Les principaux problèmes d'analyse et les plateformes régionales

C. Gautier UMR 5558, Bamboo, PRABI

Les NGS: 40 ans d'histoire

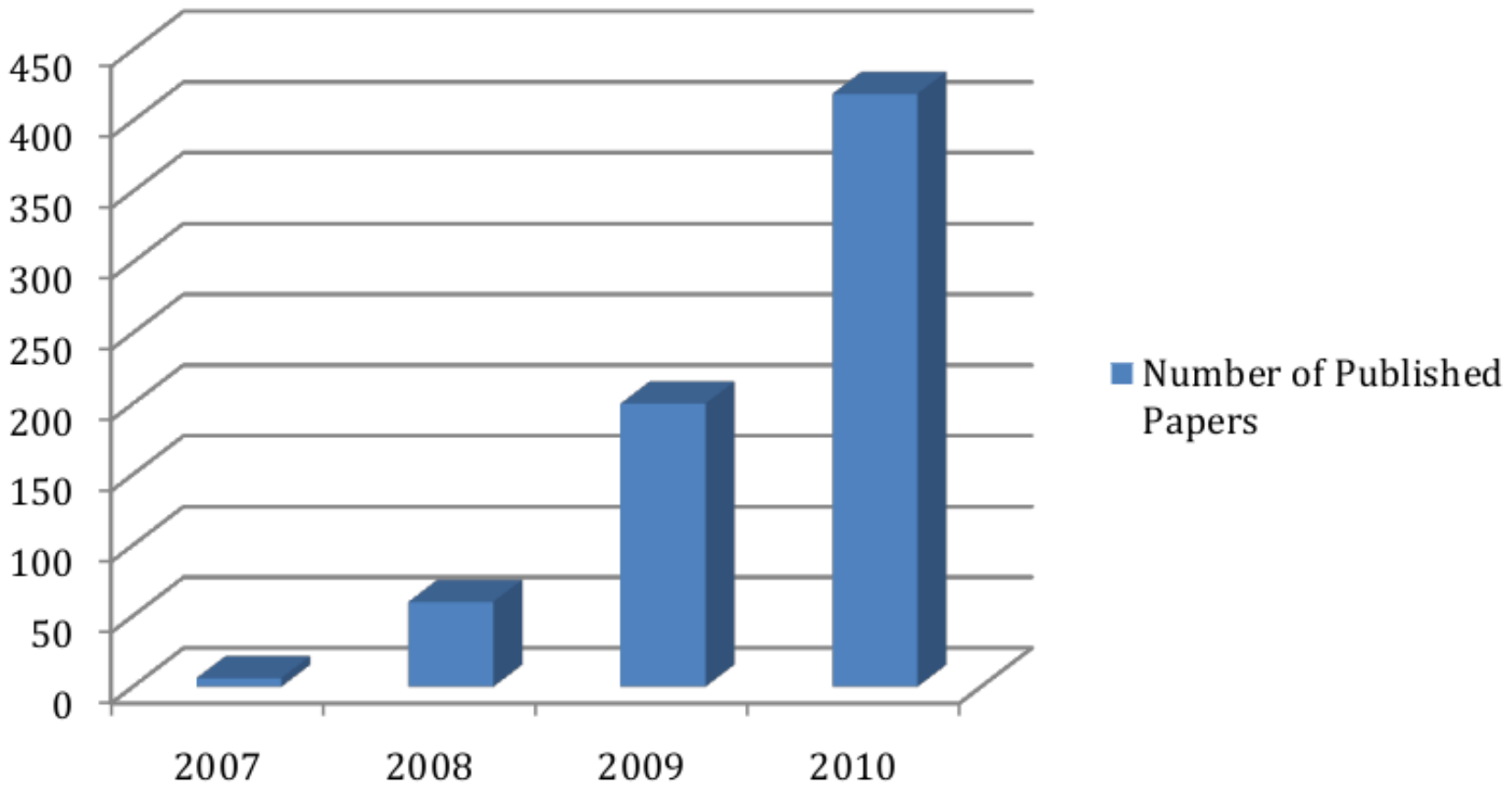
- 1972 séquençage du premier génome complet
- Des sauts qualitatifs importants
- Vers les NNNNNNGS
- Des avancées conceptuelles essentielles en biologie

Accompagnement par Statistiques et Informatique

- Une transition importante avec l'arrivée des puces
  - Avant : bases de données, représentation des données, assemblage
  - Après : explosion des travaux de statistiques
- Nécessité de traitements lourds de mise en forme des données
- Explosion de la bibliographie

Genes (2010),**1**,317-334

Figure 1. Number of publications by year deposited in PubMed on "Next generation sequencing" (Year 2010 figure is projected).



Un grand nombre d'objectifs possibles

- Les gènomes, leur organisation linéaire et spatiale
- Les gènes, leurs structures, leurs expressions
- Épigénétique
- Biodiversité, réponse à la sélection
- Trait d'histoire de vie (régime alimentaire)

