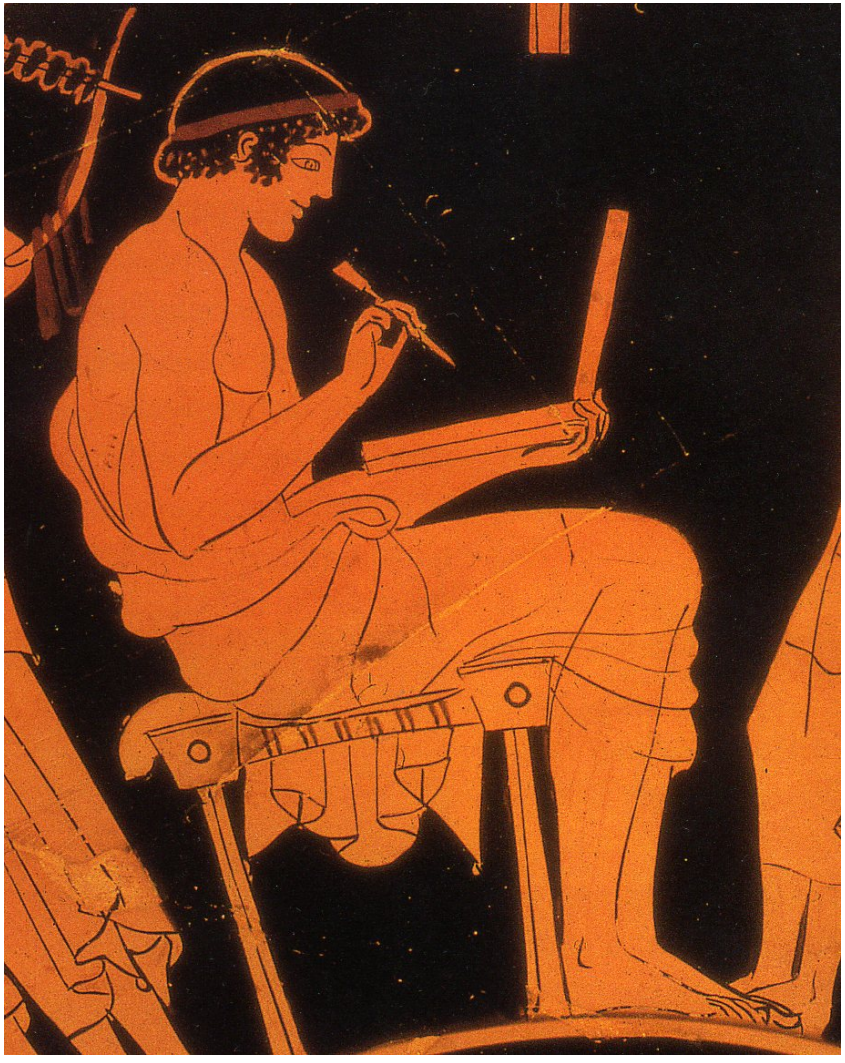
Homology and pairwise alignment

James B. Pease

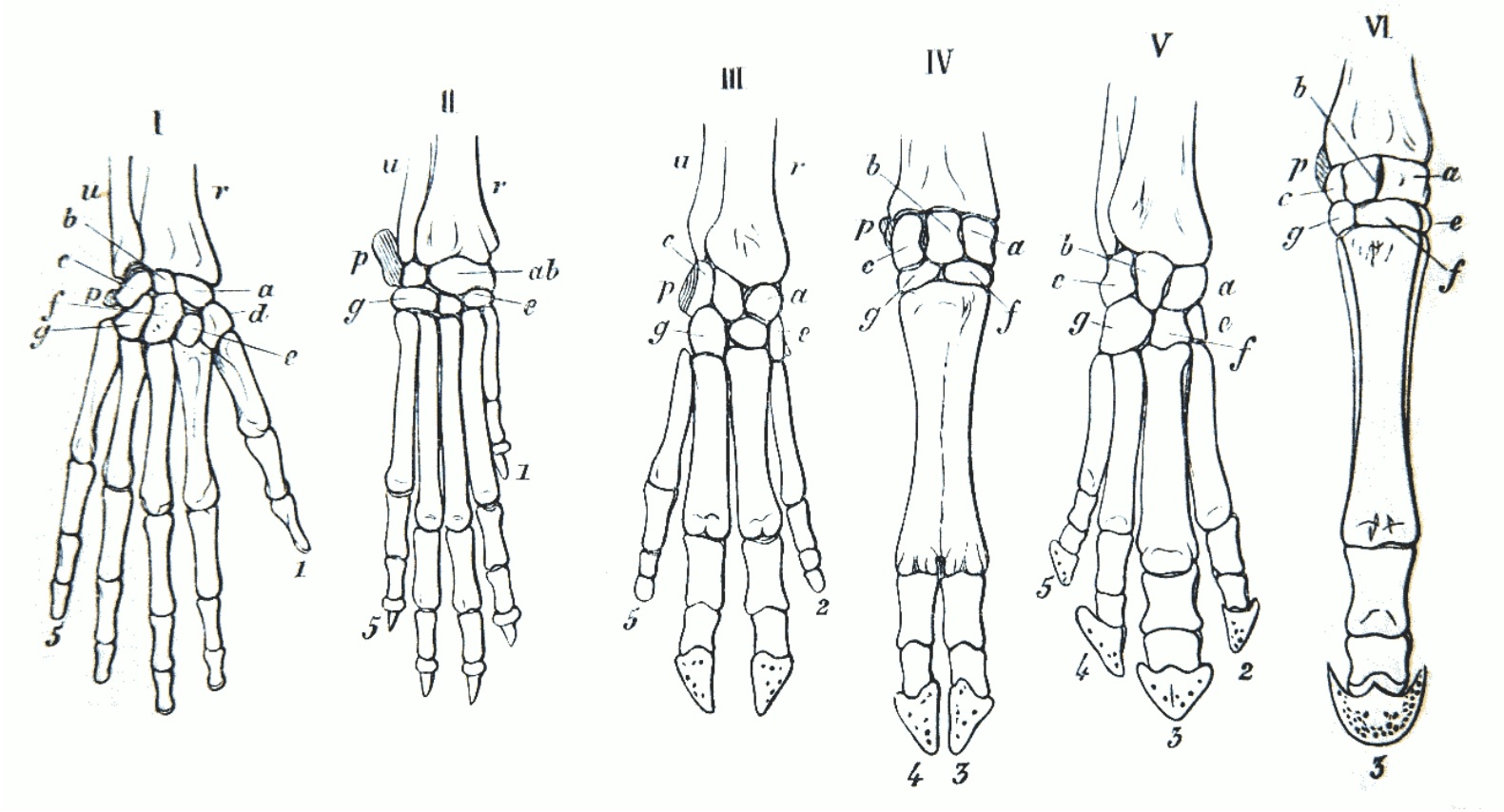
Wake Forest University



Ὁ βίος βραχύς,
ἡ δὲ τέχνη μακρή,
ὁ δὲ καιρὸς ὀξύς,
ἡ δὲ πεῖρα σφαλερή,
ἡ δὲ κρίσις χαλεπή.
— Ἱπποκράτης, Αφ. 1.1

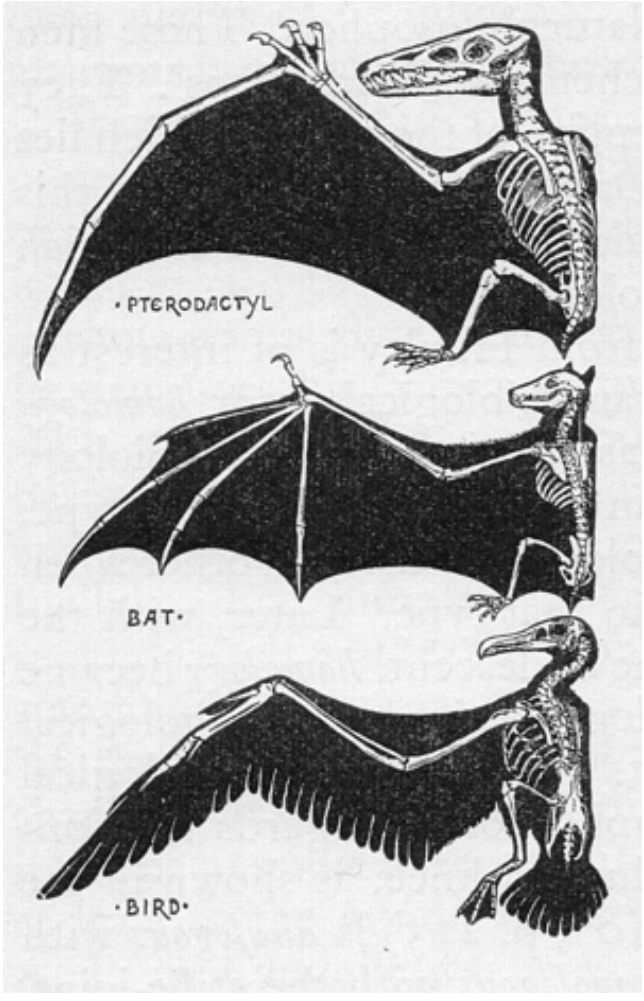
Life is short, and art long,
opportunity fleeting,
experience perilous,
and decision difficult.
— Hippokrates, Aph. 1.1

Homologous traits descend from a common ancestor



Gegenbaur 1870

Homologous traits descend from a common ancestor



Forelimbs are **homologous**
(all vertebrates have forelimbs because their ancestors did)

Wings are **analogous**
(wings evolved separately in birds, bats, flying reptiles)
(common function)

Romanes 1892

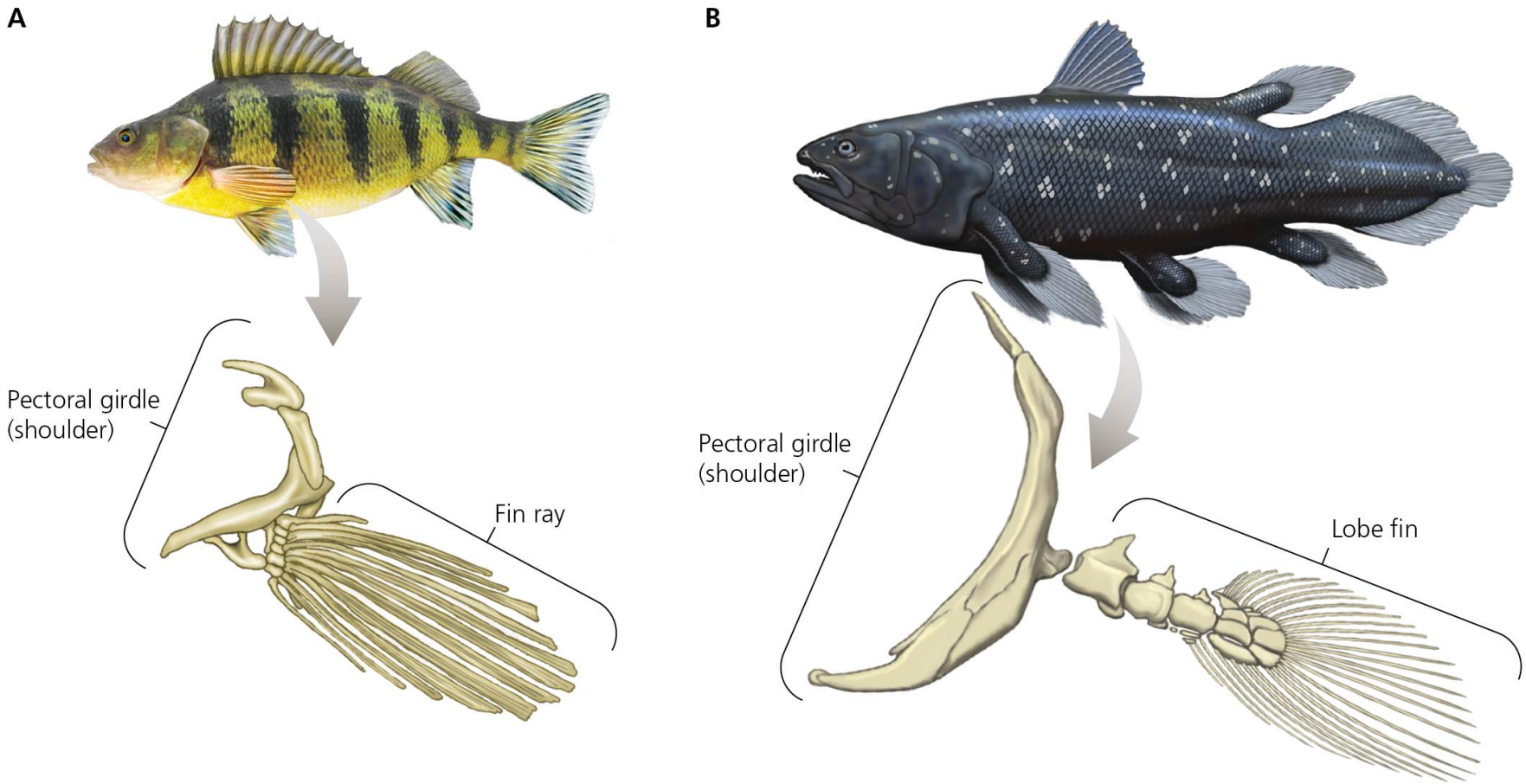
How can we determine homology?

Test hypotheses based on observations of:

- morphological similarity (*briefly* discuss)
- molecular sequence similarity (**mostly** discuss)

Verify experimentally

Coelacanth

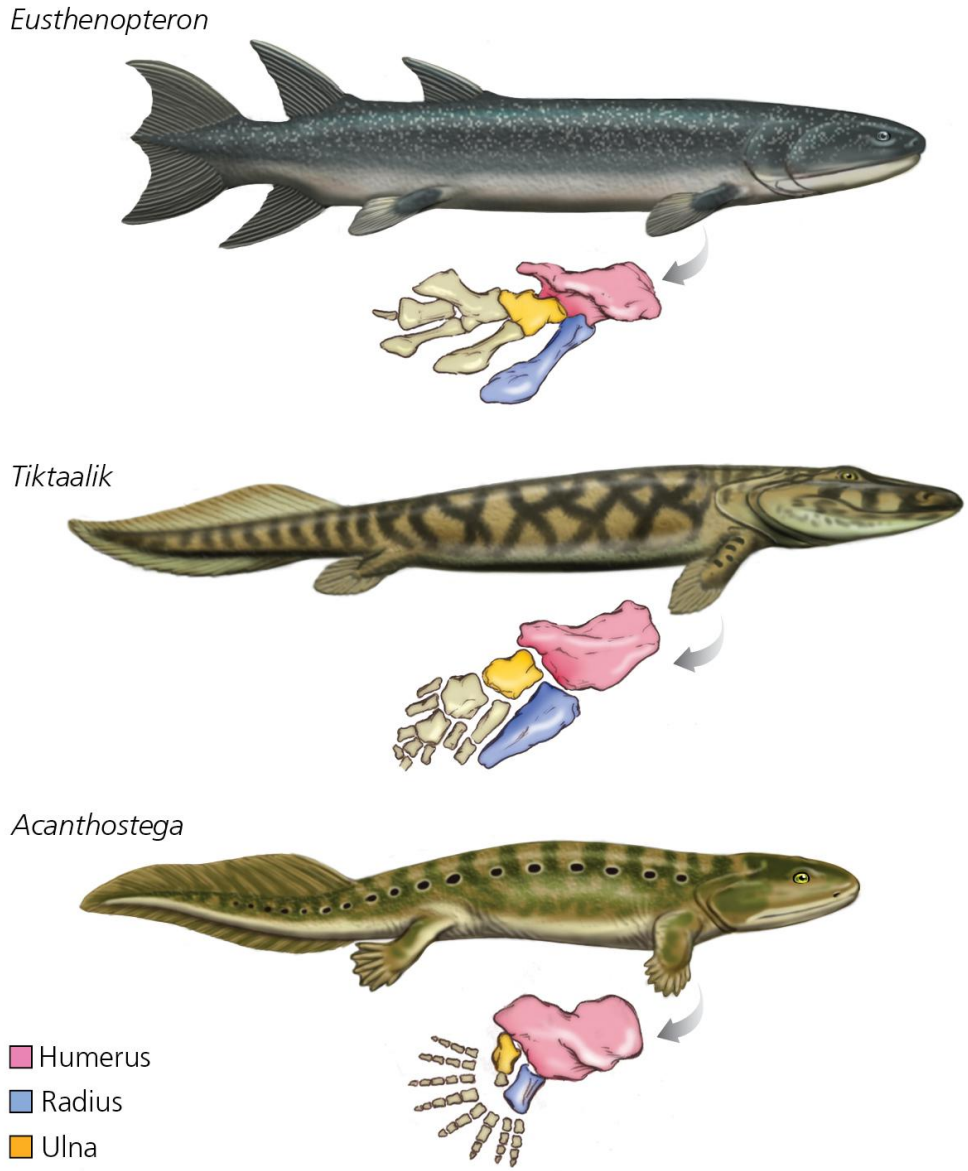

A
Pectoral girdle
(shoulder)
Fin ray
B
Pectoral girdle
(shoulder)
Lobe fin

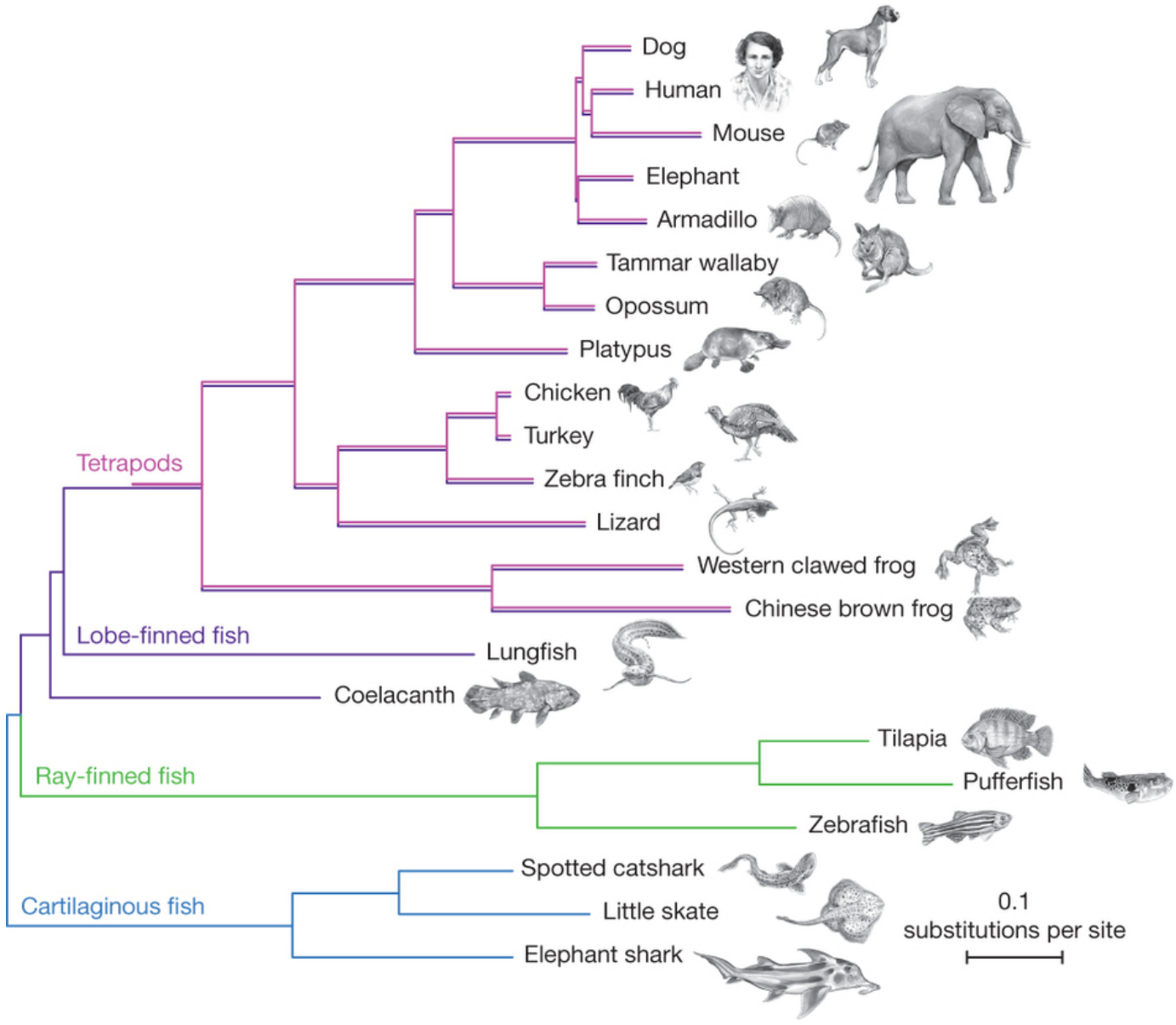
A 'Missing Link' Is Found

With the discovery of fossils of the Tiktaalik, or "large shallow water fish," scientists have found a missing connection between fishes and walking land creatures.

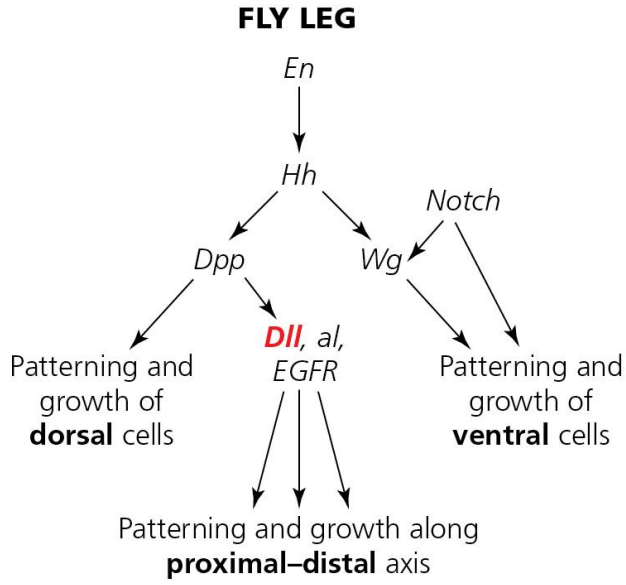
Sources: "Book of Life," edited by Stephen Jay Gould; Nature

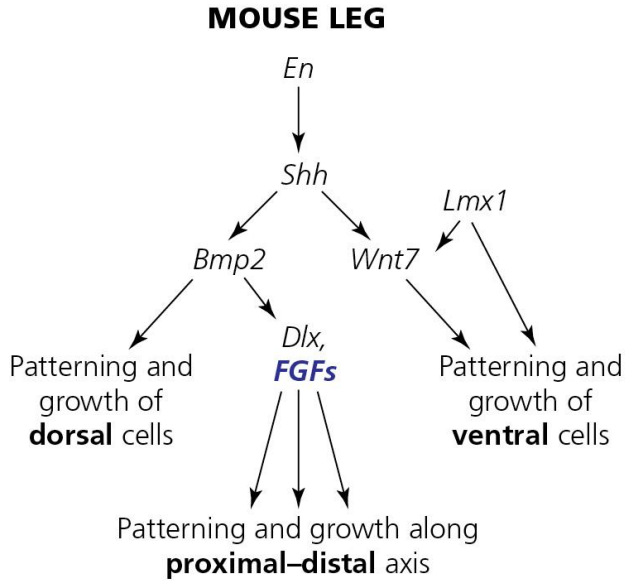
The New York Times; illustrations by Graham Roberts

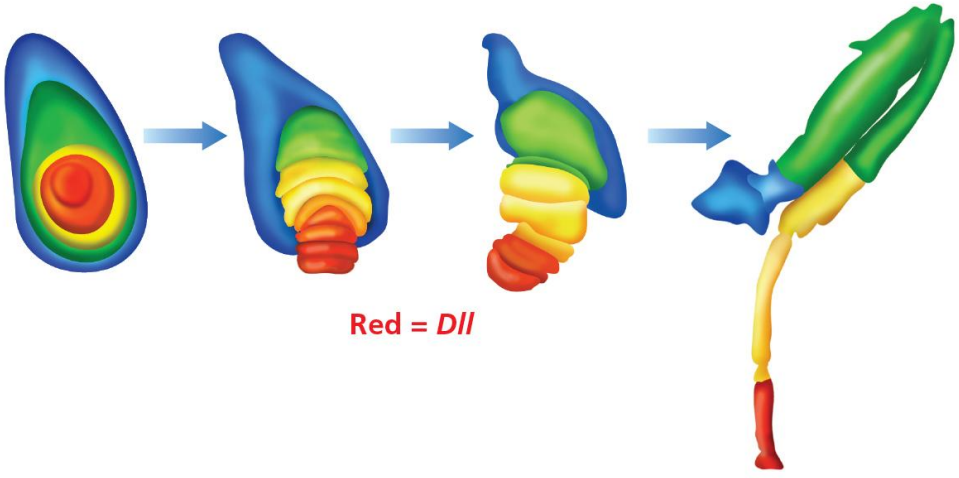

Eusthenopteron
Tiktaalik
Acanthostega
Humerus
Radius
Ulna

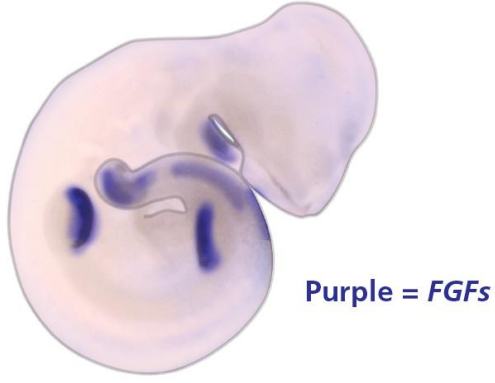

Dog
Human
Mouse
Elephant
Armadillo
Tammar wallaby
Opossum
Platypus
Chicken
Turkey
Zebra finch
Lizard
Western clawed frog
Chinese brown frog
Lungfish
Coelacanth
Tilapia
Pufferfish
Zebrafish
Spotted catshark
Little skate
Elephant shark
Tetrapods
Lobe-finned fish
Ray-finned fish
Cartilaginous fish
0.1
substitutions per site

Developmental and genetic pathway homology

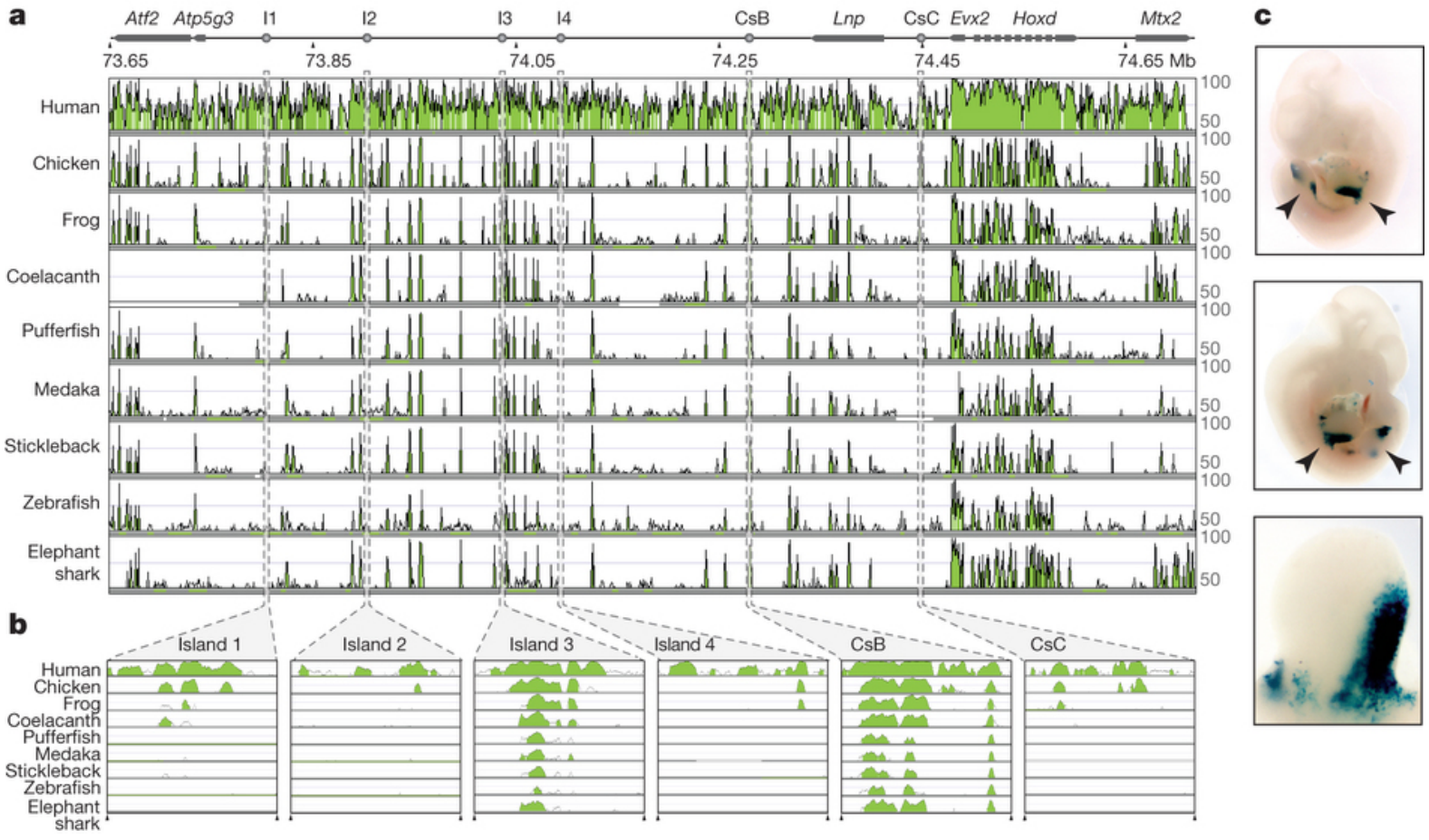




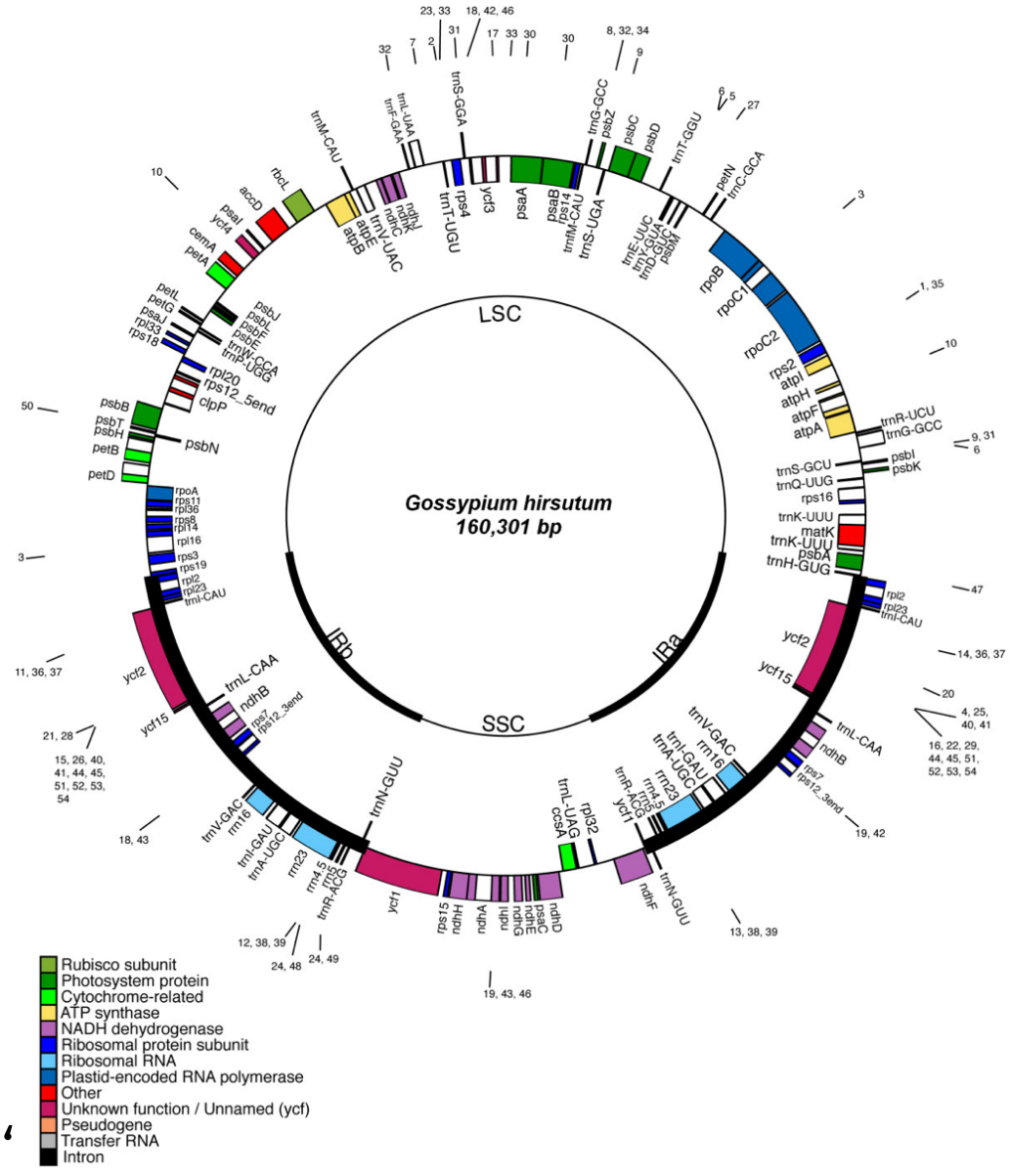




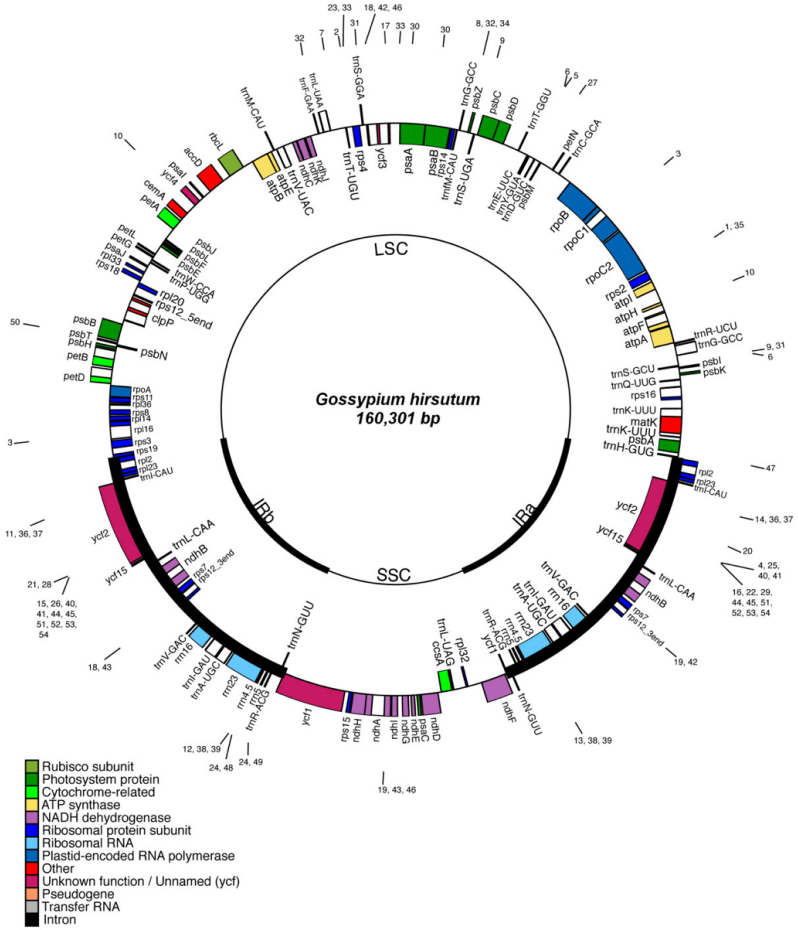
Developmental and genetic pathway homology

