Bionformatique

D. Puthier Inserm U1090
Polytech Biotech III, 2014

L’informatique est omniprésente dans notre sociéte. La biologie ne fait pas exception

Bioinformatique ?

Utiliser l’information numérisée pour comprendre le fonctionnement du vivant

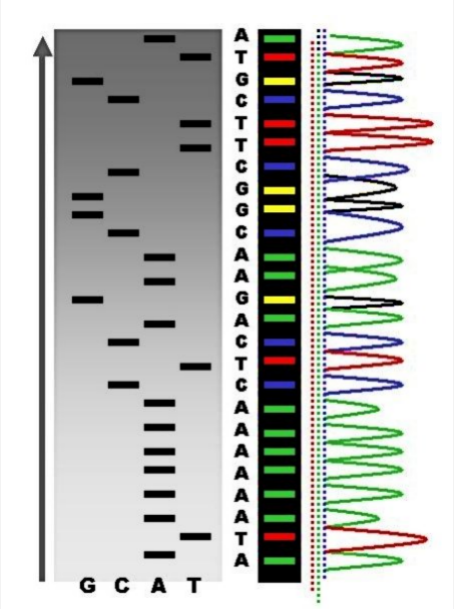
Quelques applications...

Chuck, s’il te plait...

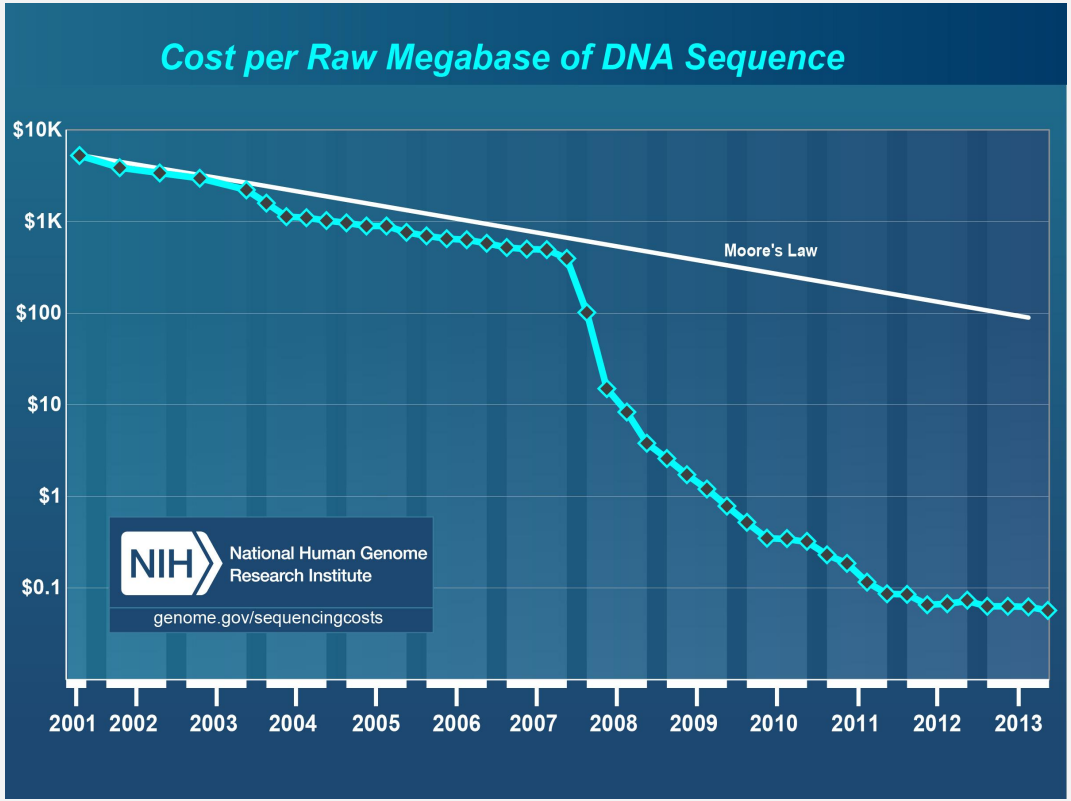
Analyse de séquences

Le séquençage c’est plus ce que c’était

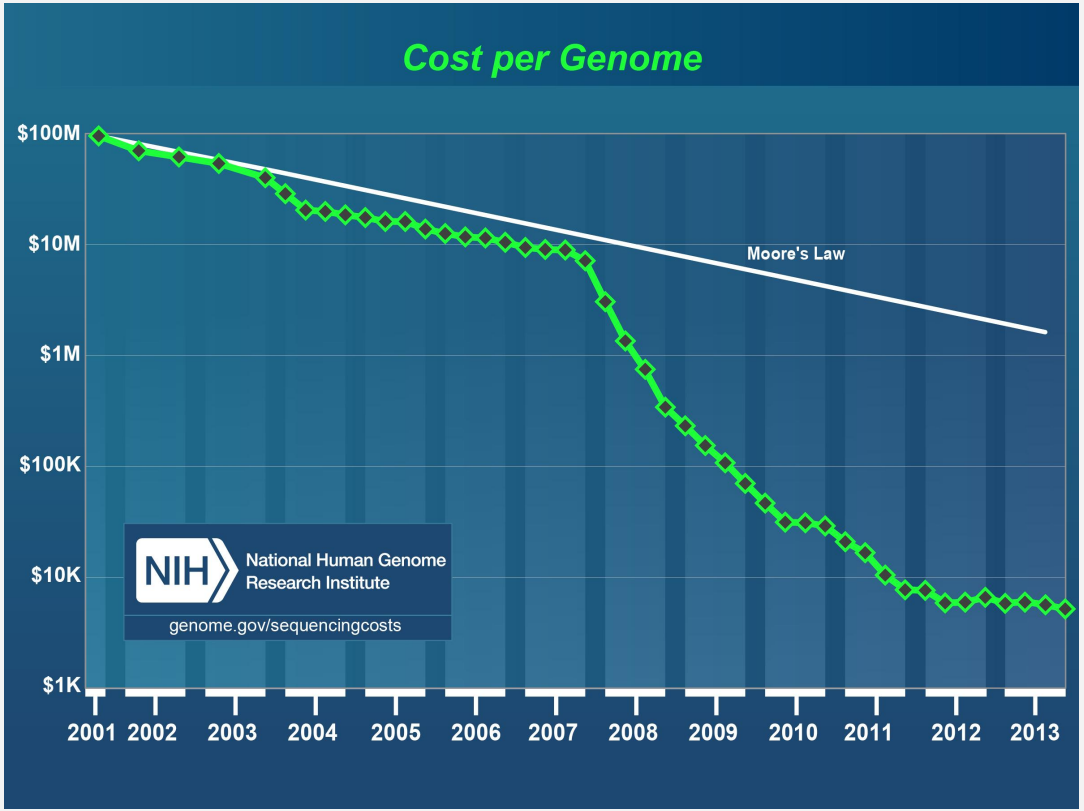
- 1977-1990, 500bp, analyses manuelles
- 1990-2000, 500Bp, analyses assistées (“1D capillary sequencers”)
- 2005-2014, 20-1000bp (“2D sequencers “Next Generation Sequencing.”)



