CS 68: BIOINFORMATICS

Prof. Sara Mathieson
Swarthmore College
Spring 2018

Outline: Jan 22

- Bioinformatics overview
- Introductions
- Syllabus highlights
- Intro to molecular biology and the central dogma

Bioinformatics Overview

Why take Bioinformatics?

Why take Bioinformatics?

- In the last 20 years, genome sequencing costs have plummeted and as a result, we have amazing “big data”

Why take Bioinformatics?

- In the last 20 years, genome sequencing costs have plummeted and as a result, we have amazing “big data”
- We have data from tens of thousands of species and hundreds of thousands of individuals from our species

Why take Bioinformatics?

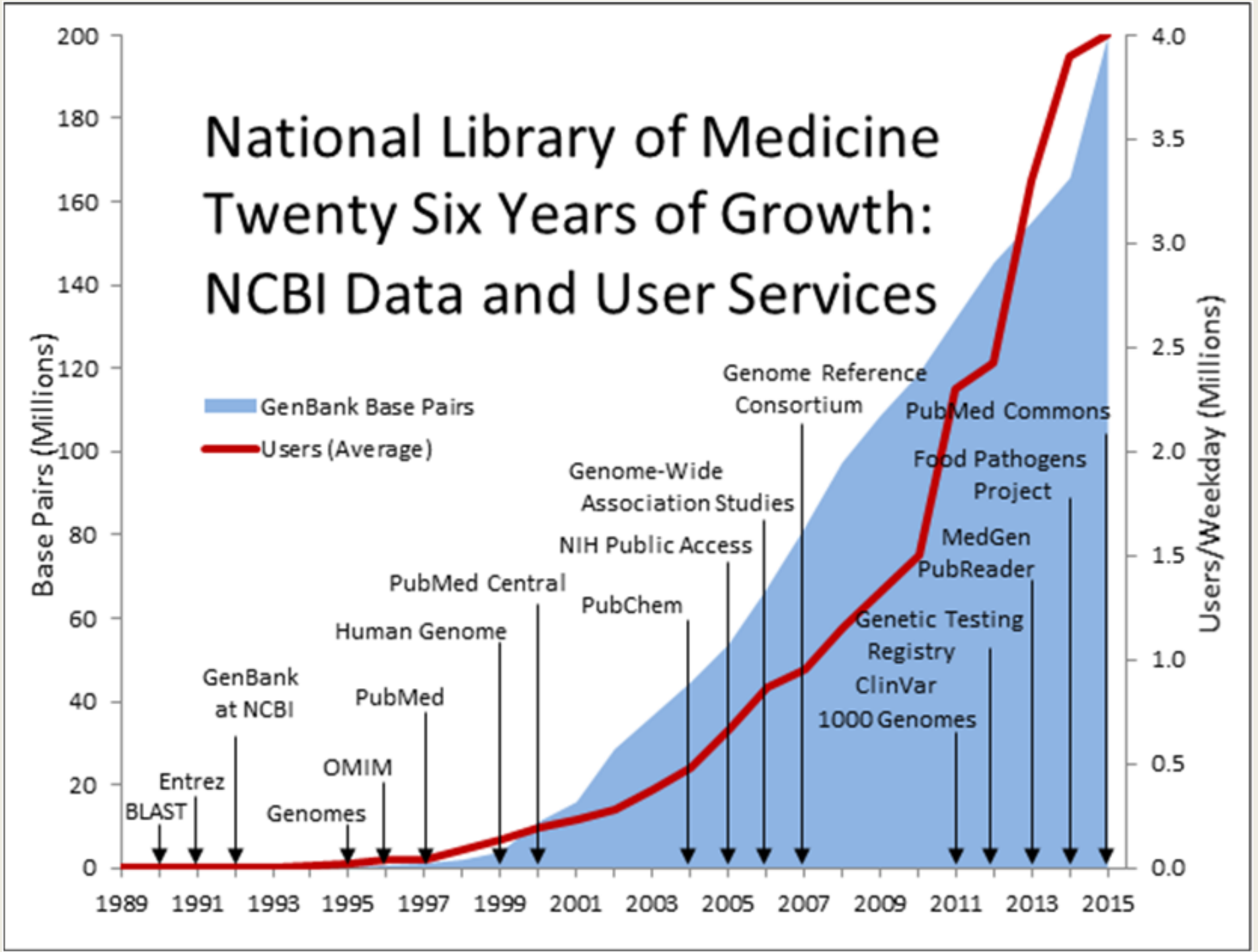
- In the last 20 years, genome sequencing costs have plummeted and as a result, we have amazing “big data”
- We have data from tens of thousands of species and hundreds of thousands of individuals from our species
- We are now in a position to answer biological questions with this data, but algorithms for analyzing and learning from this data have not developed at the same pace

Why take Bioinformatics?

- In the last 20 years, genome sequencing costs have plummeted and as a result, we have amazing “big data”
- We have data from tens of thousands of species and hundreds of thousands of individuals from our species
- We are now in a position to answer biological questions with this data, but algorithms for analyzing and learning from this data have not developed at the same pace
- CS 68 is an opportunity to learn how biological data has driven algorithm development, and how existing algorithms have been repurposed for biology

Why take Bioinformatics?

- In the last 20 years, genome sequencing costs have plummeted and as a result, we have amazing “big data”
- We have data from tens of thousands of species and hundreds of thousands of individuals from our species
- We are now in a position to answer biological questions with this data, but algorithms for analyzing and learning from this data have not developed at the same pace
- CS 68 is an opportunity to learn how biological data has driven algorithm development, and how existing algorithms have been repurposed for biology
- We will also discuss the future of bioinformatics and challenging problems that remain unsolved


National Library of Medicine
Twenty Six Years of Growth:
NCBI Data and User Services
Base Pairs (Millions)
Users/Weekday (Millions)
GenBank Base Pairs
Users (Average)
BLAST
Entrez
GenBank at NCBI
Genomes
OMIM
PubMed
Human Genome
PubMed Central
PubChem
Genome-Wide Association Studies
NIH Public Access
Genome Reference Consortium
1000 Genomes
ClinVar
Genetic Testing Registry
MedGen
PubReader
Food Pathogens Project
PubMed Commons
0
20
40
60
80
100
120
140
160
180
200
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
1989
1991
1993
1995
1997
1999
2001
2003
2005
2007
2009
2011
2013
2015

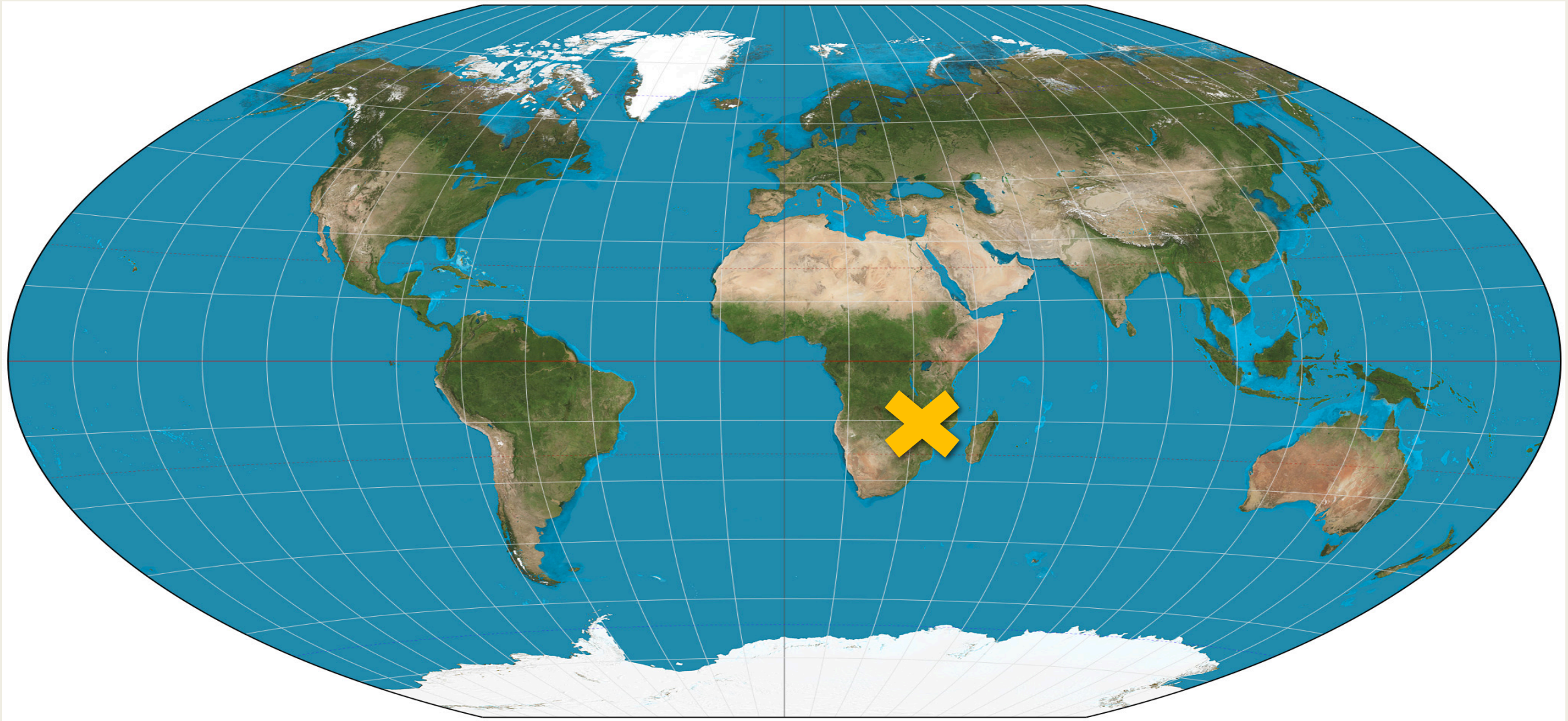
Example questions we could answer with data

- What percentage of sites in the human genome are different from chimp? Which sites make us “human”?
- How do healthy cells become cancerous and proliferate into tumors?
- Which genetic variants cause diseases such as diabetes, bipolar disorder, autism?
- When did humans first reach various regions (Middle East, Europe, Asia, Australia, the Americas)?

Example questions we could answer with data

- What percentage of sites in the human genome are different from chimp? Which sites make us “human”?
- How do healthy cells become cancerous and proliferate into tumors?
- Which genetic variants cause diseases such as diabetes, bipolar disorder, autism?
- When did humans first reach various regions (Middle East, Europe, Asia, Australia, the Americas)?

