TD 3 Bioinformatique - M1 SPGF
Traitement bioinformatique de puces ADN
Corrections

Maupetit Julien

5 décembre 2007

1 Rappels de statistiques

1.1 Moyenne

$$\mu(x) = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{1}$$

1.2 Variance

$$V(x) = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu(x))^2 \tag{2}$$

1.3 Covariance

$$Cov(x,y) = \frac{1}{n}\sum_{i=1}^{n} (x_i - \mu(x)) * (y_i - \mu(y)) \tag{3}$$

1.4 Régression linéaire

$$y = a * x + b \tag{4}$$

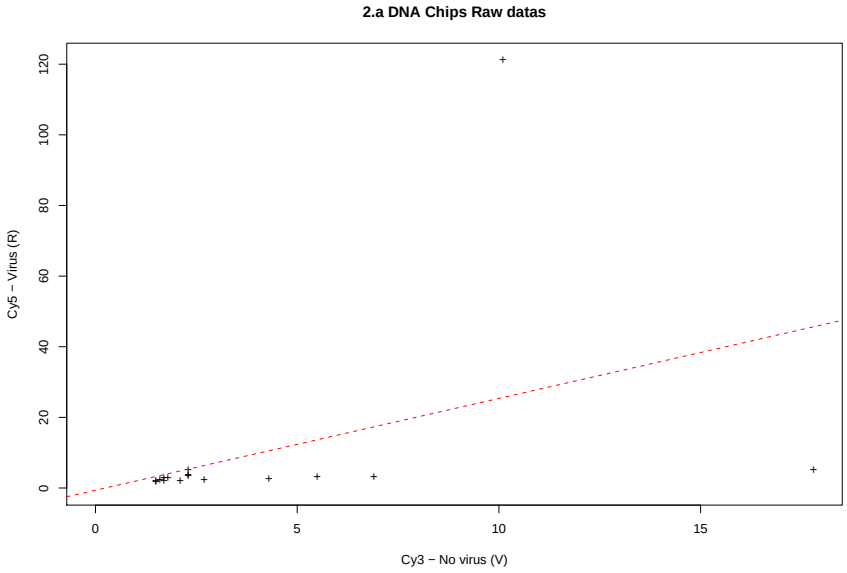
$$a = \frac{Cov(x,y)}{V(x)} \tag{5}$$

$$b = \mu(y) - a * \mu(x) \tag{6}$$

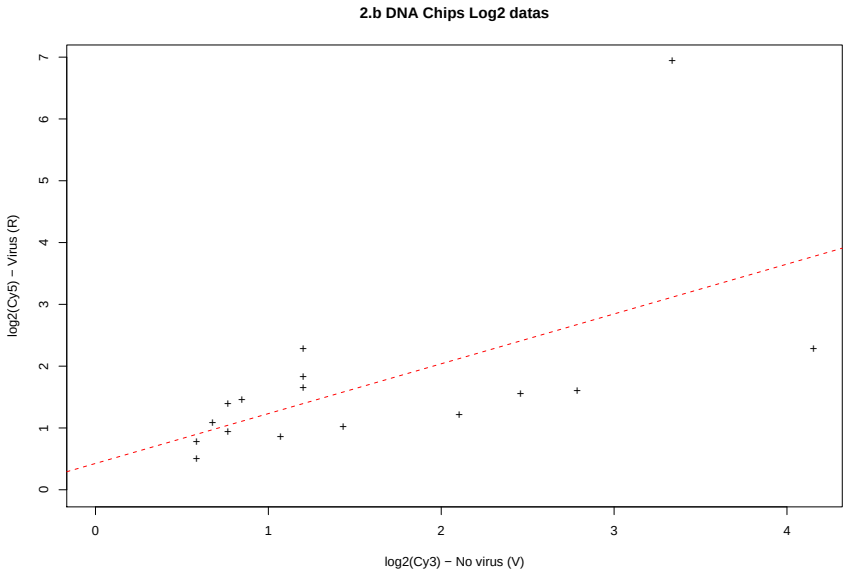
2 Représentation des données et normalisation intra-puce

