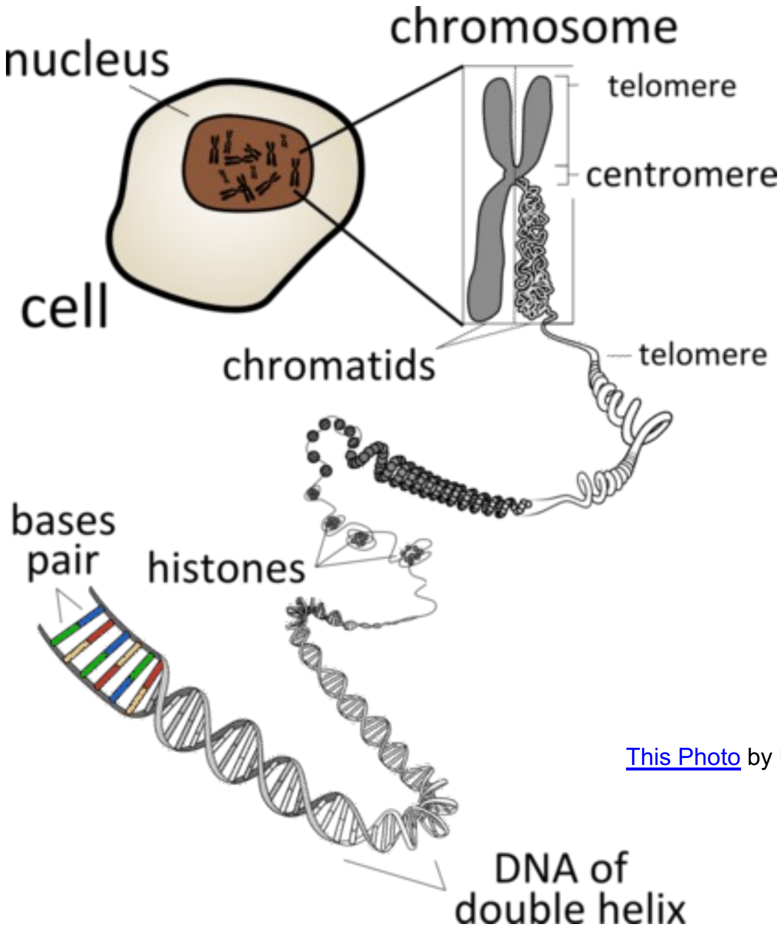
Introduction to Bioinformatics

S. Ravichandran, PhD, PMP
Hood College

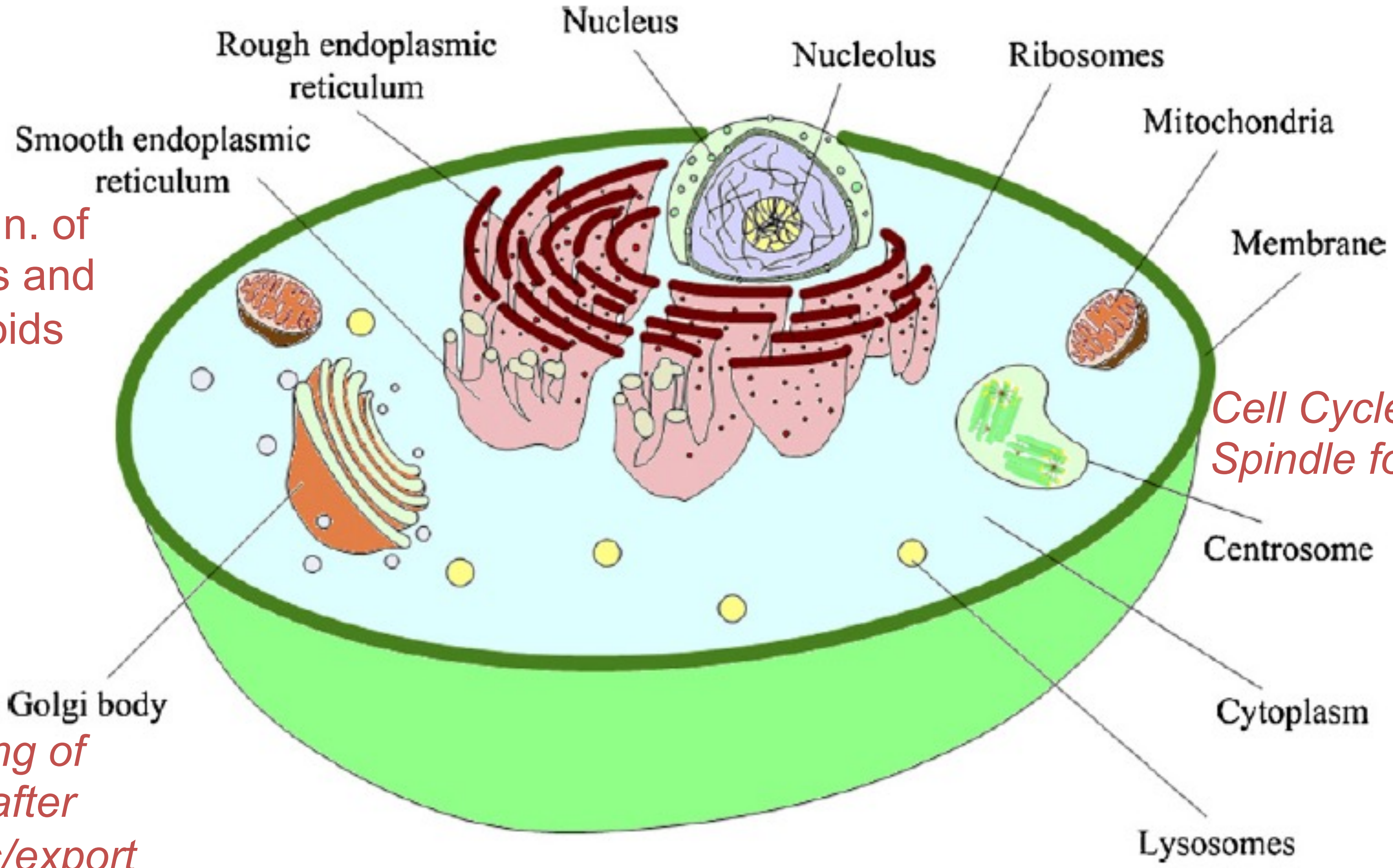
Online servers that require user accounts

- NCBI
  - myNBCI (user account)
  - https://www.ncbi.nlm.nih.gov/
- Ensembl
- UCSC etc.
- Galaxy
  - https://usegalaxy.org/ (user account)
- Create the accounts ahead of time



Human Cell As a Factory

K. Khoshmanesh et al. / Computational Biology and Chemistry 32 (2008) 315–331



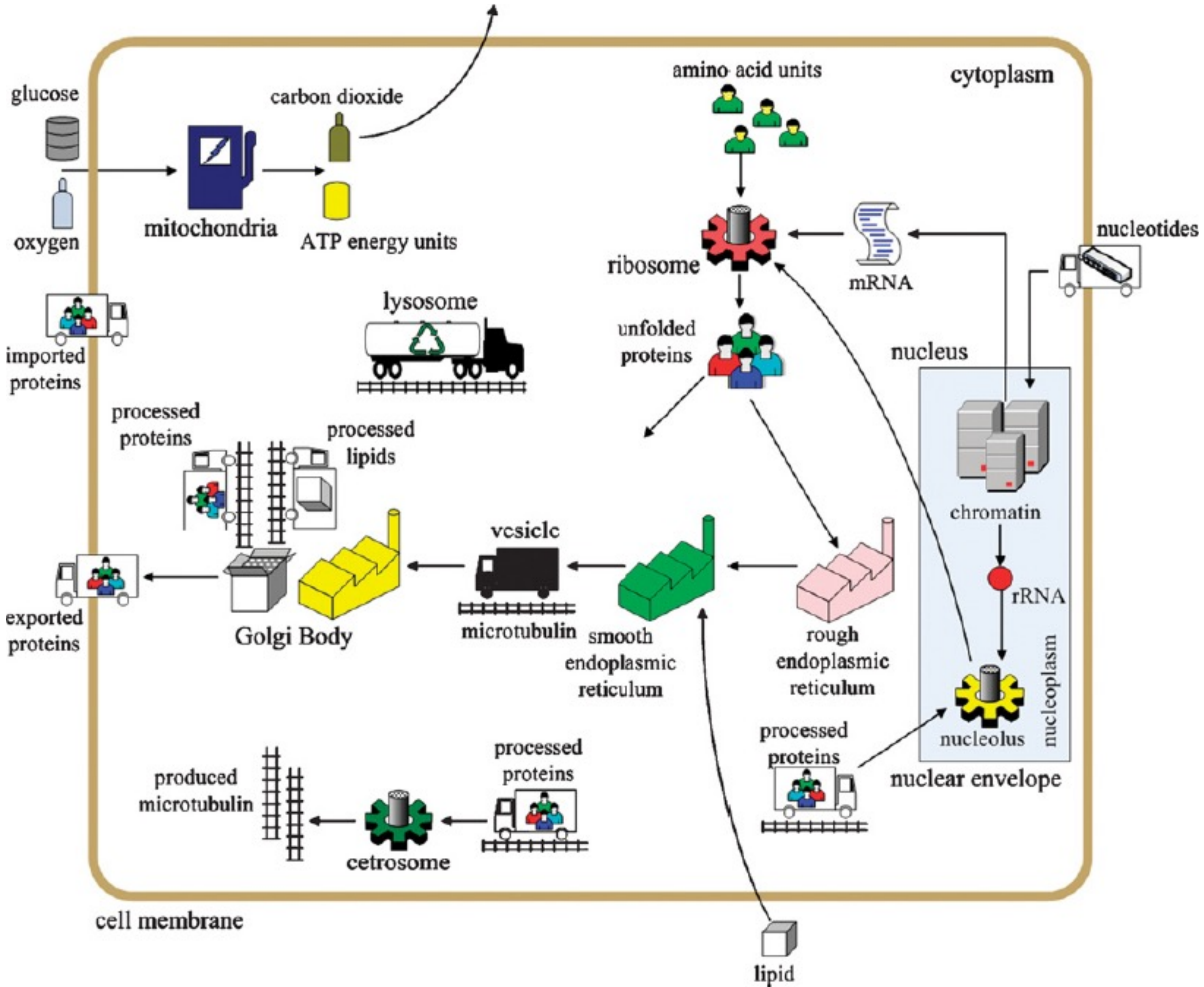
Prodn. of lipids and steroids

Cell Cycle- Spindle formation

Processing of proteins after synthesis/export

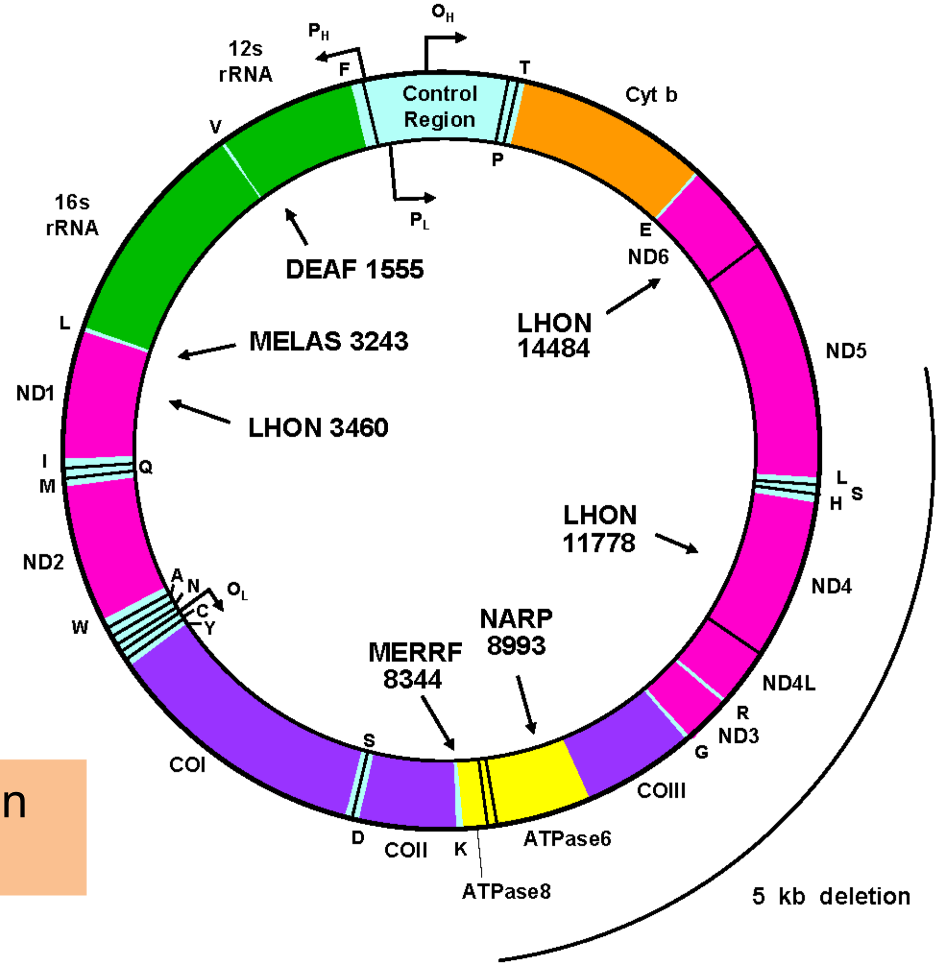
Roughly 37.2 trillion cells in our body

Typical cell (across length) 10 x 10^{-6}m


glucose
oxygen
mitochondria
carbon dioxide
ATP energy units
amino acid units
cytoplasm
ribosome
mRNA
nucleotides
lysosome
imported proteins
unfolded proteins
nucleus
processed proteins
processed lipids
chromatin
vesicle
rRNA
exported proteins
Golgi Body
microtubulin
smooth endoplasmic reticulum
rough endoplasmic reticulum
nucleoplasm
nucleolus
processed proteins
nuclear envelope
processed proteins
produced microtubulin
cetrosome
cell membrane
lipid

