An Overview of Bioinformatics

BERDC Special Topics Talk 17

DaCCoTA
DAKOTA COMMUNITY COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Dr. Mark Williamson

Biostatistics, Epidemiology, and Research Design Core

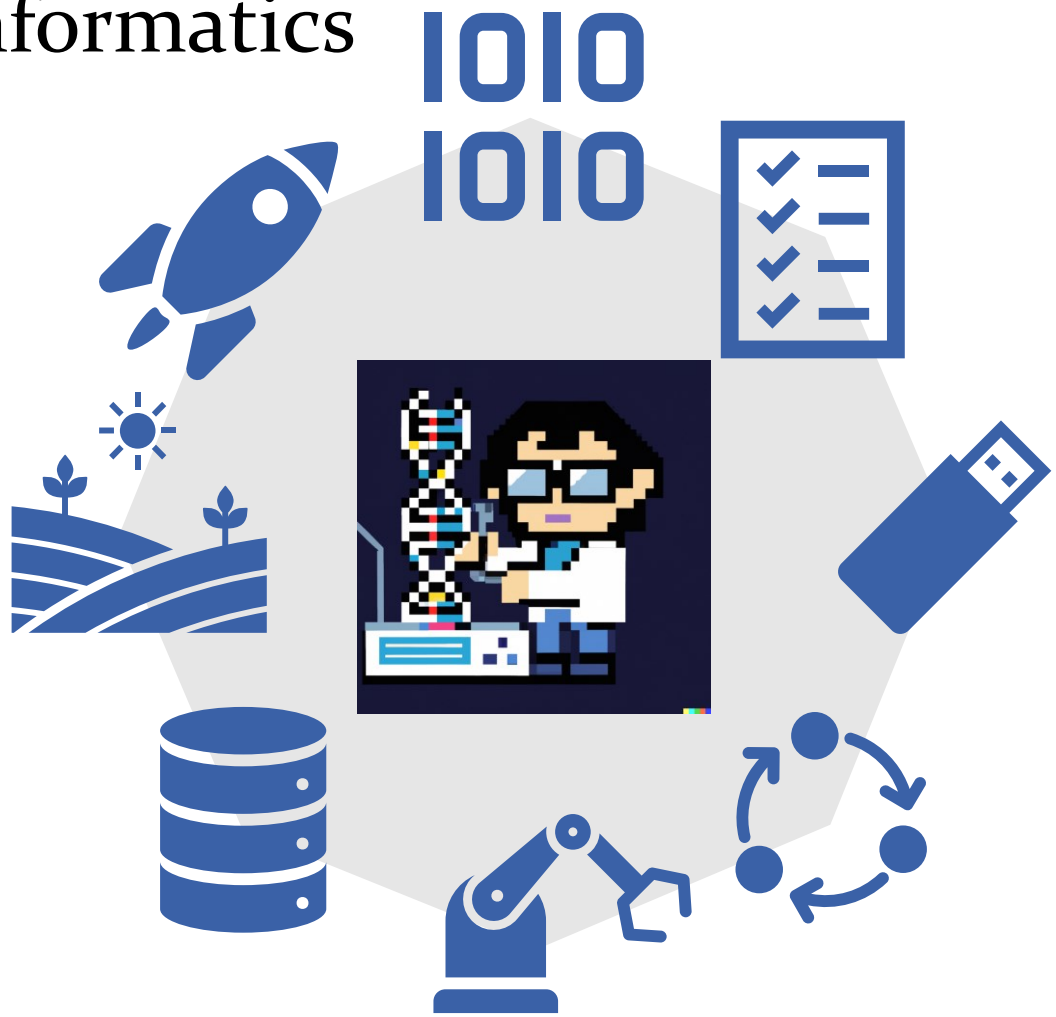
DaCCoTA
DAKOTA COMMUNITY COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Introduction

UND UNIVERSITY OF NORTH DAKOTA

Goal: give an overview of Bioinformatics

- Definition & History
- Field & Topics
- Data
- Methods
- Machines & Software
- Databases
- Example Applications
- The Future



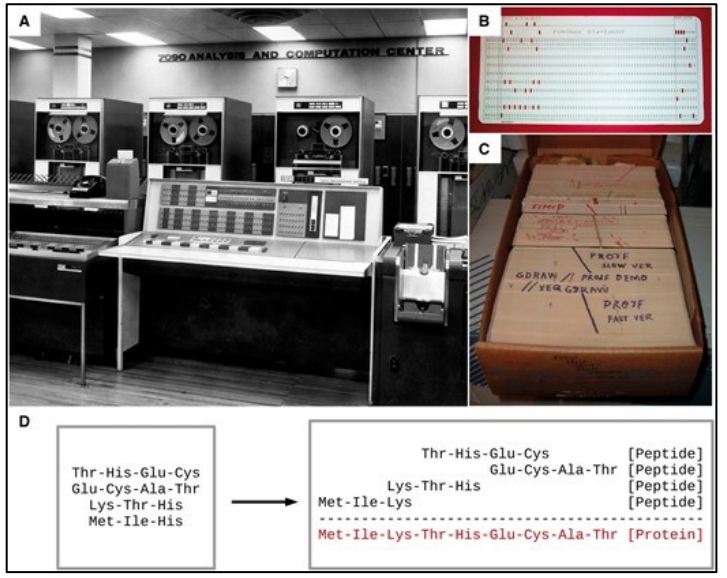
DaCCoTA
DAKOTA COMMUNITY COLLABORATIVE ON TRANSLATIONAL ACTIVITY

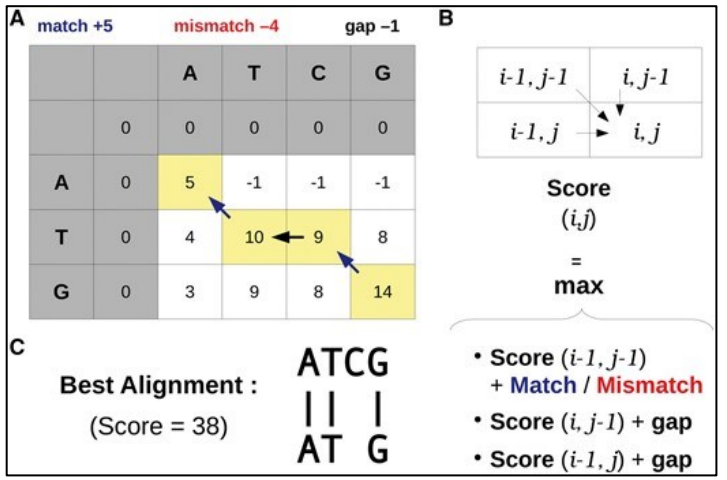
Definition & History

UND UNIVERSITY OF NORTH DAKOTA

"the collection, classification, storage, and analysis of biochemical and biological information using computers especially as applied to molecular genetics and genomics" –Merriam-Webster

- 1960-1970: Protein analysis origins and Margaret Dayhoff [1]
- 1970-1980: Shift from protein to DNA analysis
- 1980-1990: Advances in biology and computing
- 1990-2000: Genomics and the internet
- 2000-2010: High-throughput bioinformatics (2nd gen sequencing)
- 2010-Today: Big data, role of bioinformaticians, systems biology





