INSA
135 Avenue de Rangueil
31400 Toulouse

INRA
24 Chemin de Borde-rouge
31320 Auzeville-Tolosane

Rapport de stage de 4ème année
Mathématiques Appliquées

Étude des effets de polluants alimentaires sur les cellules

Fanny Mathevet

Stage effectué du 1er juillet 2019 au 27 septembre 2019
à l'INRA, unité MIAT

Sous la direction de :
Nathalie Vialaneix (tutrice de stage)
Gaëlle Lefort (tutrice de stage)
Béatrice Laurent-Bonneau (enseignante-tutrice)

Remerciements

Je tiens à remercier mes encadrantes de stage Nathalie Vialaneix et Gaëlle Lefort, pour leur disponibilité, leurs conseils avisés et la confiance qu'elles m'ont accordée tout au long de mon stage.

Je remercie également Béatrice Laurent-Bonneau, ma tutrice INSA, ainsi que l'ensemble des enseignants du département GMM de l'INSA, dont les enseignements m'ont été précieux dans la réalisation des missions qui m'ont été confiées.

Je souhaite aussi remercier Annick Moisan pour son accueil au sein du bureau MIAT23, et pour sa bonne humeur.

Enfin, je remercie l'ensemble de l'équipe MIAT pour son accueil chaleureux pendant ces 3 mois.

Table des matières

1 L'INRA et l'unité MIAT

L'Institut National de la Recherche Agronomique (INRA)
L'Institut National de la Recherche Agronomique (INRA) est un organisme de recherche scientifique publique, placé sous la double tutelle du ministère de l'Enseignement supérieur et de la Recherche, et du ministère de l'Alimentation, de l'Agriculture et de la Pêche. Il est constitué de 13 départements scientifiques, répartis sur 17 centres de recherche régionaux. L'INRA a pour missions de :

- produire et diffuser des connaissances scientifiques
- concevoir des innovations et des savoir-faire pour la société
- éclairer, par son expertise, les décisions des acteurs publics et privés
- développer la culture scientifique et technique
- former à la recherche

Il mène des recherches sur les thèmes de l'agriculture, l'alimentation et la sécurité des aliments, l'environnement et la gestion des territoires. Toutes ces recherches sont effectuées dans une perspective de développement durable.

Le département de Mathématiques et Informatique Appliquées (MIA)
Le département de Mathématiques et Informatique Appliquées réalise des recherches dans le domaine des maths-info pour répondre à des problématiques en lien avec la science du vivant et de l'environnement, et d'accompagner le développement des maths-infos. Il se compose de chercheurs et d'ingénieurs, qui développent les outils et logiciels nécessaires à l'exploitation de données biologiques et environnementales recueillies par l'INRA. Ses recherches portent sur la bio-informatique pour la biologie des systèmes et de synthèse (MIA-bio), les mathématiques et informatique pour la biologie des populations, l'écologie et l'épidémiologie (MIA-pop), et le développement du numérique pour l'agriculture, l'environnement et l'alimentation (MIA-num).

L'Unité de Mathématiques et Informatique Appliquées de Toulouse (MIAT)
L'unité de Mathématiques et Informatique Appliquées de Toulouse est chargée de mettre au point des méthodes mathématiques et informatiques et de les mettre à disposition de l'INRA, favorisant ainsi les collaborations entre départements. Le domaine de compétences de l'unité s'étend aux statistiques, probabilités, algorithmique, intelligence artificielle et sciences de la décision. L'unité comporte, depuis janvier 2011, deux équipes de recherches :

- MAD (Modélisation des Agro-écosystèmes et Décision) : modélisation des systèmes compltexes dans les champs de l'agriculture, de l'environnement et de l'analyse des risques alimentaires et des procédés industriels.
- SaAB (Statistique et Algorithmique pour la Biologie) : développement de méthodes relevant des mathématiques, de la statistique et de l'informatique destinées à l'exploitation de données de génomique et de post génomique.

L'unité compte aussi sur l'activité de trois plateformes :

- Plateforme GENOTOUL : Plateforme bioinformatique du GIS GENOTOUL, dont l'activité est centrée sur l'analyse de séquences.
- Plateforme RECORD (Rénovation et Coordination de la modélisation des cultures pour la gestion des agro-écosystèmes) : Plateforme issue du partenariat des départements Environnement et Agronomie (EA) et Mathématiques et Informatiques Appliquées (MIA). Elle vise à offrir un cadre et des outils informatiques communs aux modélisateurs des différentes disciplines (agronomie, bioclimatologie, sciences de gestion, mathématiques, ...) pour la modélisation et la simulation des systèmes de culture.
- Plateforme SIGENAE (Système d'information des génomes des Animaux d'Elevage). Elle se compose d'ingénieurs en bio-informatique qui accompagnent les biologistes de départements « animaux » (Génétique Animale, Physiologie Animale et Système d'Elevage, Santé Animale) de l'INRA dans le traitement de leurs données à haut débit.

Mon stage a été encadré par Mme Nathalie Vialaneix, Directrice de Recherche au sein de l'unité MIAT, et Mme Gaëlle Lefort, doctorante.

2 Contexte et objectif du stage

2.1 Contexte biologique

Avec l'industrialisation de la société, l'utilisation d'additifs, de conservateurs et autres produits de synthèse dans le secteur agroalimentaire et agricole s'est accrue ces dernières décennies. Plusieurs dizaines de milliers de molécules chimiques, développées dans l'après-guerre, ont été répandues dans la nature, les sols, ou encore les cours d'eau [Couturier, 2019]. À toute étape de la chaîne alimentaire, ces molécules, communément appelées « polluants alimentaires », sont susceptibles de s'ajouter à l'aliment originel et de lui conférer des propriétés toxicologiques, parfois liées à l'apparition du cancer comme celui du colon.
L'étude des effets des polluants alimentaires sur nos cellules représente un enjeu sanitaire de taille : pouvoir caractériser le comportement d'une lignée cellulaire face à un polluant nous permettrait de combattre ses éventuels effets néfastes, et ainsi d'agir sur le cancer du colon.

Les impacts des polluants alimentaires peuvent être mesurés à plusieurs niveaux de fonctionnement de la cellule, tels que la respiration cellulaire ou la quantification de l'expression de gènes.

2.1.1 La respiration cellulaire

Lors de la respiration cellulaire, les cellules consomment le dioxygène (O_2) provenant de la circulation sanguine et rejettent du dioxyde de carbone (CO_2) et de la vapeur d'eau (H_2O). La respiration se découpe en trois étapes, toutes représentées sur la figure 1.

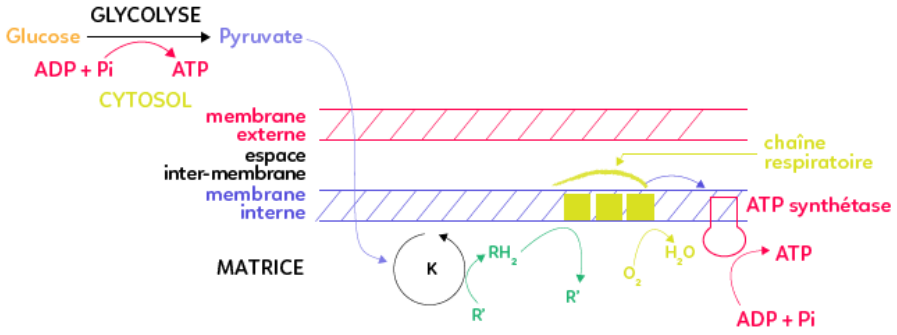


FIGURE 1: Schéma explicatif des mécanismes de la respiration cellulaire [LesBonsProfs, 2014]

- La glycolyse correspond à la première étape de la respiration cellulaire. Elle permet la transformation, dans le cytoplasme de la cellule, du glucose (sucre possédant 6 atomes de Carbone) en pyruvate (sucre possédant 3 atomes de Carbone).
- Le pyruvate pénètre dans la mitochondrie, organite intracellulaire présent dans les cellules eucaryotes, puis s'oxyde à nouveau lors du « cycle de Krebs ». Au terme de ce cycle sont formés les composés R', qui subissent une phase de réduction pour passer à l'état R'H2, puis une phase d'oxydation qui les ramène à l'état R' en cédant électrons et protons à une chaîne de transmetteurs d'électrons. C'est au cours de ce cycle, lors de la phase de réduction des composés R', qu'est produit le dioxyde de carbone [kartable, 2019].
- La chaîne de transmetteurs obtenue, également appelée « chaîne respiratoire », conduit les électrons jusqu'au dioxygène. L'acceptation des électrons et protons par le dioxygène enclenche sa consommation et la production de dioxyde de carbone et de molécules d'eau [LesBonsProfs, 2014].

2.1.2 Mesure de l'expression de gènes avec la technique qPCR

L'Acide Désoxyribo-Nucléique (ADN) est une double hélice constituée de successions de paires de nucléotides, connus sous les noms d'Adénine (A), Cytosine (C), Guanine (G), et Thymine (T). L'Adénine est toujours couplé à la Thymine, et la Cytosine à la Guanine. Un gène correspond à une succession particulière de paires de nucléotides, autrement dit, à un fragment d'ADN (figure 2).
Certains gènes encodent des protéines. La synthèse des protéines est rendue possible par l'ARN messager, créé dans le noyau lors de l'étape de transcription de l'ADN. Lors de celle-ci, l'enzyme ARN polymérase écarte les deux brins de la molécule d'ADN et synthétise le brin complémentaire au « brin codant » de

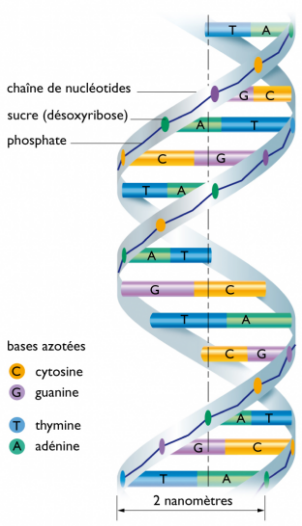


FIGURE 2: Schéma d'un fragment de brin d'ADN [Agr'OGM, 2019]

l'ADN, autrement dit, au brin du fragment d'ADN que l'on souhaite traduire en protéine [Futura, 2019]. Pour ce faire, l'ARN polymérase attire et associe les nucléotides complémentaires à ceux présents sur ce brin codant : le nucléotide de base G face à C, et vice-versa, le nucléotide de base A face à T, et le nucléotide de base Uracile (U), propre à l'ARN, face à A. La molécule obtenue, appelée ARN messager, est traduite en protéine lors de l'étape de traduction, qui a lieu dans le cytoplasme.
La mesure de l'activité d'une protéine permet d'obtenir des informations sur le comportement de la cellule. Or, comme précédemment expliqué, les protéines sont issues de l'expression de gènes. Quantifier le niveau d'expression d'un gène nous donne une idée du niveau de traduction de la protéine associée, bien qu'il n'existe pas de relation simple entre ces deux niveaux (le processus entre ARN messager et protéine étant complexe). On peut, pour cela, quantifier le niveau d'ARN messager d'intérêt. Plusieurs méthodes existent. Nous ne nous intéresserons qu'à la Réaction de Polymérisation en Chaîne quantitative, dénotée qPCR, réalisable à moindres coûts. En contrepartie, le nombre de gènes ciblés se doit d'être faible.

La PCR (Polymerase Chain Reaction) est une technique de réplication de fragments d'ADN. Elle consiste en une répétition d'étapes de synthèse, chacune d'entre elles comprenant trois parties : dénaturation, appariement, et élongation.

- La dénaturation consiste en la rupture, à haute température (95°C), des liaisons hydrogènes reliant les deux brins de l'ADN.
- Des « amorces » complémentaires de chaque brin, présentes dans le milieu réactionnel, se fixent à leurs brins complémentaires lors de la phase d'appariement. Ces amorces correspondent respectivement aux successions nucléotidiques initiales et finales du fragment d'ADN à amplifier.
- Lors de l'élongation, l'enzyme « taq polymérase » allonge chaque amorce dans une unique direction en y rattachant les nucléotides complémentaires au reste du brin auquel il est fixé.

Au bout du troisième cycle, de l'ADN double-brins borné par les amorces apparaît. Il s'agit des « fragments courts » mentionnés sur la figure 3. Cet ADN augmente en quantité au fil des cycles. Les produits de chaque étape de synthèse étant réutilisés pour les étapes suivantes, l'amplification se réalise de manière exponentielle.

Contrairement à la PCR qualitative, la PCR quantitative autorise la quantification du fragment amplifié [Eurogentec, 2015]. Pour cela, on fait intervenir des composés organiques fluorescents qui se lient aux acides nucléiques. Chaque fragment amplifié génère alors un signal fluorescent. L'accumulation de fluorescence nous renseigne sur le nombre de cycles de PCR effectués, et donc sur la quantité d'ARN messager présent. En résumé, la qPCR autorise le suivi du processus d'amplification en détectant la fluorescence de chaque nouveau produit de la PCR [Clinisciences, 2019].

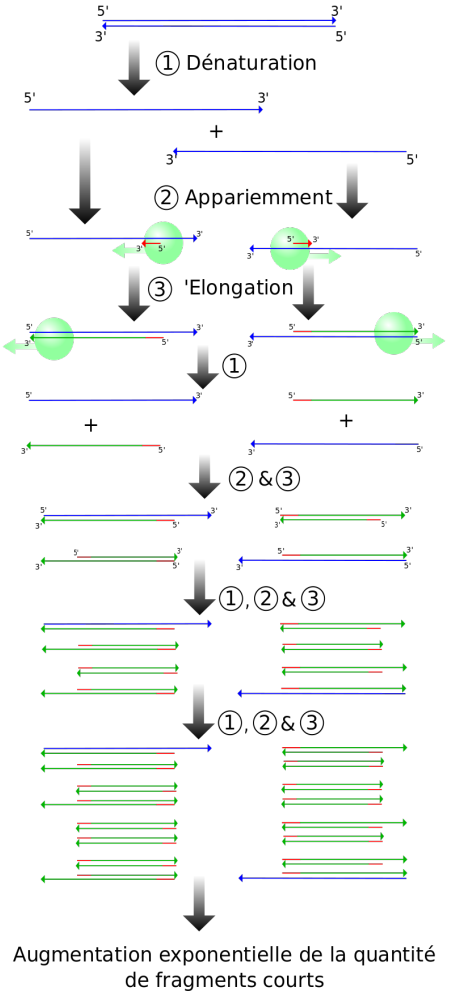


FIGURE 3: Diagramme des quatre premiers cycles de la PCR [Abdolmohammadi, 2017]

2.2 Présentation des données

Le contexte expérimental est le suivant :

- Plusieurs lignées cellulaires ont été analysées. Parmi elles, les cellules dites « contrôle », notées CT, correspondent aux cellules en conditions normales. Les lignées CTA, CTR, CTP, CTRPA, CTRPAt désignent des mutations des cellules CT, mutations dont on sait qu'elles correspondent à des susceptibilités au cancer du colon. Les cellules CTA sont les mutations les moins avancées, tandis que les cellules CTRPAt sont les mutations les plus avancées et les plus ressemblantes aux cellules cancéreuses du colon.
- Les cellules ont été exposées à différents types de polluants : 6-Benzylaminopurine, Pyrène, dioxine de Seveso, ainsi qu'un mélange de ces précédents pollants. Ils sont respectivement nommés BAP, Pyr, TCDD, et Mix. Les cellules ont également été exposée au solvant dimethyl sulfoxide, dénoté DMSO, dont le rôle est de maintenir les cellules en conditions contrôle (absence de polluant).
- Des mesures ont été prises à différents pas de temps (0 : conditions basales sans traitement, 48 heures et 120 heures d'exposition au polluant). Les mesures pour 48 heures et 120 heures d'exposition sont appariées et réalisées indépendamment des mesures en conditions basales.
- Ces mesures correspondent à plusieurs types de variables : respiratoires et génomiques.

  — Les variables respiratoires caractérisent la respiration cellulaire lors de la confrontation avec le polluant. Pendant cette période, les cellules sont introduites dans une machine et soumises à un « stress » augmentant progressivement. Le taux d'oxygène consommé pendant la respiration cellulaire (OCR) est observé à niveau de stress minimal (respiration dite basale), et à niveau

de stress maximal (respiration dite maximale). Ce taux est également mesuré pendant l'efflux de protons. Le taux d'acidification extra-cellulaire (ECAR) est, quant à lui, mesuré lors de la phase de glycolyse de la respiration cellulaire.

— Les variables dites génomiques correspondent aux expressions de 40 gènes. L'expression d'un gène a été mesurée en quantifiant le niveau de l'ARN messager produit par ce gène. Pour cela, les biologistes ont réalisé une transcription inverse suivie d'une amplification de type qPCR. Les gènes étudiés sont les suivants : *6P6D*, *ACO1*, *AhR*, *AhRR*, *ATP5IF1*, *CAT*, *CYP1A1*, *CYP1A2*, *CYP3A4*, *ENIO1*, *FH*, *G6PD*, *gene_inc*, *HK2*, *HMGCR*, *HMOX*, *IDH1*, *LDHA*, *LDHB*, *LPCAT*, *MCT4*, *MFN2*, *Mitoferrin1*, *Mitoferrin2*, *ND1*, *NHE1*, *NQO1*, *NRF2*, *PKM1*, *PKM2*, *PRDX1*, *RDK1*, *SCD1*, *SDHA*, *SDHC*, *SIRT3*, *TFAM*, *TIGAR*, *TSPO*, *UQCC3*.

Plusieurs observations simultanées des variables respiratoires, d'une part, et des variables génomiques, d'autre part, ont été réalisées. Les variables respiratoires ont été observées indépendamment des variables génomiques.

Deux dispositif expérimentaux ont été mis en place.

Dispositif 1 : Nous disposons d'une population de cellules de 6 types différents. Cette population résulte de plusieurs cultures cellulaires. Les cellules composant une même culture sont toutes issues de la même lignée, et chaque lignée comporte plusieurs cultures. Les cellules sont disposées sur des plaques, de manière à ce que chaque plaque utilisée pour l'étude en conditions basales contienne 2 (ou 3 selon les cas) cultures cellulaires complètes, et à ce que chaque plaque utilisée pour les études à 48 et 120 heures ne contienne qu'une seule culture cellulaire. Chaque plaque se divise en 24 puits, que l'on appelle « réplicats ». Les plaques sont introduites dans une machine, pour y réaliser des mesures. Les mesures effectuées sur les cellules provenant d'une même culture constituent une même « expérience ».
En définissant l'expérience de cette manière, on notera que chaque plaque regoupe plusieurs expériences, mais que chaque expérience ne peut appartenir qu'à une seule plaque.

Des mesures ont donc été réalisées à 3 niveaux :

- Mesures sur chaque puits, dénoté « réplicat », appartenant à une plaque.
- Mesures sur les réplicats d'une même expérience.
- Mesures sur l'ensemble des cellules d'une même plaque.

Le **dispositif 1** est représenté en figure 4. Il est utilisé pour mesurer les variables respiratoires.

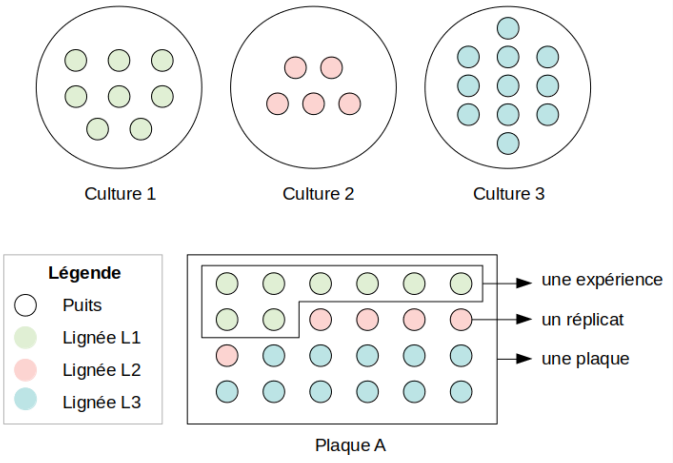


FIGURE 4: Représentation schématique du **dispositif 1**

Dispositif 2 : Nous disposons d'une population de cellules de 6 types différents. Nous y mesurons l'expression des 40 gènes précédemment définis. Chaque mesure est effectuée indépendamment des autres. Dans ces conditions, une expérience correspond à un réplicat. Le **dispositif 2** est utilisé pour mesurer les variables génomiques.

Les données nous ont été transmises par les biologistes sous la forme d'un tableur Excel organisé sur cinq onglets : « conditions basales », « conditions basales (2)», « traitement 48h », « traitement 120h », « qPCR conditions basales ». Chaque onglet contenait lui-même plusieurs tableaux, comme sur la figure 5. Les codes couleurs sont les suivants :

- Dans tous les onglets à l'exception de « conditions basales (2)», chaque couleur d'un même onglet correspond à une expérience.
- Dans l'onglet « conditions basales (2)», une couleur correspond à une plaque. Le plan de chaque plaque est joint aux tableaux de mesures. Une partie de cet onglet est représentée en figure 6.

L'onglet renseignant sur les répartitions des cellules par plaques (onglet « conditions basales (2)») ne nous a été communiqué que très tardivement. De nouvelles données, transmises elles aussi tardivement, ont dû être prises en compte en cours d'analyse. D'autres, incohérentes dans la manière dont elles étaient retranscrites, ont été signalées aux biologistes. Enfin, certains détails des dispositifs expérimentaux mis en place ne nous ont été transmis qu'en septembre. Il nous a donc fallu du temps pour parvenir à une compréhension totale des données, ce qui a nécessité de recommencer les analyses à plusieurs reprises.

À titre d'exemple, il nous a d'abord été dit qu'au sein d'un même onglet, une couleur codait pour une expérience. Nous avons donc considéré qu'une expérience pouvait comprendre plusieurs types cellulaires, comme en témoigne la figure 5. En réalité, les expériences sont indépendantes entre types cellulaires : au sein d'un même onglet, une expérience est codée par un couple *type cellulaire* × *couleur*.

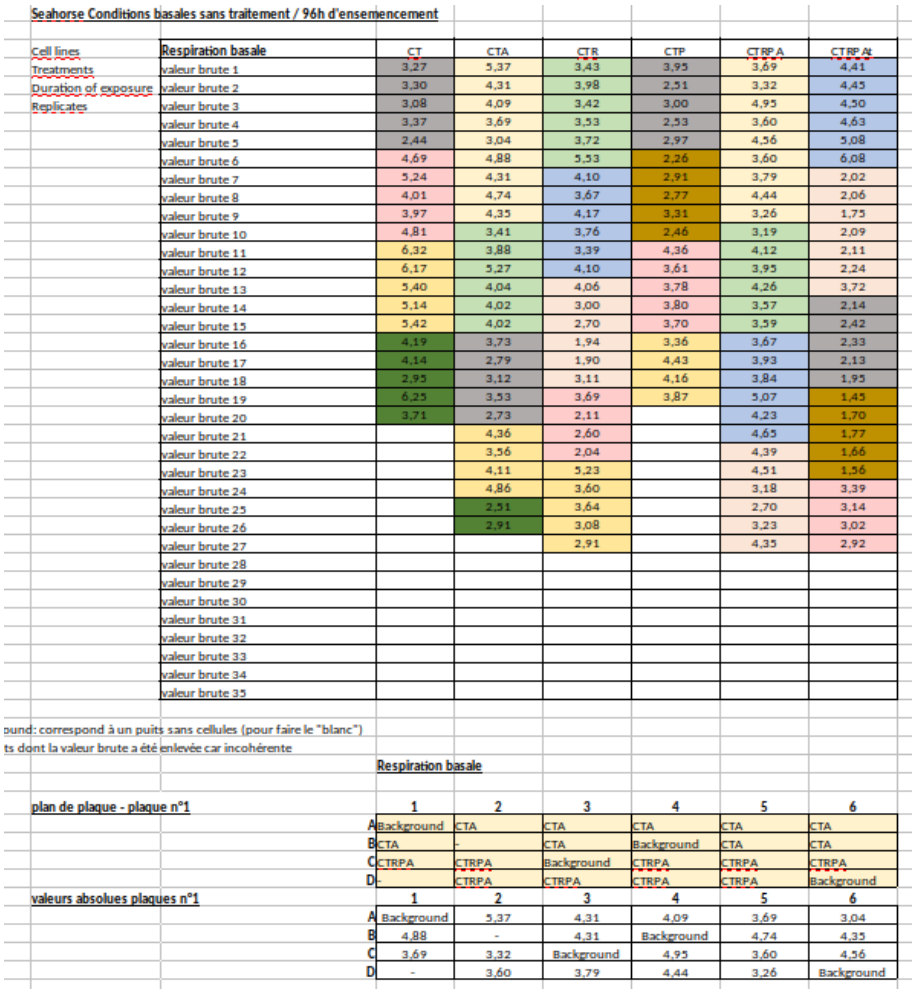
	CT				
48h	DMSO	BaP	Pyrene	TCDD	Mix
valeur brute 1	11,24	8,47	6,07	8,37	8,48
valeur brute 2	8,24	8,09	8,96	9,40	9,32
valeur brute 3	7,27		7,45	7,75	8,75
valeur brute 4	14,59	9,03			
valeur brute 5	15,73				
valeur brute 6	20,20	13,77	19,42	20,61	21,47
valeur brute 7	14,24	21,39	22,05	18,21	17,38
valeur brute 8	14,55	27,49	23,43	17,92	24,92
valeur brute 9	19,59	18,58	19,65	13,84	13,56
valeur brute 10	7,90	10,29	11,51	20,02	11,53
valeur brute 11	12,53	13,07	19,58	17,87	15,83
valeur brute 12	12,49	12,46	20,07	14,07	
valeur brute 13	11,47	16,64		11,91	17,45
valeur brute 14	12,01	10,78	15,30	13,83	11,42
valeur brute 15	7,47	7,82	9,09	10,54	10,71
valeur brute 16	9,24	10,34	11,79	10,78	10,59
valeur brute 17	7,74	8,87			12,51
valeur brute 18					
valeur brute 19					
valeur brute 20					

	CTA				
48h	DMSO	BaP	Pyrene	TCDD	Mix
valeur brute 1		38,43	33,68	48,32	30,16
valeur brute 2	31,72	31,15	28,49	46,96	42,56
valeur brute 3		34,24		43,62	25,88
valeur brute 4	36,42		30,31	48,55	
valeur brute 5	18,97	22,08	23,09	18,80	18,57
valeur brute 6	16,38	15,97	15,82	21,09	22,37
valeur brute 7		19,87	17,48	19,34	17,12
valeur brute 8	13,50	15,22	15,68	21,48	19,78
valeur brute 9	30,08	29,49	38,24	48,49	41,08
valeur brute 10	32,69	25,64	29,53	34,90	46,36
valeur brute 11	39,91	31,11	35,01		38,26
valeur brute 12	37,94				37,00
valeur brute 13	24,35	20,57	19,96	22,86	20,73
valeur brute 14	14,66	18,04	15,68	24,07	19,66
valeur brute 15		25,72	24,84	15,78	14,61
valeur brute 16	15,47		25,66		15,78
valeur brute 17					
valeur brute 18					
valeur brute 19					
valeur brute 20					

	CT				
48h	DMSO	BaP	Pyrene	TCDD	Mix
valeur brute 1	1,59	1,07	1,17	1,44	1,54
valeur brute 2	1,66	1,20	1,23	1,66	1,38
valeur brute 3	1,60		1,61	2,04	1,34
valeur brute 4	1,68	1,23			
valeur brute 5	1,74				
valeur brute 6	1,94	1,44	1,32	2,00	1,67
valeur brute 7	1,25	1,17	1,92	1,73	1,06
valeur brute 8	1,25	1,79	1,54	1,06	2,04
valeur brute 9	1,55	1,69	1,30	1,32	1,05
valeur brute 10	0,62	1,13	0,98	0,93	1,30
valeur brute 11	1,05	1,00	1,51	0,11	1,82
valeur brute 12	1,09	0,91	1,54	1,42	0,40
valeur brute 13	1,06	1,02		0,98	1,49
valeur brute 14	0,80	0,87	0,86	0,81	0,81
valeur brute 15	0,61	0,55	0,73	0,64	0,76
valeur brute 16	0,71	0,34	0,66	0,74	0,75
valeur brute 17	0,66	0,64			0,96
valeur brute 18					
valeur brute 19					
valeur brute 20					

	CTA				
48h	DMSO	BaP	Pyrene	TCDD	Mix
valeur brute 1	7,00	5,16	4,45	6,31	3,76
valeur brute 2	4,40	3,79	4,04	5,99	6,05
valeur brute 3	7,24	3,81	5,73	6,37	3,61
valeur brute 4	5,22	5,41	3,79	6,16	
valeur brute 5	3,06	3,18	3,48	2,91	3,15
valeur brute 6	2,31	2,62	2,57	2,84	3,46
valeur brute 7	1,66	3,52	2,75	2,99	3,00
valeur brute 8	2,20	2,20	2,46	3,56	2,97
valeur brute 9	1,67	1,73	2,56	4,89	5,55
valeur brute 10	4,85	4,24	1,44	1,97	3,07
valeur brute 11	3,51	5,79	1,43		2,47
valeur brute 12	2,10				2,34
valeur brute 13	1,13	1,11	0,91	1,45	1,58
valeur brute 14	0,92	0,86	0,92	1,16	1,29
valeur brute 15	2,94	0,81	2,42	1,39	1,14
valeur brute 16	1,09		1,10	1,28	1,09
valeur brute 17					
valeur brute 18					
valeur brute 19					
valeur brute 20					

FIGURE 5: Capture d'écran d'une partie de l'onglet « traitement 48h » du tableur transmis par les biologistes



Seahorse Conditions basales sans traitement / 96h d'ensemencement

Cell lines	Respiration basale	CT	CTA	CTR	CTP	CTRPA	CTRPAt
Treatments	valeur brute 1	3,27	5,37	3,43	3,95	3,69	4,41
Duration of exposure	valeur brute 2	3,30	4,31	3,98	2,51	3,32	4,45
Replicates	valeur brute 3	3,08	4,09	3,42	3,00	4,95	4,50
	valeur brute 4	3,37	3,69	3,53	2,53	3,60	4,63
	valeur brute 5	2,44	3,04	3,72	2,97	4,56	5,08
	valeur brute 6	4,69	4,88	5,53	2,26	3,60	6,08
	valeur brute 7	5,24	4,31	4,10	2,91	3,79	2,02
	valeur brute 8	4,01	4,74	3,67	2,77	4,44	2,06
	valeur brute 9	3,97	4,35	4,17	3,31	3,26	1,75
	valeur brute 10	4,81	3,41	3,76	2,46	3,19	2,09
	valeur brute 11	6,32	3,88	3,39	4,36	4,12	2,11
	valeur brute 12	6,17	5,27	4,10	3,61	3,95	2,24
	valeur brute 13	5,40	4,04	4,06	3,78	4,26	3,72
	valeur brute 14	5,14	4,02	3,00	3,80	3,57	2,14
	valeur brute 15	5,42	4,02	2,70	3,70	3,59	2,42
	valeur brute 16	4,19	3,73	1,94	3,36	3,67	2,33
	valeur brute 17	4,14	2,79	1,90	4,43	3,93	2,13
	valeur brute 18	2,95	3,12	3,11	4,16	3,84	1,95
	valeur brute 19	6,25	3,53	3,69	3,87	5,07	1,45
	valeur brute 20	3,71	2,73	2,11		4,23	1,70
	valeur brute 21		4,36	2,60		4,65	1,77
	valeur brute 22		3,56	2,04		4,39	1,66
	valeur brute 23		4,11	5,23		4,51	1,56
	valeur brute 24		4,86	3,60		3,18	3,39
	valeur brute 25		2,51	3,64		2,70	3,14
	valeur brute 26		2,91	3,08		3,23	3,02
	valeur brute 27			2,91		4,35	2,92
	valeur brute 28						
	valeur brute 29						
	valeur brute 30						
	valeur brute 31						
	valeur brute 32						
	valeur brute 33						
	valeur brute 34						
	valeur brute 35						

ound: correspond à un puits sans cellules (pour faire le "blanc")
ts dont la valeur brute a été enlevée car incohérente

Respiration basale

plan de plaque - plaque n°1

	1	2	3	4	5	6
A	Background	CTA	CTA	CTA	CTA	CTA
B	CTA	-	CTA	Background	CTA	CTA
C	CTRPA	CTRPA	Background	CTRPA	CTRPA	CTRPA
D	-	CTRPA	CTRPA	CTRPA	CTRPA	Background

valeurs absolues plaques n°1

	1	2	3	4	5	6
A	Background	5,37	4,31	4,09	3,69	3,04
B	4,88	-	4,31	Background	4,74	4,35
C	3,69	3,32	Background	4,95	3,60	4,56
D	-	3,60	3,79	4,44	3,26	Background

FIGURE 6: Capture d'écran d'une partie de l'onglet « conditions basales (2)» du tableur transmis par les biologistes

2.3 Problématique du stage

Mon stage s'inscrit dans le cadre du projet METAhCOL, financé par l'Agence Nationale de la Recherche (ANR) et impliquant plusieurs laboratoires de l'INRA et de l'Institut National de la Santé et de la Recherche Médicale (INSERM). Il repose sur l'exposition de cellules à de faibles doses de polluants, seuls ou combinés.
J'ai travaillé au sein de l'équipe SaAB de l'unité MIAT du centre Toulouse Midi-Pyrénées de l'INRA. Dans le cadre du projet METAhCOL, il m'a été proposé d'analyser des données respiratoires et génomiques provenant de la culture de cellules exposées à des polluants retrouvés en faibles quantités dans l'alimentation. Le stage se propose de répondre à plusieurs questions : Comment se distinguent les différents types de cellules avant exposition (conditions basales) ? Comment se distingue chaque type cellulaire face aux différentes expositions ? Comment se distinguent les effets des expositions entre les lignées cellulaires ?

3 Méthodologies et techniques

3.1 Organisation

3.1.1 Outils utilisés

La phase de préparation des données a été entièrement réalisée avec les logiciels Microsoft Excel et LibreOffice. La phase d'analyse a nécessité l'utilisation de R à travers RStudio, plus précisément de scripts RMarkdown. Le partage de mes travaux avec mes encadrantes s'effectuait via Git.

3.1.2 Déroulé d'une semaine-type

Mon travail à l'INRA s'est organisé selon un cycle hebdomadaire de 5 jours, et un volume horaire hebdomadaire de 35h. Chaque journée s'étendait de 8h à 16h, avec une pause-repas de 12h à 13h. Une réunion hebdomadaire avec mes encadrantes de stage, Mme Vialaneix et Mme Lefort, avait lieu chaque lundi après-midi. Lors de ces réunions, j'exposais mes travaux de la semaine passée sous la forme d'une présentation LibreOffice. Cette présentation relatait les principaux résultats obtenus suite à mes analyses. Elle décrivait aussi ce que j'avais prévu de faire, ce que j'avais réellement réussi à faire, et les difficultés rencontrées. À la fin de chaque réunion, nous discutions ensemble de ces résultats et mes encadrantes m'indiquaient la direction à suivre pour la suite de l'analyse. Le reste de la semaine, je travaillais en autonomie dans un bureau partagé avec Mme Moisan, Ingénieure d'Etude en Biostatistique. Mes encadrantes restaient disponibles par mail en cas de difficultés, et répondaient toujours rapidement à mes interrogations, me permettant de solutionner rapidement mes problèmes et de maintenir un bon rythme de travail. Si nécessaire, une rencontre était organisée au cours de la semaine pour discuter de ces problèmes. En fin de semaine, je partageais le travail effectué sur le dépôt Git et dressait la liste de mes objectifs pour la semaine suivante.
Au cours de mon stage, il m'a également été demandé de rédiger trois rapports de mes analyses à destination des biologistes du projet METAhCOL, et de prendre part à des réunions avec eux afin de leur communiquer mes résultats. Le rapport concernant la caractérisation des lignées cellulaires en conditions basales est joint en Annexe B.

3.2 Analyse exploratoire

L'objectif de l'analyse exploratoire était de caractériser les distributions de chaque variable en termes de symétrie et de variabilité. Ces informations devaient nous permettre de déterminer quelles analyses effectuer par la suite.

