

Bioinformatics and computational biology: a statistician's perspective

Shofi Andari

Departemen Statistika
Institut Teknologi Sepuluh Nopember, Surabaya

Statistics Online Seminar Series #02
July 10, 2020



IOWA STATE
UNIVERSITY

Outline

- A brief background
- What is bioinformatics, really?
- The central dogma in molecular biology
- The rise of omics fields and data
- A sneak peak on some databases
- Statistical challenges: data structure and analysis
- More examples on how statistics contributes in solving biological problem
 - 1 Phylogenetic tree for investigating SARS-CoV-2
 - 2 Hidden Markov model for sequence merging

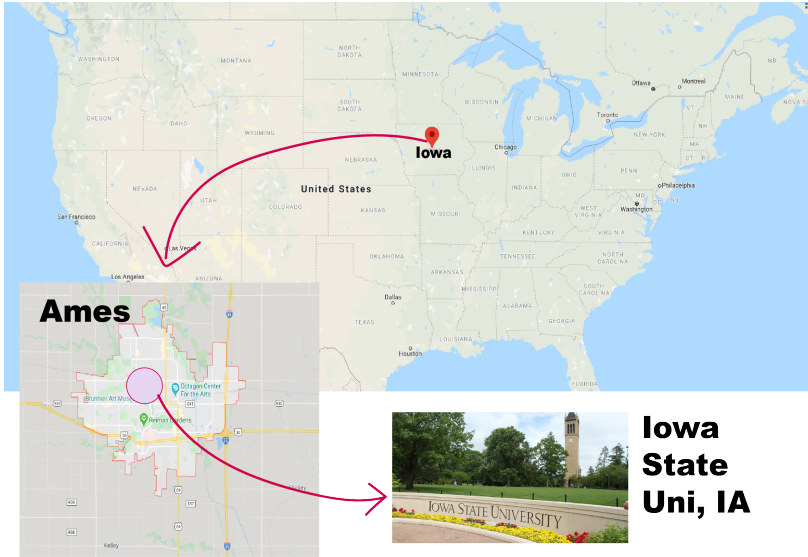
A (not so) brief background



Shofi Andari joined the Dept. of Statistics FSAD-ITS in 2013 as a teaching staff and is a member of *Lab. Statistika Lingkungan dan Kesehatan*.

She started to pursue her Ph.D. at [Bioinformatics and Computational Biology \(BCB\)](#) Program, under home department in Dept. of Statistics, Iowa State U of Science and Technology, Ames, IA, USA since Fall 2017 with a DIKTI Funded-Fulbright Foreign Program scholarship.

A (not so) brief background cont'd



About Bioinformatics and Computational Biology (BCB) Program



Figure 1: Molecular Biology Building (MBB) at Iowa State University

BCB Program at Iowa State U: ~20 departments

- an *interdepartmental* program (interdisciplinary)
- participated by ~20 departments in the university
 - Agronomy
 - Animal Science
 - Astronomy Physics
 - Biochemistry
 - Ecology, Evolution, and Organismal Biology
 - Chemistry
 - Biophysics and Molecular Biology
 - Veterinary Pathology
 - Biomedical Sciences
 - **Statistics**
 - Mathematics
 - Computer Science
 - Electrical and Computer Engineering
 - *more on <https://www.bcb.iastate.edu/>*

BCB Program at Iowa State U: the people (some of them)



Dr. Friedberg (DOGE)
Dept. of VMPM



Dr. Dan Nettleton
Dept. of STAT



Dr. Karin Dorman
Dept. of STAT / GDCB



Dr. Peng Liu
Dept. of STAT



Dr. Heike Hoffman
Dept. of STAT



Dr. Phillip Dixon
Dept. of STAT



Dr. Yandeau-Nelson
Dept. of GDCB



Dr. Tracy Heath
Dept. of EEOB



Dr. Xiaoqiu Huang
Dept. of CS



Dr. Eve Wurtele
Dept. of GDCB



Dr. Dennis Lavrov
Dept. of EEOB



Dr. Claus Kadelka
Dept. of MATH



Dr. Guiping Hu
Dept. of IMSE



Dr. Robert Jernigan
Dept. of BBMB



Dr. Adina Howe
Dept. of ABE



Dr. Andrew Severin
Dept. of EEOB



Dr. Steven Rodermel
Dept. of GDCB



Dr. Travesset
Dept. of PHYS



Dr. Jonathan Smith
Dept. of MATH



Dr. Mary Grenlee
Dept. of BIOMED



Dr. Guang Song
Dept. of CS



Dr. Jack Dekkers
Dept. of ANS



Dr. Crystal Lu
Dept. of EEOB



Dr. Suraj Khotari
Dept. of ECpE



Dr. Carson Andorf
Dept. of CS



Dr. Xueyu Song
Dept. of CHEM



Dr. Jack Lutz
Dept. of MATH / CS



Dr. Geetu Tuteja
Dept. of GDCB

What is bioinformatics, really?

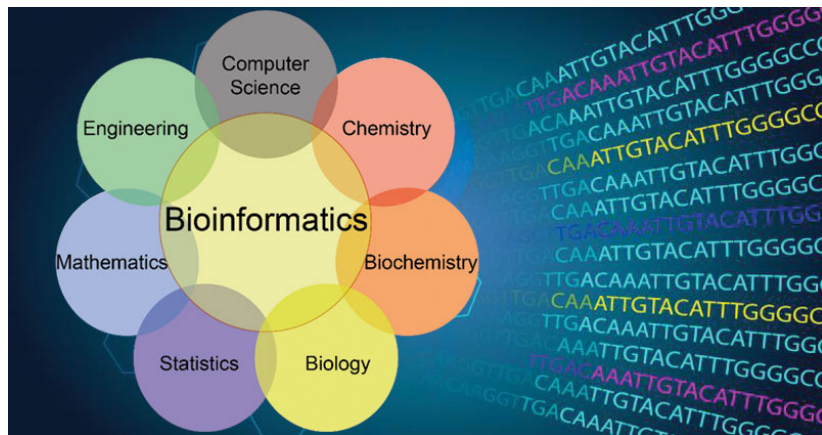


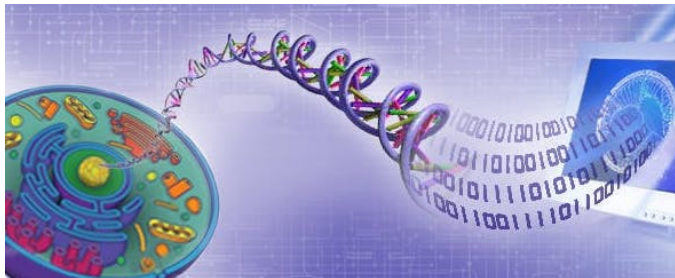
Figure 2: Disciplines building bioinformatics (but not limited to these fields!)

What is bioinformatics, really? cont'd

- "... an interdisciplinary field involving computational biologists, computer scientists, mathematical modelers, systems biologists, and statisticians exploring different facets of the data ranging from storing, retrieving, organizing and subsequent analysis of biological data." (Morris, 2017)

What is bioinformatics, really? cont'd

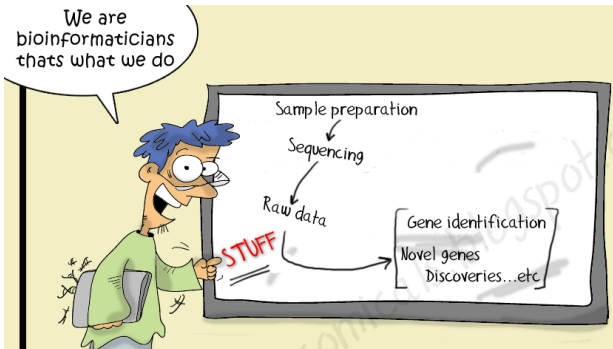
- “.. an interdisciplinary field involving computational biologists, computer scientists, mathematical modelers, systems biologists, and statisticians exploring different facets of the data ranging from storing, retrieving, organizing and subsequent analysis of biological data.” (Morris, 2017)
- “.. is the data science of biology”



What is bioinformatics, really? cont'd

Computational biology

- "... is **translating** and framing **biological** problems into **computational** problems (i.e., algorithms, math model derivations)"
- Often times, *bionformatics* and *computational biology* are used interchangeably.



The central dogma in molecular biology

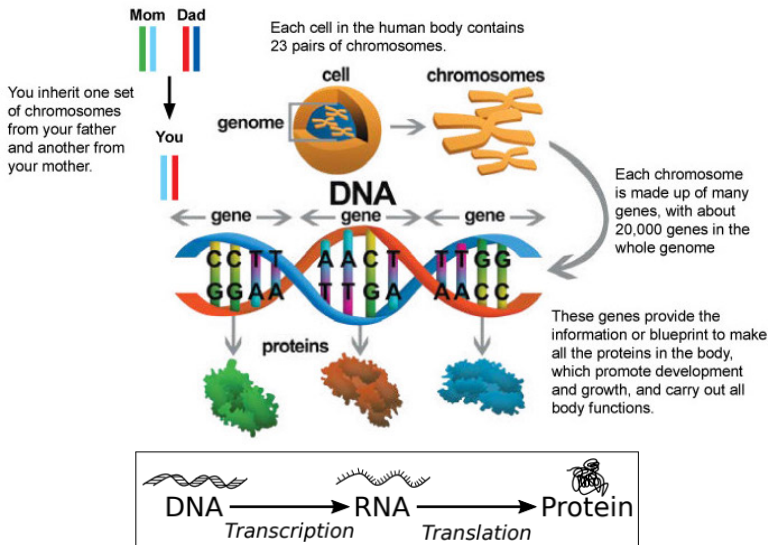


Figure 3: The central dogma of genes to protein (Ritz2018)

The central dogma in molecular biology cont'd

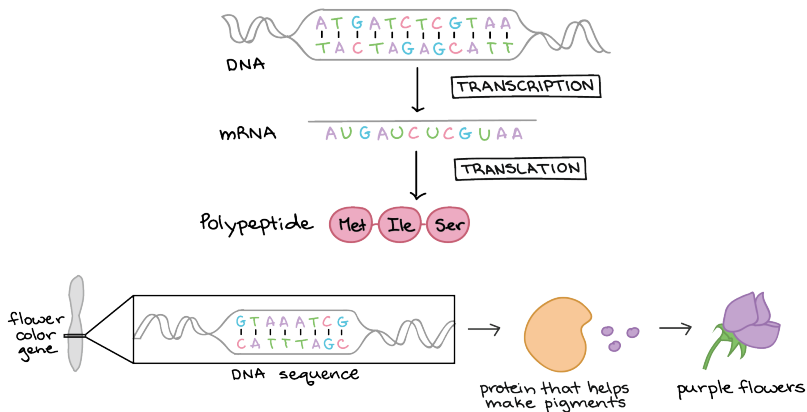


Figure 4: Mendel's flower color gene provides instructions for a protein that helps make colored molecules (pigments) in flower petals (Hellens (2010), Reece (2011), www.khanacademy.org)