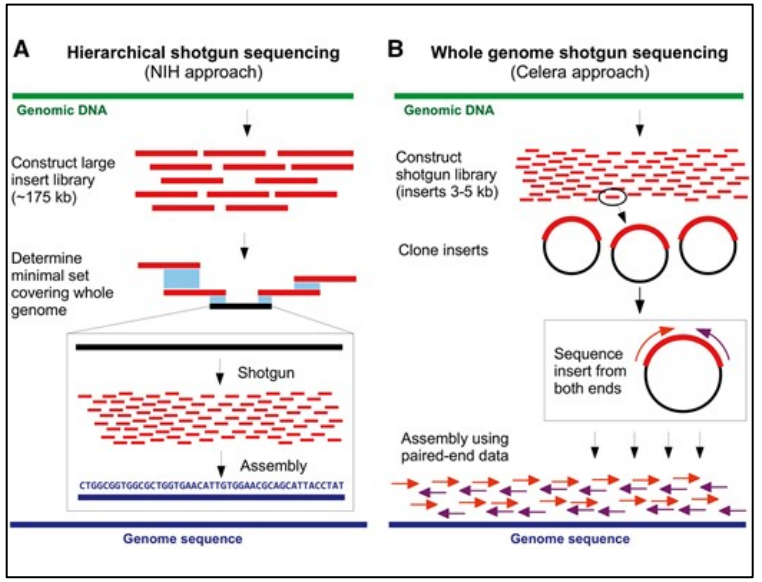
DaCCoTA
DAKOTA COMMUNITY COLLABORATIVE ON TRANSLATIONAL ACTIVITY

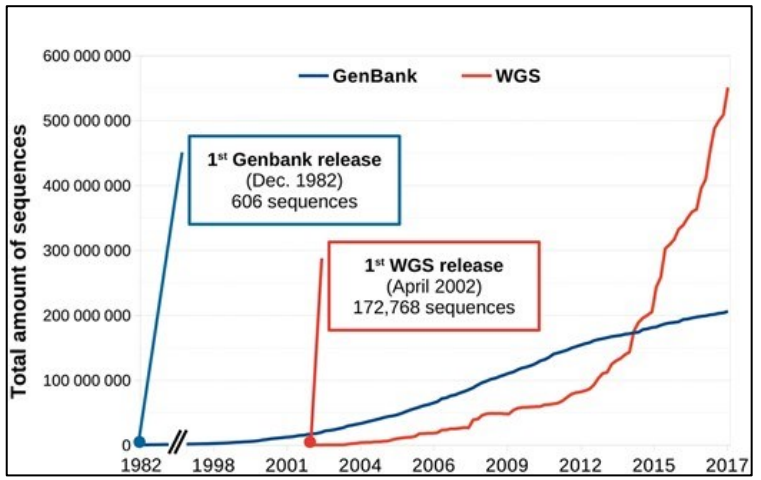
Definition & History

UNIVERSITY OF NORTH DAKOTA

"the collection, classification, storage, and analysis of biochemical and biological information using computers especially as applied to molecular genetics and genomics" –Merriam-Webster

- 1960-1970: Protein analysis origins and Margaret Dayhoff [1]
- 1970-1980: Shift from protein to DNA analysis
- 1980-1990: Advances in biology and computing
- 1990-2000: Genomics and the internet
- 2000-2010: High-throughput bioinformatics (2nd gen sequencing)
- 2010-Today: Big data, role of bioinformaticians, systems biology

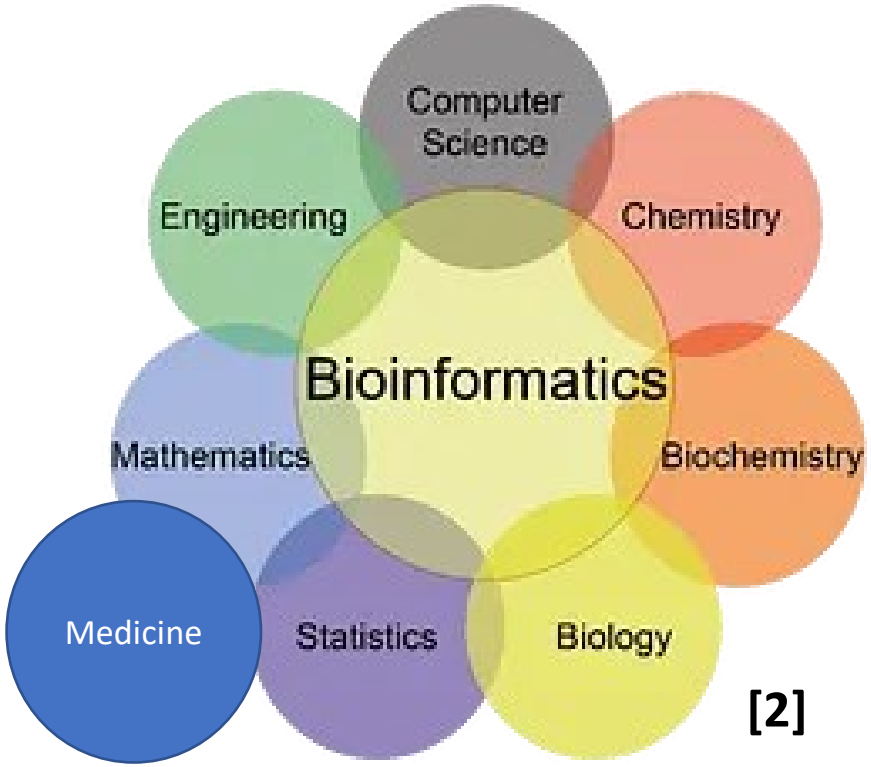




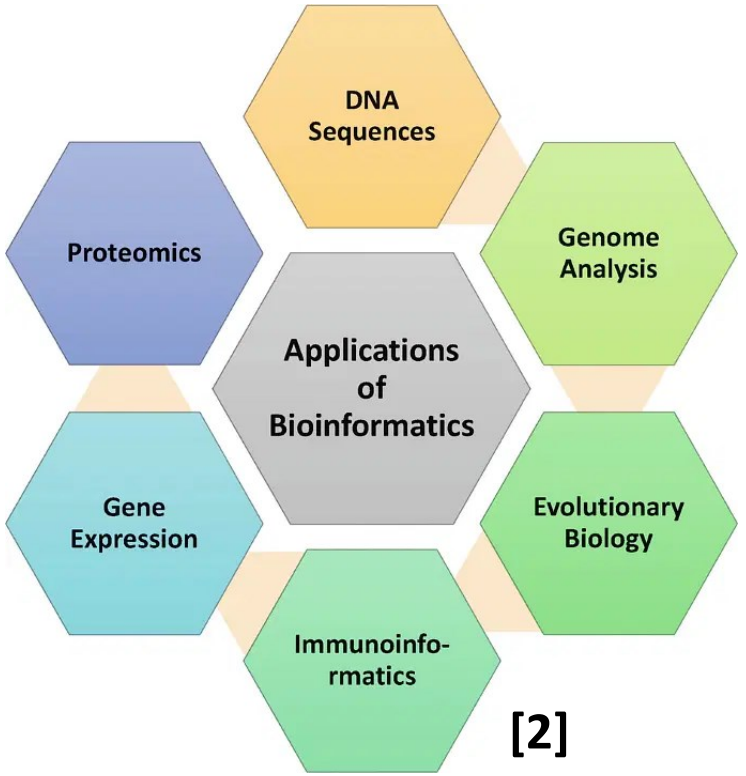
DaCCoTA
DAKOTA COMMUNITY COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Fields & Topics

UND UNIVERSITY OF NORTH DAKOTA.



[2]

