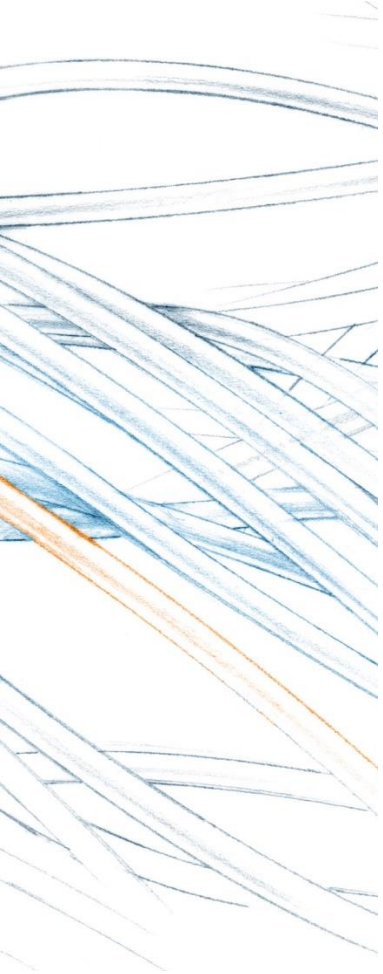


Résultats de l'analyse différentielle,

points de vigilance en statistiques

Gaëlle LELANDAIS

Résultats d'une analyse différentielle

(et après ?)

➢ Table de comptage (*counts*)

DESeq2 ⟶

➢ Table finale des résultats

RNAseq_counts_parapsilosis.txt

Normoxia_1	Normoxia_2		Normoxia_3		Normoxia_4		Normoxia_5		Normoxia_6	
Hypoxia_1	Hypoxia_2		Hypoxia_3		Hypoxia_4					
CPAR2_404750	589	498	630	612	627	642	462	611	319	619
CPAR2_212570	559	648	576	689	988	716	438	536	258	493
CPAR2_302130	167	83	230	151	180	136	112	385	103	172
CPAR2_209760	869	1438	972	1250	1588	1329	555	879	313	1206
CPAR2_403260	2549	2313	3115	2801	5834	3529	3373	3290	851	2873
CPAR2_209040	533	313	646	532	335	396	262	879	266	743
CPAR2_804040	44	42	57	44	68	43	41	76	25	71
CPAR2_211500	621	835	736	783	1156	860	529	635	293	588
CPAR2_102540	1672	2432	1845	2134	3653	3521	1470	1967	812	1970
CPAR2_109950	1772	2769	2175	2424	1629	1441	454	1485	412	920
CPAR2_800250	166	144	157	179	181	223	163	84	76	179
CPAR2_503900	1023	1435	1320	1414	1894	1664	544	1283	370	1212
CPAR2_100740	238	180	279	224	187	208	67	63	18	114
CPAR2_206160	705	933	831	823	1129	729	430	384	224	557
CPAR2_101420	383	552	448	526	612	388	246	253	100	289
CPAR2_805280	378	515	413	500	850	567	410	525	205	546
CPAR2_202350	1253	2099	1529	1826	2918	1964	1061	111		
CPAR2_703020	3668	5607	4518	5199	5891	4491	2378	266		
CPAR2_502480	219	291	284	302	233	166	117	83		
CPAR2_100810	1118	1908	1332	1616	1293	1419	415	130		
CPAR2_105300	5288	7896	6291	7105	11283	10306	9428	150		
CPAR2_407740	323	422	422	435	956	628	361	428	189	643
CPAR2_802340	640	848	822	900	1686	1236	488	622	217	543
CPAR2_106320	8875	13180	10310	11985	16549	8906	4887	2543	1266	3225
CPAR2_301210	167	191	218	204	540	451	209	280	124	315
CPAR2_302570	743	1121	856	947	1407	974	632	386	268	716
CPAR2_503570	99	137	135	101	482	213	193	234	86	300

baseMean

DESeq2_statistics.txt

baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
CPAR2_404750	557.176255877326	0.133444946455645	0.190487408943552		
0.700544708942886	0.48358719428854	0.599791972093409			
CPAR2_212570	559.374382948797	0.028735086611283 2	0.12662336890635		
0.226933518350279	0.820475438026274	0.876935520248357			
CPAR2_302130	173.841921340983	0.35023424511046	0.320288521371104		
1.09349608787466	0.274176045726641	0.389376678031686			
CPAR2_209760	954.765738326101	-0.114035759306079			
0.164726548624648	-0.692273105083535	0.488765836618937	0.604529147169239		
CPAR2_403260	2857.44347832337	0.358640646962554	0.197651989899545		
1.81450562245708	0.0695999179174566	0.137646093209841			
CPAR2_209040	491.800116175746	0.140998897015593	0.309119624371794		
0.456130526498074	0.648296119826938	0.744923394087986			
CPAR2_804040	50.1559296854689	0.345329251780373	0.261316423135524		
1.32149846395716	0.186335214355602	0.291206670418996			
CPAR2_211500	664.715757128726	0.0064962161591738	0.119238535927135		
0.054480844708992	0.956552069628808	0.968184140922596			
CPAR2_102540	1997.88804218508	0.236796415663859	0.148912116775271		
1.		67	0.1976103		
CP		5	-0.913291		
0.		0435876423	1		
CP		3	0.0920878		
0.		1457826686	0		
CP		5	-0.076710		
0.		84540483756	0		
CP		4	-0.918009		
0.		8398147556	0		
CPAR2_206160	622.942944989741	-0.399161611153727			

