Sequence Modeling: Intro

Ben Langmead

From assembling to interpreting

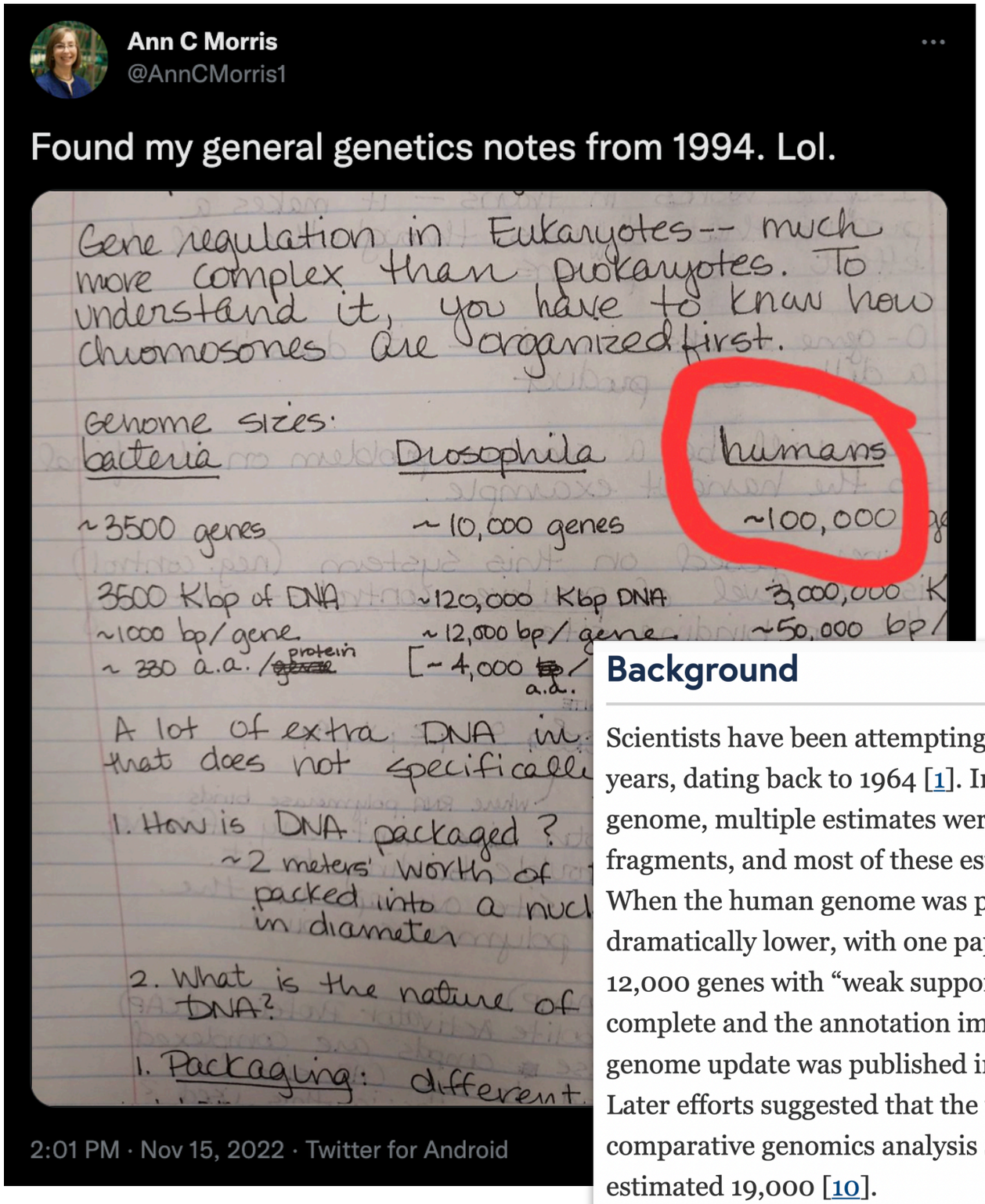
READING THE BOOK OF LIFE: THE OVERVIEW

READING THE BOOK OF LIFE: THE OVERVIEW; Genetic Code of Human Life Is Cracked by Scientists

By NICHOLAS WADE
Published: June 27, 2000

Photo: Ron Sachs/CNP/Corbis

How many genes are there?



