Sequence Modeling

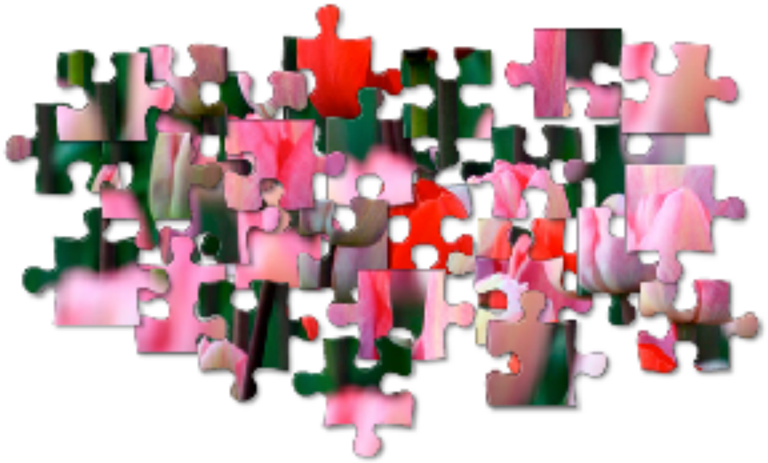
Ben Langmead

JOHNS HOPKINS
WHITING SCHOOL
of ENGINEERING

Department of Computer Science

Picking up signals

So far, we've focused on how to stitch fragmentary