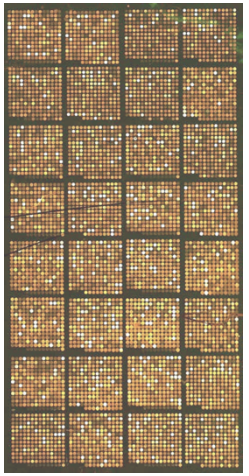
Master de génétique
UE génomique fonctionnelle
Université Denis Diderot – Janvier 2009

Normalisation

Stéphane Le Crom (lecrom@biologie.ens.fr)

Laboratoire de Génétique Moléculaire du Développement - INSERM U784
Plate-forme Transcriptome - IFR36
École normale supérieure

La bioinformatique dans une expérience biopuces



Mise au point du dessin expérimental

Hybridation

Expérimentation

Analyse des données

Fouille de données

Analyse d'image

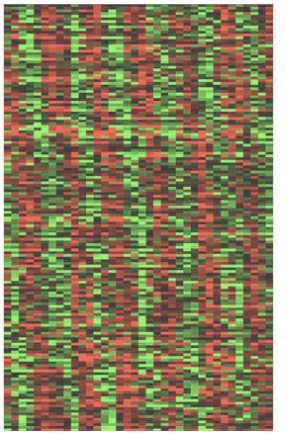
Traitement des données

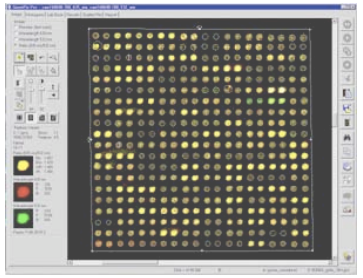
Traitement des données brutes

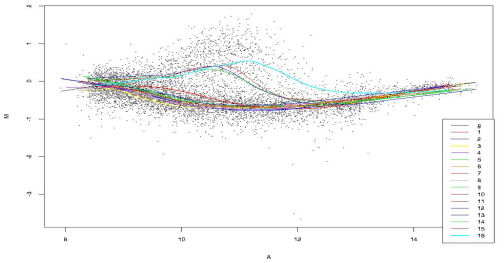
- Normalisation
- Analyse statistique

Représentation des données

- Clustering





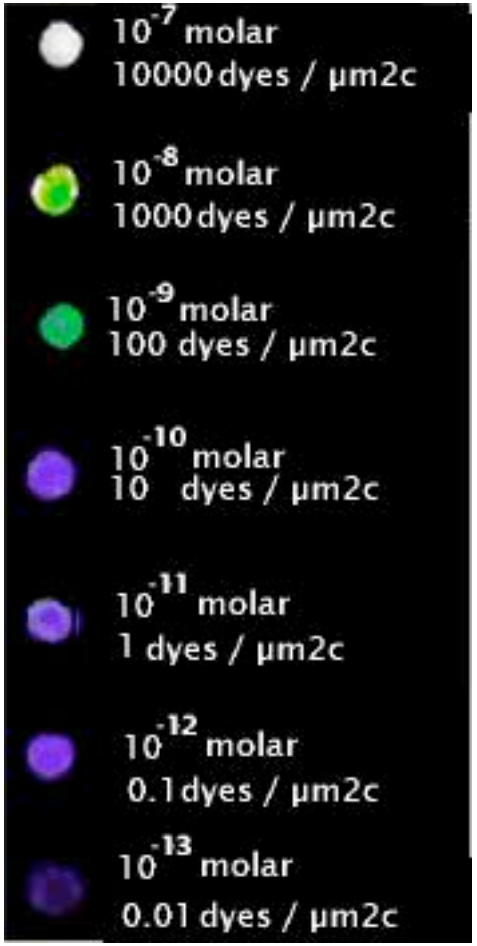


Analyse bioinformatique des puces à ADN

La segmentation du signal

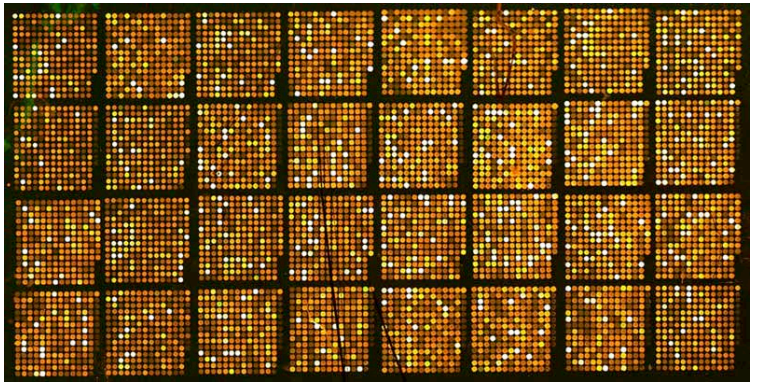
Les différents types d'images rencontrés

Echelle de couleur
=
Echelle Quantitative



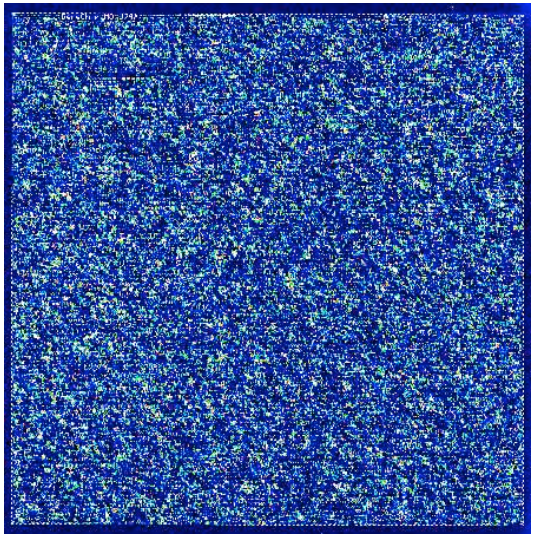
- **Puces à ADNc/oligonucleotides - ENS**

- 2 canaux
- Superposition
  (= Ratio)

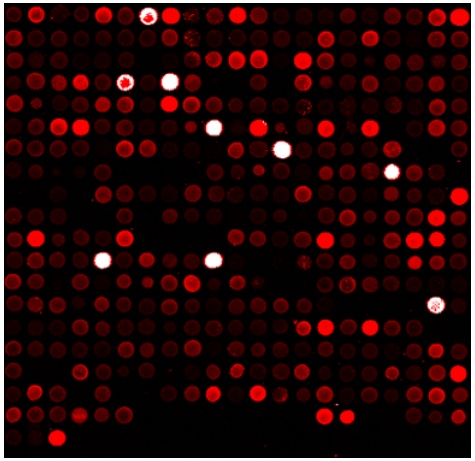


- **Puces à Oligonucleotides - Affymetrix**

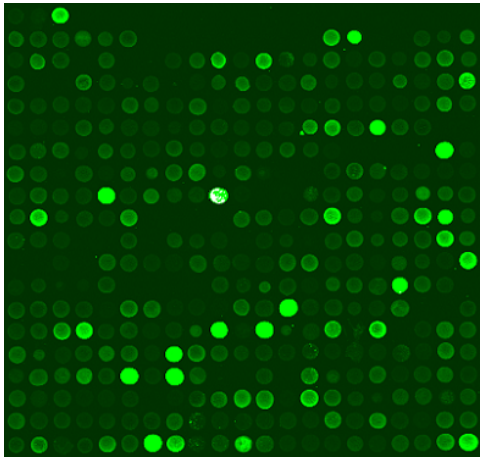
- 1 canal
- Intensité
  (= quantité d'ARN)



Obtention de l'image



Longueur d'onde du Cy5



Longueur d'onde du Cy3

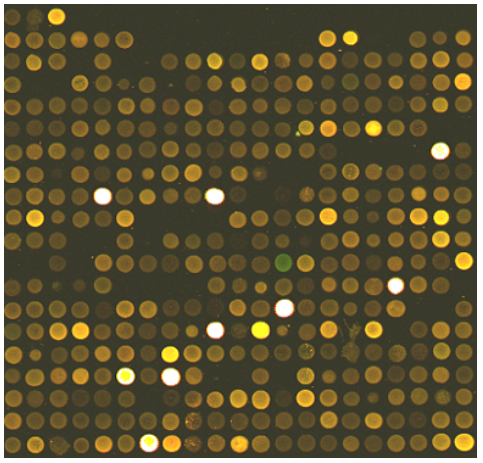
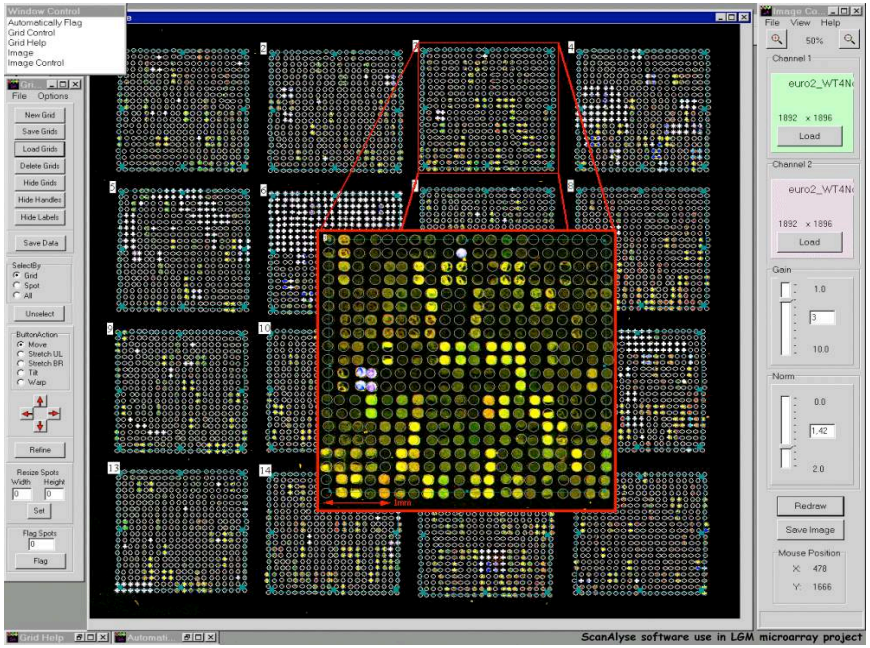


Image finale

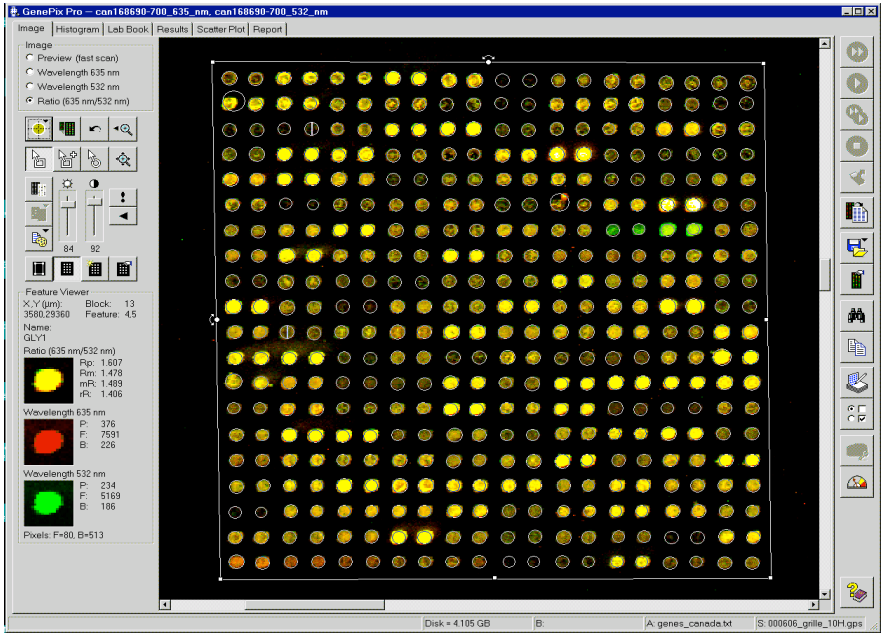
Principes généraux de l'analyse d'image

- **Convertir l'image en valeurs numériques pour quantifier l'expression**

Il existe des logiciels d'analyse d'image …



ScanAlyze
(M. Eisen Stanford University)



Genepix Pro
(Axon software)

Les différentes étapes de l'analyse d'image

1 — Localisation des spots sur la lame

Pour chaque spot

2 — Délimitation des pixels correspondant à la zone d'hybridation

3 — Délimitation des pixels pour l'estimation du bruit de fond

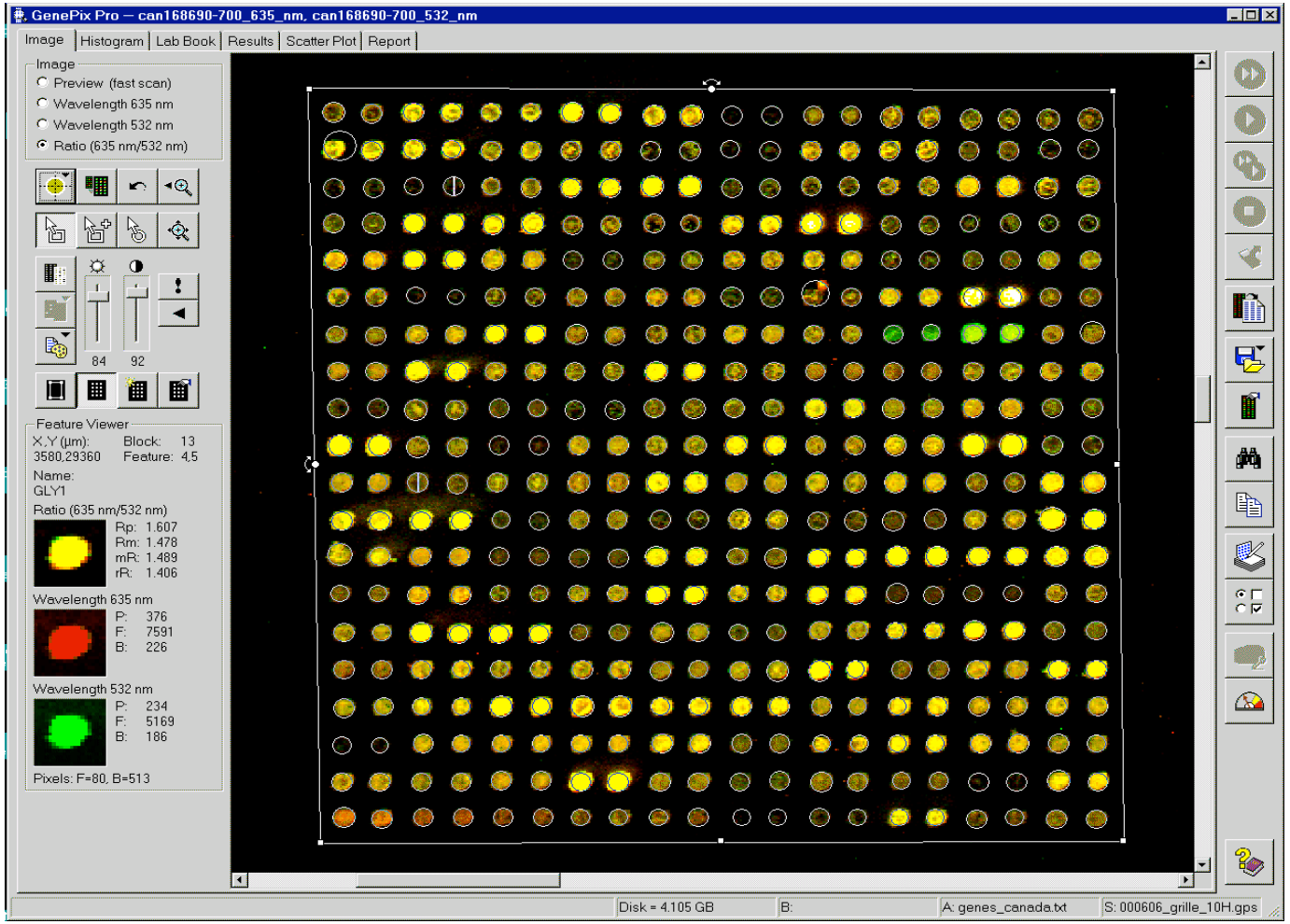
4 — Calcul de l'intensité globale de fluorescence

