Puces à ADN *(DNA chips)* & Micro-réseaux (*Microarrays*)

1-TECHNOLOGIE DES PUCES

Objectifs
Principe
Platformes technologiques
Entités mesurables
Applications
Limites

3-BIOINFORMATIQUES DU TRANSCRIPTOME

Pré-production
sélection des sondes

Production
robotique, analyse d'image, flux de données

Post-production
interprétations des résultats : expression différentielle, clustering, classification, réseaux géniques

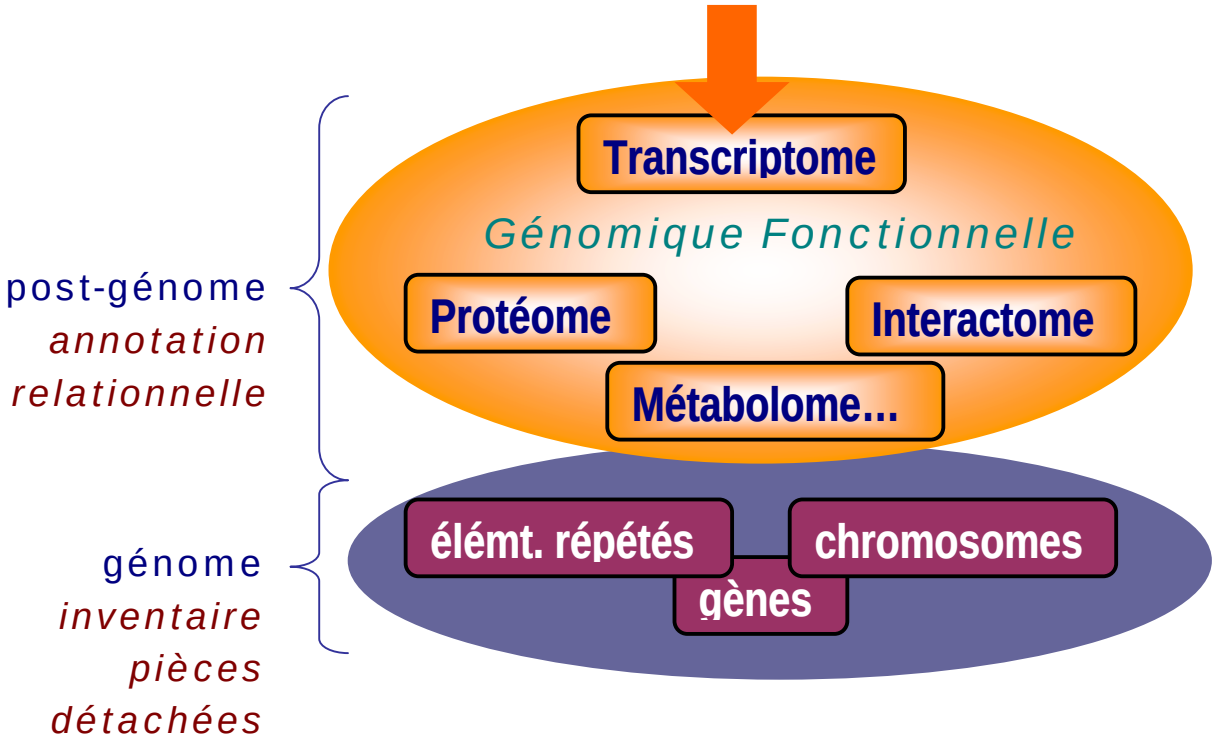
1-TECHNOLOGIE DES PUCES

1.1-Objectifs

Mesures massivement **parallèles** & **quantitatives** de l'**expression** des gènes (transcrits/ARNm) :

