INSTITUT de MATHEMATIQUES de TOULOUSE

UNIVERSITÉ TOULOUSE III PAUL SABATIER — Université de Toulouse

Rapport de stage

Master 2 MAPI^3

Intégration de données omiques pour l'étude de l'impact des contaminants alimentaires sur le métabolisme et le développement de cancers

INRAE

la science pour la vie, l'humain, la terre

24 Chemin Borde Rouge 31320 Auzeville Tolosane

Tutrice : Nathalie VIALANEIX

Professeur encadrant : Jean-Michel LOUBES

CALS Mallory

23 Mars 2020 — 31 Août 2020

Table des matières

Remerciements

Ces cinq mois de stage au sein d'INRAE Toulouse ont été une expérience très enrichissante et m'ont permis de découvrir le monde du travail.

Je tiens à remercier ici toutes les personnes qui ont contribué au bon déroulement du stage et m'ont aidé lors de la rédaction de ce mémoire.

En premier lieu, je tiens à remercier vivement mon maître de stage, Mme Nathalie VIALANEIX, Directrice de Recherche en statistique dans l'unité de Mathématiques et Informatique Appliquées de Toulouse, INRAE Toulouse, pour son accueil au sein de son unité, pour sa disponibilité, pour son accompagnement et pour son aide précieuse.

Je tiens également à remercier M. DEJEAN S. et M. RENGEL D. pour leur disponibilité et leurs explications mathématiques et biologiques tout au long de ce stage.

Je tiens à remercier toute l'équipe du projet METAhCOL et l'ensemble de l'équipe MIAT, pour avoir pris le temps de répondre à mes questions et pour leur accueil chaleureux pendant ces 5 mois.
Enfin, je remercie ma famille ainsi que les personnes qui m'ont conseillé et relu lors de la rédaction de ce rapport.

Introduction

La France fait partie des pays pour lesquels le risque de cancer colorectal est élevé, comme dans les autres pays d'Europe de l'Ouest, les États-Unis, l'Australie et, plus récemment, le Japon. En France, le cancer colorectal se situe, hommes et femmes confondus, au troisième rang des cancers les plus fréquents. Il survient en grande majorité chez les personnes âgées de 50 ans et plus. D'après les estimations de différentes études, le nombre de cancers colorectaux devrait augmenter dans les prochaines années pour atteindre 45 000 nouveaux cas annuels contre 43 336 cas détectés en 2018[1].

Le cancer colorectal fait le plus souvent suite à une tumeur bénigne qui finit par devenir cancéreuse. Le pronostic de guérison dépend largement du stade au diagnostic. Les chances de guérison sont meilleures si le cancer est découvert à un stade précoce. Le dépistage permet d'augmenter les chances de guérison grâce à une détection et une prise en charge de ce cancer à un stade précoce. La maladie se développant très rapidement, plusieurs projets de recherche, comme le projet METAhCOL, sont mis en place pour trouver des solutions pour stopper la reprogrammation métabolique [1]. Stopper la reprogrammation métabolique permettrait de stopper l'évolution de la tumeur.

A L'Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement (INRAE), plusieurs équipes de recherche se penchent sur ce sujet. Durant le stage de deuxième année de master MAPI[3], j'ai travaillé, pendant cinq mois, dans l'une de ces équipes sur l'expression des gènes marqueurs de la maladie dans six lignées cellulaires du côlon, qui couvrent les stades du cancer colorectal du stade sain aux stades cancéreux les plus avancés.

Le stage se propose de répondre à plusieurs questions à partir de données fournies par INRAE Rennes :

- Comment se caractérise chaque lignée cellulaire face aux différentes expositions à des contaminants alimentaires exerçants des effets cancérigènes ?
- Comment se distinguent les différentes expositions à des contaminants alimentaires exerçants des effets cancérigènes entre les lignées cellulaires ?
- Y a-t-il un effet d'interaction entre les traitements et les lignées cellulaires ?

Je commencerai ce rapport en décrivant INRAE et en présentant quelques notions de biologie nécessaires à la compréhension du sujet. Ensuite, je détaillerai le type de données utilisées dans cette étude et expliquerai le travail d'apurement des données puis de détection des effets. La dernière partie sera consacrée à la recherche de gènes différentiellement exprimés en fonction des traitements et des lignées cellulaires à l'aide de diverses méthodes statistiques.

1. La reprogrammation métabolique est une modification du métabolisme, c'est le résultat d'un ensemble de facteurs environnementaux et génétiques permettant à la cellule de synthétiser ses propres molécules tout en assurant sa survie.

1

INRAE, le département MathNUM, l'unité MIAT et le projet METAhCOL

1.1 L'Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement (INRAE)

INRAE est un organisme public de recherche scientifique, 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. INRAE a pour mission 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, la sécurité des aliments, l'environnement et la gestion des territoires. Toutes ses recherches sont effectuées dans une perspective de développement durable.

1.2 Le département de Mathématiques, informatique, sciences de la donnée et du NUMérique (MathNUM)

Le département MathNUM réalise des recherches dans le domaine des mathématiques-informatiques pour répondre à des problématiques en lien avec la science du vivant et de l'environnement. 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 INRAE. 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).

1.3 L'Unité de Mathématiques et Informatique Appliquées de Toulouse (MIAT)

L'unité MIAT est chargée de mettre au point des méthodes mathématiques et informatiques et de les mettre à disposition d'INRAE, favorisant ainsi les collaborations entre départements. Le domaine de compétence de l'unité s'étend de la statistique, aux probabilités, à l'algorithmique, l'intelligence artificielle et aux sciences de la décision. L'unité comporte, depuis janvier 2011, deux équipes de recherches thématiques :

- MAD (Modélisation des Agro-écosystèmes et Décision) : modélisation des systèmes complexes dans les champs de l'agriculture, de l'environnement, 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 destinée à l'exploitation de données de génomique et de post génomique. Mon activité est rattachée à cette équipe.

L'unité s'appuie 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 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 GENomes des Animaux d'Elevage). Elle se compose d'ingénieurs en bio-informatique qui accompagnent les biologistes des départements ≪ animaux ≫ (Génétique Animale, Physiologie Animale et Système d'Elevage, Santé Animale) de l'INRAe dans le traitement de leurs données à haut débit.

Ce stage a été encadré par Mme Nathalie Vialaneix, Directrice de Recherche au sein de l'unité MIAT, dans l'équipe SaAB.

1.4 Le projet METAhCOL

1.4.1 Présentation

Ce stage s'inscrit dans le cadre du projet METAhCOL, financé par l'Agence Nationale de la Recherche et impliquant la plateforme de biostatistique de Toulouse qui est une structure inter-laboratoire qui fédère les compétences en biostatistique sur Toulouse. Par ce biais, mon activité est rattachée à celle de cette plateforme, qui est co-animée par Nathalie Vialaneix, Sébastien Déjean (Institut de Mathématiques de Toulouse, Université Paul Sabatier) et David Rengel (IPBS/CNRS).

Le projet METAhCOL se propose d'étudier l'impact de faibles doses de trois contaminants alimentaires, le benzo[a]pyrène (BaP), le pyrene (Pyr) et la dioxine (TCDD), utilisées seuls ou en mélange, sur la reprogrammation métabolique associée au cancer colorectal. Il utilise un modèle cellulaire innovant composé de six lignées de cellules épithéliales [1] normales isogéniques [2] humaines des

1. Couche de cellules tapissant les organes creux et les glandes.

2. Famille de cellules qui sont sélectionnées ou modifiées pour modéliser avec précision la génétique d'une population de patients spécifique in vitro. Ils peuvent être utilisés pour modéliser une maladie avec une base génétique. Le cancer est une maladie pour laquelle les modèles de maladies humaines isogéniques ont été largement utilisés.

lignées de cellules du côlon qui récapitulent le cancer colorectal aux stades sains, pré-néoplastique [3], adénome [4] et carcinome [5]. Il s'intéresse plus précisément à l'expression de quelques gènes cibles, dont l'analyse multivariée servira à mieux comprendre les mécanismes cellulaires impliqués dans le développement du cancer.

Le cancer colorectal est fortement influencé par les contaminants environnementaux et la nutrition. Parmi les aliments contaminants exerçant des effets cancérigènes, les hydrocarbures aromatiques polycycliques (HAP) sont des xénobiotiques [6] majeurs, formés lors de la cuisson des aliments. L'homme est actuellement exposé de manière chronique aux HAP à faibles doses mais souvent en mélange. Les HAP peuvent se lier et activer les AhR [7]. Des travaux récents ont suggéré que les contaminants alimentaires pourraient participer à la cancérogenèse via des reprogrammations métaboliques.

1.4.2 Déroulé d'une semaine type

Le travail à INRAE s'est organisé selon un cycle hebdomadaire de 5 jours avec un volume horaire de 35 heures. Une réunion hebdomadaire avec les encadrants de stage, Mme N. VIALANEIX, M. DEJEAN et M. RENGEL, avait lieu chaque lundi après-midi. Lors de ces réunions, j'exposais les travaux de la semaine passée sous la forme d'une présentation LibreOffice Impress ou d'un partage d'écran de Rmarkdown. Cette présentation relatait les principaux résultats obtenus suite aux analyses. Elle décrivait aussi ce que j'avais prévu de faire, ce que j'avais réussi à faire, et les difficultés rencontrées. À la fin de chaque réunion, nous discutions ensemble de ces résultats et les encadrants m'indiquaient la direction à suivre pour la suite de l'analyse. Le reste de la semaine, je travaillais en autonomie en télétravail (lors des 3 premiers mois) ou sur site (lors des 2 derniers mois). Les encadrants restaient disponibles par mail en cas de difficultés, et répondaient toujours rapidement à mes interrogations, me permettant de solutionner les problèmes et de maintenir un bon rythme de travail. Si nécessaire, une visio-conférence était organisée. Au cours de la semaine, je partageais le travail effectué sur un dépôt Git. Il m'a également été demandé de rédiger des rapports en anglais des analyses à destination des biologistes du projet METAhCOL et de prendre part à des réunions avec eux afin de leur communiquer les résultats.
À savoir que ce rapport ne représente pas l'intégralité du travail effectué. En effet, 5 rapports en .Rmd rédigé en anglais d'un total d'environ 500 pages ont été réalisé durant ce stage.

3. Les lésions prénéoplasiques apparaissent avant la tumeur. Les lésions néoplasiques sont donc des marqueurs d'un possible développement cancéreux.

4. L'adénome est le premier stade pré-cancéreux

5. Le carcinome est le premier stade pré-cancéreux mais de taille plus grande que l'adénome.

6. Un xénobiotique est une molécule chimique polluante et parfois toxique à l'intérieur d'un organisme, y compris en faibles voire très faibles concentrations. Deux cas typiques de xénobiotiques sont les pesticides et les médicaments, en particulier les antibiotiques.

7. Le récepteur arylique d'hydrocarbures, est un facteur de transcription qui régule l'expression des gènes.

2 Notions de biologie

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 [1], ont été répandues dans la nature, les sols, ou encore les cours d'eau [2]. 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 de cancers comme le cancer colorectal. 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. C'est dans cette optique que le projet METAhCOL a été créé.

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. Pour cette étude, des données d'expression de gènes issues de la technique RT-qPCR ont été analysées. Pour bien comprendre leur signification, il est nécessaire de s'approprier quelques notions de biologie.

2.1 L'ADN et l'ARNm

L'Acide DésoxyriboNucléique (ADN) est une molécule, présente dans toutes les cellules vivantes, qui renferme l'ensemble des informations nécessaires au développement et au fonctionnement d'un organisme. Il porte l'information génétique (génotype) et constitue le génome des êtres vivants.

La structure standard de l'ADN est une double hélice, composée de deux brins complémentaires comme on peut le voir sur la Figure 2.1. Chaque brin d'ADN est constitué d'un enchaînement de nucléotides. On trouve quatre nucléotides différents dans l'ADN, notés A, C, G et T, du nom des bases correspondantes respectivement Adinine, Cytosine, Guanine et Thymine. Ces nucléotides se regroupent par paires spéciales : A avec T et C avec G. Aucune autre paire n'est possible (sauf dans le cas des mutations génétiques).

1. Années 50.

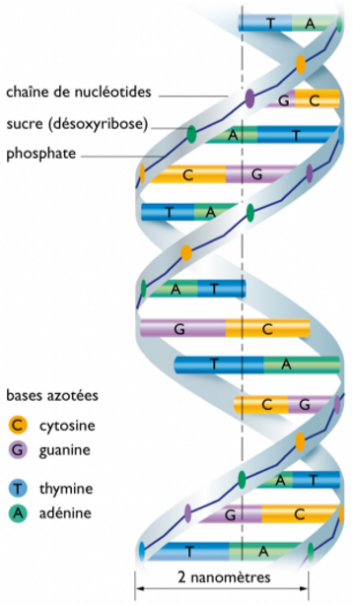


FIGURE 2.1 – Schéma d'un fragment de brin d'ADN [3]

L'ADN est à l'origine de la synthèse des protéines, par l'intermédiaire de l'Acide RiboNucléique messager (ARNm) qui est une copie transitoire d'une portion de l'ADN correspondant à un ou plusieurs gènes. La synthèse des protéines se fait en plusieurs étapes :

FIGURE 2.2 – Étapes de la synthèse des protéines[4]

- La transcription est un transfert d'information génétique de l'ADN vers l'ARNm.
- La traduction est un transfert d'information de l'ARNm vers les protéines.

L'activité des protéines détermine l'activité des cellules, qui vont ensuite déterminer le fonctionnement des organes et de l'organisme. Pour cette étude, on ne s'intéressera qu'à la transcription.

2.2 L'expression des gènes

L'expression des gènes désigne l'ensemble des processus biochimiques par lesquels l'information héréditaire stockée dans un gène est lue pour aboutir à la fabrication de molécules qui auront un rôle actif dans le fonctionnement cellulaire, comme les protéines ou les ARN. Même si toutes les cellules d'un organisme partagent le même génome, certains gènes ne sont exprimés que dans certaines cellules, à certaines périodes de la vie de l'organisme ou sous certaines conditions. La régulation de l'expression génétique est donc le mécanisme fondamental permettant la différenciation cellulaire, la morphogenèse [2] et l'adaptabilité d'un organisme vivant à son environnement.

2. Processus de développement des structures d'un organisme

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 ARNm et protéine étant complexe). On peut, pour cela, quantifier le niveau d'ARNm d'intérêt. La quantification de l'expression des gènes est une mesure quantitative de la transcription. Il serait optimal de pouvoir avoir une mesure quantitative de la traduction mais cela est compliqué et cher. Plusieurs méthodes existent pour quantifier l'expression des gènes, nous ne nous intéresserons qu'à la quantitative Real Time Polymerase Chain Reaction, dénotée RT-qPCR car c'est cette technique qui a été utilisée dans le projet METAhCOL pour générer les données qui seront analysées.

2.3 Mesure de l'expression de gènes avec la technique RT-qPCR

La PCR (Polymerase Chain Reaction) est une technique de réplication de fragments d'ADN ciblés. Elle consiste en une répétition cyclique de la synthèse de ces fragments. Chaque cycle comporte trois étapes qui ont lieu à des températures différentes :

- 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 en excès 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. Cette étape se déroule à la température optimale d'amorçage, qui est particulière à chaque amorce d'ADN.
- 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. Cette étape se déroule normalement à 72°C, qui est la température optimale pour l'activité de l'enzyme.

La donnée brute collectée est un nombre de cycles où chaque cycle correspond à une multiplication de la quantité de matière du cycle précédent (en effet, l'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 utilise un marqueur fluorescent qui se lie aux acides nucléiques, il va permettre de quantifier le fragment à l'issue de chaque cycle de PCR. En résumé, la RT-qPCR autorise le suivi du processus d'amplification en détectant la fluorescence de chaque nouveau produit de la PCR. On mesure de manière détournée la quantité d'un fragment donnée d'ARNm dans la cellule. Autrement dit : dans la RT-qPCR, le nombre de molécules cible est très variable et n'est pas figé à un par le donateur (les parents biologiques). Ainsi, si un gène est fortement exprimé, il aura un grand nombre de molécules de son ARN dans la cellule en question et la RT-qPCR le détectera après un nombre de cycles plus bas que pour un autre gène dont le nombre de molécules d'ARN était plus bas dans le tissu d'origine, c'est-à-dire dont l'expression était plus basse.

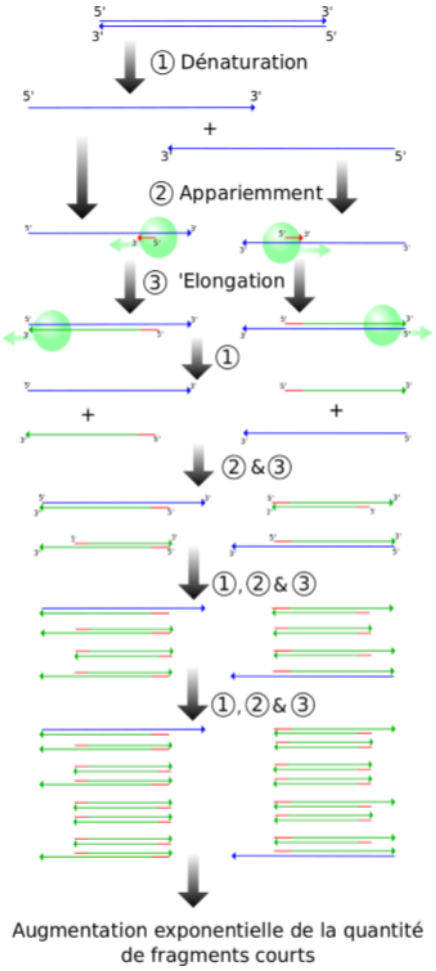


FIGURE 2.3 – Diagramme des quatre premiers cycles de la PCR [5]

3

Contrôles de la qualité et mise en forme des données

3.1 Les données

Présentation

Des biologistes ont testé l'effet de polluants alimentaires sur des lignées cellulaires. Pour cela, ils utilisent la technique RT-qPCR dans le but de mesurer le nombre de cycle de PCR pour quelques gènes obtenus à deux pas de temps différents (48h et 120h d'exposition). Ces gènes proviennent d'études précédentes et sont des gènes candidats dont on soupçonne qu'ils jouent un rôle dans le processus étudié. Le but des expériences est de savoir comment se caractérisent les lignées cellulaires face aux différents contaminants alimentaires et comment se distinguent les différents contaminants alimentaires entre les lignées cellulaires. Pour cela, un contexte expérimental a été mis en place :
Deux facteurs sont présents dans cette étude,

- Un facteur lignée cellulaire composé de six modalités, parmi elles, la lignée dite « contrôle » notées CT [1], les lignées CTA, CTP, CTR, CTRPA, CTRPAt désignent des mutations des cellules CT, mutations dont on sait qu'elles correspondent à des cellules cancéreuses typiques du cancer colorectal. Les lignées CTA sont les mutations les moins avancées, tandis que les lignées CTRPAt sont les mutations les plus avancées du cancer colorectal.
- Un facteur traitement composé de cinq modalités. Les cellules ont été exposées à cinq types de polluants : 6-Benzylaminopurine, Pyrène, dioxine de Seveso, ainsi qu'un mélange de ces précédents polluants. Ils sont respectivement nommés BaP, Pyr, TCDD, et Mix. Les cellules ont également été exposées 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 (48 heures et 120 heures d'exposition aux polluants). Les mesures pour 48 heures et 120 heures d'exposition ne sont pas appariées et sont réalisées indépendamment. Ces mesures correspondent à des expressions de 20 gènes. Les 9 gènes étudiés à 48 heures sont les suivants : *AhRR, CYP1A1, CYP3A4, HK2, HMOX, MCT4, NHE1, NQO1, NRF2*. Les 19 gènes étudiés à 120 heures sont les suivants : *ACO1, AhR, AhRR, ATP5IF1, CAT, CYP1A1, CYP3A4, G6PD, HK2, LPCAT, MCT4, MFN2, ND1, NHE1, NQO1, NRF2, PRDX1, SCD1, TFAM.*

Pour réaliser ces manipulations, il a fallu utiliser un dispositif scientifique, le dispositif est le suivant : les plaques sont introduites dans une machine pour y mesurer l'expression des gènes précédemment définis. Chaque mesure est effectuée indépendamment des autres. Dans ces conditions, Un réplicat

1. À l'issue de l'expérience, comparer les individus du groupe contrôle aux autres individus permet d'évaluer l'effet du traitement. Le groupe contrôle sert de référence.

technique correspond à une lignée cellulaire, un traitement, un gène, une expérience et un pas de temps unique. Un réplicat biologique correspond à une lignée cellulaire, un traitement, un gène et un pas de temps unique.
Les données ont été transmises par les biologistes sous la forme d'un fichier au format xlsx organisé sur sept onglets : « RT-qPCR 48h n=1 », « RT-qPCR 48h n=2 », « RT-qPCR 48h n=3 », « RT-qPCR 120h n=1 NT », « RT-qPCR 120h n=2 NT », « RT-qPCR 120h n=3 NT », « RT-qPCR 120h n=4 NT ». Chaque onglet contenait un tableau, comme sur la figure 3.1 suivante.

Condition	Treatment	Experience (biological replicate)	Plaque	Gene	Value
CT	DMSO	21	12	AhR	15,95
CT	DMSO	21	12	AhR	16,87
CT	DMSO	21	12	AhR	16,66
CT	DMSO	21	12	AhRR	17,1
CT	DMSO	21	12	AhRR	18,42
CT	DMSO	21	12	AhRR	17,76
CT	DMSO	21	12	CYP1A1	24,1
CT	DMSO	21	12	CYP1A1	24,62
CT	DMSO	21	12	CYP1A1	23,86
CT	DMSO	21	12	CYP1B1	18,06
CT	DMSO	21	12	CYP1B1	19,18
CT	DMSO	21	12	CYP1B1	18,35
CT	DMSO	21	12	CYP3A4	27,64
CT	DMSO	21	12	CYP3A4	26,22
CT	DMSO	21	12	CYP3A4	28,1
CT	DMSO	21	12	MCT4	14,25
CT	DMSO	21	12	MCT4	15,23
CT	DMSO	21	12	MCT4	14,98
CT	DMSO	21	12	NHE1	16,57
CT	DMSO	21	12	NHE1	17,62
CT	DMSO	21	12	NHE1	17,3
CT	DMSO	21	12	SDHA	15,03
CT	DMSO	21	12	SDHA	15,8
CT	DMSO	21	12	SDHA	15,55
CT	DMSO	21	12	SDHC	13,6

Condition	Treatment	Experience (biological replicate)	Plaque	Gene	Value
CT	DMSO	50	25	Mitoferrin 1	17,38
CT	DMSO	50	25	Mitoferrin 1	17,83
CT	DMSO	50	25	Mitoferrin 1	17,5
CT	DMSO	50	25	Mitoferrin 2	17,11
CT	DMSO	50	25	Mitoferrin 2	17,38
CT	DMSO	50	25	Mitoferrin 2	17,17
CT	DMSO	50	25	PRDX1	12,61
CT	DMSO	50	25	PRDX1	13,01
CT	DMSO	50	25	PRDX1	13,43
CT	Pyr	51	25	Mitoferrin 1	17,9
CT	Pyr	51	25	Mitoferrin 1	17,81
CT	Pyr	51	25	Mitoferrin 1	17,87
CT	Pyr	51	25	Mitoferrin 2	17,59
CT	Pyr	51	25	Mitoferrin 2	17,3
CT	Pyr	51	25	Mitoferrin 2	18,45
CT	Pyr	51	25	PRDX1	13,07
CT	Pyr	51	25	PRDX1	13,35
CT	Pyr	51	25	PRDX1	13,8
CT	BaP	52	25	Mitoferrin 1	17,56
CT	BaP	52	25	Mitoferrin 1	17,39
CT	BaP	52	25	Mitoferrin 1	17,5
CT	BaP	52	25	Mitoferrin 2	17,47
CT	BaP	52	25	Mitoferrin 2	17,17
CT	BaP	52	25	Mitoferrin 2	16,74
CT	BaP	52	25	PRDX1	12,89

FIGURE 3.1 – Extraits des tableaux à 48h (à gauche) et à 120h (à droite) transmis par les biologistes

Certains détails des dispositifs expérimentaux mis en place n'ont été transmis que fin mai, comme le tableau représentant la date de réalisation des plaques que nous avons demandé aux biologistes après avoir repéré des valeurs atypiques sur les analyses exploratoires. Il a donc fallu du temps pour parvenir à une validation totale des données, ce qui a nécessité de recommencer les analyses.

Mise en forme

Afin de rendre les données exploitables et de pouvoir effectuer des analyses statistiques, il a fallu restructurer les tableaux d'origine. La construction d'une matrice sous un unique onglet possédant les colonnes suivantes a été nécessaire :

- Condition : variable qualitative codant 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.
- Experience : variable qualitative codant les conditions expérimentales. Il prend la forme d'un numéro. Les observations possédant une même valeur ont été réalisées lors d'une même expérience. Toutes les expériences ont été effectuées dans des conditions expérimentales voisines.

- Plaque : variable qualitative identifiant la plaque sur laquelle a été faite la mesure. Elle prend des valeurs entières allant de 1 à 71.
- Timepoint : variable qualitative représentant le temps d'exposition au polluant. Les modalités 48 et 120 correspondent respectivement à une durée d'exposition de 48 heures et 120 heures.
- Gene : variable qualitative codant 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, 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.*
- Value : variable quantitative représentant l'écart entre le nombre de cycle obtenu et le gène de ménage [2].

Et une autre matrice (transmise ultérieurement à la suite de nos demandes) ne possédant pas le même nombre de lignes mais possédant les colonnes suivantes :

- Plate : variable qualitative identifiant la plaque sur laquelle a été faite la mesure (valeurs identiques à la colonne plaque du fichier présenté précédemment). Elle prend des valeurs entières allant de 1 à 71.
- Date : variable qualitative représentant la date à laquelle la plaque a été mesurée.

Une fois exploitables, les données ont été analysées avec R et le nettoyage et l'analyse exploratoire se sont beaucoup basés sur les packages de tidyverse.

3.2 Analyse descriptive

Nous procédons d'abord à une analyse exploratoire dont l'objectif est de préparer deux fichiers séparés (un pour chaque pas de temps car les biologistes ne sont pas intéressés par la comparaison des deux temps) en supprimant les données inutiles et en corrigeant les erreurs.
Un pré-traitement a été réalisé sur la matrice des données avant de débuter l'analyse statistique :

- Nous vérifions d'abord le nombre d'observation par gène. Tous les gènes ne présentent pas le même nombre d'observation, en effet, après 48 heures d'exposition :

Nombre de gène	1	1	5	2
Nombre d'observation	202	265	266	267

  Un gène présente beaucoup moins d'observation que les autres, il faudra par la suite faire attention a ce gène.
  après 120 heures d'exposition :

Nombre de gène	4	3	2	1	2	4	3
Nombre d'observation	268	269	270	354	357	358	359

- Nous vérifions ensuite que toutes les expériences (Condition / Traitement / Expérience / Gène / Point temporel) ont le même nombre de répliques techniques : tous les triplets ne sont pas complets. Dix-huit expériences n'ont que des valeurs manquantes, six expériences n'ont qu'une mesure et 69 expériences n'ont que deux mesures.
- Nous vérifions enfin si tous les gènes ont été quantifiés pour les cinq traitements sur chaque lignée cellulaire à un moment donné : le design est complet.

Une création de 2 sous-fichiers pour chaque point temporel (un à 48 heures et un à 120 heures) pour faire une analyse qui vise à explorer la variabilité des triplicats est ensuite réalisé. Les intervalles et le coefficient de variation (écart-type divisé par la moyenne) sont calculés pour chaque triplicat

2. Un gène de ménage est un gène qui s'exprime dans tous les types cellulaires et dont les produits assurent les fonctions indispensables à la survie des cellules. Les gènes de ménage forment l'expression des gènes ou l'expression génétique.[6]

biologique (pour tous les gènes à un moment, une lignée cellulaire et un traitement donné).

Les parties suivantes ont pour but de trouver les valeurs atypiques et les éventuels effets techniques pour les supprimer et/ou les corriger.

Après 48 heures d'exposition

On calcule une moyenne par triplicat pour n'avoir qu'une seule valeur pour celui-ci. On étudie ensuite la variance au sein d'un triplicat pour voir si cette moyenne a pu conduire à une perte d'information. Une visualisation graphique du coefficient de variation est nécessaire pour repérer des valeurs atypiques. La figure 3.2 représente la variabilité des triplicats par lignée cellulaire et par traitement.

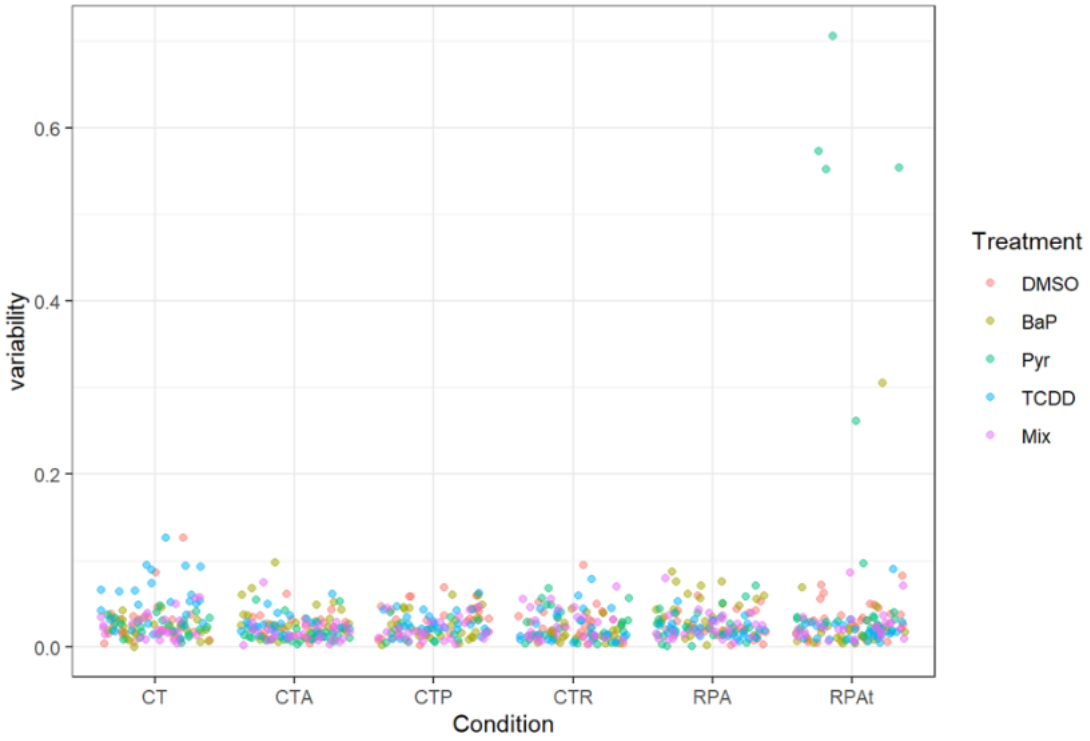


FIGURE 3.2 – Variabilité par lignée cellulaire et par traitement

Nous observons des différences de variabilité entre les gènes. De nombreuses variabilités se situent entre 0 et 0.075 mais certaines approchent 0.6, nous pouvons voir dans la figure 3.2 que ce sont principalement dans la lignée cellulaire CTRPAt sous le traitement Pyr (et une pour le traitement BaP) que des différences sont visibles.

La figure 3.3 représente la variabilité des triplicats par gène et par traitement.

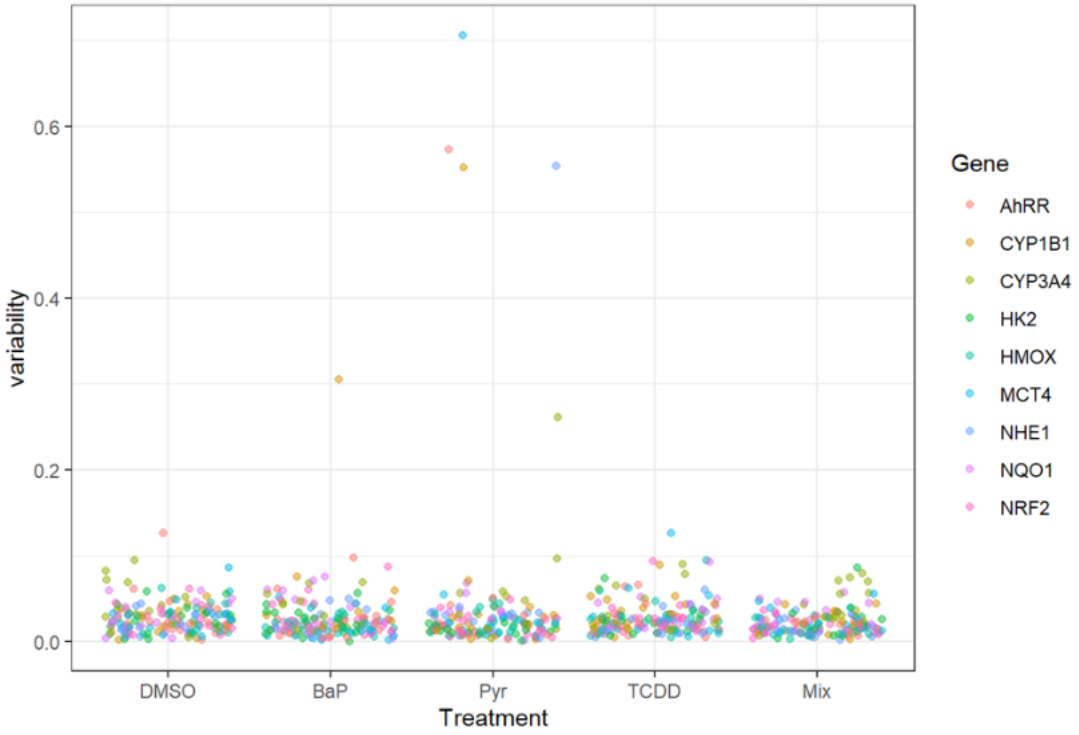


FIGURE 3.3 – Variabilité par traitement et par gène

Si l'on met en lien les figures 3.2 et 3.3, on en conclut que certaines données pour CTRPAt ont montré une variabilité aberrante. Il s'agit du gène *CYP1B1* (pour le traitement au BaP) et des gènes *CYP3A3, AhRR, NHE1, MCT4, CYP1B1* (pour le Pyr). Il est légitime de faire la moyenne des triplicats, il faut néanmoins faire attention car un risque est pris lorsque les données montrent des variabilités aberrantes.

Après 120 heures d'exposition

Des graphiques similaires à ceux produits pour les données à 48 heures montrent que l'on observe une différence de variabilité selon les gènes, beaucoup de variabilités sont comprises entre 0 et 0.05 mais certaines approchent 0.15. Tous les traitements et toutes les lignées cellulaires sont impliqués dans cette différence de variabilité. Les gènes *CYP1B1, AhR, AhRR, CYP3A4, ND1* et *NHE1* présentent une variabilité périphérique, principalement pour CTRPA avec tous les traitements. Il est la aussi légitime de faire la moyenne des triplicats, il faut néanmoins faire attention car un risque est pris sur les données montrant des variabilités aberrantes.

3.3 Formatage

Suite à ces analyses, nous voulons obtenir une unique valeur pour chaque triplicat technique, nous faisons donc la moyenne des valeurs de celui-ci en faisant attention aux triplicats présentant des variabilités fortes. La nouvelle matrice ainsi obtenue avec une unique valeur par triplicat technique est présentée dans la figure 3.4 suivante.

	Condition	Treatment	Experience	Plate	Date	AhRR	CYP1B1	CYP3A4	HK2	HMOX	MCT4	NHE1	NQO1	NRF2
1	CT	DMSO	21	12	08/10/2019	17.76000	18.53000	27.32000	NA	NA	14.820000	17.16333	NA	NA
2	CT	DMSO	21	14	09/10/2019	NA	NA	NA	15.64333	NA	NA	NA	NA	NA
3	CT	DMSO	21	15	10/10/2019	NA	NA	NA	NA	15.96000	NA	NA	10.600000	12.226667
4	CT	DMSO	130	50	03/12/2019	NA	16.56667	25.98667	15.69333	13.73667	NA	15.72667	11.016667	11.460000
5	CT	DMSO	130	60	11/12/2019	15.09667	NA	NA	NA	NA	18.983333	NA	NA	NA
6	CT	DMSO	160	53	11/12/2019	18.39333	17.95667	NA	16.43667	16.14667	23.350000	17.22000	12.106667	13.513333
7	CT	BaP	23	12	08/10/2019	17.49667	18.08667	26.44000	NA	NA	14.523333	16.65667	NA	NA
8	CT	BaP	23	14	09/10/2019	NA	NA	NA	15.78000	NA	NA	NA	NA	NA
9	CT	BaP	23	15	10/10/2019	NA	NA	NA	NA	16.00000	NA	NA	10.766667	12.636667
10	CT	BaP	132	50	03/12/2019	NA	15.99333	26.24000	15.69333	13.47667	NA	15.55333	10.826667	11.733333
11	CT	BaP	132	60	11/12/2019	15.74000	NA	NA	NA	NA	19.933333	NA	NA	NA

FIGURE 3.4 – Capture d'écran d'une partie de la matrice crée pour 48 heures

Les données sont ensuite exportées sous forme de 2 fichiers pour respectivement 48 et 120 heures.

3.4 Recherche d'effets techniques dans les données

Nous effectuons une analyse représentative complète avec des représentations univariées des distributions de gènes sur les 2 fichiers préalablement exportés. Les informations sur les conditions biologiques (lignée cellulaire et traitement) qui permettent d'identifier le signal qu'on cherche à extraire et les conditions techniques (expérience, plaque et date) afin d'identifier les éventuels biais expérimentaux sont étudiés simultanément.

Après 48 heures d'exposition

Nous visualisons les valeurs des gènes en fonction de la lignée cellulaire et du traitement pour déceler d'éventuels effets. Une plus forte valeur indique une lignée cellulaire plus grande donc une expression plus faible.

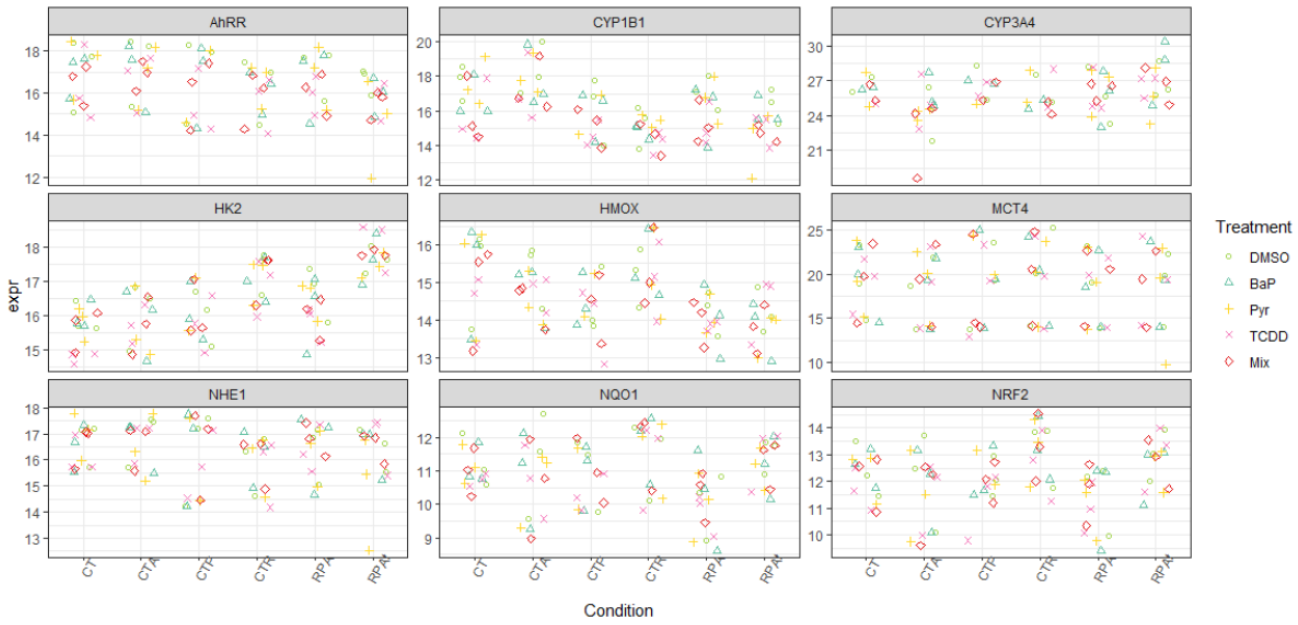


FIGURE 3.5 – Gène expression par traitement et lignée cellulaire

On observe dans la figure 3.5 que pour les gènes *CYP1B1* et *HMOX*, les cellules CT et CTA ont des expressions plus faibles donc une valeur plus importante que les autres lignées cellulaires. Le contraire est observé pour le gène *HK2*. En effet pour ces 3 gènes, la tendance des représentations est que l'expression des gènes *CYP1B1* et *HMOX* est à la hausse plus la lignée cellulaire est proche de cellules cancéreuses et que l'expression du gène *HK2* est à la baisse plus la lignée cellulaire est proche de cellules cancéreuses.

Le traitement ne semble pas influencer beaucoup l'expression des gènes entre les différentes lignées cellulaires car les couleurs représentant les traitements semblent mélangées.

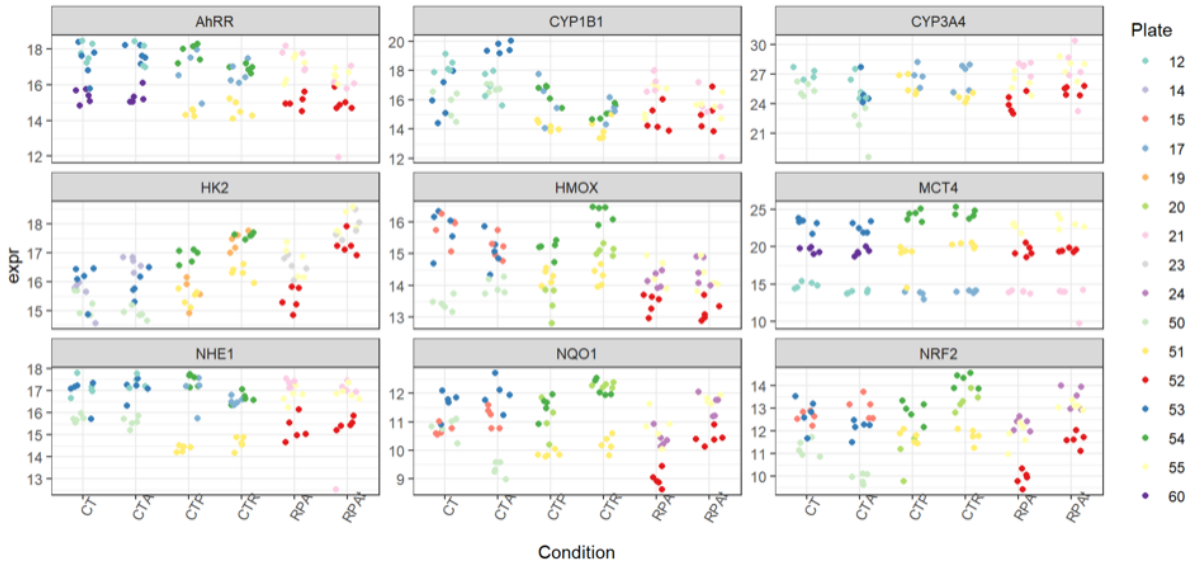


FIGURE 3.6 – Gène expression par plaque et lignée cellulaire

La figure 3.6 montre que le numéro de plaque et la lignée cellulaire sont des facteurs confondus. Le numéro de plaque a un effet très fort, en particulier, sur le gène *MCT4* avec des groupes de plaques avec des niveaux d'expression similaires. En effet, on distingue très clairement que les points sont positionnés sur des lignes horizontales correspondant à des groupes de plaques : les données des

plaques 12, 17 et 21 en bas, les plaques 51, 52, 60 au milieu et 53, 54, 55 en haut. Ces facteurs confondus sont susceptibles d'induire un biais dans l'analyse, produisant ainsi de fausses associations. Dans notre cas, nous n'avons pas réussi a déterminer l'origine de cet effet, nous avons donc décider de continuer sans modification des données.

