Finishing up databases
Loose ends
RefSeq accession numbers
NCBI tutorials
NCBI Demo
Introduction to Pairwise Sequence Alignment
BLAST
Orthologs and Paralogs
Gene duplication and mutation
Next time
Reading for next class

# Finish databases

# Start sequence alignment

## Lec˙05˙slides

# Outline

Finishing up databases
Loose ends
RefSeq accession numbers
NCBI tutorials
NCBI Demo
Introduction to Pairwise Sequence Alignment
BLAST
Orthologs and Paralogs
Gene duplication and mutation
Next time
Reading for next class

1 Finishing up databases
Loose ends
RefSeq accession numbers
NCBI tutorials
NCBI Demo
2 Introduction to Pairwise Sequence Alignment
BLAST
Orthologs and Paralogs
Gene duplication and mutation
3 Next time
4 Reading for next class

# Outline


Finishing up databases
Loose ends
RefSeq accession numbers
NCBI tutorials
NCBI Demo
Introduction to Pairwise Sequence Alignment
BLAST
Orthologs and Paralogs
Gene duplication and mutation
Next time
Reading for next class



1. Finishing up databases
   - Loose ends
   - RefSeq accession numbers
   - NCBI tutorials
   - NCBI Demo
2. Introduction to Pairwise Sequence Alignment
   - BLAST
   - Orthologs and Paralogs
   - Gene duplication and mutation
3. Next time
4. Reading for next class

# FASTA format is very common

Finishing up databases
Loose ends
RefSeq accession numbers
NCBI tutorials
NCBI Demo
Introduction to Pairwise Sequence Alignment
BLAST
Orthologs and Paralogs
Gene duplication and mutation
Next time
Reading for next class

- Line 1: ">" (greater than) followed by descriptive text
- Lines 2..n: sequence data (DNA or protein)

Examples:

```
>L19872_AhR_human mRNA
ATGAACAGCAGCAGCGCCAACATCACCTACGCCAGTCGCAAGCGGCGGAAGCCG
TGCAGAAAACAGTAAAGCCAATCCCAGCTGAAGGAATCAAGTCAAATCCTTCCA
GCGGCATAGAGACCGACTTAATACAGAGTTGGACCGTTTGGCTAGCCTGCTGCC
TTCCCACAAGATGTTATTAATAAGTTGGACAAACT

>L19872_AhR_human_1 amino acids
MNSSSANITYASRKRRKPVQKTVKPIPAEGIKSNPSKRHRDRLNTELDRLASLL
INKLDKLSVLRLSVSYLRAKSFFDVALKSSPTERNGGQDNCRAANFREGLNLQE
ALNGFVLVVTTDALVFYASSTIQDYLGFQQSDVIHQSVYELIHTEDRAEFQRQL
```

# Accession numbers

Sample GenBank record:

```
LOCUS       YP_002302326 238 aa linear BCT 04-NOV-2008
DEFINITION  green fluorescent protein
ACCESSION   YP_002302326
VERSION     YP_002302326.1  GI:211909965
AUTHORS     Srikhanta, Dowideit, ...
TITLE       Phasevarion mediated random switching ...
```

**Locus** Traditional identifier, not unique across databases, mainly historical

**Accession** Unique identifier, stable forever

**Version** An identifier that changes with each update to the sequence or annotation

**GI** Geninfo identifier

For full description:

`http://www.ncbi.nlm.nih.gov/Sitemap/samplerecord.html`

Finishing up databases
Loose ends
RefSeq accession numbers
NCBI tutorials
NCBI Demo
Introduction to Pairwise Sequence Alignment
BLAST
Orthologs and Paralogs
Gene duplication and mutation
Next time
Reading for next class

