

# A super quick introduction to molecular biology and genomics

*Héctor Corrada Bravo*  
Dept. of Computer Science  
Center for Bioinformatics and Computational Biology  
University of Maryland

*University of Maryland, Fall 2015*

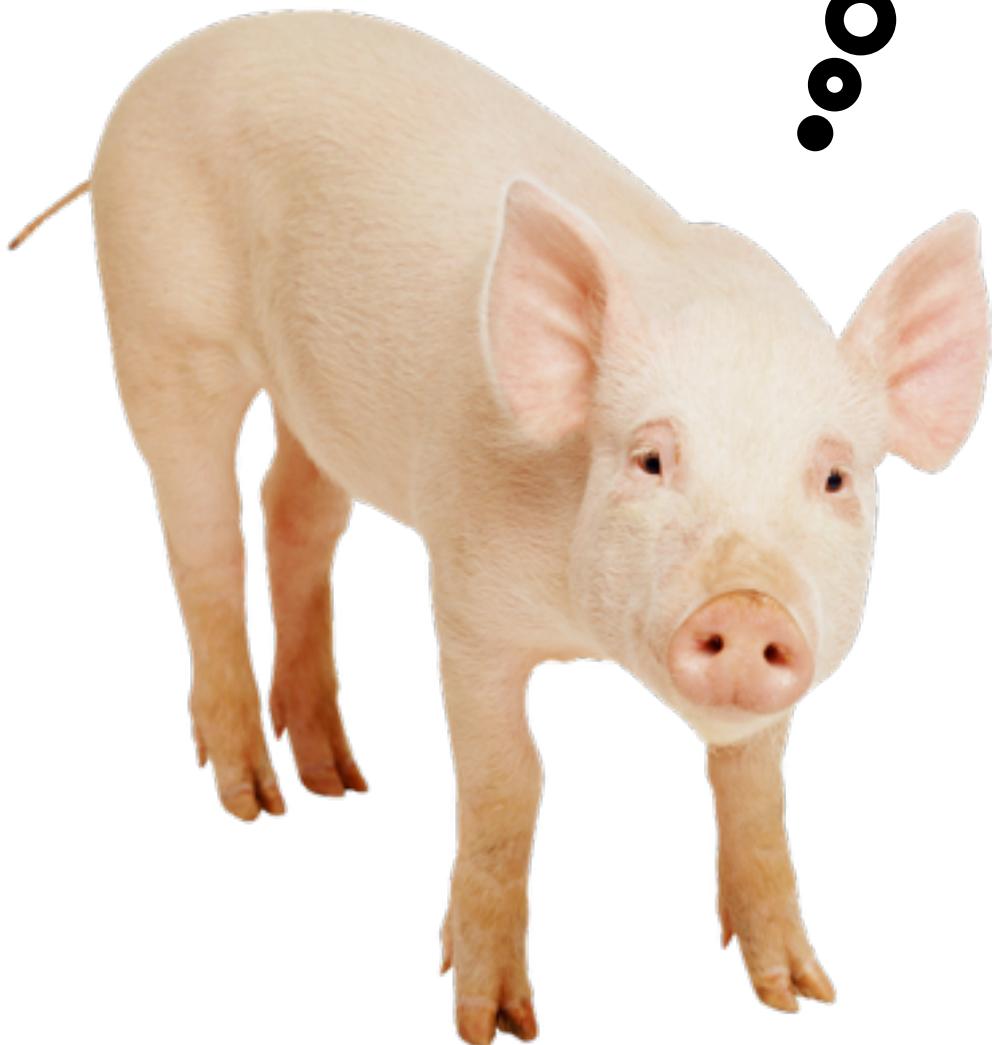
# Key terms

- Genotype/Phenotype
- Cell
- Proteins
- Evolution: inheritance, selection, variation
- DNA/RNA
- Chromosome
- Gene
- Genome
- Replication
- Transcription
- Exon/Intron
- Translation
- Codon
- Central Dogma
- Gene Expression
- Regulation
- Epigenetics

Why are my children  
such pigs?



*Why am I such a pig?*



*Phenotype, cells,  
metabolism, protein*

# Proteins

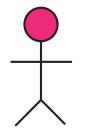
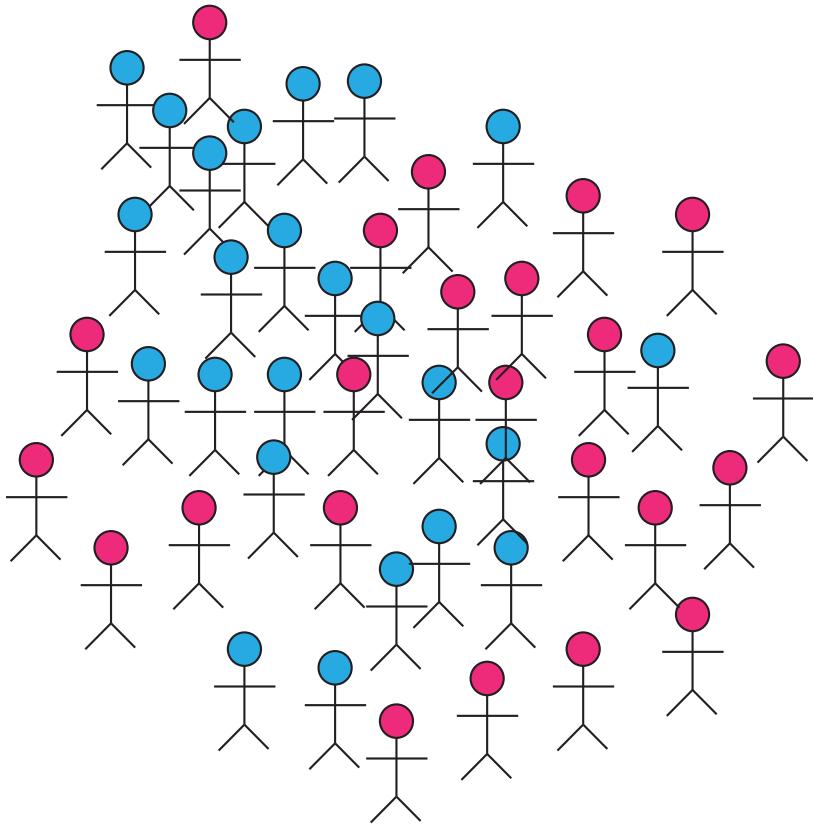
- *phenotype*: characteristics of an organism
- characteristics due to cellular structures and activities
  - mostly carried out by *proteins*
- Examples:

<i>alpha-keratin</i>	<i>component of hair</i>
<i>insulin</i>	<i>regulates blood glucose level</i>
<i>actin &amp; myosin</i>	<i>muscle contraction</i>
<i>hemoglobin</i>	<i>oxygen transport</i>
<i>DNA polymerase</i>	<i>synthesis of DNA</i>
<i>DNA glycosylases</i>	<i>DNA repair</i>
<i>matrix metalloproteinase</i>	<i>extra-cellular matrix degradation</i>

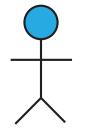
# Genetics

- *gene*: in classical genetics it was an abstract concept
  - a unit of inheritance passed from parent to offspring
  - specify proteins
- *genome* refers to the complete set of *genes*
- *genotype*: genetic characteristics of an individual