Exemple de quelques concepts

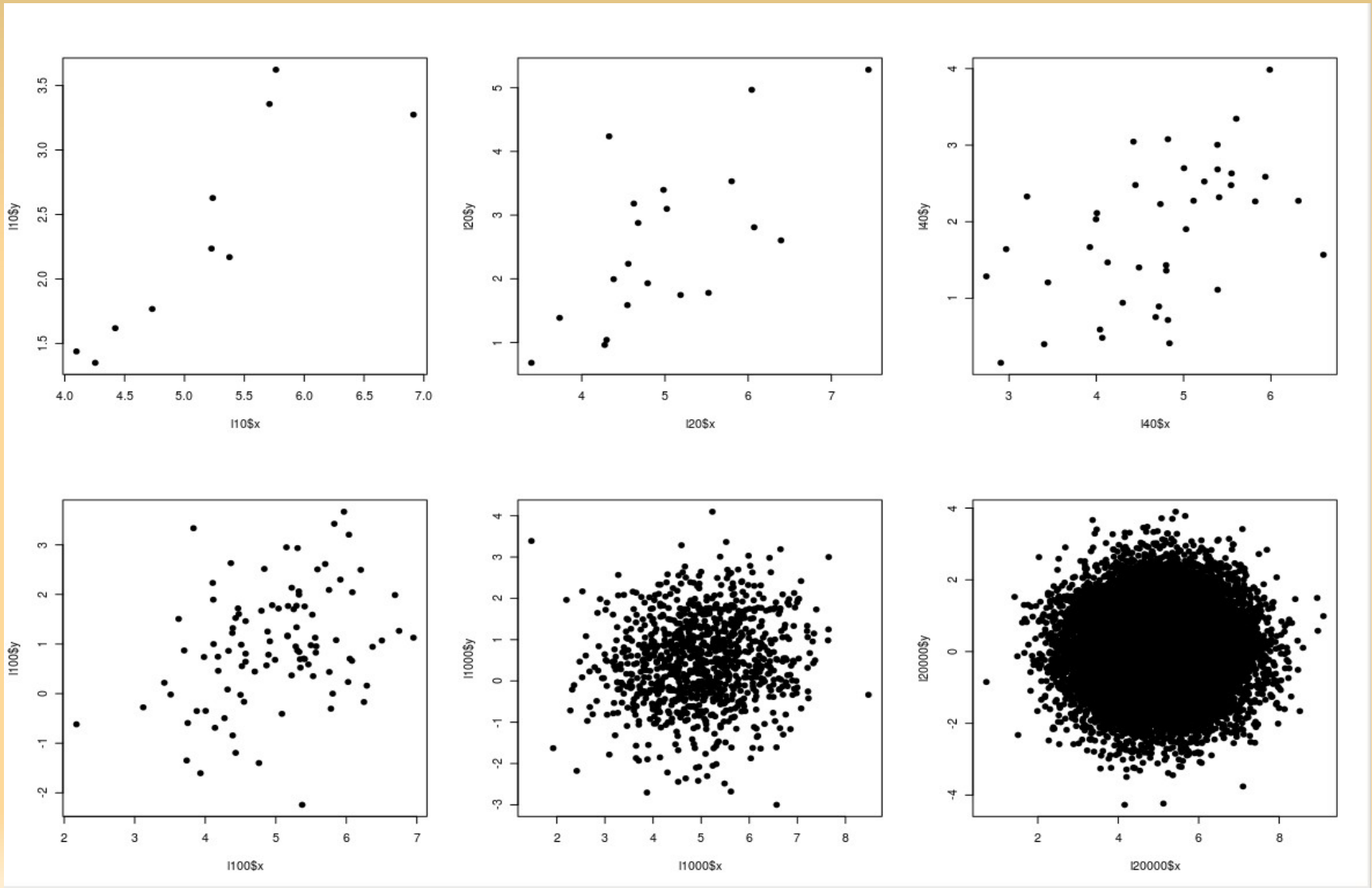
- Un nombre considérable d'outils logiciels
  - 24 assembleurs
  - 12 logiciels pour CHIPseq
  - ...
- Peu de grands concepts
- Nécessité de beaucoup d'euristiques

P-value

- Pour ne pas dire niveau critique
- Test :
  - Choix d'un risque de 1^{ere} espèce
  - Définition de la règle de décision
  - Expérience → résultat
  - Décision
- Pour la p-value on travaille à l'envers :
  - À partir du résultat on définit le risque le plus faible qui aurait entraîné le rejet
- La p-value mesure la confiance que l'on a dans l'existence d'une structure

Effet et taille d'effet

Tous les nuages correspondent à la même p-value (0.001) pour des nombres de points croissants (10, 20, 40, 100, 1000, 20000). Les corrélations sont respectivement de : 0.87 ; 0.68 ; 0.50 ; 0.33 ; 0.10 ; 0.02



Un exemple : contenu en G+C et expression des gènes

Sémon M, Mouchiroud D, Duret L (2005)
Relationship between gene expression and GC-content in mammals: statistical significance and biological relevance,Human Molecular Genetics, vol. 14 pp.421-427

Un exemple parmi les multiples résultats du papier :

Corrélation entre le C+G des introns et l'expression mesurée par EST chez la souris :
24 127 gènes ; p-value 10^{-16} ; corrélation = 0.09

De plus les auteurs montrent une forte dépendance entre les méthodes de mesures et les résultats.

Code R

```
pcor <- function(n,k,pvalmin=0.0009,pvalmax=0.0011) {
        for(i in 1:100) {
                x=rnorm(n,mean=5,sd=1)
                y=k*x+rnorm(n,0,1)
                t=cor.test(x,y)
                if(t$p.value < pvalmax & t$p.value > pvalmin) {
                        print(paste("cor = ",cor(x,y)," p-value = ",(cor.test(x,y))$p.value))
                        return(list(x=x,y=y))
                        }

                }
        print("échec  ",t$p.value)
        }

l10=pcor(10,0.5)
plot(l10$x,l10$y)
l20=pcor(20,0.5)
l40=pcor(40,0.4)
l100=pcor(100,0.2)
l1000=pcor(1000,0.1)
l20000=pcor(20000,0.01)

par(mfrow=c(2,3))
plot(l10$x,l10$y,pch=21,bg="black")
plot(l20$x,l20$y,pch=21,bg="black")
plot(l40$x,l40$y,pch=21,bg="black")
plot(l100$x,l100$y,pch=21,bg="black")
plot(l1000$x,l1000$y,pch=21,bg="black")
plot(l20000$x,l20000$y,pch=21,bg="black")
```

