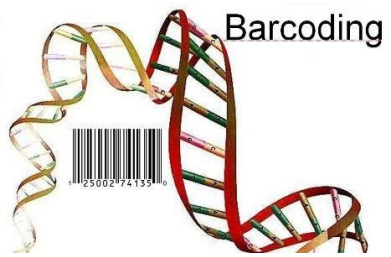


# Работа с нуклеотидными последовательностями (*теоретически-практическая часть*)

Нина Владимировна Воронова  
кандидат биологических наук, доцент  
Белорусский государственный университет



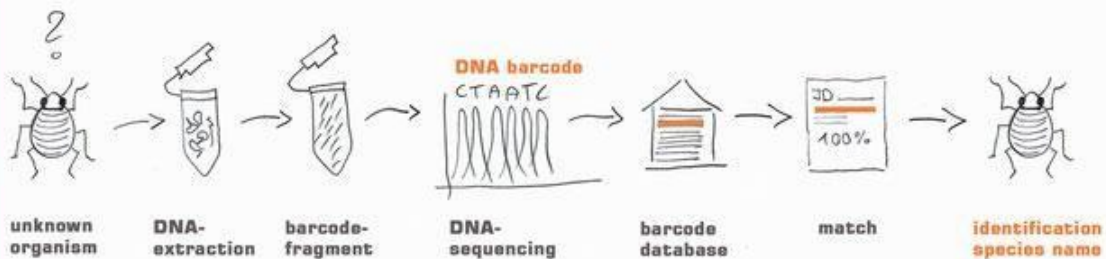
# ДНК-штрихкодирование. Суть метода

Выделение ДНК

Секвенирование  
нужного участка  
ДНК

Определение с  
использованием  
базы данных

How does DNA Barcoding work ?





# Стандарты системы BOLD

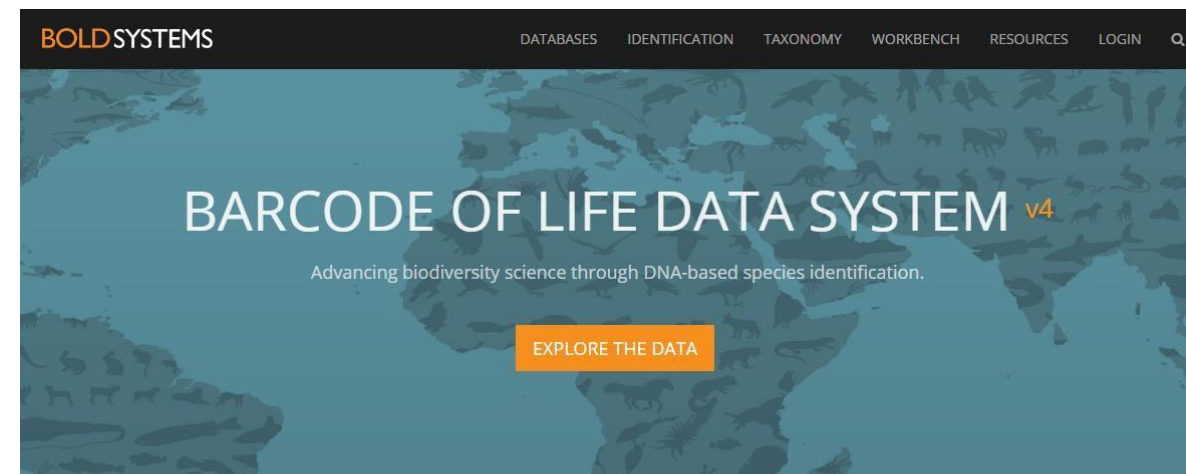
Идентификация животных – COI

Идентификация растений - *matK* и *rbcL*

Идентификация грибов – ITS-регион

Идентификация протистов - ?

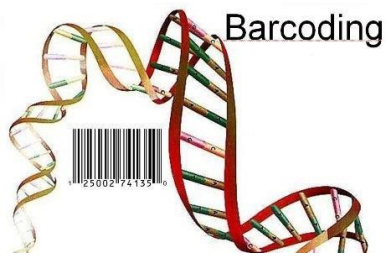
Идентификация бактерий - ? (16SrRNA)



DESIGNED TO SUPPORT THE GENERATION & APPLICATION OF DNA BARCODE DATA

BOLD is a cloud-based data storage and analysis platform developed at the Centre for Biodiversity Genomics in Canada. It consists of four main modules, a data portal, an educational portal, a registry of BINs (putative species), and a data collection and analysis workbench.

Please note that this version of BOLD is in beta and will contain bugs. Users can help address these bugs by testing the system and reporting issues to [support@boldsystems.org](mailto:support@boldsystems.org). This version is very different from the prior one but has access to all the same data.



# ДНК-штрихкодирование. Выбор маркера

*Molecular Marine Biology and Biotechnology*  
(1994) 3(5), 294-299

*Molecular Marine Biology and Biotechnology* (1994) 3(5), 294-299

## DNA primers for amplification of mitochondrial cytochrome c oxidase subunit I from diverse metazoan invertebrates

O. Folmer, M. Black, W. Hoeh, E. Lutz, and R. Vrijenhoek

Center for Theoretical and Applied Genetics, and Institute of Marine and Coastal Sciences, Rutgers University, New Brunswick, New Jersey 08903-2212

### Abstract

We describe "universal" DNA primers for polymerase chain reaction (PCR) amplification of a 710-bp fragment of the mitochondrial cytochrome c oxidase subunit I gene (COI) from 11 invertebrate phyla: Echinodermata, Mollusca, Annelida, Pogonophora, Arthropoda, Nemertea, Echinura, Sipuncula, Platyhelminthes, Tardigrada, and Coelenterata, as well as the putative phylum Vestimentifera. Preliminary comparisons revealed that these COI primers generate informative sequences for phylogenetic analysis at the species and higher taxonomic levels.

### Introduction

The purpose of this short communication is to describe "universal" DNA primers for the polymerase chain reaction (PCR) amplification of a 710-bp fragment of the mitochondrial cytochrome c oxidase subunit I gene (COI). This study was motivated by the recent discovery of more than 230 new invertebrate species, comprising new genera, families, classes, orders, and potentially a new phylum, from deep-sea hydrothermal vent and cold-water sulfide or methane seep communities (Tunnicliffe, 1991). Our goal was to develop molecular techniques for phylogenetic studies of these diverse organisms. We focused on the mitochondrial cytochrome c oxidase subunit I (COI) gene because it appears to be among the most conservative protein-coding genes in the mitochondrial genome of animals (Brown, 1985), which was preferable for the evolutionary

time depths likely to be found in our studies. We quickly became aware of the broad utility of these COI primers for broader systematic studies of metazoan invertebrates, including asclemerata, pterobranchiata, and coelenterata and deuterostomes.

### Results

To design candidate primers, we compared published DNA sequences from the following species: blue mussel, *Mytilus edulis*; fruitfly, *Drosophila yakuba*; honeybee, *Apis mellifera*; mosquito, *Anopheles gambiae*; little shrimp, *Artemia franciscana*; nematode, *Acaris suum*; and *Caenorhabditis elegans*; sea urchin, *Strongylocentrotus purpuratus*; carp, *Cyprinus carpio*; frog, *Xenopus laevis*; chicken, *Gallus gallus*; mouse, *Mus musculus*; cow, *Bos taurus*; fin whale, *Balaenoptera physalus*; and human, *Homo sapiens* (Figure 1). Several highly conserved regions of these COI genes were used as the targets for primer design.

Altogether, three coding-strand and six anti-coding-strand primers were tested (Table 1) for amplification efficiency. The following primer pair consistently amplified a 710-bp fragment of COI across the broadest array of invertebrates:

LCO1490 5'-ggtcaacaaatagaagatg-3'

HCO2198 5'-taaaacaggggcaaaaatca-3'

In the code names above, L and H refer to light and heavy DNA strands, CO refers to cytochrome oxidase, and the numbers (1490 and 2198) refer to the position of the *D. yakuba* 5' nucleotide.

We also present the primers as coding-strand sequences, along with their inferred amino acids (Figure 2). The usefulness of these primers results from the high degree of sequence conservation in their respective 3' ends across the 15 taxa. The 3' end of each primer is on a second-position nucleotide. All other pairwise primer combinations amplified fewer taxa or gave additional nonspecific products under less stringent amplification conditions.

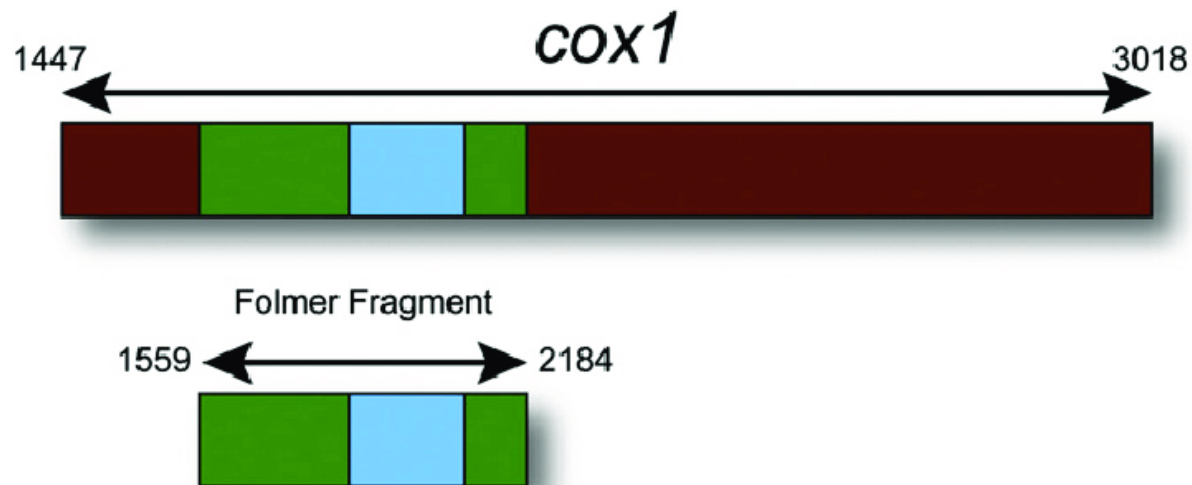
The LCO1490 and HCO2198 amplified DNA from more than 80 invertebrate species from 11

\*Present address: Department of Biology, Dalhousie University, Halifax, Nova Scotia, Canada.

Correspondence should be sent to this author.

Copyright © 1994 Blackwell Science, Inc.

Фрагмент Фолмера  
658 п.н. в 5'-области гена COI



Ратгерский университет  
Нью-Джерси, США



# Особенности маркеров, используемых для ДНК-штрихкодирования

Растения

*matK + rbcL*

Хлоропластные  
Белок-кодирующие  
Гаплоидные

Животные

COI

Митохондриальные  
Белок-кодирующие  
Гаплоидные

Грибы

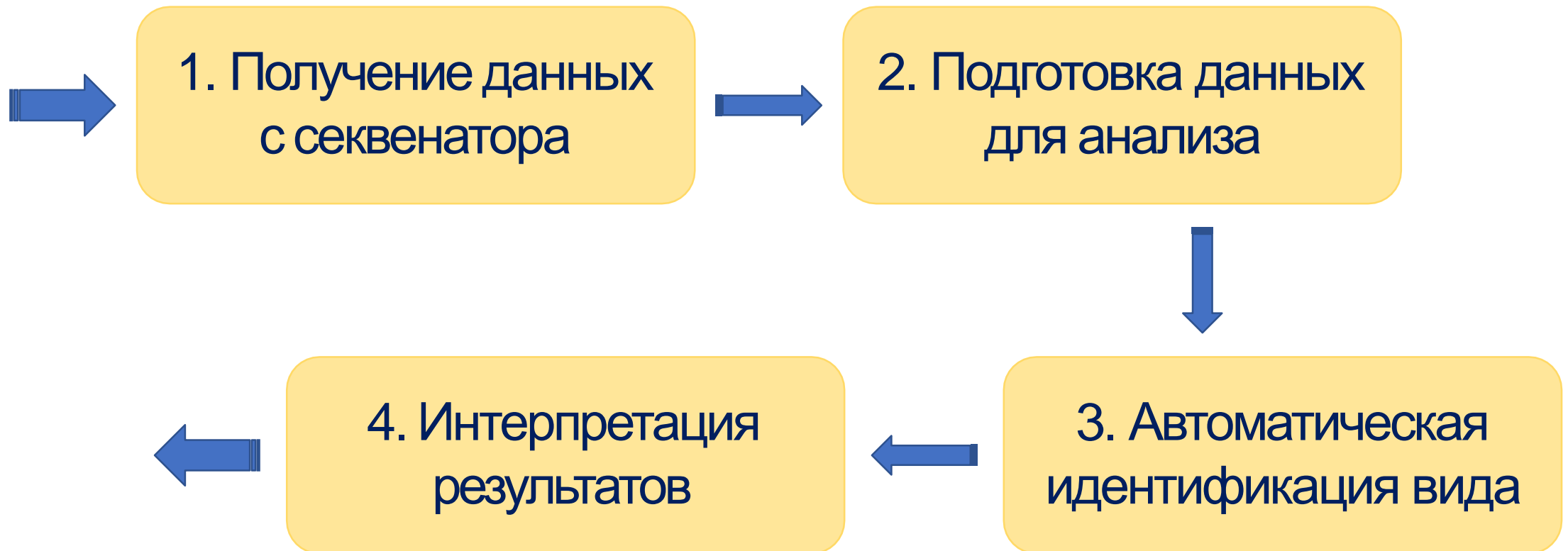
ITS1 + ITS2

Ядерные  
Некодирующие  
Гапло/диплоидные

Размер фрагмента ПЦР, используемого для ДНК-штрихкодирования, не должен превышать 800 bp



