Introduction à la bioinformatique (SSV3U15)

Chapitre 1. Présentation du cours

Jacques van Helden (Aix-Marseille Université)
ORCID 0000-0002-8799-8584

Equipe pédagogique de l'UE (2024-2025)

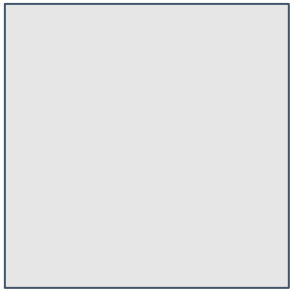
Jacques van Helden
Responsable UE

Bénédicte Wirth
Responsable site Aix

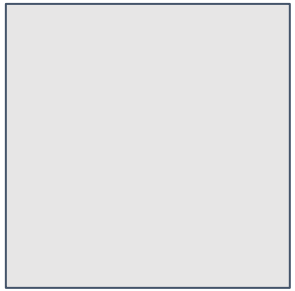
Emese Meglecz
Responsable site Saint-Charles

Aitor Gonzalez
Responsable site Luminy

Andreas Zanzoni



Yvan Perez



Alexandre Lutz

Juliette Patricio

Gael Chambonnier

Loréna Quatreville

Contenu de ce chapitre

1. Biologie et données massives
2. Quelques jalons historiques : données, modèles et découvertes en biologie
3. Organisation du cours et modalités de contrôle des connaissances
4. Réponses aux questions

La biologie, une science des données

Un changement d'échelle

- Au tournant du 20è au 21è siècle, la biologie s'est orientée vers une science qui s'appuie sur des données de plus en plus massives.
- Types de données
  - De séquençage
  - Protéomiques (quantification des protéines)
  - Métabolomiques (quantification des petites molécules)
  - Structure tridimensionnelle des protéines
  - Images
  - Phénotypiques (agriculture)
  - De santé (médecine)
  - …

Biologie et numérique

- Pour donner du sens à ces données, il faut combiner les concepts biologiques aux outils informatiques, mathématiques, statistiques, et mobiliser des moyens importants de calcul et de stockage des données.
- Exemples de domaines d'application
  - Génomique
  - Biologie évolutive
  - Médecine personnalisée
  - Biodiversité et environnement
  - Approches intégrative 'One Health"
- … en gros, tous les domaines de recherche et applications de la biologie

Quelques jalons historiques :
données, modèles et découvertes en biologie

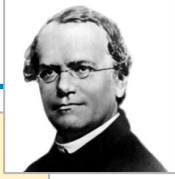
Quelques jalons historiques: données, modèles et découvertes en biologie

Les diapo suivantes présentent quelques jalons historiques de la biologie, en montrant son évolution vers une science qui s'appuie largement sur les données.

Nous faisons ici un rapide tour d'horizon, et nous reviendrons sur les exemples de façon plus approfondie lors des prochaines séances.

Ne vous inquiétez donc pas si les détails ne sont pas présentés, il ne s'agit que d'un "*teaser*" des épisodes suivants, où les explications seront fournies.

Premières lois biologiques basées sur une modélisation mathématique des données



- 1866 : premières **lois de l'hérédité** (Mendel)
- 1901 : redécouverte des lois de Mendel

EXPERIMENTS IN PLANT HYBRIDIZATION (1865)

GREGOR MENDEL

Read at the February 8th, and March 8th, 1865, meetings
of the Brünn Natural History Society

Mendel, Gregor. 1866. Versuche über Plflanzenhybriden. *Verhandlungen des naturforschenden Vereines in Brünn, Bd. IV für das Jahr 1865*, Abhandlungen, 3–47.

Generation	*A*	*Aa*	*a*	Ratios *A* : *Aa* : *a*
1	1	2	1	1 : 2 : 1
2	6	4	6	3 : 2 : 3
3	28	8	28	7 : 2 : 7
4	120	16	120	15 : 2 : 15
5	496	32	496	31 : 2 : 31
n				$2^n - 1$: 2 : $2^n - 1$

L'importance des nombres en biologie n'est pas nouvelle

En 1866, Gregor Mendel publie un volumineux article dans lequel il décrit en détail les résultats de ses expériences de croisements entre différentes variétés de pois. Au fil des générations, il dénombre les individus présentant diverses combinaisons de caractères.

En analysant ces données, il identifie des régularités numériques et en dérive **trois lois permettant de prédire les fréquences des caractères qualitatifs au fil des générations**:

1. Loi d'uniformité des caractères à la première génération.
2. Loi de ségrégation des caractères (*illustration*).
3. Loi d'indépendance des caractères.

Il s'agit sans doute de la première découverte en biologie reposant sur la modélisation mathématique de données expérimentales quantitatives.

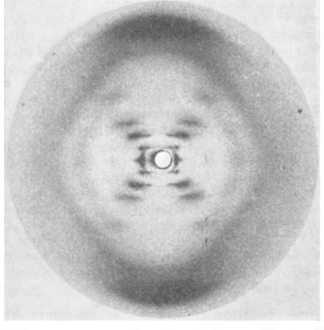
Cette publication, aujourd'hui considérée comme pionnière, passe totalement inaperçue à son époque. Les lois de Mendel seront redécouvertes en 1901, indépendamment, par trois groupes de chercheurs.

Mendel, G. (1866). Experiments in plant hybridisation.(Reprinted 1965, English translation and footnotes by W. Bateson, Cambridge, MA.

L'ADN, vecteur de l'information héréditaire

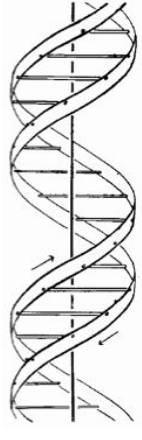
- 1866 : premières lois de l'hérédité (Mendel)
- 1901 : redécouverte des lois de Mendel
- 1910-1915: les **chromosomes** sont le **support de l'hérédité** (Morgan)
- 1944: l'**ADN** est le **support de l'hérédité** (Avery)
- 1953 : **structure de l'ADN**, la double hélice (Watson & Crick; Franklin & Gosling)

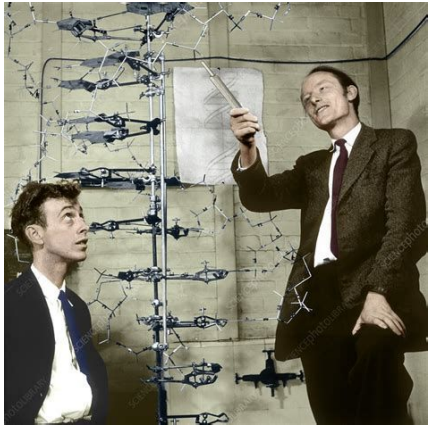
Image cristallographique l'ADN, par diffraction de rayons X (R.E. Franklin and R. Gosling, 1953)



Sodium deoxyribose nucleate from calf thymus. Structure *B*

Modèle de la structure de l'ADN (Watson and Crick, 1953b)





- Franklin,R.E. and Gosling,R.G. (1953) Molecular configuration in sodium thymonucleate. doi.org/10.1038/171740a0
- WATSON,J.D. and CRICK,F.H. (1953a) The structure of DNA. Cold Spring Harb Symp Quant Biol, 18, 123–131.doi.org/10.1101/sqb.1953.018.01.020
- Watson,J. and Crick,F. (1953b) Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature, 171, 737–738. doi.org/10.1038/171737a0
- WATSON,J.D. and CRICK,F.H. (1953c) Genetical implications of the structure of deoxyribonucleic acid. Nature, 171, 964–967. doi.org/10.1038/171964b0

Premières structures de protéines - le lien séquence - structure - fonction

- 1866 : premières lois de l'hérédité (Mendel)
- 1901 : redécouverte des lois de Mendel
- 1910-1915: les chromosomes sont le support de l'hérédité (Morgan)
- 1944: l'ADN est le support de l'hérédité (Avery)
- 1953 : structure de l'ADN, la double hélice (Watson & Crick; Franklin)
- 1958-1960 : **premières structures de protéines** (Kendrew, Perutz)
  - Les figures montrent les photos de modèles tridimensionnels qui illustraient les publications originales.
  - Pendant les cours et TP, nous présenterons les approches bioinformatiques d'analyse et de visualisation des structures de protéines.

