
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

The screenshot shows the Stanford CS273a course page on Piazza. At the top, there's a banner with two people on bicycles and the text "CS273a: A Computational Tour of The Human Genome". Below the banner, there are three main sections: "Course Description", "Prerequisites", and "Cross-listings". The "Course Description" section includes a detailed description of the course content, mentioning topics like genome sequencing, functional landscape, gene regulation, RNA genes, epigenetics, genome evolution, comparative genomics, ultraconservation, and co-option. It also notes that additional topics may include population genetics, personalized genomics, and ancient DNA. The "Prerequisites" section states that there are no biological or computational prerequisites, although a background in programming is encouraged. The "Cross-listings" section lists DBIO273. To the right of these sections, there's a sidebar for "Professors & TAs: Save Time. Teach" which encourages users to use Piazza for managing course materials and tracking student participation. It also features a "See why Piazza works" link. At the bottom, there are buttons for "Student?" (Search your classes) and "Instructor?" (Create or Join your class).

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

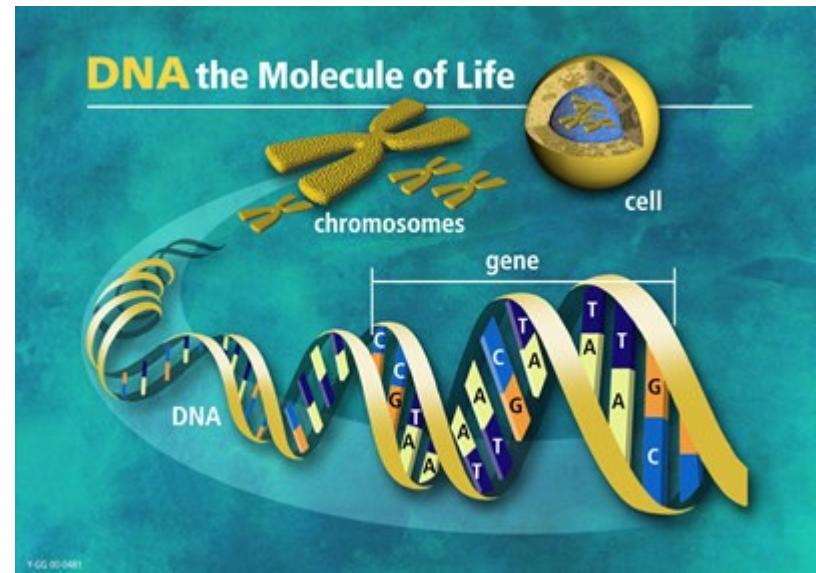
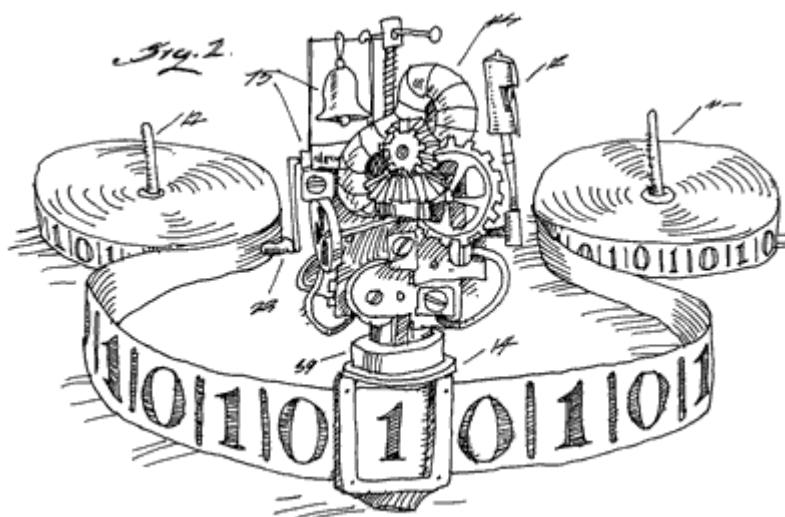
What will we study?

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

Organism – Cell - Genome

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

The cell is the basic unit of life.

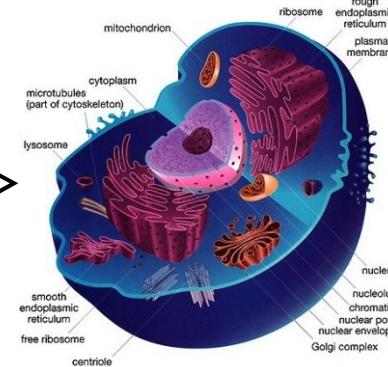
DNA = linear molecule inside the cell that carries instructions needed throughout the cell's life ~ long string(s) over a small alphabet

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

Genome:

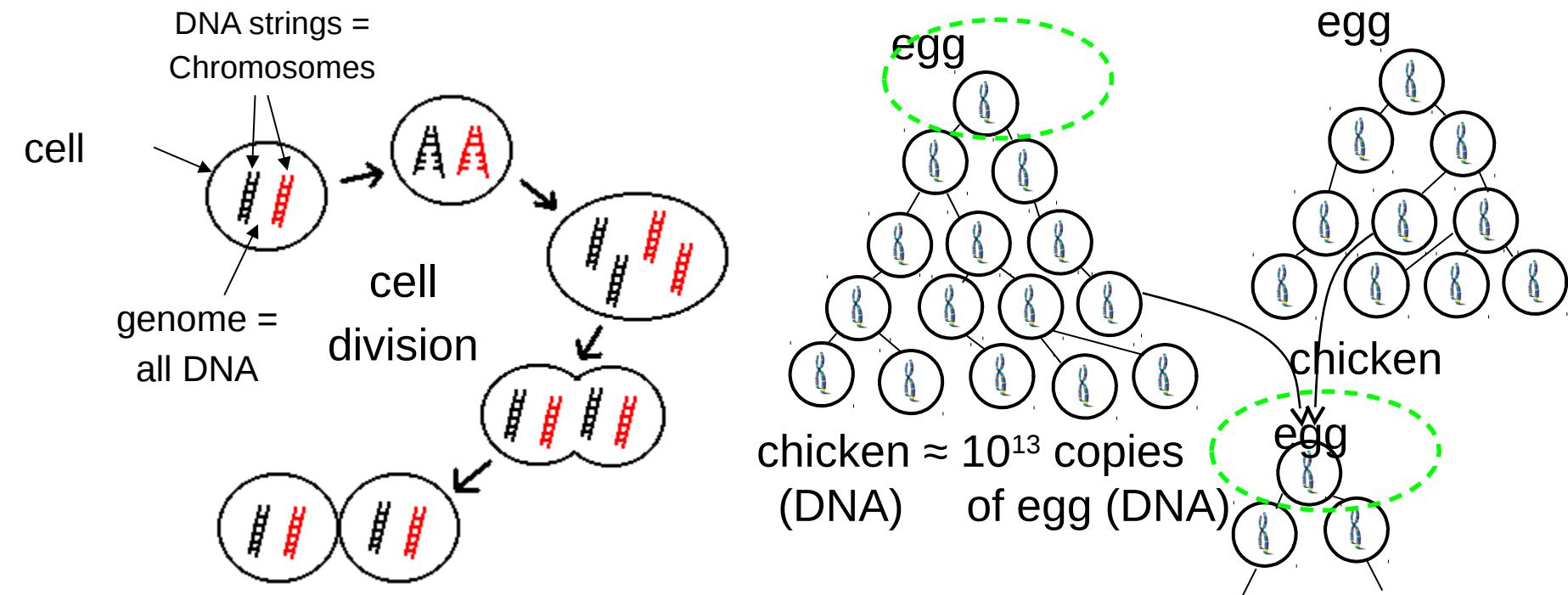
...ACGTACGACTGACTAGCATCGACTACGACTAGCAC......

 “instruction”



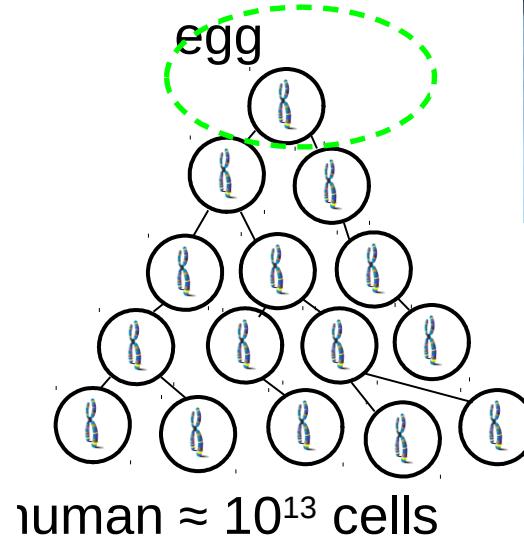
One Cell, One Genome, One Replication

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

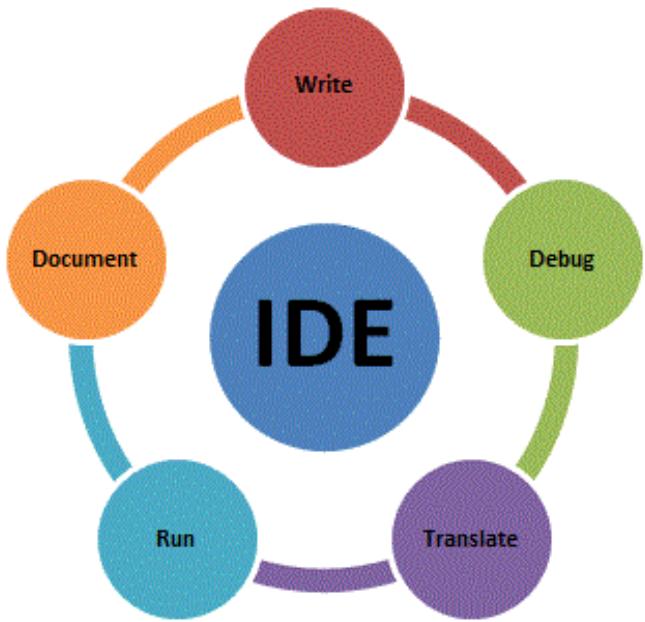


Talk about code reuse

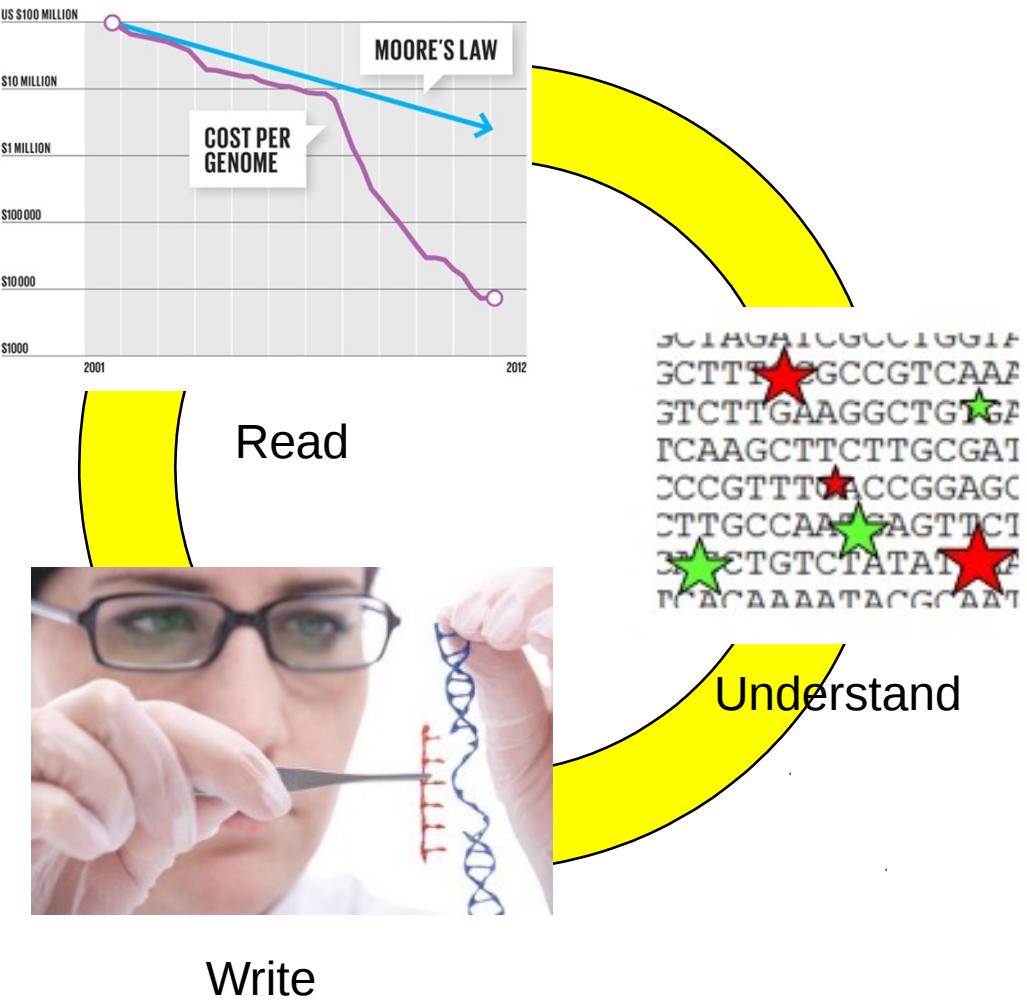
- The same genome “runs” hundreds, likely thousands of very different cell types.
- Nature vs. Nurture – If life is like a river, our genome provides the boat...



Integrated Development Environment



Bye bye natural selection
Bye bye human race ...

