Statistique en grande dimension pour les données génomiques

Franck Picard

UMR CNRS-5558, Laboratoire de Biométrie et Biologie Evolutive

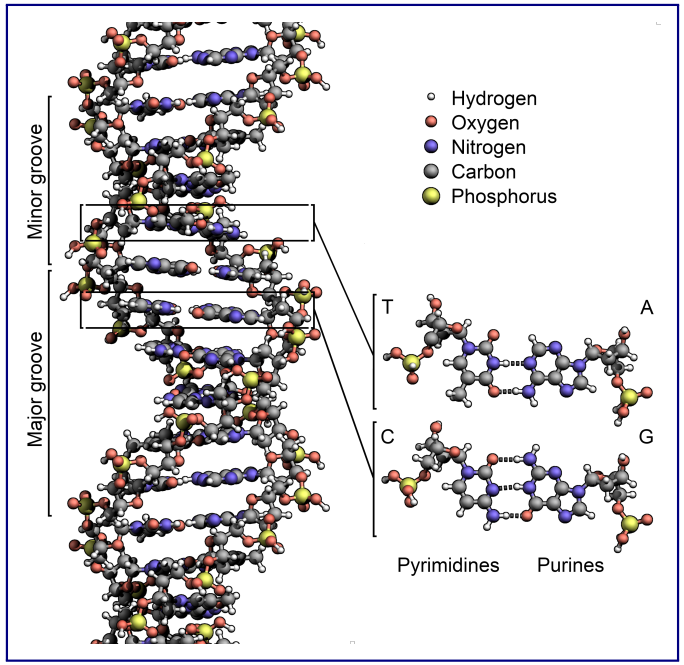
franck.picard@univ-lyon1.fr

January 5, 2017

Outline

Structure de l'ADN et bases moléculaires de l'hérédité

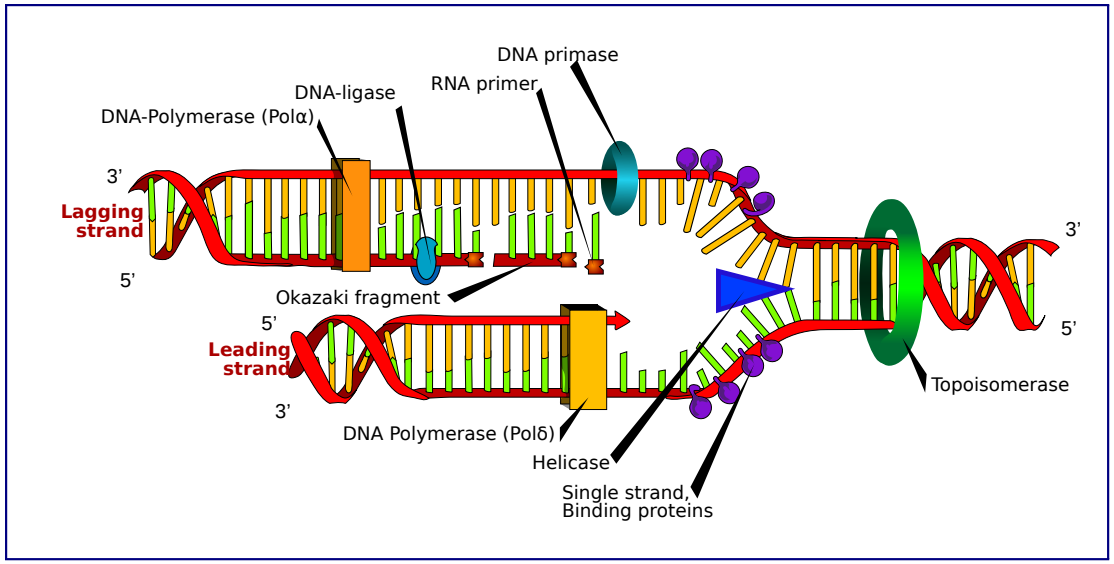
- Polymère double brins
- Monomères : nucléotides/bases
- Adénine, Cytosine, Guanine, Thymine
- Bases **complémentaires** : A-T, G-C
- Double hélice: Watson et Crick (1953)



Complémentarité des deux brins
$\Rightarrow$ on peut déduire un brin à partir de l'autre

La réplication de l'ADN

- La complémentarité des brins est à la base de la réplication
- Des enzymes ont la propriété de répliquer un brin d'ADN
- **Mutations**: erreurs de replication (variations)



Les mutations sont transmises à la descendance par l'intermédiaire de l'une des deux molécules filles

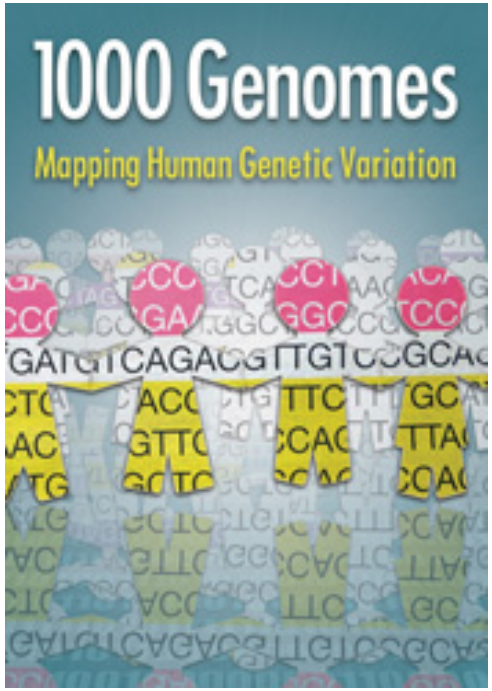
Détection des mutations par séquençage

- Détermine l'enchainement des nucléotides
- Les premières techniques 1975 (Frederick Sanger, Prix Nobel)
- Les séquences sont stockées dans des banques (GeneBank,EMBL)
- Projets de séquençage des génomes entiers (années ∼90)

Variations identifiées par alignement des séquences

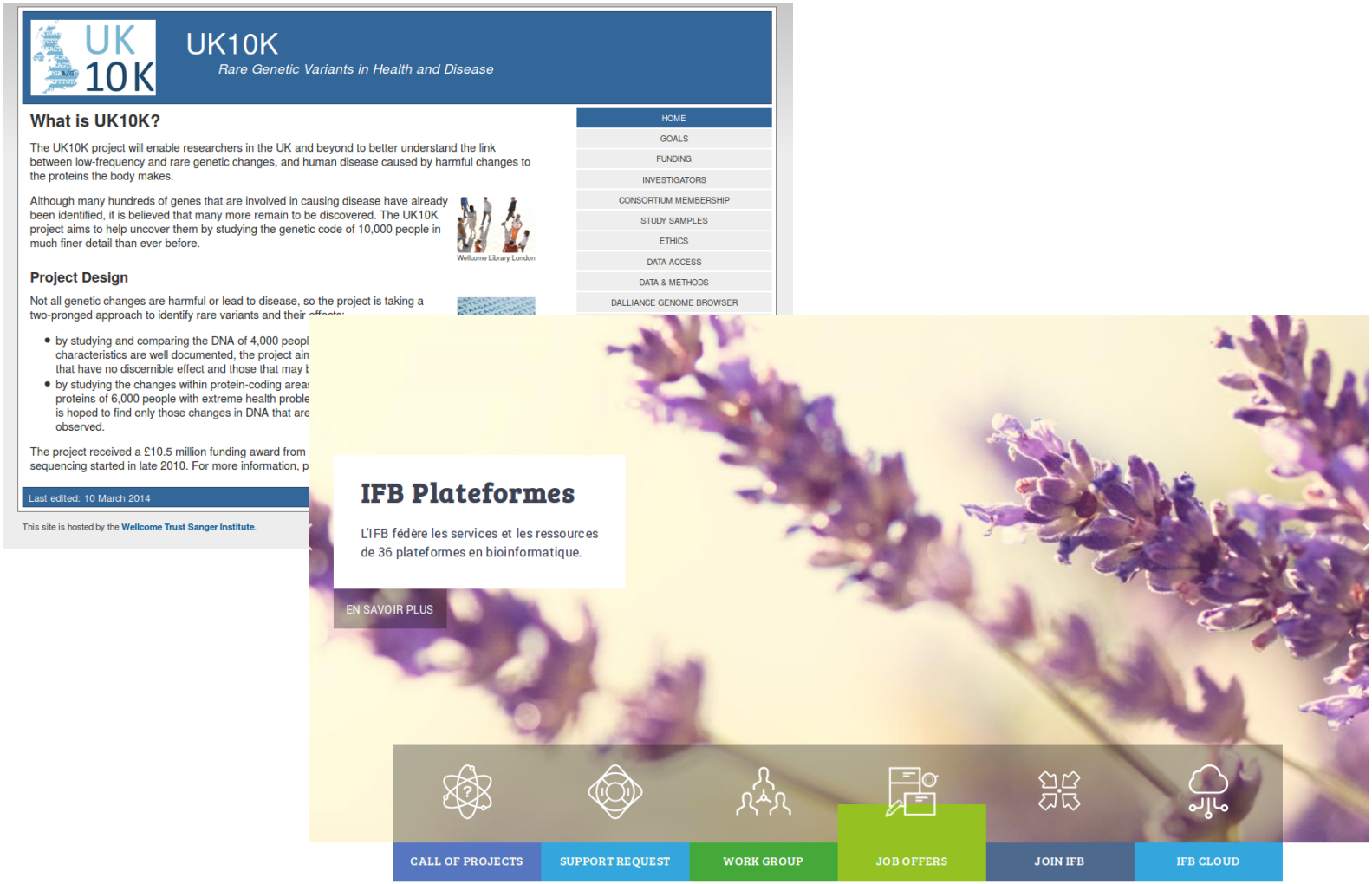
Données populationnelles de séquence

- Génome humain (3.10^9 bases) séquencé en 10 ans
- Aujourd'hui (2011) $\sim 130.10^9$ nucléotides dans $\sim 135.10^6$ séquences disponibles
- Défi technique pour le stockage et la consultation des données
- Echantillonnage des génomes entiers