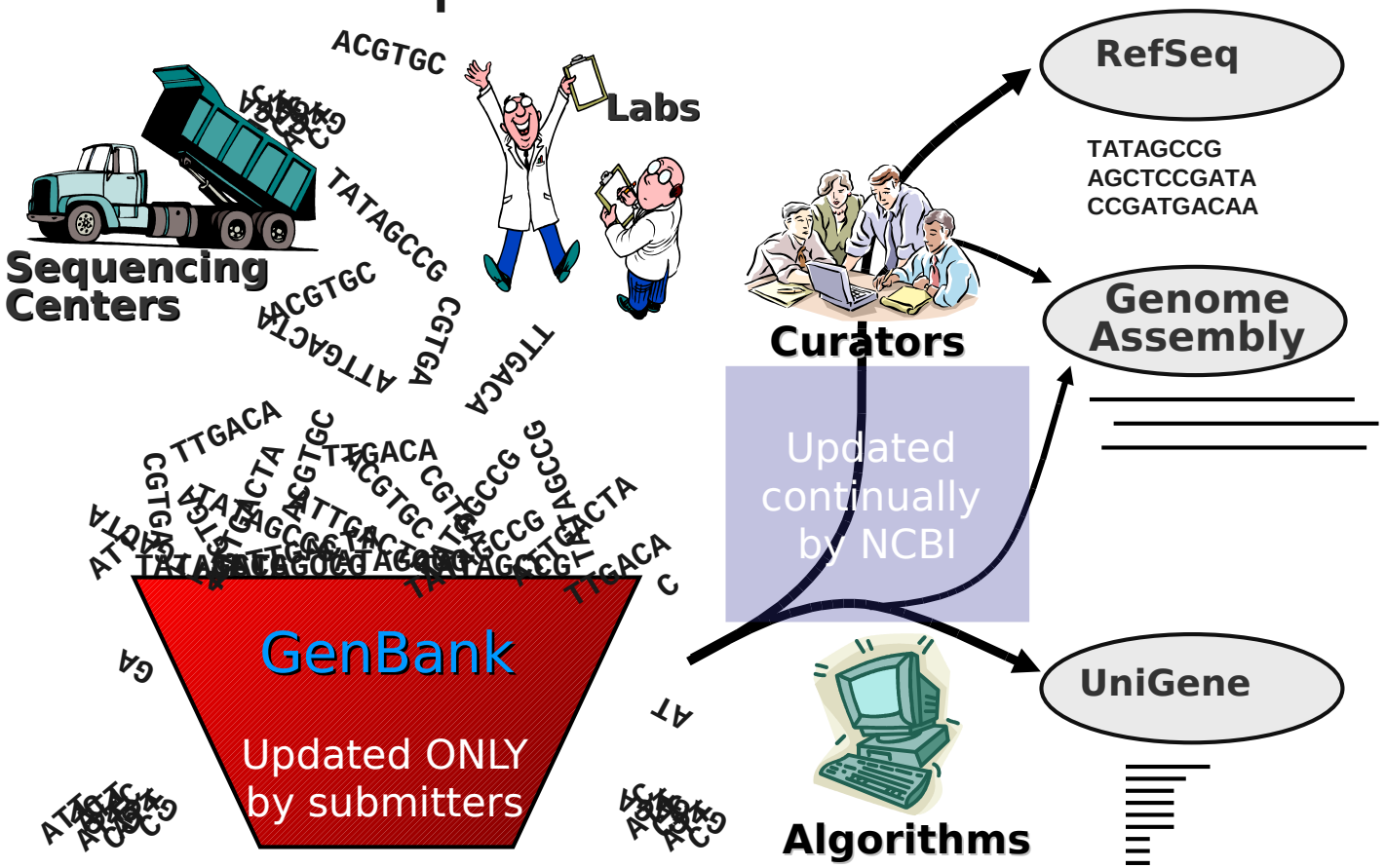



Image from ftp://ftp.ncbi.nih.gov/pub/FieldGuide/Slides/Archive/2008/USC.02.21.08/NCBI_part1.ppt

# Outline

Finishing up databases
Loose ends
RefSeq accession numbers
NCBI tutorials
NCBI Demo
Introduction to Pairwise Sequence Alignment
BLAST
Orthologs and Paralogs
Gene duplication and mutation
Next time
Reading for next class

1 Finishing up databases
Loose ends
RefSeq accession numbers
NCBI tutorials
NCBI Demo
2 Introduction to Pairwise Sequence Alignment
BLAST
Orthologs and Paralogs
Gene duplication and mutation
3 Next time
4 Reading for next class

# NCBI RefSeq accession numbers

Finishing up databases
Loose ends
RefSeq accession numbers
NCBI tutorials
NCBI Demo
Introduction to Pairwise Sequence Alignment
BLAST
Orthologs and Paralogs
Gene duplication and mutation
Next time
Reading for next class

- Accession number format indicates RefSeq
- Format: *letter letter underscore number...*
  - Example: NM_123456
- Two letters identify entry type. Important examples in red:

| | |
|---|---|
| NG_ | gene |
| NM_ | mRNA sequence |
| XM_ | computer predicted mRNA sequence |
| NP_ | protein sequence |
| XP_ | computer predicted protein sequence |
| NC_ | full length chromosome (big!) |
| (others) | See RefSeq documentation |