The rise of omics data

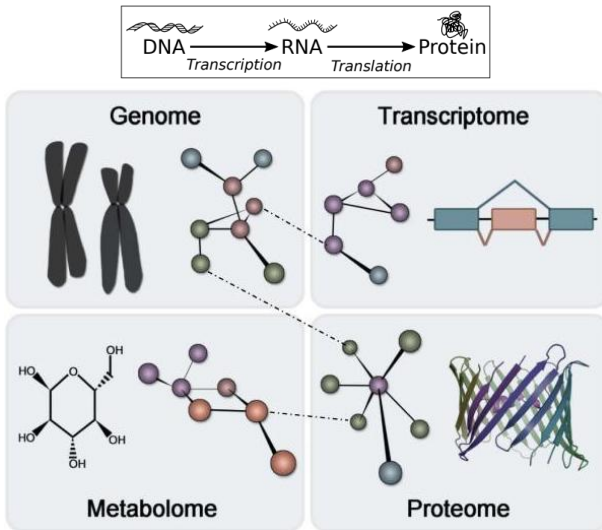


Figure 5: Different networks emerging from the central dogma (Franklin2011)

The rise of omics data cont'd

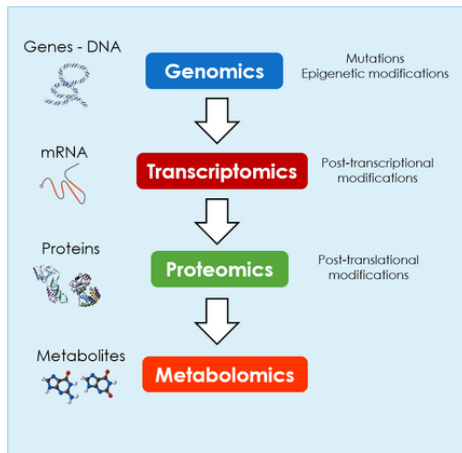


Figure 6: Four main omics fields can be distinguished: genomics (DNA), transcriptomics (mRNA), proteomics (proteins) and metabolomics (metabolites)

(<http://ch4eo.info/research/omics/>)

The rise of omics data: the history

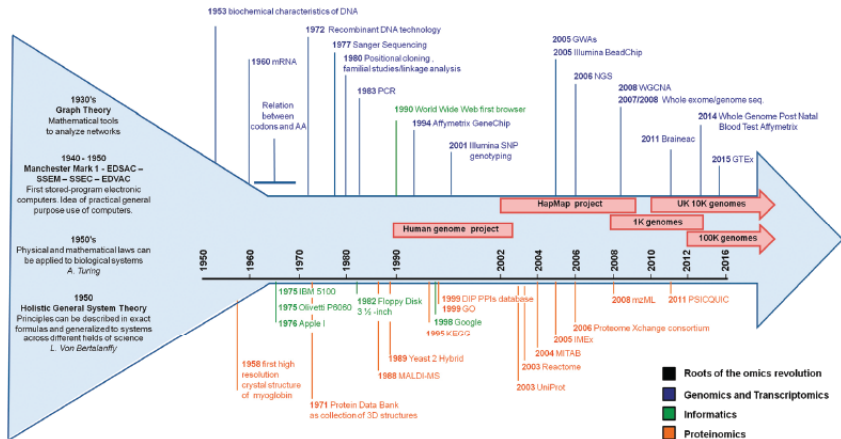
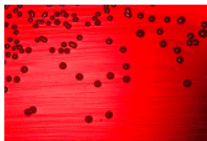
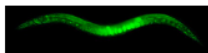


Figure 7: Progressive advance in omic-sciences (Manzoni, 2018)

The rise of omics data: walk of fame



Haemophilus influenzae



Caenorhabditis elegans: the first worm and animal to have its whole genome sequenced



Arabidopsis thaliana: the first plant genome sequenced.



Elaeis guineensis (oil palm): 10 years and 50 scientists to sequence



Human Genome Project (HGP)

Planned since 1984, launched in 1990, completed in 2003.

Remains the biggest collaborative biological project on earth.

Done in various research centers in the US, UK, Japan, France, Germany, China,

Figure 8: Whole-genome sequencing (WGS) historic lane, from bacteria to human (en.wikipedia.org)

The rise of omics data: the falling cost

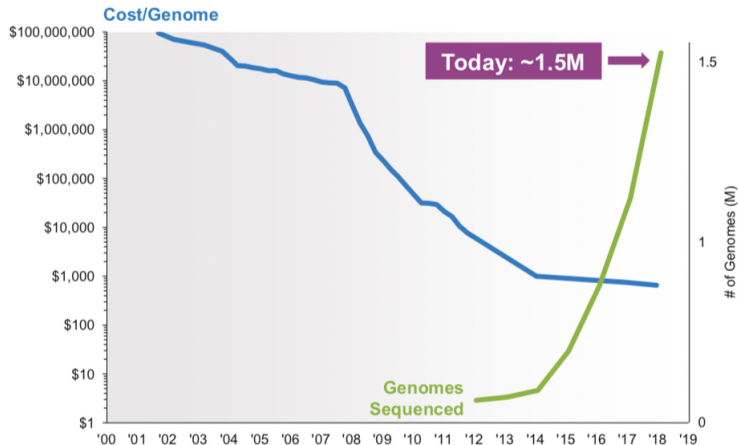


Figure 9: Advances in the field of genomics have led to substantial reductions in the cost of genome sequencing

(<https://www.forbes.com/sites/kenberman/>)

The rise of omics data: the trend

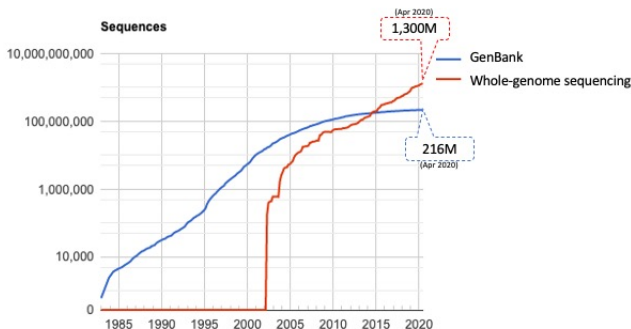


Figure 10: Volume of DNA information in GenBank

(<https://www.ncbi.nlm.nih.gov/genbank/statistics/>)

Statistical challenges



Statistical challenges: data structure and analysis

- Where to access the (benchmark) data sets (if there is any)?
- Let's talk about genomics data: DNA and/or RNA sequences
- Statistical bioinformatics

A sneak peak on some databases

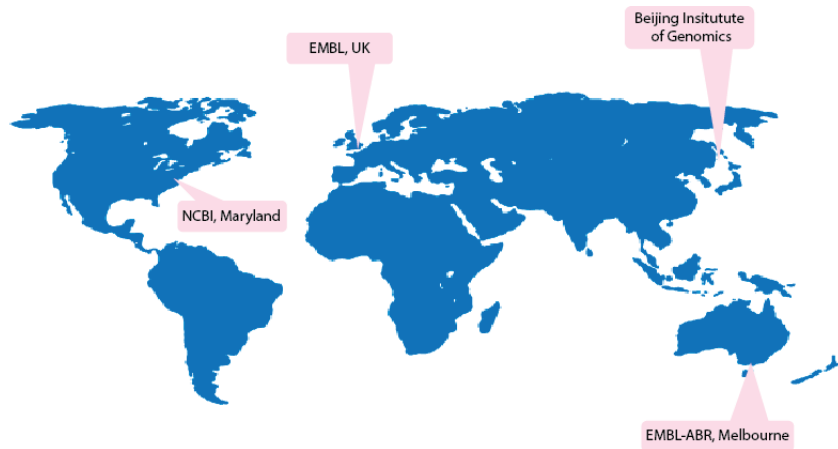


Figure 11: A few of bioinformatics labs on the globe

A sneak peak on NCBI

- National Center for Biotechnology Information (NCBI)
website: <https://www.ncbi.nlm.nih.gov/>
- formed in 1988 as a complement to the activities of the National Institutes of Health (NIH) and the National Library of Medicine (NLM)
- paramount of bioinformatics data bank and tools



NCBI website cont'd

ncbi.nlm.nih.gov

NCBI Resources How To

Sign in to NCBI

NCBI National Center for Biotechnology Information

All Databases Search

COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: <https://www.coronavirus.gov>.
Get the latest research from NIH: <https://www.nih.gov/coronavirus>.
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: <https://www.ncbi.nlm.nih.gov/sars-cov-2/>.

Welcome to NCBI
The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.
[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)

Submit
Deposit data or manuscripts into NCBI databases

Download
Transfer NCBI data to your computer

Learn
Find help documents, attend a class or watch a tutorial

Develop
Use NCBI APIs and code libraries to build applications

Analyze
Identify an NCBI tool for your data analysis task

Research
Explore NCBI research and collaborative projects

Popular Resources
PubMed
Bookshelf
PubMed Central
BLAST
Nucleotide
Genome
SNP
Gene
Protein
PubChem

NCBI News & Blog
We want to hear from you about changes to NIH's Sequence Read Archive data format and storage
30 Jun 2020
NIH's Genianne Read Archive (SRA) is
Improved access to SARS-CoV-2 data
29 Jun 2020
NCBI Datasets has a simple, new way to get Coronaviridae data, including from SARS-CoV-2 (Figure 1). The data
New GenBank submission options for SARS-CoV-2 submitters
25 Jun 2020

NCBI Home
Resource List (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature
Proteins
Sequence Analysis
Taxonomy
Training & Tutorials
Variation

NCBI website: journal papers

The screenshot shows the NCBI website homepage. At the top, there's a navigation bar with 'NCBI', 'Resources', and 'How To'. Below this is a search bar with a dropdown menu set to 'All Databases'. A prominent pink banner in the center contains COVID-19 information, including links to CDC and NIH for the latest public health information and research. On the left, a sidebar lists various resources like 'NCBI Home', 'Resource List (A-Z)', 'All Resources', 'Chemicals & Bioassays', 'Data & Software', and 'DNA & RNA'. The main content area features a 'Welcome to NCBI' message, a description of the center's mission, and links to 'About the NCBI', 'Mission', 'Organization', 'NCBI News & Blog', 'Submit', 'Download', and 'Learn'. On the right, a 'Recent Databases' section lists 'PubMed', 'Bookshelf', 'PubMed Central', 'BLAST', and 'Nucleotide'. A large red box is drawn around the 'PubMed' and 'PubMed Central' links. Below this box, a text block explains that PubMed is a database of citations and abstracts for biomedical literature from MEDLINE and other journals, and that PubMed Central provides full-text versions of articles where available. At the bottom, there are sections for 'Develop' (using NCBI APIs), 'Analyze' (identifying NCBI tools), and 'Research' (exploring NCBI research projects), each with an icon. A 'NCBI News & Blog' section on the right mentions 'Improved access to SARS-CoV-2 data' and 'New GenBank submission options for SARS-CoV-2 submitters'.

ncbi.nlm.nih.gov

NCBI Resources How To Sign In to NCBI

COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: <https://www.cdc.gov/covid-19/>.
Get the latest research from NIH: <https://www.nih.gov/covid19hub>.
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: <https://www.ncbi.nlm.nih.gov/sars-cov-2/>.

NCBI Home
Resource List (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA

Welcome to NCBI
The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.
[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)

Submit Download Learn

Recent Databases
PubMed
Bookshelf
PubMed Central
BLAST
Nucleotide

A database of citations and abstracts for biomedical literature from MEDLINE and additional life science journals. Links are provided when full text versions of the articles are available via PubMed Central (A digital archive of full-text biomedical and life sciences journal literature, including clinical medicine and public health) or other websites.