evidence into bigger pictures, i.e. genomes



Now we have more questions!

Where are the genes?

Where/what is the *functional* DNA?

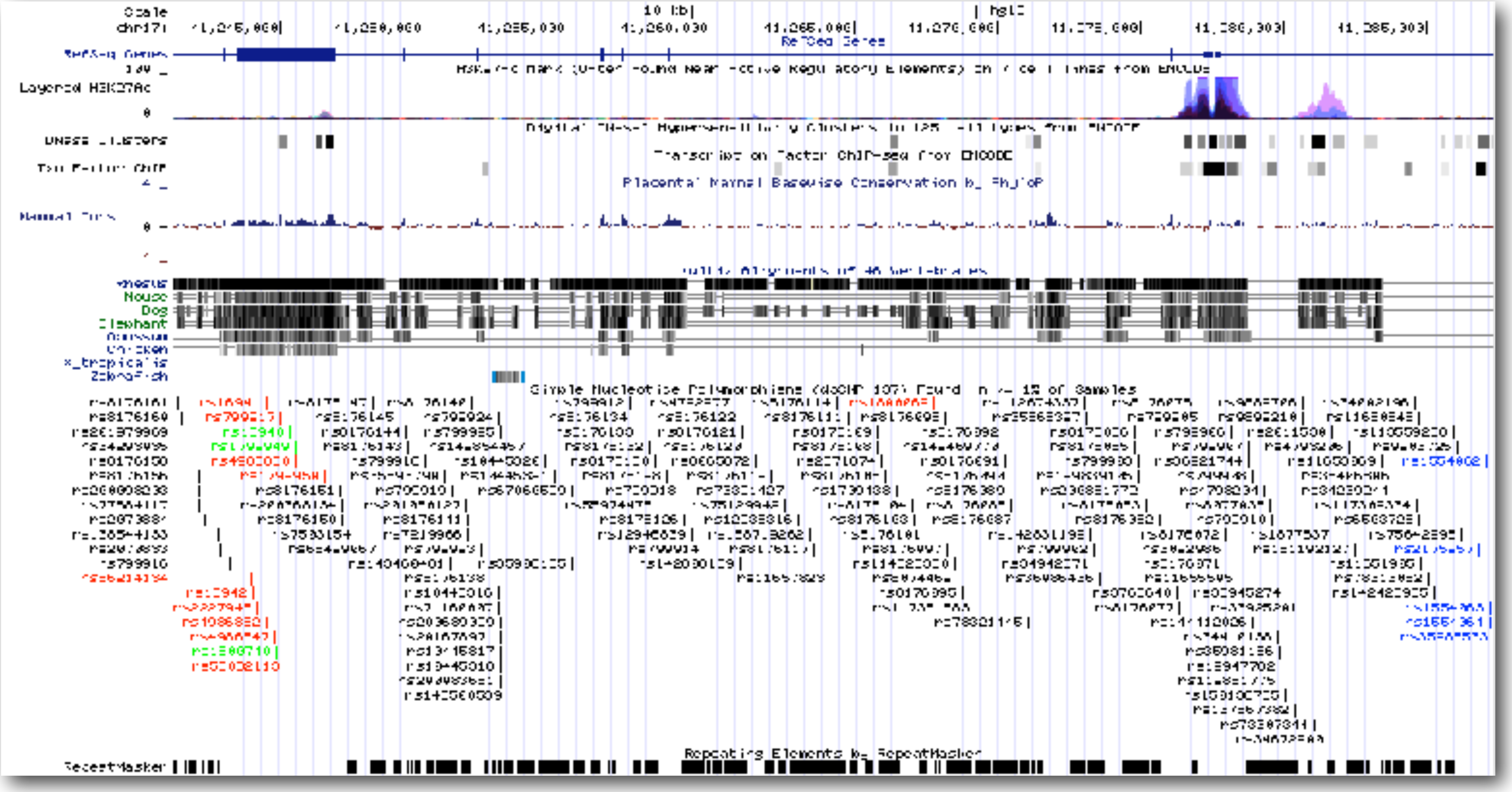
What's different about the DNA in different tissues?

