Intro to Annotation

Michael Schatz

October 2, 2024

Lecture 11. Applied Comparative Genomics

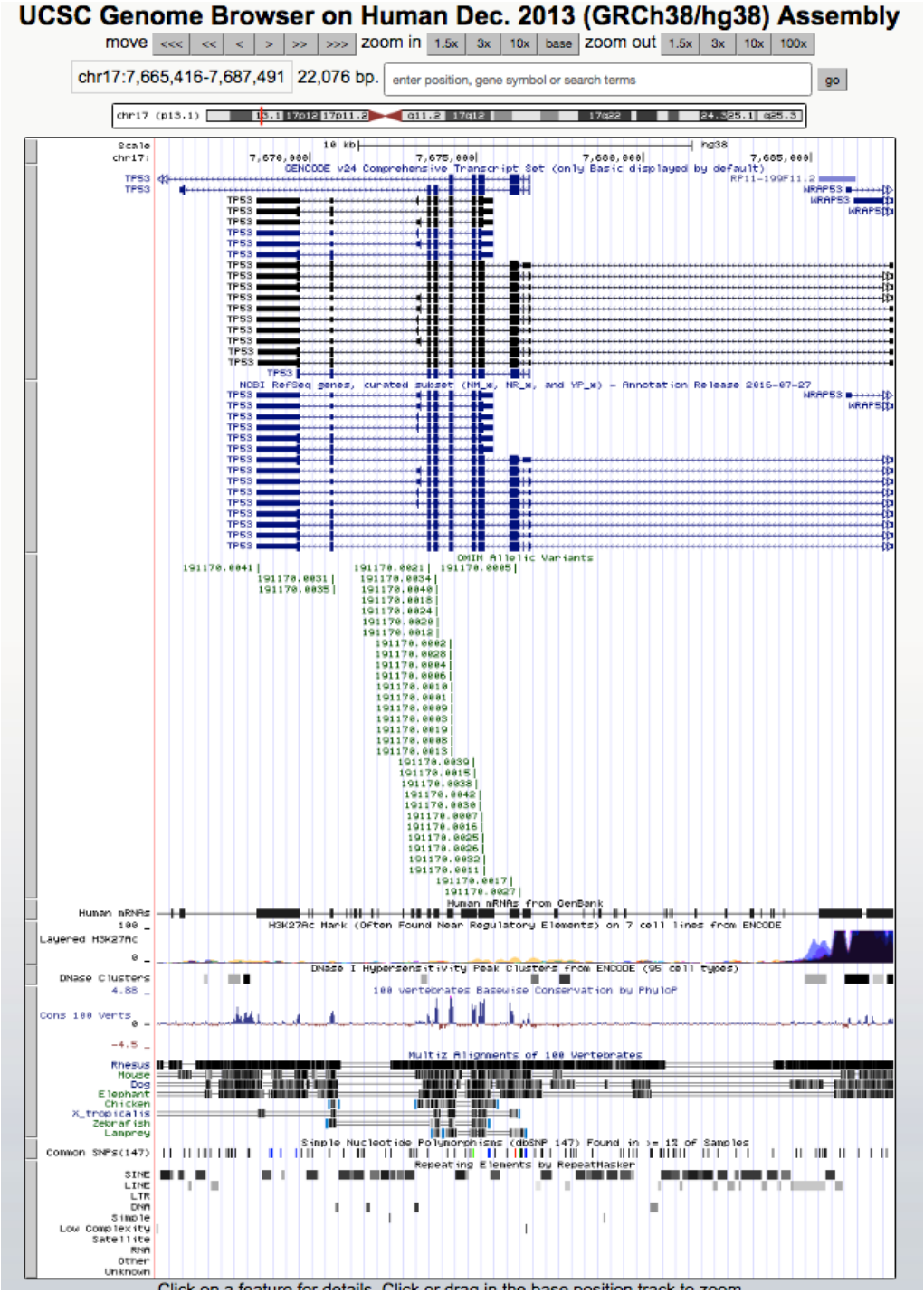
Goal: Genome Annotations

aatgcatgcggctatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaat
gcatgcggctatgcaagctgggatccgatgactatgctaagctgggatccgatgacaatgcatgcggctatgct
aatgaatggtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaa
tggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagctgggatccgatgacaatgcatgcg
gctatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctgcggctatgctaatgcatgcg
gctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatcct
gcggctatgctaatgaatggtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggct
atgctaatgaatggtcttgggatttaccttggaatatgctaatgcatgcggctatgctaagctgggaatgcatgcg
gctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatccg
atgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctcatgcggctatgctaagctgggaat
gcatgcggctatgctaagctgggatccgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctg
ggatccgatgactatgctaagctgcggctatgctaatgcatgcggctatgctaagctcggctatgctaatgaatg
gtcttgggatttaccttggaatgctaagctgggatccgatgacaatgcatgcggctatgctaatgaatggtcttgg
gatttaccttggaatatgctaatgcatgcggctatgctaagctgggaatgcatgcggctatgctaagctgggatc
cgatgacaatgcatgcggctatgctaatgcatgcggctatgcaagctgggatccgatgactatgctaagctgcg
gctatgctaatgcatgcggctatgctaagctcatgcgg

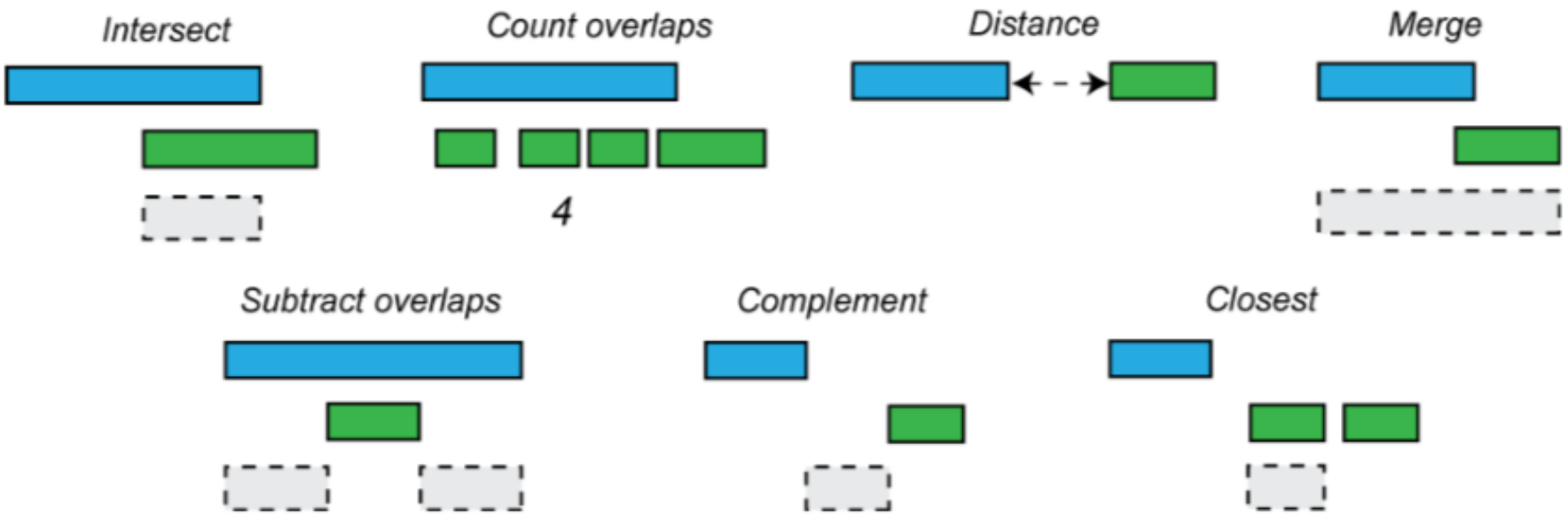
Goal: Genome Annotations

What are genome intervals?