Coût par mégabase (1 million de bases)



Coût par génome humain



- Sanger-based sequencing (average read length=500-600 bases): 6-fold coverage
- 454 sequencing (average read length=300-400 bases): 10-fold coverage
- Illumina and SOLiD sequencing (average read length=50-100 bases): 30-fold coverage

De véritables usines à séquencer

- Illumina
- 10 séquenceurs.
  - Chacun produisant 1,8 Terabases / 3 day
- 18,000 génomes / an
  - ”Factory-scale sequencing technology”
- Enfin le génome à 1000$ ...

Ou des séquenceurs sur clefs USB...

Bioinformatique: analyse de séquences

- Assemblage de génomes
- Annotation de génomes
- Recherche contre des bases de données (BLAST, Blat,...)
- Alignements multiples (Clustal,...)
- Alignements de génomes
- Recherche/découverte de motifs fonctionnels

Bioinformatique & génétique humaine

Bioinformatique: génétique humaine

- Analyse de la diversité des génomes
  - SNPs (Single Nucleotide Polymorphisms)
  - InDel (Insertion/Deletion)
  - CNV (Copy Number Variation)

- GWAS (Genome wide association studies)
  - Associer SNPs et maladies

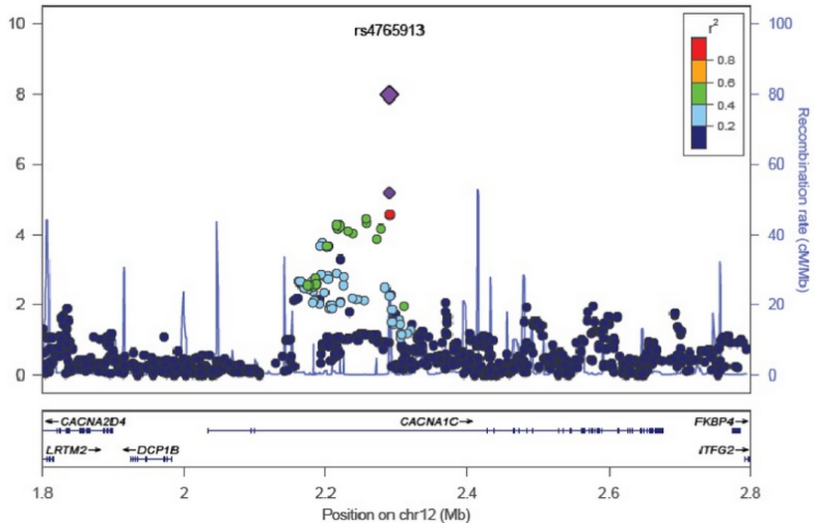
Analyse GWAS

Nat Genet. Author manuscript; available in PMC May 1, 2013.
Published in final edited form as:
Nat Genet. Oct 2011; 43(10): 977–983.
Published online Sep 18, 2011. doi: 10.1038/ng.943
PMCID: PMC3637176
HALMS: HALMS634944
INSERM Subrepository

Large-scale genome-wide association analysis of bipolar disorder identifies a new susceptibility locus near ODZ4

Bipolar disorder (BD) is a severe mood disorder affecting greater than 1% of the population[1]. Classical BD is characterized by recurrent manic episodes that often alternate with depression. Its onset is in late adolescence or early adulthood and results in chronic illness with moderate to severe impairments (...).

Genome-wide significant evidence for association was confirmed for *CACNA1C* and found for a novel gene *ODZ4* (...).
Pathway analysis identified a pathway comprised of subunits of calcium channels enriched in the bipolar disorder association intervals.



Variations genetiques chez l'humain

- Projet 1000 génomes
  - "1,092 individuals from 14 populations, constructed using a combination of low-coverage **whole-genome** and **exome Sequencing"**
- 38 millions de SNPs, 1.4 millions d'indels

An integrated map of genetic variation from 1,092 human genomes

The 1000 Genomes Project Consortium

Affiliations | Contributions | Corresponding author

Nature **491**, 56–65 (01 November 2012) | doi:10.1038/nature11632

Received 04 July 2012 | Accepted 01 October 2012 | Published online 31 October 2012

“Million Human Genomes project”

Human

The Million Human Genomes Project was launched by BGI to decode the genome of over 1 million people in November 2011. This project concludes five essential parts: Ancient genomes, Population genomes, Medical genomes, Cell genomes and Personal genomes.

The aim of this project is to establish the research baseline and reference standard for specific populations, as well as to connect the phenotypes of diseases and traits with the genetic variations to understand the disease mechanism.

The integrative genome message and scientific discoveries obtaining from the project will lay the foundation for guiding the innovative clinical diagnosis and treatment, and ultimately advancing personalized healthcare and improving human health.

华大基因 BGI

Bioinformatique & cancérologie

Bioinformatique: cancérologie

- Analyser les génomes de tumeurs
  - Définir les anomalies
    - Mutations
    - Translocations
    - Insertions
    - Déletions
    - Variations de nombre (*e.g.* amplifications)
- Applications envisageables en séquençage...