# What is Genomics?

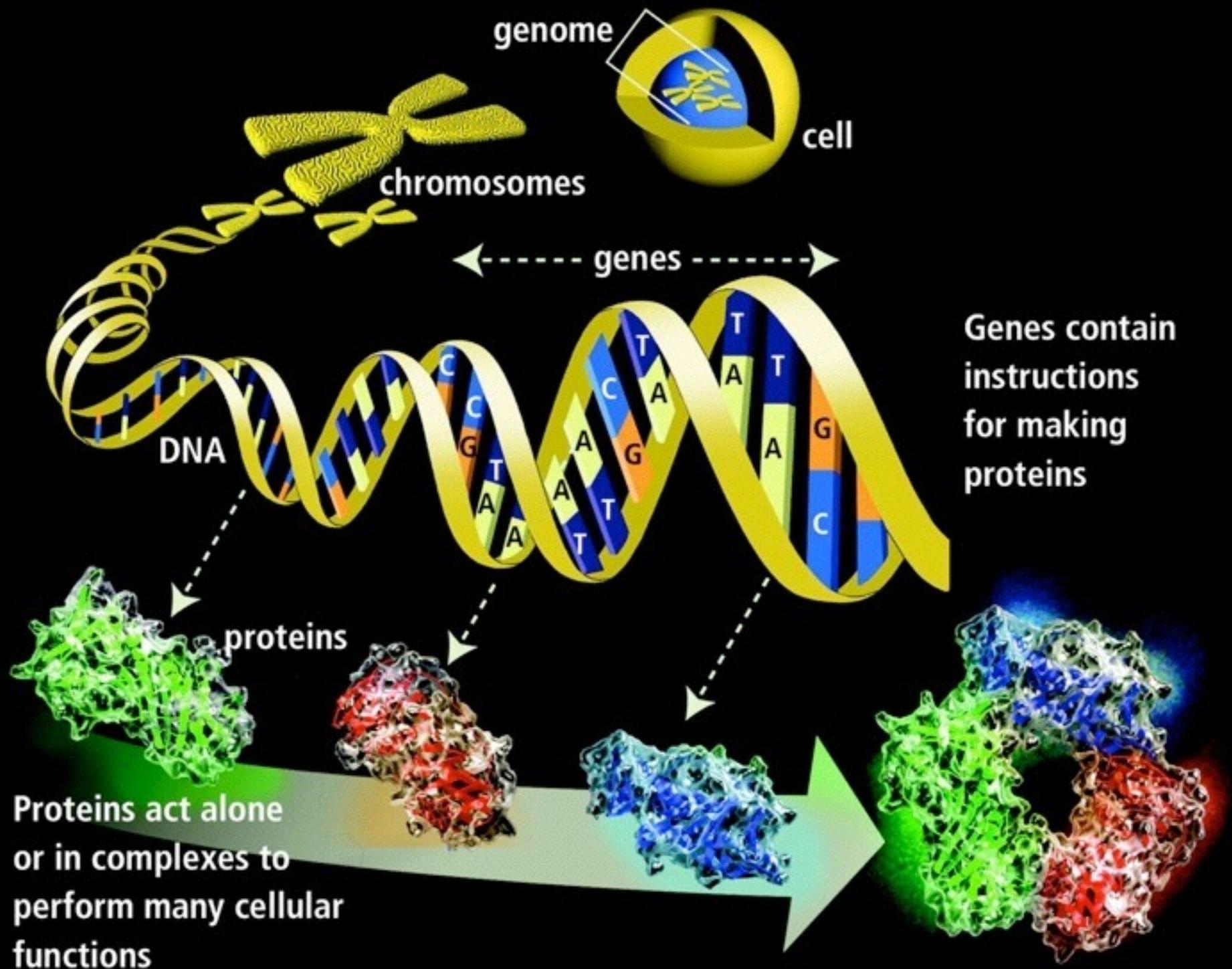


cancer



healthy

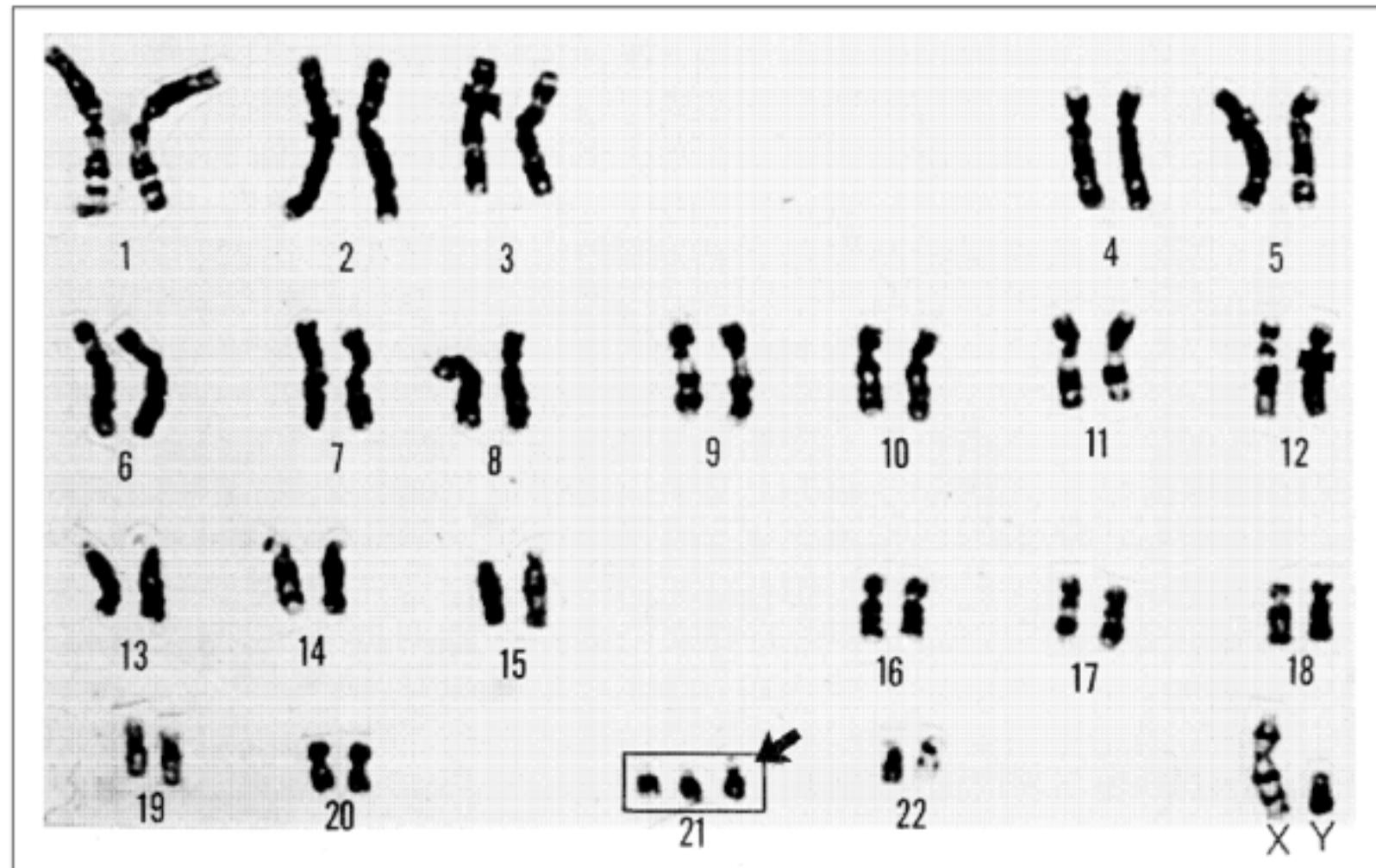
- Study the **molecular basis** of *variation* in development and disease
- Using **high-throughput** experimental methods
  - algorithms
  - ML
  - data management
  - modeling



# What is Genomics?

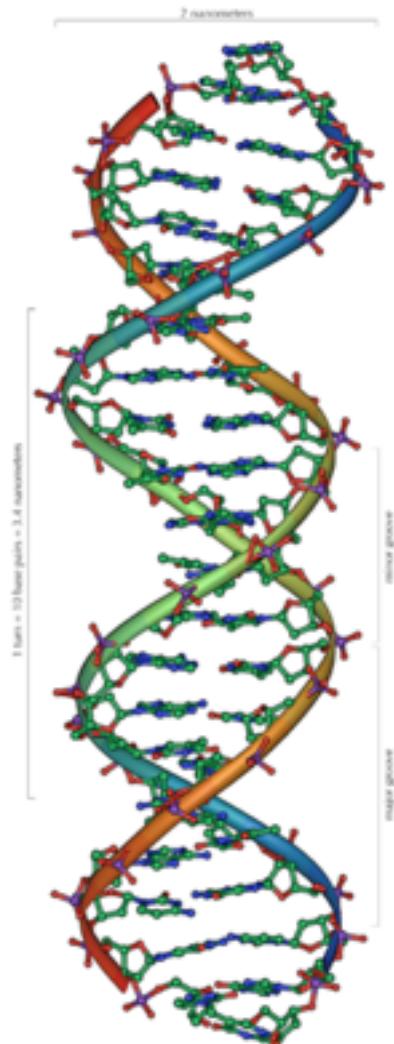
- Each cell contains a complete copy of an organism's **genome**, or blueprint for all cellular structures and activities.
- The genome is distributed along **chromosomes**, which are made of compressed and entwined **DNA**.
- Cells are of many different types (e.g. blood, skin, nerve cells), but all can be traced back to a single cell, the fertilized egg.

# Chromosomes



These are actually human. And for a down syndrome patient

# DNA



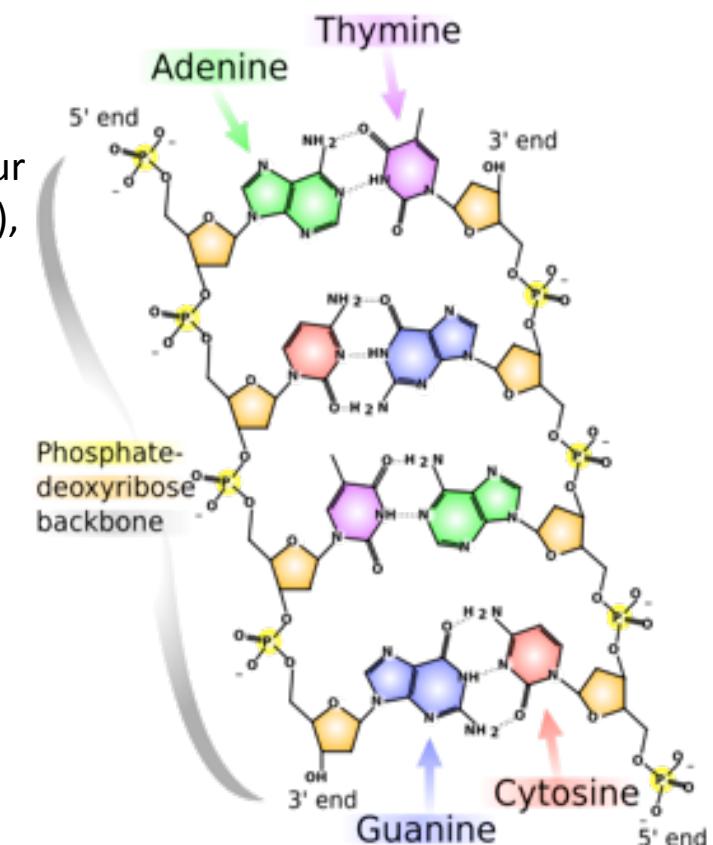
DNAs (Deoxyribonucleic acids) are molecules to store genetic information of a living organism.

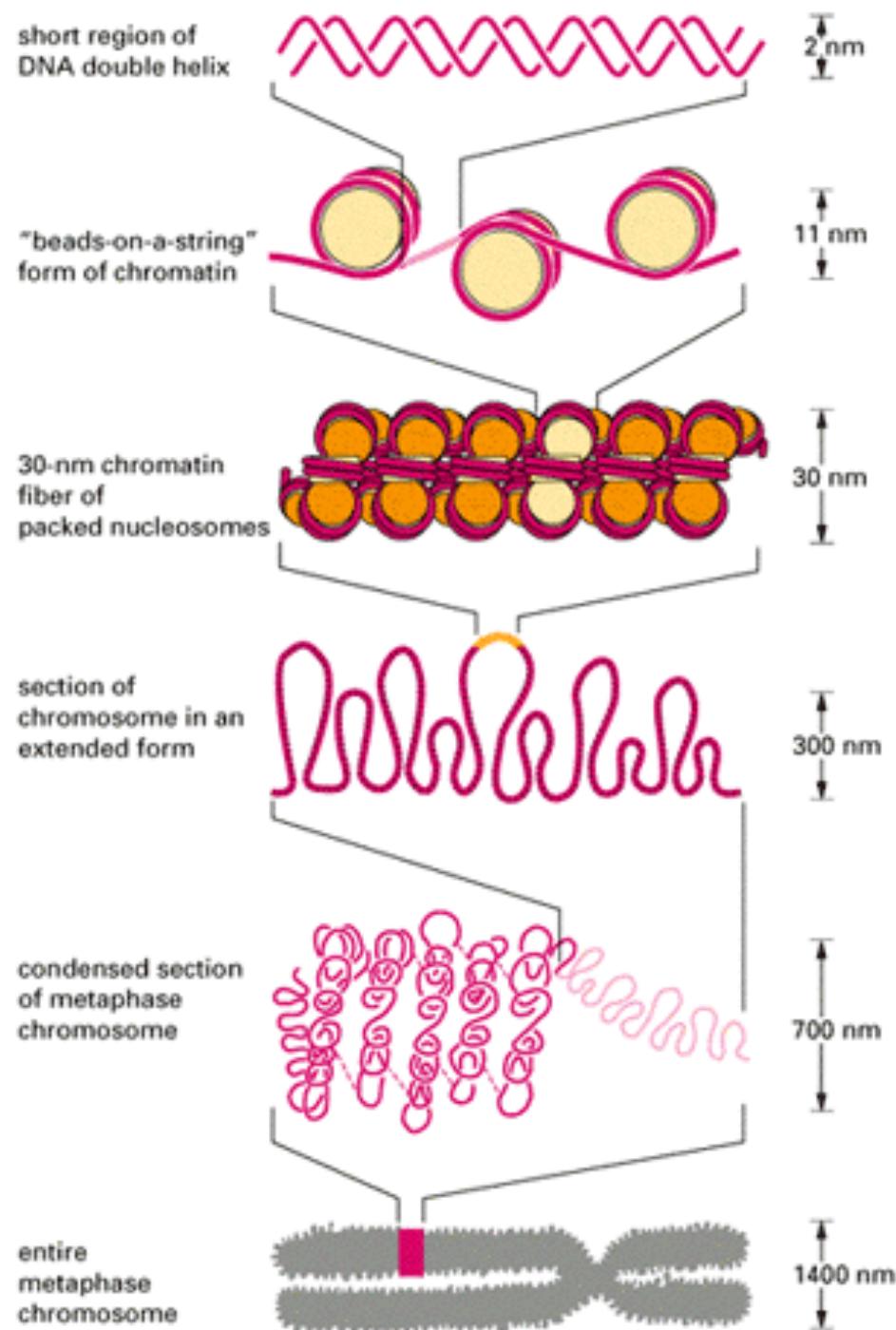
DNA consists of two polymers made from four types of nucleotides: adenine (A) guanine (G), cytosine (C) and thymine (T).

Purines: A, G; Pyrimidines: C, T

Two polymers are complementary to each other and form a double-helix structure

5' -ACCGTTCGACGGTAA-3'  
||| | | | | | | | | | |  
3' -TGGCAAGCTGCCATT-5'





# Measurement

- For a small enough piece, we can measure the sequence of bases, referred to as *sequencing*
- Human Genome Project



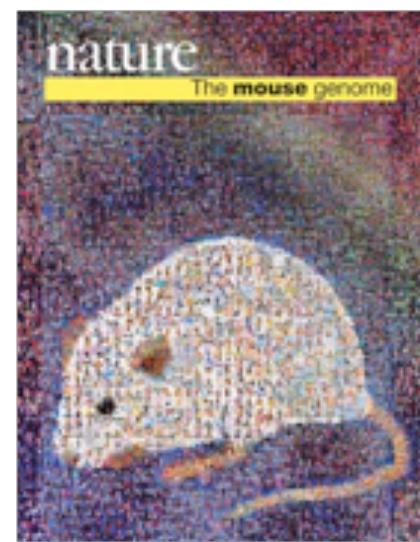
*D. melanogaster*, *Science*, 2000



*H. sapiens*, *Nature*, 2000  
and *Science*, 2000



*M. musculus*, *Nature*, 2002



# Genome

TCAGTTGGAGCTGCTCCCCACGGCCTCTCCTCACATTCCACGTCTGTAGCTCTATGACCTCCACCTTGAGTCCCTCCTCACACACTGAC  
ATGAAAAGGCACATGAGGATCCTCAAATAACCCCGTATCAGTCTCAGGGTAGCTCATAGCCTGGACAGGGCCCCCTCGGGGTTGCGCCC  
AGGTCCAGGCGGGGATGCACAGCAACAGTCACCGAAGCAGAAGCCGTACAGTGGTAGGGCTGGCAGTAGCTGGCACAGAGCTGCCAT  
GGCGGTGGACGTTGGGTTCCGAGGGTTGTGAGAACGGGCCACGGGCCCTGAGCGGTCCCTATTGCTAGGGCAGAATGCCCTCAGTAGA  
AATTCAAAAGCGTCTCTGCGCGGTCTGTAGGGGGTGGCCGAAGCCTCTAGGGGATCCCTCGAGGCTGCTGGCCTGCCGTCCAGG  
GGACAAGGAGCCAGAGTCAGGTGGGCTGTTGCCAGGGGTCAGGGAGGCTGATGTCTGGAGTCCGGATGGACCACCTGCAGAGGAGAGAC  
ATAGGTCAACACAGGGAGGTAGGATGGTGTGATGTTCCACCCACAAAAGAAAACCTATTCTTAGAACCTCCAGGATGTGAATCCTGCCT  
GCACCTGCACAGCTGGCTGGAGGCATATAGCCACTGCCCATAGATCTCAACTTACCTCACAACCAACTGCCAGGCCTAAGTTCTGCC  
TCAAAACGTCCAAGGCCTGGATAGCCAAGAGCCTGGGTCTTGGAAATATGCAACCATAAATAGTAGCTTTAGAAAGTATAAGGCTCTGTT  
TCTGGGTATATTAGTGTGTTTCACCTGTCCCCAGCCCTAACAGCCAGGTGTGGCCAGAACGAAATGTACTGTAAGAGCAGAGCAAAACTC  
CACACAGATAGTCTGTTAGGCAATACATCTCTGCCTGACTATTAGGAATCTGGTTCTGGCCTCTGTACAAAGCTGGAGCAACACAGTG  
GCCACATCAATCAAAGGACCGTGACCAACTTCAAAGTCGGTGAGCTGTACCTATTAGGCTCTGTAACAGAACAGATTCAACTA  
CAGCTCAGCAGGGCATCGTCACGGGTGTGTGTGTGTGTGTGTGTGTGTGTGGGGGGGGGGTGGACAGAGGACGGGAC  
ACAATTCACTGCCAGCCCTCTCCTCAAGGAAGGCTGCTCTACGCTGGACTGGAATACACATTCTGTAAACATGGTGGGGCCTCA  
GGCAAGCCAGAGTTGGAGCCTCCTTAACCTTCAAGGTGAGCATCTTGACTTGGAGGGTGGGGGTGGGAAGGAAGGAACCTGTGGAC  
TCCTCCCTACAAGACAGAAAAGGAATAAGCCACGAAGACAATAACGATTTGTATCAAGCGTCTCTCCATTCAAGCTTACCTGACAATGA  
AATCAAATTGGACCCCTGCAAGCATCAGTACACCCAGCAGAGTGGACACAGCACCGTCCAGAACGGAGCAAACATGTGCTCAGAGCGAGCA  
TAGCCCTGTGGTTCTGTCCCCATGGCTGTAGAAAGGCCTGAACAAAGGAGAAAATTGACACGGTCACATTCTGGGTGTGGTAAAGTGCTC  
AGCTGTGTCTATACTTGGGTTTGAT...

