In the brain of a Computational Biologist ?

Many hats : (Nature Biotechnology, Vol. 31, Number 11, Nov. 2013)
https://www.nature.com/articles/nbt.2740

« - data analyst
- data curator
- database developer
- statistician
- mathematical modeler
- bioinformatician
- software developer
- ontologist
- and many more »

For sure :

- *« computers are now essential components of modern biological research »*

A problematic :

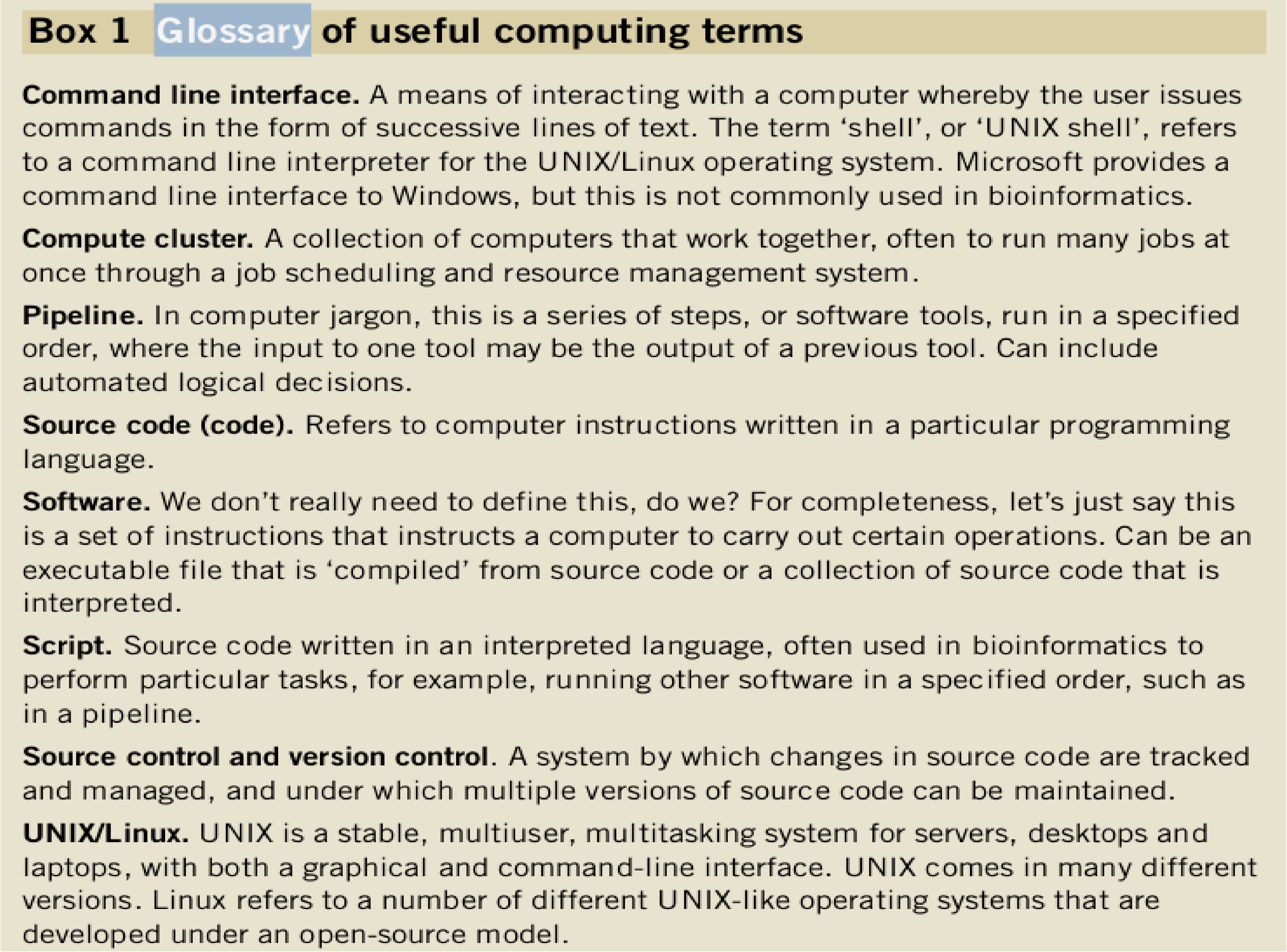
- to which extent (ethical for instance : biohacking)

A need :

- a scientist in FdV needs this kind of skills a.k.a. computational biology and needs to master different kind of terminologies

As well as

De novo transcriptome assembly	is the method of creating a transcriptome without the aid of a reference genome.
The transcriptome	is the set of all RNA molecules, including mRNA, rRNA, tRNA, and other non-coding RNA transcribed in one cell or a population of cells. It differs from the exome in that it includes only those RNA molecules found in a specified cell population, and usually includes the amount or concentration of each RNA molecule in addition to the molecular identities. http://en.wikipedia.org/wiki/Transcriptome
Megabase (Mb)	
Genome	



Box 1 Glossary of useful computing terms

Command line interface. A means of interacting with a computer whereby the user issues commands in the form of successive lines of text. The term ‘shell’, or ‘UNIX shell’, refers to a command line interpreter for the UNIX/Linux operating system. Microsoft provides a command line interface to Windows, but this is not commonly used in bioinformatics.

Compute cluster. A collection of computers that work together, often to run many jobs at once through a job scheduling and resource management system.

Pipeline. In computer jargon, this is a series of steps, or software tools, run in a specified order, where the input to one tool may be the output of a previous tool. Can include automated logical decisions.

Source code (code). Refers to computer instructions written in a particular programming language.

Software. We don’t really need to define this, do we? For completeness, let’s just say this is a set of instructions that instructs a computer to carry out certain operations. Can be an executable file that is ‘compiled’ from source code or a collection of source code that is interpreted.

Script. Source code written in an interpreted language, often used in bioinformatics to perform particular tasks, for example, running other software in a specified order, such as in a pipeline.

Source control and version control. A system by which changes in source code are tracked and managed, and under which multiple versions of source code can be maintained.

UNIX/Linux. UNIX is a stable, multiuser, multitasking system for servers, desktops and laptops, with both a graphical and command-line interface. UNIX comes in many different versions. Linux refers to a number of different UNIX-like operating systems that are developed under an open-source model.