Exome sequencing of renal cell carcinoma

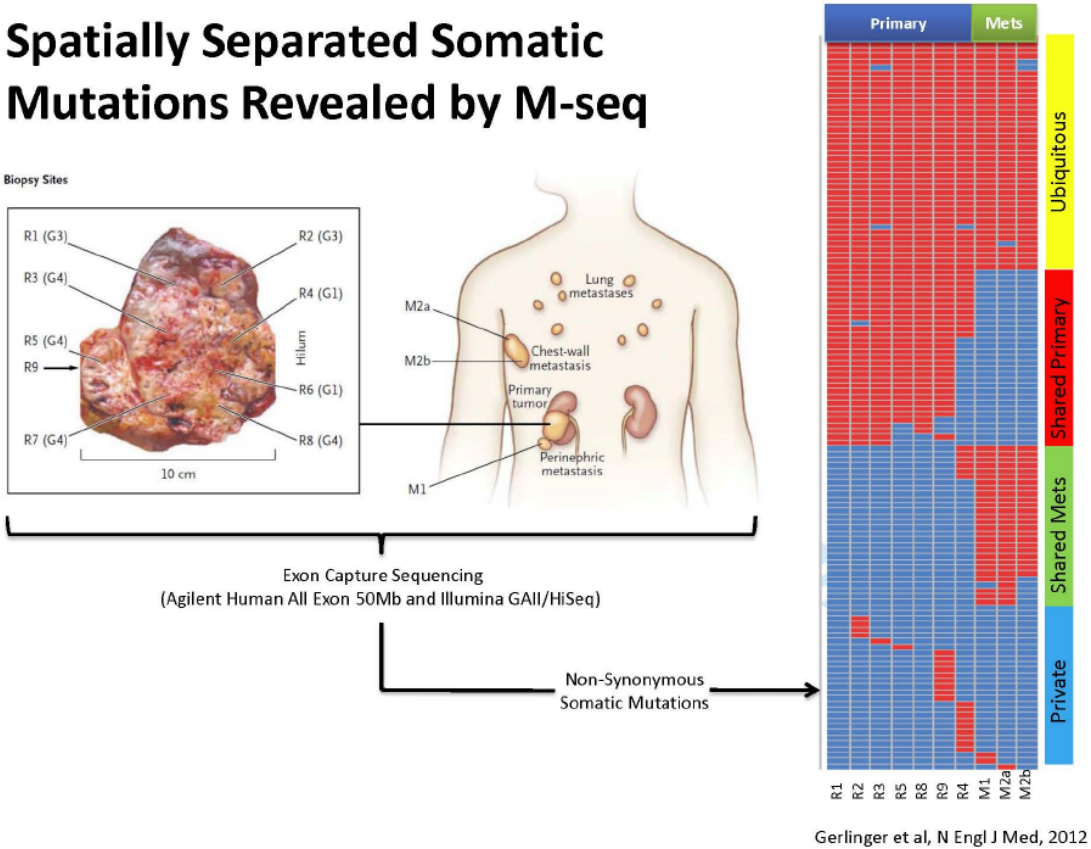
Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing

Marco Gerlinger, M.D., Andrew J. Rowan, B.Sc., Stuart Horswell, M.Math., James Larkin, M.D., Ph.D., David Endesfelder, Dip.Math., Eva Gronroos, Ph.D., Pierre Martinez, Ph.D., Nicholas Matthews, B.Sc., Aengus Stewart, M.Sc., Patrick Tarpey, Ph.D., Ignacio Varela, Ph.D., Benjamin Phillimore, B.Sc., Sharmin Begum, M.Sc., Neil Q. McDonald, Ph.D., Adam Butler, B.Sc., David Jones, M.Sc., Keiran Raine, M.Sc., Calli Latimer, B.Sc., Claudio R. Santos, Ph.D., Mahrokh Nohadani, H.N.C., Aron C. Eklund, Ph.D., Bradley Spencer-Dene, Ph.D., Graham Clark, B.Sc., Lisa Pickering, M.D., Ph.D., Gordon Stamp, M.D., Martin Gore, M.D., Ph.D., Zoltan Szallasi, M.D., Julian Downward, Ph.D., P. Andrew Futreal, Ph.D., and Charles Swanton, M.D., Ph.D.

N Engl J Med 2012; 366:883-892 | March 8, 2012 | DOI: 10.1056/NEJMoa1113205

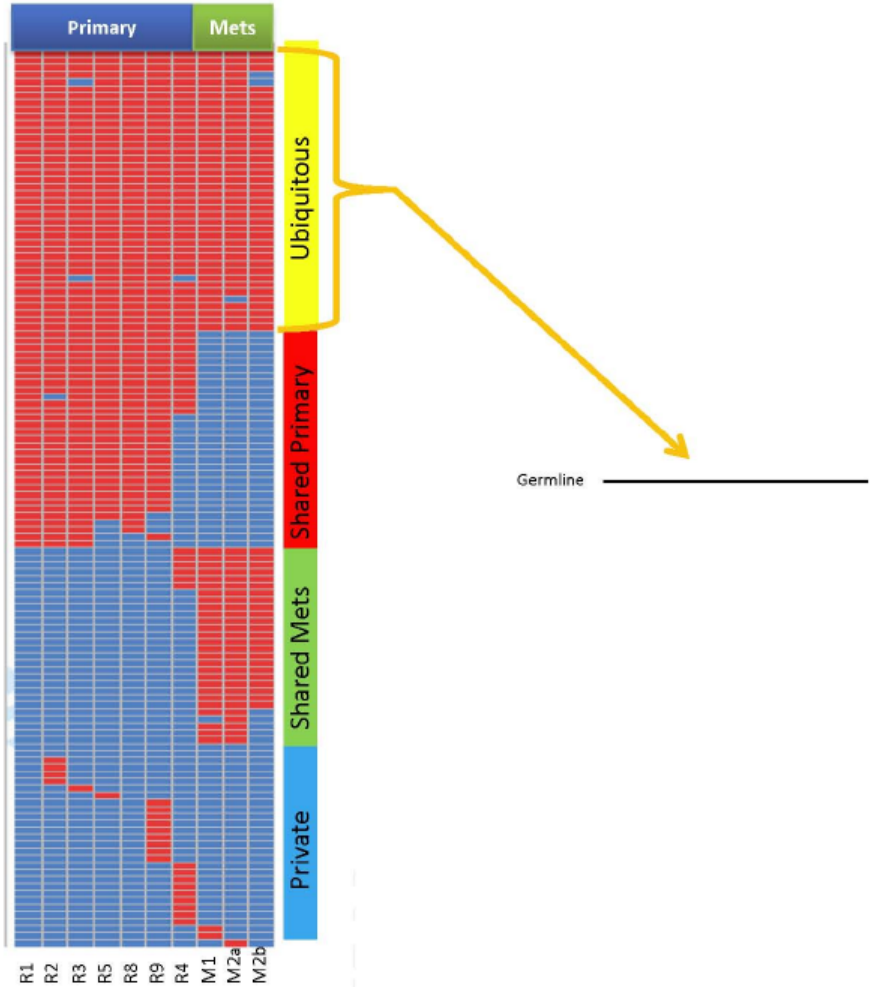
Cancer a clonal disease evolving in a linear fashion ?
What about tumor heterogeneity ?
Can we re-constitute the evolution of the tumor ?

