A super quick introduction to molecular biology and genomics

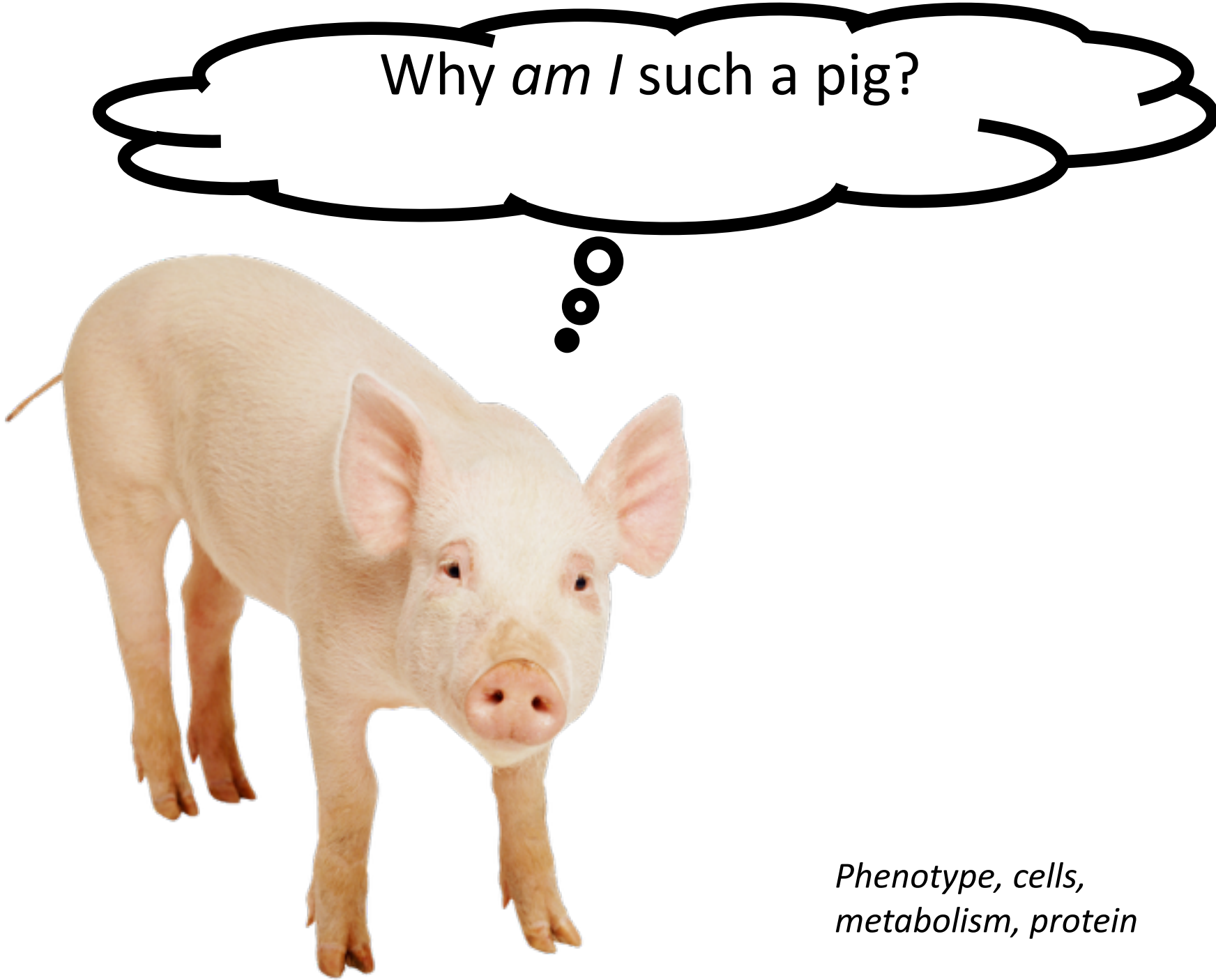
Héctor Corrada Bravo
Dept. of Computer Science
Center for Bioinformatics and Computational Biology
University of Maryland

University of Maryland, Fall 2014

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?



Phenotype, cells, metabolism, protein

Proteins

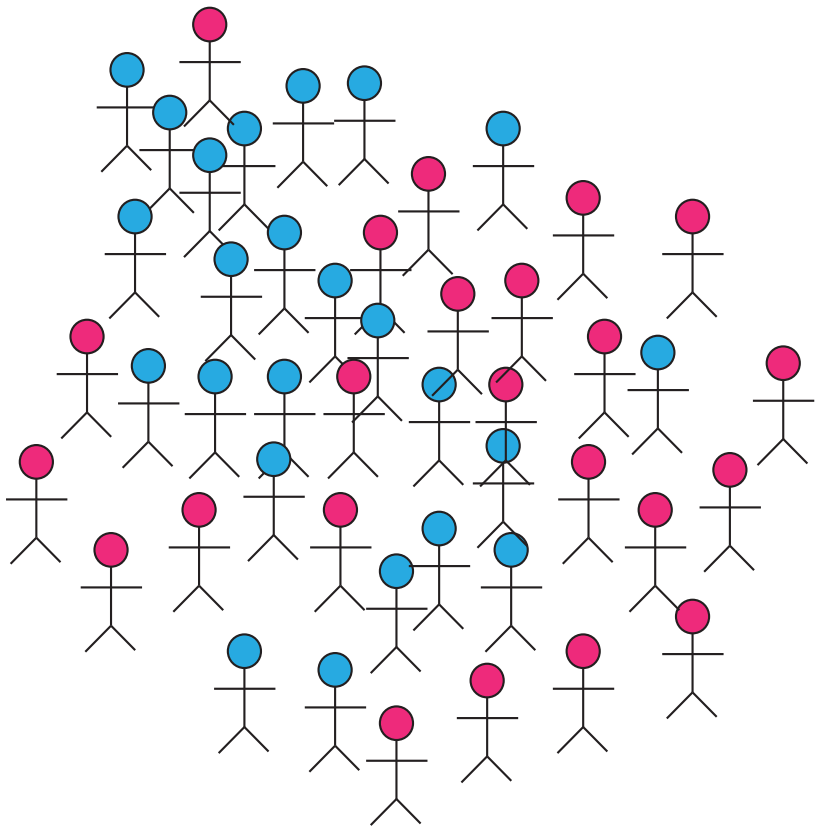
- *phenotype:* characteristics of an organism
- characteristics due to cellular structures and activities
  - mostly carried out by *proteins*
- Examples:

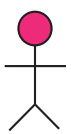
alpha-keratin	*component of hair*
insulin	*regulates blood glucose level*
actin & myosin	*muscle contraction*
hemoglobin	*oxygen transport*
DNA polymerase	*synthesis of DNA*
DNA glycosylases	*DNA repair*
matrix metalloproteinase	*extra-cellular matrix degradation*

Genetics

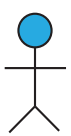
- *gene:* in classical genetics it was an abstract concept
  - a unit of inheritance passed from parent to offspring
  - specify proteins
- *genome* refers to the complete set of *genes*
- *genotype:* genetic characteristics of an individual

What is Genomics?



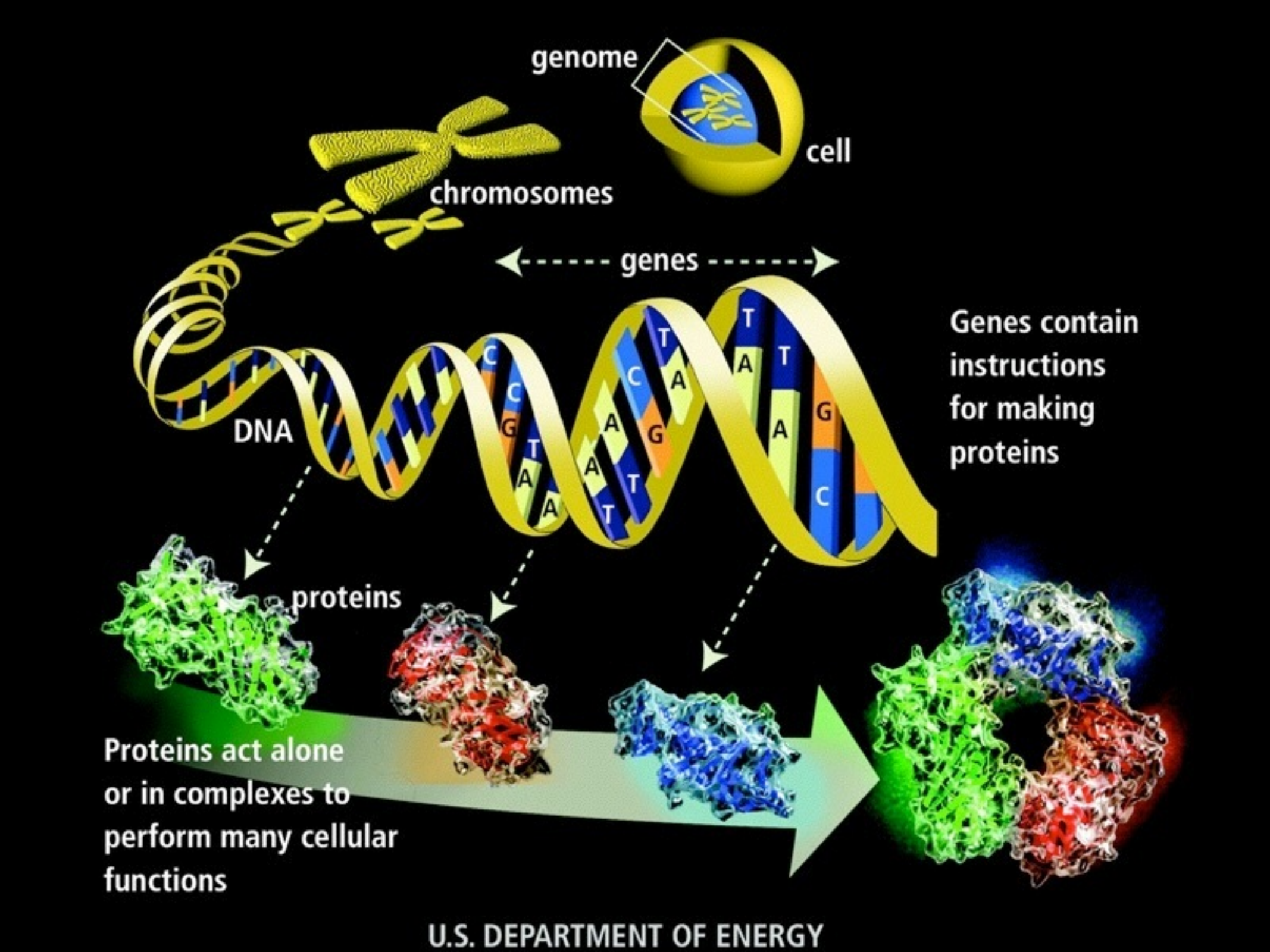


cancer



healthy

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


genome
cell
chromosomes
genes
DNA
Genes contain instructions for making proteins
proteins
Proteins act alone or in complexes to perform many cellular functions
U.S. DEPARTMENT OF ENERGY

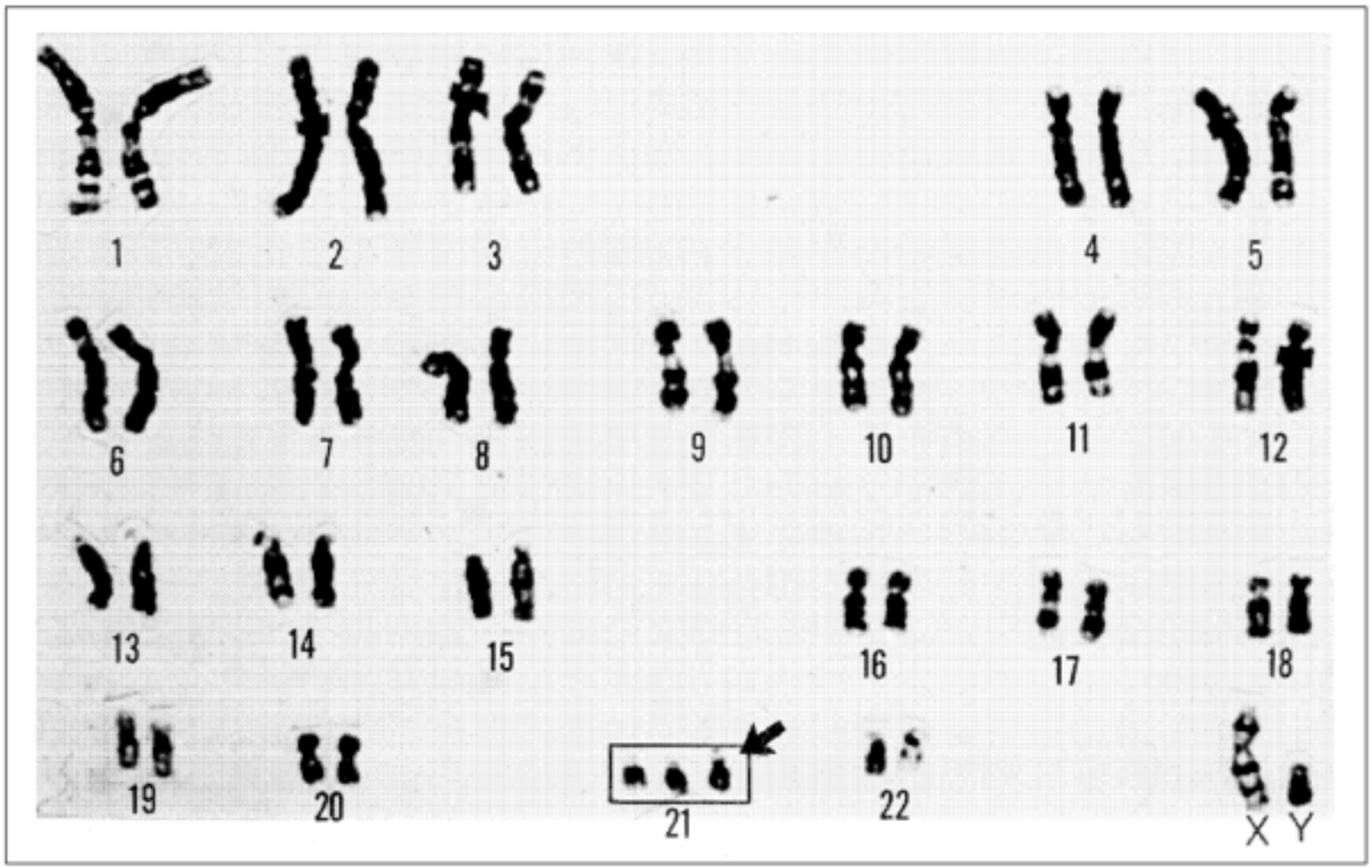
What is Genomics?

- 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.

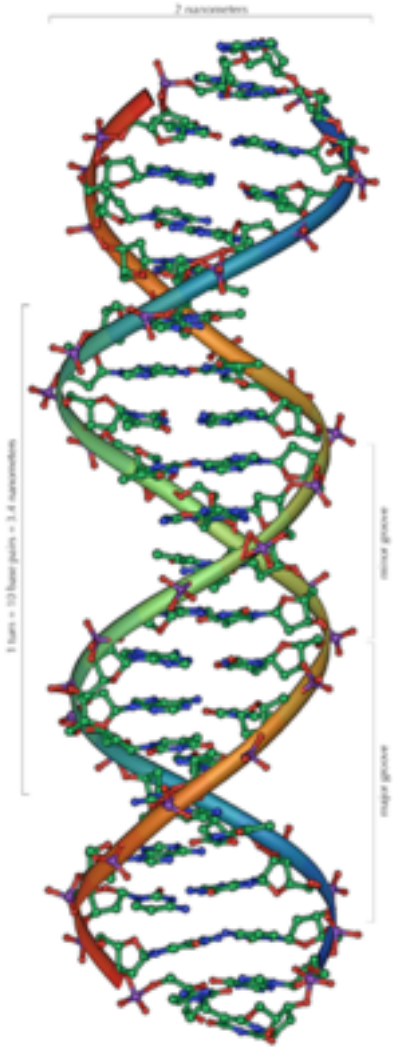
- Cells are of many different types (e.g. blood, skin, nerve cells), but all can be traced back to a single cell, the fertilized egg.