- Biological Knowledge (K) is of outmost importance for interpreting computed results
  - https://med.stanford.edu/profiles/john-ioannidis?tab=publications

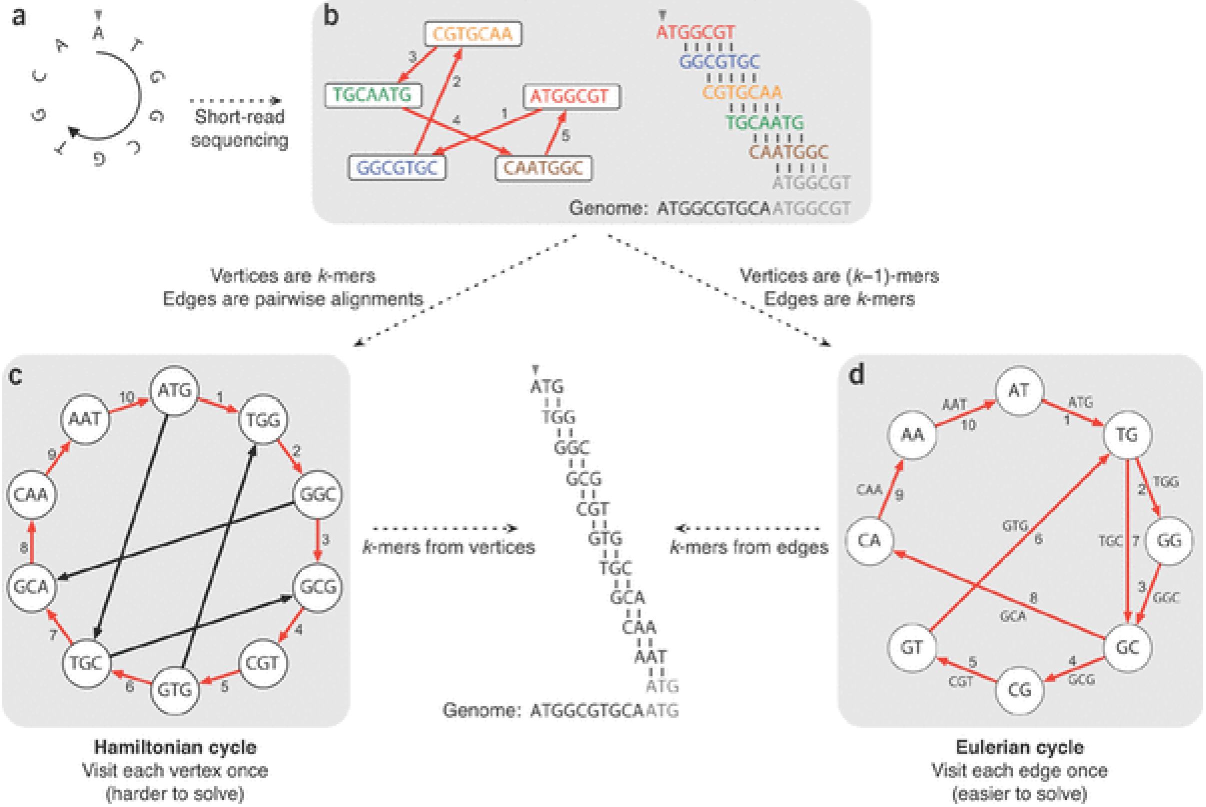
- Let’s be a data detective and able to logically manipulate them
  - http://www.biomedcentral.com/content/pdf/1471-2105-5-80.pdf

- Use all the ressources but with intelligence
  - http://www.open-bio.org/wiki/Main_Page
  - http://www.biostars.org
  - http://seqanswers.com

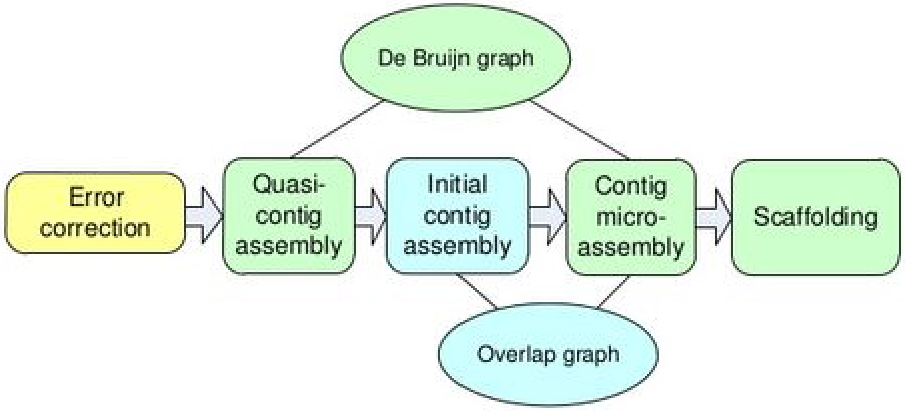
Understand the software platforms available :

- often an implementation of a more generic algorithm

Ex. : Overlap-Layout-Consensus assembler optimised for long sequences vs. Graphs of de Bruijn for short sequences