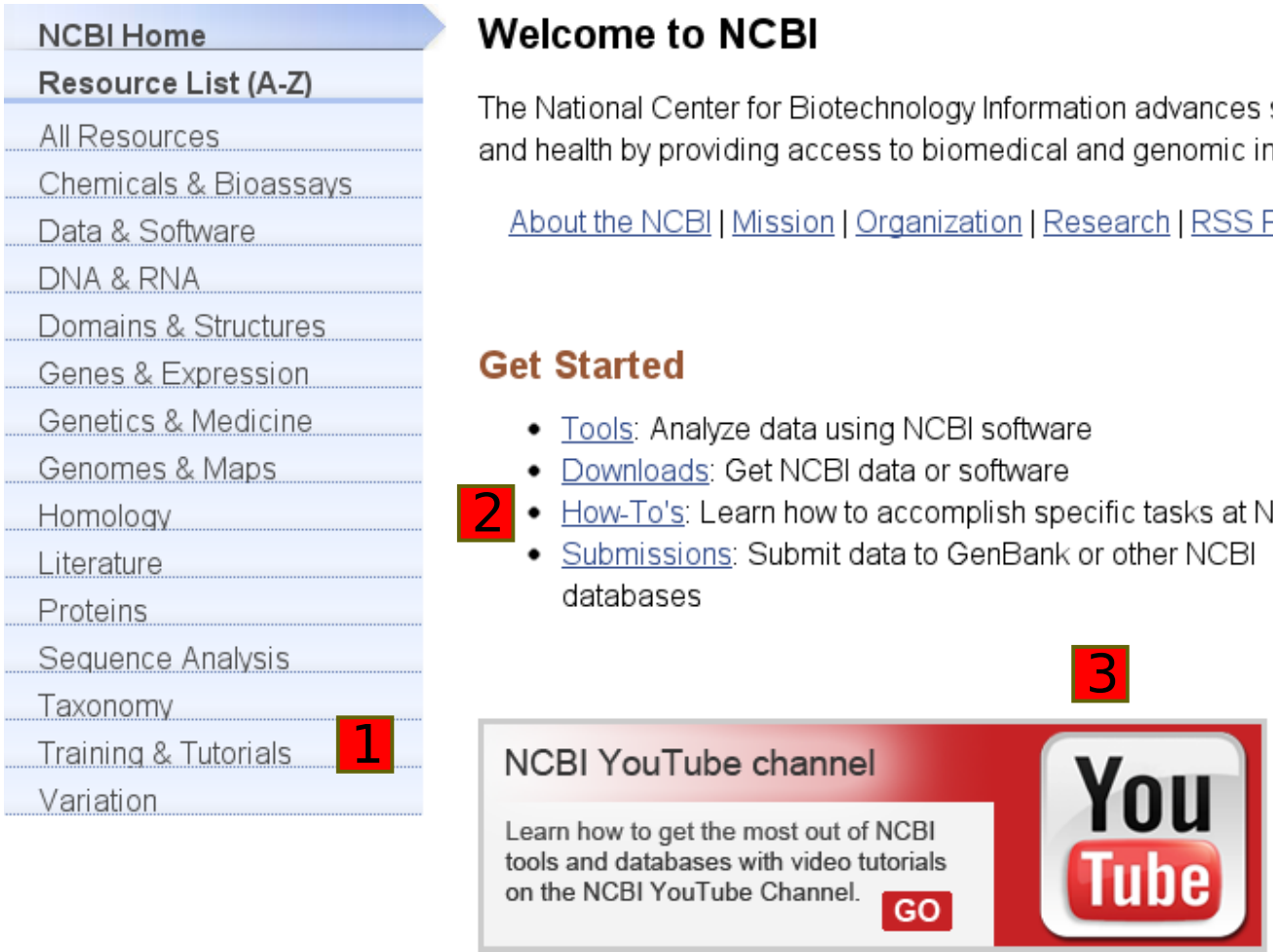
- Non-RefSeq GenBank accession numbers have no underscore. Example: AB123456

# Outline


1. Finishing up databases
   - Loose ends
   - RefSeq accession numbers
   - NCBI tutorials
   - NCBI Demo
2. Introduction to Pairwise Sequence Alignment
   - BLAST
   - Orthologs and Paralogs
   - Gene duplication and mutation
3. Next time
4. Reading for next class

# There are lots of good tutorials at NCBI


Finishing up databases
Loose ends
RefSeq accession numbers
NCBI tutorials
NCBI Demo
Introduction to Pairwise Sequence Alignment
BLAST
Orthologs and Paralogs
Gene duplication and mutation
Next time
Reading for next class


http://www.ncbi.nlm.nih.gov/

Finishing up databases
Loose ends
RefSeq accession numbers
NCBI tutorials
NCBI Demo
Introduction to Pairwise Sequence Alignment
BLAST
Orthologs and Paralogs
Gene duplication and mutation
Next time
Reading for next class