Chromosomes



These are actually human. And for a down syndrome patient

DNA



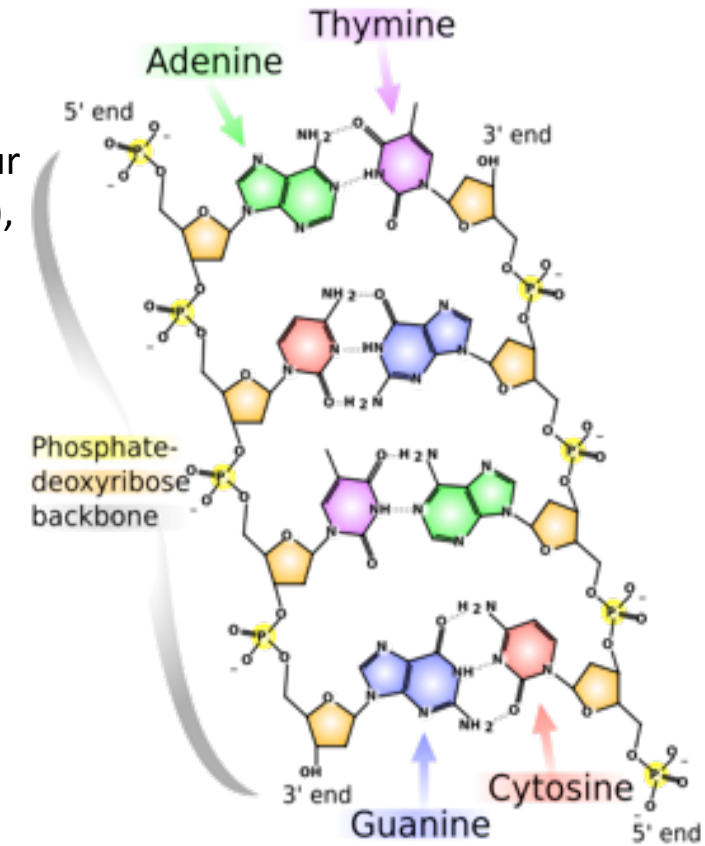
DNAs (Deoxyribonucleic acids) are molecules to store genetic information of a living organism.

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

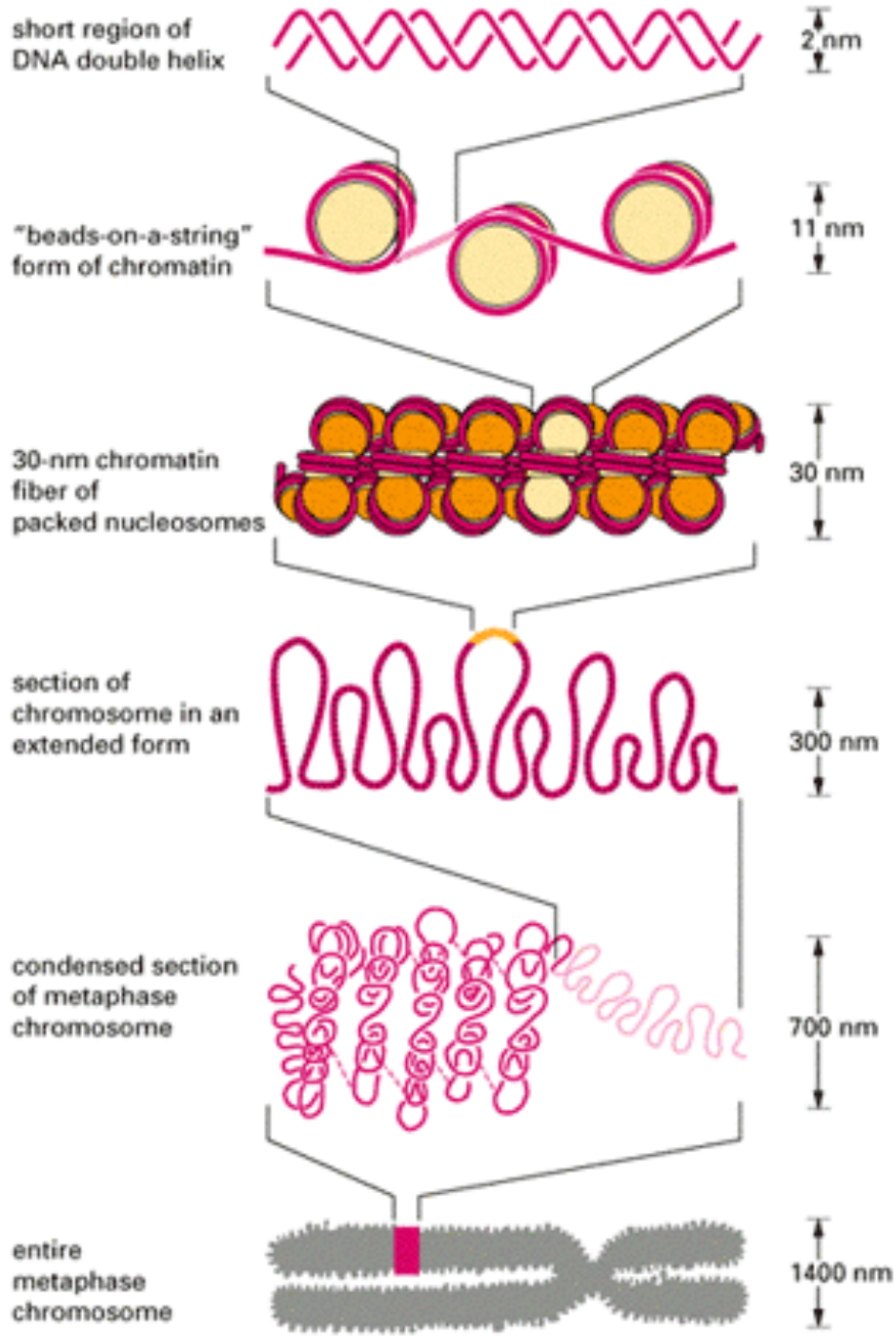
Purines: A, G; Pyrimidines: C, T

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

```
5′-ACCGTTCGACGGTAA-3′
   ||||||||||||||||
3′-TGGCAAGCTGCCATT-5′
```



Watson and Crick 1953


short region of
DNA double helix
2 nm
"beads-on-a-string"
form of chromatin
11 nm
30-nm chromatin
fiber of
packed nucleosomes
30 nm
section of
chromosome in an
extended form
300 nm
condensed section
of metaphase
chromosome
700 nm
entire
metaphase
chromosome
1400 nm

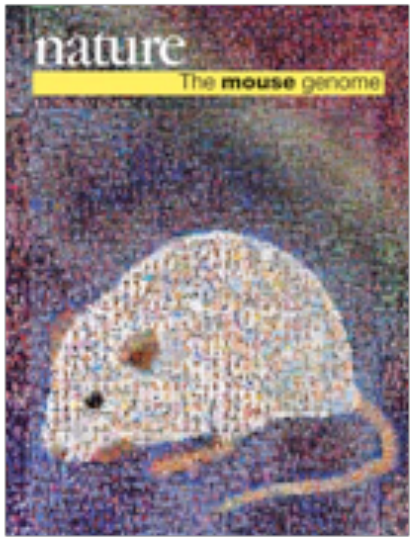
chromatin

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

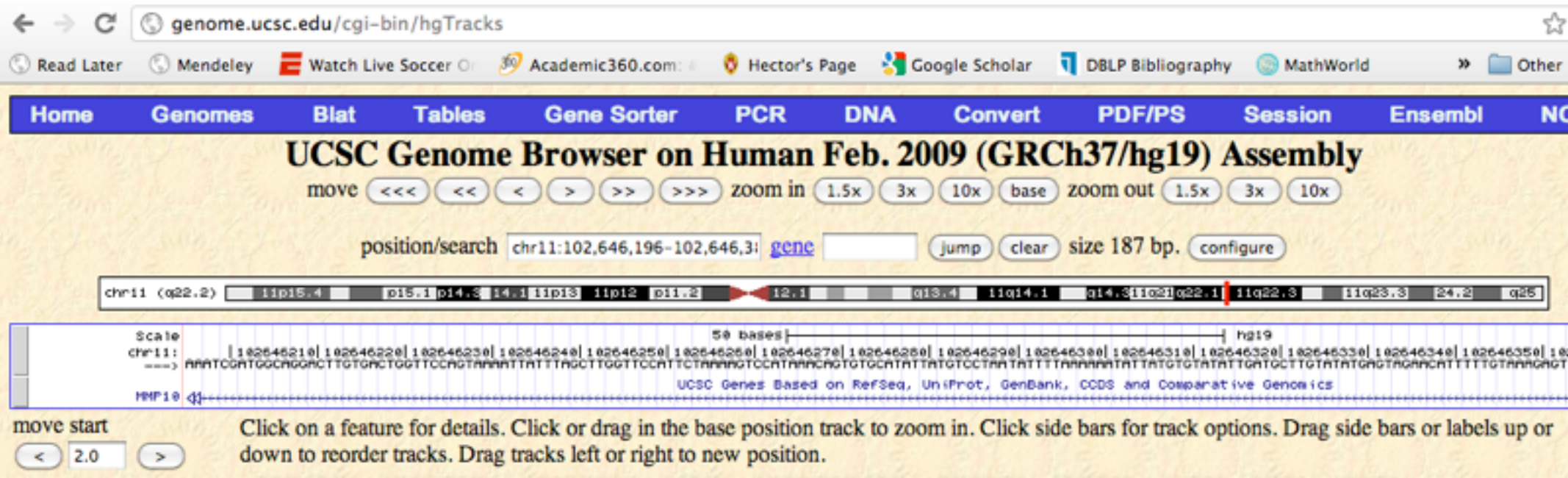


M. musculus, *Nature*, 2002

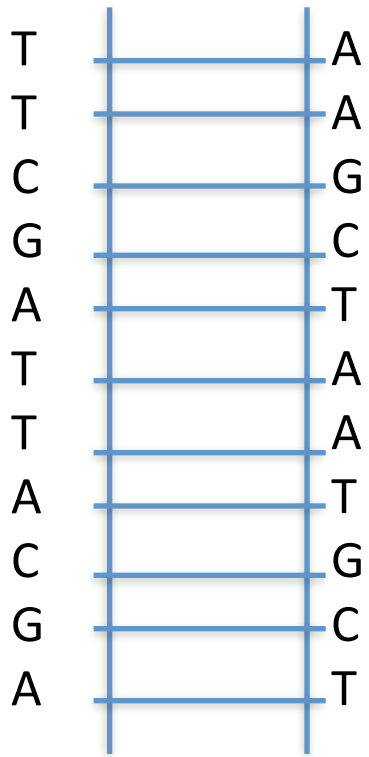
Genome

TCAGTTGGAGCTGCTCCCCCACGGCCTCTCCTCACATTCCACGTCCTGTAGCTCTATGACCTCCACCTTTGAGTCCCTCCTCTCACACCTGAC
ATGAAAAGGCACATGAGGATCCTCAAATACCCCGTGATCAGTCTCAGGGTAGCTCTCATAGCCTGGACAGGGCCCCCCTCGGGGGTTGCGCCC
AGGTCCAGGCGGGGGATGCACAGCAACAGTCACCGAAGCAGAAGCCGTCACAGTGGTGATGGGCTGGCAGTAGCTGGGCACAGAGCTGCCCAT
GGCGGTGGACGTTGGGTTCCGAGGGTTGTGAGAACGGGCCCCACGGGGCCCTGAGCGGTCCCTATTGCTAGGGCCAGAATGCCCTTCAGTAGA
AATTTCAAAAGCGTCTCTGCGCGGTCTGTAGGGGGGTGGCCGCAAGCCTTCTCTAGGGGGATCCCTTCGAGGCTGCTGGCCTTGCCGTCCAGG
GGACAAGGAGCCAGAGTCCAGGTGGGGCTGTTGCCGAGGGGTCAAGGGAGGCTGATGTCTGGAGTCCGGATGGACCACCTGCAGAGGAGAGAC
ATAGGTCAACACAGGGAGGTAGGATGGTGGTGATGTTCCACCCACAAAAGAAAACCTATTCCTTTAGAAACCTCCAGGATGTGAATCCTGCCT
GCACCTGCACAGCTGGCTGGAGGCATATAGCCACTGCCCATAGATCTCAACTTACCCTCACAACCAACTGCCCCCAGGCCTAAGTTCTCTGCC
TCAAAACTGCCAAGGCCTGGATAGCCAAGAGCCTGGGTGTCTTGGAAATATGCAACCATAAATAGTAGCTTTTAGAAGTATAAGGCTCCTGTT
TCTGGGTCATATTAGTGTTGTTTTCACCTGTCCCCAGCCCTAAGCCAGGTGTGGCCAGAAGCAAATGTACTGTAAGAGCAGAGCAAAAACTTC
CACACAGATAGTTCTGTTAGGCAATACATCTCTGCCTGACTATTAGGAATCTGGTTTCTGGGTCCTCTGTACAAAGCTCGGAGCAACACAGTG
GCCACATCAATCAAAAGGACCGTGACCAACTTCAAAGTCGGTGAGCTTGTACCTATTTTTAGGCTCCTGCTGAACAGAACCAGATTCACACTA
CAGCTCAGCAGGGCATCGTCACGGGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTGTTGGGGGGGGGGGGGTGGACAGAGGACGGGGAC
ACAATTCACTGGCCAGCCCTTCTCTCCTTCAAGGAAGGCTGCTCTAGCCTGGGACTGGAATACACATTTCCTGTAAACATGGTGGGGGCCTCA
GGCAAGCCAGAGTTTTGGAGCCTTCCTTAACTCTTCAAGGTGAGCATCTTGACTTGGAGGGTGGGGGTGCGGGTAAGGAAGGAACCTGTGGAC
TCCTCCCTACAAGACAGAAAAGGAATAAGCCACGAAGACAATAACGATTTTTGTATCAAGCGTCCTCTCCCATTTCAGCTTACCTGACAATGA
AATCAAATTCGGACCCTGCAAGCATCAGTACACCCAGCAGAGTGGACACAGCACCGTCCAGAACGGGAGCAAACATGTGCTCCAGAGCGAGCA
TAGCCCTGTGGTTCTTGTCCCCAATGGCTGTCAGAAAGGCCTGAACAAAGGAGAAAATTGACACGGTCACATTCTGGGTGTGGTAAAGTGCTC
AGCTGTGTCTATACTTGGGTTTTGTAT…

Total amount of DNA in human genome:
$3 * 10^9$ base pairs (bp)

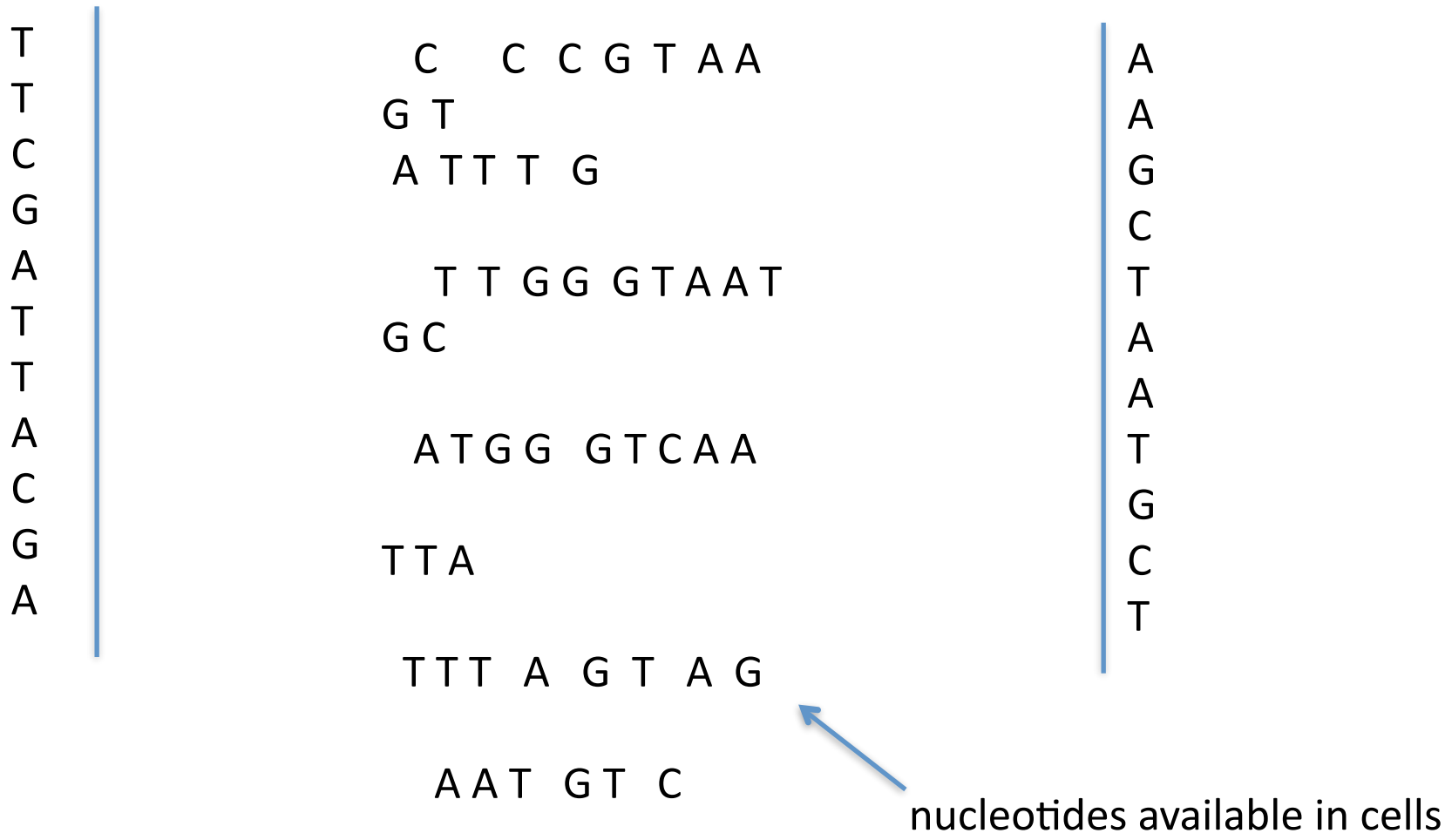

genome.ucsc.edu/cgi-bin/hgTracks
Read Later
Mendeley
Watch Live Soccer
Academic360.com
Hector's Page
Google Scholar
DBLP Bibliography
MathWorld
Other
Home
Genomes
Blat
Tables
Gene Sorter
PCR
DNA
Convert
PDF/PS
Session
Ensembl
UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Assembly
move
<<<
<<
<
>
>>
>>>
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)
Scale
chr11:
50 bases
hg19
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
MMP10
move start
2.0
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.

