Analyse statistique de la relation entre phénotype et génotype chez la levure

Xueke Bai

Institut de botanique UMR7156

RAPPORT DU STAGE

Analyse statistique de la relation entre phénotype et génotype chez la levure

M1 statistique 2012-2013

Maitre du stage : Dr Joseph Schacherer

Xueke BAI

2013/8/12

SOMMAIRE

REMERCIEMENT

Je remercie spécialement mon maître de stage, Dr Joseph Schacherer, pour sa confiance et ses conseils qu'il m'a apporté pour tout au long de ce stage dans le domaine de statistique et génétique.

Je tiens également à remercier Jing Hou pour les explications génétiques qui m'ont permis de mieux comprendre le domaine de l'étude du laboratoire.

Je remercie toute les personnes du laboratoire pour leur accueil et leur sympathie.

RESUME

La statistique appliquée est utilisée dans presque tous les domaines de l'activités: ingénierie, management, économie, biologie, informatique, etc. Maintenant parlons des méthodes statistique qui s'appliquent au domaine de la biologie montrant la puissance des statistiques.

Pour étudier la relation entre phénotype et génotype chez la levure, l'existence de QTL peut déterminer la relation. Le laboratoire collecte les données de génotype(les marqueurs, les chromosomes etc.) et phénotype (déterminant la vitesse de croissance). Après on divise le phénotype en groupe en fonction des différences marqueurs et l'on observe les différences entre ces groupes. Equivalent au test ANOVA, il s'appelle LOD Score, c'est un indicateur de QTL. Ensuite, un test de permutation et une méthode bayésienne nous aident déterminer l'intervalle de confiance de QTL. On peut supposer que l'intervalle a l'impact au phénotype.

Tous les analyses sont effectués par logiciel R package QTL.

I INTRODUCTION

I.I PRESENTATION DU LABORATOIRE

J'ai effectué mon stage à l'institut de botanique de Strasbourg, au sein de l'UMR 7156. C'est un laboratoire commun du CNRS et l'Université de Strasbourg. Le laboratoire est constitué de 6 équipes qui utilisent les micro-organismes (levures et bactéries) comme modèle de recherche. Leurs domaines d'étude sont la génétique moléculaire, la génomique, et la microbiologie.

L'institut se diverse en deux départements:

- Le département "micro-organismes, génomes et environnement"
- Le département "génétique moléculaire et cellulaire"

Pour mon stage, j'ai intégré le département "micro-organismes, génomes et environnement".

Au sein de ce département, j'ai travaillé dans l'équipe "Variation intra-spécifique et évolution des génomes " du Dr Joseph Schacherer. L'équipe se focalise sur deux espèces de levures pour leurs recherches: Saccharomyces cerevisiae et Lachancea kluyveri. Pour chaque espèce de différents régions, les phénotypes sont diverses. L'équipe s'intéresse à l'exploration de la variabilité intra-spécifique des génomes. L'objectif de la recherche est d'avoir une vue globale sur l'évolution des génomes nucléaires et mitochondriaux, et déterminer les règles qui gouvernent les relations entre génotype et phénotype.

I.IIPRESENTATION DU STAGE

I.ii.a Présentation des levures

Les levures sont un ensemble de champignons monocellulaire (cellule unique). La levure est constinué d'une multitude d'organismes vivants, appelés scientifiquement "micro-organisme" (la cellule de levure n'est visible qu'au microscope). Les deux espèces: Saccharomyces cerevisiae est utilisée en boulangerie, brasserie et vinification. La fermentation de Saccharomyces cerevisiae est réalisée dans l'environnement avec oxygène, par contre l'espèce Saccharomyces kluyveri peut le faire sans oxygène, il est plus efficace d' utiliser glucose pour la production d' énergie, c'est pour cela qu'il est utilisé pour les applications industrielle, par exemple fabrication de protéines. La reproduction pour les deux espèces est par bourgeonnement, S. Kluyveri est cependant plus ancien que S. cerevisiae. Avant la duplication de génome, il existé déjà. Durant de mon stage, mon travail était effectué sur l'espèce S. Kluyveri

I.ii.b Les significations de l'étude la levure

La levure se place au premier rang des modèles biologique se rapprochant de la cellule humaine au niveau de son organisation. Pour cela, nous lui devons à ce jour une part importante de nos connaissances sur le fonctionnement cellulaire des eucaryotes – humains, animaux, et végétaux.

I.ii.c Les avantage d'étude de la levure

Il y a beaucoup des avantages à étudier la levure:

Elle est unicellulaire, donc elle est facile et économique à manipuler.

Elle se produit vite (duplication en à peine plus de deux heures dans la plupart des cas)

Elle est le premier microorganisme eucaryote (qui contient un noyau) dont le génome ait été intégralement séquencé.

I.ii.d L'étude la levure

Comme les hommes, elles sont toutes différentes entres elles, on observe des variations de phénotypes, donc on veut chercher les causes de phénotypes – génotype, essayer de trouver la relation entre le phénotype et le génotype. Dans la prochaine partie je parlerais plus précisément du génotype et du phénotype .

I.ii.e Planification de stage

Les graphes tracés par le logiciel R nous permettent une approche significatives des résultats. On peut trouver l'information que l'on veut grâce aux graphes générés par R. Lors de mon stage, j'ai utilisé R pour les graphes . On procède en deux étapes.

Etape 1. Etude le phénotype chez la levure

Etape 2. Etude la relation entre phénotype et génotype

Au début de mon stage, j'ai essayé de comprendre le travail de l'équipe, le vocabulaire propre à la biologie, et effectué des graphes simple pour me familiariser.

La suite de mon stage portait sur l'utilisation du logiciel R, plus précisément avec le package QTL. Grâce à ce package, il est possible d'établir la relation entre phénotype et génotype. Ainsi durant mon stage, j'ai appris à effectuer des applications théoriques de statistique à biologiques, et j'ai trouvé cela impressionnant et très intéressant.

I.III PRESENTATION DES DONNEES

En totalité, j'ai reçu 3 fichiers des données sous la forme "xls"

Un fichier contient le phénotype de toutes les souches de collectées . Il y a 63 souches sous 28 conditions. Les différentes conditions testées, peuvent-être environnementales (variations de température, de pH...) et chimiques (sels, drogues...).

Les deux restes fichiers sont des fichiers confus avec des phénotypes et génotypes. Il contient les positions de marqueurs, le génotype de marqueur, le chromosome de marqueur, le phénotype etc. Il y a une différence entre les deux fichiers, l' un contient le chromosome C mais l'autre non.

II ETUDE DE PHENOTYPE DE CHEZ LA LEVURE

II.IDEFINITION DU PHENOTYPE

Le phénotype est l'ensemble des caractères observable d'un individu (la couleur de fleur, la forme de pois etc.).

Une des possibilités de phénotype pour la levure, c'est la croissance de la levure.

En effet, la vitesse de la croissance est observable au travers des étapes suivantes:

1. On met les levures et le liquide dans chaque point rond de plaque 96. Chaque condition pour une souche est testée en 4 réplicas. Grâce à la machine RoTor, il peut mettre tous les contenant de plaque 96 dans un plaque de 384, observant les 4 réplicas le même temps. RoTor

détecte également la surface de la levure, qui est un indicateur de croissance. Pour éviter l'erreur, la croissance est mesurée deux fois: une fois 24h, une fois 48h.

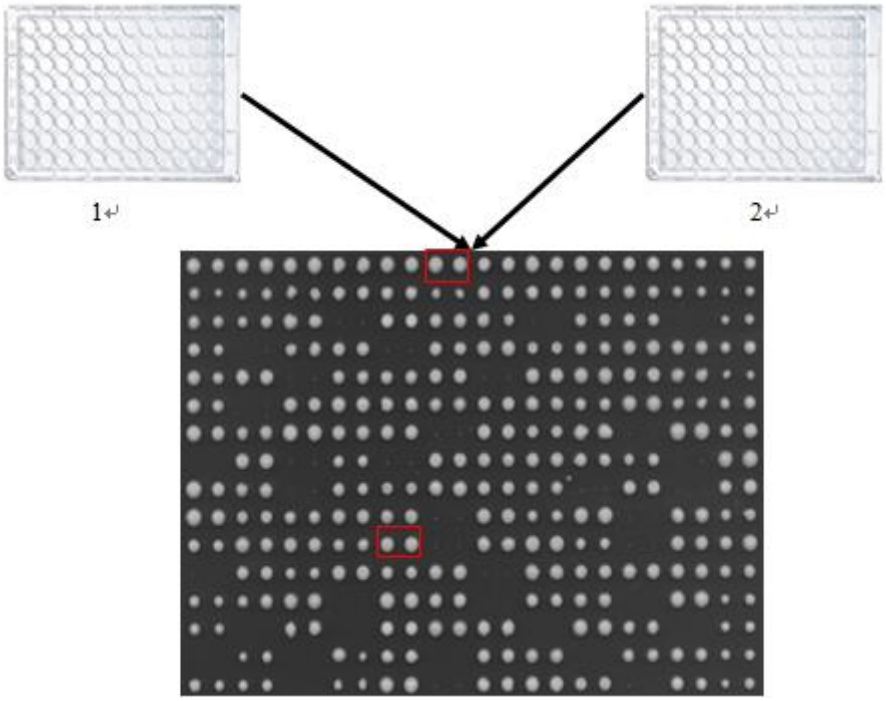


Figure II-1 la plaque 96(blanc) et plaque 384(noire)

2. Pour les souches différentes, la vitesse de croissance sont distinctes. En normalisant par rapport à la mesure en milieu YPD (milieu complet pour la croissance de la levure, chaque type de souche a son unique valeur de YPD), on peut étudier l'influence de la condition pour la levure, excluant le facteur de type de souche. Comme la normalisation de matrice, les valeurs sont sur une même base. On peut considérer que la croissance de la levure concerne que la condition.