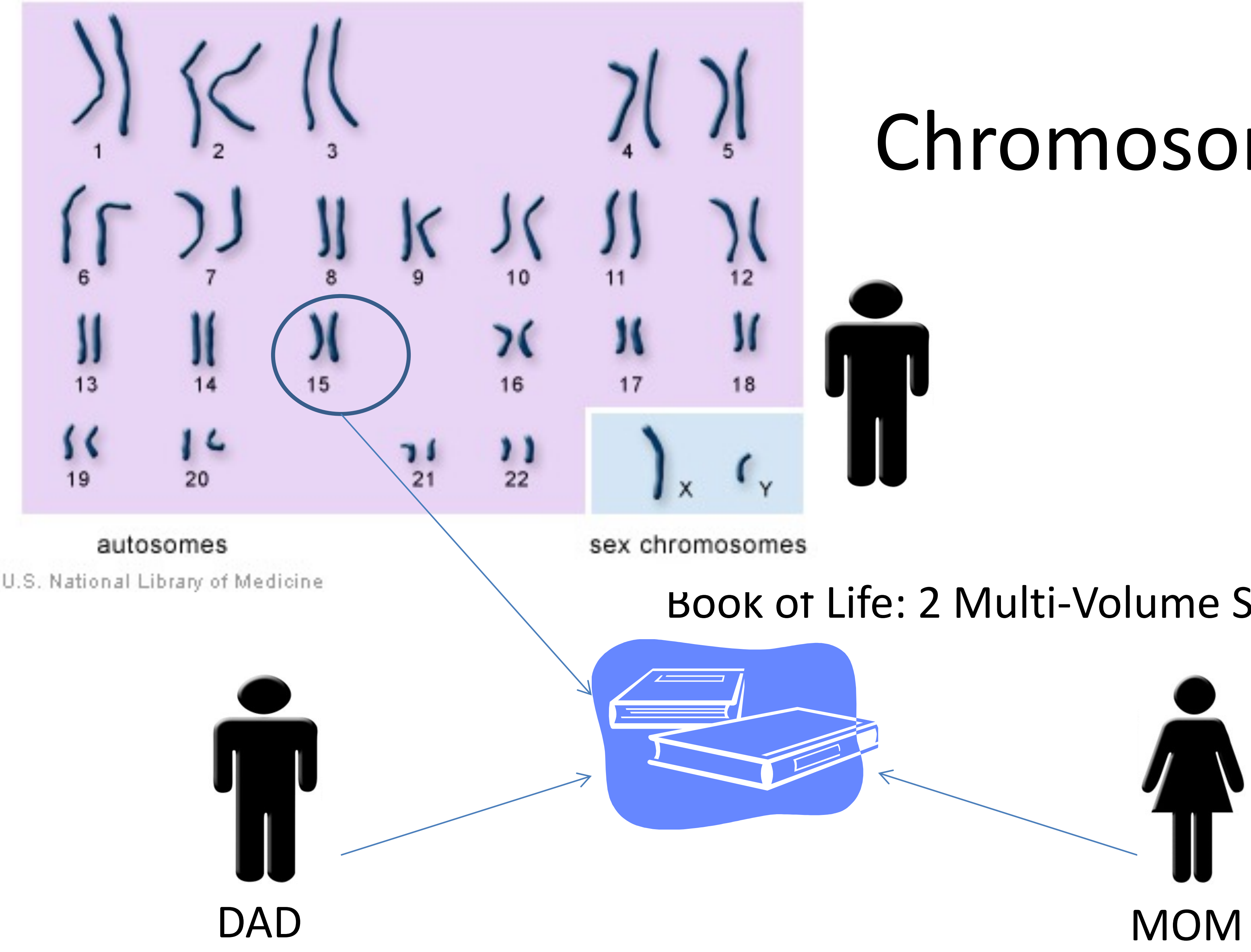
Morbid Map of the Human mtDNA Genome

Inherited from
Mom to children



Fingerprints of Evolution
Circular Genome

Chromosome



Book of Life: 2 Multi-Volume Set

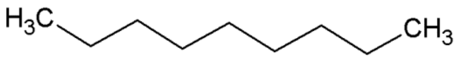
DAD

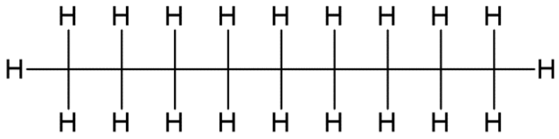
MOM

To understand the language of DNA, we need to understand some Chemistry/Biochemistry

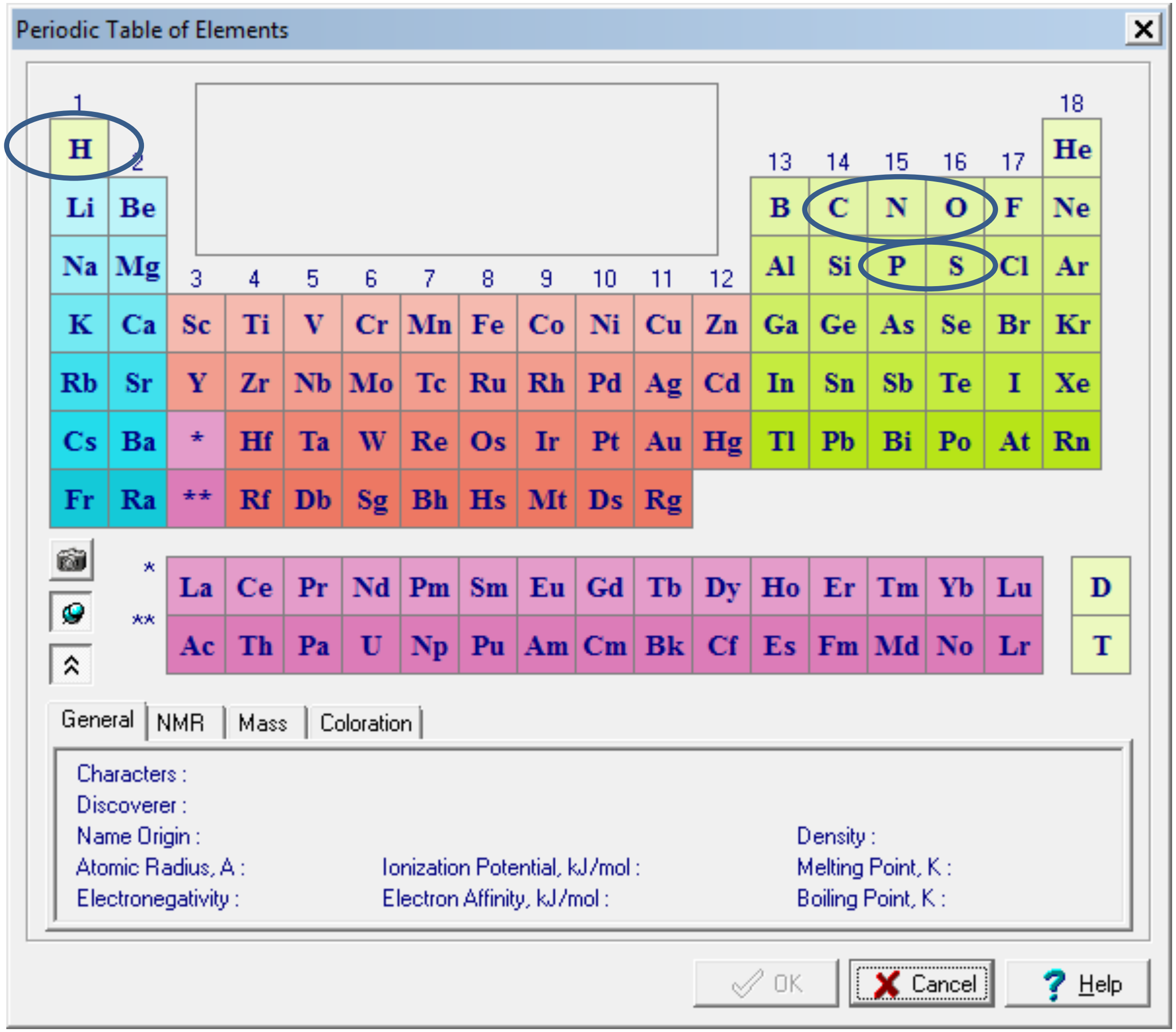
Brief Introduction

Periodic Table





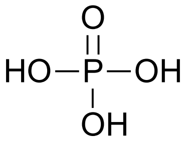
H-O-H



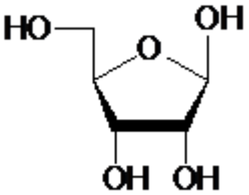
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
H																	He
Li	Be											B	C	N	O	F	Ne
Na	Mg											Al	Si	P	S	Cl	Ar
K	Ca	Sc	Ti	V	Cr	Mn	Fe	Co	Ni	Cu	Zn	Ga	Ge	As	Se	Br	Kr
Rb	Sr	Y	Zr	Nb	Mo	Tc	Ru	Rh	Pd	Ag	Cd	In	Sn	Sb	Te	I	Xe
Cs	Ba	*	Hf	Ta	W	Re	Os	Ir	Pt	Au	Hg	Tl	Pb	Bi	Po	At	Rn
Fr	Ra	**	Rf	Db	Sg	Bh	Hs	Mt	Ds	Rg							

*	La	Ce	Pr	Nd	Pm	Sm	Eu	Gd	Tb	Dy	Ho	Er	Tm	Yb	Lu	D
**	Ac	Th	Pa	U	Np	Pu	Am	Cm	Bk	Cf	Es	Fm	Md	No	Lr	T

Characters :
Discoverer :
Name Origin :
Atomic Radius, A :
Electronegativity :
Ionization Potential, kJ/mol :
Electron Affinity, kJ/mol :
Density :
Melting Point, K :
Boiling Point, K :



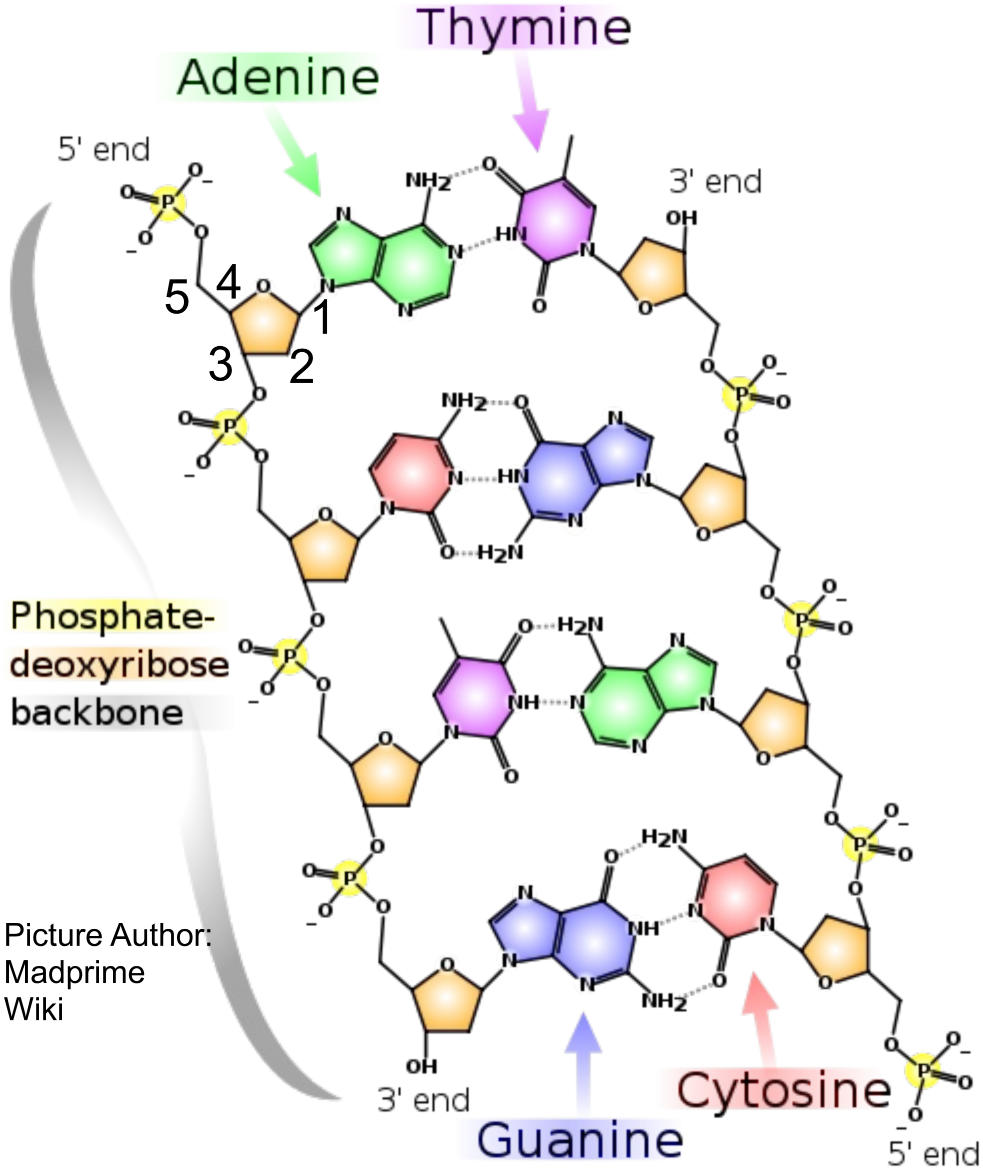
Phosphate



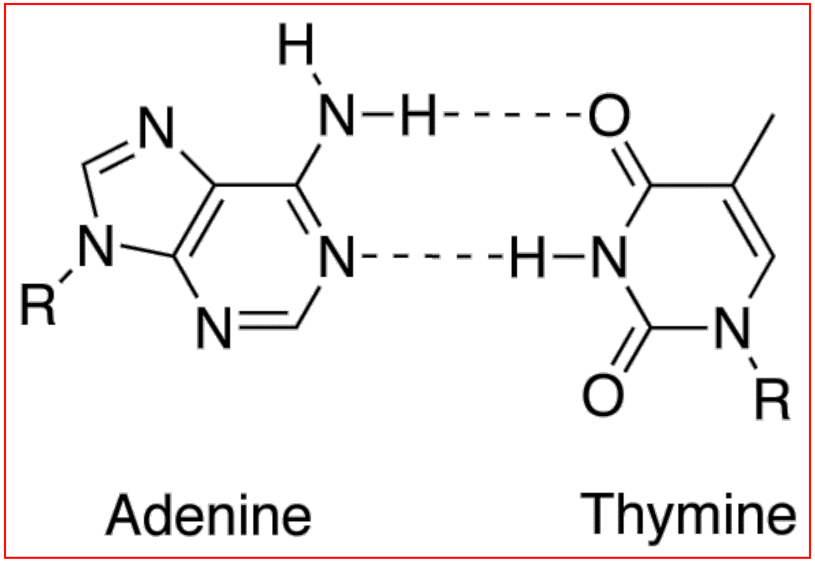
Ribose

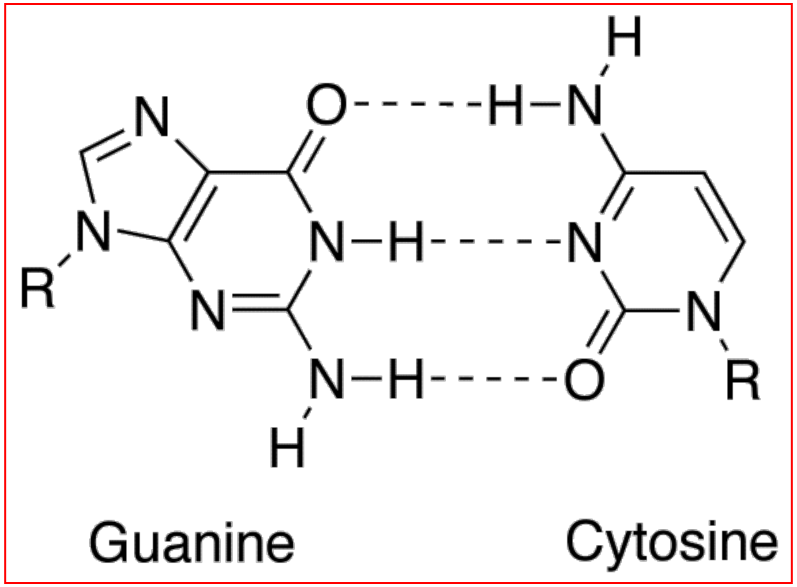
Periodic Table from ChemSketch 2015 (Freeware)

DNA has direction



Picture Author:
Madprime
Wiki





http://en.wikipedia.org/wiki/File:GC_base_pair_jypx3.png

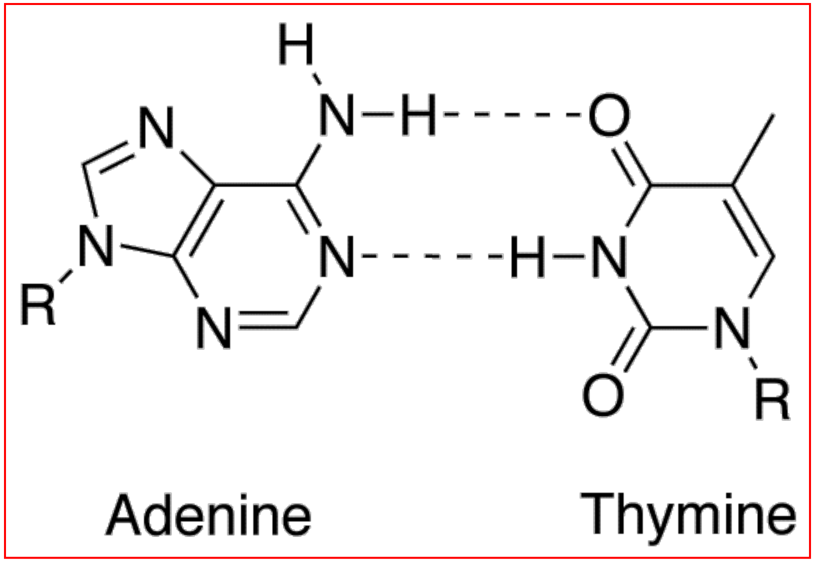
Biology is the chemistry that crawls

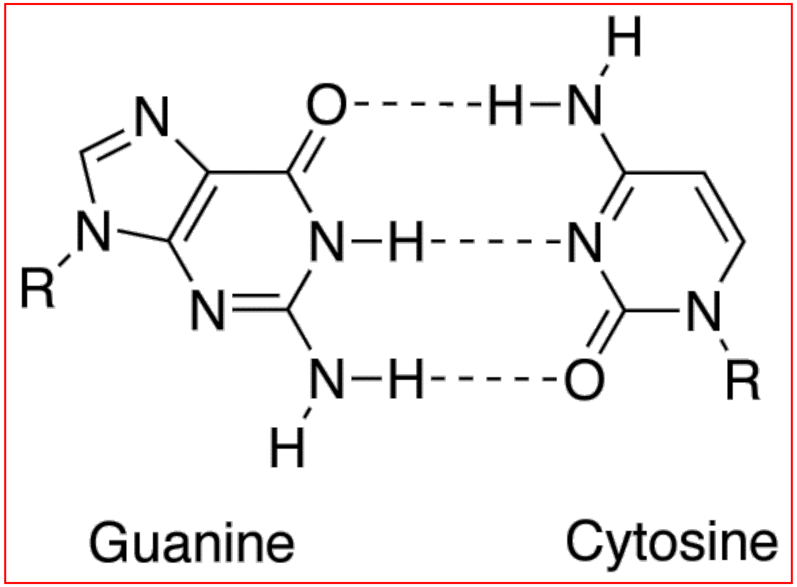
What holds the molecules together?