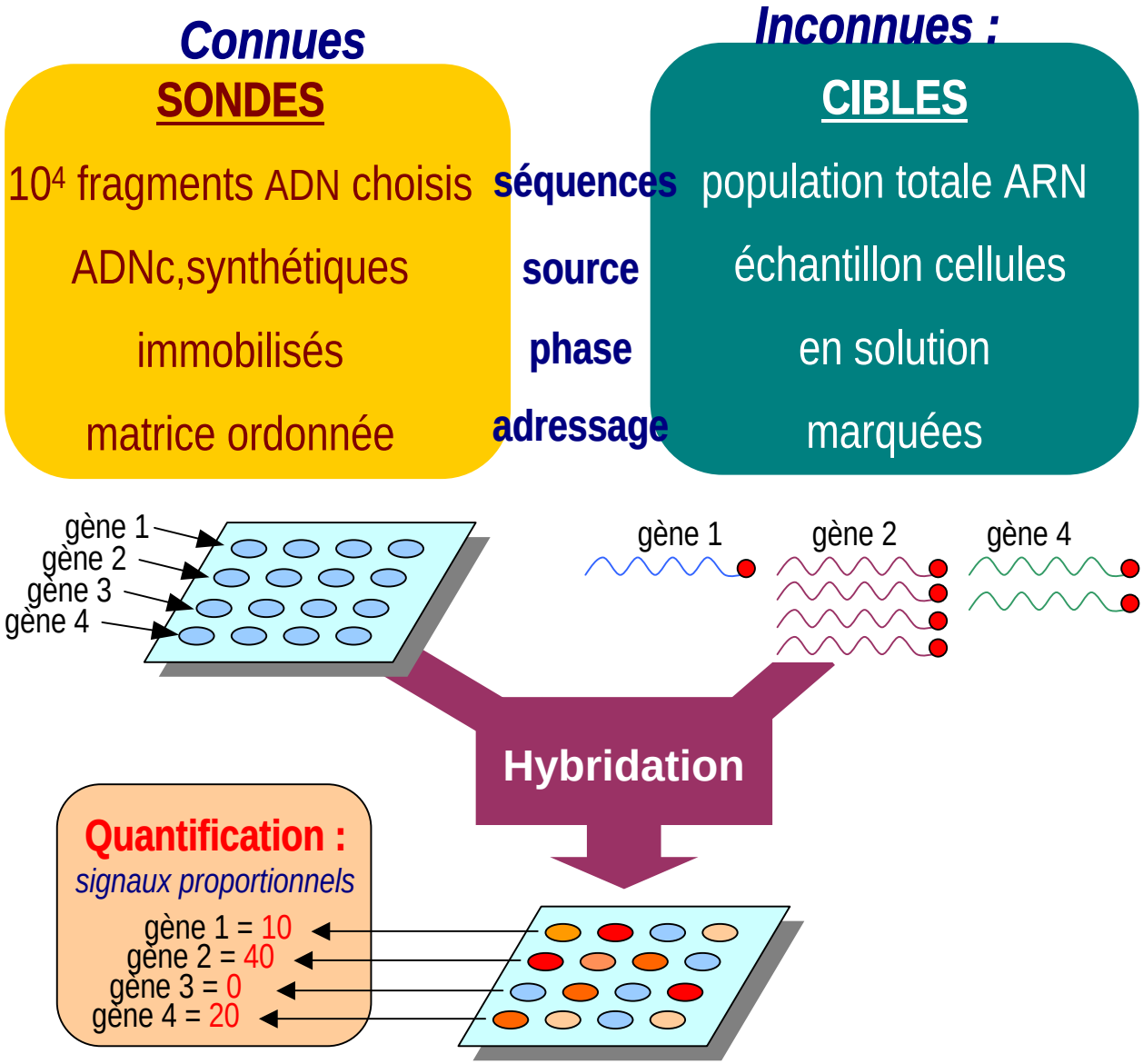
Northerns : niveau d'expression d'**1** gène / tissu

Puces ADN : expressions de **10000** gènes / tissu
voire de **tous les gènes** / organisme



1.2-Principe

Hybridation complémentaire des acides nucléiques:



Expression différentielle : ***test*** vs. ***contrôle***

canaux	marqueurs	hybridations	mesures
mono	unique *(P^{33})*	successives	normalisées
bi	double *(Cy5/Cy3)*	simultanées	ratios

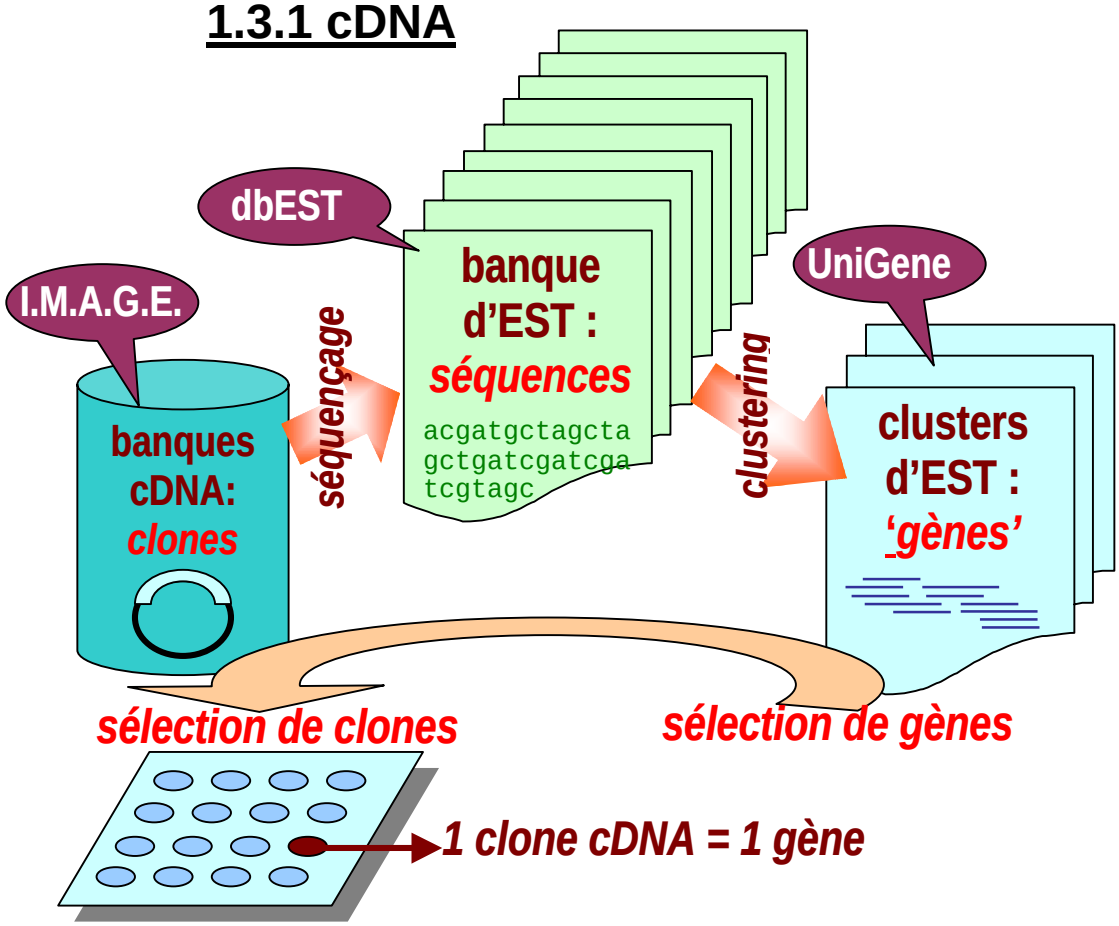
signal = **f(**marq,tps hyb,strg lav,tps exp,tps exp,Q dépt,[**cible**]**)**