Sur l'ensemble de la lame

5 — Identification des spots déformés par des artéfacts

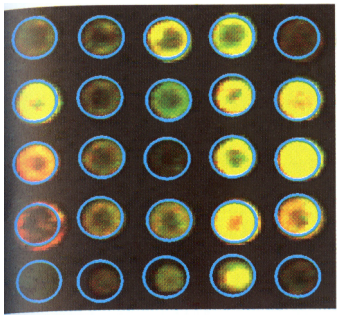
Localisation des spots sur la lame

- **Il est important d'assigner à chaque spot un identifiant correct !**

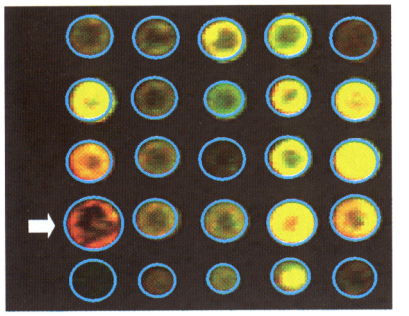


Délimitation des pixels de la zone d'hybridation

- **Différentes méthodes sont possibles (segmentation)**
  - Cercles à diamètres fixes
  - Cercles à diamètres variables
  - Histogrammes
  - Formes « adaptée »

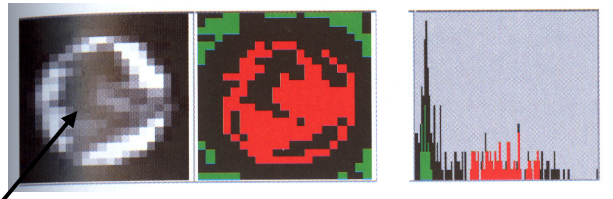


Cercles à diamètre fixe



Cercles à diamètre variable

- **Plusieurs logiciels sont disponibles**
  - GenePix Pro
  - ScanAlyze
  - QuantArray
  - ImaGene
  - Dapple



Intensité variable

Méthode par histogramme

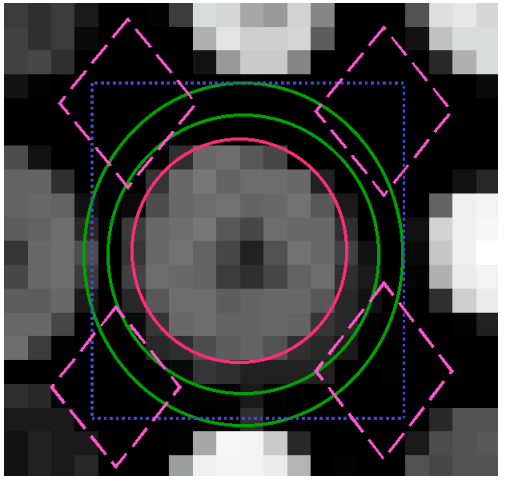
D'après « Microarray Bioinformatics », Dov Stekel

Délimitation des pixels du bruit de fond

Les signaux observés ont deux composantes

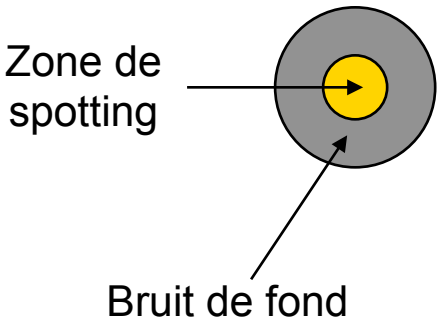
Fluorescence issue d'une hybridation spécifique

Fluorescence non spécifique : **Bruit de fond**

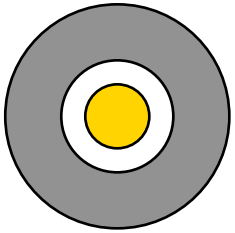


- **Analyse des pixels localisés à proximité de la région spottée**

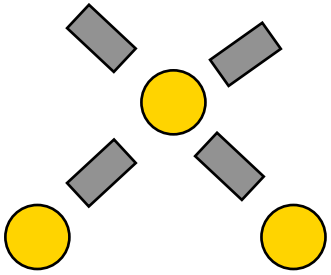
Méthode « ScanAlyze »

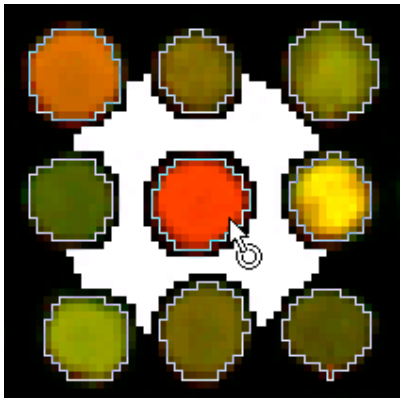


Méthode « ImaGene »

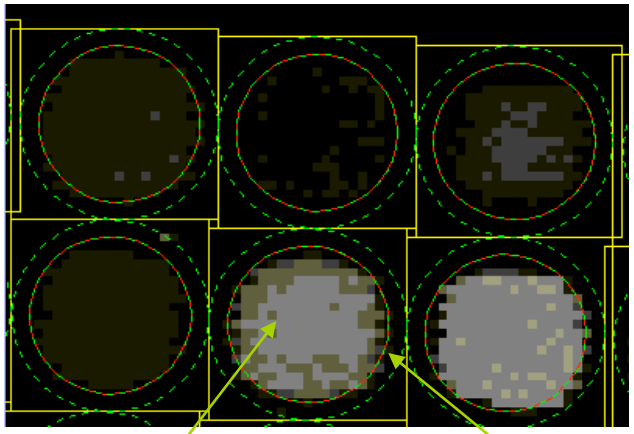


Méthode « GenePix »





Calcul de l'intensité globale de fluorescence



Intensité net
=
Intensité brute - Bruit de fond

Intensité à l'intérieur du spot

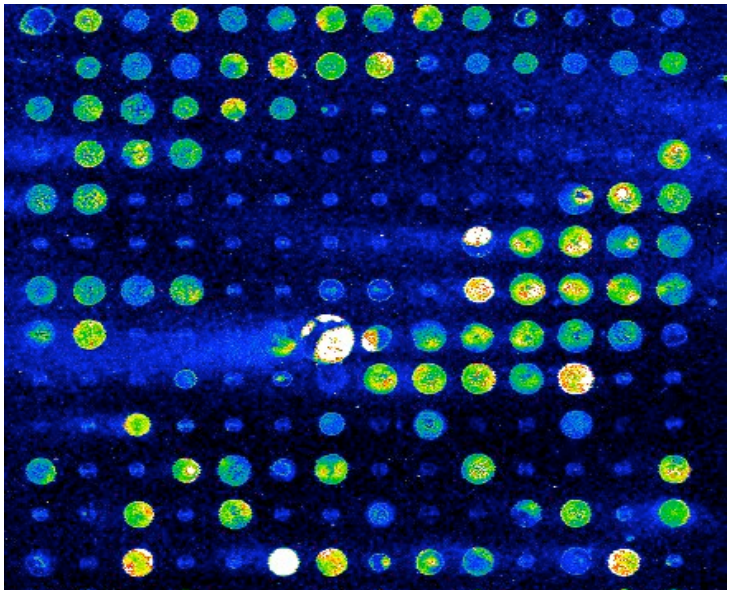
Mesure du bruit de fond

Intensité par pixel → Moyenne ou médiane ? → Intensité globale du spot

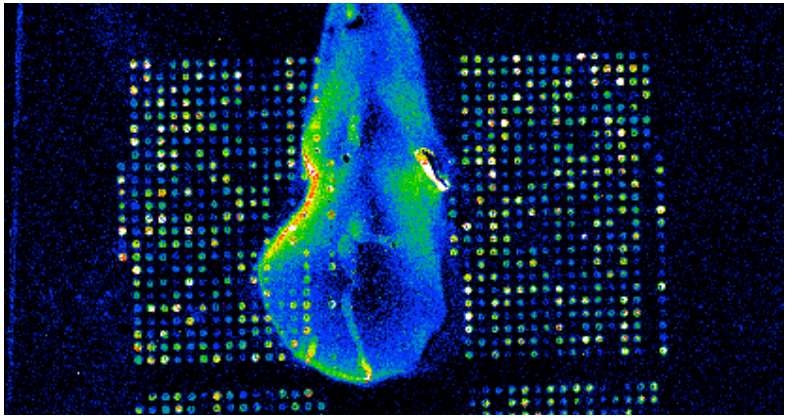
Critère qualité : taille du spot, déviation standard…

Annotation des spots non conformes

- **Exemples**
  - Effet « comète »
  - Tâches d'hybridation
  - Problèmes de lavages



- **Solutions**
  - Toujours regarder les contrôles qualité des lots de lames commandés
  - Se référer aux aides en ligne
  - Élimer les spots artéfactuels à l'aide d'un filtre manuel ou automatique

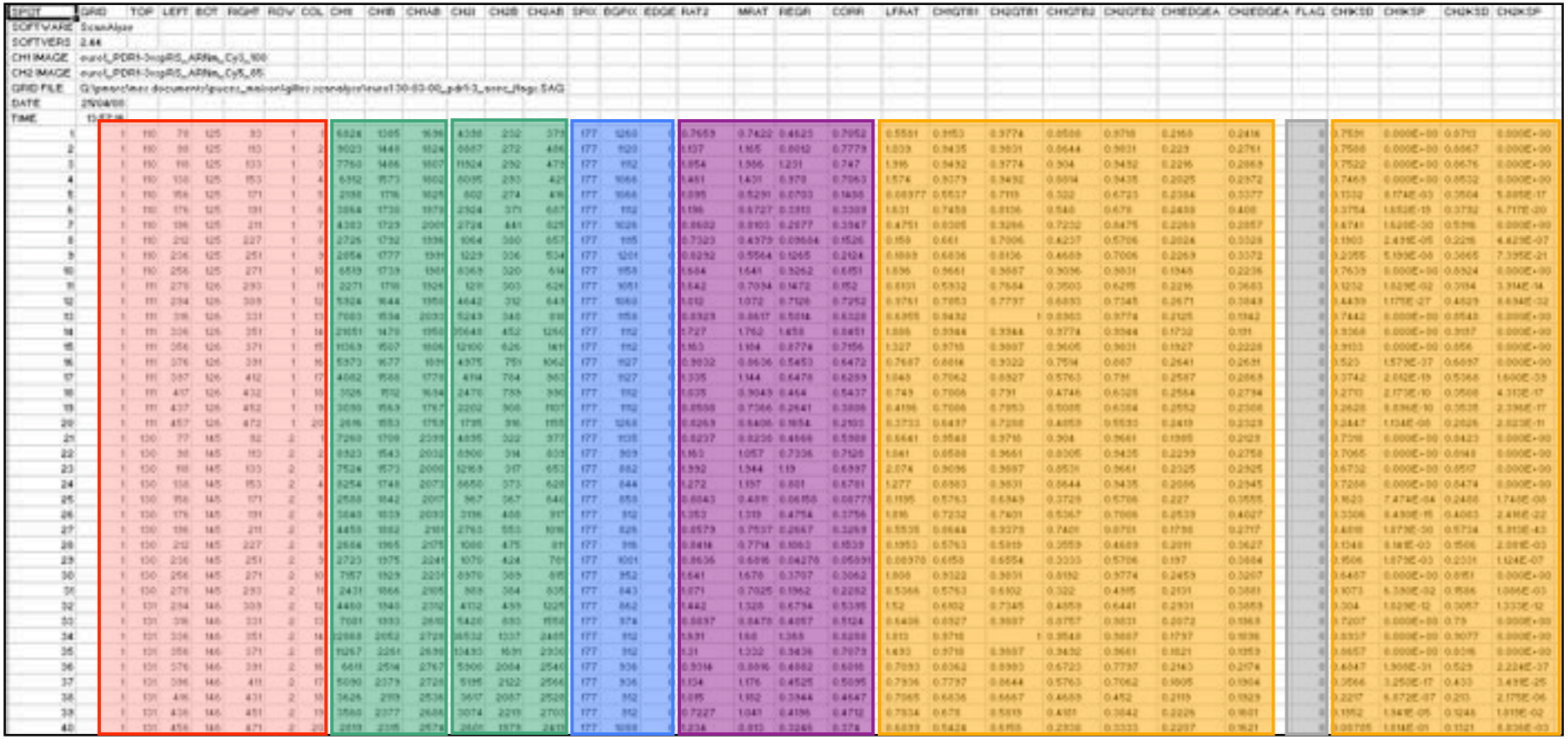


http://www.corning.com/lifesciences/technical_information/techdocs/ troubleshootingUltraGAPS_ProntoReagents.asp

Récupération des résultats bruts

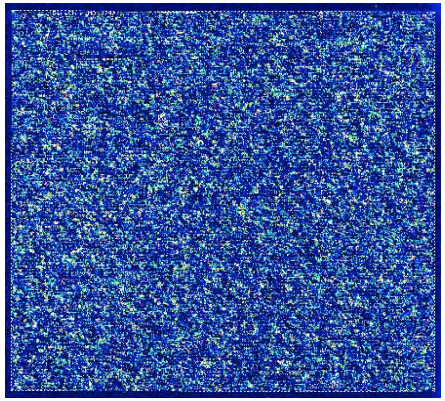
- **Exemple d'un fichier de sortie de logiciel d'analyse d'image**

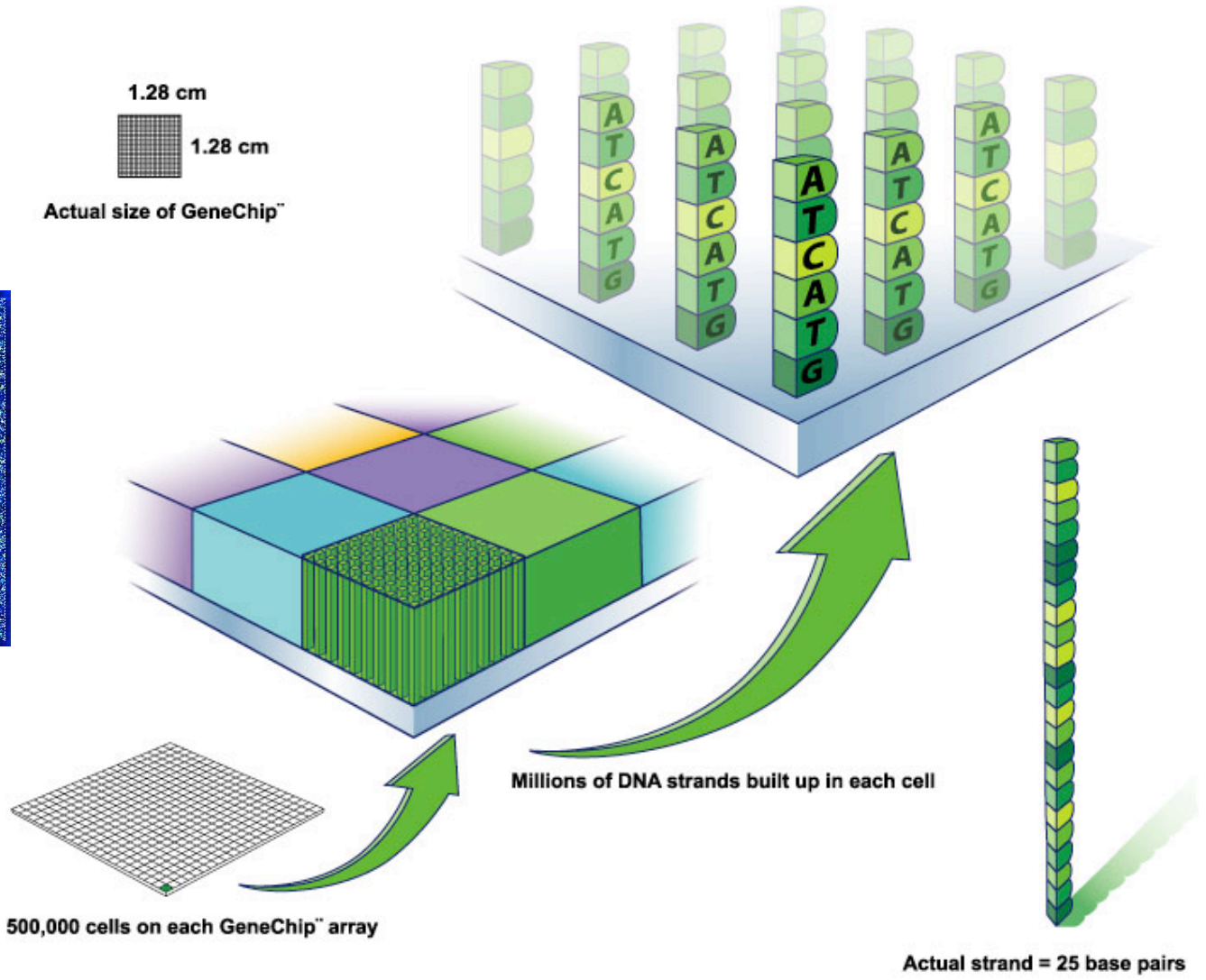
Position | Intensités Brutes | Intensités BDF | Taille | Homogénéité | Statistiques | Annotations