Structure de la myoglobine de cachalot (Kendrew, 1958)

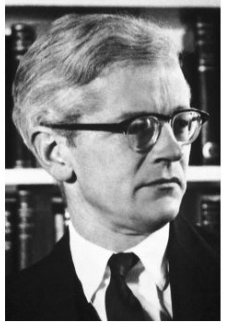
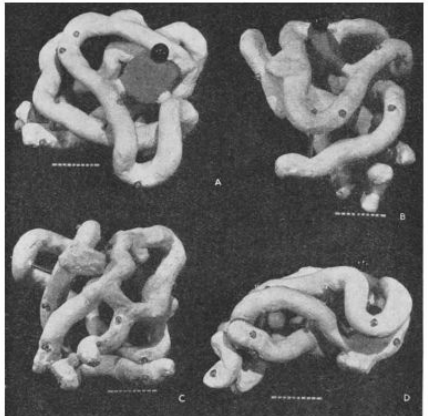


Photo from the Nobel Foundation archive.
John Cowdery Kendrew



Structure de l'hémoglobine (Perutz, 1960)

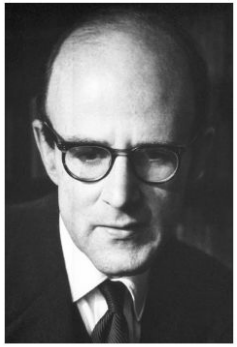
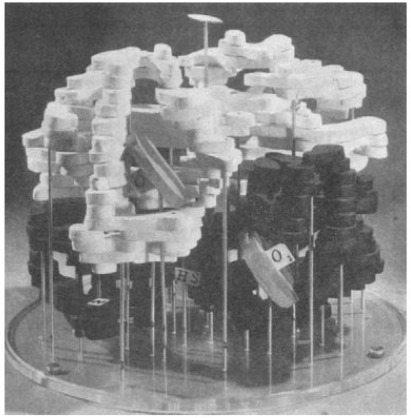


Photo from the Nobel Foundation archive.
Max Ferdinand Perutz



1. Kendrew, J. C. et al. A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. Nature 181, 662–666 (1958).
2. Perutz, M. F. et al. Structure of haemoglobin: a three-dimensional Fourier synthesis at 5.5-A. resolution, obtained by X-ray analysis. Nature 185, 416–422 (1960).
https://www.nobelprize.org/prizes/chemistry/1962/summary/

Le code universel du vivant – Correspondances entre nucléotides et acides aminés

- 1866 : premières lois de l'hérédité (Mendel)
- 1901 : redécouverte des lois de Mendel
- 1910-1915: les chromosomes sont le support de l'hérédité (Morgan)
- 1944: l'ADN est le support de l'hérédité (Avery)
- 1953 : structure de l'ADN, la double hélice (Watson & Crick; Franklin)
- 1958-1960 : premières structures de protéines (Kendrew, Perutz)
- 1961: découverte du code génétique (Nirenberg, Matthaei)

Exemples de lecture du tableau

- CAU ⟶ Arginine
- CCU ⟶ Proline
- ATG ⟶ méthionine
  (également codon start le plus fréquent)
- UAA, UAG ou UGA -> codons stop

LE CODE GENETIQUE

ARN messager Codon : première base azotée	ARN messager Codon : deuxième base azotée U	C	A	G	ARN messager Codon : troisième base azotée
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

The Nobel Prize in Physiology or Medicine 1968


Photo from the Nobel Foundation archive.
Robert W. Holley
Prize share: 1/3


Photo from the Nobel Foundation archive.
Har Gobind Khorana
Prize share: 1/3


Photo from the Nobel Foundation archive.
Marshall W. Nirenberg
Prize share: 1/3

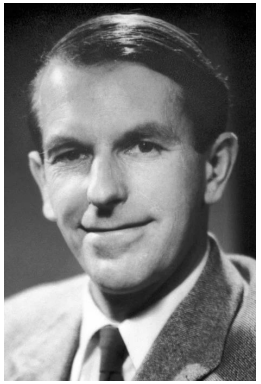
The Nobel Prize in Physiology or Medicine 1968 was awarded jointly to Robert W. Holley, Har Gobind Khorana and Marshall W. Nirenberg "for their interpretation of the genetic code and its function in protein synthesis"

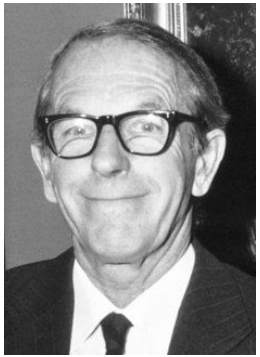
https://www.nobelprize.org/prizes/medicine/1968/summary/

Le séquençage de l'ADN

- 1866 : premières lois de l'hérédité (Mendel)
- 1901 : redécouverte des lois de Mendel
- 1910-1915: les chromosomes sont le support de l'hérédité (Morgan)
- 1944: l'ADN est le support de l'hérédité (Avery)
- 1953 : structure de l'ADN, la double hélice (Watson & Crick; Franklin)
- 1958-1960 : premières structures de protéines (Kendrew, Perutz)
- 1961: découverte du code génétique (Nirenberg, Matthaei)
- 1977: méthode de séquençage de l'ADN (Sanger)
- Note: Frederick Sanger a obtenu 2 prix Nobel
  - 1958 pour son travail sur la structure de l'insuline
  - 1977, Gilbert & Sanger pour "*leur contribution à la détermination de la séquence des bases des acides nucléiques*"

FIG. 2. Autoradiograph from an experiment using fragment R4 as primer on the complementary strand of φX174 DNA. Conditions were as in Fig. 1 with the following exceptions: ddCTP was used as inhibitor instead of araCTP. After incubation of the solutions at room temperature for 15 min, 1 μl of 0.5 mM dATP and 1 μl of restriction enzyme *Hae* III (4 units/μl) were added and the solutions were incubated at 37° for 10 min. The *Hae* III cuts close to the *Hin*dII site and it was used because it was more readily available. The electrophoresis was on a 12% acrylamide gel at 40 mA for 14 hr. The top 10.5 cm of the gel is not shown.





Frederick Sanger

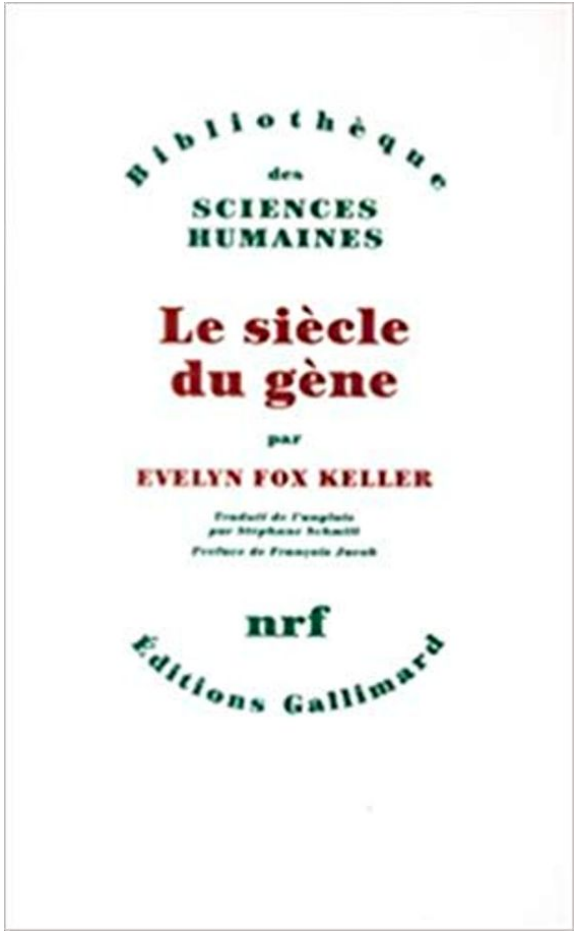
Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74, 5463–5467 (1977).
https://www.nobelprize.org/prizes/chemistry/1980/summary/
https://www.nobelprize.org/prizes/chemistry/1958/summary/

Du siècle du gène au siècle du génome

Dans son ouvrage « Le siècle du gène », l'historienne des sciences Evelyn Fox-Keller retrace l'histoire de la découverte des gènes, de leur fonction, des mécanismes moléculaires.

Elle termine le livre en soulignant que le 21ème siècle sera le siècle du génome.

Effectivement, depuis la fin des années 1990 une série de projets de séquençage ont été initiés, qui ont suscité un changement drastique de l'ensemble des approches en biologie.