Après 120 heures d'exposition

Des graphiques similaires à ceux produits pour les données à 48 heures montrent que le traitement et la lignée cellulaire peuvent avoir une légère influence sur la distribution de certains gènes. Les gènes *AhRR* et *CYP1B1* ont une valeur distincte pour les lignées cellulaires CT/CTA vs CTRPA et pour les traitements Pyr/Mix vs DMSO, ce qui est attendu d'un point de vue biologique. Deux gènes présentent néanmoins des valeurs atypiques :

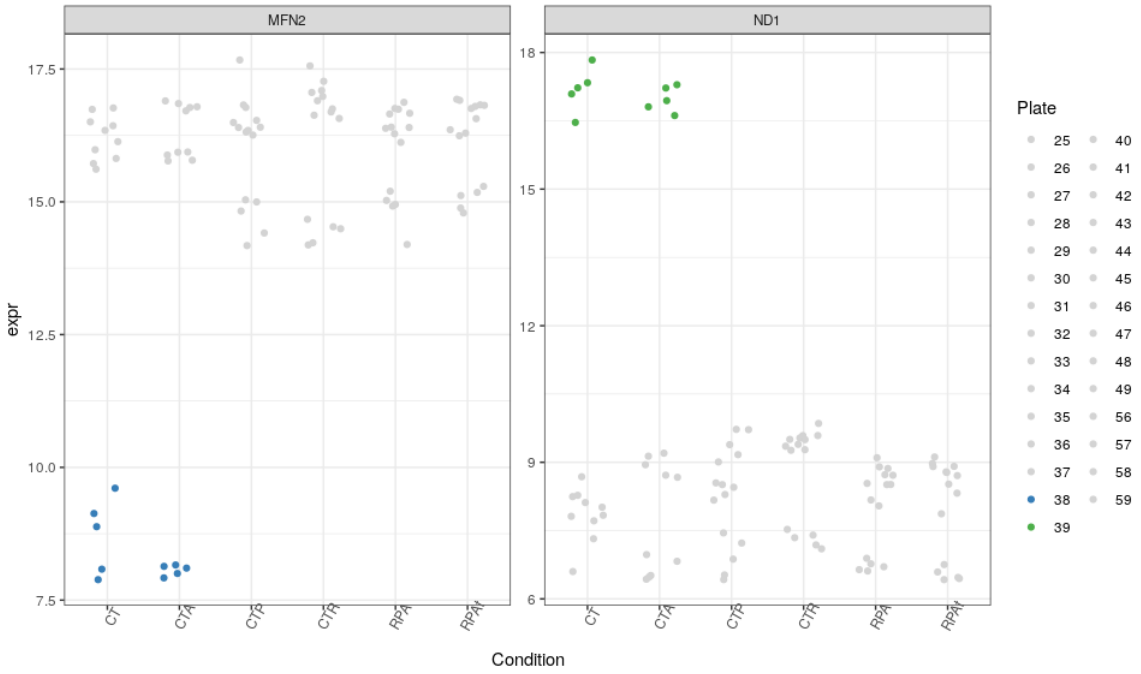


FIGURE 3.7 – Gène expression par plaque et lignée cellulaire

Deux plaques (numéros 38 et 39) ont des valeurs différentes pour respectivement les gènes *MFN2* et *ND1*. Ces deux plaques semblent avoir des valeurs essentiellement atypiques pour les deux gènes mentionnés mais pas pour les autres. Avec l'accord des biologistes, ces valeurs seront supprimées.

4 Analyse exploratoire

4.1 L'Analyse en Composantes Principales (ACP)

À chaque pas de temps, plusieurs ACP ont été réalisées : une générale sur toutes les plaques et d'autres sur des sous-groupes correspondant aux différents groupes de plaques [1]. L'ACP sur toutes les plaques est celle qui apporte le plus d'informations dans le sens où on les retrouve partiellement dans les ACP sur les petits échantillons. C'est celle-ci que l'on va détailler dans la suite.

Objectif : L'ACP a pour but de synthétiser de manière la plus pertinente possible les données initiales pour pouvoir ensuite les représenter graphiquement sur un nombre réduit de dimensions. Si on souhaite étudier les variables deux à deux la représentation graphique est possible. Ici, on voudrait observer toutes les plaques dans un plan comprenant l'ensemble des gènes. L'ACP permet cette visualisation.

En quoi elle consiste : C'est la matrice des variances-covariances (ou celle des corrélations) qui va permettre de réaliser ce résumé pertinent, car on analyse essentiellement la dispersion des données considérées. On utilise la matrice des corrélations à la place de celle des variances-covariances lorsque l'on effectue une ACP normée (c'est-à-dire, les variables ont été centrées et réduites au préalable). Cela permet d'homogénéiser les variables pour ne pas avoir de problèmes liés à des unités de mesure différentes. D'un point de vue mathématique, l'ACP est un changement de base. On passe d'une représentation dans la base canonique des variables initiales à une représentation dans la base des facteurs définis par les vecteurs propres de la matrice des corrélations. Cela va permettre de réaliser les graphiques désirés dans un espace de petite dimension (le nombre de facteurs retenus), en déformant le moins possible la configuration globale des individus selon l'ensemble des variables initiales (c'est-à-dire en choisissant les facteurs de manière à maximiser la dispersion du nuage de points projeté). C'est l'interprétation de ces graphiques qui permettra de comprendre la structure des données analysées.

Mise en œuvre sur R : Elle est très simple à mettre en œuvre sur R grâce à la fonction PCA du package FactoMineR. Cependant, avant de réaliser une ACP, on doit imputer les valeurs manquantes. Pour ce faire, nous avons utilisé la fonction imputePCA() du package missMDA qui réalise une ACP selon une démarche itérative qui permet une imputation des données manquantes. Ensuite, l'ACP standard est effectuée grâce à la fonction PCA() du package FactoMineR sur l'ensemble des données imputées (seules les valeurs manquantes sont remplacées par les résultats de la projection itérative de l'ACP)[7].

1. Un groupe de plaques contient les plaques étudiant les mêmes gènes, il y a 6 groupes à 48h et 8 groupes à 120h

4.1.1 Trois dimensions nécessaires pour résumer l'information

Avant d'interpréter les résultats de l'ACP donnés par le logiciel, il faut choisir le nombre de composantes à retenir. Il existe plusieurs méthodes que l'on peut combiner pour trouver ce nombre. La plus couramment répandue est celle de la recherche d'un « coude » sur l'éboulis des valeurs propres représenté sur la figure 4.1.

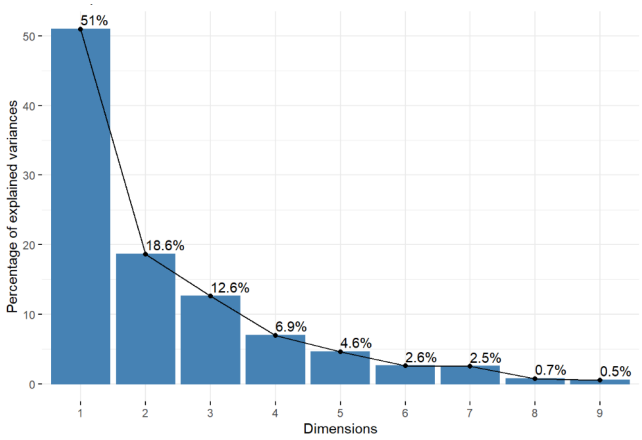


FIGURE 4.1 – Éboulis des valeurs de l'ACP réalisé sur tous les gènes à 48 heures

Un « coude » est visible au niveau de la deuxième composante et le suivant, moins marqué, au niveau de la quatrième composante. Il est aussi possible de sélectionner les axes jusqu'à obtenir un pourcentage de l'inertie totale fixé à priori, généralement 80%. Avec ces critères, seulement trois axes seront donc retenus pour essayer d'identifier des effets.

4.1.2 Identification des effets

L'ACP réalisée nous permet de représenter les données dans un espace de faible dimension pour en observer la structure et détecter des effets expérimentaux (groupe d'individus et individus aux mêmes caractéristiques répartis au même endroit du plan). Grâce à l'ajout de différentes couleurs selon les modalités de diverses caractéristiques des individus, plusieurs effets ont pu être identifiés.

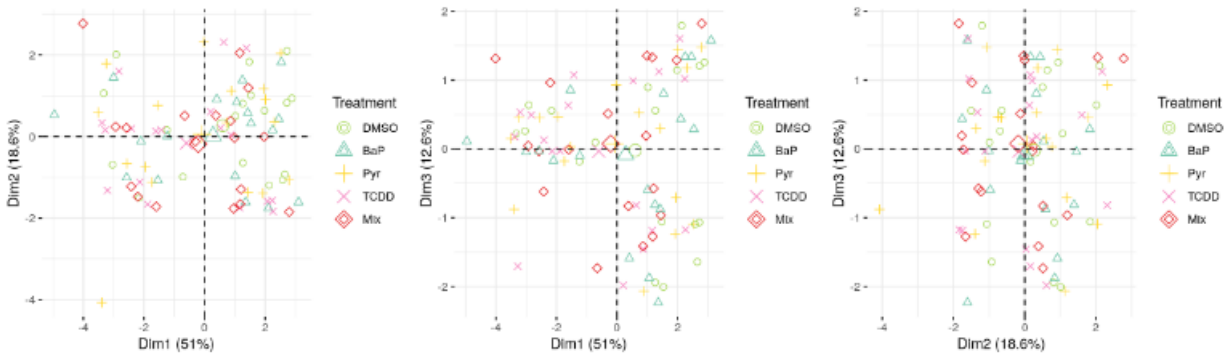


FIGURE 4.2 – Graphes des individus colorés par traitement

Sur la figure 4.2, on voit que le traitement ne semble pas avoir un impact important sur la structure de l'expérience car on constate que les points représentés ne semblent pas se regrouper en fonction du traitement.

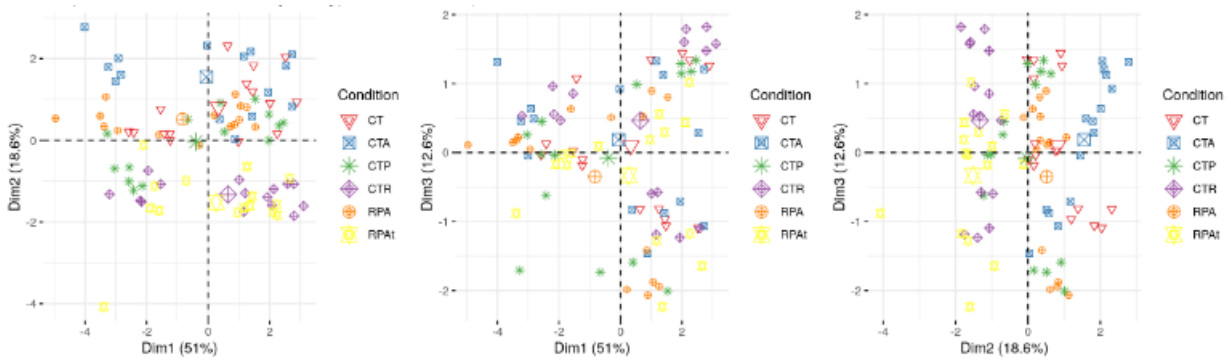


FIGURE 4.3 – Graphes des individus colorés par lignée cellulaire

Sur la figure 4.3, on voit que la lignée cellulaire semble être associée à la projection sur le deuxième axe de l'ACP. En effet, les lignées cellulaires CTRPAt/CTR sont en opposition avec CT/CTA/CTRPA. On peut voir cette opposition sur le graphique à gauche avec les lignées cellulaires CTRPAt et CTR en bas et les lignées cellulaires CT, CTA et CTRPA en haut du graphique. On retrouve aussi la même opposition sur le graphique à droite.

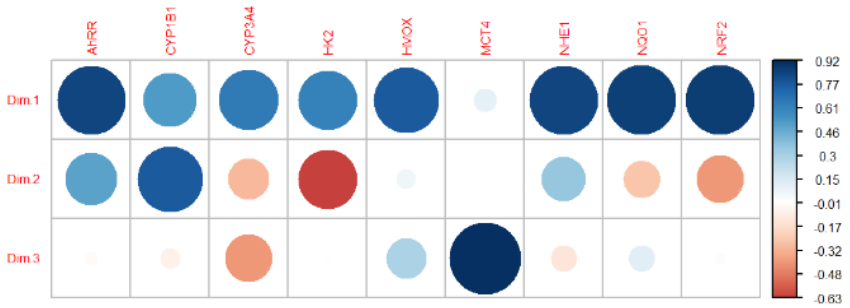


FIGURE 4.4 – Matrice de corrélation

Toutes les variables (expressions de gènes) ont une corrélation positive avec le premier axe comme on peut le voir sur la figure 4.4, ce qui indique un effet d'échelle dans les données (première ligne de la matrice de corrélation bleue).

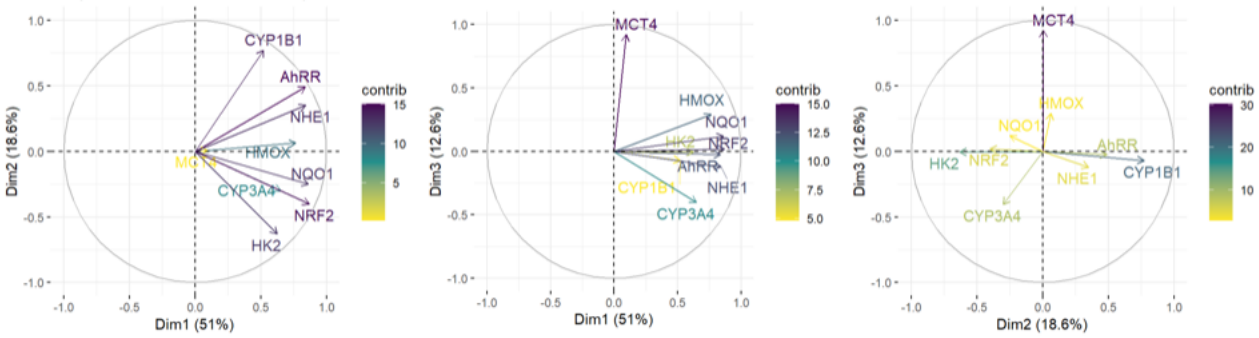


FIGURE 4.5 – Graphes des variables

Sur les graphiques de représentation des variables, on peut voir que le gène *MCT4* est faiblement corrélé avec le premier axe et est, dans l'ensemble, mal représenté sur les deux premières composantes (c'est le seul gène ayant ce comportement) mais bien représenté sur la troisième composante. Le deuxième axe présente une opposition entre l'expression de *CYP1B1* (fortement corrélée positivement avec cet axe) et l'expression de *HK2* (fortement négativement corrélée avec cet axe). Le troisième axe présente une forte corrélation avec l'expression de *MCT4* (fortement positivement corrélée avec cet axe).

4.1.3 Résumé des résultats obtenus

Après 48 heures d'exposition

Les gènes *CYP3A4* et *HK2* semblent bien caractériser les lignées cellulaires, avec une faible valeur pour les lignées cellulaires proches de la normale (CT et CTA) et une forte valeur pour les lignées cellulaires proches des cellules cancéreuses (CTRPAt). Inversement, les gènes *CYP1B1* et *HMOX* ont des valeurs élevées pour les lignées cellulaires proches de la normale (CT et CTA) et une faible valeur pour les lignées cellulaires proches des cellules cancéreuses (CTRPAt). De manière surprenante, la lignée cellulaire CTRPA est plus proche des lignées CT/CTA que des lignées CTRPAt/CTR.

Pour les traitements, les gènes *CYP1B1* et *CYP3A4* semblent avoir une valeur plus faible pour les traitements TCDD et Mix (les plus agressifs) par rapport à DMSO et BaP et *MCT4* semble avoir une valeur plus élevée pour les traitements TCDD et Mix (les plus agressifs).

Après 120 heures d'exposition

Les gènes *CYP1B1* et *ND1* semblent se comporter de manière similaire, avec une valeur élevée pour les lignées cellulaires proches de la normale (CT et CTA) et une faible expression pour les lignées cellulaires proches des cellules cancéreuses (CTRPAt). Inversement, les gènes *CYP3A4* et *HK2* ont une faible expression pour les lignées cellulaires proches de la normale (CT et CTA) et une forte expression pour les lignées cellulaires proches des cellules cancéreuses (CTRPAt). De manière surprenante, les cellules CTRPA ne sont pas similaires à CTRPAt.

Le gène *CYP1B1* est moins exprimé dans les traitements Mix et TCDD, alors que le gène *CYP3A4* est plus exprimé dans les traitements Mix et TCDD. Cependant, le contraire est observé pour le traitement Pyr, alors que les traitements DMSO et le BaP induisent des expressions intermédiaires pour ces gènes.
Le gène *SCD1* est plus exprimé avec les traitements Mix et TCDD et moins exprimé avec les traitements DMSO et BaP. L'expression de ce gène pour le traitement Pyr est intermédiaire.

4.2 L'Analyse Factorielle Discriminante

Cette section reprend les concepts et notations introduits dans Analyse Factorielle Discriminante (AFD)[8].

Objectif : L'analyse discriminante multivariée est une méthode de réduction de dimension pour l'analyse de jeux de données constitués d'une variable qualitative et de $p > 1$ variables quantitatives. Son but est de déterminer quelles combinaisons linéaires des variables quantitatives, Y_1,...,Y_p, mènent à la meilleure discrimination des groupes définis par les m modalités de la variable qualitative X. Cette nouvelle combinaison linéaire notée s est définie ici comme un vecteur de R^n, combinaison linéaire des vecteurs x^1,...,x^p (colonnes de X) :

$$s = Xu = u_1x^1+...+u_px^p,$$

où u = $(u_1, ..., u_p)' \in R^p$ est le vecteur des coefficients de cette combinaison linéaire.

Puisqu'elle est basée sur des combinaisons linéaires des variables originales et des projections linéaires, elle permet d'obtenir des représentations graphiques simplement qui aident à l'interprétation

En quoi elle consiste : L'analyse discriminante multivariée consiste à rechercher de nouvelles variables, qui sont appelées « variables discriminantes », correspondant à des directions de R^p qui séparent le mieux possible les groupes d'individus définis par les modalités de X lorsqu'ils sont projetés sur ces variables discriminantes.

Mise en œuvre sur R : Le nombre d'individus étant supérieur au nombre de variables observées, nous réalisons une PLS-DA. L'analyse discriminante multivariée a été réalisée avec la fonction plsda() du package MASS.

4.2.1 Identification des effets

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

4.2.2 Résumé des résultats obtenus

Après 48 heures d'exposition

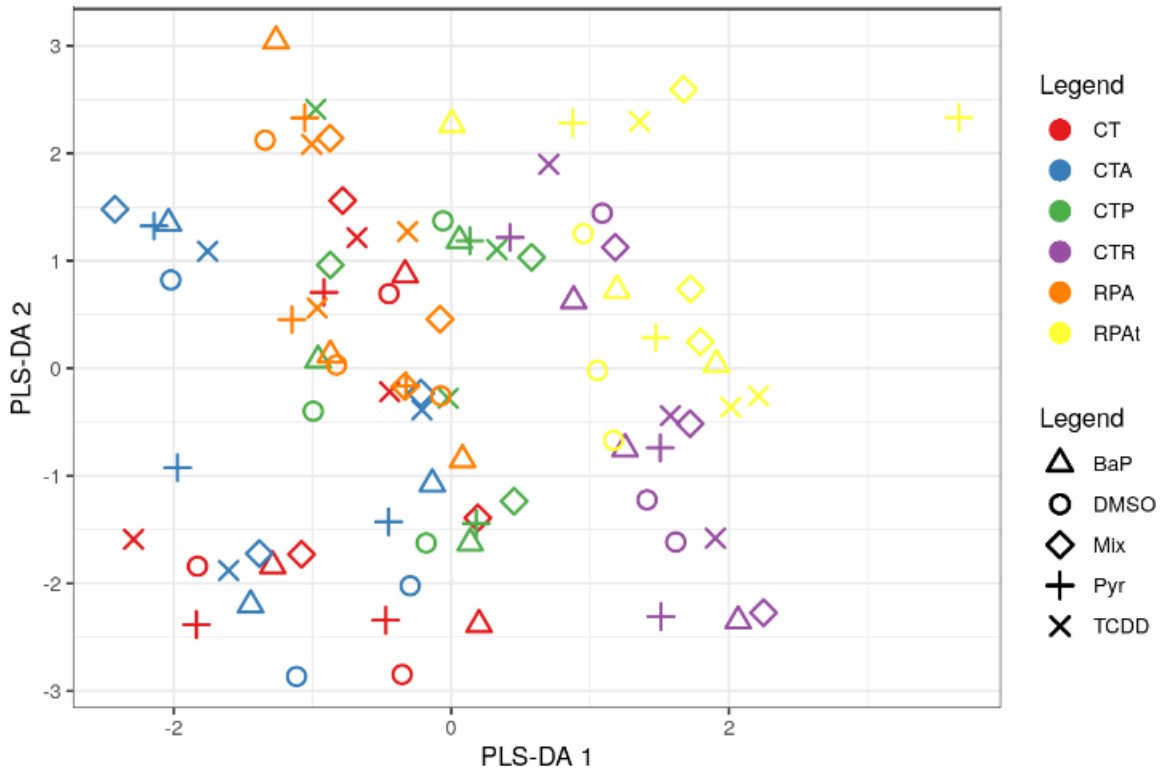


FIGURE 4.6 – PLS-DA sur tous les traitements après 48h d'exposition

Aucune conclusion claire ne peut être tirée des analyses sur l'effet d'un traitement.
Le premier axe de la figure 4.6 montre une opposition entre CTR/CTRPAt (sur le côté droit) et CT/CTA/CTRPA (sur le côté gauche).

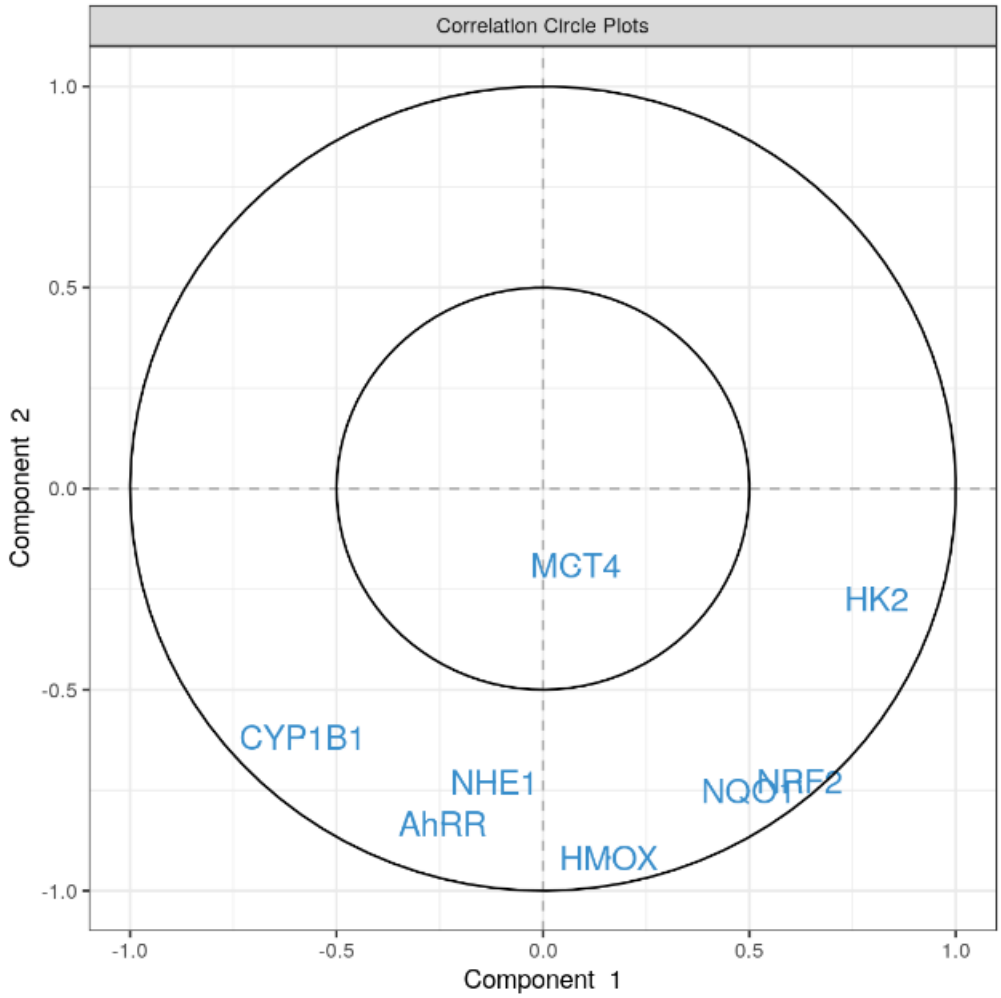


FIGURE 4.7 – Cercle de corrélation après 48 heures d'exposition

Le premier axe de la figure 4.7 présente une corrélation négative avec le gène *CYP1B1* et une corrélation positive avec les gènes *HK2* et *NRF2*. Le gène *CYP1B1* a donc tendance à être sur-exprimé pour les lignées cellulaires CT/CTA/CTRPA et sous-exprimé pour les lignées cellulaires CTR/CTRPAt. L'inverse est vrai pour les gènes *HK2* et *NRF2*.

Après 120 heures d'exposition

Le traitement Pyr se comporte de la même manière que le traitement Mix, qui est le traitement le plus agressif. Le traitement TCDD est caractérisé par la sur-expression de *CYP1B1, HK2* et *AhRR* et les traitements Pyr et Mix par la sous-expression de ces gènes.
Les lignées cellulaires CTA et CTR se comportent de manière totalement opposée : la lignée cellulaire CTA est caractérisée par une sous-expression des gènes *CAT, ACO1, NRF2* et *HK2*, et le contraire est observé pour la lignée cellulaire CTR.

4.3 Synthèse

Les seuls faits principaux qui semblent être uniformément présents dans les conclusions sont que :

- La lignée cellulaire CTRPA se comporte de manière similaire aux lignées cellulaires CT/CTA, qui sont les lignées cellulaires normales les plus proches. Les cellules CTR ont un profil similaire à celui des cellules CTRPAt.
- Les lignées cellulaires CT/CTA/CTRPA sont caractérisés par la sur-expression de *CYP1B1*, les lignées cellulaires CTR/CTRPAt ont un profil opposé. L'inverse est observé pour le gène *HK2*.
- Le traitement Pyr se comporte de la même manière que le traitement Mix, qui est le traitement le plus agressif. Le traitement TCDD est caractérisé par la sur-expression de *CYP1B1* et les traitements Pyr/Mix par la sous-expression de ces gènes.

5

Tests d'hypothèse pour la recherche des différences d'expressions

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 intergroupes des distributions de variables. Des tests de comparaison ont été effectués afin de déterminer quelles différences intergroupes pouvaient être jugées significatives. Pour rechercher les gènes différentiellement exprimés entre les traitements et les lignées cellulaires, on va utiliser plusieurs méthodes : un test de Kruskal-Wallis et des ANOVA suivi de tests post-hoc. Les tests globaux seront précédés du test de Shapiro-Wilk afin de vérifier la normalité des données.

Principe

Un test d'hypothèse est un outil statistique utilisé pour évaluer, avec un niveau de confiance 1-α, si une certaine hypothèse H_0, appelée hypothèse nulle, est compatible avec les données observées. Il s'agit d'une procédure de décision entre H_0 et son hypothèse alternative, H_1, qui est souvent complémentaire à l'hypothèse nulle. L'hypothèse nulle est celle que l'on considère, à priori, comme étant vraie et le test nous permet de vérifier la pertinence de cet à priori. Un test d'hypothèse se caractérise par :

- n observations indépendantes $(Y_1, ..., Y_n)$ d'une la variable aléatoire Y.
- Une statistique de test aussi appelée variable de décision est une variable aléatoire construite à partir d'un échantillon statistique permettant de formuler une règle de décision pour un test statistique elle est notée S.
- Une zone de rejet R_α assurant $P_{H_0}(S \in R_\alpha) = \alpha$, où α est appelée erreur de première espèce.

Nous considérons, dans ce rapport, deux tests :

- Le test de comparaison intergroupes, utilisé pour comparer un ensemble de groupes.
- Le test post-hoc, utilisé pour comparer globalement des groupes deux-à-deux est un test utilisé après le test global qui tient compte de la dépendance des hypothèses testées entre elles et de la multipicité des tests pratiqués.

5.1 Méthodes

On réalise tout d'abord un test de normalité des deux facteurs biologiques, au niveau des moyennes des triplicats biologiques dans le but de pouvoir réaliser ou non par la suite des tests ANOVA.
A la suite du test de normalité, nous programmons de réaliser des tests de comparaison intergroupes.
À la fin de ces premières analyses, nous avons établi le plan d'étude suivant :

- Tests non paramétriques et paramétriques de comparaison. Les tests non paramétriques seraient réalisés en cas de distributions non gaussiennes, les tests paramétriques en cas de distributions gaussiennes.
- Tests paramétriques de comparaison intergroupes tenant compte de l'aléa (modèles mixtes généralisés).
- Tests paramétriques de comparaison intergroupes avec effets additifs et avec effets additifs plus interactions.

5.1.1 Test de Shapiro-Wilk pour la normalité des données

Un test comparant la distribution de l'échantillon à une distribution normale peut être utilisé pour évaluer si les données montrent ou non un écart significatif par rapport à la distribution normale. En effet, de nombreux tests supposent la normalité des distributions pour être applicables comme les tests que nous effectuerons par la suite. Il existe plusieurs méthodes pour évaluer la normalité, dont le test de **Shapiro-Wilk** que nous utiliserons dans ce rapport.

Objectif : Le test de Shapiro–Wilk teste l'hypothèse nulle selon laquelle un échantillon $x_1, \ldots, x_n$ est issu d'une population distribuée selon une loi normale. Les hypothèses de ce test sont :

H_0 : la distribution de l'expression du gène est gaussienne dans
tous les groupes définis par la condition biologique
contre
H_1 : la distribution de l'expression du gène n'est pas gaussienne dans
tous les groupes définis par la condition biologique

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

Mise en œuvre sur R : Pour chaque gène, nous voulons comparer les réponses des 5 traitements et des 6 lignées cellulaires. Nous aimerions savoir si les distributions des gènes sont similaires lorsqu'on examine un traitement ou une lignée cellulaire particulière. Pour ce faire, nous effectuerons des tests de comparaison entre les modalités des facteurs. Nous devons d'abord déterminer si, pour un traitement fixe, chaque gène suit une distribution gaussienne et si pour une lignée cellulaire fixe, chaque gène suit une distribution gaussienne. Si c'est le cas, nous effectuerons une ANOVA. Si ce n'est pas le cas (c'est-à-dire, s'il y a au moins un traitement ou une lignée cellulaire pour lequel le gène ne correspond pas à la distribution gaussienne), nous effectuerons le test non paramétrique de Kruskal-Wallis. Pour cela, nous regardons la p-value de la fonction shapiro.test(gene) du package stats.

5.1.2 Test de Kruskal-Wallis pour des données non gaussiennes

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

$$
\begin{gathered}
H_0 : \text{la distribution de l'expression du gène est identique dans} \\
\text{tous les groupes définis par la condition biologique} \\
\text{contre} \\
H_1 : \text{la distribution de l'expression du gène n'est pas identique dans} \\
\text{tous les groupes définis par la condition biologique}
\end{gathered}
$$

Mise en œuvre sur R : La réalisation de ce test sous R consiste à utiliser la fonction kruskal.test (gene $\sim$ Lignee) du package stats. Cette fonction retourne plusieurs informations mais nous nous concentrerons sur la p-value.

5.1.3 Le modèle ANOVA pour des données gaussiennes

L'analyse de la variance permet de mettre en évidence si une variable numérique à des moyennes significativement différentes entre plusieurs groupes d'individus.

ANOVA à un facteur

Objectif : C'est une généralisation d'un test de Student d'égalité des moyennes lorsqu'il y a plus de deux catégories :

$$
\begin{gathered}
H_0 : \text{l'expression moyenne du gène est identique dans} \\
\text{tous les groupes définis par la condition biologique} \\
\text{contre} \\
H_1 : \text{au moins l'une des moyennes du gène est différente dans} \\
\text{les groupes définis par la condition biologique}
\end{gathered}
$$

Hypothèses : Deux hypothèses sont nécessaires pour que le test soit valide. Il faut, d'une part, que la variable numérique suive une loi normale ce qui peut-être testé à l'aide d'un test de Shapiro-Wilk. D'autre part, il faut que les variances des groupes soient égales. Le test de Bartlett est utilisé pour tester cette hypothèse.

Mise en œuvre sur R : La fonction aov(gene $\sim$ Lignee) permet de réaliser une ANOVA et d'en récupérer la p-valeur.

ANOVA à deux facteurs

Principe : Le but d'une ANOVA à deux facteurs est de tenir compte de l'influence de la lignée cellulaire lorsqu'on teste le facteur traitement et de pouvoir tester l'interaction entre les deux facteurs.

Modèle : Dans le modèle d'ANOVA à deux facteurs, la réponse y_{ijk} de l'individu k du traitement i et du type cellulaire j s'écrit :

$$y_{ijk} = \mu + \alpha_i + \beta_j + \gamma_{ij} + \epsilon_{ijk}$$

où :
μ est la moyenne générale
α_i représente l'effet du traitement i
β_j représente l'effet de la lignée cellulaire j
γ_{ij} représente l'interaction entre les deux facteurs
ϵ_{ijk} représente le résidu

Les hypothèses nulle et alternative du test relatif au facteur traitement peuvent s'écrire :

$$H_0 : \alpha_i = 0 \text{ pour tous les } i = 1, \ldots, 5$$
contre
$$H_1 : \text{tous les } \alpha_i \text{ ne sont pas égaux à } 0$$

Hypothèse relative au facteur lignée cellulaire

$$H_0 : \beta_i = 0 \text{ pour tous les } i = 1, \ldots, 6$$
contre
$$H_1 : \text{tous les } \beta_i \text{ ne sont pas égaux à } 0$$

Hypothèses pour l'évaluation de l'interaction :

$$H_0 : \gamma_{ij} = 0 \text{ pour tous les } i = 1, \ldots, 5 \text{ et } j = i, \ldots, 6$$
contre
$$H_1 : \text{tous les } \gamma_{ij} \text{ ne sont pas égaux à } 0$$

Hypothèses : Identiques aux hypothèses de l'ANOVA à 1 facteur.

Mise en œuvre sur R : La fonction identique à l'ANOVA à un facteur est utilisé pour l'ANOVA à deux facteurs, mais la formule utilisée change, pour un modèle additif nous avons utilisé la ligne de code aov(gene ∼ Lignee + Traitement) et pour un modèle testant l'interaction entre les facteurs, nous avons utilisé la ligne de code aov(gene ∼ Lignee : Traitement) .

5.1.4 Tests post-hoc

Principe : Un test post-hoc se définit comme une procédure à posteriori, permettant d'affiner un résultat positif obtenu par un test global en effectuant des tests individuels par paire de conditions biologiques. En particulier, on peut réaliser des tests post-hoc pour comparer des groupes deux-à-deux lorsque des différences significatives de distributions entre ces groupes sont mises en évidence par des tests de comparaison du type ANOVA ou Kruskal-Wallis. On appelle cela des tests « post-hoc » car ils ne sont pas planifiés d'emblée, mais suivent une conclusion préalable.

Mise en oeuvre sur R : Après une ANOVA, nous réalisons le test post hoc : TukeyHSD(aov(gene ∼ Lignee)) . Après Kruskal-Wallis, nous réalisons le test post hoc : posthoc.kruskal.nemenyi.test(aov(gene ∼ Lignee, dist = "Chisq")) .

5.1.5 Modèle mixte

Un modèle plus complet, utilisant toutes les répliques techniques avec un effet aléatoire pour le numéro de plaque, a ensuite été effectuée pour valider les résultats précédents.

Modèle : Dans un modèle de régression classique, il s'agit d'étudier la liaison statistique entre une variable à expliquer Y et des variables explicatives X non aléatoire. Soit y_i la réponse de l'individu i et x_i les valeurs prises par les variables explicatives pour cet individu. La relation entre X et Y peut s'écrire sous la forme :

où :
ϵ_i est une variable aléatoire distribuée selon une loi normale d'espérance nulle et représentant les résidus du modèle
α correspond à ce qu'on appelle l'intercept
βx_i représente les coefficients du modèle.

Mise en œuvre sur R : Nous réalisons le modèle mixte grâce à la fonction lme() du package nlme.

5.2 Résultats

Dans la suite de ce rapport, seuls les résultats obtenus sur les individus après 120 heures d'exposition seront présentés.

5.2.1 Quels sont les gènes qui s'expriment différemment selon les traitements ?

Pour répondre à cette question, 3 types de tests sont effectués, une ANOVA à un facteur, un test non paramétrique de Kruskal-Wallis et un modèle mixte, ces 3 tests en viennent le plus souvent à la même conclusion.

- Les résultats des tests paramétriques et non paramétriques sont identiques pour ***AhRR*** et ***CYP1B1***. On constate que six couples de traitements présentent une différence d'expression significative, à savoir DMSO-TCDD, BaP-TCDD, Pyr-TCDD, DMSO-Mix, BaP-Mix et Pyr-Mix. De plus, ils présentent une expression plus faible pour TCDD et Mix par rapport aux autres traitements.
- Le gène ***AhR*** s'avère avoir une expression significativement différente entre les traitements pour le test non paramétrique, qui est celui à préférer puisque l'hypothèse de distribution gaussienne n'a pas été rejetée pour ce test. Le gène ***AhR*** a une expression qui augmente régulièrement lorsque le traitement devient plus fort.
- Il existe une différence entre les tests post hoc pour le gène ***SCD1***. Comme l'hypothèse de normalité a été acceptée pour ce gène, le test paramétrique est probablement celui qui doit être préféré. Avec un risque de 5%, il conclut que le gène ***SCD1*** est différentiellement exprimé entre les traitements Pyr et TCDD. ***SCD1*** est un gène qui est sur-exprimé dans le traitement TCDD par rapport au traitement Pyr.

5.2.2 Quels gènes sont exprimés différemment entre les lignées cellulaires, une fois l'effet du traitement estimé ?

Pour répondre à cette question, une ANOVA à 2 facteurs sans interaction est effectuée.

- Pour les traitements :
  Même si le gène ***HK2*** ne montre pas de différence significative entre des traitements particuliers, l'hypothèse d'une distribution identique entre les traitements n'est pas rejetée.
- Pour les lignées cellulaires :
  Plusieurs gènes présentent une différence significative d'expression selon la lignée cellulaire, ce sont généralement les lignées cellulaires CTR et CTRPAt qui présentent des différences avec d'autres lignées cellulaires.
  ACO1 est significativement sur-exprimé dans CTR par rapport aux autres lignées cellulaires. De même, ce gène est significativement sous-exprimé dans CTA par rapport à CTRPAt.
  AhR est significativement sous-exprimé dans CTR par rapport aux autres lignées cellulaires, à l'exception de CTRPA.
  AhRR est significativement sur-exprimé dans CTP par rapport à CT et à CTA.
  CAT est significativement sur-exprimé dans CTR et CTRPAt par rapport aux autres lignées cellulaires.
  CYP1B1 est significativement sous-exprimé dans CTR par rapport aux autres lignées cellulaires. Ce gène est également significativement sous-exprimé dans CTRPA et CTRPAt par rapport à CTA.
  G6PD est significativement sur-exprimé dans CTR et CTRPAt par rapport à CT, CTA et CTR.
  HK2 est significativement sous-exprimé dans CT, CTA et CTRPA par rapport à CTP, CTR et CTRPAt comme on peut le voir sur la figure 5.1.
  LPCAT est significativement sur-exprimé dans CTRPAt par rapport aux autres lignées cellulaires. Cependant, la distribution de ce gène dans CTR et CTRPAt est bimodale, ce qui fausse quelque peu ce résultat.
  NQO1 est significativement sous-exprimé dans CT et CTA par rapport à CTRPAt. De même, ce gène est significativement sous-exprimé dans CTRPA par rapport à CTRPAt.
  NRF2 est significativement sur-exprimé dans CTR et CTRPAt par rapport à CTA. De même, CT est significativement sur-exprimé dans CTR par rapport à CTR et CTRPAt.
  PRDX1 est significativement sur-exprimé dans CTR par rapport à CTRPA.

La figure précédente est une représentation graphique du gène HK2.

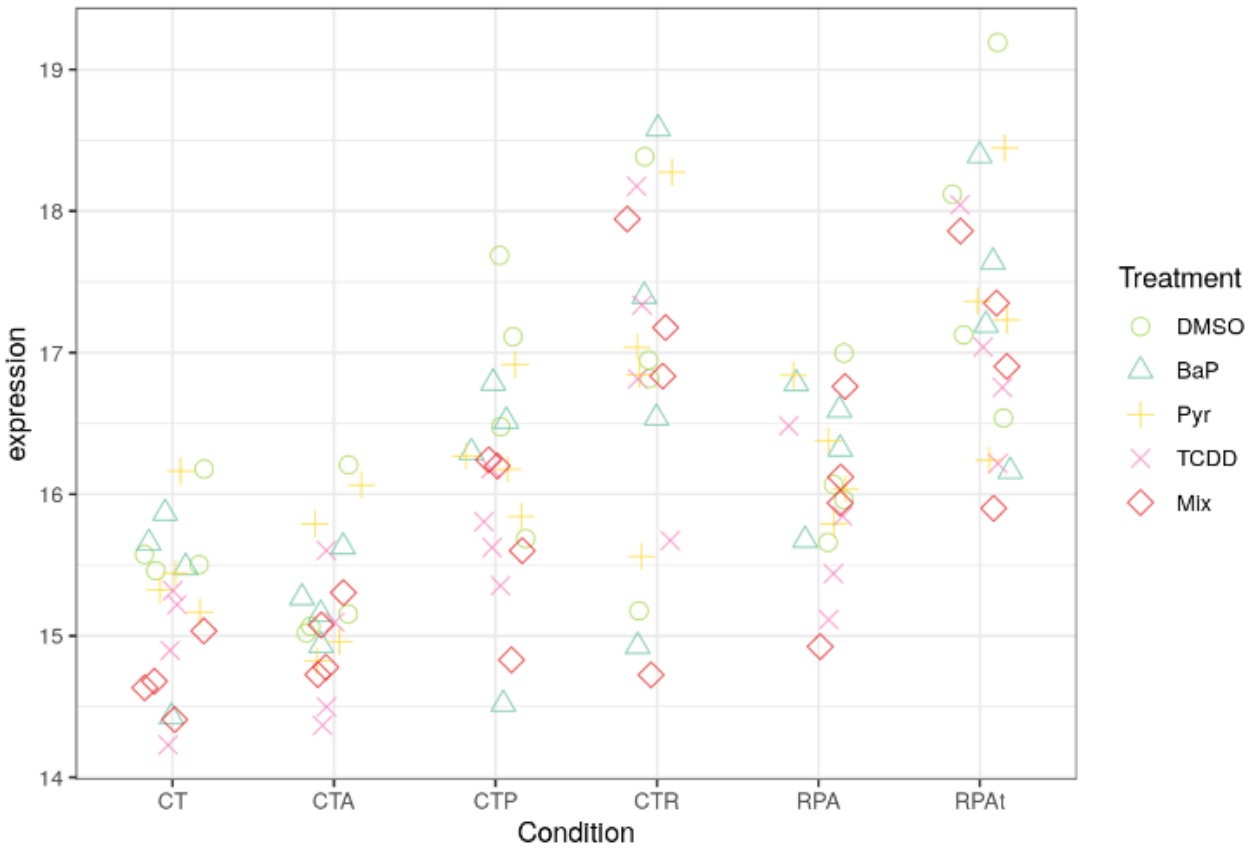


FIGURE 5.1 – Expression du gène HK2 après 120 heures d'exposition

5.2.3 Y a-t-il un effet d'interaction entre les traitements et les lignées cellulaires ?

Pour répondre à cette question, une ANOVA à 2 facteurs avec interaction est effectuée. Le tableau suivant représente les p-valeurs obtenues à la suite d'une ANOVA avec un modèle d'interaction sur les données après 120 heures d'exposition.

