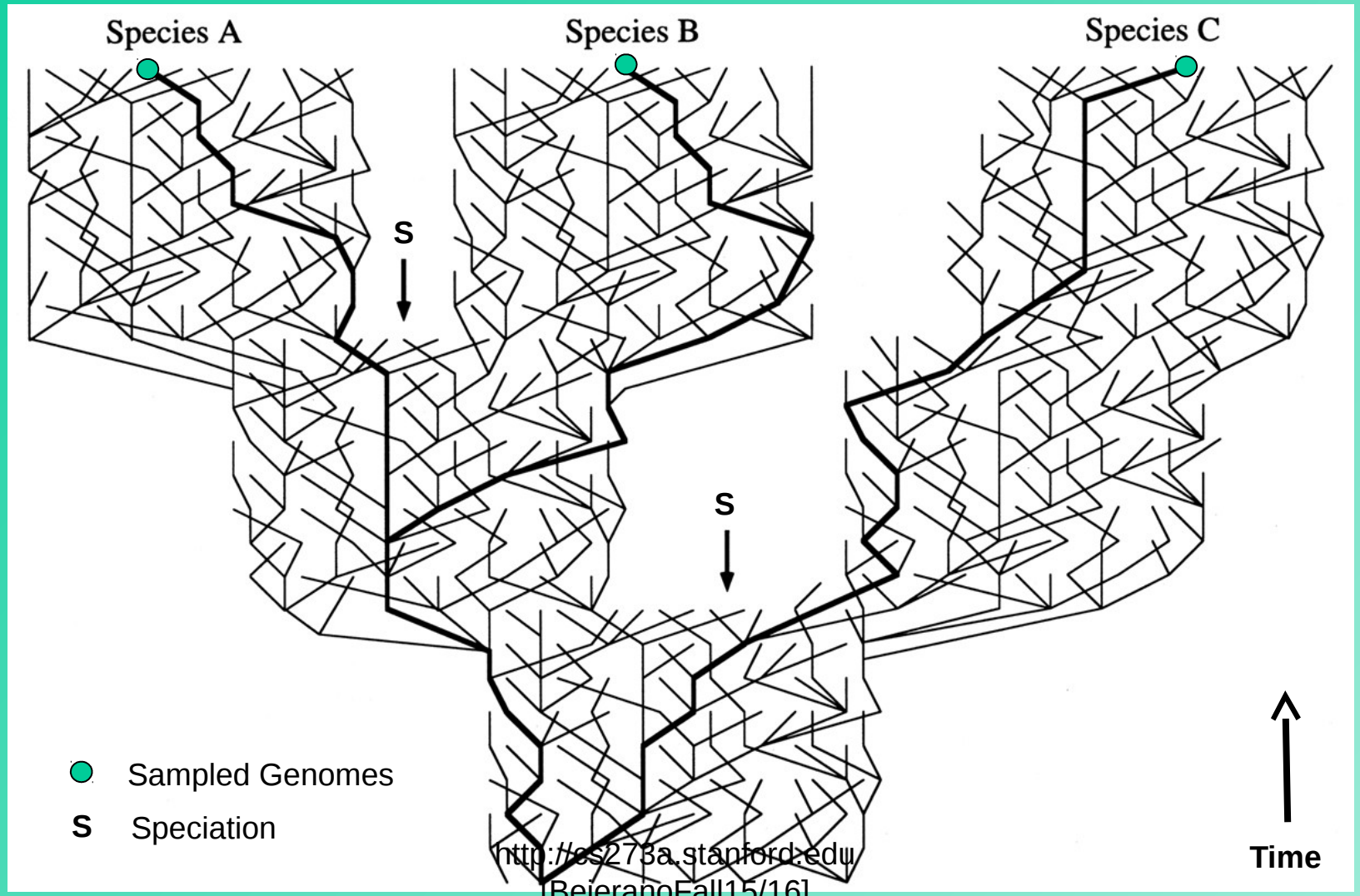
1 Get'em

2 Fix'em

3 Put'em back


Organism Evolution

Evolution is not all bad!