Parallélisation des expériences

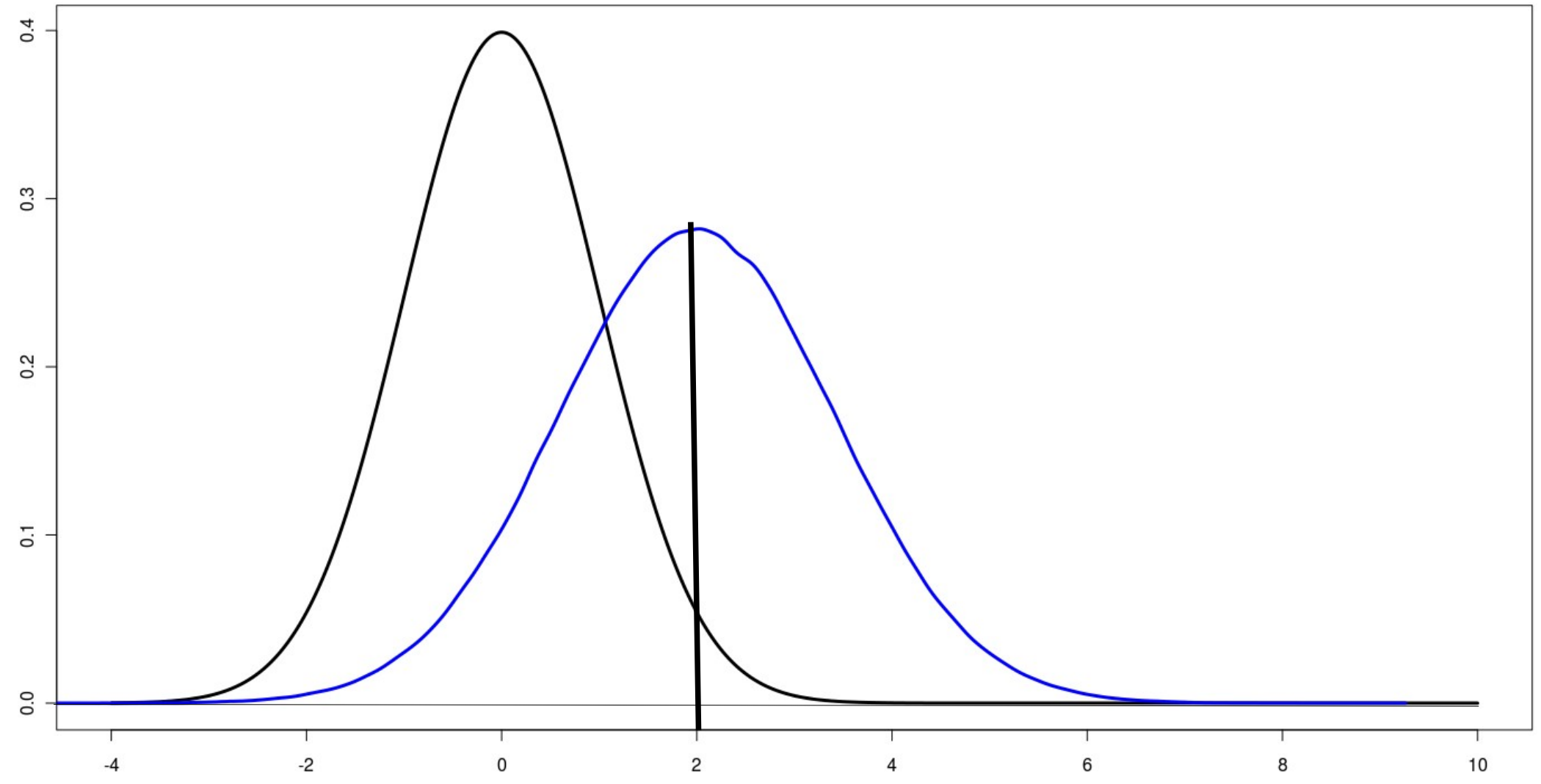
- Puces
- RNAseq
- CHIPseq

FDR (1)

- Les résultats correspondent à un mélange :
  - Simple effet des variabilités expérimentale et biologique
  - Réel effet : par exemple gènes différentiellement exprimés
- Le premier cas est beaucoup plus fréquent que le deuxième :
  - 20 000 gènes dont 100 sont différentiellement exprimés
  - Ou encore quelques centaines activés par le promoteur étudié

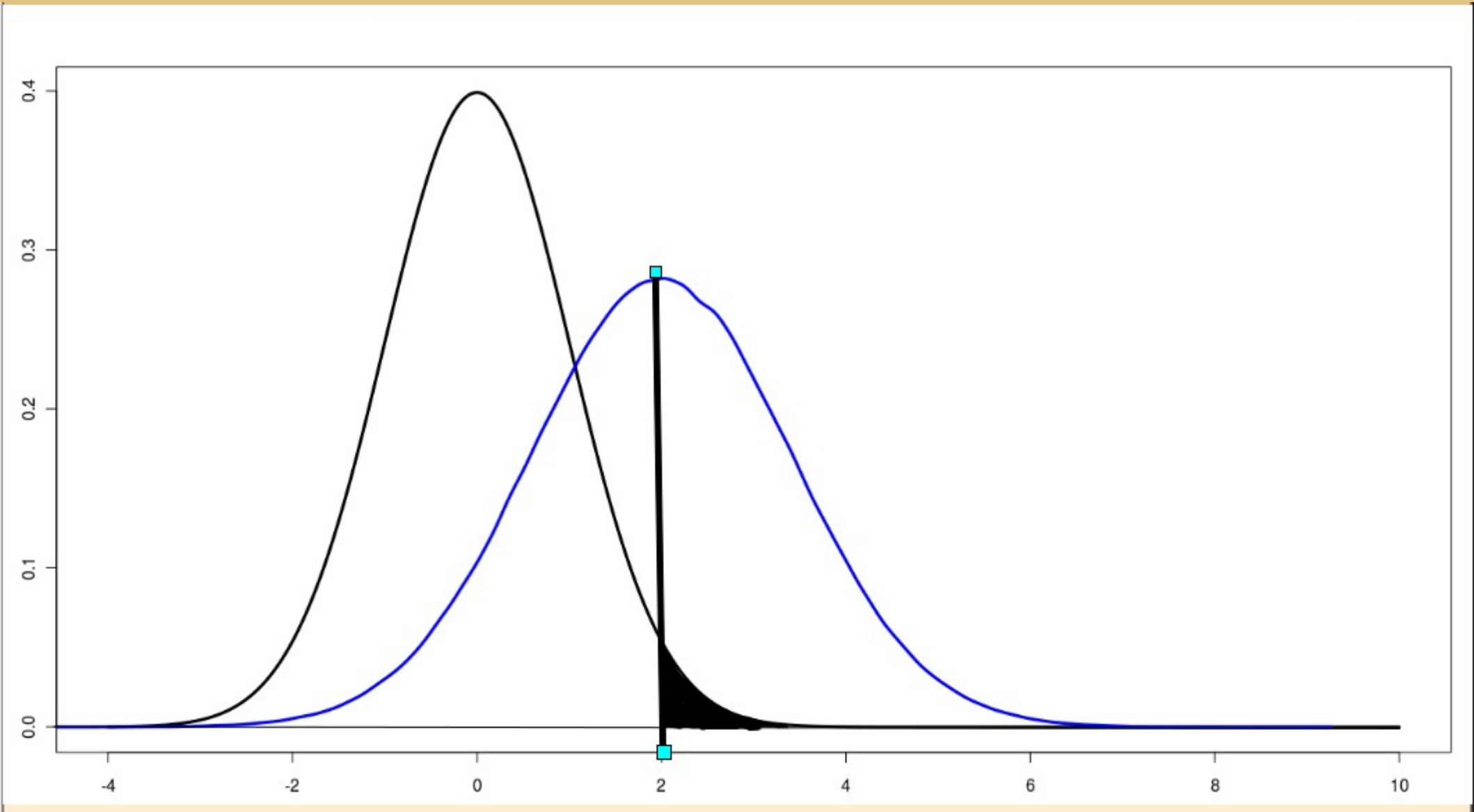
FDR

En noir la distrubition des différences d'expression observées pour des gènes ”sans différence” (H0) et en bleu celle pour les gènes ”différents” entre les deux confditions (H1).