In what abundance do we find various molecules?

What differences exist between individuals?

Picking up signals

We know much more about the genome than just its DNA sequence:

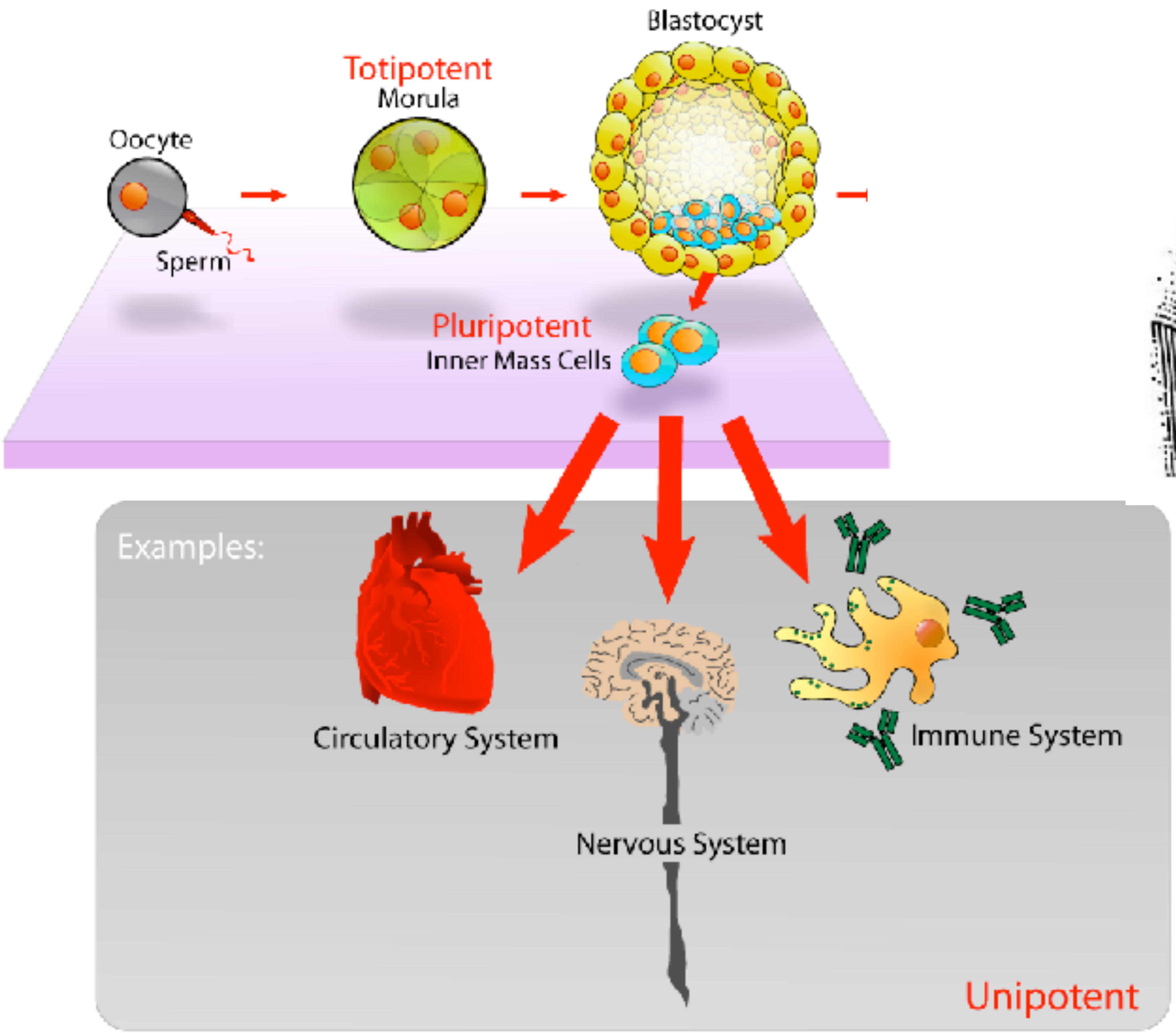


Scale
chr17:
10 kb
hg18
RefSeq Genes
Layered H3K27Ac
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
DNase Clusters
Digital DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE
Txn Factor ChIP
Transcription Factor ChIP-seq from ENCODE
Placental Mammal Basewise Conservation by PhyloP
Mammal Cons
Multiz Alignments of 46 Vertebrates
Rhesus
Mouse
Dog
Elephant
Opossum
Chicken
X_tropicalis
Zebrafish
Simple Nucleotide Polymorphisms (dbSNP 137) Found in >= 1% of Samples
Repeating Elements by RepeatMasker
RepeatMasker