niveau d'**expression** = **k** x **signal**

facteur de ***normalisation***

1.3-Plateformes technologiques

1.3.1 cDNA

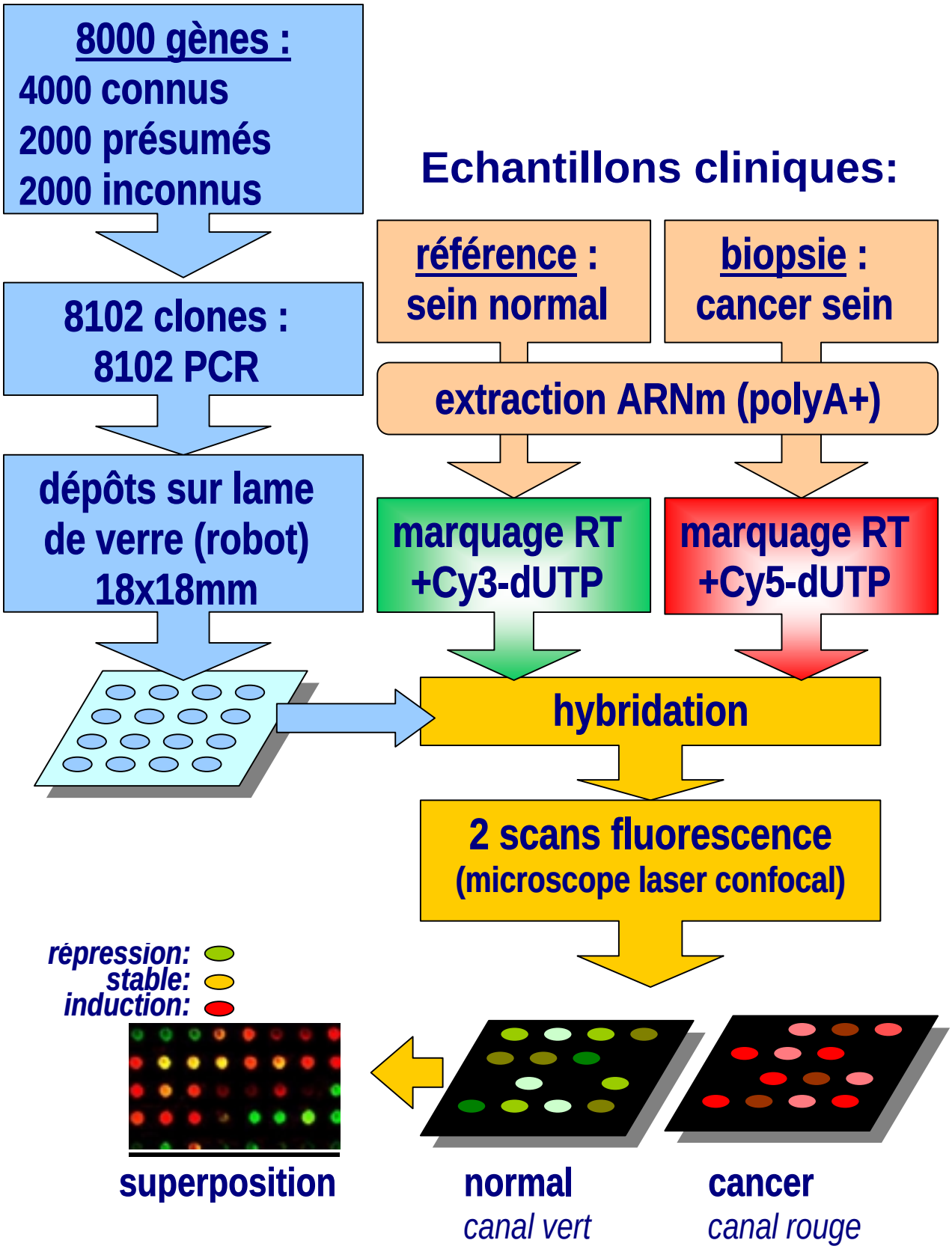


Variantes :

