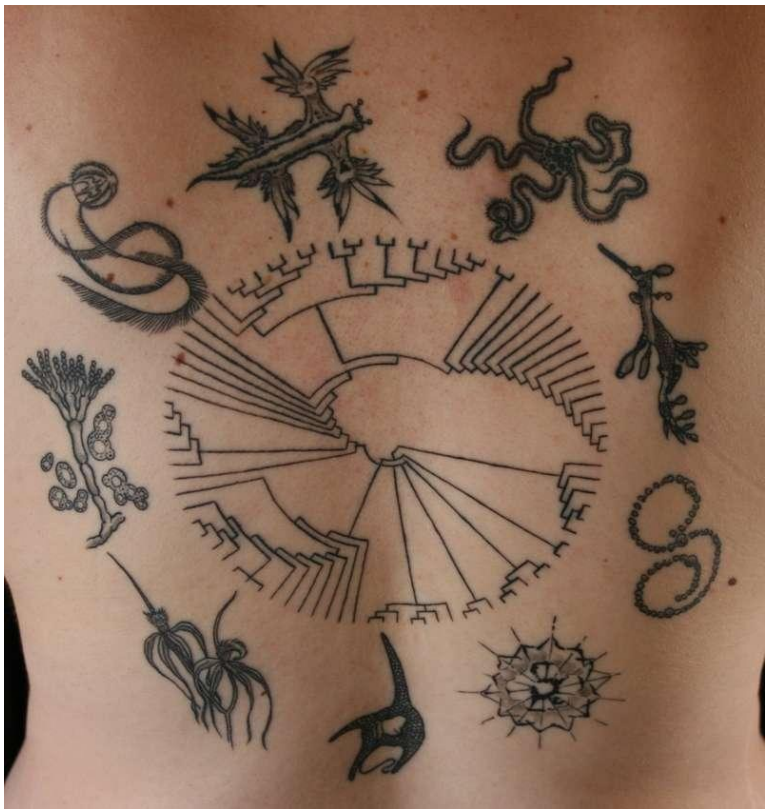
Homology and pairwise alignment

James B. Pease

University of Michigan

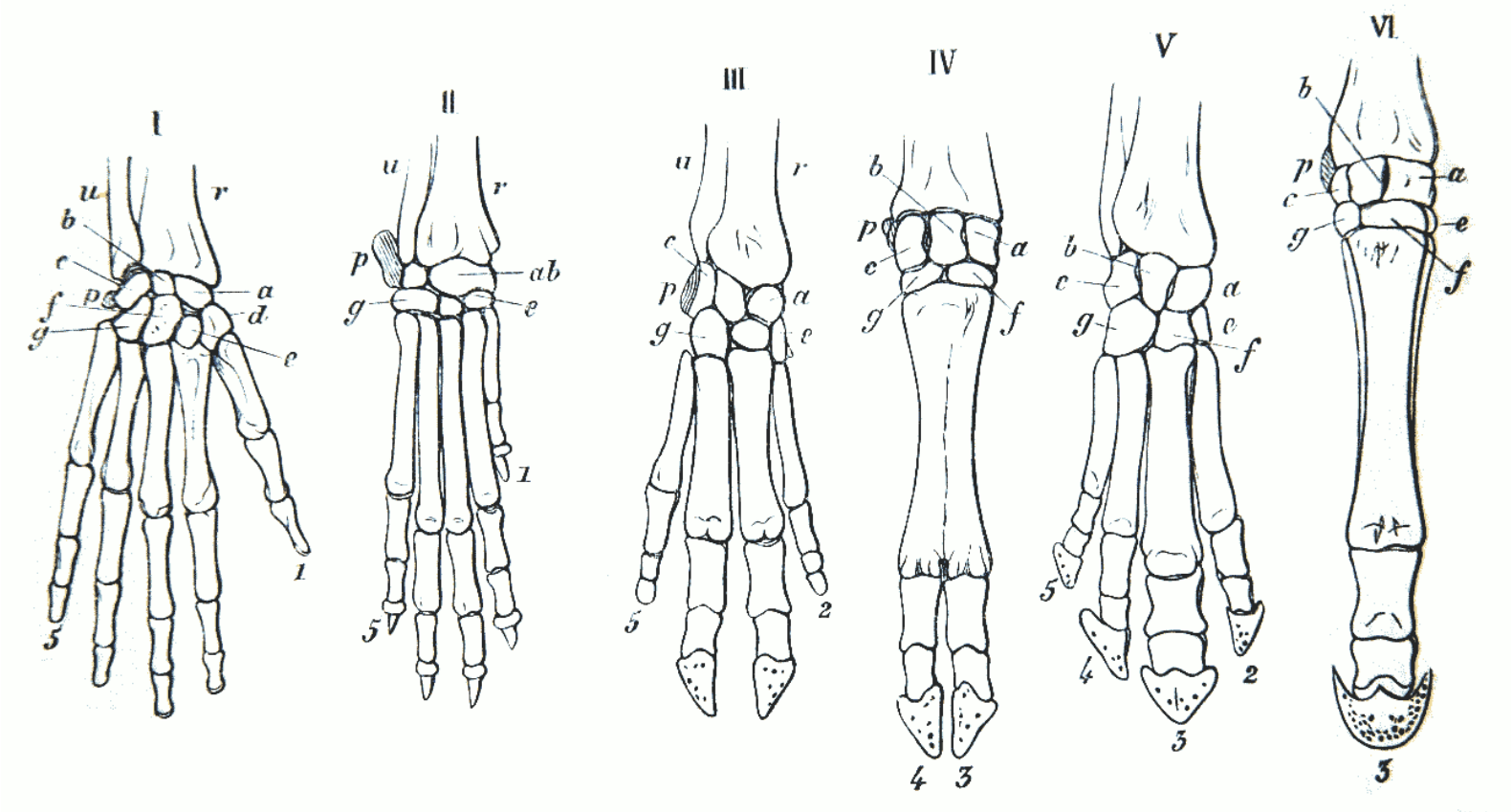


῾Ο βίος βραχύς,
ἡ δὲ τέχνη μακρή,
ὁ δὲ καιρὸς ὀξύς,
ἡ δὲ πεῖρα σφαλερή,
ἡ δὲ κρίσις χαλεπή.
— ῾Ιπποκράτης, Αφ. 1.1

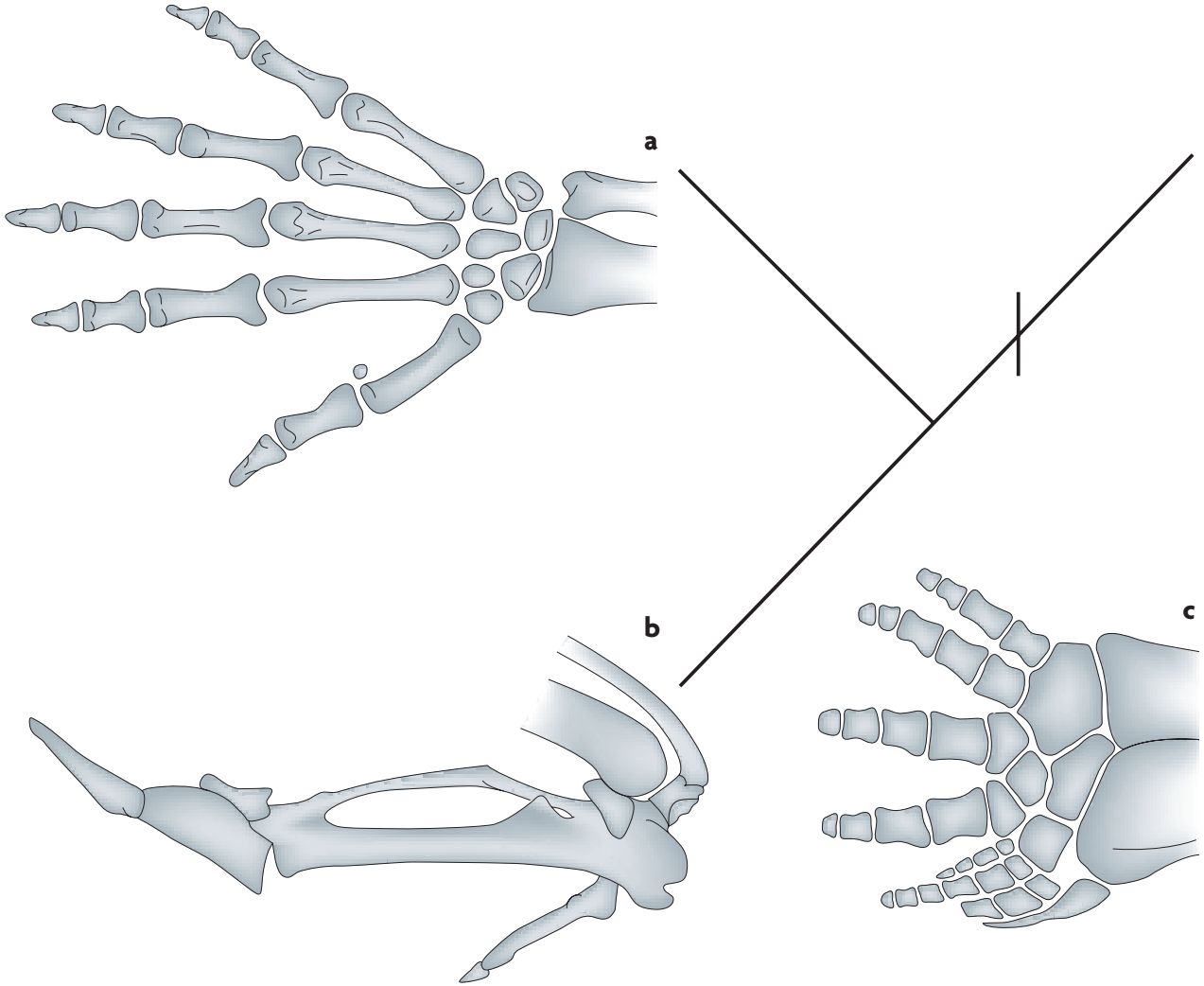
Life is short,
and art long,
opportunity fleeting,
experience perilous,
and decision difficult.
— Hippokrates, Aph. 1.1

Homology

Homology: descended from a common ancestor



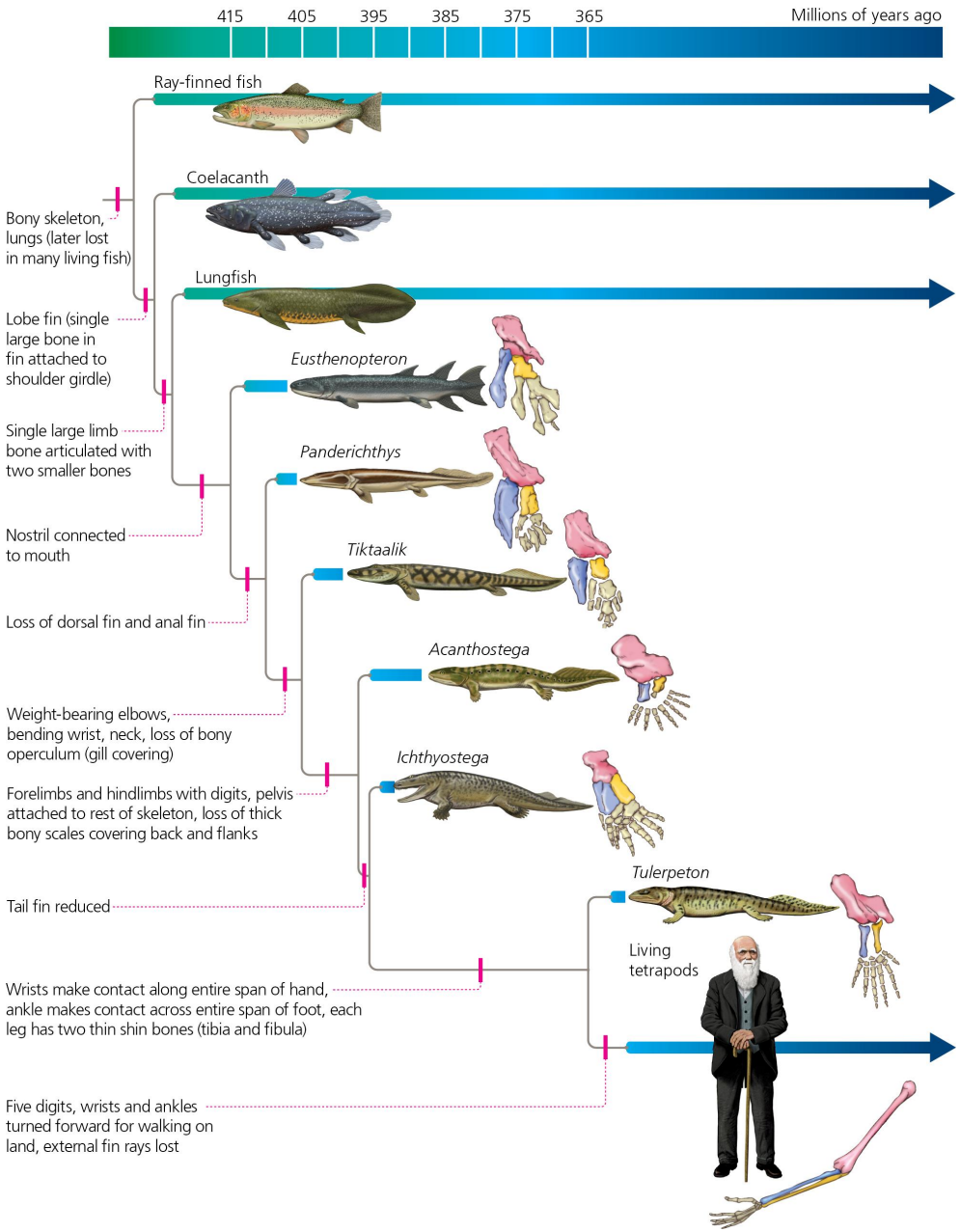
Homology: descended from a common ancestor

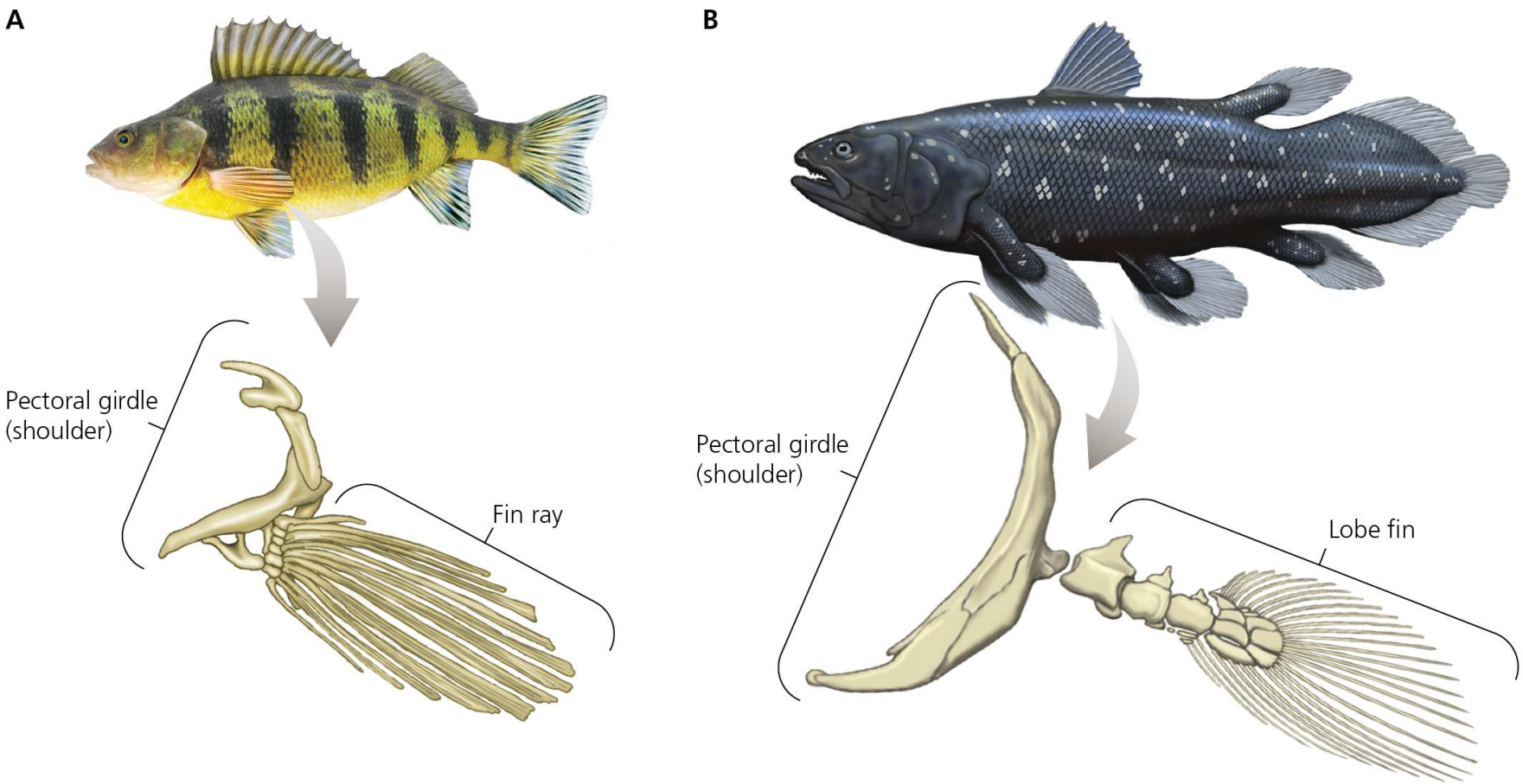


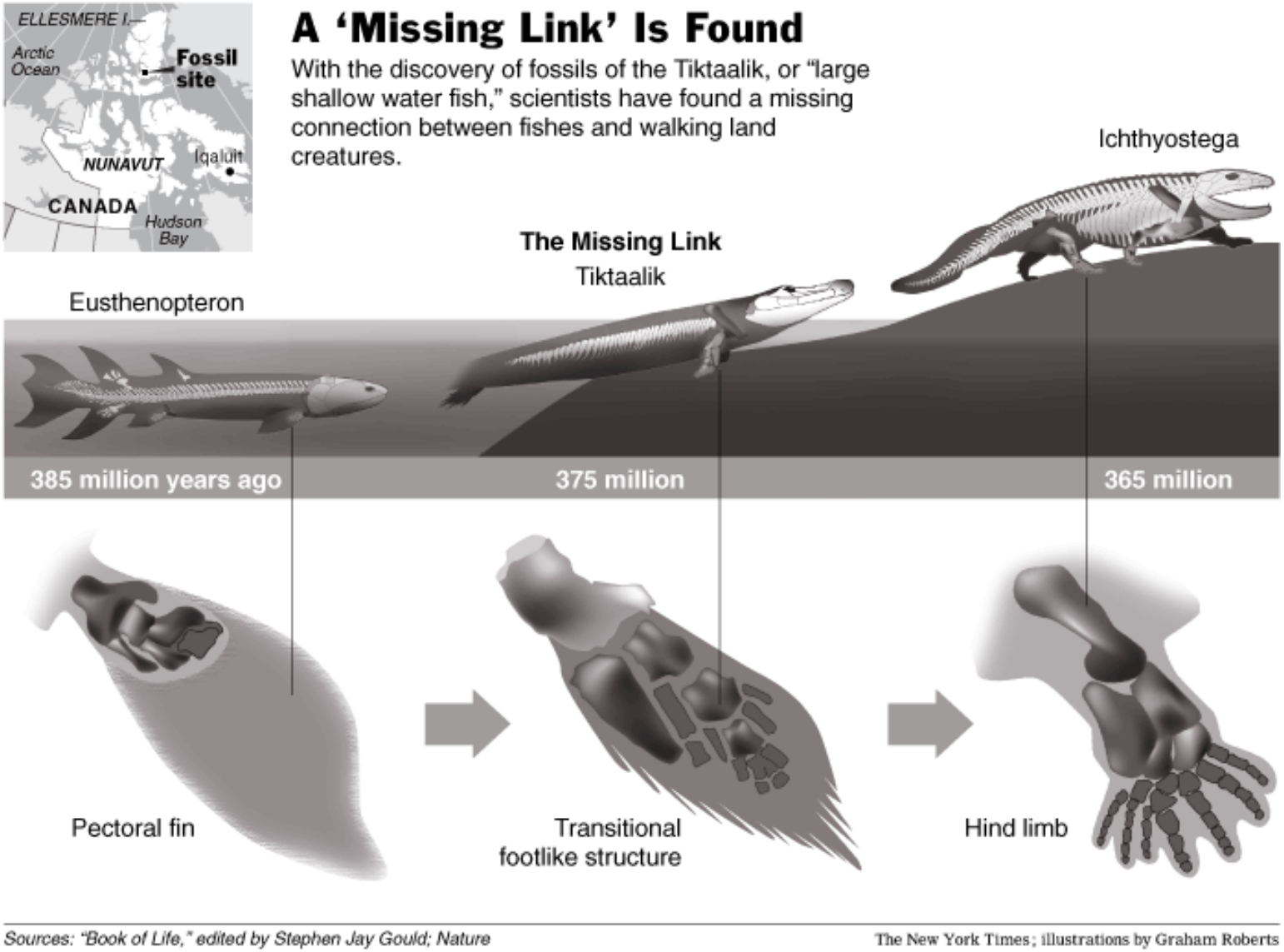
Wagner, 2007

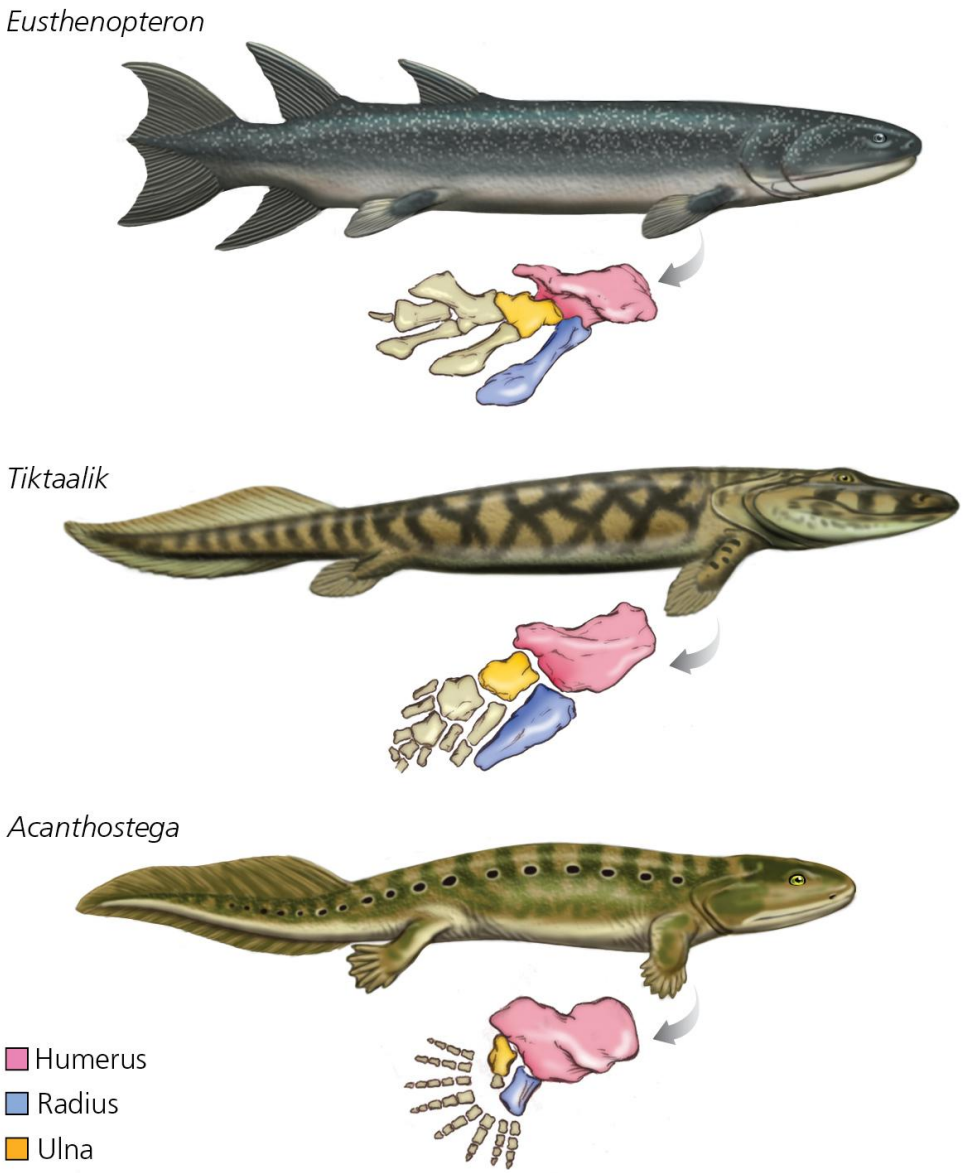
How can we determine homology?