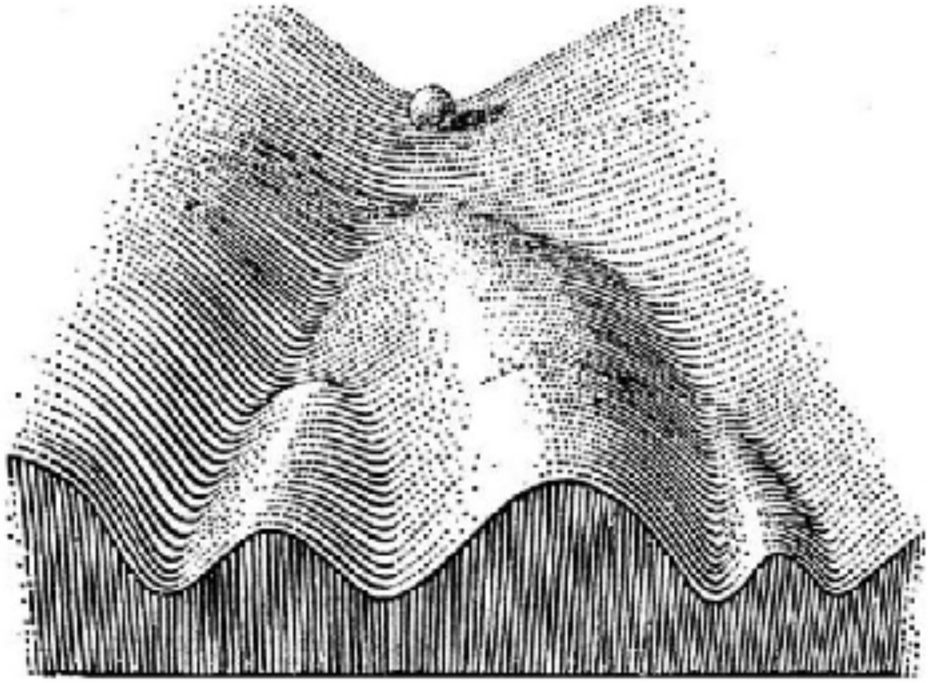
40 K nt region of chromosome 17

http://genome.ucsc.edu/cgi-bin/hgTracks

Epigenetics



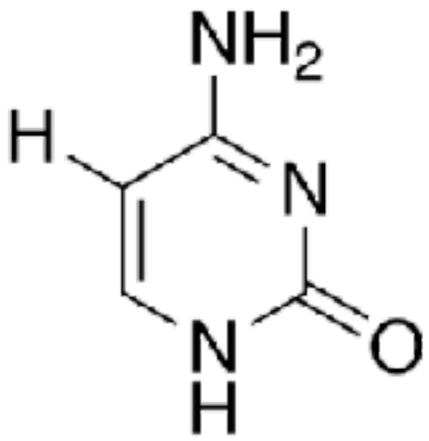
http://en.wikipedia.org/wiki/File:Stem_cells_diagram.png



"Waddington Landscape"

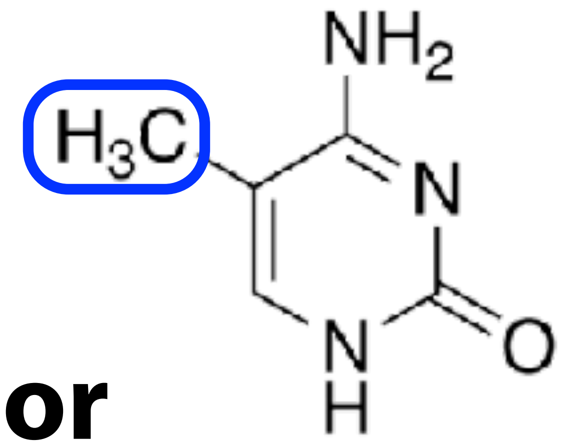
Methylation

Dinucleotide “CG” (AKA “CpG”) is special because C can have a *methyl group* attached



