Topics in Bioinformatics

Kristel Van Steen, PhD[2]

Montefiore Institute - Systems and Modeling

GIGA - Bioinformatics

ULg

kristel.vansteen@ulg.ac.be

Lecture 1: Setting the pace

1 Evolving trends in bioinformatics

Definition

Historical notes

Challenges

2 Careers in bioinformatics

Topics in bioinformatics from a journal's perspective

Becoming a bioinformatician

3 Gen-omics

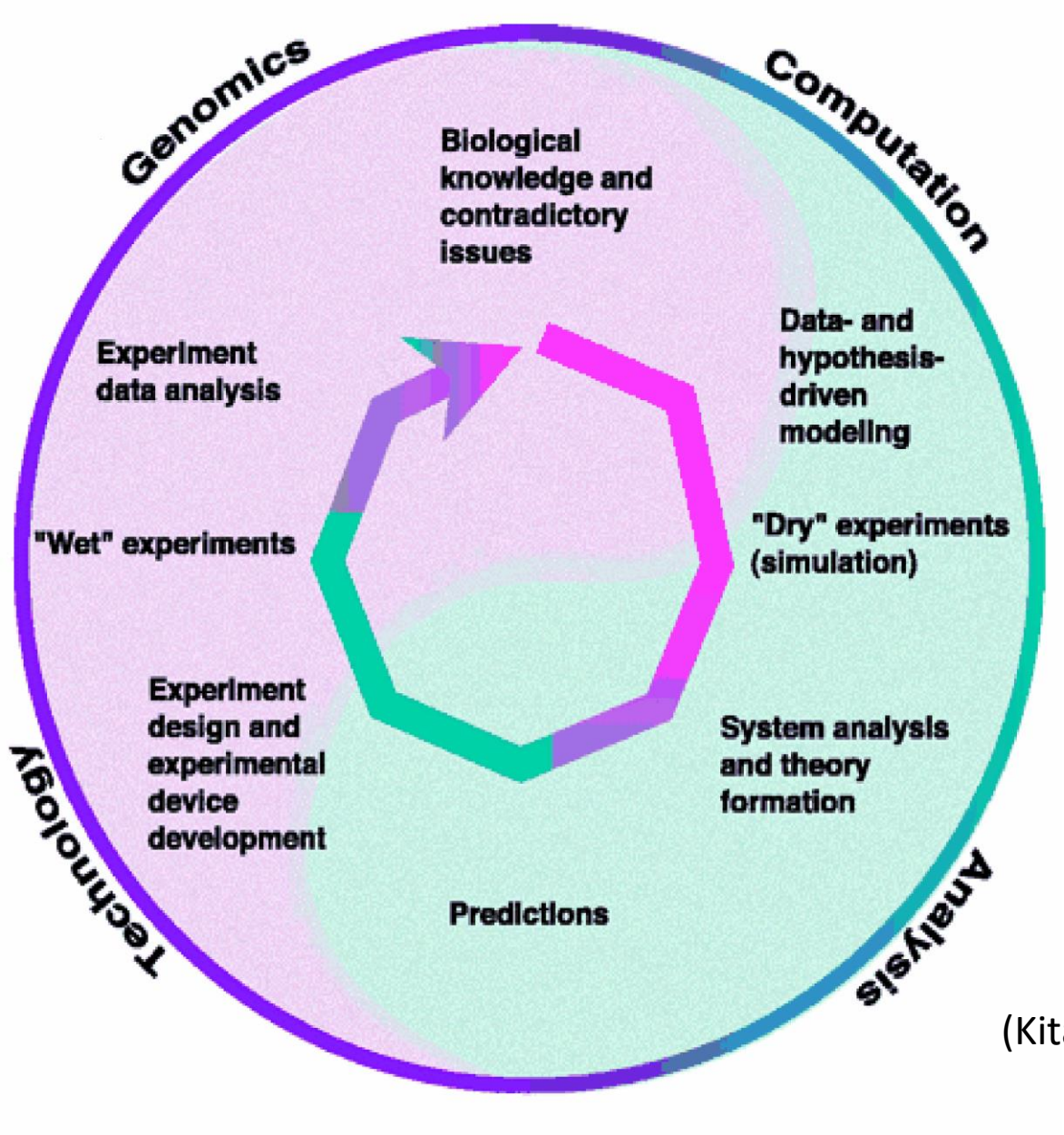
Corner stone of modern bioinformatics

1 Evolving trends in bioinformatics

Definition

- Bioinformatics as a field has arisen in parallel with the development of automated high-throughput methods of biological and biochemical discovery that yield a variety of forms of experimental data:
  - DNA sequences,
  - gene expression patterns,
  - three-dimensional models of macromolecular structure, …
- The field's rapid growth is driven by the vast potential for new understanding that can lead to new treatments, new drugs, new crops, and the general expansion of knowledge.

(http://findarticles.com/p/articles/mi_qa3886/is_200301/ai_n9182276/)


Genomics
Computation
Technology
Analysis
Biological knowledge and contradictory issues
Data- and hypothesis-driven modeling
"Dry" experiments (simulation)
System analysis and theory formation
Predictions
Experiment design and experimental device development
"Wet" experiments
Experiment data analysis

(Kitano 2002)

Oxford English Dictionary

(Molecular) bio– informatics:

bioinformatics is conceptualising *biology* in terms of molecules (in the sense of physical chemistry) and applying "*informatics techniques*" (derived from disciplines such as applied maths, computer science and statistics) to *understand and organize* the information associated with these molecules, *on a large scale*.
In short, bioinformatics is a *management information system* for molecular biology and has many practical applications.

.

(see also: Luscombe et al. 2001 – review paper)

Towards a formal definition

- Bioinformatics encompasses everything
  - from data storage
  - over data retrieval
  - to computational testing of biological hypotheses.

- Bioinformatics information systems include
  - multi-layered software,
  - hardware, and
  - experimental solutions

  bringing together a variety of tools and methods to analyze enormous quantities of “noisy” data.

(http://findarticles.com/p/articles/mi_qa3886/is_200301/ai_n9182276/)

Formal definition

Bioinformatics	**Computational biology**
Research, development or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, analyze, or visualize such data	Development and application of data-analytical, theoretical methods, mathematical modeling and computational simulation to the study of biological, behavioral, and social systems.

(BISTIC Definition Committee, NIH, 2000)

Diverse data require diverse “handling”

Libertas Academica
FREEDOM TO RESEARCH

Open Access: Full open access to this and thousands of other papers at http://www.la-press.com.

Cancer Informatics

Supplementary Issue: Classification, Predictive Modelling, and Statistical Analysis of Cancer Data (A)

Review of Current Methods, Applications, and Data Management for the Bioinformatics Analysis of Whole Exome Sequencing

Riyue Bao[1,*], Lei Huang[1,*], Jorge Andrade[1,*], Wei Tan[2,*], Warren A. Kibbe[3], Hongmei Jiang[4,§] and Gang Feng[3,§]

[1]Center for Research Informatics, The University of Chicago, Chicago, IL, USA. [2]IBM Thomas J. Watson Research Center, Yorktown Heights, New York, USA. [3]Biomedical Informatics Center (NUBIC), Clinical and Translational Sciences Institute (NUCATS), Northwestern University, Chicago, IL, USA. [4]Department of Statistics, Northwestern University, Evanston, IL, USA. *These authors contributed equally to this work. §Co-corresponding authors.

(Bao et al. 2014)

Example: The TELEVIE PDAC-xome project

- Setting / motivation

  Pancreatic ductal adenocarcinoma (PDAC), the most common type of pancreatic cancer, could be considered an orphan disease because, until present, it has not been adopted by the pharmaceutical industry. Despite its relatively low population incidence, it is the deadliest cancer worldwide with <6% 5-year survival rate. It is highly plausible that complex interactions between multiple genomic, epigenomic and environmental risk factors are involved

The PDAC-xome project

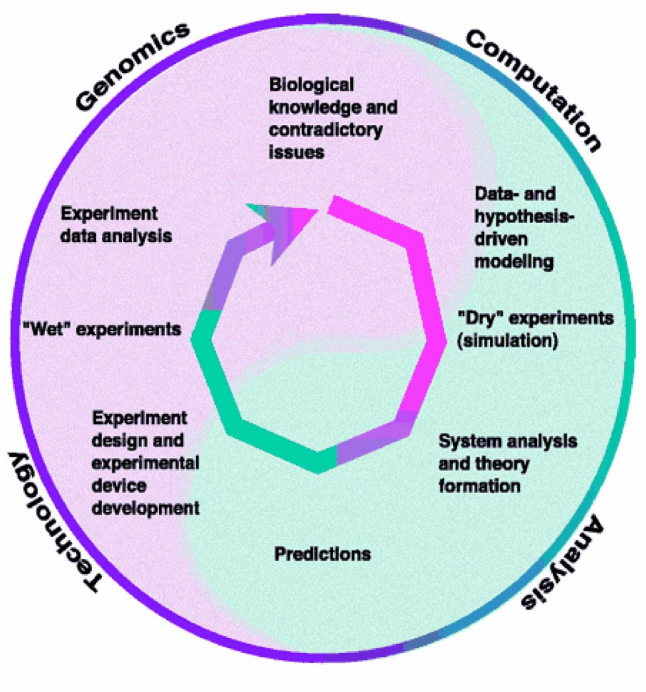
- Aims and goals

Objective 1: to perform **whole exome sequencing** (WES) on selected samples from the University Hospital in Liège (CHU-Liège) PDAC pathology repository and to describe the mutational landscape of these patients.

Objective 2: to provide a molecular characterization of PDAC using the aforementioned samples, hereby expanding the landscape of single nucleotide polymorphisms in human PDAC to include rare variants while charting patient **heterogeneity**.

