Les statistiques en biologie expérimentale : pourquoi ? Comment ? (première partie)

Au cours de ses études puis tout au long de sa carrière, un biologiste moléculaire ou un biochimiste est amené à apprendre de nouvelles techniques expérimentales, de nouvelles méthodes, et il bénéficie souvent de l'expertise de collègues expérimentés, qui lui transmettent leurs connaissances. Il est rare, cependant, qu'il reçoive les conseils d'un statisticien expert, et notre biologiste est souvent désemparé (voire, incrédule) quand on lui réclame une analyse statistique de ses résultats. S'il veut faire preuve de bonne volonté, il essaye d'apprendre de lui-même comment le faire, par ses lectures sur Internet ou dans des ouvrages spécialisés, et il est alors confronté à un jargon peu explicite, et à des concepts obscurs, peu accessibles. Au mieux, on lui sert des recettes toutes faites, auxquelles il est censé se plier sans comprendre, ce qui n'est ni satisfaisant intellectuellement, ni efficace à long terme.
L'objectif de ces deux petits articles (le premier dans ce numéro, le deuxième dans le prochain numéro de «Regard sur la biochimie») est, au contraire, d'expliquer à un public de biologistes, à quoi peuvent lui servir les statistiques, en quoi elles sont nécessaires, et comment les faire de manière convenable. Seuls les cas les plus simples seront traités ici, le but est surtout de donner au lecteur les notions essentielles pour résoudre lui-même les problèmes les plus faciles, et pour savoir quoi chercher dans les cas les plus difficiles.

À quoi servent les statistiques ?

Le problème central, dans toutes les sciences expérimentales, vient de l'irreproducibilité des mesures. Quand l'expérimentateur réalise deux fois la même mesure, sur des échantillons identiques, le résultat numérique (affiché par l'appareil de mesure) n'est jamais identique entre les deux mesures – les deux nombres ne seront jamais parfaitement égaux, avec tous leurs chiffres après la virgule.
Ce manque de reproducibilité a deux origines : d'une part, les mesures ne sont jamais réalisées de manière parfaitement identique (deux pipetages successifs n'auront jamais exactement le même volume, même si la pipette est réglée sur la même valeur ; deux incubations ne seront jamais aussi longues, à la seconde près ; ...) ; d'autre part, les échantillons mesurés ne sont eux-mêmes jamais parfaitement identiques (deux souris sauvages ne sont pas identiques génétiquement : ce ne sont pas des clones l'une de l'autre ; elles n'auront de toute façon pas eu exactement le même régime alimentaire, les mêmes infections bactériennes, *etc.* ; deux cultures de bactéries n'auront pas poussé exactement dans les mêmes conditions d'oxygénation et de température, en fonction de leur emplacement dans l'incubateur ; *etc*.). Ces phénomènes sont connus de tous ; et ils ont une conséquence inévitable : si je compare deux échantillons (mettons, un sauvage et un mutant), je vais certainement obtenir des mesures différentes (nous avons vu plus haut que deux mesures sur des échantillons *a priori* identiques ne donnaient jamais rigoureusement le même résultat, avec tous les chiffres après la virgule égaux ; donc *a fortiori*, deux échantillons sciemment différents donneront également des mesures différentes). Toute la question est alors de savoir si la différence mesurée est due au phénomène étudié (ici : si elle est due à la mutation qui distingue mon sauvage de mon mutant), ou si elle est due au fait que j'ai réalisé deux mesures (et que, chaque fois que je fais deux mesures, les résultats sont forcément, au moins un peu, différents).

Il est de la responsabilité de l'expérimentateur de s'assurer que les deux mesures ont été faites dans des conditions les plus similaires possibles (il vaut mieux que les échantillons sauvage et mutant aient été déposés sur le même gel d'électrophorèse, plutôt que sur deux gels différents ; il vaut mieux pipeter des volumes raisonnables : on ne pipette pas

0,5 µL avec une pipette de 200 µL ; *etc.*), mais même quand toutes ces précautions (nécessaires !) sont prises, les deux mesures donneront inévitablement des résultats (au moins un peu) différents, et il restera à déterminer si cette différence est due au phénomène étudié, ou à l'irreproducibilité inévitable des mesures expérimentales.

