Régression non linéaire : application à l'analyse de la rythmicité de gènes circadiens

Diariétou Sambakhe

SAMBAKHE Diariétou

Master1 : Mathématiques et Applications

Spécialité : Statistique

Université de Strasbourg

Régression non linéaire: Application à l'analyse de la rythmicité de gènes circadiens

Stage réalisé à l'institut des Neurosciences Cellulaires et Intégratives (INCI)

Département de Neurobiologie des Rythmes

Responsable du diplôme : Mme Armelle GUILLOU

Maitre de stage : Mr André Malan

Sommaire

Remerciements

Je souhaite remercier le Dr Paul Pévet, responsable du département de Neurobiologie des rythmes, de m'avoir permis de réaliser mon stage au sein du laboratoire.

Je remercie spécialement mon maitre de stage, le Dr André Malan, pour tous les précieux conseils qu'il m'a apportés tout au long de mon stage dans le domaine des statistiques mais aussi en informatique et en biologie.

Je tiens également à remercier Cristina Sandou pour les explications biologiques qui m'ont permis de mieux cerner le domaine d'étude du laboratoire.

Je remercie toutes les personnes du laboratoire pour leur accueil.

Introduction

1) Présentation du laboratoire

a)INCI

L'Institut des Neurosciences Cellulaires et Intégratives est un institut de recherche fondamental en neurobiologie qui essaye de comprendre le fonctionnement des cellules nerveuses, des cellules neuroendocrines et des circuits neuronaux.

L'institut a une approche multidisciplinaire et caractérisée par différents niveaux d'investigations : génomique, protéomique, cellulaire, intégré et comportemental.

L'INCI est un laboratoire commun du CNRS et de l'Université de Strasbourg.

b) Département de neurobiologie des rythmes

L'institut se divise en trois grands départements :

- Neurobiologie des Rythmes,

- Neurotransmission et Secrétions Neuroendocrines,

- Nociception (perception de la douleur).

Pour mon stage, j'ai intégré le département de neurobiologie des rythmes.

Au sein de ce département, les différentes équipes cherchent à comprendre les mécanismes nerveux et neuroendocrines impliqués dans le contrôle des rythmes biologiques. Ces rythmes permettent à l'organisme de s'adapter aux variations journalières et saisonnières de l'environnement. Les recherches sont effectuées sur différents mammifères.

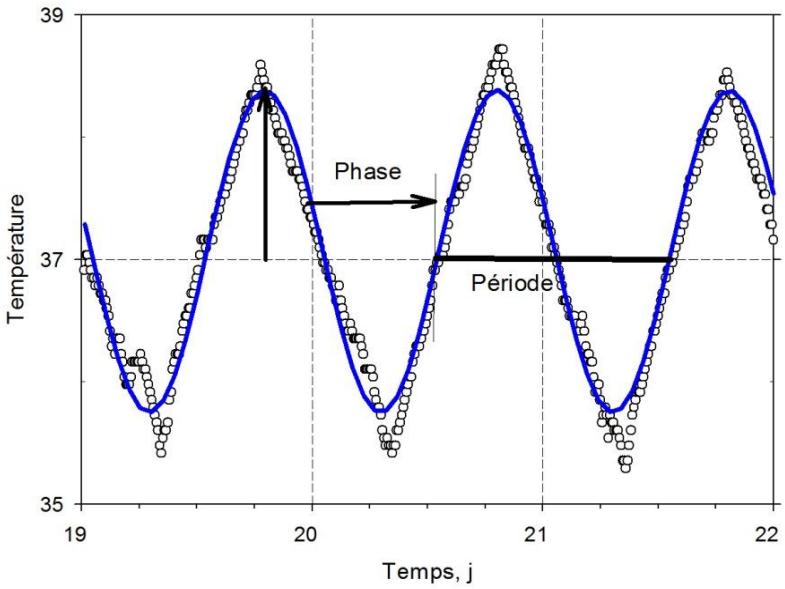
2) Quelques notions biologiques

a) Rythmes biologiques - caractérisation

Les rythmes biologiques permettent aux organismes d'anticiper et de s'adapter aux changements environnementaux pendant 24h. Ces rythmes biologiques circadiens sont générés par des horloges moléculaires.

Un rythme biologique (fig 1) se caractérise par sa période, l'emplacement de l'acrophase de la variation dans l'échelle de temps de la période, l'amplitude et le niveau moyen de la variation

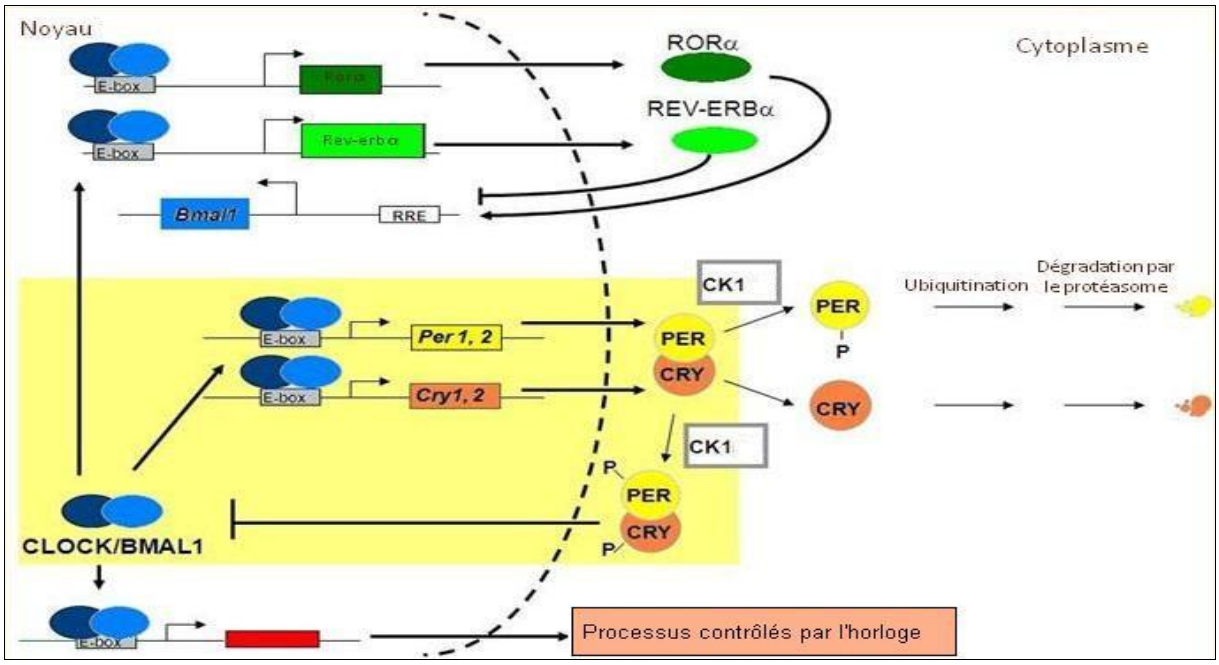
Fig 1 : rythme biologique



b) Horloge moléculaire

Une horloge moléculaire (fig 2) est composée de plusieurs gènes horloge interconnectés dans des boucles de rétrocontrôle transcriptionnel.

Fig 2 : horloge moléculaire



3) Présentation du sujet

Lors des précédentes études, il a été démontré l'existence d'une horloge moléculaire fonctionnelle dans une région particulière du cerveau(les noyaux suprachiasmatiques).

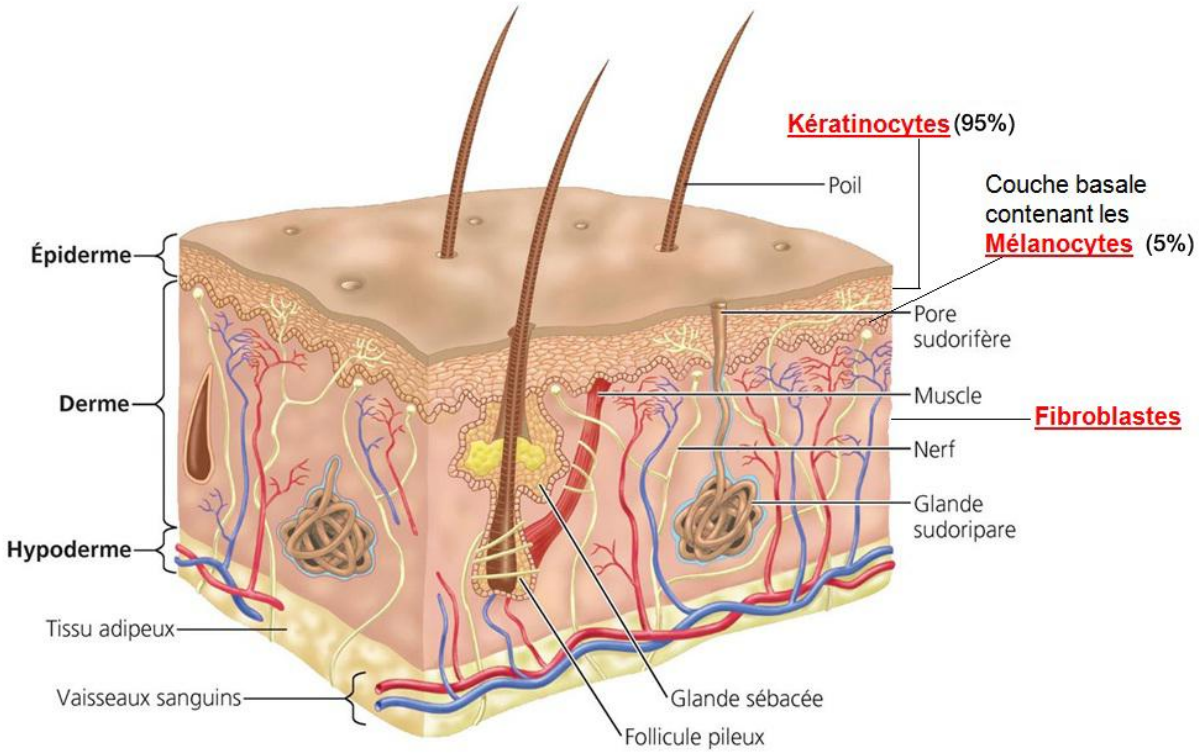
Cette horloge constitue l'horloge principale de l'organisme.

Récemment il a été démontré qu'a coté de cette horloge principale, il y a des horloges secondaires dans différentes organes (périphériques) comme l'œil, le foie, le cœur et le pancréas.

La peau est située à l'interface entre l'environnement et l'intérieur de l'organisme, elle est responsable de l'homéostasie de l'organisme et présente des fonctions rythmiques. La peau se compose de 2 couches, l'épiderme (qui contient des kératinocytes et mélanocytes) et le derme (qui contient des fibroblastes).

Chaque type cellulaire contient des gènes horloges.

Fig 3 : Représentation schématique de la structure de la peau



Le but de ce travail est de savoir si chaque type cellulaire de la peau humaine contient une horloge secondaire.

Pour cela les cellules (fibroblastes, kératinocytes, mélanocytes) sont isolées et cultivées *in vitro*.

On fait trois prélèvements toutes les 4 h pendant 56 h pour déterminer l'expression des différents gènes horloge: soit 3x14 ***tirages indépendants***.

Pour répondre a notre question, dans une première étape on va essayer de déterminer s'il y'a une variation d'expression pour chaque gène de chaque type cellulaire de la peau et si cette expression est rythmique.

Dans une deuxième partie, on va essayer de déterminer si les gènes qui s'expriment de façon rythmique pour chaque type cellulaire ont la même période.

Puis pour finir, pour chaque cellule, on va établir des différences de phases entre les gènes qui s'expriment de manière périodique avec la même période.

Afin de répondre au mieux à l'ensemble des problèmes posés, plusieurs logiciels statistiques ont été utilisés : le logiciel R, le logiciel Statistica, le logiciel Sigma plot et le tableur Excel.

I) Quelques notions pour la régression non linéaire

1) Principe de la régression non linéaire

L'objectif de la régression non linéaire est d'ajuster les valeurs des variables dans le modèle pour trouver la courbe qui prédit le mieux Y de X. Plus simplement, un modèle de régression non linéaire se présente sous la forme suivante :

$y_i = f(x_i, \theta) + \varepsilon_i \;\; i = 1, \ldots\ldots, n$ Où

Les Y_i sont les réponses, f une fonction non linéaire dépendant du vecteur $x_i = (x_{i1}, \ldots\ldots\ldots x_{ik})'$ et du paramètre $\Theta = (\theta_1, \ldots\ldots\ldots \theta_p)'$.

Les ε_i sont résidus et on fait l'hypothèse qu'ils suivent une loi normale centrée de variance σ^2.

Donc le but est de trouver les paramètres qui minimisent la somme des carrés résiduels

$SCR = \sum_{i=1}^{n} [y_i - f(x_i, \theta)]^2$.

Minimiser cette somme consiste à la dériver par rapport aux paramètres et à chercher les solutions annulant ses dérivés.

Dans le cas d'une régression linéaire, on aura un système de p équation linéaires à p inconnus simple à résoudre ce qui ne sera pas le cas dans une régression non linéaire où on aura un système de p équations non linéaires en Θ et qui ne peut pas se résoudre analytiquement.

Cependant, il existe plusieurs algorithmes itératifs pour résoudre ce problème, mais dans ce rapport on a choisit d'appliquer la méthode de Gauss – Newton.

2) Méthode numérique de Gauss – Newton

On sait que si la fonction f est différentiable autour d'un certain point θ^0 alors d'après le développement de Taylor elle peut s'écrire sous la forme :

$f(x_i, \Theta) \approx f(x_i, \theta^0) + \sum_{j=1}^{p} M_0^{ij} (\theta_j - \theta_j^0) = f(x_i, \theta^0) + M_0 . (\Theta - \theta^0)$

Où $M_0^{ij} \approx \frac{\partial f(X_i, \theta^0)}{\partial \theta_j}$.