Objective 3: to perform association analyses to identify candidate variants, genes, and pathways involved in PDAC, possibly **beyond main effects.**

The PDAC-xome project - bioinformatics



Technology:

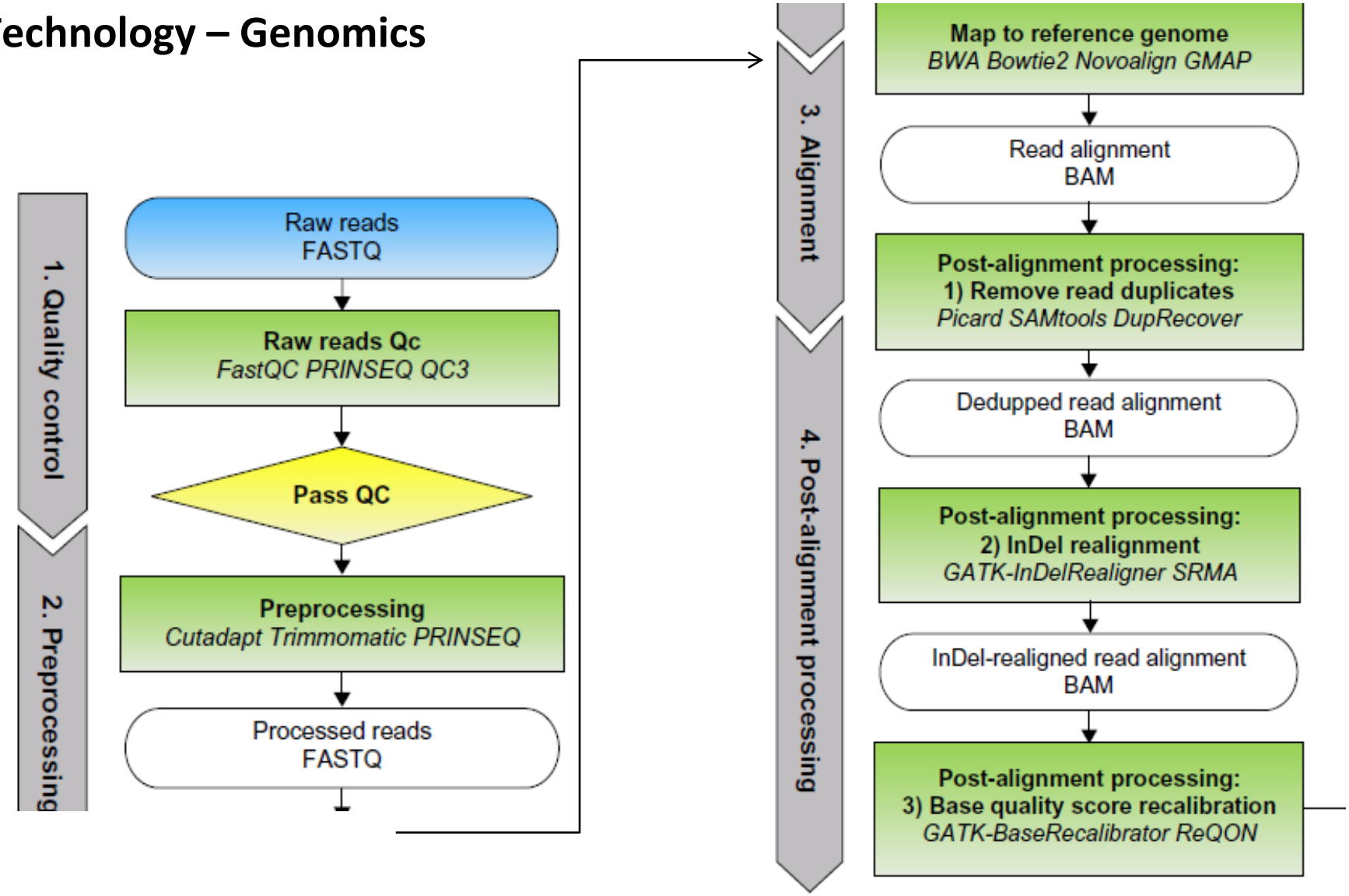
- platforms to generate the data
- internal quality control measures
- pipeline development

Genomics

Computation

Analysis

Technology – Genomics

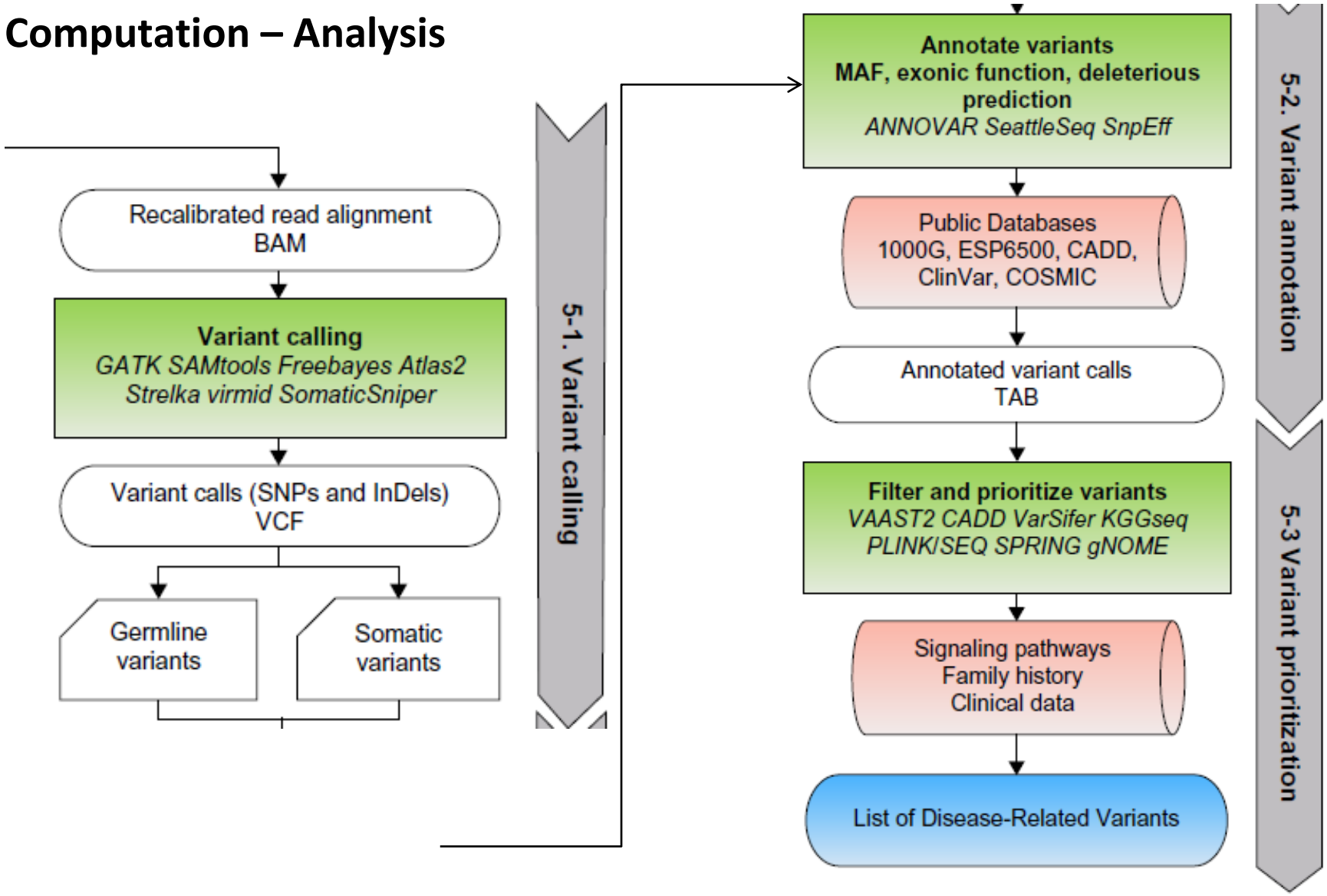


(Bao et al. 2014)

Computation - Analysis

- The previous shows part of a general framework of "**WES data analysis**" (WES = whole exome sequencing) :
  - raw reads QC,
  - preprocessing,
  - alignment,
  - post-processing, and
  - variant analysis (variant calling, annotation, and prioritization).
- FASTQ, BAM, variant call format (VCF), and TAB (tab-delimited) refer to the standard file format of raw data, alignment, variant calls, and annotated variants, respectively. A selection of tools supporting each analysis step is shown in italic.

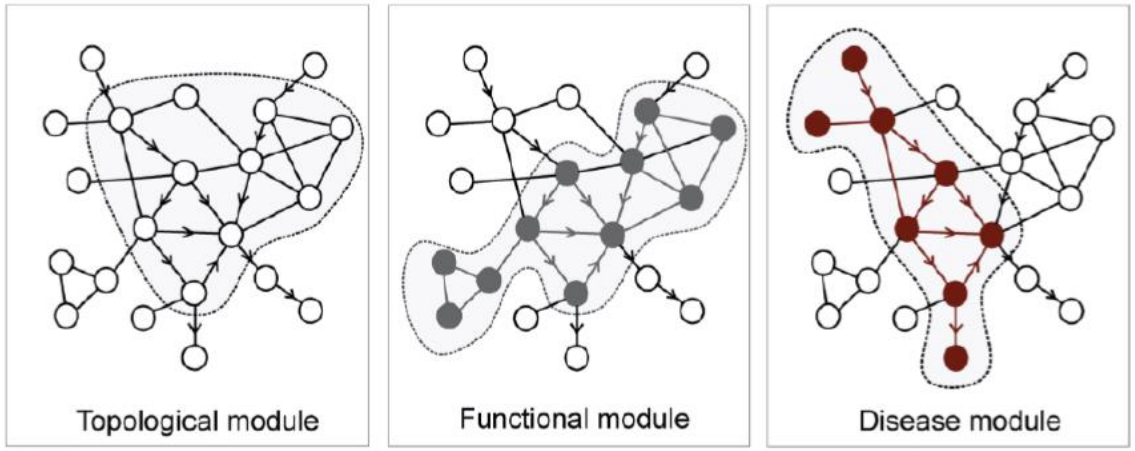
Computation – Analysis



(Bao et al. 2014)

- **The PDAC-xome project – statistical genetics**
  - Data collection and WES analysis
  - Molecular reclassification
  - Association analysis

Collaborations between statistical geneticists, biomedics, medical doctors and bioinformaticians (and many more) for this project, on the road to genome and **network medicine** …



(Barabási et al. 2011)

Bioinformatics: opportunities in drug research development

- Biofinformatics plays an increasing role in health care, for instance in new drug research and development:
  - Identification of novel drug/vaccine targets
  - Structural predictions
  - Tapping into biodiversity
  - Reconstruction of metabolic pathways
  - Systems biology

Bioinformatics in the new millennium

- Huge increase in available biological information
- Classic paradigm of “molecular biology” is altering rapidly to “genomics”
- Understanding of the new paradigms concerns more than simply “bench biology”
- Discovery requires large scale systems and broad collaborations, and dealing with global problems as well
- Funding comes in large amounts at group level, no longer a single laboratory or institution effort.
- Accountable output

Intersection of disciplines

E. Gusareva

B Dizier

F. Van Lishout

BIO

K. Chaichoompu

Bio[3]: **Bio**statistics – **Bio**medicine - **Bio**informatics

K. Bessonov

A. Tawfik

R. Fouladi

S. Pineda

(dd 2015)

Historical notes

BIOINFORMATICS REVIEW Vol. 19 no. 17 2003, pages 2176–2190
DOI: 10.1093/bioinformatics/btg309

Early bioinformatics: the birth of a discipline—a personal view

Christos A. Ouzounis[1,*] and Alfonso Valencia[2]

[1]Computational Genomics Group, The European Bioinformatics Institute, EMBL Cambridge Outstation, Cambridge CB10 1SD, UK, [2]Protein Design Group, National Center for Biotechnology, CNB-CSIC Campus U. Autonoma Cantoblanco, Madrid 28049, Spain

Received on December 13, 2002; revised on May 25, 2003; accepted on March 28, 2003

The pre-70's: pioneering computational studies

- Increased biological, structural understanding:
  - Structure of DNA (Watson and Crick, 1953)
  - Encoding of genetic info for proteins (Gamow et al. 1956)
  - Structural properties of protein molecules (Anfinsen and Scheraga 1975)
  - Evolution of biochemical pathways (Horowitz 1945)
  - Gene regulation (Britten and Davidson 1969)
  - Chemical basis for development (Turing 1952)
- Developments in fundamental computer science
  - Information theory (Shannon and Weaver 1962)
  - Game theory (Neumann and Morgenstern 1953)
- Combined: birth of computational biology

Bioinformatics	**Computational biology**
Research, development or application of computational tools and approaches for expanding the use of biological, medical, behavioral or health data, including those to acquire, store, organize, analyze, or visualize such data	Development and application of data-analytical, theoretical methods, mathematical modeling and computational simulation to the study of biological, behavioral, and social systems.

