Introduction to Bioinformatics

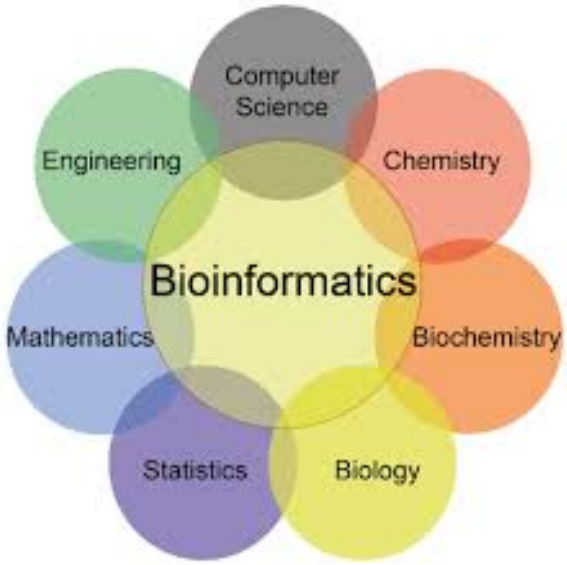
a Data Scientist perspective

Dina Machuve (PhD)

08 September 2018

What is Bioinformatics?

- The field of science in which biology, computer science, and information technology merge to form a single discipline (NCBI).
- The science of storing, retrieving and analysing large amounts of biological information (EMBL-EBI,2018)



- highly interdisciplinary field
- the ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned

Computational Biology

- The use of techniques from applied mathematics, informatics, statistics, and computer science to solve (typically noisy) biological problems
  - Multiple sequence alignment
  - Identification of functional regions or motifs
  - Classification of data
  - Phylogenetic analysis
  - Molecular structure determination and folding
  - Genetics
  - Diagnostics and medical applications
- Largely interchangable in literature with Bioinformatics

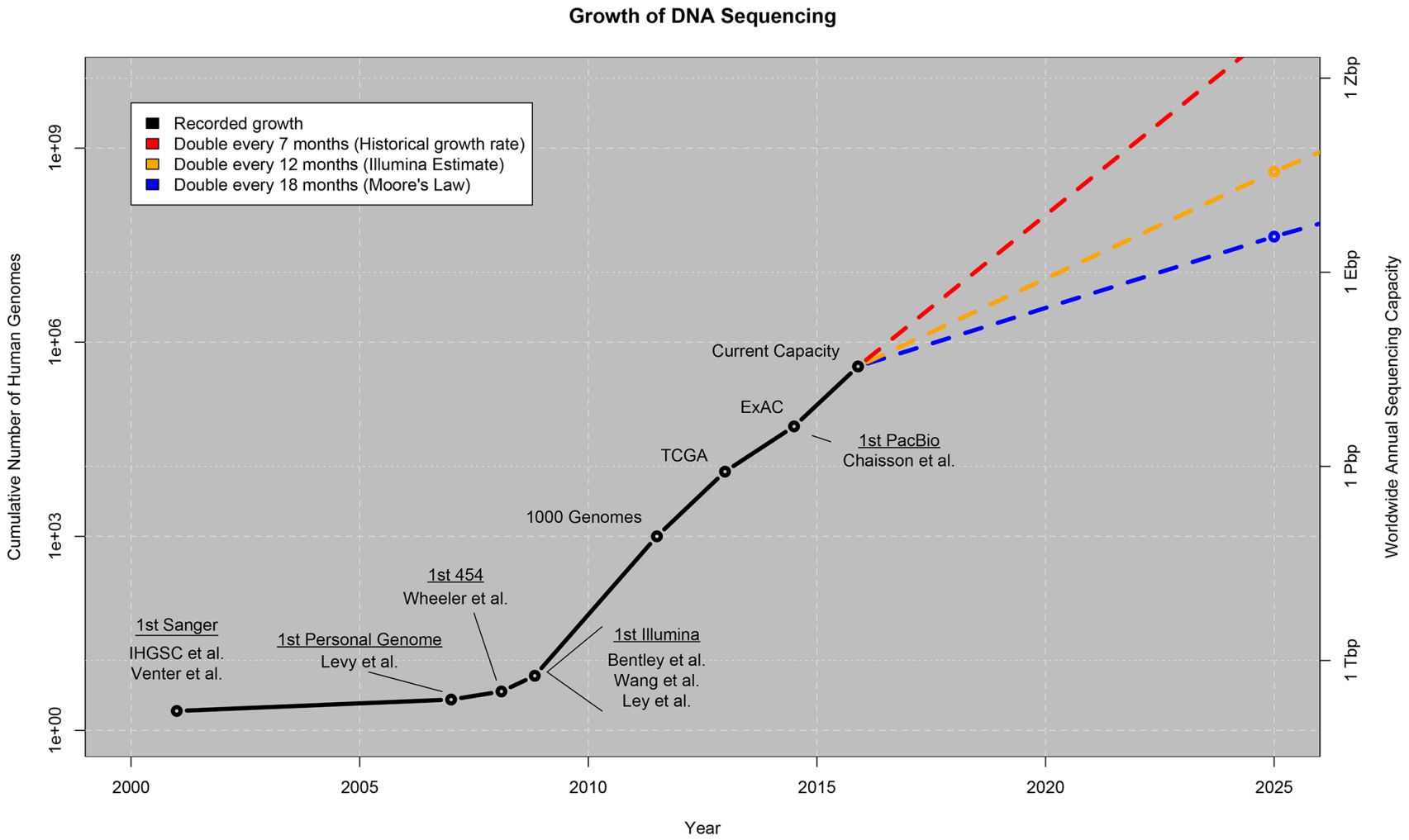
Bioinformatics - History

- Revolutionary methods in molecular biology
  - DNA sequencing
  - Protein structure determination
  - Drug design and development
- Exponential growth of biological information
- Computational requirement for
  - Database storage of information
  - Organization of information
  - Tools for analysis of data
- The transition of biology from 'wet-lab' to 'dry- lab/information science'

Why is bioinformatics important?