log2Fold Change

pvalue and padj

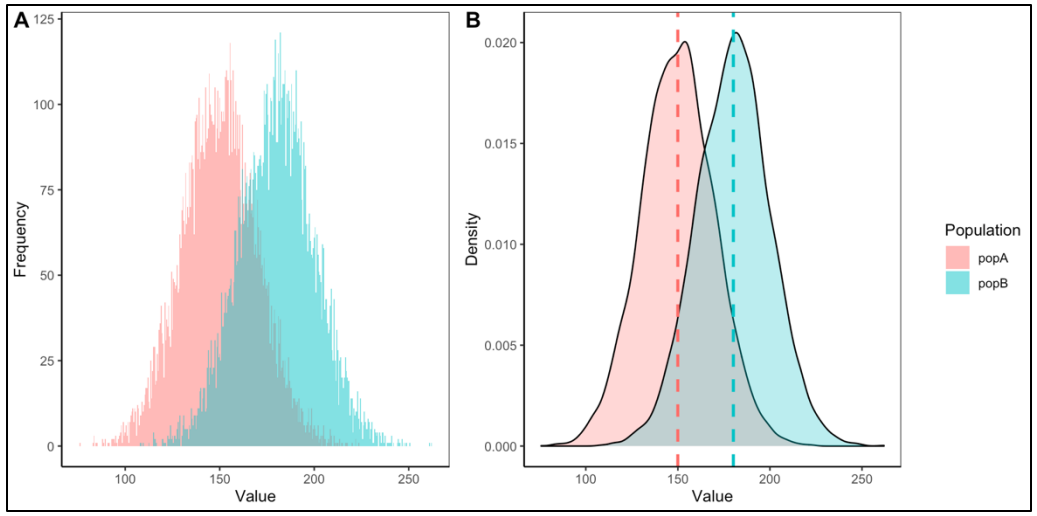
Votre travail consiste maintenant à créer votre liste de gènes candidats. Les paramètres statistiques peuvent vous aider à prendre des décisions raisonnées et argumentées. Ils ne décident pas à votre place.

Confrontation des points de vue (biologique et statistique)

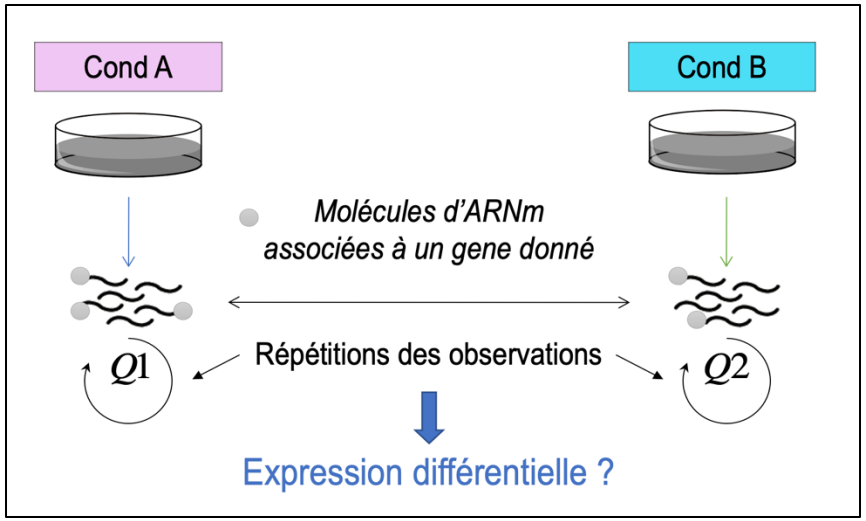
D'un point de vue biologique, un gène différentiellement exprimé est un gène dont l'expression est soumise à des mécanismes de régulations différents entre les conditions.

D'un point de vue statistique, un gène différentiellement exprimé est un gène dont la loi de probabilité (qui décrit son niveau d'expression) est différente entre les conditions.

➢ Simulation de données* :



➢ Stratégie expérimentale utilisées :



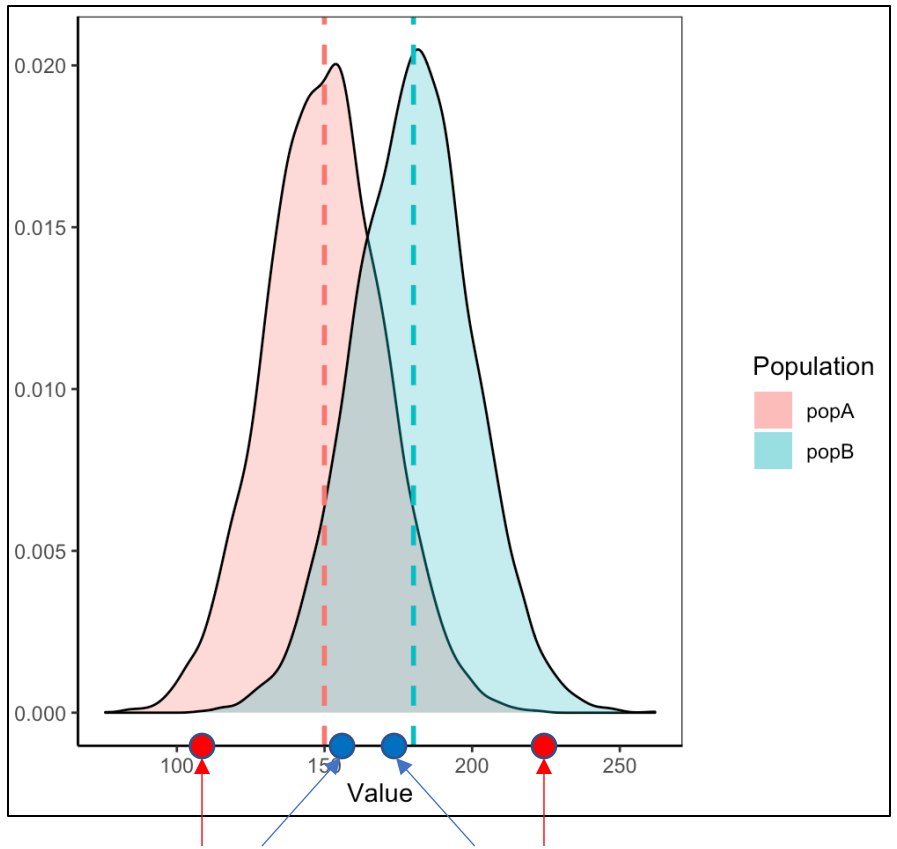
Les paramètres statistiques calculés, sont conditionnés aux modélisations statistiques sous-jacentes. Ces modèles ne correspondent pas exactement à la réalité.