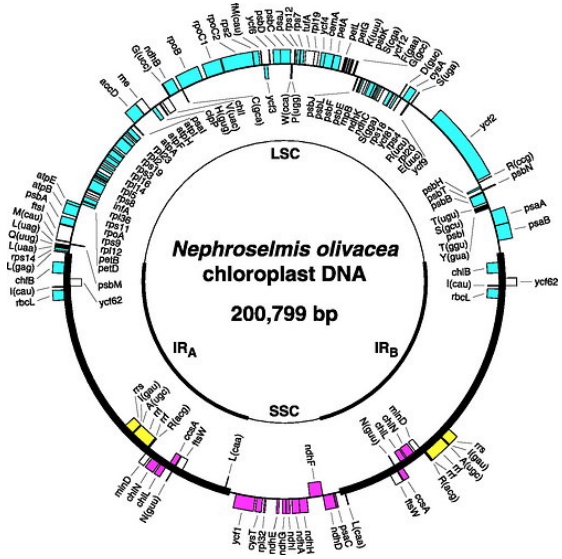


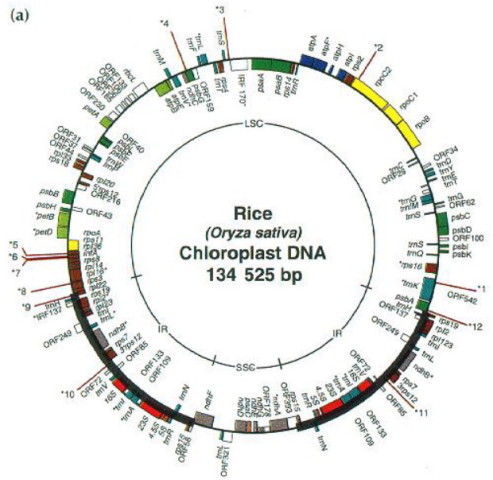
Sequence homlogy



Sequence homlogy







Common Molecular Markers for Homology

rbcL (RuBisCO, large subunit)

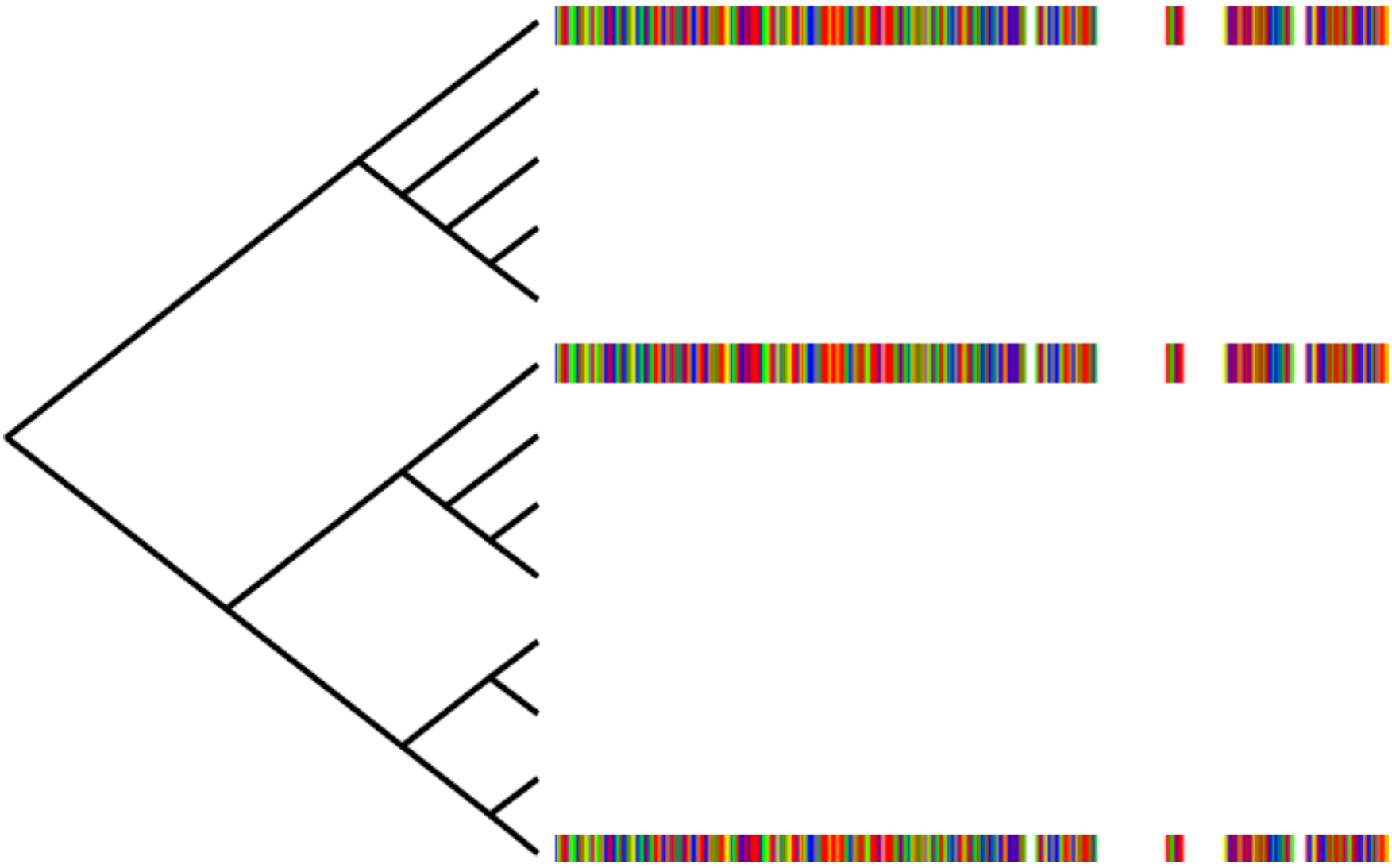
ITS (internal transcribed spacer)

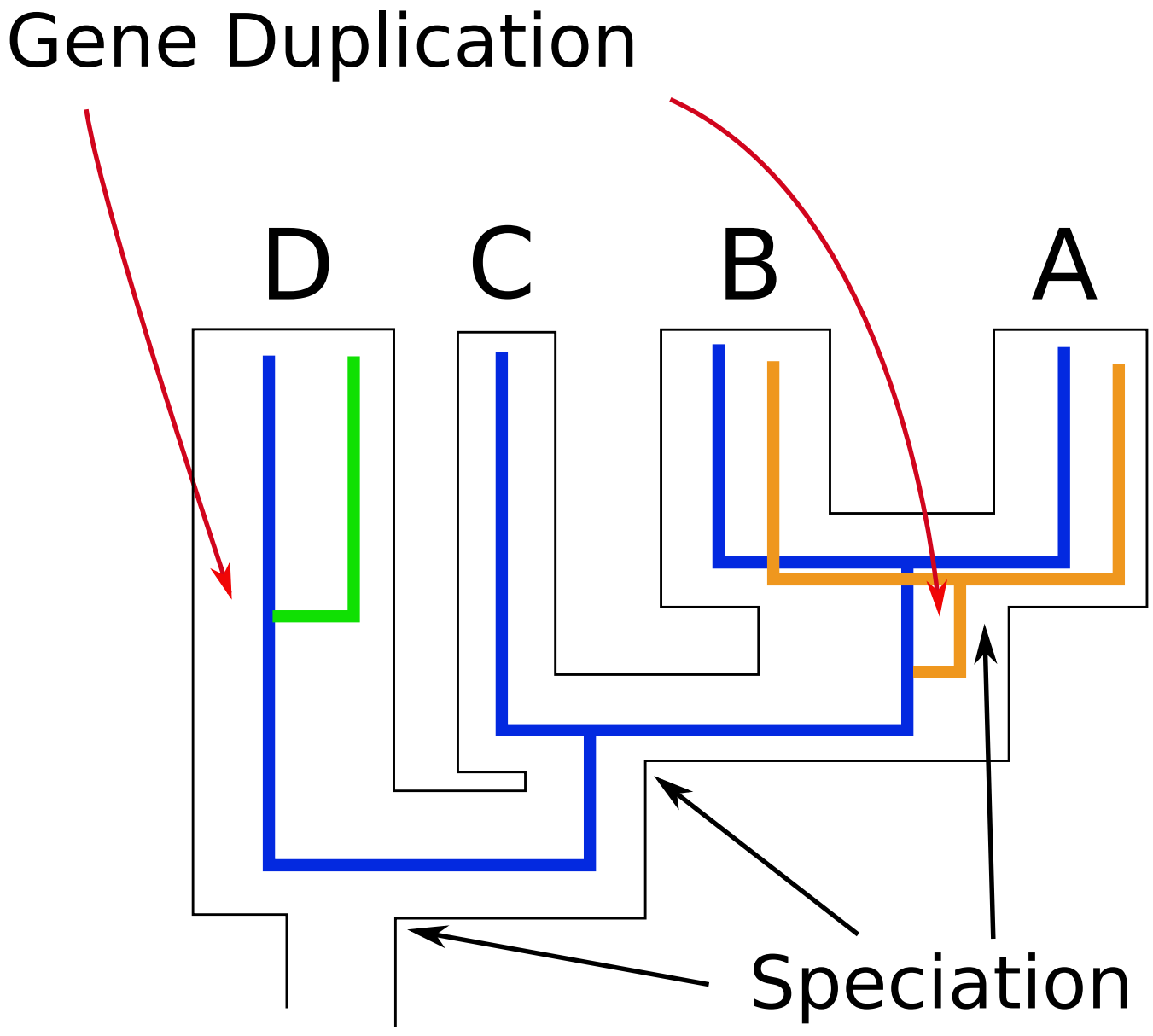
Cox1 (cytochrome C oxidase, subunit 1)

Microsatellite repeats

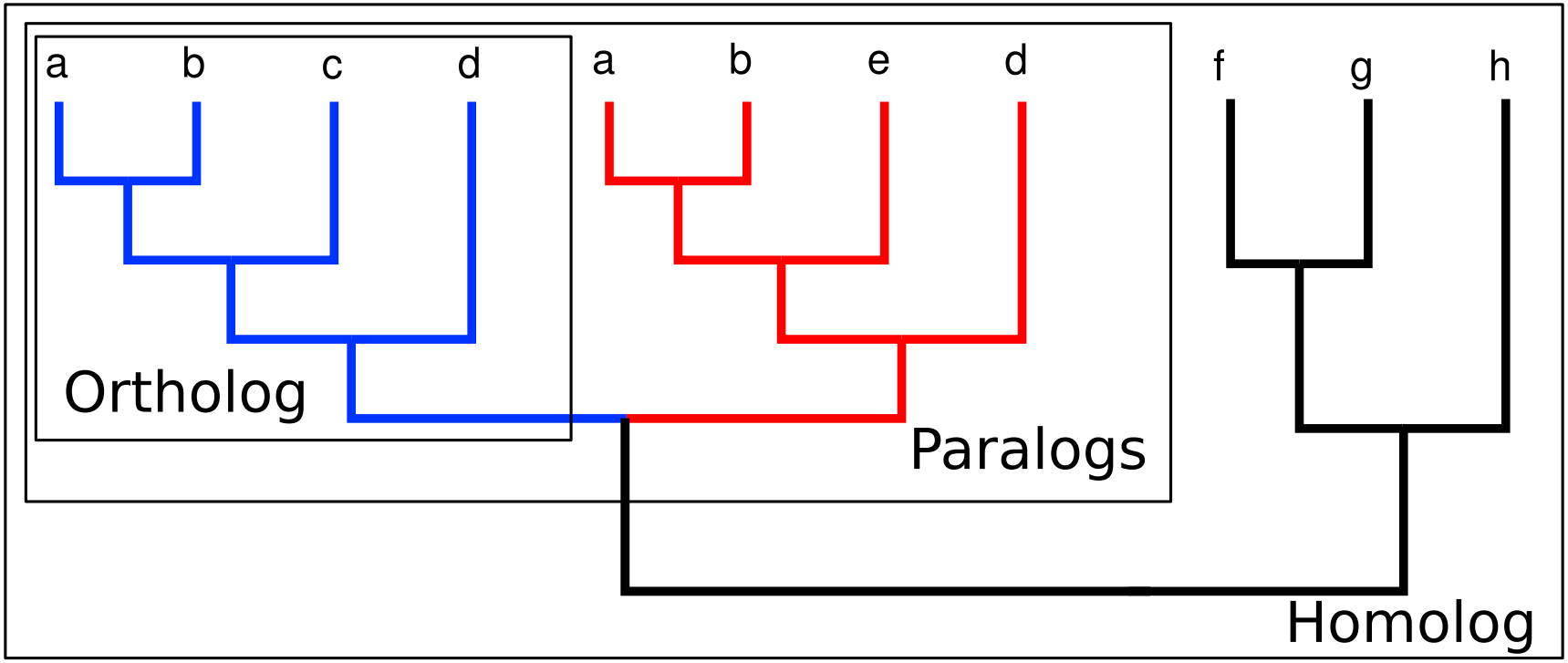
16S/18S ribosomal RNA

Molecular “states” (nucleotides or amino acids) are the characters on the tree




Gene Duplication
D
C
B
A
Speciation

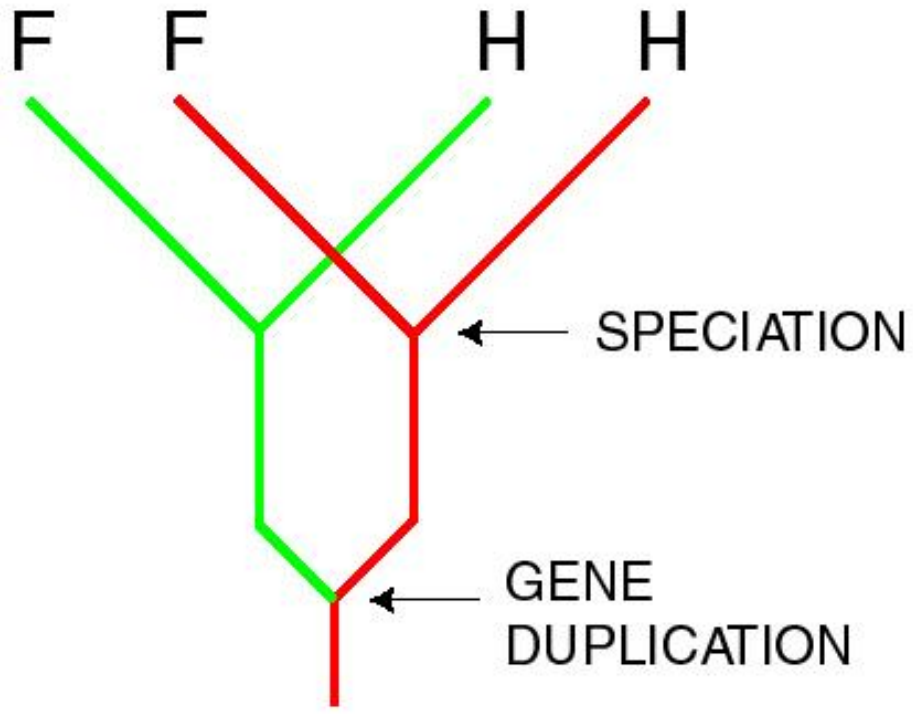
Homology, Orthology, and Paralogy

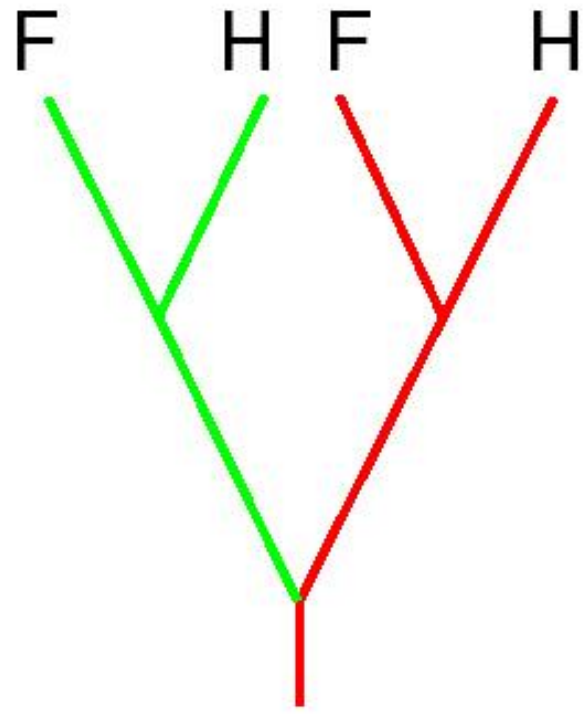


Homologs: descended from a common ancestor

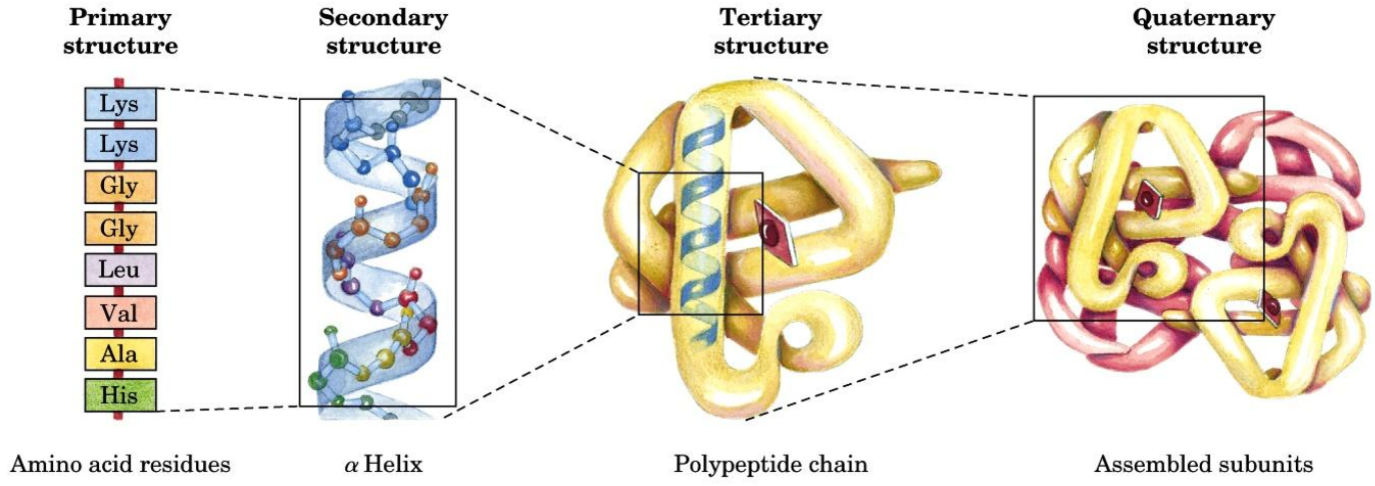
Paralogs: descended from a common ancestor and split by a gene duplication event

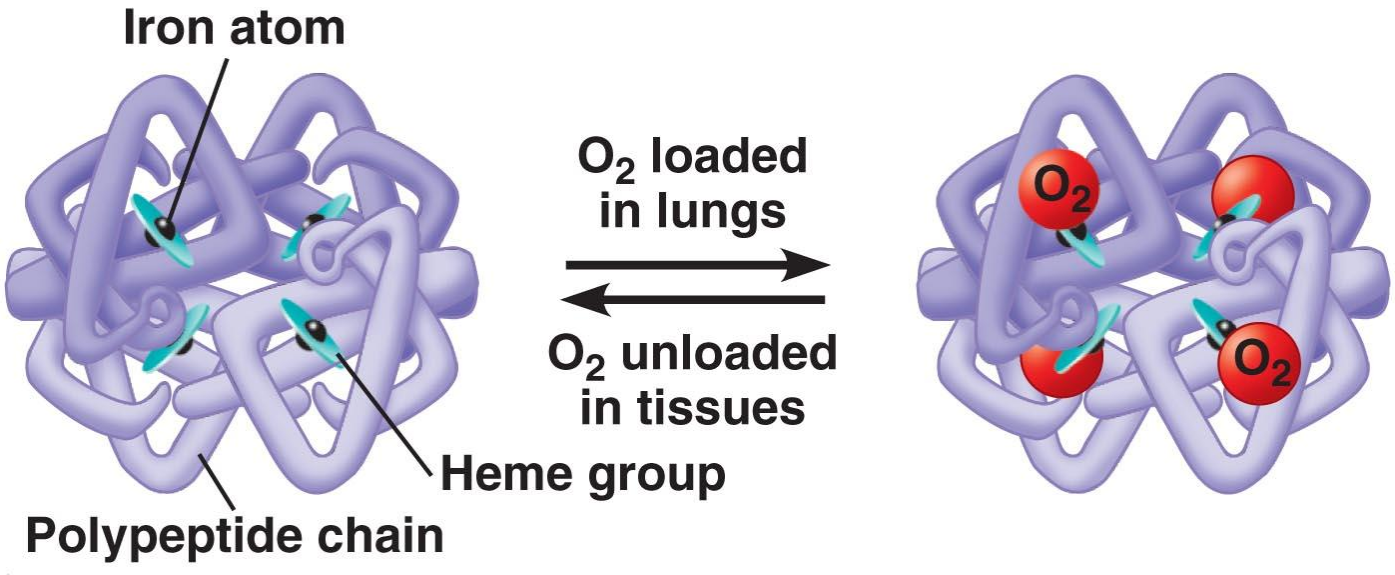
Orthologs: descended from a common ancestor and split by a speciation event