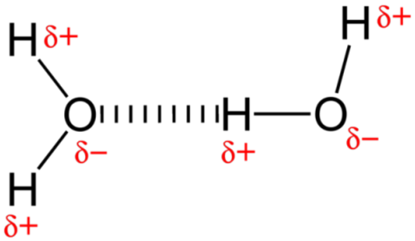
COVALENT BONDS - BY WATERMELON56
WWW.TOONDOO.COM
Covalent bonding can be modeled by kids sharing their toys...
I only need 2 more toys to be happy
We only need 1 more toy each. If we share, we can all be happy!
Now that they all have the needed amount of toys, the kids are happy.
Yay! Now we're all happy!

Atomic level: Chemistry Rules

- Bonded interactions
  - Covalent bonds
- Non-bonded interactions
  - H-bonds
    - Holds DNA
    - Makes drug binding work
  - Ionic, VDW etc.


Adenine
Thymine


Guanine
Cytosine



H-bonds

Which pair is easy to break?

A-T

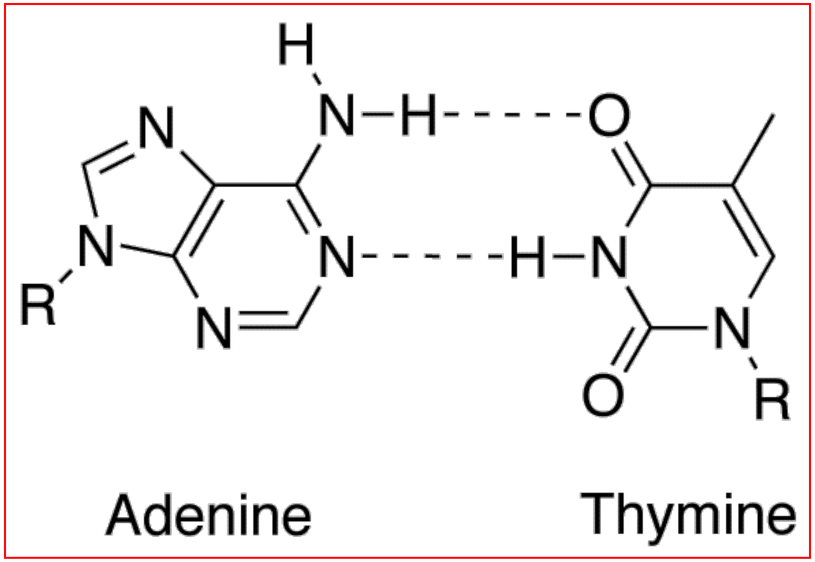
Or

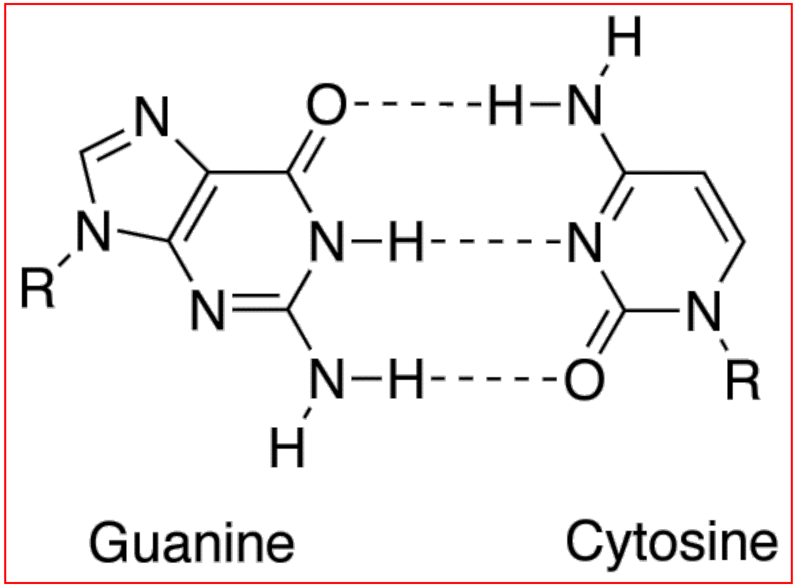
G-C

# of H-bonds

Can we think of A-T being the site for DNA actions such as double-stranded → single-stranded

Replication and Transcription etc.

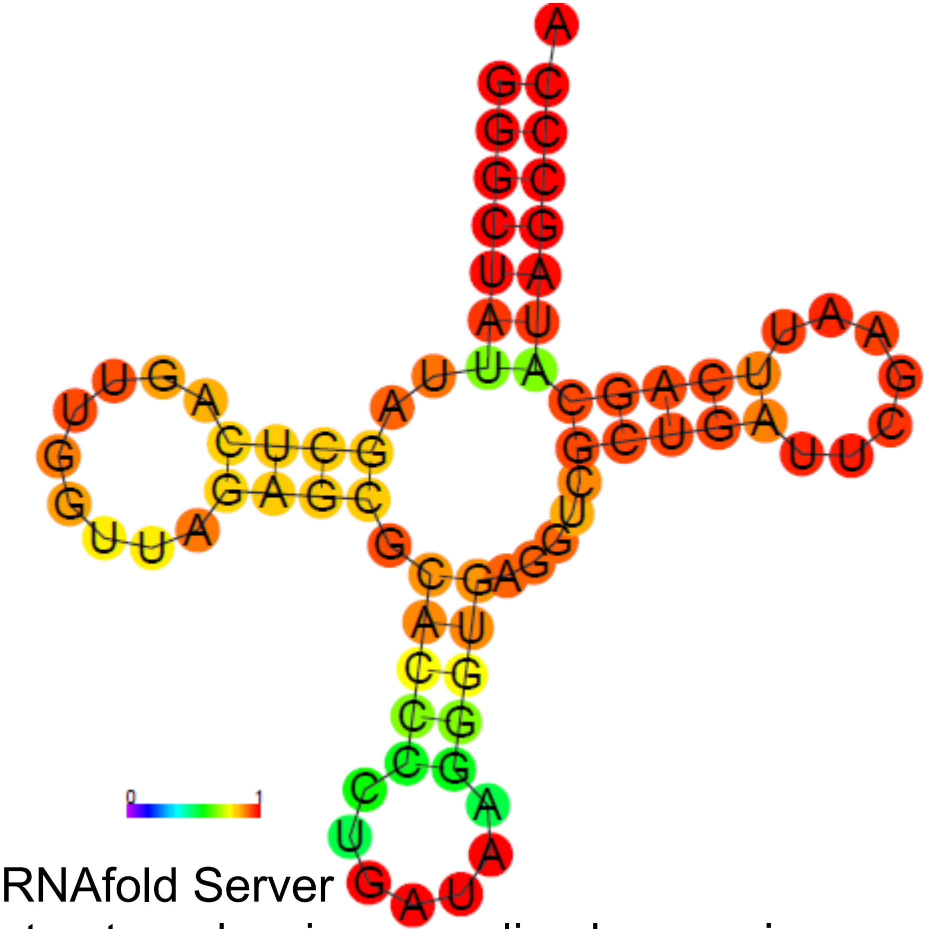




DNA/RNA 3D Structures

CGCGAATTCGGG
GCGCTTAAGCCC

GGGCUAUUAGCUCAGUUGGUUAGAGCGCACCC
CUGAUAAGGGUGAGGUCGCUGAUUCGAAUUC
AGCAUAGCCCA

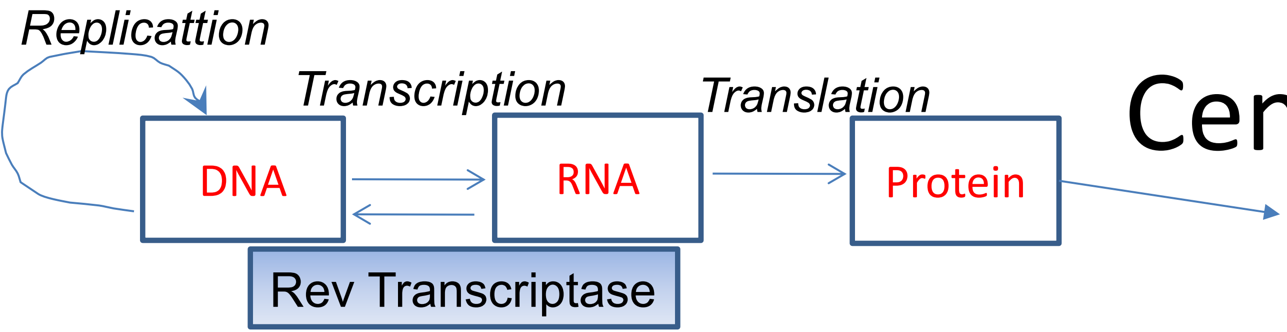


RNAfold Server
structure drawing encoding base-pair probabilities

DNA/RNA

- Different types of DNA
  - B, -Z etc
- Different RNAs
  - mRNA
    - Nucleus → Ribosomes
  - NR (non-coding; 95% of all RNAs)
    - tRNA
    - rRNA

Central Dogma



Replicattion

Transcription

Translation

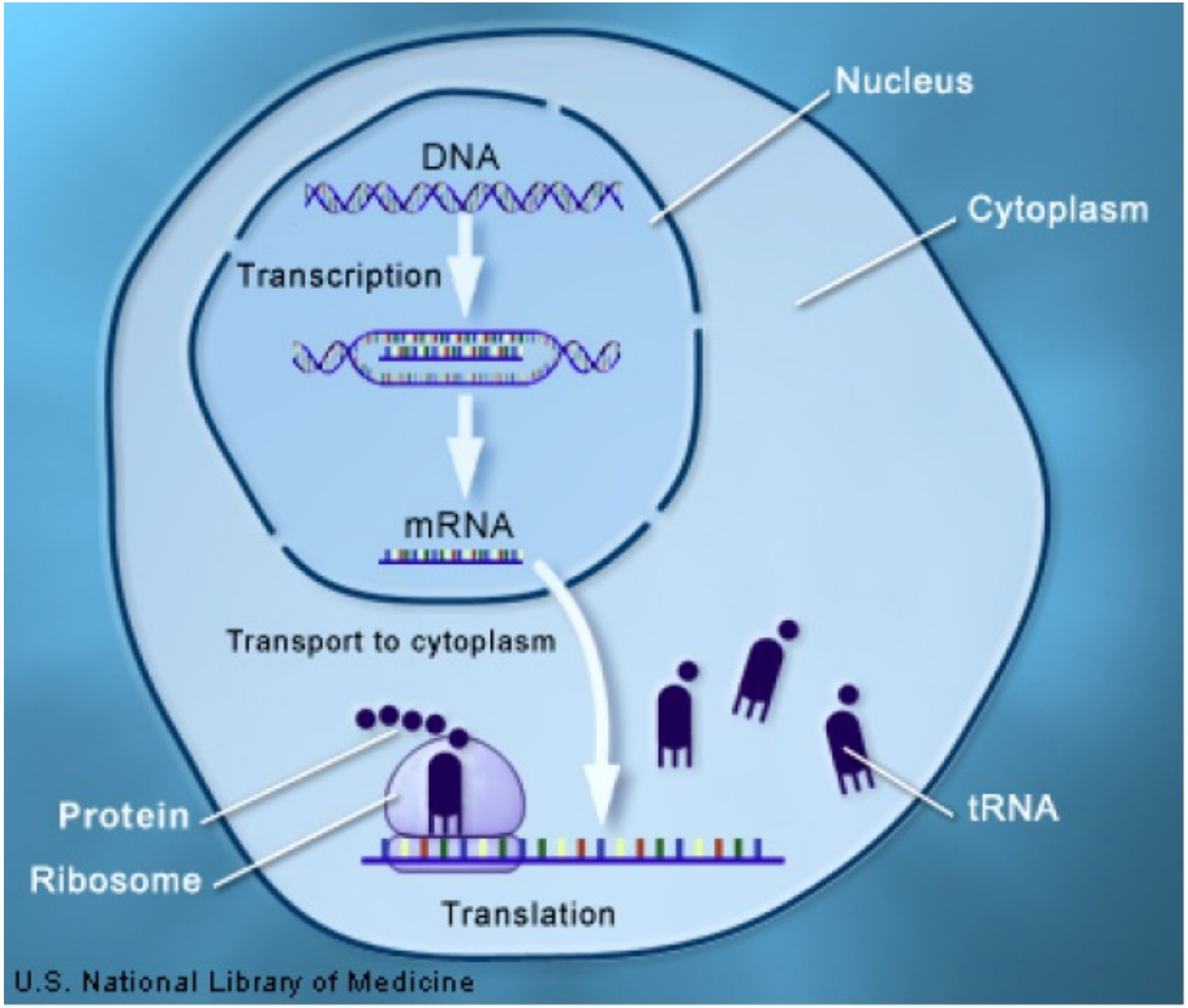
DNA

RNA

Protein

Rev Transcriptase

Cellular Phenotype



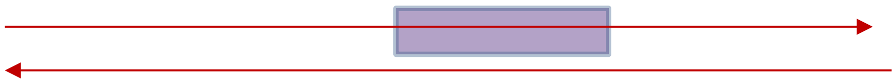
U.S. National Library of Medicine

- pre-mRNA
- 7-methylguanosime placed at 5’ end (prevent RNA degradation)
- Poly A tail is added at the 3’end (200 bps)
- splicing
- Final product, mRNA