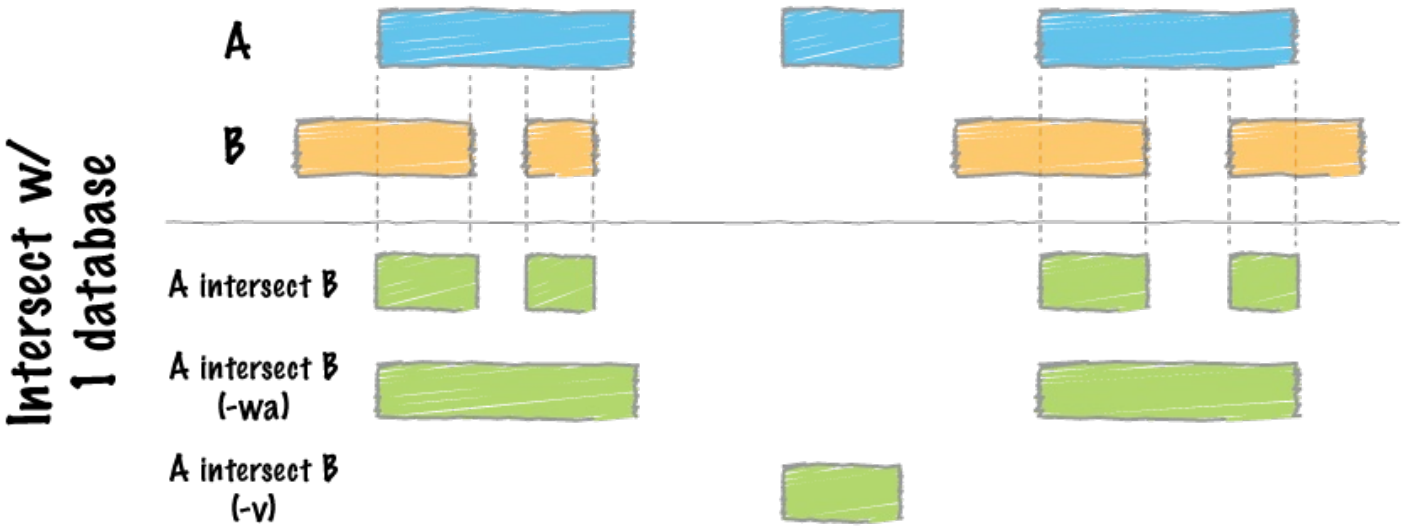
- Genetic variation:
  - SNPs: 1bp
  - Indels: 1-50bp
  - SVs: >50bp
- Genes:
  - exons, introns, UTRs, promoters
- Conservation
- Transposons
- Origins of replication
- TF binding sites
- CpG islands
- Segmental duplications
- Sequence alignments
- Chromatin annotations
- Gene expression data
- …
- ***Your own observations and data: put them into context!***



BEDTools to the rescue!



BEDTools Intersect



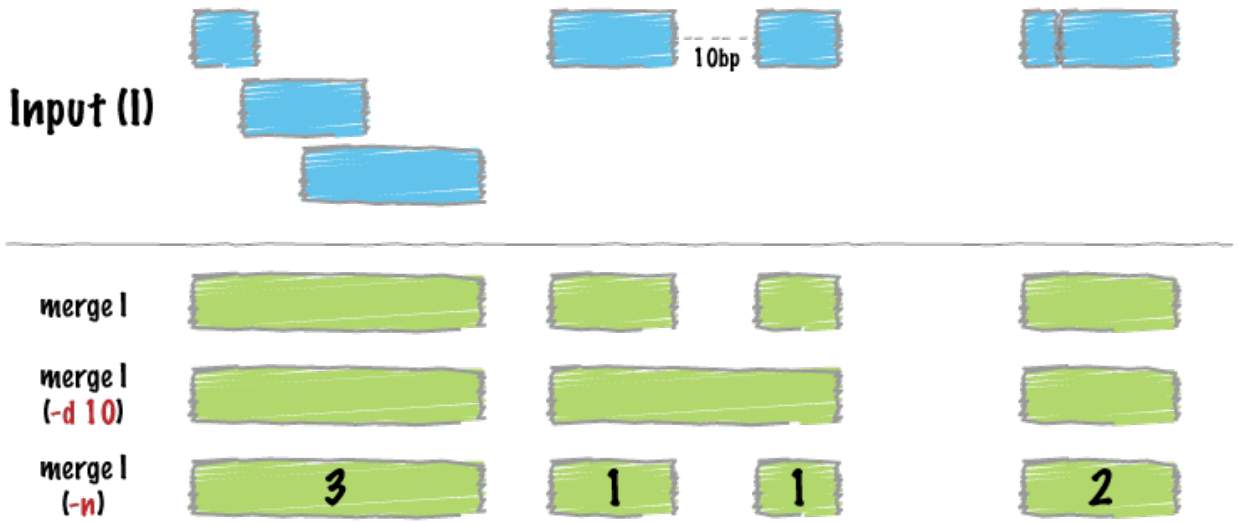
What exons are hit by SVs?

```
$ cat A.bed
chr1  10  20
chr1  30   40

$ cat B.bed
chr1  15  20

$ bedtools intersect -a A.bed -b B.bed -wa
chr1  10   20
```

What parts of exons are hit by SVs?

```
$ cat A.bed
chr1  10  20
chr1  30  40

$ cat B.bed
chr1  15   20

$ bedtools intersect -a A.bed -b B.bed
chr1  15   20
```

BEDTools Merge



What parts of the genome are exonic?

```
bedtools merge -i exons.bed | head -n 20
chr1    11873   12227
chr1    12612   12721
chr1    13220   14829
chr1    14969   15038
chr1    15795   15947
chr1    16606   16765
chr1    16857   17055
```

Note input must be sorted!

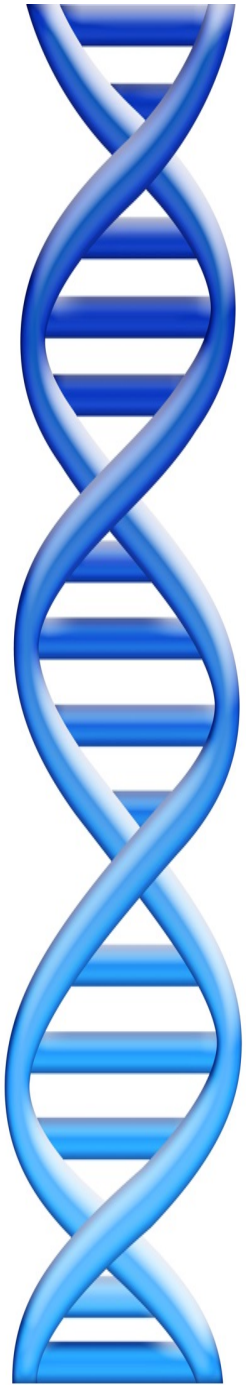
```
sort -k1,1 -k2,2n foo.bed > foo.sort.bed
```

BEDTools commands

```
annotate
bamtobed
bamtofastq
bed12tobed6
bedpetobam
bedtobam
closest
cluster
complement
coverage
expand
flank
fisher
genomecov
getfasta
groupby
groupby
igv
intersect
jaccard
links
makewindows
map
maskfasta
merge
multicov
multiinter
nuc
overlap
pairtobed
pairtopair
random
reldist
shift
shuffle
slop
sort
subtract
tag
unionbedg
window
```

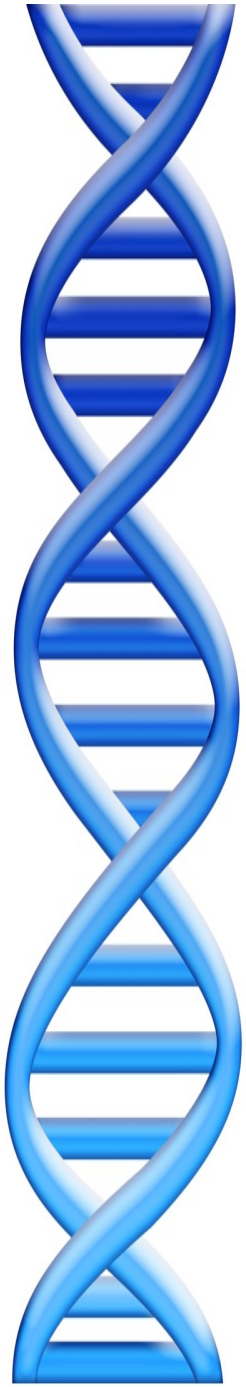
http://bedtools.readthedocs.io/en/latest/content/bedtools-suite.html

Goal: Genome Annotations



Outline

1. Alignment to other genomes
2. Prediction aka "Gene Finding"
3. Experimental & Functional Assays



Outline

1. Alignment to other genomes
2. Prediction aka "Gene Finding"
3. Experimental & Functional Assays

Basic Local Alignment Search Tool

- Rapidly compare a sequence Q to a database to find all sequences in the database with a score above some cutoff S.
  - Which protein is most similar to a newly sequenced one?
  - Where does this sequence of DNA originate?

- Speed achieved by using a procedure that typically finds "most" matches with scores > S.
  - Tradeoff between sensitivity and specificity/speed
    - Sensitivity – ability to find all related sequences
    - Specificity – ability to reject unrelated sequences

(Altschul et al. 1990)

Seed and Extend

```
FAKDFLAGGVAAAISKTAVAPIERVKLLLQVQHASKQITADKQYKGIIDCVVRIPKEQGV

FLIDLASGGTAAAVSKTAVAPIERVKLLLQVQDASKAIAVDKRYKGIMDVLIRVPKEQGV
```

- Homologous sequences are likely to contain a short high scoring word pair, a seed.
  - Smaller seed sizes make the sense more sensitive, but also (much) slower
  - Typically do a fast search for prototypes, but then most sensitive for final result

- BLAST then tries to extend high scoring word pairs to compute high scoring segment pairs (HSPs).
  - Significance of the alignment reported via an e-value

Seed and Extend

```
FAKDFLAGGVAAAISKTAVAPIERVKLLLQVQHASKQITADKQYKGIIDCVVRIPKEQGV
|   |     | |||  ||||||||||||||||||  ||| |   |   |||| |    | ||||||
FLIDLASGGTAAAVSKTAVAPIERVKLLLQVQDASKAIAVDKRYKGIMDVLIRVPKEQGV
```

- Homologous sequences are likely to contain a short high scoring word pair, a seed.
  - Smaller seed sizes make the sense more sensitive, but also (much) slower
  - Typically do a fast search for prototypes, but then most sensitive for final result

- BLAST then tries to extend high scoring word pairs to compute high scoring segment pairs (HSPs).
  - Significance of the alignment reported via an e-value

BLAST E-values

E-value = the number of HSPs having alignment score S (or higher) expected to occur by chance.