**Total amount of DNA in human genome:  
3 \* 10<sup>9</sup> base pairs (bp)**

UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

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

position/search chr11:102,646,196-102,646,3: gene jump clear size 187 bp. configure

chr11 (q22.2) 11p15.4 015.1 014.3 04-1 11p13 11p12 013.2 12.1 013.4 11q14.1 014.3 11q21 022.1 11q22.3 016q3.3 023.2 q25

move start

Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.

# Replication

T A  
T A  
C G  
G C  
A T  
T A  
T A  
A T  
C G  
G C  
A T

T  
T  
C  
G  
A  
T  
T  
A  
C  
G  
A

A  
A  
G  
C  
T  
A  
A  
T  
G  
C  
T

T  
T  
C  
G  
A  
T  
T  
A  
C  
G  
A

C C C G T A A  
G T  
A T T T G

T T G G G T A A T  
G C

A T G G G T C A A

TTA

TTT A G T A G

A A T G T C

A  
A  
G  
C  
T  
A  
A  
T  
G  
C  
T



nucleotides available in cells

T T C G A T T A C G A

A A G C T A A T G C T

T T C G A T T A C G A

A A G C T A A T G C T

T            A  
T            A  
C            G  
G            C  
A            T  
T            A  
T            A  
A            T  
C            G  
G            C  
A            T

T            A  
T            A  
C            G  
G            C  
A            T  
T            A  
T            A  
A            T  
C            G  
G            C  
A            T

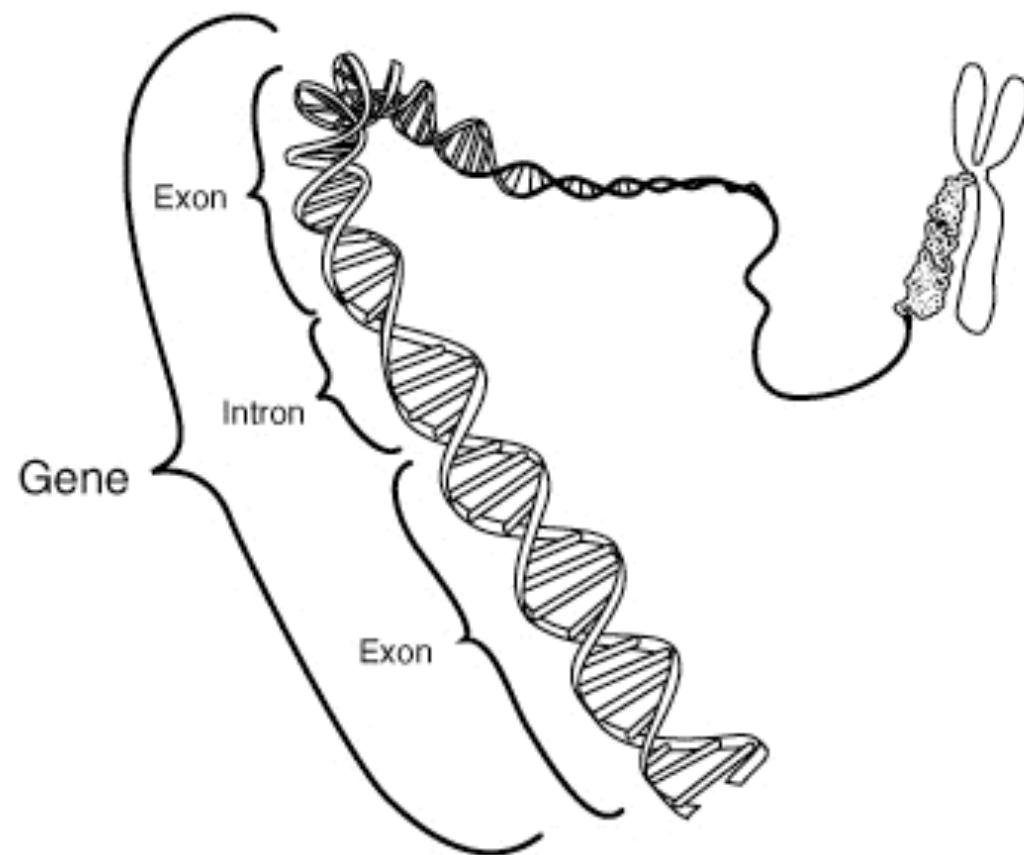
# Why are these two different?



Differences explained by 1-10% difference in genome

Similarities explained by similar genes

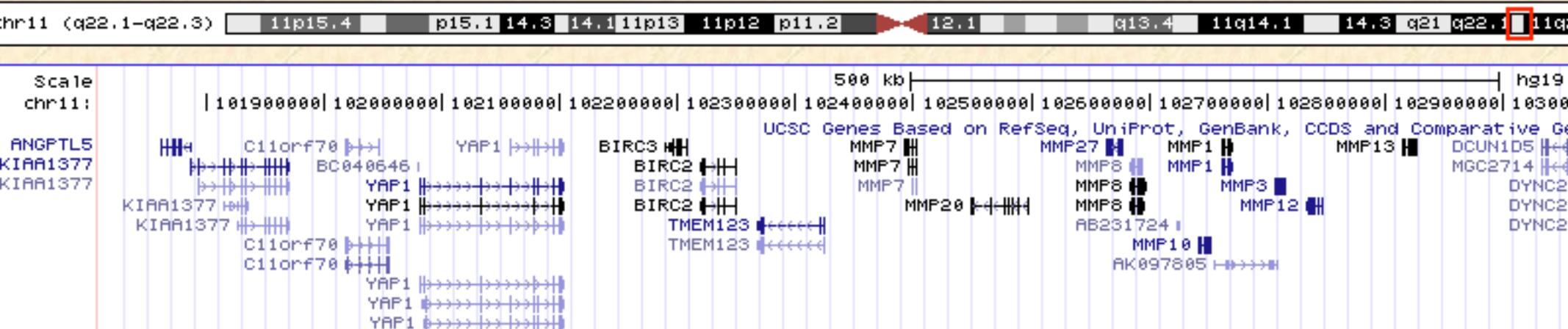
# Genes



# UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

move [<<<](#) [<<](#) [<](#) [>](#) [>>](#) [>>>](#) zoom in [1.5x](#) [3x](#) [10x](#) [base](#) zoom out [1.5x](#) [3x](#)

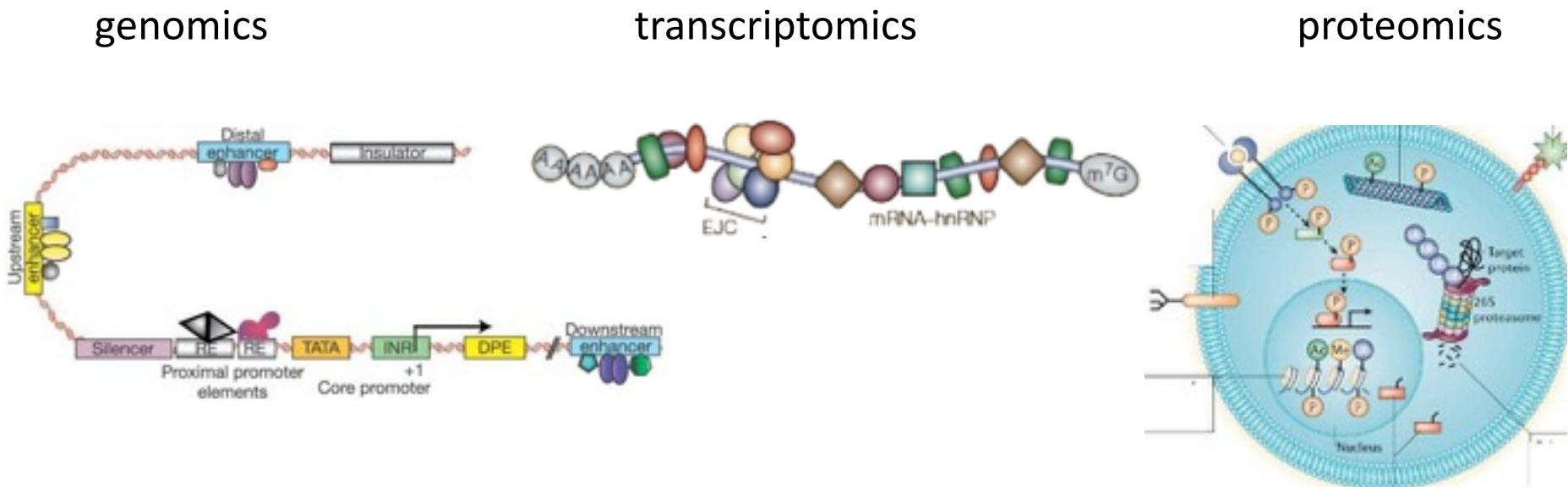
position/search [chr11:101,711,289-103,581,21](#) gene  [jump](#) [clear](#) size 1,870,000 bp. [config](#)



Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options.

# Central Dogma

Genes encode proteins which are transcribed into mRNA and translated into proteins.



**Major technological advances allow unprecedented data acquisition**

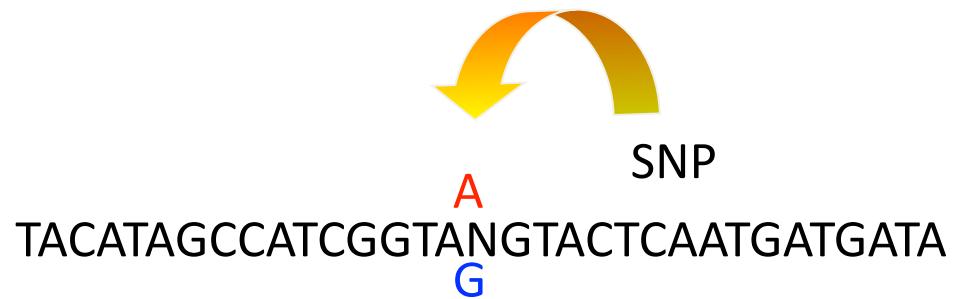
# What makes them different?