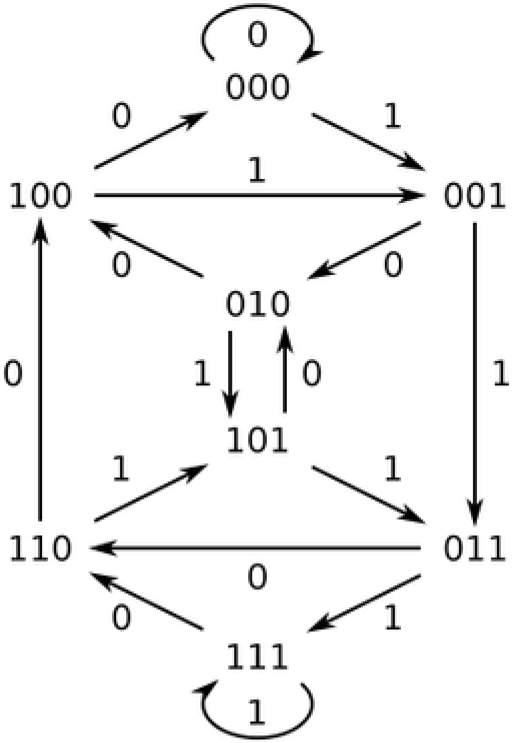


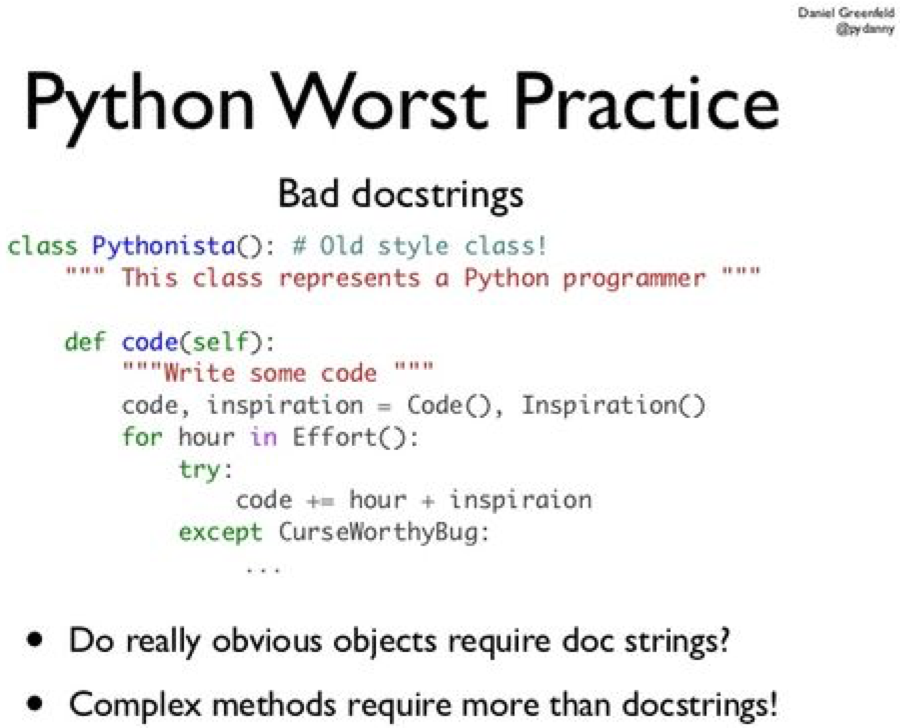
Algorithm



But you are a scientist not a programmer -> hence python
- development methodology at least (script, pipeline, software, tests on known data)
- technical doc along with biological knowledge (your added value)

Software engineering :
- versioning (Github, Subversion)
- README
- open data, open science, open source, reproductibility
- lab notebook 2.0




Daniel Greenfeld
@pydanny

Python Worst Practice

Bad docstrings

```
class Pythonista(): # Old style class!
    """ This class represents a Python programmer """

    def code(self):
        """Write some code """
        code, inspiration = Code(), Inspiration()
        for hour in Effort():
            try:
                code += hour + inspiraion
            except CurseWorthyBug:
                ...
```

- Do really obvious objects require doc strings?
- Complex methods require more than docstrings!

Be cautious with results and data : biological interpretation
- False positive, False negative, p-value, correlation vs. Causality, bias in data

https://politicalmethodology.wordpress.com/2013/04/01/p-values-are-possibly-biased-estimates-of-the-null-probability