Dans un premier temps, nous avons cherché à savoir si les distributions étaient, ou non, gaussiennes. Pour cela, nous avons effectué des tests de Shapiro-Wilk d'hypothèses :

$$H_0 : \text{La variable possède une distribution gaussienne}$$
$$\text{contre}$$
$$H_1 : \text{La variable ne possède pas de distribution gaussienne}$$

Nous rejetions l'hypothèse nulle H_0 si la p-valeur renvoyée par le test était inférieure à 5%. Dans ce cas, nous pouvions affirmer, avec un niveau de confiance de 95%, que la distribution considérée n'était pas gaussienne.

Dans un second temps, nous avons représenté nos distributions sous la forme de diagrammes-boîtes parallèles, afin de déterminer d'éventuelles différences entre les distributions. Dans le cas de différences notables, nous programmions, pour la suite de l'étude, de réaliser des tests de comparaison inter-groupes. Ces représentations nous ont également renseigné sur la présence, ou non, d'un phénomène de variations aléatoires lié à l'un des paramètres. Nous parlons d'aléa.

À la fin de ces premières analyses, nous avons établi le plan d'étude suivant :

- Tests non-paramétriques et paramétriques de comparaison inter-groupes ne tenant pas compte de l'éventuel aléa. Les tests non-paramétriques seraient réalisés en cas de distributions non-gaussiennes, les tests paramétriques en cas de distributions gaussiennes.
- Tests paramétriques de comparaison inter-groupes tenant compte de l'aléa (modèles mixtes généralisés).

3.3 Tests de comparaison

Dans notre étude, nous avons souhaité caractériser des groupes d'individus lorsqu'ils étaient soumis à des conditions (durée d'exposition, type de polluant, ...) précises. Pour ce faire, nous avons dressé des comparaisons inter-groupes des distributions de variables. Des tests de comparaison ont été effectués afin de déterminer quelles différences inter-groupes pouvaient être jugées significatives.

3.3.1 Principe d'un test d'hypothèses

Un test d'hypothèses est un outil statistique utilisé pour évaluer, avec un niveau de confiance $1-\alpha$, si une variable aléatoire vérifie une certaine hypothèse H_0 appelée *hypothèse nulle*, ou H_1, appelée *hypothèse alternative*. Il s'agit d'une procédure de décision entre H_0 et H_1. L'hypothèse nulle est celle que l'on considère, à priori, comme étant vraie. Le test nous permet de vérifier la pertinence de cet a priori. L'hypothèse alternative est complémentaire à l'hypothèse nulle.
Un test d'hypothèses se caractérise par :

- n copies indépendantes $(Y_1, ..., Y_n)$ de la variable aléatoire Y.
- Une statistique de test S dépendante de $\hat{Y}$, ainsi que sa distribution associée sous l'hypothèse nulle H_0.
- Une zone de rejet R_α assurant $P_{H_0}(S \in R_\alpha) = \alpha$.
  $P_{H_0}(S \in R_\alpha)$ est appelée *erreur de première espèce*.

Nous considérons, dans ce rapport, deux tests :

- Le test de *comparaison inter-groupes*, utilisé pour comparer un ensemble de groupes.
- Le test *post-hoc*, utilisé pour comparer des groupes deux-à-deux.

3.3.2 Tests directs de conditions expérimentales

Cette section reprend les concepts explicités dans les cours *Lecture 2 - One-Way Analysis* [Singull, 2019a], *Lecture 3 - Pairwise comparisons* [Singull, 2019b] et *Lecture 4 - Non-parametric methods* [Singull, 2019c] du module *Experimental Design and Biostatistics* enseigné à l'université de Linköping en 2018.

3.3.2.1 Test de comparaison inter-groupes

Cas de données gaussiennes : ANOVA
Soit Y une variable quantitative observée sur a groupes. Pour chaque groupe $j \in \{1, ..., a\}$, n_j mesures indépendantes de Y ont été réalisées. On note $Y_{\cdot j} = (Y_{1j}, ..., Y_{n_j j})$, Y_{ij} le $i$ème élément de $Y_{\cdot j}$, et y_{ij} la valeur observée de Y_{ij}. On définit y_{ij} comme la $i$ème observation de Y au sein du groupe j.
Nous disposons donc de a échantillons aléatoires indépendants $(Y_{11}, ..., Y_{n_1 1}), ..., (Y_{n_1 a}, ..., Y_{n_a a})$, et de leurs a séries d'observations associées $(y_{11}, ..., y_{n_1 1}), ..., (y_{n_1 a}, ..., y_{n_a a})$.

	Observations
Groupe 1	y_{11}, y_{21}, ..., $y_{n_1 1}$
Groupe 2	y_{12}, y_{22}, ..., $y_{n_2 2}$
.	.
.	.
.	.
Groupe a	y_{1a}, y_{2a}, ..., $y_{n_a a}$

Modèle
Y_{ij} peut s'écrire selon le modèle suivant :

$$Y_{ij} = \mu_j + \epsilon_{ij}$$

où :

- μ_j est *l'effet du groupe j* (moyenne des valeurs de Y pour le groupe j). Elle s'écrit aussi sous la forme $\mu_j = \mu + \tau_j$ avec μ la moyenne des valeurs sur l'ensemble des groupes, et τ_j l'effet fixe présent au sein du groupe j.
- ϵ_{ij} est une variable aléatoire représentant l'erreur, telle que $\epsilon_{ij} \sim \mathcal{N}(0, \sigma^2), \forall i = 1, ..., n_j, \forall j = 1, ..., a$. Les variables aléatoires ϵ_{ij} sont indépendantes.

Les ϵ_{ij} étant de loi normale centrée réduite, il en découle que Y_{ij} suit une loi normale d'espérance μ_j et de variance σ^2.

La condition

$$\sum_{j=1}^{a} n_j \tau_j = 0$$

assure l'unicité des estimateurs des paramètres $\tau_1, ..., \tau_a$.

Hypothèses
Nos données étant normalement distribuées, la comparaison inter-groupes est réalisée à l'aide d'une **ANOVA** (Analysis Of Variance). Nous testons les hypothèses suivantes :

$$H_0 : \mu_1 = ... = \mu_a$$

contre

$$H_1 : \exists (l, m) \in \{1, ..., a\} \times \{1, ..., a\}, l \neq m \text{ et } \mu_l \neq \mu_m$$

Estimateurs et sommes des carrés
On définit les estimateurs suivants :

- μ_j est estimé par $\hat{\mu}_j = \bar{y}_{.j} = \frac{1}{n_j} \sum_{i=1}^{n_j} y_{ij}$
- μ est estimé par $\hat{\mu} = \bar{y}_{..} = \frac{1}{n} \sum_{j=1}^{a} \sum_{i=1}^{n_j} y_{ij} = \frac{1}{n} \sum_{j=1}^{a} n_j \bar{y}_{.j}$, où $n = \sum_{j=1}^{a} n_j$
- τ_j est estimé par $\hat{\tau}_j = \hat{\mu}_j - \hat{\mu} = \bar{y}_{.j} - \bar{y}_{..}$ puisque $\mu_j = \mu + \tau_j$

La contrainte (1) mène à $E(\hat{\mu}_j) = \mu_j$ et $E(\hat{\mu}) = \mu$. Les estimateurs $\hat{\mu}_j$ et $\hat{\mu}$ sont donc sans biais.
On introduit la somme des carrées totale SS_T, définie par

$$SS_T = \sum_{j=1}^{a} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{..})^2$$

Elle peut être décomposée en

$$SS_T = SS_E + SS_{TREAT}$$

où :

- SS_E est la somme des carrés intra-groupes.

$$SS_E = \sum_{j=1}^{a} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{.j})^2 = \sum_{j=1}^{a} (n_j - 1) s_j^2 = (n - a) s^2$$

  où $s_j^2 = \frac{1}{n_j} \sum_{i=1}^{n_j} (y_{ij} - \bar{y}_{.j})^2$ représente la variance du groupe j, et $s^2 = \frac{(n_1 - 1)s_1^2 + ... + (n_a - 1)s_a^2}{n - a}$ est la variance *groupée* des a groupes.

- SS_{TREAT} est la somme des carrés inter-groupes.

$$SS_{TREAT} = \sum_{j=1}^{a} \sum_{i=1}^{n_j} (\bar{y}_{.j} - \bar{y}_{..})^2 = \sum_{j=1}^{a} (\hat{\mu}_j - \hat{\mu})^2 = \sum_{j=1}^{a} n_j \hat{\tau}_j^2$$

On peut montrer que :

1. SS_{TREAT} et SS_E sont indépendants.
2. $\frac{SS_E}{\sigma^2} \sim \chi^2(n - a)$.
3. Si $\tau_1 = ... = \tau_a = 0$, alors $\frac{SS_{TREAT}}{\sigma^2} \sim \chi^2(a - 1)$

Statistique de test
La statistique de test est donnée par

$$v = \frac{SS_{treat}/(a-1)}{SSE/(n-a)}$$

Lorsque H_0 est vraie, v suit une loi de Fisher $F(a-1, n-a)$.
En effet, $\frac{SS_E}{\sigma^2} \sim \chi^2(n-a)$ est toujours vrai, et, sous H_0, $\frac{SS_{TREAT}}{\sigma^2} \sim \chi^2(a-1)$.
v s'exprime donc comme $v = \frac{V_1/(a-1)}{V_2/(n-a)}$, où V_1 et V_2 sont deux variables aléatoires telles que $V_1 \sim \chi^2(a-1)$ et $V_2 \sim \chi^2(n-a)$.
Ainsi, on retrouve bien

$$v = \frac{\frac{SS_{treat}/\sigma^2}{a-1}}{\frac{SSE/\sigma^2}{n-a}} \sim F(a-1, n-a)$$

.

Zone de rejet
SS_{TREAT} s'exprimant comme une combinaison linéaire des carrés des τ_i, H_0 est rejetée au niveau α lorsque SS_{TREAT} est grand, et donc lorsque v est grand. On souhaite déterminer c telle que H_0 soit rejetée pour $v \geqslant c$ tout en assurant une erreur de première espèce de α. On déduit facilement que c correspond au quantile d'ordre $1-\alpha$ d'une loi de Fisher $F(a-1, n-a)$, noté $F_{1-\alpha}(a-1, n-a)$. Ce quantile est représenté sur la figure 7.

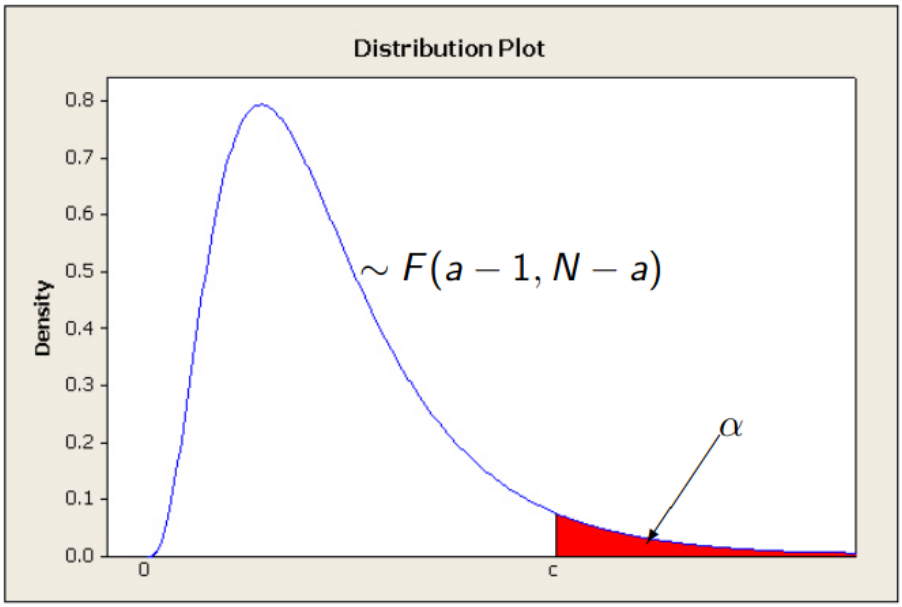


FIGURE 7: Distribution de Fisher [Singull, 2019c]

Ainsi,

$$R_\alpha = \{v \geqslant F_{1-\alpha}(a-1, n-a)\}$$

Dans R, les ANOVA ont été réalisées avec la fonction `aov()` du package `stats`.

Cas de données non-gaussiennes : test de Kruskal-Wallis
Soit Y une variable quantitative observée sur a groupes. Pour chaque groupe $j \in \{1, ..., a\}$, n_j mesures indépendantes de Y ont été réalisées. On note $Y_{.j} = (Y_{1j}, ..., Y_{n_j j})$ le vecteur des variables aléatoires Y observées sur le groupe j, Y_{ij} le $i$ème élément de $Y_{.j}$, et y_{ij} la valeur observée de Y_{ij}.
Nous disposons donc de a échantillons aléatoires indépendants $(Y_{11}, ..., Y_{n_1 1}), ..., (Y_{n_1 a}, ..., Y_{n_a a})$, et de leurs a séries d'observations associées $(y_{11}, ..., y_{n_1 1}), ..., (y_{n_1 a}, ..., y_{n_a a})$.
Supposons, cette fois, que l'hypothèse de normalité pour Y_{ij} n'est pas vérifiée. On note $\mathcal{L}_j(Y)$ la loi de la variable aléatoire Y sur le $j$ème groupe.

Hypothèses
Nos données n'étant plus gaussiennes, la comparaison de moyennes avec une ANOVA n'est plus réalisable. Nous mettons en place un test de comparaison non-paramétrique de **Kruskal-Wallis**. Ce type de test s'applique à une toute distribution continue (symétrique ou non). Nous souhaitons tester les hypothèses suivantes :

$$H_0 : \mathcal{L}_1(Y) = ... = \mathcal{L}_i(Y) = ... = \mathcal{L}_a(Y)$$

contre

$$H_1 : \text{ les lois } \mathcal{L}_1(Y), ..., \mathcal{L}_a(Y) \text{ ne sont pas toutes identiques.}$$

Hiérarchisation des observations
Le test de Kruskal-Wallis consiste à fusionner l'ensemble des groupes en un même groupes possédant $n = \sum_{j=1}^{a} n_j$ mesures, et à les ordonner de manière croissante en leur attribuant des rangs de 1 à n. On définit alors :

- r_{ij} : rang attribué à l'observation y_{ij}
- s_j la somme des rangs attribués aux observations du groupe j, définie par

$$s_j = \sum_{j=1}^{n_i} r_{ij}$$

- $S_a = \sum_{j=1}^{a} \frac{s_j^2}{n_j}$

S_a est l'équivalent non-paramétrique de l'estimateur dont dépend la statistique de test dans un test paramétrique.

Statistique de test
La statistique de test est donnée par

$$T = \begin{cases} \frac{(n-1)(S_a - C)}{S_r - C} & \text{si les données contiennent plusieurs valeurs identiques} \\ \frac{12 S_a}{n(n+1)} - 3(n+1) & \text{sinon} \end{cases}$$

avec
$S_r = \sum_{i,j} r_{ij}^2$ et $C = \frac{1}{4} n(n+1)^2$.

Zone de rejet
T peut être réécrite de la manière suivante :

$$T = \frac{12}{n(n+1)} \sum_{j=1}^{a} n_j \left(\frac{s_j}{n_j} - \frac{n+1}{2} \right)^2$$

Établir la statistique de test revient donc à comparer le rang moyen $\frac{n_j}{s_j}$ du groupe j avec le rang moyen $\frac{n+1}{2}$ de l'ensemble des groupes. On retrouve l'expression de la somme des carrés inter-groupes $SS_{TREAT} = \sum_{j=1}^{a} n_j (\bar{y_{.j}} - \bar{y_{..}})^2$.
Si tous les groupes proviennent d'une même distribution, alors tous les $\frac{s_j}{n_j}$ valent le rang moyen de l'ensemble des groupes $\frac{n+1}{2}$. H_0 est donc rejetée pour de grandes valeurs de SS_{TREAT}, et, par conséquent, de T. On souhaite déterminer c telle que H_0 soit rejetée pour $v > c$. Deux cas se présentent :

1. Pour de faibles valeurs de n_1,...,n_a, c est donné par la table de Kruskal-Wallis, pour une marge d'erreur $\alpha = P_{H_0}(T \geqslant c)$.
2. Pour de grandes valeurs de n_1,...,n_a, on considère $T \sim \chi^2(a-1)$ sous H_0. Ainsi, c correspond au quantile d'ordre $1-\alpha$ d'une loi $\chi^2(a-1)$, noté q_{a-1}. On a alors :

$$R_\alpha = \{T \geqslant q_{a-1}\}$$

Dans R, les tests de Kruskal-Wallis ont été réalisés avec la fonction `kruskal.test()` du package `stats`.

3.3.2.2 Test post-hoc

Un test post-hoc se définit comme une « procédure qui permet de comparer des groupes sans qu'une hypothèse sur la relation entre ces groupes ait été posée avant d'examiner les données » [Raufaste, 2013]. Par exemple, on peut réaliser des tests post-hoc pour comparer des groupes deux-à-deux lorsque d'importantes différences de distributions entre ces groupes sont mises en évidence par des tests de comparaison de type ANOVA ou Kruskal-Wallis.

Soit Y la variable quantitative précédemment introduite. Y est observée sur a groupes.
Pour tout groupe $j \in \{1, ..., a\}$, n_j mesures indépendantes de Y ont été réalisées.
Soient $(l, m) \in \{1, ..., a\} \times \{1, ..., a\}$, $l < m$.

Comparaisons deux-à-deux de groupes dans le cas de données gaussiennes
Supposons que les mesures effectuées sur les groupes l et m suivent des lois gaussiennes. Effectuer des comparaisons entre les groupes l et m revient à tester les hypothèses suivantes :

$$H_0 : \mu_l = \mu_m$$
contre
$$H_1 : \mu_l \neq \mu_m$$

Dans R, ces comparaisons deux-à-deux sont réalisées à l'aide de tests de Tukey. Nous utilisons la fonction `pairwise.t.test()` du package `stats`.

Comparaisons deux-à-deux de groupes dans le cas de données non-gaussiennes
Supposons que les mesures effectuées sur les groupes l et m ne suivent pas de loi gaussienne. Effectuer des comparaisons les groupes l et m revient à tester les hypothèses suivantes :

$$H_0 : \mathcal{L}_l(Y) = \mathcal{L}_m(Y)$$
contre
$$H_1 : \mathcal{L}_l(Y) \neq \mathcal{L}_m(Y)$$

Dans R, ces comparaisons deux-à-deux sont réalisées à l'aide de tests de Nemenyi. Nous utilisons la fonction `posthoc.kruskal.nemenyi.test()` du package `stats`. La p-valeur du test est, par défaut, déterminée par la méthode de Tukey. En cas de présence d'ex aequo dans les données, nous choisissons la méthode du χ^2.

3.3.3 Tests incluant des effets aléatoires

Soit Y la variable quantitative définie dans la partie *3.3.2 Tests directs de conditions expérimentales.* Lorsque l'on souhaite tenir compte, dans nos tests, de variations aléatoires liées à un phénomène connu, on construit un modèle incluant des effets dit *aléatoires* [Besse, 2013]. Le modèle est alors dit *mixte*. On définit communément l'effet aléatoire comme l'effet d'une variable dont les modalités sont sélectionnées aléatoirement parmi une population d'individus.

En plus des a groupes sur lesquels nous observions nos données, nous considérons b nouveaux groupes induits par le phénomène aléatoire. Pour chaque groupe $k \in \{1, ..., b\}$, n_k mesures indépendantes de Y ont été réalisées. On note $Y_{.jk} = (Y_{1jk}, ..., Y_{n_j n_k jk})$ le vecteur des variables aléatoires Y observées sur les groupes j et k, Y_{ijk} le $i$ème élément de $Y_{.jk}$, et y_{ijk} la valeur observée de Y_{ijk}. On définit y_{ijk} comme la $i$ème observation de Y au sein des groupes j et k.
Le modèle mixte se définit par :

$$Y_{ijk} = \mu_j + Z_k + \epsilon_{ijk}$$

où :

- μ_j est l'effet du groupe j (moyenne des valeurs de Y pour le groupe j), défini par $\mu_j = \mu + \tau_j$, avec μ la moyenne des valeurs sur les a groupes, et τ_j l'effet fixe présent au sein du groupe j.

- Z_k est l'effet aléatoire du groupe k (moyenne des valeurs de Y pour le groupe k). On note $Z = (Z_1, ..., Z_a)$ le vecteur des effets aléatoires. $Z \sim \mathcal{N}(0, \sigma_k^2)$
- ϵ_{ijk} est une variable aléatoire représentant l'erreur, telle que $\epsilon_{ijk} \sim \mathcal{N}(0, \sigma^2), \forall i = 1, ..., n_j, \forall j = 1, ..., a, \forall k = 1, ..., b$. Les variables aléatoires ϵ_{ijk} sont indépendantes.

Au contraire des modèles non-paramétriques, les modèles mixtes tiennent compte d'hypothèses. Ces derniers étant moins généraux que les modèles non-paramétriques, les tests de comparaison associés sont plus puissants que les tests de comparaison non-paramétriques. Pour obtenir des résultats fiables, il faut cependant s'assurer que les hypothèses d'application des modèles mixtes soient, dans leur majorité, vérifiées.

Étape 1 : Choix du modèle mixte d'ajustement
L'étape préliminaire à l'ajustement des données par un modèle mixte consiste à rechercher, pour la variable réponse, la distribution « connue » approchant le mieux la distribution réelle de nos données. Pour ce faire, on peut représenter les graphes de comparaison des quantiles empiriques de la distribution de cette variable avec les quantiles théoriques de diverses distributions. Sur chacun de ces graphes, une ligne continue représente les données dont nous devrions disposer pour qu'il n'y ait aucun ajustement à effectuer (i.e. les données étant exactement distribuées selon ce modèle), et des lignes discontinues représentent les bornes des intervalles dans lesquelles nos données doivent se trouver pour que le modèle soit considéré, avec un niveau de confiance de 95%, comme un modèle d'ajustement de bonne qualité. Nos données sont superposées à ces informations, sous la forme de nuages de points. Il ne reste alors qu'à choisir le modèle dont la « zone de confiance », délimitée graphiquement par les lignes discontinues, contient le maximum de points.

Dans R, les graphes de comparaison des quantiles ont été réalisés avec la fonction `qqplot()` du package `stats`.

Étape 2 : Ajustement des données par maximisation de la Quasi-Vraisemblance Pénalisée
Lorsque la distribution Γ est la distribution « connue » approchant le mieux celle de nos données, la résolution du problème de maximisation de la Vraisemblance se révèle trop complexe. Une solution est de maximiser la Quasi-Vraisemblance Pénalisée (*Penalized Quasi-Likelihood - PQL*). L'ajustement par maximisation de la Quasi-Vraisemblance Pénalisée est l'une des techniques les plus utilisée pour l'ajustement par modèles mixtes dans le cas de données non gaussiennes.
Notons que cette technique mène à la production d'estimations biaisées dans les cas suivants :

- la distribution approchant le mieux la distribution de la variable réponse considérée est discrète
- la variable réponse est binaire
- la moyenne des valeurs prises par la variable réponse est inférieure à 5

Dans R, l'ajustement par maximisation de la Quasi-Vraisemblance Pénalisée est mis en œuvre via la fonction `glmmPQL()` du package `nmle`.

3.3.3.1 Tests de comparaison

Soit le modèle mixte suivant :

$$Y_{ijk} = \mu_j + Z_k + \epsilon_{ijk}$$

Effectuer un test de comparaison inter-groupes sur ce modèle revient à effectuer un test d'hypothèses :

$$H_0 : \mu_1 = ... = \mu_a$$

contre

$$H_1 : \exists (l, m) \in \{1, ..., a\} \times \{1, ..., a\}, l \neq m \text{ et } \mu_l \neq \mu_m$$

Dans R, les tests de comparaison inter-groupes ont été réalisés avec la fonction `Anova()` du package `car`.

3.3.3.2 Tests post-hoc

Dans le cadre de modèles mixtes, les comparaisons deux-à-deux sont souvent effectuées à l'aide de Moyennes Marginales Estimées (*Estimated Marginal Means*). Considérons un modèle comprenant p variables qualitatives explicatives, q variables quantitatives explicatives, et une variable quantitative à expliquer Y. Les Moyennes Marginales Estimées de Y correspondent aux moyennes prises par Y pour chaque combinaison de modalités des p variables qualitatives explicatives [Lenth, 2018]. Pour effectuer les calculs, les q variables explicatives quantatives sont considérées comme fixes et égales à leurs moyennes sur l'ensemble des données [Lenth, 2018]. Des comparaisons de ces Moyennes Marginales Estimées deux-à-deux sont ensuite réalisées.
Dans R, les Moyennes Marginales Estimées ont été établies avec la fonction `emmeans()` du package `emmeans`. Les tests post-hoc ont ensuite été réalisés avec la fonction `pairs()` du package `emmeans`.

3.4 Analyses multivariées

Souhaitant caractériser des groupes d'individus dont les mesures se répartissent sur plusieurs variables, nous avons procédé à des analyses multivariées. L'analyse en Composantes Principales, analyse de type *descriptive*, nous a permis de résumer l'information en une représentation bidimensionnelle des données. L'analyse Factorielle Discriminante, analyse de type *explicative*, avait pour but de discriminer les groupes d'individus et de les caractériser vis-à-vis d'un ensemble de variables explicatives.

3.4.1 Analyse en Composantes Principales

Cette section reprend une partie des concepts introduits dans *Analyse en Composantes Principales (ACP)* [Besse, 2014].

Soient p variables quantitatives $Y^1, ...Y^p$, observées sur n individus. On note, $\forall i \in 1, .., n, \forall j \in 1, ..., p$: y_i^j la mesure de la variable Y^j sur l'individu i. Soit Y la matrice $n \times p$ suivante :

—	Y^1	...	Y^j	...	Y^p
1	y_1^1	...	y_1^j	...	y_1^p
·	·		·		·
·	·		·		·
·	·		·		·
i	y_i^1	...	y_i^j	...	y_i^p
·	·		·		·
·	·		·		·
·	·		·		·
n	y_n^1	...	y_n^j	...	y_n^p

Lorsque p est grand, il devient difficile de déterminer le comportement des variables et leurs corrélations deux à deux.
L'Analyse en Composantes Principales (ACP) se définit comme une « méthode factorielle de réduction de dimension pour l'exploration statistique de données quantitatives complexes » [Besse, 2014].

L'intérêt d'une telle réduction est de pouvoir visualiser les données issues de ces p variables quantitatives, dans un espace de dimension $q < p$, en déformant le moins possible la réalité. En d'autres termes, l'ACP consiste en une approximation d'une matrice de taille $n \times p$ par une matrice de même dimension mais de rang $q < p$. On souhaite obtenir :

- La représentation graphique optimale des individus, minimisant les déformations des nuages des points, dans un espace de dimension q dénoté E_q.
- La représentation graphique des variables dans un espace de dimension q dénoté F_q, explicitant au mieux les liaisons initiales entre ces variables.

On définit pour cela des composantes principales, combinaisons affines des variables $Y^1, ..., Y^p$. Chacune de ces composantes est une variable unidimensionnelle conservant le maximum de l'information contenue dans Y. Afin de fournir un résumé suffisant de cette information, il est généralement nécessaire de définir plusieurs composantes principales. Elles sont notées C^1, C^2, ... C^p, possèdent une variance maximale, et ne sont pas corrélées entre elles.
La $j$ème composante principale C^j est obtenue par la relation $C^j = (Y - E(Y))^T v_j$, où $v_j \in \mathbb{R}^p$ est solution du problème d'optimisation suivant :

$$C^j = argmax_{C=(Y-E(Y))^T v, v\in\mathbb{R}^p, vv^T=1} Var(C) \quad (1)$$

Les solutions de (1) sont données par $C^j = (Y - E(Y))^T v_j$, où v_j est le vecteur propre orthonormé de la matrice $Var(Y)$ associé à la plus grande valeur propre λ_j. Autrement dit, v_j vérifie $Var(Y)v_j = \lambda_j v_j$. De tels vecteurs v_j sont appelés *vecteurs principaux*. Ils engendrent des sous-espaces vectoriels unidimensionnels appelés *axes principaux*.
Une fois les composantes principales définies, nos données p-dimensionnelles sont représentées dans l'espace constitué des deux combinaisons principales renfermant la majorité de l'information (i.e. renfermant, à elles deux, la plus grande part de $Var(Y)$). Cette représentation bidimensionnelle facilite l'interprétation des données.

Dans R, les ACP ont été réalisées avec la fonction `PCA()` du package `FactoMineR`.

3.4.2 Analyses Discriminantes

Soient p variables explicatives $Y^1, ...Y^p$ quantitatives, et une variable à expliquer X, qualitative, possédant m modalités $\{\delta_1, ..., \delta_m\}$. n observations sont réalisées sur chaque variable.

3.4.2.1 Analyse Factorielle discriminante

Cette section reprend les concepts et notations introduits dans *Analyse Factorielle Discriminante (AFD)* [Chavent, 2015].
L'Analyse Factorielle Discriminante (AFD) est une méthode de réduction de dimension pour l'analyse de jeux de données constituées d'une variable qualitative et de $p > 1$ variables quantitatives. Son objectif est de déterminer quelles combinaisons linéaires des variables $Y^1, ...Y^p$ mènent à la meilleure discrimination des groupes définis par les m modalités de X.
On note Ω l'ensemble des n individus observés. Posons T la matrice $n \times m$ des indicatrices des modalités de X, de terme général :

$$t_i^k = \begin{cases} 1 & \text{si le } i\text{ème individu appartient au groupe } k \\ 0 & \text{sinon} \end{cases}$$

On définit, $\forall k = 1, ..., m, \Omega_k$, d'effectif n_k, comme étant l'ensemble des individus possédant la modalité τ_k pour la variable X.

Nous introduisons les notations suivantes :

- $Y(n \times p)$: matrice des données quantitatives.
- $y_i(1 \times p)$: $i$ème ligne de X (décrit le $i$ème individu).
- $y^j(n \times 1)$: $j$ème colonne de X (décrit la variable X^j).
- $\bar{Y}$: matrice des données quantitatives centrées.
- $g(1 \times p)$: barycentre global, défini par
$$g = \frac{1}{n} \sum_{i=1}^{n} y_i$$
- $g_k(1 \times p)$: barycentre de la $k$eme classe, défini par
$$g_k = \frac{1}{n_k} \sum_{i\in\Omega_k} y_i$$
- $G(m \times p)$: matrice des barycentres des classes, définie par
$$G = \begin{pmatrix} g_1' \\ . \\ . \\ . \\ g_m' \end{pmatrix}$$

- $S_r(p \times p)$: matrices de variance-covariance résiduelle (intra-classes), définie par

$$S_r = \frac{1}{n} \sum_{k=1}^{m} \sum_{i \in \Omega_k} (y_i - g_k)(y_i - g_k)'$$

- $S_e(p \times p)$: matrice de variance-covariance expliquée (inter-classes), définie par

$$S_e = \sum_{k=1}^{m} \frac{n_k}{n} (g_k - g)(g_k - g)'$$

- $S(p \times p)$: matrice de variance-covariance, définie par $S = S_r + S_e$

L'AFD vise à déterminer une combinaison linéaire $s = u_1 Y^1 + u_2 Y^2 + ... + u_p Y^p$ des vecteurs $Y^1, ..., Y^p$, de telle sorte que s discrimine au mieux les groupes définis par les m modalités de X. On pose $s = Yu$, où s et $u \in \mathbf{R}^p$ sont respectivement appelés *variable discriminante* et *facteur discriminant*.

On considère qu'un axe discrimine convenablement les classes lorsque ces dernières sont distinctement séparées les une des autres le long de cet axe. Pour visualiser une telle séparation, la variance inter-classes doit être très largement supérieure à la variance totale. Autrement dit, la variable discriminante $s = Yu$ doit être construite de telle sorte que le facteur discriminant u maximise le rapport entre la variance inter-classes et la variance totale.
u est donc solution du problème de maximisation suivant :

$$argmax_{u \in \mathbf{R}^p} \frac{u' S_e u}{u' S u} \tag{2}$$

Une solution de (2) est appelée *premier facteur discriminant* et notée u_1. La *première variable discriminante* est alors $s_1 = Yu_1$, et on définit le pouvoir discriminant de u_1 par $\lambda_1 = \frac{u_1' S_e u_1}{u_1' S u_1}$.

Une fois s_1 définie, on peut rechercher la seconde variable discriminante $s_2 = Yu_2$, non corrélée à s_1, et telle que $\frac{u_2' S_e u_2}{u_2' S u_2}$ soit maximal.
Les autres variables discriminantes s'obtiennent sur le même modèle. On peut montrer que les facteurs discriminants u_j sont les vecteurs propres de $S^{-1} S_e$, et que les valeurs propres associées $\lambda_1, ...\lambda_{m-1}$ définissent les pouvoirs discriminants de $u_1, ..., u_{m-1}$ respectivement. Le rang de la matrice $S^{-1} S_e$ étant, d'au plus, $min(p, m-1)$, avec $m - 1 < p$, on pourra construire au maximum $m - 1$ axes discriminants.

Dans R, les AFD ont été réalisées avec la fonction `lda()` du package `MASS`.

3.4.2.2 PLS discriminante

Dans le cas particulier où $n < p$, les variables de la matrice de variance-covariance S sont colinéaires. Dans ce cas, l'AFD n'est plus applicable. On réalise alors une PLS discriminante, notée PLS-DA (Partial Least Square - Discriminant Analysis), basée sur des régressions PLS.
Cette section reprend les concepts et notations introduits dans *Multivariate projection methodologies for the exploration of large biological data sets* [Déjean, 2019] et *Composantes principales et régressions PLS parcimonieuses* [Besse, 2017].

Régressions PLS
Ce paragraphe introduit deux régressions PLS, dénotées régressions PLS1 et PLS2.

La régression PLS1 s'applique dans le cas où l'on dispose d'une variable quantitative à expliquer Y et de p variables quantitatives explicatives $X^1, ..., X^p$. La régression se décompose en une série d'étapes, détaillée ci-dessous.

- Construction d'une première composante t_1 :

$$t_1 = w_{11} x_1 + ... + w_{1p} x_p$$

où $x_1, ..., x_p$ désignent les observations des variables $X^1, ..., X^p$.

- Régression simple de x sur la composante t_1 :

$$y = c_1 t_1 + y_1$$

  i.e.

$$y = c_1 w_{11} x_1 + ... c_1 w_{1p} x_p + y_1$$

- Si nécessaire, on ajoute une deuxième composante t_2, non corrélée à t_1 :

$$t_2 = w_{21} x_{11} + ... + w_{2p} x_{1p}$$

  avec $\forall j \in \{1, ..., p\}$, x_{1j} le résidu de la régression de x_j sur t_1.

- Nouvelle régression :

$$y = c_1 t_1 + c_2 t_2 + y_2$$

Le processus peut se poursuivre avec d'autres constructions de composantes.

La régression PLS2 est une généralisation de la régression PLS1 au cas d'une variable cible Y multidimensionnelle. Elle s'applique donc lorsque l'on cherche à expliquer un ensemble de q variables $Y^1, ..., Y^q$ par un ensemble de p variables explicatives $X^1, ..., X^p$. Elle consiste à « rechercher des combinaisons linéaires de chaque paquet de variables ayant la plus grande covariance possible » [Déjean, 2019].

Principe de la PLS discriminante
Le principe de la PLS-DA est le suivant : la variable à expliquer Y multidimensionnelle est remplacée par q variables indicatrices des modalités de Y. Une régression PLS2 est ensuite exécutée en considérant ces variables indicatrices comme quantitatives.

Dans R, les PLS discriminantes ont été réalisées avec la fonction `plsda()` du package `mixOmics`.