- Modeling via bioinformatics may provide answers to questions related to human health and evolution
- Novel discoveries for healthcare, agriculture, food security ($>$ 1.2 million species of plants & animals)
- Disease surveillance and response
- Management of health data (EHRs and experimental data) can inform diagnosis and treatment Precision Medicine
- Data can save lives!

DNA Sequence continues to grow



Big Data Analytics in Biology

BIOINFORMATICS

Big data versus the big C

The torrents of data flowing out of cancer research and treatment are yielding fresh insight into the disease.

BY NEIL SAVAGE

In 2013, geneticist Stephen Elledge answered a question that had puzzled cancer researchers for nearly 100 years. In 1914, German biologist Theodor Boveri suggested that the abnormal number of chromosomes — called aneuploidy — seen in cancers might drive the growth of tumours. For most of the next century, researchers made little progress on the matter. They knew that cancers often have extra or missing chromosomes or pieces of chromosomes, but they did not know whether this was important or simply a by-product of tumour growth — and they had no way of finding out.

OUTLOOK BIG DATA IN BIOMEDICINE

PERSPECTIVE

Sustaining the big-data ecosystem

Organizing and accessing biomedical big data will require quite different business models, say **Philip E. Bourne**, **Jon R. Lorsch** *and* **Eric D. Green**.

Biomedical big data offer tremendous potential for making discoveries, but the cost of sustaining these digital assets and the resources needed to make them useful have received relatively little attention. Research budgets are flat or declining in inflation-

recorded. All of this means that absolute numbers are hard to interpret.

These caveats notwithstanding, more details of data usage are needed to inform funding decisions. Over time, such usage patterns could tell us how best to target annotation and curation efforts, establish which data should receive the most attention and therefore incur the largest cost, and determine which data should be kept in the longer term. The cost of data regeneration can also influence decisions about keeping data.

Funders should encourage the development of new metrics to ascertain the usage and value of data, and persuade data resources to provide such statistics for all of the data they maintain. We can learn here from the private sector: understanding detailed data usage patterns through data analytics forms the basis of highly successful companies such as Amazon and Netflix.

FAIR AND EFFICIENT

TECHNOLOGY FEATURE

THE BIG CHALLENGES OF BIG DATA

As they grapple with increasingly large data sets, biologists and computer scientists uncork new bottlenecks.

EMBL-EBI

Extremely powerful computers are needed to help biologists to handle big-data traffic jams.

BY VIVIEN MARX

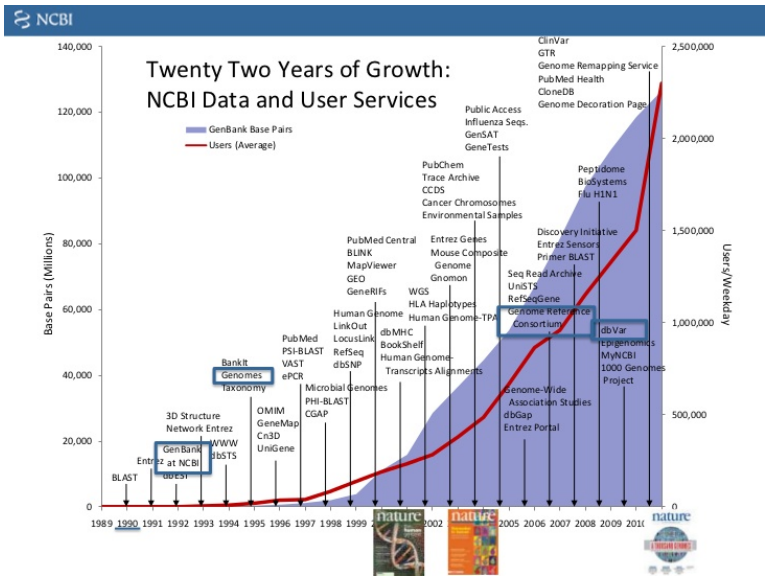
Biologists are joining the big-data club. With the advent of high-throughput genomics, life scientists are starting to grapple with massive data sets, encountering challenges with handling, processing and moving information that were once the domain of astronomers and high-energy physicists[1].

With every passing year, they turn more often to big data to probe everything from the regulation of genes and the evolution of genomes to why coastal algae bloom, what microbes dwell where in human body cavities and how the genetic make-up of different cancers influences how cancer patients fare[2]. The European Bioinformatics Institute (EBI) in Hinxton, UK, part of the European Molecular Biology Laboratory and one of the world's largest biology-data repositories, currently stores 20 petabytes (1 petabyte is 10^{15} bytes) of data and back-ups about genes, proteins and small molecules. Genomic data account for 2 petabytes of that, a number that more than doubles every year[3] (see 'Data explosion').