- Hypotheses based on morphology
- Hypotheses based on sequence similarity (what we are mostly going to be doing)
- Experimentally verify

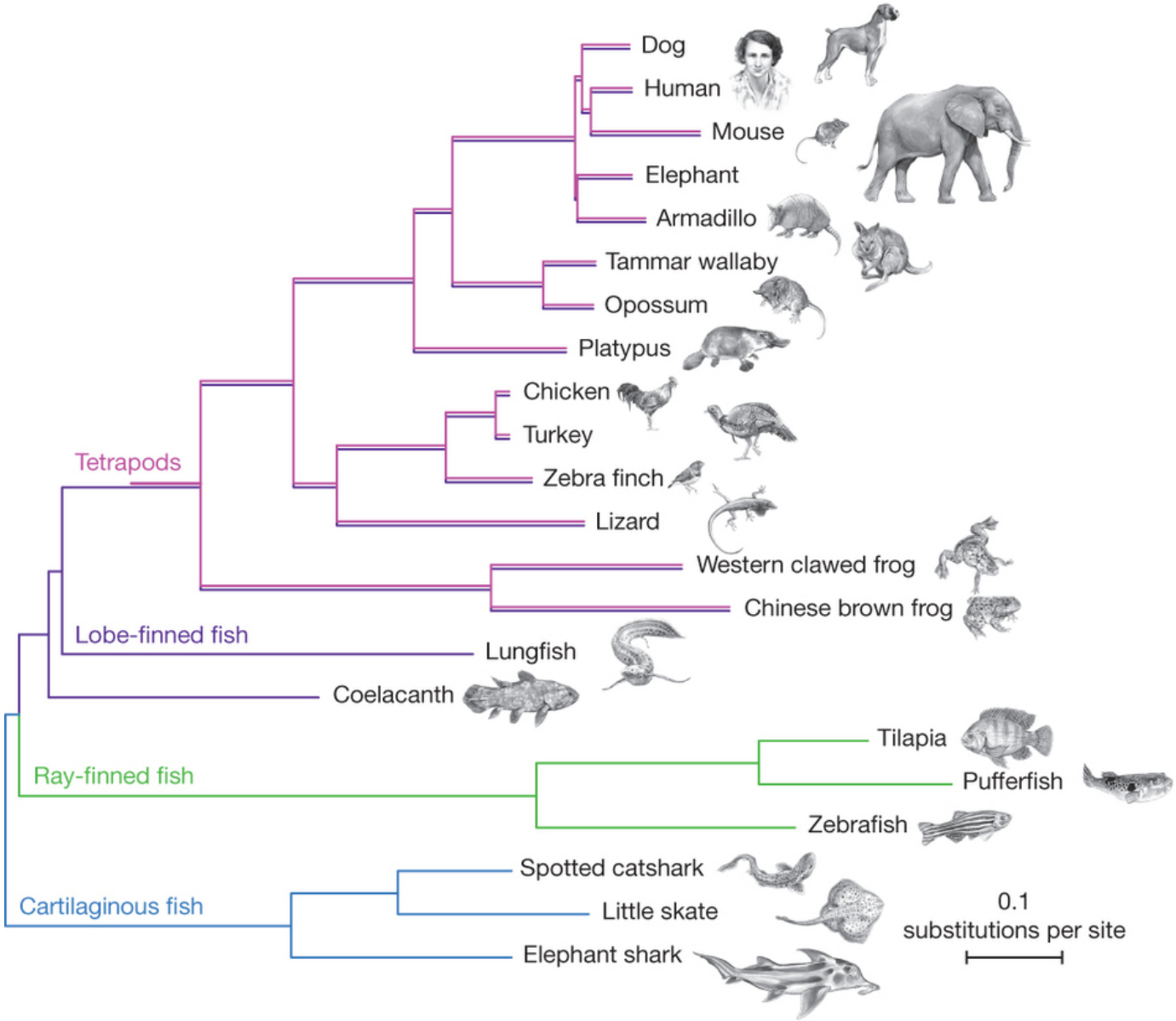

415
405
395
385
375
365
Millions of years ago
Ray-finned fish
Coelacanth
Lungfish
Eusthenopteron
Panderichthys
Tiktaalik
Acanthostega
Ichthyostega
Tulerpeton
Living tetrapods
Bony skeleton, lungs (later lost in many living fish)
Lobe fin (single large bone in fin attached to shoulder girdle)
Single large limb bone articulated with two smaller bones
Nostril connected to mouth
Loss of dorsal fin and anal fin
Weight-bearing elbows, bending wrist, neck, loss of bony operculum (gill covering)
Forelimbs and hindlimbs with digits, pelvis attached to rest of skeleton, loss of thick bony scales covering back and flanks
Tail fin reduced
Wrists make contact along entire span of hand, ankle makes contact across entire span of foot, each leg has two thin shin bones (tibia and fibula)
Five digits, wrists and ankles turned forward for walking on land, external fin rays lost

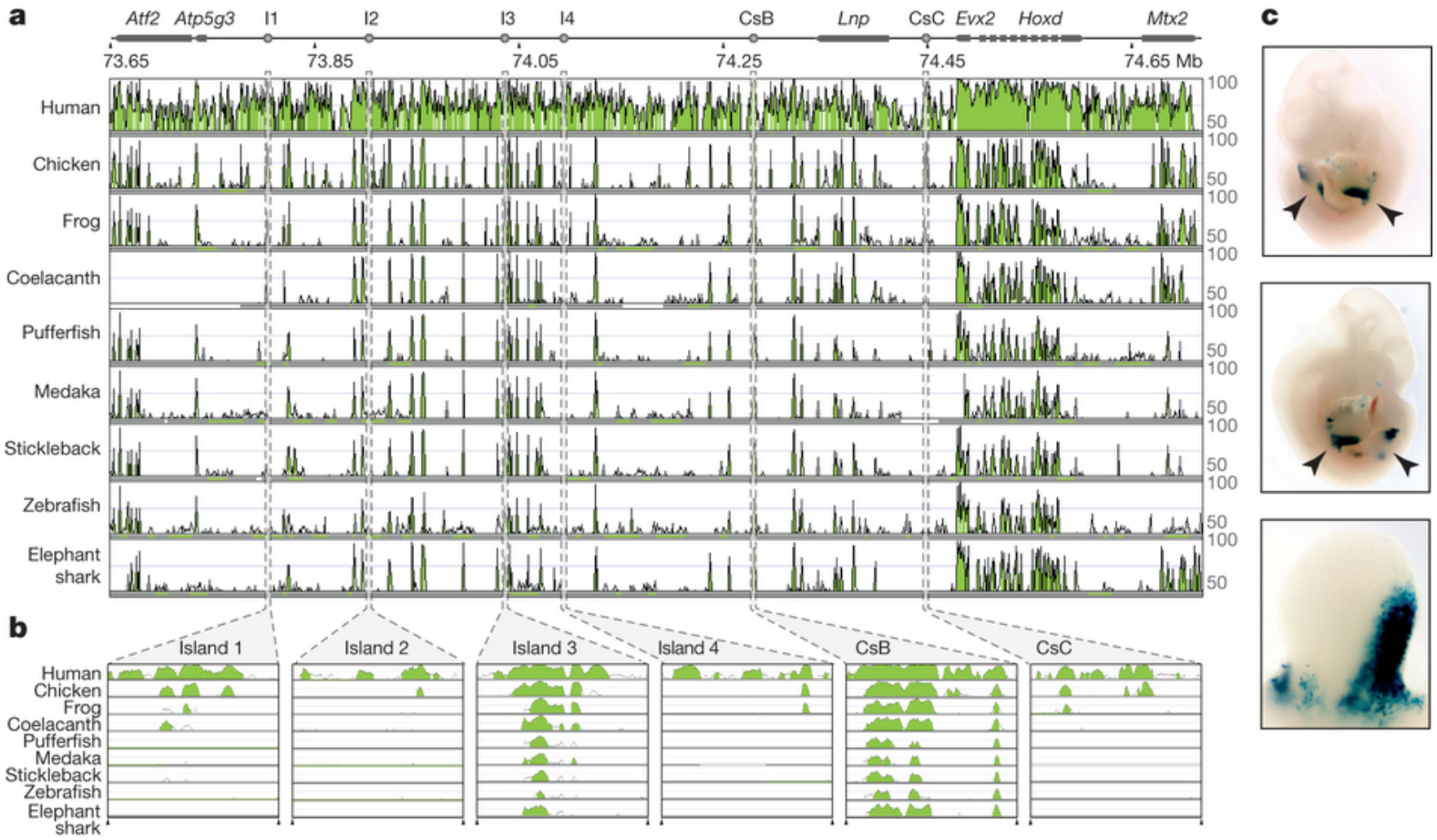

A
Pectoral girdle
(shoulder)
Fin ray
B
Pectoral girdle
(shoulder)
Lobe fin

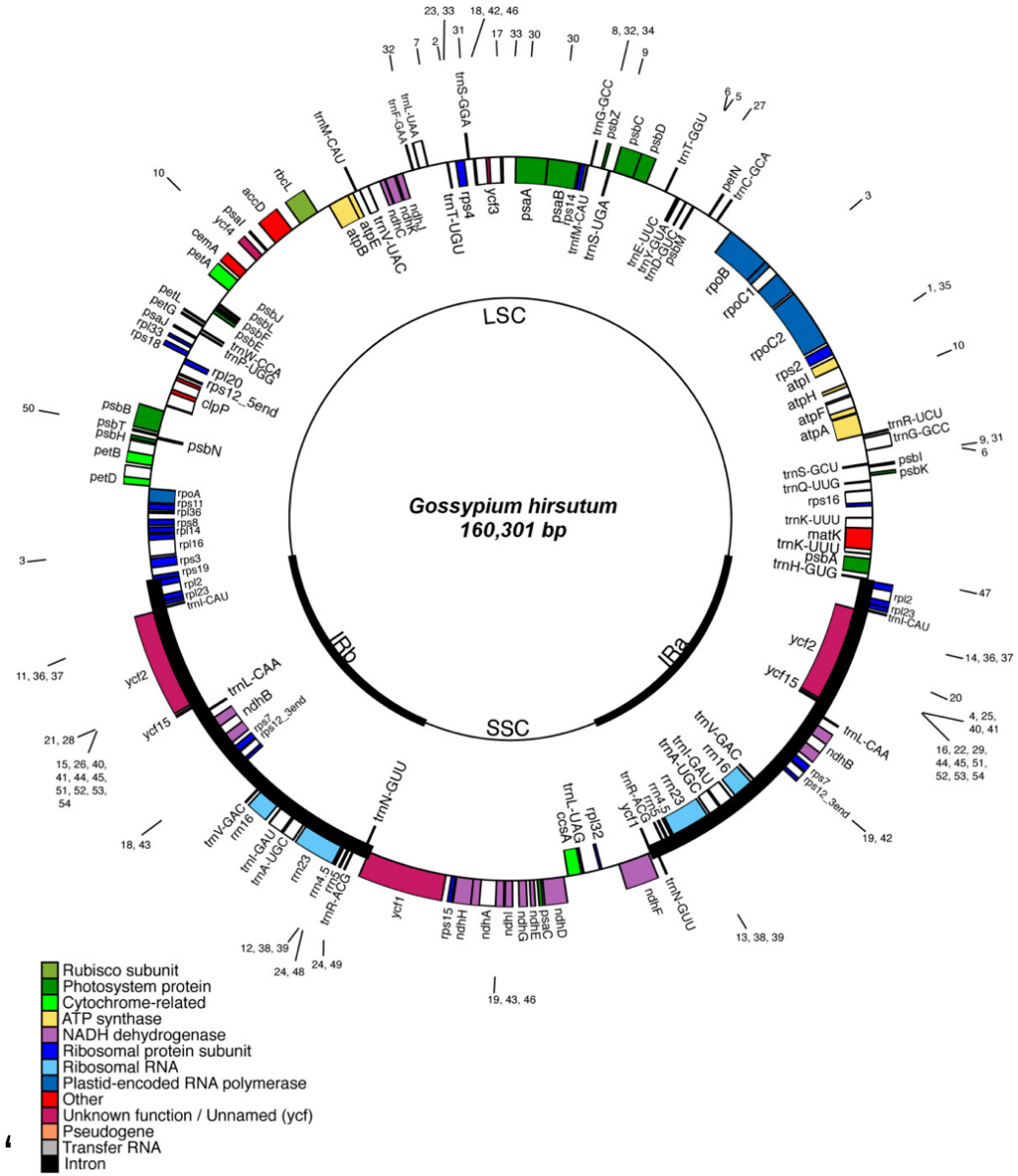

ELLESMERE I.
Arctic Ocean
Fossil site
NUNAVUT
Iqaluit
CANADA
Hudson Bay
A 'Missing Link' Is Found
With the discovery of fossils of the Tiktaalik, or "large shallow water fish," scientists have found a missing connection between fishes and walking land creatures.
Ichthyostega
The Missing Link
Tiktaalik
Eusthenopteron
385 million years ago
375 million
365 million
Pectoral fin
Transitional footlike structure
Hind limb
Sources: "Book of Life," edited by Stephen Jay Gould; Nature
The New York Times; illustrations by Graham Roberts


Eusthenopteron
Tiktaalik
Acanthostega
Humerus
Radius
Ulna

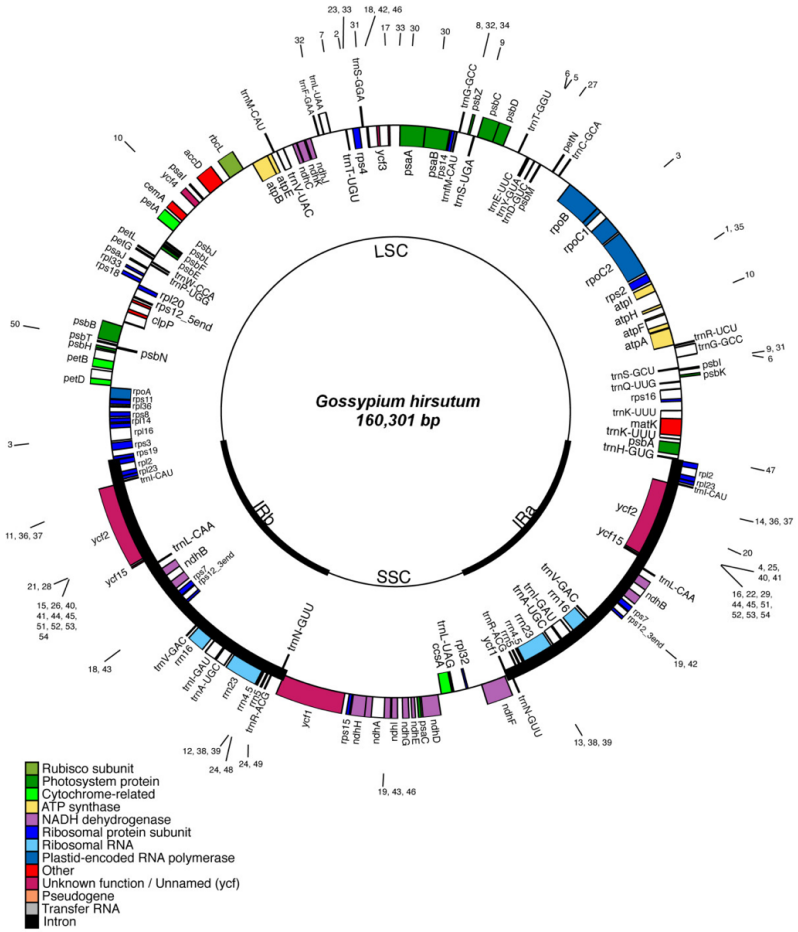
Coelacanth

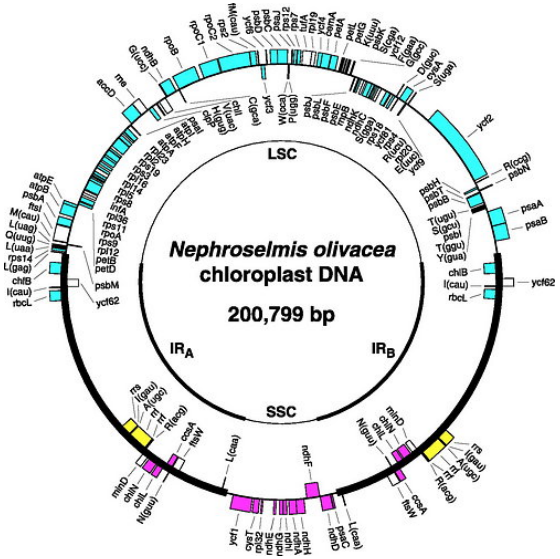

Dog
Human
Mouse
Elephant
Armadillo
Tammar wallaby
Opossum
Platypus
Chicken
Turkey
Zebra finch
Lizard
Western clawed frog
Chinese brown frog
Tetrapods
Lobe-finned fish
Lungfish
Coelacanth
Tilapia
Pufferfish
Zebrafish
Ray-finned fish
Spotted catshark
Little skate
Elephant shark
Cartilaginous fish
0.1
substitutions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

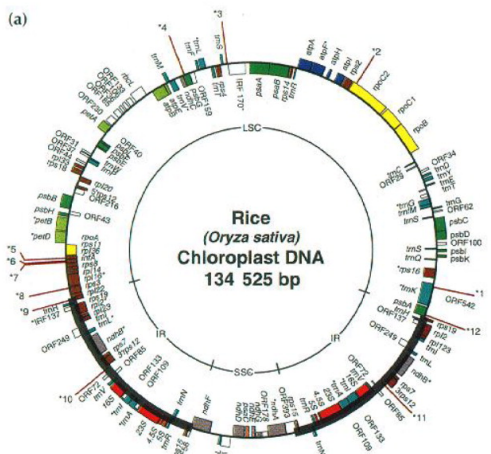

a
Atf2
Atp5g3
I1
I2
I3
I4
CsB
Lnp
CsC
Evx2
Hoxd
Mtx2
73.65
73.85
74.05
74.25
74.45
74.65 Mb
Human
Chicken
Frog
Coelacanth
Pufferfish
Medaka
Stickleback
Zebrafish
Elephant shark
100
50
b
Island 1
Island 2
Island 3
Island 4
CsB
CsC
c

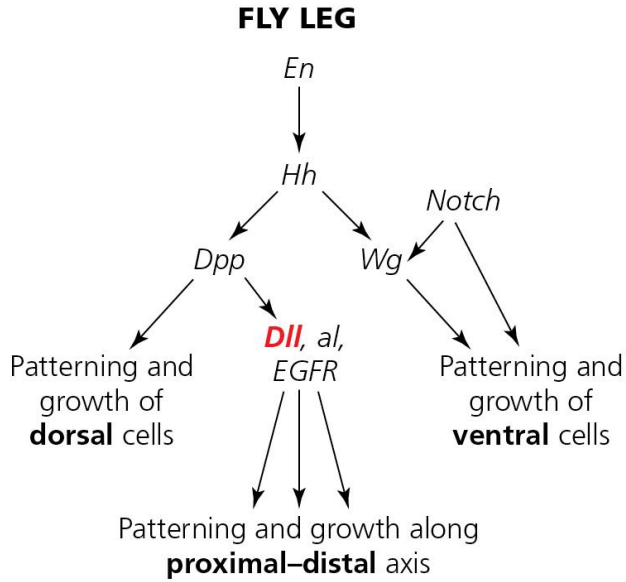

Gossypium hirsutum
160,301 bp
LSC
IRb
IRa
SSC
Rubisco subunit
Photosystem protein
Cytochrome-related
ATP synthase
NADH dehydrogenase
Ribosomal protein subunit
Ribosomal RNA
Plastid-encoded RNA polymerase
Other
Unknown function / Unnamed (ycf)
Pseudogene
Transfer RNA
Intron

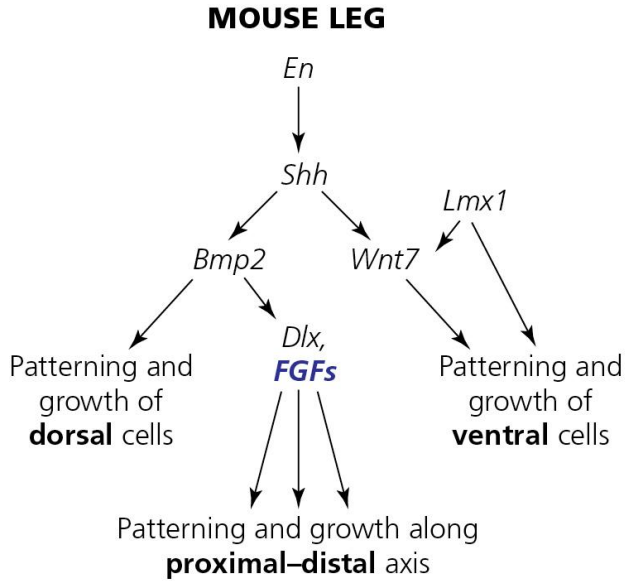
Sequence similarity

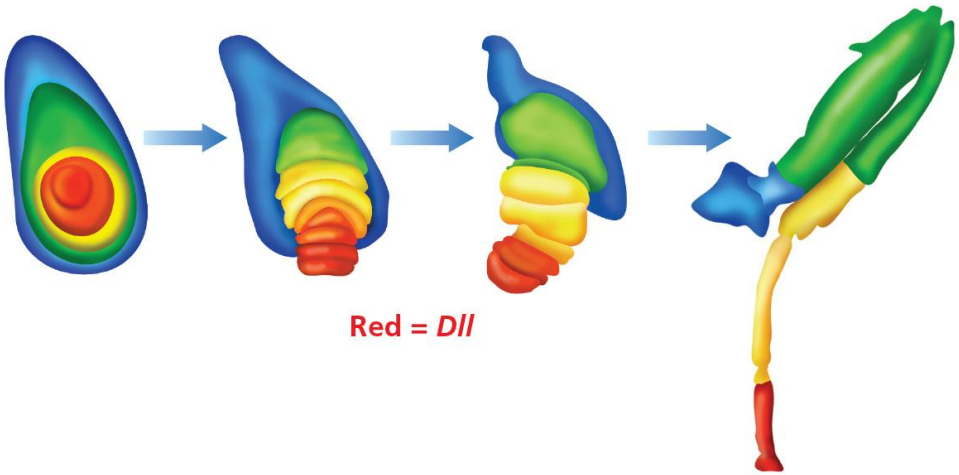


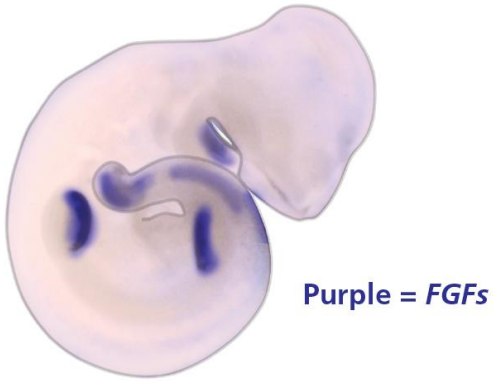




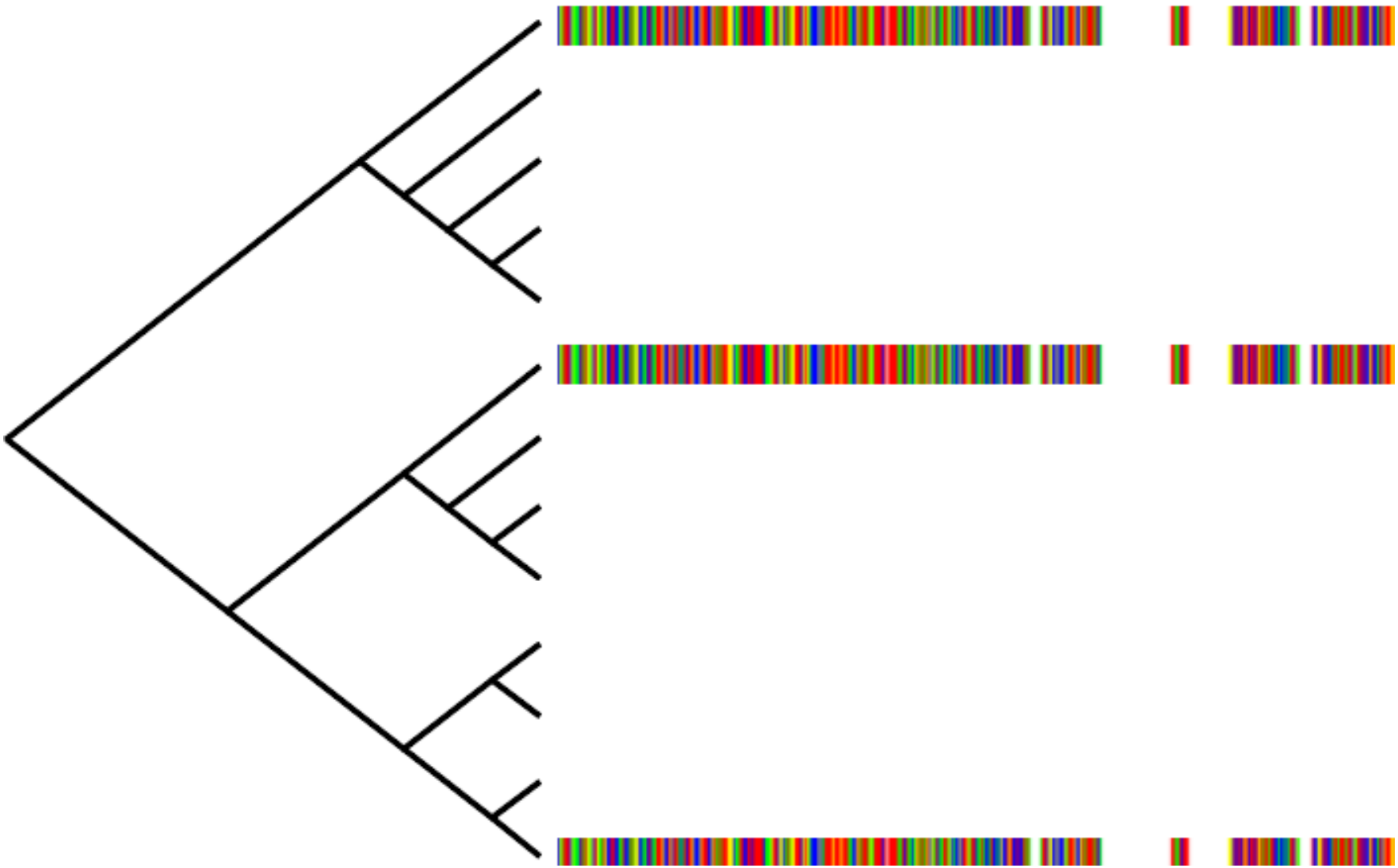

FLY LEG
En
Hh
Notch
Dpp
Wg
Dll, al, EGFR
Patterning and growth of dorsal cells
Patterning and growth of ventral cells
Patterning and growth along proximal–distal axis

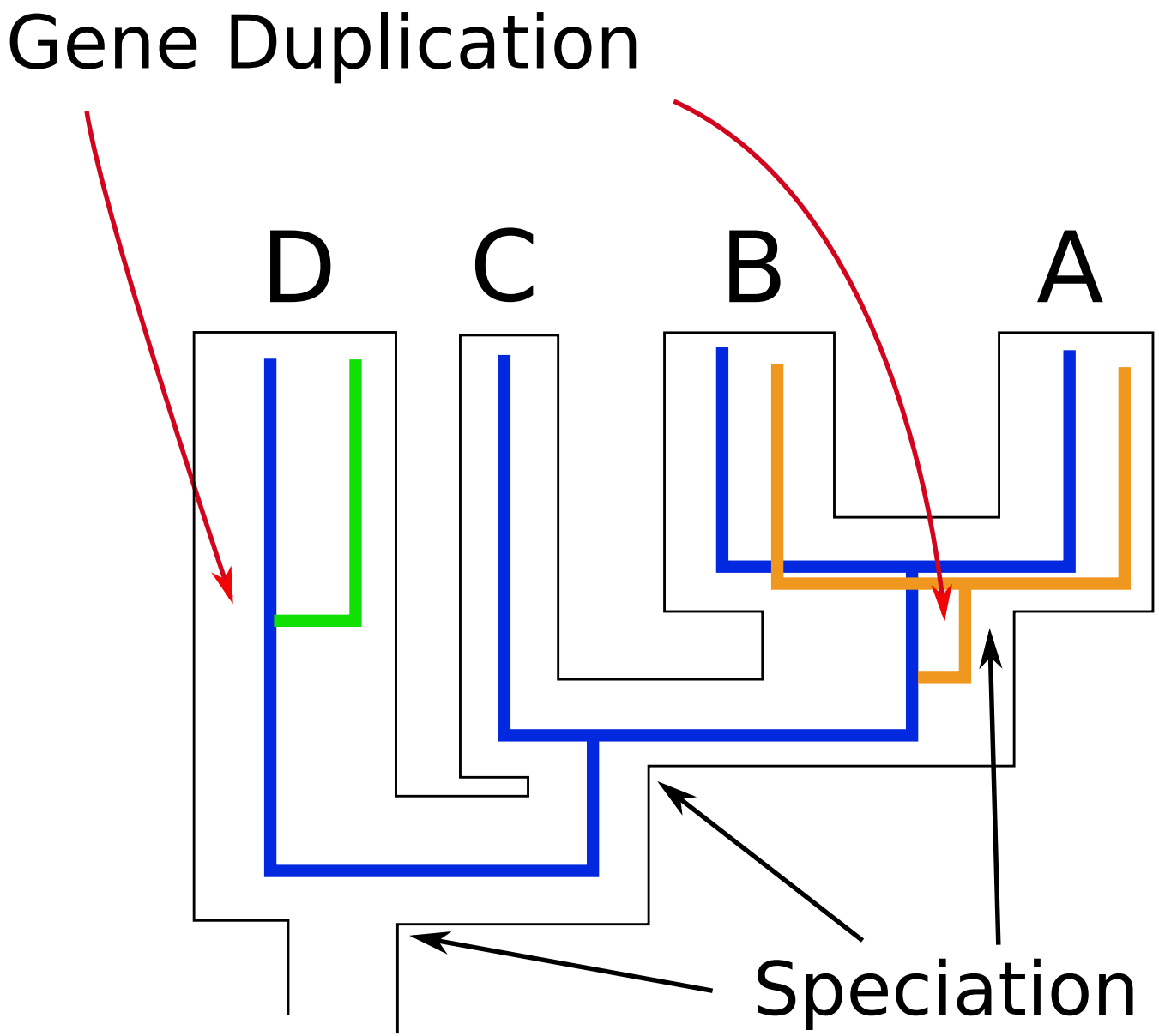

MOUSE LEG
En
Shh
Lmx1
Bmp2
Wnt7
Dlx, FGFs
Patterning and growth of dorsal cells
Patterning and growth of ventral cells
Patterning and growth along proximal–distal axis