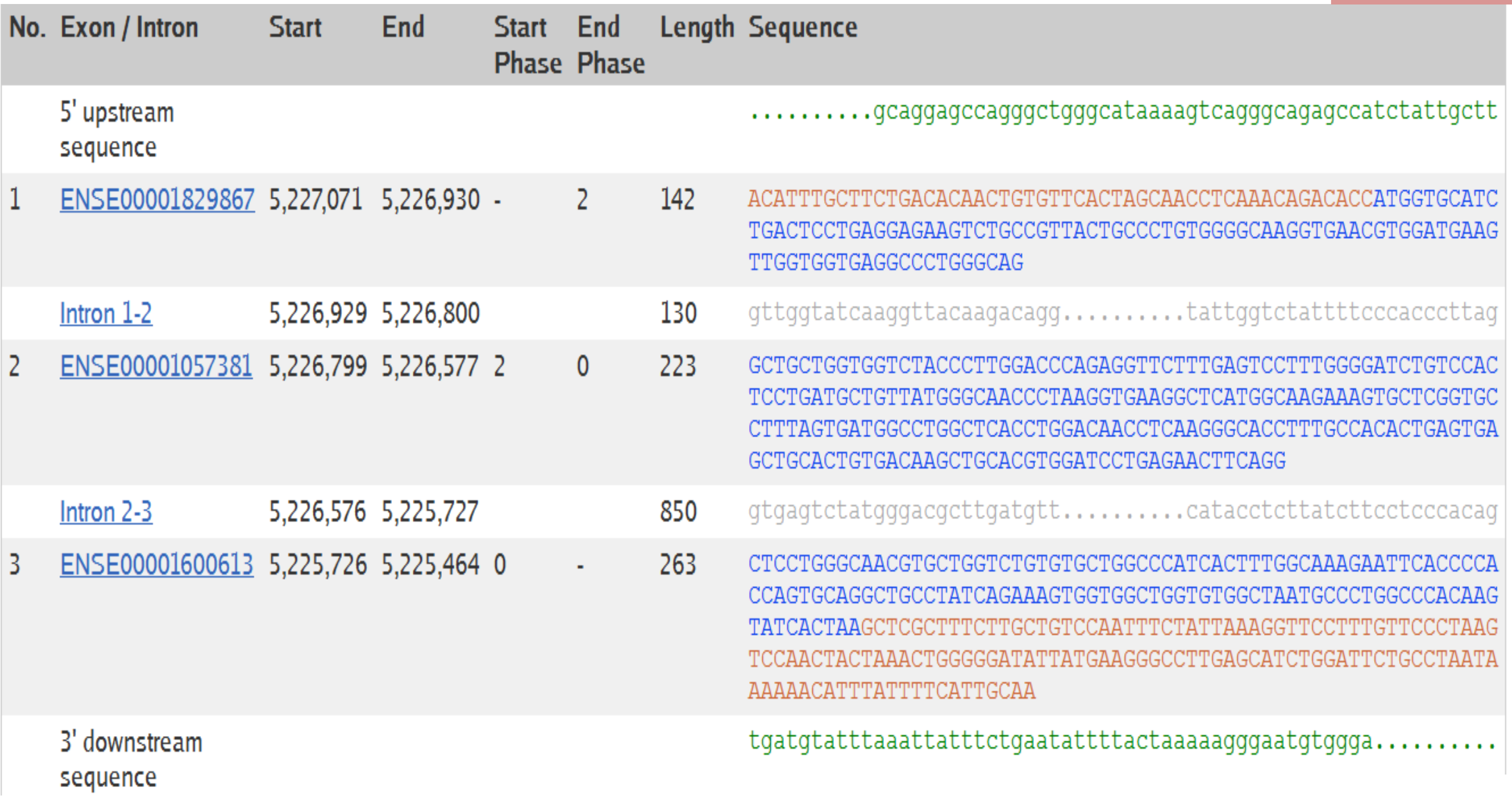
Theory published by **Crick** 1958 (Yes, the same Crick that worked with Watson)

Splicing

- Archaea and Bacteria
  - Usually have one chromosomes
  - Chromosomes are circular
    - some can be linear
- Eukaryotes
  - Multiple chromosomes
  - Linear
  - Packed into cell nucleus



Chromosome (highlight: Gene)

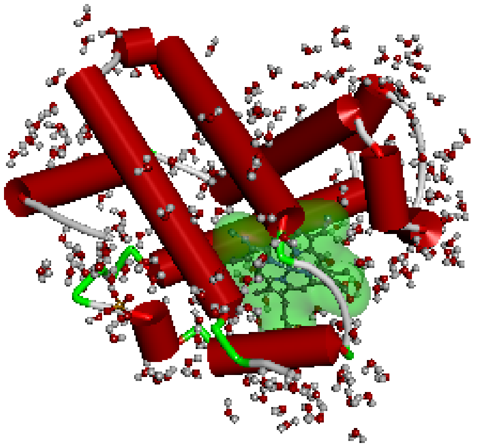


No.	Exon / Intron	Start	End	Start Phase	End Phase	Length	Sequence
	5' upstream sequence						.........gcaggagccagggctgggcataaaagtcagggcagagccatctattgctt
1	ENSE00001829867	5,227,071	5,226,930	-	2	142	ACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCATC TGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAG TTGGTGGTGAGGCCCTGGGCAG
	Intron 1-2	5,226,929	5,226,800			130	gttggtatcaaggttacaagacagg.........tattggtctattttcccacccttag
2	ENSE00001057381	5,226,799	5,226,577	2	0	223	GCTGCTGGTGGTCTACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCAC TCCTGATGCTGTTATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGC CTTTAGTGATGGCCTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGA GCTGCACTGTGACAAGCTGCACGTGGATCCTGAGAACTTCAGG
	Intron 2-3	5,226,576	5,225,727			850	gtgagtctatgggacgcttgatgtt.........catacctcttatcttcctcccacag
3	ENSE00001600613	5,225,726	5,225,464	0	-	263	CTCCTGGGCAACGTGCTGGTCTGTGTGCTGGCCCATCACTTTGGCAAAGAATTCACCCCA CCAGTGCAGGCTGCCTATCAGAAAGTGGTGGCTGGTGTGGCTAATGCCCTGGCCCACAAG TATCACTAAGCTCGCTTTCTTGCTGTCCAATTTCTATTAAAGGTTCCTTTGTTCCCTAAG TCCAACTACTAAACTGGGGGATATTATGAAGGGCCTTGAGCATCTGGATTCTGCCTAATA AAAAACATTTATTTTCATTGCAA
	3' downstream sequence						tgatgtatttaaattatttctgaatattttactaaaaagggaatgtggga.........

Gene sequence

mRNA (cDNA) and protein sequences

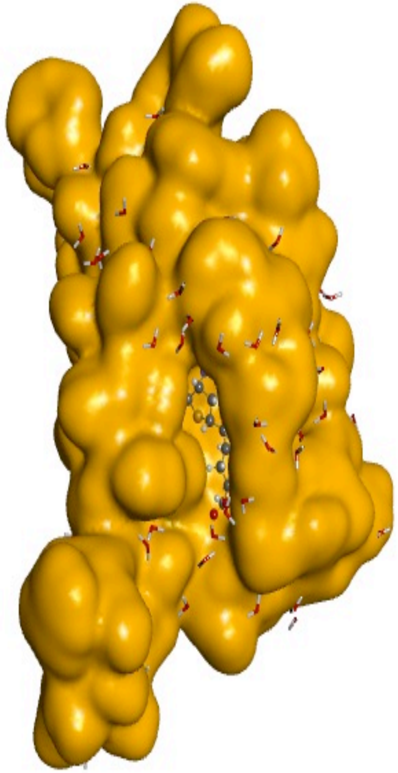
```
 1 ACATTTGCTTCTGACACAACTGTGTTCACTAGCAACCTCAAACAGACACCATGGTGCATC   60
   ..................................................ATGGTGCATC   10
   ..................................................-M--V--H--    3

61 TGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAG  120
11 TGACTCCTGAGGAGAAGTCTGCCGTTACTGCCCTGTGGGGCAAGGTGAACGTGGATGAAG   70
 4 L--T--P--E--E--K--S--A--V--T--A--L--W--G--K--V--N--V--D--E--   23
```



Proteins in 3D

Water molecules

Drug binding inhibition



Vitamin-H bound protein (183 aa protein)

Proteins vs DNA

Proteins

- Unstable
  - Seconds to months
    - Depends on protein & organism
    - <> life span of human proteins: 1 day
  - Destroyed after some time and recycled

DNA

- Stable
- DNA can be stable even for 100,000 years!!!

Related Dogmas

- Central Dogma of Genomics
  - Genome → Transcriptome → Proteome → Cellular Phenotype
- Study of organisms that inhabit the human body
  - microbiome

e-Genome
Genome → Computers

a	0	1	1	0	0	0	1	1

- Ecoli
  - – 4 million bases
    - 4,000,000 * 1 Byte = 4,000,000
      - – ~4 MB Hard drive

- Human
  - ➢ **~ 3,000,000,000** (3 Billion) bases (one copy)
    - ➢Each cell
    - ➢ 3 x 10^9 * 1 Byte = 3 x 10^9 = 3 GB

Learning Bioinformatics

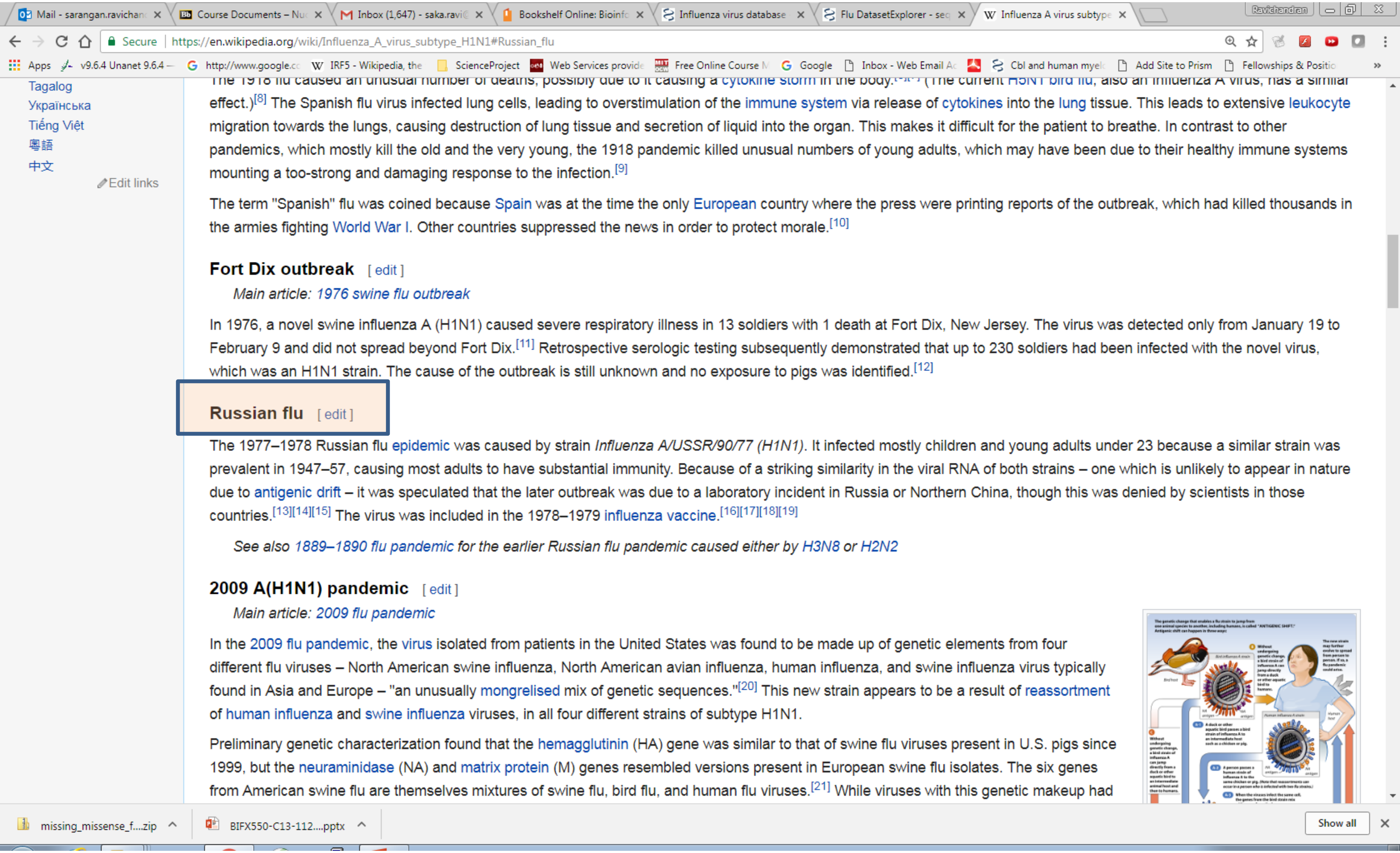
- Software/Browser/OS
  - Version issues
  - Scripts sometimes fail
  - Web connectivity issues
  - NCBI/Ensembl might change their genomic browsers without notice; Genomic browsers behave sometimes differently with different browsers/OS
- Please be patient!

"it is better 100 guilty Persons should escape than that one innocent Person should suffer“ **Benjamin Franklin**

Applications

- Criminal Justice system https://www.innocenceproject.org/dna-exonerations-in-the-united-states/
    - Genetic evidence and exoneration
    - First event happened in 1989
    - Since then many cases had been resolved using DNA exoneration
    - DNA evidence is admitted in criminal trials in almost all states in USA
  - Disease, Therapy(?)
    - Breast cancer: Mutations in BRCA1/2 genes
    - Achondroplasia (Dwarfism): Mutations in FGFR3 gene
- Evolving area
  - Systems’ view is lacking

Applications



The 1918 flu caused an unusual number of deaths, possibly due to it causing a cytokine storm in the body. (The current H5N1 bird flu, also an Influenza A virus, has a similar effect.)[8] The Spanish flu virus infected lung cells, leading to overstimulation of the immune system via release of cytokines into the lung tissue. This leads to extensive leukocyte migration towards the lungs, causing destruction of lung tissue and secretion of liquid into the organ. This makes it difficult for the patient to breathe. In contrast to other pandemics, which mostly kill the old and the very young, the 1918 pandemic killed unusual numbers of young adults, which may have been due to their healthy immune systems mounting a too-strong and damaging response to the infection.[9]

The term "Spanish" flu was coined because Spain was at the time the only European country where the press were printing reports of the outbreak, which had killed thousands in the armies fighting World War I. Other countries suppressed the news in order to protect morale.[10]

Fort Dix outbreak [edit]

Main article: 1976 swine flu outbreak

In 1976, a novel swine influenza A (H1N1) caused severe respiratory illness in 13 soldiers with 1 death at Fort Dix, New Jersey. The virus was detected only from January 19 to February 9 and did not spread beyond Fort Dix.[11] Retrospective serologic testing subsequently demonstrated that up to 230 soldiers had been infected with the novel virus, which was an H1N1 strain. The cause of the outbreak is still unknown and no exposure to pigs was identified.[12]

Russian flu [edit]

The 1977–1978 Russian flu epidemic was caused by strain *Influenza A/USSR/90/77 (H1N1)*. It infected mostly children and young adults under 23 because a similar strain was prevalent in 1947–57, causing most adults to have substantial immunity. Because of a striking similarity in the viral RNA of both strains – one which is unlikely to appear in nature due to antigenic drift – it was speculated that the later outbreak was due to a laboratory incident in Russia or Northern China, though this was denied by scientists in those countries.[13][14][15] The virus was included in the 1978–1979 influenza vaccine.[16][17][18][19]

See also 1889–1890 flu pandemic for the earlier Russian flu pandemic caused either by H3N8 or H2N2

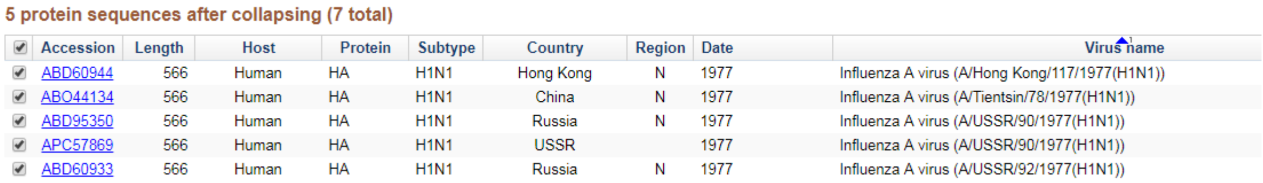
2009 A(H1N1) pandemic [edit]

Main article: 2009 flu pandemic

In the 2009 flu pandemic, the virus isolated from patients in the United States was found to be made up of genetic elements from four different flu viruses – North American swine influenza, North American avian influenza, human influenza, and swine influenza virus typically found in Asia and Europe – "an unusually mongrelised mix of genetic sequences."[20] This new strain appears to be a result of reassortment of human influenza and swine influenza viruses, in all four different strains of subtype H1N1.