# ДНК-штрихкодирование. Анализ данных





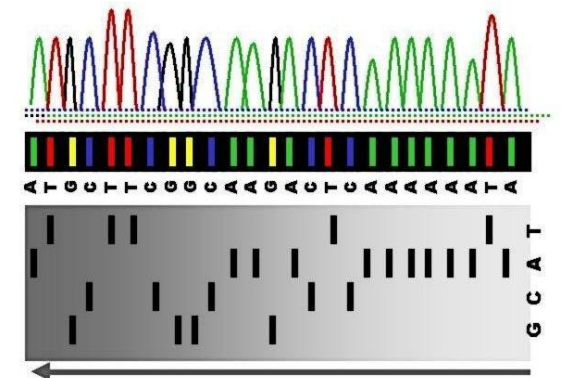
# 1. Получение данных с секвенатора

## «Классическое» ДНК-штрихкодирование

Выделение ДНК

ПЦР с универсальными  
праймерами

Секвенирование по Сэнгеру

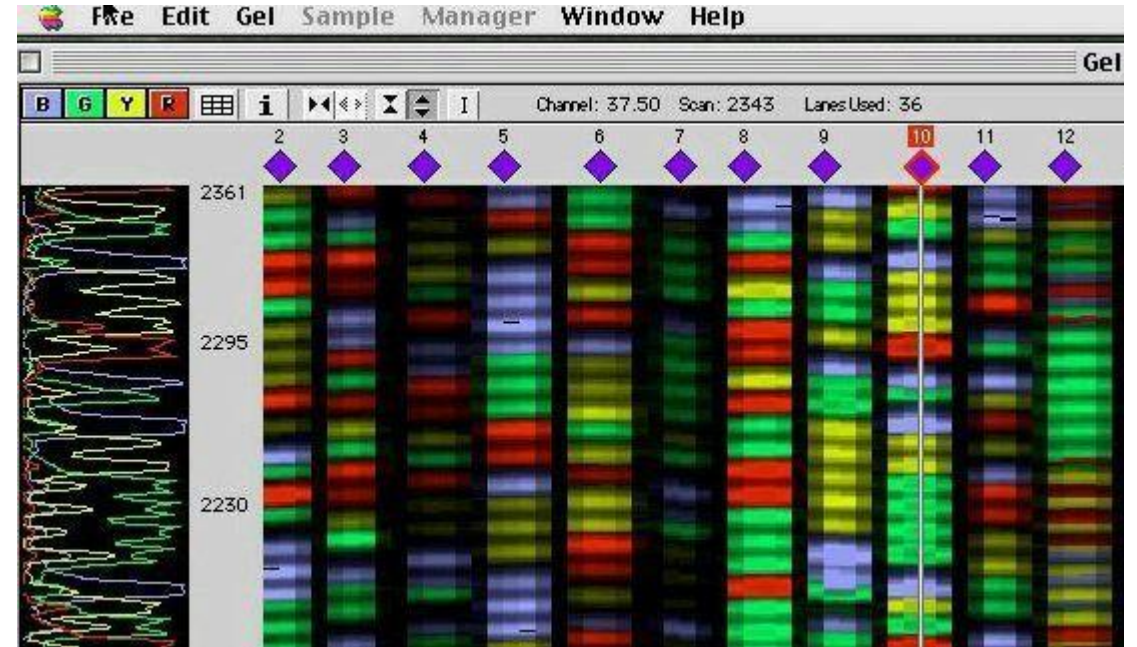
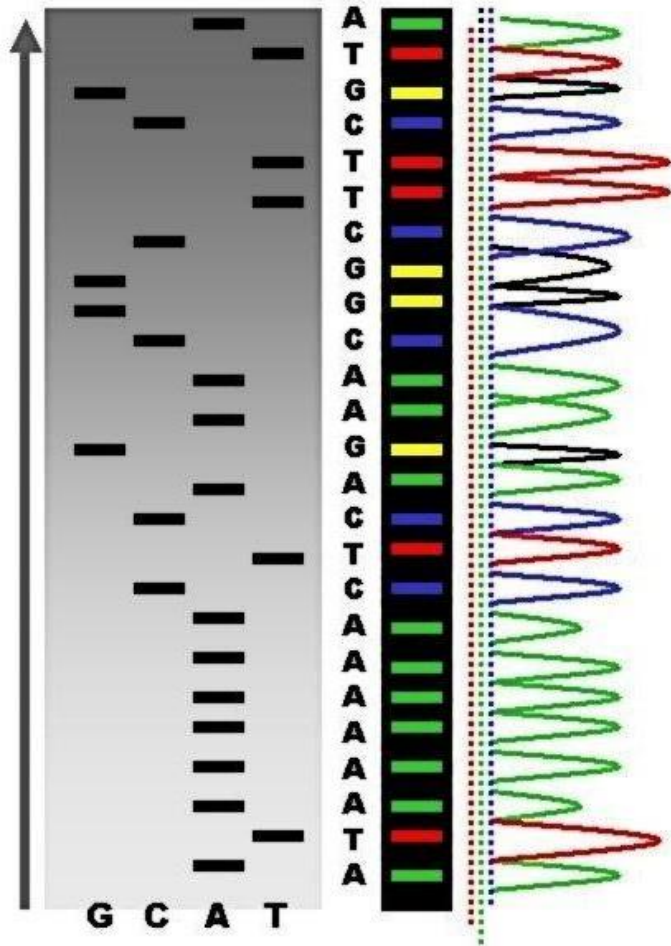


Анализ данных



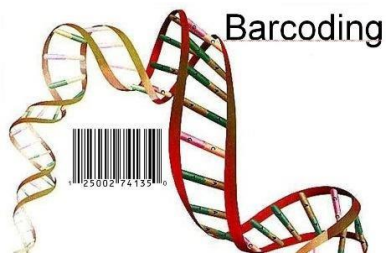
# 1. Получение данных с секвенатора

## Секвенирование по Сэнгеру



Секвенирование может быть проведено в одном (1 реакция, 1 праймер) или двух (2 реакции, 1 праймер в каждой реакции) направлениях





# 1. Получение данных с секвенатора

## Environmental DNA barcoding и метабаркодинг

Секвенирование по Сэнгеру

Секвенирование нового поколения (NGS)

Забор «сложных» проб, концентрация проб, выделение общей ДНК

ПЦР с видоспецифичными праймерами

Приготовление ПЦР ДНК-библиотеки

Секвенирование

NGS

Сборка ДНК-штрихкодов

Анализ данных



# 1. Получение данных с секвенатора

## NGS ( Next Generation Sequencing ) при метабаркодинге

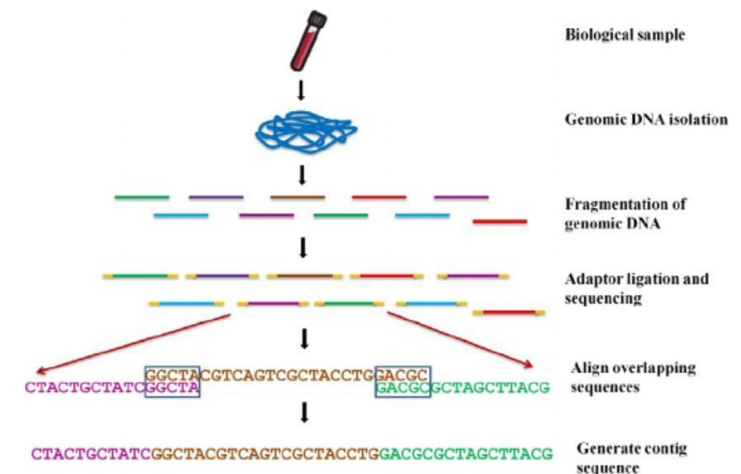
Выделение ДНК

ПЦР с универсальными  
праймерами

NGS

Приготовление библиотеки  
и «пришивание» адаптеров

Сборка индивидуальных  
ДНК-баркодов



Анализ данных



Barcoding

## 2. Подготовка данных для анализа Алгоритм действия

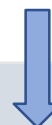
Данные с секвенатора



Оценка качества  
секвенирования



Редактирование сиквенса



Получение  
последовательности в  
текстовом формате



Получение консенсусной  
последовательности



*При необходимости*

ДНК-штрихкод



Анализ данных



## 2. Подготовка данных для анализа

### Данные, приходящие с секвенатора

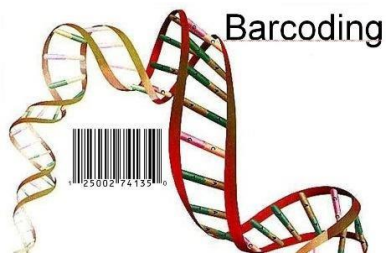
#### Секвенирование по Сэнгеру

# Одна буквенная последовательность длиной 650-700 bp (при использовании прямого праймера)

# Две последовательности длиной 650-700 bp : прямая и обратная (при использовании двух праймеров)

#### NGS

Миллионы буквенных последовательностей длиной 60-120 bp, представляющих собой неупорядоченные фрагменты ДНК-штрихкодов



## 2. Подготовка данных для анализа

# Наиболее часто используемые программы, работающие с данными секвенирования

<https://www.megasoftware.net/>

**BioEdit**  
Biological sequence alignment editor for Win95/98/NT/2K/XP

Copyright © 1997-2005  
Tom Hall  
Ibis Therapeutics  
Carlsbad, CA 92008

*Sophisticated and user-friendly software suite for analyzing DNA and protein sequence data from species and populations.*

Windows | Graphical (GUI) | MEGA X | **DOWNLOAD**

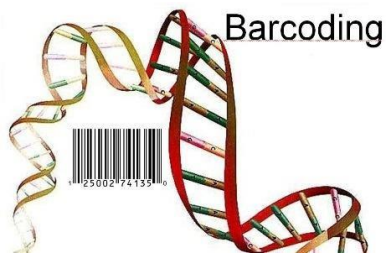
<b>BioEdit.zip</b> (full installation 12.6 Mb)
<b>Bug fixes / changes</b>
<b>BioEdit</b> General information
<b>BioDoc.pdf</b> (pdf format help doc)
<b>View Screenshots</b>

**BioEdit's features include:**

- Several modes of hand alignment
- Automated ClustalW alignment
- Automated Blast searches (local and WWW)
- Plasmid drawing and annotation
- Accessory application configuration
- Restriction mapping
- RNA comparative analysis tools
- Graphical matrix data viewing tools
- Shaded alignment figures
- Translation-based nucleic acid alignment
- ABI trace viewing, editing and printing
- Customizable ... [other features](#)

**Note:** Although BioEdit was recently updated, it is no longer being reliably maintained, and the documentation is out of date and no longer maintained. It is being updated slowly, but there is no guaranteed finish date. Until documentation is complete, play with the menus and see what happens, or email with a question.

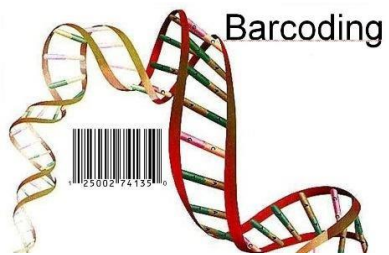
**Note:** If you have trouble launching BioEdit on Windows NT, try replacing BioEdit.exe with this version



## 2. Подготовка данных для анализа

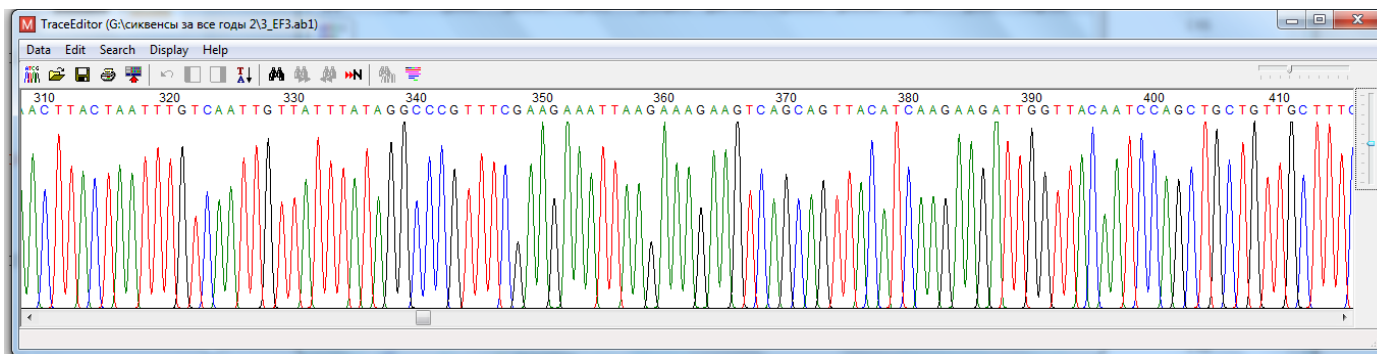
Ошибки секвенирования. Причины :

- Случайная вставка ошибочного нуклеотида
  - Присутствие неспецифичного продукта в ПЦП-образце
  - Присутствие посторонних образцов ДНК в первоначальной ДНК-пробе
- Внутриорганизменный полиморфизм
  - NUMTs (Nuclear Mitochondrial DNA segments)
  - Псевдогены