4 Traitement des données

Afin de rendre les données exploitables et de pouvoir effectuer des analyses statistiques, il a fallu restructurer le tableur d'origine. J'ai donc construit une matrice, sous un unique onglet, et possédant les colonnes suivantes :

- `cellLine` : variable qualitative codant pour le type de lignée cellulaire. Ses modalités sont notées *CT*, *CTA*, *CTP*, *CTR*, *CTRPA*, *CTRPAt*.
- `treatment` : variable qualitative représentant le polluant auquel ont été exposées les différentes lignées cellulaires. Ses modalités sont notées *DMSO*, *BAP*, *Pyr*, *TCDD*, *Mix*.
- `expe` : variable qualitative codant pour les conditions expérimentales. Il prend la forme d'un numéro. Les observations possédant une même valeur de `expe` ont été réalisées lors d'une même expérience. Toutes les expériences ont été effectuées dans des conditions expérimentales voisines.
- `plate` : variable qualitative identifiant la plaque sur laquelle a été faite la mesure. Elle prend des valeurs entières allant de 1 à 71.
- `replicate` : variable qualitative identifiant le puits d'une plaque. Ses modalités correspondent à ses coordonnées.
- `time` : variable qualitative représentant le temps d'exposition au polluant. La modalité *0* correspond aux conditions basales avant traitement, tandis que les modalités *48* et *120* correspondent à une durée d'exposition de 48 heures et 120 heures, respectivement.
- `variable` : variable qualitative codant pour la caractéristique mesurée. Les variables codant pour l'expression d'un gène sont désignées par le nom du gène en question, soit *6P6D*, *ACO1*, *AhR*, *AhRR*, *ATPSIF1*, *CAT*, *CYP1A1*, *CYP1A2*, *CYP3A4*, *ENIO1*, *FH*, *G6PD*, *gene_inc*, *HK2*, *HMGCR*, *HMOX*, *IDH1*, *LDHA*, *LDHB*, *LPCAT*, *MCT4*, *MFN2*, *Mitoferrin1*, *Mitoferrin2*, *ND1*, *NHE1*, *NQO1*, *NRF2*, *PKM1*, *PKM2*, *PRDX1*, *RDK1*, *SCD1*, *SDHA*, *SDHC*, *SIRT3*, *TFAM*, *TIGAR*, *TSPO*, *UQCC3*. Parmi les variables observées se trouvent également les variables respiratoires *basalResp*, *maxResp*, *protons*, *glyco*.
- `value` : variable quantitative codant pour la valeur d'une mesure.

Une fois exploitables, les données ont été importées dans R. Un extrait de ces données est représenté en figure 8.

cellLine <fctr>	treatment <fctr>	expe <int>	replicate <fctr>	time <int>	variable <fctr>	value <dbl>	plate <int>
CT	DMSO	31	A3	48	basalResp	7.910	12
CT	DMSO	48	A2	48	basalResp	9.410	13
CT	DMSO	48	A3	48	basalResp	5.900	13
CT	DMSO	48	A4	48	basalResp	6.540	13
CT	DMSO	48	A5	48	basalResp	8.210	13
CT	DMSO	32	A3	48	basalResp	6.120	14
CT	DMSO	32	A4	48	basalResp	5.460	14
CT	DMSO	32	A5	48	basalResp	5.370	14
CT	BAP	30	B2	48	basalResp	7.110	11
CT	BAP	30	C2	48	basalResp	6.940	11

FIGURE 8: Matrice des données

Nommons *var1*, *var2*, ... , *var44* les modalités de la variable `variable`. Une alternative à la matrice actuelle consiste à stocker nos données dans une matrice de colonnes *cellLine*, *treatment*, *expe*, *replicate*, *time*, *plate*, *var1*, *var2*, ..., *var44*. Toutes les mesures n'ayant pas été effectuées simultanément, cette matrice contient beaucoup de valeurs manquantes, et la « lisibilité » des données en est déteriorée.
Bien que moins lisible, nous choisissons d'adopter une telle représentation puisqu'elle permet d'anticiper les futures analyses multivariées, lors desquelles ne seront considérées que les variables observées simultanément. Il sera nécessaire d'avoir, sur une même ligne, les mesures de toutes les variables observées simultanément pour un même sextuplet *cellLine* × *treatment* × *expe* × *replicate* × *time* × *plate*. Tous les sextuplets *cellLine* × *treatment* × *expe* × *replicate* × *time* × *plate* ne possédant pas de mesures simultanées pour les variables d'intérêt seront supprimés des données. De cette manière, nous construirons, pour chaque modalité de `time`, la matrice du plus grand jeu de données respiratoires complet (les variables respiratoires étant observées simultanément), et, sous conditions basales uniquement, la matrice du plus grand jeu de données génomiques complet (les variables génomiques étant observées simultanément et indépendamment des variables respiratoires). Un extrait du plus grand jeu de données respiratoires complet pour 48 heures d'exposition est représenté en figure 9, à titre d'exemple.

	cellLine <fctr>	treatment <fctr>	expe <fctr>	replicate <fctr>	plate <fctr>	basalResp <dbl>	glyco <dbl>	maxResp <dbl>	protons <dbl>
40	CT	BAP	32	A6	14	5.02	1.79	10.29	1.13
42	CT	BAP	32	B1	14	5.94	1.84	13.07	1.00
44	CT	BAP	32	B2	14	5.27	1.69	12.46	0.91
45	CT	BAP	32	B3	14	6.76	2.07	16.64	1.02
47	CT	BAP	33	A6	15	4.05	1.05	10.78	0.87
49	CT	BAP	33	B1	15	3.27	0.89	7.82	0.55

FIGURE 9: Matrice du plus grand jeu de données respiratoires complet pour 48 heures d'exposition

Le pré-traitement suivant a donc été réalisé sur la matrice des données avant de débuter l'analyse statistique :

- Nous avons utilisé la fonction `spread()` du package `tidyverse` pour disperser chaque modalité de `variable` dans une colonne.
- A partir de cette nouvelle matrice ont été construites trois sous-matrices de données : une pour chaque durée d'exposition.
  Les premières lignes de ces matrices sont représentées en figures 10, 11 et 12. Notons la suppression de la colonne `treatment` dans les données en conditions basales. Par définition, ces données sont obtenues en l'absence de polluant.

	cellLine <fctr>	expe <fctr>	replicate <fctr>	plate <fctr>	6P6D <dbl>	ACO1 <dbl>	AhR <dbl>	AhRR <dbl>	ATP5IF1 <dbl>
1	CT	2	A1	10	15.50	15.53	15.54	17.85	17.76
2	CT	2	A2	10	16.44	16.56	18.11	17.92	19.83
3	CT	2	A3	10	15.45	16.23	17.20	17.84	19.17
4	CT	2	A4	10	17.49	16.89	18.35	19.38	20.04
5	CT	2	A5	10	16.66	16.03	16.76	18.08	19.04
6	CT	2	A6	10	16.83	15.17	17.39	18.72	19.51

6 rows | 1-10 of 48 columns

FIGURE 10: Matrice des données sous conditions basales `cell_basal`

	cellLine <fctr>	treatment <fctr>	expe <fctr>	replicate <fctr>	plate <fctr>	basalResp <dbl>	glyco <dbl>	maxResp <dbl>	protons <dbl>
27	CT	BAP	1	A3	50	*NA*	*NA*	*NA*	*NA*
28	CT	BAP	1	A3	51	*NA*	*NA*	*NA*	*NA*
29	CT	BAP	1	B3	50	*NA*	*NA*	*NA*	*NA*
30	CT	BAP	1	C3	50	*NA*	*NA*	*NA*	*NA*
31	CT	BAP	1	D3	50	*NA*	*NA*	*NA*	*NA*
32	CT	BAP	30	B2	11	7.11	*NA*	8.47	1.07

6 rows

FIGURE 11: Matrice des données pour 48 heures d'exposition `cell_48`

	cellLine <fctr>	treatment <fctr>	expe <fctr>	replicate <fctr>	plate <fctr>	basalResp <dbl>	glyco <dbl>	maxResp <dbl>	protons <dbl>
33	CT	BAP	30	B2	54	4.76	*NA*	11.82	0.43
35	CT	BAP	30	C2	54	5.11	*NA*	12.77	0.55
36	CT	BAP	31	A4	55	4.86	*NA*	17.14	1.32
38	CT	BAP	31	A5	55	5.47	*NA*	18.10	1.50
39	CT	BAP	32	16	56	*NA*	*NA*	17.43	0.69
41	CT	BAP	32	A6	56	3.99	1.21	*NA*	*NA*

6 rows

FIGURE 12: Matrice des données pour 120 heures d'exposition `cell_120`

5 Caractérisation des lignées cellulaires

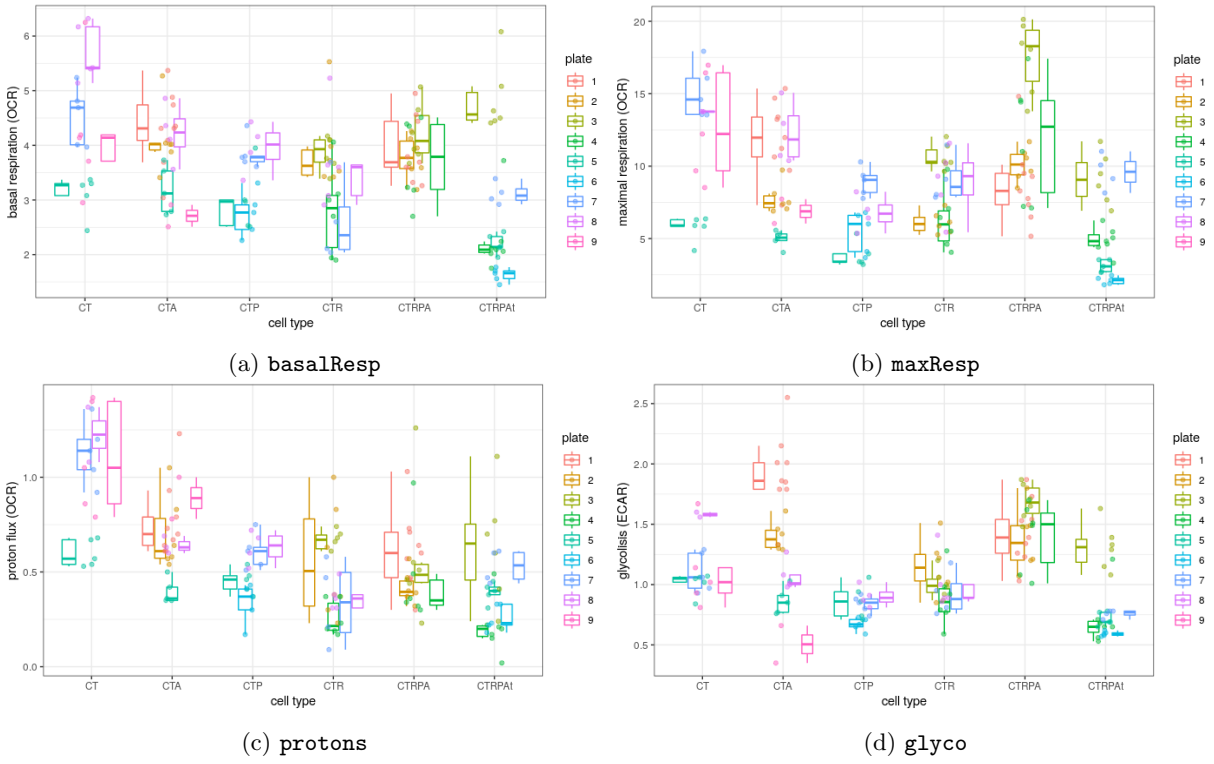
5.1 Conditions basales

L'objectif d'une telle analyse était de caractériser les symétries et variabilités des distributions.

5.1.1 Analyse exploratoire

5.1.1.1 Variables respiratoires

Variabilité des distributions en fonction de la lignée cellulaire et de la plaque
Nous avons représenté les répartitions des valeurs de chaque variable respiratoire en fonction de la lignée cellulaire et de la plaque, afin de visualiser l'impact de `plate` sur les mesures.



(a) `basalResp` (b) `maxResp`

(c) `protons` (d) `glyco`

FIGURE 13: Diagrammes-boîtes parallèles représentant les distributions de `basalResp` (a), `maxResp` (b), `protons` (c), `glyco` (d) par lignée cellulaire et plaque

Ces diagrammes-boîtes parallèles (figure 13) témoignent de la grande variabilité des données en fonction de la plaque considérée. De plus, ces variations inter-plaques ne suivent pas de schéma récurrent d'une variable à l'autre pour une même lignée, ou d'une lignée à l'autre pour une même variable. Par exemple, les distributions des variables `protons` et `glyco` obtenues pour la lignée CTA montrent que sur la plaque 9 ont été mesurées les plus hautes valeurs de `protons`, et les plus faibles valeurs de `glyco`. Nous observons donc un effet plaque, qui ne semble pas être corrigible facilement.

La figure 13 met également en avant, pour chacune des variables respiratoires, la présence de différences entre distributions pour les différentes lignées cellulaires. Pour cette raison, il sera judicieux de poursuivre l'étude par des comparaisons de distributions inter-lignées.

Symétrie des distributions
Avant de mettre en œuvre de telles comparaisons, nous avons étudié les symétries des distributions. En particulier, nous avons réalisé des tests de normalité de Shapiro sur les lignées cellulaires. Les p-valeurs résultantes ont été consignées dans la table 1.

Table 1 : Table des p-valeurs obtenues des tests de normalité de Shapiro en conditions basales

Lignée cellulaire	basalResp	maxResp	protons	glyco
CT	0.428657869	0.09911322	0.17958823	0.10305532
CTA	0.764992975	0.18901318	0.21049317	0.79236810
CTP	0.406564534	0.30346066	0.97738266	0.92901695
CTR	0.219913227	0.16527733	0.15679395	0.39045080
CTRPA	0.854082556	0.07077479	0.00131704	0.22222509
CTRPAt	0.002230869	0.02160376	0.04282964	0.01331305

Pour chaque variable mesurée, il y a au moins une lignée cellulaire pour laquelle l'hypothèse de normalité est rejetée avec un niveau de confiance de 95%. Ainsi, nous ne pouvons considérer aucune des distributions des variables respiratoires comme gaussienne. Les tests de comparaison à réaliser seront tous non-paramétriques.

5.1.1.2 Variables génomiques

Variabilité des distributions en fonction de la lignée cellulaire
Nous avons représenté, sous la forme de diagrammes-boîtes, les distributions des valeurs de chaque variable génomique en fonction de la lignée cellulaire, afin de visualiser l'impact de la variable `cellLine` sur les mesures (figure 14).

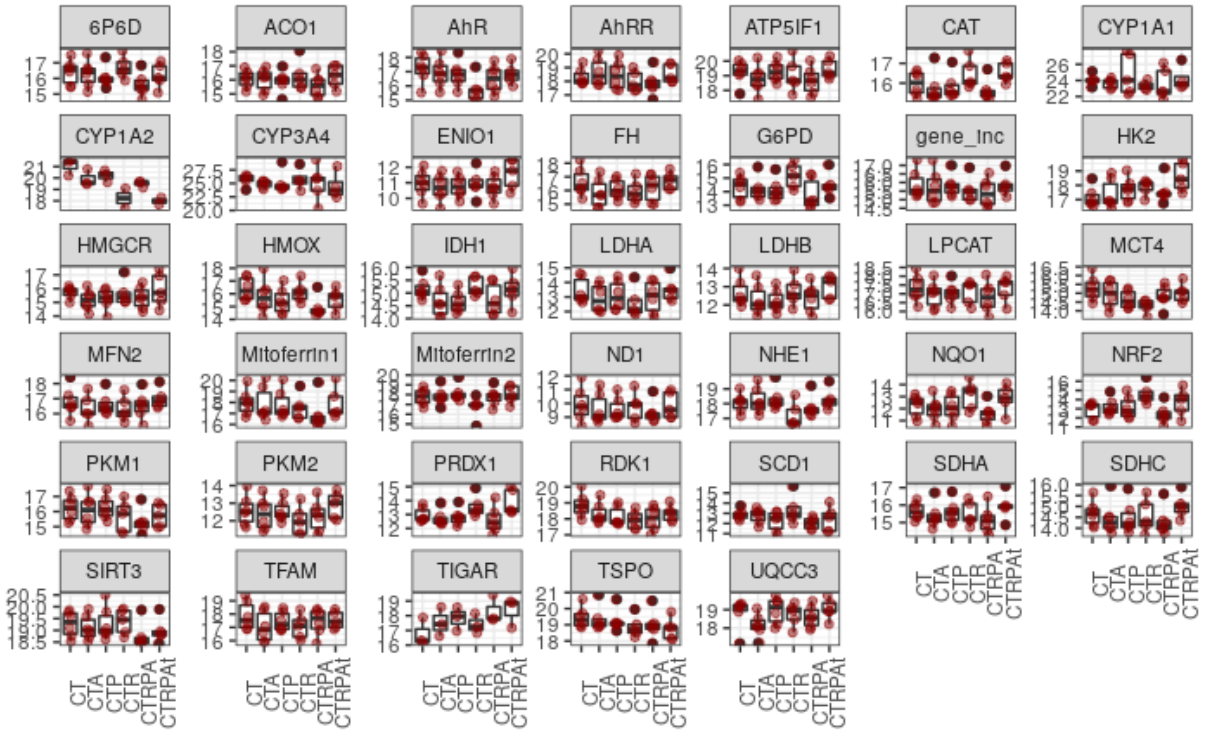


FIGURE 14: Diagrammes-boîtes parallèles représentant les distributions des variables génomiques par lignée cellulaire

La figure 14 témoigne de différences entre distributions pour les différentes lignées cellulaires. Il sera donc judicieux, ici aussi, de poursuivre l'étude par des comparaisons de distributions inter-lignées.

Symétrie des distributions
Avant de mettre en œuvre de telles comparaisons, nous avons étudié les symétries des distributions. En particulier, nous avons réalisé des tests de normalité de Shapiro sur les lignées cellulaires. Les p-valeurs résultantes ont été consignées en Annexe B.

Pour chaque variable mesurée, il y a au moins une lignée cellulaire pour laquelle l'hypothèse de normalité est rejetée avec un niveau de confiance de 95%.
De plus, la figure 14 présente, pour chaque variable génomique, au moins un diagramme-boîte pour lequel la distance entre la médiane et le premier quartile est très différente de celle entre la médiane et le troisième quartile. Pour chaque variable génomique, il existe donc au moins une distribution asymétrique (et par conséquent non-gaussienne) parmi les distributions des différentes lignées.
Ainsi, nous ne pouvons considérer aucune des distributions des variables génomiques comme gaussienne. Les tests de comparaison à réaliser seront tous non-paramétriques.

5.1.2 Tests de comparaison

L'analyse exploratoire ayant identifié la présence d'une variabilité des distributions inter-lignées, nous souhaitons à présent réaliser des tests de comparaison entre ces lignées.

5.1.2.1 Analyse non-paramétrique

5.1.2.1.a Étude des variables respiratoires

Les distributions des groupes de cellules de mêmes lignées n'étant pas gaussiennes, nous avons réalisé un test de Kruskal-Wallis, dont les hypothèses sont les suivantes : H_0 : « Toutes les distributions sont identiques », contre H_1 : « Les distributions présentent des différences ». Tous ces tests ont mené à des p-valeurs inférieures à 5 % (Annexe B), nous permettant de conclure avec une marge d'erreur de 5 % que les distributions des variables respiratoires présentent des différences significatives selon la lignée cellulaire considérée.
Dans ce contexte, il est justifié d'effectuer des comparaisons deux à deux entre lignées cellulaires. Étant dans le cas non-paramétrique, nous avons réalisé des tests de Nemenyi, afin de tester les hypothèses suivantes : H_0 : « Les deux distributions sont identiques », contre H_1 : « Les deux distributions présentent des différences ». Les résultats obtenus pour un niveau de confiance de 95 % sont joints en Annexe B et résumés dans la table 2.

Table 2 : Table résumant les différences de distributions des variables respiratoires entre les lignées cellulaires en conditions basales

Variable	basalResp	maxResp	protons	glyco
Lignées cellulaires dont les distributions sont significativement différentes (risque : 5 %)	CT-CTRPAt CTA-CTRPAt CTRPA-CTRPAt	CT-CTP CT-CTRPAt CTA-CTRPAt CTP-CTRPA CTRPA-CTRPAt	CT-CTP CT-CTR CT-CTRPA CT-CTRPAt CTA-CTR CTA-CTRPAt	CTA-CTP CTA-CTRPAt CTP-CTRPA CTR-CTRPA CTRPA-CTRPAt

D'après la table 2, il semble que la lignée CTRPAt se différencie des autres sur un plus grand nombre de caractéristiques.

5.1.2.1.b Étude des variables génomiques

Les comparaisons sur l'ensemble des groupes cellulaires ont été effectuées avec des tests de Kruskal-Wallis, et les comparaisons deux-à-deux avec des tests de Nemenyi. Toutes les p-valeurs renvoyées par ces tests, hormis pour le gène `CYP1A2`, étaient supérieures à 5 % (Annexe B). Ainsi, nous n'étions pas en mesure d'affirmer avec un niveau de confiance de 95 % que les distributions des variables génomiques (excepté `CYP1A2`) présentaient elles aussi des différences significatives selon la lignée cellulaire considérée. Le test de Nemenyi appliqué à la variable `CYP1A2` nous a mené à des p-valeurs toutes supérieures à 5 %. Cependant, en autorisant une marge d'erreur de 10 %, nous pouvons affirmer la présence de différences significatives entre les distributions des lignées CTR et CT (p-valeur de 0,086), et des lignées CTRPAt et CT (p-valeur de 0,086). Ces résultats sont joints en Annexe B et résumés dans la table 3.

Table 3 : Table résumant les différences de distributions des variables génomiques entre les lignées cellulaires en conditions basales.

Variable	CYP1A2
Lignées cellulaires dont les distributions sont significativement différentes (risque : 10 %)	CT-CTR CT-CTRPAt

5.1.2.1.c Conclusion des tests non-paramétriques

Quelques différences significatives entre lignées ont été identifiées pour les variables de respiration mais les résultats sont à prendre avec beaucoup de précaution : il y a une variabilité importante entre plaques qui rend les résultats peu fiables.

5.1.2.2 Analyse paramétrique

Nous aimerions prendre en compte, dans les analyses, l'effet plaque observé pour les données respiratoires. Nous utilisons pour cela des modèles mixtes, qui permettent de modéliser l'aléa lié à la plaque.
Cet aléa n'étant pas présent dans les données génomiques (mesurées sur une unique plaque), nous n'utilisons pas de modèles mixtes. Nous pourrions modéliser ces données par des modèles linéaires, qui tiennent uniquement compte des effets fixes, puis réaliser des tests de comparaison sur ces modèles. Cependant, cela revient à reproduire exactement les tests précédents, mais en réalisant des ANOVA, et non plus des tests de Kruskal-Wallis, sur des données que l'on sait non-gaussiennes. Les résultats seraient donc moins pertinents : nous choisissons de ne pas faire d'ajustement par modèle linéaire.

Étude préliminaire des données

- Les effets de la variable cellLine sont fixes, puisque l'ensemble des modalités *CT*, *CTA*, *CTP*, *CTR*, *CTRPA*, *CTRPAt* est fini.
- Les effets de la variable `plate` sont aléatoires : si les mesures sont reproduites un an plus tard, les plaques sur lesquelles ont été faites les mesures sous conditions basales peuvent différer des plaques actuelles (numéros de plaque différents).
- Les effets de la variable `expe` étant expliqués par ceux de la variable `plate`, nous ne considérons que les effets aléatoires engendrés par `plate`.

Ainsi, l'étude des variables respiratoires nécessite de tenir compte d'effets fixes et aléatoires. L'ajustement par modèle mixte est donc approprié.

Modèles mixtes

Comme en témoignent les graphes de comparaison des quantiles en Annexe B, la distribution approchant le mieux nos mesures de `maxResp`, `protons`, et `glyco`, est la distribution Γ. En revanche, il vaut mieux utiliser la distribution gaussienne pour `basalResp`. En modélisant nos données respiratoires par des modèles mixtes, nous avons donné de l'importance aux effets aléatoires engendrés par la variable `plate`. De ce fait, les résultats des tests de comparaison (en Annexe B) sont à considérer avec plus grand intérêt que ceux obtenus à l'issue de nos premières analyses non-paramétriques, qui ne tenaient pas compte de ces effets aléatoires. Comme en témoigne la table 4, les comparaisons deux-à-deux ont permis d'identifier un plus grand nombre de lignées aux distributions différentes.

Table 4 : Table résumant les différences de distributions des variables respiratoires entre les lignées cellulaires en conditions basales, avant et après approximations des données par des modèles mixtes (en rouge : information spécifique à l'analyse paramétrique - en bleu : information spécifique à l'analyse non-paramétrique - en noir : information commune)

Variable	basalResp	maxResp	protons	glyco
Lignées cellulaires dont les distributions sont significativement différentes (risque : 5 %)	CT-CTA CT-CTP CT-CTR CT-CTRPA CT-CTRPAt CTA-CTR CTA-CTRPAt CTP-CTR CTP-CTRPAt CTRPA-CTRPAt	CT-CTP CT-CTR CT-CTRPAt CTA-CTP CTA-CTR CTA-CTRPAt CTP-CTRPA CTR-CTRPA CTRPA-CTRPAt	CT-CTA CT-CTP CT-CTR CT-CTRPA CT-CTRPAt CTA-CTR CTA-CTRPAt	CT-CTP CT-CTR CT-CTRPAt CTA-CTP CTA-CTR CTA-CTRPAt CTP-CTRPA CTR-CTRPA CTRPA-CTRPAt

D'après la table 4, la lignée contrôle CT et la lignée CTRPAt des cellules les plus semblables aux cellules cancéreuses semblent se différencier des autres lignées sur un plus grand nombre de caractéristiques.

5.1.3 Analyses multivariées

5.1.3.1 Analyse en Composantes principales

Dans l'optique de caractériser les différentes lignées cellulaires, nous souhaitons synthétiser de manière graphique les informations de notre jeu de données multidimensionnel. Nous réalisons pour cela une Analyse en Composantes Principales, méthode de réduction de dimension préservant au mieux l'inertie des données initiales. La représentation bidimensionnelle fournie par l'ACP facilite alors l'interprétation des données.

5.1.3.1.a Variables respiratoires

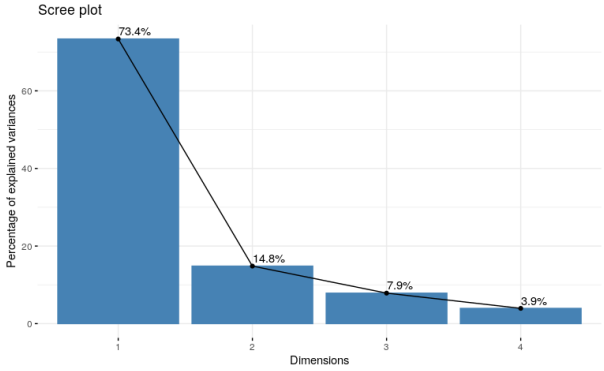


FIGURE 15: Diagramme des éboulis représentant la part de variance expliquée contenue dans chaque composante principale

Le diagramme des éboulis (figure 15) nous montre que les deux premières composantes renferment la majorité de la variance expliquée (88,2%). Ainsi, la représentation des données sur l'espace composé de ces deux composantes principales restera représentatif de la réalité.

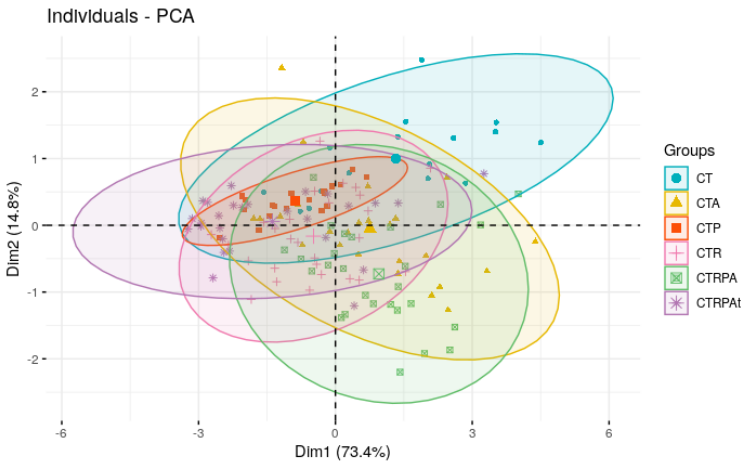


FIGURE 16: Graphe des individus groupés par lignée cellulaire

La figure 16 ne nous permet pas d'établir de distinction claire entre lignées cellulaires. Ce résultat était prévisible : nous avions observé une très grande variabilité inter-plaques des distributions pour chaque type de cellules, or l'ACP considère les données récoltées sur l'ensemble des plaques. De ce fait, l'étude d'une variable au sein d'une même lignée cellulaire peut mener à des données très éloignées. Cela se traduit, dans l'espace des composantes principales, par une ellipse étendue et potentiellement intersectée

avec les ellipses des autres lignées.

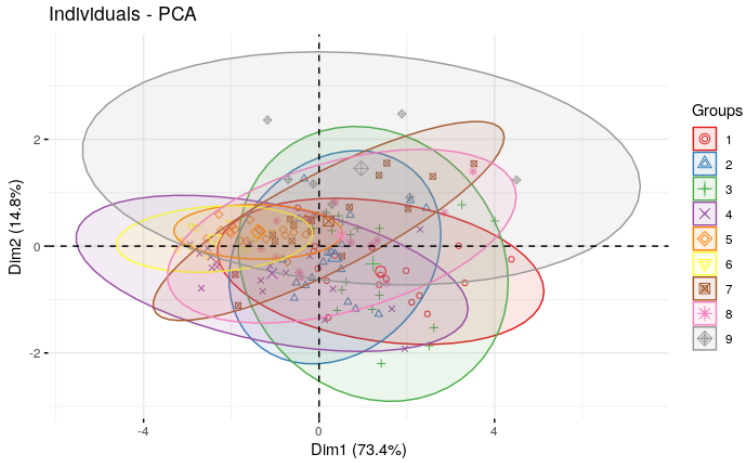


FIGURE 17: Graphe des individus groupés par plaque

La figure 17 montre que les groupes d'individus ne sont pas distinctement séparés par plaques. Les conclusions de l'analyse exploratoire sont donc consolidées : l'aléa induit par les plaques ne suit aucun « schéma ». Par conséquent, il est difficile de corriger cet aléa.

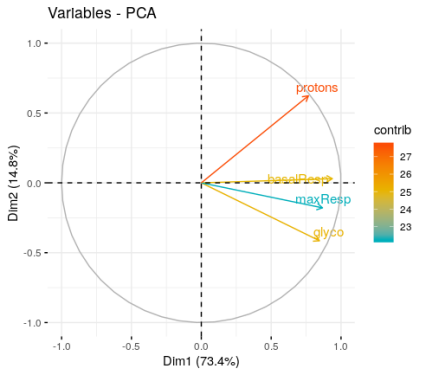


FIGURE 18: Cercle des corrélations

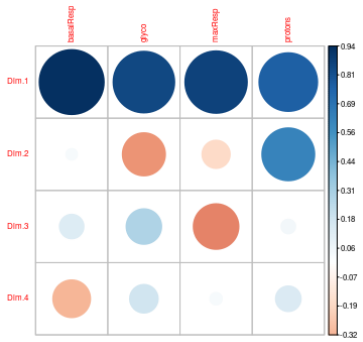


FIGURE 19: Diagramme des corrélations

Bien que cette ACP ne nous aide pas à caractériser les lignées cellulaires, elle met tout de même en avant la présence d'un effet taille entre variables. En effet, le cercle de corrélation (figure 18) montre que toutes les variables sont corrélées positivement entre elles. La première composante principale définit alors un « facteur taille ». De plus, les variables étant proches du cercle, elles possèdent une bonne qualité de représentation. Comme le confirme le diagramme des corrélations (figure 19), chaque variable est positivement et fortement corrélée avec la première composante principale.

5.1.3.1.b Variables génomiques

Les résultats des ACP obtenus pour les variables génomiques sont similaires que ceux obtenus pour les variables respiratoires. Les diagrammes des éboulis et des corrélations sont disponibles en Annexe B.
Pour les mêmes raisons que dans le cas des variables respiratoires, ce graphe des individus ne permet pas d'établir de distinction claire entre lignées cellulaires.
Le cercle des corrélations (figure 21) suggère que toutes les variables génomiques sont corrélées positivement avec la première composante principale (produit scalaire grand et positif). Toutes les variables étant également corrélées positivement entre elles, il y a un « effet taille » entre gènes. Cela signifie que certaines lignées cellulaires prennent globalement de plus grandes valeurs que d'autres.

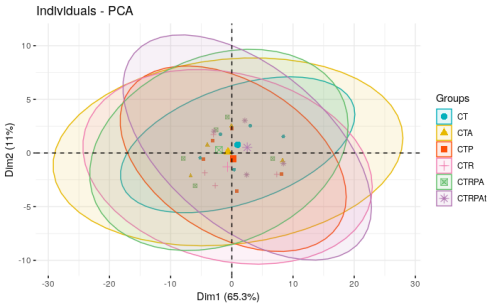


FIGURE 20: Graphe des individus groupés par lignée cellulaire

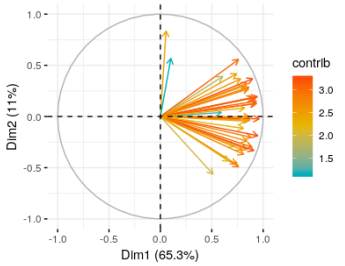


FIGURE 21: Cercle des corrélations

5.1.3.1.c Conclusion des ACP

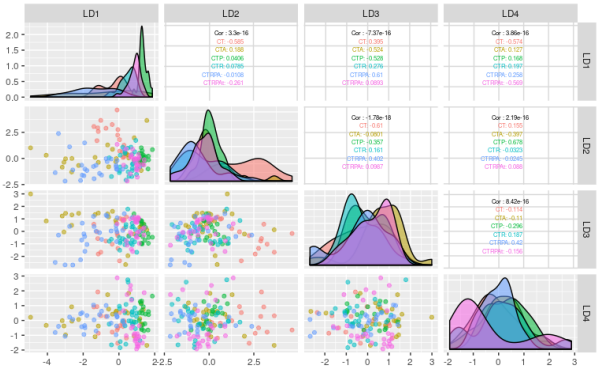
Les ACP ne permettent pas de caractériser les lignées cellulaires, à cause d'une trop grande variabilité inter-plaques des mesures. En revanche, elles soulignent la présence d'effets taille entre les variables (certaines lignées cellulaires prennent globalement de plus grandes valeurs que d'autres).

5.1.3.2 Analyses Discriminantes

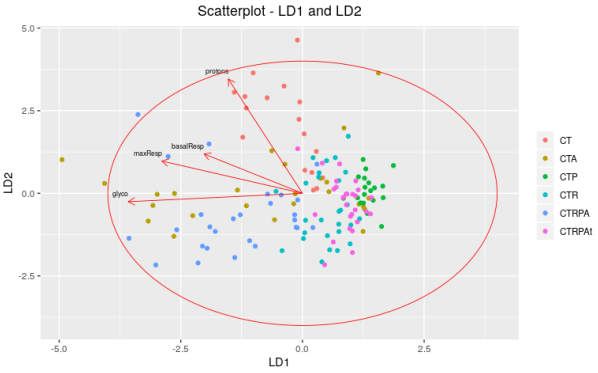
Toujours dans l'optique de caractériser les différentes lignées cellulaires, nous souhaitons construire des espaces de représentation des données permettant de discriminer au mieux ces lignées.

5.1.3.2.a Variables respiratoires

Le nombre d'individus étant supérieur au nombre de variables respiratoires observées, nous réalisons une AFD.