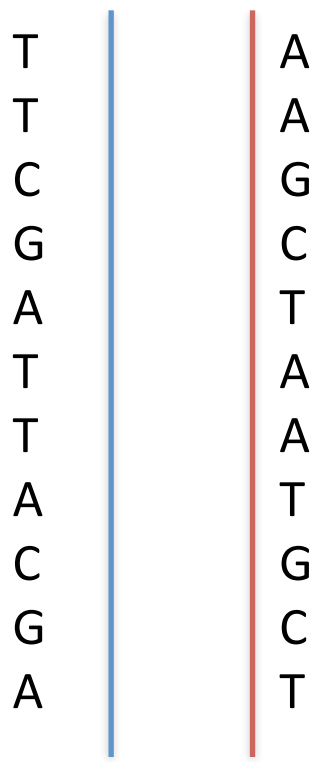
Replication

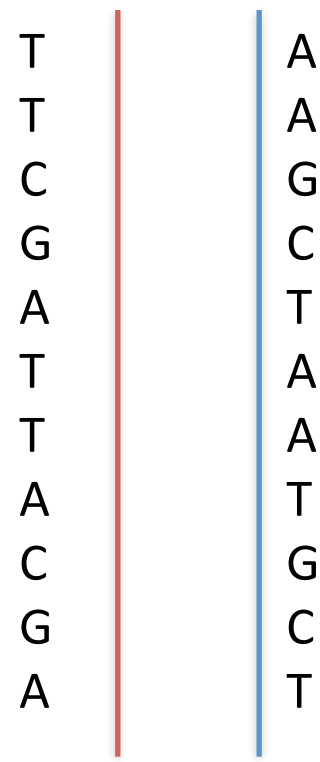


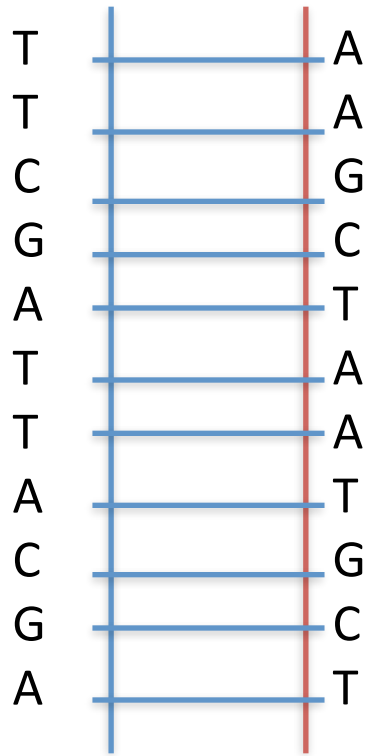
T
T
C
G
A
T
T
A
C
G
A

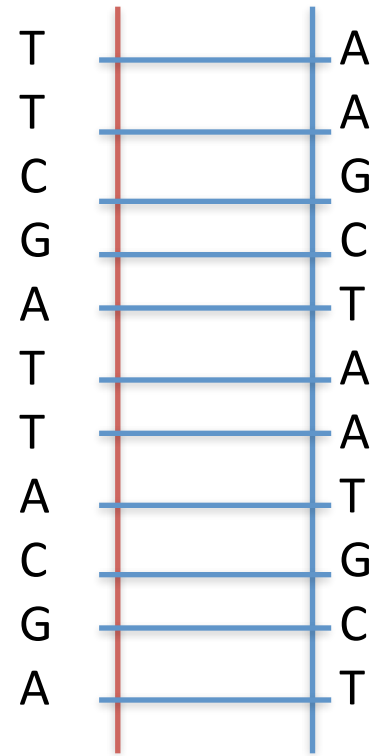
A
A
G
C
T
A
A
T
G
C
T


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
T T A
T T T 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

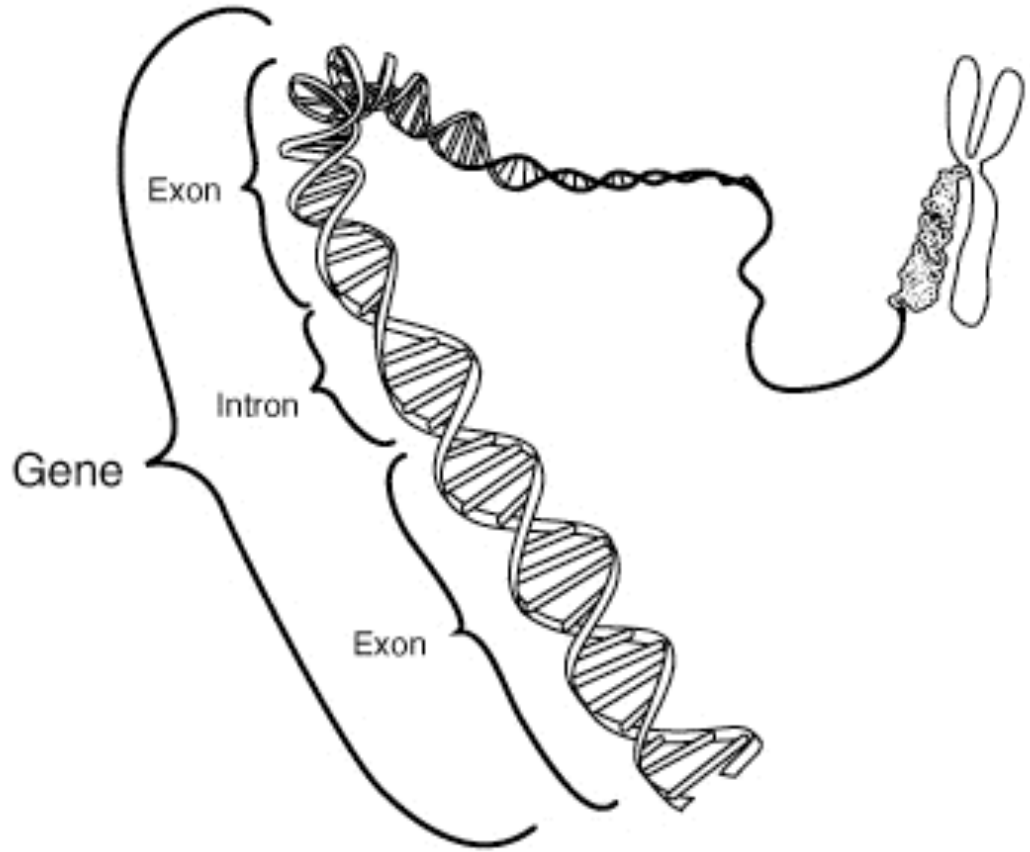

T A
T A
C G
G C
A T
T A
T A
A T
C G
G C
A T

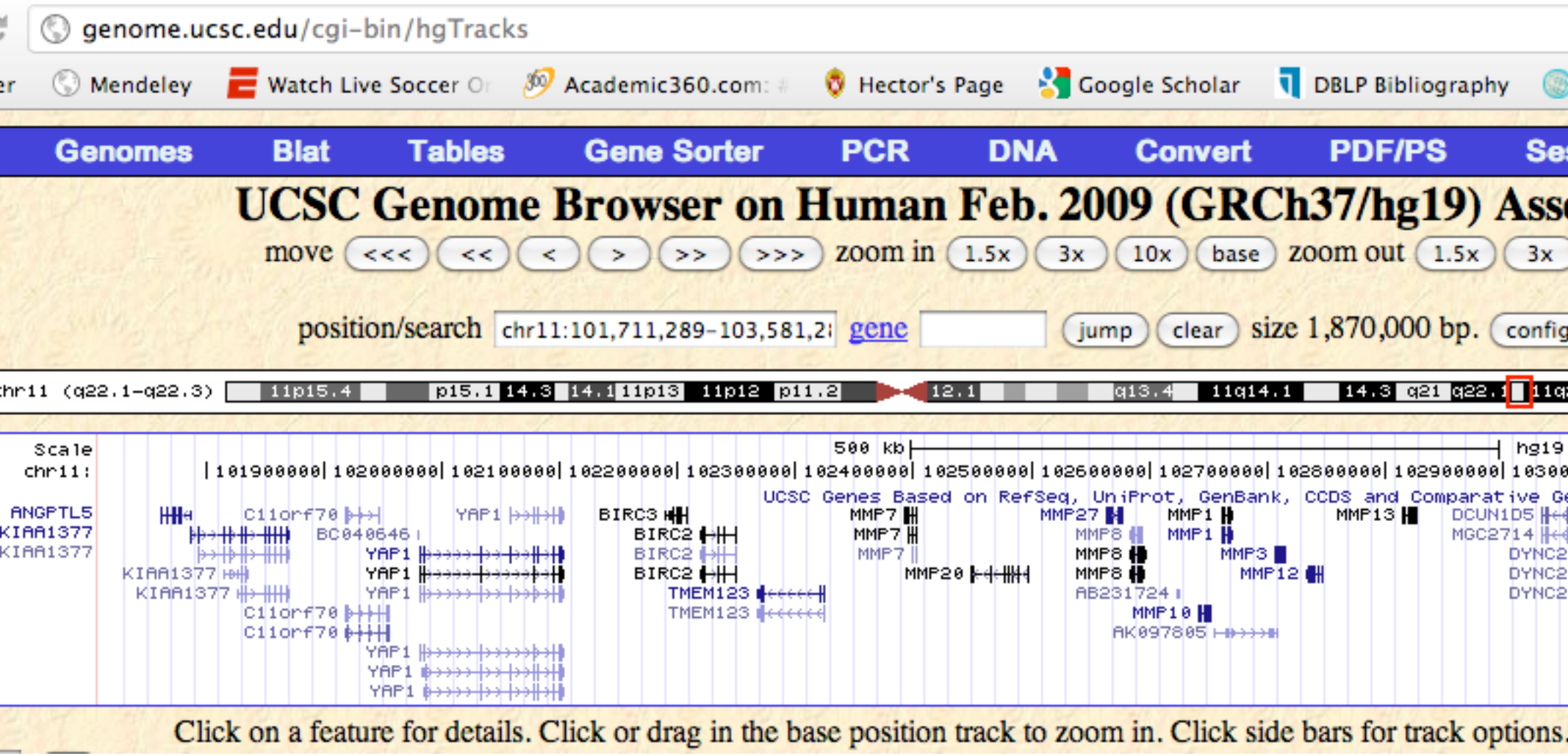
Why are these two different?

Differences explained by 1-10% difference in genome

Similarities explained by similar genes

Genes

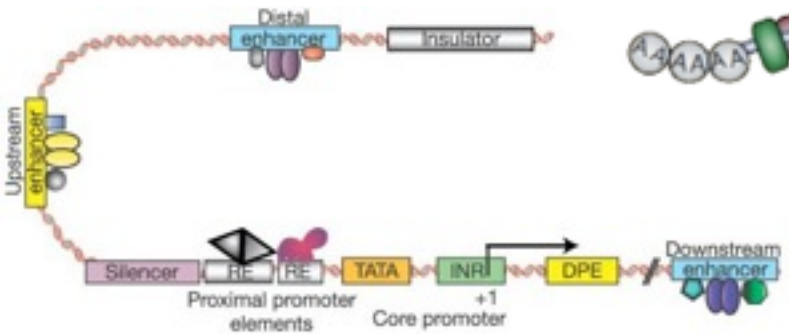


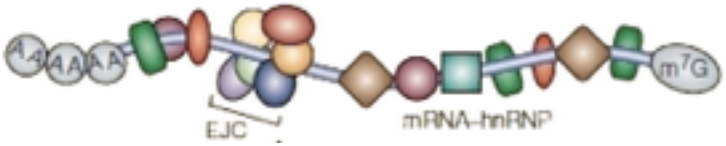

genome.ucsc.edu/cgi-bin/hgTracks
Mendeley
Watch Live Soccer
Academic360.com
Hector's Page
Google Scholar
DBLP Bibliography
Genomes
Blat
Tables
Gene Sorter
PCR
DNA
Convert
PDF/PS
UCSC Genome Browser on Human Feb. 2009 (GRCh37/hg19) Ass
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x
position/search chr11:101,711,289-103,581,2 gene jump clear size 1,870,000 bp. config
hr11 (q22.1-q22.3) 11p15.4 p15.1 14.3 14.1 11p13 11p12 p11.2 12.1 q13.4 11q14.1 14.3 q21 q22.1 11q
Scale
chr11:
500 kb
hg19
101900000 102000000 102100000 102200000 102300000 102400000 102500000 102600000 102700000 102800000 102900000 10300
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Ge
ANGPTL5
KIAA1377
KIAA1377
C11orf70
BC040646
YAP1
KIAA1377
KIAA1377
C11orf70
C11orf70
YAP1
YAP1
YAP1
YAP1
YAP1
YAP1
BIRC3
BIRC2
BIRC2
BIRC2
TMEM123
TMEM123
MMP7
MMP7
MMP7
MMP20
MMP27
MMP8
MMP8
MMP8
AB231724
MMP10
AK097805
MMP1
MMP1
MMP3
MMP12
MMP13
DCUN1D5
MGC2714
DYNC2
DYNC2
DYNC2
Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options.