# Typical roadmap for simple NCBI queries

1. Search All Databases `http://www.ncbi.nlm.nih.gov/`
2. Read NCBI Bookshelf on the subject
3. If it is a human disease, look at OMIM (Online Mendelian Inheritance in Man)
4. Look at the papers in PubMed, start with review papers.
5. Find the relevant genes
6. Search All Databases for the relevant genes, refine the query by species, etc.
7. Look in Gene for information on gene http://www.ncbi.nlm.nih.gov/gene
8. Find RefSeq, protein structure, homologues, etc
9. More details throughout the semester

# Outline


Finishing up databases
Loose ends
RefSeq accession numbers
NCBI tutorials
NCBI Demo
Introduction to Pairwise Sequence Alignment
BLAST
Orthologs and Paralogs
Gene duplication and mutation
Next time
Reading for next class



1. Finishing up databases
   - Loose ends
   - RefSeq accession numbers
   - NCBI tutorials
   - NCBI Demo
2. Introduction to Pairwise Sequence Alignment
   - BLAST
   - Orthologs and Paralogs
   - Gene duplication and mutation
3. Next time
4. Reading for next class

Finishing up databases
Loose ends
RefSeq accession numbers
NCBI tutorials
NCBI Demo
Introduction to Pairwise Sequence Alignment
BLAST
Orthologs and Paralogs
Gene duplication and mutation
Next time
Reading for next class

Remember glycolysis? Let's check out **hexokinase**.

Let's checkout how **anemia** relates to **hexokinase**.

`http://www.ncbi.nlm.nih.gov/`

# Outline

Finishing up databases
Loose ends
RefSeq accession numbers
NCBI tutorials
NCBI Demo
Introduction to Pairwise Sequence Alignment
BLAST
Orthologs and Paralogs
Gene duplication and mutation
Next time
Reading for next class

1. Finishing up databases
   - Loose ends
   - RefSeq accession numbers
   - NCBI tutorials
   - NCBI Demo
2. Introduction to Pairwise Sequence Alignment
   - BLAST
   - Orthologs and Paralogs
   - Gene duplication and mutation
3. Next time
4. Reading for next class

# Introduction to Ch. 3 - Pairwise Sequence Alignment

Finishing up databases
Loose ends
RefSeq accession numbers
NCBI tutorials
NCBI Demo
Introduction to Pairwise Sequence Alignment
BLAST
Orthologs and Paralogs
Gene duplication and mutation
Next time
Reading for next class

Biological justification for much of bioinformatics

Finishing up databases
Loose ends
RefSeq accession numbers
NCBI tutorials
NCBI Demo
Introduction to Pairwise Sequence Alignment
BLAST
Orthologs and Paralogs
Gene duplication and mutation
Next time
Reading for next class

# Think back to Biol-360 day one...

Three driving forces behind bioinformatics:

- Massive volumes of DNA sequence data
- Gene conservation between species
- Systems biology

# Outline


Finishing up databases
Loose ends
RefSeq accession numbers
NCBI tutorials
NCBI Demo
Introduction to Pairwise Sequence Alignment
BLAST
Orthologs and Paralogs
Gene duplication and mutation
Next time
Reading for next class



1 Finishing up databases
Loose ends
RefSeq accession numbers
NCBI tutorials
NCBI Demo
2 Introduction to Pairwise Sequence Alignment
BLAST
Orthologs and Paralogs
Gene duplication and mutation
3 Next time
4 Reading for next class

# Many of you have done BLAST at NCBI


Finishing up databases
Loose ends
RefSeq accession numbers
NCBI tutorials
NCBI Demo
Introduction to Pairwise Sequence Alignment
BLAST
Orthologs and Paralogs
Gene duplication and mutation
Next time
Reading for next class


1. Query NCBI for human insulin. Here's the RefSeq output:

```
LOCUS       NM_000207 450 bp mRNA linear PRI 20-DEC-2003
DEFINITION  Homo sapiens insulin (INS), mRNA.
ACCESSION   NM_000207
VERSION     NM_000207.1  GI:4557670

ORIGIN
  1 gctgcatcag aagaggccat caagcacatc actgtccttc tgccatggcc ctgtggatgc
 61 gcctcctgcc cctgctggcg ctgctggccc tctggggacc tgacccagcc gcagcctttg
121 tgaaccaaca cctgtgcggc tcacacctgg tggaagctct ctacctagtg tgcggggaac ...
```