→ Smaller E-value, more significant in statistics

→ Bigger E-value, less significant

→ Over 1 means expect this totally by chance

(not significant at all!)

The expected number of HSPs with the score at least S is :

$$E = K*n*m*e^{-\lambda S}$$

K, λ are constant depending on model

n, m are the length of query and sequence

E-values quickly drop off for better alignment bits scores

Very Similar Sequences

```
Query: HBA_HUMAN Hemoglobin alpha subunit
Sbjct: HBB_HUMAN Hemoglobin beta subunit

Score =  114 bits (285),  Expect = 1e-26
Identities = 61/145 (42%), Positives = 86/145 (59%), Gaps = 8/145 (5%)

Query  2   LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGSAQV 55
           L+P +K+ V A WGKV  +  E G EAL R+ + +P T+ +F  F       D   G+ +V
Sbjct  3   LTPEEKSAVTALWGKV--NVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVMGNPKV 60

Query  56  KGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPA 115
           K HGKKV  A ++ +AH+D++    + LS+LH  KL VDP NF+LL + L+  LA H
Sbjct  61  KAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCVLAHHFGK 120

Query  116 EFTPAVHASLDKFLASVSTVLTSKY 140
           EFTP V A+  K +A V+  L  KY
Sbjct  121 EFTPPVQAAYQKVVAGVANALAHKY 145
```

Quite Similar Sequences

```
Query: HBA_HUMAN Hemoglobin alpha subunit
Sbjct: MYG_HUMAN Myoglobin

Score = 51.2 bits (121), Expect = 1e-07,
Identities = 38/146 (26%), Positives = 58/146 (39%), Gaps = 6/146 (4%)

Query  2   LSPADKTNVKAAWGKVGAHAGEYGAEALERMFLSFPTTKTYFPHF------DLSHGSAQV  55
           LS  +   V   WGKV A    +G E L R+F   P T   F  F      D    S  +
Sbjct  3   LSDGEWQLVLNVWGKVEADIPGHGQEVLIRLFKGHPETLEKFDKFKHLKSEDEMKASEDL  62

Query  56  KGHGKKVADALTNAVAHVDDMPNALSALSDLHAHKLRVDPVNFKLLSHCLLVTLAAHLPA  115
           K HG  V  AL   +          +  L+  HA K ++     + +S C++  L +  P
Sbjct  63  KKHGATVLTALGGILKKKGHHEAEIKPLAQSHATKHKIPVKYLEFISECIIQVLQSKHPG  122

Query  116 EFTPAVHASLDKFLASVSTVLTSKYR  141
            +F      +++K L      + S Y+
Sbjct  123 DFGADAQGAMNKALELFRKDMASNYK  148
```

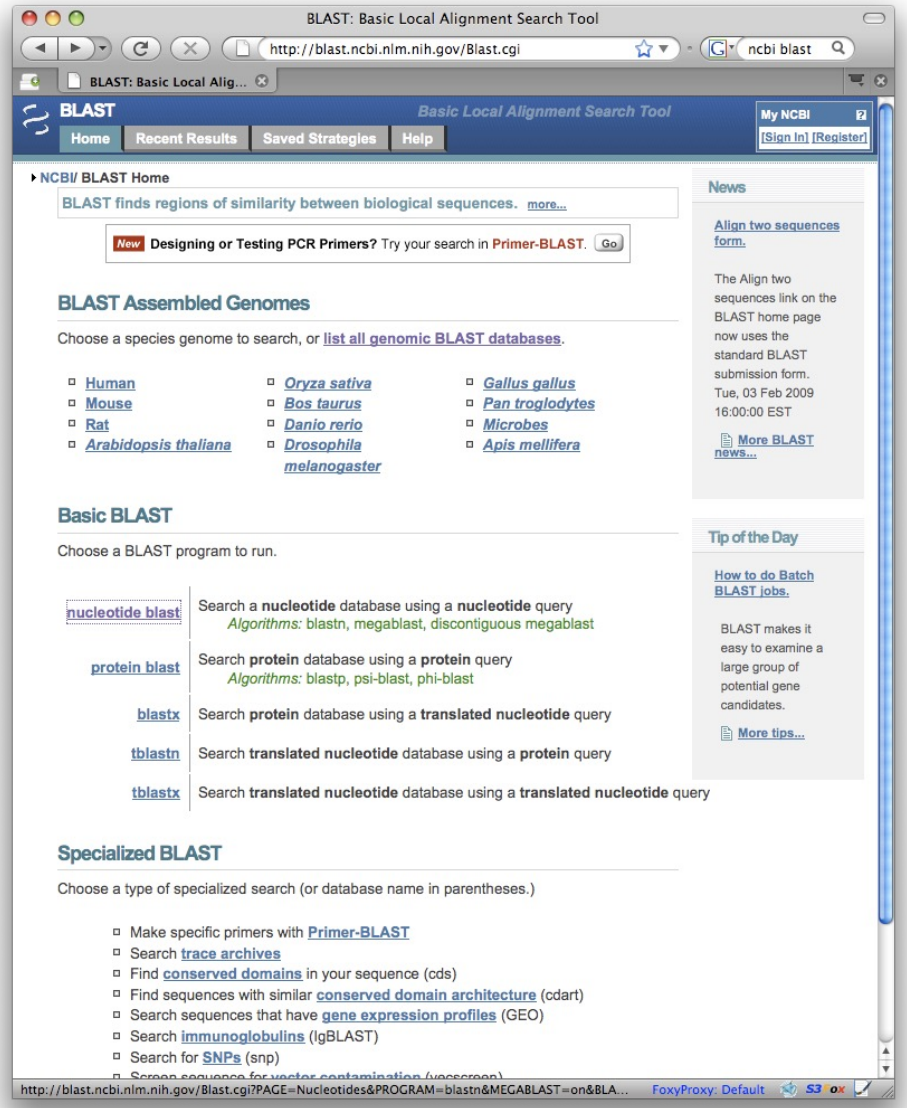
Not similar sequences

```
Query: HBA_HUMAN Hemoglobin alpha subunit
Sbjct: SPAC869.02c [Schizosaccharomyces pombe]

 Score = 33.1 bits (74),  Expect = 0.24
 Identities = 27/95 (28%), Positives = 50/95 (52%), Gaps = 10/95 (10%)

Query  30  ERMFLSFPTTKTYFPHFDLSHGSAQVKGHGKKVADALTNAVAHVDDMPNALSALSDLHAH  89
           ++M  ++P      P+F+ +H  +      + +A AL N   ++DD+  +LSA  D
Sbjct  59  QKMLGNYPEV---LPYFNKAHQISL--SQPRILAFALLNYAKNIDDL-TSLSAFMDQIVV 112

Query  90  K---LRVDPVNFKLLSHCLLVTLAAHLPAEF-TPA  120
           K   L++   ++ ++ HCLL T+   LP++  TPA
Sbjct  113 KHVGLQIKAEHYPIVGHCLLSTMQELLPSDVATPA  147
```

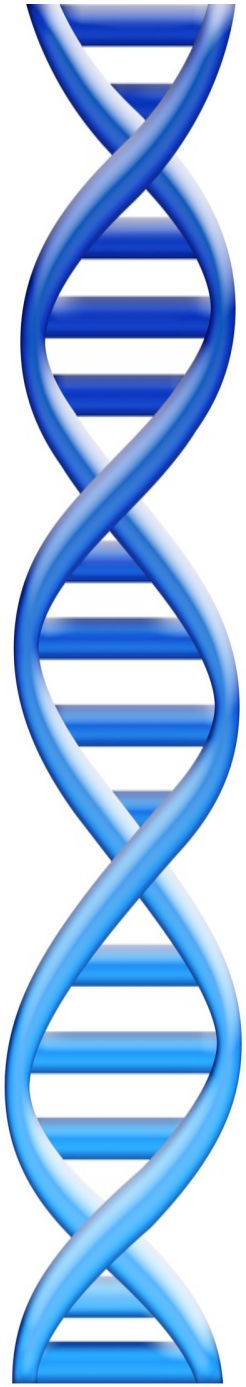
Blast Versions

Program	Database	Query
BLASTN	Nucleotide	Nucleotide
BLASTP	Protein	Protein
BLASTX	Protein	Nucleotide translated into protein
TBLASTN	Nucleotide translated into protein	Protein
TBLASTX	Nucleotide translated into protein	Nucleotide translated into protein

