25 июня 2018г

Новые горизонты алгоритмической биоинформатики

Алла Львовна Лапидус

Professor, Department of Cytology and Histology, SPbSU
Deputy director, Centre for Algorithmic Biotechnology, SPbSU

О чем

1. Центр алгоритмической биотехнологии - немного сухой статистики
2. Что такое биоинформатика?
3. Наука в Центре алгоритмической биотехнологии
4. Образование

О чем

1. Центр алгоритмической биотехнологии - немного сухой статистики
2. Что такое биоинформатика?
3. Наука в Центре алгоритмической биотехнологии
4. Образование

Center for Algorithmic Biotechnology

St. Petersburg State University

В СПбГУ с конца 2014 года - в рамках открытого конкурса СПбГУ на создание исследовательских лабораторий под руководством ведущих ученых

«Центр алгоритмической биотехнологии» (ЦАБ)

Руководитель лаборатории – П.А. Певзнер, профессор Калифорнийского университета в Сан-Диего (University of California San Diego, USA), к.ф-м.н.

Заместитель руководителя – А.Л.Лапидус, профессор кафедры цитологии и гистологии СПбГУ, к.б.н.

«Центр алгоритмической биотехнологии» (ЦАБ)

Численность лаборатории

2015 год - 14 человек, из которых 2 кандидата наук

2018 год - 20 человек, из которых 4 кандидата наук + 1 работа принята к защите на получение степени СПбГУ

Средний возраст - 27 лет

«Центр алгоритмической биотехнологии» (ЦАБ)

Научные партнеры (**Collaborators**)

- ✓ **Yale University**
- ✓ **Scripps Institution**, UCSD
- ✓ **University of British Columbia**
- ✓ **The Wellcome Trust Sanger Institute**
- ✓ **University of Hong Kong**
- ✓ **John Hopkins University**
- ✓ **UCSD**
- ✓ **JCVI**
- ✓ **Joint Genome Institute** (JGI, LBNL)
- ✓ **EMBL-EBI**

- ✓ Brain and Mind Research Institute, (**Weill Cornell Medicine**, USA)
- ✓ I**nstitute** for Agricultural and Forest Systems in the Mediterranean (Italy)

Juno Therapeutics

Genentech

AstraZeneca Oncology

«Центр алгоритмической биотехнологии» (ЦАБ)

Научные партнеры (Collaborators)

St Petersburg:

- ✓ А.С. Глотов, СПбГУ, ИТБМ
- ✓ Центр Доброжанского, СПбГУ
- ✓ А.Н. Суворов, Институт экспериментальной медицины
- ✓ ИТМО

Moscow

Dmitry Chudakov and Mikhail Shugay, Institute of Bioorganic Chemistry

Novosibirsk

Институт цитологии и генетики СО РАН

«Центр алгоритмической биотехнологии» (ЦАБ)

2015-2018 год

Публикации – 53:

Nature family – **5**; PNAS – **3**; Genome Biology – **2**; Genome research – **2**

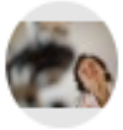
Доклады на международных конференциях – 62

Преимущественно приглашенные

РИД

Государственную регистрацию прошли **5** программных продуктов и подготовлены к регистрации еще **2**

Количество цитирований публикаций в системе Google Scholar

10. 

Alla Lapidus
St. Petersburg State University; Санкт Петербургский государственный университет
Подтвержден адрес электронной почты в домене spbu.ru
bioinformatics genomics MOOC

Цитируется: 22 464

Все позиции с 1-ой по 9-ую и с 11 по 20-ую занимают мужчины…

20. 

Alexey Gurevich
St. Petersburg State University, Center for Algorithmic Biotechnology
Подтвержден адрес электронной почты в домене spbu.ru
Bioinformatics Computational Biology Genome Assembly Metabolomics

Цитируется: 5 229

https://scholar.google.ru/citations?view_op=view_org&hl=ru&org=2752555203031532274&before_author=BXpy_5pWAAAJ&astart=0

Премии за высокую цитируемость

О чем

1. Центр алгоритмической биотехнологии - немного сухой статистики
2. Что такое биоинформатика?
3. Наука в Центре алгоритмической биотехнологии ,,,
4. Образование

Term Bioinformatics

Термин **Биоинформатика** был введен в 1970-ом году голландскими биологами Паулиеной Хогвег и Беном Хеспере, изучавшими свойства сложных динамических биологических систем

Что такое биоинформатика?