Le principe de la recherche de l'estimateur $\hat{\theta}$ de Θ par la méthode de Gauss Newton est le suivant :

$\theta^1 = \theta^0 + (M_0' M_0)^{-1} M_0' e$

e=$\big((y_1 - f(x, \theta^0)), \ldots\ldots\ldots\ldots\ldots, (y_n - f(x_n, \theta^0))'$

Et le processus se répétera, c'est à dire qu'on procédera à une nouvelle itération avec θ^0 substitué par θ^1(et M_0 substitué parM_1). Le processus itératif continuera jusqu'à ce que la convergence voulue soit vérifiée.

3) Estimation de la matrice de covariance et construction d'intervalle de confiance

Dans le cas du modèle de régression linéaire

$Y = X\theta + \varepsilon$, ε variable aléatoire indépendante identiquement distribuée (IID) suivant une loi $N(0, \sigma^2 I)$.

Il est bien connu que la matrice de covariance du vecteur des estimateurs $\hat{\theta}$ de Θ obtenu par la méthode des moindres carrés ordinaires est :

$Var(\widehat{\theta})=\sigma^2 (X'X)^{-1}$

Pour un modèle de régression non linéaire, on ne peut pas en général obtenir une expression exacte de $Var(\hat{\theta})$ dans le cas d'un échantillon de taille finie. Nous obtiendrons toutefois un résultat asymptotique qui nous permettra d'établir que

$Var(\hat{\theta})= \sigma^2 (M'M)^{-1}$ où $M_{ij}=\frac{\partial f(x_i \theta)}{\partial \hat{\theta}_j}$ (1)

Dans la pratique, nous ne pourrions bien évidemment pas faire usage de (1) car σ^2 et Θ ne sont pas connus, il nous faut les estimer. Le seul moyen raisonnable d'estimer Θ est de prendre $\hat{\theta}$, mais il y a au moins deux façons d'estimer σ^2. Il en résulte deux façons d'estimer $Var(\hat{\theta})$. La première que l'on peut utiliser est

$\widehat{Var}(\hat{\theta})=\widehat{\sigma^2}\ (\widehat{M}'\widehat{M})^{-1})$,

Où $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}[y_i - f(x_i, \theta)]^2$, et l'autre est

$\widehat{Var}(\hat{\theta})=s^2\ (\widehat{M}'\widehat{M})^{-1})$,

Où $s^2=\frac{1}{n-p}\sum_{i=1}^{n}[y_i - f(x_i, \theta)]^2$.

Le premier de ces estimateurs fait l'usage de l'estimateur σ^2 du maximum de vraisemblance qui est biaisé. Nous allons donc prendre s^2 lorsque nous estimons la matrice de covariance pour les paramètres dans les modèles de régression non linéaire, et ce en dépit qu'il n'y ait aucune justification exacte dans le cas d'un échantillon fini pour pratiquer de la sorte.

En ce qui concerne les intervalles de confiance des paramètres, supposons par exemple que le paramètre qui nous intéresse soit θ_1 , et que l'écart type estimé de l'estimateur soit

$$\widehat{s(\theta_1)} \approx s^2((\hat{M}'\hat{M})_{11})^{-1/2}$$

Il nous faut tout d'abord connaître la longueur de notre intervalle de confiance en termes de l'écart type estimé $\widehat{s(\theta_1)}$ .Nous recherchons donc la valeur de α dans une table bilatérale de la distribution normale ou de la distribution de Student, ou la valeur α/2 dans une table unilatérale. Cela nous donne la valeur critique c_α , nous trouvons donc un intervalle de confiance approximé allant de

$\widehat{\theta_1}$-$c_\alpha \widehat{s(\theta_1)}$ à $\widehat{\theta_1}$+$c_\alpha \widehat{s(\theta_1)}$

Qui comprendra la vraie valeur de θ_1 dans (1-α)% des cas.

A l'évidence, nous faisons de fortes hypothèses lorsque nous construisons un intervalle de confiance de cette façon. Premièrement, il nous faut supposer que $\widehat{\theta_1}$ est normalement distribué. Deuxièmement, nous supposons que $\widehat{s(\theta_1)}$ est la véritable déviation moyenne de $\widehat{\theta_1}$.

II) Ajustement du modèle de régression non linéaire

L'objectif de cette partie est de déterminer pour chaque type cellulaire de la peau humaine les gènes qui ont une variation d'expression rythmique avant de déterminer les paramètres de la rythmicité.

Pour ce faire pour chaque gène de chaque type cellulaire, on va essayer d'ajuster à nos données le modèle de régression non linéaire suivant :

$$\boldsymbol{y_i = y0 + b \times x_i + c \times \cos\left(\frac{2\pi(x_i - 31 - ph)}{tau}\right) + \varepsilon_i,}$$

Les variables du modèle étant :

$\boldsymbol{y_i}$: représentant le niveau d'expression du gène

$\boldsymbol{x_i}$: représentant le temps en heure.

Les paramètres du modèle:

$y0$: L'ordonnée à l'origine

b : Une éventuelle dérive linéaire des oscillations

c : L'amplitude des oscillations

phi : Le déphasage

tau : La période des oscillations

On fait l'hypothèse suivante les résidus ε_i suivent une loi normale centrée de même variance.

Pour vérifier l'hypothèse de normalité des résidus on utilise le test de Shapiro-Wilk et la droite d'Henri.

Quant à l'hypothèse d'homogénéité des variances, on la vérifie à travers le test de Bartlett.

Pour l'estimation et la significativité des paramètres de notre modèle nous allons le faire grâce à la fonction nls du logiciel R

Le choix de ce modèle découle d'une étude précédente faite sur ces données.

Le procédé pour ajuster ce modèle à nos données pour les gènes des trois cellules étant le même, dans ce rapport on va faire l'étude détaillée de cette ajustement uniquement pour les données de trois gènes horloge des kératinocytes.

En ce qui concerne l'ajustement du modèle à nos données pour les autres gènes , on va simplement donner les résultats.

1) Analyse de la variance

On va d'abord établir les tableaux d'analyse de la variance pour vérifier si le modèle est globalement significatif au seuil α=0.05.

Pour les trois ajustements, nous avons les résultats suivants :

BMAL1

Source	DF	SS	MS	P (>|t|)
Modèle	4	7.176913	1.794228	<0.0001
Résidu	36	2.224	0.06177778	
Total	40	9.400913		

Cry1

Source	DF	SS	MS	Pr (>|t|)
Modèle	4	7.1911	1.7978	<0.0001
Résidu	36	1.8552	0.0515	
Total	40	9.0463		

Rev-erbα

Source	DF	SS	MS	Pr (>|t|)
Modèle	4	66.21809	16.5545	<0.0001
Résidu	34	15.633	0.4597	
Total	38	81.85109		

Dans les trois cas, le tableau d'analyse de la variance de notre modèle nous montre que le test est globalement significatif au seuil α =5%

2) Vérification des hypothèses faites

Afin de dire si un modèle représente bien des données, il est nécessaire de s'assurer que les hypothèses sous-jacentes à l'ajustement du modèle sont respectées.

Cette réponse se fonde sur l'examen des résidus.

Les résidus sont les écarts entre valeurs observées et valeurs prédites par le modèle.

On les obtient dans R grâce à la fonction residuals ()

a) Normalité

Pour les trois cas, on utilise le test de Shapiro-Wilk pour vérifier la normalité des résidus

Bmal1

Shapiro. test(residus)

Shapiro-Wilk normality test

data: residus

W = 0.9786, p-value = 0.624

Cry1

Shapiro. test(residus)

Shapiro-Wilk normality test

data: residus

W = 0.985, p-value = 0.8572

Rev-erb α

Shapiro. test(residus)

Shapiro-Wilk normality test

data: residus

W = 0.9814, p-value = 0.7421

D'après les tests ci-dessus on accepte la normalité des résidus dans les trois cas au seuil α=0.05. En effet dans chacun des trois cas on a la p-value du test de Shapiro qui est supérieure à 0.05

b) Homogénéité des variances

On utilise le test de Bartlett pour vérifier l'homogénéité de la variance des résidus.

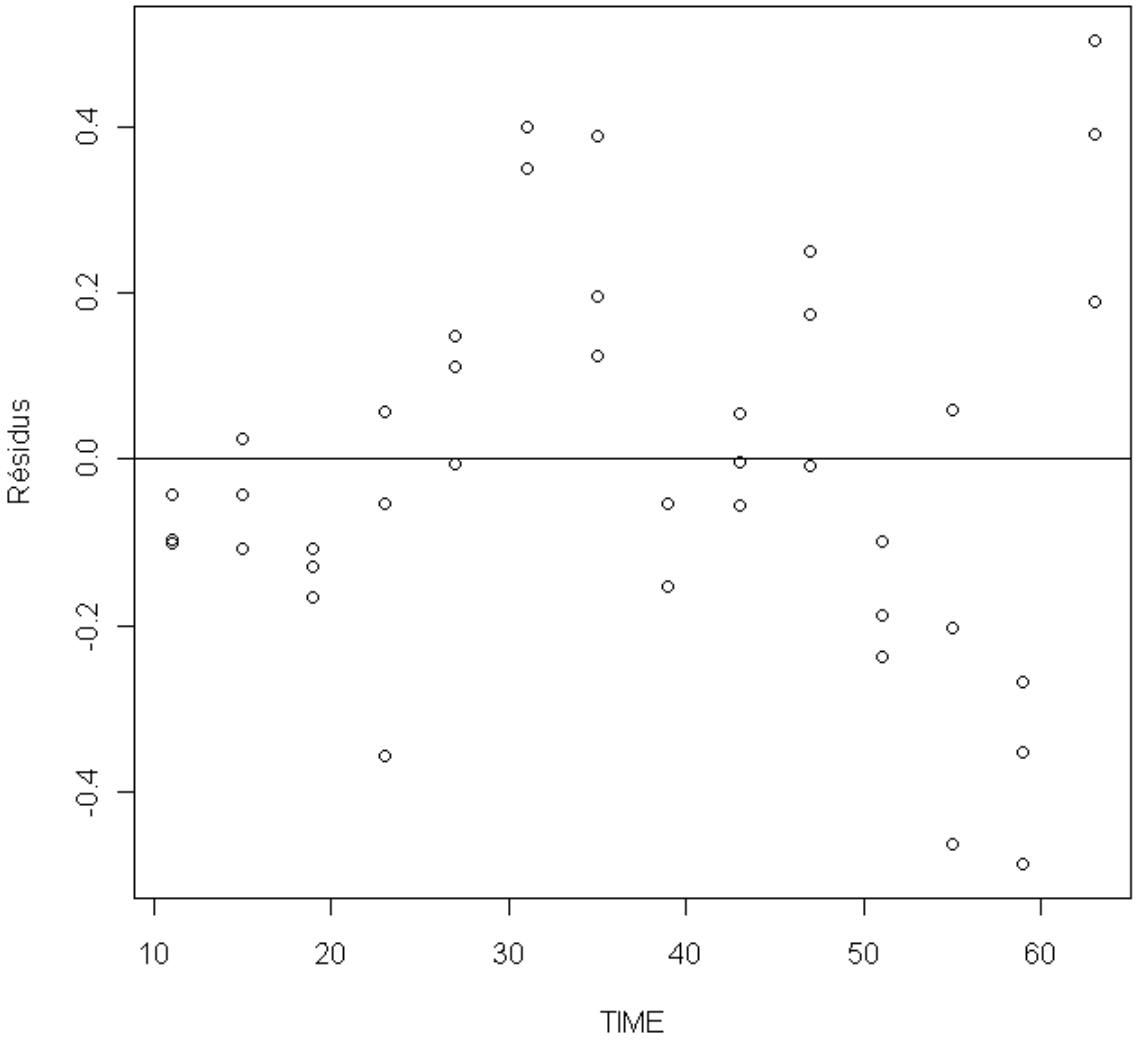
Bmal1

bartlett.test(residus)