Red = Dll

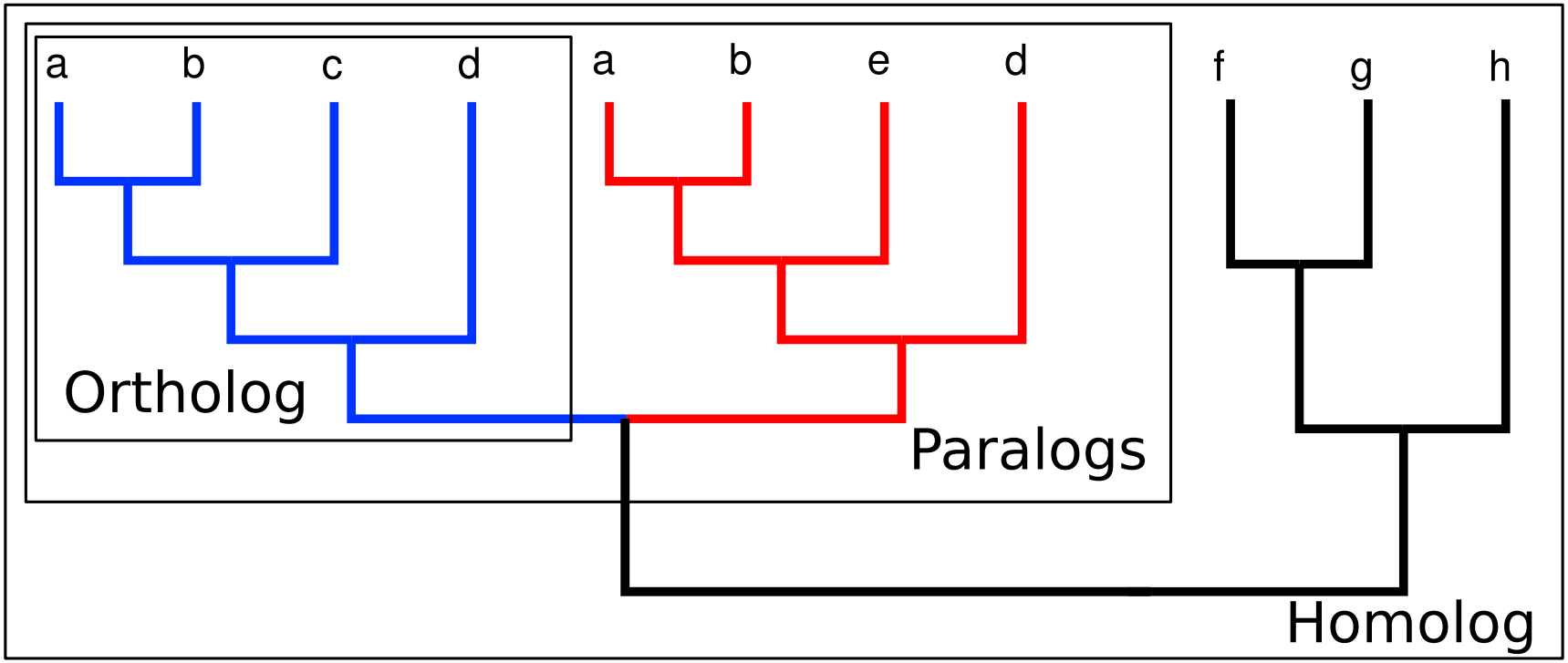

Purple = FGFs

- rbcL
  - Ribulose-1,5-bisphosphate carboxylase/oxygenase (large subunit)
  - RuBisCO (large subunit)
  - rbcL
- ITS
  - internal transcribed spacer
  - ITS
  - ITS1 ITS2

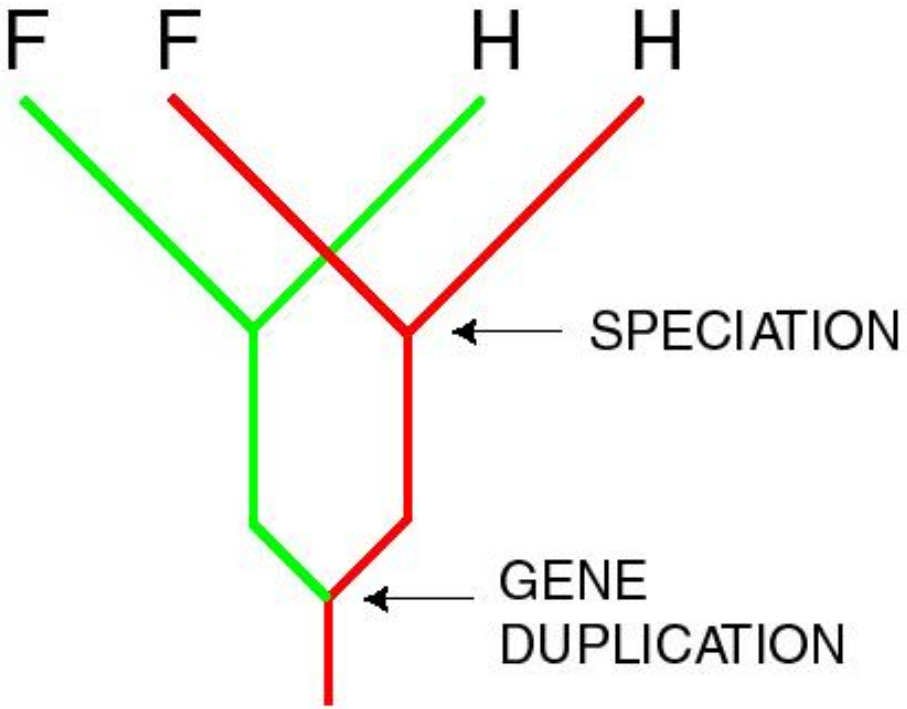


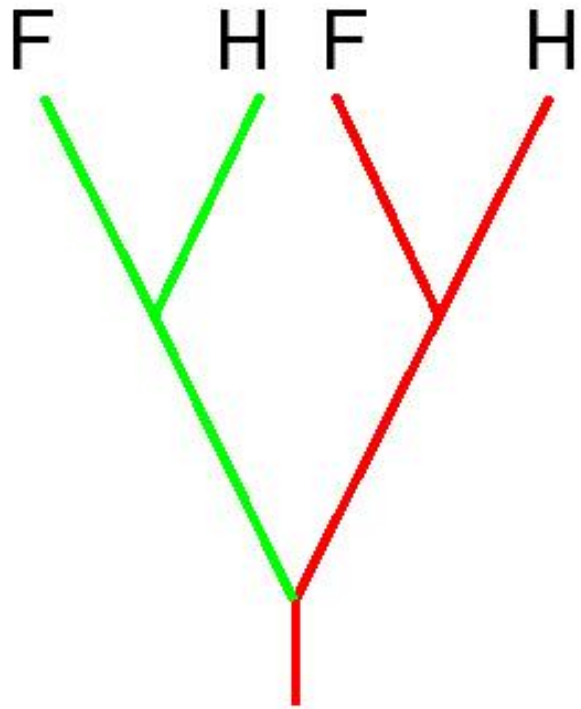

Gene Duplication
D
C
B
A
Speciation

Homology, Orthology, and Paralogy

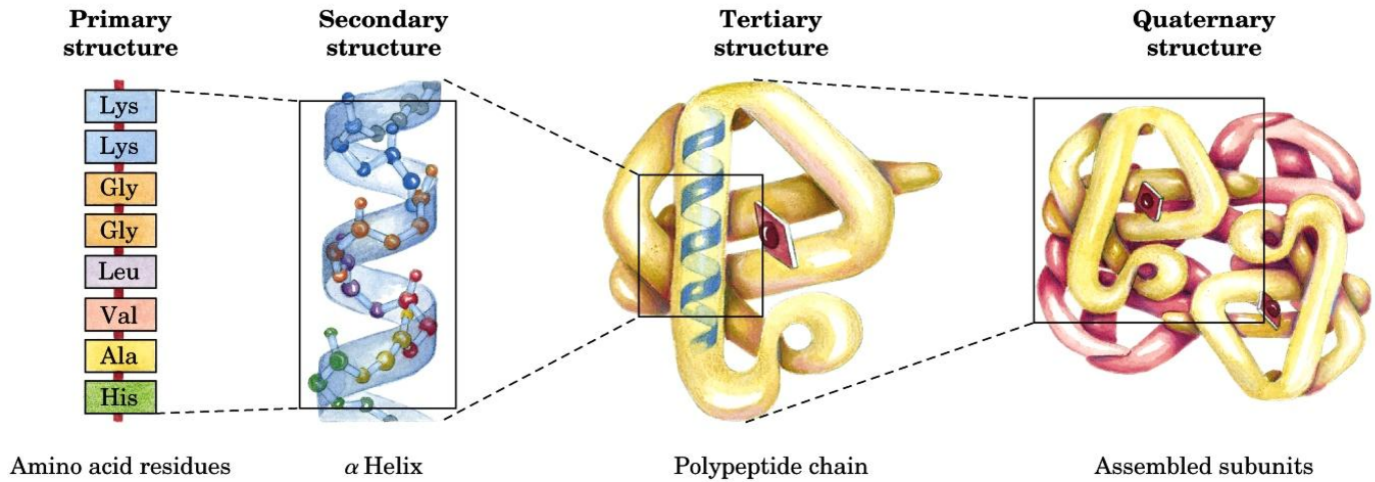


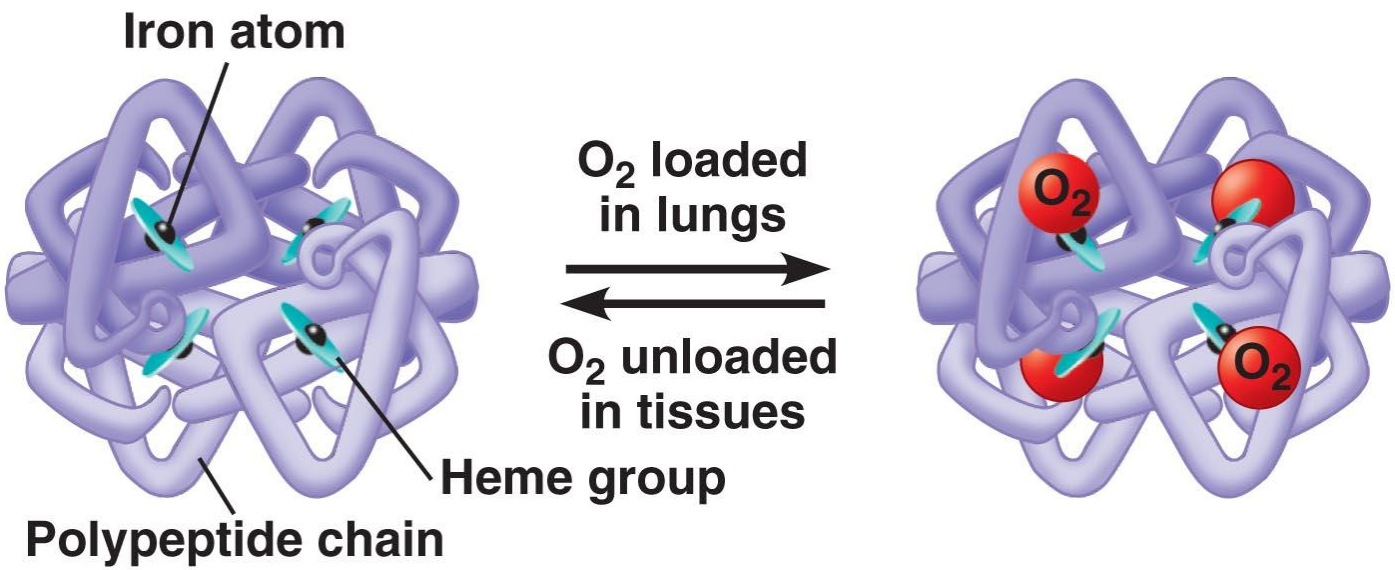
- **Homologs**: descended from a common ancestor
- **Paralogs**: descended from a common ancestor and split by a gene duplication event
- **Orthologs**: descended from a common ancestor and split by a speciation event

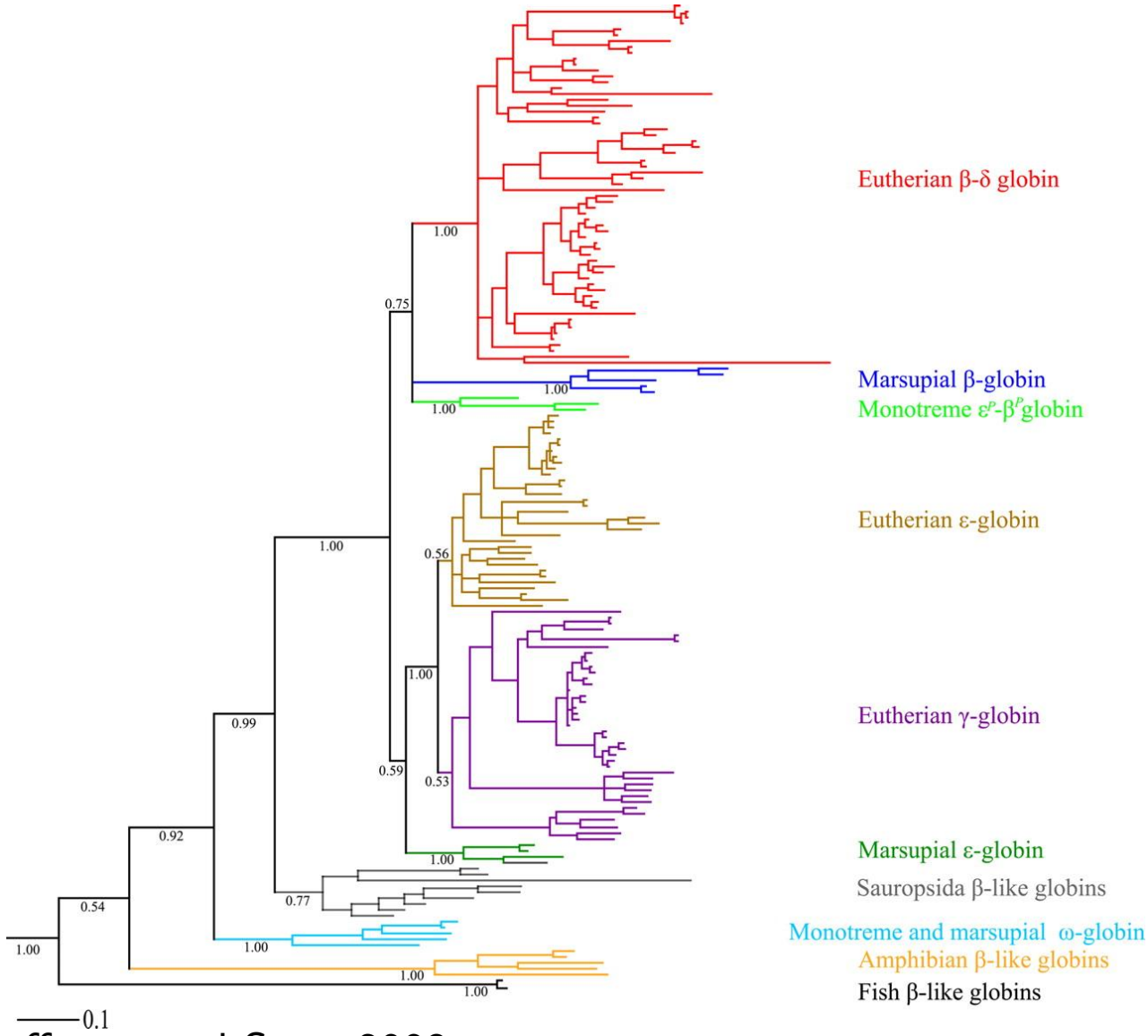




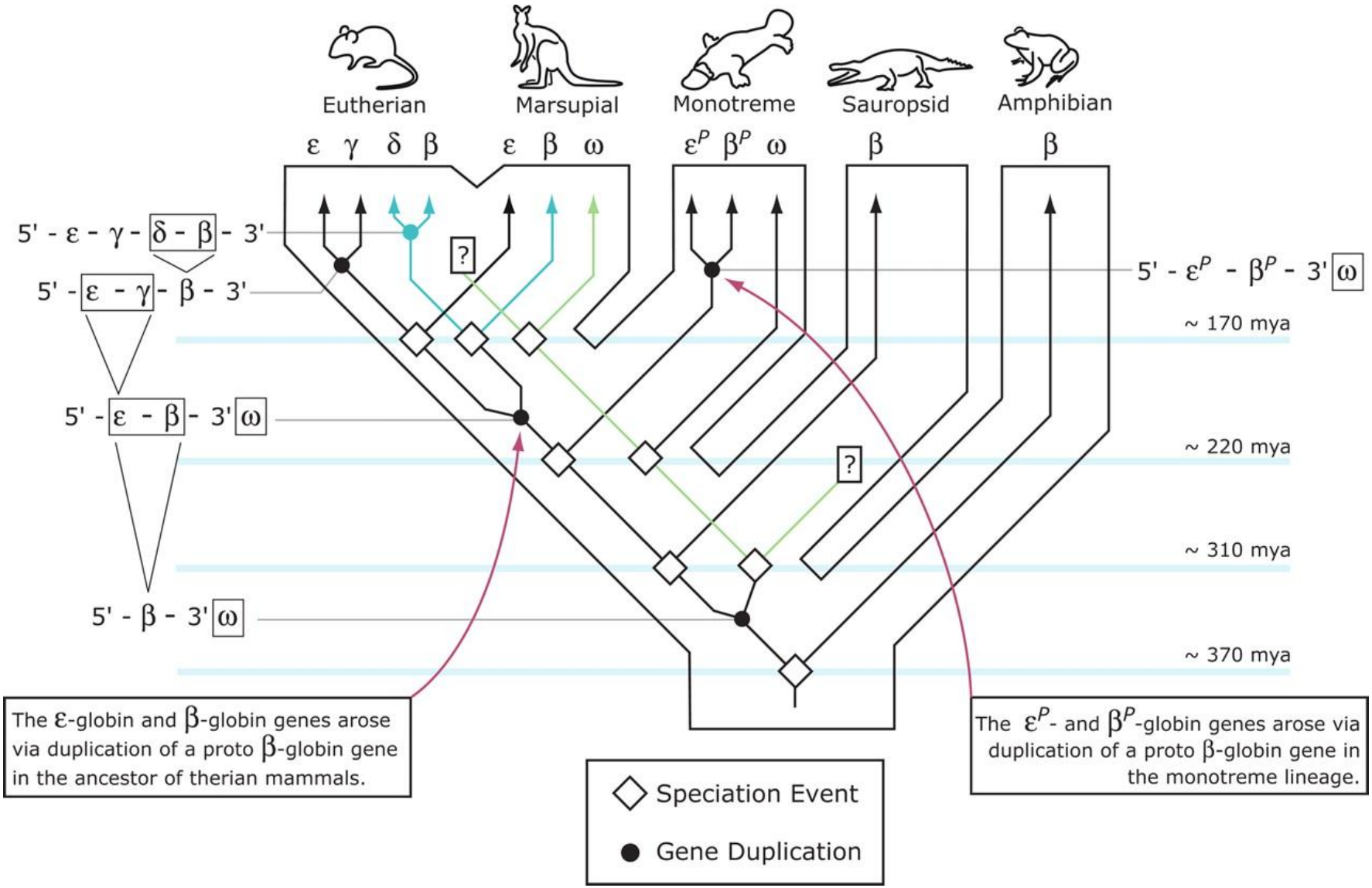
Hemoglobin



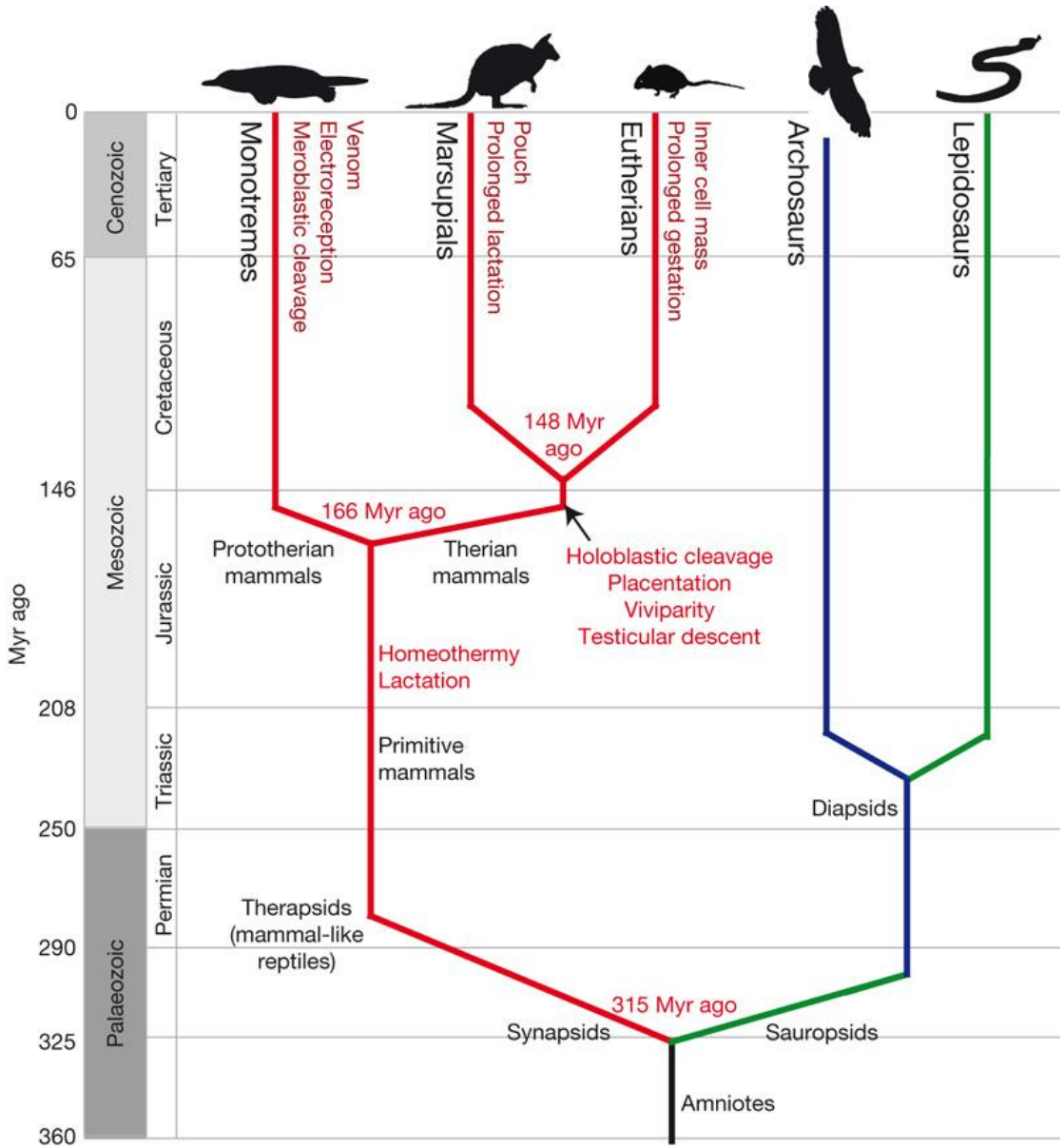


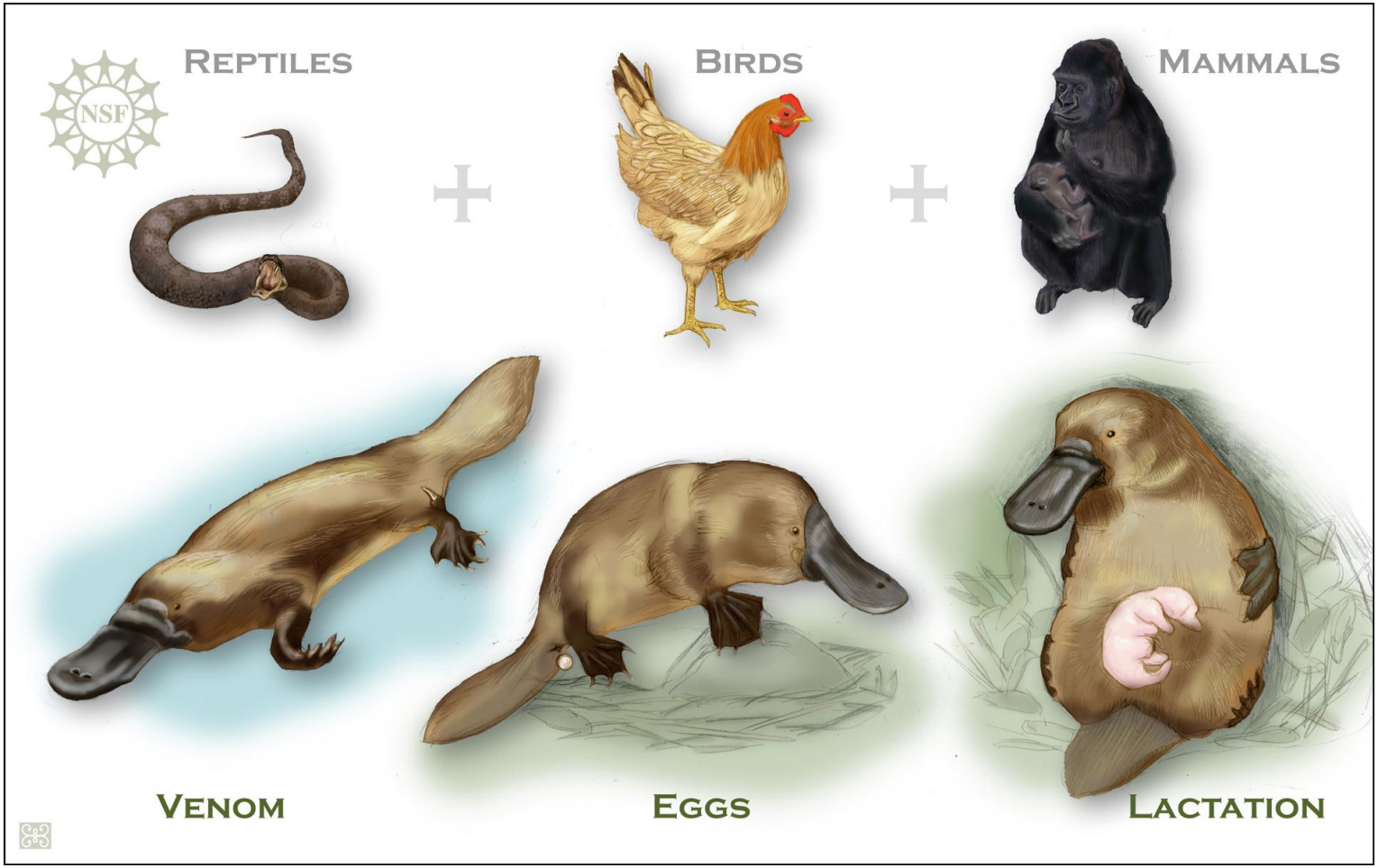


Opazo, Hoffman, and Storz 2008

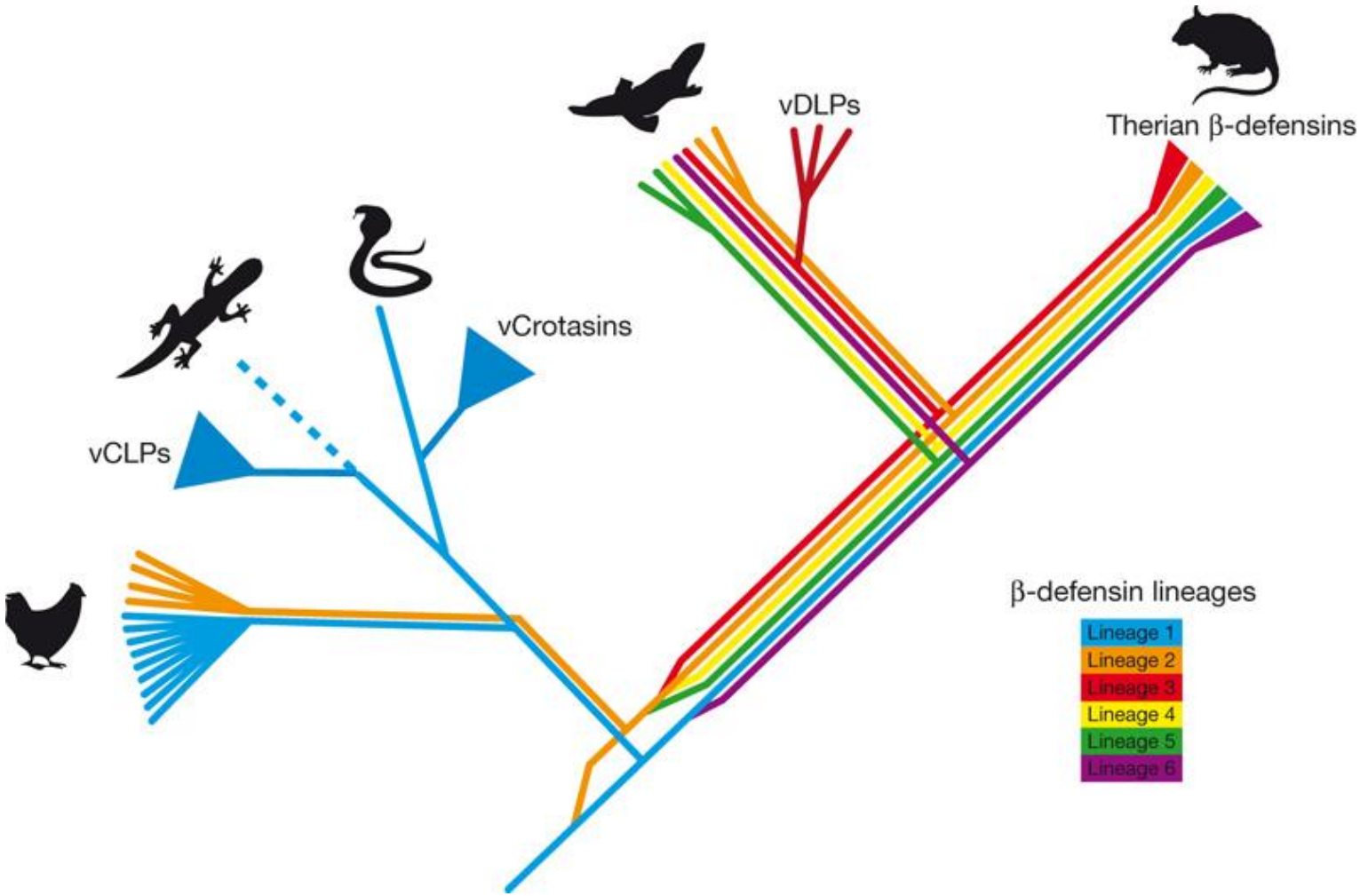


Opazo, Hoffman, and Storz 2008


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

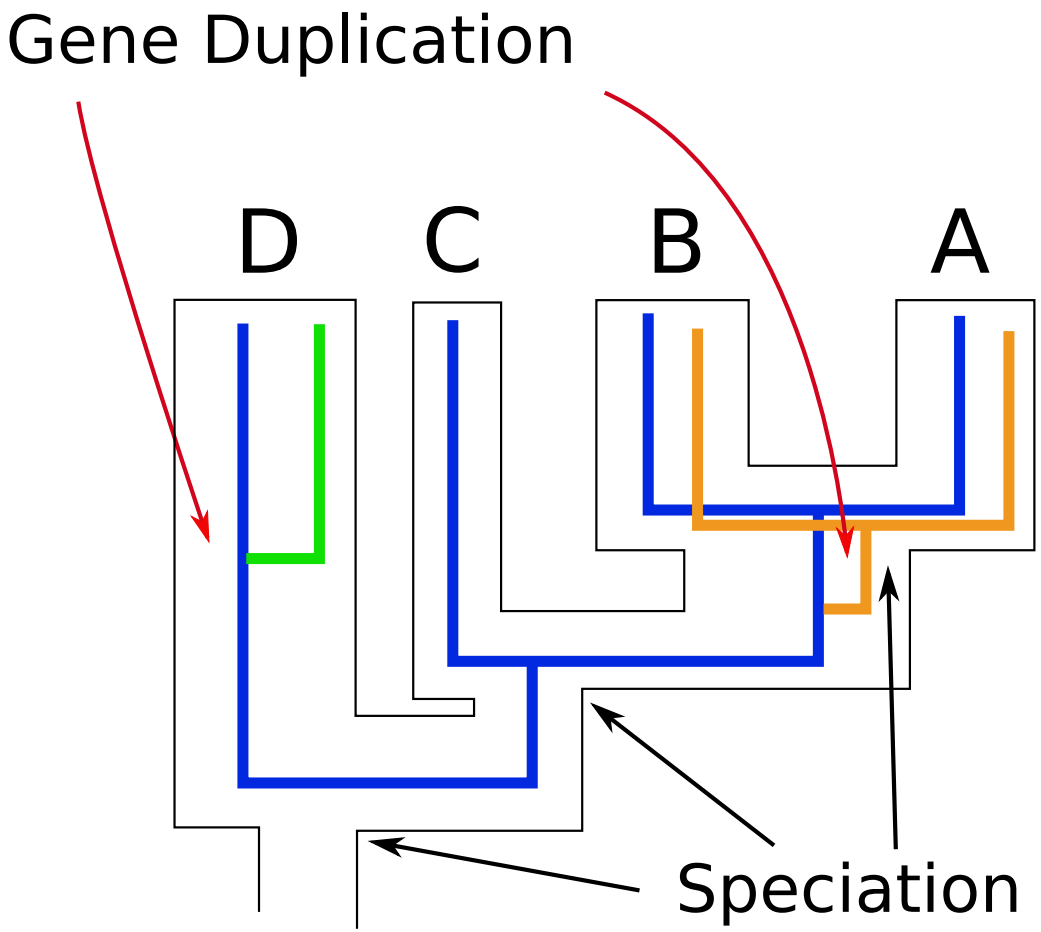

NSF
REPTILES
BIRDS
MAMMALS
VENOM
EGGS
LACTATION

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



What are we concerned with?