2. BLAST this mRNA sequence. Sample BLAST output:

83% identity

```
Human catggccctgtggatgcgcctcctgcccctgctggcgctgctggccc
      |||||||||||||||||| |||||||||||||||| |||||   ||
Mouse catggccctgtggatgcgcttcctgcccctgctggccctgctcttcc
```

Two HUGE biological questions:

1. What's going on?
2. Why do we care?

# Outline

Finishing up databases
Loose ends
RefSeq accession numbers
NCBI tutorials
NCBI Demo
Introduction to Pairwise Sequence Alignment
BLAST
Orthologs and Paralogs
Gene duplication and mutation
Next time
Reading for next class

1. Finishing up databases
   - Loose ends
   - RefSeq accession numbers
   - NCBI tutorials
   - NCBI Demo
2. Introduction to Pairwise Sequence Alignment
   - BLAST
   - Orthologs and Paralogs
   - Gene duplication and mutation
3. Next time
4. Reading for next class

# Orthologs, Paralogs, and Homologues


Finishing up databases
Loose ends
RefSeq accession numbers
NCBI tutorials
NCBI Demo
Introduction to Pairwise Sequence Alignment
BLAST
Orthologs and Paralogs
Gene duplication and mutation
Next time
Reading for next class

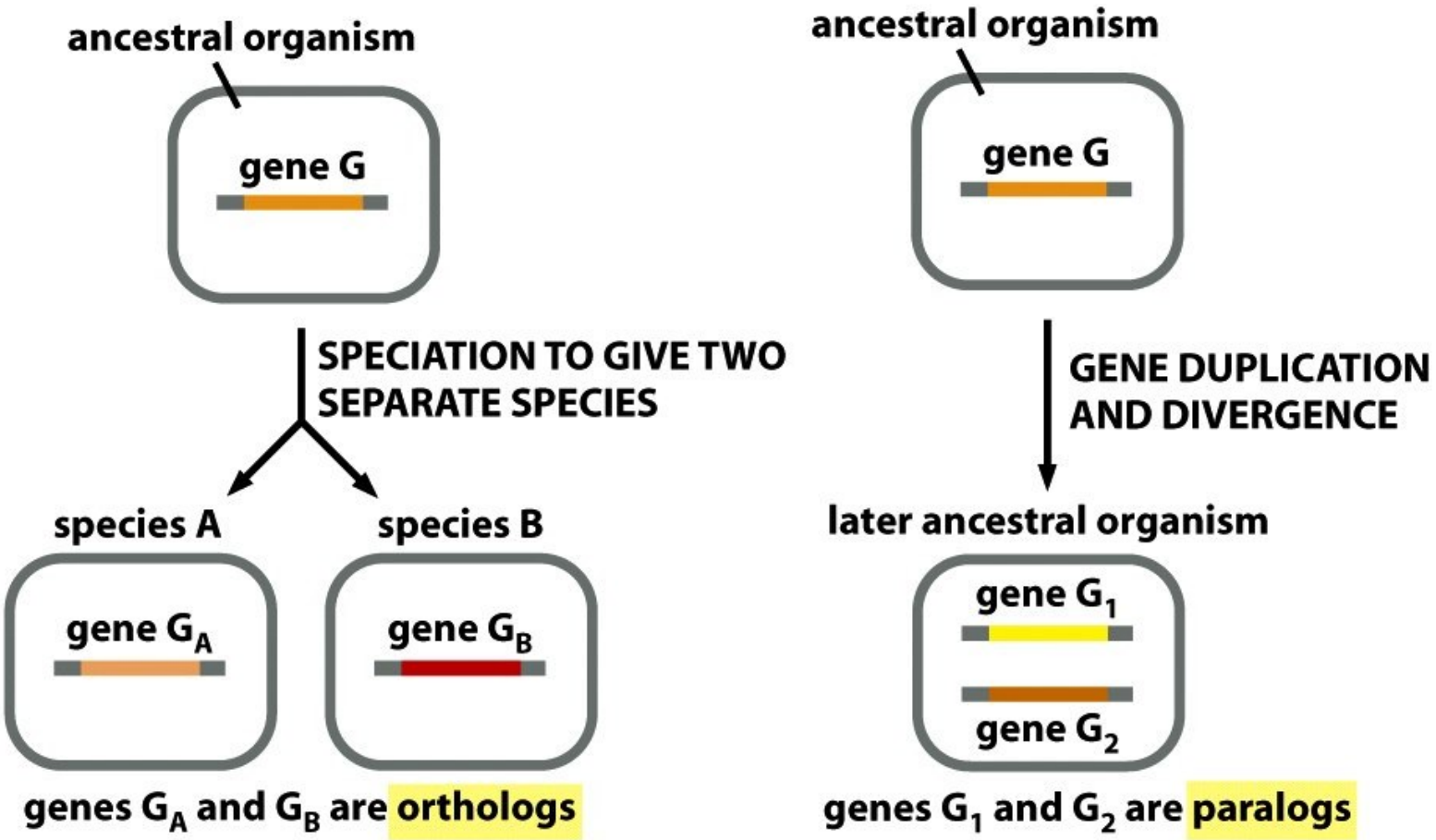




Figure 1-25 Molecular Biology of the Cell 5/e (© Garland Science 2008)

**Note to self: This would be a great iClicker question.**

# Evolution of the globin family


Finishing up databases
Loose ends
RefSeq accession numbers
NCBI tutorials
NCBI Demo
Introduction to Pairwise Sequence Alignment
BLAST
Orthologs and Paralogs
Gene duplication and mutation
Next time
Reading for next class