	ACO1	*AhR*	*AhRR*	*ARP5IF1*	*CAT*	*CYP1B1*	*G6PD*
Type cellulaire	1.91e-07	8.95e-03	3.10e-03	7.55e-01	6.78e-06	6.13e-18	1.91e-04
Traitement	3.26e-01	1.24e-01	2.43e-10	7.95e-01	8.60e-01	2.83e-14	9.55e-01
Interaction	1.00e+00	1.00e+00	9.97e-01	1.00e+00	1.00e+00	1.00e+00	1.00e+00

	HK2	*LPCAT*	*MCT4*	*MFN2*	*ND1*	*NHE1*	*NQO1*
Type cellulaire	1.41e-14	1.75e-04	9.34e-01	9.60e-01	3.45e-01	5.84e-06	7.36e-06
Traitement	4.71e-02	6.58e-01	9.89e-01	9.16e-01	8.72e-01	9.89e-01	8.44e-01
Interaction	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00

	NRF2	*PRDX1*	*SCD1*	*TFAM*
Type cellulaire	8.64e-04	1.17e-02	4.31e-01	1.38e-01
Traitement	1.69e-01	9.76e-01	3.19e-02	9.85e-01
Interaction	1.00e+00	9.45e-01	1.00e+00	1.00e+00

On peut voir que toutes les p-valeurs liés aux interaction sont grandes (de l'ordre de 10^{-4},10^{-6}) , il n'y a donc pas d'effet significatif de l'interaction entre le traitement et la lignée cellulaire.

On voit facilement sur la figure 5.3 que les lignes colorées sont quasiment parallèles pour le gène CYP1B1, ce qui illustre l'absence d'effet significatif de l'interaction entre les facteurs.

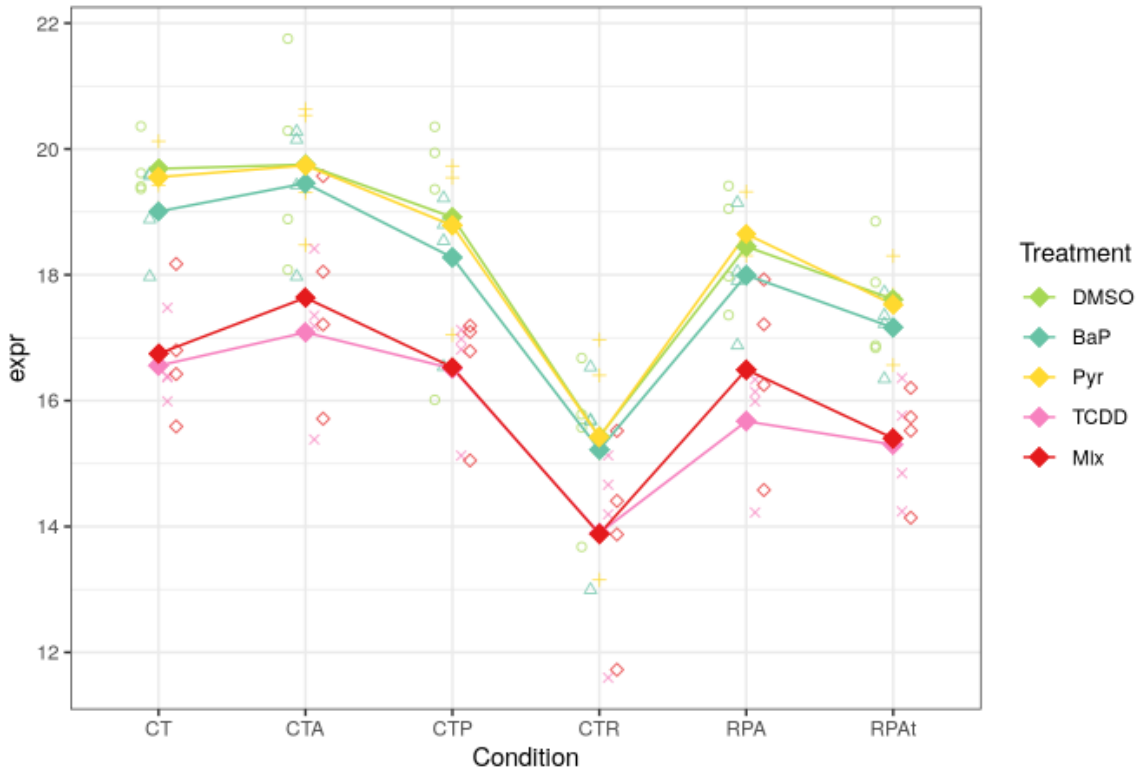


FIGURE 5.2 – Graphique d'interaction du gène *CYP1B1* après 120 heures d'exposition

5.2.4 Synthèse des différentes méthodes utilisés

1- Le gène *CYP1B1* présente toujours une différence d'expression entre TCDD et au moins un des trois traitements les plus faibles mais, comme pour tous les autres gènes, les différences sont beaucoup plus fortes à 120 heures qu'à 48 heures.
Les deux traitements TCDD et Mix présentent de très fortes similitudes et ne montrent jamais de différence significative, quel que soit le gène. Il en va de même pour DMSO et Pyr.

2- Le gène *CYP1B1* présente toujours des différences dans l'expression du traitement pour TCDD et Mix et au moins un des trois autres traitements et la lignée cellulaire pour CTA et trois autres lignées cellulaires et CTRPAt et deux autres lignées cellulaires. Les différences sont beaucoup plus importantes à 120 heures qu'à 48 heures (Résultats non montrés en détails ici).

Les deux traitements TCDD et Mix présentent de très fortes similitudes et ne présentent jamais de différence significative, quel que soit le gène. Il en va de même pour DMSO et Pyr.
La lignée cellulaire CTR présente des différences significatives avec les trois lignées cellulaires les plus proches de la normale.

3- Aucun effet significatif de l'interaction entre lignée cellulaire et traitement n'a été trouvé pour aucun gène.

Conclusion

Par l'intermédiaire de tests de comparaisons inter-groupes paramétriques et non paramétriques et d'analyses multivariées, j'ai pu identifier que le gène présentant le plus de différences entre les traitements est le gène *CYP1B1* et les gènes présentant le plus de différences entre les lignées cellulaires sont les gènes *CYP1B1, HK2, NQO1* et *NRF2*.

Une caractérisation des différences entre des lignées cellulaires mutées a ensuite été réalisé. La lignée cellulaire CTRPA, se différenciait des autres sur un grand nombre de caractéristiques, en effet, elle se comporte similairement à des cellules de stades moins avancés.

Je me suis ensuite intéressée aux effets des polluants sur ces lignées. 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 CTRPA s'est différenciée de la lignée CTRPAt. J'ai également pu conclure que les traitements TCDD et Mix étaient les plus impactants.
Ce résultat était attendu. En effet, 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. J'ai pu me familiariser avec de nouveaux packages dédiés à l'étude de données grâce à la mise en oeuvre, sous R, d'analyses exploratoires, de tests de comparaison, et d'analyses multivariées.
J'ai également pu tirer quelques enseignements, qui me seront assurément utiles dans mon futur professionnel. Je retiens notamment 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

[1] https://www.santepubliquefrance.fr/maladies-et-traumatismes/cancers/cancer-du-colon-rectum

[2] Corinne MAIRIE; Le point sur les additifs alimentaires; 15 avril 2019; Association Santé Environnement France; Disponible sur : https://www.asef-asso.fr/production/le-point-sur-les-additifs-alimentaires-la-synthese-de-corinne-mairie-dieteticienne/

[3] L'image provient de Wikimedia Commons et est attribuable à MesserWoland

[4] L'image provient de Wikimedia Commons et est attribuable à Toony

[5] L'image provient de Wikimedia Commons et est attribuable à Madprime; Disponible sur : https://fr.wikipedia.org/wiki/R%C3%A9actionencha%C3%AEnepar polym%C3%A9rase

[6] https://www.aquaportail.com/definition-10230-gene-de-menage.html

[7] AUDIGIER Vincent, HUSSON Francois et JOSSE Julie; 2012/07/02; Imputation de données manquantes pour des données mixtes via les méthodes factorielles grâce à missMDA; https://www.researchgate.net/publication/278807474 Imputationdedonneesmanquantespourdesdonneesmixtesviales methodesfactoriellesgraceamissMDA

[8] CHAVENT; 2015; http://www.math.u-bordeaux.fr/mchave100p/wordpress/wp-content/uploads/2013/10/AFD.pdf

Analysis of the effects of treatments and cell types on gene expression

CALS Mallory

03 août, 2020

Contents

Introduction

In this file, we will perform tests that are meant to answer the following questions:

- Which genes are differentially expressed between treatments?
- Which genes are differentially expressed between the cell types, once the treatment effect estimated?
- Is there an interaction effect between treatments and cell types?

```
library("tidyverse")
library("data.table")
library("stats")
library("kableExtra")
library("PMCMR")
library("ggplot2")
library("lme4")
```

```
library("nlme")
library("MASS")
library("car")
library("emmeans")
library("cowplot")
library("RColorBrewer")
```

1. Loading and basic description

Gene expressions have been copied from two Excel files `../data/qPCR_2020-01-17.xlsx` and `../data/qPCR_2020-02-14.xlsx` into a CSV file `../data/qPCR_2020-03-24.csv`. The data were then further filtered and cleaned in the file `2-Prepare-Data`, which led to export two files: `../data/qPCR_48_modified_Experience.cs` and `../data/qPCR_48_modified_Experience.csv`. These two files provide the average values of gene expression by technical triplicate (one row is one experiment for one treatment and one cell type).

Data are loaded with:

```
cell_48 <- read_delim("../data/qPCR_48_modified_Experience.csv", "\t",
                      escape_double = FALSE, trim_ws = TRUE)
cell_120 <- read_delim("../data/qPCR_120_modified_Experience.csv", "\t",
                       escape_double = FALSE, trim_ws = TRUE)
```

```
cell_48$Experience = factor(cell_48$Experience)
cell_48$Treatment = factor(cell_48$Treatment,
                           levels = c("DMSO", "BaP", "Pyr", "TCDD", "Mix"))
cell_48$Condition = factor(cell_48$Condition,
                           levels = c("CT", "CTA", "CTP", "CTR", "RPA", "RPAt"))
```

```
cell_120$Experience = factor(cell_120$Experience)
cell_120$Treatment = factor(cell_120$Treatment,
                            levels = c("DMSO", "BaP", "Pyr", "TCDD", "Mix"))
cell_120$Condition = factor(cell_120$Condition,
                            levels = c("CT", "CTA", "CTP", "CTR", "RPA", "RPAt"))
```

For this report, we need to create color and symbol palettes:

```
palettetreatment  <- c("#A6D854", "#66C2A5", "#FFD92F", "#F781BF", "#E41A1C")
palettecondition <- brewer.pal(6, "Set1")
symboltreatment  <- 1:5
symbolcondition <- 6:11
```

2. Which genes are differentially expressed between treatments?

2.1 After a 48 hour exposure

2.1.1 Normality tests

A significance test comparing the sample distribution to a normal distribution can be used to assess whether or not the data show a significant deviation from the normal distribution. There are several methods for assessing normality, including the Kolmogorov-Smirnov (K-S) and the Shapiro-Wilk test that we will use in this report.

For each gene, we want to compare the responses of the 5 treatments. We would like to know if the distributions of the 8 genes are similar when looking at a particular treatment.

To do this, we will conduct comparison tests between the cell lines. We must first determine whether, for a fixed treatment, each gene follows a Gaussian distribution. If this is the case, we will perform an ANOVA. If this is not the case (*i.e.*, if there is at least one treatment for which the gene does not fit the Gaussian distribution), we will perform the non-parametric Kruskal-Wallis test.

In statistics, the Shapiro-Wilk test tests the null hypothesis that a sample $x_1, \ldots, x_n$ comes from a normal distribution.

```
pvals_48 <- apply(cell_48[ ,-c(1:3)], 2, function(agene) {
  tapply(agene, cell_48$Treatment, function(agt)
    as.numeric(format(shapiro.test(agt)$p.value, scientific = TRUE, digits = 4)))
})
format(pvals_48, scientific=TRUE, digits=3) %>% kable() %>%
  kable_styling(bootstrap_options = "striped") %>%
  add_header_above(c(" " = 1, "p-values" = 8))
```

	p-values							
	AhRR	CYP1B1	HK2	HMOX	MCT4	NHE1	NQO1	NRF2
DMSO	3.01e-02	8.18e-01	6.76e-01	6.84e-02	5.03e-02	2.62e-02	6.34e-01	1.80e-01
BaP	6.20e-02	5.07e-01	9.79e-01	7.48e-01	3.28e-02	1.50e-02	5.15e-01	3.57e-01
Pyr	6.78e-02	7.21e-01	4.26e-01	1.93e-01	8.72e-02	1.72e-01	7.74e-01	3.13e-01
TCDD	7.18e-01	3.05e-02	1.77e-01	7.15e-01	5.50e-02	1.24e-01	3.22e-01	7.36e-01
Mix	1.83e-01	1.38e-01	1.77e-01	6.95e-01	7.29e-03	5.99e-02	7.27e-01	8.57e-01

From the output, the genes for which p-value $<$ 5e-02 are the ones for which the distribution significantly deviates from a Gaussian. For these genes, non parametric tests will be preferred.

Conclusion : Effects on the treatment on gene expressions will be performed using Kruskal-Wallis tests for AhRR, CYP1B1, MCT4 and NHE1 and using ANOVA for HK2, HMOX, NQO1 and NRF2.

2.1.2 ANOVA

ANOVA to test the treatment effect is performed for all genes. These tests test the following hypothesis:

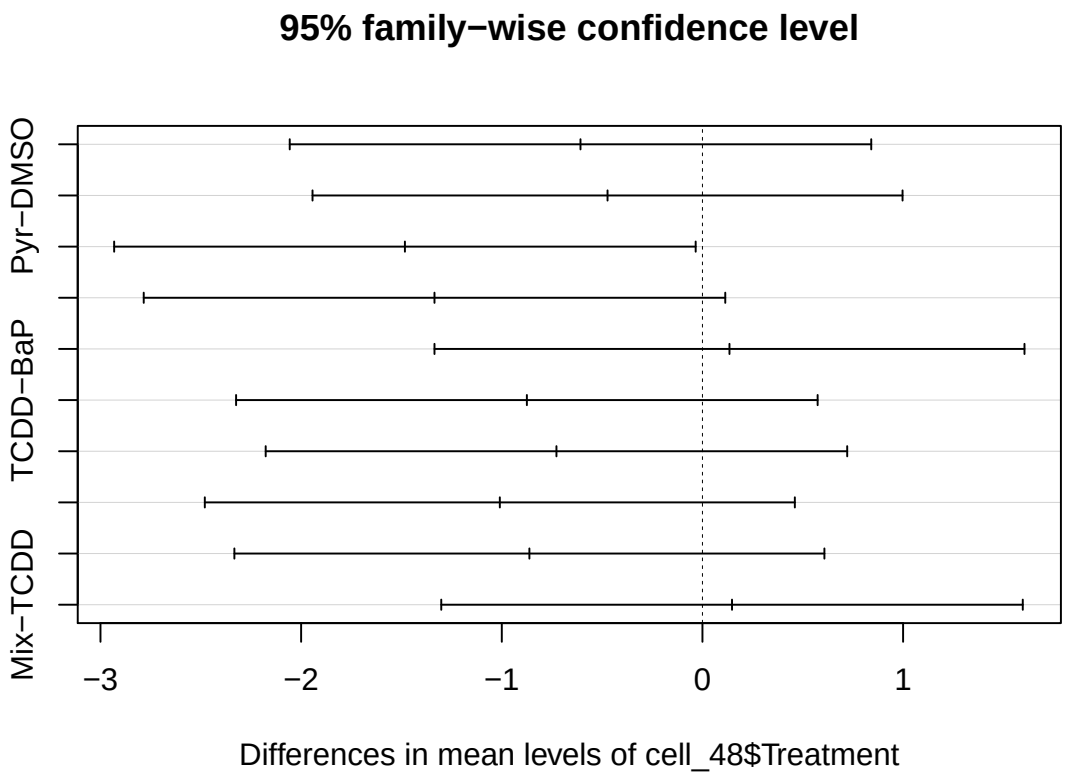
H_0: The average expression is the same for all treatments

```
anova_48 <- apply(cell_48[ ,-c(1:3)], 2, function(agene) {
  summary(aov(agene ~ cell_48$Treatment))[[1]][1,5]
})
format(anova_48, scientific=TRUE, digits=3) %>% kable() %>%
  kable_styling(bootstrap_options = "striped") %>%
  add_header_above(c(" " = 1, "p-values" = 1))
```

	p-values
	x
AhRR	2.62e-01
CYP1B1	2.96e-02
HK2	7.85e-01
HMOX	7.86e-01
MCT4	1.00e+00
NHE1	7.90e-01
NQO1	8.22e-01
NRF2	7.61e-01

The number of genes that are differentially expressed between treatments is 1 (at a 5% risk and without multiple testing correction). This (these) gene(s) are: CYP1B1.

```
selected <- which(anova_48 < 0.05)
anova_posthoc_plot_48 <- apply(cell_48[ ,-c(1:3)][ ,selected], 2, function(agene) {
  plot(TukeyHSD(aov(agene ~ cell_48$Treatment)))
})
```




```
anova_posthoc_48 <- apply(cell_48[ ,-c(1:3)][ ,selected], 2, function(agene) {
  TukeyHSD(aov(agene ~ cell_48$Treatment))[[1]][,4]
})
format(anova_posthoc_48, scientific=TRUE, digits=3) %>% kable() %>%
  kable_styling(bootstrap_options = "striped") %>%
  add_header_above(c(" " = 1, "p-values" = length(selected)))
```

	p-values
	CYP1B1
BaP-DMSO	7.69e-01
Pyr-DMSO	8.97e-01
TCDD-DMSO	4.23e-02
Mix-DMSO	8.55e-02
Pyr-BaP	9.99e-01
TCDD-BaP	4.50e-01
Mix-BaP	6.30e-01
TCDD-Pyr	3.18e-01
Mix-Pyr	4.79e-01
Mix-TCDD	9.99e-01

For each gene, the number of pairs of treatments that show significant differences in expression is:

```
sapply(1:ncol(anova_posthoc_48), function(ind) {
  cat(paste0("For gene: ", colnames(anova_posthoc_48)[ind], "\n"))
  cat(paste0("Number of significant pairs: ",
              sum(anova_posthoc_48[ ,ind] < 0.05),
             "\n"))
  cat(names(which(anova_posthoc_48[ ,ind] < 0.05)), "\n\n")
})
```

```
## For gene: CYP1B1
## Number of significant pairs: 1
## TCDD-DMSO
```

```
## [[1]]
## NULL
```

2.1.3 Kruskal-Wallis

In addition, non-parametric tests (Kruskal-Wallis) are also used to test a similar hypothesis:

H_0 : The distributions of the gene expression are the same for all treatments

```
Kruskal_48 <- apply(cell_48[ ,-c(1:3)], 2, function(agene) {
  kruskal.test(agene ~ cell_48$Treatment)[[3]]
})
format(Kruskal_48, scientific=TRUE, digits=3) %>% kable() %>% kable_styling(bootstrap_options = "stripe
  add_header_above(c(" " = 1, "p-values" = 1))
```

	p-values
	x
AhRR	1.28e-01
CYP1B1	1.03e-02
HK2	6.88e-01
HMOX	8.64e-01
MCT4	9.77e-01
NHE1	8.67e-01
NQO1	7.87e-01
NRF2	7.09e-01

The number of genes that are differentially expressed between treatments is 1 (at a 5% risk and without multiple testing correction). This (these) gene(s) are: CYP1B1.

Post-hoc tests are performed for the genes that show a positive result:

```
signif <- which(Kruskal_48 < 0.05)
Kruskal_posthoc_48 <- apply(cell_48[ ,-c(1:3)][ ,signif], 2, function(agene) {
  posthoc.kruskal.nemenyi.test(agene ~ cell_48$Treatment, dist = "Chisq")
})
Kruskal_posthoc_48
```

```
## $CYP1B1
##
##  Pairwise comparisons using Nemenyi-test with Chi-squared
##                        approximation for independent samples
##
## data:  agene by cell_48$Treatment
##
##      DMSO  BaP   Pyr   TCDD
## BaP  0.867 -     -     -
## Pyr  0.964 0.998 -     -
## TCDD 0.058 0.465 0.301 -
## Mix  0.123 0.652 0.470 0.999
##
## P value adjustment method: none
```

and for each gene, the number of pairs of treatments showing a significant difference in expression is:

```
Kruskal_phpval_48 <- lapply(Kruskal_posthoc_48, function(agene) agene$p.value)
lapply(Kruskal_phpval_48, function(agene) sum(agene < 0.05, na.rm = TRUE))
```

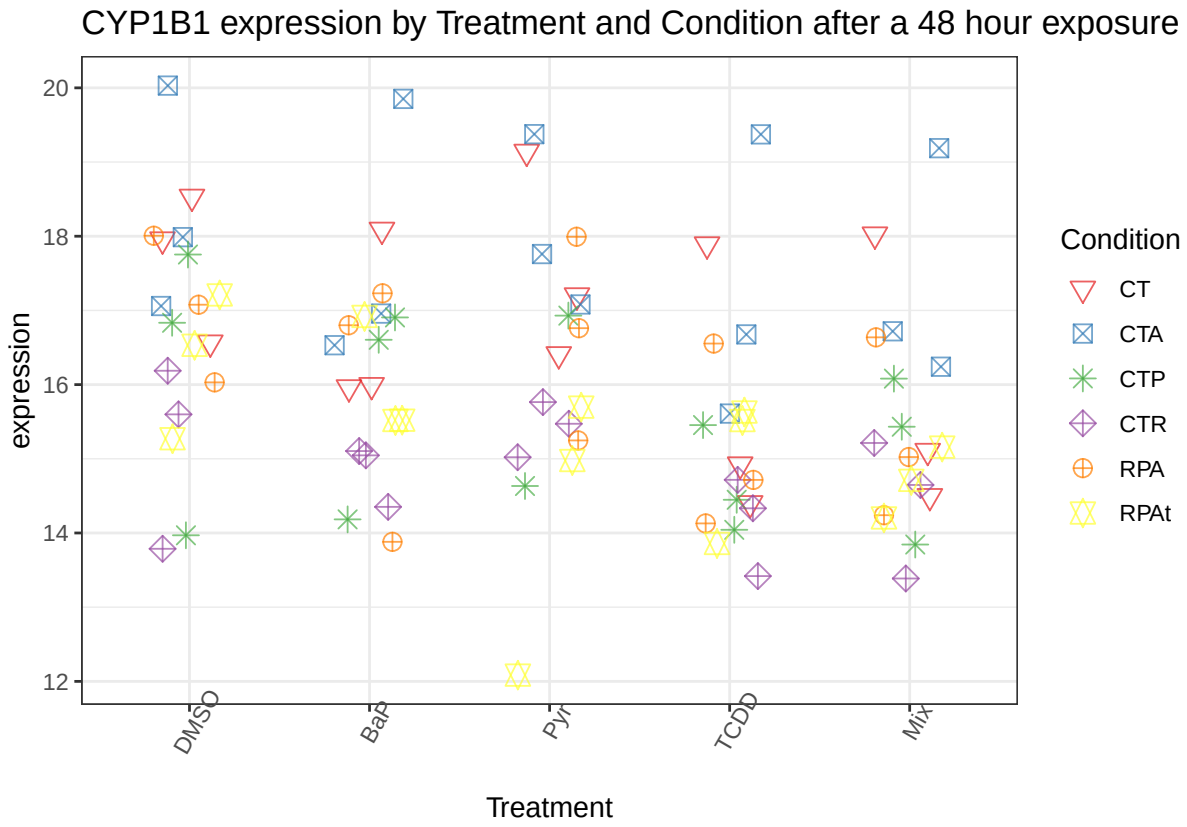
```
## $CYP1B1
## [1] 0
```

2.1.4 Conclusion

There is a minor difference between parametric and non-parametric test results for the gene CYP1B1. Since the normality assumption has been rejected for this gene, the non-parametric test is probably the one to be preferred. It concludes, at a 5% risk, in a global significant difference in gene expression between treaments. However, one couple of treatments DMSO/TCDD is found to have a significant difference in expression for this gene with parametric tests (but the expression of CYP1B1 significantly deviates from normality for TCDD) and, for the non-parametric tests, the pairs of treatments that deviate the most from the null hypothesis are DMSO/TCDD and DMSO/Mix, as expected.

```
all_signif <- names(cell_48[ ,-c(1:3)][ ,signif])
sapply(all_signif, function(agene) {
  df <- data.frame(cell_48$Treatment, cell_48[ ,agene], cell_48$Condition)
  names(df) <- c("Treatment", "expr", "Condition")
  p <- ggplot(df, aes(x = Treatment, y = expr, colour = Condition, shape = Condition)) +
  geom_jitter(alpha = 0.7, width = 0.2, size = 3) + theme_bw() +
    theme(axis.text.x = element_text(angle = 60)) + ylab("expression") +
    ggtitle(paste0(agene,
                   " expression by Treatment and Condition after a 48 hour exposure")) +
    scale_color_manual(values = palettecondition) +
    scale_shape_manual(values = symbolcondition) +
    guides(alpha = 0)
```

```
  print(p)
  invisible(NULL)
})
```




```
## $CYP1B1
## NULL
```

As we can see from the previous graph, the gene CYP1B1 has larger expression in DMSO than in TCDD and Mix.

2.2 After a 120 hour exposure

2.2.1 Normality tests

We first use the shapiro test to find out if the distributions of genes by treatment are Gaussian:

```
pvals_120 <- apply(cell_120[ ,-c(1:3)], 2, function(agene) {
  tapply(agene, cell_120$Treatment, function(agt) shapiro.test(agt)$p.value)
})
format(pvals_120[,1:9], scientific=TRUE, digits=3) %>% kable() %>%
  kable_styling(bootstrap_options = "striped") %>%
  add_header_above(c(" " = 1, "p-values" = 9))
```

	p-values								
	ACO1	AhR	AhRR	ATP5IF1	CAT	CYP1B1	G6PD	HK2	LPCAT
DMSO	8.52e-01	1.91e-01	2.85e-01	6.65e-03	6.81e-01	4.96e-01	4.81e-01	1.12e-01	1.84e-01
BaP	7.13e-01	1.26e-02	6.70e-01	3.73e-02	8.33e-01	1.70e-01	6.46e-01	6.40e-01	7.62e-01
Pyr	2.03e-01	3.47e-02	9.34e-01	2.82e-01	2.41e-01	4.64e-02	3.41e-01	3.56e-01	6.65e-01
TCDD	8.04e-01	4.24e-02	8.40e-01	1.76e-02	4.31e-01	1.95e-01	4.40e-03	3.59e-01	9.54e-01
Mix	4.59e-01	9.90e-02	9.35e-01	2.40e-02	3.27e-01	8.34e-01	7.92e-03	4.37e-02	2.49e-01

```
format(pvals_120[,10:18], scientific=TRUE, digits=3) %>% kable() %>%
  kable_styling(bootstrap_options = "striped") %>%
  add_header_above(c(" " = 1, "p-values" = 9))
```

	p-values								
	MCT4	MFN2	ND1	NHE1	NQO1	NRF2	PRDX1	SCD1	TFAM
DMSO	1.63e-01	1.41e-01	4.57e-02	2.75e-01	8.13e-02	1.86e-01	6.46e-01	3.12e-01	3.57e-01
BaP	2.42e-01	9.78e-02	9.21e-02	1.55e-02	4.00e-02	1.76e-01	4.29e-01	3.43e-01	2.70e-01
Pyr	1.47e-01	1.08e-02	6.30e-01	6.86e-02	3.96e-01	2.00e-01	5.17e-01	5.48e-01	2.72e-01
TCDD	3.68e-01	1.38e-02	4.78e-01	5.15e-02	7.12e-02	4.89e-01	7.52e-05	9.63e-01	2.91e-01
Mix	1.42e-01	1.90e-01	1.10e-02	4.09e-01	1.08e-01	8.53e-02	8.49e-01	5.81e-01	1.33e-01

Conclusion : Effects on the treatment on gene expressions will be performed using Kruskal-Wallis tests for AhR, ATP5IF1, CYP1B1, G6PD, HK2, MFN2, ND1, NHE1, NQO1 and PRDX1 and using ANOVA for ACO1, AhRR, CAT, LCPAT, MCT4, NRF2, SCD1 and TFAM .

2.2.2 ANOVA

ANOVA to test the treatment effect is performed for all genes:

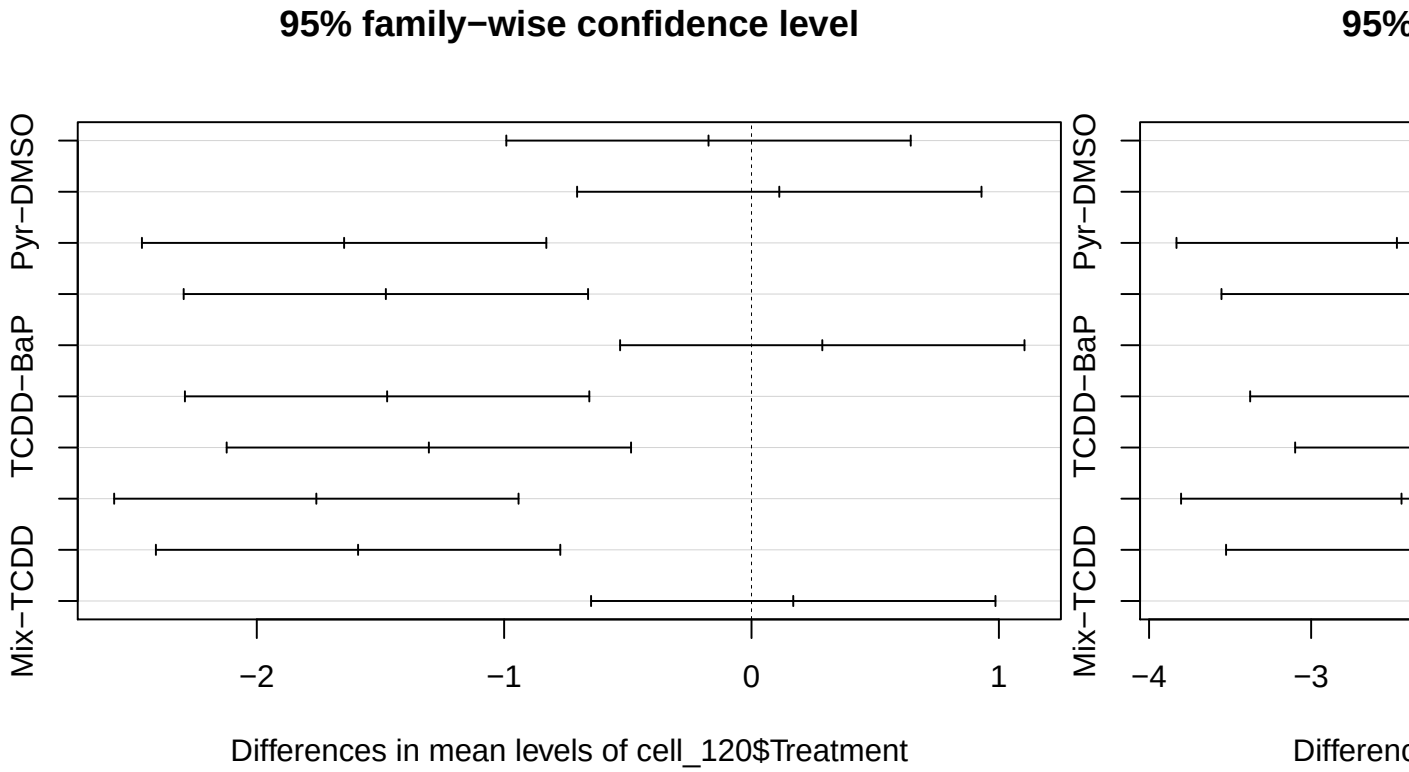
```
anova_120 <- apply(cell_120[ ,-c(1:3)], 2, function(agene) {
  summary(aov(agene ~ cell_120$Treatment))[[1]][1, 5]
})
format(anova_120, scientific=TRUE, digits=3) %>% kable() %>%
  kable_styling(bootstrap_options = "striped") %>%
  add_header_above(c(" " = 1, "p-values" = 1))
```

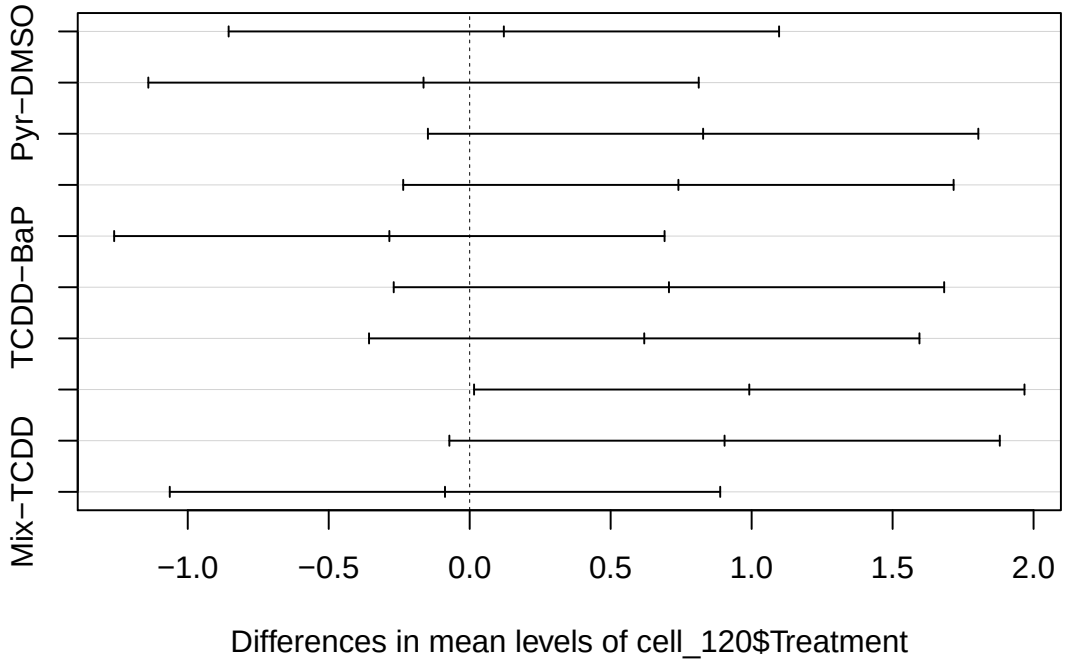
	p-values
	x
ACO1	4.75e-01
AhR	1.07e-01
AhRR	7.87e-11
ATP5IF1	7.09e-01
CAT	8.98e-01
CYP1B1	3.10e-08
G6PD	9.60e-01
HK2	2.40e-01
LPCAT	6.76e-01
MCT4	9.80e-01
MFN2	8.47e-01
ND1	8.07e-01
NHE1	9.82e-01
NQO1	8.74e-01
NRF2	1.68e-01
PRDX1	9.77e-01
SCD1	1.42e-02
TFAM	9.79e-01

The number of genes for which the null hypothesis is rejected is 3 (at 5%) and these genes are: AhRR, CYP1B1, SCD1.

For these genes, post-hoc tests are performed:

```
selected <- which(anova_120 < 0.05)
anova_posthoc_plot_120 <- apply(cell_120[ ,-c(1:3)][ ,selected], 2, function(agene) {
  plot(TukeyHSD(aov(agene ~ cell_120$Treatment)))
})
```






```
anova_posthoc_120 <- apply(cell_120[ ,-c(1:3)][ ,selected], 2, function(agene) {
  TukeyHSD(aov(agene ~ cell_120$Treatment))[[1]][ ,4]
})
format(anova_posthoc_120, scientific=TRUE, digits=3) %>% kable() %>%
```

```
  kable_styling(bootstrap_options = "striped") %>%
  add_header_above(c(" " = 1, "p-values" = length(selected)))
```

	p-values		
	AhRR	CYP1B1	SCD1
BaP-DMSO	9.76e-01	8.86e-01	9.97e-01
Pyr-DMSO	9.95e-01	1.00e+00	9.90e-01
TCDD-DMSO	1.57e-06	1.76e-05	1.37e-01
Mix-DMSO	1.93e-05	1.78e-04	2.26e-01
Pyr-BaP	8.68e-01	9.08e-01	9.27e-01
TCDD-BaP	2.07e-05	7.06e-04	2.69e-01
Mix-BaP	2.13e-04	5.12e-03	4.03e-01
TCDD-Pyr	2.74e-07	2.24e-05	4.45e-02
Mix-Pyr	3.70e-06	2.23e-04	8.35e-02
Mix-TCDD	9.79e-01	9.80e-01	9.99e-01

For each gene, the number of pairs of treatments that show significant differences in expression is:

```
sapply(1:ncol(anova_posthoc_120), function(ind) {
  cat(paste0("For gene: ", colnames(anova_posthoc_120)[ind], "\n"))
  cat(paste0("Number of significant pairs: ",
             sum(anova_posthoc_120[ ,ind] < 0.05), "\n"))
  cat(names(which(anova_posthoc_120[ ,ind] < 0.05)), "\n\n")
})
```

```
## For gene: AhRR
## Number of significant pairs: 6
## TCDD-DMSO Mix-DMSO TCDD-BaP Mix-BaP TCDD-Pyr Mix-Pyr
##
## For gene: CYP1B1
## Number of significant pairs: 6
## TCDD-DMSO Mix-DMSO TCDD-BaP Mix-BaP TCDD-Pyr Mix-Pyr
##
## For gene: SCD1
## Number of significant pairs: 1
## TCDD-Pyr
```

```
## [[1]]
## NULL
##
## [[2]]
## NULL
##
## [[3]]
## NULL
```

2.2.3 Kruskal-Wallis

In addition, non-parametric tests (Kruskal-Wallis) are also used to test:

```
Kruskal_120 <- apply(cell_120[ ,-c(1:3)], 2, function(agene) {
  kruskal.test(agene ~ cell_120$Treatment)[[3]]
})
```

```r
format(Kruskal_120, scientific=TRUE, digits=3) %>% kable() %>%
  kable_styling(bootstrap_options = "striped") %>%
  add_header_above(c(" " = 1, "p-values" = 1))
```

	p-values
	x
ACO1	6.02e-01
AhR	4.82e-02
AhRR	3.83e-09
ATP5IF1	7.75e-01
CAT	8.51e-01
CYP1B1	7.18e-08
G6PD	8.47e-01
HK2	2.06e-01
LPCAT	6.93e-01
MCT4	9.46e-01
MFN2	5.83e-01
ND1	6.80e-01
NHE1	9.77e-01
NQO1	8.14e-01
NRF2	1.83e-01
PRDX1	9.78e-01
SCD1	2.35e-02
TFAM	9.93e-01

The number of genes that are differentially expressed between treatments is 4 (at a 5% risk and without multiple testing correction). These genes are: AhR, AhRR, CYP1B1, SCD1.

Post-hoc tests are performed for the genes that show a positive result:

```r
signif <- which(Kruskal_120 < 0.05)
Kruskal_posthoc_120 <- apply(cell_120[ ,-c(1:3)][,signif], 2, function(agene) {
  posthoc.kruskal.nemenyi.test(agene ~ cell_120$Treatment, dist = "Chisq")
})
Kruskal_posthoc_120
```

```
## $AhR
##
##  Pairwise comparisons using Nemenyi-test with Chi-squared
##                          approximation for independent samples
##
## data:  agene by cell_120$Treatment
##
##      DMSO BaP  Pyr  TCDD
## BaP  0.99 -    -    -
## Pyr  1.00 1.00 -    -
## TCDD 0.44 0.22 0.29 -
## Mix  0.71 0.44 0.54 0.99
##
## P value adjustment method: none
##
## $AhRR
##
```

```
##  Pairwise comparisons using Nemenyi-test with Chi-squared
##                         approximation for independent samples
##
## data:  agene by cell_120$Treatment
##
##      DMSO    BaP     Pyr     TCDD
## BaP  1.00000 -       -       -
## Pyr  0.96860 0.96026 -       -
## TCDD 0.00099 0.00120 4.2e-05 -
## Mix  0.00583 0.00688 0.00036 0.99336
##
## P value adjustment method: none
##
## $CYP1B1
##
##  Pairwise comparisons using Nemenyi-test with Chi-squared
##                         approximation for independent samples
##
## data:  agene by cell_120$Treatment
##
##      DMSO    BaP     Pyr     TCDD
## BaP  0.97345 -       -       -
## Pyr  0.99996 0.95101 -       -
## TCDD 0.00057 0.00757 0.00034 -
## Mix  0.00450 0.03873 0.00285 0.98952
##
## P value adjustment method: none
##
## $SCD1
##
##  Pairwise comparisons using Nemenyi-test with Chi-squared
##                         approximation for independent samples
##
## data:  agene by cell_120$Treatment
##
##      DMSO BaP  Pyr  TCDD
## BaP  1.00 -    -    -
## Pyr  0.99 0.91 -    -
## TCDD 0.37 0.61 0.14 -
## Mix  0.42 0.66 0.16 1.00
##
## P value adjustment method: none
```

For each gene, the number of pairs of treatments that show significant differences in expression is:

```
Kruskal_phpval_120 <- lapply(Kruskal_posthoc_120, function(agene) agene$p.value)
sapply(1:length(Kruskal_phpval_120), function(ind) {
  cat(paste0("For gene: ", names(Kruskal_phpval_120)[[ind]], "\n"))
  nb_pairs <- sum(Kruskal_phpval_120[[ind]] < 0.05, na.rm = TRUE)
  cat(paste0("Number of significant pairs: ", nb_pairs, "\n"))
  if (nb_pairs > 0) {
    cur_mat <- Kruskal_phpval_120[[ind]]
    pvals <- cur_mat[lower.tri(cur_mat, diag = TRUE)]

    contrasts <- expand.grid(1:nlevels(cell_120$Treatment),
```

```
                            1:nlevels(cell_120$Treatment))
    contrasts <- contrasts[contrasts[ ,1] < contrasts[ ,2], ]
    contrasts <- paste(levels(cell_120$Treatment)[contrasts[ ,1]],
                       levels(cell_120$Treatment)[contrasts[ ,2]],
                       sep = "-")
    cat("Significant differences between: ", contrasts[pvals < 0.05], "\n\n")
  } else cat("\n")
})
```

```
## For gene: AhR
## Number of significant pairs: 0
##
## For gene: AhRR
## Number of significant pairs: 6
## Significant differences between:  BaP-Pyr DMSO-TCDD Pyr-TCDD DMSO-Mix BaP-Mix Pyr-Mix
##
## For gene: CYP1B1
## Number of significant pairs: 6
## Significant differences between:  BaP-Pyr DMSO-TCDD Pyr-TCDD DMSO-Mix BaP-Mix Pyr-Mix
##
## For gene: SCD1
## Number of significant pairs: 0
```

```
## [[1]]
## NULL
##
## [[2]]
## NULL
##
## [[3]]
## NULL
##
## [[4]]
## NULL
```

2.2.4 Conclusion

The following table represents all genes for which at least one global test is significant

- first column: which test(s) is (are) significant?
- second column: for which treatment(s) is normality rejected for this gene?