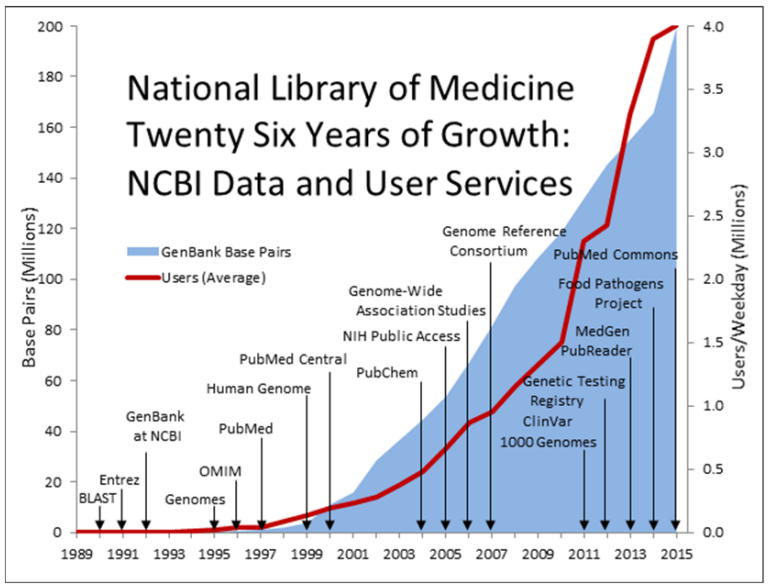
This data pile is just one-tenth the size of the data store at CERN, Europe's particle-physics laboratory near Geneva, Switzerland. Every

Biological data is big data

European Bioinformatics Institute (EMBL-EBI): total disk capacity 75 petabytes (Dec 2015)



In 2010



In 2015

Big data explosion

1.845e+16

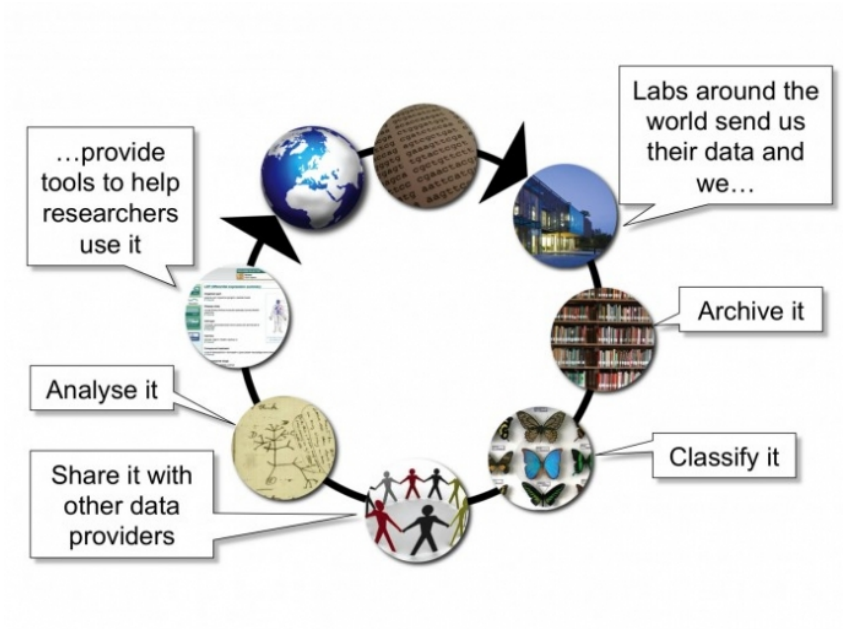
Number of publicly available bases in the NCBI Sequence Read Archive (SRA) as of July 1, 2018. This is the equivalent of 6,153,232 human genomes (which is 3e+9 bases).

30TB

Approximate amount of public sequence data received and processed ***daily*** by the NCBI Sequence Read Archive (SRA).

Bioinformatics Centres of Excellence

- A few in number worldwide with responsibility to collect, catalogue and provide open access to published biological data.
- Among these centers include:
  - The EMBL-European Bioinformatics Institute (EMBL-EBI)
  - The US National Center for Biotechnology Information (NCBI)
  - The National Institute of Genetics in Japan (NIG)



Challenges and Opportunities

Data Phase	Astronomy	Twitter	YouTube	Genomics
Acquisition	25 zetta-bytes/year	0.5–15 billion tweets/year	500–900 million hours/year	1 zetta-bases/year
Storage	1 EB/year	1–17 PB/year	1–2 EB/year	2–40 EB/year
Analysis	In situ data reduction	Topic and sentiment mining	Limited requirements	Heterogeneous data and analysis
	Real-time processing	Metadata analysis		Variant calling, ~2 trillion central processing unit (CPU) hours
	Massive volumes			All-pairs genome alignments, ~10,000 trillion CPU hours
Distribution	Dedicated lines from antennae to server (600 TB/s)	Small units of distribution	Major component of modern user's bandwidth (10 MB/s)	Many small (10 MB/s) and fewer massive (10 TB/s) data movement