Proteins
Sequence Analysis
Taxonomy
Training & Tutorials
Variation

Develop
Use NCBI APIs and code libraries to build applications

Analyze
Identify an NCBI tool for your data analysis task

Research
Explore NCBI research and collaborative projects

NCBI News & Blog
We want to hear from you about changes to NIH's Sequence Read Archive data format and storage
30 Jun 2020
NIH's Genomic Data Commons (GDC) is
Improved access to SARS-CoV-2 data
29 Jun 2020
NCBI Datasets has a simple, new way to get Coronaviridae data, including from SARS-CoV-2 (Figure 1). The data
New GenBank submission options for SARS-CoV-2 submitters
28 Jun 2020

NCBI website: BLAST

The screenshot shows the NCBI website homepage. At the top, there's a navigation bar with 'NCBI', 'Resources', and 'How To'. A search bar is present with a dropdown menu set to 'All Databases'. A prominent pink banner in the center contains COVID-19 related information and links. Below the banner, the 'Welcome to NCBI' section describes the center's mission. To the right, under 'Popular Resources', the 'BLAST' link is highlighted with a red rectangle. A red-bordered text box is overlaid on the page, stating: 'an algorithm and program for comparing primary biological sequence information, such as the amino-acid sequences of proteins or the nucleotides of DNA and/or RNA sequences'. The bottom of the page features sections for 'Develop', 'Analyze', and 'Research', each with a brief description and an icon.

ncbi.nlm.nih.gov

NCBI Resources How To Sign In to NCBI

NCBI
National Center for
Biotechnology Information

All Databases Search

COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: <https://www.cdc.gov/covid/>.
Get the latest research from NIH: <https://www.nih.gov/coronavirus>.
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: <https://www.ncbi.nlm.nih.gov/sars-cov-2/>.

NCBI Home
Resource List (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Proteins
Sequence Analysis
Taxonomy
Training & Tutorials
Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)

[Submit](#) [Download](#) [Learn](#) [Nucleotide](#)

BLAST

an algorithm and program for comparing primary biological sequence information, such as the amino-acid sequences of proteins or the nucleotides of DNA and/or RNA sequences

Popular Resources
PubMed
Bookshelf
PubMed Central
BLAST
Nucleotide

NCBI News & Blog

We want to hear from you about changes to NIH's Sequence Read Archive data format and storage

30 Jun 2020

NIH's Genomes, Data, and Bioinformatics (GDB) is

Improved access to SARS-CoV-2 data

29 Jun 2020

NCBI Datasets has a simple, new way to get Coronaviridae data, including from SARS-CoV-2 (Figure 1). The data

New GenBank submission options for SARS-CoV-2 submitters

28 Jun 2020

Develop
Use NCBI APIs and code libraries to build applications

Analyze
Identify an NCBI tool for your data analysis task

Research
Explore NCBI research and collaborative projects

NCBI website: databases

The screenshot shows the NCBI website homepage. At the top, there's a navigation bar with 'NCBI', 'Resources', and 'How To'. A search bar is present with a dropdown menu set to 'All Databases'. A prominent pink banner in the center provides COVID-19 information, including links to CDC and NIH resources. On the left, a 'NCBI Home' sidebar lists various resource categories like 'Resource List (A-Z)', 'All Resources', 'Chemicals & Bioassays', 'Data & Software', 'DNA & RNA', 'Domains & Structures', 'Genes & Expression', 'Genetics & Medicine', 'Genomes & Maps', 'Homology', 'Literature', 'Proteins', 'Sequence Analysis', 'Taxonomy', 'Training & Tutorials', and 'Variation'. The main content area is titled 'Welcome to NCBI' and describes the center's mission. It features six interactive tiles: 'Submit' (deposit data), 'Download' (transfer data), 'Learn' (find help documents), 'Develop' (use NCBI APIs), 'Analyze' (identify NCBI tools), and 'Research' (explore research projects). On the right, a 'Popular Resources' section lists 'PubMed', 'Bookshelf', 'PubMed Central', and 'BLAST'. Below this, a 'Nucleotide' section is highlighted with a red box, containing links for 'Nucleotide', 'Genome', 'SNP', 'Gene', and 'Protein'. Further down, there's a 'NCBI News & Blog' section with recent updates, including 'Improved access to SARS-CoV-2 data' and 'New GenBank submission options for SARS-CoV-2 submitters'.

ncbi.nlm.nih.gov

NCBI Resources How To Sign In to NCBI

NCBI
National Center for
Biotechnology Information

All Databases Search

COVID-19 is an emerging, rapidly evolving situation.
Get the latest public health information from CDC: <https://www.cdc.gov/covid-19/>.
Get the latest research from NIH: <https://www.nih.gov/coronavirus>.
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: <https://www.ncbi.nlm.nih.gov/sars-cov-2/>.

NCBI Home

Resource List (A-Z)

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

Sequence Analysis

Taxonomy

Training & Tutorials

Variation

Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)

Submit
Deposit data or manuscripts
into NCBI databases

Download
Transfer NCBI data to your
computer

Learn
Find help documents, attend a
class or watch a tutorial

Develop
Use NCBI APIs and code
libraries to build applications

Analyze
Identify an NCBI tool for your
data analysis task

Research
Explore NCBI research and
collaborative projects

Popular Resources

PubMed

Bookshelf

PubMed Central

BLAST

Nucleotide

Genome

SNP

Gene

Protein

NCBI News & Blog

We want to hear from you about changes
to NIH's Sequence Read Archive data
format and storage

30 Jun 2020

NIH's Genomes Data Bank (GDB) is

Improved access to SARS-CoV-2 data

29 Jun 2020

NCBI Datasets has a simple, new way to
get Coronaviridae data, including from
SARS-CoV-2 (Figure 1). The data

New GenBank submission options for
SARS-CoV-2 submitters

28 Jun 2020

A sneak peak on EMBL - EBI



A sneak peak on EMBL - EBI cont'd

ebi.ac.uk/services

Overview

A to Z

Data submission

Support

The European Bioinformatics Institute (EMBL-EBI) maintains the world's most comprehensive range of freely available and up-to-date molecular data resources.

Developed in collaboration with our colleagues worldwide, our services let you share data, perform complex queries and analyse the results in different ways. You can work locally by downloading our data and software, or use our [web services](#) to access our resources programmatically.

— You can read more about our services in the journal *Nucleic Acids Research*

Tools & Data Resources

Search all tools & data resources

Tools

Clustal Omega



Multiple sequence alignment of DNA or protein sequences. Clustal Omega replaces the older ClustalW alignment tools.

Multiple sequence alignment

InterProScan



InterProScan searches sequences against InterPro's predictive protein signatures.

Protein feature detection

Sequence motif recognition

Data resources

Ensembl



Genome browser, API and database, providing access to reference genome annotation

UniProt



A comprehensive resource for protein sequence and functional annotation.

PDB



The European resource for the collection, organisation and dissemination of 3D structural data (from PDB and EMDB) on biological macromolecules and their complexes.

Browse by type

DNA & RNA	Gene Expression	Proteins
Structures	Systems	Chemical biology
Ontologies	Literature	Cross domain

Programmatic access

EMBL-EBI web services allow you to query our large biological data resources programmatically, so that you can develop data analysis pipelines or integrate public data with your own applications. The Web Services

A sneak peak on GISAID

- .. a global science initiative and primary source for genomic data of influenza viruses and the novel coronavirus 2019.
- HQ: Munich, Germany
- Website: <https://www.gisaid.org/>



RESEARCH ARTICLE

Global Health

Data, disease and diplomacy: GISAID's innovative contribution to global health


Stefan Elbe¹ and Gemma Buckland-Merrett²

¹Centre for Global Health Policy, School of Global Studies, University of Sussex, Brighton BN1 9SN, UK

²Centre for Global Health Policy, University of Sussex, Brighton BN1 9SN, UK

(Elbe, 2016, doi:10.1002/gch2.1018)

A sneak peak on GISAID cont'd



gisaidd.org

Login

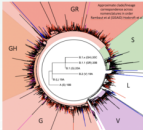
About usDatabase FeaturesEventsCollaborationsReferencesRegistrationHelp

In Focus

Clade and lineage nomenclature aids in genomic epidemiology of active hCoV-19 viruses

Due to the naturally expanding phylogenetic diversity of hCoV-19 viruses in late February 2020, GISAID named for consistent reporting larger clades, based on marker mutations within 6 high-level phylogenetic groupings from the early split of S and L, to the further evolution of L into V and G and later of G into GH and GR.

GISAID clades were eventually augmented with more detailed lineages assigned by the Phylogenetic Assignment of Named Global Outbreak Lineages (PANGOLIN) tool, and a third effort using a Year-Letter nomenclature to facilitate discussion of large-scale diversity patterns of hCoV-19, to aid in the understanding of patterns and determinants of the global spread of the pandemic strain causing COVID-19. [> read more](#)



EpiCoV Data Curation Team

UNIVERSIDAD NACIONAL DE LA PLATA

Universidad Nacional de La Plata, Buenos Aires

Richard Mitter

The Francis Crick Institute, London

Mariana Viegas

National Council of Scientific & Technical Research, Buenos Aires

Claudia Chica

CNRS & Institut Pasteur, Paris

Recent hCoV-19 data submissions

[hCoV-19/India/NCDC7762_CSIR-IGIB/2020](#)

[hCoV-19/Bangladesh/BCSIR-NILMRC_149/2020](#)


[hCoV-19/Hangzhou/HZCDC7378/2020](#)

[hCoV-19/Peru/01/2020](#)

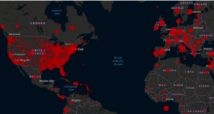
[hCoV-19/Canada/MB_80/2020](#)

Number of hCoV-19 genomic sequences: 60,069


Genomic epidemiology of hCoV-19



COVID-19 Global Cases



GISAID Resources



Free Access Credentials

Register here and join thousands of researchers around the globe.

Shofi Andari

Stats Online Seminar #02: Statistical perspective in BCB

29 / 83

A sneak peak on GISAID cont'd

epicov.org/epi3/frontend#14bf4f

GISAID

© 2008 - 2020 | [Terms of Use](#) | [Privacy Notice](#) | [Contact](#)

You are logged in as **Shofi Andari** - [Logout](#)

Registered Users

EpiFlu™

EpiCoV™

My profile

EpiCoV™

Browse

Downloads

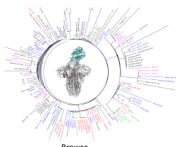
Upload

My Unreleased

Pandemic coronavirus causing COVID-19


A previously unknown human coronavirus (hCoV-19) was first detected in late 2019 in patients in the City of Wuhan, who suffered from respiratory illnesses including atypical pneumonia, an illness that has become known as coronavirus disease (COVID-19). The coronavirus originated from an animal host and is closely related to the virus responsible for the Severe Acute Respiratory Syndrome coronavirus (SARS).

On 10. January 2020, the first virus genomes and associated data were publicly shared via GISAID. The World Health Organization announced on 11. March 2020 the first coronavirus pandemic. As the pandemic progresses, scientists from around the globe are tracking the virus and its genome sequences to ensure optimal virus diagnostic tests, to track and trace the ongoing outbreak and to identify potential intervention options. Several analyses to assist with these efforts are offered here, including sequence alignments, diagnostic primer and probe coordinates, 3D protein models, drug targets, phylogenetic trees and many more.