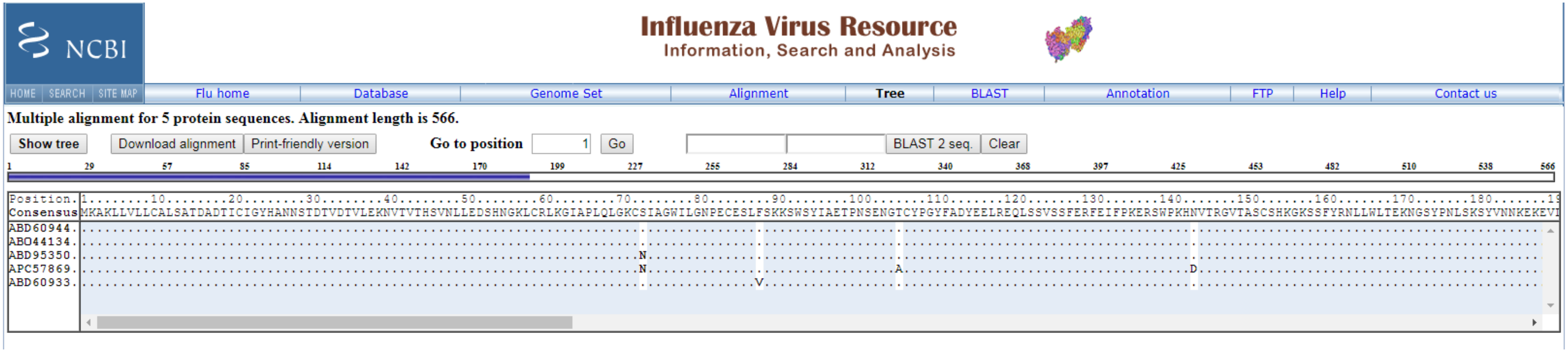
Preliminary genetic characterization found that the hemagglutinin (HA) gene was similar to that of swine flu viruses present in U.S. pigs since 1999, but the neuraminidase (NA) and matrix protein (M) genes resembled versions present in European swine flu isolates. The six genes from American swine flu are themselves mixtures of swine flu, bird flu, and human flu viruses.[21] While viruses with this genetic makeup had

Year 1977
Host Human
Protein HA
Subtype H1N1



5 protein sequences after collapsing (7 total)

☑	Accession	Length	Host	Protein	Subtype	Country	Region	Date	Virus name
☑	ABD60944	566	Human	HA	H1N1	Hong Kong	N	1977	Influenza A virus (A/Hong Kong/117/1977(H1N1))
☑	ABO44134	566	Human	HA	H1N1	China	N	1977	Influenza A virus (A/Tientsin/78/1977(H1N1))
☑	ABD95350	566	Human	HA	H1N1	Russia	N	1977	Influenza A virus (A/USSR/90/1977(H1N1))
☑	APC57869	566	Human	HA	H1N1	USSR		1977	Influenza A virus (A/USSR/90/1977(H1N1))
☑	ABD60933	566	Human	HA	H1N1	Russia	N	1977	Influenza A virus (A/USSR/92/1977(H1N1))

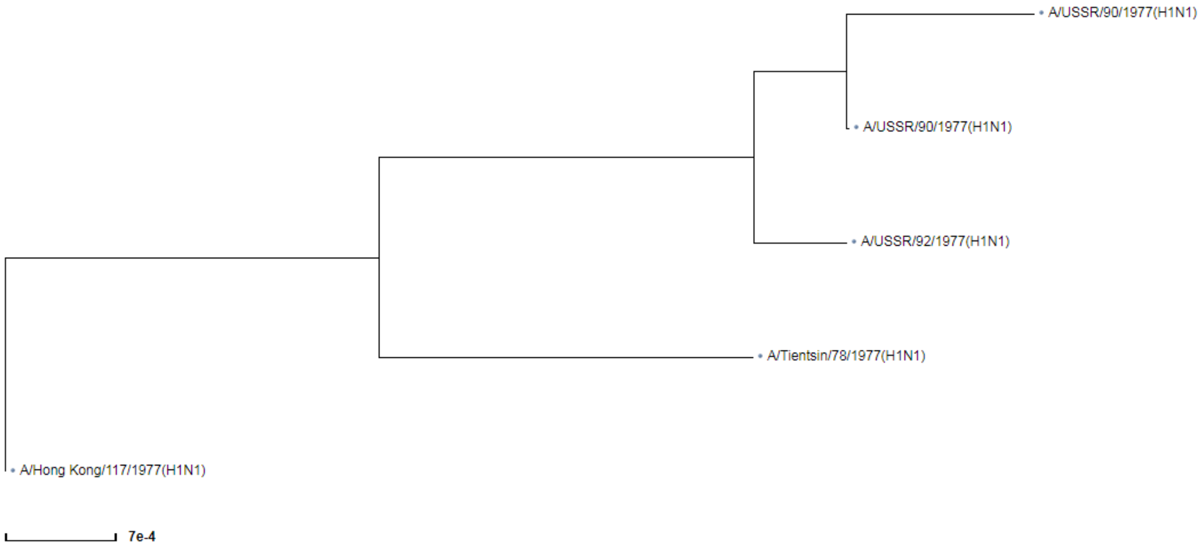


Influenza Virus Resource
Information, Search and Analysis

Multiple alignment for 5 protein sequences. Alignment length is 566.

Go to position 1

```
1    29    57    85    114    142    170    199    227    255    284    312    340    368    397    425    453    482    510    538    566

Position.1........10........20........30........40........50........60........70........80........90........100.......110.......120.......130.......140.......150.......160.......170.......180.......19
ConsensusMKAKLLVLLCALSATDADTICIGYHANNSTDTVDTVLEKNVTVTHSVNLLEDSHNGKLCRLKGIAPLQLGKCSIAGWILGNPECESLFSKKSWSYIAETPNSENGTCYPGYFADYEELREQLSSVSSFERFEIFPKERSWPKHNVTRGVTASCSHKGKSSFYRNLLWLTEKNGSYPNLSKSYVNNKEKEVI
ABD60944.............................................................................................................................................................................................
ABO44134.............................................................................................................................................................................................
ABD95350.........................................................................N...................................................................................................................
APC57869.........................................................................N................................A.....................................D............................................
ABD60933........................................................................................V....................................................................................................
```

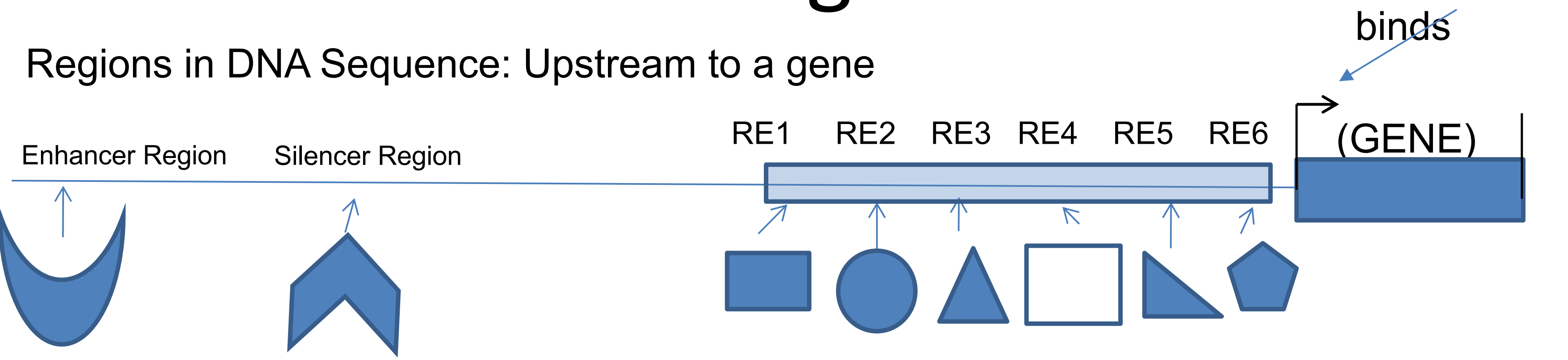


What differentiates different cells?

Gene Regulation

RNA Polymerase binds

Regions in DNA Sequence: Upstream to a gene



Enhancer Region

Silencer Region

RE1 RE2 RE3 RE4 RE5 RE6

(GENE)

Activator Activates gene (Oct-1)

Repressor- bind and block RNA polymerase binding/transcription

Transcription Factors (TF) bind to Regulatory Element (RE)

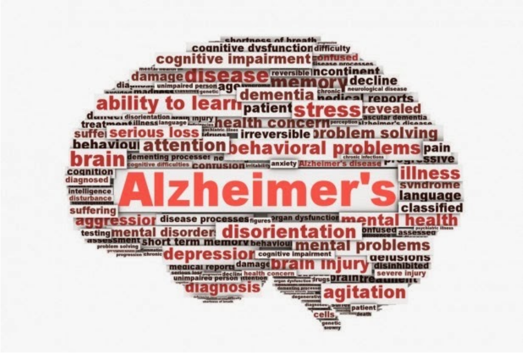
PROMOTER REGION

- Promoters, Transcriptor factor binding regions are identified by experiments.
  - Collecting samples and sequence alignments etc.

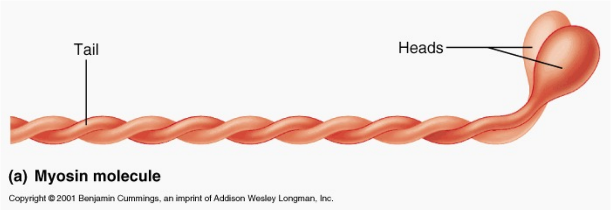
Same content (~25K Genes), different function

Active
NonActive

- **Brain Cells**
  - Amyloid, myosin, α-amylase



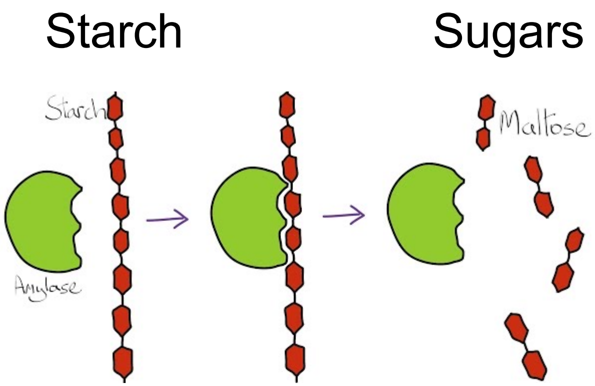
- **Muscle Cells**
  - Amyloid , myosin, α-amylase



(a) Myosin molecule

Copyright © 2001 Benjamin Cummings, an imprint of Addison Wesley Longman, Inc.

- **Salivary Gland Cells**
  - Amyloid, myosin, α-amylase



Data based on NCBI Mol Biology Review Class
Gene Expression

Approximations in Bioinformatics

- 3 → 1 Letter (don’t forget the chemistry)
  - Protein (20 alphabet)
    - Glu-Leu-Val-Ile-Ser-Thr-His-Glu-Lys-Ile-Gln-Gly
    - ELVISTHEKING
  - DNA/RNA (4 letter alphabet)

- Redundant information of DNA
  - storage

ATGGAGCTGTCTTG
TACCTCGACAGAAG

What is the key issue in Bioinformatics?

Release	Date	Bases (GB)	Seq (GB)	Bases(WGS)	Seq (WGS)
235	12/2019	388,417,258,009	215,333,020	6,277,551,200,690	1,127,023,870

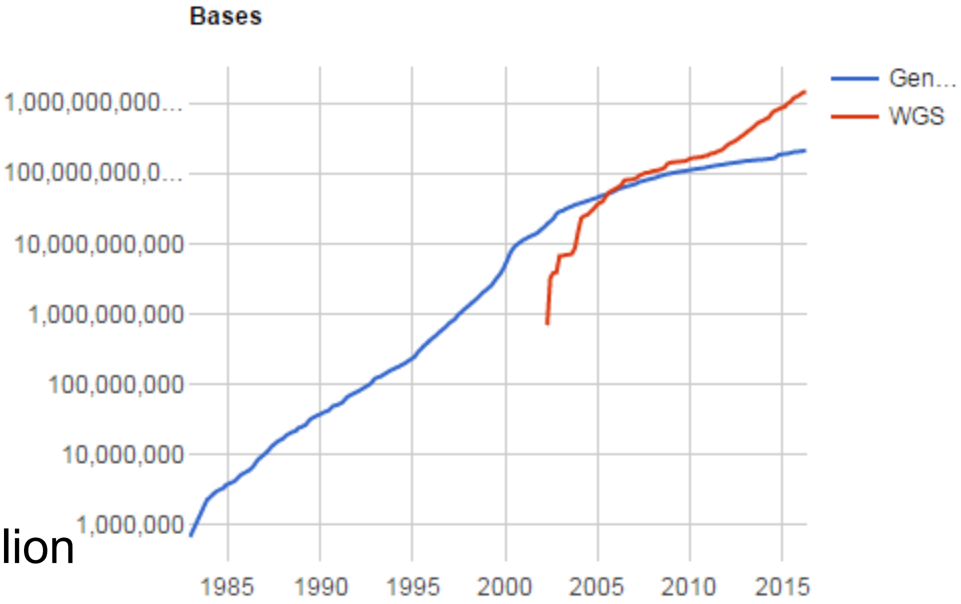
Data Growth in numbers

Enormous data

Need for algorithms to model and carry out analysis

1 Billion

Need for Computers



Storage, Retrieval and Analysis

1 Million

Whole Genome Shotgun (WGS)
High-throughput sequencing (Illumina etc.)

Image from NCBI GenBank

Storage/Retrieval

What can we do with the information?
Can we compare/align them?
What do we learn from the alignments?

Comparing 3D is one way
Not all 3D information is available

So, sequence comparison is the common approach

How to find out whether two proteins are related?

Sequence Alignment

DNA can be compared instead of protein. But, for **most cases** proteins have more information than DNA

- Relatedness (homology) among proteins/DNAs
  - Common function?
  - Homology (common ancestor)
    - When two sequences (proteins/genes) are highly similar, they might be homologous
    - Converse is not true (lack of similarity != No Homology)
  - What is homology?

How can I compare two sequences?