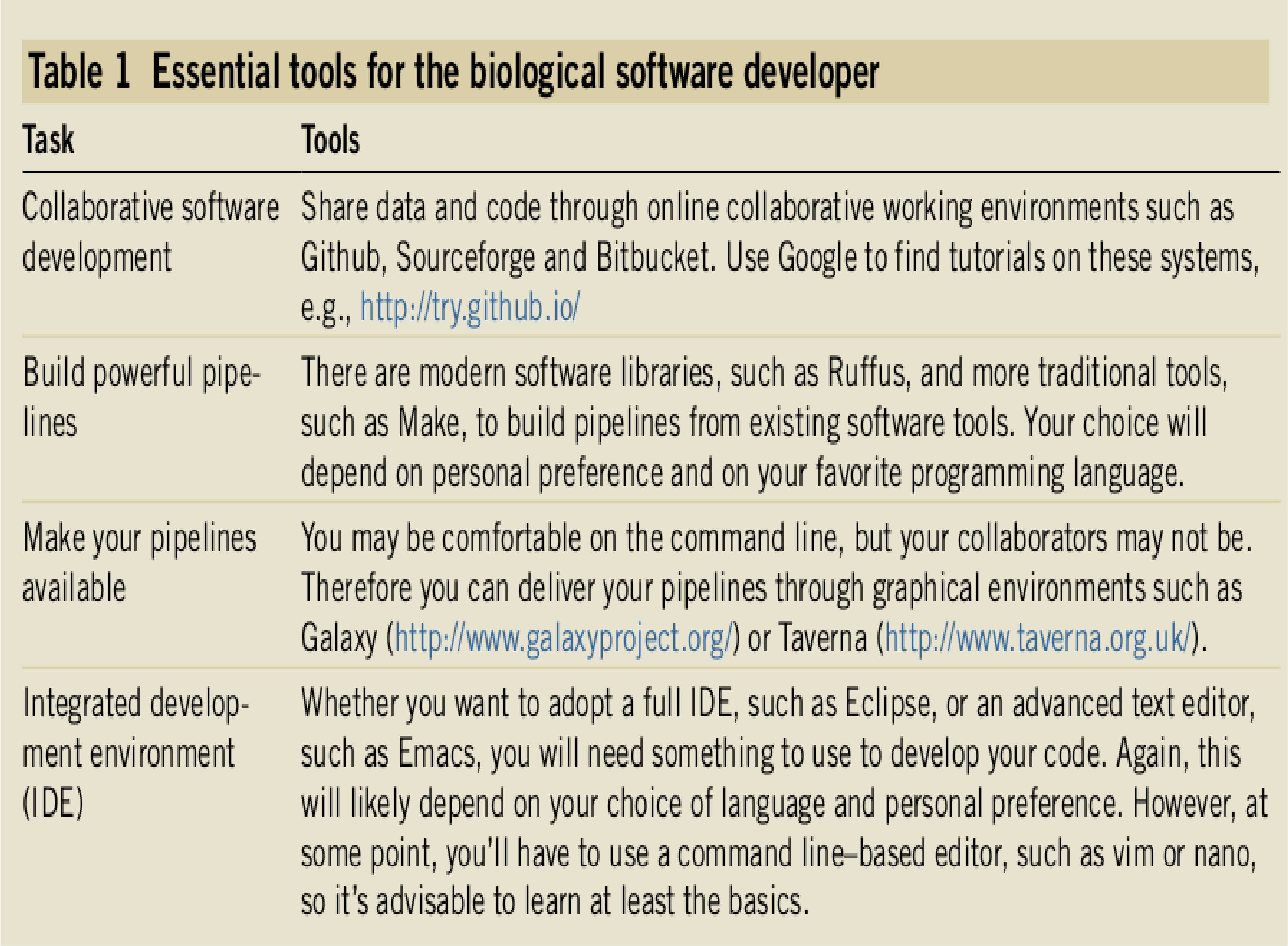


Table 1 Essential tools for the biological software developer

Task	Tools
Collaborative software development	Share data and code through online collaborative working environments such as Github, Sourceforge and Bitbucket. Use Google to find tutorials on these systems, e.g., http://try.github.io/
Build powerful pipelines	There are modern software libraries, such as Ruffus, and more traditional tools, such as Make, to build pipelines from existing software tools. Your choice will depend on personal preference and on your favorite programming language.
Make your pipelines available	You may be comfortable on the command line, but your collaborators may not be. Therefore you can deliver your pipelines through graphical environments such as Galaxy (http://www.galaxyproject.org/) or Taverna (http://www.taverna.org.uk/).
Integrated development environment (IDE)	Whether you want to adopt a full IDE, such as Eclipse, or an advanced text editor, such as Emacs, you will need something to use to develop your code. Again, this will likely depend on your choice of language and personal preference. However, at some point, you'll have to use a command line–based editor, such as vim or nano, so it's advisable to learn at least the basics.

Recently discussed at QBI 2019 http://www.biii.eu

Table 2 Useful resources for learning

Type of information	Relevant URLs
MOOCs (massive open online courses)	These are very popular at the moment and offer free training over the internet. Coursera (https://www.coursera.org/), Udacity (https://www.udacity.com/), edX (https://www.edx.org/) and the Kahn Academy (https://www.khanacademy.org/) have a range of courses relevant to bioinformatics, genomics, computing, statistics and modeling.
Learning to code	Codecademy (http://www.codecademy.com/) and Code School (https://www.codeschool.com/) are not specific to biology but do offer simple ways to learn how to code. For a more biological perspective, "Python for biologists" (http://pythonforbiologists.com/) is always popular. For examples of best practices visit http://software-carpentry.org/.
Bioinformatics problem solving	Learn bioinformatics through problem solving and pit your wits against others at http://www.rosalind.info.
Web forums	These are essential when you start out—ask questions and receive answers from experts at http://www.seqanswers.com/ and http://www.biostars.org/.
International organizations	GOBLET is the global organization for bioinformatics learning education and training (http://www.mygoblet.org/), and ELIXIR is a European organization set up to provide an infrastructure, including training, for life sciences information (http://www.elixir-europe.org/).
Blogs and lists	A variety of blogs and lists exist online that detail computational biology courses, such as http://stephenturner.us/p/edu and http://ged.msu.edu/angus/bioinformatics-courses.html.

So biotechs : what's up ?

NGS since 2012 :

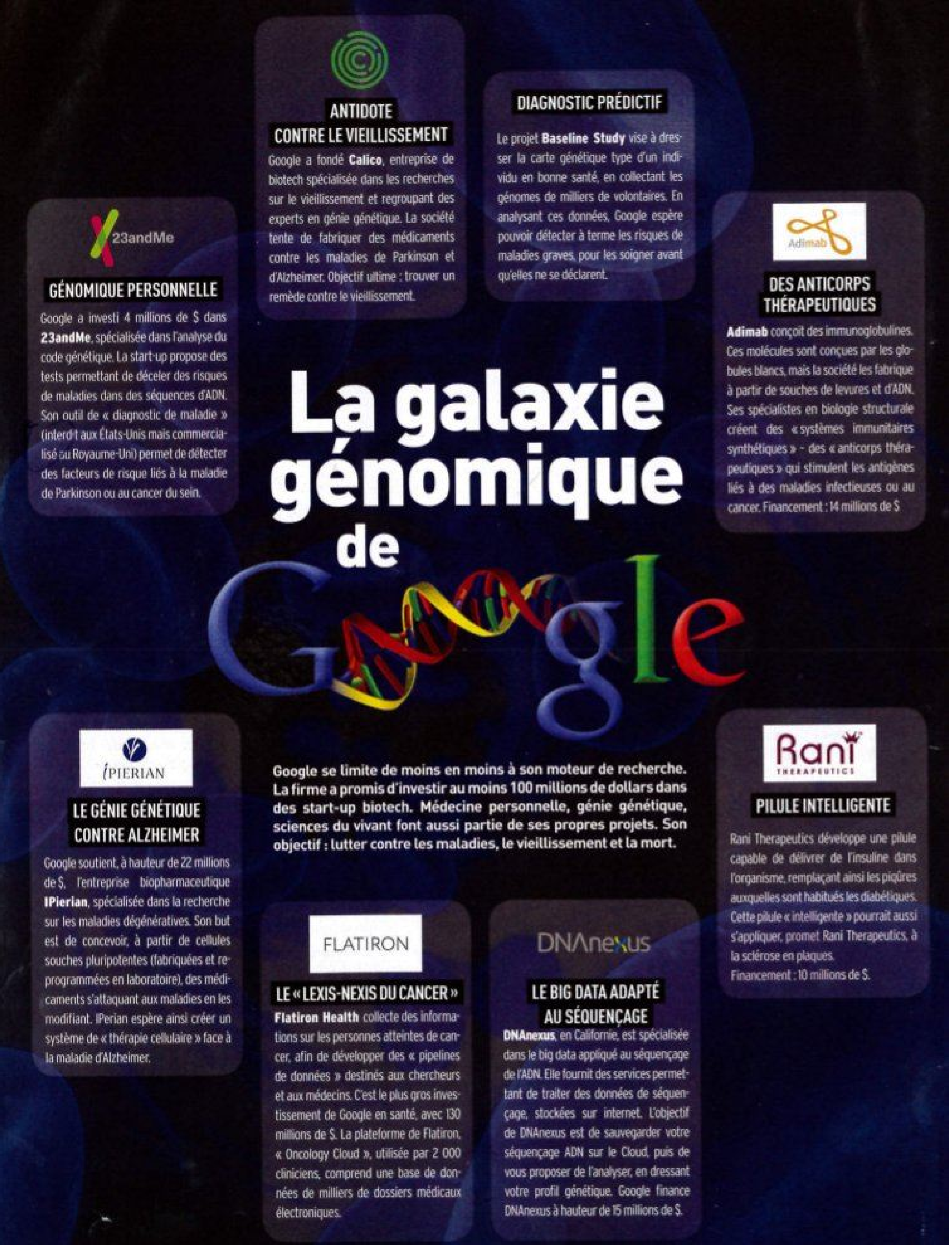
- 1st human genome : 15 years of research effort, 2.7 billion dollars

- Gain of 10^5 for megabase sequencing and 10^4 for a whole genome (2015) (few thousands dollars and hours in 2015 vs 100 million in 2001)

Post-genomic : complexity

-biohacking (DIY bio), intelligent virus (*Microbesoft* ?), synthesis biology (see Master FdV) (GMO 2.0?) (market 1 000 billion dollars in 2025, Source OCDE 2015)

- ethical questions: NGO ETC group (http://www.etcgroup.org/fr), Fondation Sciences Citoyennes -> your job as well



La galaxie génomique de Google

Google se limite de moins en moins à son moteur de recherche. La firme a promis d'investir au moins 100 millions de dollars dans des start-up biotech. Médecine personnelle, génie génétique, sciences du vivant font aussi partie de ses propres projets. Son objectif : lutter contre les maladies, le vieillissement et la mort.