(BISTIC Definition Committee, NIH, 2000)

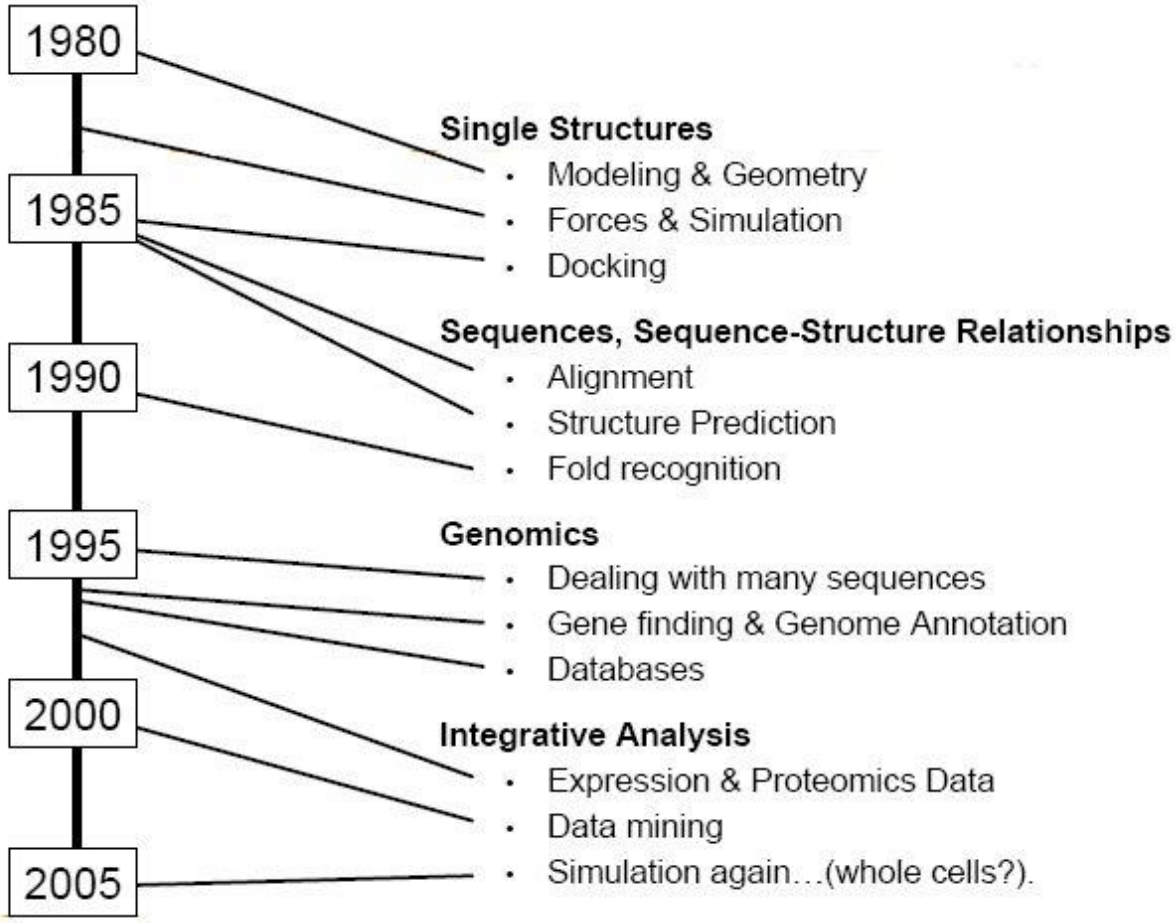
The pre-70's: theoretical foundations

- Merging of classical population genetics with molecular evolution (Ohta and Kimura 1971)
- String comparison problem in computer science (Levin 1973)

The 80's: More algorithms and resources

- Development of efficient algorithms to cope with an increasing volume of information
- Making available computer implementations for the wider scientific community
- Overall: Shaping computational biology as an independent discipline

The post 80's till 2005



(S-Star presentation; Choo)

Completion of the Human Genome Project

On June 26, 2000, the International Human Genome Sequencing Consortium announced the production of a rough draft of the human genome sequence. An essentially finished version is announced by them in April 2003.

2005+ : genomics and beyond

BIOTECHNOLOGY – Vol .XI – *Bioinformatics on Post genomic Era: From Genomes to Systems Biology* - Vihinen, Mauno

BIOINFORMATICS ON POST GENOMIC ERA: FROM GENOMES TO SYSTEMS BIOLOGY

Vihinen, Mauno
Institute of Medical Technology, University of Tampere, Finland and Research Unit, Tampere University Hospital, Tampere, Finland

Keywords: Genome, transcriptome, proteome, data mining, sequence analysis, bioinformatics, functional genomics, systems biology.

This review provides an introduction to the methods and problems tackled by bioinformatics. In addition to identifying all the genes in genomes it is crucial to store and distribute the information in databases. Annotation and identification of genes from genomes are crucial for generating useful genome databases. Ethical, legal, and social implications of genome and gene data are briefly discussed.

Challenges in bioinformatics

- Data deluge (availability, what to archive and what not?, …)
- Knowledge management (accessibility, usability, …)
- Predicting, not just explaining (what comes first: hypothesis generation, data collection? …)
- Precision medicine (alias: personalized medicine; holistic approach – correlating different causal associations – versus a reductionist approach – targeting very specific biomarkers, negative gold standards … negative controls)
- Speciation (loss in biodiversity, evolutionary units, “integrative taxonomy”: molecular, morphological, ecological and environmental information)
- Inferring the tree of life (unresolved orthology assignment, gene sampling pyramid

2 Careers in bioinformatics

Topics in bioinformatics from a journal's perspective

(source: Scope of the journal "Bioinformatics")

Data and Text Mining
This category includes: New methods and tools for extracting biological information from text, databases and other sources of information. Description of tools to organize, distribute and represent this information. New methods for inferring and predicting biological features based on the extracted information. The submission of databases and repositories of annotated text, computational tools and general methodology for the work in this area are encouraged, provided that they have been previously tested.

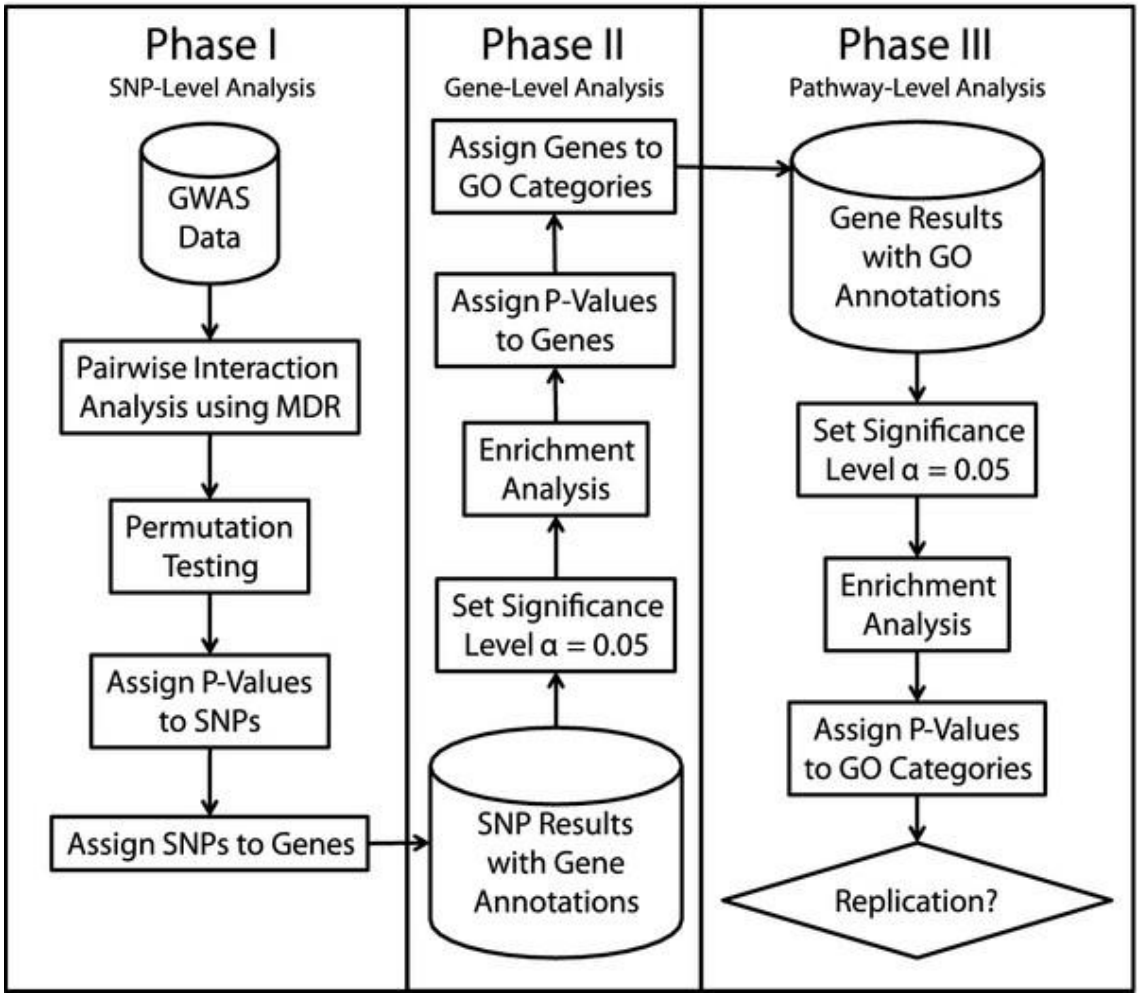
The journal ….

BioData Mining is an open access, peer reviewed, online journal encompassing research on all aspects of data mining applied to high-dimensional biological and biomedical data, focusing on computational aspects of knowledge discovery from large-scale genetic, transcriptomic, genomic, proteomic, and metabolomic data.