	Tests giving these results	Treatments whose normality Shapiro rejects
AhR	K-W test	BaP, TCDD, Pyr
AhRR	ANOVA test & K-W test	-
CYP1B1	ANOVA test & K-W test	Pyr
SCD1	ANOVA test & K-W test	-

For the genes AhRR and CYP1B1, six couples of treatments are found to have a significant difference in expression, namely: DMSO-TCDD, BaP-TCDD, Pyr-TCDD, DMSO-Mix, BaP-Mix and Pyr-Mix.

The following tables are constructed as follows:

- the upper parts of the tables are composed by the results of the post-hoc Kriskal-Wallis test. The lower part of the tables are composed by the results of the ANOVA test;
- A star means that the contrast between the two treatments is significant (p-value < 0.05);
- A dash means that the test has not been performed because the global test was not significant or because the same condition is involved in this entry of the table

AhR	DMSO	BaP	Pyr	TCDD	Mix
DMSO	-				
BaP	-	-			
Pyr	-	-	-		
TCDD	-	-	-	-	
Mix	-	-	-	-	-

AhRR	DMSO	BaP	Pyr	TCDD	Mix
DMSO	-			*	*
BaP		-		*	*
Pyr			-	*	*
TCDD	*	*	*	-	
Mix	*	*	*		-

CYP1B1	DMSO	BaP	Pyr	TCDD	Mix
DMSO	-			*	*
BaP		-		*	*
Pyr			-	*	*
TCDD	*	*	*	-	
Mix	*	*	*		-

SCD1	DMSO	BaP	Pyr	TCDD	Mix
DMSO	-				
BaP		-			
Pyr			-		
TCDD			*	-	
Mix					-

AhR is found to have a significantly different expression between treatment for the non-parametric test, which is the one to be prefered since this gene does not fit a Gaussian distribution in 3 treatments.

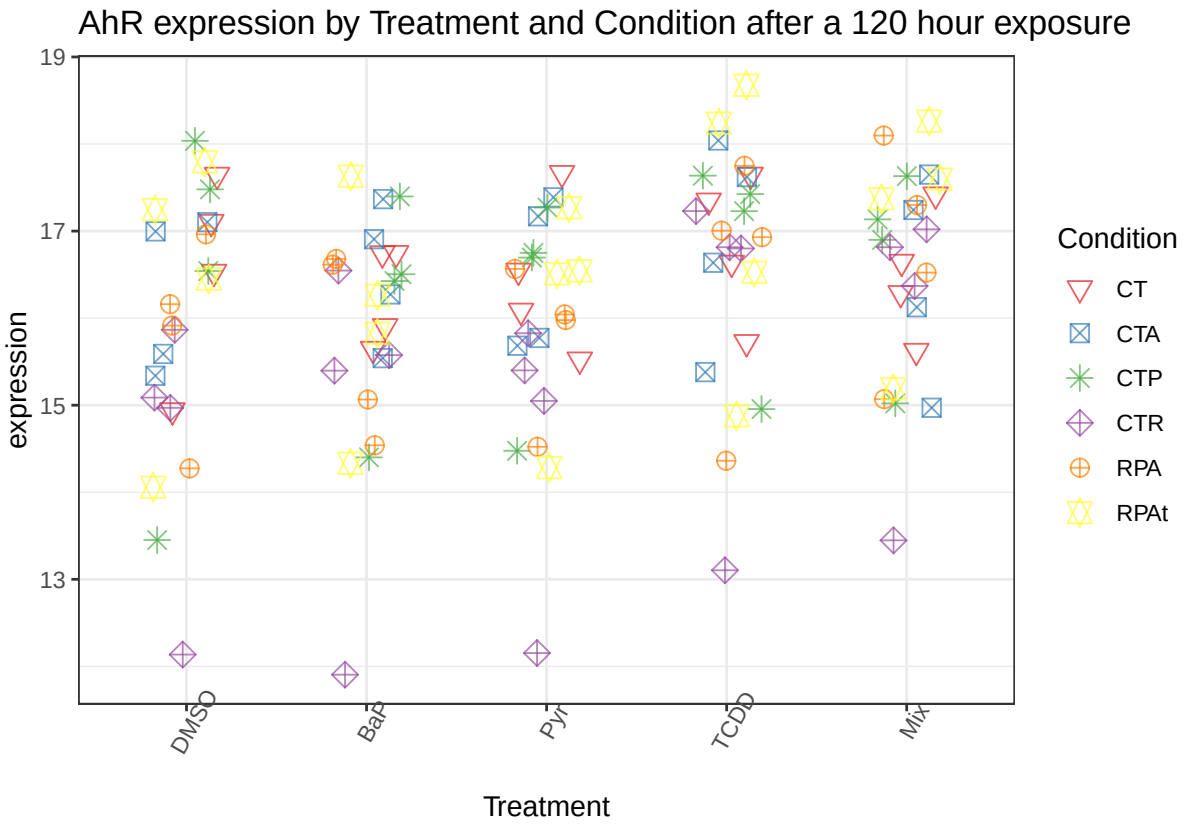
Results of the parametric and non-parametric tests are identical for AhRR and CYP1B1 that are found to be differentially expressed between TCDD or Mix and all the other treatments.

There are a difference between the post hoc tests for the gene SCD1. Since the normality assumption has been accepted for this gene, the parametric test is probably the one to be preferred. At a 5% risk, it concludes that SCD1 is differentially expressed in Pyr and TCDD.

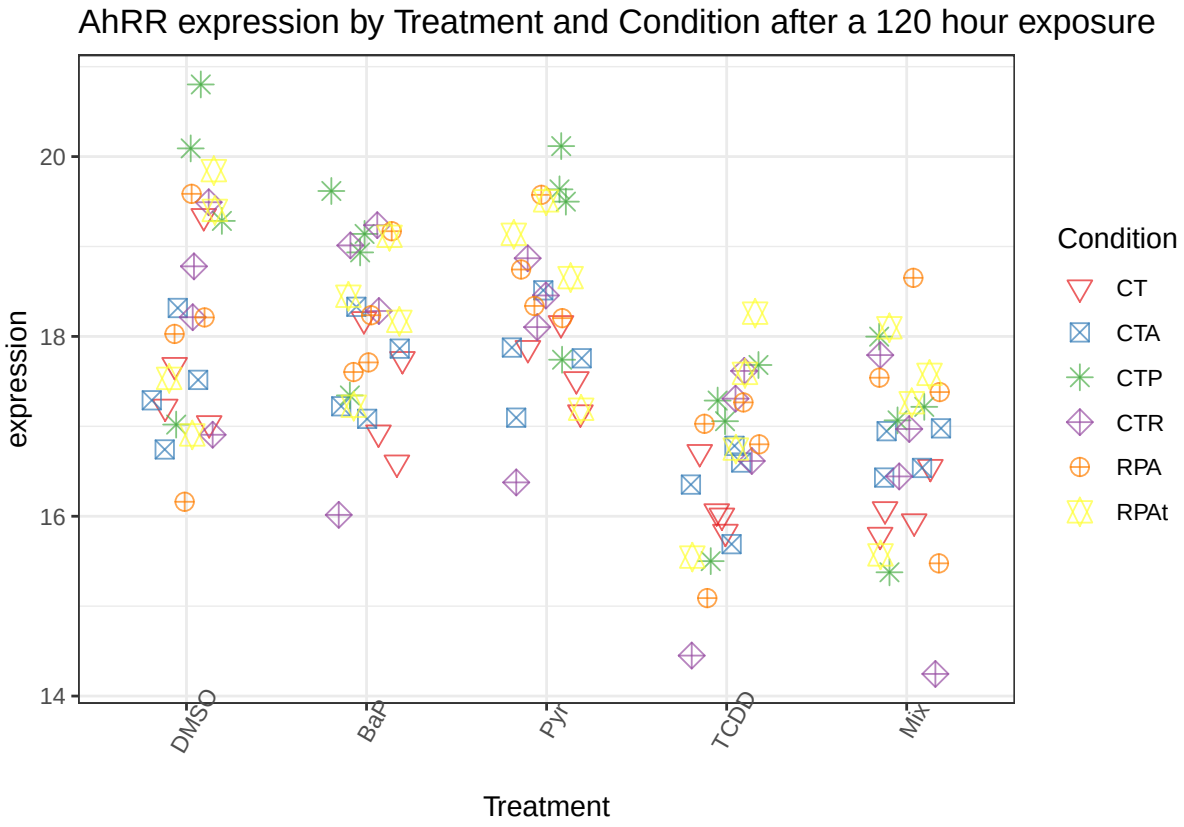
```
all_signif <- names(cell_120[ ,-c(1:3)][ ,signif])
sapply(all_signif, function(agene) {
  df <- data.frame(cell_120$Treatment, cell_120[ ,agene], cell_120$Condition)
  names(df) <- c("Treatment", "expr", "Condition")
```

```
  p <- ggplot(df, aes(x = Treatment, y = expr, colour = Condition, shape = Condition)) +
  geom_jitter(alpha = 0.7, width = 0.2, size = 3) + theme_bw() +
    theme(axis.text.x = element_text(angle = 60)) + ylab("expression") +
    ggtitle(paste0(agene,
                   " expression by Treatment and Condition after a 120 hour exposure")) +
    scale_color_manual(values = palettecondition) +
    scale_shape_manual(values = symbolcondition)
  print(p)
  invisible(NULL)
})
```

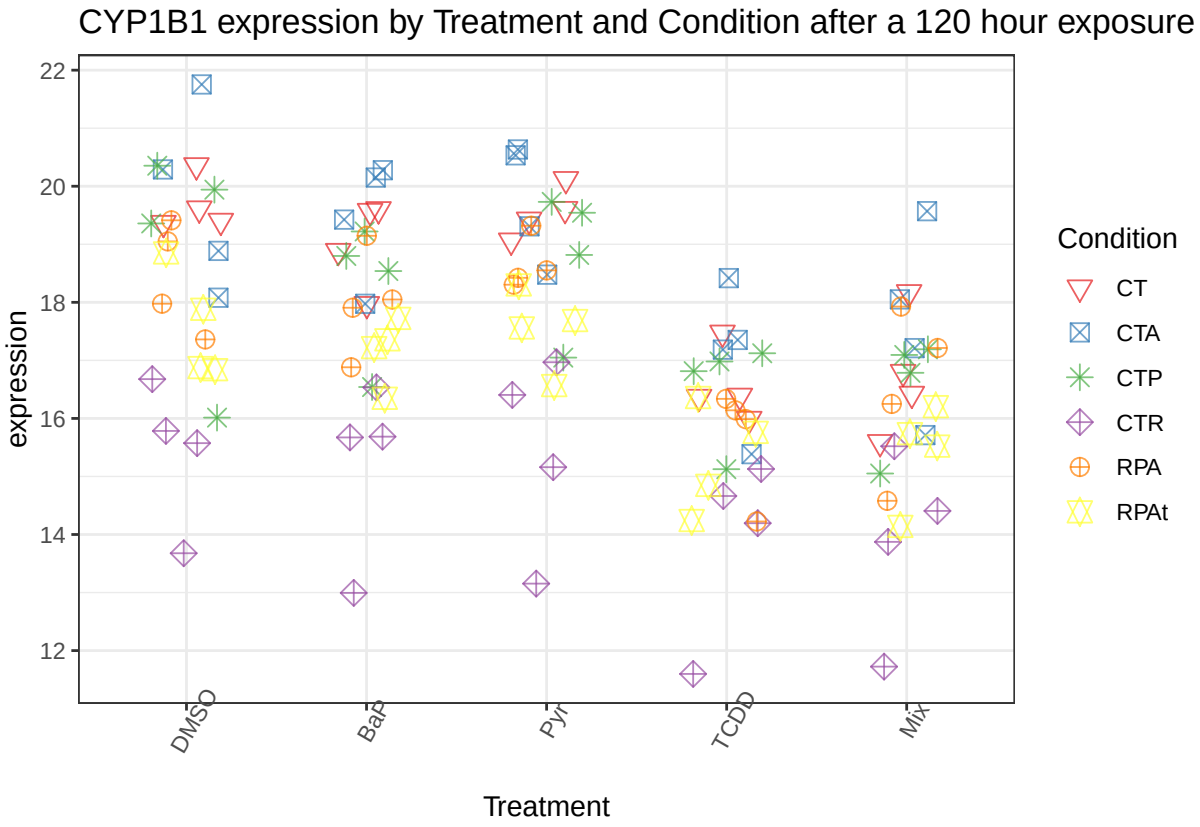
```
## Warning: Removed 10 rows containing missing values (geom_point).
```



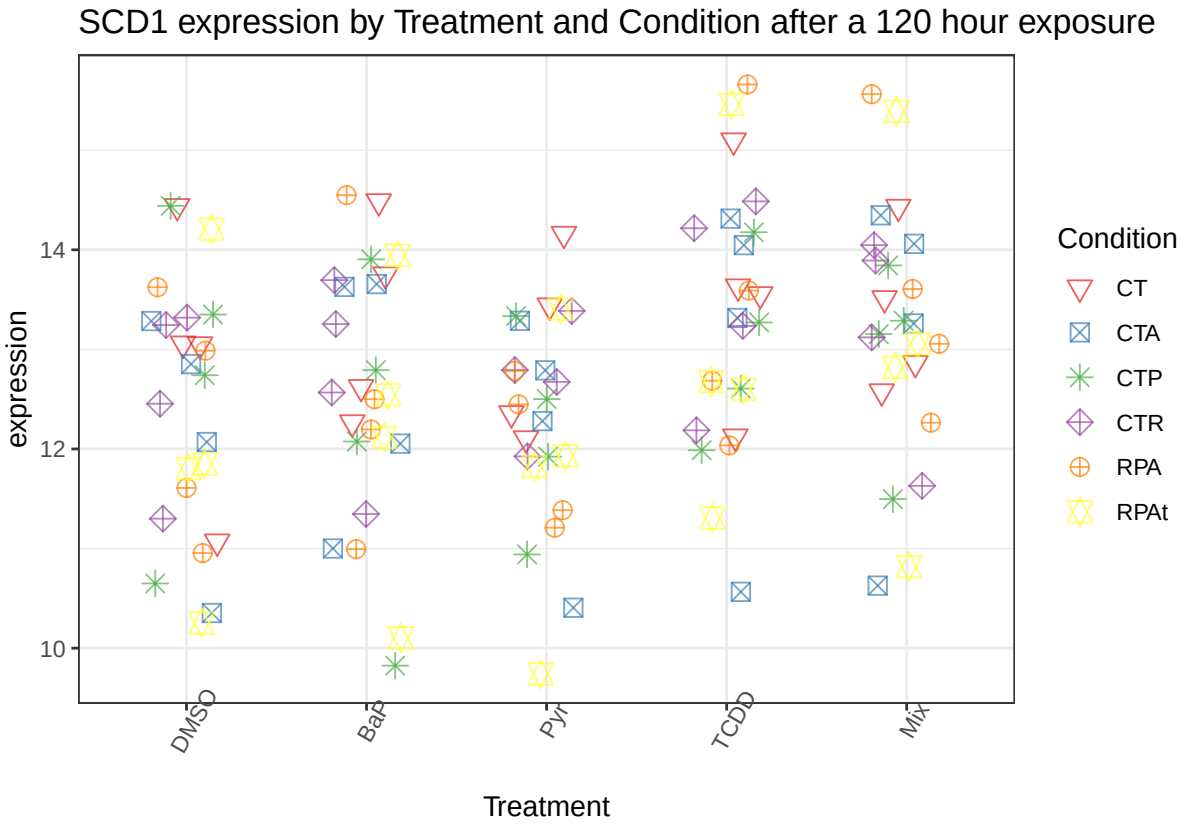

```
## Warning: Removed 10 rows containing missing values (geom_point).
```




```
## Warning: Removed 10 rows containing missing values (geom_point).
```




```
## Warning: Removed 10 rows containing missing values (geom_point).
```




```
## $AhR
## NULL
##
## $AhRR
## NULL
##
## $CYP1B1
## NULL
##
## $SCD1
## NULL
```

The gene AhR, which does not show any significant difference in its expression between any pair of treatments, has a steadily increasing expression when the treatment becomes stronger.

The genes AhRR and CYP1B1 have lower expression for TCDD and Mix compared to the other treatments.

SCD1 is a gene that is over-expressed in TCDD compared to Pyr.

2.3 Significant differences between Treatments (with a mixed model)

A more complete model, using all technical replicates in a mixed model with a random technical effect, was then performed to validate the previous findings.

2.3.1 After a 48 hour exposure

```
cells_48 <- read_delim("../data/qPCR_48.csv", "\t",
                       escape_double = FALSE, trim_ws = TRUE)
cells_48$Condition <- factor(cells_48$Condition,
                         levels = c("CT", "CTA", "CTR", "CTP", "RPA", "RPAt"))
cells_48$Experience <- factor(cells_48$Experience)
cells_48$Timepoint <- factor(cells_48$Timepoint, labels = c("48"))
cells_48$Plate <- factor(cells_48$Plate)
cells_48$Treatment <- factor(cells_48$Treatment,
                           levels = c("DMSO", "BaP", "Pyr", "TCDD",  "Mix"))

b48_cells <- cells_48 %>%
  dplyr::select(-c(Timepoint, Date))
b48_cells$Experience <- droplevels(b48_cells$Experience)
```

```
b48_cells <- b48_cells %>%
  group_by(Condition, Treatment, Experience, Plate, Gene) %>%
  summarise("Value" = mean(Value, na.rm = TRUE))
b48_cells <- b48_cells[ ,-1]

b48_cells_gene <- b48_cells %>% spread(Gene, Value)
```

```
perform_mixed_48 <- function(avar) {
  df <- b48_cells_gene[ ,c("Treatment", "Experience", avar)] %>% na.omit()
  expr <- substitute(paste("lme(", target,
                           "~ Treatment, random = ~1|Experience, data = df, method='REML')"),
                     list(target = avar))
  expr <- eval(expr)
  expr <- parse(text = expr)
  mm <- eval(expr)

  res_mm <- anova.lme(mm, type = "sequential", adjustSigma = FALSE)
  return(res_mm["Treatment", "p-value"])
}

celllines_mm <- sapply(names(b48_cells_gene[ ,-c(1:3)]), perform_mixed_48)
format(celllines_mm, scientific=TRUE, digits=3) %>% kable() %>%
  kable_styling(bootstrap_options = "striped")
```

	x
AhRR	2.62e-01
CYP1B1	2.96e-02
HK2	7.85e-01
HMOX	7.86e-01
MCT4	1.00e+00
NHE1	7.90e-01
NQO1	8.22e-01
NRF2	7.61e-01

This means that, at a 5% risk, the variables with significant differences are: CYP1B1. For these variables, post-hoc tests are performed to assess which cell types show significant differences:

```
signif_48 <- which(celllines_mm < 0.05)
celllines_postmm <- sapply(names(b48_cells_gene[ ,-c(1:3)])[signif_48], function(avar) {
  df <- b48_cells_gene[ ,c("Treatment", "Experience", avar)] %>% na.omit()
  expr <- substitute(paste("lme(", target,
                           "~ Treatment, random = ~1|Experience, data = df, method='REML')"),
                     list(target = avar))
  expr <- eval(expr)
  expr <- parse(text = expr)
  mm <- eval(expr)
  out <- summary(pairs(emmeans(mm, specs="Treatment")))[ ,c("contrast", "p.value")]
  contrasts <- as.character(out[ ,1])
  out <- out[ ,2]
  names(out) <- contrasts
  return(out)
})
format(celllines_postmm, scientific=TRUE, digits=3) %>% kable() %>%
  kable_styling(bootstrap_options = "striped")
```

	CYP1B1
DMSO - BaP	7.69e-01
DMSO - Pyr	8.97e-01
DMSO - TCDD	4.23e-02
DMSO - Mix	8.55e-02
BaP - Pyr	9.99e-01
BaP - TCDD	4.50e-01
BaP - Mix	6.30e-01
Pyr - TCDD	3.18e-01
Pyr - Mix	4.79e-01
TCDD - Mix	9.99e-01

which helps to conclude that significant differences are found for:

```
apply(celllines_postmm, 2, function(acol) names(which(acol < 0.05)))
```

```
##        CYP1B1
## "DMSO - TCDD"
```

2.3.2 After a 120 hour exposure

```
cells_120 <- read_delim("../data/qPCR_120.csv", "\t",
                        escape_double = FALSE, trim_ws = TRUE)
cells_120$Condition <- factor(cells_120$Condition,
                          levels = c("CT", "CTA", "CTR", "CTP", "RPA", "RPAt"))
cells_120$Experience <- factor(cells_120$Experience)
cells_120$Timepoint <- factor(cells_120$Timepoint, labels = c("48"))
cells_120$Plate <- factor(cells_120$Plate)
cells_120$Treatment <- factor(cells_120$Treatment,
                          levels = c("DMSO", "BaP", "Pyr", "TCDD",  "Mix"))

b120_cells <- cells_120 %>% dplyr::select(-c(Timepoint, Date))
b120_cells$Experience <- droplevels(b120_cells$Experience)
```

```
b120_cells <- b120_cells %>%
  group_by(Condition, Treatment, Experience, Plate, Gene) %>%
  summarise("Value" = mean(Value, na.rm = TRUE))
b120_cells <- b120_cells[,-1]
b120_cells_gene <- b120_cells %>% spread(Gene, Value)
```

```
perform_mixed_120 <- function(avar) {
  df <- b120_cells_gene[ ,c("Treatment", "Experience", avar)] %>% na.omit()
  expr <- substitute(paste("lme(", target,
                           "~ Treatment, random = ~1|Experience, data = df, method='REML')"),
                     list(target = avar))
  expr <- eval(expr)
  expr <- parse(text = expr)
  mm <- eval(expr)

  res_mm <- anova.lme(mm, type = "sequential", adjustSigma = FALSE)
  return(res_mm["Treatment", "p-value"])
}

celllines_mm <- sapply(names(b120_cells_gene[ ,-c(1:3,16)]), perform_mixed_120)
format(celllines_mm, scientific=TRUE, digits=3) %>% kable() %>%
  kable_styling(bootstrap_options = "striped")
```

	x
ACO1	4.75e-01
AhR	1.07e-01
AhRR	1.24e-10
ATP5IF1	7.09e-01
CAT	8.98e-01
CYP1B1	8.19e-08
G6PD	9.60e-01
HK2	2.69e-01
LPCAT	7.29e-01
MCT4	9.80e-01
MFN2	8.47e-01
ND1	8.62e-01
NQO1	8.95e-01
NRF2	2.10e-01
PRDX1	9.68e-01
SCD1	1.45e-02
TFAM	9.59e-01

This means that, at a 5% risk, the variables with significant differences are: AhRR, CYP1B1, SCD1. For these variables, post-hoc tests are performed to assess which cell types show significant differences:

```
signif_120 <- which(celllines_mm < 0.05)
celllines_postmm <- sapply(names(b120_cells_gene[ ,-c(1:3,16)])[signif_120], function(avar) {
  df <- b120_cells_gene[ ,c("Treatment", "Experience", avar)] %>% na.omit()
  expr <- substitute(paste("lme(", target,
                           "~ Treatment, random = ~1|Experience, data = df, method='REML')"),
                     list(target = avar))
  expr <- eval(expr)
  expr <- parse(text = expr)
```

```
  mm <- eval(expr)
  out <- summary(pairs(emmeans(mm, specs="Treatment")))[ ,c("contrast", "p.value")]
  contrasts <- as.character(out[ ,1])
  out <- out[ ,2]
  names(out) <- contrasts
  return(out)
})
format(celllines_postmm, scientific=TRUE, digits=3) %>% kable() %>%
  kable_styling(bootstrap_options = "striped")
```

	AhRR	CYP1B1	SCD1
DMSO - BaP	9.76e-01	8.88e-01	9.97e-01
DMSO - Pyr	9.95e-01	1.00e+00	9.90e-01
DMSO - TCDD	1.83e-06	3.24e-05	1.37e-01
DMSO - Mix	2.14e-05	2.61e-04	2.27e-01
BaP - Pyr	8.68e-01	9.09e-01	9.27e-01
BaP - TCDD	2.30e-05	1.13e-03	2.70e-01
BaP - Mix	2.28e-04	6.72e-03	4.04e-01
Pyr - TCDD	3.31e-07	4.05e-05	4.51e-02
Pyr - Mix	4.24e-06	3.22e-04	8.41e-02
TCDD - Mix	9.79e-01	9.85e-01	9.99e-01

which helps to conclude that significant differences are found for:

```
apply(celllines_postmm, 2, function(acol) names(which(acol < 0.05)))
```

```
## $AhRR
## [1] "DMSO - TCDD" "DMSO - Mix"  "BaP - TCDD"  "BaP - Mix"   "Pyr - TCDD"
## [6] "Pyr - Mix"
##
## $CYP1B1
## [1] "DMSO - TCDD" "DMSO - Mix"  "BaP - TCDD"  "BaP - Mix"   "Pyr - TCDD"
## [6] "Pyr - Mix"
##
## $SCD1
## [1] "Pyr - TCDD"
```

2.3.3 Conclusion

The results of the mixed models are identical to the conclusions obtained with the ANOVA tests.

2.4 Conclusion

The two following tables are constructed as follows:

- The upper parts of the tables are composed by the results of Kruskal-Wallis tests. The lower parts of the tables are composed by the results of the ANOVA tests (and of the mixed models).
- An empty box means that there is no significant difference between the couple of treatments whatever the gene
- A completed box means that the genes present in this box show significant differences in expression between the 2 corresponding treatments.

48h	DMSO	BaP	Pyr	TCDD	Mix
DMSO	-	-	-	-	-
BaP	-	-	-	-	-
Pyr	-	-	-	-	-
TCDD	-	CYP1B1	-	-	-
Mix	-	-	-	-	-

120h	DMSO	BaP	Pyr	TCDD	Mix
DMSO	-	-	-	AhRR, CYP1B1	AhRR, CYP1B1
BaP	AhRR, CYP1B1	-	-	AhRR, CYP1B1	AhRR, CYP1B1
Pyr	-	-	-	AhRR, CYP1B1	AhRR, CYP1B1
TCDD	-	AhRR, CYP1B1	AhRR, CYP1B1, SCD1	-	-
Mix	AhRR, CYP1B1	AhRR, CYP1B1	AhRR, CYP1B1	-	-

The CYP1B1 gene always has a difference in expression between TCDD and at least one of the three weakest treatment but, as for all the other genes, the differences are much stronger at 120h than at 48h.

The two treatments TCDD and Mix have very strong similarities and never show a significant difference, whatever the gene. The same holds for DMSO and Pyr.

3. Which genes are differentially expressed between the cell types, once the treatment effect estimated?

This section addresses the issue of jointly estimating the cell type and treatment effects in a linear model with two additive factors: expr ~ cell type + treatment. The treatment effect and the cell type effect are then sucessively analyzed in this model.

3.1 After a 48 hour exposure

3.1.1 Normality tests

We first perform a Shapiro normality test on treatments and cell types by gene.

```
pvals_Cond_48 <- apply(cell_48[ ,-c(1:3)], 2, function(agene) {
  tapply(agene, cell_48$Condition, function(agt) shapiro.test(agt)$p.value)
})
format(pvals_Cond_48, scientific=TRUE, digits=3) %>% kable() %>% kable_styling(bootstrap_options = "str
  add_header_above(c(" " = 1, "p-values" = 8))
```

	p-values							
	AhRR	CYP1B1	HK2	HMOX	MCT4	NHE1	NQO1	NRF2
CT	9.36e-02	3.64e-01	3.24e-01	1.13e-02	4.59e-02	2.77e-02	1.48e-01	3.46e-01
CTA	3.70e-02	1.25e-01	1.11e-01	4.69e-01	2.15e-02	2.62e-02	2.01e-01	4.53e-02
CTP	9.46e-03	1.41e-01	2.88e-01	6.59e-01	4.11e-02	2.89e-03	3.18e-02	3.85e-01
CTR	6.07e-02	7.33e-01	1.71e-02	2.26e-01	1.50e-02	4.90e-03	2.76e-03	2.17e-01
RPA	2.20e-01	2.55e-01	8.00e-01	9.84e-01	1.43e-02	5.72e-02	4.28e-01	3.23e-02
RPAt	1.09e-02	2.68e-01	7.03e-01	3.40e-01	1.17e-01	4.40e-03	3.65e-02	1.61e-01

Conclusion:

- For the effects of the treatment on gene expressions, the normality can not be rejected for HK2, HMOX, NQO1 and NRF2 and the normality is rejected for AhRR, CYP1B1, MCT4 and NHE1.
- For the effects of the cell type on gene expressions, the normality can not be rejected for CYP1B1 and HMOX and the normality is rejected for AhRR, HK2, MCT4, NHE1, NQO1 and NRF2.

3.1.2 Two-way ANOVA

```
anova_48 <- apply(cell_48[ ,-c(1:3)], 2, function(agene) {
  summary(aov(agene ~ cell_48$Condition + cell_48$Treatment))[[1]][1:2,5]
})
rownames(anova_48) = c("Cell type", "Treatment")
format(anova_48, scientific=TRUE, digits=3) %>%
kable() %>%
  kable_styling(latex_options = c("HOLD_position")) %>%
  column_spec(1:8)
```

	AhRR	CYP1B1	HK2	HMOX	MCT4	NHE1	NQO1	NRF2
Cell type	9.10e-02	1.56e-08	6.75e-14	1.64e-04	9.85e-01	4.67e-01	6.19e-04	2.01e-04
Treatment	2.36e-01	2.19e-03	4.34e-01	6.98e-01	1.00e+00	7.80e-01	7.55e-01	6.70e-01

Treatment The number of genes for which the null hypothesis is rejected for the treatments is 1 (at 5%) and these genes are: CYP1B1.

For these genes, post-hoc tests are performed:

```
selecttreat <- which(anova_48[2,] < 0.05)
anova_posthoc_48 <- apply(cell_48[ ,-c(1:3)][ ,selecttreat], 2, function(agene) {
  TukeyHSD(aov(agene ~ cell_48$Treatment + cell_48$Condition))[[1]][ ,4]
})
format(anova_posthoc_48, scientific=TRUE, digits=3) %>% kable() %>% kable_styling(bootstrap_options = "
  add_header_above(c(" " = 1, "p-values" = length(selecttreat)))
```

	p-values
	CYP1B1
BaP-DMSO	5.67e-01
Pyr-DMSO	7.80e-01
TCDD-DMSO	4.13e-03
Mix-DMSO	1.27e-02
Pyr-BaP	9.97e-01
TCDD-BaP	2.07e-01
Mix-BaP	3.85e-01
TCDD-Pyr	1.12e-01
Mix-Pyr	2.33e-01
Mix-TCDD	9.96e-01

For each gene, the number of pairs and the pairs of treatments that show significant differences in expression is:

```
sapply(1:ncol(anova_posthoc_48), function(ind) {
  cat(paste0("For gene: ", colnames(anova_posthoc_48)[ind], "\n"))
```

```
  cat(paste0("Number of significant pairs: ",
             sum(anova_posthoc_48[ ,ind] < 0.05), "\n"))
  cat(names(which(anova_posthoc_48[ ,ind] < 0.05)), "\n\n")
})
```

```
## For gene: CYP1B1
## Number of significant pairs: 2
## TCDD-DMSO Mix-DMSO
```

```
## [[1]]
## NULL
```

Compared to a direct ANOVA with a treatment effect, the results show an additional pair of treatments that have a significant expression in CYP1B1: TCDD versus DMSO.

Cell types The number of genes for which the null hypothesis is rejected for the cell types is 5 (at 5%) and these genes are: CYP1B1, HK2, HMOX, NQO1, NRF2.

For these genes, post-hoc tests are performed:

```
selectcell <- which(anova_48[1,] < 0.05)
anova_posthoc_48 <- apply(cell_48[ ,-c(1:3)][ ,selectcell], 2, function(agene) {
  TukeyHSD(aov(agene ~ cell_48$Treatment + cell_48$Condition))[[2]][ ,4]
})
format(anova_posthoc_48, scientific=TRUE, digits=3) %>% kable() %>%
  kable_styling(bootstrap_options = "striped") %>%
  add_header_above(c(" " = 1, "p-values" = length(selectcell)))
```

	p-values				
	CYP1B1	HK2	HMOX	NQO1	NRF2
CTA-CT	1.83e-01	9.68e-01	9.65e-01	9.78e-01	7.87e-01
CTP-CT	1.06e-01	5.81e-01	3.28e-01	9.77e-01	9.95e-01
CTR-CT	6.85e-04	7.15e-07	9.79e-01	6.41e-01	2.08e-01
RPA-CT	6.31e-01	1.47e-01	2.58e-02	3.97e-02	2.50e-01
RPAt-CT	1.85e-02	0.00e+00	1.50e-02	9.79e-01	7.61e-01
CTP-CTA	6.30e-05	9.57e-01	8.11e-01	1.00e+00	9.77e-01
CTR-CTA	5.02e-08	1.87e-05	6.50e-01	2.21e-01	7.25e-03
RPA-CTA	2.54e-03	5.54e-01	1.84e-01	2.14e-01	9.45e-01
RPAt-CTA	3.68e-06	2.86e-10	1.23e-01	7.00e-01	1.00e-01
CTR-CTP	5.97e-01	6.70e-04	7.74e-02	2.29e-01	6.93e-02
RPA-CTP	8.85e-01	9.68e-01	8.92e-01	2.40e-01	5.84e-01
RPAt-CTP	9.90e-01	2.57e-08	8.09e-01	7.03e-01	4.38e-01
RPA-CTR	7.87e-02	7.75e-03	2.83e-03	2.72e-04	3.30e-04
RPAt-CTR	9.13e-01	1.21e-01	1.49e-03	9.63e-01	9.31e-01
RPAt-RPA	5.18e-01	4.93e-07	1.00e+00	4.74e-03	8.52e-03

For each gene, the number of pairs and the pairs of cell types that show significant differences in expression is:

```
sapply(1:ncol(anova_posthoc_48), function(ind) {
  cat(paste0("For gene: ", colnames(anova_posthoc_48)[ind], "\n"))
  cat(paste0("Number of significant pairs: ",
             sum(anova_posthoc_48[ ,ind] < 0.05), "\n"))
  cat(names(which(anova_posthoc_48[ ,ind] < 0.05)), "\n\n")
})
```

```
## For gene: CYP1B1
## Number of significant pairs: 6
## CTR-CT RPAt-CT CTP-CTA CTR-CTA RPA-CTA RPAt-CTA
##
## For gene: HK2
## Number of significant pairs: 8
## CTR-CT RPAt-CT CTR-CTA RPAt-CTA CTR-CTP RPAt-CTP RPA-CTR RPAt-RPA
##
## For gene: HMOX
## Number of significant pairs: 4
## RPA-CT RPAt-CT RPA-CTR RPAt-CTR
##
## For gene: NQO1
## Number of significant pairs: 3
## RPA-CT RPA-CTR RPAt-RPA
##
## For gene: NRF2
## Number of significant pairs: 3
## CTR-CTA RPA-CTR RPAt-RPA
```

```
## [[1]]
## NULL
##
## [[2]]
## NULL
##
## [[3]]
## NULL
##
## [[4]]
## NULL
##
## [[5]]
## NULL
```

3.1.3 Conclusion

Treatment The following table represents all genes for which at least one global test is significant

- first column: which test(s) is (are) significant?
- second column: for which treatment(s) is normality rejected for this gene?

Treatment	Tests giving these results	Treatment(s) whose normality Shapiro rejects
CYP1B1	ANOVA 2 factor test	TCDD

The following tables are constructed as follows:

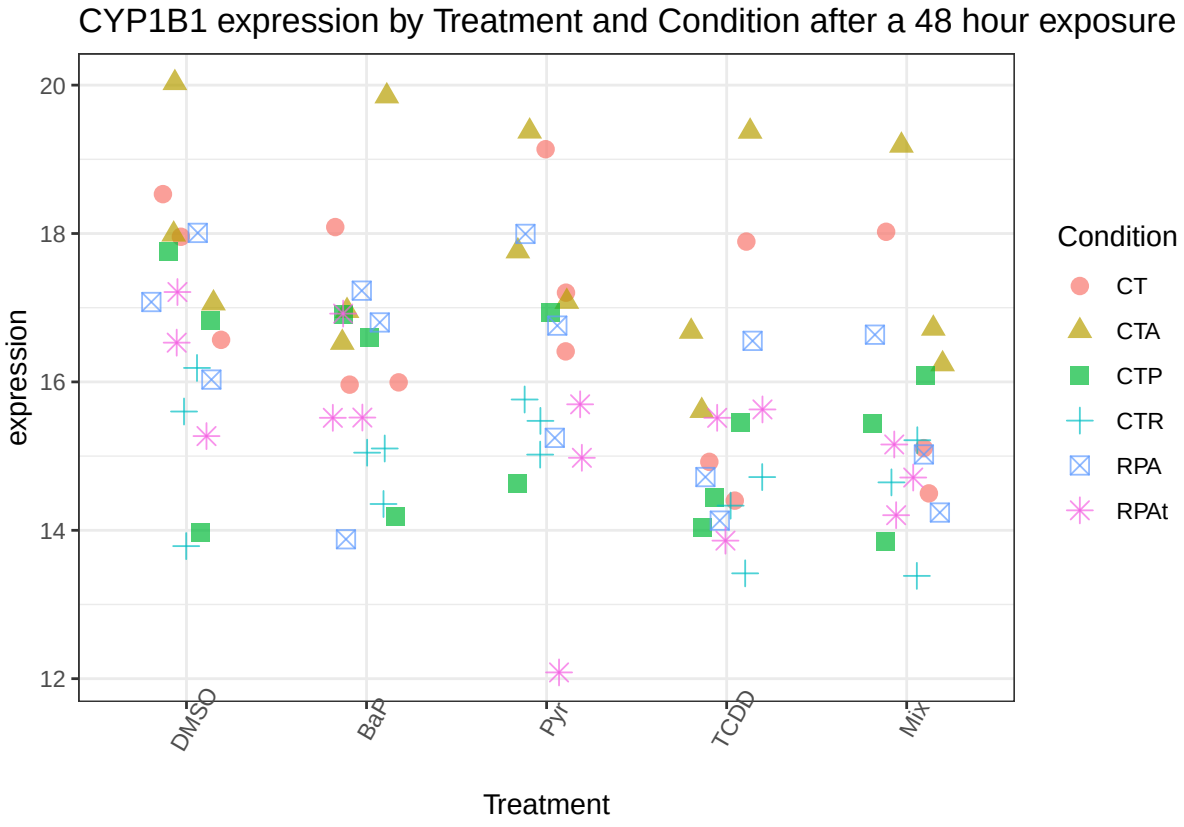
- the upper parts of the tables are composed by the results of the post-hoc ANOVA test;
- A star means that the contrast between the two treatments is significant (p-value < 0.05);

CYP1B1	DMSO	BaP	Pyr	TCDD	Mix
DMSO	-			*	*

CYP1B1	DMSO	BaP	Pyr	TCDD	Mix
BaP	-	-			
Pyr	-	-	-		
TCDD	-	-	-	-	
Mix	-	-	-	-	-

There are a minor difference with the previous test results for the gene CYP1B1. In the previous test, one couple of treatments DMSO/TCDD is found to have a significant difference in expression for this gene with parametric tests (but the expression of CYP1B1 significantly deviates from normality for TCDD) and, for the non-parametric tests, the pairs of treatments that deviate the most from the null hypothesis are DMSO/TCDD and DMSO/Mix. However, in the model with an additive effect, two couples of treatments DMSO/TCDD and Mix/DMSO are found to have a significant difference in expression for this gene, which is consistant with what was obtained by the non-parametric tests.

```
all_signif <- names(cell_48[ ,-c(1:3)][ ,selecttreat])
sapply(all_signif, function(agene) {
  df <- data.frame(cell_48$Treatment, cell_48[ ,agene], cell_48$Condition)
  names(df) <- c("Treatment", "expr", "Condition")
  p <- ggplot(df, aes(x = Treatment, y = expr, colour = Condition, shape = Condition)) +
  geom_jitter(alpha = 0.7, width = 0.2, size = 3) + theme_bw() +
    theme(axis.text.x = element_text(angle = 60)) + ylab("expression") +
    ggtitle(paste0(agene,
                   " expression by Treatment and Condition after a 48 hour exposure"))
  print(p)
  invisible(NULL)
})
```



```
## $CYP1B1
## NULL
```

- CYP1B1 is significantly over-expressed in DMSO (which is the control condition) compared to TCDD and Mix (which are the two most aggressive treatments).

Cell types The following table represents all genes for which at least one global test is significant

- first column: which test(s) is (are) significant?
- second column: for which cell type(s) is normality rejected for this gene?

