"The availability of genome sequence is just the beginning. Scientists now want to understand the genes and the role they play in the prevention, diagnosis and treatment of disease."
Dr Randy Scott, President of Incyte

Final words on assembly

- Assembly is resource intensive. As genomes get bigger, need extensive processing power (compute clusters).
- Repeats increase the time and space of these algorithms exponentially.
- Errors in the fragments from the sequencing instruments confound assembly.

de novo assembly

Sequencing

mapping reads

clustering reads

EBIO 5460

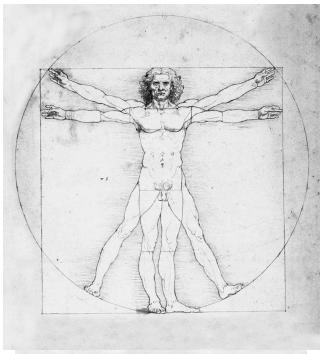
de novo assembly

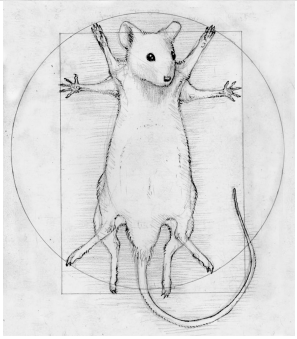
Reference Genome

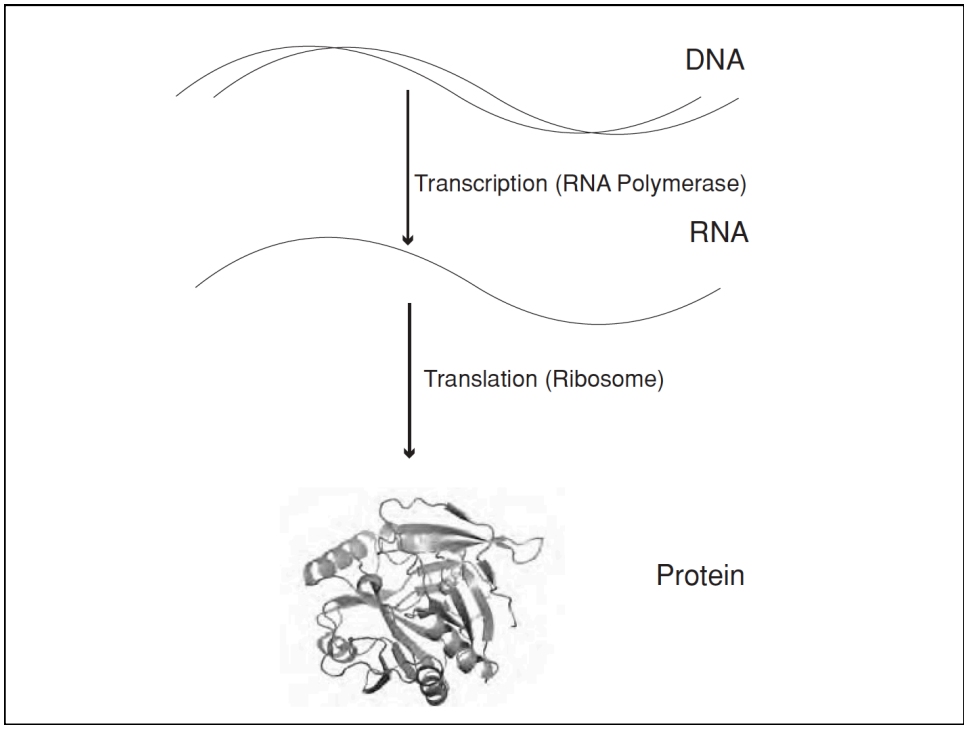
Sequencing

mapping reads

clustering reads



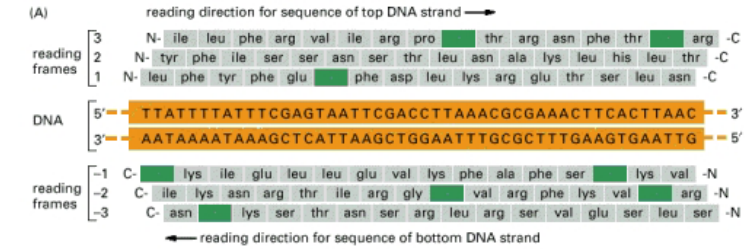


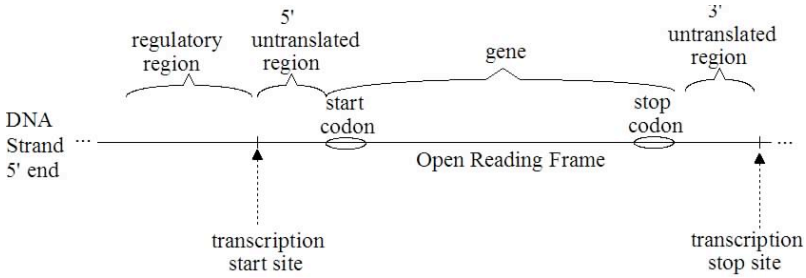


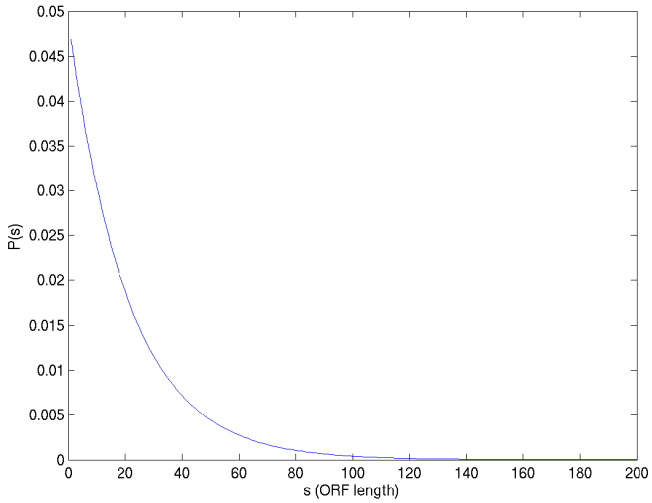
Second letter

First letter	U	C	A	G	Third letter