(a) Matrice de nuages de points



(b) Nuage de points et variables

FIGURE 22: (a) : Matrice de nuages de points résultant de l'AFD sur les données respiratoires en conditions basales, (b) : Nuage de points, vecteurs variables, et cercle de corrélation représentés dans l'espace des deux premiers facteurs discriminants, en conditions basales

Les données ont été représentées dans différents espaces bidimensionnels de facteurs discriminants, sous la forme de nuages de points. On considère qu'un facteur discriminant sépare convenablement les groupes de cellules si ces derniers sont convenablement séparés les uns des autres le long de l'axe représentant ce facteur discriminant.
La figure 22a montre que les lignées CTA et CTRPA sont séparées du reste des lignées le long de l'axe correspondant au premier facteur discriminant. Le second facteur discriminant sépare la lignée CT de toutes les autres.
En faisant apparaître les vecteurs variables et cercles de corrélation, à la manière présentée sur la figure 22b, nous avons pu attribuer quelques caractéristiques aux groupes précédemment discriminés. À titre d'exemple, la figure 22b suggère que les cellules CT possèdent de plus grandes valeurs de `protons` que

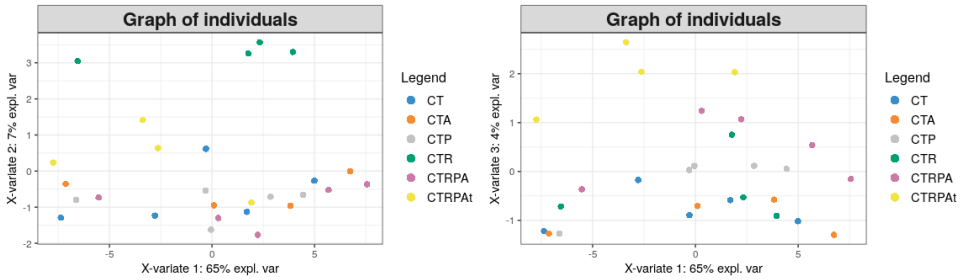
les autres lignées, et que les cellules CTA possèdent de plus grandes valeurs de `basalResp`, `maxResp` et `glyco` que les autres lignées. Les variables étant proches du cercle de corrélation, elles témoignent d'une qualité de représentation suffisamment bonne pour que les interprétations précédentes soient fiables. En répétant des analyses similaires sur l'ensemble des graphes discriminant les lignées, nous parvenons aux caractérisations suivantes :

- Les cellules CT possèdent de plus hautes valeurs de `protons` que le reste des cellules.
- Les cellules CTA possèdent de plus hautes valeurs de `glyco` et de `maxResp` que le reste des cellules.
- Les cellules CTRPA possèdent de plus hautes valeurs de `glyco` que le reste des cellules.

Les matrices de nuages de points et les représentations des individus dans les espaces de facteurs discriminants séparant convenablement les lignées sont reportés en Annexe B.

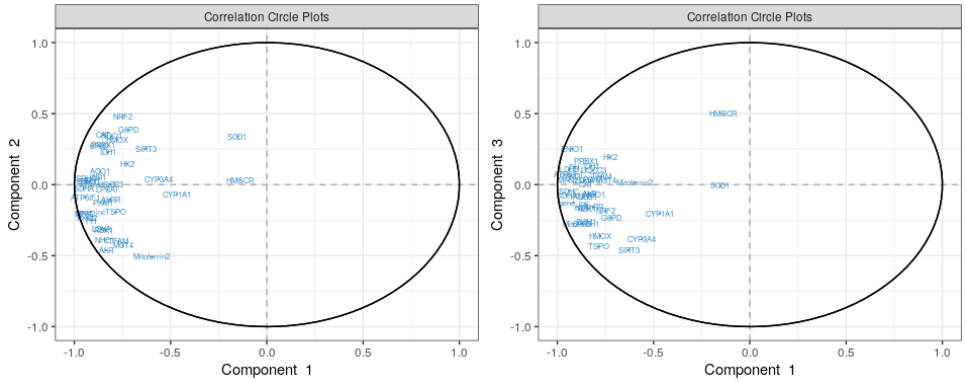
5.1.3.2.b Variables génomiques

Le nombre d'individus étant inférieur au nombre de variables génomiques observées (27 individus pour 38 variables), nous réalisons une PLS-DA.



(a) Facteurs discriminants 1 et 2 (b) Facteurs discriminants 1 et 3

FIGURE 23: Représentation des individus dans l'espace des composantes 1 et 2 (a), et dans l'espace des composantes 1 et 3 (b) après PLS-DA sur les données génomiques en conditions basales



(a) Facteurs discriminants 1 et 2 (b) Facteurs discriminants 1 et 3

FIGURE 24: Variables et cercles de corrélation dans l'espace des composantes 1 et 2 (a), et dans l'espace des composantes 1 et 3 (b) après PLS-DA sur les données génomiques en conditions basales

Les analyses simultanées des figures 23a et 24a d'une part, et des figures 23b et 24b d'autre part, mènent aux caractérisations suivantes :

- Les cellules CTR possèdent de plus hautes valeurs de `CAT`, `G6PD`, `NQO1`, `NRF2` et `SCD1` que le reste des cellules.
- Les cellules CTRPAt possèdent de plus hautes valeurs de `HMGCR` que le reste des cellules.

5.1.3.2.c Conclusion des AFD

Nous sommes parvenus à caractériser certaines lignées cellulaires. Il reste néanmoins difficile, dans la grande majorité des cas, d'obtenir de bonnes discriminations de ces lignées. En effet, la variabilité importante des observations inter-plaques ne permet pas le regroupement des mesures, mais favorise au contraire leur dispersion.

5.2 Après 48 et 120 heures d'exposition

5.2.1 Analyse exploratoire

Les distributions ont été représentées sous la formes de diagrammes-boîtes parallèles, disponibles en Annexe A. Après 48 heures comme après 120 heures d'exposition, les distributions de chaque variable respiratoire, à lignée cellulaire `cellLine` et polluant `treatment` fixés, diffèrent selon la plaque considérée. Cependant, à type cellulaire fixé, les variations inter-plaques semblent suivre un même schéma d'un traitement à l'autre et d'une variable à l'autre.

On note également, à `treatment` et `plate` fixés, la présence de différences de distributions entre lignées. De plus, les plaques utilisées pour les mesures à 48 et à 120 heures comprenant, chacune, des cellules provenant d'une unique lignée, aucun des groupes cellulaires n'a de plaque commune. L'effet lignée est donc compris dans l'effet plaque.

5.2.2 Tests de comparaison

5.2.2.1 Analyses non-paramétriques

Des analyses identiques à celles réalisées en conditions basales mènent aux résultats résumés dans les tables 5 et 6. Les tests ont été réalisés à durée d'exposition fixe et indépendamment pour chaque polluant.

Table 5 : Table résumant les polluants alimentaires à l'origine de différences de distributions entre lignées cellulaires après 48 heures d'exposition (seuil 5 %).

	basalResp	maxResp	protons	glyco
CT-CTA	BAP	DMSO, BAP, TCDD, Mix	-	BAP
CT-CTP	-	-	-	-
CT-CTR	-	-	-	-
CT-CTRPA	-	DMSO, BAP, TCDD, Mix	-	-
CT-CTRPAt	-	-	-	-
CTA-CTP	-	DMSO	-	BAP, TCDD
CTA-CTR	Tous	DMSO, Mix	Tous	BAP, Pyr, TCDD, Mix
CTA-CTRPA	-	-	-	-
CTA-CTRPAt	DMSO, BAP	Tous	DMSO, BAP, Pyr, TCDD	DMSO, BAP, TCDD
CTP-CTR	-	-	-	-
CTP-CTRPA	-	DMSO, Mix	-	DMSO, BAP
CTP-CTRPAt	-	-	-	-
CTR-CTRPA	DMSO	DMSO, Mix	DMSO, BAP, Pyr, Mix	DMSO, Mix
CTR-CTRPAt	-	-	-	-
CTRPA-CTRPAt	DMSO	DMSO, BAP, TCDD, Mix	DMSO, BAP, Pyr	DMSO

Table 6 : Table résumant les polluants alimentaires à l'origine de différences de distributions entre lignées cellulaires après 120 heures d'exposition (seuil 5 %).

	basalResp	maxResp	protons	glyco
CT-CTA	Mix	-	TCDD, Mix	TCDD, Mix
CT-CTP	-	DMSO, BAP, Pyr, TCDD	-	-
CT-CTR	-	BAP, Pyr, TCDD	-	-
CT-CTRPA	-	-	TCDD, Mix	DMSO, BAP, TCDD, Mix
CT-CTRPAt	DMSO, BAP, Pyr	DMSO, BAP, Pyr	DMSO	-
CTA-CTP	Pyr, TCDD, Mix	Tous	-	Tous
CTA-CTR	BAP, Pyr, TCDD, Mix	Tous	TCDD	-
CTA-CTRPA	-	-	-	-
CTA-CTRPAt	Tous	Tous	Tous	DMSO, BAP, Pyr, Mix
CTP-CTR	-	-	TCDD	-
CTP-CTRPA	TCDD, Mix	Tous	-	Tous
CTP-CTRPAt	-	-	Tous	-
CTR-CTRPA	Mix	Tous	TCDD	-
CTR-CTRPAt	-	-	DMSO	DMSO
CTRPA-CTRPAt	Tous	Tous	Tous	Tous

Comme en conditions basales, les analyses non-paramétriques sur les mesures effectuées après 120 heures d'exposition (table 6) nous permettent d'identifier les cellules CTRPAt comme étant particulièrement différentes des autres. Elles se distinguent sur un plus grand nombre de variables et pour un plus grand nombre d'expositions.

5.2.2.2 Analyses paramétriques

Étude préliminaire des données

Contrairement aux observations en conditions basales, les observations après 48 et 120 heures d'exposition ont été réalisées pour 5 polluants différents. On pourrait donc être tenté de réaliser les approximations par modèles mixtes de la même manière qu'en conditions basales, mais en considérant les effets supplémentaires de `treatment`. Cependant, l'objectif de nos analyses étant de caractériser les lignées cellulaires en fonction du polluant appliqué, nous réalisons les ajustements par modèles mixtes à valeurs fixes de `treatment`. Les effets de `treatment` ne sont plus à considérer.

- Les effets de la variable `cellLine` sont fixes.
- Les effets de la variable `plate` sont aléatoires.

On modélise les données par des modèles mixtes.

Modèles mixtes

Dans le cas d'expositions de 48 et 120 heures, toutes les données ont été modélisées par des modèles généralisés basés sur la distribution Γ, à l'exception des mesures de `protons` réalisées après 120 heures d'exposition à *Pyr*, qui ont été modélisées par un modèle gaussien. Les résultats des comparaisons deux-à-deux des lignées cellulaires ont été résumés dans les tables 7 et 8. Les tests de comparaisons ont été réalisés à durée d'exposition fixe et indépendamment pour chaque polluant.

Table 7 : Table résumant les polluants alimentaires à l'origine de différences de distributions entre lignées cellulaires après 48 heures d'exposition (seuil 5%), après ajustement des données par des modèles mixtes

	basalResp	maxResp	protons	glyco
CT-CTA	-	-	-	-
CT-CTP	-	-	-	-
CT-CTR	-	-	TCDD, Mix	-
CT-CTRPA	-	-	-	-
CT-CTRPAt	-	-	-	-
CTA-CTP	-	-	-	-
CTA-CTR	-	-	TCDD, Mix	-
CTA-CTRPA	-	-	-	-
CTA-CTRPAt	-	DMSO	-	-
CTP-CTR	-	-	-	TCDD, Mix
CTP-CTRPA	-	-	-	-
CTP-CTRPAt	-	-	-	-
CTR-CTRPA	-	-	-	TCDD, Mix
CTR-CTRPAt	-	-	-	TCDD, Mix
CTRPA-CTRPAt	-	DMSO	-	-

Table 8 : Table résumant les polluants alimentaires à l'origine de différences de distributions entre lignées cellulaires après 120 heures d'exposition (seuil 5%), après ajustement des données par des modèles mixtes

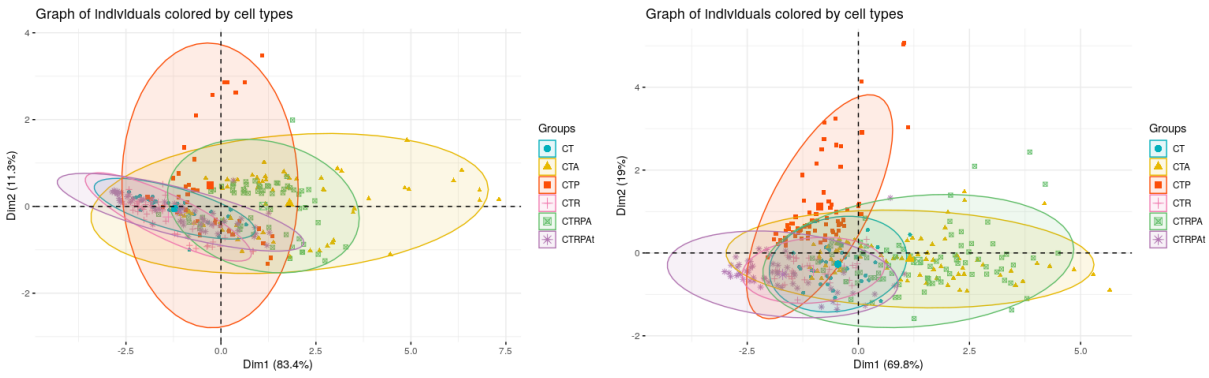
	basalResp	maxResp	protons	glyco
CT-CTA	-	-	-	-
CT-CTP	-	DMSO, BAP, Pyr, TCDD	-	Pyr, TCDD
CT-CTR	-	DMSO, TCDD	-	-
CT-CTRPA	Pyr	-	-	DMSO, TCDD, Mix
CT-CTRPAt	DMSO, BAP	DMSO, BAP, Pyr	DMSO, BAP, TCDD	-
CTA-CTP	Pyr, TCDD, Mix	Tous	-	Tous
CTA-CTR	Pyr, TCDD, Mix	DMSO, BAP, Pyr, TCDD	TCDD	-
CTA-CTRPA	-	-	-	-
CTA-CTRPAt	Tous	Tous	Tous	DMSO, BAP, Pyr
CTP-CTR	-	-	TCDD	BAP, Pyr, TCDD
CTP-CTRPA	-	Tous	-	Tous
CTP-CTRPAt	DMSO, BAP	-	Tous	TCDD
CTR-CTRPA	-	DMSO, Pyr, TCDD	TCDD	TCDD
CTR-CTRPAt	DMSO, BAP	-	DMSO, BAP, TCDD	DMSO
CTRPA-CTRPAt	DMSO, BAP, Pyr, TCDD	Tous	Tous	Tous

Pour des expositions de 48 heures, les modèles mixtes conduisent à un nombre de différences entre lignées significativement réduit par rapport au nombre de différences identifiées par les tests directs des conditions expérimentales. Cette réduction s'explique par l'inclusion de l'effet lignée dans l'effet plaque. En intégrant l'effet plaque dans nos modèles mixtes, nous y avons aussi intégré l'effet lignée. Il est donc logique que les tests paramétriques mènent à un nombre très faible de différences de distributions inter-lignées. Dans ce contexte d'étude, les modèles mixtes sont moins fiables que les modèles non-paramétriques pour caractériser les différences entre lignées.

Ce phénomène est moins visible pour une exposition de 120 heures. Les différences entre distributions inter-lignées étant certainement plus importantes qu'à 48 heures, la confusion des effets plaque et lignée a moins d'influence sur les résultats des tests. On ne perd quasiment pas d'information en passant des modèles non-paramétriques aux modèles mixtes. D'après la table 8, les cellules CTRPAt se différencient des autres, ici encore, sur un grand nombre de variables et pour un grand nombre d'expositions.

5.2.3 Analyses multivariées

5.2.3.1 Analyse en Composantes Principales

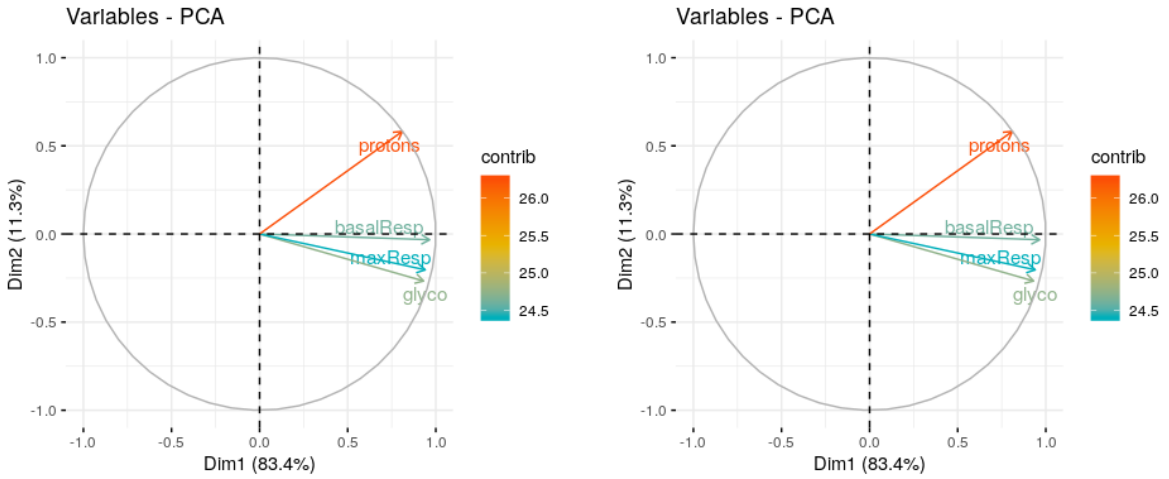


(a) 48 heures

(b) 120 heures

FIGURE 25: Graphe des individus groupés par lignée cellulaire pour 48 heures d'exposition (a) et 120 heures d'exposition (b)

Les conclusions des ACP sont les mêmes qu'en conditions basales : il est impossible d'établir de distinction claire entre lignées cellulaires sur la figure 25 à cause de la trop grande variabilité inter-plaques des distributions pour une même lignée, et les cercles de corrélations (figure 26) révèlent la présence d'effets taille entre variables.



(a) 48 heures

(b) 120 heures

FIGURE 26: Cercle des corrélations pour 48 heures d'exposition (a) et 120 heures d'exposition (b)

5.2.3.2 Analyse Factorielle Discriminante

Pré-traitement des données
Les analyses discriminantes sont réalisées sur les plus grands jeux de données respiratoires complets obtenus pour 48 et 120 heures d'exposition respectivement. Pour chaque durée d'exposition, 5 analyses sont réalisées : une par polluant. Chacune d'entre elles est réalisée en ne considérant que les colonnes `cellLine`, `basalResp`, `maxResp`, `protons` obtenues pour une valeur fixe de `treatment`.
Le nombre d'individus étant, dans chaque cas, supérieur au nombre de variables respiratoires observées, nous réalisons des AFD.

Résultats pour 48 heures d'exposition
Relativement aux autres lignées :

- Les cellules CTA et CTRPA exposées à DMSO, BAP, TCDD et Mix possèdent de hautes valeurs de `basalResp`, `maxResp`, `protons` et `glyco`.
- Les cellules CTP exposées à BAP possèdent de faibles valeurs de `protons`.

Résultats pour 120 heures d'exposition
Relativement aux autres lignées :

- Les cellules CTA et CTRPA exposées à DMSO et Mix possèdent de hautes valeurs de `basalResp`, `maxResp`, `protons` et `glyco`. Exposées à BAP, Pyr et TCDD, elles possèdent de hautes valeurs de `basalResp`, `maxResp` et `glyco`.
- Les cellules CTP exposées à Mix possèdent de hautes valeurs de `protons`.

6 Caractérisation des polluants alimentaires

Les analyses précédentes ont été reproduites afin de caractériser les polluants après 48 et 120 heures d'exposition. Nous avons observé les réponses respiratoires de cellules ayant été exposées pendant 48 et 120 heures à chaque polluant, puis identifié les lignées cellulaires dont les réponses différaient de manière significative. Les résultats des analyses exploratoires, tests de comparaison et analyses multivariées sont retranscrits dans cette section.

6.1 Analyse exploratoire

Les distributions ont été représentées sous la formes de diagrammes-boîtes parallèles, disponibles en Annexe A. Après 48 heures comme après 120 heures d'exposition, on note la présence d'un effet plaque. À type cellulaire fixé, les variations inter-plaques semblent suivre un même schéma d'un traitement à l'autre et d'une variable à l'autre. L'effet inter-plaques sera facilement corrigible.
On note également, à `cellLine` et `plate` fixés, la présence de différences de distributions entre polluants.

6.2 Tests de comparaison

Les tests ont été réalisés indépendamment pour chaque lignée cellulaire. Les résultats des comparaisons deux-à-deux des polluants après 48 heures et 120 heures d'exposition ont été retranscrits dans les tables 9 et 10 respectivement.

Table 9 : Table résumant les lignées cellulaires pour lesquelles on observe des différences de distributions entre polluants après 48 heures d'exposition (seuil 5%), avant et après approximations des données par des modèles mixtes (en rouge : information spécifique à l'analyse paramétrique - en bleu : information spécifique à l'analyse non-paramétrique - en noir : information commune)

	basalResp	maxResp	protons	glyco
DMSO-BAP	-	-	CTR	-
DMSO-Pyr	CTP, CTR	-	-	-
DMSO-TCDD	CTA, CTR	CTA	-	-
DMSO-Mix	CTR, CTRPAt	CTRPAt	-	-
BAP-Pyr	CT	-	-	-
BAP-TCDD	-	CTA	-	-
BAP-Mix	CTRPAt	-	CTRPAt	CTRPAt
Pyr-TCDD	-	CTA	-	CTA
Pyr-Mix	-	-	CTP, CTRPAt	CTRPAt
TCDD-Mix	-	-	-	-

Table 10 : Table résumant les lignées cellulaires pour lesquelles on observe des différences de distributions entre polluants après 120 heures d'exposition (seuil 5%), avant et après approximations des données par des modèles mixtes (en rouge : information spécifique à l'analyse paramétrique - en bleu : information spécifique à l'analyse non-paramétrique - en noir : information commune)

	basalResp	maxResp	protons	glyco
DMSO-BAP	-	-	-	-
DMSO-Pyr	-	CTRPA	-	CT, CTP
DMSO-TCDD	CT, CTA, CTP, CTR, CTRPAt	CT, CTA, CTRPAt	CTA	CT, CTA, CTRPA, CTRPAt
DMSO-Mix	CT, CTA, CTP, CTR, CTRPA, CTRPAt	CT, CTA, CTRPAt	CT, CTP	CT, CTA, CTP, CTRPA, CTRPAt
BAP-Pyr	-	-	CT, CTP	CTR
BAP-TCDD	CT, CTA, CTRPA, CTRPAt	CTA, CTRPA	CT, CTA, CTP	CT, CTA, CTRPA, CTRPAt
BAP-Mix	CT, CTA, CTR, CTRPA, CTR-PAt	CTA, CTRPA	CTA, CTRPA, CTRPAt	CT, CTA, CTP, CTR, CTRPA, CTRPAt
Pyr-TCDD	CT, CTA, CTRPA, CTR-PAt	CT, CTA, CTRPA	CT, CTA	CTA, CTR, CTRPA, CTRPAt
Pyr-Mix	CTA, CTP, CTR, CTRPA, CTRPAt	CT, CTA, CTP, CTRPA	CT, CTRPA, CTRPAt	CTA, CTP, CTR, CTRPA, CTRPAt
TCDD-Mix	-	-	CT, CTA	CTP

Nos comparaisons étant menées à lignée cellulaire fixe, nous n'avons pas de problème de confusion d'effet lignée et d'effet plaque lorsque l'on étudie les différences de distributions entre polluants. Les résultats obtenus des tests incluant les effets aléatoires sont donc fiables.

D'après les tables 9 et 10, les polluants TCDD et Mix semblent se différencier des autres polluants sur un grand nombre de variables et pour un grand nombre de lignées. Ainsi, pour des expositions de 48 heures comme de 120 heures, les traitements à la dioxine TCDD et au mélange Mix sont les plus impactants.

6.3 Analyses multivariées

6.3.1 Analyse en Composantes Principales

Comme dans les cas précédents, le graphe des individus obtenu de l'ACP ne permet pas d'établir de distinction claire entre les groupes de polluants. C'est, encore une fois, dû à la trop grande variabilité inter-plaques des distributions pour un même polluant. Les cercles de corrélations révèlent la présence d'effets taille entre variables.

6.3.2 Analyses Discriminantes

Le nombre d'individus étant supérieur au nombre de variables respiratoires observées, nous réalisons des AFD. Nous parvenons aux caractérisations suivantes après 48 et 120 heures d'exposition :

48 heures d'exposition
Relativement aux autres polluants, une exposition à BAP conduit à de hautes valeurs de `maxResp` et `glyco` pour les cellules CTP.

120 heures d'exposition
Relativement aux autres polluants, des expositions à TCDD et à Mix conduisent à de hautes valeurs de `basalResp`, `maxResp`, `protons` et `glyco` pour les cellules CTA et CTRPAt.

Conclusion

Par l'intermédiaire de tests de comparaisons inter-groupes paramétriques et non paramétriques (Kruskal Wallis et modèles mixtes généralisés) et d'analyses multivariées prédictives (Analyse Factorielle Discriminante et PLS discriminante), j'ai pu caractériser les différences entre des lignées cellulaires mutées avant effet du contaminant. J'ai trouvé que la lignée cellulaire la plus proche des lignées cancéreuses, à savoir CTRPAt, se différenciait des autres sur un plus grand nombre de caractéristiques.
Je me suis ensuite intéressée aux effets des polluants sur ces lignées. En réutilisant les mêmes méthodes, j'ai d'abord caractérisé les différences entre les lignées dans leurs réponses à chaque traitement, puis, pour chaque lignée, les différences de réponses aux différents traitements. De nouveau, la lignée CTRPAt s'est différenciée des autres. J'ai également pu conclure que les traitements à la toxine (TCDD) et au mélange (Mix) étaient les plus impactants.

Ces résultats étaient attendus. En effet, les cellules CTRPAt sont les plus ressemblantes aux cellules observées en cas de cancer du colon, et la toxine TCDD est connue des biologistes pour son fort impact sur les cellules, qu'elle soit seule ou combinée à d'autres polluants (comme ici, combinée à BAP et à Pyr pour former Mix).

Ce stage m'a permis de consolider mes compétences en analyse statistique et d'en acquérir de nouvelles. La mise en œuvre, sous R, d'analyses exploratoires, tests de comparaison, et analyses multivariées, m'a permis de me familiariser avec de nouveaux paquetages dédiés à l'étude de données massives. Les présentations de mes résultats, sous la forme de rapports ou d'exposés oraux, m'ont aussi fait travailler mon expression orale et écrite.

J'ai également pu tirer quelques enseignements, qui me seront assurément utiles dans mon futur professionnel. Je retiens notamment l'importance d'une bonne communication entre statisticiens et biologistes. Il est primordial, pour les statisticiens, d'être informé, dès le début de la collaboration, des détails du dispositif expérimental mis en place. J'ai aussi constaté l'importance de randomiser les données avant de commencer les analyses. Dans le cadre d'études non randomisées comme la nôtre, il n'est pas toujours aisé de corriger les éventuelles variations aléatoires liées aux conditions expérimentales.

Le fait de mettre mes connaissances théoriques au service d'un problème de santé publique, dans le cadre du projet METAhCOL, a été extrêmement stimulant et gratifiant. Je suis très reconnaissante d'avoir pu apporter ma contribution à ce projet.

Références

[Abdolmohammadi, 2017] Abdolmohammadi, A. (2017). Chapitre 6 : Pcr (polymerase chain reaction). Consulté à l'adresse https://borzuya.org/index.php/fr/universitaire-specialise/2eme-annee/biotechnologie/.

[Agr'OGM, 2019] Agr'OGM (2019). Que sont les ogm ? Consulté à l'adresse https://agr-ogm.site123.me/sommaire/2-que-sont-les-ogm.

[Besse, 2013] Besse, P. (2013). Modèles à effets aléatoires et modèles mixtes. Consulté à l'adresse https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-modmixt6-modmixt.pdf.

[Besse, 2014] Besse, P. (2014). Analyse en composantes principales (acp). Consulté à l'adresse https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-explo-acp.pdf.

[Besse, 2017] Besse, P. (2017). Composantes principales et régressions pls parcimonieuses. Consulté à l'adresse https://www.math.univ-toulouse.fr/~besse/Wikistat/pdf/st-m-app-sparse-pls.pdf.

[Chavent, 2015] Chavent, M. (2015). Analyse factorielle discriminante (afd). Consulté à l'adresse http://www.math.u-bordeaux.fr/~mchave100p/wordpress/wp-content/uploads/2013/10/AFD.pdf.

[Clinisciences, 2019] Clinisciences (2019). Pcr quantitative - qpcr. Consulté à l'adresse https://www.clinisciences.com/achat/cat-pcr-quantitative-qpcr-3511.html.

[Couturier, 2019] Couturier, E. (2019). Les polluants les plus fréquents dans notre alimentation. Consulté à l'adresse https://www.consoglobe.com/polluants-plus-frequents-alimentation-cg.

[Déjean, 2019] Déjean, S. (2019). Multivariate projection methodologies for the exploration of large biological data sets. Consulté à l'adresse http://www.nathalievialaneix.eu/teaching/doclipm/slides-mixomics.pdf.

[Eurogentec, 2015] Eurogentec (2015). qpcrguide. Consulté à l'adresse https://secure.eurogentec.com/uploads/FileBrowse/Technical%20Guides/Genomic%20Solutions/qPCR-booklet.pdf.

[Futura, 2019] Futura (2019). Transcription. Consulté à l'adresse https://www.futura-sciences.com/sante/definitions/biologie-transcription-271/).

[kartable, 2019] kartable (2019). La respiration cellulaire. Consulté à l'adresse https://www.kartable.fr/ressources/svt/cours/la-respiration-cellulaire/19434.

[Lenth, 2018] Lenth, R. (2018). Basic of estimated marginal means. Consulté à l'adresse https://mran.microsoft.com/snapshot/2018-05-29/web/packages/emmeans/vignettes/basics.html.

[LesBonsProfs, 2014] LesBonsProfs (2014). La respiration cellulaire. Consulté à l'adresse https://www.lesbonsprofs.com/svt/la-respiration-cellulaire-932).

[Raufaste, 2013] Raufaste, E. (2013). Comparaisons non planifiées : tests post-hoc. Consulté à l'adresse http://w3.uohpsy2.univ-tlse2.fr/UOHPsy2/index.php?option=com_content&task=view&id=192&Itemid=30.

[Singull, 2019a] Singull, M. (2019a). Experimental design and biostatistics : lecture 2 - one-way analysis. Consulté à l'adresse http://courses.mai.liu.se/GU/TAMS38/Dokument/fo2.pdf.

[Singull, 2019b] Singull, M. (2019b). Experimental design and biostatistics : lecture 3 - pairwise comparisons. Consulté à l'adresse http://courses.mai.liu.se/GU/TAMS38/Dokument/fo3.pdf.

[Singull, 2019c] Singull, M. (2019c). Experimental design and biostatistics : lecture 4 - non-parametric methods. Consulté à l'adresse http://courses.mai.liu.se/GU/TAMS38/Dokument/fo4.pdf.