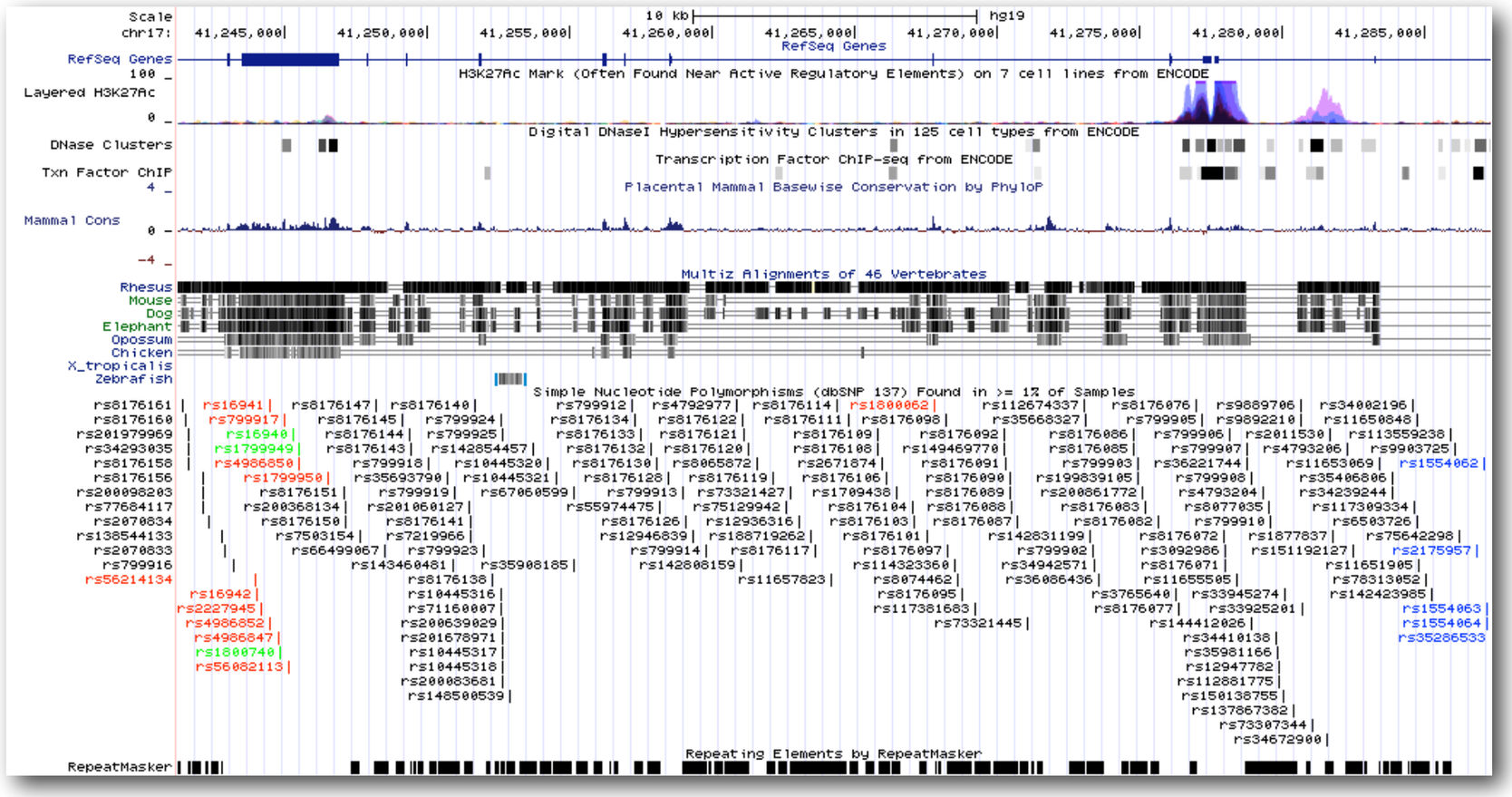
Pertea M, Shumate A, Pertea G, Varabyou A, Breitwieser FP, Chang YC, Madugundu AK, Pandey A, Salzberg SL. CHESS: a new human gene catalog curated from thousands of large-scale RNA sequencing experiments reveals extensive transcriptional noise. Genome Biol. 2018 Nov 28;19(1):208.

Background

Scientists have been attempting to estimate the number of human genes for more than 50 years, dating back to 1964 [1]. In the decade preceding the initial publication of the human genome, multiple estimates were made based on sequencing of short messenger RNA fragments, and most of these estimates fell in the range of 50,000–100,000 genes [2,3,4,5]. When the human genome was published in 2001, the estimates of the gene count were dramatically lower, with one paper reporting 31,000 genes [6] and the other 26,588 plus ~ 12,000 genes with “weak supporting evidence” [7]. As the genome was gradually made more complete and the annotation improved, the number continued to fall; when the first major genome update was published in 2004, the estimated gene count was revised to 24,000 [8]. Later efforts suggested that the true number of protein-coding genes was even smaller: a 2007 comparative genomics analysis suggested 20,500 [9], and a proteomics-based study in 2014 estimated 19,000 [10].

Picking up signals

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



genes
chromatin marks
open chromatin
transcription factor binding
sequence conservation