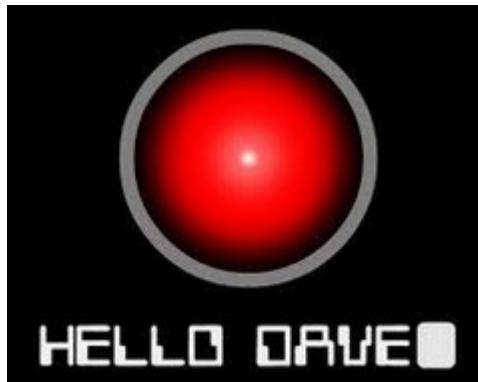
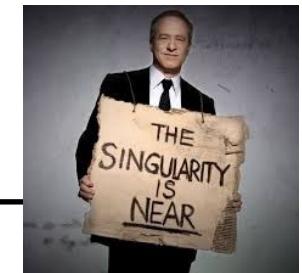


The Singularity

Singularity – (more) sentient non homo sapiens

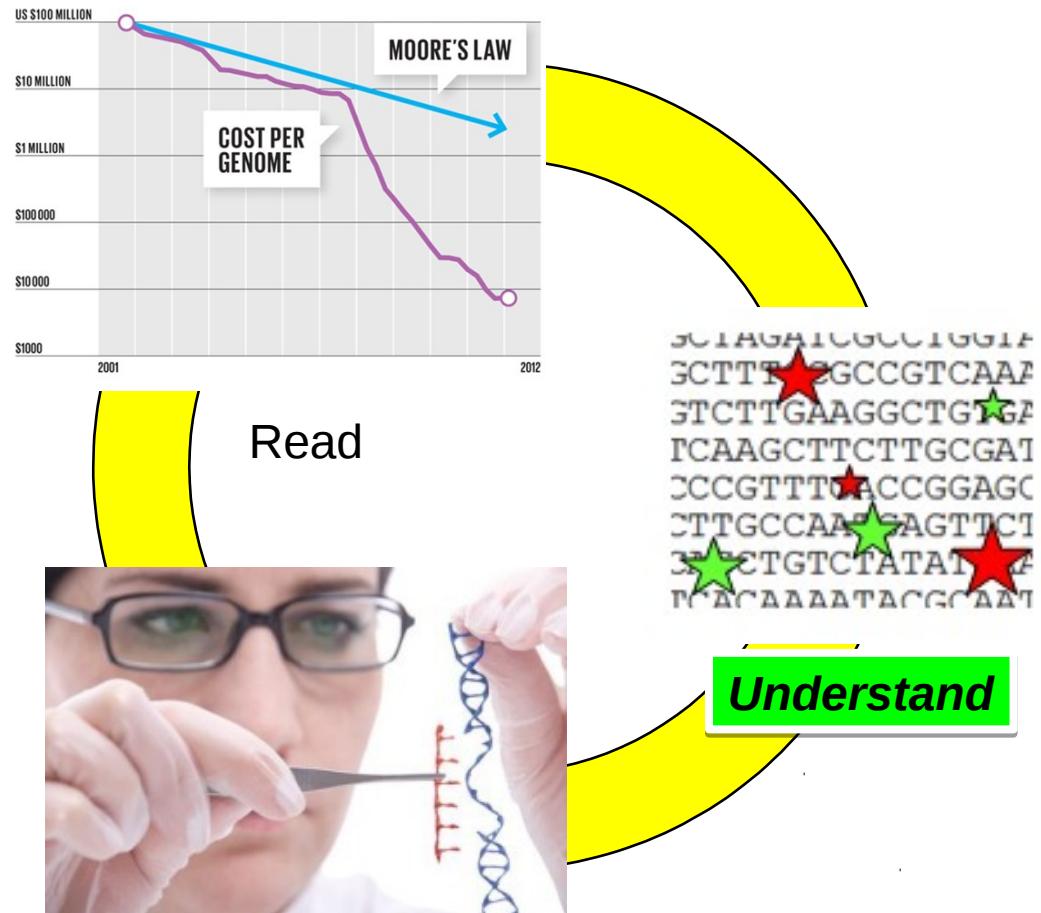
The Singularity can come from one of:

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



Write

“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

ATTTGAATTTCAAAAATTCTTAACCTTTGGATGGACGCAAAGAAGTTAACATATTACATGGCATTACCAACCATATA
ATCCATATCTAATCTTACTTATGTTGTGGAAATGTAAGAGGCCATTATCTTAGCCTAAAAAACCTCTCTGGAAACTTCA
AATACGCTTAAC TGCTCATTGCTATATTGAAGTACGGATTAGAAGCGCCGAGCAGCCCCTCCGACGGAAAGACTCTCCTC
GCCTCGTCTCACCGTCGTCCTGAAACGCAGATGTGCCTCGGCCGACTGCTCCGAACAATAAGATTCTACAATACT
TTTATGGTTATGAAGAGGAAAAATGGCAGTAACCTGGCCCCACAAACCTICAAATTAAACGAATCAAATTAAACACCATTAGGATG
ATGC GATTAGTTTTAGCCTTATTCTGGGTAA TTAAATCAGCGAAGCGATGATTTGATCTATTAAACAGATATAAAATGGAA
CTGCATAACCACCTTAACTAATACTTCACACATTTCAGTTGTATTACTTCTTATTCAAATGTCATAAAAGTATCAACAAAAAAT
TAATATACCTCTATACTTAACGTCAAGGGAGAAAAACTATAATGACTAAATCTCATTCAAAGAAGTGATTGTACCTGAGTTCAA
TAGCGCAAAGGAATTACCAAGACCATTGGCCGAAAAGTGCCCGAGCATAATTAAAGAAATTATAAGCGCTTATGATGCTAAACCGG
TTGTTGCTAGATCGCCTGGTAGAGTCAATCTAATTGGTGAACATATTGATTATTGACTTCTCGGTTTACCTTAGCTATTGAT
GATATGCTTGC GCCGTCAAAGTTGAACGATGAGATTCAAGTCTAAAGCTATATCAGAGGGCTAAGCATGTTGATTCTGAAT
TAAGAGTCTTGAAGGCTGTGAAATTAAATGACTACAGCGAGCTTACTGCCGACGAAGACTTTCAAGCAATTGGTGCCTGATG
GAGTCTCAAGCTTCTGCGATAAACTTACGAATGTTCTGAGAGATTGACAAAATTGTTCCATTGCTTGTCAAATGGATG
TGGTCCCCTGTTGACCGGAGCTGGCTGGGGTGGTTACTGTTCACTTGGTCCAGGGGCCCCAATGGCAACATAGAAAAGGTAA
AAGCCCTTGCCAATGAGTTCTACAAGGTCAAGTACCCCTAACATGACTGATGCTGAGCTAGAAAATGCTATCATGCTCTAAACCA
TTGGGCAGCTGTCTATATGAATTAGTCAAGTATACTTCTTTTACTTGTTCAGAACAACTTCTCATTTTTCTACTCATAA
TAGCATCACAAAATACGCAATAAACGAGTAGTAACACTTATAGTCATACATGCTCAACTACTAACATGATTGTATG
TGTGTTCAATGTAAGAGATTGATTCCACAAACTTAAAACACAGGGACAAAATTCTGATATGCTTCAACCGCTGCTT
TACCTATTCTGACATGATATGACTACCATTGTTA
CTTGGCAAGTTGCCAACTGACGAGATGCAGTAACACT
TTCAATGTAAGAGATTGATTCCACAAACTTAA
CTATTCTGACATGATATGACTACCATTGTTATTG
TGGCAAGTTGCCAACTGACGAGATGCAGTTCTACG
AGCGGCTCTCAAAAAGATTGAACCTCGCCAATTATGGAATCTTCCAATGAGACCTTGC
GTATAAGTCATCTCAGAGTAATATACTACCGAAGTTATGAGGCATCGAGCTTGAAGAAAAGTAAGCTCAGAAAACCTCAAT
GCTCATTCTGGAAGAAAATCTATTGAAATATGTGGTGTGACAAATCAATTCTGGGT
CAGGACTTGAAGGCCGTCAAAAAGAAAAGGGGTTGGCTGGTACAATTATTGTTACTTCTGGCTGCTGAATGTTCAATAT
CACTTGGCAAATTGCACTACAGGTCTACAACACTGGGTCTAAATTGGTGGCAGTGTGGATAACAAATTGGATTGGTACGGTT
GTGCTTTGTTGTTGGCCCTCTAGAGTTGGATCTGCTTACATTGTCATTCCCTATATCATCTAGAGCATCATTGGTATTT
TCTTATGGCCGTTATTAAACAGAGTCGTACATGGCCATCGTTGGT
GCTGAAATCTATCTTGGAAAAGATTACATGATTGACGTGGGGCAGTTGACGTCTTACATATGT
TGGCAAGTTGCCAACTGACGAGATGCAGTAACACTTATAGTCATACATGCT
TCATGTAAGAGATTGATTCCACAAACTTAAAACACAGGGACAAAATTCTG
ATTCTGACATGATATGACTACCATTGTTATTGTTATAGTCATACATGCT
TCATGTAAGAGATTGATTCCCTATAGTCATACATGCT
GATTTCGATTATCCTATAGTCATACATGCTCAACTACTTAATAAATGATTGT
TCCTTATAGTCATACATGCTCAACTACTTAATAAATGATTGT
CATACATGCTCAACTACTTAATAAATGATTGT
CAACTACTTAATAAATGATTGT
TAAATGATTGT

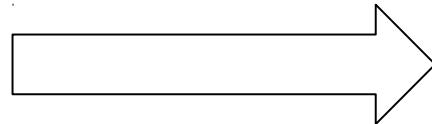
GGCAGTTGACGTCTTACATATGTCAAAG . . . TTGCGAA
CATACATGCTCAACTACTTAATAAATGATTGT
GACAAAATTCTGATATGCTTCAACCGCTGC
CAGTTGACGTCTTACATATGTCAAAGTC
ATAGGAGGGAAATATCAAGCCAGACAATCT
TACATTACAT
AGCTGAGCTTCTTCAAGTAAATGATTGT
TAC
AGCTCAGAAAACCTCAAT
GCTGTTCTATTCTGGATTCA
CAGGACTTGAAGGCCGTCAAAAAGAAAAGGGGTTGGCTGGTACA
CACTTGGCAAATTGCACTACAGGTCTACA
GTGCTTTGTTGTTGGCCCTCTAGAGTTGGATCTGCTTAC
TCTTATGGCCGTTATTAAACAGAGTCGTAC
GCTGAAATCTATCTTGGAAAAGATTACATGATTGAC
TGGCAAGTTGCCAACTGACGAGATGCAGTAACACTTATAGTCATACATGCT
TCATGTAAGAGATTGATTCCACAAACTTAAAACACAGGGACAAAATTCTG
ATTCTGACATGATATGACTACCATTGTTATTGTTATAGTCATACATGCT
TCATGTAAGAGATTGATTCCCTATAGTCATACATGCT
GATTTCGATTATCCTATAGTCATACATGCTCAACTACTTAATAAATGATTGT
TCCTTATAGTCATACATGCTCAACTACTTAATAAATGATTGT
CATACATGCTCAACTACTTAATAAATGATTGT
CAACTACTTAATAAATGATTGT
TAAATGATTGT
<http://cs276.stanford.edu> | Bejerano Fall 15 | 169

The Biggest Challenge in Genomics...

... is computational:

How does this

```
ATAGATGGCTGGTA  
GCTTGCGCCGTCAA  
GTCTTGAAGGCTGTGA  
TCAAGCTTCTTGCAT  
CCCCTTGACCGGAGC  
CTTGCCAATGAGTTCT  
CAGCTGTCTATATGAA  
TCACAAAAATAACGCAAT
```



encode *this*



Program

Output

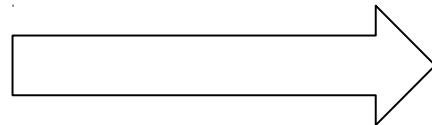
This “coding” question has profound implications for our lives

The Biggest Challenge in Genomics...

... is computational:

How does this

```
GCAGATGCGCTGGT  
GCTTGCGCGTCAA  
GTCTTGAAGGGCTGTGA  
TCAAGCTTCTTGC  
CCCCTTACCGGAGC  
CTTGCCAATGAGTTCT  
CAGCTGTCTATA  
TCACAAAAATACGGAAAT
```



encode *this*



Program

Bugs

Output

What genomic mutations predispose us to disease?

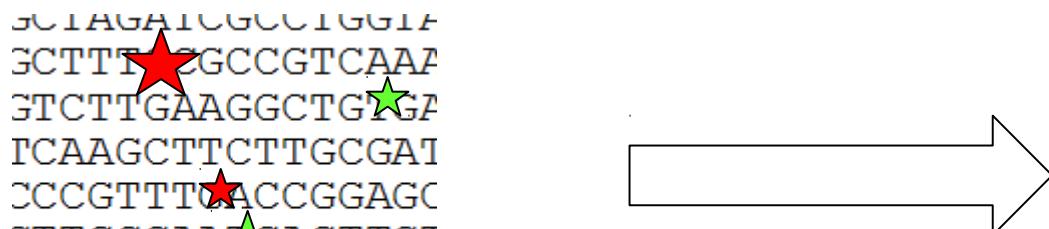
The Biggest Challenge in Genomics...

... is computational:

How does this

encode *this*

ACGAGATGGCTTGGTG
GCTTCGCCCGTCAAAT
GTCTTGAAGGGCTGGA
TCAAGCTTCTTGCAGA
CCCGTTTACCGGGAGC
CTTGCCTAACTAGTTCT
CTCTGTCTATAATTA
TCACAAAAATACGGAAAT



Program

Bugs

Patching

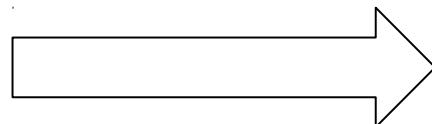
What genomic mutations determine our drug response?

The Biggest Challenge in Genomics...

... is computational:

How does this

ACATAGAATGCCCTGGT
GCTTCGCCCGTCAA
GTCTTGAAAGGCTGGA
TCAAGCTTCTTGC
CCCGTTTACCGGAGC
CTTGC
CTGTCTATA
TCAACAAAAATACGGAAAT



encode *this*



Program

Bugs

Debugging

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

The Biggest Challenge in Genomics...

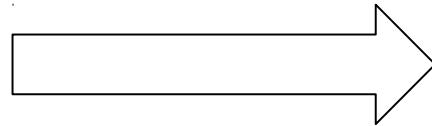
... is computational:

How does this

encode *this*

5' T A G A T C G G C T T G G T A
S C T T T G C G C C G T C A A P
G T C T T G A A G C * G T G P
T C A A G C T T C T T G C G A I
C C C G T T G A C C G G A G C
C T T G * S A A T G A G T T C I
C A G C T G T C T A T A T G A P
T C A C A A A A T A C G C A A T

5' T A G A T C G G C T T G G T A
S C T T T G C G C C G T C A A P
G T C T T G A A G C * G T G P
T C A A G C T T C T T G C G A I
C C C G T T G A C C G G A G C
C T T G C C A A T G A C * T C I
C A G C T G T C T A T A T G A P
T C A C A A A A T A C G C A A T



Program

Output

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

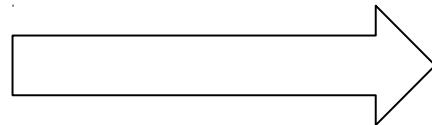
The Biggest Challenge in Genomics...