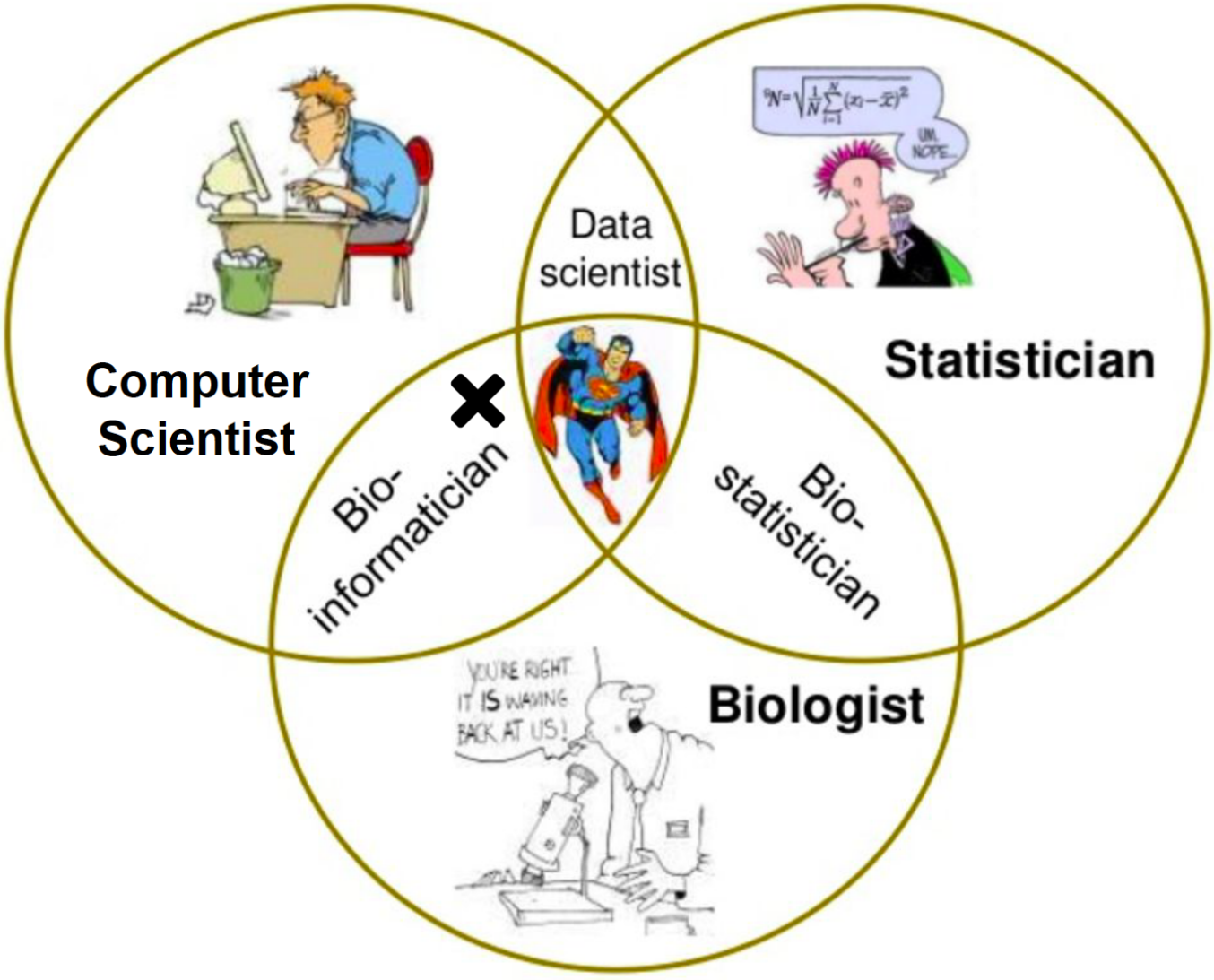
doi:10.1371/journal.pbio.1002195.t001

- Life sciences have become increasingly data driven
- Molecular life scientists work with bioinformatics experts to design, analyse and interpret their experiments (EMBL-EBI,2018).