Premiers génomes – Un génome dit “de référence” par espèce

- 1990-2000 : premiers projets de séquençage du génome d’organismes modèles: bactéries, levure du boulanger, drosophile, nématode, arabette, et … “le” génome humain

Nom d'espèce	Nom commun	Année	Taille du génome Mb	Nombre de gènes
Bactérie				
Mycoplasma genitalium	*Mycoplasma*	1995	0,6	481
Haemophilus influenzae	Bacille de Pfeiffer	1995	1,8	1 717
Escherichia coli	*Entérobactérie*	1997	4,6	4 289
Levures				
Saccharomyces cerevisiae	*Levure du boulanger*	1996	12	6 286
Animaux				
Caenorhabditis elegans	*Ver nématode*	1998	97	19 000
Drosophila melanogaster	*Mouche à vinaigre*	2000	165	16 000
Danio rerio	*Poisson zèbre*		1 527	18 957
Xenopus laevis	*Xénope (amphibien)*		1 511	18 023
Gallus gallus	*Poule*		2 961	16 736
Ornithorhynchus anatinus	*Ornithorynque*		1 918	17 951
Mus musculus	*Souris*	2002	3 421	23 493
Pan troglodytes	*Chimpanzé*		2 929	20 829
Homo sapiens	*Humain*	2001	3 200	21 528
Plantes				
Arabidopsis thaliana	*Arabette*	2001	120	27 000
Oryza sativa	*Riz*		390	37 544
Zea mais	*Maïs*		2 500	50 000
Triticum aestivum	*Blé*		16 000	
Lilium	*Lys*		120 000	
Psilotum nudum			250 000	

“Le” génome humain

- Compétition entre projet public et projet privé
- 2001 : première publication d’un génome humain
  - Version “brouillon”: 2001 (bonne couverture mais trous de séquençage)
- 2024: version parachevée

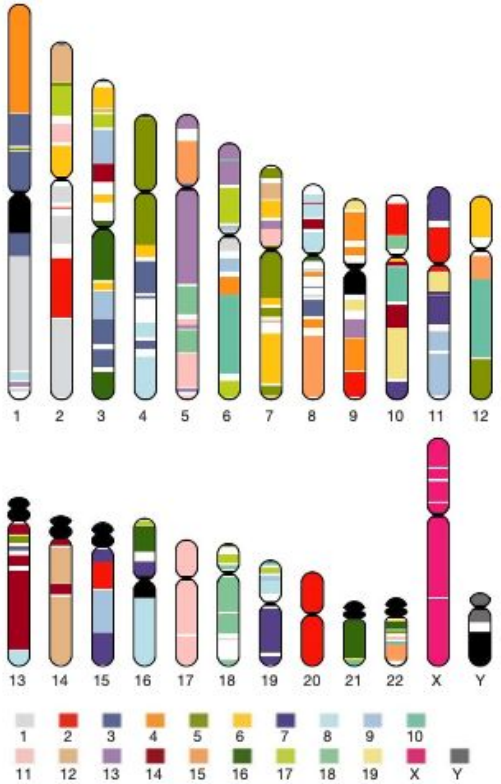


Figure 46 Conserved segments in the human and mouse genome. Human chromosomes, with segments containing at least two genes whose order is conserved in the mouse genome as colour blocks. Each colour corresponds to a particular mouse chromosome. Centromeres, subcentromeric heterochromatin of chromosomes 1, 9 and 16, and the repetitive short arms of 13, 14, 15, 21 and 22 are in black.

De la génomique à la génomique fonctionnelle

Le séquençage ne constitue qu'une toute première étape pour l'analyse des génomes.

Au terme d'un projet de séquençage, on obtient un "texte" formé des 4 lettres A, C, G, T (une par nucléotide), et il reste un énorme travail de décryptage pour pouvoir interpréter ce texte.

L'exemple ci-dessous montre un fragment de 1000 nucléotides du génome humain.

```
....CGATGCTCAAACATTTCAATTTTTTAGGTCAAAAATGCCTTAGGTTTAGCACAGCAATGTAGGTGCCAAACTC
ATCGCAGTGAATTGCAGGCGGGAGCAACAAGGACGCCTGCCTCCTTTCTGCCTGCTTTTTGCAATAGTCCGATTTGA
GAAGGGGACCCACGAGAGACACAAAATGCACGCCCCCACGCCACATCCTTTTTACCCCGCAATGGGTTAAGACTGTC
AACAGGCAGGCCACCTCGCAGCGTCCGCGGAGTTGCAGGCCCGCCCCCGCCAGGGTGTGGCGCTGTCCCCCTGGCGC
TGGGCGGGGGAGGAGGGGCGCGCGGCGGCCGAGGAGGGGCGCGCGGCGGCCGGGCGGGGCGAGCGGAGGCGAGTGGA
GGACGCGTAGACGCGCCGCGGTCCCCGCCTGCCGCTGCTCCGCCGCAGTCGCCGCTCCAGTCTATCCGGCACTAGGA
ACAGCCCCGAGCGGCGAGACGGTCCCCGCCATGTCTGCGGCCATGAGGGAGAGGTTCGACCGGTTCCTGCACGAGAA
GAACTGCATGACTGACCTTCTGGCCAAGCTCGAGGCCAAAACCGGCGTGAACAGGAGCTTCATCGCTCTTGGTGGGT
GGCCGGGGGTCGCCGCCGCTGGTAGGGCCACGGGAGCCGCCGCTGCCCCAGCTGCTGGGGAAGGAAGCAGGGAGAGG
ACTCGGGAAAGGTGGAGTCGGAGACAGACGGGACAAGCAGCATATTCAGGGATCAGGCTGGCCTCCCGGAAAGCGTG
GGCATCGGAGGACCCCGCGGGGGCTGCCCAGGCTGAGGGTCGCGGGGCTGGAGGGCAGCTGCGGCGCCGGGCGCTGG
CAGCTGGAAGGGCCAGCGCTGACGTATGTCTGCCCCGCGGCCCGGCGCCCTATTCCTGCTGTCCTGCGCGGTGGGCG
CGGACGGCGGGGCCCCTGCGGGCGGGCGCGTTGACGGAGGTACCCGGTCCTACCCGACCCTCCGTGGAGCTCCGCCC
GGAG....
```

Le génome complet comporte 3 milliards de nucléotides, 3 millions de fois plus grand.

Les premières questions qui se posent au terme du séquençage =

1. Où sont localisés les gènes ?
2. **Quelle est la fonction de ces gènes ?**

Drew Sheneman, New Jersey -- The Newark Star Ledger

Des génomes aux transcriptomes

Chez tous les êtres vivants l'expression des gènes fait l'objet d'un contrôle moléculaire à différents niveaux: transcription, maturation de l'ARN, traduction, post-traduction.

Une indication importante concernant la fonction des gènes est de savoir dans quelles conditions ils sont exprimés.

- Microbes: substrats disponibles, conditions environnementales, …
- Multicellulaires: spécificité tissulaire, stades du développement, réponse aux conditions internes et externe de l'organisme

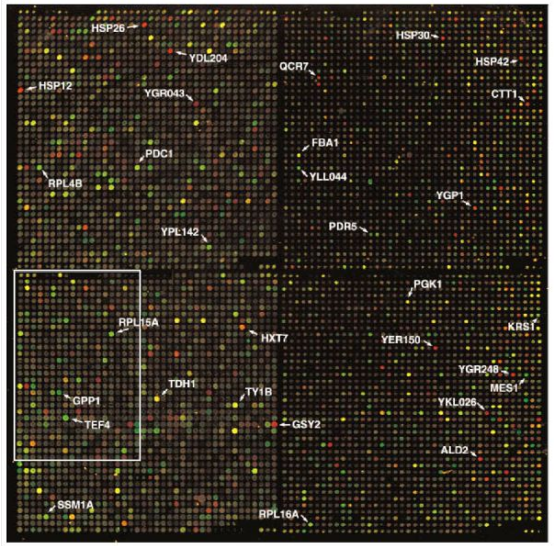
La transcriptomique consiste à mesurer simultanément l'expression de *tous* les gènes d'un échantillon prélevé sur un organisme dans des conditions particulières.

- 1997: premières approches de transcriptomiques par biopuces
- 2007: transcriptomique par séquençage massivement parallèle (RNA-seq)

La première biopuce transcriptomique (de Risi et al., 1997). Chacun des 6000 points lumineux correspond à un transcrit (ARN) de la levure du boulanger, *Saccharomyces cerevisiae*.

- L'intensité lumineuse est proportionnelle au niveau d'expression
- La couleur indique le sens de la régulation
  - Rouge: gènes sur-exprimés par rapport à l'échantillon témoin
  - Vert: gènes sous-exprimés
  - Jaune: gènes fortement exprimés dans les deux échantillons.

DeRisi et al. (1997), 10.1126/science.278.5338.680



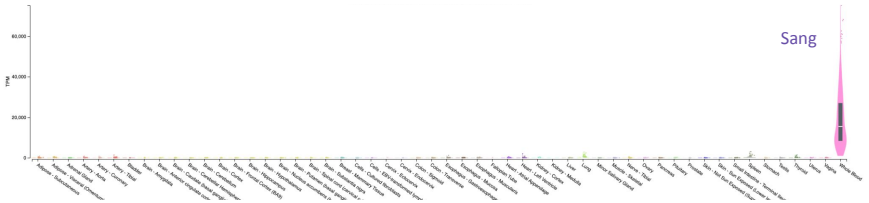
DeRisi, J. L., Iyer, V. R. & Brown, P. O. (1997). Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 278, 680-6.
doi.org/10.1126/science.278.5338.680

Dis-moi dans quels tissus tu t'exprimes, je te dirai qui tu es

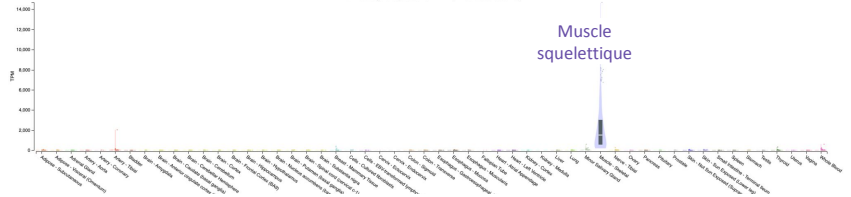
Le projet GTEX (Adult Genotype Expression)