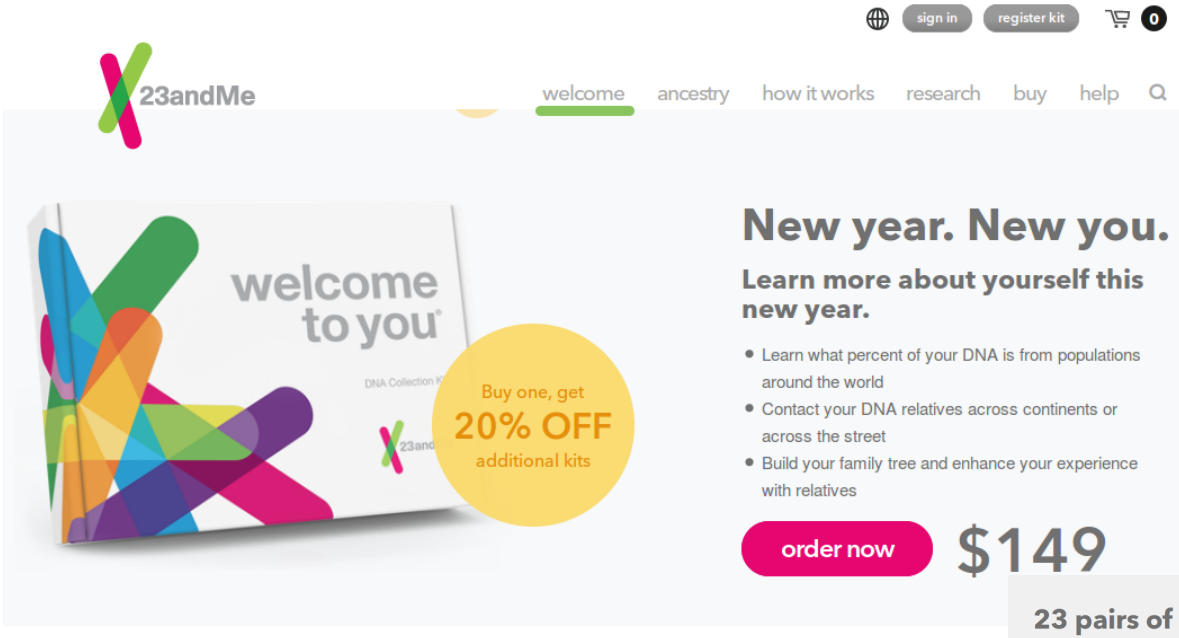


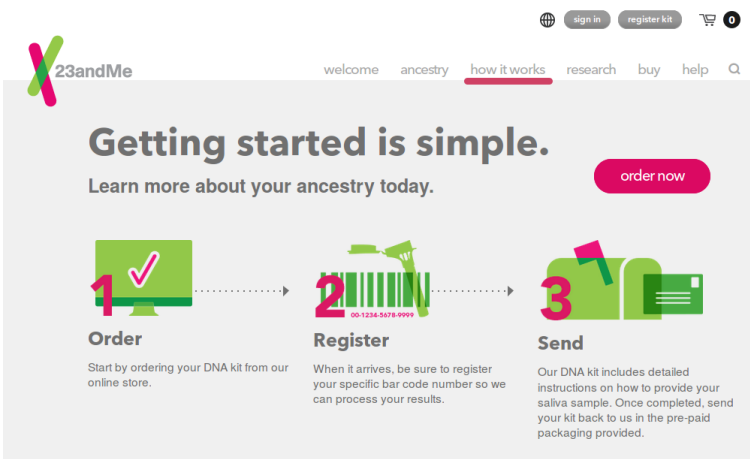
Génomique des populations: étude des variations génétiques des populations à l'échelle des génomes entiers

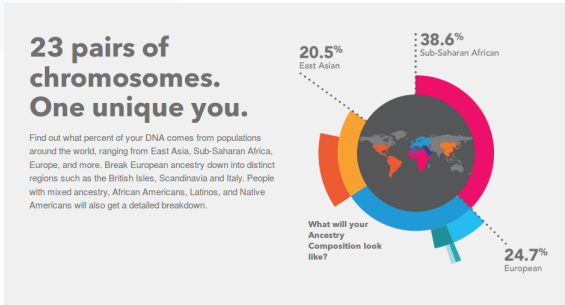
Des initiatives publiques et privées

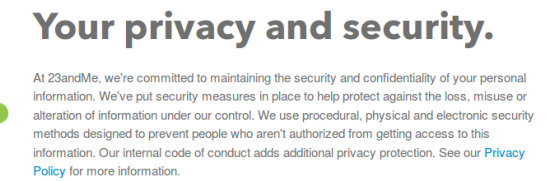


Des initiatives publiques et privées



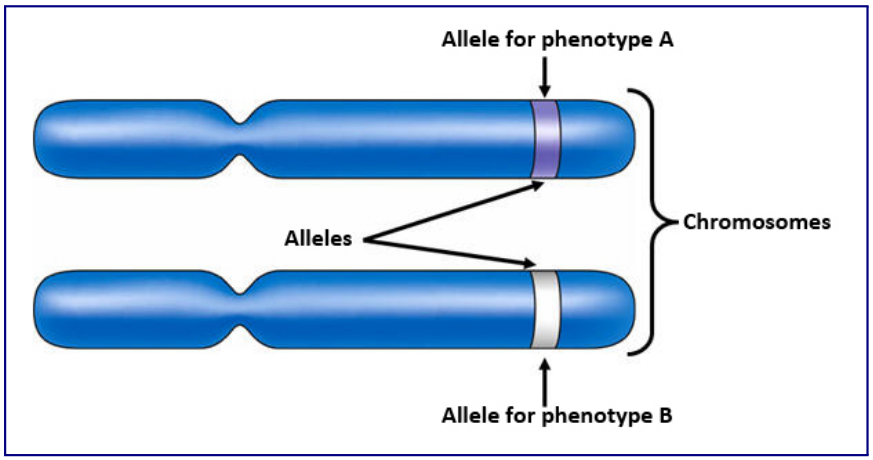






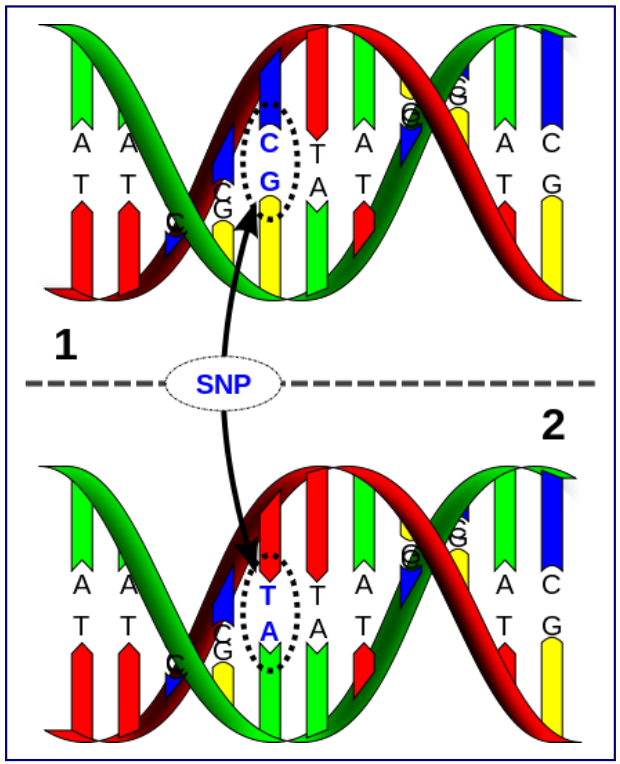
Locus - Allèles

- **Locus**: position sur un chromosome
- La plupart des organismes sont **diploides** (paire d'homologues)
- Certaines différences de séquence peuvent exister entre les deux copies.
- Les différentes formes de chaque gène sont appelées **allèles**



Variations de séquence

- Single Nucleotide Polymorphism: changement d'un unique nucléotide
- A l'origine des différences entre individus d'une même espèce
- 2 séquences humaines se ressemblent à plus de 99%
- Fréquence $> 1\%$ (convention)



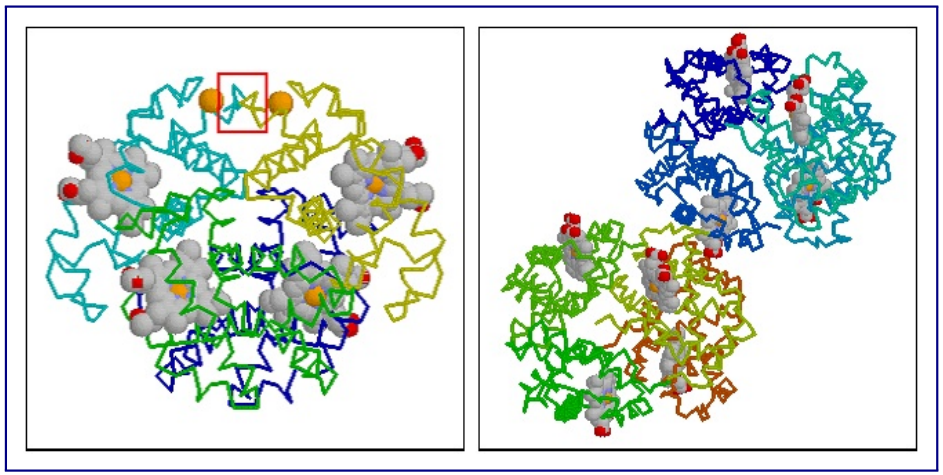
Les SNPs sont responsables de 90% des variations génétiques chez l'homme

Quelques ordres de grandeur

- En moyenne, on trouve un SNP tous les 600bp chez l'homme (régions hot-spots et régions pauvres)
- 10 millions de SNPs dans 3.2 milliards de nucléotides chez l'homme
- L'abondance des SNPs dépend des espèces (1 SNP tous les 50-100pb chez la mouche),

Variations de séquence - Variations de fonctions

- Altération de la fonction d'une protéine
- Modification de l'efficacité d'une enzyme (quantitatif)
- Aucun effet
- Modification de la régulation / épissage

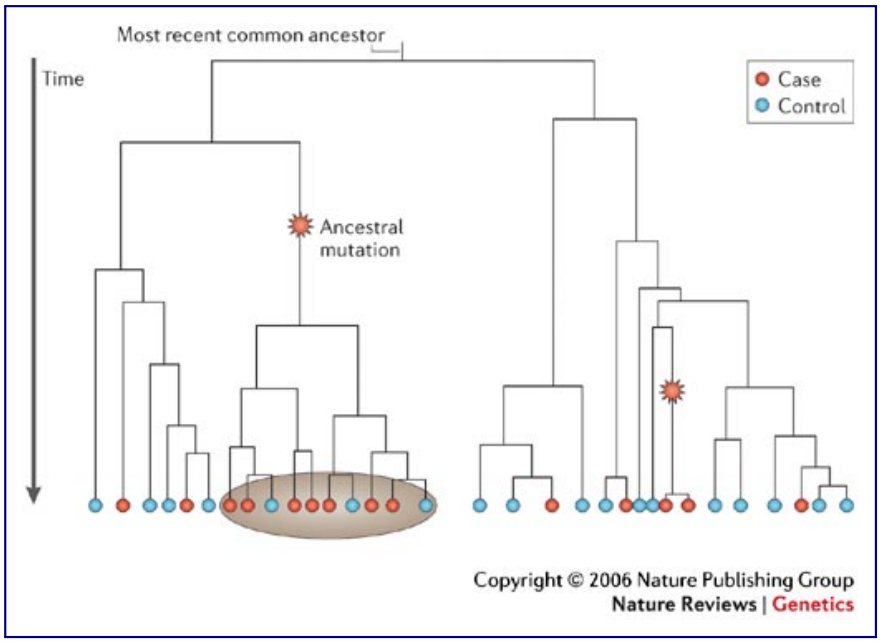


~60,000 SNPs sont dans les exons

Outline

Etudes d'association

- Identification des variations jouant un rôle dans la détermination de phénotypes mesurables
- Sur des individus non apparentés
- Identifier des SNPs qui varient systématiquement entre individus ayant différents états



Identifier la "sur-représentation" de certains variants dans les cas par rapport aux témoins.