	Cy3	Cy5	Log2(Cy3)	Log2(Cy5)	A	M	Mref	Mn
Gene 1,1	1.6	2.1	0.678	1.070	0.874	0.392	-0.174	0.567
Gene 1,2	1.5	1.4	0.585	0.485	0.535	-0.100	-0.306	0.207
Gene 1,3	6.9	3.0	2.787	1.585	2.186	-1.202	0.336	-1.538
Gene 1,4	17.8	4.8	4.154	2.263	3.208	-1.891	0.734	-2.625
Gene 2,1	5.5	2.9	2.459	1.536	1.998	-0.923	0.263	-1.186
Gene 2,2	2.3	4.8	1.202	2.263	1.732	1.061	0.159	0.902
Gene 2,3	2.3	3.1	1.202	1.632	1.417	0.431	0.037	0.394
Gene 2,4	2.7	2.0	1.433	1.000	1.216	-0.433	-0.041	-0.392
Gene 3,1	10.1	121.1	3.336	6.920	5.128	3.584	1.481	2.103
Gene 3,2	1.7	2.6	0.766	1.379	1.072	0.613	-0.098	0.710
Gene 3,3	4.3	2.3	2.104	1.202	1.653	-0.903	0.129	-1.031
Gene 3,4	1.7	1.9	0.766	0.926	0.846	0.160	-0.186	0.346
Gene 4,1	2.3	3.5	1.202	1.807	1.504	0.606	0.071	0.535
Gene 4,2	2.1	1.8	1.070	0.848	0.959	-0.222	-0.141	-0.081
Gene 4,3	1.8	2.7	0.848	1.433	1.140	0.585	-0.071	0.656
Gene 4,4	1.5	1.7	0.585	0.766	0.675	0.181	-0.252	0.432
M					1.634	0.121		
V					1.314	1.477		