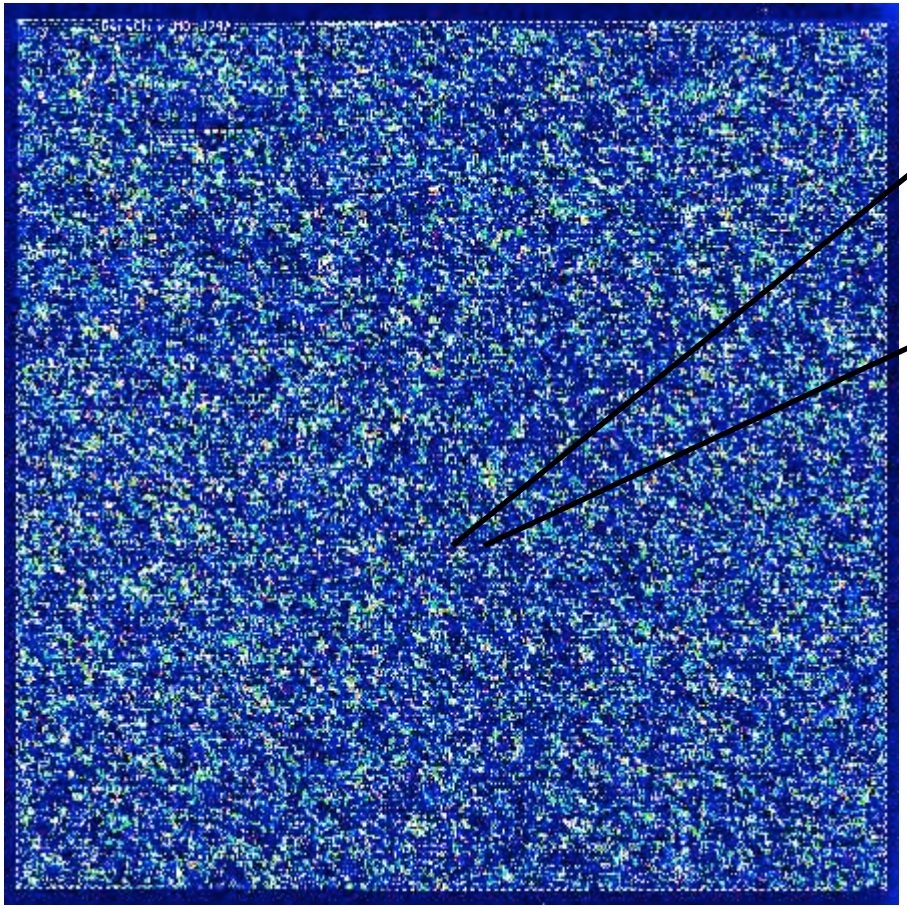
SPOT	GRID	TOP	LEFT	BOT	RIGHT	ROW	COL	CH1I	CH1B	CH1AB	CH2I	CH2B	CH2AB	SPIX	BGPIX	EDGE	RAT2	MRAT	REGR	CORR	LFRAT	CH1GTB1	CH2GTB1	CH1GTB2	CH2GTB2	CH1EDGEA	CH2EDGEA	FLAG	CH1KSD	CH1KSP	CH2KSD	CH2KSP
SOFTWARE	ScanAlyse																															
SOFTVERS	2.64																															
CH1 IMAGE	eure1_PDR1-3vsgRIS_ARNm_Cy3_100																															
CH2 IMAGE	eure1_PDR1-3vsgRIS_ARNm_Cy5_65																															
GRID FILE	G:\genoscriptes documents\puces_maison\gilles zscanalyse\eure1 30-03-06_pdr1-3_sens_flag.SAG																															
DATE	29/04/06																															
TIME	13:57:38																															
1	1	110	78	125	93	1	1	6826	1365	1636	4336	232	373	177	1260	0	0.7653	0.7422	0.4623	0.7952	0.5581	0.9153	0.3774	0.6586	0.9716	0.2169	0.2416	0	3.7531	0.000E+00	0.0703	0.000E+00
2	1	110	93	125	110	1	2	9023	1448	1624	6667	272	406	177	1120	0	1.037	1.165	0.6012	0.7779	1.033	0.9435	0.3931	0.6644	0.9931	0.223	0.2761	0	3.7566	0.000E+00	0.6667	0.000E+00
3	1	110	110	125	133	1	3	7760	1486	1667	11624	250	473	177	1112	0	0.854	1.556	1.231	0.747	1.395	0.9432	0.3774	0.904	0.9432	0.2216	0.2869	0	3.7522	0.000E+00	0.6676	0.000E+00
4	1	110	128	125	153	1	4	6352	1573	1662	5095	293	421	177	1066	0	1.451	1.431	0.978	0.7063	1.574	0.9279	0.3432	0.8894	0.9435	0.2025	0.2572	0	3.7469	0.000E+00	0.6532	0.000E+00
5	1	110	156	125	171	1	5	2038	1716	1625	602	274	418	177	1066	0	1.095	0.5291	0.0703	0.1636	0.08377	0.6537	0.7113	0.322	0.6723	0.2384	0.3377	0	3.1532	0.174E-03	0.3504	5.805E-17
6	1	110	176	125	191	1	6	3064	1730	1673	2304	371	667	177	1112	0	1.196	0.8727	0.3313	0.3308	1.631	0.7458	0.3106	0.548	0.678	0.2459	0.408	0	3.3754	1.652E-15	0.3732	6.717E-20
7	1	110	196	125	211	1	7	4383	1723	2001	2724	441	625	177	1026	0	0.8602	0.8103	0.2677	0.3367	0.4751	0.6305	0.3286	0.7232	0.8475	0.2263	0.2857	0	3.4741	1.620E-30	0.5316	0.000E+00
8	1	110	212	125	227	1	8	2725	1732	1936	1064	360	657	177	1115	0	0.7323	0.4979	0.09684	0.4526	0.159	0.661	0.7006	0.4237	0.5706	0.2024	0.3329	0	3.1903	2.439E-05	0.2216	6.482E-07
9	1	110	234	125	251	1	9	2654	1777	1911	1229	336	534	177	1201	0	0.8232	0.5564	0.1265	0.2124	0.1869	0.6836	0.8106	0.4669	0.7006	0.2263	0.3372	0	3.2055	5.193E-06	0.3665	7.335E-21
10	1	110	256	125	271	1	10	6519	1739	1561	6363	320	614	177	1153	0	1.584	1.641	0.9262	0.6851	1.636	0.9661	0.3667	0.9096	0.9831	0.1945	0.2236	0	3.7639	0.000E+00	0.6524	0.000E+00
11	1	111	278	126	293	1	11	2271	1716	1926	1211	303	626	177	1051	0	1.842	0.7034	0.1472	0.152	0.6101	0.5932	0.7564	0.3503	0.6215	0.2216	0.3663	0	3.1232	1.525E-02	0.3184	3.314E-14
12	1	111	294	126	309	1	12	5324	1644	1950	4642	312	643	177	1060	0	1.012	1.072	0.7126	0.7252	0.9751	0.7853	0.7737	0.6893	0.7345	0.2571	0.3843	0	3.4439	1.175E-27	0.4829	8.634E-32
13	1	111	316	126	331	1	13	7003	1634	2033	5243	340	618	177	1159	0	0.8328	0.8617	0.5016	0.6328	0.6305	0.9432		1	0.9363	0.2125	0.3162	0	3.7442	0.000E+00	0.6543	0.000E+00
14	1	111	336	126	351	1	14	21651	1479	1958	20648	852	1260	177	1112	0	1.727	1.762	1.658	0.8451	1.686	0.9366	0.3366	0.9774	0.9366	0.1732	0.191	0	3.3068	0.000E+00	0.3937	0.000E+00
15	1	111	356	126	371	1	15	10363	1597	1806	12100	626	1161	177	1112	0	1.163	1.164	0.8774	0.7156	1.327	0.9715	0.3997	0.9605	0.9931	0.1927	0.2228	0	3.9103	0.000E+00	0.656	0.000E+00
16	1	111	376	126	391	1	16	5373	1677	1931	4375	751	1062	177	1127	0	0.9032	0.8636	0.5453	0.6472	0.7587	0.8616	0.3322	0.7514	0.887	0.2641	0.2931	0	3.523	1.579E-27	0.6697	0.000E+00
17	1	111	397	126	412	1	17	4082	1568	1972	4114	764	1163	177	1127	0	1.335	1.144	0.6479	0.6259	1.049	0.7062	0.3927	0.5763	0.791	0.2587	0.2868	0	3.3742	2.652E-19	0.5166	1.600E-39
18	1	111	417	126	432	1	18	3926	1512	1934	2470	735	990	177	1112	0	1.035	0.9049	0.464	0.5437	0.749	0.7006	0.731	0.4746	0.6328	0.2564	0.2794	0	3.2713	2.170E-10	0.3508	4.310E-17
19	1	111	437	126	452	1	19	3050	1563	1767	2202	366	1107	177	1112	0	0.8598	0.7366	0.2641	0.3806	0.4196	0.7006	0.7353	0.5008	0.6304	0.2552	0.2306	0	3.2628	8.836E-10	0.3535	2.336E-17
20	1	111	457	126	472	1	20	2616	1553	1753	1735	316	1155	177	1258	0	0.6269	0.6406	0.1854	0.2103	0.3733	0.6497	0.7208	0.4859	0.5533	0.2419	0.2323	0	3.0447	1.134E-06	0.2026	2.023E-11
21	1	130	77	145	92	2	1	7260	1709	2339	4835	322	573	177	1035	0	0.8237	0.8236	0.4569	0.5308	0.6661	0.9568	0.3715	0.904	0.9661	0.1965	0.2129	0	3.7216	0.000E+00	0.6423	0.000E+00
22	1	130	96	145	110	2	2	6523	1643	2032	5300	314	603	177	1169	0	1.163	1.057	0.7336	0.7126	1.041	0.9588	0.3661	0.8306	0.9435	0.2239	0.2758	0	3.7065	0.000E+00	0.6148	0.000E+00
23	1	130	116	145	133	2	3	7526	1673	2000	5263	317	653	177	1162	0	1.392	1.344	1.13	0.6397	2.074	0.9096	0.3667	0.8531	0.9661	0.2325	0.2925	0	3.6732	0.000E+00	0.6597	0.000E+00
24	1	130	136	145	153	2	4	8254	1748	2073	5650	373	628	177	1144	0	1.272	1.157	0.831	0.6791	1.277	0.9583	0.3831	0.8644	0.9435	0.2086	0.2945	0	3.7256	0.000E+00	0.6474	0.000E+00
25	1	130	156	145	171	2	5	2588	1642	2017	967	267	640	177	1158	0	0.6043	0.4811	0.06155	0.08779	0.1195	0.5763	0.6349	0.3729	0.5706	0.227	0.3555	0	3.1623	7.474E-04	0.2486	1.740E-08
26	1	130	176	145	191	2	6	5843	1639	2033	3196	488	917	177	1112	0	1.353	1.319	0.4754	0.3756	1.186	0.7232	0.7601	0.5367	0.7006	0.2539	0.4027	0	3.3306	5.430E-15	0.4063	2.416E-22
27	1	130	196	145	211	2	7	4455	1662	2161	2763	553	1018	177	1128	0	0.8573	0.7537	0.2667	0.3268	0.5535	0.8648	0.3378	0.7401	0.8701	0.1798	0.2717	0	3.4818	1.873E-30	0.5734	5.313E-43
28	1	130	216	145	227	2	8	2664	1665	2075	1066	475	911	177	1116	0	0.8416	0.7714	0.5063	0.6539	0.1953	0.5763	0.5919	0.3559	0.4669	0.2011	0.3627	0	3.1048	0.181E-03	0.1506	2.091E-03
29	1	130	236	145	251	2	9	2723	1675	2241	1070	424	761	177	1001	0	0.8636	0.6696	0.84276	0.95591	0.08378	0.6458	0.6554	0.3333	0.5706	0.197	0.3684	0	3.1506	1.875E-03	0.2316	1.124E-07
30	1	130	256	145	271	2	10	7357	1929	2231	6570	369	915	177	1152	0	1.641	1.679	0.3707	0.3062	1.806	0.9122	0.3931	0.8192	0.9774	0.2459	0.3267	0	3.6457	0.000E+00	0.6951	0.000E+00
31	1	130	276	145	293	2	11	2431	1966	2195	1113	384	935	177	1143	0	1.071	0.7025	0.1962	0.2282	0.5366	0.5763	0.6102	0.322	0.4395	0.2131	0.3681	0	3.1073	5.336E-02	0.1586	1.066E-03
32	1	131	294	146	309	2	12	4450	1940	2352	4712	459	1225	177	1162	0	1.442	1.528	0.6734	0.5395	1.52	0.6102	0.7345	0.4859	0.6441	0.2301	0.3858	0	3.304	1.829E-12	0.3067	1.233E-12
33	1	131	316	146	331	2	13	7081	1933	2610	5420	693	1556	177	1179	0	0.8837	0.8473	0.4057	0.5124	0.6406	0.9327	0.3887	0.8767	0.9831	0.2072	0.3363	0	3.7207	0.000E+00	0.79	0.000E+00
34	1	131	336	146	351	2	14	22868	2052	2729	16532	1237	2485	177	1112	0	1.531	1.66	1.365	0.8288	1.833	0.9715	1	0.9549	0.9887	0.1737	0.1936	0	3.3337	0.000E+00	0.3077	0.000E+00
35	1	131	356	146	371	2	15	11267	2261	2636	10435	1631	2930	177	1112	0	1.31	1.332	0.9436	0.7979	1.493	0.9715	0.3887	0.9452	0.9661	0.1821	0.1953	0	3.6657	0.000E+00	0.6106	0.000E+00
36	1	131	376	146	391	2	16	6611	2514	2767	5906	2004	2540	177	1136	0	0.9316	0.8996	0.4862	0.6016	0.7093	0.6362	0.8393	0.6723	0.7737	0.2163	0.2174	0	3.6547	1.390E-31	0.523	2.224E-37
37	1	131	396	146	411	2	17	5090	2379	2728	5195	2122	2564	177	1136	0	1.134	1.176	0.4525	0.5095	0.7936	0.7737	0.8644	0.5763	0.7062	0.1905	0.1904	0	3.2566	3.250E-17	0.433	3.491E-25
38	1	131	416	146	431	2	18	3626	2193	2538	3617	2057	2528	177	1112	0	1.015	1.152	0.3944	0.4647	0.7065	0.6836	0.6667	0.4669	0.452	0.2115	0.1923	0	3.2207	5.672E-07	0.293	2.175E-06
39	1	131	436	146	451	2	19	3560	2327	2668	3074	2219	2703	177	1112	0	0.7227	1.041	0.4196	0.4712	0.7934	0.678	0.5919	0.4101	0.3842	0.2226	0.1801	0	3.1952	1.943E-05	0.1248	1.815E-02
40	1	131	456	146	471	2	20	2659	2316	2674	2601	1979	2491	177	1098	0	1.234	0.913	0.3265	0.378	0.6039	0.5424	0.6166	0.2936	0.3333	0.2297	0.1621	0	3.08705	1.814E-01	0.1321	0.836E-03