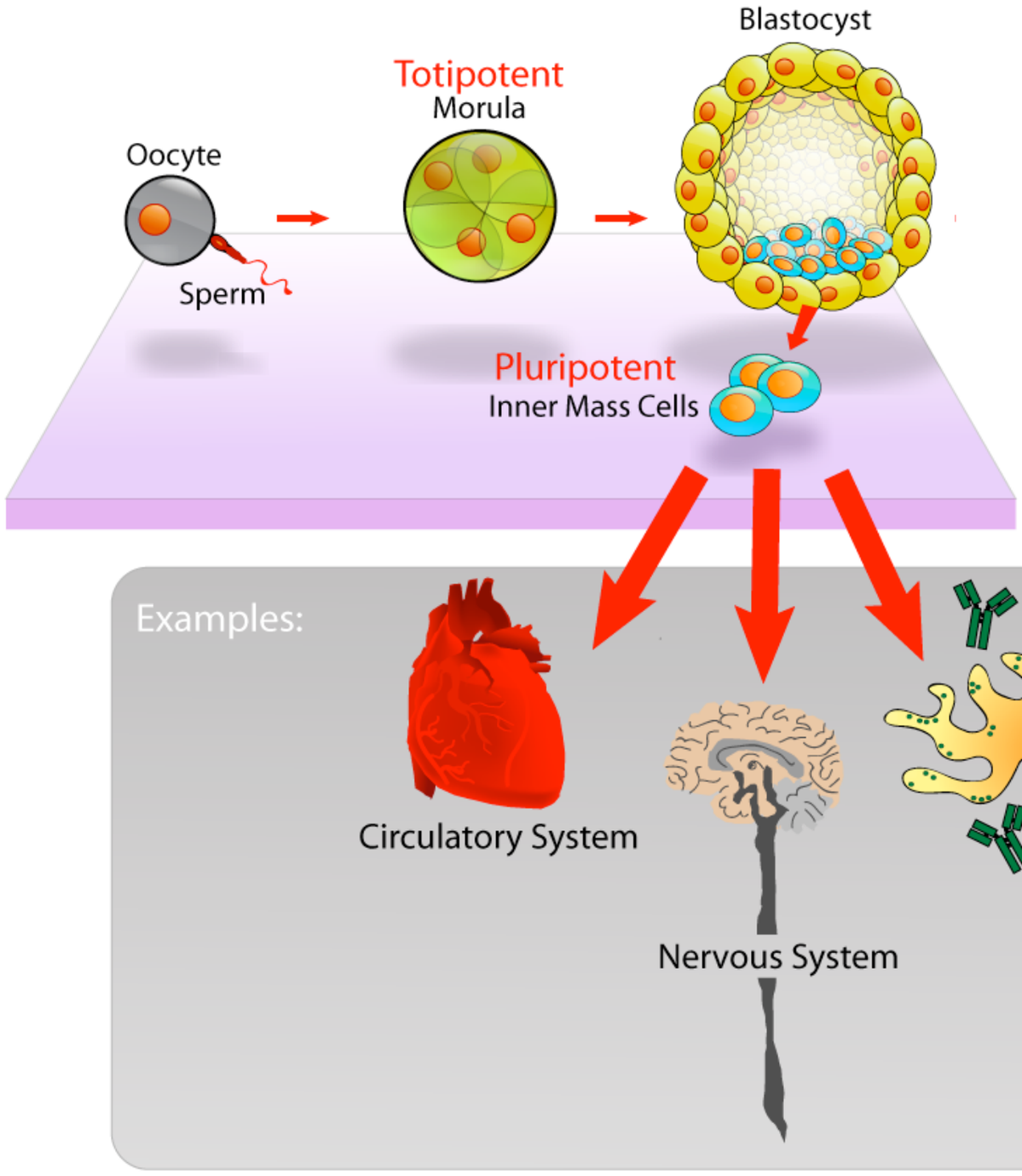
genetic variants

repetitive elements

40 K nt region of chromosome 17

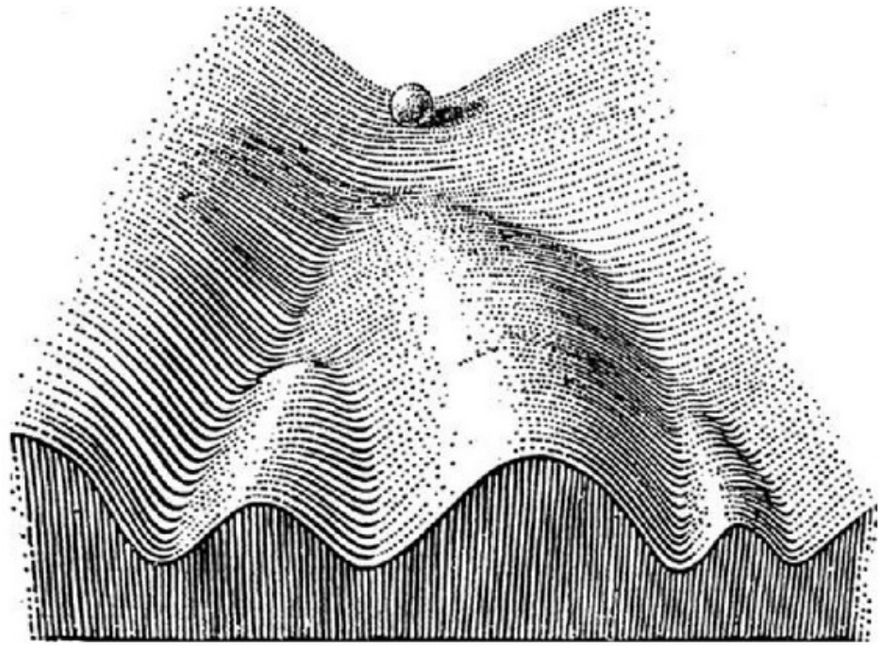
http://genome.ucsc.edu/cgi-bin/hgTracks

Epigenetics



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

"Waddington Landscape"