NCBI Blast



- Nucleotide Databases
  - nr: All Genbank
  - refseq: Reference organisms
  - wgs: All reads
- Protein Databases
  - nr: All non-redundant sequences
  - Refseq: Reference proteins



Outline

1. Alignment to other genomes
2. Prediction aka "Gene Finding"
3. Experimental & Functional Assays

Bacterial Gene Finding and Glimmer

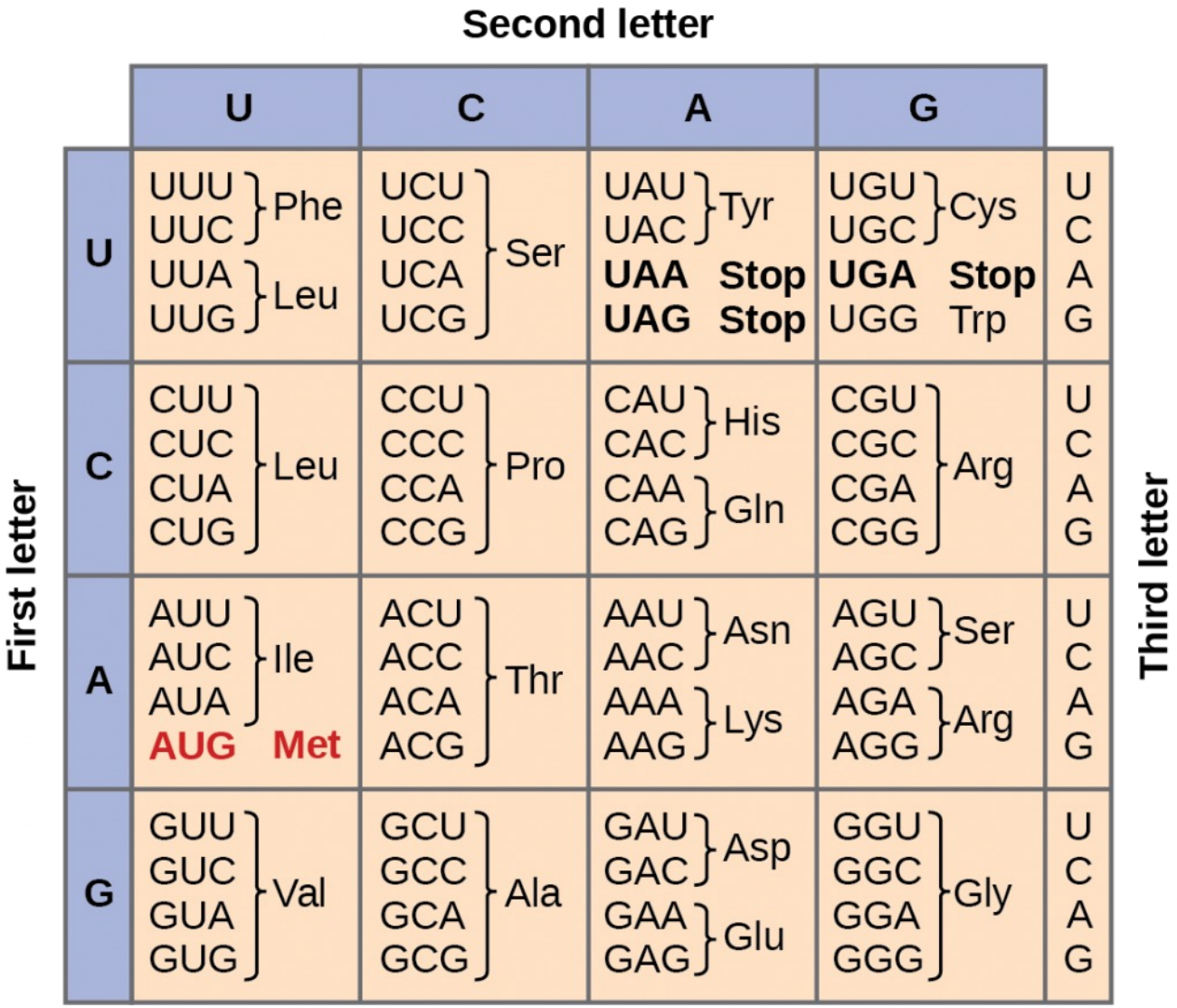
(also Archaeal and viral gene finding)

Arthur L. Delcher and Steven Salzberg

Center for Bioinformatics and Computational Biology

Johns Hopkins University

Genetic Code



Start:
- AUG

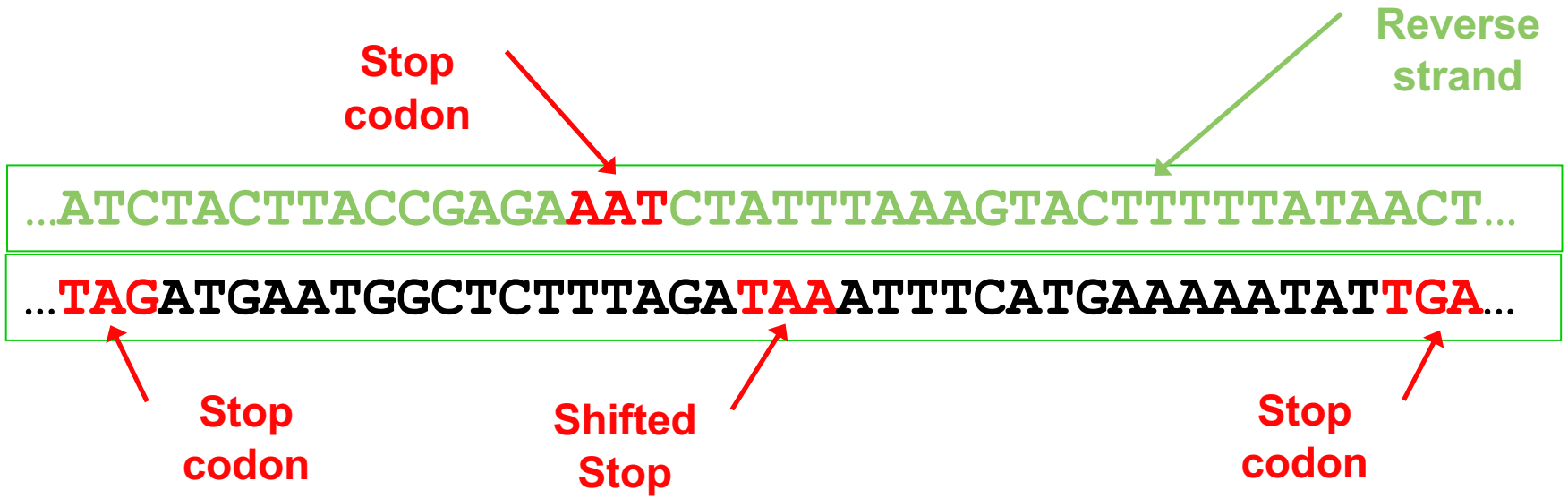
Stop:
- UAA
- UAG
- UGA

Step One

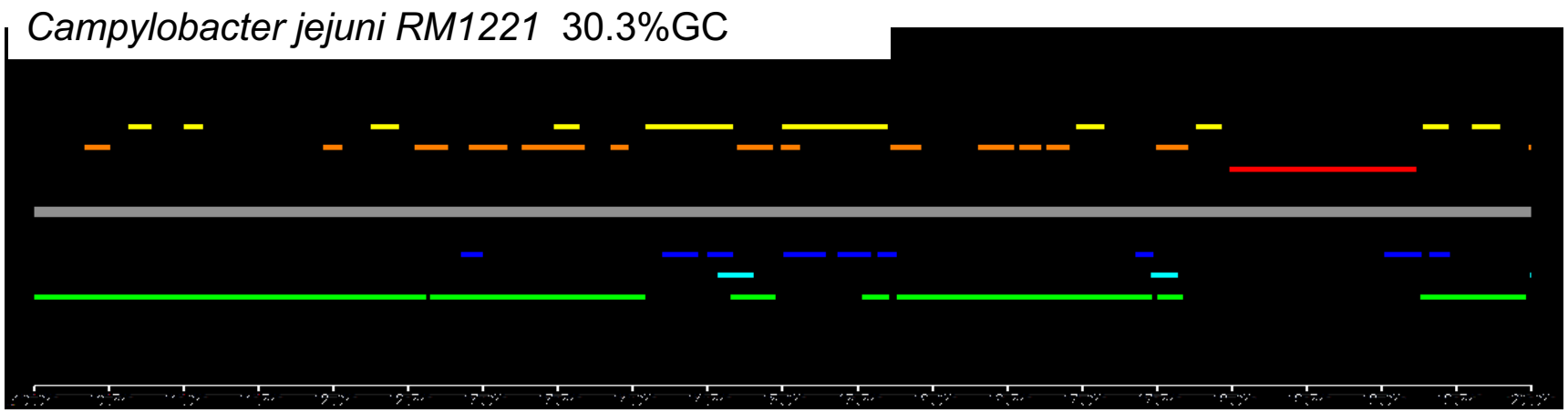
- Find open reading frames (ORFs).