Browse


Analysis Update




Full genome tree derived from all outbreak sequences (2020-07-04)




Regional clade distribution of new sequences (2020-07-04)




Common primer check for high quality genomes (2020-07-04)




Receptor binding surveillance for complete genomes (2020-07-04)




Full genome tree of hCoV-19-related precursors




Spike glycoprotein receptor binding domain (comparison pangolin, bat, human)



Spike Glycoprotein 3D Structure Model (comparison to SARS, bat)



Potential drug targets highly conserved between hCoV-19 and SARS



analysis update.pdf (2 MB)

Important note: In the [GISAID EpiFlu™ Database Access Agreement](#), you have accepted certain terms and conditions for viewing and using data regarding influenza viruses. To the extent the Database contains data relating to non-influenza viruses, the viewing and use of these data is subject to the same terms and conditions, and by viewing or using such data you agree to be bound by the terms of the [GISAID EpiFlu™ Database Access Agreement](#) in respect of such data in the same manner as if they were data relating to influenza viruses.

Shofi Andari

Stats Online Seminar #02: Statistical perspective in BCB

30 / 83

DNA/RNA sequencing: high-throughput sequencing



Figure 12: Illumina MiSeq

(<https://scientificservices.eu/item/illumina-miseq-next-generation-sequencer/5538>)

DNA/RNA sequencing: how do they do it?

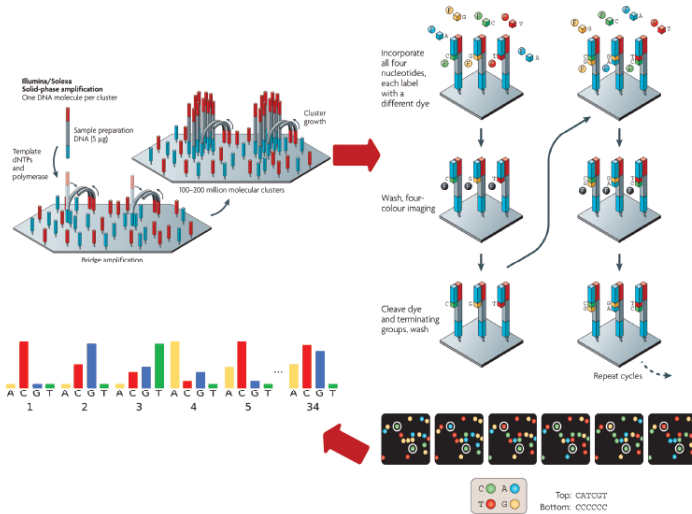


Figure 13: DNA sequencing: amplification, cycle, and base-calling (Whiteford, 2009; Metzker, 2009)

DNA sequences data format: FASTQ format

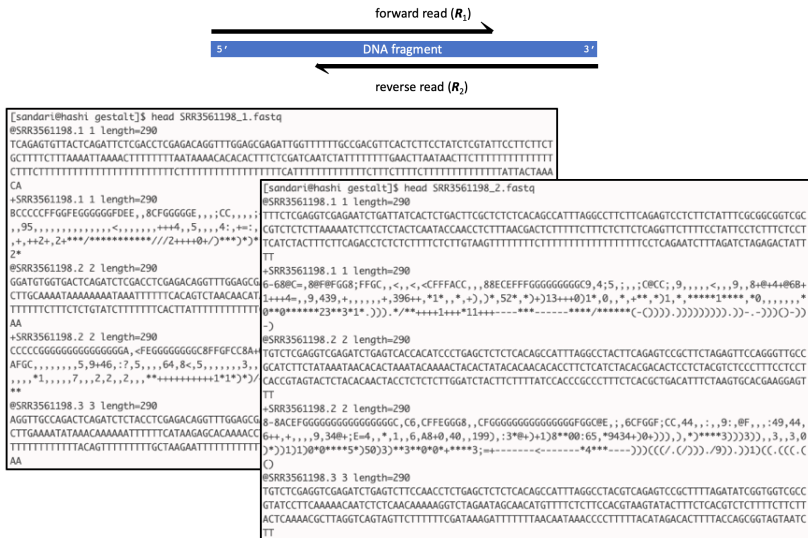


Figure 14: Paired-end sequencing results: FASTQ files

DNA sequences data format cont'd



Figure 15: An example of FASTQ files

DNA sequences data format: FASTA format

Every new entry starts with an ">" sign and an ID

The sequence:
DNA/RNA/protein

```
>hCoV-19/Indonesia/NIHRD-N206/2020|EPI_ISL_413219|2020-03-02
TTACCCAATAATACTGCGTCTTGTTCCACCGCTCTCACTCAACATGGCAAGGAAGACCTTAAATTCCTCGAGGACAAGG
CGTTCCAATTAACCAATAGCAGTCCAGATGACCAATTGGCTACTACCGAAGAGCTACCAGACGAATTCGTGGTGGTG
ACGGTAAATGAAAGATCTCAGTCCAAGATGGTATTTCTACTACCTAGGAAGCTGGGCCAGAAGCTGGACTTCCCTATGGT
GCTAACAAAGACGGCATCATATGGGTTGCAACTGAGGGAGCCTTGAATACACCAAAGATCACATTGGCACCCGC
>hCoV-19/Indonesia/NIHRD651/2020|EPI_ISL_414375|2020-03-01
GTTTGGTGGACCTCAGATTCAACTGGCAGTAACCAGAATGGAGAACGCAGTGGGGCGCGATCAAAACAACGTCGGCCCC
AAGGTTTACCCAATAATACTGCGTCTTGTTCCACCGCTCTCACTCAACATGGCAAGGAAGACCTTAAATTCCTCGAGGA
CAAGGCGTTCCAATTAACCAATAGCAGTCCAGATGACCAATTGGCTACTACCGAAGAGCTACCAGACGAATTCGTGG
TGGTGACGGTAAATGAAAGATCTCAGTCCAAGATGGTATTTCTACTACCTAGGAAGCTGGGCCAGAAGCTGGACTTCCCT
```

Figure 16: An example of FASTA files

Features of genomic data

- Data heterogeneity (various types of data is hard to create a comprehensive view of studies \Rightarrow integrative methods are necessary (Ren, 2015))

Slide courtesy to Nadeem Akhter.

Features of genomic data

- Data heterogeneity (various types of data is hard to create a comprehensive view of studies \Rightarrow integrative methods are necessary (Ren, 2015))
- Large scale data integration

Slide courtesy to Nadeem Akhter.

Features of genomic data

- Data heterogeneity (various types of data is hard to create a comprehensive view of studies \Rightarrow integrative methods are necessary (Ren, 2015))
- Large scale data integration
- High volume data

Slide courtesy to Nadeem Akhter.

Features of genomic data

- Data heterogeneity (various types of data is hard to create a comprehensive view of studies \Rightarrow integrative methods are necessary (Ren, 2015))
- Large scale data integration
- High volume data
- Uncertainty (i.e., due to genetic variants)

Slide courtesy to Nadeem Akhter.

Features of genomic data

- Data heterogeneity (various types of data is hard to create a comprehensive view of studies \Rightarrow integrative methods are necessary (Ren, 2015))
- Large scale data integration
- High volume data
- Uncertainty (i.e., due to genetic variants)
- Data curation

Slide courtesy to Nadeem Akhter.

Features of genomic data

- Data heterogeneity (various types of data is hard to create a comprehensive view of studies \Rightarrow integrative methods are necessary (Ren, 2015))
- Large scale data integration
- High volume data
- Uncertainty (i.e., due to genetic variants)
- Data curation
- Data sharing

Slide courtesy to Nadeem Akhter.

Features of genomic data

- Data heterogeneity (various types of data is hard to create a comprehensive view of studies \Rightarrow integrative methods are necessary (Ren, 2015))
- Large scale data integration
- High volume data
- Uncertainty (i.e., due to genetic variants)
- Data curation
- Data sharing
- Dynamic and subject to change

Slide courtesy to Nadeem Akhter.

Statistical challenges: statisticians' perspective

- The advent of high-throughput multi-platform genomics technologies \Rightarrow highly structured big data
- Bioinformatics is necessarily interdisciplinary in nature: clinical, biological, computational, data management, mathematical modeling, and statistical knowledge and skills
- One of the key attributes that sets statisticians apart from other quantitative scientists is their understanding of variability and uncertainty quantification

(Morris, 2017)

Statistical challenges: statisticians' perspective cont'd

With the basis of **statisticians as data scientists**:

- sampling design decisions
- multi-step processing algorithms
- reductionistic feature extraction
- inferential reasoning
- design algorithms to search high-dimensional spaces
- build predictive models while obtaining accurate measures of their predictive accuracy

(Morris, 2017)

What does it take to be a statistical bioinformatician?

The basic!

- collecting data and experiment design
- descriptive statistics and data visualization
- randomness and probability concept
- estimation: point and interval, MoM, MLE, exponential family, RBT, UMVUE, FI, CRLB
- hypothesis testing: type I and II errors, test statistics, power and sample calculations, NPL, LRT
- simulations

What does it take to be a statistical bioinformatician? cont'd

- Gene mapping and association studies: clustering, maximum likelihood estimations, QTL/eQTL, linear models



RESEARCH ARTICLE

Fast and flexible linear mixed models for genome-wide genetics

Daniel E. Runcie^{1*}, Lorin Crawford²

1 Department of Plant Sciences, University of California Davis, Davis, California, United States of America,

2 Department of Biostatistics, Brown University, Providence, Rhode Island, United States of America

* deruncie@ucdavis.edu



Abstract

Linear mixed effect models are powerful tools used to account for population structure in genome-wide association studies (GWASs) and estimate the genetic architecture of complex traits. However, fully-specified models are computationally demanding and common simplifications often lead to reduced power or biased inference. We describe *Grid-LMM* (<https://github.com/deruncie/GridLMM>), an extendable algorithm for repeatedly fitting complex linear models that account for multiple sources of heterogeneity, such as additive and non-additive genetic variance, spatial heterogeneity, and genotype-environment interactions. *Grid-LMM* can compute approximate (yet highly accurate) frequentist test statistics or Bayesian posterior summaries at a genome-wide scale in a fraction of the time compared to existing general-purpose methods. We apply *Grid-LMM* to two types of quantitative genetic analyses. The first is focused on accounting for spatial variability and non-additive genetic variance while scanning for QTL; and the second aims to identify gene expression traits affected by non-additive genetic variation. In both cases, modeling multiple sources of heterogeneity leads to new discoveries.

OPEN ACCESS

Citation: Runcie DE, Crawford L (2019) Fast and flexible linear mixed models for genome-wide genetics. *PLoS Genet* 15(2): e1007978. <https://doi.org/10.1371/journal.pgen.1007978>

Editor: Michael P. Epstein, Emory University, UNITED STATES

Received: September 7, 2018

Accepted: January 21, 2019

What else?