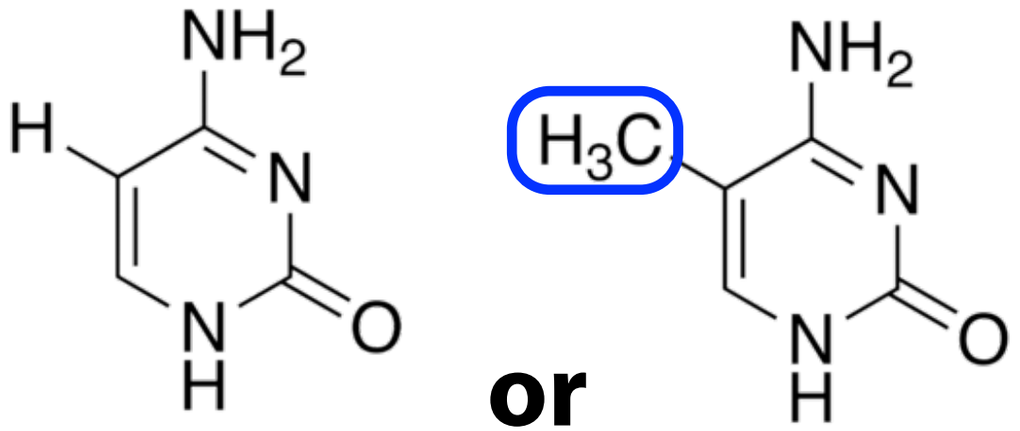
CH Waddington *The Strategy of the Genes* (Allen &Unwin, London, 1957).