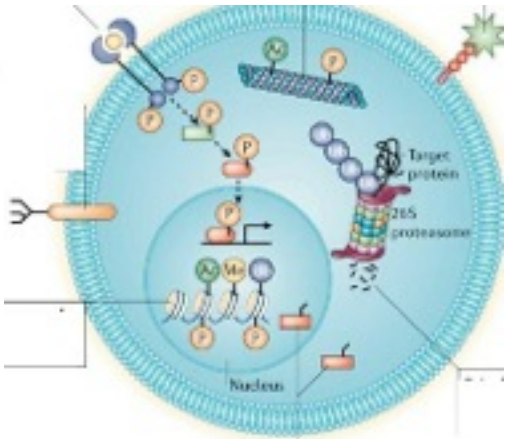
Central Dogma

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

genomics transcriptomics proteomics







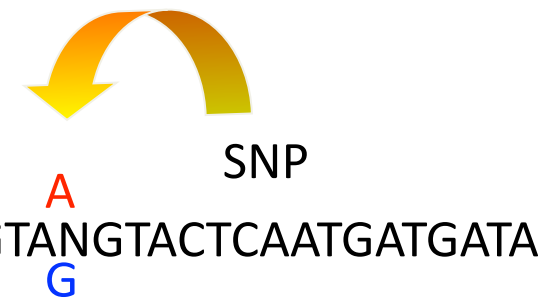
Major technological advances allow **unprecedented** data acquisition

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

TACATAGCCATCGGTAAGTACTCAATGATGATA
ATGTATCGGTAGCCATTCATGAGTTACTACTAT

Father

TACATAGCCATCGGTAGGTACTCAATGATGATA
ATGTATCGGTAGCCATCCATGAGTTACTACTAT

GG

Mother

TACATAGCCATCGGTAGGTACTCAATGATGATA
ATGTATCGGTAGCCATCCATGAGTTACTACTAT

Father

TACATAGCCATCGGTAGGTACTCAATGATGATA
ATGTATCGGTAGCCATCCATGAGTTACTACTAT

From reads to evidence

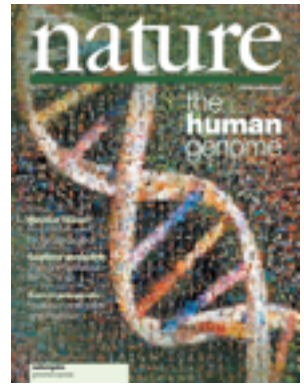
From reads to evidence

1. Comparative

Sequence-wise, individuals of a species are nearly identical

Well curated, annotated "reference" genomes exist

D. melanogaster, *Science*, 2000

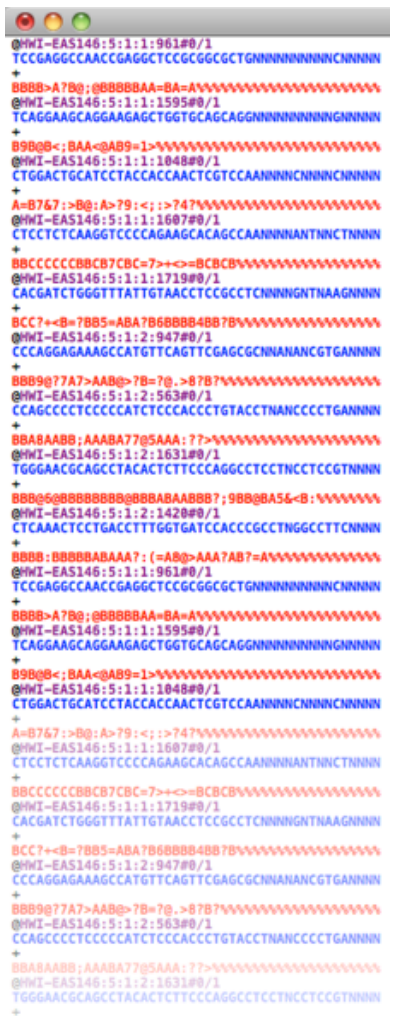


H. sapiens, *Nature*, 2000 and *Science*, 2000

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



Align

```
GTCGCAGTANCTGTCT
|||||||||  ||||||
GTCGCAGTATCTGTCT

GGATCTGCGATATACC
||||||  ||||||||||
GGATCT-CGATATACC

AATCTGATCTTATTTT
||||||||||||||||
AATCTGATCTTATTTT

ATATATATATATATAT
||||||||||||||||
ATATATATATATATAT

TCTCTCCCANNAGAGC
|||||||||  |||||
TCTCTCCCAGGAGAGC
```

"Pileup" or "Coverage plot"

"Depth of coverage" = 14

Call: HET A, G

```
      GTCGCAGTATCTGTCT
     GTCGCAGTATCTGTNN
    TGTCGCAGTATCTGTC
  TATGTCGCAGTATCTG
  TATATCGCAGTATCTT
  TATATCGCAGTATCTG
  NATATCGCAGTATNTG
 CCCTATATCGCAGTAT
ACACCCTATGTCGCA
ACACCCTATCTCGCA
ACACCCTATGTCGCA
GA-CACCCTATGTCGC
CCGGA-CACCCTATAT
CCGGA-CACCCTATAT
GCCGGA-CACCCTATG
```

"Coverage"

GTATTTTTCGTCTGGGGGGTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCTATGTCGCAGTATCTGTCTTTGATTCCTGCCTCATCCTATTATTTATCGC

Reference

(slide courtesy of Ben Langmead)

/cgi-bin/hgTracks?hgsid=240284685

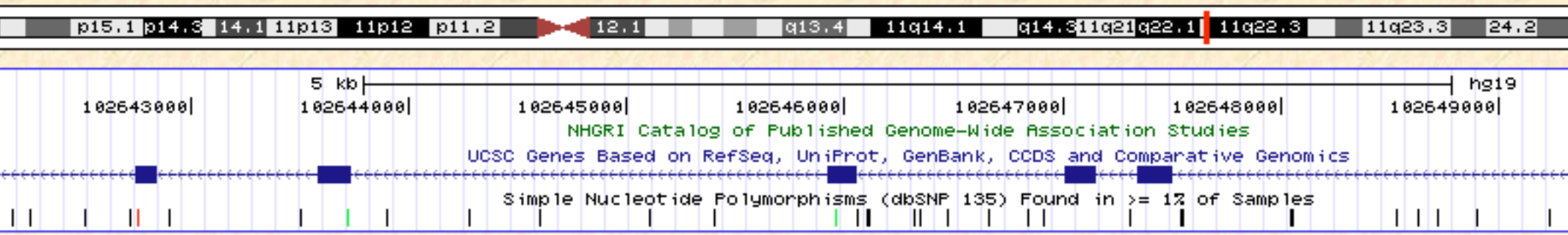
tch Live Soccer O | Academic360.com: # | Hector's Page | Google Scholar | DBLP Bibliography | MathWorld | »

lat | Tables | Gene Sorter | PCR | DNA | Convert | PDF/PS | Session | Ensembl

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

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

position/search chr11:102,641,234-102,651,3· 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
eorder tracks. Drag tracks left or right to new position.

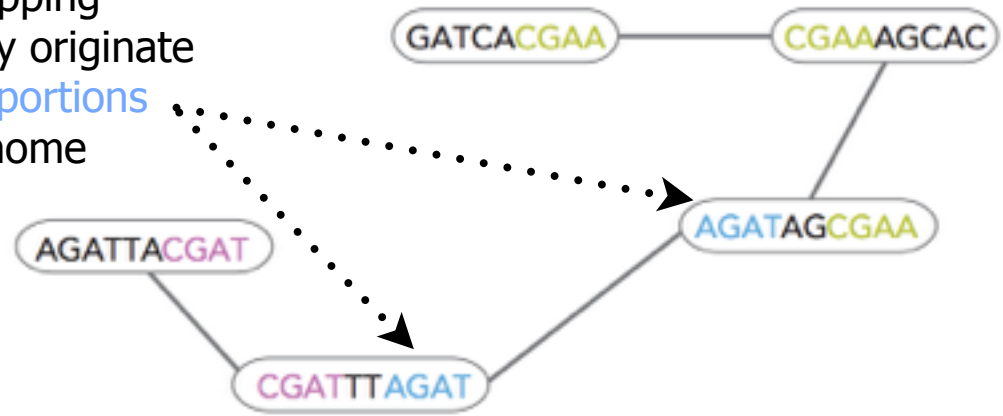
From reads to evidence

2. *de novo*

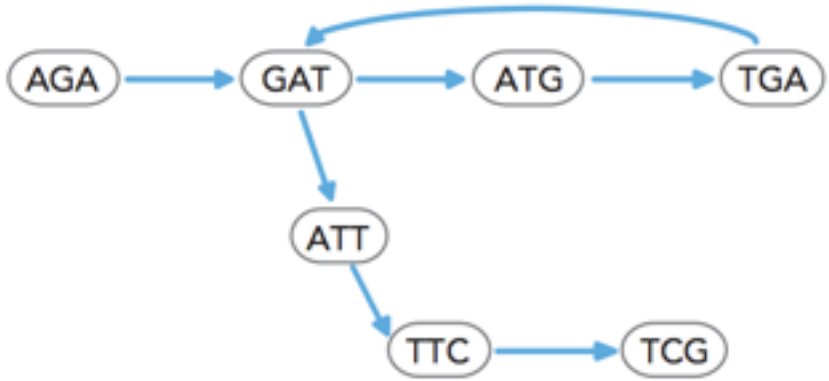
Assume nothing! - let reads tell us everything

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

Encode overlap relationships as a graph



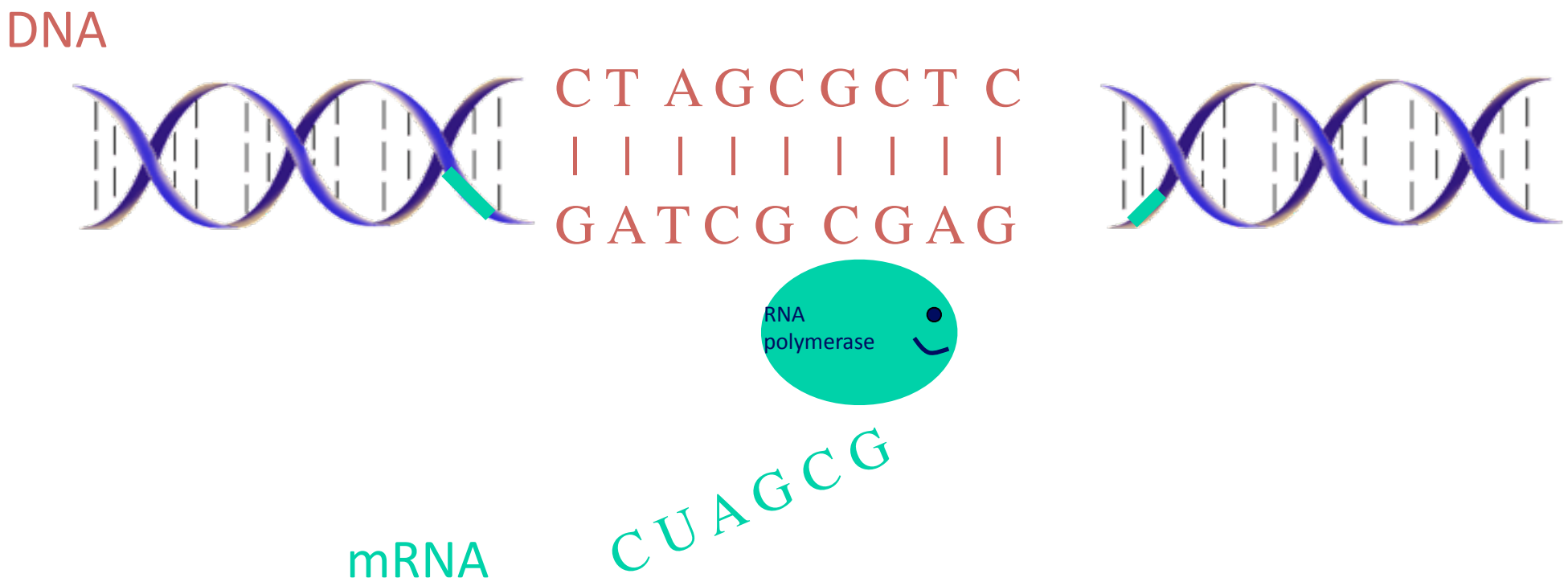
Source: De Novo Assembly Using Illumina Reads. Illumina. 2010

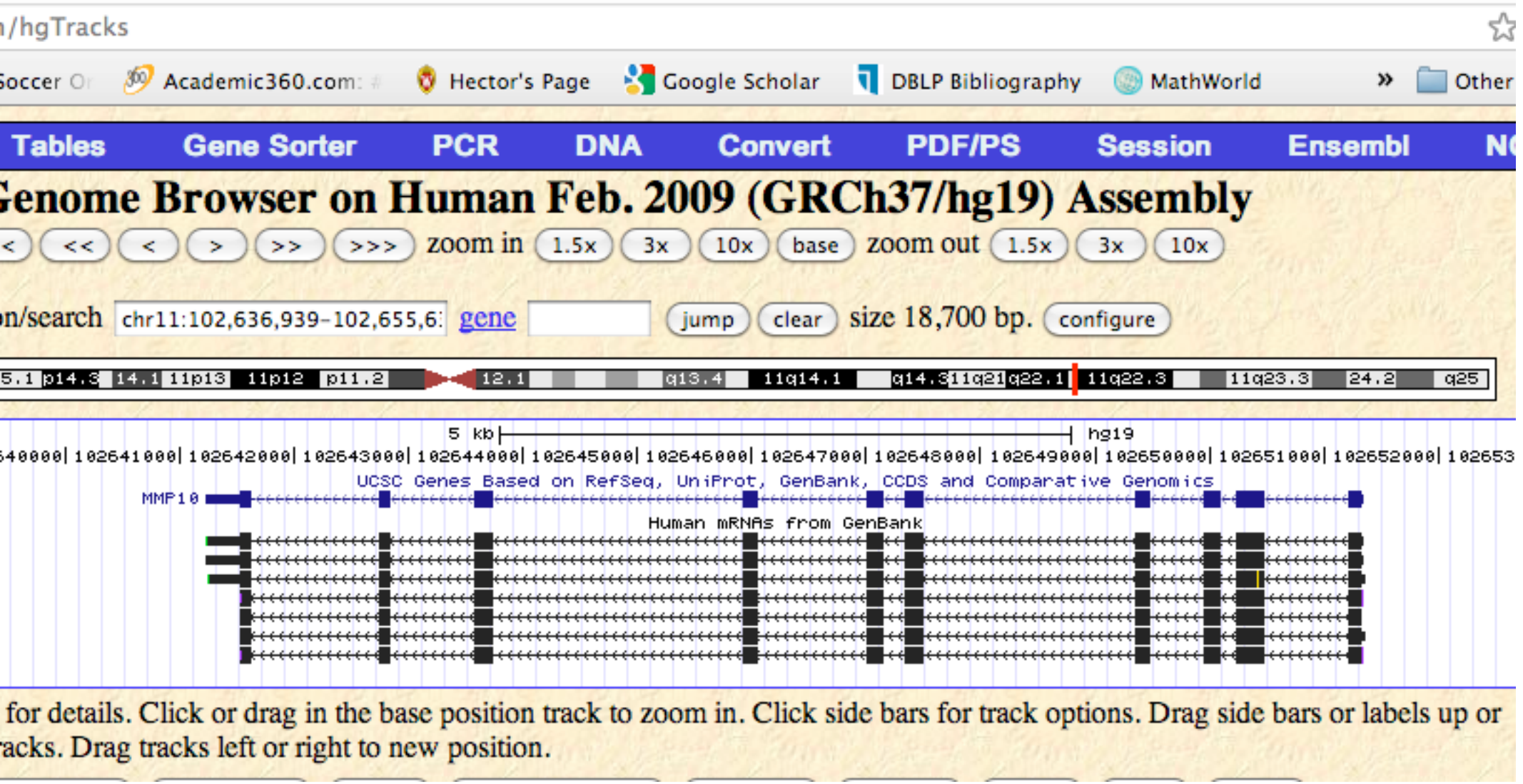


The full genome sequence is a "tour" of the graph

Source: De Novo Assembly Using Illumina Reads. Illumina. 2010
http://www.illumina.com/Documents/products/technotes/technote_denovo_assembly.pdf