Bartlett test of homogeneity of variances

data: residus and donnee1$time

Bartlett's K-squared = 18.5091, df = 13, p-value = 0.1391



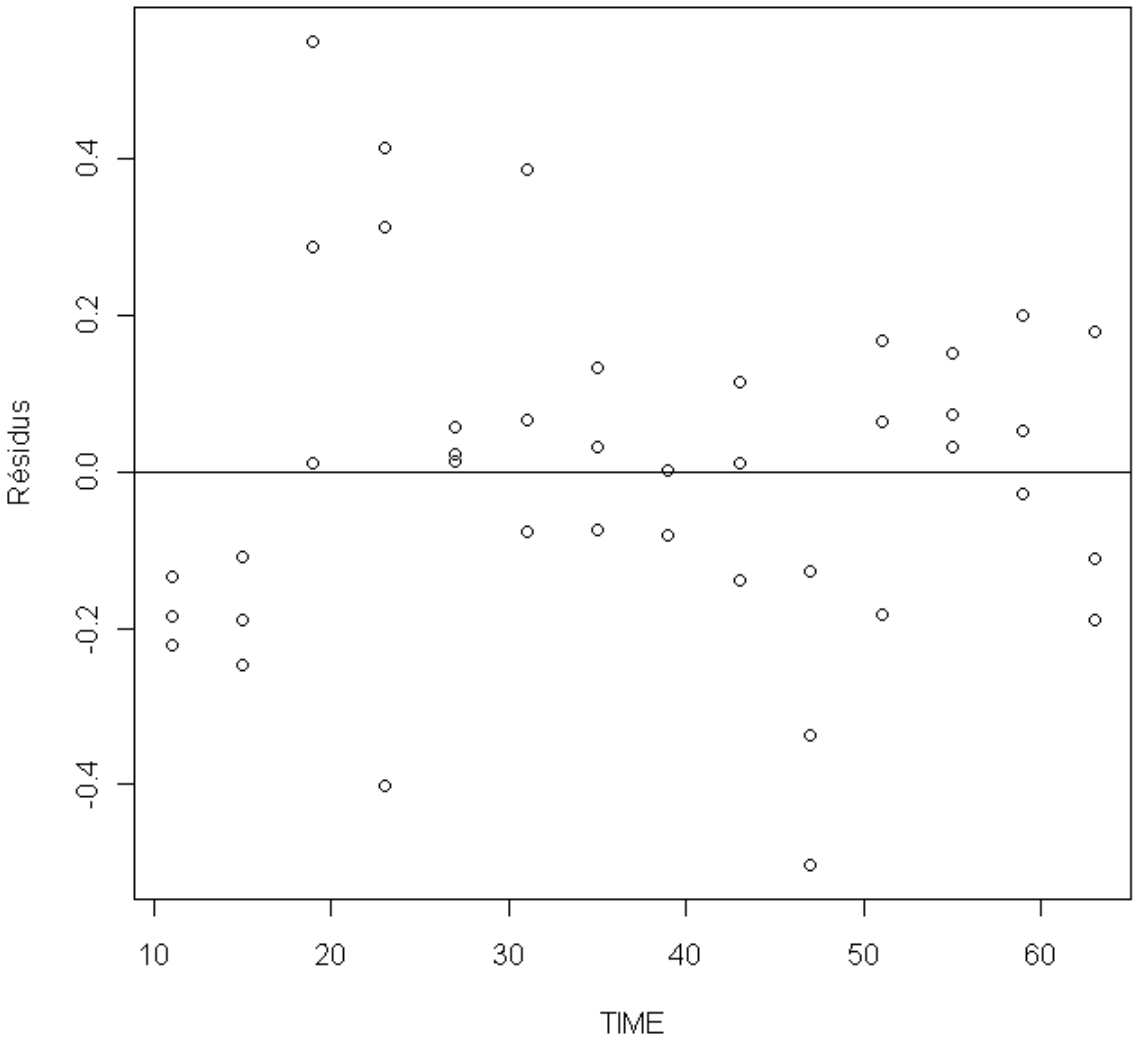
Cry1

bartlett.test(residus)

Bartlett test of homogeneity of variances

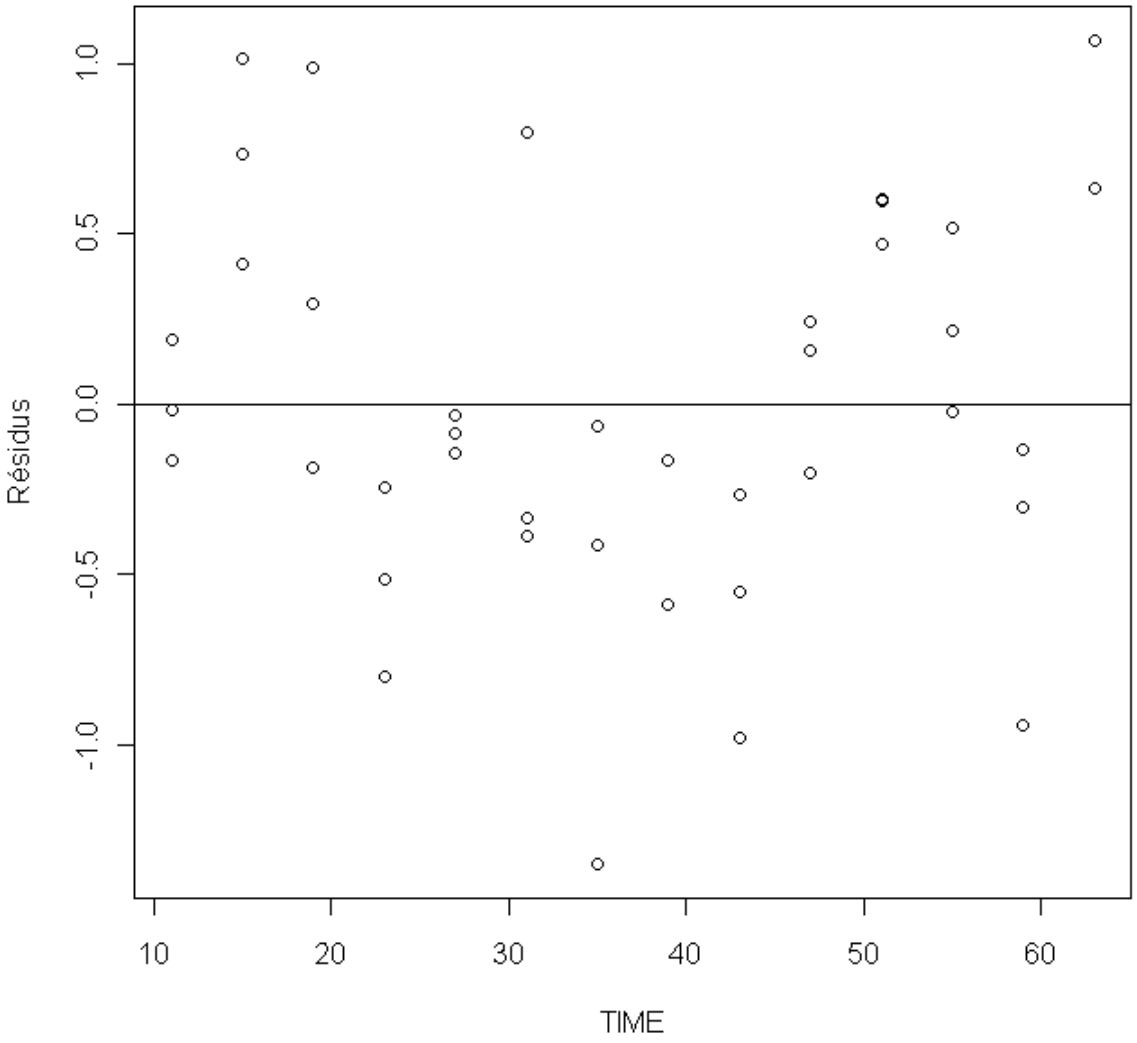
data: residus and donnee1$time

Bartlett's K-squared = 21.503, df = 13, p-value = 0.06356



Rev-erbα

bartlett.test(residus,donnée$time)
Bartlett test of homogeneity of variances
data: residus and donnee2$time
Bartlett's K-squared = 20.946, df = 13, p-value = 0.074



Le test de Bartlett dans les trois cas montre qu'on a une homogénéité de la variance des résidus au seuil α=0.05 la p-value du test étant supérieur à 0.05

3) Estimation des paramètres

La fonction gnls du logiciel R nous renseigne sur l'estimation des paramètres et la corrélation qu'il y a entre ces différents paramètres du modèle

Bmal1

	estimation	Std.Error	t-value	p-value
Y0	1.412142	0.1134194	12.45063	0e+00
b	0.013036	0.0028880	4.51377	1e-04
c	-0.493277	0.0575599	-8.56979	0e+00
phi	-12.097582	0.6551760	-18.46463	0e+00
tau	23.578344	0.6816946	34.58784	0e+00

Tous les paramètres sont significativement différents de zéro au seuil α=0.05.

Donc la mesure y de l'expression du gène Bmal1 peut s'exprimer en fonction du temps x sous la forme :

$$y = 1.412142 + 0.013036 \times x - 0.493277 \times \cos\left(\frac{2\pi(x-31+12.097582}{23.578344}\right)$$

Cry1

	estimation	Std.Error	t-value	p-value
Y0	1.845586	0.096111	19.203	< 2e-16
b	-0.004532	0.002398	-1.890	0.0668
c	-0.580370	0.050032	-11.600	1.02e-13
phi	3.074913	0.339356	9.061	8.08e-11
tau	22.895498	0.491350	46.597	< 2e-16

Rev-erb α

	estimation	Std.Error	t-value	p-value
Y0	2.908165	0.275505	10.556	2.87e-12
b	-0.001422	0.007073	-0.201	0.842
c	1.834737	0.154266	11.893	1.15e-13
phi	7.347279	0.330567	22.226	< 2e-16
tau	23.853785	0.517592	46.086	< 2e-16

Dans les deux derniers cas, nous avons les paramètres b qui sont non significatifs, mais avant de conclure on va d'abord vérifier si ce n'est pas du à une éventuelle colinéarité entre les paramètres du modèle.

4) Corrélation des paramètres

La fonction gnls(…) du logiciel R nous renseigne sur la corrélation entre les différents paramètres lors d'une régression non linéaire.

Pour nos trois ajustements on a les résultats suivants :

Bmal1

	Y0	b	c	phi
b	-0.938			
c	-0.170	0.181		
phi	-0.437	0.433	0.145	
tau	0.523	-0.561	-0.179	-0.777

Cry1

	Y0	b	c	phi
b	-0.927			
c	0.032	0.016		
phi	-0.104	0.148	0.001	
tau	-0.403	0.423	0.011	0.384

Rev-erbα

	Y0	b	c	phi
b	-0.915			
c	-0.094	0.042		
phi	-0.121	0.172	0.040	
tau	-0.176	0.254	-0.063	0.205

Nous remarquons une forte corrélation entre les paramètres y0 et b.

Cette corrélation dépassant 0.8, nous pouvons envisager un problème de multi colinéarité entre ces deux paramètres y0 et b qui entrainerait la suppression de l'un de ces paramètres.

Dans les deux derniers ajustements on remarque que le paramètre b est non significatif. Pour vérifier si la non significativité du paramètre b est dû ou non à son éventuelle colinéarité avec le paramètre y0 on va procéder ainsi :

Dans un premier temps on va refaire l'ajustement de la régression non linéaire mais cette fois-ci en fixant la valeur du paramètre y0 à sa valeur trouvée dans le premier ajustement

Ce qui donne les résultats suivants :

Cry1

	estimation	Std.Error	t-value	p-value
b	-0.0046363	0.0008851	-5.238	6.74e-06
C	-0.5798068	0.0493232	-11.755	4.68e-14
Phi	3.0689337	0.3293359	9.319	3.03e-11
tau	22.8979912	0.4441910	51.550	< 2e-16

Le paramètre b devient alors significatif, on en déduit que le niveau d'expression du gène Cry1 y s'exprime en fonction du temps x sous la forme :

$$y = 1.85 - 0.005x - 0.58 \times \cos\left(\frac{2\pi(x - 31 - 3.07)}{22.90}\right)$$

Reverb α

	estimation	Std.Error	t-value	p-value
b	-0.001422	0.002795	-0.509	0.614
C	1.834737	0.150904	12.158	4.03e-14
Phi	7.347280	0.322228	22.801	< 2e-16
tau	23.853779	0.501380	47.576	< 2e-16

Le paramètre b est toujours non significatif, donc le niveau d'expression y du gène Rev-erb α s'exprime en fonction du temps x sous la forme :

$$y = 2.9 + 1.83 \times cos\left(\frac{2\pi(x - 31 - 7.35)}{23.85}\right)$$

5) Résultats pour les autres gènes

Par le même procédé on a établit la rythmicité des neuf gènes horloge des kératinocytes, la rythmicité des sept gènes Bmal1, Per1, Per2, Cry1, Cry2, Rev-erb α et Ror α des mélanocytes et la rythmicité des gènes Bmal1, Per1, Per2, Per3, Cry1 et Cry2 des fibroblastes.

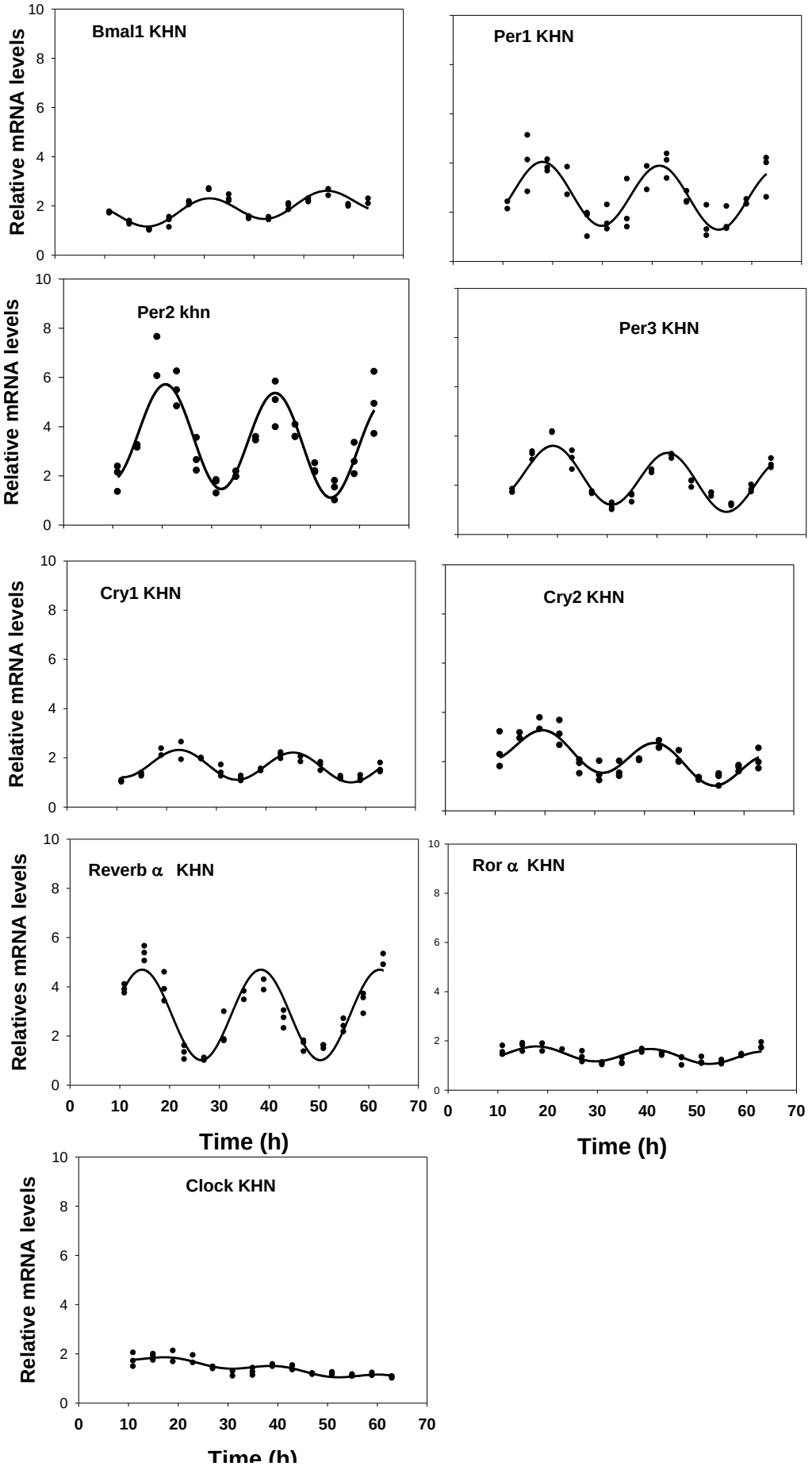
Sauf qu'il est noté que pour des problèmes de normalité des résidus, on a dû utiliser le modèle de régression non linéaire suivant :

$$log(y_i) = y0 + b \times x_i + c \times \cos\left(\frac{2\pi(x_i - 31 - ph)}{tau}\right) + \varepsilon_i$$