- Collecte d'échantillons de 54 tissus chez 1000 individus
- Extraction de l'ARN
- Séquençage et quantification dans chaque tissu (RNA-seq)

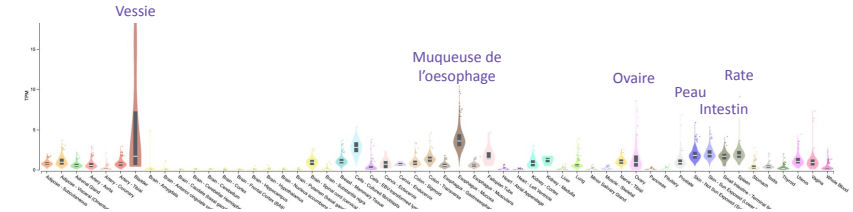
Exemples ci-contre: profils tissulaires d'expression pour quelques gènes illustratifs



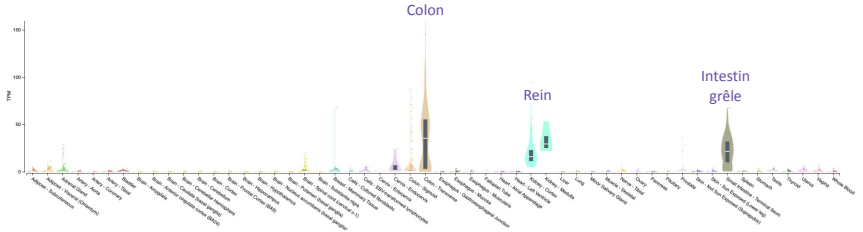
Gène HBA (chaîne alpha de l'hémoglobine)



Gène MYH1 (myoglobine)



HOX1A (gène de spécification segmentaire)



HOXB9 (gène de spécification segmentaire)

https://www.gtexportal.org/home/

Des génomes aux protéomes

Les effecteurs de la plupart des fonctions biologiques sont les protéines.

Les quantités de transcrits (ARN) ne sont qu'une indication approximative du niveau d'activité d'un gène dans une cellule, pour différentes raisons

- Régulation post-transcriptionnelle
- Régulation post-traductionnelle

Dès le milieu des années 1990, les biochimistes mettent au point des méthodes basées sur la spectrométrie de masse pour quantifier chaque protéine dans un échantillon, qui donnent naissance à la **protéomique** (caractérisation à large échelle des protéines présentes dans un échantillon).

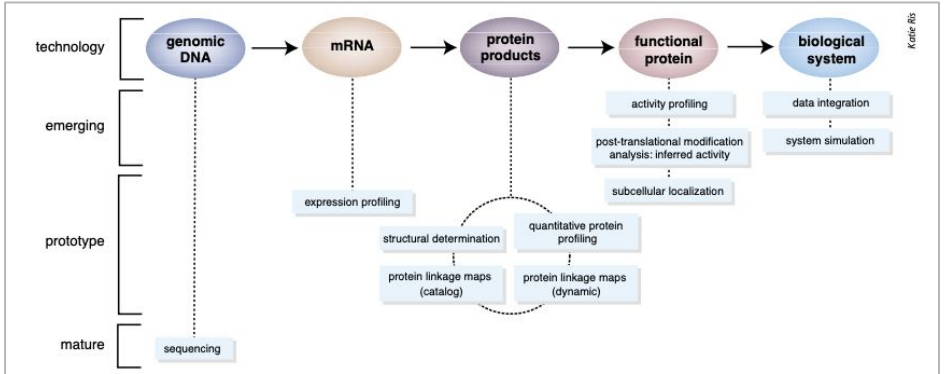


Fig. 2 The current status of proteomic technologies. The different data typically collected in proteomic research and the available technologies are listed. The relative maturity of the proteomic technologies and other key discovery science tools is apparent from the position of the respective technology on the graph.

Patterson et al. (2003). doi.org/10.1038/ng1106

Patterson, S. D. & Aebersold, R. H. Proteomics: the first decade and beyond. Nat Genet 33, 311–323 (2003). doi.org/10.1038/ng1106

Des protéomes aux interactomes

Une protéine n'agit généralement pas seule: les protéines interagissent

- De façon stable, en formant des complexes multimériques (plusieurs polypeptides)
- De façon transitoire, en établissant des liaisons temporaires qui modifient leur niveau d'activité

Au début des années 2000, plusieurs méthodes sont mises au point pour déterminer l'**interactome**, c'est-à-dire l'ensemble des interactions entre protéines d'un système biologique (organisme, tissu, échantillon).

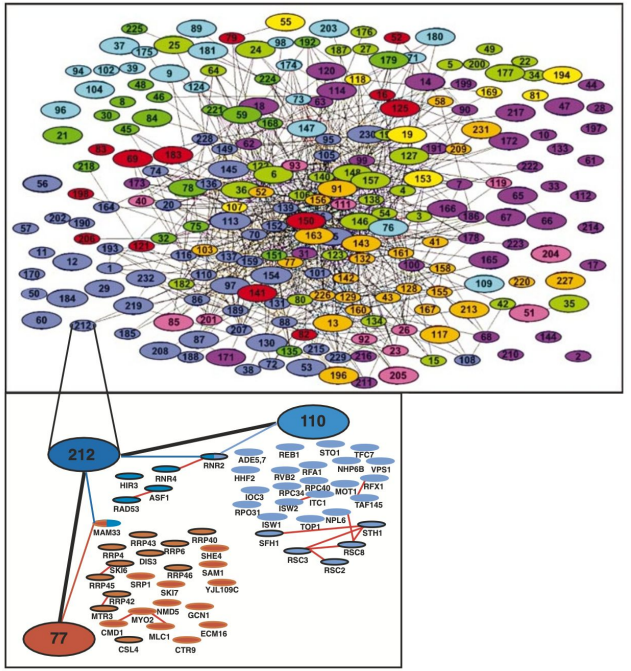


Figure 4 The protein complex network, and grouping of connected complexes. Links were established between complexes sharing at least one protein. For clarity, proteins found in more than nine complexes were omitted. The graphs were generated automatically by a relaxation algorithm that finds a local minimum in the distribution of nodes by minimizing the distance of connected nodes and maximizing distance of unconnected nodes. In the upper panel, cellular roles of the individual complexes (ascribed in Supplementary Information Table S3) are colour coded: red, cell cycle; dark green, signalling; dark blue, transcription, DNA maintenance, chromatin structure; pink, protein and RNA transport; orange, RNA metabolism; light green, protein synthesis and turnover; brown, cell polarity and structure; violet, intermediate and energy metabolism; light blue, membrane biogenesis and traffic. The lower panel is an example of a complex (yeast TAP-C212) linked to two other complexes (yeast TAP-C77 and TAP-C110) by shared components. It illustrates the connection between the protein and complex levels of organization. Red lines indicate physical interactions as listed in YPD[22].

1. Gavin, A.-C. et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415, 141–147 (2002). doi.org/10.1038/415141a

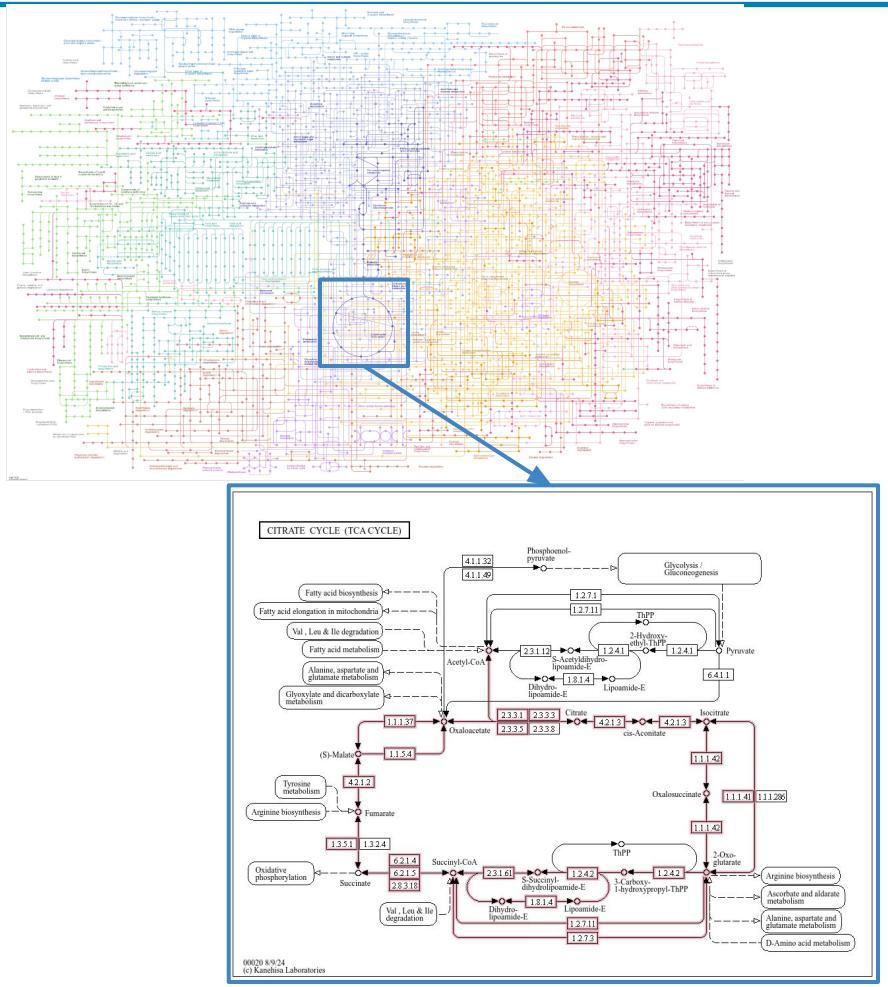
Le labyrinthe métabolique

Depuis plus d'un siècle, les biochimistes ont décrit les réactions chimiques qui permettent aux cellules de métaboliser les petites molécule, de consommer différents substrats et de produire des molécules nécessaires à leur survie.