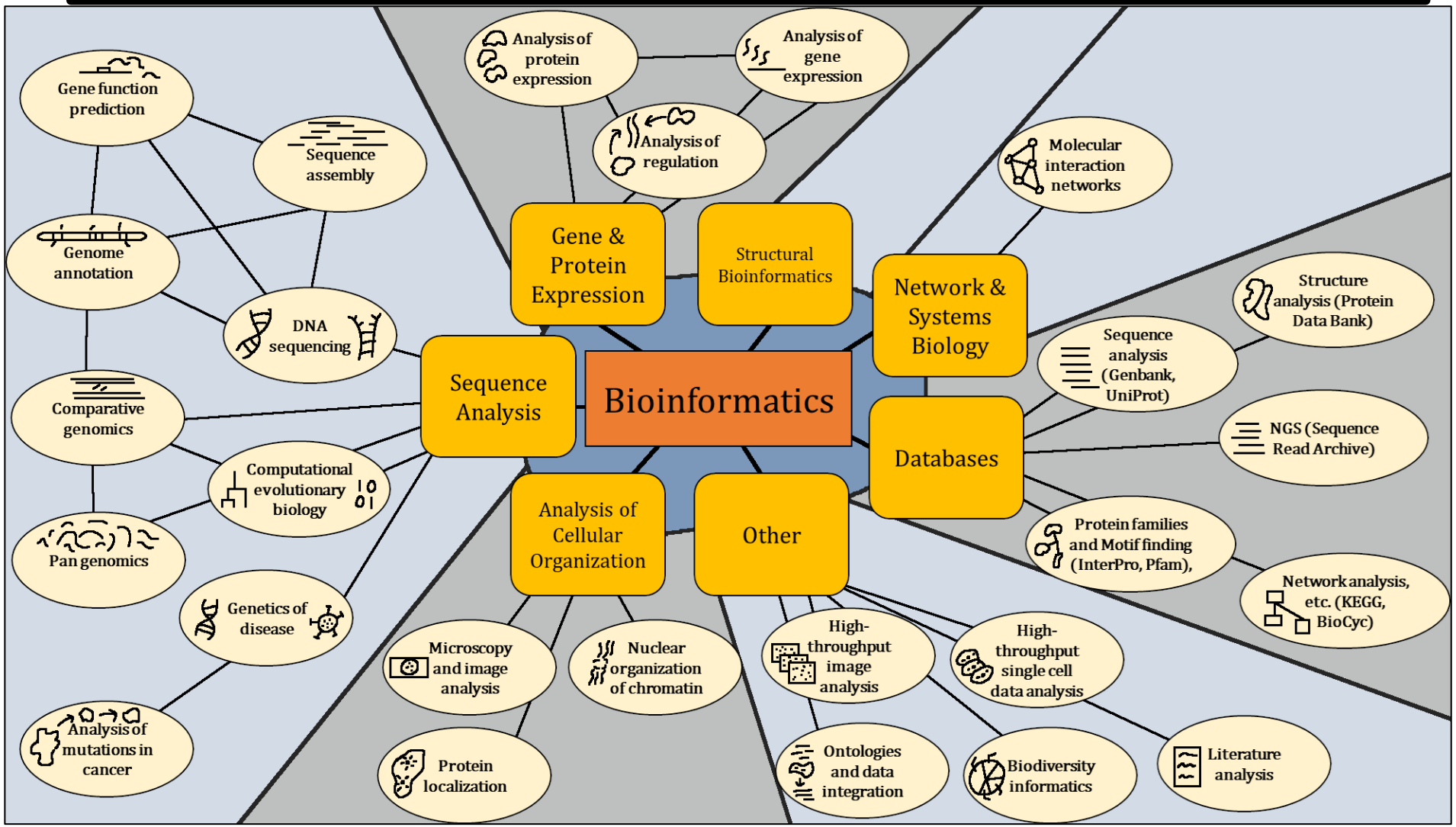


[2]

DaCCoTA
DAKOTA COMMUNITY COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Fields & Topics

UNIVERSITY OF NORTH DAKOTA

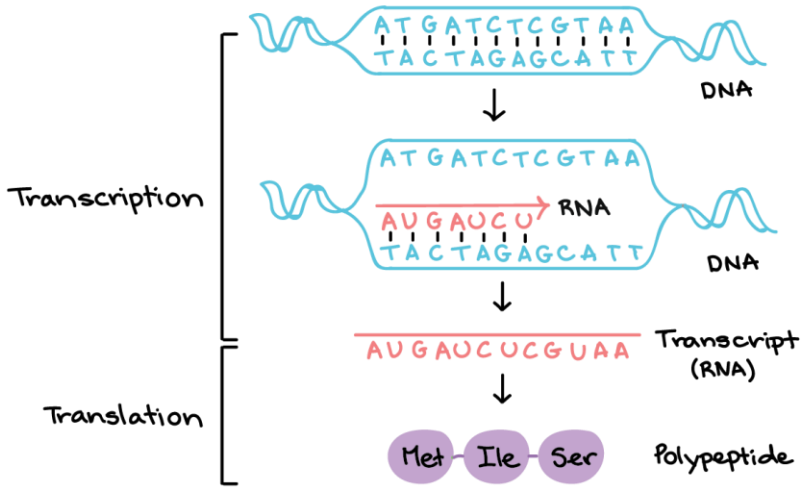


DaCCoTA
DAKOTA COMMUNITY COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Data

UND UNIVERSITY OF NORTH DAKOTA.

- Biological sources
  - DNA: sequences, genes, genomes, metagenomes
  - RNA and Proteins
  - Others [4]
    - Metadata
    - Images
    - Etc.
- File formats [5]
  - FASTA and FASTQ
  - Alignments (BAM, SAM, CRAM)
  - Other (VCF, GFF, GTF, BED, Tar.gz, PDB, PED, MAP, CSV, JSON)

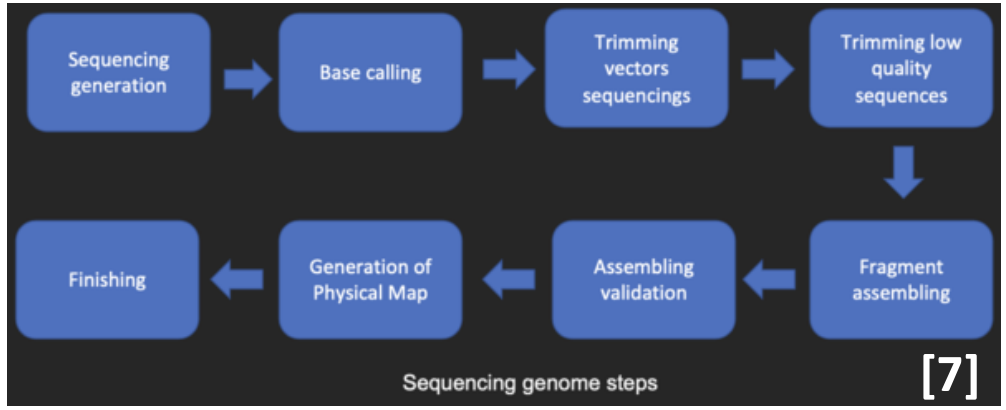


[3]

>Streptomyces sp. PAMC 26508|GCF_000364805.1|NC_021055.1|Chromosome: ANONYMOUS|3355624..3357149 -
AAAGGAGGTGATCCAGCCGCACCTTCCGGTACGGCTACCT
TGTTACGACTTCGTCCCAATCGCCAGTCCCACCTTCGAC
AGCTCCCTCCCACAAGGGGTTGGGCCACCGGCTTCGGGTG
TTACCGACTTTCGTGACGTGACGGGCGGTGTGTACAAGG
CCCGGGAACGTATTCACCGCAGCAATGCTGATCTGCGATT
ACTAGCAACTCCGACTTCATGGGGTCGAGTTGCAGACCC
CAATCCGAACTGAGACCGGCTTTTTGAGATTCGCTCCGCCT
CGCGGCATCGCAGCTCATTGTACCGGCCATTGTAGCAC
GTGTGCAGCCCAAGACATAAGGGGCATGATGACTTGACG
TCGTCCCCACCTTCCTCCGAGTTGACCCCGGCAGTCTCCT
GTGAGTCCCCATCACCCCGAAGGGCATGCTGGCAACACA
GAACAAGGGTTGCGCTCGTTGCGGGACTTAACCCAACATC
TCACGACACGAGCTGACGACAGCCATGCACCACCTGTATA
CCGACCACAAGGGGGGCACCATCTCTGATGCTTTCCGGT

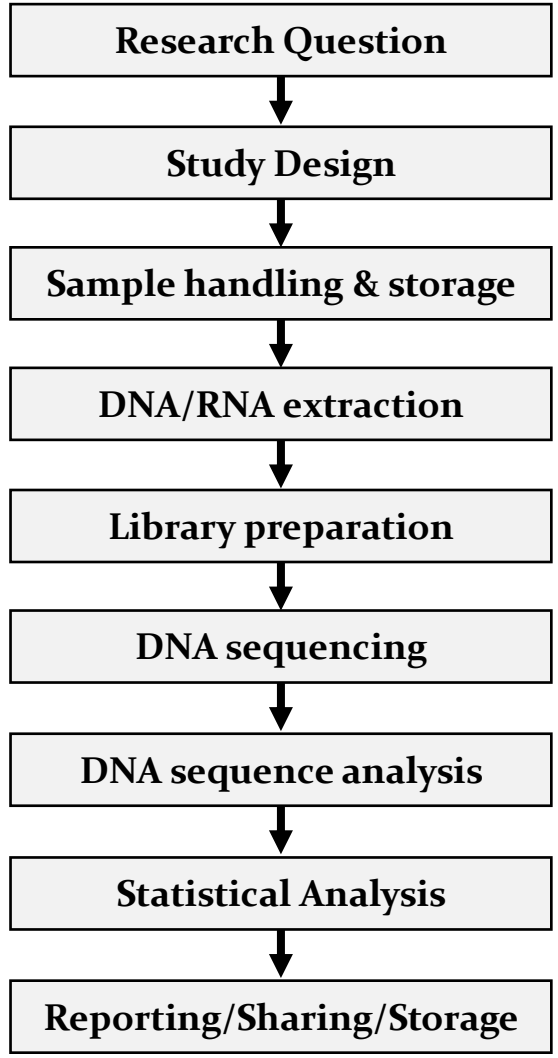
Methods

- Pipelines
- DNA Sequencing
  - Amplicon sequencing: 16/18S, single copy, etc.
  - Genome sequencing: reference based, *de novo*
  - Annotation (what's there)
  - Comparison (what's different between samples/organisms)
- Gene and Protein expression
  - RNA-Seq
  - Microarray
- Phylogeny
- Networks
- Omics