Step One

- Find open reading frames (ORFs).



- But ORFs generally overlap …



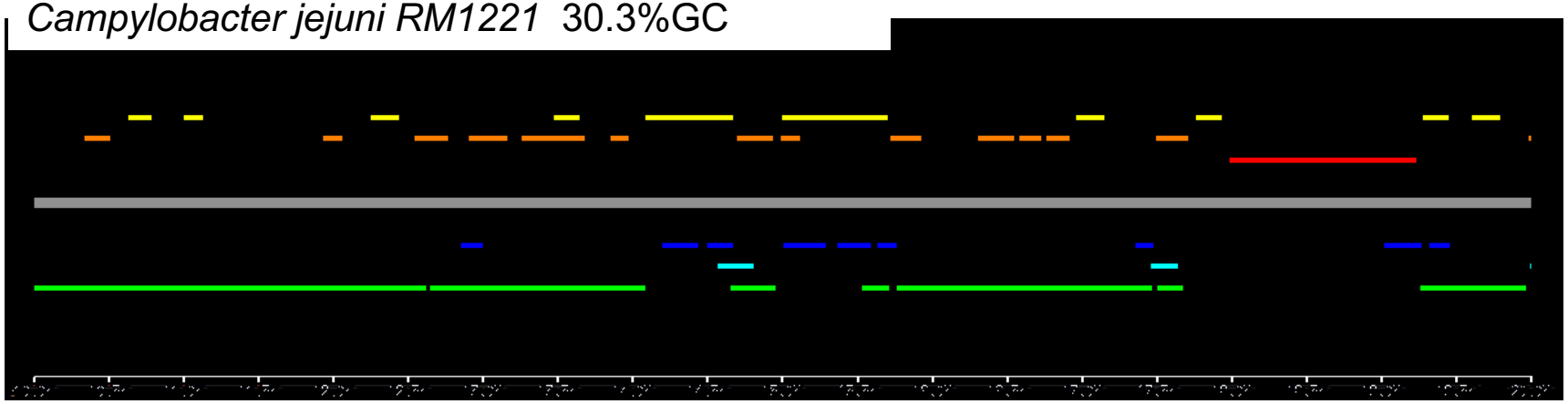
All ORFs longer than 100bp on both strands shown
- color indicates reading frame

Longest ORFs likely to be protein-coding genes

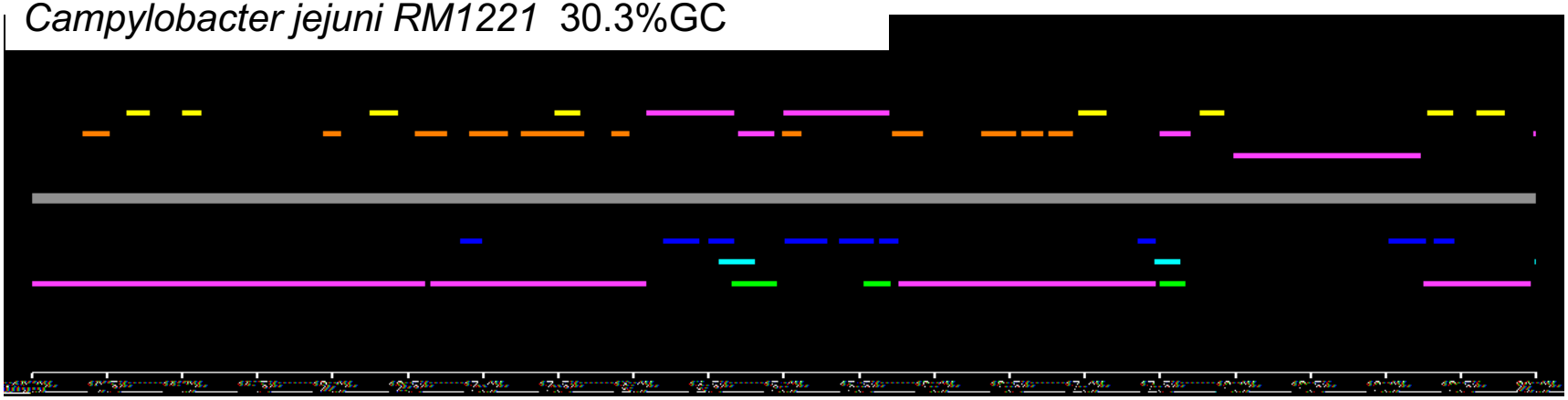
Note the low GC content

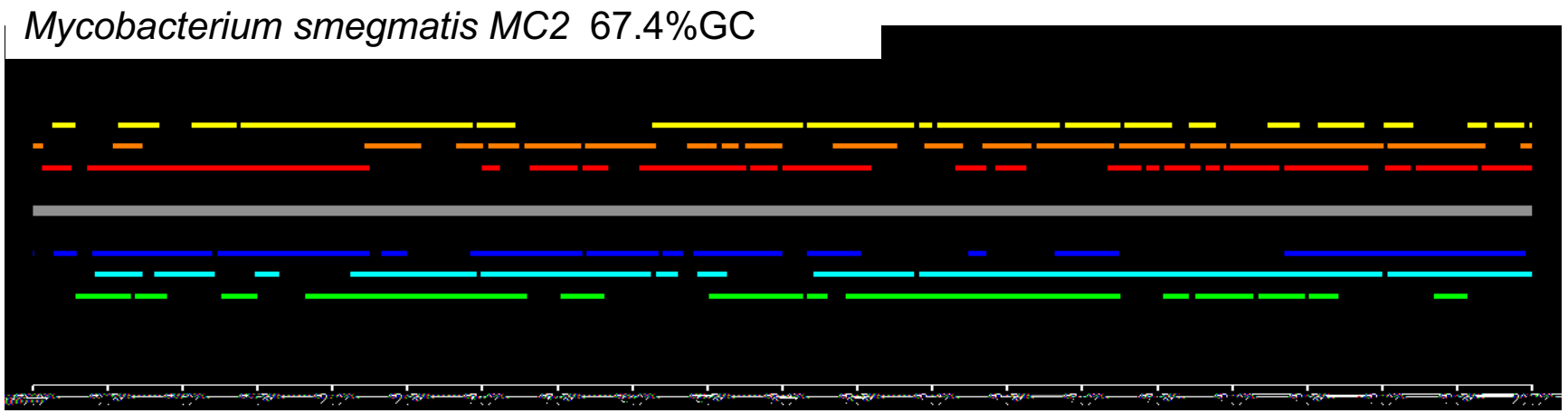
All genes are ORFs but not all ORFs are genes