U	UUU Phe UUC Phe UUA Leu UUG Leu	UCU Ser UCC Ser UCA Ser UCG Ser	UAU Tyr UAC Tyr UAA Stop UAG Stop	UGU Cys UGC Cys UGA Stop UGG Trp	U C A G
C	CUU Leu CUC Leu CUA Leu CUG Leu	CCU Pro CCC Pro CCA Pro CCG Pro	CAU His CAC His CAA Gln CAG Gln	CGU Arg CGC Arg CGA Arg CGG Arg	U C A G
A	AUU Ile AUC Ile AUA Ile AUG Met	ACU Thr ACC Thr ACA Thr ACG Thr	AAU Asn AAC Asn AAA Lys AAG Lys	AGU Ser AGC Ser AGA Arg AGG Arg	U C A G
G	GUU Val GUC Val GUA Val GUG Val	GCU Ala GCC Ala GCA Ala GCG Ala	GAU Asp GAC Asp GAA Glu GAG Glu	GGU Gly GGC Gly GGA Gly GGG Gly	U C A G

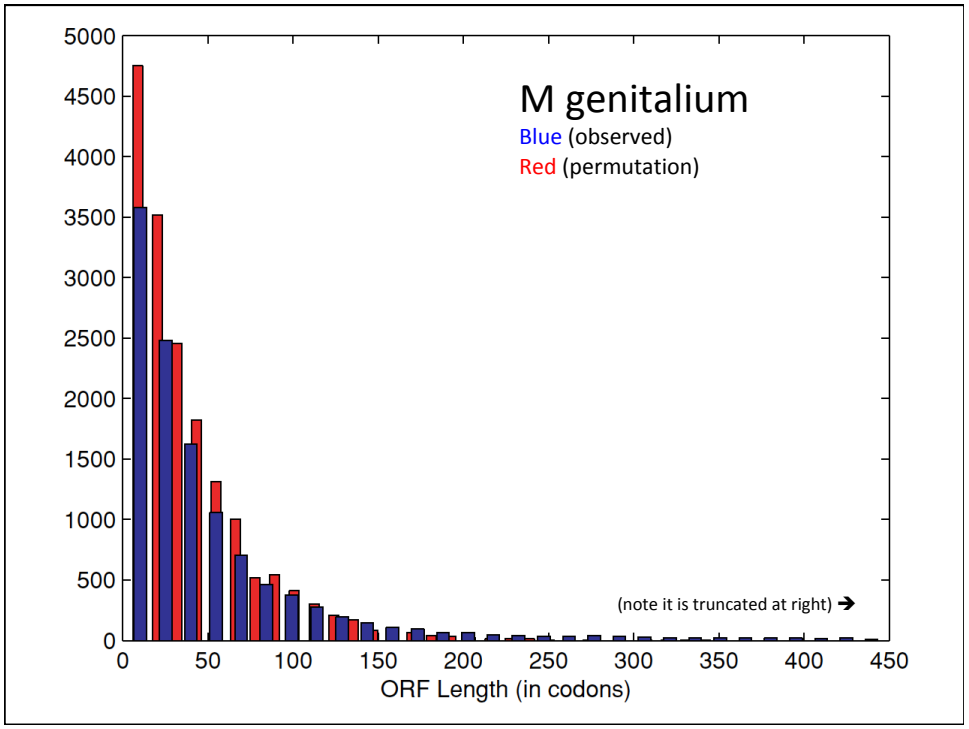
Any region of the DNA sequence can, in principle, code for six different amino acid sequences, because any one of three different reading frames can be used to interpret each of the two strands.





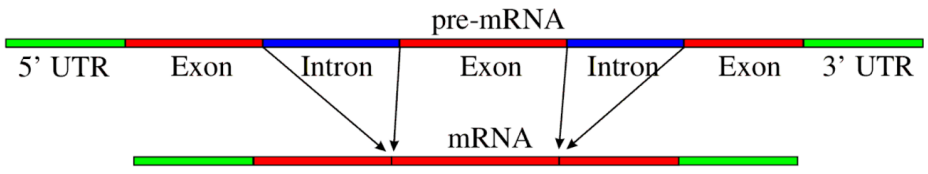


We note that P(s) follows a geometric distribution.



Eukaryotic gene finding

(harder)



- Statistical models of codon usage
- Markov models of gene structure
- Comparative genomics approaches