[7]

[6]

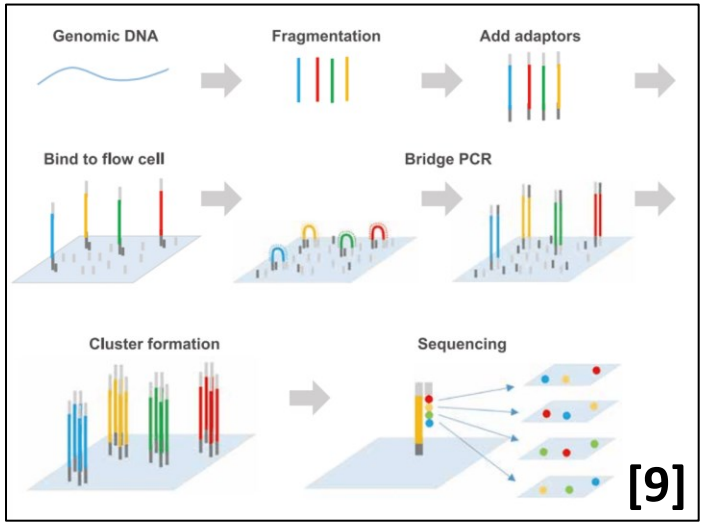


DaCCoTA
DAKOTA COMMUNITY COLLABORATIVE ON TRANSLATIONAL ACTIVITY

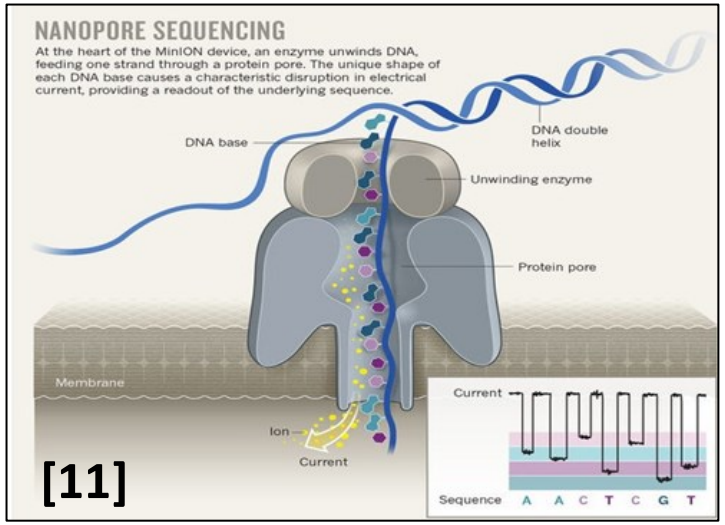
Machines

UNIVERSITY OF NORTH DAKOTA

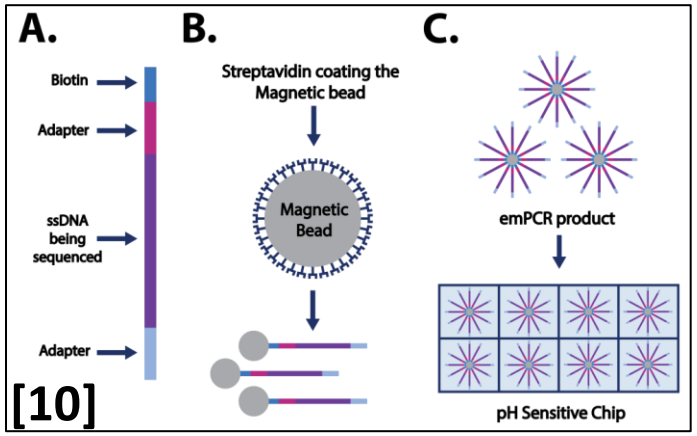
- DNA Sequencing [8]:
  - Illumina
  - Ion Torrent
  - Oxford Nanopore
  - Pacific Biosystems
- Mass Spectrometry
- Flow Cytometry



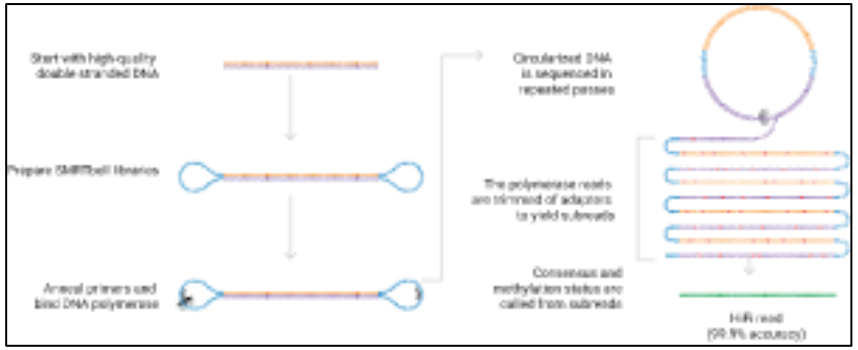
[9]



[11]



[10]


[12]

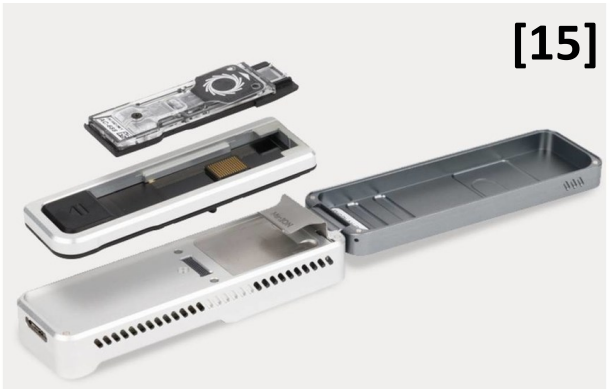
DaCCoTA
DAKOTA COMMUNITY COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Machines

UND UNIVERSITY OF NORTH DAKOTA.

- DNA Sequencing [8]:
  - Illumina
  - Ion Torrent
  - Oxford Nanopore
  - Pacific Biosystems
- Mass Spectrometry
- Flow Cytometry

[13]


[15]

[14]

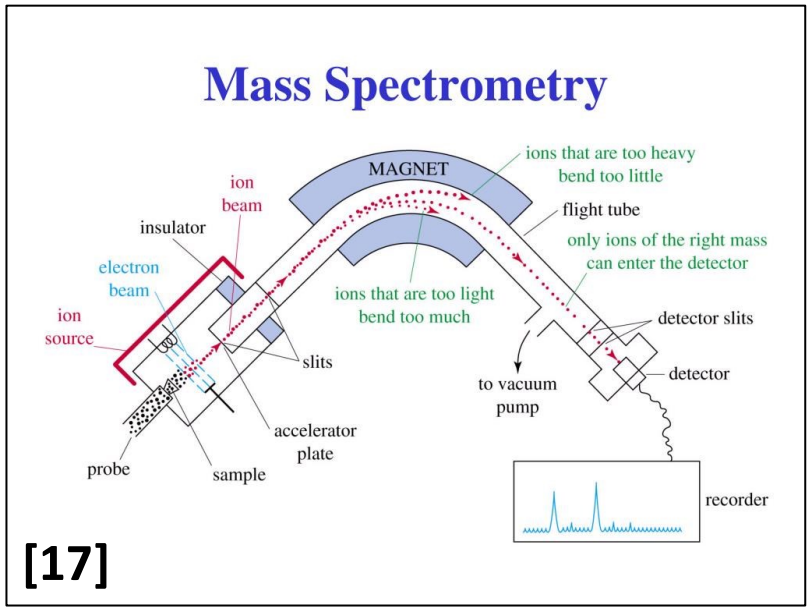

[16]