Much human variation is due to difference in ~ 6 million base pairs (0.1 % of genome) referred to as SNPs

# Single Nucleotide Polymorphism (SNP)

Genomic DNA:



Three genotypes

AA

Mother

TACATAGCCATCGGTAAGTACTCAATGATGATA  
ATGTATCGGTAGCCATTCATGAGTTACTACTAT

Father

TACATAGCCATCGGTAAGTACTCAATGATGATA  
ATGTATCGGTAGCCATTCATGAGTTACTACTAT

AG

Mother

TACATAGCCATCGGTAA**G**TACTCAATGATGATA  
ATGTATCGGTAGCCATT**C**CATGAGTTACTACTAT

Father

TACATAGCCATCGGTAG**G**TACTCAATGATGATA  
ATGTATCGGTAGCCAT**C**CATGAGTTACTACTAT

GG

Mother

TACATAGCCATCGGTAGGTACTCAATGATGATA  
ATGTATCGGTAGCCATCCATGAGTTACTACTAT

Father

TACATAGCCATCGGTAGGTACTCAATGATGATA  
ATGTATCGGTAGCCATCCATGAGTTACTACTAT

# From reads to evidence

# From reads to evidence

## I. Comparative

Sequence-wise, individuals of a species are nearly identical

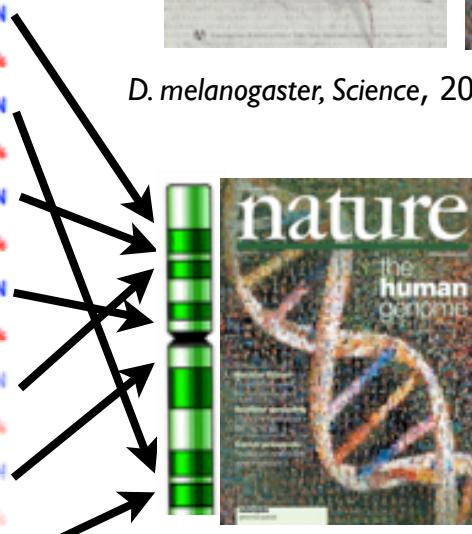
Well curated, annotated “reference” genomes exist



*D. melanogaster*, *Science*, 2000

*H. sapiens*, *Nature*, 2000  
and *Science*, 2000

*M. musculus*, *Nature*, 2002



Idea: “Map” reads to their point of origin with respect to a reference, then study differences

# **SEC-GEN SEQUENCING FOR SNPs**

A blue outlined arrow pointing right with the word "Align" inside.

```

GTCGAGTANCTGCTC
||||||| ||||| |
GTCGAGTATCTGCTC

GGATCTGCGATATAACC
||||||| - ||||| |
GGATCT-CGATATAACC

> AATCTGATCTTATTTT
||||||| |
AATCTGATCTTATTTT

ATATATATATATATATAT
||||||| |
ATATATATATATATATAT

TCTCTCCCANNAGAGGC
||||||| |
TCTCTCCCAAGGAGAGC

```

“Pileup” or “Coverage plot”

“Depth of coverage” = 14

Call: HET A, G

## “Coverage”

## Reference

(slide courtesy of Ben Langmead)

## UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly

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

position/search chr11:102,641,234-102,651,34 gene jump clear size 10,111 bp. configure



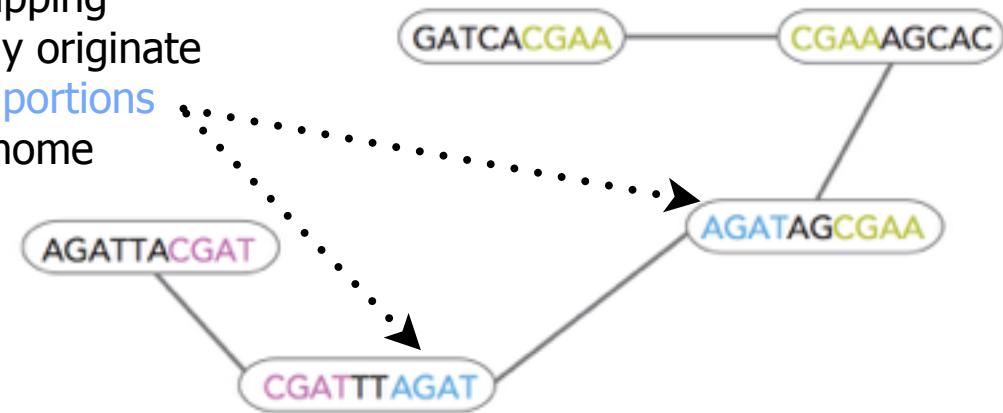
a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels to re-order tracks. Drag tracks left or right to new position.

# From reads to evidence

## 2. *de novo*

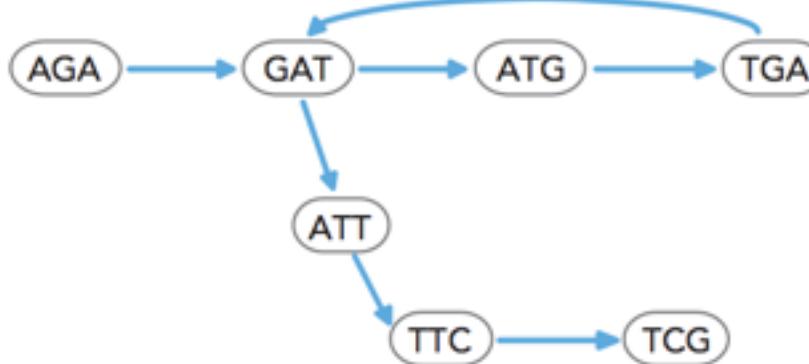
Assume nothing! - let reads tell us everything

Reads with overlapping sequence probably originate from **overlapping portions** of the subject genome



Source: De Novo Assembly Using Illumina Reads. Illumina. 2010

Encode overlap relationships as a graph

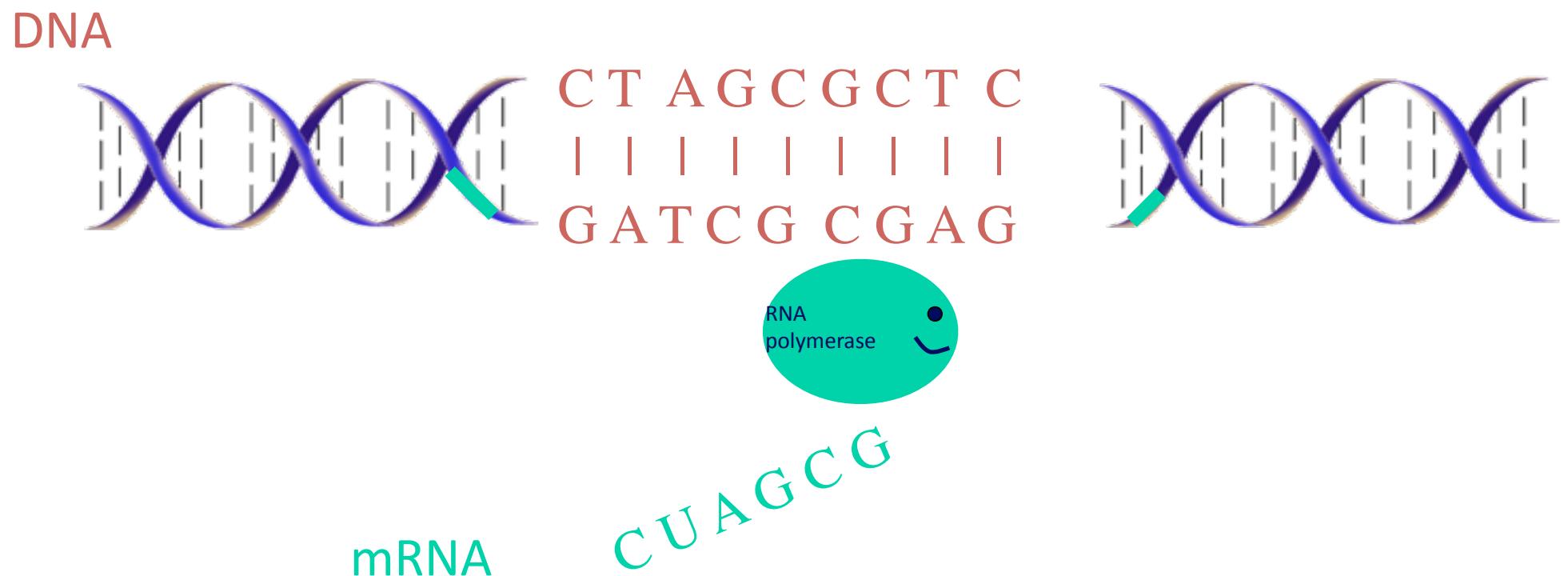


The full genome sequence is a “tour” of the graph

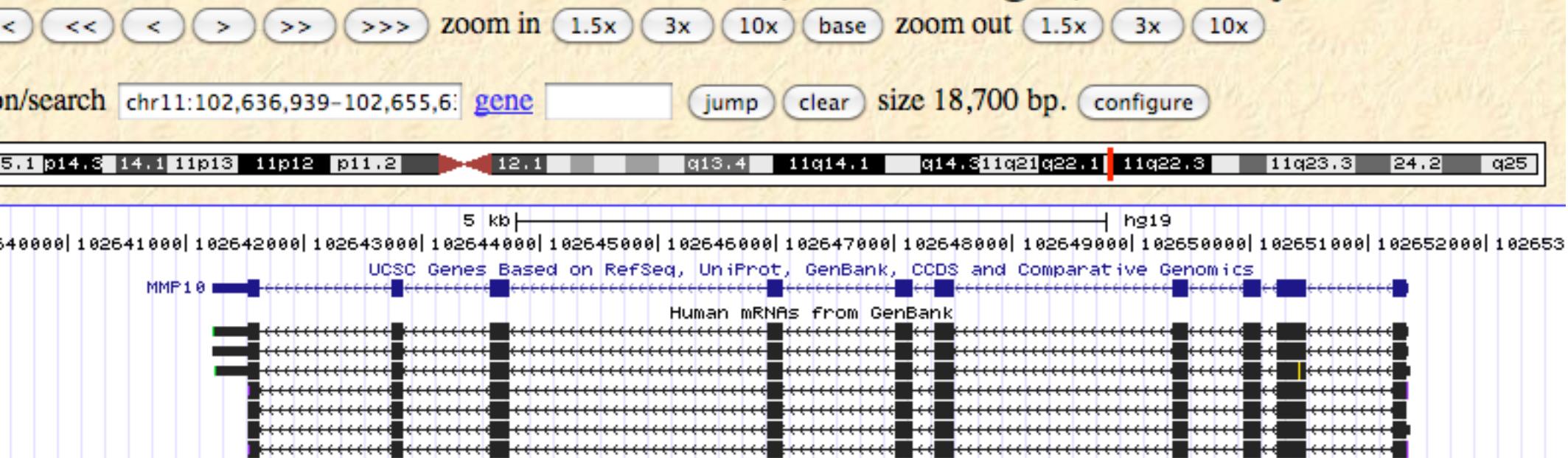
Source: De Novo Assembly Using Illumina Reads. Illumina. 2010

[http://www.illumina.com/Documents/products/technotes/technote\\_denovo\\_assembly.pdf](http://www.illumina.com/Documents/products/technotes/technote_denovo_assembly.pdf)

# Transcription

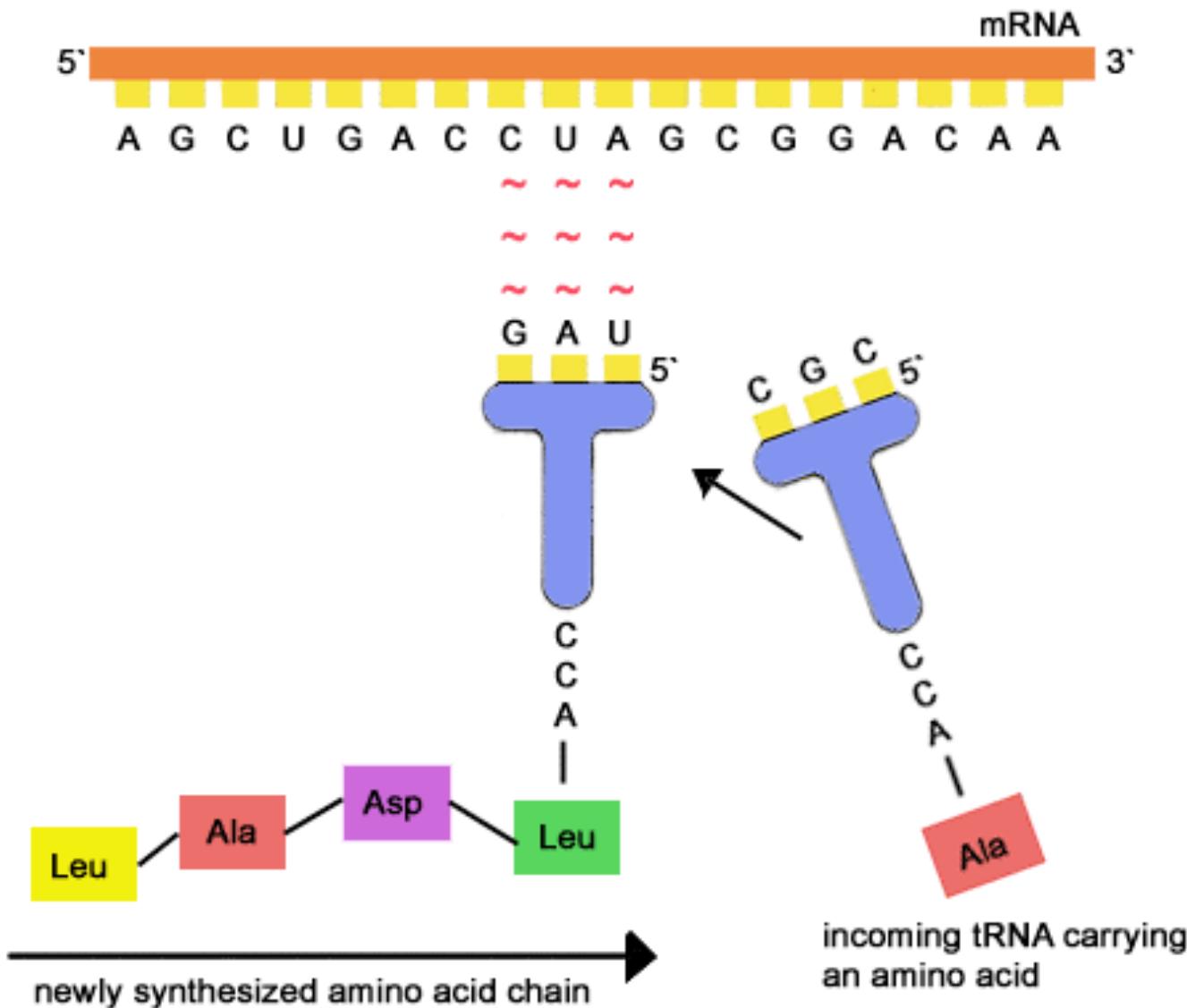


# Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly



for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down. Drag tracks left or right to new position.