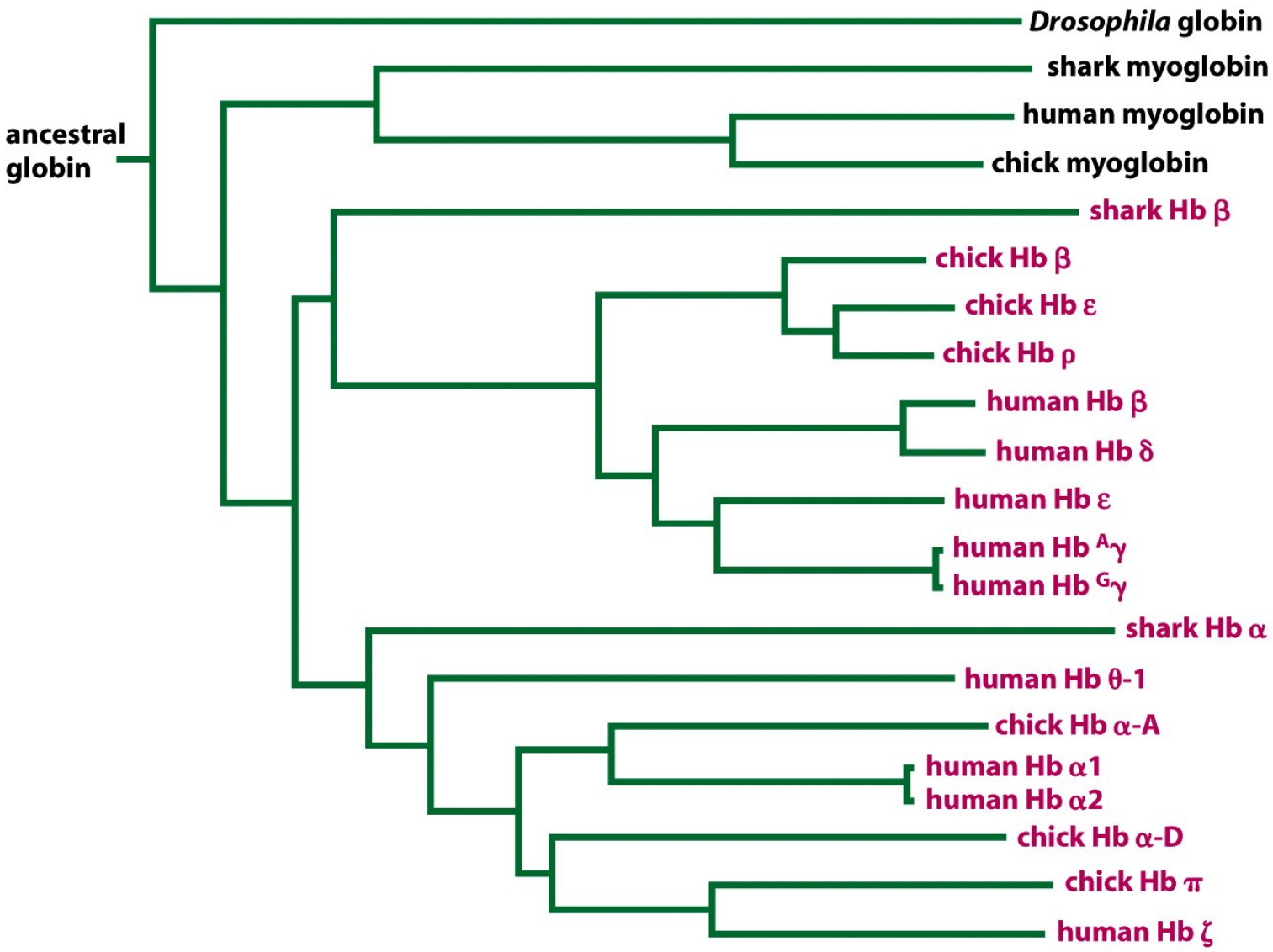




**Figure 1-26 Molecular Biology of the Cell 5/e (© Garland Science 2008)**

Finishing up databases
Loose ends
RefSeq accession numbers
NCBI tutorials
NCBI Demo
Introduction to Pairwise Sequence Alignment
BLAST
Orthologs and Paralogs
Gene duplication and mutation
Next time
Reading for next class

**Key Point:**

Gene duplication produces new genes.

Fate of new genes:

- Copy is initially identical, so (simple model):
  $\rightarrow 2\times$ amount of mRNA
  $\rightarrow 2\times$ amount of protein product.
- Is this duplicated gene beneficial for the organism?
  - If beneficial, then natural selective pressure preserves both copies. Examples: tRNA and rRNA for protein synthesis exist in many copies.
  - If detrimental, then selective pressure tends to deactivate one copy. The inactive copy is then called a pseudogene, and is no longer under any selective pressure.