Campylobacter jejuni RM1221 30.3%GC



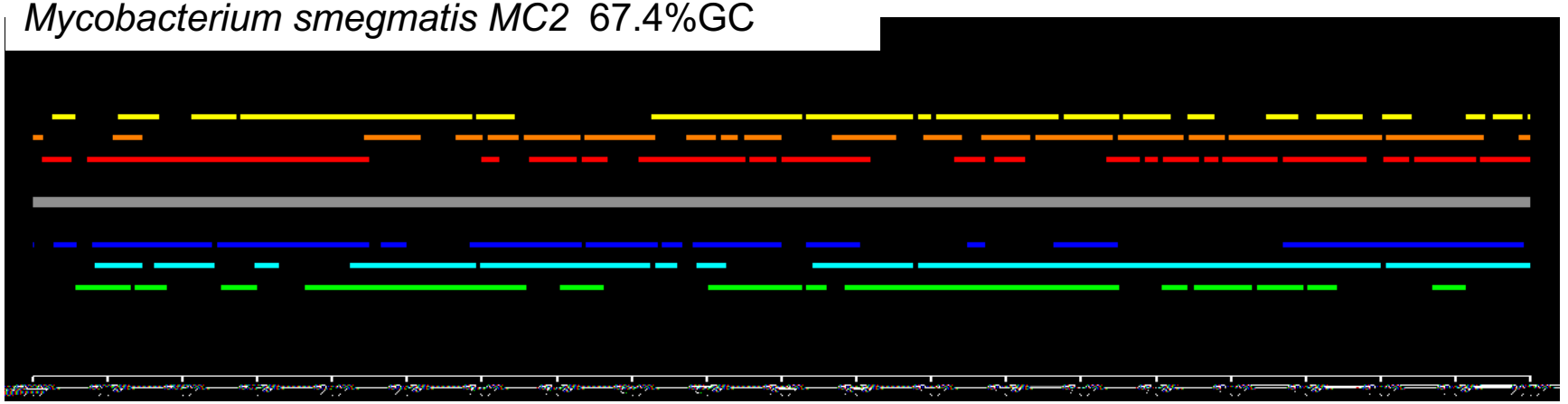
Campylobacter jejuni RM1221 30.3%GC



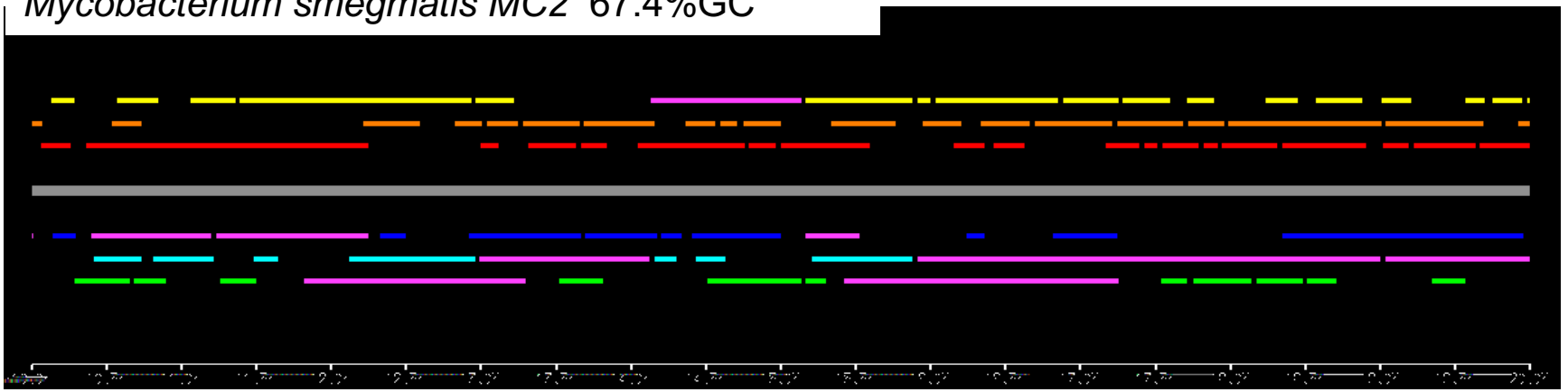


Note what happens in a high-GC genome

Mycobacterium smegmatis MC2 67.4%GC



Mycobacterium smegmatis MC2 67.4%GC



Flipping a Biased Coin

P(heads) = 61/64 (95.4%)   P(tails) = 3/64 (4.6%)

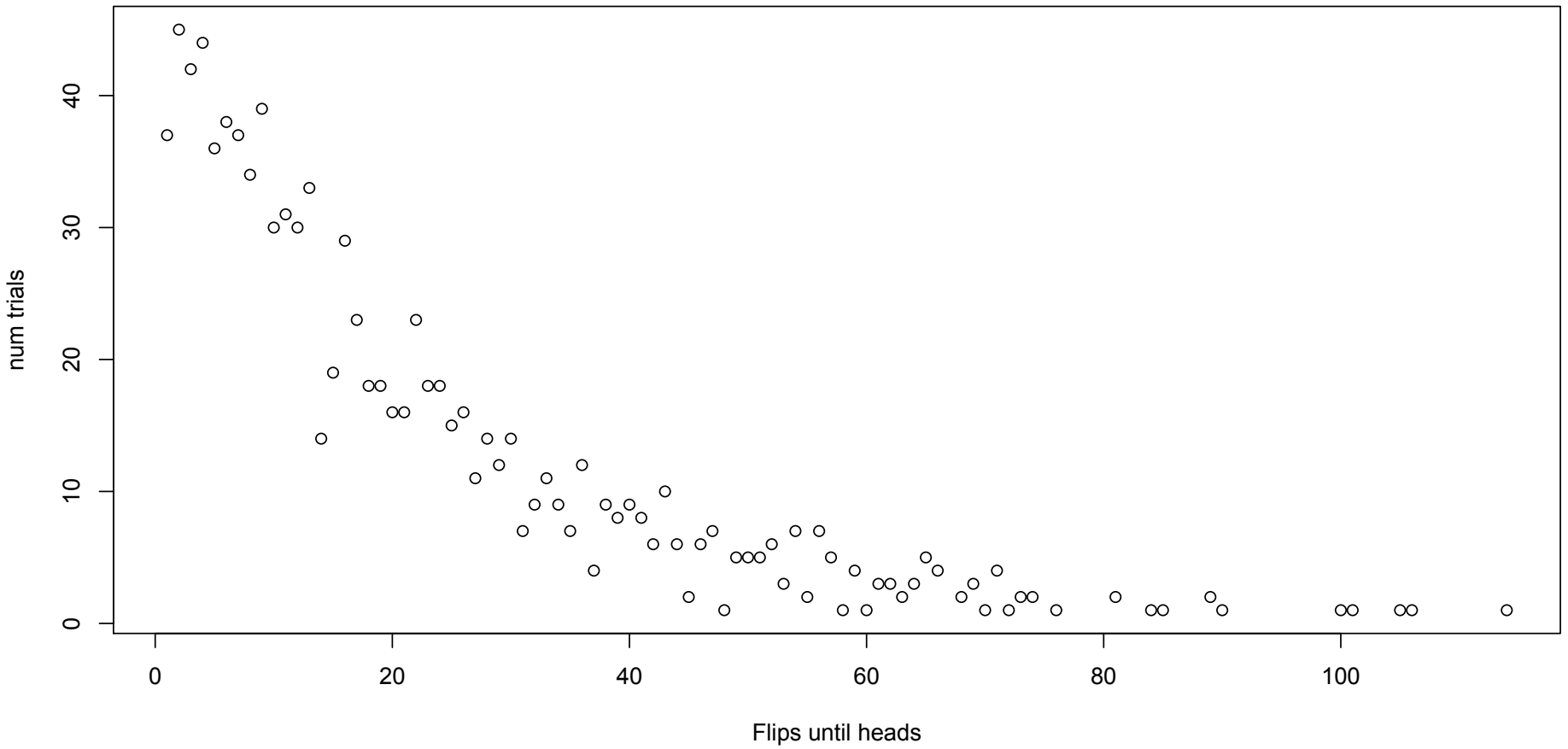
How many flips until my first tail?

```
$ ./coinflip.pl 0.046875 1000
0:   HHHHHHHHHHHHHHT   15
1:   HHHHHHT 7
2:   HHHHHHHHHHHT   12
3:   HHHHHHHHHHHHHHHHHHHHHHHT   24
4:   HT   2
5:   HHHHHHHHHHHHHT 14
6:   HHHHHHHHHT   10
7:   HHHHHHHHHHHHHT 14
8:   HHHHHT   6
9:   HHHHHHHHHHT   11
10:  HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHH
11:  HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHT   40
12:  HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHT   45
13:  HHHT   4
14:  HHHHHHHHHHHHHHT   15
15:  HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHT 39
16:  HHHHHT   6
17:  HHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHHT   38
18:  HHHHHHHHHHHHHHHHHHHHHHHHHT   26
19:  HHHHHHHHHHHT   12
```

Flipping a Biased Coin

P(heads) = 61/64 (95.4%) P(tails) = 3/64 (4.6%)

How many flips until my first tail?