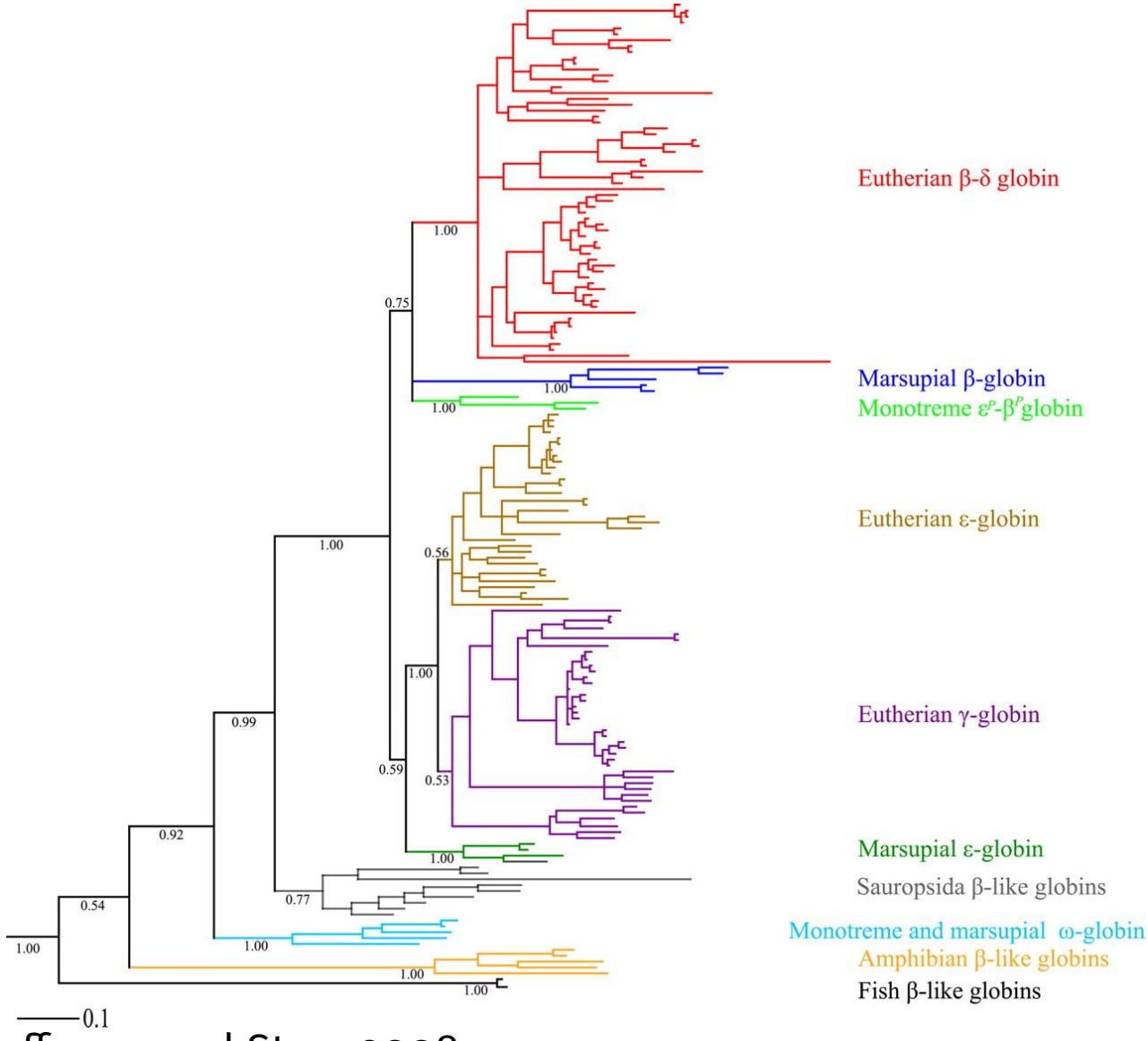




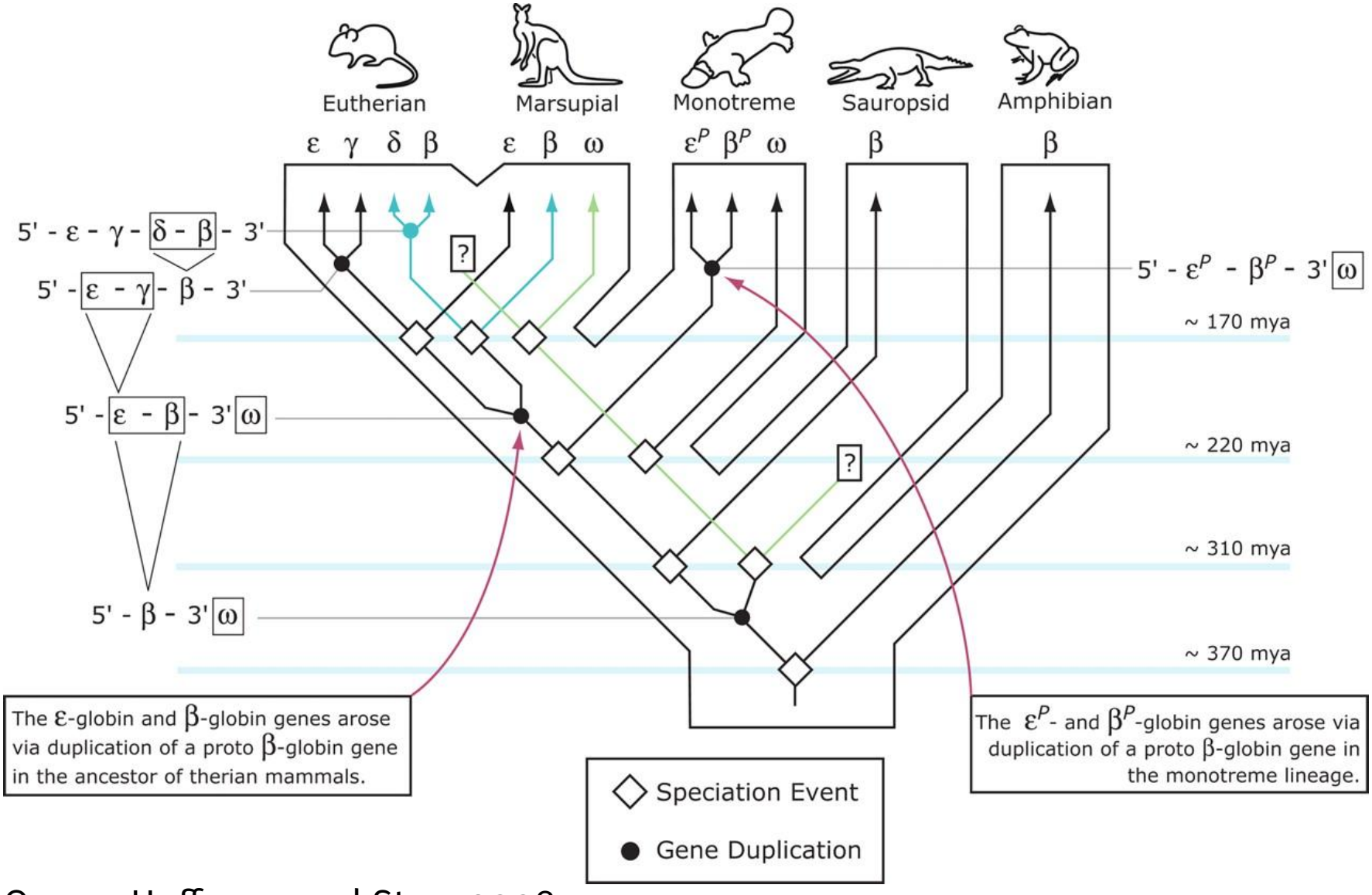
Hemoglobin



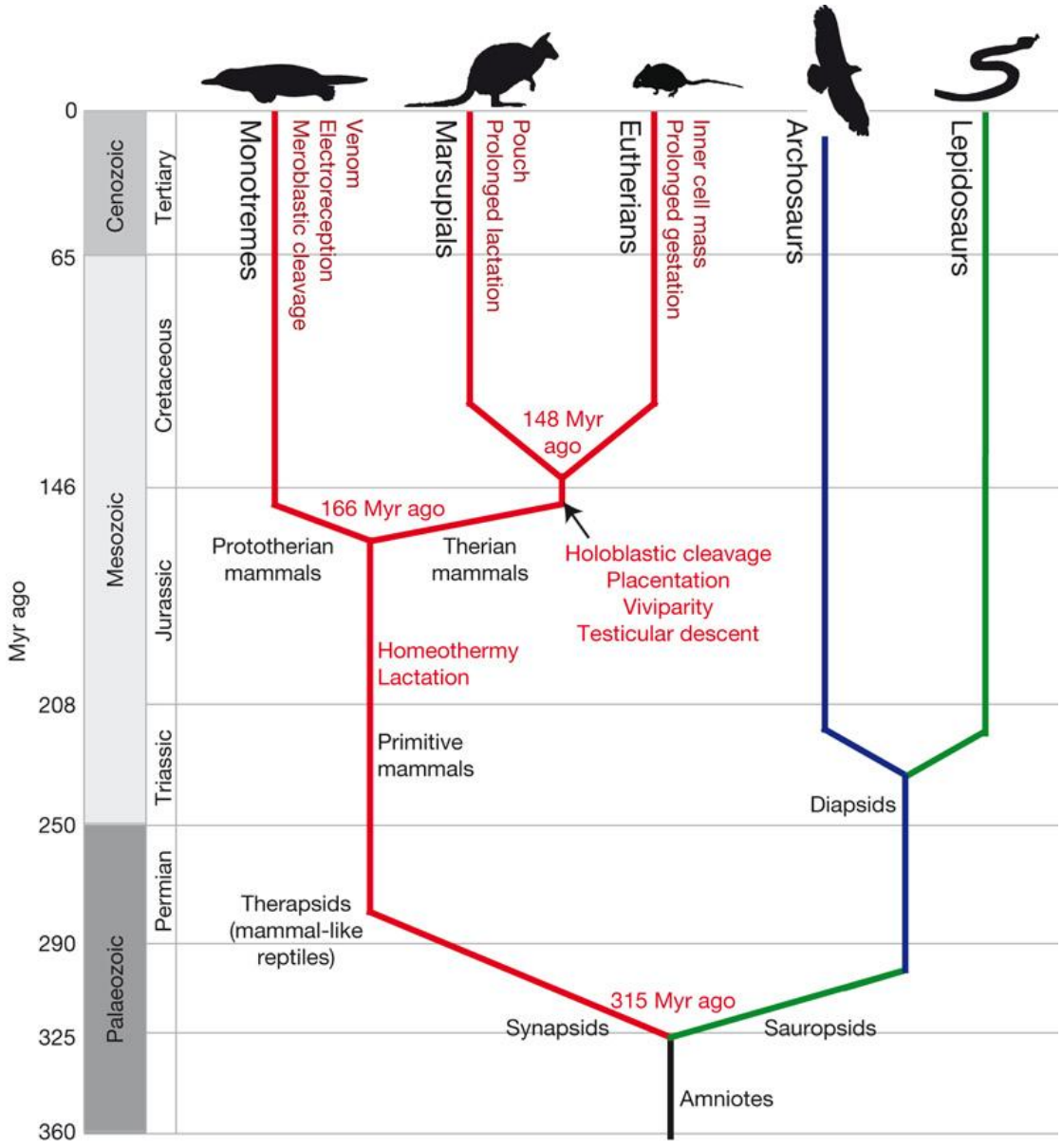


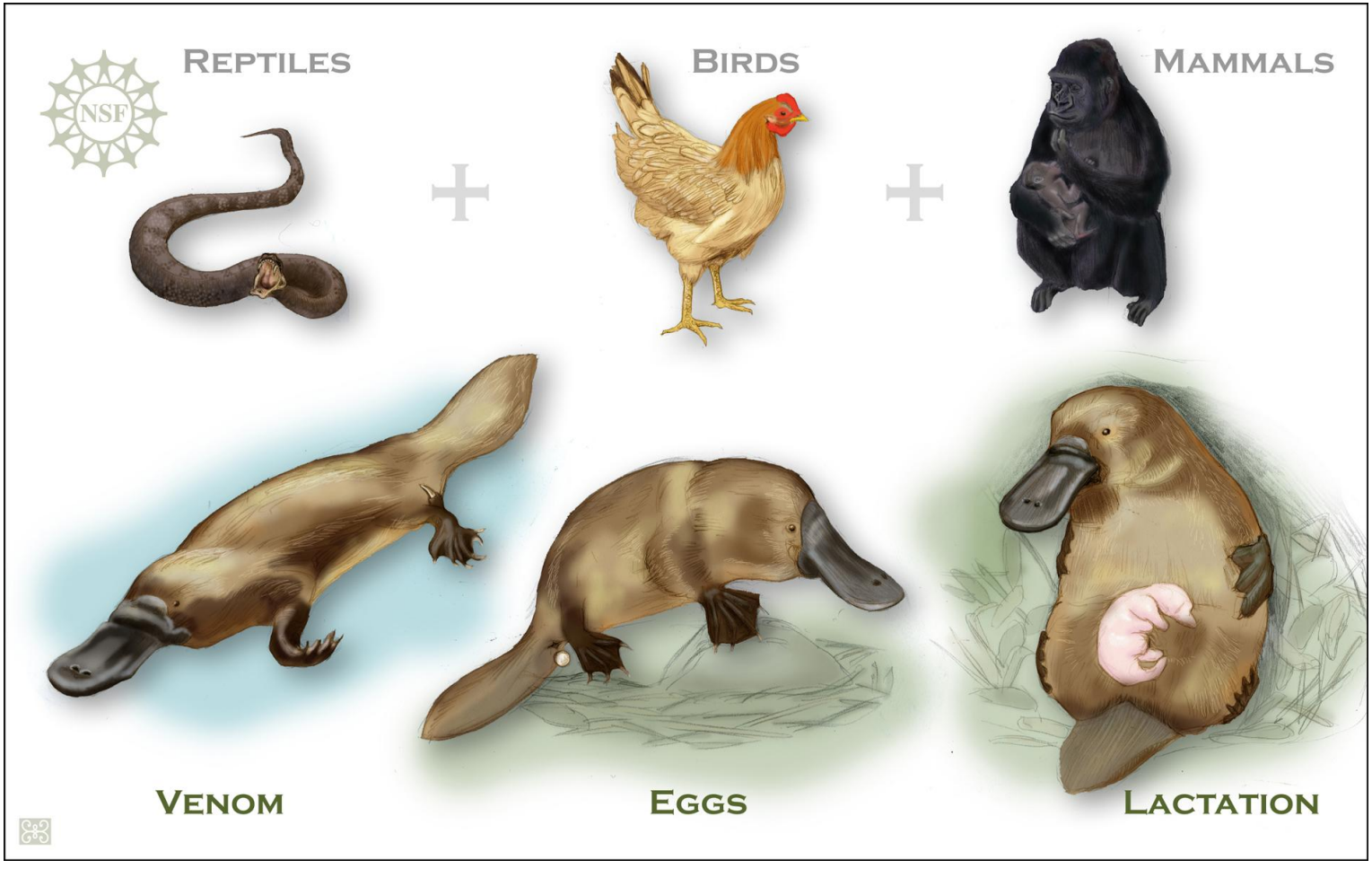


Opazo, Hoffman, and Storz 2008

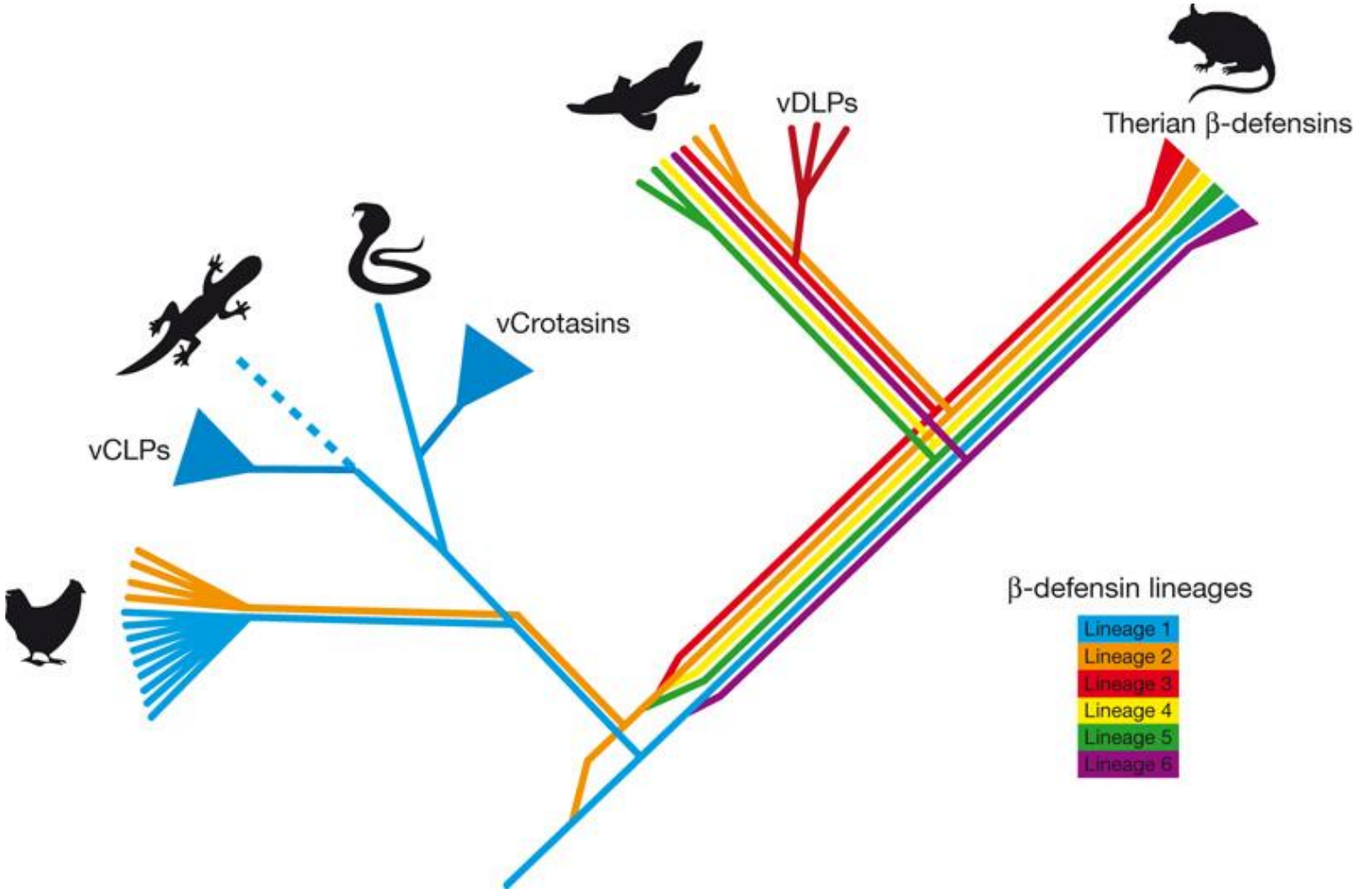


Opazo, Hoffman, and Storz 2008

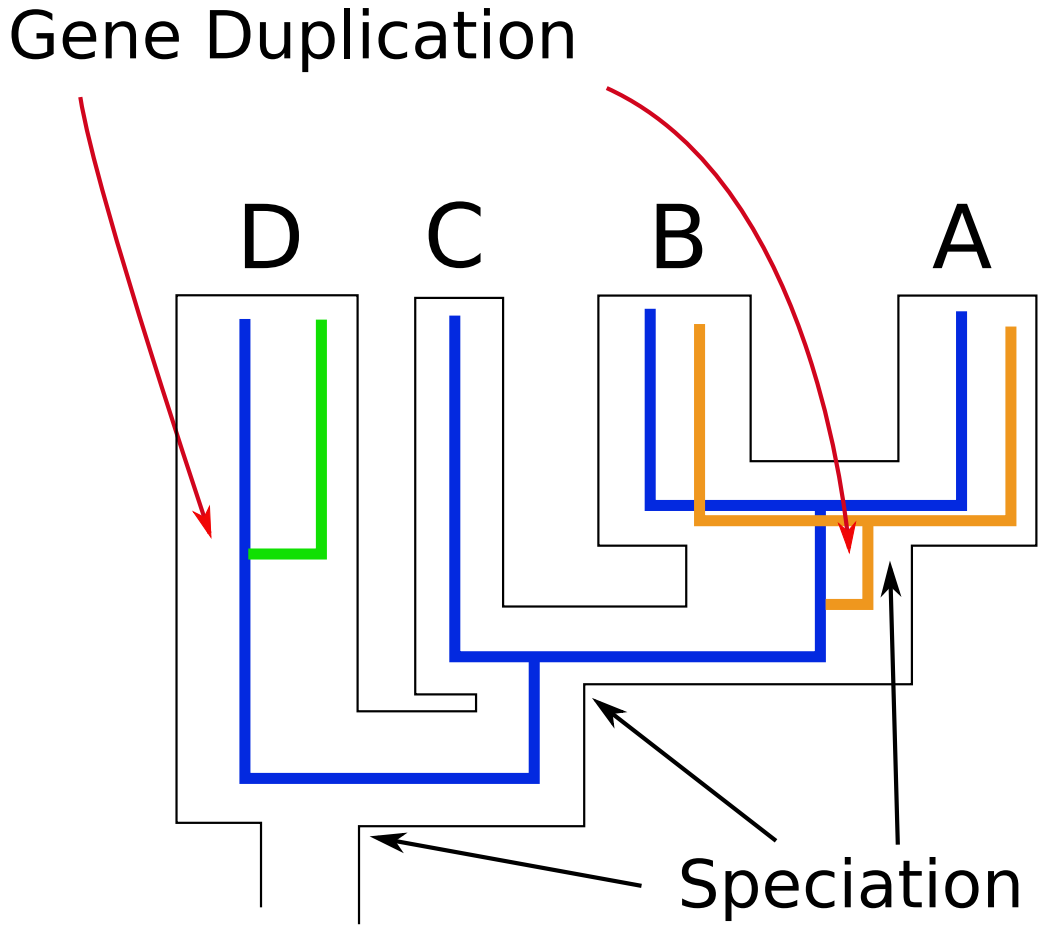

Myr ago
0
65
146
208
250
290
325
360
Cenozoic
Mesozoic
Palaeozoic
Tertiary
Cretaceous
Jurassic
Triassic
Permian
Monotremes
Marsupials
Eutherians
Archosaurs
Lepidosaurs
Venom
Electroreception
Meroblastic cleavage
Pouch
Prolonged lactation
Inner cell mass
Prolonged gestation
148 Myr ago
166 Myr ago
Prototherian mammals
Therian mammals
Holoblastic cleavage
Placentation
Viviparity
Testicular descent
Homeothermy
Lactation
Primitive mammals
Diapsids
Therapsids (mammal-like reptiles)
315 Myr ago
Synapsids
Sauropsids
Amniotes


NSF
Reptiles
Birds
Mammals
Venom
Eggs
Lactation

Platypus venom proteins are transcribed in part from duplicated beta defensin genes (vDLPs, venom defensin like proteins)



Why do we need homologs?



Why do we need homologs?

In order to **analyze changes in molecular sequences** among species/populations/individuals, we need to find **comparable sequences**.

Homology is central to:

multiple sequence alignment

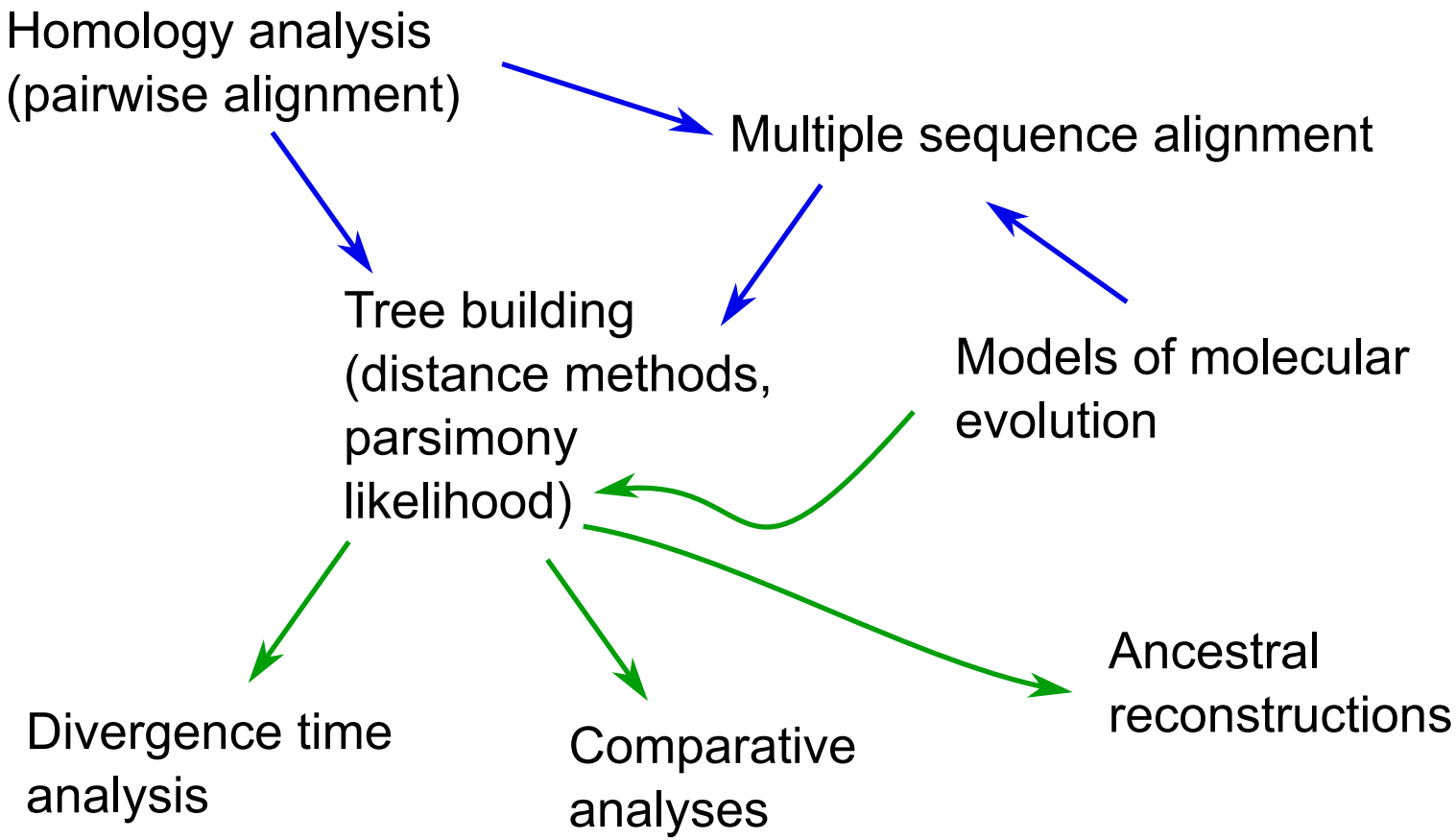
- each column is assumed to be a homologous site

phylogenetic trees

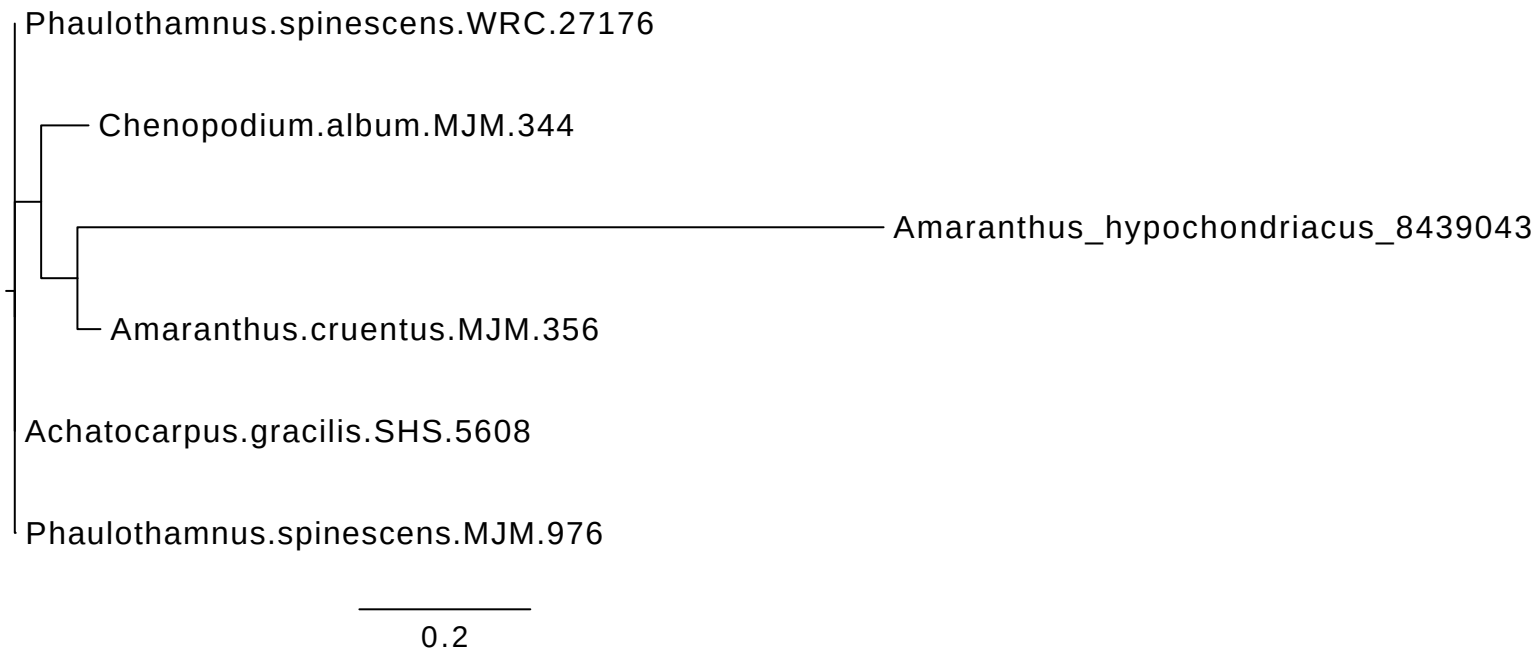
- A tree models ancestry relationships, so sequences need to be homologous

short read alignment
clustering genes

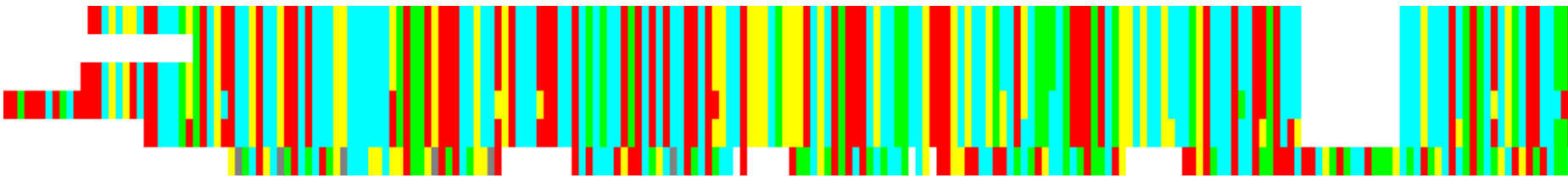
Overview


Homology analysis
(pairwise alignment)
Multiple sequence alignment
Tree building
(distance methods,
parsimony
likelihood)
Models of molecular
evolution
Divergence time
analysis
Comparative
analyses
Ancestral
reconstructions

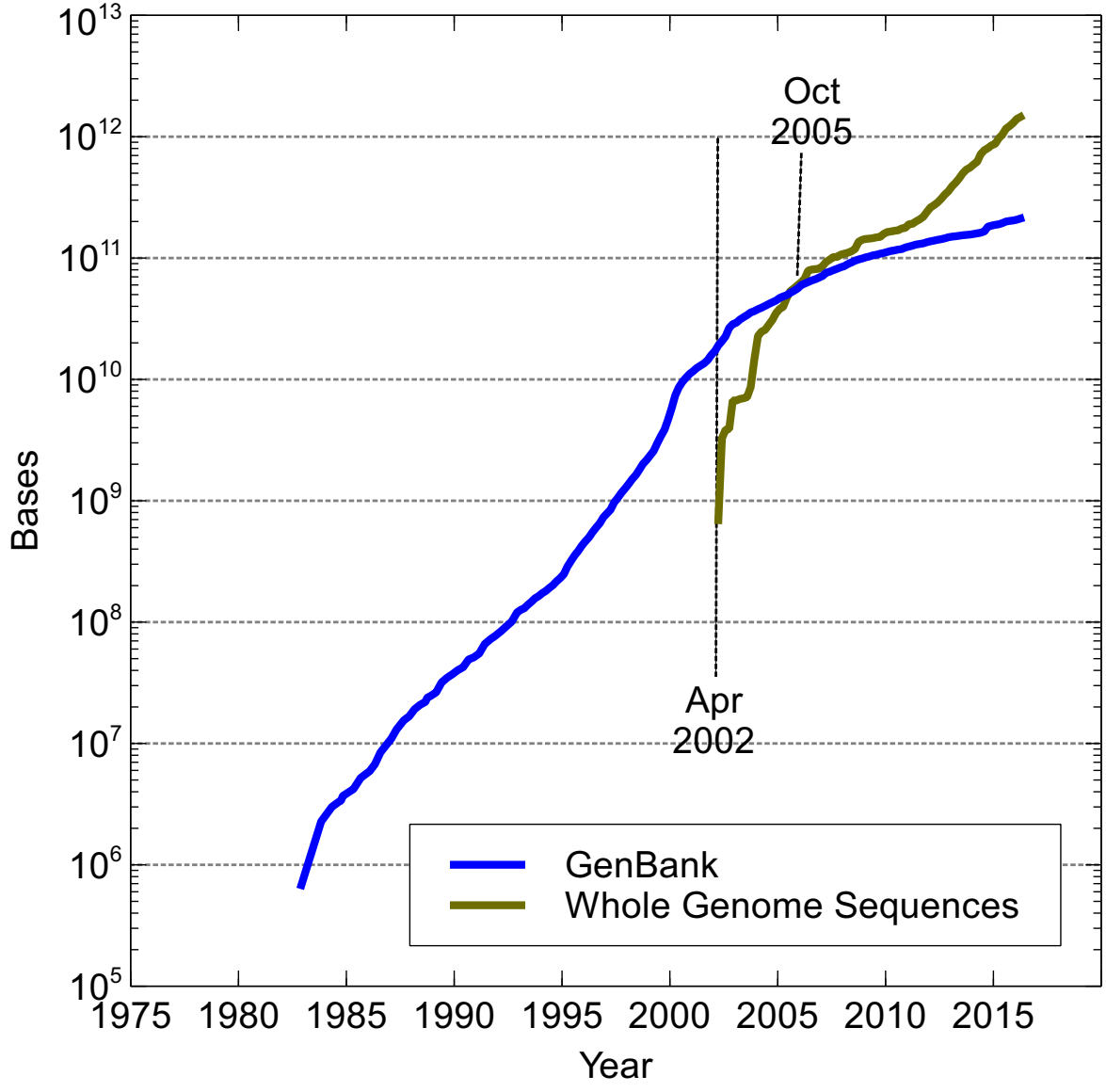
Tree


Phaulothamnus.spinescens.WRC.27176
Chenopodium.album.MJM.344
Amaranthus_hypochondriacus_8439043
Amaranthus.cruentus.MJM.356
Achatocarpus.gracilis.SHS.5608
Phaulothamnus.spinescens.MJM.976
0.2

Alignment



NCBI

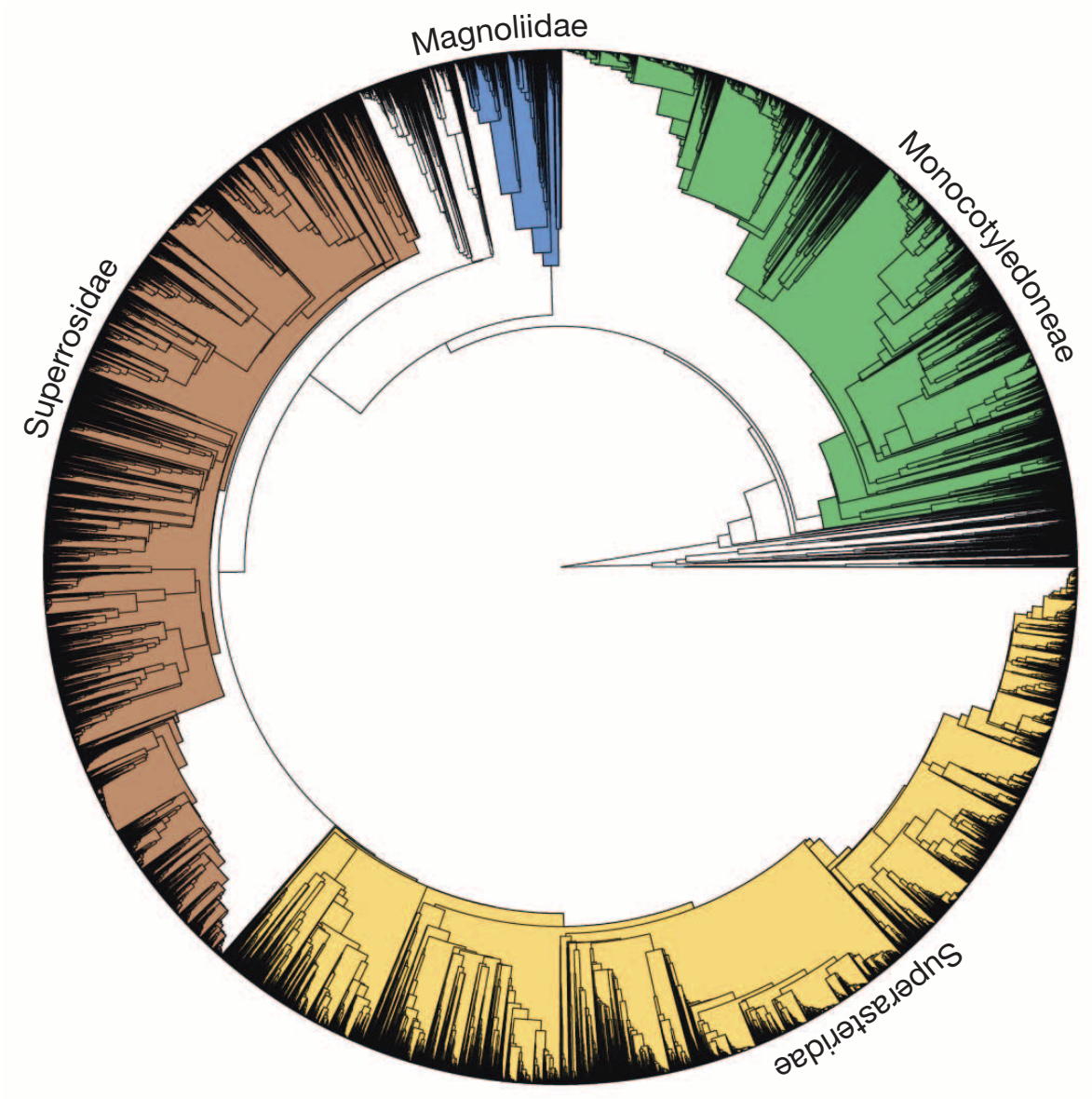

Oct
2005
Apr
2002
GenBank
Whole Genome Sequences
Bases
Year
10^{13}
10^{12}
10^{11}
10^{10}
10^{9}
10^{8}
10^{7}
10^{6}
10^{5}
1975
1980
1985
1990
1995
2000
2005
2010
2015

Can we get homology from a database?