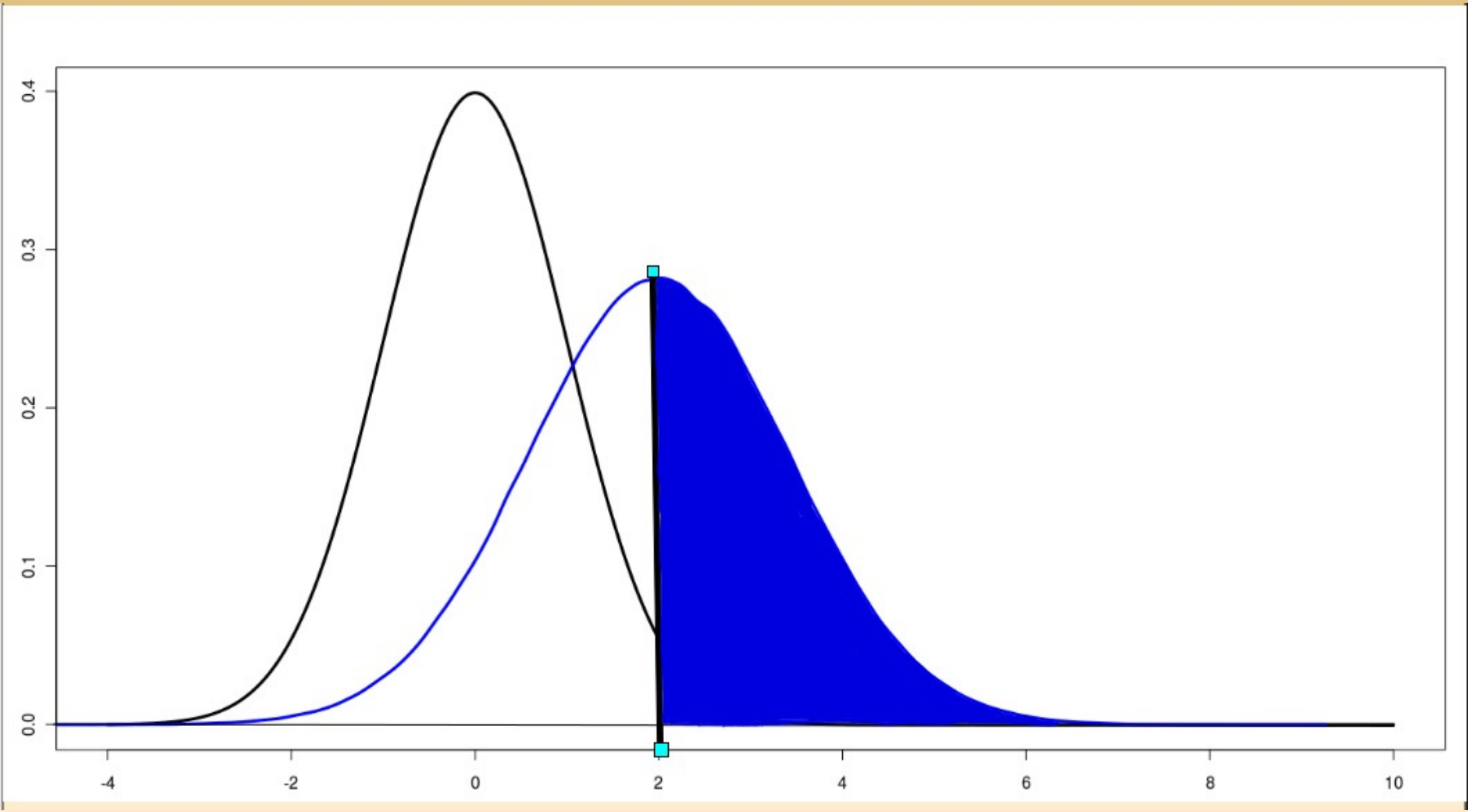
FDR

La probabilité sous H0 de dépasser 2 est de 0.025



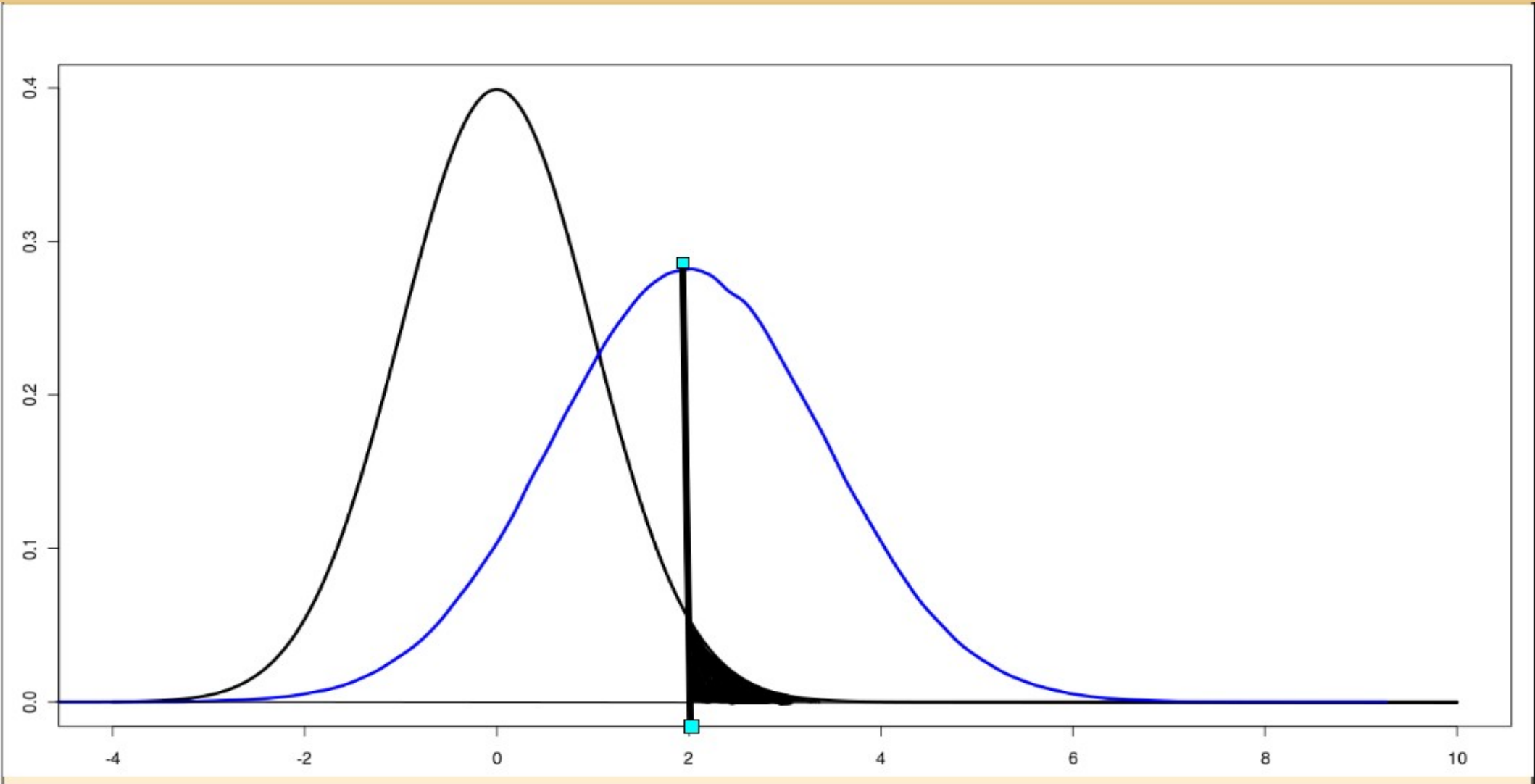
FDR

Dans le cas des gène différentiellement exprimé la probabilité de dépasser 2 est (ici) de 0.498



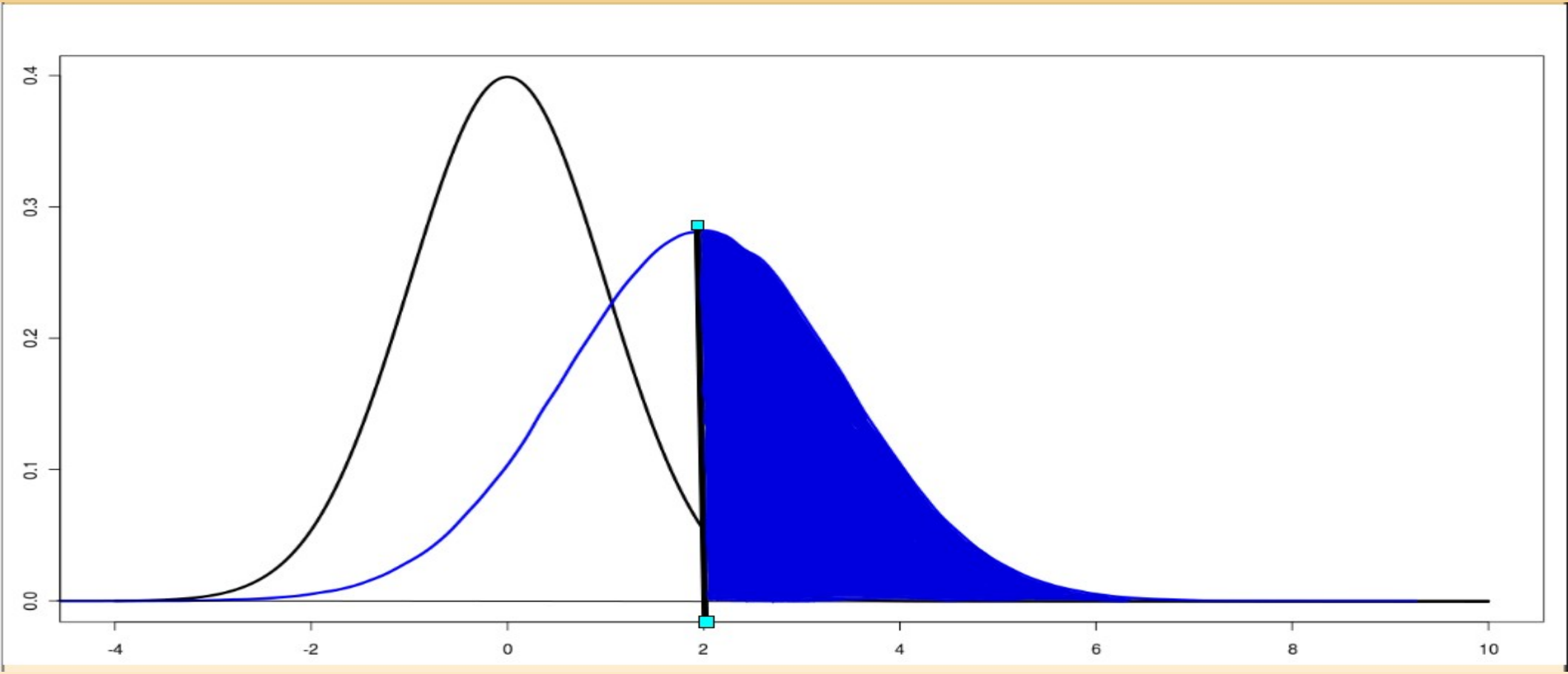
FDR

La probabilité sous H0 de dépasser 2 est de 0.025. Il y a 20 000 gènes sous H0 donc :
20 000*0.05 = 1000 gènes sont significatifs par erreur



FDR

La probabilité sous H0 de dépasser 2 est de 0.025. Il y a 20 000 gènes sous H0 donc :
20 000 * 0.05 = **1000 gènes sont significatifs par erreur**. Dans le cas des gène différentiellement exprimé la probabilité de dépasser 2 est (ici) de 0.498, si 200 gènes sont différentiellement exprimés on en détectera :
200 * 0.498 = 99 . **La proportion des gènes non différentiellement exprimés dans la liste des gènes retenus sera donc de 1000/1099 soit 90%** : c'est le FDR (False Discovery Rate)