Data Science challenges in Africa - **The opportunity!**

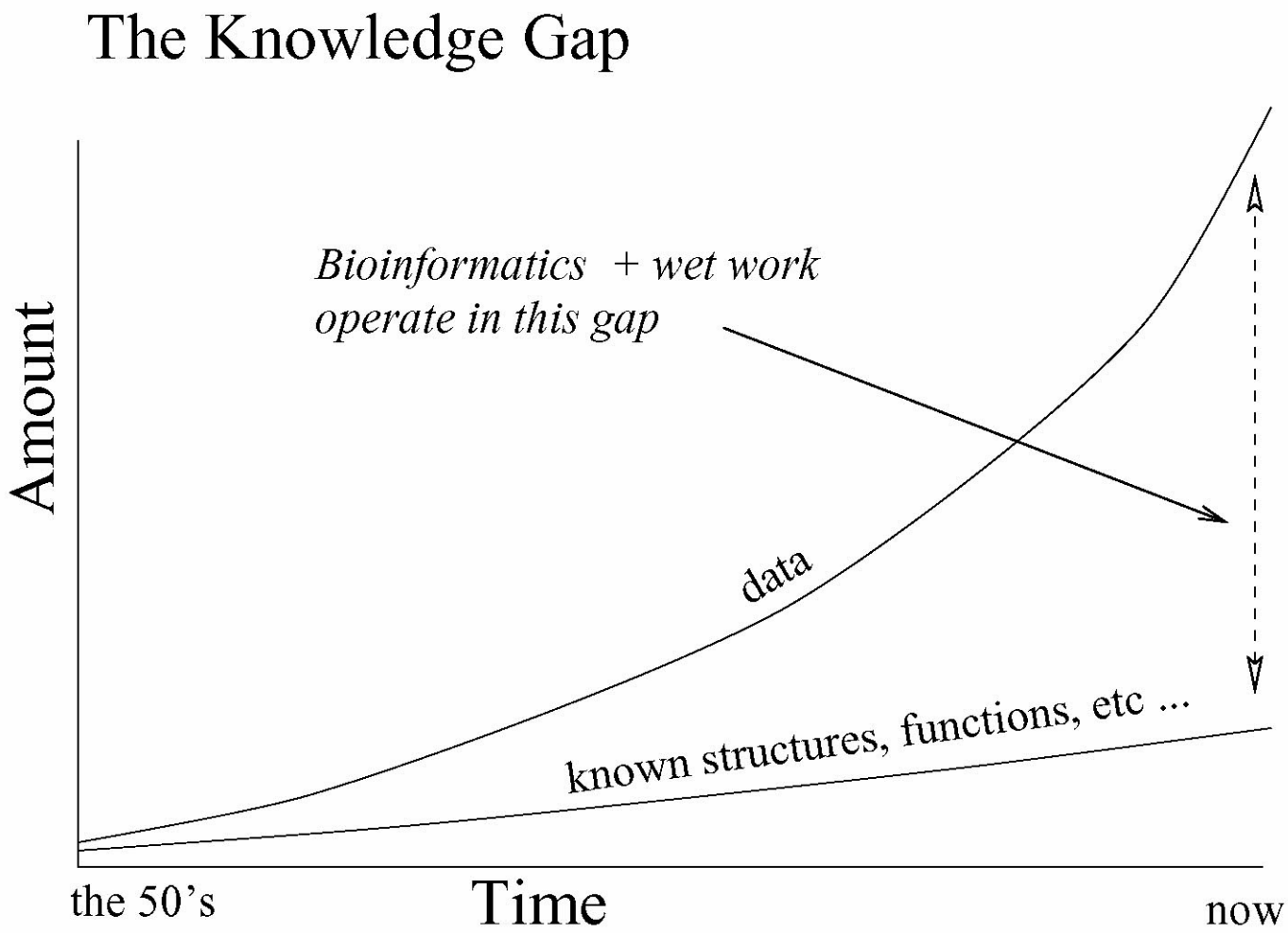
- Work on unique fauna and flora African data underrepresented
- Have to deal with ever-increasing data size and complexity
- IT challenges include:
  - Data transfer from generation site
  - Internet access
  - Adequate IT infrastructure for storage and processing
  - Long term secure storage
  - Training people at different levels on the use of this
- Not enough bioinformaticians or data scientists
- Meta data is not well curated, data quality and accuracy is not a high priority for clinicians and some researchers

The Data Scientist Challenge



Credit: Torsten Seemann

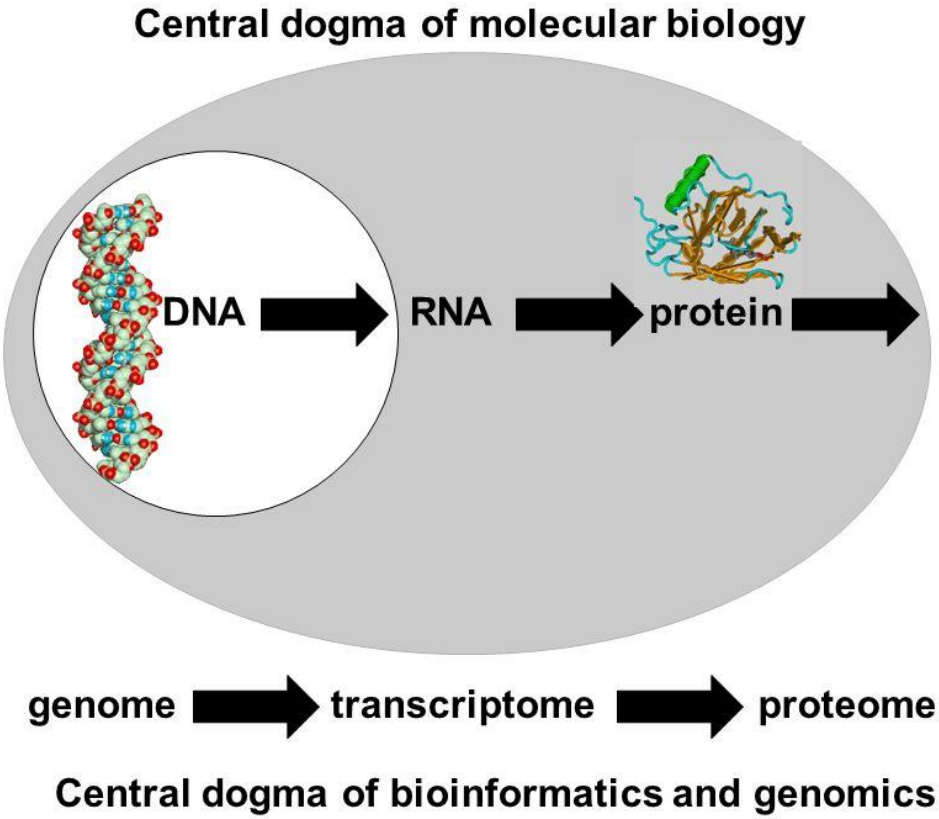
Bioinformatics



Generation of OMICS Data

NGS technologies

What is DNA?