GÉNOMIQUE PERSONNELLE

Google a investi 4 millions de $ dans **23andMe**, spécialisée dans l'analyse du code génétique. La start-up propose des tests permettant de déceler des risques de maladies dans des séquences d'ADN. Son outil de « diagnostic de maladie » (interd-t aux États-Unis mais commercialisé au Royaume-Uni) permet de détecter des facteurs de risque liés à la maladie de Parkinson ou au cancer du sein.

ANTIDOTE CONTRE LE VIEILLISSEMENT

Google a fondé **Calico**, entreprise de biotech spécialisée dans les recherches sur le vieillissement et regroupant des experts en génie génétique. La société tente de fabriquer des médicaments contre les maladies de Parkinson et d'Alzheimer. Objectif ultime : trouver un remède contre le vieillissement.

DIAGNOSTIC PRÉDICTIF

Le projet **Baseline Study** vise à dresser la carte génétique type d'un individu en bonne santé, en collectant les génomes de milliers de volontaires. En analysant ces données, Google espère pouvoir détecter à terme les risques de maladies graves, pour les soigner avant qu'elles ne se déclarent.

DES ANTICORPS THÉRAPEUTIQUES

Adimab conçoit des immunoglobulines. Ces molécules sont conçues par les globules blancs, mais la société les fabrique à partir de souches de levures et d'ADN. Ses spécialistes en biologie structurale créent des « systèmes immunitaires synthétiques » – des « anticorps thérapeutiques » qui stimulent les antigènes liés à des maladies infectieuses ou au cancer. Financement : 14 millions de $

LE GÉNIE GÉNÉTIQUE CONTRE ALZHEIMER

Google soutient, à hauteur de 22 millions de $, l'entreprise biopharmaceutique **IPierian**, spécialisée dans la recherche sur les maladies dégénératives. Son but est de concevoir, à partir de cellules souches pluripotentes (fabriquées et reprogrammées en laboratoire), des médicaments s'attaquant aux maladies en les modifiant. IPerian espère ainsi créer un système de « thérapie cellulaire » face à la maladie d'Alzheimer.

LE « LEXIS-NEXIS DU CANCER »

Flatiron Health collecte des informations sur les personnes atteintes de cancer, afin de développer des « pipelines de données » destinés aux chercheurs et aux médecins. C'est le plus gros investissement de Google en santé, avec 130 millions de $. La plateforme de Flatiron, « Oncology Cloud », utilisée par 2 000 cliniciens, comprend une base de données de milliers de dossiers médicaux électroniques.

LE BIG DATA ADAPTÉ AU SÉQUENÇAGE

DNAnexus, en Californie, est spécialisée dans le big data appliqué au séquençage de l'ADN. Elle fournit des services permettant de traiter des données de séquençage, stockées sur internet. L'objectif de DNAnexus est de sauvegarder votre séquençage ADN sur le Cloud, puis de vous proposer de l'analyser, en dressant votre profil génétique. Google finance DNAnexus à hauteur de 15 millions de $.

PILULE INTELLIGENTE

Rani Therapeutics développe une pilule capable de délivrer de l'insuline dans l'organisme, remplaçant ainsi les piqûres auxquelles sont habitués les diabétiques. Cette pilule « intelligente » pourrait aussi s'appliquer, promet Rani Therapeutics, à la sclérose en plaques.
Financement : 10 millions de $.

Socialter 2015
Start-up :

Transgène : virus intelligents contre cancer poumon et foie

Cellectis : cellules immunitaires génétiquement modifiées (France avec Pfizer)

Amyris : fabrication artificielle de l'artémisinine (anti-paludique) (US avec Sanofi)

Global Bioenergies : biocarburant (Evry avec Audi)

Abolis Biotechnologies : CAO de microorganismes

Hyasynth Bio : optiliser la production de cannabinoïdes à visée médicale (Canada)

But for sure a very wide spectrum of disciplines around math, informatics, biology, chemestry etc.