Epigenetics

GATČATCGÁCGGTATĆGTGCÄTTCGATÇTATTT

Methylation

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



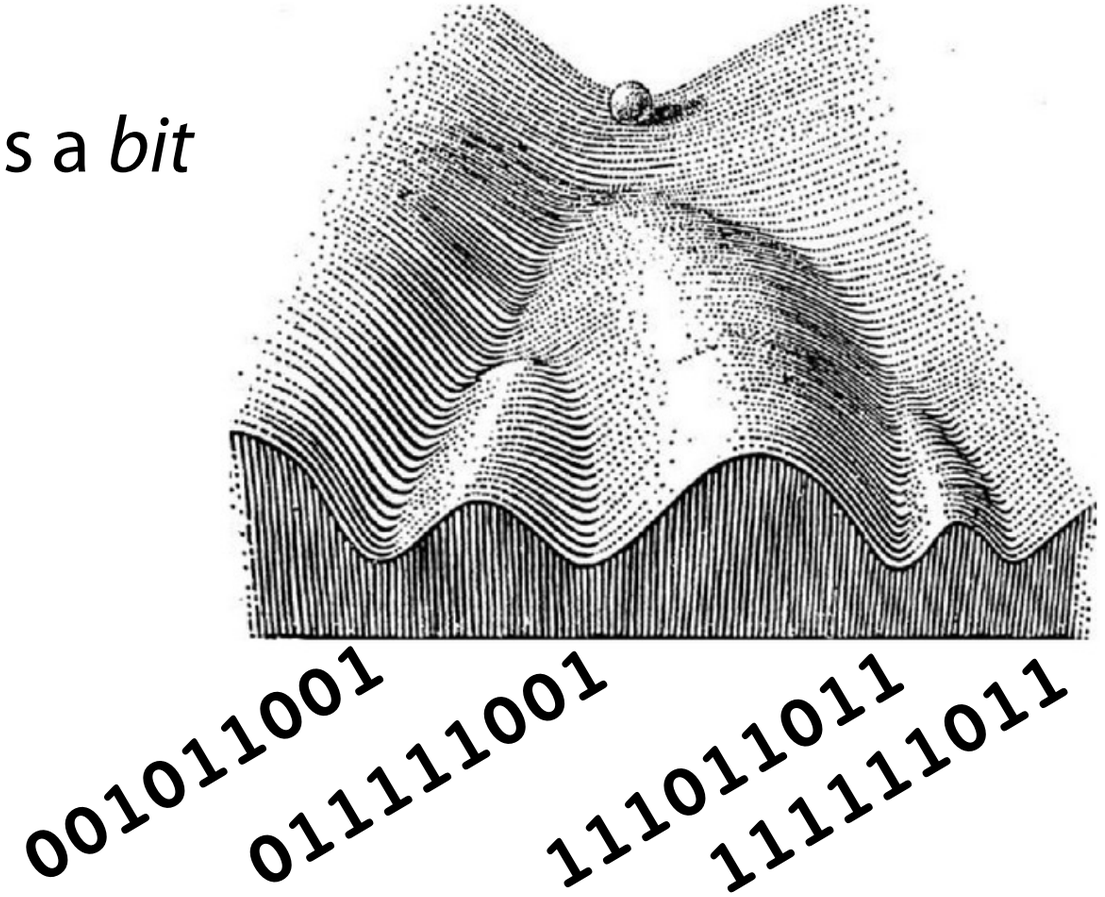
Unmethylated Methylated

Methylation

In animals, most methylation is at **C**G cytosines

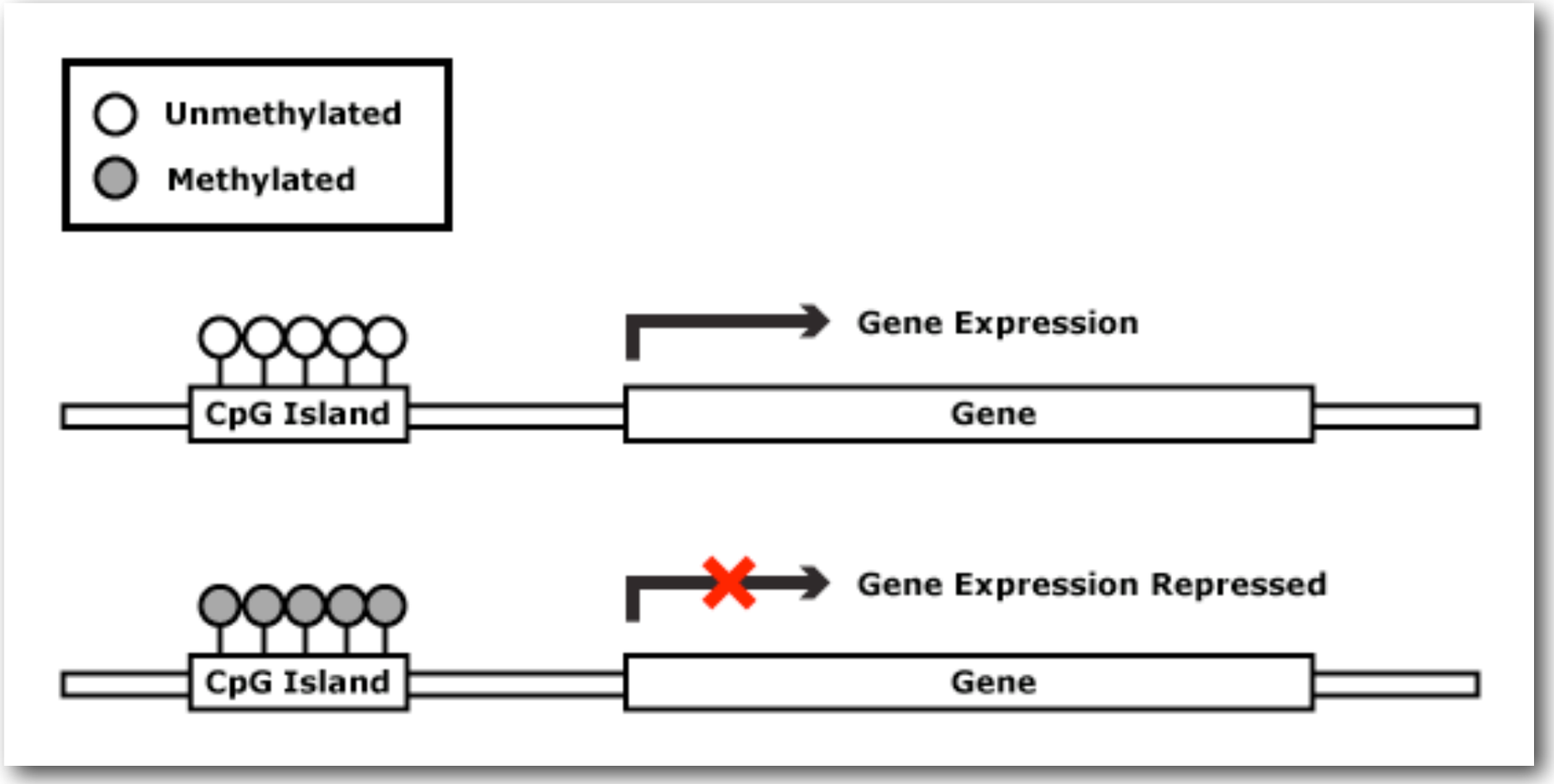
Methylation status of every CpG is a *bit*