Structure des données

	statut	snp_1	snp_2	...	snp_p	Age	Sexe	Glycémie
$i = 1$	0	0	1		0	38	F	0.8
$i = 2$	1	1	0		2	15	M	0.2
$\vdots$								
$i = N$	1	2	2		1	90	F	1.5

Pour chaque individu:

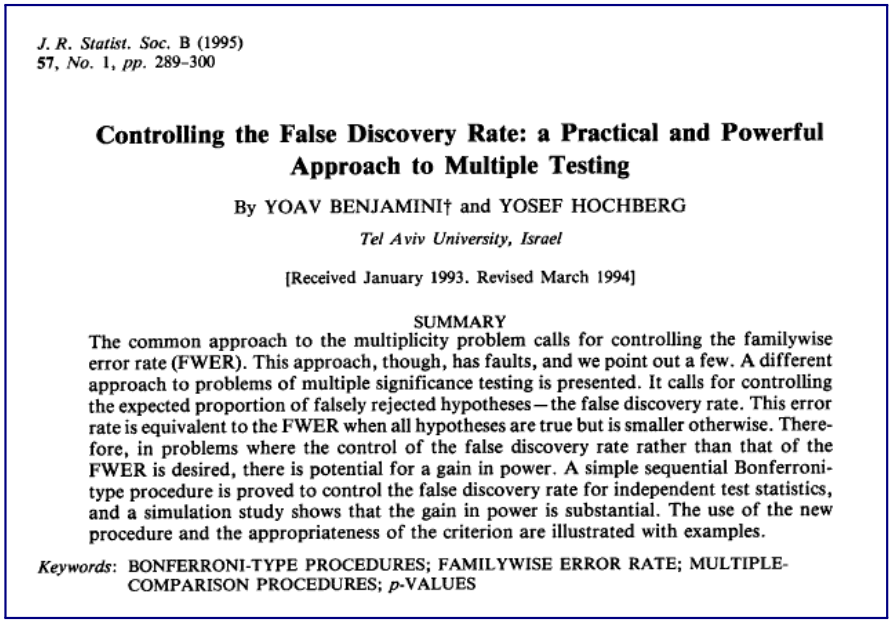
- Statut (discret: association / continu : QTL)
- Génotype mesuré sur p SNPs
- Données cliniques (non génomiques)

Objectif

Expliquer les variations d'une réponse en fonction de covariables génomiques (et cliniques)

Stratégies univariées

- Historiquement la plus utilisée
- p tests (chi-square ou Student)
- Dépendance? Tests multivariés?
- Comment inclure d'autres informations (poids, âge, clinique) ?
- Multiplicité des tests



J. R. Statist. Soc. B (1995)
57, *No.* 1, *pp.* 289–300

Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing

By YOAV BENJAMINI† and YOSEF HOCHBERG

Tel Aviv University, Israel

[Received January 1993. Revised March 1994]

SUMMARY

The common approach to the multiplicity problem calls for controlling the familywise error rate (FWER). This approach, though, has faults, and we point out a few. A different approach to problems of multiple significance testing is presented. It calls for controlling the expected proportion of falsely rejected hypotheses—the false discovery rate. This error rate is equivalent to the FWER when all hypotheses are true but is smaller otherwise. Therefore, in problems where the control of the false discovery rate rather than that of the FWER is desired, there is potential for a gain in power. A simple sequential Bonferroni-type procedure is proved to control the false discovery rate for independent test statistics, and a simulation study shows that the gain in power is substantial. The use of the new procedure and the appropriateness of the criterion are illustrated with examples.

Keywords: BONFERRONI-TYPE PROCEDURES; FAMILYWISE ERROR RATE; MULTIPLE-COMPARISON PROCEDURES; *p*-VALUES

La thématique des tests multiples (stat math) a connu un regain d'intérêt suite à l'émergence des données à haut débit

Modèles linéaires (généralisés) pour les études d'association

- On note Y_i la réponse (qualitative ou quantitative) et $\mathbf{x}_i$, $\mathbf{z}_i$ les covariables génomiques et non génomiques
- On suppose que les effets sont additifs dans un premier temps:

$$g(\mathbb{E}(Y_i|\mathbf{x}_i, \mathbf{z}_i)) = \sum_{j=1}^{p} \beta_j x_{ij} + \sum_{q=1}^{Q} \theta_{iq} z_{iq}$$

- La problématique est d'identifier les composantes du vecteur $\boldsymbol{\beta}$ qui correspondent à des SNPs ayant un effet sur la réponse
- La difficulté de l'exercice est liée au nombre important de variables p par rapport aux individus N.

Problèmes associés à la grande dimension

- Cas ou la taille du paramètre à estimer p est plus grande que le nombre d'observations n
- Inversion de $X^T X$ impossible
- Corrélation artéfactuelles entre régresseurs
- Problèmes de stockage

L'hypothèse de parcimonie

- On fera l'hypothèse que la majorité des SNPs n'ont pas d'effet, donc on supposera une structure creuse pour le vecteur de paramètres.
- On cherche à estimer le vecteur $\boldsymbol{\beta}$ en prenant en compte certaines contraintes

$$\widehat{\boldsymbol{\beta}}_0 = \underset{\boldsymbol{\beta}}{\text{Argmax}} \left\{ \log \mathcal{L}(\mathbf{Y}, \boldsymbol{\beta}, \boldsymbol{\theta}) \right\} \text{ avec } \sum_{j=1}^{p} \mathbb{I}\{\beta_j \notin 0\} \leq C$$

LASSO et Régularisation L_1

- Mais ce problème d'optimisation n'est pas convexe. On en utilise une relaxation:

$$\widehat{\boldsymbol{\beta}}_1 = \underset{\boldsymbol{\beta}}{\text{Argmax}} \{\log \mathcal{L}(\mathbf{Y}, \boldsymbol{\beta}, \boldsymbol{\theta})\} \text{ avec } \sum_{j=1}^{p} |\beta_j| \leq C$$

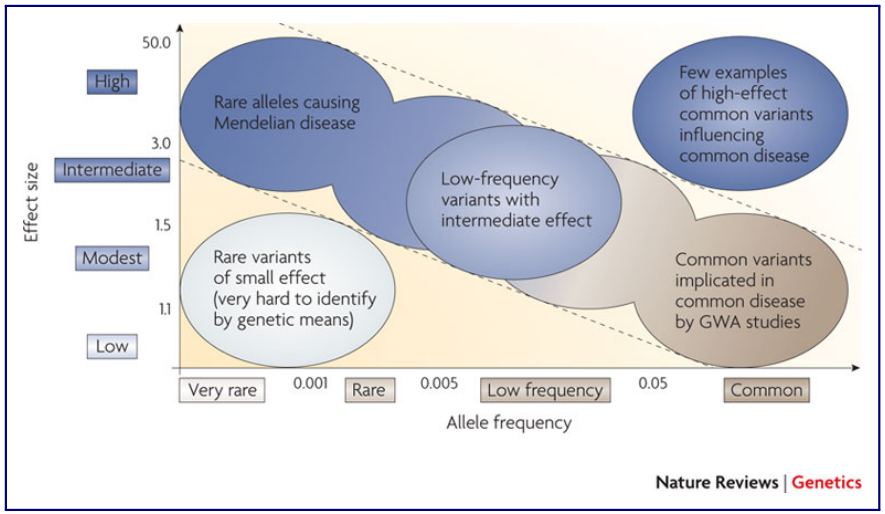
- Ce problème d'optimisation peut également s'écrire sous la forme:

$$\widehat{\boldsymbol{\beta}}_1 = \underset{\boldsymbol{\beta}}{\text{Argmax}} \left\{ \log \mathcal{L}(\mathbf{Y}, \boldsymbol{\beta}, \boldsymbol{\theta}) - \lambda_C \sum_{j=1}^{p} |\beta_j| \right\}$$

L'objectif de ce cours est d'étudier les méthodes de régression pénalisées, les techniques d'optimisation associées, et les propriétés statistiques des estimateurs

Difficultés inhérentes aux études d'association

- Cas facile: SNP causal lien direct génotype/phénotype
- Maladies "complexes": la mesure du phénotype est incertaine. Les liens génotype/phénotype sont partiellement connus.
- L'impact environnemental est supposé important



La fréquence et la taille des effets des variants sont deux composantes majeures

Modèles de régression utilisés en génétique quantitative

- Le modèle général en génétique quantitative permet de reglier un phénotype observé à des composantes génétiques et environnementales: $P = G + E + G \times E$
- Si on suppose toutes les composantes indépendantes, alors on aura une décomposition de la variance telle que: $V_P = V_G + V_E + V_{G \times E}$
- On peut définir des héritabilités au sens large comme le ratio des variances $H^2 = V_G / V_P$
- On considère également l'héritabilité au sens strict qui concerne la partie additive de la variance génétique $h^2 = V_A / V_P$

Missing Heritability ?

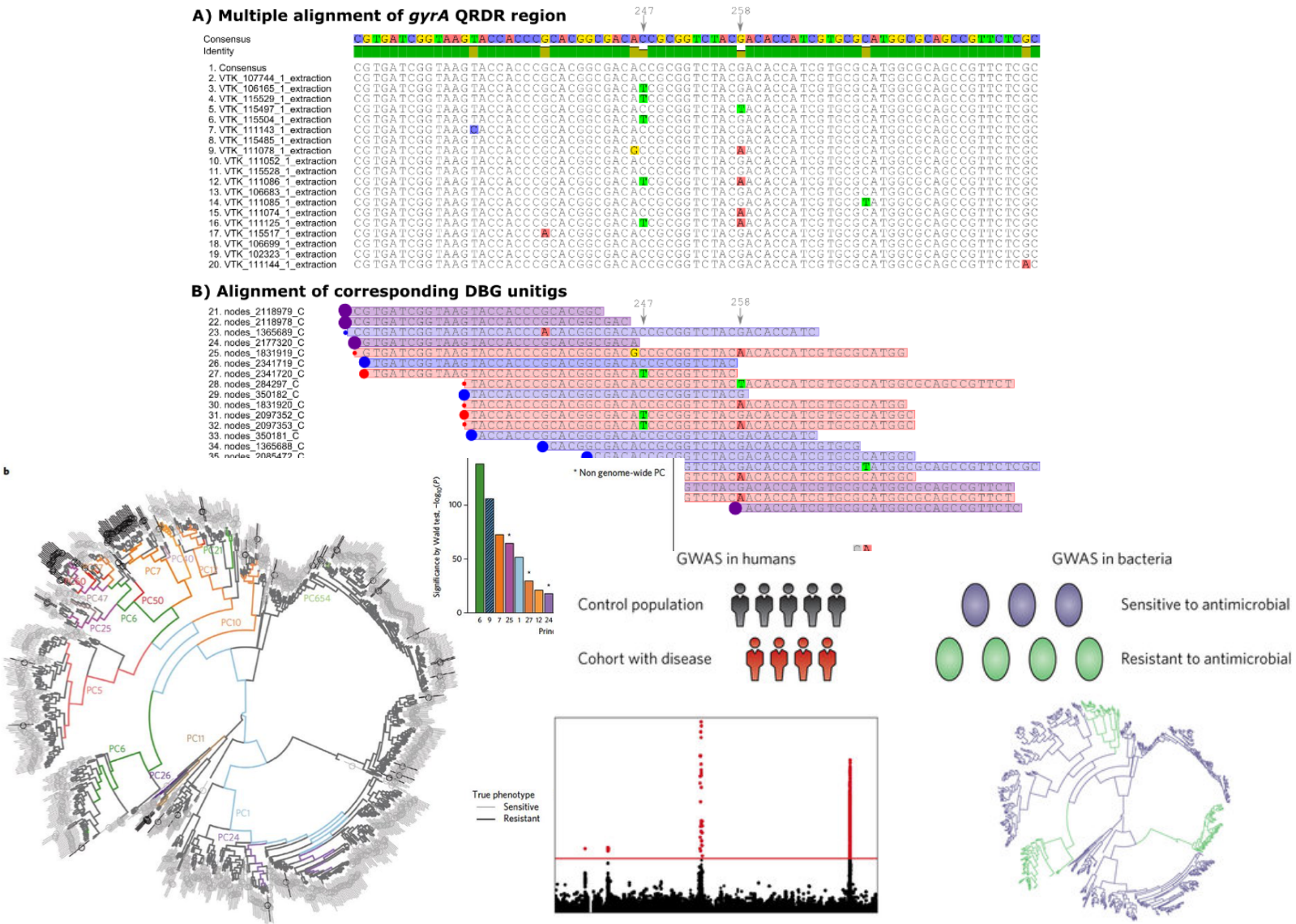
- Malgré les études d'association, il manque une part non négligeable de variabilité qui reste inexpliquée
- Prise en compte d'autres variations
- Impact des variants rares
- Dépendances entre SNPs

The case of the missing heritability

When scientists opened up the human genome, they expected to find the genetic components of common traits and diseases. But they were nowhere to be seen. **Brendan Maher** shines a light on six places where the missing loot could be stashed away.