Computational biophysics & Structural Biology

Modeling of Regulatory, signaling and metabolic networks

Pattern Recognition and Machine Learning

Data Mining / Graph Mining / Sequence Mining

Functional Genomics

Molecular Interaction **Networks** / Systems Biology + Structural Biology

Prediction of Protein-Protein and Protein-DNA interactions

Gene Expression Analysis & Prediction of Regulatory Network Structure

Study of Complex Inherited Traits

Image Analysis & Interpretation

Biomedical Ontology Development

Knowledge Extraction From Scientific Litterature & Medical Reports

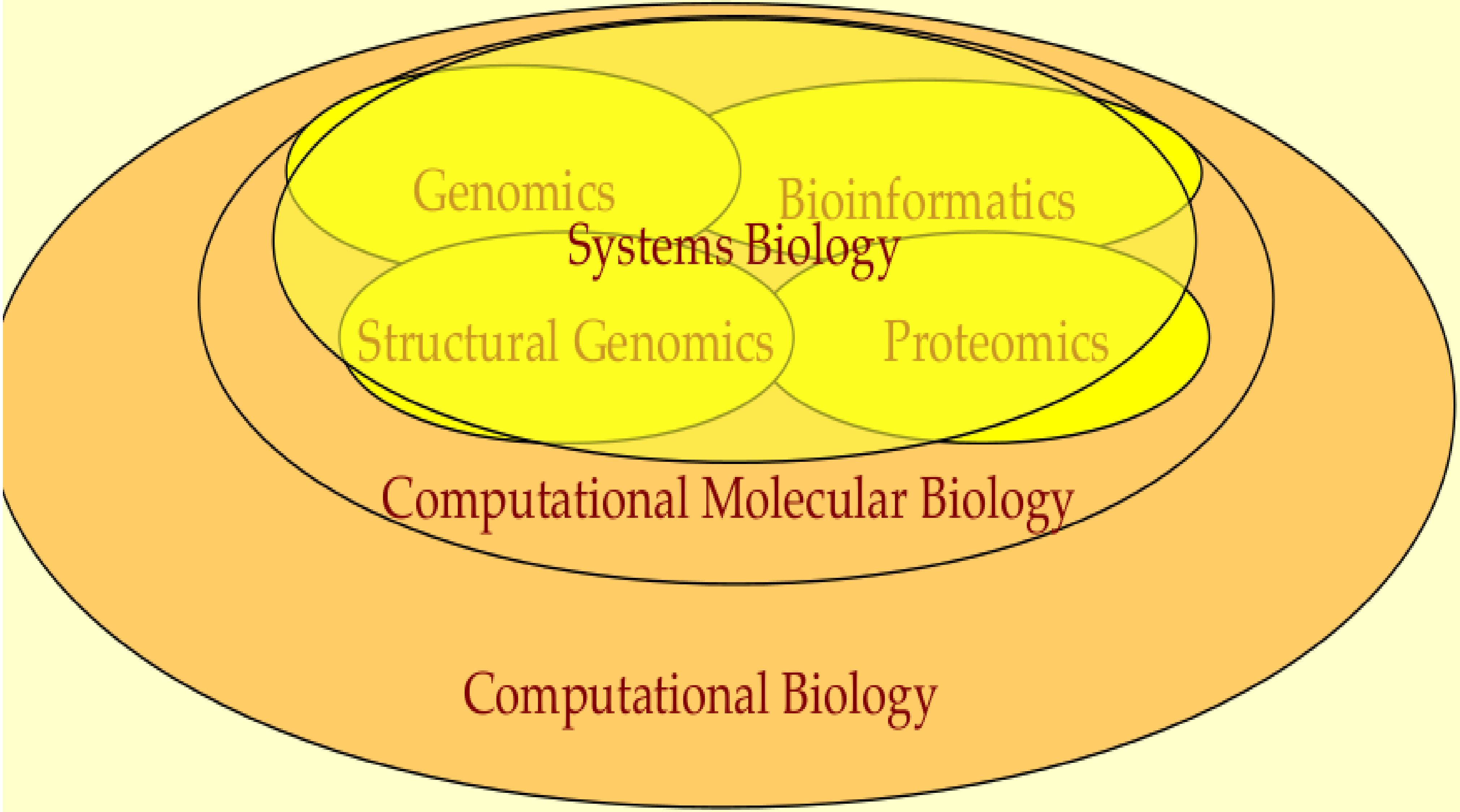
Evidence Integration

Protein Structure Modeling

Phylogenetic Tree Construction

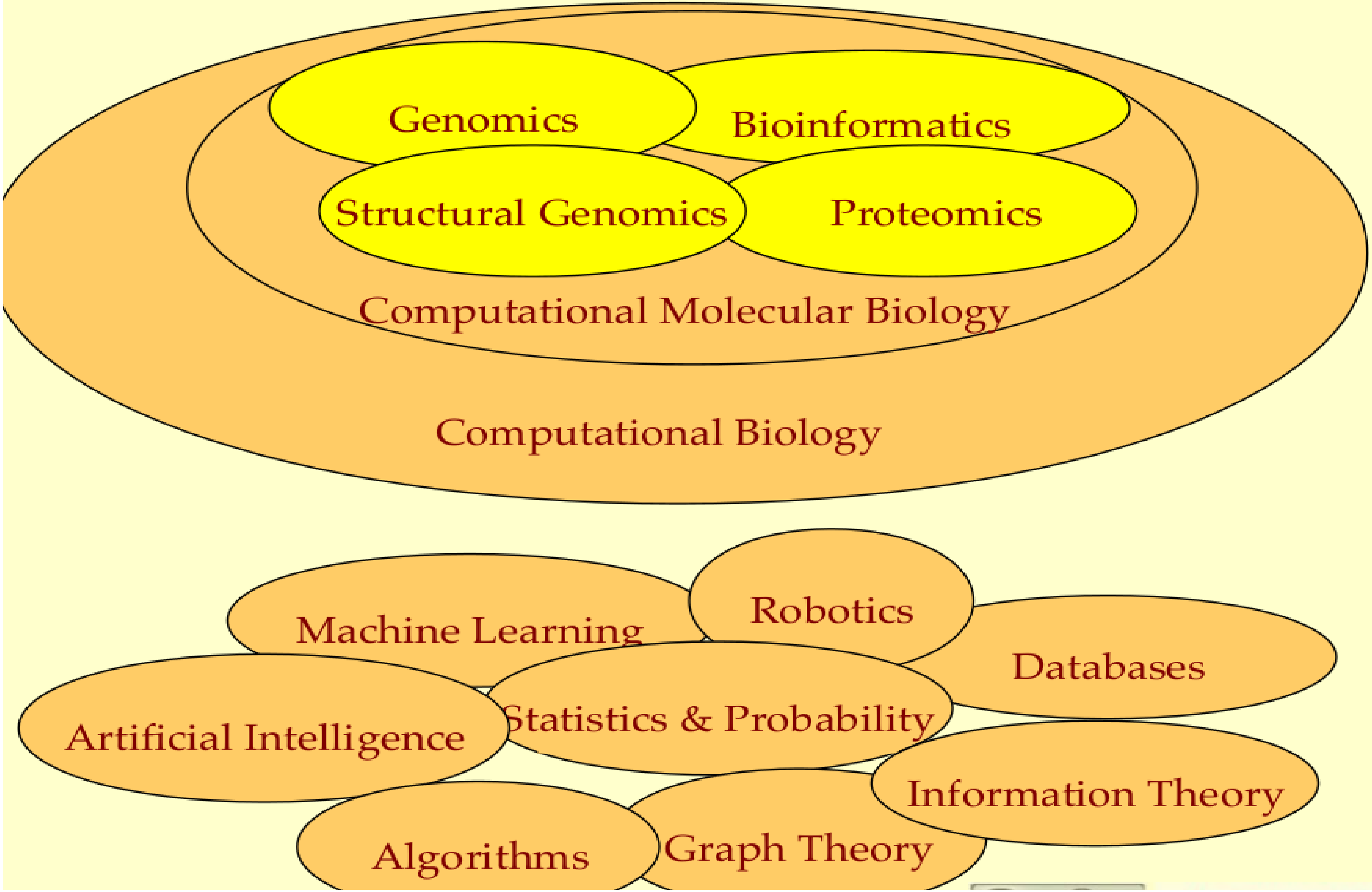
Synthetic Biology

Computational Biology



Give me a definition of Systems Biology.