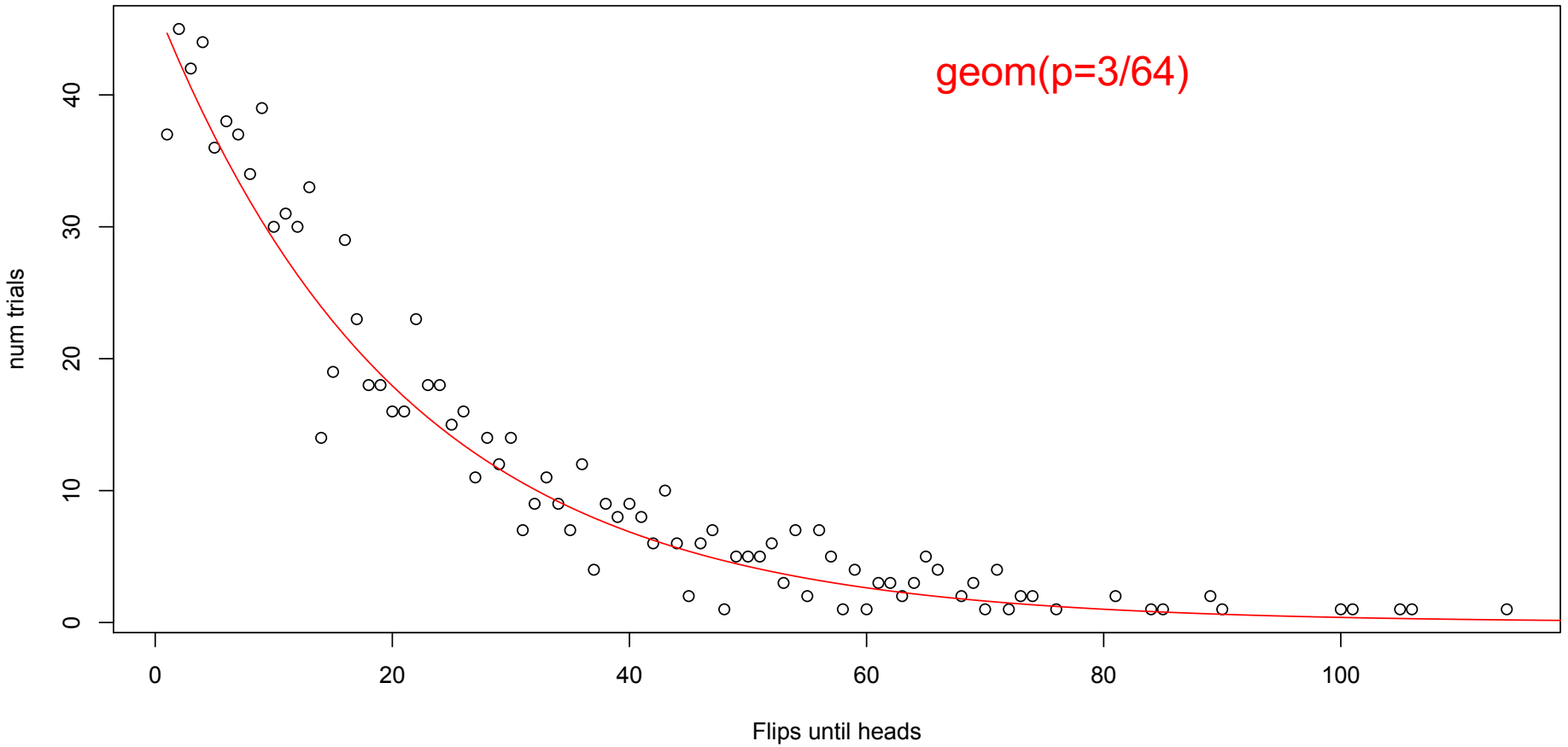


Flipping a Biased Coin

P(heads) = 61/64 (95.4%) P(tails) = 3/64 (4.6%)

How many flips until my first tail?

Geometric Distribution: $P(X=x) = p_{heads}^{x-1} p_{tails}$

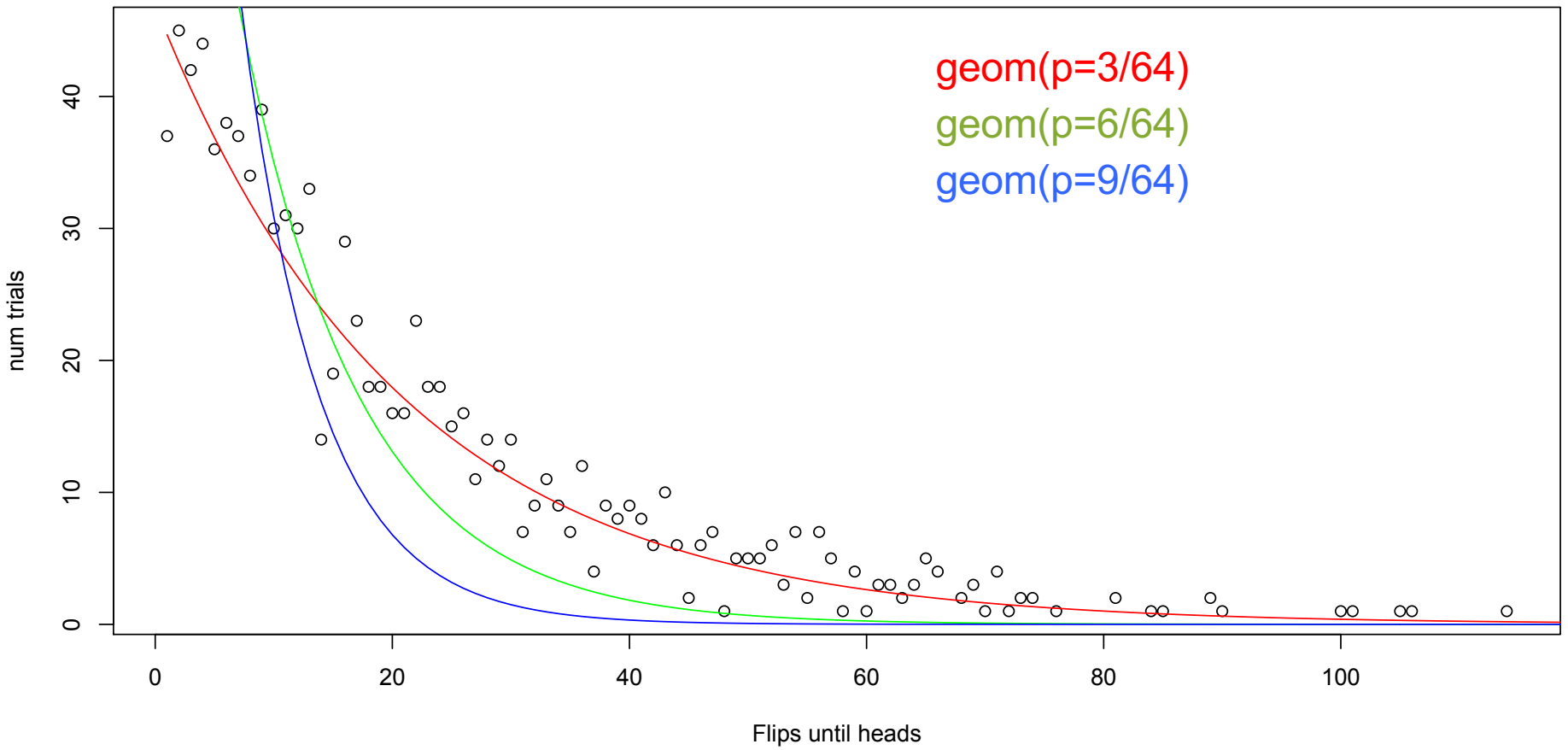


Flipping a Biased Coin

P(heads) = 61/64 (95.4%) P(tails) = 3/64 (4.6%)

How many flips until my first tail?