« Tous les modèles sont faux, mais certains sont utiles » (George Box)

* Simulations réalisées pour la formation du DU « Création, analyse et valorisation des données omiques »

Pourquoi le risque d'erreur existe ?

- Considérons un gène différentiellement exprimé (point de vue statistique). Son expression suit deux lois de probabilité distinctes (décrivent des populations).



Echantillon A Echantillon B

- Considérons que des mesures de l'expression du gène ont été réalisées (décrivent des échantillons).

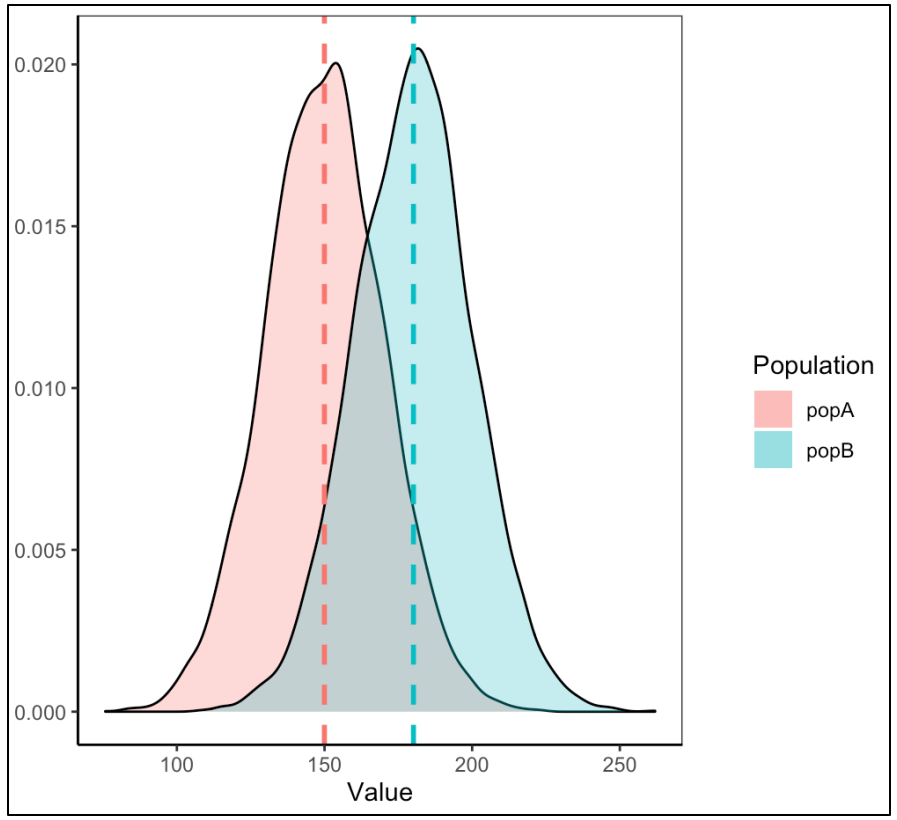
- Possibilité 1 : Echantillons A et B très différents. Ce que nous observons est conforme à la réalité, le risque de se tromper est assez faible.
- Possibilité 2 : Echantillons A et B peu différents. Pourtant le différentiel d'expression existe (!). Le risque de se tromper est alors élevé.

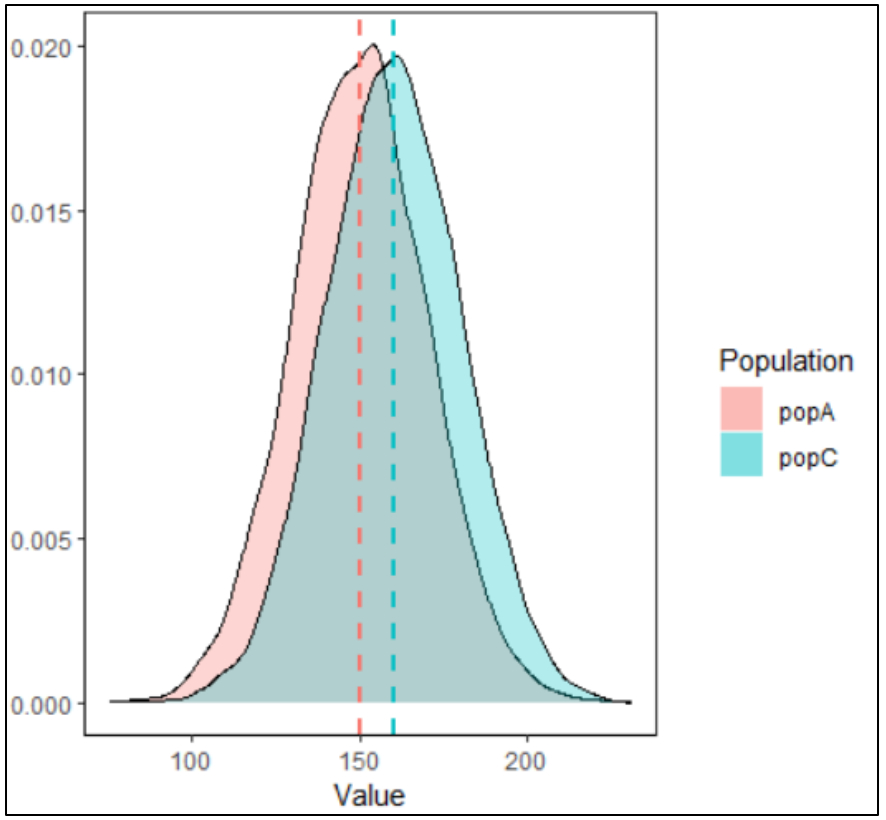
Inférence statistique consiste à « imaginer » comment sont les populations à partir des observations des échantillons.

* Simulations réalisées pour la formation du DU « Création, analyse et valorisation des données omiques »

Des situations plus ou moins complexes

- La problématique de la taille d'effet existe également au niveau des populations. Dans les deux cas, le gène est différentiellement exprimé (point de vue statistique).