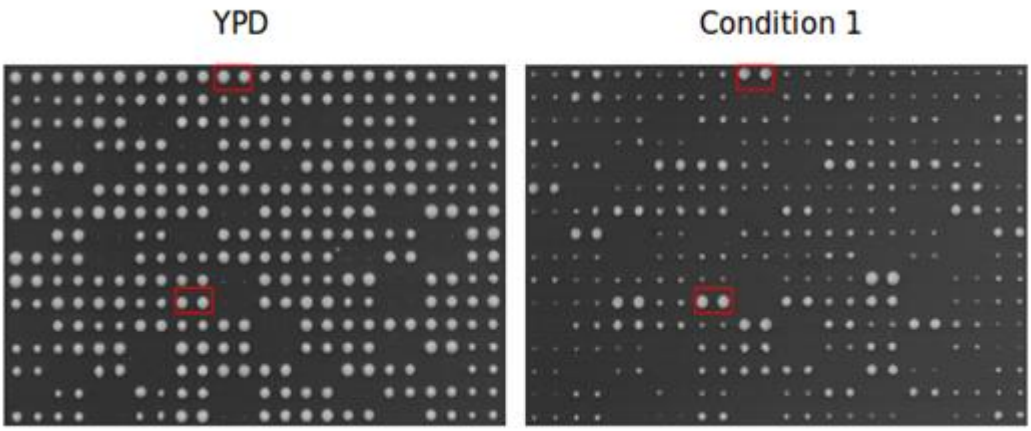


Figure II-2 LA COMPARAISON DE YPD ET UNE CONDITION

3. On prend la moyenne de 4 réplicas comme la vitesse de croissance.

4. Les vitesses de croissances sont données ci-dessous

	A	B	C	D	E	F	G	H	I	J	K
1		25°C		37°C		5FU_10-4M		5FU_10-5M		6azauracil_1mg/ml	
2		24h	48h	24h	48h	24h	48h	24h	48h	24h	48h
3	3	0.603318250	0.694444444	0.886917960	0.661157025	0.647249191	0.667779633	1.075268817	1.133144476	0.995024876	0.856531049
4	5	0.541516245	0.653594771	0.764818356	0.563380282	0.615384615	0.650406504	1.433691756	1.384083045	0.838574423	0.705882353
5	6	0.687285223	0.756143667	0.612557427	0.451467269	0.636942675	0.608828006	1.230769231	1.230769231	0.890868597	0.773694391
6	11	0.553505535	0.632911392	0.496688742	0.489396411	0.665557404	0.731261426	1.652892562	1.574803150	0.696055684	0.620155039
7	14	0.609756098	0.699300699	0.930232558	0.826446281	0.470035253	0.463499421	1.265822785	1.246105919	0.760456274	0.645161290
8	17	0.584795322	0.667779633	0.592592593	0.451467269	0.523560209	0.486618005	1.515151515	1.470588235	0.638977636	0.582524272
9	19	0.578034682	0.650406504	0.641025641	0.461538462	0.557620818	0.600600601	1.346801347	1.294498382	0.909090909	0.767263427
10	25	0.605143722	0.667779633	0.651465798	0.503144654	0.590841950	0.567107750	1.384083045	1.365187713	0.925925926	0.817438692
11	40	0.621118012	0.705467372	0.696864111	0.573065903	0.503778338	0.440044004	1.342281879	1.337792642	0.875273523	0.756143667
12	41	0.615384615	0.677966102	0.840336134	0.700525394	0.435729847	0.334821429	1.449275362	1.438848921	0.769230769	0.663349917
13	43	0.594059406	0.639658849	0.940438871	0.840336134	0.793650794	1.126760563	1.261829653	1.298701299	0.995024876	0.894854586
14	55	0.634920635	0.720720721	0.759013283	0.621118012	0.726392252	0.813008130	1.095890411	1.156069364	0.803212851	0.705467372
15	59	0.675675676	0.725952813	0.686106346	0.487210719	0.627943485	0.589101620	1.234567901	1.242236025	1.005025126	0.852878465
16	61	0.574712644	0.709219858	0.662251656	0.512820513	0.554785021	0.621118012	1.25	1.212121212	0.900900901	0.78125
17	63	0.642054575	0.694444444	0.636942675	0.427807487	0.578034682	0.606060606	1.438848921	1.444043321	0.847457627	0.724637681
18	64	0.552486188	0.639658849	0.828157350	0.770712909	0.446927374	0.468384075	1.465201465	1.544401544	0.765306122	0.684931507
19	65	0.619195046	0.679117148	0.589101620	0.425985091	0.637958533	0.582241630	1.498127341	1.454545455	0.941176471	0.793650794
20	66	0.686106346	0.666666667	0.666666667	0.527704485	0.698080279	0.634920635	1.428571429	1.680672269	0.894854586	0.773694391
21	67	0.599700150	0.662251656	0.604229607	0.431499461	0.601503759	0.581395349	1.337792642	1.342281879	0.943396226	0.8
22	68	0.681431005	0.763358779	0.617283951	0.500625782	0.637958533	0.655737705	1.346801347	1.311475410	0.890868597	0.8
23	69	0.558659218	0.642398287	0.626959248	0.584795322	0.654664484	0.701754386	1.520912548	1.486988848	0.8	0.753295669
24	70	0.688468158	0.761904762	0.613496933	0.498132005	0.574712644	0.557880056	1.294498382	1.294498382	0.847457627	0.756143667
25	71	0.662251656	0.792079208	0.722021661	0.687285223	0.828157350	1.055408971	1.286173633	1.294498382	0.865800866	0.809716599
26	77	0.630914826	0.757575758	0.943396226	0.851063830	0.657894737	0.722021661	1.186943620	1.294498382	0.917431193	0.860215054

Figure II-3Les souches et ses vitesses de croissances

En ligne, ce sont des différentes souches.

En colonne, ce sont les noms des conditions.

II.II REPRESENTATION GRAPHIQUES

II.ii.a Les boîtes à moustaches

Une boîte à moustaches est un moyen pratique et compact de visualiser la distribution d'une variable. Considérons une colonne ou une ligne comme une variable, on peut tracer la boîte à moustaches et observer

- La distribution de chaque condition sous toutes les souches
- La distribution de chaque souche sous toutes conditions

La boîte à moustaches pour une condition (25 ℃)

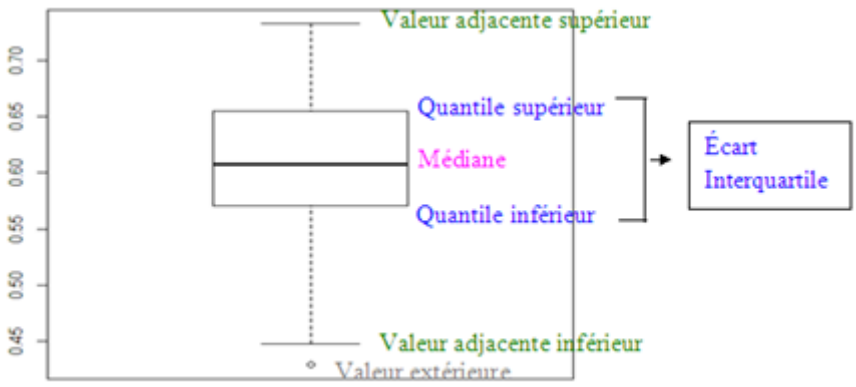


Figure II-4 l'explication de boîte à moustaches

Ecart Interquartile

Ecart Interquartile montre la vitesse de croissance comprise entre le 25e (Quantile supérieur≈0.66) et le 75e (Quantile inférieur≈0.57) percentile

Valeur Adjacentes

Les deux "moustaches" des deux côtés de la boîte montrent les valeurs adjacentes; Elles sont prévues pour représenter la maximale et la minimum de la vitesse de croissance, mais n'atteignent pas toujours la valeur maximale ou minimale.