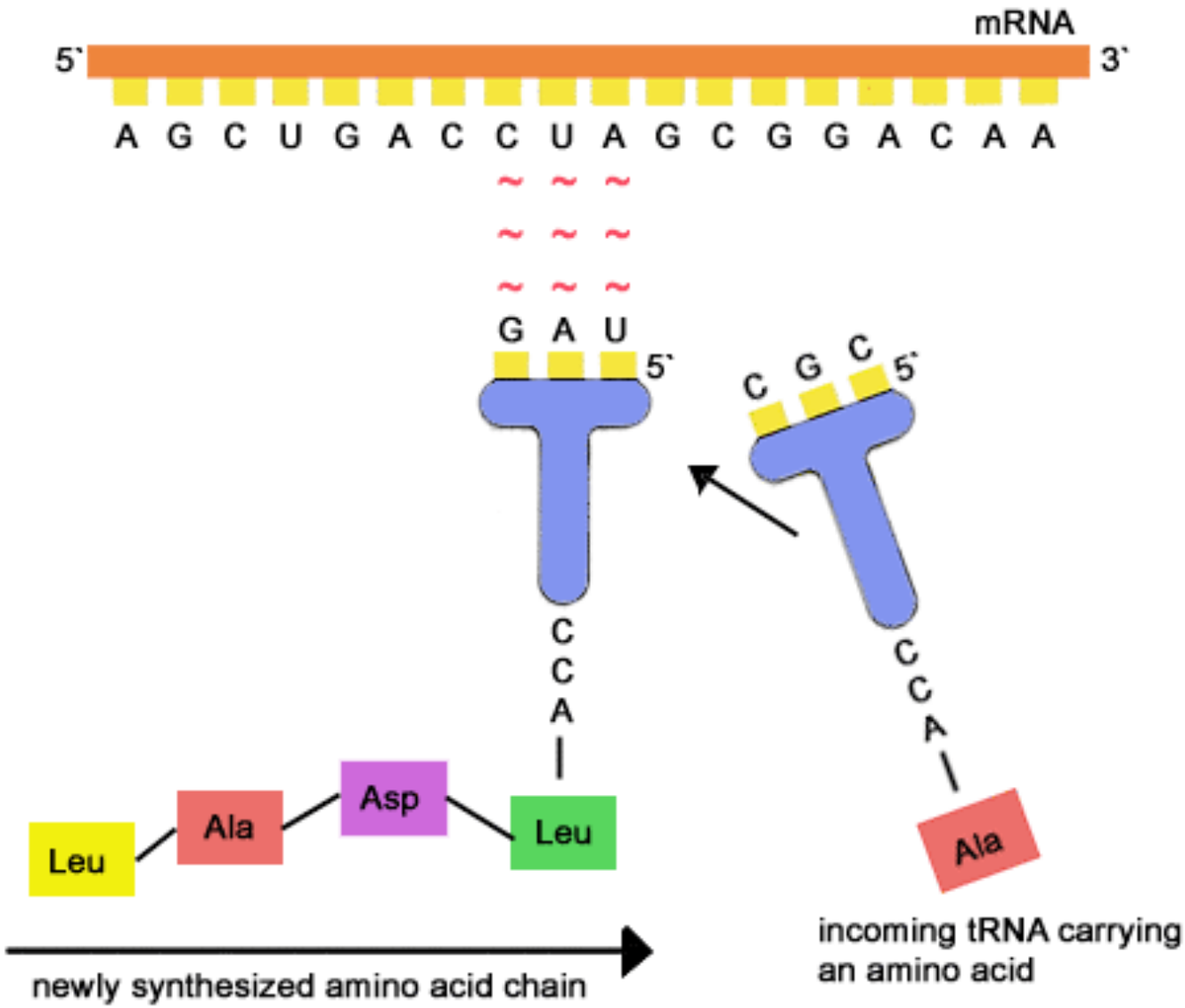
Transcription





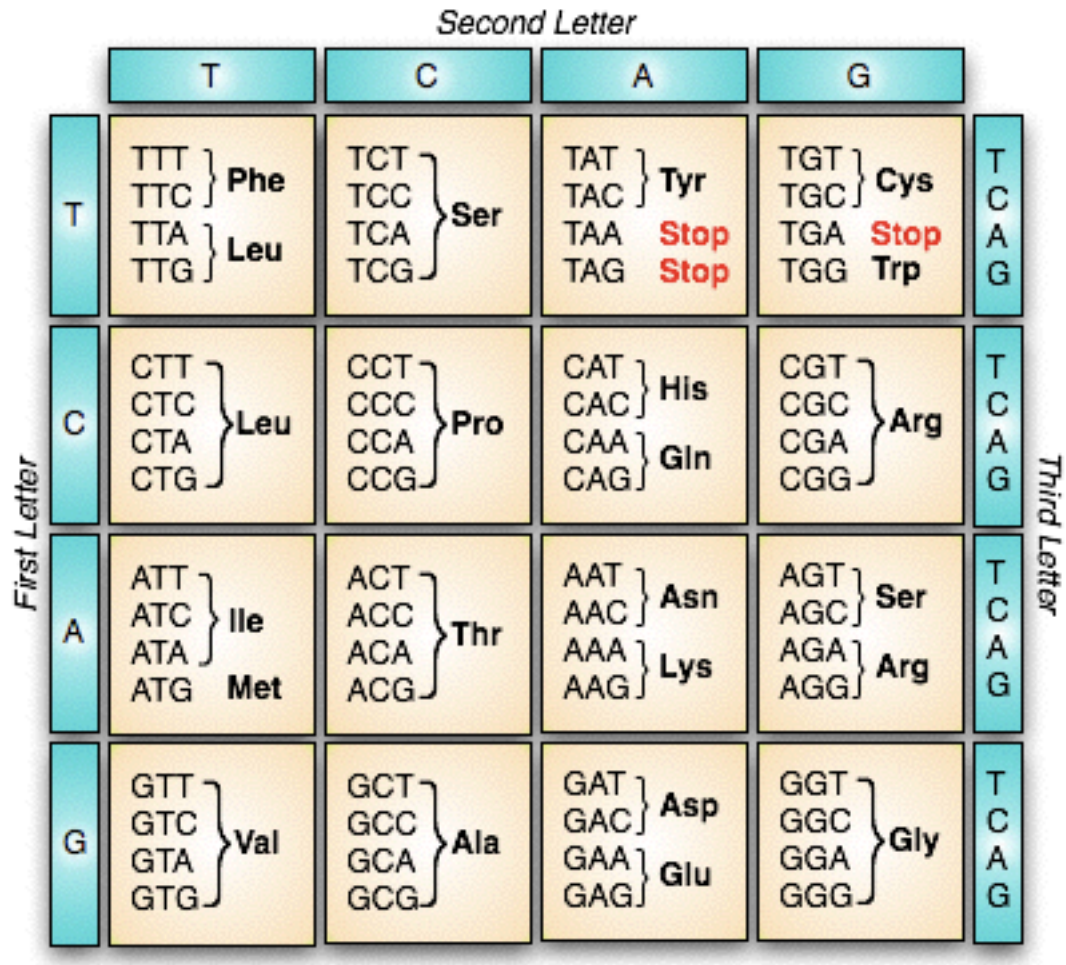
http://www.uniprot.org/uniprot/P09238

Translation

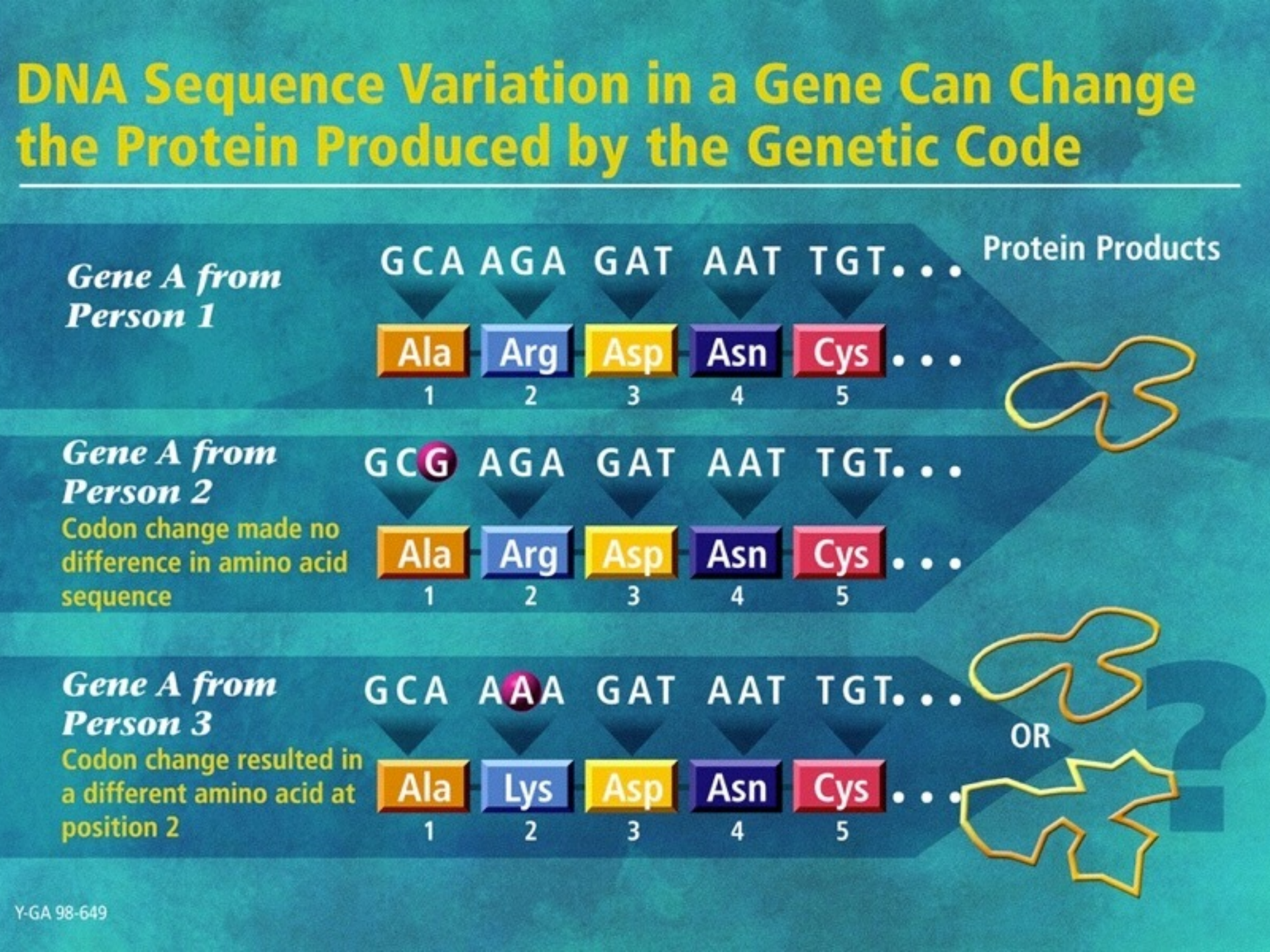


http://gel.ym.edu.tw/~ycl6/sc2005/images/translation.gif

The genetic code



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



Y-GA 98-649

Health or Disease?

DNA Sequence

Person 1 A A A T T T

Normal protein

Some DNA variations have no negative effects

Person 2 A A T T T T

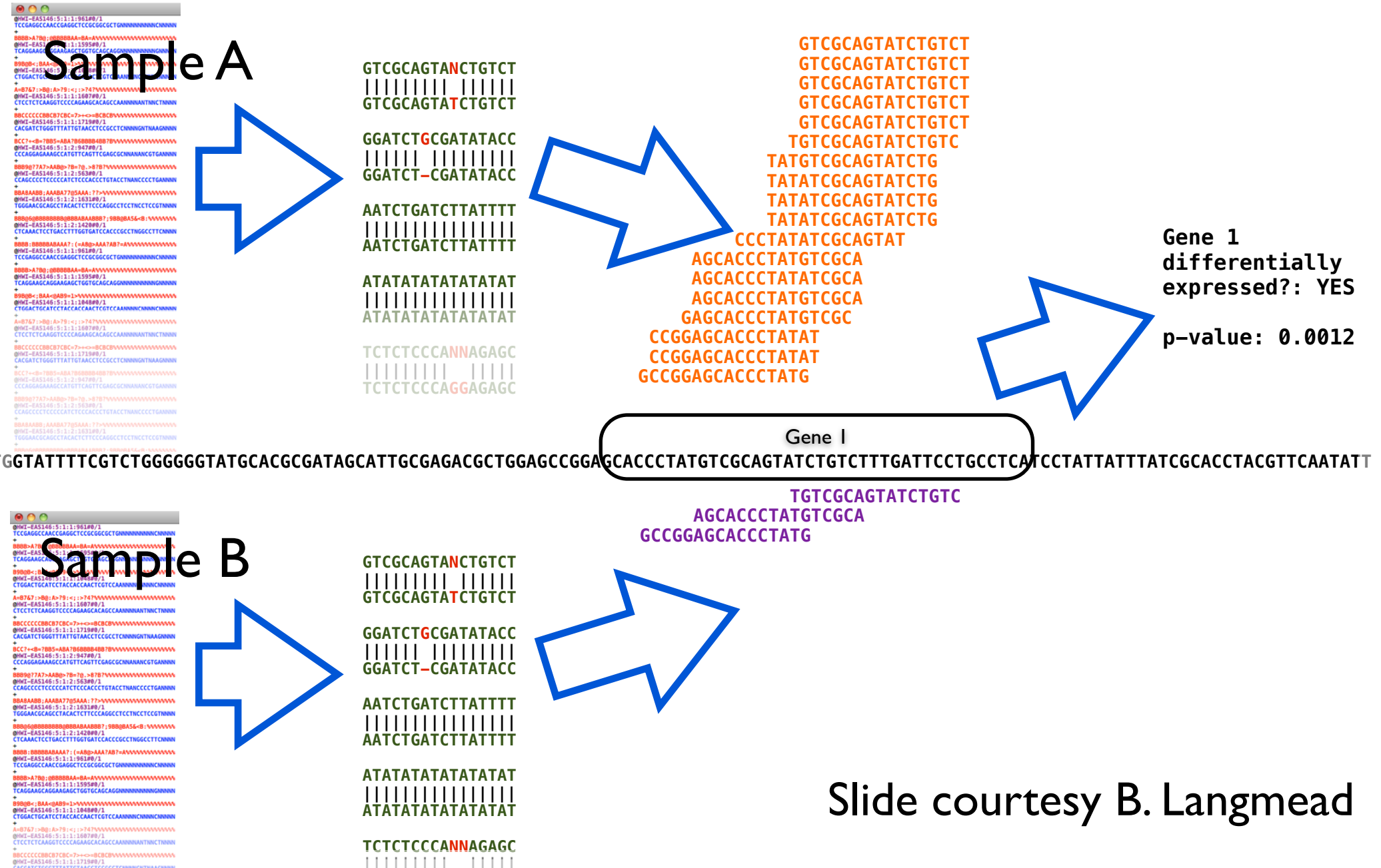
Low or nonfunctioning protein

Person 3 A A C T T T

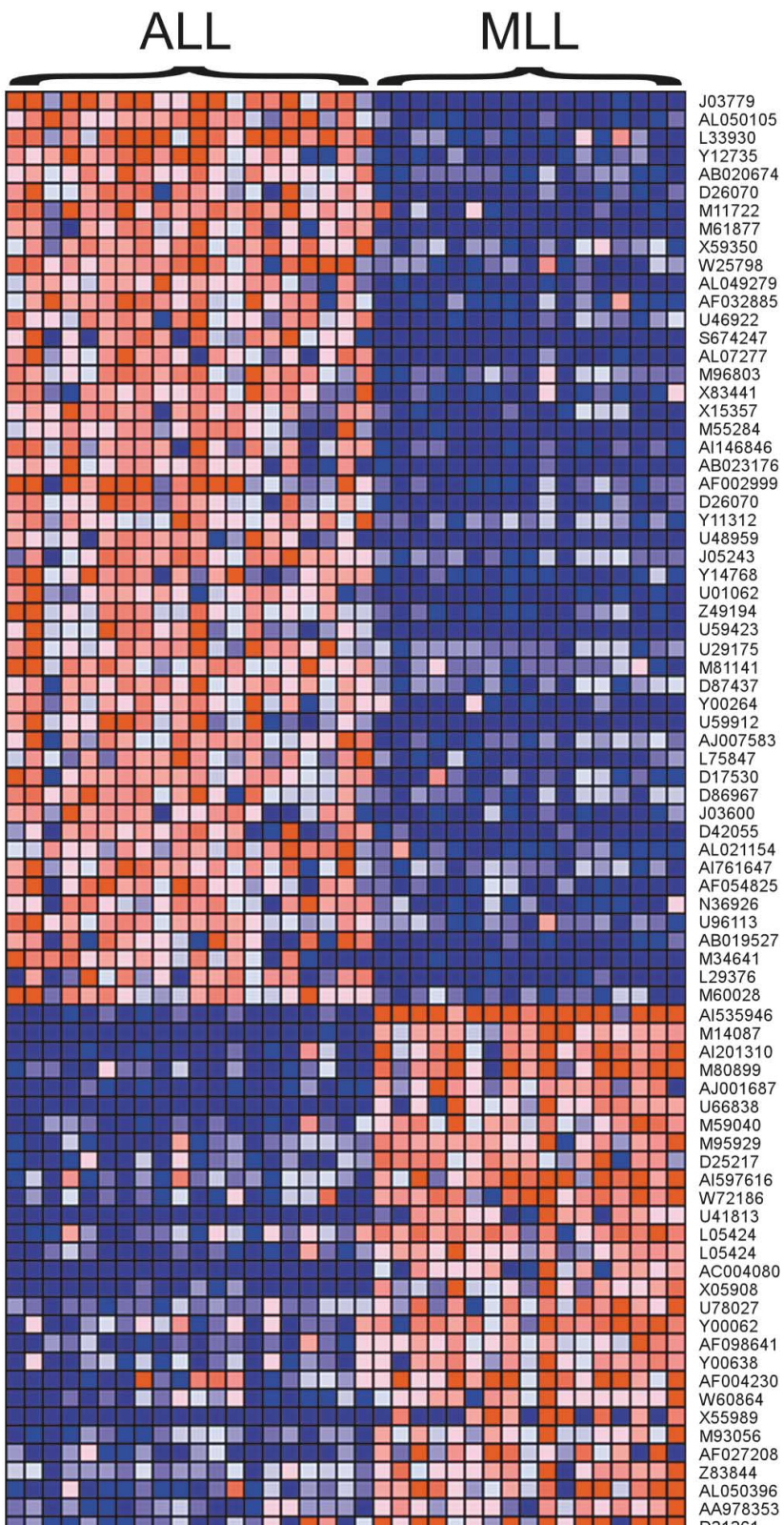
Other variations lead to disease (e.g., sickle cell) or increased susceptibility to disease (e.g., lung cancer)