Компьютеры в биологии

Программирование

Математика

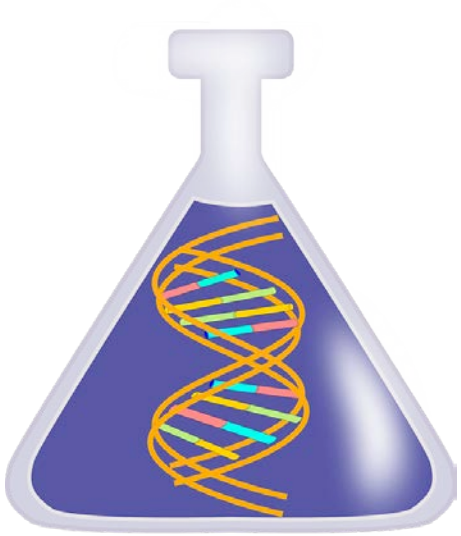
BIG data

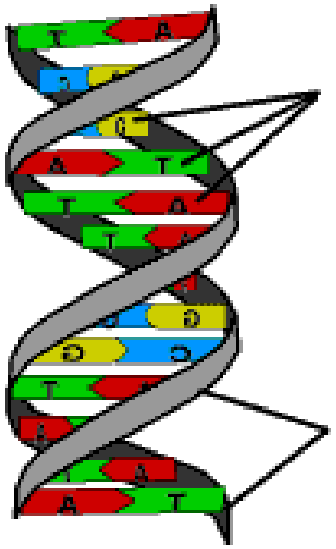
Анализ данных

Осмысление данных

Применение

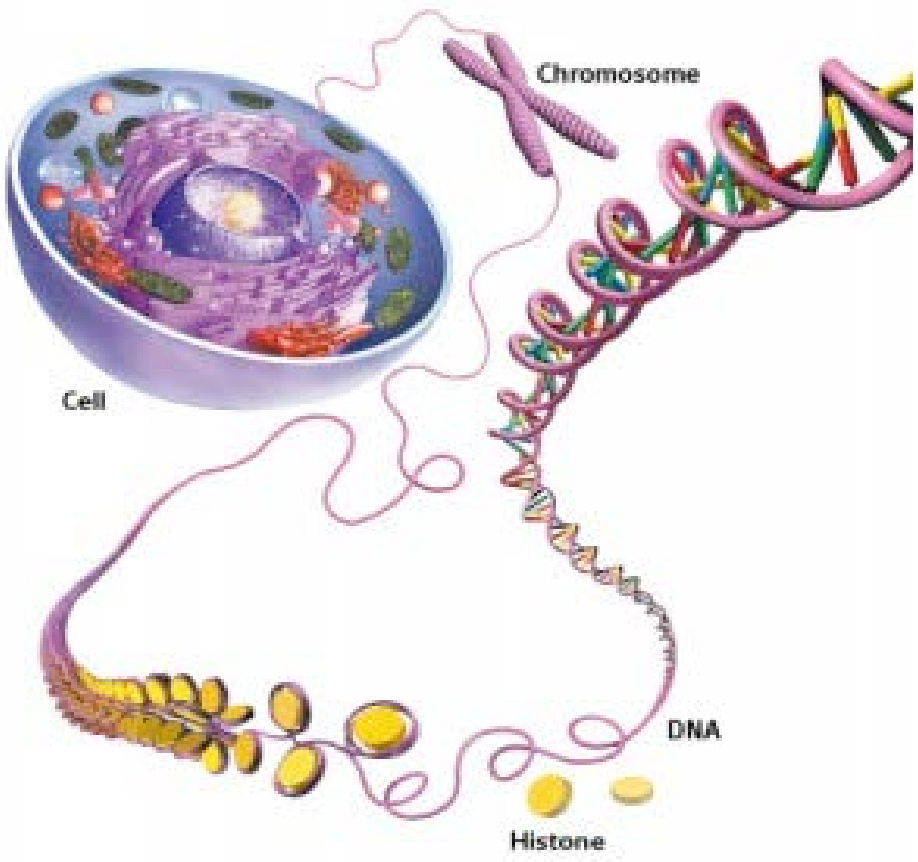
Биолог: ДНК



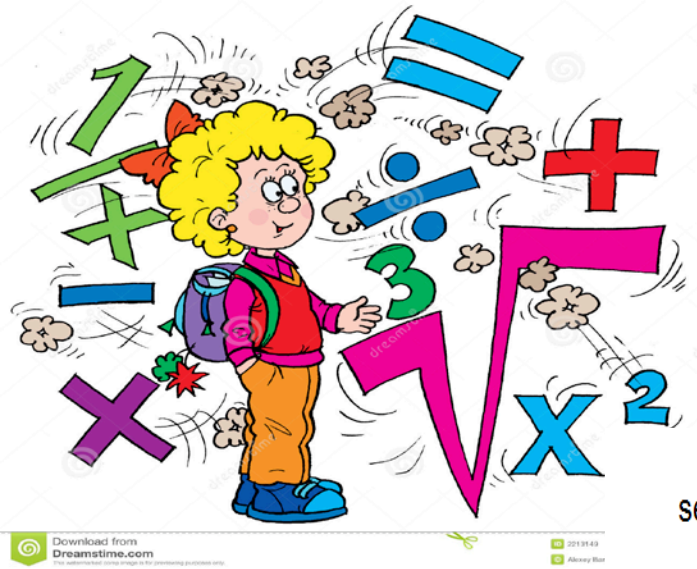


bases - varying part of molecule

side units - do not vary



Математик: ДНК

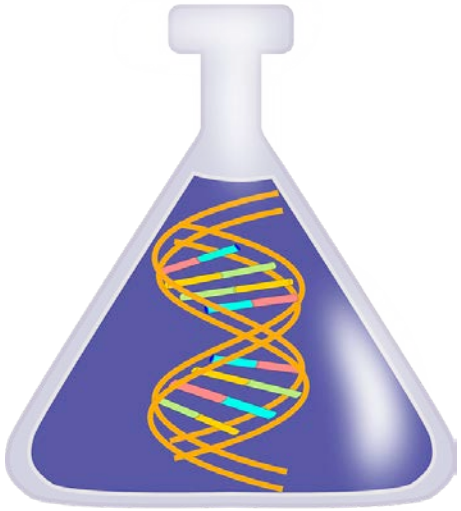


A T G C

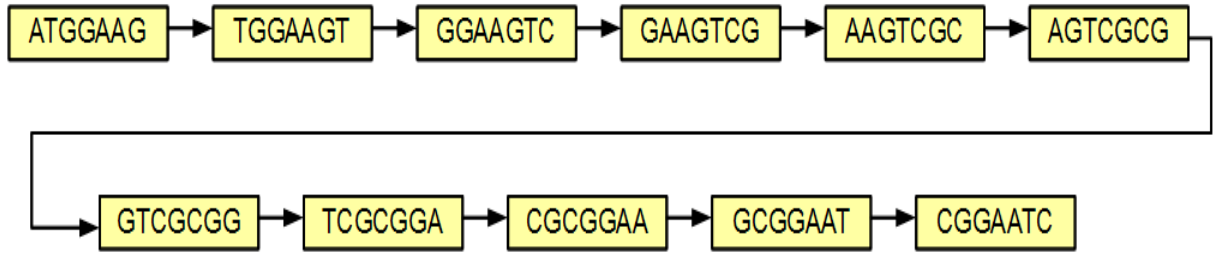
sequence ATGGAAGTCGCGGAATC

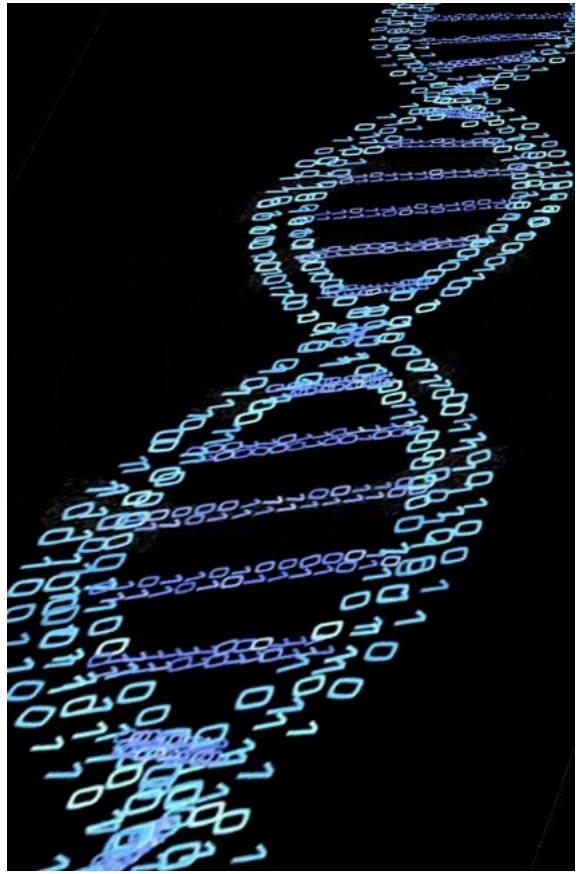
7mers

ATGGAAG
TGGAAGT
GGAAGTC
GAAGTCG
AAGTCGC
AGTCGCG
GTCGCGG
TCGCGGA
CGCGGAA
GCGGAAT
CGGAATC

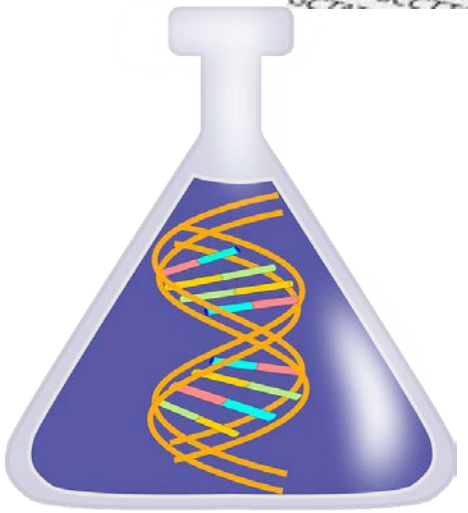


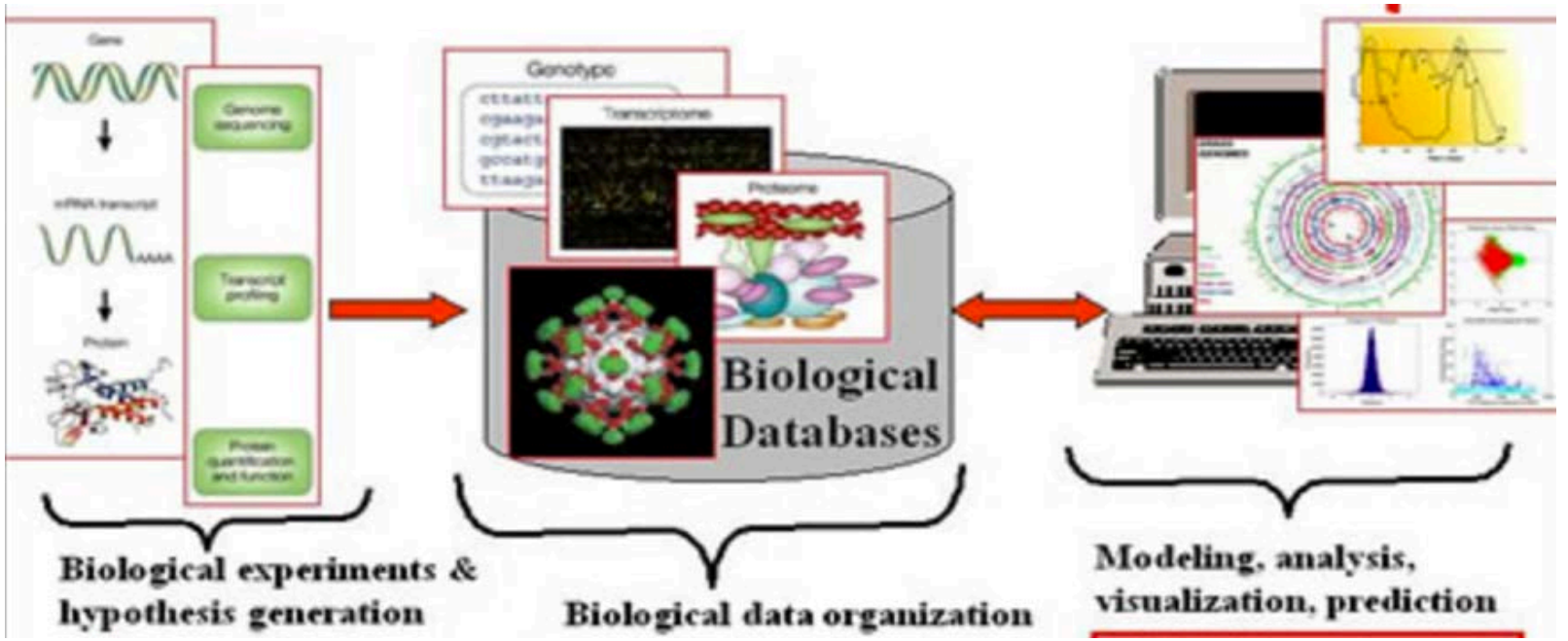
de Bruijn graph





Биоинформатик: ДНК

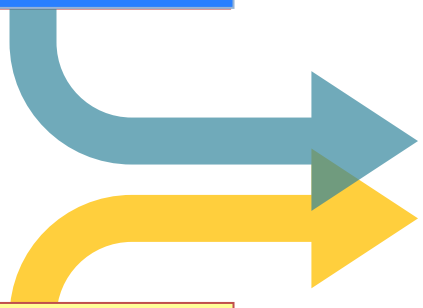




Кто такой биоинформатик?

Biologist
collects molecular data:
DNA, RNA & Protein sequences,
gene expression, etc.

Computer scientist
(+Mathematicians, Statisticians, etc.)
Develops tools, pieces of software and algorithms to store and analyze data.



Bioinformatician
Asks and answers biological questions by analyzing molecular data

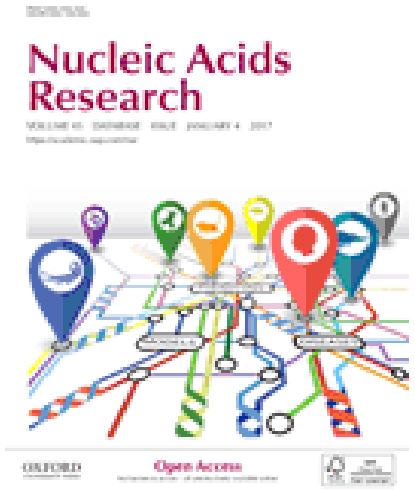
Покопались в данных и ….

1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life

Nature Biotechnology 35, 676–683 (**2017**) doi:10.1038/nbt.3886

Evolutionary drivers of thermoadaptation in enzyme catalysis.
614 protein families with currently unknown structures

Science 20 Jan **2017**: Vol. 355, Issue 6322, pp. 289-294
DOI: 10.1126/science.aah3717



IMG-ABC: new features for bacterial secondary metabolism analysis and targeted biosynthetic gene cluster discovery in thousands of microbial genomes

Nucleic Acids Res (**2017**) 45 (D1): D560-D565. DOI: https://doi.org/10.1093/nar/gkw1103

PLOS | BIOLOGY

Genomic encyclopedia of bacteria and archaea: sequencing a myriad of type strains.

PLoS Biol. **2014** Aug 5;12(8):e1001920. doi: 10.1371/journal.pbio.1001920.

Биоинформатика – это НАУКА

1-й закон биоинформатики

GARBAGE IN

GARBAGE OUT

О чем

1. Центр алгоритмической биотехнологии - немного сухой статистики
2. Что такое биоинформатика?
3. Наука в Центре алгоритмической биотехнологии
4. Образование

«Центр алгоритмической биотехнологии» (ЦАБ)

Основные научные интересы – алгоритмическая биоинформатика

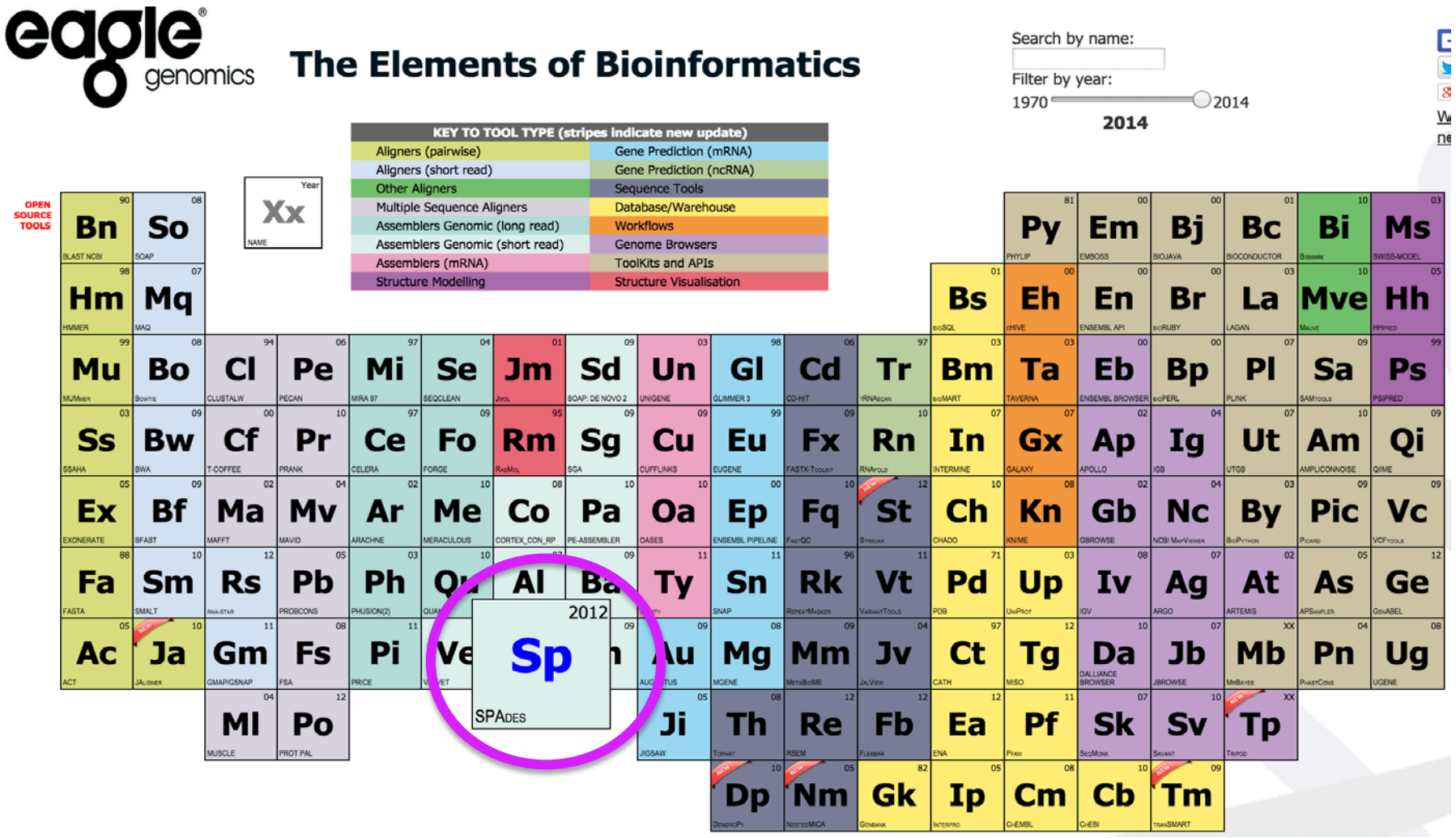
в области:

геномной сборки

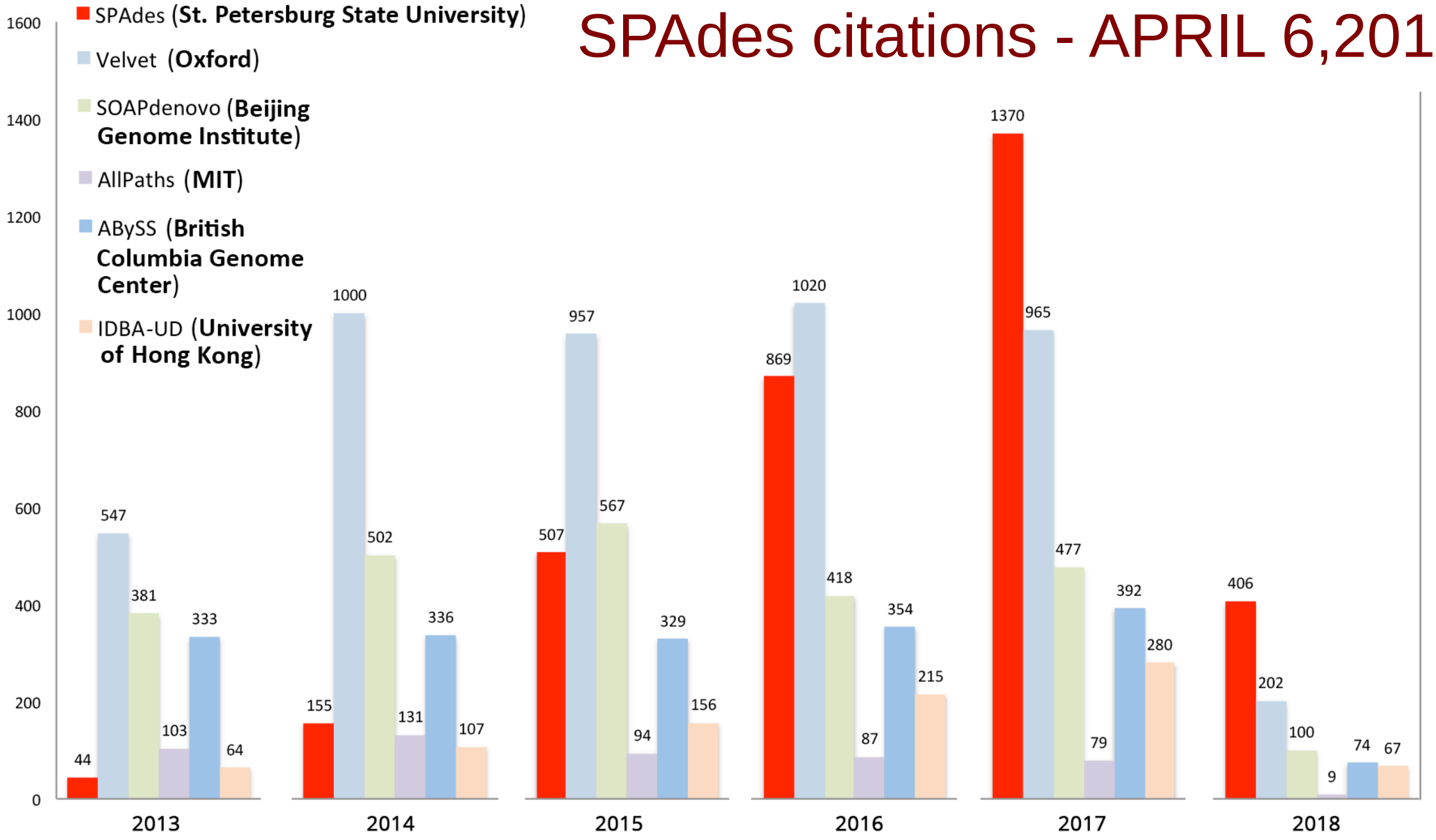
метагеномики, транскриптомики, иммуногеномики,

поиска новых антибиотиков

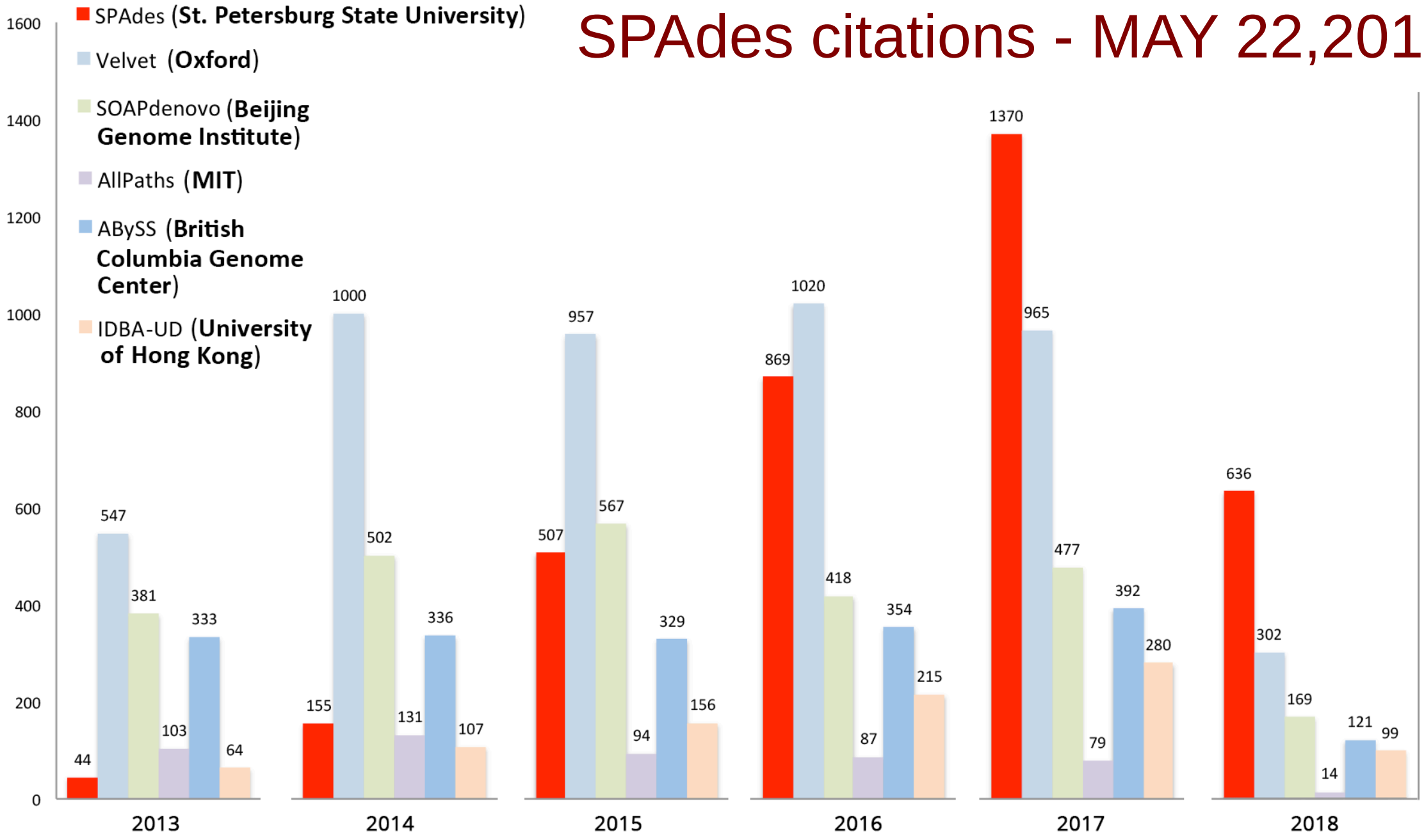
Визитная карточка лаборатории №1



SPAdes citations - APRIL 6,2018



SPAdes citations - MAY 22,2018



2. DNA fragmentation

1. DNA isolation

3. Library preparation

4. Sequencing run

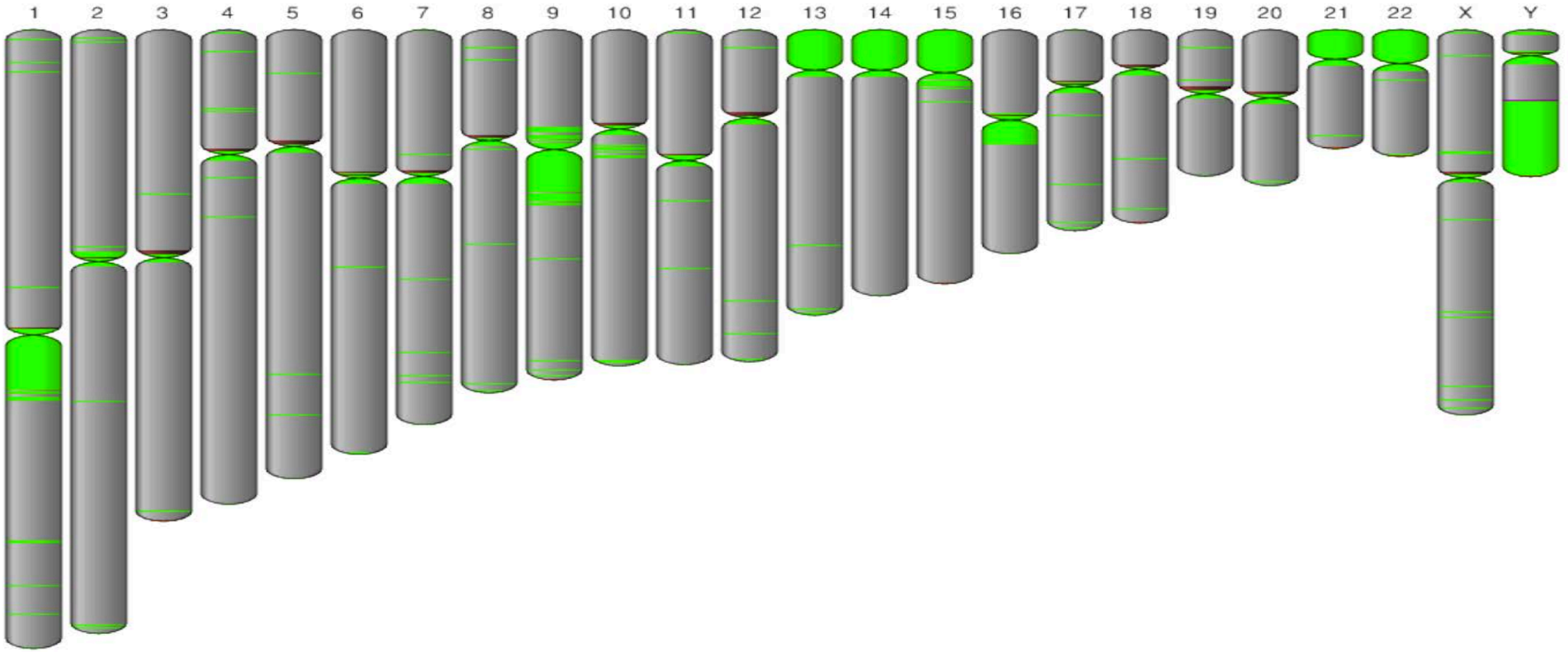
6. Genome reconstruction (assembly)

5. Sequencing reads

Why genome assembly and why it is hard…

Why genome assembly and why it is hard…

Human genome



Human genome Project - not the genome! - is finished

http://genomereference.org

Визитная карточка лаборатории №2

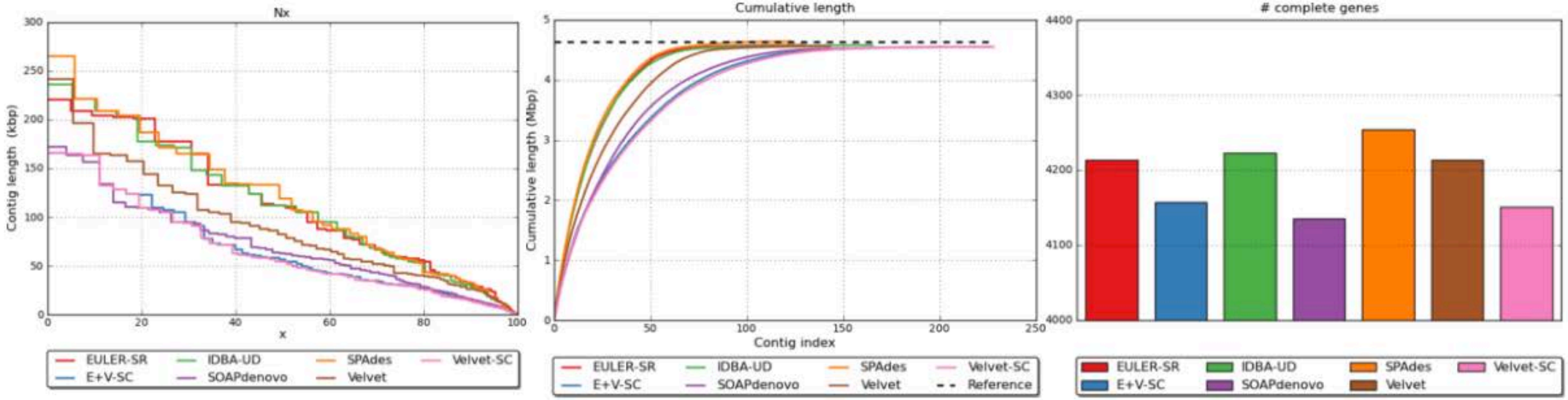
QUAST - Genome assembly evaluation tool

S.aureus dataset, reference-based evaluation, contig alignment viewer:

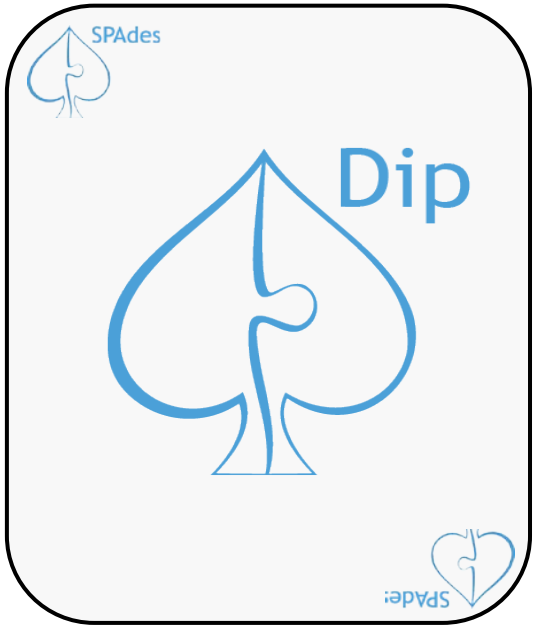
Визитная карточка лаборатории №2

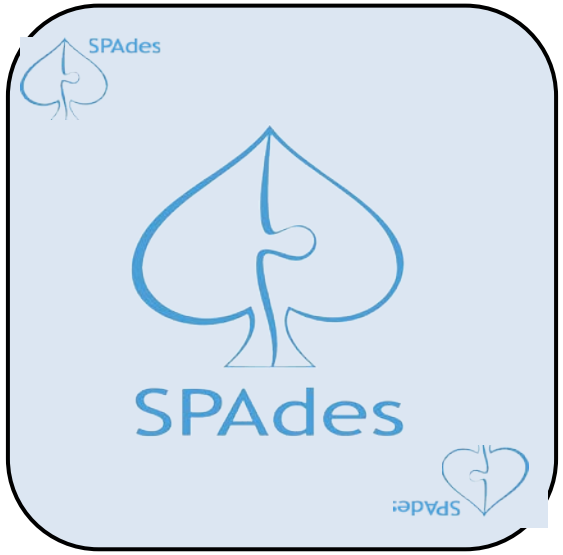
QUAST - Genome assembly evaluation tool

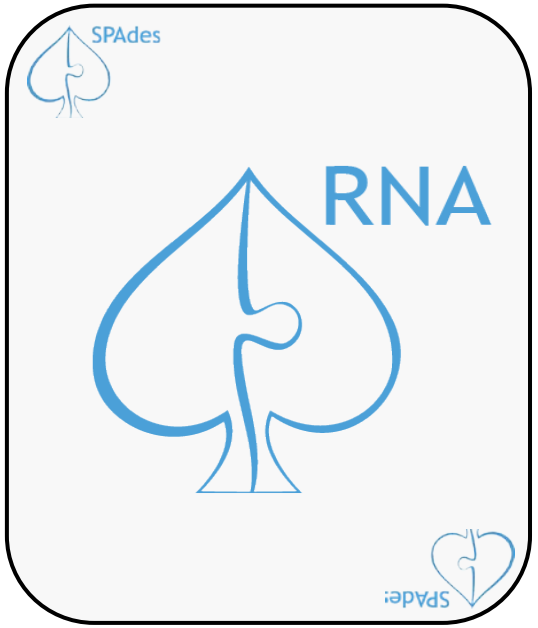
Examples of QUAST output

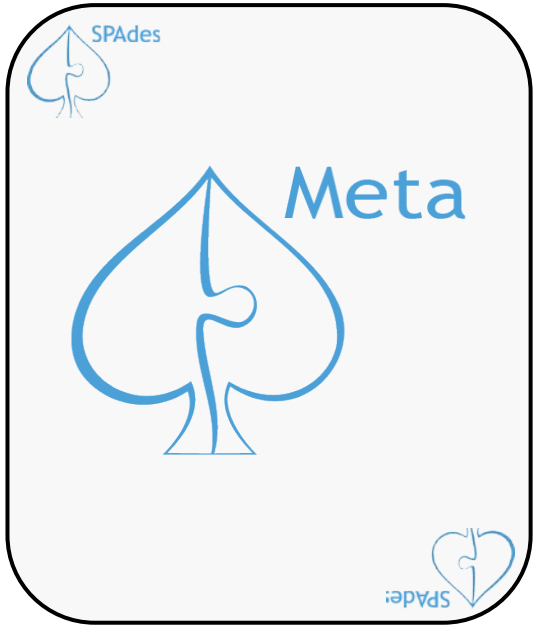


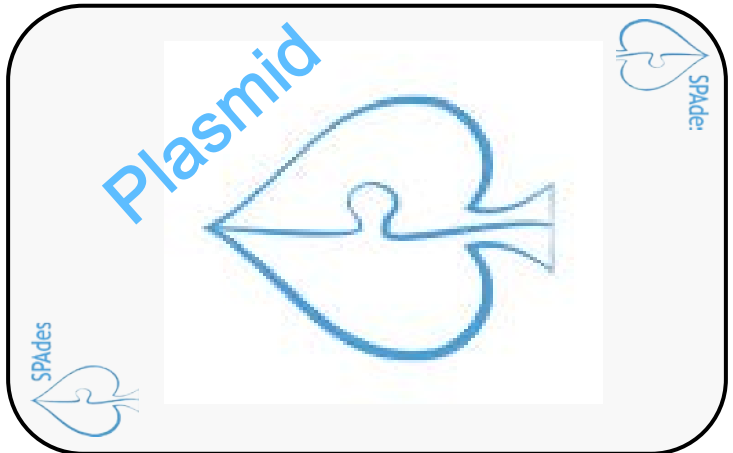
SPAdes Toolbox

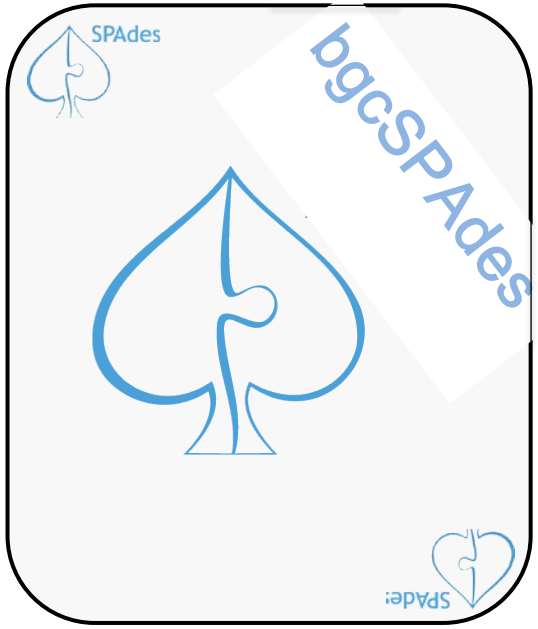












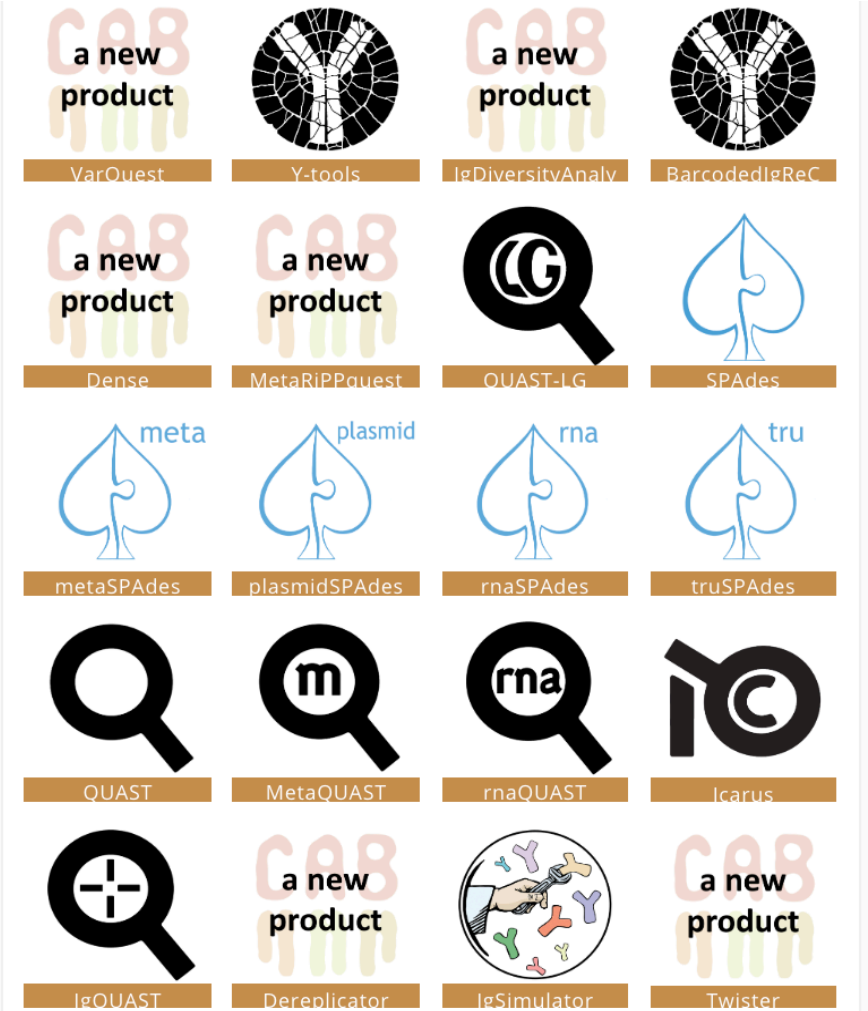
Etc…

Center for Algorithmic Biotechnology
St. Petersburg

HOME ABOUT US RESEARCH SOFTWARE MEMBERS
CONTACTS

Antibiotics Discovery | Genome Assembly | Transcriptomics | TSLR Analysis | Metagenomics | Immunoinformatics

Proteomics



Antibiotics discovery

Immunoinformatics

Education

http://cab.spbu.ru/our-software/

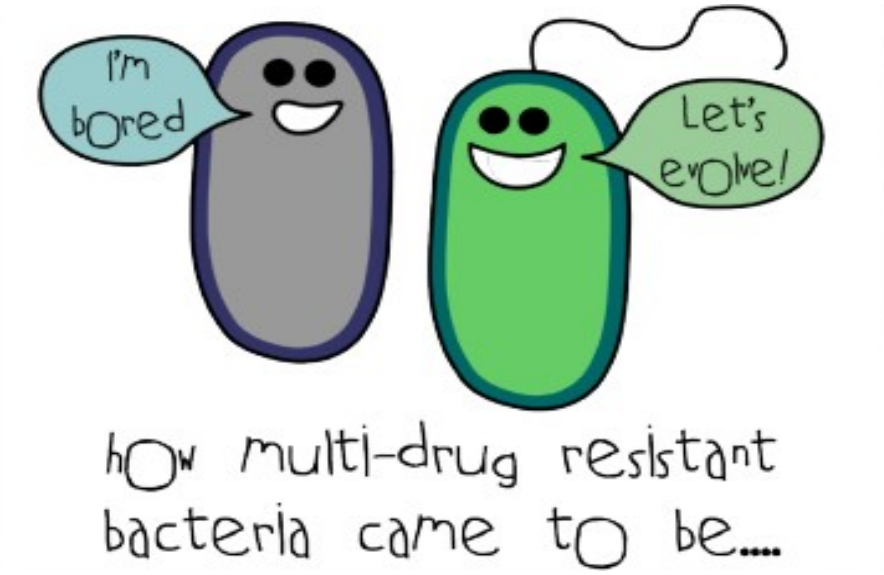
Antibiotics discovery

- Antibiotics — natural substances that kill bacteria or suppress their reproduction
- People use antibiotics more than they should (even for viruses)
- Some bacteria may survive

Antibiotics discovery

They evolve and become resistant

- Drug resistance is a major concern worldwide!
- People need new-generation drugs



Antibiotics discovery challenge

How to discover potential antibiotic among millions of proteins?