  - If neutral, then one copy is free to evolve, as long as one continues to function.

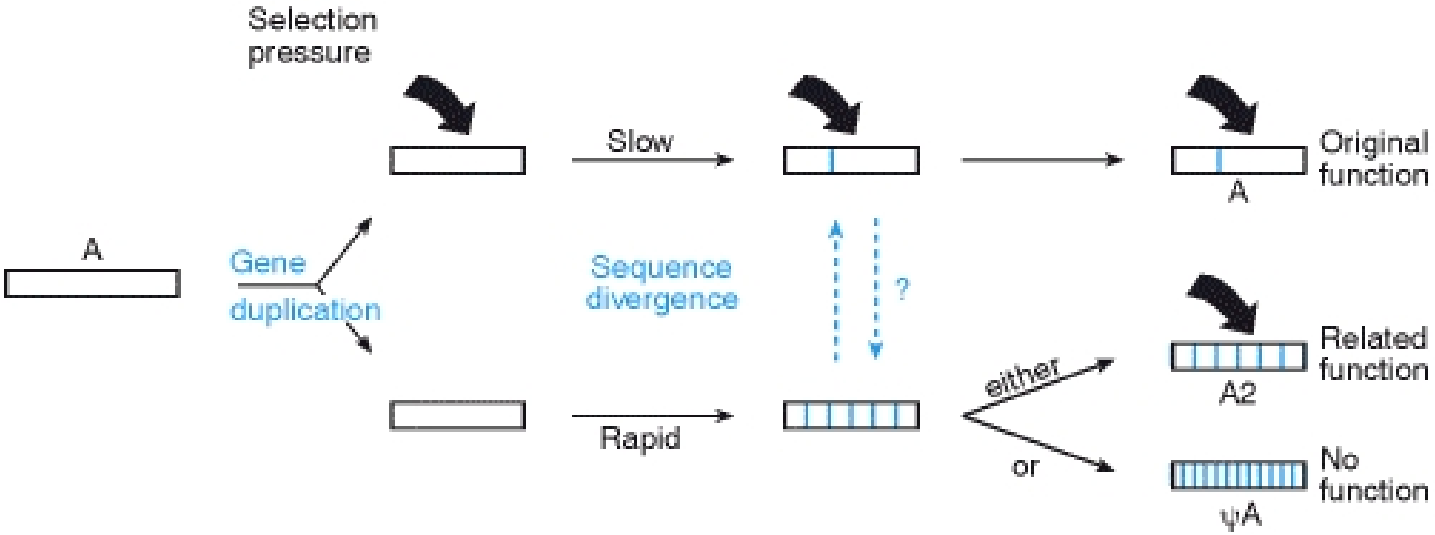
# Outline

Finishing up databases
Loose ends
RefSeq accession numbers
NCBI tutorials
NCBI Demo
Introduction to Pairwise Sequence Alignment
BLAST
Orthologs and Paralogs
Gene duplication and mutation
Next time
Reading for next class

1. Finishing up databases
   - Loose ends
   - RefSeq accession numbers
   - NCBI tutorials
   - NCBI Demo
2. Introduction to Pairwise Sequence Alignment
   - BLAST
   - Orthologs and Paralogs
   - Gene duplication and mutation
3. Next time
4. Reading for next class

# Gene duplication & evolution


Finishing up databases
Loose ends
RefSeq accession numbers
NCBI tutorials
NCBI Demo
Introduction to Pairwise Sequence Alignment
BLAST
Orthologs and Paralogs
Gene duplication and mutation
Next time
Reading for next class


Human Molecular Genetics 2. 2nd ed. Strachan, Tom and Read, Andrew P.