... is computational:

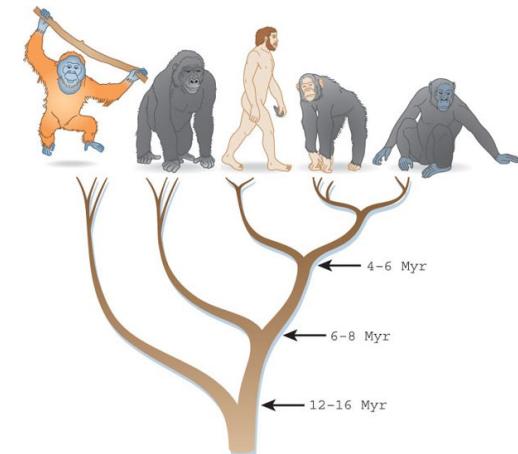
How does this

```
GCATGAACTGGCTGGT  
GCTTTGCGCCGTCAAF  
GTCCTGAAGGCTGTG  
TCAAGCTTCTTGC  
CCCGTTGACCGGAGC  
CTTGCCAATGAGTT  
CAGCTGTCTATATGAF  
TCACAAAAATAACGCAAT
```

```
GCATGAACTGGCTGGT  
GCTTTGCGCCGTCAAF  
GTCCTGAAGGCTGTG  
TCAAGCTTCTTGC  
CCCGTTGACCGGAGC  
CTTGCCAATGAGTT  
CAGCTGTCTATATGAF  
TCACAAAAATAACGCAAT
```



encode *this*



Program

Output

What in our genomes make us different from other species?

The Biggest Challenge in Genomics...

... is computational:

How does this

```
ATAGATGGCTGGT  
TTTGCAGCGTCAA  
GTCTTGAAGGCTGAGA  
TCAATCTTCCTGCGAT  
CCCGTTGACCGGGAGC  
CTTGCCAATGAGTTCT  
TAGCTATCTATATAA  
TACAAAAATCGGAAAT
```



encode *this*



Program

Output

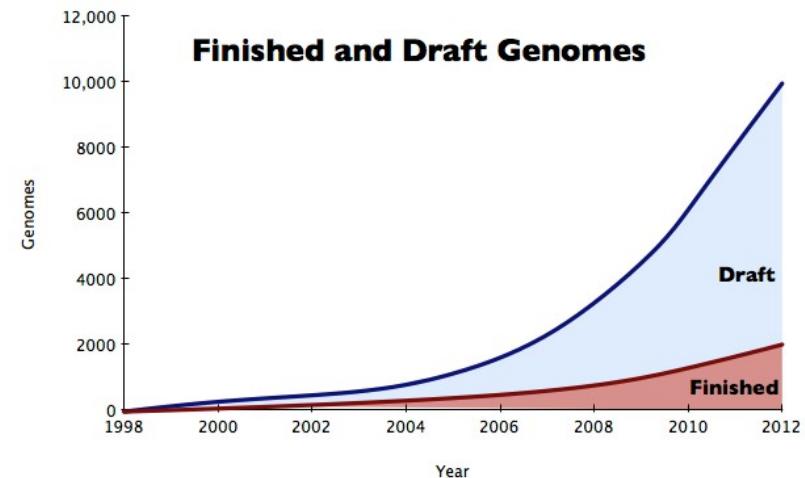
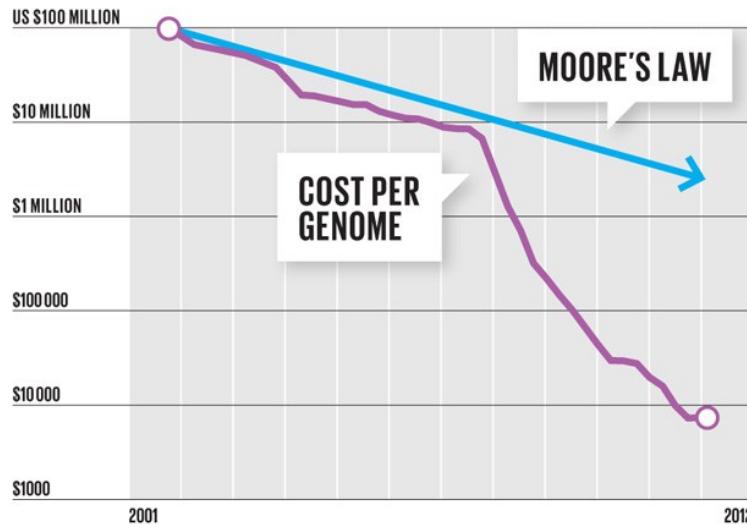
Why is our genome full of “memory leaks” and cruft?

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

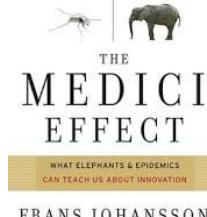


Storage
Compression
Architecture
Databases
HCI
etc.

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.

"IF YOU CAN'T BE SMART AND COME UP WITH AT LEAST A MEDIUM MOA LISA OR ELSE YOU'LL BE NOT FRIEND."
—Elmendorf, Bejerano



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



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”)

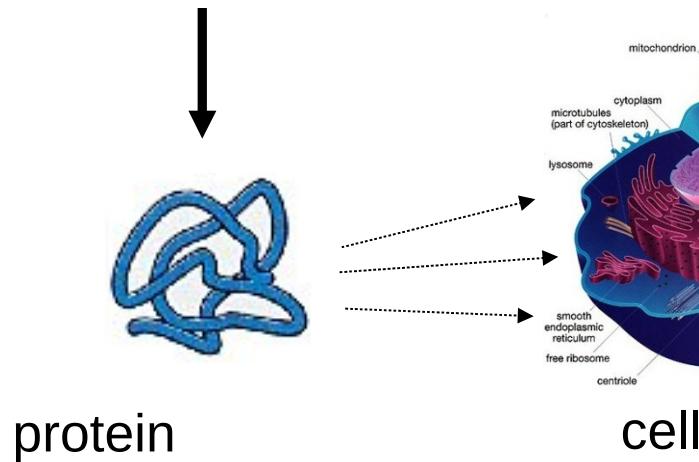
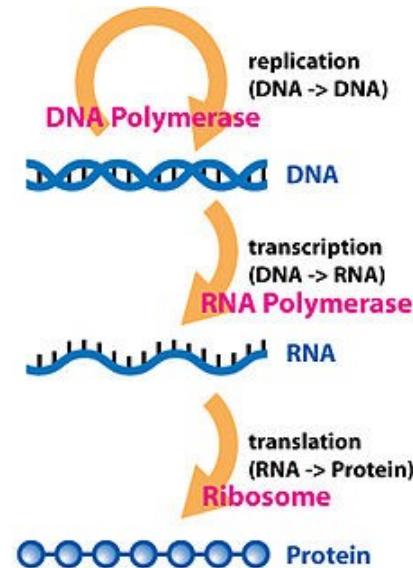
Genome content

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:

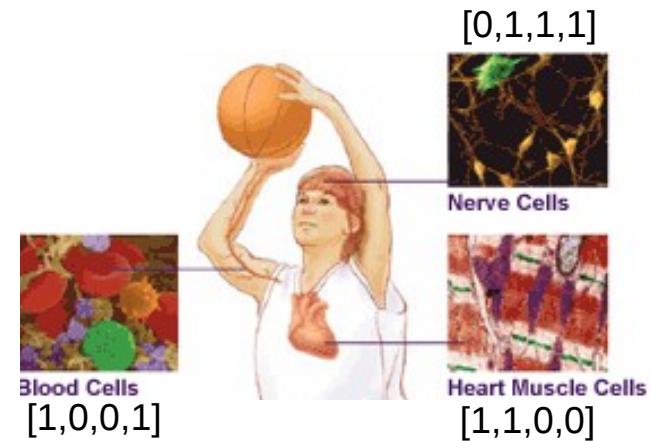
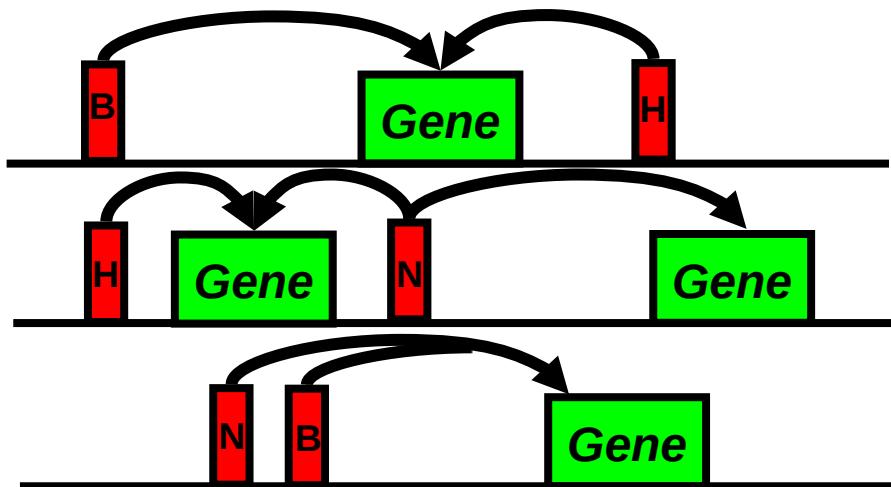
...ACGTACGACTGACTAGCATCGACTACGACTAGCAC...



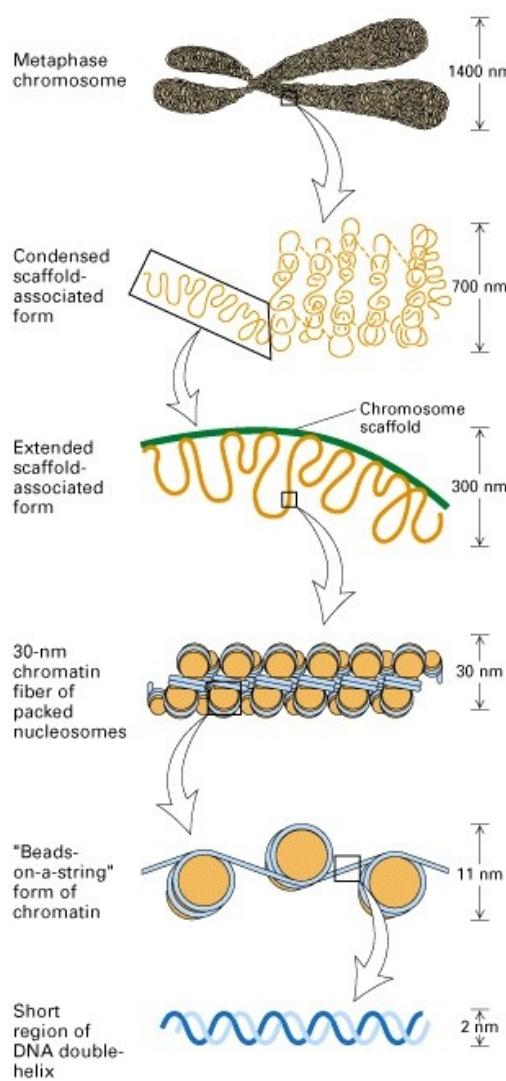
gene

Genes & Gene Regulation

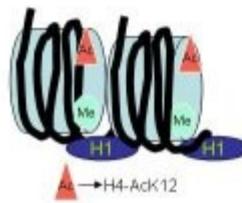
- Human genome encodes 20-25,000 genes (2% genome),
 >1,000,000 genomic switches that control genes (>10%).
- Gene = genomic substring that encodes
 HOW to make a protein.
- Genomic switch = genomic substring that encodes
 WHEN, WHERE & HOW MUCH of a protein to make.



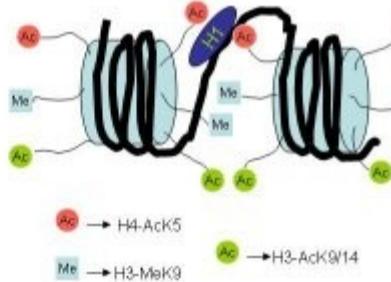
Epigenomics: transient writing “on” the genome