Inférence statistique difficile → Inférence statistique très difficile

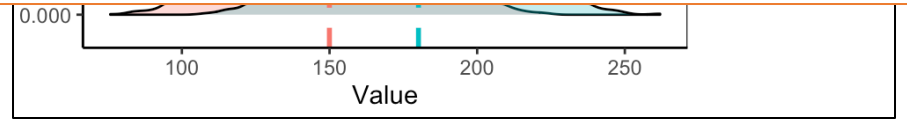
Des situations plus ou moins complexes

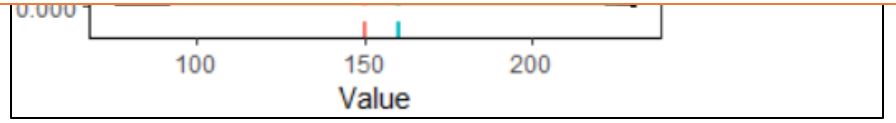
- La problématique de la taille d'effet existe également au niveau des populations. Dans les deux cas, le gène est différentiellement exprimé (point de vue statistique).

Lors d'une analyse différentielle, les inconnues sont très nombreuses :

Existe-t-il une régulation différentiée du gène ? (problématique biologique). Cette régulation a-t-elle un effet visible sur les mesures expérimentales réalisées ? Les modélisations statistiques de l'expression du gène sont-elles pertinentes pour cette régulation ? Les données sont-elles suffisantes pour estimer les paramètres des modèles ? (problématique statistique). Et bien d'autres…

Une grande vigilance est nécessaire lors de l'interprétation des résultats.



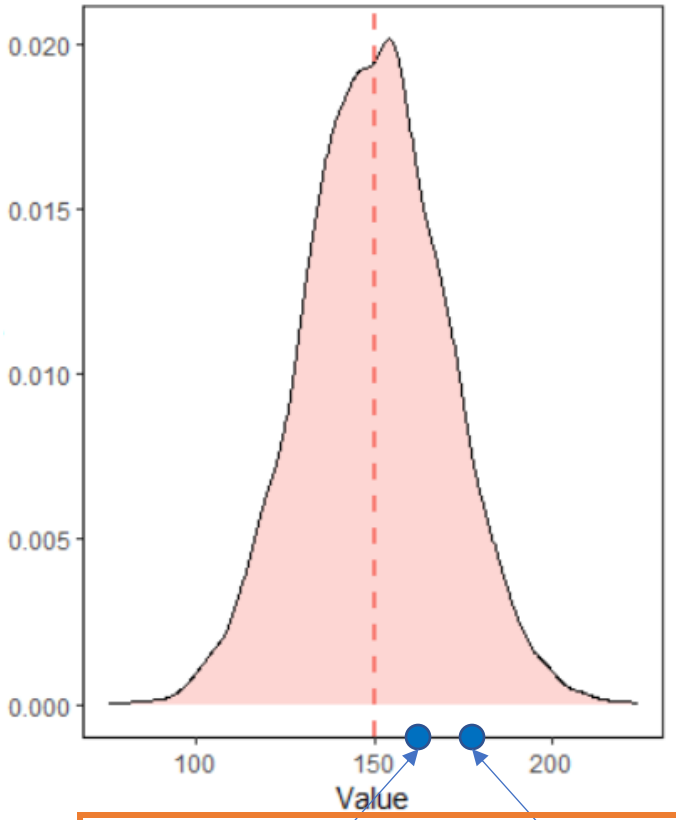


Inférence statistique difficile → Inférence statistique très difficile

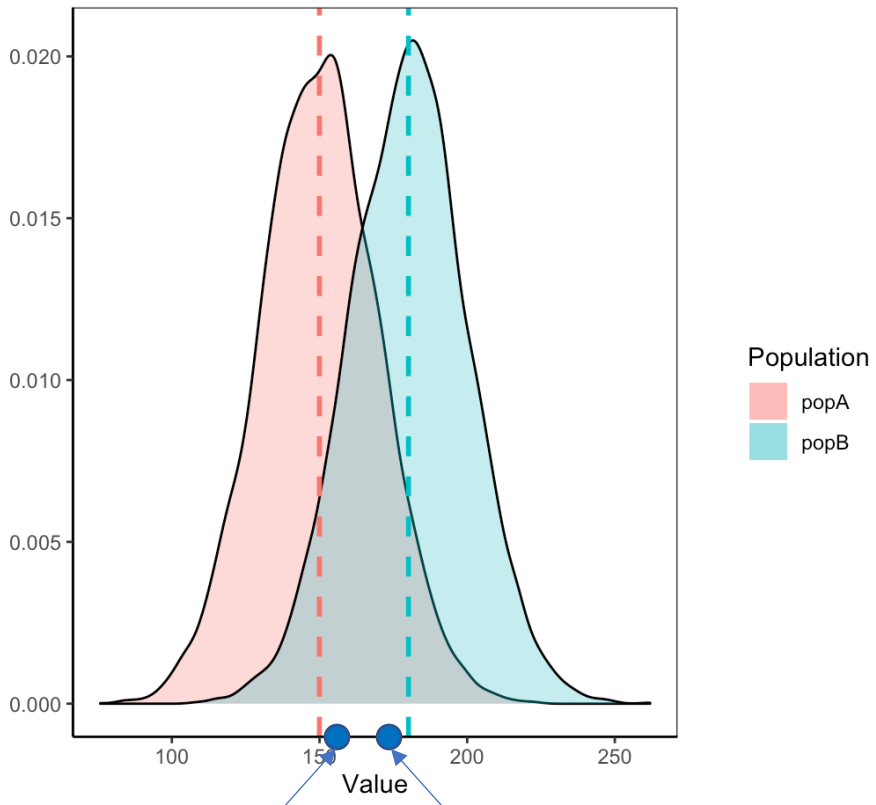
Hypothèses statistiques (rappels)

➢ Quelle est la source des différences observées entre les échantillons A et B ?