# How does all this relate to BLAST?

Finishing up databases
Loose ends
RefSeq accession numbers
NCBI tutorials
NCBI Demo
Introduction to Pairwise Sequence Alignment
BLAST
Orthologs and Paralogs
Gene duplication and mutation
Next time
Reading for next class

- BLAST looks for sequence similarity
  - Example: `AAACCG` is similar to `TAACCG`

- Homology + evolution creates sequence similarity

83% identity

```
Human catggccctgtggatgcgcctcctgcccctgctggcgctgctggccc
      |||||||||||||||||| |||||||||||||||| |||||   ||
Mouse catggccctgtggatgcgcttcctgcccctgctggccctgctcttcc
```

Therefore, we look for sequence similarity as evidence of homology

Therefore, we use BLAST to find evidence of homology.

Finishing up databases
Loose ends
RefSeq accession numbers
NCBI tutorials
NCBI Demo
Introduction to Pairwise Sequence Alignment
BLAST
Orthologs and Paralogs
Gene duplication and mutation
Next time
Reading for next class

Note to self:

Previous slide will make another great iClicker question.

Prof keeps saying this is a key point of the course.

# Outline

Finishing up databases
Loose ends
RefSeq accession numbers
NCBI tutorials
NCBI Demo
Introduction to Pairwise Sequence Alignment
BLAST
Orthologs and Paralogs
Gene duplication and mutation
Next time
Reading for next class

1. Finishing up databases
   - Loose ends
   - RefSeq accession numbers
   - NCBI tutorials
   - NCBI Demo
2. Introduction to Pairwise Sequence Alignment
   - BLAST
   - Orthologs and Paralogs
   - Gene duplication and mutation
3. Next time
4. Reading for next class

# Where we're going from here

- Ch. 3 & 4 cover sequence alignment & BLAST
- Alignment is finding the closest arrangement of two sequences:

```
--ACCTAGGA
ACCTAGGA-- →

 -ACCTAGGA
 ACCTAGGA-  →

  ACCTAGGA          We have a winner
  ACCTAGGA  →       100% identity

  ACCTAGGA-
  -ACCTAGGA  →

  ACCTAGGA--
  --ACCTAGGA
```

- Alignment is the essence of BLAST and **many** other bioinformatic tools.

# Outline for today


Finishing up databases
Loose ends
RefSeq accession numbers
NCBI tutorials
NCBI Demo
Introduction to Pairwise Sequence Alignment
BLAST
Orthologs and Paralogs
Gene duplication and mutation
Next time
Reading for next class


1. Finishing up databases
   - Loose ends
   - RefSeq accession numbers
   - NCBI tutorials
   - NCBI Demo
2. Introduction to Pairwise Sequence Alignment
   - BLAST
   - Orthologs and Paralogs
   - Gene duplication and mutation
3. Next time
4. Reading for next class

# Outline

Finishing up databases
Loose ends
RefSeq accession numbers
NCBI tutorials
NCBI Demo
Introduction to Pairwise Sequence Alignment
BLAST
Orthologs and Paralogs
Gene duplication and mutation
Next time
Reading for next class

1. Finishing up databases
   - Loose ends
   - RefSeq accession numbers
   - NCBI tutorials
   - NCBI Demo
2. Introduction to Pairwise Sequence Alignment
   - BLAST
   - Orthologs and Paralogs
   - Gene duplication and mutation
3. Next time
4. Reading for next class

Finishing up databases
Loose ends
RefSeq accession numbers
NCBI tutorials
NCBI Demo
Introduction to Pairwise Sequence Alignment
BLAST
Orthologs and Paralogs
Gene duplication and mutation
Next time
Reading for next class

## Chapter 3

Section "Scoring Matrices"
to
Section "Pairwise Alignment and Limits of Detection"

| Pages | Notes |
|---|---|
| 79–94 | Read |

Expect a few iClicker questions on these foundational ideas:

- What are homologues, orthologs, paralogs?
- What can happen to paralogs over evolutionary time?
- Important: What are the subtle distinctions between **sequence similarity** and **biological homology**?

  (If you're not **1000% clear** on these, post anonymously to the Blackboard class-wide discussion group.)