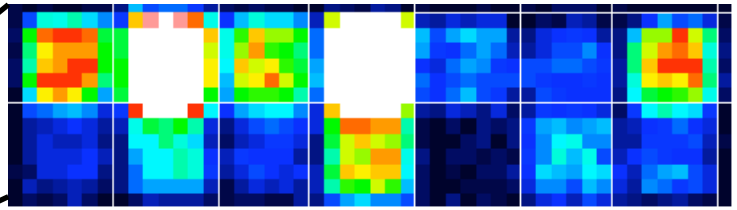
Les puces Affymetrix (GeneChip)





Obtention de l'image (GeneChip Affymetrix)





PM

MM

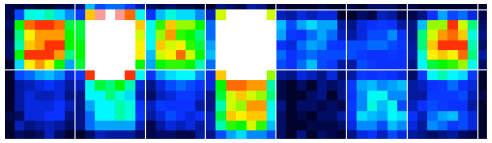
Des paires d'oligonucléotides sont crées pour chaque gènes :

- "perfect match" PM
- "mismatch" MM

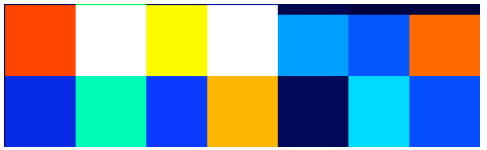
Sélection des spots corrects

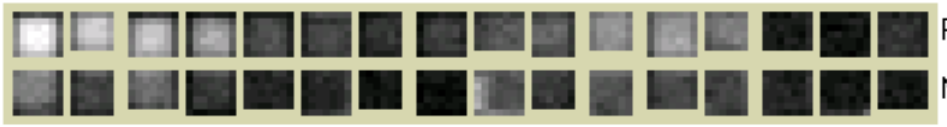
- **Estimation des valeurs moyennes**

Image obtenue pour un gène

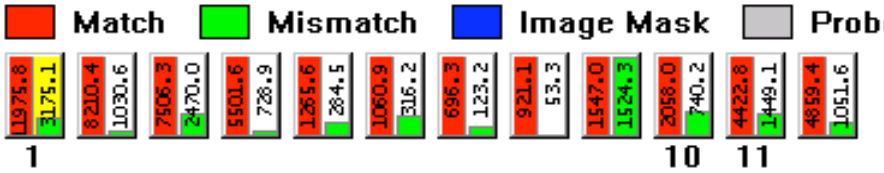


Valeurs moyennes calculées pour chaque cellule



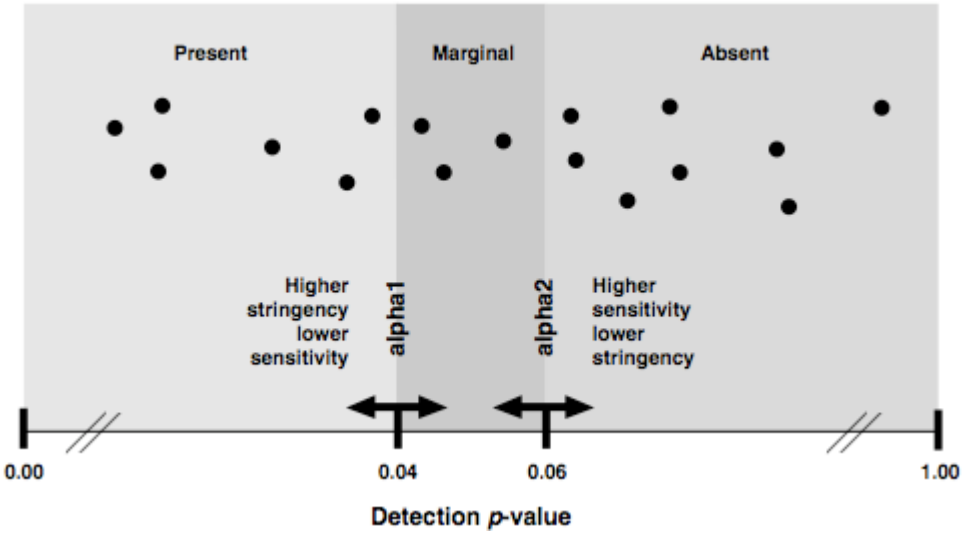


96616_at (C3MG)
Intensity: 38 - 2174



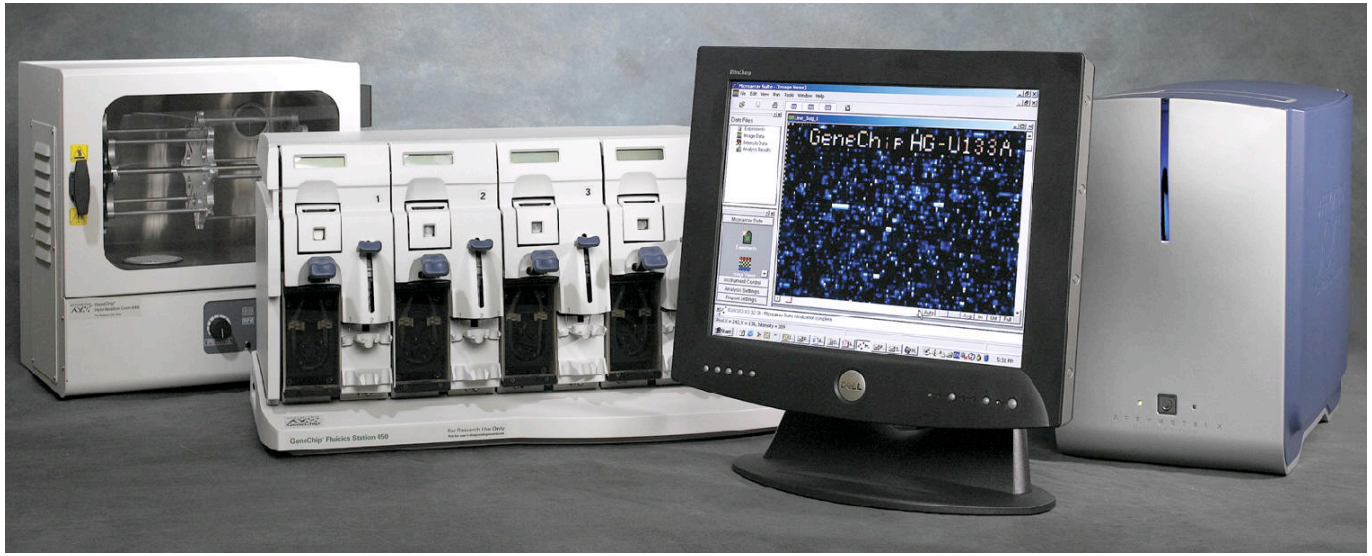
- Calcul du nombre de paires positives et négatives (matrice de décision)

- Détermination du statut des gènes (présent, marginal ou absent)



Les données brutes en sortie d'analyse

- **Différents formats de fichiers**
  - Le fichier d'expérience : .EXP
  - Le ficher d'image équivalent aux fichiers TIFF de GenePix : .DAT
  - Les données d'intensité pour chaque sonde sans traitement obtenues à partir du fichier .DAT : .CEL
  - Les données traitées (par le programme d'Affymetrix, GCOS) et rassemblées par probeset : .CHP



Analyse bioinformatique des puces à ADN

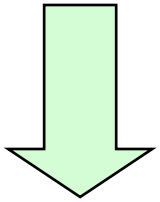
La normalisation d'une lame

Les sources de variations d'une puce à ADN

- Quantité d'ADN
- Efficacité des étapes expérimentales (extraction, RT, …)
- Les fluorochromes (Cy3 brille plus que Cy5, accrochages ≠)

Erreur systématique

- Effets similaires sur plusieurs mesures
- Les corrections peuvent se mesurer à partir des données

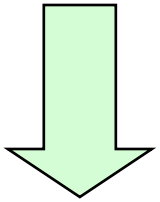


Calibration

- Rendement de la PCR
- Qualité des ADN déposés et du spotting
- Effets de cross-hybridation et hybridation non spécifique

Erreur stochastique

- Effets qui se produisent de façon trop aléatoire et qui du coup ne peuvent se mesurer comme du bruit

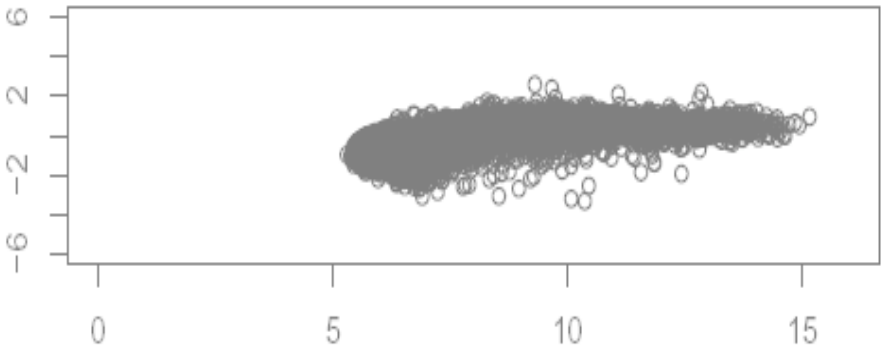


Modèle d'erreur

1ère étape : le nettoyage des données

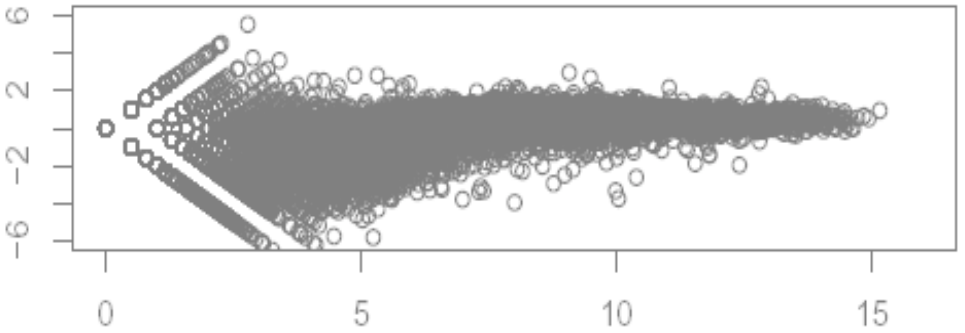
1/ Élimination des gènes dont l'annotation est différente de 0

- Éventuellement retourner voir l'image originale)



2/ Filtrage sur les intensités

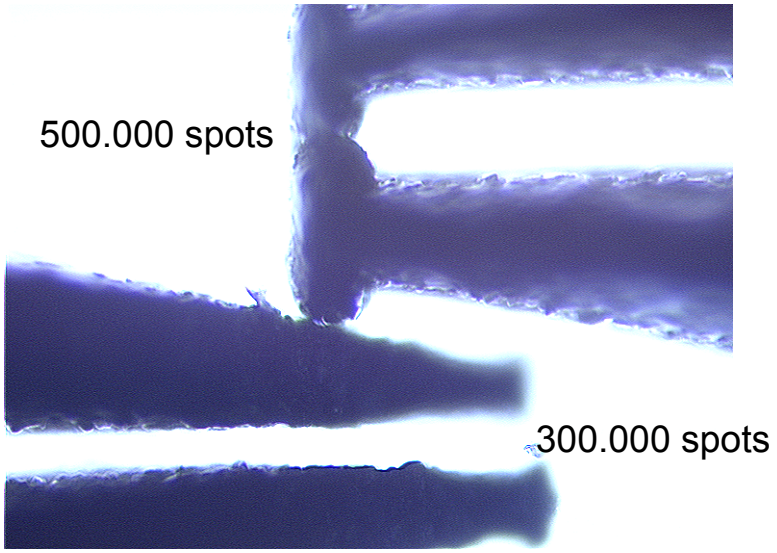
- Saturation du scanner
- Écart avec le bruit de fond trop faible



La normalisation

- **Pourquoi faire ?**
  - Pour corriger les différences systématiques entre les mesures sur la même lame qui ne représentent pas de véritables variations biologiques.

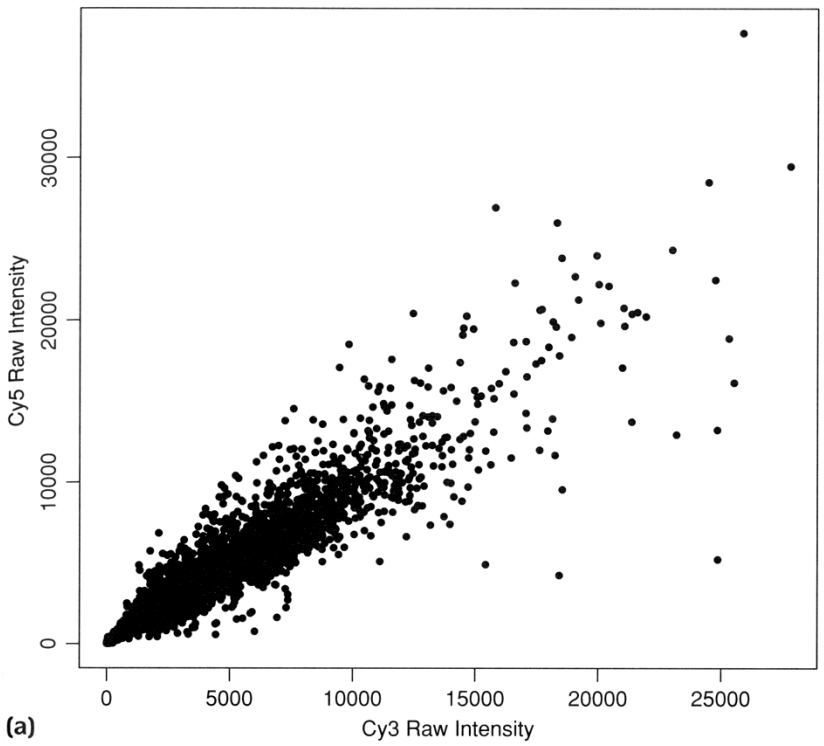
- **Pourquoi normaliser est nécessaire ?**
  - En examinant les réplicats contre le même échantillons, où de vraies différences d'expression ne doivent pas apparaître.