Code R

```
l <- list(n1=25000,n2=200,mean=2,sd1=1,sd0=1)
# distribution des H1
effet <- function(n=1000000) {
        y=numeric(n)
        for(i in 1:n) {
                m=rnorm(1,l$mean,l$sd1)
                y[i]=rnorm(1,m,1)
                }
        return(y)
        }



# le FDR pour le seuil a
seuil <- function(a) {
        pval1 <- (1-pnorm(a,0,l$sd0))
        pval2 <- sum(y>a)/1000000
        n1=l$n1*pval1
        n2=l$n2*pval2
        return(list(pval1=pval1,pval2=pval2,n1=n1,n2=n2,FDR=n1/(n1+n2)))
        }


ecrire <- function(s) {
        ligne=paste("p-val =",s$pval1," ; nb/H0 =",s$n1," ; nb/H1 =",s$n2, "; FDR=",s$FDR)
        print(ligne)
        }


# exemple d'exécution
l <- list(n1=25000,n2=200,mean=2,sd1=1,sd0=1)
v=-400:1000/100
x=dnorm(v,0,l$sd0)
plot(v,x,type="l",lwd=3)
lines(density(y),lwd=3,col="blue")
```

Comment calculer le FDR

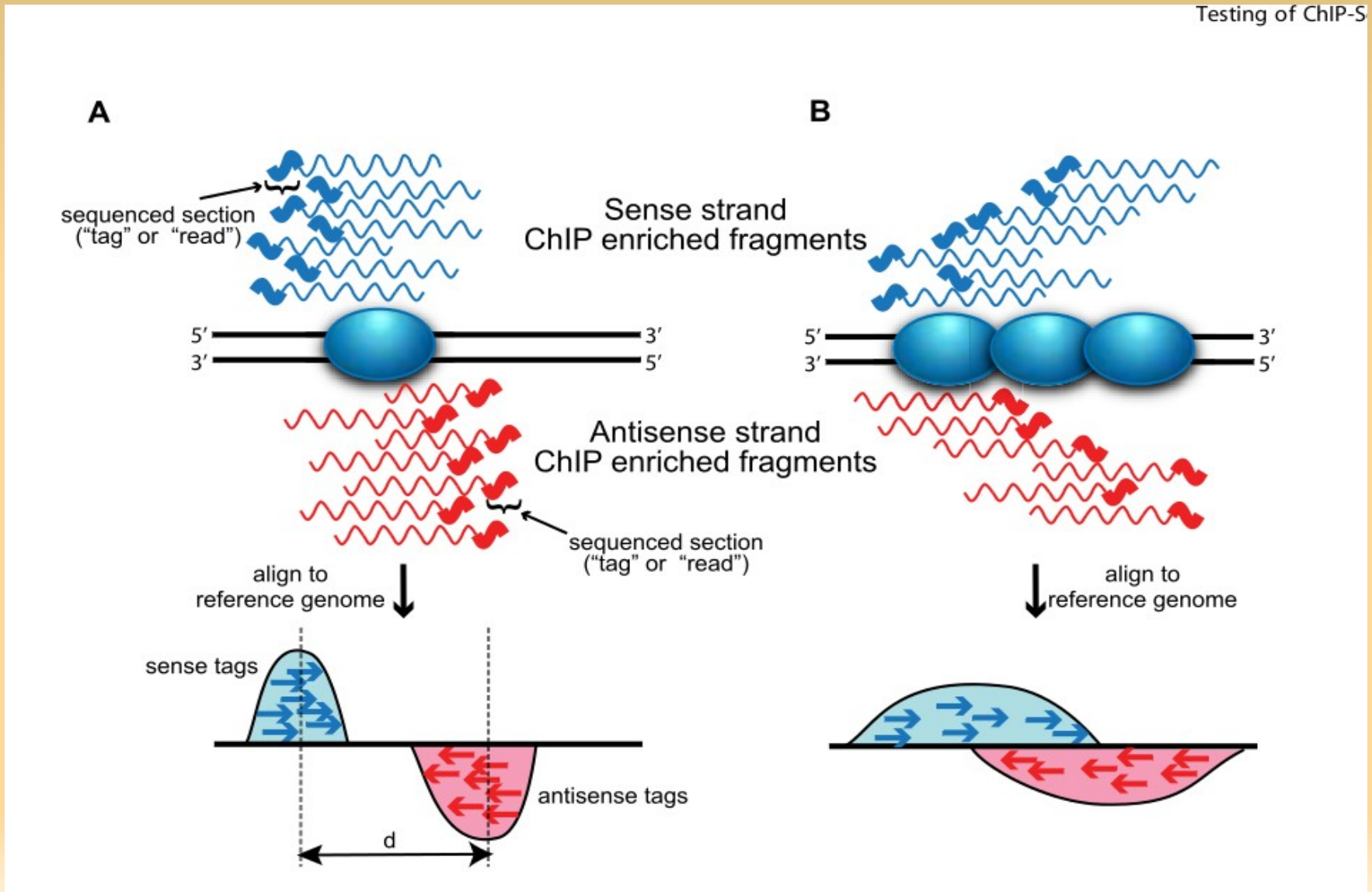
- Benjamini, Hochberg est la plus fréquemment utilisée pour les puces d'expression
- Utilisation d'un modèle (Poisson, HMM,...)
- Utiliser une expérience de ”contrôle” supposée sans situations exceptionnelles (Sc/S)