La plupart des réactions cellulaires sont catalysées par des protéines spécialisées, les enzymes.

Plusieurs bases de données répertorient l'ensemble des réactions et enzymes connues.

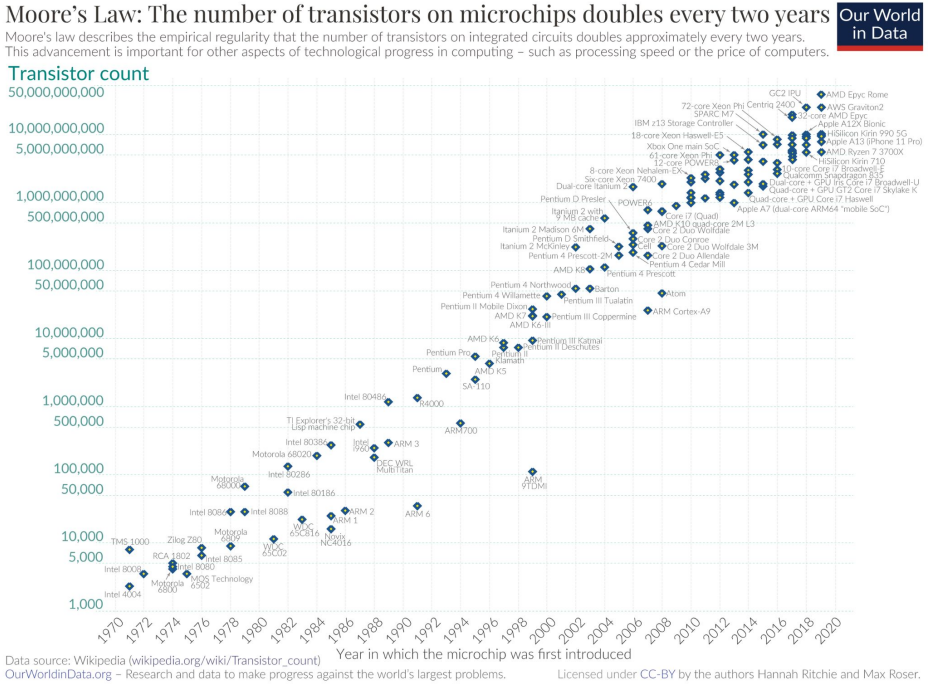
La carte métabolique (à droite) fournit une représentation *simplifiée* de l'intrication du réseau formé par l'ensemble des réactions de la base de données de voies métaboliques KEGG.



https://www.genome.jp/kegg-bin/show_pathway?map01100
https://www.genome.jp/pathway/map00020+M00009

La loi de Moore

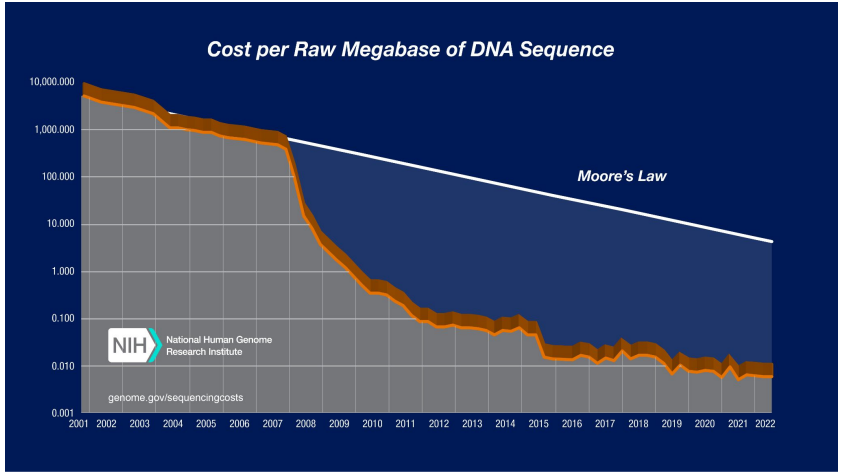
- Observation empirique : à coût constant les capacités des ordinateurs double tous les deux ans.
- Graphique: nombre de transistors (ordonnée) en fonction du temps (abscisse) de 1970 à 2020.
  - L'échelle verticale est logarithmique, la progression est donc exponentielle : à intervalles de temps constant (X), les valeurs sont *multipliées* par un facteur constant.
  - Le nombre de transistors passe de ~2.000 en 1970 à ~40.000.000.000 en 2000 → la capacité des ordinateurs est 20 millions de fois plus élevée.

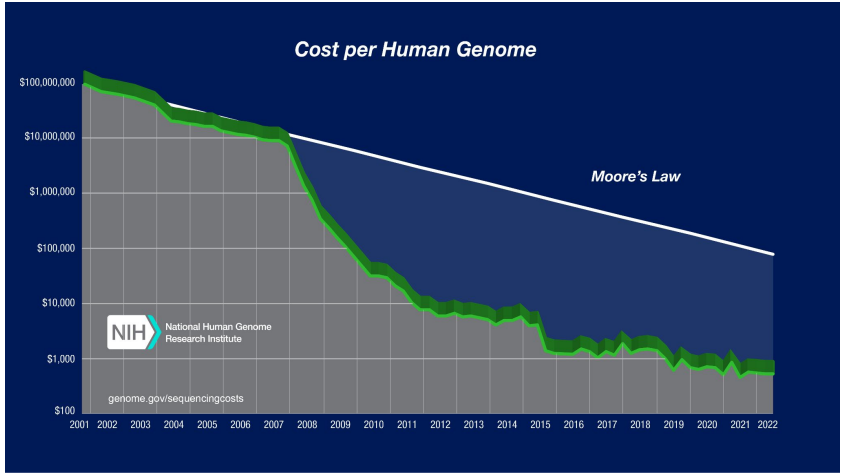


https://en.wikipedia.org/wiki/Moore%27s_law

Le séquençage massivement parallèle (“Next Generation Sequencing”, NGS)

- 1990-2000 : premiers projets de séquençage du génome d’organismes modèles: bactéries, levure du boulanger, drosophile, nématode, arabette, et … “le” génome humain
- 2001 : première publication d’un génome humain
- 2007 : technologies de séquençage massivement parallèle (“**Next Generation Sequencing**”, **NGS**)
  - De 2001 à 2007: les coûts diminuent en suivant la loi de Moore (décroissance exponentielle)
  - 2008; diminution brutale des coûts du séquençage
  - Depuis 2011: réduction plus modérée des coûts

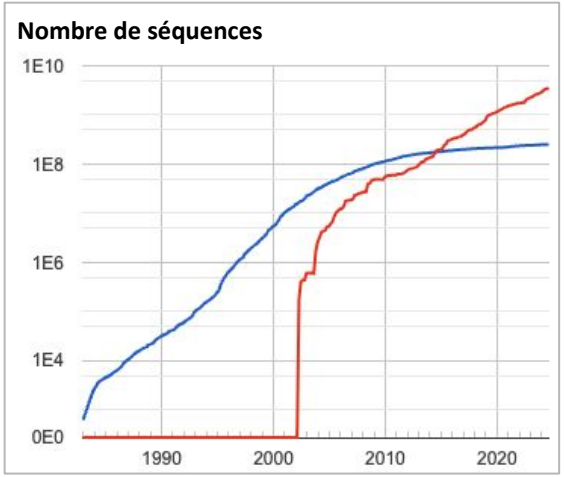


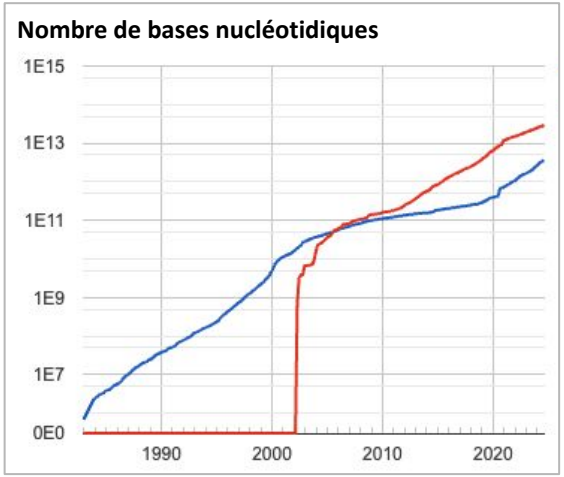


Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcostsdata. Accessed 2024-09-04.
https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

Disponibilité des séquences d'ADN

- Les séquences de macromolécules qui font l'objet de publications scientifiques sont systématiquement déposées dans des entrepôts de données internationaux, et rendues accessibles au public
  - Une exception: les séquences génomiques associées à des échantillons humains (voir cours sur la médecine génomique)
- Le nombre de séquences disponibles depuis 1980 montre une croissance exponentielle (linéaire sur un axe logarithmique).
  - Taux d'augmentation: de 1990 à 2020, x 1.48/an
- Avant 2002, il s'agissait de séquences individuelles de gènes ou de fragments génomiques (courbe bleue, Genbank).
- A partir de 2002, le séquençage de génomes complets prend le pas (courbe rouge).