Les sources de variation inter-individuelles doivent être mieux prises en compte par les modèles statistiques

GWAS Bactériennes

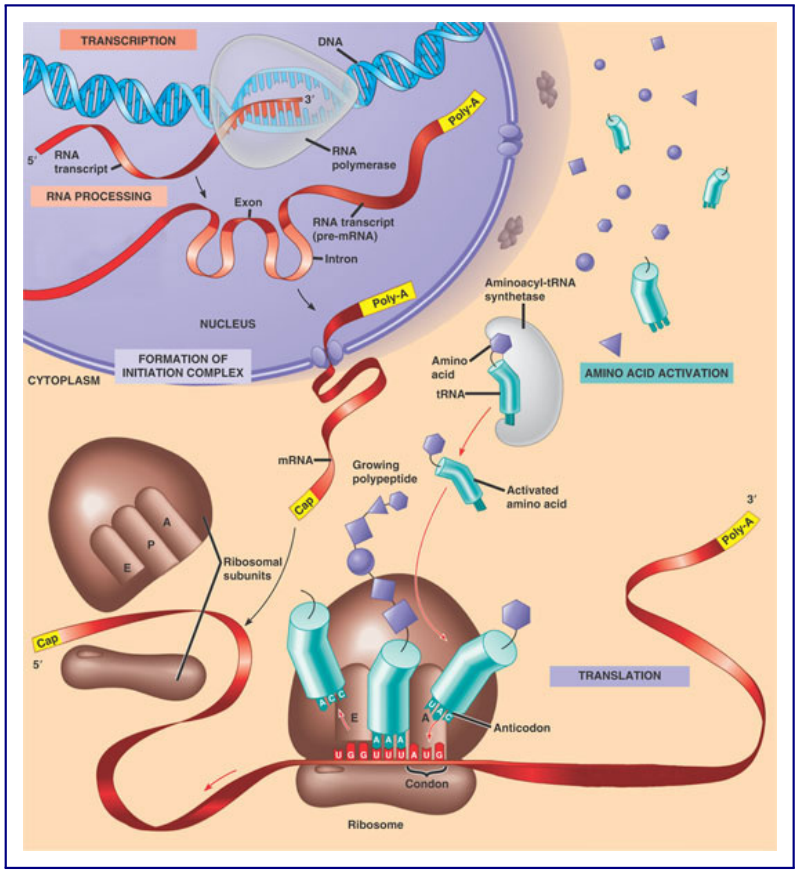


Outline

Central Dogma of molecular biology (∼1970)

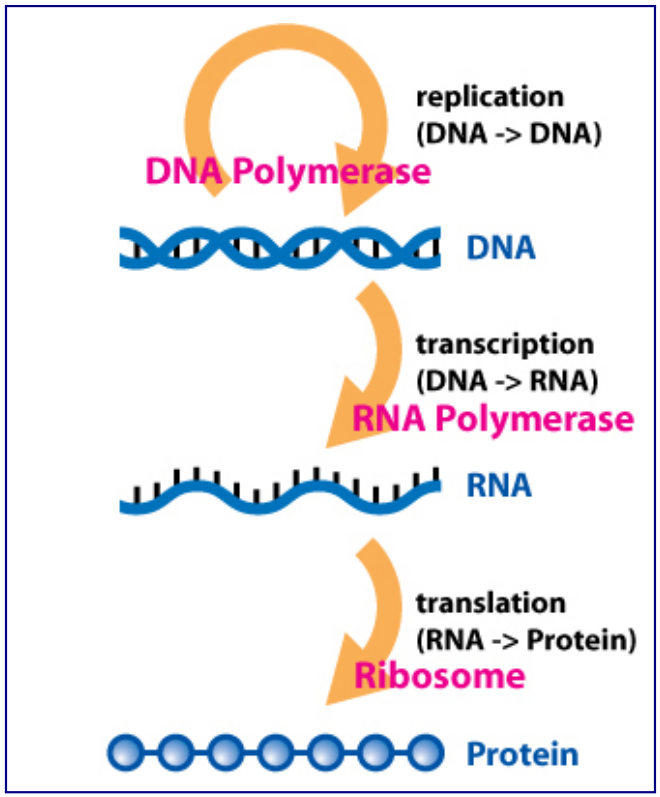
- Relier l'information génétique (ADN) au fonctionnement de la cellule (protéines) grâce au code génétique

Trois niveaux d'information génétique: ADN, ARN, protéines



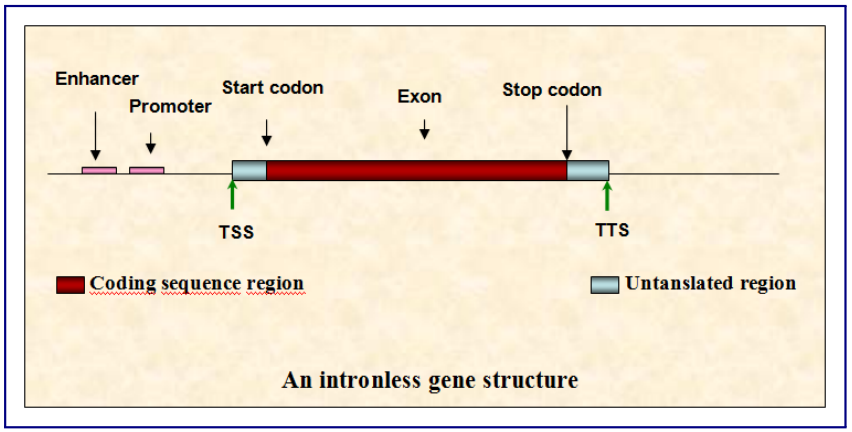
Central Dogma of molecular biology

- L'information contenue dans la séquence d'ADN est **transcrite** en ARN messager
- l'ARNm est un polymère simple brin, c'est un **vecteur**
- L'ARN messager est **traduit** en protéines grâce au code génétique
- Les protéines sont les effecteurs de l'information génétique
- polymères constitués de 20 Acides Aminés différents



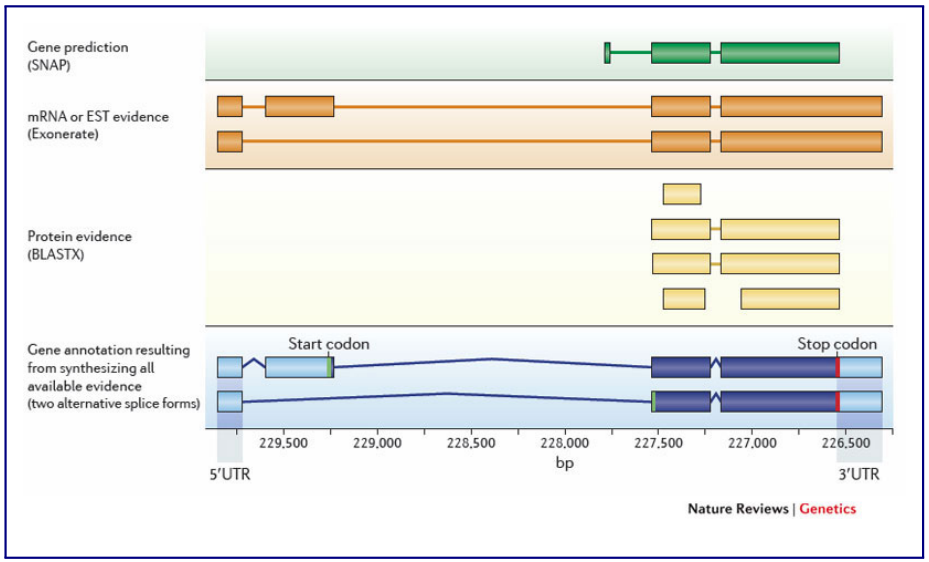
Qu'est ce qu'un gène ? (1)

- Une seule séquence codante transcrite
- Sites de fixation des régulateurs (Facteurs de Transcription)
- Vision très simpliste :
  1 gène $\Rightarrow$ 1 protéine



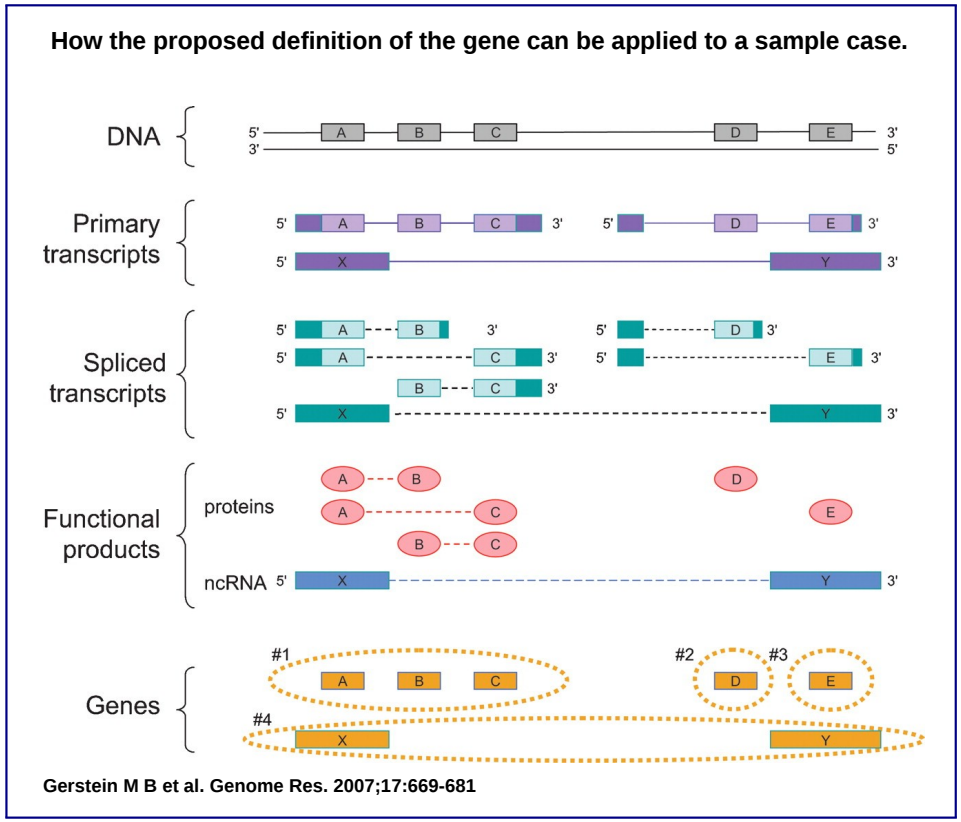
Qu'est ce qu'un gène ? (2)

- Une seule séquence codante transcrite
- Sites de fixation des régulateurs (Facteurs de Transcription)
- Vision très simpliste :
  1 gène $\Rightarrow$ 1 protéine
- Alternance de zones codantes-non codantes:
  Exons/Introns