C'est précisément à ça que vont nous servir les statistiques : à estimer la variabilité qu'il y aura dans un groupe de mesures du sauvage, la variabilité dans un groupe de mesures du mutant, et à les comparer à la différence observée entre les deux groupes. Si les sauvages entre eux (et les mutants entre eux) sont beaucoup plus semblables que ne sont semblables les sauvages aux mutants, alors la différence observée sera principalement due au phénomène étudié (ici : la mutation). Sinon, il sera impossible de conclure : l'éventuel effet de la mutation sera obscurci par la trop grande variabilité des mesures, l'expérimentateur ne saura pas dire si la mutation a effectivement un effet sur la mesure.
On le voit, pour que les tests statistiques puissent nous donner ce genre d'information, il faudra leur fournir plusieurs réplicats de la mesure du sauvage, et plusieurs réplicats de la mesure du mutant. Plus les réplicats seront nombreux, plus le test sera précis (il aura pu estimer plus précisément la variabilité dans chaque groupe, et la différence entre les groupes). Une sorte de tradition tenace veut que les biologistes réalisent leurs expériences en trois réplicats : il n'y a aucune justification théorique pour ce nombre, et si l'expérimentateur peut en faire davantage, il y a tout intérêt ! Nous verrons d'ailleurs (dans le prochain numéro de «Regard sur la biochimie») qu'un certain test statistique ne peut pas déceler de différence significative (au seuil de 0,05 sur la *p-value*) si on ne lui fournit que trois réplicats du sauvage et trois réplicats du mutant. Il n'est bien sûr pas toujours possible de faire des dizaines de réplicats de chaque mesure (chaque réplicat prend du temps, et consomme des réactifs), mais lorsque c'est possible, l'expérimentateur se donne ainsi les moyens de détecter des différences beaucoup plus subtiles qu'avec seulement trois réplicats ; son analyse sera donc plus précise. Il faut aussi insister sur un autre point : les réplicats doivent être indépendants (c'est à dire que chacun doit être une expérience entière). Imaginons que l'expérience consiste en une quantification de l'abondance d'un ARNm dans le foie de la Souris par qRT-PCR : l'expérimentateur pourrait être tenté de ne faire qu'une seule dissection de souris, de n'extraire les ARN que de ce foie, et de réaliser ensuite plusieurs RT-PCR sur cet unique échantillon d'ARN. Ces réplicats ne seront pas indépendants : ils ne se distinguent que par la dernière étape (la RT-PCR) ; comparer ces différents réplicats ne le renseignera que sur la reproducibilité de cette dernière étape, et ne lui donnera aucune indication sur la reproducibilité de son extraction d'ARN de foie, aucune indication sur la variabilité qu'il peut y avoir entre deux souris (or quand il comparera ensuite le foie sauvage au foie mutant, il comparera nécessairement deux souris différentes). De manière à capturer toutes les sources d'irreproducibilité possibles, il est indispensable que les réplicats diffèrent entre eux autant que diffèrent les deux groupes de mesure (sauvage et mutant) : ils doivent résulter de dissections différentes, d'extractions d'ARN différentes, et de RT-PCR différentes.

Qu'est-ce que le *t-test* ?

Nous avons vu que la question essentielle consiste à estimer la confiance que l'expérimentateur peut accorder à une différence qu'il mesurera, inévitablement, entre les échantillons sauvages et les échantillons mutants. C'est l'objet du test statistique appelé «*t-test*» : il faut fournir à ce test les deux séries de valeurs (la série de réplicats du sauvage, et la série de réplicats du mutant), et il calcule une valeur appelée «*p-value*». La plupart de nos collègues savent que cette «*p-value*» mesure la confiance qu'on peut accorder à la différence mesurée, et plus elle est petite, plus la différence sera

«significative» (c'est à dire qu'elle aura un sens, du point de vue de l'expérimentateur ; dans notre exemple : une *p-value* basse signifiera que la différence observée est essentiellement due à la mutation qui distingue le sauvage du mutant, plus qu'à l'irreproducibilité intrinsèque de la mesure). Il serait toutefois utile de connaître précisément la définition de cette *p-value*. La voici : la *p-value* du *t-test*, c'est la probabilité que les deux populations échantillonnées aient la même moyenne.