Valeur adjacente supérieur

=La plus grand de vitesse de croissance de quantile supérieur

+1.5 * longueur de l'écart interquartile

Valeur adjacente inférieur

=La plus petite de vitesse de croissance de quantile inférieur

-1.5 * longueur de l'écart interquartile

Médiane

La ligne dans la boîte représente la médiane. Pour 63 souches, le 32e tombe dans cette valeur.

Valeur Extérieur

La vitesse de croissance en dehors de l'intervalle des moustaches, elle est dessinée individuellement.

On peut encore mettre toutes les boîtes à moustaches ensemble et on peut comparer toutes les distributions de condition, et regarder la variation de vitesse de croissance.

On regarde un résultat de graphe qui contient plusieurs boîtes à moustaches。

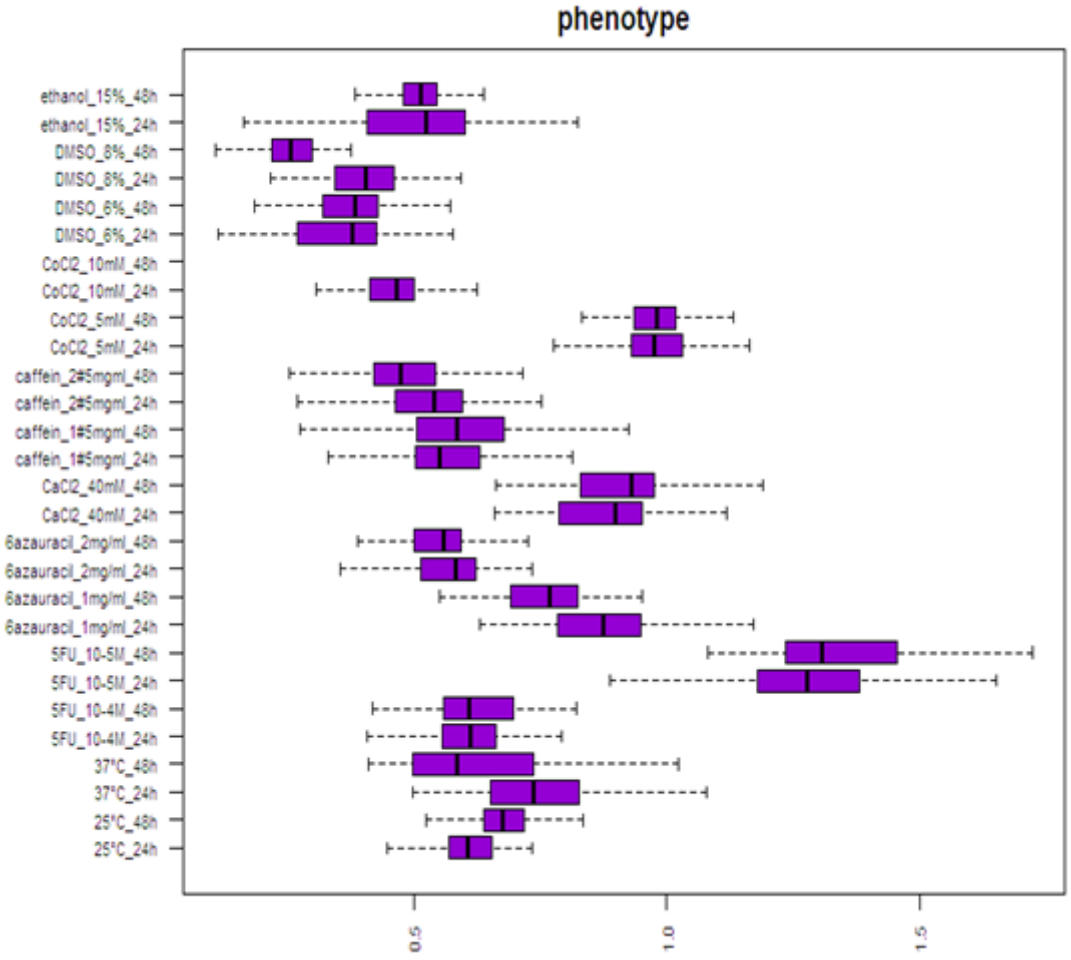


Figure II-5Les boîtes à moustaches sous toutes les conditions

Par graphe, on trouve que les vitesses de croissances sous chaque condition sont différentes. Les boîtes à moustaches montrent la variation de phénotype.

II.ii.b Heatmap

Quand on rencontre des données de matrice, on le remplace par les couleurs qui correspondent les valeurs de données. On voie que le Heatmap est une image de matrice avec les différents niveaux des couleurs. Les plus petites valeurs sont représentées par les couleurs foncées, par contre, les plus grandes sont plus claires. Avec Heatmap on trouve facilement les grandes et les petites valeurs.

Pour chaque souche, les vitesses de croissances sont différentes, on veut observer les coefficients de corrélation entre toutes les souches, par exemple, pour souche 1, calculant le coefficient de corrélation entre souche 1 et souche 2, souche 1 et souche 3, jusqu'à souche 1 et souche 63. On met les résultats dans la première ligne de matrice. On répète ce processus 63 fois, et on obtenait une matrice de 63*63.

De plus, il y a trois types de coefficient de corrélation, Pearson, Kendall et Spearman. Pearson est le calcul pour le modèle linéaire, et par contre les deux autres, sont pour le modèle non-linéaire.

La distribution entre souche semble linéaire. Donc on choisit Pearson comme coefficient de corrélation.

Le coefficient de corrélation entre deux populations X et Y est

$$r_p = \frac{\sum_{i=1}^{N}(x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2} \cdot \sqrt{\sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

Où N est le nombre total de chaque population,

Xi et Yi sont des individus de deux populations,

$\overline{x}$ et $\overline{y}$ sont des moyennes de deux populations

Maintenant on peut tracer le Heatmap de cette matrice, étudions les relations entre chaque souche qui devient visuelles.

III L'ETUDE DE LA RELATION ENTRE PHENOTYPE ET GENOTYPE

Pour l'étude de la relation entre phénotype et génotype, comparons les données précédentes, ils ont changés un peu, maintenant ils contiennent les phénotypes, et l'information de génotype (par exemple, le nom de chromosome, la position de marqueur etc.), les prochaine parties on les utilise. Pour commencer, connaissons-nous plusieurs notations de biologique.

III.I DEFINITION DE GENOTYPE

Opposition au phénotype, génotype est l'ensemble ou une partie donnée de la composition génétique d'un individu. Le phénotype détermine les caractères d'un individu, ce qui constituant le phénotype. Les génotypes entre les souches sont différents, donc on voit la variation des phénotypes (les vitesses de croissance sont distinctes).

III.II QUELQUES NOTIONS DE BIOLOGIQUES

Caractère Qualitatif

Lorsqu'un caractère est dit qualitatif, les diverses variantes de ce caractère dans une population se distinguent nettement les unes des autres (variation discontinue) et sont en nombre réduit, par exemple la couleur de fleur, le groupe sanguin etc. Un caractère qualitatif est gouverné par un seul gène et s'exprime indépendamment du milieu. On peut calculer le percentage de dépendants pour chaque caractère.

Caractère Quantitatif

Lorsqu'un caractère est dit quantitatif lorsque les diverses variantes de ce caractère dans une population sont difficilement classifiables (variation continue) et sont en très grand nombre (au point que chaque variante ne décrit qu'un faible nombre d'individus), par exemple, le rendement de production, période de maturation, la longueur de laine etc. Un caractère quantitatif dépend généralement de plusieurs gènes et est très affecté par le milieu. Un caractère phénotypique varie par degrés s'appelle l'hérédité de caractère quantitative.

Chromosome

Le chromosome est constitué de molécules d'ADN et de protéines. Donc il est le support de l'information génétique.

Pour la levure, l'espèce S. Cerevisiae a 16 paires de chromosomes, par contre l'espèce S. Kluyveri à 8 paires de chromosomes.

Marqueur

Le marqueur est un emplacement connu sur un chromosome. Il peut être utilisé pour identifier des individus et des espèces.

Les diverse souches de même espère ont des différents marqueurs, peut-être c'est une raison de variation de phénotype d'une espèce.

Un exemple simple,

Souche 1 BHAC

Souche 2 BAAC

On veut regarder si la partie HAC ou HA influe le phénotype, on ne peut pas détermine juste un marqueur, c'est une région de chromosome, c'est la notion prochaine, QTL.

QTL

Quantitative Trait Loci (en anglais), locus de caractères quantitatifs est une région du chromosome où sont localisés un ou plusieurs gènes intervenant dans l'expression d'un caractère quantitatif.