Cell types	Tests giving these results	Cell Type(s) whose normality Shapiro rejects
CYP1B1	two-factor ANOVA test	-
HK2	two-factor ANOVA test	CTR
HMOX	two-factor ANOVA test	CT
NQO1	two-factor ANOVA test	CTP, CTR, CTRPAt
NRF2	two-factor ANOVA test	CTA, CTRPA

The following tables are constructed as follows:

- the upper parts of the tables are composed by the results of the post-hoc two-factor ANOVA test;
- A star means that the contrast between the two cell types is significant (p-value < 0.05);

CYP1B1	CT	CTA	CTP	CTR	CTRPA	CTRPAt
CT	-			*		*
CTA	-	-	*	*	*	*
CTP	-	-	-			
CTR	-	-	-	-		
CTRPA	-	-	-	-	-	
CTRPAt	-	-	-	-	-	-

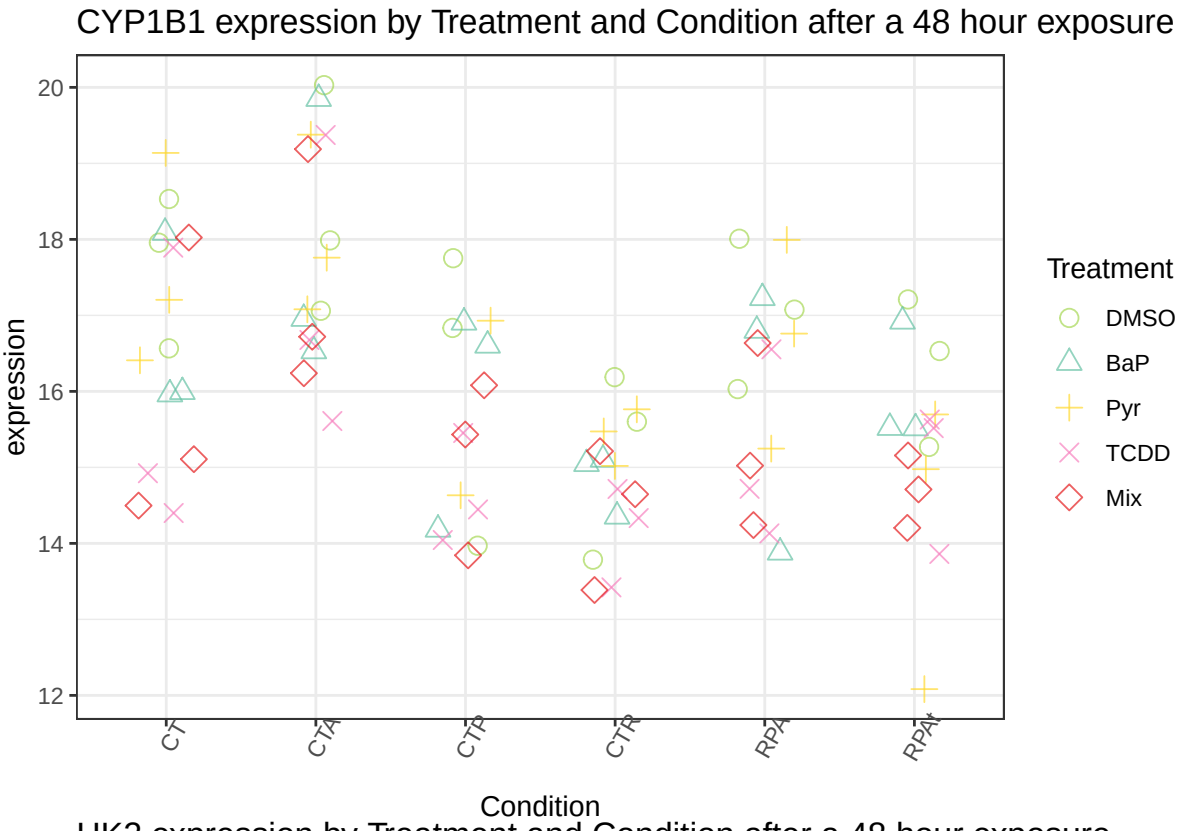
HK2	CT	CTA	CTP	CTR	CTRPA	CTRPAt
CT	-			*		*
CTA	-	-		*		*
CTP	-	-	-	*		*
CTR	-	-	-	-	*	
CTRPA	-	-	-	-	-	*
CTRPAt	-	-	-	-	-	-

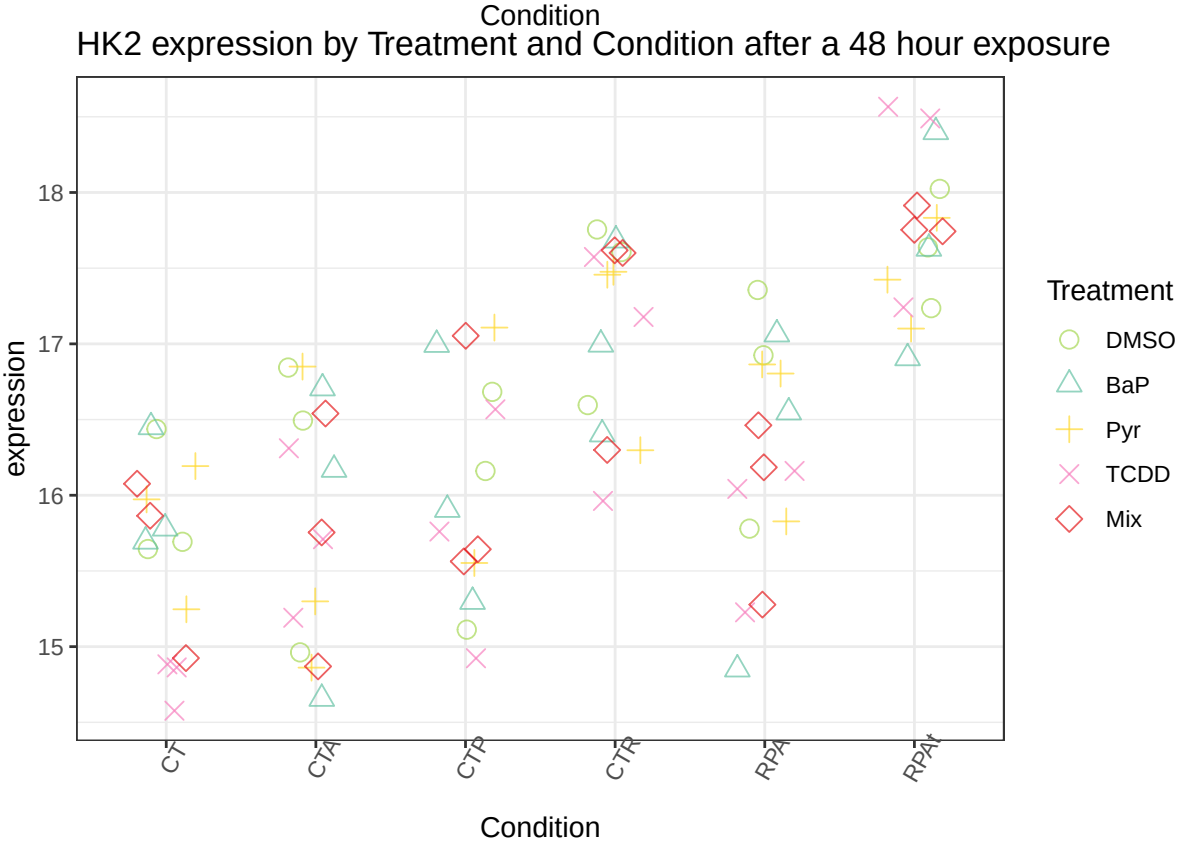
HMOX	CT	CTA	CTP	CTR	CTRPA	CTRPAt
CT	-				*	*
CTA	-	-				
CTP	-	-	-			
CTR	-	-	-	-	*	*
CTRPA	-	-	-	-	-	
CTRPAt	-	-	-	-	-	-

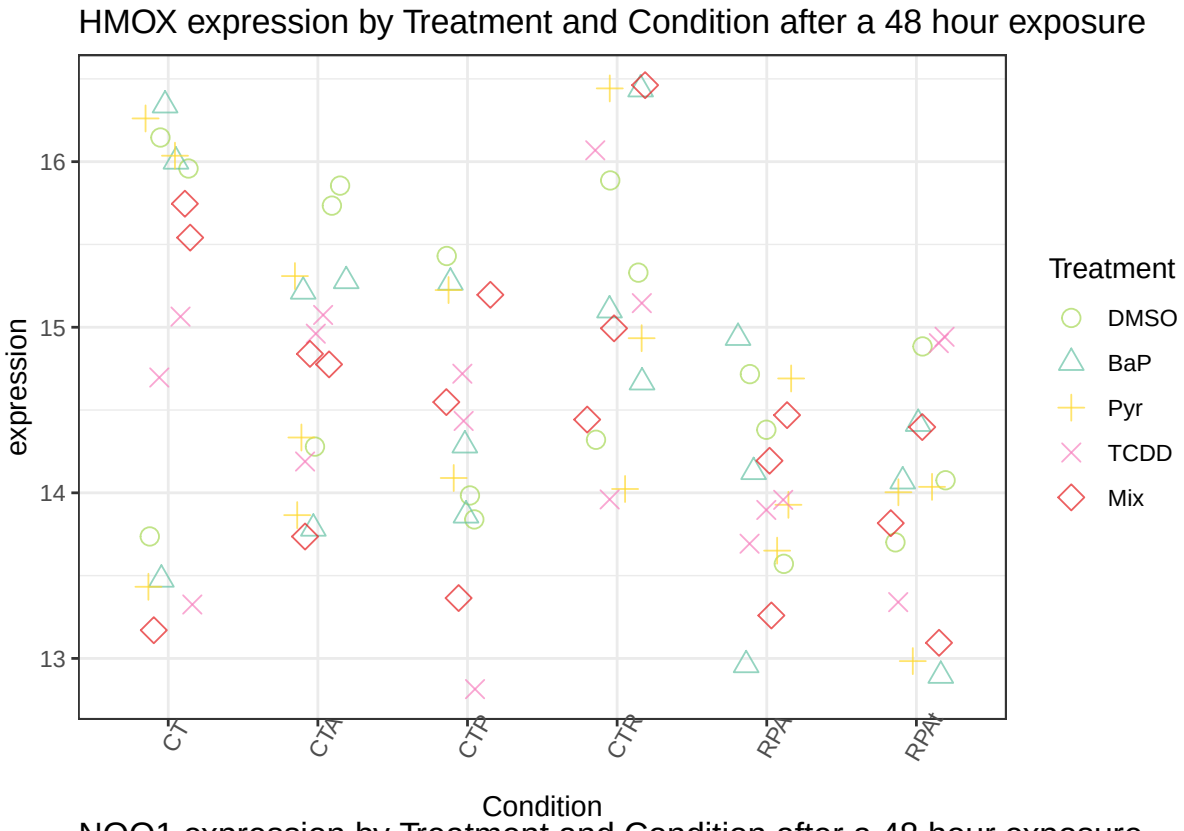
NQO1	CT	CTA	CTP	CTR	CTRPA	CTRPAt
CT	-				*	
CTA	-	-				
CTP	-	-	-			
CTR	-	-	-	-	*	
CTRPA	-	-	-	-	-	*
CTRPAt	-	-	-	-	-	-

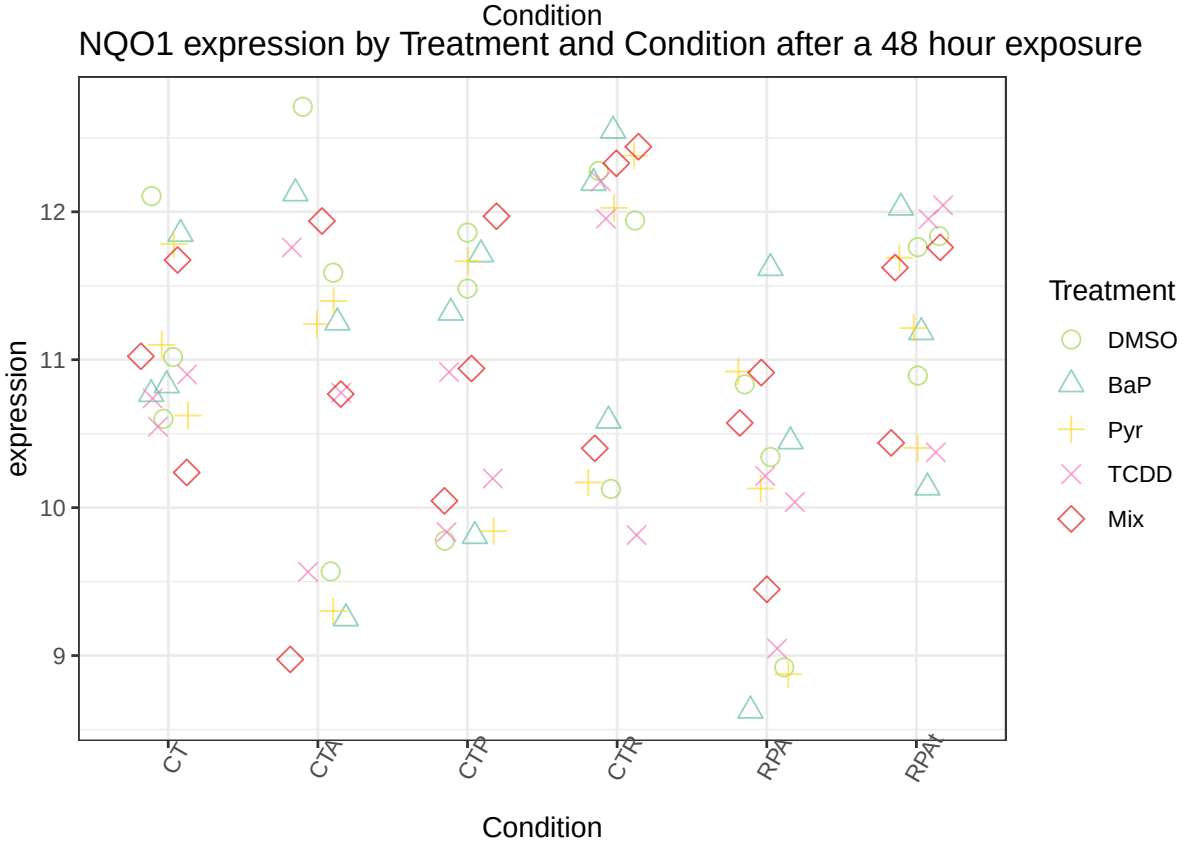
NRF2	CT	CTA	CTP	CTR	CTRPA	CTRPAt
CT	-					
CTA	-	-		*		
CTP	-	-	-			
CTR	-	-	-	-	*	
CTRPA	-	-	-	-	-	*
CTRPAt	-	-	-	-	-	-

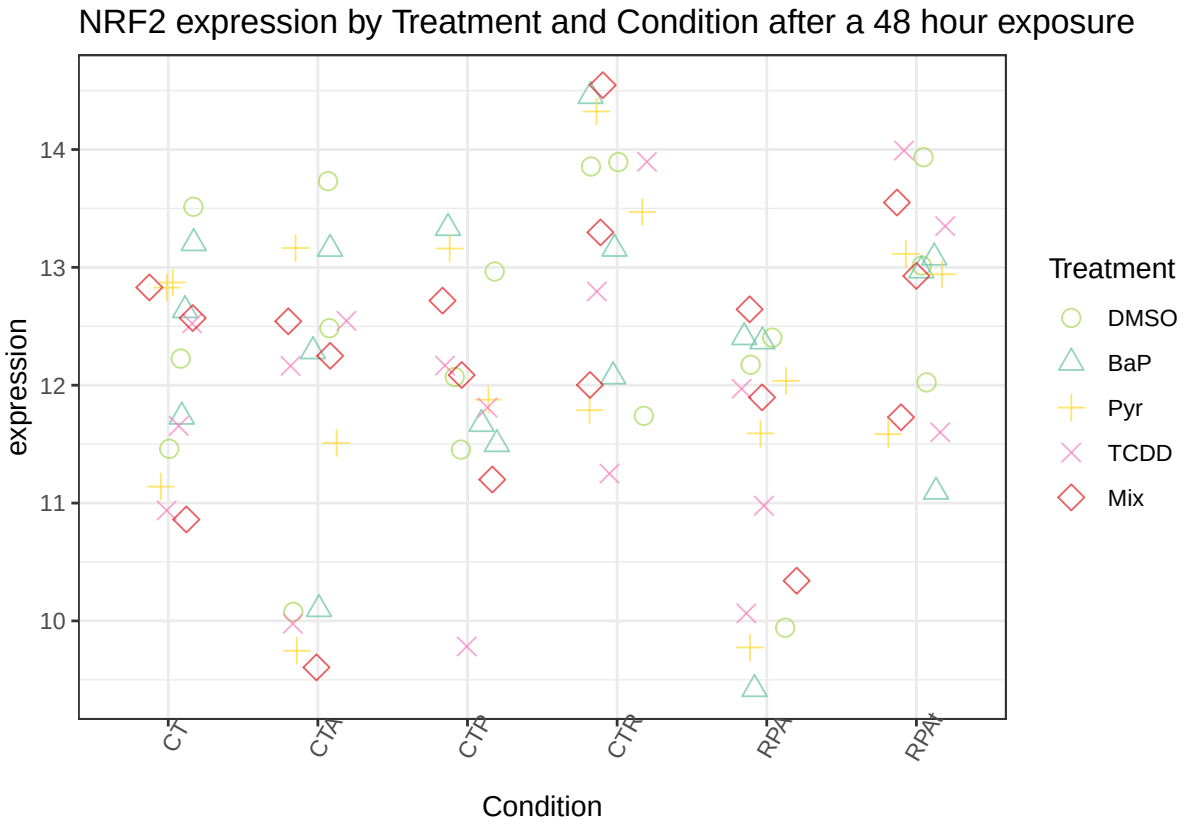
```
all_signif <- names(cell_48[ ,-c(1:3)][ ,selectcell])
sapply(all_signif, function(agene) {
  df <- data.frame(cell_48$Treatment, cell_48[ ,agene], cell_48$Condition)
  names(df) <- c("Treatment", "expr", "Condition")
  p <- ggplot(df, aes(x = Condition, y = expr, colour = Treatment, shape =Treatment)) +
    geom_jitter(alpha = 0.7, width = 0.2, size = 3) +
    theme_bw() +
    theme(axis.text.x = element_text(angle = 60)) + ylab("expression") +
    ggtitle(paste0(agene,
                   " expression by Treatment and Condition after a 48 hour exposure")) +
    scale_color_manual(values = palettetreatment) +
    scale_shape_manual(values = symboltreatment)
  print(p)
  invisible(NULL)
})
```


CYP1B1 expression by Treatment and Condition after a 48 hour exposure
expression
20
18
16
14
12
CT
CTA
CTP
CTR
RPA
RPAt
Condition
Treatment
DMSO
BaP
Pyr
TCDD
Mix


HK2 expression by Treatment and Condition after a 48 hour exposure
expression
18
17
16
15
CT
CTA
CTP
CTR
RPA
RPAt
Condition
Treatment
DMSO
BaP
Pyr
TCDD
Mix


HMOX expression by Treatment and Condition after a 48 hour exposure
expression
16
15
14
13
CT
CTA
CTP
CTR
RPA
RPAt
Condition
Treatment
DMSO
BaP
Pyr
TCDD
Mix


NQO1 expression by Treatment and Condition after a 48 hour exposure
expression
12
11
10
9
CT
CTA
CTP
CTR
RPA
RPAt
Condition
Treatment
DMSO
BaP
Pyr
TCDD
Mix




```
## $CYP1B1
## NULL
##
## $HK2
## NULL
##
## $HMOX
## NULL
##
## $NQO1
## NULL
##
## $NRF2
## NULL
```

- CYP1B1 is significantly over-expressed in CTA compared to CTP, CTR, CTRPA and CTRPAt. Similarly, CT is significantly over-expressed compared to CTR and CTRPAt.
- HK2 is significantly over-expressed in CTR and CTRPAt compared to all the other cell types (except for the difference between CTRPAt and CTRPA that is not found to be significant).
- HMOX is significantly over-expressed in CT and CTR compared to CTRPA and CTRPAt.
- NQO1 is significantly under-expressed in CTRPA compared to CT, CTR and CTRPAt.
- NRF2 is significantly over-expressed in CTR compared to CTA and CTRPA and under-expressed in CTRPA compared to CTRPAt. Note that the conclusion related to CTA is somehow flawn by the fact that this cell type has a bimodal distribution.

3.2 After a 120 hour exposure

3.2.1 Normality tests

We first perform a Shapiro normality test on treatments and cell types by gene

```
pvals_Cond_120 <- apply(cell_120[ ,-c(1:3)], 2, function(agene) {
  tapply(agene, cell_120$Condition, function(agt) shapiro.test(agt)$p.value)
})

format(pvals_Cond_120[,1:8], scientific=TRUE, digits=3) %>% kable() %>% kable_styling(bootstrap_options
  add_header_above(c(" " = 1, "p-values" = 8))
```

	p-values							
	ACO1	AhR	AhRR	ATP5IF1	CAT	CYP1B1	G6PD	HK2
CT	4.21e-01	4.14e-01	2.53e-01	8.84e-03	7.72e-02	2.23e-02	2.92e-01	6.02e-01
CTA	4.45e-02	1.95e-01	8.07e-01	1.24e-01	7.91e-01	8.05e-01	2.10e-01	4.54e-01
CTP	3.20e-01	8.46e-03	3.03e-01	3.73e-02	3.16e-01	1.89e-01	4.67e-01	9.40e-01
CTR	2.86e-01	1.61e-02	2.16e-01	1.49e-03	4.02e-02	2.68e-01	1.51e-01	2.25e-01
RPA	7.82e-03	2.76e-01	5.26e-01	2.94e-03	2.07e-01	1.99e-01	2.19e-01	8.18e-01
RPAt	9.08e-03	2.40e-01	5.91e-01	2.46e-02	3.97e-02	7.90e-01	1.30e-01	8.29e-01

```
format(pvals_Cond_120[,9:16], scientific=TRUE, digits=3) %>% kable() %>% kable_styling(bootstrap_option
  add_header_above(c(" " = 1, "p-values" = 8))
```

	p-values							
	LPCAT	MCT4	MFN2	ND1	NHE1	NQO1	NRF2	PRDX1
CT	5.56e-01	2.76e-02	5.51e-01	4.44e-01	8.82e-02	3.98e-02	3.66e-02	1.15e-06
CTA	3.76e-03	7.85e-03	7.10e-03	1.27e-02	5.62e-02	1.94e-01	1.96e-01	8.89e-01
CTP	8.65e-01	4.08e-02	8.37e-02	3.15e-01	9.06e-02	2.85e-01	3.65e-01	7.20e-02
CTR	7.70e-05	2.03e-02	3.11e-03	7.65e-04	5.52e-03	5.64e-03	2.33e-02	2.26e-02
RPA	2.02e-01	6.33e-03	1.50e-02	6.18e-03	2.68e-01	6.78e-02	1.26e-02	4.44e-01
RPAt	7.20e-03	9.58e-02	7.29e-03	2.84e-03	1.58e-01	4.45e-03	4.18e-03	9.77e-02

```
format(pvals_Cond_120[,17:18], scientific=TRUE, digits=3) %>% kable() %>% kable_styling(bootstrap_optio
  add_header_above(c(" " = 1, "p-values" = 2))
```

	p-values	
	SCD1	TFAM
CT	9.34e-01	3.56e-01
CTA	3.30e-02	1.48e-02
CTP	5.07e-01	1.47e-01
CTR	6.49e-01	8.13e-03
RPA	1.84e-01	2.29e-02
RPAt	6.78e-01	6.72e-03

Conclusion:

- For the effects on the treatment on gene expressions, the normality was accepted for ACO1, AhRR, CAT, LPCAT, MCT4, NRF2, SCD1 and TFAM and the normality was not accepted for AhR, ATP5IF1, CYP1B1, G6PD, HK2, MFN2, ND1, NHE1, NQO1 and PRDX1.

- For the effects on the cell type on gene expressions, the normality was accepted for AhRR, G6PD and HK2 and the normality was not accepted for ACO1, AhR, ATP5IF1, CAT, CYP1B1, LPCAT, MCT4, MFN2, ND1, NHE1, NQO1, NRF2, PRDX1, SCD1 and TFAM.

3.2.2 Two-way ANOVA

We now perform a 2-way ANOVA on the assumptions described in section 3.1.2.

```
anova_120 <- apply(cell_120[ ,-c(1:3)], 2, function(agene) {
  summary(aov(agene ~ cell_120$Condition + cell_120$Treatment))[[1]][1:2,5]
})
rownames(anova_120) = c("Cell type", "Treatment")
format(anova_120[,1:8], scientific=TRUE, digits=3) %>%
kable() %>%
  kable_styling(latex_options = c("HOLD_position")) %>%
  column_spec(1:8)
```

	ACO1	AhR	AhRR	ATP5IF1	CAT	CYP1B1	G6PD	HK2
Cell type	8.51e-09	2.79e-03	1.02e-03	6.52e-01	1.16e-07	2.07e-21	2.92e-05	2.47e-17
Treatment	2.50e-01	7.41e-02	5.18e-12	7.16e-01	7.98e-01	6.33e-17	9.41e-01	2.44e-02

```
format(anova_120[,9:16], scientific=TRUE, digits=3) %>%
kable() %>%
  kable_styling(latex_options = c("HOLD_position")) %>%
  column_spec(1:8)
```

	LPCAT	MCT4	MFN2	ND1	NHE1	NQO1	NRF2	PRDX1
Cell type	2.10e-05	8.86e-01	9.23e-01	1.79e-01	4.29e-01	6.56e-07	1.51e-04	4.73e-03
Treatment	5.77e-01	9.81e-01	8.59e-01	7.96e-01	9.82e-01	8.04e-01	1.05e-01	9.70e-01

```
format(anova_120[,17:18], scientific=TRUE, digits=3) %>%
kable() %>%
  kable_styling(latex_options = c("HOLD_position")) %>%
  column_spec(1:2)
```

	SCD1	TFAM
Cell type	3.29e-01	5.76e-02
Treatment	1.38e-02	9.76e-01

Treatments The number of genes for which the null hypothesis is rejected for the treatments is 4 (at 5%) and these genes are: AhRR, CYP1B1, HK2, SCD1.

For these genes, post-hoc tests are performed:

```
selecttreat <- which(anova_120[2,] < 0.05)
anova_posthoc_120 <- apply(cell_120[ ,-c(1:3)][ ,selecttreat], 2, function(agene) {
  TukeyHSD(aov(agene ~ cell_120$Treatment + cell_120$Condition))[[1]][,4]
})
format(anova_posthoc_120, scientific=TRUE, digits=3) %>% kable() %>% kable_styling(bootstrap_options =
  add_header_above(c(" " = 1, "p-values" = length(selecttreat)))
```

	p-values			
	AhRR	CYP1B1	HK2	SCD1
BaP-DMSO	9.70e-01	5.89e-01	8.33e-01	9.97e-01
Pyr-DMSO	9.94e-01	1.00e+00	9.73e-01	9.90e-01
TCDD-DMSO	2.72e-07	1.80e-11	9.13e-02	1.35e-01
Mix-DMSO	4.31e-06	1.62e-09	5.14e-02	2.23e-01
Pyr-BaP	8.36e-01	6.47e-01	9.92e-01	9.26e-01
TCDD-BaP	4.67e-06	2.60e-08	5.75e-01	2.66e-01
Mix-BaP	6.21e-05	1.64e-06	4.29e-01	4.00e-01
TCDD-Pyr	4.00e-08	2.86e-11	3.13e-01	4.36e-02
Mix-Pyr	6.97e-07	2.54e-09	2.06e-01	8.19e-02
Mix-TCDD	9.73e-01	8.99e-01	9.99e-01	9.99e-01

For each gene, the number of pairs and the pairs of treatments that show significant differences in expression is:

```
sapply(1:ncol(anova_posthoc_120), function(ind) {
  cat(paste0("For gene: ", colnames(anova_posthoc_120)[ind], "\n"))
  cat(paste0("Number of significant pairs: ",
             sum(anova_posthoc_120[ ,ind] < 0.05), "\n"))
  cat(names(which(anova_posthoc_120[ ,ind] < 0.05)), "\n\n")
})
```

```
## For gene: AhRR
## Number of significant pairs: 6
## TCDD-DMSO Mix-DMSO TCDD-BaP Mix-BaP TCDD-Pyr Mix-Pyr
##
## For gene: CYP1B1
## Number of significant pairs: 6
## TCDD-DMSO Mix-DMSO TCDD-BaP Mix-BaP TCDD-Pyr Mix-Pyr
##
## For gene: HK2
## Number of significant pairs: 0
##
##
## For gene: SCD1
## Number of significant pairs: 1
## TCDD-Pyr
```

```
## [[1]]
## NULL
##
## [[2]]
## NULL
##
## [[3]]
## NULL
##
## [[4]]
## NULL
```

These results are identical to the ones obtained with the ANOVA with a treatment effect.

Cell types The number of genes for which the null hypothesis is rejected for the cell types is 11 (at 5%) and these genes are: ACO1, AhR, AhRR, CAT, CYP1B1, G6PD, HK2, LPCAT, NQO1, NRF2, PRDX1.

For these genes, post-hoc tests are performed:

```
selectcell <- which(anova_120[1,] < 0.05)
anova_posthoc_120 <- apply(cell_120[ ,-c(1:3)][ ,selectcell], 2, function(agene) {
  TukeyHSD(aov(agene ~ cell_120$Treatment + cell_120$Condition))[[2]][,4]
})
format(anova_posthoc_120[,1:6], scientific=TRUE, digits=3) %>% kable() %>% kable_styling(bootstrap_opti
  add_header_above(c(" " = 1, "p-values" = 6))
```

	p-values					
	ACO1	AhR	AhRR	CAT	CYP1B1	G6PD
CTA-CT	9.76e-01	1.00e+00	9.92e-01	4.36e-01	8.15e-01	9.17e-01
CTP-CT	9.88e-01	1.00e+00	1.65e-03	1.00e+00	6.72e-01	1.00e+00
CTR-CT	1.68e-06	9.57e-03	6.89e-01	2.60e-05	5.61e-14	3.36e-02
RPA-CT	1.00e+00	9.04e-01	1.69e-01	6.03e-01	1.27e-01	7.11e-01
RPAt-CT	8.55e-02	1.00e+00	5.22e-02	3.74e-05	2.92e-05	9.54e-03
CTP-CTA	7.41e-01	1.00e+00	1.18e-02	6.21e-01	7.73e-02	8.68e-01
CTR-CTA	7.63e-08	8.18e-03	9.52e-01	1.76e-02	2.36e-14	1.34e-03
RPA-CTA	9.42e-01	8.86e-01	4.68e-01	1.00e+00	3.60e-03	1.57e-01
RPAt-CTA	1.19e-02	1.00e+00	2.04e-01	2.29e-02	1.12e-07	2.75e-04
CTR-CTP	2.75e-05	1.46e-02	1.26e-01	7.99e-05	2.52e-13	4.79e-02
RPA-CTP	9.97e-01	9.45e-01	6.03e-01	7.79e-01	9.06e-01	7.86e-01
RPAt-CTP	3.22e-01	1.00e+00	8.80e-01	1.14e-04	7.77e-03	1.43e-02
RPA-CTR	3.64e-06	1.56e-01	9.38e-01	8.01e-03	3.86e-11	5.87e-01
RPAt-CTR	3.95e-02	7.69e-03	7.13e-01	1.00e+00	6.30e-06	9.98e-01
RPAt-RPA	1.28e-01	8.78e-01	9.96e-01	1.06e-02	1.33e-01	3.26e-01

```
format(anova_posthoc_120[,7:11], scientific=TRUE, digits=3) %>% kable() %>% kable_styling(bootstrap_opt
  add_header_above(c(" " = 1, "p-values" = 5))
```

	p-values				
	HK2	LPCAT	NQO1	NRF2	PRDX1
CTA-CT	1.00e+00	8.85e-01	1.00e+00	9.73e-01	8.59e-01
CTP-CT	4.00e-03	9.98e-01	2.58e-03	7.17e-01	4.65e-01
CTR-CT	4.15e-09	3.42e-02	8.56e-04	1.37e-02	9.15e-01
RPA-CT	5.34e-03	9.95e-01	7.14e-01	9.84e-01	1.04e-01
RPAt-CT	4.42e-13	7.59e-03	4.67e-04	3.06e-02	1.00e+00
CTP-CTA	1.71e-03	6.45e-01	2.34e-03	2.56e-01	9.86e-01
CTR-CTA	1.24e-09	9.83e-04	7.69e-04	1.04e-03	2.72e-01
RPA-CTA	2.32e-03	5.93e-01	6.96e-01	7.03e-01	6.82e-01
RPAt-CTA	1.70e-13	1.45e-04	4.18e-04	2.76e-03	7.94e-01
CTR-CTP	2.08e-02	1.06e-01	1.00e+00	3.87e-01	6.54e-02
RPA-CTP	1.00e+00	1.00e+00	1.52e-01	9.77e-01	9.64e-01
RPAt-CTP	2.70e-05	2.91e-02	9.97e-01	5.59e-01	3.84e-01
RPA-CTR	1.61e-02	1.28e-01	7.45e-02	8.90e-02	6.47e-03
RPAt-CTR	4.54e-01	9.96e-01	1.00e+00	1.00e+00	9.52e-01
RPAt-RPA	1.87e-05	3.64e-02	4.91e-02	1.65e-01	7.55e-02

For each gene, the number of pairs and the pairs of cell types that show significant differences in expression is:

```r
sapply(1:ncol(anova_posthoc_120), function(ind) {
  cat(paste0("For gene: ", colnames(anova_posthoc_120)[ind], "\n"))
  cat(paste0("Number of significant pairs: ",
             sum(anova_posthoc_120[ ,ind] < 0.05), "\n"))
  cat(names(which(anova_posthoc_120[ ,ind] < 0.05)), "\n\n")
})
```

```
## For gene: ACO1
## Number of significant pairs: 6
## CTR-CT CTR-CTA RPAt-CTA CTR-CTP RPA-CTR RPAt-CTR
##
## For gene: AhR
## Number of significant pairs: 4
## CTR-CT CTR-CTA CTR-CTP RPAt-CTR
##
## For gene: AhRR
## Number of significant pairs: 2
## CTP-CT CTP-CTA
##
## For gene: CAT
## Number of significant pairs: 8
## CTR-CT RPAt-CT CTR-CTA RPAt-CTA CTR-CTP RPAt-CTP RPA-CTR RPAt-RPA
##
## For gene: CYP1B1
## Number of significant pairs: 9
## CTR-CT RPAt-CT CTR-CTA RPA-CTA RPAt-CTA CTR-CTP RPAt-CTP RPA-CTR RPAt-CTR
##
## For gene: G6PD
## Number of significant pairs: 6
## CTR-CT RPAt-CT CTR-CTA RPAt-CTA CTR-CTP RPAt-CTP
##
## For gene: HK2
## Number of significant pairs: 12
## CTP-CT CTR-CT RPA-CT RPAt-CT CTP-CTA CTR-CTA RPA-CTA RPAt-CTA CTR-CTP RPAt-CTP RPA-CTR RPAt-RPA
##
## For gene: LPCAT
## Number of significant pairs: 6
## CTR-CT RPAt-CT CTR-CTA RPAt-CTA RPAt-CTP RPAt-RPA
##
## For gene: NQO1
## Number of significant pairs: 7
## CTP-CT CTR-CT RPAt-CT CTP-CTA CTR-CTA RPAt-CTA RPAt-RPA
##
## For gene: NRF2
## Number of significant pairs: 4
## CTR-CT RPAt-CT CTR-CTA RPAt-CTA
##
## For gene: PRDX1
## Number of significant pairs: 1
## RPA-CTR
```

```
## [[1]]
## NULL
##
```

```
## [[2]]
## NULL
##
## [[3]]
## NULL
##
## [[4]]
## NULL
##
## [[5]]
## NULL
##
## [[6]]
## NULL
##
## [[7]]
## NULL
##
## [[8]]
## NULL
##
## [[9]]
## NULL
##
## [[10]]
## NULL
##
## [[11]]
## NULL
```

3.2.3 Conclusion

Treatments The following table represents all genes for which at least one global test is significant

- first column: which test(s) is (are) significant?
- second column: for which treatment(s) is normality rejected for this gene?

Treatments	Tests giving these results	trestment(s) whose normality Shapiro rejects
AhRR	two-factor ANOVA test	-
CYP1B1	two-factor ANOVA test	Pyr
HK2	two-factor ANOVA test	Mix
SCD1	two-factor ANOVA test	-

The following tables are constructed as follows:

- the upper parts of the tables are composed by the results of the post-hoc two-factor ANOVA test;
- A star means that the contrast between the two treatments is significant (p-value < 0.05);

AhRR	DMSO	BaP	Pyr	TCDD	Mix
DMSO	-			*	*
BaP	-	-		*	*

AhRR	DMSO	BaP	Pyr	TCDD	Mix
Pyr	-	-	-	*	*
TCDD	-	-	-	-	
Mix	-	-	-	-	-

CYP1B1	DMSO	BaP	Pyr	TCDD	Mix
DMSO	-			*	*
BaP	-	-		*	*
Pyr	-	-	-	*	*
TCDD	-	-	-	-	
Mix	-	-	-	-	-

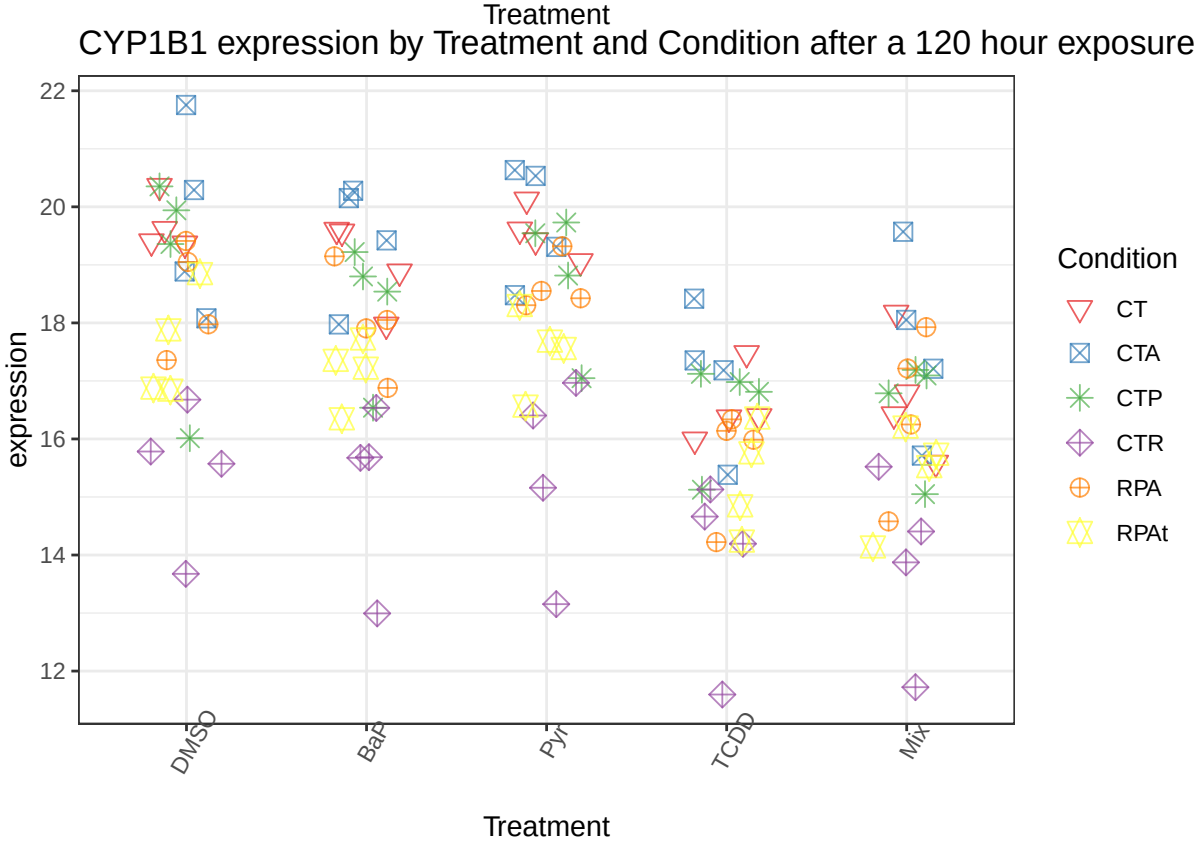
HK2	DMSO	BaP	Pyr	TCDD	Mix
DMSO	-				
BaP	-	-			
Pyr	-	-	-		
TCDD	-	-	-	-	
Mix	-	-	-	-	-

SCD1	DMSO	BaP	Pyr	TCDD	Mix
DMSO	-				
BaP	-	-			
Pyr	-	-	-	*	
TCDD	-	-	-	-	
Mix	-	-	-	-	-

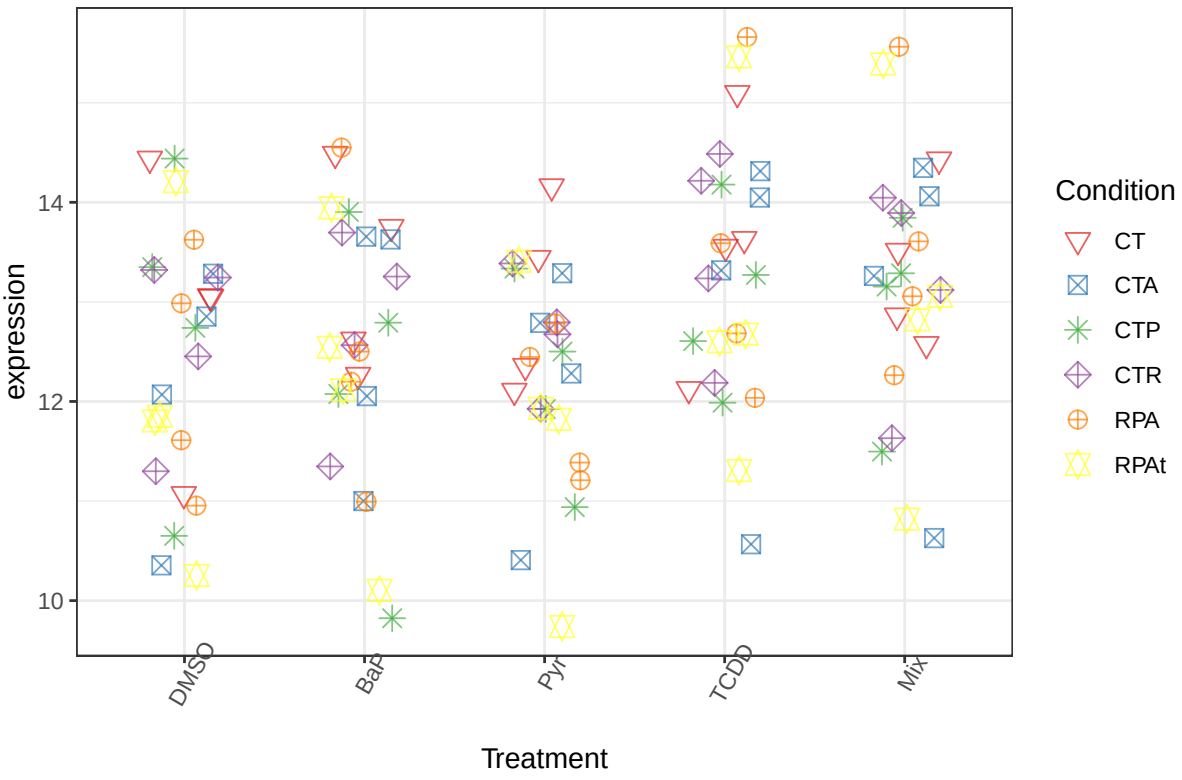
When comparing this two-factor ANOVA results with the results of the one-factor ANOVA, there is a difference from the results obtained previously. Indeed, even if the HK2 gene does not show a significant difference between particular treatments, the hypothesis of identical distribution between treatments is not rejected.
For the AhRR, CYP1B1 and SCD1 genes, the results are identical.

```
all_signif <- names(cell_120[ ,-c(1:3)][ ,selecttreat])
sapply(all_signif, function(agene) {
  df <- data.frame(cell_120$Treatment, cell_120[ ,agene], cell_120$Condition)
  names(df) <- c("Treatment", "expr", "Condition")
  p <- ggplot(df, aes(x = Treatment, y = expr, colour = Condition, shape = Condition)) +
  geom_jitter(alpha = 0.7, width = 0.2, size = 3) + theme_bw() +
    theme(axis.text.x = element_text(angle = 60)) + ylab("expression") +
    ggtitle(paste0(agene,
                   " expression by Treatment and Condition after a 120 hour exposure")) +
    scale_color_manual(values = palettecondition) +
    scale_shape_manual(values = symbolcondition)
  print(p)
  invisible(NULL)
})
```


AhRR expression by Treatment and Condition after a 120 hour exposure
expression
20
18
16
14
DMSO
BaP
Pyr
TCDD
Mix
Treatment
Condition
CT
CTA
CTP
CTR
RPA
RPAt


CYP1B1 expression by Treatment and Condition after a 120 hour exposure
expression
22
20
18
16
14
12
DMSO
BaP
Pyr
TCDD
Mix
Treatment
Condition
CT
CTA
CTP
CTR
RPA
RPAt


HK2 expression by Treatment and Condition after a 120 hour exposure
expression
19
18
17
16
15
14
DMSO
BaP
Pyr
TCDD
Mix
Treatment
Condition
CT
CTA
CTP
CTR
RPA
RPAt


SCD1 expression by Treatment and Condition after a 120 hour exposure
expression
14
12
10
DMSO
BaP
Pyr
TCDD
Mix
Treatment
Condition
CT
CTA
CTP
CTR
RPA
RPAt

```
## $AhRR
```

```
## NULL
##
## $CYP1B1
## NULL
##
## $HK2
## NULL
##
## $SCD1
## NULL
```

- AhRR and CYP1B1 are genes that are significantly over-expressed in DMSO, Bap and Pyr compared to TCDD and Mix.
- No significant difference is found in HK2 expression between any pair of treatments but the expression of this gene is found significantly different overall. The boxplot shows a steady decrease of the expression with treatment aggressiveness.
- SCD1 is significantly under-expressed in Pyr compared to TCDD.

Cell types The following table represents all genes for which at least one global test is significant

- first column: which test(s) is (are) significant?
- second column: for which cell type(s) is normality rejected for this gene?