Topical areas include, but are not limited to:

- Development, evaluation, and application of novel data mining and machine learning algorithms.
- Adaptation, evaluation, and application of traditional data mining and machine learning algorithms.
- Open-source software for the application of data mining and machine learning algorithms.
- Design, development and integration of databases, software and web services for the storage, management, retrieval, and analysis of data from large scale studies.
- Pre-processing, post-processing, modeling, and interpretation of data mining and machine learning results for biological interpretation and knowledge discovery.

Databases and Ontologies



This category includes: Curated biological databases, data warehouses, eScience, web services, database integration, biologically-relevant ontologies.

(Kim et al. 2012)

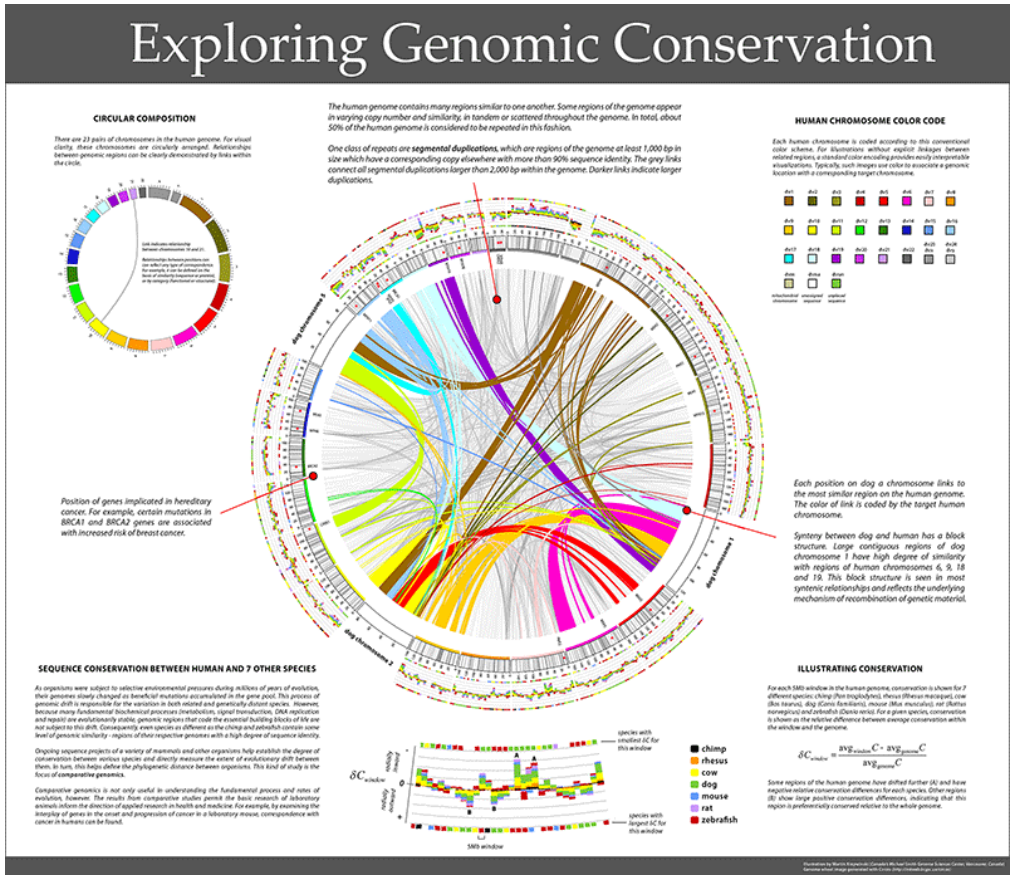
Bioimage Informatics

This category includes novel methods for the acquisition, analysis and modeling of images produced by modern microscopy, with an emphasis on the application of innovative computational methods to solve challenging and significant biological problems at the molecular, sub-cellular, cellular, and tissue levels.

This category also encourages large-scale image informatics methods/applications/software, joint analysis of multiple heterogeneous datasets that include images as a component, and development of bioimage-related ontologies and image retrieval methods.

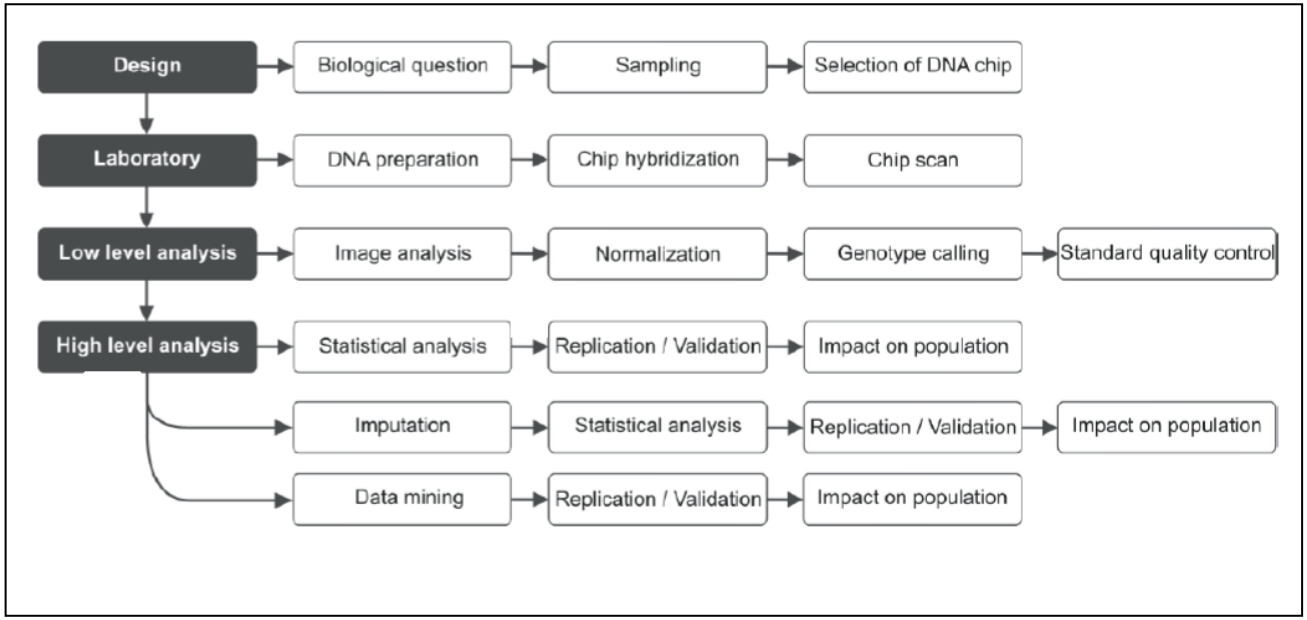
Genome analysis

This category includes: Comparative genomics, genome assembly, genome and chromosome annotation, identification of genomic features such as genes, splice sites and promoters.

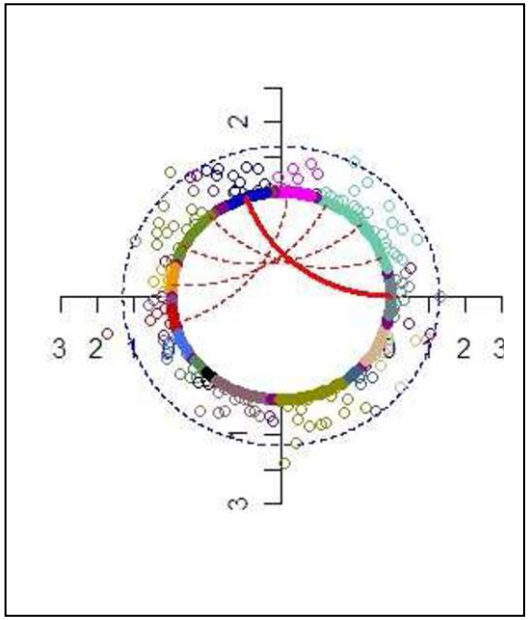


Genetics and Population Analysis

This category includes: Segregation analysis, linkage analysis, association analysis, map construction, population simulation, haplotyping, linkage disequilibrium, pedigree drawing, marker discovery, power calculation, genotype calling.



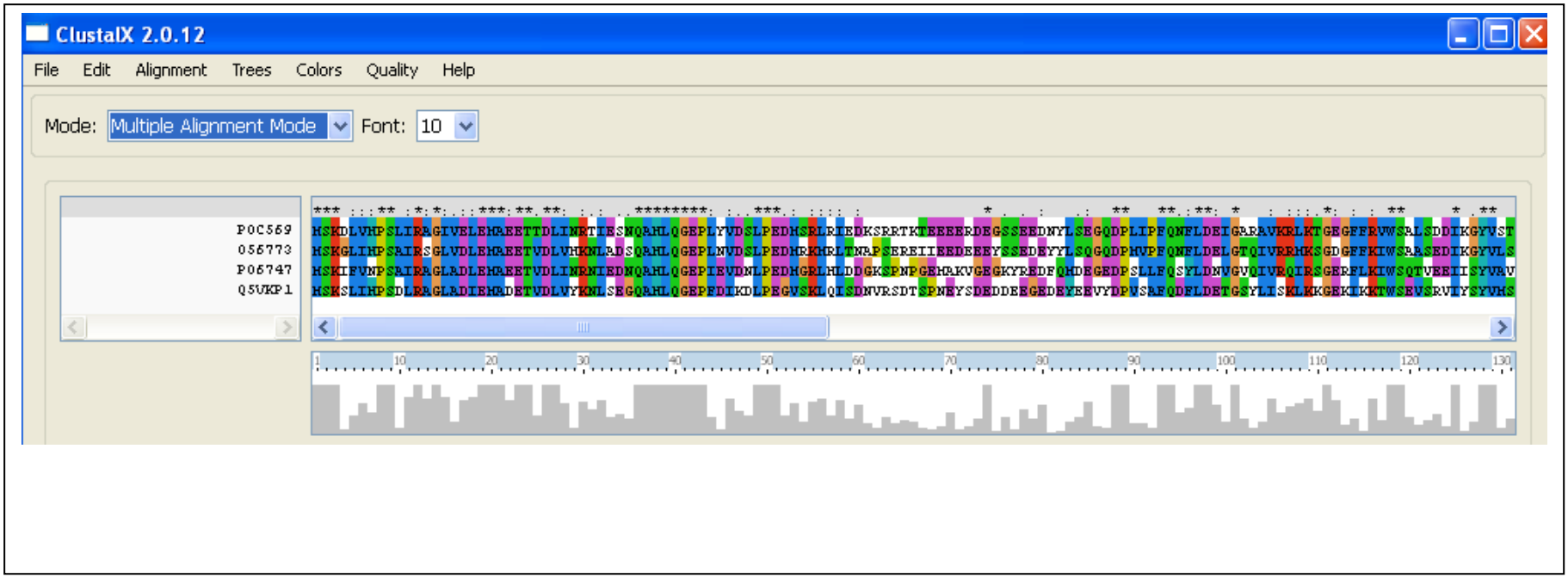
(Ziegler 2009)



(Van Steen 2011)

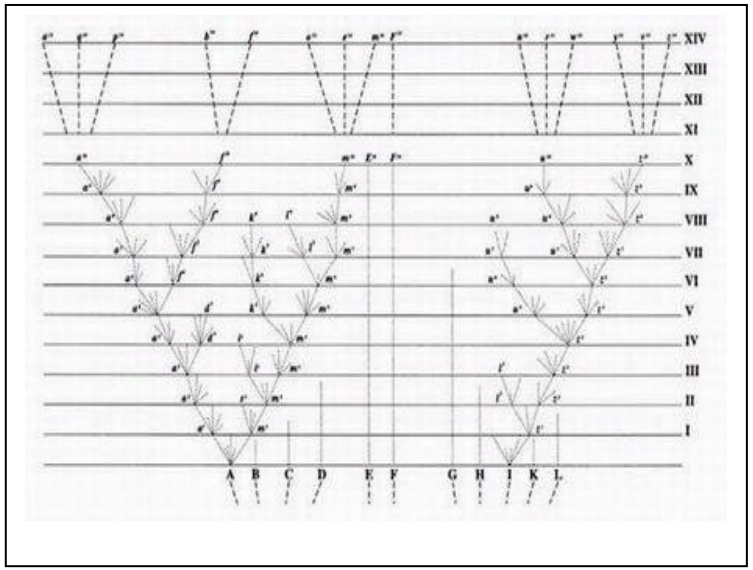
Sequence analysis

This category includes: Multiple sequence alignment, sequence searches and clustering; prediction of function and localisation; novel domains and motifs; prediction of protein, RNA and DNA functional sites and other sequence features.