Hybride

En génétique, l'hybride est un organisme issu du croisement de deux individus de deux variétés, sous-espèces (croisement inter spécifique), espèces (croisement inter spécifique) ou genres (croisement inter générique) différents. L'hybride présente un mélange des caractéristiques génétiques des deux parents.

Les levures sont le résultat d'un croisement entre deux variétés de même espèce (l'espèce S. Kluyveri).

Pour les données, Il y a 44 souches et 45 phénotype pour chaque souche--les chercheurs mettent les levures sous 23 conditions et observent la croissances de levures une fois 24h et une fois 48h. Il contient aussi les noms de marqueurs, les positions de marqueurs et lequel chromosome le marqueur appartient (Les données sont lues par R).

```
--Read the following data:
        44  individuals
        18325  markers
        45  phenotypes
--Cross type: bc
```

Parce que le chromosome C a impact sur les résultats, ensuite on manipule les deux différentes données -- avec chromosome C et sans chromosome C et compare les graphes obtenus.

III.III LE BUT DE TRACER QTL

C'est difficile de voir que quel génotype affecte quel phénotype, donc on veut chercher une méthode pour lier les deux.

L'idée fondamentale pour QTL, c'est étudier la relation du phénotype et génotype pour une population qui a les génétiques diverses. Dans le laboratoire, on étudie les génétiques associent la vitesse de croissance, c'est-à-dire on cherche l'existence de QTL sur le chromosome. S'il y a de QTL sur le chromosome, on peut supposer que cette partie de chromosome a un effet sur le phénotype.

LOD score est un indicateur de QTL, c'est une valeur numérique qui quantifie la liaison, en effet, en calculant le LOD score de chaque position sur chromosome, il est plus visuelle de montrer la relation entre phénotype et génotype sur chaque position.

III.IV ANOVA ET LOD SCORE

ANOVA (ANalyse Of VAriance) est un test statistique permettant de vérifier que plusieurs échantillons sont issus d'une même population.

Autrement dit, on peut regarder si les moyennes de chaque échantillon sont identiques. Le test peut montrer que si le groupe est un facteur influe les individus.

LOD score est un test qui ressemble à ANOVA, ou on peut dire que c'est le théorique de statistique test applique à la génétique. Pour un marqueur, on divise les individus en groupe par rapport le génotype de marqueur, ensuite on compare si les moyennes de phénotype (vitesse de croissance) de chaque groupe sont identiques. On veut regarder si les différences de marqueurs causent la variation de phénotypes.

Pour un test, on veut focaliser deux choses: l'hypothèse et la statistique de test, grâce à la statistique, on peut décider d'accepter ou rejeter l'hypothèse.

III.iv.a Le modèle statistique

Après les données sont divisées par groupe, on peut commencer à étudier les moyennes de phénotype (vitesse de croissance).

Le modèle statistique d'ANOVA est écrit de la façon suivante:

La vitesse de croissance = β0 + β1 + erreur

III.iv.b Conditions à vérifier

Pour appliquer ANOVA, il y a trois conditions à vérifier

1) Indépendance d'observateurs
2) Normalité des erreurs
3) Homoscédastique (erreur ont les même variances)

L'indépendance d'observateur est difficile de tester, mais pour la normalité et homoscedasicité, ils existent des tests à vérifier.

Sur logiciel R, le test de Shapiro-Wilk permet de tester la normalité, le test de Bartlett est possible de tester l'égalité de plusieurs variances.

Analogiquement, on regarde les conditions à vérifier pour LOD score

1) Indépendance
2) Normalité
3) Homoscédastique

Jusqu'à maintenant, il n'existe pas de test simple permet d'étudier l'indépendance.

La normalité n'est pas assez importante. L'analyse de QTL est robuste remarquable, les résidus sont supposé de suivre la loi Normale, donc le test de normalité n'est pas utile.

L'homoscédastique a un grand effet. Mais au lieu de prendre les tests, on veut regarder l'histogramme de phénotype. Pour le moment, on veut voir la distribution de phénotype. Selon l'expérience, la distribution asymétrique (skewness en anglais, figure ci-dessous) cause la non-constant de variance. Donc si la distribution est asymétrique, la condition de l'homoscédastique ne peut pas vérifier.

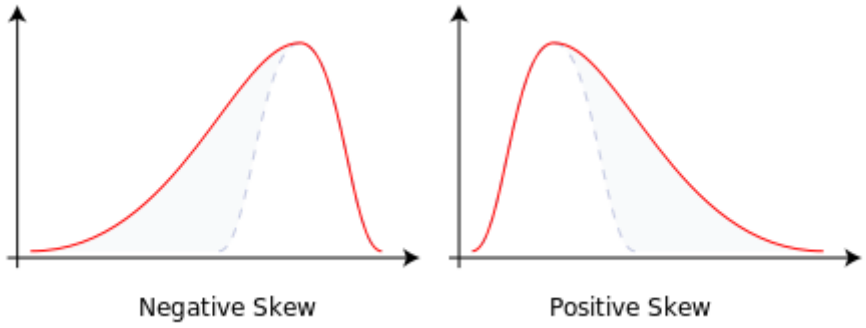


Figure III-1 distribution asymétrique

III.iv.c Hypothèse

L'hypothèse d'ANOVA:

H0: Les moyennes de tous les échantillons sont identiques

Les individus sont issus de même population

H1: Les moyennes de tous les échantillons ne sont pas toutes identiques

Les individus ne sont pas issus de même population

Pareillement, l'hypothèse de LOD score:

H0: Les phénotypes de chaque groupe sont égaux

Il n'y a pas de QTL

H1: Les phénotypes de chaque groupe ne sont pas toutes identiques

Il existe un QTL sous le test de marqueur

III.iv.d Statistique

On rejeté ou accepte l'hypothèse, ça dépend de la statistique que l'on calcule.

La statistique F est pour le test ANOVA, si les trois conditions sont satisfaites et si l'hypothèse nulle est vraie, alors

Supposons qu'on a n individus

M est le nombre d'échantillons

Degré de liberté entre échantillon: m -1

Degré de liberté intra-échantillon: n-m

i est l'indicateur de l'échantillon

m_i est le nombre de l'individu dans l'échantillon i

j est l'indicateur de l' individu dans l' échantillon

y_{ij} est individu j dans l' échantillon i

$\bar{y}$ est la moyenne empirique de tous les individus

$\bar{y}i$ est la moyenne empirique de l' échantillon i

Somme carré total: $SC_{Tot} = \sum_{i=1}^{m}\sum_{j=1}^{m_i}(y_{ij} - \bar{y})^2$

Somme carré due au facteur: $SC_F = m_i \sum_{i=1}^{m}(\bar{y}_i - \bar{y})^2$

Somme carré de résidus: $SC_R = \sum_{i=1}^{m}\sum_{j=1}^{m_i}(y_{ij} - \bar{y}_i)^2$

La relation: $SC_{Tot} = SC_F + SC_R$

$$F_{obs} = \left(\frac{\text{variance entre groupe}}{\text{variance intra-groupe}}\right)$$
$$= \left(\frac{SC_{Tot} - SC_R / m-1}{SC_R / n-p}\right)$$
$$= \left(\frac{SC_F}{SC_R}\right)\left(\frac{n-p}{m-1}\right)$$

On compare le Fobs avec le F (m-1, n-m), si Fobs est plus grand que F. On rejette l'hypothèse nulle et on suppose que les moyennes de tous les échantillons ne sont pas égales et les échantillons

sont issue de populations différents .Si la variance entre échantillon est plus grande, la valeur de F est plus grande, le test est plus significatif.

Ensuite on regarde la statistique de QTL, il s'appelle LOD Score.

yi est le phénotype de l' individu i,

$\bar{y}$ est le moyenne empirique de tous les phénotypes.

$$RSS_0 = \sum_i (y_i - \bar{y})^2$$

g_i est le génotype de l'individu i sur le marqueur

$\bar{y}_{gi}$ est la moyenne empirique de génotype g_i

$$RSS_1 = \sum_i (y_i - \bar{y}_{g_i})^2$$

Alors

$$LOD = \frac{n}{2} \log_{10}\left(\frac{RSS_0}{RSS_1}\right)$$

Si LOD score est plus grand, le test est plus significatif. On rejette l'hypothèse nulle et on suppose que il y a un QTL sur le marqueur que l'on a testé. En général, si LOD score est supérieur ou égale 3, on peut supposer l'existence de QTL.