H_0 (hasard seul)



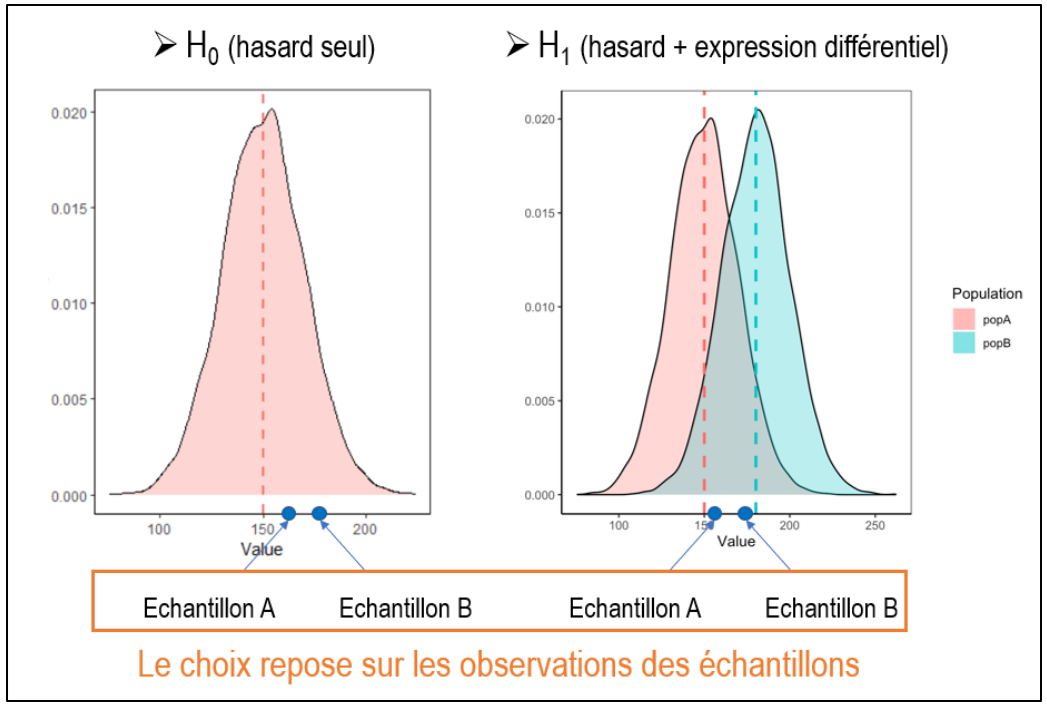
H_1 (hasard + expression différentielle)



Echantillon A Echantillon B Echantillon A Echantillon B

Le choix repose sur les observations des échantillons

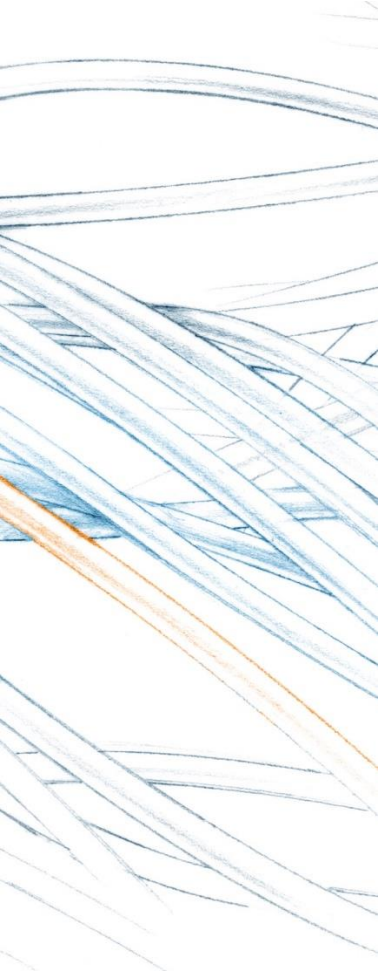
Risques d'erreurs (associés à la prise de décision)



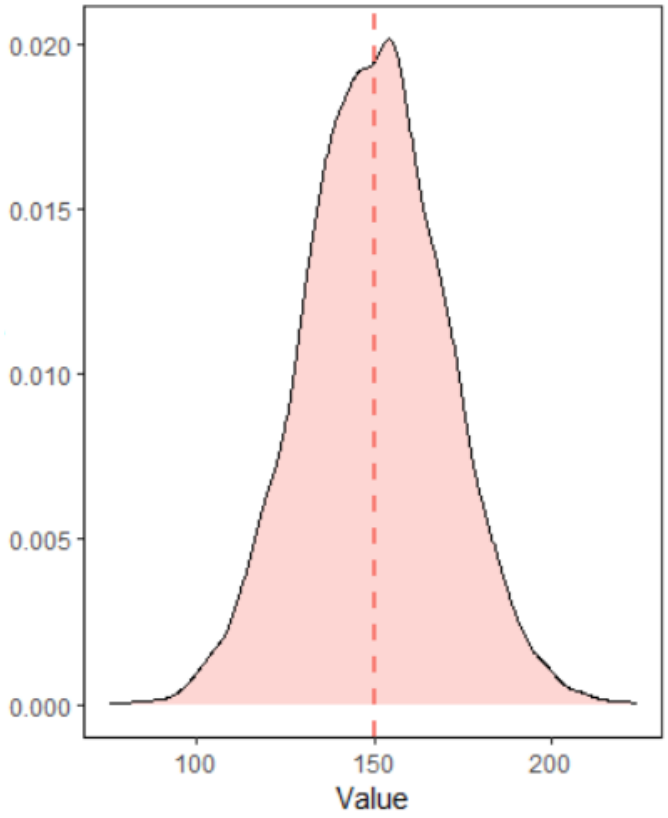
- Faux positif : Considérer que la différence est significative (choisir H_1) alors qu'elle ne l'est pas (H_0 est vraie).
- Faux négatif : Considérer que la différence n'est pas significative (choisir H_0) alors qu'elle l'est (H_1 est vraie).