Differentiated cell types have
different characteristic bit strings



CpG Islands

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

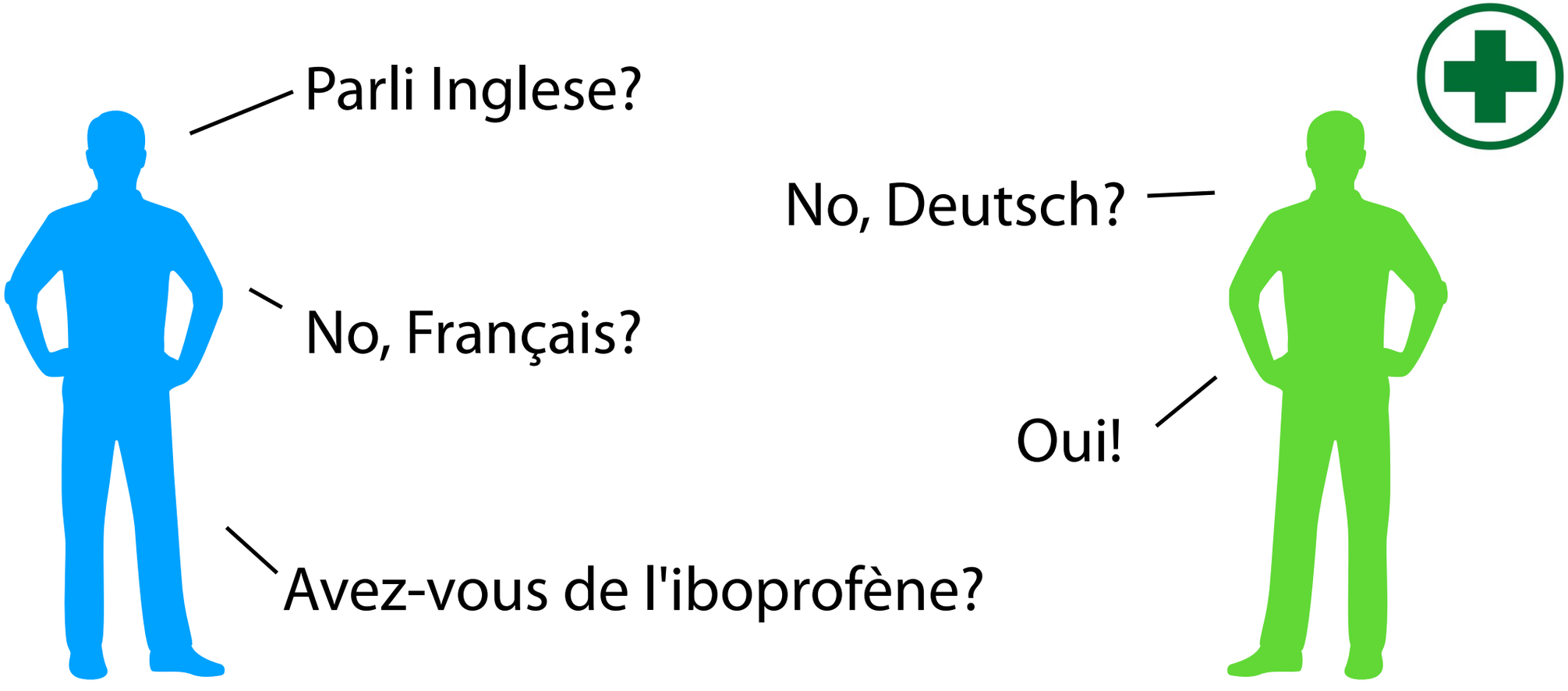


http://missinglink.ucsf.edu/lm/genes_and_genomes/methylation.html

Modeling goal

Wanted: strategy for scoring sequences according to their type, role, family; e.g. whether they belong to a CpG island

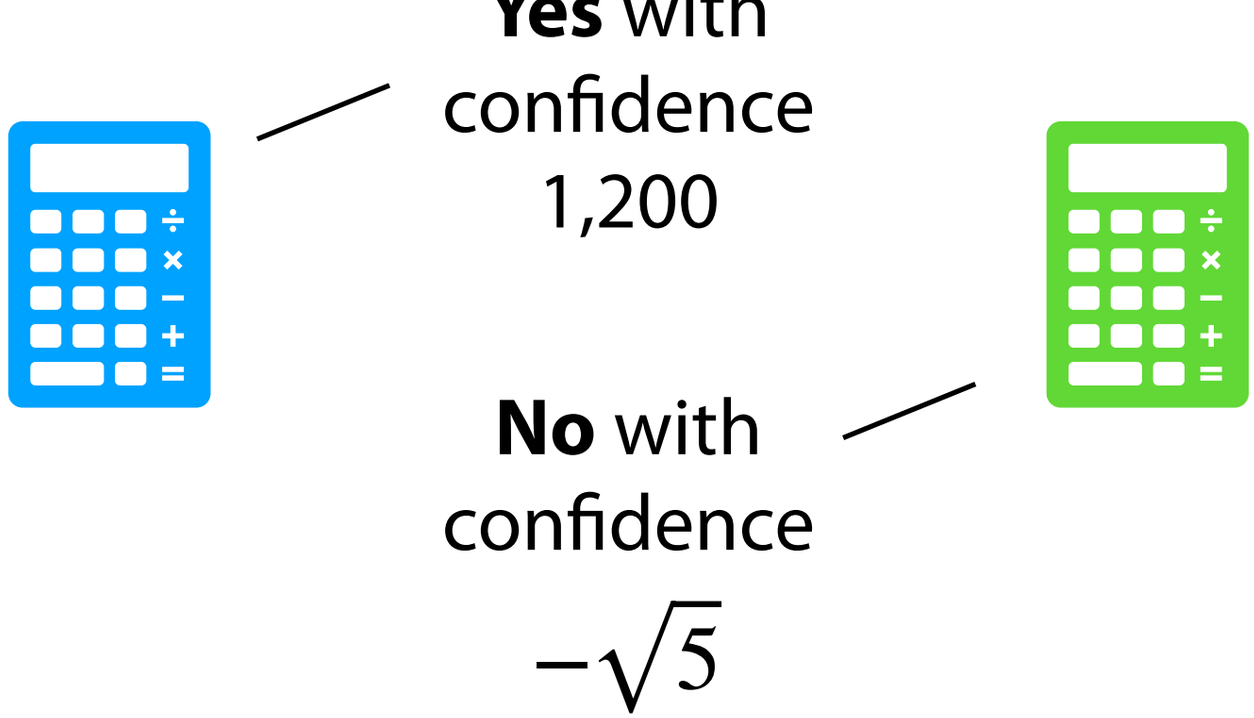
Scores should be *probabilities*



Modeling goal

Wanted: strategy for scoring sequences according to their type, role, family; e.g. whether they belong to a CpG island