- **DNA** is deoxyribonucleic acid that contains genetic information
- **RNA** is ribonucleic acid involved in protein synthesis and sometimes in transmission of genetic information

Approach

"Bioinformatics is the field of science in which **biology**, **computer science** *and* **infromation technology** *merge to form a single discipline"*

Example

1. Biological Question
2. Generate Data
3. Translate into a computer solvable task
4. Develop an algorithm
5. Implement algorithm
6. Run algorithm
7. Condense result in human readable form
8. Answer Biological Question

1. Genes regulated by protex X
2. ChIP-Seq data
3. "Align reads and identify clusters in the genome"
4. Choose data structures
5. Write source code
6. Align reads
7. Write script to summarize results genome wide
8. Report protein's binding sites

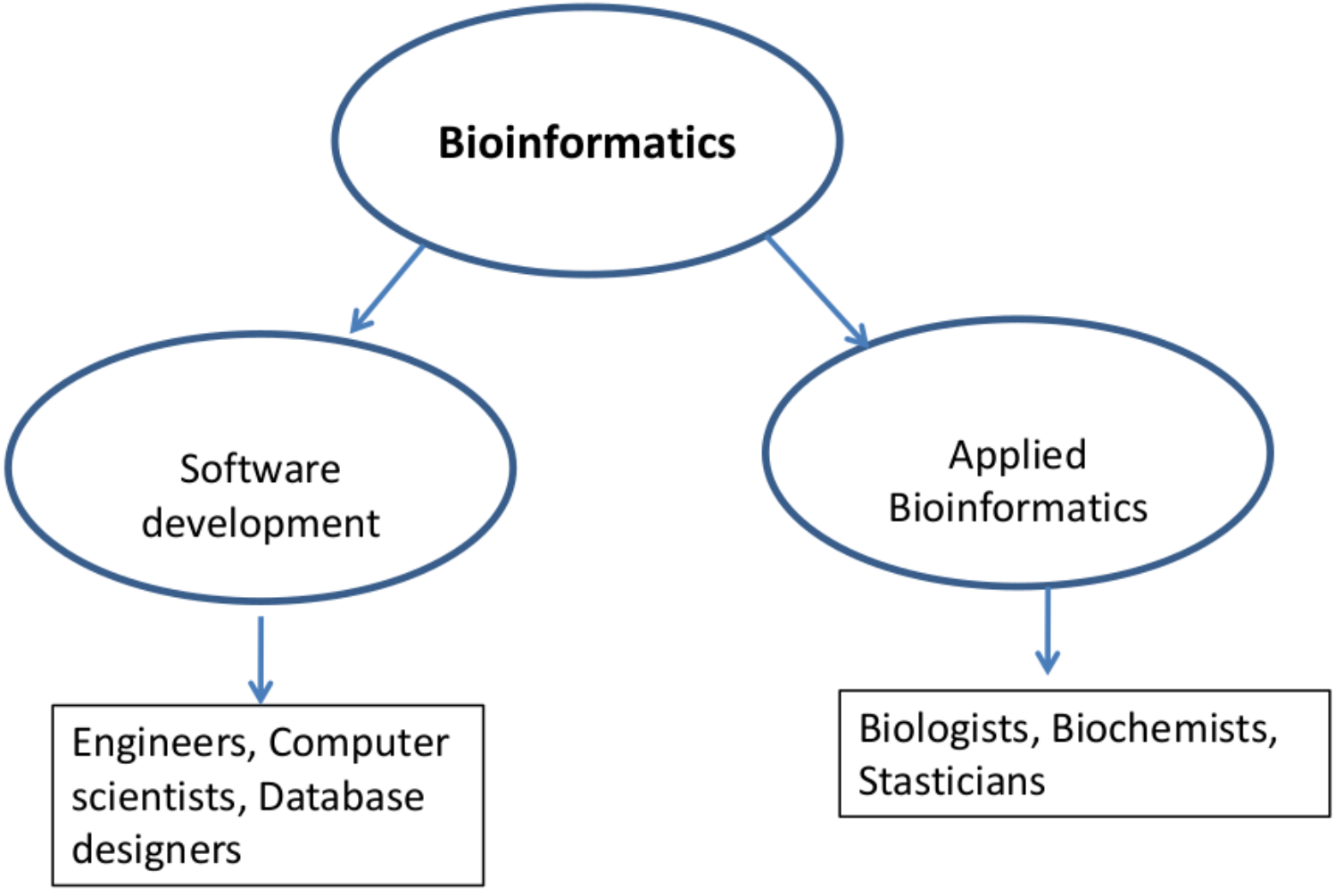
Three things to remember