Phylogenetics

This category includes: novel phylogeny estimation procedures for molecular data including nucleotide sequence data, amino acid data, whole genomes, SNPs, etc., simultaneous multiple sequence alignment and phylogeny estimation, phylogenetic approaches for any aspect of molecular sequence analysis (see Sequence Analysis scope), models of molecular evolution, assessments of statistical support of resulting phylogenetic estimates, comparative methods, coalescent theory, approaches for comparing phylogenetic trees, methods for testing and/or mapping character change along a phylogeny.



(Darwin 1859)

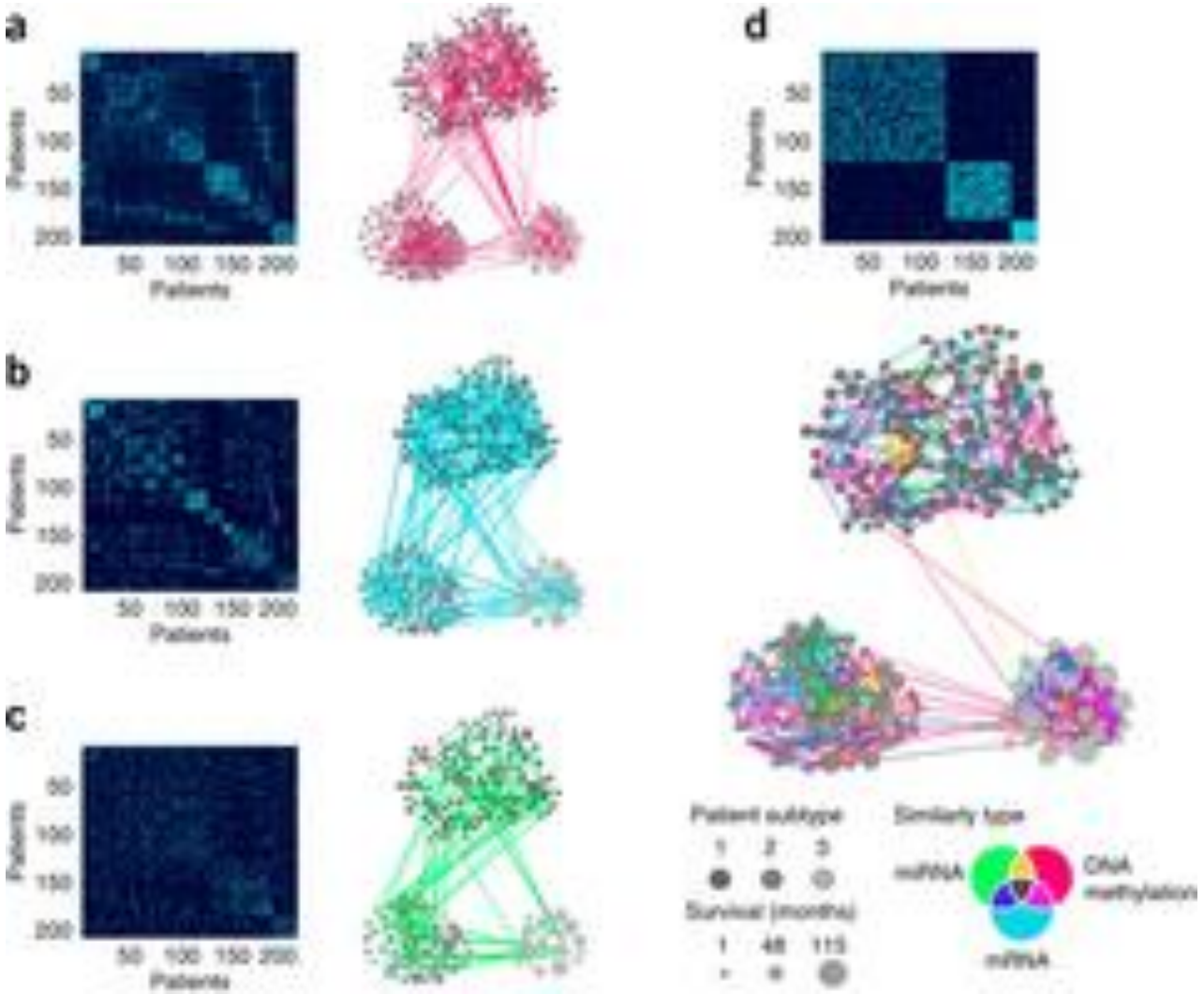
Gene Expression

This category includes a wide range of applications relevant to the high-throughput analysis of expression of biological quantities, including microarrays (nucleic acid, protein, array CGH, genome tiling, and other arrays), RNA-seq, proteomics and mass spectrometry. Approaches to data analysis to be considered include statistical analysis of differential gene expression; expression-based classifiers; methods to determine or describe regulatory networks; pathway analysis; …

Systems Biology

This category includes whole cell approaches to molecular biology. Any combination of experimentally collected whole cell systems, pathways or signaling cascades on RNA, proteins, genomes or metabolites that advances the understanding of molecular biology or molecular medicine will be considered. Interactions and binding within or between any of the categories will be considered including protein interaction networks, regulatory networks, metabolic and signaling pathways. Detailed analysis of the biological properties of the systems are of particular interest.

Omics integration



(Wang et al. 2014)

Becoming a bioinformatician

10 useful / necessary skills

- Strong background in some aspect 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

10 useful / necessary skills

- Mathematics (e.g. algorithm development)
- Engineering (e.g. robotics)
- Good knowledge of a few molecular biology software packages (molecular modeling / sequence analysis)
- Command line computing environment (Linux/Unix knowledge)
- Data administration (esp. relational database concept) and Computer Programming Skills/Experience (C/C++, Sybase, Java, Oracle) and Scripting Language Knowledge (Perl and perhaps Phython)

“Dammit Jim, I’m a doctor, not a bionformatician!”

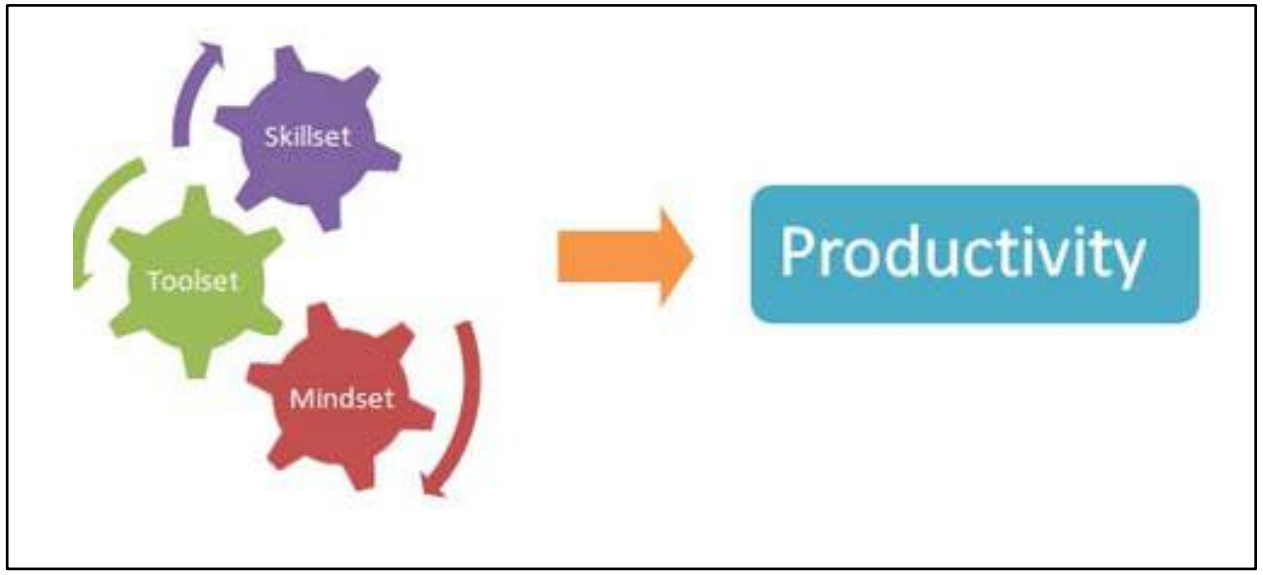
(http://www.youtube.com/watch?v=MULMbqQ9LJ8)

The following slides summarize key ideas from the white paper by Christophe Lambert – 2011, CEO and President of Golden Helix; course website: Lambert2011

Bioinformatics-driven genetic research

"your job and expertise is to do x, but to achieve your goals you also have to do y and z, which you either don't want to do or don't have the skills to do"

- It takes more than just brains to make productive advances in bioinformatics:



Skillset

- The free software tools used today require highly skilled bioinformatics professionals, which are often in short reply …
- One must have competences in several disciplines: computer science, statistics and genetics. Yet, good software overcomes limits in certain skillsets
- In practice though, one virtually has to be a computer programmer in order to perform genetics research … Why ?

Toolset

- To be able to deal with unavailable informatics expertise that can take advantage of for instance the new sequencing capabilities, some minimal requirements for the toolset are needed:
  - Robust (not only for local applications)
  - Well-documented (for application by non-informaticians)
  - Well-supported (to address questions from users with different backgrounds)
  - Maintained and updated (state-of-the art optimized software; is the case for about 9% of academic programs developed for genetic analysis)
  - Multi-platform / Cloud-friendly
  - Data format transformations (omics integration!)