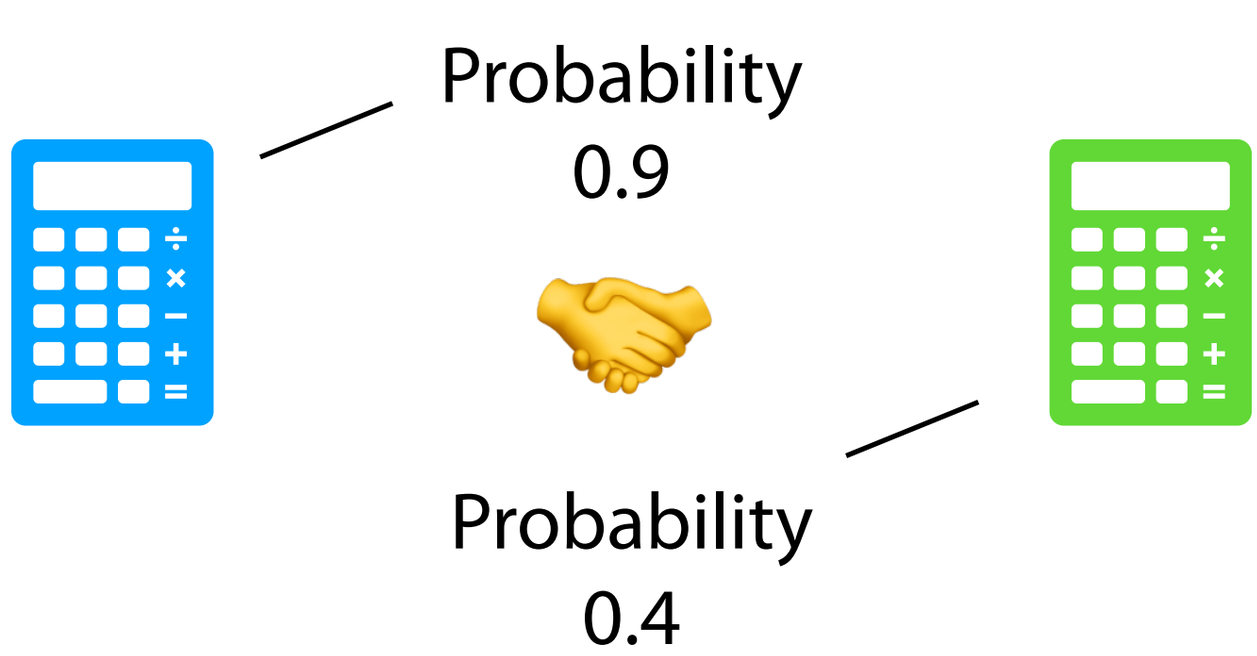
Scores should be *probabilities*



Modeling goal

Wanted: strategy for scoring sequences according to their type, role, family; e.g. whether they belong to a CpG island