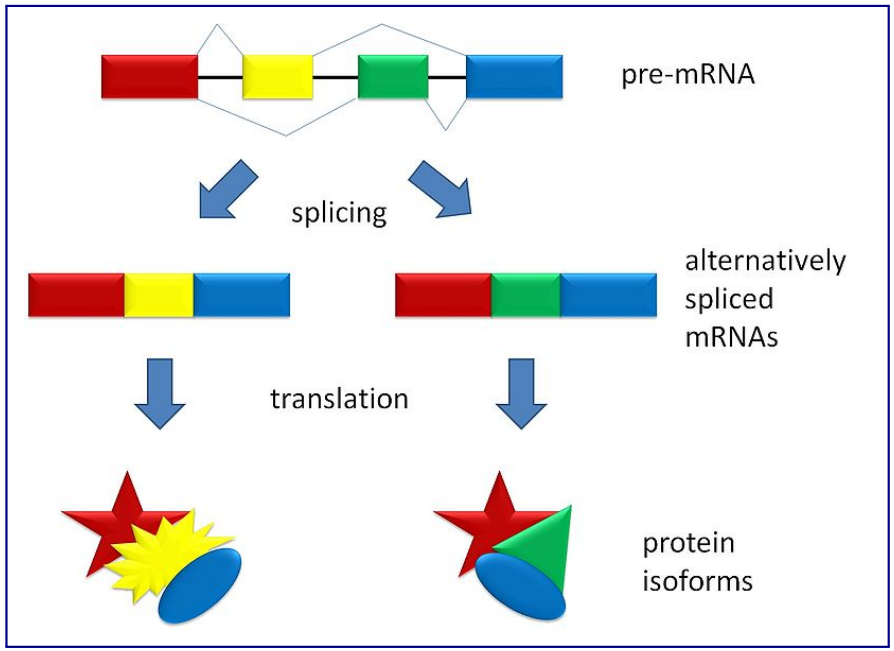
Qu'est ce qu'un gène ? (3)

- un segment d'ADN qui contribue au phénotype
- 20,000 gènes codants qui représentent 1.2% du génome humain
- 5-10% de séquences codants des produits non protéiques
- Rôle fondamental des séquences régulatrices



Epissage alternatif

- Nb gènes découverts stable
- Nb de formes différentes : 5,4 formes différentes par gène (GENCODE)
- $>$ 50% des gènes ont un site d'initiation de la transcription alternatif
- Source de diversité et d'erreurs

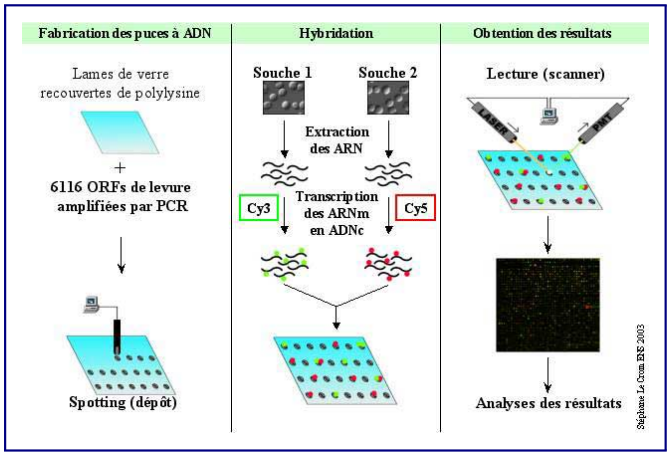


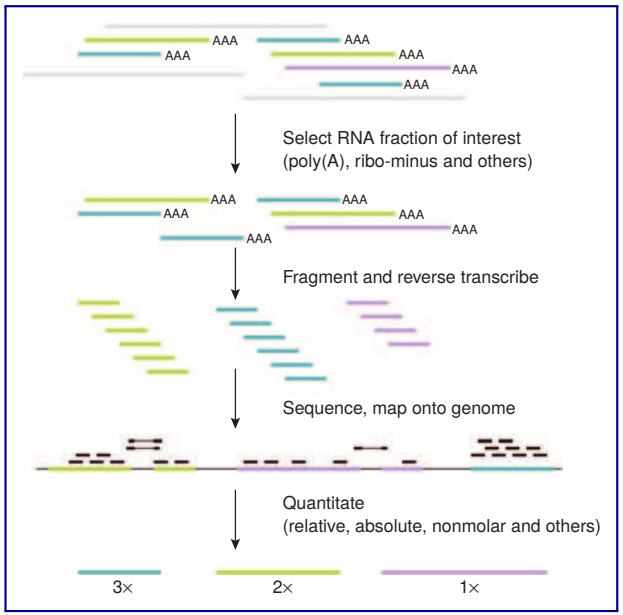
Modification de l'agencement des exons (ex: différents tissus)

Des puces à ADN au séquençage massif

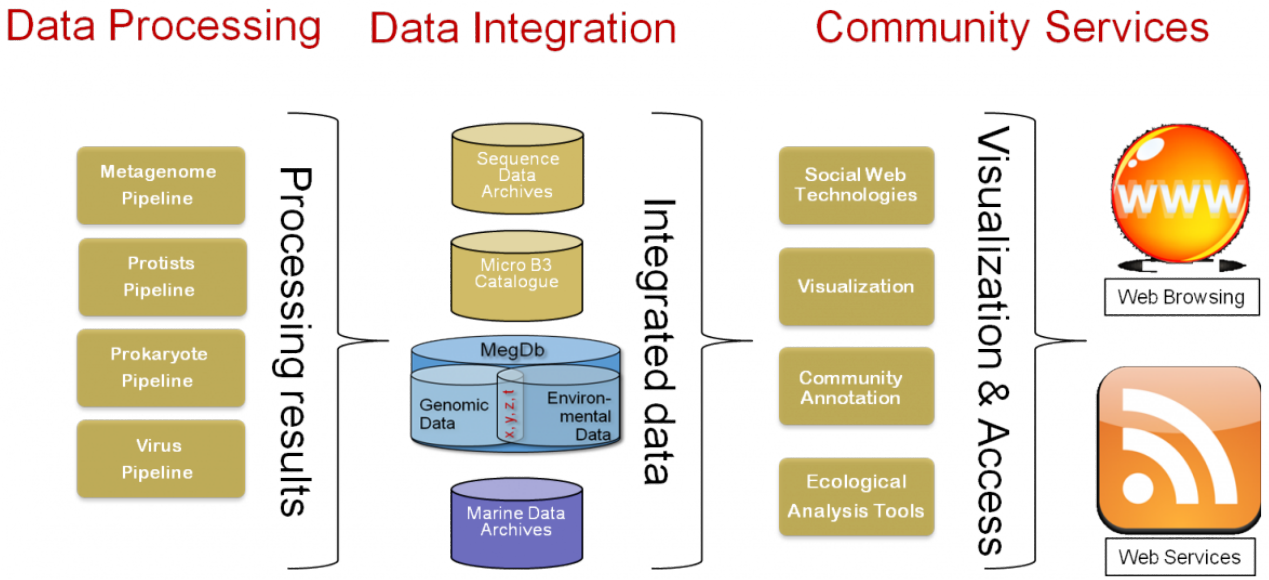
- Années 1990-2000: puces à ADN (microarrays, chips)
- Depuis 2000: quantification de l'abondance des mRNAs par séquençage massif
- NGS : next generation sequencing

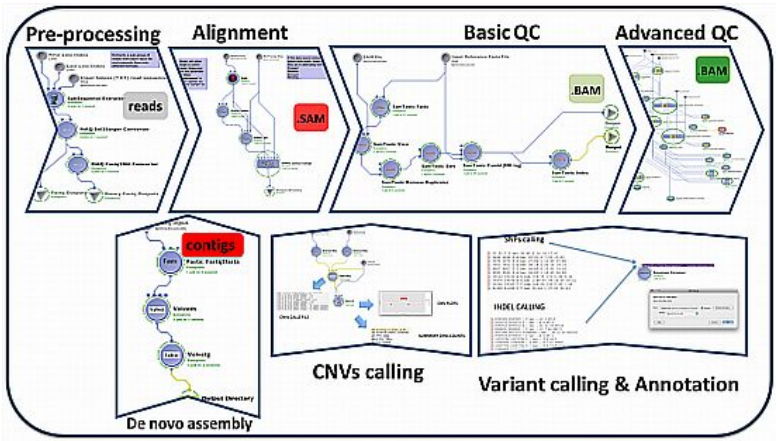
Etude du fonctionnement de la cellule à l'échelle génomique
Genome-Wide Studies

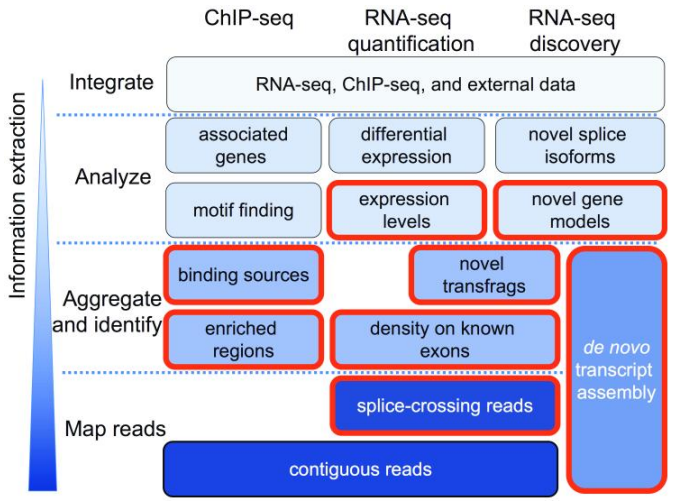




Des pipelines et encore des pipelines !







Comparaison des méthodes et nécessité de standards

BRIEFINGS IN BIOINFORMATICS. VOL 14. NO 6. 671–683
Advance Access published on 17 September 2012
doi:10.1093/bib/bbs046

A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis

Marie-Agnès Dillies, Andrea Rau*, Julie Aubert*, Christelle Hennequet-Antier*, Marine Jeanmougin*, Nicolas Servant*, Céline Keime*, Guillemette Marot, David Castel, Jordi Estelle, Gregory Guernec, Bernd Jagla, Luc Jouneau, Denis Laloë, Caroline Le Gall, Brigitte Schaëffer, Stéphane Le Crom*, Mickaël Guedj*, Florence Jaffrézic* and on behalf of The French StatOmique Consortium*

Submitted: 12th April 2012; Received (in revised form): 29th June 2012

Conesa *et al. Genome Biology* (2016) 17:13
DOI 10.1186/s13059-016-0881-8

Genome Biology

REVIEW Open Access

A survey of best practices for RNA-seq data analysis

Ana Conesa[1,2*], Pedro Madrigal[3,4*], Sonia Tarazona[2,5], David Gomez-Cabrero[6,7,8,9], Alejandra Cervera[10], Andrew McPherson[11], Michał Wojciech Szcześniak[12], Daniel J. Gaffney[3], Laura L. Elo[13], Xuegong Zhang[14,15] and Ali Mortazavi[16,17*]

OXFORD JOURNALS
Briefings in Bioinformatics
ABOUT THIS JOURNAL CONTACT THIS JOURNAL SUBSCRIPTIONS CURR
Institution: UCBL SCD Lyon 1 Sign In as Personal Subscriber
Oxford Journals › Science & Mathematics › Briefings in Bioinformatics › Volume 16, Issue 6 › Pp. 932-940.

A comparative study of RNA-seq analysis strategies

Jürgen Jänes*
Jürgen Jänes is a PhD student enrolled in the Wellcome Trust Mathematical Genomics and Medicine programme.

Fengyuan Hu*
Fengyuan Hu is a scientific programmer in the Department of Haematology, University of Cambridge.

Alexandra Lewin
Alexandra Lewin is a Lecturer in Biostatistics at the Department of Epidemiology and Biostatistics, Imperial College London.