Toolset (continued)

- Foundational to the problem is the discrepancy between:
  - The fact that academia is the birthplace of most new statistical and computational methods in genetic research
  - When funding ends, software support often ends
  - Innovation is abandoned (i.e., a new promising tool is not being used) when one cannot be sure about the accuracy of results (due to lack of documentation about what certain program options imply)

- Nevertheless, there are positive examples:

The BIOCONDUCTOR project

Mindset

- Each person has a mindset = the amalgamation of values and mental models of reality with which one engages in a goal-directed behavior.
  - Values determine goals
  - Mental models determine means of moving towards these goals
- How to measure the **distance** between current state and goal?
  - Any such metrics will drive behavior
- In academia, the prime metric is reputation:
  - Recruiting
  - Promotions
  - Grant funding
- Alternative to fuzzy metric of “reputation”: publications

Publish or perish

Impact of mindset on toolset

- Focus on papers and hence mathematical and algorithm innovation
- Ample resources (time, human resources, financial resources) for software quality

Inadequate translation from development to application

Impact of mindset on toolset

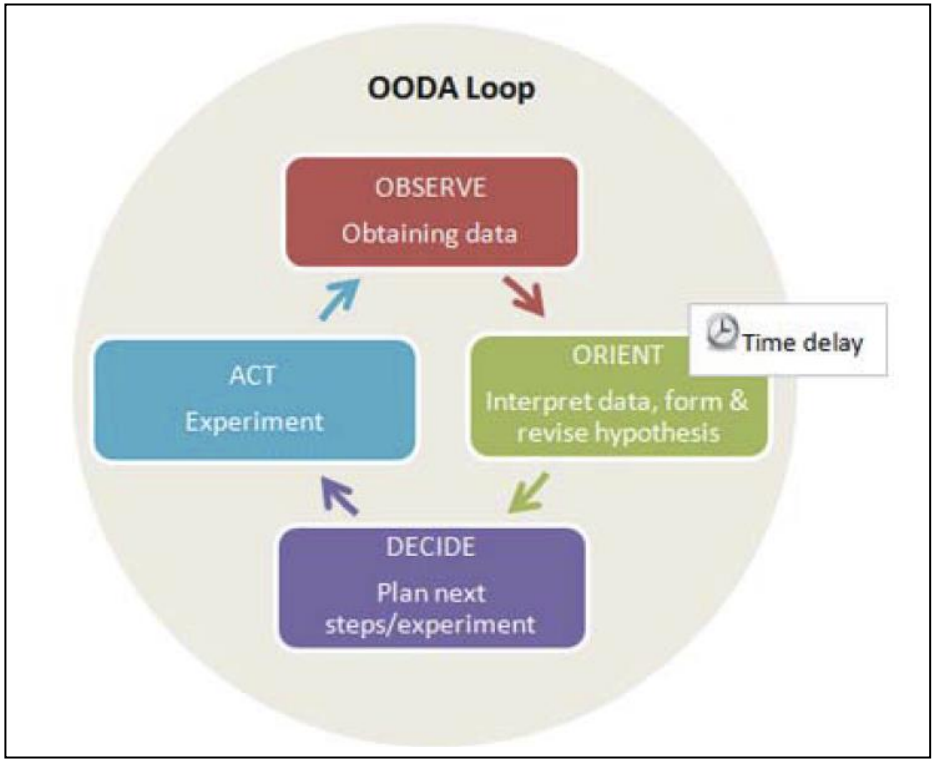
Discussion statement:

“While I do not believe it is intentional, a pattern operates where the scientific promise of the freely downloadable software is touted in the paper, and then the “consumer,” due to time pressures and challenges of learning the software, programming/statistical skillset deficits, and/or fear of improperly driving the software and making embarrassing mistakes in publications, must call in the software author and/or his students as consultants on projects to get the research done.”

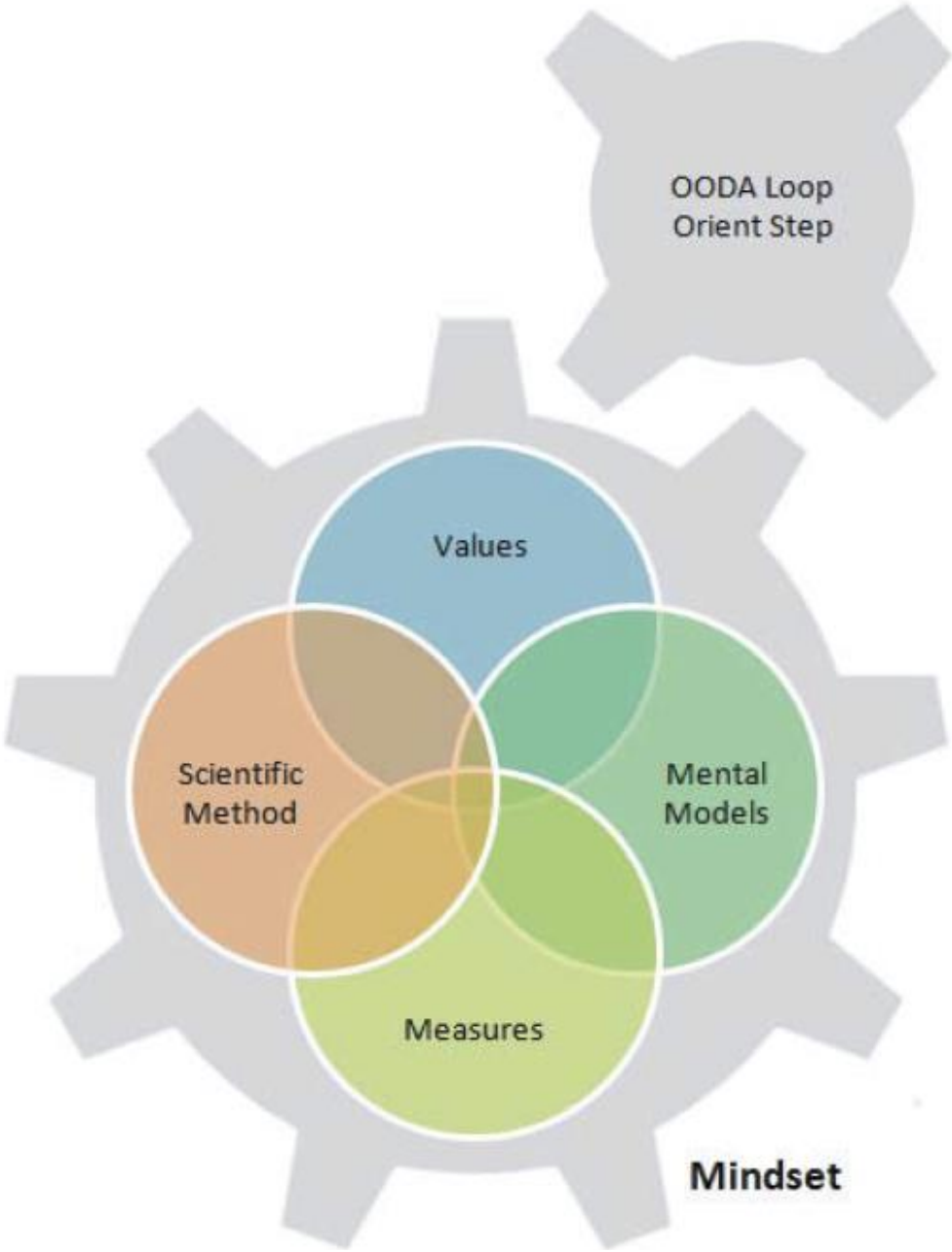
Consequences of the mindset-toolset impact

- Bioinformaticians have become the constrained bottleneck resource
  - Genomic projects too heavily rely on (in-house) bioinformaticians
  - Because of scarcity, labs may need to wait for months and months to get the data analyzed, hereby delaying or blocking next steps of the research process

- Outsourcing may be an option (when financial resources are available), but will impact the learning loop: the **OODA loop model for goal-directed activities**



Time delays


OODA Loop
Orient Step
Values
Scientific
Method
Mental
Models
Measures
Mindset

Consequences of the mindset-toolset impact

Discussion statement:

"If inadequate or missing documentation, lack of support on appropriate platforms, buggy or unstable code, and generally low usability are the norm for academic software, how can one reproduce research when the tools are unusable? Worse, as quantified earlier, a large fraction of software tools that are developed and used by a researcher to produce a publication cease to exist within a few years. Reconstruction of the steps of another researcher ranges from tedious to impossible."

In class reading: reproducible research and solutions (group discussion)

Take-home message for the practical side of this course

Checking the quality of analysis results:

"…This could be made possible with software tools that automatically log all analysis steps and parameters and saves the intermediate steps of analysis, including graphical output, for inspection at a later date. Further, the ability to annotate one's work as one might do in a laboratory notebook would allow an analyst (expert or novice) to share his or her work with a colleague for collaboration, review, and audit."

Choosing between industry and academia

OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Editorial

Ten Simple Rules for Choosing between Industry and Academia

David B. Searls*

One of the most significant decisions we face as scientists comes at the end of our formal education. Choosing between industry and academia is easy for some, incredibly fraught for others. The author has made two complete cycles between these career destinations, including on the one hand 16 years in academia, as grad student (twice, in biology and in computer science), post-doc, and faculty, and on the other hand 19 years in two different

bioinformaticists were in such short supply that any qualification would do.

If you are an old hand and have already notched up a post-doc or two, take stock of your star power. This unspoken but universally understood metric encompasses such factors as whom you've trained with, where you've published (and how much), and what recent results of yours are on everyone's lips. If you are fortunate enough to have significant capital in this

you need a quick infusion of cash, companies may offer signing bonuses, though again these were more common when bioinformatics was a rarer commodity.

Industry offers forms of compensation unavailable in academia, and you will need to consider how to value them relative to your present and future needs. Despite recent bad press, bonus systems are often part of the equation, and depending on your entry point they may

(Searls 2009)

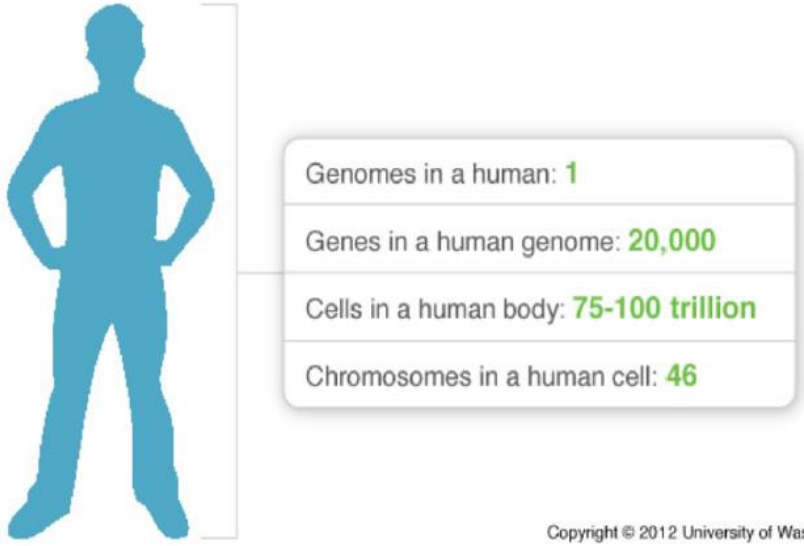
3 Gen-omics

Corner stones of bioinformatics

Genetics / Genomics

- Genetics is the study of how traits such as hair color, eye color, and risk for disease are passed ("inherited") from parents to their children. Genetics influence how these inherited traits can be different from person to person.