Exome-Seq of Renal cell carcinoma

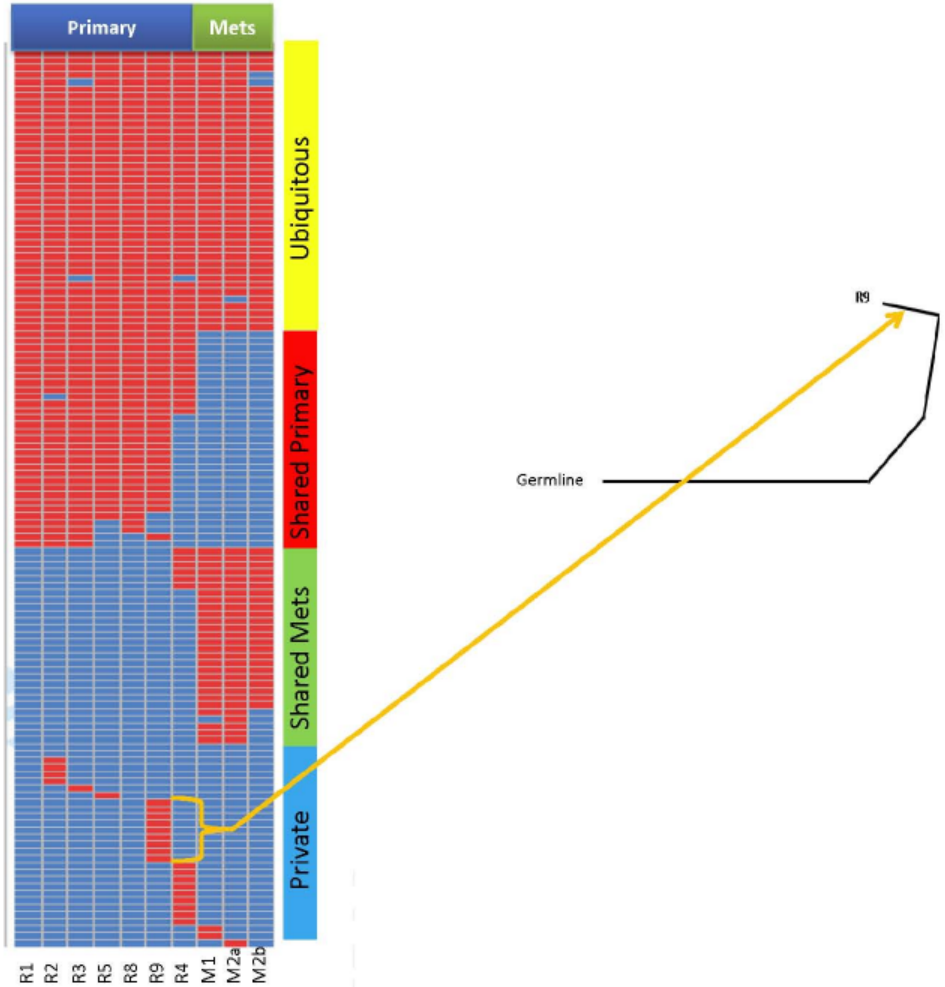


Gerlinger et al, N Engl J Med, 2012

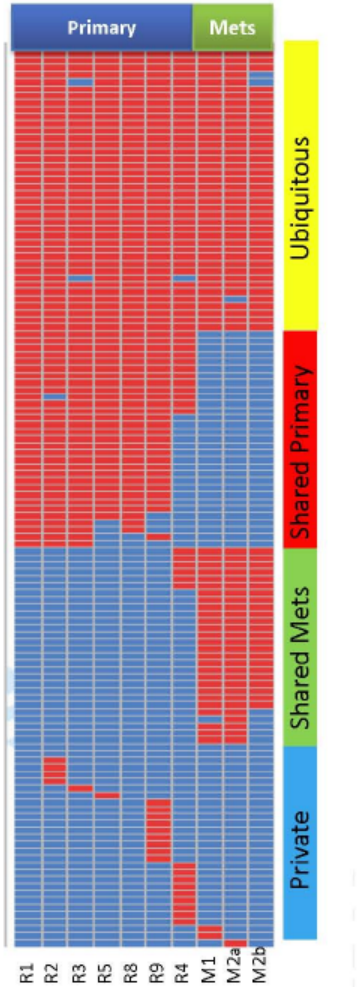
Phylogenetic reconstruction by clonal ordering