	H_0 est vraie	H_1 est vraie
H_0 est choisie	Vrai négatif ✅	Faux négatif (risque de 2nd espèce) ❌
H_1 est choisie	Faux positif (risque de 1ère espèce) ❌	Vrai positif ✅

Hypothèse nulle (seule modélisation possible)

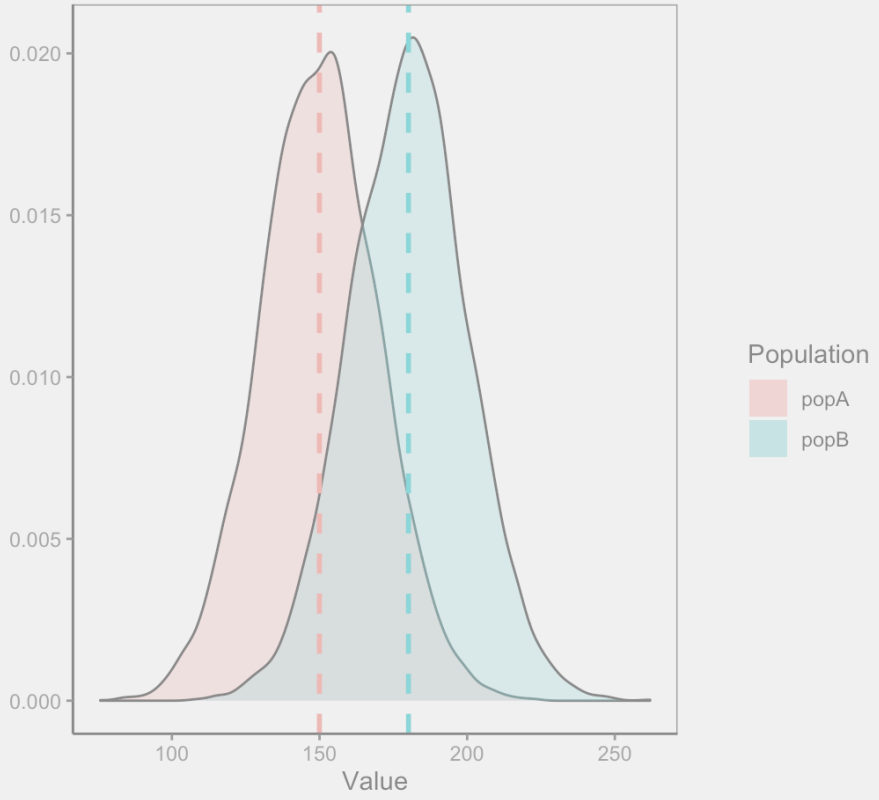


➢ H_0 (hasard seul)



Hypothèse considérée comme VRAIE pour la modélisation et les tests statistiques

➢ H_1 (hasard + expression différentiel)

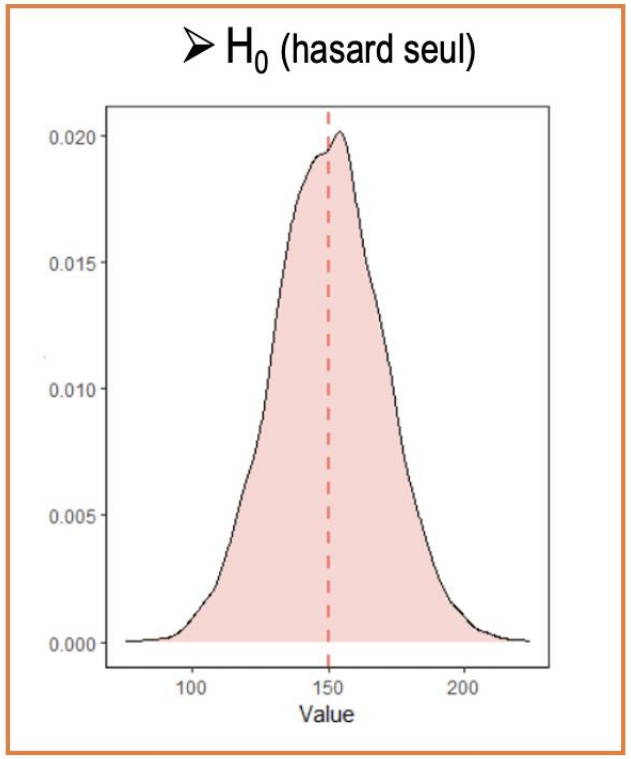


Hypothèse qui ne peut être modélisée car dépendante de paramètres inconnus

Interpréter le résultat d'un test statistique

(point de vigilance 1/2)

➢ Une règle de décision (valable pour tous les gènes) a été construite, sur la base d'une valeur de risque de 1ère espèce (0.05 ou 0.01 généralement, risque *a priori*).



Hypothèse considérée comme VRAIE pour la modélisation et les tests statistiques

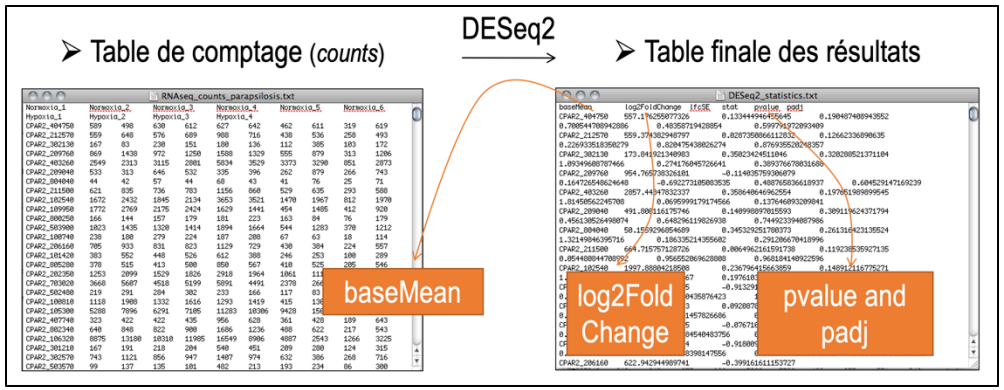
✔ Si la valeur P (pour un gène donné) est inférieure au risque *a priori*, la différence observée est « significative ». Un test statistique fournit une probabilité des observations (différences) conditionnellement à l'hypothèse nulle. C'est donc Pr(Données | H_0).