Ernest Turro
Ernest Turro is a Senior Research Associate at the Department of Haematology, University of Cambridge and Visiting Worker at the MRC Biostatistics Unit. He has developed statistical methods for analysing RNA sequencing data and has co-authored a book chapter on the subject with Alexandra Lewin. Over the past 5 years, he has taught bioinformatics on numerous courses in the UK and abroad.

Corresponding author. Ernest Turro, Department of Haematology, University of Cambridge and MRC Biostatistics Unit. Tel.: +44 (0)1223 588174; E-mail: et341@cam.ac.uk

Received December 16, 2014.
Revision received January 14, 2015.

SCIENTIFIC REPORTS

OPEN

Comparative assessment of methods for the fusion transcripts detection from RNA-Seq data

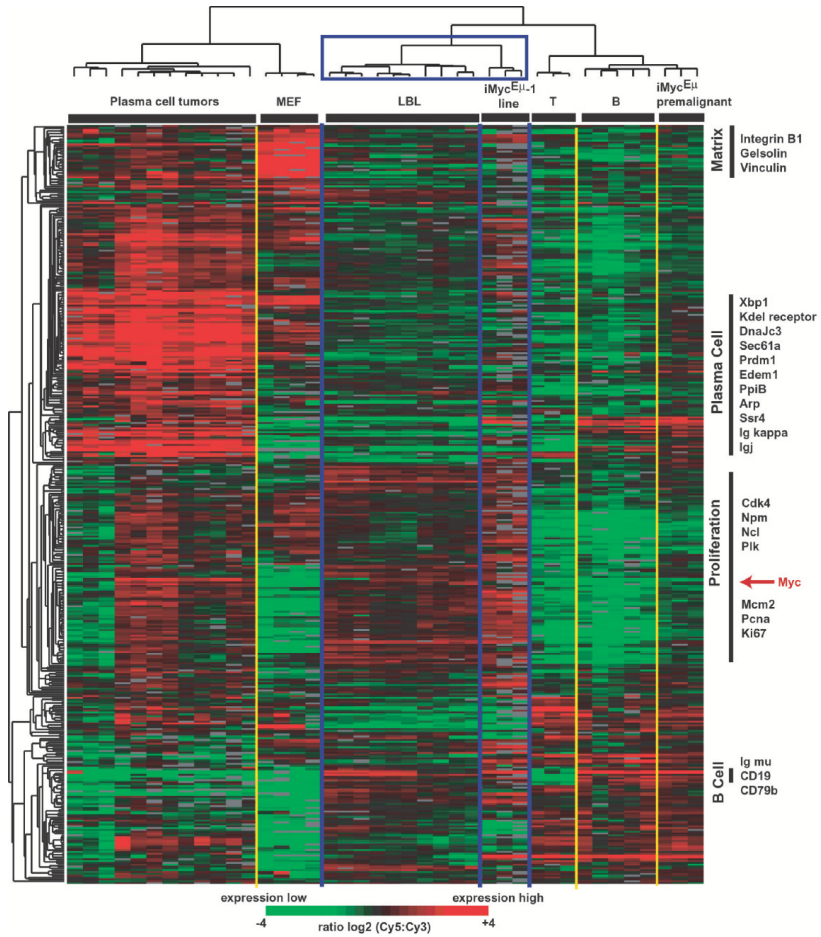
Received: 08 October 2015
Accepted: 27 January 2016
Published: 10 February 2016

Shailesh Kumar[1], Angie Duy Vo[1], Fujun Qin[1] & Hui Li[1,2]

RNA-Seq made possible the global identification of fusion transcripts, i.e. "chimeric RNAs". Even though various software packages have been developed to serve this purpose, they behave differently

Un changement d'échelle (conceptuelle)

- Avant le haut débit l'étude de l'expression des gènes se faisait gène par gène
- Les techno Haut Débit ont complètement changé le point de vue des biologistes
- en une seule expérience, on dispose du niveau d'expression de (potentiellement) tous les transcrits d'une cellule !



Structure des données

	statut	exon_1	exon_2	...	exon_p	Age	Sexe	Glycémie
$i = 1$	0	10000	50		0	38	F	0.8
$i = 2$	1	10000	30		1	15	M	0.2
$\vdots$								
$i = N$	1	20000	25		3	90	F	1.5

Pour chaque individu:

- Statut (discret/continu)
- expression des gènes mesurées (comptages ou continu)
- Données cliniques (non génomiques)

Objectif

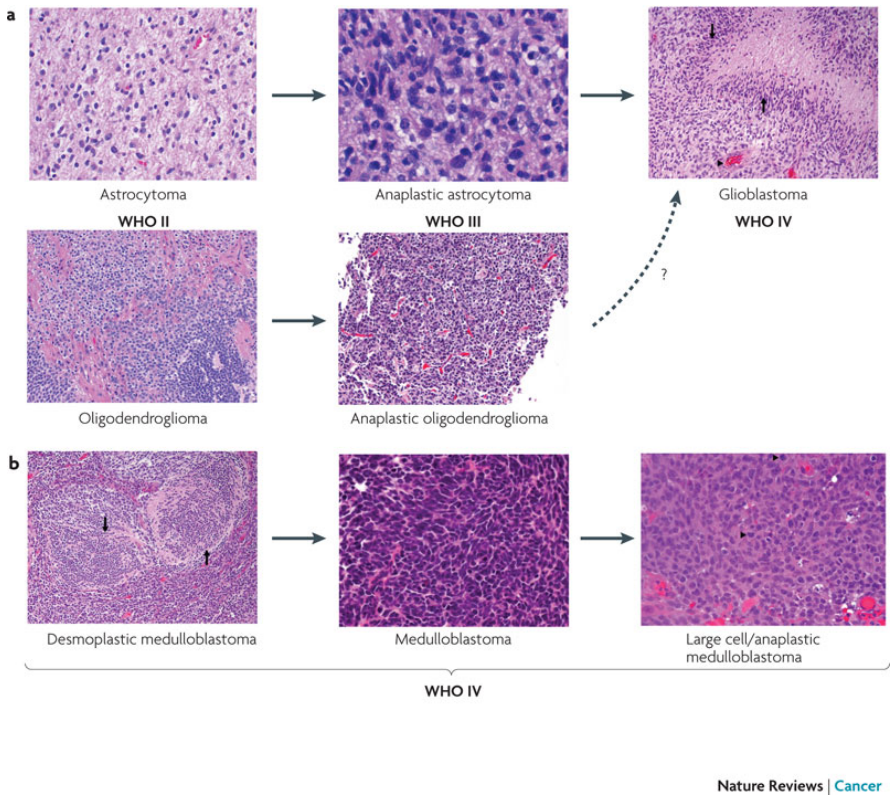
Expliquer les variations d'une réponse en fonction des niveaux d'expression des gènes (aspects fonctionnels).

Exemples de problématiques d'analyse

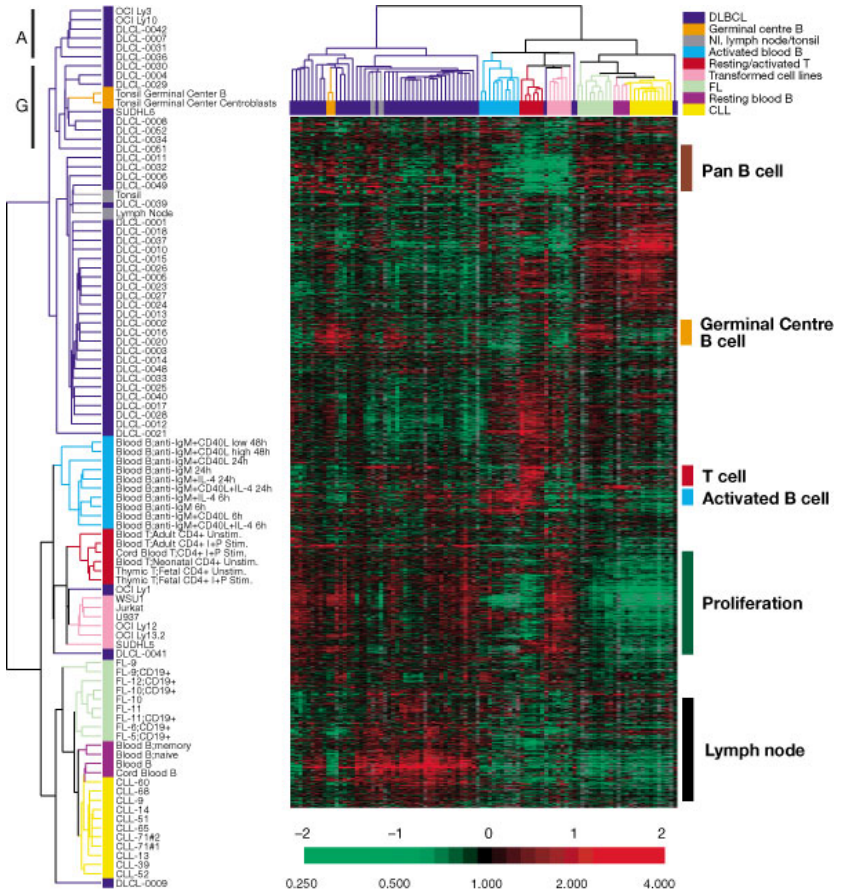
- Planification expérimentale
- Analyse différentielle d'expression
- Classification non supervisée d'individus (découverte de sous-groupes)
- Classification supervisée d'individus (prédiction de phénotype)

Tâches statistiques standard mais le nombre élevé de variables étudiées nécessite de revisiter les techniques classiques

Nouvelles classifications et traitements personnalisés



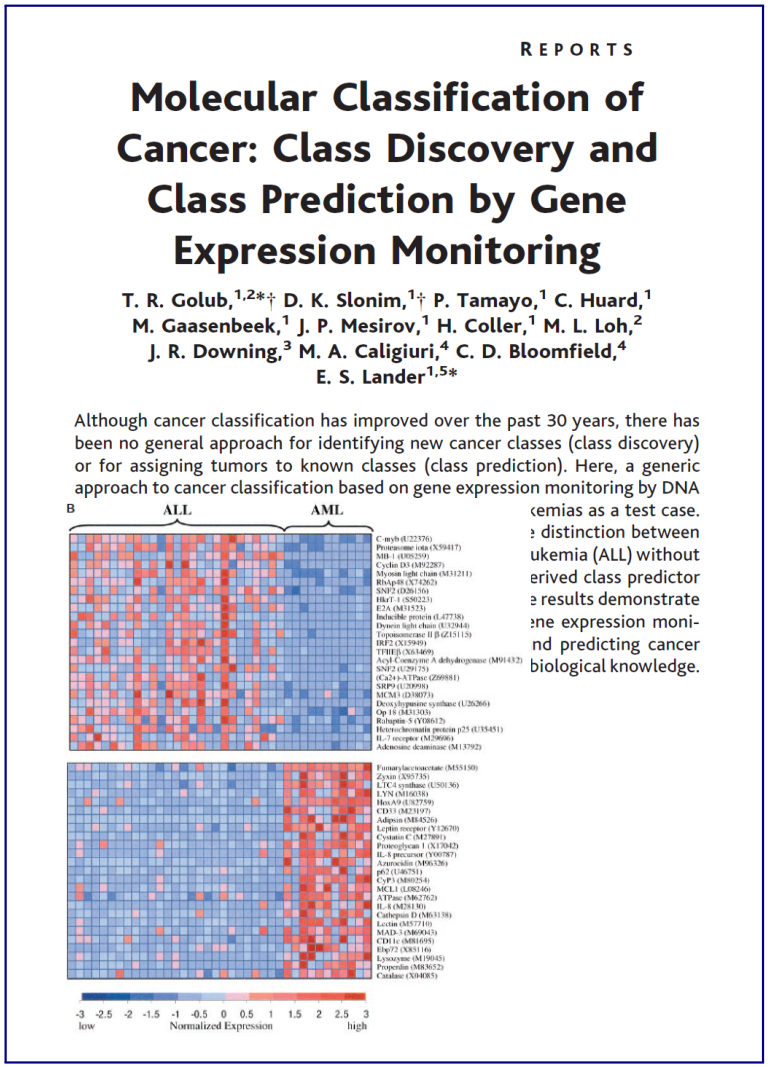
Classification sur critères histologiques



Classification sur critères moléculaires

Vers de nouvelles prédictions ?