Metabolo-genomics

the emerging approach that combines **genomic** and **proteomic** data to identify biologically active organic compounds that could be used as a natural source for novel antibiotics and other drugs

Metabolo-genomics

- DNA sequencing technologies in **genomics**
- high-throughput mass spectrometry-based analysis that includes database search of **mass spectra** against already known metabolites
- **algorithms** and **bioinformatics** tools to expedite research

Metabolo-genomics

- DNA sequencing technologies in **genomics**
- high-throughput mass spectrometry-based analysis that includes database search of **mass spectra** against already known metabolites
- **algorithms** and **bioinformatics** tools to expedite research

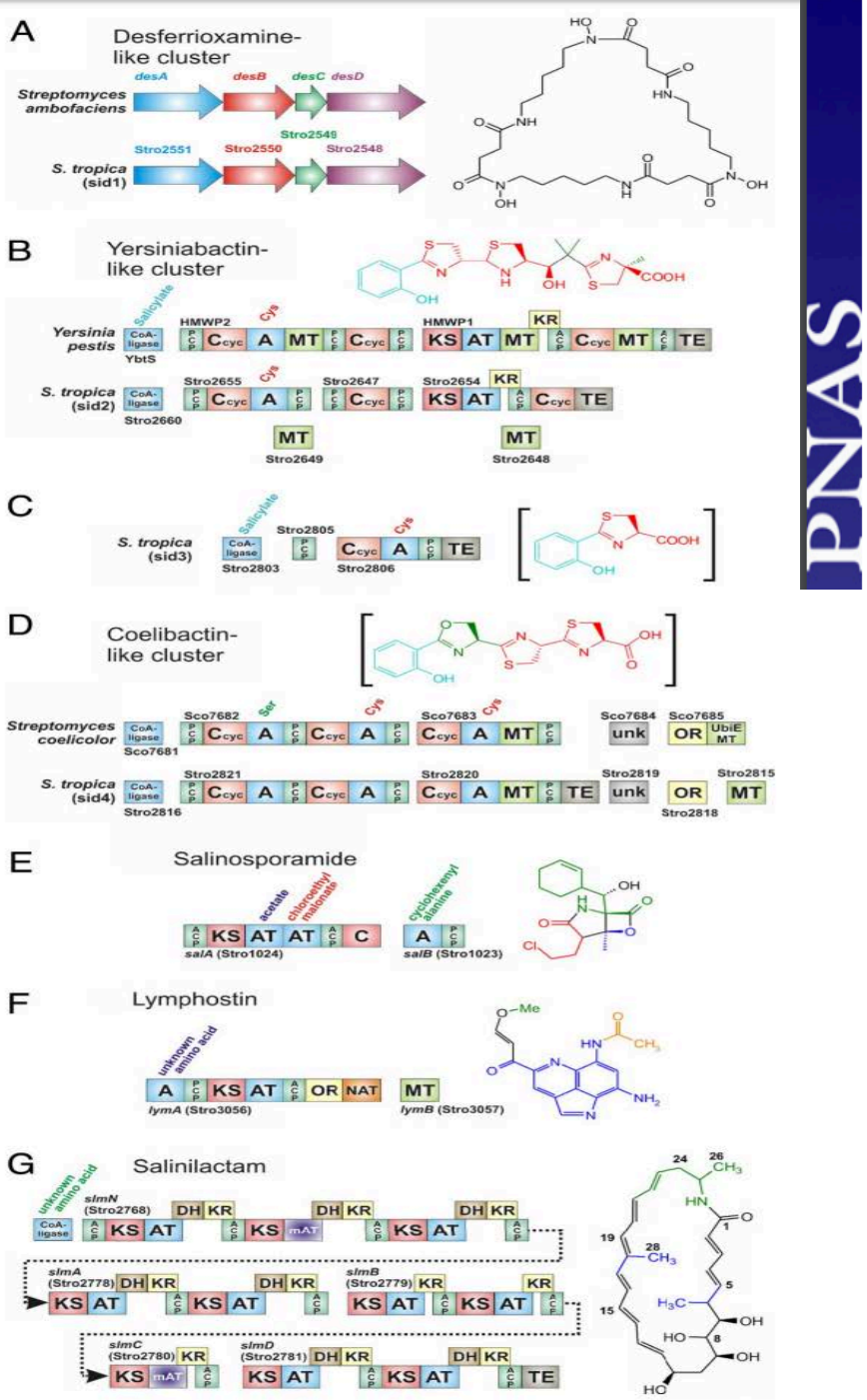


Fig. 2. Selected genes from *S. tropica* modular biosynthetic enzyme systems

2007

Genome sequencing reveals complex secondary metabolome in the marine actinomycete *Salinispora tropica*

Daniel W. Udwary*, Lisa Zeigler*, Ratnakar N. Asolkar*, Vasanth Singan†, Alla Lapidus†, William Fenical*, Paul R. Jensen*, and Bradley S. Moore*‡§

*Scripps Institution of Oceanography and ‡Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California at San Diego, La Jolla, CA 92093-0204; and †Department of Energy, Joint Genome Institute–Lawrence Berkeley National Laboratory, Walnut Creek, CA 94598

Актиномицеты - богатейший источник вторичных метаболитов, на которые приходится более половины всех антибиотиков, обнаруженных на сегодняшний день

Streptomyces coelicolor

“Antibiotics factory”

29 biosynthetic gene clusters, including 3 NRPS (nonribosomal peptide synthetase)

– Coelichelin: 3 A-domains Assembled with SPAdes in 1 contig in minutes

– Nogalamycyn: 2 A-domains Assembled with SPAdes in 1 contig in minutes

– Calcium-dependent antibiotic: 12 A-domains - But these were not ….

Data from Weber et al, Nucleic Acid Research 2015

http://themidult.com/there-is-an-upside-to-feeling-down/

Streptomyces coelicolor

“Antibiotics factory”

29 biosynthetic gene clusters, including 3 NRPS (nonribosomal peptide synthetase)

– Coelichelin: 3 A-domains Assembled with SPAdes in 1 contig in minutes

– Nogalamycyn: 2 A-domains Assembled with SPAdes in 1 contig in minutes

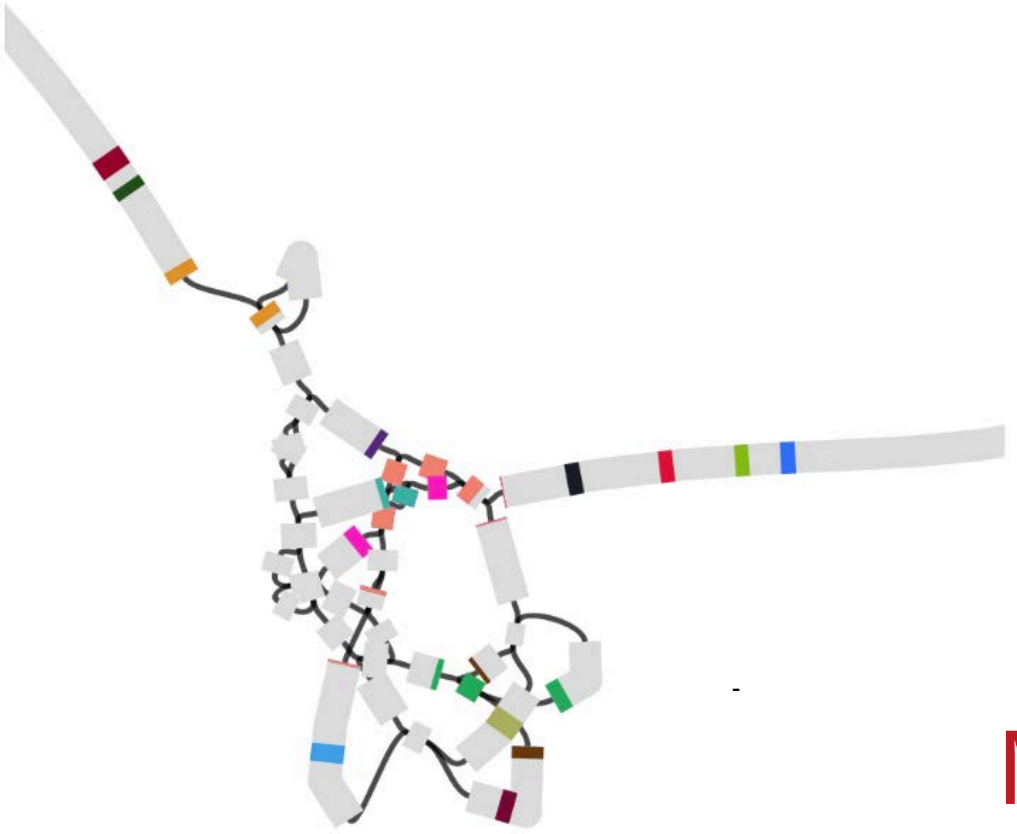
– Calcium-dependent antibiotic: 12 A-domains -> need special tool to deal with complex structures to assemble these domains!

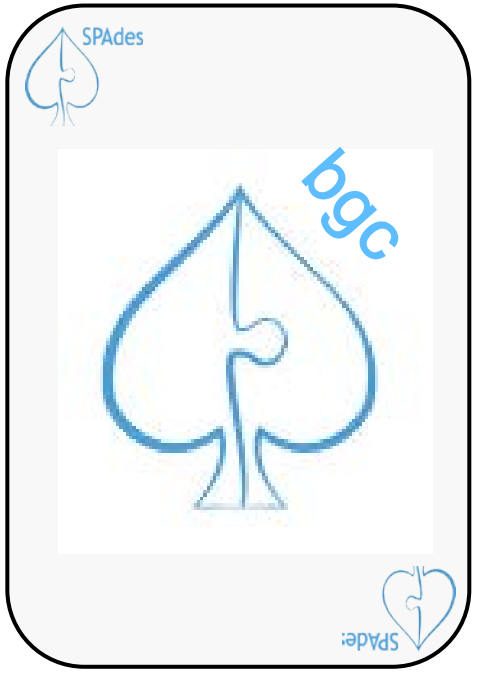
Data from Weber et al, Nucleic Acid Research 2015

New tool to solve difficult repeat structures

bgcSPAdes

converts complex subgraphs like this





Minutes vs 8 months!

Metabolo-genomics

- DNA sequencing technologies in **genomics**
- high-throughput mass spectrometry-based analysis that includes database search of **mass spectra** against already known metabolites
- **algorithms** and **bioinformatics** tools to expedite research

Global Natural Products Social (GNPS) molecular network

GNPS содержит более **миллиарда** масс-спектров биологически активных соединений, собранных во всем мире