Origin of modern humans

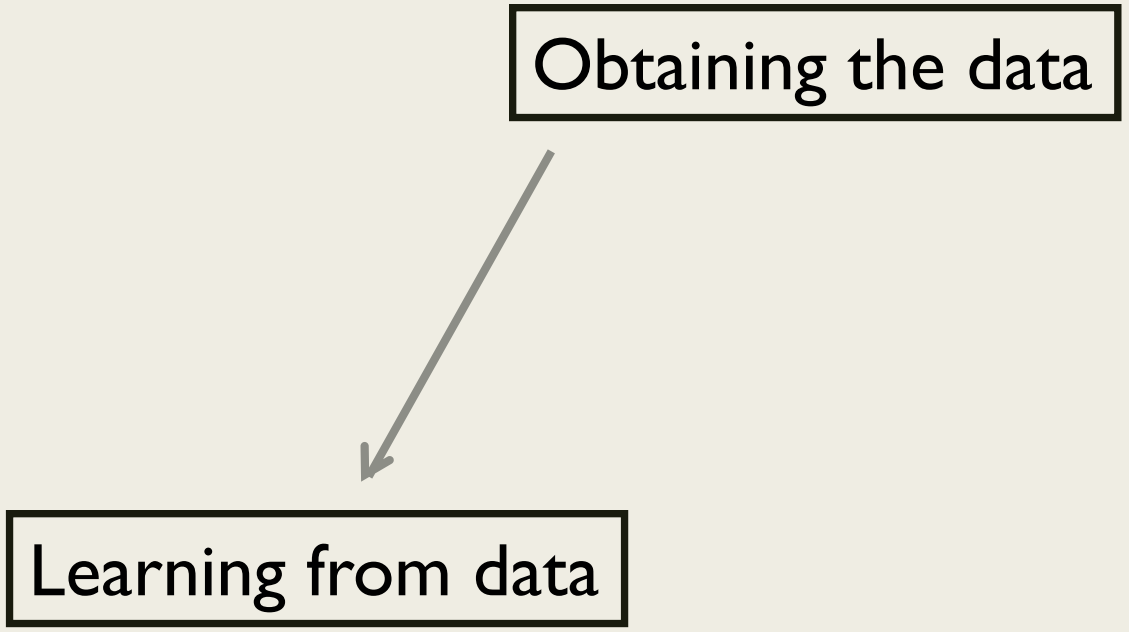


Example: how to “learn from data”

Learning from data

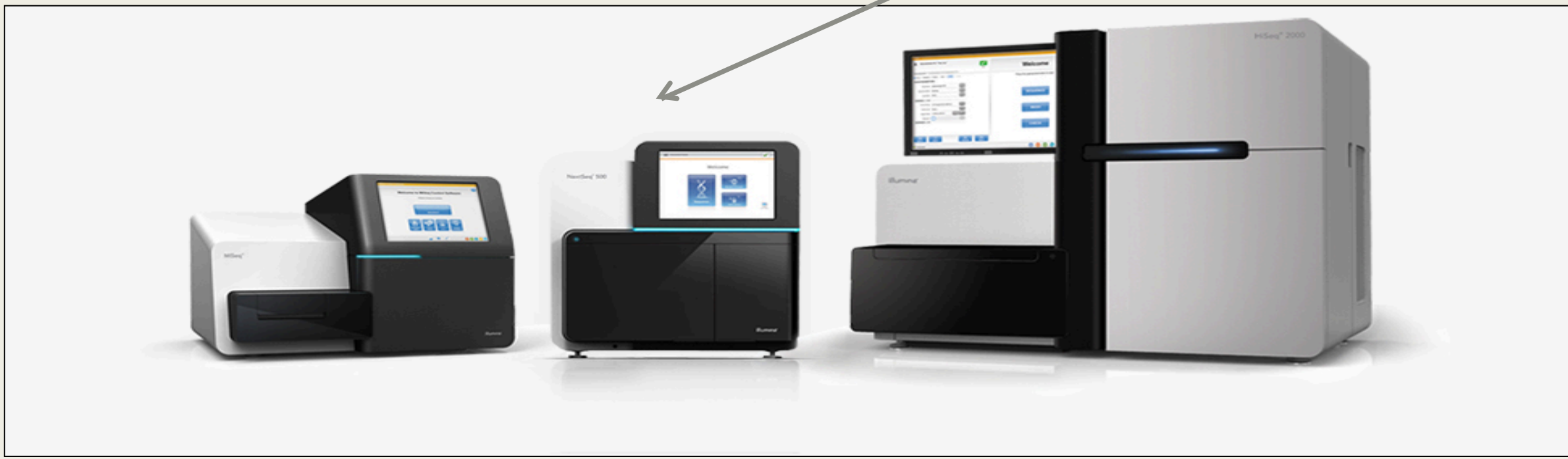
Question: How did humans move out of Africa?

Example: how to “learn from data”



Question: How did humans move out of Africa?

Obtaining the data

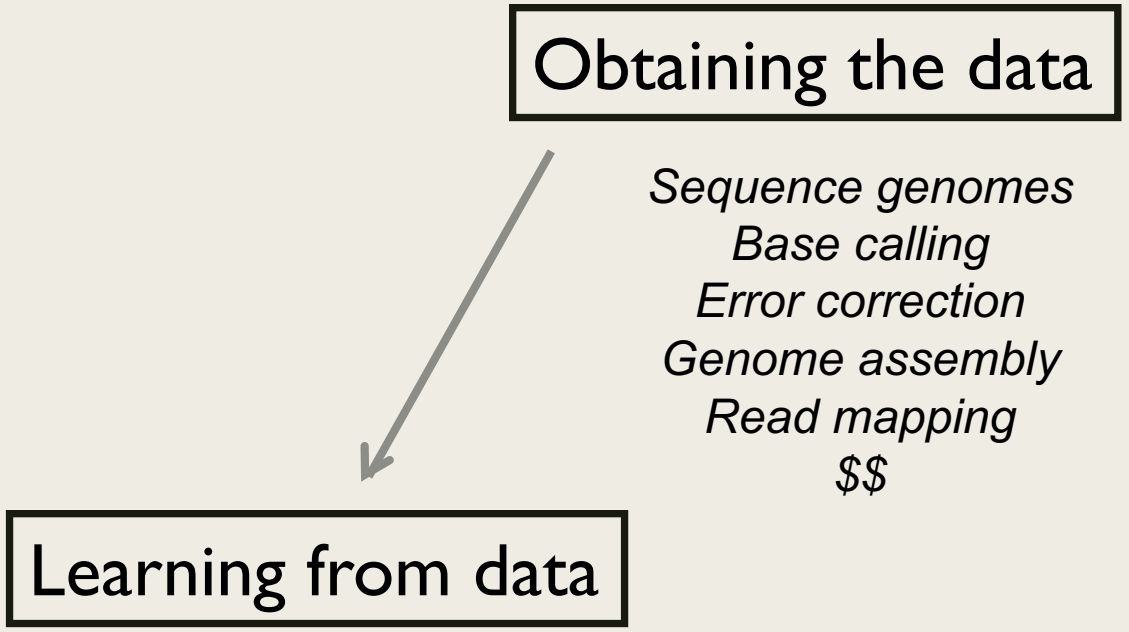


Illumina

AGCCCTAATCTAACCCTGGCCAACCTGTCTCTCAACTTACCCTCCATTACCCTGCCTCCACTCGTTACCCTGTCCCATTCAAC
CATACCACTCCGAACCACCATCCATCCCTCTACTTACTACCACTCACCCACCGTTACCCTCCAATTACCCATATCCAACCCACT
GCCACTTACCCTACCATTACCCTACCATCCACCATGACCTACTCACCATACTGTTCTTCTACCCACCATATTGAAACGCTAACA
AATGATCGTAAATAACACACACGTGCTTACCCTACCACTTTATACCACCACCACATGCCATACTCACCCTCACTTGTATACTGA
TTTTACGTACGCACACGGATGCTACAGTATATACCATCTCAAACTTACCCTACTCTCAGATTCCACTTCACTCCATGGCCCATC
TCTCACTGAATCAGTACCAAATGCACTCACATCATTATGCACGGCACTTGCCTCAGCGGTCTATACCCTGTGCCATTTACCCA
TAACGCCCATCATTATCCACATTTTGATATCTATATCTCATTCGGCGGTCCCAAATATTGTATAACTGCCCTTAATACATACGTT
ATACCACTTTTGCACCATATACTTACCACTCCATTTATATACACTTATGTCAATATTACAGAAAAATCCCCACAAAAATCACCTA
AACATAAAAATATTCTACTTTTCAACAATAATACATAAACATATTGGCTTGTGGTAGCAACACTATCATGGTATCACTAACGTAA
AAGTTCCTCAATATTGCAATTTGCTTGAACGGATGCTATTTCAGAATATTTCGTACTTACACAGGCCATACATTAGAATAATAT
GTCACATCACTGTCGTAACACTCTTTATTCACCGAGCAATAATACGGTAGTGGCTCAAACTCATGCGGGTGCTATGATACAAT
TATATCTTATTTCCATTCCCATATGCTAACCGCAATATCCTAAAAGCATAACTGATGCATCTTTAATCTTGTATGTGACACTACT
CATACGAAGGGACTATATCTAGTCAAGACGATACTGTGATAGGTACGTTATTTAATAGGATCTATAACGAAATGTCAAATAATT
TTACGGTAATATAACTTATCAGCGGCGTATACTAAAACGGACGTTACGATATTGTCTCACTTCATCTTACCACCCTCTATCTTAT
TGCTGATAGAACACT[illegible]GATATTTT
ACGTGTCAAAAAATC[illegible]ATCTTGTT
CTTAGAAGTGACGCA[illegible]TTATTTAA
GGACAAAGGTTGCGAAGCCGCACATTTCCAATTTCATTGTTGTTTATTGGACATACACTGTTAGCTTTATTACCGTCCACGTT
TTTTCTACAATAGTGTAGAAGTTTCTTTCTTATGTTCATCGTATTCATAAAATGCTTCACGAACACCGTCATTGATCAAATAGG
TCTATAATATTAATATACATTTATATAATCTACGGTATTTATATCATCAAAAAAAAGTAGTTTTTTTATTTTATTTTGTTCGTTAAT
TTTCAATTTCTATGGAAACCCGTTCGTAAAATTGGCGTTTGTCTCTAGTTTGCGATAGTGTAGATACCGTCCTTGGATAGAGC
ACTGGAGATGGCTGGCTTTAATCTGCTGGAGTACCATGGAACACCGGTGATCATTCTGGTCACTTGGTCTGGAGCAATACCG
GTCAACATGGTGGTGAAGTCACCGTAGTTGAAAACGGCTTCAGCAACTTCGACTGGGTAGGTTTCAGTTGGGTGGGCGGCT
TGGAACATGTAGTATTGGGCTAAGTGAGCTCTGATATCAGAGACGTAGACACCCAATTCCACCAAGTTGACTCTTTCGTCAG
ATTGAGCTAGAGTGGTGGTTGCAGAAGCAGTAGCAGCGATGGCAGCGACACCAGCGGCGATTGAAGTTAATTTGACCATTG
TATTTGTTTTGTTTGTTAGTGCTGATATAAGCTTAACAGGAAAGGAAAGAATAAAGACATATTCTCAAAGGCATATAGTTGAAG
CAGCTCTATTTATACCCATTCCCTCATGGGTTGTTGCTATTTAAACGATCGCTGACTGGCACCAGTTCCTCATCAAATATTCTC
TATATCTCATCTTTCACACAATCTCATTATCTCTATGGAGATGCTCTTGTTTCTGAACGAATCATAAATCTTTCATAGGTTTCGT
ATGTGGAGTACTGTTTTATGGCGCTTATGTGTATTCGTATGCGCAGAATGTGGGAATGCCAATTATAGGGGTGCCGAGGTGC
CTTATAAAACCCTTTTCTGTGCCTGTGACATTTCCTTTTTCGGTCAAAAAGAATATCCGAATTTTAGATTTGGACCCTCGTACA
GAAGCTTATTGTCTAAGCCTGAATTCAGTCTGCTTTAAACGGCTTCCGCGGAGGAAATATTTCCATCTCTTGAATTCGTACAA
CATTAAACGTGTGTTGGGAGTCGTATACTGTTAGGGTCTGTAAACTTGTGAACTCTCGGCAAATGCCTTGGTGCAATTACGT
AATTTTAGCCGCTGAGAAGCGGATGGTAATGAGACAAGTTGATATCAAACAGATACATATTTAAAAGAGGGTACCGCTAATTT
AGCAGGGCAGTATTATTGTAGTTTGATATGTACGGCTAACTGAACCTAAGTAGGGATATGAGAGTAAGAACGTTCGGCTACTC
TTCTTTCTAAGTGGGATTTTTCTTAATCCTTGGATTCTTAAAAGGTTATTAAAGTTCCGCACAAAGAACGCTTGGAAATCGCA