DaCCoTA
DAKOTA COMMUNITY COLLABORATIVE ON TRANSLATIONAL ACTIVITY

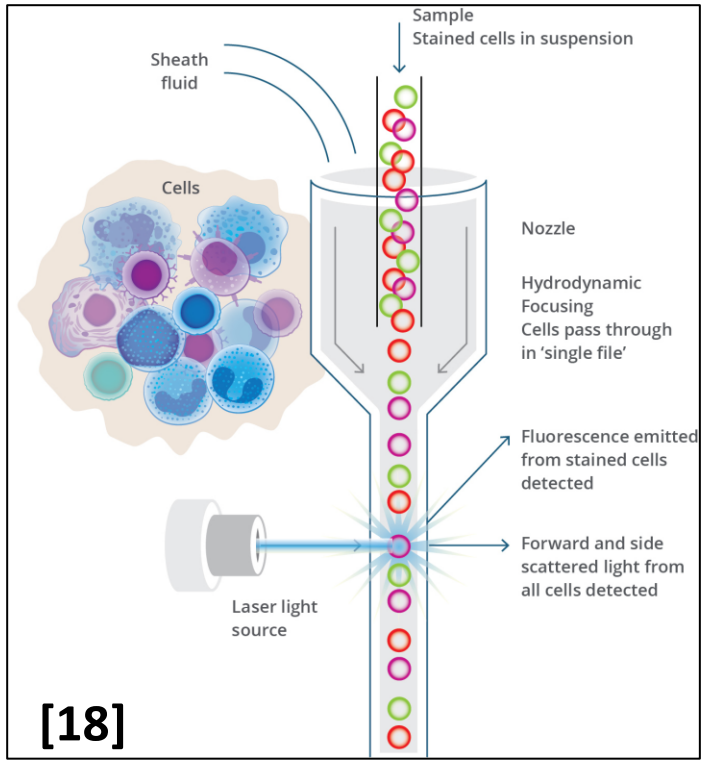
Machines

UND UNIVERSITY OF NORTH DAKOTA

- DNA Sequencing [8]:
  - Illumina
  - Ion Torrent
  - Oxford Nanopore
  - Pacific Biosystems
- Mass Spectrometry
- Flow Cytometry



[17]



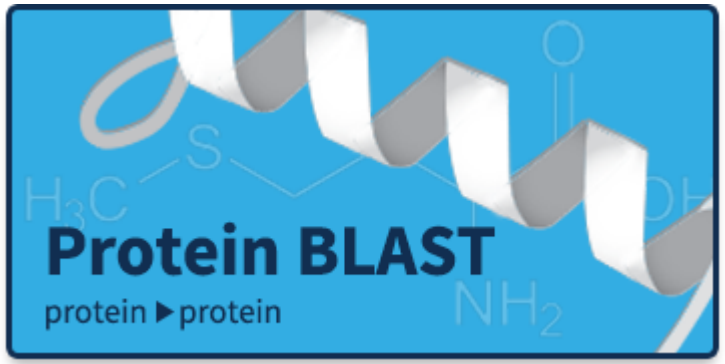
[18]

DaCCoTA
DAKOTA COMMUNITY COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Software

UND UNIVERSITY OF NORTH DAKOTA

- **Core Tools:** BLAST, EMBOSS, Clustal Omega, PROSPECT, Ensembl, Usearch, SAMtools, Velvet, Bowtie, TopHat, Muscle, ProteinTools, etc.
- **Underlying Code:** Needleman-Wunch, Borrows-Wheeler Transform, UPGMA, Neighbor-joining, etc.
- **Software Environments:** Biopython, Bioconductor, BioPerl, etc.
- **Pipelines:** Dada2, DNA-seq Analysis, Center for Genomic Epidemiology, Qiime2, Mothur, Workflow Managers, etc.



DaCCoTA
DAKOTA COMMUNITY COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Software

UND UNIVERSITY OF NORTH DAKOTA.

- **Core Tools:** BLAST, EMBOSS, Clustal Omega, PROSPECT, Ensembl, Usearch, SAMtools, Velvet, Bowtie, TopHat, Muscle, ProteinTools, etc.
- **Underlying Code:** Needleman-Wunch, Borrows-Wheeler Transform, UPGMA, Neighbor-joining, etc.
- **Software Environments:** Biopython, Bioconductor, BioPerl, etc.
- **Pipelines:** Dada2, DNA-seq Analysis, Center for Genomic Epidemiology, Qiime2, Mothur, Workflow Managers, etc.

biopython

Bioconductor
OPEN SOURCE SOFTWARE FOR BIOINFORMATICS

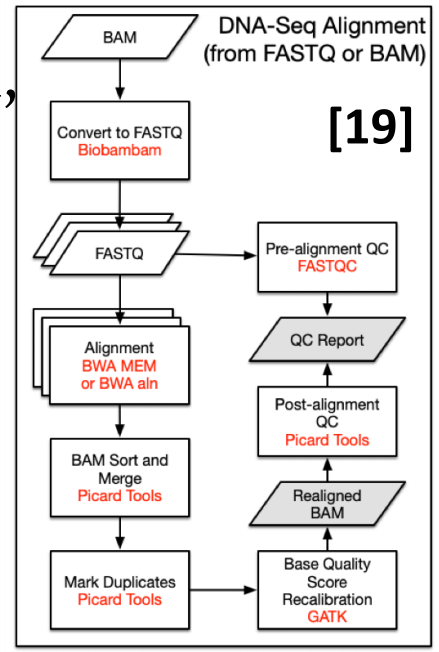
BioPerl

DaCCoTA
DAKOTA COMMUNITY COLLABORATIVE ON TRANSLATIONAL ACTIVITY

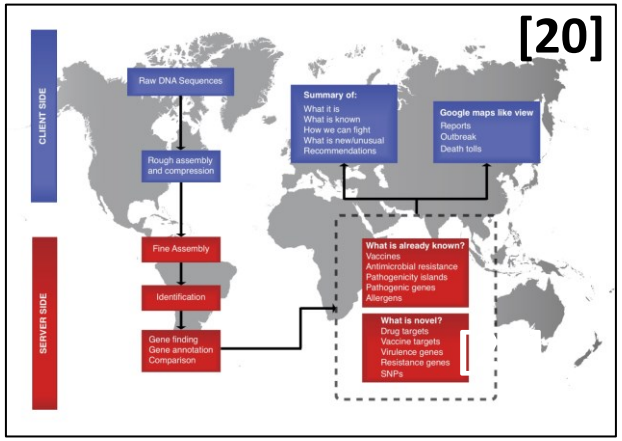
Software

UND UNIVERSITY OF NORTH DAKOTA

- **Core Tools:** BLAST, EMBOSS, Clustal Omega, PROSPECT, Ensembl, Usearch, SAMtools, Velvet, Bowtie, TopHat, Muscle, ProteinTools, etc.
- **Underlying Code:** Needleman-Wunch, Borrows-Wheeler Transform, UPGMA, Neighbor-joining, etc.
- **Software Environments:** Biopython, Bioconductor, BioPerl, etc.
- **Pipelines:** Dada2, DNA-seq Analysis, Center for Genomic Epidemiology, Qiime2, Mothur, Workflow Managers, etc.



qiime2

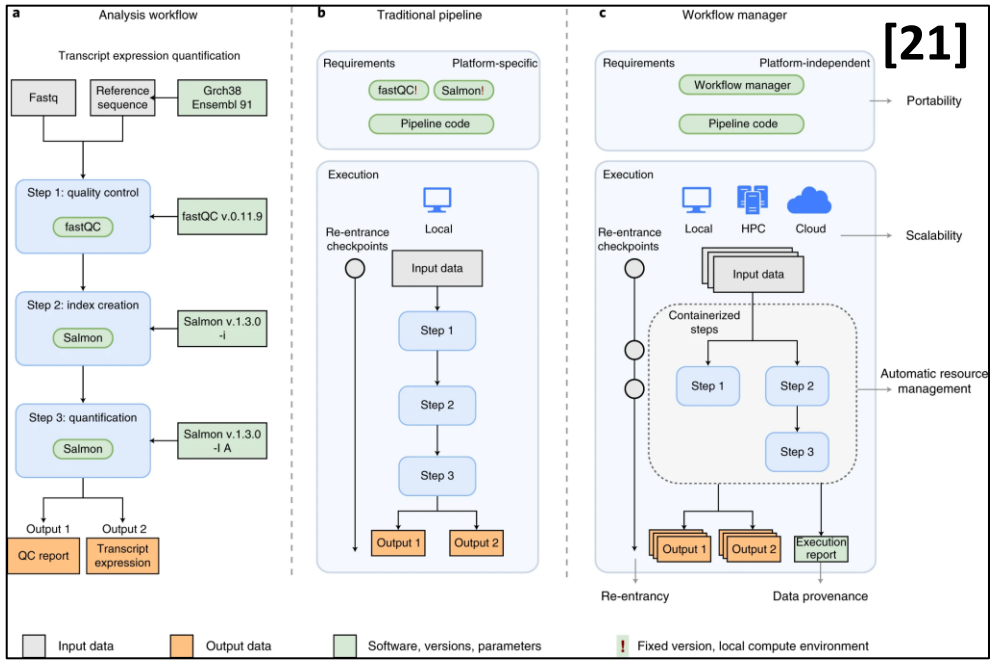


DaCCoTA
DAKOTA COMMUNITY COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Software