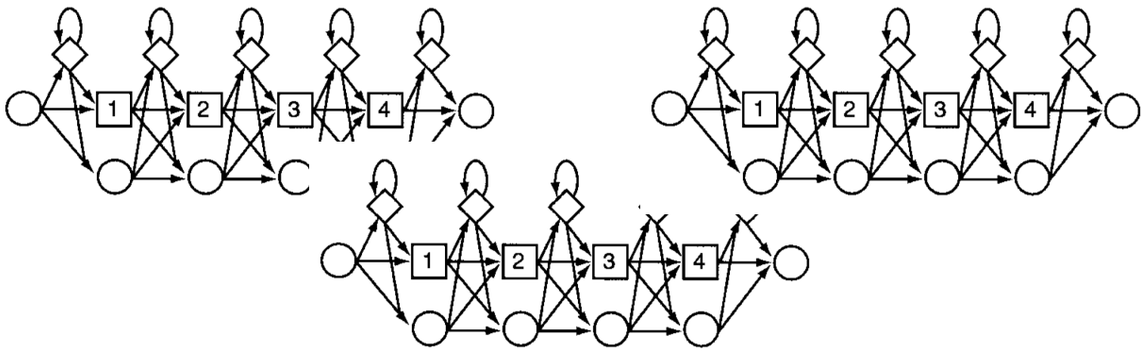
Scores should be *probabilities*



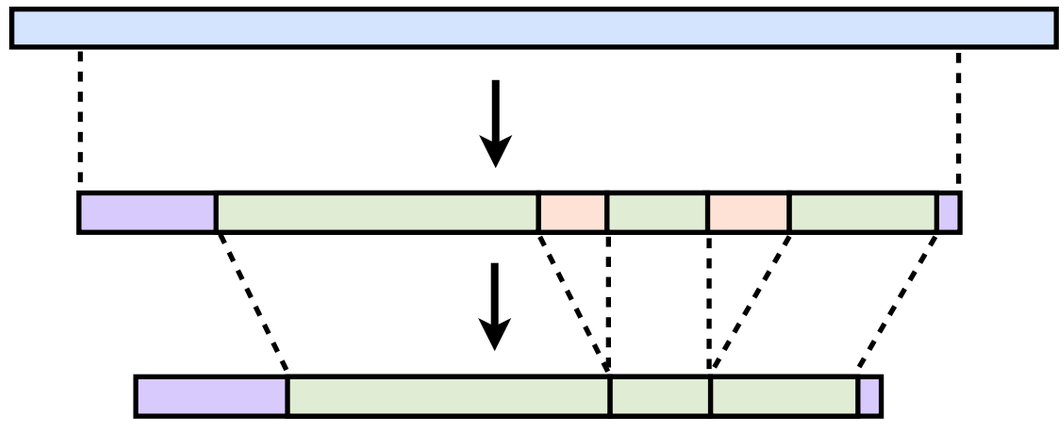
Sequence modeling strengths 💪

A natural fit for problems where we are:

- Putting sequences into ***categories*** *or* ***groups***

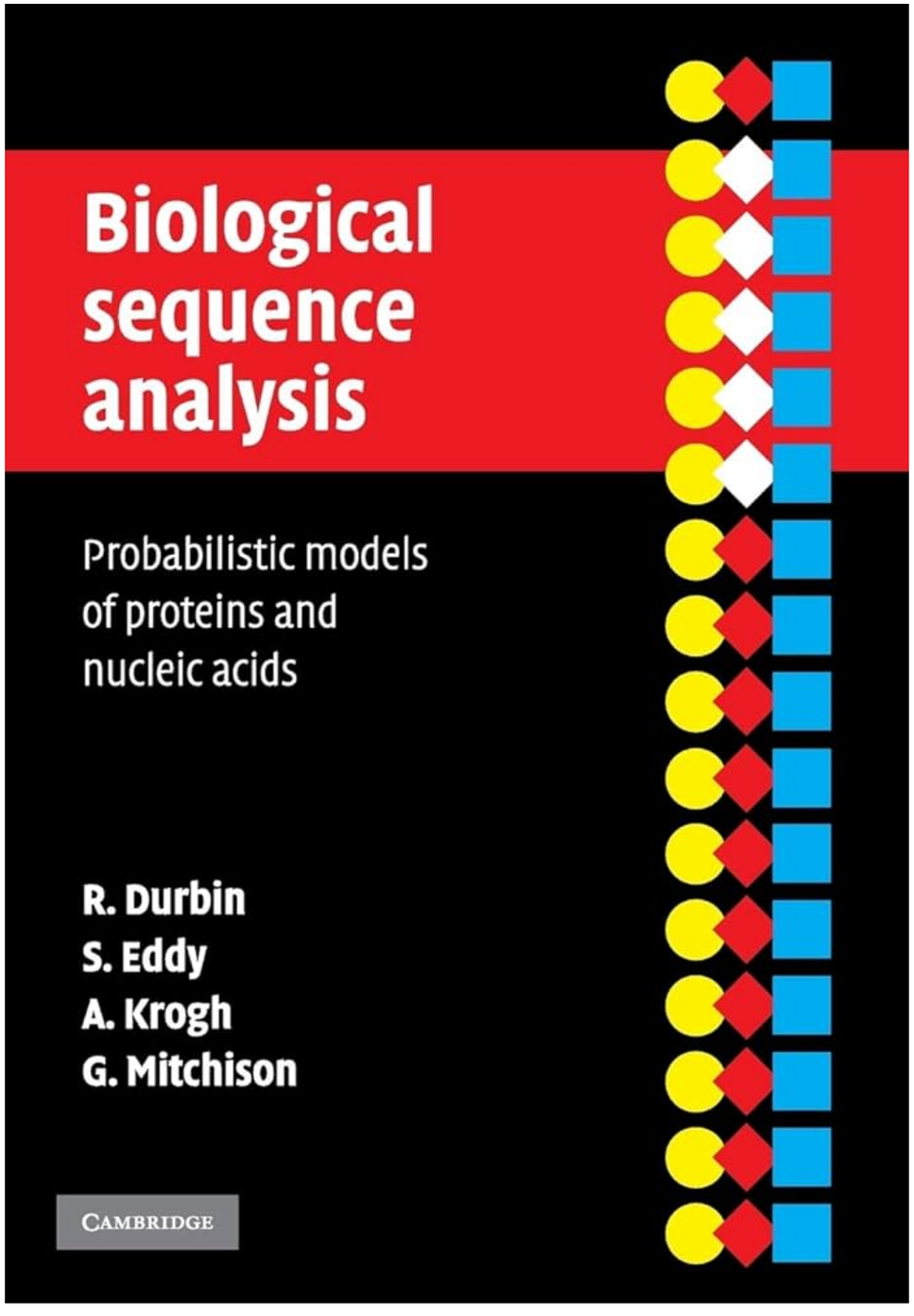


- Parsing a sequence into ***component pieces***



- **Inferring missing parts** of sequence

Reference


Biological sequence analysis
Probabilistic models of proteins and nucleic acids
R. Durbin
S. Eddy
A. Krogh
G. Mitchison
CAMBRIDGE