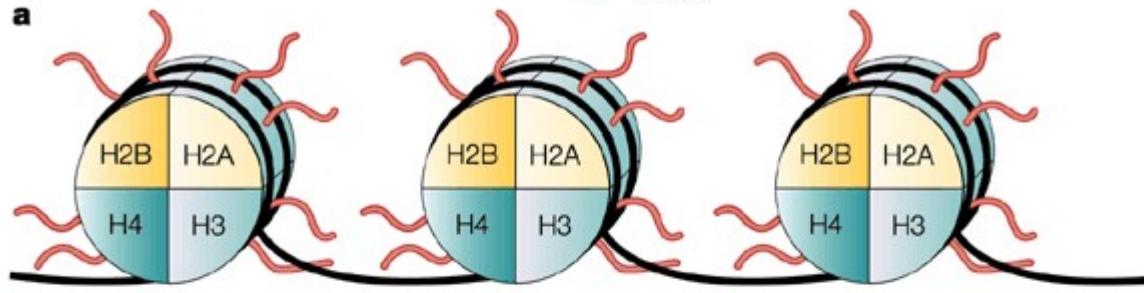
Closed/Inactive Chromatin



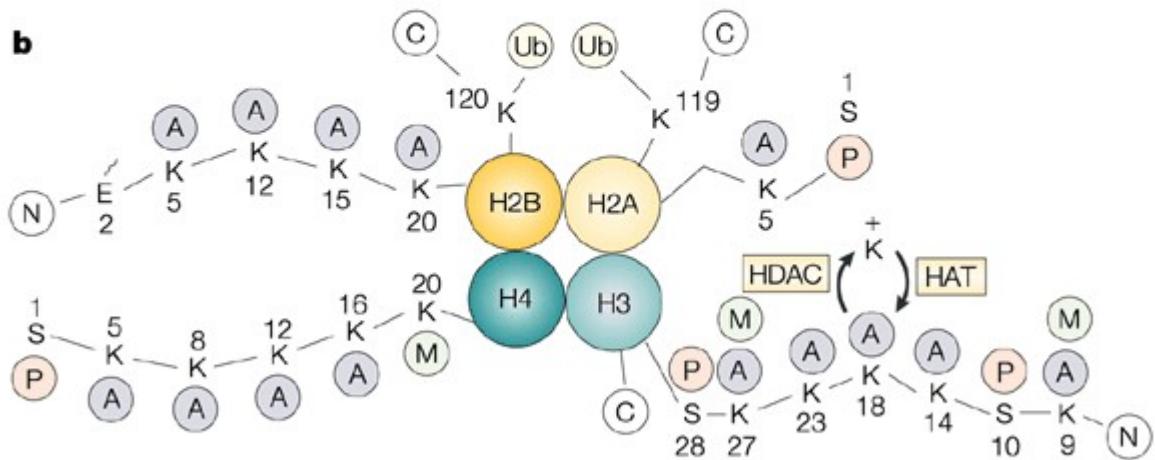
Open/Active Chromatin



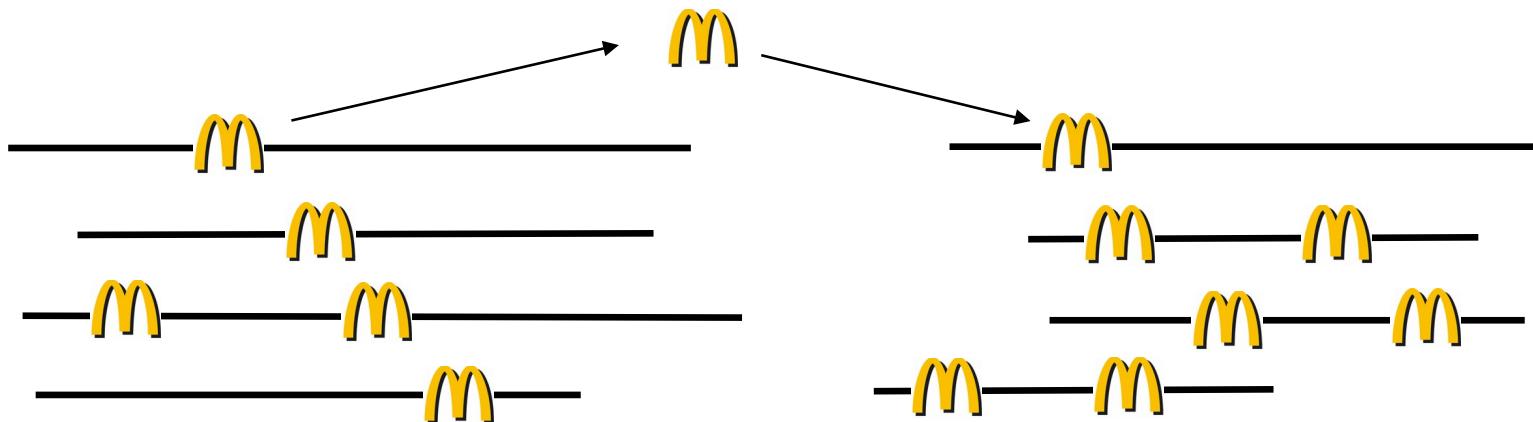
a



b



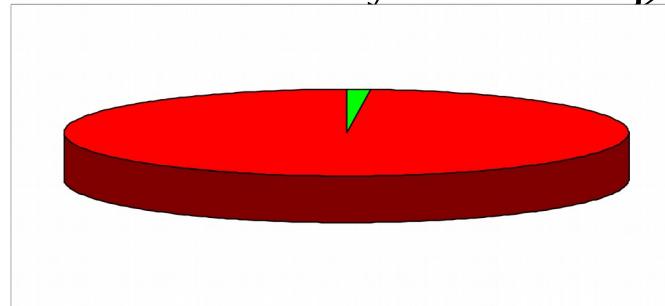
Repeats / Mobile Elements ("selfish/junk DNA")



Human
Genome:
 3×10^9 letters

2%
known
function

>50% junk



Genome: conceptual part list

Genome:

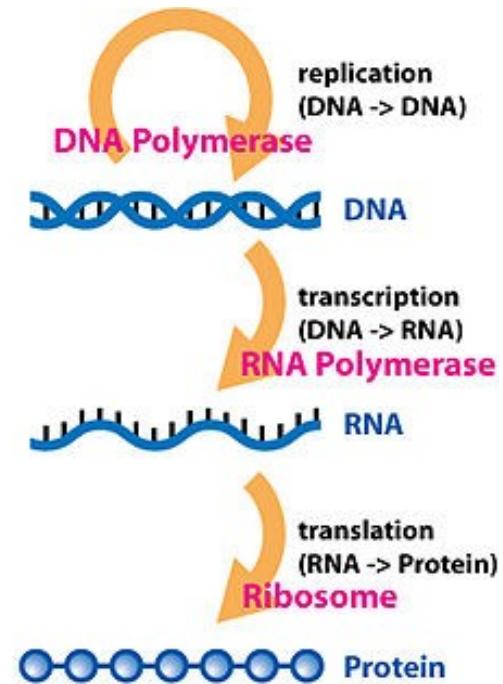
...ACGTACGACTG**ACTAGCATCGACTACGACTAGCAC...**

"instruction"

A copy (actually two) in every cell.

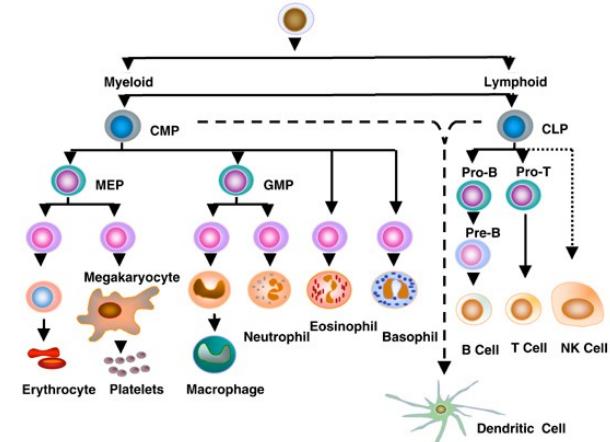
Contains:

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



Bottom line

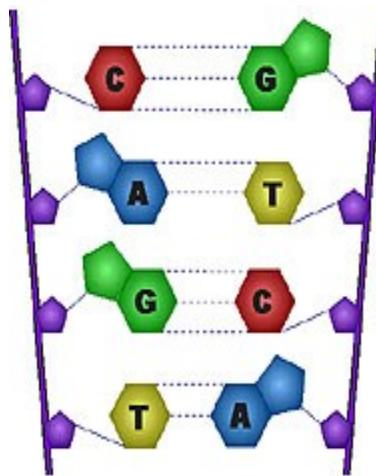
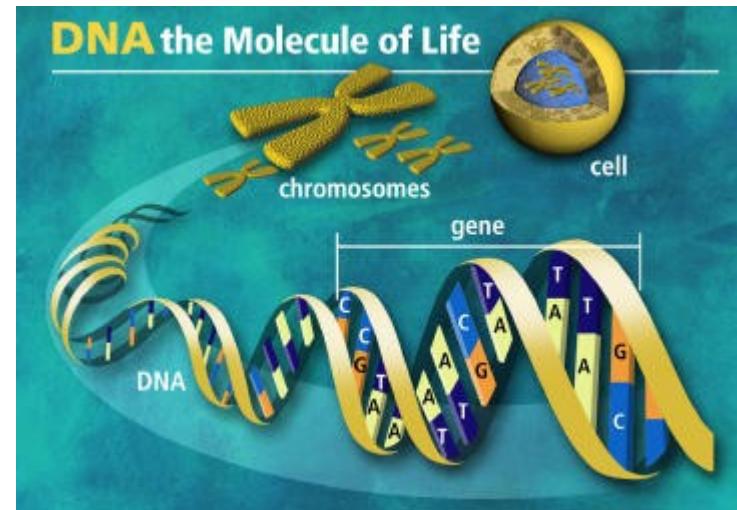
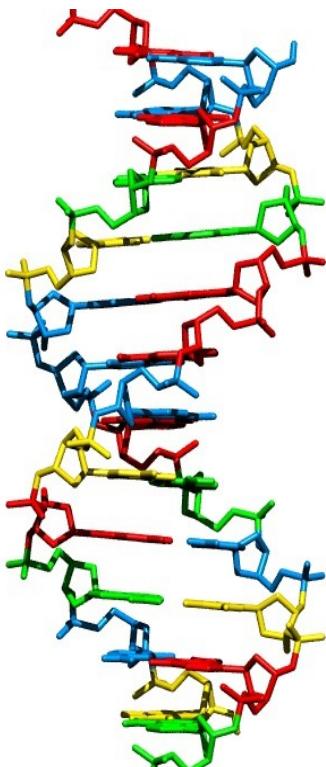
- The Genome is really simple
 - Genes
 - Gene regulation
- The Genome is really fascinating
 - One code used in many contexts
 - Lots of code reuse
 - Output is breathtaking
- Biology is a vast and deep sea
 - Humanity's biological knowledge is shallow
 - Dive anywhere and you quickly reach the frontier of human knowledge



Genome Sequencing

DNA sequencing

Genomes are awesome.
Let's get'em.



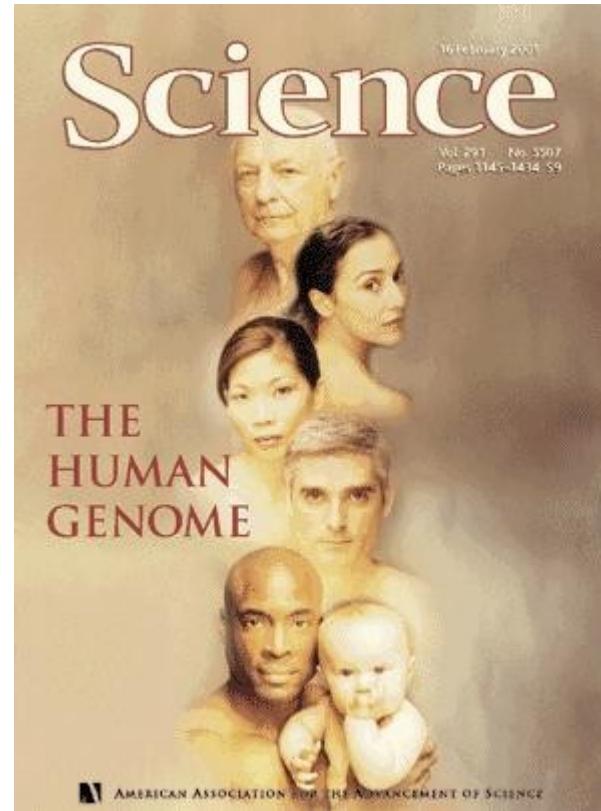
```
...ACGTGACTGAGGACC GTG  
CGACTGAGACTGACTGGGT  
CTAGCTAGACTACGTTTA  
TATATATATACGTCGT CGT  
ACTGATGACTAGATTACAG  
ACTGATTTAGATA CCTGAC  
TGATTTAAAAAAATATT...
```

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

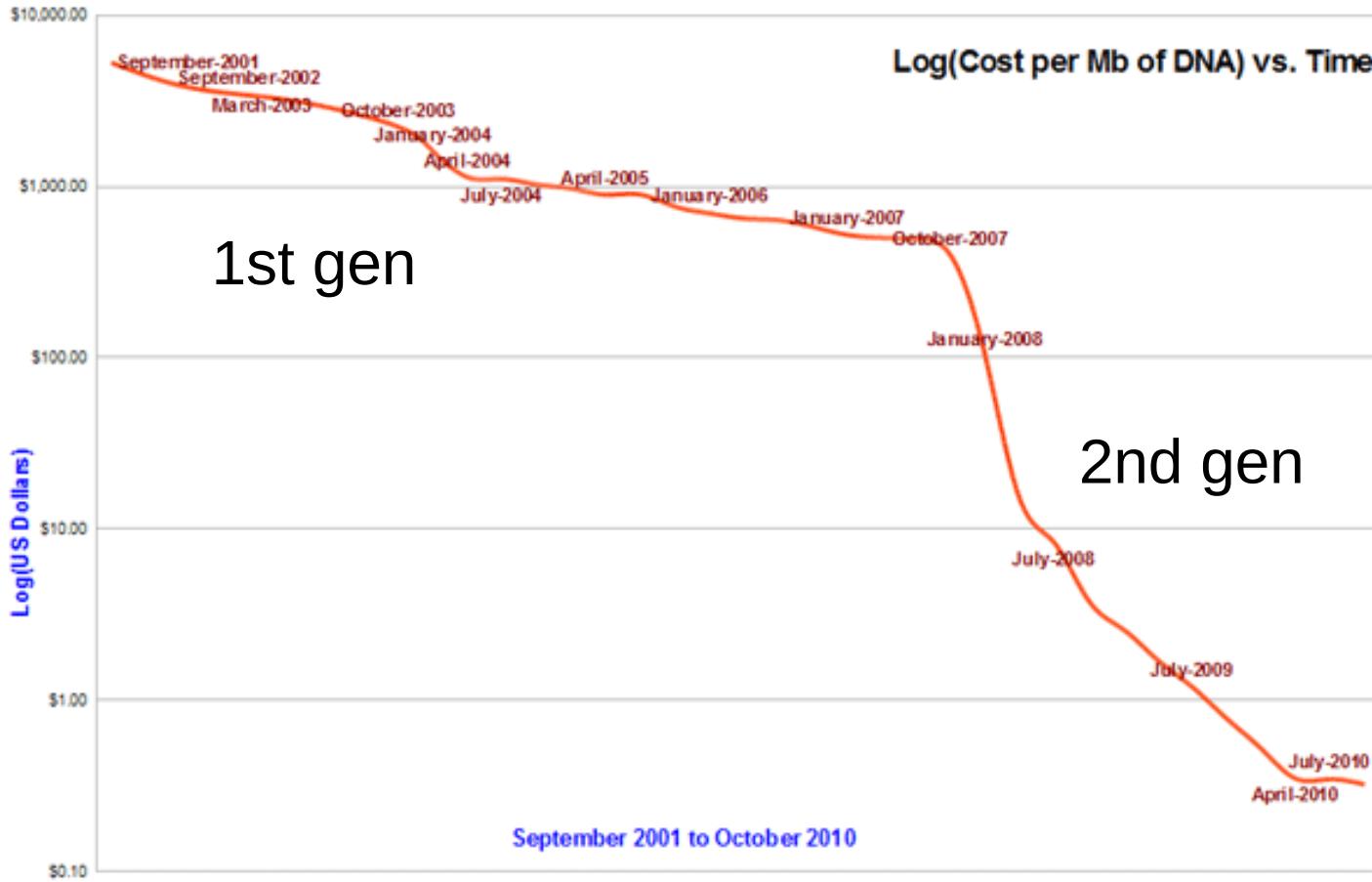


Celera

2001

Getting the “blueprint of life”

DNA sequencing costs



1st Genome Assembly

Some Terminology

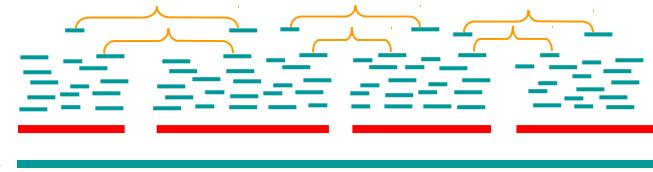
read a 500-900 long word that comes out of sequencer

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



..ACGATTACAATAGGTT..