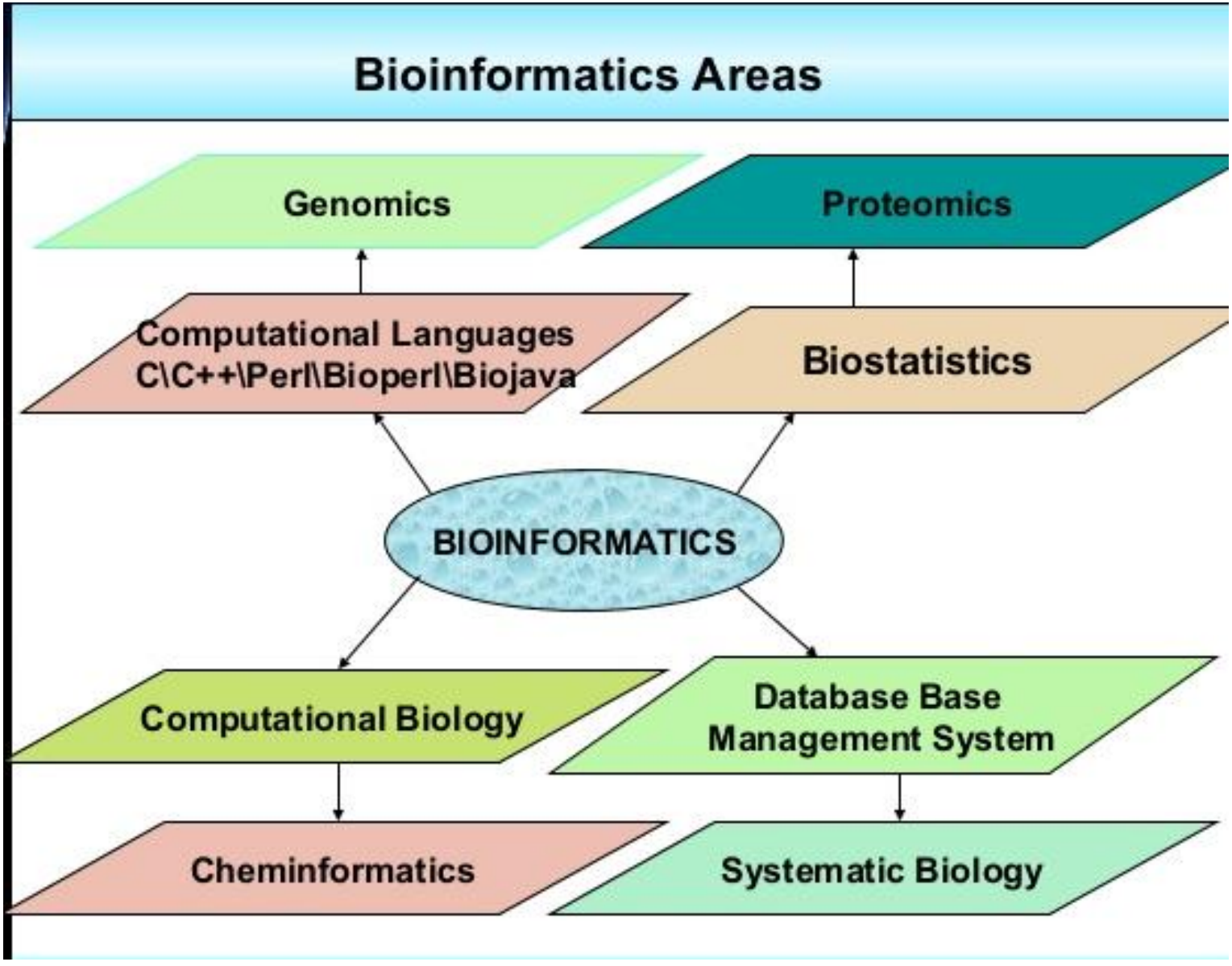
1) Bioinformatics requires dedication and continuity
2) Bioinformatics data analysis is a full research experiment in itself
3) We get the most out of our research if we work as a interdisciplinary research team throughout

Bioinformatics Scope

Bioinformatics consists of two subfields:

- The development of computational tools and databases
- The application of these tools and databases in generating biological knowledge to better understand living systems.


Bioinformatics
Software development
Applied Bioinformatics
Engineers, Computer scientists, Database designers
Biologists, Biochemists, Stasticians


Bioinformatics Areas
Genomics
Proteomics
Computational Languages
C\C++\Perl\Bioperl\Biojava
Biostatistics
BIOINFORMATICS
Computational Biology
Database Base
Management System
Cheminformatics
Systematic Biology

Why acquire bioinformatics skills?