On peut encore regarder la relation entre F et LOD score

$$F = \left(\frac{RSS_0 - RSS_1}{RSS_1}\right)\left(\frac{n - df - 1}{df}\right)$$
$$= \left(\frac{RSS_0}{RSS_1} - 1\right)\left(\frac{n - df - 1}{df}\right)$$
$$= \left(10^{\frac{2}{n}LOD} - 1\right)\left(\frac{n - df - 1}{df}\right)$$

Pour l'hybride, le df=2

Il est intéressant d'étudier la formule inverse

$$LOD = \frac{n}{2}\log_{10}\left[F\left(\frac{df}{n - df - 1}\right) + 1\right]$$

Un résultat de LOD Score sous conditions DMSO 24h:

Premièrement, on regarde la distribution de DMSO 24h

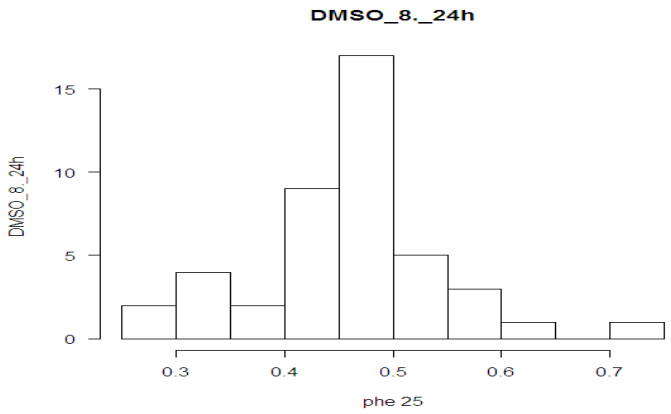


Figure III-2La distribution de DMSO 8% 24h

Elle est normalité et elle n'est pas asymétrique, donc j'ai appliqué à la méthode LOD Score et obtenu le résultat suivant:

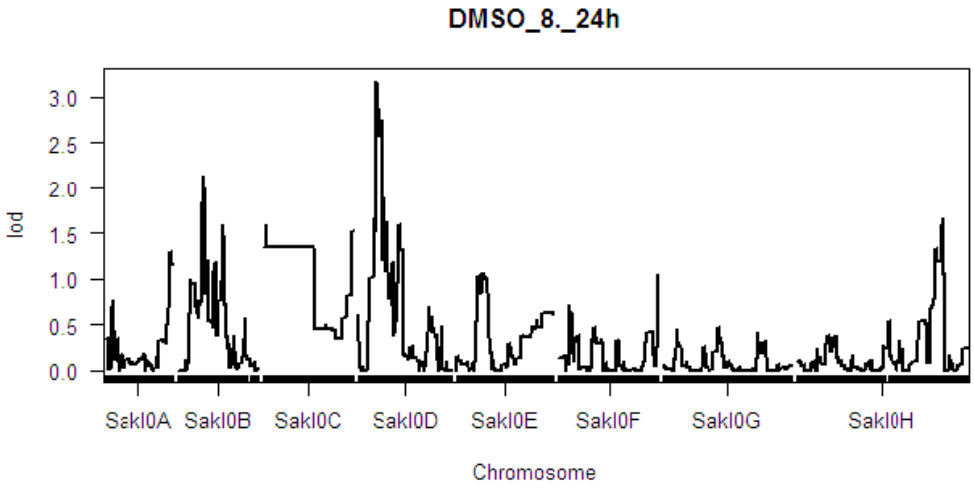


Figure III-3 LOD score de la levure sous condition DMSO 8% 24h

Sur le graphe, on voit le LOD Score sur chaque position de chromosome, et sur chromosome D, la valeur de LOD Score est supérieure à 3, donc on rejette l'hypothèse H0, et on suppose l'existence de QTL.

III.V KRUSKAL-WALLIS ET NON-PARAMETRE DE LOD SCORE

Kruskal-Wallis est un test non-paramètre qui est équivalent au test ANOVA. Comme ANOVA, il permet de vérifier que plusieurs échantillons sont issus de même population. Par contre, il ne compare pas les moyennes de chaque échantillon.

Non-paramètre de LOD Score est équivalent à Kruskal-wallis dans le domaine génétique.

III.v.a Condition à appliqué

Différents que le test ANOVA, Kruskal-Wallis ne suppose pas la distribution des échantillons ne suit la loi normale, donc on le utilise dans ce cas, ou encore quand les trois conditions fondamentales de ANOVA ne peuvent pas vérifier.

Pour non-paramètre de LOD Score, la distribution de phénotype a la forme de skewness, dichotomie ou il existe des points extrêmes, on applique la méthode.

III.v.b Condition à vérifier

Comme précédent, les échantillons doivent indépendant. De plus, pour chaque échantillon, le nombre de l'observateur doit être égal ou supérieur 3.

III.v.c Hypothèse

L'hypothèse de Kruskal-Wallis

Supposons qu'on a k échantillon

H0: L1(X)=L2(X)=...=Lk(X)

Les échantillons sont issus de même population

H1: Les lois L1(X), L2(X), Lk(X) ne sont pas toutes identiques

Les échantillons ne sont pas issus de même population

L'hypothèse de LOD Score

H0: Les phénotypes de chaque groupe sont égaux

Il n'y a pas de QTL

H1: Les phénotypes de chaque groupe ne sont pas égaux

Il existe un QTL sous le test de marqueur

III.v.d Statistique

Premièrement, on regarde le statistique K de Kruskal-Wallis (absence d'ex aequo)

Supposons qu'on a k échantillons indépendants X_1, X2...X_k

n est le nombre de individus

n_i indique le nombre de l'observateur dans le groupe i

R_{ij} est le $j^{ème}$ observateur de groupe i (le groupe est rangé par ordre croissant)

La somme de rang de par rapport à chaque échantillon: $$R_i = \sum_{j=1}^{n_i} r_{ij}$$

$\bar{R}_i$ est le moyen de somme de rang d'échantillon i

$$K = \frac{12}{n(n+1)} \sum_{i=1}^{k} n_i \left(\overline{R}_i - \frac{n+1}{2} \right)^2$$
$$= \frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^{\,2}}{n_i} - 3(n+1)$$

Si l'un de l'effectifs n_i est égal ou inférieur 4, on compare K avec la table de Kruskal-Wallis, pour un seuil α donné, le table nous offre une valeur critique Cα, on rejette l'hypothèse H0 quand K est plus grand et on suppose que tous les échantillons ne sont pas issue de même population.

Si tous les n_i est égale ou supérieur 5, $K \approx \chi^2 (k-1)$, on compare K avec la table de Khi-deux de degré K-1, pour un seuil α donné, la table nous offre une valeur critique Cα, on rejette l'hypothèse H0 quand K est plus grand et on suppose que tous les échantillons ne sont pas issue de même population.

Sous l'hypothèse H0, on suppose $W_i = n_i \overline{R}_i$ i=1...k

On peut déduire

$$E(W_i) = n_i(n+1)/2 \quad \text{Et} \quad Var(W_i) = n_i(n-n_i)(n+1)/12$$

Le statistique K peut s'écrire de la façon ci-dessous

$$K = \frac{1}{n} \sum_{i=1}^{k} (n - n_i) \frac{(W_i - E(W_i))^2}{Var(W_i)}$$

Ensuite, c'est la statistique de non-paramètre LOD Score

On range le phénotype par ordre croissant, soit y_i est le phénotype de l'individu i, R_i est le rang de l'individu i.

On suppose qu'il y a plusieurs position M_i peuvent exister QTL, alors la probabilité quand le génome est j pour l'existence de QTL à la position Mi est:

$$p_{ij} = \Pr(g_i = j | M_i)$$

La somme de rang de l'échantillon j: $S_j = \sum_i p_{ij} R_i$

$E(S_j)$ est espérance de S_j et $Var(S_j)$ est variance de S_j sous hypothèse H0, alors

$$H = \frac{1}{n} \sum_j \left(\frac{n - \sum_i p_{ij}}{n} \right) \left[\frac{(S_j - E(S_j))^2}{Var(S_j)} \right]$$

Comme précédent, en général, on suppose l'existence de QTL quand K est égal ou supérieur à 3.Un résultat de non-paramètre de LOD Score sous conditions 37°24h:

Premièrement, on voit la distribution de la condition 37°24h

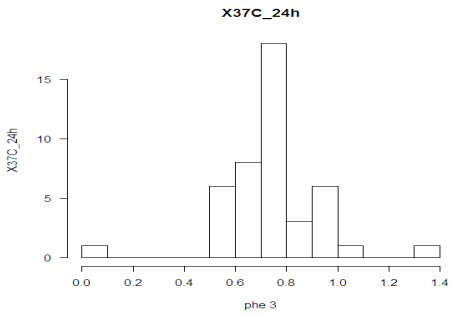


Figure III-4 la distribution de la condition 37°24h

Évidemment, elle n'est pas normalité, donc j'ai pensé de utiliser la méthode non paramètre et obtenu le résultat suivant:

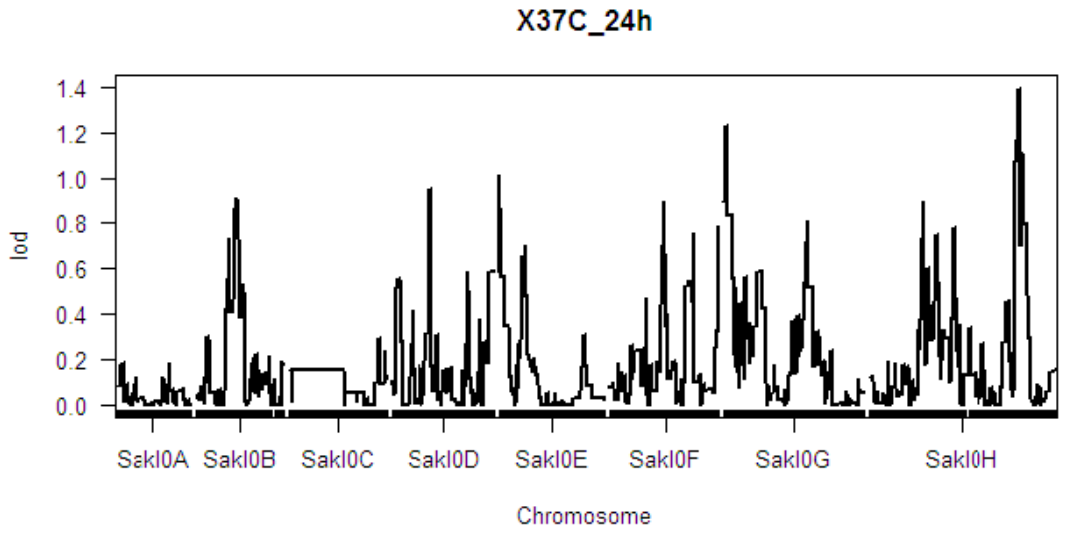


Figure III-5LOD score sous condition 37°24h

Sur le graphe, on voit que le LOD score ne dépasse pas 1.5, donc on accepte l'hypothèse H0: Il n'y a pas de QTL.

III.v.e Autre Méthodes

Si on ne veut pas utiliser le méthode Kruskal-Wallis, on a l'autre choix de reprendre le test paramètre -- transformer la forme de phénotype.

Souvent, on peut prendre la racine carrée, log, ou racine carré de log (En pratique, on peut prendre la racine carrée de log quand la distribution est skewness).

III.VI CONCLUSION DE DEUX TESTS

On utilise ANOVA au lieu de t-test et Kruskal Wallis au lieu de Wilcoxon-Mann-Whitney, parce que la propriété de génotype décide que le nombre d'échantillon est plus que 2. De plus, les deux phénotypes ne peuvent pas identiques, on choit la statistique absence d'ex aequo.

Une limitation de ANOVA et Kruskal-Wallis, on peut dire que les individus ne sont pas issue de même groupe, mais on ne sait pas quels échantillon causent les différents. Ce n'est pas un problème important pour LOD Score, parce qu'on veut juste savoir l'effet de groupe.

En effet, quand même la normalité ou l'homoscedasicité est enfreint, la méthode de LOD Score peut encore marcher. Je pense que n'importe la méthode que on choit, paramètre ou non-paramètre, la tendance de LOD Score ne change pas, de plus, l'existance de QTL est une région, on ne demande pas une position précise.

III.VII TEST DE PERMUTATION

On sait que si LOD score est supérieur ou égale 3, on déterminera l'existence de QTL, mais on peut obtenir une valeur précise, pour l'obtenir, on a besoin de considérer de chercher le QTL sur pan génomique.

Sous hypothèse nulle qu'il n'y a pas de QTL sur chromosome, on peut supposer :

Il n'y a pas un QTL sur une position particulière (hypothèse 1)

Il n'y a pas un QTL sur pan géométrie (hypothèse 2)

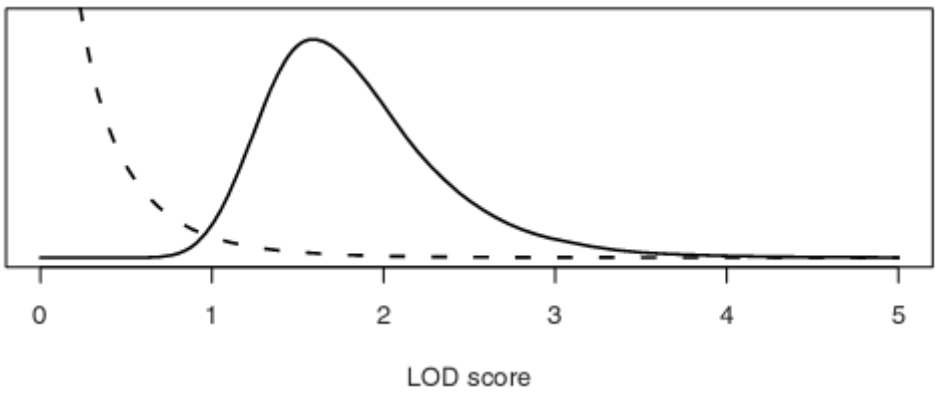


Figure III-6 La distribution de LOD Score de hypothèse 1(courbe en ligne pointillé) et hypothèse 2(courbe solide)

Sur le graphe, on a vu que pour hypothèse 1, principaux de LOD Score tombent dans la région inférieur à 1, ils sont petites comme les seuils de LOD score, par contre, à peu près 95% de LOD score de l'hypothèse 2 tombent dans le intervalle [1,3], donc ils peuvent considérer comme le seuil de LOD Score. Avec les données qu'on a obtenues maintenant, si on veut réaliser la recherche de QTL sur pan génétique, le test de permutation est nécessaire.

Test de permutation est un test non paramètre qui est utilisé vari é. Il bouleverse aléatoirement les données comme mélant les cartes. Parfois durant les expérimentaux, on a le problème du frais, du temps... On n'obtenait pas beaucoup des individus, en ce cas, on utilise le test de permutation pour prendre la correcte distribution sous hypothèse nulle. C'est un test plus computation intensive que les autres statistiques tests. Le test permutation est utilisé extensif dans le domaine génétique et chromosomique. Il est pratique quand les données sont insuffisantes .

Le principale de test de permutation est, les données de phénotype et génotype restent les même, mais lequel phénotype correspond lequel génotype sont changés. Une fois on fait un test de permutation, on trace le LOD Score et on peut obtenir un maximum de LOD Score, on répète ce

processus n fois, on obtenait n valeurs de maximum de LOD Score, et on détermine différents seuils par le α on choit.

C'est un graphe montre ce processus:

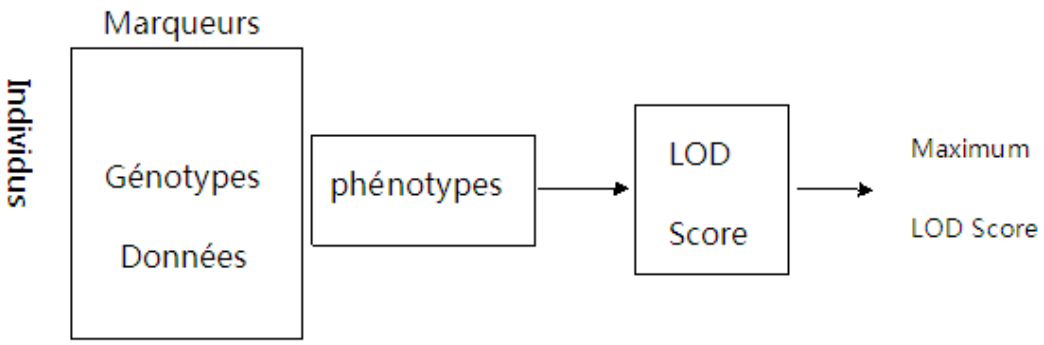


Figure III-7 Processus de test de permutation

Un exemple de test de permutation réalisé par logiciel R,

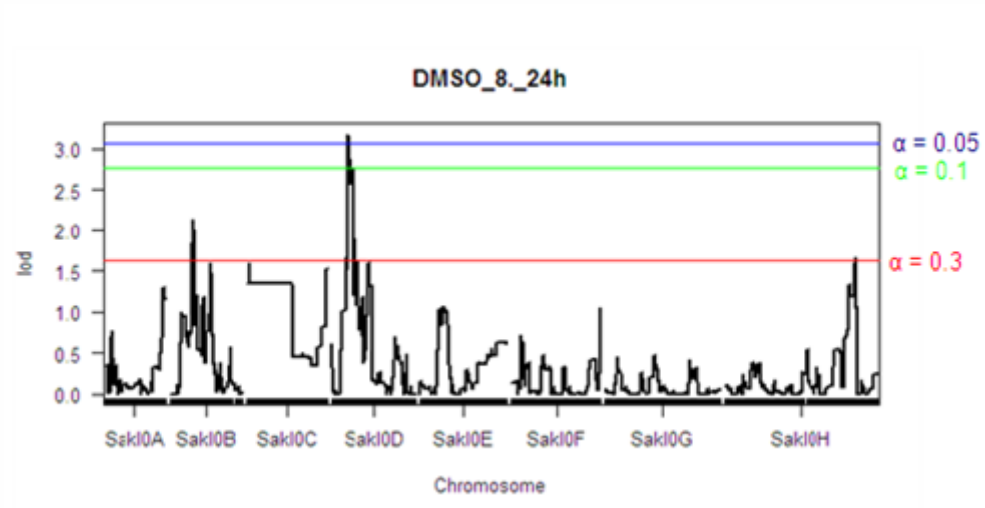


Figure III-8Les seuils de LOD score sous condition DMSO 8% 24h

On regarde les différents niveaux de seuils, les lignes rouge, verte, bleu montrent les seuils 0.3, 0.1 et 0.05. Souvent on focalise le seuil 0.1, la ligne verte, il y a des sens génétiques, 90% de maximum de Lod Score tombe à cette région, ça veut dire que si on fait le test de permutation 1000 fois, et 900 fois le maximum de LOD Score est supérieur à 2.7. De cette façon, on vue l'évidence de QTL.

III.VIII INTERVALLE DE CONFIDENCE DE QTL

Après le test de permutation, on peut commencer à déterminer un intervalle qui estimât QTL. Les deux méthodes: Lodint et Bayesient permettent de le réaliser. A ce moment, on suppose qu'il y a un et seulement un QTL sur le chromosome. On reprend l'exemple précédent, on a vu sous condition DMSO 8%, le LOD Score de chromosome D dépasse le seuil 0.1.

III.viii.a Lodint

Le 1.5-Lodint intervalle de confiance est un intervalle incluant 1.5 unité de maximum de LOD Score.

Si le LOD Score descend et remonte (comme figure ci-dessous), on a plusieurs intervalles disjoints, mais on utilise les intervalles consécutifs comme l'intervalle confiance.

Un exemple d'intervalle de confiance calculé par la méthode Lodint,

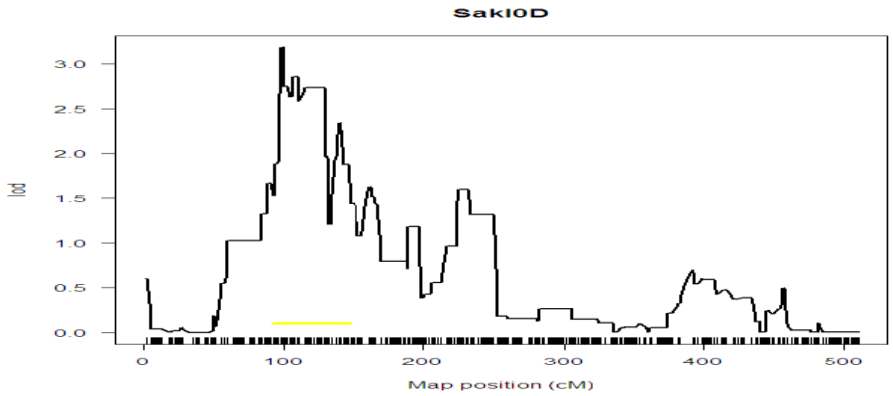


Figure III-9L'intervalle de confiance sous condition dmso 8% 24h chromosome D

	chr	pos	lod
Sakl0D 231175	Sakl0D	92.4700	1.526381
cSakl0D.loc144	Sakl0D	146.3800	1.670310

[92.4700,146.3800]

Le maximum de LOD Score est à peu près 3.1, et on a obtenu 1.6 après soustraire par 1.5. L'intervalle qui contient le LOD Score supérieur et égale 1.6 est entre 92.4700 et 146.3800.

III.viii.b Bayesint

Le théorème de Bayes

Les deux événements A et B, on veut calculer la probabilité de A sachant B, le formule est:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Autre écriture:

Si $\{A_j\}$ est une partion de l'ensemble des possibles,

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_j P(B|A_j)P(A_j)}$$

Ensuite, on regarde la méthode Bayesint,

Le 95% intervalle Bayesint est défini par la suivante, on veut chercher un intervalle I , pour $f(\theta|data)$ atteint le seuil et pour $\sum_{\theta \in I} f(\theta|data)$ supérieur ou égal 0.95.

Où $$f(\theta|data) = \frac{10^{LOD(\theta)}}{\sum_\theta 10^{LOD(\theta)}}$$

Un exemple d'intervalle de confiance calculé par la méthode bayesient

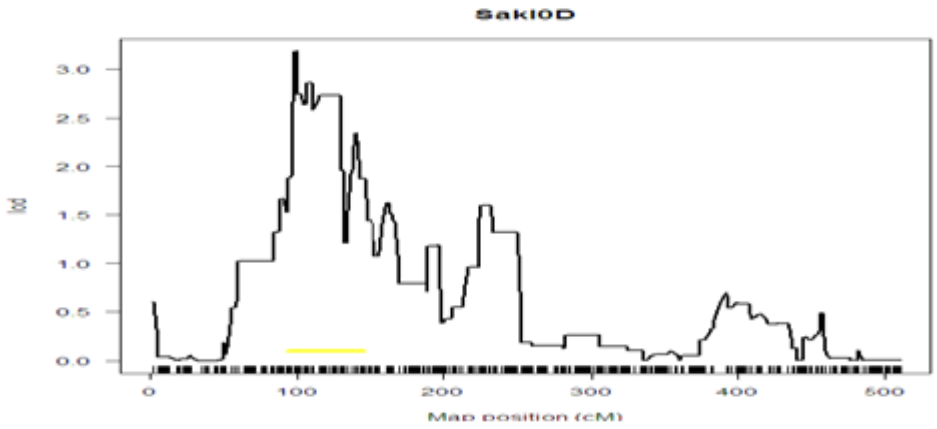


Figure III-10 l'intervalle de confiance sous condition DMSO 8% 24h chromosome D

	chr	pos	lod
Sakl0D 235364	Sakl0D	94.1456	1.883091
Sakl0D 362254	Sakl0D	144.9016	1.874354

[94.1456, 144.9016]

Après la calculation de l'ordinateur, on a obtenu l'intervalle [94.1456, 144.9016]. Les résultats de deux méthodes se ressemblent

Jusqu'à maintenant, on a déterminé l'intervalle de confiance de QTL, on a fini la recherche de la relation entre phénotype et génotype. L'existence de QTL est un facteur influent le phénotype.

IV LE PACKAGE QTL

R est un logiciel pour statistique computation et graphes. La version base de R propose la plupart des fonctionnalités utiles pour la statistique de base, mais si on veut quelque fonctions spéciales, on utilise les packages. Ils sont mis librement sur le site. R/qtl est un package qui serve dans le domaine génétique.

R/qtl est un extensif, interactif environnement pour tracer le QTL durant expérimental génétique. Un pivot de computation méthode pour le package est Modèle de Markov caché pour s'occuper les données perdues. Les programmeurs implémentent principale du algorithme.

La version actuelle de R/qtl permet de estimer la graphe génétique, identifier les erreurs génétiques, et tracer le graphe QTL.

V CONCLUSION

Premièrement, on observe le phénotype (la vitesse de croissance) de souches de l'espace S. Kluyveri grâce aux boîtes à moustaches et heatmap. On voit la variation de phénotype, on veut étudier les raisons qui causent la différence, ainsi que la relation entre phénotype et génotype chez S. Kluyveri.