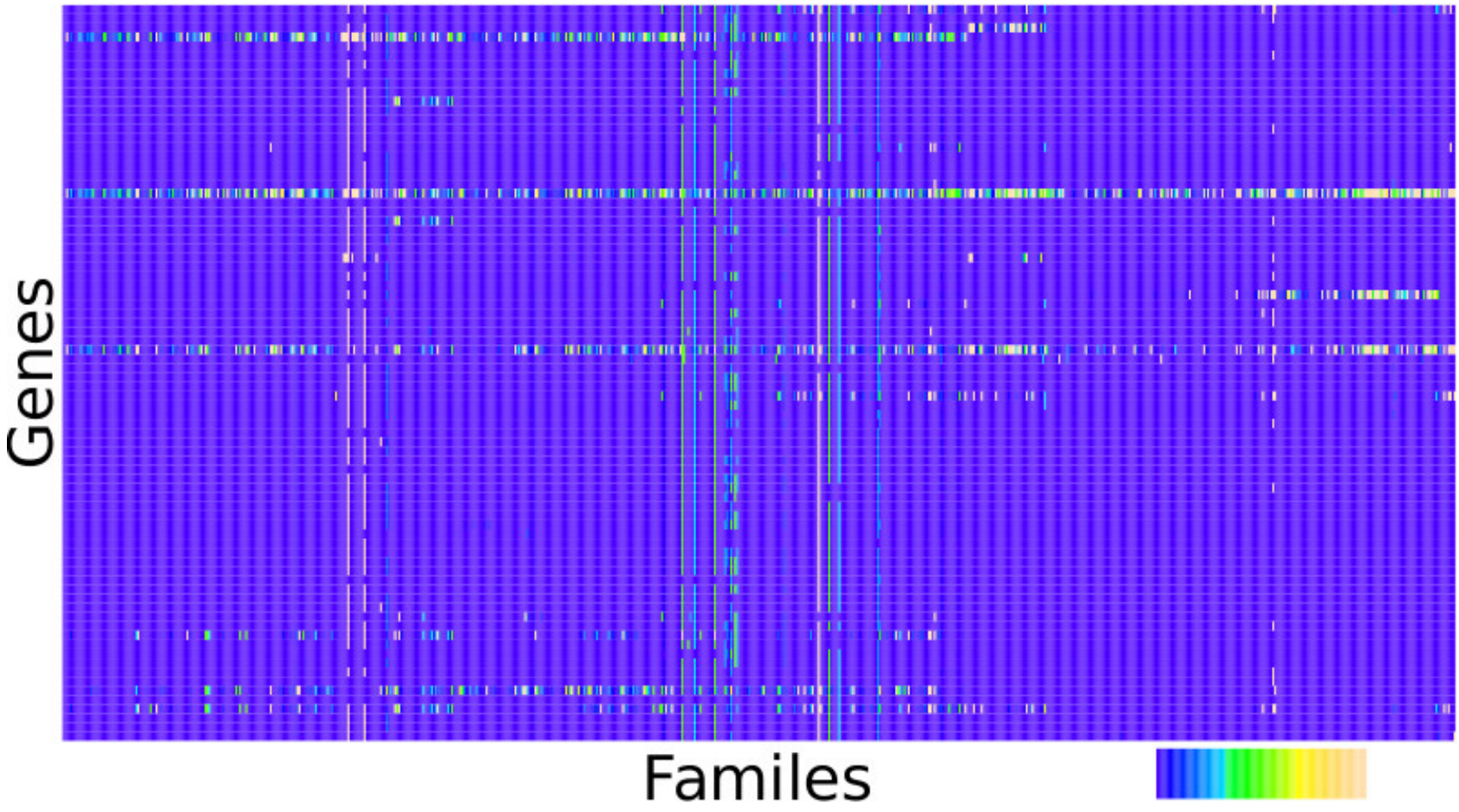
GenBank stores species taxonomy, but not gene taxonomy

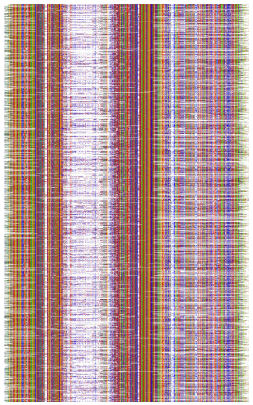
hard to get gene identifications/categories for all of GenBank (e.g., clustering)

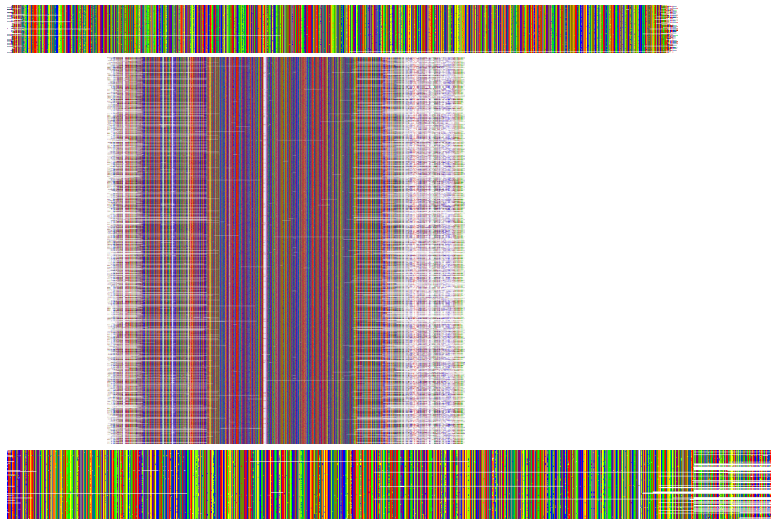
all genes for species x, but NOT all species for gene y



Zanne et al., Nature, 2013


Genes
Familes





trnL-trnF = 2485

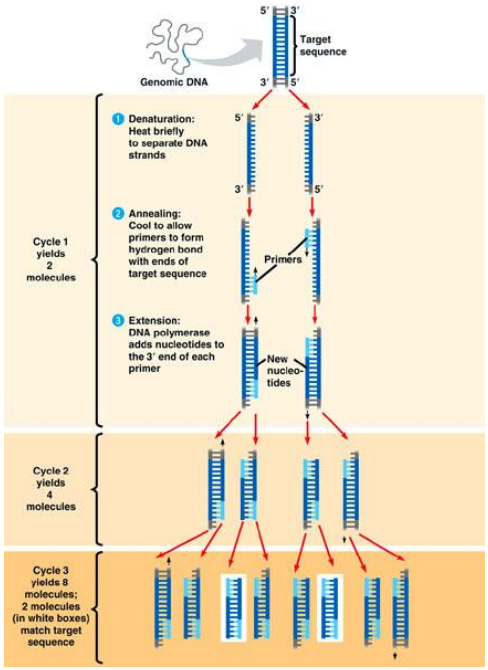
atpB = 107

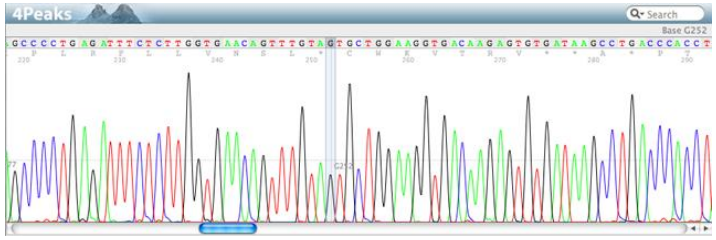
matK = 2182

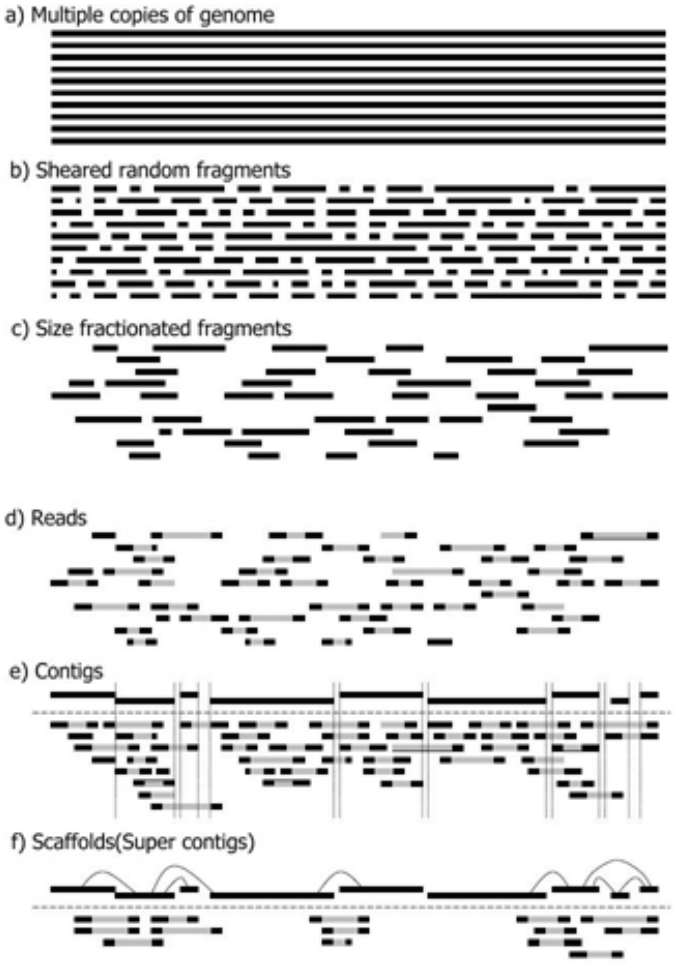
rbcL = 924

ITS = 3725

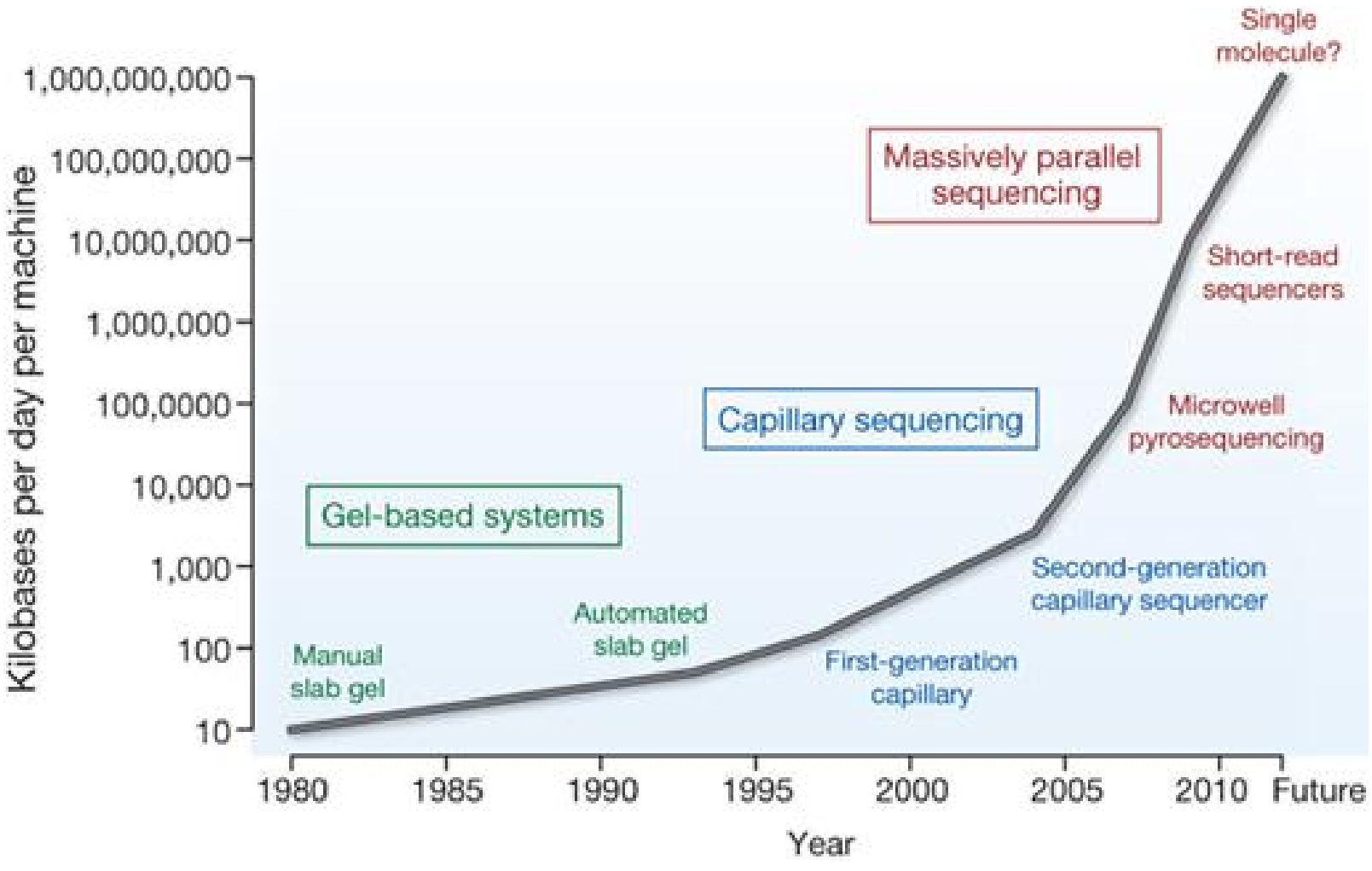
Massive parallel sequencing







Sequencing technology has improved exponentially



How many genes are used for a phylogenetic analysis?

Typical phylogenetic analyses

1-10 genes

17 genes. Plants
(Soltis et al. 2011)

19 genes. Birds
(Hackett et al. 2007)

Transcriptomic and genomic phylogenetic analyses

140 genes. Metazoa
(Dunn et al. 2008)

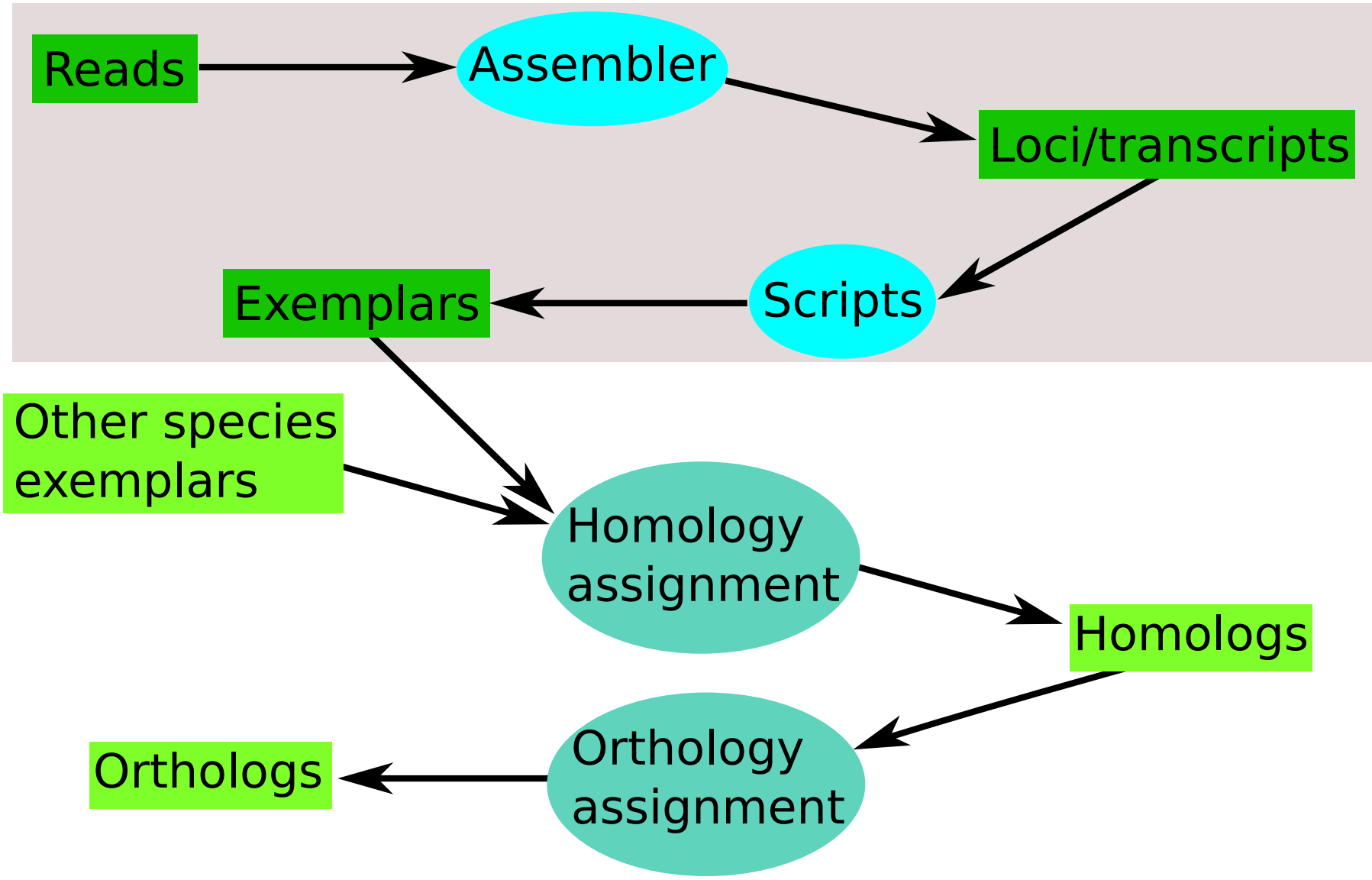
1185 genes. Molluscs
(Smith et al. 2011)

2970 genes. Seed plants
(Lee et al. 2011)

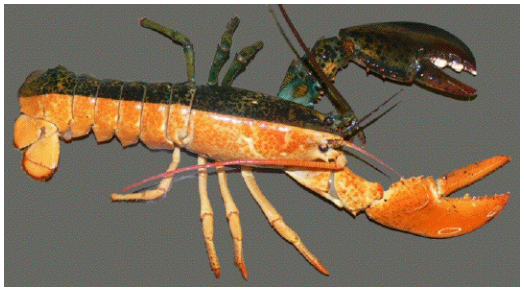
8251 genes. Birds
(Jarvis et al. 2014)

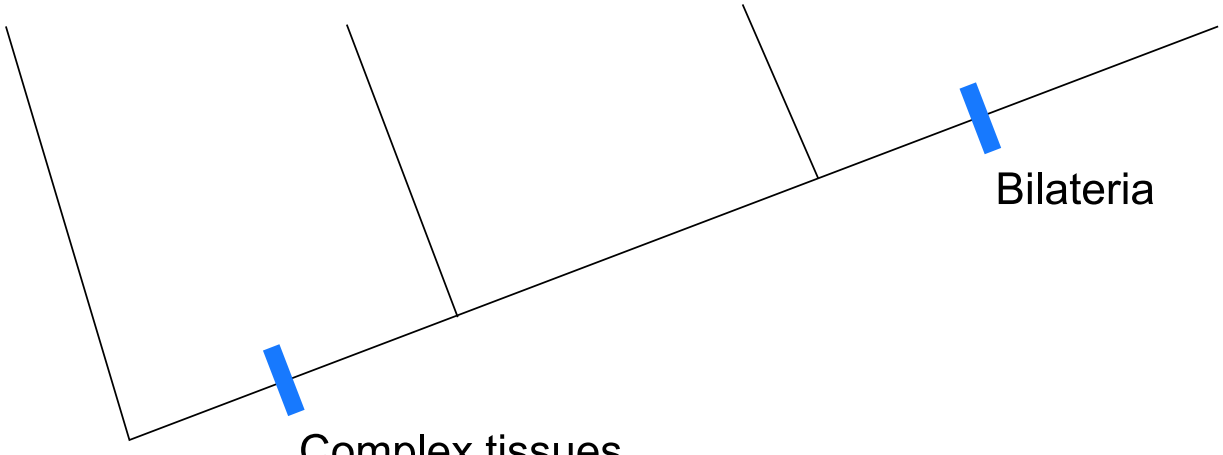
20,374 genes. Equids.
(Jónsson et al. 2014)

Pipeline

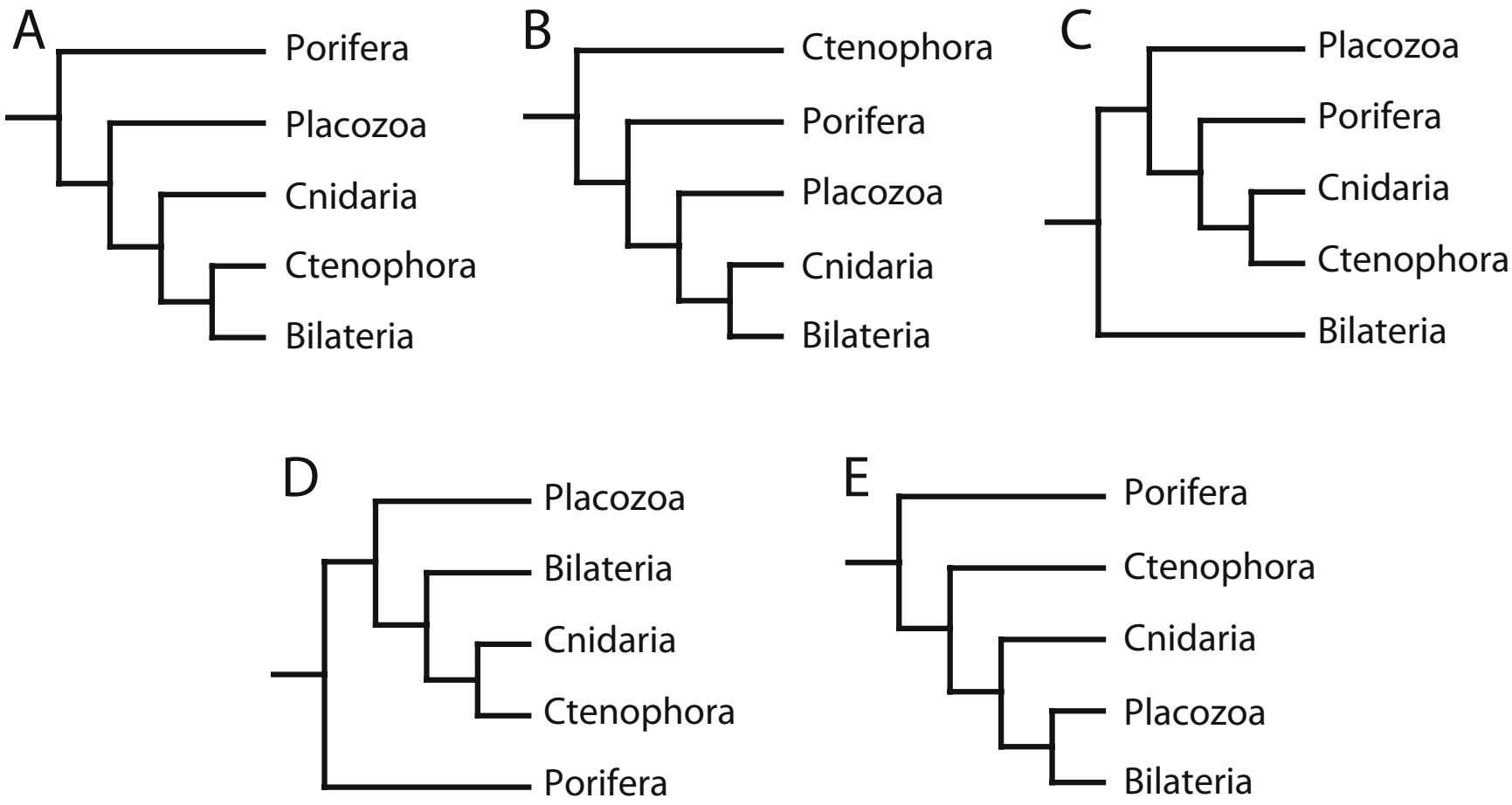

Reads
Assembler
Loci/transcripts
Scripts
Exemplars
Other species exemplars
Homology assignment
Homologs
Orthology assignment
Orthologs



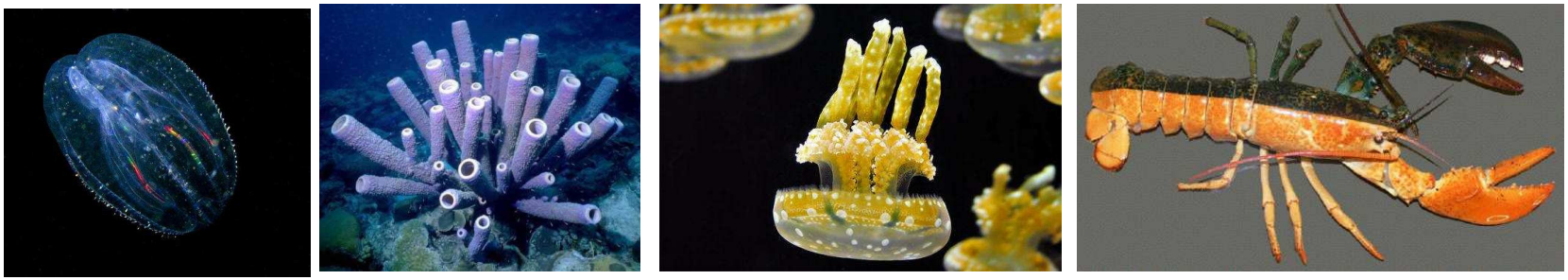

Bilateria
Complex tissues
Neural cells
Mesoderm cells

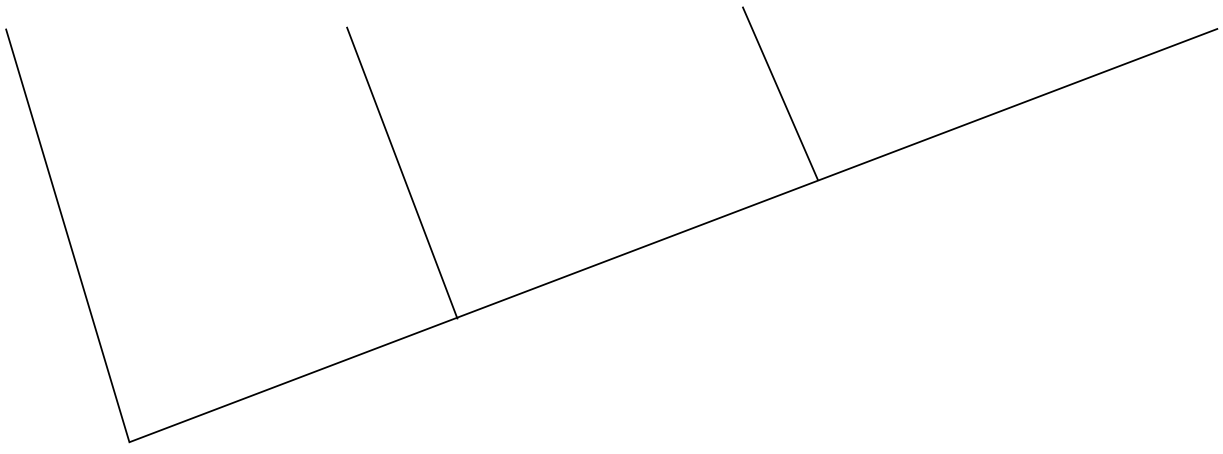
Phylogenetic discordance in animals



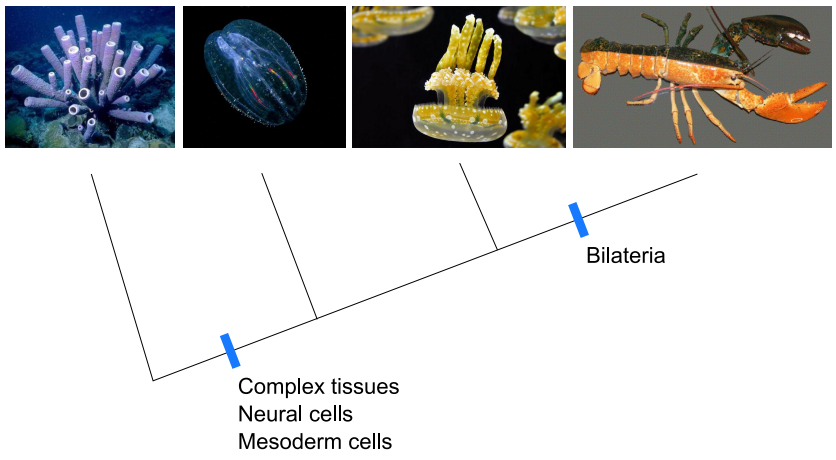
Edgecombe et al. 2011

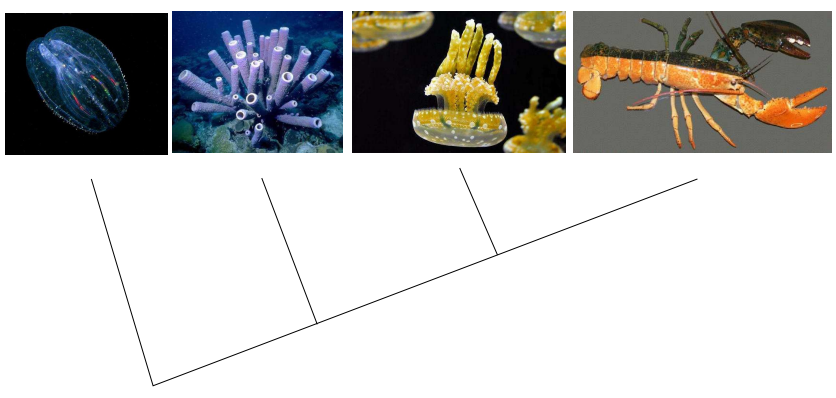
Alternative phylogeny also has support





Which is the actual order of divergence?





Traits

tissue present

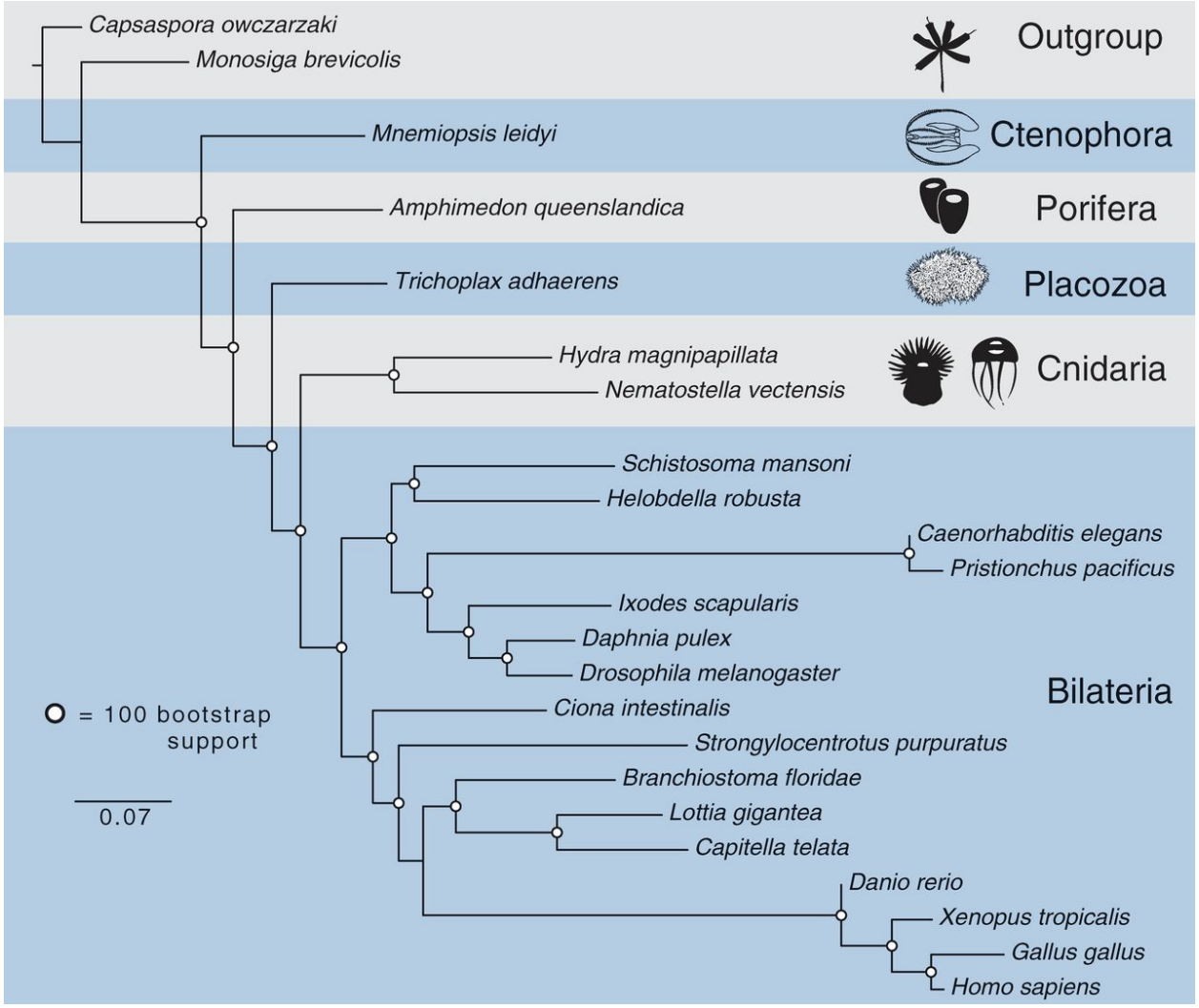
cell types

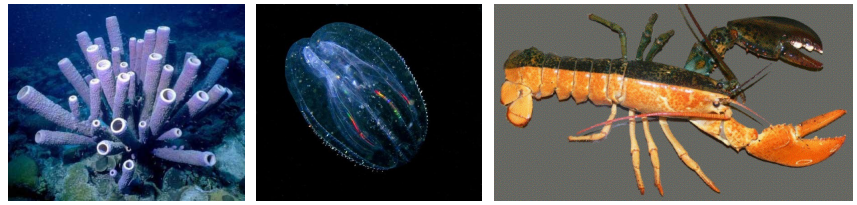
motility

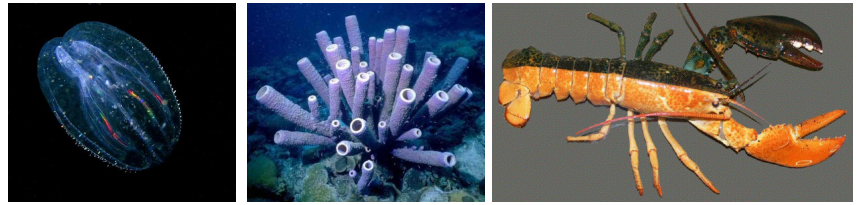
symmetry

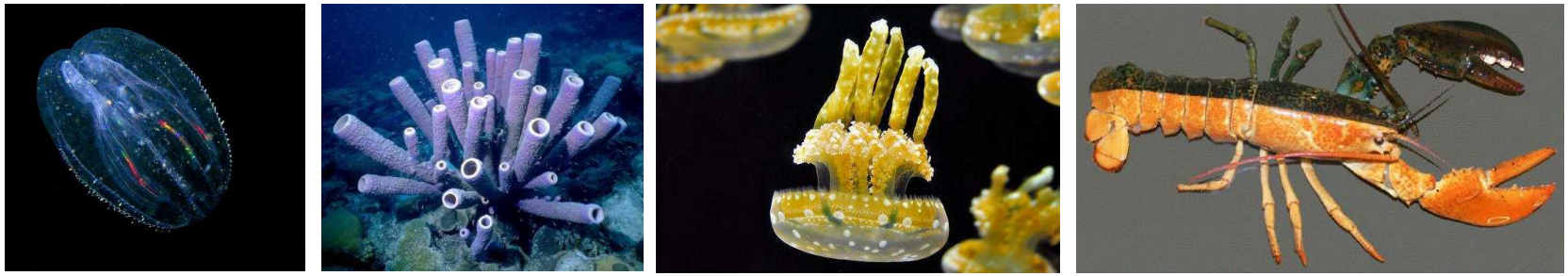
...

