Glossaire de statistique pour la génomique

Sommaire

Introduction

Ce document a pour objectif de définir un certain nombre de termes utilisés lors des analyses statistiques réalisées par la plateforme MGX sur des données issues de séquençage à haut débit. Afin de faciliter la lecture, les termes ont été regroupés en catégories :

- génomique,
- indicateurs statistiques,
- lois de probabilité,
- tests statistiques,
- tests multiples,
- méthodes de normalisation (RNA-Seq),
- graphiques.

Les termes inclus dans ce glossaire correspondent aux notions utilisées dans les analyses statistiques que nous réalisons sur la plateforme. Ce glossaire ne se veut pas exhaustif sur les notions utiles en statistique de manière générale.

Pour analyser les données, nous utilisons fréquemment des packages Bioconductor ; Biocondutor regroupe différents outils programmés en R permettant d'analyser des données génomiques. Nous faisons parfois référence à ce type d'outils dans les définitions données dans ce glossaire.

Génomique

Réplicats

Répétitions d'une même expérience biologique. L'objectif est de pouvoir réaliser des analyses statistiques pertinentes, car ils permettent d'estimer la variabilité des données. Il existe des réplicats techniques (obtenus sur le même matériel biologique) et des réplicats biologiques. Ce sont les réplicats biologiques qui sont informatifs pour les analyses statistiques.

Statistique descriptive

Médiane

La médiane d'un ensemble de valeurs (échantillon, population, distribution de probabilités) est une valeur m qui permet de couper l'ensemble des valeurs triées en deux parties égales : mettant d'un côté une moitié des valeurs, qui sont toutes inférieures ou égales à m et de l'autre côté l'autre moitié des valeurs, qui sont toutes supérieures ou égales à m (s'il y a un nombre pair de valeurs, la médiane sera la moyenne des 2 valeurs "centrales" de la distribution).

Quartile

Les quartiles sont les 3 valeurs qui divisent les données triées en 4 parts égales, de sorte que chaque partie représente 1/4 de l'échantillon de population. Il existe donc trois quartiles : Q1, Q2 (égal à la médiane) et Q3. Par exemple, Q1 est la valeur telle que 25 % des valeurs de l'échantillon lui sont inférieures, 75 % supérieures.

Dispersion

La dispersion représente la variabilité des différentes valeurs que peut prendre une variable. En statistiques, il existe différentes mesures de la dispersion. Les plus courantes sont la variance, l'écart-type ou encore l'intervalle inter-quartile. C'est une mesure peu influencée par la présence de valeurs extrêmes.
Le terme de dispersion est notamment employé dans les méthodes d'analyse différentielle en RNA-Seq pour parler de la variabilité des données.

Dispersion empirique

Dans les méthodes d'analyse différentielle en RNA-Seq, la dispersion empirique représente la dispersion estimée à partir des données, par opposition à la dispersion calculée après ajustement des données par un modèle mathématique ou à la dispersion théorique qui correspond à la "vraie" dispersion des données, qui est inconnue.

Intervalle interquartile (IQR)

Différence entre le 3ème et le 1er quartile, amplitude de l'intervalle interquartile : Q3 – Q1.

Variance, écart-type

La variance et l'écart-type sont des mesures servant à caractériser la dispersion d'un échantillon ou d'une distribution, la variance étant égale à l'écart-type au carré. Elles indiquent de quelle manière la série statistique se disperse autour de sa moyenne. C'est une mesure fortement influencée par la présence de valeurs extrêmes. Une variance de zéro signale que toutes les valeurs sont identiques. Une petite variance est signe que les valeurs sont proches les unes des autres alors qu'une variance élevée est signe que celles-ci sont très écartées.

Fold-Change

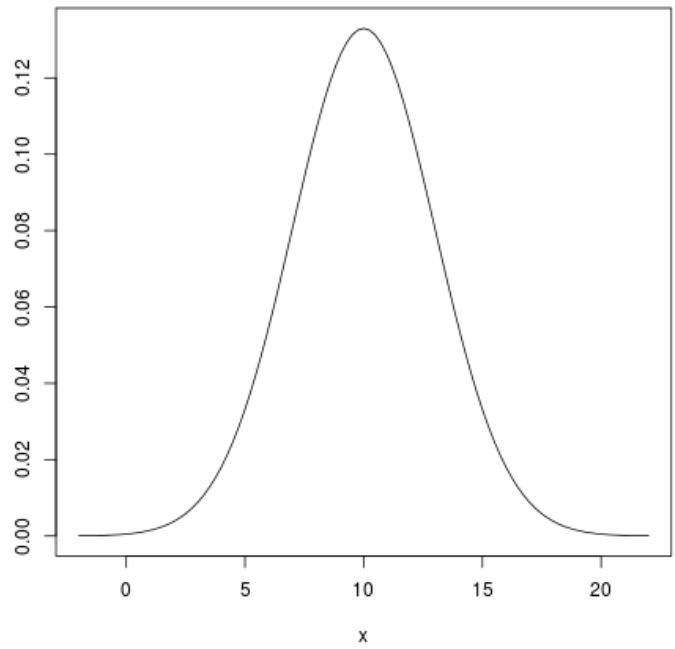
Le fold-change est le rapport du niveau moyen d'expression d'un gène dans une condition par rapport à une autre. Il est généralement exprimé en log (logarithme en base 2) afin de rendre symétriques les rapports par rapport à 1. Par exemple, un gène ayant un fold-change de 1 (respectivement -1) dans la condition A par rapport à la condition B signifie qu'il est deux fois plus (respectivement moins) exprimé dans la condition A que dans la condition B.