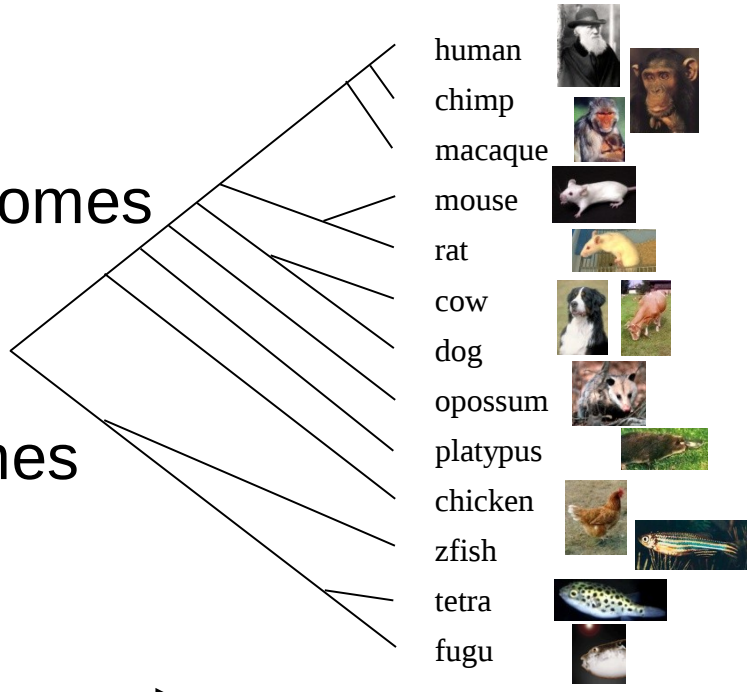
Comparative Genomics, Evo Devo

We can sample ancient genomes
(tens of thousands of years old).

We can reconstruct ancestral genomes
(tens of millions of years old).

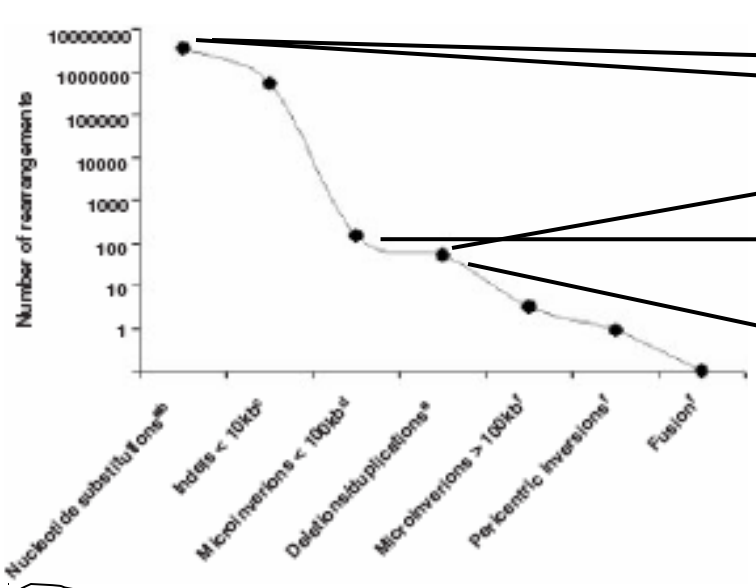
How to learn from different genomes
about their owners?

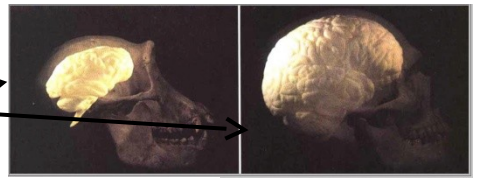
(we can get the tape,
we can play the tape,
we want to hear the music!)

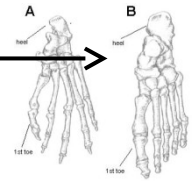


The great genotype-phenotype divide

and ways to cross it!







Genotype

Phenotype

Further Reading

These principles and tips will be revisited at course's end.

At which point we will ask ourselves:

Are we any wiser?

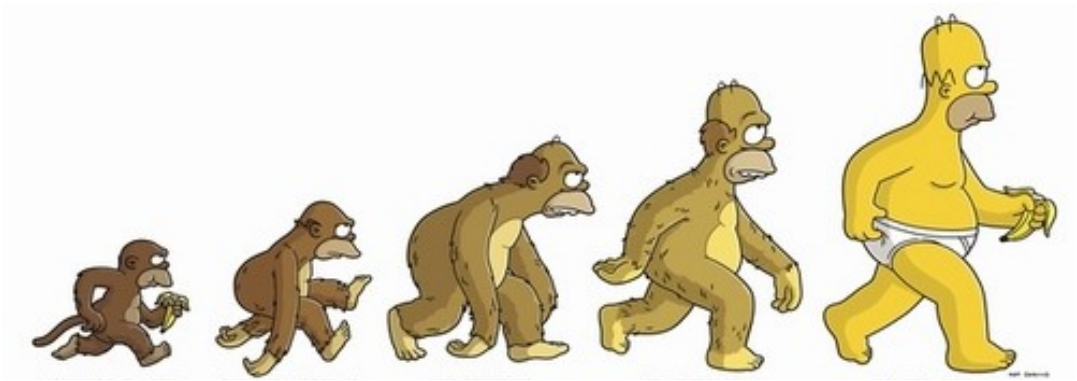
Check out the Bejerano lab "resources" page:

- Popular science books
- Core Stanford classes
- Core technical books/skills

What if I get hooked?

Classes, rotations, CURIS, honors theses, ..

Lots of genomic research on campus



To Be Continued…