Ensuite le graphe suivant montre la recherche de la relation. Le but est déterminant l'existence de QTL

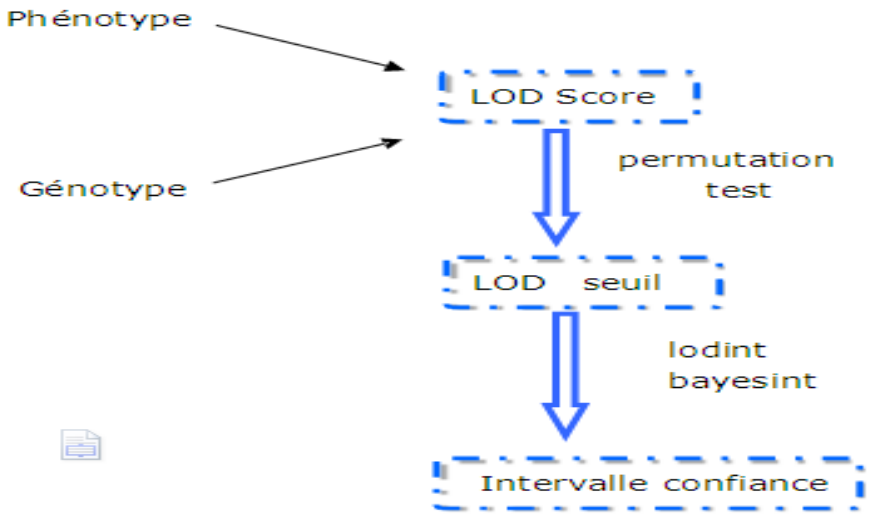


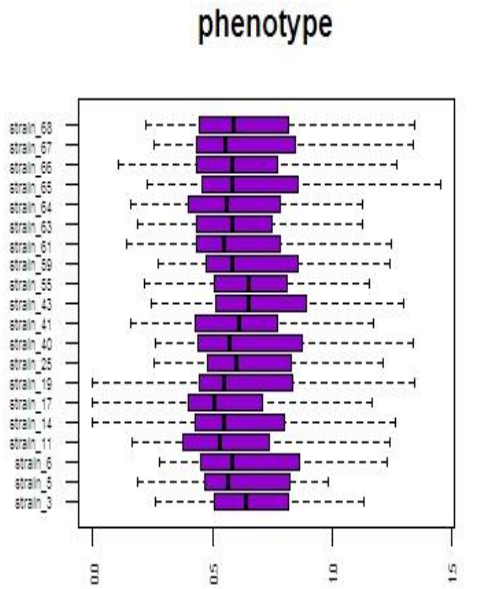
Figure V-1processus de stage

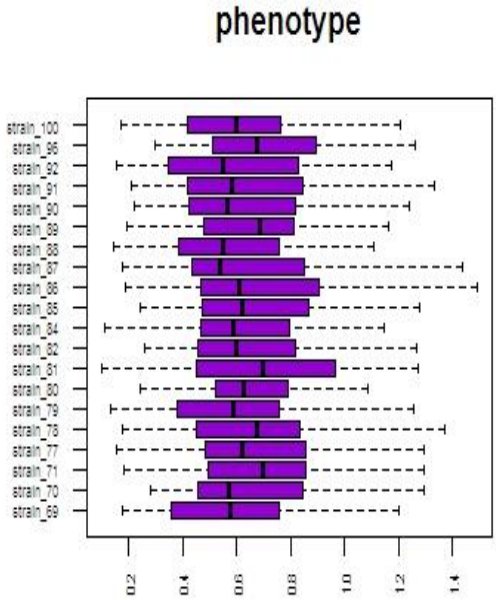
Avec les données de phénotype et génotype, on calcule le LOD Score de tous les marqueurs et trace le graphe. Ensuite on fait un test de permutation et obtenait le seuil de LOD Score. Finalement on cherche l'intervalle de confiance de QTL. Avec ce processus, on détermine la relation de phénotype et génotype.

Durant mon stage, j'ai pu appliquer les méthodes statistique vue en cours à des données réelles, et ainsi comprendre leur utilité tout en approfondissant mes connaissances. J'ai également appris une nouvelle méthode LOD Score qui est utilisé dans le domaine génétique et me familiariser avec le logiciel R. Cette expérience m'a donné une nouvelle vue du métier statisticien et de me permettra de mieux m'orienter professionnellement.

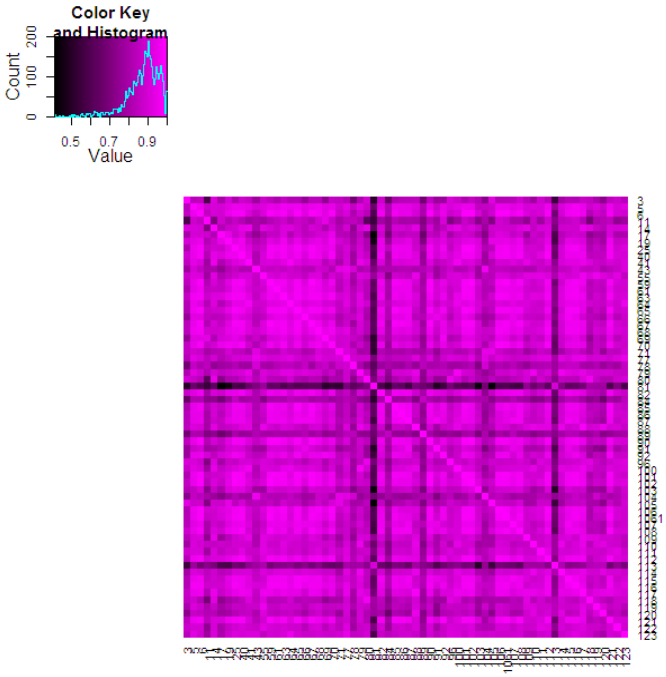
ANNEXES

V.I ANNEXE 1 boites à moustaches pour chaque souche

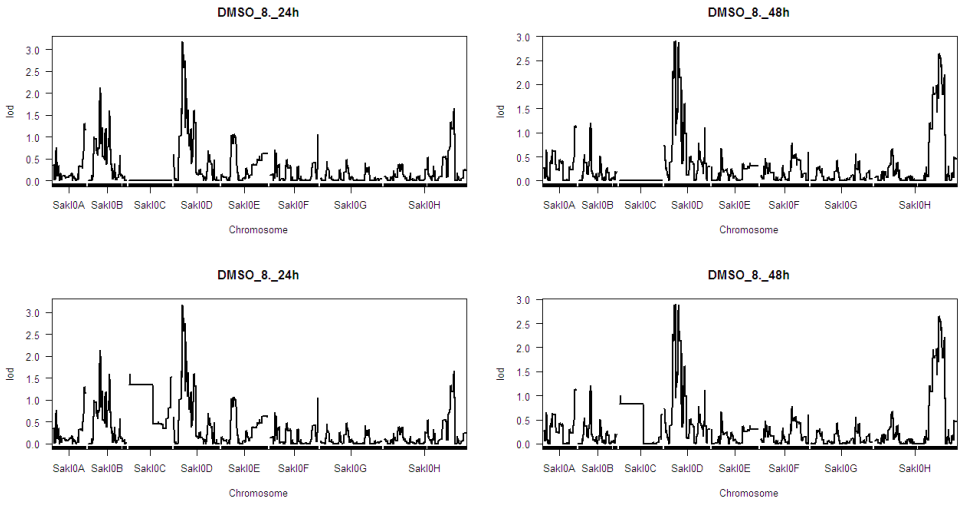




V.II Annexe2 heatmap sur la corrélation de pearson entre les souches



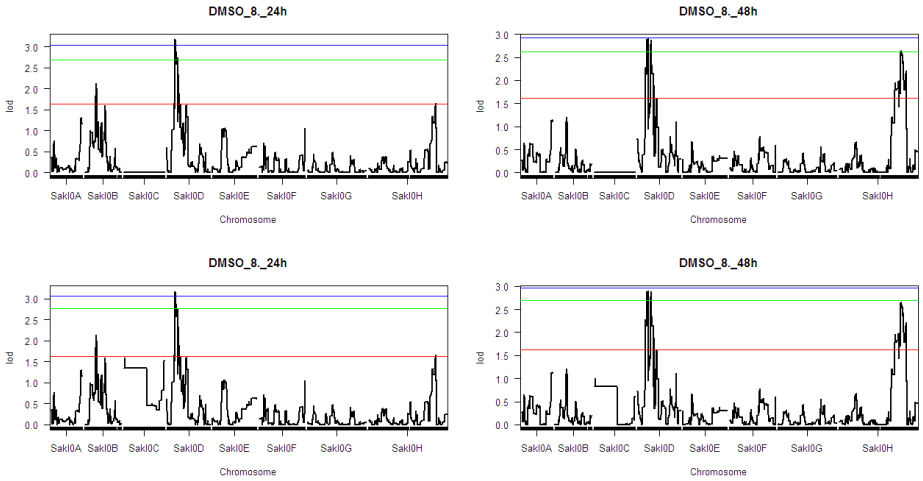
V.III Annexe 3 LOD score sous condition DMSO 8%



En haut, les LOD Scores sont calculés sans chromosome C

En bas, les LOD score sont calculés avec chromosome C

V.IV Annexe 4 les seuils sous condition DMSO 8%



La ligne rouge avec seuil 0.3

La ligue verte avec seuil 0.9

La ligne bleu avec seuil 0.95

BIBLIOGRAPHIE

1. Adler Joseph (2011), *R L'essentiel,* Pearson éducation
2. John Maindonald John Braun (2007), *Data Analysis and Graphics Using R,* Cambridge University Press
3. Karl W. Broman Saunak Sen (2009), *A Guide to QTL Mapping with R/qtl,* Springer
4. Thuriaux P., Goffeau A. (2002), *les organismes modèles: la levure,* Belin