# Translation



# The genetic code

		Second Letter					
		T	C	A	G		
First Letter	T	TTT TTC TTA TTG } Phe	TCT TCC TCA TCG } Ser	TAT TAC TAA TAG } Tyr Stop Stop	TGT TGC TGA TGG } Cys Stop Trp	T C A G	
	C	CTT CTC CTA CTG } Leu	CCT CCC CCA CCG } Pro	CAT CAC CAA CAG } His Gln	CGT CGC CGA CGG } Arg	T C A G	
	A	ATT ATC ATA ATG } Ile Met	ACT ACC ACA ACG } Thr	AAT AAC AAA AAG } Asn Lys	AGT AGC AGA AGG } Ser Arg	T C A G	
	G	GTT GTC GTA GTG } Val	GCT GCC GCA GCG } Ala	GAT GAC GAA GAG } Asp Glu	GGT GGC GGA GGG } Gly	T C A G	

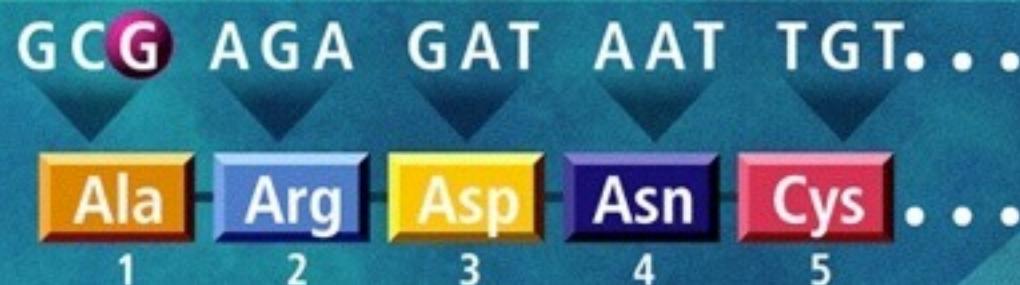
# DNA Sequence Variation in a Gene Can Change the Protein Produced by the Genetic Code

*Gene A from Person 1*



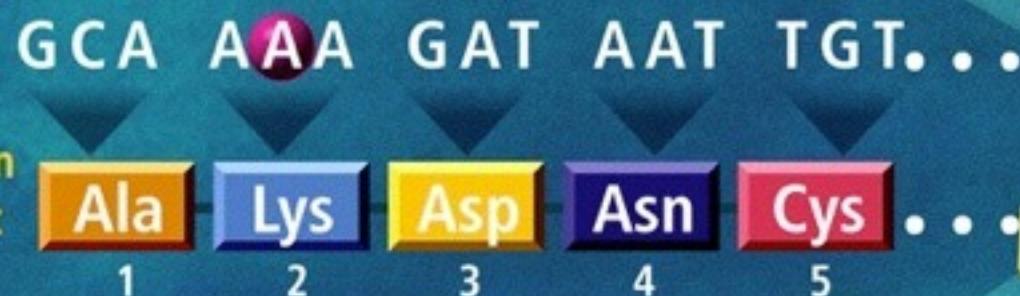
*Gene A from Person 2*

Codon change made no difference in amino acid sequence



*Gene A from Person 3*

Codon change resulted in a different amino acid at position 2



# Health or Disease?

**Person 1**

*DNA Sequence*

A A A T T T



**Normal protein**



Some DNA variations have no negative effects

**Person 2**

A A T T T T



**Low or nonfunctioning protein**



Other variations lead to disease (e.g., sickle cell) or increased susceptibility to disease (e.g., lung cancer)

**Person 3**

A A C T T T



# RNA-seq differential expression

**Sample A**

GTGCGAGTANCTGTCT  
||||||| |||||  
GTGCGAGTATCTGTCT  
  
GGATCT**G**GGATATAACC  
|||||||  
GGATCT-**C**GGATATAACC  
  
AATCTGATCTTATTTT  
|||||||  
AATCTGATCTTATTTT  
  
ATATATATATATATAT  
|||||||  
ATATATATATATATAT  
  
TCTCTCCC**A**NNAGAGC  
|||||||  
TCTCTCCC**G**GGAGAGC

The logo for AGCA, featuring a blue arrow pointing right with the acronym in orange text below it.



GTCGCACTATCTGTCT  
 GTCGCACTATCTGTCT  
 GTCGCACTATCTGTCT  
 GTCGCACTATCTGTCT  
 GTCGCACTATCTGTCT  
 GTCGCACTATCTGTCT  
 TGTCGCACTATCTGTCT  
 TATGTCGCACTATCTG  
 TATATCGCACTATCTG  
 TATATCGCACTATCTG  
 TATATCGCACTATCTG  
 CCCTATATCGCACTAT  
 AGCACCCCTATGTCGCA  
 AGCACCCCTATATCGCA  
 AGCACCCCTATGTCGCA  
 GAGCACCCCTATGTCGCA  
 CCGGAGCACCCTATAT  
 CCGGAGCACCCTATAT  
 GCCGGAGCACCCTATG

**Gene 1  
differentially  
expressed?: YES**

p-value: 0.0012

GGTATTTCTGGGGGTATGCACGCCATAGCATTGGAGACGGCTGGAGCCGGAGCACCTATGTCGAGTATCTGCTTGTATTCCTGCTCACTCTATTATATTATCGCACCTACGGTCAATT

A blue outline of a right-pointing arrow.

**GTCGCA**GTANCTGTCT  
|||||||  
**GTCGCA**GTATCTGTCT  
  
**GGATCT**CGATATAACC  
|||||||  
**GGATCT**-CGATATAACC  
  
**AATCTGATCTT**ATTTT  
|||||||  
**AATCTGATCTT**ATTTT  
  
**ATATATATATATATAT**  
|||||||  
**ATATATATATATATAT**  
  
**TCTCTCCC**ANNAGAGC

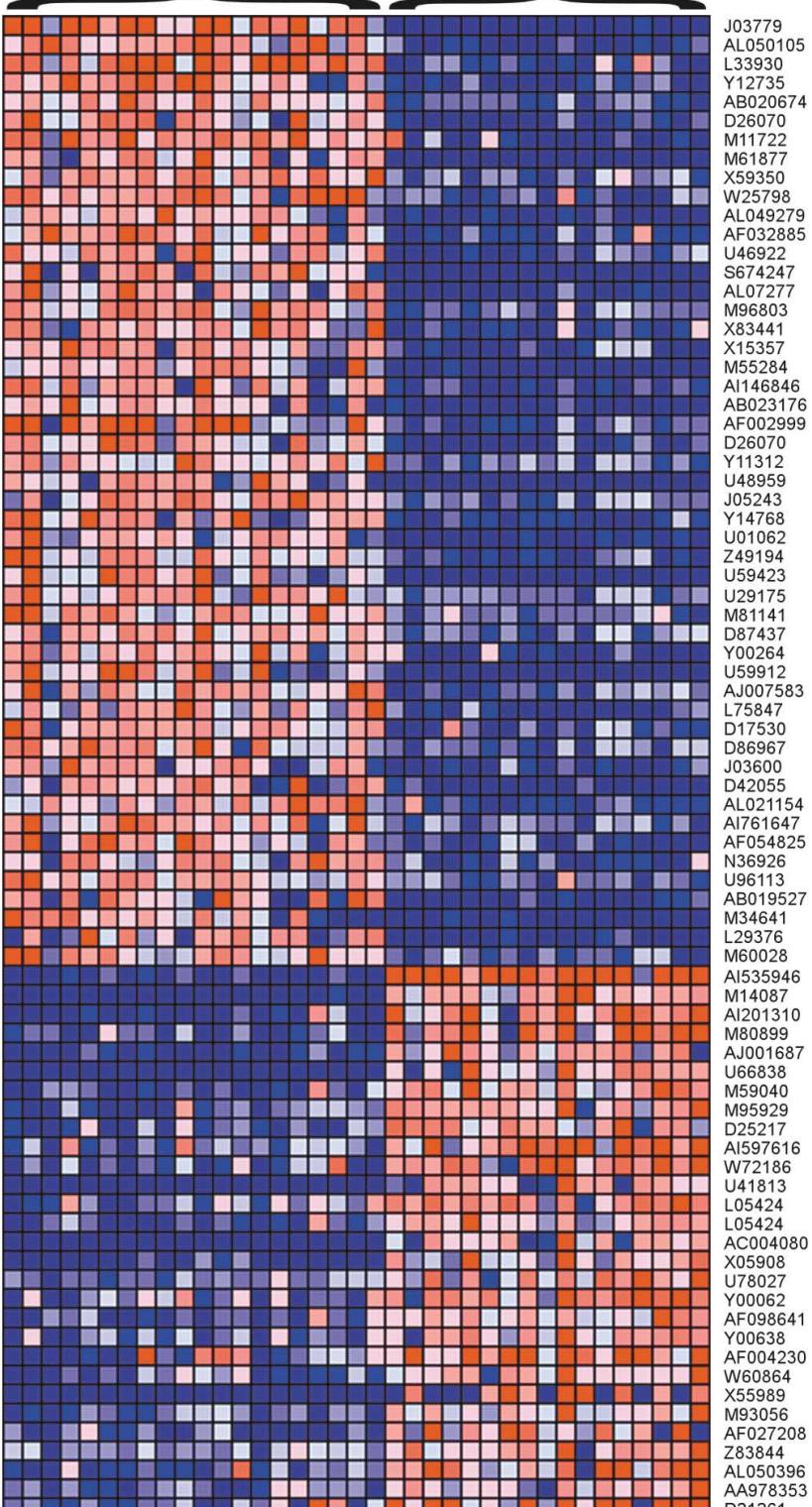
A blue outline of a right-pointing arrow.

TGTTCGAGTATCTGT  
AGCACCCATATGTCGA  
GCCGGAGCACCCATATG

Slide courtesy B. Langmead

ALL

MLL



J03779  
AL050105  
L33930  
Y12735  
AB020674  
D26070  
M11722  
M61877  
X59350  
W25798  
AL049279  
AF032885  
U46922  
S674247  
AL07277  
M96803  
X83441  
X15357  
M55284  
A1146846  
AB023176  
AF002999  
D26070  
Y11312  
U48959  
J05243  
Y14768  
U01062  
Z49194  
U59423  
U29175  
M81141  
D87437  
Y00264  
U59912  
AJ007583  
L75847  
D17530  
D86967  
J03600  
D42055  
AL021154  
AI761647  
AF054825  
N36926  
U96113  
AB019527  
M34641  
L29376  
M60028  
A1535946  
M14087  
A1201310  
M80899  
AJ001687  
U66838  
M59040  
M95929  
D25217  
A1597616  
W72186  
U41813  
L05424  
L05424  
AC004080  
X05908  
U78027  
Y00062  
AF098641  
Y00638  
AF004230  
W60864  
X55989  
M93056  
AF027208  
Z83844  
AL050396  
AA978353  
D01221