- Your genetic information is called your genetic code or genome. Your genome is made up of a chemical called deoxyribonucleic acid (DNA) and is stored in almost every cell in your body.



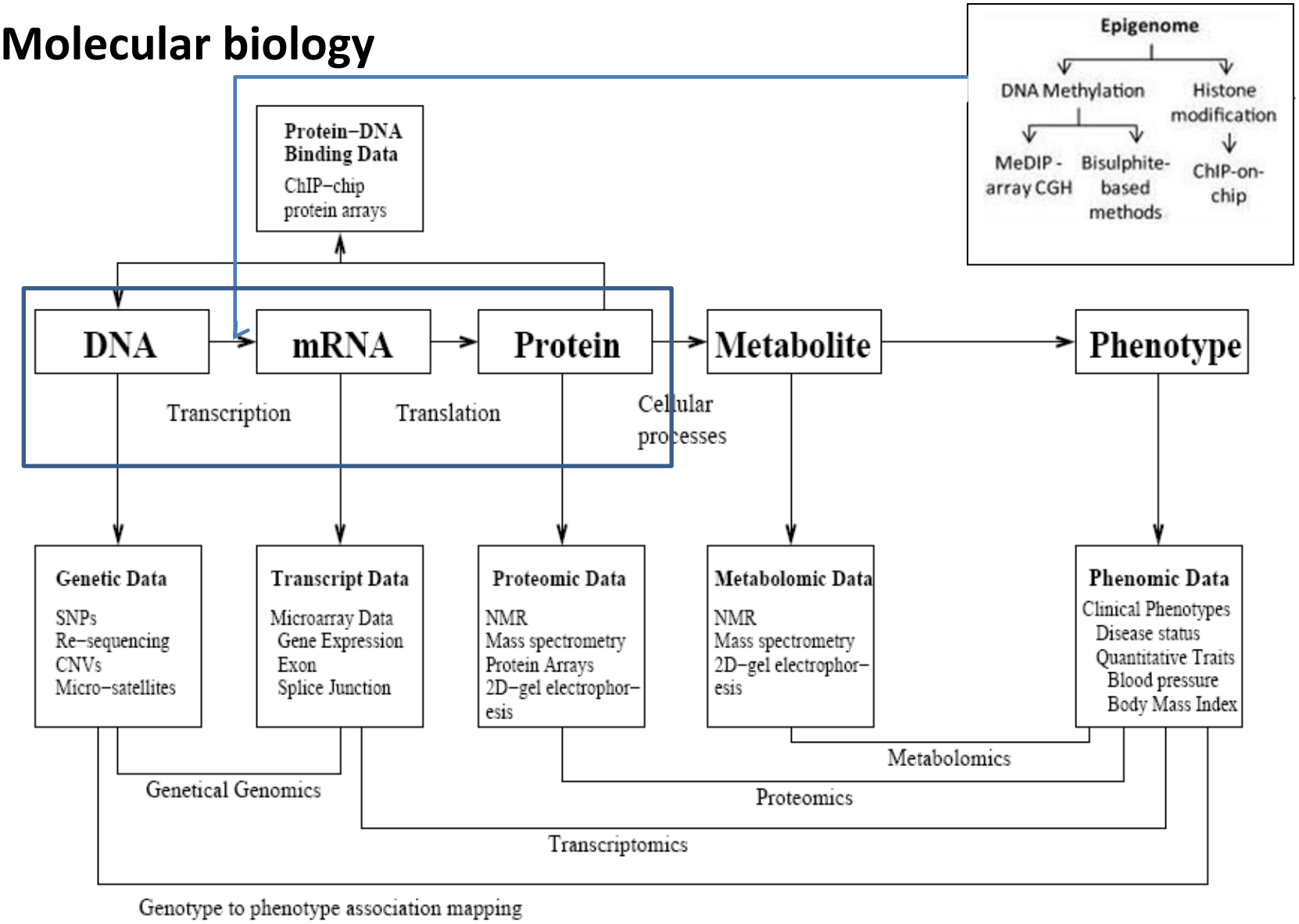
Information is everywhere: the programming of life

http://www.youtube.com/watch?v=00vBqYDBW5s

"Information:

that which can be communicated through symbolic language"

Molecular biology



(adapted from: Davies et al 2009, Integrative genomics and functional explanation)

Computational biology ("early bioinformatics")

- Biology = noun
- Computational = adjective

"When I use my method (or those of others) to answer a biological question, I am doing science. I am learning new biology. The criteria for success has little to do with the computational tools that I use, and is all about whether the new biology is true and has been validated appropriately and to the standards of evidence expected among the biological community. The papers that result report new biological knowledge and are science papers. This is computational biology." (https://rbaltman.wordpress.com)

Computational biology = the study of biology using computational techniques. The goal is to learn new biology, knowledge about living systems. It is about science.

Bioinformatics

- Bio(logy) + Informatics
- 2 nouns

"When I build a method (usually as software, and with my staff, students, post-docs–I never unfortunately do it myself anymore), I am engaging in an engineering activity: I design it to have certain performance characteristics, I build it using best engineering practices, I validate that it performs as I intended, and I create it to solve not just a single problem, but a class of similar problems that all should be solvable with the software. I then write papers about the method, and these are engineering papers. This is bioinformatics."
(https://rbaltman.wordpress.com)

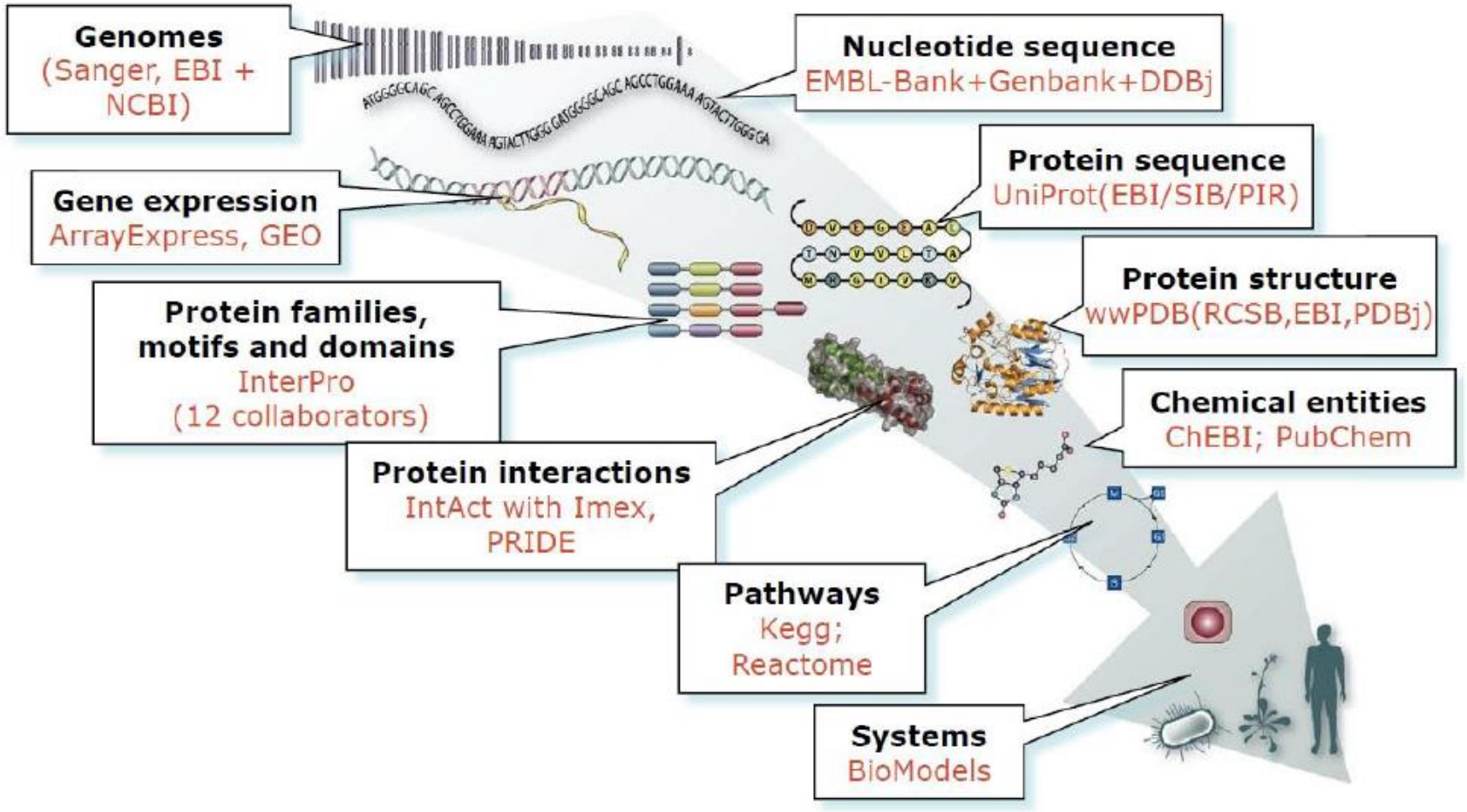
Bioinformatics = the creation of tools (algorithms, databases) that solve problems. The goal is to build useful tools that work on biological data. It is about engineering.