Il est très facile de mal comprendre cette définition ! On pourrait croire que la *p-value* est la probabilité que ma série de réplicats du sauvage, et ma série de réplicats du mutant, ont la même moyenne. Ce serait un problème trivial ! Puisque les mesures sont faites, il est facile de calculer la moyenne des réplicats du sauvage, et la moyenne des réplicats du mutant. Il est donc facile de les comparer ; si elles sont égales, alors la probabilité qu'elles soient égales vaudrait 1, et si elles sont différentes, la probabilité qu'elles soient égales vaudrait 0 (il ne pourrait donc pas y avoir de *p-value* intermédiaire entre 0 et 1). Et comme (on l'a vu plus haut) deux mesures différentes donnent toujours, inévitablement, des résultats (au moins un peu) différents, alors on peut être sûr, avant même d'avoir fait l'expérience, que ces moyennes seront différentes – donc la *p-value* vaudrait toujours 0 (elle ne serait donc pas informative).
On s'en doute, ce n'était pas la définition de la *p-value* ; la différence tient à un mot qui semble un peu incongru au milieu de la définition : le mot «échantillonnées». Voici comment interpréter cette définition : l'expérimentateur, qui sait qu'il doit faire plusieurs réplicats de sa mesure, peut faire 10 réplicats (de son Western blot, de sa qPCR, …). Mais qu'est-ce qui l'empêche d'en faire 10 de plus ? Ou 20 de plus ? Ou même, une infinité de réplicats ? Naturellement, pour toutes sortes de raisons pratiques, il ne sera pas possible de faire une infinité de réplicats (chaque réplicat prend un certain temps ; chaque réplicat du Western blot consomme un peu d'anticorps, et il n'y en a pas une quantité infinie sur Terre ; …), mais on peut imaginer que, dans un monde idéal, il serait possible de faire une infinité de réplicats. Il existe donc, quelque part dans un univers idéal, une liste de toutes les mesures de tous ces réplicats ; une liste d'effectif infini, qui donnerait tous les résultats possibles de la mesure, pendant toute la vie de l'univers, et même encore plus … Et on peut interpréter d'une nouvelle manière l'expérience de mon biologiste, qui fait 10 réplicats de son expérience : faire 10 réplicats de l'expérience, ça revient à piocher au hasard 10 valeurs, parmi cette infinité de valeurs possibles issues d'une infinité de réplicats …
Eh bien c'est cette population idéale, d'effectif infini, dans laquelle pioche l'expérimentateur quand il fait son expérience, c'est cette population, donc, qu'on appelle la «population échantillonnée». Elle est «échantillonnée», parce que l'expérimentateur y échantillonne ses 10 (ou 20, ou 30, …) réplicats quand il fait son expérience.

On le voit, la définition de la *p-value* prend alors une toute autre signification : la *p-value*, c'est la probabilité que ces deux populations idéales, d'effectif infini (celle dont sont échantillonnés les réplicats sauvages, et celle dont sont échantillonnés les réplicats mutants), aient la même moyenne. Voyez la puissance du *t-test* ! Il nous permet de comparer des populations infinies de réplicats, quelque chose qui sera, évidemment, toujours inaccessible à l'expérience ; il nous permet de raisonner sur une expérience parfaite (celle qui aurait une infinité de réplicats, aussi bien pour le sauvage, que pour le mutant), sans avoir à la réaliser.
On comprend à présent pourquoi, plus la *p-value* est basse, plus la différence est significative : quand on lui fournit une liste de réplicats du sauvage et du mutant, le *t-test* va estimer à quoi ressemblent les deux populations idéales dont ces deux séries sont échantillonnées, et il va estimer la probabilité que leurs moyennes soient égales. S'il est très peu probable que les moyennes soient égales (*p-value* très petite devant 1), il faudra

en conclure qu'une expérience parfaite (avec une infinité de réplicats dans chaque série) détecterait certainement une différence entre les moyennes des deux séries – en d'autres termes : quand le nombre de réplicats est infini (c'est encore le meilleur moyen de gommer les sources techniques d'irreproducibilité de la mesure, puisque ces fluctuations atteindront de la même manière chacun des deux groupes), il restera une différence entre le groupe des sauvages et le groupe des mutants.