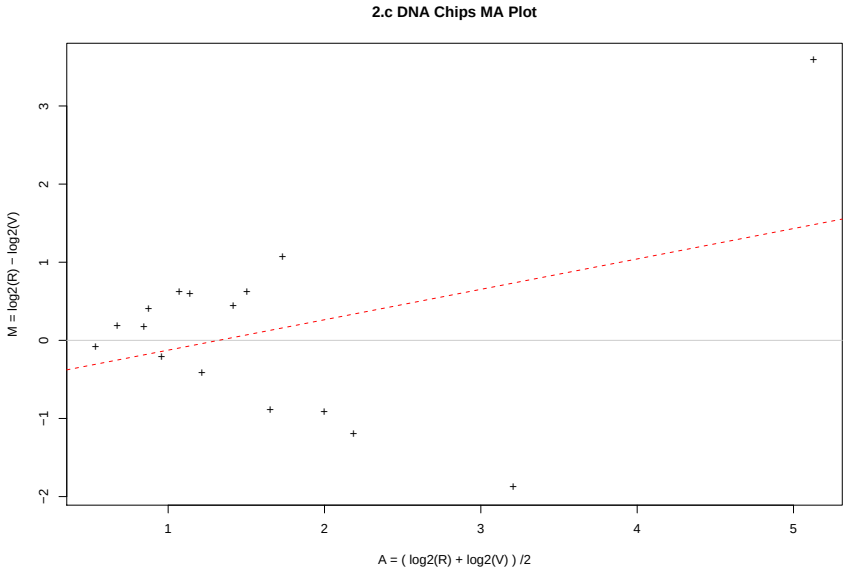
2.1 Cy5 vs Cy3



2.2 Log_2(Cy5) vs Log_2(Cy3)



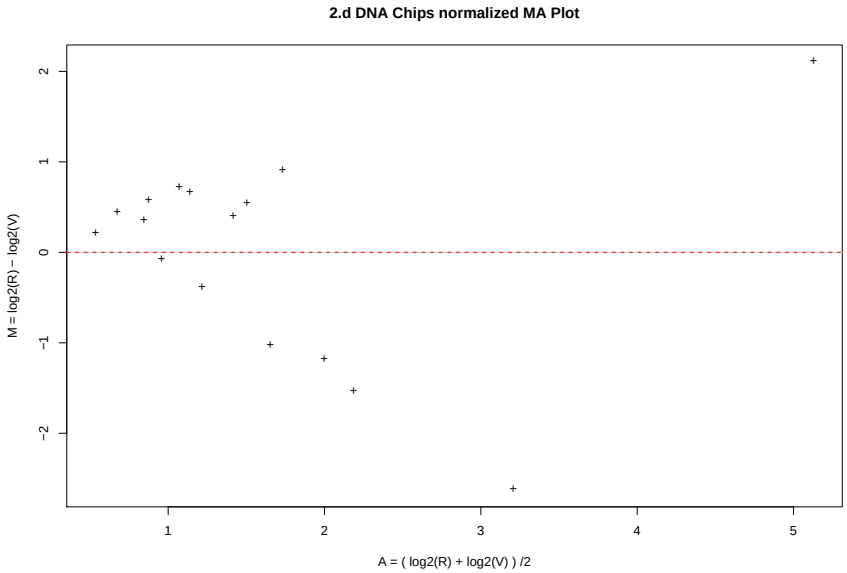
2.3 *MA plot*



- La covariance de (A,M) est 0.5114021
- L'équation de la droite de regression linéaire est

$$M = 0.389 * A + -0.515 \tag{7}$$

2.4 *MA plot*, données normalisées



Faire les boxplots avec les $\mu + / - \sigma$ et $1.5 * 2\sigma$. Normalement, premier quartile, mediane et dernier quartile. Pour déterminer les outliers $1,5 * (x_{0.25} - x_{0.75})$.

3 Normalisation inter-puces

	log2(R/V) A	log2(R/V) B	log2(R/V) C	$-\mu_A$	$-\mu_B$	$-\mu_C$	$/\sigma_A$	$/\sigma_B$	$/\sigma_C$
Gene 1,1	-0.400	1.400	1.000	-0.281	1.069	1.050	-0.229	1.425	1.382
Gene 1,2	0.000	0.000	0.100	0.119	-0.331	0.150	0.097	-0.442	0.197
Gene 1,3	1.200	0.800	1.500	1.319	0.469	1.550	1.075	0.625	2.040
Gene 1,4	1.900	-0.100	-1.100	2.019	-0.431	-1.050	1.646	-0.575	-1.382
Gene 2,1	1.000	0.600	0.400	1.119	0.269	0.450	0.912	0.358	0.592
Gene 2,2	-1.100	-0.300	-1.000	-0.981	-0.631	-0.950	-0.800	-0.842	-1.250
Gene 2,3	-0.400	0.800	-0.500	-0.281	0.469	-0.450	-0.229	0.625	-0.592
Gene 2,4	0.500	0.200	0.300	0.619	-0.131	0.350	0.504	-0.175	0.461
Gene 3,1	-3.600	0.700	-0.300	-3.481	0.369	-0.250	-2.838	0.492	-0.329
Gene 3,2	-0.700	0.500	-0.100	-0.581	0.169	-0.050	-0.474	0.225	-0.066
Gene 3,3	0.900	2.200	1.200	1.019	1.869	1.250	0.831	2.492	1.645
Gene 3,4	-0.100	-0.500	-0.600	0.019	-0.831	-0.550	0.015	-1.109	-0.724
Gene 4,1	-0.600	-0.600	-0.300	-0.481	-0.931	-0.250	-0.392	-1.242	-0.329
Gene 4,2	0.200	0.100	-0.300	0.319	-0.231	-0.250	0.260	-0.308	-0.329
Gene 4,3	-0.500	-0.500	-0.700	-0.381	-0.831	-0.650	-0.311	-1.109	-0.855
Gene 4,4	-0.200	0.000	-0.400	-0.081	-0.331	-0.350	-0.066	-0.442	-0.461
Mean	-0.119	0.331	-0.050						
sd	1.226	0.750	0.760						