- Not possible without the help of Math and Statistics

- Luckily for us the problem is addressed and some framework is available for us to use

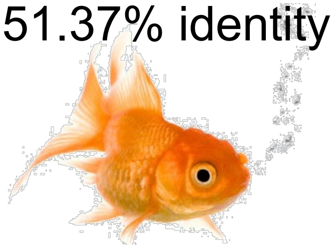
Temple F. Smith

Michael S. Waterman

82.88% identity

```
Human         MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
RT-Lemur      -TFLTPEENGHVTSLWGKVNVEKVGGEALGRLLVVYPWTQRFFESFGDLSSPDAIMGNPK
               ..*****:. **:*******::***************************:***:*****

Human         VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
RT-Lemur      VKAHGKKVLSAFSEGLHHLDNLKGTFAQLSELHCVALHVDPENFKLLGNVLVIVLAHHFG
              *********.***:** ********** ******  ********:******* *******

Human         KEFTPPVQAAYQKVVAGVANALAHKYH
RT-Lemur      NDFSPQTQAAFQKVVTGVANALAHKYH
              ::*:* .***:****:***********
```

51.37% identity



```
Human       MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPK
Goldfish    -VEWTDAERSAIIGLWGKLNPDELGPQALARCLIVYPWTQRYFATFGNLSSPAAIMGNPK
             *. *  *:**: .****:* **:* :**.* *:*******:* :**:**:* *:*****

Human       VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
Goldfish    VAAHGRTVMGGLERAIKNMDNIKATYAPLSVMHSEKLHVDPDNFRLLADCITVCAAMKFG
            * ***:.*:*.:. .: ::**:*.*:* ** :*.:******:*****.: :.   * :**

Human       -KEFTPPVQAAYQKVVAGVANALAHKYH
Goldfish    PSGFNADVQEAWQKFLSVVVSALCRQYH
             . *.   ** *:**.:: *..**.::**
```

21.74% identity

```
Human        MVHLTPEEKSAVTALWGKVNVDEV----GGEALGRLLVVYPWTQRFFESFGDLSTPDAVM
Bloodworm    -MGLSAAQRQVVASTWKDIAGSDNGAGVGKECFTKFLSAHHD---IAAVFGFSG-----A
              : *:  ::..*:: * .:  .:     * *.: ::* .:     :   **  .

Human        GNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLS-----ELHCDKLHVDPENFRLLGNVL
Bloodworm    SDPGVADLGAKVLAQIGVAVSHLGDEGKMVAEMKAVGVRHKGYGYKHIKAEYFEPLGASL
             .:* *   * ***. :. .::**.:    .* :.     .   .  *:. * *. **  *

Human        VCVLAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH-
Bloodworm    LSAMEHRIGGKMTAAAKDAWAAAYADISGALISGLQS
             :..: *::* ::*  .: *:  . *.::.**    :
```

19.85% Identity

```
Human      MVHLTPEEKSAVTALWGKVN--VDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGN
Soybean    MVAFTEKQDALVSSSFEAFKANIPQYSVVFYTSILEKAPAAKDLFSFLA----NGVDPTN
           ** :* ::.: *:: :  .:  : : .      :*   * :: :*. :.     ..   *

Human      PKVKAHGKKVLGAFSDGLAHLDNLKG--TFATLSELHCDKLHVDPENFRLLGNVLVCVLA
Soybean    PKLTGHAEKLFALVRDSAGQLKASGTVVADAALGSVHAQKAVTDPQ-FVVVKEALLKTIK
           **:..*.:*::. . *. .:*.      : *:*..:*.:*  .**: * :: :.*: .:

Human      HHFGKEFT----PPVQAAYQKVVAGVANALAHKYH
Soybean    AAVGDKWSDELSRAWEVAYDELAAAIKKA------
          .*.:::       :.**:::.*.: :*
```

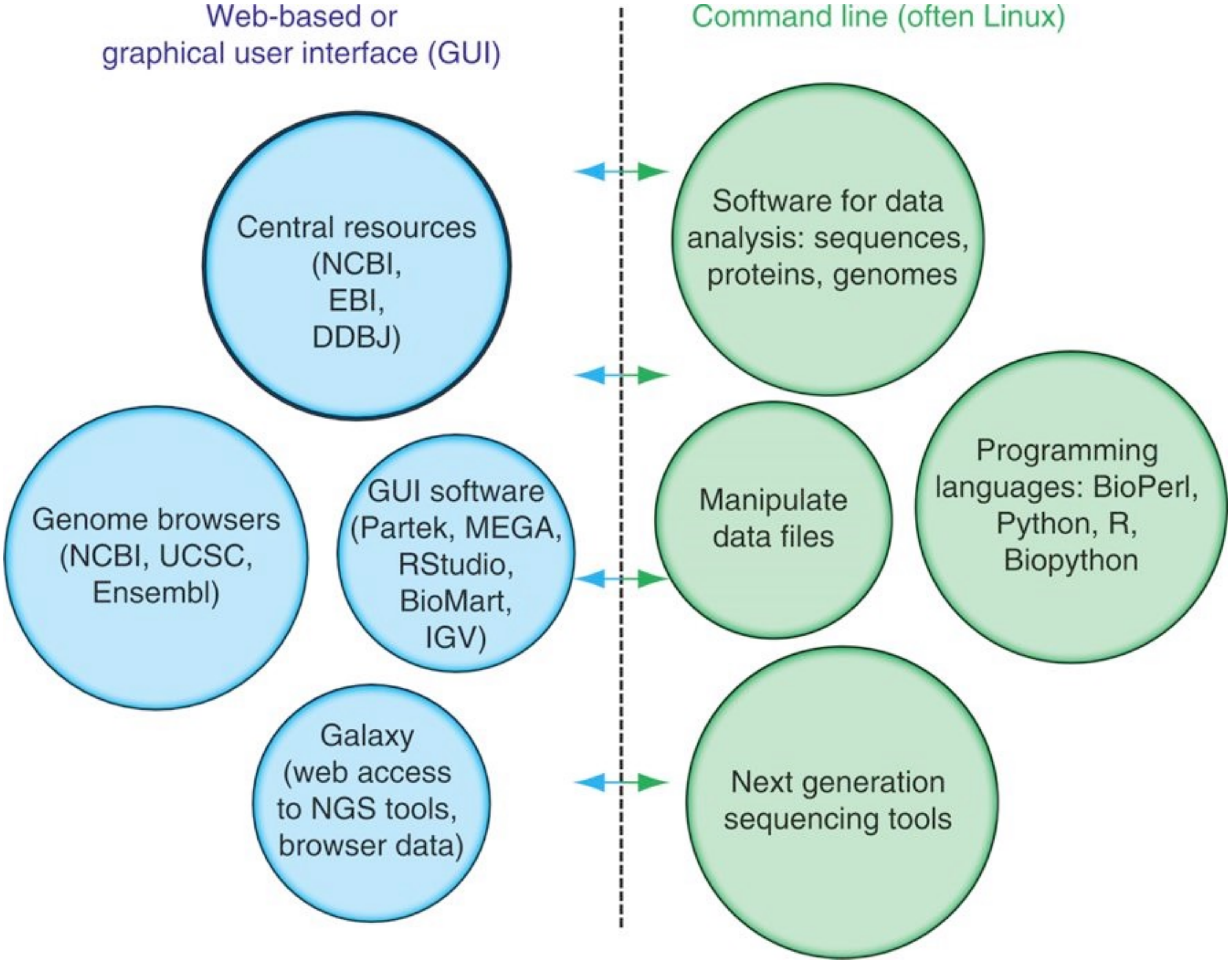
Questions

- What sequences to use and why?
- What types of alignments
  - Global and Local
- Statistics of alignments
  - Scores and matrices

Two Cultures in Bioinformatics

- Two cultures
  - Web-based
    - Point-and-Click (no programming effort)
  - Command line
    - Sometimes steep learning curve (some programming)
- Which one is better?

Fig 1.5 from Bioinformatics and functional genomics / Jonathan Pevsner.— Third edition. Copyright Figure. Please do not distribute.


Web-based or graphical user interface (GUI)
Command line (often Linux)
Central resources (NCBI, EBI, DDBJ)
Software for data analysis: sequences, proteins, genomes
Genome browsers (NCBI, UCSC, Ensembl)
GUI software (Partek, MEGA, RStudio, BioMart, IGV)
Manipulate data files
Programming languages: BioPerl, Python, R, Biopython
Galaxy (web access to NGS tools, browser data)
Next generation sequencing tools

Validity of Predictions

- Can we use a software as a black-box?
- How do we know whether a software method is working properly?
- Each software team will in most cases do self evaluation
- Sensitivity(TPR) and Specificity(TNR)
  - Sensitivity: Detecting true cases
  - Specificity: Excluding those without disease

$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

$$\mathrm{TNR} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}$$

Evaluations

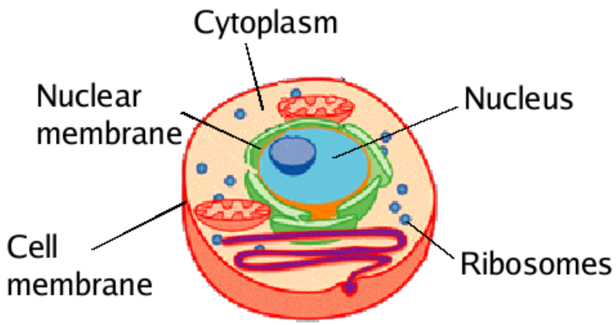
Name	Competition
Alignathon	Compare whole-genome sequence alignment methods
EGASP	ENCODE Genome Annotation Assessment Project
Assemblathon	Compare the performance of genome assemblers
GAGE	Genome Assembly Gold-standard Evaluations
ANRF	Assn. of Biomolecular Resource Facilities (ABRF) assessment of phosphorylation
CASP	Critical Assessment of Structure Prediction
CAFA	Critical Assessment of protein Function Annotation algorithms
CAGI	Critical Assessment of Genome Interpretation. Assess computational methods for predicting phenotypic impacts of genomic variation

New Paradigms for Learning Bioinformatics

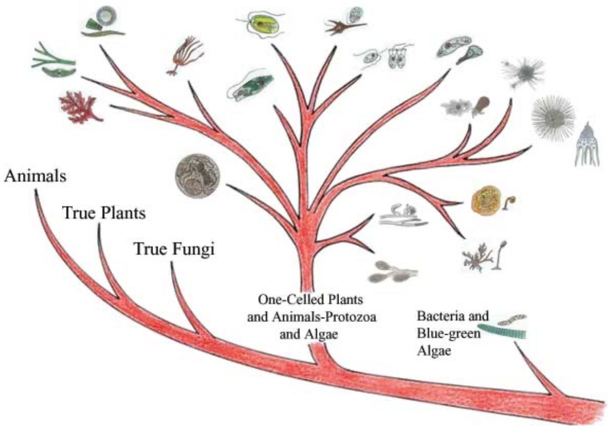
- Online resources
  - Google
- Online Classes
  - MOOC
  - EdX, Stanford, Coursersa, Udacity etc.
- Programming
  - R, Python, Perl
  - Linux Shell scripting

Two perspectives in Bioinformatics

- Cell Perspective
  - Contents of the cell
    - DNA/RNA/Protein
  - Analysis of the sequences



- Organism Perspective
  - How genes are expressed?
    - Age
    - Different tissues/cells
    - Race
    - Disease vs non-disease states


Animals
True Plants
True Fungi
One-Celled Plants and Animals-Protozoa and Algae
Bacteria and Blue-green Algae

Command-line interface

- Command line or point-click or application focused
  - What OS?
    - Windows or Linux
  - Why?
    - Cost, ability to carry out tasks

Reproducibility

“More than 70% of researchers have tried and failed to reproduce another scientist’s experiments, and more than half have failed to reproduce their own experiments”

Is there a reproducibility crisis? M.Baker, Nature, 533, 452, 2016

Reproduce another scientist’s experiments (failed to reproduce their own experiment)
Chemistry: 90% (60%)
Biology: 80% (60%)
Physics & Engineering: 70% (50%)
Medicine: 70% (60%)
Earth and Env. Science: 60% (40%)

Reproducibility in Published Papers

- Script availability
  - Supplemental pages is a good place
  - Useful for checking the results
  - Useful for learning/teaching
  - Useful for reviewers
  - Etc.

Based on ideas from Hadley Wickham

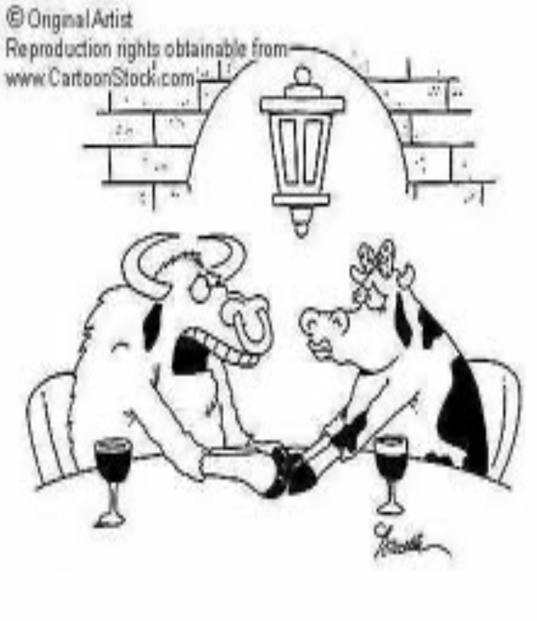
Reproducibility using R

- R session
  - sessionInfo()
- What packages was used
  - library(??)
- Show the code
  - Use R command "dput" to make the user copy and use your code
- Show comments

```r
# Reproducible Code ; S. Ravichandran, Ph.D. 01/23/2017

# load libraries at the top of the script
library(rafalib)

# if not installed
# install.packages("rafalib")
# set seed for reproducibility
set.seed(100)

# create 100 uniform normally distributed random numbers
x <- rnorm(100)
y <- rnorm(100)

# use mypar from rafalib to plot 2 figs in 1 row
mypar(1,2)
hist(x,col="red"); hist(y,col="blue") # use two lines; for lack of space used 1 line
ttest <- t.test(x,y, alternative = "two.sided", conf.level = 0.95) # p-value = 0.94

# add an outlier in y and call it my
m_y <- c(y, 150)
# use dput(data) if you want to send data via email; ex. dput(m_y)

#use two lines; for lack of space used 1 line
hist(x,col="red"); hist(m_y,col="blue")
ottest <- t.test(x, m_y, alternative = "two.sided", conf.level = 0.95) # p-value = 0.94
ottest
sessionInfo() # provide sessionInfo()
```

Application-1
Sequence-based approach

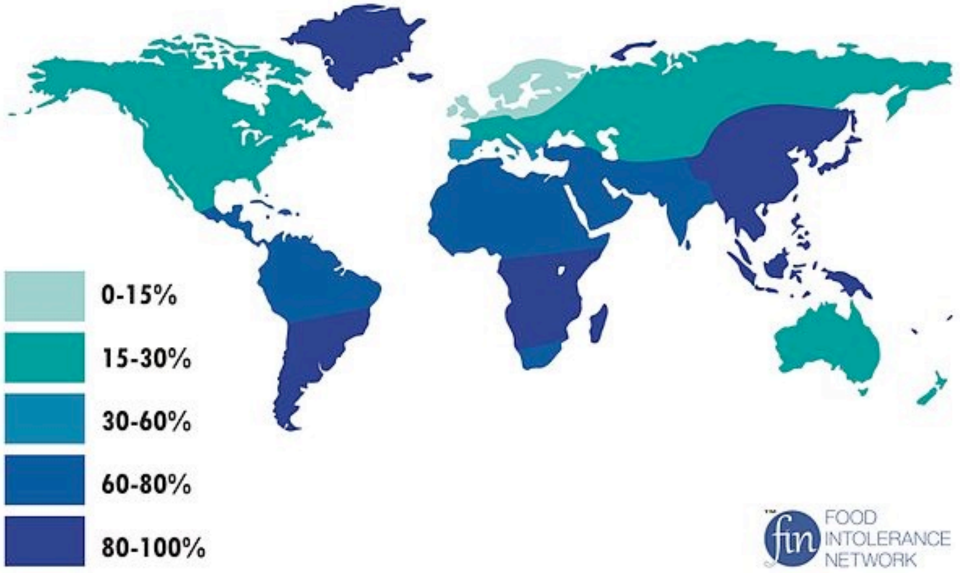
Lactose Intolerance


© Original Artist
Reproduction rights obtainable from
www.CartoonStock.com

"It has nothing to do with you, Bessie. It's just that I'm lactose intolerant."