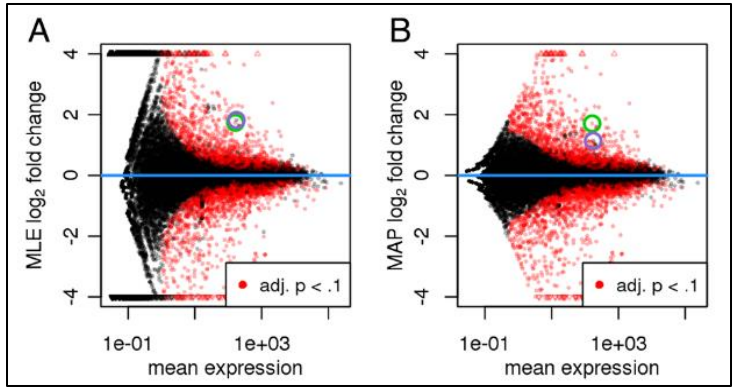
✖ Un test statistique ne nous dit pas quelle hypothèse est la plus probable compte tenu des données, c'est-à-dire Pr(H_0 | Données) et Pr(H_1 | Données).

Vigilance donc vis-à-vis des interprétations rapides telles que « ~~*plus la p-value est petite, plus j'ai de chance que le gène soit différentiellement exprimé*~~ » ou « ~~*plus la p-value est petite, plus le risque d'erreur est faible*~~ ».

Interpréter le résultat d'un test statistique

(point de vigilance 2/2)





- Un filtre sur la valeur P est réalisé pour créer la liste de gènes candidats (différentiellement exprimés). La valeur P est ainsi le seuil de signification observé, permettant d'évoquer pour certains gènes un résultat de test « très significatif » (< 0.01) ou seulement « significatif » (< 0.1)*.

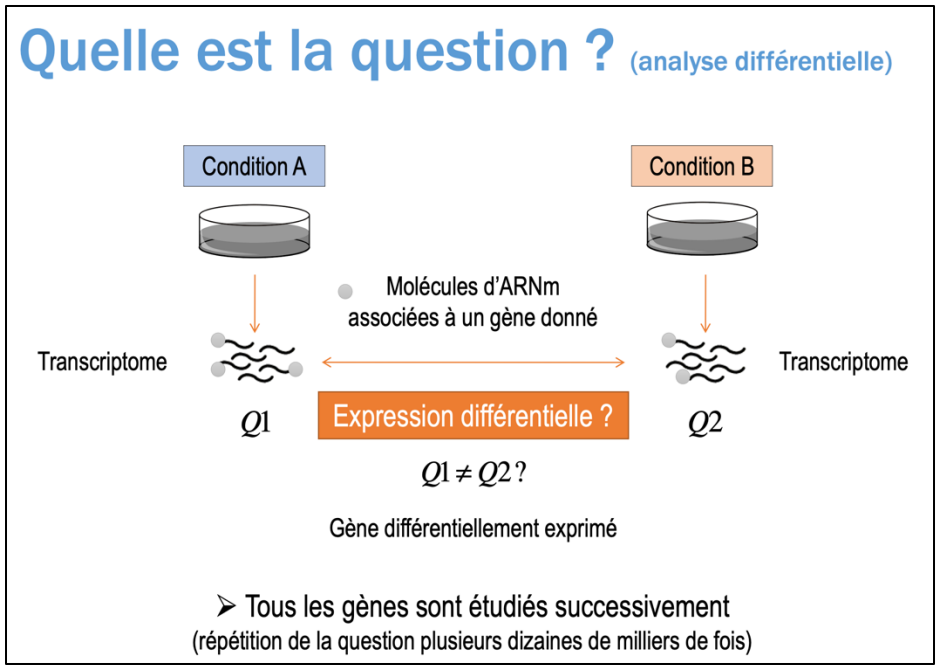
- Obtenir un résultat de test significatif ne démontre pas que les hypothèses sont ou ne sont pas vraies. Tous les calculs ont été réalisés en considérant que c'est l'hypothèse qui H_0 est vraie.

Vigilance donc vis-à-vis d'interprétations telles que « ~~le test est significatif, le gène est différentiellement exprimé~~ » ou bien « ~~le test n'est pas significatif, le gène n'est pas différentiellement exprimé~~ » .

* Notez que si la valeur P est comparée à un risque *a priori*, le test sera juste « significatif » ou « non significatif ».

Problématique des tests multiples

➢ Lors d'une analyse différentielle, la même question est posée plusieurs milliers de fois (nombre de gènes présents dans la table de comptage).



Dans le cas où les conditions A et B seraient identiques, aucun gène ne serait différentiellement exprimé (l'hypothèse H_0 devrait être choisie pour tous les gènes).

Le risque de considérer que la différence est significative (choisir H_1) alors qu'elle ne l'est pas (H_0 est vraie) est égale au risque de 1ère espèce (souvent 0.05).

Si 10 000 tests statistiques sont réalisés, il est attendu (en moyenne*) 500 gènes pour lesquels H_1 est choisie alors que H_0 est vraie.

	H_0 est vraie	H_1 est vraie
H_0 est choisie	Vrai négatif ✅	Faux négatif (risque de 2nd espèce) ❌
H_1 est choisie	Faux positif (risque de 1ère espèce) ❌	Vrai positif ✅

* Espérance de la loi binomiale de paramètre p = 0.05 et n = 10 000.

Quelle valeur P utiliser ?

➢ Table de comptage (*counts*)