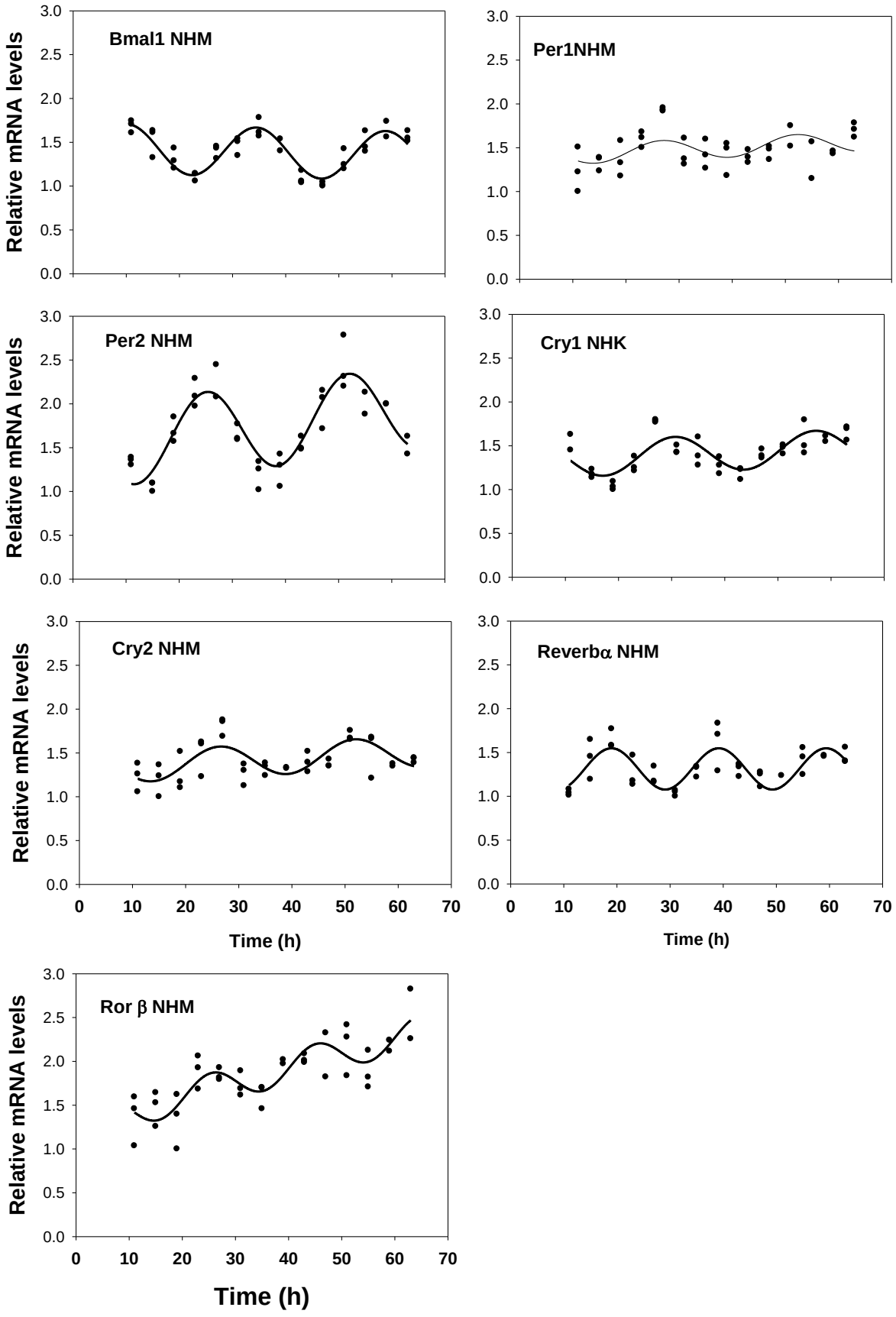
Pour l'ajuster à nos données concernant les gènes des fibroblastes.

L'ensemble des ajustements est illustré par les figures suivantes

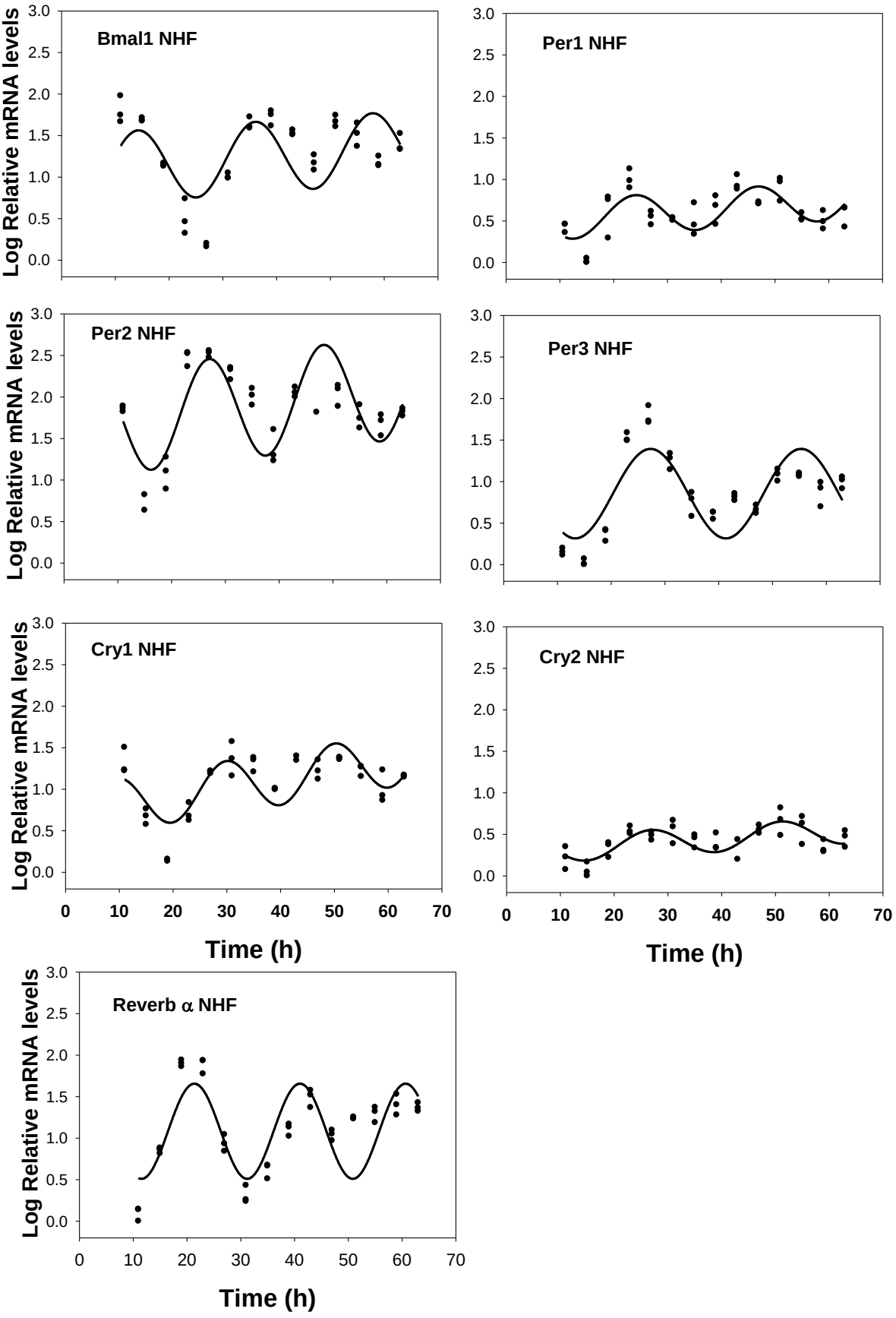
Kératinocytes


Bmal1 KHN
Per1 KHN
Per2 khn
Per3 KHN
Cry1 KHN
Cry2 KHN
Reverb α KHN
Ror α KHN
Clock KHN
Relative mRNA levels
Relatives mRNA levels
Time (h)

Mélanocytes



Fibroblastes



III) Recherche de la période commune

Pour chaque type cellulaire, on va essayer de mettre en évidence une oscillation d'ensemble avec une période commune.

Donc pour chaque type cellulaire on va déterminer l'ensemble des gènes qui s'expriment de manière périodique avec la même période.

Pour cela, si par exemple, pour un type cellulaire donné on a établi la rythmicité de p gènes parmi les neuf gènes horloge dans la première partie, dans cette partie on va essayer d'ajuster p sinusoïdes qui s'expriment de manière périodique avec la même période mais avec un déphasage, une amplitude, une ordonnée à l'origine et une éventuelle dérive linéaire spécifique pour chaque sinusoïde à l'ensemble de nos données pour les p gènes.

Pour cela on va essayer d'ajuster le modèle de régression non linéaire

$$y_i = y0[j] + b[j] \times x_i + c[j] \times \cos\left(\frac{2\pi(x_i - 31 - ph[i])}{\boldsymbol{tau}}\right) + \varepsilon_i$$

à nos données pour les p gènes.

j=1…… p, p étant le nombre de gènes qui s'expriment de façon rythmique

$y0[j]$: L'ordonnée à l'origine de la j ième sinusoïde

$b[j]$: Une éventuelle dérive linéaire

$c[j]$: L'amplitude de la sinusoïde

$phi[j]$: Le déphasage

tau : La période commune

1) Estimation des paramètres

Kératinocytes

Pour les Kératinocytes, on va essayer d'estimer les paramètres de l'ajustement de notre modèle aux données des gènes Bmal1, Per1, Per2, Per3, Cry1, Cry2, Rev-erb α ,Ror α et Clock. Ce qui nous donne les résultats suivants :

PARAMETRE	ESTIMATION	Std.Error	t-value	Pr (>|t|)
Y01	1.369343	0.225087	6.084	3.24e-09
Y02	3.535376	0.228943	15.442	< 2e-16
Y03	4.185052	0.240446	17.405	< 2e-16
Y04	2.820643	0.228844	12.326	< 2e-16
Y05	1.838489	0.224961	8.172	6.43e-15
Y06	3.483184	0.226884	15.352	< 2e-16
Y07	2.890320	0.225438	12.821	< 2e-16
Y08	1.673190	0.224614	7.449	8.19e-13
Y09	1.995798	0.224378	8.895	< 2e-16
B1	0.014256	0.005565	2.562	0.010853
B2	-0.019367	0.005682	-3.409	0.000734
B3	-0.017077	0.006013	-2.840	0.004791
B4	-0.014405	0.005677	-2.537	0.011623
B5	-0.004293	0.005559	-0.772	0.440568
B6	-0.032415	0.005619	-5.769	1.83e-08
B7	-0.001804	0.005584	-0.323	0.746802
B8	-0.006180	0.005551	-1.113	0.266369
B9	-0.015553	0.005544	-2.805	0.005320
C1	-0.485286	0.129915	-3.735	0.000221
C2	1.245029	0.130252	9.559	< 2e-16
C3	2.240781	0.130133	17.219	< 2e-16
C4	1.187273	0.130038	9.130	< 2e-16
C5	0.579702	0.127552	4.545	7.71e-06
C6	0.897145	0.130017	6.900	2.64e-11
C7	1.829760	0.129187	14.164	< 2e-16
C8	0.324252	0.130209	2.490	0.013254
C9	0.129261	0.130212	0.993	0.321583
Phi1	11.222596	0.967053	11.605	< 2e-16
Phi2	10.774594	0.379736	28.374	< 2e-16
Phi3	12.025586	0.216299	55.597	< 2e-16
Phi4	11.313839	0.398460	28.394	< 2e-16
Phi5	14.594915	0.825016	17.690	< 2e-16
Phi6	11.175300	0.524984	21.287	< 2e-16
Phi7	7.347288	0.259279	28.337	< 2e-16
Phi8	10.237451	1.442415	7.097	7.74e-12
Phi9	10.140271	3.615828	2.804	0.005338
tau	23.067520	0.211460	109.087	< 2e-16

Pour chaque gène on a des données qu'on veut ajuster à une sinusoïde.

On remarque qu'il est impossible d'ajuster neuf sinusoïdes qui s'expriment avec une période à nos données pour les neuf gènes. En effet, on a l'amplitude de la sinusoïde qu'on veut ajuster à nos données pour le gène Clock est non significative.