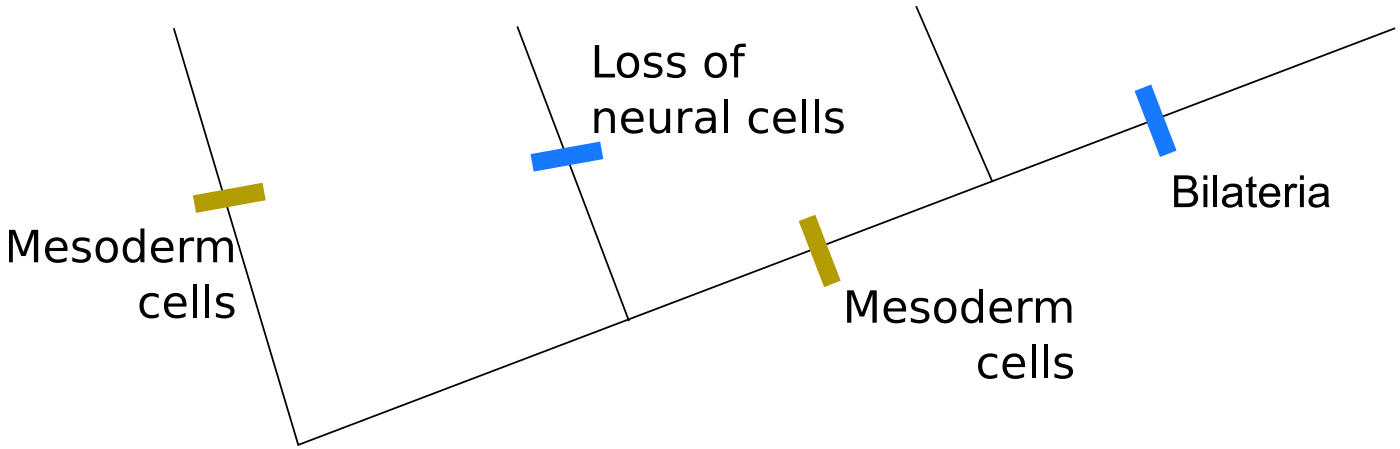
242 genes; Ryan et al., 2013



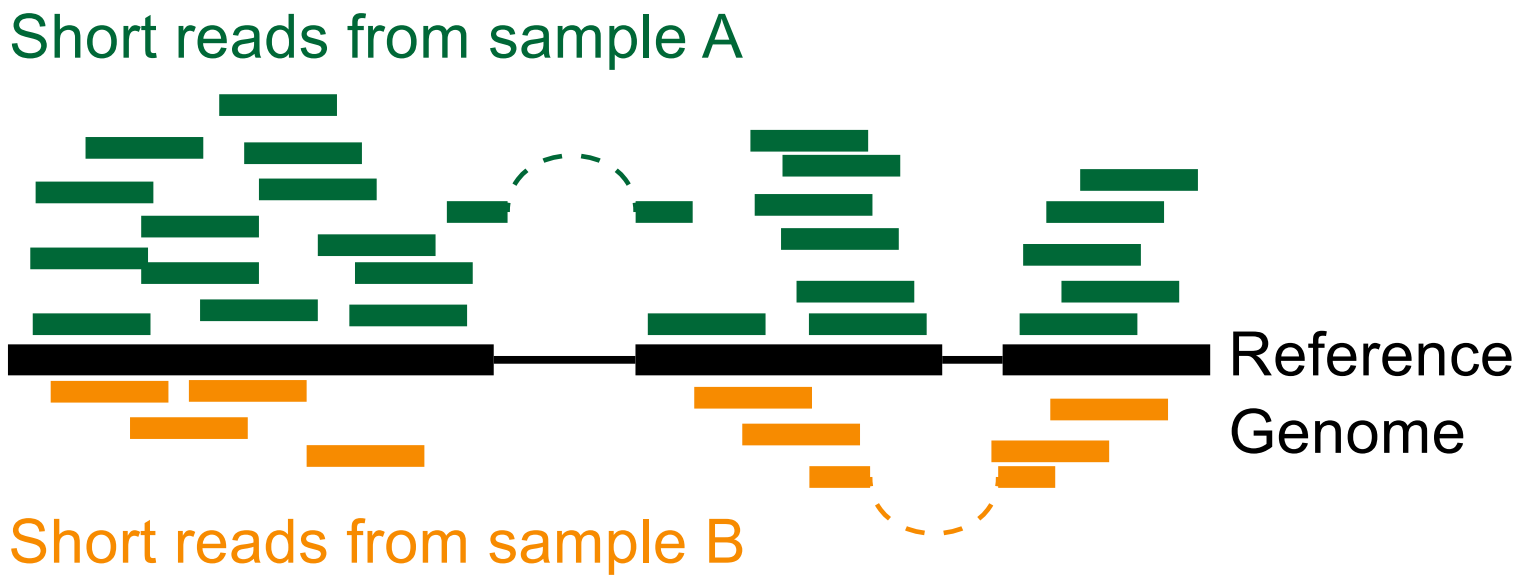




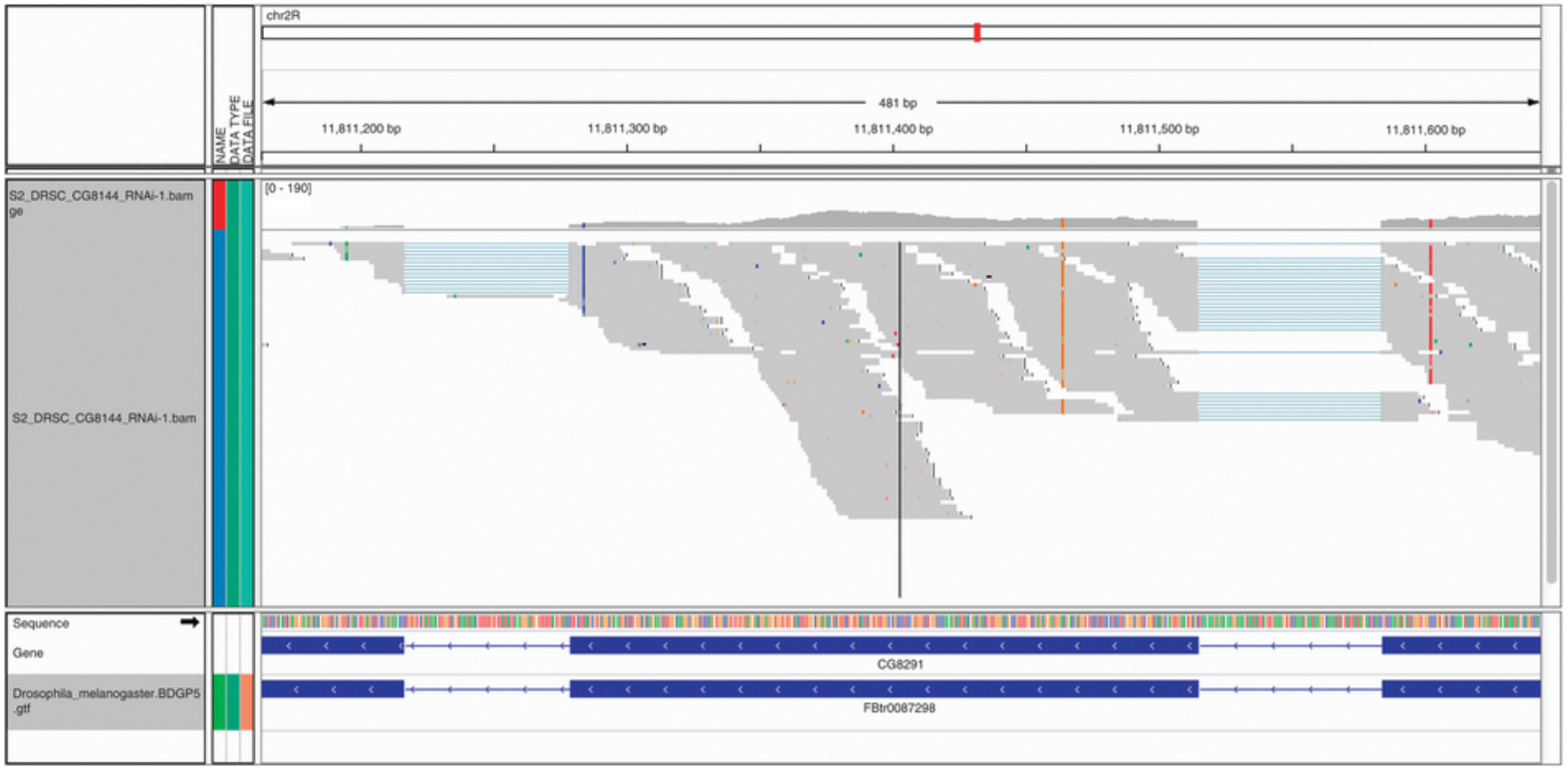



Loss of
neural cells
Bilateria
Mesoderm
cells
Mesoderm
cells

Short read “mapping” (i.e., aligning to a reference)

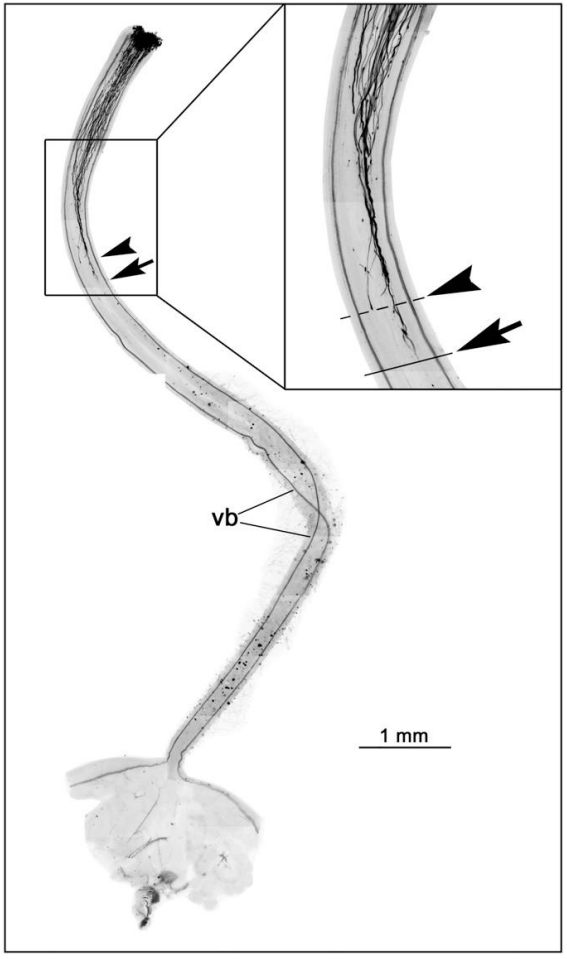


Short read mapping also used for populations to map variation

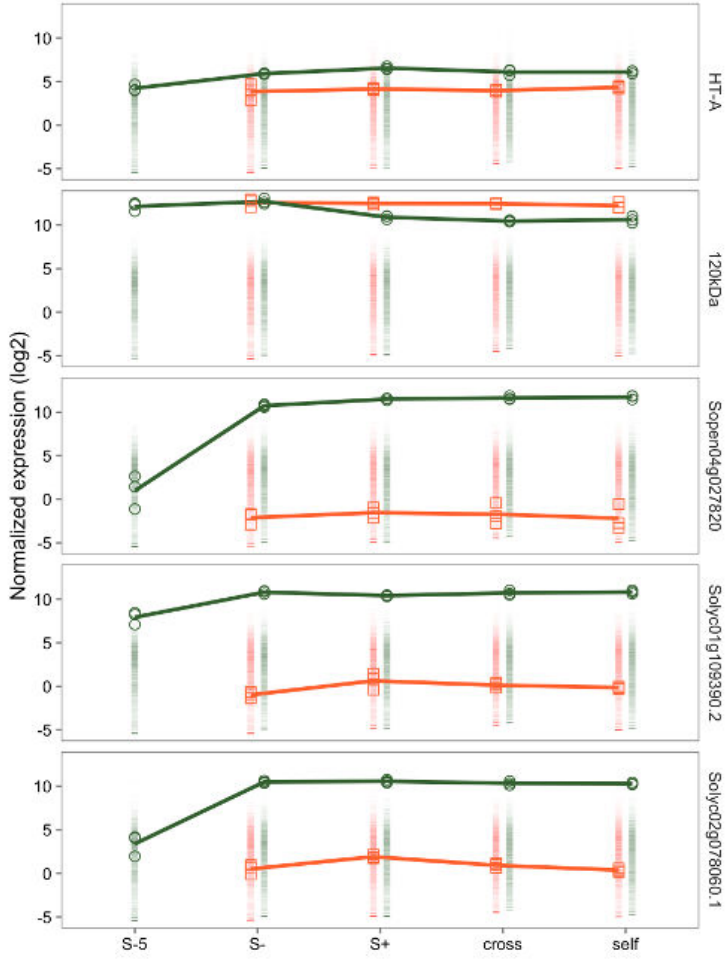


Anders et al., 2013, Nature Protocols

Using aligned read counts to determine pollen tube rejection genes



Baek et al 2015 Am J Bot



Pease et al. 2016 Molecular Ecology

All these techniques require some sort of pairwise alignment (and homology assessment)

Major topics covered

Detecting homologous sequences (one-to-one) with pairwise sequence alignment

- global → Needleman-Wunsch-Gotoh algorithm
- local → Smith-Waterman algorithm
- BLAST → heuristic search

Grouping homologs (many-to-many)

- Markov clustering
- other clustering methods

Schedule

First

- Detecting homologous sequences
- Lab for detecting homologous sequences

Second

- Clustering sequences
- Short lab for clustering sequences

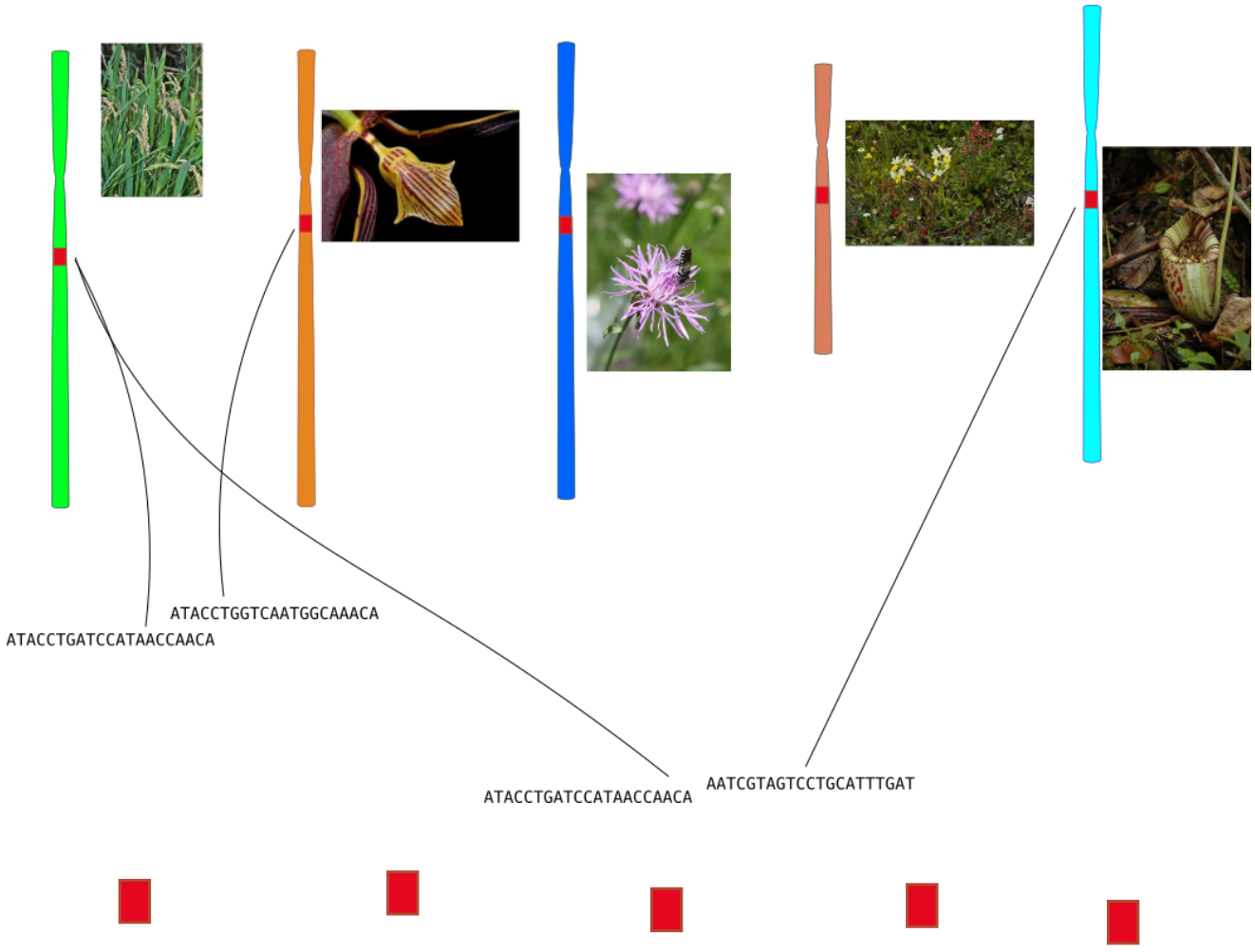
What are we going to address with pairwise alignment?

Questions from a biological perspective:

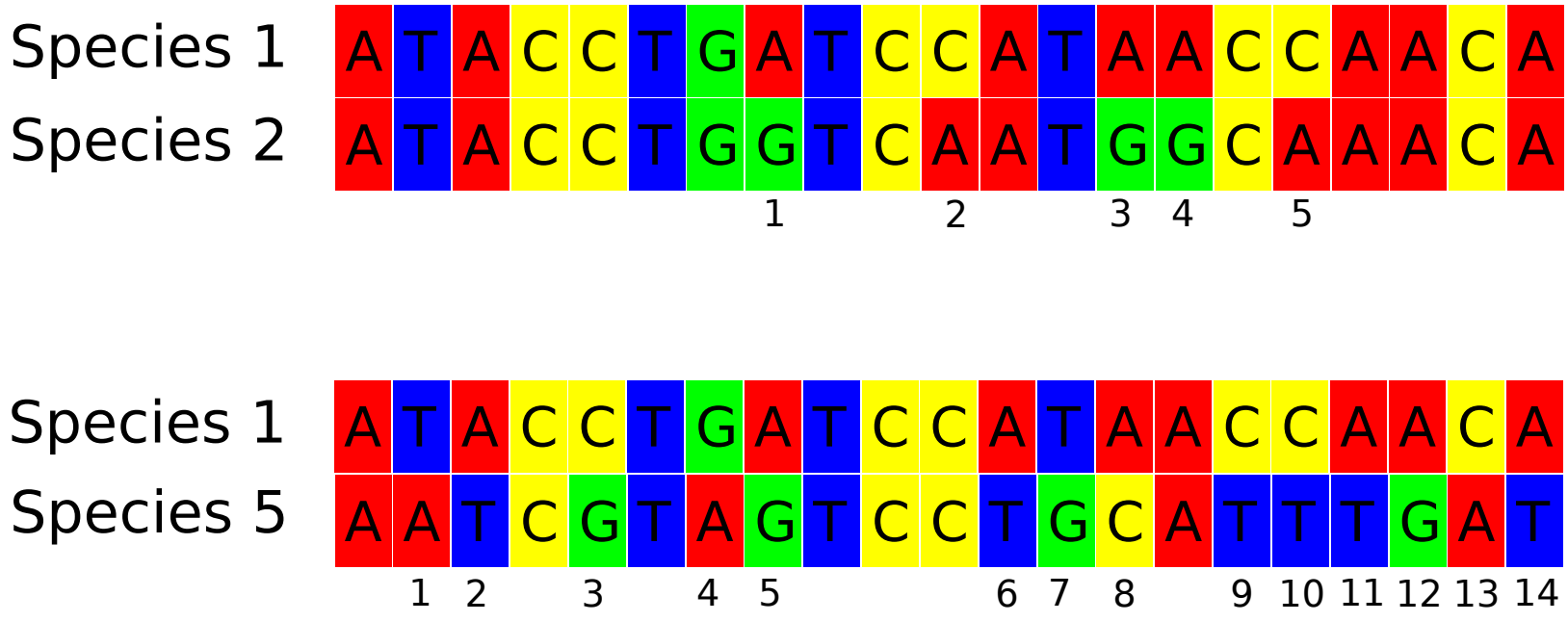
- Are these sequences homologous?
  Do they share common ancestry?

Questions from an informatics perspective:

- Should a sequence be removed? Too short or too noisy (information poor)?
  Are these sequences “similar enough” to include in analysis? (cut-offs)


ATACCTGATCCATAACCAACA
ATACCTGGTCAATGGCAAACA
ATACCTGATCCATAACCAACA
AATCGTAGTCCTGCATTTGAT

Comparisons



Aligning words can be trivial...

```
Species 1:  SOMEONE
Species 2:  AWESOME
```

Aligning words can be trivial...

```
Species 1:  SOMEONE
Species 2:  AWESOME

Species 1:  ---SOMEONE
Species 2:  AWESOME---
```

… or less trivial

```
Species 1:  ACGTTAGA
Species 2:  CGTTGAA
```

… or less trivial

```
Species 1:  ACGTTAGA
Species 2:  CGTTGAA

Species 1:  -------ACGTTAGA
Species 2:  CGTTGAA-------
```

… or less trivial

```
Species 1:  ACGTTAGA
Species 2:  CGTTGAA

Species 1:  --------ACGTTAGA
Species 2:  CGTTGAA--------

Species 1:  ACGTTAGA-
Species 2:  -CGTT-GAA
```

Alignments are evaluated by a quantitative score

```
Species 1:  --------ACGTTAGA
Species 2:  CGTTGAA---------
```

score: -15 (gaps = -1, match = 1)

```
Species 1:  ACGTTAGA-
Species 2:  -CGTT-GAA
```

score: 3

Example for you to try:

```
Species 1:  TTGGCACGTTAGA
Species 2:  TGCACCTTAGTTA
```

How do we get the “best” alignment?

We cannot enumerate all of the possible alignments

We **can** find the best alignments algorithmically

- using Dynamic Programming (no be covered here, but there are great resources if you are interested)
  solve a large problem by breaking it down and solve sub-problems

Needleman-Wunsch is the standard global alignment algorithm

- published in 1970
  cited over 7000 times

Example to try:

```
Species 1:  TTGGCACGTTAGA
Species 2:  TGCACCTTAGTTA
```

+1 match, -1 mismatch

Needleman-Wunsch result

```
Species 1:  TTGGCACGTTAGA
Species 2:  TGCACCTTAGTTA
```

```
Species 1:  TTGGCA-CGTTAG--A
Species 2:  -T-GCACC-TTAGTTA
```

Amino acids often use more complex score matrices

Don’t just have to have match and mismatch

NUC.4.4 from NCBI also known as “EDNAFULL”

ftp://ftp.ncbi.nih.gov/blast/matrices/

	A	T	G	C	S	W	R	Y	K	M	B	V	H	D	N	U
A	5	-4	-4	-4	-4	1	1	-4	-4	1	-4	-1	-1	-1	-2	-4
T	-4	5	-4	-4	-4	1	-4	1	1	-4	-1	-4	-1	-1	-2	5
G	-4	-4	5	-4	1	-4	1	-4	1	-4	-1	-1	-4	-1	-2	-4
C	-4	-4	-4	5	1	-4	-4	1	-4	1	-1	-1	-1	-4	-2	-4
S	-4	-4	1	1	-1	-4	-2	-2	-2	-2	-1	-1	-3	-3	-1	-4
W	1	1	-4	-4	-4	-1	-2	-2	-2	-2	-3	-3	-1	-1	-1	1
R	1	-4	1	-4	-2	-2	-1	-4	-2	-2	-3	-1	-3	-1	-1	-4
Y	-4	1	-4	1	-2	-2	-4	-1	-2	-2	-1	-3	-1	-3	-1	1
K	-4	1	1	-4	-2	-2	-2	-2	-1	-4	-1	-3	-3	-1	-1	1
M	1	-4	-4	1	-2	-2	-2	-2	-4	-1	-3	-1	-1	-3	-1	-4
B	-4	-1	-1	-1	-1	-3	-3	-1	-1	-3	-1	-2	-2	-2	-1	-1
V	-1	-4	-1	-1	-1	-3	-1	-3	-3	-1	-2	-1	-2	-2	-1	-4
H	-1	-1	-4	-1	-3	-1	-3	-1	-3	-1	-2	-2	-1	-2	-1	-1
D	-1	-1	-1	-4	-3	-1	-1	-3	-1	-3	-2	-2	-2	-1	-1	-1
N	-2	-2	-2	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2
U	-4	5	-4	-4	-4	1	-4	1	1	-4	-1	-4	-1	-1	-2	5

How we score gaps can vary considerably among models

- constant gap model
  - one gap = one penalty
    `Species 1:  GTTAGTTAC`
    `Species 2:  GTTA----C`
    match = 1, gap = -1, score = 4 (+5,-1)
- linear gap model (what we have done)
  - gap still has the one parameter, but take into account length
    match = 1, gap = -1, score = 1 (+5,-4)
- affine gap model
  - parameter for **opening** a gap
    AND a separate parameter for **extending** a gap
    match = 1, open = -2, ext = -1, score = -1 (+5,-2+(-1*4))

More complicated gap models can also exist (or be specified)

```
Species 1:  TTGGCACGTTAGA
Species 2:  TGCACCTTAGTTA
```

```
Species 1:  TTGGCA-CGTTAG--A
Species 2:  -T-GCACC-TTAGTTA
```

Gap open: -100, Gap extend: -0.0005

```
Species 1:  TTGGCACGTTAGA--
Species 2:  --TGCACCTTAGTTA
```

play around on
https://www.ebi.ac.uk/Tools/psa/emboss_needle/nucleotide.html

Protein

Algorithmically, protein (amino acid) alignments are not different

The major differences involve the scoring matrices

Major choices are PAM and BLOSUM

PAM: Point Accepted Mutation (Dayhoff et al.)

- this was typically used before the 1990s

BLOSUM

- has generally replaced PAM as the common matrix

PAM matrices indicate divergence

refers to evolutionary difference
(higher the number → more divergent)

based on evolutionary models and empirical data

ranges from closely related to completely random

includes {PAM250, PAM120, PAM1}

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	3	-3	-1	0	-3	-1	0	1	-3	-1	-3	-2	-2	-4	1	1	1	-7	-4	0	0	-1	-1	-8
R	-3	6	-1	-3	-4	1	-3	-4	1	-2	-4	2	-1	-5	-1	-1	-2	1	-5	-3	-2	-1	-2	-8
N	-1	-1	4	2	-5	0	1	0	2	-2	-4	1	-3	-4	-2	1	0	-4	-2	-3	3	0	-1	-8
D	0	-3	2	5	-7	1	3	0	0	-3	-5	-1	-4	-7	-3	0	-1	-8	-5	-3	4	3	-2	-8
C	-3	-4	-5	-7	9	-7	-7	-4	-4	-3	-7	-7	-6	-6	-4	0	-3	-8	-1	-3	-6	-7	-4	-8
Q	-1	1	0	1	-7	6	2	-3	3	-3	-2	0	-1	-6	0	-2	-2	-6	-5	-3	0	4	-1	-8
E	0	-3	1	3	-7	2	5	-1	-1	-3	-4	-1	-3	-7	-2	-1	-2	-8	-5	-3	3	4	-1	-8
G	1	-4	0	0	-4	-3	-1	5	-4	-4	-5	-3	-4	-5	-2	1	-1	-8	-6	-2	0	-2	-2	-8
H	-3	1	2	0	-4	3	-1	-4	7	-4	-3	-2	-4	-3	-1	-2	-3	-3	-1	-3	1	1	-2	-8
I	-1	-2	-2	-3	-3	-3	-3	-4	-4	6	1	-3	1	0	-3	-2	0	-6	-2	3	-3	-3	-1	-8
L	-3	-4	-4	-5	-7	-2	-4	-5	-3	1	5	-4	3	0	-3	-4	-3	-3	-2	1	-4	-3	-2	-8
K	-2	2	1	-1	-7	0	-1	-3	-2	-3	-4	5	0	-7	-2	-1	-1	-5	-5	-4	0	-1	-2	-8
M	-2	-1	-3	-4	-6	-1	-3	-4	-4	1	3	0	8	-1	-3	-2	-1	-6	-4	1	-4	-2	-2	-8
F	-4	-5	-4	-7	-6	-6	-7	-5	-3	0	0	-7	-1	8	-5	-3	-4	-1	4	-3	-5	-6	-3	-8
P	1	-1	-2	-3	-4	0	-2	-2	-1	-3	-3	-2	-3	-5	6	1	-1	-7	-6	-2	-2	-1	-2	-8
S	1	-1	1	0	0	-2	-1	1	-2	-2	-4	-1	-2	-3	1	3	2	-2	-3	-2	0	-1	-1	-8
T	1	-2	0	-1	-3	-2	-2	-1	-3	0	-3	-1	-1	-4	-1	2	4	-6	-3	0	0	-2	-1	-8
W	-7	1	-4	-8	-8	-6	-8	-8	-3	-6	-3	-5	-6	-1	-7	-2	-6	12	-2	-8	-6	-7	-5	-8
Y	-4	-5	-2	-5	-1	-5	-5	-6	-1	-2	-2	-5	-4	4	-6	-3	-3	-2	8	-3	-3	-5	-3	-8
V	0	-3	-3	-3	-3	-3	-3	-2	-3	3	1	-4	1	-3	-2	-2	0	-8	-3	5	-3	-3	-1	-8
B	0	-2	3	4	-6	0	3	0	1	-3	-4	0	-4	-5	-2	0	0	-6	-3	-3	4	2	-1	-8
Z	-1	-1	0	3	-7	4	4	-2	1	-3	-3	-1	-2	-6	-1	-1	-2	-7	-5	-3	2	4	-1	-8
X	-1	-2	-1	-2	-4	-1	-1	-2	-2	-1	-2	-2	-2	-3	-2	-1	-1	-5	-3	-1	-1	-1	-2	-8
*	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	1

BLOSUM matrices indicate percent identity

refers to percent identity
(higher the number → *less* divergent)

based on empirical data

BLOSUM has a narrower range than PAM

includes {BLOSUM45, BLOSUM62, BLOSUM80}

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1	-4
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1	-4
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-4
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

Differences between PAM and BLOSUM

PAM120, BLOSUM30

A-A: PAM=3, BLOSUM=4

A-R: PAM=−3, BLOSUM=−1

R-N: PAM=−1, BLOSUM=0

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	3	-3	-1	0	-3	-1	0	1	-3	-1	-3	-2	-2	-4	1	1	1	-7	-4	0	0	-1	-1	-8
R	-3	6	-1	-3	-4	1	-3	-4	1	-2	-4	2	-1	-5	-1	-1	-2	1	-5	-3	-2	-1	-2	-8
N	-1	-1	4	2	-5	0	1	0	2	-2	-4	1	-3	-4	-2	1	0	-4	-2	-3	3	0	-1	-8
D	0	-3	2	5	-7	1	3	0	0	-3	-5	-1	-4	-7	-3	0	-1	-8	-5	-3	4	3	-2	-8
C	-3	-4	-5	-7	9	-7	-7	-4	-4	-3	-7	-7	-6	-6	-4	0	-3	-8	-1	-3	-6	-7	-4	-8
Q	-1	1	0	1	-7	6	2	-3	3	-3	-2	0	-1	-6	0	-2	-2	-6	-5	-3	0	4	-1	-8
E	0	-3	1	3	-7	2	5	-1	-1	-3	-4	-1	-3	-7	-2	-1	-2	-8	-5	-3	3	4	-1	-8
G	1	-4	0	0	-4	-3	-1	5	-4	-4	-5	-3	-4	-5	-2	1	-1	-8	-6	-2	0	-2	-2	-8
H	-3	1	2	0	-4	3	-1	-4	7	-4	-3	-2	-4	-3	-1	-2	-3	-3	-1	-3	1	1	-2	-8
I	-1	-2	-2	-3	-3	-3	-3	-4	-4	6	1	-3	1	0	-3	-2	0	-6	-2	3	-3	-3	-1	-8
L	-3	-4	-4	-5	-7	-2	-4	-5	-3	1	5	-4	3	0	-3	-4	-3	-3	-2	1	-4	-3	-2	-8
K	-2	2	1	-1	-7	0	-1	-3	-2	-3	-4	5	0	-7	-2	-1	-1	-5	-5	-4	0	-1	-2	-8
M	-2	-1	-3	-4	-6	-1	-3	-4	-4	1	3	0	8	-1	-3	-2	-1	-6	-4	1	-4	-2	-2	-8
F	-4	-5	-4	-7	-6	-6	-7	-5	-3	0	0	-7	-1	8	-5	-3	-4	-1	4	-3	-5	-6	-3	-8
P	1	-1	-2	-3	-4	0	-2	-2	-1	-3	-3	-2	-3	-5	6	1	-1	-7	-6	-2	-2	-1	-2	-8
S	1	-1	1	0	0	-2	-1	1	-2	-2	-4	-1	-2	-3	1	3	2	-2	-3	-2	0	-1	-1	-8
T	1	-2	0	-1	-3	-2	-2	-1	-3	0	-3	-1	-1	-4	-1	2	4	-6	-3	0	0	-2	-1	-8
W	-7	1	-4	-8	-8	-6	-8	-8	-3	-6	-3	-5	-6	-1	-7	-2	-6	12	-2	-8	-6	-7	-5	-8
Y	-4	-5	-2	-5	-1	-5	-5	-6	-1	-2	-2	-5	-4	4	-6	-3	-3	-2	8	-3	-3	-5	-3	-8
V	0	-3	-3	-3	-3	-3	-3	-2	-3	3	1	-4	1	-3	-2	-2	0	-8	-3	5	-3	-3	-1	-8
B	0	-2	3	4	-6	0	3	0	1	-3	-4	0	-4	-5	-2	0	0	-6	-3	-3	4	2	-1	-8
Z	-1	-1	0	3	-7	4	4	-2	1	-3	-3	-1	-2	-6	-1	-1	-2	-7	-5	-3	2	4	-1	-8
X	-1	-2	-1	-2	-4	-1	-1	-2	-2	-1	-2	-2	-2	-3	-2	-1	-1	-5	-3	-1	-1	-1	-2	-8
*	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	-8	1

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1	-4
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1	-4
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-4
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

Species 1: HEAGAWGHEE

Species 2: PAWHEAE

```
Species 1:  HEAGAWGHEE
Species 2:  PAWHEAE
```

```
BLOSUM62 and -2 gap
Species 1:  HEAGAWGHE-E
Species 2:  --P-AW-HEAE
```

```
Species 1:  HEAGAWGHEE
Species 2:  PAWHEAE
```

```
BLOSUM62 and -2 gap
Species 1:  HEAGAWGHE-E
Species 2:  --P-AW-HEAE
```

```
BLOSUM30 and -2 gap
Species 1:  HEAGAWGHE-E
Species 2:  -P--AW-HEAE
```

How do we measure whether an alignment is significant?