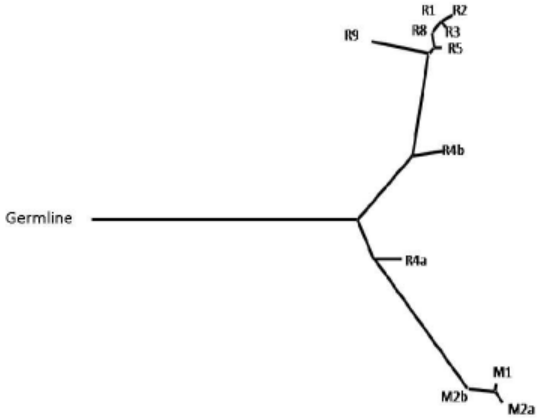


Phylogenetic reconstruction by clonal ordering

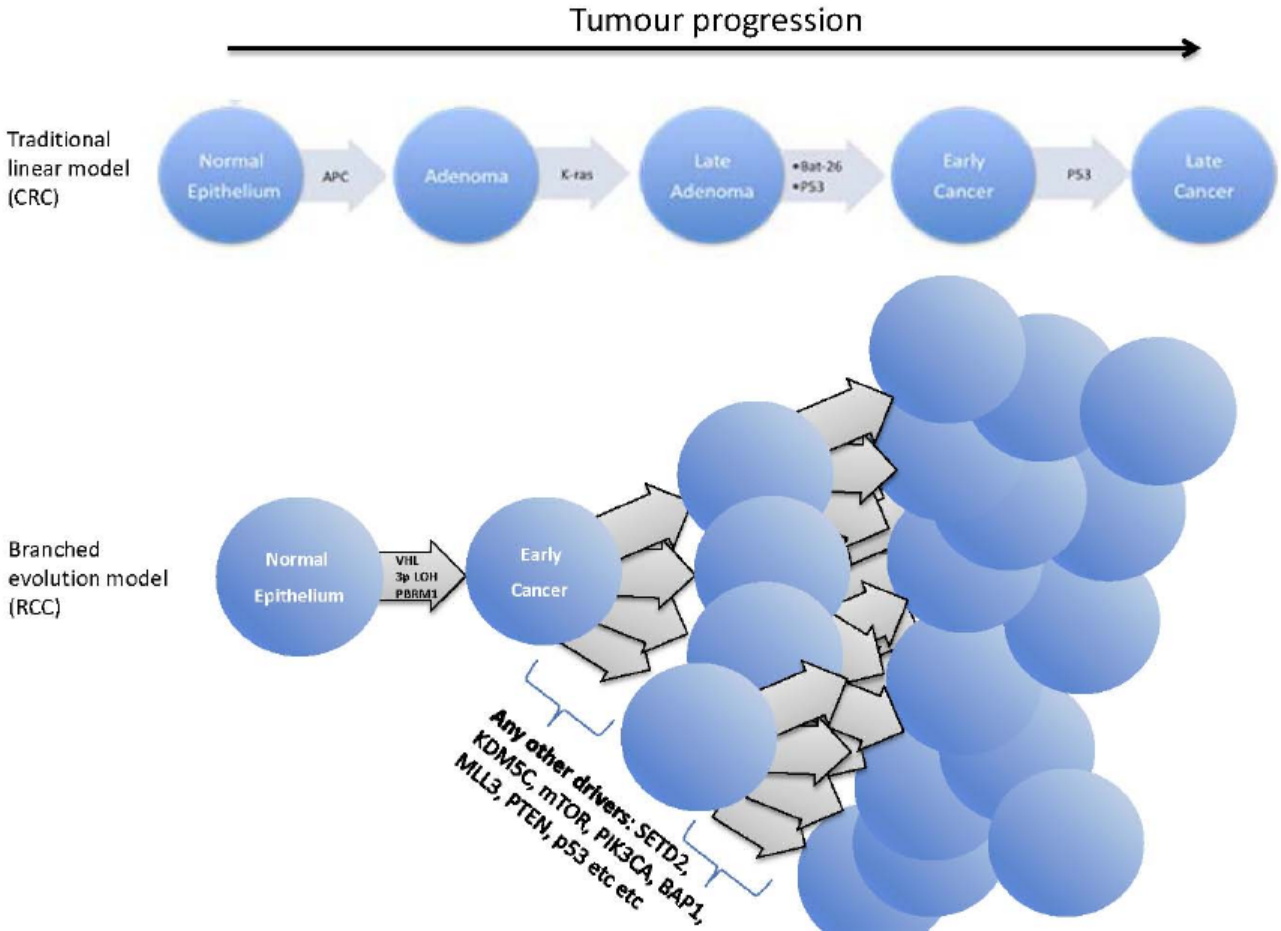


Phylogenetic reconstruction by clonal ordering





Cancer: A clonal disease evolving in a linear fashion?



Séquençage de tumeurs

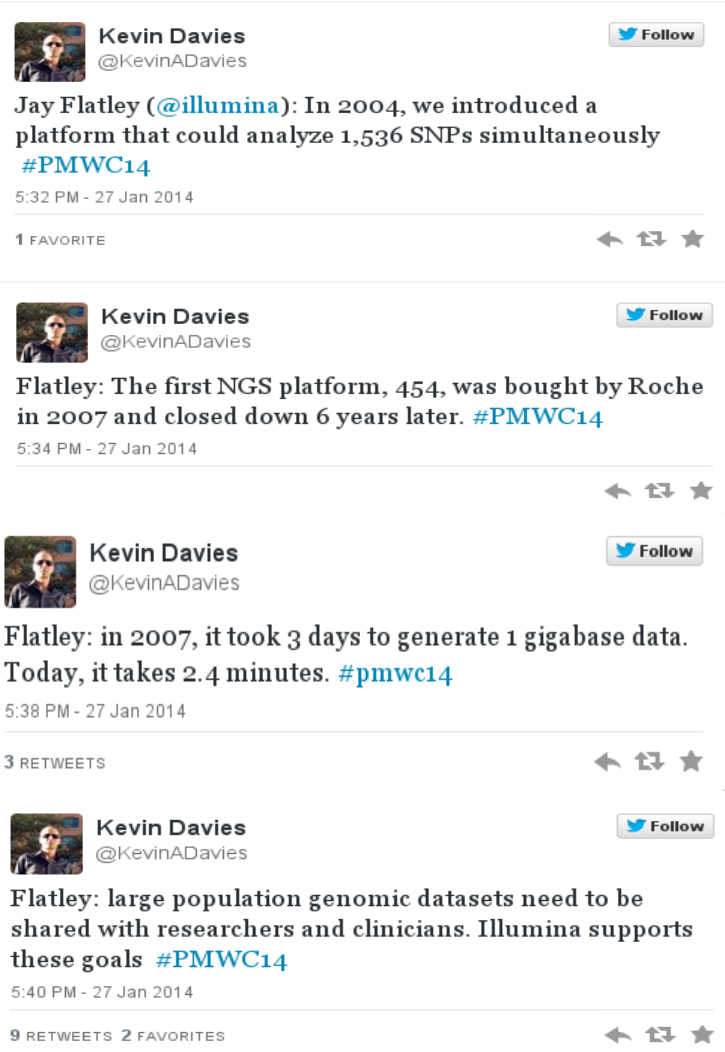
Illumina's Jay Flatley at #PMWC14: Get Sequence of 1 million cancer patients in next 5 years

January 27, 2014 by nextgenseek · 1 Comment

Like 3 Tweet 23 Pin it Share 11

Illumina's Jay Flately said at #PMWC14 that Illumina wants to have the sequence of 1 million cancer patients in a database in the next five years. And one of his personal goal is to make cancer a "chronic" disease within 10 years. Jay Flatley said Illumina support the goals of sharing large population genomic datasets with researchers and clinicians. This is the gist of Jay Flately's talk at #PMWC14 happening right now at Mountain View, CA.

Thanks to awesome live tweets by Kevin Davies, @DivaBioTech, and Theral Timpson. Here are the links to the original tweets.