- Gene regulatory network & multi-omics data integration: gene expression \Leftarrow heatmap, feature extraction (PCA, PLS)

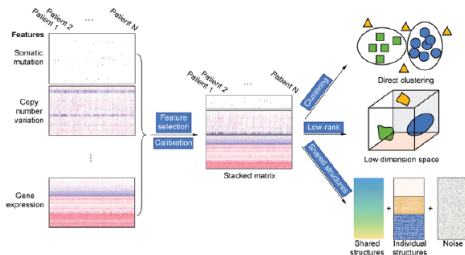
Quantitative Biology 2016, 4(1): 56-67
DOI 10.1007/s40484-015-0063-4

Integrative clustering methods of multi-omics data for molecule-based cancer classifications

Dongfang Wang and Jin Gu*

Ministry of Education Key Laboratory of Bioinformatics and Bioinformatics Division, Center for Synthetic and Systems Biology, Tsinghua National Laboratory for Information Science and Technology/Department of Automation, Tsinghua University, Beijing 100084, China

* Correspondence: jgu@tsinghua.edu.cn



What else? cont'd

- High-throughput data processing: genome assembly, alignment

BMC Bioinformatics



Methodology article

Open Access

A statistical score for assessing the quality of multiple sequence alignments

Virpi Ahola^{*1,2}, Tero Aittokallio^{3,6}, Mauno Vihinen^{4,5} and Esa Uusipaikka²

Address: ¹Biotechnology and Food Research, MTT Agrifood Research Finland, Jokioinen, Finland, ²Department of Statistics, University of Turku, Turku, Finland, ³Department of Mathematics, University of Turku, Turku, Finland, ⁴Institute of Medical Technology, University of Tampere, Tampere, Finland, ⁵Research Unit, Tampere University Hospital, Tampere, Finland and ⁶Systems Biology Unit, Institut Pasteur, Paris, France

Email: Virpi Ahola^{*} - virpi.ahola@mtt.fi; Tero Aittokallio - tero.aittokallio@utu.fi; Mauno Vihinen - mauno.vihinen@uta.fi; Esa Uusipaikka - esa.uusipaikka@utu.fi

^{*} Corresponding author

Published: 03 November 2006

Received: 11 April 2006

BMC Bioinformatics 2006, 7:484 doi:10.1186/1471-2105-7-484

Accepted: 03 November 2006

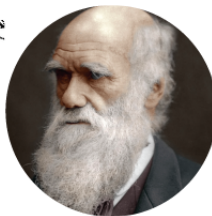
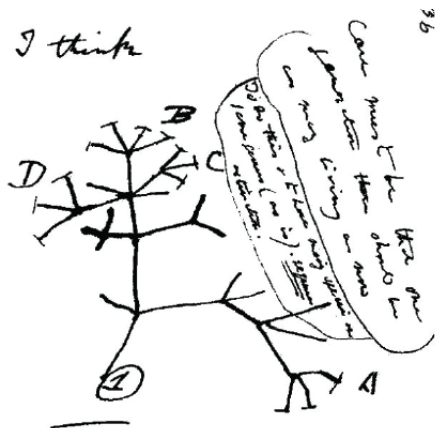
This article is available from: <http://www.biomedcentral.com/1471-2105/7/484>

© 2006 Ahola et al; licensee BioMed Central Ltd.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

What else? cont'd

- Evolutionary genomics: clustering concept, Bayesian framework, CTMC



Charles Darwin's sketches is considered to be the first metaphor of a tree to represent evolutionary relationships. (Image source: Wikimedia Commons.)

What does it take to be a statistical bioinformatician? cont'd

Softwares:

- R, Python
- Web resources

Further readings:

- Handbook of Statistical Genomics - Balding, 2019
- A Guide to QTL Mapping with R/qlt - Broman, 2009
- Statistical Contributions to Bioinformatics: Design, Modeling, Structure Learning, and Integration - Morris, 2017
- Orchestrating high-throughput genomic analysis with Bioconductor - Huber, 2014

More examples on how statistics contributes in solving biological problem

- 1 Phylogenetics tree for tracing the origin of SARS-CoV-2
- 2 Paired-end sequences alignment via pair-hidden Markov model

Phylogenetics tree for tracing the origin of SARS-CoV-2



Figure 17: SARS-CoV-2 has a shape like a dandelion, but ...

(Image source: CDC/Alissa Eckert & Dan Higgins)

Phylogenetics tree for SARS-CoV-2 cont'd



Figure 18: Covid-19's spread around the globe (<https://nextstrain.org/>)

Phylogenetics tree for SARS-CoV-2 cont'd



virus type ^

■ SARS-like CoV

 SARS-CoV

■ SARS-CoV-2

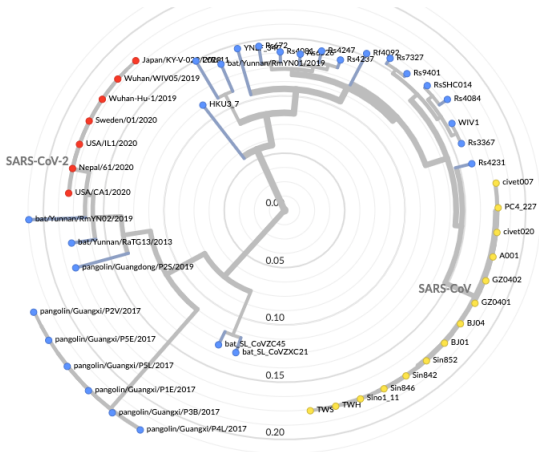


Figure 19: Covid-19's spread around the globe (<https://nextstrain.org/>)

Phylogenetics tree for SARS-CoV-2 cont'd

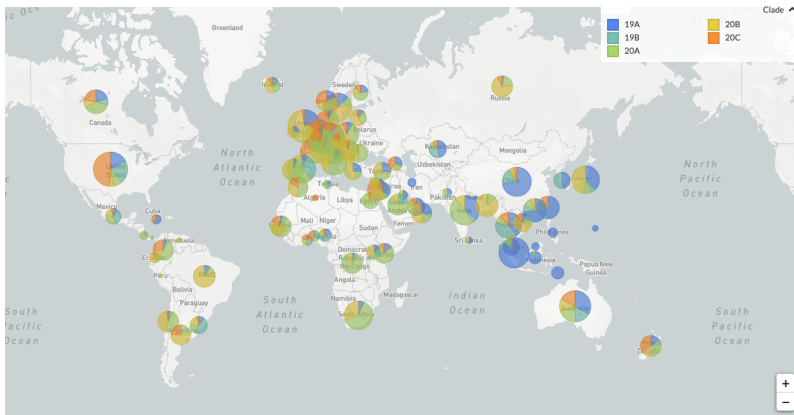


Figure 20: Covid-19's spread around the globe (<https://nextstrain.org/>)

Phylogenetics tree for SARS-CoV-2 cont'd

Phylogeny

Clade ^

19A

19B

20A

20B

20C

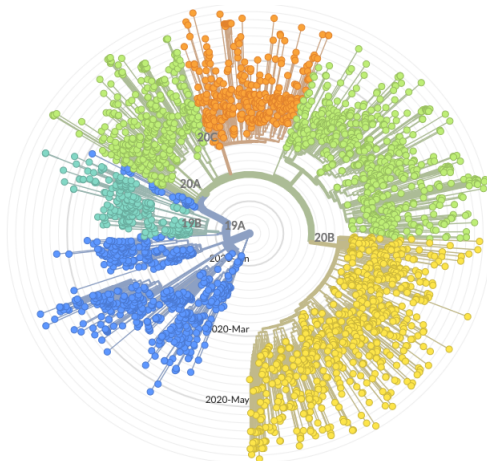


Figure 21: Covid-19's spread around the globe : radial phylogenetic tree
(<https://nextstrain.org/>)

Phylogenetics tree for SARS-CoV-2 cont'd

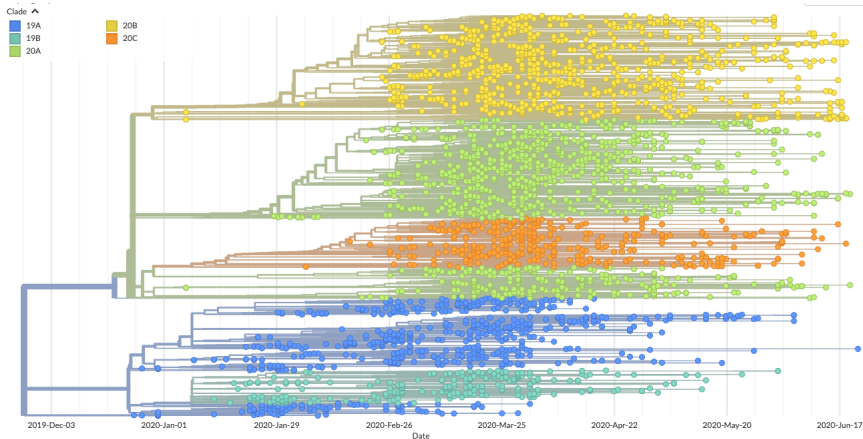


Figure 22: Covid-19's spread around the globe: rectangular phylogenetic tree (<https://nextstrain.org/>)

Phylogenetics tree for SARS-CoV-2: INA cases

Revisiting GISAID home page:

epicov.org/epi3/frontend#14bf4f

© 2008 - 2020 | Terms of Use | Privacy Notice | Contact

You are logged in as **Shofi Andari** - [logout](#)

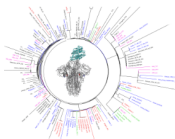
Registered Users EpiFlu™ EpiCoV™ My profile

EpiCoV™ Browse Downloads Upload My Unreleased

Pandemic coronavirus causing COVID-19










A previously unknown human coronavirus (hCoV-19) was first detected in late 2019 in patients in the City of Wuhan, who suffered from respiratory illnesses including atypical pneumonia, an illness that has become known as coronavirus disease (COVID-19). The coronavirus originated from an animal host and is closely related to the virus responsible for the Severe Acute Respiratory Syndrome coronavirus (SARS).

On 10. January 2020, the first virus genomes and associated data were publicly shared via GISAID. The World Health Organization announced on 11. March 2020 the first coronavirus pandemic. As the pandemic progresses, scientists from around the globe are tracking the virus and its genome sequences to ensure optimal virus diagnostic tests, to track and trace the ongoing outbreak and to identify potential intervention options. Several analyses to assist with these efforts are offered here, including sequence alignments, diagnostic primer and probe coordinates, 3D protein models, drug targets, phylogenetic trees and many more.



[Browse](#)

Analysis Update

 Full genome tree derived from all outbreak sequences (2020-07-04)	 Regional clade distribution of new sequences (2020-07-04)	 Common primer check for high quality genomes (2020-07-04)	 Receptor binding surveillance for complete genomes (2020-07-04)	 Full genome tree of hCoV-19-related precursors	 Spike glycoprotein receptor binding domain (comparison pangolin, bat, human)
 Spike Glycoprotein 3D Structure Model (comparison to SARS, bat)	 Potential drug targets highly conserved between hCoV-19 and SARS	 analysis update.pdf (2 MB)			

Important note: In the [GISAID EpiFlu™ Database Access Agreement](#), you have accepted certain terms and conditions for viewing and using data regarding influenza viruses. To the extent the Database contains data relating to non-influenza viruses, the viewing and use of these data is subject to the same terms and conditions, and by viewing or using such data you agree to be bound by the terms of the [GISAID EpiFlu™ Database Access Agreement](#) in respect of such data in the same manner as if they were data relating to influenza viruses.