YGG-00-0480

RNA-seq differential expression



Slide courtesy B. Langmead

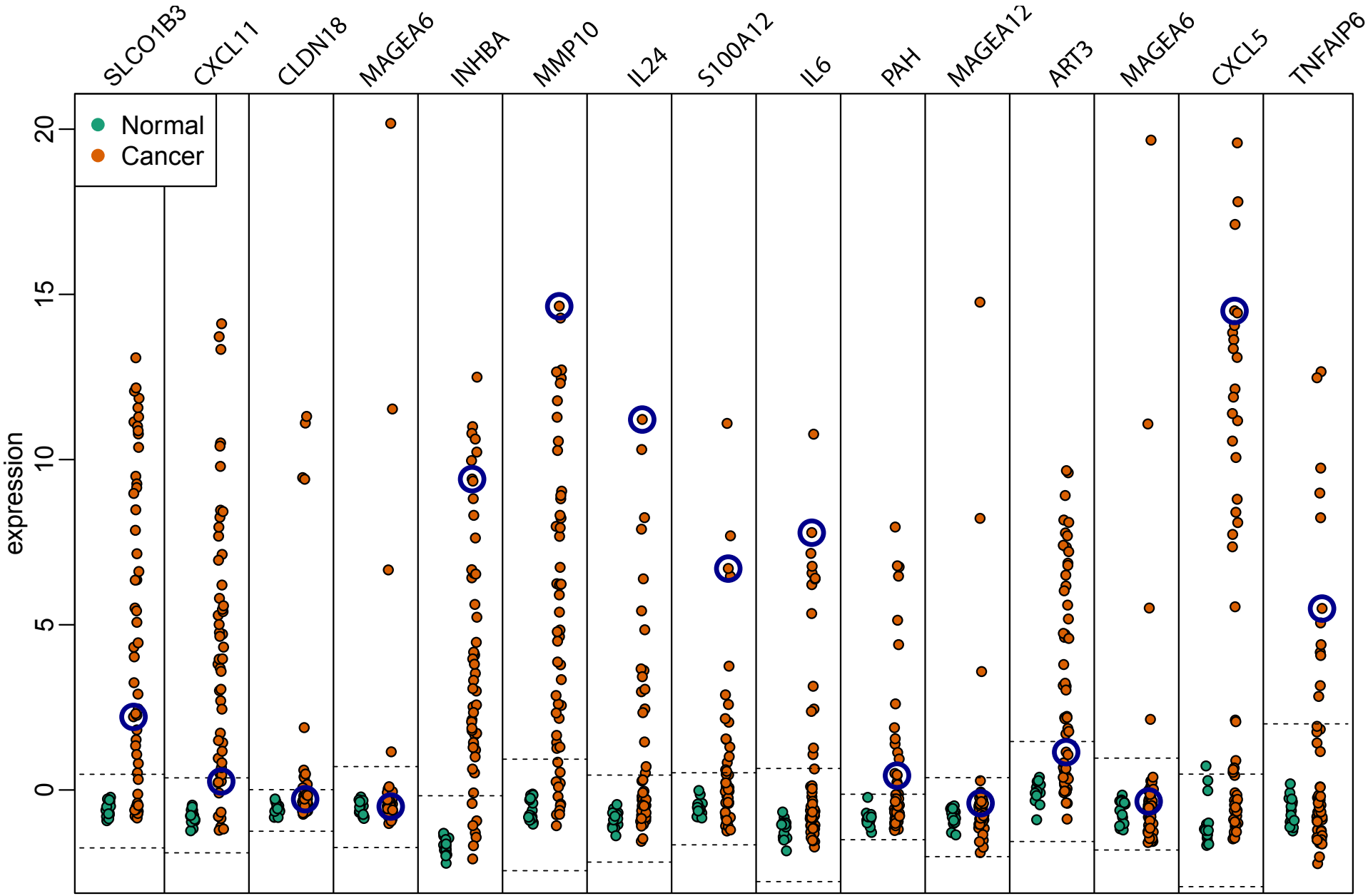


MLL translocations specify a distinct gene expression profile that distinguishes a unique leukemia

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

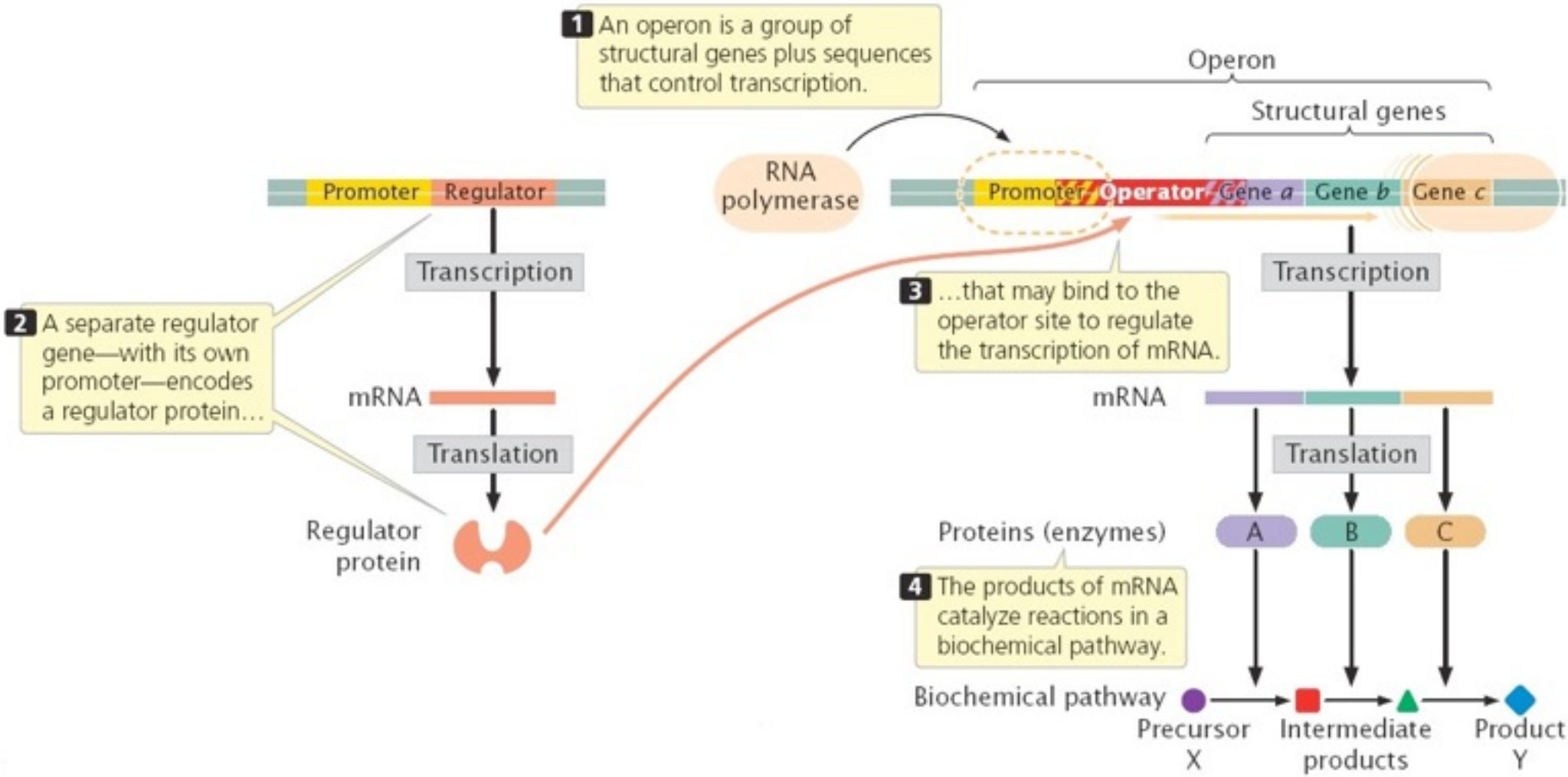
**These authors contributed equally to this work.*

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

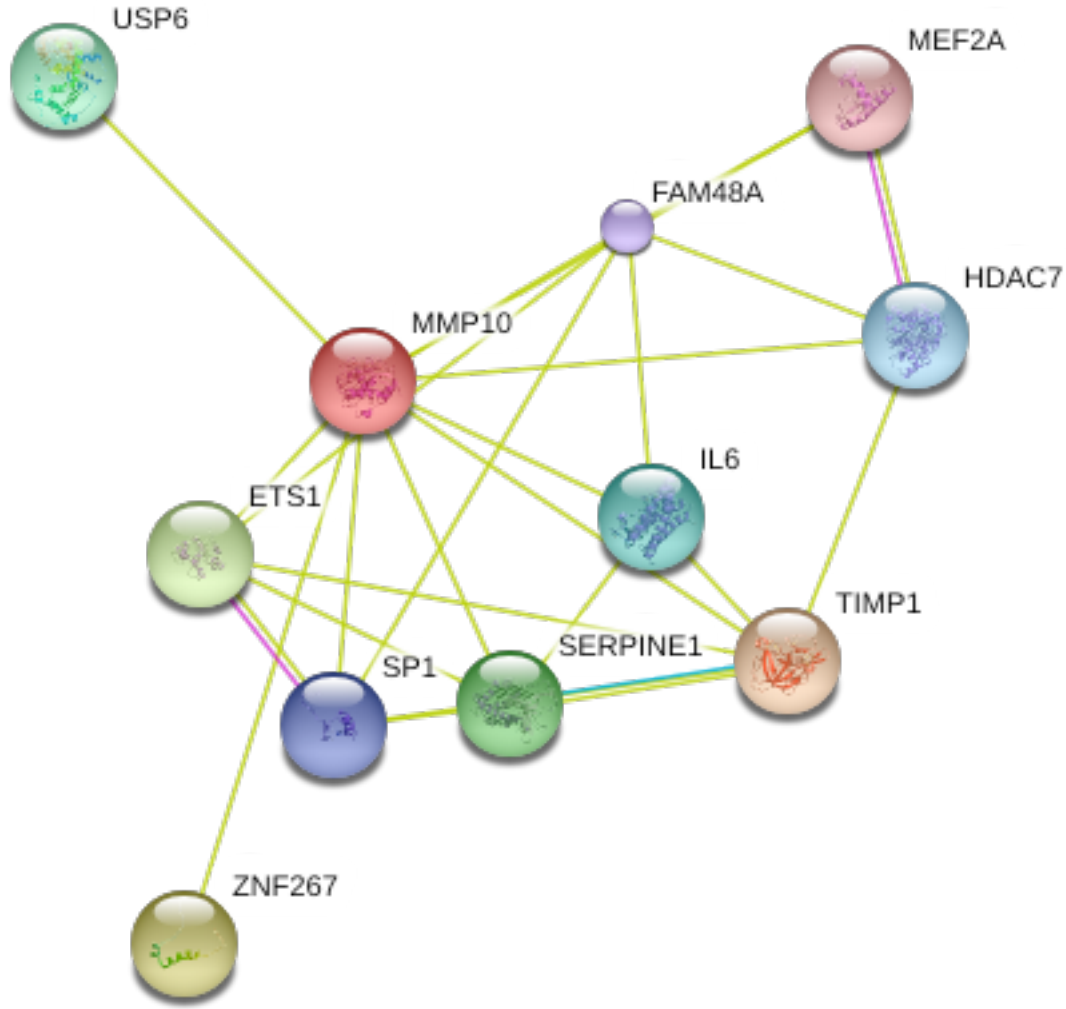


[Corrada Bravo, et al. Gene expression anti-profiles. *BMC Bioinformatics*.]

gene regulation

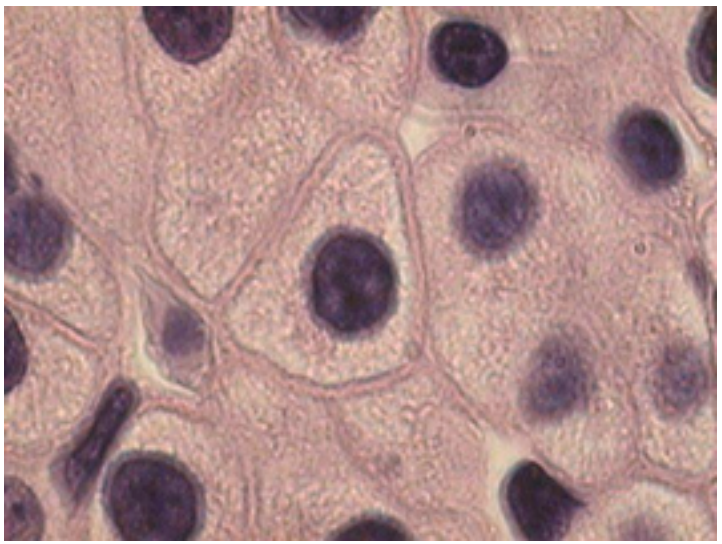


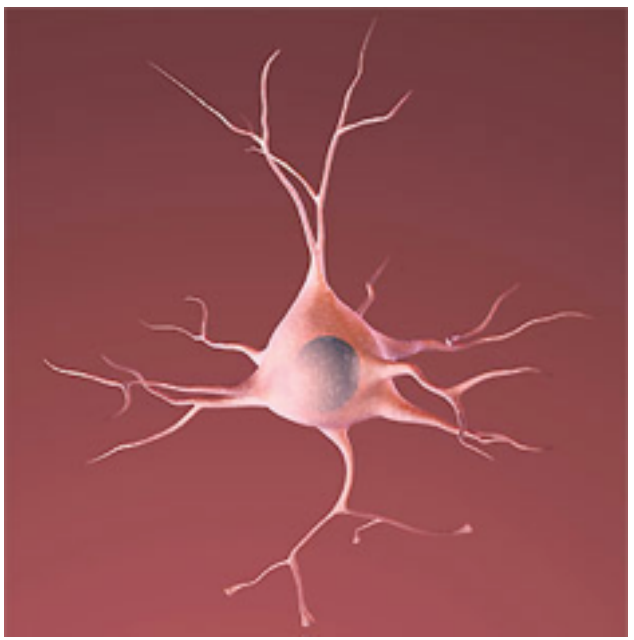
gene regulation

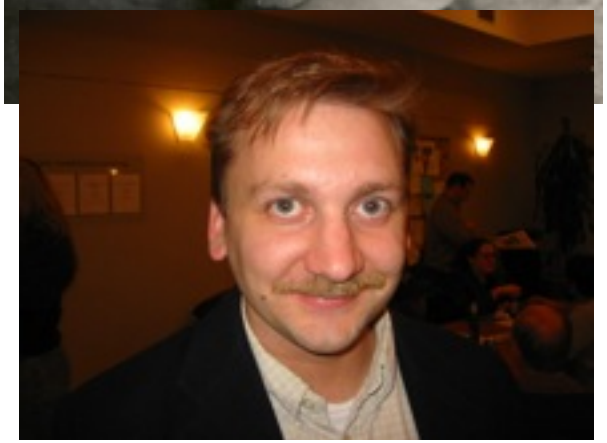


http://string-db.org/version_9_0/newstring_cgi/show_network_section.pl?identifier=9606.ENSP00000279441&all_channels_on=1&network_flavor=evidence&targetmode=proteins