"It has nothing to do with you, Bessie. It's just that I am Lactose intolerant"

Worldwide prevalence of lactose intolerance in recent populations (schematic)



FOOD INTOLERANCE NETWORK

Finland: 1/60K inborns have LCT intolerance

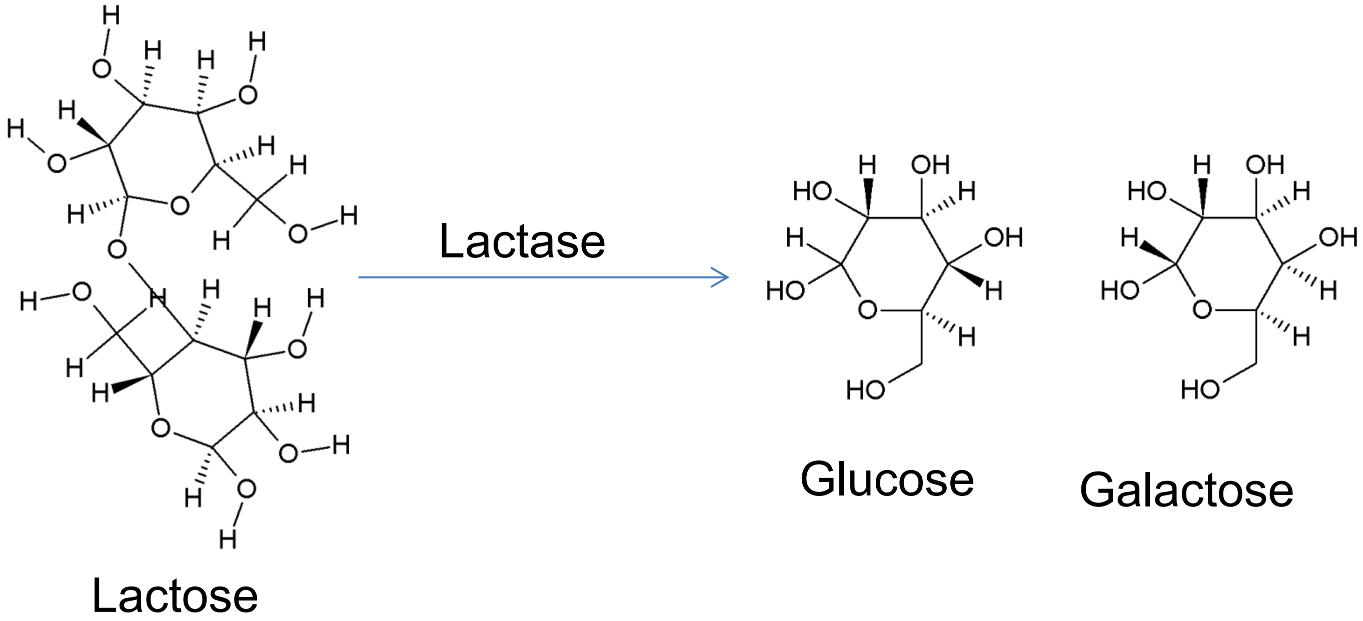
Very common in people of
-West African,
-Arabs,
-Jewish,
-Greek and
-Italian descent.

(ghr.nlm.nih.gov)

Lactose intolerance

- Lactase is the gene that produces Lactase (protein enzyme)

- Lactase digests Lactose → simple sugars

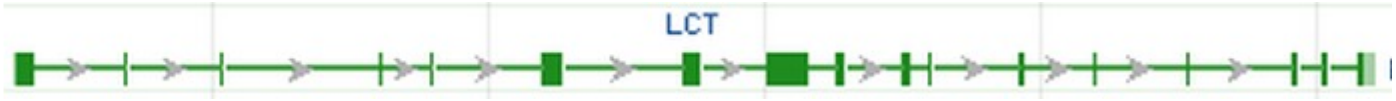


LCT Gene

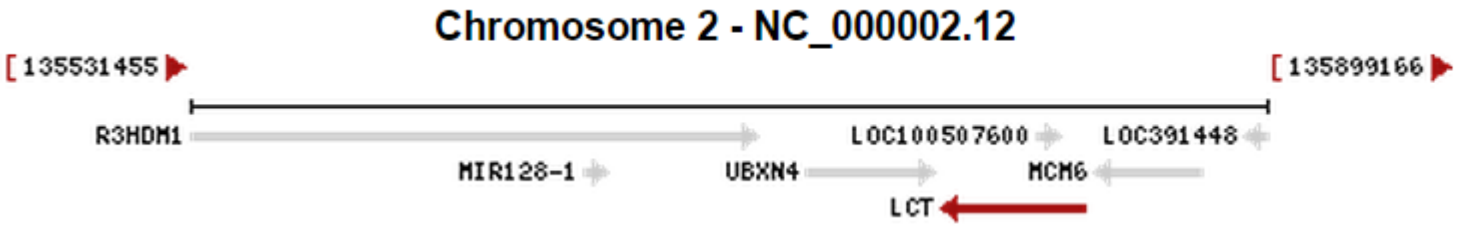
- Intolerant
  - Gene turned off
- Tolerant (adult)
  - Persistence
- Remedy
  - Lactase pill
  - Which does not interfere with transcription but just provide a supply of Lactase enzyme
  - Need to be taken before the Lactose food

LCT Gene

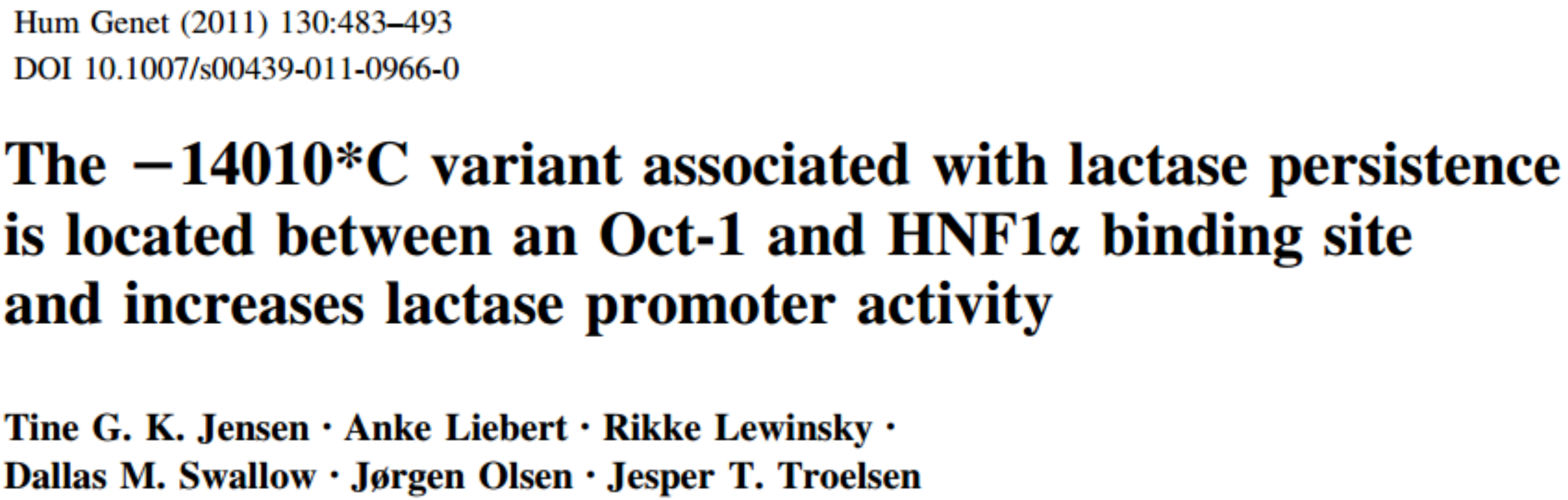
- Lactase is active during childhood but slows or stops when child grows up for some people
- LCT (short name for the Lactase gene)



  - Chr 2; 17 exons



Note the gene direction?



Hum Genet (2011) 130:483–493
DOI 10.1007/s00439-011-0966-0

The −14010*C variant associated with lactase persistence is located between an Oct-1 and HNF1α binding site and increases lactase promoter activity

Tine G. K. Jensen · Anke Liebert · Rikke Lewinsky · Dallas M. Swallow · Jørgen Olsen · Jesper T. Troelsen

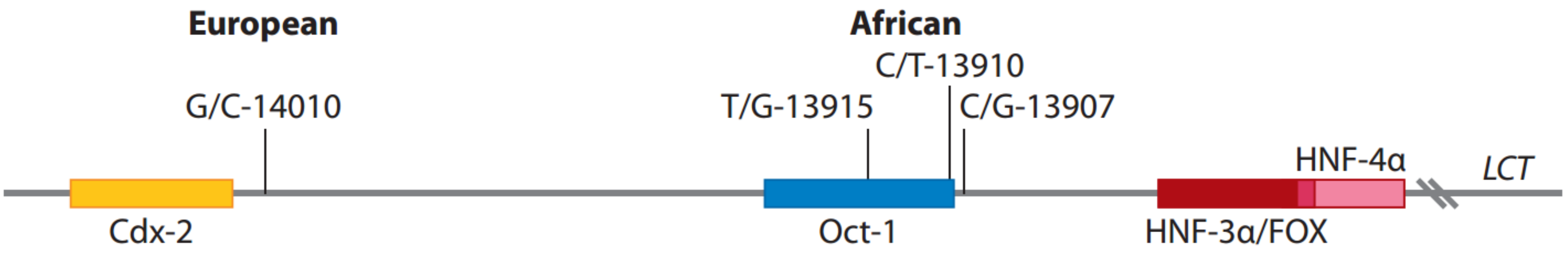


Figure 3

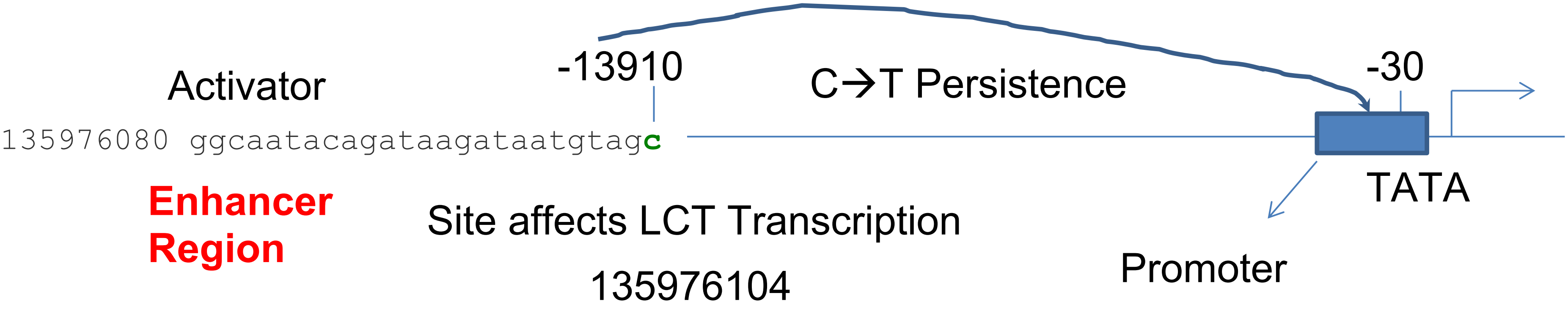
Locations of transcription factor-binding sites and predicted adaptive alleles upstream of *LCT*, the lactase gene. Three alleles were identified as potentially causal alleles in the African pastoral populations, whereas C/T-13910 was predicted to be the causal allele in Northern Europeans. Additionally, the T/G-13915 allele is correlated with lactase persistence in the Saudi Arabian population. The transcription factors and the sequence they bind in a supershift assay (48) are: HNF-4α (−13854 to −13830), HNF-3α and FOX (−13872 to −13848), Oct-1 and GAGA (−13933 to −13909), and Cdx-2 (−14040 to −14016).

150 *Kelley • Swanson*

Kelley and Swanson, Human Genet, 2008:9:143-160

Possible Mechanism

Protein (Activator) that bind in Enhancer regions far away from Promoter (non-binding) and bends and interacts with Promoter (TATA region) to positively affect transcription



Activator

-13910

C→T Persistence

-30

135976080 ggcaatacagataagataatgtagc

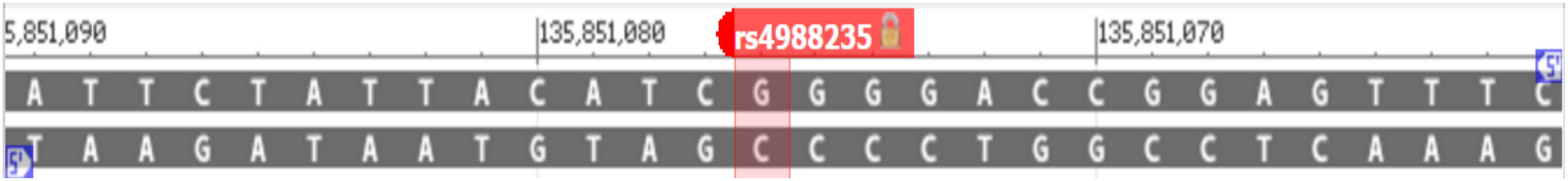
TATA

Enhancer Region

Site affects LCT Transcription

135976104

Promoter



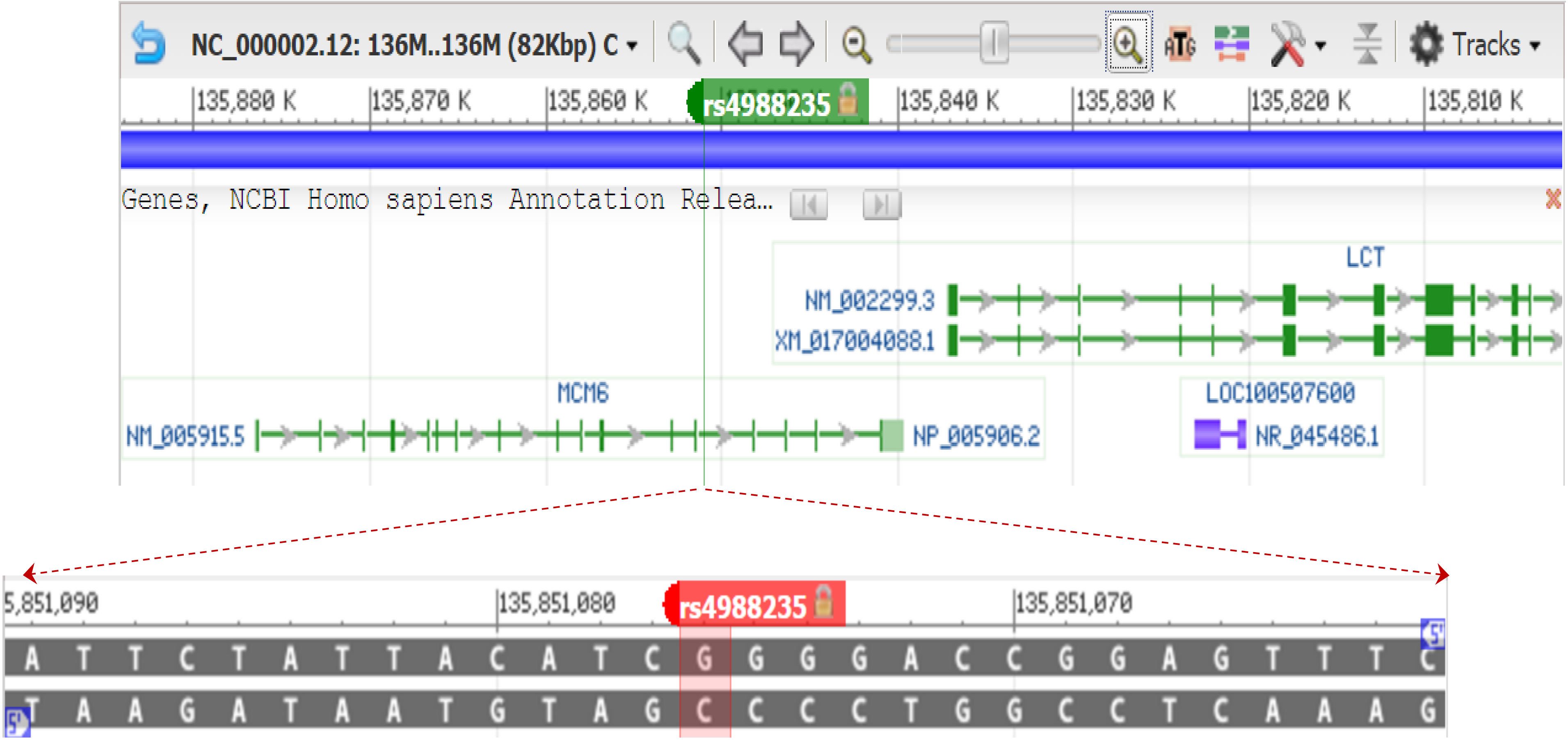
5,851,090 135,851,080 rs4988235 135,851,070

5' A T T C T A T T A C A T C G G G G A C C G G A G T T T C

T A A G A T A A T G T A G C C C C T G G C C T C A A A G 5'

http://www.fda.gov/forconsumers/consumerupdates/ucm094550.htm

Where is the variation (or lack of) that causes Lactose Persistence (Lactose intolerance)?