On va alors essayer d'ajuster le modèle de régression non linéaire

$$y_i = y0[j] + b[j] \times x_i + c[j] \times \cos\left(\frac{2\pi(x_i - 31 - ph[i])}{\boldsymbol{tau}}\right) + \varepsilon_i$$

à nos données pour les 8 gènes bmal1, per1, per2, per3, cry1, cry2, Rev-erb α et Ror α.

PARAMETRE	ESTIMATION	Std.Error	t-value	Pr (>|t|)
Y01	1.369414	0.237250	5.772	1.98e-08
Y02	3.535192	0.241319	14.649	< 2e-16
Y03	4.184714	0.253460	16.510	< 2e-16
Y04	2.820466	0.241215	11.693	< 2e-16
Y05	1.838424	0.237117	7.753	1.46e-13
Y06	3.483054	0.239146	14.565	< 2e-16
Y07	2.890224	0.237619	12.163	< 2e-16
Y08	1.673146	0.236751	7.067	1.15e-11
B1	0.014254	0.005866	2.430	0.015689
B2	-0.019362	0.005989	-3.233	0.001365
B3	-0.017067	0.006339	-2.693	0.007495
B4	-0.014400	0.005984	-2.406	0.016722
B5	-0.004291	0.005860	-0.732	0.464576
B6	-0.032411	0.005922	-5.473	9.48e-08
B7	-0.001802	0.005886	-0.306	0.759761
B8	-0.006179	0.005851	-1.056	0.291808
C1	-0.485300	0.136937	-3.544	0.000458
C2	1.245049	0.137292	9.069	< 2e-16
C3	2.240803	0.137169	16.336	< 2e-16
C4	1.187299	0.137067	8.662	3.10e-16
C5	0.579700	0.134444	4.312	2.21e-05
C6	0.897161	0.137045	6.546	2.61e-10
C7	1.829783	0.136165	13.438	< 2e-16
C8	0.324248	0.137246	2.363	0.018800
Phi1	11.223032	1.019290	11.011	< 2e-16
Phi2	10.774801	0.400259	26.920	< 2e-16
Phi3	12.025908	0.227995	52.746	< 2e-16
Phi4	11.314186	0.419990	26.939	8.64e-11
Phi5	14.595300	0.869629	16.783	< 2e-16
Phi6	11.175727	0.553350	20.196	< 2e-16
Phi7	7.347446	0.273301	26.884	< 2e-16
Phi8	10.237685	1.520398	6.734	8.64e-11
tau	23.068371	0.223036	103.429	< 2e-16

Nous remarquons que les huit sinusoïdes ont chacune une amplitude significativement différente de zéro

Mélanocytes

Pour les mélanocytes, on va essayer d'estimer les paramètres de l'ajustement de notre modèle aux données des gènes Bmal1, Per1, Per2, Cry1, Cry2 et Rev-erb α. Ce qui nous donne les résultats suivants.

PARAMETRE	ESTIMATION	Std.Error	t-value	Pr(>|t|)
Y01	1.472001	0.080859	18.205	< 2e-16
Y02	1.397726	0.079307	17.624	< 2e-16
Y03	1.541442	0.079298	19.439	< 2e-16
Y04	1.348849	0.080585	16.738	< 2e-16
Y05	1.316342	0.079274	16.605	< 2e-16
Y06	1.169379	0.079697	14.673	< 2e-16
B1	-0.002472	0.002024	-1.221	0.223353
B2	0.002680	0.001995	1.344	0.180461
B3	0.006341	0.002005	3.163	0.001780
B4	0.002052	0.002029	1.011	0.312918
B5	0.003058	0.001994	1.534	0.126532
B6	0.004885	0.002005	2.436	0.015623
C1	0.276982	0.044718	6.194	2.83e-09
C2	-0.112623	0.043442	-2.592	0.010165
C3	-0.496636	0.042433	-11.704	< 2e-16
C4	-0.210823	0.047093	-4.477	1.21e-05
C5	-0.144492	0.042659	-3.387	0.000836
C6	0.126680	0.046922	2.700	0.007475
Phi1	3.079627	0.650739	4.733	3.96e-06
Phi2	8.419819	1.634736	5.151	5.74e-07
Phi3	7.267014	0.380958	19.076	< 2e-16
Phi4	12.624998	0.812125	15.546	< 2e-16
Phi5	7.599286	1.294237	5.872	1.57e-08
Phi6	37.118757	1.472966	25.200	< 2e-16
tau	25.216392	0.488528	51.617	< 2e-16

Nous remarquons que les six sinusoïdes ont chacune une amplitude significativement différente de zéro

Fibroblastes

Pour les Fibroblastes, on va essayer d'estimer les paramètres de l'ajustement de notre modèle $\log(y_i) = y0[j] + b[j] \times x_i + c[j] \times \cos\left(\frac{2\pi(x_i - 31 - ph[i])}{tau}\right) + \varepsilon_i$

aux données des gènes Bmal1, Per1, Per2, Per3, Cry1 et Cry2. Ce qui nous donne les résultats suivants :

PARAMETRE	ESTIMATION	Std.Error	t-value	Pr (>|t|)
Y01	1.075409	0.123014	8.742	5.17e-16
Y02	0.473911	0.123085	3.850	0.000153
Y03	1.638259	0.123667	13.247	< 2e-16
Y04	0.569555	0.123067	4.628	6.21e-06
Y05	0.738512	0.124292	5.942	1.05e-08
Y06	0.258412	0.122969	2.101	0.036705
B1	0.004394	0.003041	1.445	0.149874
B2	0.004331	0.003043	1.423	0.156002
B3	0.005802	0.003057	1.898	0.058958
B4	0.009244	0.003041	3.040	0.002644
B5	0.009678	0.003077	3.145	0.001882
B6	0.004862	0.003039	1.600	0.110999
C1	-0.424679	0.070756	-6.002	7.65e-09
C2	-0.236194	0.070381	-3.356	0.000927
C3	-0.496846	0.070575	-7.040	2.26e-11
C4	-0.434548	0.070779	-6.140	3.66e-09
C5	-0.289525	0.069253	-4.181	4.15e-05
C6	-0.289525	0.070433	-2.170	0.031008
Phi1	-5.889963	0.616998	-9.546	< 2e-16
Phi2	4.349375	1.040086	4.182	4.13e-05
Phi3	7.371904	0.495256	14.885	< 2e-16
Phi4	6.703179	0.562686	11.913	< 2e-16
Phi5	10.073125	0.866439	11.626	< 2e-16
Phi6	7.493187	1.604016	4.672	5.12e-06
tau	22.241148	0.390746	56.920	< 2e-16

Nous remarquons que les huit sinusoïdes ont chacune une amplitude significativement différente de zéro.

2) Comparaison de modèles

Pour chaque type cellulaire après avoir déterminer une période commune pour la rythmicité de p gènes

Nous allons tester l'hypothèse nulle.

H0 : il n'existe pas de différence significative entre l'ajustement du modèle de régression non linéaire $y_i = y0 + b \times x_i + c \times \cos\left(\frac{2\pi(x_i - 31 - ph)}{tau}\right) + \varepsilon_i$ à nos données pour chacun des p gènes et l'ajustement du modèle de régression non linéaire $y_i = y0[j] + b[j] \times x_i + c[j] \times \cos\left(\frac{2\pi(x_i - 31 - ph[i])}{tau}\right) + \varepsilon_i$ à données pour ces p gène.

Contre l'alternative

H1 : il existe une différence significative entre les deux ajustements.

Sous H0 F= $\frac{(SS1 - SS2)/(DF1 - DF2)}{SS2/DF2}$ suit une loi de Fisher à $DF1 - DF2$ degrés de liberté au numérateur et $DF2$ degrés de liberté au dénominateur avec :

$SS2 = \sum_{j=1}^{p} SCR_j$

$DF2 = \sum_{i=1}^{p} DF_j$

SCR_j étant la somme des carrés résiduels du à l'ajustement du modèle de régression non linéaire$y_i = y0 + b \times x_i + c \times \cos\left(\frac{2\pi(x_i - 31 - phi)}{tau}\right)$ à nos données pour le gène j, DF_j le degré de liberté qui lui est associé et $SS1$ la somme des carrés résiduels du à l'ajustement du modèle de régression non linéaire $y_i = y0[j] + b[j] \times x_i + c[j] \times \cos\left(\frac{2\pi(x_i - 31 - ph[i])}{tau}\right) + \varepsilon_i$ à nos données pour les p gènes .

Kératinocytes

Tableau d'analyse de la variance pour la régression commune des 8 gênes

Source	DF	SS
Régression	25	284.8967
Erreur	295	109.71
Total	320	394.6067

Tableau d'analyse de la variance pour la régression des 8 gênes individuelles

source	DF	SS
Régression	32	287.8366
Erreur	288	106.7701
Total	320	394.6067

Tableau d'analyse de la variance de la différence entre les deux régressions

Source	DF	SS	MS	
Régression combinée	25	284.8967		
Régression individuel	32	287.8366		P(>|t|)
Différence des 2 reg	7	2.9399	0.4199857	0.342357
Erreur	288	106.7701	0.3707295	
Total	320			

On a la p-value du test qui est égale à 0.342357 donc on ne rejette pas l'hypothèse H0

Mélanocytes

Tableau d'analyse de la variance pour la régression des 6 gènes combinés

Source	DF	SS
Régression	19	9.2054
Erreur	221	5.2211
Total	240	14.4265

Tableau d'analyse de la variance pour la régression des 6 gênes individuelles

Source	DF	SS
Régression	24	9.2556
Erreur	216	5.1709
Total	240	14.4265

Tableau d'analyse de la variance de la différence entre les deux régressions

Source	DF	SS	MS	
Régression combi	19	9.2054		
Régression indiv	24	10.038		
Différence	5	0.0502	0.01673333	P(t>|t|)
Erreur	216	5.1709	0.03590903	0.7064615
Total	240			

On a la p-value du test qui est égale à 0.7064615 donc on ne rejette pas l'hypothèse H0

Fibroblastes

Tableau d'analyse de la variance pour la régression commune des 6 gênes

Source	DF	SS
Régression	19	18.38724

Erreur	221	20.015
Total	240	38.48885

Tableau d'analyse de la variance pour la régression des 6 gênes individuellement

Source	DF	SS
Régression	24	18.85925
Erreur	216	19.6296
Total	240	38.48885

Tableau d'analyse de la variance de la différence entre les deux régressions

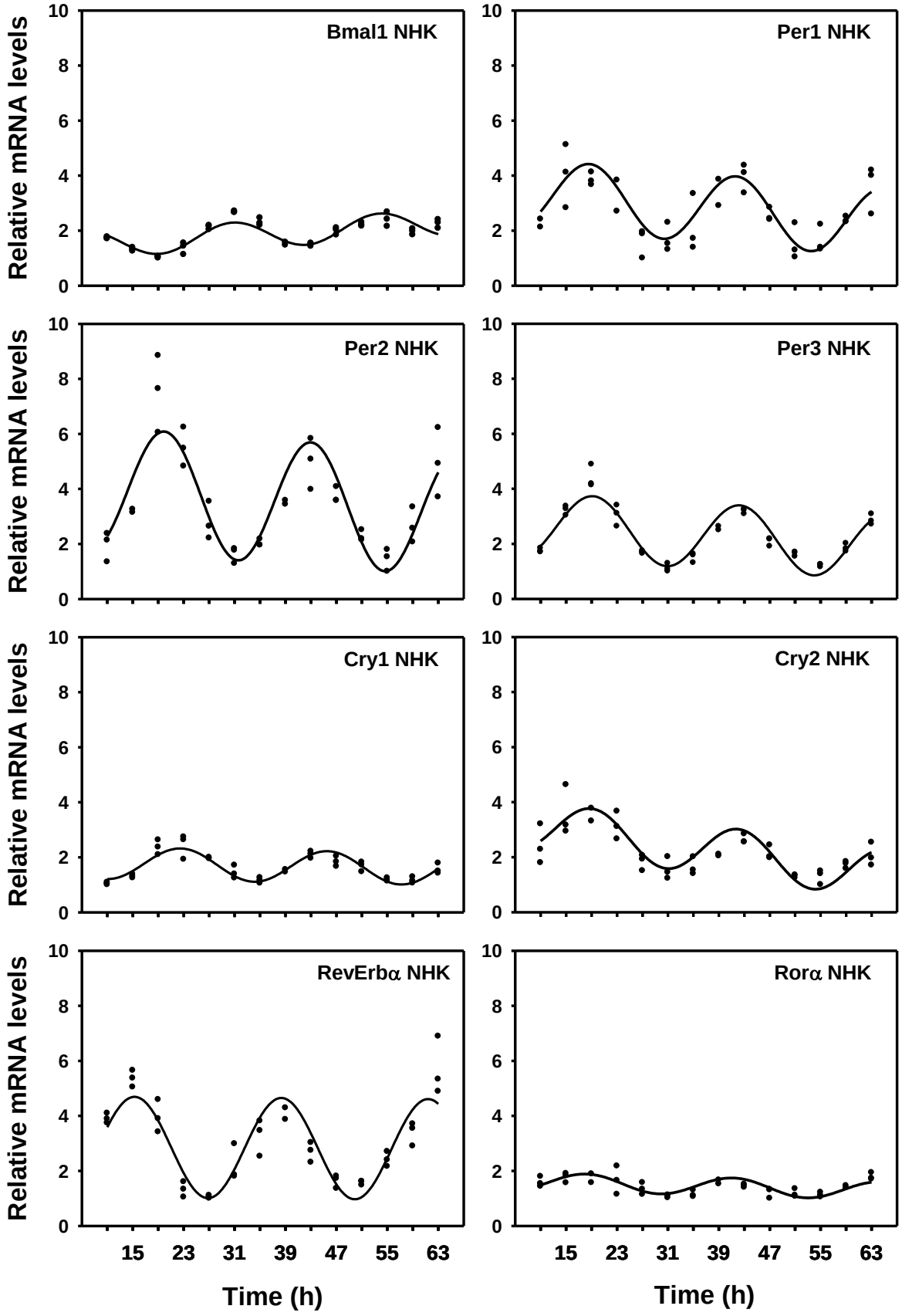
Source	DF	SS	
Régression combinée	19	18.38724	
Régression individuelle	24	18.85925	Pr (>|t|)
Différence des deux reg	5	0.47201	0.3170854
Erreur	216	19.6296	
Total	240	38.48885	