Calico

Larry Page at Google's headquarters

being

MOUNTAIN VIEW, CA – September 18, 2013 – Google today announced Calico, a new company that will focus on health and well-being, in particular the challenge of aging and associated diseases. Arthur D. Levinson, Chairman and former CEO of Genentech and Chairman of Apple, will be Chief Executive Officer and a founding investor.

Announcing this new investment, Larry Page, Google CEO said: "Illness and aging affect all our families. With some longer term, moonshot thinking around healthcare and biotechnology, I believe we can improve millions of lives. It's impossible to imagine anyone better than Art—one of the leading scientists, entrepreneurs and CEOs of our generation—to take this new venture forward." Art said: "I've devoted much of my life to science and technology, with the goal of improving human health. Larry's focus on outsized improvements has inspired me, and I'm tremendously excited about what's next."

Art Levinson will remain Chairman of Genentech and a director of Hoffmann-La Roche, as well as Chairman of Apple.

Commenting on Art's new role, Franz Humer, Chairman of Hoffmann-La Roche, said: "Art's track record at Genentech has been exemplary, and we see an interesting potential for our companies to work together going forward. We're delighted he'll stay on our board."

Tim Cook, Chief Executive Officer of Apple, said: "For too many of our friends and family, life has been cut short or the quality of their life is too often lacking. Art is one of the crazy ones who thinks it doesn't have to be this way. There is no one better suited to lead this mission and I am excited to see the results."

Yet another ongoing project : HLI

Human Longevity Inc. (HLI) Launched to Promote Healthy Aging Using Advances in Genomics and Stem Cell Therapies

HLI is Building World's Largest Genotype/Phenotype Database by Sequencing up to 40,000 Human Genomes/Year Combined with Microbiome, Metabolome and Clinical Data to Develop Life Enhancing Therapies

HLI has Purchased Two Illumina HiSeq X Ten Sequencing Systems

SAN DIEGO, CA (March 4, 2014)—Human Longevity Inc. (HLI), a genomics an cell therapy-based diagnostic and therapeutic company focused on extendin the healthy, high performance human life span, was announced today by co-founders J. Craig Venter, Ph.D., Robert Hariri, M.D., Ph.D., and Peter H. Diamandis, M.D.

The company, headquartered in San Diego, California, is being capitalized wi an initial $70 million in investor funding.

HLI's funding is being used to build the largest human sequencing operation in the world to compile the most comprehensive and complete human genotype, microbiome, and phenotype database available to tackle the diseases associated with aging-related human biological decline. HLI is also leading the development of cell-based therapeutics to address age-related decline in endogenous stem cell function. Revenue streams will be derived

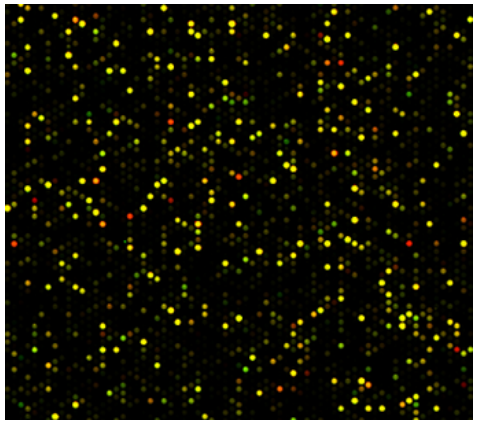
HLI has initially purchased two Illumina HiSeq X Ten Sequencing Systems (with the option to acquire three additional systems) to sequence up to 40,000 human genomes per year, with plans to rapidly scale to 100,000 human genomes per year. HLI will sequence a variety of humans—children, adults and super centenarians and those with disease and those that are healthy.

HLI is uniquely positioned to identify therapeutic solutions to preserve the healthy, high performing body by focusing on some of the most prevalent an actionable areas. HLI is concentrating on cancer, diabetes and obesity, heart and liver diseases, and dementia with its team of expert scientists and clinicians. The company has established strategic collaborations with Metabolon Inc., University of California, San Diego, and the J. Craig Venter Institute (JCVI).

Bioinformatique & régulation des gènes

Bioinformatique: transcriptomique

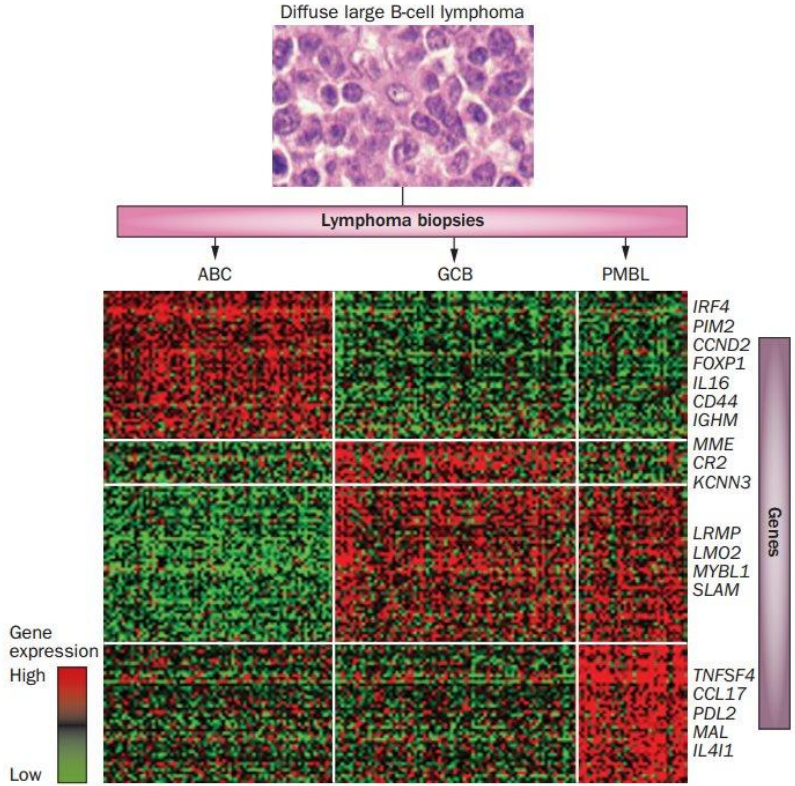
- Transcriptome: l’ensemble des transcrits d’ une cellule
  - Analyse par microarrays ou séquençage (RNA-Seq)