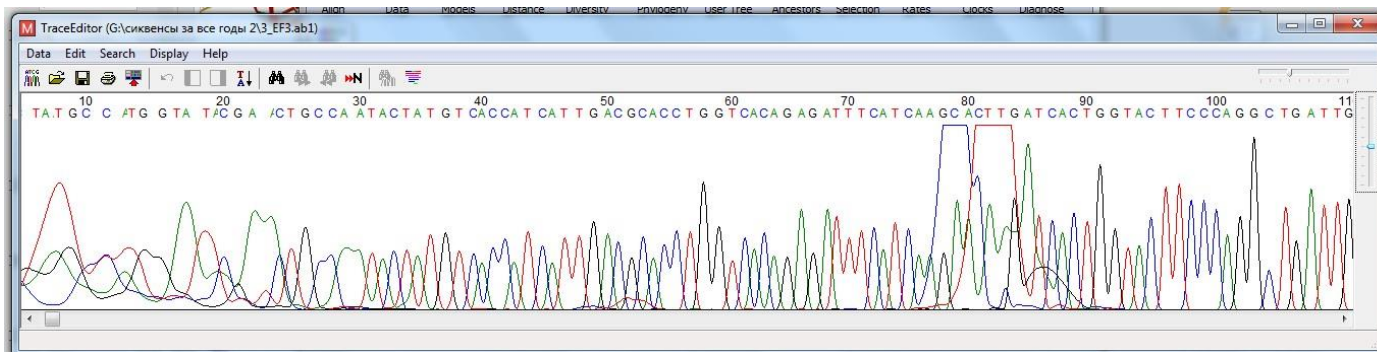


## 2. Подготовка данных для анализа

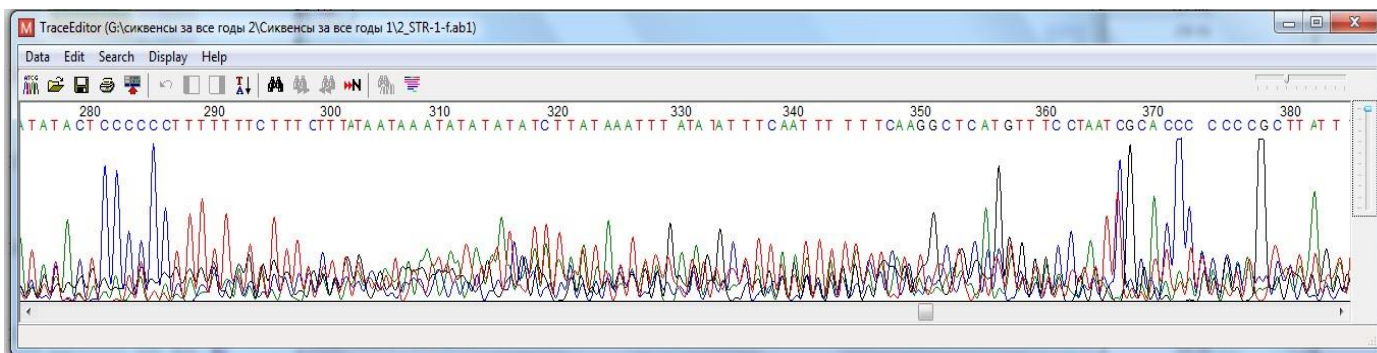
### Анализ качества сиквенса



Высокое качество



Качество, достаточное для редактирования



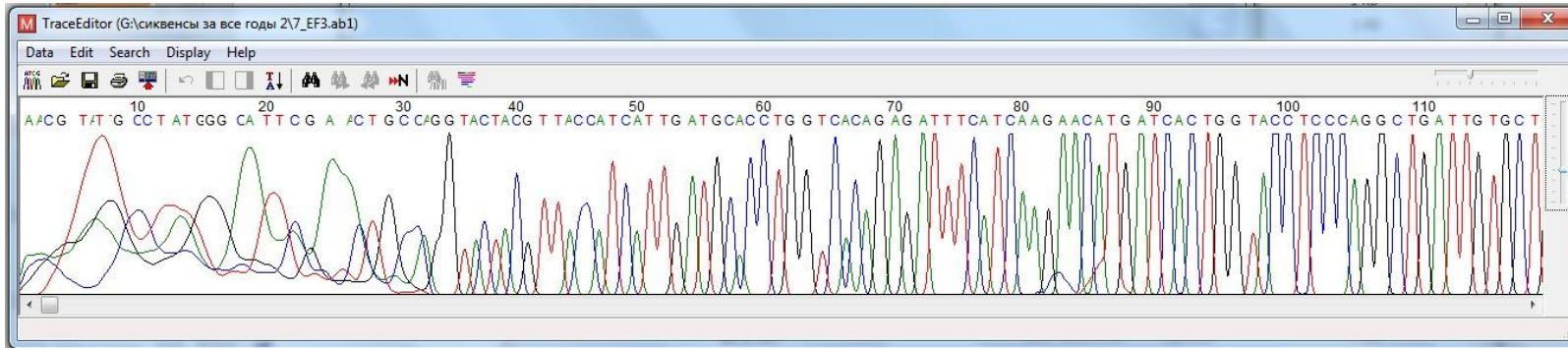
Низкое качество

Анализ данных

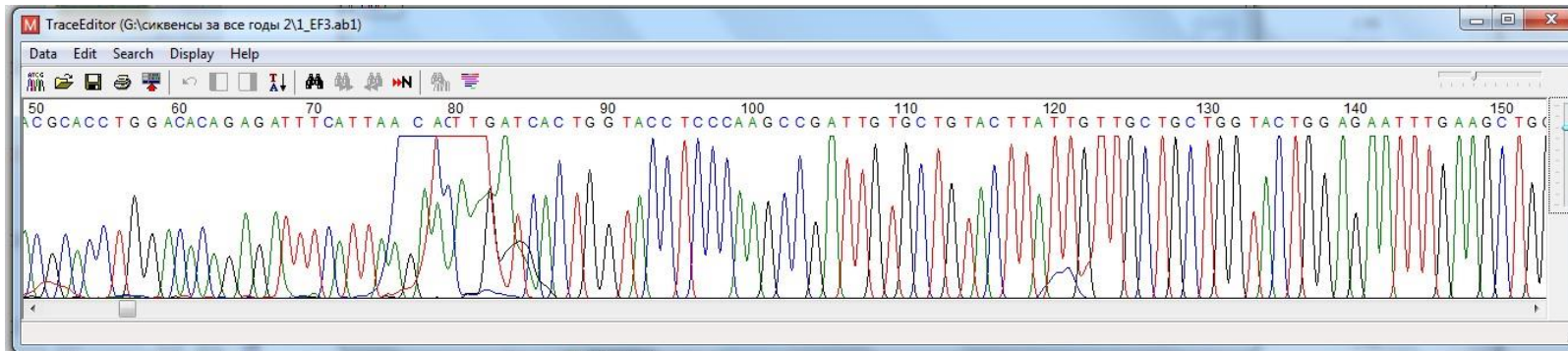


## 2. Подготовка данных для анализа

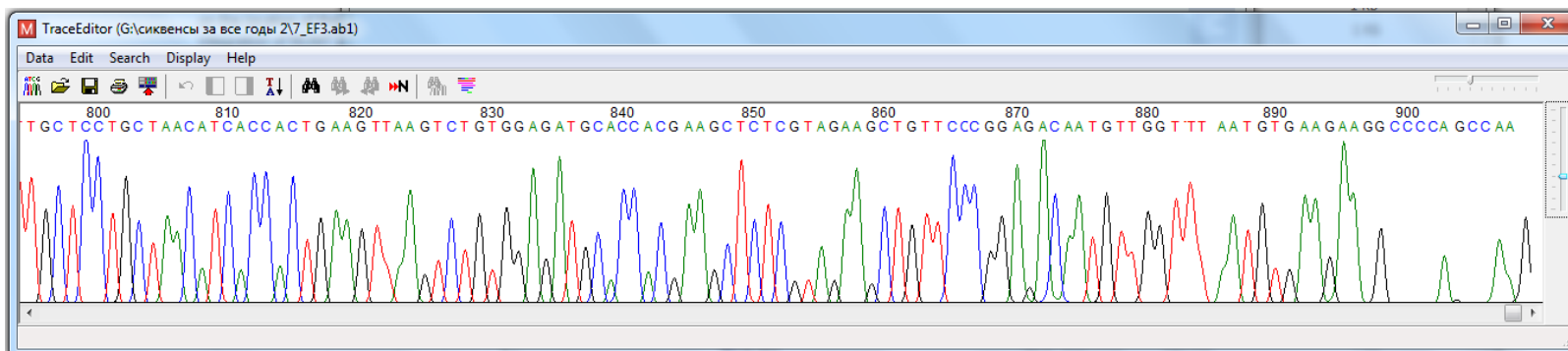
### Редактирование сиквенсов



Удаление  
некачественного  
участка в начале



Редактирование  
коротких  
некачественных  
участков



Удаление  
некачественного  
участка в конце

Анализ данных





## 2. Подготовка данных для анализа

Формат файла,  
приходящего с  
секвенатора



Формат файла для  
дальнейшего анализа

`.ab1`

Хроматограмма и буквенная  
последовательность

`.fas`

Только буквенная  
последовательность



## 2. Подготовка данных для анализа

### Получение консенсусной последовательности (при необходимости)

3'-attccgaatcATGCCTAGCTAGGCT-5'

Прямая последовательность

3'-cctaggcattAGCCTAGCTAGGCAT-5'

Обратная последовательность

3'-attccgaatcATGCCTAGCTAGGCTaatgcctagg-5'

Консенсусная последовательность

Получение обратной  
комплементарной  
последовательности

Выравнивание  
последовательностей

Получение  
консенсусной  
последовательности

Для  
последовательности,  
полученной с  
обратным  
праймером

*Консенсусная последовательность* – «обобщающая»  
последовательность, полученная из нескольких  
последовательностей



## 2. Подготовка данных для анализа

### Получение консенсусной последовательности (при необходимости)

The screenshot displays a software window titled 'Assemble' with a menu open. The menu options include 'Automatically', 'Assemble to Reference', 'Interactively...', 'Build Reference Database or Index', 'Align Using', 'Align Data Files to Ref Using', 'Assemble Data Files Using', 'RNA-Seq Using Cufflinks...', 'Merge Cufflinks Alignments with Cuffmerge...', 'RNA-Seq Differential Expression Using Cuffdiff...', 'Quantify RNA-Seq Data Using Cuffquant...', and 'Normalization Using Cuffnorm...'. The 'Align Using' menu is expanded, showing 'Clustal' and 'MUSCLE' as options. Below the menu, a large window shows a multiple sequence alignment of DNA sequences from various bacterial samples (e.g., Bacterial sample13, Bacterial sample14, etc.). The sequences are color-coded by nucleotide (A, C, G, T). At the bottom, a consensus sequence is shown: `TCWATTTCTCKACYGGMTCRYTWWGGBTAGATGTHGCRCITDGGDATYGGYGGYTTTRCCYATGGGWCGAATTGTWGA`. A legend below the consensus sequence indicates: 'D) highlight base call disagreements', '(+) highlight ambiguities', and '• highlight ambiguities'.

Выравнивание последовательностей



Получение консенсусной последовательности

Анализ данных



Barcoding

ДНК-штрихкод



Выравнивание  
последовательности в  
системах BOLD или BLAST



Оценка качества  
выравнивания



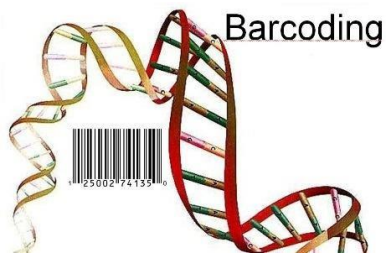
Вид

## 2. Идентификация вида

Алгоритм действия

Анализ данных





### 3. Идентификация вида

## Основные алгоритмы выравнивания последовательностей

MUSCLE

Допускает множественное  
выравнивание

Быстрый

Расслабленный

CLUSTAL

Допускает множественное  
выравнивание

Медленный

Строгий



# 3. Идентификация вида

## Оценка результатов выравнивания



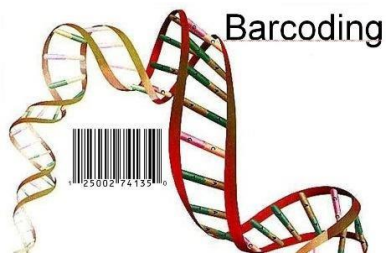
Визуальная оценка

Параметральная оценка

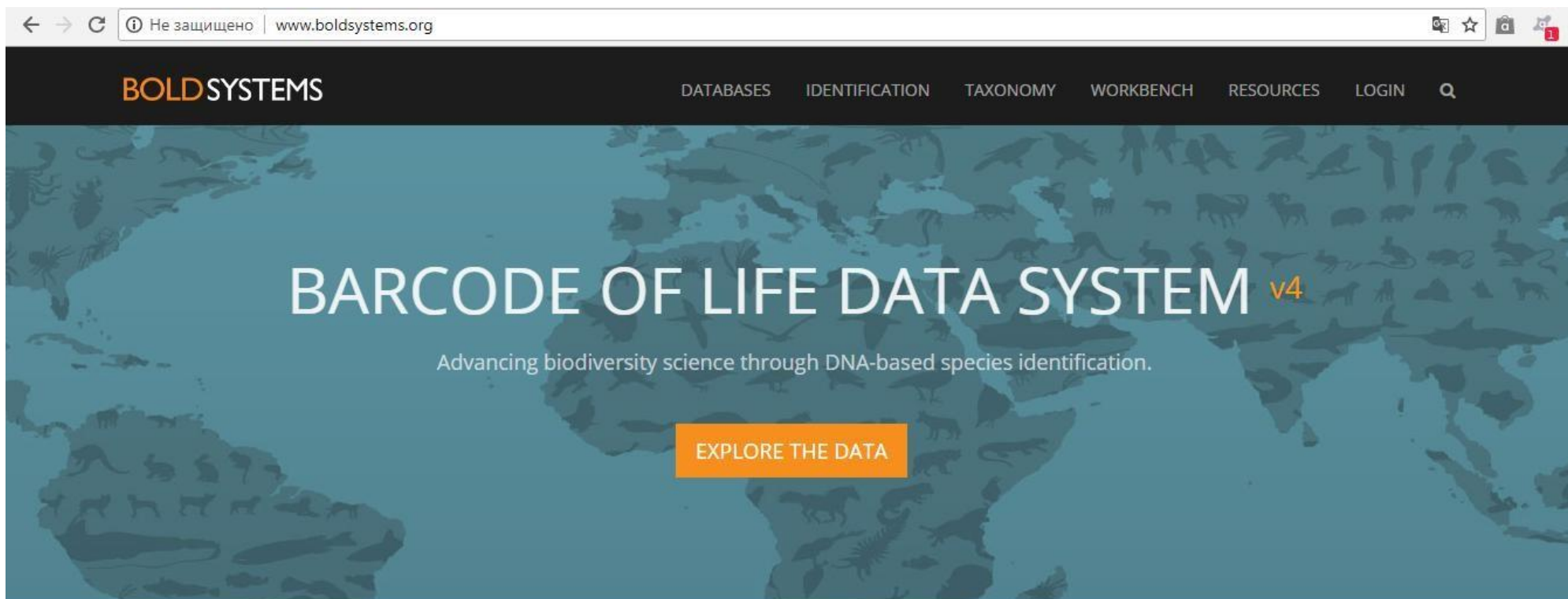
*Nucleotide identity* - процент совпадающих нуклеотидов

*Query coverage* - процент перекрытия анализируемой последовательности с последовательностями сравнения

Анализ данных



# 3. Идентификация вида BOLD

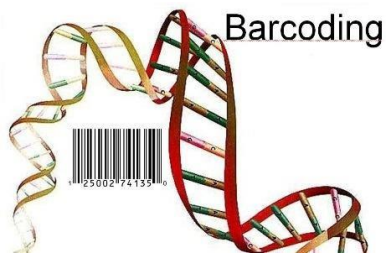


## DESIGNED TO SUPPORT THE GENERATION & APPLICATION OF DNA BARCODE DATA

BOLD is a cloud-based data storage and analysis platform developed at the Centre for Biodiversity Genomics in Canada. It consists of four main modules, a data portal, an educational portal, a registry of BINs (putative species), and a data collection and analysis workbench.