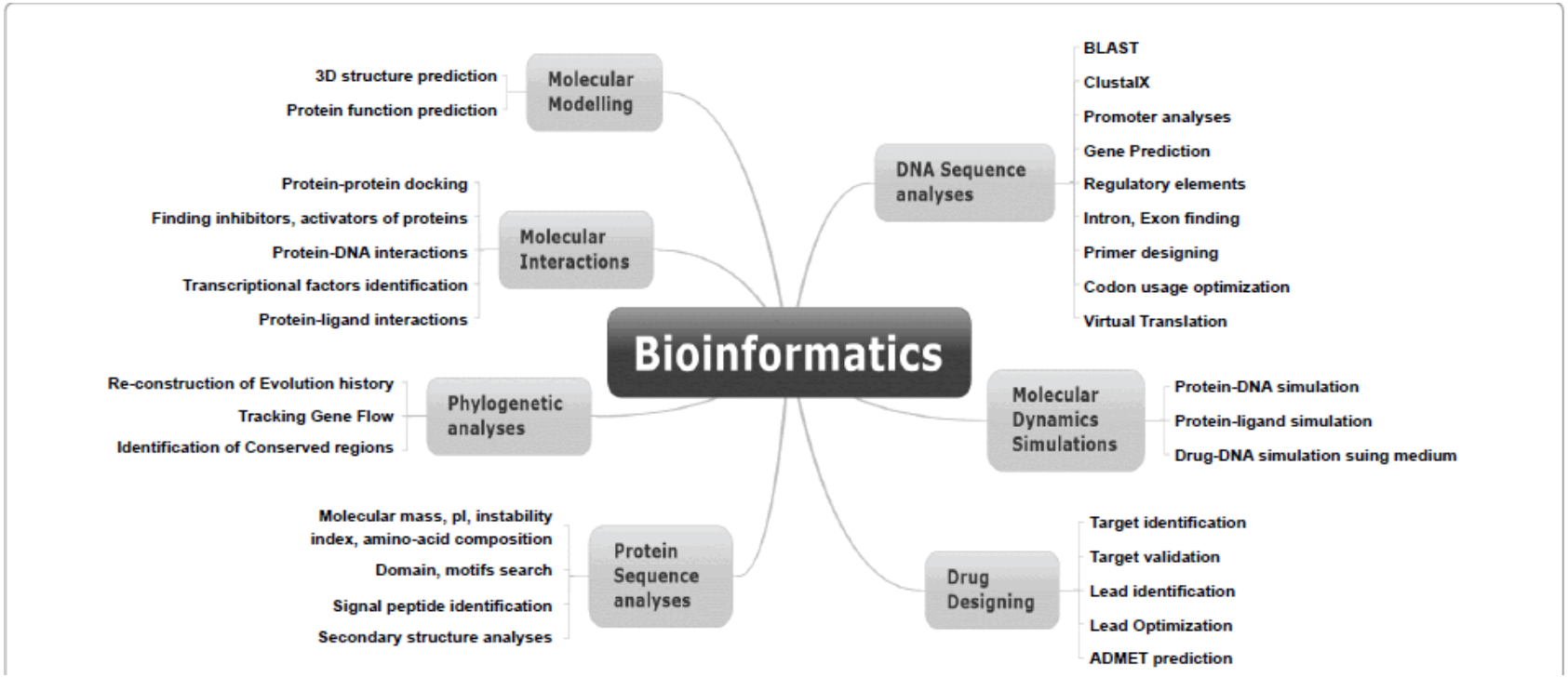
Learn how to:

- Store/retrieve biological information from databases
- Retrieve/compare gene sequences
- Predict function of unknown genes/proteins
- Search for previously known functions of a gene
- Compare data with other researchers
- Compile/distribute data for other researchers

Pre-requisite Skills for Bioinformatics

- Basic background in some aspects of molecular biology
- Ability to communicate biological questions comprehensibly to computer scientists
- Thorough comprehension of the problem in the bioinformatics field
- Statistics (association studies, clustering, sampling)
- Ability to filter, parse, and munge data and determine the relationships between the data sets
- Mathematics (e.g. algorithm development using ML or DL)
- Engineering (e.g. robotics)
- Good knowledge of a few molecular biology software packages (molecular modeling / sequence analysis)
- Command line computing environment (Linux/Unix knowledge)
- Database administration (SQL, Oracle etc)
- Computer Programming Skills/Experience (C/C++, Sybase, Java and Scripting)
- Programming Language Knowledge (Python)

Application of Bioinformatics Tools



Neil Lawrence in Computational Biology

- "Bridging the Gap Between Computational Biology and Systems Biology", at Pathologists Society Summer Meeting, Sheffield (2012)
- "Between Systems and Data-driven Modeling for Computational Biology: Target Identification with Gaussian Processes", at ABCD 2011, Ravenna, Italy
- "Gaussian Processes in Computational Biology Tutorial: Multioutput Gaussian Processes and Mechanistic Models", at BioPreDyn Workshop, CRG, Barcelona, Spain (2012)
- "An Introduction to Systems Biology from a Machine Learning Perspective", at TISE Summer School, Tampere, Finland (2009)

Ongoing Work

Development of a Sanger Sequence Automatic Assembly and Analysis Tool

- There have been a few attempts at building free, open source software to deal with the analysis of Sanger sequencing data
  - sangerseqR
  - sangeranalyseR
  - ASAP tool in Python
- To assess the usability of these various software
- To develop a software that provides users with the ability to analyze efficiently and automatically **large datasets** of alignable Sanger sequence data.

Where to start

Never too late to start to learn!

1. Brooksbank, C. and Cowley, A.(2018), Bioinformatics for the terrified, EMBL-EBI Train Online
2. Isaev, A. (2006) Introduction to Mathematical Models in Bioinformatics, Springer
3. An Introduction To Applied Bioinformatics Interactive lessons in bioinformatics
4. Mulder, N. (2017), "H3ABioNet enabling bioinformatics and big data research in Africa", CHPC Conference
5. https://github.com/crazyhottommy/getting-started-with-genomics-tools-and-resources
6. @DNAed_tech; @EANBiT_Project; @emblebi; @EBItraining; @GA4GH; @IDeAL_KEMRI_WT