Biais techniques

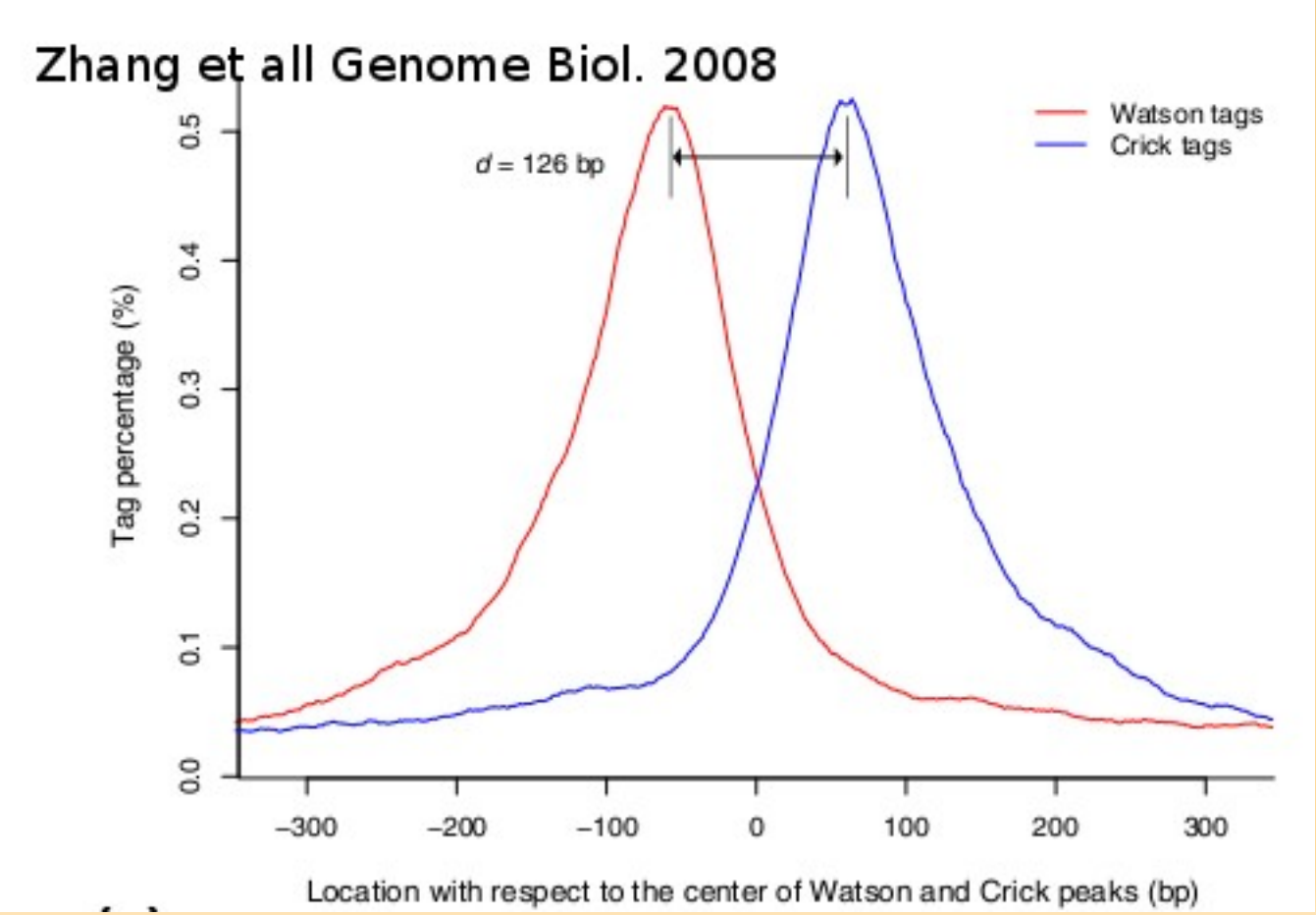
- Présents à toutes les étapes d'un projet NGS
- Quelques exemples :
  - Baisse de qualité le long des fragments
  - Empilement de fragments
  - Directement liés à la pratique expériementale

Biais techniques, un exemple : CHIPseq

Elizabeth G. Wilbanks, Marc T. Facciotti (2010) Plos One **5** (7) e11471



CHIPseq (MACS)



Chipseq : Echantillon de méthodes

Program	Reference	Version	Graphical user interface?	Window-based scan	Tag clustering	Gaussian kernel density estimator	Strand-specific scoring	Peak height or fold enrichment (FE)	Background subtraction	Compensates for genomic duplications or deletions	False Discovery Rate	Compare to normalized control data (FE)	Compare to statistical model fitted with control data	Statistical model or test
CisGenome	28	1.1	X*	X				X	X		X		X	conditional binomial model
Minimal ChipSeq Peak Finder	16	2.0.1			X			X				X		
E-RANGE	27	3.1			X			X				X	X	chromsome scale Poisson dist.
MACS	13	1.3.5		X				X			X		X	local Poisson dist.
QuEST	14	2.3				X		X			X**		X	chromsome scale Poisson dist.
HPeak	29	1.1		X				X					X	Hidden Markov Model
Sole-Search	23	1	X	X				X		X			X	One sample t-test
PeakSeq	21	1.01			X			X					X	conditional binomial model
SISSRS	32	1.4		X			X					X		
spp package (wtd & mtc)	31	1.7		X			X		X	X'	X			
				Generating density profiles			Peak assignment		Adjustments w. control data		Significance relative to control data			

X* = Windows-only GUI or cross-platform command line interface

X** = optional if sufficient data is available to split control data

X' = method exludes putative duplicated regions, no treatment of deletions

Elizabeth G. Wilbanks, Marc T. Facciotti (2010) Plos One **5** (7) e11471

Biais techniques : Conclusion

- Très divers
- Dépendant de la méthode expérimentale et pas de l'objectif biologique
- Une cause majeure de l'explosion du nombre d'outils logiciel disponible

Masse des données

- Ressource informatique de traitement
  - Calcul, contrainte sur le type de machines
  - Stockage
- Archivage
- Des solutions algorithmique adaptées : un exemple graphes de de Bruijn (Pevzner,Tang, Waterman (2001) PNAS 98(17) 9748-9753)

Structure de données et assemblage

PhD de Daniel Robert Zerbino, Darwin College, 2009 (l'auteur de Velvet)