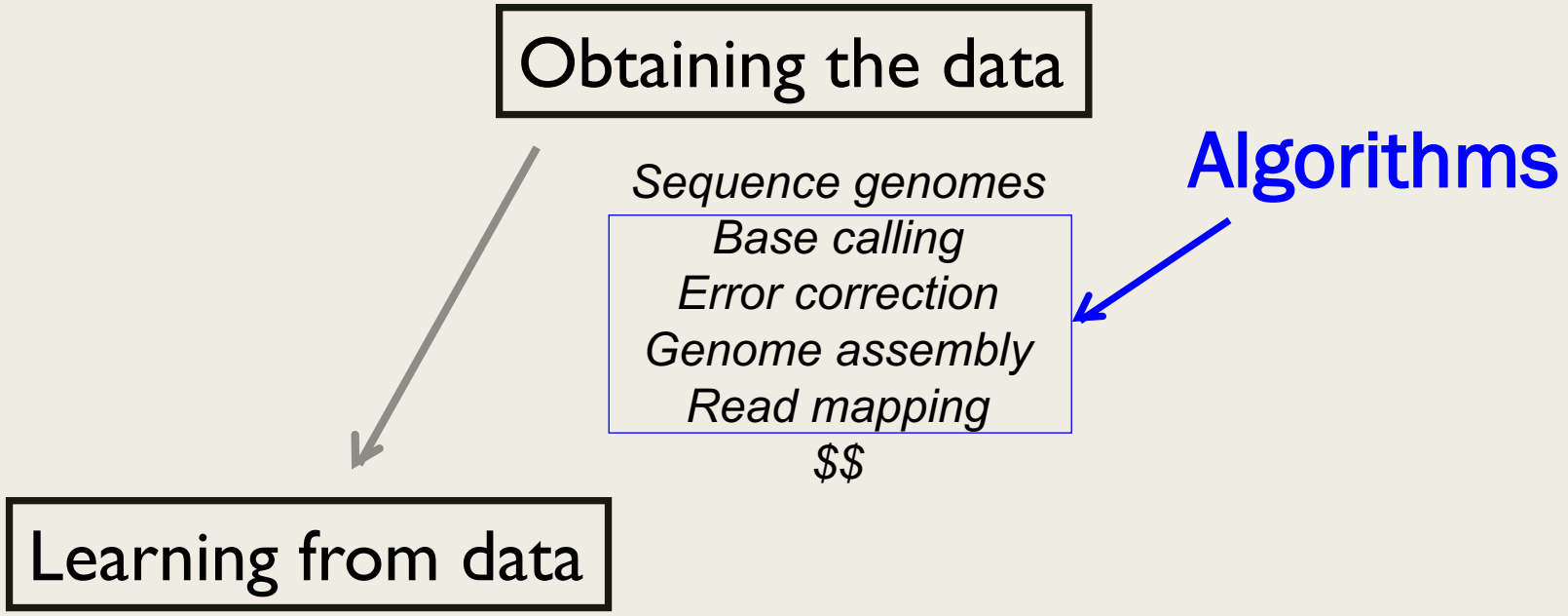
We obtain a “string” of *bases* (A,C,G,T)

Example: how to “learn from data”



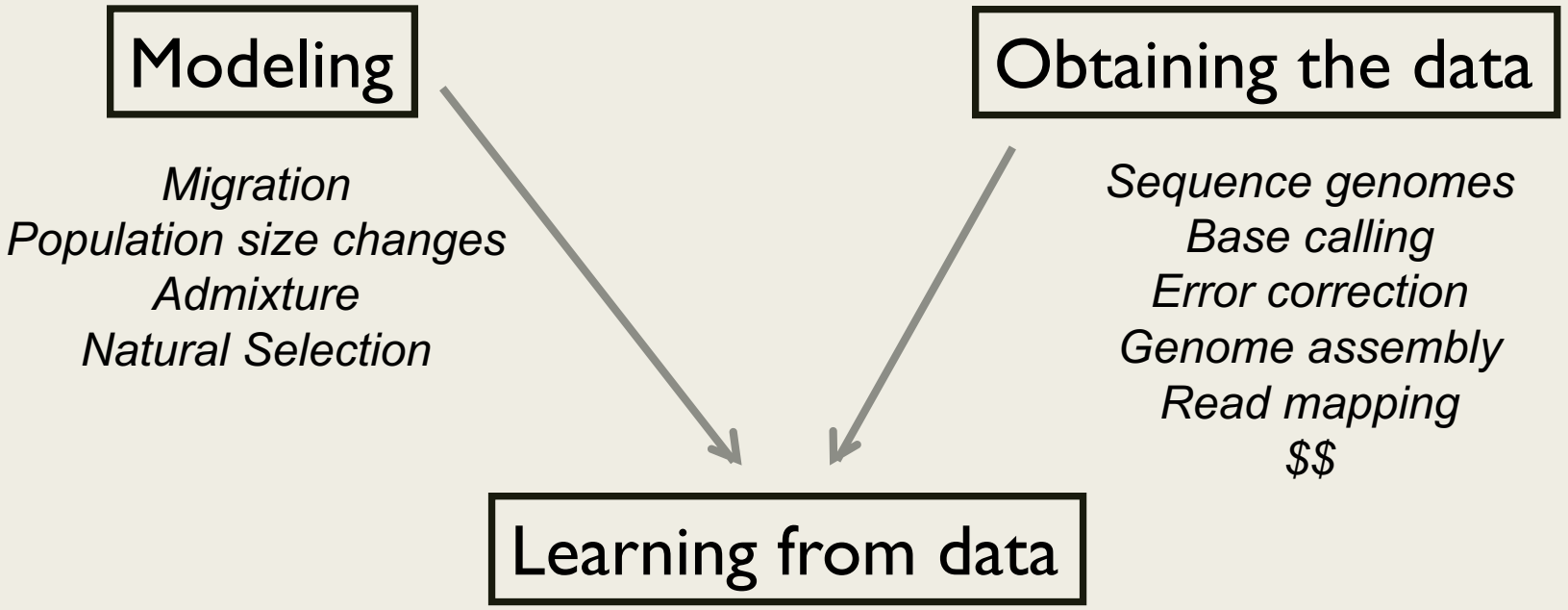
Question: How did humans move out of Africa?

Example: how to “learn from data”



Question: How did humans move out of Africa?

Example: how to “learn from data”



Question: How did humans move out of Africa?