Please note that this version of BOLD is in beta and will contain bugs. Users can help address these bugs by testing the system and reporting issues to [support@boldsystems.org](mailto:support@boldsystems.org). This version is very different from the prior one but has access to all the same data.





# 3. Идентификация вида BOLD

www.barcodinglife.org/index.php/IDS\_OpenIdEngine

**BOLDSYSTEMS** Databases | Taxonomy | Identification | Workbench | Resources

## Identification Request Print

**Animal Identification [COI]** | **Fungal Identification [ITS]** | **Plant Identification [rbcL & matK]**

The BOLD Identification System (IDS) for COI accepts sequences from the 5' region of the mitochondrial Cytochrome c oxidase subunit I gene and returns a species-level identification when one is possible. Further validation with independent genetic markers will be desirable in some forensic applications.

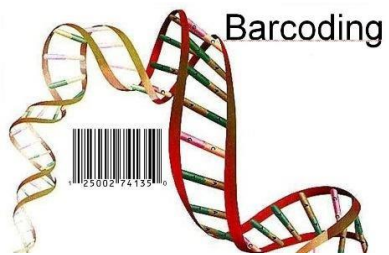
**Historical Databases:** [Jul-2013](#) [Jul-2012](#) [Jul-2011](#) [Jul-2010](#) [Jul-2009](#)

Search Databases:

- All Barcode Records on BOLD (2,240,132 Sequences)**  
Every COI barcode record on BOLD with a minimum sequence length of 500bp (warning: unvalidated library and includes records without species level identification). This includes many species represented by only one or two specimens as well as all species with interim taxonomy. This search only returns a list of the nearest matches and does not provide a probability of placement to a taxon.
- Species Level Barcode Records (1,540,956 Sequences/138,653 Species/56,354 Interim Species)**  
Every COI barcode record with a species level identification and a minimum sequence length of 500bp. This includes many species represented by only one or two specimens as well as all species with interim taxonomy.
- Public Record Barcode Database (451,983 Sequences/57,169 Species/14,126 Interim Species)**  
All published COI records from BOLD and GenBank with a minimum sequence length of 500bp. This library is a collection of records from the published projects section of BOLD.
- Full Length Record Barcode Database (1,232,554 Sequences/126,054 Species/49,225 Interim Species)**  
Subset of the Species library with a minimum sequence length of 640bp and containing both public and private records. This library is intended for short sequence identification as it provides maximum overlap with short reads from the barcode region of COI.

Enter sequences in fasta format:

Анализ данных



# 3. Идентификация вида BOLD

www.barcodinglife.org/index.php/IDS\_IdentificationRequest

**BOLD**SYSTEMS Databases | Taxonomy | Identification | Workbench | Resources

Specimen Identification Request [Print](#)

▼ Query: unlabeled\_sequence Top Hit: Arthropoda - Hemiptera - Brachycaudus sp. A rgf-2008 (100%)

Search Request:

Type: COI FULL DATABASE (includes records without species designation)

Search Result:

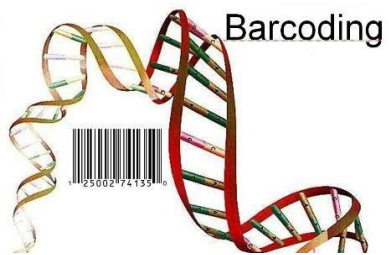
**Tree Based Identification**

Similarity scores of the top 99 matches:

TOP 20 Matches: Display option: Top 20 ▼

Phylum	Class	Order	Family	Genus	Species	Similarity (%)	Status
Arthropoda	Insecta	Hemiptera	Aphididae	Brachycaudus	<i>sp. A rgf-2008</i>	100	Published <a href="#">↗</a>
Arthropoda	Insecta	Hemiptera	Aphididae	Brachycaudus	<i>tragopogonis</i>	100	Early-Release
Arthropoda	Insecta	Hemiptera	Aphididae	Brachycaudus	<i>tragopogonis</i>	100	Early-Release
Arthropoda	Insecta	Hemiptera	Aphididae	Brachycaudus	<i>tragopogonis</i>	100	Early-Release
Arthropoda	Insecta	Hemiptera	Aphididae	Brachycaudus	<i>tragopogonis</i>	100	Published <a href="#">↗</a>
Arthropoda	Insecta	Hemiptera	Aphididae	Brachycaudus	<i>prunicola</i>	100	Published <a href="#">↗</a>
Arthropoda	Insecta	Hemiptera	Aphididae	Brachycaudus	<i>schwartzi</i>	100	Published <a href="#">↗</a>
Arthropoda	Insecta	Hemiptera	Aphididae	Brachycaudus	<i>schwartzi</i>	100	Published <a href="#">↗</a>
Arthropoda	Insecta	Hemiptera	Aphididae	Brachycaudus	<i>schwartzi</i>	100	Published <a href="#">↗</a>
Arthropoda	Insecta	Hemiptera	Aphididae	Brachycaudus	<i>schwartzi</i>	100	Published <a href="#">↗</a>
Arthropoda	Insecta	Hemiptera	Aphididae	Brachycaudus	<i>schwartzi</i>	100	Published <a href="#">↗</a>
Arthropoda	Insecta	Hemiptera	Aphididae	Brachycaudus	<i>schwartzi</i>	100	Published <a href="#">↗</a>
Arthropoda	Insecta	Hemiptera	Aphididae	Brachycaudus	<i>schwartzi</i>	100	Published <a href="#">↗</a>
Arthropoda	Insecta	Hemiptera	Aphididae	Brachycaudus	<i>schwartzi</i>	100	Published <a href="#">↗</a>
Arthropoda	Insecta	Hemiptera	Aphididae	Brachycaudus	<i>schwartzi</i>	100	Published <a href="#">↗</a>
Arthropoda	Insecta	Hemiptera	Aphididae	Brachycaudus	<i>schwartzi</i>	100	Published <a href="#">↗</a>
Arthropoda	Insecta	Hemiptera	Aphididae	Brachycaudus	<i>schwartzi</i>	100	Published <a href="#">↗</a>
Arthropoda	Insecta	Hemiptera	Aphididae	Brachycaudus	<i>schwartzi</i>	100	Published <a href="#">↗</a>
Arthropoda	Insecta	Hemiptera	Aphididae	Brachycaudus	<i>schwartzi</i>	100	Published <a href="#">↗</a>

Анализ данных



# 3. Идентификация вида

# BLAST

Анализ данных

NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information Sign in to NCBI

BLAST® Home Recent Results Saved Strategies Help

### Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. [Learn more](#)

**NEWS**  
The BLAST widget - integrating your BLAST results into NCBI's Genome Data Viewer!  
Tue, 31 Jul 2018 18:00:00 EST [More BLAST news...](#)

### Web BLAST

**blastx**  
translated nucleotide ► protein

**tblastn**  
protein ► translated nucleotide

**Nucleotide BLAST**  
nucleotide ► nucleotide

NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information Sign in to NCBI

BLAST® » blastn suite Home Recent Results Saved Strategies Help

### Standard Nucleotide BLAST

blastn blasto blastx tblastn tblastx

BLASTN programs search nucleotide databases using a nucleotide query. [more...](#) [Reset page](#) [Bookmark](#)

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#) Query subrange [From](#) [To](#)

Or, upload file [Выберите файл](#) [Файл не выбран](#)

Job Title [Enter a descriptive title for your BLAST search](#)

Align two or more sequences

Choose Search Set

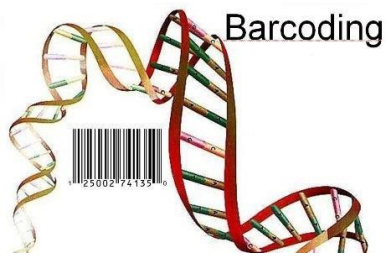
Database  Human genomic + transcript  Mouse genomic + transcript  Others (nr etc.):  
Nucleotide collection (nr/nt)

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](#)

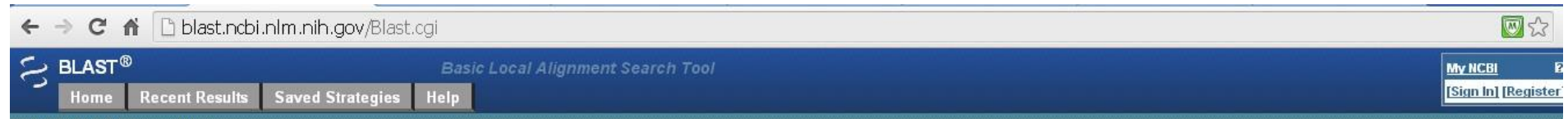
Exclude Optional  Models (XM/XP)  Uncultured/environmental sample sequences

Limit to Optional  Sequences from type material

Entrez Query Optional [You Tube](#) [Create custom database](#)  
[Enter an Entrez query to limit search](#)



# 3. Идентификация вида. BLAST



NCBI/BLAST/blastn suite/ Formatting Results - DZHAG2MB01N

## Nucleotide Sequence (462 letters)

RID [DZHAG2MB01N](#) (Expires on 01-23 20:19 pm)  
Query ID |cl|125421  
Description None  
Molecule type nucleic acid  
Query Length 462

Database Name nr  
Description Nucleotide collection (nt)  
Program BLASTN 2.2.29+ [Citation](#)

Other reports: [Search Summary](#) [Taxonomy reports](#) [Distance tree of results](#)

## 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	Ident	Accession
<input type="checkbox"/>	<a href="#">Macrosiphum gei isolate 09-11 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial</a>	854	854	100%	0.0	100%	<a href="#">JF340102.1</a>
<input type="checkbox"/>	<a href="#">Macrosiphum euphorbiae voucher CNC#HEM051851 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial</a>	837	837	100%	0.0	99%	<a href="#">EU701729.1</a>
<input type="checkbox"/>	<a href="#">Macrosiphum melampyri isolate 10-487 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial</a>	832	832	100%	0.0	99%	<a href="#">JF340104.1</a>
<input type="checkbox"/>	<a href="#">Macrosiphum gei isolate 09-13 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial</a>	832	832	98%	0.0	99%	<a href="#">JF340096.1</a>
<input type="checkbox"/>	<a href="#">Macrosiphum daphnidis voucher CNC#HEM054303 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial</a>	826	826	100%	0.0	99%	<a href="#">EU701724.1</a>
<input type="checkbox"/>	<a href="#">Macrosiphum sp. C1774 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial</a>	826	826	100%	0.0	99%	<a href="#">EU189671.1</a>

*Macrosiphum (Macrosiphum) gei* (Koch, 1855)  
462 bp. Код доступа GenBank: JF340102

```

1 ggtataattg gatgctctct tagaatttta attcgattag aattaagaca aattaattct
61 attattaata ataataaatt atataatgta attgttacaa ttcatgcttt tattataaatt
121 tttttataa ctataccaat tgtaattggt ggatttgaa attgattaat tcoataata
181 ataggatgc ctgatatac atttccaagt ttaaataata ttagattttg attattaact
241 ccatcattaa taataaatt ttttagattt ttaataata atggaacagg aacaggatga
301 acaatttacc cccctttatc aaacaatatt gcacataata acatttcagt tgatttaact
361 atttttctt tacatttagc aggaattcoa tcaattttg gagcaattaa ctttattttg
421 acaattctta atataatacc aaacaatata aaattaaatc aa

```

blast.ncbi.nlm.nih.gov/Blast.cgi

Download GenBank Graphics

Macrosiphum gei isolate 09-11 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial  
Sequence ID: [gb|JF340102.1](#) Length: 462 Number of Matches: 1

Score	Expect	Identites	Gaps	Strand	Plus/Minus
854 bits(462)	0.0	462/462(100%)	0/462(0%)		

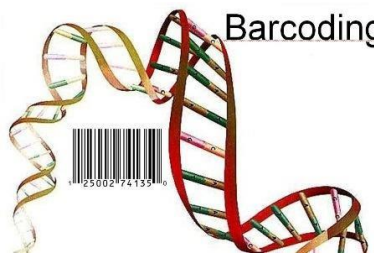
Range 1: 1 to 462 [GenBank](#) [Graphics](#) [Next Match](#) [Previous Match](#)

```

Query 1  GGTATAATTGGATCCTCTCTAGAAATTTCAATGGATTGGAAATTTGAATTCCTATAATA
Sbjct 1  GGTATAATTGGATCCTCTCTAGAAATTTCAATGGATTGGAAATTTGAATTCCTATAATA
Query 61  ATTAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATA
Sbjct 61  ATTAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATA
Query 121  CCAATCATTAAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATA
Sbjct 121  CCAATCATTAAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATA
Query 181  ATAGGATGCTCTGATATATCATTCACGTTAAATATATAGATTTTGATTTATTAACCT
Sbjct 181  ATAGGATGCTCTGATATATCATTCACGTTAAATATATAGATTTTGATTTATTAACCT
Query 241  CCATCATTAAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATA
Sbjct 241  CCATCATTAAATAATAATAATAATAATAATAATAATAATAATAATAATAATAATA
Query 301  ACAATTTACCCCTTTATCAACAGATATGACAGATATACATTTCAAGTTGATTAAC
Sbjct 301  ACAATTTACCCCTTTATCAACAGATATGACAGATATACATTTCAAGTTGATTAAC
Query 361  ATTTTCTTTACATTTAGCAGAAATCTCATCAATTTAGAGCAATTAACCTATTTT
Sbjct 361  ATTTTCTTTACATTTAGCAGAAATCTCATCAATTTAGAGCAATTAACCTATTTT
Query 421  ACAATTTCAATATAATCAACAATAATAATAATAATAATAATAATAATAATAATA
Sbjct 421  ACAATTTCAATATAATCAACAATAATAATAATAATAATAATAATAATAATAATA