=> *необходимы алгоритмы поиска* Натуральных пептидных продуктов (*PNP*) *в этой массе данных*

https://gnps.ucsd.edu/ProteoSAFe/static/gnps-splash.jsp

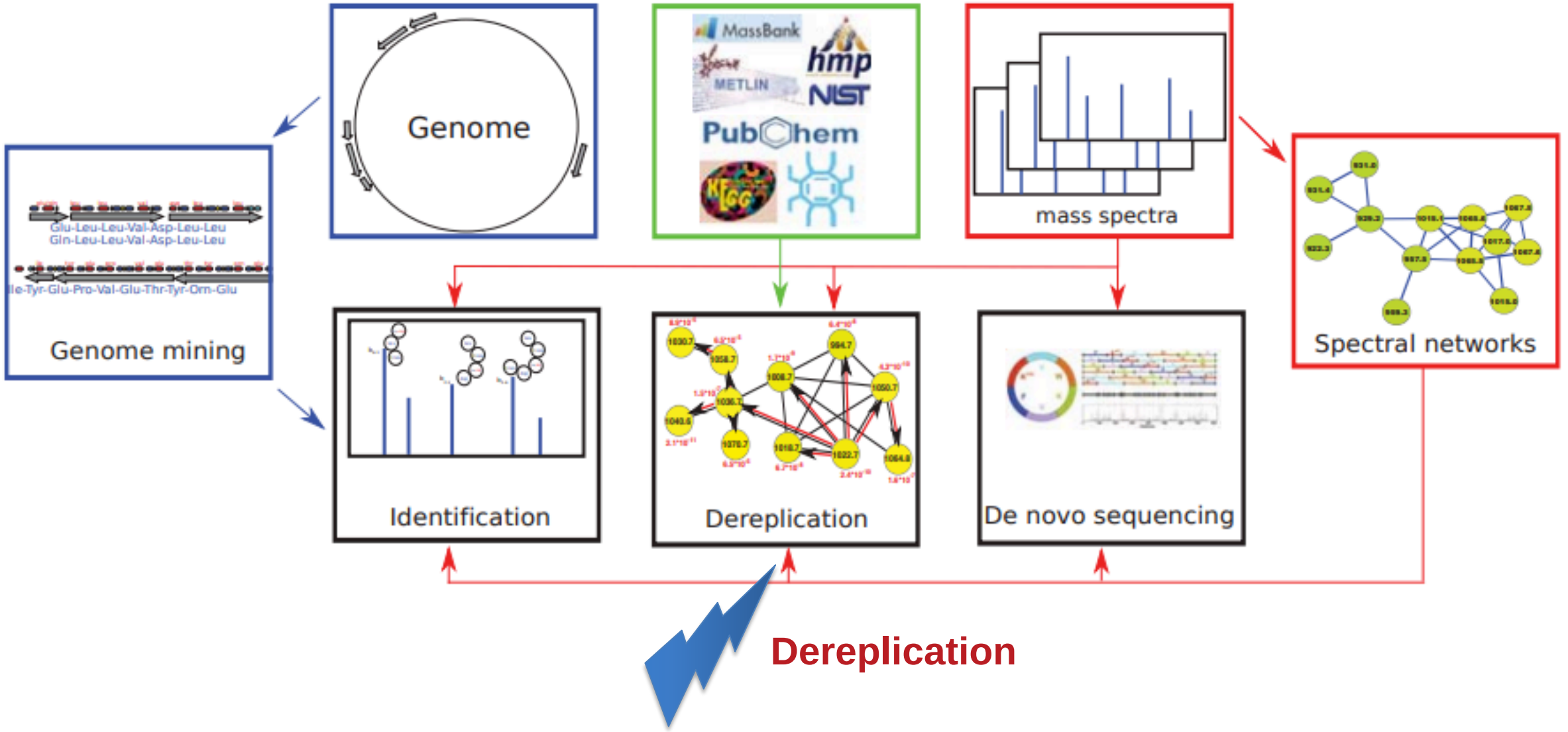
Antibiotics discovery challenge

How to discover potential antibiotic among millions of proteins?

PNP discovery challenge

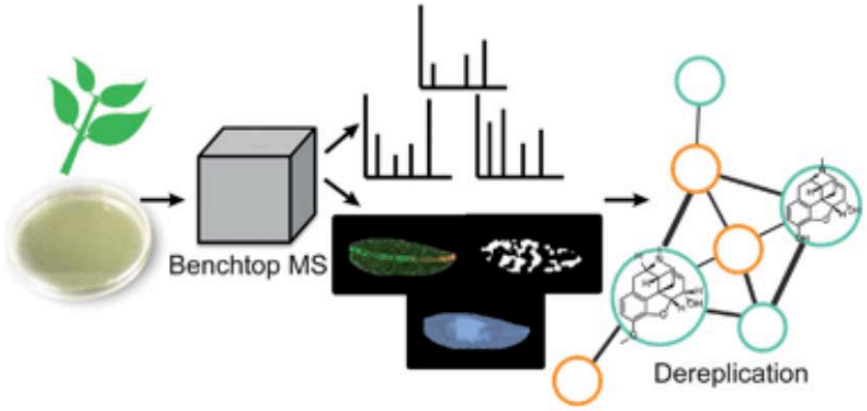
How to discover new PNP among millions of proteins?

Metabolo-genomics



Dereplication

- тестирование образцов смесей для распознавания и исключения из рассмотрения тех активных веществ, которые уже были ранее охарактеризованы



http://pubs.rsc.org/en/journals/articlecollectionlanding?themeid=4d49-6d56-14a6-2b79-c1d6-d55f-221c-93b2

http://cab.spbu.ru/software/dereplicator/

Dereplicator

- a computational tool for identification of known natural products from LC/MS-MS data.

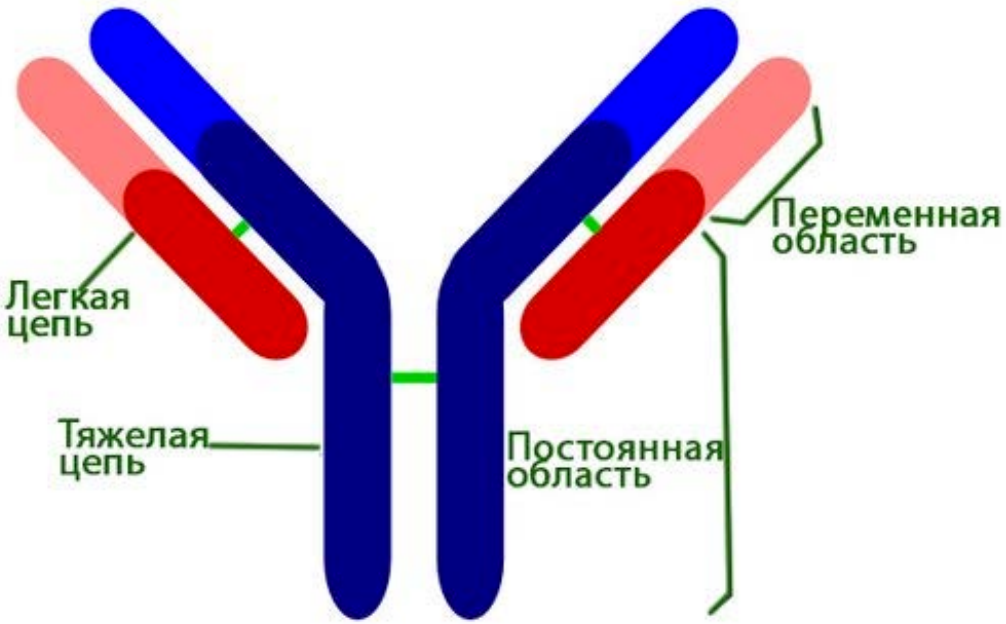
- ✓ использует базу данных **химических** структур
- ✓ генерирует in-silico **масс-спектры** соединений, предсказывая, как они фрагментируются во время масс-спектрометрии
- ✓ объединяет их с **экспериментальными** LC / MS-MS для обнаружения сходства
- ✓ конвертирует оценки подобия в статистическую значимость

Collaborative Development with University of California, San Diego

https://www.nature.com/articles/nchembio.2219

http://cab.spbu.ru/software/dereplicator/

Immunoinformatics at CAB



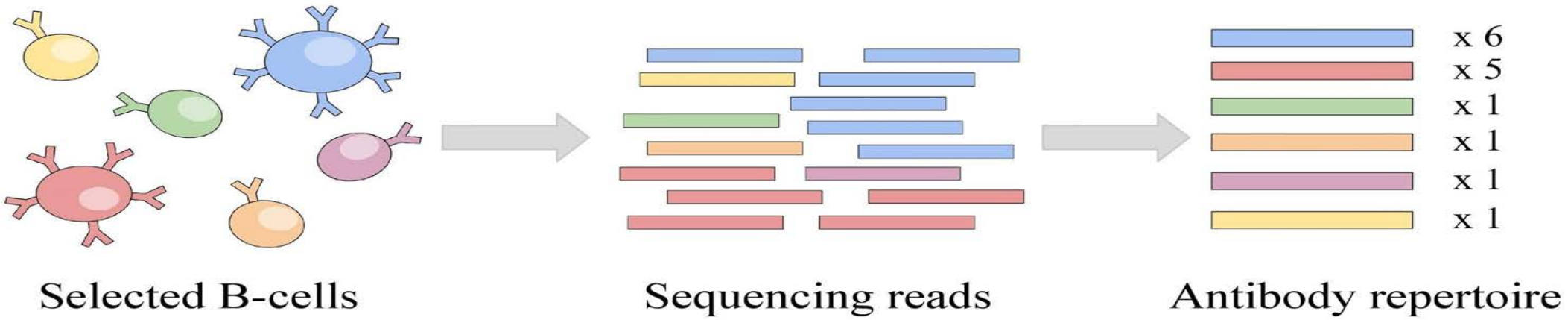
Антитело (англ. antibody) — белок (иммуноглобулин), синтезируемый В-лимфоцитами в организме человека и животного в ответ на попадание в него чужеродного вещества и обладающий специфическим сродством к этому веществу.

Антитела продуцируются иммунной системой для нейтрализации антигенов (чужеродных молекул)

http://cab.spbu.ru/research/immunoinformatics/

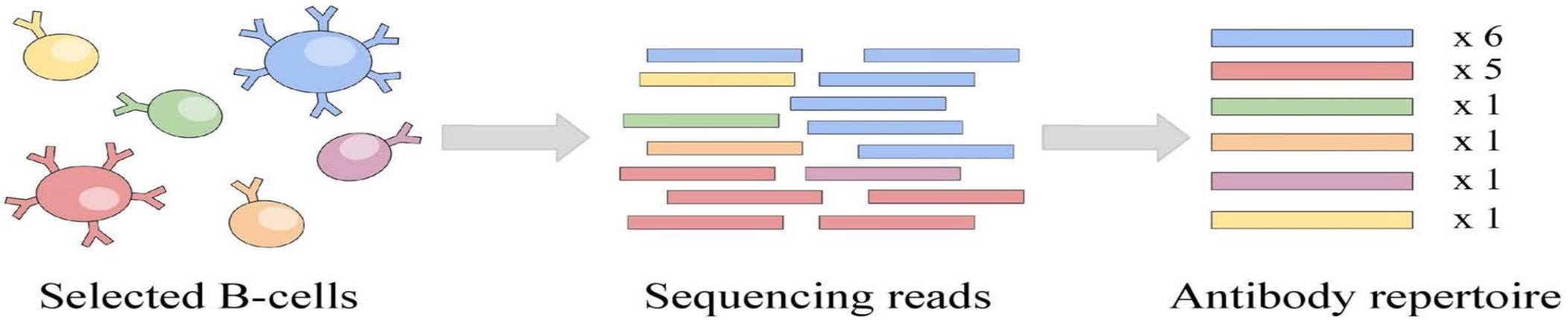
Репертуар антител

Репертуар антител — суммарное количество типов антител, характеризующих функциональное состояние иммунной системы индивидуума в данный момент



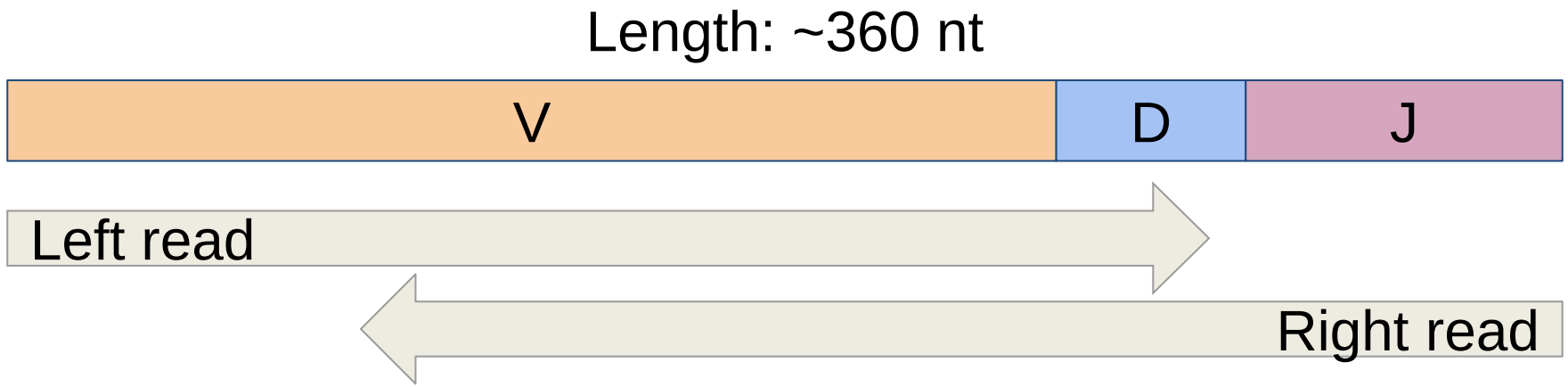
Репертуар антител

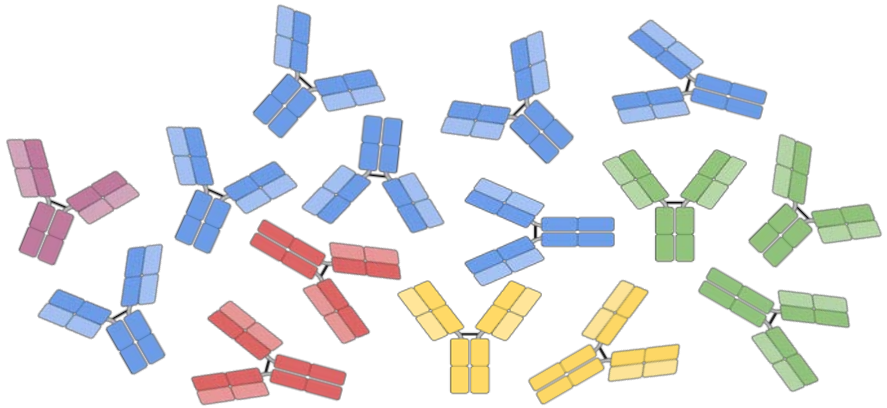
Построение репертуара антител из данных сиквенса ДНК (NGS) и его анализ являются важными шагами в разработке лекарственных препаратов на основе антител и клинических исследований