Genomics, Bioinformatics & Computational Biology

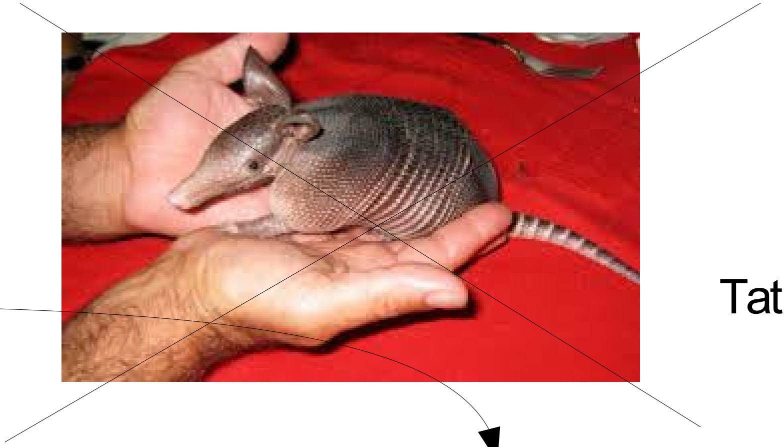


Biology and *computer sciences : a perfect paradigm for interdisciplinarity*

Human genetic variability : 0.1 % (vs. 0.2 % in chimpanzee)

Mean length of a protein sequence : 300-500 amino-acids
The longuest : Human titin protein with 34 350 amino acids, fiber muscular elasticity