```

Анализ данных



# 3. Идентификация вида. BLAST

www.ncbi.nlm.nih.gov/nuccore/JF340102

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide Search Limits Advanced Help

Display Settings: GenBank Send:

**Macrosiphum gei isolate 09-11 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial**

GenBank: JF340102.1  
[FASTA](#) [Graphics](#) [PopSet](#)

Go to:

LOCUS JF340102 462 bp DNA linear INV 14-MAR-2011  
 DEFINITION *Macrosiphum gei* isolate 09-11 cytochrome oxidase subunit 1 (COI) gene, partial cds; mitochondrial.  
 ACCESSION JF340102  
 VERSION JF340102.1 GI:325560635  
 KEYWORDS .  
 SOURCE mitochondrion *Macrosiphum gei*  
 ORGANISM [Macrosiphum gei](#)  
 Eukaryota; Metazoa; Ecdysozoa; Arthropoda; Hexapoda; Insecta;  
 Pterygota; Neoptera; Paraneoptera; Hemiptera; Sternorrhyncha;  
 Aphidiformes; Aphidoidea; Aphididae; Macrosiphini; Macrosiphum.  
 REFERENCE 1 (bases 1 to 462)  
 AUTHORS Voronova,N.V., Kurchenko,V.P. and Buga,S.V.  
 TITLE The nucleotide partial sequences of cytochrome c oxidase subunit I (COI) gene of some aphid species from Belarus region  
 JOURNAL Unpublished  
 REFERENCE 2 (bases 1 to 462)  
 AUTHORS Voronova,N.V., Kurchenko,V.P. and Buga,S.V.  
 TITLE Direct Submission  
 JOURNAL Submitted (12-FEB-2011) Zoology, BSU, Kurchatova 10, Minsk, Minsk Region 220000, Belarus  
 FEATURES Location/Qualifiers

Change region shown

Customize view

Analyze this sequence

Run BLAST  
 Pick Primers  
 Highlight Sequence Features  
 Find in this Sequence

Related information

Related Sequences  
 PopSet  
 Protein  
 Taxonomy

Recent activity

Turn Off Clear

Macrosiphum gei isolate 09-11 cytochrome oxidase subunit 1 (COI) gene, part Nucleotide  
 trid436495(Organism:popul(3))



## 2. Идентификация вида. Алгоритм действия

Анализ данных

Определение вида по  
морфологии,  
ДНК-штрихкодирование,  
помещение штрихкода в  
базу данных

*Работа других исследователей*

ДНК-штрихкод

Выравнивание  
последовательности в  
системах BOLD или BLAST

Оценка качества  
выравнивания

Вид

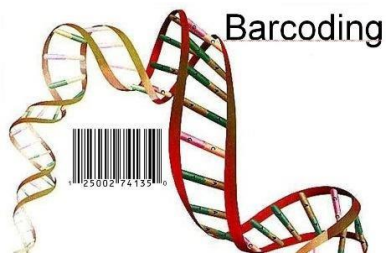


## 3. Идентификация вида

В базе нет референса

Что делать ?!!





### 3. Идентификация вида

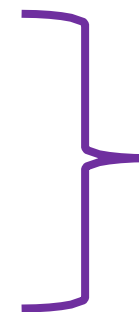
В базе нет референса ?

**Иди к систематикам !!**

Анализ данных



1. Таксономист
2. Ваучерный образец
3. Хорошая этикетка



решают  
ВСЕ проблемы



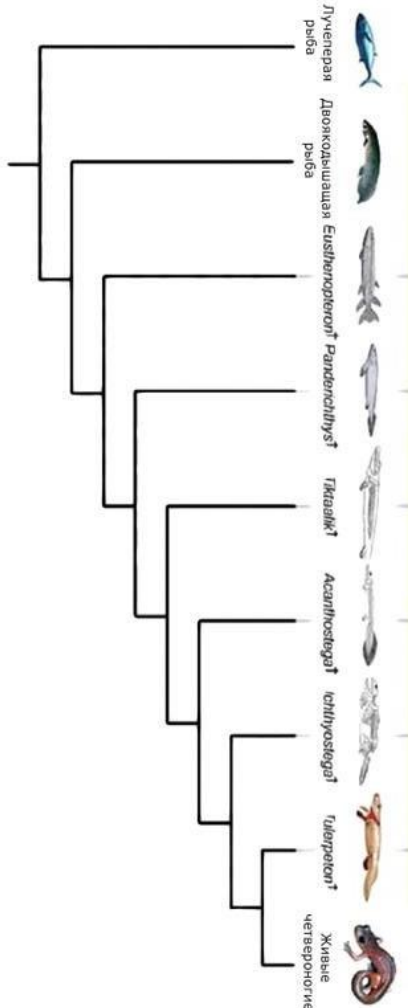


# 3. Идентификация вида

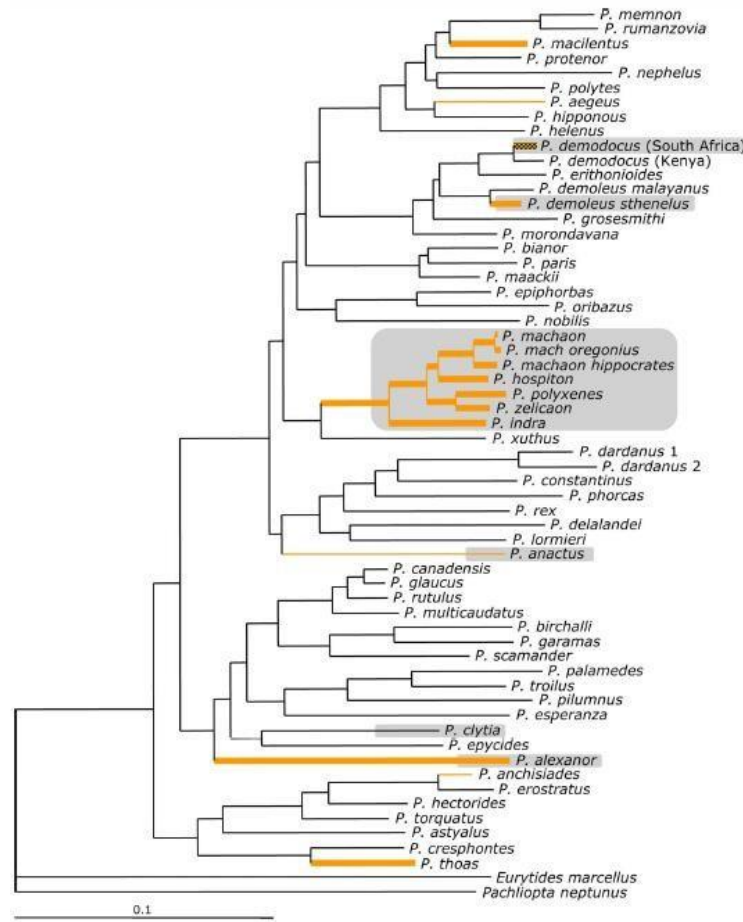
## Использование филогенетического анализа