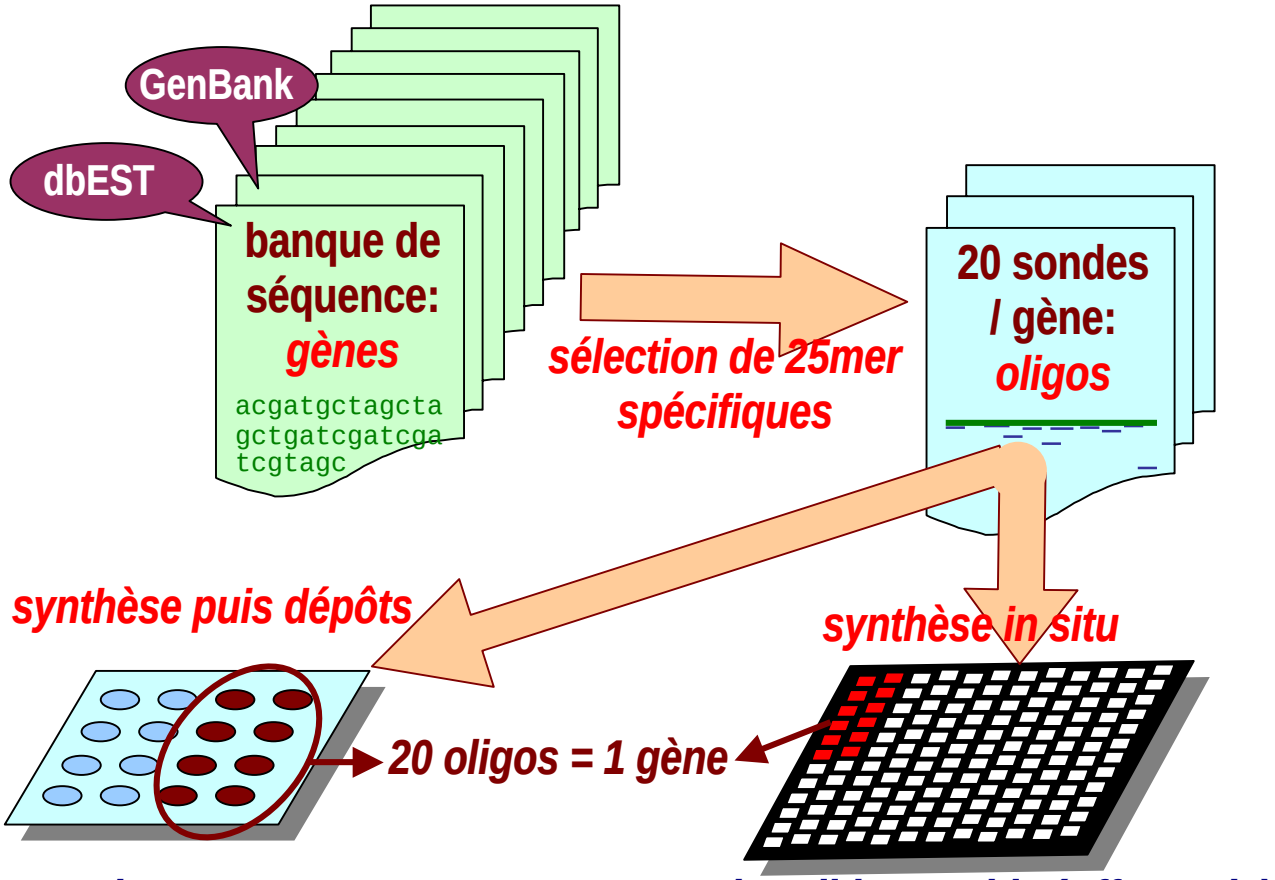
	macro-	micro-réseaux		
supports	'filtres' nylon	lames verre	nylon	nylon
détection	P^{32}	fluorescence	colorimét.	P^{33}
densité élts/cm^2	≈10-100	≈10000	≈5000	≈1000
spots (lxL mm)	3500 (80x120)	6400 (18x18)	9600 (27x18)	8500 (50x20)
publié	1990	1995	1998	1999

Exemple d'expérience:

cDNAs, lame de verre, fluorescence bi canaux



1.3.2 Oligonucléotides de synthèse

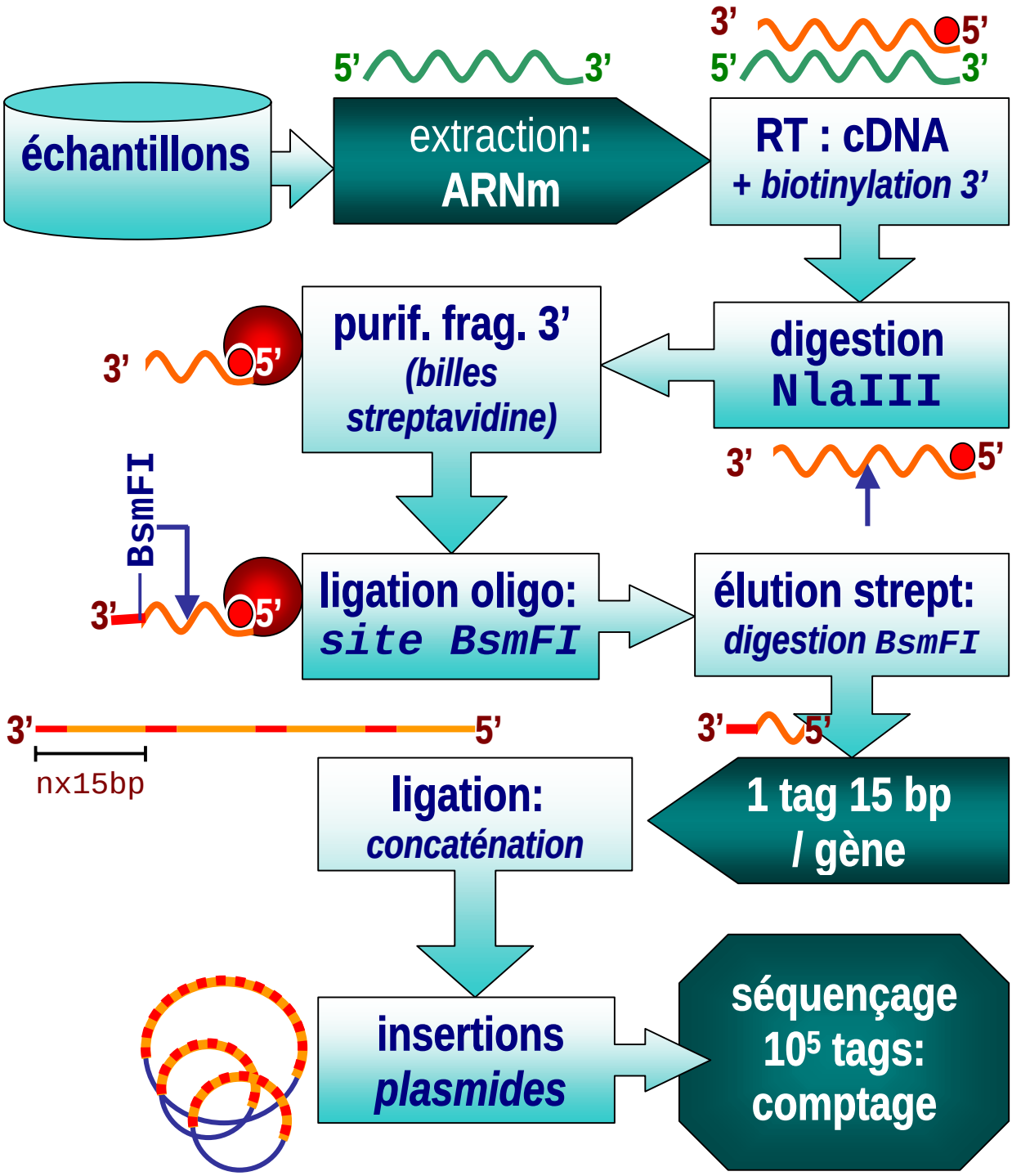


- *robots spotteurs*
- *imprimantes 'jet d'encre'*

- *photolithographie (Affymetrix)*
- *imprimantes 'jet d'encre'*

	micro-réseaux	puces à ADN
supports	lames verre	silicium
détection	fluo bi canaux	fluo mono canal
densité élts/cm^2	≈10 000	≈200 000
gènes (lxL mm)	6400 (18x18)	15000 (13x13)
publié	1995	1996

1.3.3 SAGE (Serial Analysis of Gene Expression)



1.3.4 Futur

- Microscopie Force Atomique (AFM)
- Micro-balances
- Adressage électronique
- Microbilles & fibres optiques

UniGene Build #120

Sequences Included in UniGene
==========

Known genes are from **GenBank** 117 (April 15,2000)
ESTs are from **dbEST** through 26-Aug-2000

37458	mRNAs + gene CDSs
1044880	EST, 3'reads
471665	EST, 5'reads
+ 293644	EST, other/unknown

1847647	total sequences in clusters

Final Number of Clusters (sets)
==========

83566	sets total
13959	sets contain at least one known gene
82395	sets contain at least one EST
12788	sets contain both genes and ESTs

Histogram of cluster sizes for UniGene Hs build 120

Cluster size	Number of clusters
1	**27193**
2	14045
3-4	13025
5-8	9222
9-16	5608
17-32	3819
33-64	3786
65-128	3574
129-256	2218
257-512	748
513-1024	222
1025-2048	70
2049-4096	31
4097-8192	4
8193-**16384**	1

translation elongation factor 1 alpha 1

1.4-Entités mesurées

Echelles :

- expressions **absolues**
  nb copies ARNm / cellule
- expressions **relatives**

induction ou *répression* / état référence (ratio)

échelle	contrôle	échantillon 1	échantillon 2
absolue	18.50	46.25	7.4
intuitive	référence	induction **x 2.5**	répression **÷ 2.5**
relative brute	1.0	2.5	0.4
relative $\log_2$	0	1.32	-1.32

Variables mesurables:

- abondance des transcrits (statique) :
  pulse
- taux de transcription instantané (dynamique) :
  pulse/chase
- taux de traduction instantané (dyn) :
  IP / anti-ribosomaux
- interactions ADN/protéines :
  IP (ex : facteurs transcription)