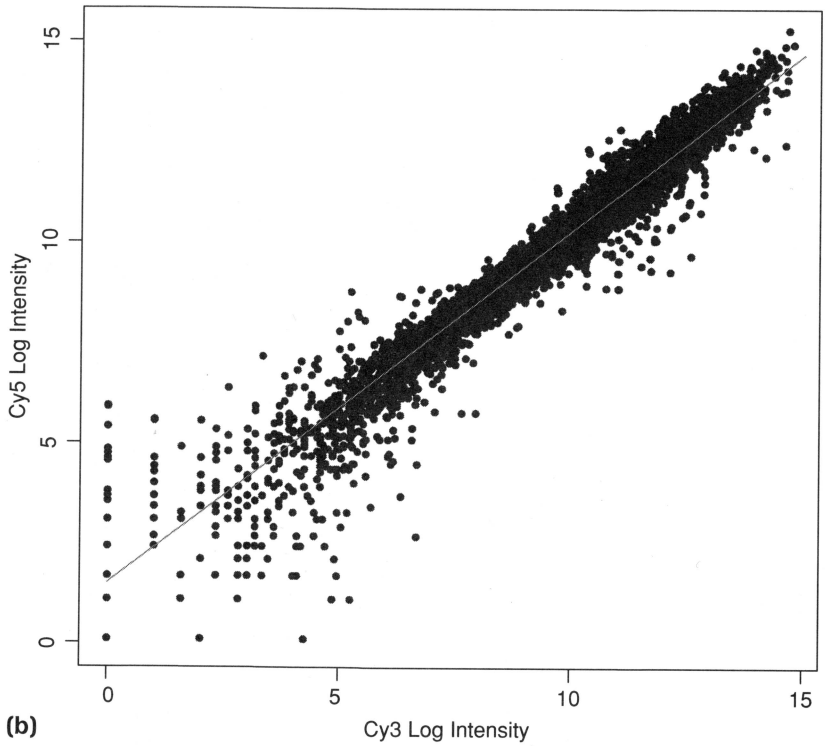
Extrémités des pointes de dépôt

- La normalisation des résultats des puces à ADN permet la comparaison de plusieurs expériences (référence commune)
- La normalisation calibre les erreurs systématiques (et non stochastique)
- Il est indispensable d'effectuer deux transformations mathématiques sur les résultats bruts avant de normaliser les données

La transformation logarithmique



Les effectifs sont plus important vers les faibles intensités

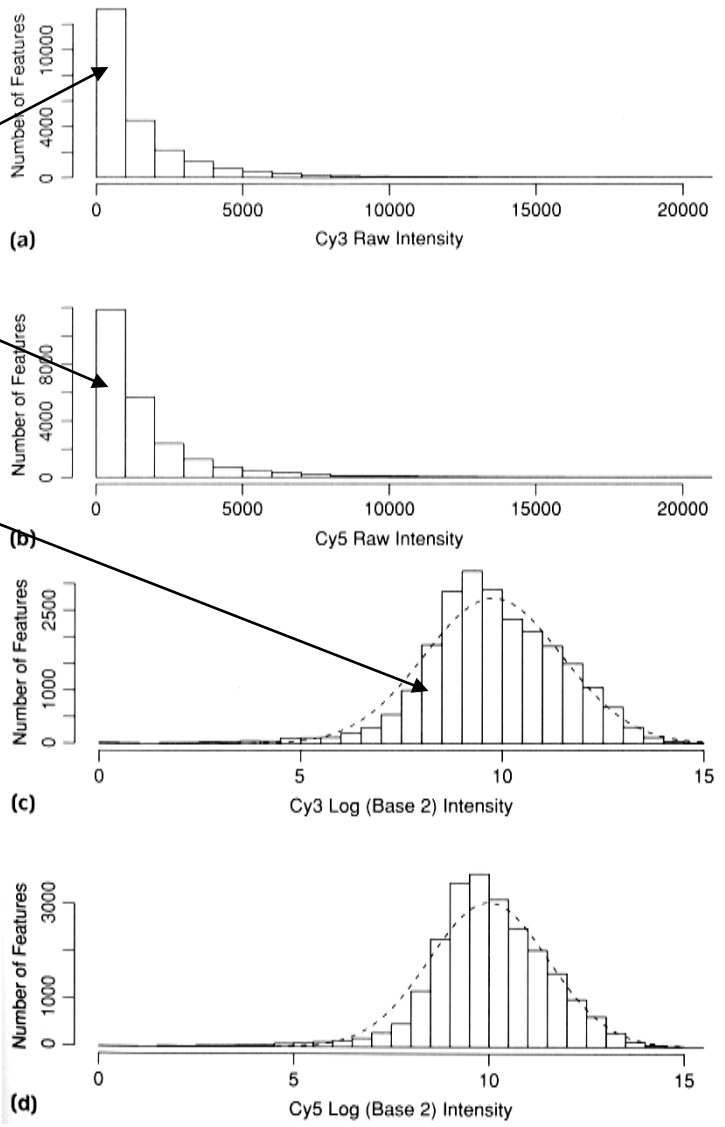


Les intensités sont distribuées de façon uniforme

Remarque : On choisira le logarithme en base 2 pour les analyses

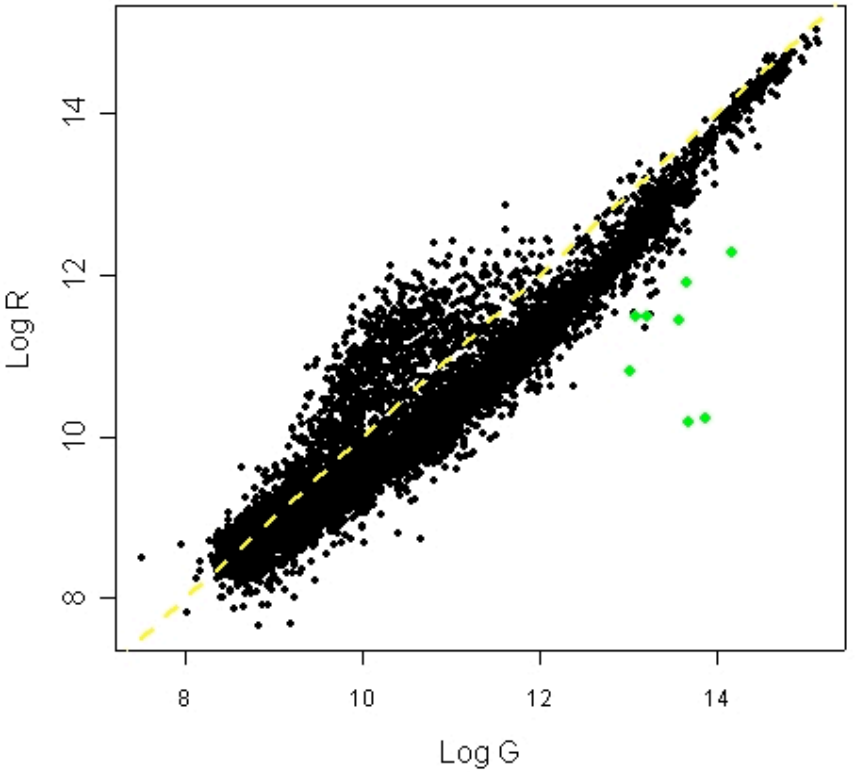
La transformation logarithmique

- **Effet sur la distribution des intensités**
  - La plupart des intensités mesurées sont faibles
  - Distribution « en cloche »
  - Recentrage de la distribution
  - Rend symétrique les distribution
  - Facilite l'utilisation des statistiques…

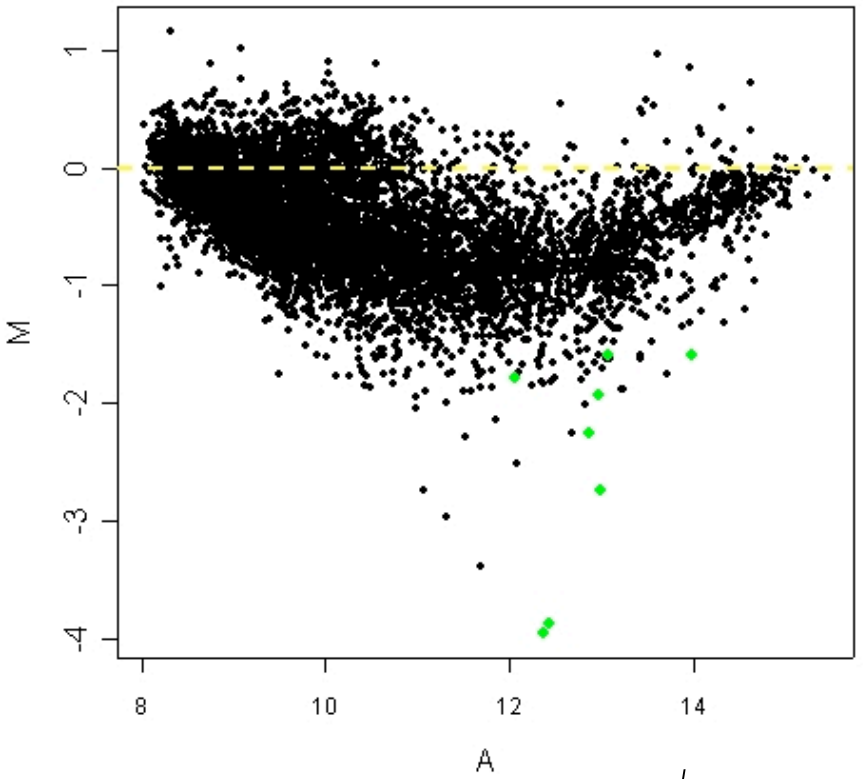


D'après « Microarray Bioinformatics », Dov Stekel

La rotation des graphiques (MA plot)



$\log_2 R$ vs $\log_2 G$



$M=\log_2 R/G$ vs $A=\log_2 \sqrt{RG}$

- **Différences des intensités :**

M = log ratio = log R/G = log R - log G

- **Contre la moyenne des intensités :**

A = log moyenne = log $\sqrt{RG}$ = [log R + log G]/2

Quels spots utiliser pour normaliser ?

- **Utilisation de contrôles positifs**
  - Prise en compte de spots contenant des gènes de ménages ou de l'ADN génomique
  - Les contrôles positifs doivent être détectables, posséder une expression stable et tomber dans la gammes de détection du scanner

Avantage : peu de gènes sont nécessaires

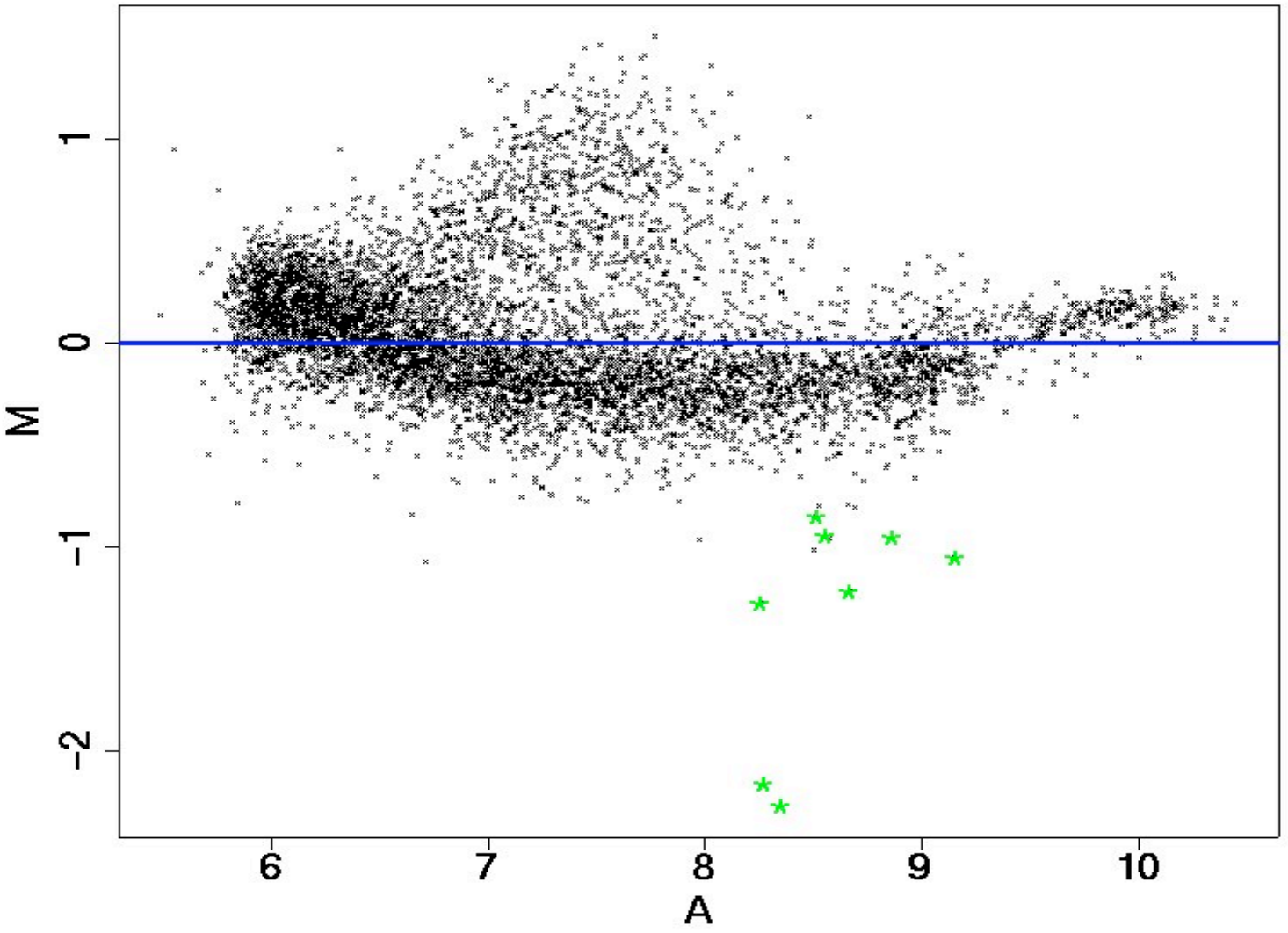
Inconvénient : ces gènes subissent trop de fluctuations non contrôlées dans les systèmes biologiques

- **Mesure de l'intensité globale**
  - Utilisation de l'intensité globale sur toute la membrane, mesurée pour tous les spots
  - La mesure de l'intensité globale doit s'effectuer sur un nombre suffisant de spots et doit utiliser des valeurs homogènes

Avantage : mesure efficace sur un grand nombre de spots

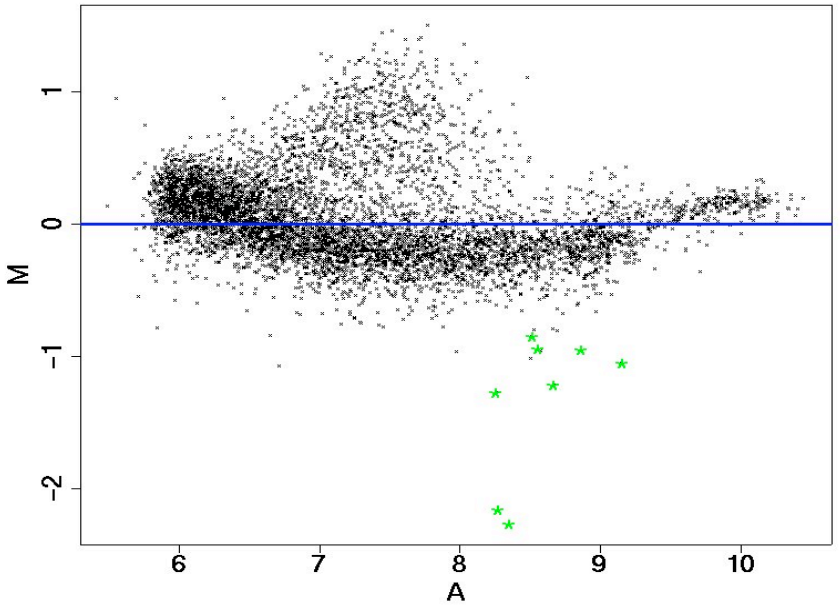
Inconvénient : il est nécessaire que la majorité des gènes analysés n'ait pas une expression modifiée

Les différentes méthodes de normalisation



Les différentes méthodes de normalisation

- **Normalisation basée sur un ajustement global**
  - Médiane ou moyenne des log de ratios pour un gène particulier ou un ensemble de gènes (gènes de ménage)
  - Normalisation en utilisant les intensités totales