Needleman-Wunsch will always give the best alignment (within its model assumptions)

In theory:

No method allows us to predict the distribution of alignment scores from random sequences

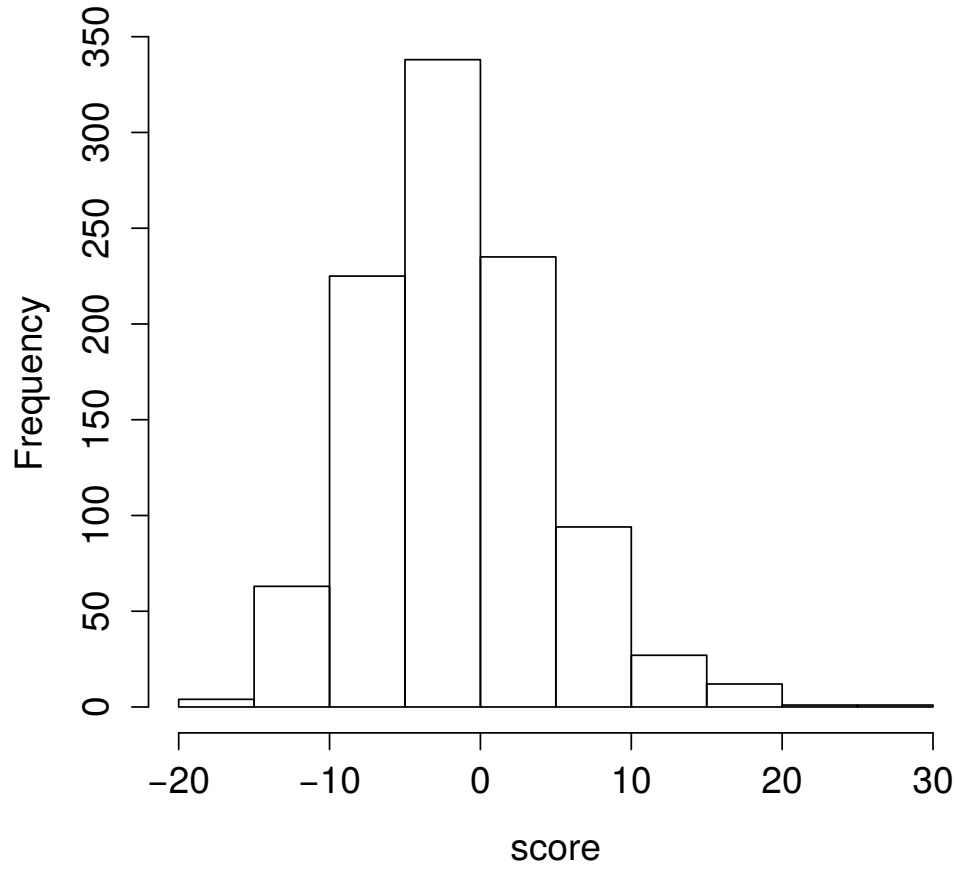
In practice:

generate a distribution simulated alignments and get scores
compare our score to the distribution
if 100 random alignments give scores that are lower than the observed alignment score, then $P < 0.01$.

```
Species 1:  HEAGAWGHE-E
Species 2:  --P-AW-HEAE
```

score = 22 with BLOSUM62 and -2 gap

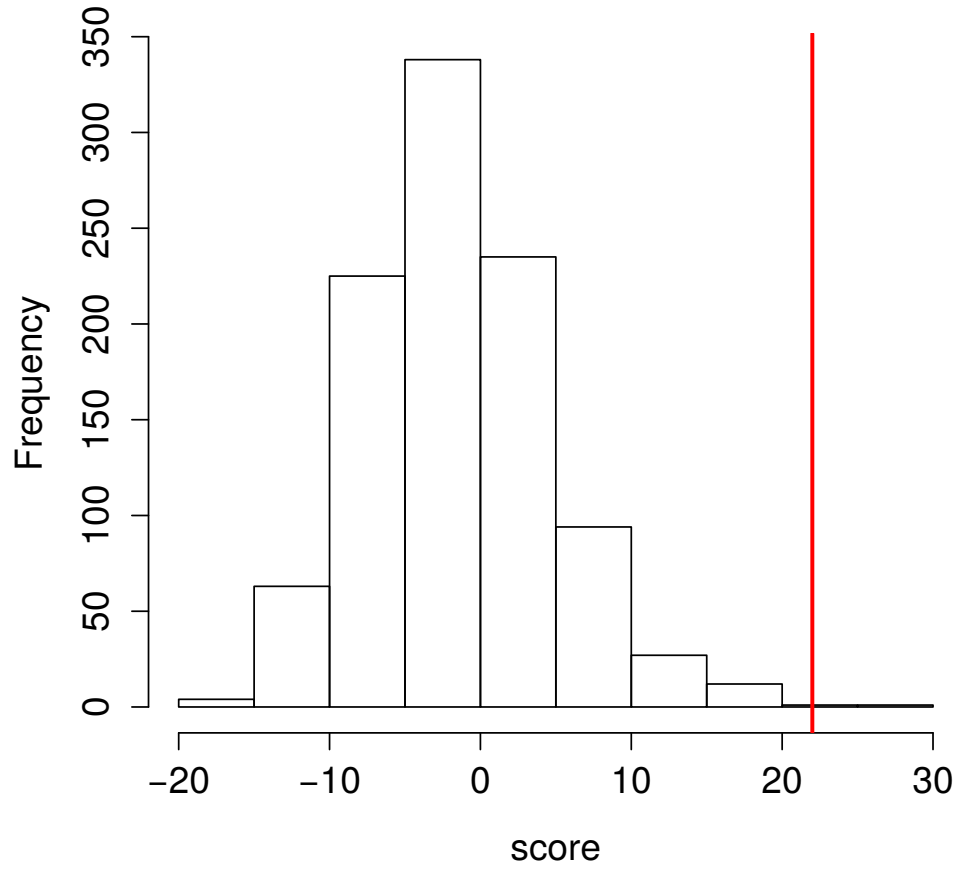
1/1000 > 22



```
Species 1:  HEAGAWGHE-E
Species 2:  --P-AW-HEAE
```

score = 22 with BLOSUM62 and -2 gap

1/1000 > 22



Global alignment is not the same as local alignment

```
Species 1:  SOMEONE
Species 2:  AWESOME
```

Global alignment is not the same as local alignment

```
Species 1:  SOMEONE
Species 2:  AWESOME

Species 1:  ---SOMEONE
Species 2:  AWESOME---
```

Global alignment is not the same as local alignment

```
Species 1:  SOMEONE
Species 2:  AWESOME

Species 1:  ---SOMEONE
Species 2:  AWESOME---

Species 1:  SOME
Species 2:  SOME
```

Pairwise alignment

Needleman-Wunsch is the standard **global alignment** algorithm

- published 1970, cited over 9600 times

Smith-Waterman is the standard **local alignment** algorithm

- published 1981, cited over 8800 times
  also a dynamic programming algorithm

Difference

```
Species 1:  ACGTTAGA
Species 2:  CGTTGAA
```

```
Species 1:  ACGTTAGA
Species 2:  -CGTTGAA
```

```
Species 1:  CGTTAGA
Species 2:  CGTTGAA
```

Another example

(aminio acids with BLOSUM 62)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V	B	Z	X	*
A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	1	0	-3	-2	0	-2	-1	0	-4
R	-1	5	0	-2	-3	1	0	-2	0	-3	-2	2	-1	-3	-2	-1	-1	-3	-2	-3	-1	0	-1	-4
N	-2	0	6	1	-3	0	0	0	1	-3	-3	0	-2	-3	-2	1	0	-4	-2	-3	3	0	-1	-4
D	-2	-2	1	6	-3	0	2	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3	4	1	-1	-4
C	0	-3	-3	-3	9	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1	-3	-3	-2	-4
Q	-1	1	0	0	-3	5	2	-2	0	-3	-2	1	0	-3	-1	0	-1	-2	-1	-2	0	3	-1	-4
E	-1	0	0	2	-4	2	5	-2	0	-3	-3	1	-2	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3	-1	-2	-1	-4
H	-2	0	1	-1	-3	0	0	-2	8	-3	-3	-1	-2	-1	-2	-1	-2	-2	2	-3	0	0	-1	-4
I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	2	-3	1	0	-3	-2	-1	-3	-1	3	-3	-3	-1	-4
L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4	-2	2	0	-3	-2	-1	-2	-1	1	-4	-3	-1	-4
K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2	0	1	-1	-4
M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5	0	-2	-1	-1	-1	-1	1	-3	-1	-1	-4
F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6	-4	-2	-2	1	3	-1	-3	-3	-1	-4
P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2	-2	-1	-2	-4
S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4	1	-3	-2	-2	0	0	0	-4
T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5	-2	-2	0	-1	-1	0	-4
W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11	2	-3	-4	-3	-2	-4
Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	-1	-3	-2	-1	-4
V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4	-3	-2	-1	-4
B	-2	-1	3	4	-3	0	1	-1	0	-3	-4	0	-3	-3	-2	0	-1	-4	-3	-3	4	1	-1	-4
Z	-1	0	0	1	-3	3	4	-2	0	-3	-3	1	-1	-3	-1	0	-1	-3	-2	-2	1	4	-1	-4
X	0	-1	-1	-1	-2	-1	-1	-1	-1	-1	-1	-1	-1	-1	-2	0	0	-2	-1	-1	-1	-1	-1	-4
*	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	-4	1

```
Species 1:   HEAGAWGHEE
Species 2:   PAWHEAE
```

Global means full-sequence alignment, local focuses on best aligned subregion

```
Species 1:  HEAGAWGHEE
Species 2:  PAWHEAE
```

```
Species 1:  AWGHE-E
Species 2:  AW-HEAE
```

```
Species 1:  HEAGAWGHE-E
Species 2:  --P-AW-HEAE
```

When are sequences homologous? (decision is difficult)

- What properties are used to qualify/quantify homology?
  - sequences that are similar ***enough*** are homologous
  - find the best alignment defining similarity
  - calculate alignment scores
  - Other properties?
- What constitutes similar enough?
  - more similar than by chance?
  - how is the probability of this chance determined?

How do we calculate the significance of alignments?

We have calculated the optimal alignment (best score) called the “maximum segment pair” (MSP)

(This does not depend on sequences relatedness)

Significance: How many MSPs do we expect with at least the same score by chance?

How do we calculate the significance of alignments?

If query sequence “GeneQ” and a given target sequence “GeneT” have MSP score S, then what is the probability an MSP between GeneQ to other random target sequences in the database with MSP $\geq S$?

E-value (expectation)

How do we calculate the significance of alignments?

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

K and λ = scaling parameters calculated based on the search space (K) and scoring scheme (λ)
m = length of query sequence
n = total length of database sequences

The probability of finding at least one alignment with our score (the P value) is $P = 1 - e^{-E(S)}$

So E and P decrease exponentially as score increases, but increase as the database increase in size.

Calculating significance

$m = 980; n = 10,030,834,086; m \times n \approx 10^{13}$

$K = 1.37; \lambda = 0.711$

score	*E*	*P*
39	12	0.99
41	2.9	0.95
42	1.4	0.76
46	0.08	0.08
49	0.01	0.01
55	0.0001	0.0001

Warning about E-values

E-values are calculated based on the query and the database, so they are **not directly comparable between searches**.

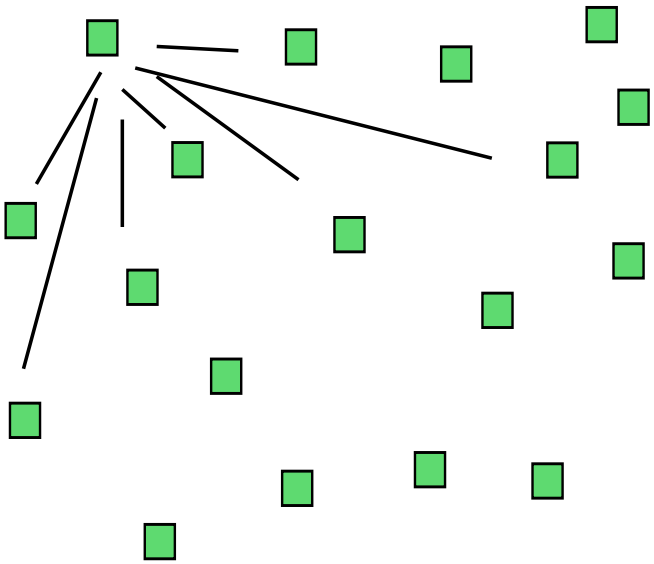
They have **no absolute meaning**, since they depend on database context.

However, bit-scores are normalized by the database parameters, and can be compared between searches.

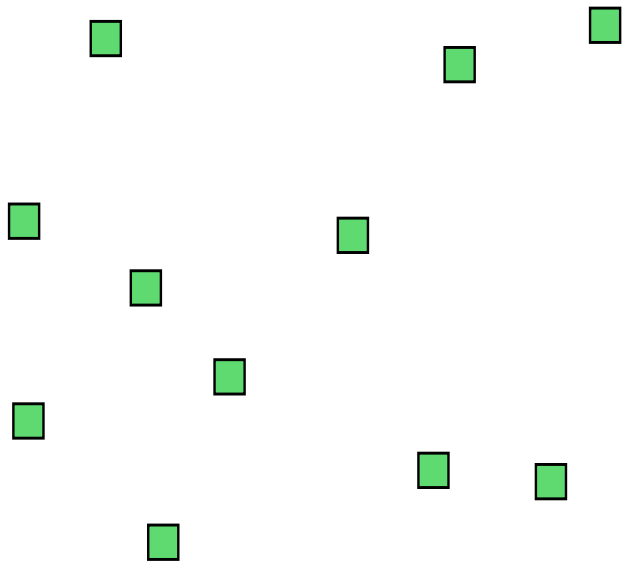
$$S' = (\lambda S - \ln K)/\ln 2$$

Sets of sequences

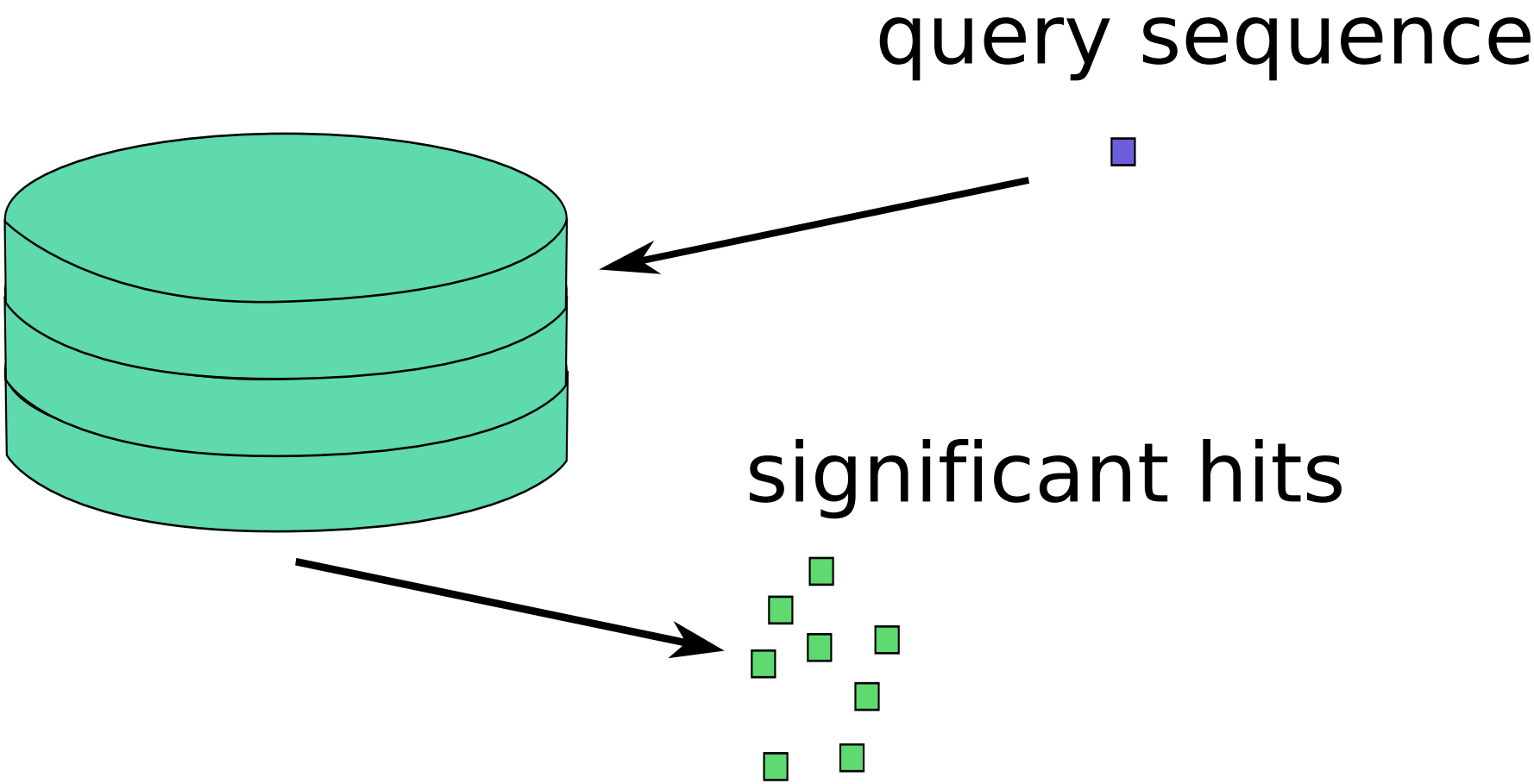
sequences



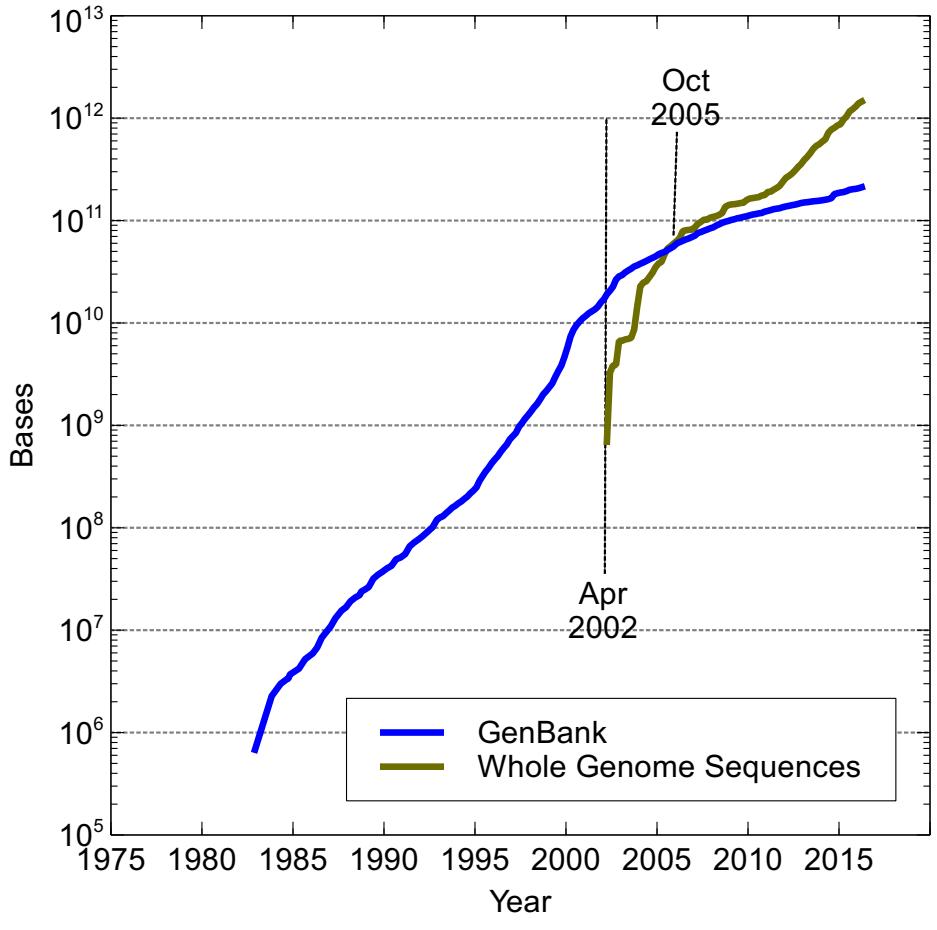
significant hits



From a database



Sequence databases growing exponentially (genomic data even faster)



GSAFDACFDEADRKANTAYKNEQHPNDMFQTYEBLUEBERRYPANCAKESSHLEGQSC

CTFQTYEBLUEBERRYPANCAKESQNPFEEQGPQMKVEVQAFNQDFWDLFRFFPQWHK

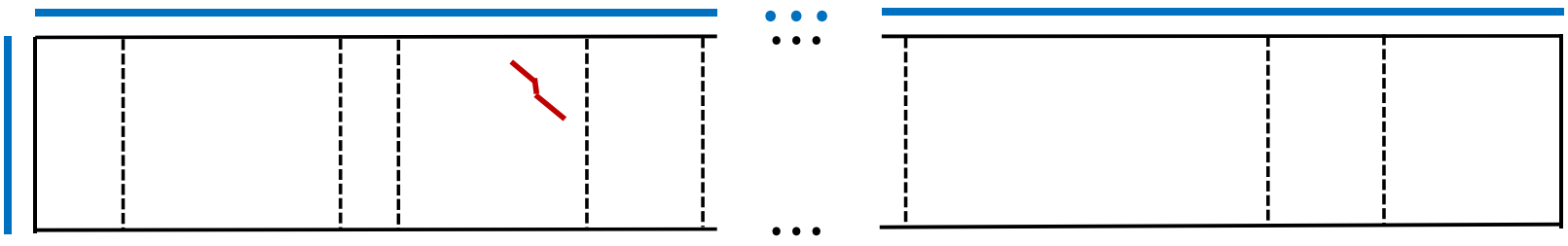
GSAFDACFDEADRKANTAYKNEQHPNDMFQTYEBLUEBERRYPANCAKESSHLEGQSC

CTFQTYEBLUEBERRYPANCAKESQNPFEEQGPQMKVEVQAFNQDFWDLFRFFPQWHK

GSAFDACFDEADRKANTAYKNEQHPNDMFQTYEBLUEBERRYPANCAKESSHLEGQSC

CTFQTYEBLUEBERRYPANCAKESQNPFEEQGPQMKVEVQAFNQDFWDLFRFFPQWHK

Smith-Waterman does not always focus on biologically-relevant regions



BLAST: Basic Local Alignment Search Tool

- Smith-Waterman is often slow for local alignment
  - While getting to the human genome, needed something faster (50-100$\times$)
- BLAST (Altschul et al., 1990, cited over 128,000 times)
  - Heuristic that produces an **approximate** best match (S-W is a guarantee)
    Calculate the high scoring matches instead of the maximum scoring matches (HSP instead of MSP)

BLAST

filter out repeats and low complexity regions (this is different than FASTA)

>gi|195593191|gb|EU940837.1| Zea mays clone 1158441 mRNA sequence

GTTCACATCATCCTGCGGACTGCCTGAGGAGGGGATCACACTGTCCTCTCAGGCTTTCAGTTAGTGCTTA

TGGCGTTCTGTTAGAACCTGTATTTGTATTCCTGCTGGGAGCCTGAGTTCCATTCCGTGCAAAGATAAAT

CATGTGTGACGACACGTTTGCAACAGCATTTATGCTTAAACTGCATTAATGATGATGCGTTGAGCTCC**AA**

AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

ATTTTTTTTTTTTTTTTAAAAAAAAAAAAAAAAAAAAA

BLAST

filter out repeats and low complexity regions

break the sequence into “words” of a length (default for nucleotides is 28, we will look at 4)

```
GTTCACATCATCCTGC

    GTTC
     TTCA
      TCAC
       CACA
        ACAT
         CATC
          ATCA
    ...
```

BLAST

for each word, look at “likely” mutants (based on scoring matrices)

you could call this the word’s sequence “neighborhood”

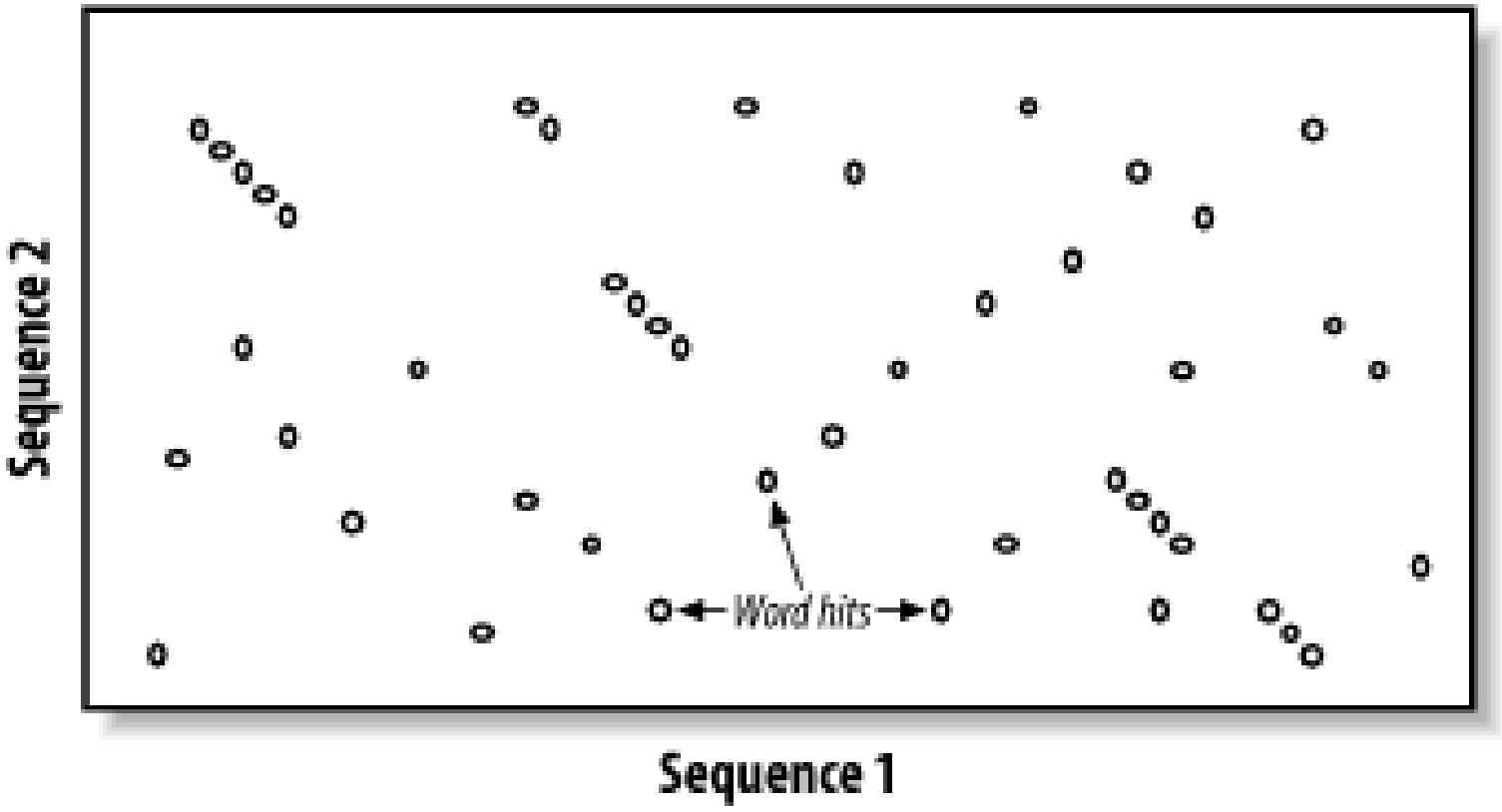
```
GTTCACATCATCCTGC

    GTTC: CTTC,GTTC,GATC...
    TTCA: TTCT,TTGA,TTGT...
    TCAC: AGAC,CCAC,TCTG...
    CACA: ...
    ACAT: ...
    CATC: ...
    ATCA: ...
    …
```

BLAST

organize the words into a form best for searching

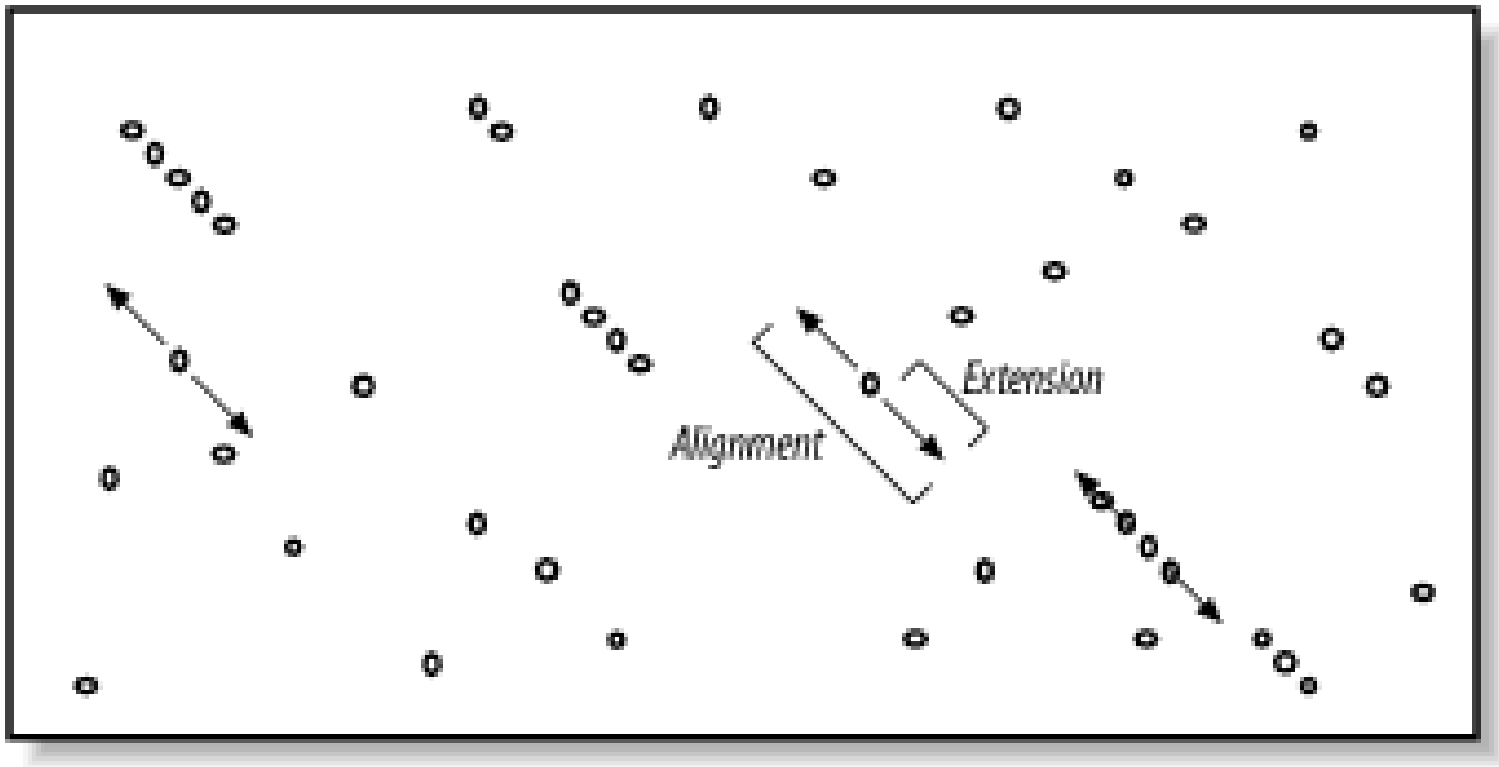
scan the other sequence for words that match



from Korf et al.