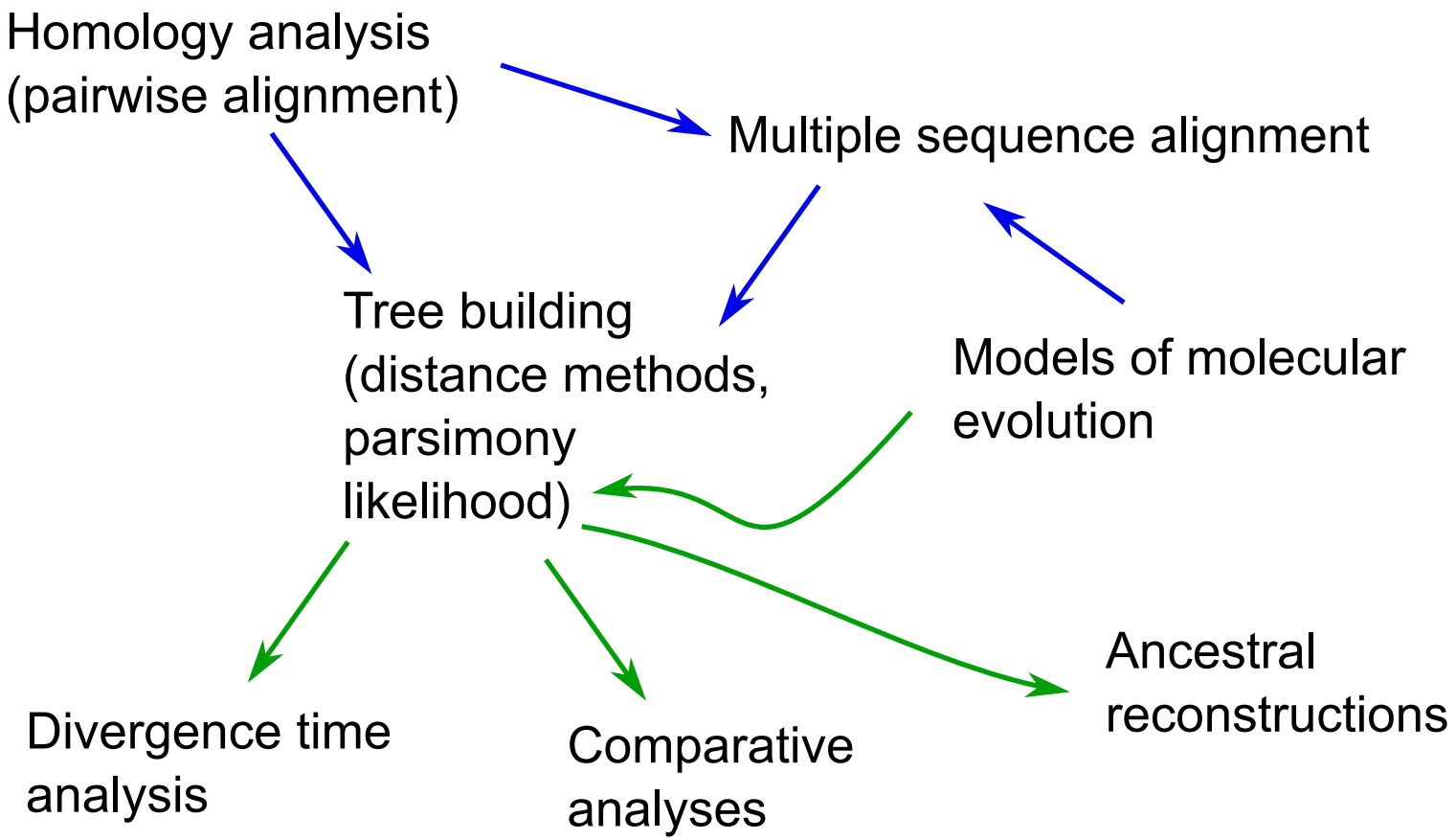
- We need to be able to identify **homologous** gene regions
  - gene regions that share a common ancestor



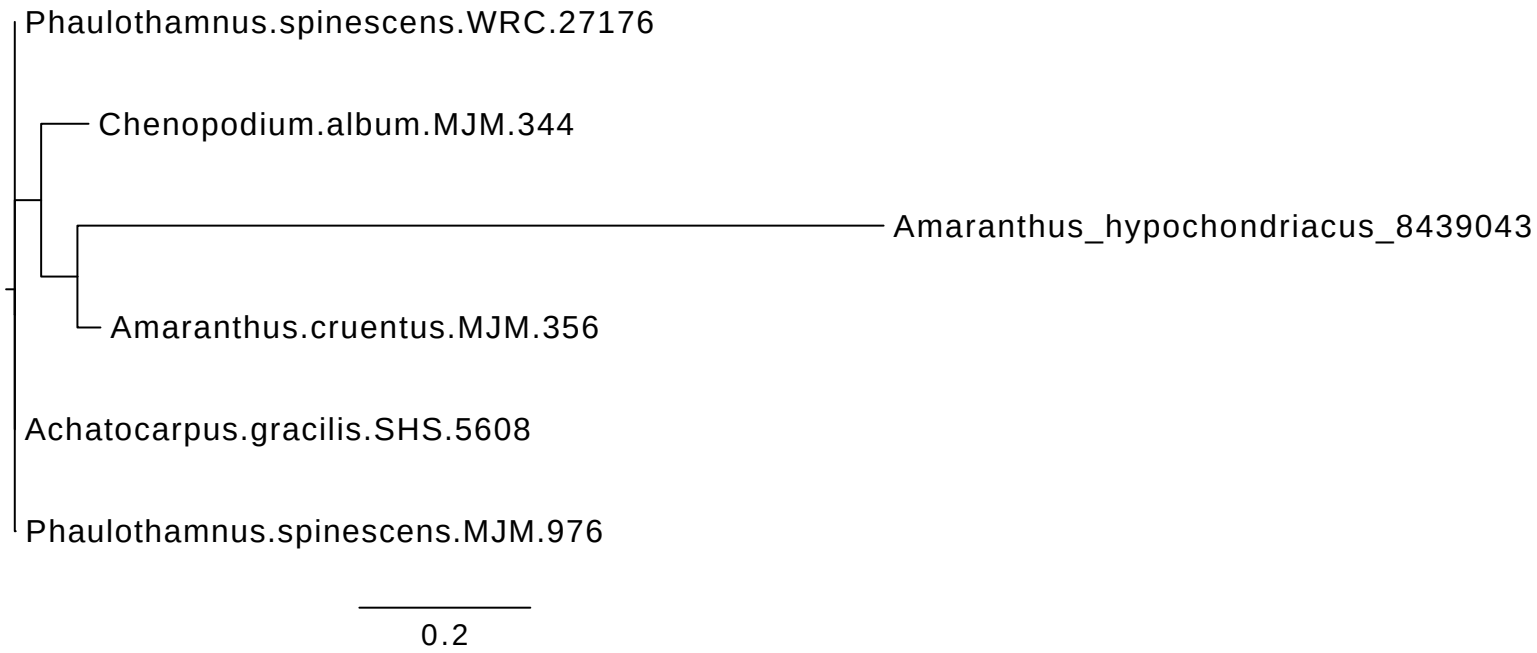
Why are homologs important to identify?

- Identifying homologous gene regions is the first task in most evolutionary and phylogenetic questions
- Homology is central to:
  - multiple sequence alignments
    - each column is assumed to be a homologous site
  - phylogenetic trees
    - A tree models common ancestry relationships, so sequences need to be homologous
  - short read alignments (NGS)
  - clustering genes (NGS)

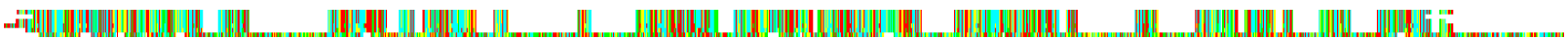
Overview


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

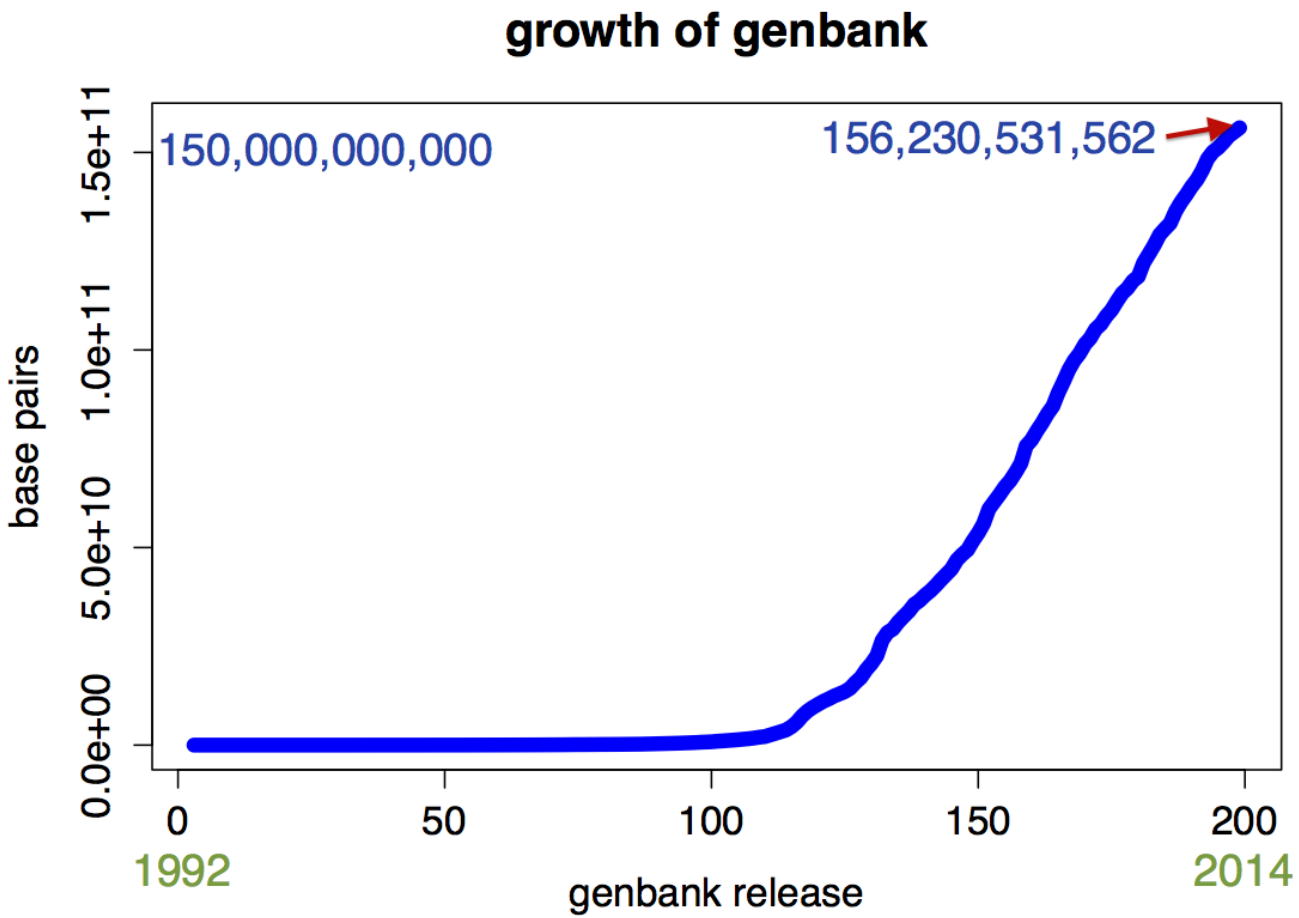
Tree


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

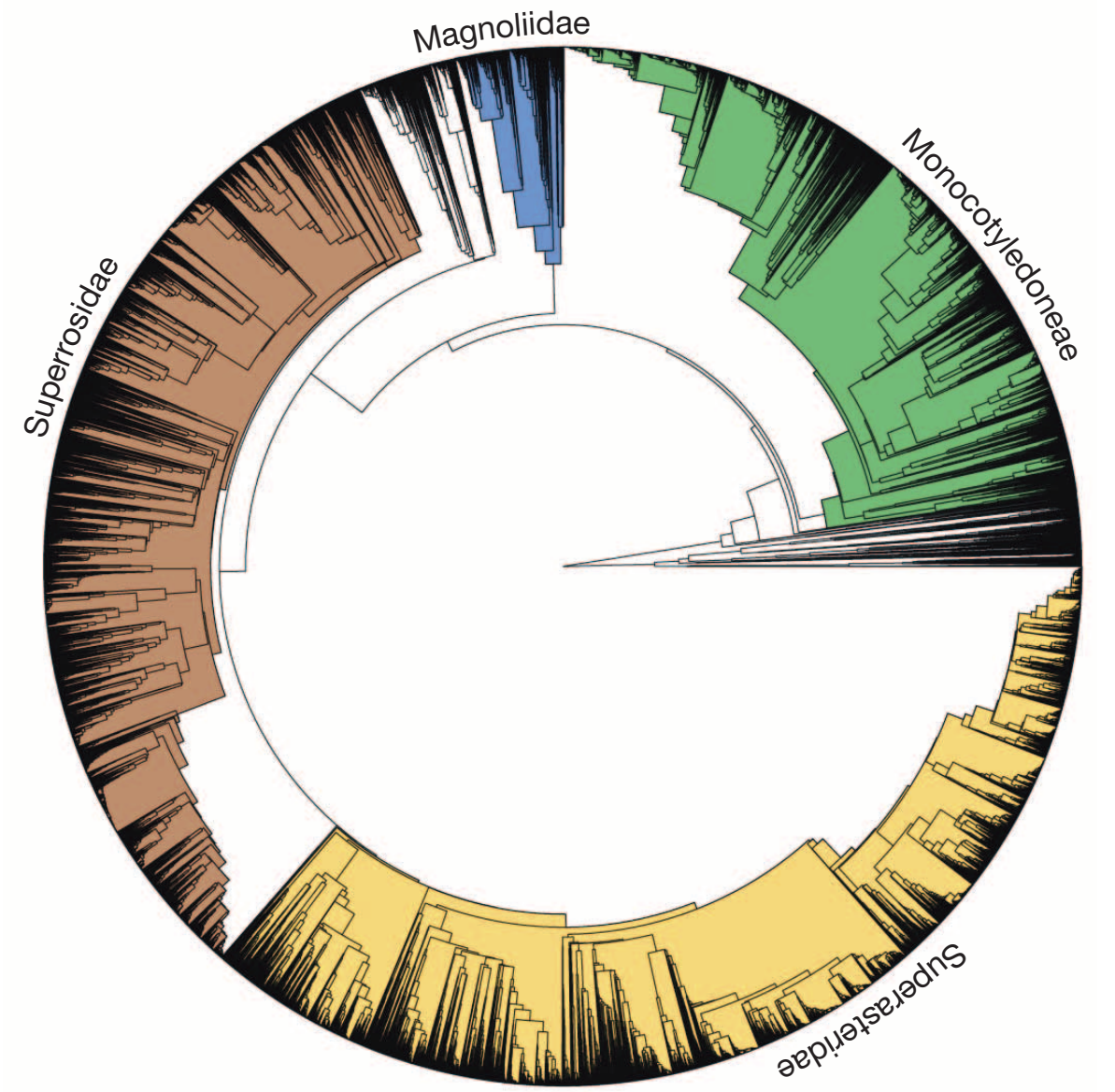


NCBI

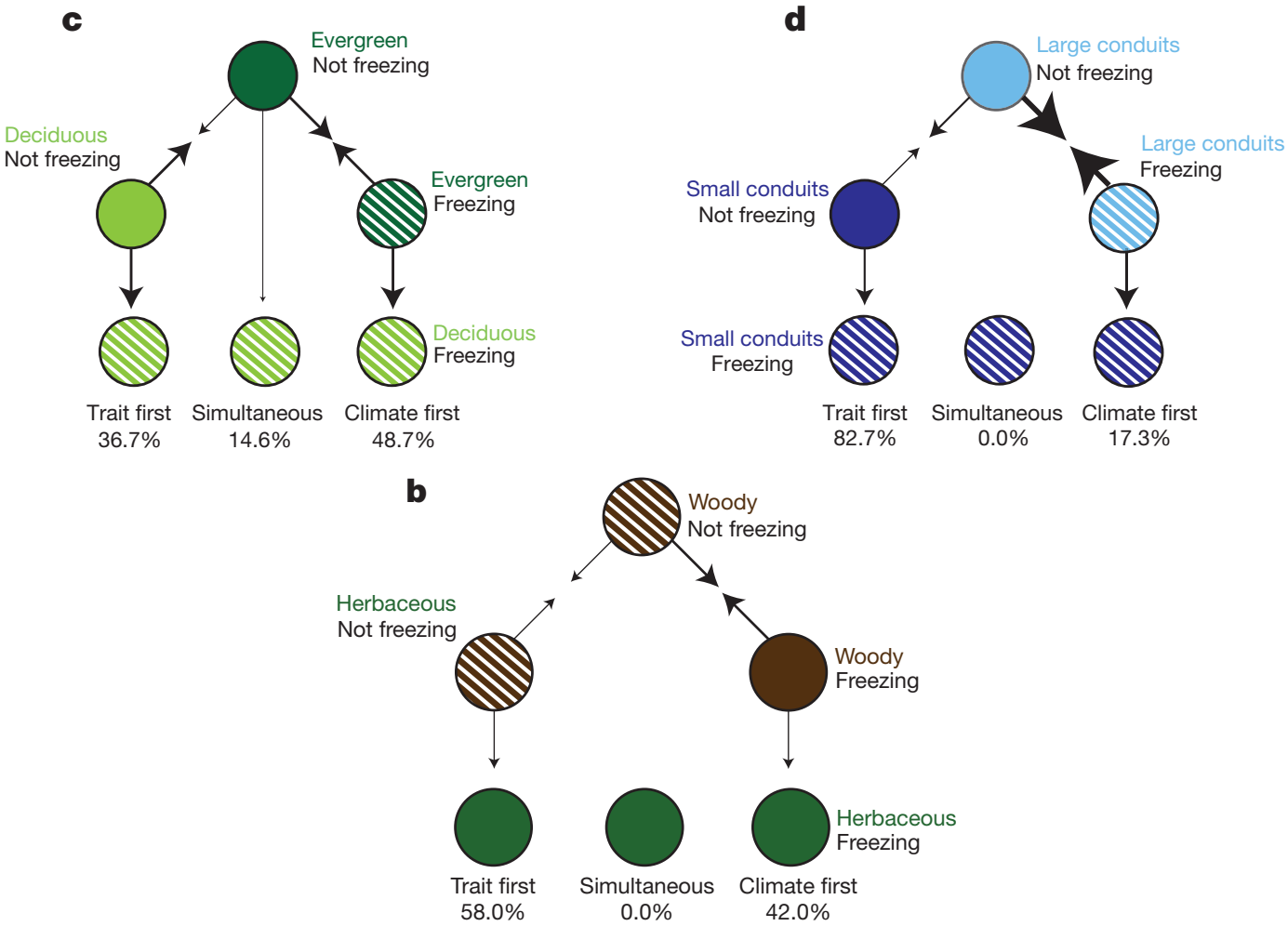

growth of genbank
150,000,000,000
156,230,531,562
base pairs
0.0e+00
5.0e+10
1.0e+11
1.5e+11
0
50
100
150
200
1992
2014
genbank release

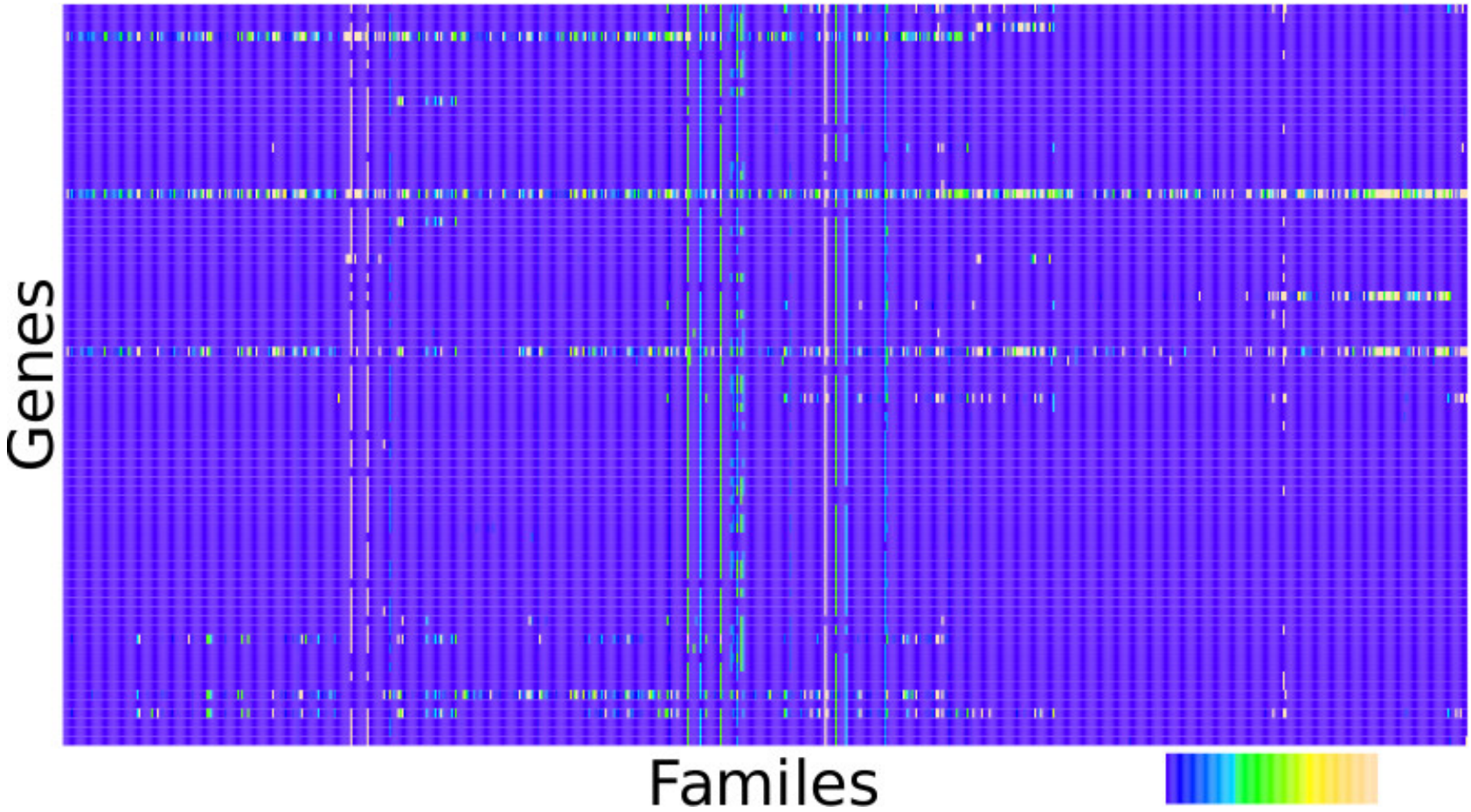
Do we have homology?