Суть метода

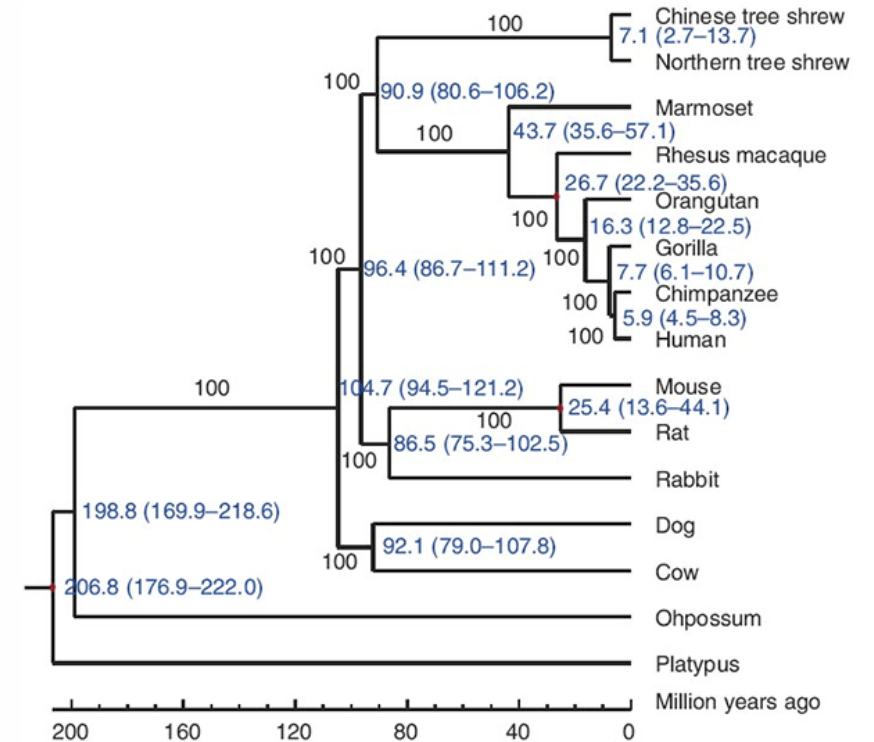
Кладограмма



Филогенетическое дерево



Хронограмма

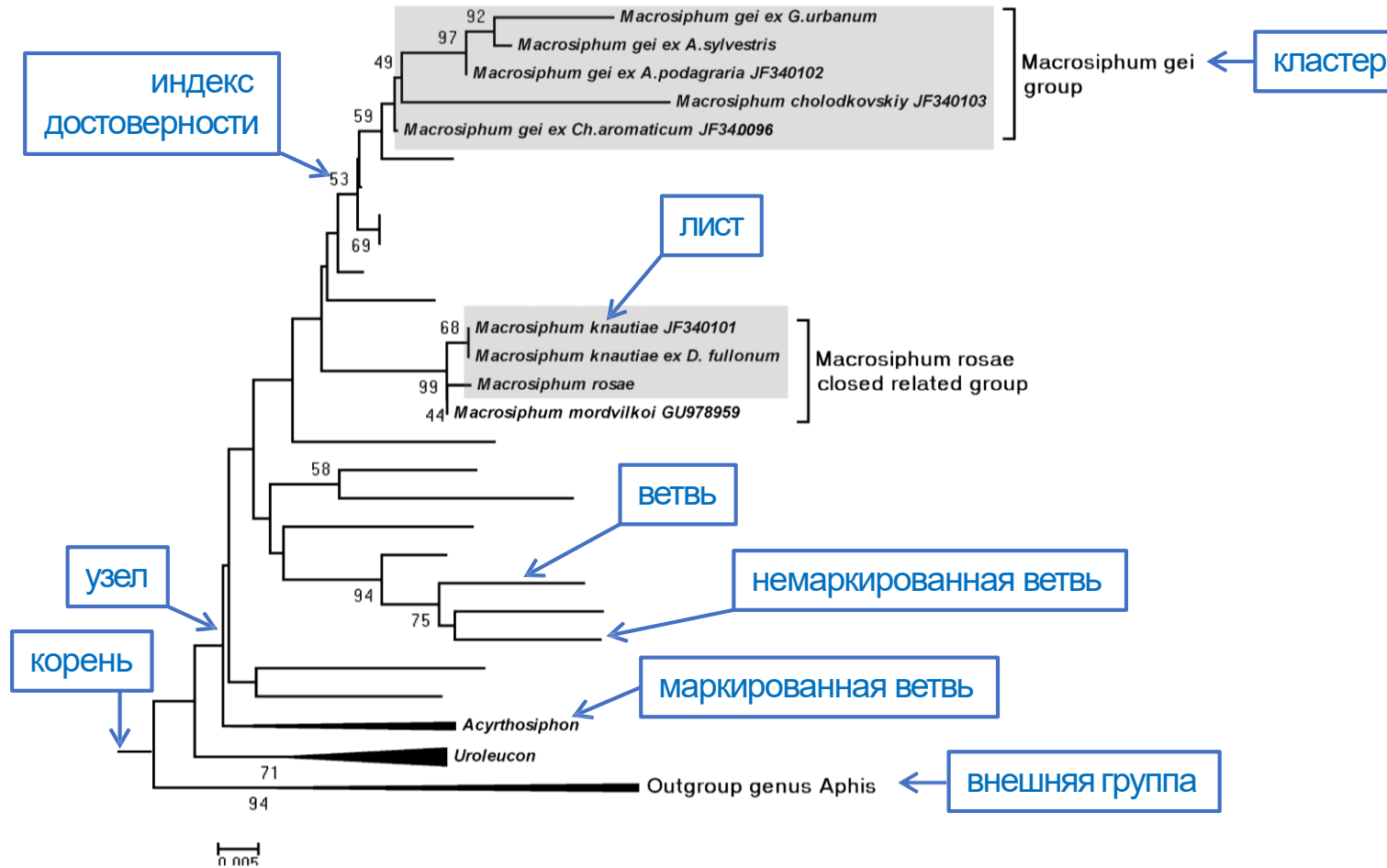


Анализ данных



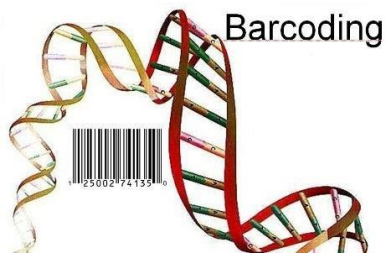
# 3. Идентификация вида

## Использование филогенетического анализа



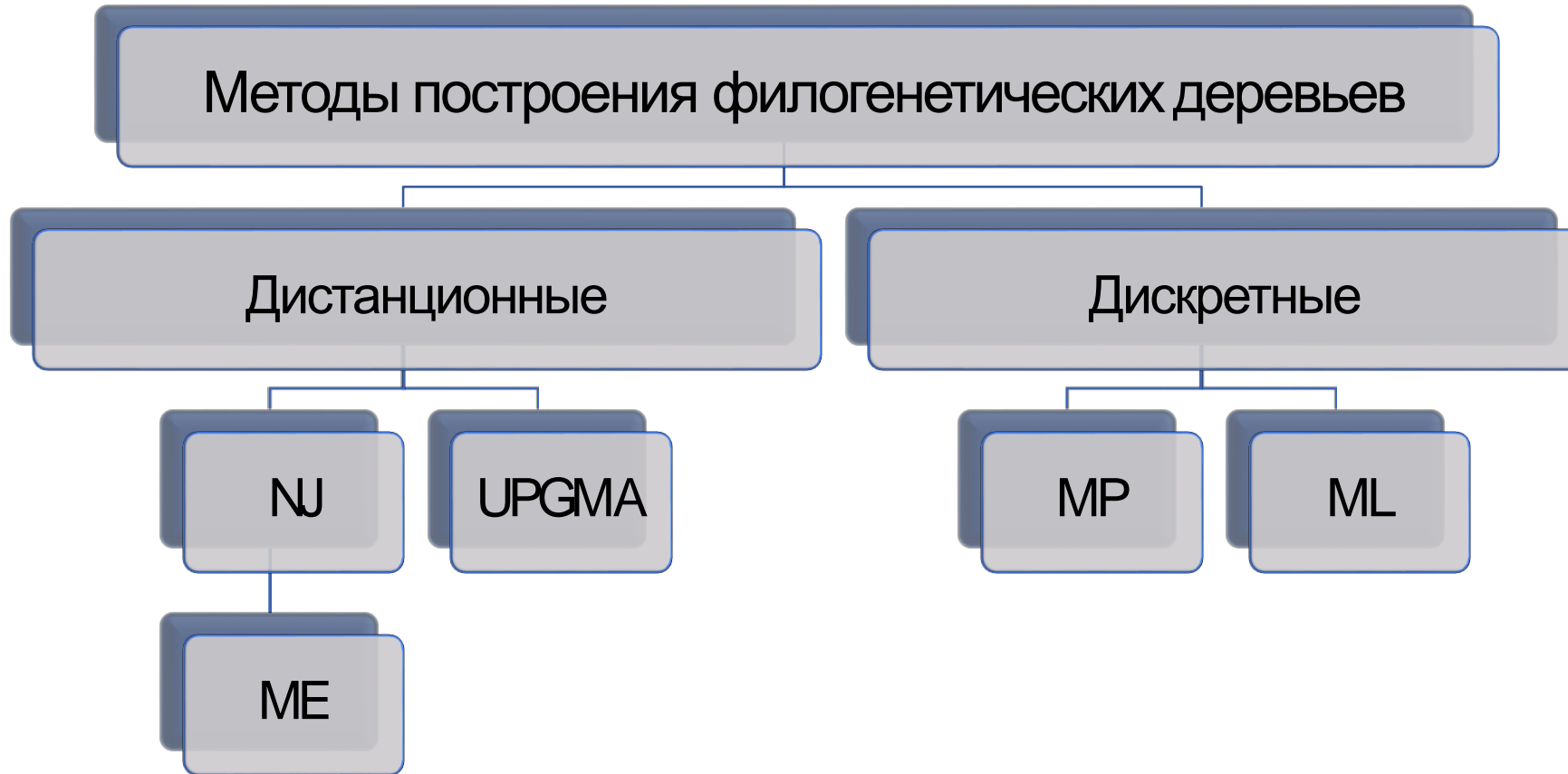
Анализ данных

Филогенетическое дерево



# 3. Идентификация вида

## Использование филогенетического анализа



Анализ данных

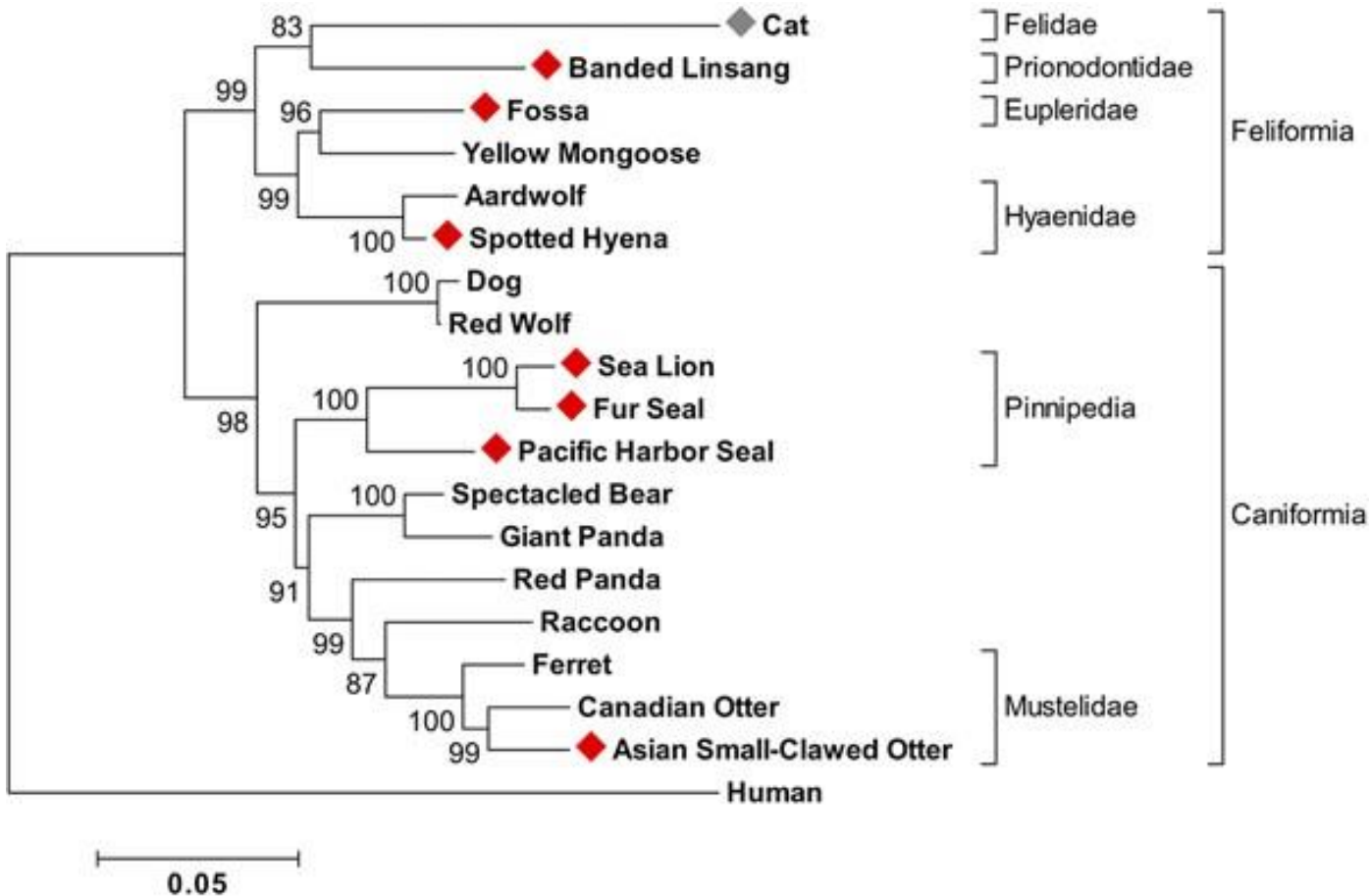
Основные методы построения деревьев



# 3. Идентификация вида

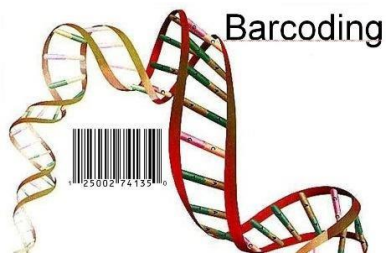
## Использование филогенетического анализа

### Статистическая проверка топологии дерева



**Бутстрэп-анализ** — практический компьютерный метод определения статистик вероятностных распределений, основанный на многократной генерации выборок на базе имеющейся выборки

Анализ данных



# 3. Идентификация вида

## Использование филогенетического анализа

Исходные последовательности

```
000000000111111111122  
123456789012345678901  
1AATCGTTCGATGACTTCGGAG  
2AAGGACTCGTCGACCTCAGAG
```



```
000000000111111111122  
1256787801212345678801  
1AAGTTCATGTGACTTCGGAG  
2AAACTCTCTCGCGACCTCAAAG
```



```
000000000111111122122  
456789789014545901901  
1CGTTCGTCGATCTCTGAGGAG  
2GACTCGTCGTCGCCGAGGAG
```



Симулированные последовательности  
(реплики)



Статистическая проверка топологии дерева

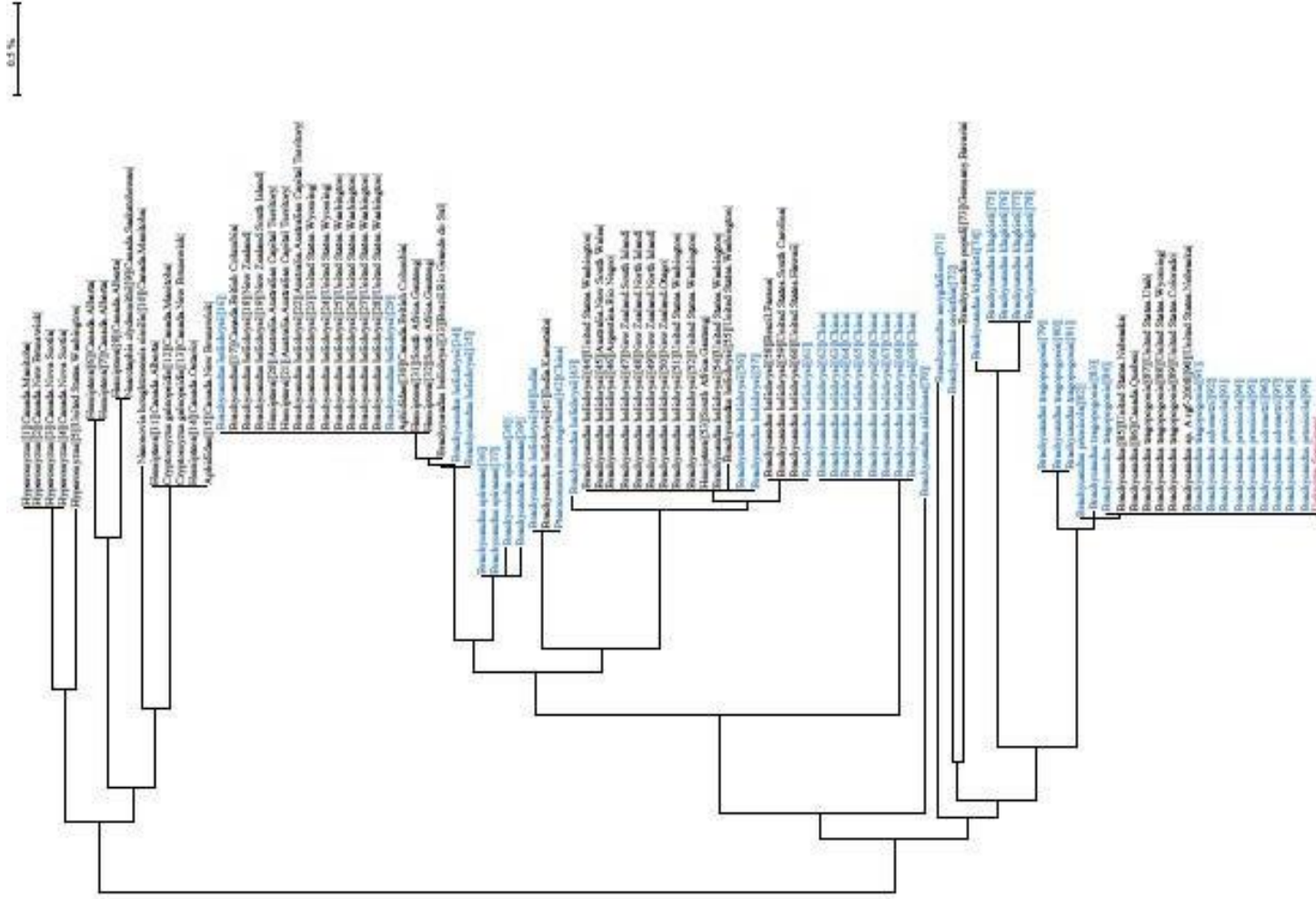
Анализ данных

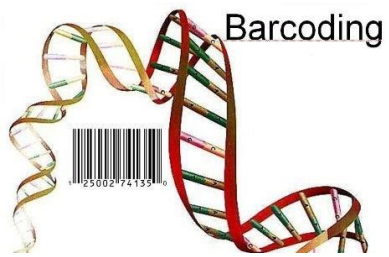


## 3. Идентификация вида

### Использование филогенетического анализа

### Филогенетическое дерево, построенное в BOLD





# Выделение нового вида на основе ДНК-штрихкода

## Криптические виды и механизмы их возникновения

**Близкие виды** – виды, обладающие определенным сходством морфологии и экологии вследствие общности происхождения

**Сестринские виды** – обычно молодые виды, обладающие высоким сходством морфологии вследствие единства происхождения

**Криптические (скрытые) виды** – виды, не различимые по морфологическим признакам, но имеющие различия в маркерных областях геномов