Human migration out of Africa: hypothesis

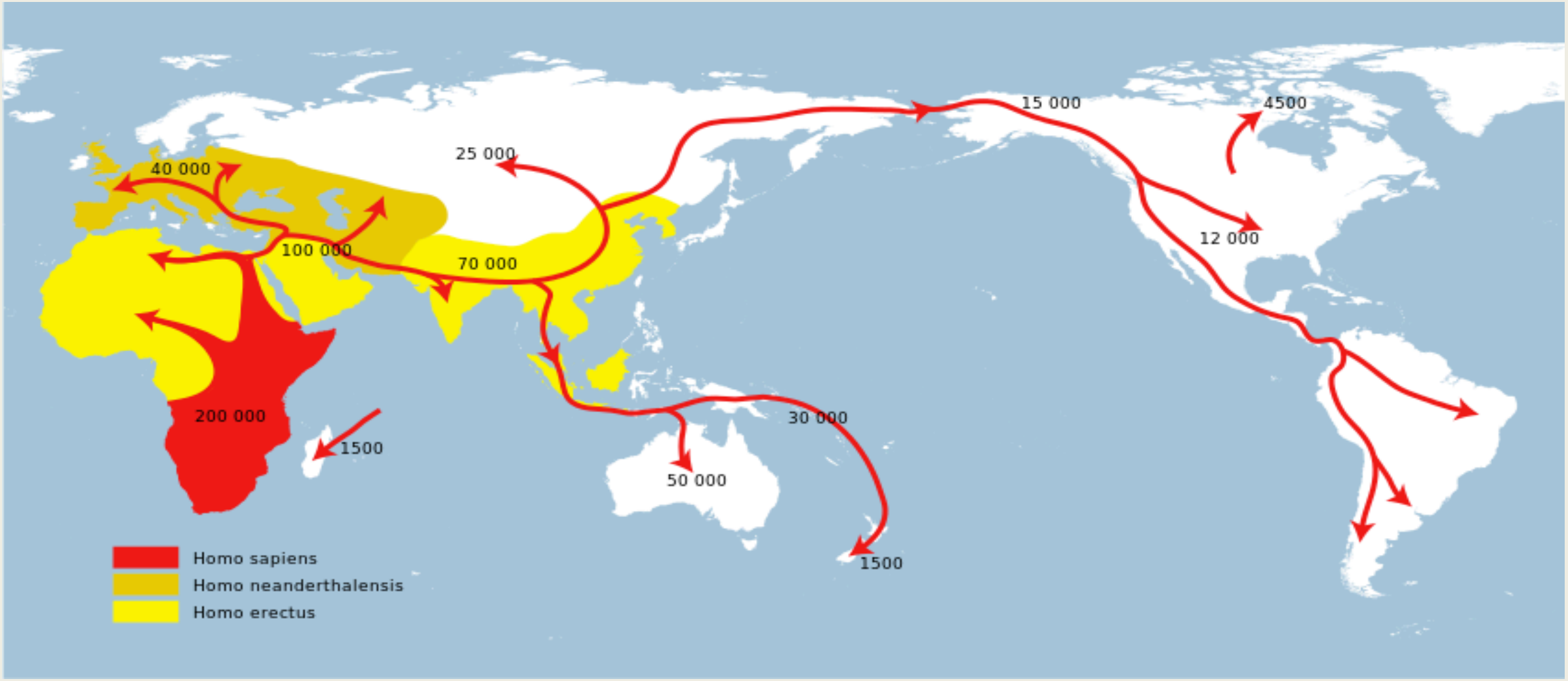


Image: NordNordWest

Introductions

With a partner briefly discuss…

1. How long is the human genome in base pairs?

   (A) 3 thousand (B) 3 million (C) 3 billion

2. If I compare two human genomes, approximately how often will there be a difference?

   (A) Every 10 bases (B) Every 100 bases (C) Every 1000 bases

3. How much does it cost to sequence a human genome?

   (A) $100 (B) $1,000 (C) $10,000

4. When did humans and chimp last share a common ancestor?

   (A) 1 thousand years ago (B) 1 million years ago (C) 10 million years ago

5. How long has life on earth been evolving?

With a partner briefly discuss…

1. How long is the human genome in base pairs?

   (A) 3 thousand (B) 3 million (C) 3 billion

2. If I compare two human genomes, approximately how often will there be a difference?

   (A) Every 10 bases (B) Every 100 bases (C) Every 1000 bases

3. How much does it cost to sequence a human genome?

   (A) $100 (B) $1,000 (C) $10,000

4. When did humans and chimp last share a common ancestor?

   (A) 1 thousand years ago (B) 1 million years ago (C) 10 million years ago

5. How long has life on earth been evolving?

With a partner briefly discuss…

1. How long is the human genome in base pairs?

   (A) 3 thousand (B) 3 million **(C) 3 billion**

2. If I compare two human genomes, approximately how often will there be a difference?

   (A) Every 10 bases (B) Every 100 bases **(C) Every 1000 bases**

3. How much does it cost to sequence a human genome?

   (A) $100 (B) $1,000 (C) $10,000

4. When did humans and chimp last share a common ancestor?

   (A) 1 thousand years ago (B) 1 million years ago (C) 10 million years ago

5. How long has life on earth been evolving?

With a partner briefly discuss…

1. How long is the human genome in base pairs?

(A) 3 thousand (B) 3 million **(C) 3 billion**

2. If I compare two human genomes, approximately how often will there be a difference?

(A) Every 10 bases (B) Every 100 bases **(C) Every 1000 bases**

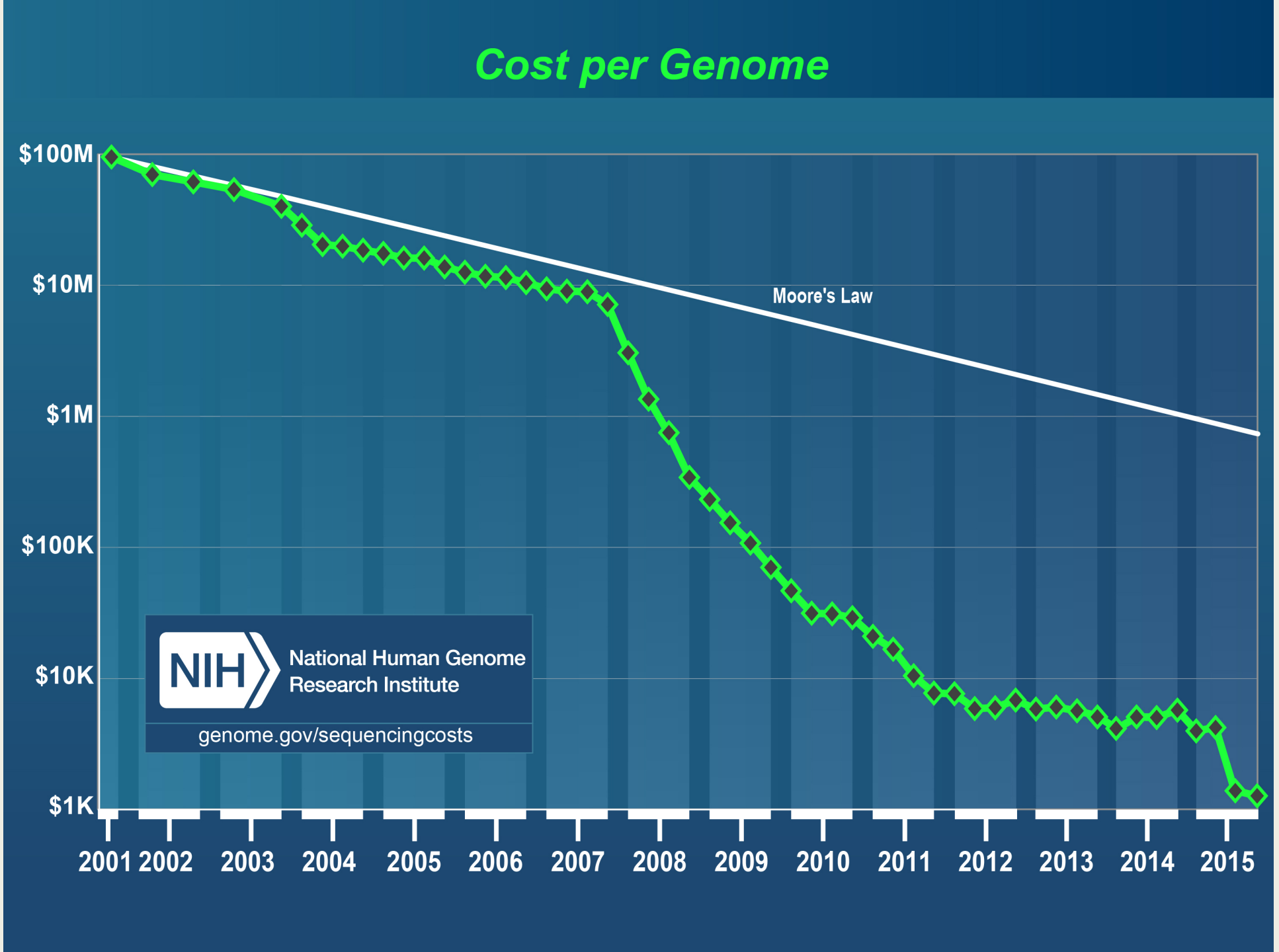
3. How much does it cost to sequence a human genome?

(A) $100 important variants **(B) $1,000** full sequence (C) $10,000

4. When did humans and chimp last share a common ancestor?

(A) 1 thousand years ago (B) 1 million years ago (C) 10 million years ago

5. How long has life on earth been evolving?


Cost per Genome
$100M
$10M
$1M
$100K
$10K
$1K
Moore's Law
NIH
National Human Genome Research Institute
genome.gov/sequencingcosts
2001 2002 2003 2004 2005 2006 2007 2008 2009 2010 2011 2012 2013 2014 2015