Lois de probabilité

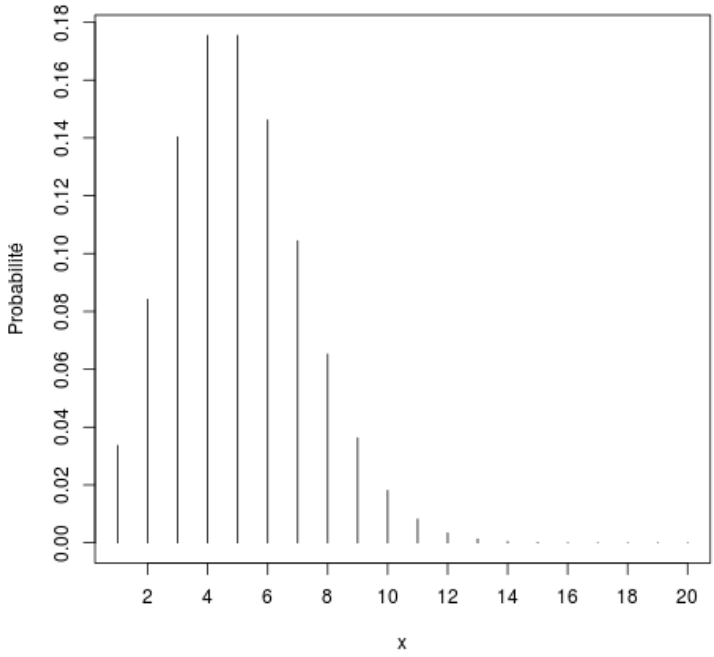
Distribution

Ensemble des valeurs, modalités ou classes d'une variable statistique, et des effectifs ou fréquences associées. La distribution d'une variable peut être représentée sous forme d'un diagramme en bâtons (variables discrètes) ou sous forme d'une fonction de densité (variables continues).

Exemple 1 (variable continue) : distribution d'une loi normale de moyenne 10 et d'écart-type 3



Exemple 2 (variable discrète) : distribution d'une loi de Poisson de paramètre 5



Loi de Poisson

La loi de Poisson est une loi de probabilité discrète connue pour être la loi des événements rares. Elle décrit le comportement du nombre d'événements se produisant dans un intervalle spatial ou temporal fixé. Elle possède un paramètre, souvent noté λ (lambda), représentant la moyenne et la variance.

C'est la loi utilisée par le logiciel MACS pour modéliser le nombre de reads dans un intervalle génomique donné.

Loi binomiale négative

La loi binomiale négative est une distribution de probabilité discrète. Elle possède deux paramètres (r et p).
C'est une alternative intéressante à la loi de Poisson. Elle est particulièrement utile pour des données discrètes dont la variance empirique excède la moyenne empirique. Si une loi de Poisson est utilisée pour modéliser de telles données, la moyenne et la variance doivent être égales. Dans ce cas, les observations sont «sur-dispersées» par rapport au modèle Poisson. Puisque la loi binomiale négative possède un paramètre supplémentaire, il peut être utilisé pour ajuster la variance. C'est pour cette raison qu'elle est souvent utilisée en génomique pour modéliser l'expression des gènes à partir des données de comptage obtenues en séquençage (c'est le cas dans les packages edgeR, DESeq, DESeq2).

Modèle Linéaire Généralisé (GLM)

Le modèle linéaire généralisé (GLM) est une généralisation souple de la régression linéaire. Le GLM généralise la régression linéaire en permettant aux variables explicatives du modèle linéaire (les valeurs d'expression des gènes par exemple) d'être reliées à la variable réponse (appartenance à différents groupes ou score continu par exemple) *via* une fonction de lien et en autorisant l'amplitude de la variance de chaque mesure d'être une fonction de sa valeur prévue.
Ce modèle est implémenté dans plusieurs packages Bioconductor permettant de réaliser l'analyse de données RNA-Seq sous R, en particulier dans les packages edgeR et DESeq2. Il permet notamment de traiter des design expérimentaux complexes, à plusieurs facteurs ou à plus de deux conditions.