Unmethylated

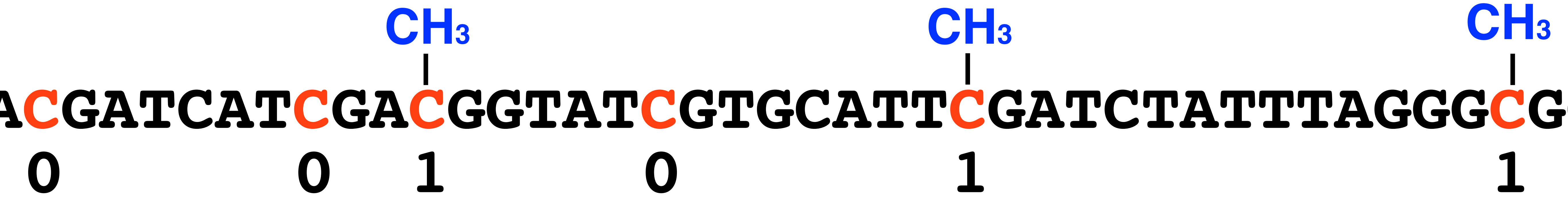
or



Methylated

Methylation

In animals, most methylation is at CG cytosines



CH_3 CH_3 CH_3

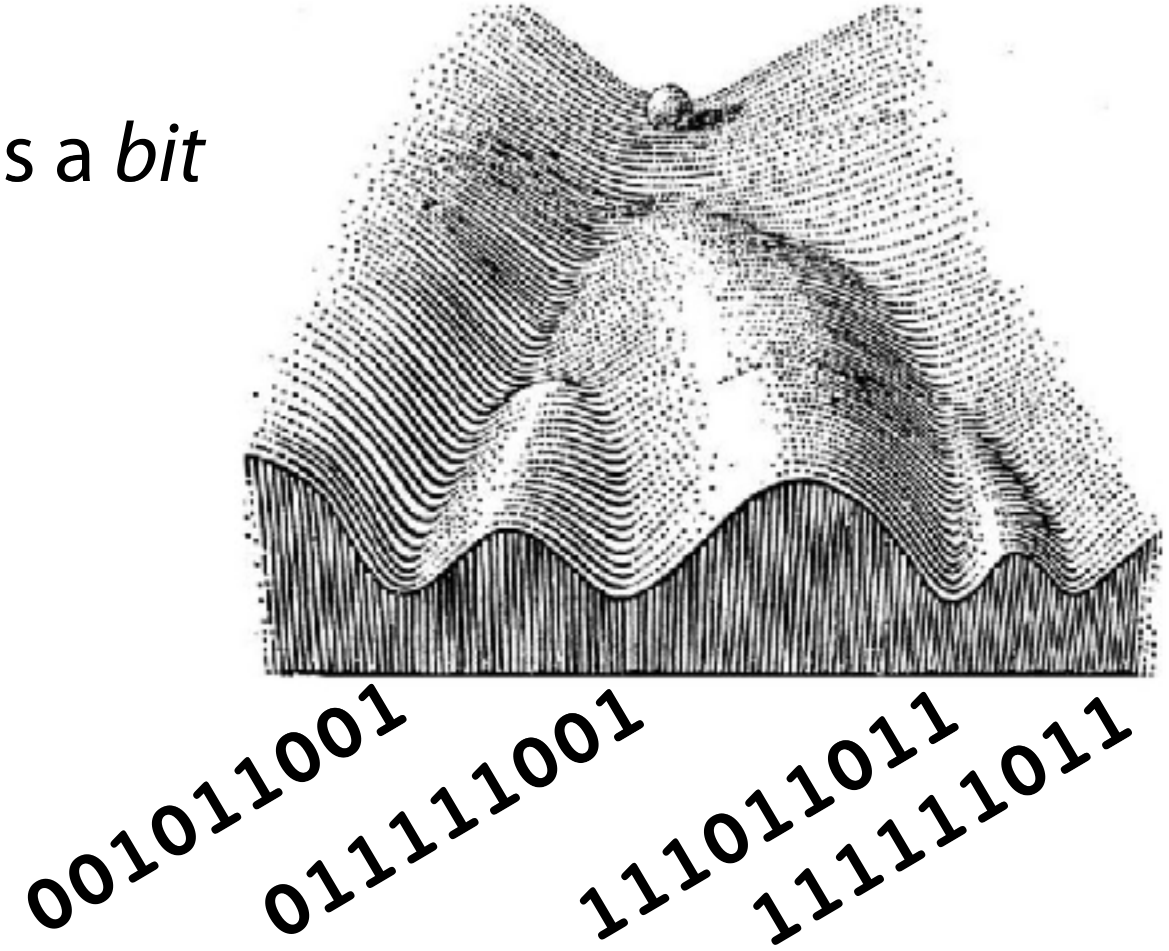
ACGATCATCGACGGTATCGTGCATTCGATCTATTTAGGGCG