With a partner briefly discuss…

1. How long is the human genome in base pairs?

(A) 3 thousand (B) 3 million **(C) 3 billion**

2. If I compare two human genomes, approximately how often will there be a difference?

(A) Every 10 bases (B) Every 100 bases **(C) Every 1000 bases**

3. How much does it cost to sequence a human genome?

(A) $100 important variants **(B) $1,000** full sequence (C) $10,000

4. When did humans and chimp last share a common ancestor?

(A) 1 thousand years ago (B) 1 million years ago **(C) 10 million years ago**

5. How long has life on earth been evolving?

Human-chimp divergence

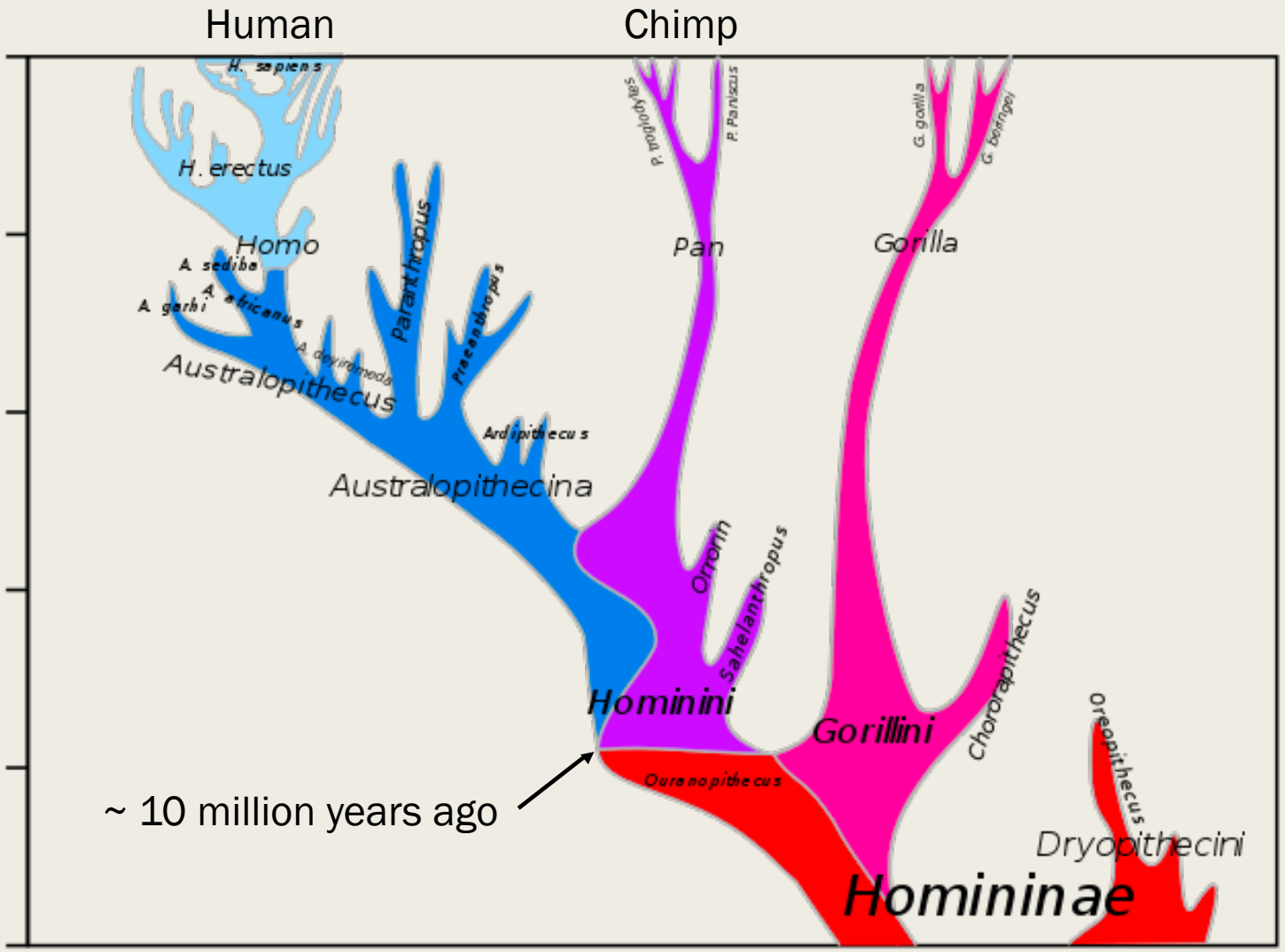


Image: modified from wikipedia

With a partner briefly discuss…

1. How long is the human genome in base pairs?

(A) 3 thousand (B) 3 million **(C) 3 billion**

2. If I compare two human genomes, approximately how often will there be a difference?

(A) Every 10 bases (B) Every 100 bases **(C) Every 1000 bases**

3. How much does it cost to sequence a human genome?

(A) $100 ← important variants **(B) $1,000** ← full sequence (C) $10,000

4. When did humans and chimp last share a common ancestor?

(A) 1 thousand years ago (B) 1 million years ago **(C) 10 million years ago**

5. How long has life on earth been evolving? ~ 4 billion years

Earliest life on earth

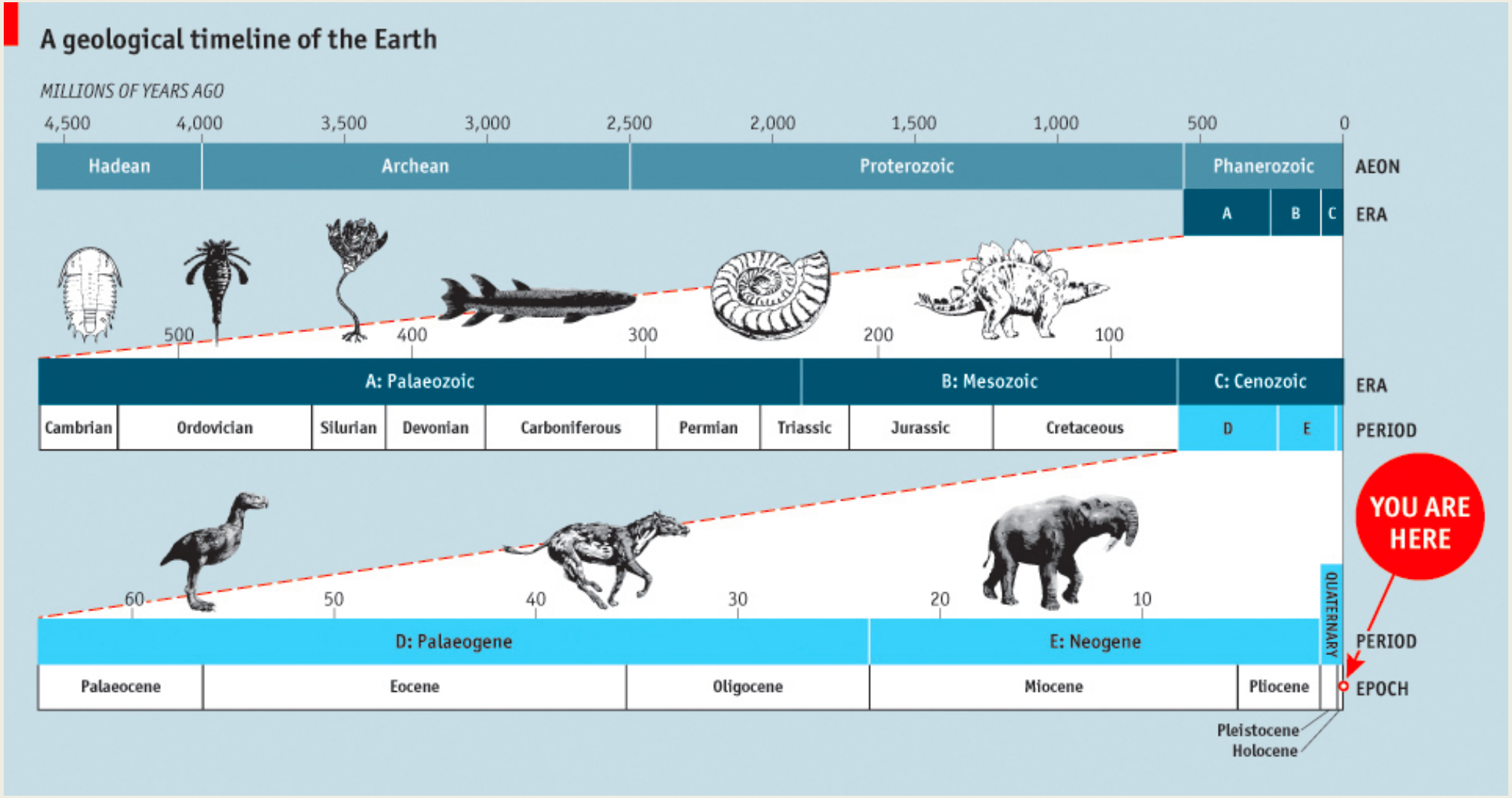


Image: *The Economist*

Syllabus (highlights)

https://www.cs.swarthmore.edu/~smathieson/teaching/s18/

Course Goals

- Handle “real-world” data sets (large and noisy)
- Connect core bioinformatics algorithms to CS
- Understand, implement, and apply core bioinformatics algorithms
- Learn to model uncertainty using probability
- Communicate ideas effectively
- Understand the scientific method (asking a biological question, forming a hypothesis, designing a computational experiment, implementing and applying algorithms, iterating the process, drawing conclusions and communicating the results)
- Develop an appreciation for questions that require interdisciplinary skills to answer

Topics (tentative)

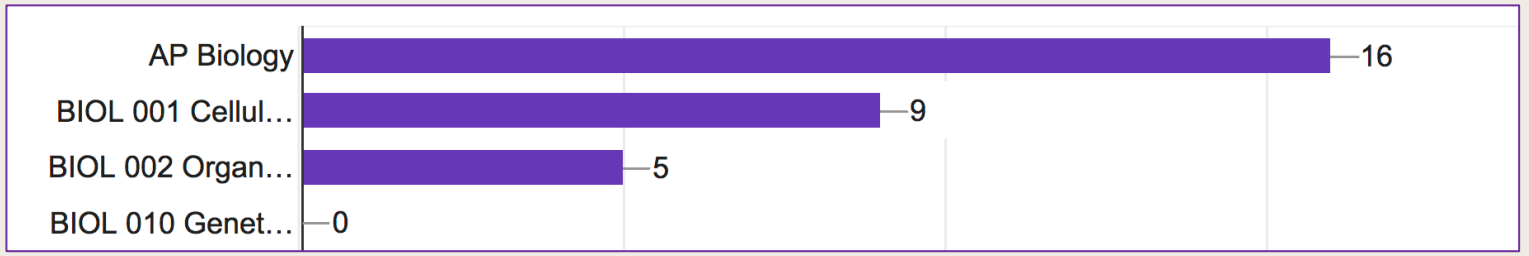
- Sequences and the central dogma
- Genome assembly
- Sequence alignment (dynamic programming)
- Read mapping and Burrows-Wheeler
- Phylogenetic tree algorithms (clustering)
- Ancestral reconstruction
- Population genetics and sequence diversity
- Hidden Markov models (HMMs)
- Deep learning in biology
- Cancer genomics
- RNA folding and non-linear structures
- Ethics and the genome

Prerequisites

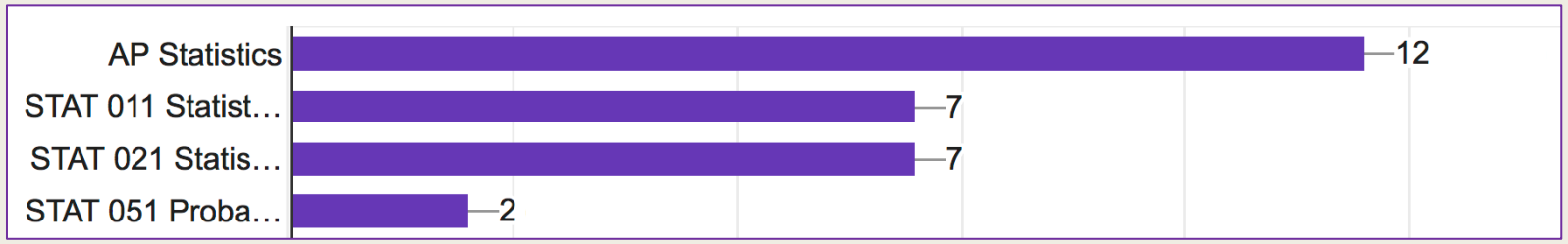
- No biology prerequisite
- CS 21: Introduction to Computer Science
- CS 35: Data Structures
- (helpful but not required) Linear Algebra