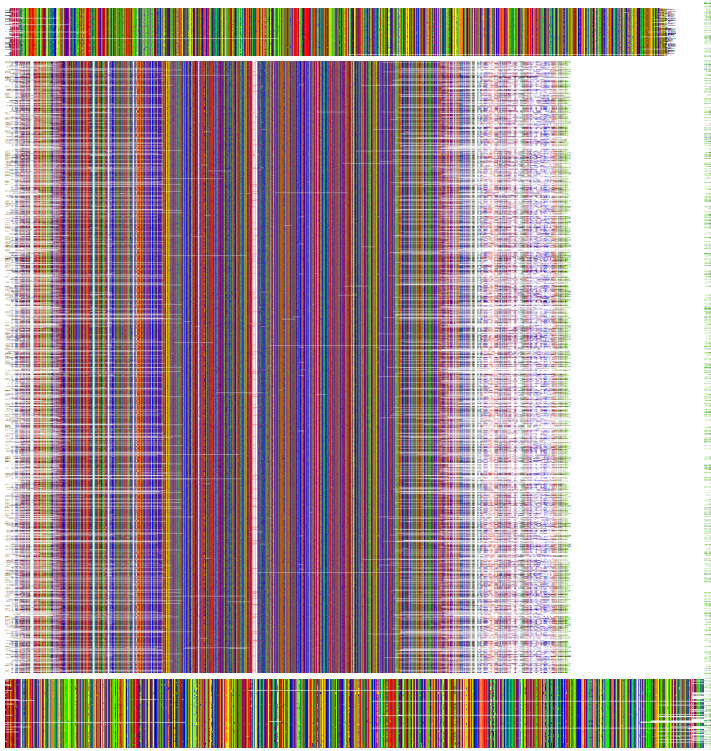
- GenBank stores species taxonomy but not gene taxonomy
- hard to get gene identifications for all of GenBank (e.g., clustering)
- all genes for species x
- NOT all species for gene y

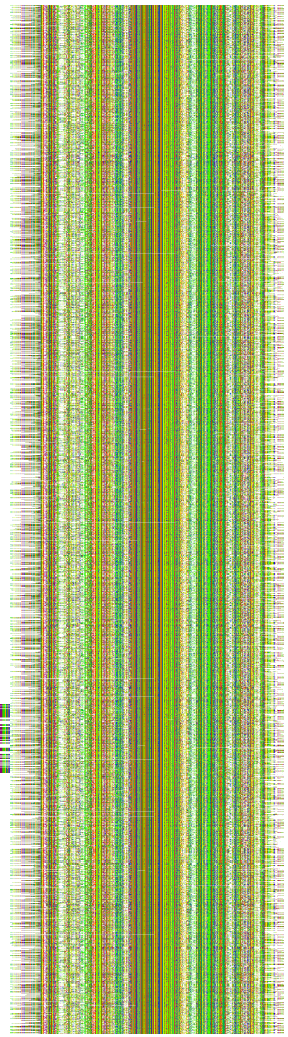


Zanne et al., Nature, 2013


c
Evergreen
Not freezing
Deciduous
Not freezing
Evergreen
Freezing
Deciduous
Freezing
Trait first
36.7%
Simultaneous
14.6%
Climate first
48.7%
d
Large conduits
Not freezing
Large conduits
Freezing
Small conduits
Not freezing
Small conduits
Freezing
Trait first
82.7%
Simultaneous
0.0%
Climate first
17.3%
b
Woody
Not freezing
Herbaceous
Not freezing
Woody
Freezing
Herbaceous
Freezing
Trait first
58.0%
Simultaneous
0.0%
Climate first
42.0%


Genes
Familes





trnL-trnF = 2485

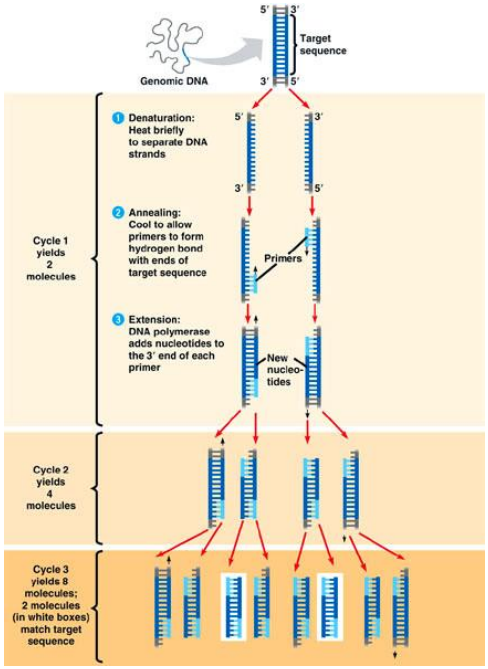
atpB = 107

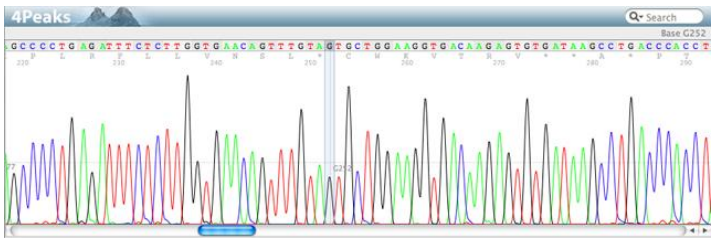
matK = 2182

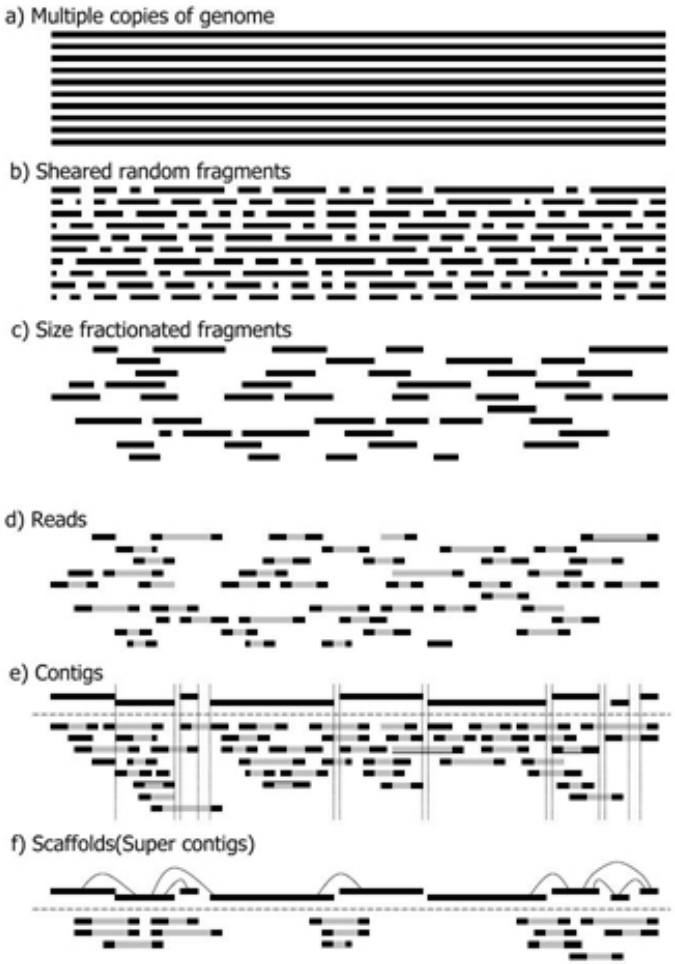
rbcL = 924

ITS = 3725

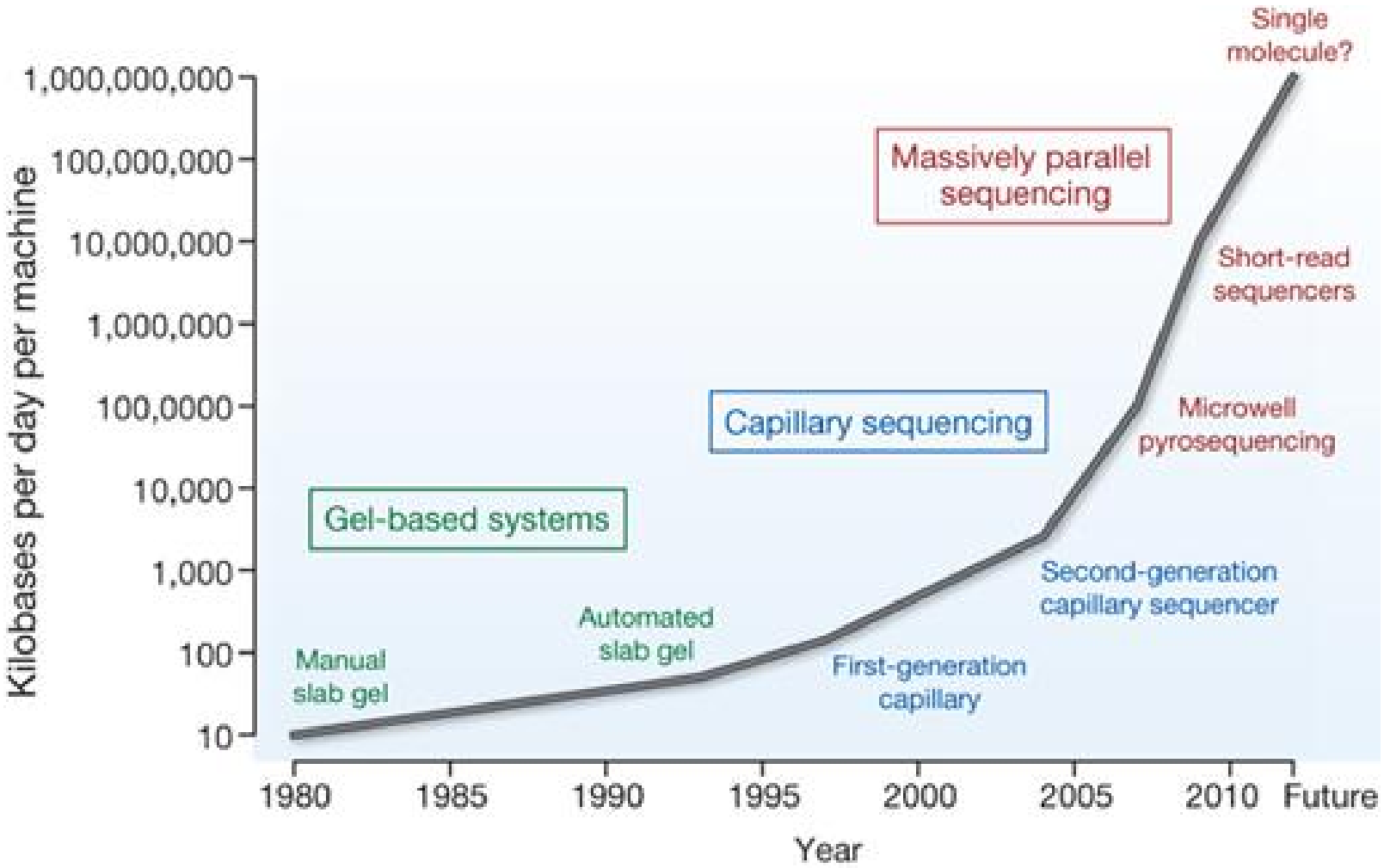
Next generation sequencing







New data


Kilobases per day per machine
1,000,000,000
100,000,000
10,000,000
1,000,000
100,0000
10,000
1,000
100
10
1980
1985
1990
1995
2000
2005
2010
Future
Year
Single molecule?
Massively parallel sequencing
Short-read sequencers
Capillary sequencing
Microwell pyrosequencing
Gel-based systems
Second-generation capillary sequencer
Automated slab gel
Manual slab gel
First-generation capillary

Number of genes

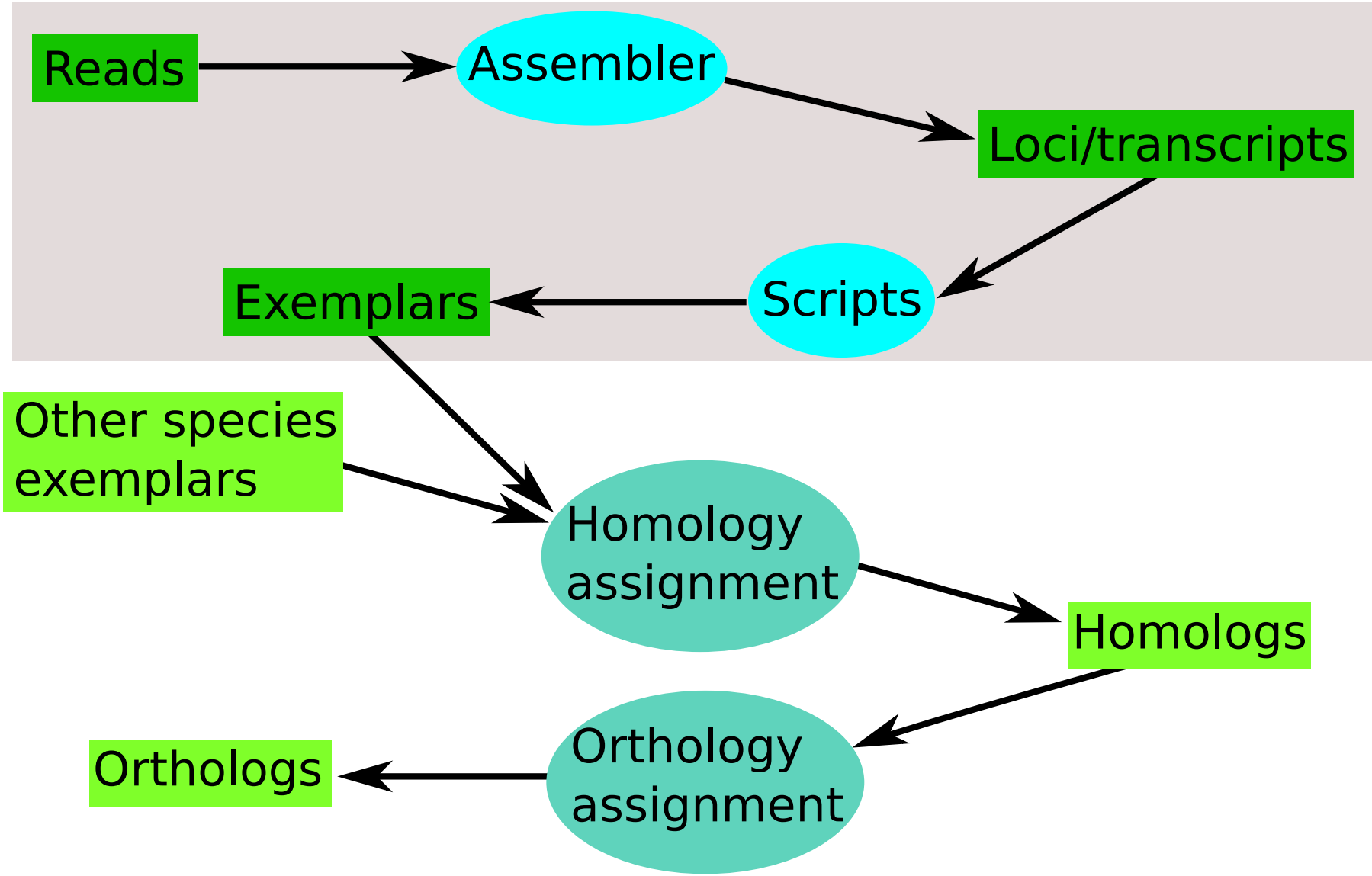
Typical phylogenetic analyses

- 1-10 genes
- 17 genes. Plants (Soltis et al. 2011)
- 19 genes. Birds (Hackett et al. 2007)

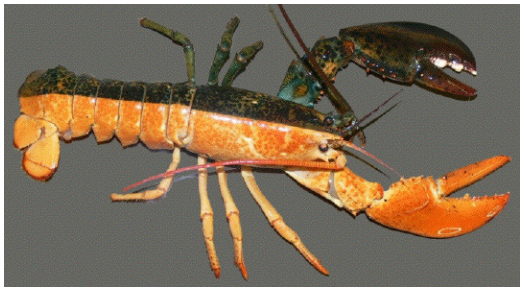
Transcriptomic and genomic phylogenetic analyses

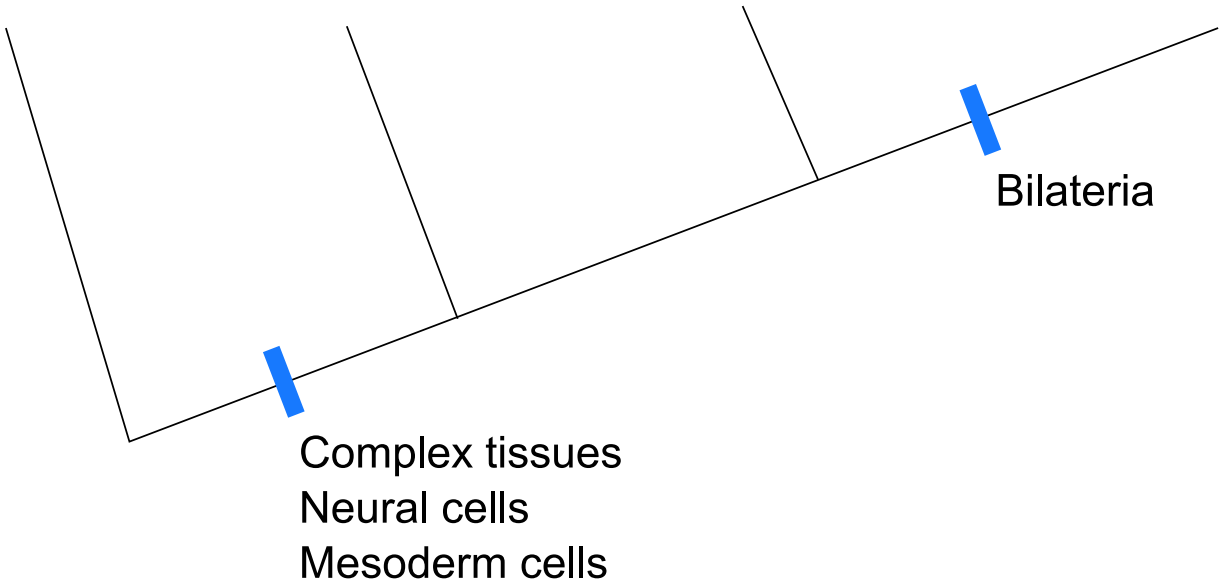
- 140 genes. Metazoa (Dunn et al. 2008)
- 248 genes. Turtles (Chiari et al., 2012)
- 1185 genes. Molluscs (Smith et al. 2011)
- 1720 genes. Rice (Cranston et al. 2007)
- 2970 genes. Seed plants (Lee et al. 2011)

Pipeline

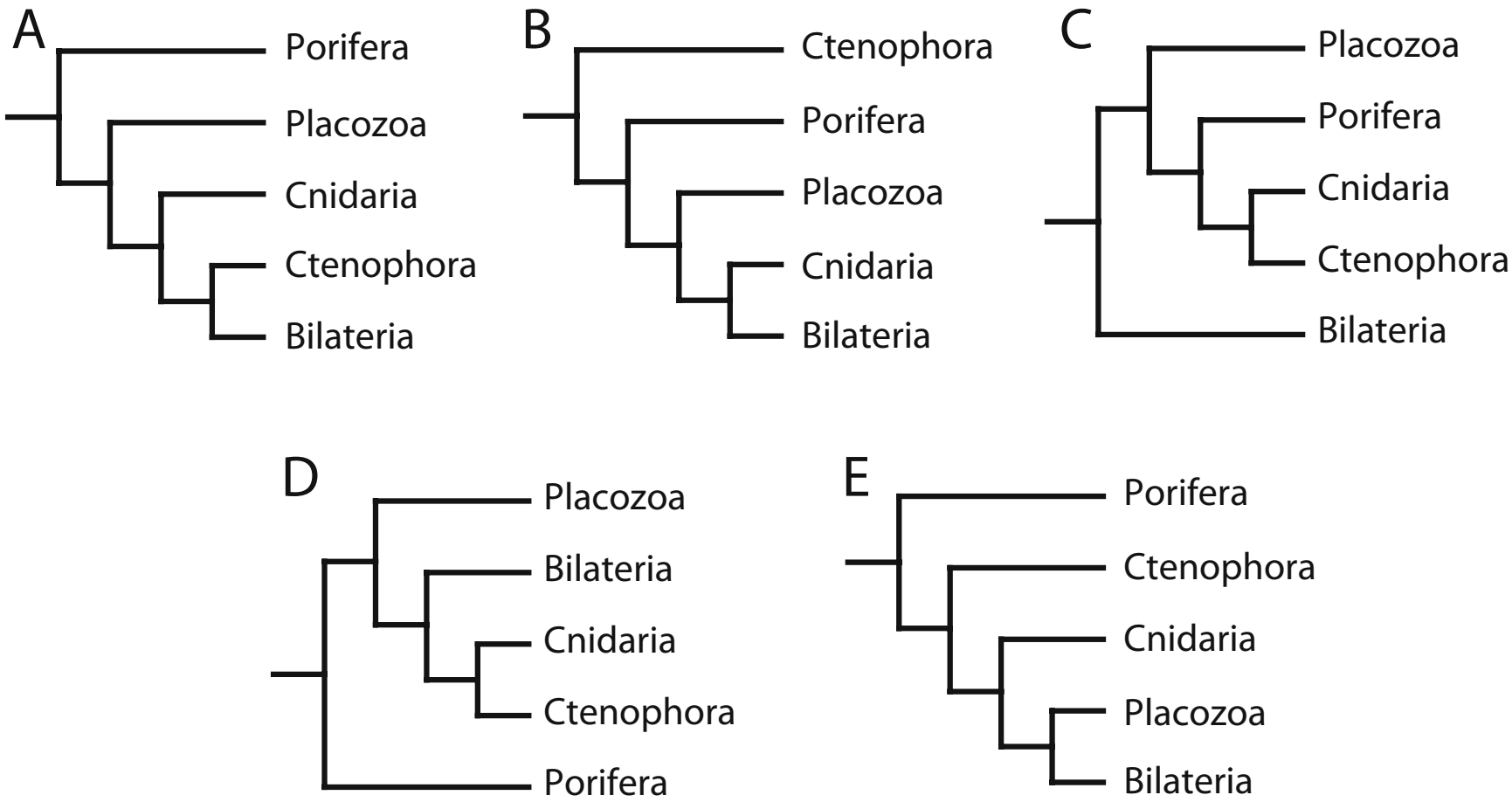

Reads
Assembler
Loci/transcripts
Scripts
Exemplars
Other species exemplars
Homology assignment
Homologs
Orthology assignment
Orthologs



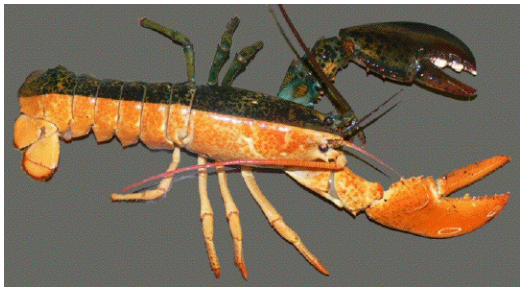

Bilateria
Complex tissues
Neural cells
Mesoderm cells

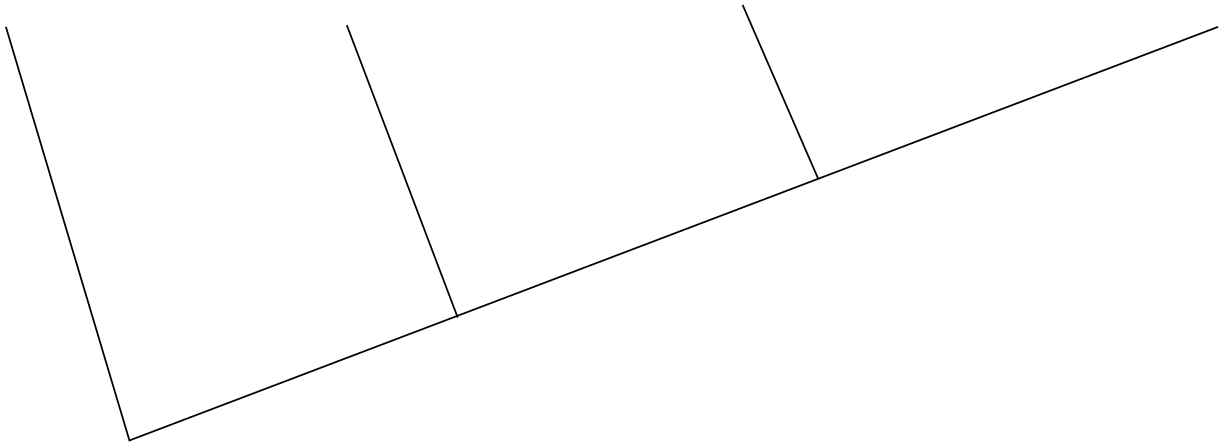
Disagreement in animals



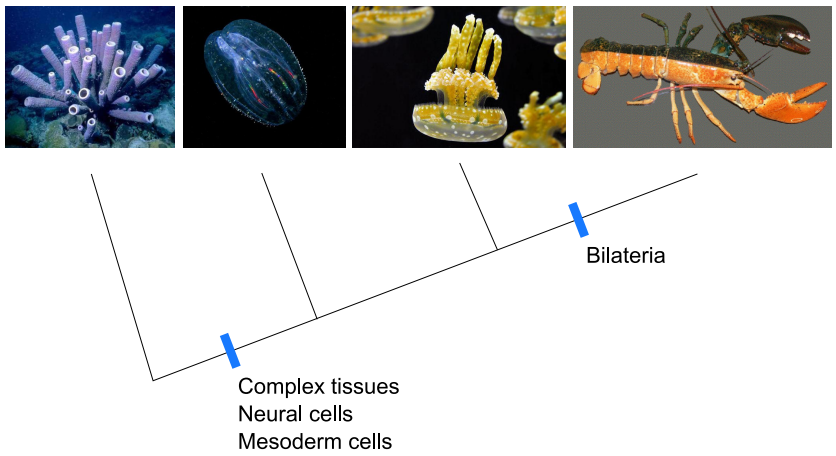
Edgecombe et al. 2011

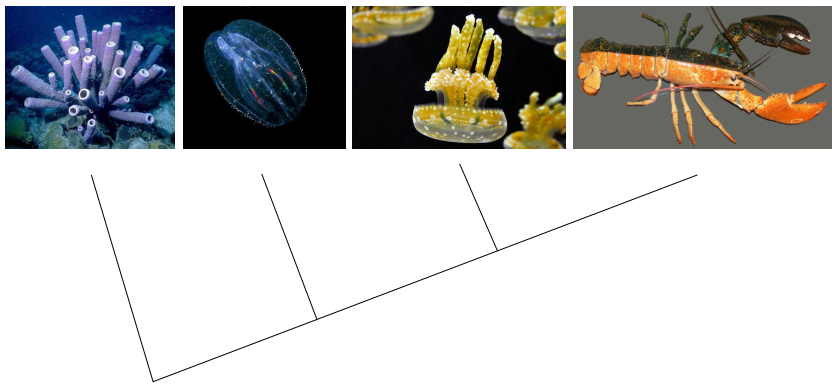




Alternatives

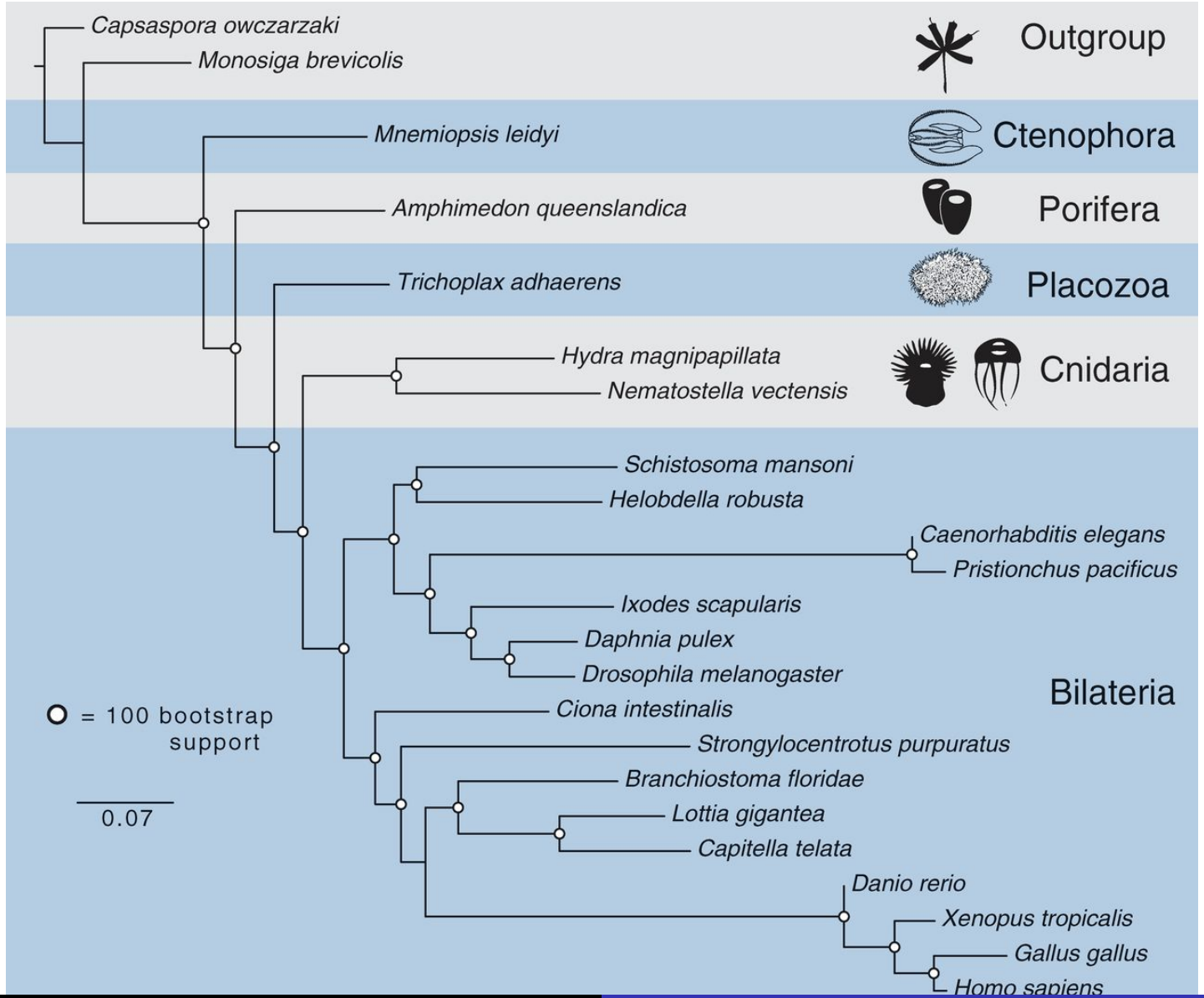


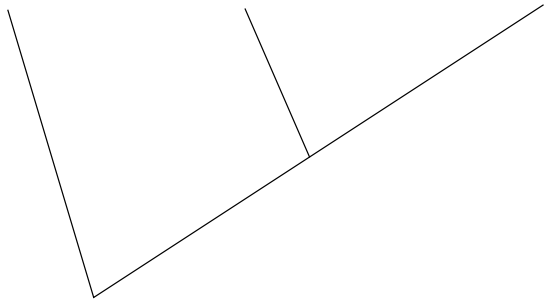


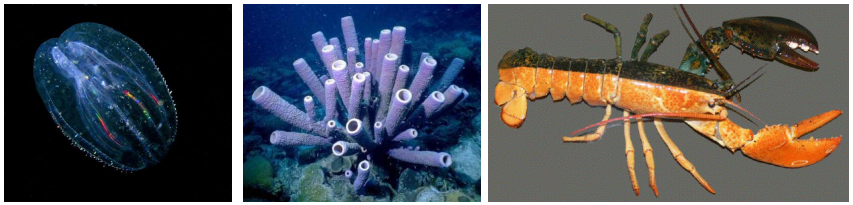
Traits

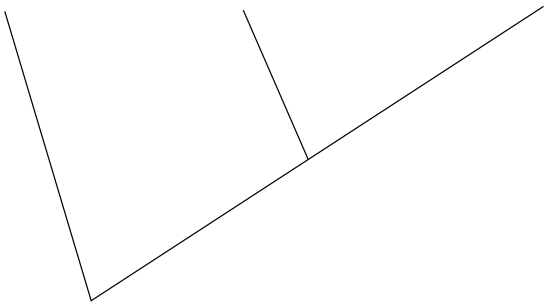
- tissue present
- cell types
- motility
- symmetry
- ...