Application-2
Sickle Cell Disease (SCD):
Structure-based approach

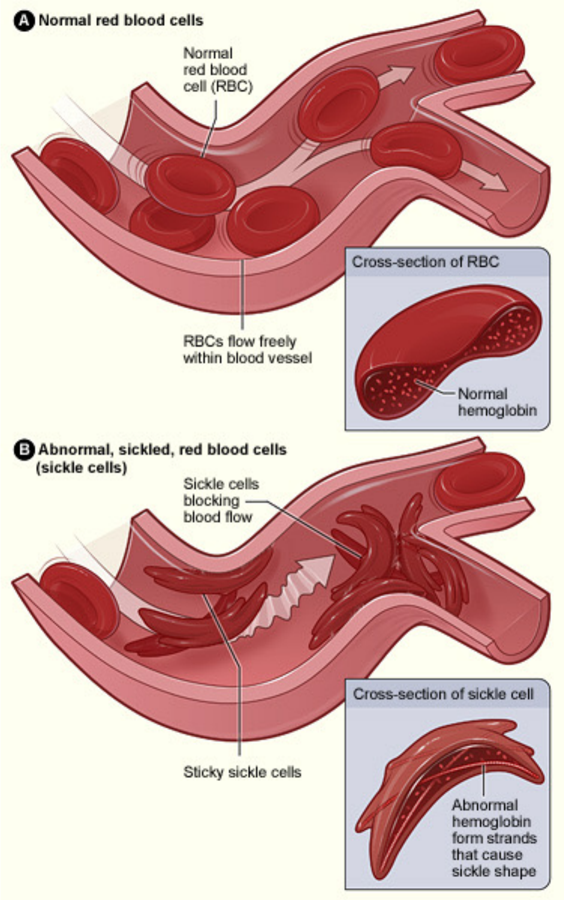

A Normal red blood cells
Normal red blood cell (RBC)
RBCs flow freely within blood vessel
Cross-section of RBC
Normal hemoglobin
B Abnormal, sickled, red blood cells (sickle cells)
Sickle cells blocking blood flow
Sticky sickle cells
Cross-section of sickle cell
Abnormal hemoglobin form strands that cause sickle shape

Figure taken from NCBI/NIH

Data Source CDC

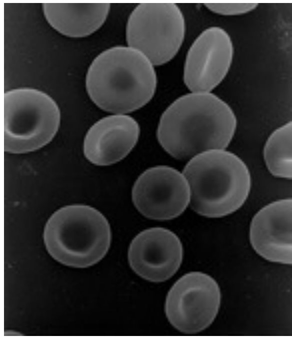
Malaria (2015)

214M (World-wide)
~1500 cases in US every year

Sickle-Cell Disease (SCD)

affects ~ 100,000 in US
occurs 1/365 black or African-American births
Occurs 1/16,300 Hispanic-American births

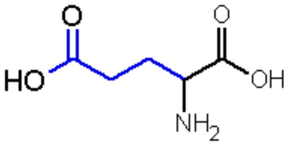
Biology of RBC



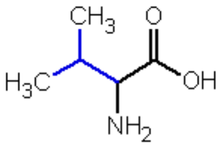
S.E.M of RBC Wikipedia

- RBC (a.k.a. erythrocytes, haematids etc.)
- Nucleus lacking in human
- 2.4 M raw RBC are produced/second
- Produced in bone marrow and travel all over body carrying O_2 (& CO_2)
- Carrier protein complex: Hemoglobin
  - Not a single gene product ; 2 α and 2 β chains
  - HBA: α ; HBB:β

https://en.wikipedia.org/wiki/File:Redbloodcells.jpg



Glutamic Acid (Glu)



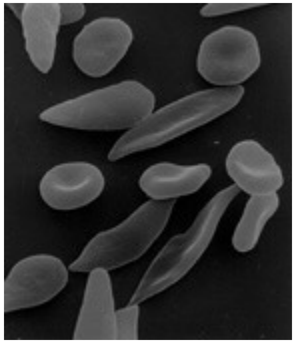
Valine (Val)

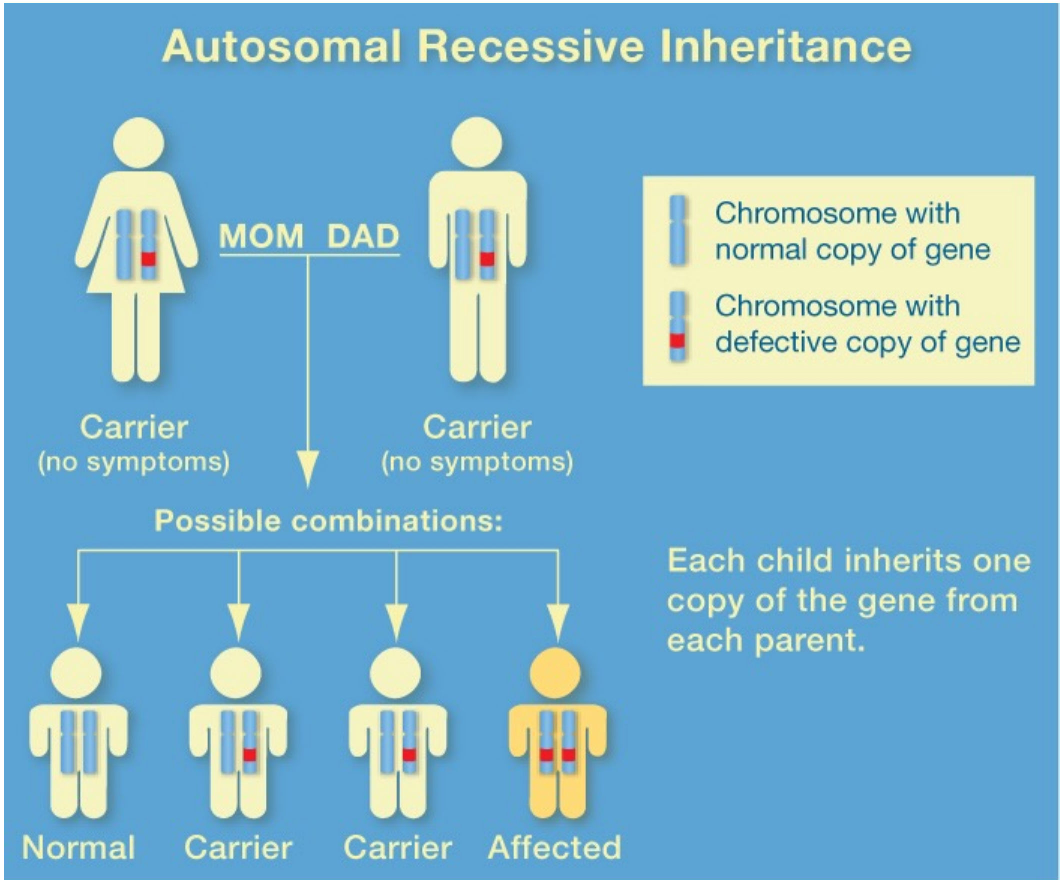
SCD

2 drops of oxygenated/deoxygenated blood

https://en.wikipedia.org/wiki/Red_blood_cell

- Genetic disorder
- HBB:Glu7→Val
- Cell sickle & dies sooner





Autosomal Recessive Inheritance

Chromosome with normal copy of gene

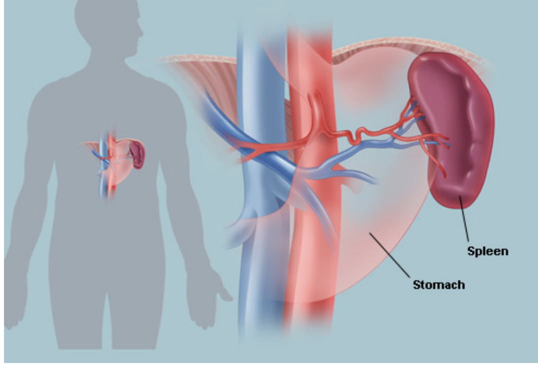
Chromosome with defective copy of gene

Each child inherits one copy of the gene from each parent.

Picture taken from

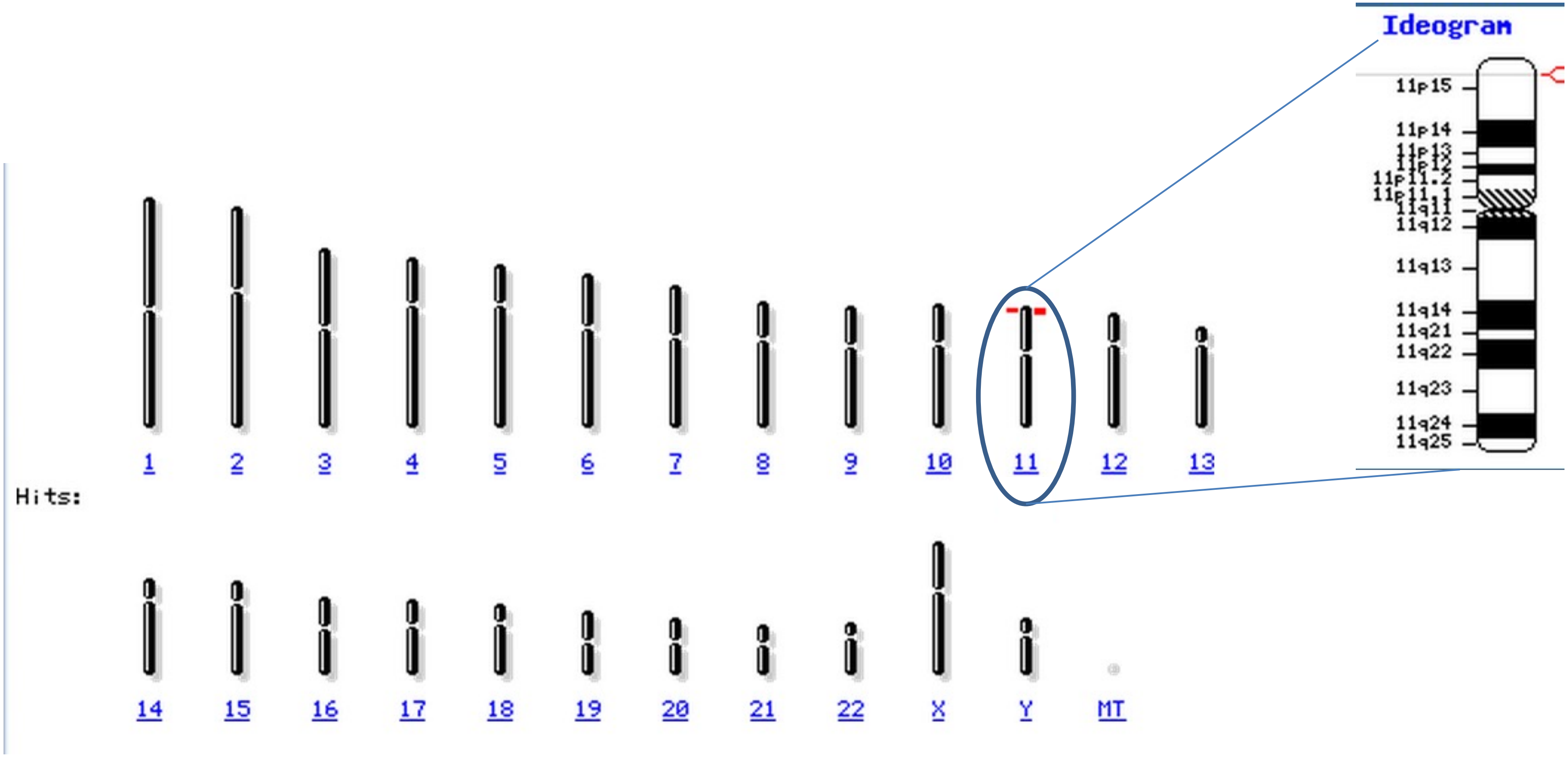
http://learn.genetics.utah.edu/content/disorders/singlegene/sicklecell/

Related Disease



- SCD → Anemia
- Spleen the blood filter will be clogged by the sickle cells and damages → infections
- **Malaria** is caused by mosquito bites. Parasite invades the RBC and **destroys** it.
- In US disease most commonly affects African American

HBB: Where is it?



NCBI Database

Hemoglobin alpha/beta sequences

```
        10         20         30         40         50
MVLSPADKTN VKAAWGKVGA HAGEYGAEAL ERMFLSFPTT KTYFPHFDLS
        60         70         80         90        100
HGSAQVKGHG KKVADALTNA VAHVDDMPNA LSALSDLHAH KLRVDPVNFK
       110        120        130        140
LLSHCLLVTL AAHLPAEFTP AVHASLDKFL ASVSTVLTSK YR
```

Alpha

E→V → Sickle Cell Anemia

```
        10         20         30         40         50         60
MVHLTPEEKS AVTALWGKVN VDEVGGEALG RLLVVYPWTQ RFFESFGDLS TPDAVMGNPK
        70         80         90        100        110        120
VKAHGKKVLG AFSDGLAHLD NLKGTFATLS ELHCDKLHVD PENFRLLGNV LVCVLAHHFG
       130        140
KEFTPPVQAA YQKVVAGVAN ALAHKYH
```

Beta

```
CLUSTAL multiple sequence alignment by MUSCLE (3.8)


sp|P68871|HBB_HUMAN      VKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFG
sp|P69905|HBA_HUMAN      -----KKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLP
                              *** .*:::.:**:*:: .::::**:**. **.*** **.**.: *: .** *:

sp|P68871|HBB_HUMAN      KEFTPPVQAAYQKVVAGVANALAHKYH
sp|P69905|HBA_HUMAN      AEFTPAVHASLDKFLASVSTVLTSKYR
                          ****.*:*: :*.:*.*:..*: **.
```

43.9% identity

Summary

- What is Bioinformatics?
- Basics of Bioinformatics
- Cell biology
  - DNA/Proteins/RNA
- Central Dogma
- Expression → Function
- Homology, sequence alignment (1D → 3D)
- Applications

Thanks

S. Ravichandran, Ph.D
ravichandran@hood.edu