Tests statistiques

Test Statistique

Un test d'hypothèse est une démarche consistant à rejeter ou à ne pas rejeter (rarement accepter) une hypothèse statistique, appelée hypothèse nulle, en fonction d'un jeu de données observées (échantillon). Il s’agit de statistique inférentielle : à partir de calculs réalisés sur des données observées, nous émettons des conclusions sur la population, en leur rattachant des risques de se tromper. En général, on connaît le risque de se tromper en rejetant à tort l'hypothèse nulle mais on ne connaît pas le risque de se tromper en ne rejetant pas, à tort, l'hypothèse nulle.

Échantillons appariés

Des échantillons sont appariés s'ils proviennent de mêmes individus ou cultures cellulaires. Par exemple, on peut avoir produit des échantillons d'une même culture cellulaire à différents temps ; ou bien, afin d'avoir des réplicats biologiques, plusieurs cultures cellulaires ont été réalisées. Ainsi, les échantillons provenant d'une même culture sont appariés.

Hypothèses (nulle et alternative)

Lors d'un test statistique, l'hypothèse posée « par défaut » est l'hypothèse nulle. L'hypothèse alternative est celle dont on cherche à prouver la véracité. Par exemple, lors d'une analyse différentielle en RNA-Seq, l'hypothèse nulle est l'hypothèse selon laquelle le gène n'est pas différentiellement exprimé entre les deux conditions étudiées ; l'objectif du test est de prouver l'hypothèse alternative, c'est-à-dire que le gène est différentiellement exprimé.

P-value

La p-value est la probabilité d'obtenir une valeur au moins aussi extrême que celle que l'on observe si l'hypothèse nulle était vraie. Pour un test visant à déterminer si un gène est différentiellement exprimé, la p-value représente la probabilité que le gène ait été déclaré différentiellement exprimé par erreur alors qu'il ne l'est pas en réalité.

Test exact de Fisher

Le Test exact de Fisher est un test statistique utilisé pour comparer deux proportions. Ce test est utilisé en général avec des faibles effectifs mais il est valide pour toutes les tailles d'échantillon. C'est un test qualifié d'exact car les probabilités peuvent être calculées exactement plutôt qu'en s'appuyant sur une approximation qui ne devient correcte qu'asymptotiquement comme pour le test du chi2.
C'est ce test qui est utilisé par le package Bioconductor topGO, qui permet de faire une analyse d'enrichissement sur les termes Gene Ontology.

Test paramétrique / non paramétrique

On parle de tests paramétriques lorsque l'on stipule que les données sont issues d'une distribution paramétrée. Dans ce cas, les caractéristiques des données peuvent être résumées à l'aide de paramètres estimés sur l'échantillon, la procédure de test subséquente ne porte alors que sur ces paramètres. L'hypothèse de normalité sous-jacente des données est le plus souvent utilisée, la moyenne et la variance suffisant ainsi pour caractériser complètement la distribution. Concernant les tests d'homogénéité par exemple, pour éprouver l'égalité des distributions, il suffira de comparer les moyennes et/ou les variances. Les tests non paramétriques ne font aucune hypothèse sur la distribution sous-jacente des données. On les qualifie souvent de tests *distribution free*. L'étape préalable consistant à estimer les paramètres des distributions avant de procéder au test d'hypothèse proprement dit n'est plus nécessaire. Lorsque les données sont quantitatives, les tests non paramétriques transforment souvent les valeurs en rangs. L'appellation « tests de rangs » est d'ailleurs souvent rencontrée. Lorsque les données sont qualitatives, seuls les tests non paramétriques sont utilisables. La distinction paramétrique – non paramétrique est essentielle. Elle est systématiquement mise en avant dans la littérature. Les tests non paramétriques, en ne faisant aucune hypothèse sur les distributions des données, élargissent le champ d'application des procédures statistiques. En contrepartie, ils sont moins puissants lorsque ces hypothèses sont compatibles avec les données.
Les tests non paramétriques seront utilisés dans les cas suivants :
- les données sont qualitatives,

- les effectifs sont trop petits (on considère souvent n<10) pour pouvoir tester si les données sont bien issues d'une distribution donnée (normale la plupart du temps),
- les effectifs sont suffisant pour pouvoir tester si les données sont bien issues d'une distribution donnée mais le test a rejeté cette hypothèse.

Tests multiples

Correction pour tests multiples

Correction appliquée aux p-values obtenues en résultats de tests statistiques lorsque l'on a réalisé plusieurs tests simultanément. Cela permet de contrôler le taux de faux positif. Cette correction est indispensable lorsque l'on réalise un grand nombre de tests simultanément, comme lors d'une analyse différentielle en RNA-Seq, où l'on teste généralement plusieurs milliers de gènes. En effet, chaque test étant associé à une erreur possible, cumuler les tests conduit à augmenter cette erreur. Il est donc nécessaire de prendre en compte le nombre de tests. La correction est d'autant plus forte que le nombre de tests est grand. Une des méthodes les plus couramment utilisées est celle de Benjamini et Hochberg appelée souvent correction FDR (*False Discovery Rate*).