UNIVERSITY OF NORTH DAKOTA

- **Core Tools:** BLAST, EMBOSS, Clustal Omega, PROSPECT, Ensembl, Usearch, SAMtools, Velvet, Bowtie, TopHat, Muscle, ProteinTools, etc.
- **Underlying Code:** Needleman-Wunch, Borrows-Wheeler Transform, UPGMA, Neighbor-joining, etc.
- **Software Environments:** Biopython, Bioconductor, BioPerl, etc.
- **Pipelines:** Dada2, DNA-seq Analysis, Center for Genomic Epidemiology, Qiime2, Mothur, Workflow Managers, etc.



[21]

Galaxy PROJECT [22]

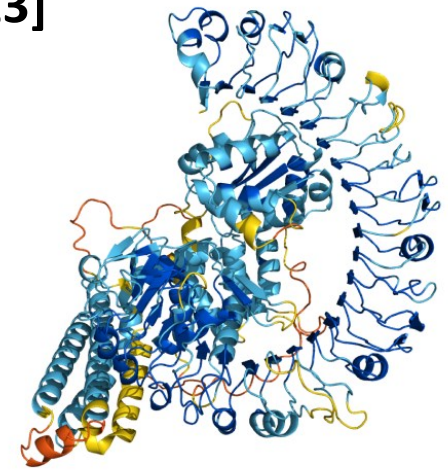
DaCCoTA
DAKOTA COMMUNITY COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Databases

- **Entities:** NCBI, EMBL, Japanese Institute of Genetics
- **DNA:** GenBank, RefSeq
- **Proteins:** UniProt, Alphafold
- **Pathways:** KEGG
- **Taxonomy:** NCBI Taxonomy, Catalogue of Life
- **Interesting:** rrnDB, GTDB, silva

NCBI

[23]



KEGG Kyoto Encyclopedia of Genes and Genomes

UNIVERSITY OF NORTH DAKOTA

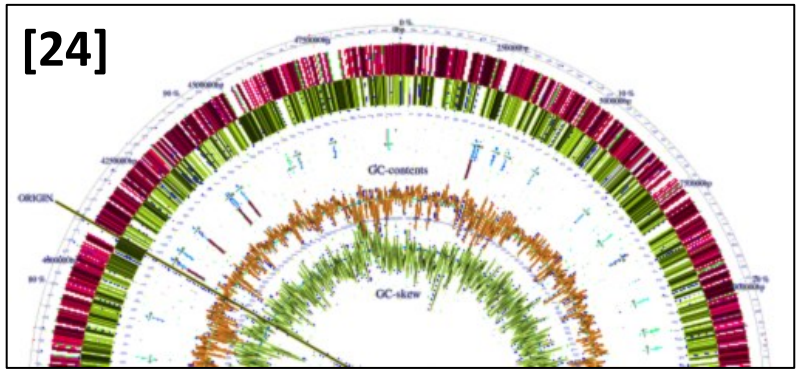
DaCCoTA
DAKOTA COMMUNITY COLLABORATIVE ON TRANSLATIONAL ACTIVITY

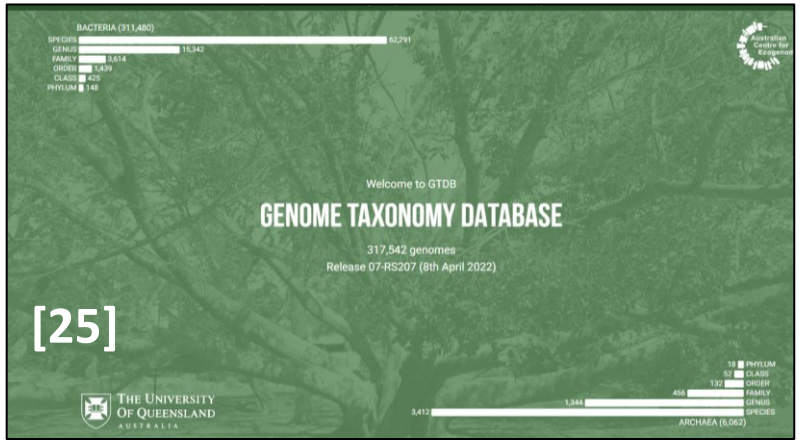
Databases

UND UNIVERSITY OF NORTH DAKOTA.

- **Entities:** NCBI, EMBL, Japanese Institute of Genetics
- **DNA:** GenBank, RefSeq
- **Proteins:** UniProt, Alphafold
- **Pathways:** KEGG
- **Taxonomy:** NCBI Taxonomy, Catalogue of Life
- **Interesting:** rrnDB, GTDB, silva

[24]





[25]

silva
high quality ribosomal RNA databases

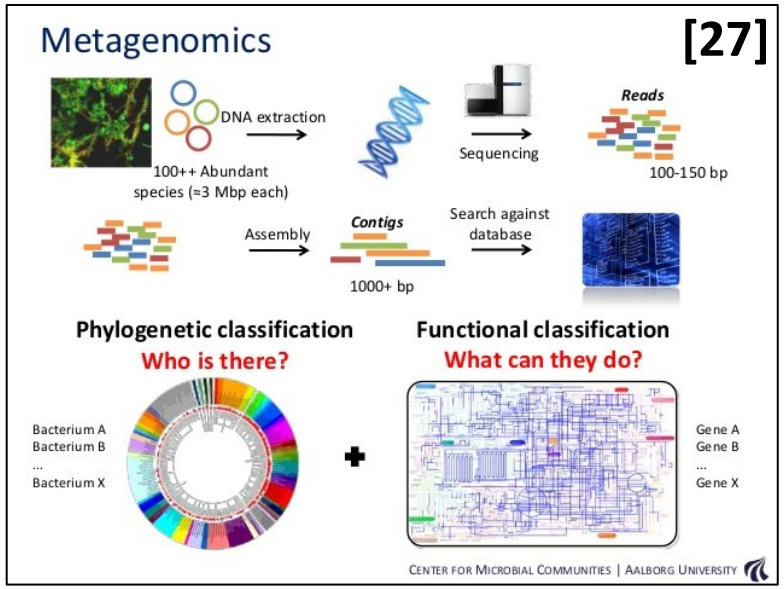
[26]