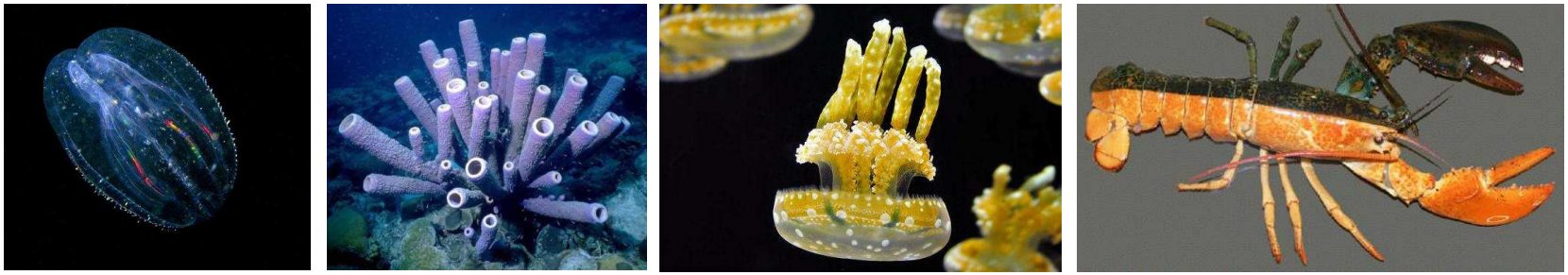
242 genes; Ryan et al., 2013

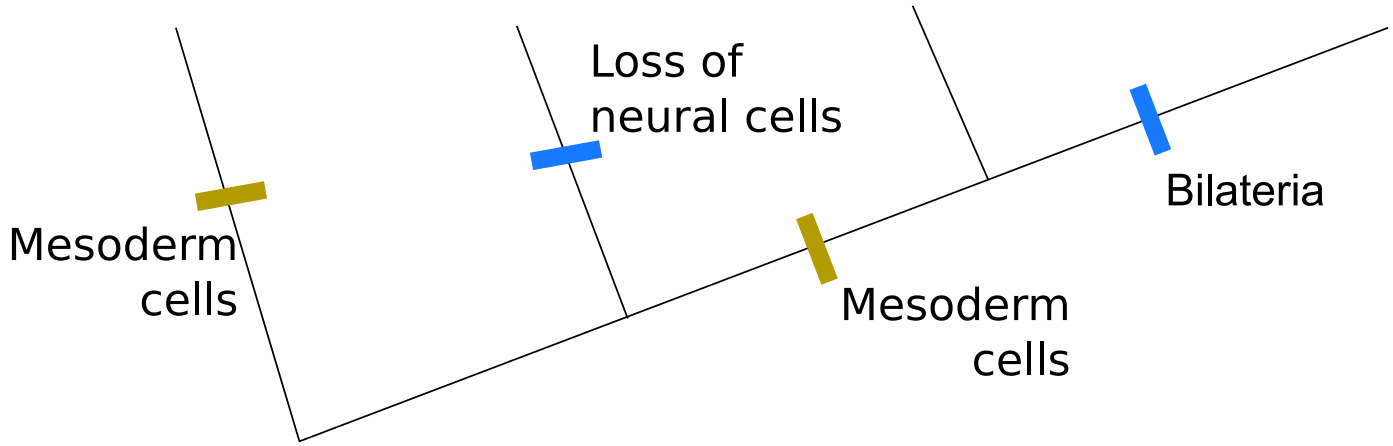






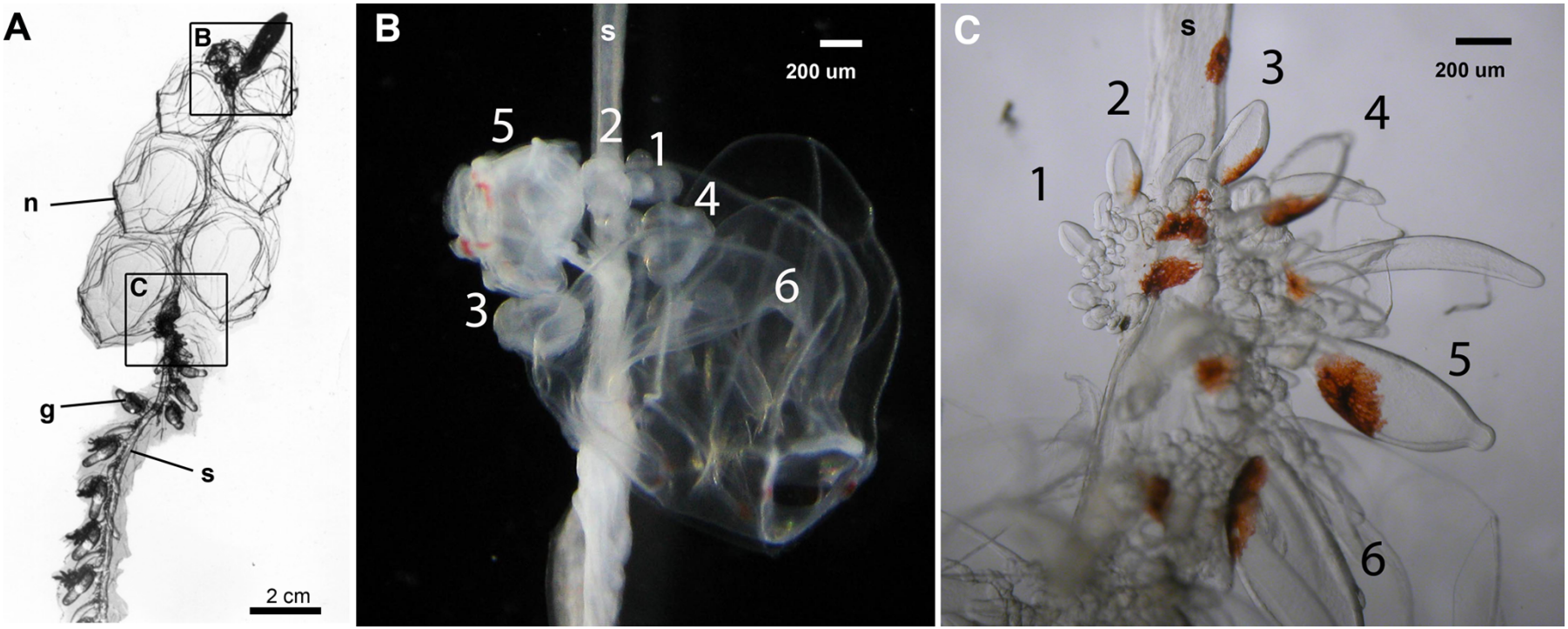





Loss of
neural cells
Bilateria
Mesoderm
cells
Mesoderm
cells

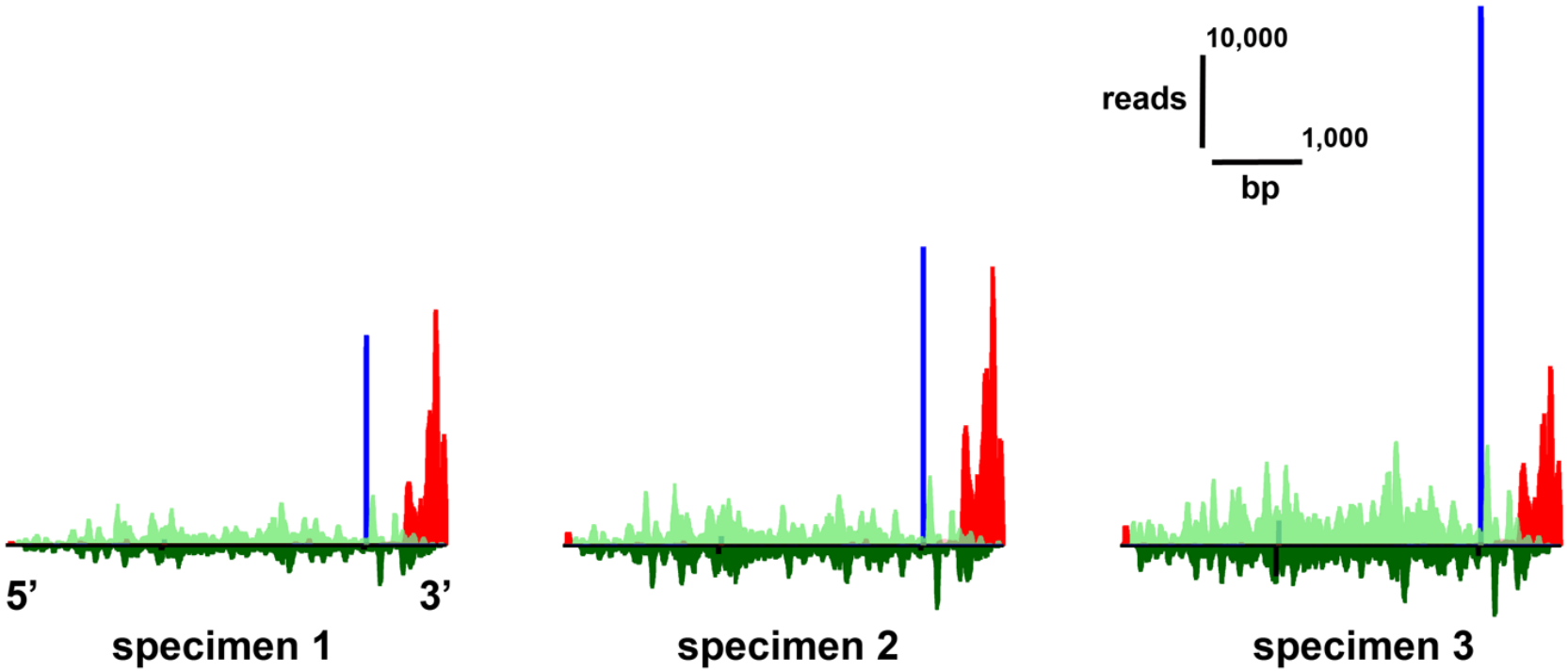
Next generation mapping

Next generation mapping



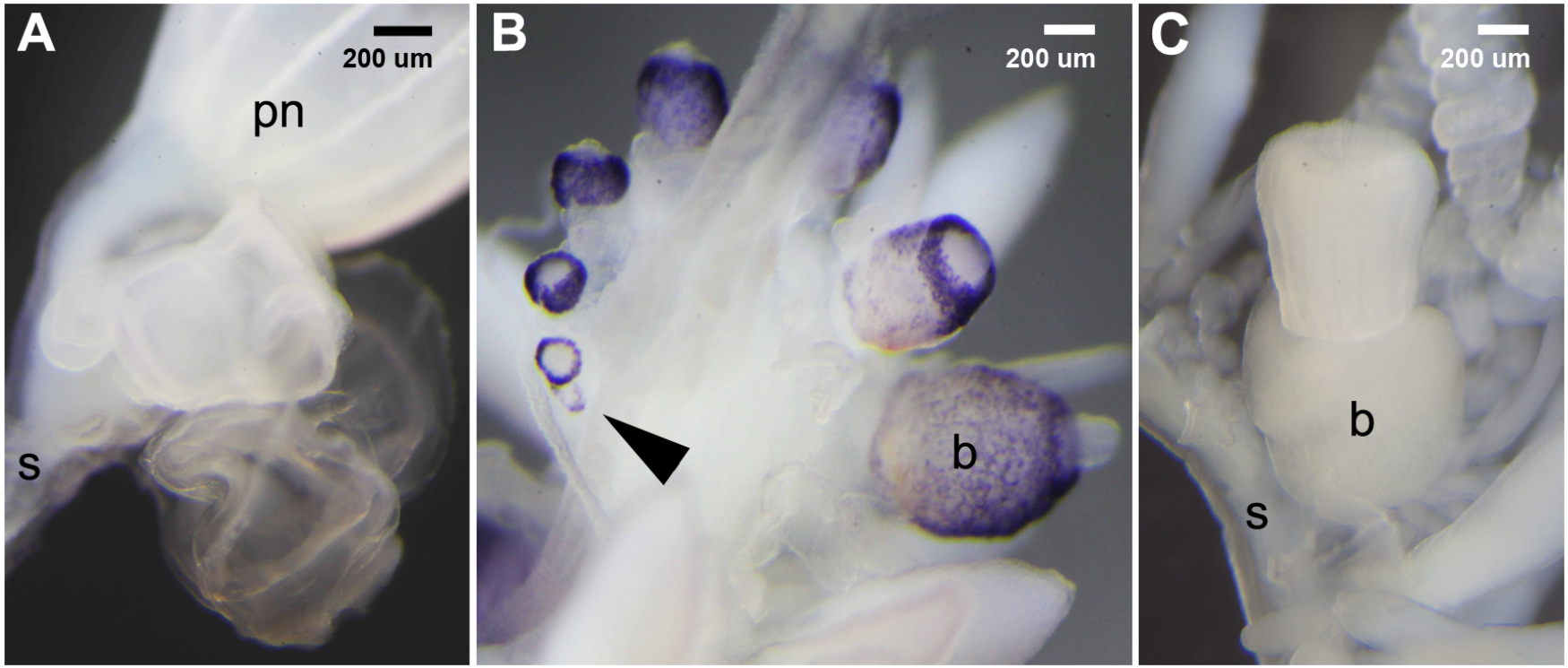
Seibert et al., 2011, PLoS

Next generation mapping



Seibert et al., 2011, PLoS

Next generation mapping



Seibert et al., 2011, PLoS

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

Major topics covered

- Detecting homologous sequences
  - pairwise sequence alignment
    - global $\rightarrow$ Needleman-Wunsch-Gotoh algorithm
    - local $\rightarrow$ Smith-Waterman algorithm
    - BLAST $\rightarrow$ heuristic search
- Grouping homologs
  - clustering
    - MCL and general clustering

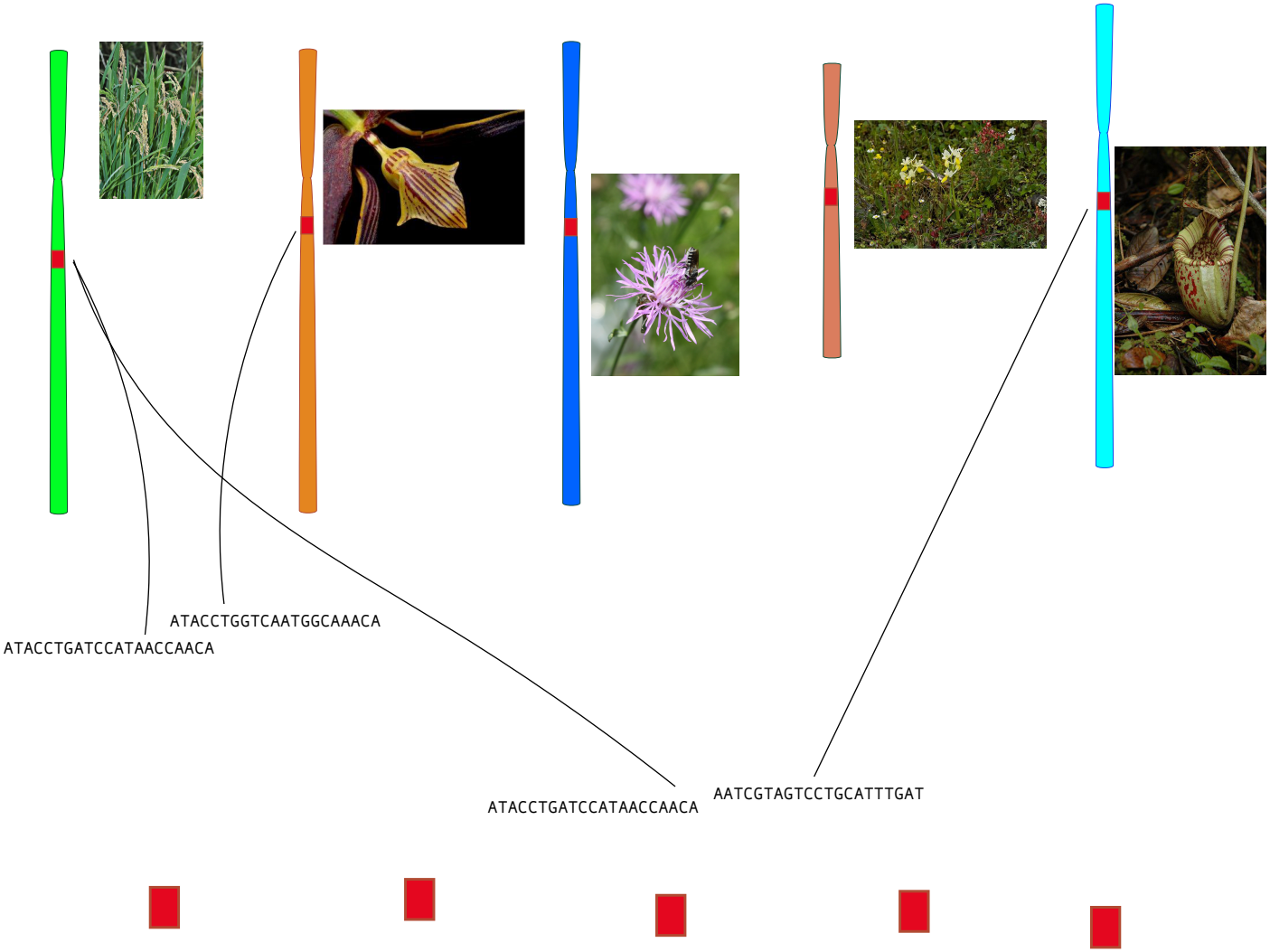
Schedule

- First
  - Detecting homologous sequences
  - Lab for detecting homologous sequences
- Second
  - Clustering sequences
  - Short lab for clustering sequences

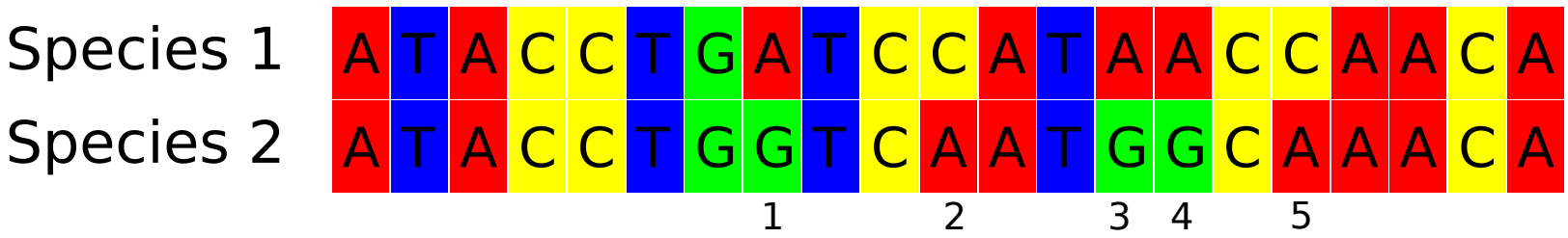
Pairwise alignment

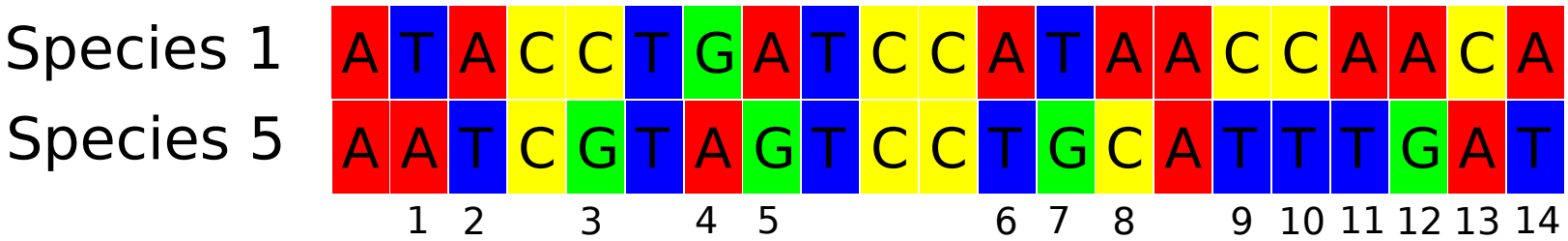
What are we going to address with pairwise alignment?

- Homology
  - Are these sequences homologous?
  - Do they share common ancestry?
  - Are they sufficiently similar?
- Bioinformatics
  - Are these sequences similar enough to include in the analysis?
  - Should a sequence be removed? Too short or too noisy (information poor)?


ATACCTGATCCATAACCAACA
ATACCTGGTCAATGGCAAACA
ATACCTGATCCATAACCAACA
AATCGTAGTCCTGCATTTGAT

Comparisons





Word aligning

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

Word aligning

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

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

Less Trivial

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

Less Trivial

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

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

Less Trivial

- Species 1: ACGTTAGA
- Species 2: CGTTGAA

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

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

Less Trivial

- `Species 1:  -------ACGTTAGA`
- `Species 2:  CGTTGAA--------`
  - score: -15 (gaps = -1, match = 1)
- `Species 1:  ACGTTAGA-`
- `Species 2:  -CGTT-GAA`
  - score: 3

You do one

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

Pairwise 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

You do one

- `Species 1:  TTGGCACGTTAGA`
- `Species 2:  TGCACCTTAGTTA`
- +1 match, -1 mismatch

NW

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

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

More complicated scoring matrices

- don't just have to have match and mismatch
- this is the 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

More complicated gap models

- constant gap model
  - a gap has 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
  - parameter for **extension** of a gap
  - match = 1, open = −2, ext = −1, score = -1 (+5,-2+(-1*4))
  - you will see these parameters and this is what they mean

More complicated gap models

- `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

- Protein (amino acid) alignments are algorithmically no 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 1990's
  - BLOSUM
    - has generally replaced PAM as the common matrix

PAM

- includes PAM 250, PAM 120, PAM 1
- based on evolutionary models and empirical data
- refers to evolutionary difference
  - higher the number, more divergent
- ranges from closely related to completely random

	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

- includes BLOSUM 45, BLOSUM 62, BLOSUM 80
- based on empirical data
- refers to percent identity
  - higher the number, *less* divergent
- narrower range than PAM

	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

- PAM120, BLOSUM30
- **A**-**A**: 3 (P), 4 (B)
- **A**-**R**: -3 (P), -1 (B)
- **R**-**N**: -1 (P), 0 (B)

	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`

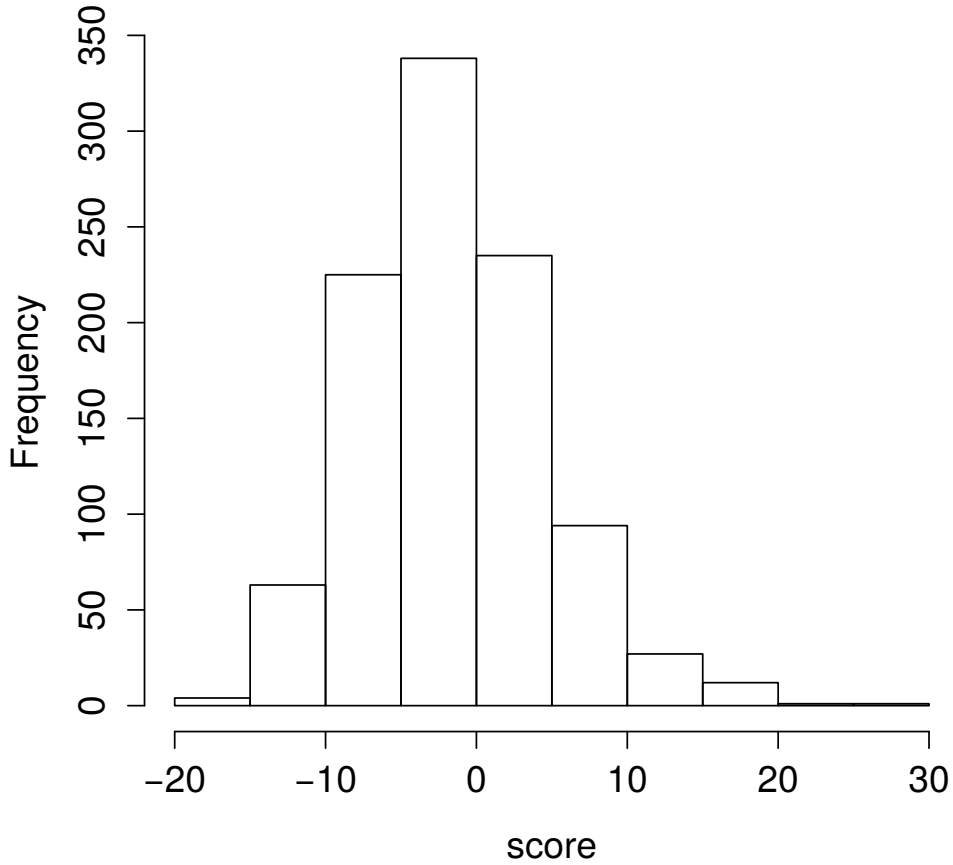
Significance

- Needleman-Wunsch will always give the best alignment (given the assumptions)
- How do we measure whether an alignment is significant?