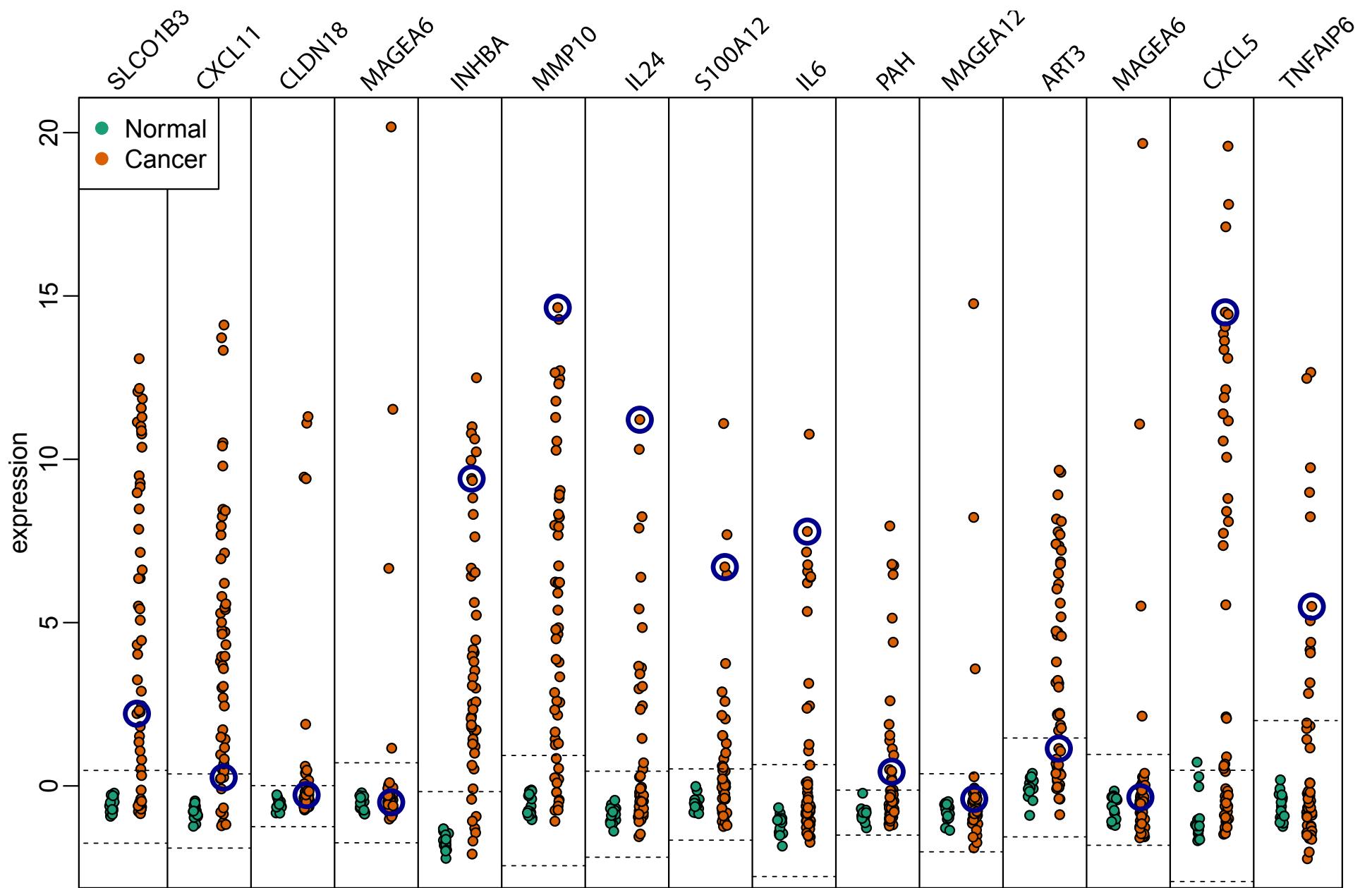
article

## MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia

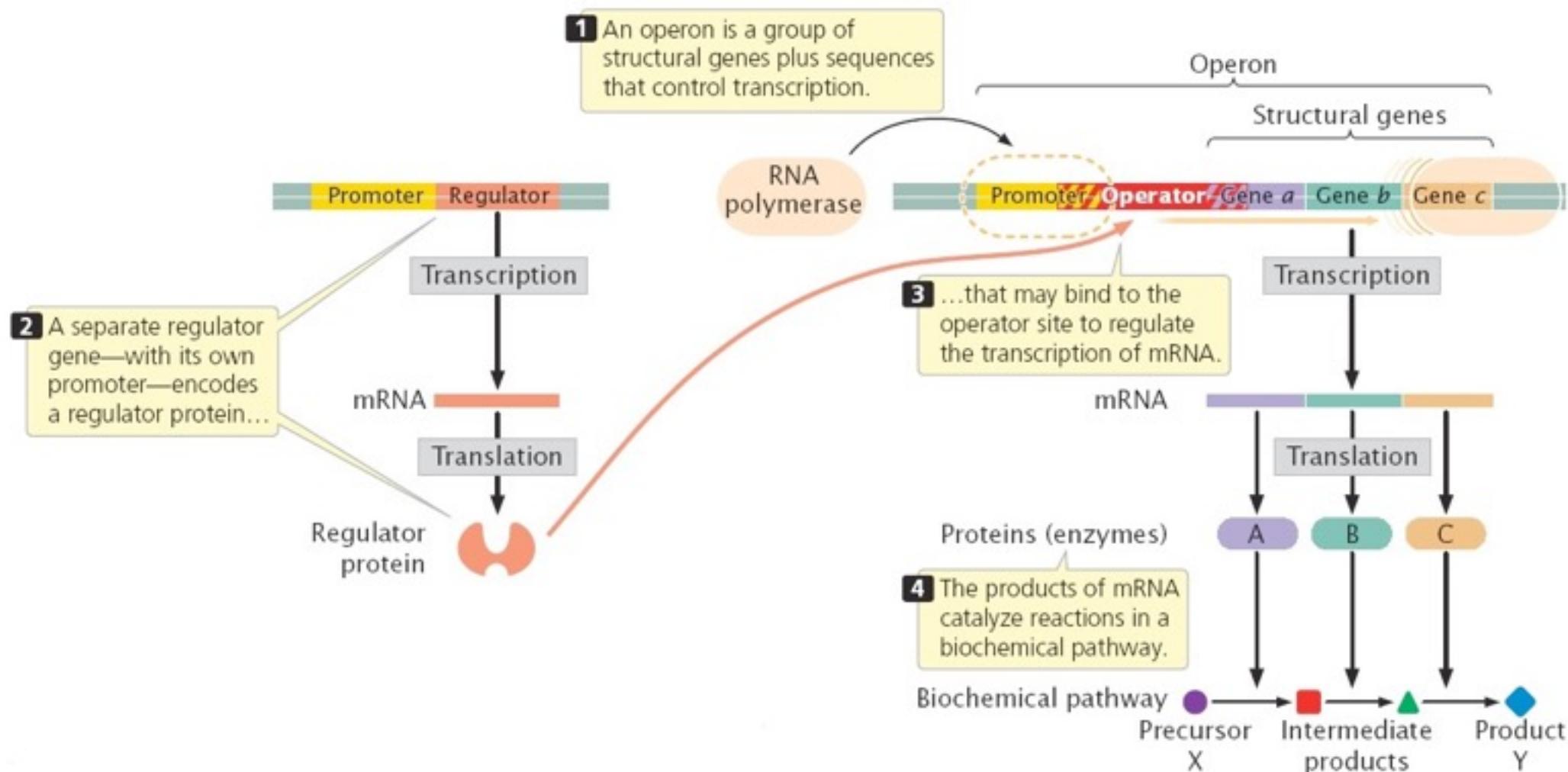
Scott A. Armstrong<sup>1–4</sup>, Jane E. Staunton<sup>5</sup>, Lewis B. Silverman<sup>1,3,4</sup>, Rob Pieters<sup>6</sup>, Monique L. den Boer<sup>6</sup>, Mark D. Minden<sup>7</sup>, Stephen E. Sallan<sup>1,3,4</sup>, Eric S. Lander<sup>5</sup>, Todd R. Golub<sup>1,3,4,5\*</sup> & Stanley J. Korsmeyer<sup>2,4,8\*</sup>

\*These authors contributed equally to this work.

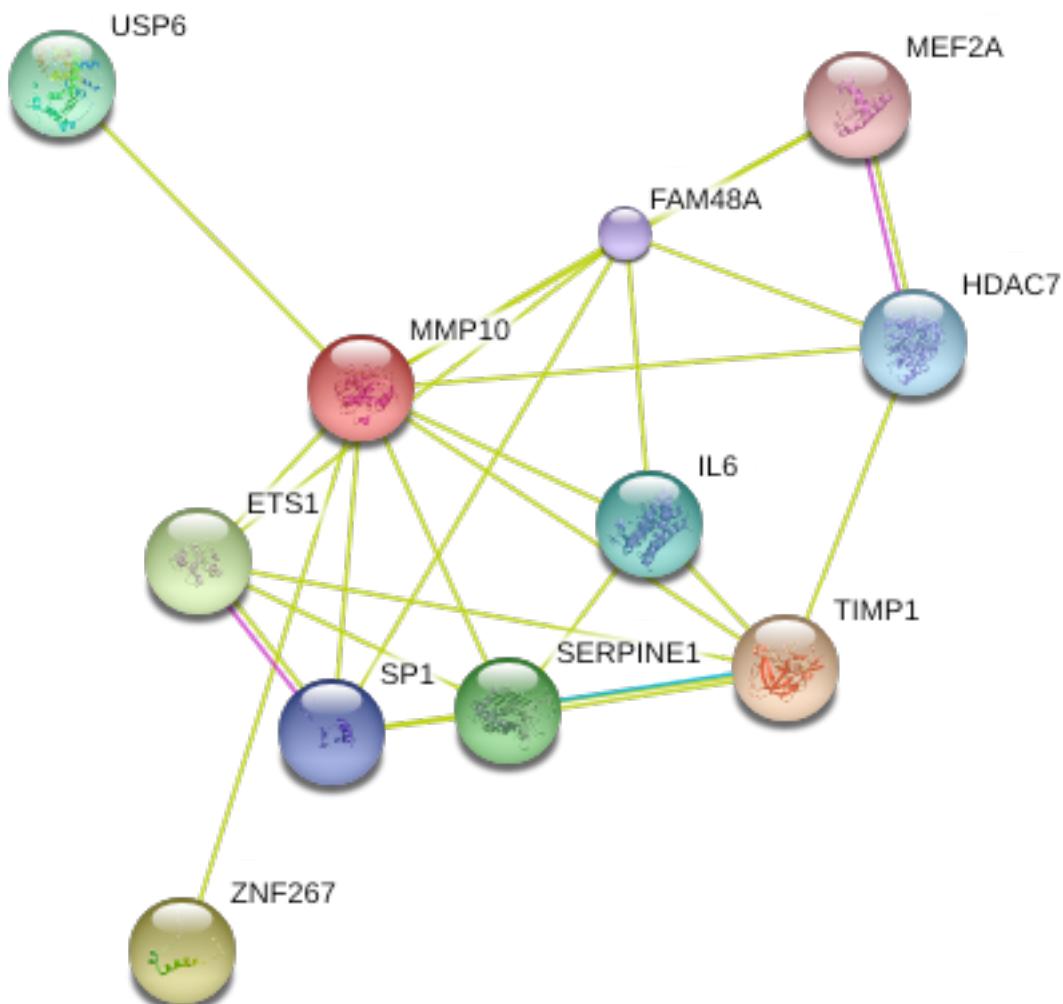
Published online: 3 December 2001, DOI: 10.1038/ng765



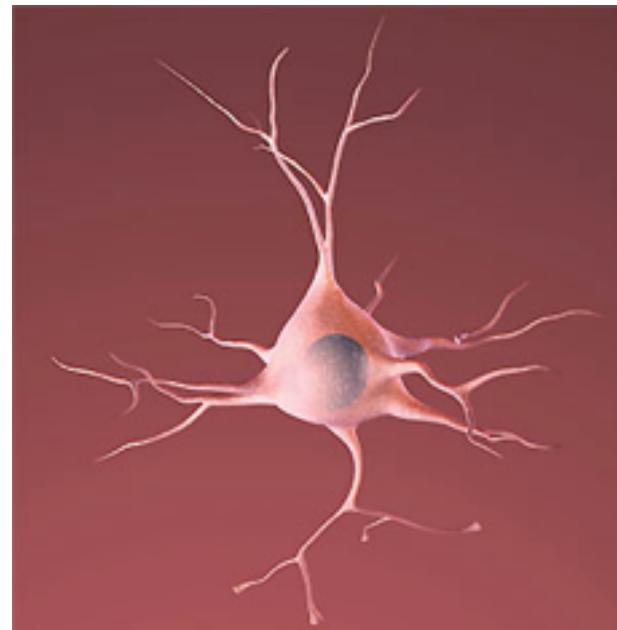
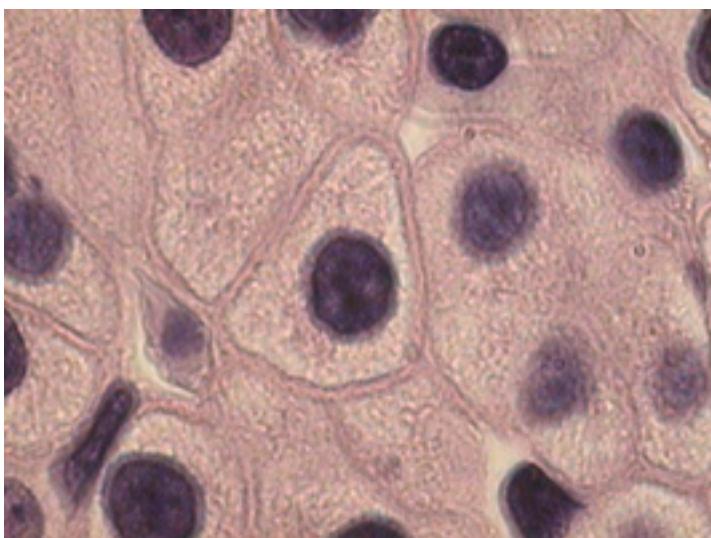
# gene regulation