- En 1999 un article fait sensation en proposant une prédiction du statut moléculaire d'individus atteints de deux types de leucémie à partir de signatures génomiques
- Le nombre d'individus est 38 pour 6817 gènes étudiés !
- Développement de la thématique de l'apprentissage statistique aux données génomiques



REPORTS

Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring

T. R. Golub,[1,2*†] D. K. Slonim,[1†] P. Tamayo,[1] C. Huard,[1] M. Gaasenbeek,[1] J. P. Mesirov,[1] H. Coller,[1] M. L. Loh,[2] J. R. Downing,[3] M. A. Caligiuri,[4] C. D. Bloomfield,[4] E. S. Lander[1,5*]

Although cancer classification has improved over the past 30 years, there has been no general approach for identifying new cancer classes (class discovery) or for assigning tumors to known classes (class prediction). Here, a generic approach to cancer classification based on gene expression monitoring by DNA

B ALL AML

-3 -2.5 -2 -1.5 -1 -0.5 0 0.5 1 1.5 2 2.5 3
low Normalized Expression high

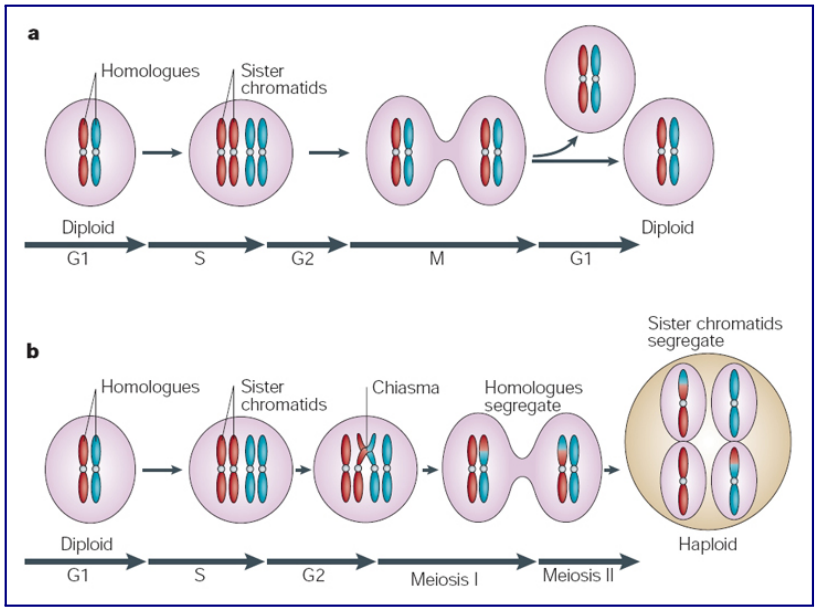
Outline

Prise en compte de la dépendance entre SNPs

- La modélisation prend-elle suffisamment en compte les connaissances disponibles ?
- Le modèle qui suppose l'indépendance entre variables génomiques est-il pertinent ?
- Les processus biologiques à l'origine des variations génétiques sont fortement structurés le long du génome (réplication / recombinaison)

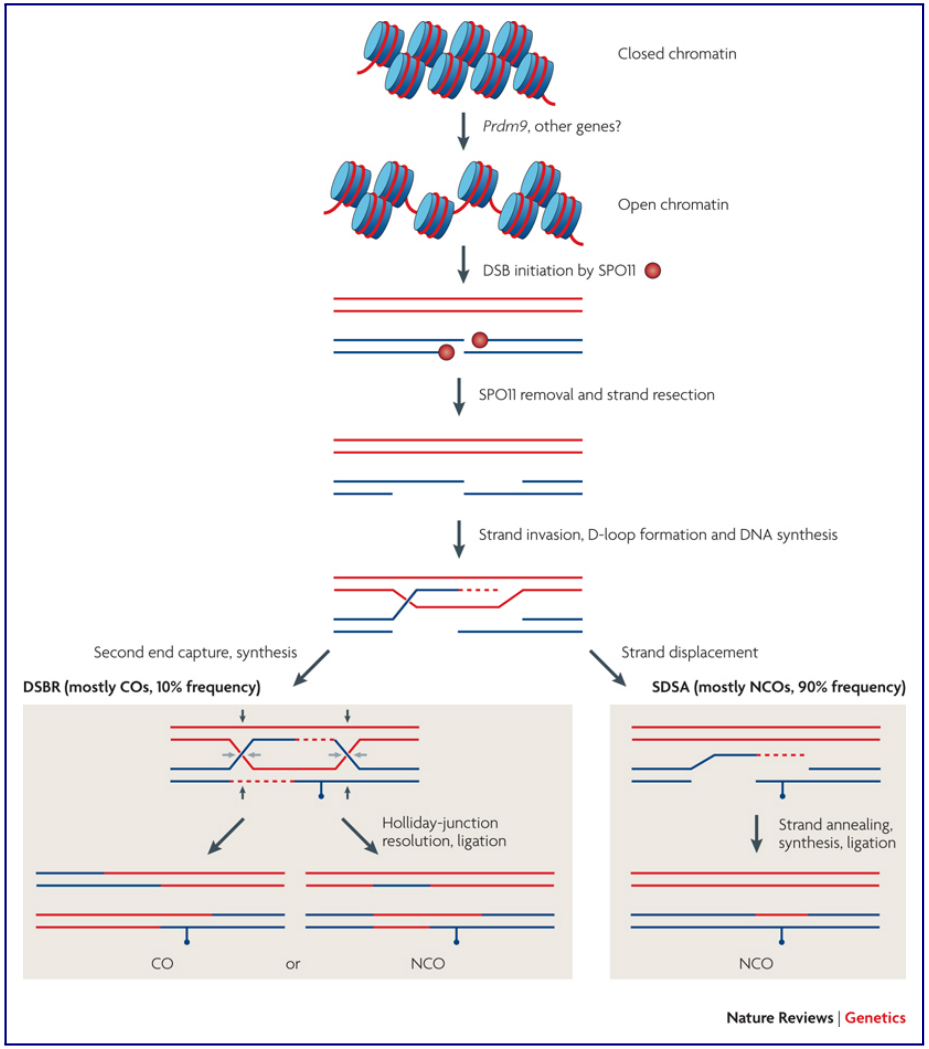
Meiose et reproduction sexuée

- L'information génétique transmise par les gamètes (haploides)
- L'état diploide est restauré au moment de la fécondation par la fusion des gamètes
- Parents homozygotes produiront un seul type de gamètes,
- Parents hétérozygotes produiront deux types de gamètes



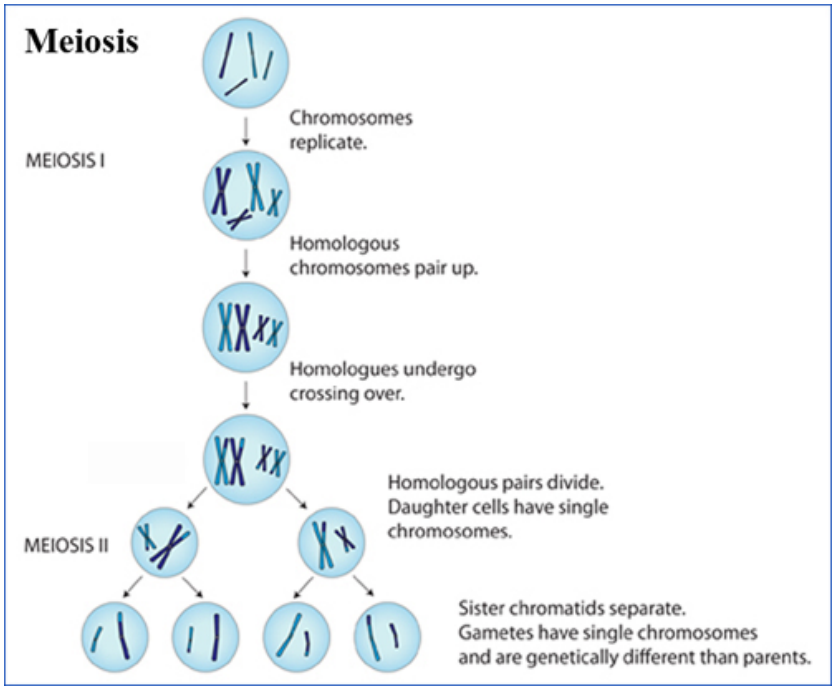
Meiose et recombinaison

- Les deux chromosomes homologues sont côte à côte (alignement des loci)
- Chaque chromatide est dupliquée (4 chromatides forment une tétrade)
- Attachement entre chromatides (plutôt loin du centromère)
- Crossing over permet d'échanger des morceaux de chromatides



Recombinaison et liaison

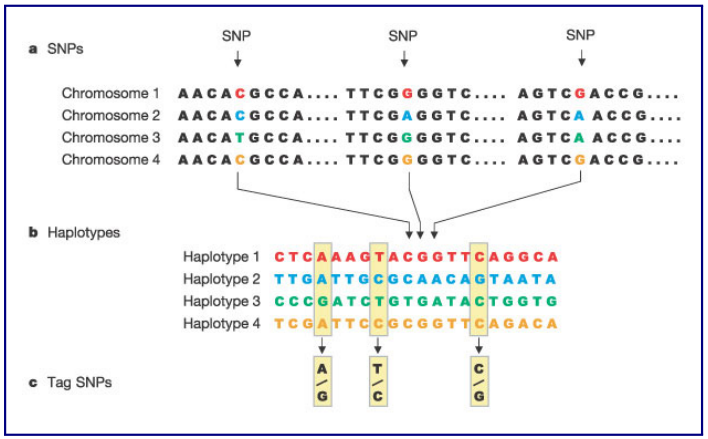
- Le crossing over modifie les combinaisons d'allèles présentes sur chaque chromatide
- On obtient des gamètes recombinants ou non recombinants
- La transmission des loci ne se fait pas indépendamment les uns des autres



Certains SNPs peuvent être transmis conjointement avec d'autres marqueurs, et on dit dans ce cas qu'ils sont liés.