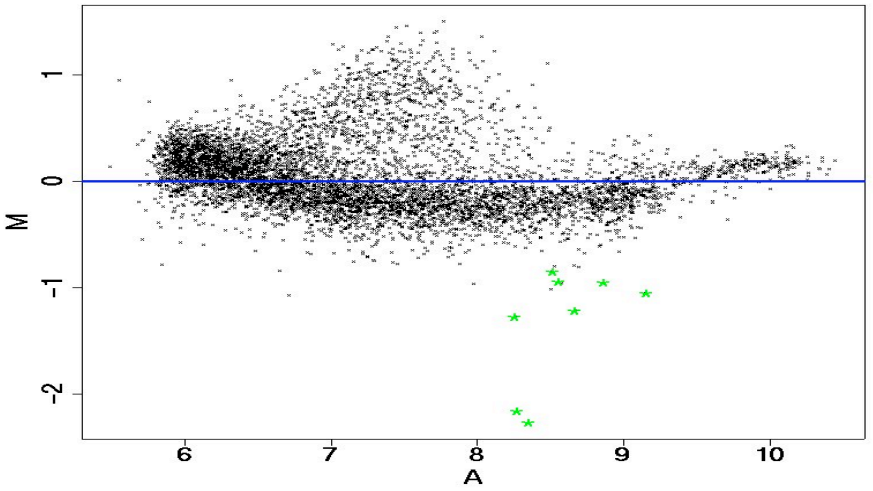


MA plot avec médianes ≈ 0

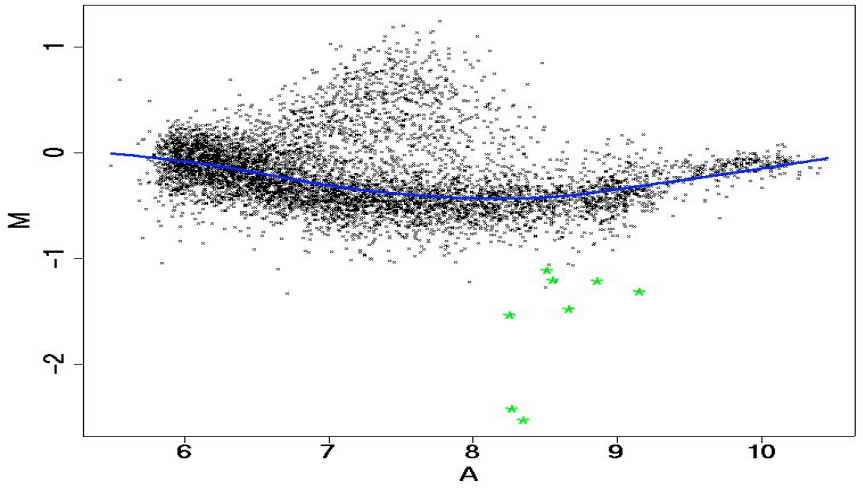
Les différentes méthodes de normalisation

- **Normalisation dépendante des intensités**
  - Utilisation d'une méthode de régression qui utilise la fonction lowess de Cleveland (1979) :

Lowess = LOcally WEighted Scatterplot Smoothing

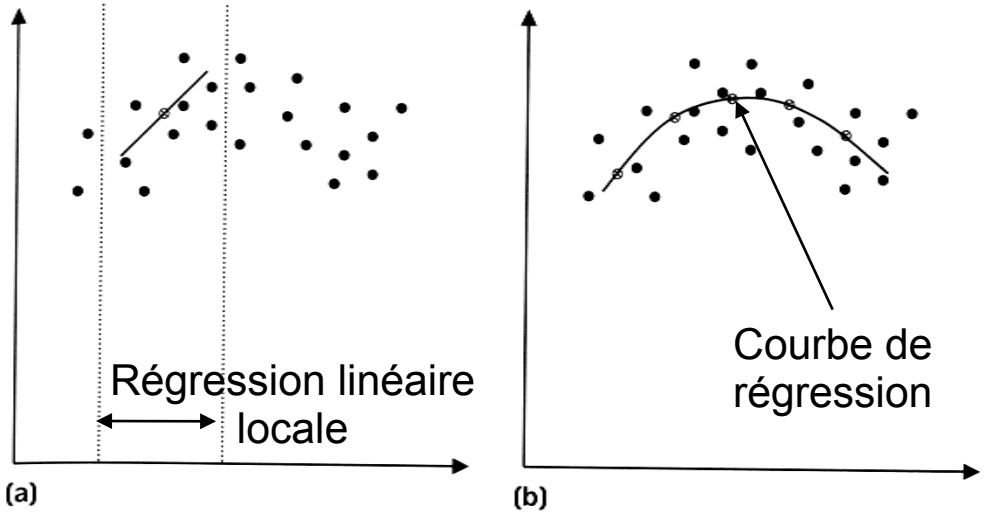


Normalisation sur les intensités globales



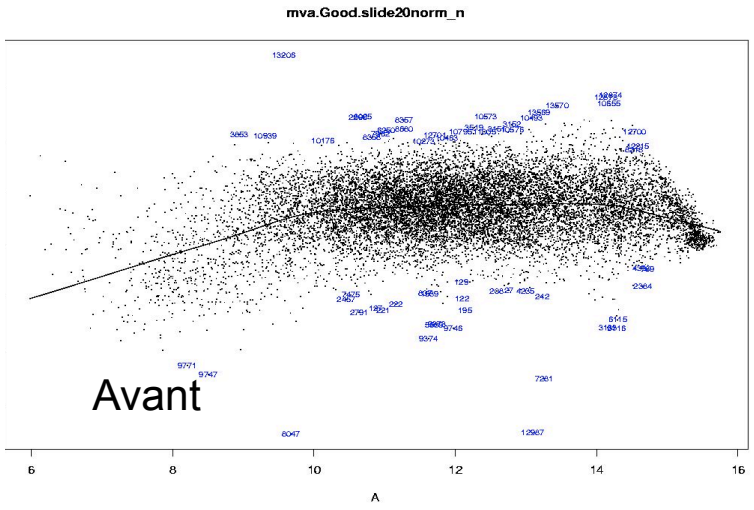
Normalisation dépendant des intensités

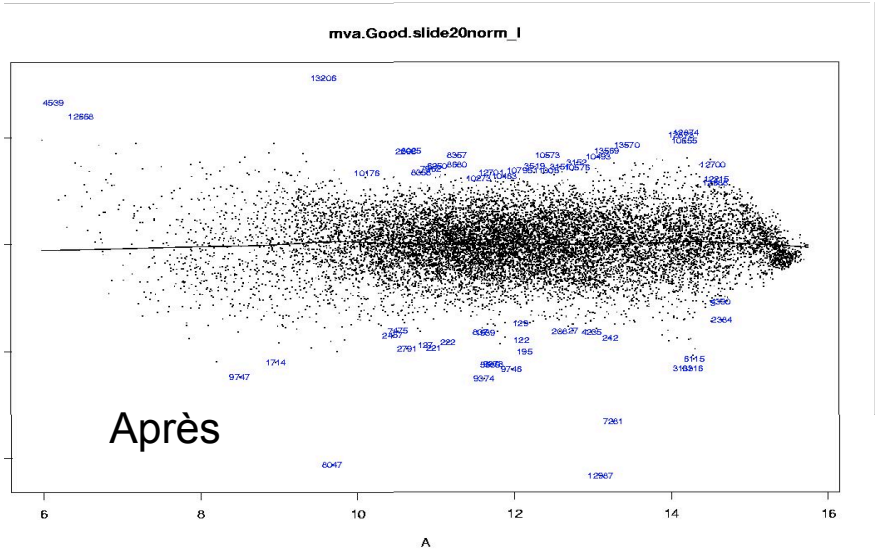
La normalisation par Lowess



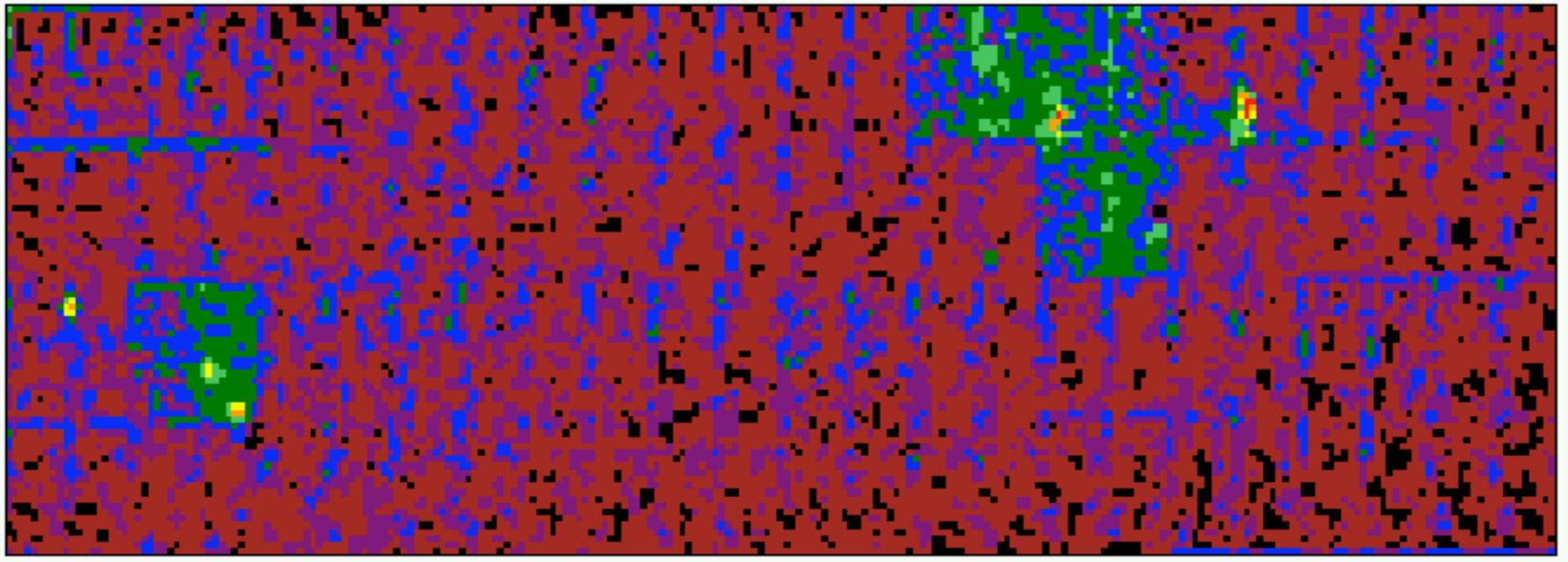
Les paramètres à prendre en compte :

- La taille des fenêtres
- Le chevauchement des fenêtres





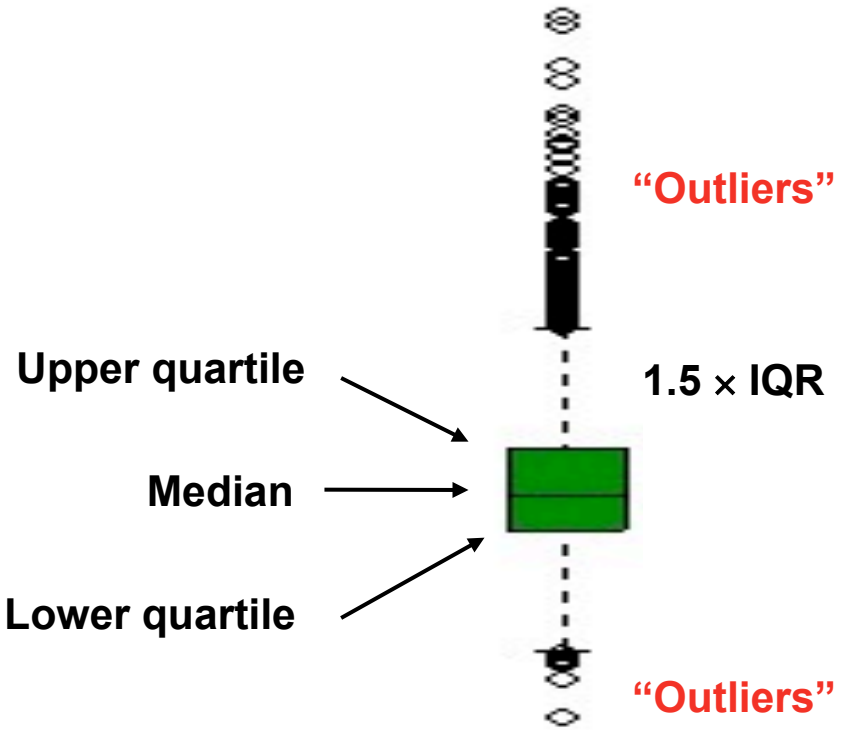
Comment s'affranchir des effets spatiaux ?

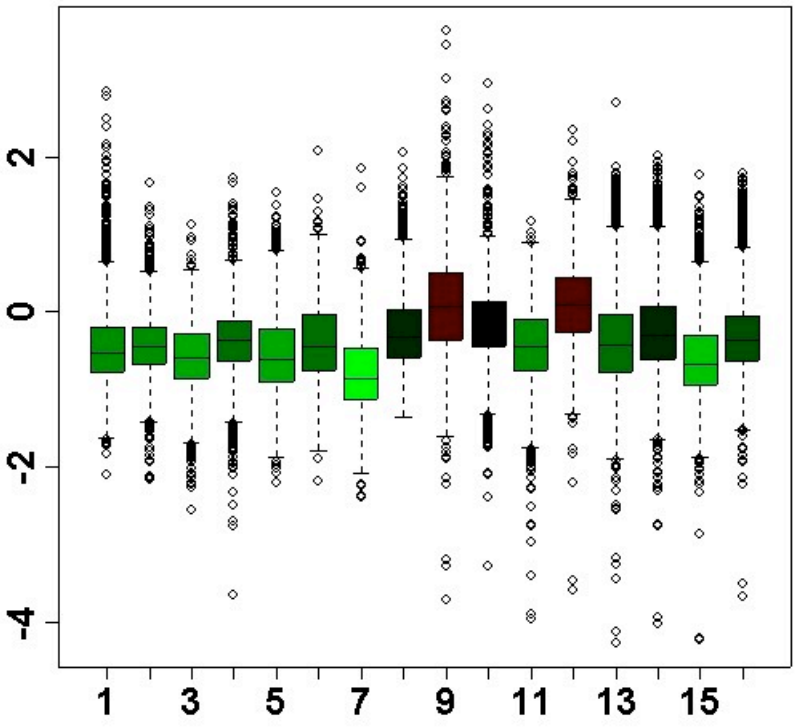


Effet de bloc ou de pointe de dépôt

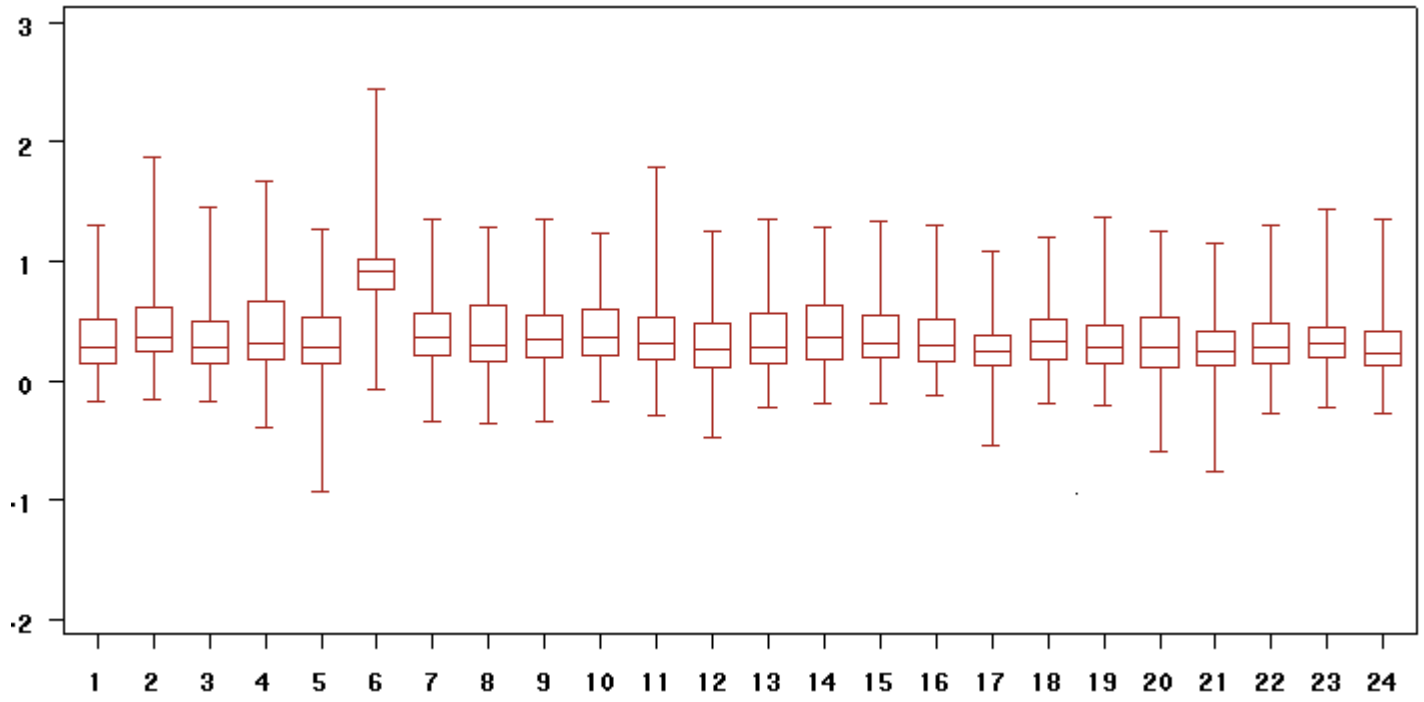
Box plots

- Chaque distribution des ratios (M) d'un bloc est représentée par une boîte
- On peut visualiser directement la forme globale de la distribution (moyenne, écart-type) et comparer facilement et rapidement les $\log_2$(ratios)





Visualisation de l'effet aiguille par box-plots



Les différentes méthodes de normalisation

- **Normalisation en prenant en compte les pointes de spotting**

  - En plus des variations dépendantes de l'intensité, un biais spatial peut aussi être une source importante d'erreur systématique.