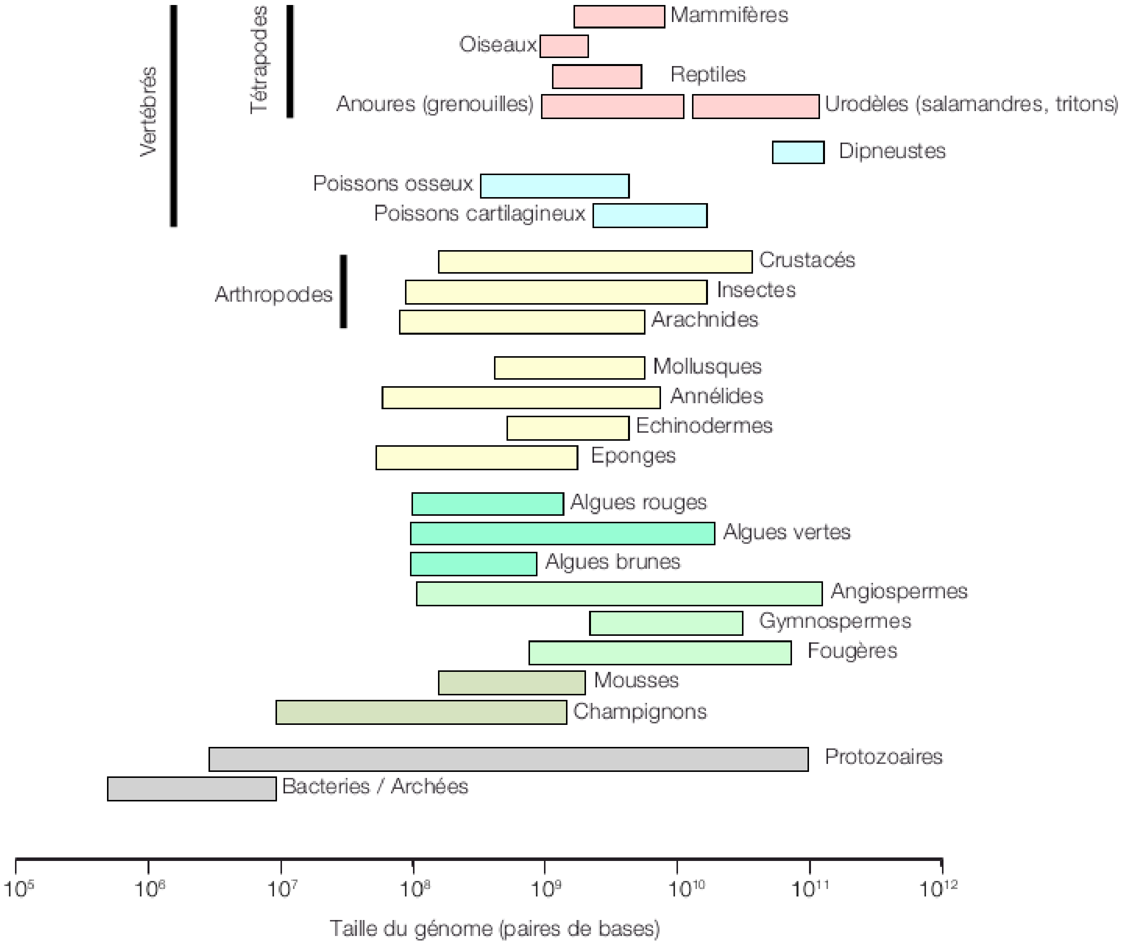
pangolin

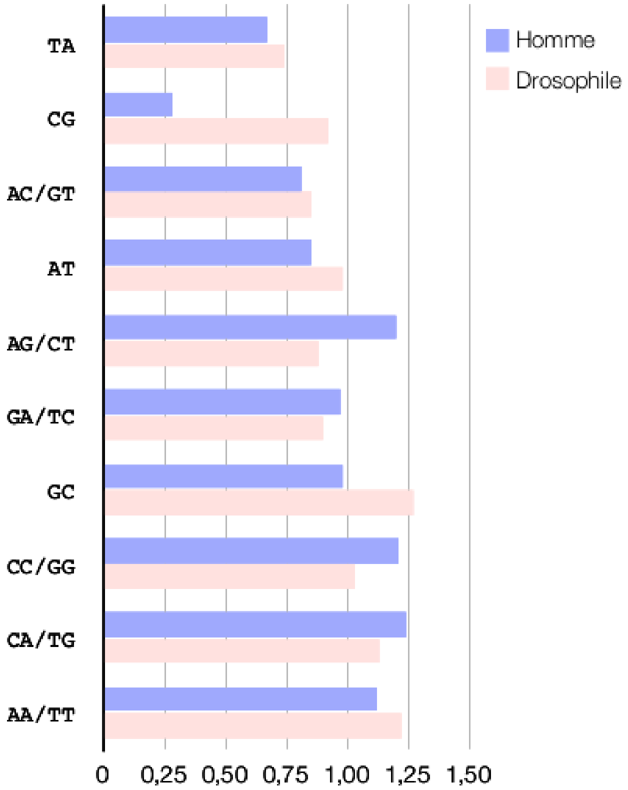


Tatou / armadillo

feline

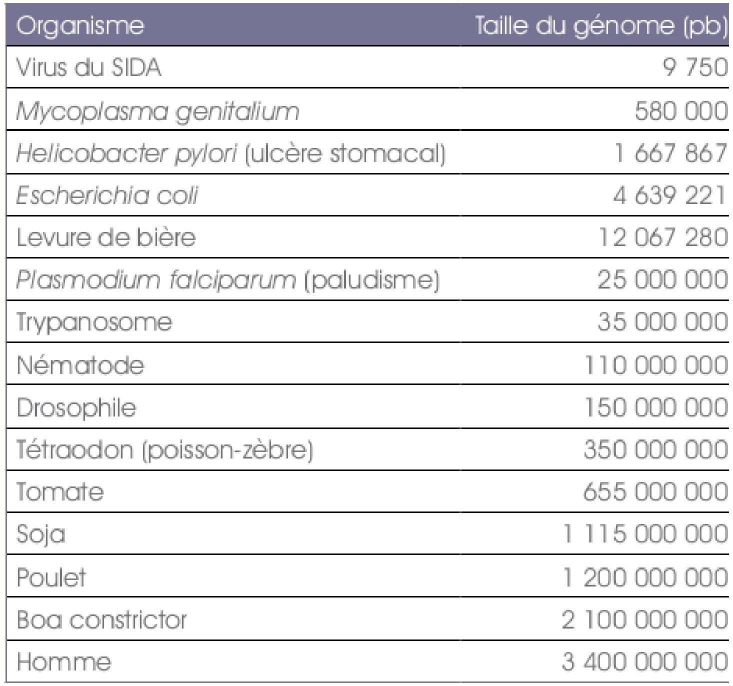
Biology and *computer sciences : perfect paradigm for interdisciplinarity*



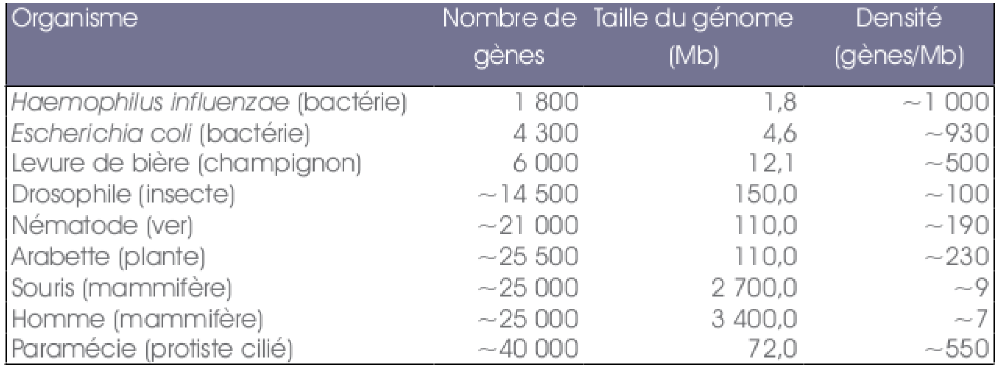


Which species has the longuest genome ?

Biology and *computer sciences : perfect paradigm for interdisciplinarity*

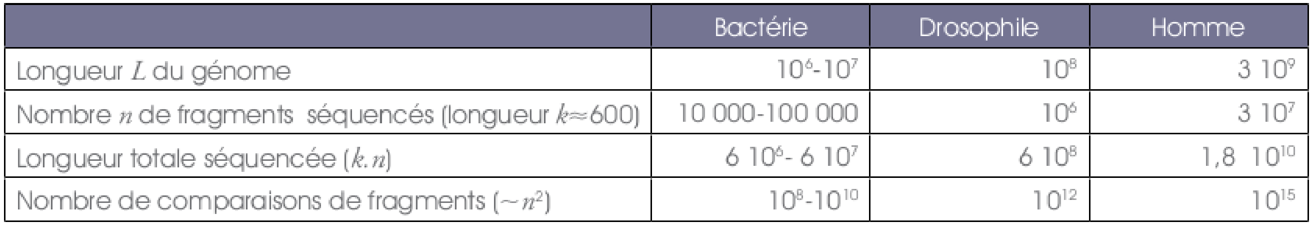


Organisme	Taille du génome (pb)
Virus du SIDA	9 750
Mycoplasma genitalium	580 000
Helicobacter pylori (ulcère stomacal)	1 667 867
Escherichia coli	4 639 221
Levure de bière	12 067 280
Plasmodium falciparum (paludisme)	25 000 000
Trypanosome	35 000 000
Nématode	110 000 000
Drosophile	150 000 000
Tétraodon (poisson-zèbre)	350 000 000
Tomate	655 000 000
Soja	1 115 000 000
Poulet	1 200 000 000
Boa constrictor	2 100 000 000
Homme	3 400 000 000



Organisme	Nombre de gènes	Taille du génome (Mb)	Densité (gènes/Mb)
Haemophilus influenzae (bactérie)	1 800	1,8	~1 000
Escherichia coli (bactérie)	4 300	4,6	~930
Levure de bière (champignon)	6 000	12,1	~500
Drosophile (insecte)	~14 500	150,0	~100
Nématode (ver)	~21 000	110,0	~190
Arabette (plante)	~25 500	110,0	~230
Souris (mammifère)	~25 000	2 700,0	~9
Homme (mammifère)	~25 000	3 400,0	~7
Paramécie (protiste cilié)	~40 000	72,0	~550

Paris japonica
Fleur de « parisette »
> 150 10^9 pb



	Bactérie	Drosophile	Homme
Longueur L du génome	10^6-10^7	10^8	3 10^9
Nombre n de fragments séquencés (longueur $k \approx 600$)	10 000-100 000	10^6	3 10^7
Longueur totale séquencée ($k.n$)	6 10^6- 6 10^7	6 10^8	1,8 10^{10}
Nombre de comparaisons de fragments ($\sim n^2$)	10^8-10^{10}	10^{12}	10^{15}

Salma Barkaoui

Nicolas Loménie
Université Paris Descartes,
Systèmes Intelligents de Perception
http://w3.mi.parisdescartes.fr/sip-lab/

Nicolas.lomenie@parisdescartes.fr

Cédric Gageat (now ANEO)
Institut de Biologie Physico-Chimique
Laboratoire de Biochimie Théorique
http://www-lbt.ibpc.fr

Urszula Czerwińska (now Post Doc)
Institut Pasteur

Systems Biology Lab
http://proteomics.fr/Sysbio/