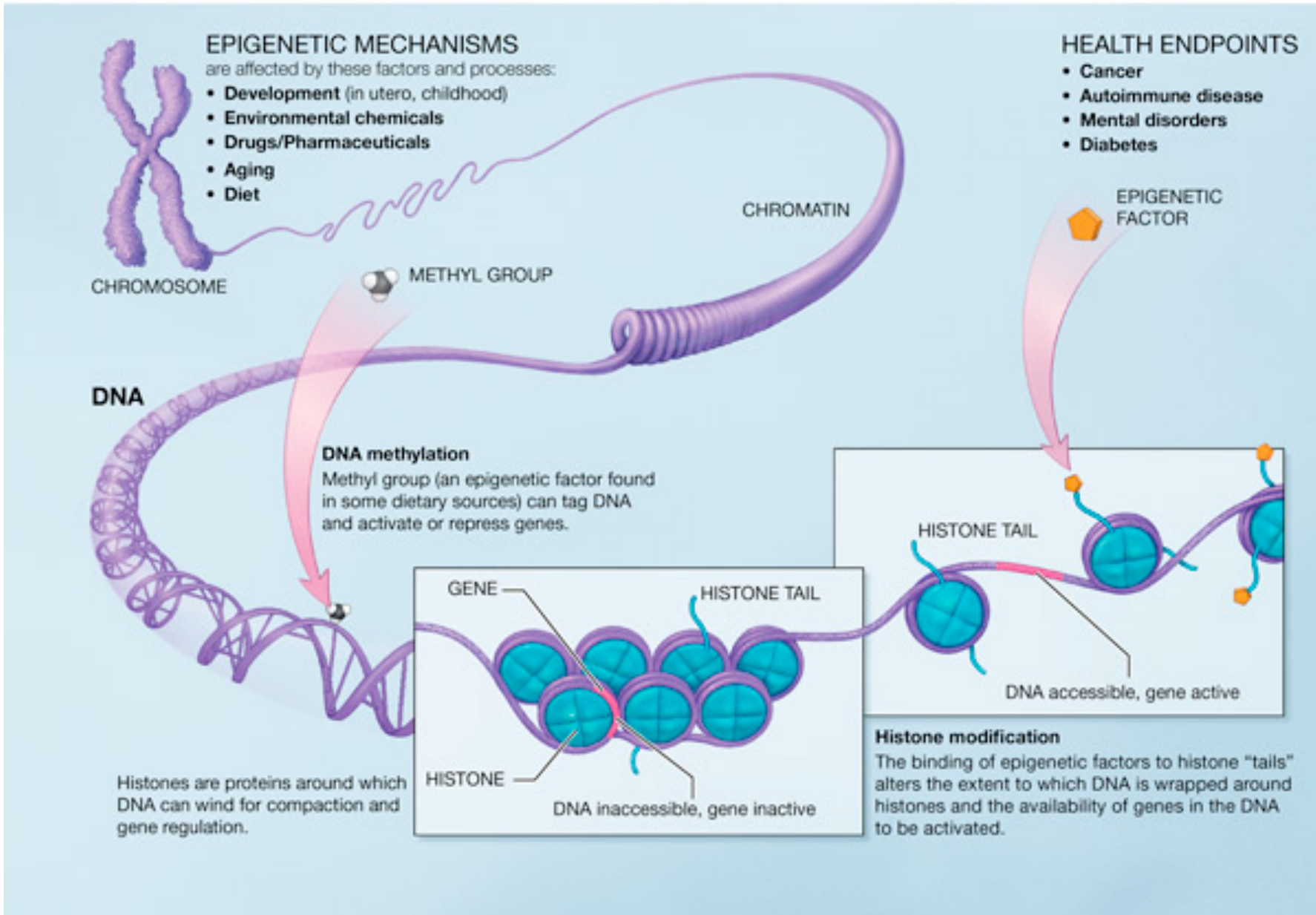
How many basepair differences?





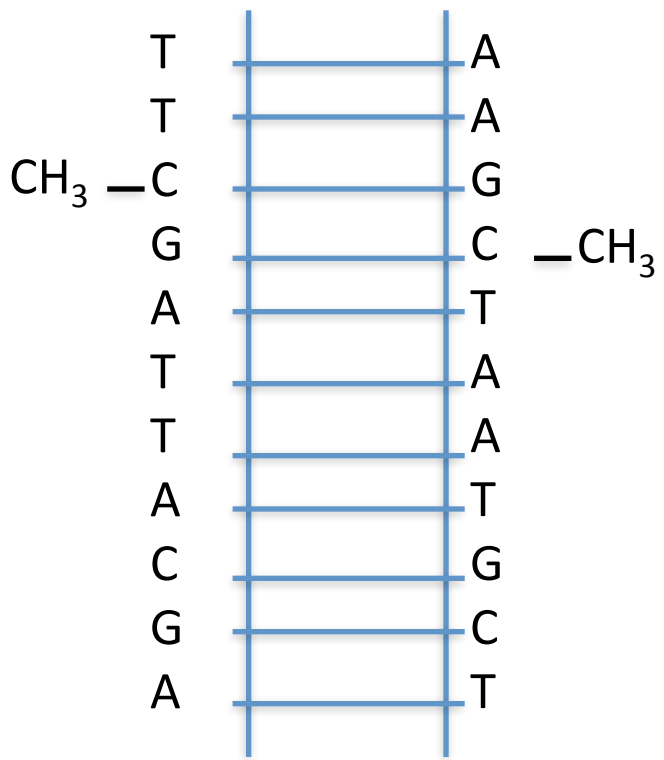


Epigenetics

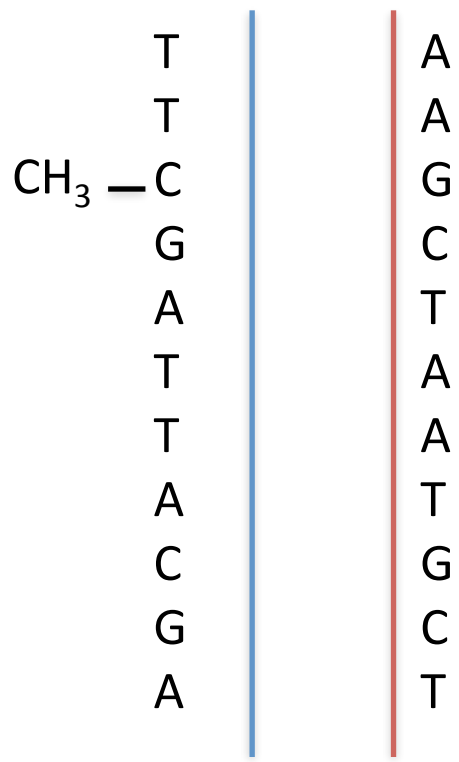


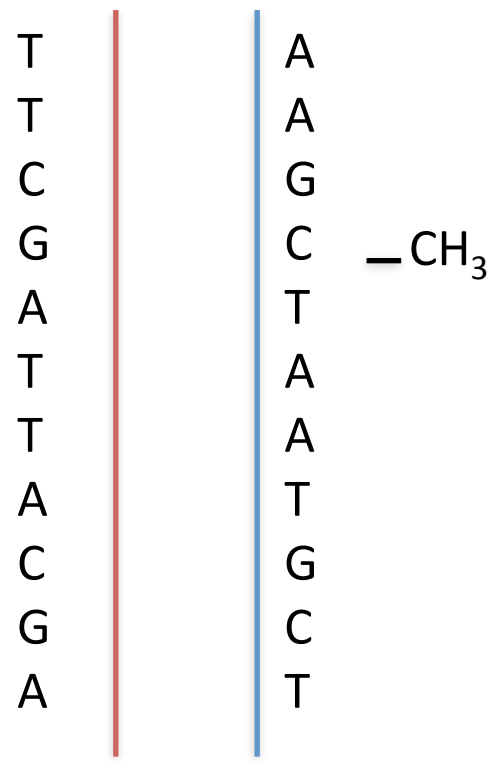
http://nihroadmap.nih.gov/EPIGENOMICS/images/epigeneticmechanisms.jpg