Annexes

Annexe A : Représentation des distributions

Caractérisation des lignées cellulaires après 48 heures d'exposition

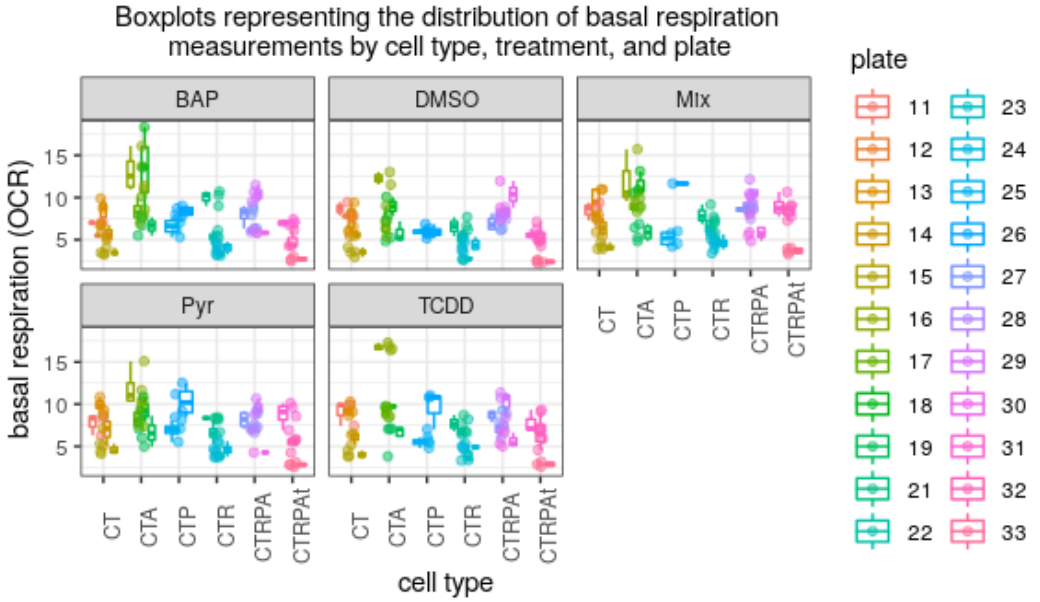


FIGURE 27: Diagrammes-boîtes parallèles représentant les distributions de `basalResp` par traitement, lignée cellulaire et plaque

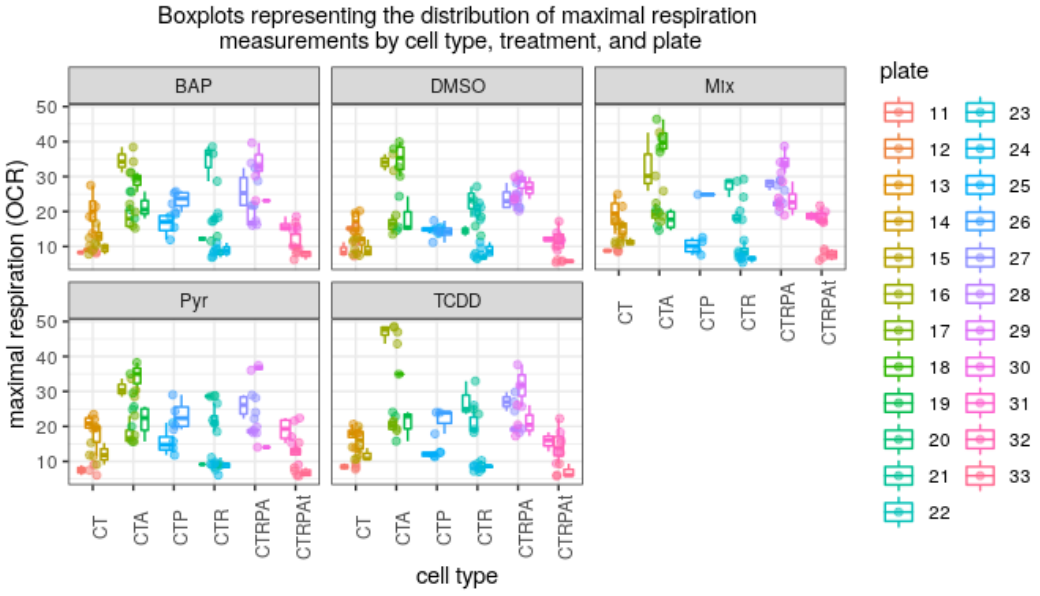


FIGURE 28: Diagrammes-boîtes parallèles représentant les distributions de `maxResp` par traitement, lignée cellulaire et plaque

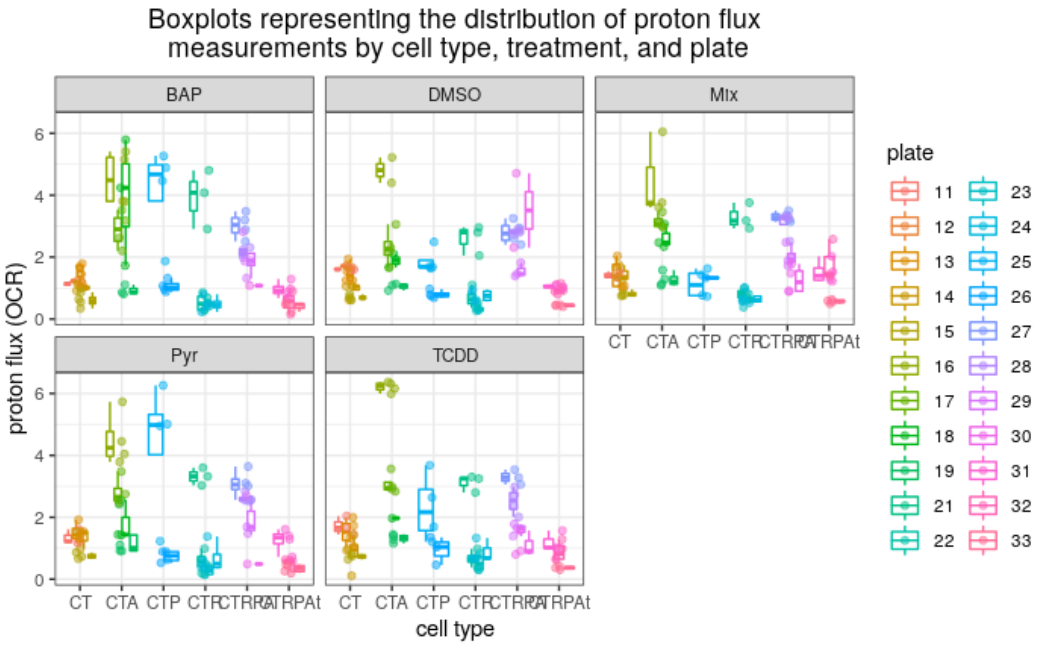


FIGURE 29: Diagrammes-boîtes parallèles représentant les distributions de `protons` par traitement, lignée cellulaire et plaque

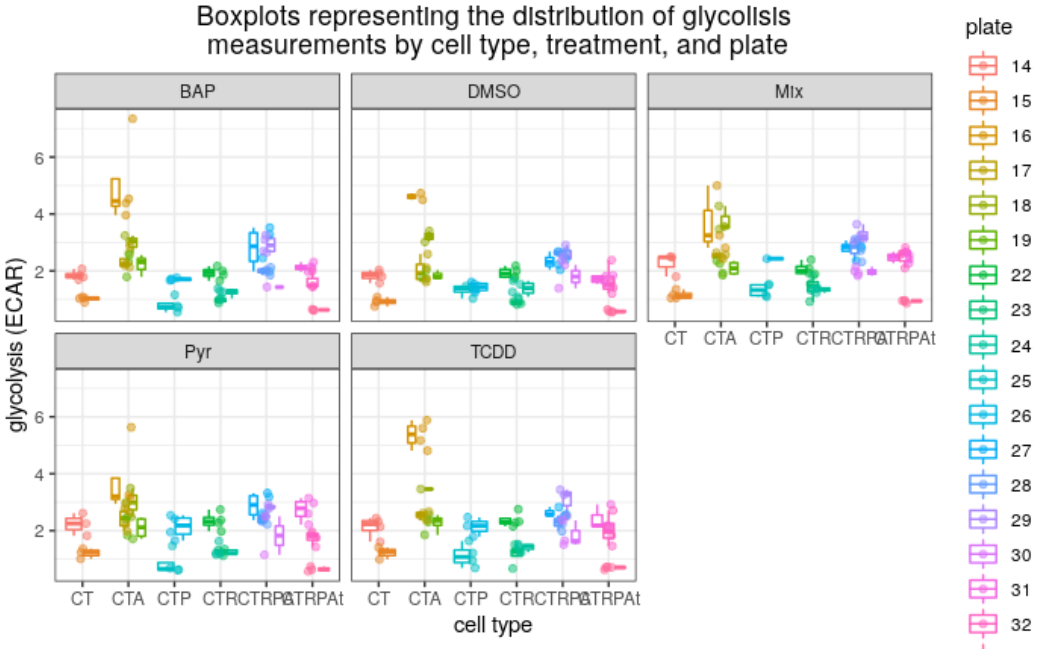


FIGURE 30: Diagrammes-boîtes parallèles représentant les distributions de `glyco` par traitement, lignée cellulaire et plaque

Caractérisation des lignées cellulaires après 120 heures d'exposition

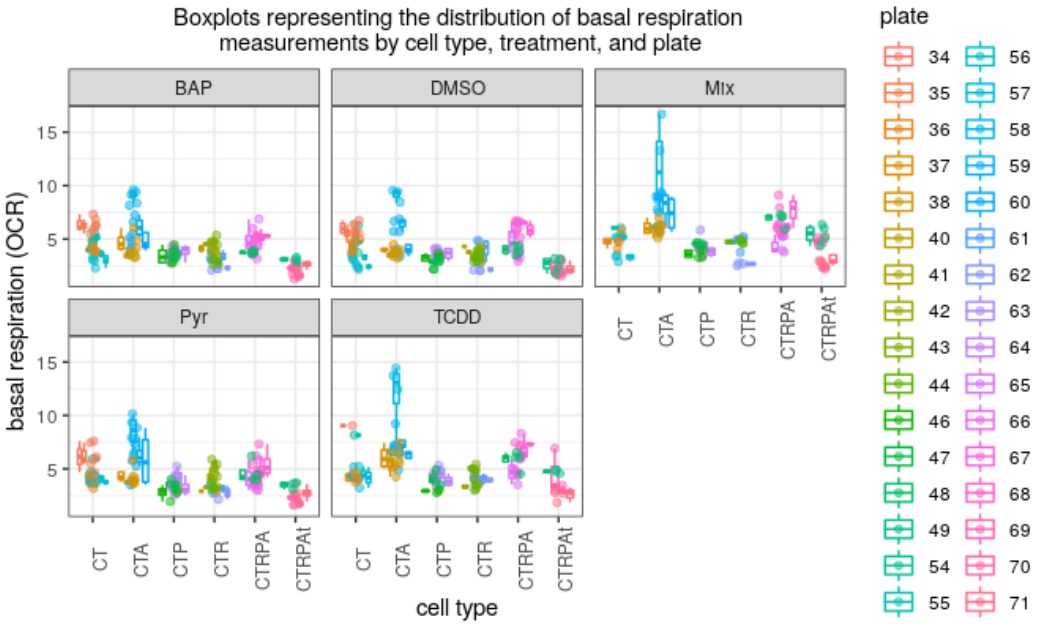


FIGURE 31: Diagrammes-boîtes parallèles représentant les distributions de `basalResp` par lignée cellulaire et plaque, pour chaque traitement

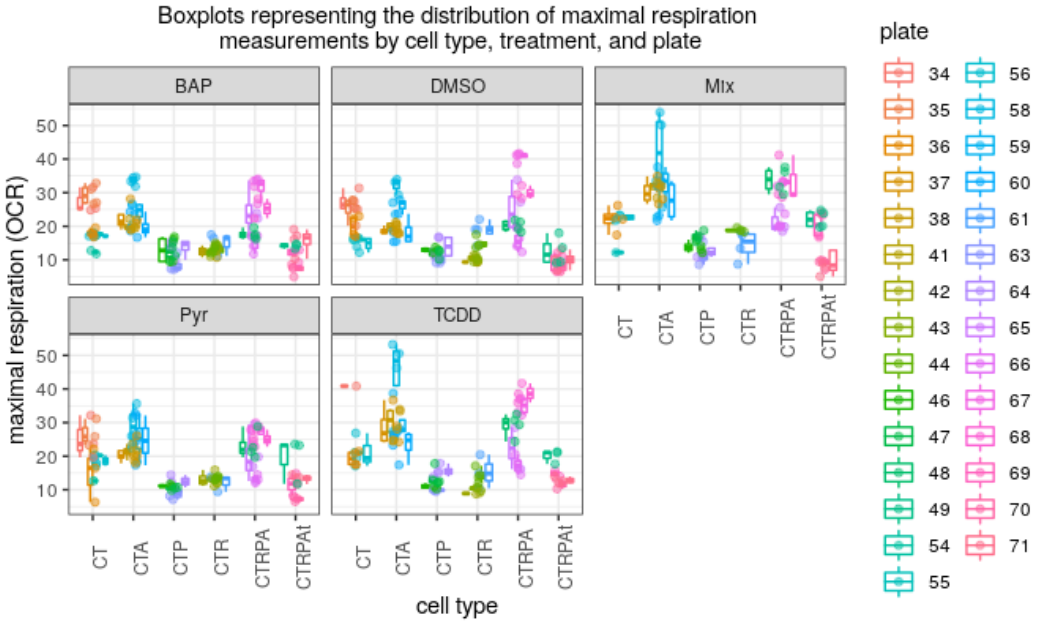


FIGURE 32: Diagrammes-boîtes parallèles représentant les distributions de `maxResp` par lignée cellulaire et plaque, pour chaque traitement

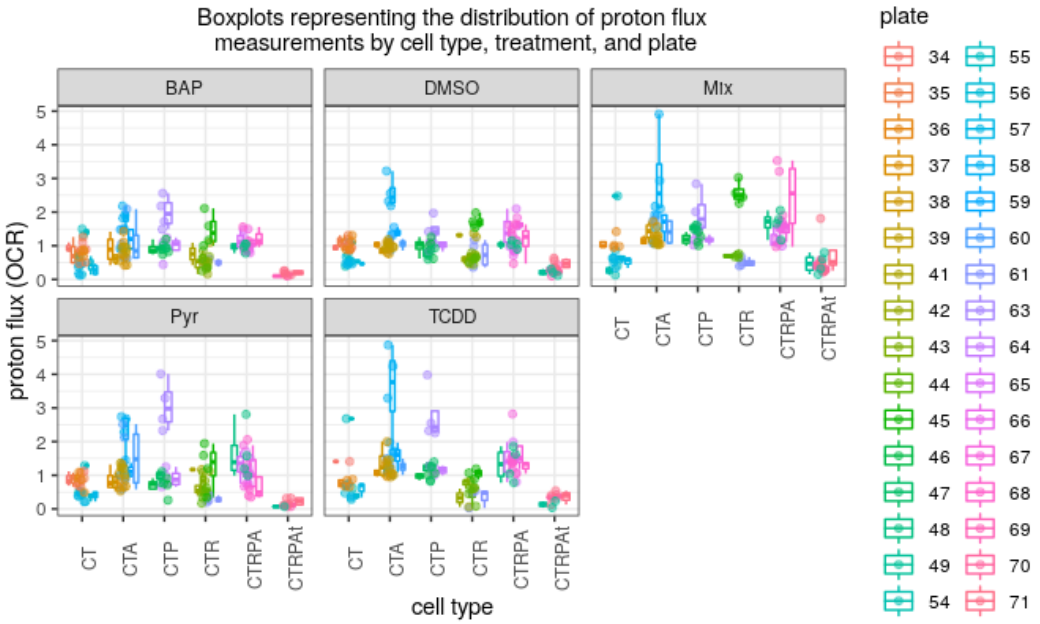


FIGURE 33: Diagrammes-boîtes parallèles représentant les distributions de `protons` par lignée cellulaire et plaque, pour chaque traitement

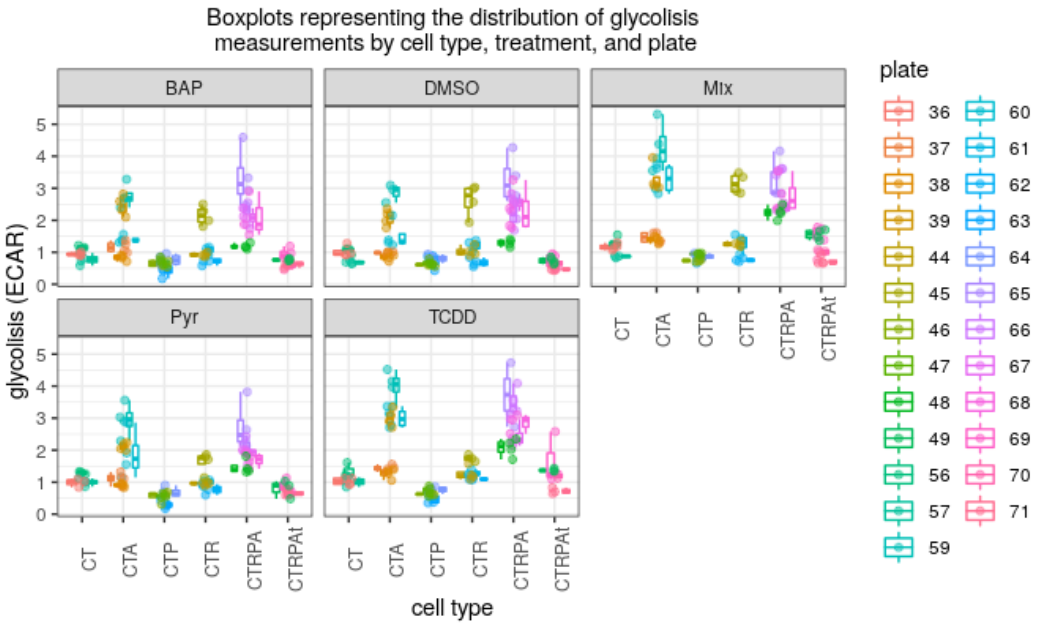


FIGURE 34: Diagrammes-boîtes parallèles représentant les distributions de `glyco` par lignée cellulaire et plaque, pour chaque traitement

Caractérisation des polluants après 48 heures d'exposition

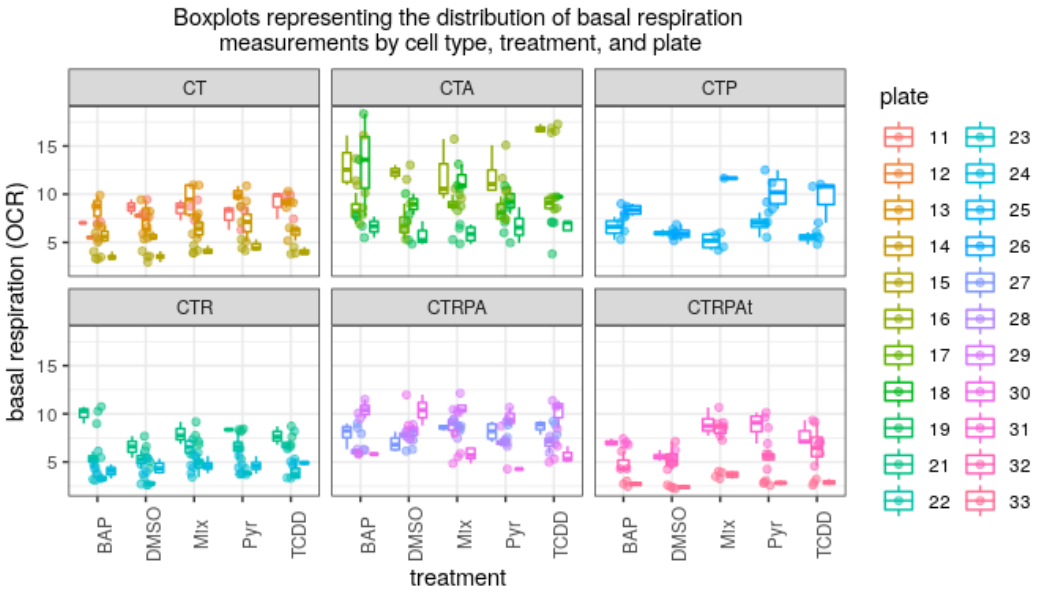


FIGURE 35: Diagrammes-boîtes parallèles représentant les distributions de `basalResp` par traitement et plaque, pour chaque lignée cellulaire

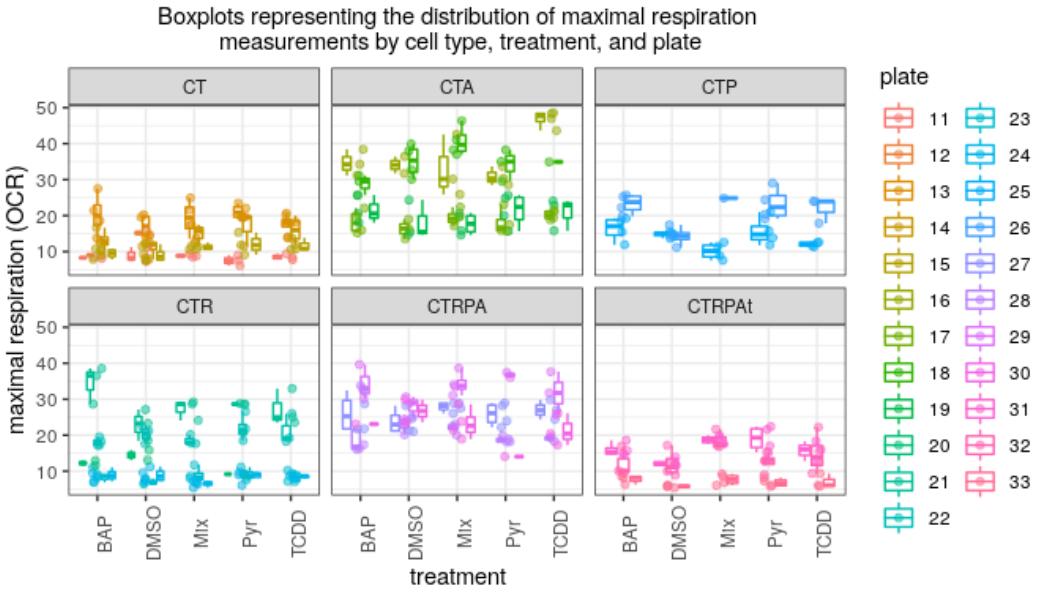


FIGURE 36: Diagrammes-boîtes parallèles représentant les distributions de `maxResp` par traitement et plaque, pour chaque lignée cellulaire

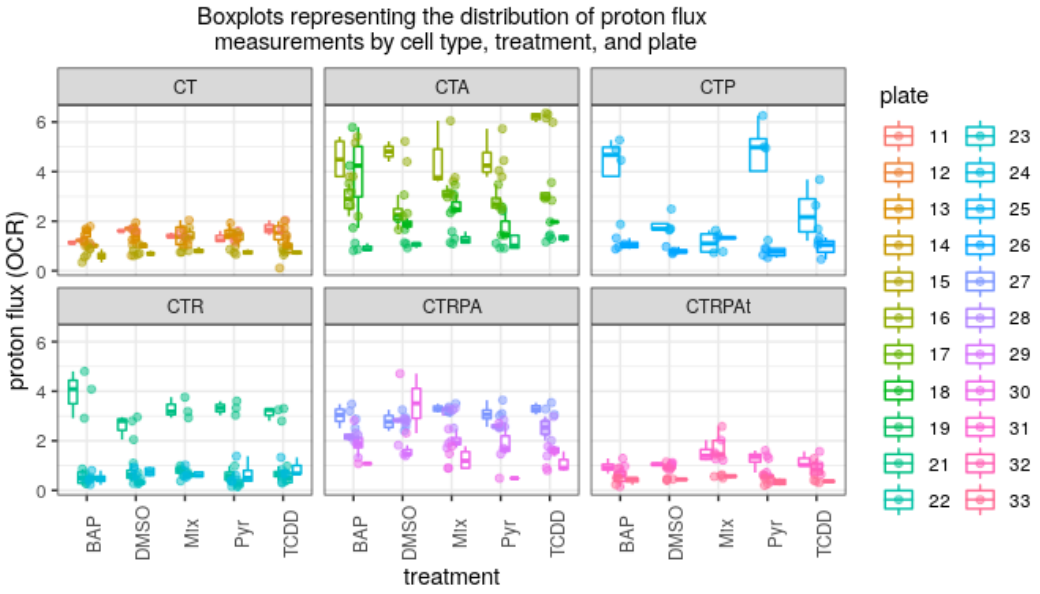


FIGURE 37: Diagrammes-boîtes parallèles représentant les distributions de `protons` par traitement et plaque, pour chaque lignée cellulaire

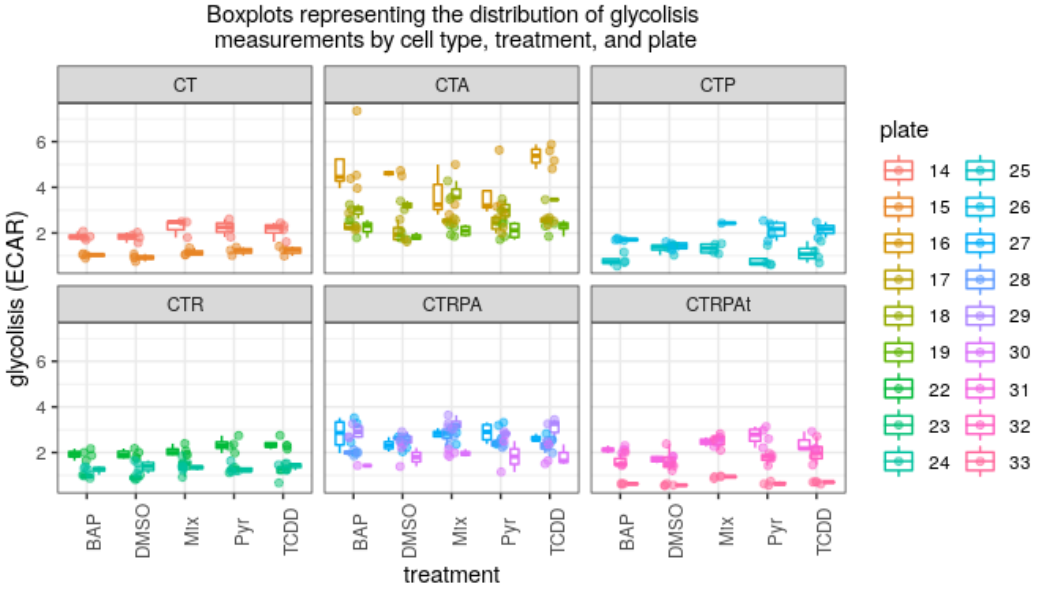


FIGURE 38: Diagrammes-boîtes parallèles représentant les distributions de `glyco` par traitement et plaque, pour chaque lignée cellulaire

Caractérisation des polluants après 120 heures d'exposition

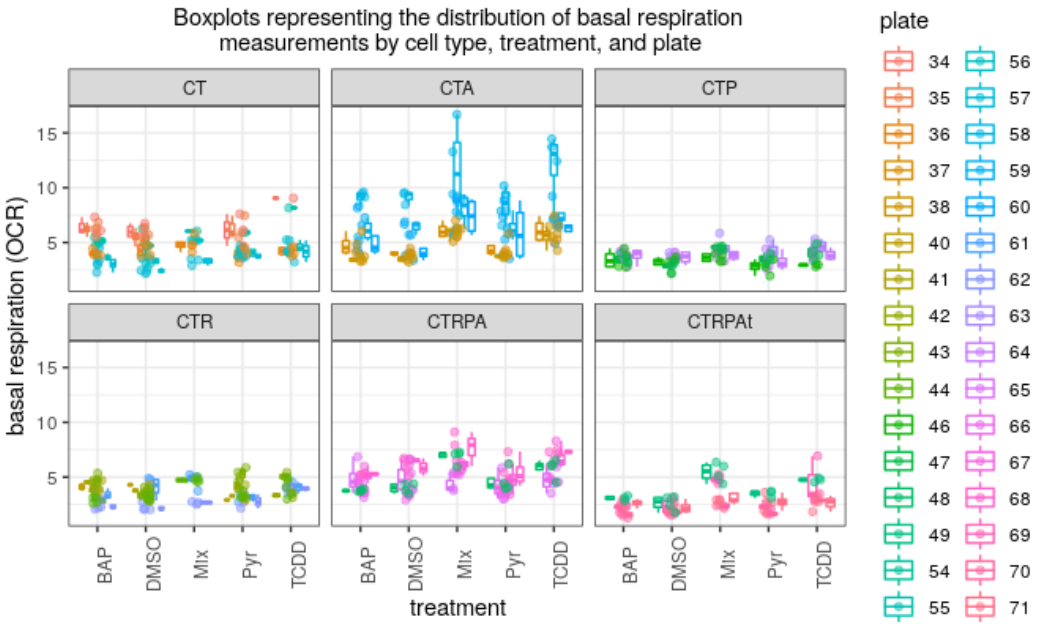


FIGURE 39: Diagrammes-boîtes parallèles représentant les distributions de `basalResp` par traitement et plaque, pour chaque lignée cellulaire

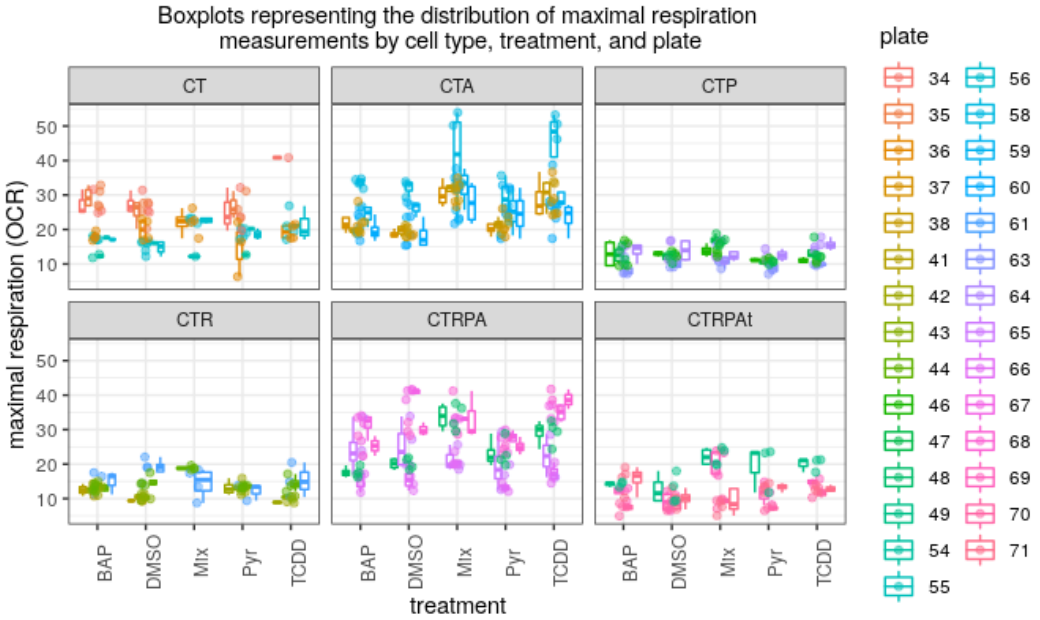


FIGURE 40: Diagrammes-boîtes parallèles représentant les distributions de `maxResp` par traitement et plaque, pour chaque lignée cellulaire

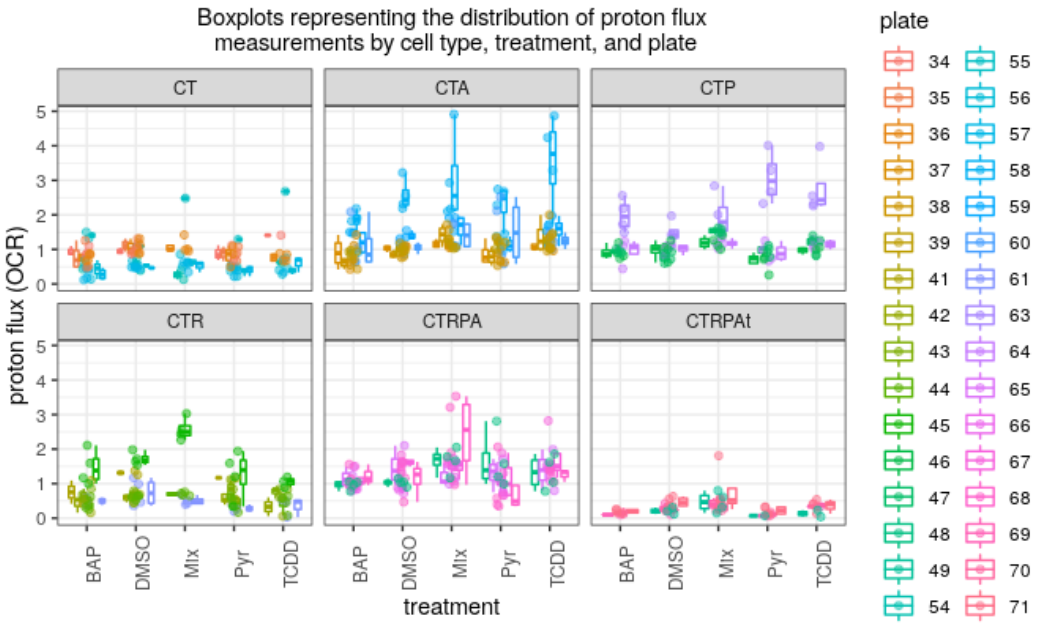


FIGURE 41: Diagrammes-boîtes parallèles représentant les distributions de `protons` par traitement et plaque, pour chaque lignée cellulaire

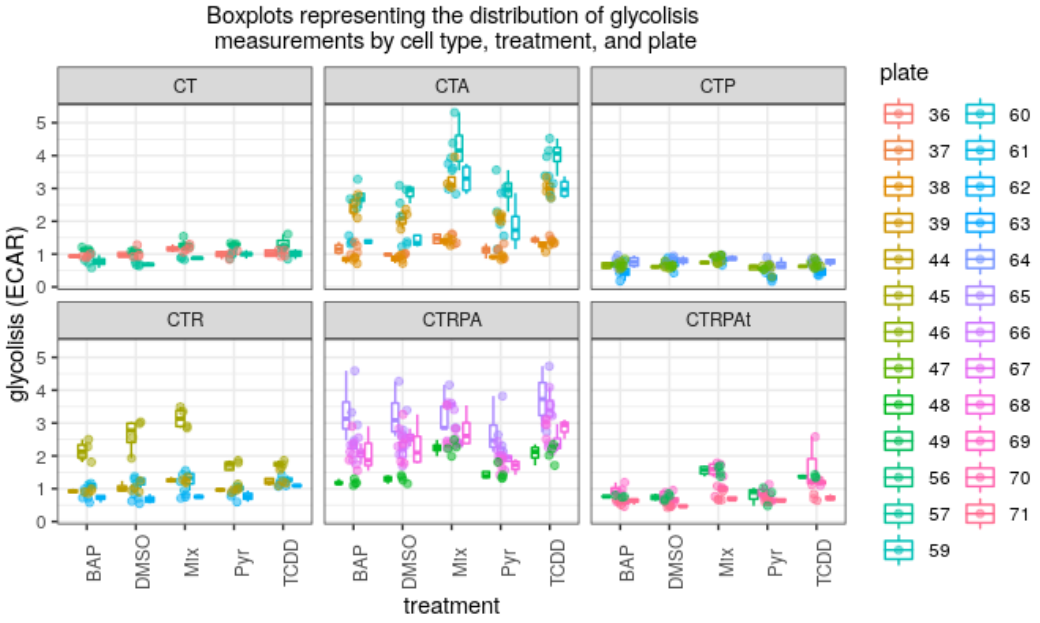


FIGURE 42: Diagrammes-boîtes parallèles représentant les distributions de `glyco` par traitement et plaque, pour chaque lignée cellulaire

Annexe B : Compte-rendu à destination des biologistes

Ce rapport reprend le détail des analyses effectuées pour la caractérisation des lignées cellulaires en conditions basales. Des analyses similaires ont été menées pour la caractérisation des lignées cellulaires et des polluants alimentaires après 48 et 120 heures d'exposition.

Characterization of cell types under basal conditions

Fanny Mathevet

27 septembre, 2019

Contents

Introduction

In this report, we aim at characterizing 6 types of cells under basal conditions. Measurements of basal and maximal respirations, glycolysis, proton flux and 40 gene expressions have been made on these cells in several replicates and in several experimental contexts. We first conduct an exploratory analysis, then perform comparison tests, and finally focus on two multivariate analysis : the Principal Component Analysis (PCA), and the Factorial Discriminant Analysis (FDA).

This document focuses on basal conditions.