Cell types	Tests giving these results	Cell Type(s) whose normality Shapiro rejects
ACO1	two-factor ANOVA test	CTA, CTRPA, CTRPAt
AhR	two-factor ANOVA test	CTP, CTR
AhRR	two-factor ANOVA test	-
CAT	two-factor ANOVA test	CTR, CTRPAt
CYP1B1	two-factor ANOVA test	CT
G6PD	two-factor ANOVA test	-
HK2	two-factor ANOVA test	-
LCPAT	two-factor ANOVA test	CTA, CTR, CTRPAt
NQO1	two-factor ANOVA test	CT, CTR, CTRPAt
NRF2	two-factor ANOVA test	CT, CTR, CTRPA, CTRPAt
PRDX1	two-factor ANOVA test	CT, CTR

The following tables are constructed as follows:

- the upper parts of the tables are composed by the results of the post-hoc two-factor ANOVA test;
- A star means that the contrast between the two cell types is significant (p-value < 0.05);

ACO1	CT	CTA	CTP	CTR	CTRPA	CTRPAt
CT	-			*		
CTA	-	-		*		*
CTP	-	-	-	*		
CTR	-	-	-	-	*	*
CTRPA	-	-	-	-	-	
CTRPAt	-	-	-	-	-	-

AhR	CT	CTA	CTP	CTR	CTRPA	CTRPAt
CT	-			*		
CTA	-	-		*		
CTP	-	-	-	*		
CTR	-	-	-	-		*
CTRPA	-	-	-	-	-	
CTRPAt	-	-	-	-	-	-

AhRR	CT	CTA	CTP	CTR	CTRPA	CTRPAt
CT	-		*			
CTA	-	-	*			
CTP	-	-	-			
CTR	-	-	-	-		
CTRPA	-	-	-	-	-	
CTRPAt	-	-	-	-	-	-

CAT	CT	CTA	CTP	CTR	CTRPA	CTRPAt
CT	-			*		*
CTA	-	-		*		*
CTP	-	-	-	*		*
CTR	-	-	-	-	*	
CTRPA	-	-	-	-	-	*
CTRPAt	-	-	-	-	-	-

CYP1B1	CT	CTA	CTP	CTR	CTRPA	CTRPAt
CT	-			*		*
CTA	-	-		*	*	*
CTP	-	-	-	*		*
CTR	-	-	-	-	*	*
CTRPA	-	-	-	-	-	
CTRPAt	-	-	-	-	-	-

G6PD	CT	CTA	CTP	CTR	CTRPA	CTRPAt
CT	-			*		*
CTA	-	-		*		*
CTP	-	-	-	*		*
CTR	-	-	-	-		
CTRPA	-	-	-	-	-	
CTRPAt	-	-	-	-	-	-

HK2	CT	CTA	CTP	CTR	CTRPA	CTRPAt
CT	-		*	*	*	*
CTA	-	-	*	*	*	*
CTP	-	-	-	*		*

HK2	CT	CTA	CTP	CTR	CTRPA	CTRPAt
CTR	-	-	-	-	*	
CTRPA	-	-	-	-	-	*
CTRPAt	-	-	-	-	-	-

LPCAT	CT	CTA	CTP	CTR	CTRPA	CTRPAt
CT	-			*		*
CTA	-	-		*		*
CTP	-	-	-			*
CTR	-	-	-	-		
CTRPA	-	-	-	-	-	*
CTRPAt	-	-	-	-	-	-

NQO1	CT	CTA	CTP	CTR	CTRPA	CTRPAt
CT	-		*	*		*
CTA	-	-	*	*		*
CTP	-	-	-			
CTR	-	-	-	-		
CTRPA	-	-	-	-	-	*
CTRPAt	-	-	-	-	-	-

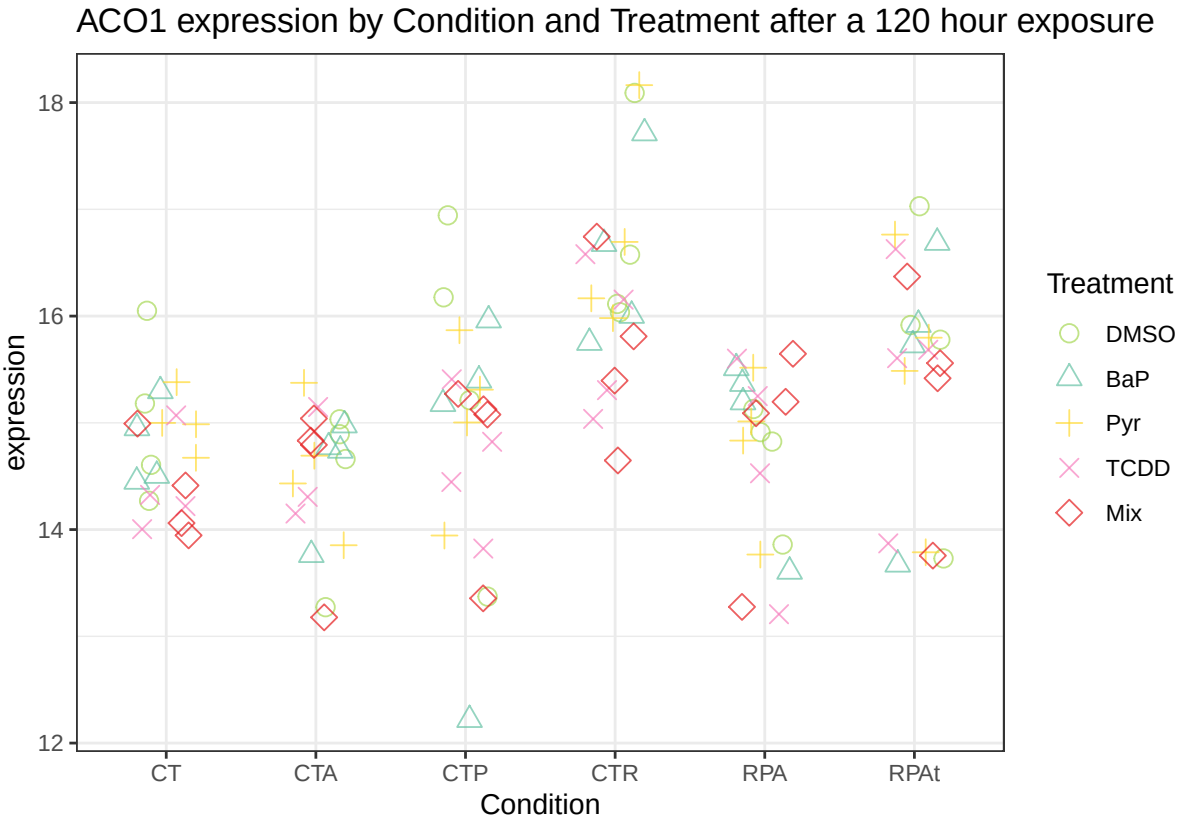
NRF2	CT	CTA	CTP	CTR	CTRPA	CTRPAt
CT	-			*		*
CTA	-	-		*		*
CTP	-	-	-			
CTR	-	-	-	-		
CTRPA	-	-	-	-	-	
CTRPAt	-	-	-	-	-	-

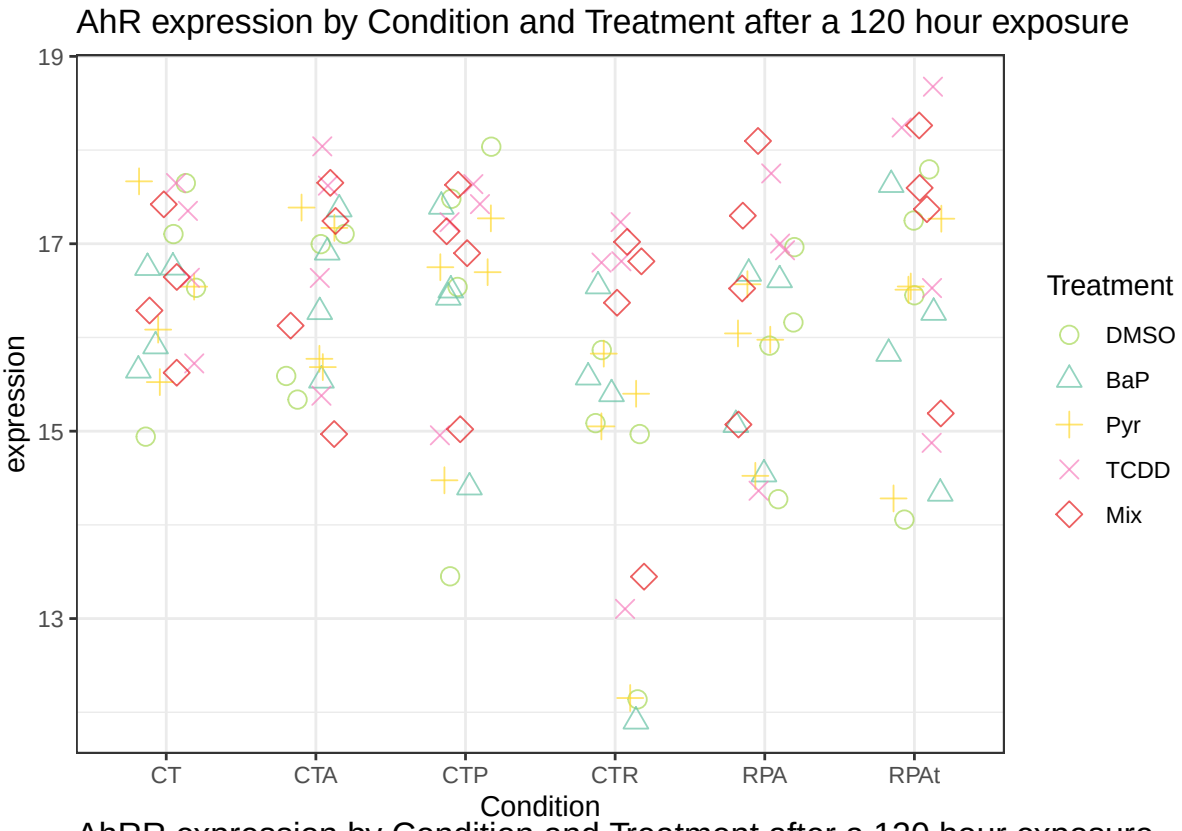
PRDX1	CT	CTA	CTP	CTR	CTRPA	CTRPAt
CT	-					
CTA	-	-				
CTP	-	-	-			
CTR	-	-	-	-	*	
CTRPA	-	-	-	-	-	
CTRPAt	-	-	-	-	-	-

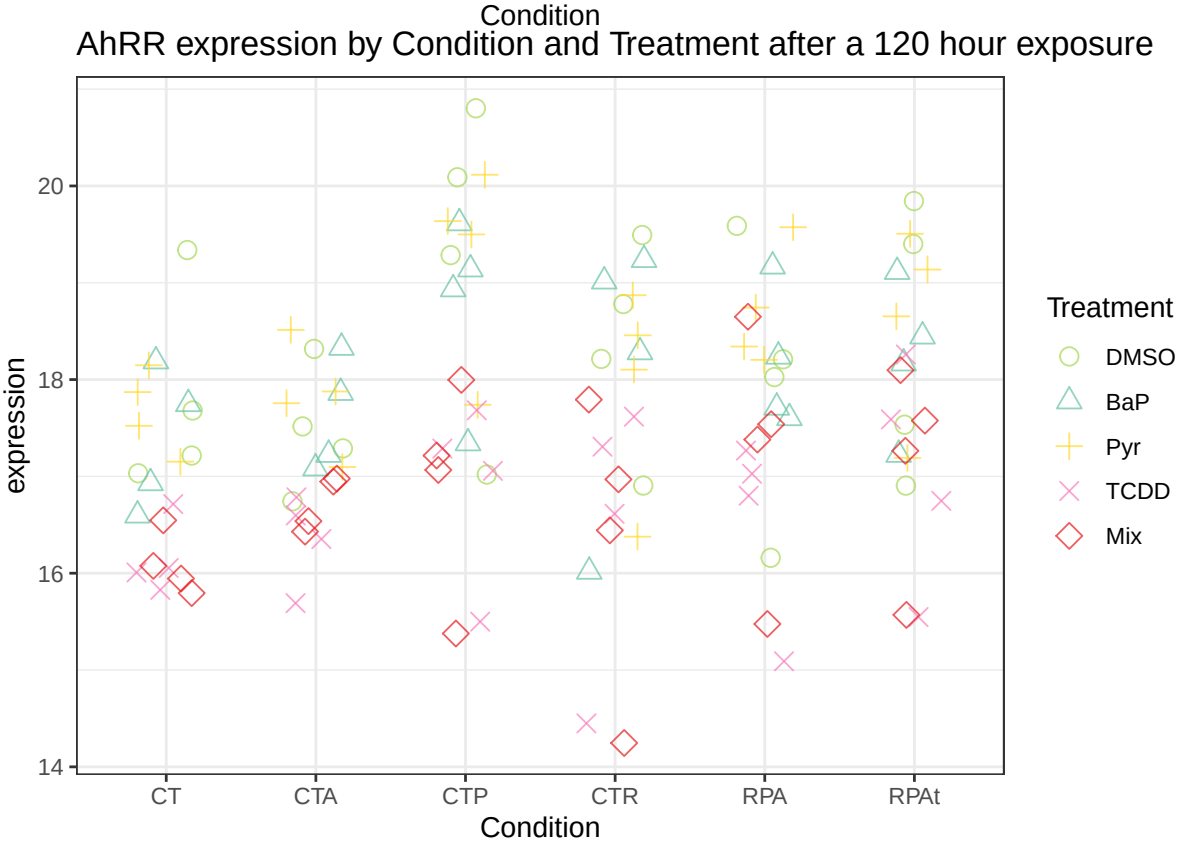
Several genes show a significant difference in expression depending on the cell type, it is generally the cell types CTR and CTRPAt that show differences with other cell types.

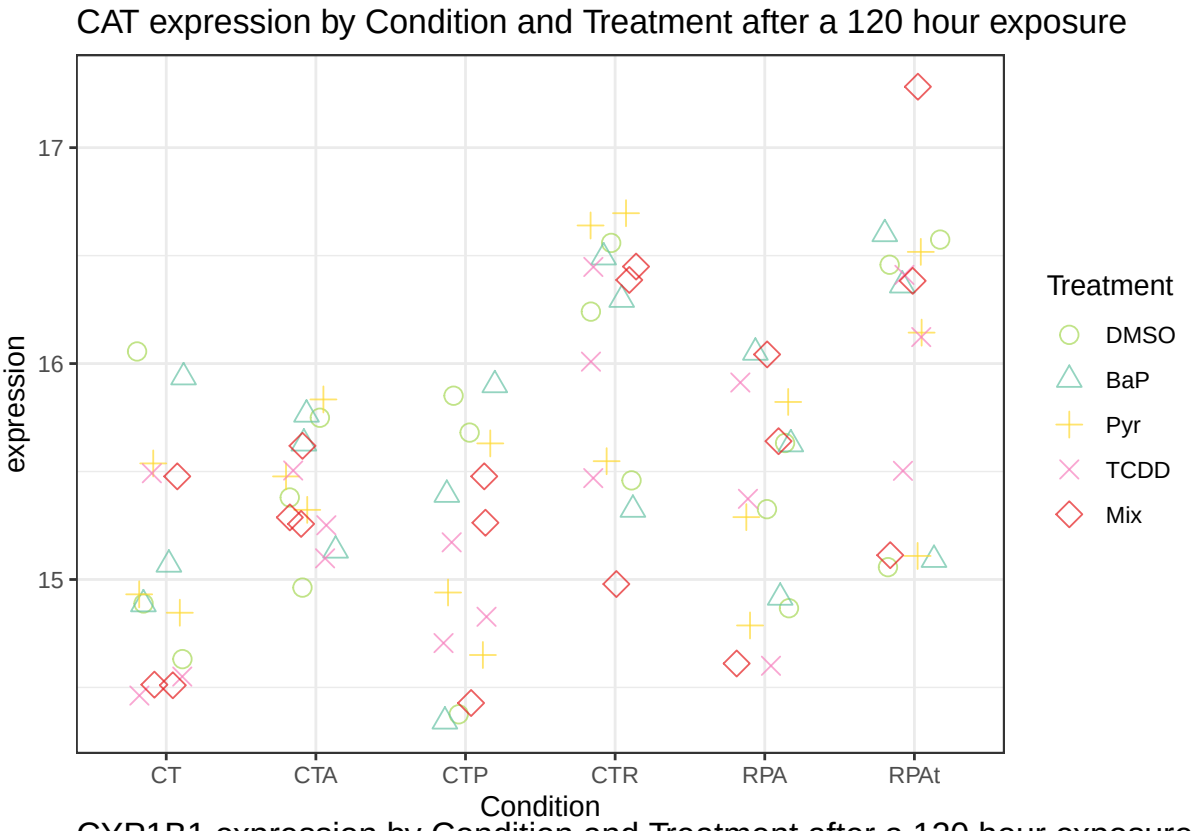
```
all_signif <- names(cell_120[ ,-c(1:3)][ ,selectcell])
sapply(all_signif, function(agene) {
  df <- data.frame(cell_120$Treatment, cell_120[ ,agene], cell_120$Condition)
  names(df) <- c("Treatment", "expr", "Condition")
  p <- ggplot(df, aes(x = Condition, y = expr, colour = Treatment, shape = Treatment)) +
```

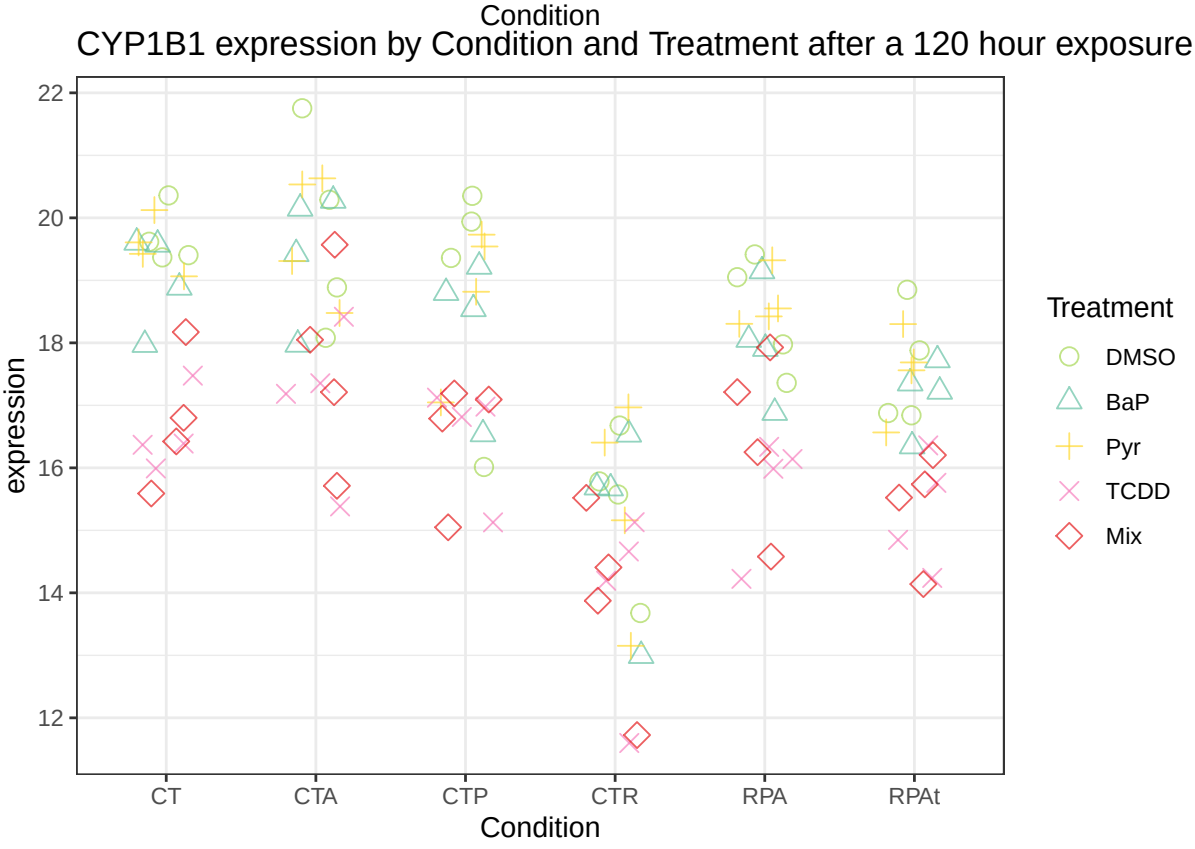
```
    geom_jitter(alpha = 0.7, width = 0.2, size = 3) +
    theme(axis.text.x = element_text(angle = 60)) +
    ylab("expression") + theme_bw() +
    ggtitle(paste0(agene,
                   " expression by Condition and Treatment after a 120 hour exposure")) +
    scale_color_manual(values = palettetreatment) +
    scale_shape_manual(values = symboltreatment)
  print(p)
  invisible(NULL)
})
```

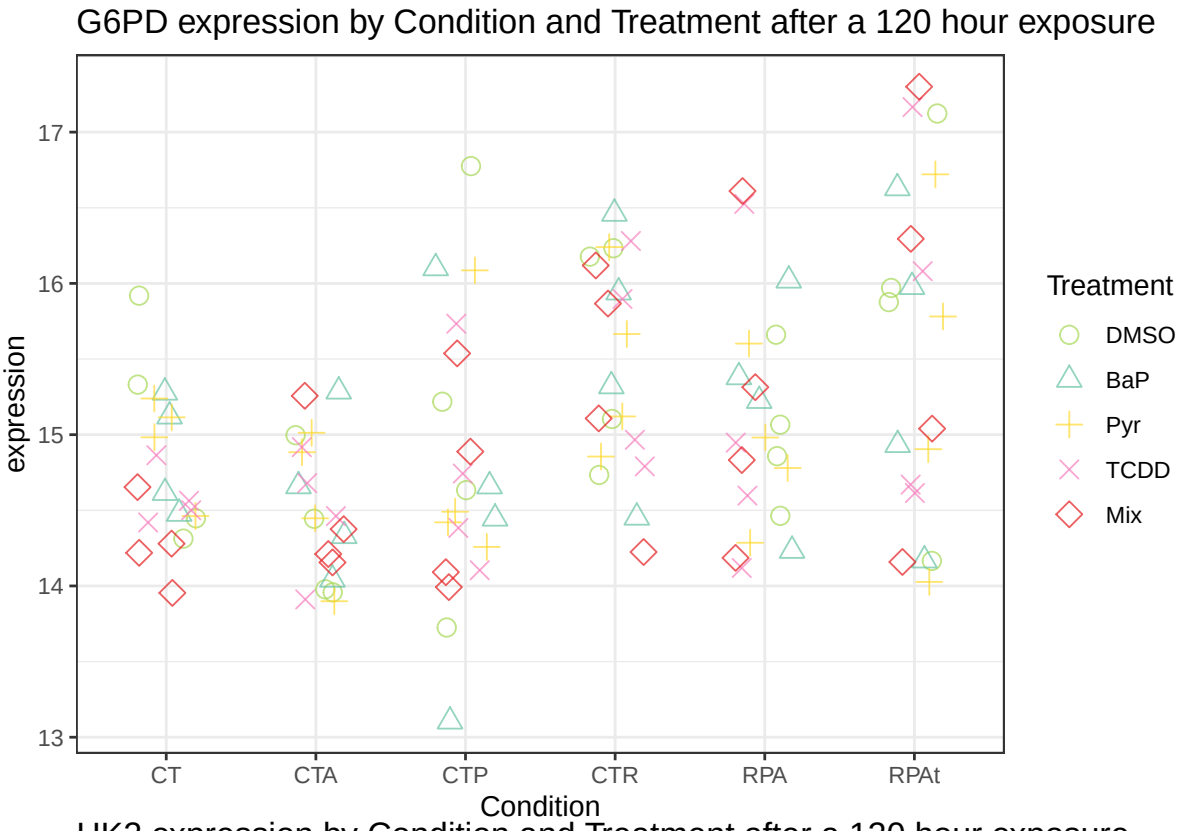


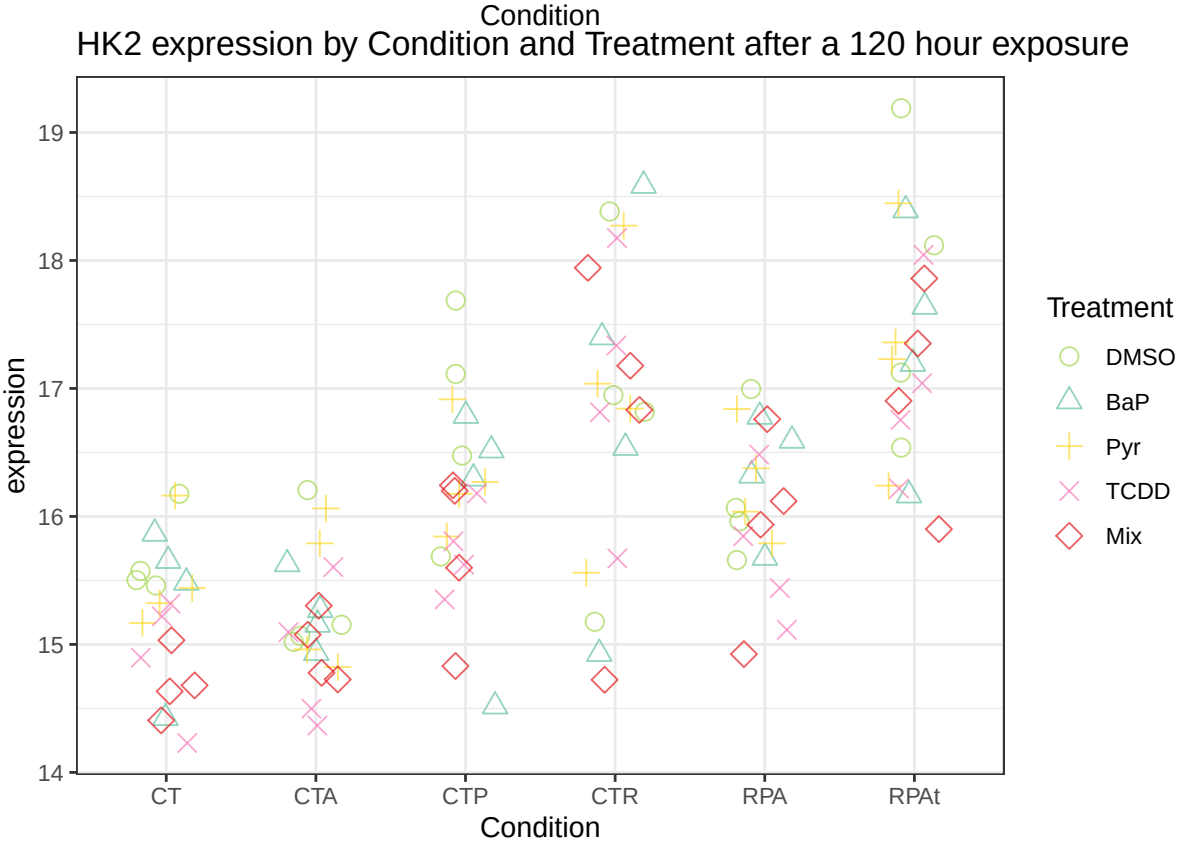

AhR expression by Condition and Treatment after a 120 hour exposure
expression
19
17
15
13
CT
CTA
CTP
CTR
RPA
RPAt
Condition
Treatment
DMSO
BaP
Pyr
TCDD
Mix

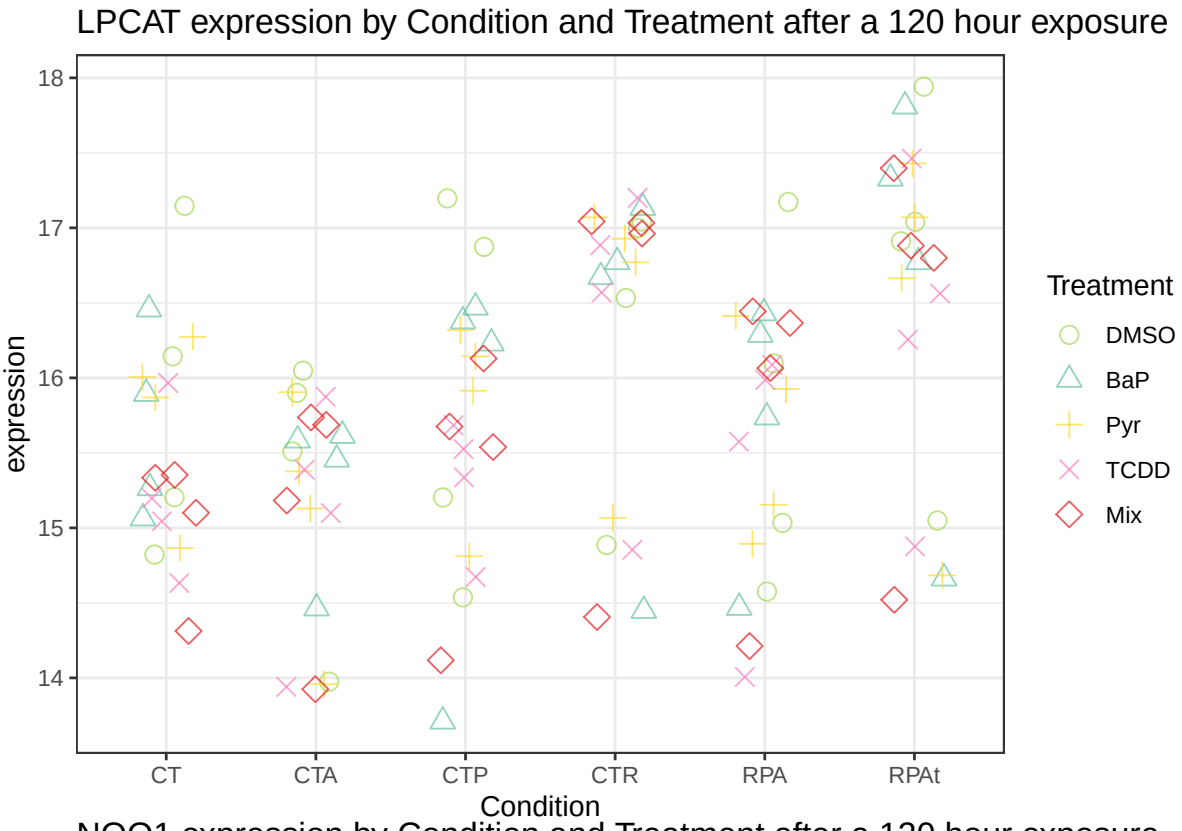

AhRR expression by Condition and Treatment after a 120 hour exposure
expression
20
18
16
14
CT
CTA
CTP
CTR
RPA
RPAt
Condition
Treatment
DMSO
BaP
Pyr
TCDD
Mix

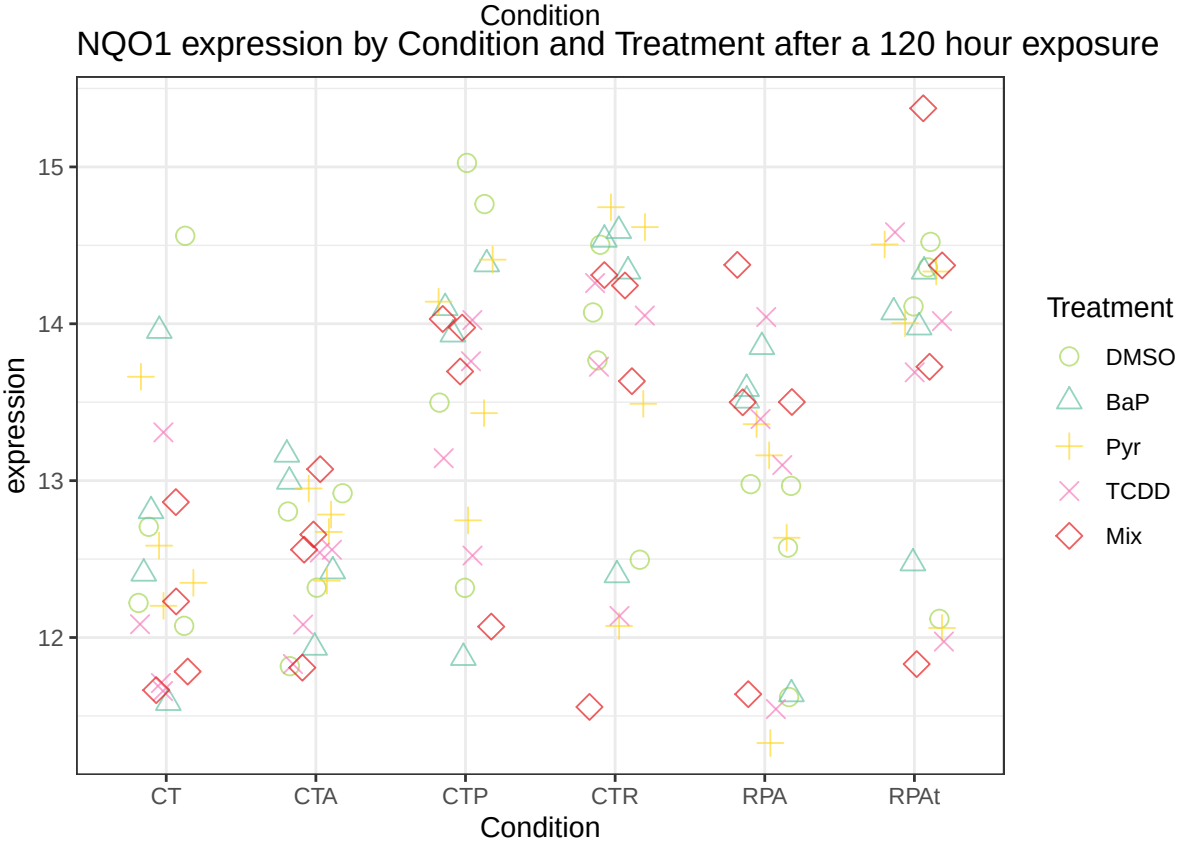

CAT expression by Condition and Treatment after a 120 hour exposure
expression
17
16
15
CT
CTA
CTP
CTR
RPA
RPAt
Condition
Treatment
DMSO
BaP
Pyr
TCDD
Mix

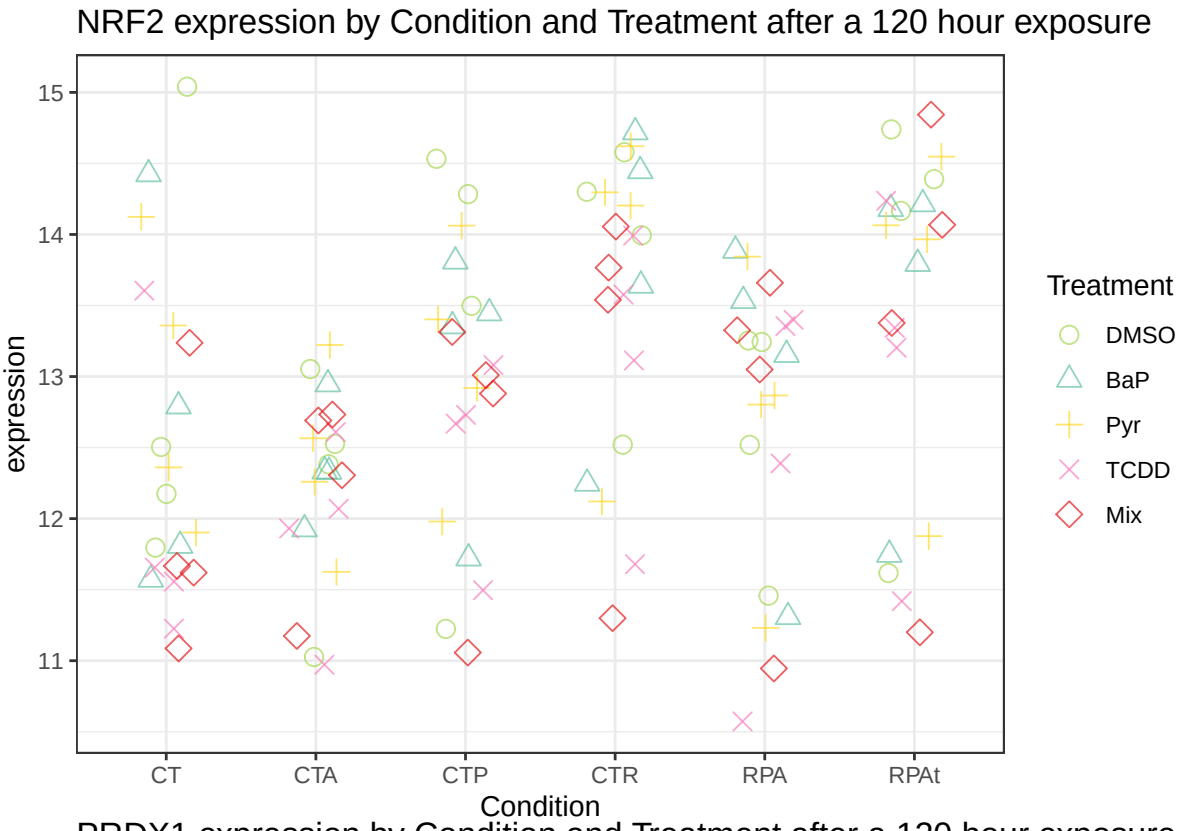

CYP1B1 expression by Condition and Treatment after a 120 hour exposure
expression
22
20
18
16
14
12
CT
CTA
CTP
CTR
RPA
RPAt
Condition
Treatment
DMSO
BaP
Pyr
TCDD
Mix

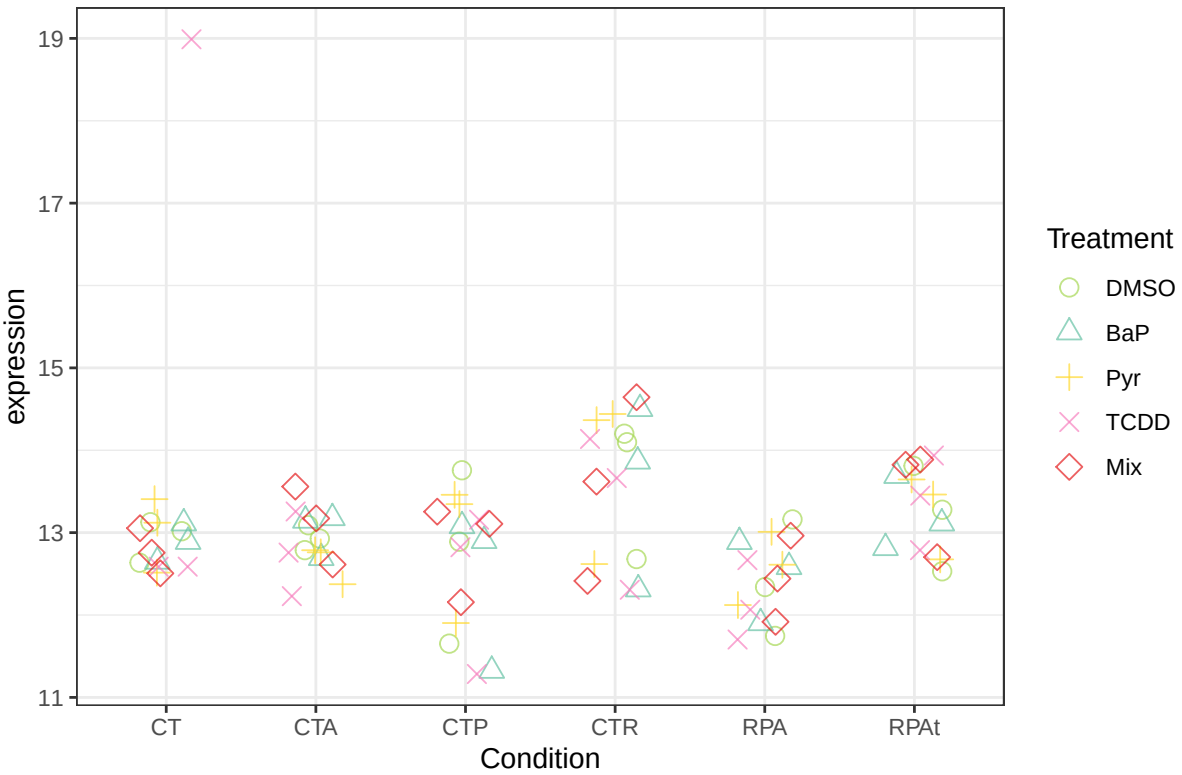

G6PD expression by Condition and Treatment after a 120 hour exposure
expression
17
16
15
14
13
CT
CTA
CTP
CTR
RPA
RPAt
Condition
Treatment
DMSO
BaP
Pyr
TCDD
Mix


HK2 expression by Condition and Treatment after a 120 hour exposure
expression
19
18
17
16
15
14
CT
CTA
CTP
CTR
RPA
RPAt
Condition
Treatment
DMSO
BaP
Pyr
TCDD
Mix


LPCAT expression by Condition and Treatment after a 120 hour exposure
expression
18
17
16
15
14
CT
CTA
CTP
CTR
RPA
RPAt
Condition
Treatment
DMSO
BaP
Pyr
TCDD
Mix


NQO1 expression by Condition and Treatment after a 120 hour exposure
expression
15
14
13
12
CT
CTA
CTP
CTR
RPA
RPAt
Condition
Treatment
DMSO
BaP
Pyr
TCDD
Mix


NRF2 expression by Condition and Treatment after a 120 hour exposure
expression
15
14
13
12
11
CT
CTA
CTP
CTR
RPA
RPAt
Condition
Treatment
DMSO
BaP
Pyr
TCDD
Mix


PRDX1 expression by Condition and Treatment after a 120 hour exposure
expression
19
17
15
13
11
CT
CTA
CTP
CTR
RPA
RPAt
Condition
Treatment
DMSO
BaP
Pyr
TCDD
Mix

```
## $ACO1
```

```
## NULL
##
## $AhR
## NULL
##
## $AhRR
## NULL
##
## $CAT
## NULL
##
## $CYP1B1
## NULL
##
## $G6PD
## NULL
##
## $HK2
## NULL
##
## $LPCAT
## NULL
##
## $NQO1
## NULL
##
## $NRF2
## NULL
##
## $PRDX1
## NULL
```

- ACO1 is significantly over-expressed in CTR compared to the other cell types. Similarly this gene is significantly under-expressed in CTA compared to CTRPAt.
- AhR is significantly under-expressed in CTR compared to the other cell types except for CTRPA.
- AhRR is significantly over-expressed in CTP compared to CT and CTA.
- CAT is significantly over-expressed in CTR and CTRPAt compared to the other cell types.
- CYP1B1 is significantly under-expressed in CTR compared to the other cell types. This gene is also significantly under-expressed in CTRPA and CTRPAt compared to CTA.
- G6PD is significantly over-expressed in CTR and CTRPAt compared to CT, CTA and CTR.
- HK2 is significantly under-expressed in CT, CTA and CTRPA compared to CTP, CTR and CTRPAt.
- LPCAT is significantly over-expressed in CTRPAt compared to the other cell types. However, the distribution of this gene in CTR and CTRPAt is bimodal which somehow flaws a bit this result.
- NQO1 is significantly under-expressed in CT and CTA compared to CTRPAt. Similarly this gene is significantly under-expressed in CTRPA compared to CTRPAt.
- NRF2 is significantly over-expressed in CTR and CTRPAt compared to CTA. Similarly, CT is significantly over-expressed compared to CTR and CTRPAt.
- PRDX1 is significantly over-expressed in CTR compared to CTRPA.

3.3 Conclusion

The four following tables are constructed as follows:

- The upper parts of the tables are composed by the results of the two-way ANOVA tests.
- An empty box means that there is no significant difference between the couple whatever the gene
- A completed box means that the genes present in this box show significant differences in expression.