Déséquilibre de liaison (LD)

- Le LD existe lorsque la probabilité d'observer un couple d'allèles sur un chromosome n'est pas égale au produit des probabilités
- Le taux de recombinaison étant hétérogène le long du génome, il existe des blocs de génome qui sont en désequilibre
- Il existe des blocs entiers qui sont transmis de générations en générations



L'essentiel de l'information concernant le motif de variation génétique au sein d'un bloc peut se résumer par un sous ensemble de loci

Le Fused Lasso

- Cette stratégie consiste à structurer la pénalisation, en prenant en compte une information a priori
- Dans le cas des SNPs, on peut prendre en compte l'ordre le long du génome
- Le cadre général est toujours la régression pénalisée:

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{Argmax}} \left(\log \mathcal{L}(\mathbf{Y}; \boldsymbol{\beta}) - \text{pen}(\boldsymbol{\beta}) \right)$$

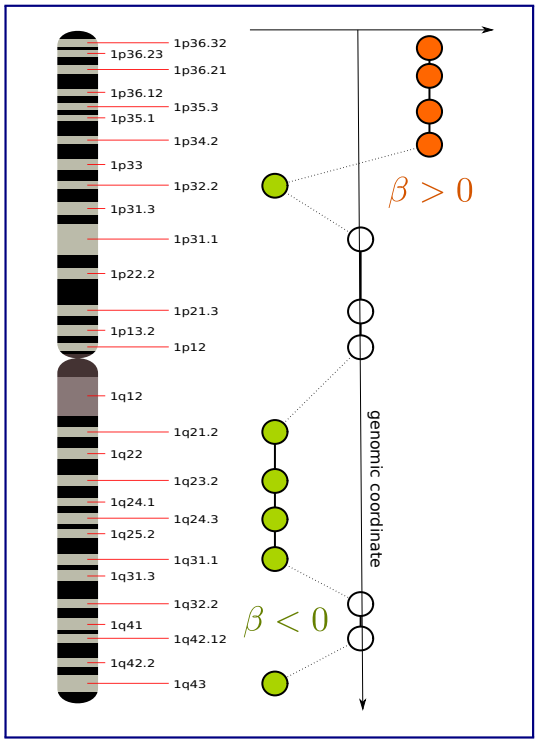
- La pénalité contient deux termes :

$$\text{pen}(\boldsymbol{\beta}) = \lambda_1 \sum_{j=1}^{p} |\beta_j| + \lambda_2 \sum_{j=2}^{p} |\beta_{j-1} - \beta_j|$$

Fused Lasso et prédiction

- $Y_i \in \{\text{cancer}, \text{no-cancer}\}$ avec la probabilité conditionnelle $\pi(\mathbf{x}_i)$
- $\mathbf{x}_i$ le génotype du patient i
- $(\beta_1^*, ...\beta_p^*)$ sont les log odd-ratios d'être malade

$$\log\left(\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right) = \beta_0^* + \sum_{j=1}^{p} \beta_j^* x_{ij}$$

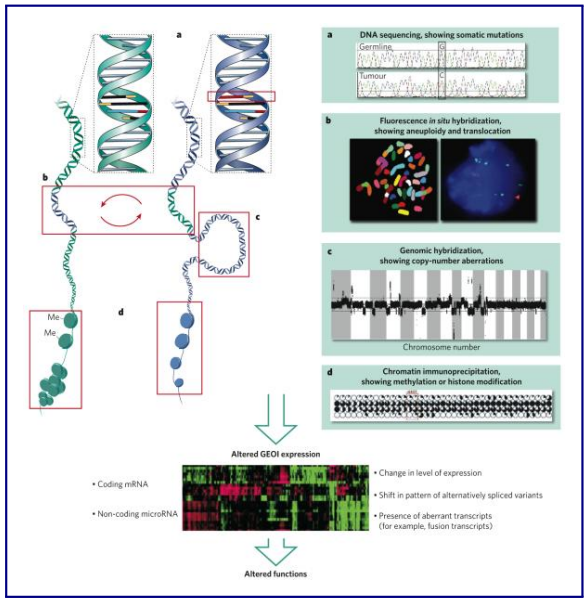


Des loci adjacents sont supposés partager des effets similaires sur la probabilité d'être malade

Outline

Une vision intégrée des phénomènes moléculaires

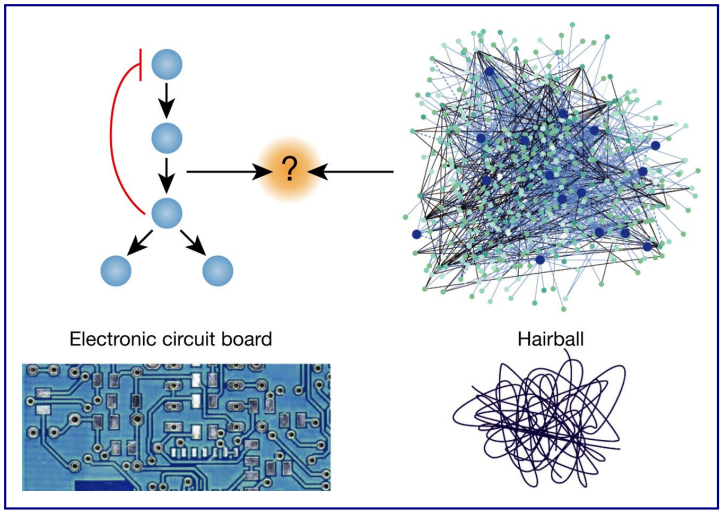
- Les phénomènes moléculaires sont envisagés dans leur ensemble
- La vision multivariée est désormais l'approche privilégiée
- La tâche est désormais d'intégrer différents types de données



Après avoir étudié séparemment plusieurs phénomènes moléculaires, on cherche désormais à les combiner

Les réseaux et la biologie des systèmes

- Suite à l'émergence des réseaux sociaux et l'essor d'internet
- Les données sont constituées d'agents dont les interactions permettent le fonctionnement d'un système
- Les technologies à haut débit ont permis de collecter des ensembles de données sur le génome, transcriptome, régulome, métabolome ...

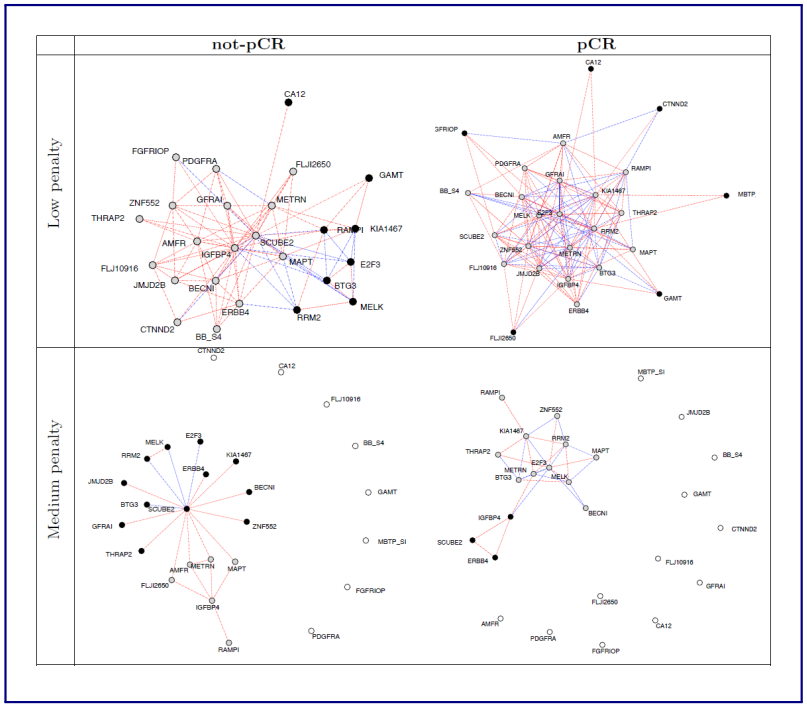


L'avènement de la science de la complexitude !

Les réseaux de régulation

- $\mathbf{X} = (X^1, \ldots X^n)$, n réplicats des expressions de p gènes
- On peut commencer par inférer le réseau de co-expression
- Modèle gaussien graphique $\mathbf{X} \sim \mathcal{N}(0, \mathbf{\Sigma})$
- On s'intéresse aux corrélations partielles:

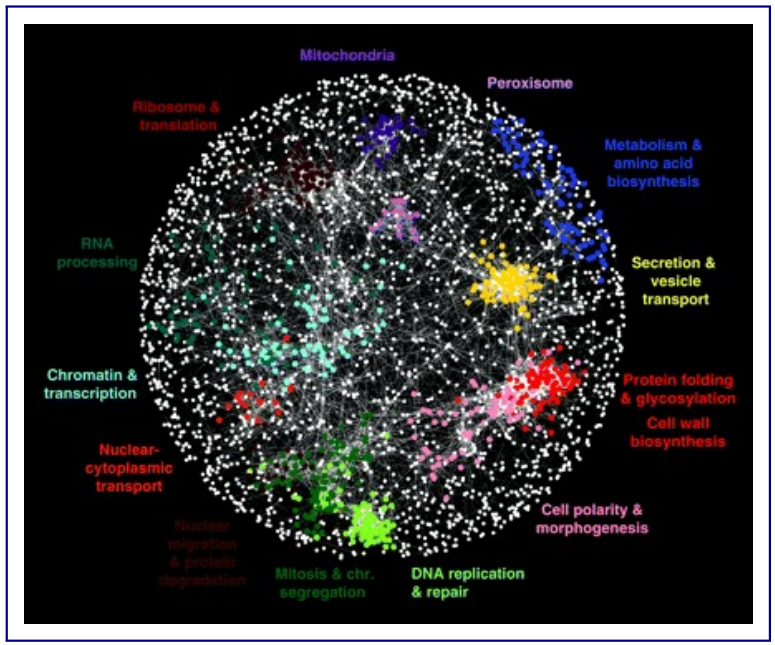
$$X_i \perp X_j | \left\{ \mathbf{X}_{-\{i,j\}} \right\} \Leftrightarrow \Sigma^{-1}_{ij} = 0$$



Un terme de pénalité permet d'obtenir des réseaux parcimonieux

Les réseaux comme information pour sélectionner des gènes discriminants

- On connait désormais certains des liens entre l'expression de différents gènes
- On peut utiliser un graphe $G = (V, E)$ qui décrit les connections entre gènes
- G peut être utilisé comme information a priori pour la sélection



$$\text{pen}(\boldsymbol{\beta}) = \lambda_1 \sum_{j \in V} |\beta_j| + \lambda_2 \sum_{(j,k) \in E} |\beta_j - \beta_k|$$

Sparse CCA pour l'intégration de données

- Considérons $\mathbf{X}_1 \in \mathbb{R}^{n \times p_1}$ la matrice des expressions de p_1 gènes, et $\mathbf{X}_2 \in \mathbb{R}^{n \times p_2}$ la matrice des nombre de copies des gènes
- L'objectif est de rechercher des combinaisons linéaires de $\mathbf{X}_1$ et $\mathbf{X}_2$ qui sont corrélées entre elles:

$$\max_{\mathbf{w}_1, \mathbf{w}_2} \left\{ \mathbf{w}_1^T \mathbf{X}_1^T \mathbf{X}_2 \mathbf{w}_2 \right\}, \text{ avec } \|\mathbf{X}_1 \mathbf{w}_1\|^2 = \|\mathbf{X}_2 \mathbf{w}_2\|^2 = 1$$

- Dans ce cas on peut introduire des contraintes supplémentaires:

$$\begin{aligned} \text{pen}(\mathbf{w}_1) &= \lambda_1^{(1)} \sum_{j=1}^{p_1} |w_{1,j}| \\ \text{pen}(\mathbf{w}_2) &= \lambda_1^{(2)} \sum_{j=1}^{p_2} |w_{2,j}| + \lambda_2^{(2)} \sum_{j=1}^{p_2} |w_{2,j} - w_{2,j-1}| \end{aligned}$$

Outline

Un besoin croissant en méthodologie

- Le domaine de la génomique a connu une explosion de la masse de données générées en 15 ans
- Les phénomènes moléculaires sont désormais envisagés genome-wide à l'échelle des populations
- La modélisation de la variabilité inter-individuelle devient centrale
- Les données génomiques ont permis de décrire plus en détail la complexité des phénomènes moléculaires (médecine personalisée ?)
- Point de vue statistique: méthodes pénalisées pour l'analyse des données génomiques
- Enjeu de la grande dimension qui dépasse la génomique (image, réseaux, physique, ...)