0 0 1 0 1 1

Methylation status of every CpG is a *bit*

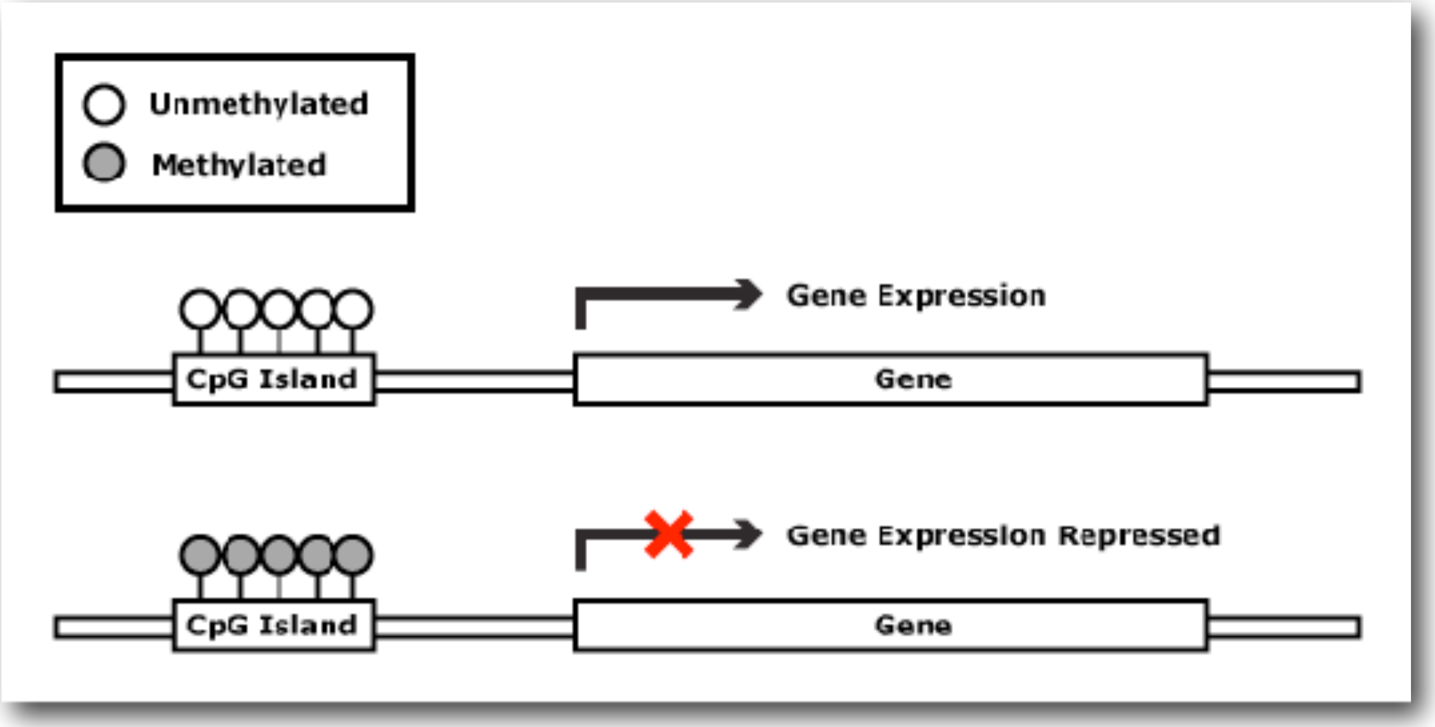
Differentiated cell types have different characteristic bit strings

But every cell type has same *genome*



CpG Islands

CpG island: part of the genome where CG occurs particularly frequently



http://missinglink.ucsf.edu/lm/genes_and_genomes/methylation.html

CpG Islands

Wanted: a strategy for scoring a *k*-mer according to how confident we are it belongs to a CpG island

Scores should be *probabilities*

(This is a simple problem, but real-world tools do use these kinds of techniques to find CpG islands & genes)