DESeq2 ⟶

➢ Table finale des résultats

RNAseq_counts_parapsilosis.txt

	Normoxia_1	Normoxia_2	Normoxia_3	Normoxia_4	Normoxia_5	Normoxia_6
	Hypoxia_1	Hypoxia_2	Hypoxia_3	Hypoxia_4		
CPAR2_404750	589	498	630	612	627	642	462	611	319	619
CPAR2_212570	559	648	576	689	988	716	438	536	258	493
CPAR2_302130	167	83	230	151	180	136	112	385	103	172
CPAR2_209760	869	1438	972	1250	1588	1329	555	879	313	1206
CPAR2_403260	2549	2313	3115	2801	5834	3529	3373	3290	851	2873
CPAR2_209040	533	313	646	532	335	396	262	879	266	743
CPAR2_804040	44	42	57	44	68	43	41	76	25	71
CPAR2_211500	621	835	736	783	1156	860	529	635	293	588
CPAR2_102540	1672	2432	1845	2134	3653	3521	1470	1967	812	1970
CPAR2_109950	1772	2769	2175	2424	1629	1441	454	1485	412	920
CPAR2_800250	166	144	157	179	181	223	163	84	76	179
CPAR2_503900	1023	1435	1320	1414	1894	1664	544	1283	370	1212
CPAR2_100740	238	180	279	224	187	208	67	63	18	114
CPAR2_206160	705	933	831	823	1129	729	430	384	224	557
CPAR2_101420	383	552	448	526	612	388	246	253	100	289
CPAR2_805280	378	515	413	500	850	567	410	525	205	546
CPAR2_202350	1253	2099	1529	1826	2918	1964	1061	111		
CPAR2_703020	3668	5607	4518	5199	5891	4491	2378	266		
CPAR2_502480	219	291	284	302	233	166	117	83		
CPAR2_100810	1118	1908	1332	1616	1293	1419	415	130		
CPAR2_105300	5288	7896	6291	7105	11283	10306	9428	150		
CPAR2_407740	323	422	422	435	956	628	361	428	189	643
CPAR2_802340	640	848	822	900	1686	1236	488	622	217	543
CPAR2_106320	8875	13180	10310	11985	16549	8906	4887	2543	1266	3225
CPAR2_301210	167	191	218	204	540	451	209	280	124	315
CPAR2_302570	743	1121	856	947	1407	974	632	386	268	716
CPAR2_503570	99	137	135	101	482	213	193	234	86	300

baseMean

DESeq2_statistics.txt

baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
CPAR2_404750	557.176255877326	0.133444946455645	0.190487408943552		
0.700544708942886	0.48358719428854	0.599791972093409			
CPAR2_212570	559.374382948797	0.028735086611283	0.12662336890635		
0.226933518350279	0.820475438026274	0.876935520248357			
CPAR2_302130	173.841921340983	0.35023424511046	0.320288521371104		
1.09349608787466	0.274176045726641	0.389376678031688			
CPAR2_209760	954.765738326101	-0.114035759306079			
0.164726548624648	-0.692273105083535	0.488765836618937	0.604529147169239		
CPAR2_403260	2857.44347832337	0.358640646962554	0.197651989899545		
1.81450562245708	0.0695999179174566	0.137646093209841			
CPAR2_209040	491.800116175746	0.140998897015593	0.309119624371794		
0.456130526498074	0.648296119826938	0.744923394087986			
CPAR2_804040	50.1559296854689	0.345329251780373	0.261316423135524		
1.32149846395716	0.186335214355602	0.291206670418996			
CPAR2_211500	664.715757128726	0.0064962161591738	0.119238535927135		
0.054480844708992	0.956552069628808	0.968184140922596			
CPAR2_102540	1997.88804218508	0.236796415663859	0.148912116775271		
1.		67	0.1976103		
CP		5	-0.913291		
0.		0435876423	1		
CP		3	0.0920878		
0.		1457826686	0		
CP		5	-0.076710		
0.		84540483756	0		
CP		4	-0.918009		
0.		8398147556	0		
CPAR2_206160	622.942944989741	-0.399161611153727			

baseMean

log2Fold Change

pvalue and padj

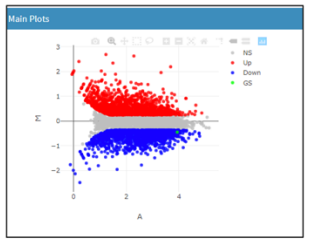
Il existe différentes stratégies pour « ajuster » la valeur P, et ainsi tenir compte de la problématique des tests multiples. Si les valeurs P ajustées sont plus élevées que les valeurs P initiales, elles ne changent pas l'ordre des gènes. Celui qui avait la plus petite valeur P aura toujours la plus petite valeur P ajustée.

Bilan

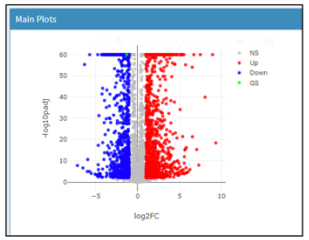
- La mise en application de méthodes statistiques et très utile pour identifier (sans *a priori*) des gènes intéressants, bons candidats pour être différentiellement exprimés.

Bilan des représentations graphiques
(application WEB)

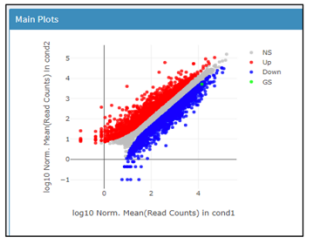
- Filtre de sélection par défaut : padj < 0.01 ou FC > 2



MA plot



Volcano plot



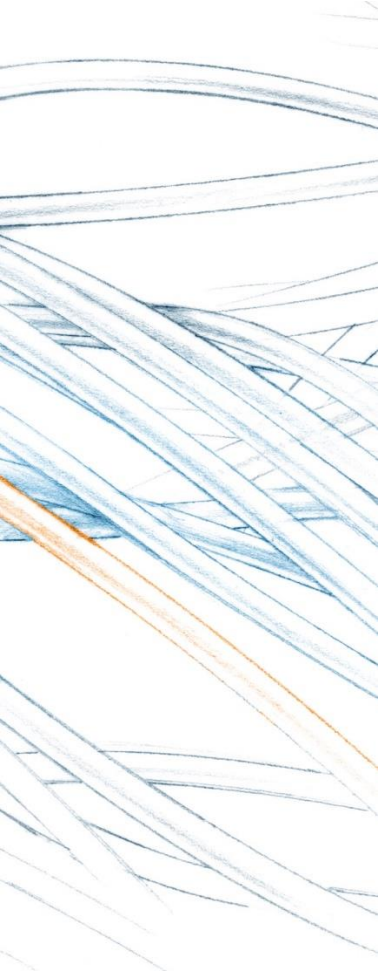
Scatter plot

- Données initiales, pour un gène donné

Vigilance vis-à-vis de la sur-interprétation des résultats
(positivement ou négativement)

Les statistiques sont une aide précieuse à la prise de décision
(mais vous seul(e) restez responsable)

Ressource(s) complémentaire(s)



- Vidéo de présentation du cours (replay, 16 minutes)
  https://youtu.be/CiP7tO0jOsM