Phylogenetics tree for SARS-CoV-2: INA cases cont'd

Samples from Indonesia (26 submissions)

epicov.org/epi3/frontend#1e7175

GISAID

© 2008 - 2020 | Terms of Use | Privacy Notice | Contact

You are logged in as Shofi Andari - logout

Registered Users EpiFlu™ EpiCoV™ My profile

EpiCoV™ Browse Downloads Upload My Unreleased

Search

Accession ID Virus name ☐ complete ☐ high coverage ☐
 Location Host ☐ low coverage excl ☐ w/Patient status ☐
 Collection date To Submission date To

<input type="checkbox"/>	Virus name	Passage de	Accession ID	Collection da	Submission C	Length	Host	Location	Originating lab
<input type="checkbox"/>	hCoV-19/Indonesia/JKT-EIJK07/2020	Original	EPI_ISL_467376	2020-04-24	2020-06-13	29,842	Human	Asia / Indonesia	RSUP Fatm
<input type="checkbox"/>	hCoV-19/Indonesia/MND-EIJK06/2020	Original	EPI_ISL_467375	2020-03-23	2020-06-13	29,767	Human	Asia / Indonesia	RSUP Prof. I
<input type="checkbox"/>	hCoV-19/Indonesia/SMR-EIJK05/2020	Original	EPI_ISL_467374	2020-03-18	2020-06-13	29,864	Human	Asia / Indonesia	Dinkes Sam
<input type="checkbox"/>	hCoV-19/Indonesia/E-JITD3101NT/2020	Original	EPI_ISL_458083	2020-04-11	2020-06-03	29,903	Human	Asia / Indonesia	Adi Husada I
<input type="checkbox"/>	hCoV-19/Indonesia/E-JITD2766NT/2020	Original	EPI_ISL_458082	2020-04-09	2020-06-03	29,903	Human	Asia / Indonesia	Universitas F
<input type="checkbox"/>	hCoV-19/Indonesia/E-JITD1273NT/2020	Original	EPI_ISL_458081	2020-03-30	2020-06-03	29,903	Human	Asia / Indonesia	RSUD Bangi
<input type="checkbox"/>	hCoV-19/Indonesia/E-JITD1238Sp/2020	Original	EPI_ISL_458079	2020-03-30	2020-06-03	29,903	Human	Asia / Indonesia	Mitra Kelan
<input type="checkbox"/>	hCoV-19/Indonesia/MRINUPH_02-461/2020	Original	EPI_ISL_438549	2020-03	2020-05-12	894	Human	Asia / Indonesia	Siloam Hosp
<input type="checkbox"/>	hCoV-19/Indonesia/MRINUPH_02-456/2020	Original	EPI_ISL_438548	2020-03	2020-05-12	900	Human	Asia / Indonesia	Siloam Hosp
<input type="checkbox"/>	hCoV-19/Indonesia/MRINUPH_01-461/2020	Original	EPI_ISL_438547	2020-03	2020-05-12	858	Human	Asia / Indonesia	Siloam Hosp
<input type="checkbox"/>	hCoV-19/Indonesia/MRINUPH_01-456/2020	Original	EPI_ISL_438546	2020-03	2020-05-12	850	Human	Asia / Indonesia	Siloam Hosp
<input type="checkbox"/>	hCoV-19/Indonesia/JKT-EIJK04/2020	Original	EPI_ISL_437192	2020-04-01	2020-05-08	29,903	Human	Asia / Indonesia	Mitra Kelan
<input type="checkbox"/>	hCoV-19/Indonesia/JKT-EIJK03/2020	Original	EPI_ISL_437191	2020-03-27	2020-05-08	29,903	Human	Asia / Indonesia	RS Pondok I
<input type="checkbox"/>	hCoV-19/Indonesia/JKT-EIJK02/2020	Original	EPI_ISL_437190	2020-03-26	2020-05-08	29,903	Human	Asia / Indonesia	RS Pondok I

Total: 26 viruses

<< first < prev 1 next > last >>

4 labs submitting SARS-CoV-2 from Indonesia: Eijkman Institute, Litbangkes RI, Mochtar Riady Institute (UPH), and ITD UNAIR.

Phylogenetics tree for SARS-CoV-2: INA cases cont'd

Details of the first row:

Virus detail	
Virus name:	hCoV-19/Indonesia/JKT-EIJK07/2020
Accession ID:	EPI_ISL_467376
Type:	betacoronavirus
Lineage (<i>GISAI</i> D Clade):	B (L)
Passage details/history:	Original
Sample information	
Collection date:	2020-04-24
Location:	Asia / Indonesia / Jakarta
Host:	Human
Additional location information:	
Gender:	Male
Patient age:	74
Patient status:	Hospitalized
Specimen source:	Nasopharyngeal and Oro-pharyngeal swab
Additional host information:	
Outbreak:	
Last vaccinated:	
Treatment:	
Sequencing technology:	Illumina MiSeq
Assembly method:	Bowtie2 + SPAdes
Coverage:	860x
Comment:	
Institute information	
Originating lab:	RSUP Fatmawati
Address:	Jl. RS. Fatmawati Raya No.4,Kota Jakarta Selatan, Daerah Khusus Ibukota Jakarta 12430, Indonesia
Sample ID given by the sample provider:	
Submitting lab:	Eijkman Institute for Molecular Biology, Ministry of Research and Technology/National Agency for Research and Innovation
Address:	Jalan Diponegoro 69, Jakarta 10430, Indonesia

Phylogenetics tree for SARS-CoV-2: INA cases cont'd

Continuing the details of the first row:

Sample ID given by the submitting laboratory:	EIJK0007
Authors:	Edison Johar, Frilasita A Yudhaputri, Hidayat Trimarsanto, David H Muljono, Safarina G Malik, Khin Saw Myint, Amin Soebandrio
Submitter Information	
Submitter:	Soebandrio, Amin
Submission Date:	2020-06-13
Address:	Jl. Diponegoro 69 10430 Jakarta

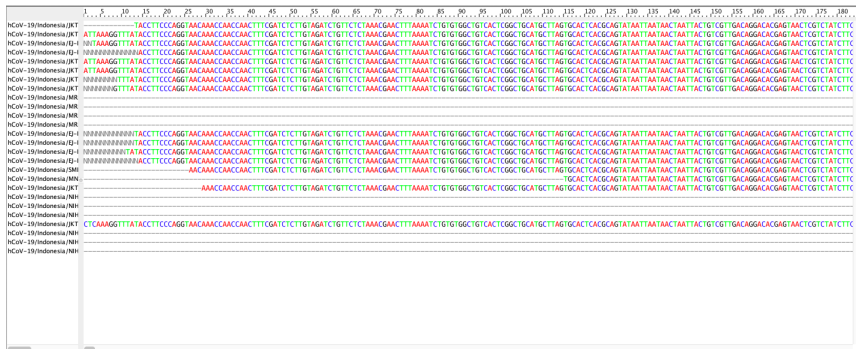
FASTA

```
>hCoV-19/Indonesia/JKT-EIJK07/2020|EPI_ISL_467376|2020-04-24
AAACCAACCACTTTGCATCTCTGTAGATCTGTTCTCAAACGAACCTTTAAATCTGTGTGGCTGTCACCTCGGCTGCAT
GCTTGTAGTCACTCAGCAGATAAATAAATACTAATTAAGTCTGCTGACAGGACACGAGTAACCTGCTATCTCTGCAG
GCTGCTTACGGTTTGTGTCGCTGTGACGCGCATCATCAGCAGCATCTAGGTTTGTGTCGGGTGTGACGGAAGGTAAGATG
GAGAGCTTTGTCCTGGTTTCAACGAGAAAAACACAGTCCAACTCAGTTTGCTGTTTACAGGTTCCGACAGTGTCTGT
ACGTGCTTTGGAGACTCGGTGGAGGAGGCTTATCAGAGGCACGTCAACATCTTAAAGATGGCAGCTTGTGGCTTAGTAG
AAGTTGAAAAAGCGGTTTGGCTCAACTTGAACAGCCCTATGTGTTTATCAAAACGTTCCGATGCTCGAACTGCACCTCAT
GGTCATGTTATGTTGAGCTGTGACGAGAACTCGAAGGACCTTCACTACGGTCTGAGTGGTGAGACATCTGGTGTCTCTGT
CCCTCATGTGGCGGAATACCAAGTGGCTTACCGAAGGTTCTTCTCTGTAAGAACGGTAATAAAGAGAGCTGGTGGCCATA
GTTACGGCGCCGATCTAAAGTCATTGTGACTTAGGCGACGAGCTTGGCACTGATCCTTATGAAGATTTTCAAGAAAACTGG
AACATAAATAGATCAGTGGTGTATCCCGTGAACCTCATGCTGAGCTTAAACGAGGGGCATACACTCGCTATGTGATATA
CAACTCTGTGGGCTGATGGCTACCTCTTGTAGTGCAATTAAGACCTTCTAGCAGCTGCTGGTAAGGCTTCATGACCTT
TGTCCCAACAACTGACCTTATTAACACTTAGAGGGGTGTATACCTGCTGCGGTGAACATGAGACATGAATTCCTGGTAC
ACCGGAAGCTTCTGAAAGAGCTATGAATTTGACAGACACTTTGAATAATTAATTTGGCAAGAAATTTGACACCTTCAATGG
GGAATGTCCAAATTTGTATTTTCCCTTAAATTCATTAATCAAGACTTCAACCAAGGTTTGAAGAAAAAGCTTGTATG
GCTTTATGGGTGAAATTCATCTGTCTATCCAGTTGCGCTACCAAAATGAATGCAACAAATGTGCTTTCAACTCTCATG
AAGTGTGATCATTTGGGTGAACTTTCATGGCAGACGGGGGATTTTGTAAAGCCACTTGCAGAAATTTGTGGCAGCTGAGAA
TTTGACTAAGAAGGTGCCACTACTTGTGGTTACTTACCCAAATATGCTGTTGTTAAATTTATTGTCCAGCATGTACATA
ATTCAAGAGTAGGACCTGAGCATAGTCTTGGCGAATACCAATGAATCTGGCTTGAACACCATTTCTGTAAGGGTGGT
CGCATATTTGCTTTGGAGGCTGTGTGTTCTTATGTTGTTGCCATAACAAGTGTGCTTATTTGGGTCCACGCTGAG
CGCTAACATAGGTTGTAAACATACAGGTGTTGTTGGAGAAGGTCGGAAGGCTTAAATGACAACTCTTGTGAATACTCC
AAAAAGAGAGAGTCAACATATTTGTTGGTGACTTTAACTTAATGAAGAGATCGCCATTATTTTGGCATCTTTTCT
GCTTCCACAGTGCTTTTGTGGAACACTTGAAGAGGTTTGGATTATAAGGCACTCAAAACAAATTTGTGAATCTGTGGTAA
TTTTAAGGTTACAAAAGGAAAGCTTGAAGAAAGTGGCTTGAATATTTGGTGAACAGAAATCAATACTGACTTCTTTATG
CATTTGCTATCAAGAGCTGTGTTACAGCAATTTTCTCCGCACTTGTGAACCTGCTGAACTGCTGAACTGCTGAACTG
TTACAGAAGGCCGCTATAACAACTAGATGGAATTTACACATATCTACTGAGACTTATGATGCTATGATGTTCAACAT
TGATTTGGCTACTAACAACTAGTTGTTAATGGCTTACATTACAGTGGTGTGTTTCAAGTTGACTTCCGAGTGGCTAACTA
ACATCTTTGGCACTGTTTATGAAAACTCAAAACCGTCTTGTATGCTTGAAGAGAAAGTTTAAAGAGGTTAGAGATTT
CTTAGAGAGCTTGGGAAAGTTTAAATTTATCTCAACCTGTGCTTGTGAAATTTGCGGTGGACAAATTTGTCACTGTGC
AAGGAAATTAAGGAGAGTGTTCAGACATTTCTTAAGCTTGAATAAATTTTGTGCTTGTGTGCTGACTCTATCATTA
TTAGTGGAGCTTAACTTAAAGCTTGAATTTAGGTGAACATTTGTGCACGCACTCAAGGGAATTTACAGAAAGTGTGTT
AAATTCAGAGAGAAATGTGCTACTCATGCTCTTAAAGGCCCAAAAGAAATTTATCTTCTAGAGGGAGAAACACTTCC
CACAGAGTGTGTTACAGAGGAAGTTGTCTTGAACCTGGTGATTTACAACCATTAAGAACACTTACTAGTGAAGCTGTTG
AAGTCCCATTTGGTTGGTACACAGCTTGTATTAACGGGCTTATGTTGCTCGAAATCAAGACACAGAAAGTACTGTGCC
CTTGCACCTTAATATGATGGTGAACAACAACTTACACACTAAAGGCGGTGACCAACAAAGGTTACTTTTGGTGATGA
CACTGTGATAGAAGTGAAGGTTACAAGAGTGAATATCATTGTAACCTTGTATGAAGAGGATGATAAAGTACTTAATG
```