1. Data import and first handling

Data under basal conditions (first lines of the data frame) :

```
##   cellLine expe replicate plate  6P6D  ACO1   AhR  AhRR ATP5IF1 basalResp
## 1       CT    2        A1    10 15.50 15.53 15.54 17.85   17.76        NA
## 2       CT    2        A2    10 16.44 16.56 18.11 17.92   19.83        NA
## 3       CT    2        A3    10 15.45 16.23 17.20 17.84   19.17        NA
## 4       CT    2        A4    10 17.49 16.89 18.35 19.38   20.04        NA
## 5       CT    2        A5    10 16.66 16.03 16.76 18.08   19.04        NA
## 6       CT    2        A6    10 16.83 15.17 17.39 18.72   19.51        NA
##     CAT CYP1A1 CYP1A2 CYP3A4 ENIO1    FH  G6PD gene_inc glyco   HK2 HMGCR
## 1 15.47  23.91  21.40  25.49  9.66 15.39 13.28    15.20    NA 16.69 15.47
## 2 15.64  24.08  21.61  26.38 11.55 17.42 14.33    16.46    NA 16.92 16.20
## 3 15.21  23.15  20.22  25.96 10.92 16.05 13.36    15.41    NA 16.77 15.97
## 4 16.65  25.12     NA  26.49 12.10 18.26 15.40    17.33    NA 18.46 15.96
## 5 16.52  24.03     NA  23.81 10.43 16.07 14.92    15.31    NA 17.45 15.76
## 6 16.40     NA     NA     NA 11.08 16.55 14.75    15.60    NA 16.55 14.98
##    HMOX  IDH1  LDHA  LDHB LPCAT maxResp  MCT4  MFN2 Mitoferrin1
## 1 15.56 14.71 12.62 11.91 16.26      NA 14.42 15.52       16.66
## 2 15.80 15.36 14.61 13.21 17.87      NA 15.35 17.22       17.91
## 3 15.49 14.90 13.14 12.23 17.53      NA 15.69 16.40       17.02
## 4 17.45 15.88 14.52 14.00 18.44      NA 16.40 18.42       19.94
## 5 16.65 15.06 12.83 12.45 16.88      NA 15.05 16.45       18.51
## 6 17.39 15.03 12.86 12.22 16.92      NA 14.75 16.66       17.97
##   Mitoferrin2   ND1  NHE1  NQO1  NRF2  PKM1  PKM2 PRDX1 protons  RDK1
## 1       16.64  8.37 17.16 11.31 11.70 15.25 11.63 12.52      NA 17.93
## 2       18.23  9.65 18.15 13.19 13.41 16.81 13.09 13.76      NA 19.24
## 3       17.45  9.05 17.55 10.96 11.73 15.40 11.80 12.71      NA 18.31
## 4       18.84 11.89 19.04 12.89 13.65 17.34 13.97 13.49      NA 20.05
## 5       17.12 10.74 18.15 12.41 13.45 16.48 12.57 12.65      NA 18.88
## 6       18.40  9.88 18.29 12.60 13.34 15.96 12.45 12.64      NA 18.69
##    SCD1  SDHA  SDHC SIRT3  TFAM TIGAR  TSPO UQCC3
## 1 12.29 15.05 14.13 18.77 16.90    NA 19.08 17.16
## 2 13.05 16.04 14.60 18.53 18.95    NA 19.53 19.27
## 3 12.91 15.29 14.04 19.60 17.79    NA 18.84 18.88
## 4 13.74 16.34 15.64 19.76 19.44 17.88 20.60 19.32
## 5 12.49 15.36 14.73 19.85 16.87 16.11 18.93 19.22
## 6 12.76 15.82 14.92 19.09 17.27 16.32 19.72 19.14
```

Data summary :

```
##    cellLine       expe       replicate       plate          6P6D      
##  CT    :26   2      : 34   A3     : 14   10     :34   Min.   :14.66  
##  CTA   :33   7      :  9   A4     : 14   5      :20   1st Qu.:15.63  
##  CTP   :25   21     :  9   A5     : 14   4      :19   Median :16.05  
##  CTR   :32   26     :  7   A6     : 14   8      :19   Mean   :16.20  
##  CTRPA :33   8      :  6   A2     : 13   1      :18   3rd Qu.:16.77  
##  CTRPAt:32   12     :  6   B3     :  9   2      :18   Max.   :17.77  
##              (Other):110   (Other):103   (Other):53   NA's   :147    
##       ACO1            AhR             AhRR          ATP5IF1      
##  Min.   :14.61   Min.   :15.06   Min.   :16.68   Min.   :17.38  
##  1st Qu.:15.44   1st Qu.:15.80   1st Qu.:17.74   1st Qu.:18.52  
```

```
##  Median :16.02   Median :16.77   Median :18.00   Median :19.07
##  Mean   :16.00   Mean   :16.70   Mean   :18.31   Mean   :19.06
##  3rd Qu.:16.54   3rd Qu.:17.30   3rd Qu.:18.96   3rd Qu.:19.73
##  Max.   :18.07   Max.   :18.46   Max.   :20.22   Max.   :20.70
##  NA's   :147     NA's   :147     NA's   :147     NA's   :147
##    basalResp          CAT            CYP1A1          CYP1A2
##  Min.   :1.450   Min.   :15.18   Min.   :21.64   Min.   :17.35
##  1st Qu.:2.955   1st Qu.:15.43   1st Qu.:22.95   1st Qu.:19.14
##  Median :3.670   Median :15.82   Median :23.86   Median :19.64
##  Mean   :3.617   Mean   :16.00   Mean   :23.95   Mean   :19.66
##  3rd Qu.:4.253   3rd Qu.:16.49   3rd Qu.:24.18   3rd Qu.:20.39
##  Max.   :6.320   Max.   :17.63   Max.   :27.68   Max.   :21.61
##  NA's   :35      NA's   :147     NA's   :153     NA's   :165
##      CYP3A4          ENIO1             FH             G6PD
##  Min.   :20.41   Min.   : 9.36   Min.   :14.70   Min.   :12.78
##  1st Qu.:23.95   1st Qu.:10.33   1st Qu.:15.59   1st Qu.:13.42
##  Median :25.26   Median :10.87   Median :16.14   Median :14.30
##  Mean   :25.18   Mean   :10.92   Mean   :16.22   Mean   :14.34
##  3rd Qu.:25.98   3rd Qu.:11.51   3rd Qu.:16.98   3rd Qu.:15.17
##  Max.   :29.40   Max.   :12.49   Max.   :18.26   Max.   :16.36
##  NA's   :153     NA's   :147     NA's   :147     NA's   :148
##     gene_inc         glyco             HK2            HMGCR
##  Min.   :14.50   Min.   :0.3500   Min.   :16.40   Min.   :13.93
##  1st Qu.:15.19   1st Qu.:0.7975   1st Qu.:16.90   1st Qu.:14.99
##  Median :15.60   Median :1.0200   Median :17.48   Median :15.46
##  Mean   :15.71   Mean   :1.1133   Mean   :17.68   Mean   :15.50
##  3rd Qu.:15.96   3rd Qu.:1.3525   3rd Qu.:18.27   3rd Qu.:15.97
##  Max.   :17.33   Max.   :2.5500   Max.   :19.79   Max.   :17.47
##  NA's   :147     NA's   :37       NA's   :147     NA's   :147
##       HMOX            IDH1            LDHA            LDHB
##  Min.   :14.18   Min.   :14.06   Min.   :11.62   Min.   :11.34
##  1st Qu.:14.86   1st Qu.:14.48   1st Qu.:12.12   1st Qu.:11.98
##  Median :15.58   Median :14.91   Median :12.89   Median :12.36
##  Mean   :15.68   Mean   :14.92   Mean   :13.10   Mean   :12.53
##  3rd Qu.:16.34   3rd Qu.:15.35   3rd Qu.:14.01   3rd Qu.:13.16
##  Max.   :17.98   Max.   :15.96   Max.   :14.96   Max.   :14.00
##  NA's   :147     NA's   :147     NA's   :147     NA's   :147
##      LPCAT          maxResp            MCT4            MFN2
##  Min.   :15.68   Min.   : 1.810   Min.   :13.65   Min.   :15.19
##  1st Qu.:16.45   1st Qu.: 5.540   1st Qu.:14.35   1st Qu.:16.05
##  Median :17.16   Median : 8.100   Median :14.73   Median :16.48
##  Mean   :17.08   Mean   : 8.702   Mean   :14.83   Mean   :16.61
##  3rd Qu.:17.62   3rd Qu.:10.950   3rd Qu.:15.26   3rd Qu.:16.89
##  Max.   :18.44   Max.   :20.130   Max.   :16.40   Max.   :18.42
##  NA's   :147     NA's   :38       NA's   :147     NA's   :148
##   Mitoferrin1     Mitoferrin2         ND1             NHE1
##  Min.   :15.87   Min.   :14.81   Min.   : 8.370   Min.   :16.49
##  1st Qu.:16.71   1st Qu.:17.15   1st Qu.: 8.965   1st Qu.:17.55
##  Median :17.09   Median :17.80   Median : 9.245   Median :17.93
##  Mean   :17.68   Mean   :17.73   Mean   : 9.672   Mean   :17.99
##  3rd Qu.:18.53   3rd Qu.:18.16   3rd Qu.:10.553   3rd Qu.:18.30
##  Max.   :20.28   Max.   :19.77   Max.   :11.890   Max.   :19.79
##  NA's   :147     NA's   :147     NA's   :147      NA's   :147
##       NQO1            NRF2            PKM1            PKM2
```

```
##  Min.   :10.56   Min.   :11.13   Min.   :14.47   Min.   :11.21
##  1st Qu.:11.38   1st Qu.:12.36   1st Qu.:15.21   1st Qu.:11.89
##  Median :12.21   Median :13.14   Median :15.79   Median :12.40
##  Mean   :12.27   Mean   :13.27   Mean   :15.89   Mean   :12.45
##  3rd Qu.:13.15   3rd Qu.:14.08   3rd Qu.:16.62   3rd Qu.:13.01
##  Max.   :14.61   Max.   :16.43   Max.   :17.68   Max.   :13.97
##  NA's   :147     NA's   :147     NA's   :147     NA's   :147
##      PRDX1           protons            RDK1            SCD1
##  Min.   :11.55   Min.   :0.0200   Min.   :17.01   Min.   :10.89
##  1st Qu.:12.47   1st Qu.:0.3625   1st Qu.:17.78   1st Qu.:12.01
##  Median :12.88   Median :0.5150   Median :18.20   Median :12.62
##  Mean   :13.06   Mean   :0.5613   Mean   :18.27   Mean   :12.60
##  3rd Qu.:13.45   3rd Qu.:0.6875   3rd Qu.:18.75   3rd Qu.:13.09
##  Max.   :15.04   Max.   :1.4200   Max.   :20.05   Max.   :15.68
##  NA's   :147     NA's   :39       NA's   :147     NA's   :147
##       SDHA            SDHC            SIRT3            TFAM
##  Min.   :14.30   Min.   :13.70   Min.   :18.44   Min.   :15.78
##  1st Qu.:15.05   1st Qu.:14.03   1st Qu.:18.70   1st Qu.:16.85
##  Median :15.39   Median :14.35   Median :18.95   Median :17.27
##  Mean   :15.53   Mean   :14.57   Mean   :19.13   Mean   :17.41
##  3rd Qu.:15.96   3rd Qu.:14.98   3rd Qu.:19.72   3rd Qu.:18.22
##  Max.   :17.09   Max.   :15.93   Max.   :20.50   Max.   :19.44
##  NA's   :147     NA's   :147     NA's   :147     NA's   :147
##      TIGAR            TSPO            UQCC3
##  Min.   :16.11   Min.   :17.84   Min.   :17.16
##  1st Qu.:17.01   1st Qu.:18.82   1st Qu.:18.13
##  Median :17.77   Median :19.00   Median :18.89
##  Mean   :17.70   Mean   :19.13   Mean   :18.71
##  3rd Qu.:18.47   3rd Qu.:19.34   3rd Qu.:19.26
##  Max.   :19.41   Max.   :20.85   Max.   :19.80
##  NA's   :163     NA's   :147     NA's   :147
```

Data for Multivariate Analysis :

We will perform the Multivariate Analysis using the largest complete data sets under basal conditions. These data sets contain simultaenous observations of some chosen variables. \

We will consider two data sets :

- The largest complete **respiration** data set under basal conditions (containing simultaneous observations of the respiration variables)
- the largest complete **genomic** data set under basal conditions (containing simultaneous observations of the genomic variables)

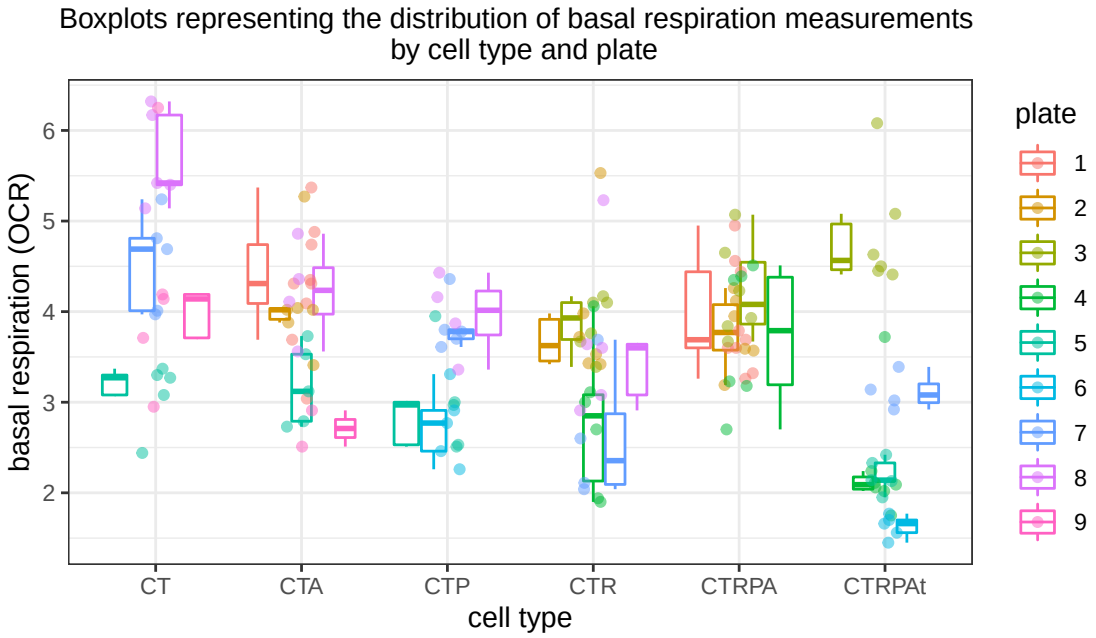
2. Exploratory analysis

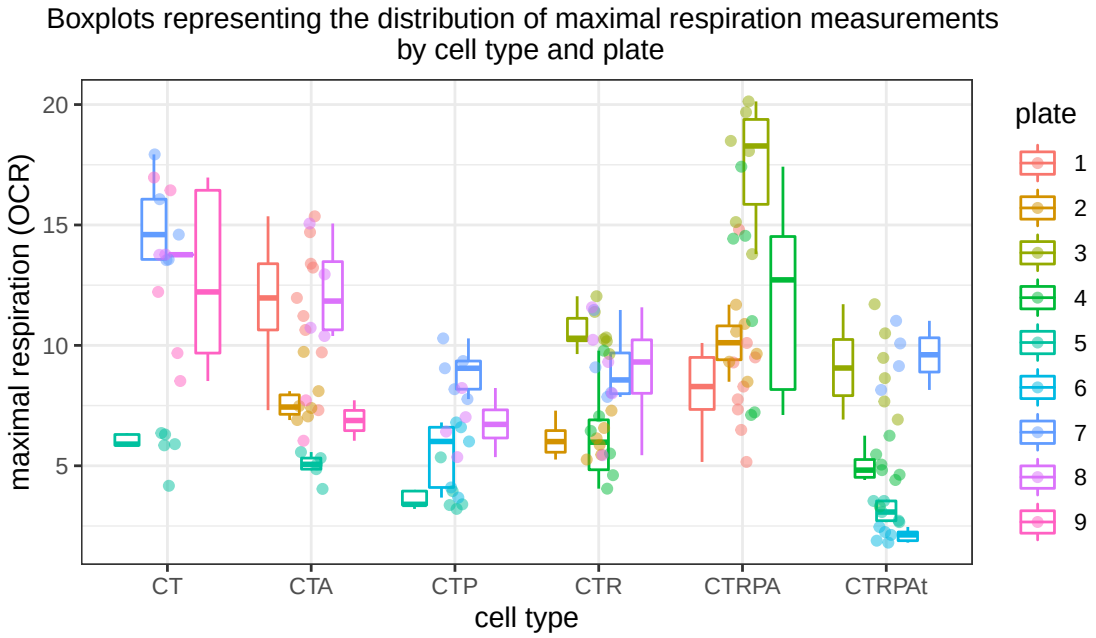
2.1. Considering the respiration variables

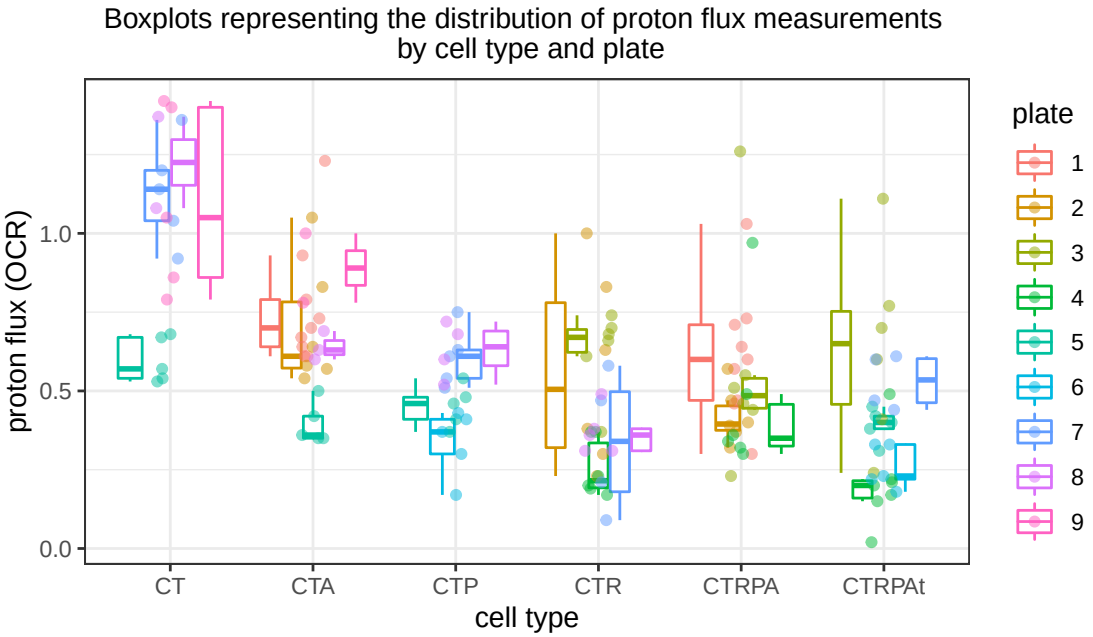
2.1.1. Visualization of the distributions

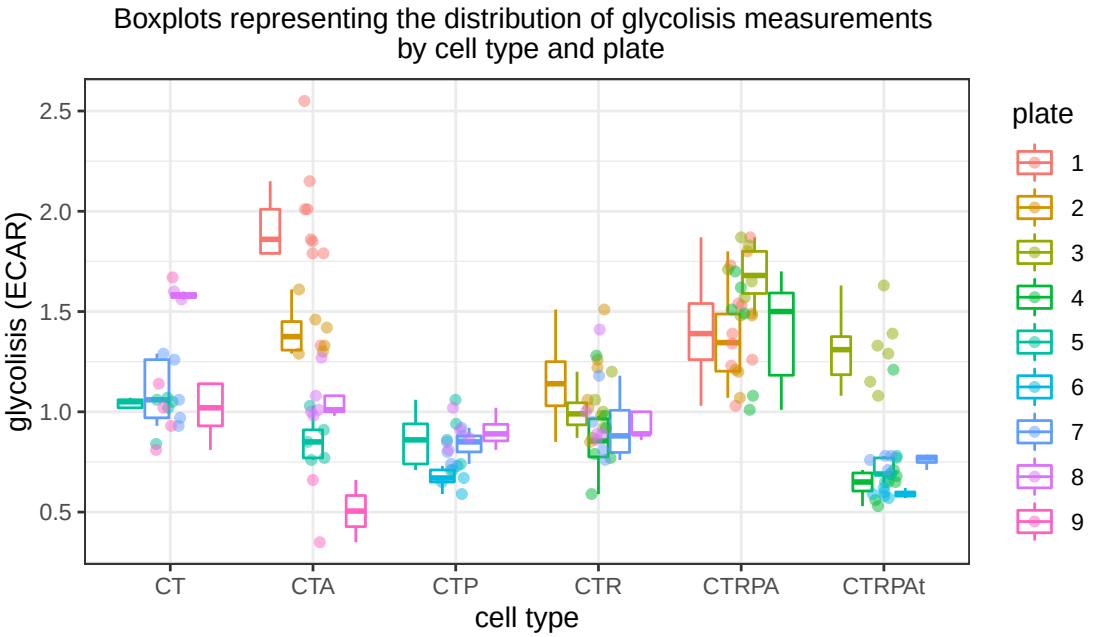
Distributions of the respiration variables given the cell type and the plate

To visualize the effects of plates on the measurements, we first plot the distributions of basalResp, maxResp, protons and glyco given the cell type and the plate.





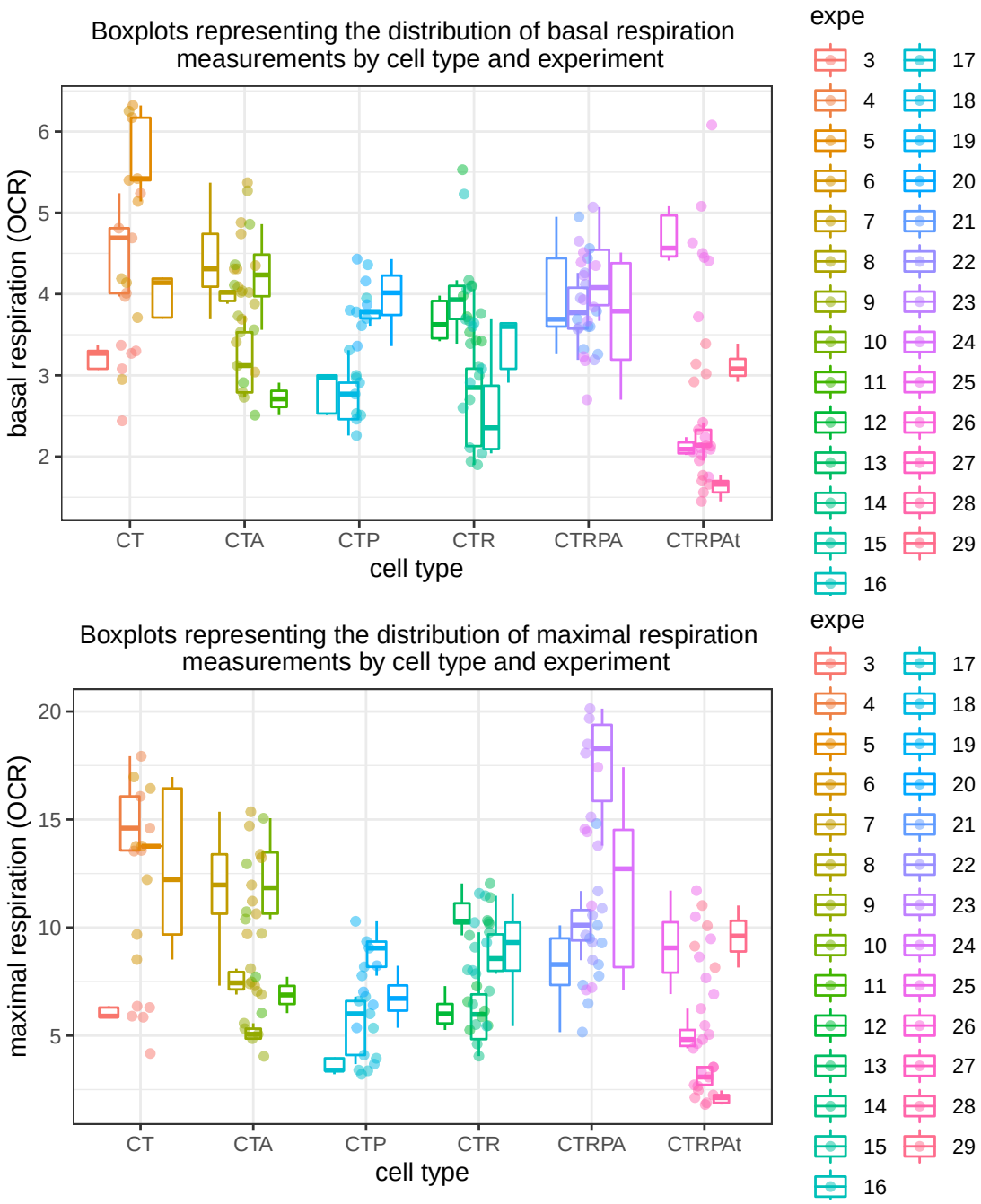


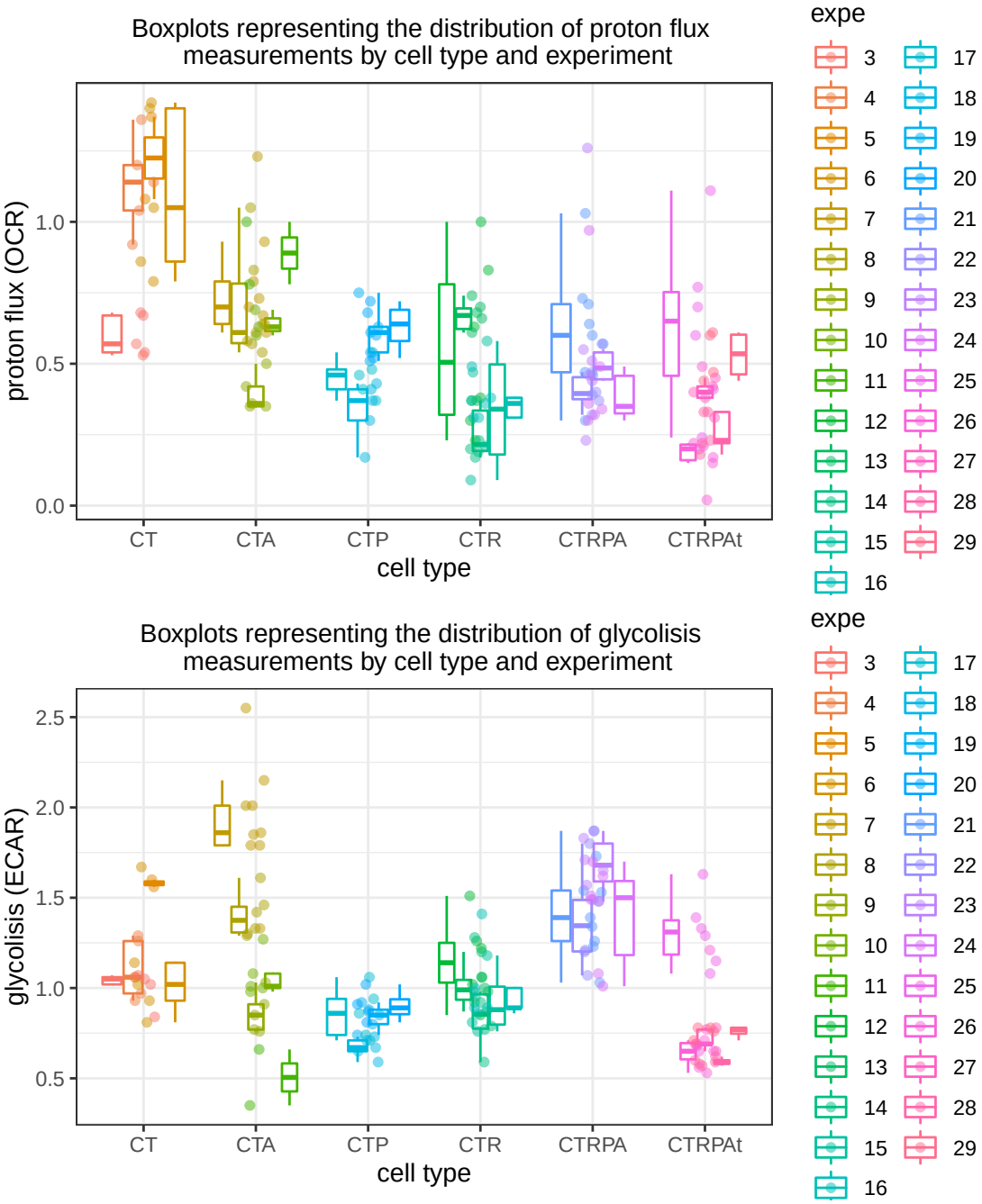


When fixing both the variable and the cell type, the distributions differ given the plate. Besides, the varability between plates does not follow any general pattern for all cell types. We would like to know if these differences are significant. We will perform comparison tests. To do so, we will first determine whether the cell samples are normally distributed or not using Shapiro tests. If they are, we will perform an ANOVA. Otherwise, we will perform the associated non-parametric Kruskal-Wallis test.

Distributions of the respiration variables given the cell type and the experiment

In order to visualize the effects of experiments on the measurements, we also plot the distributions of basalResp, maxResp, protons and glyco given the cell type and the experiment.





We note that these boxplots are identical to the previous ones. We also note that 1 plate corresponds to several experiments, whereas 1 experiment corresponds to 1 single plate. This is perfectly normal (it corresponds to the experimental design we used) and helps us visualizing that the experiment-effects are included in the plate-effects.

2.1.2. Characterization of the distributions with Shapiro tests

basalResp :

```
##            p-value
## CT     0.428657869
## CTA    0.764992975
## CTR    0.219913227
## CTP    0.406564534
## CTRPA  0.854082556
## CTRPAt 0.002230869
```

With a 95% confidence level, we can not reject the assumption that the distributions of "basalRep" for CT, CTA, CTR, CTP, and CTRPA samples are gaussian (associated p-values greater than 5%). However, we reject this assumption for CTRPAt cells (associated p-value less than 5%).

maxResp :

```
##           p-value
## CT     0.09911322
## CTA    0.18901318
## CTR    0.16527733
## CTP    0.30346066
## CTRPA  0.07077479
## CTRPAt 0.02160376
```

With a 95% confidence level, we can not reject the assumption that the distributions of "maxResp" for CT, CTA, CTR, CTP, and CTRPA samples are gaussian (associated p-values greater than 5%). However, we reject this assumption for CTRPAt cells (associated p-values less than 5%).

protons :

```
##           p-value
## CT     0.17958823
## CTA    0.21049317
## CTR    0.15679395
## CTP    0.97738266
## CTRPA  0.00131704
## CTRPAt 0.04282964
```

With a 95% confidence level, we can not reject the assumption that the distributions of "protons" for CT, CTA, CTR, and CTP samples are gaussian (associated p-values greater than 5%). However, we reject this assumption for CTRPA and CTRPAt cells (associated p-values less than 5%).

glyco :

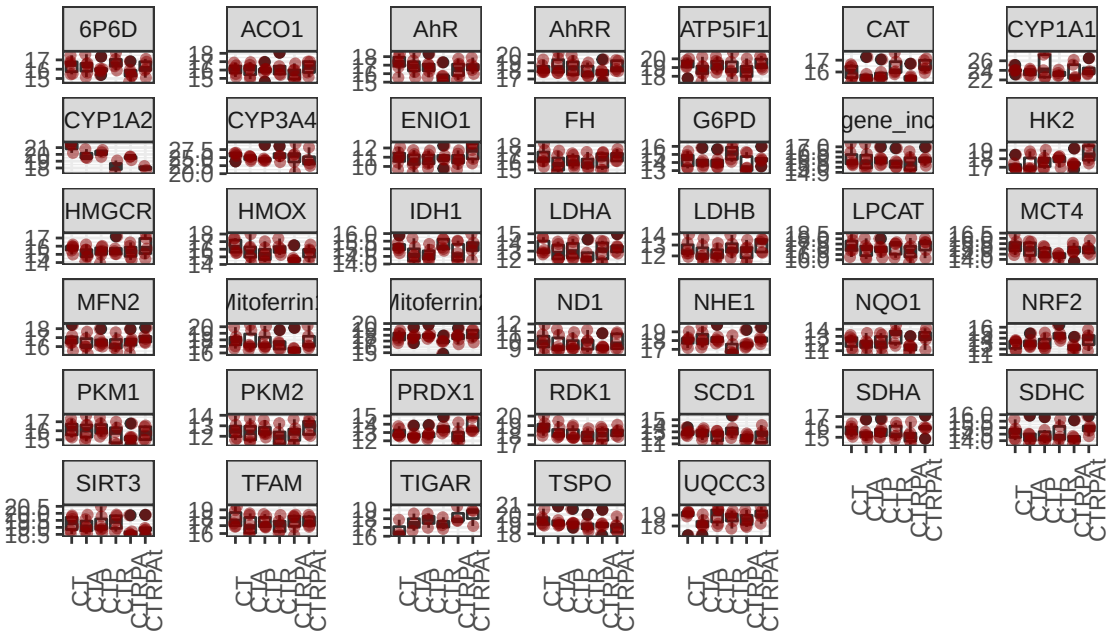
```
##             p-value
## CT     0.0242503686
## CTA    0.7923680980
## CTR    0.3904508021
## CTP    0.9290169490
## CTRPA  0.2222250891
## CTRPAt 0.0001701184
```

With a 95% confidence level, we can not reject the assumption that the distributions of "glyco" for CTA, CTR, CTP, and CTRPA samples are gaussian (associated p-values greater than 5%). However, we reject this assumption for CT and CTRPAt cells (associated p-values less than 5%).

Hence, for each measured variable, there is at least one cell-type for which the Gaussian assumption is rejected. Thereore we cannot perform ANOVA on the cell groups. We perform non-parametric Kruskal-Wallis tests.

2.2. Considering the genomic variables

Vizualisation and characterization of the distributions



For each gene, there is at least one boxplot whose distance between the median and the first quartile is different from the distance between the median and the third quantile. In other words, there is at least one cell sample that is not symetricaly distributed, and thus not normaly distributed. Hence, we will perform the non-parametric Kruskal-Wallis test to analyse differences between the cell samples.

3. Comparison tests

We remind the Kruskal-Wallis test hypothesis :

$$H_0 : \textit{All samples come from the same continuous distribution}$$

$$H_1 : \textit{there are differences between the distributions}$$

.

3.1. Non-parametric analysis

3.1.1. Considering the respiration variables

3.1.1.1. Detection of differences among the treatment samples

```
## [1] "Kruskal-Wallis on cell groups and for the variable 'basalResp'"
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  basalResp by cellLine
## Kruskal-Wallis chi-squared = 28.52, df = 5, p-value = 2.88e-05
```

```
## [1] "Kruskal-Wallis on cell groups and for the variable 'maxResp'"
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  maxResp by cellLine
## Kruskal-Wallis chi-squared = 42.207, df = 5, p-value = 5.348e-08
```

```
## [1] "Kruskal-Wallis on cell groups and for the variable 'protons'"
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  protons by cellLine
## Kruskal-Wallis chi-squared = 45.934, df = 5, p-value = 9.37e-09
```

```
## [1] "Kruskal-Wallis on cell groups and and for the variable 'glycolisis'"
```

```
##
##  Kruskal-Wallis rank sum test
##
## data:  glyco by cellLine
## Kruskal-Wallis chi-squared = 63.659, df = 5, p-value = 2.126e-12
```

All of the Kruskal-Wallis tests lead to extremely small p-values. Therefore, the null hypothesis is always rejected. In other words, there are significant differences between cell groups for each fixed variable. Thus, it is meaningful to apply post-hoc tests.

3.1.1.2. Pairwise comparisons

After a Kruskal-Wallis test, we perform Nemenyi tests with hypothesis :

$$H_0 : \textit{The two samples have similar distributions}$$
$$H_1 : \textit{There are differences between the two distributions}$$

```
##
##  Pairwise comparisons using Nemenyi-test with Chi-squared
##                         approximation for independent samples
##
## data:  basalResp by cellLine
##
##         CT     CTA    CTP    CTR    CTRPA
## CTA     0.9616 -      -      -      -
## CTP     0.1859 0.5978 -      -      -
## CTR     0.1622 0.5910 1.0000 -      -
## CTRPA   0.9711 1.0000 0.5483 0.5356 -
## CTRPAt 0.0016 0.0205 0.8474 0.6961 0.0146
##
## P value adjustment method: none
```

The p-values for the pairs CT-CTRPAt, CTA-CTRPAt, CTRPA-CTRPAt are less than 5%. Therefore we affirm with a 95% confidence level that there are no significant differences between all possibles pairs of cell types except from CT-CTRPAt, CTA-CTRPAt, CTRPA-CTRPAt.

```
##
##  Pairwise comparisons using Nemenyi-test with Chi-squared
##                         approximation for independent samples
##
## data:  maxResp by cellLine
##
##         CT    CTA   CTP   CTR   CTRPA
## CTA     0.902 -     -     -     -
## CTP     0.019 0.179 -     -     -
## CTR     0.472 0.967 0.613 -     -
## CTRPA   1.000 0.710 0.002 0.196 -
## CTRPAt 0.002 0.035 0.999 0.274 4.9e-05
##
## P value adjustment method: none
```

The p-values for the pairs CT-CTP, CT-CTRPAt, CTA-CTRPAt, CTP-CTRPA, CTRPA-CTRPAt are less than 5%. Therefore we affirm with a 95% confidence level that there are no significant differences between all possibles pairs of cell types except from CT-CTP, CT-CTRPAt, CTA-CTRPAt, CTP-CTRPA, CTRPA-CTRPAt.

```
##
##  Pairwise comparisons using Nemenyi-test with Chi-squared
##                         approximation for independent samples
##
## data:  protons by cellLine
##
##         CT      CTA    CTP    CTR    CTRPA
## CTA     0.5466  -      -      -      -
## CTP     0.0142  0.5102 -      -      -
## CTR     7.8e-05 0.0362 0.9450 -      -
## CTRPA   0.0038  0.3411 1.0000 0.9456 -
```

```
## CTRPAt 3.5e-06 0.0041 0.6926 0.9916 0.6585
##
## P value adjustment method: none
```

The p-values for the pairs CT-CTR, CT-CTP, CT-CTRPA, CT-CTRPAt, CTA-CTR, CTA-CTRPAt are less than 5%. Therefore we affirm with a 95% confidence level that there are no significant differences between all possibles pairs of cell types except from CT-CTR, CT-CTP, CT-CTRPA, CT-CTRPAt, CTA-CTR, CTA-CTRPAt.

```
##
##  Pairwise comparisons using Nemenyi-test with Chi-squared
##                        approximation for independent samples
##
## data:  glyco by cellLine
##
##        CT      CTA     CTP     CTR     CTRPA
## CTA    0.97415 -       -       -       -
## CTP    0.09756 0.00198 -       -       -
## CTR    0.89253 0.27559 0.51263 -       -
## CTRPA  0.30044 0.68746 1.3e-06 0.00266 -
## CTRPAt 0.05758 0.00038 1.00000 0.40760 4.6e-08
##
## P value adjustment method: none
```

The p-values for the pairs CTA-CTP, CTA-CTRPAt, CTP-CTRPA, CTR-CTRPA, CTRPA-CTRPAt are less than 5%. Therefore we affirm with a 95% confidence level that there are no significant differences between all possibles pairs of cell types except from CTA-CTP, CTA-CTRPAt, CTP-CTRPA, CTR-CTRPA, CTRPA-CTRPAt.