# Выделение нового вида на основе ДНК-штрихкода

## Генетическая (геномная) концепция вида

Генетический вид – группа особей, генетически изолированная от других таких же групп

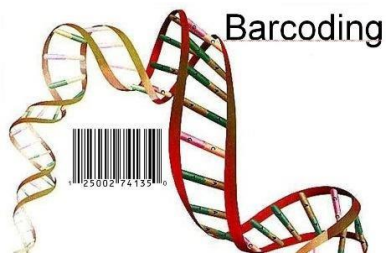


Свидетельство существования генетической изоляции: наличие «генетической дистанции» между группами



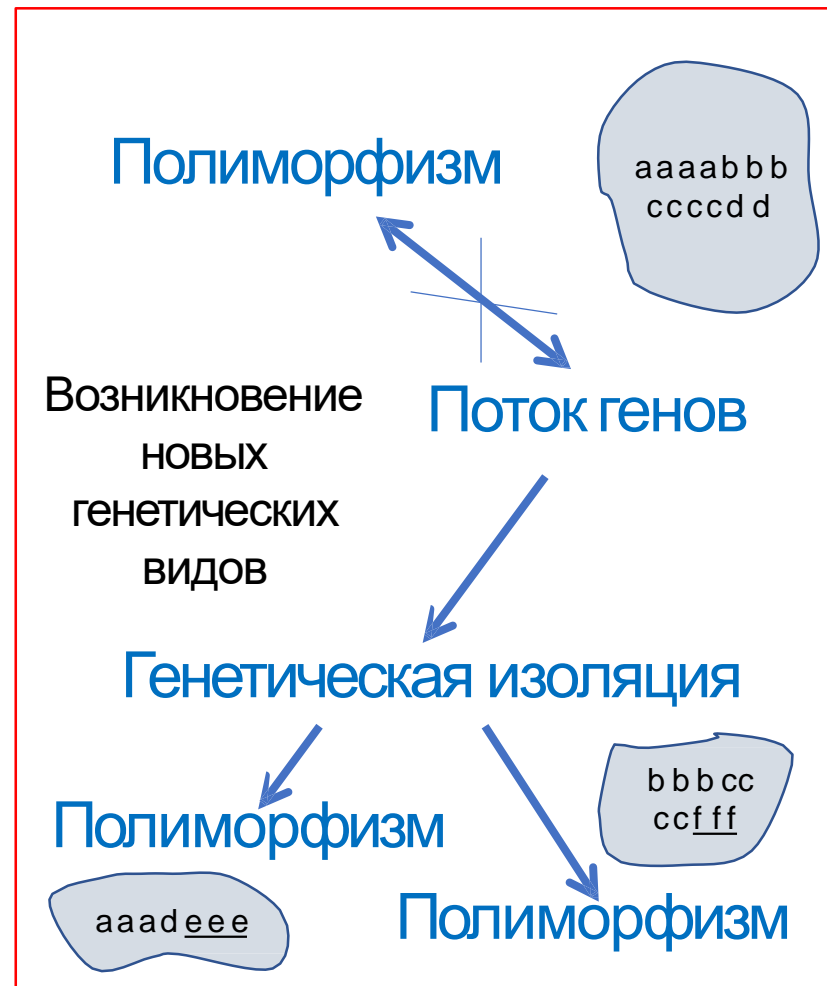
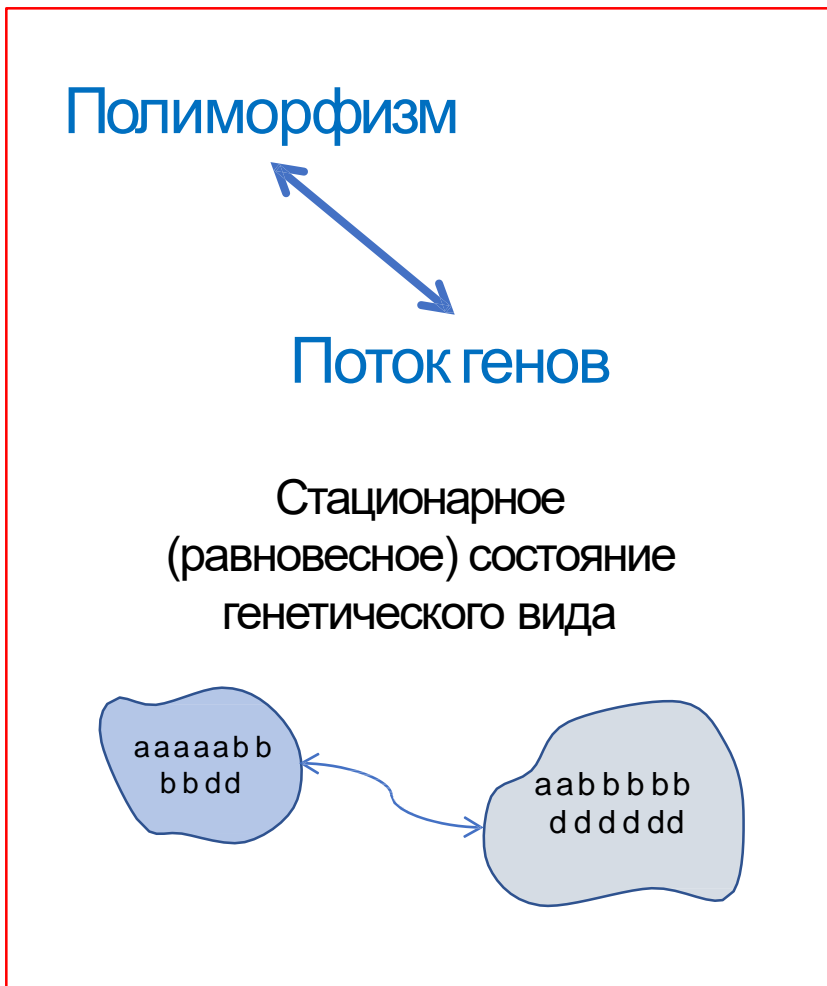
Генетическая дистанция – мера любых генетических различий между двумя особями





# Выделение нового вида на основе ДНК-штрихкода

## Генетические различия между видами





# Выделение нового вида на основе ДНК-штрихкода

## Криптические виды и механизмы их возникновения

**Генетический вид** – группа особей, генетически изолированная от других таких же групп



Свидетельство существования генетической изоляции : наличие «генетической дистанции» между группами



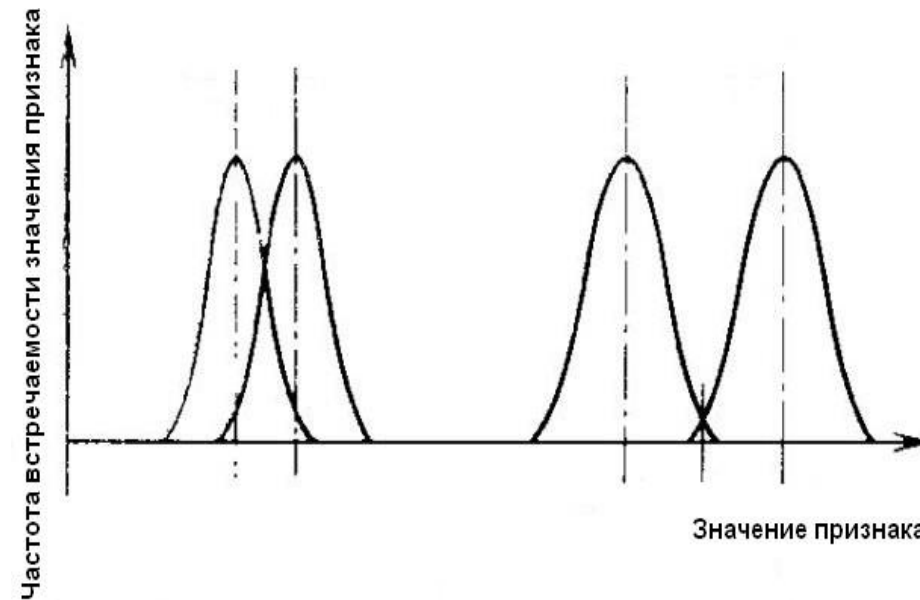
**Генетическая дистанция** – мера любых генетических различий между двумя особями

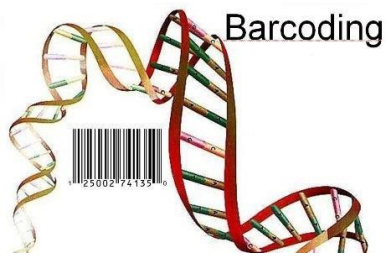


# Выделение нового вида на основе ДНК-штрихкода

## Концепция вида. Существование хиатусов между видами в значении или состоянии признака

<u>Длина тела</u>	
A	B
2	4 min
1	5
2	5
max 3	6
1	7
2	6
2	6
1	5
2	4
2	6
1	6





# Выделение нового вида на основе ДНК-штрихкода

## Генетическая дистанция

SNP – Single Nucleotide Polymorphism

$$GD = Np/N$$

$$d_{MTN} = -\frac{2\pi_A\pi_G}{\pi_R} \log\left(1 - \frac{\pi_R}{f_{AG}} p_{AG} - \frac{1}{2\pi_R} p_{RY}\right) - \frac{2\pi_T\pi_C}{\pi_Y} \log\left(1 - \frac{\pi_Y}{f_{TC}} p_{TC} - \frac{1}{2\pi_Y} p_{RY}\right) - 2\left(\pi_R\pi_Y - \frac{\pi_A\pi_G\pi_Y}{\pi_R} - \frac{\pi_T\pi_C\pi_R}{\pi_Y}\right) \times \log\left(1 - \frac{1}{f_{RY}} p_{RY}\right)$$

*Aphis fabae* group

		190	200	210
Aphis_fabae_fabae	181	ATTATAATTA	TTTTTATAAC	TATACCAATT
Aphis_fabae_philadelfi	181	.....T	.....	.....
Aphis_fabae_cirsiiacanthoidis	181	.....T	.....	.....
Aphis_fabae_solanella	181	.....T	.....	.....
		310	320	330
Aphis_fabae_fabae	301	TTATACCAC	CATCACTAAT	AATAATGATT
Aphis_fabae_philadelfi	301	.....	.....	.....A..
Aphis_fabae_cirsiiacanthoidis	301	.....	.....	.....A..
Aphis_fabae_solanella	301	.....	.....	.....A..
		400	410	420
Aphis_fabae_fabae	391	AATAATATTG	CCATAATAA	TCTTCAGTT
Aphis_fabae_philadelfi	391	.....	.....C	.....A
Aphis_fabae_cirsiiacanthoidis	391	.....	.....C	.....A
Aphis_fabae_solanella	391	.....	.....C	.....A

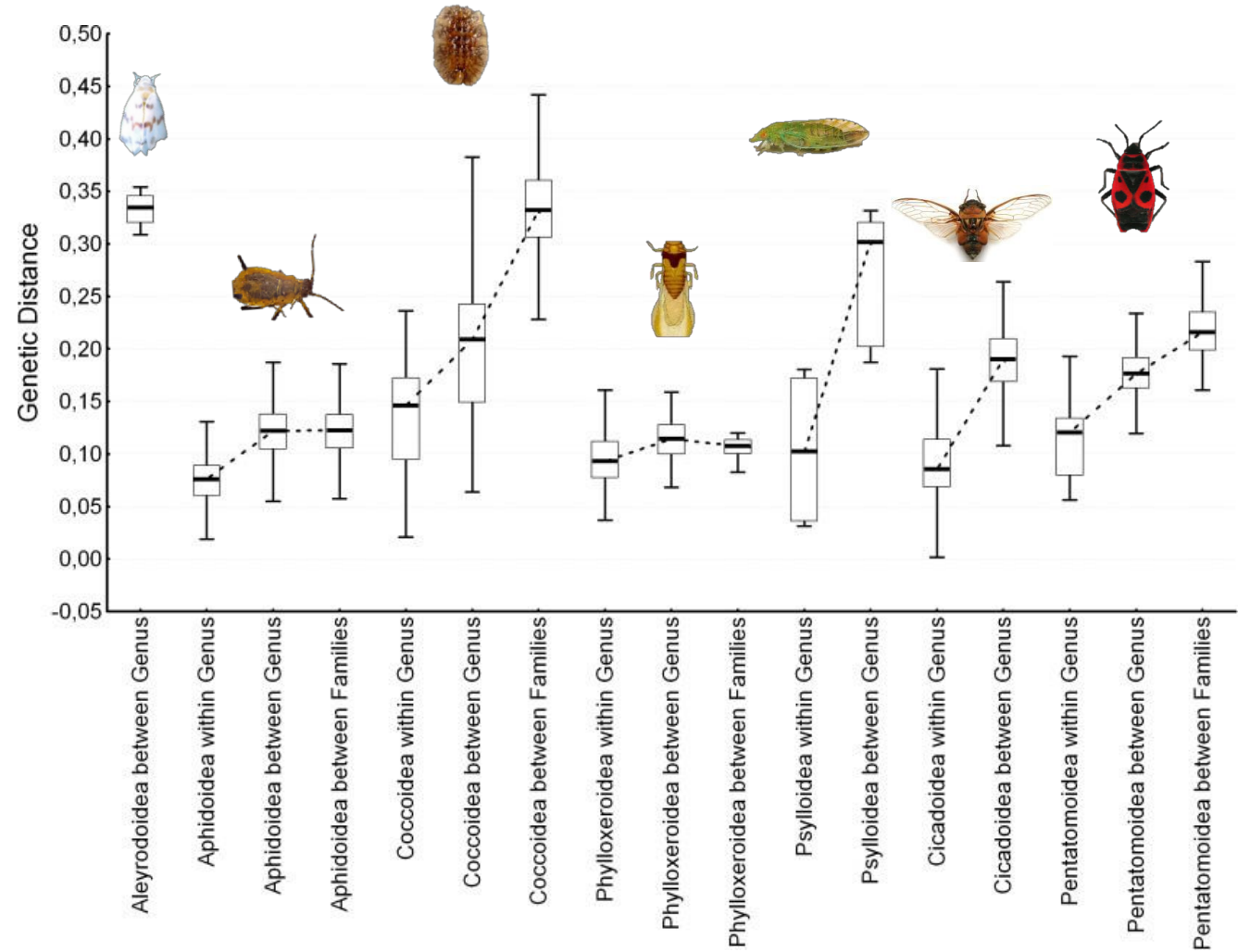
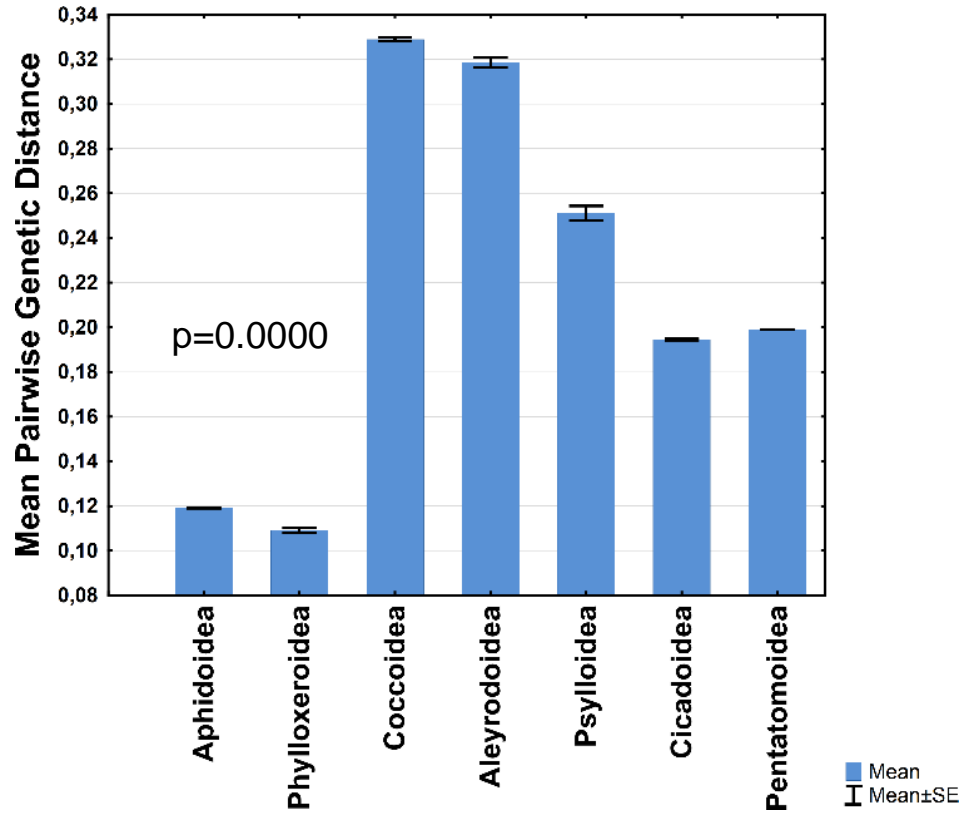
Генетическая дистанция между двумя последовательностями – доля полиморфных сайтов в последовательностях заданной длины