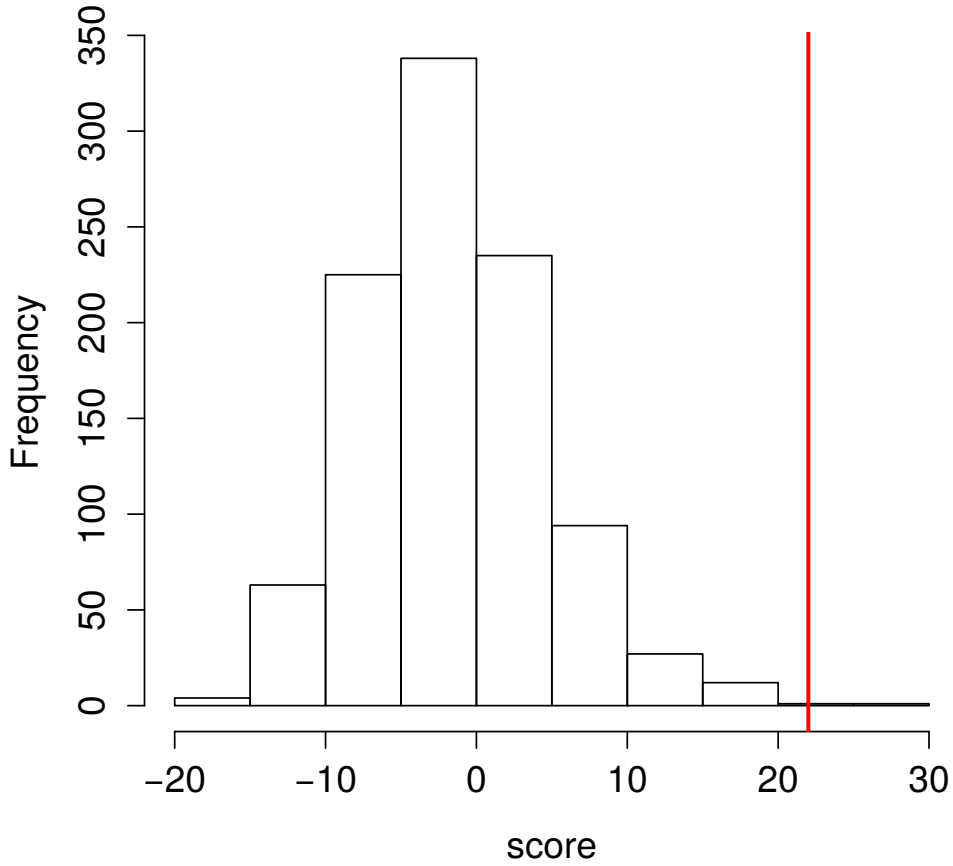
Significance

- How do we measure whether an alignment is significant?
- There is no good theory allowing us to predict the distribution of alignment scores from random sequences
- In practice, we create our own distribution
  - generate random alignments and get scores
  - compare our score to the random distribution
  - if 100 random alignments give scores that are lower than the observed alignment score, our p-value should be less than 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 vs Local

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

Global vs Local

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

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

Global vs Local

- `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 7700 times
- Smith-Waterman is the standard local alignment algorithm
  - published 1981
  - cited over 6800 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

- looking at amino acids
- 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`

Difference

- Species 1: HEAGAWGHEE
- Species 2: PAWHEAE

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

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

How can we detect homology?

- What would be the properties of homologous sequences?
  - sequences that are similar ***enough*** are homologous
  - find the best alignment defining similarity
  - calculate alignment scores
- What constitutes similar enough?
  - more similar than by chance?
  - what defines chance?

Calculating significance

- We have calculated the optimal alignment
  - the alignment with the best score
  - this doesn't depend on whether the sequences are related or not
  - call this the maximum segment pair (MSP)
- How many MSPs do we expect with at least the same score by chance?

Calculating significance

- We make use of the extreme value distribution (EVD) to calculate the number of alignments between random sequences that we expect given our score or better
- This is known as the *E*-value or the number of distinct alignments with a score equal or better than our score
  - $E(S) = Kmne^{-\lambda S}$
    - K and λ = scaling parameters calculated based on the search space (K) and scoring scheme (λ)
    - $m * n$ = size of the search space
- The probability of finding at least one alignment with our score (the p value)
  - $1 - e^{-E(S)}$
- As both E and p values decrease, the biological significance increases

Calculating significance

- $m = 980, n = 10,030,834,086, K = 1.37, \lambda = 0.711$
- $m \times n \sim 10^{13}$

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


SONY
PS3
SONY
SELECT
START


SWEET BEAN

Software

Highly accessed Open Access

CBESW: Sequence Alignment on the Playstation 3

Adrianto Wirawan*, **Chee K Kwoh**, **Nim T Hieu** and **Bertil Schmidt**

* Corresponding author: Adrianto Wirawan adri0004@ntu.edu.sg

▾ Author Affiliations

School of Computer Engineering, Nanyang Technological University, Singapore
For all author emails, please log on.

BMC Bioinformatics 2008, **9**:377 doi:10.1186/1471-2105-9-377

The electronic version of this article is the complete one and can be found online at:
http://www.biomedcentral.com/1471-2105/9/377

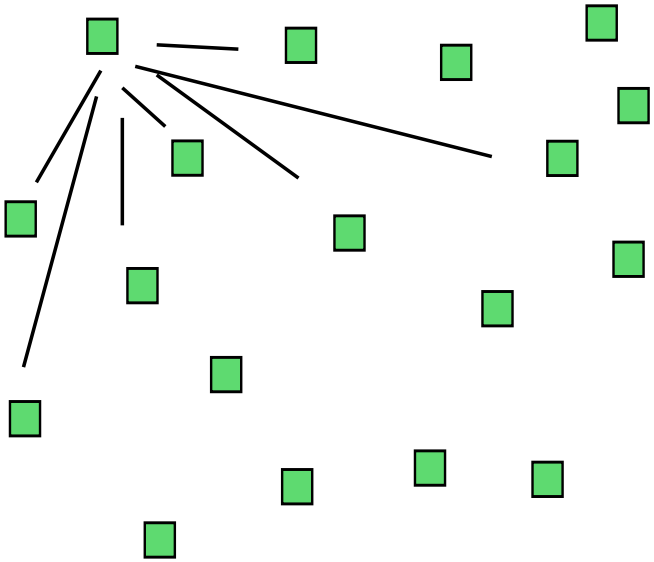
Received: 22 April 2008
Accepted: 17 September 2008
Published: 17 September 2008

Abstract

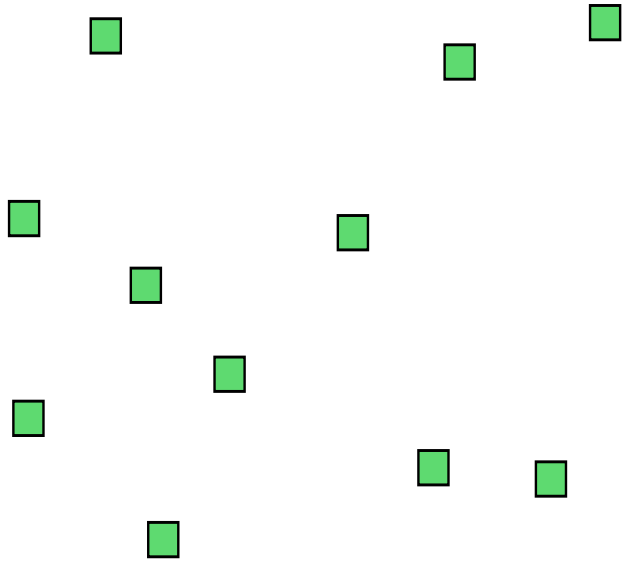
Formula display: MathJax

Sets of sequences

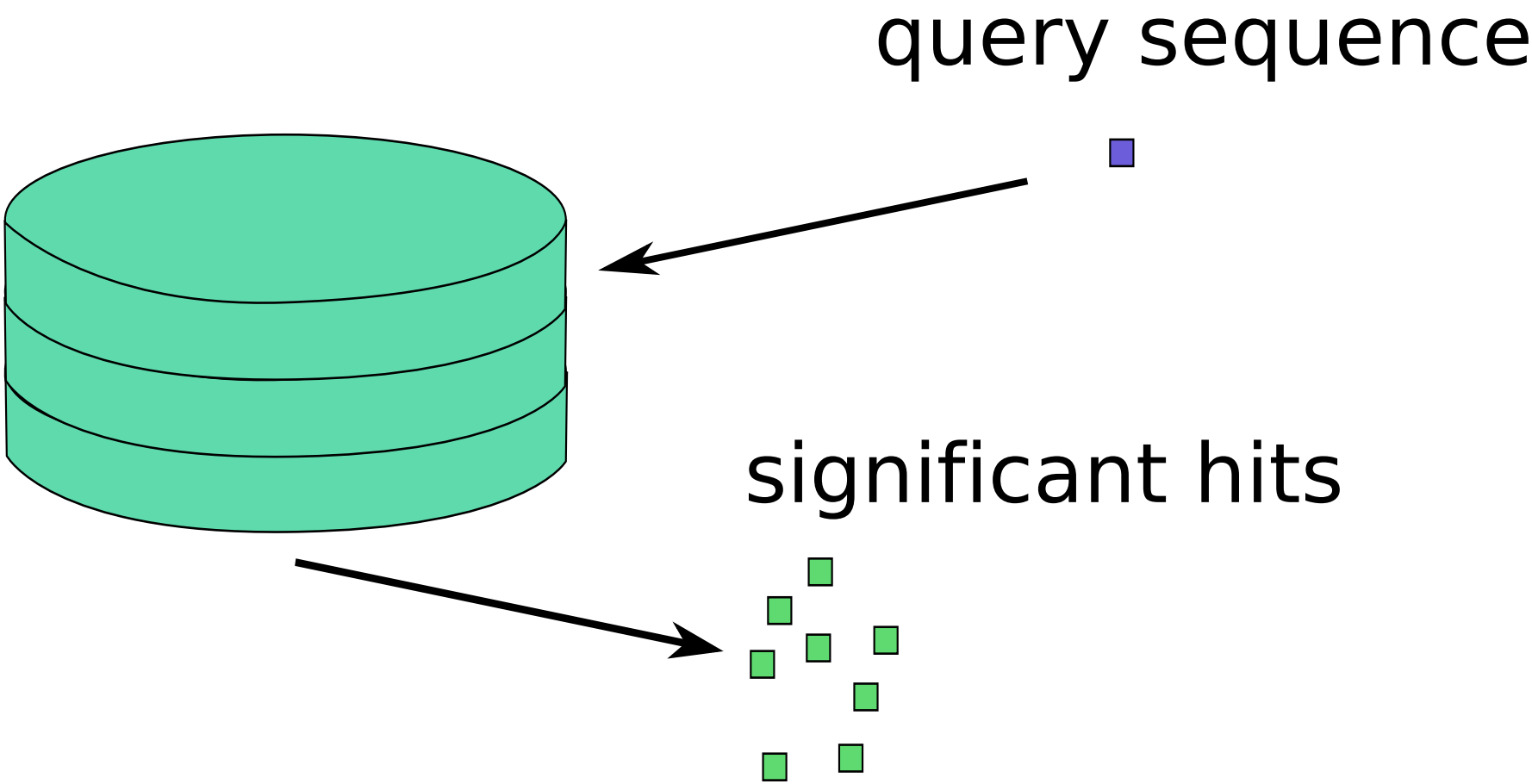
sequences



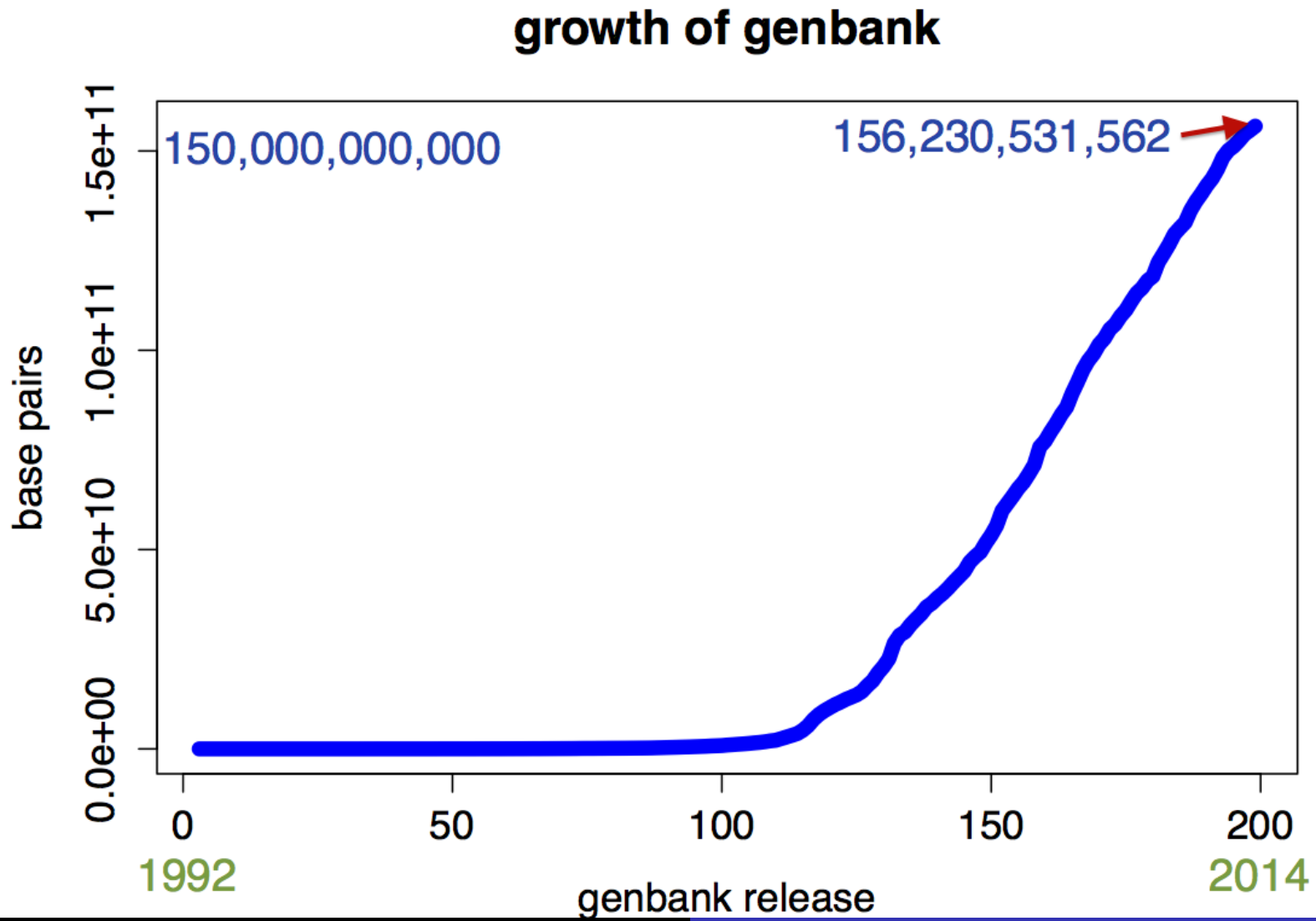
significant hits



From a database



NCBI


growth of genbank
150,000,000,000
156,230,531,562
base pairs
0.0e+00
5.0e+10
1.0e+11
1.5e+11
0
50
100
150
200
1992
2014
genbank release

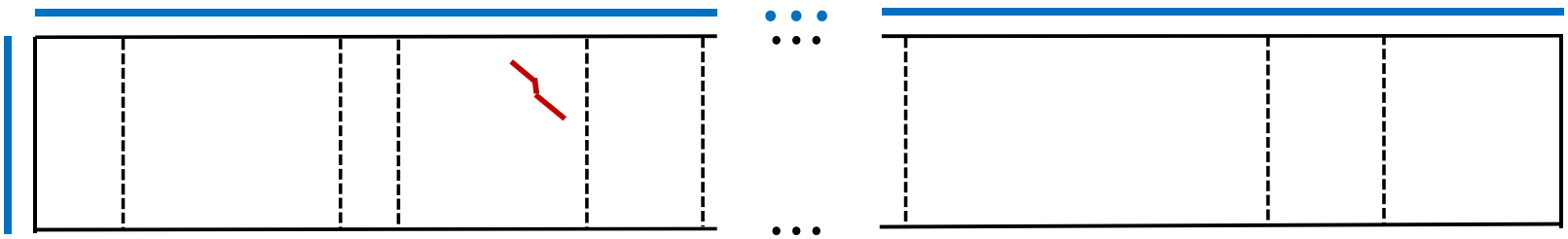
- abfghjlkdfmwholovesjetpacksetkfderteiuosdfjlkjasdfevcvxvjz
- zfdlkreoivpoqweoflfgytmgnhslkfdfmwholovesjetpackspldggjkjj

- abfghjlkdfmwholovesjetpacksetkfderteiuosdfjlkjasdfevcvxvjz
- zfdlkreoivpoqweoflfgytmgnhslkfdfmwholovesjetpackspldggjkjj

- abfghjlkdfmwholovesjetpacksetkfderteiuosdfjlkjasdfevcvxvjz
- zfdlkreoivpoqweoflfgytmgnhslkfdfmwholovesjetpackspldggjkjj

Smith-Waterman problems

- with Smith-Waterman, we know we are getting the best alignment
- spends much time in areas of the alignment that have no biological relevance



BLAST

- Basic Local Alignment Search Tool
  - Before the development of PS3, SW was too slow for local alignment
  - While getting to the human genome, needed something faster (50-100×)
  - Developed BLAST (Altschul et al., 1990, cited over 43,000 times)
  - Heuristic that produces an **approximate** best match (SW is a guarantee)
    - calculate the high scoring matches instead of the maximum scoring matches (HSP instead of MSP)

BLAST (1)

- 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**

AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

ATTTTTTTTTTTTTTTTAAAAAAAAAAAAAAAAAAAAA

BLAST (2)

- 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`
  - ...

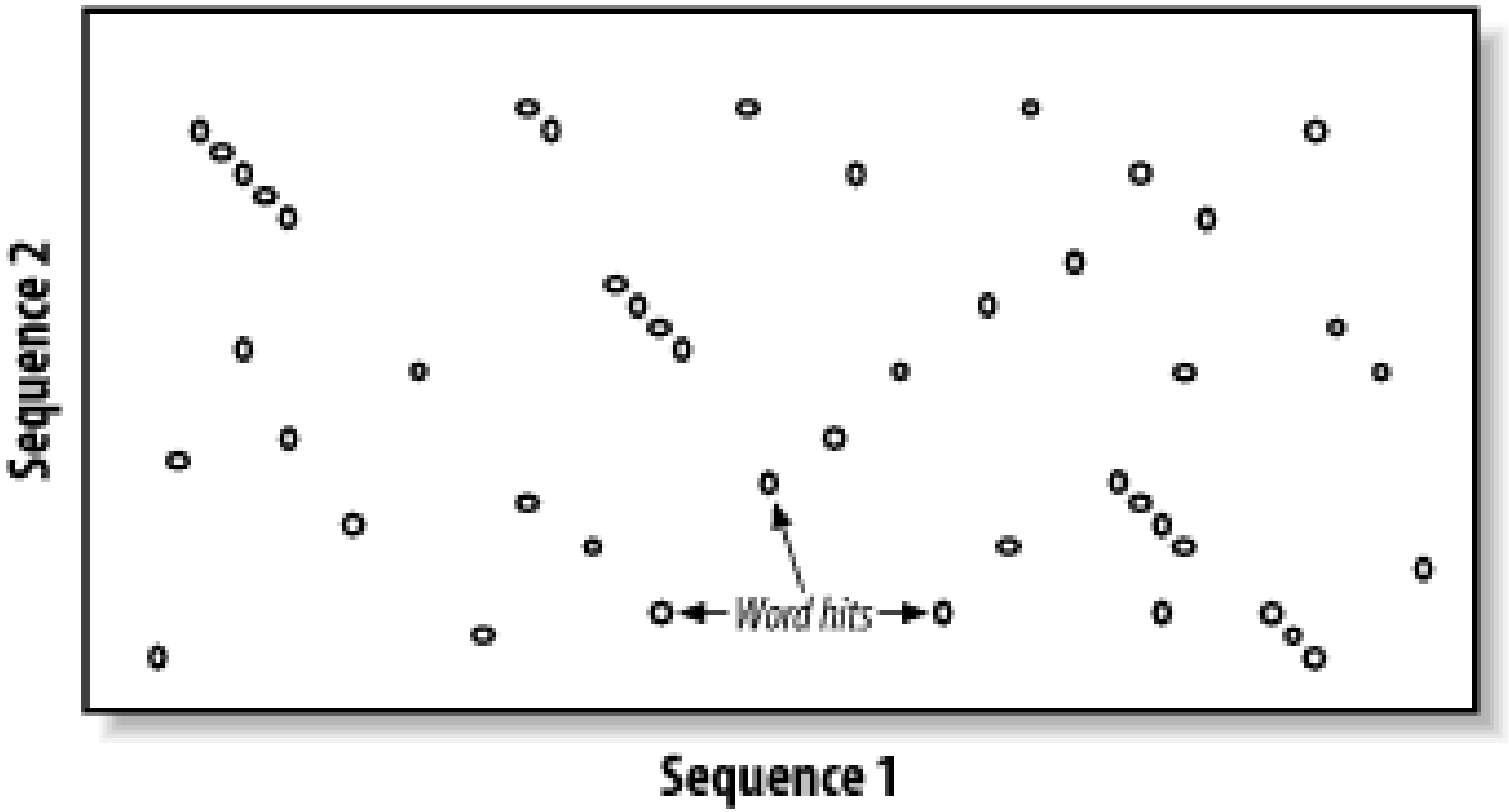
BLAST (3)

- for each of the words, look at "likely" mutants (based on scoring matrices)
- you could call this the "neighborhood"

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

BLAST (4)

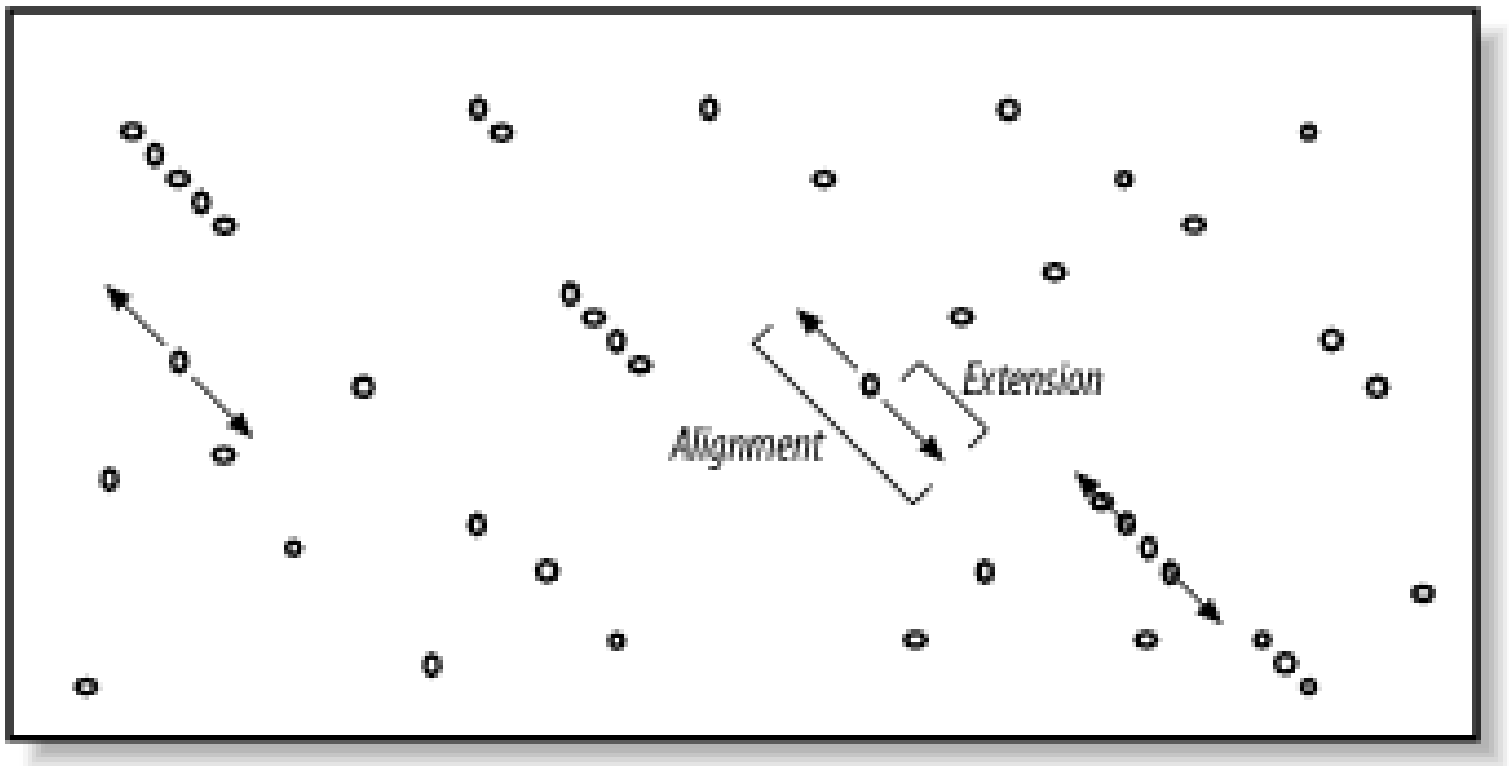
- organize the words into a form best for searching
- scan the other sequence for words that match



from Korf et al.