Prerequisites

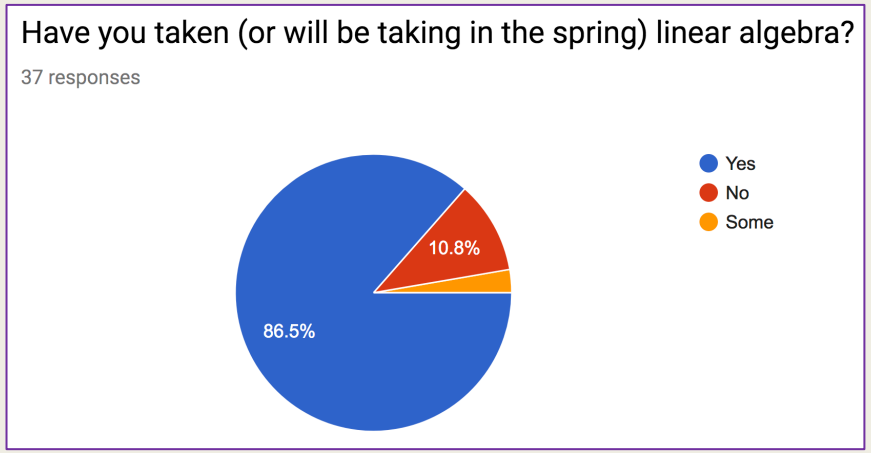
- No biology prerequisite



- CS 21: Introduction to Computer Science
- CS 35: Data Structures



- (helpful but not required) Linear Algebra



Course Components

- Labs (roughly 8-10 total): 35%
- Midterms (2 in-lab, week 6 and week 12): 40% (20% each)
- Final project: 15% (includes an oral presentation and “lab notebook”)
- Participation: 10%

My expectations

- Come to class (M/W/F) and lab (Th), and actively participate during both
- Complete the weekly reading *before* lab
- Come to office hours **(Mon. 3-5pm, Wed. 1-3pm)**
- Post questions on Piazza

WEEK	DAY	ANNOUNCEMENTS	TOPIC & READING	LABS
1	Jan 22		**Introduction to Bioinformatics and Molecular Biology** • Central Dogma of molecular biology • Basics of evolution • History of sequencing • Example applications and goals of computational biology Reading: • (required) **Life and Its Molecules** • (required) **The Central Dogma of Molecular Biology** • (optional) **Primer on Molecular Genetics** • (optional) **Molecular Biology for Computer Scientists**	Mon/Wed/Fri **Lab 1**: Working with sequences
	Jan 24			
	Jan 26			

Syllabus highlights

(Note: you are responsible for reading the entire syllabus on the course webpage)

Syllabus highlights

(Note: you are responsible for reading the entire syllabus on the course webpage)

- Notes and slides will be posted *after* class on the course webpage

Syllabus highlights

(Note: you are responsible for reading the entire syllabus on the course webpage)

- Notes and slides will be posted *after* class on the course webpage
- Durbin *et al* book is required, but could be shared

Syllabus highlights

(Note: you are responsible for reading the entire syllabus on the course webpage)

- Notes and slides will be posted *after* class on the course webpage
- Durbin *et al* book is required, but could be shared
- Lab is **mandatory** (attendance will be taken)

Syllabus highlights

(Note: you are responsible for reading the entire syllabus on the course webpage)

- Notes and slides will be posted *after* class on the course webpage
- Durbin *et al* book is required, but could be shared
- Lab is **mandatory** (attendance will be taken)
- Labs will be done in **pairs** (randomly assigned for the first lab, after that you can choose)

Syllabus highlights

(Note: you are responsible for reading the entire syllabus on the course webpage)

- Notes and slides will be posted *after* class on the course webpage
- Durbin *et al* book is required, but could be shared
- Lab is **mandatory** (attendance will be taken)
- Labs will be done in **pairs** (randomly assigned for the first lab, after that you can choose)
- You will get **2 late days** during the semester (counts for both partners)

Syllabus highlights

(Note: you are responsible for reading the entire syllabus on the course webpage)

- Notes and slides will be posted *after* class on the course webpage
- Durbin *et al* book is required, but could be shared
- Lab is **mandatory** (attendance will be taken)
- Labs will be done in **pairs** (randomly assigned for the first lab, after that you can choose)
- You will get **2 late days** during the semester (counts for both partners)
- Extensions beyond these two days must be arranged with your class dean, **no exceptions**

Syllabus highlights

(Note: you are responsible for reading the entire syllabus on the course webpage)

- Notes and slides will be posted *after* class on the course webpage
- Durbin *et al* book is required, but could be shared
- Lab is **mandatory** (attendance will be taken)
- Labs will be done in **pairs** (randomly assigned for the first lab, after that you can choose)
- You will get **2 late days** during the semester (counts for both partners)
- Extensions beyond these two days must be arranged with your class dean, **no exceptions**
- Email: allow 24 hours for a response (will not respond during evenings/weekends)

Syllabus highlights

(Note: you are responsible for reading the entire syllabus on the course webpage)

- Notes and slides will be posted *after* class on the course webpage
- Durbin *et al* book is required, but could be shared
- Lab is **mandatory** (attendance will be taken)
- Labs will be done in **pairs** (randomly assigned for the first lab, after that you can choose)
- You will get **2 late days** during the semester (counts for both partners)
- Extensions beyond these two days must be arranged with your class dean, **no exceptions**
- Email: allow 24 hours for a response (will not respond during evenings/weekends)
- Piazza: should be used for all content/logistics questions

Participation (10% of course grade)

What counts as participation?

- Asking and answering questions in class (very important!)
  - Raise your hand (just because some people are more/less comfortable shouting out answers)
  - Will occasionally “cold call”, but only after giving you a few minutes to think out loud in pairs

Participation (10% of course grade)

What counts as participation?

- Asking and answering questions in class (very important!)
  - Raise your hand (just because some people are more/less comfortable shouting out answers)
  - Will occasionally “cold call”, but only after giving you a few minutes to think out loud in pairs
- Actively participating in in-class activities (group work, handouts, etc)

Participation (10% of course grade)

What counts as participation?

- Asking and answering questions in class (very important!)
  - Raise your hand (just because some people are more/less comfortable shouting out answers)
  - Will occasionally “cold call”, but only after giving you a few minutes to think out loud in pairs
- Actively participating in in-class activities (group work, handouts, etc)
- Working well with your lab partner during lab
  - Switching “driver” and “navigator” frequently (at least every 30 min)
  - Discussing details instead of just trying to get to the end of the lab

Participation (10% of course grade)

What counts as participation?

- Asking and answering questions in class (very important!)
  - Raise your hand (just because some people are more/less comfortable shouting out answers)
  - Will occasionally “cold call”, but only after giving you a few minutes to think out loud in pairs
- Actively participating in in-class activities (group work, handouts, etc)
- Working well with your lab partner during lab
  - Switching “driver” and “navigator” frequently (at least every 30 min)
  - Discussing details instead of just trying to get to the end of the lab
- Asking and answering questions on Piazza
  - Avoid long blocks of code and giving away answers
  - Only non-anonymous posts count toward participation grade

Participation (10% of course grade)

What counts as participation?

- Asking and answering questions in class (very important!)
  - Raise your hand (just because some people are more/less comfortable shouting out answers)
  - Will occasionally “cold call”, but only after giving you a few minutes to think out loud in pairs
- Actively participating in in-class activities (group work, handouts, etc)
- Working well with your lab partner during lab
  - Switching “driver” and “navigator” frequently (at least every 30 min)
  - Discussing details instead of just trying to get to the end of the lab
- Asking and answering questions on Piazza
  - Avoid long blocks of code and giving away answers
  - Only non-anonymous posts count toward participation grade
- Attending office hours

Participation (10% of course grade)

What counts as participation?

- Asking and answering questions in class (very important!)
  - Raise your hand (just because some people are more/less comfortable shouting out answers)
  - Will occasionally “cold call”, but only after giving you a few minutes to think out loud in pairs
- Actively participating in in-class activities (group work, handouts, etc)
- Working well with your lab partner during lab
  - Switching “driver” and “navigator” frequently (at least every 30 min)
  - Discussing details instead of just trying to get to the end of the lab
- Asking and answering questions on Piazza
  - Avoid long blocks of code and giving away answers
  - Only non-anonymous posts count toward participation grade
- Attending office hours

Sometimes participation goes too far…

- Try to avoid dominating class discussion, office hours, Piazza, pair-programming, etc

Academic Integrity

Discussing ideas and approaches to problems with others on a general level is fine (in fact, we encourage you to discuss general strategies with each other), but you should never read anyone else's code or let anyone else read your code.

- No code from online
- No code from students who took this course previously

Class Deans

Class	Dean	To Schedule an appointment with your dean
First-Year	Dean Karen Henry	Betsy Durning 610-690-5744 edurnin1@swarthmore.edu
Sophomore	Dean Jason Rivera	Stephanie Holznagel (assists with schedule only) 610-690-3999 sholzna1@swarthmore.edu
Junior	Dean Dion Lewis	Bonnie Lytle 610-328-8456 dlytle1@swarthmore.edu
Senior	Dean Nathan Miller	Stephanie Holznagel 610-690-3999 sholzna1@swarthmore.edu

Disability Services

http://www.swarthmore.edu/academic-advising-support/welcome-to-student-disability-service

Registering with the Student Disability Service

Please contact Leslie Hempling, Director of Student Disability Services, at lhempli1@swarthmore.edu or 610-690-5014 to arrange an intake appointment. We are happy to hold initial appointments for incoming students by phone. If at all possible, please submit documentation of your disability in advance so that we can review it prior to talking with you. We recommend that you contact us as early as possible since some accommodations (e.g., electronic books, interpreters, etc.) can take a several weeks to arrange. We want to be sure that your needs are met in time for classes.