Treatment

48h	DMSO	BaP	Pyr	TCDD	Mix
DMSO	-			CYP1B1	CYP1B1
BaP	-	-			
Pyr	-	-	-		
TCDD	-	-	-	-	
Mix	-	-	-	-	-

120h	DMSO	BaP	Pyr	TCDD	Mix
DMSO	-			AhRR, CYP1B1	AhRR, CYP1B1
BaP	-	-		AhRR, CYP1B1	AhRR, CYP1B1
Pyr	-	-	-	AhRR, CYP1B1, SCD1	AhRR, CYP1B1
TCDD	-	-	-	-	
Mix	-	-	-	-	-

Cell type

48h	CT	CTA	CTP	CTR	CTRPA
CT	-			CYP1B1, HK2	HMOX, NQO1
CTA	-	-	CYP1B1	CYP1B1, HK2, NRF2	CYP1B1
CTP	-	-	-	HK2	
CTR	-	-	-	-	HK2, HMOX, NQO1, NRF2
CTRPA	-	-	-	-	-
CTRPAt	-	-	-	-	-

120h	CT	CTA	CTP	CTR	CTRPA
CT	-		AhRR, HK2, NQO1	ACO1, AhR, CAT, CYP1B1, G6PD, HK2, LPCAT, NQO1, NRF2	HK2
CTA	-	-	AhRR, HK2, NQO1	ACO1, AhR, CAT, CYP1B1, G6PD, HK2, LPCAT, NQO1, NRF2	CYP1B1, HK2
CTP	-	-	-	ACO1, AhR, CAT, CYP1B1, G6PD, HK2	

120h	CT	CTA	CTP	CTR	CTRPA
CTR	-	-	-	-	ACO1, CAT, CYP1B1, HK2, PRDX1,
CTRPA	-	-	-	-	-
CTRPAt	-	-	-	-	-

Global The CYP1B1 gene always shows differences in expression of treatment with TCDD and Mix and at least one of the other three treatments and a difference in cell type between CTA and 3 other cell types and between CTRPAt and 2 other cell types. There are much more differences at 120h than at 48h.

The two treatments TCDD and Mix show very strong similarities and never show any significant difference, regardless of the gene. The same goes for the treatments DMSO and Pyr.

The CTR cell type shows regularly significant differences with the three cell types closest to normal.

4. Is there an interaction effect between treatments and cell types?

An interaction can occur when considering the relationship between three or more variables. The term "interaction" is therefore used to describe a situation in which the simultaneous influence of two variables on a third is not additive. If two variables of interest interact, the relationship between each of the variables and a third "dependent" variable depends on the value of the interacting variables.

In this section a complete model with two additive effects (cell line and treatment) and their interaction is fitted. The significance of the interaction effect is then tested.

4.1 After a 48 hour exposure

4.1.1 Two-way ANOVA with interaction

```
anova_48 <- apply(cell_48[ ,-c(1:3)], 2, function(agene) {
  summary(aov(agene ~ cell_48$Condition : cell_48$Treatment +
                cell_48$Condition * cell_48$Treatment))[[1]][1:3,5]
})
rownames(anova_48) = c("Cell type", "Treatment", "Cell type : Treatment")
format(anova_48, scientific=TRUE, digits=3) %>%
kable() %>%
  kable_styling(latex_options = c("HOLD_position")) %>%
  column_spec(1:8)
```

	AhRR	CYP1B1	HK2	HMOX	MCT4	NHE1	NQO1	NRF2
Cell type	1.80e-01	8.25e-07	2.89e-11	1.39e-03	9.92e-01	5.87e-01	4.90e-03	2.00e-03
Treatment	3.50e-01	8.31e-03	5.16e-01	7.76e-01	1.00e+00	8.38e-01	8.31e-01	7.62e-01
Cell type : Treatment	1.00e+00	9.98e-01	9.80e-01	1.00e+00	1.00e+00	9.99e-01	1.00e+00	1.00e+00

There is no interaction between cell types and treatments, their effects are additive.

Treatment The number of genes for which the null hypothesis is rejected for the treatments is 1 (at 5%) and these genes are: CYP1B1.

For these genes, post-hoc tests are performed:

```
selecttreat <- which(anova_48[2,] < 0.05)
anova_posthoc_48 <- apply(cell_48[ ,-c(1:3)][ ,selecttreat], 2, function(agene) {
  TukeyHSD(aov(agene ~ cell_48$Condition + cell_48$Treatment +
                cell_48$Condition * cell_48$Treatment))[[2]][,4]
})
format(anova_posthoc_48, scientific=TRUE, digits=3) %>% kable() %>%
  kable_styling(bootstrap_options = "striped") %>%
  add_header_above(c(" " = 1, "p-values" = length(selecttreat)))
```

	p-values
	CYP1B1
BaP-DMSO	6.55e-01
Pyr-DMSO	8.02e-01
TCDD-DMSO	1.30e-02
Mix-DMSO	3.21e-02
Pyr-BaP	9.99e-01
TCDD-BaP	2.99e-01
Mix-BaP	4.86e-01
TCDD-Pyr	2.09e-01
Mix-Pyr	3.63e-01
Mix-TCDD	9.97e-01

For each gene, the number of pairs and the pairs of treatments that show significant differences in expression is:

```
sapply(1:ncol(anova_posthoc_48), function(ind) {
  cat(paste0("For gene: ", colnames(anova_posthoc_48)[ind], "\n"))
  cat(paste0("Number of significant pairs: ",
             sum(anova_posthoc_48[ ,ind] < 0.05), "\n"))
  cat(names(which(anova_posthoc_48[ ,ind] < 0.05)), "\n\n")
})
```

```
## For gene: CYP1B1
## Number of significant pairs: 2
## TCDD-DMSO Mix-DMSO
```

```
## [[1]]
## NULL
```

Cell types The number of genes for which the null hypothesis is rejected for the cell types is 5 (at 5%) and these genes are: CYP1B1, HK2, HMOX, NQO1, NRF2.

For these genes, post-hoc tests are performed:

```
selectcell <- which(anova_48[1,] < 0.05)
anova_posthoc_48 <- apply(cell_48[ ,-c(1:3)][ ,selectcell], 2, function(agene) {
  TukeyHSD(aov(agene ~ cell_48$Condition + cell_48$Treatment +
                cell_48$Condition * cell_48$Treatment))[[1]][,4]
})
format(anova_posthoc_48, scientific=TRUE, digits=3) %>% kable() %>%
```

```
  kable_styling(bootstrap_options = "striped") %>%
  add_header_above(c(" " = 1, "p-values" = length(selectcell)))
```

	p-values				
	CYP1B1	HK2	HMOX	NQO1	NRF2
CTA-CT	2.80e-01	9.77e-01	9.78e-01	9.87e-01	8.60e-01
CTP-CT	1.67e-01	6.69e-01	4.55e-01	9.88e-01	9.97e-01
CTR-CT	3.27e-03	9.08e-06	9.87e-01	7.53e-01	3.37e-01
RPA-CT	7.20e-01	2.18e-01	6.47e-02	1.01e-01	3.86e-01
RPAt-CT	4.58e-02	1.02e-09	4.23e-02	9.88e-01	8.41e-01
CTP-CTA	4.14e-04	9.71e-01	8.71e-01	1.00e+00	9.86e-01
CTR-CTA	1.61e-06	1.32e-04	7.46e-01	3.57e-01	2.72e-02
RPA-CTA	9.35e-03	6.36e-01	2.93e-01	3.49e-01	9.67e-01
RPAt-CTA	4.95e-05	1.76e-08	2.16e-01	7.99e-01	1.97e-01
CTR-CTP	7.18e-01	2.42e-03	1.52e-01	3.73e-01	1.50e-01
RPA-CTP	9.06e-01	9.75e-01	9.28e-01	3.72e-01	7.04e-01
RPAt-CTP	9.96e-01	5.69e-07	8.68e-01	8.07e-01	5.76e-01
RPA-CTR	1.44e-01	1.93e-02	1.15e-02	2.32e-03	2.54e-03
RPAt-CTR	9.40e-01	1.87e-01	6.94e-03	9.78e-01	9.58e-01
RPAt-RPA	6.22e-01	6.70e-06	1.00e+00	2.05e-02	3.07e-02

For each gene, the number of pairs and the pairs of cell types that show significant differences in expression is:

```
sapply(1:ncol(anova_posthoc_48), function(ind) {
  cat(paste0("For gene: ", colnames(anova_posthoc_48)[ind], "\n"))
  cat(paste0("Number of significant pairs: ",
             sum(anova_posthoc_48[ ,ind] < 0.05), "\n"))
  cat(names(which(anova_posthoc_48[ ,ind] < 0.05)), "\n\n")
})
```

```
## For gene: CYP1B1
## Number of significant pairs: 6
## CTR-CT RPAt-CT CTP-CTA CTR-CTA RPA-CTA RPAt-CTA
##
## For gene: HK2
## Number of significant pairs: 8
## CTR-CT RPAt-CT CTR-CTA RPAt-CTA CTR-CTP RPAt-CTP RPA-CTR RPAt-RPA
##
## For gene: HMOX
## Number of significant pairs: 3
## RPAt-CT RPA-CTR RPAt-CTR
##
## For gene: NQO1
## Number of significant pairs: 2
## RPA-CTR RPAt-RPA
##
## For gene: NRF2
## Number of significant pairs: 3
## CTR-CTA RPA-CTR RPAt-RPA
```

```
## [[1]]
## NULL
##
## [[2]]
```

```
## NULL
##
## [[3]]
## NULL
##
## [[4]]
## NULL
##
## [[5]]
## NULL
```

4.1.2 Conclusion

Treatment The following table represents all genes for which at least one global test is significant

- first column: which test(s) is (are) significant?
- second column: for which treatment(s) is normality rejected for this gene?

Treatment	Tests giving these results	Treatment(s) whose normality Shapiro rejects
CYP1B1	ANOVA 2 factor test	TCDD

The following tables are constructed as follows:

- the upper parts of the tables are composed by the results of the post-hoc ANOVA test;
- A star means that the contrast between the two treatments is significant (p-value < 0.05);

CYP1B1	DMSO	BaP	Pyr	TCDD	Mix
DMSO	-			*	*
BaP	-	-			
Pyr	-	-	-		
TCDD	-	-	-	-	
Mix	-	-	-	-	-

There is no difference with the test results obtained with a two-way ANOVA without the interaction effect for the direct treatment effect.

Cell types The following table represents all genes for which at least one global test is significant

- first column: which test(s) is (are) significant?
- second column: for which cell type(s) is normality rejected for this gene?

Cell types	Tests giving these results	Cell Type(s) whose normality Shapiro rejects
CYP1B1	two-factor ANOVA test	-
HK2	two-factor ANOVA test	CTR
HMOX	two-factor ANOVA test	CT
NQO1	two-factor ANOVA test	CTP, CTR, CTRPAt
NRF2	two-factor ANOVA test	CTA, CTRPA

The following tables are constructed as follows:

- the upper parts of the tables are composed by the results of the post-hoc two-factor ANOVA test;
- A star means that the contrast between the two cell types is significant (p-value < 0.05);

HMOX	CT	CTA	CTP	CTR	CTRPA	CTRPAt
CT	-					*
CTA	-	-				
CTP	-	-	-			
CTR	-	-	-	-	*	*
CTRPA	-	-	-	-	-	
CTRPAt	-	-	-	-	-	-

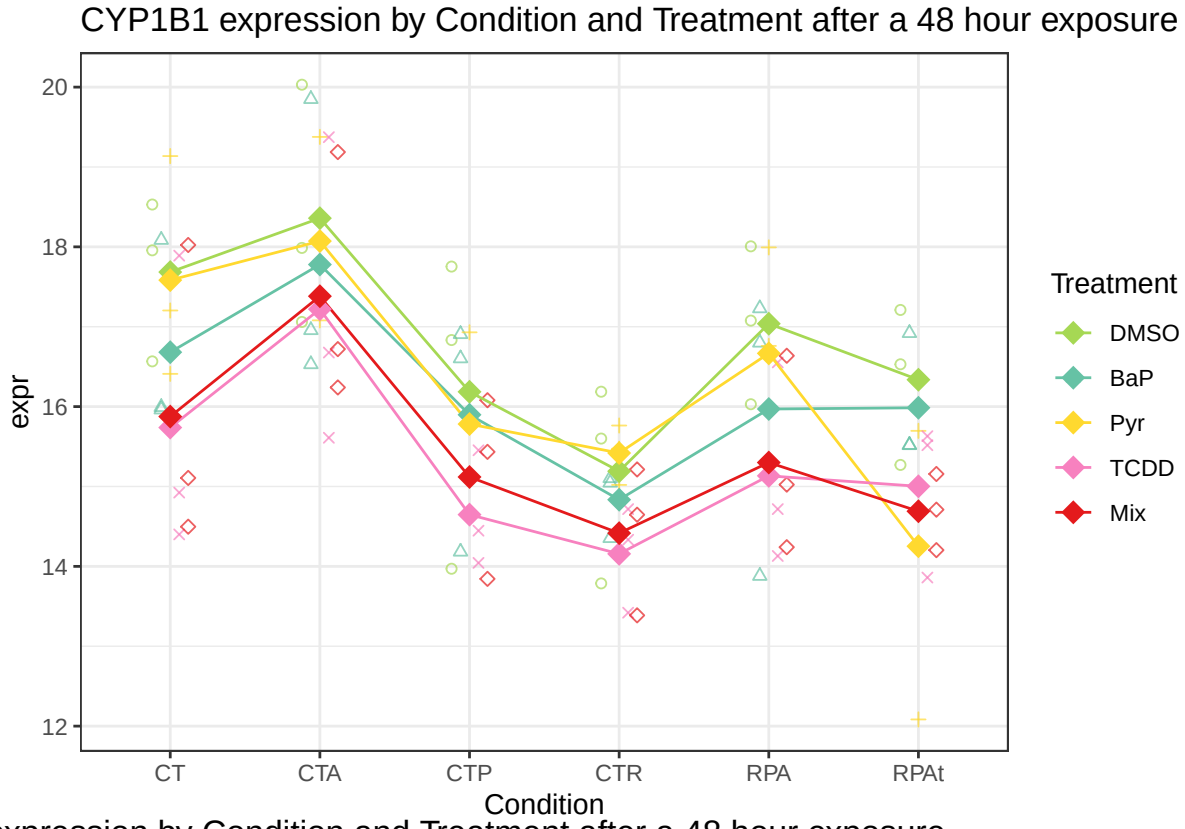
NQO1	CT	CTA	CTP	CTR	CTRPA	CTRPAt
CT	-					
CTA	-	-				
CTP	-	-	-			
CTR	-	-	-	-	*	
CTRPA	-	-	-	-	-	*
CTRPAt	-	-	-	-	-	-

Some genes exhibited a significant difference in expression between pairs of cell lines in the two-way ANOVA without interactions, which are no longer significant in the model with interactions. They are:

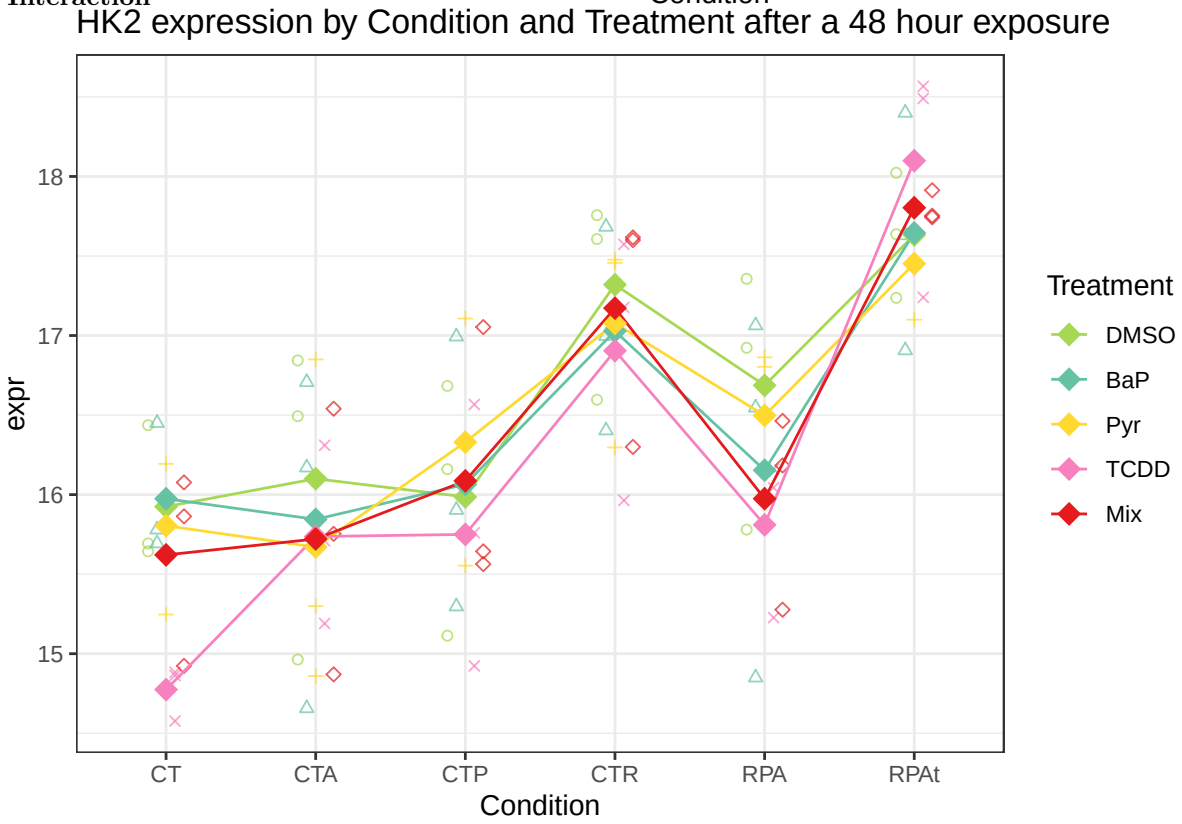
- HMOX expression is no longer found significantly different between CTPRA and CT cell lines.
- NQO1 expression is no longer found significantly different between CTPRA and CT cell lines.

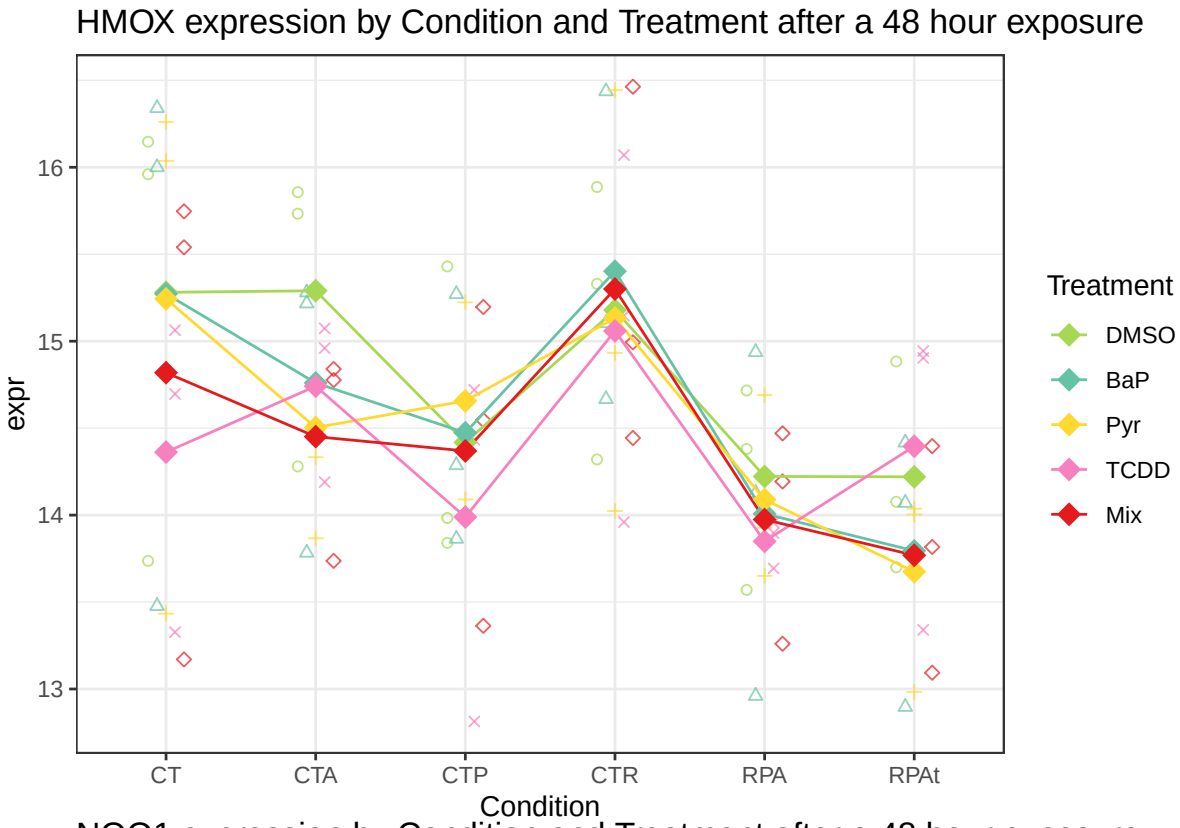
For the other genes, the results are identical.

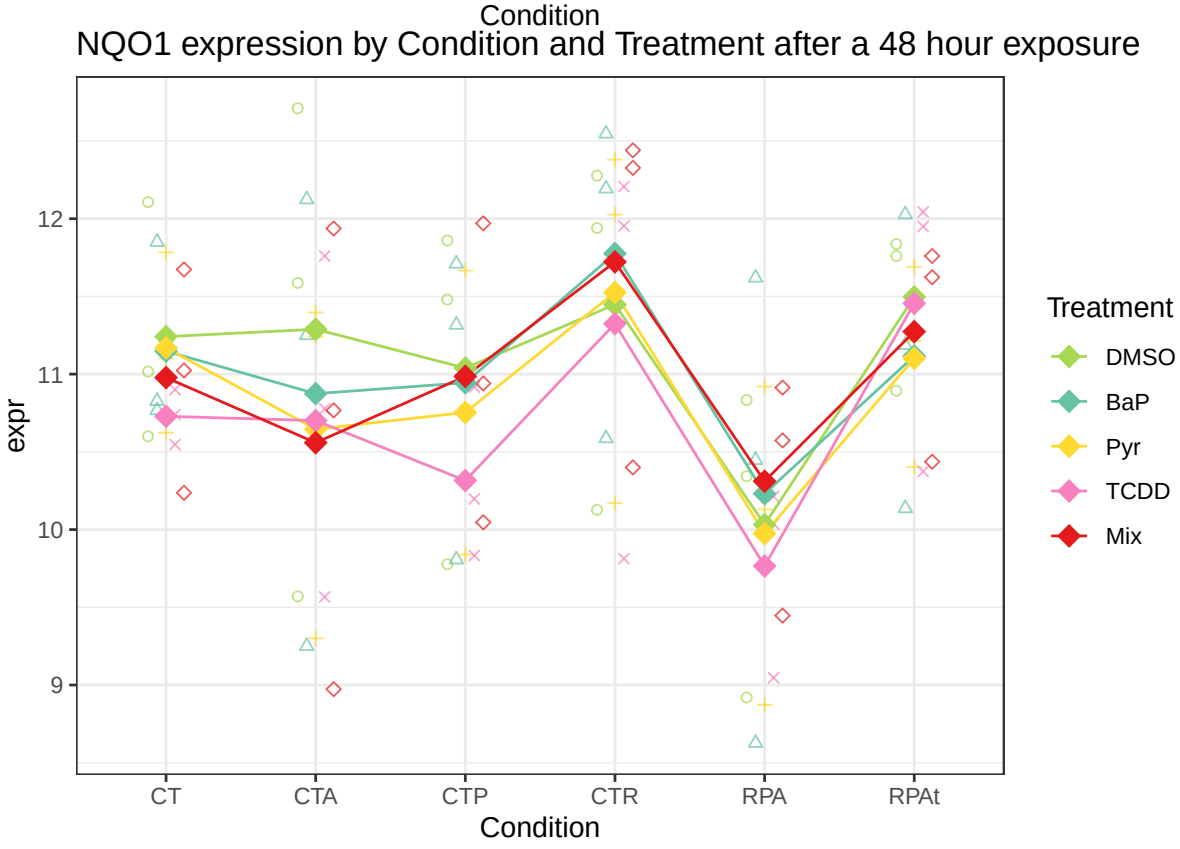
```
all_signif <- names(cell_48[ ,-c(1:3)][ ,selectcell])
sapply(all_signif, function(agene) {
  df <- data.frame(cell_48$Treatment, cell_48[ ,agene], cell_48$Condition)
  names(df) <- c("Treatment", "expr", "Condition")
  p <- ggplot(df, aes(x = Condition, y = expr, group = Treatment,
                      colour = Treatment, shape = Treatment)) +
    geom_jitter(position=position_dodge(0.3), alpha = 0.7) +
    stat_summary(geom = "line", fun = "mean") +
    stat_summary(fun = mean, geom = "point", shape = 18, size = 4) +
    theme_bw() + ggtitle(paste0(agene,
                                " expression by Condition and Treatment after a 48 hour exposure")) +
    scale_color_manual(values = palettetreatment) +
    scale_shape_manual(values = symboltreatment)
  print(p)
  invisible(NULL)
})
```

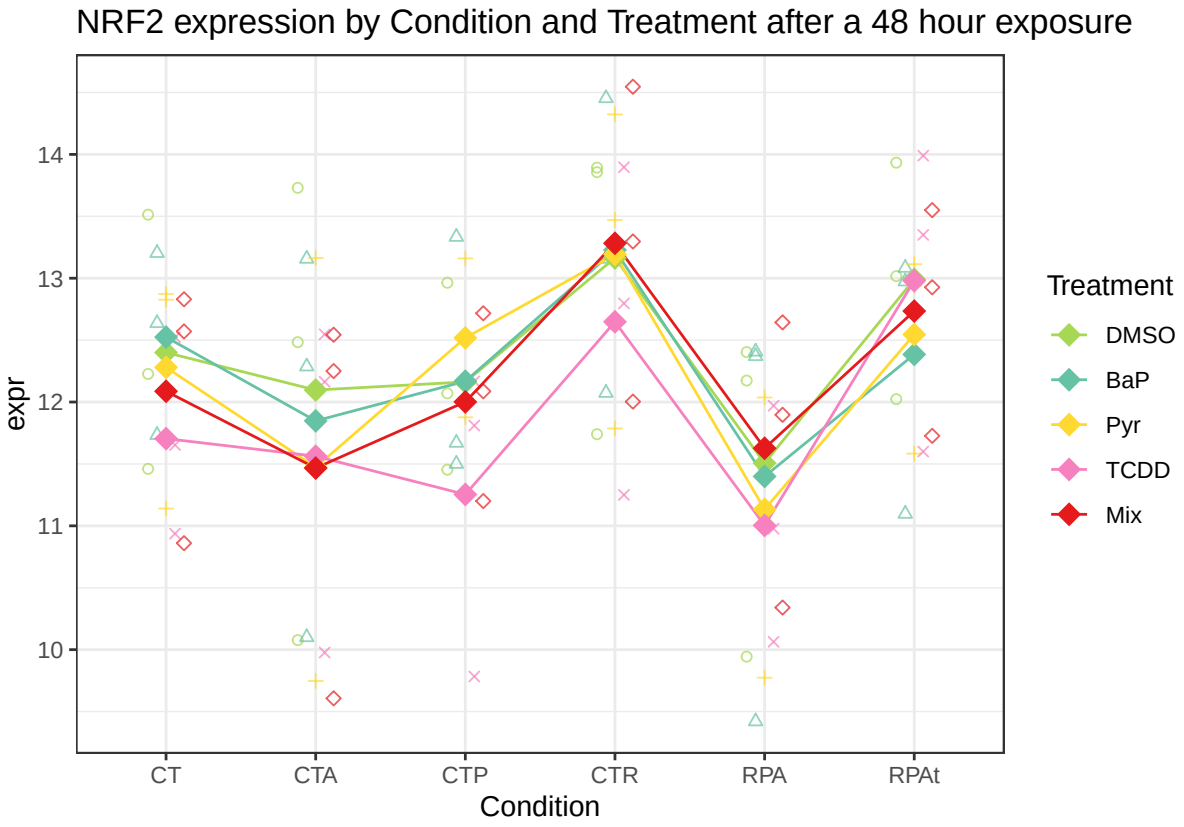


Interaction




HMOX expression by Condition and Treatment after a 48 hour exposure
expr
16
15
14
13
CT
CTA
CTP
CTR
RPA
RPAt
Condition
Treatment
DMSO
BaP
Pyr
TCDD
Mix


NQO1 expression by Condition and Treatment after a 48 hour exposure
expr
12
11
10
9
CT
CTA
CTP
CTR
RPA
RPAt
Condition
Treatment
DMSO
BaP
Pyr
TCDD
Mix




```
## $CYP1B1
## NULL
##
## $HK2
## NULL
##
## $HMOX
## NULL
##
## $NQO1
## NULL
##
## $NRF2
## NULL
```

There is no significant interaction between treatments and cell lines.

- HMOX is significantly over-expressed in CT and CTR compared to CTRPAt, similarly CTR is over -expressed compared to CTRPA.
- NQO1 is significantly under-expressed in CTRPA compared to CTR and CTRPAt.

4.2 After a 120 hour exposure

4.2.2 Two-way ANOVA with interaction

```
anova_120 <- apply(cell_120[ ,-c(1:3)], 2, function(agene) {
  summary(aov(agene ~ cell_120$Condition + cell_120$Treatment +
                cell_120$Condition * cell_120$Treatment))[[1]][1:3,5]
})
rownames(anova_120) = c("Cell type", "Treatment", "Cell type : Treatment")
format(anova_120[,1:8], scientific=TRUE, digits=3) %>%
kable() %>%
  kable_styling(latex_options = c("HOLD_position")) %>%
  column_spec(1:8)
```

	ACO1	AhR	AhRR	ATP5IF1	CAT	CYP1B1	G6PD	HK2
Cell type	1.91e-07	8.95e-03	3.10e-03	7.55e-01	6.78e-06	6.13e-18	1.91e-04	1.41e-14
Treatment	3.26e-01	1.24e-01	2.43e-10	7.95e-01	8.60e-01	2.83e-14	9.55e-01	4.71e-02
Cell type : Treatment	1.00e+00	1.00e+00	9.97e-01	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00

```
format(anova_120[,9:16], scientific=TRUE, digits=3) %>%
kable() %>%
  kable_styling(latex_options = c("HOLD_position")) %>%
  column_spec(1:8)
```

	LPCAT	MCT4	MFN2	ND1	NHE1	NQO1	NRF2	PRDX1
Cell type	1.75e-04	9.34e-01	9.60e-01	3.45e-01	5.84e-01	7.36e-06	8.64e-04	1.17e-02
Treatment	6.58e-01	9.89e-01	9.16e-01	8.72e-01	9.89e-01	8.44e-01	1.69e-01	9.76e-01
Cell type : Treatment	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.45e-01

```
format(anova_120[,17:18], scientific=TRUE, digits=3) %>%
kable() %>%
  kable_styling(latex_options = c("HOLD_position")) %>%
  column_spec(1:2)
```

	SCD1	TFAM
Cell type	4.31e-01	1.38e-01
Treatment	3.19e-02	9.85e-01
Cell type : Treatment	1.00e+00	1.00e+00

Treatment The number of genes for which the null hypothesis is rejected for the treatments is 4 (at 5%) and these genes are: AhRR, CYP1B1, HK2, SCD1.

For these genes, post-hoc tests are performed:

```
selecttreat <- which(anova_120[2,] < 0.05)
anova_posthoc_120 <- apply(cell_120[ ,-c(1:3)][ ,selecttreat], 2, function(agene) {
  TukeyHSD(aov(agene ~ cell_120$Condition + cell_120$Treatment +
                cell_120$Condition * cell_120$Treatment))[[2]][,4]
})
format(anova_posthoc_120, scientific=TRUE, digits=3) %>% kable() %>%
  kable_styling(bootstrap_options = "striped") %>%
  add_header_above(c(" " = 1, "p-values" = length(selecttreat)))
```

	p-values			
	AhRR	CYP1B1	HK2	SCD1
BaP-DMSO	9.76e-01	6.58e-01	8.68e-01	9.98e-01
Pyr-DMSO	9.95e-01	1.00e+00	9.80e-01	9.93e-01
TCDD-DMSO	2.23e-06	1.26e-09	1.40e-01	2.02e-01
Mix-DMSO	2.42e-05	3.77e-08	8.64e-02	3.06e-01
Pyr-BaP	8.66e-01	7.10e-01	9.94e-01	9.45e-01
TCDD-BaP	2.60e-05	3.92e-07	6.44e-01	3.52e-01
Mix-BaP	2.44e-04	1.32e-05	5.07e-01	4.89e-01
TCDD-Pyr	4.28e-07	1.66e-09	3.91e-01	7.99e-02
Mix-Pyr	5.02e-06	5.50e-08	2.77e-01	1.34e-01
Mix-TCDD	9.79e-01	9.22e-01	1.00e+00	9.99e-01

For each gene, the number of pairs and the pairs of treatments that show significant differences in expression is:

```
sapply(1:ncol(anova_posthoc_120), function(ind) {
  cat(paste0("For gene: ", colnames(anova_posthoc_120)[ind], "\n"))
  cat(paste0("Number of significant pairs: ",
             sum(anova_posthoc_120[ ,ind] < 0.05), "\n"))
  cat(names(which(anova_posthoc_120[ ,ind] < 0.05)), "\n\n")
})
```

```
## For gene: AhRR
## Number of significant pairs: 6
## TCDD-DMSO Mix-DMSO TCDD-BaP Mix-BaP TCDD-Pyr Mix-Pyr
##
## For gene: CYP1B1
## Number of significant pairs: 6
## TCDD-DMSO Mix-DMSO TCDD-BaP Mix-BaP TCDD-Pyr Mix-Pyr
##
## For gene: HK2
## Number of significant pairs: 0
##
##
## For gene: SCD1
## Number of significant pairs: 0
##
```

```
## [[1]]
## NULL
##
## [[2]]
## NULL
##
## [[3]]
## NULL
##
## [[4]]
## NULL
```

Cell types The number of genes for which the null hypothesis is rejected for the cell types is 5 (at 5%) and these genes are: CYP1B1, HK2, HMOX, NQO1, NRF2.

For these genes, post-hoc tests are performed:

```
selectcell <- which(anova_120[1,] < 0.05)
anova_posthoc_120 <- apply(cell_120[ ,-c(1:3)][ ,selectcell], 2, function(agene) {
  TukeyHSD(aov(agene ~ cell_120$Condition + cell_120$Treatment +
                cell_120$Condition * cell_120$Treatment))[[1]][,4]
})
format(anova_posthoc_120[,1:6], scientific=TRUE, digits=3) %>%
kable() %>%
  kable_styling(latex_options = c("HOLD_position")) %>%
  column_spec(1:6)
```

	ACO1	AhR	AhRR	CAT	CYP1B1	G6PD
CTA-CT	9.86e-01	1.00e+00	9.94e-01	5.73e-01	8.59e-01	9.40e-01
CTP-CT	9.91e-01	1.00e+00	4.37e-03	1.00e+00	7.39e-01	1.00e+00
CTR-CT	1.37e-05	2.33e-02	7.46e-01	3.47e-04	4.35e-10	6.40e-02
RPA-CT	1.00e+00	9.31e-01	2.31e-01	7.18e-01	1.91e-01	7.74e-01
RPAt-CT	1.36e-01	1.00e+00	8.52e-02	4.60e-04	1.62e-04	2.23e-02
CTP-CTA	8.14e-01	1.00e+00	2.39e-02	7.32e-01	1.27e-01	9.02e-01
CTR-CTA	1.14e-06	2.04e-02	9.63e-01	5.26e-02	4.35e-10	4.26e-03
RPA-CTA	9.64e-01	9.17e-01	5.43e-01	1.00e+00	9.63e-03	2.29e-01
RPAt-CTA	2.84e-02	1.00e+00	2.72e-01	6.44e-02	1.44e-06	1.12e-03
CTR-CTP	1.48e-04	3.31e-02	1.80e-01	8.29e-04	4.54e-10	8.60e-02
RPA-CTP	9.98e-01	9.61e-01	6.69e-01	8.54e-01	9.30e-01	8.36e-01
RPAt-CTP	4.08e-01	1.00e+00	9.06e-01	1.09e-03	1.84e-02	3.13e-02
RPA-CTR	2.64e-05	2.32e-01	9.53e-01	2.89e-02	2.12e-09	6.67e-01
RPAt-CTR	7.11e-02	1.94e-02	7.66e-01	1.00e+00	4.40e-05	9.99e-01
RPAt-RPA	1.91e-01	9.11e-01	9.97e-01	3.59e-02	1.99e-01	4.17e-01

```
format(anova_posthoc_120[,7:11], scientific=TRUE, digits=3) %>%
kable() %>%
  kable_styling(latex_options = c("HOLD_position")) %>%
  column_spec(1:5)
```

	HK2	LPCAT	NQO1	NRF2	PRDX1
CTA-CT	1.00e+00	9.18e-01	1.00e+00	9.82e-01	8.88e-01
CTP-CT	1.04e-02	9.99e-01	7.11e-03	7.85e-01	5.38e-01
CTR-CT	8.51e-08	6.95e-02	2.78e-03	3.25e-02	9.34e-01
RPA-CT	1.32e-02	9.97e-01	7.74e-01	9.89e-01	1.52e-01
RPAt-CT	4.68e-10	2.02e-02	1.66e-03	6.30e-02	1.00e+00
CTP-CTA	5.04e-03	7.26e-01	6.53e-03	3.52e-01	9.89e-01
CTR-CTA	3.08e-08	3.71e-03	2.54e-03	3.84e-03	3.43e-01
RPA-CTA	6.53e-03	6.81e-01	7.59e-01	7.73e-01	7.37e-01
RPAt-CTA	4.46e-10	7.55e-04	1.51e-03	8.62e-03	8.33e-01
CTR-CTP	4.18e-02	1.75e-01	1.00e+00	4.88e-01	1.03e-01
RPA-CTP	1.00e+00	1.00e+00	2.21e-01	9.84e-01	9.73e-01
RPAt-CTP	1.48e-04	6.08e-02	9.98e-01	6.50e-01	4.58e-01
RPA-CTR	3.36e-02	2.03e-01	1.22e-01	1.50e-01	1.44e-02
RPAt-CTR	5.40e-01	9.97e-01	1.00e+00	1.00e+00	9.63e-01
RPAt-RPA	1.08e-04	7.32e-02	8.58e-02	2.48e-01	1.16e-01

For each gene, the number of pairs and the pairs of cell types that show significant differences in expression is:

```r
sapply(1:ncol(anova_posthoc_120), function(ind) {
  cat(paste0("For gene: ", colnames(anova_posthoc_120)[ind], "\n"))
  cat(paste0("Number of significant pairs: ",
             sum(anova_posthoc_120[ ,ind] < 0.05), "\n"))
  cat(names(which(anova_posthoc_120[ ,ind] < 0.05)), "\n\n")
})
```

```
## For gene: ACO1
## Number of significant pairs: 5
## CTR-CT CTR-CTA RPAt-CTA CTR-CTP RPA-CTR
##
## For gene: AhR
## Number of significant pairs: 4
## CTR-CT CTR-CTA CTR-CTP RPAt-CTR
##
## For gene: AhRR
## Number of significant pairs: 2
## CTP-CT CTP-CTA
##
## For gene: CAT
## Number of significant pairs: 6
## CTR-CT RPAt-CT CTR-CTP RPAt-CTP RPA-CTR RPAt-RPA
##
## For gene: CYP1B1
## Number of significant pairs: 9
## CTR-CT RPAt-CT CTR-CTA RPA-CTA RPAt-CTA CTR-CTP RPAt-CTP RPA-CTR RPAt-CTR
##
## For gene: G6PD
## Number of significant pairs: 4
## RPAt-CT CTR-CTA RPAt-CTA RPAt-CTP
##
## For gene: HK2
## Number of significant pairs: 12
## CTP-CT CTR-CT RPA-CT RPAt-CT CTP-CTA CTR-CTA RPA-CTA RPAt-CTA CTR-CTP RPAt-CTP RPA-CTR RPAt-RPA
##
## For gene: LPCAT
## Number of significant pairs: 3
## RPAt-CT CTR-CTA RPAt-CTA
##
## For gene: NQO1
## Number of significant pairs: 6
## CTP-CT CTR-CT RPAt-CT CTP-CTA CTR-CTA RPAt-CTA
##
## For gene: NRF2
## Number of significant pairs: 3
## CTR-CT CTR-CTA RPAt-CTA
##
## For gene: PRDX1
## Number of significant pairs: 1
## RPA-CTR
```

```
## [[1]]
## NULL
##
```

```
## [[2]]
## NULL
##
## [[3]]
## NULL
##
## [[4]]
## NULL
##
## [[5]]
## NULL
##
## [[6]]
## NULL
##
## [[7]]
## NULL
##
## [[8]]
## NULL
##
## [[9]]
## NULL
##
## [[10]]
## NULL
##
## [[11]]
## NULL
```

4.2.2 Conclusion

Treatments The following table represents all genes for which at least one global test is significant

- first column: which test(s) is (are) significant?
- second column: for which treatment(s) is normality rejected for this gene?

Treatments	Tests giving these results	trestment(s) whose normality Shapiro rejects
AhRR	two-factor ANOVA test	-
CYP1B1	two-factor ANOVA test	Pyr
HK2	two-factor ANOVA test	Mix
SCD1	two-factor ANOVA test	-

The following tables are constructed as follows:

- the upper parts of the tables are composed by the results of the post-hoc two-factor ANOVA test;
- A star means that the contrast between the two treatments is significant (p-value < 0.05);

SCD1	DMSO	BaP	Pyr	TCDD	Mix
DMSO	-				
BaP	-	-			

SCD1	DMSO	BaP	Pyr	TCDD	Mix
Pyr	-	-	-		
TCDD	-	-	-	-	
Mix	-	-	-	-	-

Two way ANOVA without interaction showed a significant difference in expression between TCDD and Pyr treatment that is not found in the model with interaction for the gene SCD1. For the genes AhRR, CYP1B1 and HK2 the results are identical.

Cell types The following table represents all genes for which at least one global test is significant

- first column: which test(s) is (are) significant?
- second column: for which cell type(s) is normality rejected for this gene?