2nd Gen: Next Generation (re)Sequencing

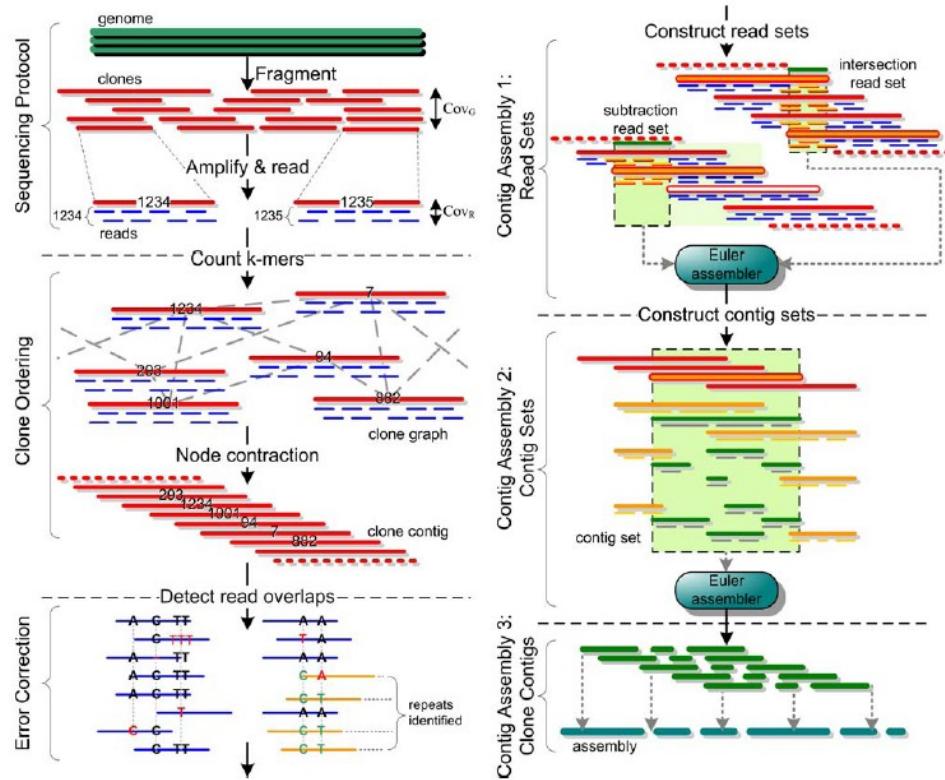


Figure 1. Sequencing protocol and assembly methodology. Reads are obtained in a hierarchical sequencing protocol with high genome-clone coverage and low clone-read coverage. From the k-mer content of each clone we construct a clone graph whose edge weights reflect the likely clone proximities, and from 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.

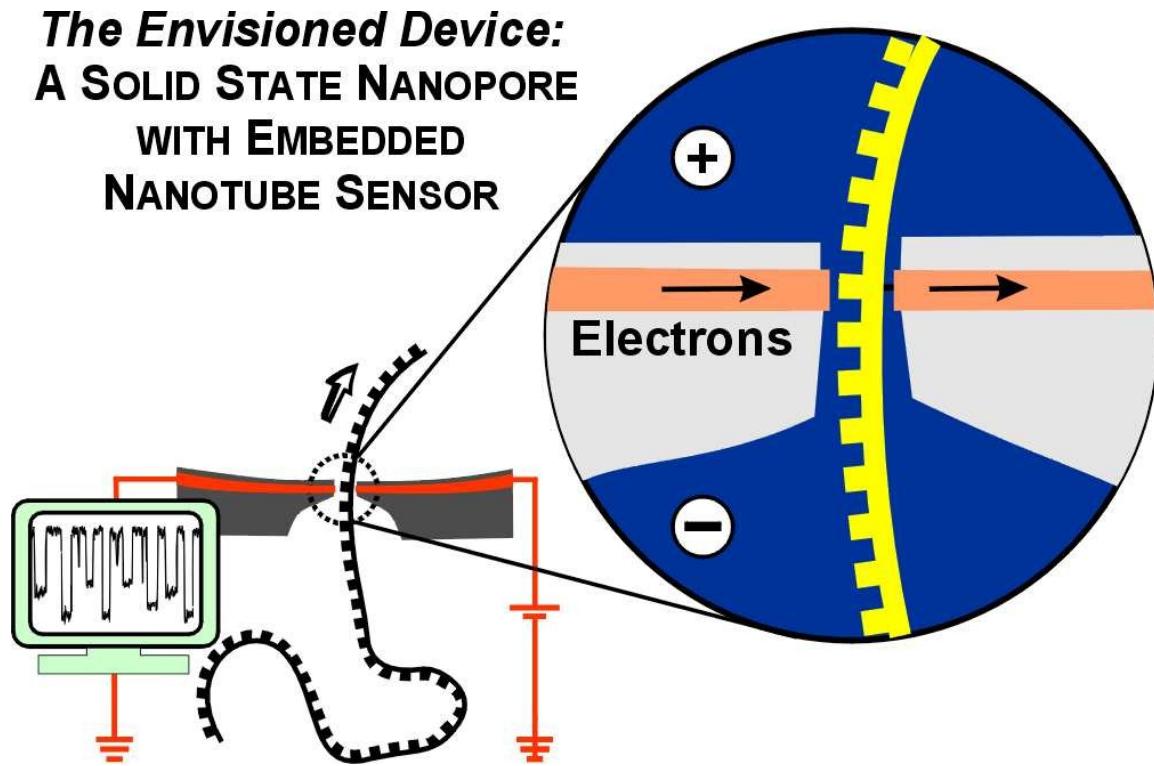
doi:10.1371/journal.pone.0000484.g001

Output = massive amounts of short, lower quality reads.

New Technologies + New Algorithms = New Opportunities

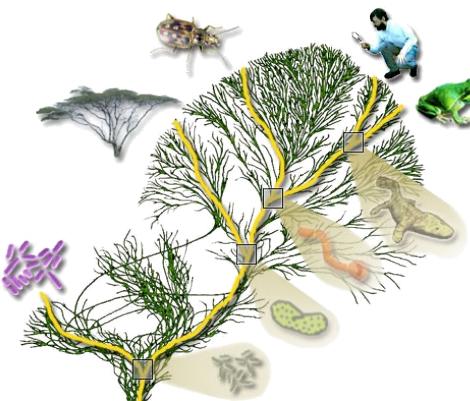
3rd Gen: cost effective, long reads

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

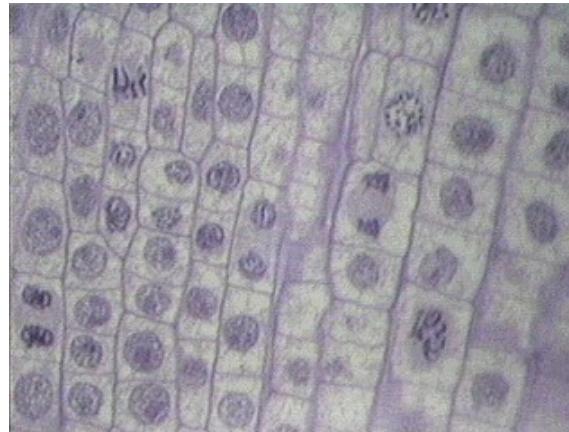
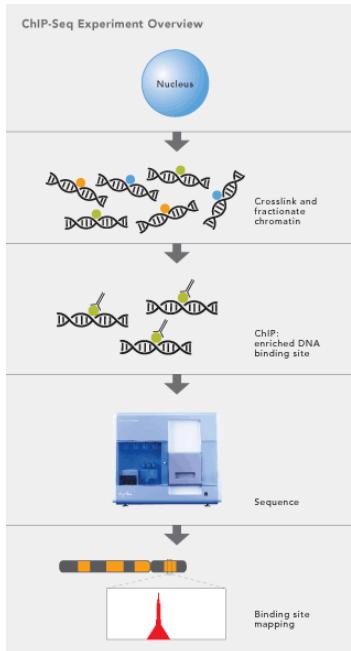


100 million species

7 billion
individuals



or sequence just
an active portion

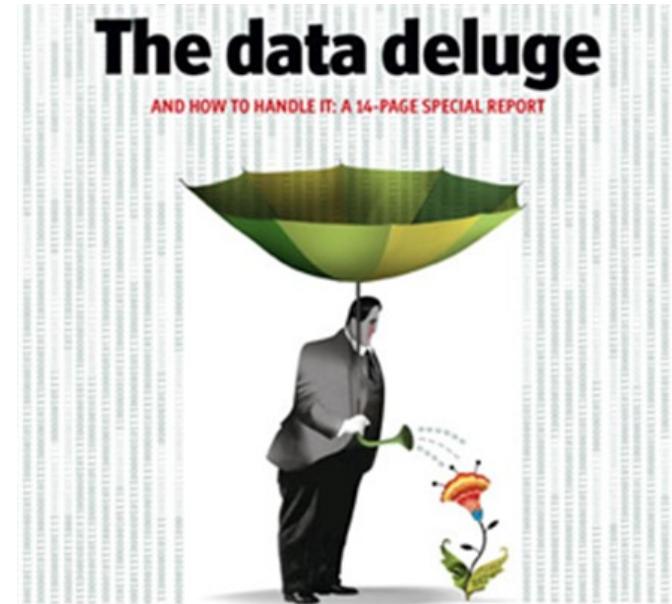


10^{13} cells
in a human

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 Visualization

<http://cs273a.stanford.edu> | Bejerano Fall 11 5 / 169

Portals to the Human Genome

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

The image shows a composite view of several genome portal websites. At the top left is the NCBI Map Viewer interface, featuring a blue header with the NCBI logo and a search bar. To its right is a vertical sequence of DNA bases (A, T, C, G) in grey. Below these are the UCSC Genome Bioinformatics and Ensembl Genome Browser interfaces. The UCSC interface has a yellow header with 'UCSC Genome Bioinformatics' and a blue navigation bar with links like 'Genomes', 'Blat', 'Tables', etc. The Ensembl interface has a red header with 'Ensembl' and a yellow navigation bar with 'Ensembl Genome Browser'. Both interfaces include search bars and various genomic data visualization tools.

Genome Browser Database

Home Genomes Blat Tables Gene Sorter PCR DNA Convert PDF/PS Help

UCSC Genome Browser on Human Mar. 2006 Assembly

move << << < > >> zoom in 1.5x 3x 10x base zoom out 1.5x 3x 10x

position/search chrX:151,073,054-151,383,976 jump clear size 310,923 bp. configure

chrX (228) [chrX:151,073,054-151,383,976] 28,045 1 28

STS Markers on Genetic (blue) and Radiation Hybrid (black) Maps

UCSC Known Genes Based on UniProt, RefSeq, and GenBank mRNA

RefSeq Genes

Human Gene Collection Full CDS mRNAs

BC028629

BC028629 Human mRNAs from

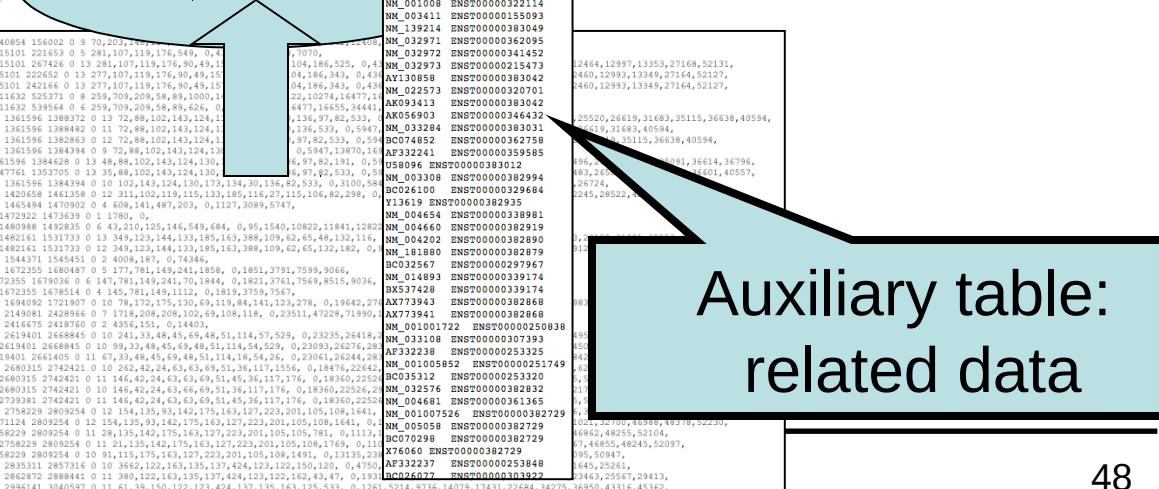
BC028629 S62988

Spliced ESTs Vertebrate Multiz Alignment 2

Conservation Human ESTs (17 SPECIES)