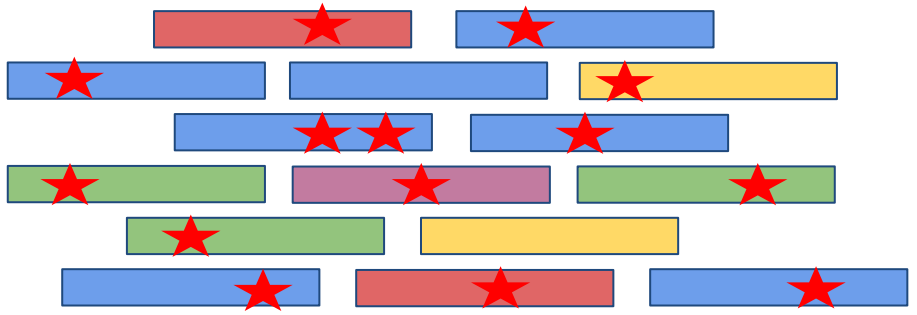
http://cab.spbu.ru/research/immunoinformatics/

Antibody repertoire sequencing (Rep-seq)





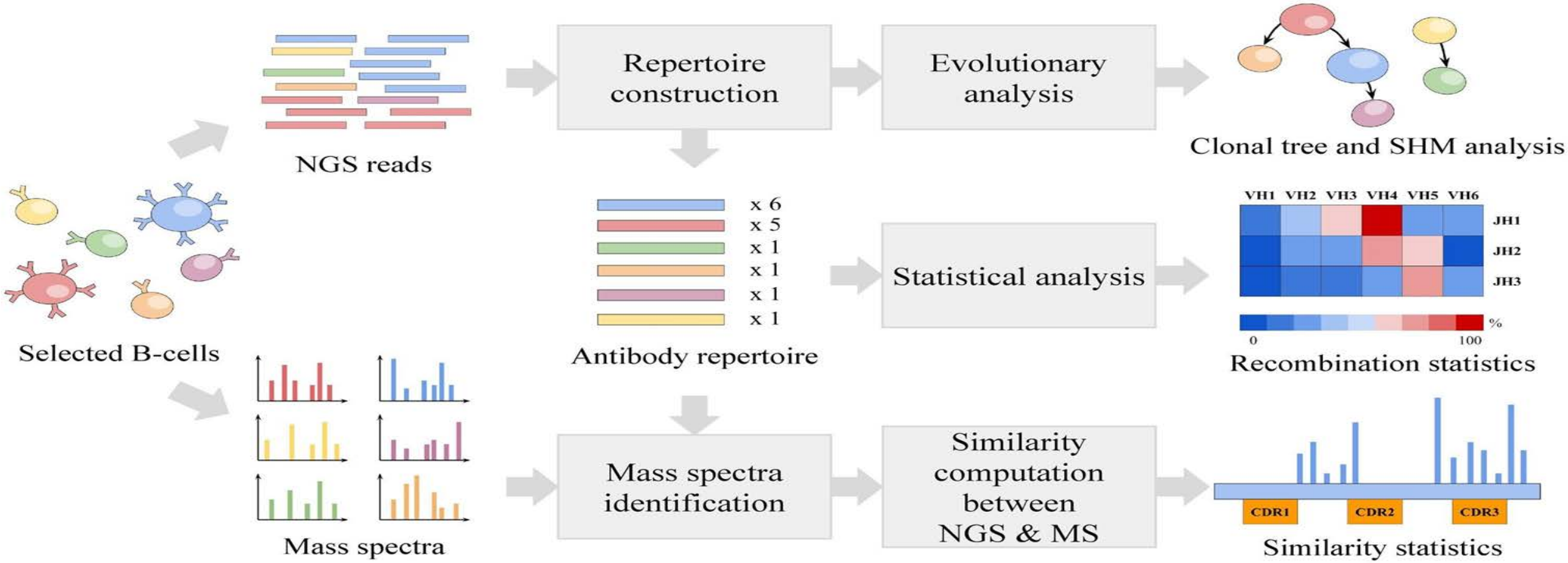
VDJ from DNA or RNA



Error-prone immunosequencing reads

Turchaninova et al, *Nat Protocols*, 2016

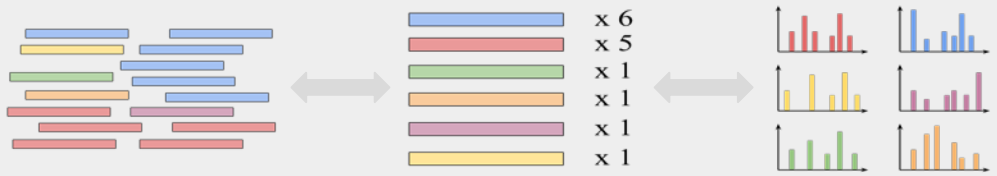
Immunoinformatics analysis



Immunosequencing data opens new bioinformatics challenges

Immunoinformatics at CAB

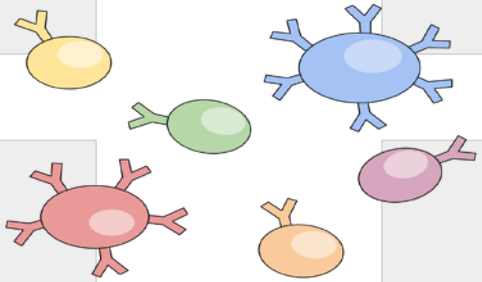
IgRepertoireConstructor



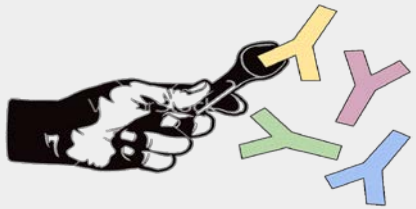
tool for construction of antibody repertoire using mass spectra

IgAnalyzer & IgQUAST

tools for “quality” assessment of antibody repertoire

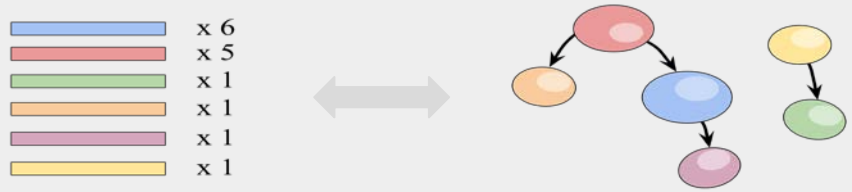


IgSimulator



versatile immunosequencing simulator

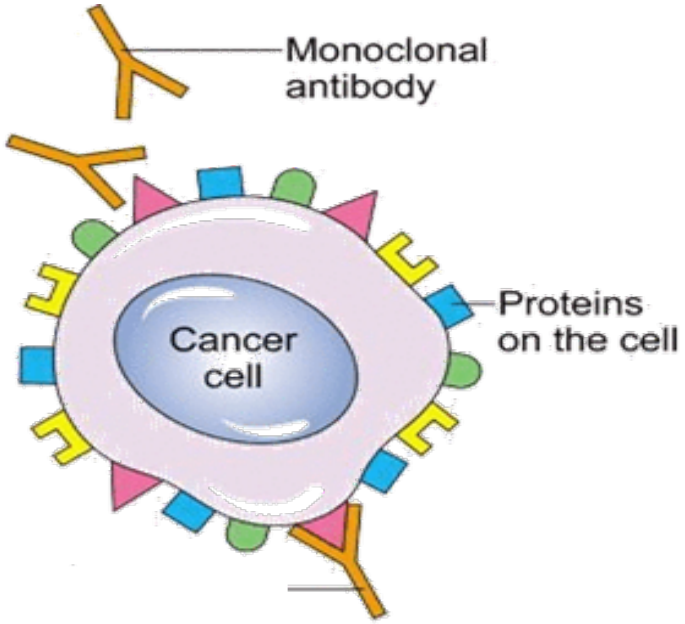
AntEvolo



tool for construction of clonal trees and analysis of somatic hypermutation (SHM)

http://cab.spbu.ru/research/immunoinformatics/

Биомедицинские применения



Cancer treatment based on monoclonal antibody therapy

Studying of autoimmune disorders and immunodeficiency

Immunization analysis and monitoring of treatment

Trends in Immunology

Review

Bioinformatic and Statistical Analysis of Adaptive Immune Repertoires

Victor Greiff,[1] Enkelejda Miho,[1] Ulrike Menzel,[1] and Sai T. Reddy[1,*]

Yaari and Kleinstein *Genome Medicine* (2015) 7:121
DOI 10.1186/s13073-015-0243-2

Genome Medicine

REVIEW Open Access

Practical guidelines for B-cell receptor repertoire sequencing analysis

Gur Yaari[1*] and Steven H. Kleinstein[2,3*]

Sequencing and bioinformatics analysis of adaptive immune repertoire are becoming routine in clinical studies

HHS Public Access
Author manuscript
Nat Rev Rheumatol. Author manuscript; available in PMC 2015 April 01.

Published in final edited form as:
Nat Rev Rheumatol. 2015 March ; 11(3): 171–182. doi:10.1038/nrrheum.2014.220.

Sequencing the functional antibody repertoire—diagnostic and therapeutic discovery

William H. Robinson
Division of Immunology and Rheumatology, CCSR 4135, 269 Campus Drive, Stanford, CA 94305, USA. wrobins@stanford.edu

REVIEW

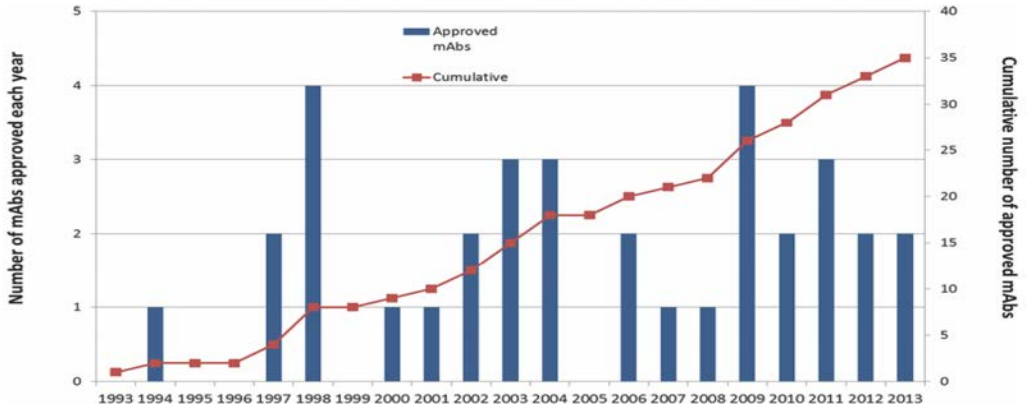
nature biotechnology

The promise and challenge of high-throughput sequencing of the antibody repertoire

George Georgiou[1–4], Gregory C Ippolito[3,4], John Beausang[5,6], Christian E Busse[7], Hedda Wardemann[7] & Stephen R Quake[5,6,8,9]