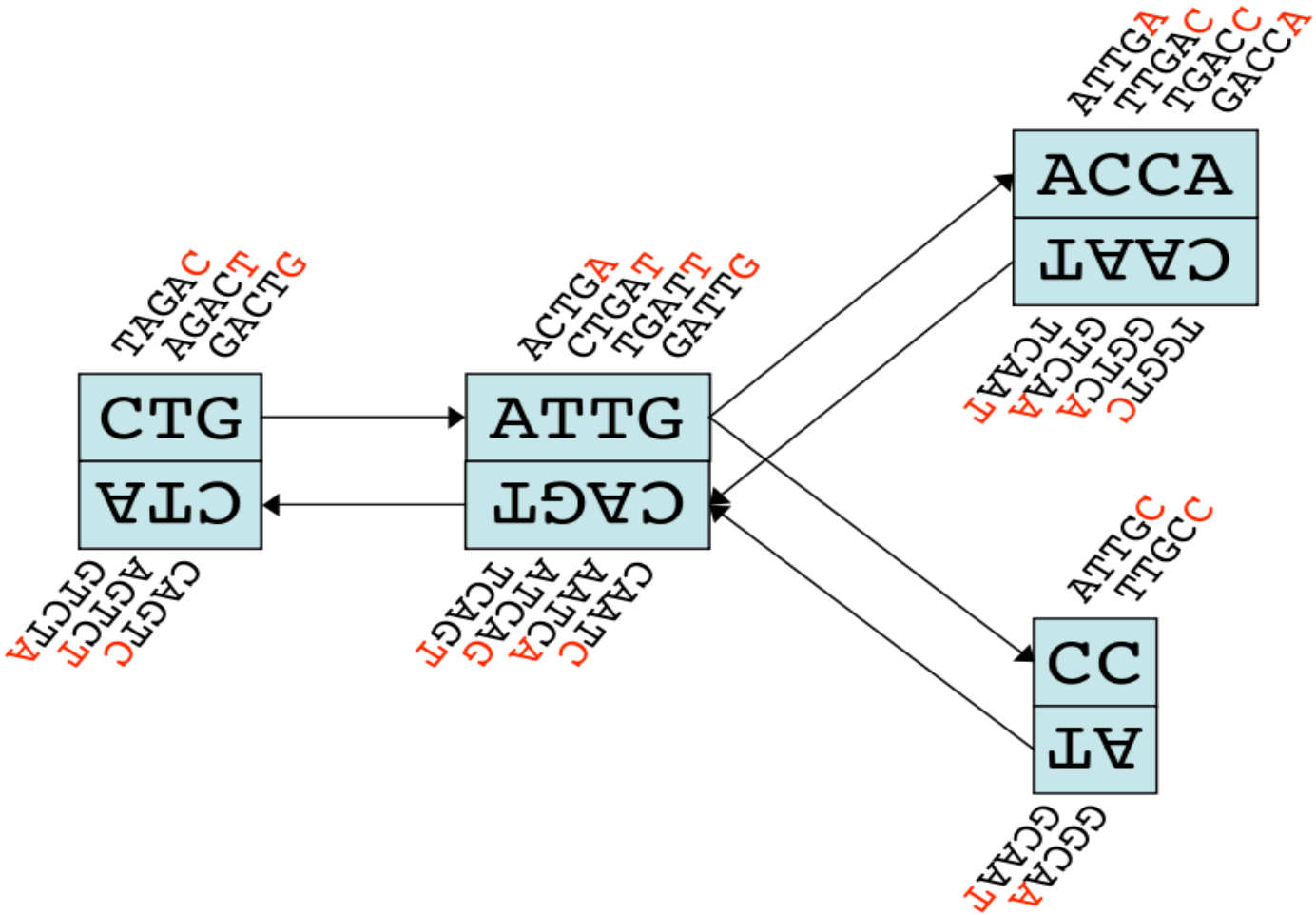


Figure 2.1: Schematic diagram of the de Bruijn graph implementation

Conclusions

- Autres problémes classiques
  - Le fléau de la dimension
  - Biodiversité (identification, taxonomie, polymorphisme, ….)
- Grande evolutivité de la discipline
- Vers l'intégration de données hétérogènes
- Et la biologie systémique ?

Le Pôle Rhône-Alpes de Bioinformatique

www.cnrs.fr

http://www.prabi.fr

Présentation

Origine remontant à 1998 avec la mise en place du Pôle Bioinformatique Lyonnais (PBIL) :

Une composante « Doua » (BBE) et une composante « Gerland » (IBCP).

Plate-forme labellisée IBiSA (Infrastructures Biologie-Santé et Agronomie).

Membre du Réseau National des plates-formes en Bioinformatique (ReNaBi) :

Six centres régionaux correspondant à 24 plates-formes :

- Structuration en un grand Institut Français de Bioinformatique (IFB).

Organisation

Découpage en six sites officiels :

PRABI-Doua.

PRABI-Gerland.

PRABI-Lyon Sud.

PRABI-Grenoble.

PRABI-Analyse et Modélisation*.

PRABI-Synergie Lyon Cancer*.

(* Intégration/création en 2010)

Laboratoires associés

Laboratoire	Équipe	Site
LBBE	- Bioinformatique et Génomique Évolutive	Doua
	- Écologie Évolutive des Populations	Doua
	- Eléments Transposables, Évolution, Populations	Doua
+INRIA-RA	- Baobab/Bamboo	Doua
	- Biostatistiques Santé	Lyon Sud
INCa	- Synergie Lyon Cancer	SLC
BF2I	- Génomique Fonctionnelle des Interactions Trophiques	Doua
IBCP	- Bioinformatique : Structures et Interactions	Gerland
	- Biocristallographie	Gerland
LECA	- Génomique des Populations et Biodiversité	Grenoble
INRIA-RA	- Ibis	Grenoble
IRTSV	- Bioinformatique Moléculaire	Grenoble

Douze équipes appartenant à sept laboratoires/instituts

Quelques outils

Nom	Publication la plus récente	Citations
SeaView	Gouy *et al.* (2010) *Mol. Biol. Evol.* **27**:221-224	1221 (2)
ADE-4	Thioulouse *et al.* (1997) *Stat. Comput.* **7**:75-83	834 (1)
WWW-Query	Perrière et Gouy (1996) *Biochimie* **78**:364-369	740 (1)
NPS@	Combet *et al.* (2000) *Trends Biochem. Sci.* **25**:147-150	645 (1)
ProDom	Bru *et al.* (2005) *Nucleic Acids Res.* **33**:D212-215	559 (6)
HOVERGEN *et al.*	Penel *et al.* (2009) *BMC Bioinformatics* **10**(**S6**):S3	291 (4)
ACNUC	Gouy et Delmotte (2008) *Biochimie* **90**:555-562	256 (4)
AntheProt	Deléage *et al.* (2001) *Comp. Biol. Med.* **31**:259-267	219 (5)
Geno3D	Combet *et al.* (2002) *Bioinformatics* **18**:213-214	156 (1)
Oriloc	Frank et Lobry (2000) *Bioinformatics* **16**:560-561	51 (1)
BIBI	Devulder *et al.* (2003) *J. Clin. Microbiol.* **41**:1785:1787	47 (1)
NuReBase	Ruau *et al.* (2004) *Nucleic Acids Res.* **32**:D165-167	42 (2)
Total		**5061**

Aspects formations/services

Activités présentes sur l'ensemble des sites :

Au total, ~12 formations proposées chaque année :

- Biostatistiques, phylogénie, analyse des données « omiques », initiation aux NGS, etc.

Personnel dédié à ces aspects pour la composante PRABI-Analyse et Modélisation (dir. C. Gautier) :

- Clothilde Deschamps (IE CDI Ezus).
- Vincent Navratil (IR UCBL).
- Christine Oger (IR UCBL).