Visit the Accommodations Process and the Documentation Guidelines sections in the "For Students" section of this website for all details.

Intro to Molecular Biology and The Central Dogma

Chromosomes

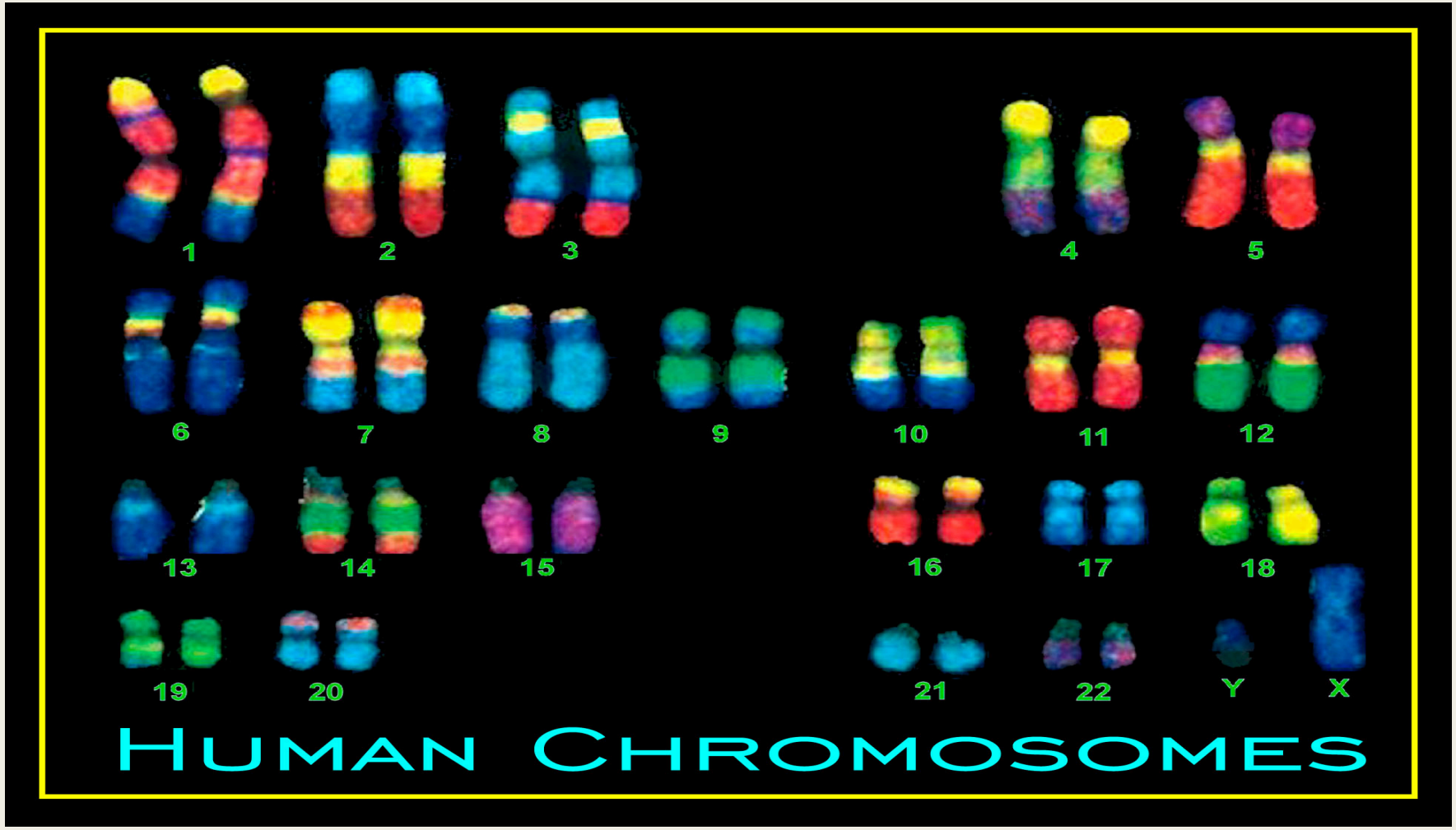
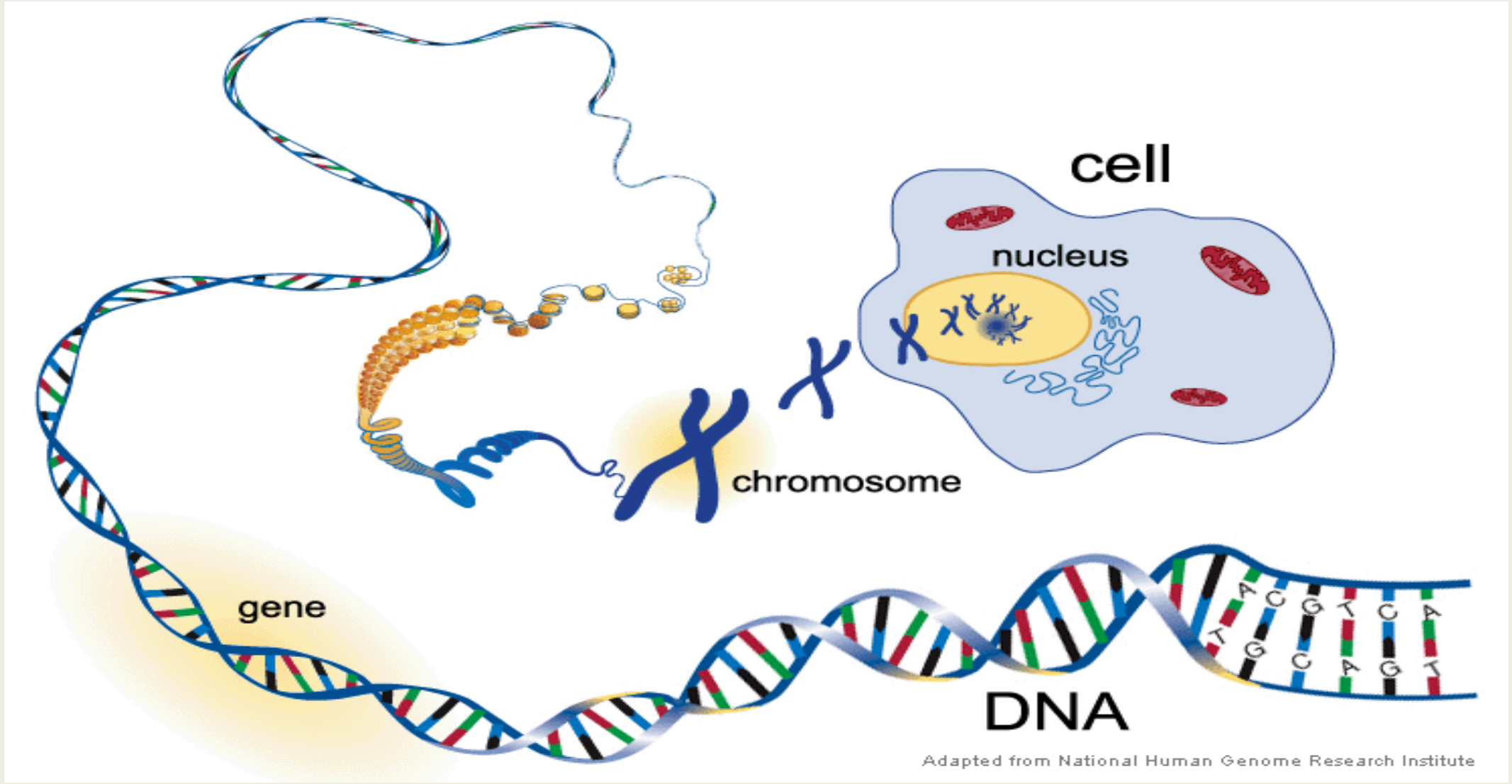


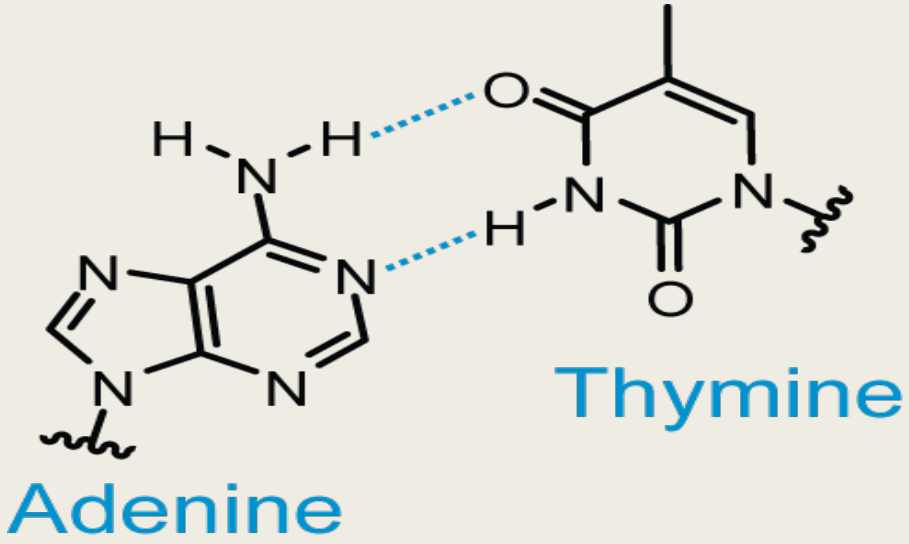
Image: Darwin, Then and Now

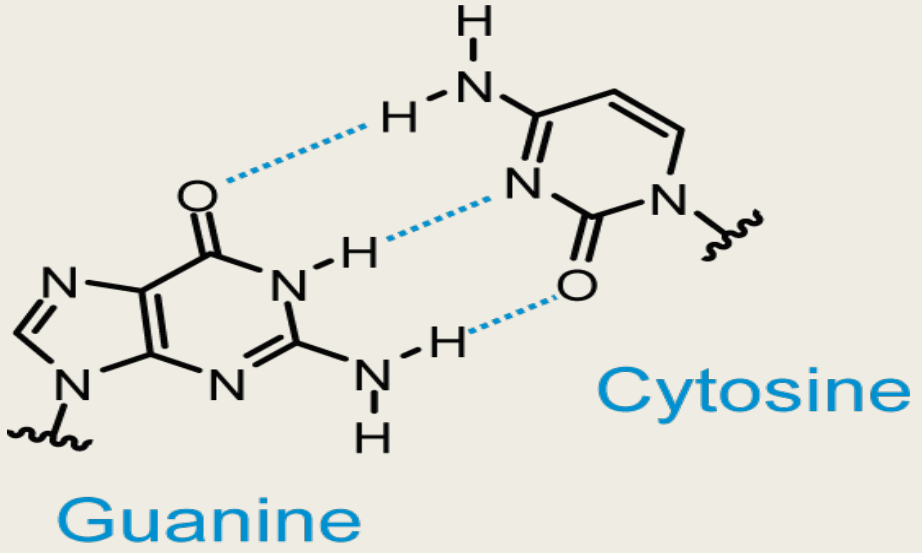
Zooming in on a chromosome



Base-pairing

- “A” with “T”
- “G” with “C”
- Humans: 3 billion base pairs (bp)