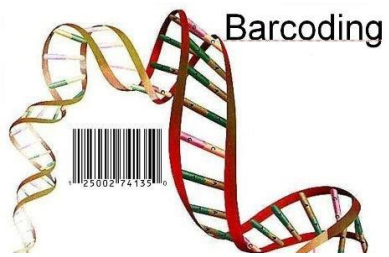
«Золотой стандарт» молекулярной таксономии – генетическая дистанция между близкими видами приблизительно равна **2%**



# Выделение нового вида на основе ДНК-штрихкода

## Особенности изменчивости COI в родственниках насекомых





# Выделение нового вида на основе ДНК-штрихкода

## Алгоритм действия

Выделение в том же роде «хороших» видов с полученными ДНК-штрихкодами



Расчет значения генетических дистанций между ними



Сравнение значений генетических дистанций



Расчет значения генетических дистанций между потенциально новым видом и сестринскими



Заключение о существовании криптического вида (*о соответствии различий межвидовому уровню*)



?



# Возможные проблемы, которые могут возникнуть в процессе ДНК-штрихкодирования

1. Отсутствие референсной последовательности для определения (отсутствие ДНК-штрихкода в базах данных)
2. Ошибочно определенный референс (ДНК-штрихкод первоначально был получен для образца, определенного неверно)

*Построение дендрограммы способно выявить проблему*

3. Контаминация образца ДНК

*Решение - только перевыделение ДНК и повторение процедуры*



# Возможные проблемы, которые могут возникнуть в процессе ДНК-штрихкодирования

## Типы контаминации образцов :

- Паразиты (особенно, при выделении из тотального образца)
- Смешанные колонии
- Механическое загрязнение посторонним биоматериалом или «перенос» микрочастиц ДНК из образца в образец



- Перепутанные (плохо маркированные) образцы





# Возможные проблемы, которые могут возникнуть в процессе ДНК-штрихкодирования

## Особые случаи

- Разные маркеры определяют один образец как разные виды
- Стандартные маркеры не работают для организмов целевого таксона
- Получить качественный сиквенс не удастся, несмотря на многократные попытки



# Возможные проблемы, которые могут возникнуть в процессе ДНК-штрихкодирования

Разные маркеры определяют один и тот же образец как разные виды. Причины

- Перепутанные образцы
- Контаминированный образец
- Гибрид





# Возможные проблемы, которые могут возникнуть в процессе ДНК-штрихкодирования

## Изучение гибридов





## Возможные проблемы, которые могут возникнуть в процессе ДНК-штрихкодирования

Стандартные маркеры не работают для организмов целевого таксона

- Слишком консервативный маркер – выявляемых различий не достаточно
- Слишком переменный маркер – выявляется много внутривидовых различий, которые снижают уверенность в получаемых результатах
- Универсальные праймеры не работают

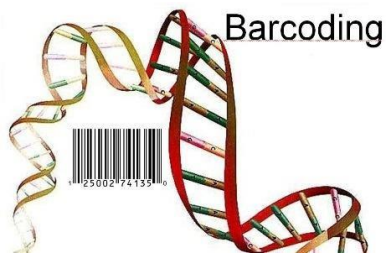


# Возможные проблемы, которые могут возникнуть в процессе ДНК-штрихкодирования

Слишком консервативный маркер

Как выглядит : *филогенетическое дерево остается «не разрешенным»*

1. Оценка вариабельности разных митохондриальных маркеров
2. Смена белок-кодирующего маркера на некодирующий



# Возможные проблемы, которые могут возникнуть в процессе ДНК-штрихкодирования

Слишком вариабельный маркер

Как выглядит : *неправдоподобно большие значения генетических дистанций между видами, последовательности на дереве «перемешиваются»*

1. Использование ядерного маркера
2. Смена митохондриального маркера или анализируемого региона
3. Работа с аминокислотной последовательностью



# Возможные проблемы, которые могут возникнуть в процессе ДНК-штрихкодирования

Универсальные праймеры не работают

*Как выглядит : ПЦР-продукт отсутствует, в то время как реагенты остаются рабочими*

1. Смена универсальных праймеров
2. Секвенирование мтДНК и дизайн новых праймеров



# Возможные проблемы, которые могут возникнуть в процессе ДНК-штрихкодирования

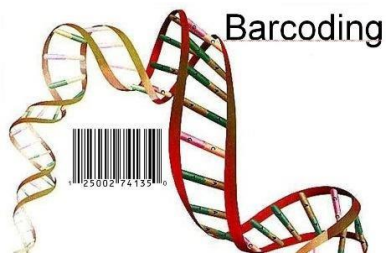
Получить качественный сиквенс не удастся

Как выглядит : *Несмотря на усилия, каждый сиквенс целевого региона оказывается «плохим»*

Причины :

1. NUMTs
2. Другие псевдогены





# Возможные проблемы, которые могут возникнуть в процессе ДНК-штрихкодирования

**Псевдогены** (англ. *pseudogenes*) - нефункциональные аналоги структурных генов, утратившие способность кодировать белок и не экспрессирующиеся в клетке. Происходят от обычных функциональных генов, утрачивают способность экспрессии в результате нонсенс-мутаций, отсутствия функциональных участков и т.п.

**Numt** (от “nuclear mitochondrial DNA”), псевдогены митохондриального происхождения, предположительно перенесенные транспозонами в ядерный геном

Species	mt genome size (bp)	Database size (bp) <sup>b</sup>	% of genome checked	No. of Numts <sup>b</sup>	Total length of Numts (bp)	Numt % of genome <sup>b</sup>
<i>Saccharomyces cerevisiae</i>	85 779	12 069 247	100	17	1389	0.012
<i>Plasmodium falciparum</i>	5967	28 718 804	50	3	228	0.0008
<i>Caenorhabditis elegans</i>	13 794	106 660 070	~100	2	212	~0.0002
<i>Drosophila melanogaster</i>	19 496	122 655 632	70	3	724	0.0006
<i>Homo sapiens</i>	16 569	2 853 531 108	84	354	418 552	0.012

<sup>a</sup>Abbreviations: bp, base pairs; Numt, nuclear mitochondrial pseudogenes.

<sup>b</sup>Summary of Numts in GenBank genome projects. Mitochondrial genomes were queried against nuclear mapped genome databases using the NCBI BLAST search facilities (<http://www.ncbi.nlm.nih.gov/BLAST>, <http://www.ncbi.nlm.nih.gov/Malaria/plasmodiumblcus>, <http://www.ncbi.nlm.nih.gov/genome/seq/HsBlast>) set at default parameters (16 February 2001). Only data with < 0.0001 probability of occurring by chance were included.

Количество NUMTs в геномах разных видов

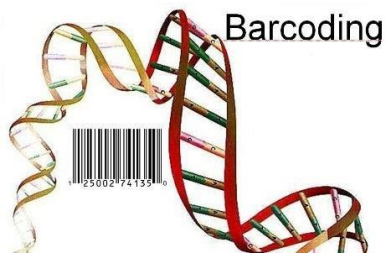
Bensasson et al.  
2001



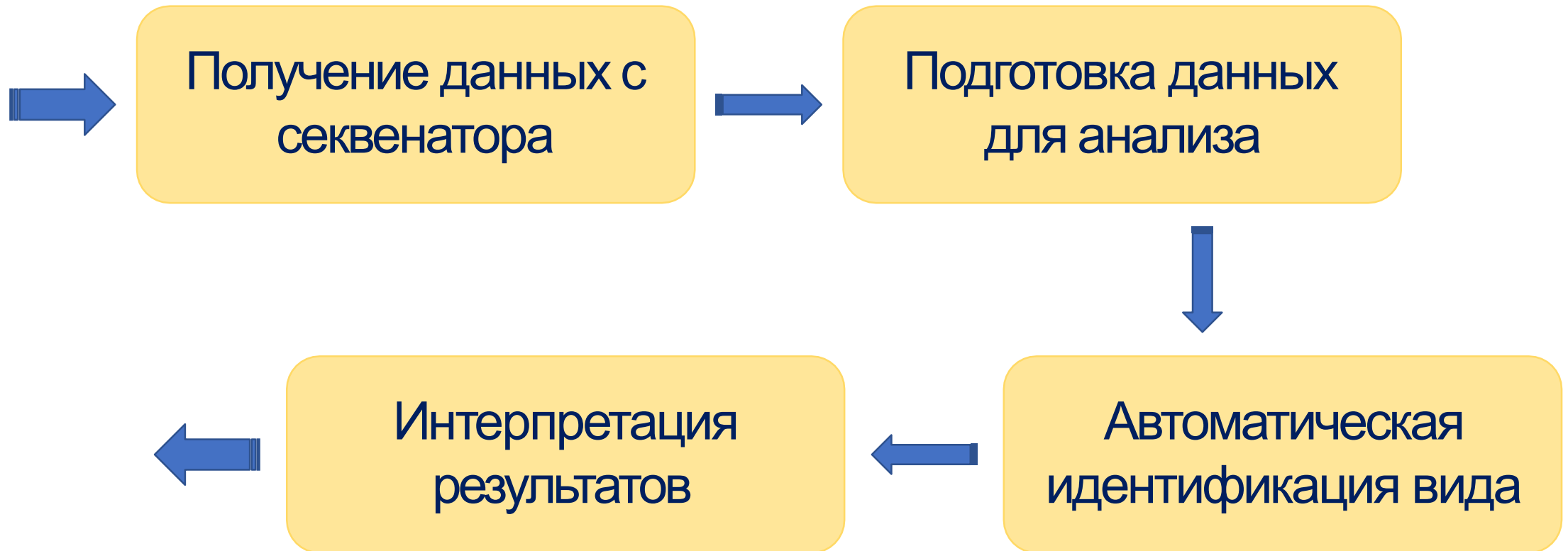
# Правило «разумной аккуратности» применения ДНК-штрихкодирования

1. Оценка разрешающей способности метода в конкретном таксоне живых организмов
2. Оценка необходимости включения в идентификацию дополнительных ДНК-маркеров
3. Оценка достоверности определения видов референсных ДНК-штрихкодов

Отклонение от правила 2 %различий по ДНК-штрихкоду между близкими видами



# Практическая часть занятия. Порядок действий





# Практическая часть занятия

## Блок 1. Идентификация

1. Отобрать из набора качественные сиквенсы
2. Провести редактирование сиквенсов и перевести их в текстовый формат
3. Провести идентификацию видов с использованием системы BOLD и BLAST.  
Сравнить получаемые результаты
4. Построить филогенетические деревья в BOLD, оценить достоверность результатов идентификации



# Практическая часть занятия

## Блок 2. Анализ полиморфизма

1. Скачать из BOLD выборку нуклеотидных последовательностей целевого таксона
2. Провести очистку данных
3. Провести выравнивание последовательностей, оценить его результаты
4. Рассчитать генетические дистанции между последовательностями
5. Выбрать лучшую модель нуклеотидных замещений
6. Построить филогенетические деревья разными методами с оценкой статистической значимости топологии, с задействованием разных позиций в кодоне