[Back](#)[Contact Submitter](#)[Download Metadata](#)[Download FASTA](#)

Phylogenetics tree for SARS-CoV-2: INA cases cont'd

Multiple sequence alignment:



Phylogenetics tree for SARS-CoV-2: INA cases cont'd

Choosing model and distance matrix:

- Jukes-Cantor (JC69): equal base frequencies, all substitutions equally likely
- Felsenstein (F81): variable base frequencies, all substitutions equally likely
- Hasegawa-Kishino-Yano (HKY): variable base frequencies, one transition rate and one transversion rate
- Kimura-2params (K80): equal base frequencies, one transition rate and one transversion rate
- General time reversible (GTR): variable base frequencies, symmetrical substitution matrix

More on substitution models: <http://evomics.org/resources/substitution-models/nucleotide-substitution-models/>

Phylogenetics tree for SARS-CoV-2: INA cases cont'd

Choosing model and distance matrix:

- Jukes-Cantor (JC69): equal base frequencies, all substitutions equally likely
- Felsenstein (F81): variable base frequencies, all substitutions equally likely
- Hasegawa-Kishino-Yano (HKY): variable base frequencies, one transition rate and one transversion rate
- Kimura-2params (K80): equal base frequencies, one transition rate and one transversion rate
- General time reversible (GTR): variable base frequencies, symmetrical substitution matrix

+ Gamma dist (G) / proportion of invariable sites (I): describing rate variation among sites in a sequence.

Phylogenetics tree for SARS-CoV-2: INA cases cont'd

Choosing model and distance matrix on R (packages: phylogram, phangorn, seqinr)

```
> modelTest(cov_phyDat)
negative edges length changed to 0!
[1] "JC+I"
[1] "JC+G"
[1] "JC+G+I"
[1] "F81+I"
[1] "F81+G"
[1] "F81+G+I"
[1] "K80+I"
[1] "K80+G"
[1] "K80+G+I"
[1] "HKY+I"
[1] "HKY+G"
[1] "HKY+G+I"
[1] "SYM+I"
[1] "SYM+G"
[1] "SYM+G+I"
[1] "GTR+I"
[1] "GTR+G"
[1] "GTR+G+I"
```

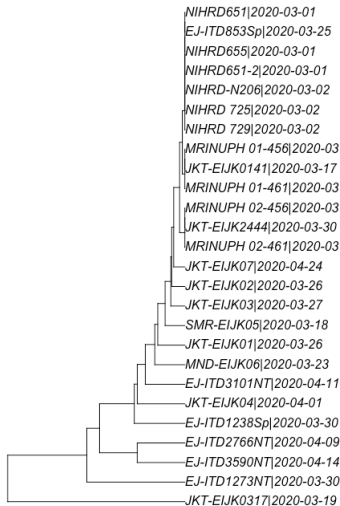

Phylogenetics tree for SARS-CoV-2: INA cases cont'd

Choosing model and distance matrix on R

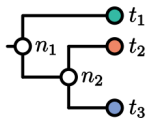
	Model	df	logLik	AIC	AICw	AICc	AICcw	BIC
1	JC	49	-42308.54	84715.08	0.000000e+00	84715.25	0.000000e+00	85122.10
2	JC+I	50	-42303.53	84707.06	0.000000e+00	84707.23	0.000000e+00	85122.38
3	JC+G	50	-42306.04	84712.09	0.000000e+00	84712.26	0.000000e+00	85127.41
4	JC+G+I	51	-42303.54	84709.08	0.000000e+00	84709.26	0.000000e+00	85132.70
5	F81	52	-41406.31	82916.63	4.408795e-05	82916.81	4.504991e-05	83348.56
6	F81+I	53	-41401.48	82908.95	2.049250e-03	82909.14	2.086538e-03	83349.19
7	F81+G	53	-41403.83	82913.66	1.940636e-04	82913.86	1.975948e-04	83353.90
8	F81+G+I	54	-41401.49	82910.97	7.460532e-04	82911.17	7.568843e-04	83359.52
9	K80	50	-42304.75	84709.51	0.000000e+00	84709.68	0.000000e+00	85124.83
10	K80+I	51	-42299.67	84701.34	0.000000e+00	84701.52	0.000000e+00	85124.97
11	K80+G	51	-42302.25	84706.50	0.000000e+00	84706.68	0.000000e+00	85130.12
12	K80+G+I	52	-42299.68	84703.37	0.000000e+00	84703.55	0.000000e+00	85135.30
13	HKY	53	-41401.76	82909.53	1.536213e-03	82909.72	1.564166e-03	83349.77
14	HKY+I	54	-41396.92	82901.85	7.137944e-02	82902.05	7.241571e-02	83350.39
15	HKY+G	54	-41399.28	82906.56	6.762151e-03	82906.76	6.860322e-03	83355.11
16	HKY+G+I	55	-41396.94	82903.87	2.592739e-02	82904.08	2.620702e-02	83360.73
17	SYM	54	-42296.71	84701.42	0.000000e+00	84701.61	0.000000e+00	85149.96
18	SYM+I	55	-42290.76	84691.52	0.000000e+00	84691.73	0.000000e+00	85148.37
19	SYM+G	55	-42294.14	84698.29	0.000000e+00	84698.50	0.000000e+00	85155.14
20	SYM+G+I	56	-42290.79	84693.58	0.000000e+00	84693.79	0.000000e+00	85158.74
21	GTR	57	-41395.94	82905.89	9.466819e-03	82906.11	9.496713e-03	83379.35
22	GTR+I	58	-41390.77	82897.54	6.167872e-01	82897.77	6.163339e-01	83379.31
23	GTR+G	58	-41393.43	82902.87	4.292927e-02	82903.10	4.289772e-02	83384.64
24	GTR+G+I	59	-41390.79	82899.58	2.221780e-01	82899.82	2.211384e-01	83389.66

Phylogenetics tree for SARS-CoV-2: INA cases cont'd

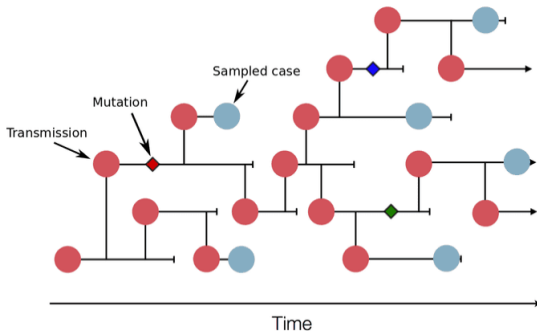
The tree:



Phylogenetics tree for SARS-CoV-2: the tree, the meaning

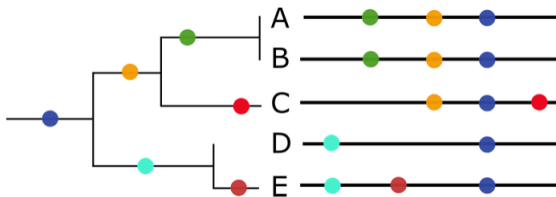


A rooted tree with 3 tips and 2 nodes
(Hall & Colijn, 2019: doi:10.1093/molbev/msz058)



<https://nextstrain.org/help/general/how-to-read-a-tree>

Phylogenetics tree for SARS-CoV-2: the tree, the meaning cont'd



<https://nextstrain.org/help/general/how-to-read-a-tree>

Studying the variability among the sequences (e.g., due to random mutations) \Rightarrow

- tracking the spread of the pathogen,
- learning its transmission routes and dynamics.

Paired-end sequences merging via pair-Hidden markov model

Illumina's paired-end reads

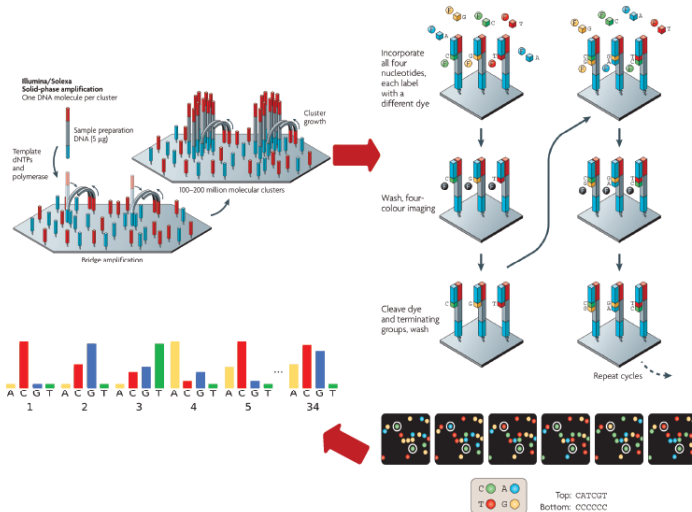


Figure 23: DNA sequencing: amplification, cycle, and base-calling (Whiteford, 2009; Metzker, 2009)

Illumina's paired-end reads cont'd

- All Illumina NGS systems are capable of paired-end sequencing

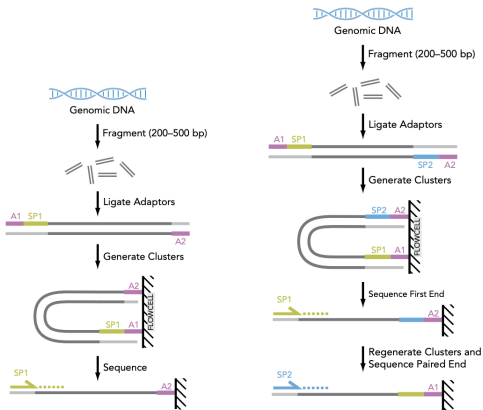


Figure 24: Illumina's single-end vs. paired-end sequencing, ©Illumina

Merging PE reads: why it matters?

- Merging PE reads can **substantially improve** various subsequent bioinformatics processes, including genome assembly, binning, mapping, annotation, and clustering for taxonomic analysis (Bushnell, 2017).

Merging PE reads: why it matters?

- Merging PE reads can **substantially improve** various subsequent bioinformatics processes, including genome assembly, binning, mapping, annotation, and clustering for taxonomic analysis (Bushnell, 2017).
- Existing tools:
 - SHERA (Rodrigue et al. 2010)
 - FLASH (Magoc and Salzberg 2011)
 - COPE (Liu et al. 2012)
 - PANDAsseq (Masella et al. 2012)
 - PEAR (Zhang et al. 2014)
 - AdapterRemoval v2 (Schubert, Lindgreen, and Orlando 2016)
 - MeFiT (Parikh et al. 2016)
 - NGmerge (Gaspar 2018)

All of **these methods** either **ignore the quality scores** or **assume all nucleotides are equally likely**.

Merging PE reads: why it matters?