Bioinformatique: transcriptomique

- Analyse du transcriptome
  - Analyse systématiques des transcrits cellulaires
    - Classification des patients
      - Comparaison des classes
        - Recherche de gènes spécifiques, biomarqueurs
      - Prédiction de classes
        - Prédire le type de maladie sur la base de l'expression des gènes
      - Découverte de classes
        - Décrire de nouvelles classes de patients
    - Recherche fondamentale

Bioinformatique: transcriptomique



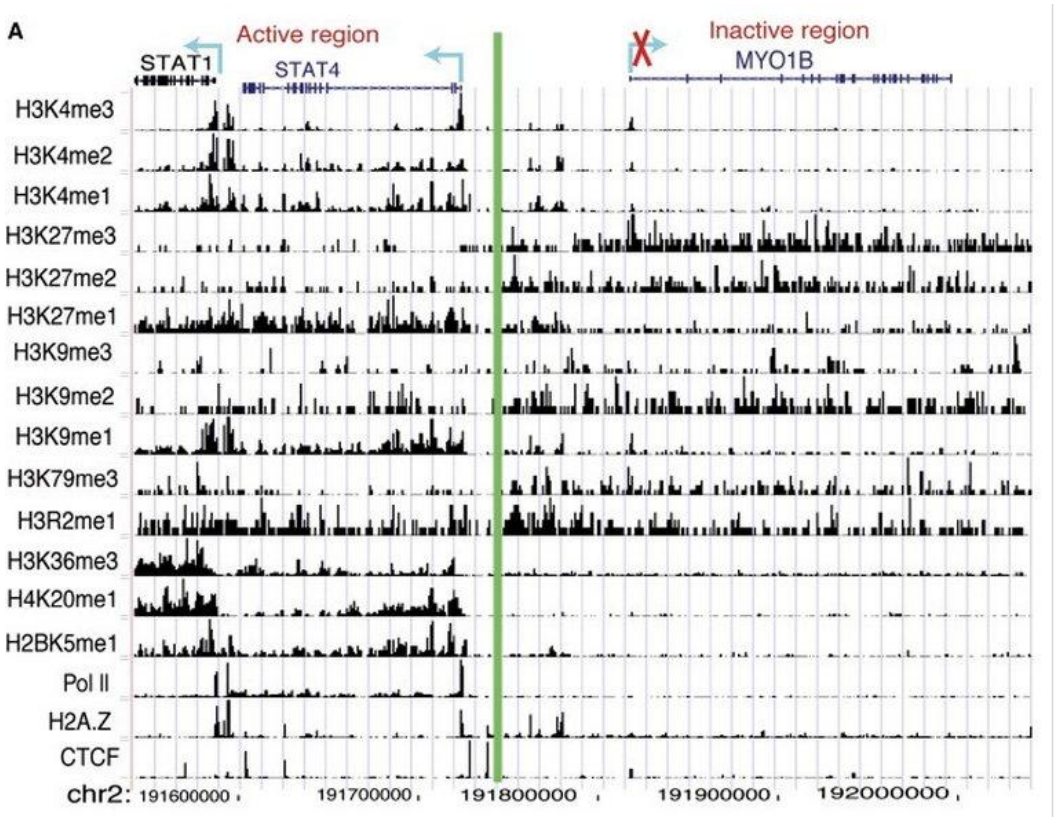
Diffuse large B-cell lymphoma—treatment approaches in the molecular era

Mark Roschewski, Louis M. Staudt and Wyndham H. Wilson

Bioinformatique: régulation génique

- ChIP-Seq (Chromatine Immuno-Precipitation and sequencing)
  - Recherche des régions régulatrices
    - Recherche de sites de liaison de facteurs de transcription
    - Analyse des modifications épigénétiques
  - Influence sur le transcriptome ? (*e.g.* cancer)

ChIP-Seq


A
Active region
Inactive region
STAT1
STAT4
MYO1B
H3K4me3
H3K4me2
H3K4me1
H3K27me3
H3K27me2
H3K27me1
H3K9me3
H3K9me2
H3K9me1
H3K79me3
H3R2me1
H3K36me3
H4K20me1
H2BK5me1
Pol II
H2A.Z
CTCF
chr2: 191600000 191700000 191800000 191900000 192000000

Bioinformatique & évolution

Bioinformatique: évolution

- Phylogénie
  - Vise à proposer des modèles d'évolution des organismes vivants
- Génomique comparative
  - Etude comparative de la structure des génomes de différentes espèces.
  - Effet de la sélection sur l'organisation et l'évolution des génomes

Million plants and animals genome project

Plant & Animal

The Million Plant & Animal Genomes Project aims to generate reference genomes for thousands of economically and scientifically important plant/animal species and resequence millions of plant/animal specimens. This enormous project, to be carried out in collaboration with scientists worldwide, will ultimately generate a huge database of genetic information, allow dramatic improvement in the research of biodiversity conservation, evolutionary mechanism studies, gene function analyses, and help to build animal models for diseases, accelerate molecular breeding, etc. The primary goal for this project is to use genome sequencing and bioinformatics technologies to accelerate the development of practical mechanisms to ensure food security, promote medical applications, improve ecological conservation, and develop new energy sources.

GENOME 10K

Genome 10K Project

The Genome 10K project aims to establish a genomic 'zoo' — a collection of DNA sequences representing the genomes of 10,000 vertebrate species, approximately one for every vertebrate genus. Capturing the genetic diversity of vertebrate species will create an unprecedented resource for the life sciences and for worldwide conservation efforts.

i5k

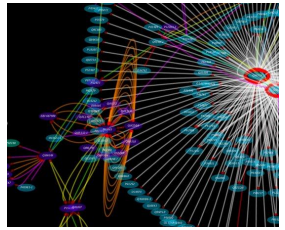
i5k Initiative

The i5k initiative plans to sequence the genomes of 5,000 insect and related arthropod species over the next 5 years. It aims to sequence the genomes of all insect species known to have worldwide importance in agriculture, food safety, medicine, and energy production, and those with important scientific value in evolution and phylogeny research.