Geometric Distribution: $P(X=x) = p_{heads}^{x-1} p_{tails}$

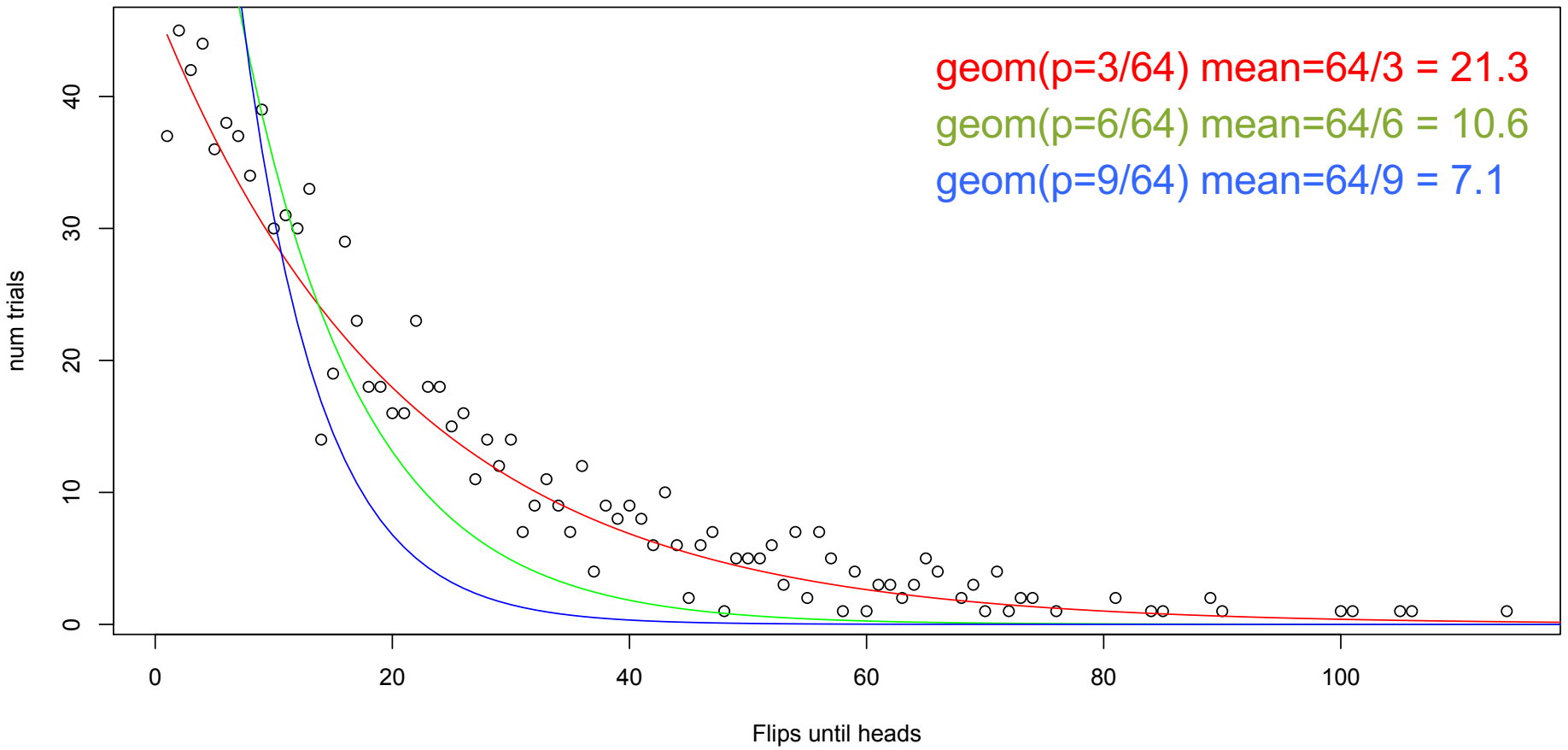


Flipping a Biased Coin

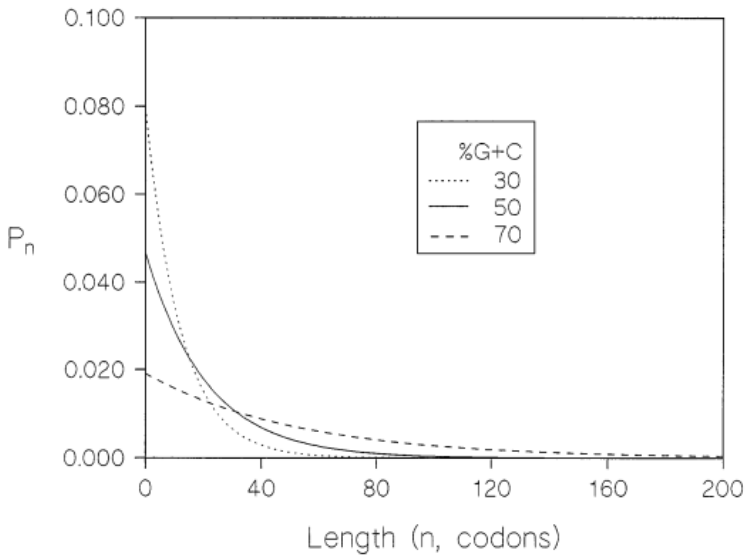
P(heads) = 61/64 (95.4%) P(tails) = 3/64 (4.6%)

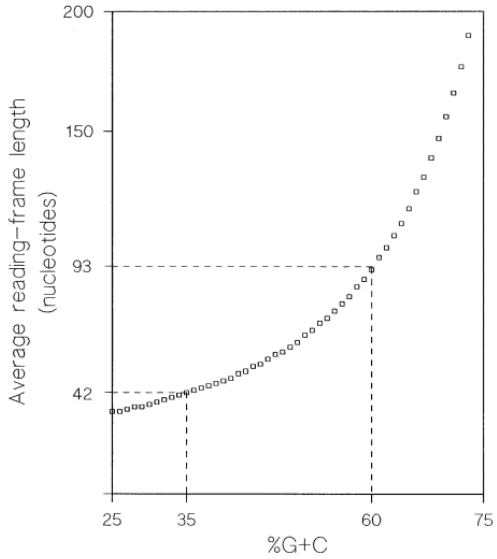
How many flips until my first tail?

Geometric Distribution: $P(X=x) = p_{heads}^{x-1} p_{tails}$



Stop Codon Frequencies



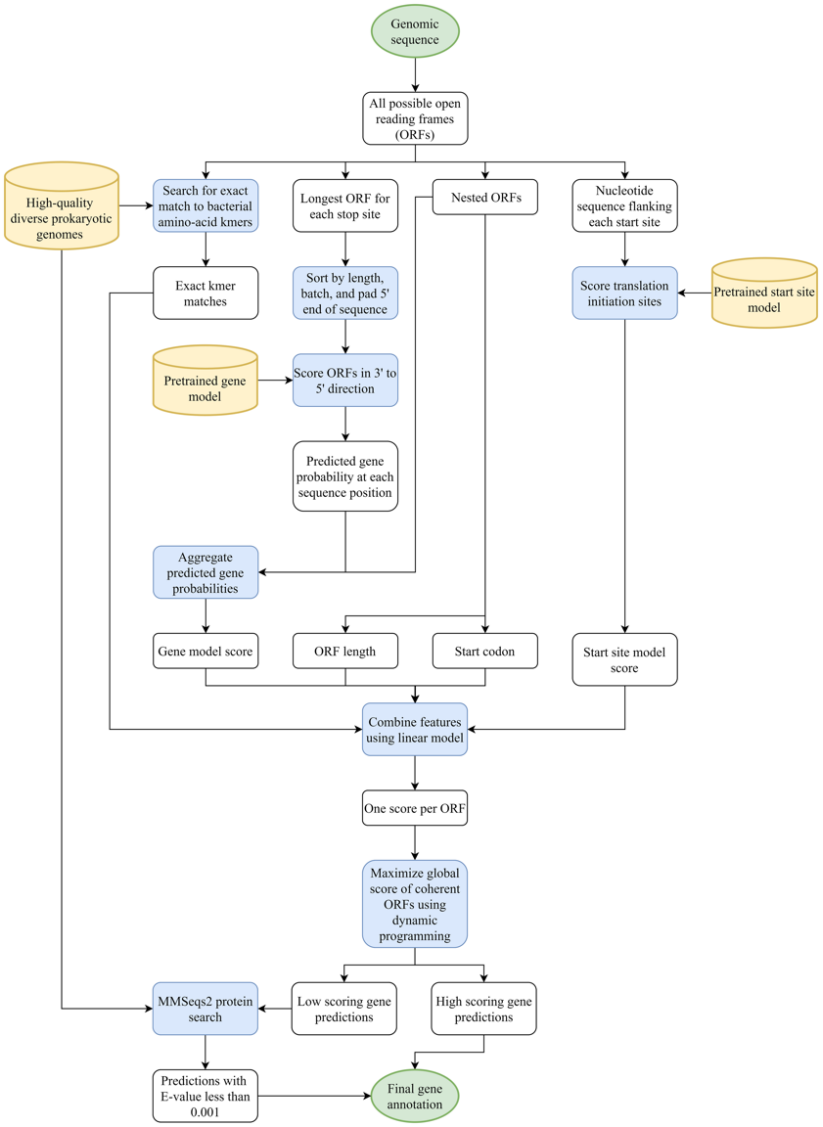


If the sequence is mostly A+T, then likely to form stop codons by chance!

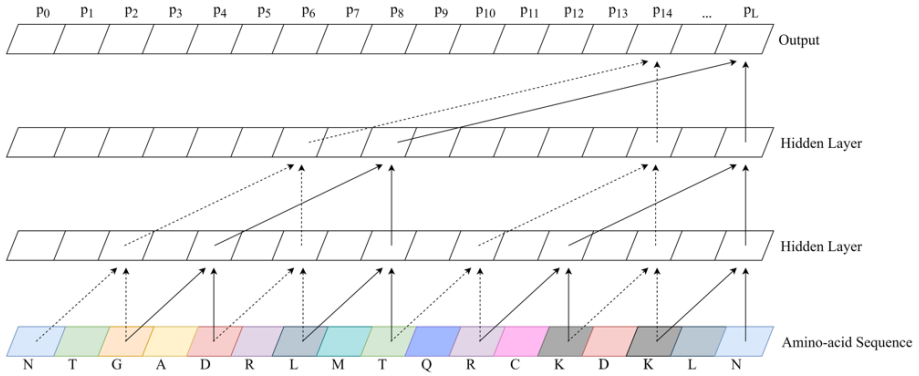
In High A+T (Low G+C):
Frequent stop codons; Short Random ORFs; long ORFs likely to be true genes

In High G+C (Low A+T):
Rare stop codons; Long Random ORFs; harder to identify true genes

A relationship between GC content and coding-sequence length.
Oliver & Marín (1996) J Mol Evol. 43(3):216-23.



Temporal Convolutional Network



Balrog: A universal protein model for prokaryotic gene prediction

Sommer, MJ, Salzberg, SL (2021) PLOS Comp. Bio. doi: 10.1371/journal.pcbi.1008727

Probabilistic Methods

- Create models that have a probability of generating any given sequence.
  - Evaluate gene/non-genome models against a sequence
- Train the models using examples of the types of sequences to generate.
  - Use RNA sequencing, homology, or "obvious" genes
- The "score" of an orf is the probability of the model generating it.
  - Most basic technique is to count how kmers occur in known genes versus intergenic sequences
  - More sophisticated methods consider variable length contexts, "wobble" bases, other statistical clues