Antibody drugs

In 2017, global sale revenue for antibody drugs was $90 billion

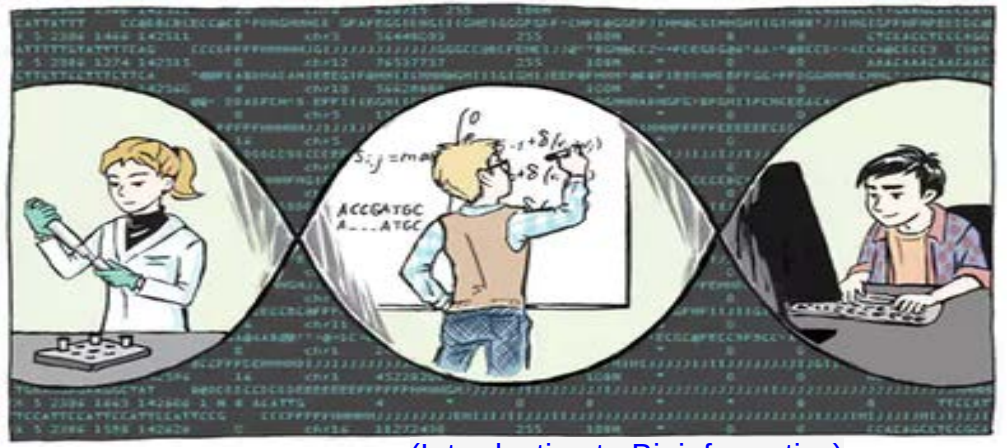


О чем

1. Центр алгоритмической биотехнологии - немного сухой статистики
2. Что такое биоинформатика?
3. Наука в Центре алгоритмической биотехнологии
4. Образование

Bioinformatics education: **online**

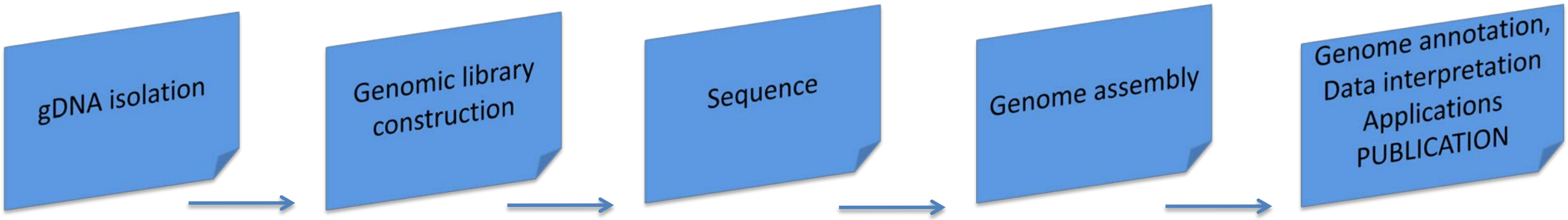
- **Introduction to bioinformatics for biologists**
  - Coursera (2014)
- **Introduction to bioinformatics: Metagenomics** (2016)
  - Coursera, Opened Education Platform, Stepik
- **Introduction into Linux for biologists** (2015)
  - Stepik



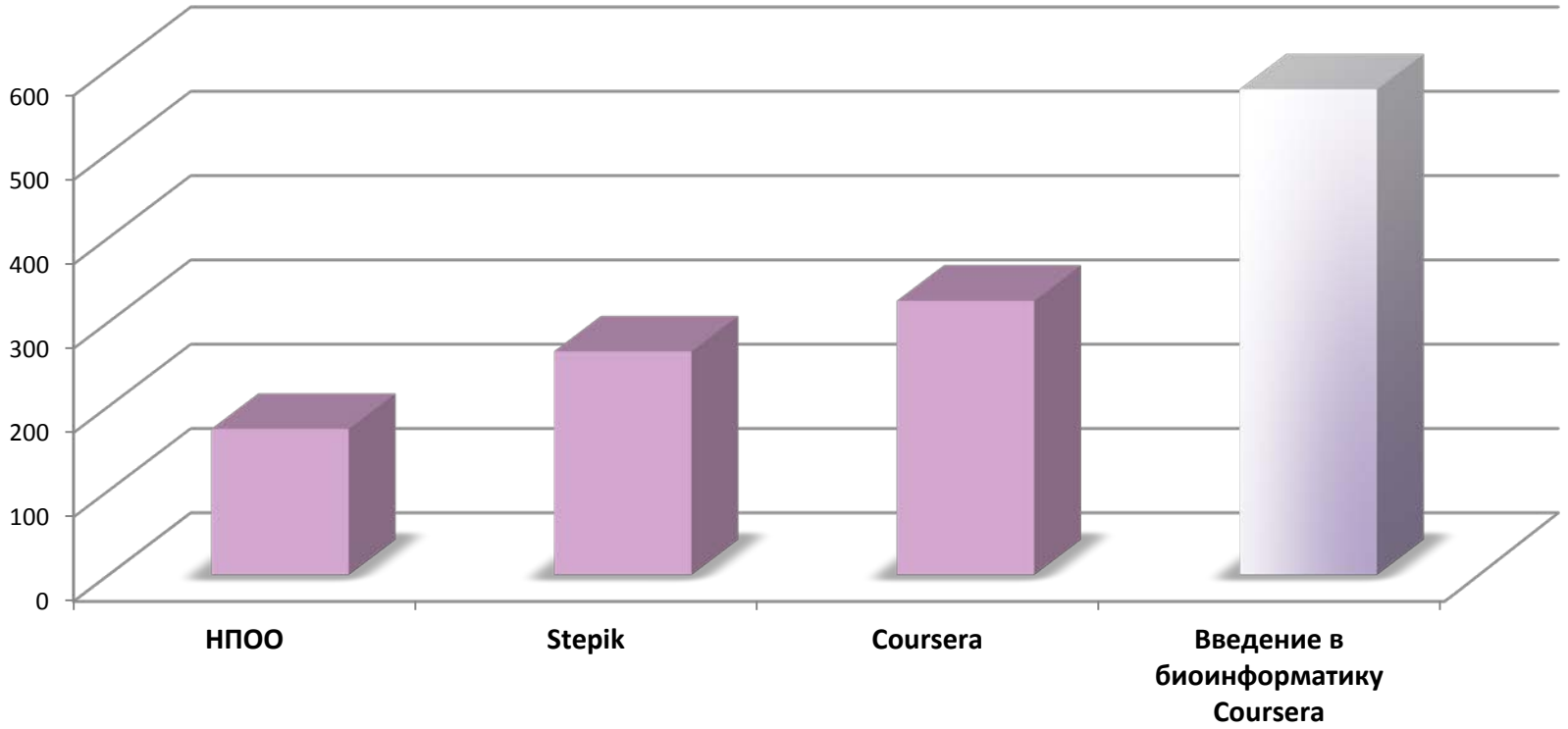
(Introduction to Bioinformatics)

How they differ from other bioinformatics courses on Coursera (and not only Coursera)

- На русском языке
- New - “along the project” – wet lab + algorithms
- Practical



“Введение в биоинформатку”
&
“Введение в биоинформатку: МЕТАГЕНОМИКА”

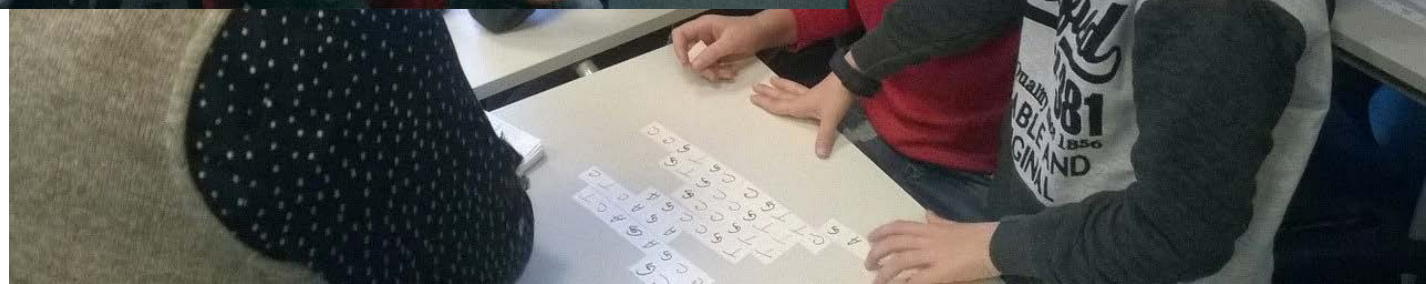


TOTAL: 1339 new students/month

Bioinformatics education: **offline**

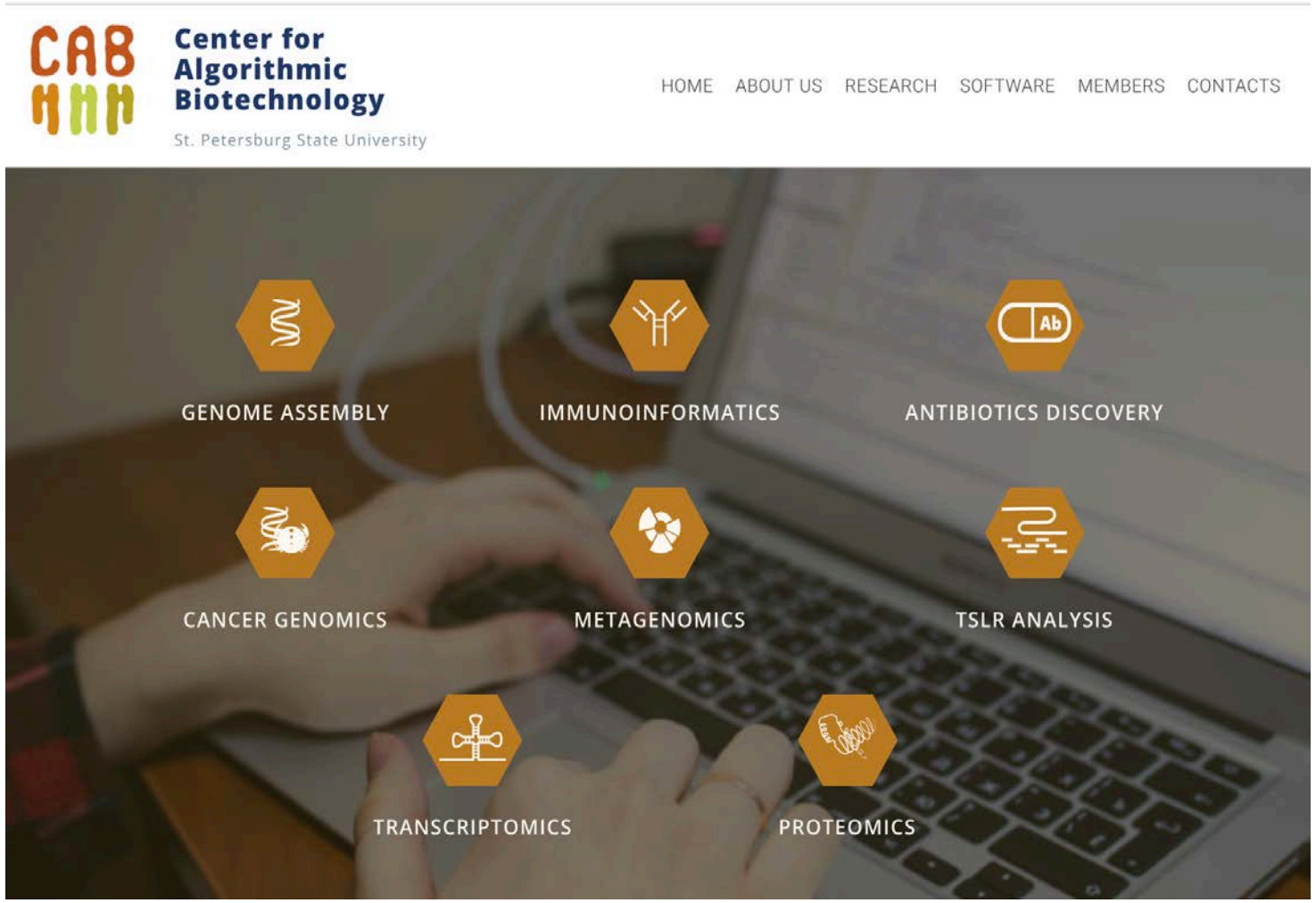
- Bioinformatics for biologists
  - Introduction to bioinformatics (1 semester main course, MS)
  - Two year MS program “Bioinformatics” (**will start September 2018**)
- Algorithms, statistics and sequencing data analysis for computer scientists
- Genomics workshops (Russia, Finland, Italy)
- Summer schools for undergraduate students
- Summer internships at CAB

Bioinformatics education: **offline**



http://biata2018.spbu.ru/

Acknowledgments



http://cab.spbu.ru/

RSF | Russian Science Foundation

(grant #14-50-00069)