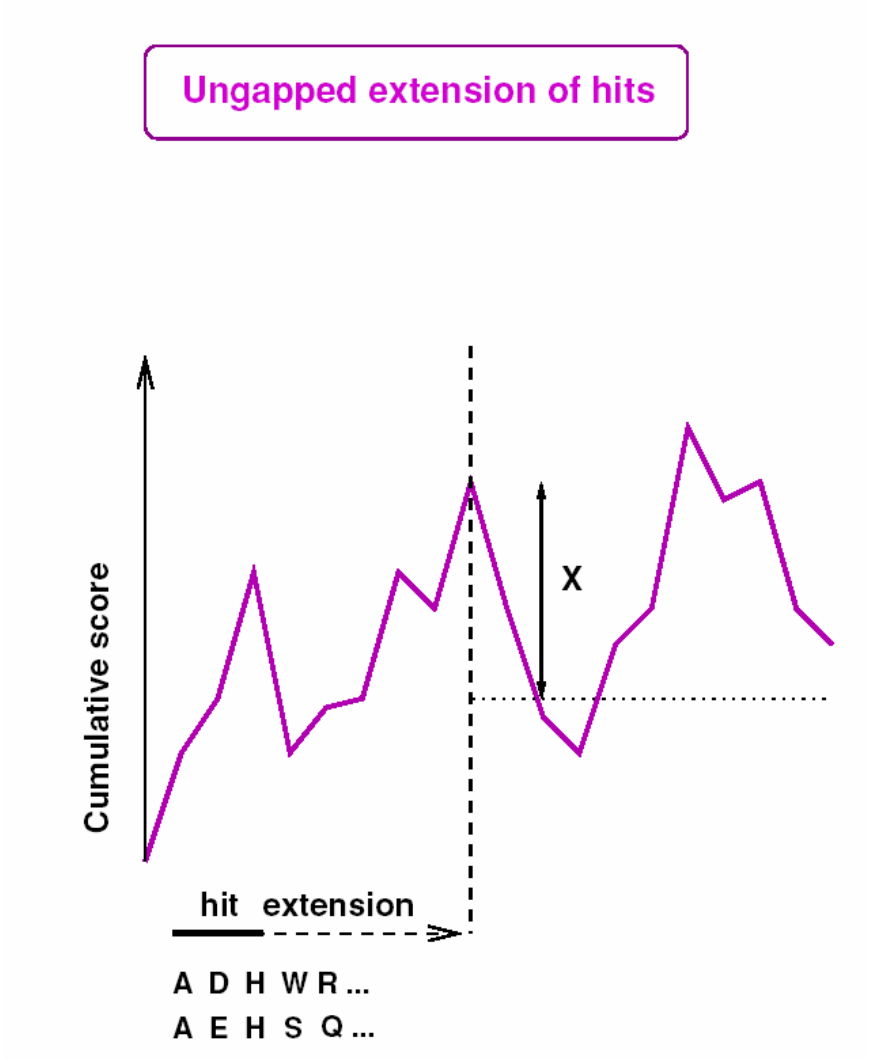
BLAST

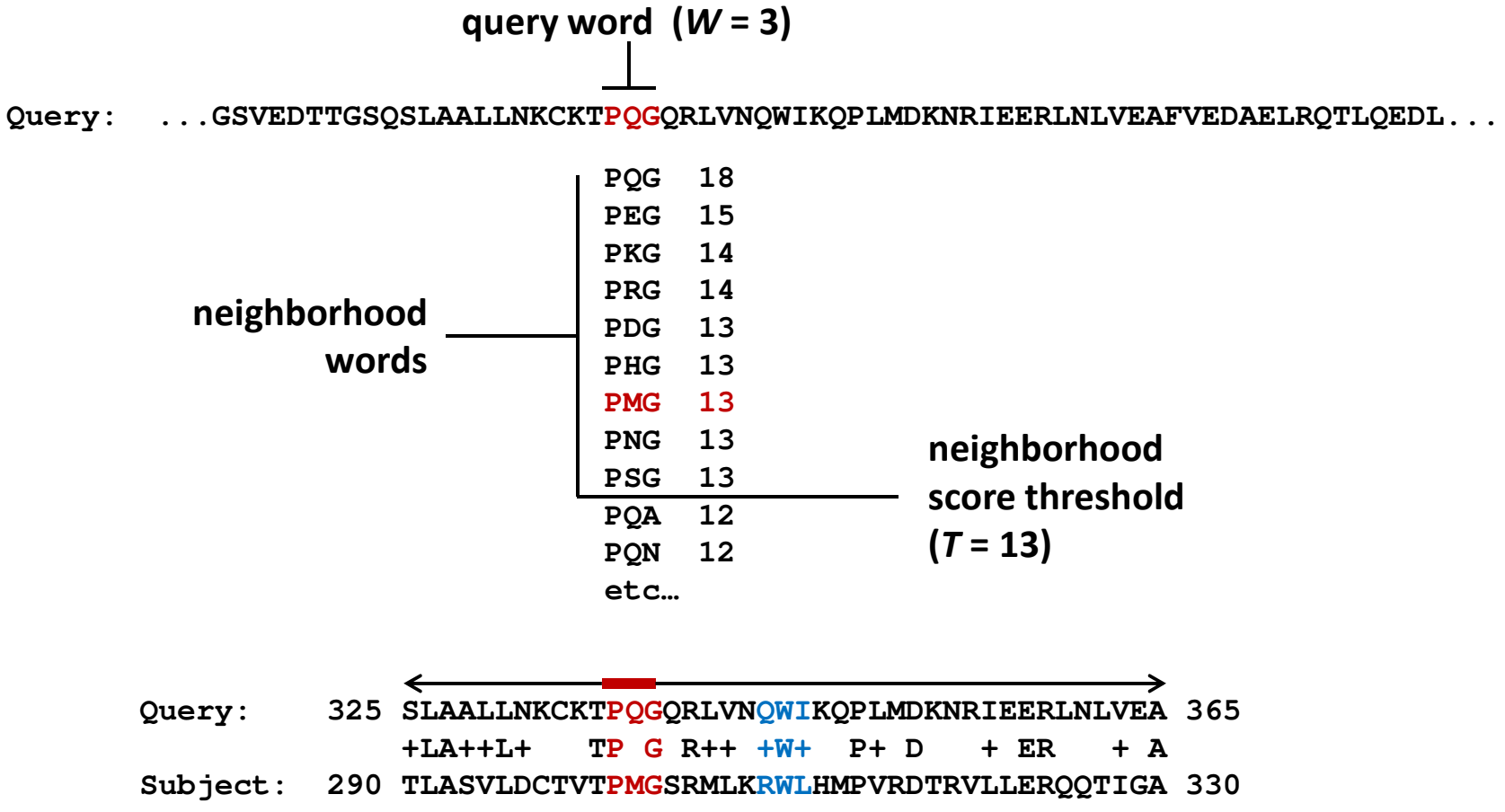
extend these matches in the local neighborhood (these are going to be HSP or high scoring segment pairs)



from Korf et al.

Extension stops when the score decreases past a certain point (X) when compared to the highest score





High-scoring Segment Pair (HSP)

figure from Altschul

BLAST

Calculate *E*-values and return significant results

- Expectation that you would get that alignment by change given the database of sequences

We already talked about these *E*-values and *p*-values with Smith-Waterman significance

BLAST

Because of the speed, BLAST has been used in many different ways

identification of homologs
organism identification
translation (at least the first steps)
putative function

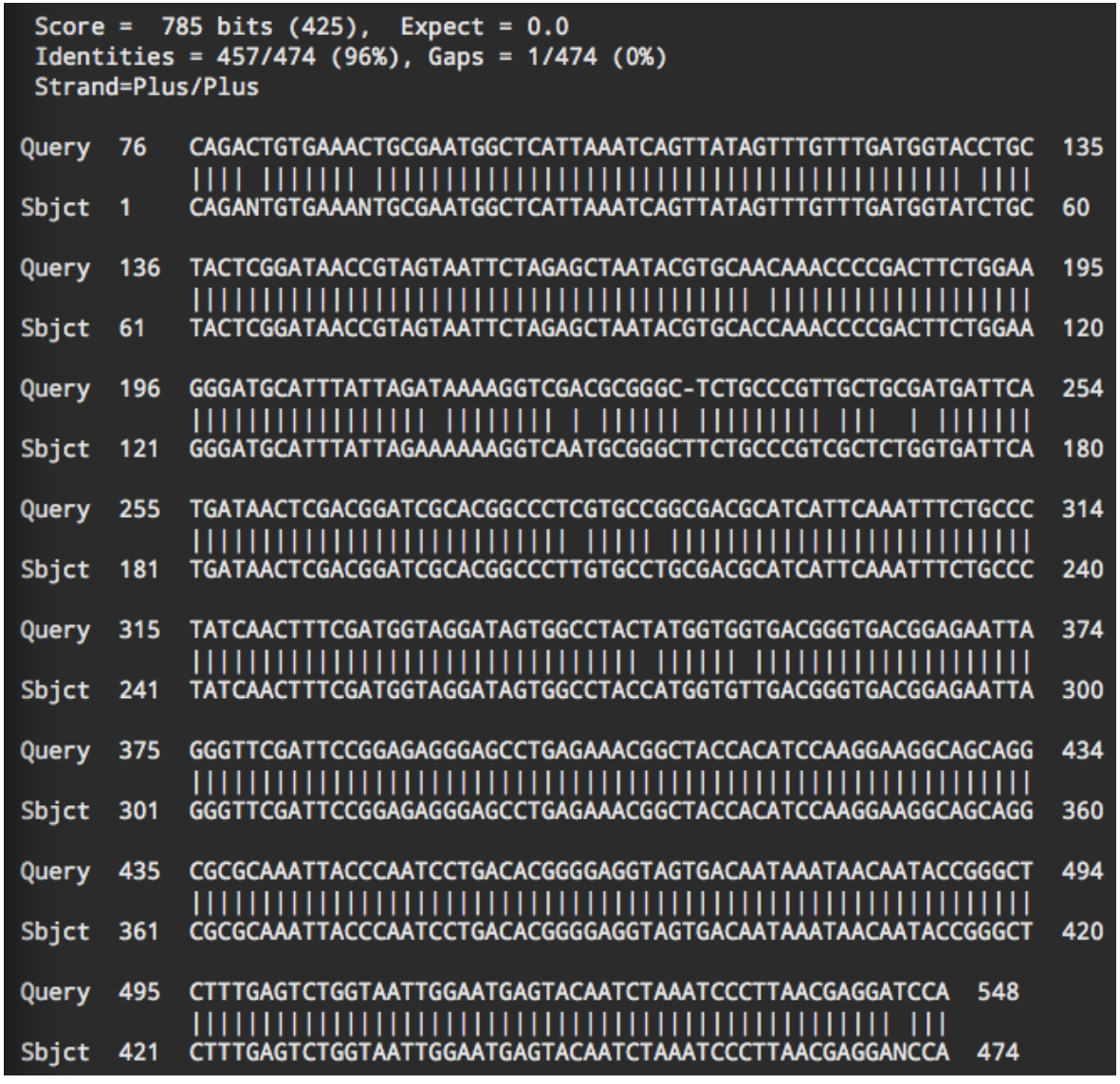
Here, we mainly search for sequences that will have significant matches

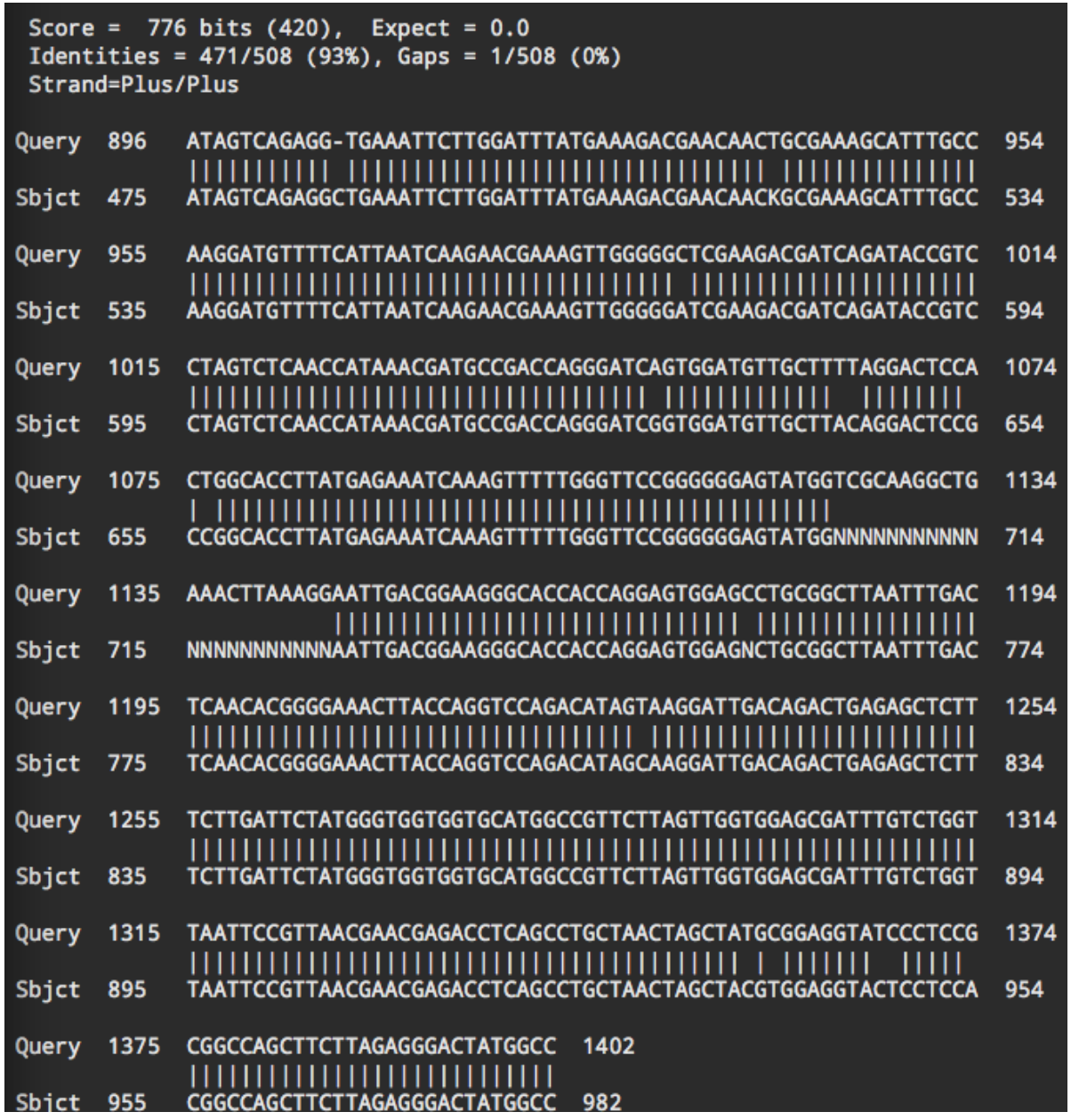
Can search a custom set of sequences or a standard database of sequences

First lets look at results from our own sequences

```
>18S_Abelia
NNNNNNNNNNNNNNNNNNGTAGTCATATGCTTGTCTCAAAGATTAAGCCATGCATGTGTAAGTATGAACTAATTCAGACTGTGAAAC
TGCGAATGGCTCATTAAATCAGTTATAGTTTGTTTGATGGTACCTGCTACTCGGATAACCGTAGTAATTCTAGAGCTAATACGTGCAA
CAAACCCCGACTTCTGGAAGGGATGCATTTATTAGATAAAAGGTCGACGCGGGCTCTGCCCGTTGCTGCGATGATTCATGATAACTCG
ACGGATCGCACGGCCCTCGTGCCGGCGACGCATCATTCAAATTTCTGCCCTATCAACTTTCGATGGTAGGATAGTGGCCTACTATGGT
GGTGACGGGTGACGGAGAATTAGGGTTCGATTCCGGAGAGGGAGCCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCA
AATTACCCAATCCTGACACGGGGAGGTAGTGACAATAAATAACAATACCGGGCTCTTTGAGTCTGGTAATTGGAATGAGTACAATCTA
AATCCCTTAACGAGGATCCATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATATTTAAGTTGTTG
CAGTTAAAAAGCTCGTAGTTGGACTTTGGGTTGGGTCGGCCGGTCCGCCTATCGGTGTGCACCGGTCGTCTCGTCCCTTCTGCCGGCG
ATGCGCTCCTGGCCTTAACTGGTCGGGTCGTGCCTCCGGCGCTGTTACTTTGAAGAAATTAGAGTGCTCAAAGCAAGCCTACGCTCTG
GATACATTAGCATGGGATAACATCATAGGATTTCGGTCCTATTACGTTGGCCTTCGGGATCGGAGTAATGATTAACAGGGACAGTCGG
GGGCATTCGTATTTCATAGTCAGAGGTGAAATTCTTGGATTTATGAAAGACGAACAACTGCGAAAGCATTTGCCAAGGATGTTTTCAT
TAATCAAGAACGAAAGTTGGGGGCTCGAAGACGATCAGATACCGTCCTAGTCTCAACCATAAACGATGCCGACCAGGGATCAGTGGAT
GTTGCTTTTAGGACTCCACTGGCACCTTATGAGAAATCAAAGTTTTTGGGTTCCGGGGGGAGTATGGTCGCAAGGCTGAAACTTAAAG
GAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGGAAACTTACCAGGTCCAGACATAGTAAG
GATTGACAGACTGAGAGCTCTTTCTTGATTCTATGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTC
CGTTAACGAACGAGACCTCAGCCTGCTAACTAGCTATGCGGAGGTATCCCTCCGCGGCCAGCTTCTTAGAGGGACTATGGCCTTTCAG
GCCACGGAAGTTTGAGGCAATAACAGGTCTGTGATGCCCTTAGATGTTCTGGGCCGCACGCGCGCTACACTGATGTATTCAACGAGCC
TATAGCCTTGGCCGACAGGCCCGGGAAATCTTTGAAATTTCATCGTGATGGGGATAGATCATTGCAATTGTTGGTCTTAAACGAAGAA
TTCCTAGTAAGCGCGAGTCATCAGCTCGCGTTGACTACGTCCCTGCCCTTTGTACACACCGCCCGTCGCTCCTACCGATTGAATGGTC
CGGTGAAGTGTTCGGATCGCGGCGACGTGGGCGGTTCGCTGCCGGCGACGTCGCGAGAAGTCCACTGAACCTTATCATTGAGAGGAAG
GAGAGTC

>18S_Acorus
CAGANTGTGAAANTGCGAATGGCTCATTAAATCAGTTATAGTTTGTTTGATGGTATCTGCTACTCGGATAACCGTAGTAATTCTAGAG
CTAATACGTGCACCAAACCCCGACTTCTGGAAGGGATGCATTTATTAGAAAAAAGGTCAATGCGGGCTTCTGCCCGTCGCTCTGGTGA
TTCATGATAACTCGACGGATCGCACGGCCCTTGTGCCTGCGACGCATCATTCAAATTTCTGCCCTATCAACTTTCGATGGTAGGATAG
TGGCCTACCATGGTGTTGACGGGTGACGGAGAATTAGGGTTCGATTCCGGAGAGGGAGCCTGAGAAACGGCTACCACATCCAAGGAAG
GCAGCAGGCGCGCAAATTACCCAATCCTGACACGGGGAGGTAGTGACAATAAATAACAATACCGGGCTCTTTGAGTCTGGTAATTGGA
ATGAGTACAATCTAAATCCCTTAACGAGGANCCAATAGTCAGAGGCTGAAATTCTTGGATTTATGAAAGACGAACAACKGCGAAAGCA
TTTGCCAAGGATGTTTTCATTAATCAAGAACGAAAGTTGGGGGATCGAAGACGATCAGATACCGTCCTAGTCTCAACCATAAACGATG
CCGACCAGGGATCGGTGGATGTTGCTTACAGGACTCCGCCGGCACCTTATGAGAAATCAAAGTTTTTGGGTTCCGGGGGGAGTATGGN
NNNNNNNNNNNNNNNNNNNNNAATTGACGGAAGGGCACCACCAGGAGTGGAGNCTGCGGCTTAATTTGACTCAACACGGGGAAACTTA
CCAGGTCCAGACATAGCAAGGATTGACAGACTGAGAGCTCTTTCTTGATTCTATGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGA
GCGATTTGTCTGGTTAATTCCGTTAACGAACGAGACCTCAGCCTGCTAACTAGCTACGTGGAGGTACTCCTCCACGGCCAGCTTCTTA
GAGGGACTATGGCCGCTTAGGCCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAAACTTTCATCGTGAGTGGGGATGAGMCGTGNGA
AATTGTTGGTCTTCAAGGAGGAGAGTTCCTANTAAKYRYNTGAAGTAATTNCAGSCCNNGTTGACKACKKCCTCTGCTSCWWTGWNNN
NNNNNNNNNNNTCCTACCGATTGAATGGTCCGGTGAAGTGTTCGGATCGCGGCGACAGGGCGGTTSCGCCGGCGACGTTGTGAGAAGT
CCATT
```


Score = 785 bits (425), Expect = 0.0
Identities = 457/474 (96%), Gaps = 1/474 (0%)
Strand=Plus/Plus
Query 76 CAGACTGTGAAACTGCGAATGGCTCATTAAATCAGTTATAGTTTGTTTGATGGTACCTGC 135
Sbjct 1 CAGANTGTGAAANTGCGAATGGCTCATTAAATCAGTTATAGTTTGTTTGATGGTATCTGC 60
Query 136 TACTCGGATAACCGTAGTAATTCTAGAGCTAATACGTGCAACAAACCCCGACTTCTGGAA 195
Sbjct 61 TACTCGGATAACCGTAGTAATTCTAGAGCTAATACGTGCACCAAACCCCGACTTCTGGAA 120
Query 196 GGGATGCATTTATTAGATAAAAGGTCGACGCGGGC-TCTGCCCGTTGCTGCGATGATTCA 254
Sbjct 121 GGGATGCATTTATTAGAAAAAAGGTCAATGCGGGCTTCTGCCCGTCGCTCTGGTGATTCA 180
Query 255 TGATAACTCGACGGATCGCACGGCCCTCGTGCCGGCGACGCATCATTCAAATTTCTGCCC 314
Sbjct 181 TGATAACTCGACGGATCGCACGGCCCTTGTGCCTGCGACGCATCATTCAAATTTCTGCCC 240
Query 315 TATCAACTTTCGATGGTAGGATAGTGGCCTACTATGGTGGTGACGGGTGACGGAGAATTA 374
Sbjct 241 TATCAACTTTCGATGGTAGGATAGTGGCCTACCATGGTGTTGACGGGTGACGGAGAATTA 300
Query 375 GGGTTCGATTCCGGAGAGGGAGCCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGG 434
Sbjct 301 GGGTTCGATTCCGGAGAGGGAGCCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGG 360
Query 435 CGCGCAAATTACCCAATCCTGACACGGGGAGGTAGTGACAATAAATAACAATACCGGGCT 494
Sbjct 361 CGCGCAAATTACCCAATCCTGACACGGGGAGGTAGTGACAATAAATAACAATACCGGGCT 420
Query 495 CTTTGAGTCTGGTAATTGGAATGAGTACAATCTAAATCCCTTAACGAGGATCCA 548
Sbjct 421 CTTTGAGTCTGGTAATTGGAATGAGTACAATCTAAATCCCTTAACGAGGANCCA 474


Score = 776 bits (420), Expect = 0.0
Identities = 471/508 (93%), Gaps = 1/508 (0%)
Strand=Plus/Plus
Query 896 ATAGTCAGAGG-TGAAATTCTTGGATTTATGAAAGACGAACAACTGCGAAAGCATTTGCC 954
Sbjct 475 ATAGTCAGAGGCTGAAATTCTTGGATTTATGAAAGACGAACAACKGCGAAAGCATTTGCC 534
Query 955 AAGGATGTTTTCATTAATCAAGAACGAAAGTTGGGGGCTCGAAGACGATCAGATACCGTC 1014
Sbjct 535 AAGGATGTTTTCATTAATCAAGAACGAAAGTTGGGGGATCGAAGACGATCAGATACCGTC 594
Query 1015 CTAGTCTCAACCATAAACGATGCCGACCAGGGATCAGTGGATGTTGCTTTTAGGACTCCA 1074
Sbjct 595 CTAGTCTCAACCATAAACGATGCCGACCAGGGATCGGTGGATGTTGCTTACAGGACTCCG 654
Query 1075 CTGGCACCTTATGAGAAATCAAAGTTTTTGGGTTCCGGGGGGGAGTATGGTCGCAAGGCTG 1134
Sbjct 655 CCGGCACCTTATGAGAAATCAAAGTTTTTGGGTTCCGGGGGGGAGTATGGNNNNNNNNNNN 714
Query 1135 AAACTTAAAGGAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGAC 1194
Sbjct 715 NNNNNNNNNNNAATTGACGGAAGGGCACCACCAGGAGTGGAGNCTGCGGCTTAATTTGAC 774
Query 1195 TCAACACGGGGAAACTTACCAGGTCCAGACATAGTAAGGATTGACAGACTGAGAGCTCTT 1254
Sbjct 775 TCAACACGGGGAAACTTACCAGGTCCAGACATAGCAAGGATTGACAGACTGAGAGCTCTT 834
Query 1255 TCTTGATTCTATGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGT 1314
Sbjct 835 TCTTGATTCTATGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGT 894
Query 1315 TAATTCCGTTAACGAACGAGACCTCAGCCTGCTAACTAGCTATGCGGAGGTATCCCTCCG 1374
Sbjct 895 TAATTCCGTTAACGAACGAGACCTCAGCCTGCTAACTAGCTACGTGGAGGTACTCCTCCA 954
Query 1375 CGGCCAGCTTCTTAGAGGGACTATGGCC 1402
Sbjct 955 CGGCCAGCTTCTTAGAGGGACTATGGCC 982

```
 Score =  113 bits (61),  Expect = 1e-28
 Identities = 76/83 (92%), Gaps = 3/83 (4%)
 Strand=Plus/Plus

Query  1654  TCCTACCGATTGAATGGTCCGGTGAAGTGTTCGGATCGCGGCGACGTGGGCGGTTCGCTG  1713
             ||||||||||||||||||||||||||||||||||||||||||||||||  ||||||||  | |
Sbjct  1244  TCCTACCGATTGAATGGTCCGGTGAAGTGTTCGGATCGCGGCGACA-GGGCGGTTS-C-G  1300

Query  1714  CCGGCGACGTCGCGAGAAGTCCA  1736
             |||||||||| | ||||||||||
Sbjct  1301  CCGGCGACGTTGTGAGAAGTCCA  1323



Lambda      K        H
    1.33    0.621     1.12

Gapped
Lambda      K        H
    1.28    0.460    0.850

Effective search space used: 2298183



Matrix: blastn matrix 1 -2
Gap Penalties: Existence: 0, Extension: 2.5
```