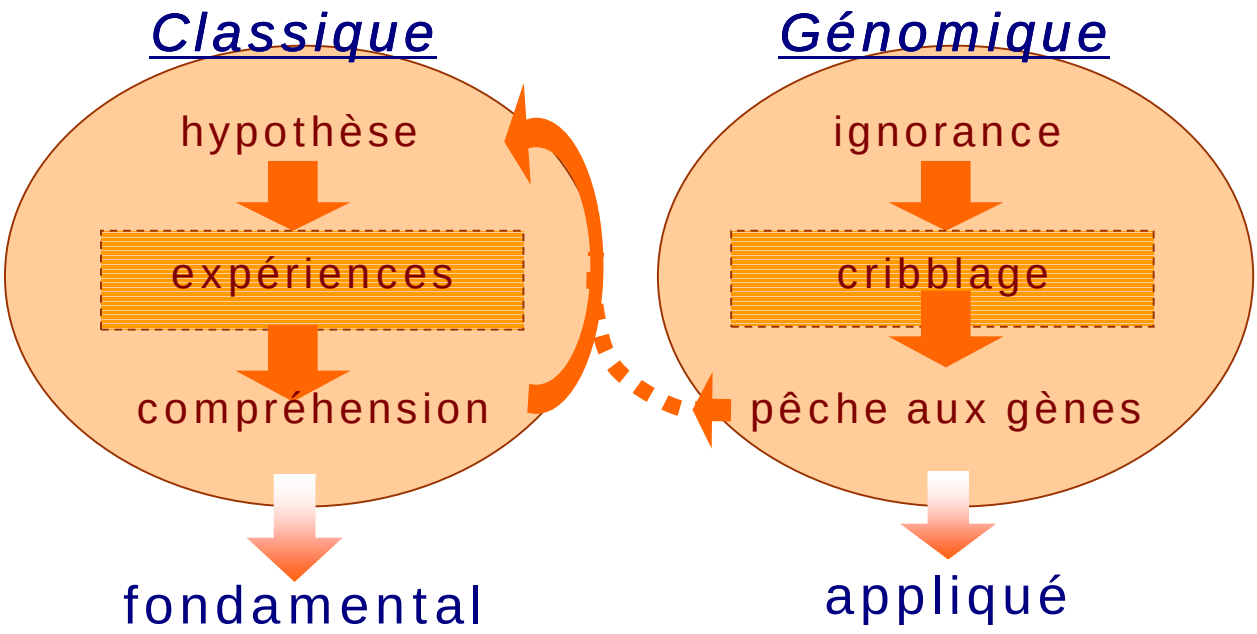
1.5-Applications

Recherche fondamentale

- caractérisations des gènes de fonction inconnue
- étude régulation transcription:➡ promoteurs etc.
- annotation relationnelle: ➡ réseaux géniques

Recherche appliquée

- identification de cibles thérapeutiques ➡ expression différentielle
- diagnostique / pronostique clinique ➡ classification des pathologies
- microbiologie, écologie ➡ impact environnement



1.6-Limites

Légitimité extrapolation niveaux **mRNA** ➡ activité **fonctionnelle** produits protéiques ?

Régulations transcription / traduction / activité enzymatique

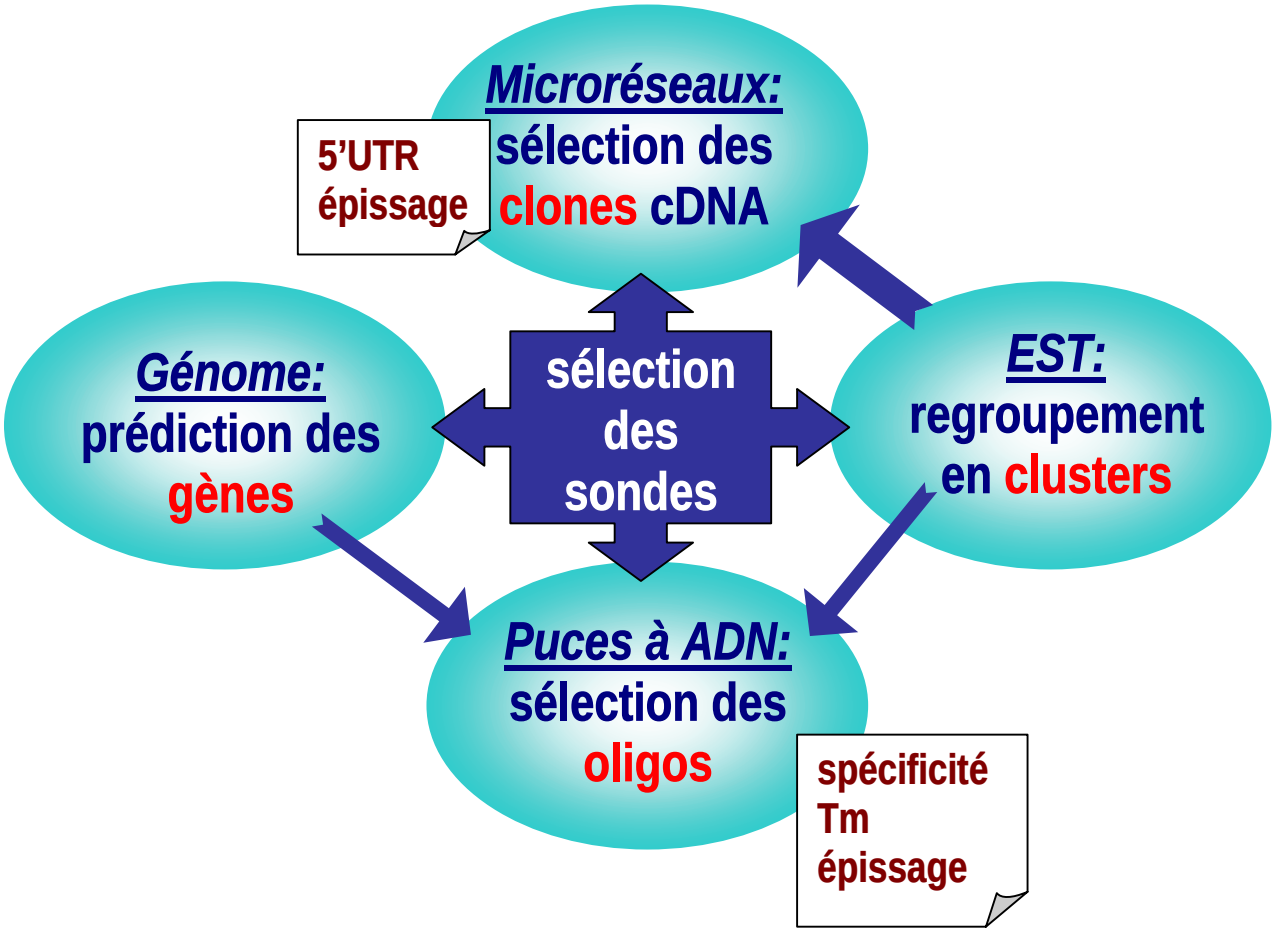
Fiabilité : variations expérimentales, estimation des erreurs, significativité, reproductibilité, hybridation non-spécifique (contrôle qualité? ➡ taille Northerns)

Artefacts liés aux mesures **globales :** méconnaissance des constantes & variables expérimentales ➡ fausses interprétations

Homogénéité des échantillons : types cellulaires variés (tissus complexes, cellules normales/pathologiques etc.)