- Merging PE reads can **substantially improve** various subsequent bioinformatics processes, including genome assembly, binning, mapping, annotation, and clustering for taxonomic analysis (Bushnell, 2017).
- Existing tools:
 - SHERA (Rodrigue et al. 2010)
 - FLASH (Magoc and Salzberg 2011)
 - COPE (Liu et al. 2012)
 - PANDAsseq (Masella et al. 2012)
 - PEAR (Zhang et al. 2014)
 - AdapterRemoval v2 (Schubert, Lindgreen, and Orlando 2016)
 - MeFiT (Parikh et al. 2016)
 - NGmerge (Gaspar 2018)

All of **these methods** either **ignore the quality scores** or **assume all nucleotides are equally likely**.

- What we are seeking: a merging tool that could provide **accurate** merged sequences (and fast!)

Illumina's paired-end reads: finding overlap

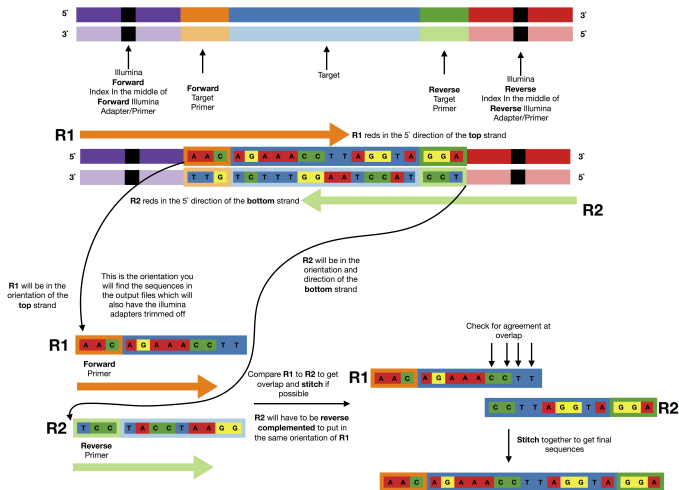


Figure 25: Illumina paired-end sequencing

(Source: https://seekdeep.brown.edu/illumina_paired_info.html)

Merging PE reads: an illustration

Obtaining a final (merged) sequence via alignment

forward read (R_1)	C	A	T	T	G	A	C	A
Q scores (q_1)	32	34	20	20	28	16	14	10

reverse read (R_2)	A	A	T	G	T	C	T	A
Q scores (q_2)	40	38	20	12	8	4	5	2

Quantifying Q-scores: [https:](https://support.illumina.com/help/BaseSpace_OLH_009008/Content/Source/Informatics/BS/QualityScoreEncoding_swBS.htm)

[//support.illumina.com/help/BaseSpace_OLH_009008/Content/Source/Informatics/BS/QualityScoreEncoding_swBS.htm](https://support.illumina.com/help/BaseSpace_OLH_009008/Content/Source/Informatics/BS/QualityScoreEncoding_swBS.htm)

Merging PE reads: an illustration cont'd

Obtaining a final (merged) sequence via alignment

R_1	C	A	T	T	G	A	C	A		
q_1	32	34	20	20	28	16	14	10		
w_2			T	A	G	A	C	A	T	T
q_2			2	5	4	8	12	20	38	40
Final (consensus)	C	A	T	T	G	A	C	A	T	T
Posterior Q score	32	34	22	16	35	28	30	34	38	40

Should consider gaps in the alignment to represent insertions/deletions (indels).

Merging PE reads: an illustration cont'd

The thing about sequence alignment...

forward read (R_1)	C	A	T	T	G	A	C	A
Q scores (q_1)	32	34	20	20	28	16	14	10

reverse read (R_2)	A	A	T	T	G	T	C	A
Q scores (q_2)	40	38	24	10	8	5	6	2

Suppose the reverse read does not match as good as our previous illustration.

Merging PE reads: an illustration cont'd

The thing about sequence alignment...

R_1	C	A	T	T	G	A	C	A	—	—
W_2	—	—	T	G	A	C	A	A	T	T

Is it a good alignment?

Merging PE reads: an illustration cont'd

The thing about sequence alignment...

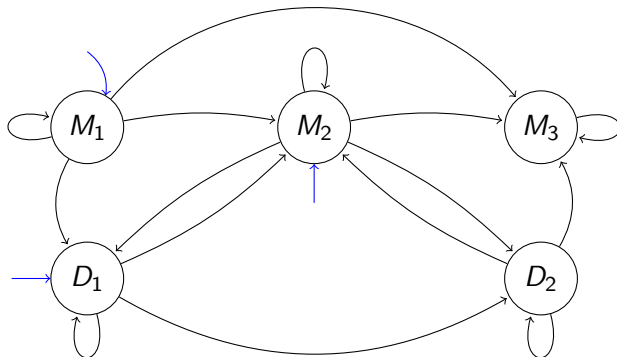
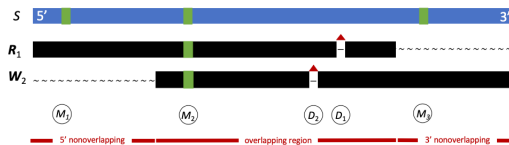
R_1	C	A	T	T	G	A	C	A	—	—	—
W_2	—	—	T	—	G	A	C	A	A	T	T

How about this one?

We introduce a gap after the first match.

It can be "costly", but now we have more matches.

PE merging via pair-HMM cont'd



Merging PE reads: pair-HMM approach

The blue fragment here is the reference.

Sadly, we do not always have the reference (genome). In fact, these FASTQ files are going to be used to rebuild the genome (genome assembly).

Nucleotide pairing in R_{1j} and W_{2j} = observable state

Nucleotide pairing in the alignment = hidden state

⇒ pair-hidden Markov model (Rabiner, 1989; Durbin, 1998).

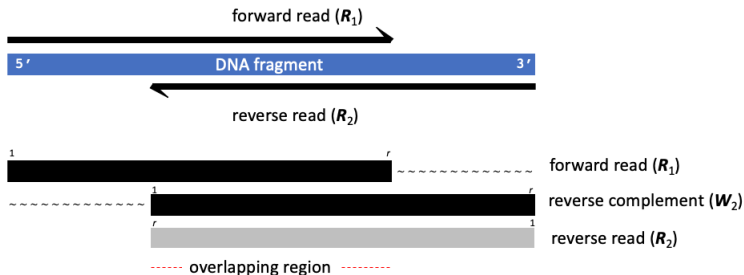


Figure 26: Alignment between forward and reverse read

Courtesy to Prof. Mardia (Dept. of Statistics, University of Leeds, UK) at the 2006 LASR Conference



"Statisticians need to be more open, more ready to learn *molecular biology*, more computationally aware, more ready to understand databanks, ...

But above all, **we always need great scientist friends!!**

This all is a part of solving great questions in life sciences of taming the nature and immortality, etc!"



ISCBacademy Webinar Series

ISCBacademy Webinar Series



Upcoming Webinars

Archived Webinars

Other ISCB Webinars



ISCBacademy Webinar Series

Welcome to the ISCBacademy Webinar Series. In conjunction with the [communities of special interest \(COSIs\)](#), select presentations are invited to give a live-streamed talk about their research. Access to the webinar series is complimentary for all. Non ISCB Members must register.

Be sure to check back regularly for information about upcoming webinars or to watch recordings of previous presentations.

Please use the links below to find more information or to register for an upcoming webinar:

- **July 21, 2020 at 11:00AM EDT, Pooled CRISPR screens with imaging on microRaft arrays reveals stress granule-regulatory factors** by Emily Wheeler, hosted by iRNA COSI and the RNA Society
- **July 30, 2020 at 9:00AM EDT, Southern African Human Population Structure - an Opportunity to Expand Genomics Research Worldwide** by Caitlin Uren, hosted by ASBCB
- **August 24, 2020 at 11:00AM EDT, Unravelling the mystery of orphan genes to understand the origins of genetic novelty** by Nikos Vekrellis, hosted by EvolCompGen and SMBE
- **September 30, 2020 at 11:00AM EDT, RAXML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference** by Alexey Kozlov, hosted by EvolCompGen COSI

<https://www.iscb.org/iscbacademy-webinars>



28TH CONFERENCE ON
**Intelligent Systems
for Molecular Biology**
JULY 13-16, 2020

ISMB 2020

Virtual Conference

REGISTRATION

JOIN ISCB

NEWS

KEY DATES

Home

GENERAL INFO

SUBMIT RESEARCH

ISMB 2020 is going virtual - July 13-16
Still Time to Register

Conference will be conducted in the Eastern Daylight Time Zone
(convert schedule to your time zone here)

<https://www.iscb.org/ismb2020-registration>

eccb2020.info



Contact us

**19th European Conference
on Computational Biology**

August 31st - September 8th, 2020 *Virtual!*

ABOUT

KEY DATES

REGISTRATION

PROGRAMME

SPONSORING

CALLS

NEW TRENDS IN BIOINFORMATICS BY ECCB

ECCB2020 GOES VIRTUAL

<https://eccb2020.info/>

A banner for the RECOMB/ISCB Conference on Regulatory & Systems Genomics with DREAM Challenges. The left side features a circular logo with 'RSG with DREAM 2020' and a network diagram. The right side has a blue background with the text 'VIRTUAL ISCB EVENT Nov 16 - 18, 2020' and 'Mark your calendars!'. Below the banner is an orange navigation bar with four buttons: 'CONTACT', 'JOIN ISCB', 'KEY DATES', and 'REGISTER'.


RECOMB/ISCB CONFERENCE on
REGULATORY & SYSTEMS GENOMICS
with DREAM CHALLENGES

VIRTUAL ISCB EVENT
Nov 16 - 18, 2020
Mark your calendars!

CONTACT JOIN ISCB KEY DATES REGISTER

[RSG with DREAM 2020 | November 16-18, 2020 | DREAM Submissions](https://www.iscb.org/recomb-regsysgen2020)

<https://www.iscb.org/recomb-regsysgen2020>

A banner for the ROCKY 2020 Bioinformatics Conference. The left side shows a logo with 'ROCKY 2020' and a mountain graphic. The right side has a dark background with a person in a red jacket and the text 'Aspen/Snowmass Colorado December 3 - 5, 2020' and 'Mark your calendars!'. Below the banner is a blue navigation bar with four buttons: 'JOIN ISCB', 'KEY DATES', 'REGISTER', and 'HOUSING'.

ROCKY 2020
Bioinformatics Conference

Aspen/Snowmass Colorado
December 3 - 5, 2020
Mark your calendars!

JOIN ISCB KEY DATES REGISTER HOUSING

[ROCKY 2020 | Dec 3 - 5, 2020 | Aspen/Snowmass, CO | HOME - ROCKY 2020](https://www.iscb.org/rocky2020)

<https://www.iscb.org/rocky2020>

GIW/ISCB-Asia 2020

Genome Informatics Workshop

To be announced | Tainan , Taiwan

Due to the pandemic of COVID-19, the conference will postpone to 2021.

<https://giw2020.ncku.edu.tw/>

© FEBRUARY 10, 2020

We are delighted to invite you to the 19th InCoB. The conference will be virtually hosted for the first time. InCoB 2020 took on the theme of **"Bioinformatics and the translation of data-driven discoveries"**, and will include presentations of original research results, discussions in plenary sessions, poster sessions, workshops, software demos and panel discussions related to the field of bioinformatics. This is a great opportunity for you to showcase your research!

It was originally planned to be held in Kunming, Yunnan, China and hosted by Kunming University of Science and Technology (KMUST), Kunming, China. However, due to the global Covid-19 crisis, it was decided that the conference will be held virtually. More details will be made available as soon as possible.



<https://www.apbionet.org/international-conference-on-bioinformatics-2020-incob2020/>

Terima kasih.

shofi.andari@statistika.its.ac.id
shfandari@gmail.com