On a la p-value du test qui est égale à 0.3170854 donc on ne rejette pas l'hypothèse H0

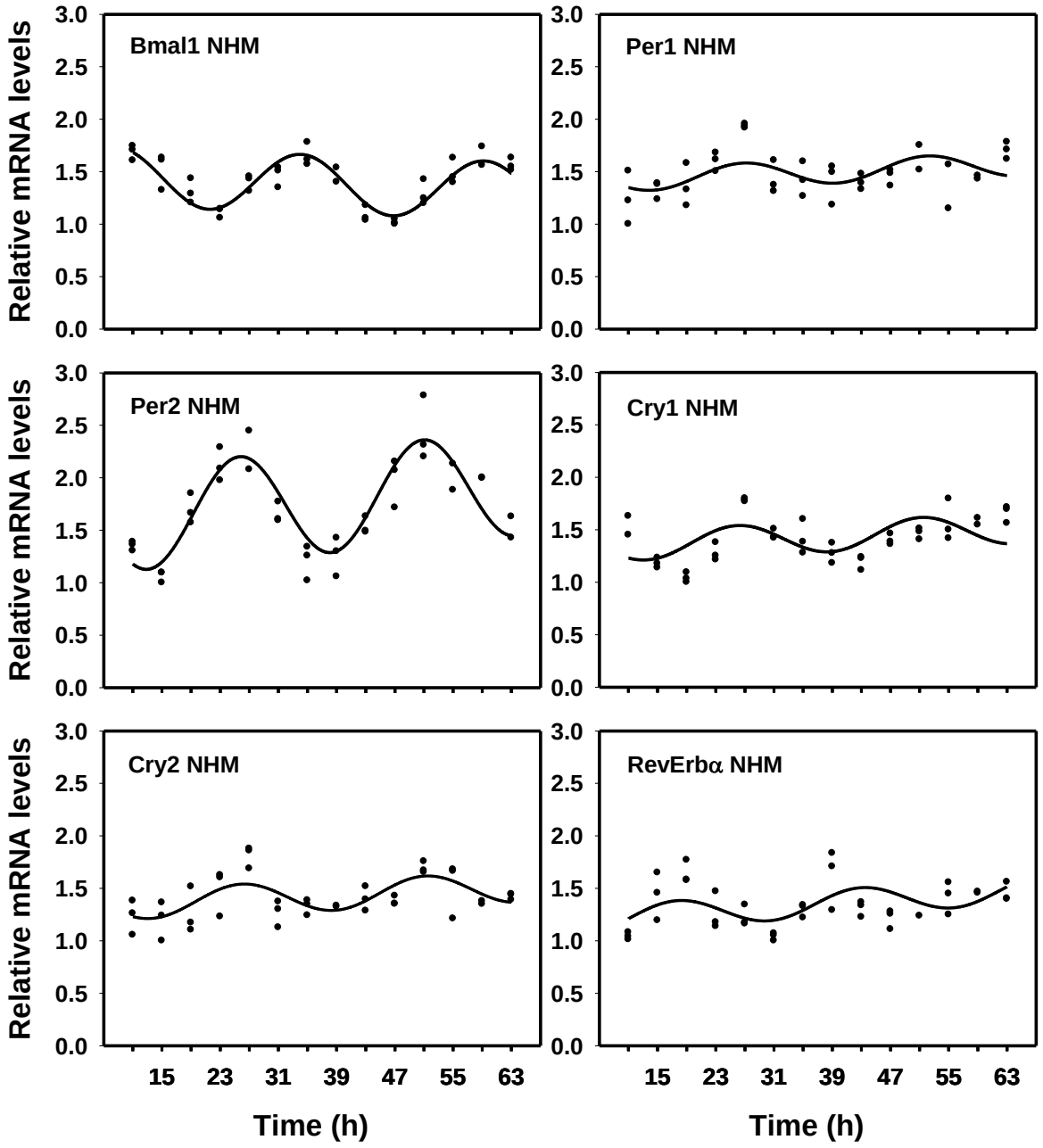
on déduit des résultats ci-dessus que les gènes horloge Bmal1,Per1, Per2, Per3, Cry1, Cry2, Rev-erb α et Ror α des kératinocytes s'expriment de manière périodique avec une même période égale à 23.07 ± 0.22, les gènes Bmal1, Per1, Per2, Cry1, Cry2 et Rev-erb α des mélanocytes s'expriment également de manière périodique avec la même période égale à 25.22±0.48 et enfin les gènes Bmal1, Per1,Per2, Per3, Cry1, Cry2 des fibroblastes s'expriment de manière périodique avec la même période qui est égale 22.24 ±0.39.

Ces résultats sont illustrés par les figures suivantes :

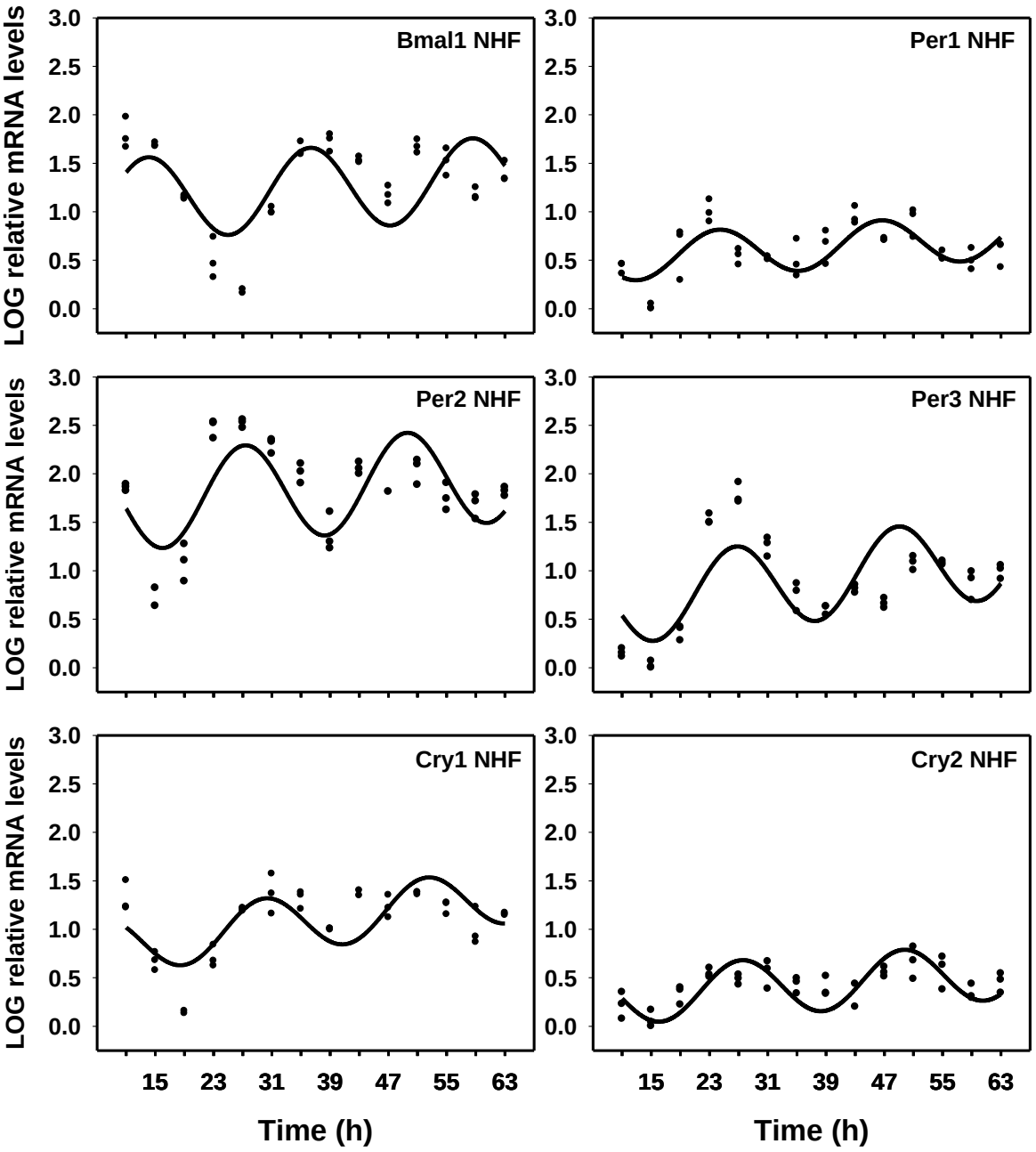
Kératinocytes


Bmal1 NHK
Per1 NHK
Per2 NHK
Per3 NHK
Cry1 NHK
Cry2 NHK
RevErbα NHK
Rorα NHK
Relative mRNA levels
Time (h)
0
2
4
6
8
10
15
23
31
39
47
55
63

Mélanocytes



Fibroblastes



IV) Analyse des phases

Pour chaque type cellulaire, une fois l'existence d'une période commune établi pour un ensemble de gènes donnés, notre objectif dans cette partie est de déterminer les relations de phase entre ces gènes et l'ordre dans lequel ils fonctionnent au cours du temps.

Dans ce rapport dans les trois types cellulaires, le gène Bmal1 étant l'activateur on va le prendre comme référence.

On suit ici la convention des biologistes : les phases sont exprimées en heures circadiennes, c'est-à-dire en 24ième de la période circadienne observée. Donc on convertie les phases astronomiques en phases circadiennes en utilisant la formule suivante :

$$phase\ circadienne = \frac{phase\ astronomique \times 24}{tau}$$

Kératinocytes

	Temps astronomique	Ecart type	Temps circadien	Temps circadien	Ecart type
Phi1	22.75722	1.019290	23.676283	0	1.060455
Phi2	10.774801	0.400259	11.209947	11.533664	0.4164237
Phi3	12.025908	0.227995	12.511581	12.835298	0.2372027
Ph14	11.314186	0.419990	11.771116	12.094833	0.4369515
Phi5	14.5953	0.869629	15.184739	15.508456	0.9047495
Phi6	11.175727	0.553350	11.627065	11.950782	0.5756973
Phi7	7.347446	0.273301	7.644177	7.967894	0.2843384
Phi8	10.237685	1.520398	10.651140	10.974857	1.5818

Mélanocytes

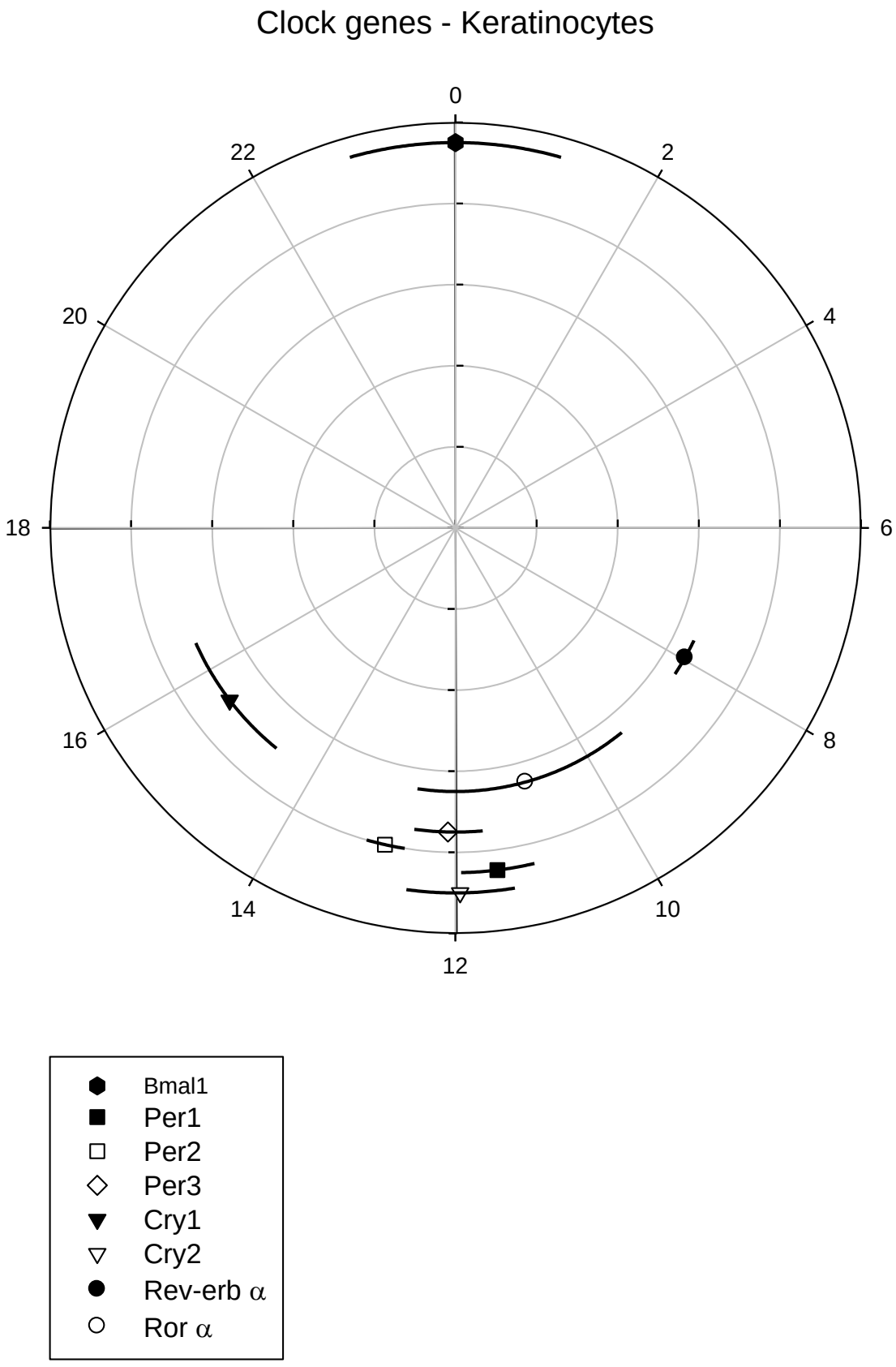
	Temps astronomique	Ecart type astronomique	Temps circadien	Temps circadien	Ecart type circadien
Phi1	3.079627	0.650739	2.931072	0	0.6193486
Phi2	21.02801	1.634736	20.013658	17.08259	1.5558794
Phi3	19.87521	0.380958	18.916467	15.98539	0.3625813
Ph14	25.23319	0.812125	24.015988	21.08492	0.7729496
Phi5	20.20748	1.294237	19.232709	16.30164	1.2318054
Phi6	11.90237	1.472966	11.328222	8.397150	1.4019129