3-BIOINFORMATIQUES DU TRANSCRIPTOME

3.1-Préproduction: *SONDES*



3.2 Production : *LIMS*

Laboratory Information Management System

Robotique

- repiquage clones
- PCR
- dépôts microréseaux

Analyse d'image

- quantification scans d'hybridation

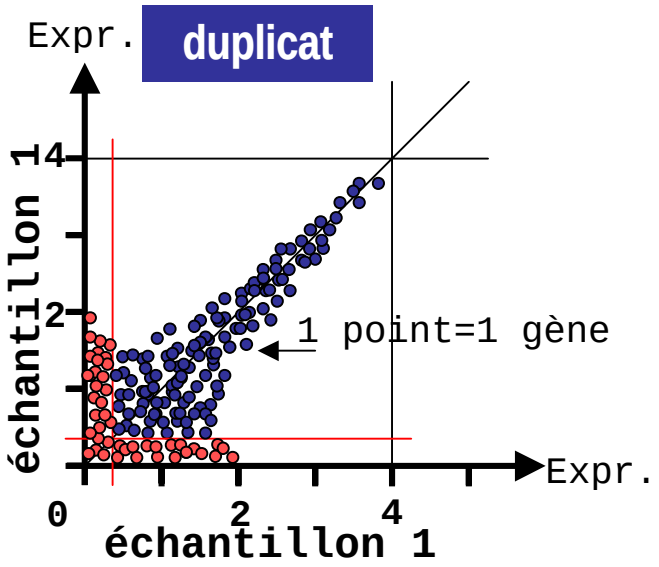
Flux de données

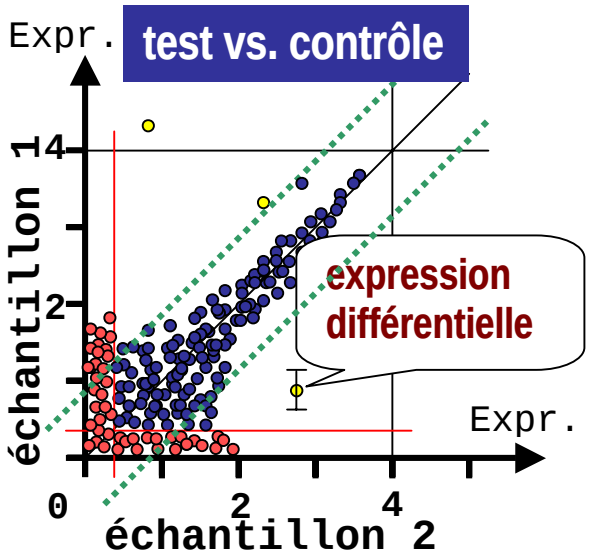
- bases de données
- XML
- annotation

10 exp
50 microréseaux
10000 gènes
id/signal/disp/bruit/
double canaux
= 40 millions données

3.3 Postproduction: *interprétation des résultats*

3.3.1 Comparaison 2 échantillons





3.3.2 Analyse n>2 échantillons

matrice n*k mesures d'expressions

	tissu 1	tissu 2	…	tissu n
gène 1	1.02	.04	…	1.14
gène 2	.972	1.57	…	.783
gène 3	.672	4.31	…	1.71
…	…	…	…	…
gène k	2.03	.947	…	.002

vecteurs d'expression

comparaisons ***profils*** (gènes / échantillons)

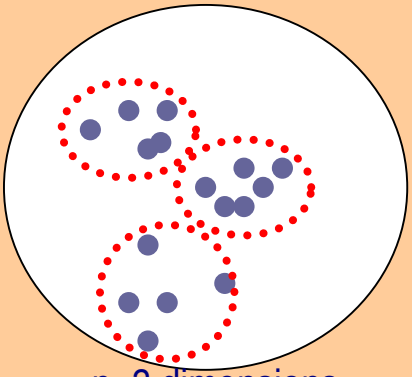
outils

analyse ***vecteurs*** / espace ***n dimensions***

Outils de classification

non-supervisée

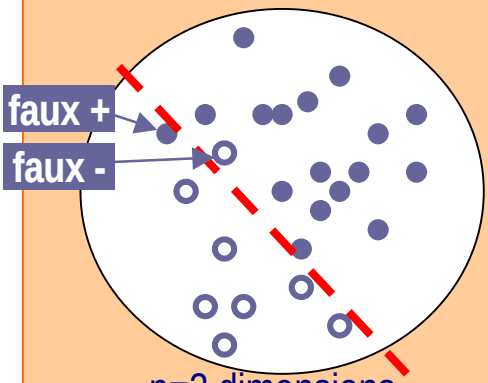
détecter sous-groupes



n=2 dimensions
3 clusters

supervisée

discriminer ≠ classes



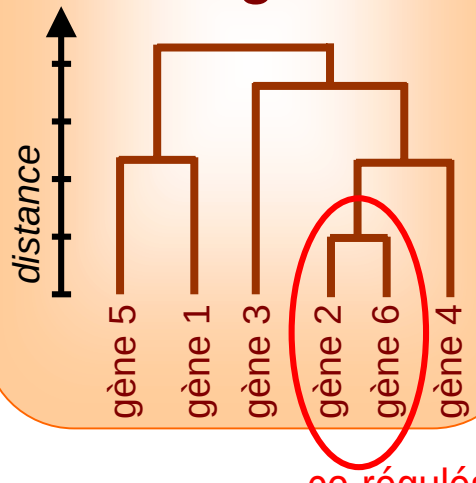
n=2 dimensions
2 classes: ○ & ●

ordonner par similitude

prédire appartenance

représentation clusters:

dendrogramme



co-régulés ➡ similitude fonction?