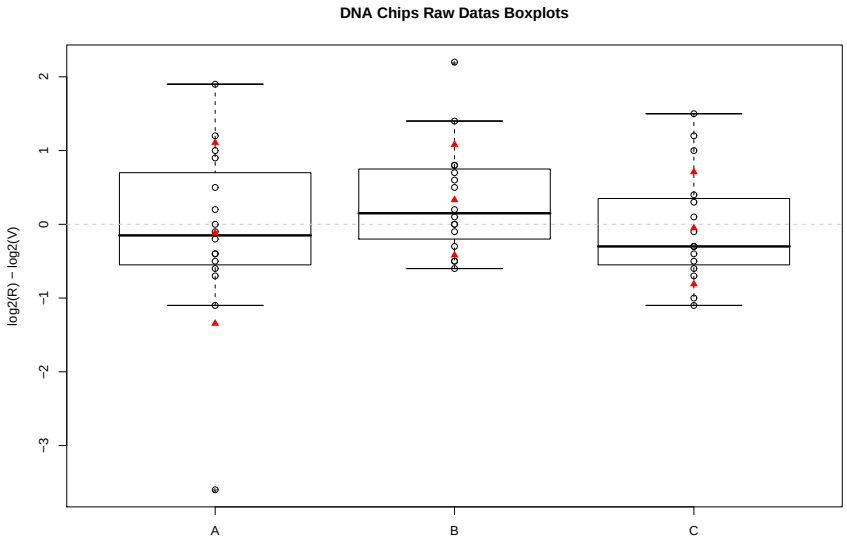
Nous en déduisons les valeurs des premiers et troisièmes quantiles en plus de la médiane.

Erratum : Nous ne pouvons avoir qu'une seule valeur de quantile par expérience, calculée sur les valeurs normalisées. Le tableau ci-dessous remplace celui de la version précédente de cette correction.

	A	**B**	**C**
$x_{0.25}$	-0.331	-0.642	-0.625
$x_{0.50}$	0.0255	-0.242	-0.329
$x_{0.75}$	0.586	0.525	0.494

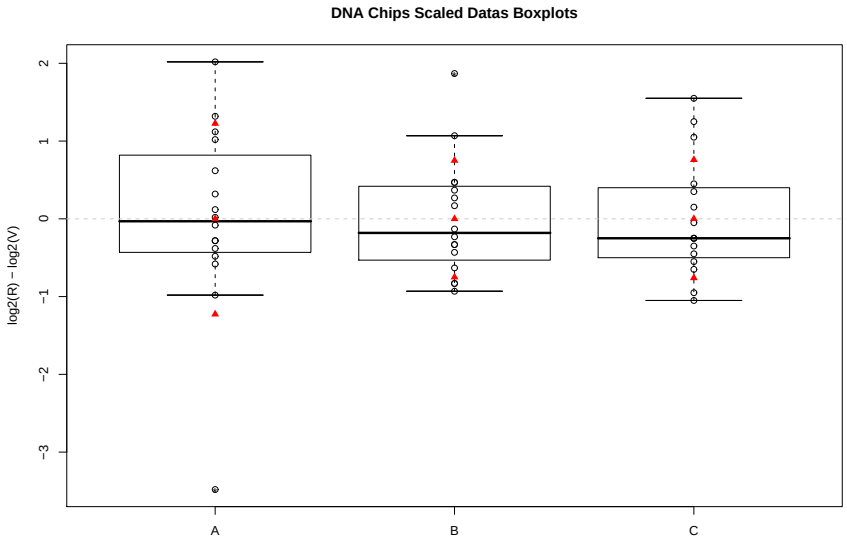
Nota bene : les valeurs décrites ci-dessus sont extrapolées par le logiciel R qui suppose que notre variable suit une loi normale. Vous pouvez déduire ces valeurs "à la main" en ordonnant les 16 valeurs pour chaque expérience, et en déterminant quelle valeur laisse 25, 50 et 75% de la distribution à droite.