mouse rat rabbit dog armadillo opossum chicken X. tropicalis tettigideus Repeating Elements by RepeatMasker

visualize



Genome Evolution

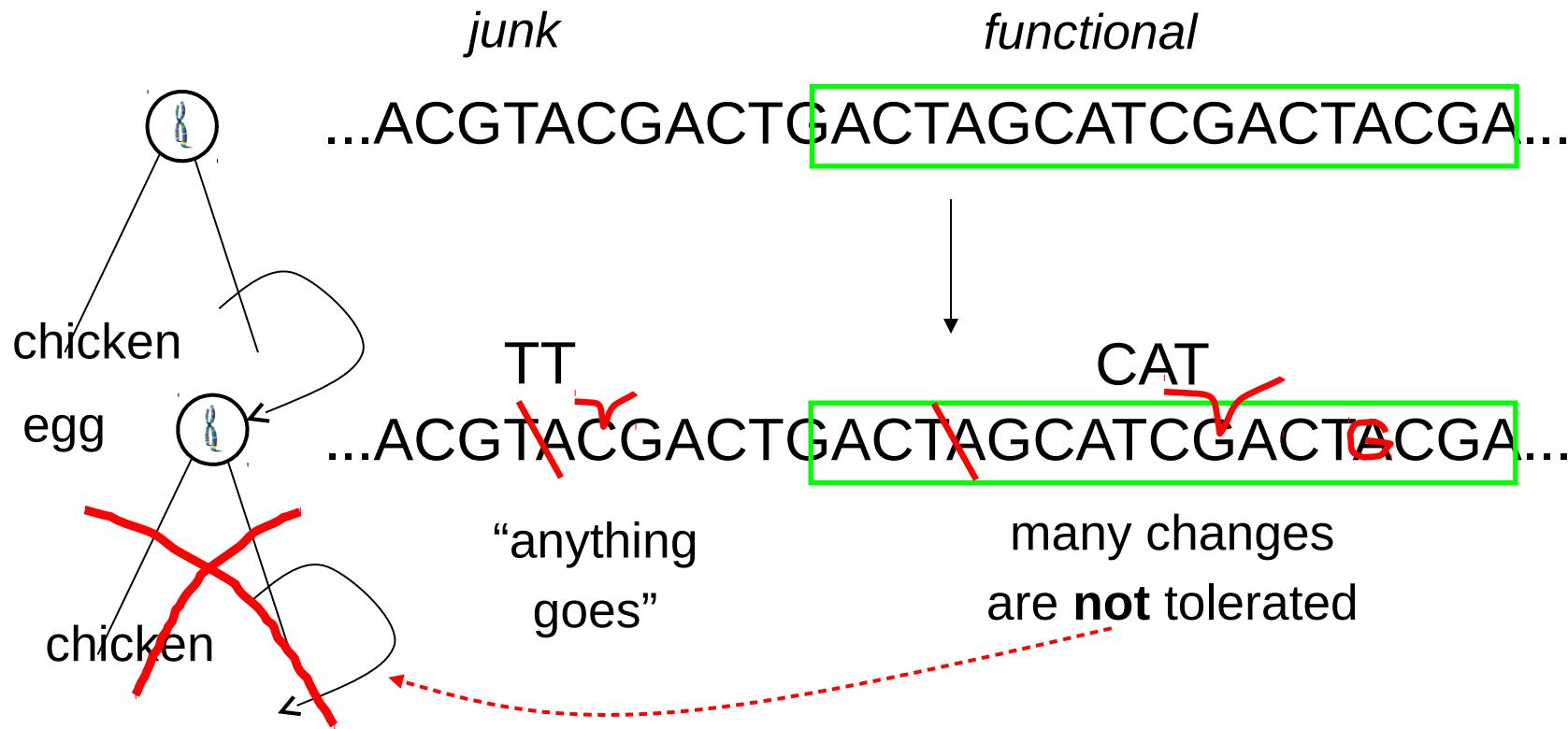
Genome Evolution

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.

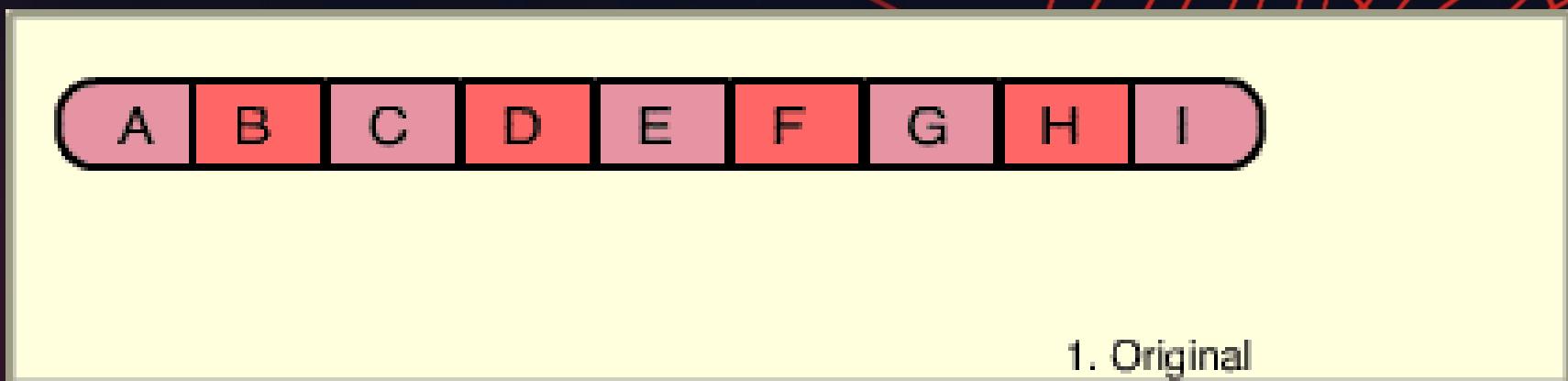
Genome mutation types: anything you can do to a string

Deletion

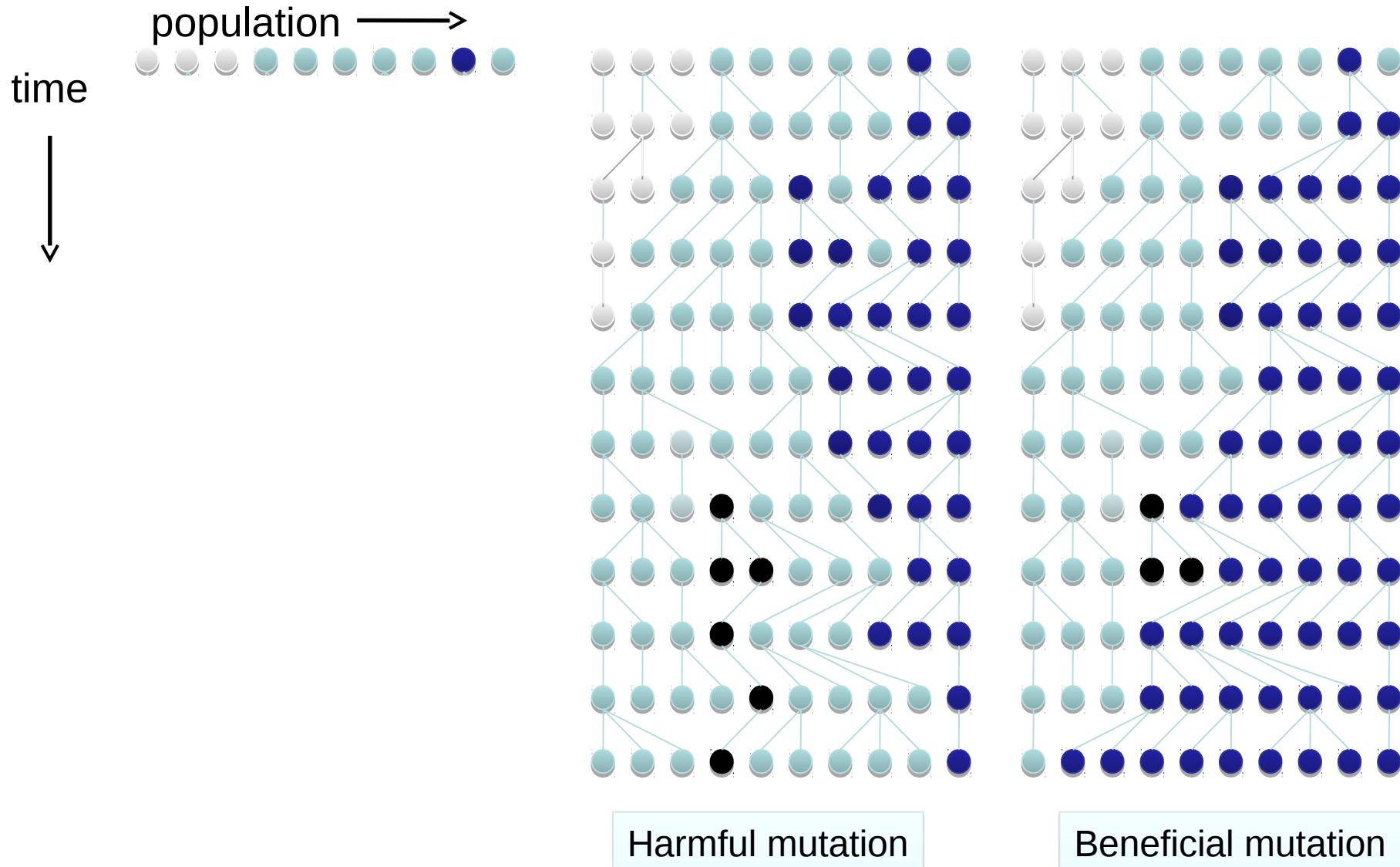
Inversion

Translocation

Duplication

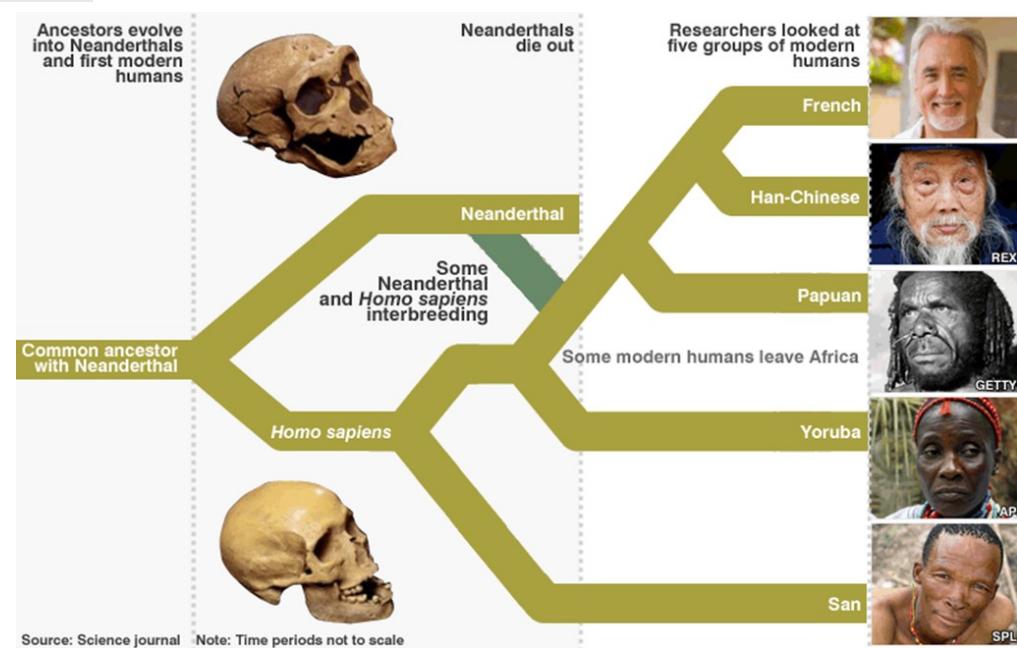
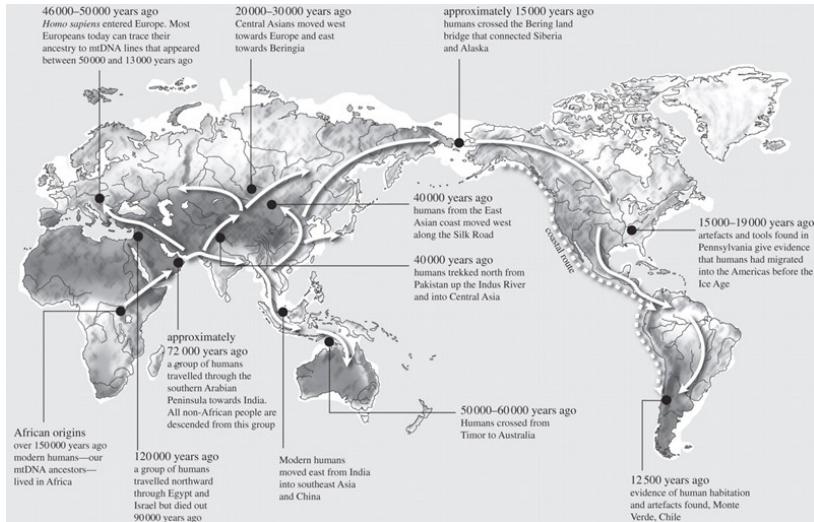