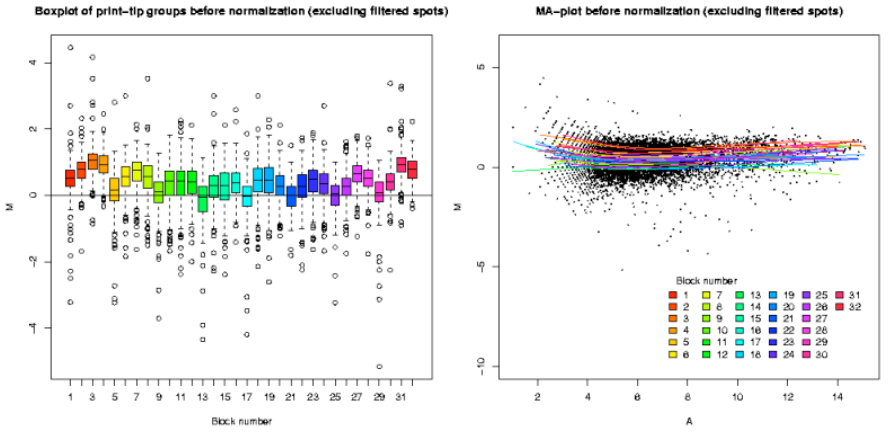
  - Peu de méthodes de normalisation corrigent les effets spatiaux qui produisent des artéfacts d'hybridation comme les pointes de spotting ou les différentes plaques lors de la production des lames.

  - Il est possible de corriger en même temps les biais dû à l'intensité et aux différentes pointes utilisées en effectuant une régression par Lowess sur les données à l'intérieur de chaque groupe de pointes soit :

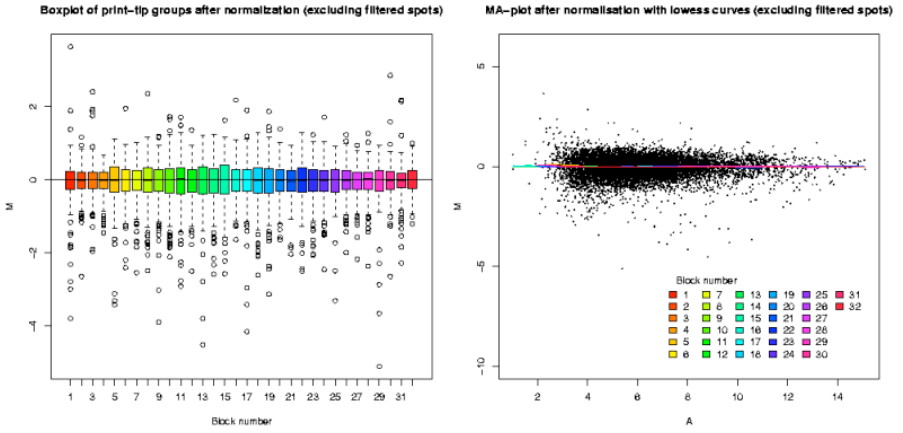
  $log_2 R/G \Rightarrow log_2 R/G - c_i(A) = log_2 R/(k_i(A)G)$

  où $c_i(A)$ est le coefficient de régression Lowess sur le MA plot pour la grille i.

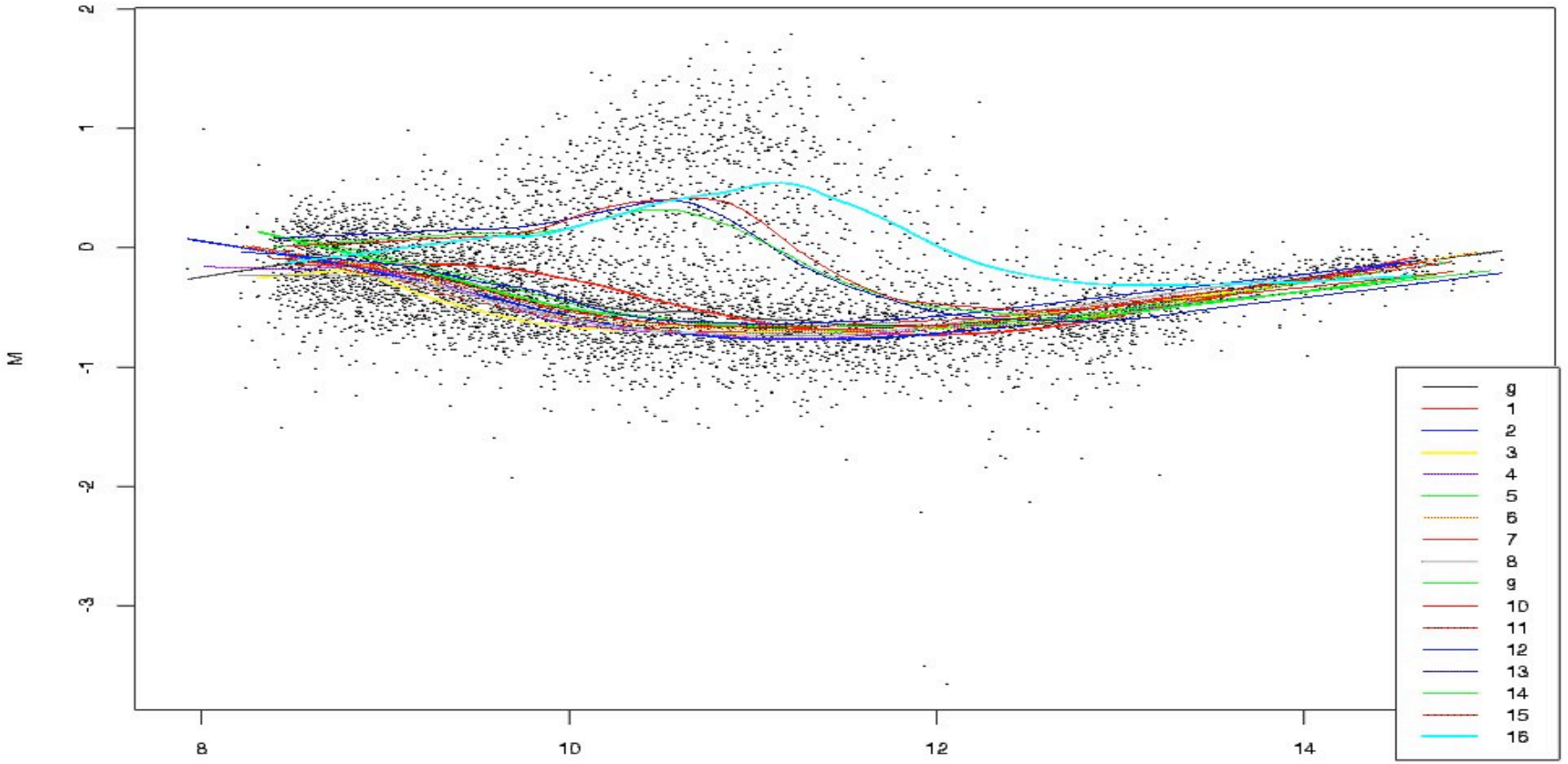
Avant normalisation

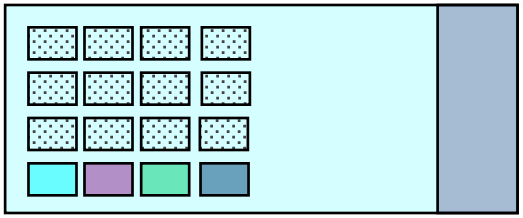


Après normalisation



Lowess par groupes de pointes





Lame de verre

Lames de verre avec ADNc

4x4 blocs = 16 groupes de pointes

Les différentes méthodes de normalisation

Avantages	Inconvénients
----- Régression linéaire -----	
• Méthode très simple • Alignement sur une horizontale	• Asymétrie entre Cy5 et Cy3 • Peu efficace sur les nuages déformés
----- Lowess -----	
• Marche bien sur les nuages déformés • Correction à la fois des intensités en Cy5 et Cy3	• Requiert l'utilisation de logiciels statistiques • Nécessite de faire attention aux paramètres de la régression
----- Lowess par bloc -----	
• Corrige les nuages déformés • Prend en compte les effets locaux	• Risque de sur-correction du signal • Il est nécessaire d'avoir assez de spots par bloc pour que la normalisation marche

La normalisation « deux couleurs »

- **Il y a quelques hypothèses à garder à l'esprit**
  - Utiliser la méthode Loess global implique qu'à l'échelle de l'abondance des ARNm :
    - seule une minorité des gènes est différentiellement exprimée
    - il y a un nombre égal de gènes différentiellement exprimés induits ou réprimés
  - Pour les méthodes spécifiques par bloc, il est nécessaire que les conditions précédentes soient respectées pour chaque bloc. D'un point de vue statistique, le nombre de spots concerné par la méthode ne doit pas être trop petit.
  - Utiliser un sous-ensemble de gènes spécifiques pour la normalisation (control, gènes de ménage) implique des hypothèses similaires.

- **Il y a des améliorations à apporter**
  - Utilisation d'une méthode de normalisation adaptée aux données utilisées.
  - Il est important de ne pas écraser les variation et de ne pas créer de faux positifs.

La normalisation « deux couleurs »

- **On recommande**
  - D'effectuer la transformation en log2 ratios (MA plot)
  - D'utiliser la normalisation Loess globale pour corriger le biais de fluorochrome
  - D'utiliser une normalisation par pointe (médiane) pour prendre en compte les biais spatiaux.
  - De garder à l'esprit que la normalisation change les données brutes : il est donc nécessaire d'adapter la méthode de normalisation aux données.
  - De garantir les mêmes conditions techniques pour toutes les lames qui seront utilisées dans vos expériences (même manipulateur, même scanner, même lot de lames …)
  - De ne pas hésiter à passer du temps sur les contrôles qualité (lames jaunes)

- **En principe**
  - Les spots mauvais seront éliminés en faisant plusieurs réplicats

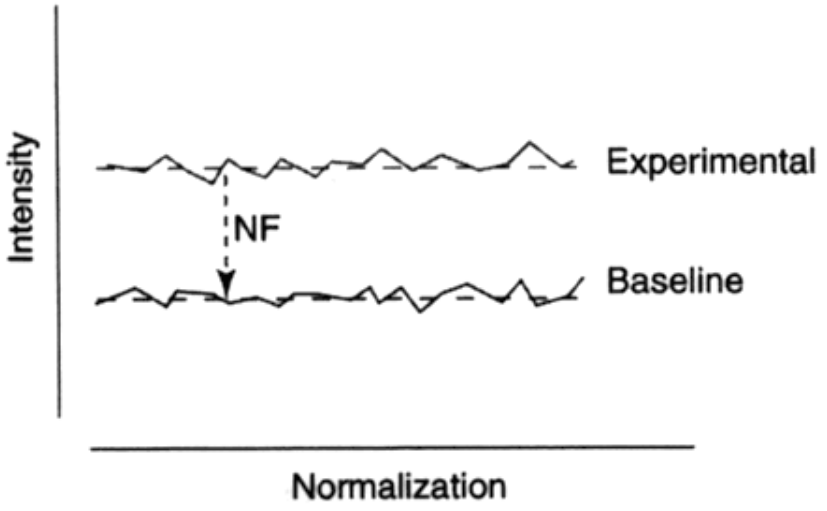
=> La chose la plus difficile c'est de corriger les biais techniques sans rien changer au signal étudié

Analyse bioinformatique des puces à ADN

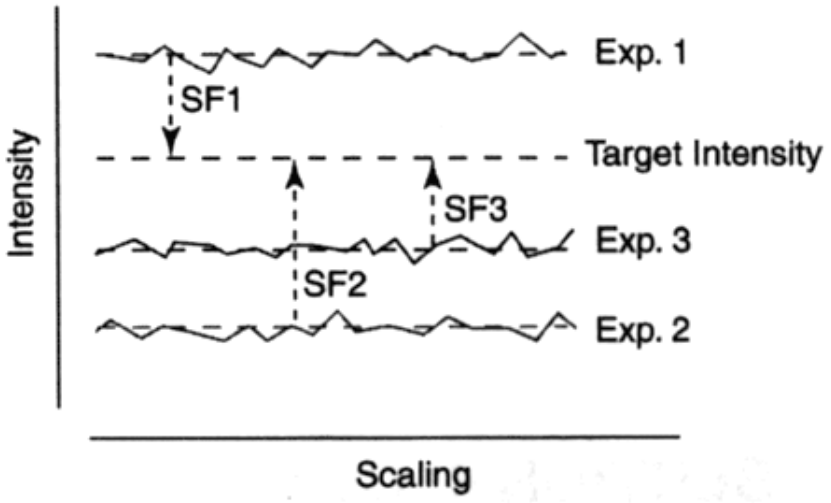
La normalisation entre lames

La normalisation avec les puces Affymetrix

- **Normalisation**
  - Comparaison entre deux expériences :
  - Chaque condition est hybridée sur une seule membrane/lame.
  - L'utilisation des intensités globales permet la normalisation de toutes les valeurs et donc la comparaison des deux expériences

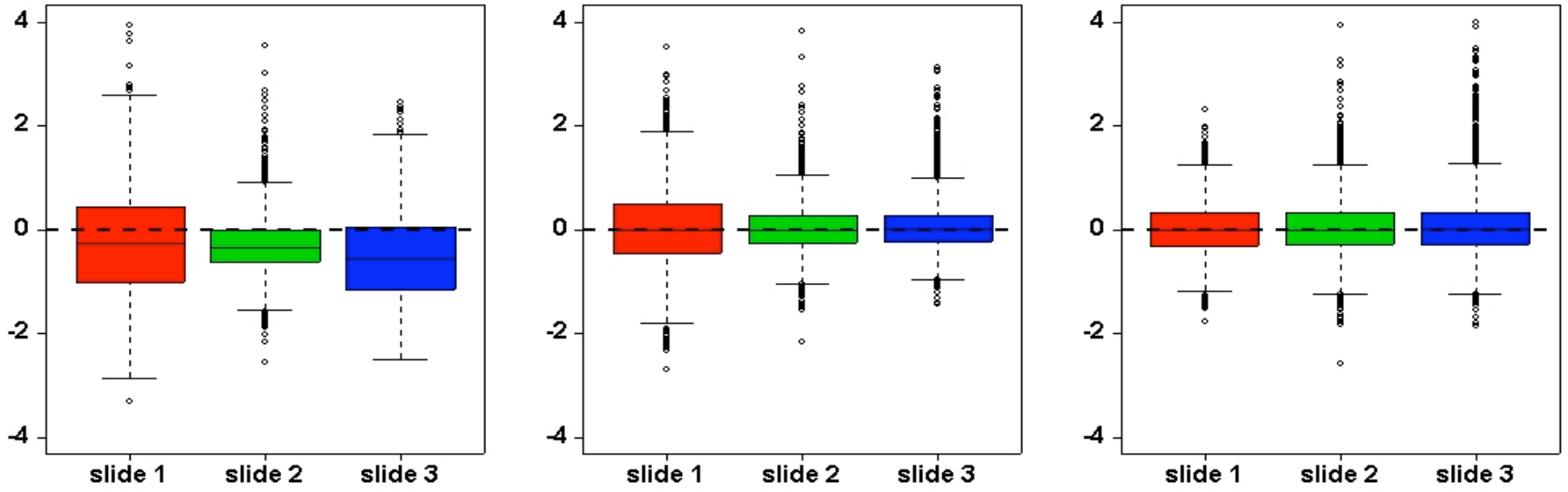


- **Standardisation**
  - Comparaison entre plusieurs expériences :
  - Toutes les expériences doivent être comparées à une condition contrôle.
  - Il est nécessaire d'utiliser une intensité cible, fixée arbitrairement ou obtenue de façon absolue (gènes de ménage).