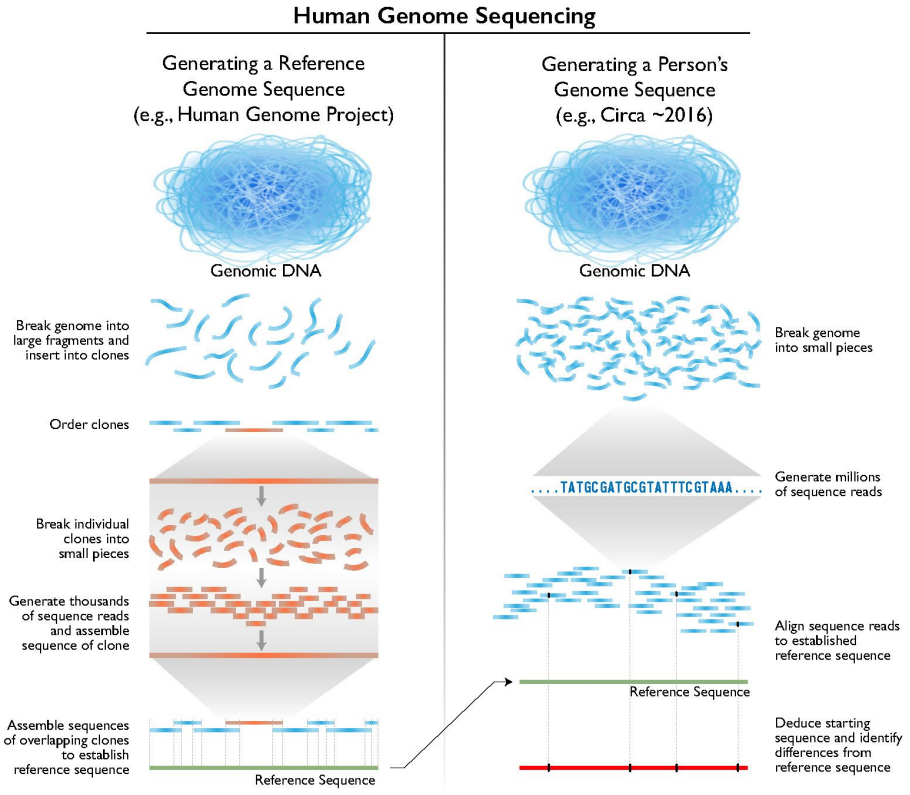




https://www.ncbi.nlm.nih.gov/genbank/statistics/

Le coût d'un génome humain

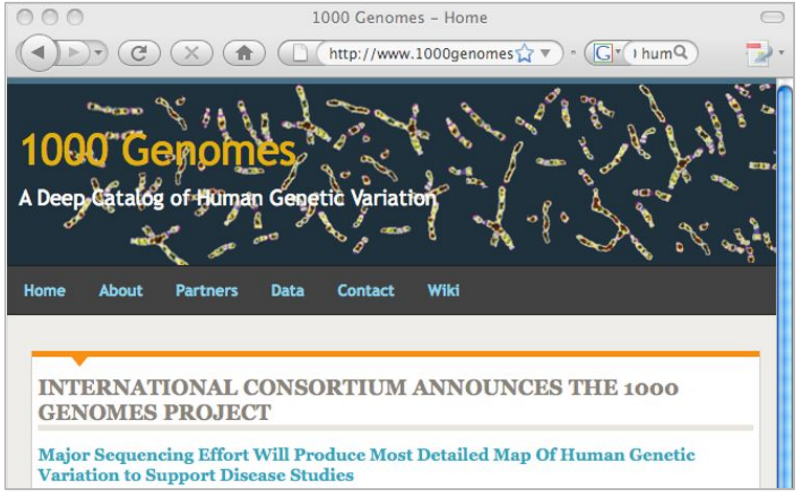
- 1990-2000 : premiers projets de séquençage du génome d'organismes modèles: bactéries, levure du boulanger, drosophile, nématode, arabette, et … "le" génome humain
- 2001 : première publication d'un génome humain
- 2007 : technologies de séquençage massivement parallèle ("Next Generation Sequencing", NGS)
- 2001: **premier génome humain de référence**, version "brouillon"
- 2004: version "propre" du premier génome humain de référence
- Coût de séquençage d'un génome humain
  - Le premier génome humain (2001): ~3 milliards US $
  - 2006 (avant le NGS): 16 millions US $
  - 2016 (après le NGS) : 1.500 US $
  - 2022 : 800 US $

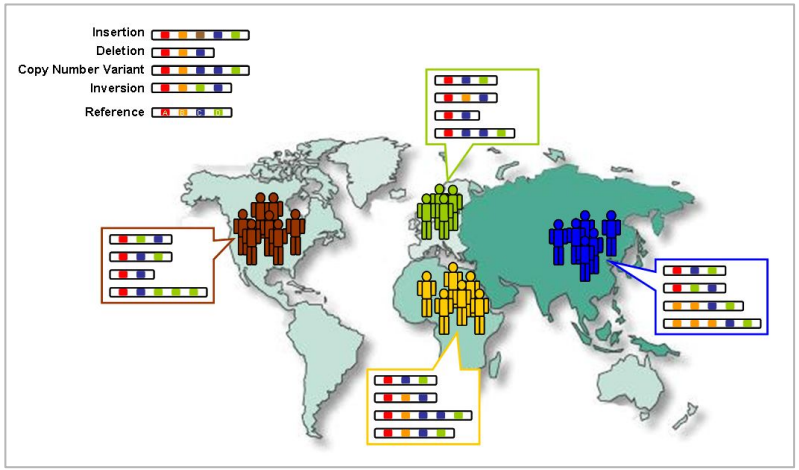


https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

Du génome aux 1.000 génomes

- 1990-2000 : premiers projets de séquençage du génome d'organismes modèles: bactéries, levure du boulanger, drosophile, nématode, arabette, et … "le" génome humain
- 2001 : première publication d'un génome humain
- 2007 : technologies de séquençage massivement parallèle ("Next Generation Sequencing", NGS)
- 2001: premier génome humain, version "brouillon"
- 2004: premier génome humain, version "propre"
- 2008: projet 1.000 génomes (humains)
  - But: caractériser la diversité génotypique dans les populations humaines

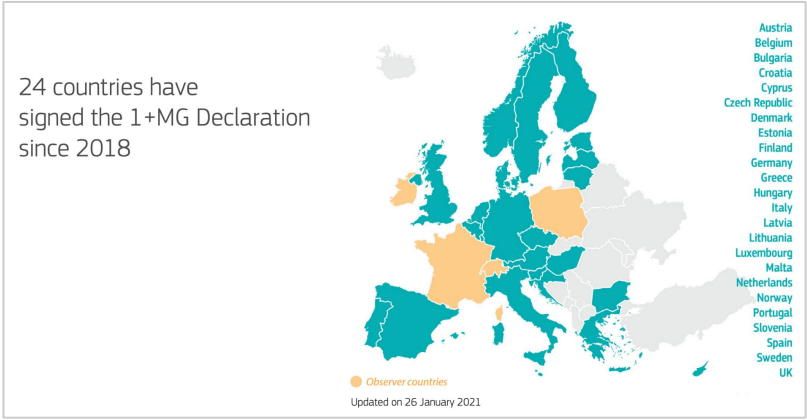




1. The 1000 Genomes Project Consortium et al. A global reference for human genetic variation. Nature 526, 68–74 (2015). doi.org/10.1038/nature15393