***sensibilité*:**

$$sb = \frac{VP}{VP + FN}$$

***spécificité*:**

$$sp = \frac{VP}{VP + FP}$$

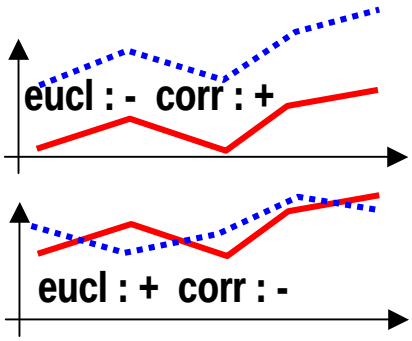
***fiabilité*:**

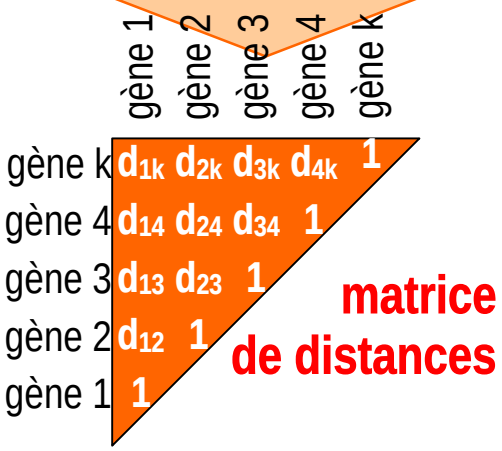
distance à ligne démarcation

Clustering (classification non-supervisée)

calcul **distances** entre chaque paire de vecteurs (gènes ou tissus) :

- distance euclidienne
- degré de corrélation





matrice de distances

Clustering :

- *k-means*
- *SOM's*
- *hierarchique*

paire gènes les + proches

fondent 1 noeud

n-1 itérations

recalcul matrice de distances

dimension -1

2 gènes remplacés par 'gène virtuel' = moyenne des 2 vecteurs

gène 4 gène 6

Classification supervisée

Classificateurs construits sur :

- discrimination linéaire (LDA)
- arbres décisionnels
- machines à support vectoriel (SVM)
- analyse de voisinage (NJ)

Appliquée aux :

- **gènes** ➡ prédiction fonctionnelle
- **échantillons** ➡ diagnostique, cibles thérapeutiques

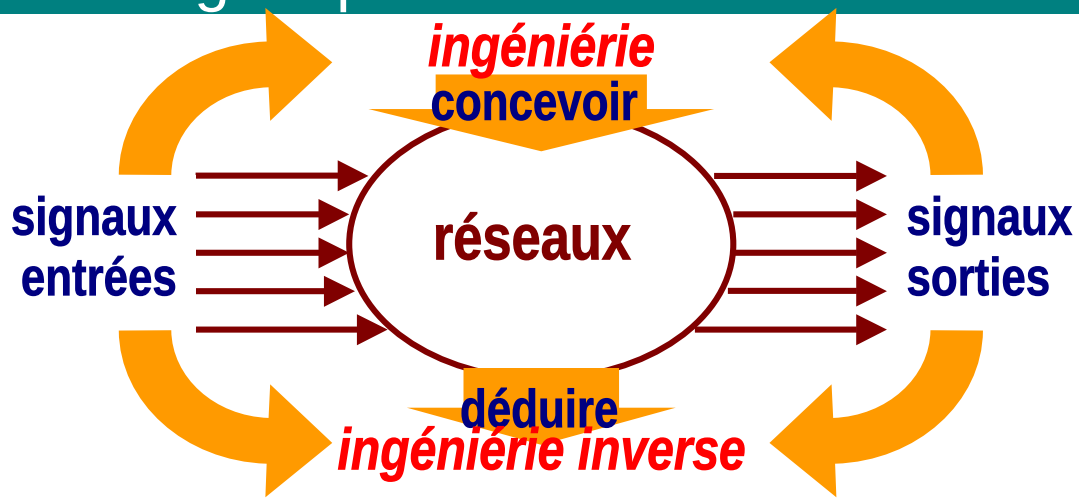
Risque / matrice dissymétrique:

nb attributs (gènes) **>>** nb objets (échantillons)

toutes classes ➡ *existe discriminateurs parfaits*

contrainte de complexité minimale

Réseaux géniques



Puces à ADN : synoptique des technologies

	cDNA			oligonucleotides synth.			SAGE
	radioactif	colorimetrie	fluorescent	photolithogr.	'jet d'encre'	dépôts	séquençage
cDNA cloné	oui	oui	oui	non	non	non	non
gène séquencé	non	non	non	oui	oui	oui	oui
nouveaux gènes	+	+	+	-	+++	++	n/a tous inclus
spécificité	+++	+++	+++	+ court	++	++	+++
sensibilité	+++	+	+	++	++	++	variable
gamme dynam.	+++	-	+	++	++	++	variable
discrimin. fine	-	-	-	+++	+++	+++	+
nb. canaux	1 (3?)	1	2	1	1+	1+	n/a
rapidité	+	+++	++	++	++	++	-
coûts	-/+	--/+	+	++++	++?	+++	++