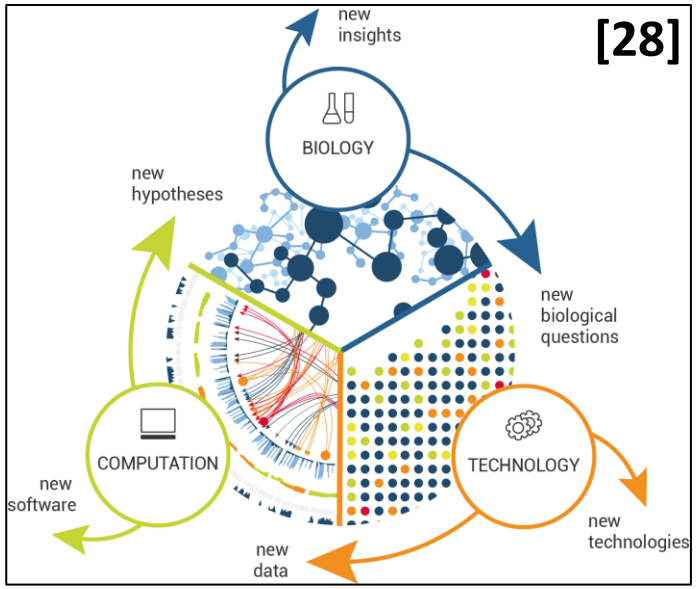
DaCCoTA
DAKOTA COMMUNITY COLLABORATIVE ON TRANSLATIONAL ACTIVITY

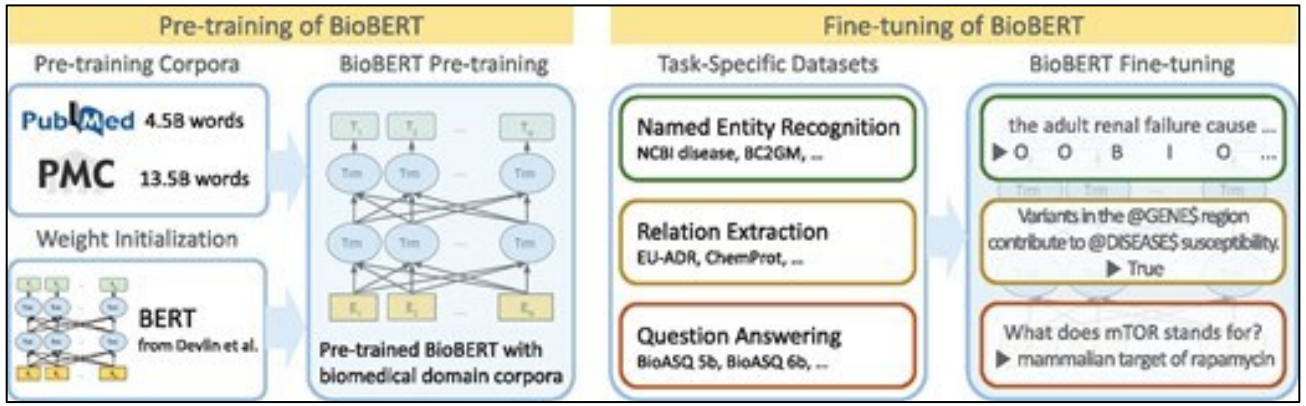
Example Applications

UND UNIVERSITY OF NORTH DAKOTA

- Metagenomics
- Systems Biology
- Literature Analysis
- Microbiome





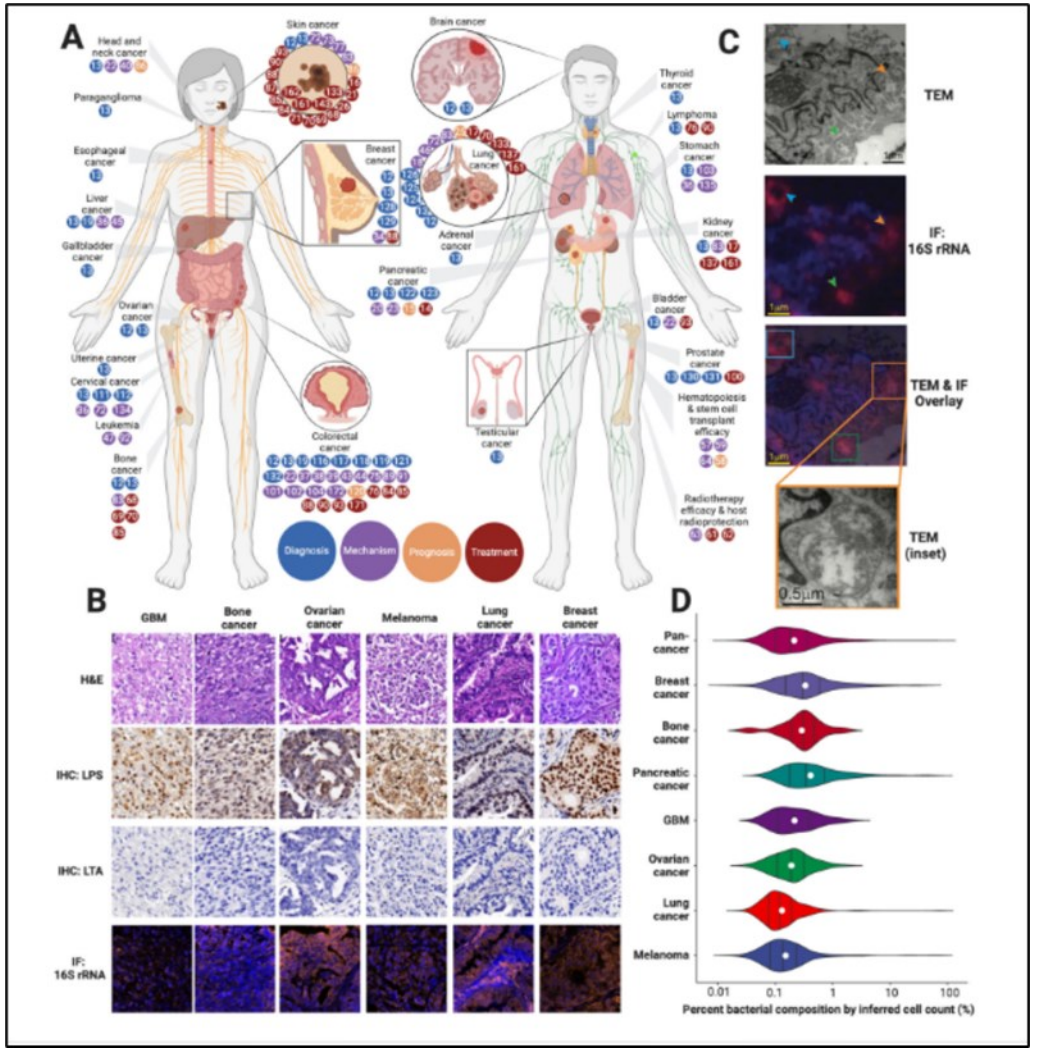


DaCCoTA
DAKOTA COMMUNITY COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Example Applications

UND UNIVERSITY OF NORTH DAKOTA

- Metagenomics
- Systems Biology
- Literature Analysis
- Microbiome



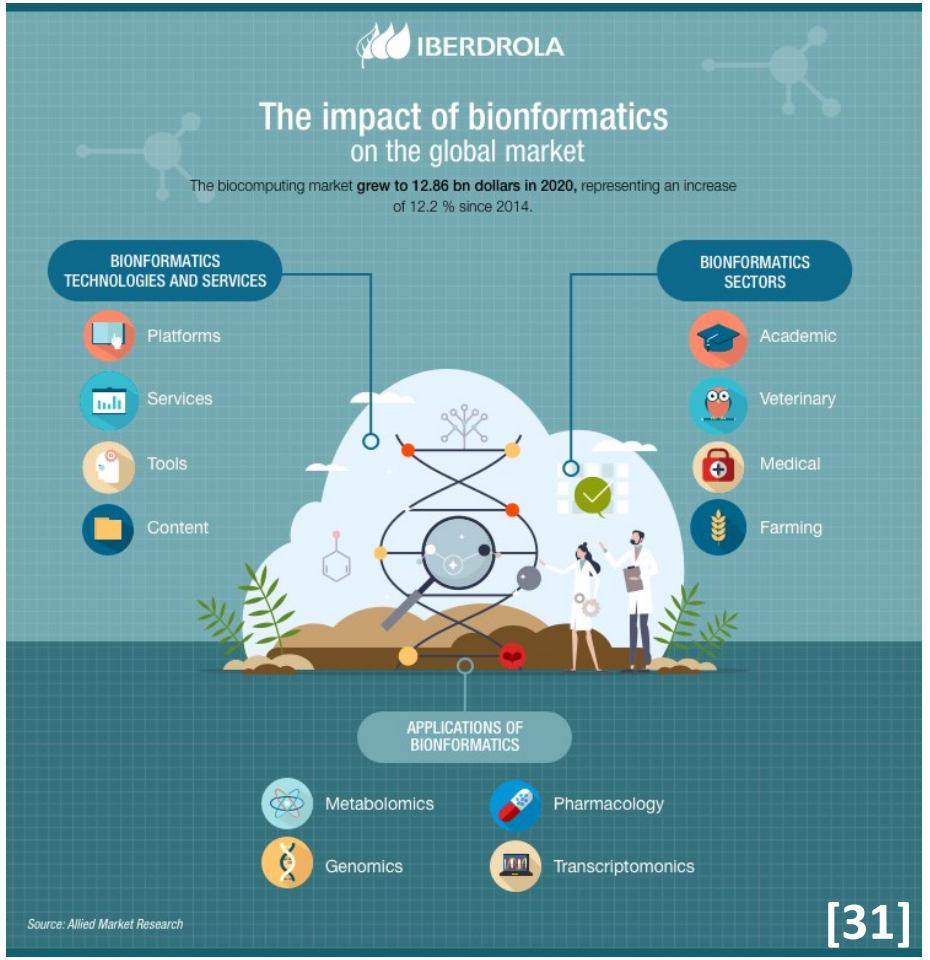
[30]

DaCCoTA
DAKOTA COMMUNITY COLLABORATIVE ON TRANSLATIONAL ACTIVITY

The Future

UND UNIVERSITY OF NORTH DAKOTA

- Increasing market share
- Use of robotics, machine learning, deep learning, AI, etc. [32]
- More data, more integration
- The future is long? [33]



[31]

DaCCoTA
DAKOTA COMMUNITY COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Resources

UND UNIVERSITY OF NORTH DAKOTA.