Conclusion:

	pairs of cell samples with significantly different distributions
basalResp	CT-CTRPAt, CTA-CTRPAt, CTRPA-CTRPAt
maxResp	CT-CTP, CT-CTRPAt, CTA-CTRPAt, CTP-CTRPA, CTRPA-CTRPAt
protons	CT-CTP, CT-CTR, CT-CTRPA, CT-CTRPAt, CTA-CTR, CTA-CTRPAt
glyco	CTA-CTP, CTA-CTRPAt, CTP-CTRPA, CTR-CTRPA, CTRPA-CTRPAt

3.1.2. Considering the genomic variables

3.1.2.1. Detection of differences among the treatment samples

```
## Kruskal-Wallis on cell groups and and for the gene 6P6D :
## p-value = 0.3477539
##
## Kruskal-Wallis on cell groups and and for the gene ACO1 :
## p-value = 0.7085096
##
## Kruskal-Wallis on cell groups and and for the gene AhR :
## p-value = 0.3332004
##
## Kruskal-Wallis on cell groups and and for the gene AhRR :
## p-value = 0.5276
##
## Kruskal-Wallis on cell groups and and for the gene ATP5IF1 :
```

```
## p-value = 0.6112426
##
## Kruskal-Wallis on cell groups and and for the gene CAT :
## p-value = 0.08805606
##
## Kruskal-Wallis on cell groups and and for the gene CYP1A1 :
## p-value = 0.8578327
##
## Kruskal-Wallis on cell groups and and for the gene CYP1A2 :
## p-value = 0.04664244 -> We reject H0
##
## Kruskal-Wallis on cell groups and and for the gene CYP3A4 :
## p-value = 0.7515375
##
## Kruskal-Wallis on cell groups and and for the gene ENIO1 :
## p-value = 0.5881631
##
## Kruskal-Wallis on cell groups and and for the gene FH :
## p-value = 0.5639261
##
## Kruskal-Wallis on cell groups and and for the gene G6PD :
## p-value = 0.276479
##
## Kruskal-Wallis on cell groups and and for the gene gene_inc :
## p-value = 0.8454389
##
## Kruskal-Wallis on cell groups and and for the gene HK2 :
## p-value = 0.1416976
##
## Kruskal-Wallis on cell groups and and for the gene HMGCR :
## p-value = 0.7358828
##
## Kruskal-Wallis on cell groups and and for the gene HMOX :
## p-value = 0.1615693
##
## Kruskal-Wallis on cell groups and and for the gene IDH1 :
## p-value = 0.2451705
##
## Kruskal-Wallis on cell groups and and for the gene LDHA :
## p-value = 0.5398886
##
## Kruskal-Wallis on cell groups and and for the gene LDHB :
## p-value = 0.4105791
##
## Kruskal-Wallis on cell groups and and for the gene LPCAT :
## p-value = 0.9017387
##
## Kruskal-Wallis on cell groups and and for the gene MCT4 :
## p-value = 0.1287255
##
## Kruskal-Wallis on cell groups and and for the gene MFN2 :
## p-value = 0.6847832
##
## Kruskal-Wallis on cell groups and and for the gene Mitoferrin1 :
```

```
## p-value = 0.3081309
##
## Kruskal-Wallis on cell groups and and for the gene Mitoferrin2 :
## p-value = 0.4482336
##
## Kruskal-Wallis on cell groups and and for the gene ND1 :
## p-value = 0.9249324
##
## Kruskal-Wallis on cell groups and and for the gene NHE1 :
## p-value = 0.2228725
##
## Kruskal-Wallis on cell groups and and for the gene NQO1 :
## p-value = 0.2568203
##
## Kruskal-Wallis on cell groups and and for the gene NRF2 :
## p-value = 0.09756824
##
## Kruskal-Wallis on cell groups and and for the gene PKM1 :
## p-value = 0.3709262
##
## Kruskal-Wallis on cell groups and and for the gene PKM2 :
## p-value = 0.6346606
##
## Kruskal-Wallis on cell groups and and for the gene PRDX1 :
## p-value = 0.05249079
##
## Kruskal-Wallis on cell groups and and for the gene RDK1 :
## p-value = 0.2441343
##
## Kruskal-Wallis on cell groups and and for the gene SCD1 :
## p-value = 0.2903639
##
## Kruskal-Wallis on cell groups and and for the gene SDHA :
## p-value = 0.5720092
##
## Kruskal-Wallis on cell groups and and for the gene SDHC :
## p-value = 0.3247357
##
## Kruskal-Wallis on cell groups and and for the gene SIRT3 :
## p-value = 0.3028975
##
## Kruskal-Wallis on cell groups and and for the gene TFAM :
## p-value = 0.8250137
##
## Kruskal-Wallis on cell groups and and for the gene TIGAR :
## p-value = 0.4187416
##
## Kruskal-Wallis on cell groups and and for the gene TSPO :
## p-value = 0.2864442
##
## Kruskal-Wallis on cell groups and and for the gene UQCC3 :
## p-value = 0.2999814
```

With a 95% confidence level, we only rejet H0 for the gene CYP1A2. It means that there are significant

differences between the distributions of the 6 cell samples. In the other cases, the distributions of all cell samples are similar. To get further details about these differences between the cell samples distributions for CYP1A2, we perform Nemenyi post-hoc tests.

3.1.2.2. Pairwise comparisons

After a Kruskal-Wallis test, we perform Nemenyi tests with hypothesis :

$$H_0 : \textit{The two samples have similar distributions}$$
$$H_1 : \textit{There are differences between the two distributions}$$

We remind the test hypothesis: H0: "The two distributions are similar", versus H1: "There are differences between the two distributions".

```
##
##  Pairwise comparisons using Tukey and Kramer (Nemenyi) test
##                    with Tukey-Dist approximation for independent samples
##
## data:  value by cellLine
##
##        CT    CTA   CTP   CTR   CTRPA
## CTA    0.793 -     -     -     -
## CTP    0.972 0.996 -     -     -
## CTR    0.086 0.667 0.368 -     -
## CTRPA  0.636 1.000 0.972 0.804 -
## CTRPAt 0.086 0.667 0.368 1.000 0.804
##
## P value adjustment method: none
```

When considering a 95% confidence level, there is no significant differences between any pairs of cell-samples. Nevertheless, we could admit with a 90% confidence level that both the pairs CTR-CT and CTRPAt-CT have significantly different distributions.

Conclusion:

For each gene except from CYP1A2, the distributions of the cell samples are quite similar. When allowing a 10% margin of error, the pairs CTR-CT and CTRPAt-CT can be considered as having significantly different distributions for the CYP1A2-gene expression.

3.2. Parametric analysis

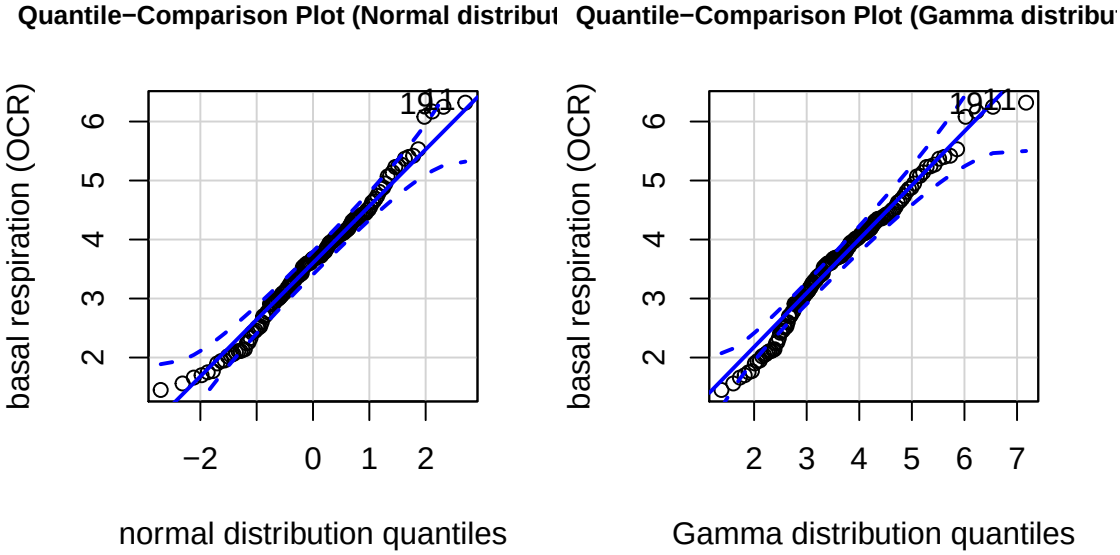
We have seen that the values of the respiration variables strongly depend on the plate. For that reason, we choose to approach our respiration data with models that will actually take this variability into account throughout the analysis. The results will be more precise (since we consider some random effects that were ignored in the non-parametric analysis), but conditionned to the goodness-of-fit of the distribution probability assumed by the corresponding model.

A mixed model estimates the effects of explanatory variables on a response variable. It has both fixed and random effects. The so-called "fixed effects" are the effects of variables with finite level sets, whereas the "random effects" are the effects of variables whose levels are randomly selected among a larger population. In our case, the fixed effects are those of `cellLine` and the random effects are those of `plate`.

3.2.1. Probability distribution best fitting the data

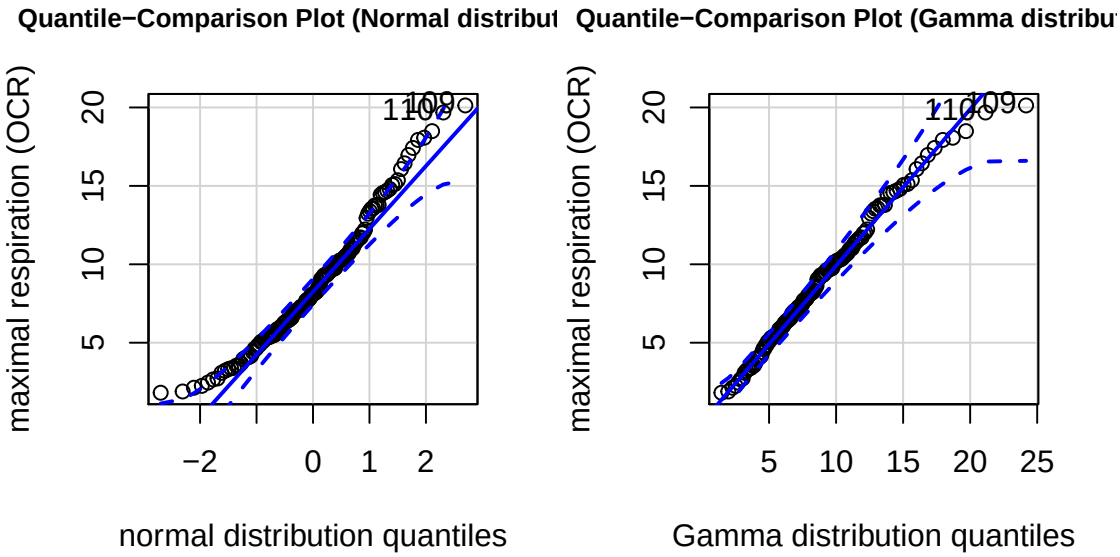
```
## [1] 11 19
```

```
## [1] 11 19
```



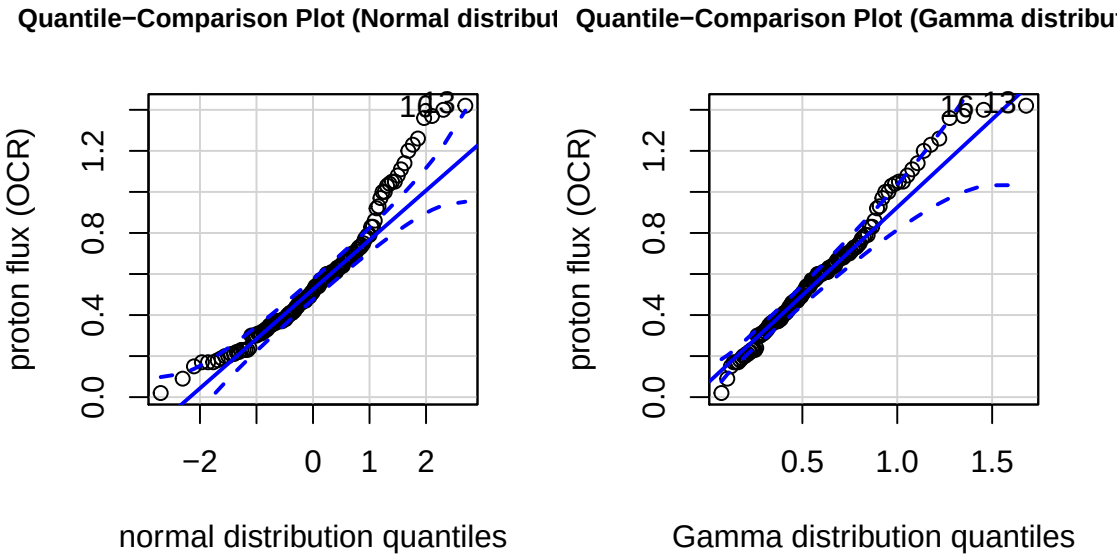

```
## [1] 109 110
```

```
## [1] 109 110
```



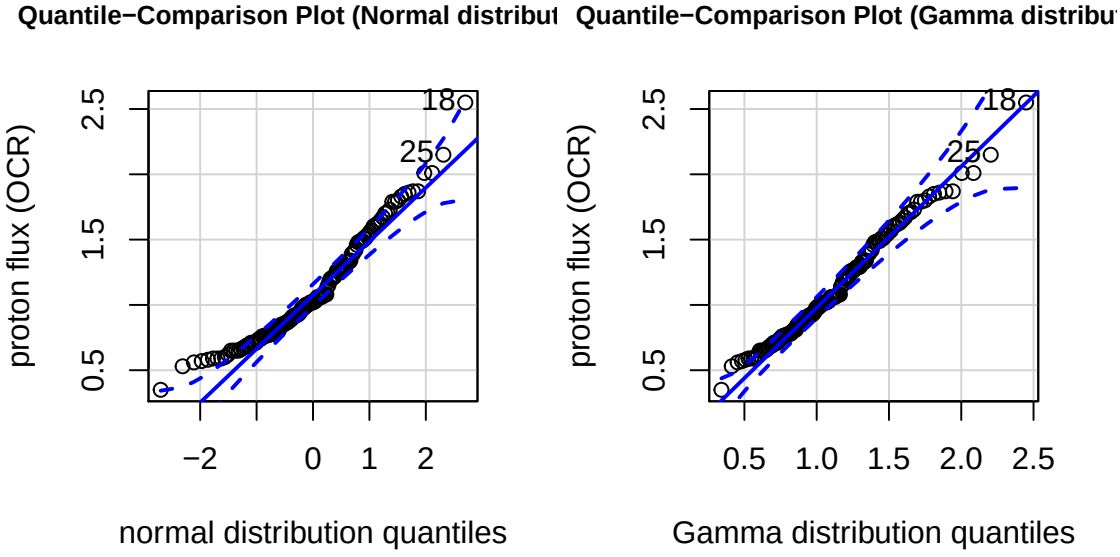

```
## [1] 13 16
```

```
## [1] 13 16
```




```
## [1] 18 25
```

```
## [1] 18 25
```



In the previous plots, the solid blue line represents the perfect distribution fit, the dashed blue lines are the bounds of the confidence intervals of the perfect distribution fit. We added scatterplots of our data. The distribution to be chosen must be the one including the highest amount of points between the dashed lines. Therefore, the normal distribution seems to be the one best fitting the data when considering basalResp. For maxResp, protons, and glyco, we should rather choose the Gamma distribution.

3.2.2. Mixed model fit

When the data is not normally distributed, we can fit the data using the so-called Penalized Quasilikelihood (PQL). This technique produces biased estimates if the response variable fits a discrete count distribution, like Poisson or binomial, and the mean is less than 5 - or if the response variable is binary. We are not under any of these conditions so we can use PQL without producing bias.

We have previously seen that for all observed variables, there was at least one cell-type sample for which the

Gaussian assumption was rejected. Hence, for all observed variables, we will not consider our data to be normally distributed.

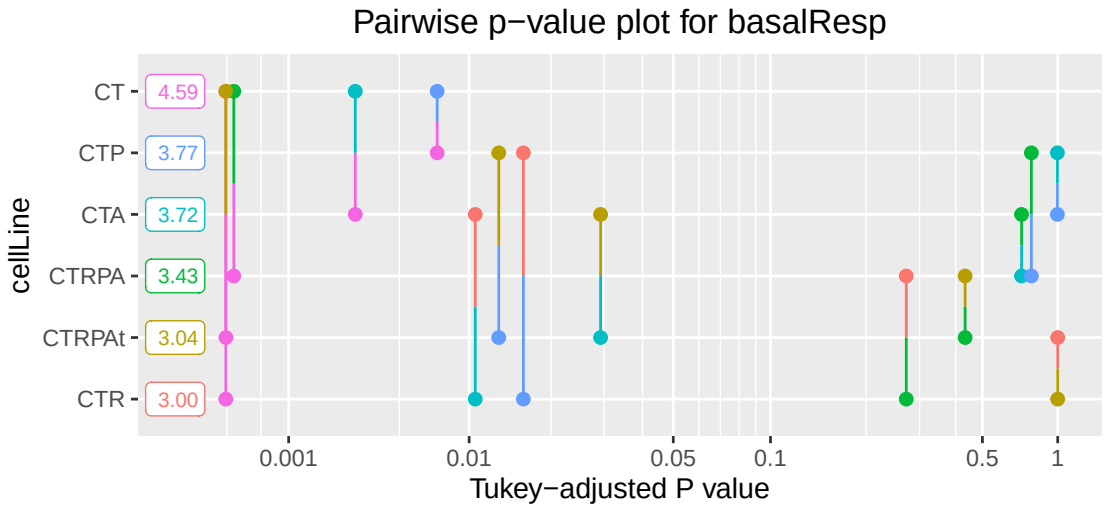
3.2.3. Comparison of multiple independent cell samples

basalResp

```
## Analysis of Deviance Table (Type III tests)
##
## Response: zz
##               Chisq Df Pr(>Chisq)
## (Intercept) 243.640  1  < 2.2e-16 ***
## cellLine     61.339  5  6.426e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The resulting p-value being extremely small, we reject H_0 with a 95% confidence level. Hence, we can affirm, with a 5% margin of error, that the value of basalResp depends on the cell type. It is then meaningful to perform post-hoc pairwise comparisons.

```
##  contrast       estimate    SE  df t.ratio p.value
##  CT - CTA         0.8722 0.229 132  3.801  0.0029
##  CT - CTP         0.8255 0.234 132  3.531  0.0074
##  CT - CTR         1.5857 0.232 132  6.846  <.0001
##  CT - CTRPA       1.1597 0.254 132  4.565  0.0002
##  CT - CTRPAt      1.5541 0.232 132  6.702  <.0001
##  CTA - CTP       -0.0467 0.239 132 -0.196  1.0000
##  CTA - CTR        0.7136 0.209 132  3.420  0.0106
##  CTA - CTRPA      0.2876 0.201 132  1.431  0.7081
##  CTA - CTRPAt     0.6819 0.221 132  3.084  0.0293
##  CTP - CTR        0.7602 0.231 132  3.288  0.0160
##  CTP - CTRPA      0.3343 0.254 132  1.316  0.7760
##  CTP - CTRPAt     0.7286 0.217 132  3.355  0.0130
##  CTR - CTRPA     -0.4260 0.198 132 -2.147  0.2702
##  CTR - CTRPAt    -0.0317 0.197 132 -0.160  1.0000
##  CTRPA - CTRPAt   0.3943 0.212 132  1.858  0.4325
##
## P value adjustment: tukey method for comparing a family of 6 estimates
```



On the previous graph, each vertical line segment represents a comparison. The ends of each segment correspond to the two Estimated Marginal Means (EMM) being compared. The horizontal position of the segment depends on the p-value of the pairwise comparison.

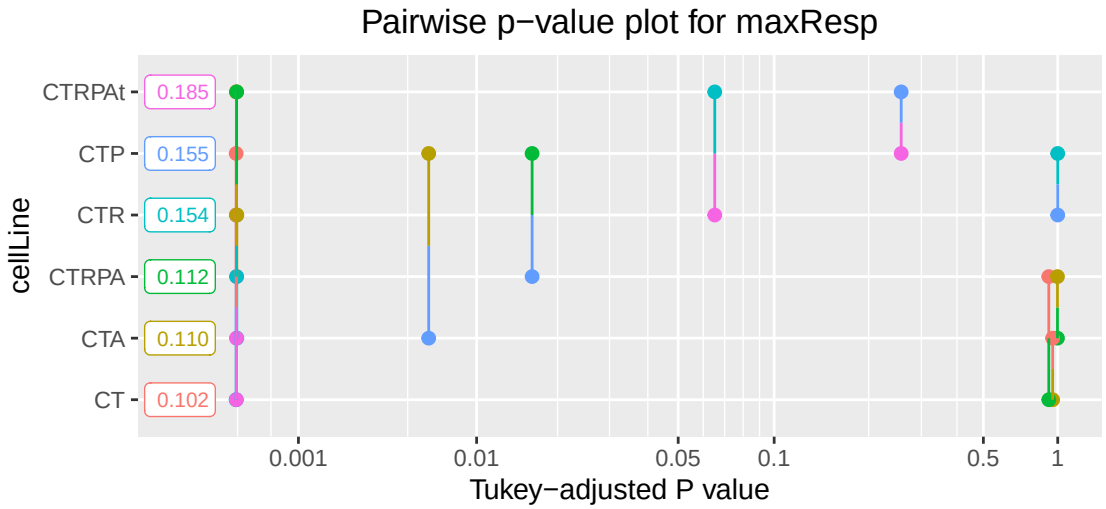
With a 95% confidence level, we affirm that there are signficant differences between the distributions of basalResp for CT and CTA, CT and CTP, CT and CTR, CT and CTRPA, CT and CTRPAt, CTA and CTR, CTA and CTRPAt, CTP and CTR, CTP and CTRPAt (associated p-values less than 5%).

maxResp

```
## Analysis of Deviance Table (Type III tests)
##
## Response: zz
##               Chisq Df Pr(>Chisq)
## (Intercept)  32.343  1  1.292e-08 ***
## cellLine    103.117  5  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The resulting p-value being extremely small, we reject H_0 with a 95% confidence level. Hence, we can affirm, with a 5% margin of error, that the value of maxResp depends on the cell type. It is then meaningful to perform post-hoc pairwise comparisons.

```
##  contrast        estimate      SE  df t.ratio p.value
##  CT - CTA       -0.008708 0.00999 129 -0.872  0.9525
##  CT - CTP       -0.053302 0.01127 129 -4.728  0.0001
##  CT - CTR       -0.052786 0.01018 129 -5.184  <.0001
##  CT - CTRPA     -0.010752 0.01072 129 -1.002  0.9164
##  CT - CTRPAt    -0.083507 0.01202 129 -6.946  <.0001
##  CTA - CTP      -0.044594 0.01245 129 -3.583  0.0063
##  CTA - CTR      -0.044078 0.00936 129 -4.709  0.0001
##  CTA - CTRPA    -0.002044 0.00838 129 -0.244  0.9999
##  CTA - CTRPAt   -0.074799 0.01175 129 -6.364  <.0001
##  CTP - CTR       0.000516 0.01254 129  0.041  1.0000
##  CTP - CTRPA     0.042550 0.01296 129  3.284  0.0163
##  CTP - CTRPAt   -0.030205 0.01388 129 -2.176  0.2562
##  CTR - CTRPA     0.042034 0.00809 129  5.198  <.0001
##  CTR - CTRPAt   -0.030721 0.01100 129 -2.792  0.0652
##  CTRPA - CTRPAt -0.072756 0.01032 129 -7.047  <.0001
##
## P value adjustment: tukey method for comparing a family of 6 estimates
```



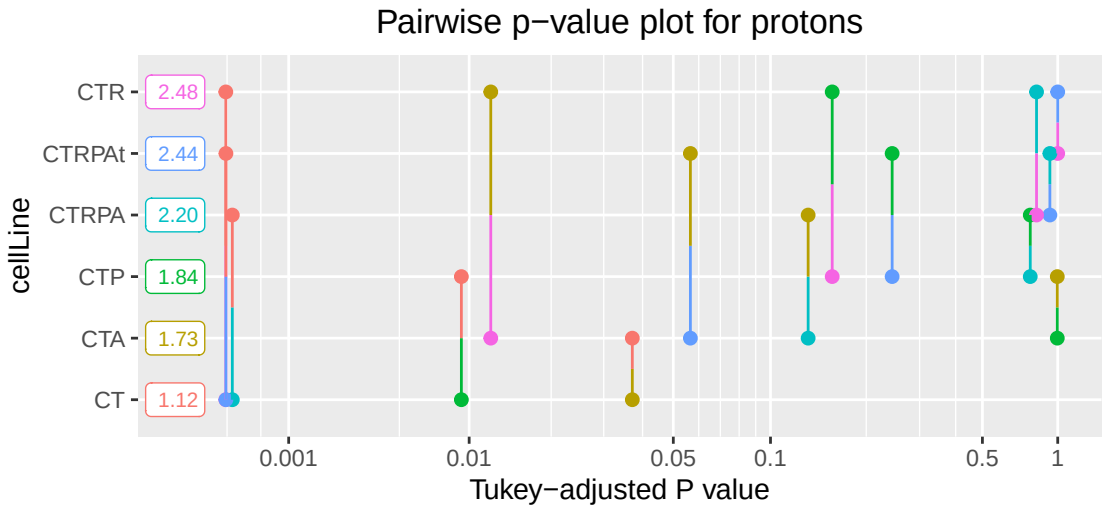
With a 95% confidence level, we affirm that there are signficant differences between the distributions of maxResp for CT and CTP, CT and CTR, CT and CTRPAt, CTA and CTP, CTA and CTR, CTA and CTRPAt, CTP and CTRPA, CTR and CTRPA, CTRPA and CTRPAt (associated p-values less than 5%).

protons

```
## Analysis of Deviance Table (Type III tests)
##
## Response: zz
##              Chisq Df Pr(>Chisq)
## (Intercept) 24.849  1  6.201e-07 ***
## cellLine    50.790  5  9.549e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The resulting p-value being extremely small, we reject H_0 with a 95% confidence level. Hence, we can affirm, with a 5% margin of error, that the value of protons depends on the cell type. It is then meaningful to perform post-hoc pairwise comparisons.

```
##  contrast       estimate    SE  df t.ratio p.value
##  CT - CTA        -0.6097 0.203 128 -3.003  0.0371
##  CT - CTP        -0.7235 0.209 128 -3.462  0.0093
##  CT - CTR        -1.3639 0.229 128 -5.954  <.0001
##  CT - CTRPA      -1.0852 0.236 128 -4.594  0.0001
##  CT - CTRPAt     -1.3231 0.243 128 -5.448  <.0001
##  CTA - CTP       -0.1138 0.247 128 -0.461  0.9973
##  CTA - CTR       -0.7542 0.223 128 -3.379  0.0121
##  CTA - CTRPA     -0.4756 0.190 128 -2.503  0.1309
##  CTA - CTRPAt    -0.7134 0.251 128 -2.847  0.0565
##  CTP - CTR       -0.6404 0.264 128 -2.424  0.1557
##  CTP - CTRPA     -0.3618 0.272 128 -1.331  0.7670
##  CTP - CTRPAt    -0.5996 0.272 128 -2.205  0.2428
##  CTR - CTRPA      0.2786 0.224 128  1.244  0.8141
##  CTR - CTRPAt     0.0408 0.250 128  0.163  1.0000
##  CTRPA - CTRPAt  -0.2379 0.245 128 -0.970  0.9265
##
## P value adjustment: tukey method for comparing a family of 6 estimates
```



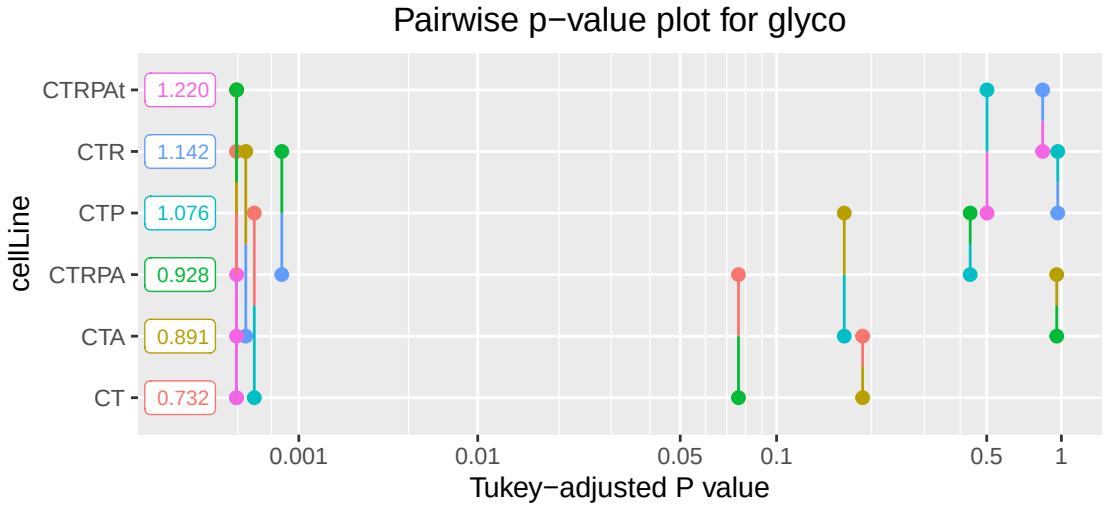
With a 95% confidence level, we affirm that there are signficant differences between the distributions of protons for CT and CTA, CT and CTP, CT and CTR, CT and CTRPA, CT and CTRPAt, CTA and CTR (associated p-values less than 5%).

glyco

```
## Analysis of Deviance Table (Type III tests)
##
## Response: zz
##              Chisq Df Pr(>Chisq)
## (Intercept) 58.454  1  2.081e-14 ***
## cellLine    70.056  5  9.977e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The resulting p-value being extremely small, we reject H_0 with a 95% confidence level. Hence, we can affirm, with a 5% margin of error, that the value of glyco depends on the cell type. It is then meaningful to perform post-hoc pairwise comparisons.

```
##  contrast       estimate     SE  df t.ratio p.value
##  CT - CTA        -0.1588 0.0680 130 -2.337  0.1871
##  CT - CTP        -0.3442 0.0775 130 -4.440  0.0003
##  CT - CTR        -0.4103 0.0725 130 -5.658  <.0001
##  CT - CTRPA      -0.1961 0.0718 130 -2.731  0.0761
##  CT - CTRPAt     -0.4878 0.0763 130 -6.392  <.0001
##  CTA - CTP       -0.1854 0.0772 130 -2.402  0.1633
##  CTA - CTR       -0.2515 0.0553 130 -4.548  0.0002
##  CTA - CTRPA     -0.0373 0.0432 130 -0.863  0.9545
##  CTA - CTRPAt    -0.3290 0.0655 130 -5.024  <.0001
##  CTP - CTR       -0.0660 0.0802 130 -0.823  0.9628
##  CTP - CTRPA      0.1481 0.0799 130  1.855  0.4347
##  CTP - CTRPAt    -0.1435 0.0820 130 -1.751  0.5013
##  CTR - CTRPA      0.2141 0.0509 130  4.205  0.0007
##  CTR - CTRPAt    -0.0775 0.0640 130 -1.210  0.8312
##  CTRPA - CTRPAt  -0.2917 0.0603 130 -4.835  0.0001
##
## P value adjustment: tukey method for comparing a family of 6 estimates
```



With a 95% confidence level, we affirm that there are signficant differences between the distributions of glyco for CT and CTP, CT and CTR, CT and CTRPAt, CTA and CTR, CTA and CTRPAt, CTR and CTRPA, CTRPA and CTRPAt (associated p-values less than 5%).

Conclusion

Differences between the distributions of respiration variables under basal conditions and using mixed models :

	Pairs of cell samples with significantly different distributions
basalResp	CT-CTA, CT-CTP, CT-CTR, CT-CTRPA, CT-CTRPAt, CTA-CTR, CTA-CTRPAt, CTP-CTR, CTP-CTRPAt
maxResp	CT-CTP, CT-CTR, CT-CTRPAt, CTA-CTP, CTA-CTR, CTA-CTRPAt, CTP-CTRPA, CTR-CTRPA, CTRPA-CTRPAt
protons	CT-CTA, CT-CTP, CT-CTR, CT-CTRPA, CT-CTRPAt, CTA-CTR
glyco	CT-CTP, CT-CTR, CT-CTRPAt, CTA-CTR, CTA-CTRPAt, CTR-CTRPA, CTRPA-CTRPAt

Differences between the distributions of respiration variables under basal conditions and without mixed models (Recall) :

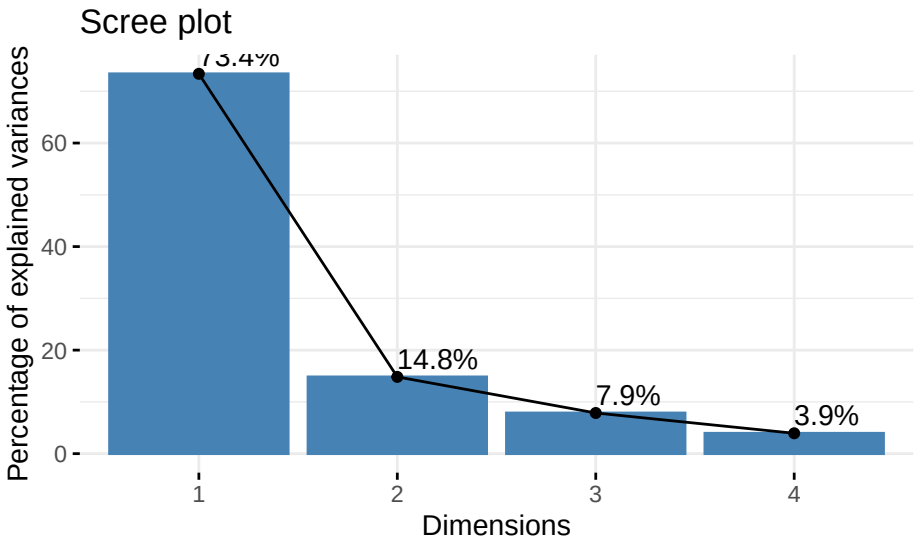
	Pairs of cell samples with significantly different distributions
basalResp	CT-CTRPAt, CTA-CTRPAt, CTRPA-CTRPAt
maxResp	CT-CTP, CT-CTRPAt, CTA-CTRPAt, CTP-CTRPA, CTRPA-CTRPAt
protons	CT-CTP, CT-CTR, CT-CTRPA, CT-CTRPAt, CTA-CTR, CTA-CTRPAt
glyco	CTA-CTP, CTA-CTRPAt, CTP-CTRPA, CTR-CTRPA, CTRPA-CTRPAt

4. Multivariate Analysis

4.1. Principal Component Analysis

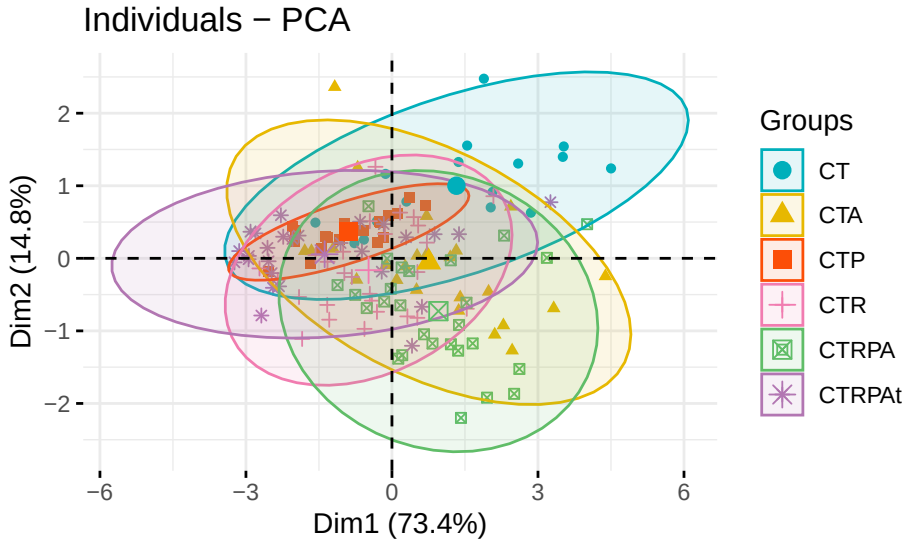
4.1.1. Considering the respiration variables

Scree plot :

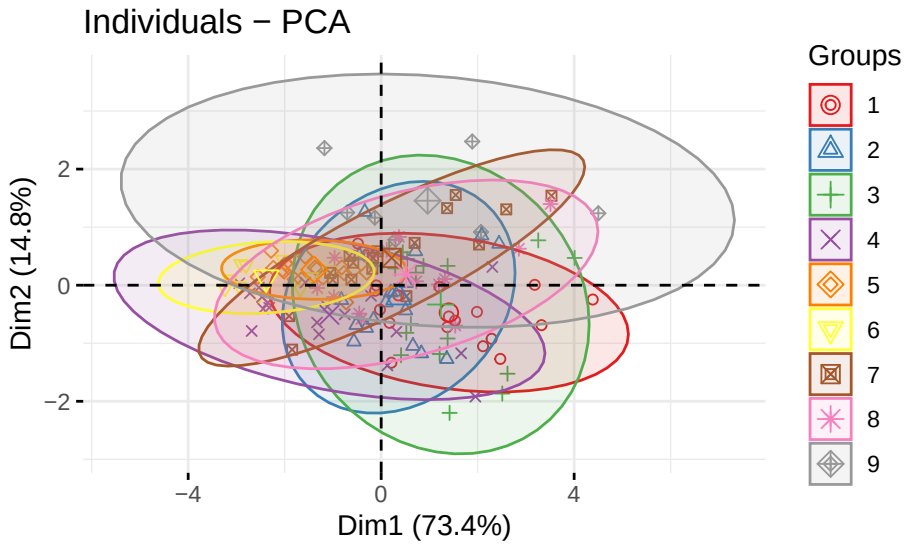


The two first principal components contain most of the explained variance (88.2%). Therefore, a representation of the data in the 2-dimensional space made of these 2 components still reflects the reality.