La normalisation entre lames

- **Hypothèse : les variations des distributions observées ne sont pas des changements biologiques réels**



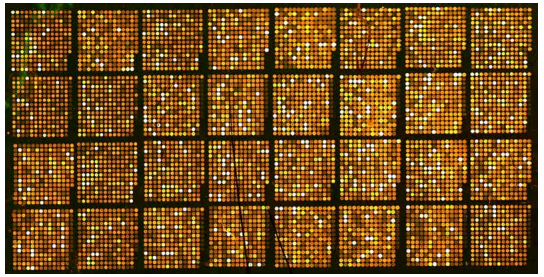
Box plot des distributions des $\log_2$(ratios) pour 3 hybridations identiques (réplicats) :

- Gauche : sans aucune normalisation
- Centre : après une normalisation Loess par bloc (centrage)
- Droite : après une normalisation entre lames (réduction)

Analyse bioinformatique des puces à ADN

La gestion des données

La gestion du flux de données



Images obtenues avec le scanner

Serveur de Fichiers

Publication Web

Base de données publiques

Internet

Intranet

Images

Données brutes

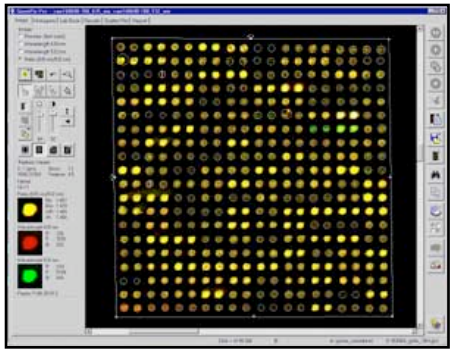
Données brutes

Données normalisées

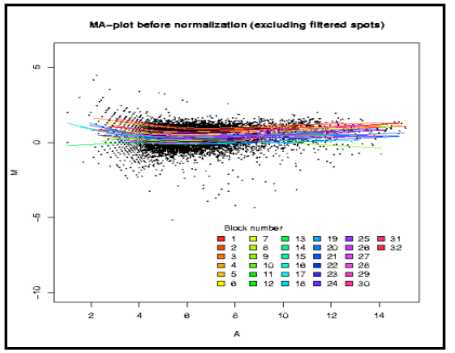
Données normalisées

Données publiées

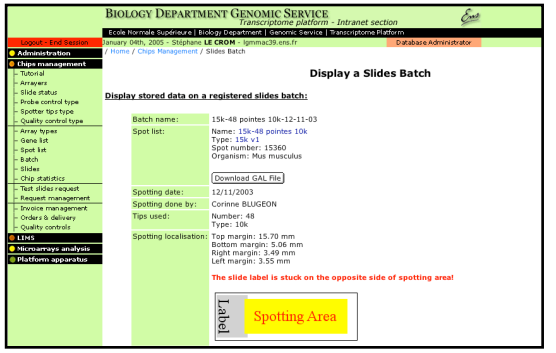
Analyse d'images



Normalisation



Interface Web



Les systèmes de stockage des données

- **Il y a trois niveaux différents de gestion des données**
  - Les dépôts de données publics

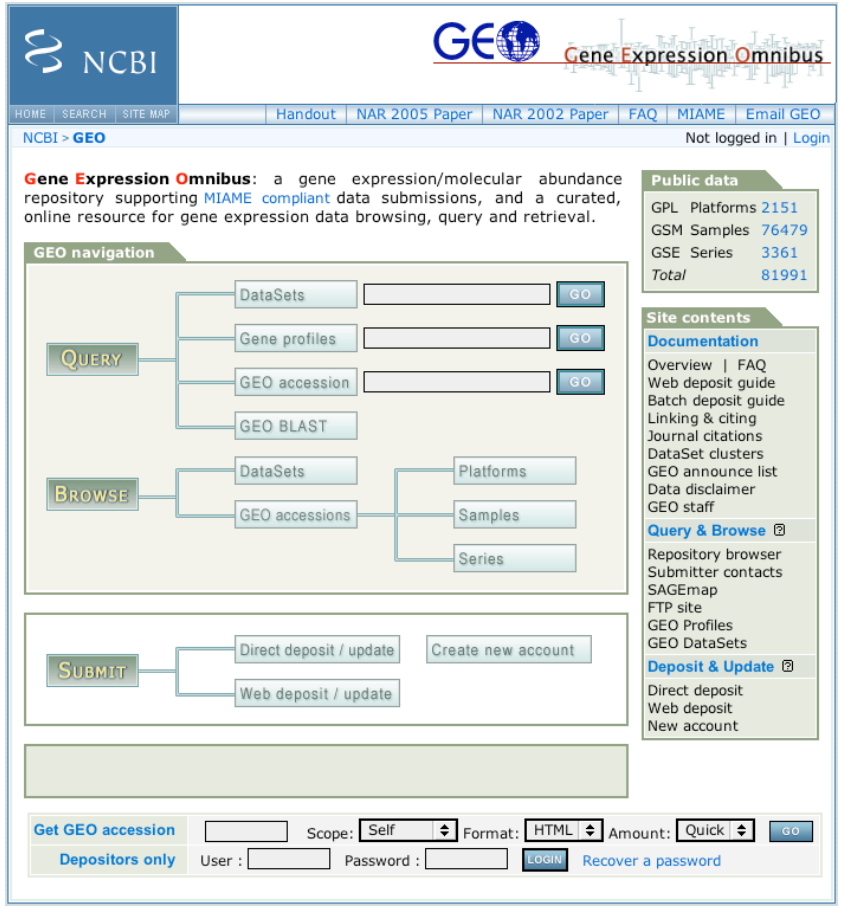
    Construits sur un schéma le plus flexible possible pour assurer le stockage de données hétérogènes comme les données provenant de différents organismes ou obtenues avec différents processus d'analyse

  - Les bases de données institutionnelles

    Construites afin d'aider un groupe d'utilisateurs sur une plate-forme technique dédiée ou pour répondre à un projet spécifique

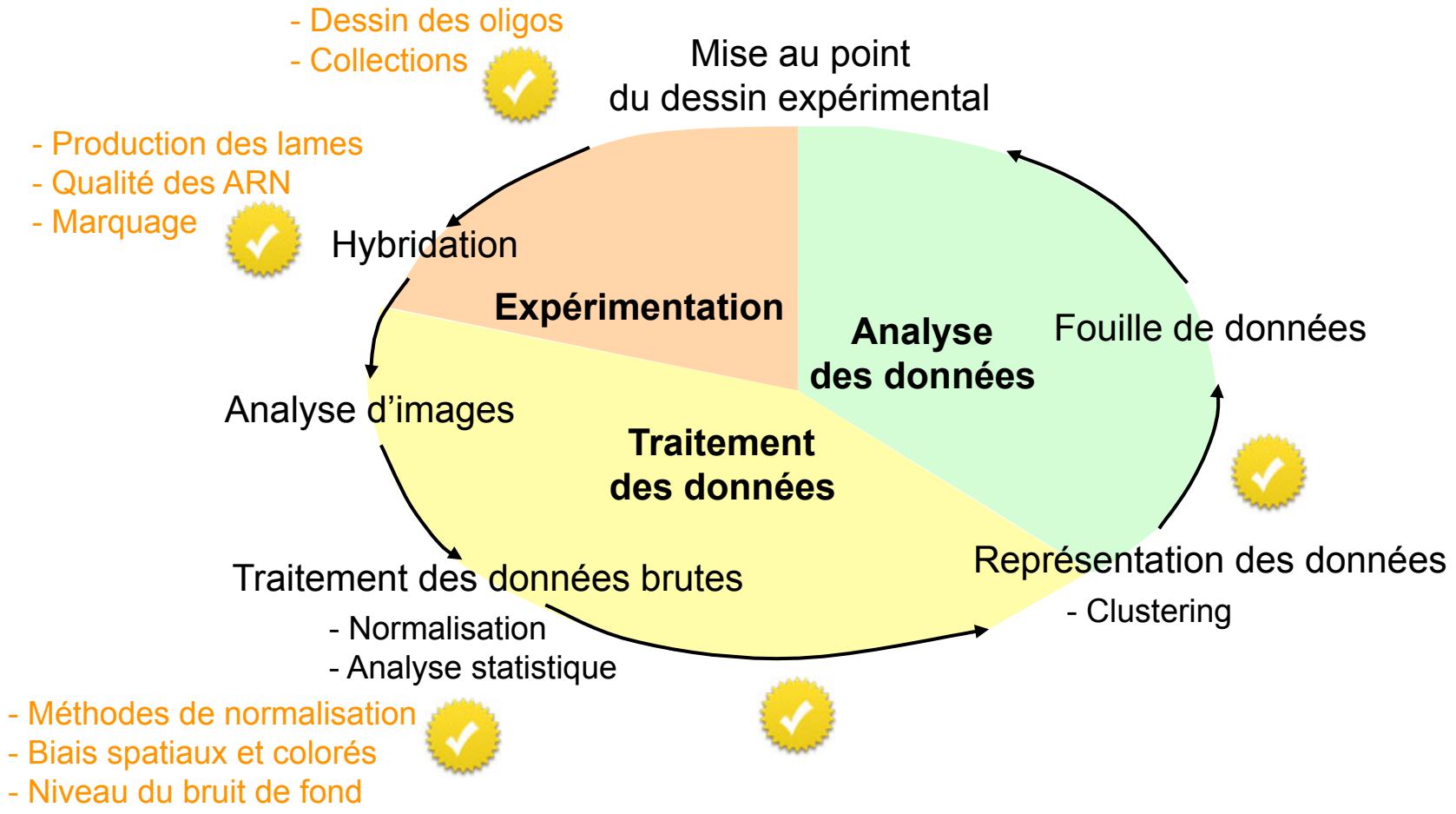
  - Les bases de données locales

    Construites et installées pour un petit nombre d'utilisateurs et pour répondre à des questions très spécifiques et précises



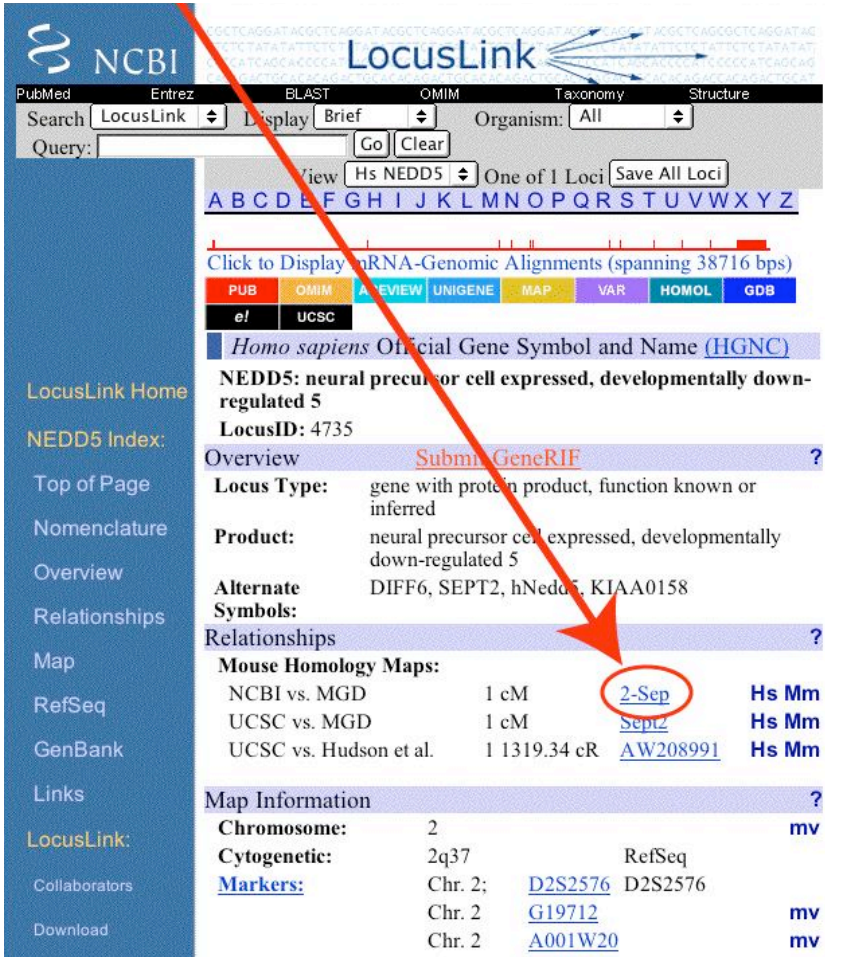
http://www.ncbi.nlm.nih.gov/geo/

Être attentif à la qualité des données



Excel introduit des erreurs dans le nom des gènes

- **La conversion automatique**
  - La conversion automatique par défaut des dates introduit des erreurs. Par exemple, le gène suppresseur de tumeurs DEC1 est converti en « 1-DEC » (premier décembre).
  - La conversion par défaut des nombres affecte les identifiants de clones de la forme nnnnnnnEnn, où n indique un chiffre. Par exemple, le clone RIKEN « 2310009E13 » est converti en nombre à virgule flottante « 2.31E+13 ». Une recherche a identifié plus de 2000 identifiants de ce type sur un total de 60770 clones RIKEN.
  - Ces conversions sont irréversible, le nom de gène original ne peut plus être retrouvé.



Zeeberg BR et al. Mistaken identifiers: gene name errors can be introduced inadvertently when using Excel in bioinformatics. *BMC Bioinformatics*. 2004 **5**:80.

Quelques références

• Normalisation

- Quackenbush J. Microarray data normalization and transformation. *Nat Genet.* 2002 **32** Suppl:496-501.
- Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* 2002 **30**(4):e15.
- Leung YF, Cavalieri D. Fundamentals of cDNA microarray data analysis. *Trends Genet.* 2003 **19**(11):649-59.

• Statistiques

- Saporta. Probabilités, analyse des données et statistiques. *Editions Technip*
- Daudin, Robin et Vuillet. Statistique inférentielle, idées, démarches, exemples. *Presses Universitaires de Rennes*
- Tassi. Méthodes statistiques. *Economica*