Genetic epidemiology

- Genetic epidemiology is a particular sub-discipline of epidemiology which considers genetic influences on human traits.
- As with any other epidemiologic studies, genetic epidemiology studies aim
  - to assess the public health importance of diseases,
  - to identify the populations at risk,
  - to identify the causes of the disease,
  - and to evaluate potential treatment or prevention strategies based on those findings
- Strategies of analysis include population studies and family studies.
  - Huge challenge is to combine big data repositories from population or family-studies (e.g., genomics, transcriptomics, metagenomics, metabolomics, epigenomics, ….) with clinical and demographic data:

BIOINFORMATICS

Integration with high-throughput omics (data-bases)



(Janet Thornton, EBI)

Course toolset repository

Principal internet resources for genome browsers and databases

Resource	Web address	Description	Sponsoring organizations
Open Helix	http://www.openhelix.com/tutorials.shtml	On-line tutorial material for all of the genome databases.	OpenHelix, LLC
UCSC Genome Browser	http://genome.ucsc.edu	Comprehensive, multi-species genome database providing genome browsing and batch querying.	Genome Bioinformatics Group, University of California, Santa Cruz
Ensembl Browser	http://www.ensembl.org	Comprehensive, multi-species genome database providing genome browsing and batch querying.	European Bioinformatics Institute (EBI) and the Sanger Center
NCBI MapViewer	http://www.ncbi.nlm.nih.gov/mapview	Multi-species genome browser focusing especially on genome mapping applications.	National Center for Biotechnology Information (NCBI)

Biomart	http://www.biomart.org/	Genome-database, batch-querying interface used by Ensembl and several single-genome databases.	Ontario Institute for Cancer Research and European Bioinformatics Institute
Galaxy	http://main.g2.bx.psu.edu	Integrated toolset for analyzing genome batch-querying data.	Center for Comparative Genomics and Bioinformatics. Penn State University
Taverna	http://taverna.sourceforge.net	Toolset for creating pipelines of bioinformatics analyses implemented via the Web services protocol.	Open Middleware Infrastructure Institute, University of Southampton (OMII-UK)
GMOD	http://www.gmod.org	Repository of software tools for developing generic genome databases.	A consortium of organizations operating as the Generic Model Organism Database project

(Schattner et al. 2009)

Bioconductor (TA- sessions)

Method

Open Access

Bioconductor: open software development for computational biology and bioinformatics

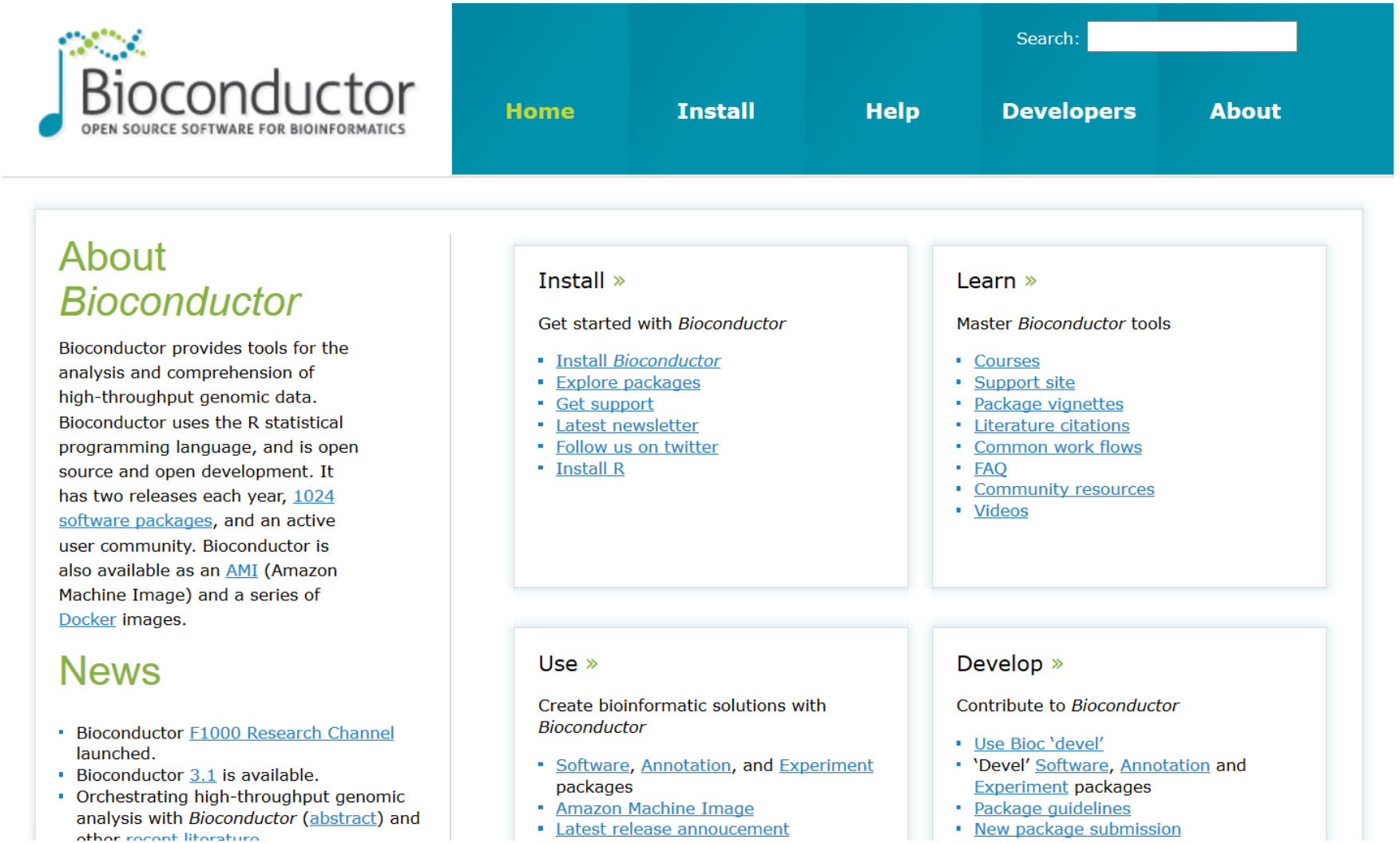
Robert C Gentleman[1], Vincent J Carey[2], Douglas M Bates[3], Ben Bolstad[4], Marcel Dettling[5], Sandrine Dudoit[4], Byron Ellis[6], Laurent Gautier[7], Yongchao Ge[8], Jeff Gentry[1], Kurt Hornik[9], Torsten Hothorn[10], Wolfgang Huber[11], Stefano Iacus[12], Rafael Irizarry[13], Friedrich Leisch[9], Cheng Li[1], Martin Maechler[5], Anthony J Rossini[14], Gunther Sawitzki[15], Colin Smith[16], Gordon Smyth[17], Luke Tierney[18], Jean YH Yang[19] and Jianhua Zhang[1]

Published: 15 September 2004

Genome ***Biology*** 2004, **5**:R80

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2004/5/10/R80

Received: 19 April 2004
Revised: 1 July 2004
Accepted: 3 August 2004



(http://www.bioconductor.org/)

R (TA- sessions)

- R is a freely available language and environment for statistical computing and graphics which provides a wide variety of statistical and graphical techniques: linear and nonlinear modelling, statistical tests, time series analysis, classification, clustering, etc.
  - Consult the R project homepage for further information.
  - The “R-community” is very responsive in addressing practical questions with the software (but consult the FAQ pages first!)
- CRAN is a network of ftp and web servers around the world that store identical, up-to-date, versions of code and documentation for R.

[Home]

Download

CRAN

R Project

About R
Contributors
What's New?
Mailing Lists
Bug Tracking
Conferences
Search

R Foundation

Foundation
Board
Members
Donors
Donate

Documentation

The R Project for Statistical Computing

Getting Started

R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS. To **download R**, please choose your preferred CRAN mirror.

If you have questions about R like how to download and install the software, or what the license terms are, please read our answers to frequently asked questions before you send an email.

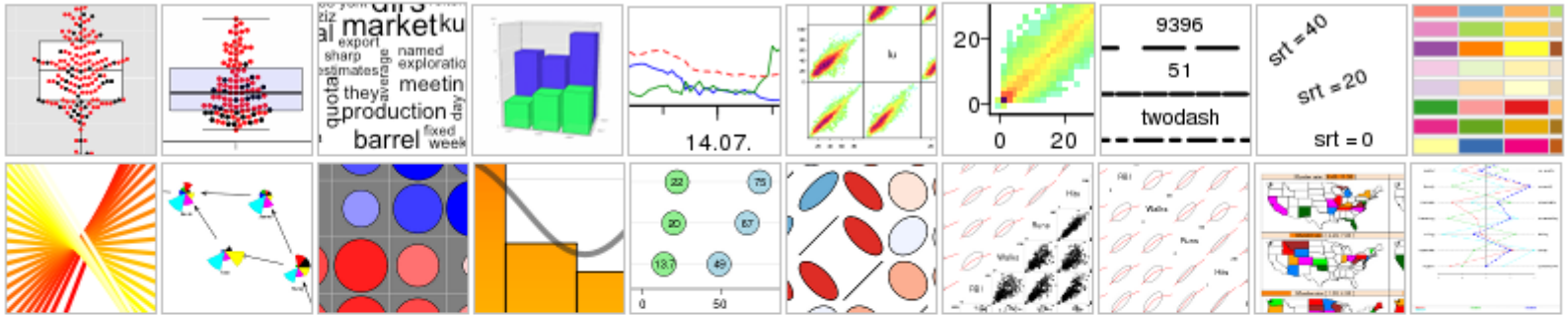
News

- **R version 3.2.2 (Fire Safety)** has been released on 2015-08-14.
- **The R Journal Volume 7/1** is available.
- **R version 3.1.3 (Smooth Sidewalk)** has been released on 2015-03-09.
- **useR! 2015**, will take place at the University of Aalborg, Denmark, June 30 - July 3, 2015.
- **useR! 2014**, took place at the University of California, Los Angeles, USA June 30 - July 3, 2014.

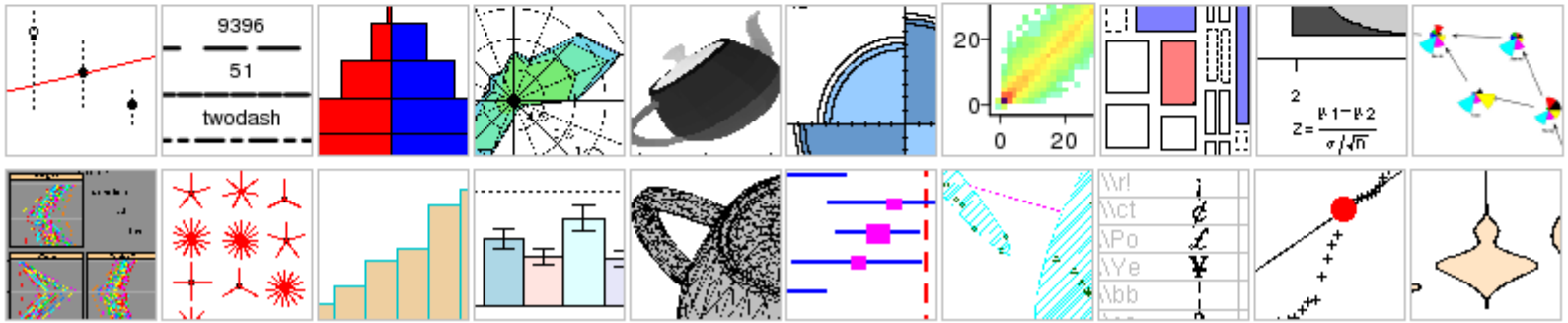
(https://www.r-project.org/)

The R graph gallery (https://www.r-graph-gallery.com/)

» Last entries ...



» Random entries



- One of R's strengths is the ease with which well-designed publication-quality plots can be produced …