3.1 *Box-plot*

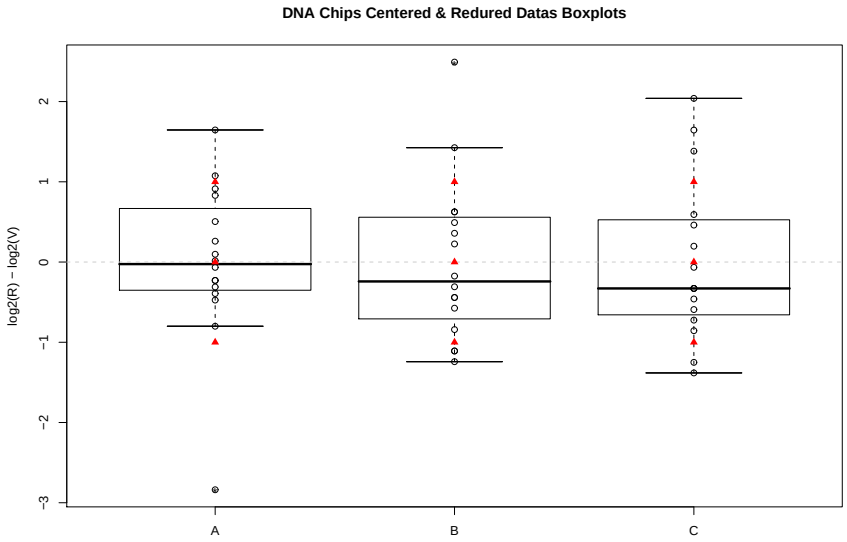


3.2 *Box-plot normalisé*

3.2.1 centré



3.2.2 centré réduit



4 Analyse des données

4.1 Test t

Dans ce cas précis, les données sont appariées (puces identiques). Nous allons donc ici tester si le $\log_2$ ratio diffère de 0. Nous réalisons donc un test de Student, la formule est dans ce cas :

$$t = \frac{\sqrt{(n)} * \mu}{\sigma} \tag{8}$$

Le nombre d de degrés de liberté est $n - 1$. n étant égal à 3, la valeur seuil pour que le test soit significatif au risque α = 5% est donc 4.303.

	μ	σ	t value	Expression
Gene 1,1	0.859	0.889	1.578	0
Gene 1,2	-0.049	0.118	-0.248	0
Gene 1,3	1.247	0.523	2.987	0
Gene 1,4	-0.104	2.459	-0.115	0
Gene 2,1	0.621	0.077	3.869	0
Gene 2,2	-0.964	0.062	-6.712	-
Gene 2,3	-0.065	0.391	-0.181	0
Gene 2,4	0.263	0.145	1.199	0
Gene 3,1	-0.892	3.010	-0.890	0
Gene 3,2	-0.105	0.123	-0.517	0
Gene 3,3	1.656	0.690	3.452	0
Gene 3,4	-0.606	0.326	-1.837	0
Gene 4,1	-0.654	0.260	-2.224	0
Gene 4,2	-0.126	0.112	-0.652	0
Gene 4,3	-0.758	0.166	-3.222	0
Gene 4,4	-0.323	0.049	-2.514	0

+ : sur expression - : sous expression 0 : inchangée

Le gène 2,2 est donc sous exprimé en réponse à l'infection.

4.2 Classification hiérarchique

Pour ceux qui ont le courage de faire la classification hiérarchique envoyez moi le résultat par courriel. Je mettrai ensuite la correction en ligne.