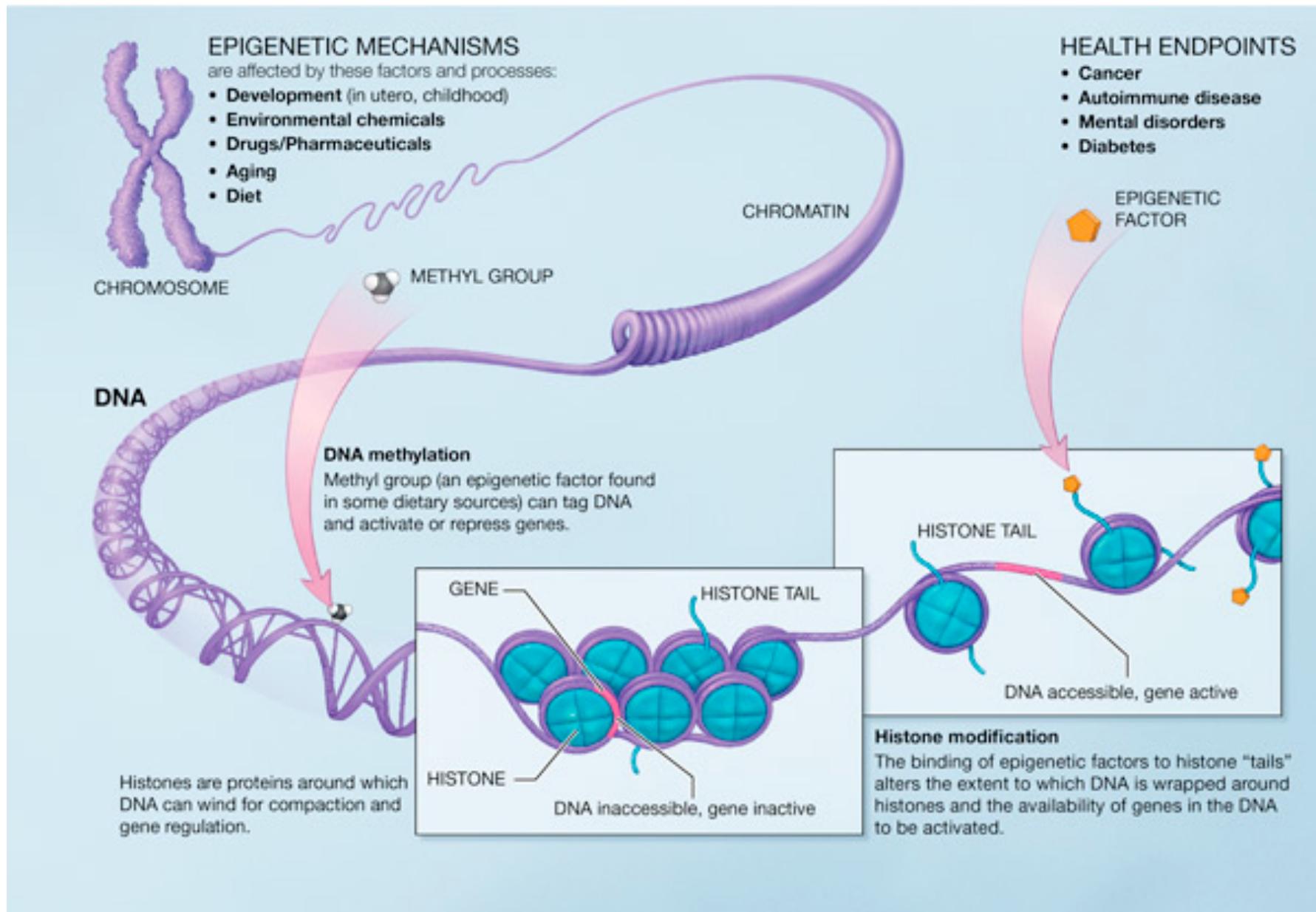
# gene regulation



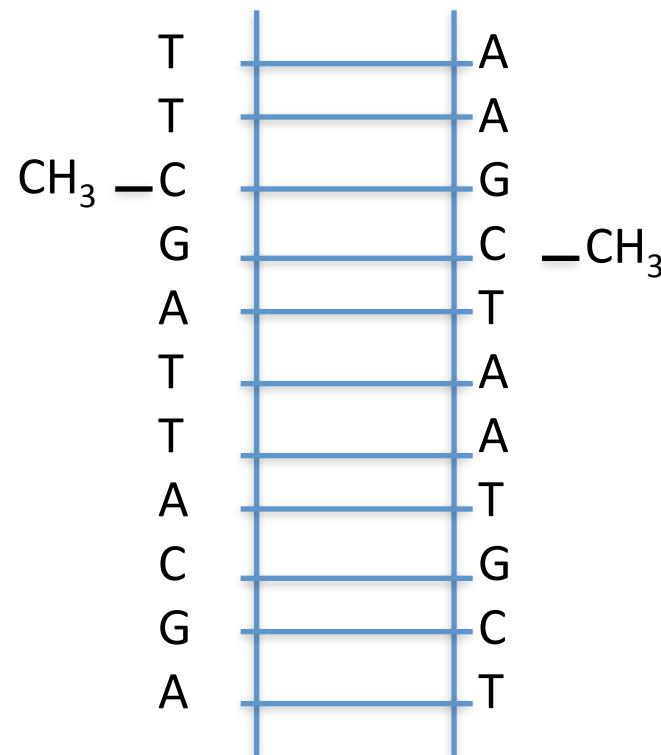
# How many basepair differences?



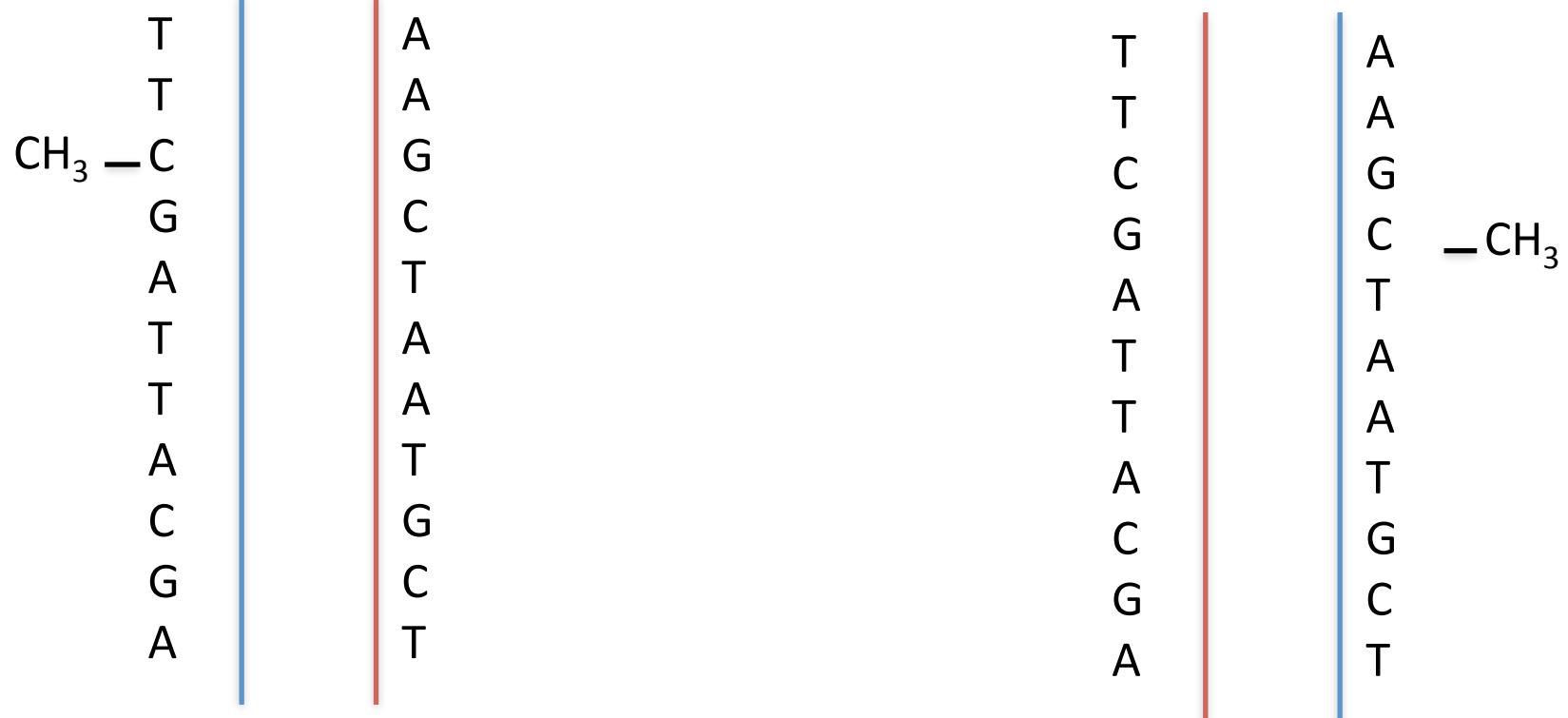
# Epigenetics

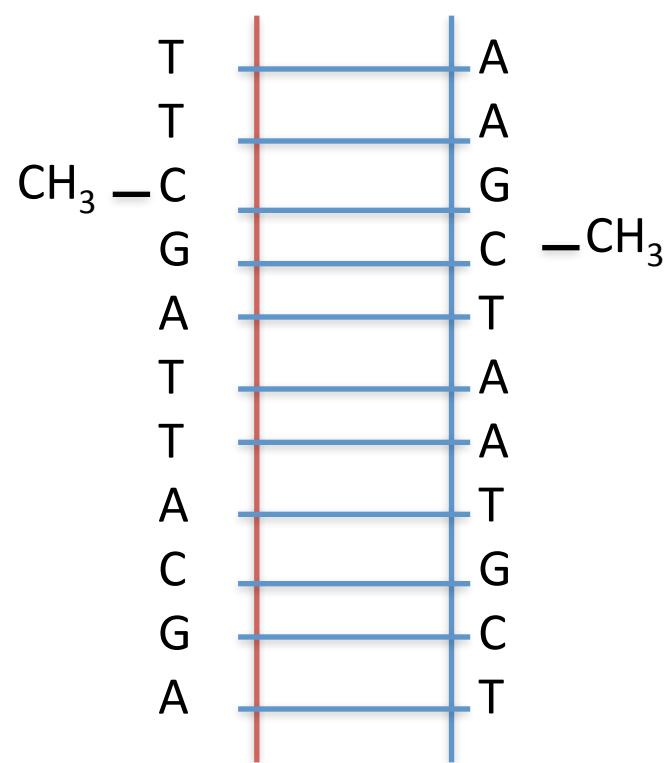
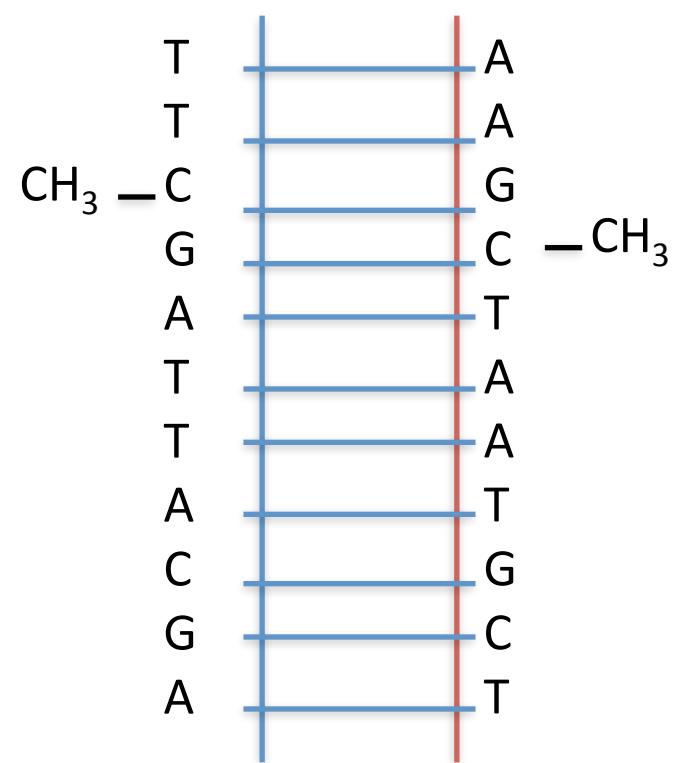