Du génome aux millions de génomes

- 1990-2000 : premiers projets de séquençage du génome d'organismes modèles: bactéries, levure du boulanger, drosophile, nématode, arabette, et … "le" génome humain
- 2001 : première publication d'un génome humain
- 2007 : technologies de séquençage massivement parallèle ("Next Generation Sequencing", NGS)
- 2001: premier génome humain, version "brouillon"
- 2004: premier génome humain, version "propre"
- 2008: projet 1.000 génomes (humains)
  - But: caractériser la diversité génotypique
- 2018: projet Européen 1.000.000 génomes (1+MG)
  - But: découvrir les mutations associées aux maladies rares et au cancer
- Initiatives similaires dans d'autres régions du monde.
  - Royaume Uni : 100.000 génomes
  - France : 200 000 génomes / an (annonce PFMG)
  - Chine : 100 millions de génome !
  - USA, Australie, … des centaines de millions de $



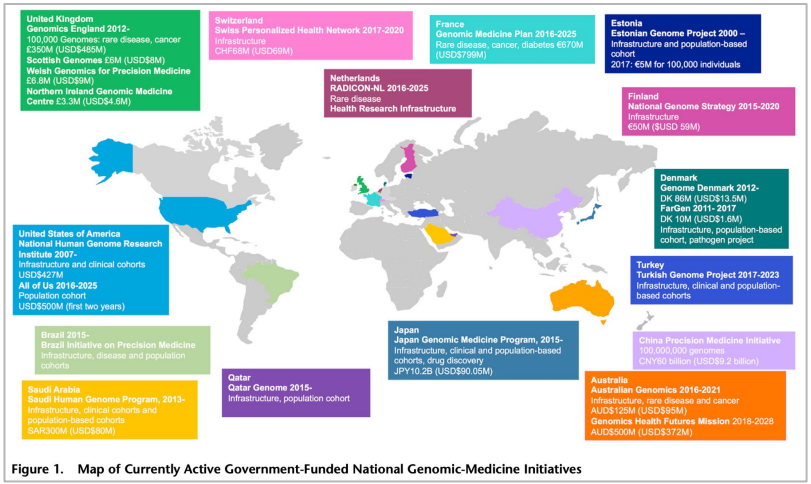


Figure 1. Map of Currently Active Government-Funded National Genomic-Medicine Initiatives

https://digital-strategy.ec.europa.eu/en/policies/1-million-genomes

Des génomes aux métagénomes

- La **métagénomique** consiste à séquencer des échantillons provenant de divers milieux (océans, flore intestinale, …) pour échantillonner les espèces vivantes dans leur milieu naturel.
- « Génomique classique », on isole une espèce microbienne, on la met en culture, et on séquence ensuite son génome (si la culture fonctionne).
- « Métagénomique », on séquence directement tout l'ADN extrait de l'écosystème. On peut ensuite identifier les espèces présentes, caractériser leur abondance, découvrir de nouvelles protéines,
- Exemples
  - **Métagénomique océanique** : l'expédition TARA a échantillonné de la biodiversité dans les eaux océaniques de 2010 à 2012.
  - **Microbiote intestinal** : séquençage de tout l'ADN d'un échantillon fécal, et caractérisation de la flore bactérienne et virale, établissement des liens avec la santé et l'alimentation.
  - Diversité microbienne dans les **fromages** AOP.
  - …

http://oceans.taraexpeditions.org/
https://www.inrae.fr/actualites/fromages-laits-aop-analyse-grande-ampleur-revele-leur-diversite-microbienne

Biologie et santé

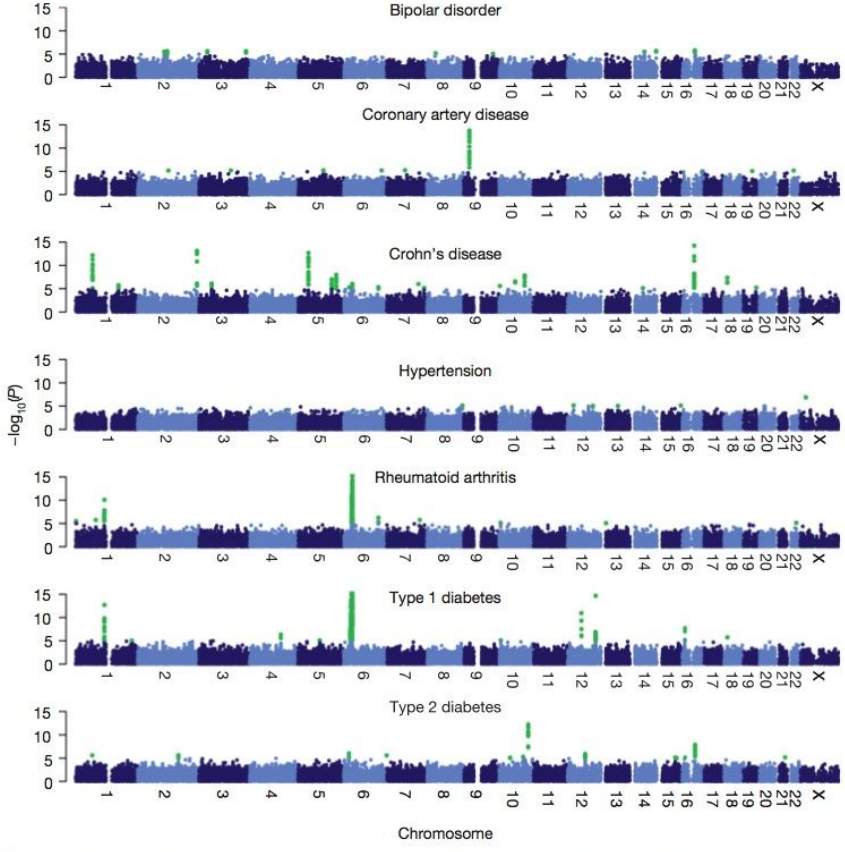
Des grands projets nationaux et internationaux visent à collecter des données médicales à des fins de recherche (découverte des facteurs influençant la santé), de prévention et de soin.

Ces projets combinent différents types de données

- Génomes des patients
- Génomes microbiens
- Métabolites (petites molécules)
- Imagerie médicale
- Données de soin
- Données d'environnement, …

Quelques exemples

- The Cancer Genome Atlas (2005-2018): détection de mutations associées à différents types de cancers
- Études d'associations à l'échelle génomique
  - Une des premières études remarquables : régions génomiques associées à 7 maladies (TWTCCC 2007): 2000 patients pour chaque maladie + 3000 contrôles (figure à droite)
  - Septembre 2024 : >47.000 publications "genome-wide association studies"

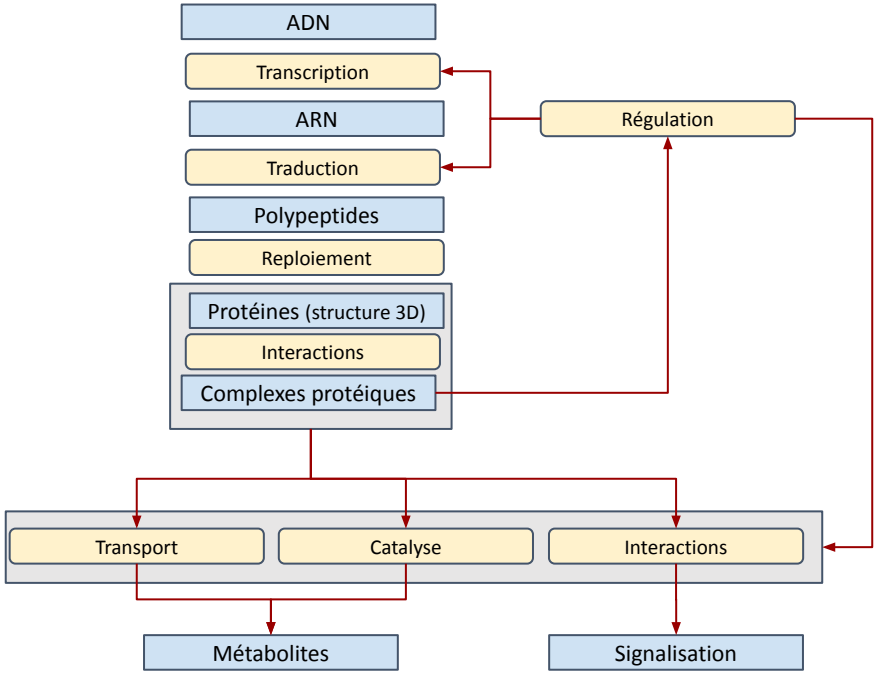


doi.org/10.1038/nature05911

The Wellcome Trust Case Control Consortium et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447, 661–678 (2007). doi.org/10.1038/nature05911

Une vision holistique des systèmes biomoléculaires

- Les cellules vivantes sont des systèmes complexes dont le fonctionnement repose sur l'action coordonnée de milliers de molécules.
- Depuis le début du 20è siècle, des technologies à haut débit ont été développées pour mesurer la quantité et l'activité de ces molécules de façon systématique : génome, transcriptome, protéome, interactome, métabolome, …
- Ces approches holistiques font désormais partie intrinsèque de la façon dont les biologistes analysent les systèmes vivants.
- Elles ouvrent également le champ à de nombreuses applications, dans les domaines de la médecine, des biotechnologies, de l'agriculture, de l'environnement.