La figure 1 illustre le processus : considérons deux séries de valeurs (la série n°1, et la série n°2). Ces deux séries sont centrées sur des valeurs moyennes voisines de 15 (panneau du haut), et les populations idéales dont elles sont échantillonnées ont donc vraisemblablement, chacune, une moyenne voisine de 15 (les populations idéales, échantillonnées, sont représentées en bleu sur le panneau du milieu). Dans cet exemple précis, la *p-value* calculée par le *t-test* vaut à peu près 0,74. Si maintenant l'un des deux jeux de données (le jeu n°2) est décalé de deux unités vers la droite (panneau du bas, à droite), la probabilité que les populations idéales aient la même moyenne est beaucoup plus faible ($6{,}9.10^{-8}$).

Il est important de comprendre qu'aucune relation mathématique ne permet de calculer la *p-value* à partir de la différence entre les moyennes des deux séries de données : la *p-value* dépend également d'autres paramètres :

- La forme des distributions des deux séries de valeurs : si les distributions sont très serrées (dans chaque série, toutes les valeurs sont très voisines les unes des autres, et l'histogramme montre un pic haut et étroit), alors le *t-test* dira avec beaucoup plus de certitude que les populations idéales ont des moyennes différentes (il aura pu estimer avec davantage de confiance que les deux populations idéales étaient écartées l'une de l'autre, parce qu'elles seront, chacune, mieux résolues). Réciproquement, si les deux séries de valeurs sont très étalées, le *t-test* aura du mal à dire que les moyennes des populations idéales sont différentes : la *p-value* sera plus grande.
- Le nombre de réplicats dans chaque série de valeurs : plus les réplicats sont nombreux, plus le *t-test* pourra estimer avec précision la population idéale, d'effectif infini – donc plus facilement il pourra déceler une différence, même minime, dans leurs moyennes. C'est la raison pour laquelle, dans le panneau en bas à droite de la figure 1, la *p-value* est si basse (avec seulement 3 ou 4 réplicats, une différence de ~10% entre les moyennes des deux séries de valeurs, sur des séries de valeurs assez étalées comme ici, le *t-test* calculerait certainement une *p-value* beaucoup plus grande). Ici, le grand nombre de réplicats (200 dans la série n°1, 100 dans la série n°2) permet au *t-test* d'estimer avec précision les populations idéales, et donc, d'être d'autant plus affirmatif sur la différence entre leurs moyennes.

Ces considérations permettent de relativiser l'importance de la *p-value* : une *p-value* peut être très basse, même quand la différence entre les deux séries de valeurs est faible (si les distributions sont suffisamment fines, et si le nombre de réplicats est suffisamment grand). Il est d'ailleurs possible de faire baisser la *p-value* autant qu'on veut, simplement en multipliant le nombre de réplicats (le nombre de réplicats requis pour faire baisser la *p-value* en-dessous d'un seuil quelconque dépendra de la forme des deux distributions). Il ne faut donc pas considérer que la *p-value* mesure l'amplitude de la différence (l'amplitude, elle, se mesure en calculant le rapport ou la différence entre les moyennes, ou médianes, ou autres estimateurs, des deux séries de valeurs) : la *p-value* se contente de mesurer la confiance qu'on peut accorder à cette amplitude de la différence.

En résumé, chacune de ces deux grandeurs (l'amplitude de la différence, et : la *p-value*)

permet de mesurer une caractéristique de la différence entre les jeux de données ; chacune des deux est nécessaire pour la caractériser, et aucune n'est suffisante seule.