FDR

False Discovery Rate (« taux de fausses découvertes », Benjamini et Hochberg, 1995). Méthode permettant de contrôler le taux de faux positifs parmi les tests ayant conduit à rejeter H0 (par exemple parmi les gènes considérés comme différentiellement exprimés) lors de la réalisation simultanée d'un grand nombre de tests statistiques.

P-value ajustée

P-value ajustée après la correction pour tests multiples (cf. «Correction pour tests multiples »).

Méthodes de normalisation (RNA-Seq)

Référence : Dillies et al., 2012

Normalisation, facteur de normalisation

La normalisation, en statistique, consiste en la transformation des données en vue de les rendre comparables entre différentes conditions.

RLE

La normalisation RLE (*Relative Log Expression*) a été développée dans le package Bioconductor DESeq. Elle se base sur l'hypothèse selon laquelle la plupart des gènes ne sont pas différentiellement exprimés. Le facteur de normalisation pour un échantillon est obtenu en calculant pour chaque gène la médiane des ratio de ses comptages par rapport à

sa moyenne géométrique entre les différents échantillons. L'idée sous-jacente est que les gènes non différentiellement exprimés doivent avoir des comptages similaires entre différents échantillons, et donc un ratio proche de 1. Si l'on suppose que la plupart des gènes ne sont pas différentiellement exprimés, la médiane des ratio constitue une estimation du facteur correctif qui doit être appliqué à l'ensemble des comptages.
C'est cette normalisation que nous utilisons généralement lors des analyses différentielles que nous réalisons sur la plateforme.

RPKM

La normalisation RPKM (*Reads Per Kilobase per Million*) a été introduite initialement pour faciliter les comparaisons entre gènes d'un même échantillon ; elle combine donc une normalisation inter et intra-échantillons. Ainsi, les comptages sont corrigés pour prendre en compte la taille de la librairie et la longueur des gènes. Cependant, il a été montré que la correction de la longueur des gènes a pour effet d'introduire un biais dans la variance par gène, en particulier pour les gènes faiblement exprimés. Cette méthode reste toutefois très populaire dans de nombreuses applications.
Nous ne l'utilisons pas sur la plateforme.

TMM

La normalisation TMM (*Trimmed Mean of M-values*) est implémentée dans le package Bioconductor edgeR. Elle se base sur l'hypothèse selon laquelle la plupart des gènes ne sont pas différentiellement exprimés. Le facteur TMM est calculé pour chaque échantillon, l'un d'eux étant considéré comme l'échantillon de référence et les autres comme des échantillons test. Pour chaque échantillon test, le facteur TMM est la moyenne pondérée des log-ratios entre ce test et la référence, après exclusion des gènes les plus exprimés et des gènes ayant les plus forts log-ratios. D'après l'hypothèse selon laquelle il y a peu de gènes différentiellement exprimés, le facteur TMM doit être proche de 1. S'il ne l'est pas, sa valeur donne une estimation du facteur correctif à appliquer aux tailles des librairies (et pas aux comptages bruts) afin de rendre l'hypothèse vraie.

Upper Quartile

Les comptages par gène sont divisés par le 3e quartile des comptages non nuls de l'échantillon, puis multipliés par la moyenne des 3e quartiles de tous les échantillons.

Graphiques

Box-plot (boîte à moustache)

Le box plot est un graphe qui permet de résumer graphiquement certaines caractéristiques d'une distribution : médiane, quartiles et minimum/maximum ou déciles. Il permet également de repérer les valeurs extrêmes (*outliers*) d'une distribution.

Diagramme de Venn

Représentation schématique des intersections entre différentes listes de valeurs.

MA plot

Le MA plot est un graphe qui était initialement utilisé dans les analyses de puce à ADN. C'est un nuage de points représentant en abscisse l'expression moyenne du gène à travers les différents échantillons, et en ordonnée le log-ratio des expressions moyennes d'une condition par rapport à l'autre. En RNA-Seq, après normalisation, on s'attend à ce que les points soient répartis symétriquement autour de 0 en ordonnée (c'est-à-dire un ratio de 1).

Nuage de points (*scatter-plot*)

Un nuage de points est une représentation de données dépendant de plusieurs variables. Il permet de mettre en évidence le degré de corrélation entre au moins deux variables liées.
Nous l'utilisons en RNA-Seq pour représenter les comptages (exprimés en log2) observés pour deux échantillons. Un nuage de points proche de la diagonale montrera que les deux échantillons sont globalement proches.