Les défis de la biologie à l'ère des données massives

La biologie contemporaine couvre les mêmes problématiques et questions que celle du 20è siècle, mais elle les aborde de façon holistique, en s'appuyant sur des technologies productrices de données massives.

Défis numériques

- Stockage des données en croissance exponentielle
- Puissance de calcul
- Efficacité algorithmique
- Intelligence artificielle

Défis scientifiques

- Modélisation des objets biologique
- Extraction de l'information pertinente dans un océan de donnés (rôle crucial des statistiques)
- Représentation des connaissances (bases de données, visualisation)

Défis sociétaux

- Choix politiques concernant les applications, pour qu'elles soient au bénéfice de l'humain et de l'environnement
- Impact environnemental des moyens numériques
- Protection des données à caractère personnel

La bioinformatique, qu'est-ce que c'est ?

Les définitions varient fortement selon les sources, souvent influencées par le point de vue de la personne qui définit (son domaine de recherche, d'intérêt).

Quelques définitions assez consensuelles

- *In general terms, the application of computers and computational techniques to biological data* . [...] Bioinformatics can be seen as a synonym for Computational Biology. (J. M. Hancock, *in* Concise Encyclopaedia of Bioinformatics)
- *An interdisciplinary field of science that develops methods and software tools for understanding biological data, especially when the data sets are large and complex. Bioinformatics uses biology, chemistry, physics, computer science, computer programming, information engineering, mathematics and statistics to analyze and interpret biological data. The subsequent process of analyzing and interpreting data is referred to as computational biology.* (en.wikipedia.org)

Quelques remarques concernant ces définitions.

La première focalise sur les données, la seconde sur les méthodes et outils.

La première est trop restrictive: la bioinformatique ne se limite pas à l'application de méthodes de calcul à des données. Elle inclut notamment

- La **modélisation statistique** des données
- La **modélisation mathématique** des systèmes biologiques
- Le **développement d'outils logiciels** pour répondre aux questions biologiques.
- Le développement de **bases de données**
- L'annotation ("curation") des données par des experts, pour produire des **bases de connaissances**.

La seconde définition établit une distinction entre "bioinformatics" and "computational biology", mais cette séparation est arbitraire et non-consensuelle.

Concise encyclopaedia of bioinformatics and computational biology. (Wiley Blackwell, 2014). ISBN: 978-0-470

Organisation du cours et modalités de contrôle des connaissances

Plan du cours

CM (7 x 2h)

- Panorama des principales approches bioinformatiques et de leurs applications à différents domaines de la biologie
- Accent sur l'apport des des données massives pour la compréhension des mécanismes du vivant.
- Exemples d'application à différents domaines de la biologie, en particulier évolution, santé, biodiversité.

Chapitres des CM

1. Introduction
2. Séquence → structure → fonction des protéines
3. Des gènes aux génomes
4. Retracer l'évolution à partir des séquences
5. Génomique personnelle
6. Exploration de la biodiversité
7. Réseaux et systèmes biologiques
8. L'information au coeur du vivant

TP (8x2h)

- Sur ordinateur
- Utilisation des outils bioinformatiques conviviaux pour analyser des données de différents types (séquences macromoléculaires , génomes, structures, réseaux biologiques).
- Aucun compétence prérequise en informatique

Séances de TP

1. Séquence, structure, fonction
2. Du gène à la protéine
3. Du gène au génome et au protéome
4. Alignements par paires et alignements multiples
5. Inférence phylogénétique
6. Variants génétiques
7. Systèmes et réseaux biologiques
8. Récapitulation, questions / réponses

Acquis d'apprentissages / objectifs d'UE (d'après les définitions des compétences de la licence SV)

1.1 Connaître les concepts de la biologie au niveau moléculaire (Structure, fonction des biomolécules, Flux d'information génétique)

1.2 Connaître les concepts de la biologie au niveau cellulaire (Organisation et fonctionnement des génomes, génomique)

1.4 Intégrer les différents niveaux d'organisation du vivant (Diversité et unicité du vivant, Organisation et fonctionnement de réseaux biologiques, Grands principes de l'analyse génomique et in silico des séquences, Spécificité et complexité des systèmes biologiques)

1.5 Situer les connaissances actuelles en biologie dans le contexte de l'évolution des questions, concepts et théories (Grands jalons de l'histoire de la biologie)

1.6 Acquérir et mobiliser les connaissances de base des disciplines connexes aux sciences du vivant pour analyser des résultats biologiques (Probabilités et statistiques)

1.7 Connaître et mobiliser les méthodologies et technologies de la biologie (Outils et méthodes bioinformatiques, Organismes et systèmes modèles, Principales techniques à haut débit: génomique, transcriptomique, protéomique)

1.8 Identifier les enjeux éthiques, environnementaux et sociétaux liés à l'application de la biologie (Enjeux éthiques et sociétaux de la biologie pour la recherche et la santé, Environnement et écologie)

2.6 Choisir et utiliser des outils d'analyse et de traitement des données dans différents domaines de la biologie (Analyse statistique, Probabilités, outils bioinformatiques)

Modalités de contrôle des connaissances

Critères d'évaluation

- Acquisition des concepts de bioinformatique
- Compréhension du rôle des données massives et de la bioinformatique dans différents domaines de la biologie (santé humaine, biodiversité)
- Compréhension des outils bioinformatiques utilisés pendant les TP et interprétation des résultats

La présence aux TP est obligatoire

- Une marge de 20% d'absence est tolérée
- Au-delà de ce seuil, absence injustifiée (ABI) pour l'UE dans son ensemble ⟶ passage en deuxième session

Première session

- QCM hors séance (20%)
  - A réaliser en cours de semestre, en dehors des séances de CM et TP
  - Questionnaires communiqués au fil de l'eau durant le cours
  - But: auto-évaluation et entraînement au QCM final
  - Note d'assiduité: points attribués en fonction du taux de réponse plutôt que de leur correction
- Examen terminal (80%)
  - Sur table en QCM
  - Inclura des **questions de cours** et des **questions de TP**

Seconde session

- Examen sur table en QCM
- Pondération: pour chaque étudiant, la note finale sera la note maximale entre deux formules
  - 20%CC + 80% examen de seconde session
  - 100% examen de seconde session

Des questions ?

Foire aux questions (réponses de l'enseignant aux questions posées lors du sondage de fin de séance)

Modalités de contrôle des connaissances

- *Les QCM sont durs ?*
  - Nous veillerons à ce qu'ils aient un niveau de difficulté adéquat pour les étudiants. Le but est d'évaluer votre acquisition des connaissances et compétences associées aux cours, et pas de vous mettre en difficulté.
- *Le QCM comportera-t-il des points négatifs ?*
  - En cours d'évaluation par l'équipe pédagogique
- *Cette séance d'introduction fait-elle partie de la matière d'examen ?*
  - Oui, ainsi que toutes les séances de CM et de TP
- *Faut-il retenir les dates ?*
  - Non, mais vous devez avoir une idée approximative (décennie) des grands jalons de l'histoire de la biologie et des méthodes bioinformatiques
- *Doit-on apprendre tout ce qui a été dit ou juste ce qui figure sur les diapo ?*
  - Le contenu du cours est ce qui a été dit. Les diapo ne sont qu'un support graphique (et pour l'enseignant, un guide pour le déroulé du cours)
- *Les TP seront-ils notés et inclus dans la note finale ?*
  - Les TP ne seront pas notés, mais il sera nécessaire de les suivre pour deux raisons: ils sont obligatoires (>20% d'absence → seconde session d'office) + réponse suivante
- *Y a-t-il des applications des TP dans les QCM ?*
  - Oui. Le QCM comportera plus ou moins 50% de questions sur les CM, et 50% sur les méthodes et résultats des TP

Supports de cours

- *Les diapo seront-elles mises en ligne ?*
  - Oui, dans la mesure du possible avant le cours, et sinon juste après

Autres questions

- *Peut-on s'entraîner chez nous ?*
  - Oui, et nous vous y encourageons. Tous les outils logiciels utilisés aux TP sont accessibles en ligne gratuitement
- *Pour les statistiques, est-ce que nous allons avoir des cours de math ?*
  - Il n'y a pas de cours de statistique en L2 ou L3 SV AMU, mais les TP intégrés du second semestre incluront une prise en main des outils statistiques pour analyser vos données expérimentale
- *On est d'accord que c'est pas le même cours que l'année dernière ?*
  - Effectivement. Certaines notions et certains outils se retrouveront dans cette UE, mais la perspective a été révisée en profondeur.
- *Les annales de l'année dernière sont ils utiles pour cette année avec la réforme*
  - Non, les supports de cours seront modifiés en profondeur
- *Pourquoi vous filmez ?*
  - Parce que cette UE est ouverte en enseignement à distance (téléenseignement)
- *Le cours portera-t-il plus sur le fonctionnement des bio informatiques ou c'est uniquement sur son histoire ?*
  - La séance d'introduction donnait une perspective historique, mais les séances suivantes seront consacrées aux approches bioinformatiques pour l'analyse des données biologiques.