Modes of evolution = Mutation + Selection



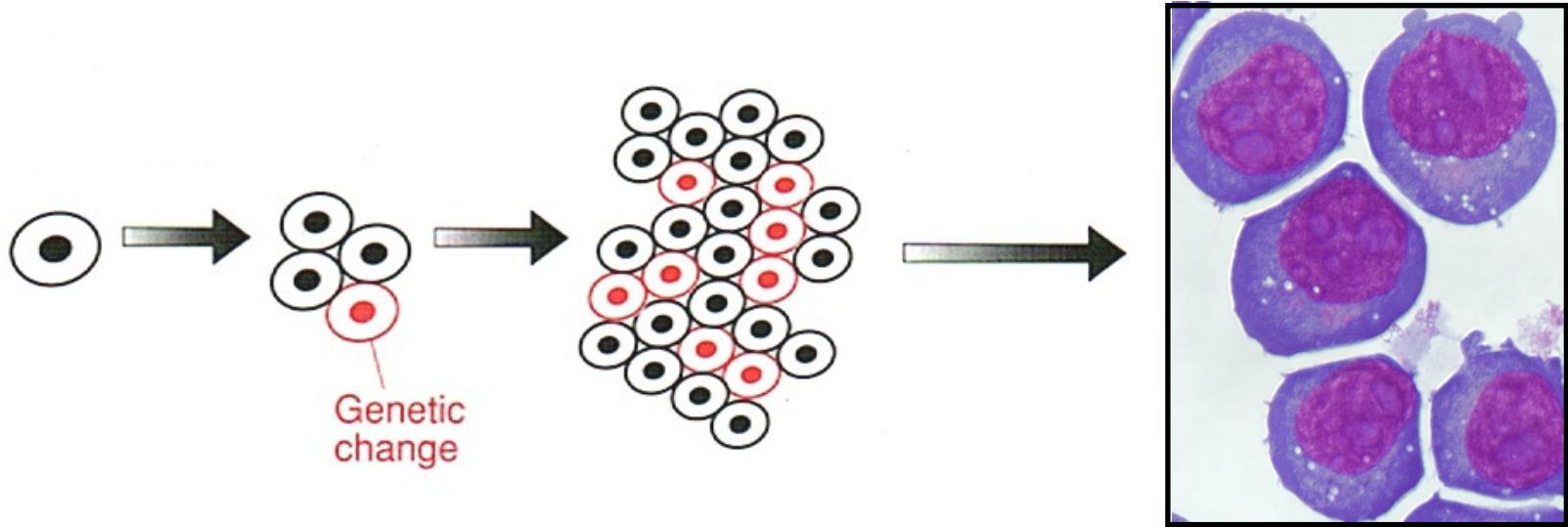
Population Genomic

From neutral evolution alone



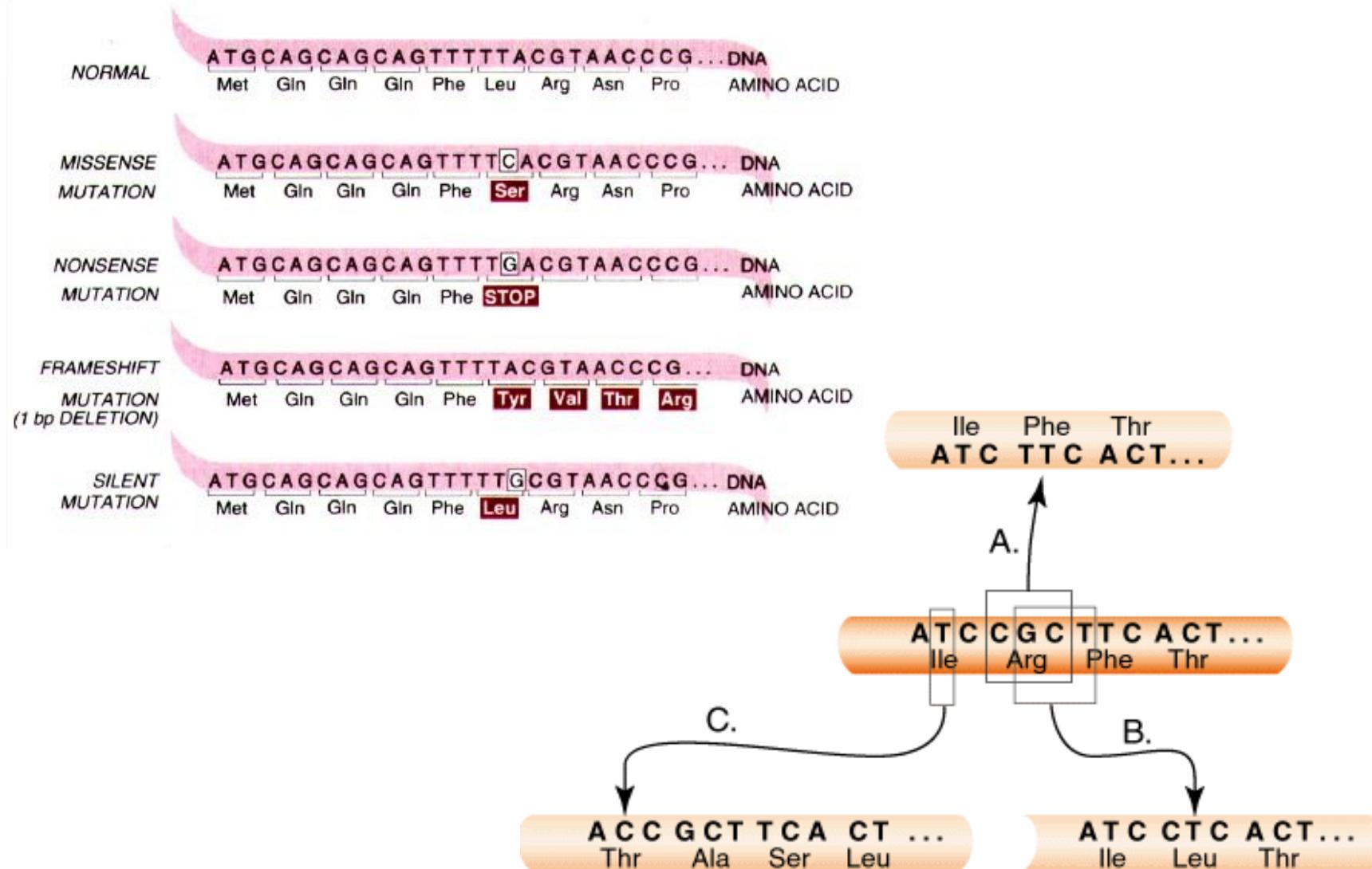
Genomics & Disease

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

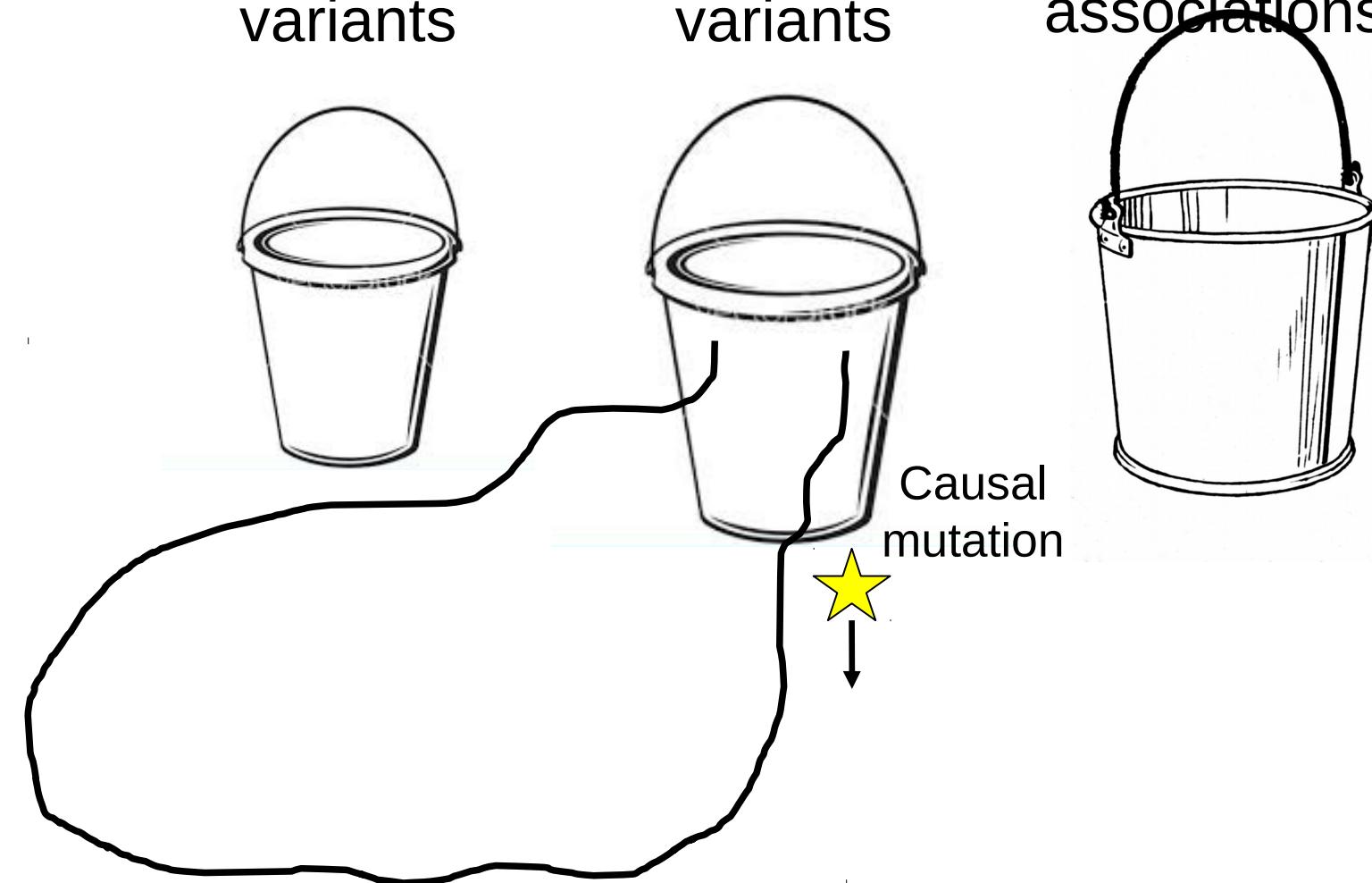


Mendelian Diseases

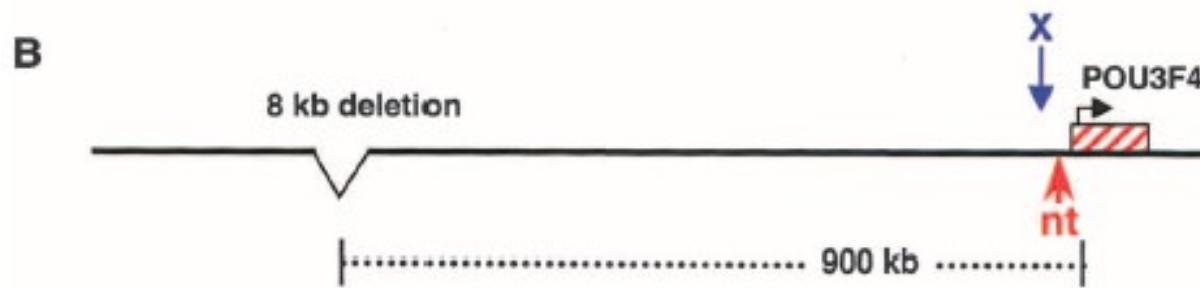
Affected
variants

Unaffected
variants

Gene-disease
associations



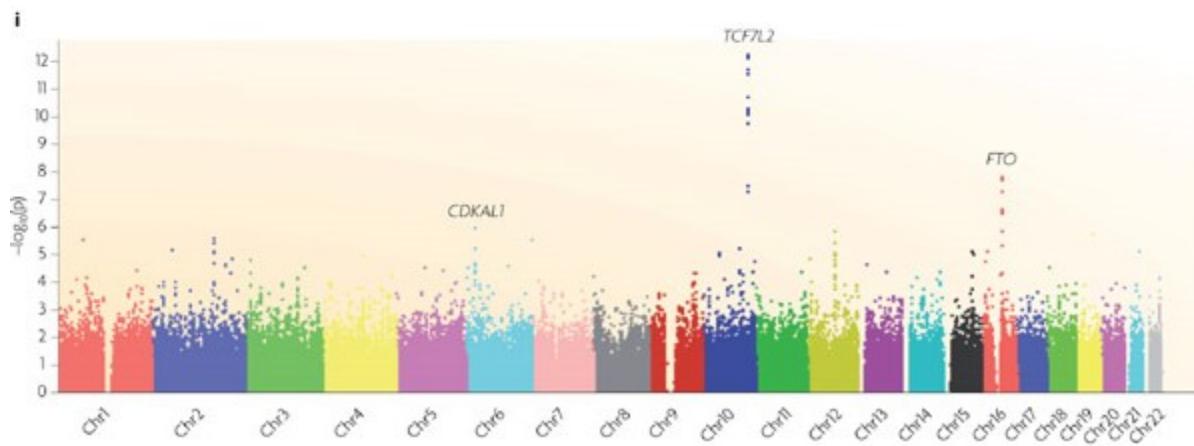
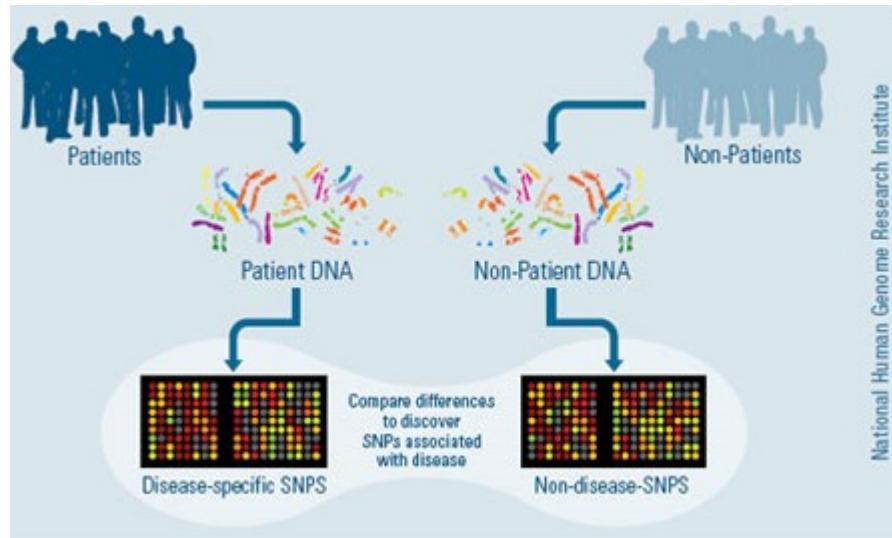
“Non-coding” mutations can be detrimental