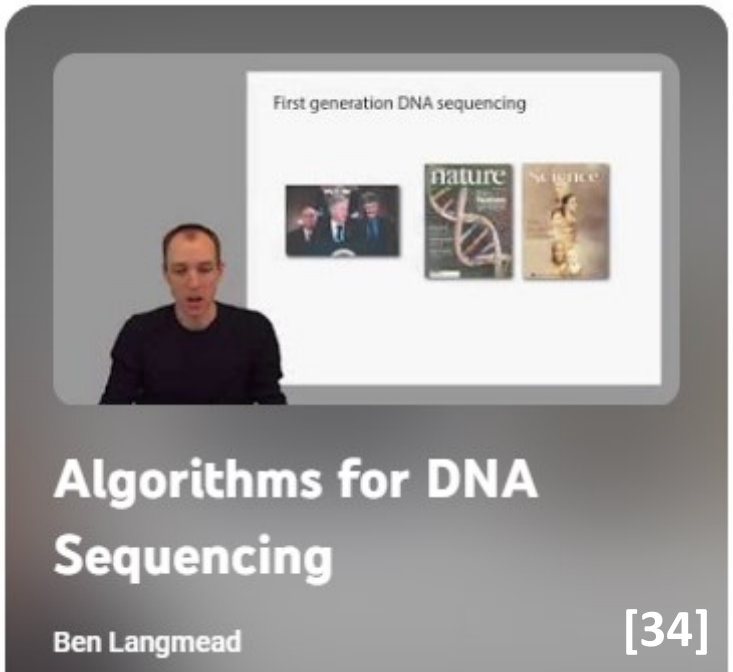
- Algorithms for DNA Sequencing
- Joint Genome Institute
  - NeLLi 2021 [35]
  - VEGA Symposium
  - Annual Genomics of Energy & Environment Meeting
- Center for Genomic Epidemiology (TDU)
  - Metagenomics applied to surveillance of pathogens and antimicrobial resistance
  - Whole genome sequencing of bacterial genomes-tools and applications
- Bioinformatics.ca [36]

bioinformatics.ca

JGI JOINT GENOME INSTITUTE
A DOE OFFICE OF SCIENCE USER FACILITY

Center for Genomic Epidemiology

Home Services Publications Contact



DaCCoTA
DAKOTA COMMUNITY COLLABORATIVE ON TRANSLATIONAL ACTIVITY

References 1

[1] https://pubmed.ncbi.nlm.nih.gov/30084940/

[2] https://collegedunia.com/exams/bioinformatics-biology-articleid-1454

[3] https://sites.google.com/site/obenscience7e/_/rsrc/1490190066963/unit-7/from-dna-to-protein/dna%20to%20rna%20to%20protein.png

[4] https://www.ebi.ac.uk/ols/ontologies/edam

[5] https://www.formbio.com/blog/your-essential-guide-different-file-formats-bioinformatics

[6] https://www.coursera.org/learn/wgs-bacteria/

[7] https://en.wikipedia.org/wiki/Bioinformatics

[8] https://genohub.com/ngs-instrument-guide/

[9] https://blogs.iu.edu/ncgas/files/2021/09/Picture1.png

[10] https://i0.wp.com/apollo-institute.org/wp-content/uploads/2021/11/IonTSeq-Figure-1-01-01.png?resize=1024%2C636&ssl=1

[11] https://www.medgadget.com/wp-content/uploads/2020/04/nanopore.png

[12] https://www.pacb.com/technology/hifi-sequencing/

DaCCoTA
DAKOTA COMMUNITY COLLABORATIVE ON TRANSLATIONAL ACTIVITY

References 2

[13] https://img.medicalexpo.com/images_me/photo-g/83632-12919744.jpg

[14] http://genearrays.com/wp-content/uploads/2014/03/pgm.jpg

[15] https://mma.prnewswire.com/media/834434/Oxford_Nanopore_Flongle.jpg?p=facebook

[16] https://www.pacb.com/revio/

[17] https://image.slideserve.com/464232/mass-spectrometry-l.jpg

[18] https://www.streck.com/wp-content/uploads/2020/11/flow-cytometry-light-scatter-illustration.jpg

[19] https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/DNA_Seq_Variant_Calling_Pipeline/

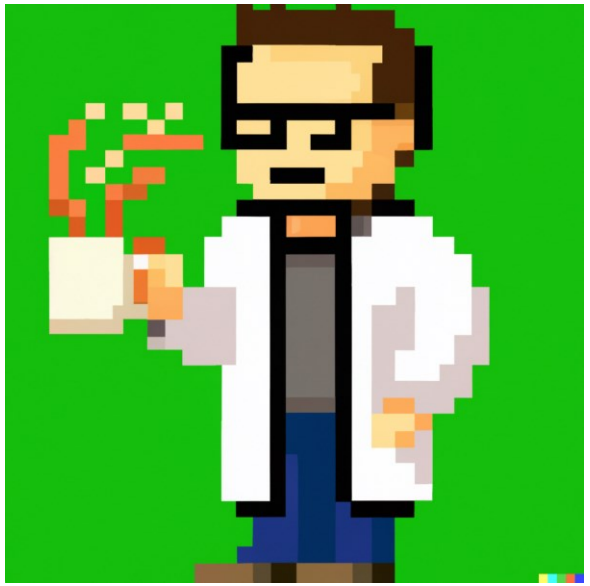
[20] https://www.genomicepidemiology.org/

[21] https://www.nature.com/articles/s41592-021-01254-9/figures/1

[22] By Anton Nekrutenko, http://galaxyproject.org/, Fair use, https://en.wikipedia.org/w/index.php?curid=32778899

[23] https://alphafold.ebi.ac.uk/

[24] https://rrndb.umms.med.umich.edu/



UND
UNIVERSITY OF NORTH DAKOTA

DaCCoTA
DAKOTA COMMUNITY COLLABORATIVE ON TRANSLATIONAL ACTIVITY

References 3

[25] https://gtdb.ecogenomic.org/

[26] https://www.arb-silva.de/

[27] http://myriverside.sd43.bc.ca/gracynk2015/files/2016/11/METagenomics-1-245mzc9.jpg

[28] https://www.omicscouts.com/media/files/blockcontent/2016-09/SystemsBiology.jpg

[29] https://academic.oup.com/bioinformatics/article/36/4/1234/5566506?login=false

[30] https://pubmed.ncbi.nlm.nih.gov/33766858/

[31] https://www.iberdrola.com/innovation/bioinformatics

[32] https://www.linkedin.com/pulse/future-bioinformatics-trends-predictions-venkatesh-chellappa/

[33] https://www.youtube.com/watch?v=KVz6UtNaWbE&list=PL2mpR0RYFQsBiCWVJSvVAO3OJ2t7DzoHA&index=54

[34] https://www.youtube.com/playlist?list=PL2mpR0RYFQsBiCWVJSvVAO3OJ2t7DzoHA

[35] https://www.youtube.com/playlist?list=PLkxZMDuKlaKs80mdPGKRl1r7DKII1O8L1

[36] https://bioinformaticsdotca.github.io/

DaCCoTA
DAKOTA COMMUNITY COLLABORATIVE ON TRANSLATIONAL ACTIVITY

Acknowledgements

The DaCCoTA is supported by the National Institute of General Medical Sciences of the National Institutes of Health under Award Number U54GM128729.

For the labs that use the Biostatistics, Epidemiology, and Research Design Core in any way, including this Module, please acknowledge us for publications: ***"Research reported in this publication was supported by DaCCoTA (the National Institute of General Medical Sciences of the National Institutes of Health under Award Number U54GM128729)"***

DaCCoTA
DAKOTA COMMUNITY COLLABORATIVE ON TRANSLATIONAL ACTIVITY

1010 1010
1010 1010