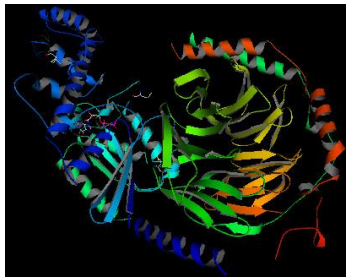
华大基因 BGI

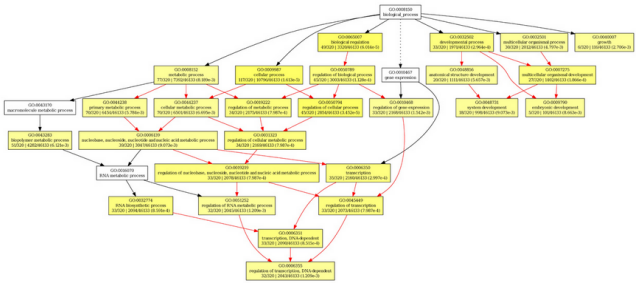
Bioinformatique & visualisation

Bioinformatique et visualisation



- Nécessité de visualiser les informations
  - Des structures
  - Des graphes d'interactions
  - Des séquences
  - Des alignements
  - Des génomes
  - Des annotations / ontologies





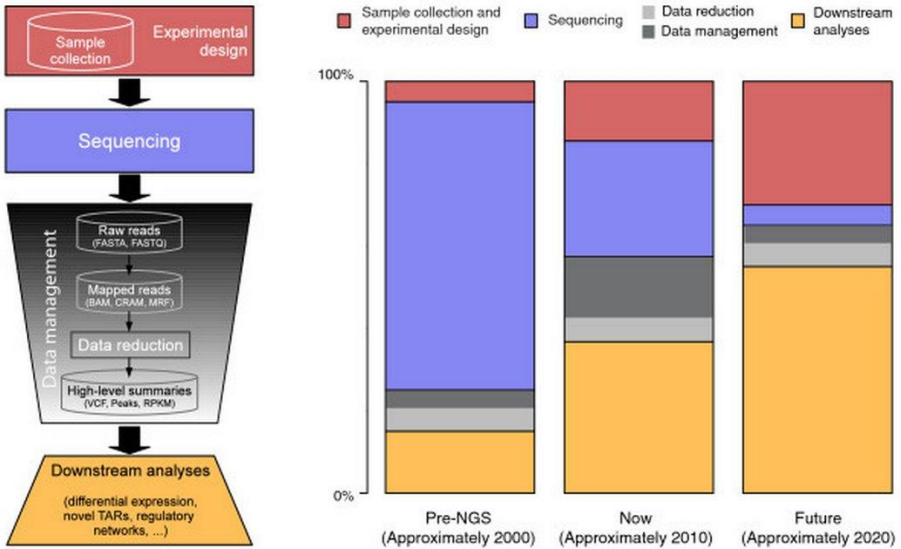
Bioinformatique:
mais encore...

Bioinformatique: mais encore

- Réseau d'interaction (*e.g.* protéine-protéine)
- Modélisation dynamique
- Datamining (*e.g.* analyse de la littérature)
- Biochimie structurale
- ...

The real cost of sequencing: higher than you think!

Andrea Sboner[1,2], **Xinmeng Jasmine Mu**[1], **Dov Greenbaum**[1,2,3,4,5], **Raymond K Auerbach**[1] and **Mark B Gerstein**[1,2,6]*

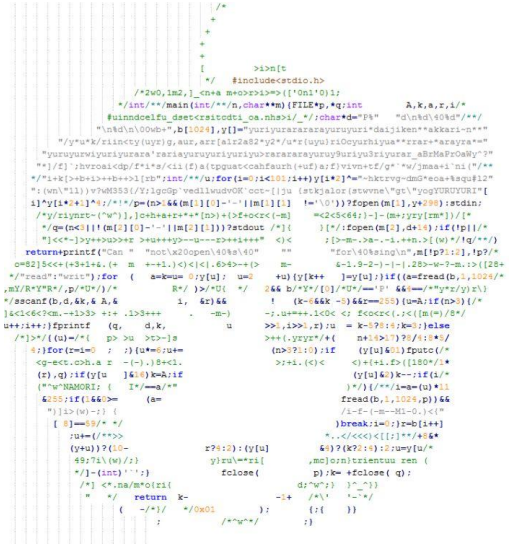


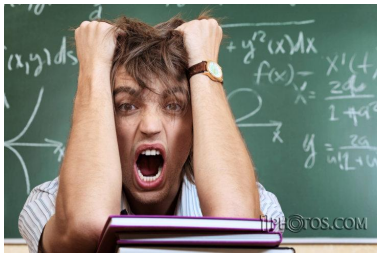
Contribution of different factors to the overall cost of a sequencing project across time. Left, the four-step process: (i) experimental design and sample collection, (ii) sequencing, (iii) data reduction and management, and (iv) downstream analysis. Right, the changes over time of relative impact of these four components of a sequencing experiment. BAM, Binary Sequence Alignment/Map; BED, Browser Extensible Data; CRAM, compression algorithm; MRF, Mapped Read Format; NGS, next-generation sequencing; TAR, transcriptionally active region; VCF, Variant Call Format.

Langages & bioinformatique

Langages et lisibilité

- Attention au code abscons
  - “obfuscating code”
- Nécessité de produire un code lisible
  - Utilisez des commentaires !





Perl

- Développé au départ pour l'analyse de fichiers au format texte (Larry Wall, linguiste).
  - Bénéficie d'une large bibliothèque de modules (i.e; ensemble de fonctions).
    - *e.g.* modules BioPerl pour la bioinformatique
      - Analyse de séquences
  - Moteur d'expressions régulières très puissant

R

- Issu du langage S…
  - Particulièrement utile pour l'analyse statistique des données
    - Microarrays, NGS, Cytométrie de flux, protéomique, analyse d'image,...
    - Bioconductor

      “open source and open development. It has two releases each year, more than 380 packages, and an active user community“
  - Bénéficie d'outils graphiques assez évolués.

SQL (Structured Query Language)

- Permet de communiquer avec des bases de données
  - Interrogation des données via un langage permet de mettre en relation des tables (bases de données relationnelles)

Python

- Une bonne alternative à Perl
  - Syntaxe plus lisible / naturelle
  - Facile à apprendre
  - Nombreux outils pour l'analyse de séquences
  - Nombreux outils pour l’analyse de données (statistiques)
  - Interface graphique développée (e.g. matplotlib)
  - Devenu très populaire
  - Permet le développement d’applications fenêtrées + web

Bash

- Bourne-Again shell
  - Permet d’interagir avec des serveurs sous Unix/Linux
  - Permet de piloter une machine à l’aide d’instructions simples
  - Permet de produire des traitements complexes à partir de commandes simples

La liste complète des langages...

http://fr.wikipedia.org/wiki/Liste_des_langages_de_programmation

Merci

Concentrons nous maintenant sur Bash….