Cell types	Tests giving these results	Cell Type(s) whose normality Shapiro rejects
ACO1	two-factor ANOVA test	CTA, CTRPA, CTRPAt
AhR	two-factor ANOVA test	CTP, CTR
AhRR	two-factor ANOVA test	-
CAT	two-factor ANOVA test	CTR, CTRPAt
CYP1B1	two-factor ANOVA test	CT
G6PD	two-factor ANOVA test	-
HK2	two-factor ANOVA test	-
LCPAT	two-factor ANOVA test	CTA, CTR, CTRPAt
NQO1	two-factor ANOVA test	CT, CTR, CTRPAt
NRF2	two-factor ANOVA test	CT, CTR, CTRPA, CTRPAt
PRDX1	two-factor ANOVA test	CT, CTR

The following tables are constructed as follows:

- the upper parts of the tables are composed by the results of the post-hoc two-factor ANOVA test;
- A star means that the contrast between the two cell types is significant (p-value < 0.05);

ACO1	CT	CTA	CTP	CTR	CTRPA	CTRPAt
CT	-			*		
CTA	-	-		*		*
CTP	-	-	-	*		
CTR	-	-	-	-	*	
CTRPA	-	-	-	-	-	
CTRPAt	-	-	-	-	-	-

CAT	CT	CTA	CTP	CTR	CTRPA	CTRPAt
CT	-			*		*
CTA	-	-				
CTP	-	-	-	*		*
CTR	-	-	-	-	*	
CTRPA	-	-	-	-	-	*
CTRPAt	-	-	-	-	-	-

G6PD	CT	CTA	CTP	CTR	CTRPA	CTRPAt
CT	-					*
CTA	-	-		*		*
CTP	-	-	-			*
CTR	-	-	-	-		
CTRPA	-	-	-	-	-	
CTRPAt	-	-	-	-	-	-

LPCAT	CT	CTA	CTP	CTR	CTRPA	CTRPAt
CT	-					*
CTA	-	-		*		*
CTP	-	-	-			
CTR	-	-	-	-		
CTRPA	-	-	-	-	-	
CTRPAt	-	-	-	-	-	-

NQO1	CT	CTA	CTP	CTR	CTRPA	CTRPAt
CT	-		*	*		*
CTA	-	-	*	*		*
CTP	-	-	-			
CTR	-	-	-	-		
CTRPA	-	-	-	-	-	
CTRPAt	-	-	-	-	-	-

NRF2	CT	CTA	CTP	CTR	CTRPA	CTRPAt
CT	-			*		
CTA	-	-		*		*
CTP	-	-	-			
CTR	-	-	-	-		
CTRPA	-	-	-	-	-	
CTRPAt	-	-	-	-	-	-

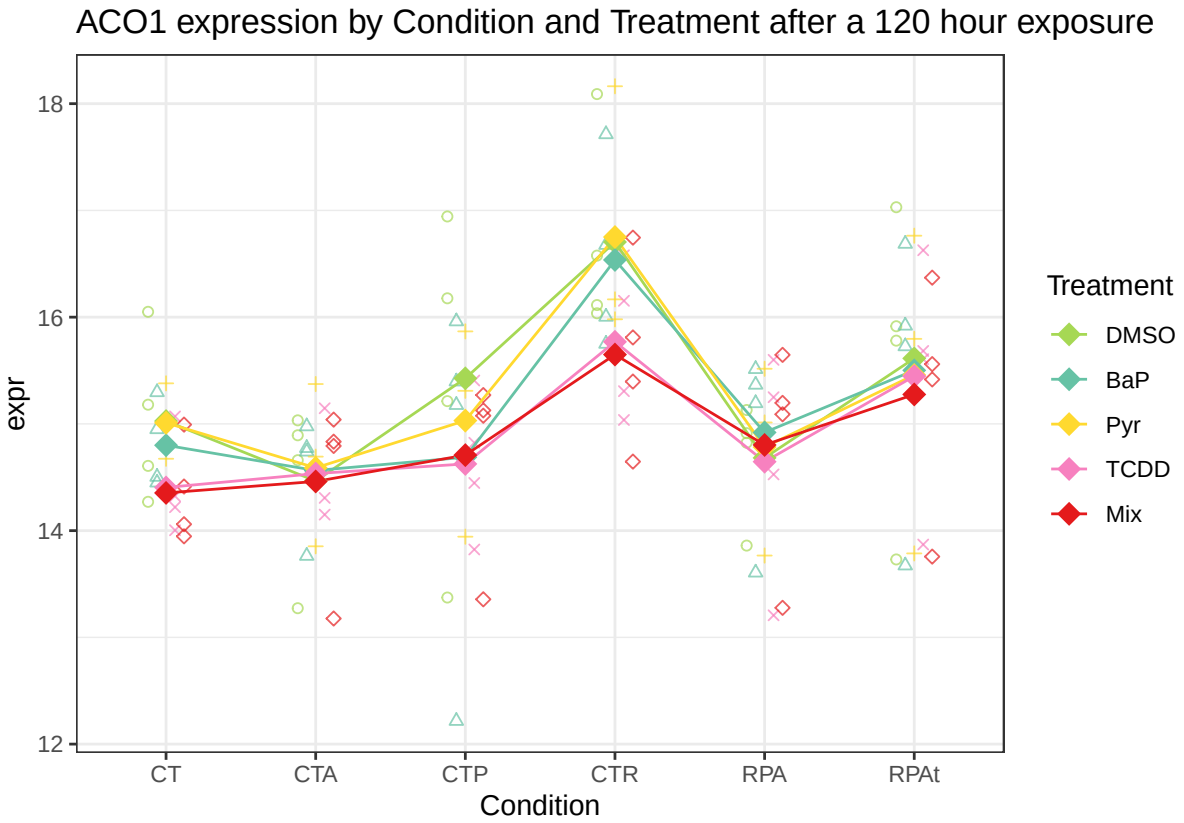
Some genes exhibited a significant difference in expression between pairs of cell lines in the two-way ANOVA without interactions, which are no longer significant in the model with interactions. They are:

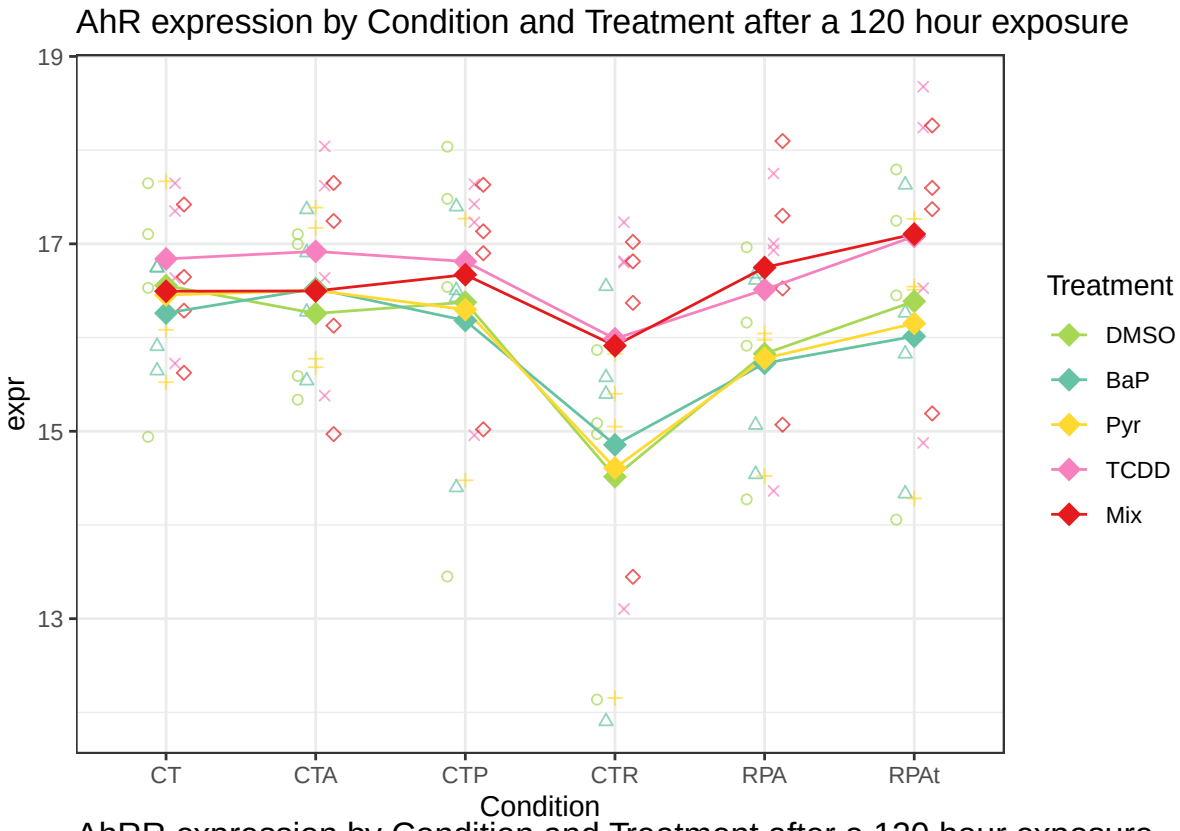
- for ACO1, CTR/CTRPAt;
- for CAT, CTA/CTR and CTA/CTRPAt;
- for G6PD, CT/CTR and CTP/CTR;
- for LPCAT, CT/CTR, CTP/CTRPAt and CTRPA/CTRPAt;
- for NQO1, CTRPA/CTRPAt;
- for NRF2, CT/CTRPAt;

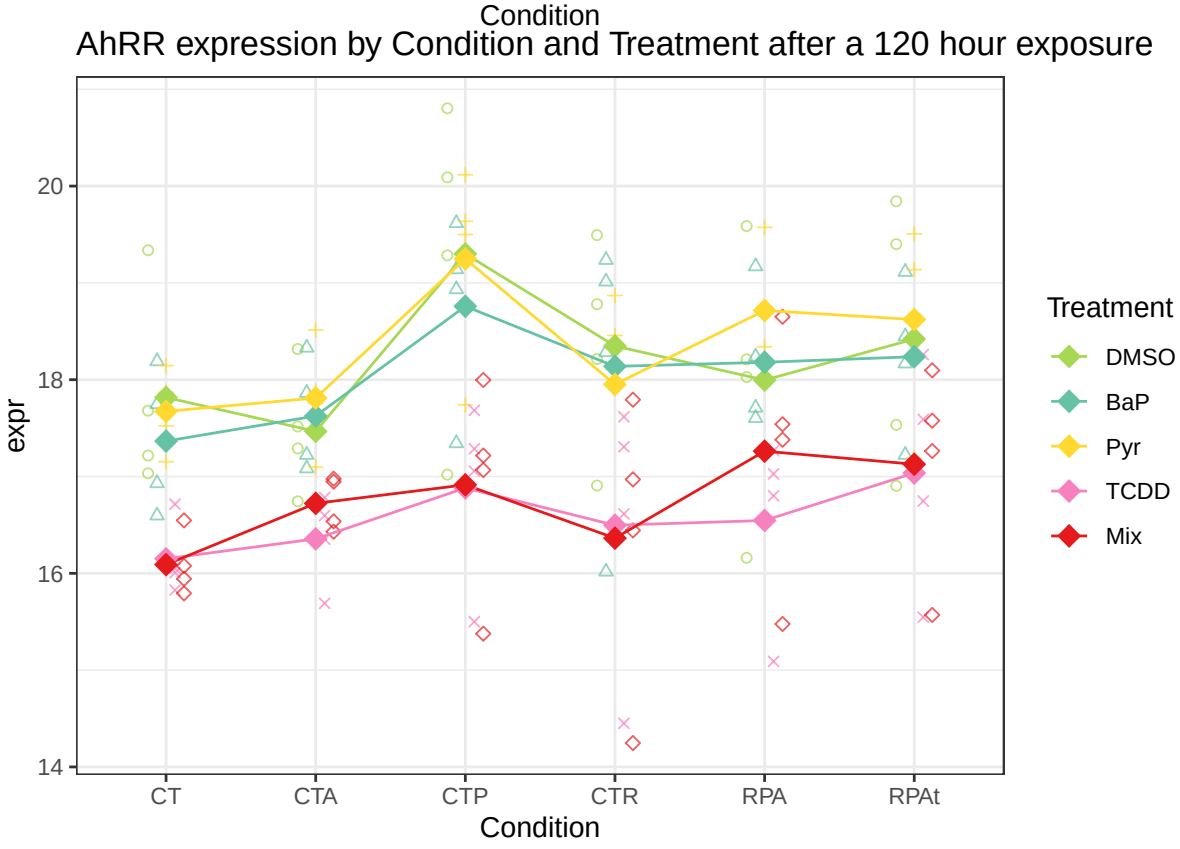
For the other genes, the results are identical.

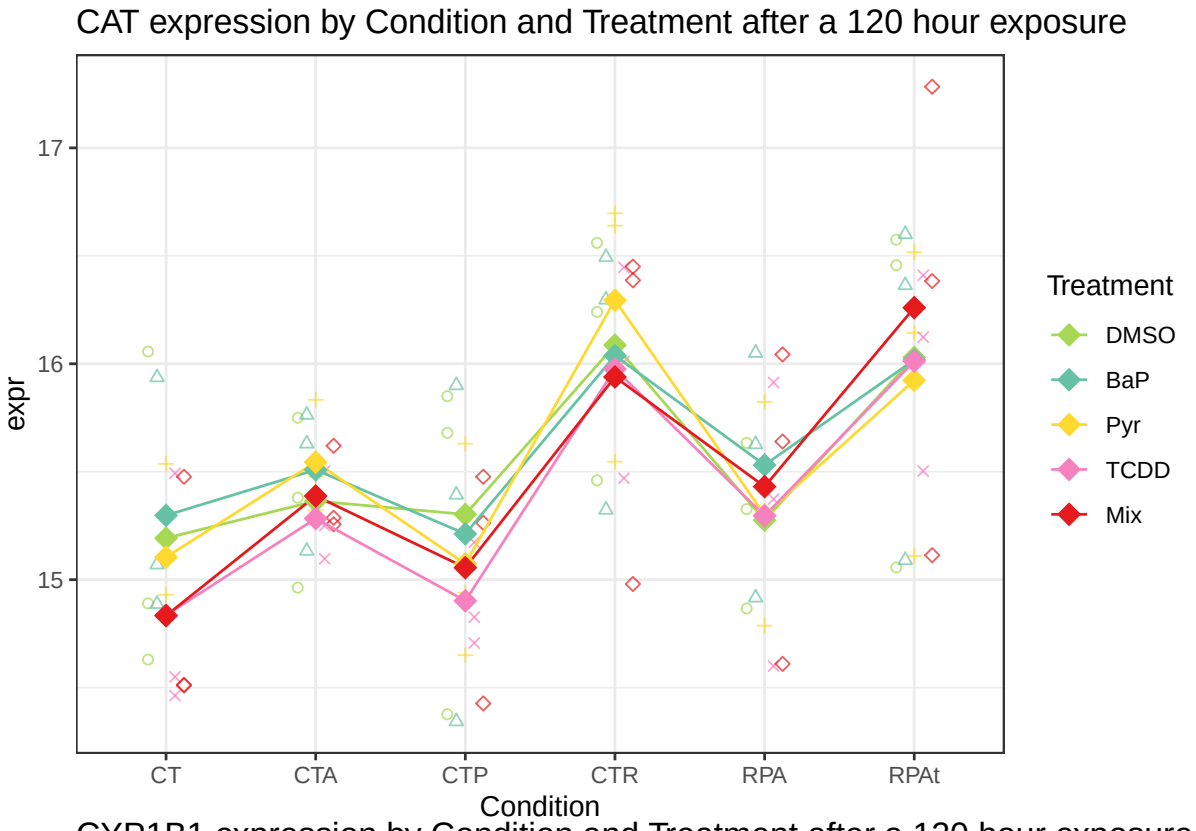
```
all_signif <- names(cell_120[ ,-c(1:3)][ ,selectcell])
sapply(all_signif, function(agene) {
  df <- data.frame(cell_120$Treatment, cell_120[ ,agene], cell_120$Condition)
  names(df) <- c("Treatment", "expr", "Condition")
```

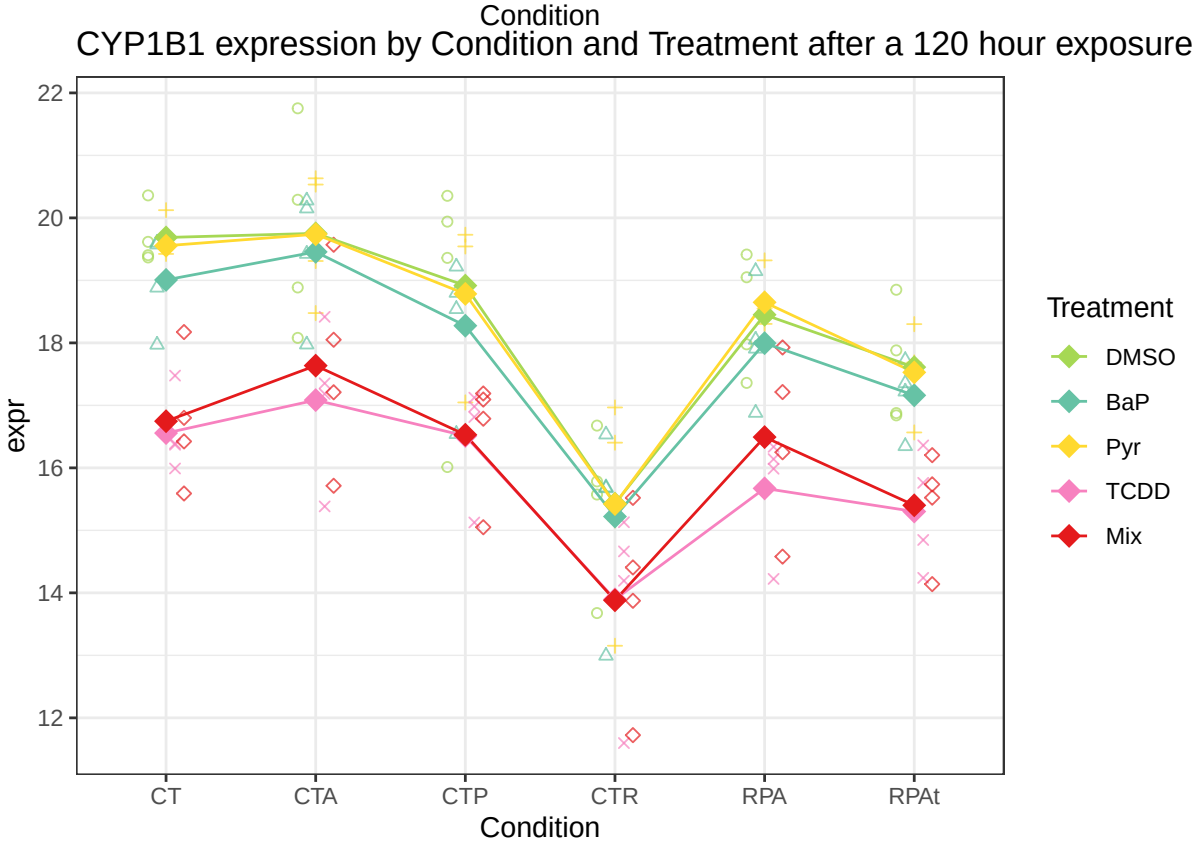
```
  p <- ggplot(df, aes(x = Condition, y = expr, group = Treatment,
                      colour = Treatment, shape = Treatment)) +
    geom_jitter(position=position_dodge(0.3), alpha = 0.7) +
    stat_summary(geom = "line", fun = "mean") +
    stat_summary(fun = mean, geom = "point", shape = 18,size = 4) +
    theme_bw() + ggtitle(paste0(agene,
                  " expression by Condition and Treatment after a 120 hour exposure")) +
    scale_color_manual(values = palettetreatment) +
    scale_shape_manual(values = symboltreatment)

  print(p)
  invisible(NULL)
})
```

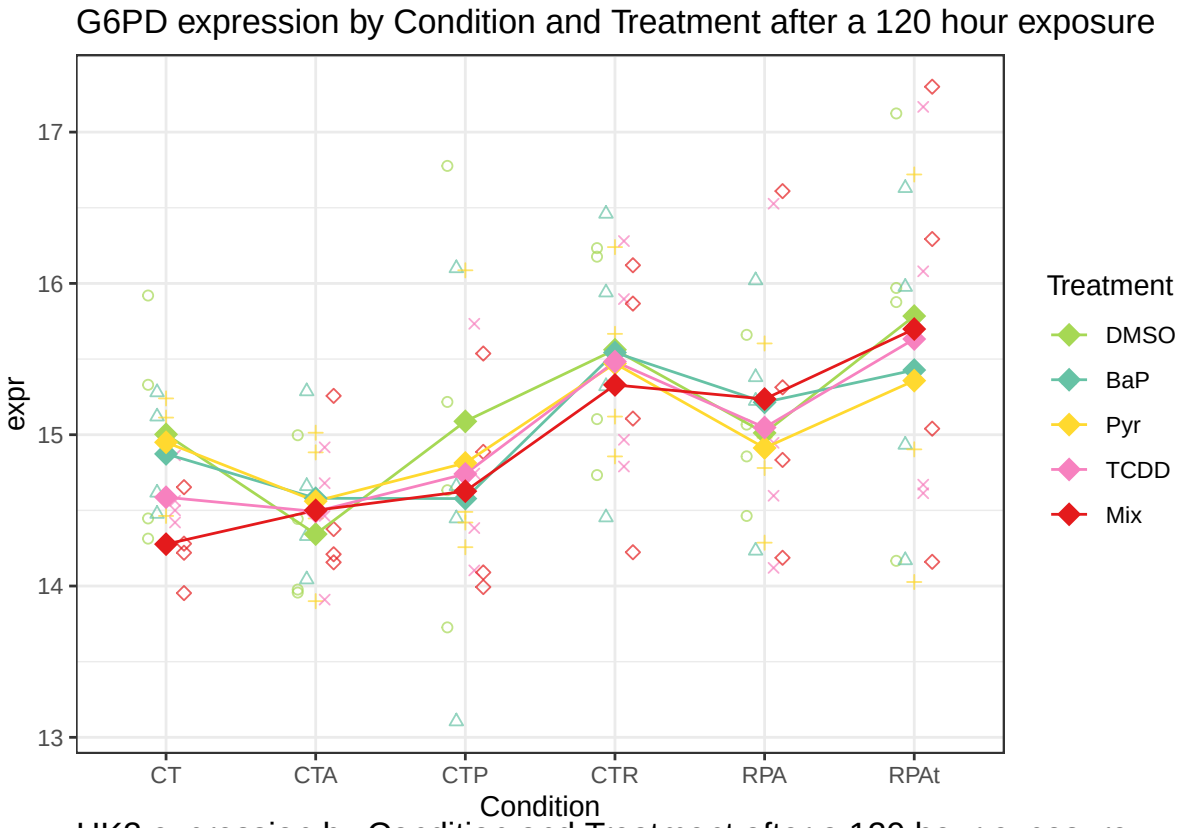


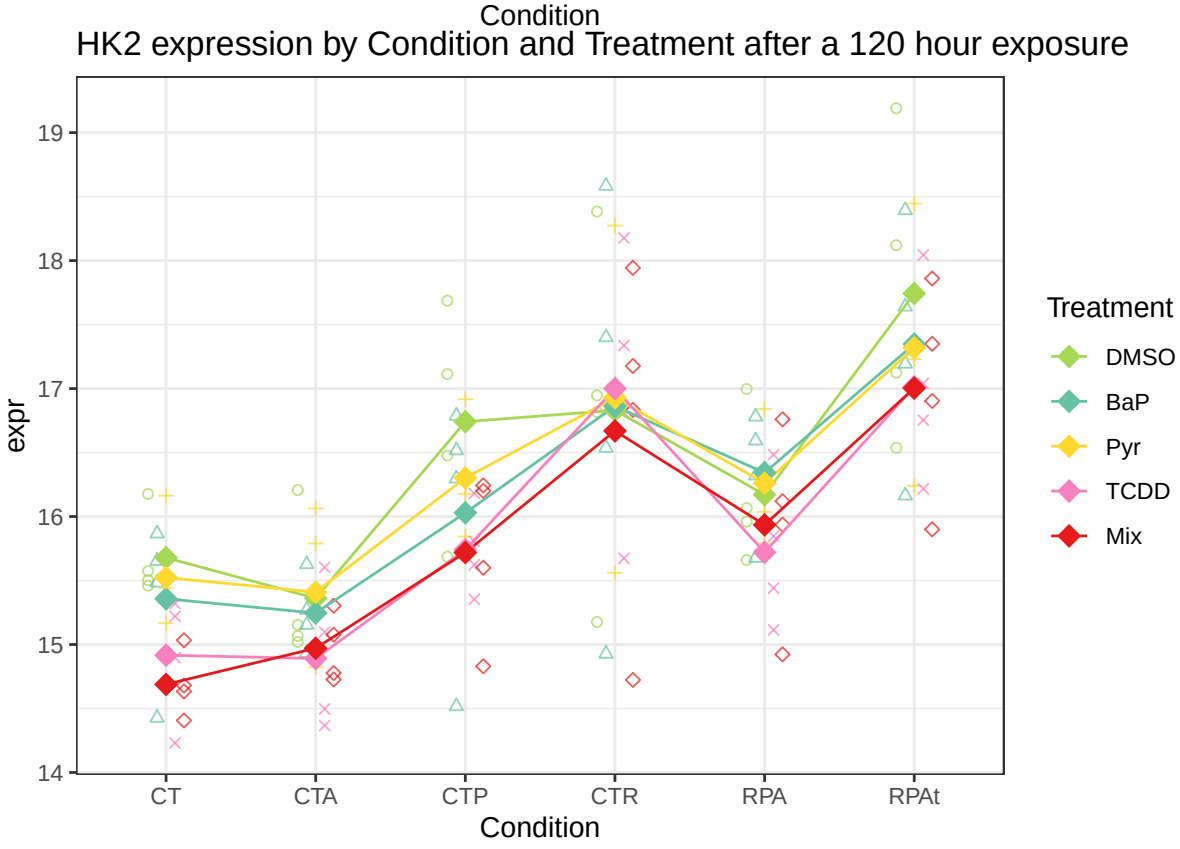

AhR expression by Condition and Treatment after a 120 hour exposure
19
17
15
13
expr
CT
CTA
CTP
CTR
RPA
RPAt
Condition
Treatment
DMSO
BaP
Pyr
TCDD
Mix

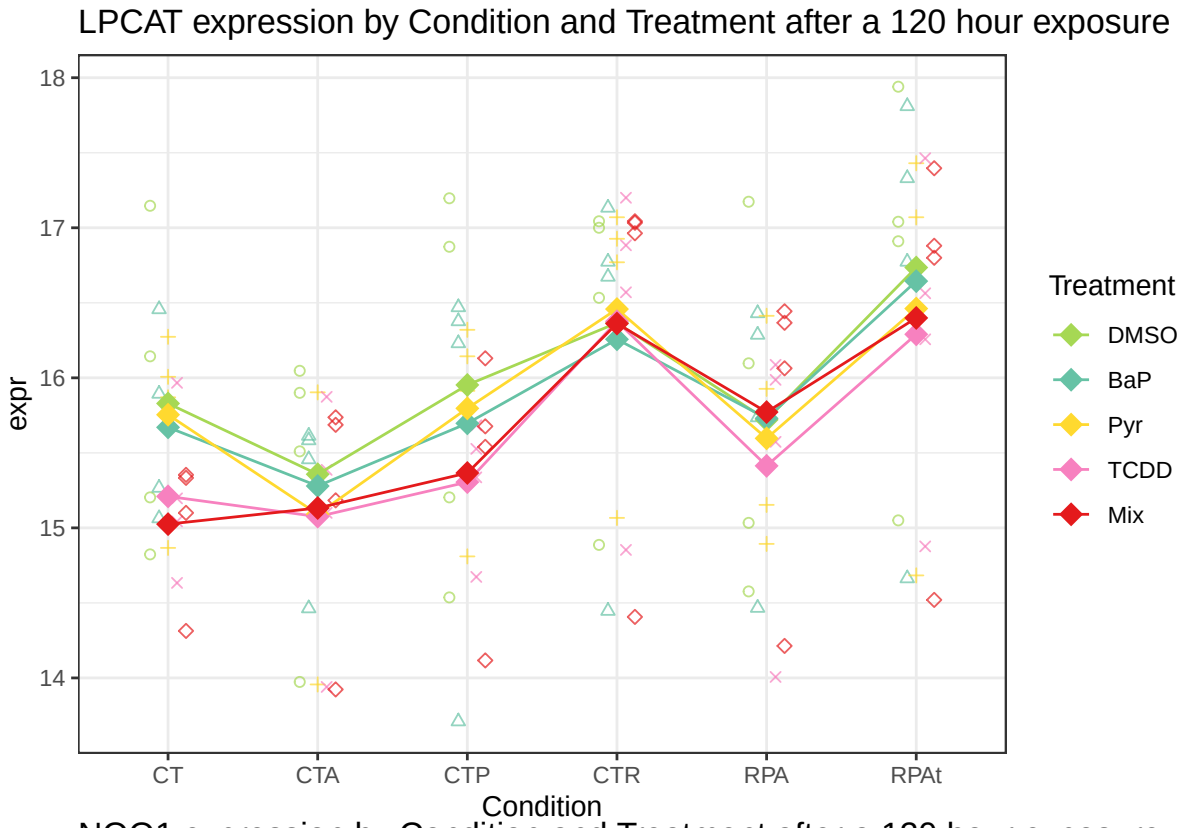

AhRR expression by Condition and Treatment after a 120 hour exposure
20
18
16
14
expr
CT
CTA
CTP
CTR
RPA
RPAt
Condition
Treatment
DMSO
BaP
Pyr
TCDD
Mix

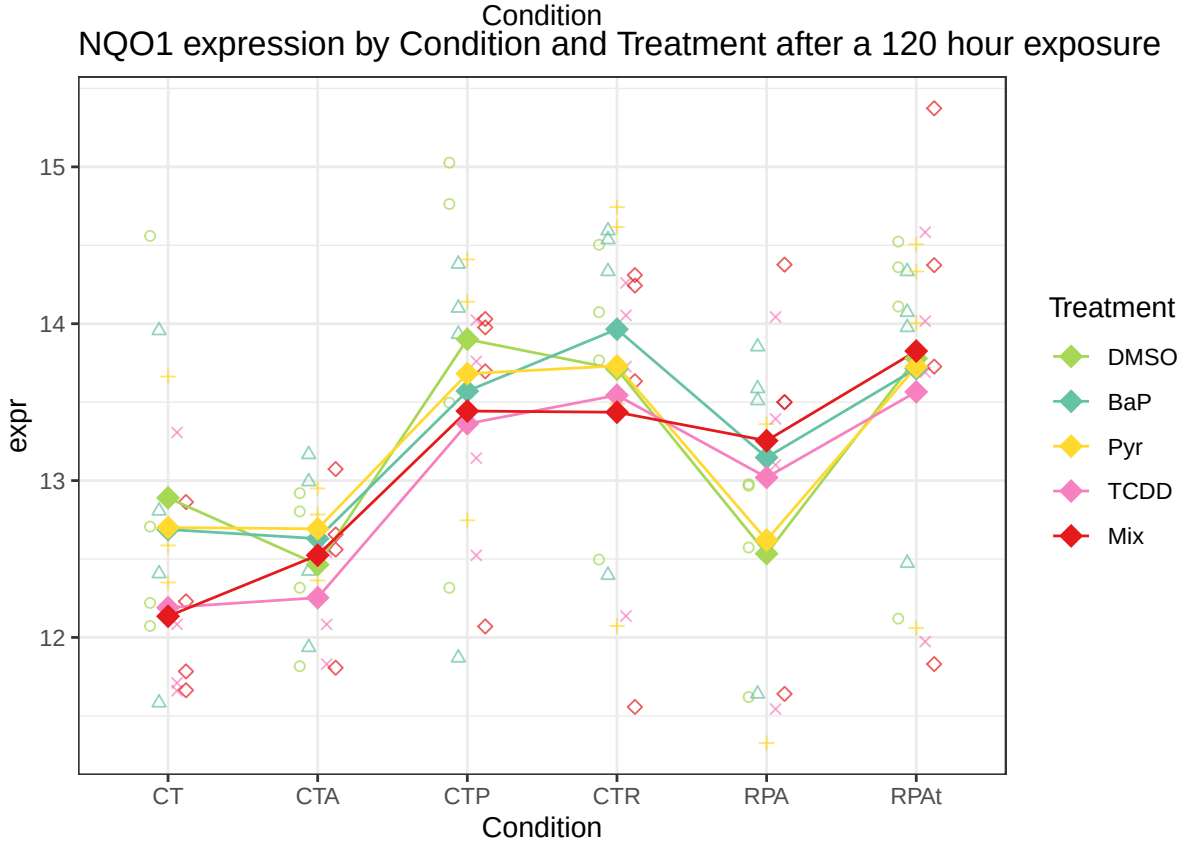

CAT expression by Condition and Treatment after a 120 hour exposure
expr
17
16
15
CT
CTA
CTP
CTR
RPA
RPAt
Condition
Treatment
DMSO
BaP
Pyr
TCDD
Mix

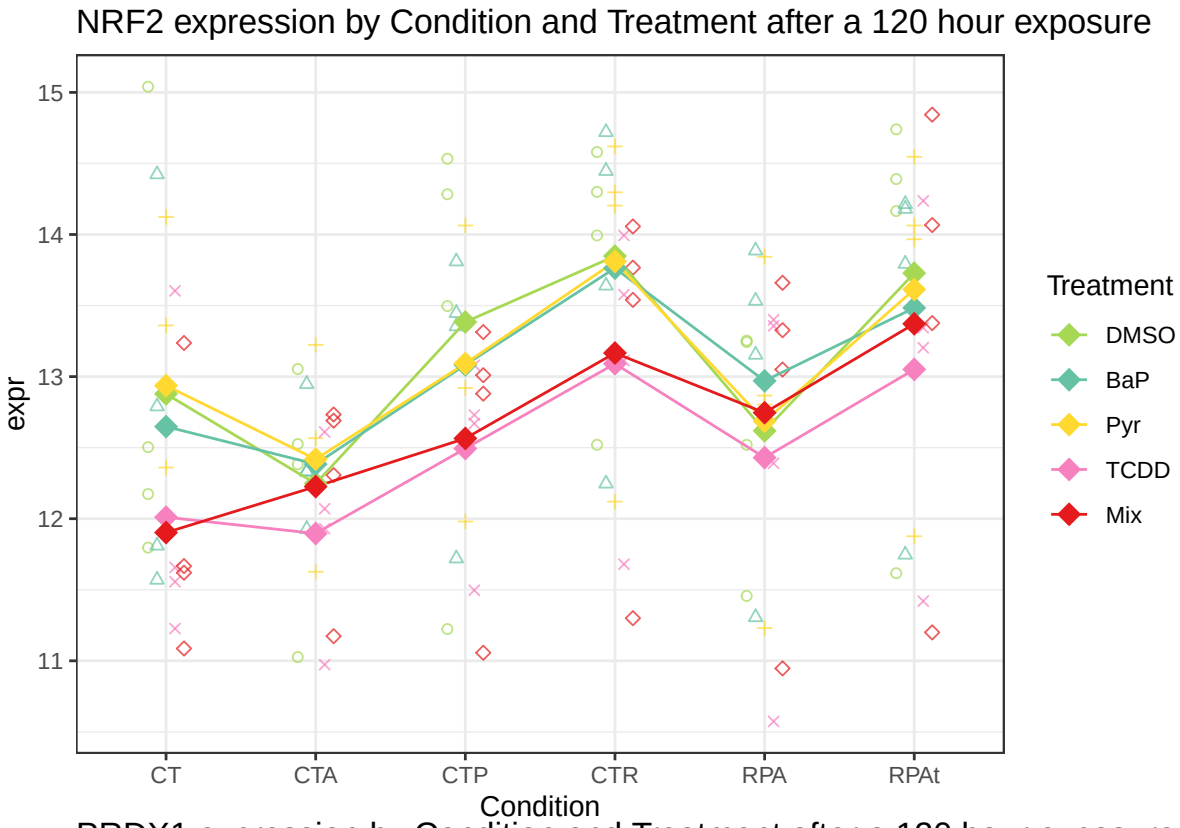

CYP1B1 expression by Condition and Treatment after a 120 hour exposure
expr
22
20
18
16
14
12
CT
CTA
CTP
CTR
RPA
RPAt
Condition
Treatment
DMSO
BaP
Pyr
TCDD
Mix

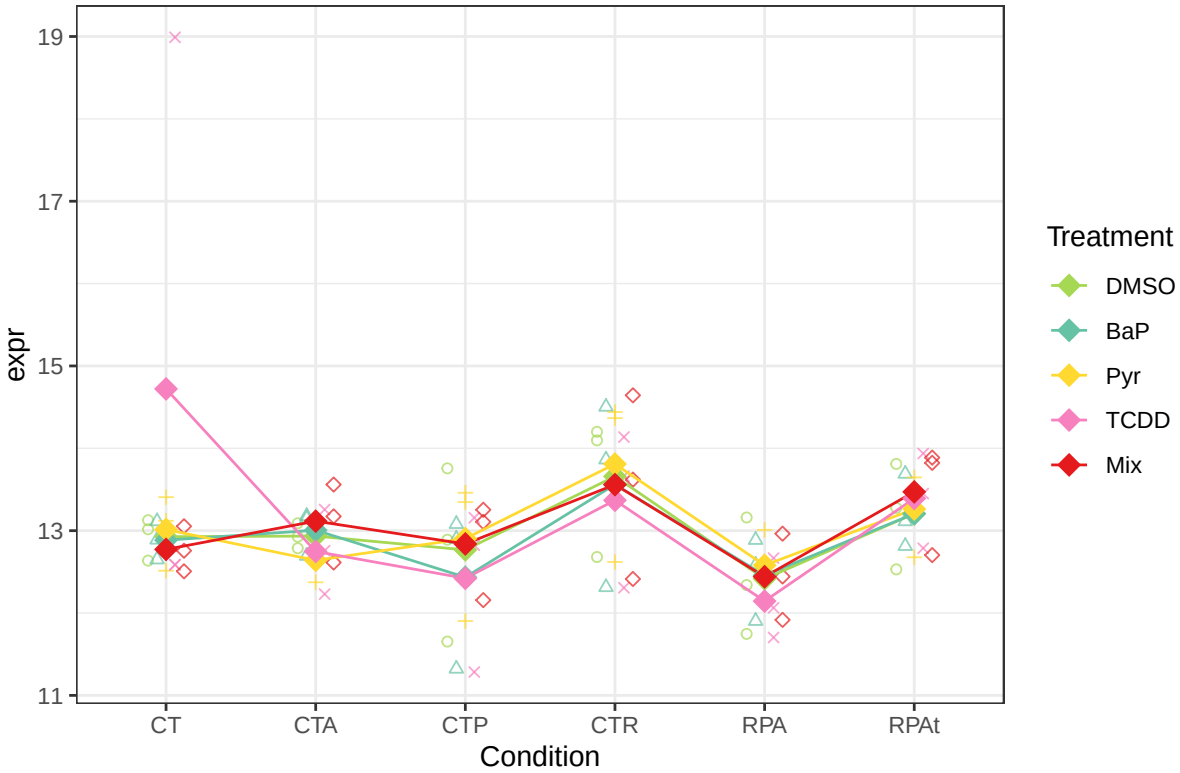

G6PD expression by Condition and Treatment after a 120 hour exposure
expr
17
16
15
14
13
CT
CTA
CTP
CTR
RPA
RPAt
Condition
Treatment
DMSO
BaP
Pyr
TCDD
Mix


HK2 expression by Condition and Treatment after a 120 hour exposure
expr
19
18
17
16
15
14
CT
CTA
CTP
CTR
RPA
RPAt
Condition
Treatment
DMSO
BaP
Pyr
TCDD
Mix


LPCAT expression by Condition and Treatment after a 120 hour exposure
18
17
16
15
14
expr
CT
CTA
CTP
CTR
RPA
RPAt
Condition
Treatment
DMSO
BaP
Pyr
TCDD
Mix


NQO1 expression by Condition and Treatment after a 120 hour exposure
15
14
13
12
expr
CT
CTA
CTP
CTR
RPA
RPAt
Condition
Treatment
DMSO
BaP
Pyr
TCDD
Mix


NRF2 expression by Condition and Treatment after a 120 hour exposure
expr
15
14
13
12
11
CT
CTA
CTP
CTR
RPA
RPAt
Condition
Treatment
DMSO
BaP
Pyr
TCDD
Mix


PRDX1 expression by Condition and Treatment after a 120 hour exposure
expr
19
17
15
13
11
CT
CTA
CTP
CTR
RPA
RPAt
Condition
Treatment
DMSO
BaP
Pyr
TCDD
Mix

```
## $ACO1
```

```
## NULL
##
## $AhR
## NULL
##
## $AhRR
## NULL
##
## $CAT
## NULL
##
## $CYP1B1
## NULL
##
## $G6PD
## NULL
##
## $HK2
## NULL
##
## $LPCAT
## NULL
##
## $NQO1
## NULL
##
## $NRF2
## NULL
##
## $PRDX1
## NULL
```

There is no significant interaction between treatments and cell lines.

- ACO1 is significantly over-expressed in CTR compared to the other cell types except CTRPAt, similarly CTA is significantly under-expressed compared to CTRPAt.
- CAT is significantly over-expressed in CTR and CTRPAt compared to the other cell types except CTA.
- G6PD increasing increase according to the oncology of the cell, CTRPAt are significantly over-expressed compared to CT, CTA and CTR, similarly CTA is significantly under-expressed compared to CTR.
- LPCAT is significantly over-expressed in CTRPAt compared to CT and CTA and significantly over-expressed in CTR compared to CTA.
- NQO1 is significantly over-expressed in CTP, CTR and CTRPAt compared to CTA.
- NRF2 is significantly over-expressed in CTR and CTRPAt compared to CTA, similarly CT is significantly slightly over-expressed compared to CTR.

4.3 Conclusion

No significant interaction effect was found for any gene between cell type and treatment. This means that the values taken by one factor do not vary with the values taken by the other factor.

Session information

```
sessionInfo()
```

```
## R version 3.6.3 (2020-02-29)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 18.04.4 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.7.1
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.7.1
##
## locale:
##  [1] LC_CTYPE=fr_FR.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=fr_FR.UTF-8        LC_COLLATE=fr_FR.UTF-8
##  [5] LC_MONETARY=fr_FR.UTF-8    LC_MESSAGES=fr_FR.UTF-8
##  [7] LC_PAPER=fr_FR.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=fr_FR.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] RColorBrewer_1.1-2 cowplot_1.0.0      emmeans_1.4.8      car_3.0-8
##  [5] carData_3.0-3      MASS_7.3-51.6      nlme_3.1-147       lme4_1.1-21
##  [9] Matrix_1.2-18      PMCMR_4.3          kableExtra_1.1.0   data.table_1.12.8
## [13] forcats_0.5.0      stringr_1.4.0      dplyr_0.8.3        purrr_0.3.3
## [17] readr_1.3.1        tidyr_1.0.0        tibble_3.0.2       ggplot2_3.3.2
## [21] tidyverse_1.3.0
##
## loaded via a namespace (and not attached):
##  [1] fs_1.4.2          lubridate_1.7.9   webshot_0.5.2     httr_1.4.1
##  [5] tools_3.6.3       backports_1.1.5   R6_2.4.1          DBI_1.1.0
##  [9] colorspace_1.4-1  withr_2.1.2       tidyselect_1.1.0  curl_4.3
## [13] compiler_3.6.3    cli_2.0.1         rvest_0.3.5       xml2_1.3.2
## [17] sandwich_2.5-1    labeling_0.3      scales_1.1.0      mvtnorm_1.0-12
## [21] digest_0.6.23     foreign_0.8-76    minqa_1.2.4       rmarkdown_2.3
## [25] rio_0.5.16        pkgconfig_2.0.3   htmltools_0.4.0   dbplyr_1.4.4
## [29] rlang_0.4.6       readxl_1.3.1      rstudioapi_0.10   farver_2.0.3
## [33] generics_0.0.2    zoo_1.8-7         jsonlite_1.6      zip_2.0.4
## [37] magrittr_1.5      Rcpp_1.0.3        munsell_0.5.0     fansi_0.4.1
## [41] abind_1.4-5       lifecycle_0.2.0   stringi_1.4.5     multcomp_1.4-12
## [45] yaml_2.2.0        plyr_1.8.5        grid_3.6.3        blob_1.2.1
## [49] crayon_1.3.4      lattice_0.20-41   haven_2.3.1       splines_3.6.3
## [53] hms_0.5.3         knitr_1.27        pillar_1.4.3      boot_1.3-25
## [57] estimability_1.3  codetools_0.2-16  reprex_0.3.0      glue_1.3.1
## [61] evaluate_0.14     modelr_0.1.8      vctrs_0.3.1       nloptr_1.2.1
## [65] cellranger_1.1.0  gtable_0.3.0      assertthat_0.2.1  xfun_0.12
## [69] openxlsx_4.1.5    xtable_1.8-4      broom_0.5.6       coda_0.19-3
## [73] survival_3.1-12   viridisLite_0.3.0 TH.data_1.0-10    ellipsis_0.3.0
```