Fibroblastes

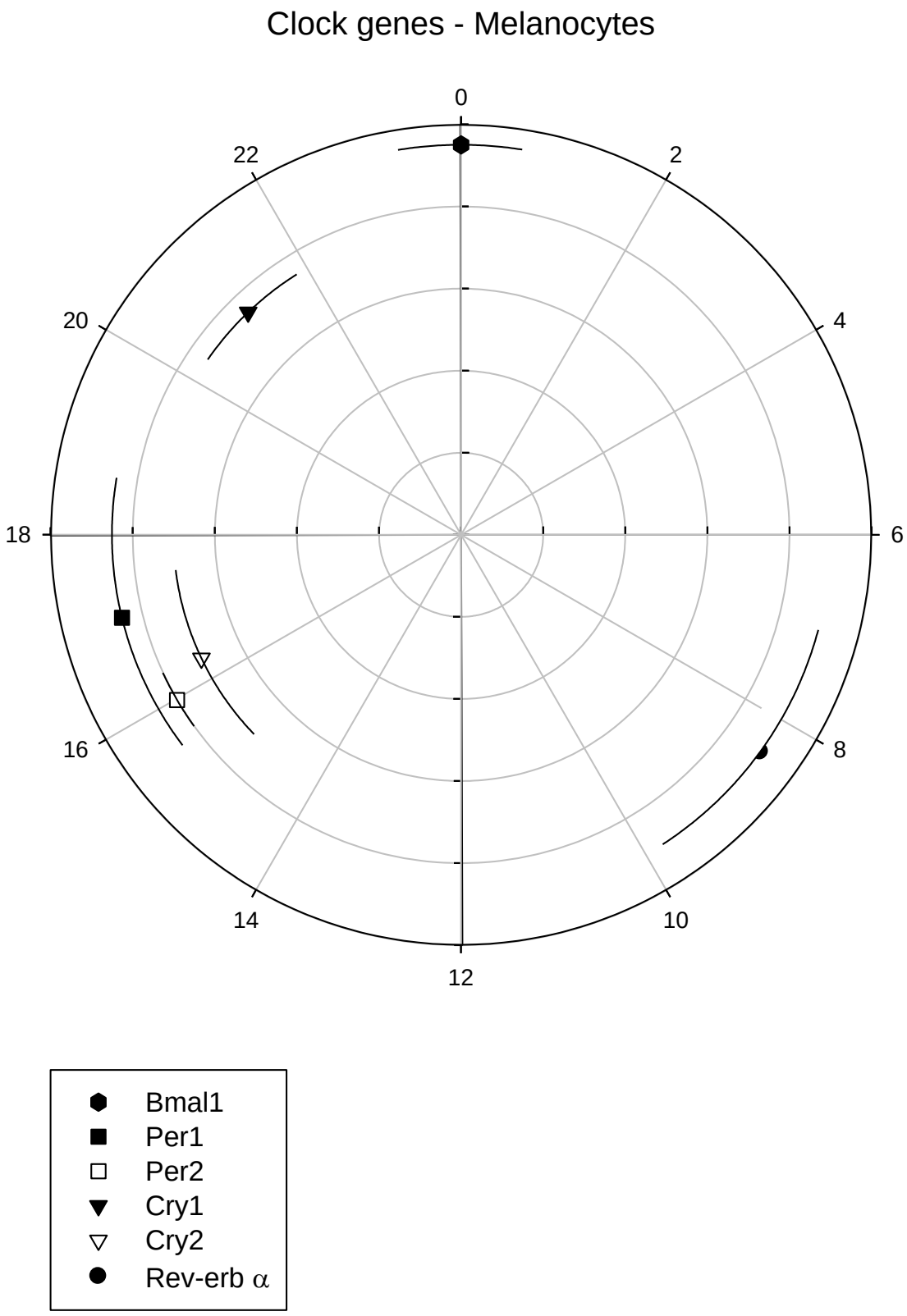
	Temps astronomique	Ecart type astronomique	Temps circadien	Temps circadien	Ecart type circadien
Phi1	5.230611	0.616998	5.644253	0	0.6657908
Phi2	15.469949	1.040086	16.693328	11.04907	1.1223370
Phi3	18.492478	0.495256	19.954881	14.31063	0.5344213
Ph14	17.823753	0.562686	19.233273	13.58902	0.6071838
Phi5	21.193699	0.866439	22.869718	17.22546	0.9349579
Phi6	18.613761	1.604016	20.085756	14.44150	1.7308632

A l’aide du logiciel de statistique sigma plot on a pu représenter les phases circadiennes et leurs écarts types sur un cercle, ce qui nous permet de mettre en évidence les relations de phases.

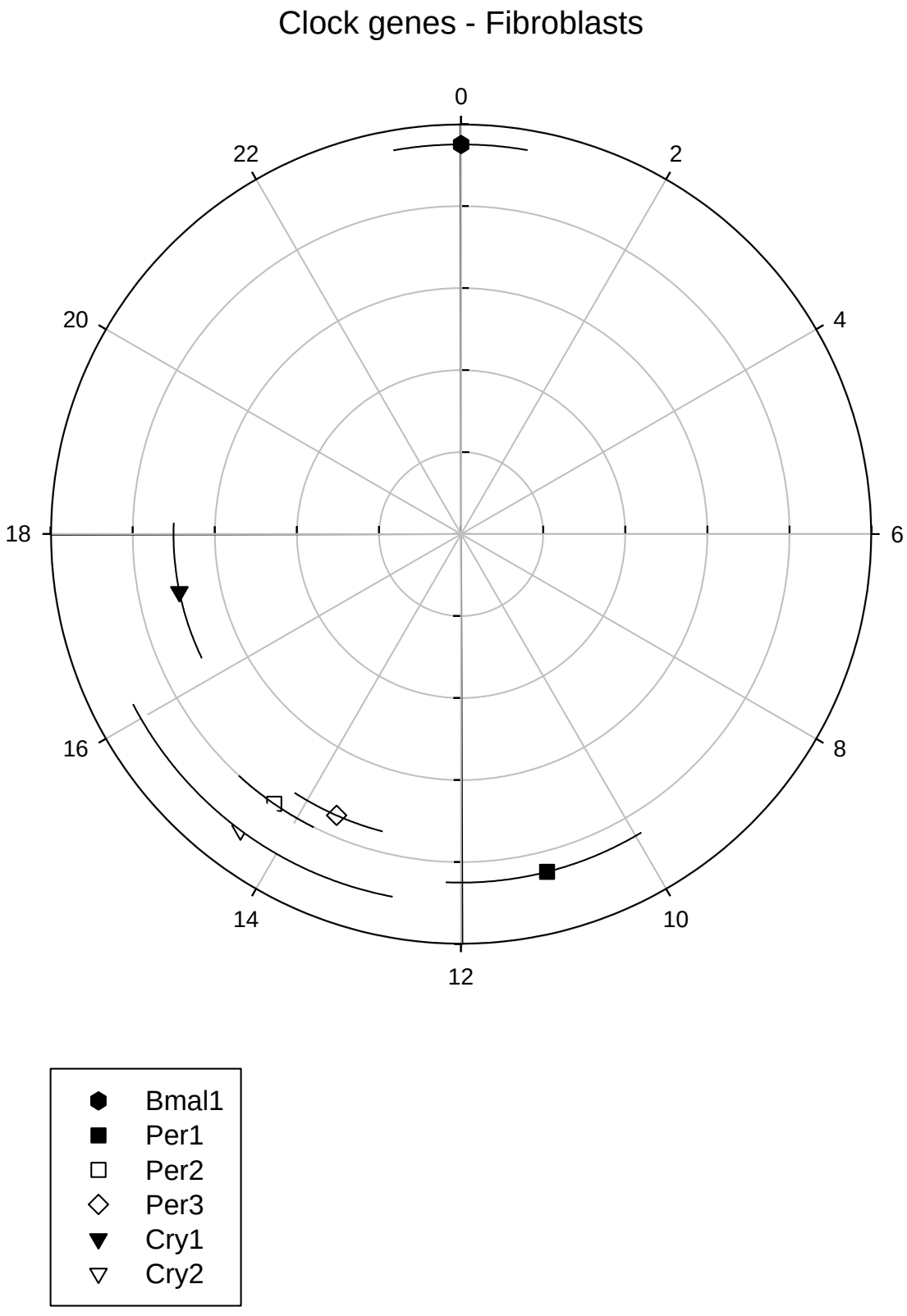
Kératinocytes



Mélanocytes



Fibroblastes



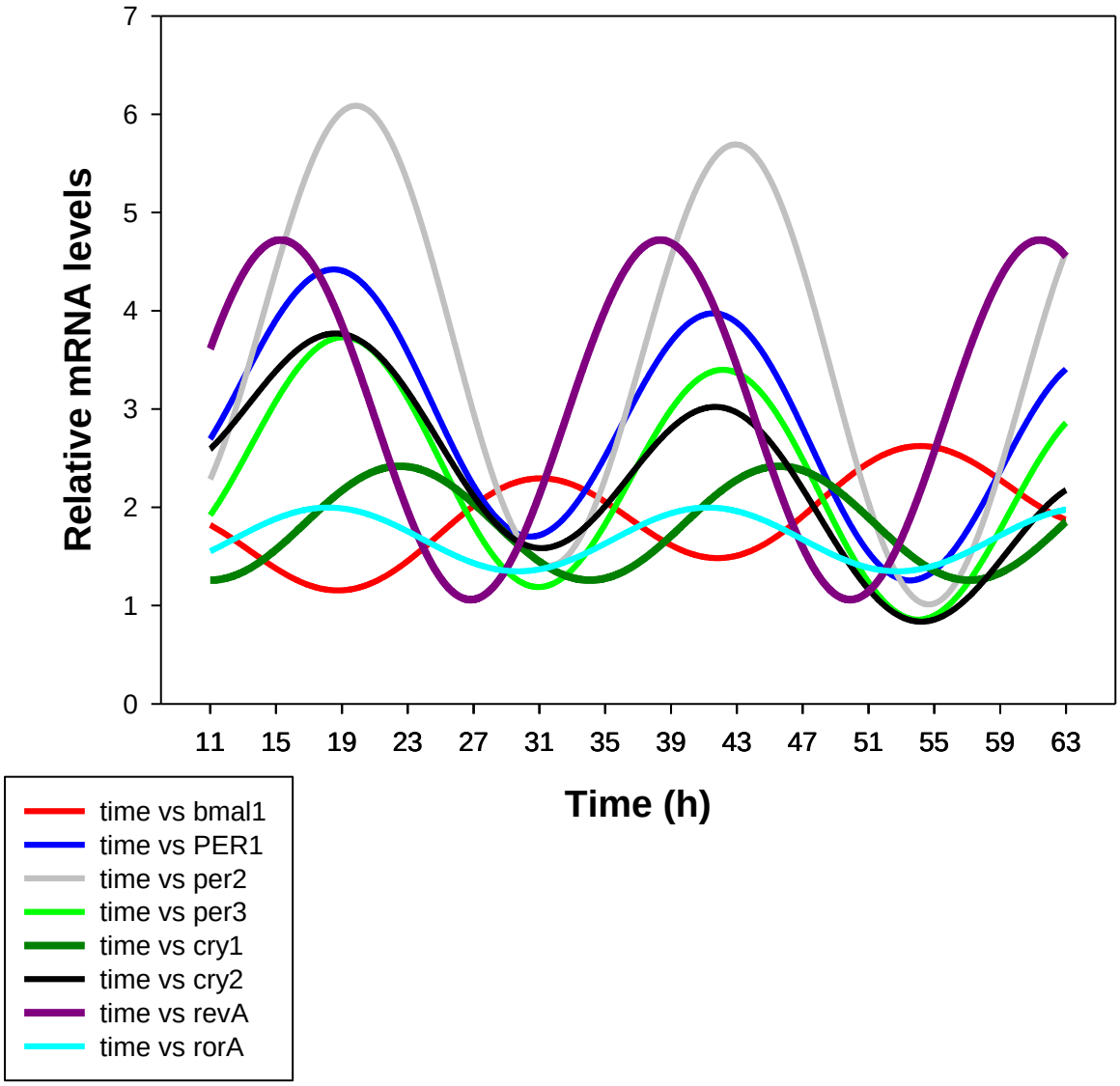
Pour les kératinocytes, on a le gène Rev-erb α qui est en retard de phase d'un tiers de période par rapport au gène Bmal1, les gènes Per1, Per2, Per3, Cry2 et Ror α sont en antiphase par rapport à Bmal1 et le gène Cry1 qui est en avance de phase d'un tiers de période par rapport à Bmal1.

Pour les mélanocytes, on a le gène Rev-erb α est en retard de phase d'un tiers de période par rapport à Bmal1, les gènes Per1, Per2 et Cry2 sont en avance de phase d'un tiers de période par rapport à Bmal1 et le gène Cry1 est avance de phase d'un sixième de période par rapport à Bmal1.