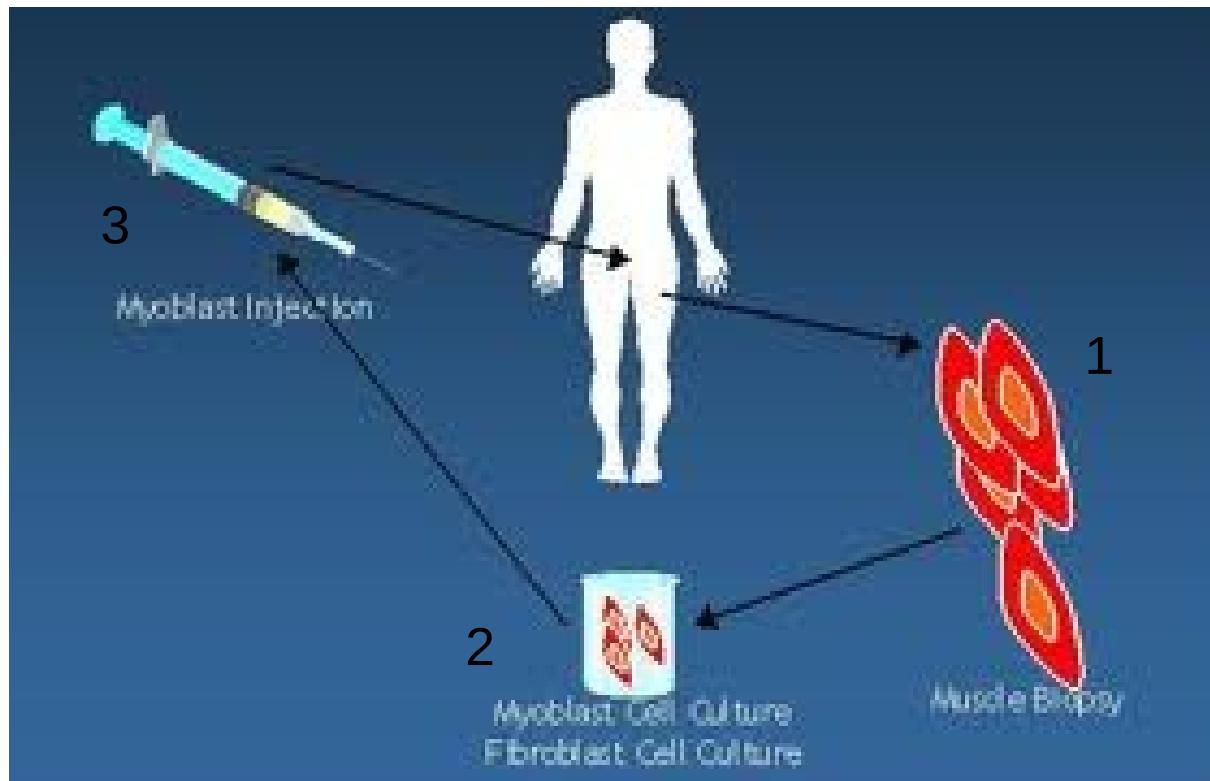
retinal expression. B, The human POU3F4 deafness locus. The microdeletion of an 8-kb region located 900 kb upstream of the gene contains a conserved noncoding sequence, the loss of which leads to congenital deafness. The mouse slf inversion breakpoint X leaves the neural tube

[de Kok et al, 1996]

Finding Disease Loci



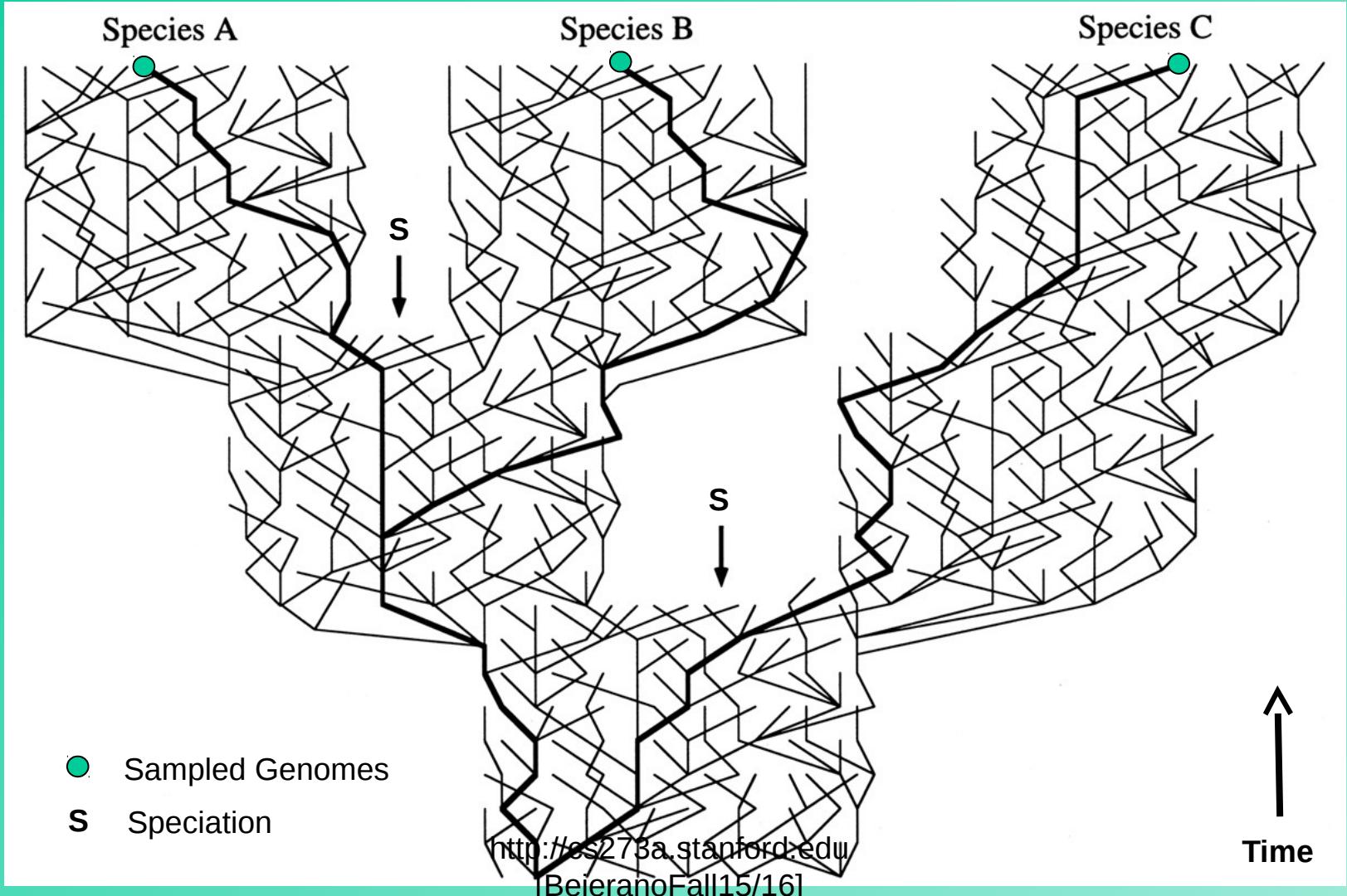
Gene/Cell Therapy: Curing Genomics Defects



- 1 Get'em
- 2 Fix'em
- 3 Put'em back

Organism Evolution

Evolution is not all bad!



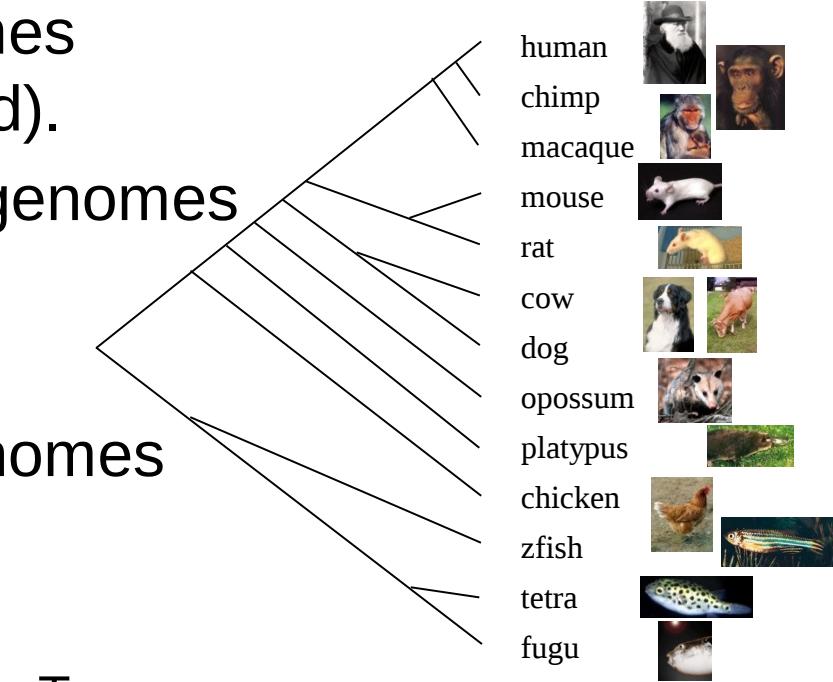
Comparative Genomics, Evo Devo

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

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

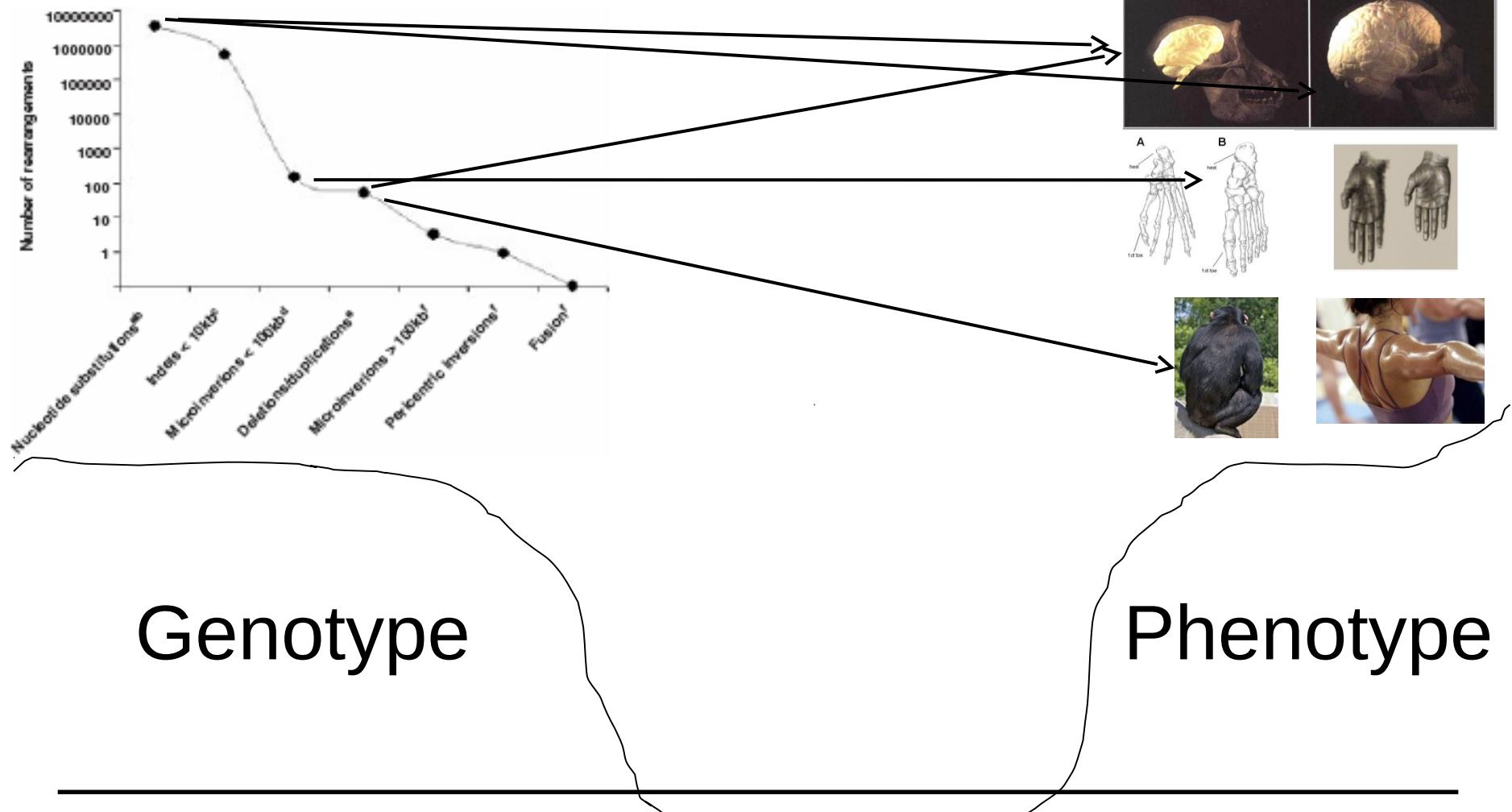
How to learn from different genomes
about their owners?

(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!



Further Reading

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

At which point we will ask ourselves:

Are we any wiser?

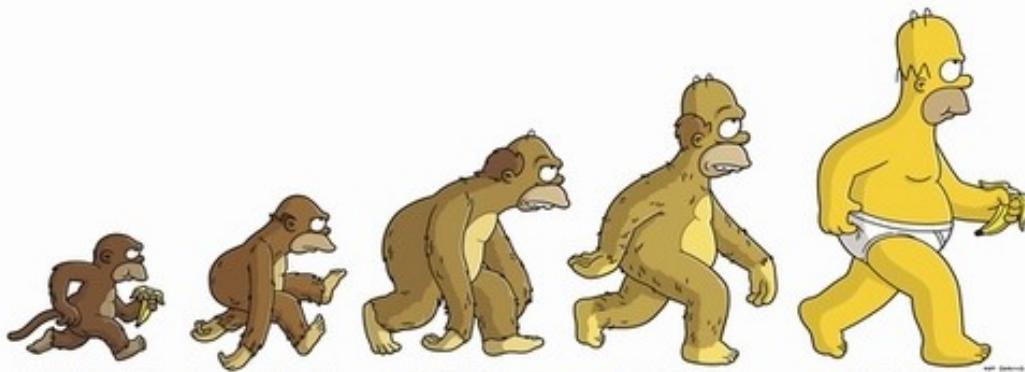
Check out the Bejerano lab “resources” page:

- 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...