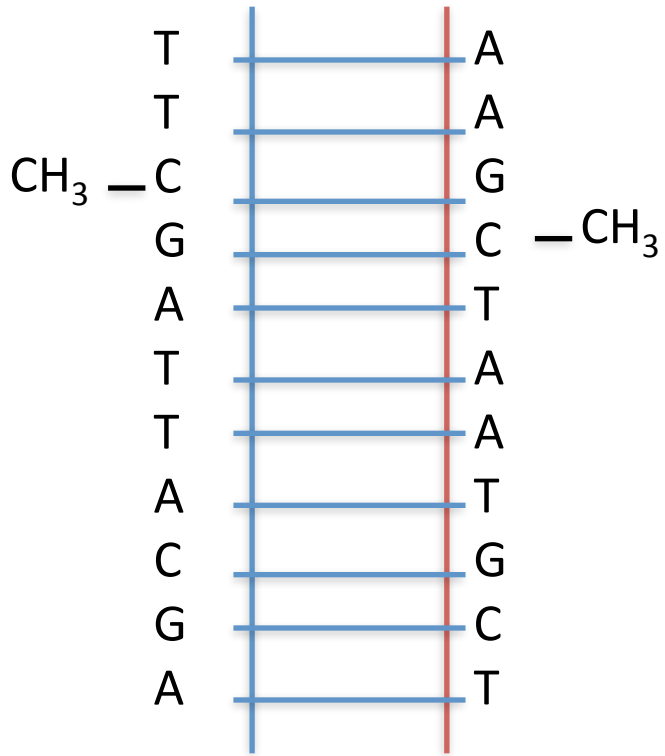
DNA Methylation

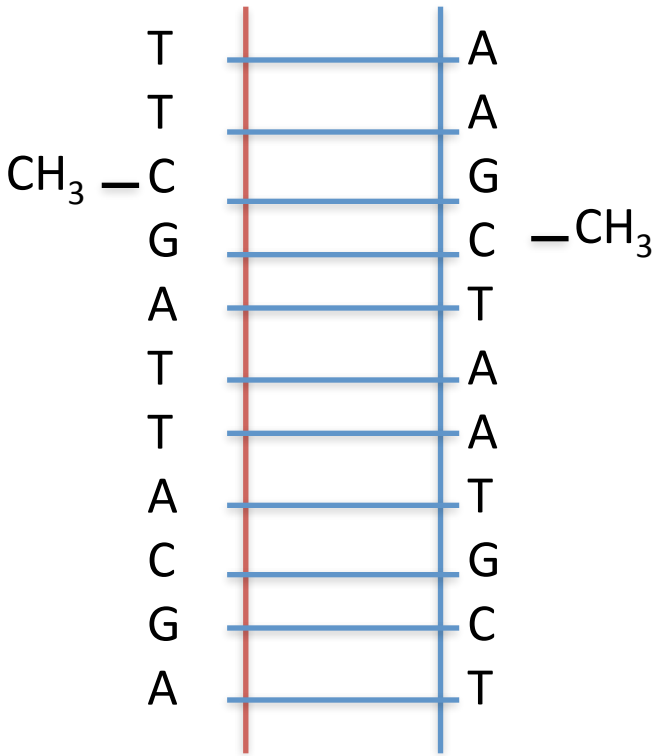



T
T
CH3 C
G
A
T
T
A
C
G
A
A
A
G
C CH3
T
A
A
T
G
C
T

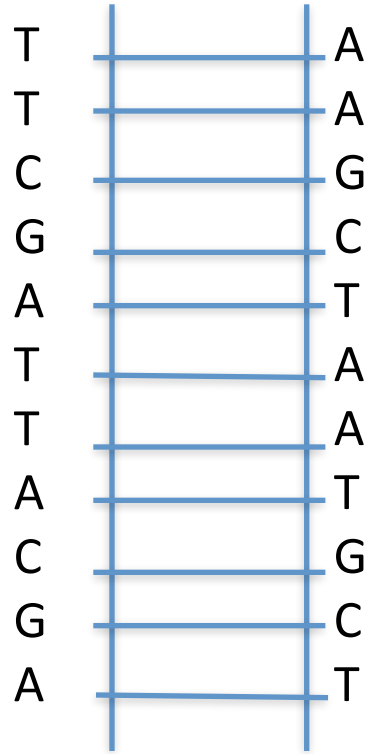

T
T
CH3 C
G
A
T
T
A
C
G
A
A
A
G
C
T
A
A
T
G
C
T


T
T
C
G
A
T
T
A
C
G
A
A
A
G
C CH3
T
A
A
T
G
C
T

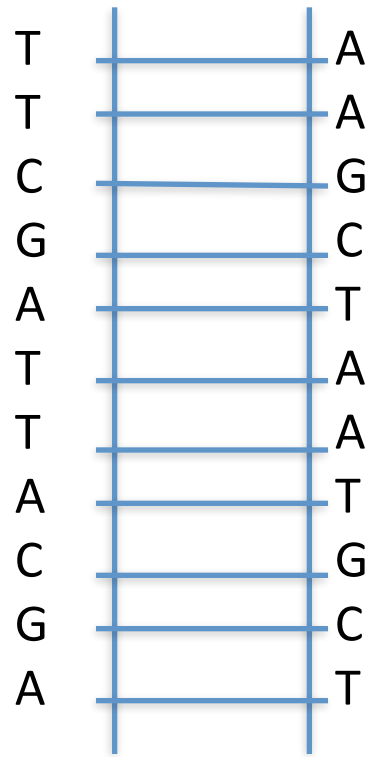

T
T
CH3 C
G
A
T
T
A
C
G
A
A
A
G
C CH3
T
A
A
T
A
T
A


T
T
CH3 C
G
A
T
T
A
C
G
A
A
A
G
C CH3
T
A
A
T
A
T
A

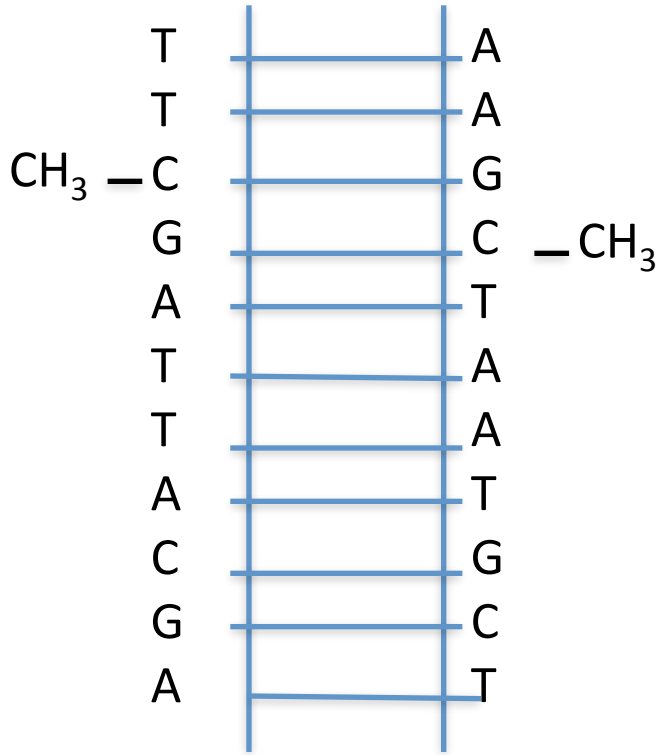
Liver



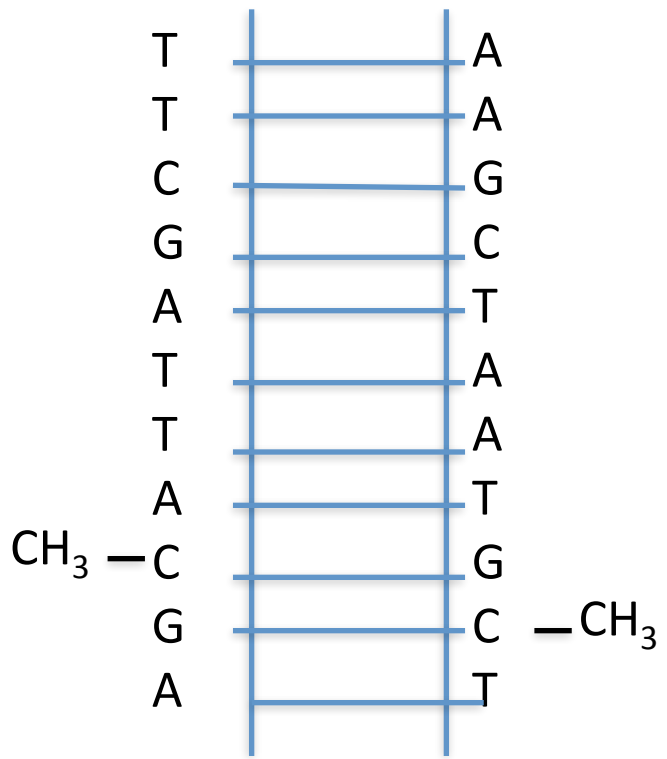
Brain



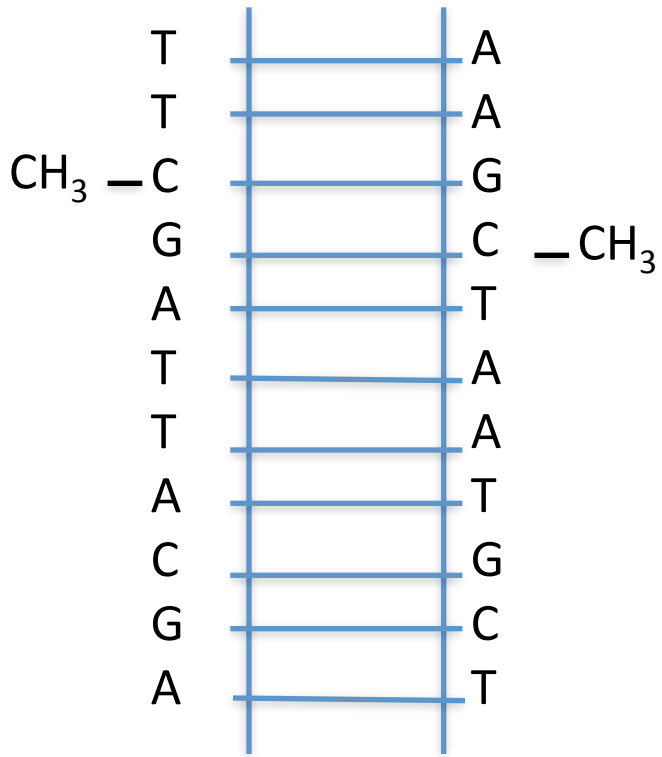
Liver

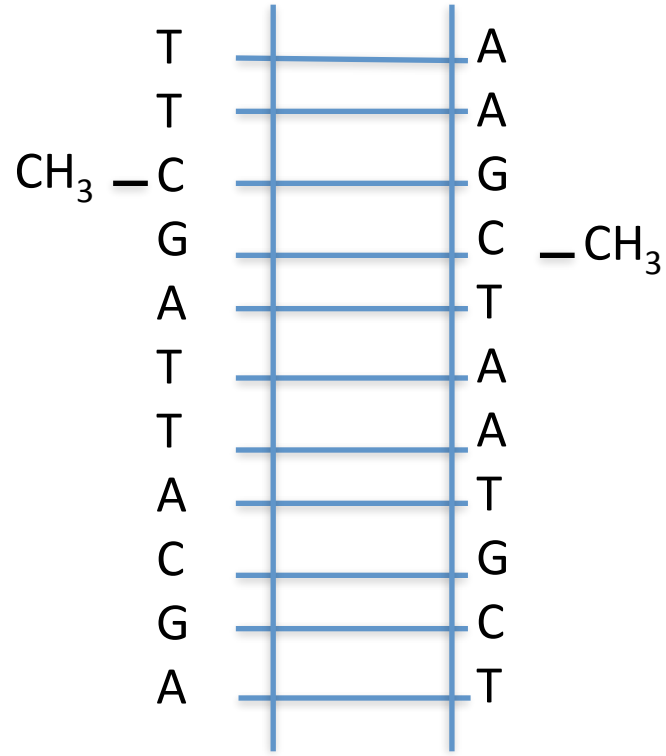

T A
T A
CH3 C G
G C CH3
A T
T A
T A
A T
C G
G C
A T

Brain

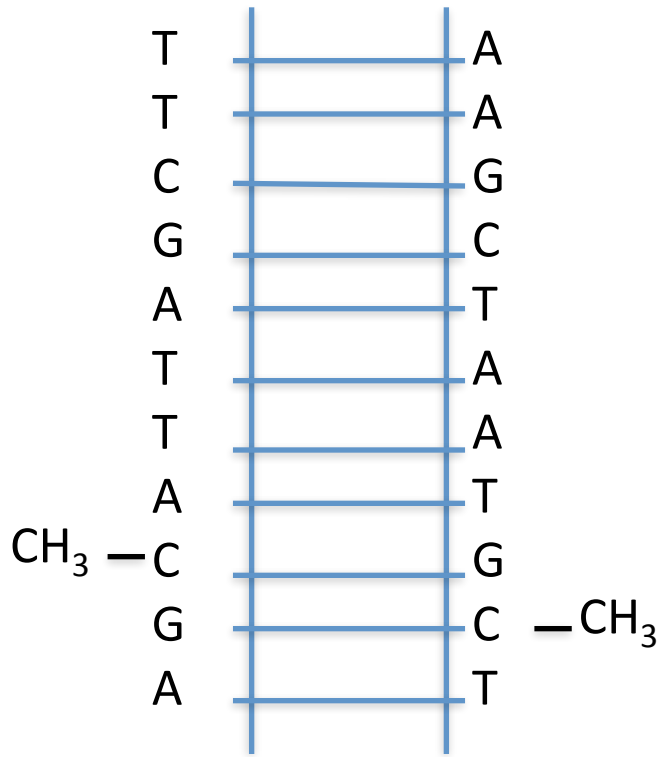

T A
T A
C G
G C
A T
T A
T A
A T
CH3 C G
G C CH3
A T

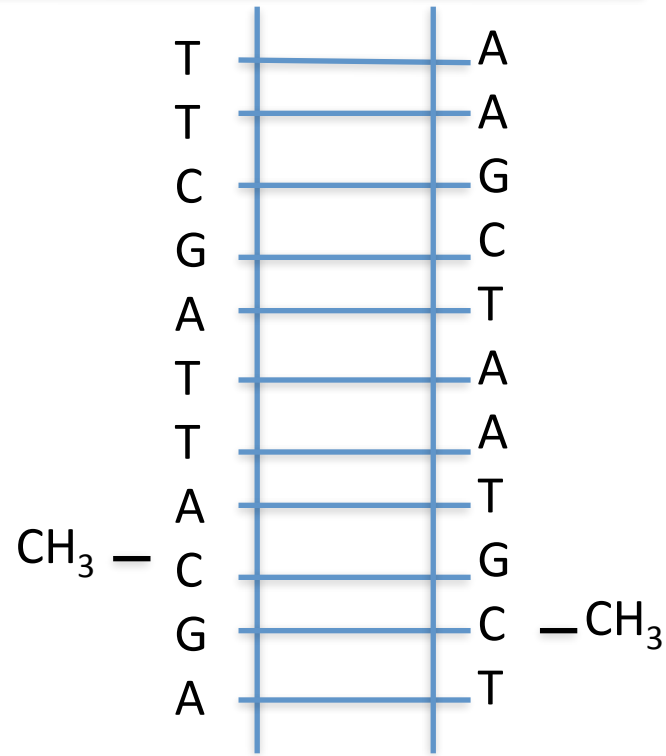
Liver


T
T
CH3 — C
G
A
T
T
A
C
G
A
A
A
G
C
T
A
A
T
G
C
T
— CH3

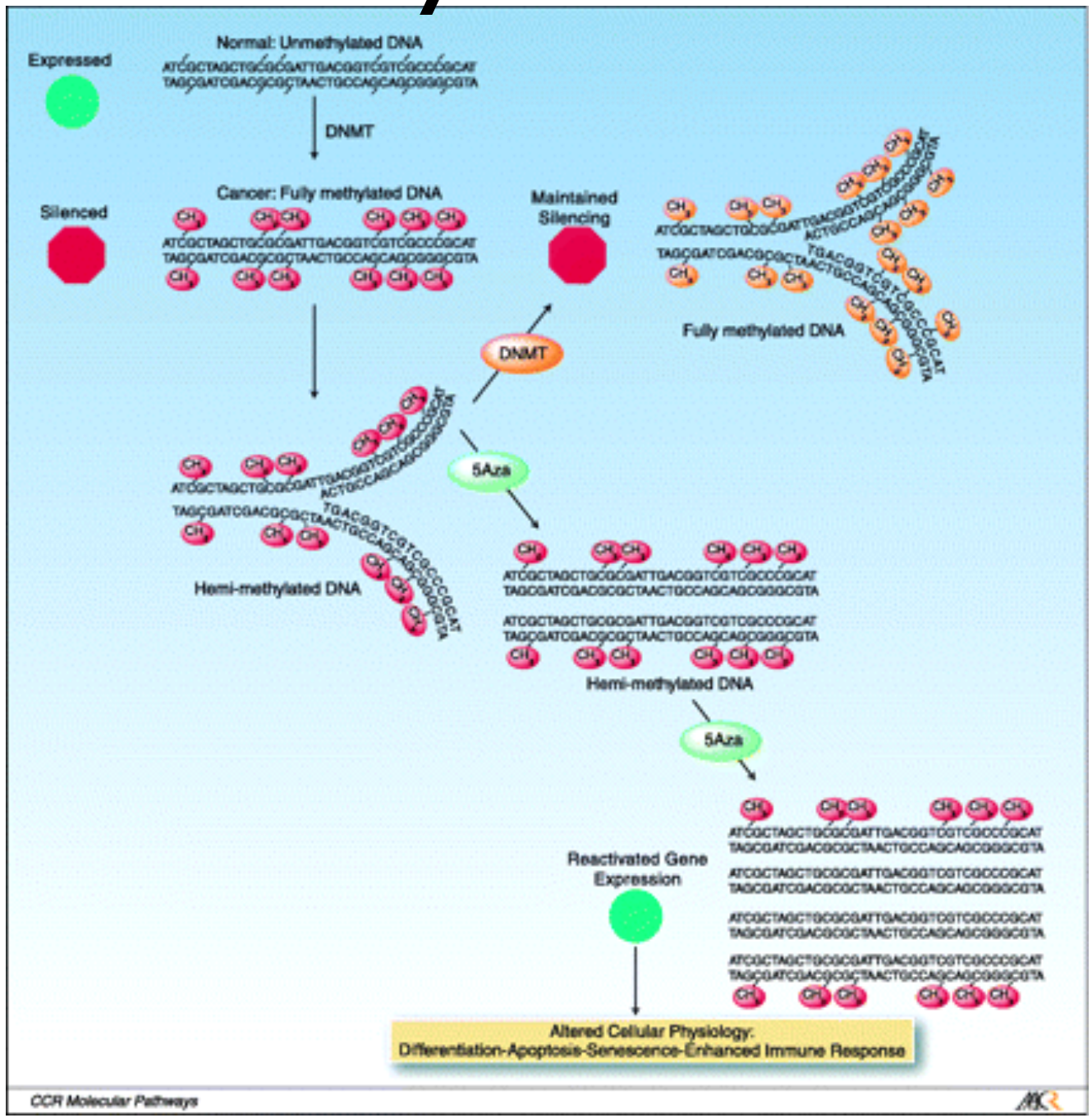

T
T
CH3 — C
G
A
T
T
A
C
G
A
A
A
G
C
T
A
A
T
G
C
T
— CH3

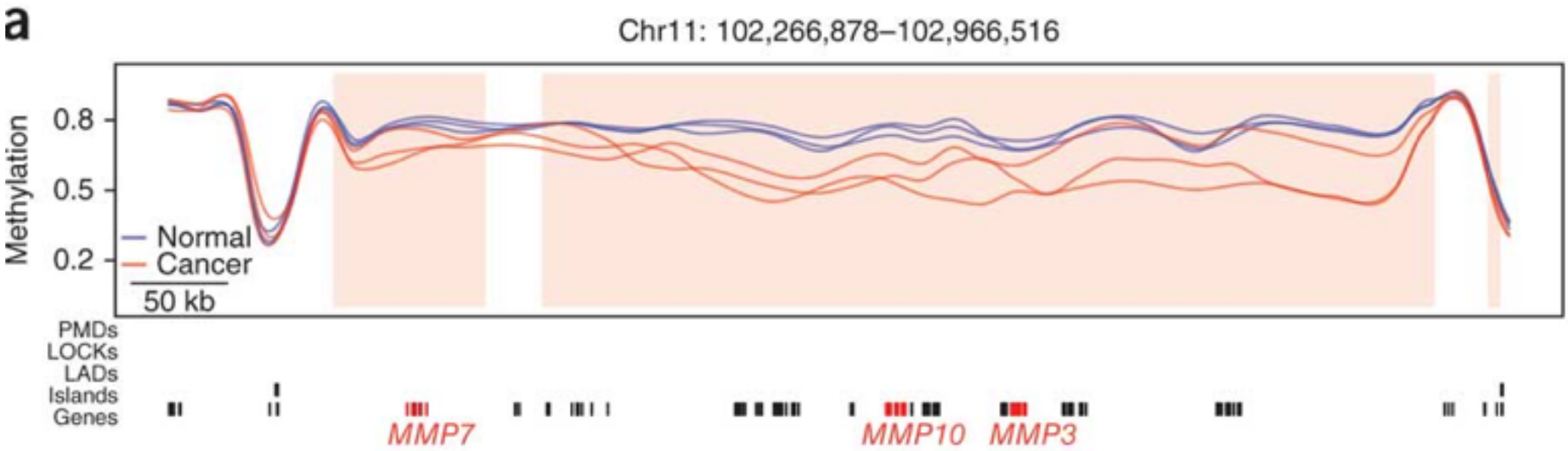
Brain


T
T
C
G
A
T
T
A
CH3 — C
G
A
A
A
G
C
T
A
A
T
G
C
T
— CH3

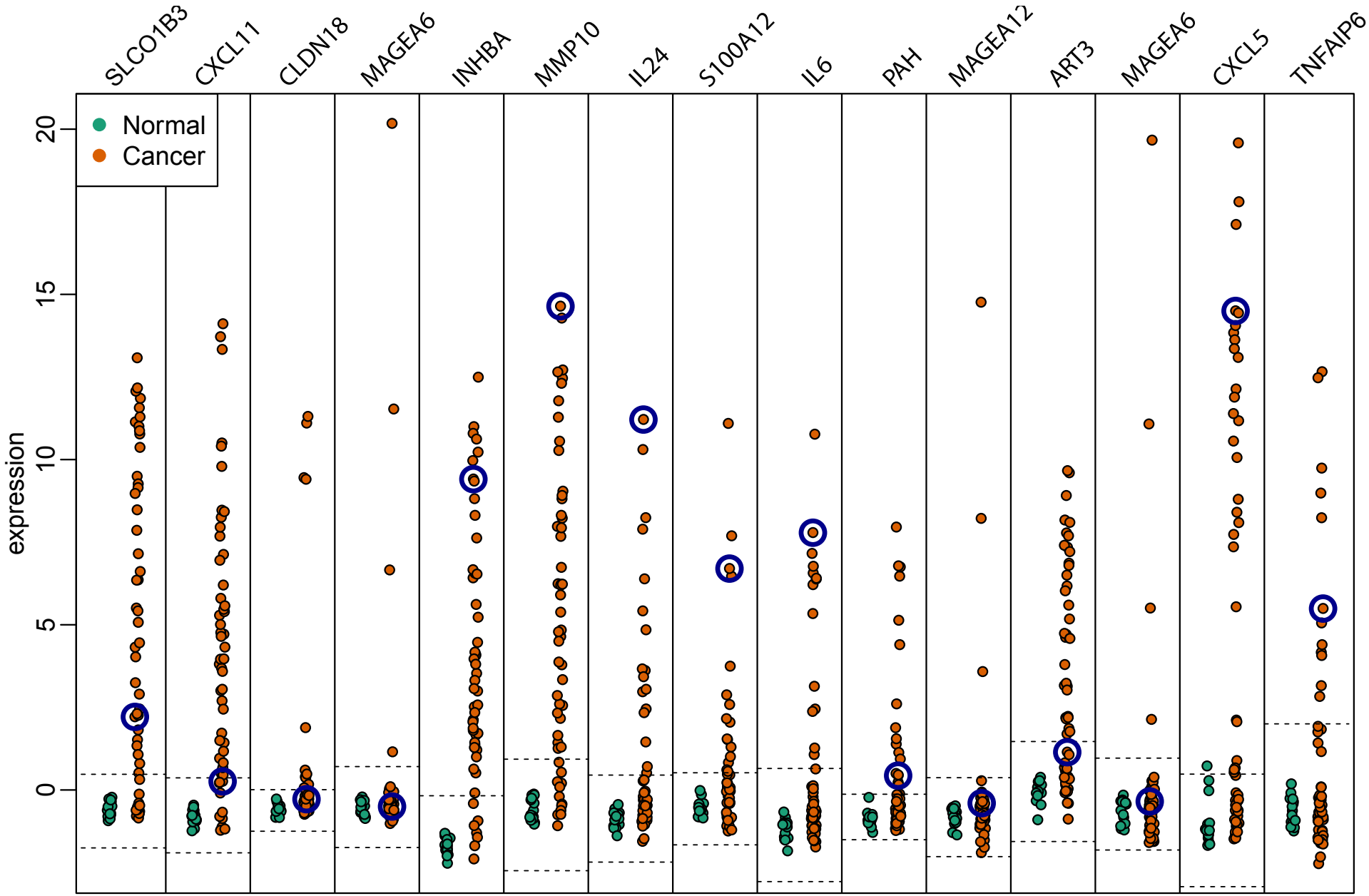

T
T
C
G
A
T
T
A
CH3 — C
G
A
A
A
G
C
T
A
A
T
G
C
T
— CH3

DNA Methylation in cancer





[Hansen et al. Nature Genetics 2011]



[Corrada Bravo, et al. Gene expression anti-profiles. *BMC Bioinformatics*]