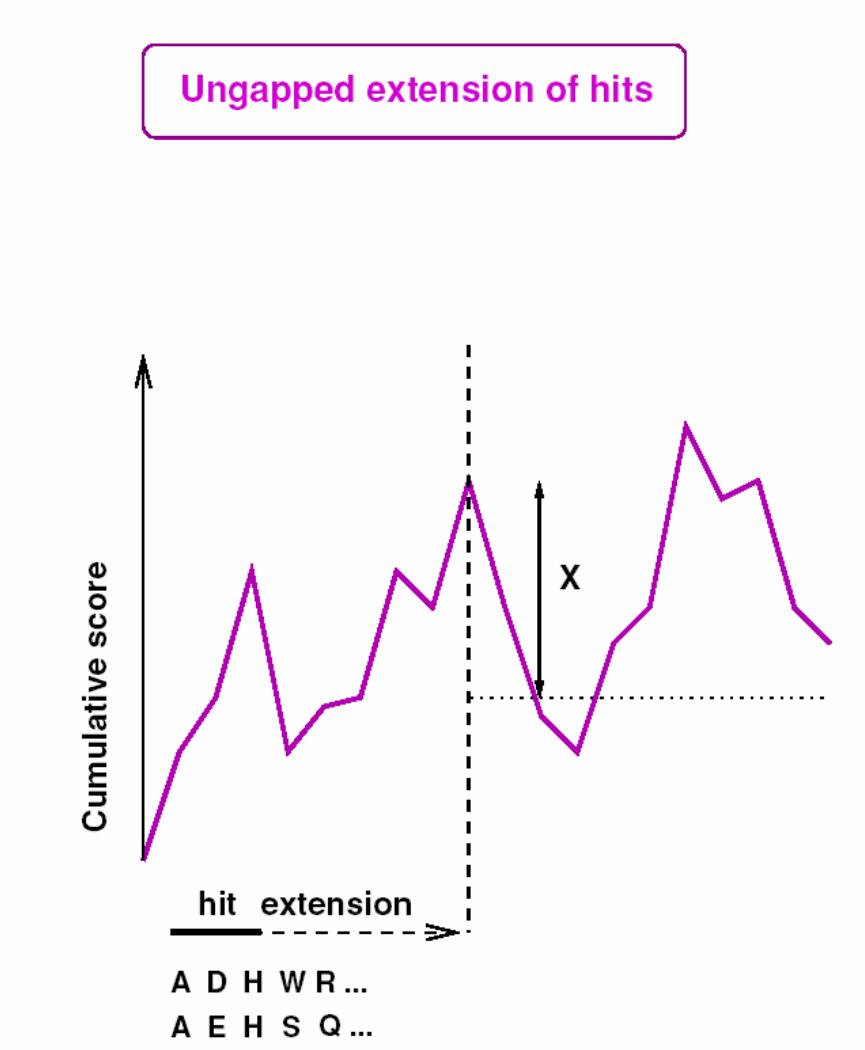
BLAST (5)

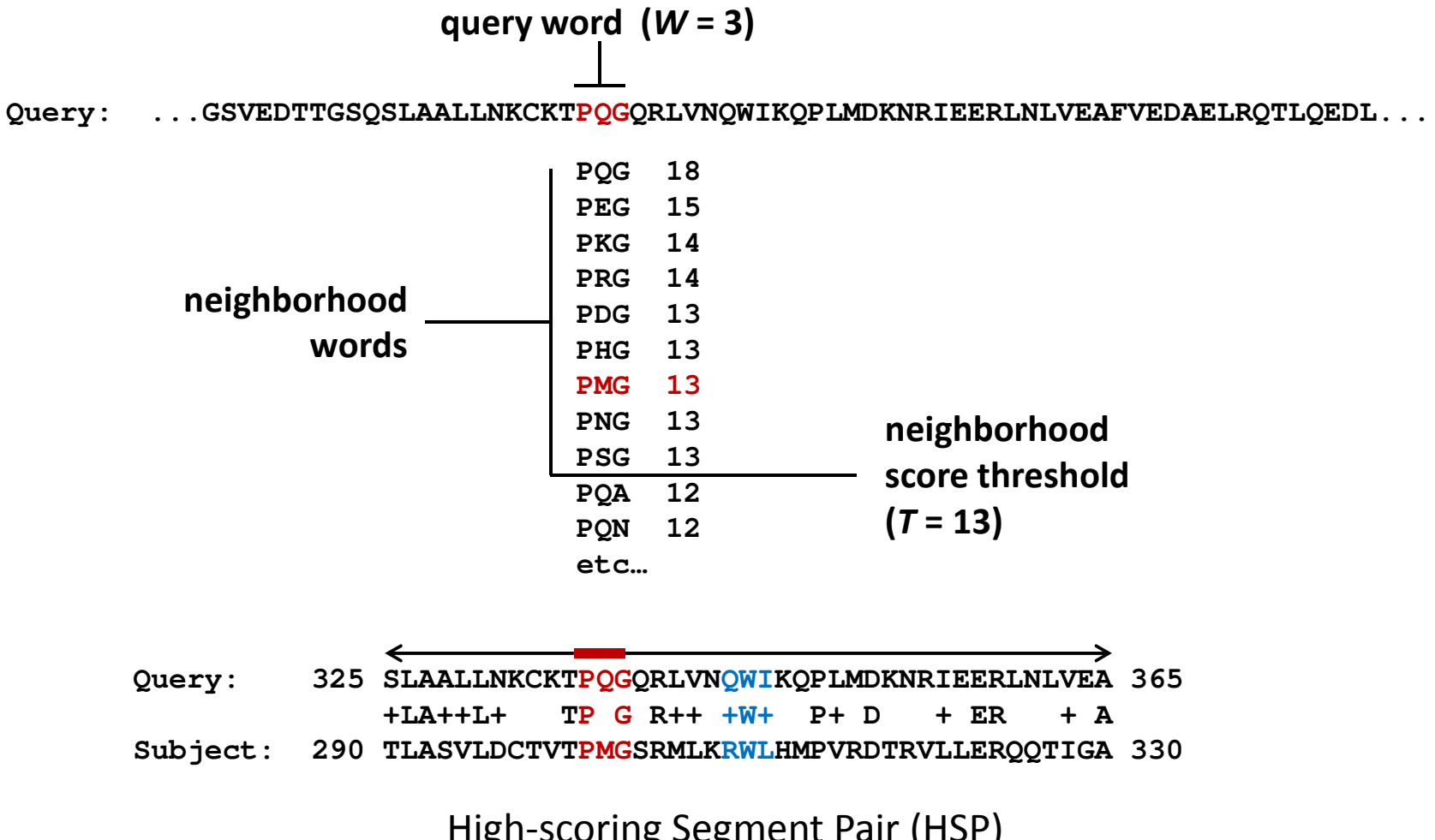
- extend these matches in the local neighborhood (these are going to be HSP or high scoring segment pairs)



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 (6)

- calculate *E*-values
  - expectation that you would get that alignment by change given the database of sequences
- return significant results
- 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
- These searches can be between a set of sequences we have determined earlier or they can be a database of sequences (that are not ours)
- First lets look at results from our own sequences

```
>18S_Abelia
NNNNNNNNNNNNNNNNNNNGTAGTCATATGCTTGTCTCAAAGATTAAGCCATGCATGTGTAAGTATGAACTAATTCAGACTGTGAAAC
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
NNNNNNNNNNNNNNNNNNNNAATTGACGGAAGGGCACCACCAGGAGTGGAGNCTGCGGCTTAATTTGACTCAACACGGGGAAACTTA
CCAGGTCCAGACATAGCAAGGATTGACAGACTGAGAGCTCTTTCTTGATTCTATGGGTGGTGGTGCATGGCCGTTCTTAGTTGGTGGA
GCGATTTGTCTGGTTAATTCCGTTAACGAACGAGACCTCAGCCTGCTAACTAGCTACGTGGAGGTACTCCTCCACGGCCAGCTTCTTA
GAGGGACTATGGCCGCTTAGGCCNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
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 (1)

- 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**

AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA

ATTTTTTTTTTTTTTTTAAAAAAAAAAAAAAAAAAAAA

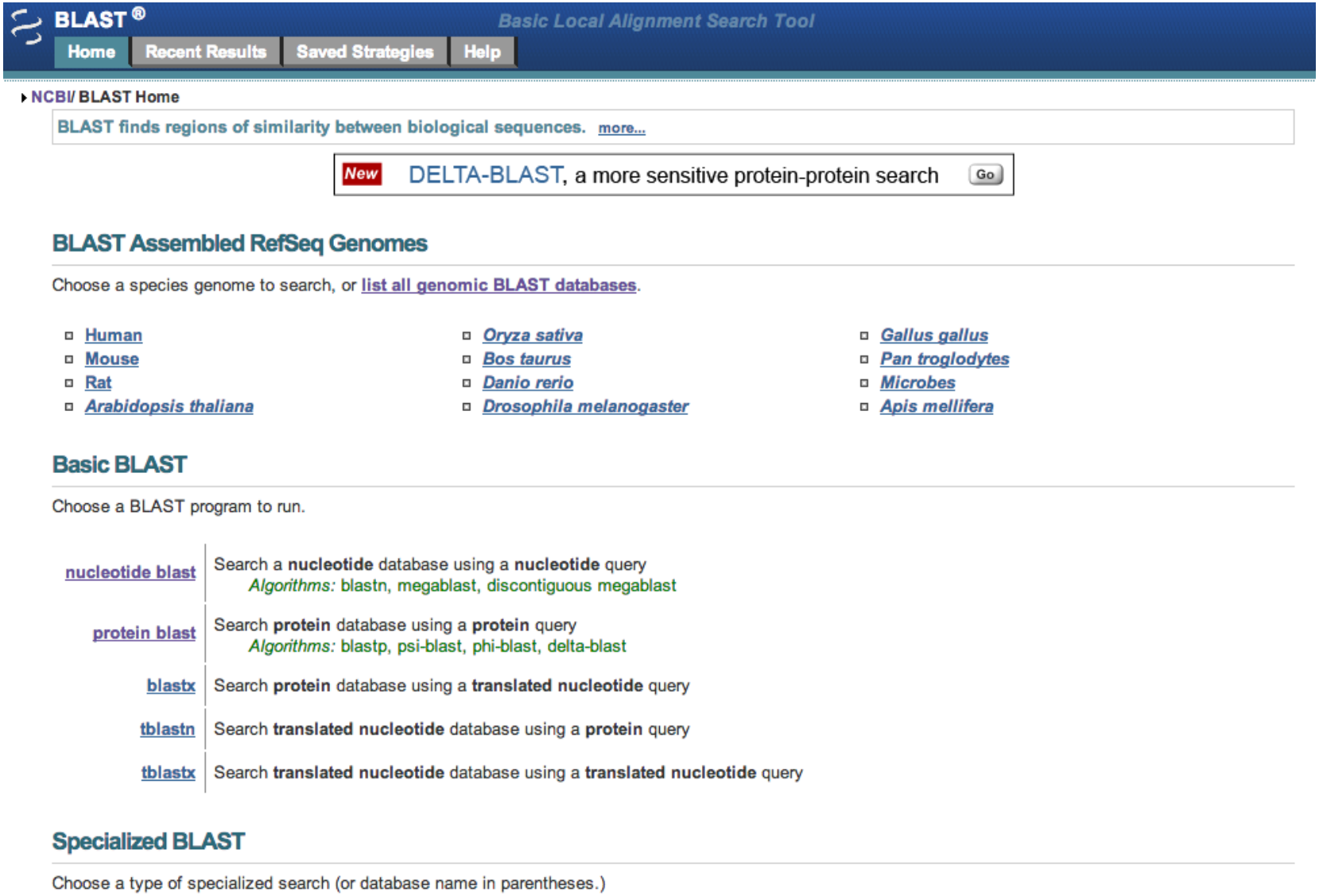
BLAST databases

- 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 Home

BLAST finds regions of similarity between biological sequences. more...

New DELTA-BLAST, a more sensitive protein-protein search Go

BLAST Assembled RefSeq Genomes

Choose a species genome to search, or list all genomic BLAST databases.

- Human
- Mouse
- Rat
- *Arabidopsis thaliana*
- *Oryza sativa*
- *Bos taurus*
- *Danio rerio*
- *Drosophila melanogaster*
- *Gallus gallus*
- *Pan troglodytes*
- *Microbes*
- *Apis mellifera*

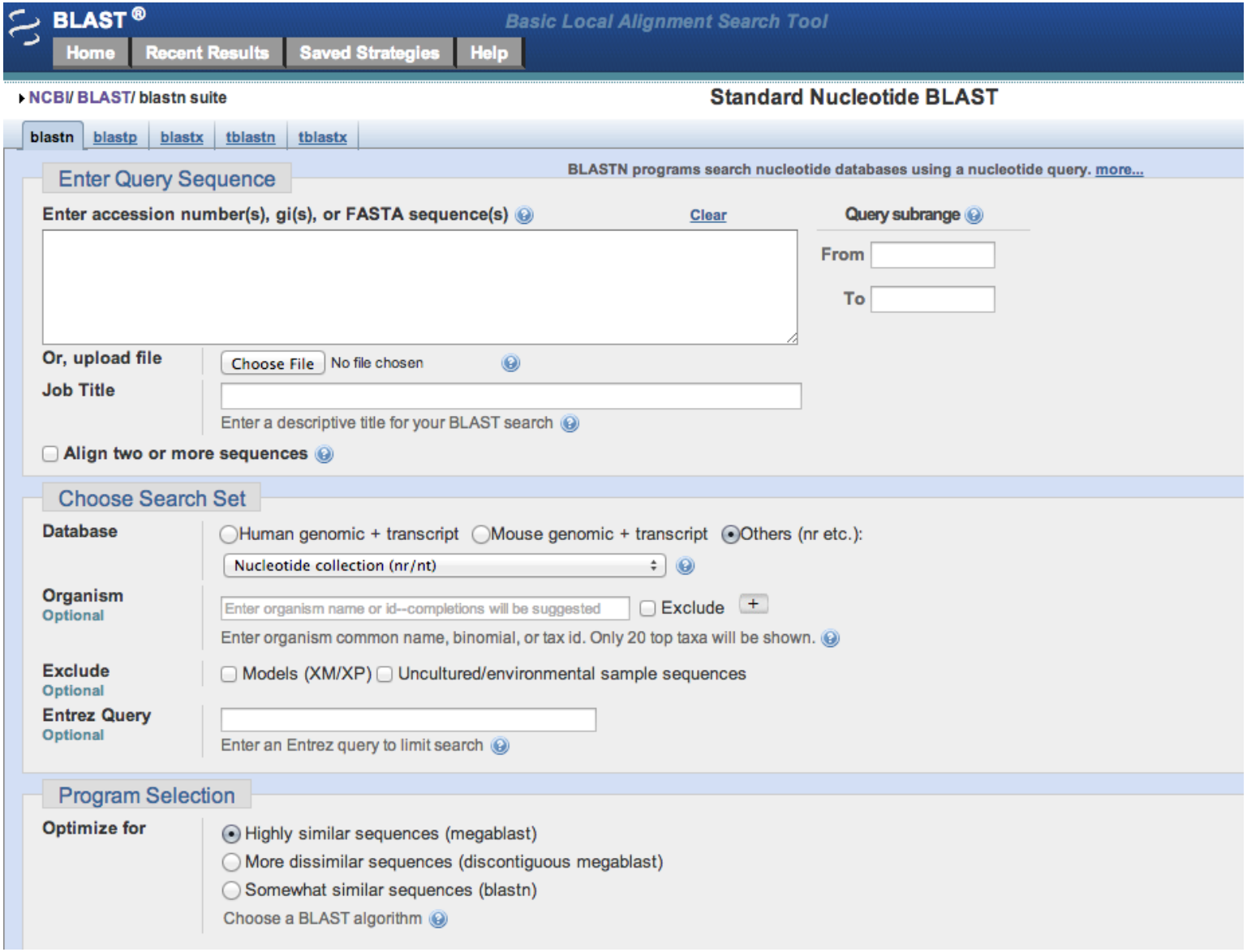
Basic BLAST

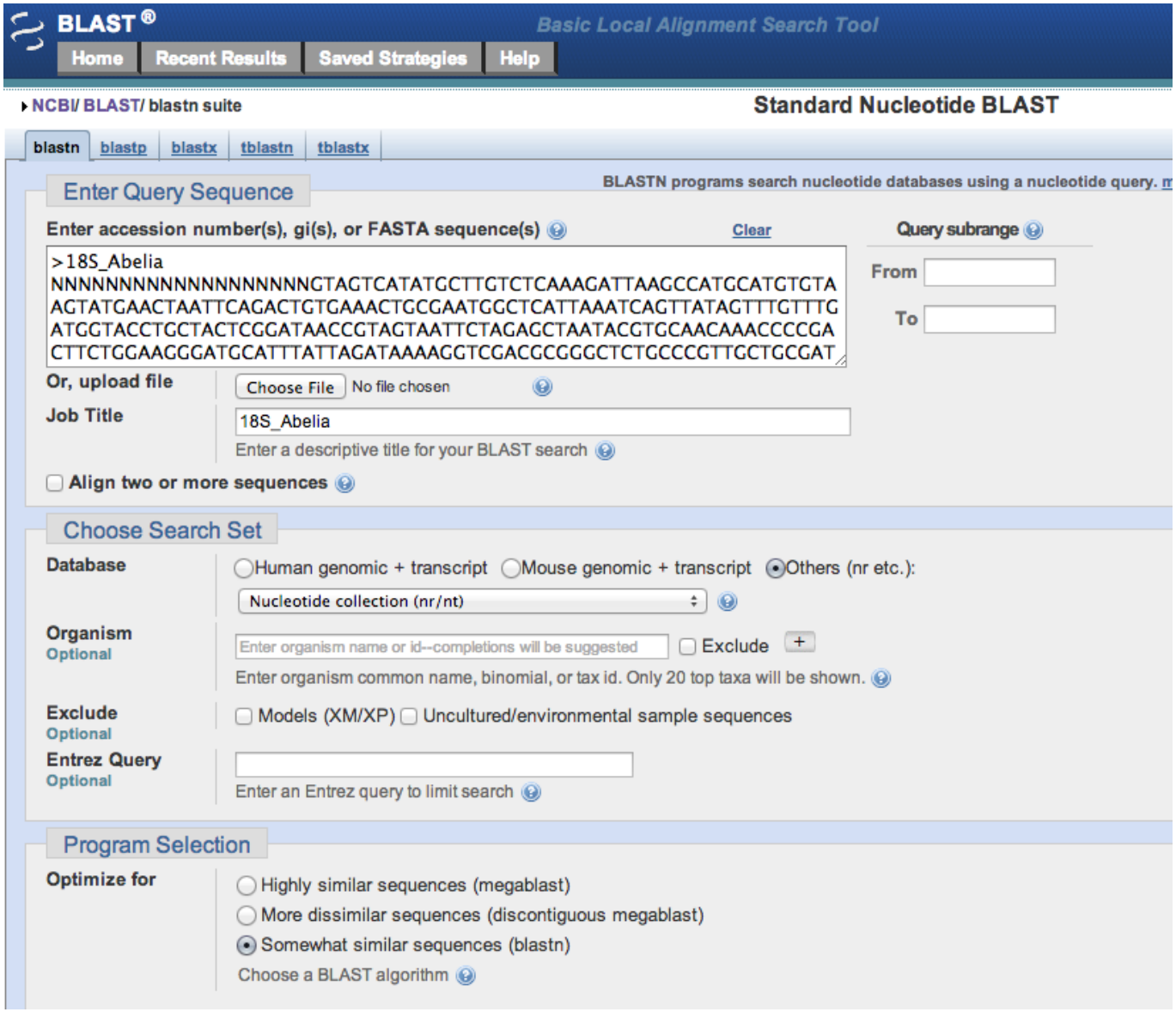
Choose a BLAST program to run.

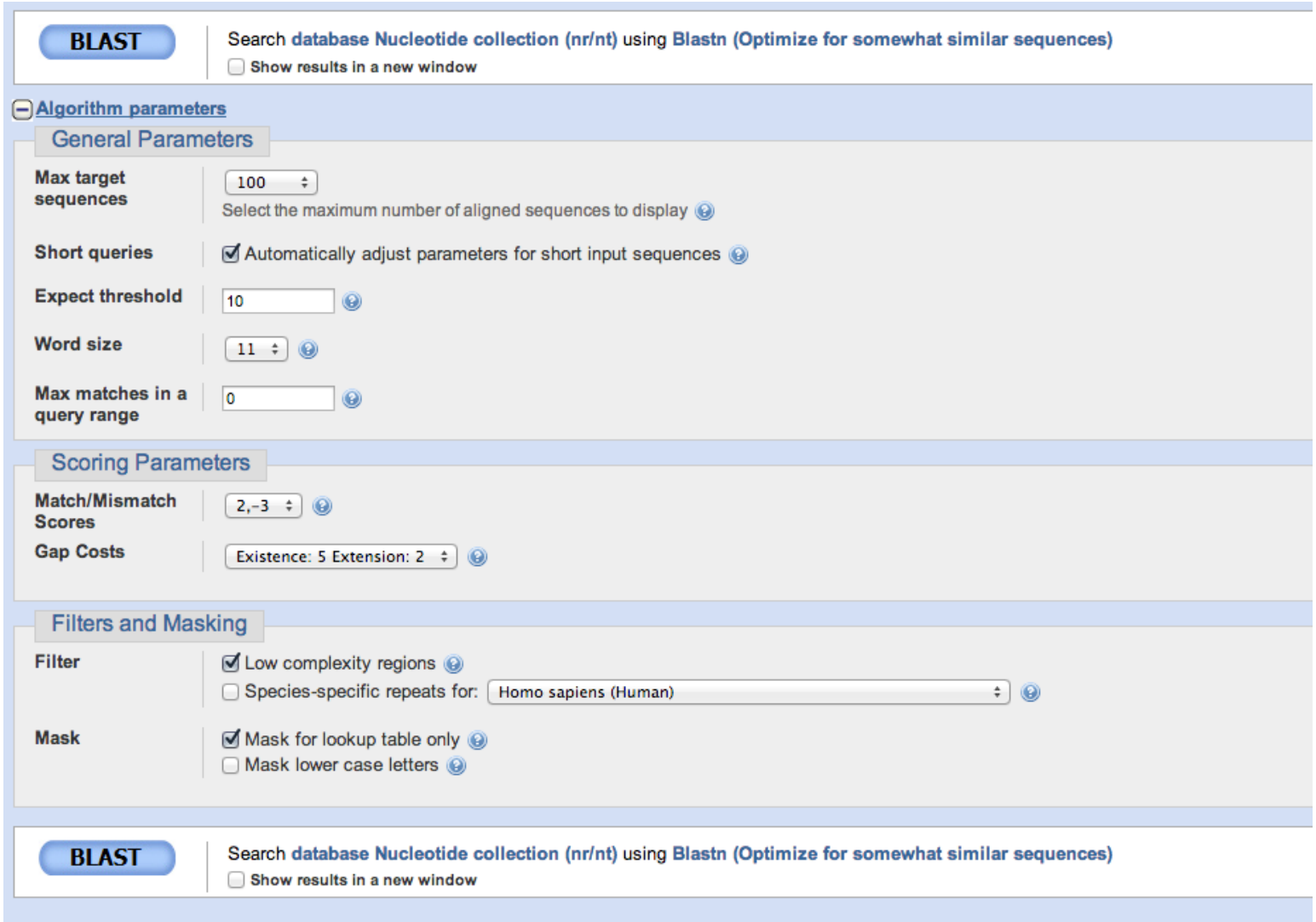
nucleotide blast	Search a **nucleotide** database using a **nucleotide** query *Algorithms:* blastn, megablast, discontiguous megablast
protein blast	Search **protein** database using a **protein** query *Algorithms:* blastp, psi-blast, phi-blast, delta-blast
blastx	Search **protein** database using a **translated nucleotide** query
tblastn	Search **translated nucleotide** database using a **protein** query
tblastx	Search **translated nucleotide** database using a **translated nucleotide** query

Specialized BLAST

Choose a type of specialized search (or database name in parentheses.)


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
Enter Query Sequence
BLASTN programs search nucleotide databases using a nucleotide query.
Enter accession number(s), gi(s), or FASTA sequence(s)
Clear
>18S_Abelia
NNNNNNNNNNNNNNNNNNNGTAGTCATATGCTTGTCTCAAAGATTAAGCCATGCATGTGTA
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

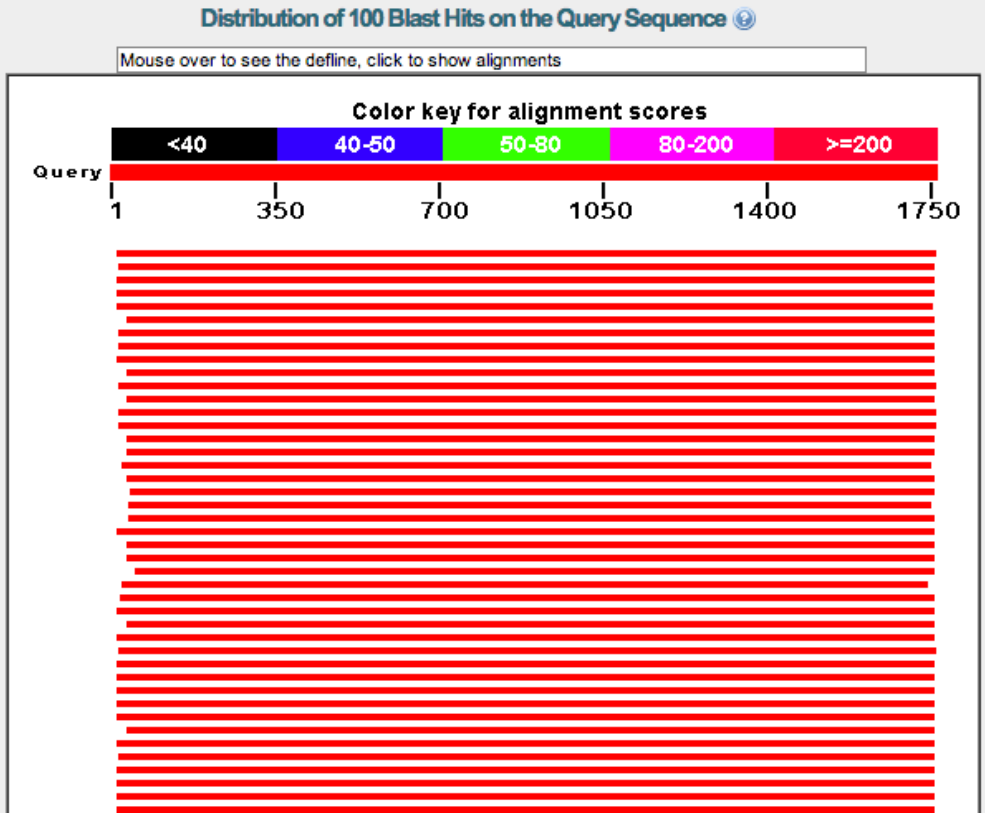

BLAST
Search database Nucleotide collection (nr/nt) using Blastn (Optimize for somewhat similar sequences)
Show results in a new window
Algorithm parameters
General Parameters
Max target sequences
100
Select the maximum number of aligned sequences to display
Short queries
Automatically adjust parameters 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

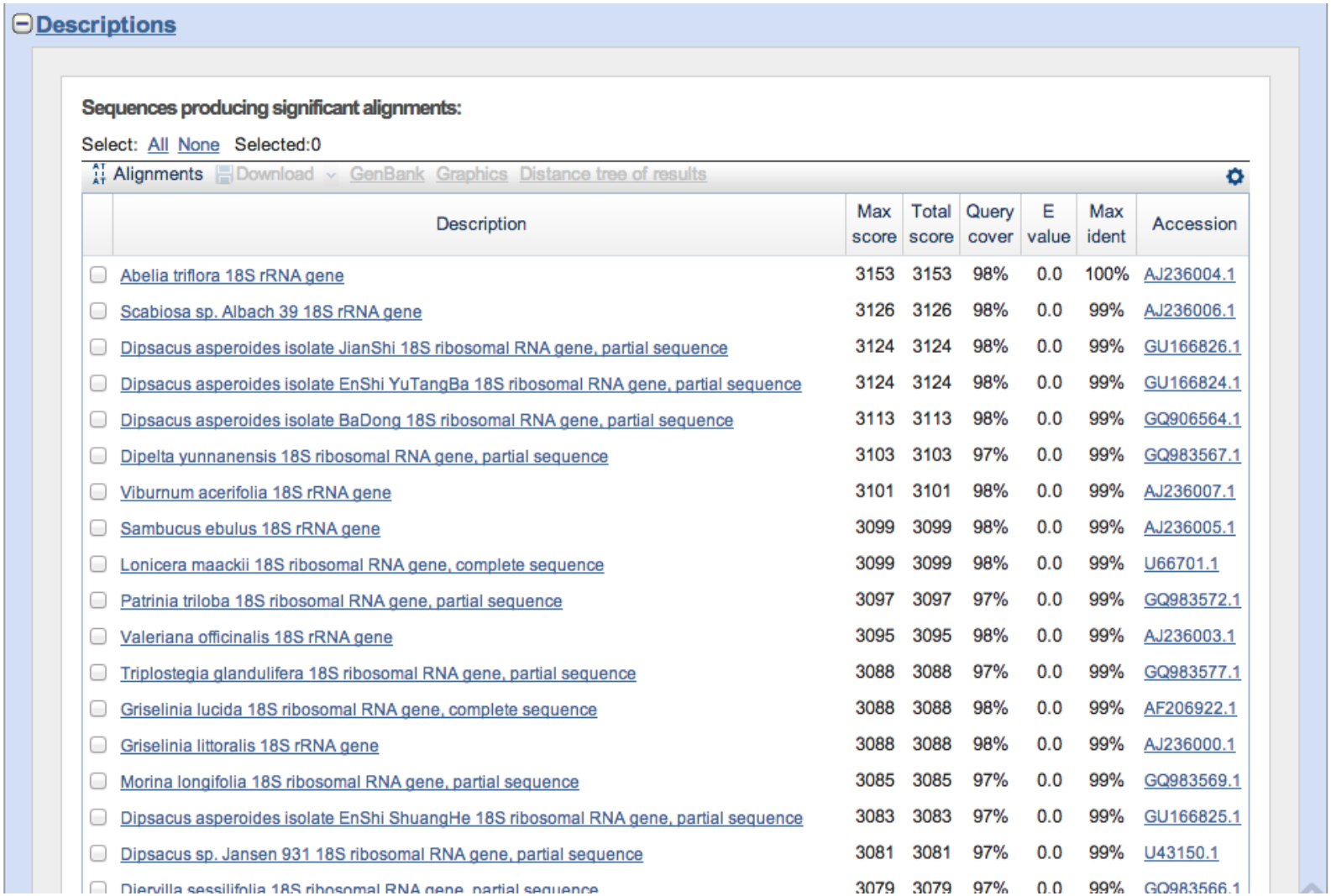
18S_Abelia

Query ID	lcl|27385	**Database Name**	nr
Description	18S_Abelia	**Description**	Nucleotide collection (nt)
Molecule type	nucleic acid	**Program**	BLASTN 2.2.28+ ▷Citation
Query Length	1767		

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?
- There are 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 $\rightarrow$ 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

Terminology matters

- **Can anything be “90% homologous”?**

Terminology matters

- 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

Therefore

- sequence similarity can suggest homology
- a significant alignment over 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

Rules

- There are no hard and fast rules
- Nucleotides
  - it has been suggested that sequence identity of more than 70% suggests homology
  - *E*-values of 10^{-6} or less = nope
- Proteins
  - 25% or more sequence identity
  - *E*-values of 10^{-3} or less = hmm... nope
- you must verify, and in a high-throughput large-scale analysis, there will be a (inescapable) margin of error

Conclusions

- What are you asking? What can pairwise sequence alignment address?
  - Are there any homologous sequences?
  - Are a set of sequences I have homologous?
- What can you not ask?
  - What are the relationships between these sequences?
  - Is there shared function with this sequence?
- What database should you search?
  - DNA or protein
- Which program should you run?
  - When possible it is best to search protein databases
  - Use NR and general GenBank for exploration or specific queries, but best to narrow down if possible