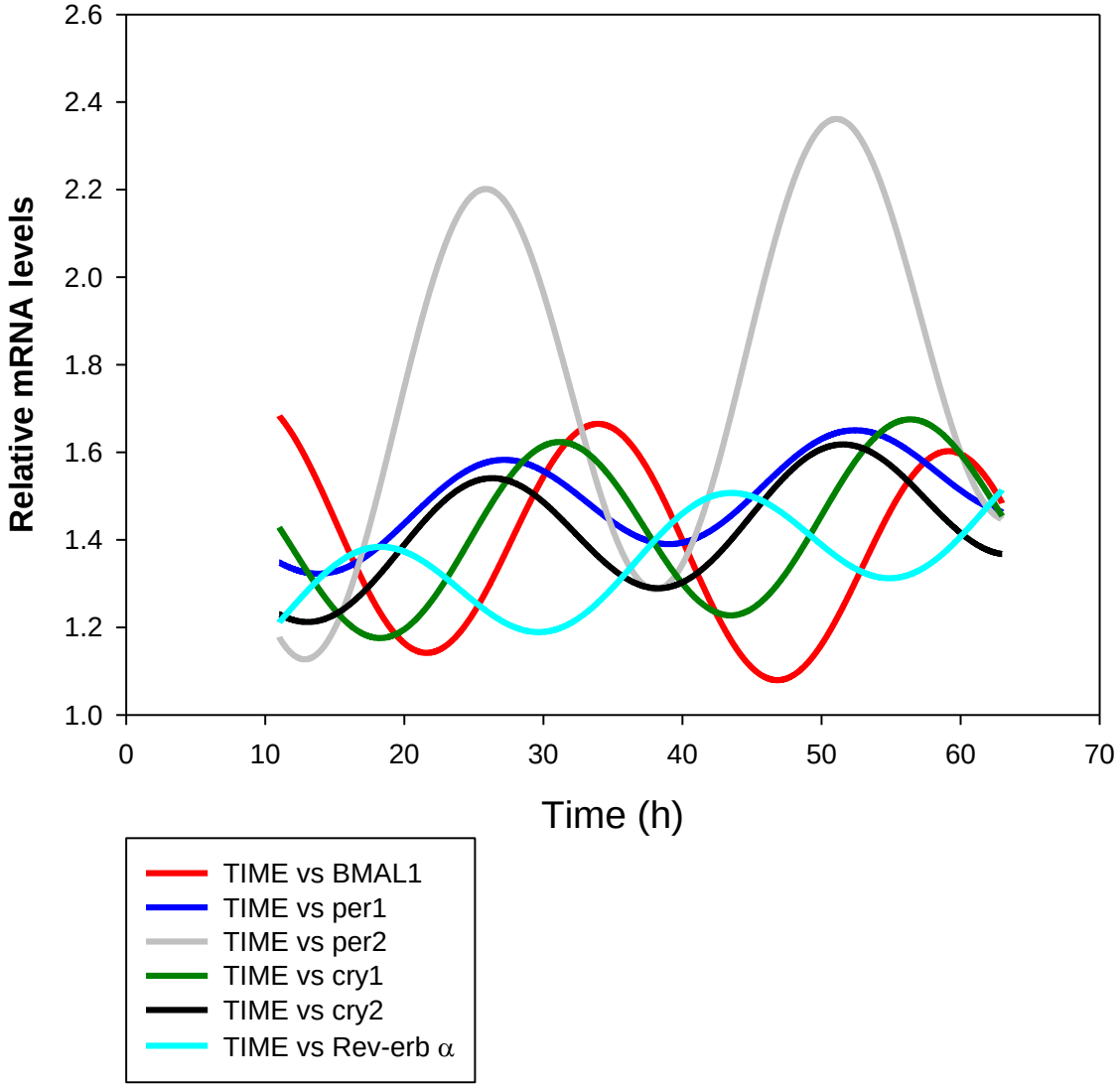
Pour les fibroblastes, le gène Per1 est en antiphase par rapport à Bmal1, les gènes Per2, Per3 et Cry2 sont en avance de phase de cinq douzième de période par rapport à Bmal1 et le gène Cry1 est en avance de phase d'un quart de période par rapport à Bmal1.

Pour chaque type cellulaire, on retrouve ces résultats qui sont plus compliqués à traduire dans les figures ci-dessous où on représente sur une même figure les p sinusoïdes qu'on ajuste à l'ensemble de nos données pour les p gènes qui s'expriment avec la même période.

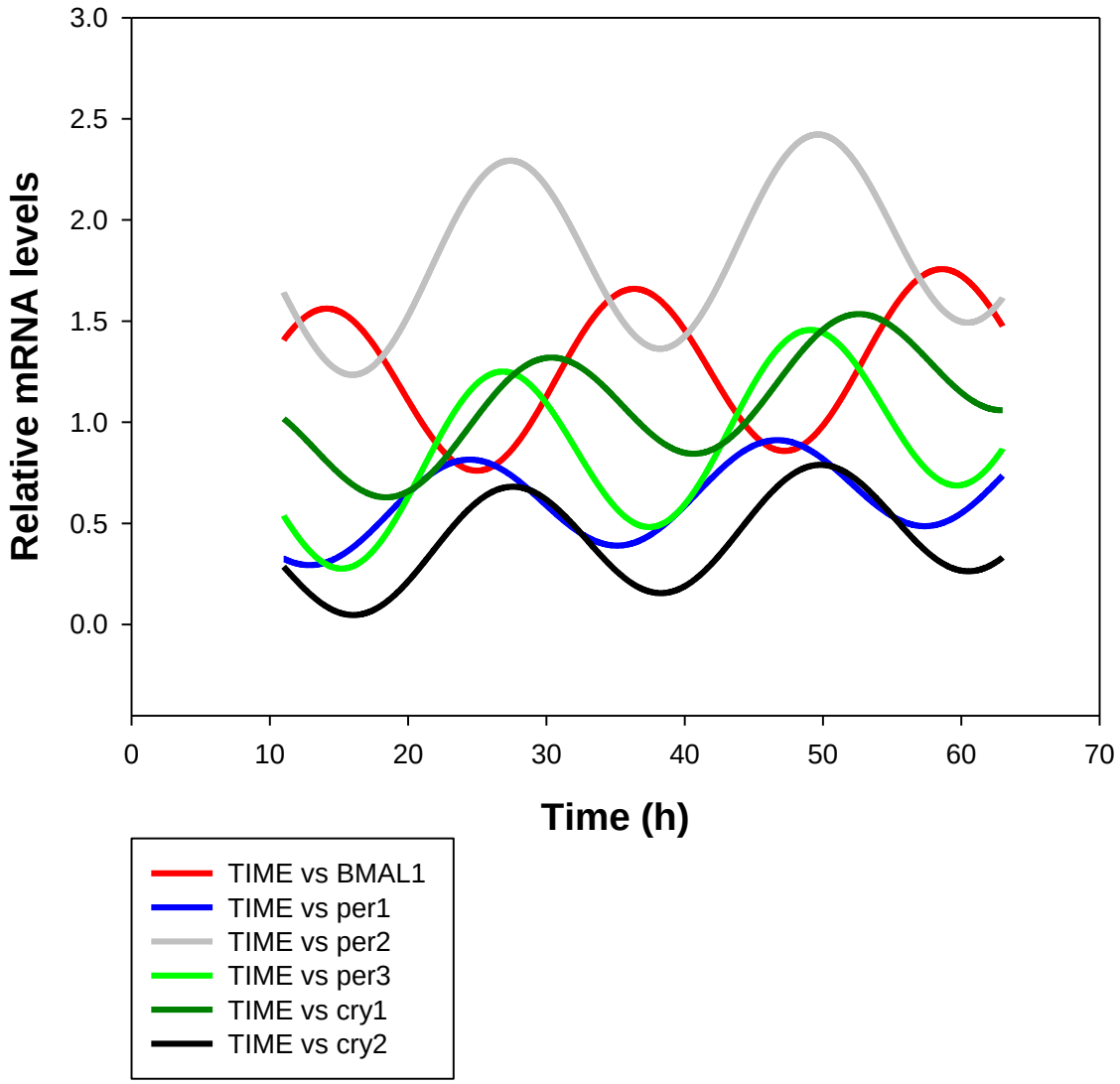
Mise en évidence des relations de phase pour les gènes des kératinocytes



Mise en évidence des relations de phase pour les gènes des mélanocytes



Mise en évidence des relations de phase pour les gènes des fibroblastes



Conclusion

L'analyse de régression non linéaire a permis de mettre en évidence l'expression rythmique de gènes horloge dans tous les types cellulaire. Elle a aussi permis de mettre en évidence une oscillation d'ensemble de tous les gènes concernés, avec une période commune, ce qui indique la présence d'une horloge moléculaire dans chaque type cellulaire.

Pour chaque type cellulaire la période commune qu'on a trouvé est circadienne (supérieure à 20 h et inférieure à 28 h).

Par contre les périodes communes sont différentes donc on a une spécificité cellulaire, ce qui a aussi été démontré dans les tissus avec une structure complexe.

Les relations de phase entre les gènes horloge et le gène Bmal1 (la référence) sont conformes au modèle de l'horloge circadienne décrit avant dans d'autres tissus.

La peau humaine contient donc trois horloges moléculaires autonomes (dans les mélanocytes, les fibroblastes et les kératinocytes) qui sont responsables du contrôle des fonctions rythmiques.

Ce stage m'a permis de mettre en pratique les méthodes statistiques vues en cours à des données réelles, d'approfondir mes connaissances. J'ai également pu me familiariser avec les logiciels comme R, Statistica et Sigma Plot, souvent utilisés par les biologistes. Pour finir, ce stage réalisé dans un environnement de chercheurs m'a permis de comprendre le fonctionnement d'un laboratoire et de situer la place des statistiques et leur utilité. Il me permettra de mieux choisir le milieu professionnel dans lequel j'exercerai ma future profession.

L'ensemble de ces résultats statistiques va être incorporé dans une publication sur les gènes horloge des cellules de la peau.

Bibliographie

Les livres

Encyclopedia of Statistical Sciences Samuel *Kotz, N. Balakrishnan, Campell B.Read*,

Brani Vidakovics, Norman L.Johnson John Wiley & Sons United Stated 2006

Moderne Regression Methods *Thomas P. Ryan Wiley New York 1997*

Non Linear Regression: *G.A.F. Saber, C.J. Wild Wiley United Stated 1989*

Non Linear Regression with R *Cristian Ritz, Jen Carl Streibig*

Statistical Tools for Non linear Regression *S.Huet, A.Bouvier*

M.-A Poursat E.Jolivet

Statistique Théorique et Application

Les sites internet:

http://www.r-project.org/

*cran.r-project.org/doc/contrib/Fox.../appendix-**nonlinear**-**regression**.pdf*

Russell.vcharite.uni-mrs.fr/EIE/fchap3.pdf

*en.wikipedia.org/wiki/**Nonlinear_regression***

www.statsci.org/smyth/pubs/eoe_nr.pdf

Annexe

Annexe1(Etude gene par gène)

```
donnee=read.table("kgene1.csv",dec=".",sep=";",quote="\"",header=T,na.strin
g="*")

boxplot(donnee$RQ,main="RQ")

head(donnee)

donnee1=na.omit(donnee)

plot(donnee$time,donnee$RQ,type="p",ylab="RQ.BMAL1")

plot(donnee1$time,donnee1$RQ,type="l",ylab="RQ.BMAL1")

ini=list(y0=1.42,b=0.01,c=0.5,phi=7,th=24)

md=RQ~y0+b*time+c*cos(2*3.1416*(time-phi-31)/th)

fm=nls(md,start=ini,data=donnee1,trace=TRUE,algorithm="port")

summary(fm)

lines(donnee1$time,predict(fm,donnee1$time),col='red',lwd=3)

residus=residuals(fm)

plot(donnee1$time,residus,ylab="Résidus",xlab="TIME",abline(h=0))

shapiro.test(residus)


qqnorm(residus,ylab="Quantiles observés",xlab="Quantiles theoriques")

qqline(residus,col="blue")

bartlett.test(residus,donnee1$time)

library(nlme)

reg=gnls(md,start=ini,data=donnee1)

summary(reg)

ini=list(b=0.01,c=0.5,phi=7,th=24)

mdbis=RQ~1.412142+b*time+c*cos(2*3.1416*(time-31-phi)/th)

fmbis=nls(mdbis,start=ini,data=donnee1,trace=TRUE,algorithm="port")

summary(fmbis)

regbis=gnls(mdbis,start=ini,data=donnee1)

summary(regbis)
```

Annexe2 (Recherche de période commune)

```
donnee=read.table("bmal1per123cry12revArorAclock.csv",dec=".",sep=";",quote
="\"",header=T,na.string="*")

boxplot(donnee$RQ,main="RQ")

head(donnee)

donnee1=na.omit(donnee)
ini=list(y0=1.42,b=0.01,c=0.5,phi=7,th=24)

md=y~y0+b*x+c*cos(2*3.1416*(x-phi-31)/th)

fm=nls(md,start=ini,data=donnee1,trace=TRUE,algorithm="port")

summary(fm)

d=coef(fm)

fm1=nls(y~y0[i]+b[i]*x+c[i]*cos(2*3.1416*(x-phi[i]-
31)/th),start=list(y0=rep(d[1],9)
,b=rep(d[2],9),c=rep(d[3],9),phi=rep(d[4],9),th=d[5]),
data=donnee1,algorithm="port")
summary(fm1)
```

Annexe3 (Relation de phases)

col(4)=rayon

col(5)=phi

col(6)=ecart type

for i=1 to 8 step 1 do

col(5+2*i)=data(cell(4,i),cell(4,i),300)

col(6+2*i)=data((cell(5,i)-cell(6,i)),(cell(5,i)+cell(6,i)),0.02)

end for