Probability mini-review

Sample space (Ω) is set of all possible outcomes

E.g. Ω = { all possible rolls of 2 dice }

An *event* (*A*, *B*, *C*, ...) is a subset of Ω

A = { rolls where first die is odd }, *B* = { rolls where second die is even }

We're often concerned with assigning a probability to an event

P(*A*): fraction of all possible outcomes that are in *A*

P(*A*) = | *A* | / | Ω | = 18 / 36 = 0.5

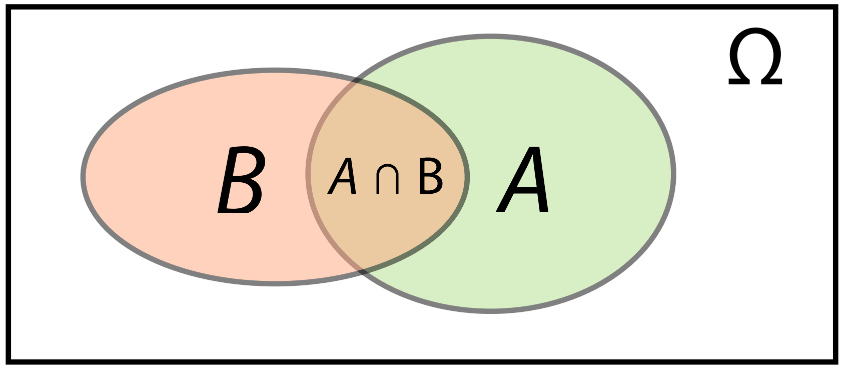
Probability mini-review

P(*A*, *B*): fraction of all possible outcomes in both *A* and *B*

P(*A*, *B*) = | A ∩ B | / | Ω | = 9 / 36 = 0.25

Also written: P(*A* ∩ *B*) or P(*AB*)

Joint probability of *A* and *B*



P(*A* | *B*): fraction of outcomes in *B* that are also in *A*

conditional probability of A given B

P(*A* | *B*) = | A ∩ B | / | B | = 9 / 18 = 0.5

P(*A* | *B*) = P(*A*, *B*) / P(*B*) <— *Bayes rule*

P(*A*, *B*) = P(*A* | *B*) • P(*B*) <— *multiplication rule*

Probability mini-review

Multiplication rule for joint prob with many variables:

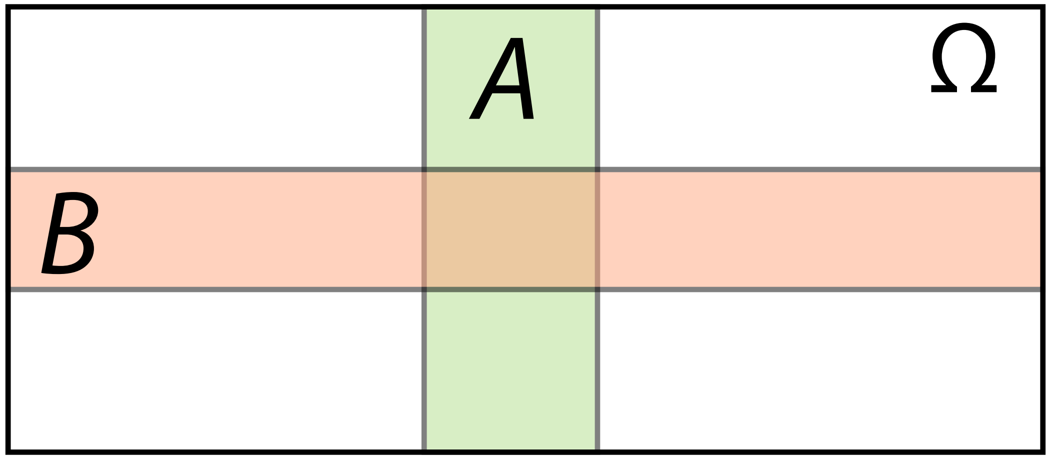
$$\mathrm{P}(A, B, C, D) = \mathrm{P}(A \mid B, C, D) \bullet \mathrm{P}(B, C, D)$$

$$\mathrm{P}(A, B, C, D) = \mathrm{P}(B \mid A, C, D) \bullet \mathrm{P}(A, C, D)$$

$$\mathrm{P}(A, B, C, D) = \mathrm{P}(A \mid B, C, D) \bullet \mathrm{P}(B \mid C, D) \bullet \mathrm{P}(C, D)$$

$$\mathrm{P}(A, B, C, D) = \underbrace{\mathrm{P}(A \mid B, C, D) \bullet \mathrm{P}(B \mid C, D) \bullet \mathrm{P}(C \mid D)}_{\text{conditional probabilities}} \bullet \underset{\text{marginal prob}}{\mathrm{P}(D)}$$