Images: MSOE Center for BioMolecular Modeling

Not only humans…

Chimp

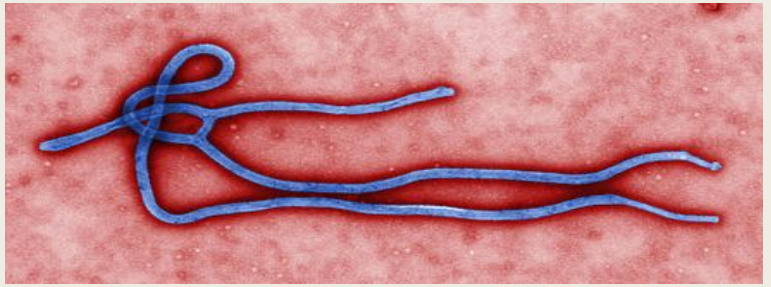
Melitaea cinxia

Buffalo

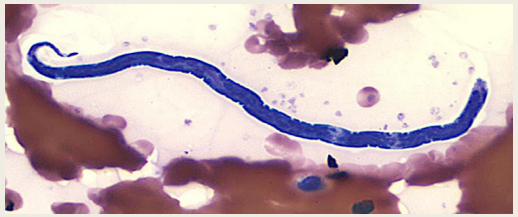
Images: wikipedia

Maize

Chinese liver worm



Ebola



Loa loa (eye worm)

DNA as a string

GCCTAGCTAGGTTACGTACG

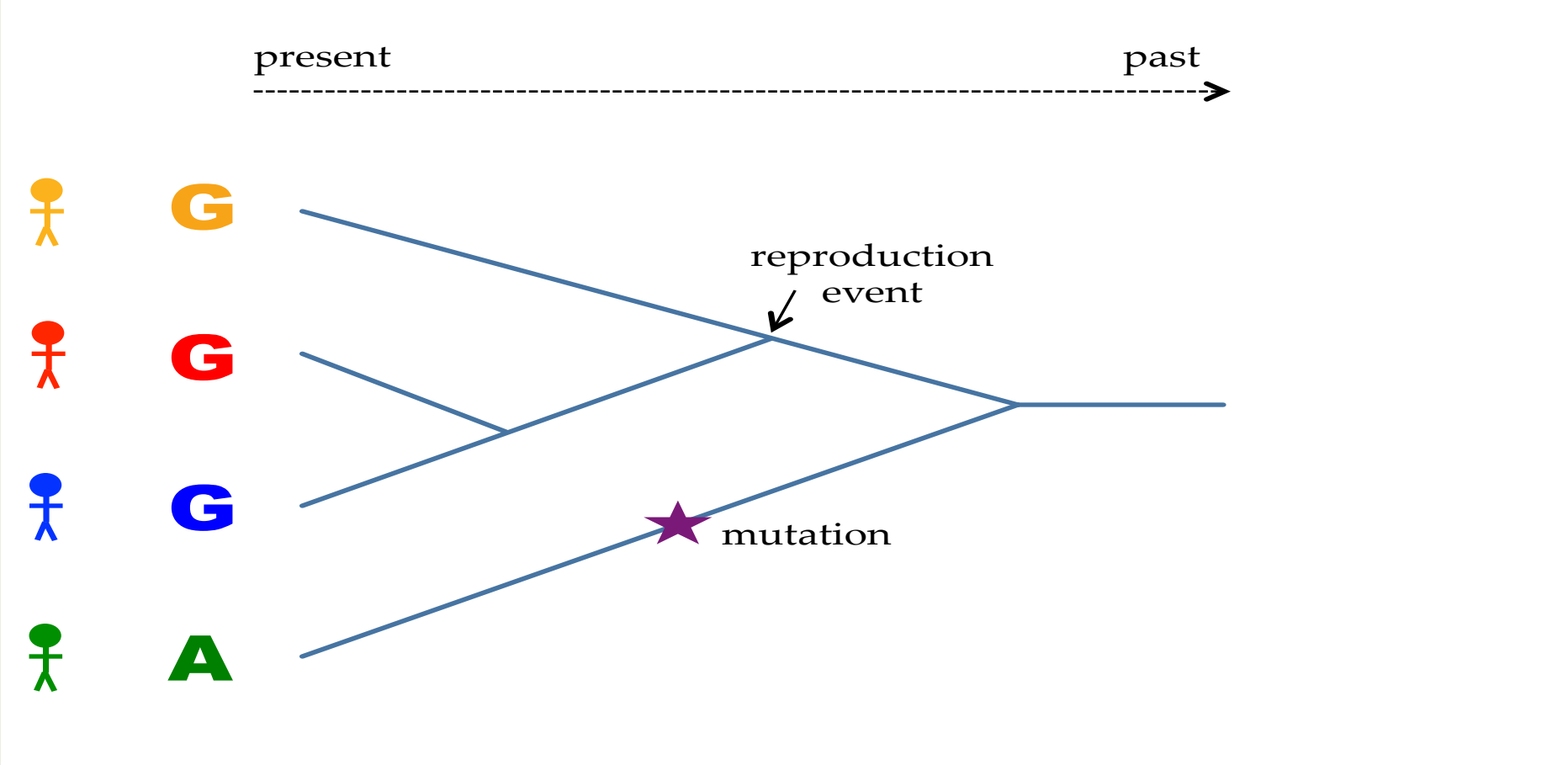
GCCTAGCTAGGTTACGTACG

GCCTAGCTAGGTTACGTACG

ACCTAGCTAGGTTACGTACG

SNP: single-nucleotide polymorphism

Mutation



Central Dogma of Molecular Biology

- More correctly stated: "*The central dogma states that information in nucleic acid can be perpetuated or transferred but the transfer of information into protein is irreversible.*" (B. Lewin, 2004)

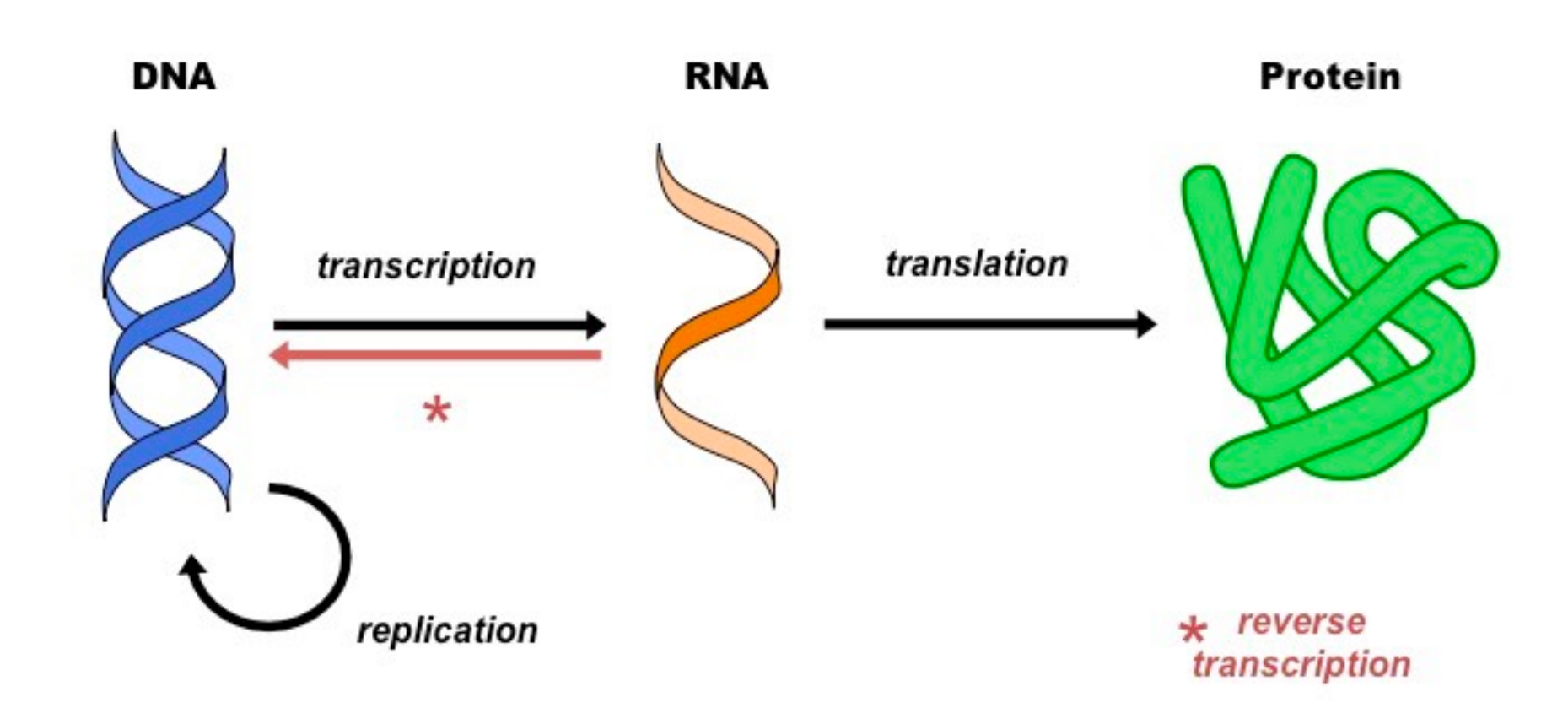


Image: http://ib.bioninja.com.au/

Central Dogma of Molecular Biology

- More correctly stated: "*The central dogma states that information in nucleic acid can be perpetuated or transferred but the transfer of information into protein is irreversible.*" (B. Lewin, 2004)

DNA	ATGCAATCAGATTAG
RNA	CAAUCAGAU
protein	Q S D

GFP protein example