(le deuxième article, publié dans le prochain numéro de «Regard sur la biochimie», décrira les différentes options du *t-test*, et ses conditions d'application*).*

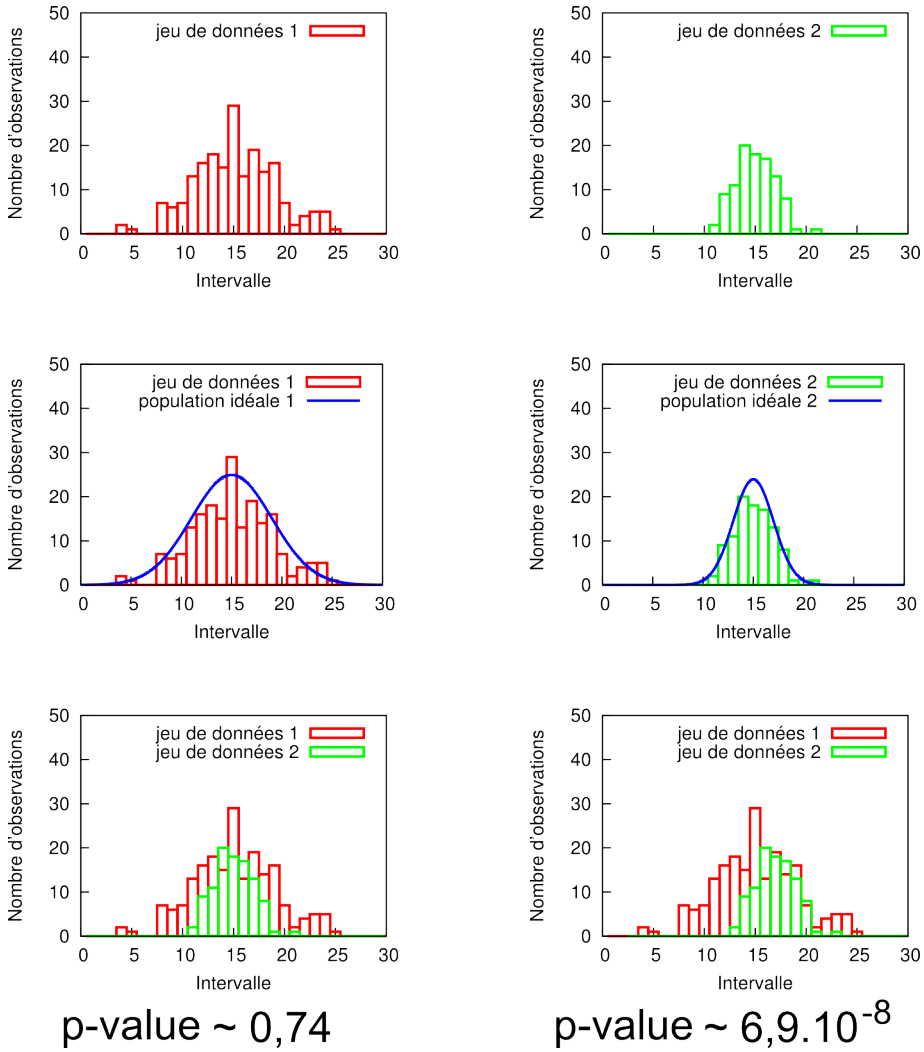

Figure 1 :
Soient deux jeux de données : le jeu n°1 (constitué de 200 réplicats, centrés sur une valeur voisine de 15 ; panneau en haut à gauche) et le jeu n°2 (constitué de 100 réplicats, centrés sur une valeur voisine de 15 également ; panneau en haut à droite). Ces histogrammes représentent les nombres d'observations qui tombent dans des intervalles de largeur 1 unité (par exemple, 18 valeurs du jeu de données n°1 tombaient entre 12,5 et 13,5, donc la barre centrée sur 13 a une hauteur de 18). Ici, les deux séries de valeurs ont été générées par ordinateur, elles ont été volontairement échantillonnées à partir de populations idéales d'effectifs infinis et de moyenne 15, représentées en bleu sur le panneau du milieu. Le *t-test* va estimer la probabilité que les populations idéales (qui lui sont inconnues), dont sont issues les deux séries, ont la même moyenne : cette probabilité vaut 0,74, ce qui indique que, d'après le *t-test*, il est très probable que ces deux séries soient échantillonnées à partir de populations infinies qui ont la même moyenne (ce qui est effectivement le cas). Si l'une des deux séries est décalée de deux unités vers la droite (panneau en bas à droite ; la série n°2 est maintenant centrée sur une valeur voisine de 17), le *t-test* estime très peu probable que les populations infinies, échantillonnées, aient la même moyenne (probabilité de $6{,}9.10^{-8}$).