# DNA Methylation









Liver

T	A
T	A
C	G
G	C
A	T
T	A
T	A
A	T
C	G
G	C
A	T

---

Brain

T	A
T	A
C	G
G	C
A	T
T	A
T	A
A	T
C	G
G	C
A	T

Liver

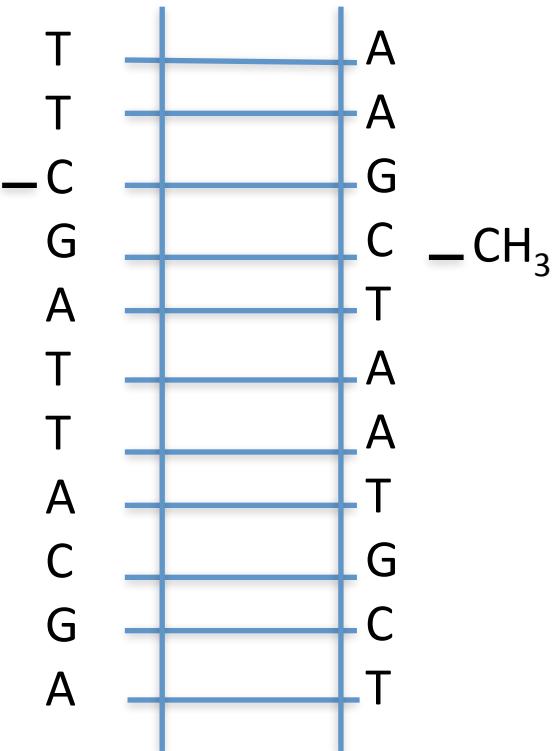
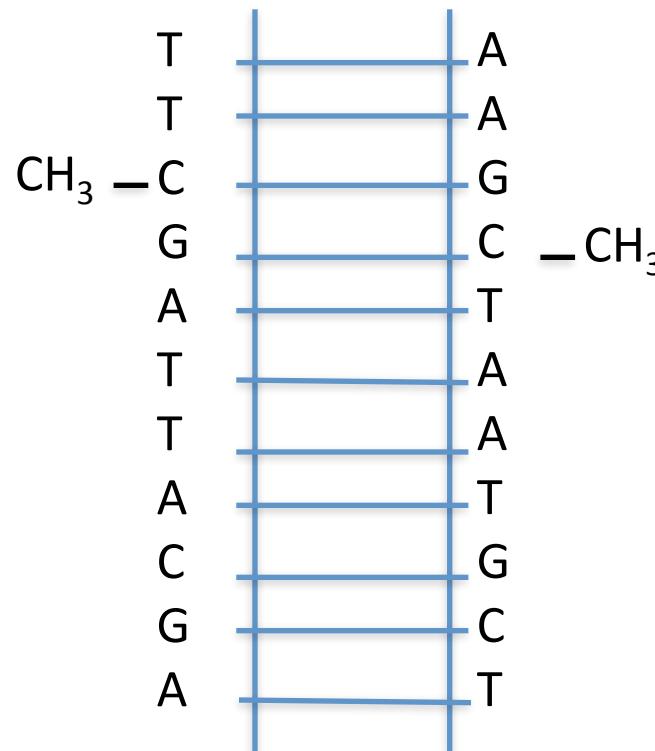
T A  
T A  
CH<sub>3</sub> - C G  
G C - CH<sub>3</sub>  
A T  
T A  
T A  
A T  
C G  
G C  
A T

---

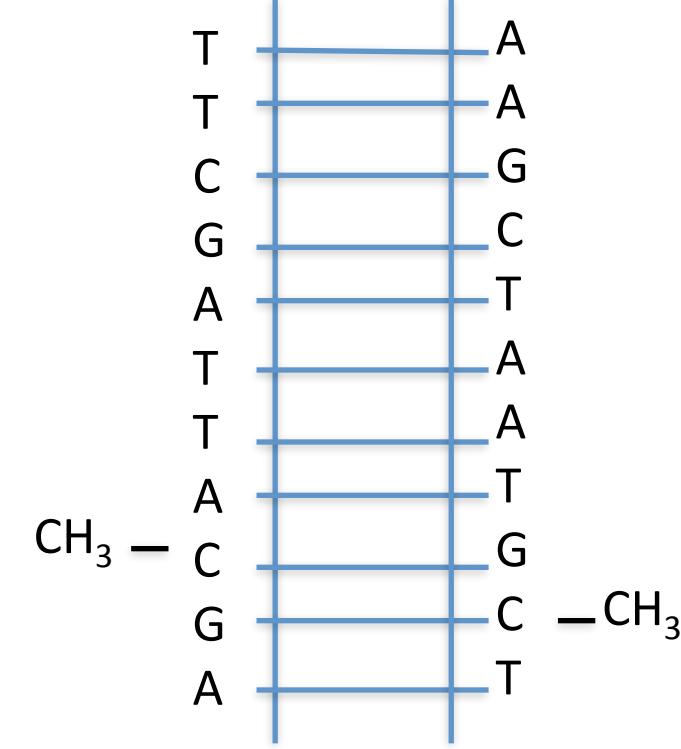
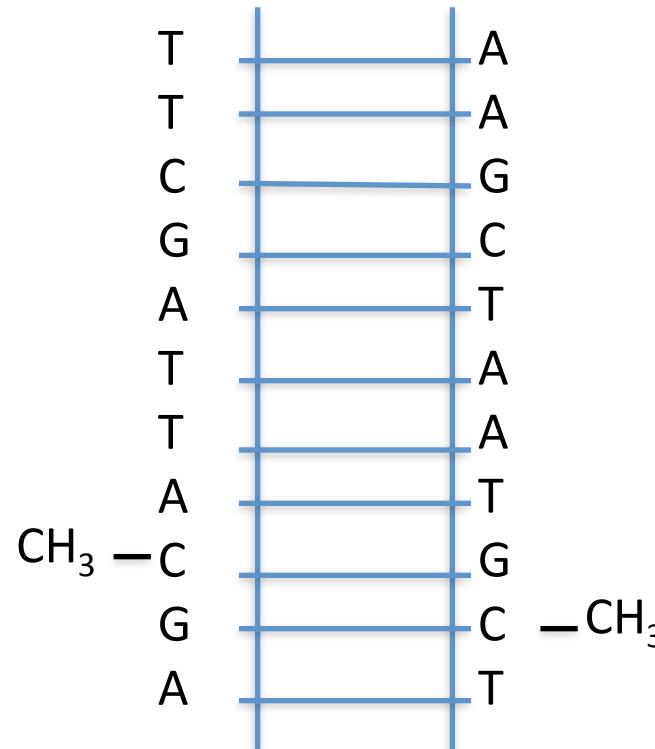
Brain

T A  
T A  
C G  
G C  
A T  
T A  
T A  
A T  
CH<sub>3</sub> - C G  
G C - CH<sub>3</sub>  
A T

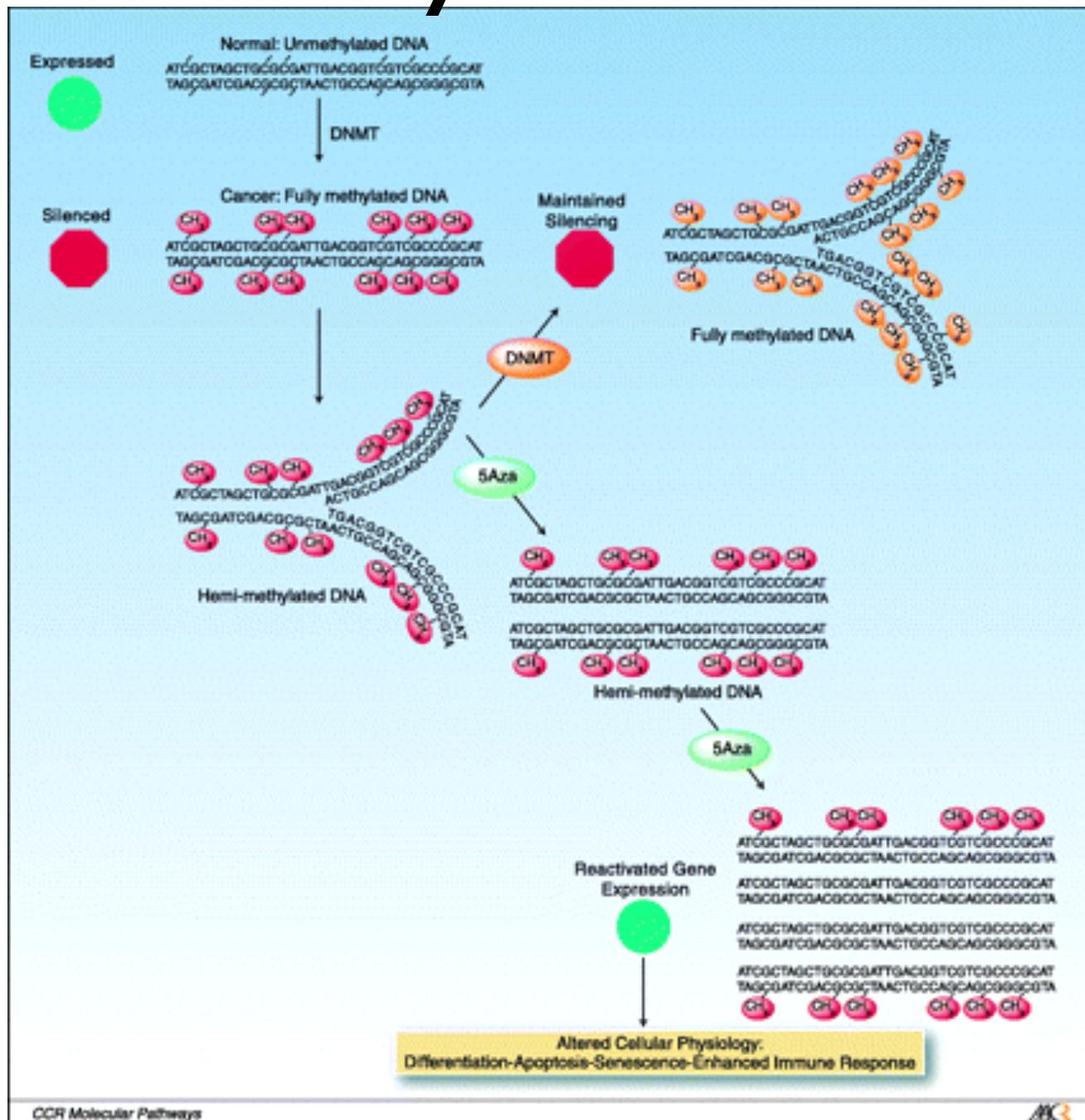
Liver

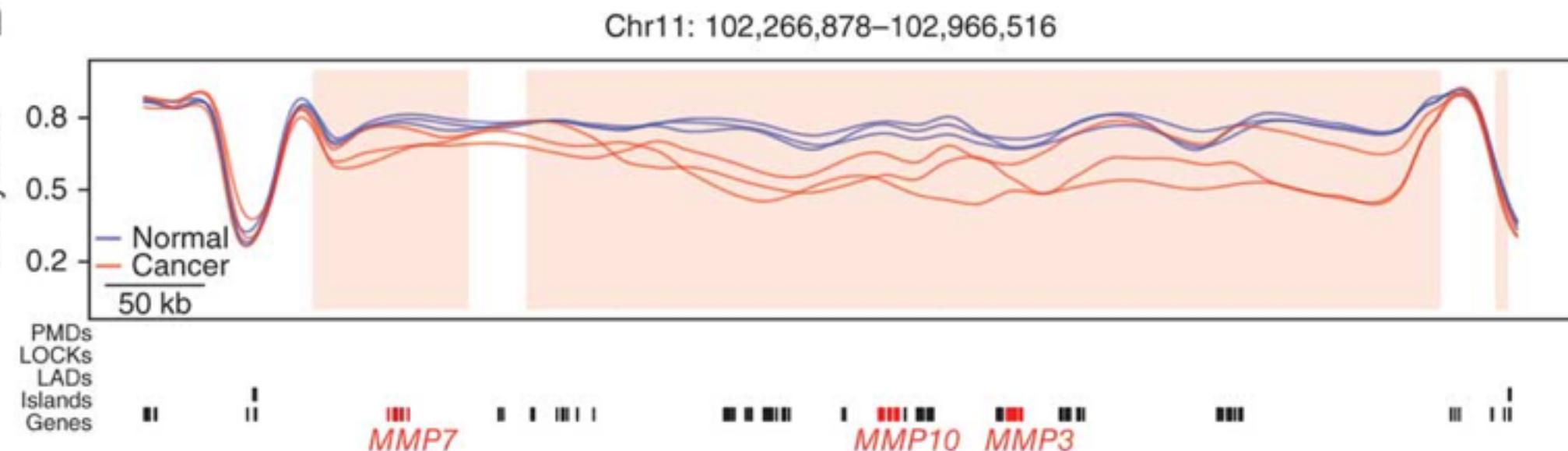


Brain



# DNA Methylation in cancer



**a**

[Hansen et al. Nature Genetics 2011]