Against a set of sequences

- BLAST is not limited to pairwise comparisons
  - in fact, pairwise is definitely not the default means of interaction
- We can compare the same sequence to a set of sequences
- In this case:
  - one query sequence (sequence of interest, our own)
    3167 subject sequences (a set of sequences from a bunch of other genes)
    included the query sequence in the set of subject sequences
    otherwise, just a standard BLAST

query	subject	%ident	length	#mismat	#gp_open	que_sta	que_end	sub_sta	sub_end	evalue	score
18S_Abelia	18S_Abelia	100	1748	0	0	20	1767	20	1767	0	3229
18S_Abelia	18S_Acorus	96.41	474	16	1	76	548	1	474	0	785
18S_Abelia	18S_Acorus	92.72	508	36	1	896	1402	475	982	0	776
18S_Abelia	18S_Acorus	91.57	83	4	3	1654	1736	1244	1323	1.00E-28	113
18S_Abelia	18S_Aextoxicon	98.22	1741	31	0	24	1764	1	1741	0	3044
18S_Abelia	18S_Agave	97.3	1742	44	3	24	1763	1	1741	0	2955
18S_Abelia	18S_Ailanthus	98.09	1728	33	0	37	1764	1	1728	0	3009
18S_Abelia	18S_Alisma	95.91	1734	69	2	20	1751	10	1743	0	2808
18S_Abelia	18S_Alnus	97.15	1508	41	2	204	1710	3	1509	0	2545
18S_Abelia	18S_Amborella	96.04	1742	66	3	22	1761	1	1741	0	2832
18S_Abelia	18S_Angelica	98.51	1749	25	1	20	1767	17	1765	0	3085
18S_Abelia	18S_Anisophyllea	96.93	1106	33	1	24	1129	1	1105	0	1853
18S_Abelia	18S_Anisophyllea	97.7	609	13	1	1160	1767	1136	1744	0	1046
18S_Abelia	18S_Anisoptera	97.2	1747	46	3	23	1767	1	1746	0	2953
18S_Abelia	18S_Annona	95.23	1069	40	11	74	1141	39	1097	0	1681
18S_Abelia	18S_Aphanopetalum	97.71	1750	36	4	20	1767	2	1749	0	3009
18S_Abelia	18S_Arabidopsis	97.04	1046	29	2	723	1767	453	1497	0	1759
18S_Abelia	18S_Arabidopsis	96.4	444	13	3	20	462	20	461	0	728
18S_Abelia	18S_Aristolochia	93.32	449	16	5	91	539	1	435	0	652

the BLAST of the sequence against itself starts at base 20

any guesses why?

```
>18S_Abelia
NNNNNNNNNNNNNNNNNNGTAGTCATATGCTTGTCTCAAAGATTAAGCCATGCATGTGTAAGTATGAACTAATTCAGACTGTGAAAC
TGCGAATGGCTCATTAAATCAGTTATAGTTTGTTTGATGGTACCTGCTACTCGGATAACCGTAGTAATTCTAGAGCTAATACGTGCAA
CAAACCCCGACTTCTGGAAGGGATGCATTTATTAGATAAAAGGTCGACGCGGGCTCTGCCCGTTGCTGCGATGATTCATGATAACTCG
ACGGATCGCACGGCCCTCGTGCCGGCGACGCATCATTCAAATTTCTGCCCTATCAACTTTCGATGGTAGGATAGTGGCCTACTATGGT
GGTGACGGGTGACGGAGAATTAGGGTTCGATTCCGGAGAGGGAGCCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGCA
AATTACCCAATCCTGACACGGGGAGGTAGTGACAATAAATAACAATACCGGGCTCTTTGAGTCTGGTAATTGGAATGAGTACAATCTA
AATCCCTTAACGAGGATCCATTGGAGGGCAAGTCTGGTGCCAGCAGCCGCGGTAATTCCAGCTCCAATAGCGTATATTTAAGTTGTTG
CAGTTAAAAAGCTCGTAGTTGGACTTTGGGTTGGGTCGGCCGGTCCGCCTATCGGTGTGCACCGGTCGTCTCGTCCCTTCTGCCGGCG
ATGCGCTCCTGGCCTTAACTGGTCGGGTCGTGCCTCCGGCGCTGTTACTTTGAAGAAATTAGAGTGCTCAAAGCAAGCCTACGCTCTG
GATACATTAGCATGGGATAACATCATAGGATTTCGGTCCTATTACGTTGGCCTTCGGGATCGGAGTAATGATTAACAGGGACAGTCGG
GGGCATTCGTATTTCATAGTCAGAGGTGAAATTCTTGGATTTATGAAAGACGAACAACTGCGAAAGCATTTGCCAAGGATGTTTTCAT
TAATCAAGAACGAAAGTTGGGGGCTCGAAGACGATCAGATACCGTCCTAGTCTCAACCATAAACGATGCCGACCAGGGATCAGTGGAT
GTTGCTTTTAGGACTCCACTGGCACCTTATGAGAAATCAAAGTTTTTGGGTTCCGGGGGGAGTATGGTCGCAAGGCTGAAACTTAAAG
GAATTGACGGAAGGGCACCACCAGGAGTGGAGCCTGCGGCTTAATTTGACTCAACACGGGGAAACTTACCAGGTCCAGACATAGTAAG
GATTGACAGACTGAGAGCTCTTTCTTGATTCTATGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGAGCGATTTGTCTGGTTAATTC
CGTTAACGAACGAGACCTCAGCCTGCTAACTAGCTATGCGGAGGTATCCCTCCGCGGCCAGCTTCTTAGAGGGACTATGGCCTTTCAG
GCCACGGAAGTTTGAGGCAATAACAGGTCTGTGATGCCCTTAGATGTTCTGGGCCGCACGCGCGCTACACTGATGTATTCAACGAGCC
TATAGCCTTGGCCGACAGGCCCGGGAAATCTTTGAAATTTCATCGTGATGGGGATAGATCATTGCAATTGTTGGTCTTAAACGAAGAA
TTCCTAGTAAGCGCGAGTCATCAGCTCGCGTTGACTACGTCCCTGCCCTTTGTACACACCGCCCGTCGCTCCTACCGATTGAATGGTC
CGGTGAAGTGTTCGGATCGCGGCGACGTGGGCGGTTCGCTGCCGGCGACGTCGCGAGAAGTCCACTGAACCTTATCATTGAGAGGAAG
GAGAGTC
```

BLAST

filter out repeats and less complex regions (this is different than FASTA)

>gi|195593191|gb|EU940837.1| Zea mays clone 1158441 mRNA sequence

GTTCACATCATCCTGCGGACTGCCTGAGGAGGGGATCACACTGTCCTCTCAGGCTTTCAGTTAGTGCTTA

TGGCGTTCTGTTAGAACCTGTATTTGTATTCCTGCTGGGAGCCTGAGTTCCATTCCGTGCAAAGATAAAT

CATGTGTGACGACACGTTTGCAACAGCATTTATGCTTAAACTGCATTAATGATGATGCGTTGAGCTCC**AA**

AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

ATTTTTTTTTTTTTTTTAAAAAAAAAAAAAAAAAAAAA

BLAST databases

We can also search against a database

NR: non-redundant amino acid sequences
many many model organisms
BCT: bacterial sequences
ENV: environmental sequences
EST: expressed sequence tags
GSS: genome survey sequences
HTC: high throughput genomic sequencing
INV: invertebrate sequences
MAM: “other” mammal sequences
PAT: patent sequence
PLN: plant sequences
ROD: rodent sequences
PRI: primate sequences
VR: viral sequences
VRT: vertebrate sequences

There are *more* (look for the ncbi ftp)

BLAST programs

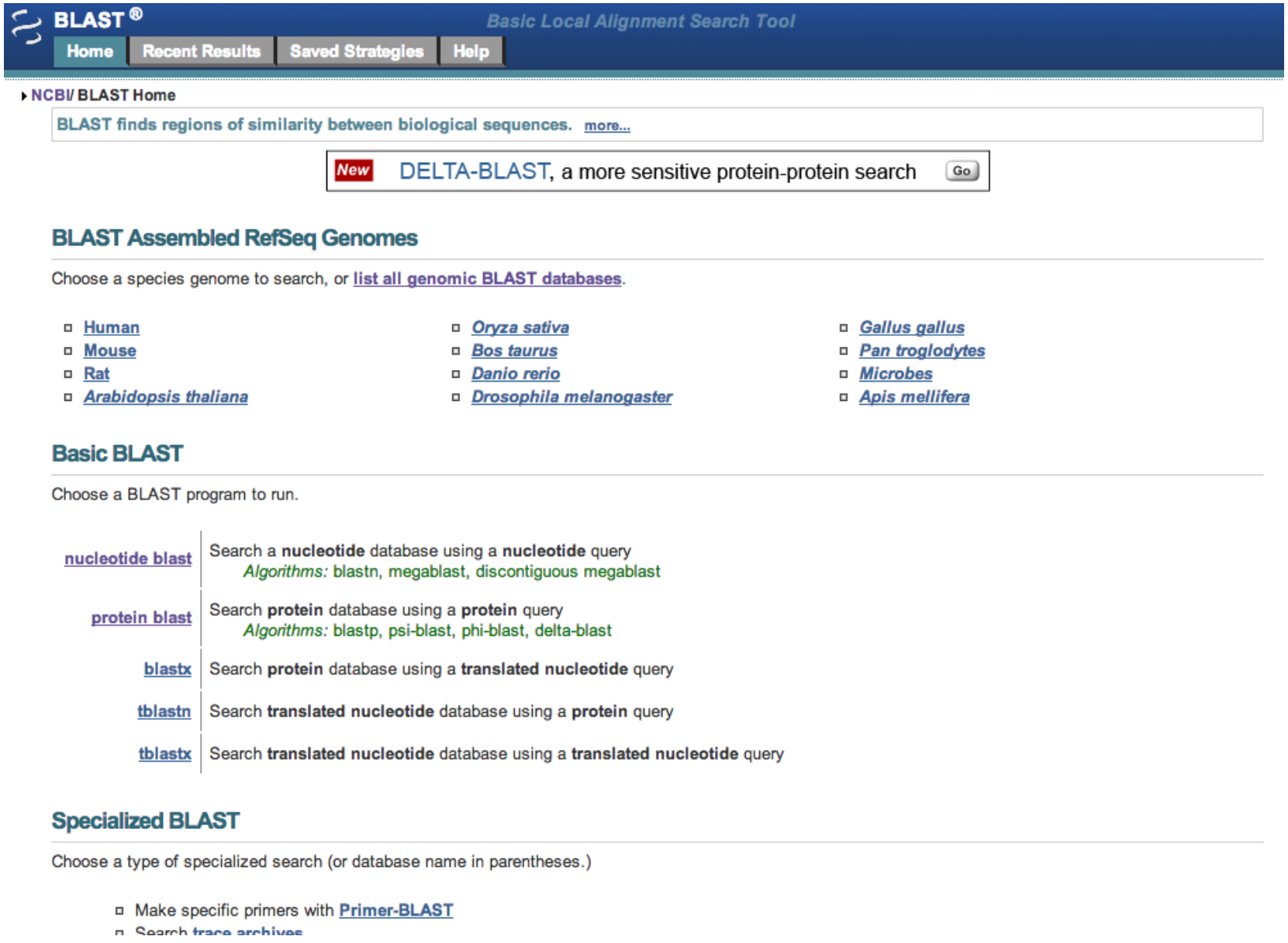
BLAST has different programs

blastn: nucleotide BLAST to other nucleotides
blastp: protein BLAST to protein sequences
blastx: translated nucleotides searching against a protein database
tblastn: proteins searching against translated nucleotide database
tblastx: translated nucleotides searching against translated nucleotide database

There are many other specialized BLAST variants:

conserved domains
vector screening
MegaBLAST - essentially identical sequences
many specialized versions are just specific parameterizations of regular BLAST searches

Web BLAST example



BLAST®

Basic Local Alignment Search Tool

Home | Recent Results | Saved Strategies | Help

NCBI/ BLAST/ blastn suite

Standard Nucleotide BLAST

blastn | blastp | blastx | tblastn | tblastx

Enter Query Sequence

BLASTN programs search nucleotide databases using a nucleotide query. more...

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

Query subrange

From

To

Or, upload file Choose File No file chosen

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database Human genomic + transcript | Mouse genomic + transcript | Others (nr etc.):

Nucleotide collection (nr/nt)

Organism Optional

Enter organism name or id--completions will be suggested Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Optional

Models (XM/XP) Uncultured/environmental sample sequences

Entrez Query Optional

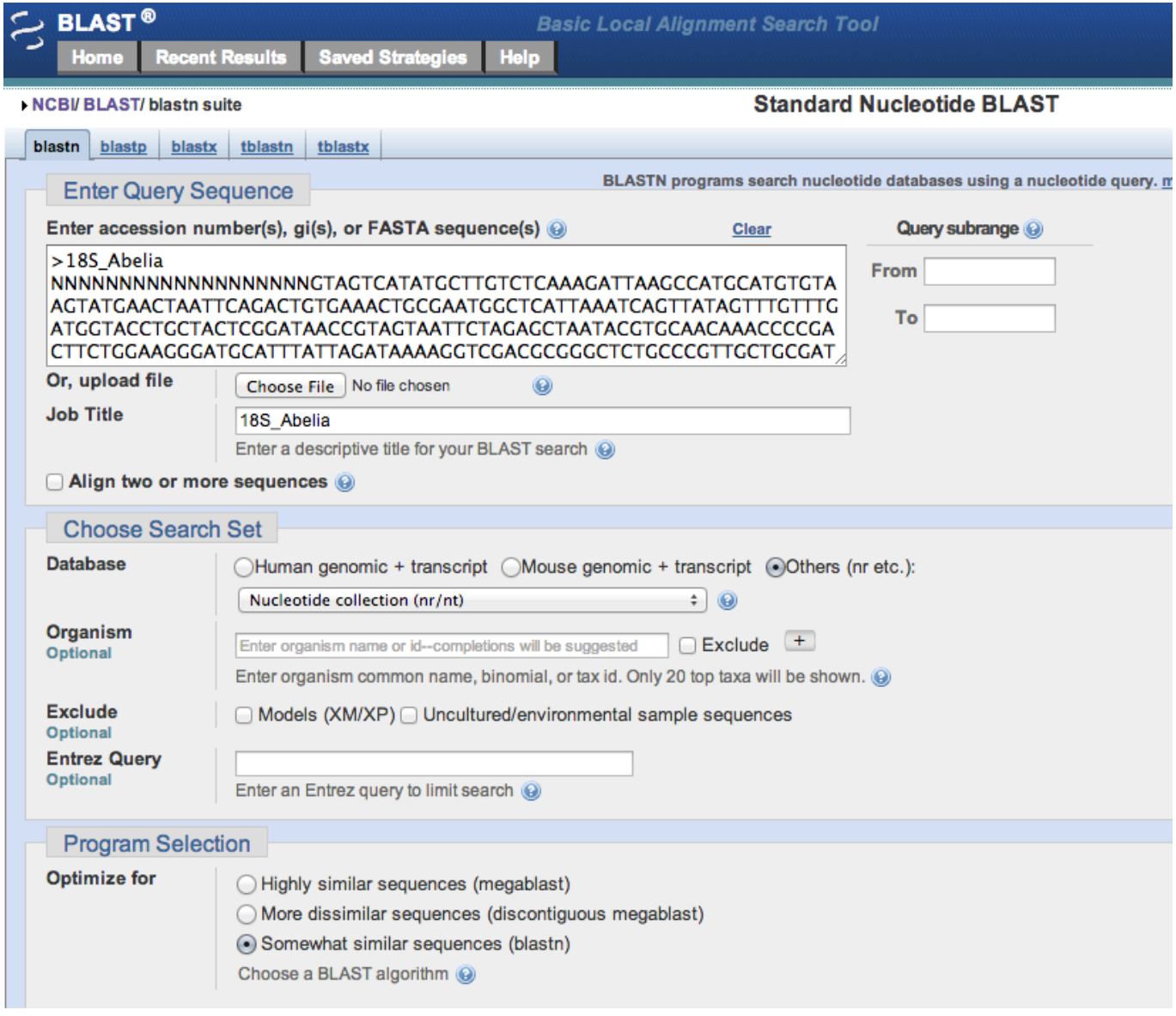
Enter an Entrez query to limit search

Program Selection

Optimize for

- Highly similar sequences (megablast)
- More dissimilar sequences (discontiguous megablast)
- Somewhat similar sequences (blastn)

Choose a BLAST algorithm


BLAST®
Basic Local Alignment Search Tool
Home
Recent Results
Saved Strategies
Help
NCBI/ BLAST/ blastn suite
Standard Nucleotide BLAST
blastn
blastp
blastx
tblastn
tblastx
BLASTN programs search nucleotide databases using a nucleotide query.
Enter Query Sequence
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
>18S_Abelia
NNNNNNNNNNNNNNNNNNGTAGTCATATGCTTGTCTCAAAGATTAAGCCATGCATGTGTA
AGTATGAACTAATTCAGACTGTGAAACTGCGAATGGCTCATTAAATCAGTTATAGTTTGTTTG
ATGGTACCTGCTACTCGGATAACCGTAGTAATTCTAGAGCTAATACGTGCAACAAACCCCGA
CTTCTGGAAGGGATGCATTTATTAGATAAAAGGTCGACGCGGGCTCTGCCCGTTGCTGCGAT
Query subrange
From
To
Or, upload file
Choose File
No file chosen
Job Title
18S_Abelia
Enter a descriptive title for your BLAST search
Align two or more sequences
Choose Search Set
Database
Human genomic + transcript
Mouse genomic + transcript
Others (nr etc.):
Nucleotide collection (nr/nt)
Organism
Optional
Enter organism name or id--completions will be suggested
Exclude
+
Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.
Exclude
Optional
Models (XM/XP)
Uncultured/environmental sample sequences
Entrez Query
Optional
Enter an Entrez query to limit search
Program Selection
Optimize for
Highly similar sequences (megablast)
More dissimilar sequences (discontiguous megablast)
Somewhat similar sequences (blastn)
Choose a BLAST algorithm

BLAST

Search database Nucleotide collection (nr/nt) using Blastn (Optimize for somewhat similar sequences)

Show results in a new window

Algorithm parameters

General Parameters

Max target sequences: 100

Select the maximum number of aligned sequences to display

Short queries: Automatically adjust parameters for short input sequences

Expect threshold: 10

Word size: 11

Max matches in a query range: 0

Scoring Parameters

Match/Mismatch Scores: 2,-3

Gap Costs: Existence: 5 Extension: 2

Filters and Masking

Filter: Low complexity regions

Species-specific repeats for: Homo sapiens (Human)

Mask: Mask for lookup table only

Mask lower case letters

BLAST

Search database Nucleotide collection (nr/nt) using Blastn (Optimize for somewhat similar sequences)

Show results in a new window

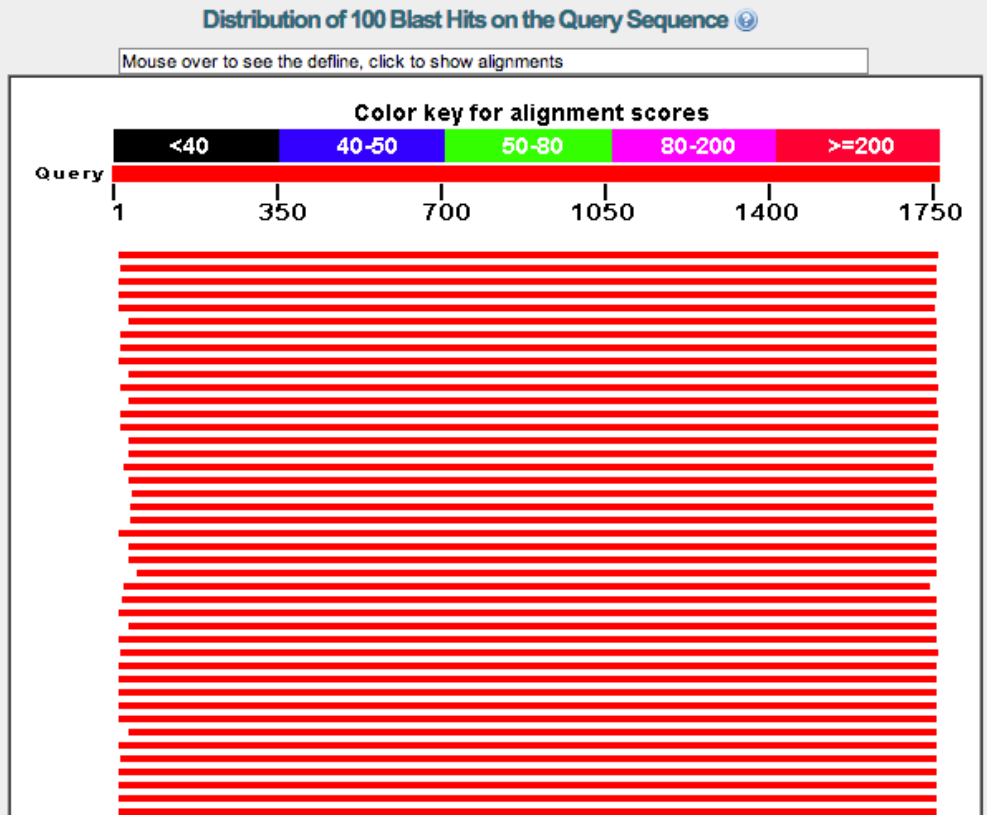
18S_Abelia

Query ID	lcl|27385
Description	18S_Abelia
Molecule type	nucleic acid
Query Length	1767

Database Name	nr
Description	Nucleotide collection (nt)
Program	BLASTN 2.2.28+ ▷ Citation

Other reports: ▷ Search Summary [Taxonomy reports] [Distance tree of results]

Graphic Summary



Descriptions

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenBank Graphics Distance tree of results

Description	Max score	Total score	Query cover	E value	Max ident	Accession
Abelia triflora 18S rRNA gene	3153	3153	98%	0.0	100%	AJ236004.1
Scabiosa sp. Albach 39 18S rRNA gene	3126	3126	98%	0.0	99%	AJ236006.1
Dipsacus asperoides isolate JianShi 18S ribosomal RNA gene, partial sequence	3124	3124	98%	0.0	99%	GU166826.1
Dipsacus asperoides isolate EnShi YuTangBa 18S ribosomal RNA gene, partial sequence	3124	3124	98%	0.0	99%	GU166824.1
Dipsacus asperoides isolate BaDong 18S ribosomal RNA gene, partial sequence	3113	3113	98%	0.0	99%	GQ906564.1
Dipelta yunnanensis 18S ribosomal RNA gene, partial sequence	3103	3103	97%	0.0	99%	GQ983567.1
Viburnum acerifolia 18S rRNA gene	3101	3101	98%	0.0	99%	AJ236007.1
Sambucus ebulus 18S rRNA gene	3099	3099	98%	0.0	99%	AJ236005.1
Lonicera maackii 18S ribosomal RNA gene, complete sequence	3099	3099	98%	0.0	99%	U66701.1
Patrinia triloba 18S ribosomal RNA gene, partial sequence	3097	3097	97%	0.0	99%	GQ983572.1
Valeriana officinalis 18S rRNA gene	3095	3095	98%	0.0	99%	AJ236003.1
Triplostegia glandulifera 18S ribosomal RNA gene, partial sequence	3088	3088	97%	0.0	99%	GQ983577.1
Griselinia lucida 18S ribosomal RNA gene, complete sequence	3088	3088	98%	0.0	99%	AF206922.1
Griselinia littoralis 18S rRNA gene	3088	3088	98%	0.0	99%	AJ236000.1
Morina longifolia 18S ribosomal RNA gene, partial sequence	3085	3085	97%	0.0	99%	GQ983569.1
Dipsacus asperoides isolate EnShi ShuangHe 18S ribosomal RNA gene, partial sequence	3083	3083	97%	0.0	99%	GU166825.1
Dipsacus sp. Jansen 931 18S ribosomal RNA gene, partial sequence	3081	3081	97%	0.0	99%	U43150.1
Diervilla sessilifolia 18S ribosomal RNA gene, partial sequence	3079	3079	97%	0.0	99%	GQ983566.1

Alignments

Download GenBank Graphics ▼ Next ▲ Previous Descriptions

Abelia triflora 18S rRNA gene

Sequence ID: emb|AJ236004.1| **Length:** 1767 **Number of Matches:** 1

Related Information

Range 1: 20 to 1767 GenBank Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Identities	Gaps	Strand
3153 bits(3496)	0.0	1748/1748(100%)	0/1748(0%)	Plus/Plus

```
Query  20    GTAGTCATATGCTTGTCTCAAAGATTAAGCCATGCATGTGTAAGTATGAACTAATTCAGA  79
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  20    GTAGTCATATGCTTGTCTCAAAGATTAAGCCATGCATGTGTAAGTATGAACTAATTCAGA  79

Query  80    CTGTGAAACTGCGAATGGCTCATTAAATCAGTTATAGTTTGTTTGATGGTACCTGCTACT  139
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  80    CTGTGAAACTGCGAATGGCTCATTAAATCAGTTATAGTTTGTTTGATGGTACCTGCTACT  139

Query  140   CGGATAACCGTAGTAATTCTAGAGCTAATACGTGCAACAAACCCCGACTTCTGGAAGGGA  199
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  140   CGGATAACCGTAGTAATTCTAGAGCTAATACGTGCAACAAACCCCGACTTCTGGAAGGGA  199

Query  200   TGCATTTATTAGATAAAAGGTCGACGCGGGCTCTGCCCGTTGCTGCGATGATTCATGATA  259
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  200   TGCATTTATTAGATAAAAGGTCGACGCGGGCTCTGCCCGTTGCTGCGATGATTCATGATA  259

Query  260   ACTCGACGGATCGCACGGCCCTCGTGCCGGCGACGCATCATTCAAATTTCTGCCCTATCA  319
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  260   ACTCGACGGATCGCACGGCCCTCGTGCCGGCGACGCATCATTCAAATTTCTGCCCTATCA  319

Query  320   ACTTTCGATGGTAGGATAGTGGCCTACTATGGTGGTGACGGGTGACGGAGAATTAGGGTT  379
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  320   ACTTTCGATGGTAGGATAGTGGCCTACTATGGTGGTGACGGGTGACGGAGAATTAGGGTT  379

Query  380   CGATTCCGGAGAGGGAGCCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGC  439
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  380   CGATTCCGGAGAGGGAGCCTGAGAAACGGCTACCACATCCAAGGAAGGCAGCAGGCGCGC  439

Query  440   AAATTACCCAATCCTGACACGGGGAGGTAGTGACAATAAATAACAATACCGGGCTCTTTG  499
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  440   AAATTACCCAATCCTGACACGGGGAGGTAGTGACAATAAATAACAATACCGGGCTCTTTG  499
```

Other reports: ▽ Search Summary [Taxonomy reports] [Distance tree of results]

Search Parameters	
Program	blastn
Word size	11
Expect value	10
Hitlist size	100
Match/Mismatch scores	2,-3
Gapcosts	5,2
Low Complexity Filter	Yes
Filter string	L;m;
Genetic Code	1

Database	
Posted date	Apr 22, 2013 2:26 PM
Number of letters	48,635,782,348
Number of sequences	25,407,946
Entrez query	none

Karlin-Altschul statistics		
Lambda	0.633731	0.625
K	0.408146	0.41
H	0.912438	0.78

Results Statistics	
Length adjustment	37
Effective length of query	1730
Effective length of database	47695688346
Effective search space	82513540838580
Effective search space used	82513540838580

DNA *vs.* Protein

Should you use blastn or blastp?

DNA *vs.* Protein

Should you use blastn or blastp?

Four potential nucleotides $\{A, C, G, T\}$ and therefore four potential states

There are 22 amino acids states (including stops)

blastp should be more sensitive than blastn (larger state space → lower chance of a random hit)

If sequences are highly similar, DNA works well

If no translated sequences available, DNA is required

- intergenic spacers
- RNA genes

Is homology a quantitative measure?

Can anything be “90% homologous”?

Homology is a boolean term (with quantitative support measures)

nothing is “90% homologous”

things are either homologous or not
there are no “degrees” of homology
there may be a degree of your support for homology

statistical significance depends on the size of the alignments and the database

E-value increases as database gets bigger

- more chance for a random hit

E-value decreases as alignments get longer

- more significant the longer the alignment

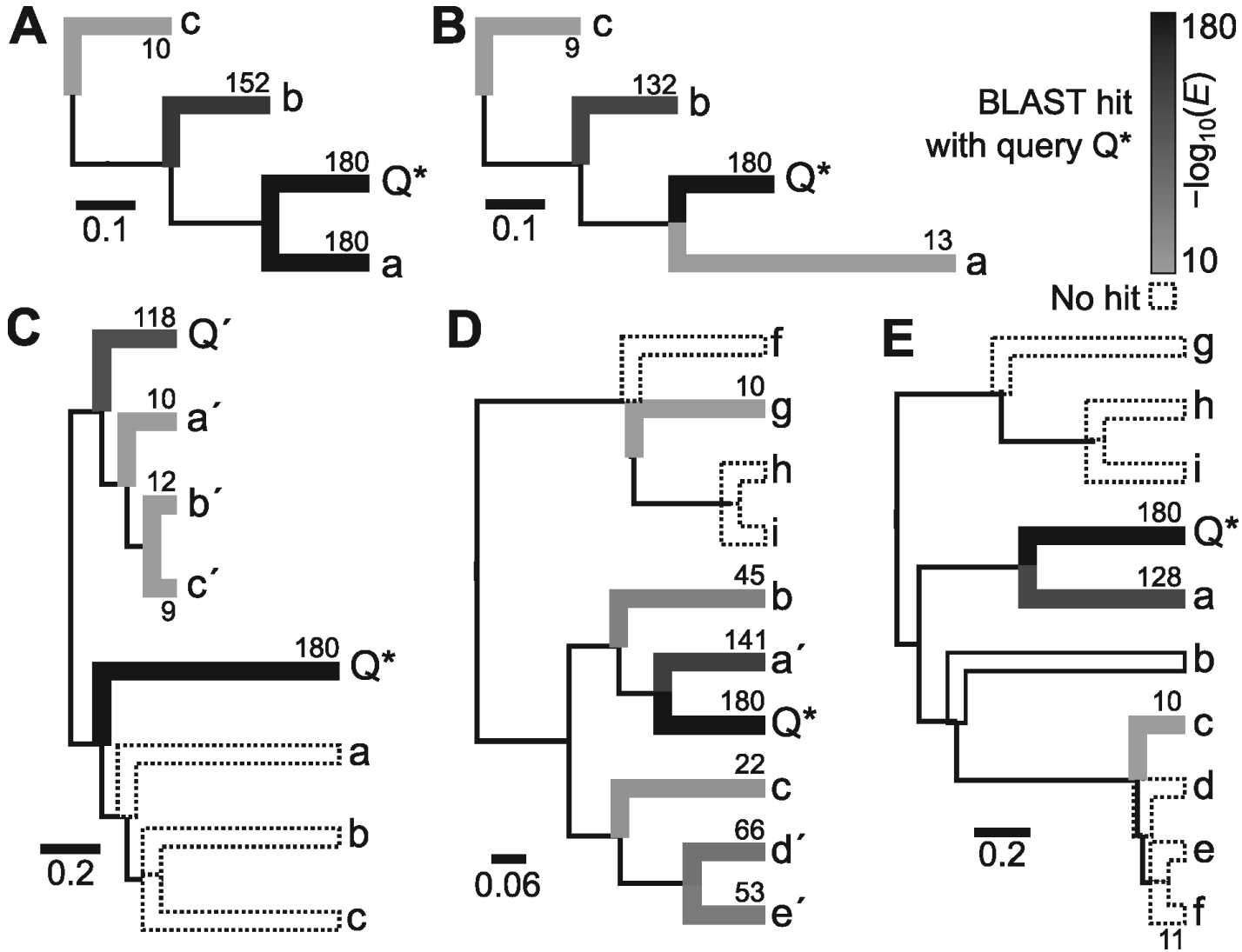
Sequence similarity can suggest homology

(1) significant alignment over (2) the majority of the length of both sequences → strongly suggests homology

homologous sequences do not always produce significant alignments (!)

regions with low complexity (but that are not cleaned out by initial steps in BLAST) can produce significant alignments with virtually no homology

Sequence similarity can suggest homology



Pease and Smith. 2017. *Briefings in Bioinformatics.*

So what are the “rules” for determining homology?

(There are no easy or standarized rules)

Nucleotides

- Some have suggested that sequence identity > 70% is the standard for homology
  E-values of 10^{-6} or less = nope

Proteins

- sequence identity > 25% has been suggested
  E-values of 10^{-3} or less = hmm... nope

You must verify and thoroughly explore your own dataset.

In a high-throughput large-scale analysis, there will be a (inescapable) margin of error

So what are the “rules” for determining homology?

(There are no easy or standarized rules)

Nucleotides

- Some have suggested that sequence identity > 70% is the standard for homology
  E-values of 10^{-6} or less = nope

Proteins

- sequence identity > 25% has been suggested
  E-values of 10^{-3} or less = hmm... nope

You must verify and thoroughly explore your own dataset.

In a high-throughput large-scale analysis, there will be a (inescapable) margin of error

Conclusions and Questions

Can pairwise sequence alignment address your question?

- Are there any homologous sequences?
- Are a set of sequences I have homologous?

What questions are not approachable?

- What are the **relationships** among these sequences?
- Is there shared **function** with this sequence?

Which database should you search?

Which program should you run?

- When possible, it is best to search protein databases
- Use NR and general GenBank for exploration or specific queries,
- Best to narrow down to a smaller database, if possible

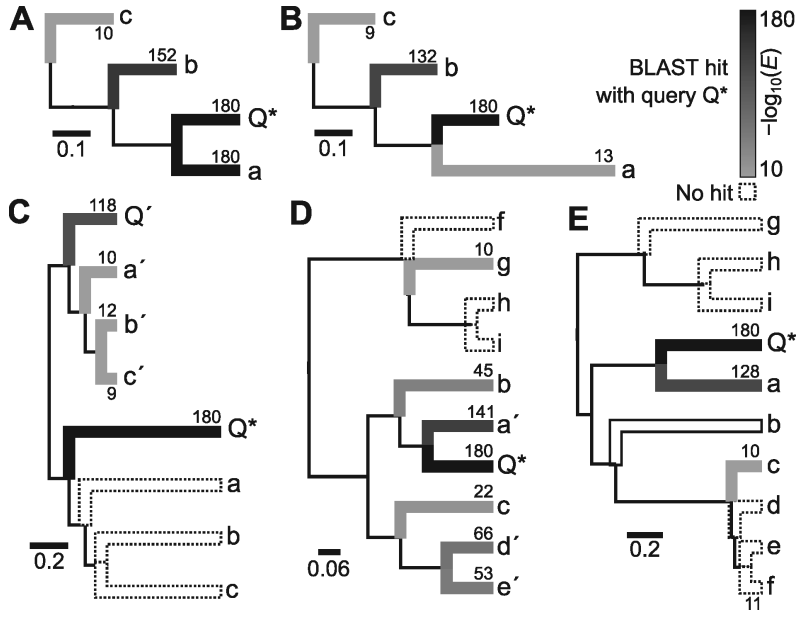
As you assemble your sequence sets from databases:

Be critical and skeptical of sequences!

Watch our for repetitive regions, paralogs, or fragmentary sequences.

Recursively search for matches using your growing pool of sequences.

Keep track of your ids as you go and automate as much as you can (manual entry leads to manual error).



Keep careful records and automate, automate, automate.

Keep track of your database sequence ids in a table as you go and automate as much as you can throughout.

Manual data entry leads to manual data error.