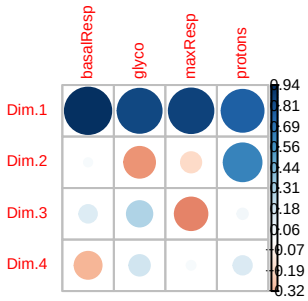
Graph of individuals colored by cell type :



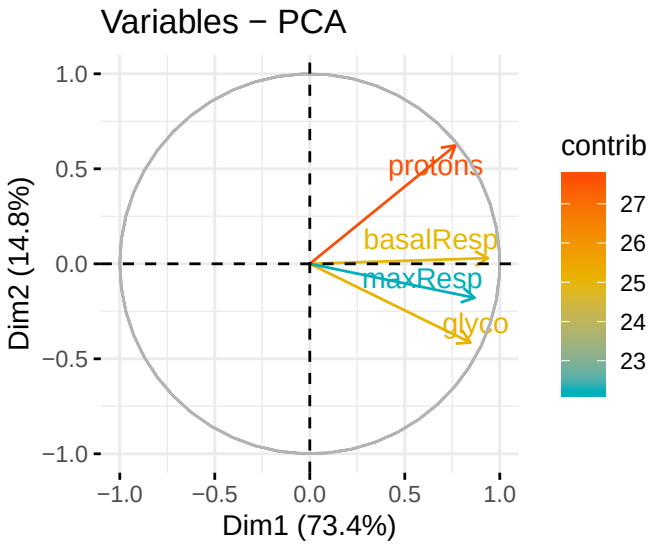
Graph of individuals colored by plate :



Correlations of variables :



Graph of variables :

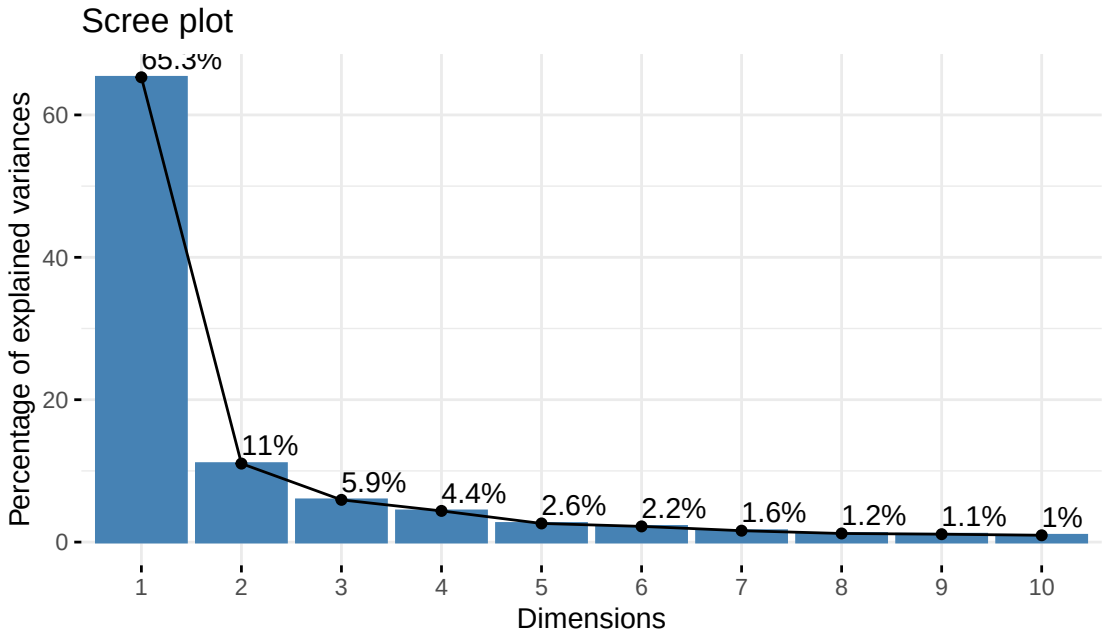


All variables are on the same side of the circle and positively correlated: There are size effects.

4.1.2. Considering the genomic variables

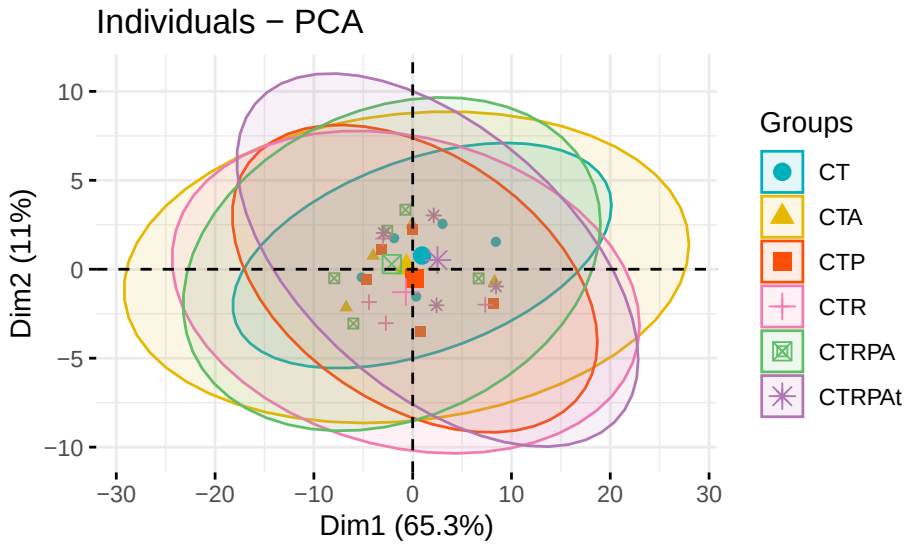
We perform the PCA on the largest complete genomic data set under basal conditions. All of the genes can be simultaneously observed, except from "CYP1A2" and "TIGAR". For this reason, we do not include any of these genes in our analysis.

Scree plot :

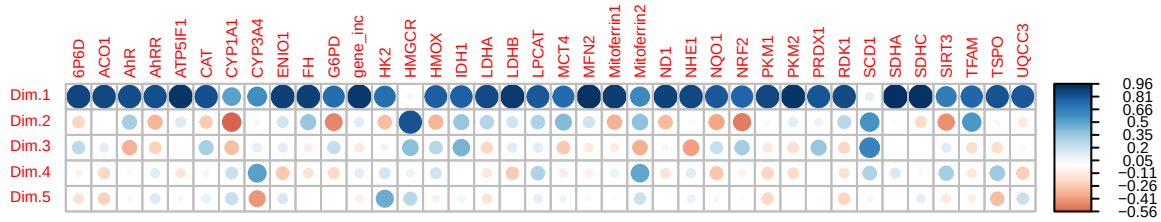


The two first principal components contain most of the explained variance (76.3%). Therefore, a representation of the data in the 2-dimensional space made of these 2 components still reflects the reality.

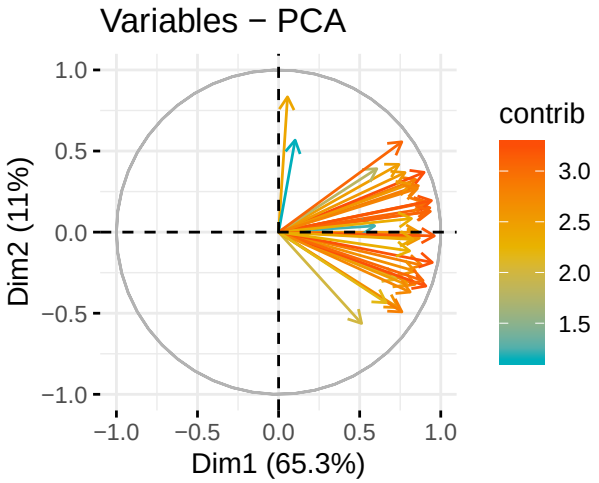
Graph of individuals colored by cell type :



Correlations of variables :



Graph of variables :



All variables are on the same side of the circle and positively correlated: There are size effects.

Conclusion of the PCAs :

For respiration as for genomic variables :

- The cell type is not the main factor
- The main factor seems to be a size effect (ie, certain cells having globally a larger expression than others for all cells)

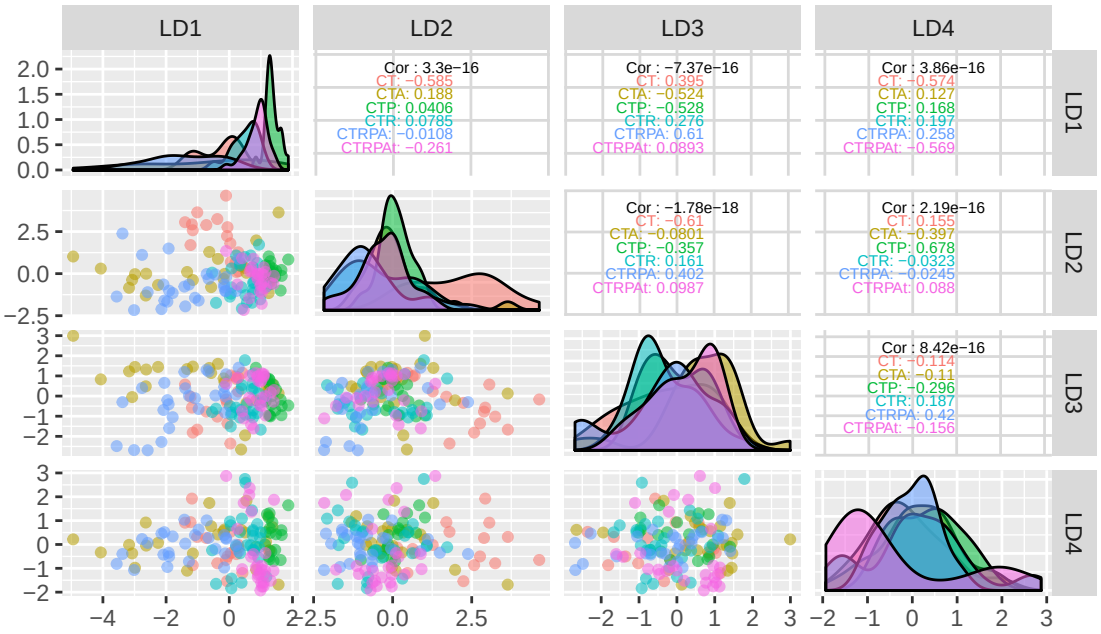
4.2. Factorial Discriminant Analysis

The objective of the FDA (also called Linear Discriminant Analysis - LDA) is to find the linear combination of the observed quantitative variables that gives the best possible separation between the cell groups. To perform a FDA, the data need to be made of one qualitative variable divided into classes, and several quantitative variables characterizing the qualitative one.

4.2.1. Considering the respiration variables

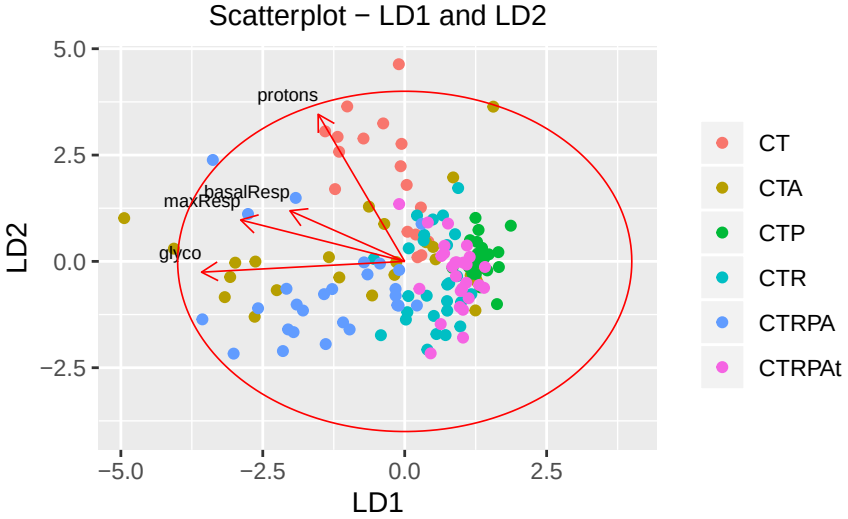
We perform the FDA on the largest complete respiration data set under basal conditions.

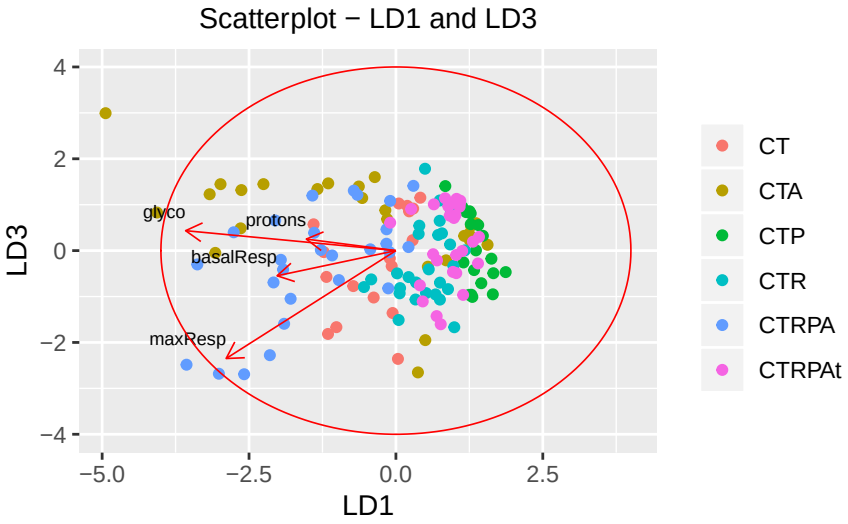
Scatterplot matrix and correlation circles :

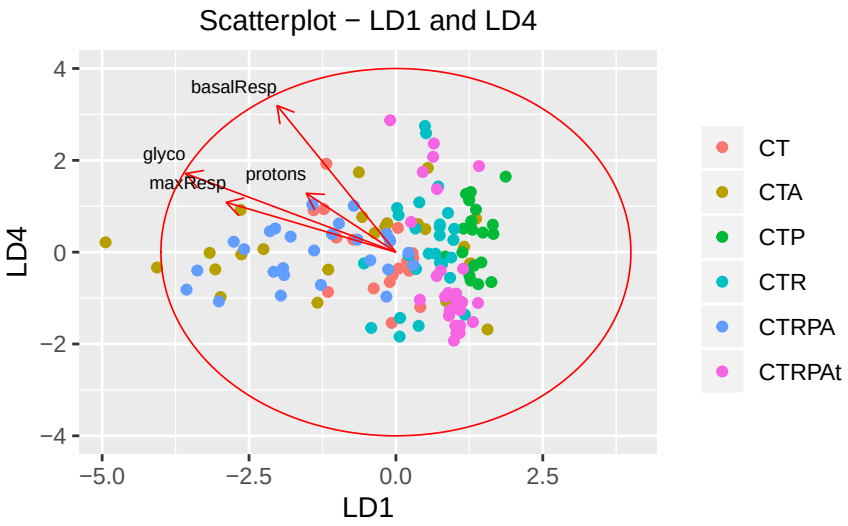


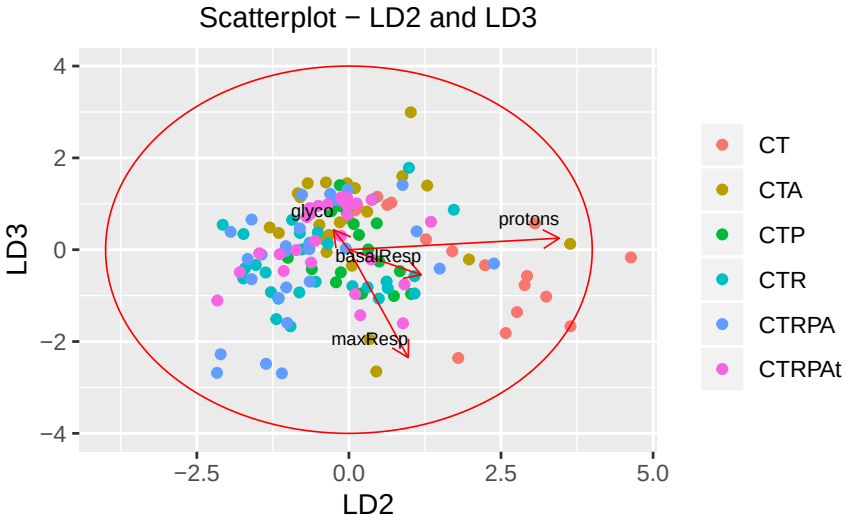
There is no configuration where all cell groups are distinctly separated. However, the first discriminant function separates quite correctly the groups CTA and CTRPA from the other groups (the separation of these cell groups along the x axis on the first column of the scatterplot matrix is acceptable), and the second discriminant function separates correctly CT from the others.

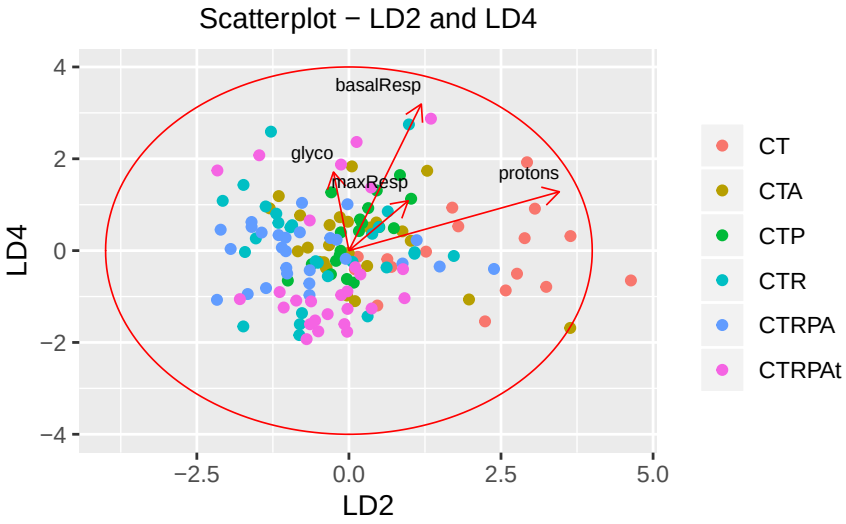
We now plot the correlations between the respiration variables and the discriminant functions. This way, we may be able to characterize some of the previously discriminated cells.




Scatterplot − LD1 and LD3
LD3
LD1
glyco
protons
basalResp
maxResp
CT
CTA
CTP
CTR
CTRPA
CTRPAt


Scatterplot − LD1 and LD4
LD4
LD1
basalResp
glyco
maxResp
protons
CT
CTA
CTP
CTR
CTRPA
CTRPAt


Scatterplot − LD2 and LD3
LD3
LD2
glyco
protons
basalResp
maxResp
CT
CTA
CTP
CTR
CTRPA
CTRPAt



Conclusion of the FDA:

- CT cells : Particularly high values of protons.
- CTA cells : Particularly high values of basalResp, maxResp, protons and glyco.
- CTRPA cells : Particularly high values of basalResp, maxResp, protons and glyco.

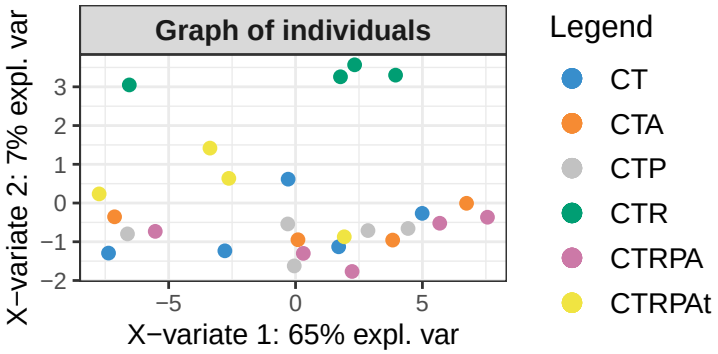
4.2.2. Considering the genomic variables

We perform the FDA on the largest complete genomic data set under basal conditions.

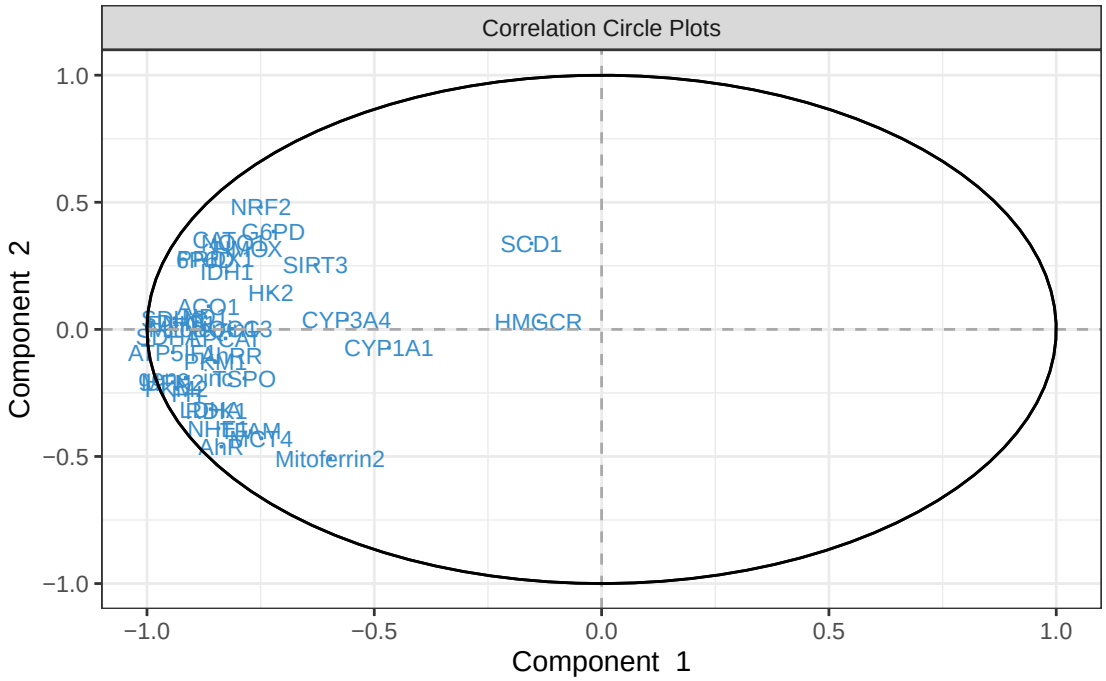
Class summary :

```
##      CT     CTA     CTP     CTR  CTRPA CTRPAt
##       5       4       5       4       5       4
```

Graphs of individuals and variables in the space of components 1 and 2 (72% of the explained variance) :



The second component seems to separate the CTR cells from the others (discriminant ++).

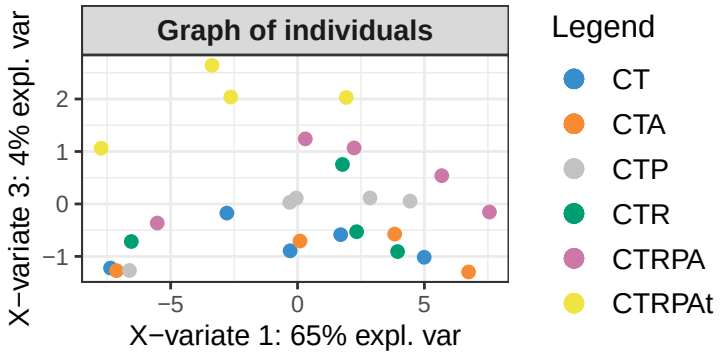


Given that the CTR-individuals have very high y-coordinates on the graph of individuals, we know that any variable with high y-coordinate will take particularly high values for CTR cells. We need to fix a minimum threshold. We take the y-coordinate of SCD1.\ Variables having higher y-coordinates than the y-coordinate of SCD1 :

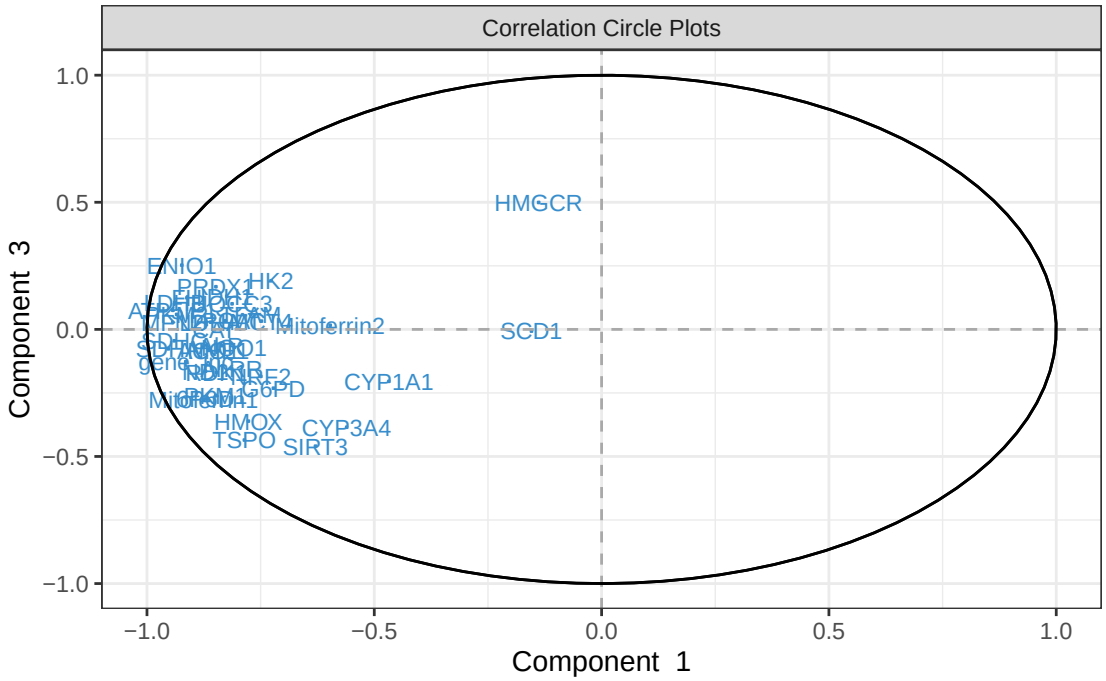
```
## [1] "CAT"  "G6PD" "NQO1" "NRF2"
```

Therefore, the CTR group has particularly high values of CAT, G6PD, NQO1, NRF2 and SCD1.

Graphs of individuals and variables in the space components 1 and 3 (69% the of explained variance):



The third component seems to separate the CTRPAt cells from the others (discriminant +).

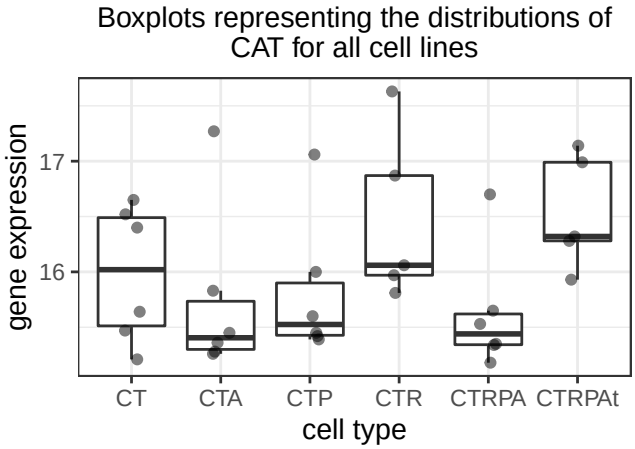


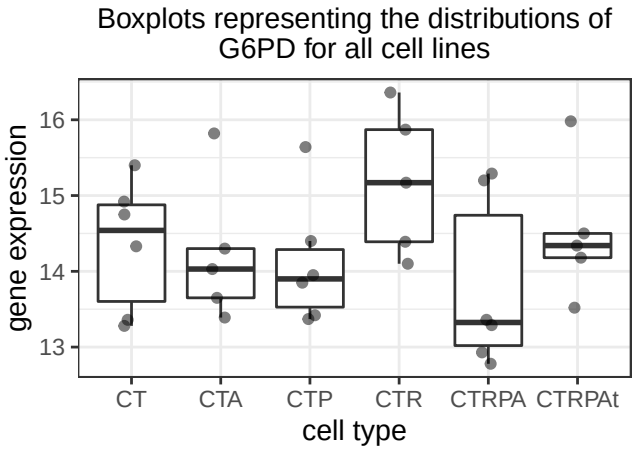
The CTRPAt group has particularly high values of HMGCR.

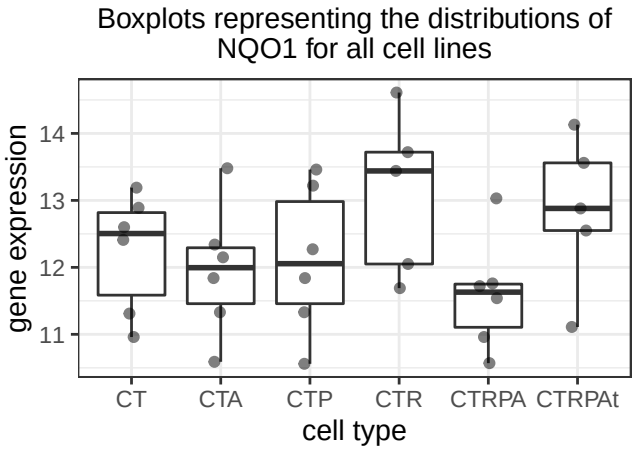
Conclusion of the PLS-DA:

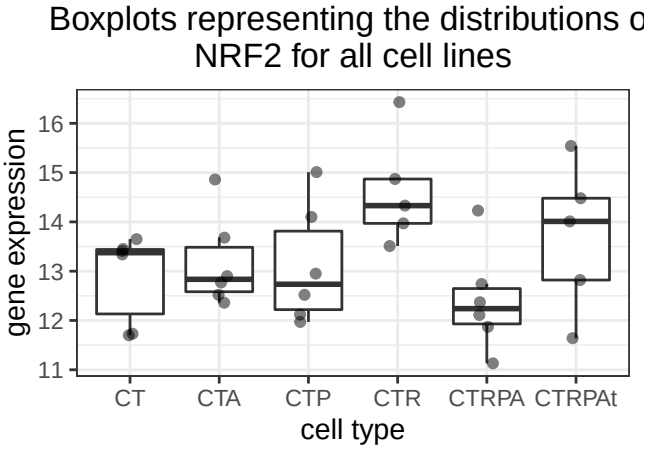
- On the previous graphs of individuals, the cell groups are not clearly separated from one another. There is no component that discriminates all cell groups correctly.
- The graphs of variables show that the first component axis is a size effect axis. It does not help us characterizing cells.
- We get the following characterizations:
  - CAT, G6PD, NQO1, NRF2 and SCD1 take particularly high values for CTR cells
  - HMGCR take particularly high values for CTRPAt cells

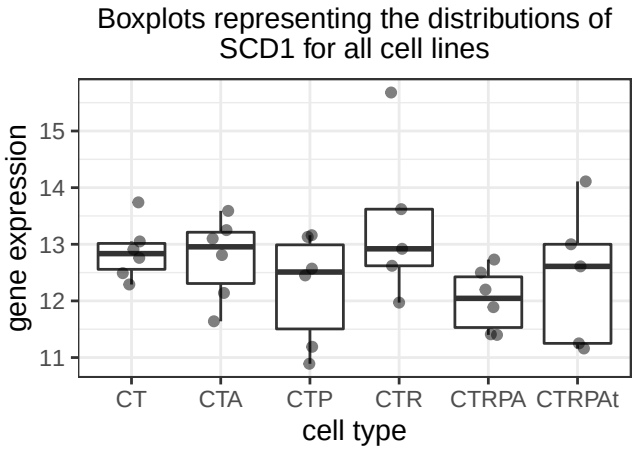
We check these assertions by plotting the distributions of CAT, G6PD, NQO1, NRF2, SCD1 and HMGCR for all cell lines.

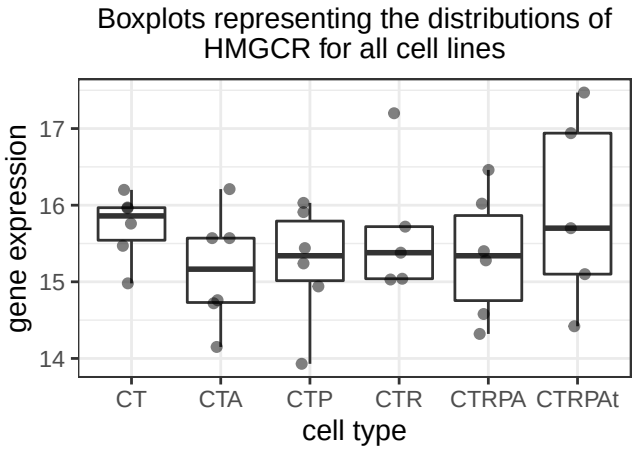



Boxplots representing the distributions of
G6PD for all cell lines
gene expression
16
15
14
13
CT
CTA
CTP
CTR
CTRPA
CTRPAt
cell type


Boxplots representing the distributions of
NQO1 for all cell lines
gene expression
14
13
12
11
CT
CTA
CTP
CTR
CTRPA
CTRPAt
cell type


Boxplots representing the distributions o
NRF2 for all cell lines
gene expression
16
15
14
13
12
11
CT
CTA
CTP
CTR
CTRPA
CTRPAt
cell type





The boxplots confirm that :

- compared to the other cell lines, CTR has particularly high values of CAT, G6PD, NQO1, NRF2 ans SCD1
- compared to the other cell lines, CTRPAt has particularly high values of HMGCR.

5. Conclusion

- No differences between genes (almost)
- The PCAs show that the cell type is not the main factor driving the differences between expressions, and the main factor seems to be a size effect.
- The FDAs allow us to discriminate and characterize some of the cell types but the high variability of data between the plates still makes the discrimination of the cell types difficult.