Probability mini-review



Events A and B are independent if $\mathrm{P}(A \mid B) = \mathrm{P}(A)$

So $\mathrm{P}(A, B) = \mathrm{P}(B)\, \mathrm{P}(A \mid B) = \mathrm{P}(A)\, \mathrm{P}(B)$

Probability mini-review

More probability review, courtesy of Prof. Joe Blitzstein and others:

http://projects.iq.harvard.edu/stat110/youtube

http://j.mp/CG_prob_cheatsheet

Sequence models

Sequence model is a *probabilistic model* that associates probabilities with sequences

What *k*-mers do I see inside versus outside of a CpG island?

What's the probability of next character
being A if previous characters were GATTAC?

Given a genome, where are the genes?

Right: model for eukaryotic gene finding

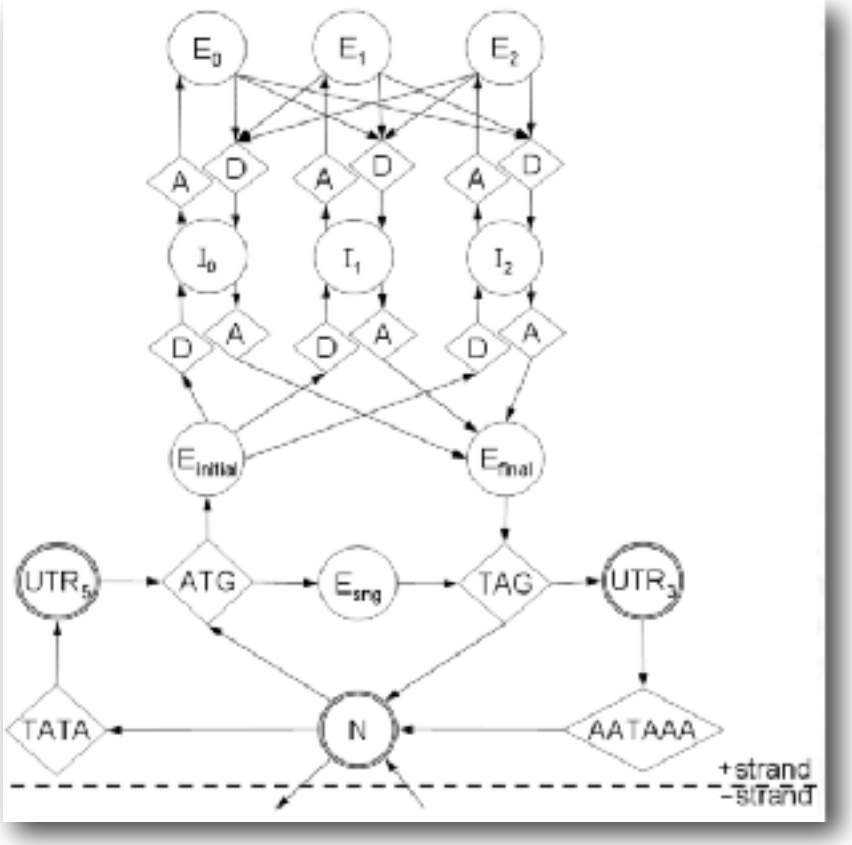


Image by Bill Majoros:
http://www.genezilla.org/design.html

Sequence models

We'll use sequence models that *learn from examples*

Say we sample 100K 5-mers from *inside* CpG islands and 100K 5-mers from *outside*

We're given a new 5-mer: CGCGC. Can we guess whether it came from a CpG island?

# CGCGC inside	315
# CGCGC outside	12

$p(\text{inside}) = 315/(315 + 12)= 0.963$

Sequence models

P(*x*) = probability of sequence *x*

$$P(\,x\,) = \underbrace{P(\,x_k, x_{k-1}, \ldots x_1\,)}$$

Joint probability of all
bases at all positions

Estimating P(*x*): # occurrences *inside* ÷ # occurrences total

For large *k*, might see few or no occurrences of *x*. Joint probabilities for rare events are hard to estimate well!